2019

Predicting Violent Crime Reports from Geospatial and Temporal Attributes of US 911 Emergency Call Data

Vincent Corcoran
Technological University Dublin

Recommended Citation

Corcoran, V. (2019) Predicting Violent Crime Reports from Geospatial and Temporal Attributes of US 911 Emergency Call Data, Masters Thesis, Technological University Dublin.

Predicting Violent Crime Reports from Geospatial and Temporal Attributes of US 911 Emergency Call Data

Vincent Corcoran

A dissertation submitted in partial fulfilment of the requirements of

Technical University Dublin for the degree of

M.Sc. in Computing (Data Analytics)

20'th September, 2019

Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technical University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institutes guidelines for ethics in research.

Signed: Vincent Corcoran

Date: 15/6/2019

Abstract

Increasingly, US police forces are overwhelmed with the volume of 911 Request for Service calls. However, the majority of these 911 calls do not result in any crime reports, as the police officers who respond to the calls often determine that no crime has occurred. Furthermore, no analysis is performed to determine which 911 calls result in crime reports. The aim of this study is to create a model to predict which 911 calls will result in crime reports of a violent nature. Such a prediction model could be used by the police to prioritise calls which are most likely to lead to violent crime reports. The model will use geospatial and temporal attributes of the call to predict whether a crime report will be generated. To create this model, a dataset of characteristics relating to the neighbourhood where the 911 call originated will be created and combined with characteristics related to the time of the 911 call. Geospatial and temporal analysis of past 911 calls and crime reports will be applied to determine which 911 calls resulted in crime reports (the dependent variable) so that supervised learning can be performed.

The primary aim of the study will be to investigate if characteristics relating to the neighbourhood where the 911 call originated, for example social, demographic or economic characteristics, combined with attributes related to the time of the call such as time of day, can be used as predictors in the model. This study will apply geospatial analysis using census data and open street maps to determine the characteristics of the neighbourhoods where the 911 calls originated. Temporal analysis will be used to extract data relating to the time of the 911 call, for example weather at the time of the call. Finally, machine learning will be used to predict if crime reports are associated with each 911 call.

Keywords: *Crime Prediction, Criminology, 911, Geospatial Analysis, GIS, Temporal Analysis, US Census, Demographics, Open Street Maps, Supervised Learning, Machine Learning, Support Vector Machines, Random Forests and Logistic Regression*

Acknowledgments

I would like to thank my supervisor Dr John Gilligan for his help, guidance and expertise throughout this project. I would also like to acknowledge the help of Deirdre Lawless who provided me with prompt and detailed responses to several statistical queries I had during the project. I would like to also mention Mark Foley who sparked my interest in the area of Geographical Information Systems.

I would like to acknowledge the assistance of law enforcement in Michigan, in particular, Ms Diana Flora, Executive Manager of the Crime Intelligence Unit for the Detroit Police Department and also Ms Wendy Easterbrook, Department Crime Specialist, Michigan State Police.

Finally, I would like to acknowledge the support of my wife, Linda, and our children who showed great patience whilst I was busy with my studies.

Contents

List of Figures

List of Tables

List of Acronyms

AUC	Area under the Curve
CFS/RFS	Call/Request for Service
DPD	Detroit Police Department
EMS	Emergency Medical Services
FBI	Federal Bureau of Investigation
GEOID	Geographic Identifiers
GIS	Geographical Information System
GPS	Global Positioning System
OSM	Open Street Maps
PCA	Principal Component Analysis
POI	Points of Interest
PBF Format	Protocolbuffer Binary Format
QGIS	Open Source Geographical Information System
ROC	Receiver Operating Curve
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
UCR	Uniform Crime Reporting

Chapter 1

Introduction

The three digit telephone number 911 is designated in the US as the *"Universal Emergency Number"* [1] for citizens requiring emergency assistance when there is a threat to life or property. This number connects the caller to an emergency dispatcher who based on the nature of the reported emergency as provided by the caller, may invoke support from one of three primary emergency services; Police, Fire or EMS (Emergency Medical Services). When police response is required, the dispatcher decides the priority of the call as per a localised police priority system, and the call is then assigned to the appropriate available police personnel which maybe administrative for low priority calls. This document, when referring to 911 calls, specifies calls which have been determined by the dispatcher as calls requesting police assistance.

The objective of this study is to create a predictive model of crime reports, of a violent nature, using 911 call data from Detroit city. The model will use geospatial and temporal attributes of the 911 call to predict whether a crime report of a violent incident will be generated. Violent and non-violent crimes are categorised by the attribute *offence description* in each crime report. Crime reports which report violent incidents, for example *assault and battery*, are filtered from the crime report dataset during data preparation and are listed in figure 4.20. Henceforth in this document a violent crime report will refer to a crime report reporting a violent incident.

This paper starts by researching the sociology of crime looking for factors which have

[1] National Emergency Number Association www.nena.org

been linked to crime and which can potentially be linked to 911 calls and then included as predictors in the model. Based on this research, various geospatial and temporal characteristics which can be related to the 911 calls are identified as potential indicators of future crime reports. Some of these features can be extracted from the 911 call dataset, whilst others will be extracted from US Census Data, US Meteorological Data, US Federal holidays and Open Street Maps.

A large part of the project will involve the creation of geospatial and temporal datasets which can be related to each 911 call. Physical features and demographic characteristics relating to each neighbourhood where the 911 calls originate will be extracted from US Census Data. An algorithm will be developed to extract point of interest (POI) data from Open Street Maps creating a geographical grid of Detroit with an index for each cell indicating the number of nightclubs and bars in that cell and neighbouring cells. Crime hotspots will be created by geospatially aggregating historical crimes. Temporal features related to the time of the 911 call such as the weather, time of day and whether it is a holiday or weekend will be determined. The final dataset will be tagged to indicate if a violent crime report is associated with each 911 call by cross referencing more than 600,000 911 calls against 61,000 violent crime reports using temporal and geospatial proximity to determine if they are related.

Models will then be created using support vector machines, random forests and logistic regression and validated against unseen test data.

1.1 Background

A major challenge to society is the reduction of crime due to its significant economic and social cost. Economic costs can include damage to or loss of property, policing, criminal judiciary and penal costs or reduced business due to the fear of potential crime. Social costs include the human effects for victims of crime and the lost potential of young people who become involved in crime.

Sociologists in the research of crime have been able to show correlation between crime and a variety of features in the neighbourhood such as demographics and physical environmental factors but they are unable to draw causal relationships between such features and crime. With the advent of machine learning methods capable of analysing large amounts of data, researchers have applied predictive analytics using these features to create models capable of predicting crime, to better target resources at reducing and stopping crime.

A *Request for Service Call (911)* in the US is a call made to an emergency operator requesting emergency assistance. In the case of calls requesting police assistance, the 911 call reports suspected criminal activity. A crime report is an official record of a criminal act. Either one of these two incidents can occur without the other and police forces typically do not look for relationships between the two incidents. Most 911 police calls do not result in crime reports, there can be 10 times as many 911 calls as crime reports. It is proposed to examine if a model using 911 call data to predict Violent Crimes can be improved by adding data from heterogeneous source related to the location and time of the 911 call.

1.2 Research Problem

Currently police response to 911 calls is determined by the priority of the calls as assigned by the 911 emergency dispatcher, based only on information obtained from the caller. The objective of this research is to investigate if it is it possible to improve on this by using supervised learning to build a model with 911 call data to predict if a violent crime report will result from that 911 call.

To use supervised learning, the dependent variable, in this case violent crime report *(yes/no)* must be known. This will be determined by relating 911 calls to crime reports by their geospatial and temporal proximity, i.e. if a crime report is created close to the location and time of the 911 call, then the 911 call will be considered to have resulted in a crime report. A model can then be created using only 911 call data to predict the dependent variable, violent crime report.

Expanding on this, it will be investigated if other features, related to the time and neighbourhood of the 911 call, can be found and included in the model to improve the prediction of violent crime reports. Such features can include data on past crimes in the neighbourhood, characteristics of the neighbourhood, for example demographics, weather conditions at the time of the call or characteristics relating to the time of the incident, for example Saturday nights after bars and nightclubs close. It is hoped that the model could be used as an aid when deciding police response to 911 calls.

Figure 1.1, shows the locations of 911 calls (black dots) and crime reports (red diamonds) for one random day in down town Detroit city. As can be seen there are many more 911 calls than crime reports, and they are rarely co-located in exactly the same location. Significant processing using geospatial and temporal analysis will be needed to link 911 calls with crime reports.

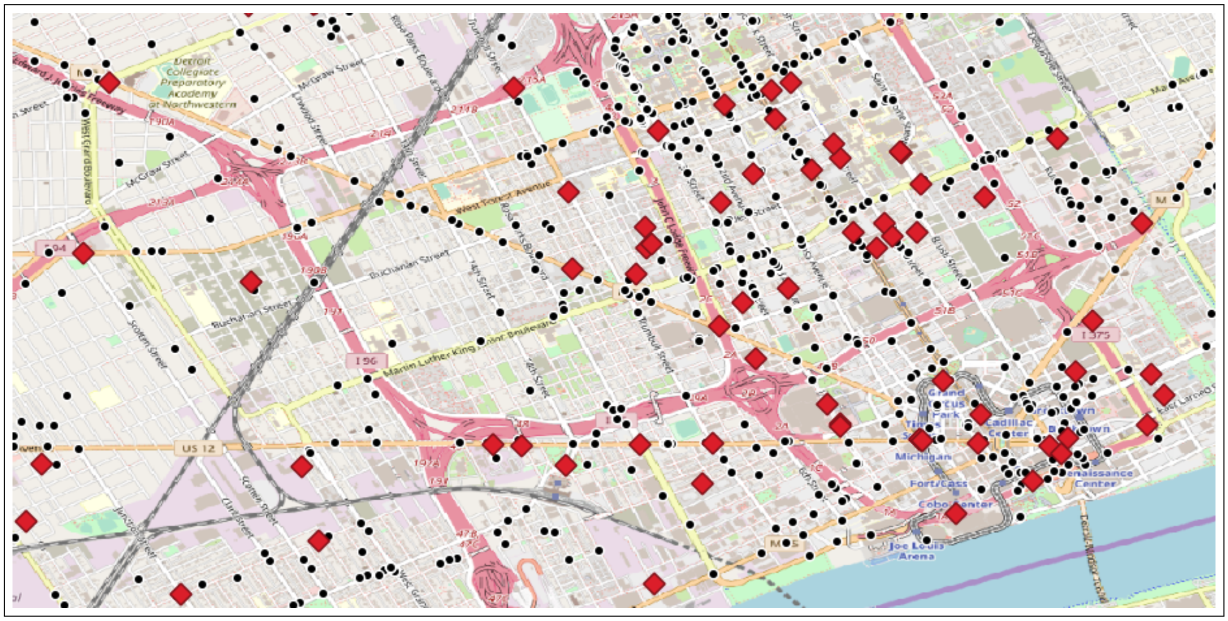


Figure 1.1: A Random Day in Detroit, 911 Calls and Crimes Reports

1.3 Research Question

Can a Machine Learning model to predict Violent Crime Reports using 911 Call Data be improved by adding Geospatial and Temporal Features related to the location and time of the 911 call?

1.3.1 Research Hypotheses

H_0: *The Sensitivity of a Machine Learning Model to predict Violent Crime Reports using 911 Call Data will not be improved by adding geospatial and temporal data from heterogeneous data sources, related to the location and time of the 911 call.*

H_a: *The Sensitivity of a Machine Learning Model to predict Violent Crime Reports using 911 Call Data will be improved by adding geospatial and temporal data from heterogeneous data sources, related to the location and time of the 911 call.*

1.4 Research Objectives

Crime prediction is a binary classification problem so there are a variety of algorithms which can be applied. It is proposed to use supervised learning to construct several models to predict violent crime as defined by a crime report of a violent incident.

1. Analyse the research into 911 calls and violent crime, with the aim of identifying potential causal factors, their associated data and the relationship between violent crime and 911 calls.
2. Filter DPD crime reports to remove non-violent crime reports.
3. Determine a mechanism using temporal and geospatial analysis to link past 911 calls and violent crime reports to produce a tagged dataset which classifies the 911 calls by including a new binary dependent variable which indicates if the 911 call is associated with a crime report or not.
4. Splitting the dataset into training and test data, build models to predict the creation of violent crime reports based only on 911 call data and apply these models to previously unseen test data.
5. Create a geographical dataset with characteristics related to neighbourhoods in Detroit and a temporal dataset with holiday and weather data. Expand the 911 dataset to include these geospatial and temporal features related to the neighbourhood and time of the 911 call, to create a dataset that can then be used to build additional models to predict crime reports.

6. Apply these models to test data, experiment with different sampling ratios and compare the sensitivity from each experiment.
7. Using k-fold cross validation compare the best performing algorithm and sampling ratio.
8. Accept or reject the null hypothesis.

1.5 Research Methodologies

This research will be quantitative in nature with the objective of creating a mathematical model by applying quantitative research to construct a solution to an existing problem using existing data. Reasoning will be deductive in that the ability of geospatial and temporal to improve a model to predict violent crime reports will be explored. A literature review will be undertaken with the aim of identifying suitable features which can be combined with 911 call data in predicting crime and examining the various methods previously used to predict crime. It is expected that data preparation will form a significant part of the research with geospatial analysis a key component. The implementation and experimental phase consists of the following steps:-

1. Identify and source data which supports the research.
2. Prepare and integrate the data into a dataset which can be used for modelling.
3. Identify correlated features which can be removed from the model.
4. Perform principal component analysis to further reduce the dataset whilst still explaining the variance.
5. Apply sampling methods to address the expected class imbalance.
6. Create the models and apply the models to test data.
7. Compare and evaluate the performance of the models.

1.6 Scope and Limitations

Several sources for 911 and crime data in the US were analysed with Detroit having the most complete available 911 data so Detroit city will be the subject of research for this project. The scope of the analysis will be focused only on predicting violent crimes, a crime being defined by the creation of a crime report. It will be investigated if it is possible to improve the predictive power of 911 data in predicting violent crime reports using other publicly available data.

A major limitation to the study is the lack of research into the relationship between 911 calls and crime reports. Further, internal police documentation on the 911 call handling process from the Detroit police was difficult to find but some references to it were sourced. Documentation from other police forces was found but the handling of 911 calls, in particular how calls are prioritised, is not consistent across police forces in the US. During the research it was discovered that the 911 data for Detroit includes administrative (non-emergency) data relating to police officers. This can be completely removed by removing all calls initiated by the police, however an attempt was made to keep the police initiated calls by filtering out administrative call data from the dataset. There will be an element of bias in constructing the historical crime hotspots as the same source as used for crime reports is used. It was originally intended to use FBI data but a query to Michigan State Police confirmed that the local Police Department (in this case Detroit PD) are ultimately the source of FBI crime data (personal communication, March 18, 2019).

1.7 Document Outline

Chapter 2 - Literature Review: This chapter starts by reviewing the 911 call for service emergency process. It analyses in depth the geospatial and temporal factors which have been linked to crime, focusing on violent crime. It examines crime prediction techniques used in the past in particular hotspot techniques using geospatial analysis. It examines how these techniques have been combined with machine learning methods to predict crime. It addresses the common problem of imbalanced data which will be an issue when linking 911 calls to crime and concludes by looking at the research that has been done in linking 911 calls and crime.

Chapter 3 - Design and Methodology: This chapter describes how the project will follow the CRISP-DM methodology to systematically explore, process and extract features from the various data sources. It describes key concepts relating to geospatial analysis, types of machine learning algorithms and various analytical techniques such as PCA that will be used during the project. It gives a brief description of the various tools that will be employed. It then describes the metrics and statistical analysis will be used to determine the outcome of the research question.

Chapter 4 - Implementation and Results: This chapter is the main body of the project. It starts by describing the various data sources to be used and the features which are associated with each source. The data sources are explored, described and then processed to extract feature which can be combined into datasets to be used for machine learning. The significant effort that was required to relate the 911 calls with the crime reports and to extract POI information from Open Street Maps is described in detail. The processing of the other data in particular, US Census data is described. Dimension reduction using PCA is used to prepare the dataset for modelling. Finally, a variety of models are built and the best two performing models are experimented with further using k-fold cross validation. The results are statistically analysed to determine the outcome for the research question and a conclusion on the hypothesis is reached.

Chapter 5 - Conclusion: This chapter reviews the research performed and the reasons for the research, that is the problem being investigated. It briefly describes the work which was undertaken to address the problem and then lists the contributions from this research. The document concludes by suggesting future areas which could be explored.

Chapter 2

Literature Review

This chapter reviews literature on the 911 call for service emergency process and the relationship between crime and 911 calls. It analyses in depth the geospatial and temporal factors which have been linked to crime, focusing on violent crime. It examines crime prediction techniques in particular hotspot techniques using geospatial analysis and how these techniques have been combined with machine learning methods to predict crime.

2.1 Introduction

It is estimated [1] that there are 240 million calls made to 911 in the US every year requesting emergency assistance from police, fire department or medical emergency services. A 911 call may require more that one emergency service, for example a call reporting an assault may require police and medical assistance but this paper will only consider 911 calls which have been routed to the police. This chapter will start by examining 911 police calls and look at how they are handled by emergency dispatch and the police. It will examine the reasons for the higher number of 911 calls compared to crime reports to try and understand which 911 calls are likely to lead to crime reports. It will examine the factors which sociologists consider when undertaking a study of crime, the features that contribute to crime, the various types

[1] National Emergency Number Association www.nena.org

of crime, temporal and geospatial variation in crime patterns, environmental factors which impact crime, attitudes towards the police and variations across society in the reporting of crime. In the review of the literature that follows, researchers have linked crime to a number of geospatial and temporal factors. The chapter will conclude by examining the application of predictive analytics to crime.

2.2 911 Police Request for Service Calls

The 911 Request/Call for Service was set up in 1967 for citizens to call when they required emergency police, medical or fire department assistance. The system was later improved to *Enhanced 911* whereby the system routed calls automatically to the nearest dispatcher based on the caller's location. However in the case of police 911 calls, as will be shown later, most of these calls do not result in crime reports. So, what exactly is a 911 call, how is it handled and what is the relationship between 911 calls and crime reports?

Emergency Dispatch

In his book on violent crime victims, (Turvey, 2013) examines the emergency response system and the challenges faced by emergency dispatchers when dealing with 911 calls. Dealing with calls requiring assistance from the police, fire department and medical services, the dispatcher is expected to quickly triage these calls. Ultimately the dispatcher taking the call decides on the priority of the call based on what they are told by the caller, so this will impact on whether the caller receives immediate assistance. For example (Turvey, 2013) describes how in Chicago lower priority crimes such as vehicle theft are now transferred to police officers on light duties who then create crime reports over the phone. (Turvey, 2013) points out the three main challenges faced by 911 dispatchers are personnel, budgets and false alarms; all of which will have an impact on what happens after a 911 call is made.

Response Times

A key metric that the police are expected to perform well on is 911 response times. In fact when declaring bankruptcy in 2013, Detroit cited it's 58 minute average in responding to 911 calls as a factor (Bialik, n.d.). A subsequent DPD plan of action in 2014 [2] to improve police performance explicitly specified measures that would improve 911 response times. Since 2013, the DPD are responding to priority 1 calls faster (Wilkinson, n.d.), however a consequence of this is that response times for other call priorities have declined. Other than measurement metrics, two other important points are raised by (Vidal & Kirchmaier, 2015). Faster response times lead to increased citizen satisfaction and faster response time leads to an increase in crime detection, with (Vidal & Kirchmaier, 2015) stating that a 10% **increase** in response time **decreases** the chances of solving the crime by 4.7% particularly for violent crimes.

What is a Crime?

Crime statistics in the US are measured by the FBI using the Uniform Crime Reporting system (UCR) [3] which aims to provide a common framework when reporting crime, to improve the overall quality of data and aid comparison of crime between different police forces. However, as pointed out by (Klinger & Bridges, 1997) the use of crime reports in compiling crime statistics has been questioned by a number of criminologists who would prefer to use 911 call for service (CFS) data to generate statistics on crime. The proponents of this believe that 911 calls are a proxy for actual crimes and this theory is explored by (Asher, n.d.) who believes that 911 calls are better at measuring crime than crime reports. He believes it is possible to get more detailed information from the 911 call data and that the UCR is too slow to release the data. However, as Asher does admit, there is a large discrepancy between the numbers of 911 calls and crime reports, for example Chicago has a ratio of 10:1 for CFS data (911) to UCR (crime reports) (Asher, n.d.). A similar ratio was found in this research for Detroit. (Klinger & Bridges, 1997) contends that it would be incorrect to use 911 calls and

[2] http://www.justiceacademy.org/iShare/Library-StrategicPlans/DetroitPD–StrategicPlan.pdf

[3] https://www.fbi.gov/services/cjis/ucr

that the use of crime reports to measure crime should continue. In either case, as things stand a crime has only occurred when a crime report has been written and it is included in UCR statistics and not when a 911 call is made.

In the case of Detroit, it was confirmed by Ms Wendy Easterbook of the Michigan State Police (personal communication, March 18, 2019) that they do not collect or store 911 data and that the crime data they supply to the FBI is the crime reports as supplied by the DPD.

911 Call Priorities

Each police department has a criteria for prioritising 911 calls. The DPD like other police forces, class calls which are urgent or life threatening as priority 1. Their prioritisation is as follows [4]:-

1. An emergency; the perpetrator is still on the scene, emergency medical service is needed or evidence preservation is urgent
2. The situation is stabilized, but serious. The crime is in progress or has happened within 15 minutes; likelihood of apprehension is high.
3. Assistance is needed, but not urgent. The incident occurred less than 15 minutes earlier, such as a break-in.
4. Not considered serious. The incident is not in progress, occurring more than 15 minutes earlier.
5. Telephone Crime Reporting Unit (T.C.R.U.) handles these situations. The incident occurred more than 15 minutes prior, apprehension is unlikely, and damage is less than $10,000.

911 Temporal and Geospatial Attributes

Associated with each 911 call record is data relating to the time and location of the 911 call. Temporal data includes the time and date that the call was made and data relating to the time involved in handling the call, for example the response time

[4] https://www.securitymagazine.com/articles/82845-detroit-police-department-aims-to-prioritize-911-calls

and police time on the scene. Geospatial data on the origin of the call includes the neighbourhood, census block, council district, address and GPS coordinates.

911 to Crime Relationship

Crime reports often follow 911 calls, but a large amount of 911 calls do not result in crime reports and there are a number or reasons for this. In their book, (Turvey, Savino, & Baeza, 2017) describe the three main types of non-emergency 911 calls:-

1. Accidental
2. Nuisance
3. False Reports

(Turvey, 2013) speculates that up to 40% of 911 calls can be accidental such as accidental dials or miss-dials. These will not lead to a 911 call report but can congest the system. Nuisance calls are when the caller has misjudged the situation or the purpose of the 911 process or the caller is suffering from some form of impairment, for example alcohol. False Reports can be malevolent and made for a variety of reasons such as crime concealment by misdirecting police, establishing an alibi or as revenge.

(Klinger & Bridges, 1997) discusses the discrepancies in volume between 911 calls and crime reports in detail and they suggest several reasons why 911 calls made in good faith do not result in crime reports. The public are often not aware of what exactly constituents a crime or they may mislead or exaggerate the details to prompt a faster response. The legal characteristic of events may change between the time of the call and police arriving meaning that there is no evidence of a crime.

In conclusion, 911 callers may not be qualified to determine if a crime is happening or may just indicate that police presence is required and as stated by Ms Easterbrook of the Michigan state police (personal communication, May 21, 2019), *"trained police officers are best qualified to decide if a crime report is required"*. That said, it is possible that 911 calls can be a good indication that crime has taken place or is about to happen. For example, suspicious behaviour which is not in itself criminal could indicate a future crime.

2.3 Geospatial Factors relating to Crime

In the review of the literature that follows, it will be discussed how researchers have linked crime to a number of geospatial and temporal factors. It will be referenced how demographic and economic attributes such as gender distribution by age, race, education, income, employment and family structure, all characteristics of the population at neighbourhood level, may influence the amount of crime in the area. Attributes of the physical environment such as the number of bars, the amount of vacant and run down buildings have also been shown to be predictive indicators of potential crime. Even the ambient light conditions of the area play a role. For these reasons geospatial analysis of crime is a common approach in crime prediction. This data is not available directly from 911 call datasets and will be generated from other sources such as Open Street Maps and US Census Data through extensive geospatial analysis. Geospatial data mining is therefore a considerable part of this work.

Neighbourhood in this document will refer to the location from which the 911 call is made and as will be described later, it is defined as a US Census Tract.

2.3.1 Neighbourhood Economic Factors

Poverty

Poverty is generally considered to be one of the most significant drivers linked to crime and there are many research articles on the subject. (Sampson, Raudenbush, & Earls, 1997) established that poverty and deprivation lead to increased crime. What is perhaps less known is that certain types of crime are more often associated with poverty than others. (Hsieh & Pugh, 1993) found that assault and homicide was more likely to be associated with poverty than rape and robbery. (Lipton et al., 2013) when analysing block groups of Boston with the highest levels of violent crime, found that they were generally poorer and had a higher number of alcohol outlets.

Employment

It is generally accepted and the subject of several studies that lack of employment opportunities can encourage crime particularly among youths (Good, Pirog-Good, & Sickles, 1986). (Freeman, 1991) demonstrated that a large number of high school dropouts particularly amongst black students subsequently developed criminal records. Studies have also established that employment reduces the tendency of criminals to re-offend (Tripodi, Kim, & Bender, 2010).

Collective Efficacy

In the study of crime, collective efficacy refers to the ability of a community to control individuals and groups to prevent antisocial behaviour and crime. It is demonstrated by (Sampson et al., 1997) that financial investment (**home ownership**) and tenure *"promote collective efforts to maintain social control"* leading to decreased crime. Contrary to this they also noted that areas with concentrated **poverty** and concentrations of **immigrants** were less likely to have collective efficacy in addressing crime. When considering racial make-up of areas, poverty rather than race led to decreased collective efficacy with resource deprivation leading to decreased collective efficacy.

2.3.2 Neighbourhood Demographic Factors

Racial and Matriarchal Family Structures

There are a number of factors to be considered when studying different racial behaviours relating to crime. It has been documented that there are significantly higher levels of violence perpetuated by and on blacks compared to whites, for example blacks are six times more likely than whites to die by violence (Sampson, Morenoff, & Raudenbush, 2005). (Sampson et al., 2005) explores various theories to explain this, such as a higher number of single parent families and matriarchal family structures and lower socio economic status amongst black families. (Sampson et al., 2005) refers to statistics which indicate that violent crimes are higher in disadvantaged areas which also have large concentrations of ethnic minorities although he refers to a paper which suggests that bias in the criminal justice system may contribute to these statistics. (Sampson et al., 2005) concludes that some of the ethnic differences in crime can be explained by immigration status, marital status, residential tenure, education and neighbourhood environment.

Minorities and Trust in police

A topical subject is ethnic distrust of the police which may lead to under reporting of crime. (Kwak, Dierenfeldt, & McNeeley, 2019) built on a 1999 study by (Anderson, 2000) which suggested that racial distrust of the police reduced the likelihood of victim cooperation with the police. (Kwak et al., 2019) concluded that poor perceptions of the police can lead to significant decreases in crime reporting by black victims of crime. This distrust in the police could lead to 911 calls being made whilst crime is in progress in black areas but not followed up with by crime reports. It could also mean that an area could actually have high crime and high 911 calls but relatively lower crime reporting. (Hagan, McCarthy, Herda, & Chandrasekher, 2018) explores the confounding nature of race in predicting crime concluding that due to an absence of alternatives and despite perceived police ineffectiveness, minorities in disadvantaged areas continue to call 911. (Small, 2018) when considering the relationship between the police and African Americans, immigrants and other minorities, explored conscious cynicism about the police against an unconscious desire to be protected. (Desmond, Papachristos, & Kirk, 2016) found a temporary one year reduction in 911 calls from black neighbourhoods after a well-publicised alleged case of police misconduct.

Gender and Age

As shown by (Freeman, 1991) in a tight labour market, for less educated male youths crime is an attractive alternative. While acknowledging that social factors play a huge part in crime (Levitt & Lochner, 2001) examined the gender and age of individuals involved in crime noting that gender is a major predictor of crime and in particular for violent crime with males five times more likely to be arrested for violent crime compared to females. They also found that age is relevant, as an 18 year old is five times more likely than a 35 year old to be arrested for property crime and that criminal behaviour spikes in the teenage years before peaking at age 18.

Education

Several studies have looked into the relationship between education and individuals participating in crime, (Lochner & Moretti, 2004) found that education significantly reduces the likelihood of involvement in crime.

Population Density

Higher population density is often associated with higher crime density but as (Harries, 2006) point out, this relationship is generally moderated by socio economic status so for example a densely populated neighbourhood in a wealthy area would have increased guardianship and lower crime.

2.3.3 Neighbourhood Physical Environment

Environmental criminology a relatively new branch of criminology looks at environmental conditions as a factor in criminal behaviour. "*Environmental criminology is the study of crime, criminality, and victimization as they relate, first, to particular places, and secondly, to the way that individuals and organizations shape their activities geospatially, and in so doing are in turn influenced by place-based or spatial factors.*"[5] Environmental criminologists believe crime should be understood as a confluence of offenders and victims/targets at particular times and places and they look for crime patterns that they can be explained in terms of environmental influences (Wortley & Townsley, 2016). A new concept in criminology, *risk terrain modelling* examines how certain environmental features, such as bars, contribute to the risk of crime (Caplan & Kennedy, 2010).

[5] https://en.wikipedia.org/wiki/Environmental_criminology

Points of Interest (POIs) - Bars

The presence of facilities/points of interest (POIs) have also been linked to crime (Wortley & Townsley, 2016). It can be shown that crime can be clustered close to Bars, Restaurants, Retail and other forms of entertainment (Ingilevich & Ivanov, 2018). Anecdotally people would associate the misuse of alcohol with crime as alcohol can cause heightened aggression or reduced awareness leading to opportunities for criminals and many of the papers reviewed back up this belief. As far back as 1981, (Roncek & Bell, 1981) examined the effects of bars on crime in residential areas looking at two alternative theories, one being that areas without bars tend to be quiet or deserted making pavements unsafe or alternatively that bars encourage more people in the areas making them safer. However, (Roncek & Bell, 1981) concluded that the evidence supports the hypothesis that areas with bars will have higher crime levels. (Chainey, Tompson, & Uhlig, 2008) and (Belesiotis, Papadakis, & Skoutas, 2018) when constructing a crime hotspot map of London noted that there were clusters of crimes close to bars. (Lipton et al., 2013) when analysing violent crime in Boston found that there was a geospatial relationship between alcohol outlets density and violent crime in both the block group and neighbouring areas.

Empty Buildings and Urban Decay

Urban planners and sociologists have linked urban decay and vacant buildings to crime leading to many cities investing in programs to either rehabilitate or demolish vacant buildings. Detroit has a program for such demolitions [6]. The premise being that criminals can operate undetected in these empty buildings or that signs of neglect indicate that an area is unlikely to detect criminals. (Spader, Schuetz, & Cortes, 2016) looked at Federal housing revitalisation programs in three US cities targeted at neighbourhoods with high numbers of foreclosed and vacant properties to analyse if there was a decrease in crime after demolition/revitalisation. They found that there was a localised but short-term improvement in burglary and theft within 250 ft of

[6] https://data.detroitmi.gov/Government/Detroit-Demolitions-Map/79fx-a8xj

demolished properties. (Aliprantis & Hartley, 2015) examined the dispersion effect of crime to neighbouring areas after demolition of high rise public housing, finding that the increased crime in neighbouring areas was less than the reduction of crime in the area of demolition leading to a net reduction in crime. Of particular relevance to this study, (Larson, Xu, Ouellet, & Klahm IV, 2019) explored the effects of 9,398 demolitions in Detroit on various types of crime, between 2010 and 2014, see figure 2.1). Examining the effects on property crime, violent crime, drug crime and total crime, (Larson et al., 2019) noted a reduction in property and violent crime in block groups which had demolitions. They added the caveat that areas with the most demolitions were often poorer.

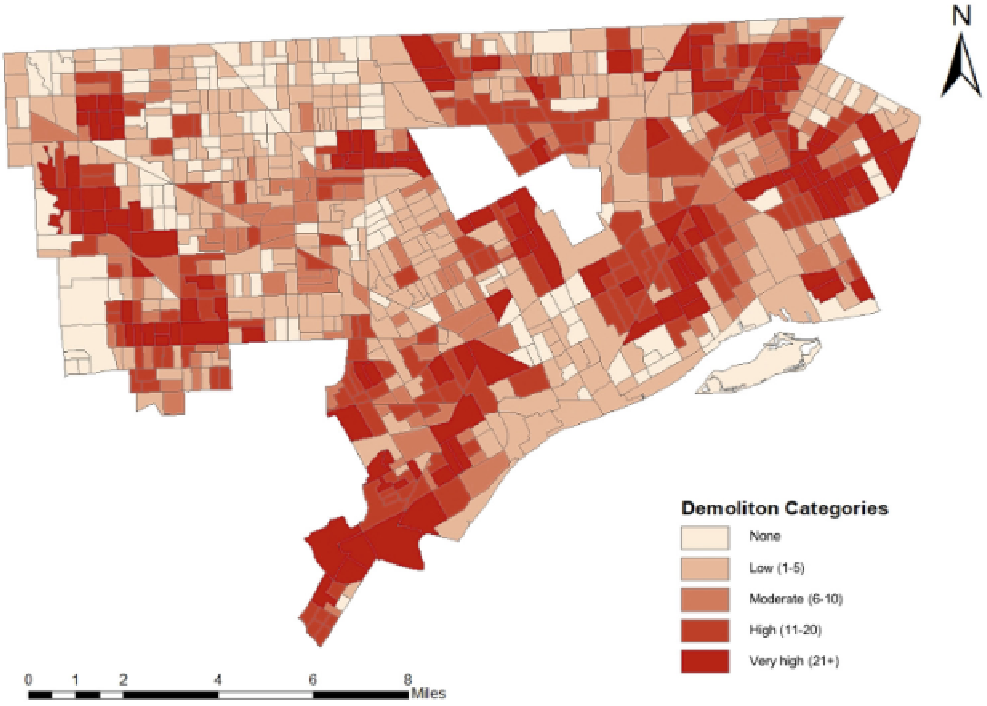


Figure 2.1: Frequency of Demolitions Detroit City

Street Lights

As (Xu, Fu, Kennedy, Jiang, & Owusu-Agyemang, 2018) state, crime patterns show correlation with environmental conditions and although not included in this study (Xu et al., 2018) examined the links between Detroit street lights and crime, concluding that better street lighting is associated with fewer crimes. They do however give the caveat that other variables such as the weather and alcohol outlets should be considered when undertaking geospatial analysis of crime.

2.4 Temporal Variations in Crime

Like geospatial data, temporal data also plays a part in predicting crime. Again, not all of this data is available directly in the 911 call dataset and will need to be sourced from external datasets such as the US Meteorological data or US Federal holidays. Beyond the common geospatial hotspot analysis, Ratcliffe (Ratcliffe, 2004) also considered the temporal aspects of crime within crime hotspots describing *spatio-temporal characteristics of hotspots* for example crimes in a shopping center car park during the day or as (Merlo, Hong, & Cottler, 2010) discusses crime around a football stadium during sporting events.

Weekend and Holidays

Building on the relationship between alcohol and crime, the relationship between binge drinking at **weekends** and holidays is well known and explored further by (Mäkelä, Martikainen, & Nihtilä, 2005) who concluded that death by intoxication peaked around weekends and during festive periods. Also, well-known is the increase in certain types of crime during **holidays** or at certain times of the day and much of the research reviewed notes a temporal aspect to variations in crime.

Weather

A less well-known relationship has been established between crime and the **weather**, in particular temperature. (Ranson, 2014) reported an apparently linear relationship between temperature and violent crime and suggested that higher temperatures may lead to heightened aggression. (Butke & Sheridan, 2010) also found higher levels of aggressive crime occurring during the summer months compared to the winter weather in Cleveland, Ohio but differentiated between crimes such as rape and homicide and aggravated assault speculating that hotter temperatures cause people to become aggressive but that more severe crime like homicide are driven by other motivations such as revenge or financial gain and so less likely to be influenced by the weather. (Horrocks & Menclova, 2011) applied research from the US and found that in New Zealand both temperature and precipitation are correlated with violent crime with violent crime decreasing during rain.

Time of Day

(Butke & Sheridan, 2010) when analysing crime in Cleveland noted the effects of temperature on crime were most prevalent around mid-day and early night (3 am). (Zhao & Tang, 2017) investigated temporal and geospatial correlations noting that crime numbers were correlated temporally with the day of the week or holiday periods.

2.5 Crime Hotspot Analysis to Predict Crime

Historically, crime hotspot analysis was used as a form of crime prediction model whereby police analysed historical crime data based on geography to identify areas prone to crime, the premise being that crimes are likely to re-occur in the same locations (Venturini & Baralis, 2016). As stated by (Chainey et al., 2008), crime tends to concentrate at locations due to the interaction between offenders and victims and the opportunities for offenders to commit crime. This type of intelligence led policing appeals to police management and local political leaders as it allows them to quickly determine areas in most need of police resources. Historically such analysis was performed manually using pins and maps but shortcomings in the manual collation of data led to the Illinois Criminal Justice Information Authority initiating the Geospatial and Temporal Analysis of Crime (STAC) program in 1987 [7] with the aim of identifying crime clusters.

[7] http://www.ncjrs.gov/App/publications/abstract.aspx?ID=105748

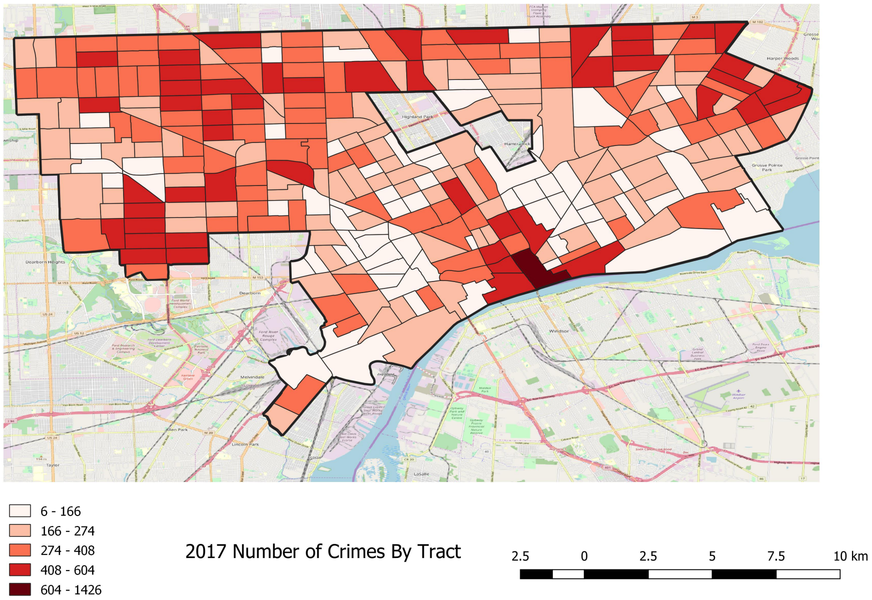


Figure 2.2: Detroit Crime Hot Spots - 2017

Geographic Information Systems

Geographical Information Systems, where crime addresses are geocoded to x and y co-ordinates, are now commonly used by police forces to produce choropleth maps which show the incidences of crime. For example see figure 2.2 produced as part of this research showing Detroit crimes for 2017 aggregated by census tract. Beyond the appeal of hot spot mapping to police, the use of hot spots is also employed by environmental sociologists who analyse geospatial distributions of crime and look for theoretical reasons behind higher incidences of crimes in particular locations. This analysis can then be applied to better understand offender behaviour.

2.6 Machine Learning to Predict Crime

In parallel with the social study of crime there is significant research into the use of modern information systems and predictive analytics as tools in combating crime. Combining traditional hot spot mapping with various machine learning methods using additional features to predict crime numbers at particular locations has become very popular.

An early (1997) and innovative approach to combining GIS systems with predictive analytics was explored by (Olligschlaeger, 1997). They looked at combining the Pittsburgh DMAP computer system, a fully integrated GIS, with 911 calls using a neural network to predict emerging drug selling locations at street level. While their results were encouraging they were limited by the then computing power available and they correctly predicted that future advances in computing power would increase the possibilities of deep learning.

Aspects to be considered when creating any prediction model are the features to be used, how to obtain the features and the type of model to be created. With crime, the potential features also depend on the type of crime as sociologists have demonstrated (Belesiotis et al., 2018) that features show differing correlation depending on the type of crime and that combining features derived from multiple heterogeneous data sources can improve crime prediction accuracy. Features often used include demographics for an area and other features such as the proximity to Bars, the weather, holidays, day of week and time of day. This analysis is performed using a variety of algorithms. (Marzan, Baculo, de Dios Bulos, & Ruiz Jr, 2017) geocoded and aggregated historical crime reports against a geographical grid of Manila to create hotspots, they then performed **association rules analysis** for each location to construct a model that could predict the type of crime under various circumstances on a particular day. (Belesiotis et al., 2018) when predicting geospatial crime in London combined several heterogeneous data sources and then used univariate feature selection to reduce the dataset before applying **random forest, support vector machines and ridge regression** against 14 different crime types finding that random forests gave the best results. (Bogomolov et al., 2014) experimented with a variety of classifiers, **logistic regression, support vector machines, neural networks, decision trees and random forests** to predict whether a particular area of London was likely to experience crime, finding that random forests yielded the best results. They used the smallest geographical area from the UK census and predicted whether a crime was likely to occur in the following month.

An unusual and innovative use of crime hot spots was attempted by (Stec & Klabjan, 2018) where they used **recurrent neural networks** (RNN) and **convolutional neural networks** (CNN) to predict crime in Portland and Chicago using a mix of crime data, census data and weather data. RNNs are typically used with temporal classification where what went before is relevant, for example with language modelling while CNNs are used in image recognition to look for feature/patterns in an image. They applied RNNs to very recent crime, speculating for example that pick pocketers in an area one day could indicate that a criminal gang is targeting that area so an RNN would have memory of the recent past. They split the city into a grid, with each cell equivalent to a pixel, containing census data relevant to that area of the city. They then applied CNN to the grid and combined the outputs of the RNN and CNN layers into a single network with the intention of capturing the temporal and geospatial elements of crime. While their recorded results were not as good as other papers, their models are certainly worth further study and refinement.

A weakness with traditional crime hot spot maps is that the hotspots are often created by aggregating crimes to administrative areas such as police precincts/beats or geographical areas such as census tracts meaning that crime hot spots will look different depending on the geographical areas chosen. Another issue which causes problems is when crime statistics are based on population numbers which will be low in commercial areas but high in residential areas so possibly indicating higher crime in commercial areas. (Ratcliffe, 2004) demonstrated a number of geospatial techniques to address these problems for example the use of kernel density surface maps where the area of study is divided into fine grids and crimes in each grid and in surrounding grids are counted to result in a hotspot surface map similar to a weather report. A similar method is used in this research to quantify the influence of bars and nightclubs to an area. (Corcoran, Wilson, & Ware, 2003) looked at creating clusters of crime that transcend predefined boundaries using geospatial analysis.

It may be that different algorithms have different results depending on the metric used. (Sohony, Pratap, & Nambiar, 2018) observed that **random forest** were more accurate in detecting normal instances, while **neural networks** were better at detecting fraud instances. **Logistic regression** was found by (Ingilevich & Ivanov, 2018) to give the best result compared to **linear regression** and **gradient boosting**.

The features used in crime prediction have evolved beyond the traditional predictors to include new types of features. Human mobility can now be tracked using mobile phones since mobiles are ubiquitous. (Bogomolov et al., 2014) used aggregated data from the mobile network to predict crime in London, stating that phones can be *"seen as sensors of aggregated human activity"*. (Belesiotis et al., 2018) also considered land use, transport and mobility data and unusually flickr images. They point out that different crime types occur in different spaces and are linked with different features. (Alves, Ribeiro, & Rodrigues, 2018) used a random forest algorithm to identify the most useful predictors then applied regression to predict the number of homicides in a Brazilian city for the following year. Of interest is their use of sanitation and literacy as predictors, features which would not be so useful in developed countries, illustrating that local knowledge can be important. (Ingilevich & Ivanov, 2018) investigating crime in Russia, included population density, the number of homeless people on the streets and even street lights but they point out that feature selection is important and is a factor in the performance of the model. (Stalidis, Semertzidis, & Daras, 2018) looked at using deep learning for crime classification and prediction by using deep learning to create hotspots. They suggest that deep learning can be superior to other machine learning algorithms, which are reliant on the experience of the analyst to prepare the data, because deep learning can extract the most useful features from the raw data. However logistic regression is also proficient at identifying features which contribute to a model. (Gerber, 2014) explored the use of another new source of data, Twitter, the usage of which has seen huge growth in the last few years. Tweets have the added benefit of being tagged with precise temporal and geospatial data, however their content is not so easily used due to misspellings and character constraints so semantic analysis is needed which leads to another important subject, feature extraction. Potentially valuable data is often not structured for ready use and requires processing to harvest the features. (Ghosh, Chun, Shafiq, & Adam, 2016) looked at extracting specific features from narrative crime reports using natural language processing and then applied a variety of algorithms to classify these reports.

2.7 911 Request for Service Calls

Compared to the research into crime there is relatively little research into the relationship if any between police request for service (911) calls and crimes reported. (Wu & Frias-Martinez, 2018) when examining the relationship between 911 calls and crimes, proposed pairing 911 calls and crimes based on geospatial temporal proximity. They then performed further analysis to look at reporting rates by 911 call per crime type. With the geospatial temporal parameters that they chose, they were able to match 35% of crimes and 7% of calls. Assuming their algorithm and parameters were correct, this would mean that 65% of crimes were not preceded by a 911 call and 93% of 911 calls did not result in a crime. Similar results are found in this study. Analysing by type of crime, they observed high 911 calls for shootings and lower calls for arson and rape. They also concluded that poorer neighbourhoods call 911 more frequently than wealthy neighbourhoods, but they didn't state how many crimes resulted from these calls.

Despite the small number of 911 calls leading to crimes, the police are expected to respond promptly to 911 calls and so they need to be resourced accordingly. With this in mind, (Chohlas-Wood, Merali, Reed, & Damoulas, 2015) applied hotspot analysis for New York 911 calls, looking for patterns in 911 calls and to predict resource requirements and potential emergency events.

2.8 Feature Selection and Correlation

As (Næs & Mevik, 2001) explain, collinearity between variables in a model can be a problem for prediction and classification and a common approach to address this is by using principal component analysis (PCA). For this project, as a first step the dataset will be reduced by removing features which are highly correlated with other features, but presented in different formats, for example percentages and absolute numbers.

Then PCA is applied to further reduce the dataset.

2.9 Imbalanced Data

Imbalanced data is a common problem in machine learning and occurs when there is a disproportionate ratio of each class. The problem is often seen in medical diagnosis, fraud detection, spam detection and with crime. With significant imbalance an algorithm can appear to have good accuracy simply by predicting the majority class each time. For this reason, with imbalanced data attention must be paid to the metric chosen to report on the model. Also, certain algorithms such as random forests have been shown to perform better. The main method to counter imbalanced data is to experiment with the sampling, for example oversampling the minority class and under sampling the majority class although this should only be done with the training data and not the validation or test data. Another technique proposed by (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) is to generate new synthetic data for the minority class using a method called SMOTE (Synthetic Minority Oversampling Technique). SMOTE is an alternative to oversampling with replacement of the minority class where synthetic samples of the minority class are created by taking samples of the minority class and then looking for its k nearest neighbours and creating the synthetic data points along the vector in the feature space between the nearest neighbours. Detecting credit card fraud is a good example of an imbalanced data problem. (Awoyemi, Adetunmbi, & Oluwadare, 2017) focussed on sampling techniques when a variety of machine learning techniques were applied to a large credit card dataset containing 284,807 transactions of which only 0.172% were fraudulent. The fraudulent data was over sampled and non-fraudulent data under sampled into 2 distributions, 34:66 and 10:90. Then naive bayes, k-nearest neighbour and logistic regression were performed. The models were measured using several metrics, k-nearest neighbour algorithm performed significantly better than the other algorithms across all the metrics except for accuracy with the 34:66 sampled data.

2.10 Conclusion of Chapter

This chapter reviewed the 911 process and determined geospatial and temporal features which have been related to crime. It looked at past crime prediction techniques using geospatial analysis and how they have been adapted to machine learning. Based on the research as described in this chapter, the following features relating to the neighbourhood of the 911 calls and features related to the time of the 911 calls can be used in the prediction models.

Geospatial Features

- Poverty Indicators
- Employment levels
- Median Income
- Home Ownership
- Rental Costs
- Vacant buildings
- Racial
- Immigrants
- Gender and Age
- Educational levels
- Population density
- Bars in Neighbourhood
- Past Crime Levels

Temporal Features

- Time of Day *
- Weekend *
- Holidays
- Weather Conditions

* Will be derived directly from the 911 Call Data Set.

The next chapter will examine the methodology and tools which will be employed during this project.

Chapter 3

Design and Methodology

This chapter describes the methodology followed and key concepts, in particular relating to geography, which are used during the project. It gives a brief description on the machine learning algorithms applied and the tools used during the project.

3.1 Introduction

The aim of this study is to build models to predict if crime reports will be created following 911 calls to the police. The models will be constructed using supervised learning first with only 911 data and the binary dependent variable *Crime Report* (Yes/No) and then by creating a second extended dataset by adding features from several disparate data sources linked to the location and time of the 911 calls. Features such as data on local demographics, the weather, vacant buildings and nearby points of interest such as bars will be added. This model can then be applied to future 911 calls to predict if a crime report will occur.

3.2 CRISP-DM methodology

The CRISP-DM (CRoss-Industry Standard Process for Data Mining) is a methodology, see figure 3.1 which provides a structured iterative approach to Data Mining Projects. The methodology describes a sequence of several distinct steps which can be performed and which may be repeated several times. This project generally follows this process with the exception of deployment which is outside the scope.

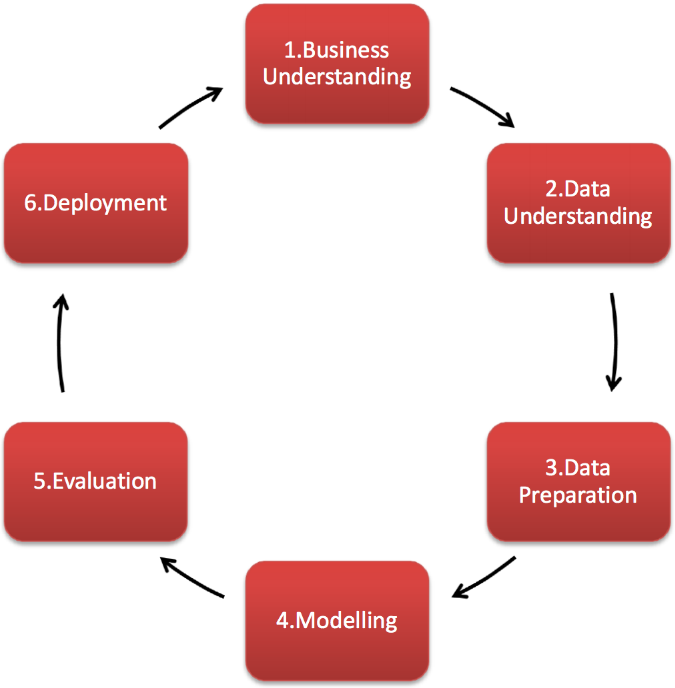


Figure 3.1: CRISP-DM Workflow

3.3 Detroit City

Detroit City, located in Michigan State in the mid-west of the USA, see figure (3.2), is a medium sized city with a population of almost 700,000, land area $360km^2$ and water area $10km^2$. It is located in Wayne County which is in the South East of Michigan. The city is located on the northern border of Wayne County and also shares a border with Windsor City in Canada. Detroit City totally encloses two other cities, Highland Park City and Hamtramck City. Detroit was chosen because of the good quality data made available by the city and because the size and demographic mix of the city were considered suitable for exploring the factors related to crime as identified in the research.

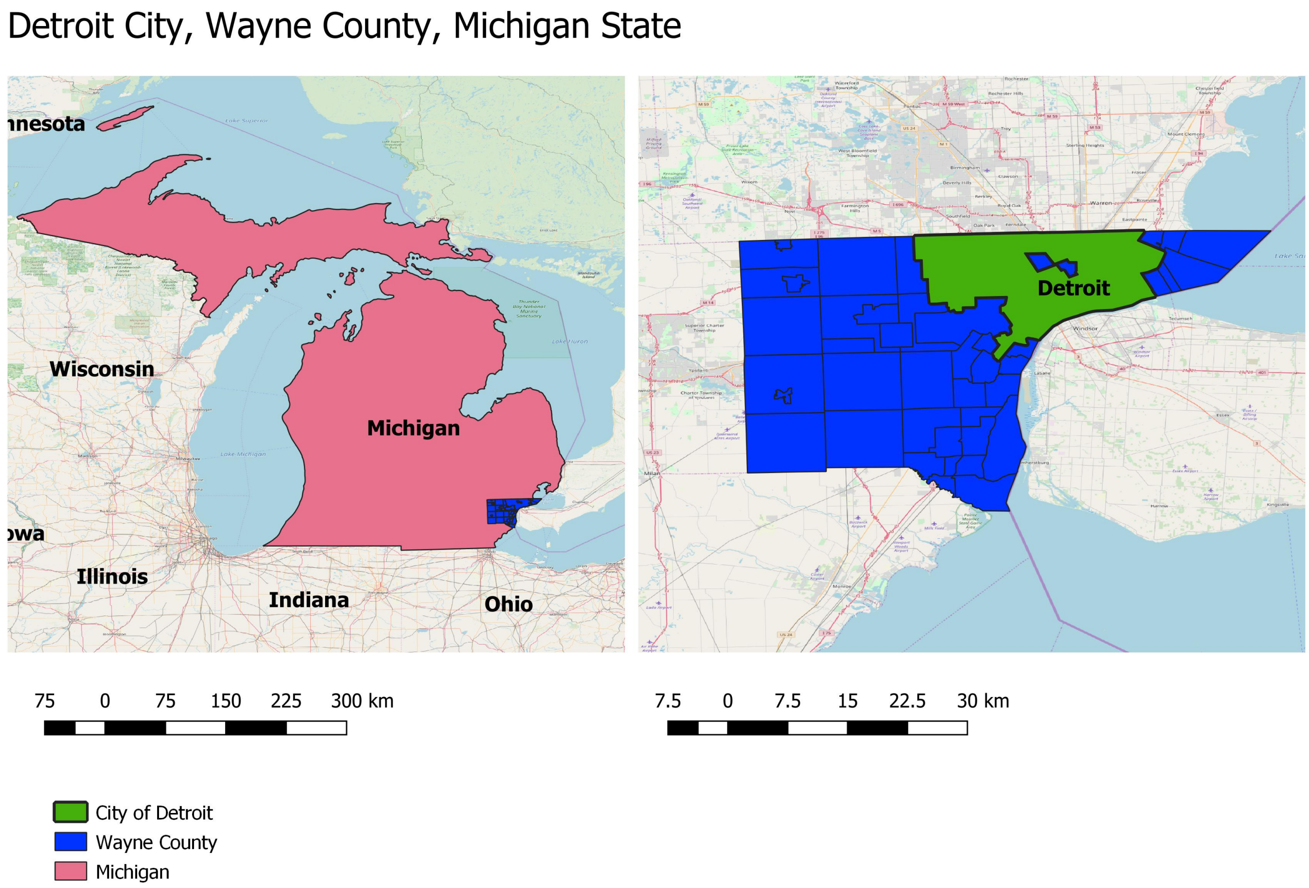


Figure 3.2: Detroit, Michigan

3.4 Key Concepts

3.4.1 Census Geography

Geography is a key component with census analysis as the data must generally be associated with an area in order to fulfil the main function of the census. Data is aggregated, tabulated and presented per geographical area. In the case of US Census data there are two types of geography, legal and statistical. Legal areas include state, local and tribal areas whilst statistical areas are defined by the US Census Bureau and used in order to support their functions in tabulating the data. A numbering system is used whereby a unique structured identification, called a GEOID, is used to identify each geographical area. The boundaries of these areas can change for a variety of reasons, so it is important to associate the correct geography for the relevant year.

Census Hierarchy

There is a hierarchy associated with geographical units with the largest being Nation and the smallest being Census Blocks, see figure 3.3. [1] Not all data is made available for the smaller entities as doing so could enable identification and breach confidentiality.

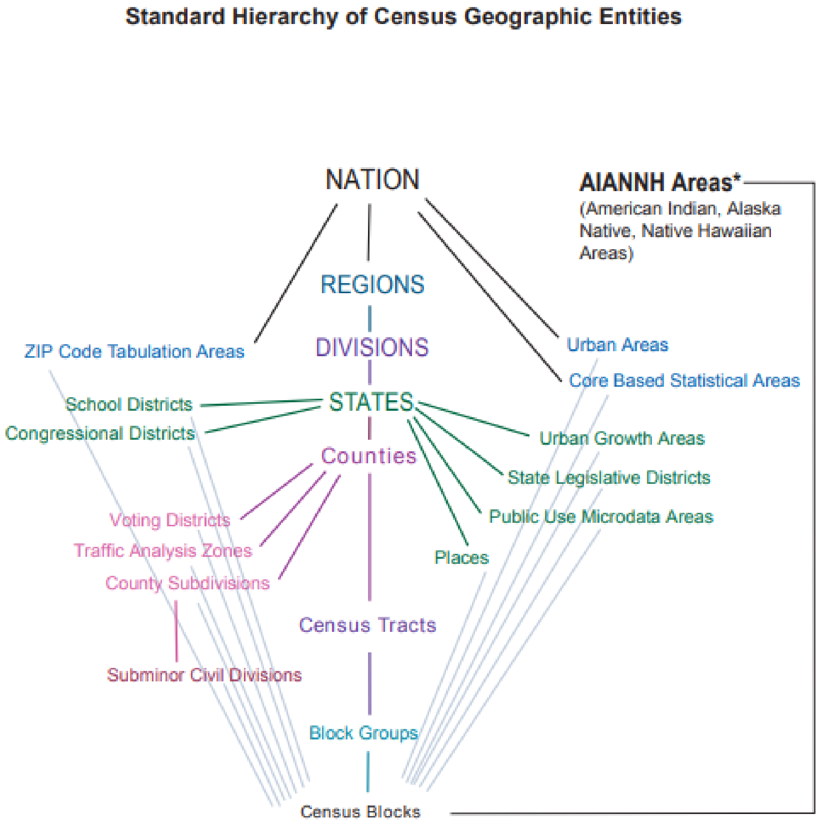


Figure 3.3: US Census Bureau Geographical Hierarchy

[1] diagram taken from US Census Bureau

Census Blocks

The smallest geographical unit which will be small in a suburban area but larger in rural areas. Often bounded by legal lines such as city or county lines, by streets or by geographical features such as rivers. The population is normally several hundred and there are more than eight million Census Blocks in the US and Puerto Rico. They never cross legal lines and are contained in Block Groups.

Block Groups

These are clusters of blocks and generally contain between 600 and 3,000 people with the exception of *special places* such as American Indian reservations, correctional institutions, military installations, college campuses and various types of residences such as care homes. Special places must contain a minimum of 300 people in order to be considered as a Block Group.

Census Tract

These generally contain between 1,500 and 8,000 people with an optimum number of 4,000. They are designed in order to be somewhat homogeneous in terms of the population's characteristics so that comparisons can be made from census to census. There are 60,000 Census Tracts in the US and Puerto Rico. Census tracts are defined with the aim of combining group characteristic for that area. For this reason, **census tracts** will be used in this research to measure the characteristics related to the neighbourhoods where the 911 calls occurred.

Geographic Identifiers

Geographic Identifiers (GEOIDs) see table 3.1 are used to identify geographical entities and are usually used for associating the objects with related data. In the case of US Census data, a number of Federal bodies are responsible for maintaining and determining the GEOIDs which are generated based on the hierarchy. As will be seen when exploring the 911 call and crime data, GEOIDs for census blocks are provided with the 911 and crime report datasets to identify the locations.

Area Type	**GEOID Structure**	**Number of Digits**	**Example GEOID**
Census Tract	STATE+COUNTY+TRACT	2+3+6=11	48201223100
Block Group	STATE+COUNTY+TRACT + BLOCK GROUP	2+3+6+1=12	482012231001
Block	STATE+COUNTY+TRACT + BLOCK	2+3+6+4=15	482012231001050

Table 3.1: US Census GEOIDs

3.4.2 Principal Component Analysis

Principal Component Analysis (PCA) is a mathematical procedure which can be used to reduce the number of dimensions in a dataset, which has a large number of features, by including as much of the information as possible in fewer variables. Often there can be correlation between features in a dataset meaning that there is redundancy of information in the data. PCA uses an orthogonal transformation to convert these possibly correlated features into sets of linearly uncorrelated variables called principal components, each orthogonal to the last where the first principal component has the largest variance in that it accounts for as much of the variance in the data as possible. Each principal component is a normalised linear combination made up from the original features in the dataset. For example, the data shown in figure 3.4 shows a PCA transformation of three dimensions to two dimensions using PCA. PCA will be used for dimension reduction, once the final dataset is created.

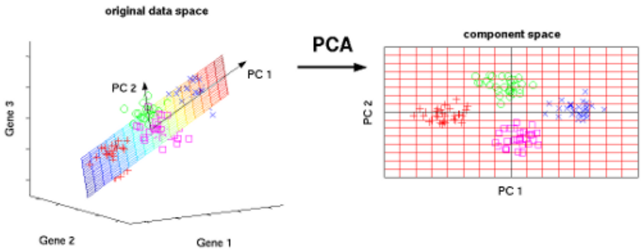


Figure 3.4: PCA Transformation. Source www.nlpca.org

3.4.3 K-fold Cross Validation

In supervised machine learning, the data is split into training and test data with the model built using the training data and then validated/tested against the test data. A potential flaw in this approach is that there may be interesting information in the data reserved for validation and so not included when building the model. One approach to overcome this issue, is k-fold cross validation where the data is randomly split into k-folds of data. The data is iteratively trained on k-1 folds of data and then validated against the unused k^{th} fold, see figure 3.5. The average of the metrics used to evaluate the model is the performance metric for the model.

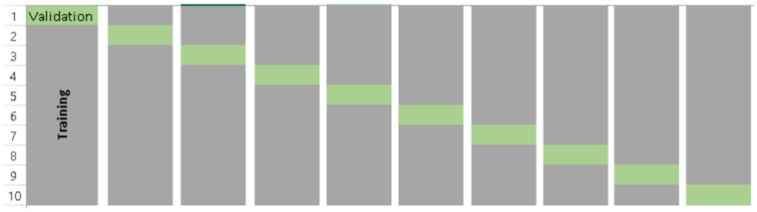


Figure 3.5: 10 Fold Cross Validation.

Stratified K-fold Cross Validation

An extension of k-fold cross validation is stratified k-fold cross validation where each fold is chosen so that it has a good representation of the data as a whole. For example in our imbalanced data, it is desired that each fold will have a similar ratio of the dependent variable *crime report*, see figure 3.6. Source of both diagrams [2]

[2] https://www.analyticsvidhya.com

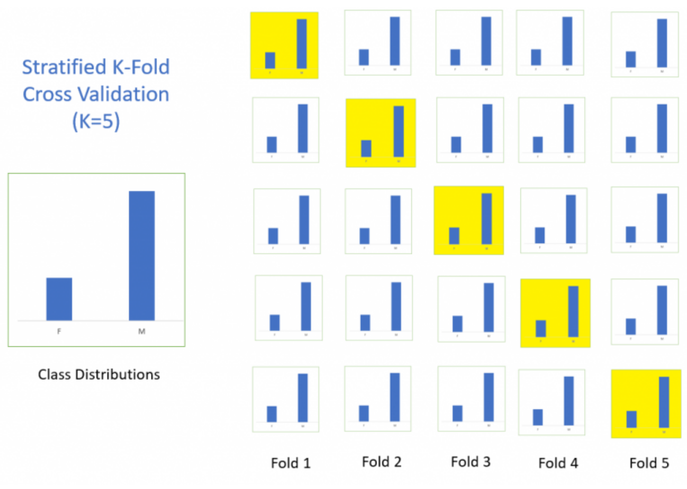


Figure 3.6: Stratified K-fold Validation.

3.4.4 Logistic Regression

Logistic Regression is a commonly used statistical method to predict a categorical dependent variable from one of more categorical and/or continuous predictor variables. Logistic Regression does not try to predict the actual value of the dependent variable but rather a probability based on the inputs. As (Kleinbaum, Dietz, Gail, Klein, & Klein, 2002) point out, logistic regression is a popular model because it is based on the logistic function,

$$f(z) = \frac{1}{1 + exp(-z)} \tag{3.1}$$

which outputs the result as show in figure 3.7, the output of which ranges in value between 0 and 1 but never reaching either. For this reason, it is suitable for binary classification.

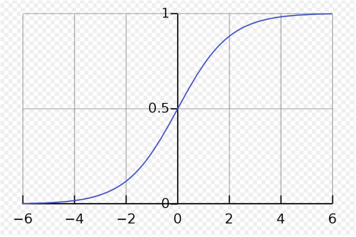
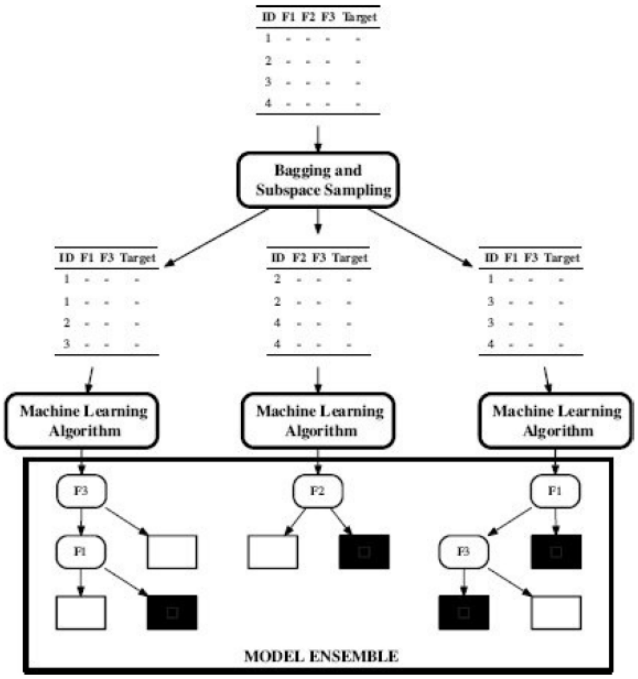


Figure 3.7: Logistic Function

3.4.5 Random Forests

Decision trees are a method for regression or classification where models are created in the form of a tree so that each branch of the tree breaks the data into into subsets based on the value of a feature in the data. The final result is the class of value for that subset of data. Random Forests introduced by (Breiman, 2001), are a form of ensemble learning using decision trees with bagging and subspace sampling. Ensemble learning, is a process where multiple models are created and applied to a problem, so that each model makes a prediction for the data by voting on the outcome, the majority vote being the overall result. With bagging, each of the models is trained on a sample of the training data which is created by randomly sampling the training data, with replacement. The final size of each of these training samples is equal to the training dataset so, because of the replacement there are duplicates in the training data. With subspace sampling different combinations of the features are used for each model. Figure 3.8, sourced from (Kelleher, Mac Namee, & D'arcy, 2015), describes the random forest process.



ID	F1	F2	F3	Target
1	-	-	-	-
2	-	-	-	-
3	-	-	-	-
4	-	-	-	-

ID	F1	F3	Target
1	-	-	-
1	-	-	-
2	-	-	-
3	-	-	-

ID	F2	F3	Target
2	-	-	-
2	-	-	-
4	-	-	-
4	-	-	-

ID	F1	F3	Target
1	-	-	-
3	-	-	-
3	-	-	-
4	-	-	-

Bagging and Subspace Sampling
Machine Learning Algorithm
Machine Learning Algorithm
Machine Learning Algorithm
F3
F1
F2
F1
F3
MODEL ENSEMBLE

Figure 3.8: Random Forest

3.4.6 Support Vector Machines

Support Vector Machines (SVM) are a kernel based supervised learning method for classification (Vapnik, 2013) and (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997). SVMs attempt to create the maximum decision boundary called a hyperplane between samples so the data can then be classified. For data which is not linearly separable, the data points are projected, figure 3.9, using a kernel algorithm into a higher dimensional feature space where they can then be linearly separated, figure 3.10.
Source of diagrams [3]

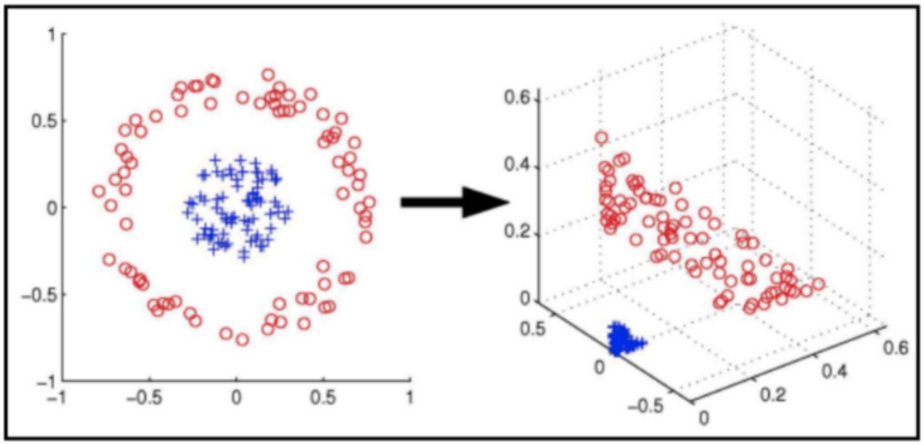


Figure 3.9: SVM Kernel Transformation

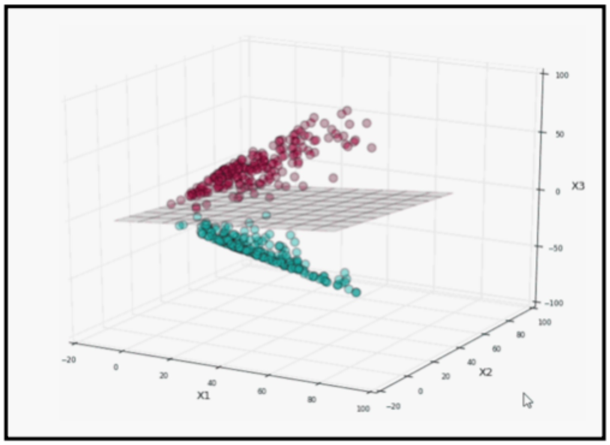


Figure 3.10: SVM Hyperplane in Higher Dimension Feature Space

[3] https://medium.com/@zachary.bedell/support-vector-machines-explained-73f4ec363f13

3.5 Methods and Tools

3.5.1 RStudio

RStudio, an open source development environment for R, is heavily used during the project for data wrangling, geospatial, temporal and statistical analysis. Many common libraries are used but of note are the libraries **caret** for classification and cross validation, **e1071** for support vector machines and **DMwR** for SMOTE.

3.5.2 Geographical Information System

A Geographical Information System (GIS) is a computer system designed to capture, store, manage, analyse and present geospatial/geographic data. Typically geospatial data is represented either by vector data such as shape files or by raster data which contains pixels. Each pixel represents data for a geographic feature. Various statistical data can then be associated with geospatial objects. There are a number of commercial and open source GIS systems available and QGIS is used for this project.
Note that shapefiles will be described in more detail in the next chapter.

TIGER

For the 1990 Census, the US Census Bureau in collaboration with the US Geological Survey, developed a nationwide map/format of the US and Puerto Rico called Topologically Integrated Geographic Encoding and Referencing (TIGER) database [4]. Using TIGER to link geospatial data with Census Data, the Census Bureau release a variety of tools and formats to access the Census Data. They release shapefiles associated with the various geographical objects without any associated statistical data. They also release several TIGER products per geographical object incorporated with detailed demographic and economic data.

[4] https://www.census.gov/geo/maps-data/data/tiger.html

3.5.3 Geospatial Measurement of Distance

To determine geospatial proximity between 911 call and crime report locations, it will be required to measure the distance between their GPS positions. To measure the distance between two GPS positions, it could be assumed that the earth is a perfect sphere and that both locations are at an equal distance from the centre of the sphere, that is they have the same altitude. Then, spherical trigonometry can be applied by using the **law of cosines** to calculate the distance. Rounding errors at small distances meant an alternative solution was to use the **Haversine formula**, considered more robust at small distances. Since the earth is not a perfect sphere but an imperfect ellipsoid shape, the latitude of the two points and indeed position above sea level has an impact on the calculations. For these reasons, the **Vincenty inverse formulae** were published in 1975, two related methods which are much more accurate and so widely used in geodesy (a branch of mathematics dealing with the shape of the earth). These formulae consider the earth to be an oblate spheroid and so consider the actual latitude of the positions to be measured.

Comparing all three methods using R Code[5] the distance between Dublin (53.349487, -6.260216) and Detroit (42.330358, -83.045648) is calculated, then the distance between two points in Dublin (-6.172660, 53.367175 to -6.223230, 53.371466). Examining the results in table 3.2, there is a variation of more than 10 km for the Dublin to Detroit calculations but only 11 meters within Dublin.

Method	**Dublin - Detroit**	**Within Dublin**
Haversine	5573.384 km	3.388823 km
Spherical law of cosines	5573.384 km	3.388823 km
Vincenty	5588.977 km	3.399795 km

Table 3.2: Distance Measurements using R

The Vincenty formula is the most accurate but more complex and there will be a large number of calculations (5 million) to make. Further since both locations have been

[5] R https://www.r-bloggers.com/great-circle-distance-calculations-in-r/

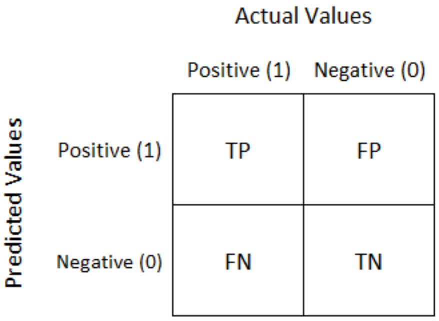
anonymised there will be a small inherent error anyway. So, for these reasons, it was decided to use the Spherical law of cosines when measuring the differences in position between 911 calls and crime reports.

3.6 Performance Metrics

With significant imbalance an algorithm can appear to have good accuracy simply by predicting the majority class each time. For this reason, with imbalanced data, attention must be paid to the metric chosen to report on the model. As noted by (Chawla et al., 2002), the cost of misclassifying the minority class can often be much higher than misclassifying the majority class. For example, incorrectly diagnosing no cancer is worse than incorrectly diagnosing cancer. In the case of crime prediction, if a model were to incorrectly predict that no crime was likely (a Type II error) after a high priority call was made to 911 and the police were slow to respond, then the model would rightly be considered to have performed badly.

The starting point when evaluating a classification model is the confusion matrix. In the case of binary classification, it is a two by two matrix, showing:-

- True Positives (TP), the model correctly predicted positive.
- False Positives (FP), the model incorrectly predicted positive.
- False Negative (FN), the model incorrectly predicted negative.
- True Negative (TN), the model correctly predicted negative.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 3.11: Confusion Matrix

Simple metrics that can be derived from the confusion matrix are the error rate which is the percentage of incorrect predictions, (false positives and false negatives) and the accuracy being the percentage of correct predictions (true positives and true negatives). Sensitivity/recall and specificity are metrics that focus on either the true positive rate or true negative rate. A potential use of this model if successful, would be the prioritisation of certain 911 calls over other calls so it is important to perform well in predicting crimes rather than predicting no crime. Like medical diagnosis, it is preferred to reduce the number of false negatives, so the best measure for this study is sensitivity and sensitivity is the metric chosen to judge our models.

$$Sensitivity = \frac{True\ Positives}{True\ Positives\ +\ False\ Negatives}$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives\ +\ False\ Positives}$$

3.6.1 Receiver Operating Curve

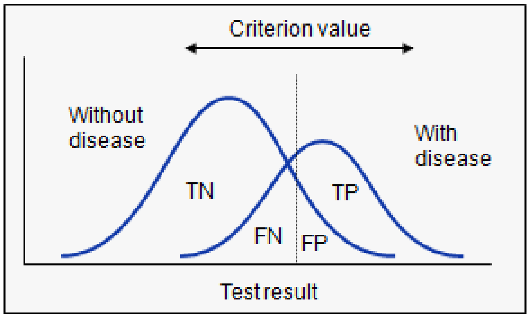
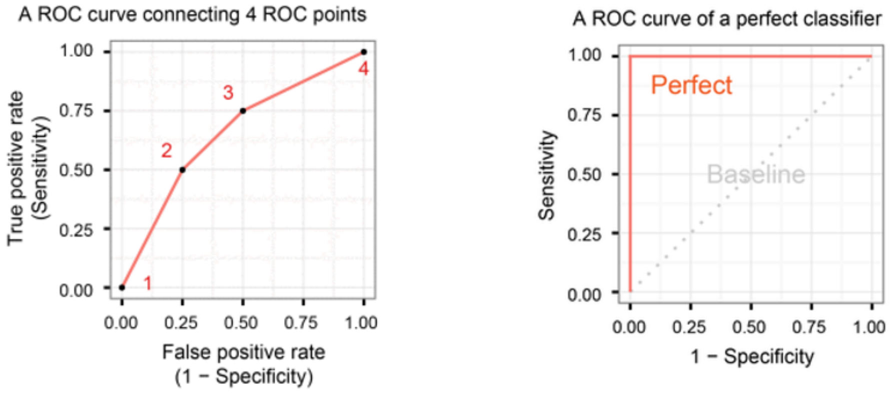


Figure 3.12: Distribution of Classification Predictions

Consider the graph in figure 3.12, source [6], which shows two distributions of people, those with a disease and those without a disease. Based on some test, the vertical

[6] https://www.medcalc.org/manual/roc-curves.php

line illustrates that a criteria or threshold is used to decide on a positive or negative result from this test. Samples which lie on the wrong side of the threshold will either be false positives or false negatives. A model with high sensitivity is best at avoiding false negatives, while a model with high specificity is best at avoiding false positives. By moving the threshold to the left, there will be an improvement in sensitivity but a decrease in specificity. Moving it to the right would improve specificity but decrease sensitivity. The ROC (Receiving Operating Characteristic) is a graphic designed to show this trade off, see figure 3.13. source [7] It plots sensitivity and specificity (1 - specificity) for different thresholds.



A ROC curve connecting 4 ROC points

A ROC curve of a perfect classifier

Figure 3.13: A ROC Curve

A perfect classifier would have 0 on the 1-specificity axis and 1 on the sensitivity axis. The ROC is measured using the Area under the Curve (AUC), with a range of AUC score between 0 and 1, although anything below 0.5 is considered poor.

[7] https://classeval.wordpress.com/introduction/introduction-to-the-roc-receiver-operating-characteristics-plot/

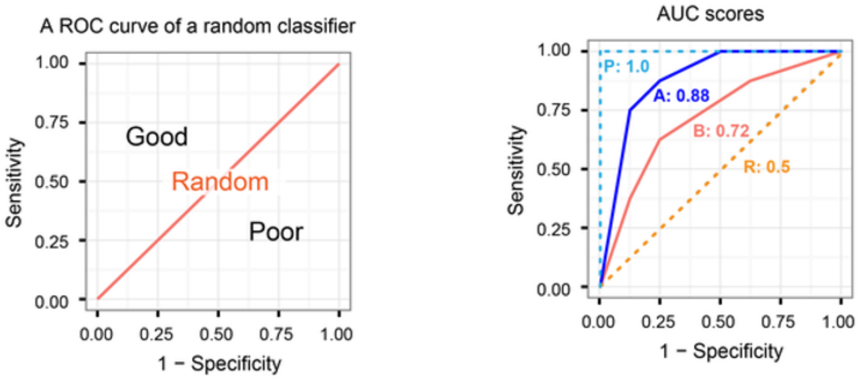


Figure 3.14: Area Under the Curve

3.6.2 Independent T-test

The models for both hypotheses with the highest sensitivity will be tested further using k-fold stratified cross validation to generate and test 10 further models (k=10). For each iteration of the validation, k-1 groups will be down sampled to reduce the imbalance and then used to construct the models. The models will be applied to the k^{th} sample and classification results recorded. This will be repeated k times and sensitivity will be calculated for each iteration.

The distributions of Sensitivity results for both datasets and for both models will be analysed using an independent t-test to compare the means of the various groups.

If there is a statistically significantly difference in group means indicating they are from different populations and there is a higher mean sensitivity for the extended dataset, then the null hypothesis can be rejected and the alternative hypothesis accepted that temporal and geospatial data related to the location of the 911 call improves the sensitivity of the model to predict violent crime.

3.7 Conclusion of Chapter

Using QGIS and R Studio, geospatial and temporal analysis will be applied to various datasets associated with Detroit to extract features associated with crime as identified in the research. These features will then be matched with 911 calls using temporal and geospatial attributes. Geospatial and temporal analysis will be used to relate 911 Calls with crime reports to determine if each 911 call resulted in a crime report. Models using Random Forest, Logistic Regression and Support Vector Machine algorithms will be created to predict crime and then tested using k-fold cross validation. Finally, a statistical analysis, using independent t-test, of sensitivity results will be used to prove or disprove the hypotheses.

As described in the literature review, a variety of algorithms have been tested to predict crime with no predominant algorithm. The algorithms chosen for this paper were chosen because they are all suitable for binary classification, i.e. a crime will occur or will not occur. Random Forests and Support Vector Machines will split the data into two distinct classes while in the case of Logistic Regression, the outcome will be a probability so that a threshold will be used to decide the outcome.

The next chapter will be the main body of the project. It will describe the data used and technical details of what was done with the data and the results of the modelling.

Chapter 4

Implementation and Results

This chapter is the main body of the project. It explores the various data sources used and describes how the features are extracted. It will describe in detail how the 911 calls are linked with the crime reports and how POI information is extracted from Open Street Maps. The processing of the other data in particular, US Census data is described. Dimension reduction using PCA is applied to the data prior to the creation and evaluation of several models. Further experimentation is performed using k-fold cross validation. The results are statistically analysed to determine the outcome for the research question and a conclusion on the hypothesis is reached.

4.1 Business Understanding

Public opinion of the police is often influenced by the perceived response to 911 calls while a key performance indicator for police is response time to priority 1 911 calls. However, due to a lack of resources, police departments have been forced in some cases to reduce the level of response to lower priority 911 calls, sometimes taking crime reports solely over the phone. A review of existing research shows that while there have been significant efforts to apply machine learning to predict crime, mostly the predictions are not in real time and at a high temporal or geospatial granularity. For example, predicting the number of murders in a city for the following year. Attempts at real time predictions have been based on using sources such as Twitter or by the use of proxies such as cellular telephony usage to indicate crowd movements. Research indicates that very little use is made of 911 call data, other than to track 911 response times. The purpose of this research is to determine if value can be extracted from 911 calls, a potentially data rich source of information, when combined with other relatively static/constant data related to the location and time of the 911.

It is important to note that while a positive correlation with actual crime reports is expected, a 911 call does not mean a crime has happened or will subsequently be reported. The State authorities do not collect, store or classify 911 calls and use only crime reports to report statistics on crime. This was confirmed by Ms Easterbrook of the Michigan State police (personal correspondence, March 18, 2019).

It is planned to use 911 call data combined with other data extracted from publicly available data sources as described below to create a real time prediction model which police forces could utilise to better allocate their resources. Features which have been linked to crime as described during the literature review in chapter 2, will be sourced from the following datasets.

4.1.1 911 Call Data and Crime Reports

The city of Detroit through it's Open Portal Initiative, makes available to the public data concerning City operations from several local government agencies [1]. Among the contributors to the open data are the Detroit Police Department (DPD). The DPD provide several datasets such as citizen complaints, information on precincts and stations and various datasets relating to crime such as crime reports and 911 request for service emergency calls. This data is available to download in csv format. There are also interactive maps available relating to this data.

Historical Crime Hotspots

For historical crime hot spots, it was originally planned to use FBI data, however Ms Easterbook of the Michigan State police (personal communication, March 18, 2019)

[1] https://data.detroitmi.gov/

stated that Michigan State police provide to the FBI, DPD crime statistics. For this reason, the original data from the DPD is used, but this is a potential source of bias.

Dependent Variable

The 911 data will be linked with crime reports to determine if a violent crime report resulted from the 911 call.

Hour, Day and Weekend

The time of day, day of week and if the 911 call occurred on a weekend will be extracted directly from the 911 call data.

4.1.2 US Census Data

The research has indicated that demographic and economic factors can lead to crime. The US Census Bureau provides detailed demographic and economic data for the US, so this will be used a source of data. The following features will be extracted from US Census data.

- The percentage of people receiving food stamps will be an indicator of poverty.
- Unemployment levels will be calculated from census features.
- Various measures of income will be extracted.
- The percentage of home ownership per neighbourhood.
- Mean rental costs per neighbourhood.
- The number of vacant buildings in the neighbourhood.
- Racial breakdown in terms of white, black and Asian.
- The number of immigrants and date when they arrived in the US.
- Various measures of gender and age.
- Educational levels.
- Population density.

4.1.3 Open Street Maps

Open Street Maps will be used to determine the concentration of bars in the neighbourhood.

4.1.4 Weather Data

The National Centers for Environmental Information will be used to source past weather in Detroit.

4.1.5 Holiday Data

A list of Federal holidays will be sourced from a US government website to determine if the 911 call happened during or immediately before a holiday.

4.1.6 Risks

There are two main risks relating to the data used. Both 911 call data and crime reports are anonymised to protect privacy. The algorithm used to associate 911 calls and crime reports will consider this to reduce the risk. The DPD 911 call dataset contains administrative data on police officer's activity as officers use the 911 system to log their activity meaning the 911 call data has non-emergency incidents included. An attempt to filter this data out was made but there could be some non-emergency call data not removed.

4.1.7 Terminology

The literature uses the terms *call for service* and *request for service* to indicate an emergency call. In this document 911 call is intended to refer to an emergency call requesting police assistance. A crime report which refers to a violent incident will be referred to as a violent crime report.

4.1.8 Benefits

Potential benefits of this project would be the ability to prioritise resources more efficiently based on the prediction model. Potentially this could lead to better resolution of crimes and better perception of police performance in dealing with crime.

4.2 Data Understanding

The following data sources will be used, and features identified by the research as linked to crime in chapter 2 will be extracted and then used to create the dataset used for the models.

4.2.1 911 Request for Service Calls

The DPD dataset of 911 calls requesting police presence will be used. This dataset also contains 911 calls which are officer initiated. The DPD makes available to the public via the Detroit open data portal [2], two datasets with 911 call data, an up to date dataset titled *"DPD: 911 Calls for Service, September 20, 2016 – Present"* which is continuously updated and an older dataset from 2016 with a slightly different format. The more recent dataset was downloaded on March 9, 2019.

911 Data Description

See **Appendix A** for the complete list of fields.
The 911 data has 25 fields including:-

- Unique incident ID, agency (mostly Detroit PD).
- Location of each 911 call by address, GPS co-ordinates and census block GEOID.
- Call codes, assigned priority, category and description of the call.
- Time and date of call.
- DPD specific data such as responding unit, precinct and response times.

[2] https://data.detroitmi.gov/

The initial dataset has 1.7 million rows each representing a 911 request for service call to the DPD with calls from September 20th, 2016 to March 2019. To reduce the amount of data and in order to get one full year of data, it was decided to include calls only from 2018. This resulted in a dataset of 800,226 calls.

Administrative 911 calls

There is a field, indicating if an officer initiated the 911 call and of the 800,226 calls from 2018, 478,884 (59.84%) of these 911 calls are officer initiated. There are also a number of fields (Call Code, Call Description and Category) which give further information as to the nature of the 911 Calls. An analysis shows that there are more than 100 categories of call type, several of which do not seem to indicate an emergency situation, for example *"START OF SHIFT INFORMATION"*.

This was checked with Ms Flora, DPD (personal communication, March 27, 2019), who stated that these calls are not emergencies, but administrative whereby the officers are logging their activity. However, there are also a number of call types where the officers are undertaking proactive actions which may result in a crime report.

911 Call Locations

Although not stated within the dataset, it was confirmed by Ms Flora, DPD (personal communication, March 5, 2019) that the 911 data is location anonymised to protect the identity of the callers. Further, as can be seen in figure 4.1, many of the 911 Calls originate in locations outside the Detroit City boundary.

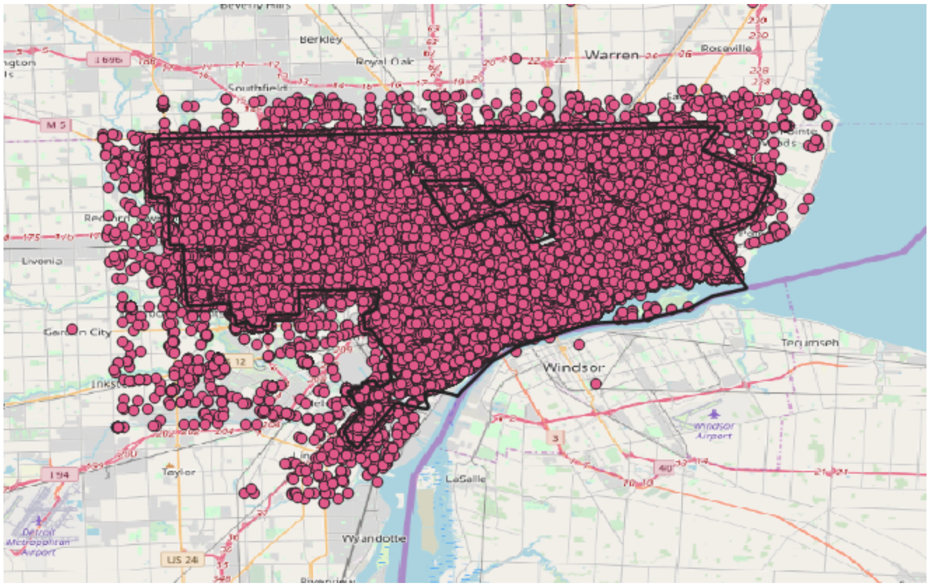


Figure 4.1: All 911 Calls 2018

This issue was raised with Ms Flora (personal communication, March 1, 2019) and she confirmed that this can occur for one of several reasons such as:-

- The proximity of the caller to Detroit's border.
- The quality of location services on the caller's cell phone.
- Whether the caller is driving at the time of the call.

In each case the DPD co-ordinates the calls with the surrounding administrations to ensure that the appropriate agency is notified. Ms Flora also confirmed that the appropriate agency would take the lead in investigating any crimes that were determined to have occurred. Ms Flora also explained that generally, the GPS co-ordinates are determined by geocoding the address provided by the caller using the DPD Computer Aided Dispatch System.

For this reason, it was decided that the datasets would be clipped to only include 911 calls and crime reports located inside of the Detroit City boundaries. QGIS will be used to link the geospatial attributes of 911 calls with the concentrations of bars in Detroit which will be extracted from Open Street Maps also using QGIS.

Data Quality

Generally, the quality of the data is good. There are a number of 911 calls where some of the fields are missing such as the location or the priority of the call. Of the 911 calls for 2018, 25521 (3.1%) do not have the call location and may be discarded from the analysis if they are found to temporally match any crimes. The priority of calls with the field missing are imputed using the median priority. The main quality issue is the inclusion of administrative data in the 911 dataset. An attempt is made to rectify this by removing calls which are not emergency related.

4.2.2 Crime Report Data

The DPD also release several datasets relating to reported criminal offences [3]. They release datasets on homicides, car jackings, non-fatal shootings, major crimes and a dataset for all crime incidents. They also provide interactive maps for these crimes. The dataset titled *DPD: All Crime Incidents, December 6, 2016 - Present* will be used for this project and was also downloaded on March 9, 2019.

Crime Data Description

The Crime Data contains 185 thousand rows each referring to a reported criminal offence with at least one row and 30 columns for each crime event. If a single crime incident results in more than one crime type, then each crime is separately recorded in the dataset using the same reference identifier, but with the different charge description. Of the crimes recorded for 2018, 4582 rows out of 83,028 (5.5%) are duplicates. For example, table 4.2.2 shows five crime reports which relate to only two incidents with crime IDs 3284303 and 3285191 :-
Data included for each entry includes:-

1. Crime reference id. and management system reference which may be repeated.
2. Anonymised location, including address, GPS co-ordinates and US census block.
3. Several fields relating to local government and police codes for administration.

[3] https://data.detroitmi.gov/browse?q=crime&sortBy=relevance

Crime.ID	Report	Category	Charge.Description
3284303	1810310355	BURGLARY	BURGLARY - FORCED ENTRY
3284303	1810310355	ASSAULT	ASSAULT AND BATTERY/SIMPLE ASSAULT
3285191	1811020305	KIDNAPPING	KIDNAPPING / ABDUCTION
3285191	1811020305	STOLEN PROPERTY	STOLEN PROPERTY
3285191	1811020305	AGGRAVATED ASSAULT	AGGRAVATED / FELONIOUS ASSAULT

Table 4.1: Example of Duplicated Crime Reports

4. Description, category, DPD and state codes for the offence.
5. Time and date of the offence plus derived fields such as day of week.

See Appendix B for the complete list of fields. It is explicitly stated in the dataset that the location of the crimes are masked to the 100-block level or an intersection.

Crime Data Exploration

The crime reports date from 1920 to March 7'th, 2019. There are a small number of historically dated crime reports with one each from 1920, 1951 and 1963 and several in the 1970's 1980's, 1990's and 2000's. These historic crimes would seem to indicate data quality issues, perhaps related to data entry. However, checking with Ms Flora (personal communication, March 11, 2019) she confirmed that there is no entry errors and the crimes are present either due to historic crimes being recently reported or for example a crime being upgraded from assault to homicide due to the victim dying several decades after the crime. Data included for each report includes, unique references for the DPD Record Management System, data relating to the location and time of the reported crime, details about the actual offence reported and various DPD related data. It is stated that the locations are anonymised by is masking to the 100 block-level or an intersection.

Crime Report Locations

Similar to the 911 calls, there are a small (but lesser) number of crime reports located outside the City limits. This can be seen in figure 4.2 which for comparison shows both crimes and 911 calls. The crimes, in red dots, are overlaid on the 911 calls in grey dots.

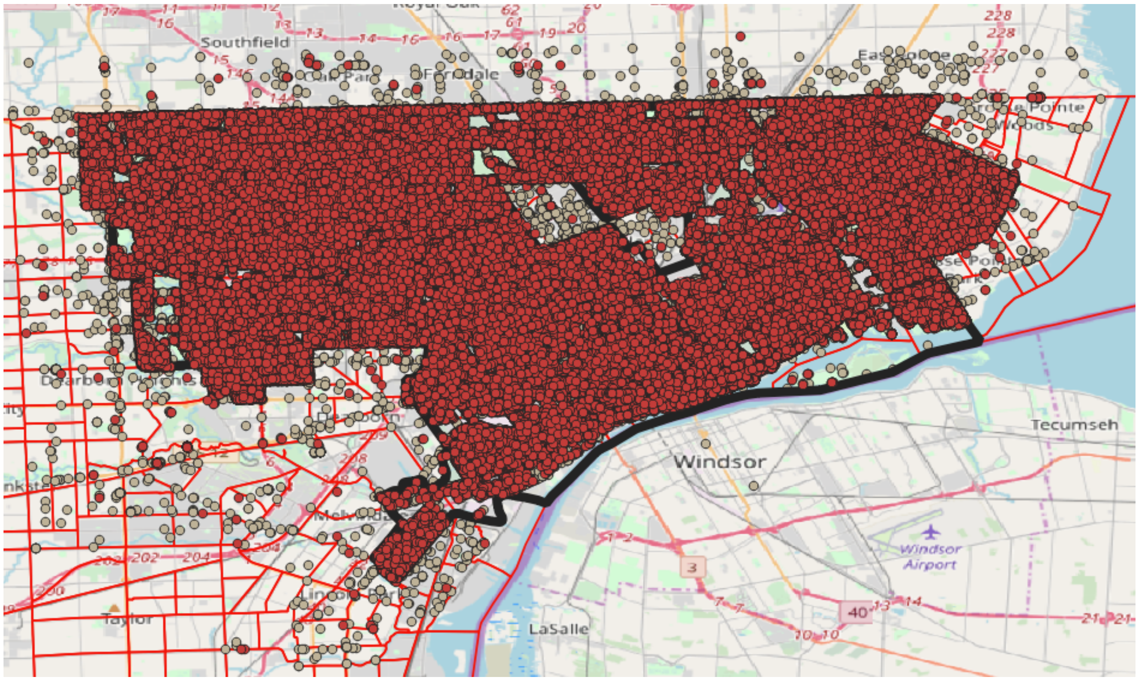
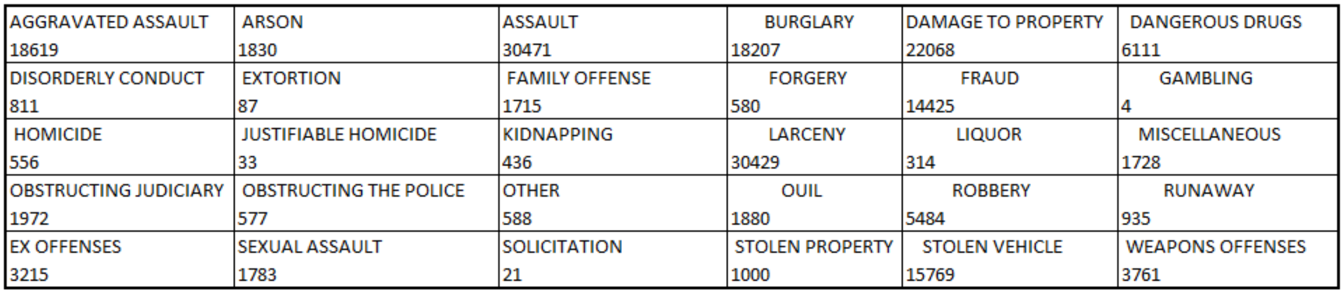


Figure 4.2: 2018 - Crime Reports and 911 Call Locations

As previously stated, 911 calls may occur from locations outside the Detroit City limits and they are investigated by the relevant agency, confirmed by Ms Flora (personal communication, March 1, 2019). It is unclear why there would be crime reports from outside the city limits. These calls and crime reports will be considered as outliers and so they will not be included when building the model.

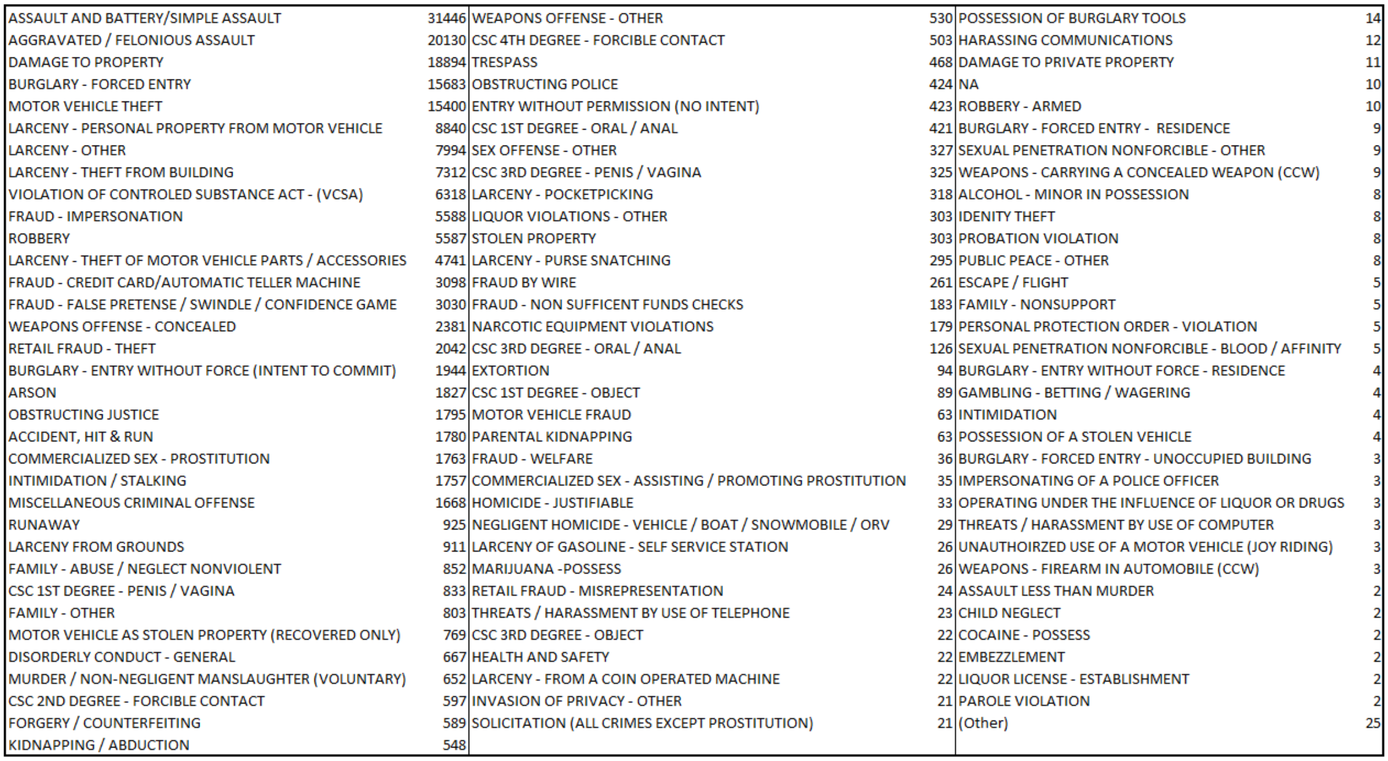
There are a number of fields which indicate the type of crime reported. There are 30 offence categories, see categories of crime and the number of crimes per category in figure 4.3.



AGGRAVATED ASSAULT 18619	ARSON 1830	ASSAULT 30471	BURGLARY 18207	DAMAGE TO PROPERTY 22068	DANGEROUS DRUGS 6111
DISORDERLY CONDUCT 811	EXTORTION 87	FAMILY OFFENSE 1715	FORGERY 580	FRAUD 14425	GAMBLING 4
HOMICIDE 556	JUSTIFIABLE HOMICIDE 33	KIDNAPPING 436	LARCENY 30429	LIQUOR 314	MISCELLANEOUS 1728
OBSTRUCTING JUDICIARY 1972	OBSTRUCTING THE POLICE 577	OTHER 588	OUIL 1880	ROBBERY 5484	RUNAWAY 935
EX OFFENSES 3215	SEXUAL ASSAULT 1783	SOLICITATION 21	STOLEN PROPERTY 1000	STOLEN VEHICLE 15769	WEAPONS OFFENSES 3761

Figure 4.3: Number of Crimes per Category in Crime Dataset

These categories are further sub divided into 122 Offence Descriptions and 121 Charge Descriptions which give further details on each crime. There is very little difference between Offence Description and Charge Description details. Figure 4.4 shows the volume of each crime type recorded.



ASSAULT AND BATTERY/SIMPLE ASSAULT	31446	WEAPONS OFFENSE - OTHER	530	POSSESSION OF BURGLARY TOOLS	14
AGGRAVATED / FELONIOUS ASSAULT	20130	CSC 4TH DEGREE - FORCIBLE CONTACT	503	HARASSING COMMUNICATIONS	12
DAMAGE TO PROPERTY	18894	TRESPASS	468	DAMAGE TO PRIVATE PROPERTY	11
BURGLARY - FORCED ENTRY	15683	OBSTRUCTING POLICE	424	NA	10
MOTOR VEHICLE THEFT	15400	ENTRY WITHOUT PERMISSION (NO INTENT)	423	ROBBERY - ARMED	10
LARCENY - PERSONAL PROPERTY FROM MOTOR VEHICLE	8840	CSC 1ST DEGREE - ORAL / ANAL	421	BURGLARY - FORCED ENTRY - RESIDENCE	9
LARCENY - OTHER	7994	SEX OFFENSE - OTHER	327	SEXUAL PENETRATION NONFORCIBLE - OTHER	9
LARCENY - THEFT FROM BUILDING	7312	CSC 3RD DEGREE - PENIS / VAGINA	325	WEAPONS - CARRYING A CONCEALED WEAPON (CCW)	9
VIOLATION OF CONTROLED SUBSTANCE ACT - (VCSA)	6318	LARCENY - POCKETPICKING	318	ALCOHOL - MINOR IN POSSESSION	8
FRAUD - IMPERSONATION	5588	LIQUOR VIOLATIONS - OTHER	303	IDENITY THEFT	8
ROBBERY	5587	STOLEN PROPERTY	303	PROBATION VIOLATION	8
LARCENY - THEFT OF MOTOR VEHICLE PARTS / ACCESSORIES	4741	LARCENY - PURSE SNATCHING	295	PUBLIC PEACE - OTHER	8
FRAUD - CREDIT CARD/AUTOMATIC TELLER MACHINE	3098	FRAUD BY WIRE	261	ESCAPE / FLIGHT	5
FRAUD - FALSE PRETENSE / SWINDLE / CONFIDENCE GAME	3030	FRAUD - NON SUFFICENT FUNDS CHECKS	183	FAMILY - NONSUPPORT	5
WEAPONS OFFENSE - CONCEALED	2381	NARCOTIC EQUIPMENT VIOLATIONS	179	PERSONAL PROTECTION ORDER - VIOLATION	5
RETAIL FRAUD - THEFT	2042	CSC 3RD DEGREE - ORAL / ANAL	126	SEXUAL PENETRATION NONFORCIBLE - BLOOD / AFFINITY	5
BURGLARY - ENTRY WITHOUT FORCE (INTENT TO COMMIT)	1944	EXTORTION	94	BURGLARY - ENTRY WITHOUT FORCE - RESIDENCE	4
ARSON	1827	CSC 1ST DEGREE - OBJECT	89	GAMBLING - BETTING / WAGERING	4
OBSTRUCTING JUSTICE	1795	MOTOR VEHICLE FRAUD	63	INTIMIDATION	4
ACCIDENT, HIT & RUN	1780	PARENTAL KIDNAPPING	63	POSSESSION OF A STOLEN VEHICLE	4
COMMERCIALIZED SEX - PROSTITUTION	1763	FRAUD - WELFARE	36	BURGLARY - FORCED ENTRY - UNOCCUPIED BUILDING	3
INTIMIDATION / STALKING	1757	COMMERCIALIZED SEX - ASSISTING / PROMOTING PROSTITUTION	35	IMPERSONATING OF A POLICE OFFICER	3
MISCELLANEOUS CRIMINAL OFFENSE	1668	HOMICIDE - JUSTIFIABLE	33	OPERATING UNDER THE INFLUENCE OF LIQUOR OR DRUGS	3
RUNAWAY	925	NEGLIGENT HOMICIDE - VEHICLE / BOAT / SNOWMOBILE / ORV	29	THREATS / HARASSMENT BY USE OF COMPUTER	3
LARCENY FROM GROUNDS	911	LARCENY OF GASOLINE - SELF SERVICE STATION	26	UNAUTHOIRZED USE OF A MOTOR VEHICLE (JOY RIDING)	3
FAMILY - ABUSE / NEGLECT NONVIOLENT	852	MARIJUANA -POSSESS	26	WEAPONS - FIREARM IN AUTOMOBILE (CCW)	3
CSC 1ST DEGREE - PENIS / VAGINA	833	RETAIL FRAUD - MISREPRESENTATION	24	ASSAULT LESS THAN MURDER	2
FAMILY - OTHER	803	THREATS / HARASSMENT BY USE OF TELEPHONE	23	CHILD NEGLECT	2
MOTOR VEHICLE AS STOLEN PROPERTY (RECOVERED ONLY)	769	CSC 3RD DEGREE - OBJECT	22	COCAINE - POSSESS	2
DISORDERLY CONDUCT - GENERAL	667	HEALTH AND SAFETY	22	EMBEZZLEMENT	2
MURDER / NON-NEGLIGENT MANSLAUGHTER (VOLUNTARY)	652	LARCENY - FROM A COIN OPERATED MACHINE	22	LIQUOR LICENSE - ESTABLISHMENT	2
CSC 2ND DEGREE - FORCIBLE CONTACT	597	INVASION OF PRIVACY - OTHER	21	PAROLE VIOLATION	2
FORGERY / COUNTERFEITING	589	SOLICITATION (ALL CRIMES EXCEPT PROSTITUTION)	21	(Other)	25
KIDNAPPING / ABDUCTION	548				

Figure 4.4: Number of Crimes per Descriptions in Crime dataset

4.2.3 Historical Crime Data

As has been previously stated the crime hot spots will be created by using historical crime reports from the DPD crime reports dataset. As can be seen in figure 4.5, the monthly volumes are generally similar with the exception of February and December.

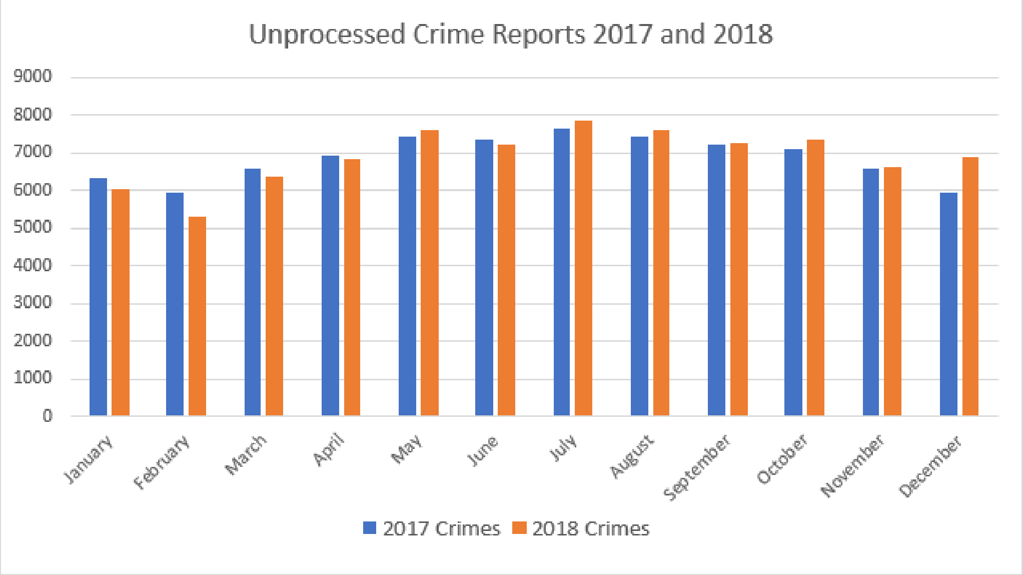


Figure 4.5: Detroit City Crimes - 2017

4.2.4 911 Calls and Recorded Crime Trends for 2018

Figure 4.6 shows the monthly volumes of 911 requests for service and crime reports for 2018 with the same y-axis showing the relative volumes. Figure 4.7 presents the data with different scales for the y-axis giving an indication of the trends for 911 calls and crime reports.

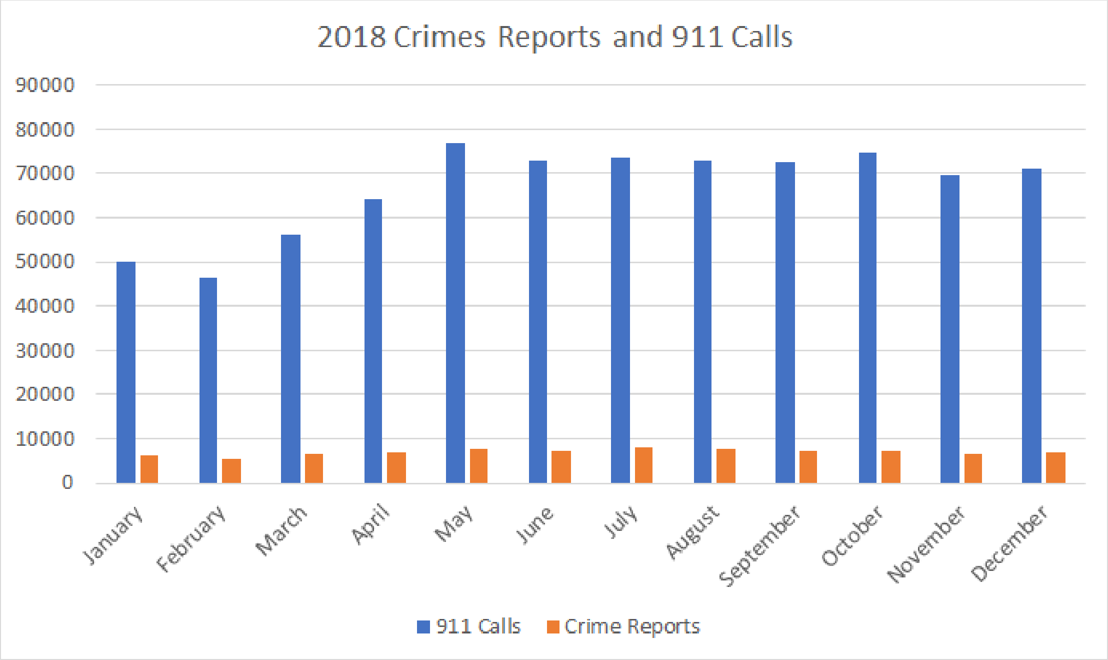


Figure 4.6: 2018 - Raw Data for Crime and 911 Calls

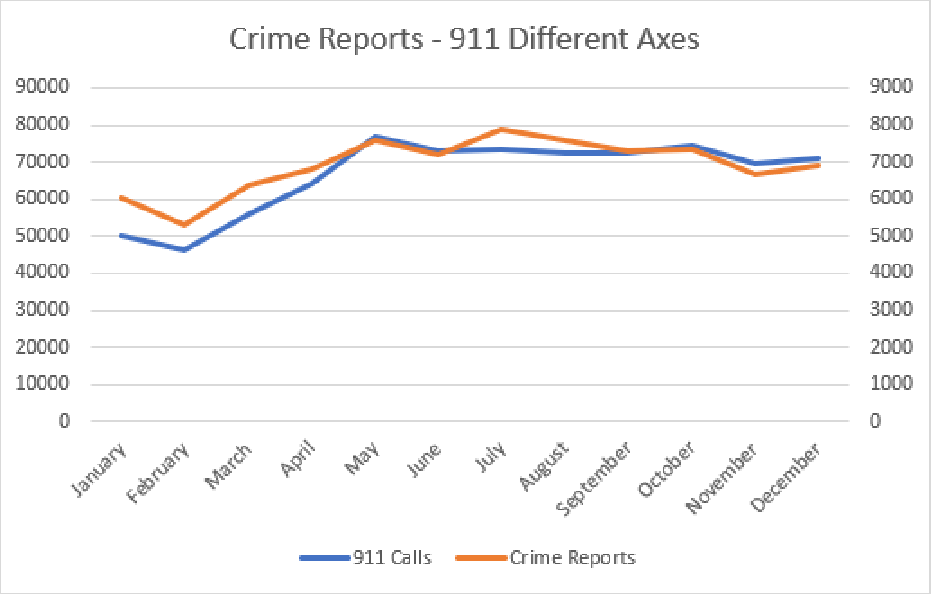


Figure 4.7: 2018 - Crime and 911 Calls Dual Y Axis

4.2.5 Demographic and Economic Data - US Census

The US Census Bureau is the primary source of official information for the US population and economy and they are tasked with providing the most current facts regarding people, places and the economy. This data is used for a variety of functional purposes for government, for example, in allocating government funding or to allocate seats in the US House of Representatives. [4] Various types of data are collected and processed, relating to the population, government and economy. The US Census Bureau undertakes a full Census for each resident every ten years collecting data on the population and housing with the next census due in 2020.

To supplement the decennial census, the US Census Bureau annually undertake a sample survey of 3.5 million US households where the households chosen at random are legally obliged to participate in the survey. This survey, called the American Community Survey (ACS) [5], takes place all year every year and the data is then used to create estimates which are released yearly in the form of tables and analytical reports.

American Community Survey

Previously three different types of estimates were released, 1-year, 3-year and 5-year with the 1-year data being the most current but also for the largest geographical areas (populations of 65,000 or more). The 3-year estimates have been discontinued. 5-year estimates use data collected over 60 months for all areas and is the most reliable, but the least current. The 5-year estimates data will be used for this project as it has data for the geographical area of study, census tracts.

The data collected relates to four subject areas:-

- Social Characteristics - Education, marital status, relationships, fertility, grandparents.
- Economic Characteristics - Income, employment, occupation, commuting to work.

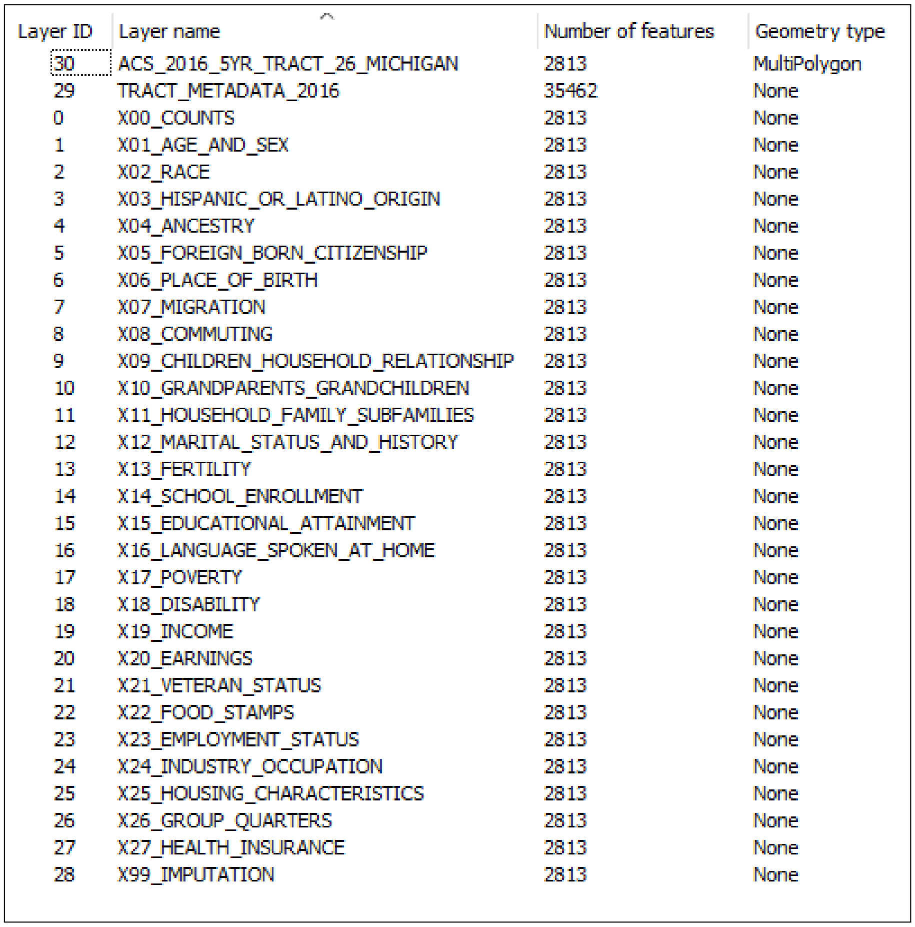
[4] https://www.census.gov/about.html

[5] https://www.census.gov/programs-surveys/acs

- Housing Characteristics - Occupancy and structure, housing value and costs, utilities.
- Demographic Characteristics - gender, age, race and immigrant origin.

Census Tract Data

The Michigan census tract file *ACS_2016_5YR_TRACT_26.gdb* was downloaded on February 20, 2019. The file is a Geodatabase file consisting of one vector geometry file, one file containing metadata and 28 data files each with 2813 features, see figure 4.8. The metadata file TRACT_METADATA_2016 describes in detail the contents of the other files. Each of the data files has 735 rows of data, one each for the 735 census tracts in Michigan, there are 297 census tracts in Detroit. The data files need to be imported before processing and exploration is possible.



Layer ID	Layer name	Number of features	Geometry type
30	ACS_2016_5YR_TRACT_26_MICHIGAN	2813	MultiPolygon
29	TRACT_METADATA_2016	35462	None
0	X00_COUNTS	2813	None
1	X01_AGE_AND_SEX	2813	None
2	X02_RACE	2813	None
3	X03_HISPANIC_OR_LATINO_ORIGIN	2813	None
4	X04_ANCESTRY	2813	None
5	X05_FOREIGN_BORN_CITIZENSHIP	2813	None
6	X06_PLACE_OF_BIRTH	2813	None
7	X07_MIGRATION	2813	None
8	X08_COMMUTING	2813	None
9	X09_CHILDREN_HOUSEHOLD_RELATIONSHIP	2813	None
10	X10_GRANDPARENTS_GRANDCHILDREN	2813	None
11	X11_HOUSEHOLD_FAMILY_SUBFAMILIES	2813	None
12	X12_MARITAL_STATUS_AND_HISTORY	2813	None
13	X13_FERTILITY	2813	None
14	X14_SCHOOL_ENROLLMENT	2813	None
15	X15_EDUCATIONAL_ATTAINMENT	2813	None
16	X16_LANGUAGE_SPOKEN_AT_HOME	2813	None
17	X17_POVERTY	2813	None
18	X18_DISABILITY	2813	None
19	X19_INCOME	2813	None
20	X20_EARNINGS	2813	None
21	X21_VETERAN_STATUS	2813	None
22	X22_FOOD_STAMPS	2813	None
23	X23_EMPLOYMENT_STATUS	2813	None
24	X24_INDUSTRY_OCCUPATION	2813	None
25	X25_HOUSING_CHARACTERISTICS	2813	None
26	X26_GROUP_QUARTERS	2813	None
27	X27_HEALTH_INSURANCE	2813	None
28	X99_IMPUTATION	2813	None

Figure 4.8: American Community Survey Michigan 2016

Files which will be utilised are:-

- X01 Age and Sex
- X02 Race
- X15 Educational Attainment
- X19 Income
- X22 Food Stamps
- X23 Employment Status
- X25 Housing Characteristics

4.2.6 Detroit City Census Geography

In Detroit City, there are 1010 Block Groups and 297 census tracts, see figure 4.9. The Block Groups range in size from $0.1km^2$ to $4.7km^2$ and census tracts range from $0.2km^2$ to $5.1km^2$. The population ranges in Block Groups from 0 to 7190 with a median population of 1081 and for census tracts from 0 to 9622 with a median population of 3178.

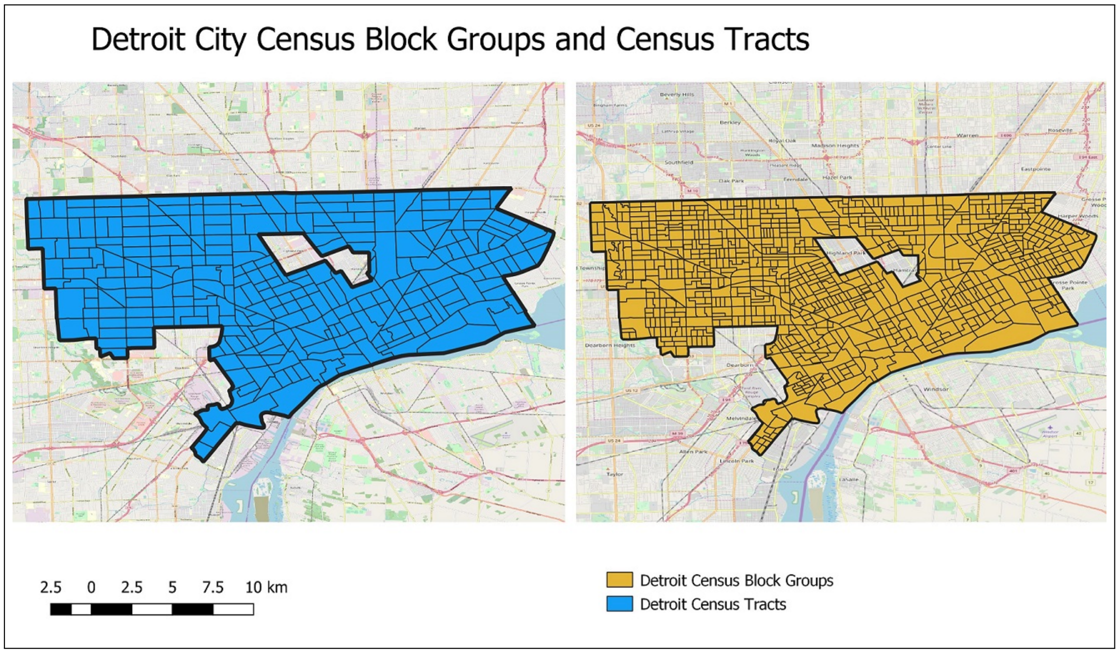

Detroit City Census Block Groups and Census Tracts
2.5 0 2.5 5 7.5 10 km
🟧 Detroit Census Block Groups
🟦 Detroit Census Tracts

Figure 4.9: Detroit Geography

A design decision was made to choose the resolution of census tract as the neighbourhood as it is felt that the population size is optimum for building the model and there will be a lower processing cost when compared to Block Groups. In addition, as previously stated, census tracts are created with the aim of combining group characteristic for that area.

4.2.7 Points of Interest - Bars

The research has shown a link between crime and proximity to bars and nightclubs. For that reason, the proximity of each 911 call to bars/nightclubs is included in the model. There were two possible sources of data for data on bars and nightclubs sourced.

1. Detroit list of Liquor Licenses
2. Open Street Maps

Michigan state release the list of venues licensed to sell alcohol with the name, type of venue and addresses.

OpenStreetMap (OSM) is a collaborative Wikipedia type effort to create a free downloadable map of the world where local mappers construct the maps. OSM was chosen because it was considered that this data was most likely to be complete although there are certain challenges, which will be described later, in utilising this data.

OpenStreetMap

A key aspect to OSM is the data which represents a wide variety of geographical objects ranging from bus benches to entertainment venues, bridges, roads, rivers etc. This data is available in a variety of formats such as OSM format, pbf format and shape files. QGIS will be used to extract geospatial data on bars in the Detroit area and to then link this data on bars with the locations of 911 calls. Initially, data for Michigan in the pbf format [6] was considered. However, the pbf format is data intense, see figure 4.10 where every possible type of object is represented by a yellow dot. Due to the computation effort which was observed in QGIS when attempting to filter the data for the Detroit area, this format was considered unsuitable for the project.

[6] https://download.geofabrik.de/north-america/us/michigan.html

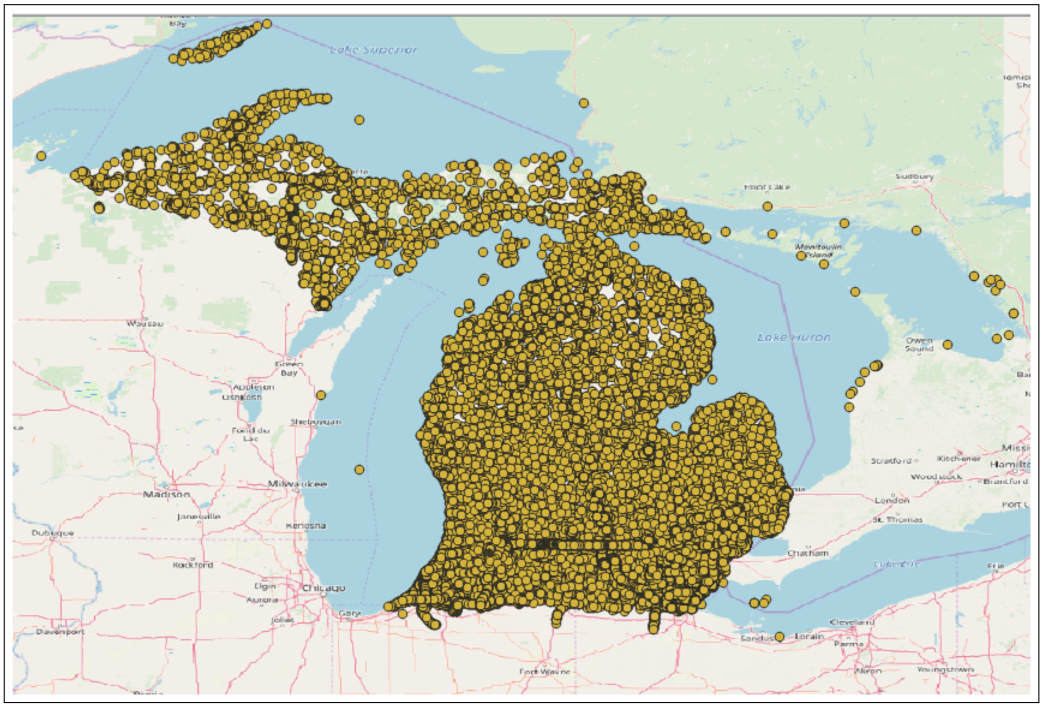


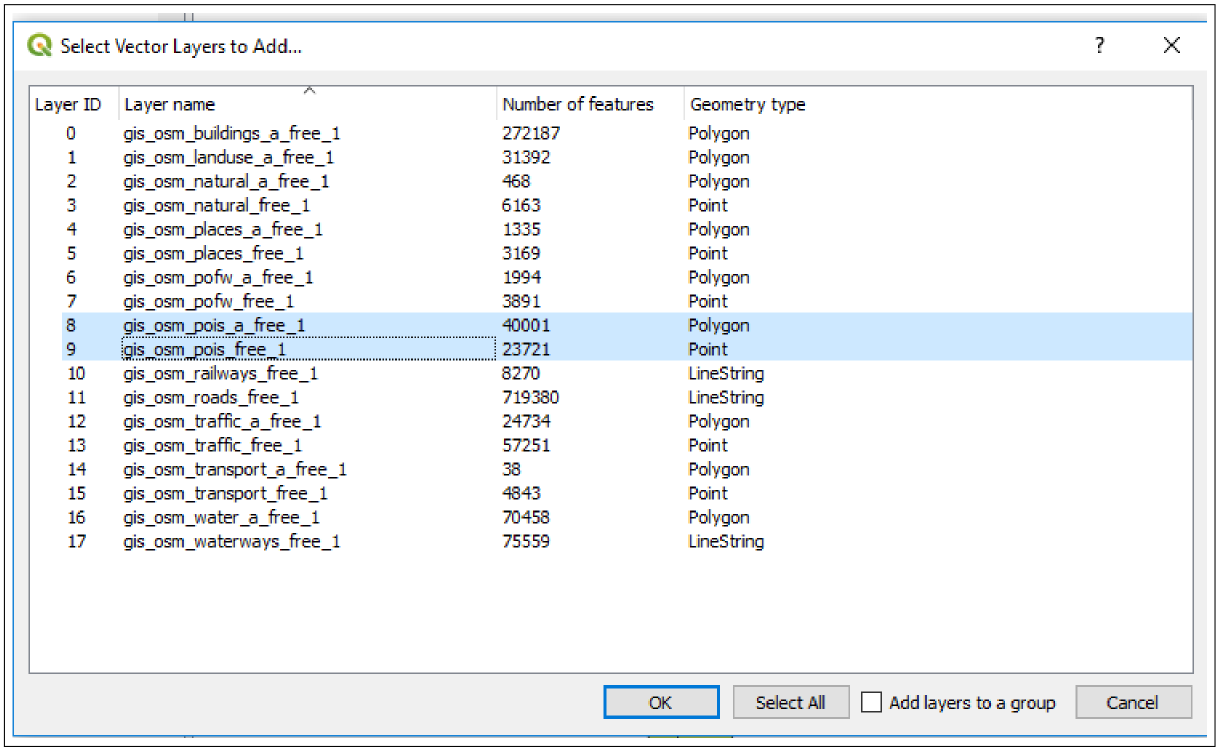
Figure 4.10: Michigan OSM Data in PBF Format

Shapefiles

Shapefiles are a widely used format originally developed by ESRI, a leading supplier of GIS systems. Shapefiles are a collection of files which provide location and other information regarding geographical objects. Each shapefile contains one type of feature, a point, a polygon or a line. A point is a single object with one geographical location which can represent anything from a city to a bus stop depending on the resolution. A line is a series of points which can represent for example, a road or a river. A polygon is a line where its end point is the same as it's start point and can represent objects such as lakes or building footprints.

Examining the shapefile for Michigan [7], figure 4.11, shows that it contains 17 layers, note the two highlighted files which contain POI data and will be discussed later.

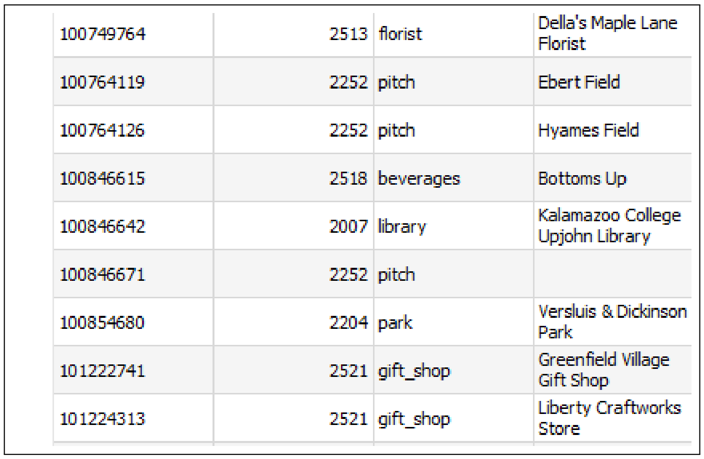
[7] http://download.geofabrik.de/north-america/us/michigan.html


Select Vector Layers to Add...

Layer ID	Layer name	Number of features	Geometry type
0	gis_osm_buildings_a_free_1	272187	Polygon
1	gis_osm_landuse_a_free_1	31392	Polygon
2	gis_osm_natural_a_free_1	468	Polygon
3	gis_osm_natural_free_1	6163	Point
4	gis_osm_places_a_free_1	1335	Polygon
5	gis_osm_places_free_1	3169	Point
6	gis_osm_pofw_a_free_1	1994	Polygon
7	gis_osm_pofw_free_1	3891	Point
8	gis_osm_pois_a_free_1	40001	Polygon
9	gis_osm_pois_free_1	23721	Point
10	gis_osm_railways_free_1	8270	LineString
11	gis_osm_roads_free_1	719380	LineString
12	gis_osm_traffic_a_free_1	24734	Polygon
13	gis_osm_traffic_free_1	57251	Point
14	gis_osm_transport_a_free_1	38	Polygon
15	gis_osm_transport_free_1	4843	Point
16	gis_osm_water_a_free_1	70458	Polygon
17	gis_osm_waterways_free_1	75559	LineString

Figure 4.11: Michigan OSM Data Layers in Shapefile Format

An example of this data is shown in figure 4.12.



100749764	2513	florist	Della's Maple Lane Florist
100764119	2252	pitch	Ebert Field
100764126	2252	pitch	Hyames Field
100846615	2518	beverages	Bottoms Up
100846642	2007	library	Kalamazoo College Upjohn Library
100846671	2252	pitch	
100854680	2204	park	Versluis & Dickinson Park
101222741	2521	gift_shop	Greenfield Village Gift Shop
101224313	2521	gift_shop	Liberty Craftworks Store

Figure 4.12: Michigan OSM Shapefile, Example of data

4.2.8 Weather Data

The National Centers for Environmental Information (formerly National Climatic Data Center) [8] collects and makes available data relating to the climate. It is possible to download historic data from selected data stations. The weather station at Detroit airport was selected for this study.

Weather Data Description

The data was downloaded in csv format using millimetres for rainfall and Celsius for temperature, for the US Weather station at Detroit City airport designation USW00014822. Daily data was downloaded for:-

- Precipitation (PRCP)
- Maximum temperature (TMAX)
- Minimum temperature (TMIN)

2018 Weather Summary

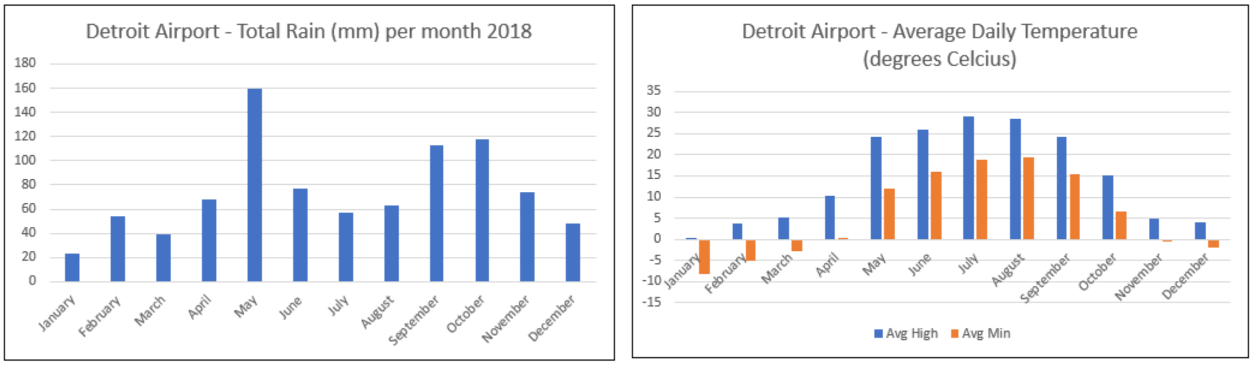


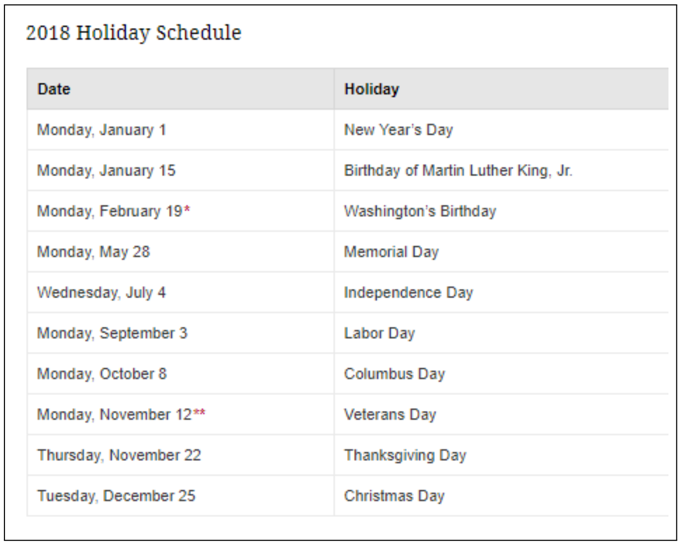
Figure 4.13: 2018 Weather

4.2.9 Holiday Data

It is intended to include in the model variables to indicate the time of day and day of week that the 911 call occurs. In addition, it will be indicated if the 911 call occurs

[8] https://www.ncdc.noaa.gov

during a US Federal holiday [9] or the day preceding a holiday by introducing two new variables, *isHoliday* and *isHolidayEve*.



2018 Holiday Schedule

Date	Holiday
Monday, January 1	New Year's Day
Monday, January 15	Birthday of Martin Luther King, Jr.
Monday, February 19*	Washington's Birthday
Monday, May 28	Memorial Day
Wednesday, July 4	Independence Day
Monday, September 3	Labor Day
Monday, October 8	Columbus Day
Monday, November 12**	Veterans Day
Thursday, November 22	Thanksgiving Day
Tuesday, December 25	Christmas Day

Figure 4.14: 2018 Federal Holidays

In addition, the days after Thanksgiving and the 4th of July will also be marked as holidays for this study as it is common for both days to be taken as a holiday.

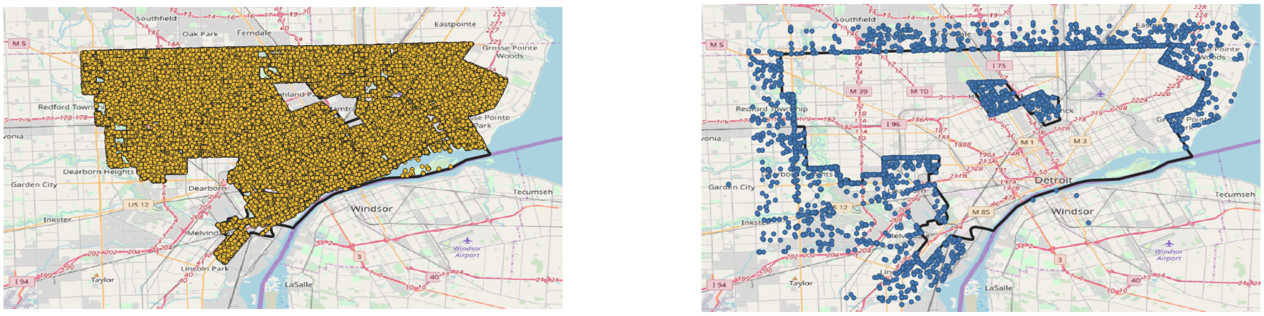
4.3 Data Preparation - Primary Dataset

The primary dataset will contain only 911 Call data and the dependent variable Crime Report (Yes/No) and is intended for the testing of the null hypothesis. This dataset will then be extended by adding additional features to test the alternative hypothesis.

4.3.1 911 Calls

The 911 dataset is first reduced from 1.7 million calls to slightly over 800,000 calls by including only data from 2018. Several columns which are not required are discarded to reduce memory required during processing. Information relating to the following is retained, unique reference id, address and location, priority, call codes and descriptions, time and date, officer initiated, response time and census block code.

[9] https://www.opm.gov/policy-data-oversight/snow-dismissal-procedures/federal-holidays



(a) 911 Calls inside Detroit

(b) 911 Calls outside Detroit

Figure 4.15: 2018 Detroit PD 911 Calls by City Boundary

911 Call Location - GEOID

The location of each 911 call is given by its address, GPS co-ordinates and census block GEOID. As previously stated, analysis will be performed at census tract resolution so census tracts must be computed from the census blocks. The census tract can be calculated by stripping away the last four digits of the census block GEOID, see table 4.2. Note that this reduces the location areas being analysed from 17,821 census blocks to 297 census tracts.

Area Type	GEOID Structure	Number of Digits	Example GEOID
Census Tract	STATE+COUNTY+TRACT	2+3+6=11	48201223100
Block	STATE+COUNTY+TRACT + BLOCK	2+3+6+4=15	482012231001050

Table 4.2: US Census GEOIDs

Calls Outside Detroit City

As previously shown, there are calls from outside the Detroit City boundary. Each of the 911 calls from 2018 (800,226 calls) are checked against a list of Detroit City GEOIDs and the calls outside the city boundary are removed leaving 767,224 Calls (33,002 calls were from outside the city boundary) see figure 4.15.

911 Data Cleaning

The remaining 911 Calls were further filtered to remove some calls which were unlikely to be associated with Violent Crime and or administrative, for example calls which are administrative or medical.

Ms Flora of the DPD (personal communication, March 28, 2019), confirmed that the DPD can distinguish between administrative 911 calls and non-administrative but they do not release this information.

The following call types were dropped as these calls are probably not emergencies but administrative whereby the officers are logging their activity.

1. START OF SHIFT INFORMATION - 56938 calls - Start of officer shift
2. TOWING DETAIL - 22349 calls - Car being towed
3. MISCELLANEOUS TRAFFIC - 4875 calls
4. INFORMATION/NON-CRIMINAL RPT - 4592 calls
5. RECOVER AUTO - 3879 calls
6. BUS BOARDING - 7273 calls - routine boarding of bus
7. MT EMS-TRO/ENTRY 2089 calls - Medical/EMS code
8. ADMIT OR E/E - 1605 calls - Medical/EMS code
9. BUILDING CHECK - 5934 calls - routine administrative
10. LOST PROPERTY - 1502 calls - routine call
11. TRANSPORT PRISONER - 1192 calls - administrative

This reduced the number of 911 calls to 653,764 calls.

911 Calls Features

The following features are retained for model building:-

- Priority
- Officer.Initiated
- Day
- Hour

- Month
- GEOID
- Call.Code

Census tracts with 297 levels are used instead of census blocks with 17821 levels. There are three variables which represent the type of call for example, Call.Code 34201, is described by Call Description "Shots Fired IP" and has category "SHOTS IP". There is collinearity between these three features and Call.Code has the most information with 326 levels so is kept. The call codes descriptions can be found on the Michigan state police website [10].

911 Calls by Month

Examining the monthly 911 calls, see figure 4.16, there is a trend where the total number of 911 calls start the year at slightly over 40,000 calls per month but then increase to approximately 60,000 calls per month. This trend is mirrored by the officer initiated calls, but the public calls peak during the summer months.

911 Calls by Day

The daily call rates, see figure 4.17, show that there is very little difference in the number of calls initiated by the public but that officer initiated calls peak on a Wednesday and are lowest at the weekend.

911 Calls by Hour of the Day

The hourly call rates, see figure 4.18, show that the officer initiated and public trends are similar with the number of call peaking in mid afternoon and reducing in the early morning.

The number of officer initiated calls may be influenced by administrative factors such as overtime. Or, it could be reduced when officers are busy with public initiated calls and so unable to make proactive calls.

[10] https://www.michigan.gov/msp/0,4643,7-123-72297_24055-253478–,00.html.

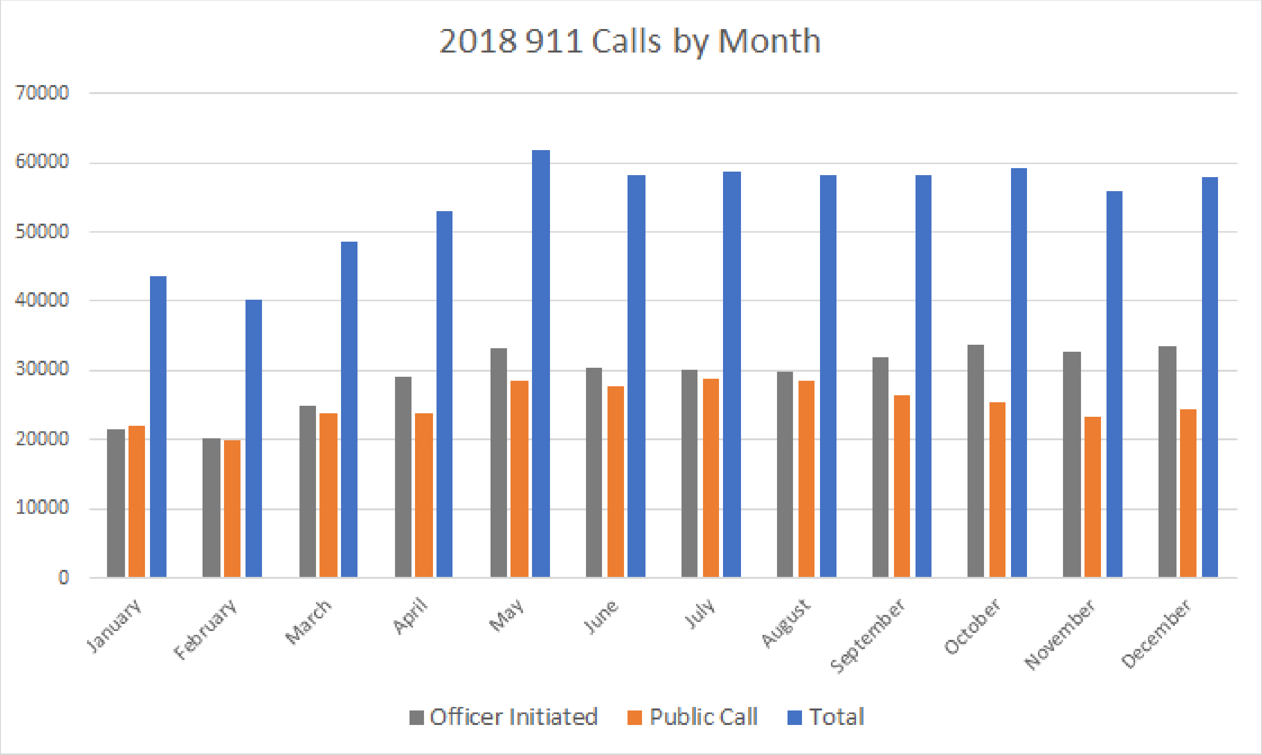


Figure 4.16: Filtered 911 Calls - 2018 Monthly

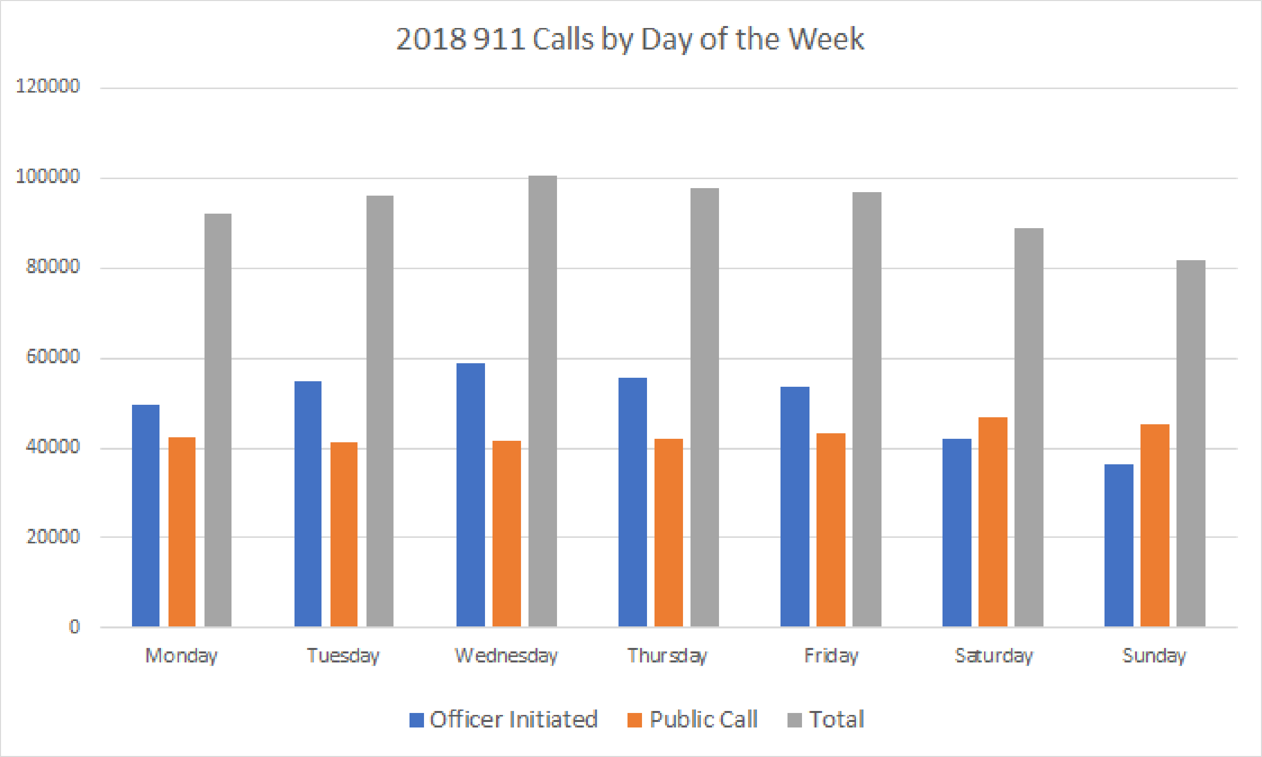


Figure 4.17: Filtered 911 Calls - 2018 Daily

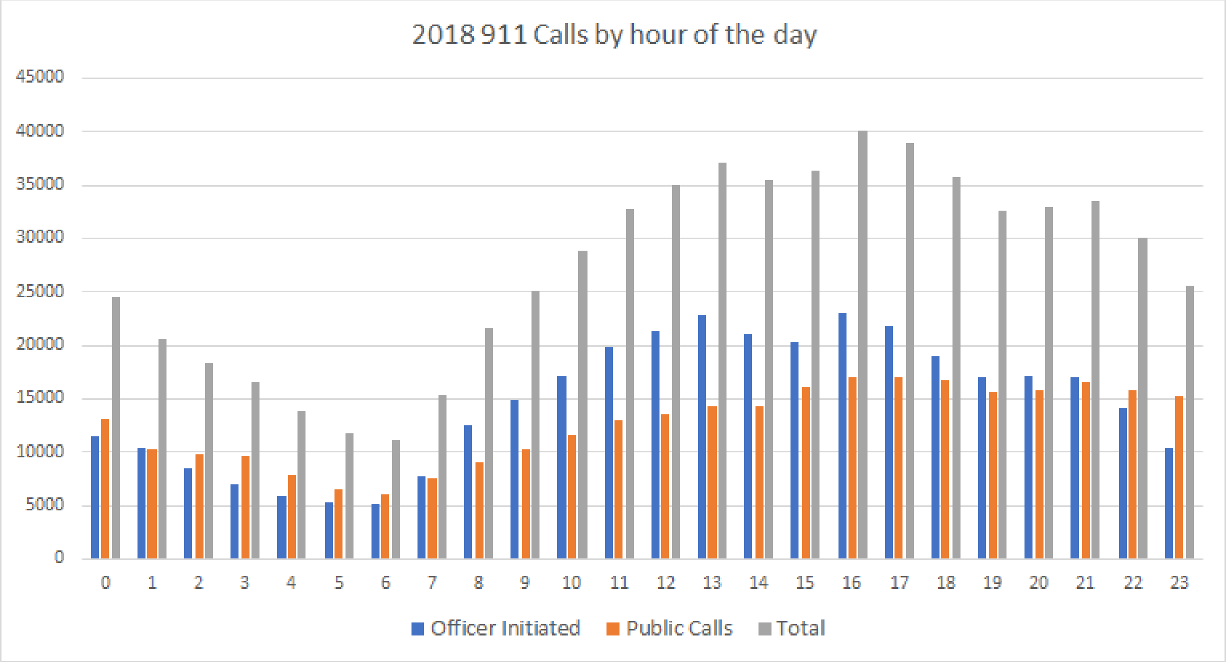


Figure 4.18: Filtered 911 Calls - 2018 Hourly

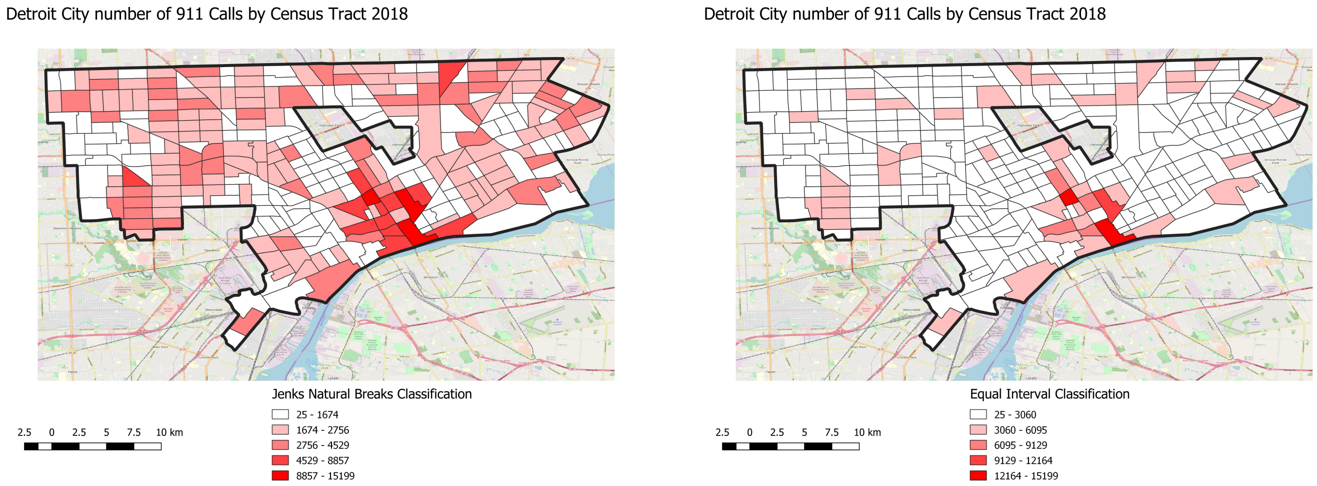
911 Calls by Location

Grouping the calls by their census tract location see figure 4.19, as expected a higher number of calls are received in the south of the city which is the city center. Jenks natural breaks classification and equal interval classification are both shown in the choropleth map and it can be seen that Jenks natural breaks is more effective at describing the 911 call distribution.

4.3.2 Crime Reports

The crime report dataset will be analysed by comparing it geospatially and temporally to the 911 dataset, in order to create the dependent variable *crime report (yes/no)*. It is first filtered to include only Crimes from 2018 plus crimes on the first day of 2019 and last day of 2017. The two extra days are included as it is possible that 911 calls on January 1st could be related to crimes before midnight and likewise, 911 calls on December 31st may be related to crimes the next day.

This reduces the crime dataset from 185,409 crime reports to 83,210 crime reports.


Detroit City number of 911 Calls by Census Tract 2018
Jenks Natural Breaks Classification
25 - 1674
1674 - 2756
2756 - 4529
4529 - 8857
8857 - 15199
2.5 0 2.5 5 7.5 10 km
Detroit City number of 911 Calls by Census Tract 2018
Equal Interval Classification
25 - 3060
3060 - 6095
6095 - 9129
9129 - 12164
12164 - 15199
2.5 0 2.5 5 7.5 10 km

(a) Jenks Natural Breaks Classification

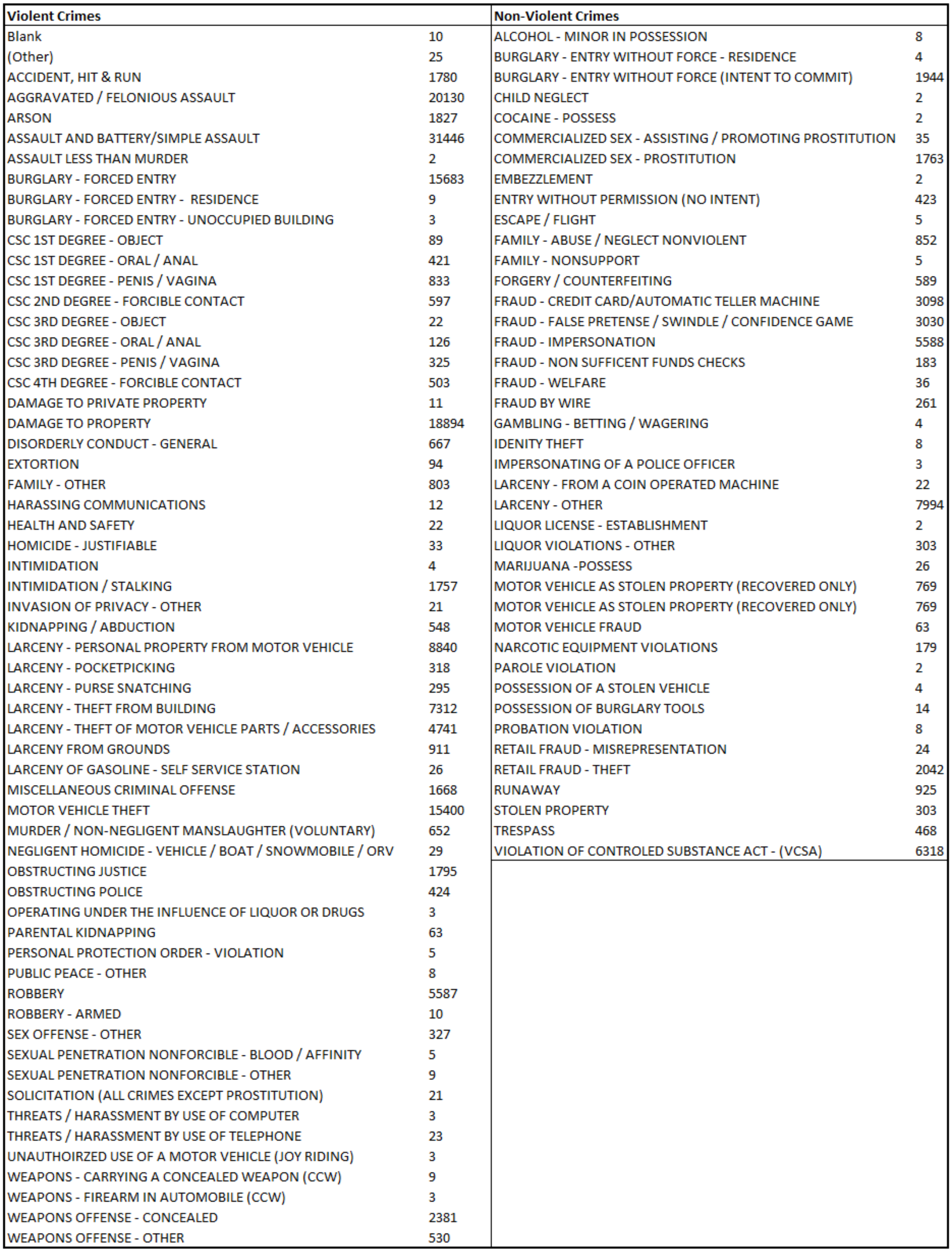
(b) Equal Interval Classification

Figure 4.19: 2018 Detroit PD Filtered 911 Calls by Census Tract

The dataset size is then further reduced in size by removing columns containing data which will not be used or is available in other forms. 1090 Crimes located outside the Detroit City boundary are also removed reducing the dataset to 82,120 rows.

Violent Crime Exploration

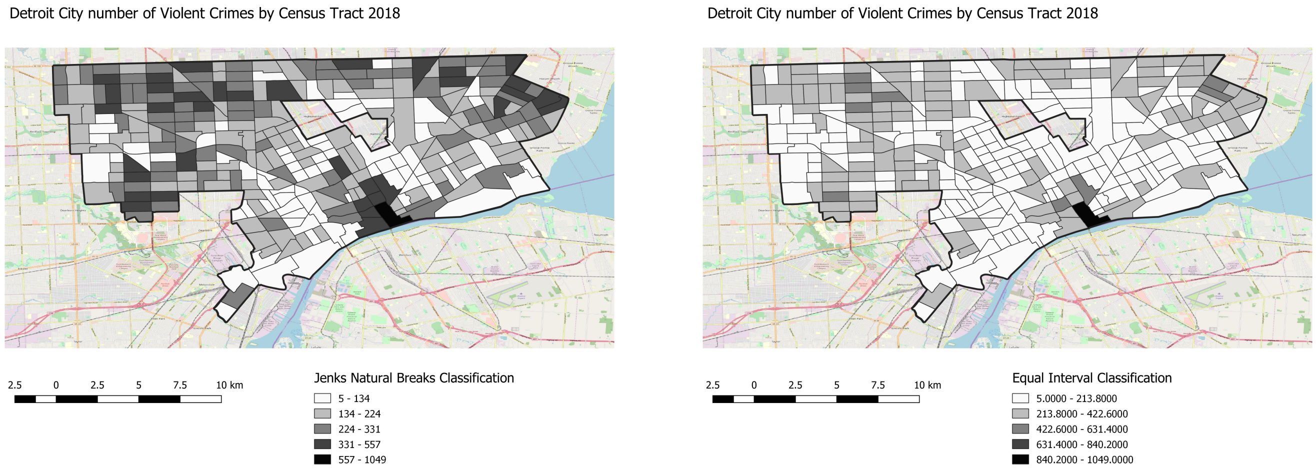
As it is intended to predict violent crimes only, the Crimes are subdivided in Violent/ Not Violent based on the Offence Description see figure 4.20 splitting the dataset into 65492 violent crimes and 16628 non-violent crimes.



Violent Crimes		Non-Violent Crimes	
Blank	10	ALCOHOL - MINOR IN POSSESSION	8
(Other)	25	BURGLARY - ENTRY WITHOUT FORCE - RESIDENCE	4
ACCIDENT, HIT & RUN	1780	BURGLARY - ENTRY WITHOUT FORCE (INTENT TO COMMIT)	1944
AGGRAVATED / FELONIOUS ASSAULT	20130	CHILD NEGLECT	2
ARSON	1827	COCAINE - POSSESS	2
ASSAULT AND BATTERY/SIMPLE ASSAULT	31446	COMMERCIALIZED SEX - ASSISTING / PROMOTING PROSTITUTION	35
ASSAULT LESS THAN MURDER	2	COMMERCIALIZED SEX - PROSTITUTION	1763
BURGLARY - FORCED ENTRY	15683	EMBEZZLEMENT	2
BURGLARY - FORCED ENTRY - RESIDENCE	9	ENTRY WITHOUT PERMISSION (NO INTENT)	423
BURGLARY - FORCED ENTRY - UNOCCUPIED BUILDING	3	ESCAPE / FLIGHT	5
CSC 1ST DEGREE - OBJECT	89	FAMILY - ABUSE / NEGLECT NONVIOLENT	852
CSC 1ST DEGREE - ORAL / ANAL	421	FAMILY - NONSUPPORT	5
CSC 1ST DEGREE - PENIS / VAGINA	833	FORGERY / COUNTERFEITING	589
CSC 2ND DEGREE - FORCIBLE CONTACT	597	FRAUD - CREDIT CARD/AUTOMATIC TELLER MACHINE	3098
CSC 3RD DEGREE - OBJECT	22	FRAUD - FALSE PRETENSE / SWINDLE / CONFIDENCE GAME	3030
CSC 3RD DEGREE - ORAL / ANAL	126	FRAUD - IMPERSONATION	5588
CSC 3RD DEGREE - PENIS / VAGINA	325	FRAUD - NON SUFFICENT FUNDS CHECKS	183
CSC 4TH DEGREE - FORCIBLE CONTACT	503	FRAUD - WELFARE	36
DAMAGE TO PRIVATE PROPERTY	11	FRAUD BY WIRE	261
DAMAGE TO PROPERTY	18894	GAMBLING - BETTING / WAGERING	4
DISORDERLY CONDUCT - GENERAL	667	IDENITY THEFT	8
EXTORTION	94	IMPERSONATING OF A POLICE OFFICER	3
FAMILY - OTHER	803	LARCENY - FROM A COIN OPERATED MACHINE	22
HARASSING COMMUNICATIONS	12	LARCENY - OTHER	7994
HEALTH AND SAFETY	22	LIQUOR LICENSE - ESTABLISHMENT	2
HOMICIDE - JUSTIFIABLE	33	LIQUOR VIOLATIONS - OTHER	303
INTIMIDATION	4	MARIJUANA -POSSESS	26
INTIMIDATION / STALKING	1757	MOTOR VEHICLE AS STOLEN PROPERTY (RECOVERED ONLY)	769
INVASION OF PRIVACY - OTHER	21	MOTOR VEHICLE AS STOLEN PROPERTY (RECOVERED ONLY)	769
KIDNAPPING / ABDUCTION	548	MOTOR VEHICLE FRAUD	63
LARCENY - PERSONAL PROPERTY FROM MOTOR VEHICLE	8840	NARCOTIC EQUIPMENT VIOLATIONS	179
LARCENY - POCKETPICKING	318	PAROLE VIOLATION	2
LARCENY - PURSE SNATCHING	295	POSSESSION OF A STOLEN VEHICLE	4
LARCENY - THEFT FROM BUILDING	7312	POSSESSION OF BURGLARY TOOLS	14
LARCENY - THEFT OF MOTOR VEHICLE PARTS / ACCESSORIES	4741	PROBATION VIOLATION	8
LARCENY FROM GROUNDS	911	RETAIL FRAUD - MISREPRESENTATION	24
LARCENY OF GASOLINE - SELF SERVICE STATION	26	RETAIL FRAUD - THEFT	2042
MISCELLANEOUS CRIMINAL OFFENSE	1668	RUNAWAY	925
MOTOR VEHICLE THEFT	15400	STOLEN PROPERTY	303
MURDER / NON-NEGLIGENT MANSLAUGHTER (VOLUNTARY)	652	TRESPASS	468
NEGLIGENT HOMICIDE - VEHICLE / BOAT / SNOWMOBILE / ORV	29	VIOLATION OF CONTROLED SUBSTANCE ACT - (VCSA)	6318
OBSTRUCTING JUSTICE	1795		
OBSTRUCTING POLICE	424		
OPERATING UNDER THE INFLUENCE OF LIQUOR OR DRUGS	3		
PARENTAL KIDNAPPING	63		
PERSONAL PROTECTION ORDER - VIOLATION	5		
PUBLIC PEACE - OTHER	8		
ROBBERY	5587		
ROBBERY - ARMED	10		
SEX OFFENSE - OTHER	327		
SEXUAL PENETRATION NONFORCIBLE - BLOOD / AFFINITY	5		
SEXUAL PENETRATION NONFORCIBLE - OTHER	9		
SOLICITATION (ALL CRIMES EXCEPT PROSTITUTION)	21		
THREATS / HARASSMENT BY USE OF COMPUTER	3		
THREATS / HARASSMENT BY USE OF TELEPHONE	23		
UNAUTHOIRZED USE OF A MOTOR VEHICLE (JOY RIDING)	3		
WEAPONS - CARRYING A CONCEALED WEAPON (CCW)	9		
WEAPONS - FIREARM IN AUTOMOBILE (CCW)	3		
WEAPONS OFFENSE - CONCEALED	2381		
WEAPONS OFFENSE - OTHER	530		

Figure 4.20: Violent/Non-violent Crimes

As previously mentioned, there can be multiple crimes committed during a single incidence, so the duplicate entries are removed reducing the final dataset of violent crimes to 61618 crime reports whose geographical distribution can be seen in figure 4.21. Again, the Jenks natural breaks classification gives a better indication of the violent crime spread across Detroit.


Detroit City number of Violent Crimes by Census Tract 2018
2.5 0 2.5 5 7.5 10 km
Jenks Natural Breaks Classification
5 - 134
134 - 224
224 - 331
331 - 557
557 - 1049
Detroit City number of Violent Crimes by Census Tract 2018
2.5 0 2.5 5 7.5 10 km
Equal Interval Classification
5.0000 - 213.8000
213.8000 - 422.6000
422.6000 - 631.4000
631.4000 - 840.2000
840.2000 - 1049.0000

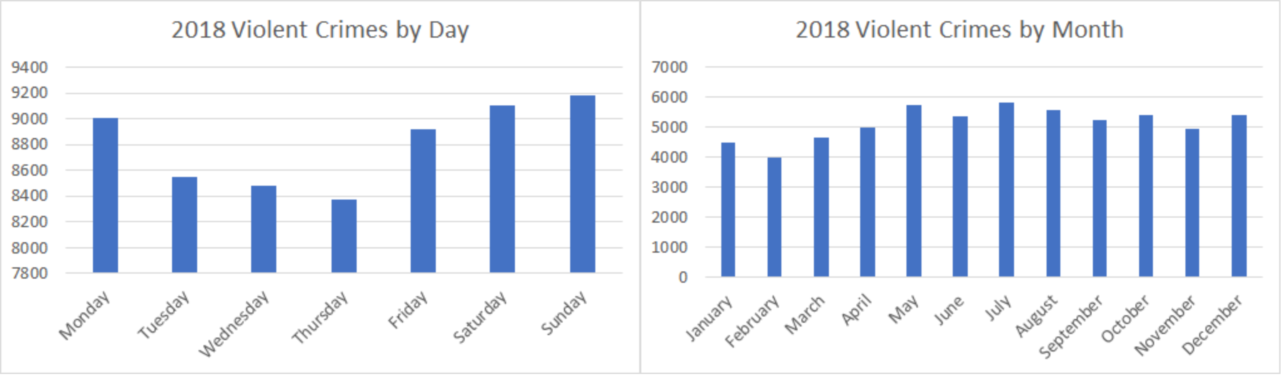
(a) Jenks Natural Breaks Classification

(b) Equal Interval Classification

Figure 4.21: 2018 Detroit PD Violent Crimes by Census Tract

Violent Crime Trends

Looking at the daily figure 4.22a, monthly figure 4.22b and hourly figure 4.23 violent crime trends, it can be seen that crimes peak at the weekend, during the summer months and just after midnight. This would seem to support the research linking violent crime with temperature and alcohol.



(a) Daily Totals

(b) Monthly Totals

Figure 4.22: 2018 Detroit Violent Crimes

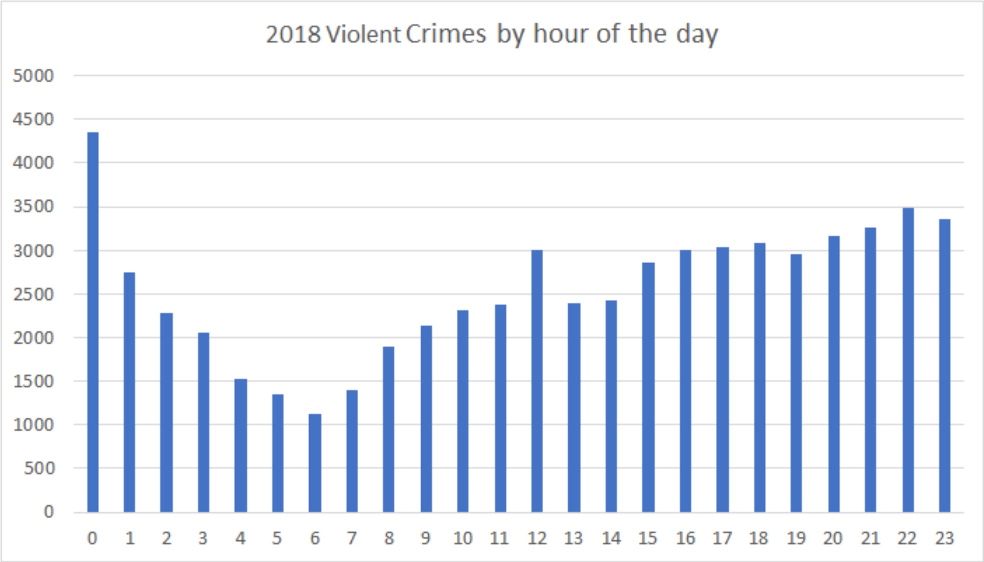


Figure 4.23: 2018 Violent Crimes by hour

4.3.3 Matching 911 Calls to Crimes

As previously stated, in order to perform supervised learning, it must be known whether each 911 call resulted in a crime report or not. For the purpose of this project it is intended to match 911 calls and crimes in a similar manner as (Wu & Frias-Martinez, 2018) by applying geospatial and temporal analysis to determine geospatial and temporal proximity of 911 calls and crime reports. It was better to temporally match 911 calls and crime reports first before checking for geospatial proximity, because to check geospatial proximity would require matching every 911 call with every crime report.

Temporal Relationship between 911 Calls and Crime

In order to match a crime report with a 911 call, both incidents must occur in the same time frame. To match these two incidents, the time of each 911 call is compared to the crime dataset and crimes within 30 minutes (either before or after the 911 call) are considered to be possible matches. Due to the large datasets, 653,764 911 calls and 61,477 crime reports, the logic and processing required to do that was challenging and did not work using standard R data frames. Using SQL with a MySQL server was investigated as a potential solution using a small sample of the data. It was possible to write simple SQL to relate 911 calls and crime reports by their time. However, there were issues in importing the entire datasets into MySQL. The default import wizard was too slow and there were security settings when using external tools which blocked their use. For this reason, R was investigated further with a solution eventually being found using the following design steps:-

1. R list objects were used instead of standard data frames as suggested by (Burns, 2011).
2. Reverse Processing. The temporal matching was performed in reverse, i.e. the smaller crime report dataset was matched against the larger 911 calls.
3. The data was split by month and temporal matching performed one month at a time, an extra day was included at either end to include matches past midnight at the start and end of each month. Attempting to temporally match the complete dataset did not complete after several hours.

The results of the temporal matching, table 4.3, show that there was a seven-fold matching, i.e. each 911 call in Detroit had seven potential matching crimes within ±30 minutes time frame. So, approximately 600,000 911 calls when temporally matched with 61,000 Crimes resulted in 4.9 million potential pairs within 30 minutes and 2.5 million matches within 15 minutes. For the purpose of this research, it was decided to consider that 911 calls and crime reports were potential matches if they occurred within 15 minutes of each other.

Of course, these incidents could be in different parts of the city and not related, so the next step was to geospatially match the incidents.

Month	911 Calls	Crimes	+/- 30 min	+/- 15 min
Jan	43522	4487	281649	146362
Feb	40209	3972	254646	131832
Mar	48674	4652	326569	168818
Apr	53060	4990	395401	204912
May	61756	5725	514140	266019
Jun	58185	5365	469252	242861
Jul	58787	5829	491156	254914
Aug	58245	5574	474510	246172
Sep	58314	5241	463311	240519
Oct	59197	5413	466351	241708
Nov	55931	4943	414610	214651
Dec	57884	5286	443108	229328
Total	653764	61477	4994703	2588096

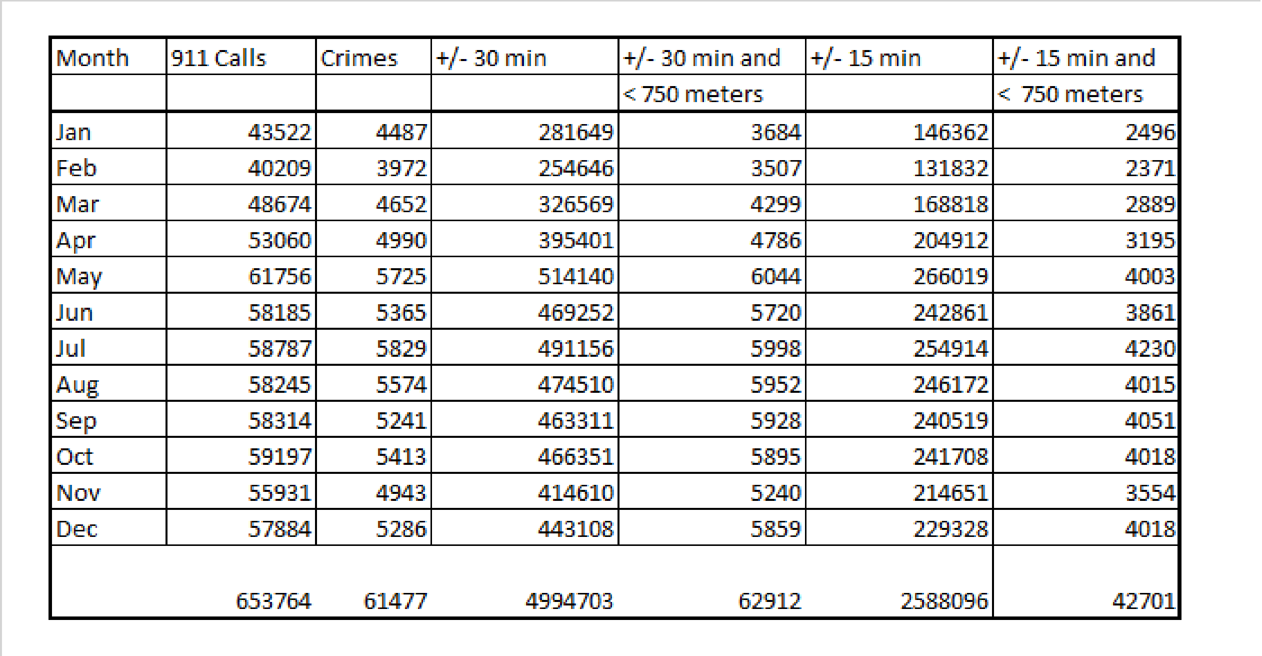
Table 4.3: 911 to Violent Crime Report Temporal Matches

Geospatial Proximity of 911 Data and Crime Data

In addition to temporal proximity, in order to further assume a relationship between the 911 calls and crimes, the distance between the two events is considered. Both the crime and 911 datasets are anonymised by the DPD to 100 block level. This does not translate to an exact measurement, but on examining several streets in Detroit, it was decided to consider the 100 block distance to be one tenth of a mile which is slightly over 300 meters. Allowing for a maximum possible anonymisation error of 650 meters and 100 meters between the 911 caller and the actual crime, a 911 call and crime will be considered to be geospatially related if they are within 750 meters of each other. Using the list of 911 to crime report temporal matches, the spherical law of cosines algorithm is applied to each pair of 911 call and crime report incidents. Crimes and 911 calls which are more than 750 meters apart are dropped as they are considered to be unrelated.

Temporal and Geospatial Matches

After running the initial analysis, matching 653,764 911 calls with 61,477 Crimes, there are 4,994,703 matches within ±30 minutes. Reducing, the metric to ±15 minutes, the number of matches reduced by almost 50% to 2,588,096. Applying a geospatial filtering of 750 meters, the number of matches is massively reduced to 62,912 for ±30 minutes and to 42,701 for ±15 minutes. Removing duplication, where 911 calls are matched with more than one crime, the number of matches is reduced from 42,701 to 41,066. Note that for this report, the metric of ±15 minutes will be used to link 911 calls with crime reports, i.e. the two events must occur within 15 minutes of each other.



Month	911 Calls	Crimes	+/- 30 min	+/- 30 min and	+/- 15 min	+/- 15 min and
				< 750 meters		< 750 meters
Jan	43522	4487	281649	3684	146362	2496
Feb	40209	3972	254646	3507	131832	2371
Mar	48674	4652	326569	4299	168818	2889
Apr	53060	4990	395401	4786	204912	3195
May	61756	5725	514140	6044	266019	4003
Jun	58185	5365	469252	5720	242861	3861
Jul	58787	5829	491156	5998	254914	4230
Aug	58245	5574	474510	5952	246172	4015
Sep	58314	5241	463311	5928	240519	4051
Oct	59197	5413	466351	5895	241708	4018
Nov	55931	4943	414610	5240	214651	3554
Dec	57884	5286	443108	5859	229328	4018
	653764	61477	4994703	62912	2588096	42701

Figure 4.24: Temporal Geospatial Matches of 911 Calls to Violent Crime reports

Temporal and Geospatial Matches by Priority

Examining the 911 calls which have been determined to have resulted in violent crime reports and grouping by priority, it can be seen that priority 1 911 calls are more likely to result in crime reports, see table 4.4.

911 Priority	Number of calls	Associated Crimes	% Crime Reports Created
Priority 1	72603	12264	16.89%
Priority 2	246385	13934	5.65%
Priority 3	286750	11958	4.17%
Priority 4	40649	2663	6.55%
Priority 5	7073	234	3.3%

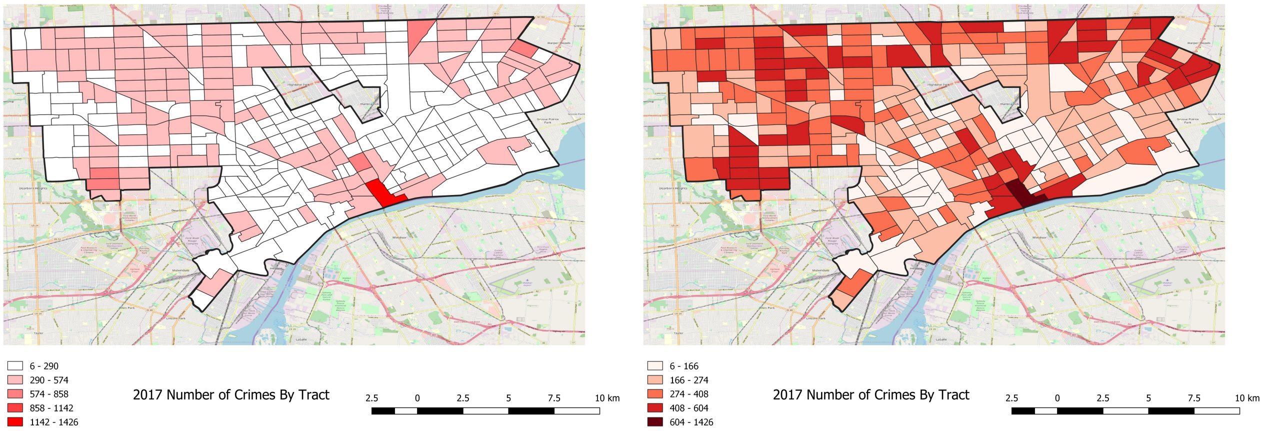
Table 4.4: Violent Crime Reports by 911 Call Assigned Priority

4.4 Data Preparation - Extended Dataset

The initial dataset containing only 911 Call Data and the Crime Report dependent variable is now extended to included temporal and geospatial features related to the location and time of the 911 calls in order to investigate the research question.

4.4.1 Crime Hotspot

Geospatial analysis is applied to historic crime in order to construct a neighbourhood crime hotspots map. The crime dataset is cut to include only crimes from 2017. The census tracts for each crime are computed and then the number of crimes per census tract is calculated. Historical crime hot spots by census tract are shown in figure 4.25. The historic crime volume can then be referenced to each 911 call using the census tract GEOID.



(a) Equal Interval Classification

(b) Jenks Natural Breaks Classification

Figure 4.25: 2017 Detroit Crimes

4.4.2 US Census data

In order to use and explore the US census data it requires significant processing to extract the required variables. The US census releases the data in geodatabase format [11] which is a collection of files with both geospatial and non-geospatial data. The file is first imported to QGIS and then seven data files, out of the 28 available, related to each of the data types to be utilised are exported to csv files for further processing using R Studio. The seven files contain data related to age, race, immigration status, education, housing, poverty and employment. Each of the seven csv files contains data for the entire state of Michigan, so it is first filtered by performing a left outer join between the Detroit City objects and the Michigan census Data therefore resulting only in Detroit census data, leaving 297 rows each representing one of the Detroit City census tracts. The original data is presented in detailed raw form, for example, it specifies the exact number of 18 and 19 year old males by tract so the data requires significant processing in order to utilise it further.

Note that four census tracts have a population of zero meaning that subsequent calculations will lead to zero or NA.

[11] http://desktop.arcgis.com/en/arcmap/10.3/manage-data/administer-file-gdbs/file-geodatabases.htm

Age and Gender Data

The following data per census tract is derived and extracted from the Age and Sex table for Detroit. The population density in people per km^2 see figure 4.26, median age of the population and median age for both males and females. The percentage of males and females, the ratio of females to males, the number of juvenile males where juvenile males are considered to be males aged 15 to 24 years old. Summary statistics for this data can be seen in figure 4.27.

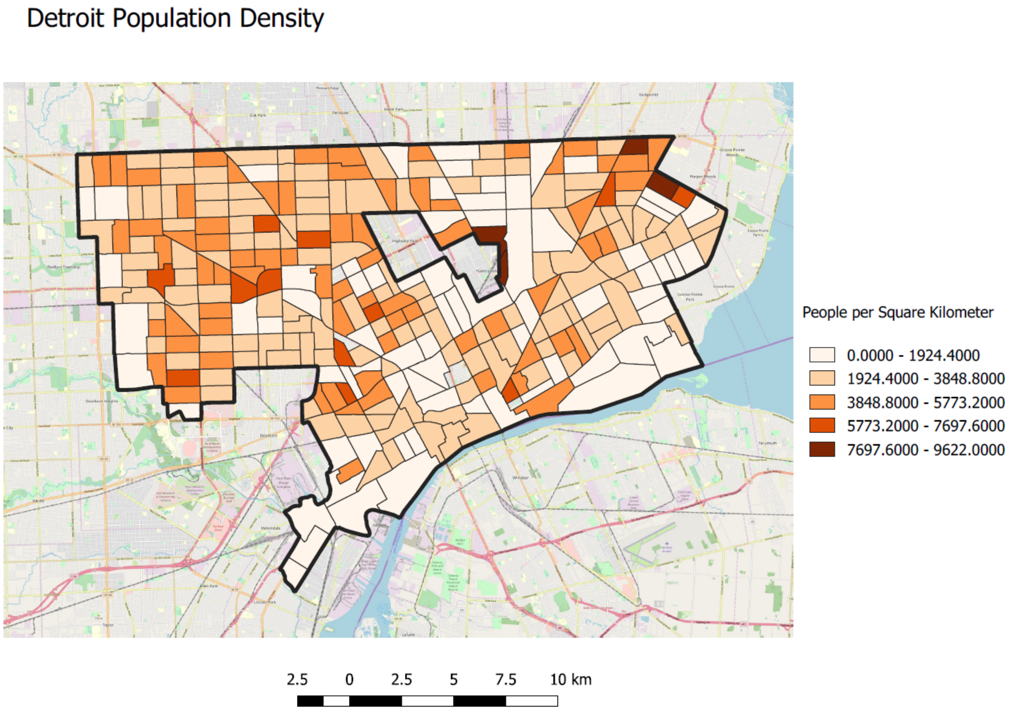
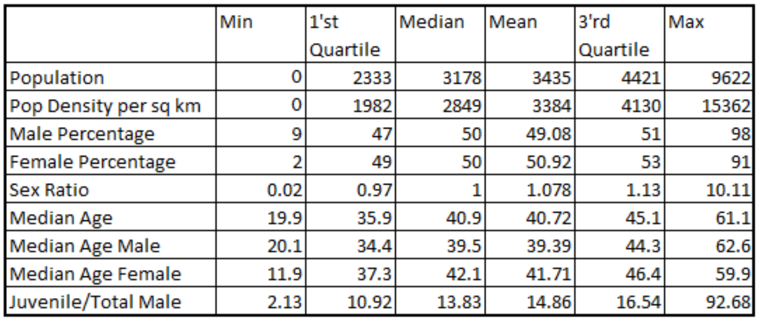


Figure 4.26: Detroit Population Density



	Min	1'st Quartile	Median	Mean	3'rd Quartile	Max
Population	0	2333	3178	3435	4421	9622
Pop Density per sq km	0	1982	2849	3384	4130	15362
Male Percentage	9	47	50	49.08	51	98
Female Percentage	2	49	50	50.92	53	91
Sex Ratio	0.02	0.97	1	1.078	1.13	10.11
Median Age	19.9	35.9	40.9	40.72	45.1	61.1
Median Age Male	20.1	34.4	39.5	39.39	44.3	62.6
Median Age Female	11.9	37.3	42.1	41.71	46.4	59.9
Juvenile/Total Male	2.13	10.92	13.83	14.86	16.54	92.68

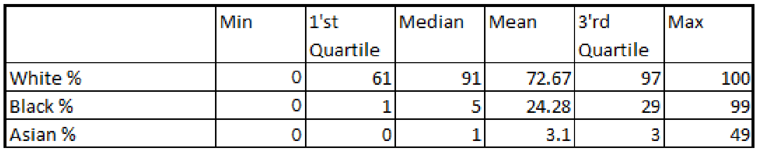
Figure 4.27: Detroit Census Tracts Age and Gender

Racial Data

The US Census Bureau present a very detailed breakdown of statistics on race allowing a person to indicate several races and nationalities in their ethnicity. For the purpose of this project, this data is simplified to three races/ethnicities, white, black and Asian derived from the census variables:-

- B02008e1 - WHITE ALONE OR IN COMBINATION WITH ONE OR MORE OTHER RACES
- B02009e1 - BLACK OR AFRICAN AMERICAN ALONE OR IN COMBINATION WITH ONE OR MORE OTHER RACES
- B02011e1 - ASIAN ALONE OR IN COMBINATION WITH ONE OR MORE OTHER RACES

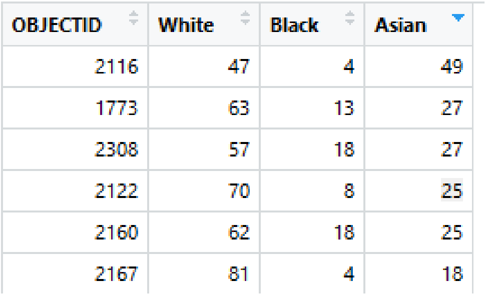
For Detroit, the summary of race percentages per census tract is in figure 4.28. As can be seen most census tracts in Detroit are predominantly white, although there are some where the predominant race is black or Asian.



	Min	1'st Quartile	Median	Mean	3'rd Quartile	Max
White %	0	61	91	72.67	97	100
Black %	0	1	5	24.28	29	99
Asian %	0	0	1	3.1	3	49

Figure 4.28: Detroit Census Tracts by Race

Examining the Asian percentages, five census tracts are greater than 25% Asian. The following, figure 4.29, shows the census tracts with the highest Asian populations in Detroit.

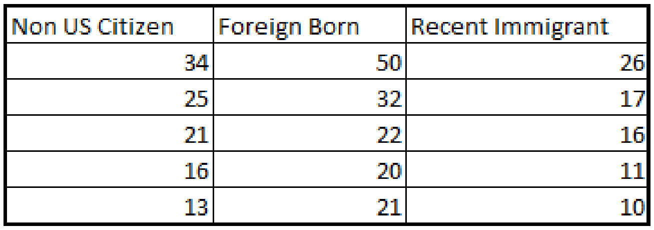


OBJECTID	White	Black	Asian
2116	47	4	49
1773	63	13	27
2308	57	18	27
2122	70	8	25
2160	62	18	25
2167	81	4	18

Figure 4.29: Detroit by Asian Race

Foreign Born Data

The Census Bureau collect information on citizenship, place of birth and when immigrants first entered the US. This data is processed to indicate the percentage of foreign born and non-US citizens per tract and the number of recent (since 2010) immigrants to the US. As would be expected, the majority of tracts have low numbers of foreign born and/or non-US citizens, although there are several tracts, with substantial numbers of immigrant/non-US citizens, see figure 4.30.



Non US Citizen	Foreign Born	Recent Immigrant
34	50	26
25	32	17
21	22	16
16	20	11
13	21	10

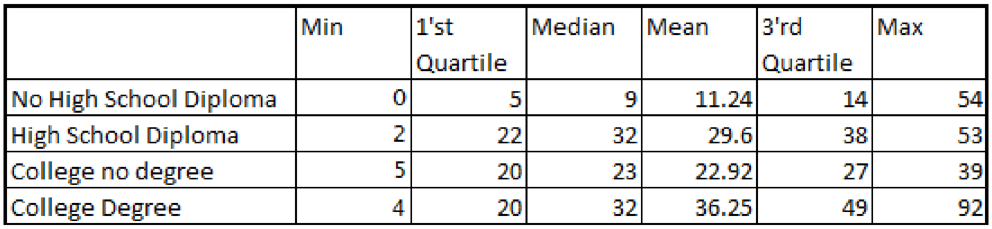
Figure 4.30: Selected Tracts and Immigration Status by % of Tract Population

Education Data

The data presents the number of individuals at **their highest level** of education achievement. Detailed data is presented for each tract to 25 different education levels, for example highest education to 10th grade. This data is analysed and grouped to four approximate educational level as follows:-

- Education to 12th Grade or less
- High School Diploma
- College - No Degree
- College - Graduated

The percentages at each education level are calculated as a percentage of people in each tract aged who are over 25 years of age, in order to discount people from the calculations who are possibly still in education. The US national average for having obtained at least a high school diploma is 87.3% and for obtaining a degree is 30.9% which is comparable to the Detroit figures as shown in figure 4.31.



	Min	1'st Quartile	Median	Mean	3'rd Quartile	Max
No High School Diploma	0	5	9	11.24	14	54
High School Diploma	2	22	32	29.6	38	53
College no degree	5	20	23	22.92	27	39
College Degree	4	20	32	36.25	49	92

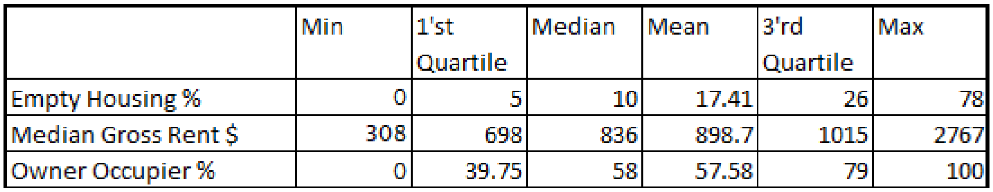
Figure 4.31: Detroit Educational Achievement

Housing Data

There are a number of variables associated with housing which will give good indications as to the overall housing characteristics for the neighbourhood. The following data is calculated and derived.

- Median Gross Rent
- Percentage of empty housing
- Amount of owner occupiers

The number of empty buildings in certain parts of Detroit due to economic problems is well publicised and can be seen in figure 4.33. There are several areas in the city with high levels of vacant buildings. Comparing this to the map of Detroit demolitions shown earlier in figure 2.1, it is obvious that certain areas of the city are in decline, for example the south west corner of the city. Summary statistics relating to housing per census tract are shown in 4.32.



	Min	1'st Quartile	Median	Mean	3'rd Quartile	Max
Empty Housing %	0	5	10	17.41	26	78
Median Gross Rent $	308	698	836	898.7	1015	2767
Owner Occupier %	0	39.75	58	57.58	79	100

Figure 4.32: Detroit Housing indicators

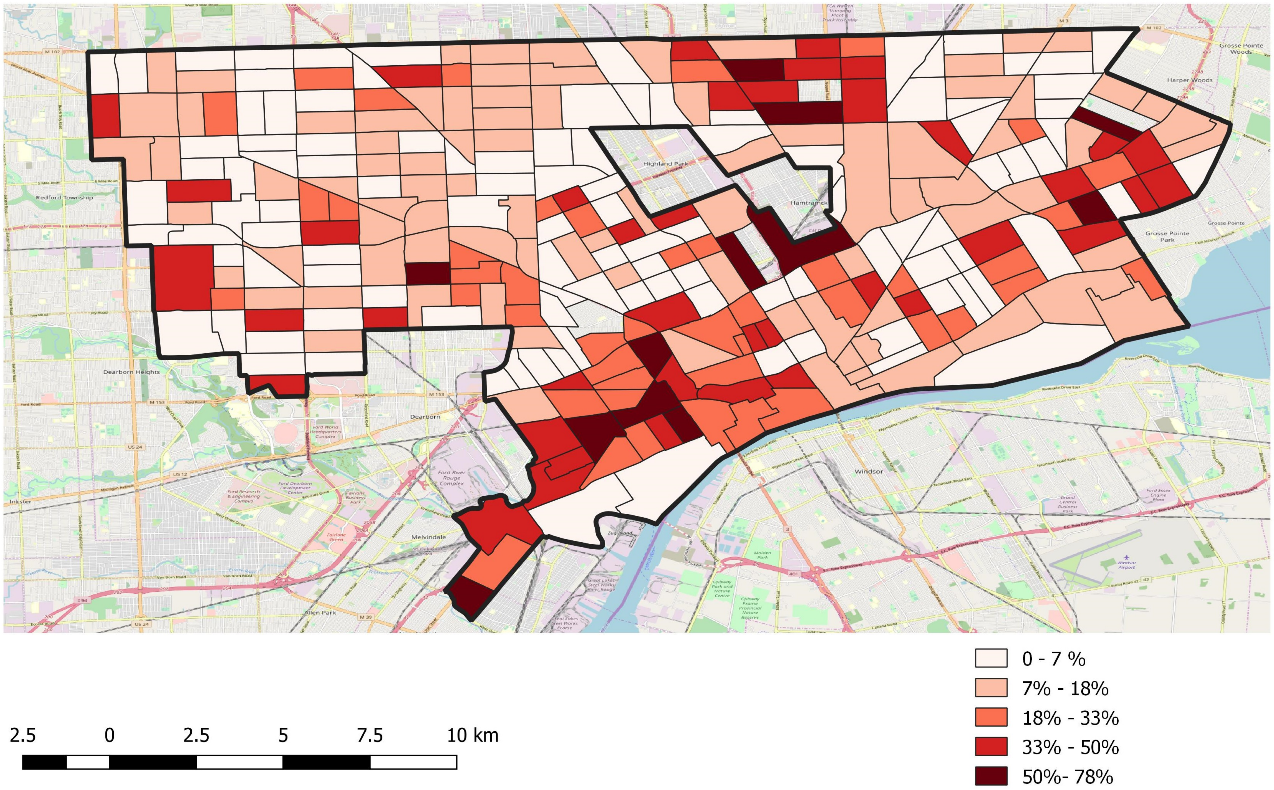
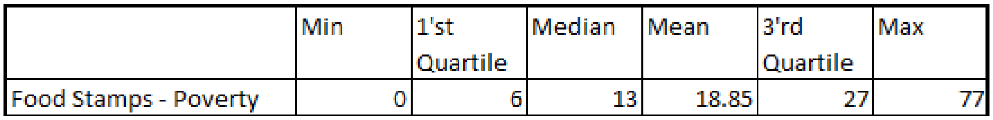


Figure 4.33: Detroit Vacant Housing

Poverty Indicators

The census indicates the number of households which are in receipt of food stamps due to poverty. Grouping by census tracts gives the following statistics, see figure 4.34.

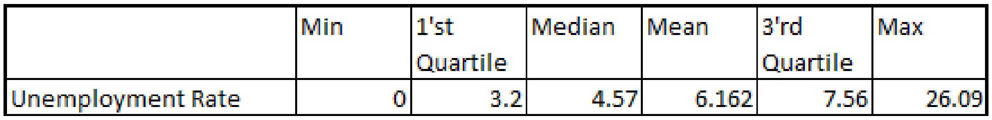


	Min	1'st Quartile	Median	Mean	3'rd Quartile	Max
Food Stamps - Poverty	0	6	13	18.85	27	77

Figure 4.34: Households in Receipt of Food Stamps %

Unemployment Data

Statistics relating to employment are given by gender for several age ranges. This includes numbers available for work, in the military, working and unemployed. The numbers unemployed for each age group are added to calculate the total unemployment rate per tract, see figure 4.35 .



	Min	1'st Quartile	Median	Mean	3'rd Quartile	Max
Unemployment Rate	0	3.2	4.57	6.162	7.56	26.09

Figure 4.35: Unemployment Rate per Tract %

Household Income Data

The median income per household and percentage of households on an income of less than $60,000 is calculated, figure 4.37 gives summary statistics of income, while figure 4.36 shows the income distribution across the city. Again, these figures can be matched to each 911 call using the census tract GEOID.

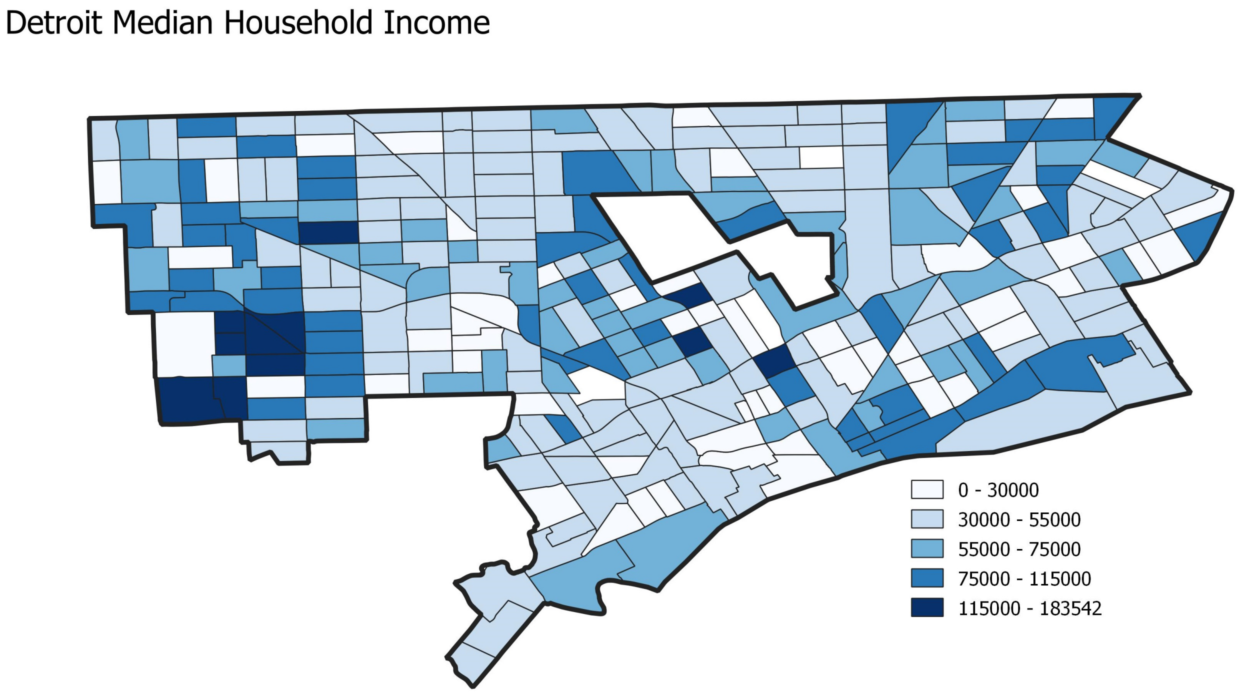
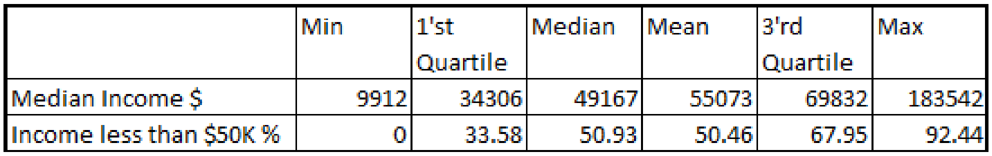


Figure 4.36: Household Income $ per Tract



	Min	1'st Quartile	Median	Mean	3'rd Quartile	Max
Median Income $	9912	34306	49167	55073	69832	183542
Income less than $50K %	0	33.58	50.93	50.46	67.95	92.44

Figure 4.37: Incomes per Tract

4.4.3 Points of Interest/Bars

OSM shapefiles will be used to extract POIs (Points of Interest) for Detroit. Examining the OSM shapefile layers, there are two files for POIs. Combining the files and displaying the data, see figure 4.38, you can see that the difference between the two files is that one contains polygons (irregular shapes) and the other contains points, both of which are used to represent bars. The irregular shape most likely illustrates the footprint of the buildings.

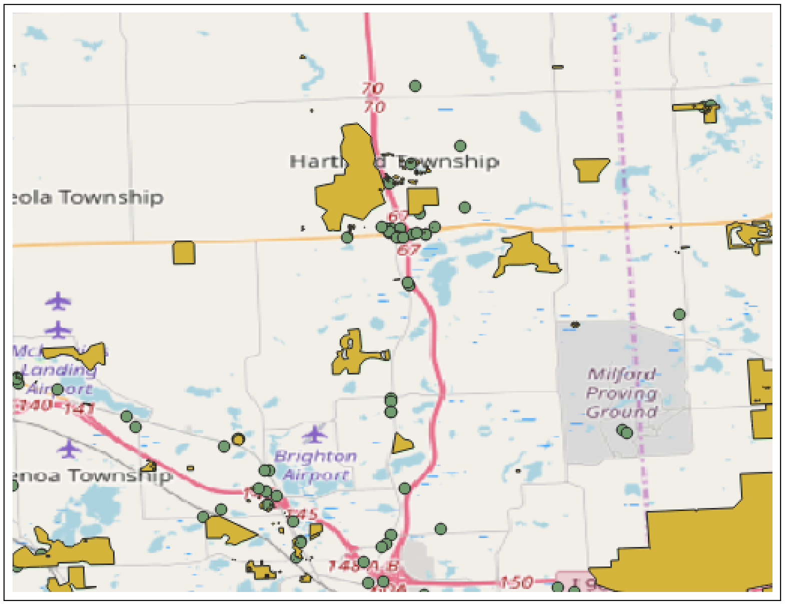


Figure 4.38: Points of Interest - Points and Polygons

Using QGIS, these two shapefiles containing points and polygons were clipped to include only locations in the Detroit area. Bars immediately outside the city are included as they may have an influence on neighbourhoods just inside the City limits. They were further filtered by type, to include only entertainment venues and the polygons converted to centroid points resulting in a list of entertainment venues in point format, see figure 4.39.

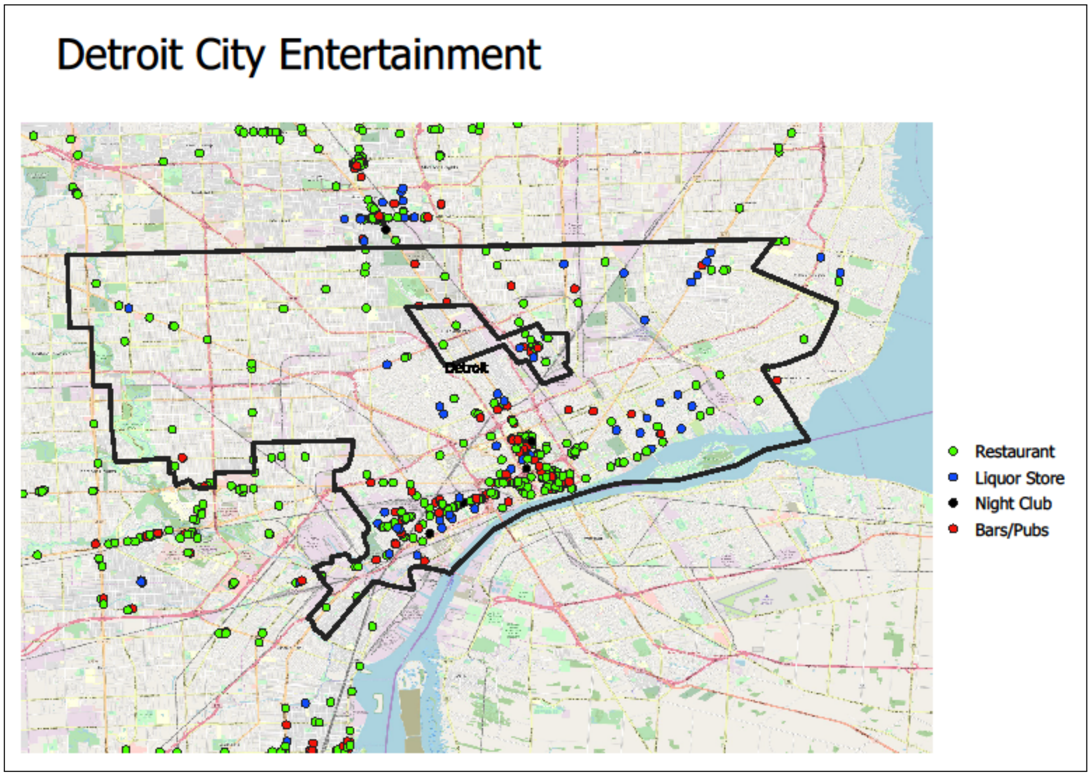


Figure 4.39: Entertainment Venues in the Detroit Area

The objective is to determine the number of bars close to each 911 call.

One option would be to create a circular buffer around each 911 point and then for each 911 point determine the number of bars in that buffer. But with several hundred thousand 911 calls, this would require a large computational effort.

Another option would be to map the bar locations to their census tracts and calculate the number of bars in each census tract and then link this to the 911 calls. However, census tracts vary in size, so the metrics would not be equivalent in each location. Plus, this would not account for bars and 911 calls which are close to the boundaries of the census tracts.

A better option is to divide the city into a number of small equally sized cells and aggregate the number of bars in each cell. Then for each cell construct an index which represents the number of bars in that cell and in the neighbouring cells. The 911 points could then be linked to this index by determining which cell of the grid that each 911 call is located.

Using QGIS, a grid is superimposed on Detroit see figure 4.40.

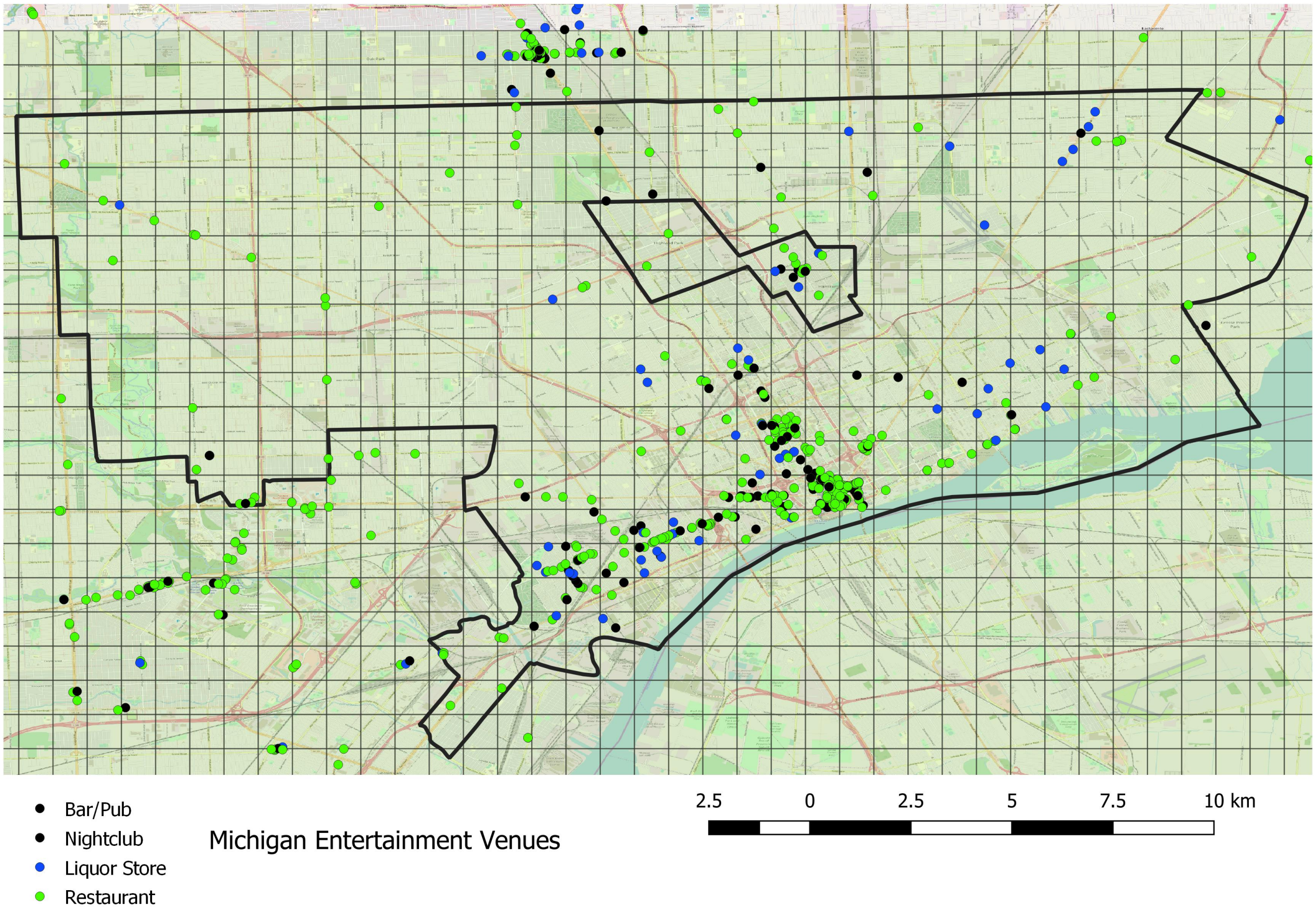


Figure 4.40: Detroit Area Entertainment and Grid

The bars, pubs and nightclubs are then grouped according to their location in the grid and counted per cell. It was decided not to include restaurants and liquor stores as the research associated only bars to crime. An algorithm was then developed in R, to count how many bars were in each cell and how many bars were in each of the eight neighbouring cells. For example, see downtown Detroit in figure 4.41 where the first number indicates the number of bars in that cell and the second indicates the number of bars in the adjacent cells.

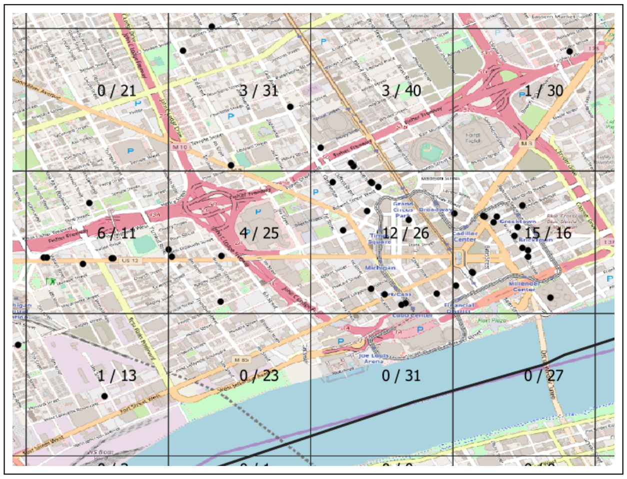


Figure 4.41: Number of Bars in Cell/Neighbours

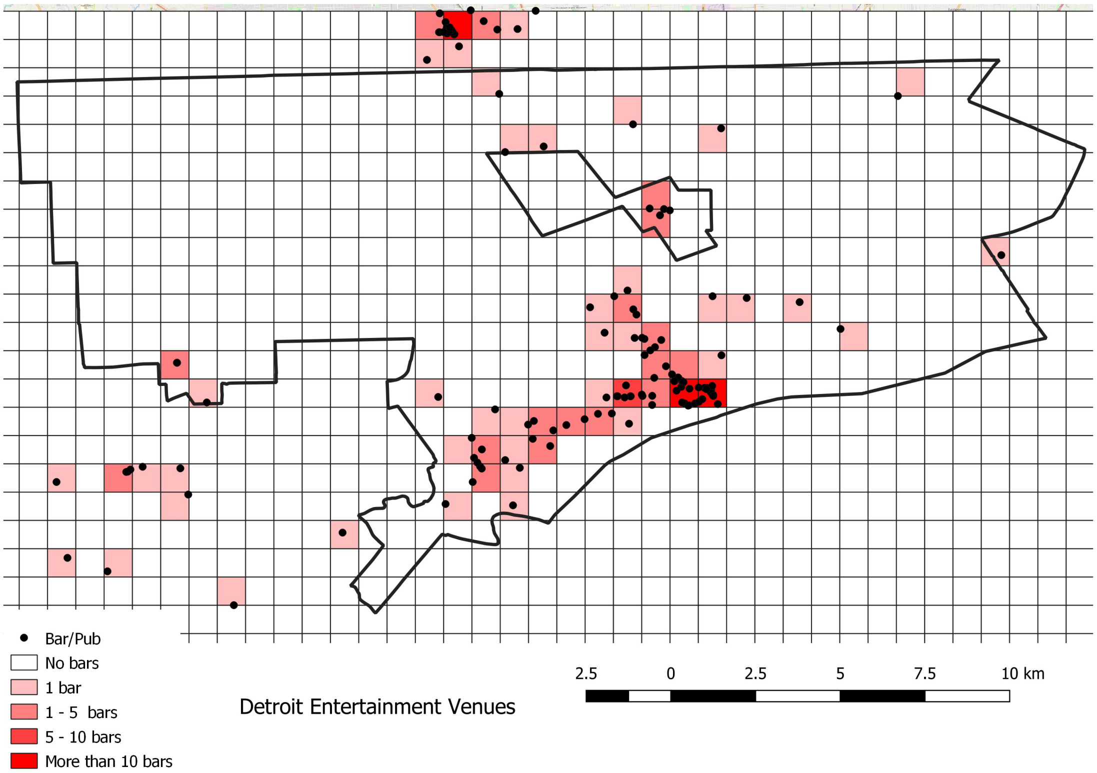


Figure 4.42: Absolute number of Bars per Cell Graduated

Using a graduated legend, the number of bars per cell in Detroit is shown in figure 4.42. This shows that bars are mainly clustered in the south, however, there is a cluster of bars in the north which are outside the Detroit City boundary.

To account for bars in neighbouring areas, an index for each cell, which indicates the number of nearby bars, is created equal to the number of cells in that cell plus the number of bars in each adjacent cell divided by two. This bar index is illustrated in figure 4.43. Note the high concentration of bars in down town Detroit on the south-eastern border. To consider bars just outside the Detroit City boundary, when constructing the grid, a full cell outside the boundary is included. Each grid is 0.01 Degrees which is approximately 1100 meters so bars within 1.1km of the Detroit City boundary are included in the analysis as neighbouring cells.

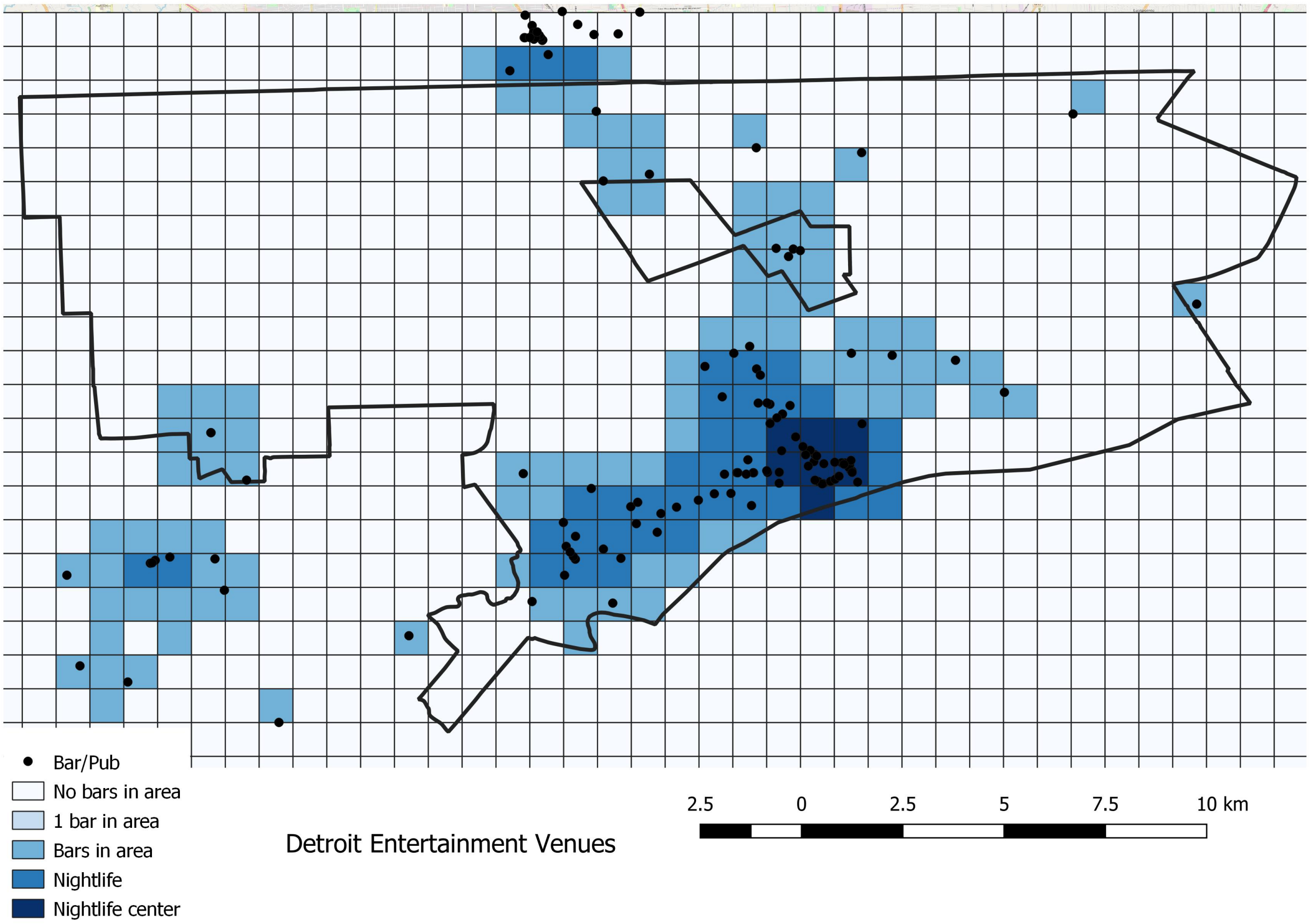


Figure 4.43: Bars/Nightlife Index

Mapping of 911 Calls to Bars Index

To relate each 911 call to the index of bars, QGIS is used to map the location of each 911 Call to the grid location by performing an intersection operation between the Detroit Grid and the 911 Call Data. The resulting join can be exported as a csv file containing the 911 reference and grid reference enabling a join of the 911 data to the bar indices. Each 911 call will then have a feature which indicates the number of bars in the area of the 911 call.

4.4.4 Weather Data

The weather data from 2018 was downloaded from National Centers for Environmental Information and matched to each 911 Call by the date. Each 911 call record will then contain, features for the precipitation, maximum and minimum temperature for the day of the 911 call.

4.4.5 Holiday Data

Using a list of Federal holidays, figure 4.14, each 911 call record is checked to see if the call happened on the day of a holiday or the day preceding a holiday. New variables *isHoliday, isHolidayEve* and *isWeekend* are imputed from the date of the 911 call.

4.5 Final datasets

Both datasets contain 653,764 rows, whilst the extended dataset is made up of 76 features and the dependent variable *violent crime report* from six diverse data sources. See appendix C for a complete list of the features in the extended dataset.

However, there are several factors to consider before creating a model from these datasets.

- Officer Initiated Calls
- Highly Correlated Features
- Imbalanced Data

4.5.1 Officer Initiated Calls

As previously discussed, the DPD use the 911 call system to record administrative events which are marked as officer initiated records, but officer initiated records can also be for potential criminal activity. An attempt was made to filter the administration events from the dataset. However, as can be seen in table 4.5 the 911 / crime ratio is 9.54% for public initiated calls versus 3.47% for officer initiated calls. This is discussed further in this document.

	Total	Crime		No Crime	
Total 911 Calls	653,764	41,066	6.28%	612,698	93.72%
Officer Initiated 911 Calls	351,104	12,179	3.47%	338,925	96.53%
Public Initiated 911 Calls	302,660	28,887	9.54%	273,773	90.46%

Table 4.5: Breakdown of 911 Calls

4.5.2 Correlated Features

To this point, there has been little feature selection performed and it is probable that there is a certain amount of collinearity in the variables. As (Næs & Mevik, 2001) explain, collinearity between variables in a model can be a problem for prediction and classification leading to poor results. A common approach and one suggested by (Næs & Mevik, 2001) is to use principal component analysis to reduce collinearity, however as a first step, the dataset is reduced by removing features which are highly correlated with other features but presented in different formats, for example percentages and absolute numbers.

Feature Selection

The dataset is first reduced by removing obviously correlated and redundant features, also removed are features such as unique reference ids and dates which are of no relevance in a model, see Appendix E for a list of features removed.

Correlation Matrix

Using R, a correlation matrix for the remaining features was created, which indicated some correlation between the features see Appendix D.

Temperature

As you would expect and as can be seen in figure 4.13 there is a strong positive correlation (0.95) between the daily maximum and minimum temperatures. The research indicates that violent crime is associated with warm temperatures so it should not affect the model if the minimum temperature is removed from consideration.

Bars

There are two features to measure the number of bars, one which is an absolute measure of the number of bars in the immediate area of the call and the second, an index which applies a weighting to bars in the adjacent areas. Interestingly, the number of crimes in 2017, per census tracts shows high positive correlation with both of these features which represent the number of bars, 0.65 for *NumBars* and 0.51 for *BarIndex*. This would support the research and neither of these features are dropped.

Median Age

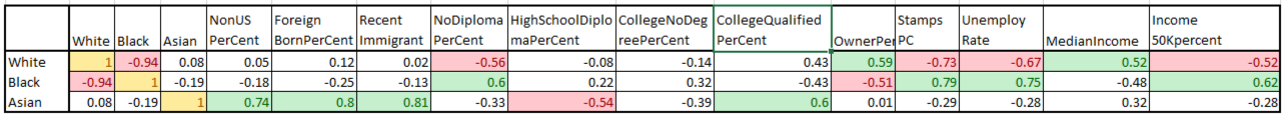
There are three features for the median age in the census tract, one each for male and female and one for the total population. Also, included is the number of juvenile males in each census tract. The correlation between the three features representing median age is high but low (-0.29) between *malejuveniles* and *medianagemale*. For this reason, only *malejuveniles* and *medianAge* are kept.

Gender

The proportion of male and females are not shown as correlated but obviously there is a direct relationship. However there is a moderately strong correlation between average household size and proportion of males, so for this reason, proportion female is kept and proportion male dropped. In addition, the research linked racial crime to matriarchal family structures.

Race, Immigrant Status, Education, Income and Poverty

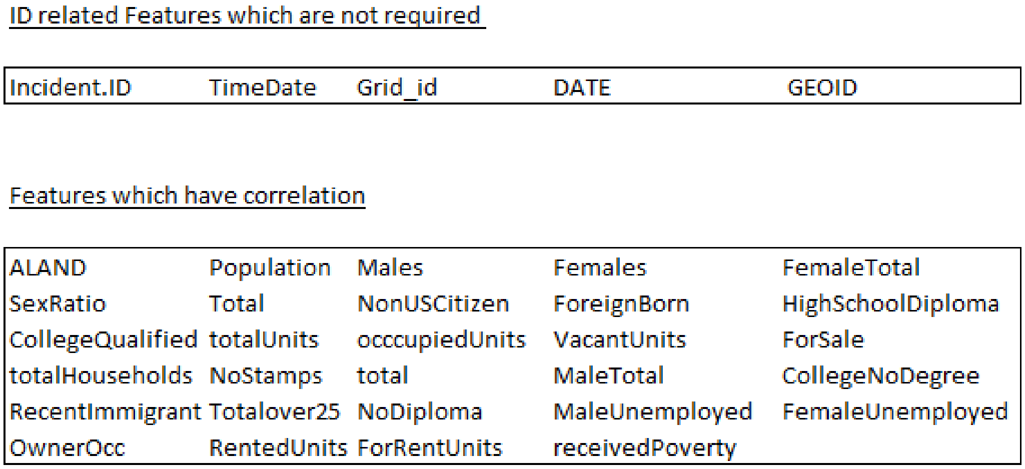
Based on the racial profile of Detroit, three racial groups are categorised, white, black and Asian. However, most census tracts in Detroit are mainly white or black, so there is high correlation between black and white. There is also high correlation between the Asian feature and nationality, immigration status and education which seems to indicate that a high number of immigrants are Asian with higher education. This observation is seen in other research, for example (Hagy & Staniec, 2002).
As the correlations are complicated, it is decided not to remove any of these features with the exception of *Income50Kpercent*. This is the percentage of households with an income less than $50,000. This information will be represented by *MedianIncome* for the census tract.



	White	Black	Asian	NonUS PerCent	Foreign BornPerCent	Recent Immigrant	NoDiploma PerCent	HighSchoolDiplomaPerCent	CollegeNoDegreePerCent	CollegeQualified PerCent	OwnerPer	Stamps PC	Unemploy Rate	MedianIncome	Income 50Kpercent
White	1	-0.94	0.08	0.05	0.12	0.02	-0.56	-0.08	-0.14	0.43	0.59	-0.73	-0.67	0.52	-0.52
Black	-0.94	1	-0.19	-0.18	-0.25	-0.13	0.6	0.22	0.32	-0.43	-0.51	0.79	0.75	-0.48	0.62
Asian	0.08	-0.19	1	0.74	0.8	0.81	-0.33	-0.54	-0.39	0.6	0.01	-0.29	-0.28	0.32	-0.28

Figure 4.44: Racial Correlations

The features removed are shown in figure 4.45.



ID related Features which are not required

Incident.ID	TimeDate	Grid_id	DATE	GEOID

Features which have correlation

ALAND	Population	Males	Females	FemaleTotal
SexRatio	Total	NonUSCitizen	ForeignBorn	HighSchoolDiploma
CollegeQualified	totalUnits	occcupiedUnits	VacantUnits	ForSale
totalHouseholds	NoStamps	total	MaleTotal	CollegeNoDegree
RecentImmigrant	Totalover25	NoDiploma	MaleUnemployed	FemaleUnemployed
OwnerOcc	RentedUnits	ForRentUnits	receivedPoverty	

Figure 4.45: Features Removed from Initial Dataset

4.5.3 Dimension Reduction using PCA

The extended dataset is further reduced in dimensions by applying principal component analysis. In order to perform PCA, the dataset is split into training and testing and then further divided into two datasets each containing only categorical or only numeric data since PCA is applied only to numeric data. It is also important to note, that the principal components should be calculated using the **training dataset**, but that the PCA transformation should then be applied to the **test data**. The principal components are calculated using the numerical test data creating 31 principal components, see Appendix F for first ten principal components. This shows the contribution from each of the original features to each of the first ten principal components.

Examining the principal components, see figure 4.46 PC1 (the red star) accounts for 27.9% of the variance in the dataset and PC2 (red circle) explains 11.98% .

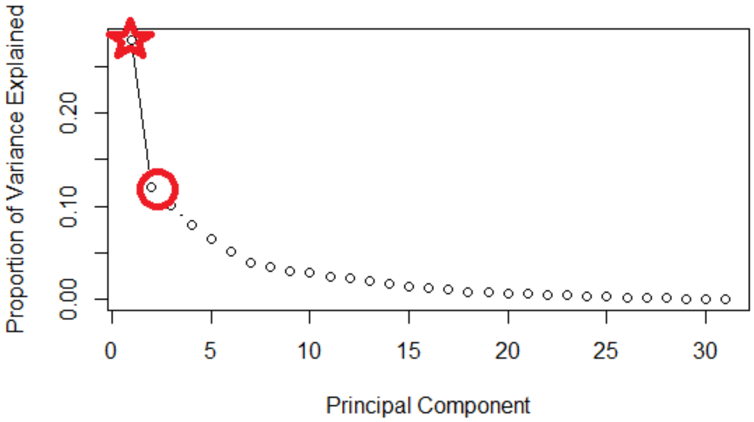


Figure 4.46: Variance Explained by Principal Components

Cumulatively, figure 4.47, shows that the first 25 principal components explain over 95% of the variance. For this research it was decided that it is reasonable to replace the 31 numerical features with 22 principal components.

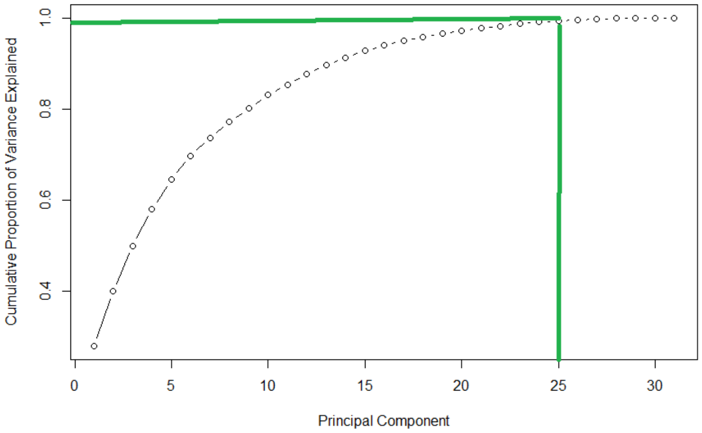


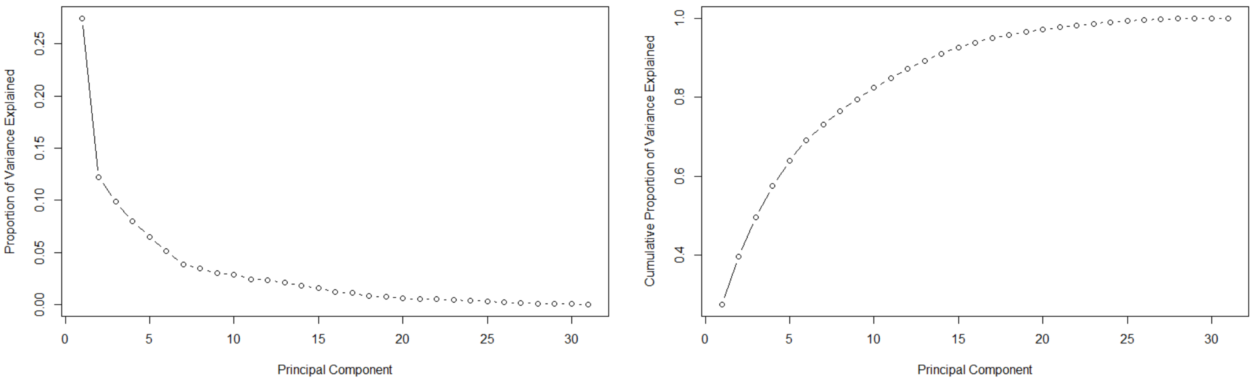
Figure 4.47: Cumulative Variance by Principal Components

Code for PCA graphs was obtained from [12].

[12] www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/

4.5.4 Public Initiated 911 Calls Only

In order to create models without the officer initiated calls, PCA is separately applied to the public only dataset (numeric data). The results are similar, see figure 4.48. Looking at the cumulative plot figure 4.48b, the amount of variance continues to increase until 25 principal components.



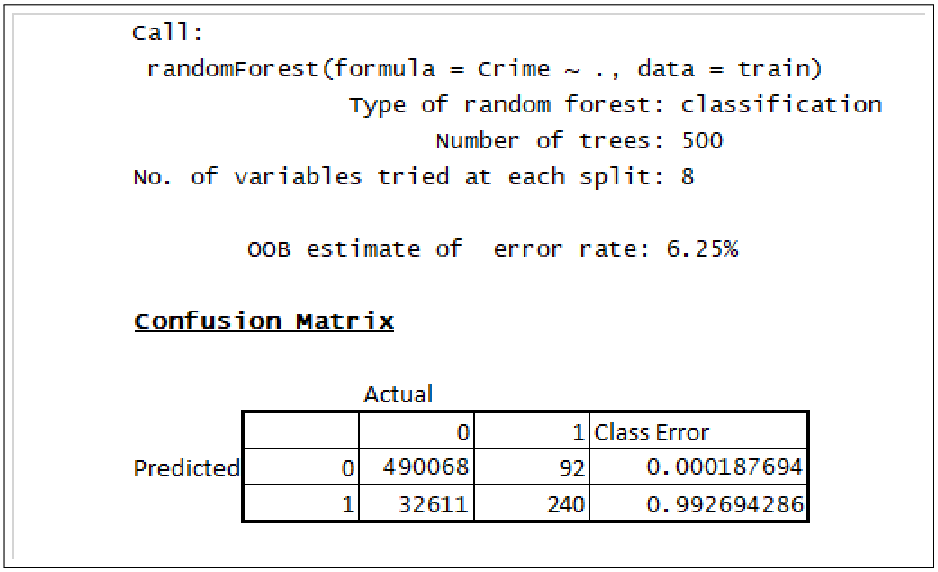
(a) Variance by Principal Component

(b) Cumulative Variance

Figure 4.48: Non Officer Initiated Calls

4.5.5 Imbalanced Data

As previously discussed, imbalanced data is a common problem in machine learning. The consequence of imbalanced data can be illustrated by an initial random forest model built with training data but no sampling, see figure 4.49.



```
Call:
 randomForest(formula = Crime ~ ., data = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 8

        OOB estimate of  error rate: 6.25%
```

Confusion Matrix

		Actual		
		0	1	Class Error
Predicted	0	490068	92	0.000187694
	1	32611	240	0.992694286

Figure 4.49: Baseline Random Forest Model

This model contains 500 trees and appears to have a low error, 6.25%. However, examining the class error the error for predicting Crime is 99%. The reason for this is the imbalance of the dataset. From our 523,011 rows in the 911 training set only 32,851 (6.3%) resulted in a crime. In effect, if the model was to predict zero (no crime) for every sample, it would achieve an accuracy of 93.7%. This problem is known as imbalanced data and is the subject of several papers.

4.6 Modelling

Multiple models were built and tested with variations in sampling, algorithm and by removing officer initiated calls. Experiments were performed as follows:

Officer Initiated Calls

Both the full dataset and the dataset with officer initiated calls removed.

Sampling and SMOTE

Five different sampling ratios (crime/no crime) including a dataset with synthetically created crime instances (SMOTE).

- Down sampling to 1:1 ratio
- Down sampling to 1:2 ratio
- Down sampling to 1:4 ratio
- SMOTE to create yes samples to obtain 10:13 ratio
- Full Imbalanced dataset - 93.7% No crime

Algorithms

Algorithms used included Random Forest, Logistic Regression and Support Vector Machines.
When building models for the extended dataset using SVMs, there were serious performance issues and very poor results when the data was imbalanced. Creation of an SVM model using the entire dataset took more than five days and predicted only the majority class i.e. *No Crime*. For this reason, experimentation with SVM models were limited and concentrated on down sampled training data with both officer initiated and public calls included.

Metrics

The primary metric for the project is sensitivity. Accuracy and specificity were also calculated for completeness.

Categorical Variables

Note also that when building random forests and logistic regression models, there are limitations in the number of categorical variables, so categorical features with high numbers of levels were dropped. For this reason, census tracts and call codes are dropped when building models with random forests and with logistic regression but retained for SVMs.

4.7 Receiver Operating Curve

The Receiver Operating Curve is a good way to explore the relationship between sensitivity and specificity. Note that SVMs do not provide probabilities (Platt et al., 1999). Plotting the ROC curve with the random forest model using public only 911 calls, the results are shown in figure 4.50. The diagonal line shows where the true positive rate is the same as the true negative rate and this model is better than that. The area under the curve is 63.9% which would be classed as *good*.

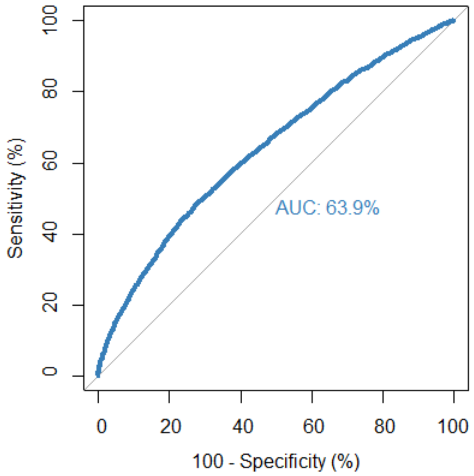


Figure 4.50: ROC Curve and AUC for Random Forest Model Public Calls

Sensitivity

If the priority was the True Positive Rate/Sensitivity then the threshold to predict crime could be reduced. For example, reducing the threshold of probability from 0.5 to 0.3 the following results are obtained.

		Actual	
		No Crime	Crime
Predicted	No Crime	5684	287
	Crime	48955	5606

This gives a sensitivity of 95.13% and specificity of 10.40% so the model would be good at predicting crime, but very poor at predicting no crime.

Specificity

Increasing the threshold from 0.5 to 0.8 would prioritise the negatives and get the following results. Conversely, this threshold gives a sensitivity of 9.06% and specificity of 97.61% so the model would be good at predicting no crime, but very poor at predicting crime.

		Actual	
		No Crime	Crime
Predicted	No Crime	53334	5359
	Crime	1305	534

4.8 Results

4.8.1 Tables of Results

The table 4.6 shows the metrics for the primary dataset (null hypothesis) which contains only 911 data and the dependent variable crime. Accuracy, sensitivity and specificity are shown. The table 4.7 shows the metrics for the extended dataset intended to prove or disprove the alternative hypothesis. The models are built with various sampling ratios and percentages of the yes class, as shown. The metrics are calculated using the confusion matrix which is obtained after the models are applied to the unseen test data. The best sensitivity results for both datasets are highlighted in green.

Sensitivity, accuracy and specificity are plotted for both datasets and it can be seen that sensitivity improves when down sampling reduces the imbalance but both accuracy and specificity decrease with down sampling to increase the minority class.

4.8.2 Graph of Sensitivity

The results for sensitivity with each sampling ratio and for each algorithm are shown in figure 4.51 for the 911 only data (null hypothesis) and in figure 4.52 for the extended dataset (alternative hypothesis).

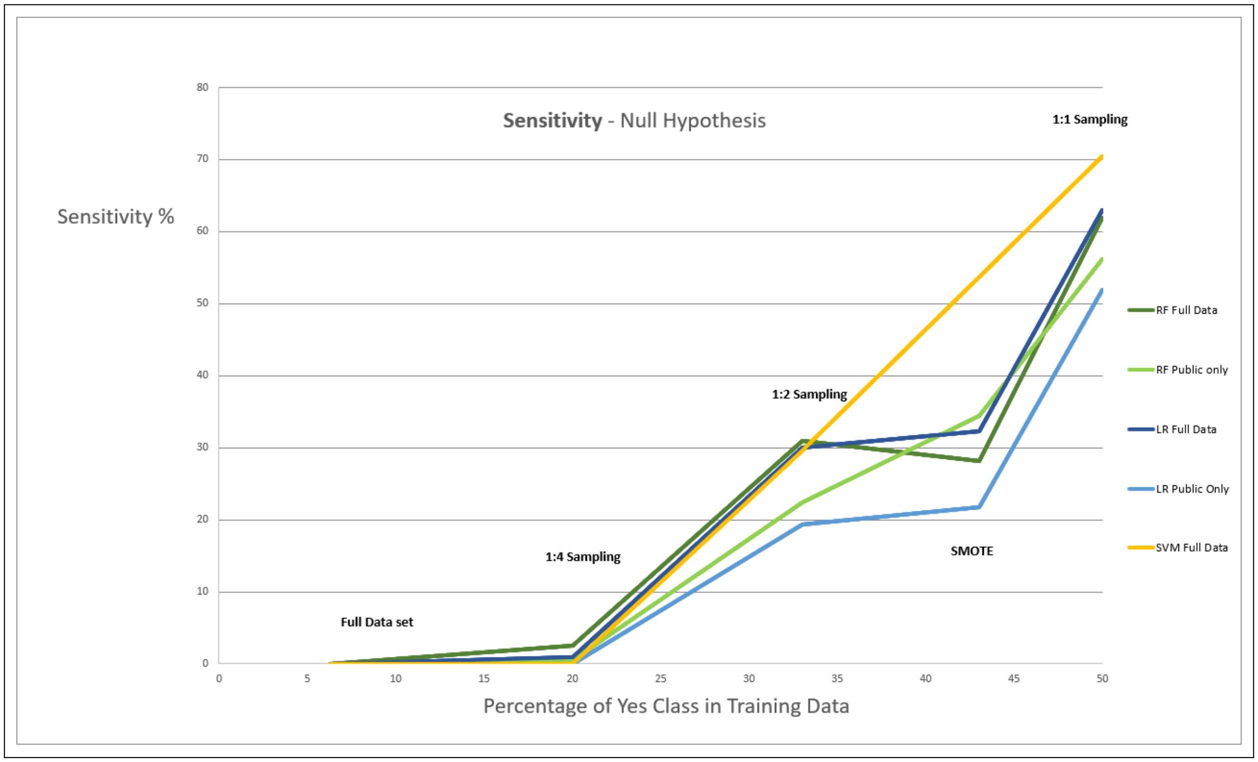


Figure 4.51: Sensitivity of Models for the Primary Dataset - Null Hypothesis

Algorithm	**Sampling**	**No. of**	**%**	**Test Data**		
(Officer Init.)	**Method**	**Samples**	**Yes**	**Accuracy**	**Sensitivity**	**Specificity**
Random Forest	Down 1:1	65704	50%	65.23%	61.97%	65.45%
(Officer Calls	Down 1:2	98556	33%	85.69%	30.98%	89.36%
included)	Down 1:4	164260	20%	93.28%	2.57%	99.36%
	Full Training	523011	6.3%	93.72%	0.00%	100.00%
	SMOTE	229964	43%	80.79%	28.12%	84.32%
Random Forest	Down 1:1	46186	50%	61.08%	56.19%	61.60%
(Public Calls	Down 1:2	69279	33%	82.62%	22.47%	88.98%
only)	Down 1:4	115465	20%	90.27%	0.47%	99.77%
	Full Training	242372	9.5%	90.43%	0.02%	99.99%
	SMOTE	160958	43%	75.75%	34.47%	78.52%
Log. Regression	Down 1:1	65704	50%	64.37%	62.98%	64.46%
(Officer Calls	Down 1:2	98556	33%	86.16%	29.97%	89.93%
included)	Down 1:4	164260	20%	93.52%	0.88%	99.73%
	Full Training	523011	6.3%	93.72%	0%	100.00%
	SMOTE	229964	43%	79.73%	32.23%	82.91%
Log. Regression	Down 1:1	46816	50%	67.51%	51.85%	48.15%
(Public Calls	Down 1:2	69279	33%	84.31%	19.41%	91.18%
only)	Down 1:4	115465	20%	90.43%	0.00%	100.00%
	Full Training	242372	9.5%	90.43%	0.00%	100.00%
	SMOTE	160958	43%	78.20%	21.69%	84.18%
SVM	Down 1:1	65704	50%	56.32%	70.43%	55.37%
(Officer Calls	Down 1:2	98556	33%	86.32%	29.63%	90.12%
included)	Down 1:4	164260	20%	93.72%	0.0%	100.00%
	Full Training	523011	6.3%	93.72%	0.0%	100.00%
	SMOTE *	-	-	-	-	-

Table 4.6: Null Hypothesis Models Built using Primary Dataset.

Algorithm (Officer Init.)	**Sampling Method**	**No. of Samples**	**% Yes**	**Accuracy**	**Sensitivity**	**Specificity**
Random Forest (Officer Calls included)	Down 1:1	65704	50%	60.38%	70.04%	59.73%
	Down 1:2	98556	33%	82.03%	42.91%	84.65%
	Down 1:4	164260	20%	90.78%	20.16%	95.51%
	Full Training	523011	6.3%	93.70%	5.59%	99.61%
	SMOTE	229964	43%	84.59%	30.56%	88.21%
Random Forest (Public Calls only)	Down 1:1	46186	50%	57.56%	63.35%	56.94%
	Down 1:2	69279	33%	80.26%	31.26%	85.54%
	Down 1:4	115465	20%	88.42%	12.10%	96.66%
	Full Training	242372	9.5%	90.28%	5.75%	99.40%
	SMOTE	160958	43%	84.00%	21.67%	90.72%
Log. Regression (Officer Calls included)	Down 1:1	65704	50%	65.56%	63.18%	65.72%
	Down 1:2	98556	33%	85.85%	32.12%	89.45%
	Down 1:4	164260	20%	93.06%	3.77%	99.05%
	Full Training	523011	6.3%	93.72%	0%	100.00%
	SMOTE	229964	43%	84.66%	22.24%	88.84%
Log. Regression (Public Calls only)	Down 1:1	46816	50%	66.55%	54.06%	67.90%
	Down 1:2	69279	33%	84.28%	20.97%	91.10%
	Down 1:4	115465	20%	90.24%	0.15%	99.96%
	Full Training	242372	9.5%	90.24%	0.15%	99.96%
	SMOTE	160958	43%	82.25%	16.15%	89.38%
SVM (Officer Calls included)	Down 1:1	65704	50%	56.40%	73.02%	55.28%
	Down 1:2	98556	33%	86.31%	30.93%	90.03%
	Down 1:4	164260	20%	93.72%	0%	100%
	Full Training	523011	6.3%	93.72%	0%	100%
	SMOTE *	-	-	-	-	-

Table 4.7: Alternative Hypothesis Models Built using Extended Data.

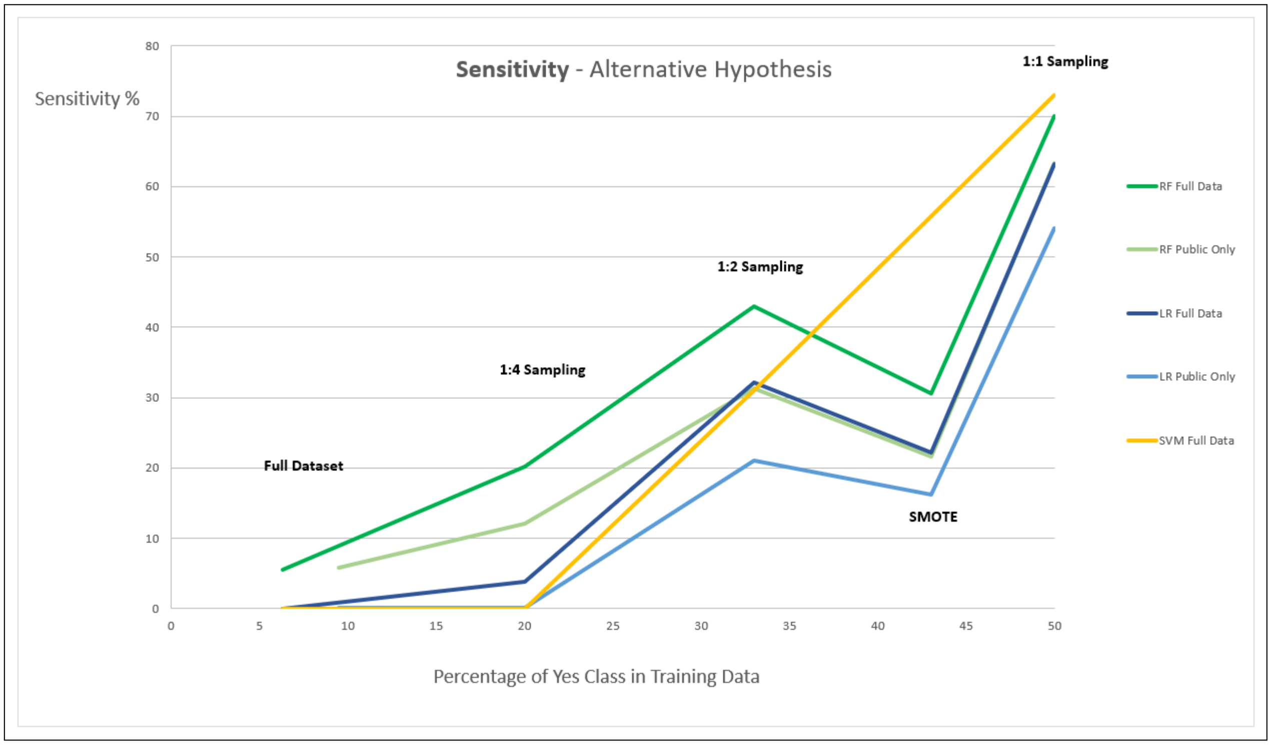


Figure 4.52: Sensitivity of Models for Extended dataset - Alternative Hypothesis

4.8.3 Graph of Accuracy

The results for accuracy are graphed in figure 4.53 for the 911 only data (null hypothesis), and in figure 4.54 for the extended dataset (alternative hypothesis).

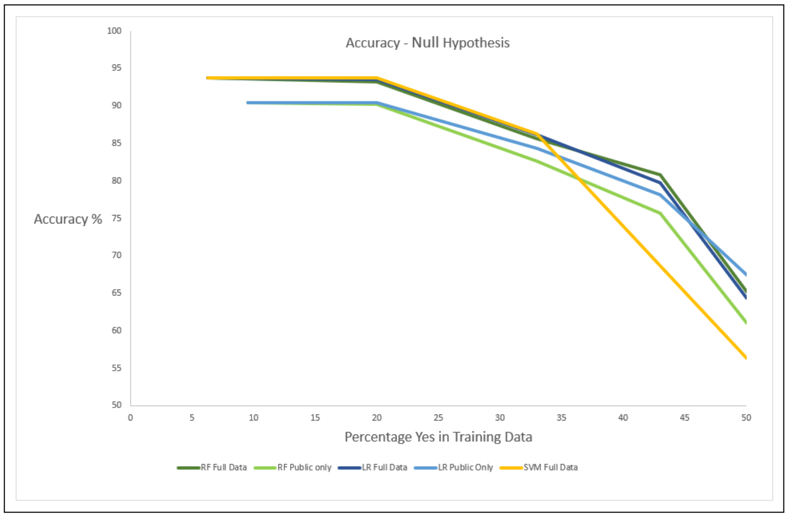


Figure 4.53: Accuracy of Models for the Primary Dataset - Null Hypothesis

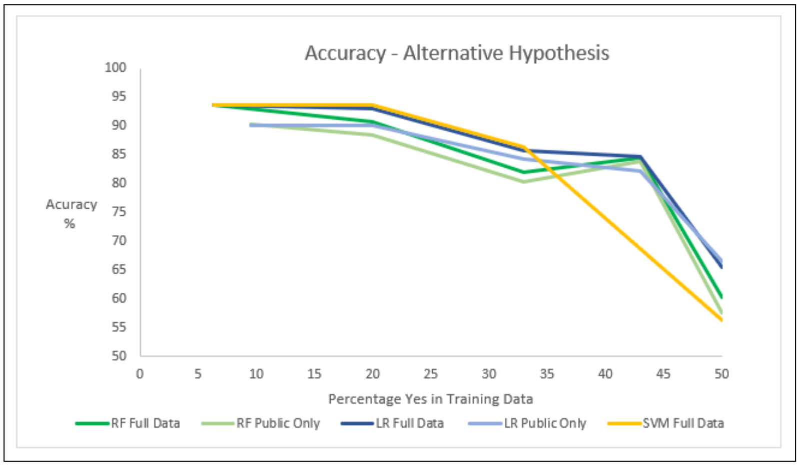


Figure 4.54: Accuracy of Models for Extended Dataset

4.8.4 Graph of Specificity

The specificity results are graphed in figure 4.55 for the 911 only data (null hypothesis), and figure 4.56 for the extended dataset (alternative hypothesis).

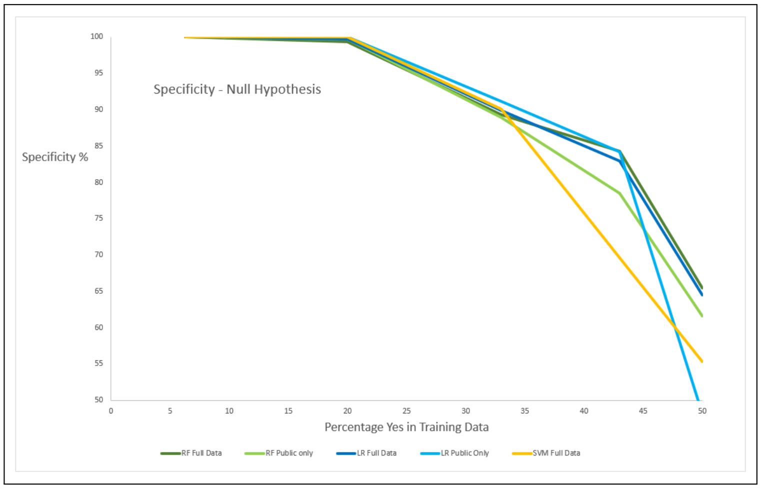


Figure 4.55: Specificity of Models for the Primary Dataset - Null Hypothesis

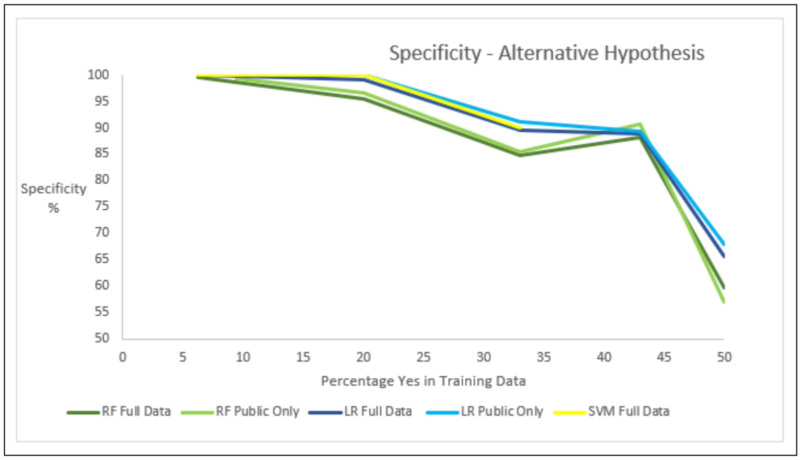


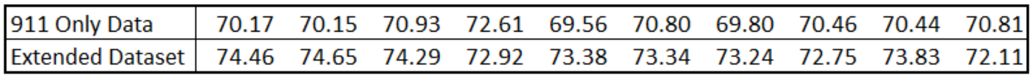
Figure 4.56: Specificity of Models for Extended Dataset

4.9 K-Fold testing using SVM

In total 24 models were constructed using the two datasets, one dataset intended for each hypothesis. For both datasets, the SVM model with 1:1 sampling achieved the highest sensitivity, 70.43% for the primary dataset and 73.02% for the extended dataset. This would support the alternative hypothesis. However, to rule out chance in the results due to sampling, k-fold testing with a sampling ratio of 1:1 was performed using SVMs with both datasets.

4.9.1 Sensitivity Results after K-Fold Stratified Validation

The results for sensitivity after k-fold testing are shown in figure 4.57 and plotted in figure 4.58.



911 Only Data	70.17	70.15	70.93	72.61	69.56	70.80	69.80	70.46	70.44	70.81
Extended Dataset	74.46	74.65	74.29	72.92	73.38	73.34	73.24	72.75	73.83	72.11

Figure 4.57: Sensitivity Results with K-Fold Validation

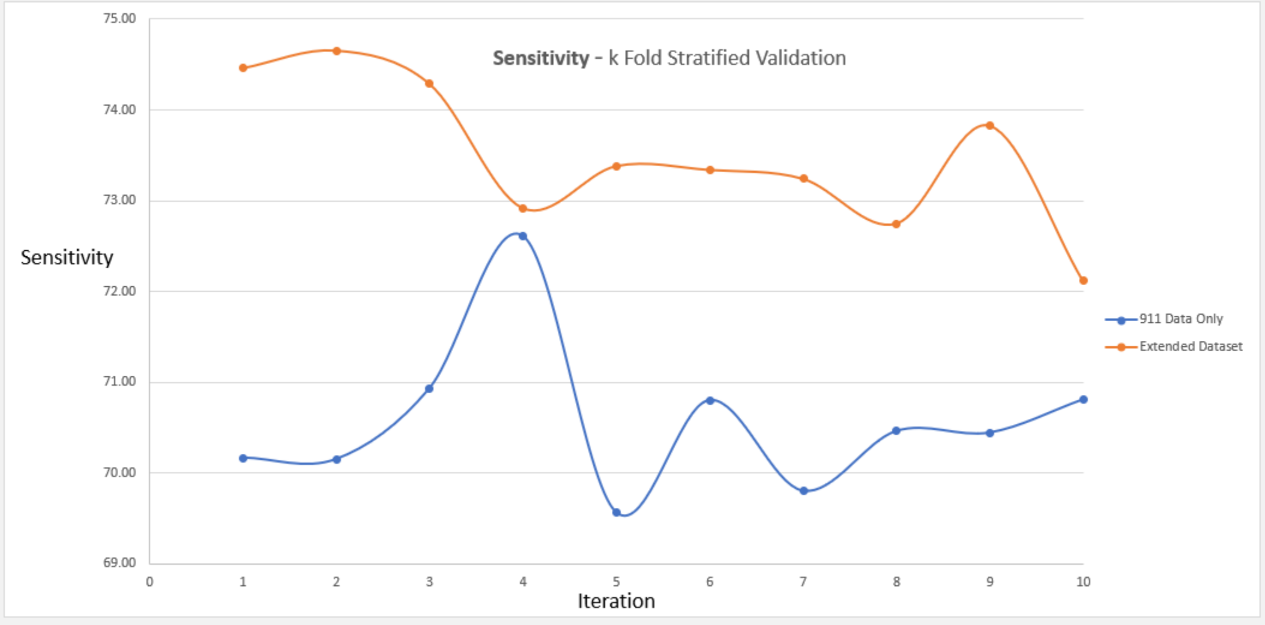


Figure 4.58: Sensitivity of Models with K-Fold Validation

When using k-fold cross validation, there is variation in the sensitivity of the modelsas expected. The primary dataset (for the null hypothesis) has a mean sensitivity of 70.57% and the extended dataset (for the alternative hypothesis) has a mean sensitivity of 73.49%.

Examining the results closely illustrates how sampling could impact results. Note that iteration 4 which applies to the null hypothesis achieved a sensitivity of 72.61% which is better than iteration 10 using the extended dataset which had a sensitivity of 72.11%. It is possible that these results could have been achieved by sampling chance leading to a possibly incorrect conclusion. For this reason, to rule out chance in the results an independent t-test is performed to determine whether there is statistical evidence that their means are significantly different.

4.9.2 Independent T-test

The independent t-test will determine if the results from both models are from the same population with regard to sensitivity by calculating the t statistic as shown below. The calculated value will be compared against the t-distribution table for a 2-tail test

using the value for degrees of freedom of 18 (2 * 10 samples - 2) and statistical level $\alpha = 0.05\%$.
If the calculated value of t is higher, then the null hypothesis that both means are from the same population can be rejected.

The t statistic can be calculated using the results in 4.57 as follows:-

$$calculated\, t \; = \; \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \tag{4.1}$$

where $\bar{x_1}$ is the mean of sample 1 and $\bar{x_2}$ is the mean of sample 2
S_1^2 is the variance of sample 1 and S_2^2 is the variance of sample 2
n_1 and n_2 are the number of samples in sample 1 and sample 2.

$$t \; = \; \frac{73.50 - 70.57}{\sqrt{\dfrac{0.656}{10} + \dfrac{0.707}{10}}} \; = \; \mathbf{7.92} \tag{4.2}$$

A t value of 2.101 which is what would be expected by chance, is found in the T-distribution table[13] for a 2-tail test with statistical level of 0.05 and df = 18. As the calculated t statistic is greater at 7.92, the null hypothesis that the means are from the same population, can be rejected. Therefore it was concluded that the mean sensitivities were from different populations.

4.10 Research Hypothesis Conclusion

The sensitivity of models to predict violent crime reports were tested using 911 call data only and a dataset of 911 call data combined with geospatial and temporal data related to the location or time of the 911 calls. Using k-fold stratified cross validation, models were created and the sensitivities using both datasets calculated. An independent t-test was then conducted on the sensitivity results after k-fold cross

[13] https://www.easycalculation.com/statistics/t-distribution-critical-value-table.php

validation indicating that the results were from different populations with regard to sensitivity.

As the mean sensitivity for the extended dataset at 73.50% is greater than the sensitivity for the primary dataset at 70.57%, **the null hypothesis can be rejected** and it can be concluded that a model using geospatial and temporal data related to the 911 call data will improve the sensitivity of the model.

4.11 Conclusion of Chapter

Two distinct datasets were created to test the two hypotheses. The primary dataset including only 911 call data was tagged as either *crime/no crime report* to indicate if a violent crime report could be associated with the 911 call. This dataset was extended to create a second dataset which included features from other data sources which the research had linked to violent crime. These features were associated with the 911 calls by either location or time. After dimension reduction, modelling was performed separately for both datasets with varied sampling ratios using random forests, logistic regression and support vector machines. For both data sets, the priority of the call was seen as the strongest categorical predictor. Dimension reduction was performed on the numerical features with education, income, race and home ownership accounting for the most variation in the data. This would suggest that the experiments could be repeated and sub divided by priority level. The support vector machines with a sampling ratio of 1:1 (crime/no crime) for both datasets showed the highest sensitivities. In order to discount sampling error, stratified k-fold cross validation was performed with each dataset and the results statistically analysed using an independent t-test which indicated that the mean sensitivity from the extended dataset was higher.

Therefore the alternative hypothesis that geospatial and temporal data relating to the 911 calls would improve prediction was accepted.

Chapter 5

Conclusion

This chapter summarises the work undertaken during the project, suggests possible future work and describes the results achieved.

5.1 Research Overview

The project attempted to solve a traditional problem, crime prediction, by approaching the problem from a different angle, using a different source of data as the basis for prediction. Currently, police response to 911 calls are reactive and determined solely by the assigned priority of the call and resources available at the time of the call. Anecdotally, it is assumed that a 911 call precedes or follows crime, but as no attempt is made to correlate 911 calls with crime reports or to look for other factors which may influence the outcome of a 911 call, this has not been formally tested. This research project examined the relationship between 911 calls and crime reports looking for features that could be linked to the 911 calls in order to improve the prediction model.

5.1.1 Summary of the Thesis

Chapter 1: This chapter introduced the research problem which was to determine if 911 call data combined with features relating to the neighbourhood and time of the 911 calls could be used to predict if violent crime would happen. The research problem was formalised into a directional research hypothesis which could be tested with a measurable metric.

Chapter 2: This chapter reviewed literature describing previous research into crime and it's causes. It discussed methods for crime prediction and how machine learning has been used as a tool in crime prediction. It identified several features which could be associated with 911 call data to be used in the prediction model.

Chapter 3: This chapter described several key concepts that would be used, in particular geospatial analysis. It described the algorithms that would be employed, the methodology, tools and the metrics used to formally answer the research question.

Chpater 4: This chapter described in detail the main body of work which was undertaken. It described the various datasets and how they were mined to extract features for the models. In particular, it described an approach which was used to extract geospatial data from Open Street Maps. It also described how the dependent variable was calculated by correlating 911 calls with crime reports. It described the modelling and tests which were carried out detailing the results which led to the alternative hypothesis being accepted.

5.2 Problem Definition

The research problem was to determine features which could be included with 911 call data in a model to predict if crimes would result from the 911 calls. The problem was formalised as follows:-

Can the Sensitivity of a Machine Learning Model to predict Violent Crime Reports using 911 Call Data be improved by adding geospatial and temporal data from heterogeneous data sources, related to the location and time of the 911 call?

The main challenges in the project were:-

1. Match 911 calls and crime reports using an efficient algorithm.
2. Determine and extract features from US Census Data.

3. Extract POI data from Open Street Maps and geospatially associate it to 911 calls.
4. Address imbalanced data with traditional and synthetic sampling techniques.

5.3 Design/Experimentation, Evaluation & Results

5.3.1 Design

In order to investigate the two hypotheses, two datasets were created, one for each hypothesis. The first dataset contained only data from the 911 call data and the dependent variable *Crime/No Crime*. The second dataset also included features which were temporally and geospatially linked to the 911 calls.

To determine the dependent variable, previous research suggested that temporal and geospatial proximity could be used to match 911 calls and crime reports. Using R code, an algorithm was written to first temporally match 911 calls and crime reports occurring within 15 minutes of each other and then to measure the distance between the resulting matches. Incidents were considered to be matched if they occurred within 15 minutes and within 750 meters of each other.

For the extended dataset, US Census Data was used as a source for several demographic and economic features related to the neighbourhoods. Open Street Maps contains POI data such as bars and nightclubs so a geospatial algorithm was developed using a grid to extract and calculate a bar index to indicate the bar concentration for each cell in the grid. The 911 calls were then geospatially assigned to the grid to determine their bar index. Weather and holiday features were more easily obtained.

5.3.2 Evaluation and Results

Using the two datasets, 24 different models, with various algorithms and sampling ratios, were created and their sensitivities compared. Further experimentation using k-fold cross validation was performed with the extended data set achieving a mean sensitivity of 73.49%, while the 911 only dataset had a mean sensitivity of 70.57%.

These results were statistically analysed leading to the acceptance of the alternative hypothesis, that data linked to the location and time of 911 calls would improve the prediction of crime reports.

5.4 Contributions and Impact

It is common with crime prediction, as seen in past research, to use extracted temporal and geospatial features from a variety of data sources to predict crime. What is different with this research, is that data relating to the location and time of 911 calls is used in the models. Features were extracted from US Census Data and Open Street Maps using geospatial analysis, historical crime hotspots were created by geospatially analysing past crime reports. Temporally, the weather at the time of the 911 calls was extracted from US Meteorological data, holiday matched to the time and weekend data imputed.

911 as a Predictor of Crime

Anecdotally, people assume that a 911 call is the same thing as a crime report. As the analysis has shown a crime report does not automatically follow a 911 call. In past research, 911 data has not been used as a predictor of crime, but this research demonstrated how 911 data can be combined with other data to predict crime reports with reasonable results.

Algorithm to Relate 911 Calls with Crime Reports

There is very little research into the relationship between 911 calls and crime reports, the exception being (Wu & Frias-Martinez, 2018) which proposes how they can be linked but without suggesting a practical solution. This research developed an algorithm which temporally and geospatially linked 650,000 911 calls with 60,000 crime reports as shown in figure 4.24.

Algorithm to Extract POIs from Open Street Maps

A more challenging problem was linking POIs with 911 calls. An algorithm to create an index which related POIs, in this case bars and nightclubs, to the location of 911 calls was created. Using a combination of QGIS and R studio, the algorithm, extracted POI data from Open Street Maps, combined that data with a geographical grid of Detroit, then quantified the number of near and neighbouring bars to the 911 call locations.

This algorithm could be used to extract any type of feature from OSM.

5.5 Future Work and Recommendations

The results demonstrated the potential in using 911 data combined with geospatial and temporal data to predict crime but with more time some of the analysis and methods could be adapted or modified.

5.5.1 Priority 1 Calls

The analysis during this project has shown that priority 1 calls are more likely, than lower priority calls, by a factor of three to result in a crime report, see table 4.4. The research indicates that there may be a variety of reasons to explain this. It would be worthwhile to repeat the modelling using datasets subdivided by the priority of the calls. In particular, it would be interesting to see how the models perform with only priority 1 calls.

5.5.2 Officer Initiated Calls

The 911 call dataset includes the variable *Officer.Initiated* which differentiates 911 calls from the public and 911 calls from police officers. The officer initiated calls include administrative events for those officers and an attempt was made to remove these from the dataset. However, as can be seen in table 4.5, a much lower percentage of officer calls 3.47% result in crime reports compared to calls from the public 9.54%. This could indicate that administrative calls are still included in the officer initiated subset of calls or that the pre-emptive work done by police officers stops potential crimes from happening. The sensitivity results using random forest and logistic regression are better with both officer and public initiated calls but due to performance issues when building SVM models, an SVM model with calls only from the public was not tested.

It would be worthwhile to focus only on models which contain the public only 911 calls, as the potential benefits would be greater when the model was applied to calls from the public.

5.5.3 Impact Coding of Categorical Variables

Both logistic regression and random forests have limitations when using categorical variables with a large number of levels. In the case of this project our locations and types of crime have many levels. A possible method that can be used to get around this limitation, impact coding is described in the blog [1] and a case study [2]. It would be worthwhile to investigate the benefits of **impact coding**.

5.5.4 Extra Features

The research identified several feature that could be included in the extended dataset but not all were included in this research. In particular, other types of POI information could be used. For example, the impact of sporting events such as football matches related to the time of 911 calls and locations of entertainment venues could be included. The zoning of neighbourhoods, residential, business or industrial could be included. Using the methods employed to extract bar data from OSM, these other features could be extracted.

[1] http://www.win-vector.com/blog/2012/08/a-bit-more-on-impact-coding/

[2] http://www.win-vector.com/blog/2012/07/modeling-trick-impact-coding-of-categorical-variables-with-many-levels/

5.5.5 Time Correlation of 911 Calls and Crime Reports

For the purpose of this research, it was decided to consider that 911 calls and crime reports were potential matches if they occurred within 15 minutes of each other. It would be worthwhile to analyse the various types of crime to determine a suitable time window per crime type. When matching 911 calls and crimes, this could be applied as a final test to determine if the two incidents are indeed related.

5.5.6 Algorithms

Support Vector Machines had the best values for sensitivity and handle categorical variables. As previously discussed, the SVM models were not fully explored due to performance issues with some of the SVM models taking several days to create. But when using a down sampling ratio of 1:1 they achieved the best results for sensitivity. It would be worthwhile repeating this on a more powerful machine or by investigating solutions to optimise SVMs.

5.6 Concluding Summary

This project shows that it is possible to apply machine learning using 911 calls to predict crime. Although the results were not excellent, with model tuning and experimentation using different features, these results would likely improve. In addition, this research has demonstrated a method to relate 911 calls with crime reports to determine the percentages of 911 calls which result in crimes. This method could be used in the allocation of resources. A method was developed to extract POI information from Open Street Maps and to apply this information to a geographical location. It would be worthwhile to experiment with other data sources available from OSM.

Due to political and public considerations it is unlikely that 911 calls at priority 1 would ever be handled differently using a prediction model, but the lower priority calls could be handled differently by the police based on a model similar to this one.

References

Aliprantis, D., & Hartley, D. (2015). Blowing it up and knocking it down: The local and city-wide effects of demolishing high concentration public housing on crime. *Journal of Urban Economics*, *88*, 67–81.

Alves, L. G., Ribeiro, H. V., & Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications*, *505*, 435–443.

Anderson, E. (2000). *Code of the street: Decency, violence, and the moral life of the inner city.* WW Norton & Company.

Asher, J. (n.d.). Numbers racket there's great crime data for nearly every city in the united states. why is nobody using it? *Slate*. Retrieved from `https://slate.com/news-and-politics/crime`

Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 international conference on computing networking and informatics (iccni)* (pp. 1–9).

Belesiotis, A., Papadakis, G., & Skoutas, D. (2018). Analyzing and predicting spatial crime distribution using crowdsourced and open data. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, *3*(4), 12.

Bialik, C. (n.d.). Detroit police response times no guide to effectiveness. *Wall Street Journal*. Retrieved from `https://www.wsj.com`

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014). Once upon a crime: towards crime prediction from demographics and mobile data. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 427–434).

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Burns, P. (2011). *The r inferno.* Lulu. com.

Butke, P., & Sheridan, S. C. (2010). An analysis of the relationship between weather and aggressive crime in cleveland, ohio. *Weather, Climate, and Society*, *2*(2), 127–139.

Caplan, J. M., & Kennedy, L. W. (2010). *Risk terrain modeling manual: Theoretical framework and technical steps of spatial risk assessment for crime analysis.* Rutgers Center on Public Security.

Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security journal*, *21*(1-2), 4–28.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chohlas-Wood, A., Merali, A., Reed, W., & Damoulas, T. (2015). Mining 911 calls in new york city: temporal patterns, detection, and forecasting. In *Workshops at the twenty-ninth aaai conference on artificial intelligence.*

Corcoran, J. J., Wilson, I. D., & Ware, J. A. (2003). Predicting the geo-temporal variations of crime and disorder. *International Journal of Forecasting*, *19*(4), 623–634.

Desmond, M., Papachristos, A. V., & Kirk, D. S. (2016). Police violence and citizen crime reporting in the black community. *American Sociological Review*, *81*(5), 857–876.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems* (pp. 155–161).

Freeman, R. B. (1991). *Crime and the employment of disadvantaged youths* (Tech. Rep.). National Bureau of Economic Research.

Gerber, M. S. (2014). Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, *61*, 115–125.

Ghosh, D., Chun, S., Shafiq, B., & Adam, N. R. (2016). Big data-based smart city platform: Real-time crime analysis. In *Proceedings of the 17th international digital government research conference on digital government research* (pp. 58–66).

Good, D. H., Pirog-Good, M. A., & Sickles, R. C. (1986). An analysis of youth crime and employment patterns. *Journal of Quantitative Criminology*, *2*(3), 219–236.

Hagan, J., McCarthy, B., Herda, D., & Chandrasekher, A. C. (2018). Dual-process theory of racial isolation, legal cynicism, and reported crime. *Proceedings of the National Academy of Sciences*, *115*(28), 7190–7199.

Hagy, A. P., & Staniec, J. F. O. (2002). Immigrant status, race, and institutional choice in higher education. *Economics of Education Review*, *21*(4), 381–392.

Harries, K. (2006). Property crimes and violence in united states: An analysis of the influence of population density. *International Journal of Criminal Justice Sciences*, *1*(2).

Horrocks, J., & Menclova, A. K. (2011). The effects of weather on crime. *New Zealand Economic Papers*, *45*(3), 231–254.

Hsieh, C.-C., & Pugh, M. D. (1993). Poverty, income inequality, and violent crime: a meta-analysis of recent aggregate data studies. *Criminal justice review*, *18*(2), 182–202.

Ingilevich, V., & Ivanov, S. (2018). Crime rate prediction in the urban environment using social factors. *Procedia Computer Science*, *136*, 472–478.

Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies.* MIT Press.

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression.* Springer.

Klinger, D. A., & Bridges, G. S. (1997). Measurement error in calls-for-service as an indicator of crime. *Criminology*, *35*(4), 705–726.

Kwak, H., Dierenfeldt, R., & McNeeley, S. (2019). The code of the street and cooperation with the police: Do codes of violence, procedural injustice, and police ineffectiveness discourage reporting violent victimization to the police? *Journal of Criminal Justice*, *60*, 25–34.

Larson, M., Xu, Y., Ouellet, L., & Klahm IV, C. F. (2019). Exploring the impact of 9398 demolitions on neighborhood-level crime in detroit, michigan. *Journal of Criminal Justice*, *60*, 57–63.

Levitt, S. D., & Lochner, L. (2001). The determinants of juvenile crime. In *Risky behavior among youths: An economic analysis* (pp. 327–374). University of Chicago Press.

Lipton, R., Yang, X., A. Braga, A., Goldstick, J., Newton, M., & Rura, M. (2013). The geography of violence, alcohol outlets, and drug arrests in boston. *American journal of public health*, *103*(4), 657–664.

Lochner, L., & Moretti, E. (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American economic review*, *94*(1), 155–189.

Mäkelä, P., Martikainen, P., & Nihtilä, E. (2005). Temporal variation in deaths related to alcohol intoxication and drinking. *International journal of epidemiology*, *34*(4), 765–771.

Marzan, C. S., Baculo, M. J. C., de Dios Bulos, R., & Ruiz Jr, C. (2017). Time series analysis and crime pattern forecasting of city crime data. In *Proceedings of the international conference on algorithms, computing and systems* (pp. 113–118).

Merlo, L. J., Hong, J., & Cottler, L. B. (2010). The association between alcohol-related arrests and college football game days. *Drug and alcohol dependence*, *106*(1), 69–71.

Næs, T., & Mevik, B.-H. (2001). Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *15*(4), 413–426.

Olligschlaeger, A. M. (1997). Artificial neural networks and crime mapping. *Crime mapping and crime prevention*, 313–348.

Platt, J., et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*(3), 61–74.

Ranson, M. (2014). Crime, weather, and climate change. *Journal of environmental economics and management*, *67*(3), 274–302.

Ratcliffe, J. H. (2004). The hotspot matrix: A framework for the spatio-temporal targeting of crime reduction. *Police practice and research*, *5*(1), 5–23.

Roncek, D. W., & Bell, R. (1981). Bars, blocks, and crimes. *Journal of Environmental Systems*, *11*(1), 35–47.

Sampson, R. J., Morenoff, J. D., & Raudenbush, S. (2005). Social anatomy of racial and ethnic disparities in violence. *American journal of public health*, *95*(2), 224–232.

Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, *277*(5328), 918–924.

Small, M. L. (2018). Understanding when people will report crimes to the police. *Proceedings of the National Academy of Sciences*, *115*(32), 8057–8059.

Sohony, I., Pratap, R., & Nambiar, U. (2018). Ensemble learning for credit card fraud detection. In *Proceedings of the acm india joint international conference on data science and management of data* (pp. 289–294). New York, NY, USA: ACM. Retrieved from `http://doi.acm.org/10.1145/3152494.3156815` doi: 10.1145/3152494.3156815

Spader, J., Schuetz, J., & Cortes, A. (2016). Fewer vacants, fewer crimes? impacts of neighborhood revitalization policies on crime. *Regional Science and Urban Economics*, *60*, 73–84.

Stalidis, P., Semertzidis, T., & Daras, P. (2018). Examining deep learning architectures for crime classification and prediction. *arXiv preprint arXiv:1812.00602*.

Stec, A., & Klabjan, D. (2018). Forecasting crime with deep learning. *arXiv preprint arXiv:1806.01486*.

Tripodi, S. J., Kim, J. S., & Bender, K. (2010). Is employment associated with reduced recidivism? the complex relationship between employment and crime. *International Journal of Offender Therapy and Comparative Criminology*, *54*(5), 706–720.

Turvey, B. E. (2013). *Forensic victimology: Examining violent crime victims in investigative and legal contexts.* Academic Press.

Turvey, B. E., Savino, J. O., & Baeza, J. J. (2017). *False allegations: Investigative and forensic issues in fraudulent reports of crime.* Elsevier.

Vapnik, V. (2013). *The nature of statistical learning theory.* Springer science & business media.

Venturini, L., & Baralis, E. (2016). A spectral analysis of crimes in san francisco. In *Proceedings of the 2nd acm sigspatial workshop on smart cities and urban analytics* (p. 4).

Vidal, J. B. I., & Kirchmaier, T. (2015). The effect of police response time on crime detection.

Wilkinson, M. (n.d.). Detroit police improve response times. but not all neighborhoods are equal. *The Guardian*. Retrieved 2019-05-24, from `https://www.bridgemi.com/detroit-journalism-cooperative`

Wortley, R., & Townsley, M. (2016). *Environmental criminology and crime analysis*. Taylor & Francis.

Wu, J., & Frias-Martinez, V. (2018). An analysis of the relationship between crime incidents and 911 calls. *Proceedings of the Association for Information Science and Technology*, *55*(1), 933–935.

Xu, Y., Fu, C., Kennedy, E., Jiang, S., & Owusu-Agyemang, S. (2018). The impact of street lights on spatial-temporal patterns of crime in detroit, michigan. *Cities*, *79*, 45–52.

Zhao, X., & Tang, J. (2017). Modeling temporal-spatial correlations for crime prediction. In *Proceedings of the 2017 acm on conference on information and knowledge management* (pp. 497–506).

Appendix A

911 Dataset Description

1. Incident.ID - Unique Reference for call in plain text
2. Agency - Police department for call (DPD - 1719765 rows, Winston Salem PD - 91 rows)
3. Incident.Address - Address anonymised to 100 block level
4. Zip.Code
5. Priority - priority assigned to call, range 1 to 5, 7, 9 and NA
6. Call.Code - Internal Police Code for call 110 categories
7. Call.Description - Description related to Call.Code
8. Category - further internal police categorisation
9. Call.Time - Date and Time of Call
10. Time.of.Call - Time of call
11. Precinct.Scout.Car.Area - DPD precinct area code
12. Responding.Unit - DPD car code
13. Officer.Initiated - Call Initiated by public or by officer
14. Intake.Time - time to process call, generally 0
15. Dispatch.Time - time before car dispatched
16. Travel.Time - time to travel to location
17. Total.Response.Time - total time for response
18. Time.On.Scene - time at scene
19. Total.Time - total time for call

20. Neighborhood - Detroit neighbourhood
21. Census.Block.GEOID - US Census Code for the block
22. Council.District - Council District code
23. Longitude - gps co-ordinates
24. Latitude - gps co-ordinates
25. Incident.Location - gps co-ordinates

Appendix B

Crime Dataset Description

1. Crime.ID - unique reference in DPD Record Management System
2. Report - unique id
3. Incident.Address - Address anonymised to 100 block level
4. Offense.Description - Description of offense
5. Offense.Category - Category
6. Charge.Description - Duplicate of Offense Description
7. Offense.Code - DPD 5 digit code for offence
8. State.Offense.Code - State 4 digit code for offense
9. Incident.Date...Time - Date and Time
10. Incident.Time..24h - Time
11. Day.of.Week - Day of the week represented by a number
12. Hour.of.Day -
13. Year
14. Scout.Car.Area - DPD code
15. Precinct.Number - DPD code
16. Census.Block.GEOID - US Census Code
17. Neighbourhood - name of neighbourhood
18. Council.District - code for district
19. Zip.Code - missing
20. Longitude - gps

21. Latitude - gps
22. IBR.Report.Date - date report sent to State Reporting system
23. Location - gps codes
24. uniq - id used as part of anonymisation process
25. Hardest.Hit.Fund.Areas - Code relating to Federal assistance program
26. City.Council.Districts - Code for district
27. Detroit.Neighborhoods - code for neighbourhood
28. Scout.Car.Areas - DPD code for area
29. Counties - county code
30. Zip.Codes - zip code

Appendix C

Initial Full Dataset Description

1. Incident.ID - Unique ID for 911 call
2. Priority - Priority assigned by operator to call
3. TimeDate - Time and Date of call
4. Officer.Initiated - Call initiated by police officer
5. Day - Day of the week for 911 call
6. Hour - Hour of the day for 911 call
7. Month - Month of 911 call
8. Grid_id - Grid ID for calculating bar index
9. DATE - Date of call
10. PRCP - amount of precipitation on day of call
11. TMAX - maximum temperature on day of call
12. TMIN - minimum temperature on day of call
13. isHoliday - Was call on a holiday
14. isHolidayEve - Was call on day before a holiday
15. NumBars - Number of bars in grid
16. BarIndex - Index to represent number of bars in surrounding grids
17. GEOID - Census Tract GEOID for 911 call
18. ALAND - Are of Census Tract
19. Population - Population of census tract
20. Males - Number of males in census tract

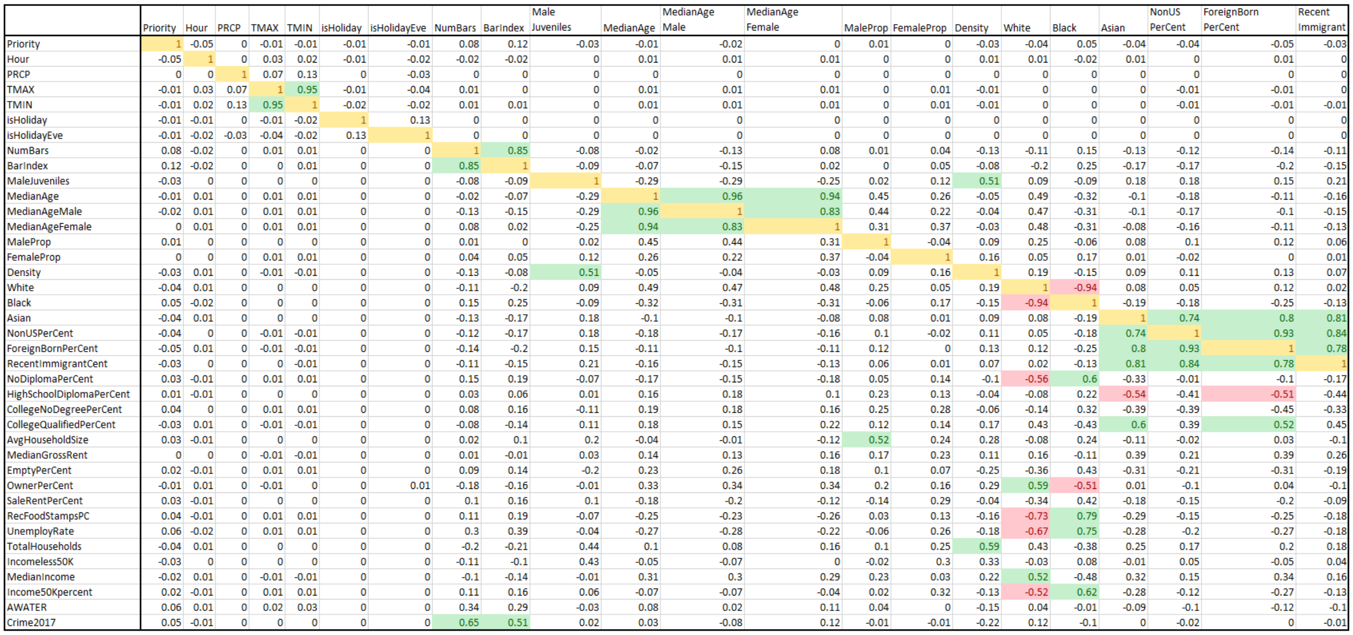
21. MaleJuveniles - Number of males aged 15 to 24
22. Females - Number of females in census tract
23. MedianAge - Median age in census tract
24. MedianAgeMale - Median age of males in census tract
25. MedianAgeFemale - Median age of females in census tract
26. MaleProp - Percentage of males
27. FemaleProp - Percentage of females
28. SexRatio - Ratio female to males
29. Density - Population Density
30. White - Percentage white
31. Black - Percentage black
32. Asian - Percentage Asian
33. Total - total population for census tract
34. NonUSCitizen - number not a US citizen
35. ForeignBorn - number foreign born
36. RecentImmigrant - number immigrant after 2010
37. NonUSPerCent - percentage of census tract not US citizen
38. ForeignBornPerCent - percentage of census tract foreign born
39. RecentImmigrantCent - percentage of census tract recent immigrant
40. Totalover25 - total number of people aged over 25
41. NoDiploma - number educated to high school or less without diploma
42. HighSchoolDiploma - number educated to high school with diploma
43. CollegeNoDegree - number educated to college without qualification
44. CollegeQualified - number educated to college with qualification
45. NoDiplomaPerCent - percentage of over 25's without high school diploma
46. HighSchoolDiplomaPerCent - percentage of over 25's to high school diploma
47. CollegeNoDegreePerCent - percentage of over 25's who attended college
48. CollegeQualifiedPerCent - percentage of over 25's college qualified
49. totalUnits - housing total number of units in census tract
50. occcupiedUnits - housing total number of occupied units in census tract

51. VacantUnits - housing total number of occupied units in census tract
52. OwnerOcc - housing total number of owner occupied units in census tract
53. RentedUnits - housing total number of rented units in census tract
54. ForRentUnits - housing total number of units for rent in census tract
55. ForSale - housing total number of units for sale in census tract
56. AvgHouseholdSize - average number per household
57. MedianGrossRent - median gross rent for census tract
58. EmptyPerCent - percentage of empty units
59. OwnerPerCent - percentage of owner occupied units
60. SaleRentPerCent - percentage for sale units
61. receivedPoverty - number of households that received food stamps due to poverty
62. totalHouseholds - total number of households in census tract
63. NoStamps - number of households that did not receive food stamps due to poverty
64. RecFoodStampsPC - percentage of households receiving stamps due to poverty
65. total - Employment. Total available workforce.
66. MaleTotal - Total available male workforce.
67. FemaleTotal - Total available female workforce.
68. MaleUnemployed - Number of males unemployed
69. FemaleUnemployed - Number of females unemployed
70. UnemployRate - Unemployment rate %
71. TotalHouseholds - Total number of households
72. Incomeless50K - Number of households with income less than $50K
73. MedianIncome - Median income for census tract
74. Income50Kpercent - Percentage of tract with income less than $50K
75. AWATER - Amount of water area in census tract in
76. Crime2017 - Number of crimes in census tract during 2017
77. Crime - Violent Crime associated with 911 call

Appendix D

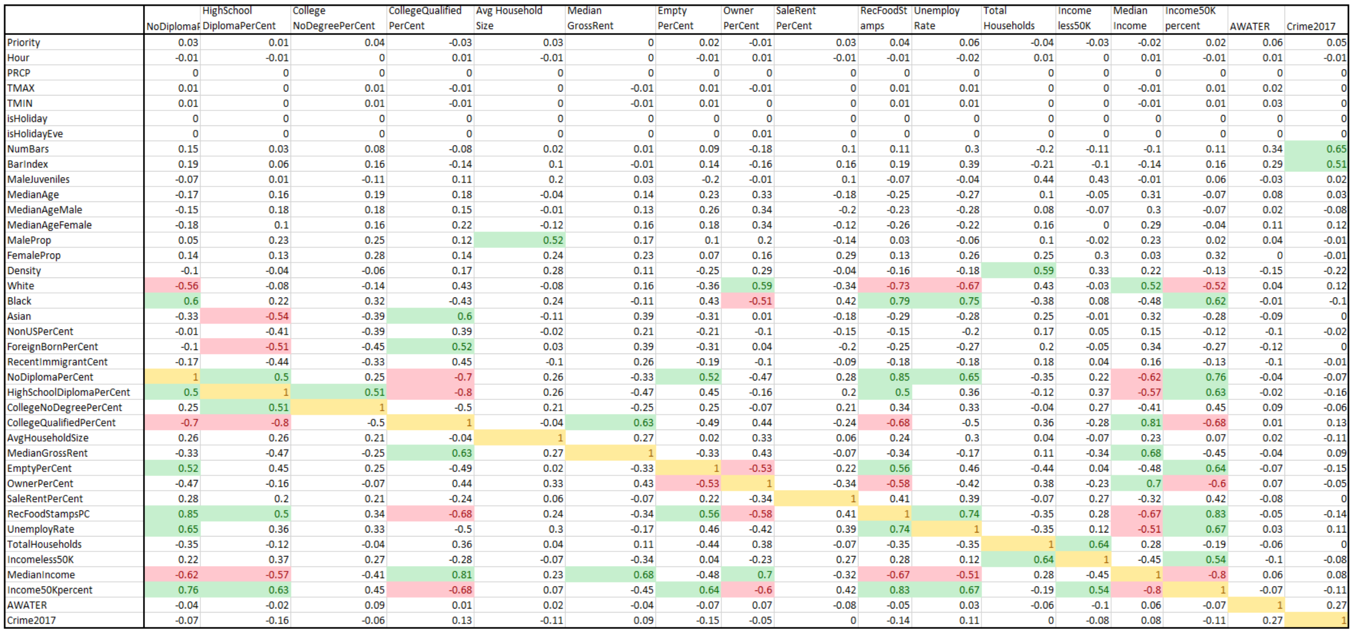
Correlation of Dataset

Note highlighted in green indicates positive correlation above 0.5, while highlighted in red indicates negative correlation below -0.5.



	Priority	Hour	PRCP	TMAX	TMIN	isHoliday	isHolidayEve	NumBars	BarIndex	Male Juveniles	MedianAge	MedianAge Male	MedianAge Female	MaleProp	FemaleProp	Density	White	Black	Asian	NonUS PerCent	ForeignBorn PerCent	Recent Immigrant
Priority	1	-0.05	0	-0.01	-0.01	-0.01	-0.01	0.08	0.12	-0.03	-0.01	-0.02	0	0.01	0	-0.03	-0.04	0.05	-0.04	-0.04	-0.05	-0.03
Hour	-0.05	1	0	0.03	0.02	-0.01	-0.02	-0.02	-0.02	0	0.01	0.01	0.01	0	0	0.01	0.01	-0.02	0.01	0	0.01	0
PRCP	0	0	1	0.07	0.13	0	-0.03	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TMAX	-0.01	0.03	0.07	1	0.95	-0.01	-0.04	0.01	0	0	0.01	0.01	0.01	0	0.01	-0.01	0	0	0	-0.01	-0.01	0
TMIN	-0.01	0.02	0.13	0.95	1	-0.02	-0.02	0.01	0.01	0	0.01	0.01	0.01	0	0.01	-0.01	0	0	0	-0.01	-0.01	-0.01
isHoliday	-0.01	-0.01	0	-0.01	-0.02	1	0.13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
isHolidayEve	-0.01	-0.02	-0.03	-0.04	-0.02	0.13	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NumBars	0.08	-0.02	0	0.01	0.01	0	0	1	0.85	-0.08	-0.02	-0.13	0.08	0.01	0.04	-0.13	-0.11	0.15	-0.13	-0.12	-0.14	-0.11
BarIndex	0.12	-0.02	0	0	0.01	0	0	0.85	1	-0.09	-0.07	-0.15	0.02	0	0.05	-0.08	-0.2	0.25	-0.17	-0.17	-0.2	-0.15
MaleJuveniles	-0.03	0	0	0	0	0	0	-0.08	-0.09	1	-0.29	-0.29	-0.25	0.02	0.12	0.51	0.09	-0.09	0.18	0.18	0.15	0.21
MedianAge	-0.01	0.01	0	0.01	0.01	0	0	-0.02	-0.07	-0.29	1	0.96	0.94	0.45	0.26	-0.05	0.49	-0.32	-0.1	-0.18	-0.11	-0.16
MedianAgeMale	-0.02	0.01	0	0.01	0.01	0	0	-0.13	-0.15	-0.29	0.96	1	0.83	0.44	0.22	-0.04	0.47	-0.31	-0.1	-0.17	-0.1	-0.15
MedianAgeFemale	0	0.01	0	0.01	0.01	0	0	0.08	0.02	-0.25	0.94	0.83	1	0.31	0.37	-0.03	0.48	-0.31	-0.08	-0.16	-0.11	-0.13
MaleProp	0.01	0	0	0	0	0	0	0.01	0	0.02	0.45	0.44	0.31	1	-0.04	0.09	0.25	-0.06	0.08	0.1	0.12	0.06
FemaleProp	0	0	0	0.01	0.01	0	0	0.04	0.05	0.12	0.26	0.22	0.37	-0.04	1	0.16	0.05	0.17	0.01	-0.02	0	0.01
Density	-0.03	0.01	0	-0.01	-0.01	0	0	-0.13	-0.08	0.51	-0.05	-0.04	-0.03	0.09	0.16	1	0.19	-0.15	0.09	0.11	0.13	0.07
White	-0.04	0.01	0	0	0	0	0	-0.11	-0.2	0.09	0.49	0.47	0.48	0.25	0.05	0.19	1	-0.94	0.08	0.05	0.12	0.02
Black	0.05	-0.02	0	0	0	0	0	0.15	0.25	-0.09	-0.32	-0.31	-0.31	-0.06	0.17	-0.15	-0.94	1	-0.19	-0.18	-0.25	-0.13
Asian	-0.04	0.01	0	0	0	0	0	-0.13	-0.17	0.18	-0.1	-0.1	-0.08	0.08	0.01	0.09	0.08	-0.19	1	0.74	0.8	0.81
NonUSPerCent	-0.04	0	0	-0.01	-0.01	0	0	-0.12	-0.17	0.18	-0.18	-0.17	-0.16	0.1	-0.02	0.11	0.05	-0.18	0.74	1	0.93	0.84
ForeignBornPerCent	-0.05	0.01	0	-0.01	-0.01	0	0	-0.14	-0.2	0.15	-0.11	-0.1	-0.11	0.12	0	0.13	0.12	-0.25	0.8	0.93	1	0.78
RecentImmigrantCent	-0.03	0	0	0	-0.01	0	0	-0.11	-0.15	0.21	-0.16	-0.15	-0.13	0.06	0.01	0.07	0.02	-0.13	0.81	0.84	0.78	1
NoDiplomaPerCent	0.03	-0.01	0	0.01	0.01	0	0	0.15	0.19	-0.07	-0.17	-0.15	-0.18	0.05	0.14	-0.1	-0.56	0.6	-0.33	-0.01	-0.1	-0.17
HighSchoolDiplomaPerCent	0.01	-0.01	0	0	0	0	0	0.03	0.06	0.01	0.16	0.18	0.1	0.23	0.13	-0.04	-0.08	0.22	-0.54	-0.41	-0.51	-0.44
CollegeNoDegreePerCent	0.04	0	0	0.01	0.01	0	0	0.08	0.16	-0.11	0.19	0.18	0.16	0.25	0.28	-0.06	-0.14	0.32	-0.39	-0.39	-0.45	-0.33
CollegeQualifiedPerCent	-0.03	0.01	0	-0.01	-0.01	0	0	-0.08	-0.14	0.11	0.18	0.15	0.22	0.12	0.14	0.17	0.43	-0.43	0.6	0.39	0.52	0.45
AvgHouseholdSize	0.03	-0.01	0	0	0	0	0	0.02	0.1	0.2	-0.04	-0.01	-0.12	0.52	0.24	0.28	-0.08	0.24	-0.11	-0.02	0.03	-0.1
MedianGrossRent	0	0	0	-0.01	-0.01	0	0	0.01	-0.01	0.03	0.14	0.13	0.16	0.17	0.23	0.11	0.16	-0.11	0.39	0.21	0.39	0.26
EmptyPerCent	0.02	-0.01	0	0.01	0.01	0	0	0.09	0.14	-0.2	0.23	0.26	0.18	0.1	0.07	-0.25	-0.36	0.43	-0.31	-0.21	-0.31	-0.19
OwnerPerCent	-0.01	0.01	0	-0.01	0	0	0.01	-0.18	-0.16	-0.01	0.33	0.34	0.34	0.2	0.16	0.29	0.59	-0.51	0.01	-0.1	0.04	-0.1
SaleRentPerCent	0.03	-0.01	0	0	0	0	0	0.1	0.16	0.1	-0.18	-0.2	-0.12	-0.14	0.29	-0.04	-0.34	0.42	-0.18	-0.15	-0.2	-0.09
RecFoodStampsPC	0.04	-0.01	0	0.01	0.01	0	0	0.11	0.19	-0.07	-0.25	-0.23	-0.26	0.03	0.13	-0.16	-0.73	0.79	-0.29	-0.15	-0.25	-0.18
UnemployRate	0.06	-0.02	0	0.01	0.01	0	0	0.3	0.39	-0.04	-0.27	-0.28	-0.22	-0.06	0.26	-0.18	-0.67	0.75	-0.28	-0.2	-0.27	-0.18
TotalHouseholds	-0.04	0.01	0	0	0	0	0	-0.2	-0.21	0.44	0.1	0.08	0.16	0.1	0.25	0.59	0.43	-0.38	0.25	0.17	0.2	0.18
Incomeless50K	-0.03	0	0	0	0	0	0	-0.11	-0.1	0.43	-0.05	-0.07	0	-0.02	0.3	0.33	-0.03	0.08	-0.01	0.05	-0.05	0.04
MedianIncome	-0.02	0.01	0	-0.01	-0.01	0	0	-0.1	-0.14	-0.01	0.31	0.3	0.29	0.23	0.03	0.22	0.52	-0.48	0.32	0.15	0.34	0.16
Income50Kpercent	0.02	-0.01	0	0.01	0.01	0	0	0.11	0.16	0.06	-0.07	-0.07	-0.04	0.02	0.32	-0.13	-0.52	0.62	-0.28	-0.12	-0.27	-0.13
AWATER	0.06	0.01	0	0.02	0.03	0	0	0.34	0.29	-0.03	0.08	0.02	0.11	0.04	0	-0.15	0.04	-0.01	-0.09	-0.1	-0.12	-0.1
Crime2017	0.05	-0.01	0	0	0	0	0	0.65	0.51	0.02	0.03	-0.08	0.12	-0.01	-0.01	-0.22	0.12	-0.1	0	-0.02	0	-0.01

Figure D.1: Correlation of Dataset - Part 1

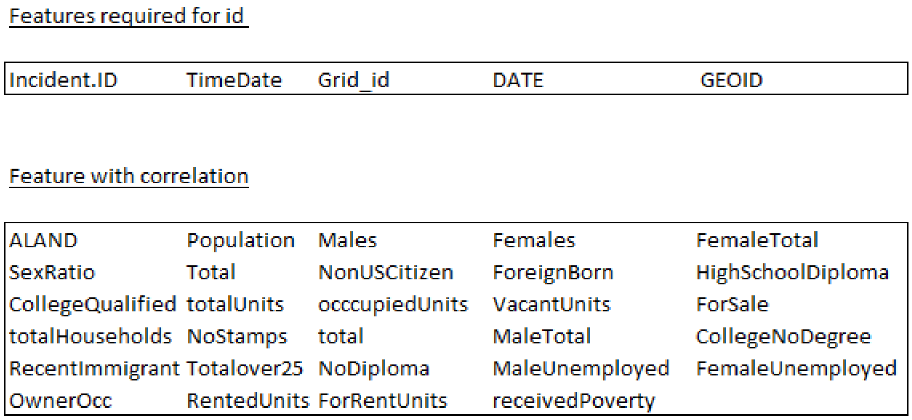


		HighSchool	College	CollegeQualified	Avg Household	Median	Empty	Owner	SaleRent	RecFoodSt	Unemploy	Total	Income	Median	Income50K		
	NoDiplomaF	DiplomaPerCent	NoDegreePerCent	PerCent	Size	GrossRent	PerCent	PerCent	PerCent	amps	Rate	Households	less50K	Income	percent	AWATER	Crime2017
Priority	0.03	0.01	0.04	-0.03	0.03	0	0.02	-0.01	0.03	0.04	0.06	-0.04	-0.03	-0.02	0.02	0.06	0.05
Hour	-0.01	-0.01	0	0.01	-0.01	0	-0.01	0.01	-0.01	-0.01	-0.02	0.01	0	0.01	-0.01	0.01	-0.01
PRCP	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TMAX	0.01	0	0.01	-0.01	0	-0.01	0.01	-0.01	0	0.01	0.01	0	0	-0.01	0.01	0.02	0
TMIN	0.01	0	0.01	-0.01	0	-0.01	0.01	0	0	0.01	0.01	0	0	-0.01	0.01	0.03	0
isHoliday	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
isHolidayEve	0	0	0	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0
NumBars	0.15	0.03	0.08	-0.08	0.02	0.01	0.09	-0.18	0.1	0.11	0.3	-0.2	-0.11	-0.1	0.11	0.34	0.65
BarIndex	0.19	0.06	0.16	-0.14	0.1	-0.01	0.14	-0.16	0.16	0.19	0.39	-0.21	-0.1	-0.14	0.16	0.29	0.51
MaleJuveniles	-0.07	0.01	-0.11	0.11	0.2	0.03	-0.2	-0.01	0.1	-0.07	-0.04	0.44	0.43	-0.01	0.06	-0.03	0.02
MedianAge	-0.17	0.16	0.19	0.18	-0.04	0.14	0.23	0.33	-0.18	-0.25	-0.27	0.1	-0.05	0.31	-0.07	0.08	0.03
MedianAgeMale	-0.15	0.18	0.18	0.15	-0.01	0.13	0.26	0.34	-0.2	-0.23	-0.28	0.08	-0.07	0.3	-0.07	0.02	-0.08
MedianAgeFemale	-0.18	0.1	0.16	0.22	-0.12	0.16	0.18	0.34	-0.12	-0.26	-0.22	0.16	0	0.29	-0.04	0.11	0.12
MaleProp	0.05	0.23	0.25	0.12	0.52	0.17	0.1	0.2	-0.14	0.03	-0.06	0.1	-0.02	0.23	0.02	0.04	-0.01
FemaleProp	0.14	0.13	0.28	0.14	0.24	0.23	0.07	0.16	0.29	0.13	0.26	0.25	0.3	0.03	0.32	0	-0.01
Density	-0.1	-0.04	-0.06	0.17	0.28	0.11	-0.25	0.29	-0.04	-0.16	-0.18	0.59	0.33	0.22	-0.13	-0.15	-0.22
White	-0.56	-0.08	-0.14	0.43	-0.08	0.16	-0.36	0.59	-0.34	-0.73	-0.67	0.43	-0.03	0.52	-0.52	0.04	0.12
Black	0.6	0.22	0.32	-0.43	0.24	-0.11	0.43	-0.51	0.42	0.79	0.75	-0.38	0.08	-0.48	0.62	-0.01	-0.1
Asian	-0.33	-0.54	-0.39	0.6	-0.11	0.39	-0.31	0.01	-0.18	-0.29	-0.28	0.25	-0.01	0.32	-0.28	-0.09	0
NonUSPerCent	-0.01	-0.41	-0.39	0.39	-0.02	0.21	-0.21	-0.1	-0.15	-0.15	-0.2	0.17	0.05	0.15	-0.12	-0.1	-0.02
ForeignBornPerCent	-0.1	-0.51	-0.45	0.52	0.03	0.39	-0.31	0.04	-0.2	-0.25	-0.27	0.2	-0.05	0.34	-0.27	-0.12	0
RecentImmigrantCent	-0.17	-0.44	-0.33	0.45	-0.1	0.26	-0.19	-0.1	-0.09	-0.18	-0.18	0.18	0.04	0.16	-0.13	-0.1	-0.01
NoDiplomaPerCent	1	0.5	0.25	-0.7	0.26	-0.33	0.52	-0.47	0.28	0.85	0.65	-0.35	0.22	-0.62	0.76	-0.04	-0.07
HighSchoolDiplomaPerCent	0.5	1	0.51	-0.8	0.26	-0.47	0.45	-0.16	0.2	0.5	0.36	-0.12	0.37	-0.57	0.63	-0.02	-0.16
CollegeNoDegreePerCent	0.25	0.51	1	-0.5	0.21	-0.25	0.25	-0.07	0.21	0.34	0.33	-0.04	0.27	-0.41	0.45	0.09	-0.06
CollegeQualifiedPerCent	-0.7	-0.8	-0.5	1	-0.04	0.63	-0.49	0.44	-0.24	-0.68	-0.5	0.36	-0.28	0.81	-0.68	0.01	0.13
AvgHouseholdSize	0.26	0.26	0.21	-0.04	1	0.27	0.02	0.33	0.06	0.24	0.3	0.04	-0.07	0.23	0.07	0.02	-0.11
MedianGrossRent	-0.33	-0.47	-0.25	0.63	0.27	1	-0.33	0.43	-0.07	-0.34	-0.17	0.11	-0.34	0.68	-0.45	-0.04	0.09
EmptyPerCent	0.52	0.45	0.25	-0.49	0.02	-0.33	1	-0.53	0.22	0.56	0.46	-0.44	0.04	-0.48	0.64	-0.07	-0.15
OwnerPerCent	-0.47	-0.16	-0.07	0.44	0.33	0.43	-0.53	1	-0.34	-0.58	-0.42	0.38	-0.23	0.7	-0.6	0.07	-0.05
SaleRentPerCent	0.28	0.2	0.21	-0.24	0.06	-0.07	0.22	-0.34	1	0.41	0.39	-0.07	0.27	-0.32	0.42	-0.08	0
RecFoodStampsPC	0.85	0.5	0.34	-0.68	0.24	-0.34	0.56	-0.58	0.41	1	0.74	-0.35	0.28	-0.67	0.83	-0.05	-0.14
UnemployRate	0.65	0.36	0.33	-0.5	0.3	-0.17	0.46	-0.42	0.39	0.74	1	-0.35	0.12	-0.51	0.67	0.03	0.11
TotalHouseholds	-0.35	-0.12	-0.04	0.36	0.04	0.11	-0.44	0.38	-0.07	-0.35	-0.35	1	0.64	0.28	-0.19	-0.06	0
Incomeless50K	0.22	0.37	0.27	-0.28	-0.07	-0.34	0.04	-0.23	0.27	0.28	0.12	0.64	1	-0.45	0.54	-0.1	-0.08
MedianIncome	-0.62	-0.57	-0.41	0.81	0.23	0.68	-0.48	0.7	-0.32	-0.67	-0.51	0.28	-0.45	1	-0.8	0.06	0.08
Income50Kpercent	0.76	0.63	0.45	-0.68	0.07	-0.45	0.64	-0.6	0.42	0.83	0.67	-0.19	0.54	-0.8	1	-0.07	-0.11
AWATER	-0.04	-0.02	0.09	0.01	0.02	-0.04	-0.07	0.07	-0.08	-0.05	0.03	-0.06	-0.1	0.06	-0.07	1	0.27
Crime2017	-0.07	-0.16	-0.06	0.13	-0.11	0.09	-0.15	-0.05	0	-0.14	0.11	0	-0.08	0.08	-0.11	0.27	1

Figure D.2: Correlation of Dataset - Part 2

Appendix E

Features Removed



Features required for id

Incident.ID	TimeDate	Grid_id	DATE	GEOID

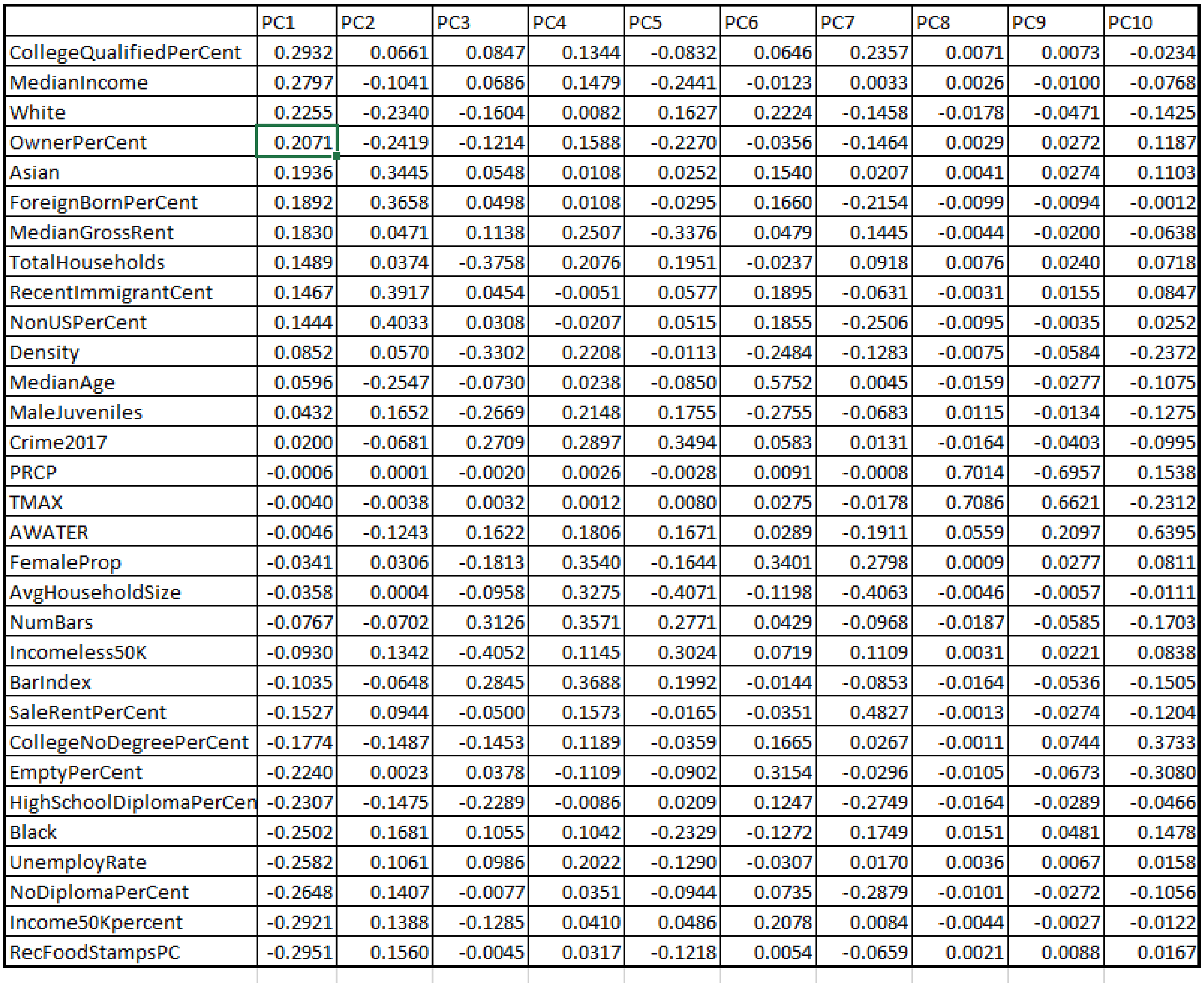
Feature with correlation

ALAND	Population	Males	Females	FemaleTotal
SexRatio	Total	NonUSCitizen	ForeignBorn	HighSchoolDiploma
CollegeQualified	totalUnits	occcupiedUnits	VacantUnits	ForSale
totalHouseholds	NoStamps	total	MaleTotal	CollegeNoDegree
RecentImmigrant	Totalover25	NoDiploma	MaleUnemployed	FemaleUnemployed
OwnerOcc	RentedUnits	ForRentUnits	receivedPoverty	

Figure E.1: First Feature Removal List

Appendix F

Principal Components 1 to 10

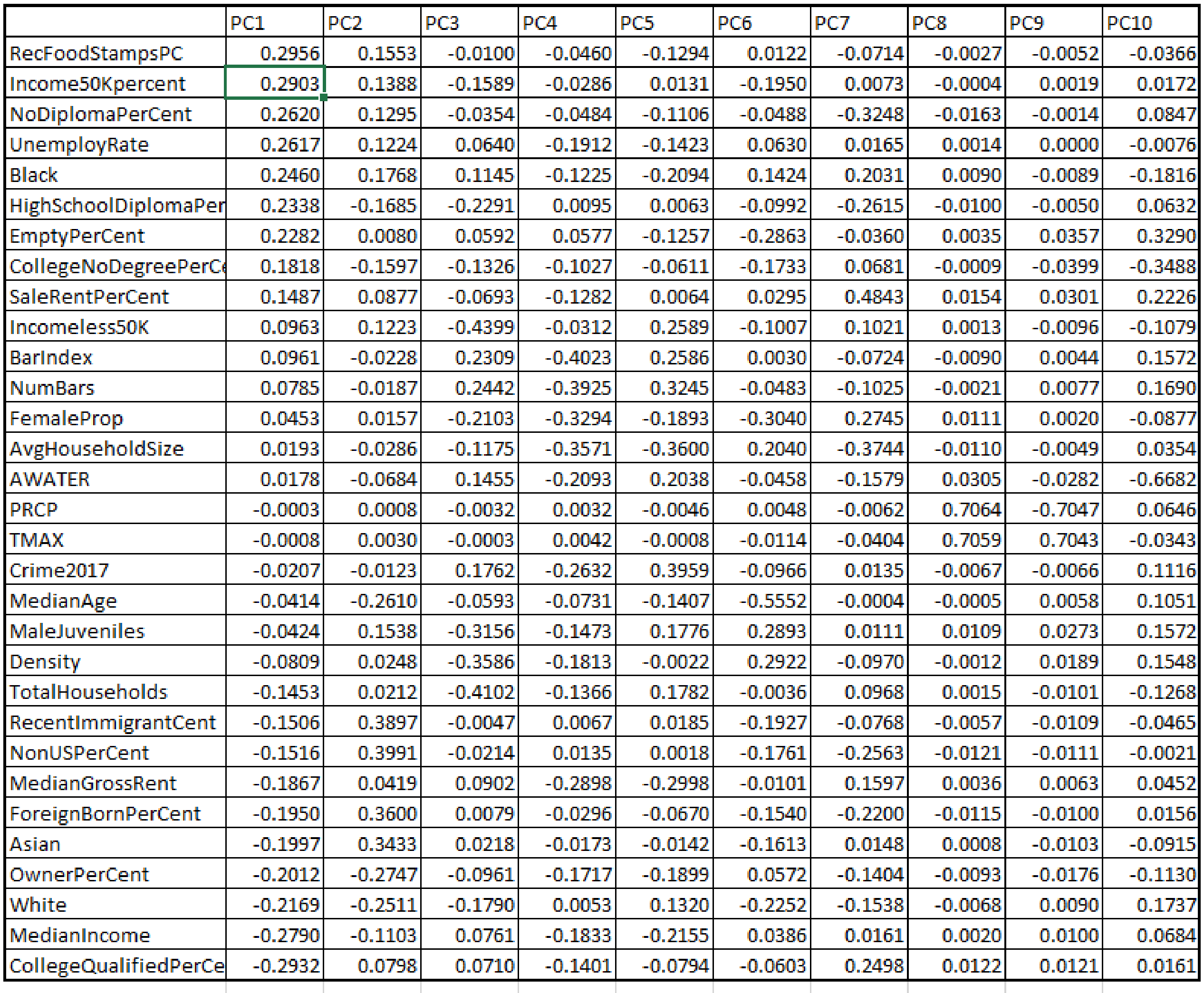


	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
CollegeQualifiedPerCent	0.2932	0.0661	0.0847	0.1344	-0.0832	0.0646	0.2357	0.0071	0.0073	-0.0234
MedianIncome	0.2797	-0.1041	0.0686	0.1479	-0.2441	-0.0123	0.0033	0.0026	-0.0100	-0.0768
White	0.2255	-0.2340	-0.1604	0.0082	0.1627	0.2224	-0.1458	-0.0178	-0.0471	-0.1425
OwnerPerCent	0.2071	-0.2419	-0.1214	0.1588	-0.2270	-0.0356	-0.1464	0.0029	0.0272	0.1187
Asian	0.1936	0.3445	0.0548	0.0108	0.0252	0.1540	0.0207	0.0041	0.0274	0.1103
ForeignBornPerCent	0.1892	0.3658	0.0498	0.0108	-0.0295	0.1660	-0.2154	-0.0099	-0.0094	-0.0012
MedianGrossRent	0.1830	0.0471	0.1138	0.2507	-0.3376	0.0479	0.1445	-0.0044	-0.0200	-0.0638
TotalHouseholds	0.1489	0.0374	-0.3758	0.2076	0.1951	-0.0237	0.0918	0.0076	0.0240	0.0718
RecentImmigrantCent	0.1467	0.3917	0.0454	-0.0051	0.0577	0.1895	-0.0631	-0.0031	0.0155	0.0847
NonUSPerCent	0.1444	0.4033	0.0308	-0.0207	0.0515	0.1855	-0.2506	-0.0095	-0.0035	0.0252
Density	0.0852	0.0570	-0.3302	0.2208	-0.0113	-0.2484	-0.1283	-0.0075	-0.0584	-0.2372
MedianAge	0.0596	-0.2547	-0.0730	0.0238	-0.0850	0.5752	0.0045	-0.0159	-0.0277	-0.1075
MaleJuveniles	0.0432	0.1652	-0.2669	0.2148	0.1755	-0.2755	-0.0683	0.0115	-0.0134	-0.1275
Crime2017	0.0200	-0.0681	0.2709	0.2897	0.3494	0.0583	0.0131	-0.0164	-0.0403	-0.0995
PRCP	-0.0006	0.0001	-0.0020	0.0026	-0.0028	0.0091	-0.0008	0.7014	-0.6957	0.1538
TMAX	-0.0040	-0.0038	0.0032	0.0012	0.0080	0.0275	-0.0178	0.7086	0.6621	-0.2312
AWATER	-0.0046	-0.1243	0.1622	0.1806	0.1671	0.0289	-0.1911	0.0559	0.2097	0.6395
FemaleProp	-0.0341	0.0306	-0.1813	0.3540	-0.1644	0.3401	0.2798	0.0009	0.0277	0.0811
AvgHouseholdSize	-0.0358	0.0004	-0.0958	0.3275	-0.4071	-0.1198	-0.4063	-0.0046	-0.0057	-0.0111
NumBars	-0.0767	-0.0702	0.3126	0.3571	0.2771	0.0429	-0.0968	-0.0187	-0.0585	-0.1703
Incomeless50K	-0.0930	0.1342	-0.4052	0.1145	0.3024	0.0719	0.1109	0.0031	0.0221	0.0838
BarIndex	-0.1035	-0.0648	0.2845	0.3688	0.1992	-0.0144	-0.0853	-0.0164	-0.0536	-0.1505
SaleRentPerCent	-0.1527	0.0944	-0.0500	0.1573	-0.0165	-0.0351	0.4827	-0.0013	-0.0274	-0.1204
CollegeNoDegreePerCent	-0.1774	-0.1487	-0.1453	0.1189	-0.0359	0.1665	0.0267	-0.0011	0.0744	0.3733
EmptyPerCent	-0.2240	0.0023	0.0378	-0.1109	-0.0902	0.3154	-0.0296	-0.0105	-0.0673	-0.3080
HighSchoolDiplomaPerCen	-0.2307	-0.1475	-0.2289	-0.0086	0.0209	0.1247	-0.2749	-0.0164	-0.0289	-0.0466
Black	-0.2502	0.1681	0.1055	0.1042	-0.2329	-0.1272	0.1749	0.0151	0.0481	0.1478
UnemployRate	-0.2582	0.1061	0.0986	0.2022	-0.1290	-0.0307	0.0170	0.0036	0.0067	0.0158
NoDiplomaPerCent	-0.2648	0.1407	-0.0077	0.0351	-0.0944	0.0735	-0.2879	-0.0101	-0.0272	-0.1056
Income50Kpercent	-0.2921	0.1388	-0.1285	0.0410	0.0486	0.2078	0.0084	-0.0044	-0.0027	-0.0122
RecFoodStampsPC	-0.2951	0.1560	-0.0045	0.0317	-0.1218	0.0054	-0.0659	0.0021	0.0088	0.0167

Figure F.1: First 10 Principal Components

Appendix G

Public Calls Principal Components 1 to 10



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
RecFoodStampsPC	0.2956	0.1553	-0.0100	-0.0460	-0.1294	0.0122	-0.0714	-0.0027	-0.0052	-0.0366
Income50Kpercent	0.2903	0.1388	-0.1589	-0.0286	0.0131	-0.1950	0.0073	-0.0004	0.0019	0.0172
NoDiplomaPerCent	0.2620	0.1295	-0.0354	-0.0484	-0.1106	-0.0488	-0.3248	-0.0163	-0.0014	0.0847
UnemployRate	0.2617	0.1224	0.0640	-0.1912	-0.1423	0.0630	0.0165	0.0014	0.0000	-0.0076
Black	0.2460	0.1768	0.1145	-0.1225	-0.2094	0.1424	0.2031	0.0090	-0.0089	-0.1816
HighSchoolDiplomaPer	0.2338	-0.1685	-0.2291	0.0095	0.0063	-0.0992	-0.2615	-0.0100	-0.0050	0.0632
EmptyPerCent	0.2282	0.0080	0.0592	0.0577	-0.1257	-0.2863	-0.0360	0.0035	0.0357	0.3290
CollegeNoDegreePerC	0.1818	-0.1597	-0.1326	-0.1027	-0.0611	-0.1733	0.0681	-0.0009	-0.0399	-0.3488
SaleRentPerCent	0.1487	0.0877	-0.0693	-0.1282	0.0064	0.0295	0.4843	0.0154	0.0301	0.2226
Incomeless50K	0.0963	0.1223	-0.4399	-0.0312	0.2589	-0.1007	0.1021	0.0013	-0.0096	-0.1079
BarIndex	0.0961	-0.0228	0.2309	-0.4023	0.2586	0.0030	-0.0724	-0.0090	0.0044	0.1572
NumBars	0.0785	-0.0187	0.2442	-0.3925	0.3245	-0.0483	-0.1025	-0.0021	0.0077	0.1690
FemaleProp	0.0453	0.0157	-0.2103	-0.3294	-0.1893	-0.3040	0.2745	0.0111	0.0020	-0.0877
AvgHouseholdSize	0.0193	-0.0286	-0.1175	-0.3571	-0.3600	0.2040	-0.3744	-0.0110	-0.0049	0.0354
AWATER	0.0178	-0.0684	0.1455	-0.2093	0.2038	-0.0458	-0.1579	0.0305	-0.0282	-0.6682
PRCP	-0.0003	0.0008	-0.0032	0.0032	-0.0046	0.0048	-0.0062	0.7064	-0.7047	0.0646
TMAX	-0.0008	0.0030	-0.0003	0.0042	-0.0008	-0.0114	-0.0404	0.7059	0.7043	-0.0343
Crime2017	-0.0207	-0.0123	0.1762	-0.2632	0.3959	-0.0966	0.0135	-0.0067	-0.0066	0.1116
MedianAge	-0.0414	-0.2610	-0.0593	-0.0731	-0.1407	-0.5552	-0.0004	-0.0005	0.0058	0.1051
MaleJuveniles	-0.0424	0.1538	-0.3156	-0.1473	0.1776	0.2893	0.0111	0.0109	0.0273	0.1572
Density	-0.0809	0.0248	-0.3586	-0.1813	-0.0022	0.2922	-0.0970	-0.0012	0.0189	0.1548
TotalHouseholds	-0.1453	0.0212	-0.4102	-0.1366	0.1782	-0.0036	0.0968	0.0015	-0.0101	-0.1268
RecentImmigrantCent	-0.1506	0.3897	-0.0047	0.0067	0.0185	-0.1927	-0.0768	-0.0057	-0.0109	-0.0465
NonUSPerCent	-0.1516	0.3991	-0.0214	0.0135	0.0018	-0.1761	-0.2563	-0.0121	-0.0111	-0.0021
MedianGrossRent	-0.1867	0.0419	0.0902	-0.2898	-0.2998	-0.0101	0.1597	0.0036	0.0063	0.0452
ForeignBornPerCent	-0.1950	0.3600	0.0079	-0.0296	-0.0670	-0.1540	-0.2200	-0.0115	-0.0100	0.0156
Asian	-0.1997	0.3433	0.0218	-0.0173	-0.0142	-0.1613	0.0148	0.0008	-0.0103	-0.0915
OwnerPerCent	-0.2012	-0.2747	-0.0961	-0.1717	-0.1899	0.0572	-0.1404	-0.0093	-0.0176	-0.1130
White	-0.2169	-0.2511	-0.1790	0.0053	0.1320	-0.2252	-0.1538	-0.0068	0.0090	0.1737
MedianIncome	-0.2790	-0.1103	0.0761	-0.1833	-0.2155	0.0386	0.0161	0.0020	0.0100	0.0684
CollegeQualifiedPerCe	-0.2932	0.0798	0.0710	-0.1401	-0.0794	-0.0603	0.2498	0.0122	0.0121	0.0161

Figure G.1: Public Initiated Calls, First 10 Principal Components