Georgia Tech School of City & Regional Planning
COLLEGE OF DESIGN

A Text-Mining and GIS Approach to Understanding Transit Customer Satisfaction

Soo Huey Yap

MS-GIST Capstone Project

July 24, 2020

CONTENTS

A Text Mining and GIS Approach to Understanding Transit Customer Satisfaction

1. INTRODUCTION

1.1 TRANSIT PERFORMANCE EVALUATION

Performance evaluation is a concept that most can understand. Examples of performance evaluation include evaluating the performance of students in schools via assignments and exams, and corporations and boards evaluating departmental and corporation-wide performance. In many of these instances, the objectives of performance evaluation are clear. In our first example, the aim of schools may be the education of students, and therefore performance evaluation is conducted to measure students' understanding and learning of the syllabi. In our second example, the aim of corporations may be to improve efficiency (reduce costs) and increase income. Performance measures used by private sector corporations may include number of sales, customer satisfaction ratings, and number of clicks on advertisements.

The Federal Transit Administration (FTA) recommends that transit agencies establish a performance evaluation system, and in fact, mandates annual reporting of certain performance measures. Recipients of FTA funding from the Urbanized Area Formula Program (5307) or Rural Formula Program (5311) are required to submit financial, asset, ridership and operation data to the National Transit Database (NTD) every year (FTA 2020). Other performance measurement that are mandatory for transit agencies include requirements for municipal budgeting and reporting, ADA compliance, and risk and liability assessment for insurance.

FTA helps transit agencies develop a performance evaluation system that is suited for each agency through the publication of a series of reports and guidelines, including Kittleson & Associates et. al. (2003), Ryus et. al. (2010), Boyle (2019), and Unger et. al. (2019). Kittleson & Associates et. al. (2003) described

more than 400 performance measures, which were later expanded to more than 600 measures by Unger et. al. (2019) by incorporating additional measures of social and economic impact. These FTA guidance documents outline steps to help agencies select from the 400+ performance measures proposed. Selection of performance measurement would also be based on factors such as availability and feasibility of data type and technology.

However, a recent survey (Boyle 2019) reveal that some performance measures were markedly more common than others; i.e. seven performance measures were used by at least 10 of a total of 23 respondents: Schedule adherence (20 agencies), Policy headways (19 agencies), Loading standard (18 agencies), Passengers per revenue hour (13 agencies), Service span (12 agencies), Bus shelter policy (10 agencies), and Bus stop spacing (10 agencies). This tendency for agencies to report the same measures may be related to resource constraints. When asked about challenges faced in the development and use of performance measures, more than half of all respondents rated 'Limited staff time' as the major challenge (Boyle 2019). Many transit agencies tend to report only exactly the performance measures required by regulation or grant/funding requirements (Kittleson & Associates 2003).

The influence of resource constraints and practicality on the selection and use of performance measures can be inferred from the work of Unger et. al. (2019) that established a panel of eight transit agencies to be interviewed to first validate the importance of performance measures from a list of measures identified as important measures from literature review, and then in a second round of evaluation, to refine the list based on evaluation of applicability to all agency types and sizes, realistic/attainability, reasonableness of tracking over time, and understandability by stakeholders. In their report, *Number of customer complaints responded to by type of complaint* and *Overall satisfaction with the transit system by user group* were respectively identified as the *first* and *third most important measures* to achieve the goal of community-building and engagement. However, when the list of important performance measures was refined in the second round, all seven community-building and

engagement performance measures had been removed. This suggests that while these measures are important, transit agencies perceive difficulties in measuring and reporting them.

1.2 USING TEXT-MINING AND SENTIMENT ANALYSIS TO MEASURE TRANSIT CUSTOMER SATISFACTION

In 2003, Kittleson & Associates et. al. reported that tracking and measuring customer satisfaction or loyalty had not "taken hold as quickly in the transit industry as in the private sector" because transit agencies are driven by different objectives. They suggested that transit agencies are limited in taking customer satisfaction into account because they are not profit-oriented and thus have limited financial resources to pay attention to customer satisfaction. There is also the perception that many people riding transit are transit-dependent riders, thus agencies focus on performing well for those riders rather than "choice" or non-transit-dependent riders.

Recognizing the importance but also difficulty faced by transit agencies in collecting information on customer satisfaction, various groups have begun exploring the use of text mining on social media as a mechanism to help transit agencies measure customer satisfaction. Collin et. al. (2013) collected and analyzed the sentiment of Twitter messages to quantify and compare the performance of rail lines operated by the Chicago Transit Authority. Their work with a limited number of data (557 Twitter messages) is a proof of concept that sentiment analysis of Twitter messages can be used to evaluate transit rider satisfaction; Albeit their results showed a predominance of negative sentiments and suggest that transit riders are more inclined to Tweet negative sentiments than positive sentiments. Luong and Houston (2015), using 8,515 Twitter messages about light rail transit services in Los Angeles, expanded on the work by incorporating word clustering analysis to improve understanding of positive and negative

sentiments (i.e. identify topics associated with different sentiments), and also analyzing retweet relationships to understand the virality of Twitter for the spread of transit information.

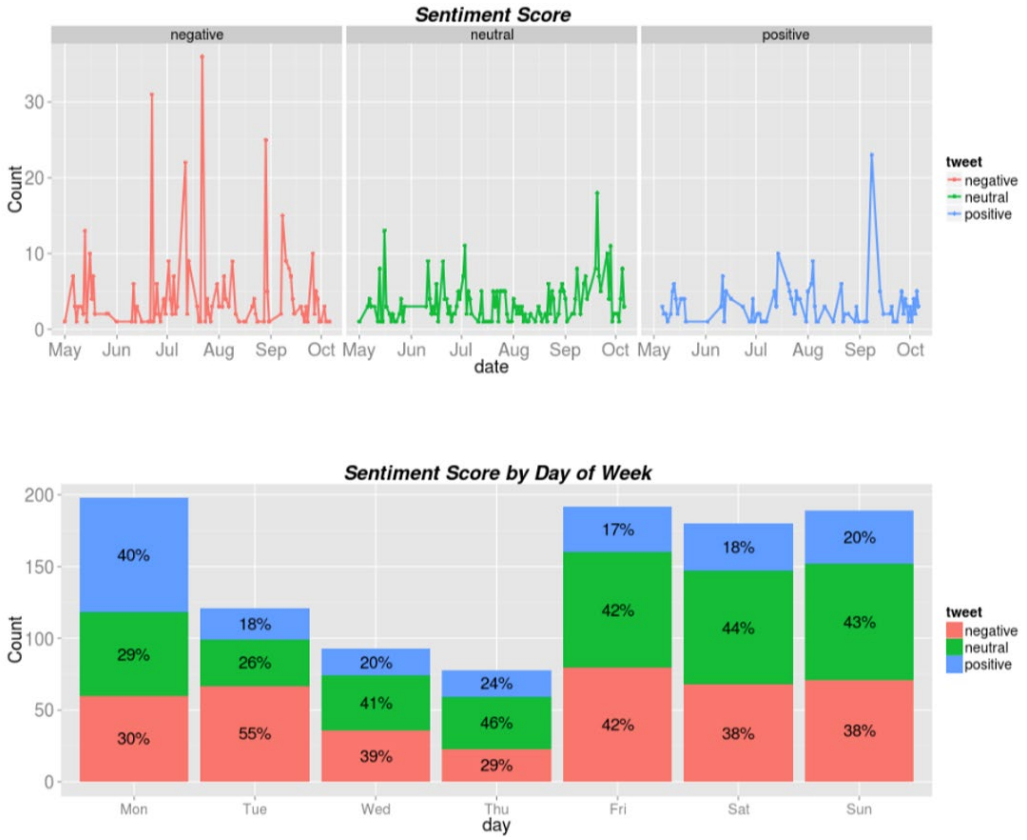


FIGURE 1. Interactive web-application developed by Luong and Houston (2015) to facilitate comparison of sentiment analysis over time and between rail lines

Luong and Houston (2015) also developed an interactive web-application for the reporting of their analysis to facilitate comparison of sentiment analysis over time and between rail lines (FIGURE 1). Mendez et. al. (2019) used more than 9,000 Twitter messages from 2014 to 2016 to study community satisfaction with bus transit services in Santiago, Chile. As with earlier studies, tweets about transit services in Santiago were predominantly negative. Nonetheless, text mining from Twitter data was effective in identifying and locating trends and issues over a larger geographic area than can be achieved by traditional surveys. However, Twitter data mining for transit rider satisfaction measurement should be

used with caution due to social justice concerns as Mendez et. al. (2019) noted that communes with a higher socioeconomic level tweeted more than those of lower incomes.

Due to the predominance of negative sentiments in Twitter messages, the work described above demonstrate the potential of Twitter data mining for the identification and understanding of *dis*satisfaction of the community with transit services. This may be a concern if the analyses were to be reported publicly; However, if the analyses were only reported internally within the transit agency, text mining of Twitter messages may be useful for transit agencies to identify and track issues and weaknesses within their system based on customer feedback and sentiment. In the longer term, the predominance of negative sentiment may also be overcome by increasing community engagement. It is noteworthy that an analysis of 64,000 Twitter messages about transit found that transit agencies that engage in more interactive dialog with Twitter users receive statistically significantly more positive statements, and fewer racist/sexist comments, than agencies that blast out announcements without interaction (Schweitzer 2014).

In this paper, we report results from a Master of Science in Geographic Information Science and Technology (MS-GIST) Capstone Project that combined Geographic Information System (GIS) analysis and text mining to measure transit customer satisfaction and to visualize locations where particular issues and concerns may be most prevalent in a transit system. As Twitter data may be confounded by negative sentiments that may not always be constructive, we use transit rider surveys collected by transit agency staff positioned at stations and stops throughout the transit system. This study allows identification of problems at specific transit stations, neighborhoods, cities, and other locations that may benefit from targeted interventions. If studied temporally, this study would also allow transit agencies to evaluate the effectiveness of interventions, E.g. To objectively determine whether driver training, vehicle maintenance, or service changes on particular routes lead to reduction in complaints about routes being late.

In order to ensure that methods and analyses developed in this study may be used by transit agencies independent of size and resource availability, only open-source and free software were used for all data preparation, analyses and mapping.

TABLE 1. Sociodemographic Information of the Three Counties Served by MARTA[†]

	FULTON	DEKALB	CLAYTON
Land area in square miles, 2010	526.64	267.58	141.57
Population Estimates	1,063,937	759,297	292,256
Population per square mile	2,020	2,838	2,064
Persons 65 years and over, percent	11.7%	12.4%	9.3%
White alone, percent	45.6%	35.8%	19.6%
Black or African American alone, percent	44.5%	54.9%	72.1%
American Indian and Alaskan Native alone, percent	0.3%	0.5%	0.6%
Asian alone, percent	7.5%	6.6%	5.2%
Native Hawaiian and Other Pacific Islander alone, percent	<0.5%	0.1%	0.1%
Two or More Races, percent	2.1%	2.2%	2.4%
Hispanic or Latino, percent	7.3%	8.6%	13.3%
White alone, not Hispanic or Latino, percent	39.7%	29.2%	9.6%
Foreign born persons, percent, 2014-2018	12.7%	16.4%	13.5%
Language other than English spoken at home, percent of persons age 5 years+, 2014-2018	15.9%	19.0%	19.7%
Mean travel time to work (minutes), workers age 16 years+, 2014-2018	28.5	32.2	31.5
Median Household Income (in 2018 dollars), 2014-2018	$64,787	$59,280	$45,778
Persons in poverty, percent	13.5%	14.3%	17.6%

[†] Values are based on 2010 to 2019 estimates reported by the US Census Bureau QuickFacts, unless other years are specified. All values reported in this table are from the US Census Bureau (2020).

2. METHODOLOGY

2.1 STUDY SITE AND TRANSIT AUTHORITY

This study was conducted with permission from the Research and Analysis Department of the Metropolitan Atlanta Rapid Transit Authority (MARTA). MARTA is the largest transit agency within the Metropolitan Atlanta region in Georgia. It provides bus and rail services over three counties – Fulton county, DeKalb county and Clayton county. Sociodemographic information for the three counties served by MARTA are outlined in TABLE 1 below. A map of MARTA's rail and bus routes is illustrated in FIGURE 2.

2.2 DESCRIPTION OF DATA

MARTA conducted a ridership trends survey in personal interview format at rail stations and onboard trains and buses in 2019. Responses from the ridership trend survey collected over 48 days (January 23rd to March 12th, 2019) were provided in comma-separated values (csv) format by the MARTA Research and Analysis Department for the purpose of this study. 18,487 survey responses were collected from throughout MARTA's transit system within this period. The primarily multiple choice format survey was comprised of 25 questions that included questions on transit use, travel mode preference, origin and destination, income, household, demography, home zip code, and MARTA rail station most convenient to home. The survey also included an open-ended "Other Comments" question. Responses for "Other Comments" was used in this study for text-mining and sentiment analysis to examine transit customer satisfaction.

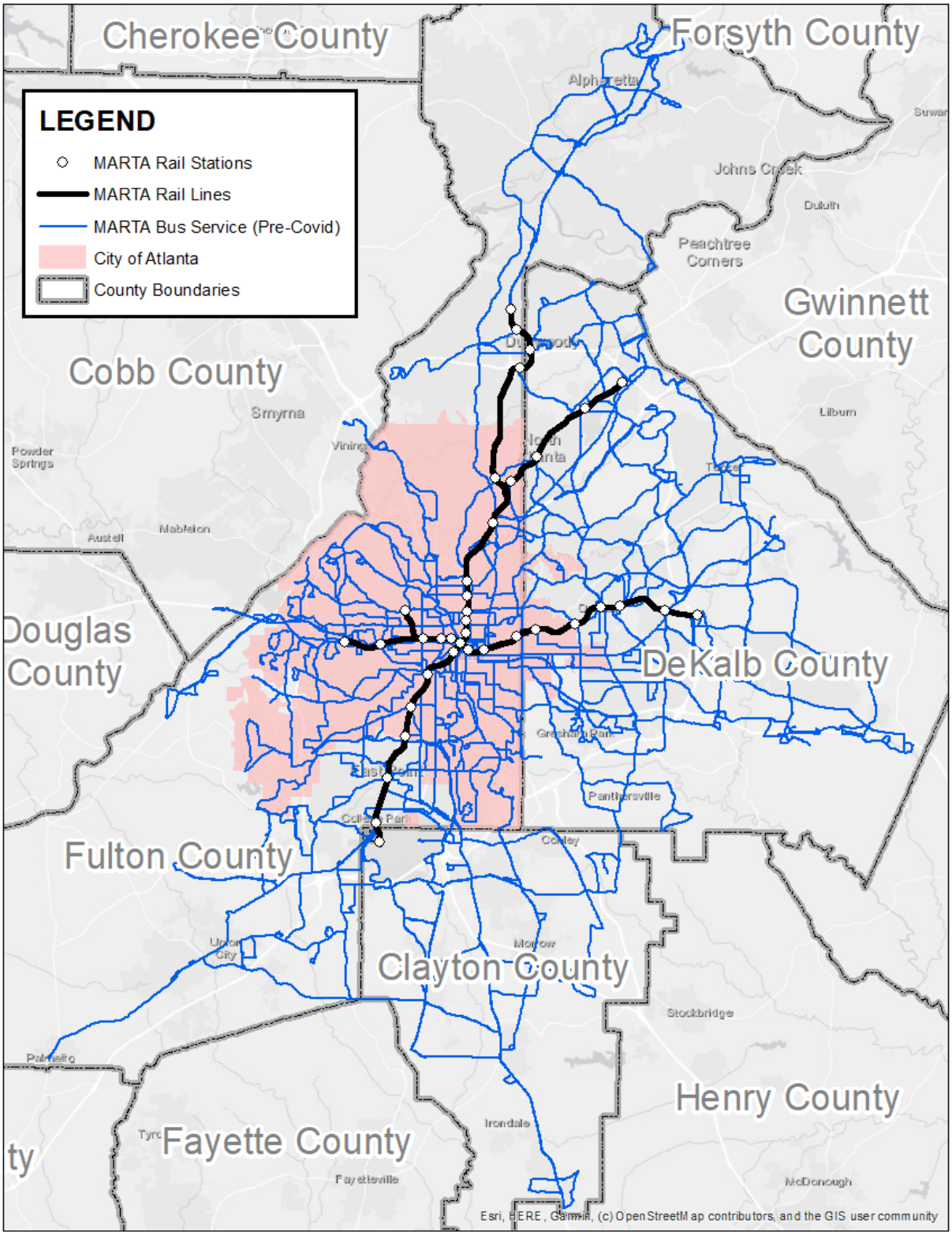


FIGURE 2. Metropolitan Atlanta Rapid Transit Authority (MARTA) provides rail and bus transit services in Fulton, DeKalb and Clayton counties in Georgia

2.3 TEXT MINING AND SENTIMENT ANALYSIS

A combination of R and Python 3 programming languages were used to prepare and analyze our data. R was used in the open-source RStudio Desktop version 1.2.5042 integrated development environment (IDE), while Python was used in JetBrain's PyCharm version 2019.2.6 (Community Edition) IDE. Two Natural Language Processing (NLP) Python libraries – Natural Language Tool Kit version 3 (NLTK) and TextBlob – were used for word frequency analysis and sentiment analysis. The Matplotlib and Wordcloud Python libraries were used to plot and visualize data.

2.3.1 Data Preparation

R was used for exploratory examination of the survey responses data provided by MARTA. As we are interested in GIS or location-based examination of transit customer satisfaction, we first performed an exploratory examination of responses to the questions on 'home zip code' and 'MARTA rail station most convenient to home'. Responses to 'home zip code' appeared less reliable as there were 200 responses that provided answers that did not begin with 3 (the first number of all zip codes in Georgia) or that were less than five digits in length. This may be because some respondents were unwilling to provide home zip code information due to privacy concerns. All respondents provided an answer for 'MARTA rail station most convenient to home'. Consequently, the MARTA rail station most convenient to home was selected to represent the location of transit customers.

A data frame that contained responses for "MARTA rail station most convenient to home" ("Home Station") and "Other Comments" was created using R. Survey responses that did not provide answers to these two questions were removed from the dataset. The remaining data frame of 5,393 responses was saved in csv format. In subsequent data cleaning and analyses, these responses were further aggregated in Python according to "Home Station".

2.3.2 Determining most frequent words

The NLTK Python library was used to further clean and prepare data for word frequency analysis. Data cleaning was improved iteratively before final word frequency analysis. Briefly, all responses to the "Other Comments" survey question were first transformed to lower case and tokenized (strings of sentences separated and recognized as single words), and all punctuations were removed. Next, stop words from the NLTK library were removed from our data. Stop words are words that are very common in a language, such as 'the', 'is', 'a', 'there', 'my', 'very' and 'most'. Different NLP tools define different lists of stop words. Stop words defined in the NLTK library for the English language (Loper, Klein & Bird, 2019) was removed from our data in the first round of data cleaning.

After each iterative round of data cleaning, the most frequently occurring words in transit customer comments aggregated by "Home Station" was determined using the Collections, Matplotlib and Wordcloud modules in Python. Results from each round of analysis was examined to identify words that were common for all Home Stations and therefore did not provide location-specific information. Consequently, in subsequent rounds of data cleaning, common words such as 'marta', 'ride', 'route', 'bus', 'station', 'train' and 'rider' were removed from our data. Words that are part of station names such as 'lenox', 'brookhaven', 'civic', 'center', 'college', 'park', 'east' and 'point' were also removed as their frequent occurrence at their respective stations resulted in an over-emphasis of station name and the overshadowing of other words. Nonetheless, 'airport' was retained because it provided information on locations where people boarded the train to go to the airport. Other common words such as 'work', 'service', 'time', 'late', 'stop' and 'driver' were also retained to avoid removing too many words that may provide context or insights for interpreting the significance of other words. Ultimately, while removing

words that occur too frequently to provide meaningful information, we also avoided removing too many words that might in fact provide meaningful information.

Results from the first round of analysis is attached in APPENDIX 2 to demonstrate results before the custom removal of MARTA-specific stop words. Before the final round of analysis, the Porter Stemming Algorithm (Porter, 2006) that is provided as a module in the NLTK library was also used to collapse words that have a common root word; E.g. 'writer', 'writing', 'wrote' and 'written' would be stemmed or transformed to their root word of 'write'. The final Python script used for this analysis is attached in APPENDIX 1.

2.3.3 Sentiment Analysis

Sentiment analysis was performed using NLTK and TextBlob libraries, and the results from these analyses were compared. Both the NLTK Vader module and TextBlob library analyze sentiments by assigning positive or negative values to each word. Transit customer sentiment at each Home Station is represented by the mean of positive and negative sentiment values. In order to assign whether a Home Station is associated with a positive or negative sentiment, the sentiment score for each station is compared to the median sentiment for all stations. If the sentiment score for a station is greater than the median sentiment score, the customer sentiment at the station is labelled and mapped as "positive". Conversely, if the sentiment score for a station is lower than the median sentiment score, the customer sentiment at the station is labelled and mapped as "negative".

The TextBlob library provides additional analysis for subjectivity. Subjectivity is a measure of whether comments are more factually stated, or more opiniated. Comments are considered more subjective or opiniated when there is more use of "intensifier" words such as 'very' and 'great'. Intensifiers are scored and the mean score for each Home Station is recorded as a subjectivity measure. A scatterplot of results is generated using the Matplotlib library, and sentiments are mapped as described below. Python scripts

developed for sentiment analysis using the NLTK Vader module and the TextBlob library are attached in APPENDIX 3.

2.4 OPEN-SOURCE VISUALIZATION AND MAPPING

All analyses and mapping in this study were performed using Python to remove commercial software dependency and enable similar analyses and mapping to be reproduced from the Python scripts by any transit agency independent of software and resources. The Folium Python library was selected for mapping in this study to provide an interactive map that may be embedded into a website for publication and distribution. The open-data and open-source Open Street Map service was used as basemap.

In addition to rail station symbology based on sentiment values obtained from the Sentiment Analysis described above, Folium interactivity was further enhanced using the 'Tooltip' and 'Popup' Folium functions to provide additional information to map users. 'Tooltip' was programmed to display the name of rail stations when a mouse hovers over a marker. 'Popup' was programmed to fetch WordCloud images from the working directory using Home Station as index so that map users may concurrently examine the results from the Frequent Words Analysis described above.

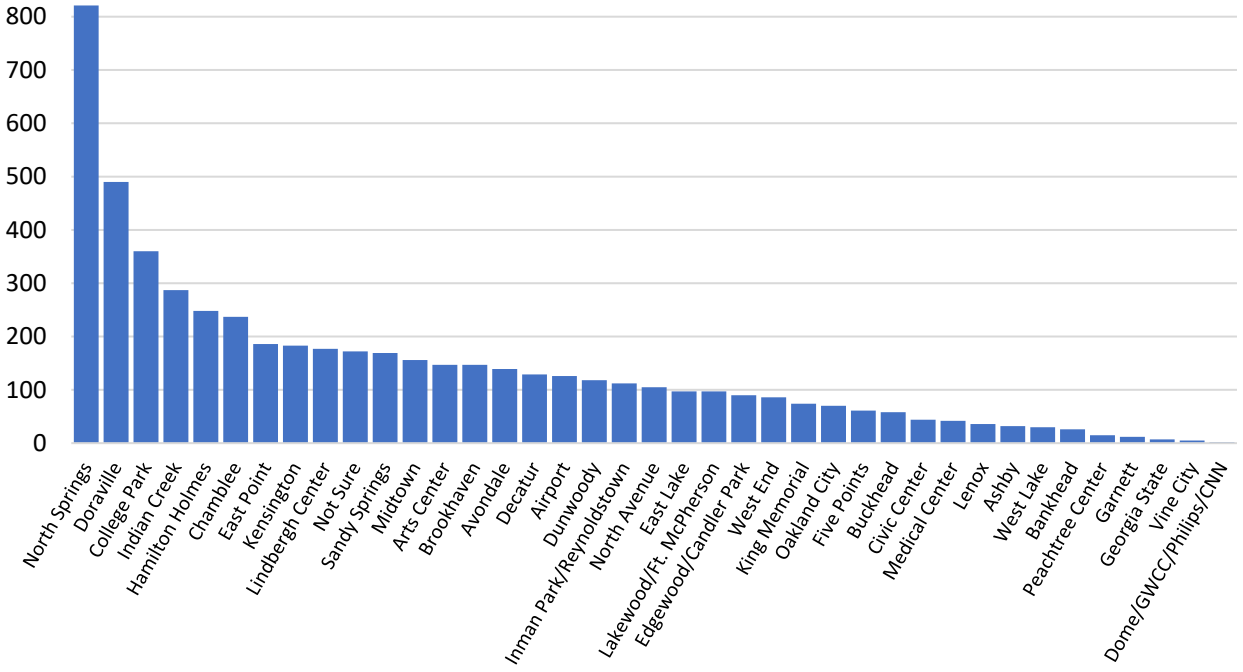
The Geopandas Python library was additionally used to package and display census boundary shapefiles that were fetched by our Python script from the US Census Bureau using File Transfer Protocol (FTP). In our script (APPENDIX 4), workers separately download their desired census data at the census tract level and provide the location of the census data in the script. The Python code automatically joins the census data to the TIGER/LINE census tract boundary by 'GEOID'. For Python scripts that would automate fetching census data and census boundaries with user input for Year, State FIP, County FIP, and desired census field, please refer to Yap (2019). Even though Yap (2019) performed data analysis and mapping using the ESRI-based Arcpy Python library, Python code that provide user-input functionality to

fetch census data and boundaries for any year and county in the US may be merged with our current study. In the current study, we focused on providing open-source and interactive visualization and mapping. The Python libraries Request, Geopandas and Folium were combined to fetch census information and provide an overlay of 'Percentage Households with Zero Vehicles' by census tract in our interactive Folium map to provide greater location-based and sociodemographic context for sentiments and frequently used words around each MARTA station. The Anaconda (Individual Edition) Python Package Manager was used to overcome the problems frequently associated with specification conflicts and dependencies for the Geopandas library.

3. RESULTS AND DISCUSSION

Survey responses were organized according to the answer given for "MARTA rail station most convenient to home" ("Home Station"). Frequent words analysis and sentiment analysis were carried out with assumption that the Home Station represents the location of survey respondents, and therefore, representing the location that is associated with specific words highlighted by the frequent words analysis and associated with customer sentiment. It is important to note that the significance of words and sentiments observed in this study varies between station due to differences in sample size. The number of responses received for each MARTA Home Station are highlighted in FIGURE 3 and TABLE 2. The most number of survey responses (821 responses) was received with North Springs as Home Station, and the lowest number of survey responses (2 responses) was received with Dome/GWCC/Phillips/CNN as Home Station. 172 responses replied "Not Sure" to the Home Station survey question.

FIGURE 3. Number of responses after respondents who did not provide "Other Comments" were removed from dataset



3.1 DETERMINING MOST FREQUENT WORDS

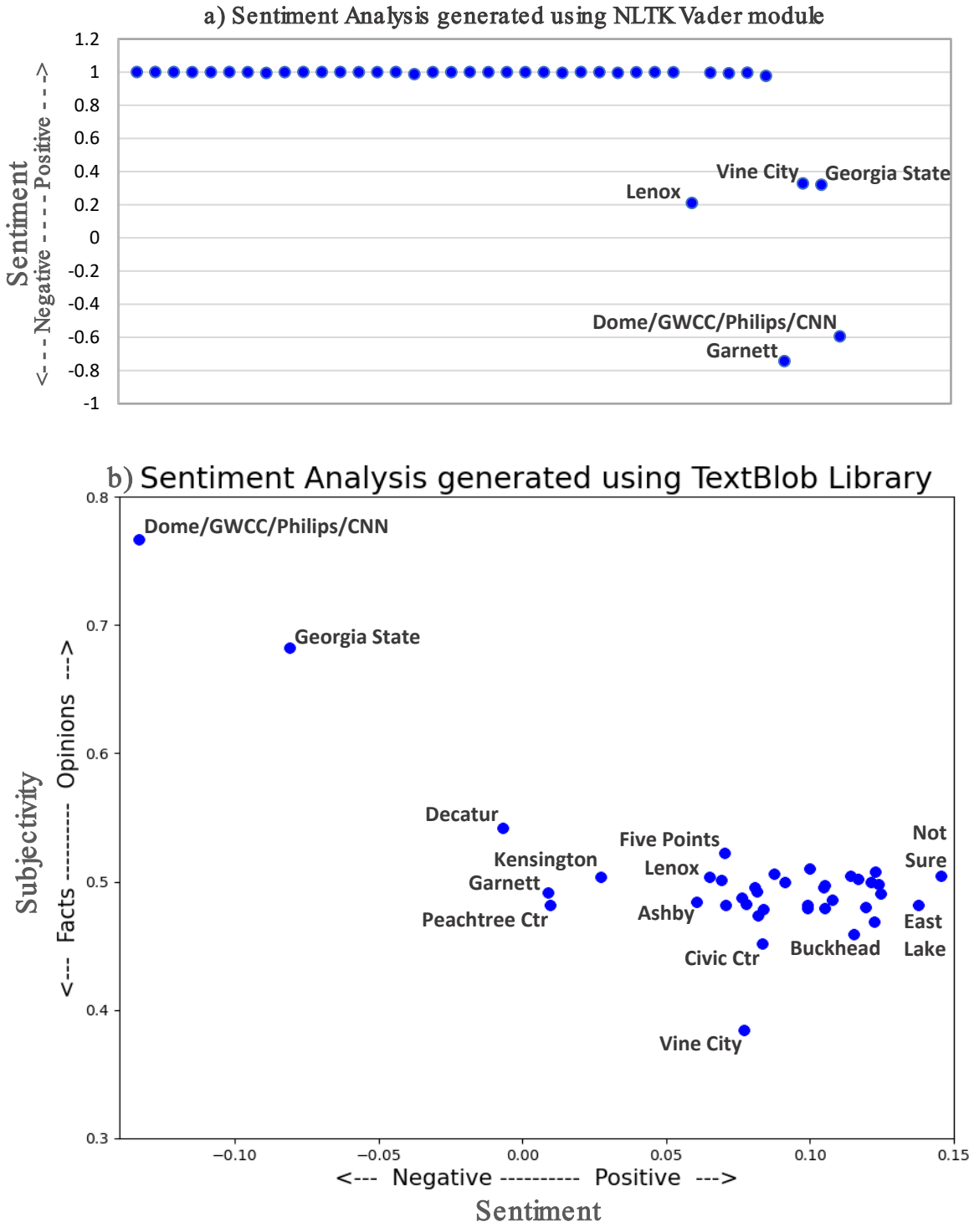
In the Frequent Words Analysis, we attempted to elucidate information about specific Home Stations based on words that are commonly used in survey comments. In our first attempt, we discovered that station names and words such as ‘marta’ and ‘route’ were the most common words for most stations, and did not provide meaningful information. Consequently, we removed these words in an iterative process to try to elucidate words that may be specific to a station, or have particular significance for a station. However, it was difficult to differentiate the importance of a word. For example, ‘late’ and ‘work’ occurred commonly for all stations, but removing them may result in the removal of words that may provide important context or information. The final list of words that were removed (APPENDIX 1) is the result of a compromised approach to this analysis.

A quick scan of wordclouds generated from this analysis (APPENDIX 5) may suggest that several words (E.g. ‘time’, ‘use’ and ‘service’) were still occurring consistently or commonly for all stations. However, while it was difficult to discern words that were specific for each Home Station throughout the transit system, a close inspection of the wordclouds did indeed reveal words that were more prominent or specific for certain stations (TABLE 2). Examples include ‘urine’ for Doraville Station, ‘panhandle’ for Chamblee Station and Brookhaven Station, ‘breeze’ for Lindbergh Center Station, ‘escalator’ for Lenox Station, ‘map’ for Bankhead Station, and ‘stroller’ for Georgia State Station.

3.2 SENTIMENT ANALYSIS

The suitability of different NLP tools can differ for different datasets depending on the context and culture that may affect the choice of words and sentence structure used in a dataset. Therefore, sentiment analysis was performed using two Python libraries (NLTK and TextBlob) in this study, and the results from these analyses were compared (FIGURE 4 and TABLE 2). The standard NLTK Vader module (FIGURE 4a) appears less suitable for our dataset as it produced sentiment scores of 0.98 or higher for 34 out of 39 stations. Nonetheless, we note that variations in scores produced in the TextBlob sentiment analysis (FIGURE 4b) may be affected by sample size. Home Stations with the smallest sample size (Dome/GWCC/Philips/CNN: 2 responses, Vine City: 5 responses, and Georgia State: 7 responses) appear to be outliers in subjectivity scoring.

FIGURE 4. Comparison of sentiment analysis using NLTK Vader and TextBlob



If results from outlier stations (Stations with less than 10 survey responses) are discarded, the five stations with the most negative customer sentiments (in order of most to least negative) are Decatur, Garnett, Peachtree Center, Kensington and Ashby stations, and the five stations with the most positive sentiments (in order of most to least positive) are East Lake, Inman Park/Reynoldstown, Bankhead, Airport and Medical Center stations. The five stations with the most opiniated sentiments (in order of most to least opiniated) are Decatur, Five Points, College Park, Airport and Indian Creek stations, and the five stations with the most factual sentiments (in order of most to least factual) are Civic Center, Buckhead, Medical Center, North Avenue and Dunwoody stations.

3.3 LOCATION-BASED ANALYSIS

Table 2 provides a summary of sentiments and frequently used words determined for each MARTA Station. However, it does not provide spatial or geographic context. In this study, we adopt the capabilities of the Folium and Geopandas Python libraries to demonstrate the use and combination of text-mining and natural language processing with GIS visualization as a methodology for exploring and understanding transit customer satisfaction by location. Importantly, the HTML output of the Folium map may be embedded into websites for easy publication and distribution of an interaction map. The interactivity of the Folium map also allows users to pan and zoom as required, which greatly enhances user ability to explore the study area and examine the surrounding environment. Figures 6 and 7 illustrate and outline the strengths and benefits of using Folium combined with Geopandas for visualizing transit customer sentiment.

TABLE 2. Summary of results

Home Station	Number of Responses	NLTK Sentiment	TextBlob Sentiment	TextBlob Subjectivity	Words that were more prominent, or specific to a station
North Springs	821	pos	neg	Opiniated	airport, delay, wait, expand, extend, alpharetta
Doraville	490	pos	pos	Opiniated	downtown, access, smell, urine, expand
College Park	360	pos	pos	Opiniated	driver, late, homeless, weekend
Indian Creek	287	pos	neg	Opiniated	late, delay, expand, driver
Hamilton Holmes	248	pos	pos	Opiniated	cobb, drive, extend, service
Chamblee	237	pos	pos	Factual	airport, driver, panhandle, announce
East Point	186	pos	neg	Factual	driver, late, one, wait
Kensington	183	neg	neg	Opiniated	schedule, driver, late
Lindbergh Center	177	pos	neg	Factual	machine, late, breeze, airport, uber
Not Sure	172	pos	pos	Opiniated	
Sandy Springs	169	pos	pos	Factual	cobb, track, expand, drive
Midtown	156	pos	pos	Opiniated	late, weekend, expand
Arts Center	147	pos	pos	Factual	schedule, expand, one, cobb
Brookhaven	147	pos	pos	Opiniated	airport, schedule, panhandle, drive
Avondale	139	pos	pos	Opiniated	driver, enough, park, wait
Decatur	129	neg	neg	Opiniated	horrible, driver, late, work
Airport	126	neg	pos	Opiniated	card, convenient
Dunwoody	118	pos	neg	Factual	airport, schedule
Inman Park/Reynoldstown	112	pos	pos	Factual	wait, commute, bridge
North Avenue	105	neg	neg	Factual	schedule, driver, late, reliable
East Lake	97	pos	pos	Factual	delay, commute, convenient
Lakewood/Ft. McPherson	97	neg	neg	Factual	late, driver, traffic, money, homeless
Edgewood/Candler Park	90	pos	pos	Factual	atlanta, card, one, commute, schedule
West End	86	neg	neg	Factual	driver, schedule, expand
King Memorial	74	pos	pos	Opiniated	commute, bike, car, extreme
Oakland City	70	neg	pos	Factual	driver, wait, side, 24 hour, weekend

Home Station	Number of Responses	NLTK Sentiment	TextBlob Sentiment	TextBlob Subjectivity	Words that were more prominent, or specific to a station
Five Points	61	neg	neg	Opiniated	driver, late, stand, wait, app, smoke
Buckhead	58	neg	pos	Factual	schedule, delay, app
Civic Center	44	neg	neg	Factual	live, police, secure, expand
Medical Center	42	neg	pos	Factual	airport, secure, announce
Lenox	36	neg	neg	Opiniated	clean, escalator, convenient, homeless
Ashby	32	neg	neg	Factual	driver, late, schedule, disable, 25, app
West Lake	30	neg	neg	Opiniated	driver, route, late, 10, mobile
Bankhead	26	neg	pos	Opiniated	show, map, might, close
Peachtree Center	15	neg	neg	Factual	ride, arrive, app, hour
Garnett	12	neg	neg	Factual	emory, inside, empty, rule, patient, weekend
Georgia State	7	neg	neg	Opiniated	stroller, house, home, work, drive
Vine City	5	neg	neg	Factual	public, politician, moremarta
Dome/GWCC/Philips/CNN	2	neg	neg	Opiniated	customer, service, agent, kelly, number, gate

FIGURE 5. The HTML map may be uploaded and embedded into websites for easy publication and distribution of an interactive map without the need of commercial software licenses

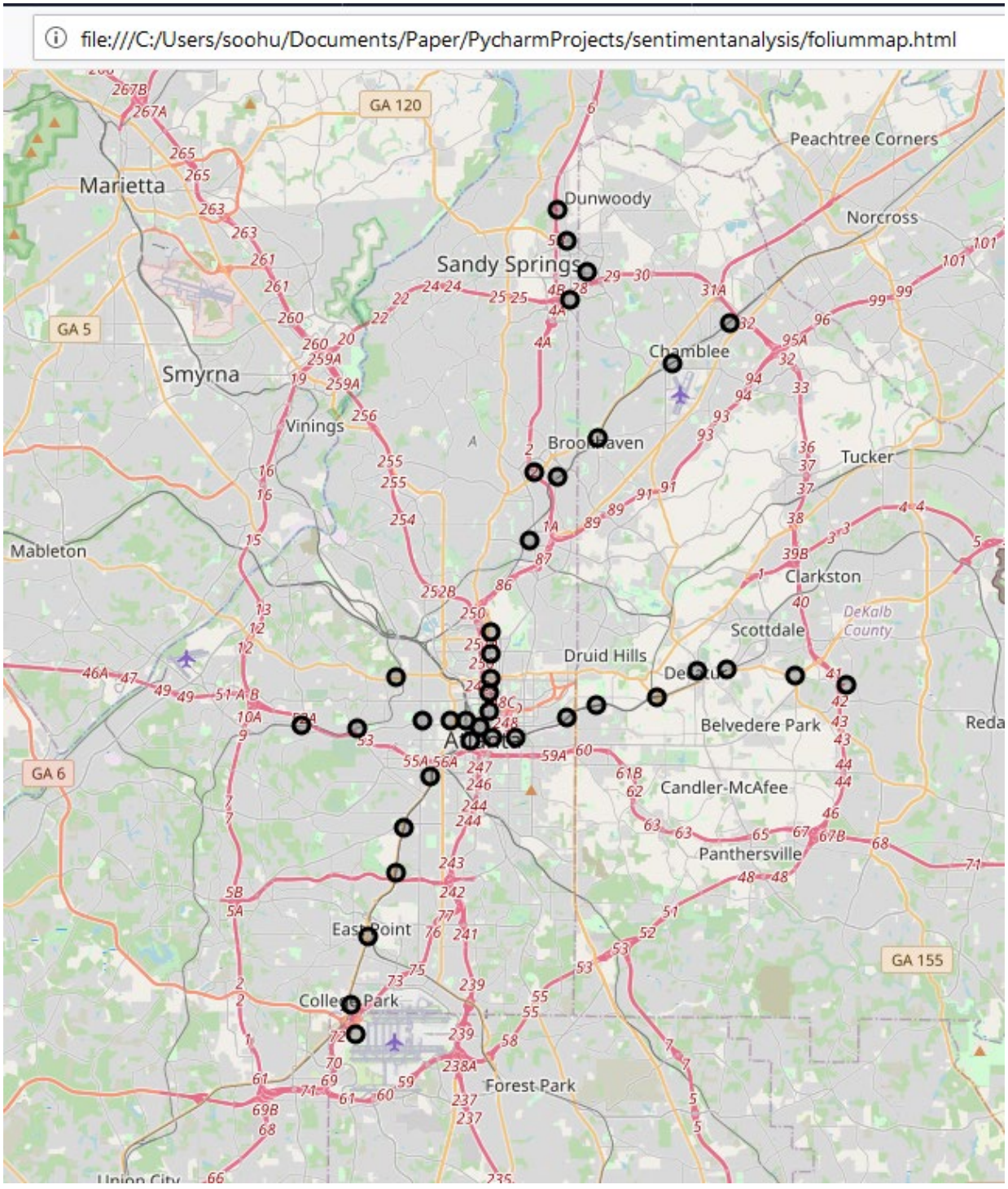
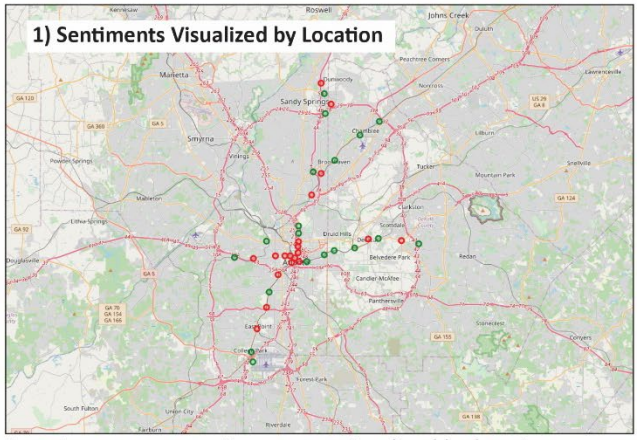
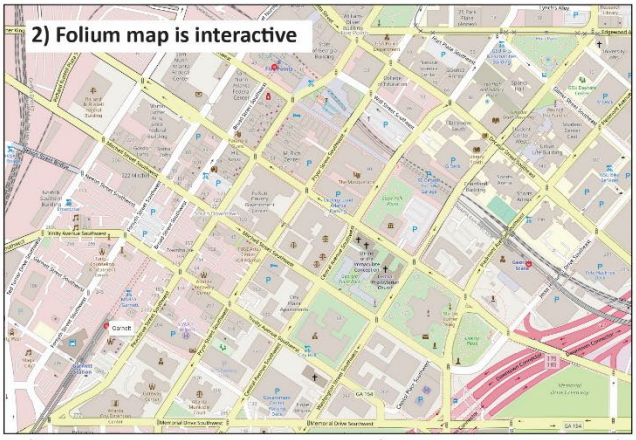


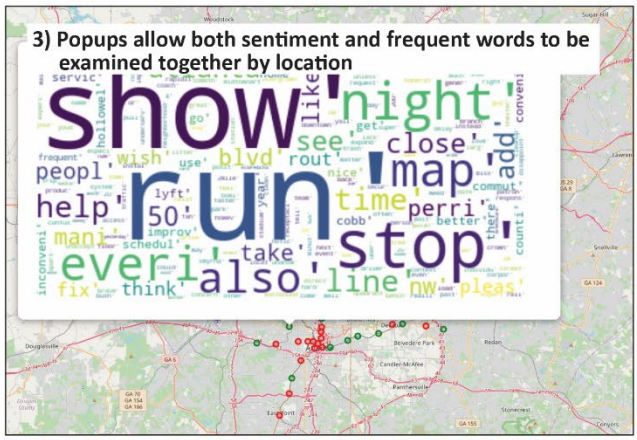
FIGURE 6. Benefits of combining Text-Mining analyses with GIS using the Folium Python functions



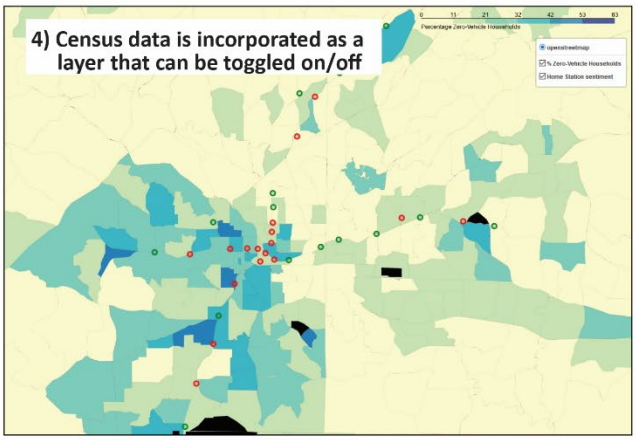
Transit customer sentiments are visualized by location. Positive sentiments are symbolized in green, and negative sentiments are symbolized in red.



Folium generates interactive maps where users may pan and zoom as required to understand the study area and surrounding environment. Additional information may be provided using the 'Tooltip' function where a box with information automatically appears when the user hovers his/her mouse over a station. Tooltip can be coded to provide demographic, sentiment or other data as required. In this screenshot, Home Station name (Garnett) on bottom left shows when mouse is placed above the station.



The results of Word Cloud analysis are linked to the Folium map by Home Station name. Using the 'Popup' function, we allow map users to click on any station of interest to examine words that are associated with the station. This allows users to combine the study of location, sentiment and frequent words to gain a better understanding of transit customer satisfaction and concerns at each location.



The Geopandas and Request python libraries are combined with Folium to access census boundary shapefiles by FTP, to join US Census Bureau American Community Survey data with census boundaries, to perform sociodemographic calculations, and to overlay socio-demographic data in the Folium map. Here, the percentage of transit dependent households (Households without vehicles) is mapped to demonstrate use of Geopandas and the ability to incorporate study of sociodemographic data. Map layers may be toggled on/off as required. Python's computing power allows the functionality of this map to be expanded to provide additional calculations and analysis.

4. CHALLENGES AND FUTURE WORK

The aim of this study was to develop a text-mining and GIS-based methodology that would allow transit agencies to objectively and systematically evaluate transit system performance from a customer perspective. The customer perspective was measured using sentiment scores and frequently used words. These were visualized on a Folium map to provide spatial and geographic context, and which may be explored interactively.

Our study successfully demonstrates the potential for using Python-based NLP analyses and mapping to evaluate transit system performance from a customer perspective. However, we note that some subjectivity was necessary in both Word Cloud / Frequent Word Analysis and Sentiment Analysis. Namely, in evaluating word frequency, we determined that standard stop word libraries were insufficient for transit-related analyses as the standard NLTK stop word library did not account for words frequently used in communication about transit; For example, 'route', 'stop', 'bus', 'rail', 'station' and 'driver' are neutral words that would occur very frequently in communication about transit. Without removing transit-specific stop words from the analyses, the significance of other words may be diluted. Nonetheless, the inclusion of words such as 'driver' and 'station' may also be necessary to help identify causal or transit factors that are associated with positive or negative sentiments. This introduces the need for a "judgement call" to decide on which transit-specific words to include or omit from the analysis. The author made the judgement call in this study. However, in adopting such an analysis for performance measurement, each transit agency may have to evaluate and make their own judgement calls, which would impact the objectivity of the analysis as well as the comparability of results across different transit agencies. Alternatively, the present study may be expanded by working collaboratively with multiple transit agencies with the aim of producing a standard for transit-specific stop words that may be adopted by all American transit agencies to enable comparability and benchmarking between transit agencies.

Similarly, sentiment analysis libraries are developed for generic text. It is expected that the way people phrase sentences and the choice of words used in a transit rider survey would differ from sentence structure and choice of words used in other context. Also, transit agencies may decide to use assign different sentiment polarity (positive vs negative) and values for each word in the sentiment analysis library; For example, the word 'stop' may be assigned a positive or neutral sentiment in a transit study rather than a negative sentiment, and transit agencies may want to place higher negative weightage for words such as 'late', 'dangerous' and 'unsafe'. Therefore, a sentiment analysis library specific for the study of transit surveys may be useful to maximize the applicability of this methodology for transit agencies. It is also important to note that the NLTK library include the Naïve Bayes Classifier and SklearnClassifier modules that may be used for machine-learning to further improve Sentiment Analysis using transit rider survey.

5. CONCLUSION

It is anticipated that community-building and engagement will become markedly more important to transit agencies in the near future as the effects of the Coronavirus Disease 2019 (COVID-19) pandemic will force transit agencies to pivot from a narrow emphasis on increasing ridership, and to work on gaining trust and support from community. A better understanding of transit customer concerns and sentiment would allow transit agencies to improve community engagement. Deepening this understanding according to location is necessary for more targeted and effective efforts to address shortcomings in agency facilities and services.

This study developed and demonstrated a methodology for text-mining and location-based analysis of transit customer perspective and sentiment. The GIS component includes interactive visualization of

sentiment locations couple with Word Cloud popup functionality to provide exploration of issues and concerns related to each sentiment and location. Overlays with census data provide further context.

Nonetheless, in order for such an analysis to become a standard for performance evaluation by transit agencies, more work may be required to develop transit-specific standards for stop words list and sentiment analysis parameters so that they may be applicable and standardized across transit agencies to allow comparability and benchmarking between agencies. The development of this methodology on 100% open-source platforms facilitate collaborative development and improvement of this methodology including acquiring feedback and joint-development of stop words, algorithms and sentiment analysis parameters on platforms such as GitHub and Jupyter Notebook.

With improved transit-specific stop words and sentiment analysis, the computational capacity of Python may also be utilized for statistical analysis to explore correlation between sentiment values and socio-demographic factors.

6. REFERENCES

1. Allen-Connelly, C. (2020). "COVID-19 & Public Transit". A Better City: https://www.abettercity.org/news-and-events/blog/covid-19-response-how-is-the-public-transit-system-measuring-up

2. Ali, F., et al. (2019). "Transportation sentiment analysis using word embedding and ontology-based topic modeling." Knowledge-Based Systems **174**: 27-42.

3. Baratian-Ghorghi, F., and H. Ahmadianyazdi. (2017). "Recommendation of a New Transit Performance Measure in the National Transit Database." Journal of Public Transportation **20**(2): 90 - 102.

4. Boyle, D. (2019). "Transit Service Evaluation Standards." Transportation Research Board, Washington, D.C.

5. Chow, W. (2014). Evaluating Online Surveys for Public Transit Agencies Using a Prompted Recall Approach (Master's thesis). Massachusetts Institute of Technology.

6. Collins, C., Hasan, S., and Ukkusuri, S.V. (2013). "A Novel Transit Rider Satisfaction Metric: Rider Sentiments Measured from Online Social Media Data." Journal of Public Transportation **16**(2): 21-45.

7. Federal Transit Administration (FTA) (2020). "The National Transit Database." Federal Transit Administration, Washington, D.C.: https://www.transit.dot.gov/ntd

8. Fletcher, K., Amarakoon, S., Haskell, J., Penn, P., Wilmoth, M., Matherly, D., and Langdon, N. (2014). "TCRP 769: A Guide for Public Transportation Pandemic Planning and Response." Transportation Research Board, Washington, D.C.

9. Haberle, M., et al. (2019). "Geo-spatial text-mining from Twitter — a feature space analysis with a view toward building classification in urban regions." European Journal of Remote Sensing **52**: 2-11.

10. Hassan M, Hawas Y, Ahmed K. (2012). "A Multi-Dimensional Framework for Evaluating the Transit Service Performance." Transportation Research Part A: Policy and Practice **46**(7): 1066-1085.

11. Karimzadeh, M., et al. (2019). "GeoTxt: A scalable geoparsing system for unstructured text geolocation." Transactions in GIS **23**: 118-136.

12. Kittelson & Associates Inc., Urbitran Inc., LKC Consulting Services Inc., MORPACE International Inc., Queensland University of Technology, and Nakanishi, Y. (2003). "A Guidebook for Developing a Transit Performance-Measurement System." Transportation Research Board, Washington, D.C.

13. Loper, E., Klein, E. and S. Bird. (2019). "Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit." Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 United States License: http://www.nltk.org/book/

14. Luong, T. T. B. and D. Houston (2015). Public opinions of light rail service in Los Angeles, an analysis using Twitter data. iConference 2015.

15. Mendez, J. T., Lobel, H., Parra, D. and Herrera, J.C. (2019). "Using Twitter to Infer User Satisfaction With Public Transport: The Case of Santiago, Chile." IEEE Access **7**: 60255-60263.

16. Nakanishi, Y.J., and G.F. List. (2000). "Regional Transit Performance Indicators: A Performance Measurement Model." Rensselaer Polytechnic Institute, Troy, NY.

17. NCHRP Synthesis 300. (2001). "Performance Measures for Research, Development, and Technology Programs." Transportation Research Board Executive Committee

18. Porter, M. (2006). "The Porter Stemming Algorithm." https://tartarus.org/martin/PorterStemmer/

19. Resch, B., et al. (2018). "Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment." Cartography and Geographic Information Science **45**(4): 362-376.

20. Ryus, P., K. Coffel, J. Parks, V. Perk, L. Cherrington, J. Arndt, Y. Nakanishi, and A. Gan. (2010). "A Methodology for Performance Measurement and Peer Comparison in the Public Transportation Industry." Transportation Research Board, Washington, D.C.

21. Schweitzer, L. (2014). "Planning and Social Media: A Case Study of Public Transit and Stigma on Twitter." Journal of the American Planning Association **80**(3): 218-238.

22. Smith, S. (2020). "Coronavirus Outbreak Triggers New Thinking About Transit's Essential Role." Next City: https://nextcity.org/daily/entry/coronavirus-outbreak-triggers-new-thinking-about-transits-essential-role

23. Yap, S.H. (2019). "Automation of Mandatory Triennial Title VI and Environmental Justice Service Monitoring (Transit Access) Analysis for Transit Agencies Receiving Funding from the US FTA." Project submitted for CP6581 Programming for GIS coursework at Georgia Institute of Technology.

24. Unger, H., Heller, A., Lane, L.B., and D. Matherly. (2019). "Social and Economic Sustainability Performance Measures for Public Transportation: Final Guidance Document." Transportation Research Board, Washington, D.C.

25. US Census Bureau (2020). "Quick Facts". Retrieved on July 13th, 2020 from https://www.census.gov/quickfacts/fact/table/fultoncountygeorgia,dekalbcountygeorgia,claytoncountygeorgia/PST045219.

26. Utomo, M. N. Y., et al. (2018). "Geolocation Prediction in Social Media Data Using Text Analysis: A Review." International Conference on Information and Communications Technology**:** 84-89.

27. Walker, J. (2020). "In a Pandemic, We're All 'Transit Dependent'." Citylab: https://www.citylab.com/perspective/2020/04/coronavirus-public-transit-subway-bus-ridership-revenue/609556/

28. Zane, D. and Ohland, G. (2020). "Public Health's Impact on Future Transit Behavior & Urban Boulevards." The Planning Report: https://www.planningreport.com/2020/03/30/public-health-s-impact-future-transit-behavior-urban-boulevards

APPENDIX 1:

FINAL PYTHON SCRIPT FOR FREQUENT WORDS ANALYSIS

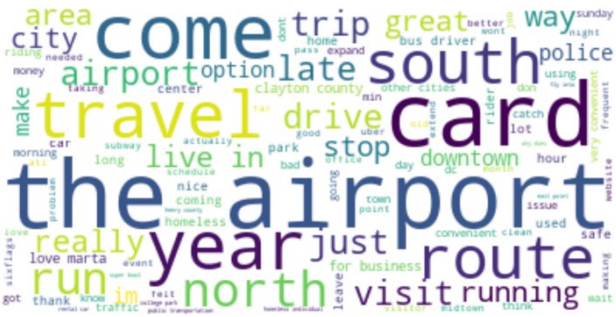
```
import string
from collections import Counter
from nltk.tokenize import word_tokenize
from nltk.stem.porter import PorterStemmer
from nltk.corpus import stopwords
import pandas as pd

df =
pd.read_csv(r'C:\Users\soohu\Documents\Paper\PycharmProjects\sentimentanalysis\Survey
1.csv',header=0)
data_df = df.groupby(['Home_Station'])['Comments'].apply(' '.join).reset_index()
data_df = data_df.sort_index()

stopwords = ['ourselves', 'hers', 'between', 'yourself', 'but', 'again', 'there',
'about', 'once', 'during',
             'out', 'very', 'having', 'with', 'they', 'own', 'an', 'be', 'some',
'for', 'do', 'its', 'yours',
             'such', 'into', 'of', 'most', 'itself', 'other', 'off', 'is', 's',
'am', 'or', 'who', 'as', 'from',
             'him', 'each', 'the', 'themselves', 'until', 'below', 'are', 'we',
'these', 'your', 'his', 'through',
             'don', 'nor', 'me', 'were', 'her', 'more', 'himself', 'this', 'down',
'should', 'our', 'their',
             'while', 'above', 'both', 'up', 'to', 'ours', 'had', 'she', 'all',
'no', 'when', 'at', 'any','station ',
             'before', 'them', 'same', 'and', 'been', 'have', 'in', 'will', 'on',
'does', 'yourselves', 'then',
             'that', 'because', 'what', 'over', 'why', 'so', 'can', 'did', 'not',
'now', 'under', 'he', 'you',
             'herself', 'has', 'just', 'where', 'too', 'only', 'myself', 'which',
'those', 'i', 'after', 'few',
             'whom', 't', 'being', 'if', 'theirs', 'my', 'against', 'a', 'by',
'doing', 'it', 'how', 'further',
             'was', 'here',
'than','bus','marta','ride','route','buses','station','train','trains','rider','rider
s',

'Arts','Center','Ashby','Avondale','Bankhead','Brookhaven','Buckhead','Chamblee','Civ
ic','College','Park',

'Decatur','Dome','GWCC','Philips','CNN','Doraville','Dunwoody','East','Lake','Point',
'Edgewood','Candler',

'Five','Points','Garnett','Georgia','State','Hamilton','Holmes','Indian','Creek','Inm
an','Reynoldstown',

'Kensington','King','Memorial','Lakewood','Fort','Ft','McPherson','Lenox','Lindbergh'
,'Medical','Midtown',

'North','Avenue','Springs','Oakland','City','Peachtree','Sandy','Vine','West','End','
```

```
Lake']

def clean_text_round1 (text):
    text = text.lower()
    text = text.translate(str.maketrans(' ', ' ', string.punctuation))
    text = word_tokenize(text,"english")
    return text

round1 = lambda x: clean_text_round1(x)
data_clean = pd.DataFrame(data_df.Comments.apply(round1))

data_clean['Comments'] = data_clean['Comments'].apply(lambda x: [item for item in x
if item not in stopwords])

data = data_clean.transpose()
top_dict = {}
for c in data.columns:
    top = data[c].sort_values(ascending=False).head(30)
    top_dict[c] = list(zip(top.index, top.values))

words = []
for index in data.columns:
    top = [word for (word, count) in top_dict[index]]
    for t in top:
        words.append(t)

Counter(words).most_common()
add_stop_words = [word for word, count in Counter(words).most_common() if count > 20]

data_clean['Comments'] = data_clean['Comments'].apply(lambda x: [item for item in x
if item not in add_stop_words])

ps = PorterStemmer()
data_clean['Comments'] = data_clean['Comments'].apply(lambda x: [ps.stem(y) for y in
x])
data_clean['Home_Station'] = data_df['Home_Station']
print(data_clean)

from wordcloud import WordCloud
import matplotlib.pyplot as plt

def generate_wordcloud(c, title = None):
    wc = WordCloud(background_color="white",
                           max_words=200,
                           max_font_size=100,
                           random_state=9).generate(str(c))
    plt.rcParams['figure.figsize'] = [12, 12]
    plt.axis("off")
    plt.imshow(wc, interpolation="bilinear")

generate_wordcloud(data_clean.Comments[31])
plt.show()
```

APPENDIX 2:

RESULTS FROM 1ST ROUND DATA CLEANING AND FREQUENT WORDS ANALYSIS

Airport Station



Arts Center Station

Ashby Station

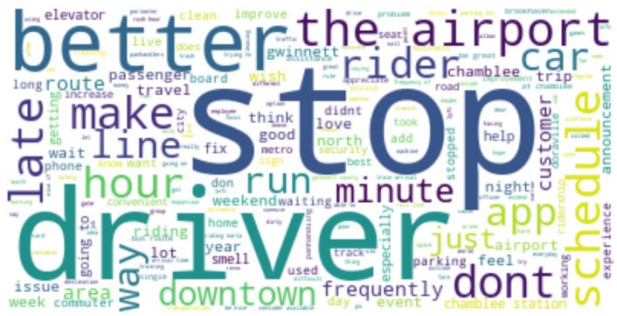
Avondale Station

Bankhead Station

Brookhaven Station

Buckhead Station

Chamblee Station

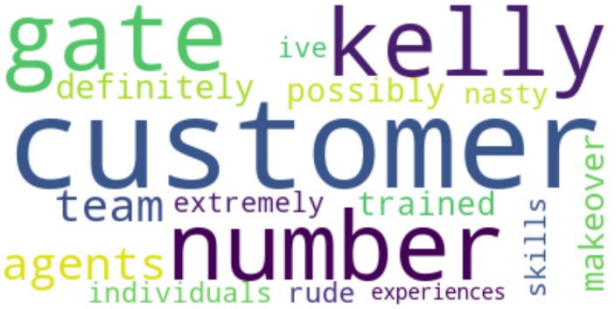


Civic Center Station

College Park Station

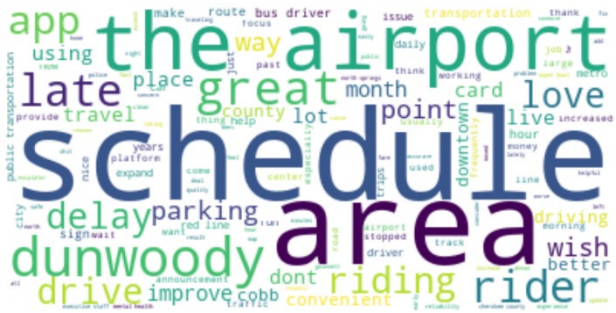
Decatur Station

Dome/GWCC/Philips/CNN Station



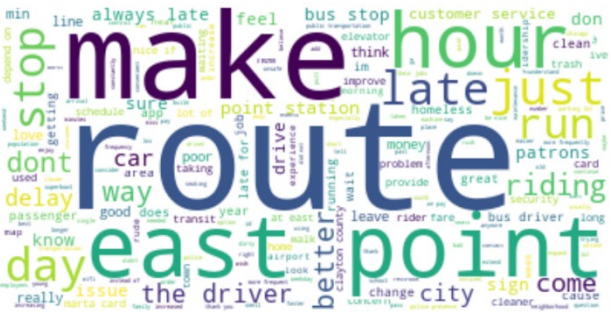
Doraville Station

Dunwoody Station



East Lake Station

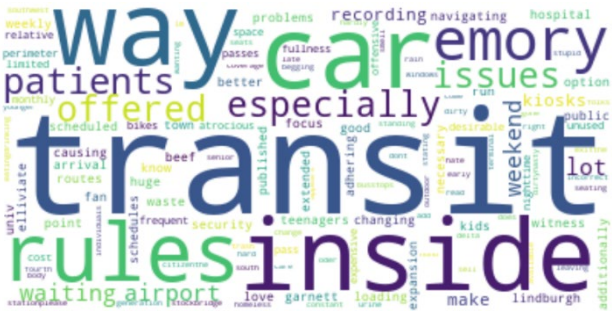
East Point Station



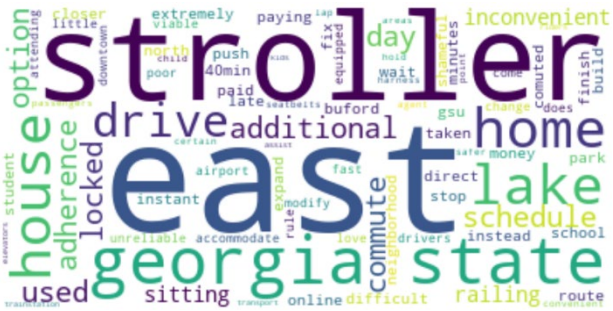
Edgewood/Candler Park Station

Five Points Station

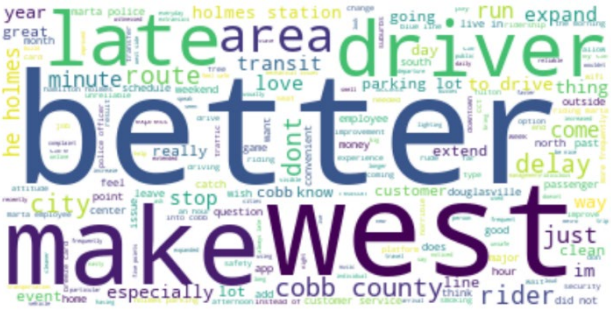
Garnett Station



Georgia State Station



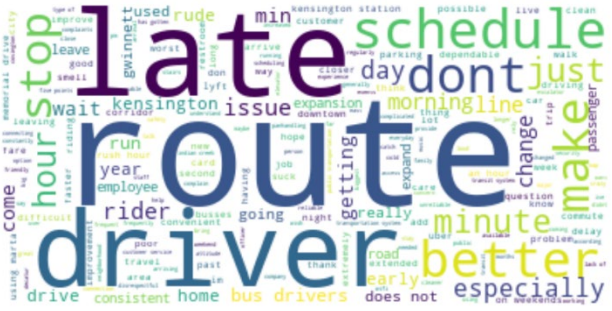
Hamilton Holmes Station



Indian Creek Station

Inman Park/Reynoldstown Station

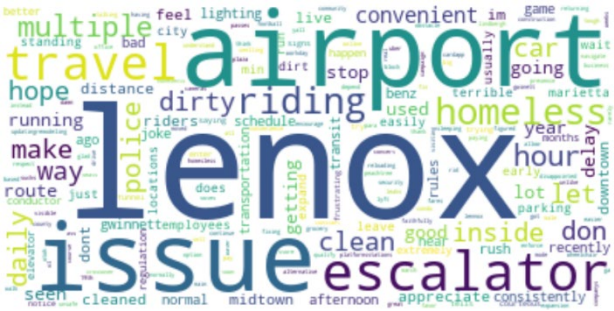
Kensington Station



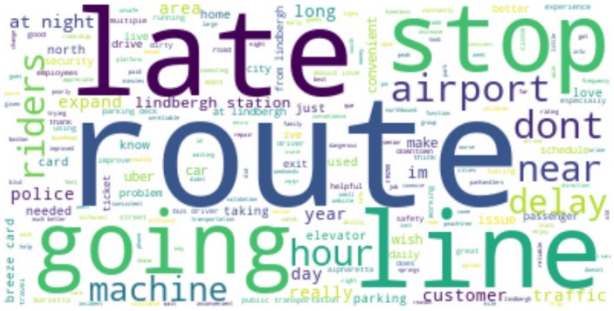
King Memorial Station

Lakewood/Ft. McPherson Station

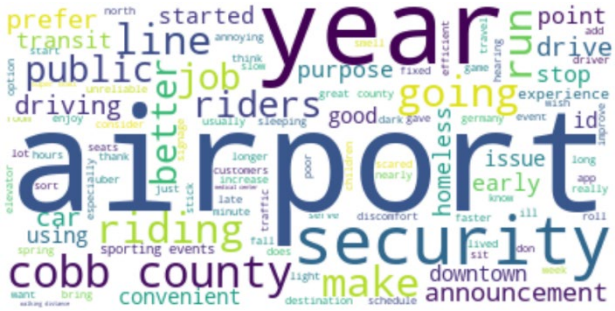
Lenox Station



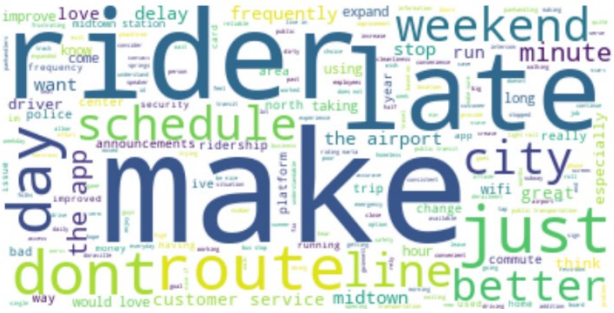
Lindbergh Center Station



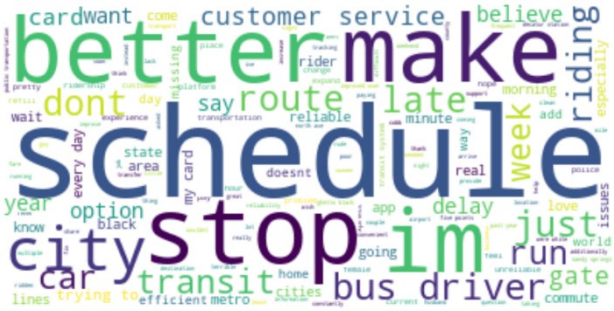
Medical Center Station



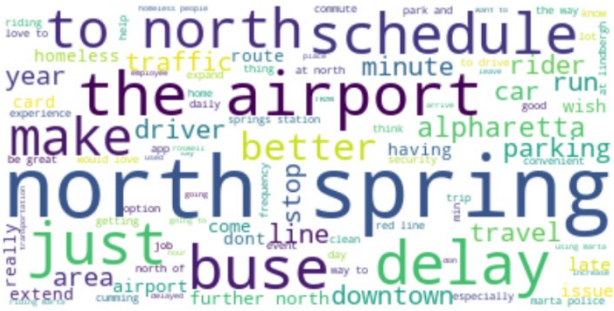
Midtown Station



North Avenue Station



North Springs Station

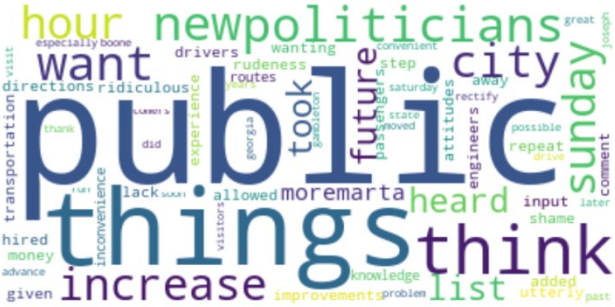


Oakland City Station

Peachtree Center Station

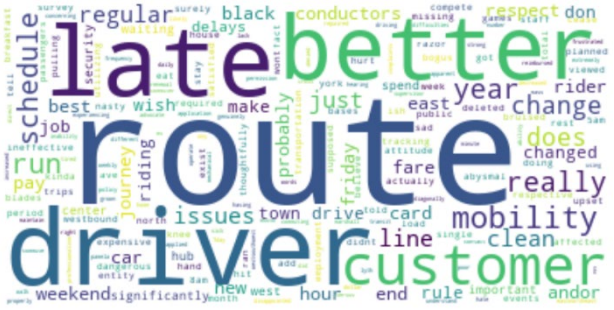
Sandy Springs Station

Vine City Station



West End Station

West Lake Station



Not Sure

APPENDIX 3:

PYTHON SCRIPT FOR SENTIMENT ANALYSIS USING THE NLTK VADER MODULE

```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import matplotlib.pyplot as plt, pandas as pd

df =
pd.read_csv(r'C:\Users\soohu\Documents\Paper\PycharmProjects\sentimentanalysis\Survey
1.csv',header=0)
data_df = df.groupby(['Home_Station'])['Comments'].apply(' '.join).reset_index()
data_df = data_df.sort_index()

def nltk_sentiment(sentence):
    nltk_sentiment = SentimentIntensityAnalyzer()
    score = nltk_sentiment.polarity_scores(sentence)
    return score

nltk_results = [nltk_sentiment(row) for row in data_df.Comments]
results_df = pd.DataFrame(nltk_results)
nltk_df = data_df.join(results_df)
nltk_df['sentiment'] = nltk_df['compound'].apply(lambda c: 'pos' if c >=0 else 'neg')

nltk_df.to_csv('sentiment_nltk.csv')
```

PYTHON SCRIPT FOR SENTIMENT ANALYSIS USING TEXTBLOB

```
import pandas as pd
from textblob import TextBlob
import matplotlib.pyplot as plt

df =
pd.read_csv(r'C:\Users\soohu\Documents\Paper\PycharmProjects\sentimentanalysis\Survey
1.csv',header=0)
data_df = df.groupby(['Home_Station'])['Comments'].apply(' '.join).reset_index()
data_df = data_df.sort_index()

pol = lambda x: TextBlob(x).sentiment.polarity
sub = lambda x: TextBlob(x).sentiment.subjectivity

data_df['polarity'] = data_df['Comments'].apply(pol)
data_df['subjectivity'] = data_df['Comments'].apply(sub)

textblob_df = data_df
textblob_df.to_csv('sentiment_txtblob.csv')

plt.rcParams['figure.figsize'] = [10, 8]

for Station in zip(data_df['Home_Station']):
    x = data_df.polarity
    y = data_df.subjectivity
    plt.scatter(x, y, color='blue')
    plt.xlim(0, .15)
    plt.ylim(0.3, 0.6)
```

```
plt.title('Sentiment Analysis generated using TextBlob Library', fontsize=20)
plt.xlabel('<---  Negative ----------  Positive  --->', fontsize=15)
plt.ylabel('<---  Facts ----------  Opinions  --->', fontsize=15)
plt.savefig('TxtBlobSentimentPlot2.png')
```

APPENDIX 4:

FINAL PYTHON SCRIPT FOR FOLIUM MAP

```
import pathlib
import urllib.request
import folium
import pandas as pd
import geopandas as gpd
import base64
from folium import IFrame

senti =
pd.read_csv(r'C:\Users\soohu\Documents\Paper\PycharmProjects\sentimentanalysis\tbsent
imentlatlon2.csv')
census =
pd.read_csv(r'C:\Users\soohu\Documents\Paper\PycharmProjects\sentimentanalysis\Census
Data.csv', dtype={'GEOID':str})
map = folium.Map(location=[senti['Lat'].mean(), senti['Lon'].mean()], zoom_start=11,
tiles='OpenStreetMap')

tracts_filename = "tl_2018_13_tract.zip"
tracts_url = f"https://www2.census.gov/geo/tiger/TIGER2018/TRACT/{tracts_filename}"
tracts_file = pathlib.Path(tracts_filename)

for data_file, url in zip([tracts_file], [tracts_url]):
    if not data_file.is_file():
        with urllib.request.urlopen(url) as resp, \
                open(data_file, "wb") as f:
            f.write(resp.read())
tracts = gpd.read_file(f"zip://{tracts_file}")

datatracts = tracts.merge(census, on='GEOID', how='right')
datatracts['percNoVeh'] = (datatracts['TOTNoVeh']/datatracts['TOTALHH'])*100

folium.Choropleth(
    datatracts,
    data = datatracts,
    name = '% Zero-Vehicle Households',
    key_on = 'properties.GEOID',
    columns=['GEOID', 'percNoVeh'],
    fill_color = 'YlGnBu',
    fill_opacity = 1,
    line_opacity = 0.05,
    legend_name = 'Percentage Zero-Vehicle Households').add_to(map)
```

```
fg = folium.FeatureGroup(name="Home Station sentiment")

medsenti = senti['polarity'].median()
def color(polarity):
    if polarity < medsenti:
        col = 'red'
    elif polarity > medsenti:
        col = 'green'
    else:
        col = 'yellow'
    return col

for lat, lon, station, polarity, wordcloud in zip(senti['Lat'], senti['Lon'],
senti['Home_Station'], senti['polarity'], senti['Wordcloud']):
    encoded = base64.b64encode(open(wordcloud, 'rb').read()).decode()
    html = '<img src="{}">'.format
    iframe = IFrame(html(encoded), width="980", height="490")
    popup = folium.Popup(iframe)
    fg.add_child(folium.vector_layers.CircleMarker(location=[lat, lon], radius=5,
popup=popup, tooltip=(folium.Tooltip(station)), color=color(polarity),
fill_color=color(polarity)))
map.add_child(fg)

folium.LayerControl().add_to(map)
map.save('finalmap.html')
```

APPENDIX 5:

RESULTS FROM FINAL WORDS ANALYSIS

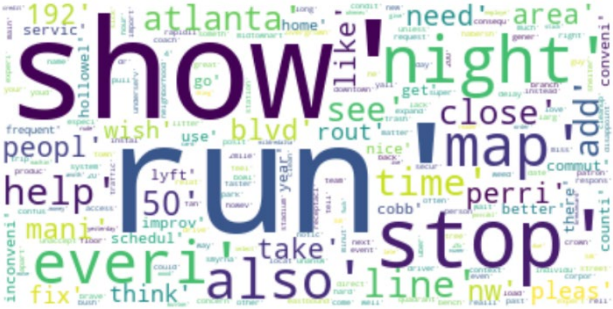
Airport Station

Arts Center Station

Ashby Station

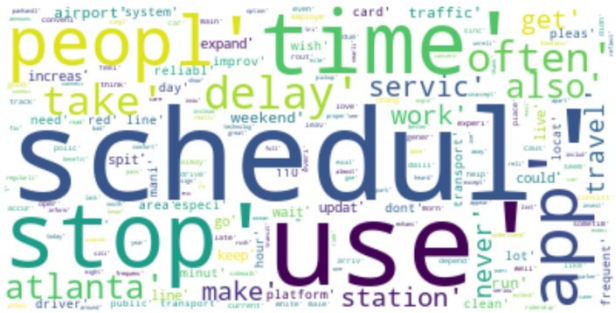
Avondale Station

Bankhead Station



Brookhaven Station

Buckhead Station



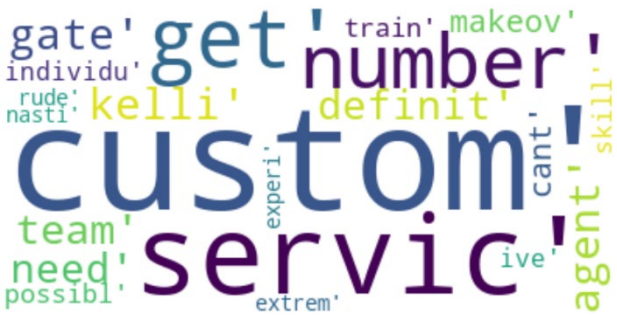
Chamblee Station

Civic Center Station

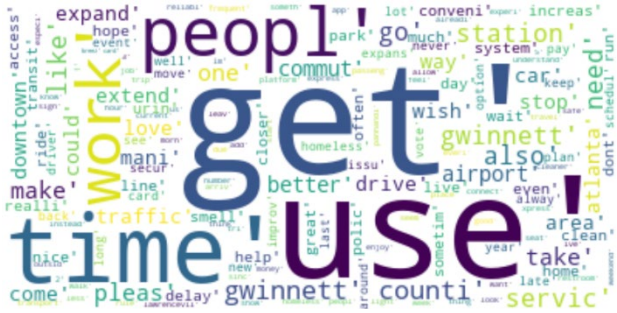
College Park Station

Decatur Station

Dome/GWCC/Philips/CNN Station

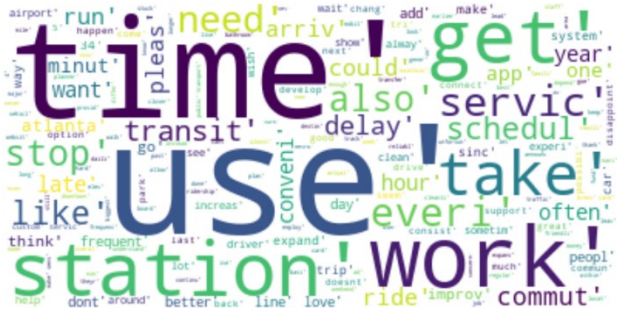


Doraville Station

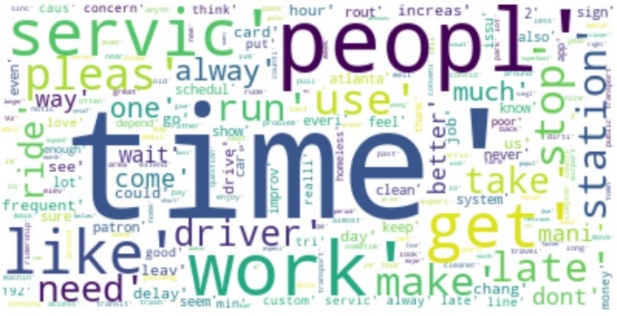


Dunwoody Station

East Lake Station



East Point Station

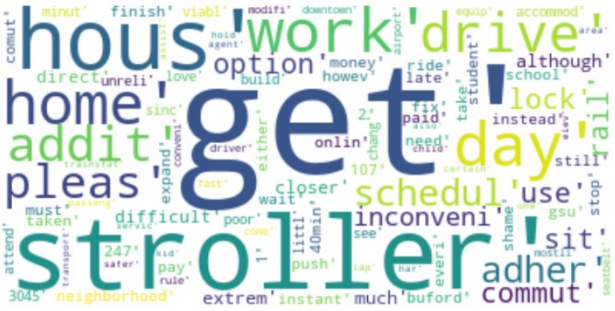


Edgewood/Candler Park Station

Five Points Station

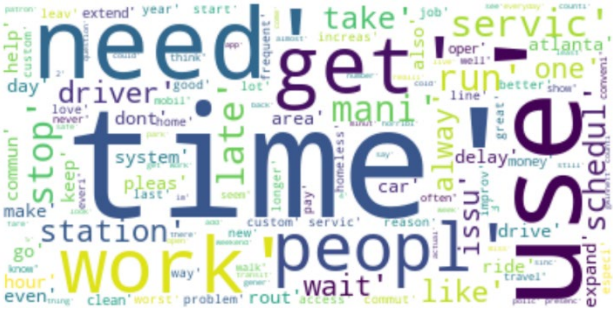
Garnett Station

Georgia State Station

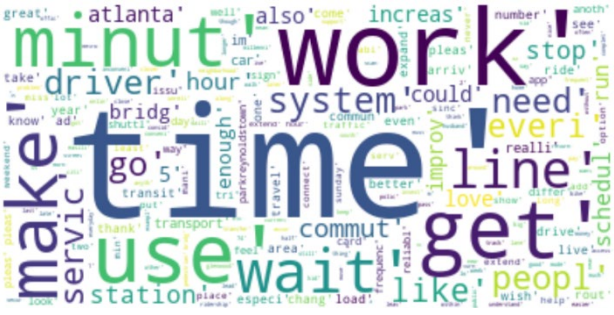


Hamilton Holmes Station

Indian Creek Station



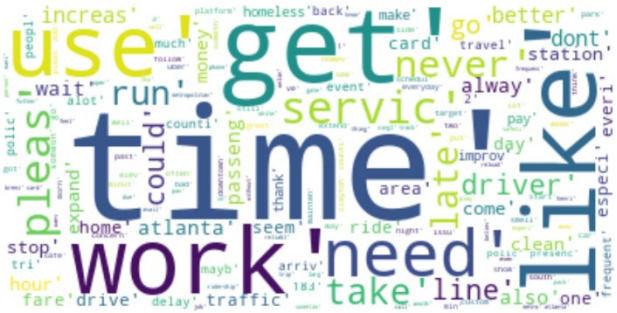
Inman Park/Reynoldstown Station



Kensington Station

King Memorial Station

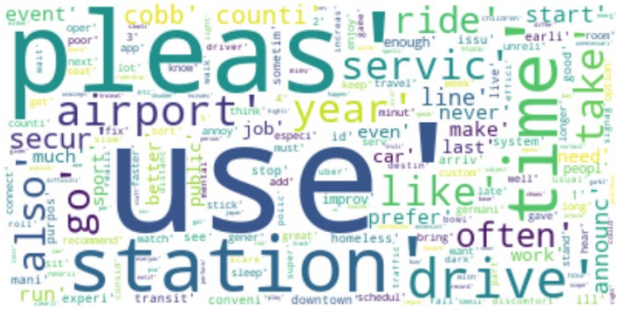
Lakewood/Ft. McPherson Station



Lenox Station

Lindbergh Center Station

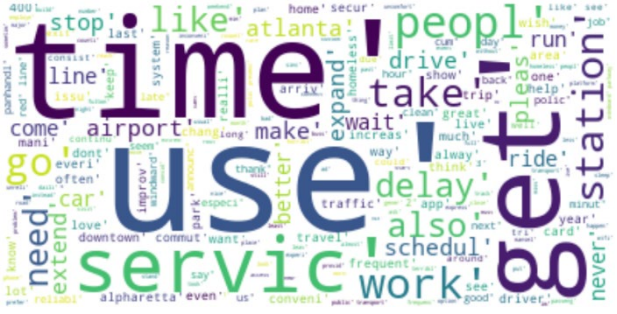
Medical Center Station



Midtown Station

North Avenue Station

North Springs Station



Oakland City Station

Peachtree Center Station

Sandy Springs Station

Vine City Station

West End Station

West Lake Station

Not Sure