Aalto University
School of Business

PREDICTING REAL ESTATE SALES VOLUME IN FINLAND

Building a Predictive Model for The Sales Volume of Old Apartments

Master's Thesis
Tuukka Pesonen
Aalto University School of Business
Information and Service Management
Spring 2018

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Author Tuukka Pesonen

Title of thesis PREDICTING REAL ESTATE SALES VOLUME IN FINLAND

Degree Master of Science in Economics and Business Administration

Degree programme Information and Service Management

Thesis advisor(s) Pekka Malo

Year of approval 2018 **Number of pages** 97 **Language** English

Abstract

The aim of this Master's Thesis is to find an optimal set of explanatory variables affecting the real estate market in order to build a robust and accurate predictive model that forecasts the development of the real estate sales volume for the next 12 months. In more detail, this research examines the prior literature concerning the factors affecting the real estate market and predictive models based on which the initial variable set is constructed and the model is built. Two interviews are conducted interviewing industry experts in order to gain deeper knowledge of the field.

The research aims to answer the following research questions: *(1) What factors/input variables to involve when predicting the real estate sales volume, more accurately the sales volume of old apartments, in Finland, (2) What modelling method will give the best result when predicting real estate sales for the next 12 months given the nature of the data and (3) How does the sales volume of old apartments differ based on the apartment's location and type.* Thus, the research tries to build a robust predictive model that can predict the number of old apartments sold in Finland for the next 12 months as accurately as possible. This research is conducted as a both quantitative and qualitative study. In order to connect the results of this study to the existing literature and theoretical framework, five hypotheses were created. The hypotheses in order: (H1), *the number of sold old apartments in total will increase within the next 12 months, (H2) the sales volume for old apartments will increase more in the capital region (Helsinki, Espoo and Vantaa) than in other regions, (H3) the sales volume for smaller studio apartments will increase more than for other apartment types, (H4) the economic variables have the biggest impact on the number of house sold and (H5) search query data from Google Trends enhances the model and serves as an important predictor variable.*

Four models were created to predict the sales volume. Poisson regression and Negative Binomial regression were chosen as the modelling methods given that the response variable represented count data. Based on the results Negative Binomial regression model using predictor variables from Lasso variable selection was the best model as it had the best goodness of fit and thus the best prediction accuracy. Based on the forecasts it seems that the total sales volume of old apartments will increase overall within the next 12 months regardless of the location or type. The growth will be strongest in the capital region followed by Tampere and Turku. Variables related to economy or finance seems to be the most important ones in terms of predicting the sales volume of apartments.

Keywords real estate factors, Finnish real estate market, count data, Poisson regression, Negative Binomial regression, GLM, predicting sales volume

Aalto-yliopisto
Kauppakorkeakoulu

Aalto-yliopisto, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Maisterintutkinnon tutkielman tiivistelmä

Tekijä Tuukka Pesonen

Työn nimi ASUNTOJEN KAUPPAMÄÄRÄN ENNUSTAMINEN SUOMESSA

Tutkinto Kauppatieteiden maisteri

Koulutusohjelma Tieto- ja palvelujohtaminen

Työn ohjaaja(t) Pekka Malo

Hyväksymisvuosi 2018 **Sivumäärä** 97 **Kieli** Englanti

Tiivistelmä

Tämän maisterityön tarkoituksena on löytää paras mahdollinen määrä Suomen kiinteistömarkkinoilla vaikuttavia riippumattomia muuttujia sekä rakentaa niiden pohjalta robusti ja tarkka tilastollinen malli vanhojen asuntojen kauppamäärän ennustamiseksi seuraavan 12 kuukauden aikana. Tarkemmin sanottuna tässä tutkimuksessa perehdytään aikaisempaan kirjallisuuteen liittyen kiinteistömarkkinnoilla vaikuttaviin tekijöihin, joita käytetään alustavan muuttujasetin valitsemisessa tilastollista mallinnusta ja ennustamista varten. Alan tuntemuksen lisäämiseksi, kahta alan asiantuntijaa on myös haastateltu.

Tämä tutkimus pyrkii vastaamaan seuraaviin tutkimuskysymyksiin: *(1) Mitä riippumattomia muuttujia tai tekijöitä pitää ottaa huomioon, kun ennustetaan vanhojen asuntojen kauppamäärää Suomessa, (2) Mikä mallintamisen lähestymistapa toimii parhaiten, kun ennustetaan asuntojen kauppamäärää ottaen huomioon datan luonteen ja (3) Miten vanhojen asuntojen kauppamäärä eroaa asunnon sijainnin tai tyypin perusteella*. Tutkimuksessa käytetään sekä kvantitatiivisia että kvalitatiivissa tutkimusmenetelmiä. Viisi hypoteesia pyrkii yhdistämään tutkimustulokset aikaisempaan kirjallisuuteen sekä teoreettiseen viitekehykseen: *(H1) vanhojen asuntojen kauppamäärä kokonaisuudessaan kasvaa seuraavan 12 kuukauden aikana, (H2) vanhojen asuntojen kauppamäärä kokonaisuudessaan kasvaa enemmän pääkaupunkiseudulla (Helsinki, Espoo, Vantaa) kuin muilla alueilla, (H3) vanhojen yksiöiden myyntimäärä kasvaa enemmän kuin muiden asuntotyyppien myyntimäärä, (H4) talouteen liittyvillä muuttujilla on suurin vaikutus asuntojen kauppamäärään ja (H5) Google Trendin tilastotiedot Googlen kautta tehdyistä hauista parantavat mallin suorituskykyä sekä ovat tärkeä selittävä muuttujaryhmä ennustettaessa asuntojen kauppamäärää.*

Neljä tilastollista mallia rakennettiin ennustamaan vanhojen asuntojen kauppamäärää. Poisson regressio sekä Negatiivinen binomi regressio valittiin mallintamisen tavoiksi, koska ennustettava muuttuja sisälsi ei-negatiivia kokonaislukuja. Tulosten perusteella Negatiivinen binomi regressio, jossa riippumattomina selittävinä ennustajina käytettiin Lasso regression muuttujavalinnan jälkeen jäljelle jääneitä muuttujia, on paras malli yhteensopivuus sekä ennustustarkkuus huomioon ottaen. Tutkimuksen ennusteiden perusteella vanhojen asuntojen kauppamäärä näyttäisi kasvavan seuraavan 12 kuukauden aikana riippumatta asunnon sijainnista tai tyypistä. Kasvu on kovinta pääkaupunkiseudulla, Tampereella sekä Turussa. Yleiseen tai asuntokuntien talouteen liittyvät muuttujat olivat tutkimustulosten perusteella tärkeimpiä selittäviä ennustajia.

Avainsanat kiinteistömarkkinoiden tekijät, Suomen kiinteistömarkkinat, Poisson regressio, negatiivinen binomi regressio, myyntimäärän ennustaminen

Acknowledgements

I want to thank my thesis supervisor Pekka Malo, who has helped me with this project and guided along the way. In addition, I want to thank Tommi Vilkamo and Jarkko Hänninen from my workplace, who have supported me whenever needed and answered my questions whether they were industry specific or related to technical execution or statistics.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Research Problem

Real estate market has always been one of the most important field of studies as it is highly correlated with the current state as well as the changes in the economy. According to Straszheim (1975) real estate market has an important role when developing the city infrastructure, planning household's budget and building the overall standard of living. Many studies try to explain the changes in the housing prices in terms of their impact and correlation with the economy. Gottlieb (1976) studied the changes in the real estate prices and building constructions and found out that they are correlated with the swings in economic activity. In their study, Campbell & Cocco (2007) also stated that housing price changes are correlated with the changes in household consumption thus affecting the economy and vice versa. Sirmans & Turnbull (1997) proved that real estate agent's commission fees vary based on the economic conditions. Salo (2009) also studied the changes in housing prices and the impact it has on household spending. According to her, a one per cent increase in property value increases the consumption by 0,03 to 0,12 per cent. In fact, Case et al (2005) proved that changes in dwelling prices have a more significant impact on US consumption than the stock market has. This was later solidified to be true for other countries as well (Reinhart & Rogoff, 2009).

This study will try to predict the real estate market, more accurately the real estate completions by building a predictive model from a reliable and robust variable set instead. Hence the research topic is closely related to data science and predictive analytics. All in all, one could say that the research topic has elements from the fields such as real estate management, econometrics and data science. The motivation behind the target variable comes from the nature of the real estate broker business. As explained by the CFO of Kiinteistömaailma, the profits and revenues of a real estate agency are directly related to the number of houses they sell. A real estate agent receives a commission fee for houses they sell. The commission does not always follow the changes in the housing price. Therefore, the sales volume of houses in housing units is a more important measure than the housing price when looking from the perspective of a real estate agency.

Many prior studies have tried to forecast the development in the future housing prices with different modelling methods. Brown et al (1997) predicted the housing prices in UK using a Time Varying Coefficient regression. Kusan et al (2009) predicted the housing prices in Eskisehir, Turkey, using fuzzy training and logic systems. Kain & Quigley (1970) developed the frequently used hedonic model for predicting the house prices that has been modified further based on the need. Ottensmann et al (2008) built a hedonic regression model to forecast the housing prices taking into account the location of the apartment. Chica-Olmo et al (2013) created a multi-equational hedonic regression model taking into account the coregionalized disturbance and heterotopic data.

Although hedonic regression seems to have been the most used method to forecast house prices, there have also been some other approaches. Limsombunchao (2004) studied the artificial neural network as a model for predicting housing prices in Christchurch, New Zealand, and compared it to the traditional hedonic model. Furthermore, Ng (2015) built a mobile application that forecasts the future dwelling prices in London. Although testing several modelling techniques such as linear regression, Bayesian linear regression and relevance vector machines, Gaussian regression was chosen for the best model. (Ng, 2015)

As discussed there have been a lot of prior academic and empirical studies explaining or predicting the housing prices but almost close to none predicting the number of houses sold. Because of the lacking research, there is not a clear evaluation or consensus which modelling method or approach is the most accurate one when predicting the real estate completions, thus the number of houses sold.

In addition to build a robust model for predicting the sales volume of houses, there is a need to assess more carefully what kind of factors affect the real estate market. As one can guess the real estate market is affected by various different factors. Although there are prior researches related to the determinants affecting the industry, that will be presented in the literature review, there has not been deeper statistical analysis or ranking of these variables in terms of their importance. In addition, the existing literature is mostly highlighting factors affecting the real estate prices not directly the house sales volume. Moreover, the rationale or the analysis behind the predictors or input variables is not usually emphasized but they are rather chosen based on author's opinions, experience or availability of the data.

The lack of Finnish research is one of the biggest research problems. There have not been any prior academic papers studying the factors affecting the development of real estate sales volumes in Finland. In fact, most of the studies have focused on predicting the housing prices in a specific country other than Finland such as some of the studies discussed previously. Most of the Finnish studies are also focusing on housing price instead of the sales volume and often explaining it rather than predicting. Kuosmanen (1997) forecasted prices in the Finnish house market by comparing econometric and time series models such as a demand model and ARIMA. Hannonen (2015) predicted the housing prices in Helsinki submarket by building a demand potential model. Takala (2016) studied the impact of buyer's residence on the selling price of an apartment. Brotherus (2011) used hedonic pricing to find out the factors affecting the price of an apartment in the capital region. Laine (2015) explored the correlation between the quality of public transportation and the housing prices in Tampere and Turku regions. Hence one can see that although there have been studies related to the Finnish house market and especially the formation of housing prices in Finland, there is a clear lack of academic studies predicting or explaining the sales volume of houses in Finland and whether location or the type of the real estate property affect it.

1.2 Research Objectives and Questions

The purpose of this research is to find an optimal set of explanatory variables affecting the real estate market in order to build a robust and accurate predictive model that forecasts the sales volume of old houses for the next 12 months. Hence, there are two primary research objectives. First objective is to build a robust predictive model for predicting real estate completions, meaning the sales volume of the houses.

The second objective is to choose the best set of input variables for the model that affect the target variable, that is the number of houses sold. The underlying reason behind this objective is to find academic and statistical evidence on what variables to include when forecasting the sales volume of houses. This, in turn, would give more credibility to the research and hopefully generate more accurate results in the long term.

The secondary research objective for the study is to distinguish the differences in sales volume based on the geographical location and the type of the apartment. In this study we

are exploring old row houses as well as one-room, two-room and three-room apartments. The term "old" refers to dwellings that have been built at least 10 years ago. The apartments that are taken into account are from the following Finnish cities: Helsinki, Espoo, Vantaa, Turku, Tampere, Oulu, Jyväskylä, Lahti and Kuopio.

Three research questions are formed based on the research objectives that this study is aiming to answer:

1) *What factors/input variables to involve when predicting the real estate sales volume, more accurately the sales volume of old apartments, in Finland?*

There are many factors and variables that affect the real estate market. Hence it is important to determine which variables to include when building a predictive model in order to generate the most accurate and best results as possible.

2) *What modeling method will give the best result when predicting the sales volume of old apartments for the next 12 months given the nature of the data?*

Although there has been literature about predicting housing prices, there is no existing academic research about forecasting the number of houses sold. Hence there is a need to evaluate what kind of modelling method or approach would give the most robust model and most accurate results in the long term.

3) *How does the sales volume of old apartments differ based on the apartment's location and type?*

We will try to distinguish the differences in sales volume based on the geographical location and the type of the apartment in order to gain even deeper knowledge of the future market.

1.3 Structure of the Thesis

This thesis has 6 main chapters that are briefly presented in the Table 1 below.

Table 1: Structure of the Thesis

Chapter	Topic	Description
1	Introduction	Introducing the research topic and problem as well as the research objectives and questions
2	Literature Review	Presenting and discussing prior studies and literature related to the real estate market factors, real estate market in Finland and predictive models
3	Methodology	Presenting the research and statistical methods used as well as the theoretical framework and hypotheses.
4	Findings and the Model	Exploring the data, performing statistical tests, building the models and presenting the in-sample and out-of-sample results.
5	Discussion and Analysis	Selecting the best model, testing the hypotheses and answering the research questions.
6	Managerial Implications and Limitations	Discussing managerial contributions, limitations of the study and similar future studies.

The first chapter is introduction to the topic of the study and presents the research problem, motivation of the study, research objectives and research questions. Second chapter presents and discusses the prior studies and literature related to the real estate market factors and characteristics, real estate market conditions and expectations in Finland and differences between predictive and explanatory models. Third chapter discusses the research methods and also briefly explains the statistical methods and tests used. It also includes the theoretical framework, a procedure and guideline used to build the predictive model given the topic and context of this study as well as the hypotheses created to better connect the existing literature discussed in the literature review to the results of the study. Fourth chapter first explores the data and performs some statistical tests in order to know what kind of data we are dealing with. Moreover, the models are built and their results, meaning the goodness of fit, are presented for both the in-sample and out-of-sample data. In the fifth chapter the best model is selected based on the results presented in the fourth chapter. Moreover, the hypotheses are then tested against the results of the chosen model. In addition, the research questions are answered. In the sixth and last chapter, the results and contributions of the thesis are discussed from the managerial perspective. Furthermore, the limitations of the study and suggestions for similar future studies are presented.

2 Literature Review

This chapter presents the most important literature related to our research. The topics can be roughly categorized into studies and papers related to real estate market characteristics and factors, real estate market conditions in Finland and characteristics of predictive models. Literature related to used statistical methods are discussed later in the methodology chapter.

2.1 Real Estate Market

2.1.1 The Common Characteristics

A housing market can be defined as a market in which either a direct or indirect transaction between a seller and a buyer of a house occurs. (Jones & Watkins, 2009) In this research housing market and real estate market are referring to the same market. The real estate market has its own specific characteristics that differ from other markets. Arnott (1987) highlights 10 characteristics typical to a house as a commodity.

1. Necessity

   The need to find a shelter is one of the basic human needs thus a necessity. Housing satisfies this need.

2. Importance

   House is often the most important item of consumption within a household.

3. Durability

   Houses are the most durable major commodities.

4. Spatial Fixity

   It is often really difficult or extremely costly to transport the house to a different location. Hence a house is almost always fixed to its location.

5. Indivisibility

   A house is typically sold as a whole unit not in fractions.

6. Heterogeneity

   A house typically has many different characteristics that may or may not correlate with each other.

7. Thinness of the market

   Housing units and household's characteristics are spares and often do not match.

8. Nonconvexities in production

   Production of new houses is often slow so there is a discontinued change in the market. In the short term housing supply is not flexible but demand might change quickly.

9. Information asymmetry

   Parties related often have different amount of information or are not aware of all of the characteristics.

10. Transaction costs

    Selling and buying a new house include many different costs not related to the price such as the search costs for searching the house, moving costs for moving and transactions costs paid to the agency.

According to Arnott (1987) there is also an eleventh characteristic that is "*the near-absence of relevant insurance and future market"***.** However, this characteristic is not taken into account in this research as it is more related to the financial industry and thus is out of the study scope.

Skurnik & Summa (1978) divide the real estate market into submarkets based on which apartments for buying or rental are distinguished from the apartment production market. The first one is affected by the potential and effective demand whereas the latter is influenced by the supplement effective demand. These terms are further discussed later.

2.1.2 Real Estate Market in Finland

The urbanization has continuously increased in Finland. According to KTI Finland (2017) around 69% of the population lived in the 14 largest towns in 2015. KTI Finland (2017) also forecasted that, by 2040, this number is 75% or more. It is also worth to mention that these 14 cities, including the areas surrounding them, currently make up to 75% of Finland's GDP. Moreover 70% of all jobs are located in these regions. The Table 2 below shows the numbers for the largest regions.

Table 2: Biggest City Regions in Finland (KTI Finland, 2017:18)

Significance of 14 biggest city regions in Finland, % of total

	Helsinki region	The regions of Tampere, Turku and Oulu	14 biggest city regions
Population (2015)	27.3	17.5	68.3
Jobs (2014)	32.0	17.5	71.7
Private-sector jobs (2014)	35.1	17.9	74.3
GDP (2013)	36.4	17.1	74.1
Research and development expenditure (2013)	46.8	28.6	89.7
Completed dwellings (2010-2015)	30.4	21.1	76.1

Source: Statistics Finland

Because of the urbanization, the demand for smaller and better located apartments has increased in contrast to larger apartments (KTI Finland, 2017). In fact, Wang et al (2015) highlighted that the rapid urbanization has also had a positive impact on housing demand in China. Vehviläinen (2016) studied the development of the Finnish real estate market between the years 2000 and 2016 and also emphasizes the significance of the urbanization in terms of affecting the market. Although moving to cities, people tend to find houses from city borders instead of the center. This is because of the high prices that continues to rise. As the location of the houses are further away from the city center the apartment's distance from public transport is now more crucial than before. Vehviläinen (2016) also found out that the supply of smaller apartments has increased in contrast to bigger apartments because of the changes in demographics such as the family size or young people's willingness to live alone. Moreover, the mortgage interest has also decreased although this has not had a significant positive effect on real estate's transaction volumes. Pellervon taloustutkimus (PTT) is an association that contributes to the relevant fields by conducting studies and analysis. According to their newest publication, the number of old house in Finland will be increasing in the near future regardless of the clear decrease since 2012. In fact, the sales volume was higher in 2016 compared to 2015. The authors emphasized that this proves that the real estate industry has been able to

adapt to the low economic growth through the sales volume of the apartments instead of housing prices. In 2016, especially the number of newly constructed apartments sold increased and the future outlook of construction predicts that the number of houses sold will increase also in the future. (Kekäläinen, Tähtinen & Vuori, 2017) Figure 1 presents the development of the sales volume for the old apartments.

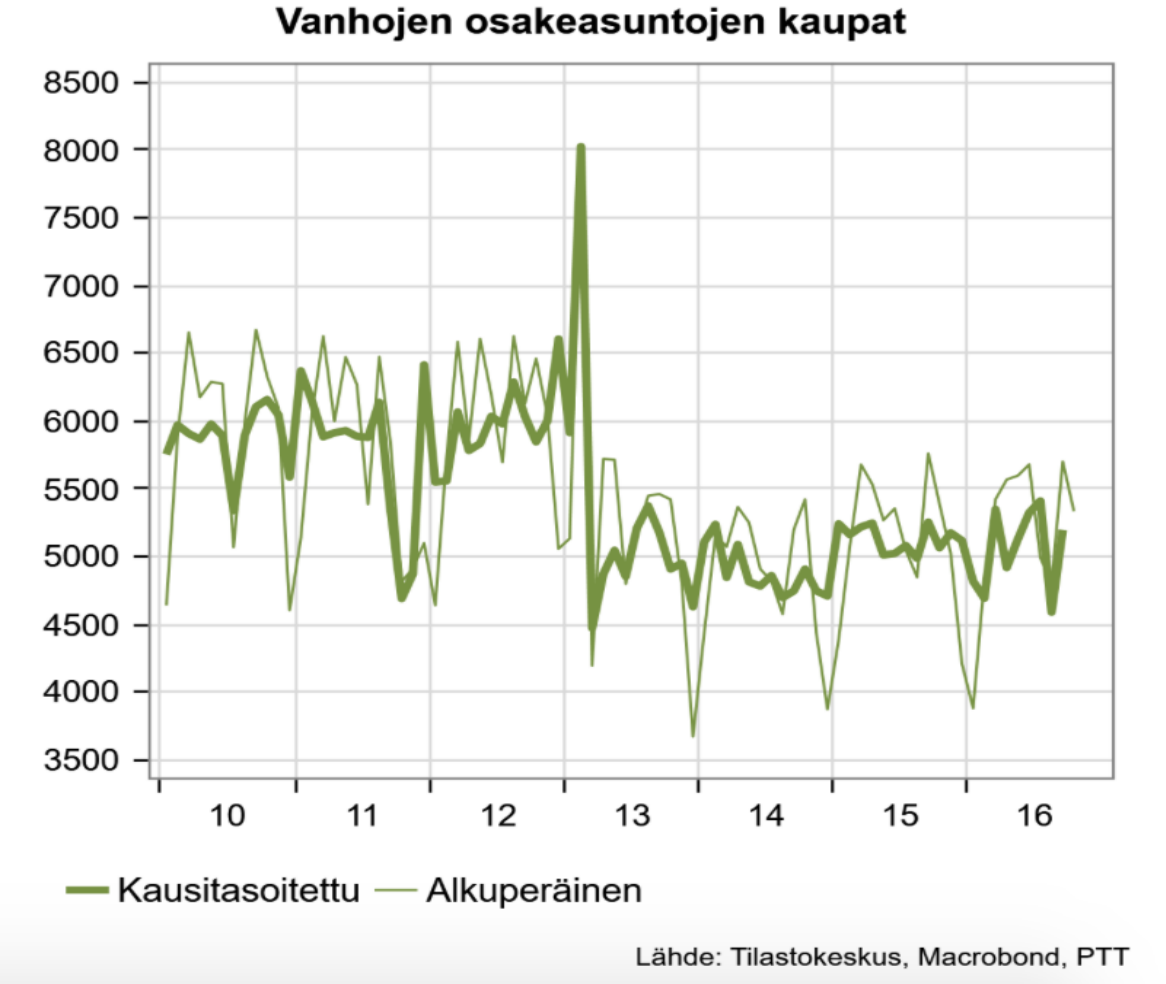


Figure 1. The Sales Volume of the Old Apartments.

Hypo, a Finnish credit institution specializing in housing, publishes a report about the outlook and conditions in the Finnish real estate market four times a year. Based on the newest report, published in November 2017, the sales volume of apartments has been increasing throughout the whole year and is expected to increase even more. The author, Brotherus (2017), mentioned that the growth has been strong in bigger cities, especially in Helsinki, Tampere and Turku but also in some middle-sized ones such as in Oulu and Jyväskylä. The institution's apartment index proved that the demand for old apartments has

increased in the capital region, Tampere and Turku although the demand for newly constructed apartments has been even higher. In fact, the sales volume of old apartments decreased in the early 2017. Brotherus (2017) also emphasized that there is a gap between the bigger cities and middle-sized ones where the sales volume has not been increasing as much. The urbanization has also been huge and for example young people tend to buy their first apartments mostly from the biggest cities. Brotherus (2017) pointed out that the capital region, Tampere and Turku cover now over half of the sales volume when it comes to the first apartment sales. Although the sales volume for big cities have been increasing overall, the capital region has distinguished itself and this gap is expected to grow even more in the future. Figure 2 presents the demand for apartments based on the Hypo's apartment index in the capital region, Tampere and Turku.

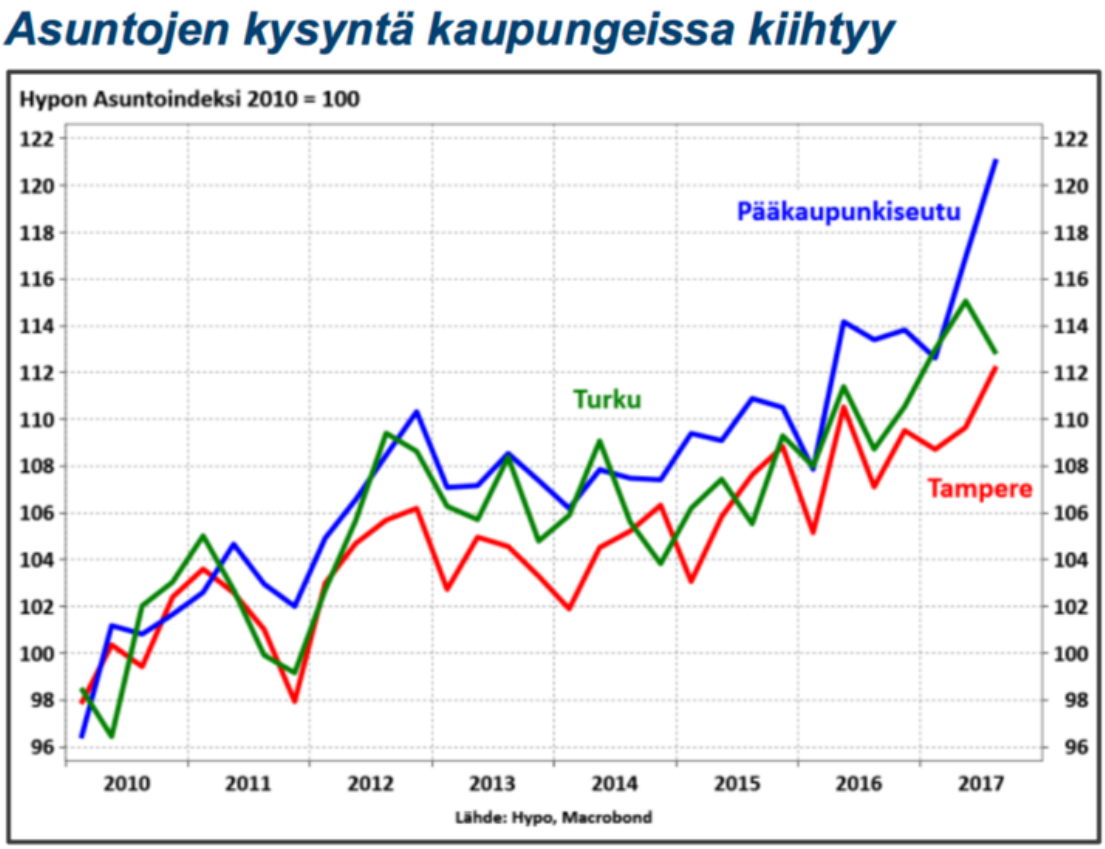


Figure 2. The Demand for Apartments in Helsinki, Tampere and Turku.

Brotherus (2017) also stressed that the average mortgage rate is currently as low as it has ever been, 1%. The author expected that the Euribor interest rate will be negative also in 2018 and that the European Central Bank will rise the interest rate moderately during the next two years.

2.2 Factors Affecting the Real Estate Market

2.2.1 Housing Supply and Demand

Mourouzi-Sivitanidou (2011) stressed that, although the urban real estate market differs from most other markets, it still follows the same basic economic principle of supply and demand. Hence, the following sections present the different factors affecting the housing demand and supply based on the prior studies and literature.

2.2.1.1 Housing Demand

Mourouzi-Sivitanidou (2011) defined housing demand as follows: "the quantity of space or number of units demanded at various prices". Although demand is often typically highly correlated with the price of the unit, real estate demand is also affected by other factors. According to Arnott (1987), housing demand depends also on the buyer's income level, location of the apartment and apartment's distance from work as well as demographic characteristics such as the number of family members. In fact, according to Mourouzi-Sivitanidou, real estate demand factors can be divided in two different categories:1) Endogenous determinants and 2) Exogenous determinants. An endogenous factor for real estate demand is the actual price or rent of the housing unit. Exogenous determinants, in contrast, are the non-price related factors that affect the changes and development of the real estate demand. These determinants can be further divided into four subcategories (Mourouzi-Sivitanidou, 2011):

1. Market Size

   Market size factors include employment and population among others. They are variables that typically have a positive effect on real estate demand. The impact of different market size factors often depends on the target variable. For example, the number of households is more significant factor when predicting the housing

demand whereas the size of office employment is better factor when measuring the demand for office spaces.

2. Income/Wealth

   Income level of the household has a direct and positive effect on the real estate demand.

3. Price of Substitutes

   Although related to the price or rent determinant that is an endogenous factor, price of substitutes does not consider the real estate market as a whole but based on different housing alternatives. For instance, if the prices for rent apartments increase when the prices for non-rental apartments of the same size are staying constant, this may increase the demand of non-rental apartments instead of rental ones.

4. Expectations

   Consumers or firms' overall expectations about future whether it is related to the economy or income level, real estate prices or company's own growth possibilities among other alternatives.

Siikanen (1992) highlighted other important non-price related factors affecting the real estate demand such as the society's subsidy, living preferences and the price of mortgage. Demographic factors such as migration, fertility and the willingness to start a family are also among the important determinants. Jansen et al (2011) have gathered a set of variables to take into account when choosing the predictors for housing demand. In addition to the sales price, the demand is affected by people's relocation behavior that is the number of moves to new houses. Relocation behavior, in turn, is partly influenced by the demographic and social-economic variables, supply of the houses and price changes in the real estate market as well as in the mortgage market. This set include explanatory variables such as growth of the population within an age category, average income of households, inflation, average house or rental prices, number of newly built houses and changes in the interest rate. (Jansen et al, 2011). However, Jansen et al (2011) also pointed out that some of these factors might have a delay effect on the demand. Hence the authors suggest time-series and lagged variables as an option for modelling.

As mentioned before the real estate market can be divided into submarkets. Similarly, the demand for real estate can divided into three groups: 1) Potential demand, 2) Effective demand and 3) Supplemental effective demand. (Skurnik & Summa, 1978) Skurnik & Summa (1978) defined potential demand as demand that is quite easy to predict. Its factors include household income level, relative house prices and demographics. Effective demand is demand that is relative to economic resources. There cannot be effective demand without potential demand. In other words, households have the economic resources to actually buy the house in addition to their willingness to purchase it. These two types of demand are the factors that affect the market for rental and non-rental apartments. In contrast, supplemental effective demand refers to the demand that the existing housing supply cannot meet and often leads to building more new apartments. (Skurnik & Summa, 1978)

Ostamo (1997) has categorized the factors affecting the housing demand. First category is the changes in population that include factors such as the population structure, migration, household structure and population growth. According to Ostamo (1997) the second category of housing demand is the income and employment level as well as changes in these factors. Third category is the living costs that include factors such as the development of price and rent levels, price and availability of mortgage, interest rate level, taxes and housing allowance. Fourth category is the consumers' housing preferences that is what kind of apartment they are looking for and why. The fifth and last category is the housing politics and legislation. Heiskanen (2008) also emphasized politics and demographics' importance as factors affecting the housing demand and thus the real estate market.

Tiwari (2000) studied the housing choices and demand in Tokyo. Similarly, to Mourouzi-Sivitanidou (2011), the author also stressed the substitution of the price as a determinant for the demand. For instance, a large increase in a rental house's price when the price of an ownership house is decreasing may increase the demand for the ownership house compared to the rental one. Tiwari (2000) also listed culture, policy and demographic behavior as determinants of housing in Tokyo. For example, households tend to live in rental apartments when they are younger but move to ownership apartments as they get older and the size of the household increases. As already highlighted above, household's income level seems to be one of the key variables affecting the housing demand. Tiwari

(2000) also found out that the income elasticity differs between the ownership and rental houses. The first one has a positive income elasticity whereas the latter has a negative one. Therefore, people tend to purchase a house instead of renting it when their income level rises.

Pirounakis (2013) also created a comprehensive set of determinants for housing demand and supply based on prior literature and studies. These factors are mostly same as the ones discussed before but also include some new ones. For instance, cost of dwelling maintenance is distinguished as a separate determinant from other costs such as transportation costs. Moreover, the expected rise in the house price is also emphasized as a separate determinant. The factors are presented in the Table 3 below.

Table 3: Determinants of Housing Demand and Supply (Pirounakis, 2013:212)

Demand determinant	*Supply determinant*
1 Current income	1 Land availability (which depends on land ownership pattern and the laxity or otherwise of the planning function)
2 Expected (permanent) income	2 Expected price at time of completion (in practice, a function of past prices, as it typically takes one to two years to construct a house or apartment building).
3 Proximity to work	3 National or area-specific zoning and building regulations
4 Proximity to amenities (social and environmental)	4 Cost of construction (labour and materials; current borrowing cost; insurance and regulatory costs
5 Transport costs (monetary and time-related)	5 Cost of land (the developer-paid purchase price of land, which is roughly the difference between expected revenue from the development and the sum of construction cost and the developer's required return; also, the opportunity cost of keeping the land undeveloped)
6 Dwelling characteristics (other than location), e.g., house or flat; size; etc.	6 Developer's required rate of return
7 Tenure, expressed as a tenure-choice factor	7 Taxes (like VAT on newly constructed dwellings)
8 Mortgage interest rates, and other loan terms (relevant to those considering owner-occupation)	8 Building technology
9 Cost of dwelling maintenance	9 Long-term real interest rates (Levin and Pryce, 2009
10 Expected price appreciation (relevant to current or prospective owner-occupiers)	
11 Demographic factors (rate of population growth, rate of household formation, age distribution of population, dependency ratio*)	
12 Non-housing wealth of households	
13 Social characteristics of households, including educational level	
14 Saving interest rates, as they affect the speed with which households can save towards house purchase	
15 Capital-gains tax, to the extent that capital gains matter to home-owners	
16 Property taxes: if, e.g., owner-occupation is heavily taxed, then renting may be an alternative but only if there is enough rented accommodation in the area; otherwise, people may decide o stay put (i.e., share with relatives)	

* Dependency ratio = typically, the number of people younger than 15 and older than 64 divided by the number of people aged 15–64

2.2.1.2 Housing Supply

Housing supply can be defined as "the quantity of commercial space or housing units supplied at various prices." (Mourouzi-Sivitanidou, 2011) Mourouzi-Sivitanidou (2011) emphasized the differences within the housing supply. The real estate supply can be either aggregated in the long-run, aggregated in the short-run or based on the new construction. When long-run supply describes the supply over a long period of time, the short-run supply represents the supply at the given point of time and is often better for the market analysis as it takes better into account the construction lag that is how much time it is required to plan or develop a building. However, Mourouzi-Sivitanidou (2011) also argued that new construction is the most crucial type of housing supply when doing market analysis and when it comes to understanding the development of the market over time. Although new construction could be the most important factor, Dipasquale (1999) states that the supply is also affected by the house owners and agents' decisions about the existing stock. As the objective of this study is to predict the real estate completions of old existing apartments, new constructions will not have a huge weight as a predictor in our model.

Arnott (1987) pointed out that housing supply can be described as four different processes: 1) Construction, 2) Maintenance, 3) Rehabilitation and 4) Conversion. Construction is straightforward and means producing housing units given the required land and capital to do it. Maintenance is the process of slowing down the deterioration of the house's quality also requiring capital. In contrast, rehabilitation is maintenance but with sudden use of the capital. Conversion, that can be either downward or upward, is the process of changing the size of the house that is often done discontinuously (Arnott, 1987). Jansen et al (2011) highlighted that housing supply is also a side effect of the consumer preferences and decisions. In fact, the authors stated that housing market itself present some opportunities and constraints that affect individual's search for a house. For example, buyer's income level can be seen as both an opportunity or a constraint similarly to the distance from school or workplace.

Housing supply is typically non-flexible in the short-run as adapting and meeting the demand in the market is usually slow (Skurnik & Summa, 1978). Lönnqvist & Vaattovaara (2004) divided housing supply into two categories: 1) Reserve and 2) Flow. Reserve refers to the existing housing stock whereas flow means the new construction. Siikanen & Tyrkkö (1993) also stated that although new construction is one of the main factors

affecting the housing supply it only covers roughly just a couple of percent of the whole housing stock. Hence one could say that the housing stock is almost entirely represented by the existing stock, reserve, that is increased and affected by basic improvements and expansions as well as deployment of empty apartments.

2.2.2 Internet Search Queries

There has been a lot of research showing that data related to internet search queries improves the forecasts of predictive models (Norros, 2014). Norros (2014) pointed out that paying for the apartment is a long process. Therefore, a rationale buyer always gathers as much information as possible before making the purchase. Brynjolfsson & Wu (2009) also estimated that 90 percent of the transaction parties, whether it is a seller or a buyer of the house, use internet search queries to support their decision making. Moreover, internet search queries are the primary information source for 80 percent of the house buyers (Hohenstatt & Kauesbauer, 2014).

The most robust and accurate real estate market forecasts are based on government and other institution's official and published reports. However, these publications are often delayed and published months after the actual event (Choi & Varian, 2009). Therefore, Brynjolfsson & Wu (2014) also emphasized that using only these kind of reports is not the most optimal solution when predicting future development as the delay might skew the results. Norros (2014) also stressed that, based on all the prior studies, the statistical and predictive models' accuracy can be improved by combining real-time data concerning consumer's online search behavior and actual data about the past development of the real estate market. Asiktas & Zimmermann (2011) pointed out that internet search usually offers consumers an anonymity that they otherwise could not have. Therefore, internet search query data often better represent consumer's real willingness to purchase as it is not affected by the external factors as much. For instance, the social pressure might affect respondent's answers when filling a questionnaire survey on the spot.

Choi & Varian's (2009) study was the first to demonstrate the relationship between google search queries and the real estate market. Predicting the sales of apartments, retail, motor vehicles and trips, squared R for every model improved when using search query data. Norros (2017) studied the nowcasting of the Finnish Real estate market, more accurately the housing price index and transaction volumes using internet search queries. To do this

the author added a search query data variable, Home Financing as the search word, to the baseline model. The results showed that the adjusted squared R increased for both models, 6,8 percent for the housing price index model and 2,4 percent for the transaction volume model. Furthermore, the mean-squared error decreased for both models, 5,9 percent for the housing price index model and 5,5 percent for the transaction volumes model. Although Norros (2017) nowcasted the Finnish real estate market instead of forecasting it, the results can be applied to this study. For example, Bartlett (2017) defined nowcasting as "forecasting an economic variable in the very near term by incorporating data of varying frequencies or timeless as well as data from official and unofficial sources". According to the author, nowcasting has increased in popularity in recent years. Norros (2017) also emphasized that internet search queries can improve the models estimating both the current situation as well as the near future.

2.3 Predictive Models

Shmueli (2010) defined predictive modeling as "the process of applying statistical model or data mining algorithm to data for the purpose of predicting new or future observations". In other words, a predictive model is any model generating forecasts despite the method or algorithm used. In contrast, explanatory models aim to explain the causality and relationship between the independent variables and the dependent variable. (Shmueli, 2010)

Shmueli (2010) presented four major aspects that need to be considered whenever choosing the method for either a predictive or explanatory model.

1. Causation-Association

   Explanatory models aim to model an underlying causal function in which a dependent variable Y is caused by an independent variable X. In predictive models this function is the association between the input and target variable instead of the causality.

2. Theory-Data

   Explanatory models always require that the function is built supporting the test hypotheses and the estimated relationship between the independent variables and

the dependent variable. In contrast, predictive models do not require direct interpretability and are often built form the data. However, this does not mean that the transparency or interpretability of the model should be ignored but it is less important in terms of predicting accuracy.

3. Retrospective-Prospective

   Predictive model is looking forward, in other words predicting new observations whereas explanatory model is prospective. It tests existing hypotheses determined beforehand.

4. Bias-Variance

   Bias refers to the outcome of misspecifiying the statistical model. While an explanatory model aims to minimize the bias, a predictive model stresses on minimizing the combination of bias and estimation variance. Hence, predictive models sometimes sacrifice theoretical accuracy for empirical one. In other words, a wrong model can forecast better than a correct one.

Given these four aspects and the interpretability of the explanatory model Shmueli (2010) stated that algorithmic methods are not suitable for explanatory modeling. In contrast, a predictive model can apply both statistical models and data mining algorithms. However, all methods are not suitable taken into account the retrospective-prospective aspect, especially when dealing with time series. Shmueli (2010) also introduced two methods, based on the bias-variance aspect, that can be exploited when predicting but not in explanatory models: 1) Shrinkage methods and 2) Ensemble methods. The first one refers to methods such as ridge or lasso regression that is methods shrinking and eliminating the predictors to decrease estimation variance. The second one, ensemble methods, refers to combining multiple models such as random forests and boosting.

The modeling approach and method should always be chosen based on the data and especially its characteristics. Hence the characteristics of the data should be taken into account also when choosing the modeling technique in this research. As mentioned before there are countless of different methods and techniques to use in the predictive model. These machine learning techniques can be classified roughly into four different approaches based on the target and input variables of the model as well as the aim of the modelling: 1)

Supervised learning, 2) Unsupervised learning, 3) Semi-supervised learning and 4) Reinforcement learning. Bronwlee (2016) defined supervised learning as predicting the output variables by using an algorithm or statistical model to approximate the mapping function. In other words, the model or machine learns to predict the output from the input data that is labeled beforehand. In contrast, unsupervised learning refers to predicting without the output variable. The model aims to find underlying patterns, structures or distribution in the data without teaching it. Moreover, the data is not labeled beforehand as it is when using supervised learning. Semi-supervised learning is a mixture of both techniques where some of the data is labeled and the rest is not. Reinforcement learning is the newest technique that similarly to the supervised learning aims to predict output variables based on the input data by approximating the mapping function. However, there is a reward function that the machine learns from and that the supervised technique lacks (Shaikh, 2017). The Figure 3 below presents the four different groups with further categories.

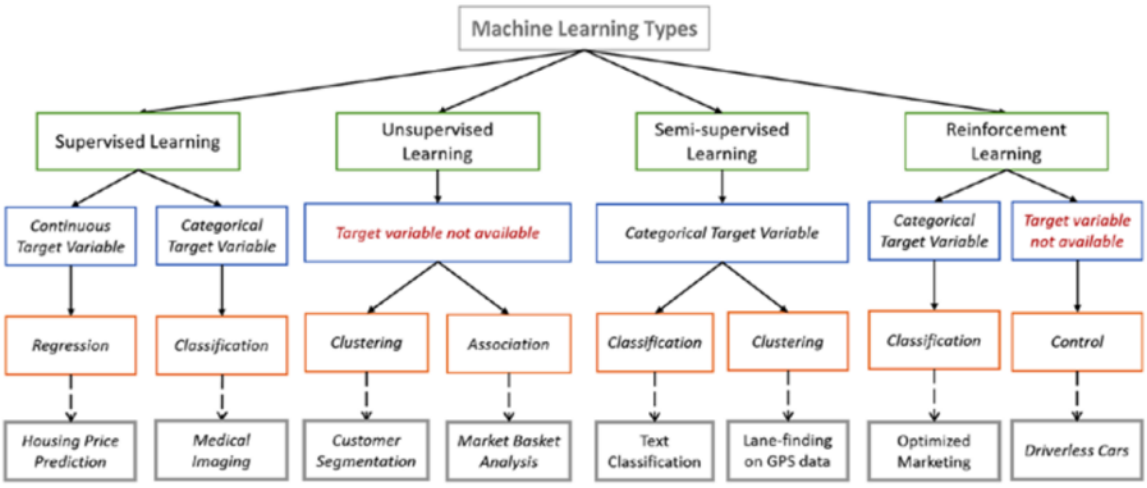


Figure 3. Machine Learning Types (ProftMe, 2015).

When choosing the correct modelling approach, we have to consider the nature of the independent variables, research questions and the target variable. As we can see from the Figure 3 housing prices are predicted using regression as a form of supervised learning. Although our response variable is not continuous, regression seems to be the best method given that predicting housing prices do not differ too much from predicting the number of houses sold in terms of the industry and possible predictors. Although the housing sales

volume could also be predicted using a different method or even a machine learning algorithm, regression is usually the most straightforward technique with results easy to interpret. To be more accurate, our target variable, the number of houses sold, can be classified as count data or a discrete variable. Hence Poisson regression is going to be the baseline method to which other methods are compared. In addition, we will apply a shrinkage method such as lasso or ridge in order to reduce the number of predictors in the model. Moreover, elastic net is also considered as an alternative. In the end another regression approach might be taken after performing statistical tests and testing assumptions. However, machine learning algorithms are left out in order to limit the scope of the data. All of the statistical methods will be further discussed in the methodology chapter.

3 Methodology

This chapter presents the methodological techniques and methods used in this research to answer the research questions. The research primarily used a quantitative approach that means statistical methods to choose the right input variable set for the model and the correct modeling technique for predicting. However, a qualitative method was also used in terms of semi-structured interviews that allowed to gain more perspective and information about the real estate industry as well as the possible predictor variables to include. The following sections presents the different statistical techniques used in the research as well as the collection of the data and other methodological aspects. Moreover, qualitative method that is the semi-structured interviews are covered briefly.

3.1 Data Collection

Both quantitative and qualitative methods were used for data collection. However, qualitative method, thus the semi-structured interviews, was mainly supportive and served as a basis for collecting the right data and variables for the analysis and the model. Hence one could say that the primary data for this research is the quantitative data whereas the interviews and the data gathered from the prior academic literature are secondary in terms of only supporting the collection of the primary data.

3.1.1 Quantitative Data

The quantitative data was collected based on the literature review and semi-structured interviews. The aim was to collect the best data available that represent the input variables for the models as well as the target variable. For this three major sources of data were used: 1) Statistics Finland (Tilastokeskus), 2) Kiinteistövälitysalan keskusliiton hintaseurantapalvelu (HSP) and 3) Google Trends. Moreover, Research Institute of the Finnish Economy (ETLA) provided data related to Finland's gross domestic product (GDP).

Statistics Finland is a Finnish public authority established for statistics. Currently employing over 800 industry experts it publishes the majority of the official statistics in Finland. In this research, most of the data including the target variable was collected from Statistics Finland's publications. Google Trends is a Google's service that offers information related to Google search-queries in real-time (Norros, 2014). In addition to be

able to select data based on the geographical location and time on a higher level, the service also allows to examine the search queries of a certain specific city and on a daily basis (Chamberlin, 2010). As millions of people are using Google, the service allows to assess the consumer behavior and search queries very precisely and thus enhance the predictions and their timeliness in the other context (Brynjolfsson & Wu, 2014). Tuhkuri (2014) also pointed out that the database of Google Trends is very broad as there are more than 3 million Google search queries daily worldwide. Google Trends does not actually represent the actual number of searches for a specific word or sentence but the ratio of the search word compared to the overall volume of search queries with the same selections and criteria (Tuhkuri, 2014). For this research, Google Trends was used to collect search query data based on the most important keywords correlating with the real estate sales volume found out in the semi-structured interviews. These search terms included 1) Apartments for Sale (Myytävät asunnot), 2) Brokerage (Kiinteistönvälitys) and 3) Real estate broker (Kiinteistönvälittäjä). The search terms were in Finnish as we can assume that the majority of Finnish consumers write their search queries in Finnish especially when it comes to buying or selling a house in Finland.

3.1.2 Semi-structured Interviews

In order to support the selection of the model variables there was a need to interview industry experts and get their professional opinions. Hence two interviews were conducted. The interviews were semi-structured so that the questions were just leading the conversation to right direction. The two interviewees were the CFO and CEO of Kiinteistömaailma who both have had a long career in the real estate industry.

The questions were divided into 5 category based on the prior literature. These categories were: 1) Locations and type of the housing unit, 2) Overall economy and financing of the household, 3) Finland versus the rest of the world, 4) Demographics and 5) The most important factors of the industry. The categories were also introduced and discussed in the same order as above. This allowed the interviewee to think about all the possible factors before ranking them in the end. First, second and fourth categories were straightforward and directly based on the existing literature and studies. First one consisted of questions related to the importance of the housing location or its type, second one related to the influence of the current or future state of the economy and household's financing and fourth one related to the buyer and other demographic aspects. Third one, Finland versus

the rest of the world, was just to highlight whether there are differences in the Finnish real estate industry and its determinants compared to the other countries as this should have been taken into account when building the model. The interview questions are in the appendix.

3.2 Research Variables

The target variable also known as the dependent variable was the number of sold houses (lukumäärä) that comes from the Statistics Finland dataset. The other variables, independent or explanatory, are the ones explaining or predicting the target variable. Table 4 presents the most important research variables chosen based on the prior literature. It is worth of mentioning that some of these variables have been excluded from the model after performing statistical tests and exploring them more closely.

3.2.1 Dependent Variables

Number of sold old apartments (lukumäärä)

The dependent, hence the target variable was the number of sold old apartments. The number differs based on the city, year and year quarter. The data type is count data as the number of sold houses can only be a non-negative integer. The apartments are labeled as old that is it has been at least 10 years since they have been built.

3.2.2 Independent Variables

Independent variables are the variables explaining the target variable and eventually predicting it. Table 4 below presents the main explanatory variables initially included to the model. The data types vary from nominal and ratio variables to continuous variables and count data. Data sources are Research Institute of the Finnish Economy (ETLA), Google Trends, The Price Tracking Service of the Central Federation of Finnish Real Estate Agencies (HSP) and Statistics Finland (Tilastokeskus). As we can see from the Table 4, most of the variables are related to either overall economy, household finances or demographics. Many of the variables are also exactly same as presented in the literature review such as the average size of a household or the number of new started constructions.

Table 4: Initial Variable Set

Variable:	Description:	Data type:	Data source:
BKT volyymin vuosimuutos	The change per year in the gross domestic product of Finland	Ratio	ETLA
BKT 2010 hinnoin	Finland's GDP with 2010 price level	Continuous	ETLA
Myytävät asunnot haut	Index describing the search volume with the search term: "Myytävät asunnot"	Count data	Google Trends
Kiinteistönvälitys haut	Index describing the search volume with the search term: "Kiinteistönvälitys"	Count data	Google Trends
Kiinteistönvälittäjä haut	Index describing the search volume with the search term: "Kiinteistönvälittäjä"	Count data	Google Trends
HSP_kappaleet	The amount of sold houses in the HSP database.	Count data	HSP
HSP median myyntiaika	The median sales time of an apartment	Continuous	HSP
Asunnon ostoaikomus 12 kk sisällä	The respondent's intention to buy a house within 12 months	Continuous	Tilastokeskus
Kuluttajien luottamusindikaattori	The respondent's trust to the future and economy	Continuous	Tilastokeskus
Alue	City	Nominal (Categorical)	Tilastokeskus

Tyyppi	The apartment's type	Nominal (Categorical)	Tilastokeskus
Väkiluvun muutos edellisestä vuodesta	The change in population from last year	Ratio	Tilastokeskus
Neliövuokra	The rent for a squared meter. Takes into account the price of the substitute in the model.	Continuous	Tilastokeskus
Asuntokunnan kesikoko henkilöä	The average of members in a household	Continuous	Tilastokeskus
Työllisyysaste	Employment rate	Ratio	Tilastokeskus
Rakennusaloitukset	The number of started building constructions	Continuous	Tilastokeskus
Vuokra-asunnoissa asuvien asuntokuntien osuus	The ratio of households living in a rental apartment	Ratio	Tilastokeskus
Asuntokuntien lukumäärä	The number of households	Count data	Tilastokeskus
Väkiluku	The population in Finland	Count data	Tilastokeskus
Vuosi Vuosineljännes	Year Year quarter	Date	Tilastokeskus, HSP Google Trends ETLA

3.3 Modelling Methods and Statistical Tests

As discussed in the literature review, regression will be the main modelling approach given the type of the target variable and the input data. Moreover, regression is the most straightforward method to conduct and has several different sub-alternatives to choose from. The baseline model will be a Poisson multiple linear regression that will be further processed in terms of adding and deleting variables. As the number of independent variables is large, Lasso is used as a shrinkage method in order to generate a better and more robust model. Below the methods are presented briefly. As we are dealing with time series data it is important to also test if there are issues with autocorrelation or heteroscedasticity and if the data is stationary or not. Hence three tests, Augmented Dickey Fuller, Breusch-Pagan and Breusch-Godfrey are conducted. These tests are briefly explained in this chapter as well. First, the basics of the multiple linear regression are introduced.

3.3.1 Multiple Linear Regression

Bewick et al (2003) defined a linear regression as an equation that expresses the correlation, that is the strength of the relationship, between two variables. Multiple linear regression is the same as a simple linear regression but with more than two variables one of which is the dependent variable the others, independent variables, are explaining. Hurlin (2013) wrote a basic multiple linear regression equation as follows:

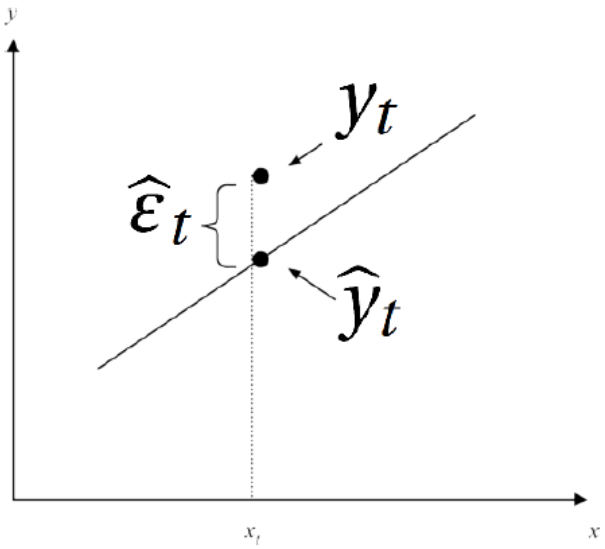
$$y = \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k + \varepsilon. \quad (1)$$

In the equation y represents the dependent and explained variable whereas x represents one of the independent variables explaining the dependent variable. The last term, ε, is the error term. β is the correlation coefficient in other words the strength of the relationship between the specific independent variable and the dependent variable. For instance, when β_1 of x_1 is 2, we can say that while x_1 increases by 1 unit, y increases by 2 units. Lof (2016) mentioned that typically the first dependent variable is normalized in order to get a constant term that is the intercept where the line crosses the y-axis. Hence the multiple linear regression model after normalizing looks as follows where β_1 is the intercept term:

$$y = \beta_1 + \beta_2 x_2 + .. + \beta_k x_k + \varepsilon. \tag{2}$$

Lof (2016) defined the purpose of the regression as explaining the variation in target variable y as a function of a dependent variable x. This can be seen as a linear regression line trying to fit the data observations. The residual, ε, is the difference between the fitted regression line and each of the observations. In order to enhance the variation explained by the independent variables, regression coefficients are chosen so that the line is as close to the actual observations as possible. This is done by minimizing the sum of squared residuals (RSS) that is the distance between the regression line and all of the observation values. In other words, RSS is the sum of the squared errors where errors refer to the difference between the predicted and actual values. RSS is defined as follows (Jain, 2016):

$$RSS = \sum_{t=1}^{T} \varepsilon_t^2. \tag{3}$$



Brooks (2008) defined the squared R (R2) as a statistic that evaluates how well the regression fits the data. In other words, R2 is a number between 0 and 1 representing the percentage of the variation in the dependent variable explained by the model or the independent variables. The equation for R2 is defined as follows (Lof, 2016):

$$R^2 = \frac{Var(y)}{Var(y)} = 1 - \frac{Var(\varepsilon)}{Var(y)}. \quad (4)$$

In this study the target variable will be the sales volume of old apartments. The initial set of independent variables is chosen based on the existing literature and the possible multicollinearity problems. R2 of the test set will be the best numerical indicator whether the model works for predicting or not.

3.3.2 Poisson Regression

Poisson regression is typically used with count data. Coxe et al (2009) defined count data as "the number of occurrences of a behavior in a fixed period of time". Count data can only include non-negative integers (Karazsia et al, 2008). Hence Poisson regression is a generalized linear regression model with a logarithm link function. Agresti (2013) defined generalized linear models (GLM) as "models extending the ordinary regression model to encompass non-normal response distributions and modeling function of the mean." Poisson regression as a generalized linear model has three major assumptions (Durrant, 2016; Rodriguez, 2009):

1. Target variable follows a Poisson distribution
2. Logarithm of the target variable's expected value can be modelled by unknown linear parameters
3. Mean of the distribution equals to the variance of the distribution.

Hence, Poisson regression formula can be expressed in two ways (Fan, 2016):

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + .. + \beta_k x_{ik}$$

or

$$\lambda_i = e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}}. \quad (5)$$

As all generalized linear models, Poisson regression has a random component representing the target variable and its distribution, a systematic component consisting of the explanatory variables in the linear function and a link function that is the expected value of the target variable equated to the linear function. Here the link function is the log link deriving from the Poisson distribution that we will not be discussing more in this study. (Agresti, 2013)

3.3.3 Lasso Regression

Introduced in 1996, Least Absolute Shrinkage and Selection Operator, also known as Lasso, was a new method for linear model estimation. The inventor Tibshirani (1996) described it as follows: "The Lasso minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant". In other words, Lasso is a regression shrinkage method typically used in models with large number of variables but relatively few observations. A large number of features may cause overfitting of the model thus making the regression coefficients, R2 and p-values misleading. Although overfitting does not affect the unbiasedness, a large number of variables also makes the model complex and computationally slow (Frost 2017; Jain, 2016; Lof, 2016). Hence Lasso regression's main purpose is to perform variable selection while fitting the regression line to the data. Similarly, to the Ridge regression, this is done by shrinking certain coefficients but in addition setting some of the coefficients also to zero. Lasso performs a L1 regularization by adding a penalty to the objective under optimization. This penalty is the sum of absolute value of coefficients and determines which coefficients to shrink and how much. The Lasso optimization objective is defined as follows:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|. \quad (6)$$

RSS is the residual sum of squares and lambda is the parameter deciding the weight on minimizing the RSS compared to the penalty term that is the sum of absolute value of coefficients. Jain (2016) generalized the scenarios for different values of lambda:

1. $\lambda = 0$: Same coefficients as in simple linear regression

2. $\lambda = \infty$: All coefficients zero
3. $0 < \lambda < \infty$: coefficients between 0 and that of simple linear regression

The penalty term's purpose is to penalize the model for overfitting. Lasso method aims to keep the good attributes of both Ridge regression and best subset selection methods by keeping as many variables as possible without the risk of overfitting or reduced prediction accuracy.

3.3.4 Augmented Dickey-Fuller Test

Augmented Dickey-Fuller (ADF) tests the stationarity of the data. Time series data assumes that the data is stationary that is its distribution shape does not change over time. ADF is often conducted in terms of lags. In other words, ADF adds lagged differences to the regression model and tests whether there is a unit root or not. The null hypothesis is that there is a unit root and thus the data is not stationary. If the p-value of the test is smaller than 0.05 we can reject the null hypothesis and say that the data is stationary. Hence when forecasting with time series data it must be stationary. (Glen, 2016; Zhang, 2015)

3.3.5 Breusch-Pagan Test

Breusch-Pagan test is a test for heteroscedasticity. Lof (2016) defined heteroscedasticity as a phenomenon in which the error variance varied across observations violating the assumption of a constant variance. The null hypothesis of the test is that there is no heteroscedasticity. If the t-value exceeds the critical value that is if the p-value is smaller than the previously determined significance level the null hypothesis can be rejected and there is heteroscedasticity. Lof (2016) also mentioned that sometimes log transforming either dependent or independent variables can reduce the heteroscedasticity. It is also worth to point out that only variables with strictly positive values can be log transformed.

3.3.6 Breusch-Godfrey Test

Breusch-Godfrey test is a test for autocorrelation of the error terms in the data. Correlation is normally calculated as follows:

$$Corr(x_t, y_t) = \frac{Cov(x_t, y_t)}{s.d.(x_t) \times s.d.(y_t)}. \tag{7}$$

So the correlation is the covariance of the independent and the dependent variable divided by the multiplication of the independent variable and dependent variables' standard deviations. The purpose of an econometric model is to describe a variable in terms of an equation consisting of the fitted part and the error term. The ordinary least squares (OLS) assumption is that the errors cannot be predicted and that they are uncorrelated. The formula for this assumption is defined as follows (Lof, 2016):

$$Cov(\varepsilon_t, \varepsilon_s) = 0. \tag{8}$$

Hence the autocorrelation of the data means that current values may help to predict future values. This will, for example, skew the traditional hypothesis testing based on regular standard errors and t-values. The mathematical formula for autocorrelation is defined as follows:

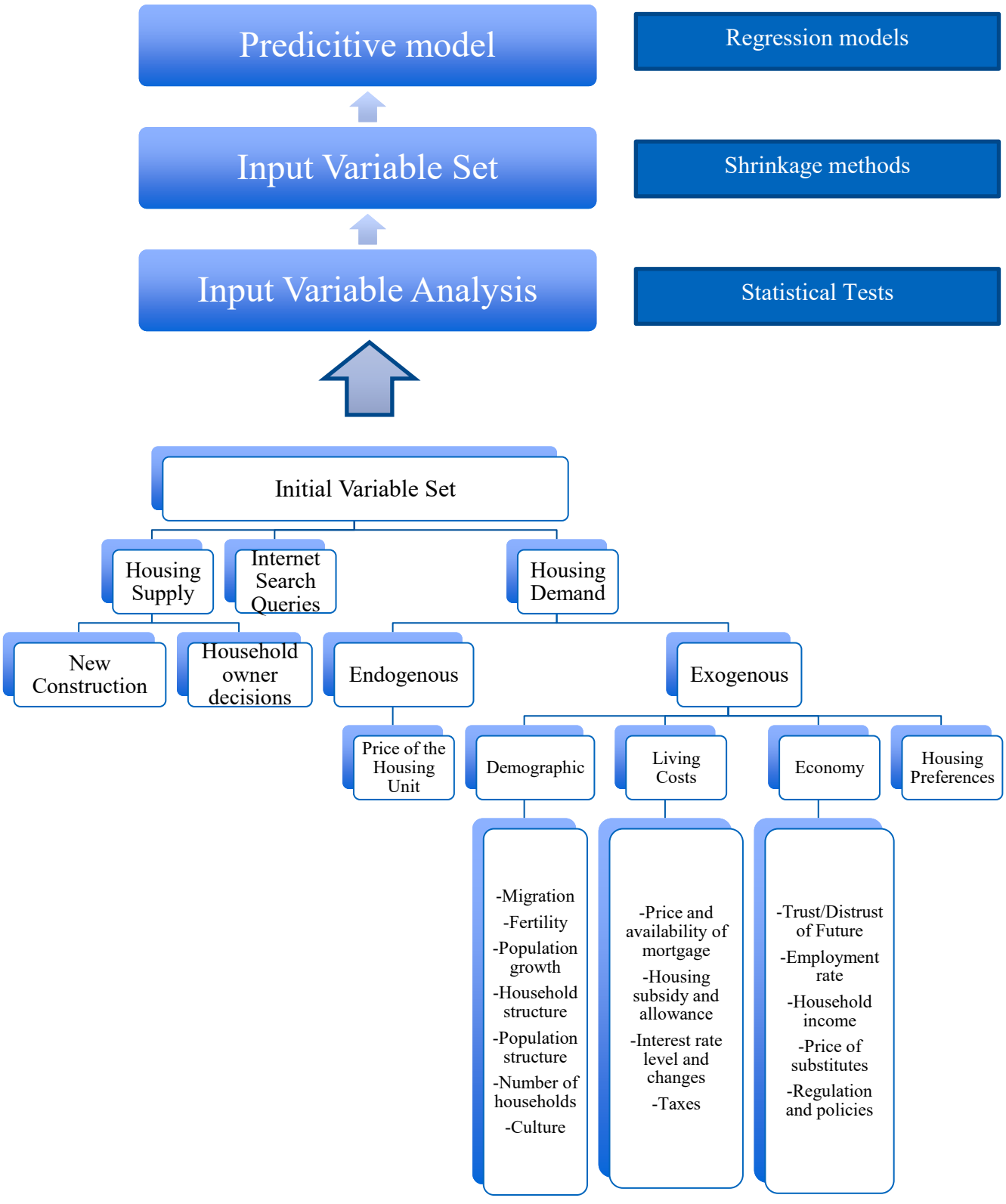
$$\rho_s = Corr(y_t, y_{t-s}) = \frac{Cov(y_t - y_{t-s})}{s.d.(y_t) \times s.d.(y_{t-s})} = \frac{Cov(y_t - y_{t-s})}{Var(y_t)}. \tag{9}$$

The null hypothesis for Breusch-Godfrey test is that there is no autocorrelation up to p where p is the degrees of freedom. The statistic is a distributed chi-square. If the t-value exceeds the critical value or in other words if the p-value is smaller than the decided significance level, then we can reject the null hypothesis and say that the data is autocorrelated. (Lof, 2016). Although Maddala et al (2009) pointed out that Durbin-Watson is often the most used test for autocorrelation, Breusch-Godfrey, also known as LM test, can be used when there are lagged variables involved too.

3.4 Theoretical Framework

In this part a theoretical framework is created on a basis of the literature review highlighting important findings that will be examined and exploited also in the empirical research. This study tries to find an optimal set of predictor variables affecting the real estate market in order to build a robust and accurate predictive model that forecasts the sales volume of old apartments for the next 12 months. The empirical research will focus on building the predictive model based on the chosen input variable set. Therefore, the research is aiming to build the best model as possible in terms of the prediction accuracy and decreasing the root mean-squared error given the nature, quality and availability of the data, industry constraints as well as the target variable.

This theoretical framework links together the most important determinants presented in the literature review and affecting the housing demand and housing supply that are the biggest factors affecting the sales volume of apartments in turn. The determinants are categorized based on their nature. Housing demand is the single most important factor when predicting the sales volume of houses. Hence the majority of the model's input variables will be determinants of demand instead of supply that can be seen from the framework as well. The categories for the housing demand factors are derived from the prior literature. The higher level divides demand factors into 1) Endogenous and 2) Exogenous factors. Exogenous that is indirectly price related factors, are further categorized as follows: 1) Demographic, 2) Living costs, 3) Economy and 4) Housing preferences. In addition to housing demand and supply, internet search queries are lifted as one of the main input variable alternatives based on the information highlighted in the literature review. The factors form the initial input variable set that is then further enhanced with variable analysis, such as variance inflation factor analysis, correlation matrix and other statistical tests, thus creating the input variable set for the model. However, the number of predictors might will still be reduced even more using shrinkage methods such as Lasso when building the model. Overall the theoretical framework describes the flow and process of selecting the optimal input variable set for the model based on the determinants and factors presented in the prior literature and taking into account the number of houses sold as the target variable. Two interviews were also conducted in order to gain deeper industry knowledge and the results of these interviews are also taken into account building the framework.


Predicitive model
Regression models
Input Variable Set
Shrinkage methods
Input Variable Analysis
Statistical Tests
Initial Variable Set
Housing Supply
Internet Search Queries
Housing Demand
New Construction
Household owner decisions
Endogenous
Exogenous
Price of the Housing Unit
Demographic
Living Costs
Economy
Housing Preferences
-Migration
-Fertility
-Population growth
-Household structure
-Population structure
-Number of households
-Culture
-Price and availability of mortgage
-Housing subsidy and allowance
-Interest rate level and changes
-Taxes
-Trust/Distrust of Future
-Employment rate
-Household income
-Price of substitutes
-Regulation and policies

3.5 Hypotheses

In order to connect the results of this study to the existing literature and theoretical framework, five hypotheses are created. These hypotheses are then tested and argued in the discussion and analysis chapter. The first three hypotheses are directly related to the prediction results derived from the current, past and expected conditions in the Finnish real estate market discussed previously. The remaining two hypotheses aim to answer questions about the predictor importance and are based on the existing literature as well as the conducted interviews that will be elaborated more in the methodology chapter.

Hypothesis 1 (H1): *The number of sold old apartments in total will increase within the next 12 months.*

As discussed, many industry expert predicts that the sales volume of apartments will increase also in the future (Brotherus, 2017; Kekäläinen, Tähtinen & Vuori, 2017). Although the demand for newly constructed apartments have been higher, it can be assumed that the demand and thus the sales volume of old apartments will also be increasing. Hence the first hypothesis states that the number of sold old apartments will increase within the next 12 months. The next 12 months are counted from Q4 2017.

Hypothesis 2 (H2): *The sales volume for old apartments will increase more in the capital region (Helsinki, Espoo and Vantaa) than in other regions.*

Brotherus (2017) stated that the demand for old apartments has been growing in 2017 especially in Helsinki and the capital region. Furthermore, he pointed out that there is a gap in sales volume between the bigger cities such as Tampere and Turku and middle-sized cities such as Jyväskylä, Oulu and Kuopio. Vehviläinen (2016) also emphasized the impact of urbanization on the Finnish real estate market. Therefore, the second hypothesis states that the sales volume for old apartments will increase more in the capital region compared to the other cities taken into account in this research.

Hypothesis 3 (H3): *The sales volume for smaller studio apartments will increase more than for other apartment types.*

Vehviläinen (2016) found out that the supply of smaller apartments has increased in contrast to bigger apartments. This was mainly because of the changes in demographics such as the family size or young people's willingness to live alone. Brotherus (2017) supported this argument and also justified the growth of smaller apartment sales volumes with the decrease in household size. Hence the third hypotheses tests whether the sales volume for smaller studio apartments will increase more compared to the other apartment types taken into account in this study. Studio here refers to a one-room flat.

Hypothesis 4 (H4): *The economic variables have the biggest impact on the number of houses sold.*

As showed in the literature review most of the real estate demand factors are related to either economy or household finances. Moreover, the interviewed industry experts also stressed the importance of economy and especially consumer's trust in economy when ranking the possible predictor variables. The theoretical framework divided the exogenous real estate demand factors into four different categories from which the economic one including its variables is considered to be the most important in terms of the prediction power. Therefore, the fourth hypothesis tests whether the economic variables have the biggest impact on the target variable that is the number of houses sold.

Hypothesis 5 (H5): *Search query data from Google Trends enhances the model and serves as an important predictor variable.*

Choi & Varian (2009) were able to improve the squared R when adding search query data to their models. In fact, Norros (2014) emphasized that, based on all the prior studies, the statistical and predictive models' accuracy can be improved by combining real-time data concerning consumer's online search behavior and actual data about the past development of the real estate market. The theoretical framework includes internet search queries as one of the main data and predictor variable sources of the model. Hence, the fifth hypothesis tests whether the search queries of specific search terms actually enhance our model.

4 Findings and the Model

In this chapter the results of the different methods and tests are presented. Moreover, the procedures based on these results are highlighted and argued. We start with exploring the data and variable correlation. This research follows the CRISP-DM methodology that is the cross-industry process for data mining. As presented in the Figure 4, CRISP-DM has 6 phases. Literature review of the study was the first phase, understanding the business, in this case the real estate industry and what factors affect it. This chapter will consist of the 3 following phases. First we explore the data in order to understand it and then the data is prepared for the modeling phase. The modeling phase means testing different methods and variables in order to build a robust model with accurate prediction power.

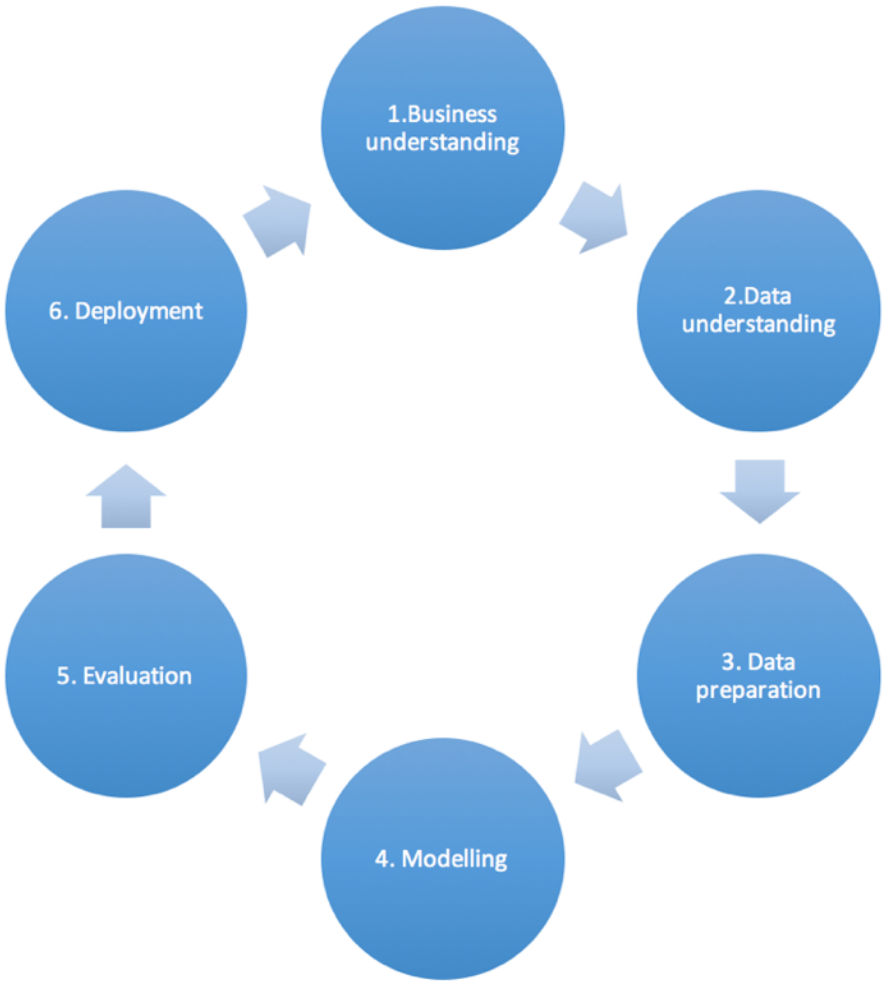


Figure 4. CRISP-DM.

4.1 Data Exploration

As presented previously in the methodology chapter, the data has several explanatory variables from several data types and sources. Hence it is important to explore them to get a better understanding what kind of data we are dealing with.

Figure 5 presents the past development of the total sales volume of old apartments per year quarter and location. As we can see the graphs slightly differ based on the type and location. The overall trend, regardless of the location and apartment type and besides couple of exceptions, seems to be that the sales numbers were decreasing since 2010 until they slowly began to rise around 2014.

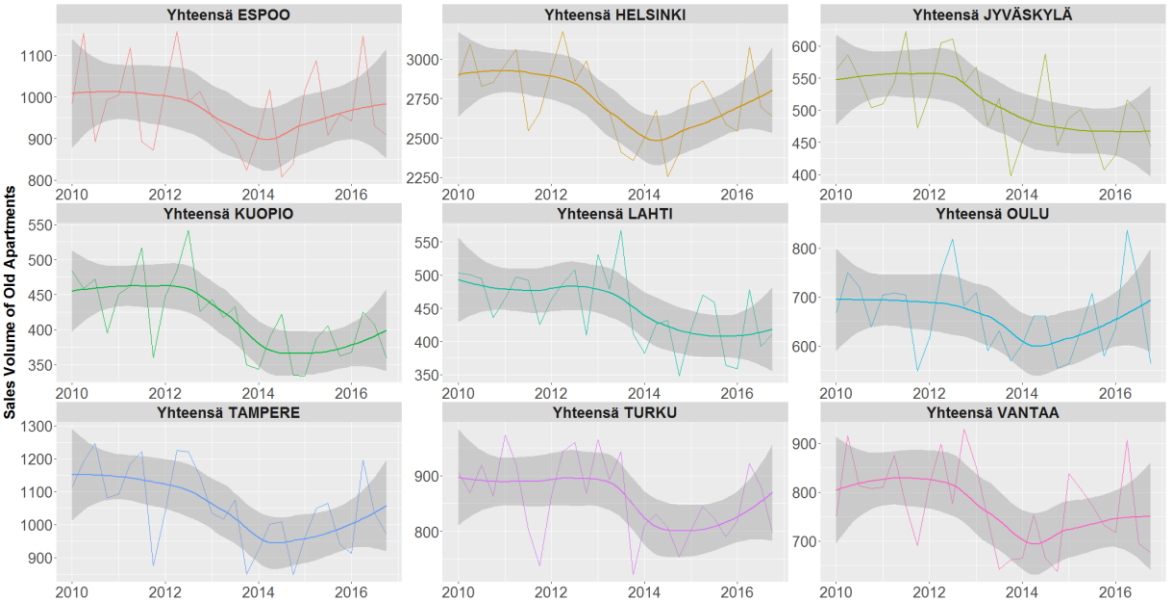


Figure 5. The Past Sales Volume Development of the Old Apartments.

Figure 6 compares the changes in respondent's trust to the future and economy versus respondent's intention to buy a house within 12 months over time and per location. Based on the interviews and as discussed in the literature review financial and economic measures and especially the consumer's trust in economy are perhaps the most important factors affecting the housing demand. Hence it is interesting to see whether the trust is correlated with the consumer's actual intention of buying an apartment. As we can see from the Figure 6, consumer's intentions to buy does not strictly follow the trust in economy. In fact, the intention seems to have been rather stable over time.

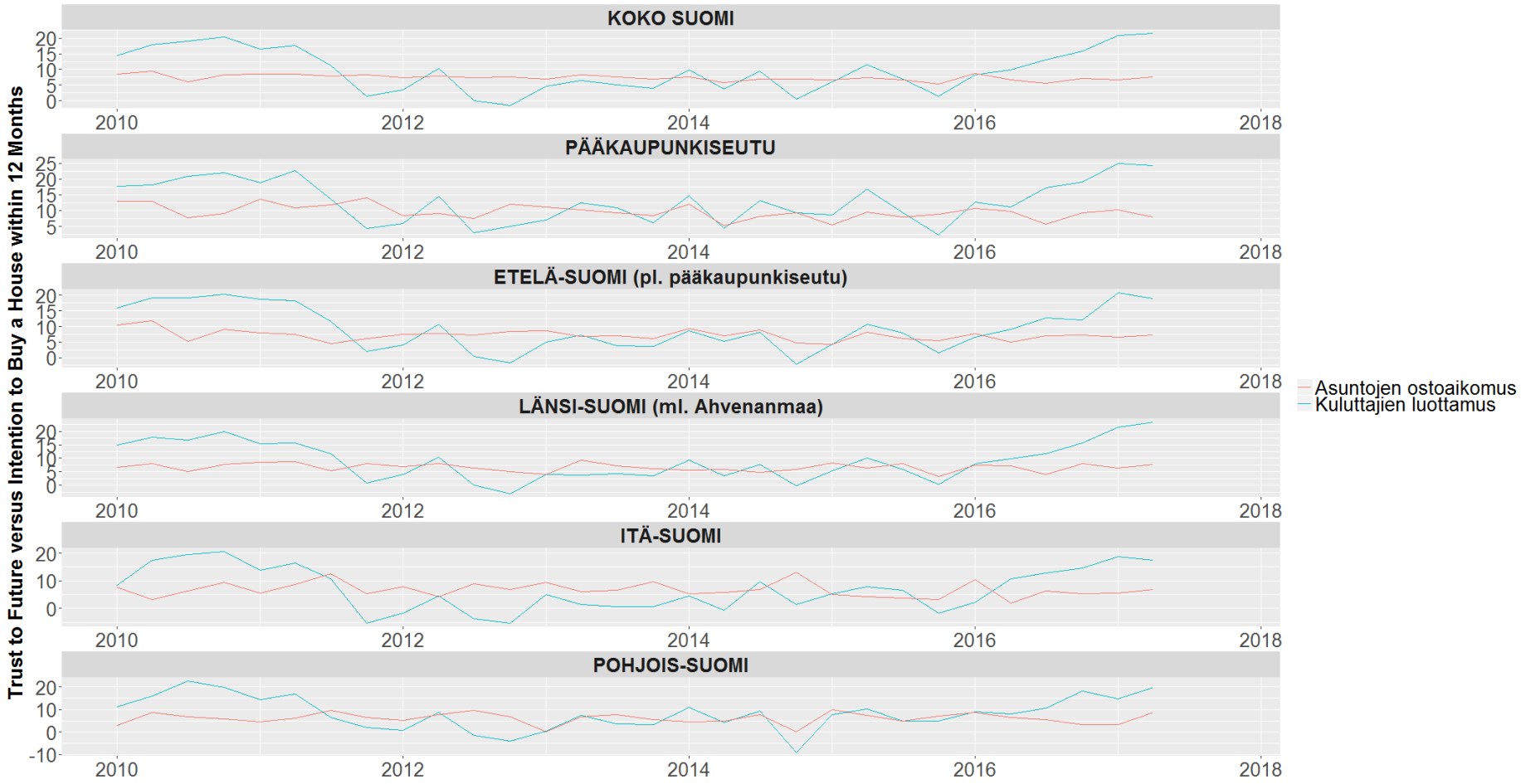


Figure 6. The Changes in Respondent's Trust to the Future and Economy versus Respondent's Intention to Buy a House within 12 Months.

The following Figures aim to demonstrate the changes in the consumer's buying intention and trust in two different graphs. A colored line represents one of the geographic locations. As we can see the capital region seems to have had both the highest trust and the intention to buy apartments whereas northern and eastern Finland have had the lowest ones. Overall we can see that the consumer's trust has begun to rise again since the end of 2015 whereas there is no clear pattern or trend for the intention of buying an apartment.

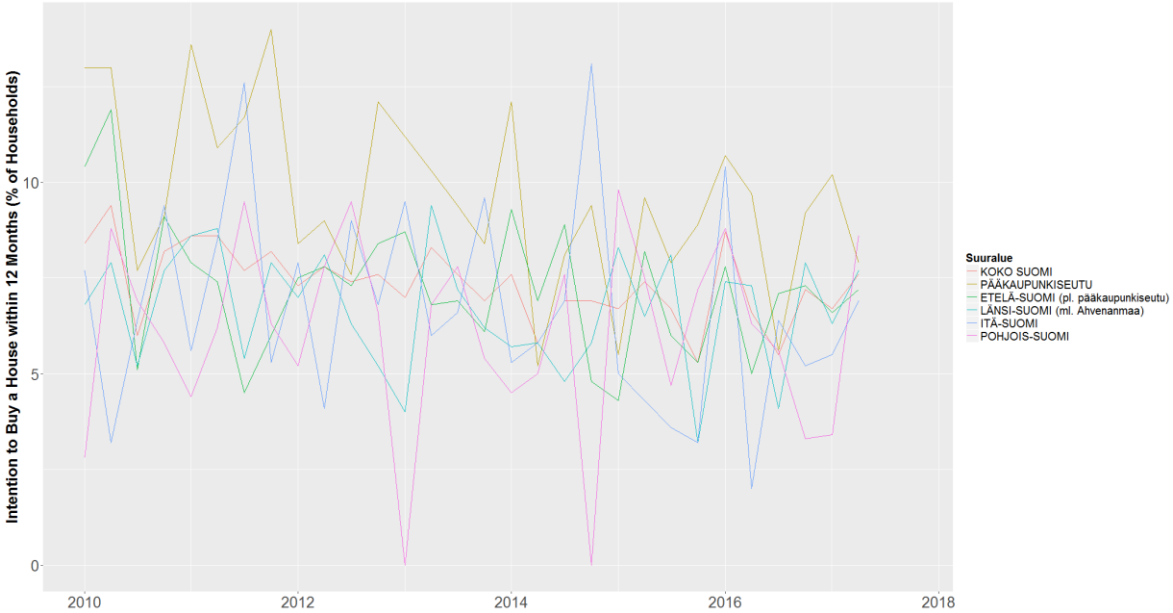


Figure 7. Consumer's Intention to Buy an Apartment within the next 12 Months Over Time per County.

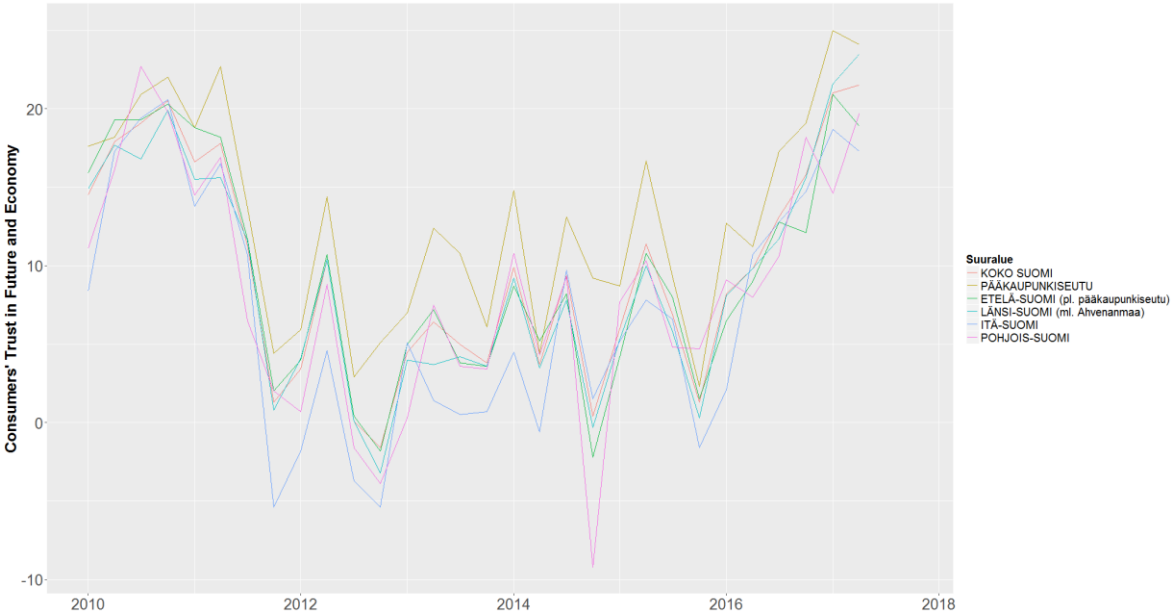


Figure 8. Consumer's Trust Over Time per County.

Three search terms were chosen based on the interviews: Myytävät asunnot (Apartments for sale), Kiinteistönvälitys (Real estate brokerage) and Kiinteistönvälittäjä (Real estate

agent). Figure 9 describes the development of the search volumes with these search terms over time. As we can see Real estate brokerage used to have highest search volume until late 2014 when Real estate agent passed it. Since early 2016 also Apartments for sale has overtaken Real estate brokerage in search volume. What is interesting it seems that both Apartments for sale and Real estate agent as search terms have been increasing since mid 2014 whereas Real estate brokerage has been decreasing.

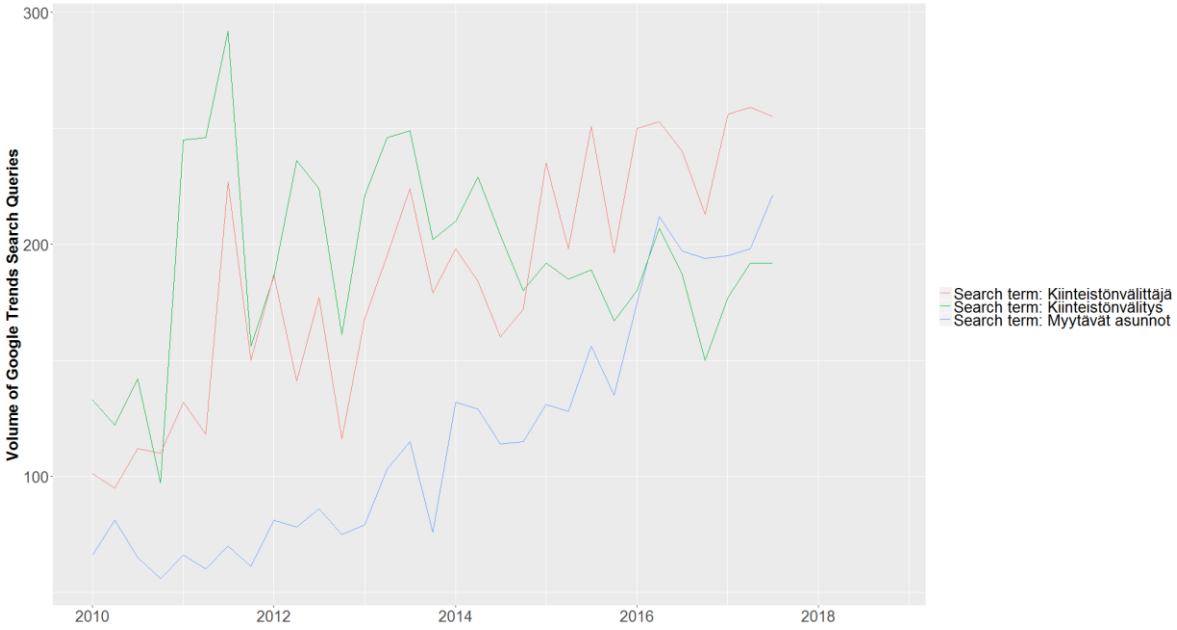


Figure 9. Google Trends Search Volume: Kiinteistönvälitys (green), Kiinteistönvälittäjä (red), Myytävät asunnot (blue)

Following three Figures show the changes in search volumes for each individual search term in their own graphs. We can see that the search volume for the apartments for sale has been steeply increasing since 2011. The search volume for Real estate agents has also been increasing. However, the increase has not been as strict and clear as with the apartments for sale. In contrast, the search volume for real estate brokerage has been slowly decreasing since early 2013. This is rather interesting when thinking about the fact that the other two search terms have been increasing at the same time. Possible explanations could be that potential buyers rather search apartments directly on their own or specific real estate agents instead of whole companies.

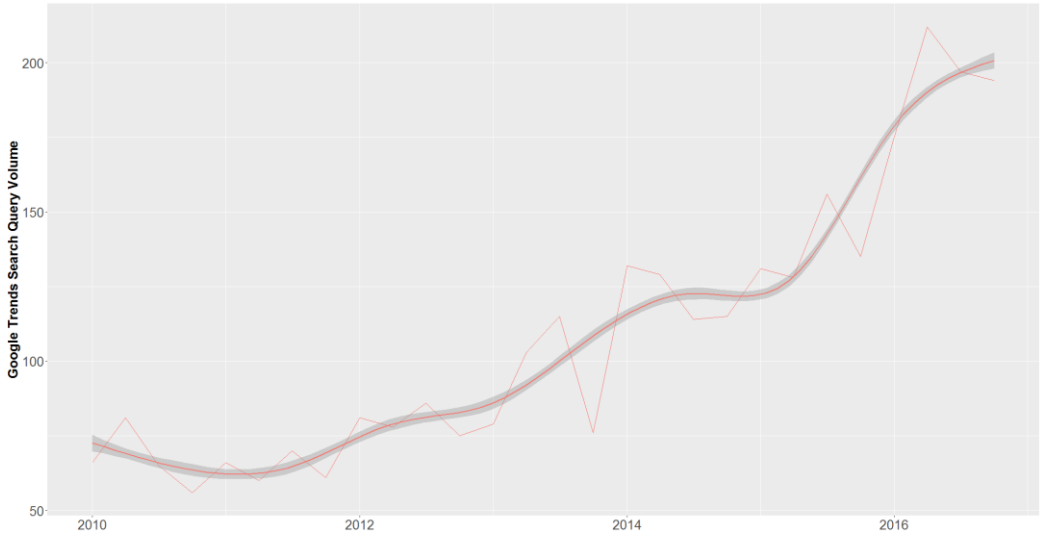


Figure 10. Google Trends Search Volume for "Myytävät asunnot" (Apartments for sale).

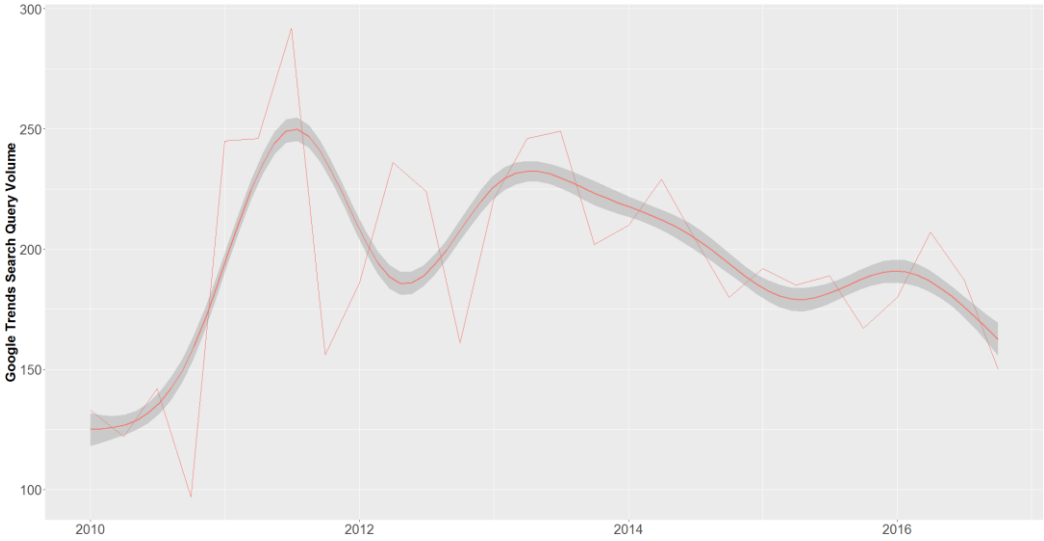


Figure 11. Google Trends Search Volume for "Kiinteistönvälitys" (Real estate brokerage).

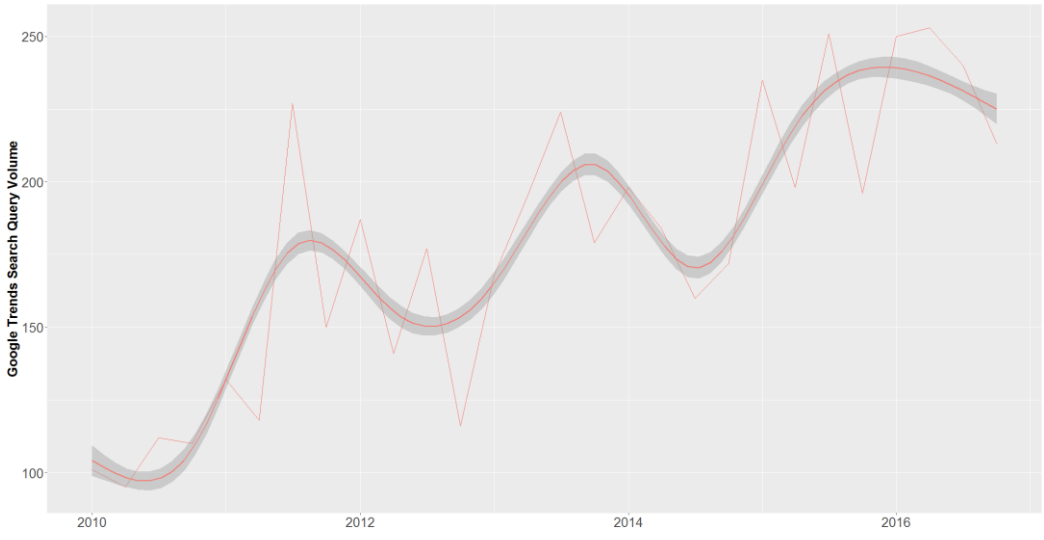


Figure 12. Google Trends Search Volume for "Kiinteistönvälittäjä" (Real estate agent).

Figure 13, 14 and 15 describes the development of the population, ratio of households living in a rental apartment and the average size of a household. The amount of population has been log transformed as it is easier to visualize it with smaller numbers. As we can see from Figure 13 the population has been slowly but steadily growing. Although there are not any small cities and every city is from an urban area, there seems to be a clear pattern that the population growth is strongest in the capital region such as in Helsinki and Espoo. This could be a sign of migration to bigger cities already discussed in the literature review.

Figure 14 shows a pattern that the ratio of households living in a rental apartment seems to been increasing. The steepest increase has been in Tampere although the increase has been strict in every city since 2016. This could affect the sales volume of old apartments as well being a substitute offering. However, there is not a way to deduce whether this increase is a consequence of people willing to live in rental apartments instead of purchased apartments or rather just because of the population growth overall.

Figure 15 describes the average size of a household over time per city and apartment type. As we can see the increase or decrease has been quite stable regardless of the location or type. However, while the household size in Helsinki and capital region has been increasing, the household size has been decreasing in middle-sized cities such as in Kuopio, Jyväskylä and Lahti. Furthermore, Espoo and Turku has been exceptions where the average household size has stayed more or less the same since 2010.

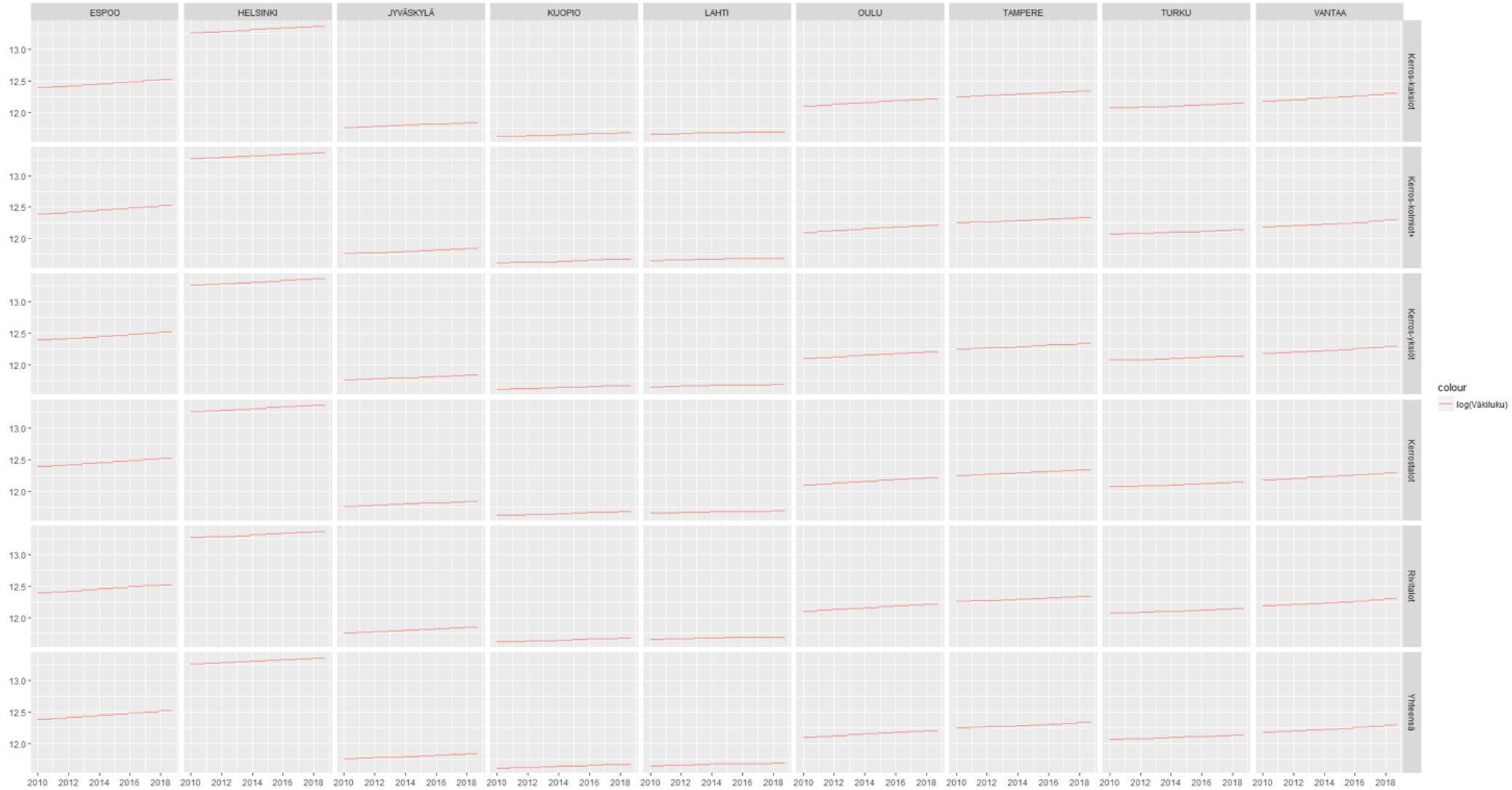


Figure 13. The Population Growth Over Time per City and Apartment Type.

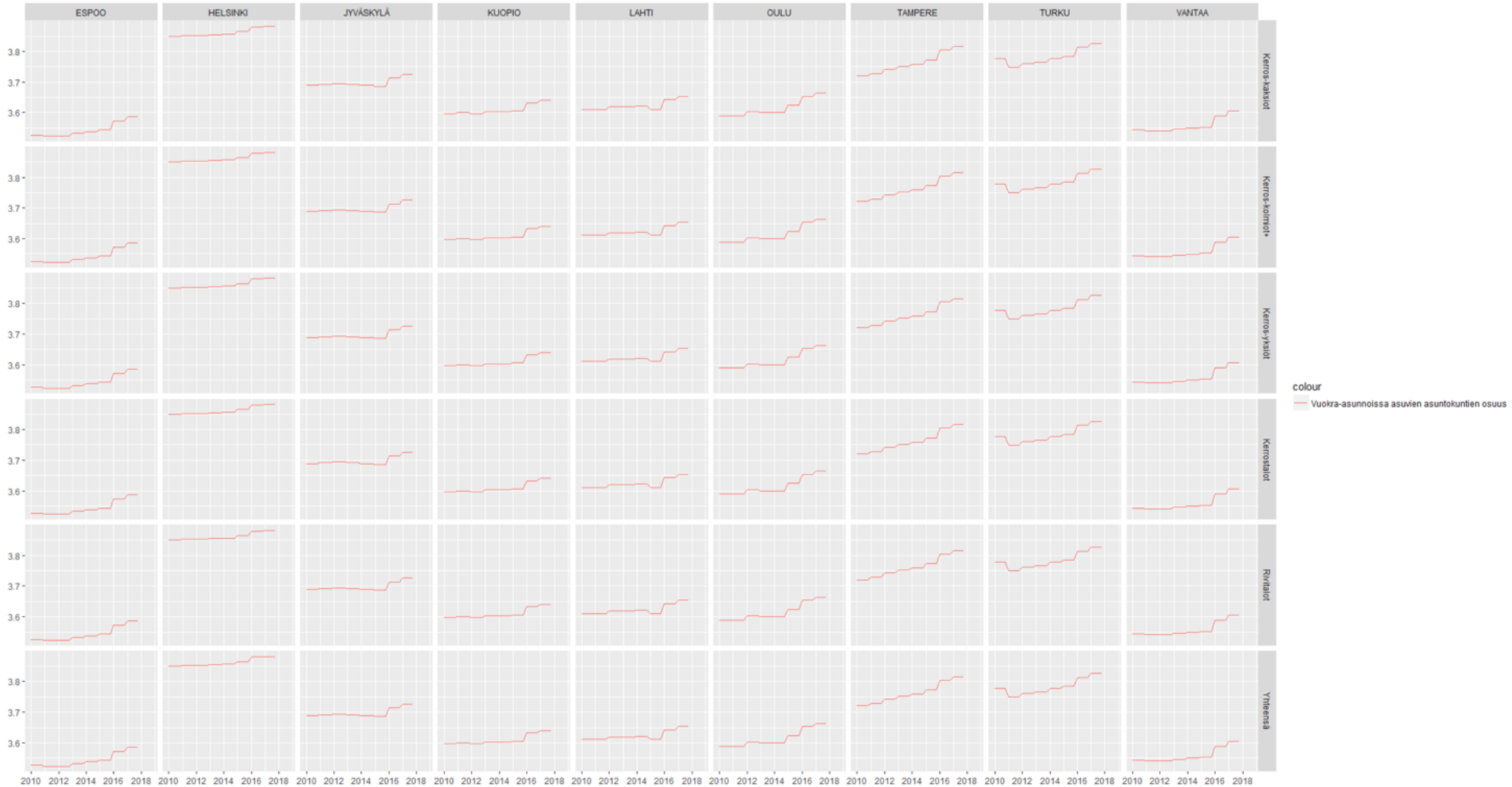


Figure 14. The Ratio of Households Living in a Rental Apartment Over Time per City and Apartment Type.

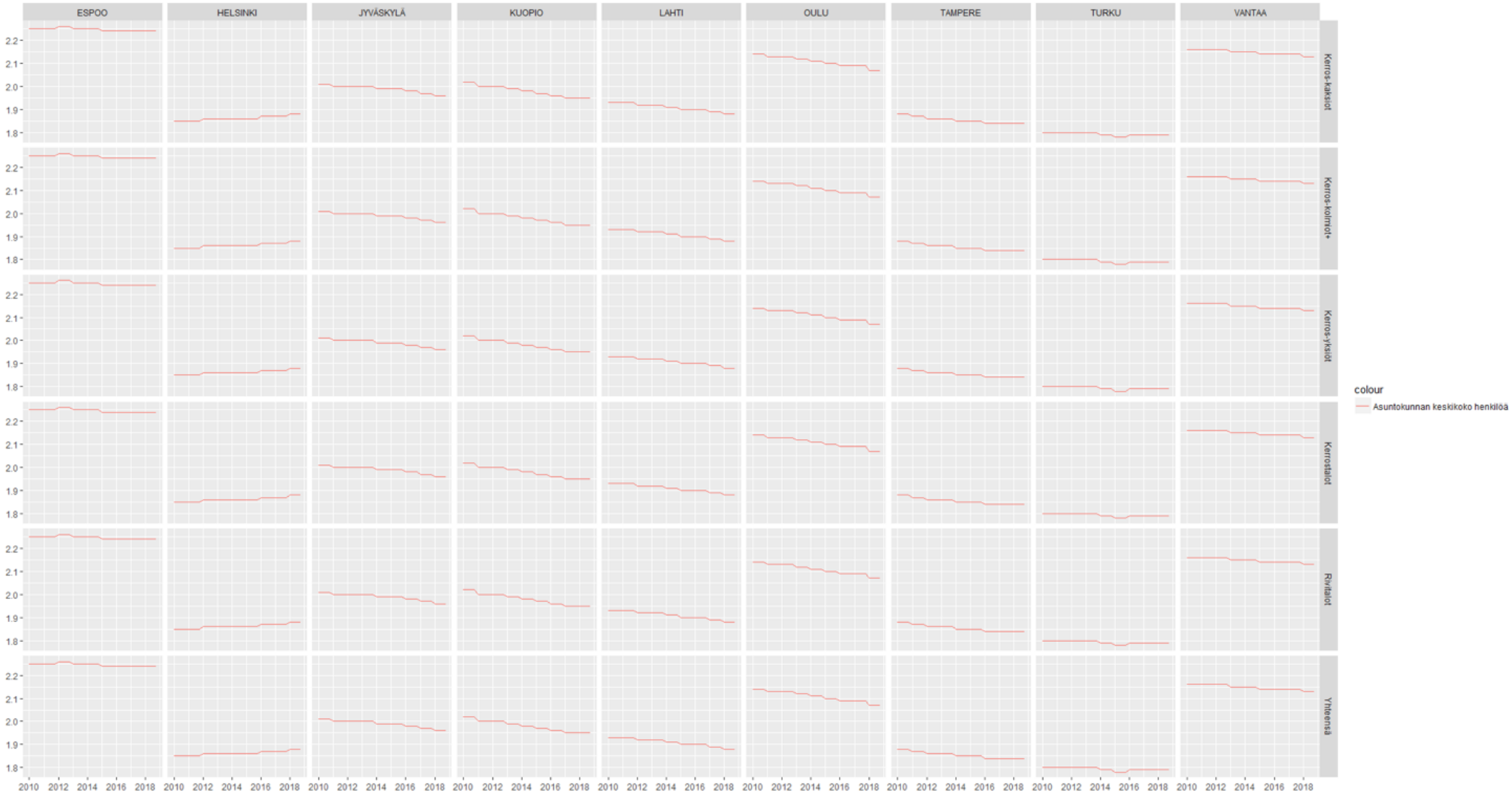


Figure 15. The Average Size of a Household Over Time per City and Apartment Type.

4.2 Multicollinearity

Chen (2008) defined multicollinearity as "a situation in which two or more independent variables are perfectly or nearly perfectly correlated". Although multicollinearity does not affect the overall fit of the model, it increases and makes the standard errors unstable and biased thus leading to unstable p-values, low statistical significance and less precise OLS coefficients. (Lof, 2016; Vatcheva et al, 2016) Lof (2016) pointed out that the perfect multicollinearity is such that an independent variable can be expressed as a linear function of the other explanatory variables. This can be mathematically illustrated as follows:

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + \varepsilon_t,$$
$$x_{3,t} = a + bx_{2,t}. \quad (10)$$

First we look at the correlation between the independent variables in a correlation matrix. Figure 16 presents the correlation matrix in which the circle size refers to the level of collinearity. The bigger the circle is higher the correlation is between those variables. The color refers to the negativity or positivity of the correlation. Figure 17 presents the same correlation matrix but instead of a circle we have the actual level of correlation as number. As we can see from the Figures it seems that the highest positive correlation, above 0,9, is between the "Real estate agents" search volume and "Apartments for sale" search volume, the ratio of households living in a rental apartment and the employment rate as well as the GDP during economic growth (BKT scenario nousukausi) and the GDP with the 2010 price level (BKT 2010 hinnoin). It also seems that most of the highest negative correlations are between the variable "GDP during economic growth" and another variable. Hence it seems wise to drop at least this independent variable off the model in order to prevent multicollinearity and the issues it brings. However, only looking at the correlation matrix is not often enough to determine which variables should be included into the model or which ones should not. Hence a Variance Inflation factor (VIF) test is performed in order to better determine what variables have a high multicollinearity given the dataset.

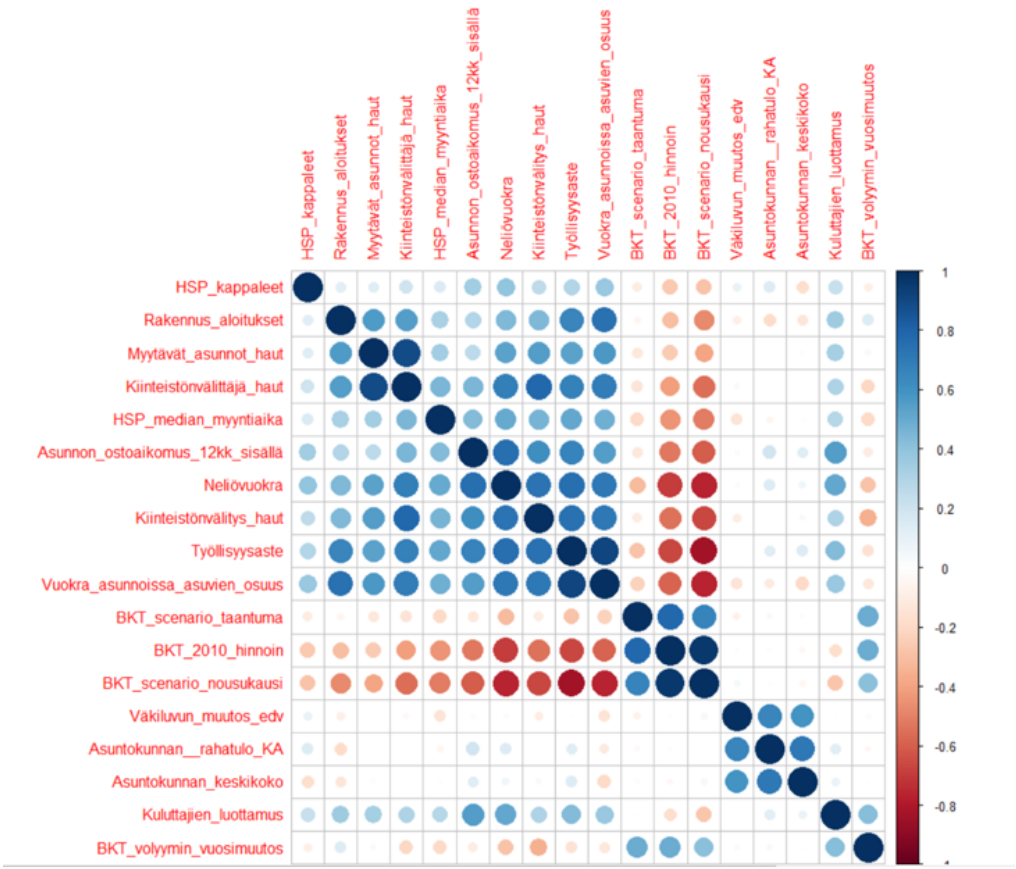


Figure 16. Correlation Matrix with Circle Size as the Measure.

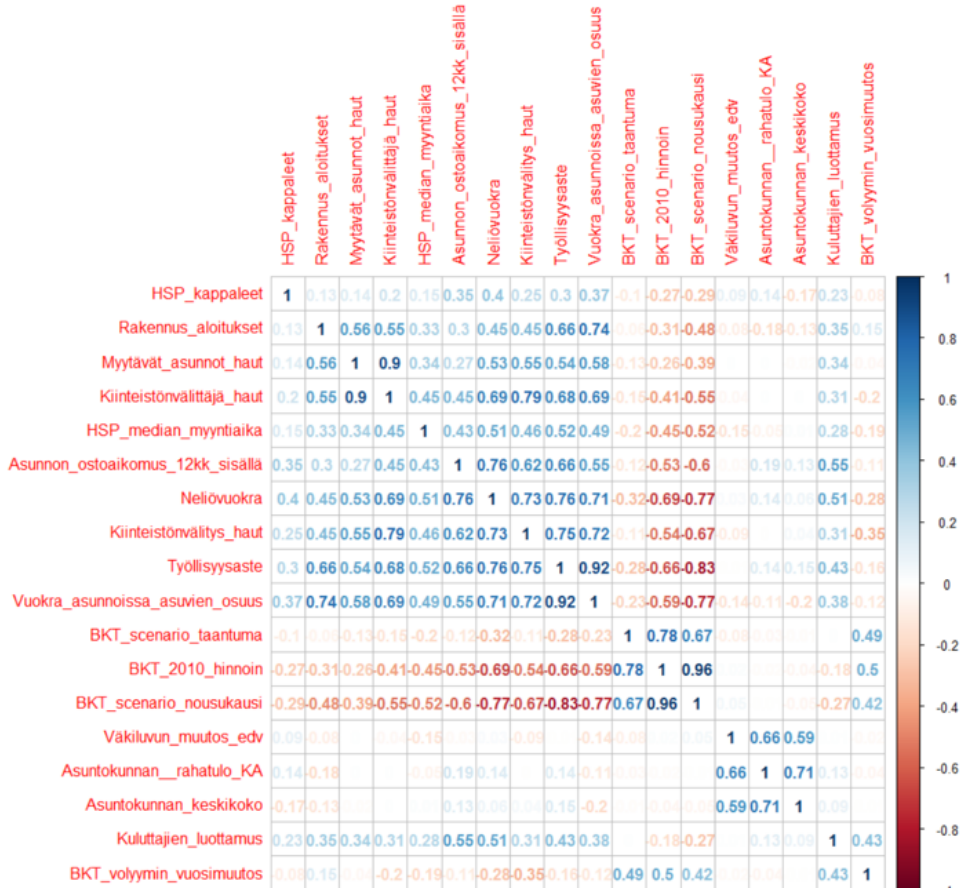


Figure 17. Correlation Matrix.

Variance Inflation Factor (VIF) is a statistic measuring the level of multicollinearity. The VIF is defined as follows:

$$VIF_i = \frac{1}{1 - R_i^2}. \tag{11}$$

The prior literature suggested different levels of VIF based on which the multicollinearity problem is either critical or not. According to Ringle et al (2015) the maximum acceptable level of VIF should be smaller than 5. However, more popular view is the one offered by Hair et al (2015) that is the maximum acceptable level of VIF is smaller than 10. Hence it seems that the chosen level follows the decisions and data in every specific research. Therefore, in this study, we will use the more traditional maximum level of VIF that is 10. Table 5 presents the VIF levels for four different iterations. The initial variable set is chosen based on the literature review. As we can see there are two variables exceeding the threshold of 10, the ratio of households living in a rental apartment (Vuokra-asunnoissa asuvien asuntokuntien osuus) and the average size of a household (Asuntokunnan kesikoko henkilöä).

Table 5: Variance Inflation Factor (VIF) Statistics

Variable	Initial variable Set	Second Iteration	Third Iteration	Fourth Iteration
HSP kappaleet	1.748	1.748	1.744	1.748
Myytävät asunnot haut	4.715	4.713	4.714	4.702
Kiinteistönvälitys haut	2.665	2.654	2.641	2.642
Kiinteistönvälittäjä haut	5.259	5.255	5.258	5.258
Kuluttajien luottamusindikaattori	2.711	2.711	2.711	2.703
Asunnon ostoaikomus 12 kk sisällä	1.656	1.581	1.593	1.656
Neliövuokra	2.995	2.993	2.984	2.986
Työllisyysaste	6.391	6.057	6.350	6.391
Väkiluvun muutos edellisestä vuodesta	3.900	3.232	2.261	**Omitted**
Väkiluku	9.898	5.044	9.056	9.865
Asuntokunnan käytettävissä oleva rahatulo keskimäärin	7.667	7.583	6.625	7.663
Rakennus aloitukset	1.723	1.524	1.690	1.711
Vuokra-asunnoissa asuvien asuntokuntien osuus	21.724	**Omitted**	7.296	18.0045
HSP median myyntiaika	1.579	1.575	1.577	1.571
Asuntokunnan keskikoko henkilöä	20.157	6.770	**Omitted**	11.686
BKT volyymin vuosimuutos	3.732	3.730	3.732	3.727
BKT 2010 hinnoin	2.398	2.358	2.268	2.237

The second iteration represents the VIF levels when the ratio of households living in a rental apartment (Vuokra-asunnoissa asuvien asuntokuntien osuus), having the highest level before, has been omitted. As we can see, now every variable has a smaller VIF than 10 including the average size of a household that previously had a VIF larger than 10. Third iteration represents the VIF levels when the average size of a household (Asuntokunnan keskikoko henkilöä) is omitted instead of the ratio of households living in a rental apartment. All of the variables have a smaller VIF than 10 also in this case. However, it seems that the VIF for most of the variables is smaller when omitting the ratio of households living in a rental apartment instead of the average size of a household. The fourth iteration represents the VIF levels when the change in population compared to the previous year (Väkiluvun muutos edellisestä vuodesta) is omitted. This decreases all of the variable VIFs compared to the original VIF levels of the initial variable set. For instance, the VIF for the ratio of households living in a rental apartment is now 18 compared to 21 and the VIF for the average size of a household is 11 instead of 20. The other variables' VIF levels only slightly decrease.

Although Variance Inflation Factor is valid and common measure of multicollinearity it has its challenges. Maddala et al (2009) pointed out that VIF only explore the intercorrelations between the independent variables. Hence a change or transformation in an explanatory variable might change the correlations and thus the VIF levels. Therefore, methods such as shrinkage like Ridge or Lasso regression, principal component regression or simple variable dropping is suggested and promoted as a better way to deal with multicollinearity problems. The results of Variance Inflation Factor tests are still taken into account in this study and especially when building the baseline model without using a shrinkage method.

4.3 Statistical Tests

When building a statistical model, it is important to test if the data meets the assumptions set beforehand. Three tests were performed: 1) Augmented Dickey-Fuller, 2) Breusch-Pagan and 3) Breusch-Godfrey. The results of this tests are presented next. All of the tests are performed with a test dataset "Testi" consisting of the initial variables presented in Table 4.

4.3.1 Augmented Dickey-Fuller

As explained in the methodology chapter, Augmented Dickey-Fuller tests for the stationarity of the data. Table 6 presents the results of this test with different number of lags. As we can see the p-value seems to be same, 0,01, for each scenario differing based on the lag number. Hence we can reject the null hypothesis of having a unit root and say that the data is stationary at the 5% significance level. This is one of the major statistical assumptions when dealing with time series data.

Table 6: Augmented Dickey Fuller Test Results

Dickey-Fuller	Number of Lags	P-value
-26,378	1	0,01
-21,834	2	0,01
-18,029	3	0,01
-15,535	4	0,01
-14,082	5	0,01
-12,876	6	0,01
-11,864	7	0,01
-10,800	8	0,01

4.3.2 Breusch-Pagan

Breusch Pagan tests for the heteroscedasticity of the data. Table 7 presents the initial results of the test without performing any alterations.

Table 7: Breusch-Pagan Test Results

BP	Degrees of Freedom	P-value
354.690	16	< 2.2e-16

As we can see the p-value of the test is smaller than 0,05. Hence we can reject the null hypothesis and deduce that there is some heteroscedasticity. However, as Rodriguez (2009) pointed out a lot of issues can be fixed by using a GLM approach like Poisson regression

instead of ordinary linear models. Hence the Poisson regression model, assuming that the dependent variable follows a Poisson distribution, takes into account the heteroscedasticity already in the first place. Therefore, no other alterations, like log transformation of independent variables, is needed to deal with heteroscedasticity. (Rodriguez, 2009) In fact, response variables from count or binary response dataset are typical examples of data with heteroscedasticity. (Lockhart, 1997). However, when using Poisson regression there is always a potential issue with overdispersion that will be discussed more later.

4.3.3 Breusch-Godfrey

Breusch Godfrey tests for the autocorrelation of the data. Similar to the Breusch-Pagan test the null hypothesis is that there is no autocorrelation. Tables 8 and 9 below presents the results of the tests.

Table 8: Breusch-Godfrey Test results without Log Transformation

LM	Degrees of Freedom	P-value
29.450	1	5.73e-08

Table 9: Breusch-Godfrey Test Results with Log Transformation

LM	Degrees of Freedom	P-value
0.630	1	< 0.43

As we can see the p-value of the first test is smaller than 0,05. Hence we can reject the null hypothesis and say that there is autocorrelation. However, the second test uses a regression model where the dependent variable is log transformed in order to simulate the Poisson regression model. Now the p-value is clearly larger than 0,05 and we can say that there is not autocorrelation that needs to be fixed. We have used 5% as the significance level for all of the tests that is the typical level for also many prior studies. The same significance level is used also when evaluating the statistical significance of the regression coefficients.

4.4 Modelling

In this section four models are built using two different regression techniques and the prediction results presented. The first technique used for the baseline model and the second model is Poisson regression. The second technique is Negative Binomial regression that is used for the third and fourth model. Lasso regression is also used as a shrinkage method to determine what variables should be included to the model. The data has been split into training dataset and testing dataset. James et al (2013) defined the purpose of this process as "applying a statistical learning method to the training data in order to estimate the unknown function f". In other words, training data is used to train the model that is after tested with the unseen test dataset. In this study the prediction point, dividing the data, is Q1 2016. Hence the training data dates from the beginning of the 2010 until the first quarter of 2016 whereas the test data begins from the first quarter of 2016 and stops at the first quarter of 2017. The purpose is to predict the dependent variable for the next four quarters. As it is already the end of year 2017 the models forecast even beyond that until the third quarter of 2018.

4.4.1 Models

4.4.1.1 Baseline Model – Poisson Regression

The baseline model is built using a Poisson regression discussed earlier in the methodology chapter. Poisson regression was chosen as the method given that the response variable is count data and thus follows a Poisson distribution. The independent variables were chosen, after exploring the data and performing statistical tests, from the initial variable set presented in Table 4. Table 10 below contains the variables chosen for the first model and their Beta estimates. As we can see some of the coefficients are null. We have used elastic net approach to benefit from the shrinkage already in this model. R package called "glmnet" offers an easy way to include elastic net to the model. Alpha parameter controls the penalty of the objective under optimization similar to Lasso and Ridge. In fact, alpha bridges the gap between these two methods.

1. $\alpha = 1$: Full Lasso Regression. This is the default.
2. $\alpha = 0$: Full Ridge regression

Hence elastic net is a mixture of these two methods. We have used here 0,3 for alpha and thus the model is more like a Ridge than Lasso Regression. Moreover, all of the predictor variables are lagged terms in order to take into account the time series aspect of the data. We want to predict how the past data and trends affect the future. Lag 8 was chosen as the appropriate number of lags dating back to two years. For example, the predictor's value for the first quarter of 2016 is actually its value for the first quarter of 2014. This way we can make sure that there is enough data to predict beyond one year from today. In fact, the original research objective was to forecast sales volume of old apartments for the next two years thus eight following year quarters. If lags from one to seven were also included to the model it could have forecasted the sales volume only for the next year quarter as there was not enough input or training data available yet.

In addition, couple of predictors were log-transformed given the fact that they had high standard deviation and taking a logarithm seemed to improve the overall R2 of the model. The number of constructions was also normalized. Some of the independent variables were eliminated from the model based on their VIF score. These included the change in population compared to the previous year (Väkiluvun muutos edellisestä vuodesta), GDP during economic growth (BKT scenario nousukausi), GDP during economic decline, (BKT scenario taantuma) and the change per year in the GDP (BKT volyymin vuosimuutos). Moreover, search volume for apartments for sale ("Myytävät asunnot_haut") was removed as it tend to increase the target variable too much. As we saw from Figure 10, the search volume for the apartments for sale has been steeply increasing since 2011. This, considering the small amount of data, would skew the results as there has not been a period of time where this search index has been decreasing thus falsely giving too strict increase for the predictions. Finland's GDP with 2010 price level (BKT 2010 hinnoin) was kept in the model as it contained the absolute values of GDP and served as the predictor for future economy.

Table 10: Baseline Model Regression Coefficients Poisson Regression

Variable	Estimate
Intercept	1.515e+00
Log(HSP_kappaleet_lag8)	7.851e-01
Väkiluku_lag8	4.246e-07
E4_ostoaikomus_lag8	2.855e-03
A1_luottamus_lag8	7.546e-03
Kiinteistönvälitys_haut_lag8	1.960e-04
Kiinteistönvälittäjä_haut_lag8	3.495e-05

Log(Asuntokunnan_käytettävissä_olevat_rahatulot_keskiarvo_lag8)	2.673e-01
Vuokra_asunnoissa_asuvat_lag8	9.865e-03
Työllisyysaste_lag8	.
Asuntokunnan_keskikoko_henkilöä_lag8	.
Neliövuokra_lag8	.
Log(Rakennusaloitukset_normalized_lag1y + 1)	-8.953e-02
Log(HSP_median_myyntiaika_lag8)	.
BKT_2010_hinnoin_lag8	-1.823e-05
AlueHELSINKI	.
AlueJYVÄSKYLÄ	.
AlueKUOPIO	-1.343e-01
AlueLAHTI	-1.384e-01
AlueOULU	.
AlueTAMPERE	.
AlueTURKU	1.580e-02
AlueVANTAA	-1.033e-03
tyyppiKerros-kolmiot+	-3.122e-02
tyyppiKerros-yksiöt	-1.926e-02
tyyppiKerrostalot	1.934e-01
tyyppiRivitalot	-2.666e-02
tyyppiYhteensä	2.526e-01
quarters(vuosineljannes)Q2	.
quarters(vuosineljannes)Q3	.
quarters(vuosineljannes)Q4	-3.167e-02

4.4.1.2 Poisson Regression Model with Lasso Variable Selection

As discussed before, the large number of predictors created a need to use a shrinkage method for variable selection. Hence a Lasso regression model was built. The model is not used for prediction by itself but only to the variable selection. The variables selected by the model were then used in Poisson regression and the results compared to the baseline Poisson regression model were the variables were chosen only based on the prior literature, interviews, variable correlation and VIF score. As mentioned earlier in this chapter, VIF is not usually the best measure for selecting the independent variables and thus a more sophisticated alternative like Lasso or Ridge is needed.

Figure 18 presents the lambda values, the weight of the regularization term controlling the penalty as discussed earlier. We can see that a very small lambda retains more variables in the model compared to a larger lambda. As the lambda grows more coefficients are shrank to zero thus resulting in fewer variables in the model. The top of the plot shows the number of predictors in the model for given value of lambda. For example, there are 21 predictors when lambda has a value of 2 but only 1 predictor with the value of 6. Each colored line represents one of the model coefficients. The bigger the lambda closer to the Ordinary Least Square equation the optimization objective is.

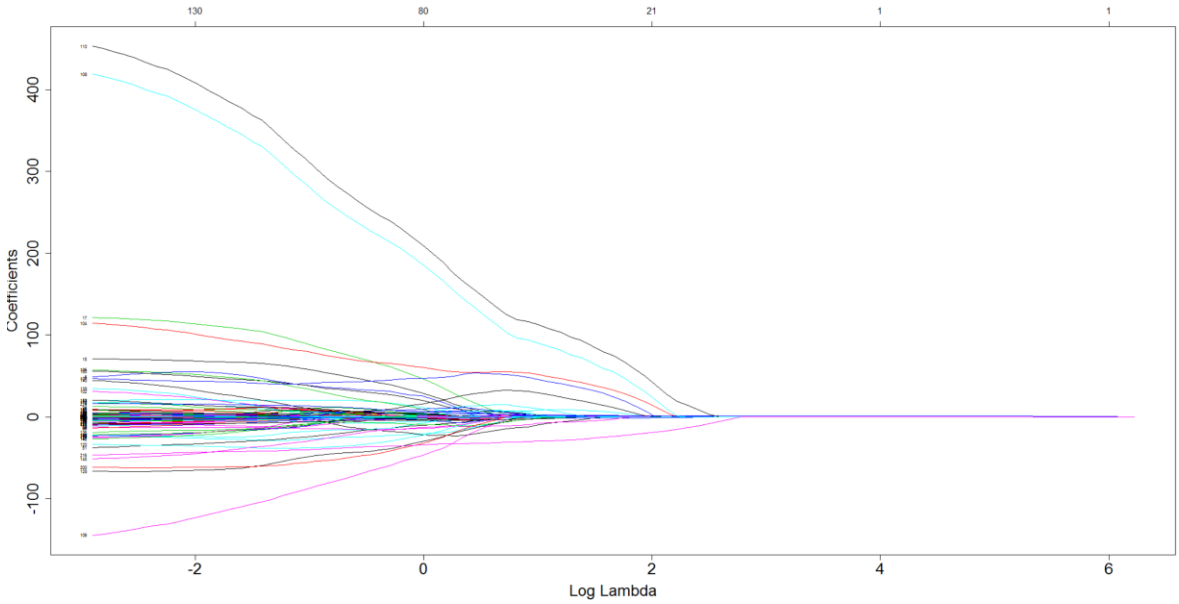


Figure 18. Log Lambda versus Coefficients and Non-Zero Independent Varaiables

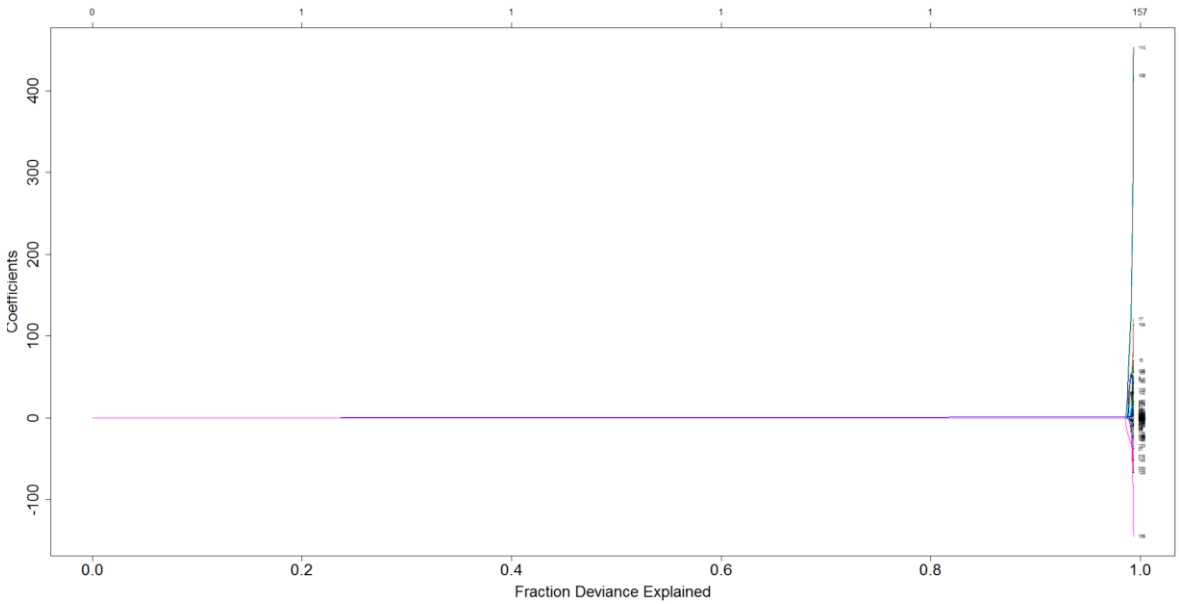


Figure 19. The Percentage of Deviance Explained by the Coefficients

Figure 19 presents the percentage of deviance explained by the model coefficients. Rodriguez (2009) defined deviance as "a measure of discrepancy between observed and fitted values." In other words, it is another statistic measuring the goodness of fit. Compared to the previous plot we can see that an increase in lambda decreases most of the coefficients. We can also see that the coefficients increase when the fraction of deviance explained increases. Therefore, more predictors there are, higher the fraction of deviance

explained is. This is not new as R2 also increases when more predictors are added to the model although this does not always indicate that the model is performing better.

Table 11 presents the coefficients chosen for the model based on the Lasso variable selection. Similarly, to the baseline model we have used elastic net approach with alpha of 0,3 also for the chosen set of variables as the number of variables after the Lasso selection was still relatively large. As we can see most of the independent variables are new compared to the ones in the baseline model. The absolute numbers for the GDP and population have also been replaced by the relative change. Moreover, there are now more variables related to the household finances such as household economy now (Oma talous nyt), household economy in 12 months (Oma talous 12 kuukauden kuluttua), favorable timing for saving money (Ajankohdan otollisuus säästämiseen) and favorable timing for taking loan (Ajankohdan otollisuus lainanottoon). This variable set also contains more demographic variables such as the ratio of jobs related to primary production (Alkutuotannon työpaikkojen osuus), net migration based on the city in persons (Kuntien välinen muuttovoitto tappiohenkilöä) and the ratio of at least 15-year-old persons with at least secondary education (Vähintään keskiasteen tutkinnon suorittaneiden osuus 15 vuotta täyttäneistä). The model also has the same log-transformed predictors as the baseline model.

Table 11: Regression Coefficients Poisson Lasso Regression (Model 2)

Variable	Estimate
Intercept	-4.009e-01
Log(HSP_kappaleet_lag8)	8.118e-01
Väkiluvun_muutos_lag8	1.008e-01
Alkutuotannon_työpaikkojen_osuus_lag8	-5.380e-02
Kiinteistönvälittäjä_haut_lag8	.
Kuntien_välinen_muuttovoitto_tappiohenkilöä_lag8	1.137e-05
Vuokra-asunnoissa_asuvat_lag8	4.570e-03
Oma_talous_Nyt_lag8	-2.399e-03
Oma_talous_12kk_kuluttua_lag8	3.329e-03
Suomen_talous_nyt_lag8	2.391e-03
Ajankohdan_otollisuus_säästämiseen_lag8	-2.018e-03
Ajankohdan_otollisuus_lainanottoon_lag8	8.644e-04
Kotitalouden_rahatilanne_nyt_lag8	.
Kotitalouden_säästämismahdollisuudet_12kk_sisällä_lag8	.
Rahankäyttö_kestotavaroihin_lag8	.
Vähintään_keskiasteen_tutkinnon_suorittaneiden_osuus_15_vuotta_täyttäneistä_lag8	2.626e-03
Palvelujen_työpaikkojen_osuus_lag8	1.623e-02
Log(rakennusaloitukset_normalized_lag_1y +1)	.
Log(HSP_median_myyntiaika_lag8)	.

Sosiaali_ja_terveystoiminta_yhteensänettokäyttökustannuksetеuroa_asukas_lag8	.
BKT_volyymin_vuosimuutos_lag8	1.563e-03
AlueHELSINKI	1.504e-02
AlueJYVÄSKYLÄ	.
AlueKUOPIO	.
AlueLAHTI	-1.928e-02
AlueOULU	-1.559e-02
AlueTAMPERE	.
AlueTURKU	.
AlueVANTAA	.
tyyppiKerros-kolmiot+	2.020e-02
tyyppiKerros-yksiöt	.
tyyppiKerrostalot	1.705e-01
tyyppiRivitalot	-1.393e-02
tyyppiYhteensä	2.232e-01
quarters(vuosineljannes)Q2	2.479e-03
quarters(vuosineljannes)Q3	.
quarters(vuosineljannes)Q4	-4.916e-02

4.4.1.3 Negative Binomial Regression

As discussed in the methodology chapter, one of the assumptions of the Poisson distribution is that the mean and variance are equal (Rodriguez, 2009; Rodriguez 2013). The data is said to be overdispersed if the variance exceeds the mean and underdispersed if the mean exceeds the variance instead (Cameron, 2013). The overdispersion problem often arises when dealing with real-life count data. Procházka (2017) stated that Negative Binomial Regression is one of the most used methods to handle overdispersion. It adds an extra parameter to the variance expression thus enabling more accurate results as the mean and variance do not have to be equal anymore (Reese, 2016). The formula for this adjusted variance is defined as follows:

$$Var(Y) = \frac{pr}{(1-p)^2} = \mu + \frac{1}{r}\mu^2. \tag{12}$$

The mean for our dependent variable is 432,16 and the variance 255662,8. Hence the data is clearly overdispersed. Two alternative Negative Binomial Regression (NBR) models were built from the variable sets introduced previously to take into account the overdispersion. First NBR model used the predictors of the baseline model whereas the second one used the variables selected from the Lasso regression. Almost all of the

coefficients of the first NBR model are statistically significant given the 5% significance level as we can see from Table 12. The AIC for this model is 60821. The coefficients of the second NBR model with the predictors from lasso regression (Table 13) have less statistically significant variables than the first NBR model. In addition, the AIC, 60898, is slightly larger than the AIC of the first model.

Table 12: Regression Coefficients Negative Binomial regression (Model 3)

Variable	Estimate	Std. Error	Z value	Pr(>Z)
Intercept	-3.614e+01	3.409e+00	-10.603	<2e-16
Log(HSP_kappaleet_lag8)	8.467e-01	6.283e-03	134.771	<2e-16
välikuluvun_muutos_lag8	1.427e-01	1.063e-02	13.427	<2e-16
E4_ostoaikomus_lag8	5.761e-03	8.281e-04	6.957	3.470e-12
A1_luottamus_lag8	8.873e-03	3.236e-04	27.415	<2e-16
Kiinteistönvälittäjä_haut_lag8	6.420e-05	6.209e-05	1.034	0.301
Log(Asuntokunnan_käytettävissä_olevat_rahatulot_keskiarvo_lag8)	2.519e+00	1.935e-01	13.015	<2e-16
Vuokra_asunnoissa_asuvat_lag8	5.322e-02	8.533e-03	6.236	4.480e-10
Työllisyysaste_lag8	8.851e-03	4.249e-01	2.231	0.026
Asuntokunnan_keskikoko_henkilöä_lag8	2.708e+00	3.681e-03	6.373	1.860e-10
Neliövuokra_lag8	5.569e-03	2.178e-02	1.513	0.130
Log(Rakennusaloitukset_normalized_lag1y + 1)	-7.499e-02	4.582e-02	-3.443	0.001
Log(HSP_median_myyntiaika_lag8)	1.616e-01	4.582e-02	3.527	0.000
HSP_median_myyntiaika_lag8	-3.459e-03	1.111e-03	-3.114	0.002
BKT_2010_hinnoin_lag8	2.640e-06	2.868e-06	0.921	0.357
AlueHELSINKI	1.185e+00	1.556e-01	7.617	2.600e-14
AlueJYVÄSKYLÄ	1.537e+00	1.584e-01	9.708	<2e-16
AlueKUOPIO	1.584e+00	1.748e-01	9.063	<2e-16
AlueLAHTI	1.772e+00	2.055e-01	8.626	<2e-16
AlueOULU	1.118e+00	1.130e-01	9.888	<2e-16
AlueTAMPERE	1.793e+00	2.003e-01	8.950	<2e-16
AlueTURKU	2.120e+00	2.301e-01	9.213	<2e-16
AlueVANTAA	7.609e-01	7.311e-02	10.407	<2e-16
tyyppiKerros-kolmiot+	-5.393e-02	7.666e-03	-7.035	1.990e-12
tyyppiKerros-yksiöt	-1.242e-02	1.522e-02	-0.816	0.414
tyyppiKerrostalot	1.287e-01	6.622e-03	19.436	<2e-16
tyyppiRivitalot	-4.642e-02	6.390e-03	-7.265	3.740e-13
tyyppiYhteensä	1.784e-01	8.016e-03	22.256	<2e-16
quarters(vuosineljannes)Q2	8.793e-03	4.229e-03	2.079	0.038
quarters(vuosineljannes)Q3	1.692e-02	4.588e-03	3.687	0.000
quarters(vuosineljannes)Q4	-3.939e-02	4.373e-03	-9.007	<2e-16

Table 13: Regression Coefficients Negative Binomial Lasso Regression (Model 4)

Variable	Estimate	Std. Error	Z value	Pr(>Z)
Intercept	4.900e+00	6.497e-01	7.542	4.620e-14
Log(HSP_kappaleet_lag8)	8.558e-01	6.183e-03	138.402	<2e-16
Väkiluvun_muutos_lag8	6.723e-02	1.890e-02	3.558	0.000
Alkutuotannon_työpaikkojen_osuus_lag8	4.637e-02	4.425e-02	1.048	0.295

Kiinteistönvälittäjä_haut_lag8	1.671e-04	8.727e-05	1.915	0.055
Kuntien_välinen_muuttovoitto_tappiohenkilöä_lag8	2.799e-05	1.113e-05	2.514	0.012
Vuokra-asunnoissa_asuvat_lag8	-2.558e-02	6.791e-03	-3.767	0.000
Oma_talous_12kk_kuluttua_lag8	-1.478e-03	9.754e-04	-1.515	0.130
Suomen_talous_nyt_lag8	4.541e-03	1.948e-04	23.308	<2e-16
Ajankohdan_otollisuus_säästämiseen_lag8	-2.758e-03	5.463e-04	-5.049	4.450e-07
Ajankohdan_otollisuus_lainanottoon_lag8	-1.474e-03	3.569e-04	-4.129	3.640e-05
Kotitalouden_rahatilanne_nyt_lag8	1.371e-03	8.183e-04	1.675	0.094
Kotitalouden_säästämismahdollisuudet_12kk_sisällä_lag8	1.433e-03	3.816e-04	3.754	0.000
Rahankäyttö_kestotavaroihin_lag8	-2.390e-03	6.152e-04	-3.886	0.000
Vähintään_keskiasteen_tutkinnon_suorittaneiden_osuus_15_vuotta_täyttäneistä_lag8	-1.653e-03	7.313e-03	-0.226	0.821
Palvelujen_työpaikkojen_osuus_lag8	1.602e-02	3.174e-03	5.047	4.490e-07
Log(rakennusaloitukset_normalized_lag_1y +1)	-5.346e-02	2.253e-02	-2.372	0.018
Log(HSP_median_myyntiaika_lag8)	-1.785e-02	1.183e-02	-1.509	0.131
Sosiaali_ja_terveystoiminta_yhteensänettokäyttökustannukseteuroa_asukas_lag8	7.857e-05	5.021e-05	1.565	0.118
BKT_volyymin_vuosimuutos_lag8	-2.240e-05	1.498e-06	-14.947	<2e-16
AlueHELSINKI	2.329e-01	9.861e-02	2.362	0.018
AlueJYVÄSKYLÄ	6.912e-02	5.682e-02	1.216	0.224
AlueKUOPIO	-3.921e-01	1-788e-01	-2.193	0.028
AlueLAHTI	-1.036e-01	8.109e-02	-1.278	0.201
AlueOULU	-7.362e-02	4.457e-02	-1.652	0.099
AlueTAMPERE	1.256e-01	6.286e-02	1.997	0.046
AlueTURKU	1.342e-01	7.853e-02	1.708	0.088
AlueVANTAA	-7.992e-02	6.326e-02	-1.263	0.206
tyyppiKerros-kolmiot+	-5.044e-02	6.232e-03	-8.093	5.800e-16
tyyppiKerros-yksiöt	-8.748e-03	8.920e-03	-0.981	0.327
tyyppiKerrostalot	1.269e-01	6.608e-03	19.213	<2e-16
tyyppiRivitalot	-3.275e-02	6.141e-03	-5.334	9.630e-08
tyyppiYhteensä	1.761e-01	8.076e-03	21.810	<2e-16
quarters(vuosineljannes)Q2	9.036e-03	4.559e-03	1.982	0.048
quarters(vuosineljannes)Q3	-1.653e-02	5.764e-03	-2.868	0.004
quarters(vuosineljannes)Q4	-5.819e-02	4.283e-03	-13.587	<2e-16

4.4.2 Model Fit and Results

In this section the goodness of fit, both in-sample and out-of-sample, is evaluated for every model. Our training dataset represents the in-sample that is data from the beginning of 2010 until the beginning of 2016 whereas the test dataset represents the out-of-sample that is from the beginning of 2016 until the beginning of 2017 not including it. Hence in-sample only includes actual data that is used to train the model and we do not predict the sales volume for this period of time. In contrast, out-of-sample represents the period of time that we are trying to predict and the predicted values are compared with the actual ones for this set. In addition, the models forecast the future that is observations that are

neither in-sample or out-of-sample. Hence these are the predicted observations after the test dataset ends that is beyond the first quarter of 2017.

R2, the squared R, of the test set is usually the best measure to evaluate numerically how well your model fit the data as discussed before. Another commonly used measure of the model fit is the mean squared error also known as MSE. The formula of MSE is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2. \tag{13}$$

MSE calculates the sum of the squared difference between the actual observation y and the prediction f(x), the mean of the distribution, divided by the number of predictions n. The objective is to minimize the MSE. The smaller MSE is, closer the predictions are to the actual observations. (James et al, 2013; Jokinen, 2016) In other words, MSE is the average of the squared errors that is the difference between the fitted values and predictions. If the number of test observations is large, the test MSE can be calculated as follows:

$$Avg(f(x_0) - y_0)^2, \tag{14}$$

where x_0 and y_0 are test observations. In other words, the test MSE is the average squared prediction error. (James et al, 2013) However, Root Mean Squared Error, RMSE, is often preferred as a measure of fit compared to MSE because it represents the absolute fit having the same scale and unit as the response variable (Hyndman, 2005). RMSE is calculated by taking the square root of MSE and thus it represents the standard deviation of the errors. Similarly, to the MSE, the smaller RMSE means better model fit. However, RMSE, having the same unit as the response variable, should be compared with the same scale as y.

When it comes to models with count data such as Poisson regression or Negative binomial regression r-square is often computed using residual deviance and is the best measure for the model fit compared to the traditional r-square. In fact, r-square was originally developed as a goodness of fit measure for linear regression models without heteroscedasticity. (Cameron & Windmeijer, 1996). Hence we are going to calculate this

deviance R-square for every model. To save from confusion the deviance R-square is referred as R-square in the future. Although Bartlett (2014) stated that deviance itself can be considered as an overall measure of the fit for Poisson regression models, we will not compare the deviances in this study but focus on r-square instead. Cameron & Windmeijer (1996) also found out that the deviance residual R2 is the best one for all standard count data models. Friedman (2010) defined deviance R2 as follows:

$$1 - \frac{Deviance}{Null\ deviance}. \tag{15}$$

Deviance is defined as 2*(loglike_sat - loglike), where loglike_sat is the log- likelihood for the saturated model whereas null deviance is defined as 2*(loglike_sat -loglike(Null)) that is the model with only the intercept (Friedman, 2010) However, this will not be explained in more detail as it is out of the study scope.

4.4.2.1 In-sample Fit and Results

The R2 for the baseline model using the train dataset is 0,9900159 whereas the RMSE for this model is 43,84573. The R2 for the second model, that is Poisson regression using predictor variables selected from Lasso regression, is 0,9896019 and the RMSE is 48,57919.

For the first Negative Binomial regression model the R2 for the train dataset is 0,9861383 while the R2 for the second Negative Binomial regression model using the predictor variables from Lasso regression is 0,9859421. The RMSE for the first NBR model is 39,66451 and for the second one 43,40253. Table 14 presents the minimum, maximum and mean of the target variable, number of houses sold, in the training dataset.

Table 14: Train Dataset Response Variable Diagnostics

Maximum	Minimum	Mean
3178	30	433,40

Table 15 presents the measures for goodness of model fit for each of the models using in-

sample thus train data. We can see that the r-square is really high, over 98% for each model. Therefore, one could say that each model explains at least 98% of the variation in the data. Moreover, the RMSE for Negative Binomial regression models are close to 40 and given that the minimum value of the response variable in the train data is 30 and maximum 3178 this seems to be good value. Although train RMSE could be even better test RMSE is far more important measure in terms of the prediction power and accuracy of predictions (James et al, 2013). The really high train r-squares might lead to overfitting if the test r-squares for the models are clearly lower. This proves again why it is crucial to test the model against unseen out-of-sample dataset and calculate the goodness of fit measures also for the test dataset.

Table 15: Train Dataset Measures for Goodness of Fit

	Poisson Regression (Baseline model)	Poisson-Lasso Regression	Negative Binomial Regression	Negative Binomial-Lasso Regression
R2	99,0%	98,9%	98,6%	98,6%
RMSE	43,85	48,58	39,66	43,40

4.4.2.2 Out-of-sample Fit and Results

The test R2 of the baseline Poisson regression model is 0,9881457 and hence the model explains 98% of the variation in the data. The test R2 for the Poisson regression model using predictor variables from Lasso selection is 0,9892417. Hence the model explains close to 99% of the variation in the data. For the baseline model test RMSE is 54,68 whereas it is 52,09 for the second model. The test R2 of the first NBR model is 0,9841956 thus explaining 98% of the model variance whereas the test R2 of the second NBR model using Lasso variables is even higher, 0,9925287, thus explaining more than 99% of the model variance. The test RMSE of the first NBR model is 63,14 (63,13909) while the test RMSE of the second NBR model is even better, 43,41 (43,41176). The minimum, maximum and mean of the target variable, number of houses sold, in the test dataset are presented in the Table 16 and the measures of model fit in Table 17. We can see that for each model R2 is more than 98%. Furthermore, RMSE for each model is close to the minimum value of the target variable in the test dataset that is 41.

Table 16: Test Dataset Response Variable Diagnostics

Maximum	Minimum	Mean
3081	41	424,68

Table 17: Test Dataset Measures for Goodness of Fit

	Poisson Regression (Baseline model)	Poisson-Lasso Regression	Negative Binomial Regression	Negative Binomial-Lasso Regression
R2	98,8%	98,9%	98,4%	99,3%
RMSE	54,68	52,09	63,14	43,41

The fit of the models given the out-of-sample dataset can also be illustrated with following Figures. We use total sales volume in Helsinki (Figure 20), total sales volume in Tampere (Figure 21) and the sales of studios in Helsinki (Figure 22) to compare the models' actual observations and predictions. Top left is always the baseline Poisson regression whereas top right is the Poisson regression model using the predictor variables from Lasso variable selection. Bottom left is the Negative binomial regression model whereas bottom right is the Negative binomial regression model using the predictor variables from Lasso variable selection. The prediction is with black line whereas the colored line represents the actual values. The out-of-sample data period we are exploring is the one where we have data for both of them. Hence from the beginning of 2016 until the beginning of 2017.

As we can see from Figure 20 the shape of the prediction line for all of the models is similar to the actual observations. However, NBR with Lasso variables (bottom right) has clearly the best fit almost equaling the actual observations. The basic Poisson regression model with Lasso variables (top right) also seems to have the second best fit. It is worth of mentioning that the predictions do not need to be exactly same as the actual observations as long as they are similar. Figure 21 presents the total sales volume in Tampere. Similarly, Negative Binomial regression with Lasso variables (bottom right), has the predictions closest to the actual values. However, the predictions are not as accurate as they were in Helsinki. The other models' prediction accuracy seems to be better now and closer to the

actual observations. For example, baseline Poisson regression model's (top left) prediction is now better and a little bit closer to the actual observations than in Helsinki although still not the best.

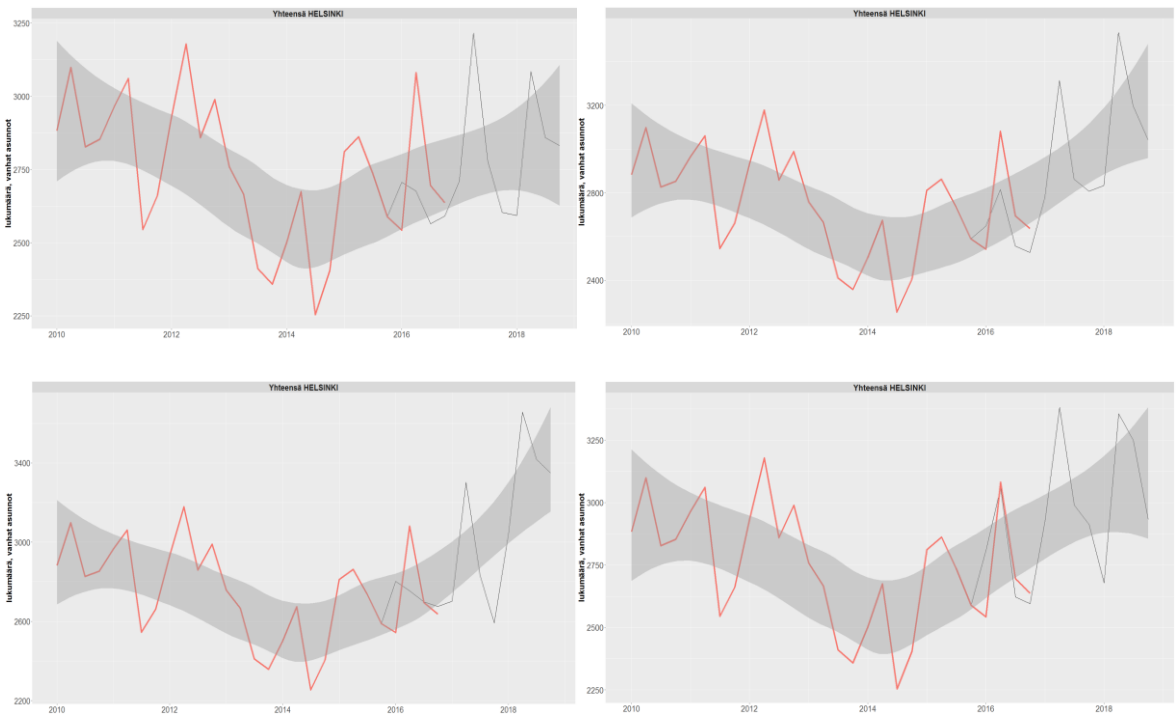


Figure 20. Total Sales Volume of Old Apartments in Helsinki

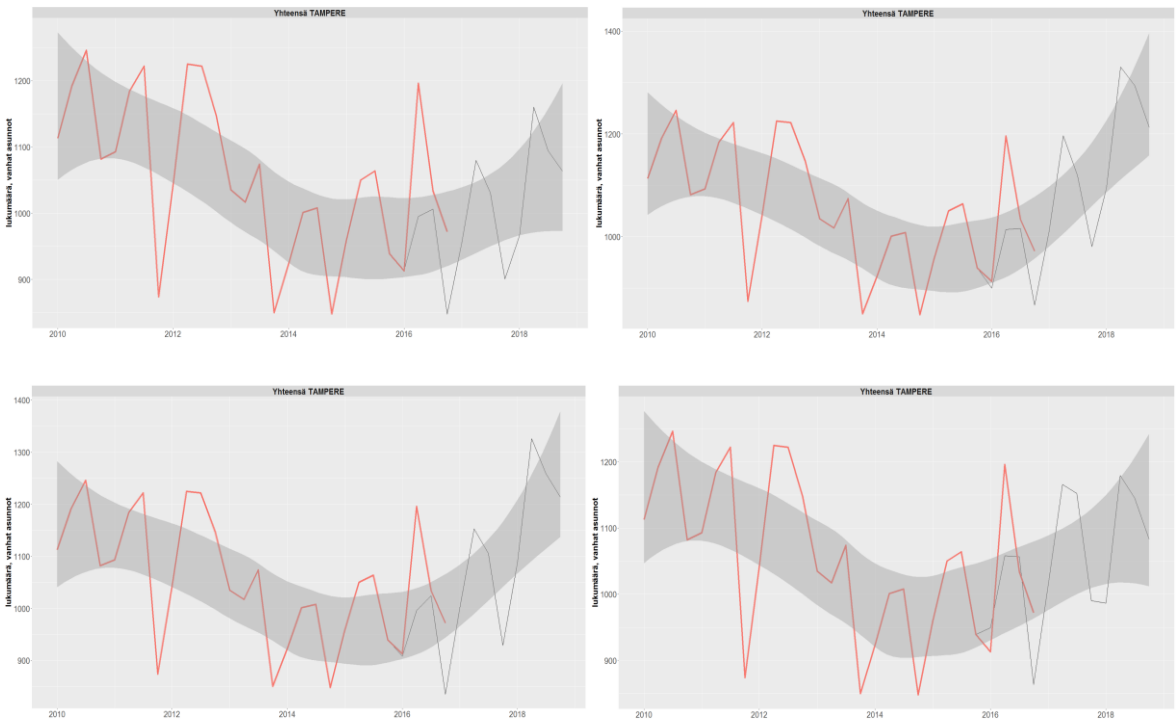


Figure 21. Total Sales Volume of Old Apartments in Tampere

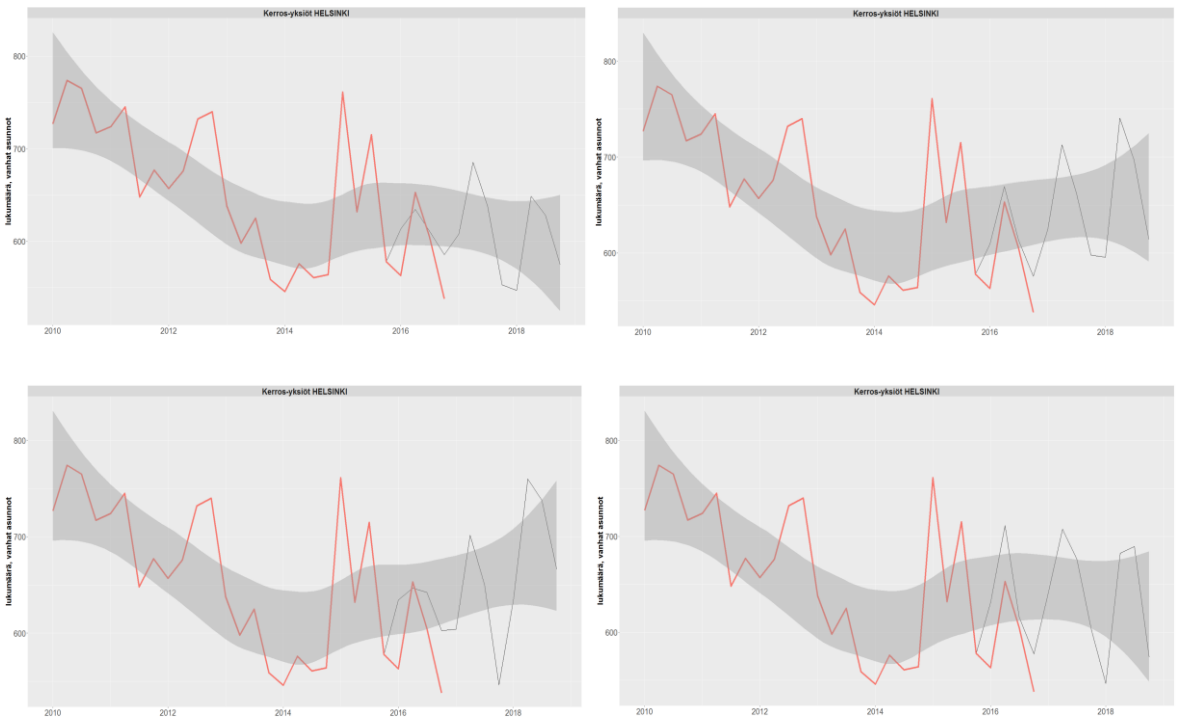


Figure 22. The Sales Volume of Old Studios in Helsinki

Figure 22 presents the sales volume of studios in Helsinki. We have taken it as an example to see whether the type increases the model's prediction accuracy instead of predicting the total sales volume regardless of the type. Now the out-of-sample is better for every model compared to the predictions in the Figure 21. In fact, Poisson regression model with Lasso variables (top right) seems to have now the best prediction closest to the actual observations although Negative Binomial regression model with lasso variables (bottom right) is close second. The other two models without Lasso predictor variables also have quite good predictions compared to the actual observations in the test dataset. The numerical goodness of fit of every model is compared against each other in the next chapter based on which the best model is selected for further analysis of the results.

5 Discussion and Analysis

In this chapter the results and goodness of fit of the models are discussed and further analyzed. The best model will be selected based on the fit measures and the predictor importance is calculated for this model. Finally, the hypotheses and research questions introduced in the beginning of this study are answered based on the discussion and analysis.

5.1 Best Goodness of Fit

Table 18 presents both train and test r-square and RMSE statistics for each model. Moreover, the maximum, minimum and mean of the response variable is presented for both datasets. As we can see the models' r-squares are very close to each other and really high, at least 98% for both train and test models. This could be due to the fact that, regardless of trying to limit the number of predictors, multiple variables were still included to the model even after Lasso selection. We will elaborate this more in the next chapter. As discussed in the methodology chapter more predictors there are higher the r-square usually is. However, we cannot use adjusted r-square to compare the models as they differ from each other based on the predictors and the approach. Moreover, test r-squares are close to train r-squares implying that there are not issues with overfitting. The test r-square for the Negative Binomial regression model using Lasso predictor variables has even larger test r-square 99,3% than its train r-square 98,6%.

RMSE values of the models were also calculated. The train RMSE statistics are here to show that they are relatively larger than test RMSE values of the models. However, train RMSE is not remotely as important measure of prediction accuracy than test RMSE is (James et al, 2013). As we can see from Table 16 the mean for our target variable in the test dataset is 424,68 and the minimum value thus the number of houses sold is 41. The scale of the RMSE is same as the scale of the response variable. Hence closer the test RMSE is to the minimum response variable value better the model fit is. (Hyndman, 2005). As we can see the models using predictors chosen based on the Lasso variable selection have better test RMSE statistics than the other two models including the baseline model. Even though Poisson-Lasso regression model has a good test RMSE of 52,09, the test RMSE of Negative Binomial-Lasso regression, 43,41 is even better and really close to the

minimum value of 41. All together we can see that this model has also the highest test r-square and therefore could be considered as the best model for prediction from the models tested. Hence the results of the Negative Binomial-Lasso Regression model will be discussed and analyzed more closely. Therefore, every chart and analysis in the future are based on the results of this model.

Table 18: Model Results

	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
	Poisson Regression Baseline	Poisson Regression Baseline	Poisson-Lasso Regression	Poisson-Lasso Regression	NBR	NBR	NBR Lasso	NBR Lasso
R2	99,0%	98,8%	98,9%	98,9%	98,6%	98,4%	98,6%	99,3%
RMSE	43,85	54,68	48,58	52,09	39,66	63,14	43,40	43,41
Max	3178	3081	3178	3081	3178	3081	3178	3081
Min	30	41	30	41	30	41	30	41
Mean	433,4	424,7	433,4	424,7	433,4	424,7	433,4	424,7

5.2 Predictions (H1, H2, H3)

We come back to the hypotheses created in the beginning of the study. The first three hypotheses were related to the forecasting the sales volume and conditions in the Finnish real estate market. They are now tested against the results from the Negative Binomial regression with Lasso predictors that was chosen as the best model based on its fit.

First hypothesis (H1) stated that *the number of sold old apartments in total will increase within the next 12 months.* This was based on the existing studies and literature covered in the literature review (Brotherus, 2017; Kekäläinen, Tähtinen & Vuori, 2017; Vehviläinen, 2016). Our results show that in overall the sales volume for old apartments will grow within the next 12 months given the cities and apartment types taken into account in this study. Figure 23 presents the total sales volume of old apartments in Helsinki, Tampere and Turku. The reason these three cities are highlighted comes from the literature. Brotherus (2017) mentioned in his latest report that the demand for old apartments has

increased in the capital region, Tampere and Turku during 2017. We have used a function called “geo smoothing” to better visualize the trend and the direction of the sales volume. We can clearly see that the sales volume will increase for each of the three cities based on the forecast. The strongest growth will be in Helsinki.

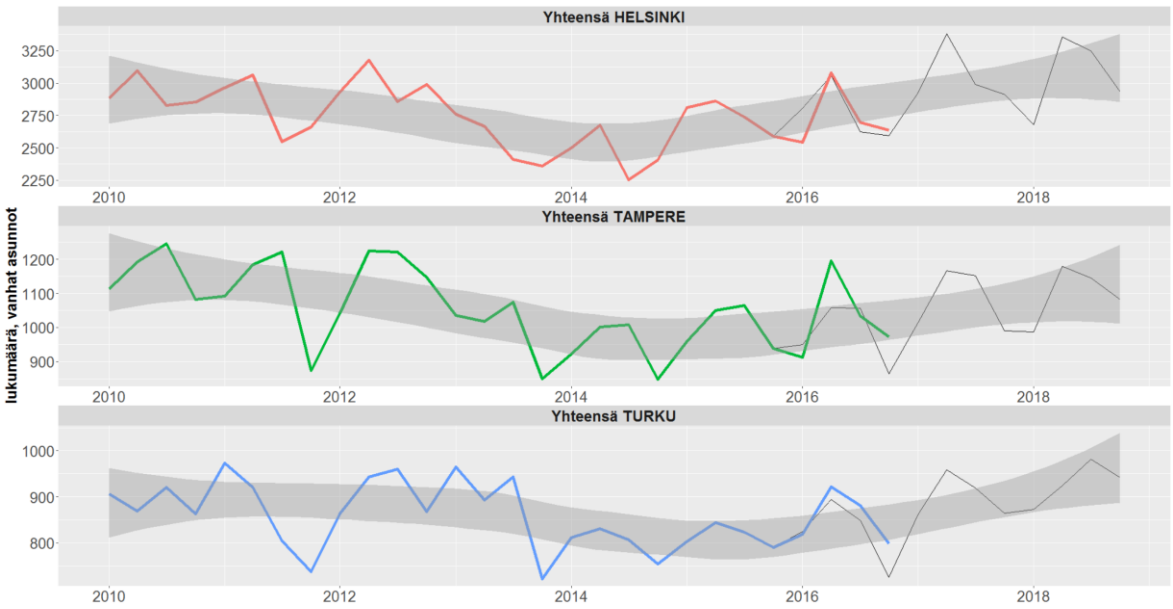


Figure 23. Total Sales Volume of Old Apartments in Helsinki (red), Tampere (green) and Turku (blue)

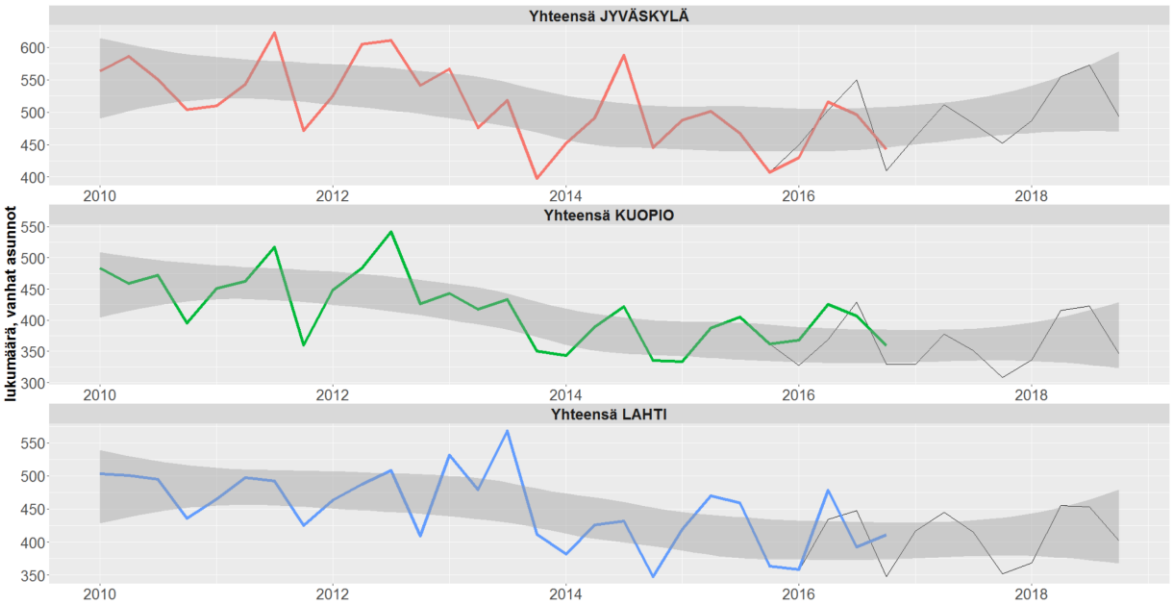


Figure 24. Total Sales Volume of Old Apartments in Jyväskylä (red), Kuopio (green) and Lahti (blue)

Figure 24 shows the total sales volume of old apartments in Jyväskylä, Kuopio and Lahti that are considered as middle-sized cities. In fact, Brotherus (2017), also emphasized that although the growth has been strong in bigger cities there is a gap between the bigger cities and middle-sized ones where the sales volume has not been increasing as much. We can

clearly see this from the Figure 24 where the forecast predicts only a slight increase in sales volume of old apartments in Jyväskylä while the sales volume in Kuopio and Lahti will stay more or less the same. Therefore, overall we could say that our first hypothesis is true and that the total sales volume of old apartments will increase within the next 12 months although the growth will vary based on the location focusing on bigger cities.

The second hypothesis (H2) was as follows: *The sales volume for old apartments will increase more in the capital region (Helsinki, Espoo and Vantaa) than in other regions.* As Vehviläinen (2016) proved in his study, the capital region has distinguished itself from other cities in terms of sales volume mostly due to the urbanization. Overall the capital region has been in its own category due to very strong sales volume growth in Helsinki that has even separated itself from other cities with strong growth numbers such as Tampere and Turku. (Brotherus, 2017) Figure 25 presents the total sales volume of old apartments in every city of this study. As we can see the forecasted numbers and growth is on another level in Helsinki. Turku and Tampere have the second strongest growth within the next 12 months based on the forecasts. Espoo is also forecasted to increase its sales volume of old apartments and to sell close to the same numbers as Tampere. Vantaa will also increase its sales volume of old apartments although more moderately than the other cities in the capital region. Therefore, we could say that the second hypothesis that the sales volume for old apartments will increase more in the capital region than in other regions is true even though it is mostly due to the extreme growth in Helsinki.

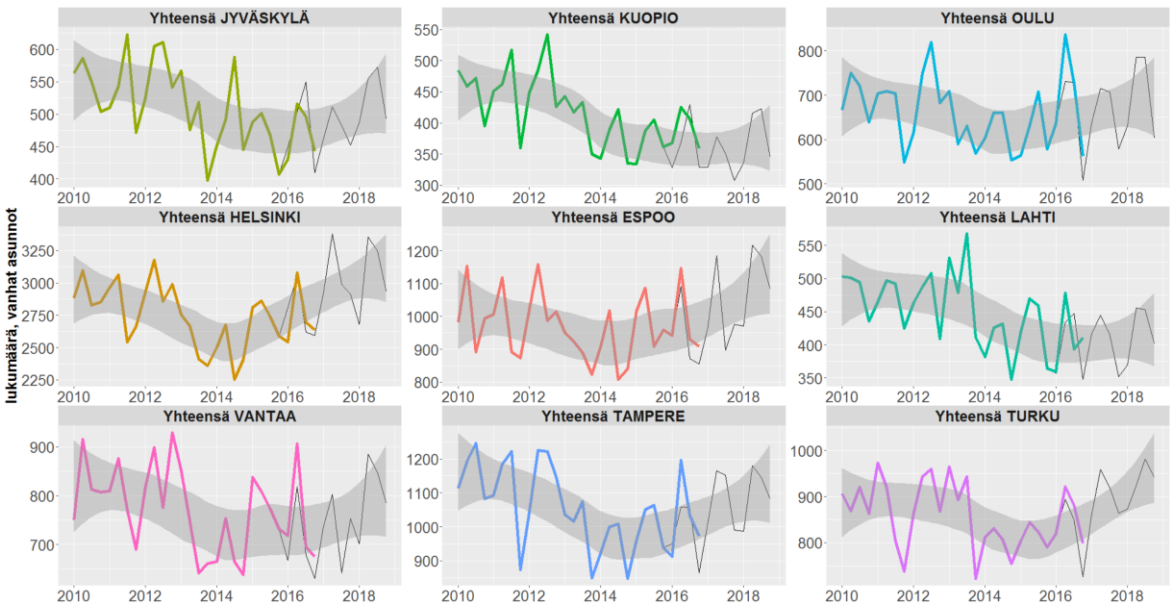


Figure 25. Total Sales Volume of Old Apartments (from top to bottom) in Jyväskylä, Kuopio, Oulu, Helsinki, Espoo, Lahti, Vantaa, Tampere and Turku

The third hypothesis (H3) argued that *the sales volume for smaller studio apartments will increase more than for other apartment types.* The supply of smaller apartments has increased in contrast to bigger apartments mainly because of the changes in demographics such as the household size or young people's willingness to live alone (Vehviläinen, 2016) The urbanization has also been huge and for example young people tend to buy their first apartments mostly from the biggest cities. In fact, the capital region, Tampere and Turku cover now over half of the sales volume when it comes to the first apartment sales. The first apartment is usually a smaller one. (Brotherus, 2017) Hence Helsinki was chosen as the example city to analyze and test this phenomenon and hypothesis. Figure 26 shows the sales volume of old row houses and flats from apartment buildings. One could assume that row houses are larger in square meters in general. Surprisingly the forecasted sales volume will grow almost at the same rate for both apartment types and even slightly more for row houses. Therefore, the types of apartment building flats will be analyzed further. In this study, there are three flat types: studios having one room, two-room flats and flats with at least three rooms.

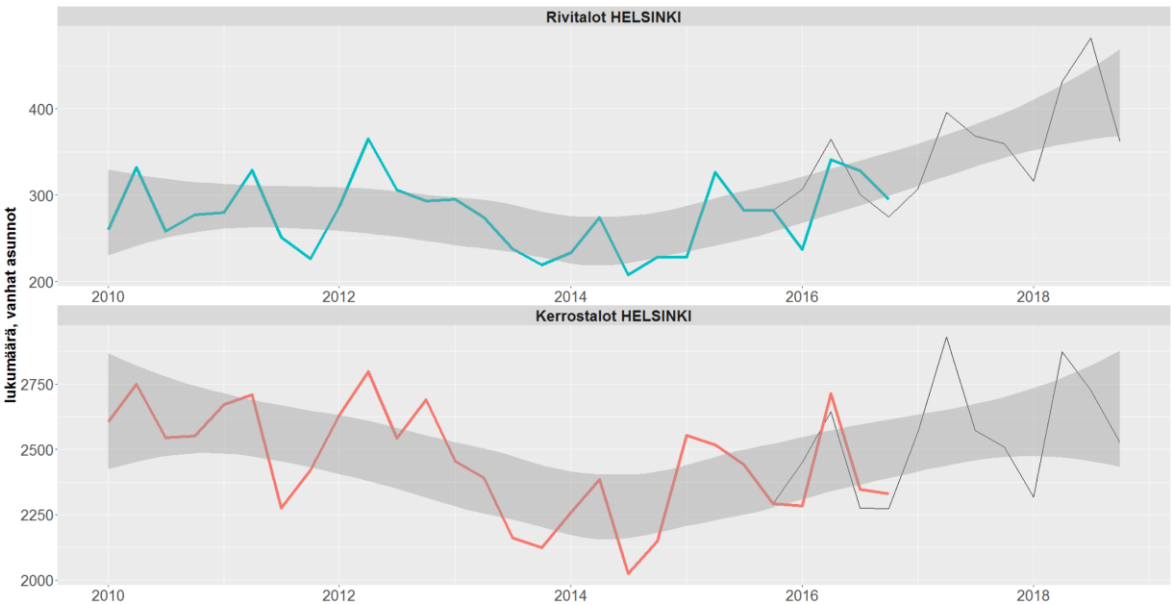


Figure 26. The Sales Volume of Old Row Houses (turquoise) and Apartments (red) in Helsinki

Figure 27 presents the sales volume of these apartment types in Helsinki. We can clearly see that the sales volume of studios in Helsinki will more or less stay the same if not slightly decrease whereas the sales volume for two-room flats and flats with at least three rooms will increase. The sales volume of two-room flats in Helsinki will increase most based on the forecast and two-room flats are also clearly selling more than the other two

flat types. Therefore, we could say that the third hypothesis, the sales volume for smaller studio apartments will increase more than for the other apartment types, is false. The sales volume of smaller apartments will increase in terms of two-room flats but the sales volume of studios won't increase as much. However, we have to remember that we only used the forecasts for Helsinki and that the overall sales volume of studios could increase when taking into account also the forecasts for other cities.

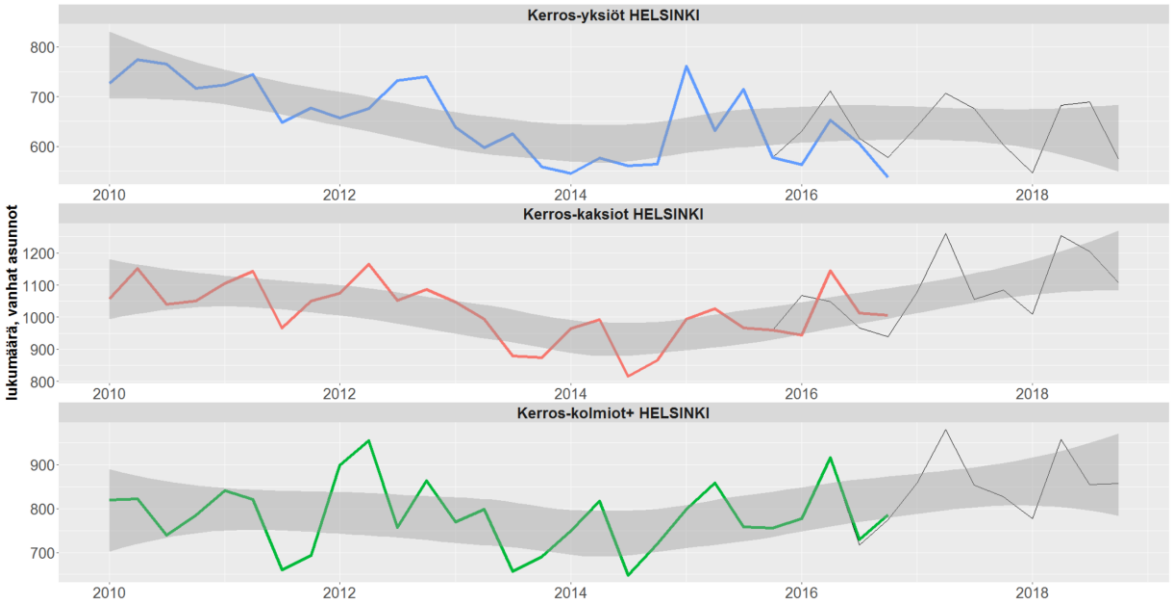


Figure 27. The Sales Volume of Old Studios (blue), Two-room Flats (red) and Three-room Flats (green) in Helsinki

5.3 Predictor Importance (H4, H5)

The fourth and fifth hypotheses were related to the predictor importance when predicting the sales volume of old apartments:

(H4) *The economic variables have the biggest impact on the number of houses sold.*
(H5) *Search query data from Google Trends enhances the model and serves as an important predictor variable.*

Most of the real estate demand and supply factors discussed in the literature review are related to economy or household finance. (Pirounakis, 2013; Ostamo, 1997). Both CFO and CEO of Kiinteistömaailma thought that the trust or distrust in economy is one of the most important factors for real estate markets. Choi & Varian (2009) were able to improve R2 when adding search query data to their models. Based on all the prior studies, the

statistical and predictive models' accuracy can be improved by combining real-time data concerning consumer's online search behavior and actual data about the past development of the real estate market (Norros, 2014). In order to test these two hypotheses, the predictor importance for the chosen model, Negative Binomial regression with Lasso predictor variables, was calculated using function called "varimp". Table 19 presents the results where the numbers represent the scaled importance of the predictors. Higher the number is more important the predictor variable is in this model and for this target variable. Table 20 has ranked the variables in decreasing order based on their predictor importance, first being the most important predictor. Not surprisingly, the log transformed number of sold houses from HSP is the most important predictor with a huge margin. This is because it basically represents the same variable as the target variable itself but is included to the model as a predictor because the data is more accurate and on time in HSP compared to the data in Statistics Finland. Statistics Finland however has more data of all transactions. Hence the number of houses sold from Statistics Finland was our target variable and the number of houses sold from HSP just a predictor trying to take into account the delays in the data.

When ignoring it we can see that the most important predictor is actually Finnish economy now (Suomen talous nyt lag8). In fact, half of the variables in the top 10 are somewhat economical or related to finance whereas the other half are related to the apartment type or period of time. Therefore, we could say that the fourth hypothesis is true and economic and household finance variables are at least among the most important ones if not the most important variable category for predicting sales volume of old apartments. Moreover, the model's variables were chosen based on the Lasso variable selection so Lasso itself chose variables from this category.

Originally the initial variable set included three variables based on the internet search query data. After Lasso selection only one them remained in the model, search volume for "Real estate agent" as the search term. We can see that this variable is ranked as 23rd most important predictor that is not as high as one would expect. However, it still remained in the model and thus enhanced the model performance and prediction accuracy. One could wonder if more similar variables were included to the model how many of them would have been important predictors. Therefore, we could say that the fifth and last hypothesis that search query data from Google Trends enhances the model and serves as an important

predictor variable is only partly true. This hypothesis should be still further studied in future.

Table 19: Predictor Importance

Predictor	Relative Importance
Log(HSP_kappaleet_lag8)	138.402
Väkiluvun_muutos_lag8	3.558
Alkutuotannon_työpaikkojen_osuus_lag8	1.048
Kiinteistönvälittäjä_haut_lag8	1.915
Kuntien_välinen_muuttovoitto_tappiohenkilöä_lag8	2.514
Vuokra-asunnoissa_asuvat_lag8	3.767
Oma_talous_12kk_kuluttua_lag8	1.515
Suomen_talous_nyt_lag8	23.308
Ajankohdan_otollisuus_säästämiseen_lag8	5.049
Ajankohdan_otollisuus_lainanottoon_lag8	4.129
Kotitalouden_rahatilanne_nyt_lag8	1.675
Kotitalouden_säästämismahdollisuudet_12kk_sisällä_lag8	3.754
Rahankäyttö_kestotavaroihin_lag8	3.886
Vähintään_keskiasteen_tutkinnon_suorittaneiden_osuus_15_vuotta_täyttäneistä_lag8	0.226
Palvelujen_työpaikkojen_osuus_lag8	5.047
Log(rakennusaloitukset_normalized_lag_1y +1)	2.372
Log(HSP_median_myyntiaika_lag8)	1.509
Sosiaali_ja_terveystoiminta_yhteensänettokäyttökustannukseteuroa_asukas_lag8	1.565
BKT_volyymin_vuosimuutos_lag8	14.947
AlueHELSINKI	2.362
AlueJYVÄSKYLÄ	1.216
AlueKUOPIO	2.193
AlueLAHTI	1.278
AlueOULU	1.652
AlueTAMPERE	1.997
AlueTURKU	1.708
AlueVANTAA	1.263
tyyppiKerros-kolmiot+	8.093
tyyppiKerros-yksiöt	0.981
tyyppiKerrostalot	19.213
tyyppiRivitalot	5.334
tyyppiYhteensä	21.810
quarters(vuosineljannes)Q2	1.982
quarters(vuosineljannes)Q3	2.868
quarters(vuosineljannes)Q4	13.587

Table 20: Predictor Importance in Order

```
 [1] "log(HSP_kappaleet_lag8)"
 [2] "Suomen_talous_nyt_lag8"
 [3] "tyyppiYhteensä"
 [4] "tyyppiKerrostalot"
 [5] "BKT_2010_hinnoin_lag8"
 [6] "quarters(vuosineljannes)Q4"
 [7] "tyyppiKerros-kolmiot+"
 [8] "tyyppiRivitalot"
 [9] "Ajankohdan_otollisuus_säästämiseen_lag8"
[10] "Palvelujen_työpaikkojen_osuus_lag8"
[11] "Ajankohdan_otollisuus_lainanottoon_lag8"
[12] "Rahankäyttö_kestotavaroihin_lag8"
[13] "Vuokra_asunnoissa_asuvat_lag8"
[14] "Kotitalouden_säästämismahdollisuudet_12kk_sisällä_lag8"
[15] "vakiluvun_muutos_lag8"
[16] "quarters(vuosineljannes)Q3"
[17] "Kuntien_välinen_muuttovoitto_tappiohenkilöä_lag8"
[18] "log(rakennusaloitukset_normalized_lag_1y + 1)"
[19] "AlueHELSINKI"
[20] "AlueKUOPIO"
[21] "AlueTAMPERE"
[22] "quarters(vuosineljannes)Q2"
[23] "Kiinteistönvälittäjä_haut_lag8"
[24] "AlueTURKU"
[25] "Kotitalouden_rahatilanne_nyt_lag8"
[26] "AlueOULU"
[27] "Sosiaali_ja_terveystoiminta_yhteensänettokäyttökustannukseteuroa_asukas_lag8"
[28] "Oma_talous_12kk_kuluttua_lag8"
[29] "log(HSP_median_myyntiaika_lag8)"
[30] "AlueLAHTI"
[31] "AlueVANTAA"
[32] "AlueJYVÄSKYLÄ"
[33] "Alkutuotannon_työpaikkojen_osuus_lag8"
[34] "tyyppiKerros-yksiöt"
[35] "Vähintään_keskiasteen_tutkinnon_suorittaneiden_osuus_15_vuotta_täyttäneistä_lag8"
```

5.4 Answering the Research Questions

This study tried to answer three research questions presented in the introduction chapter:

1) What factors/input variables to involve when predicting the real estate sales volume, more accurately the sales volume of old apartments, in Finland?

2) What modeling method will give the best result when predicting the sales volume of old apartments for the next 12 months given the nature of the data?

3) How does the sales volume of old apartments differ based on the apartment's location and type?

5.4.1 Variables for Predicting Real Estate Sales Volume

The first question was partly answered in terms of the theoretical framework that was built based on prior literature and interviews in turn. The Real Estate market and thus the sales volume of old apartments is mostly affected by the housing demand and supply. Housing demand can be further divided into endogenous and exogenous factors. Endogenous factors include variables directly related to the price of the house whereas exogenous factors are not related or are at least indirectly related to the housing price. Exogenous factors have four major categories: 1) Demographic, 2) Living costs, 3) Economy and 4) Housing Preferences. Based on the interviews economic factors are the most important ones although all of the categories together form a strong basis for the prediction. Housing supply can be divided into two major factors that are new construction and household owner decisions. For, instance the normalized amount of new constructions has been included to the baseline model as a predictor. However, it is not as important predictor in this study as it could have been in others given the fact that the target variable is the number of old houses sold, old referring to the houses that have been constructed more than 10 years ago. As discussed in the literature review the new construction also only covers roughly just a couple of percent of the whole housing stock (Siikanen & Tyrkkö,1993). In addition to housing supply and demand, the third major factor affecting the real estate sales is the internet search queries. As discussed before, prior studies show that the statistical and predictive models' accuracy can be improved by adding real-time data to the model (Norros, 2014). Hence, in this study, search volumes for specific search terms from Google Trends were added to the model. The following terms were chosen based on the interviews: 1) Myytävät asunnot (the apartments for sale), 2) Kiinteistönvälitys (Real estate brokerage) and 3) Kiinteistönvälittäjä (Real estate agent).

Selecting predictors to the model from the three main categories of real estate determinants should be able to create a sufficient pool of independent variables for a robust model. However, it is still important to perform statistical tests and variable selection even after choosing the initial variables in order to reduce the number of predictors and thus avoiding skewing R2. Moreover, it is crucial to include and at least somehow take into account the future of the economy as it is closely correlated with the real estate market and sales. This study tried to solve this problem by adding the GDP forecast as a predictor to the model in addition of having a lot of economic variables.

All in all, based on the results of this study, there should always be variables from each main category when predicting the sales volume of old apartments. The most important predictor variables for our model are either economical or financial ones. Demographic variables were also well represented when ranking the predictors based on their importance. Moreover, constructions or more accurately, construction starts is an important predictor of sales volume and a determinant of housing supply. However, the number of constructions would be a more crucial predictor when predicting the sales volume of new apartments instead of old apartments. Finally, as discussed before, combining real-time data about buyer's online behavior with other data seems to enhance the model performance. Internet search query data is easy and straightforward way to integrate real-time data to the model.

5.4.2 Best Model for Prediction

Based on the statistics presented in Table 18, Negative Binomial Regression seems to be the best modelling approach for predicting the number of old houses sold. It also had the best out-of-sample predictions in terms of how close they were to the actual observations when visually comparing the models. Although Poisson regression is a relevant and good approach to predict count data that our target variable represents, it does not take into account the possible overdispersion or underdispersion. Furthermore, it seemed that the best method was to combine Lasso's variable selection to the Negative Binomial Regression. As there was no alpha to determine the weight of elastic net, the variables selected by the Lasso regression were used as predictors in this model. This model clearly had the best fit in terms of RMSE of the test data and also had the highest test R2.

It is impossible to generalize that negative binomial regression with predictor variables form Lasso variable selection is always the best modelling approach when predicting the sales volume of old apartments. However, we could say that it is a method that need to be taken into account when predicting similar target variable or building similar models in the future.

5.4.3 The Differences between the Location and Apartment Type

The differences between the cities could be best explored when looking at the total sales per city in Figure 25. As we discussed before with the first hypothesis the total sales volume of old apartments has been growing and will also grow in the near future regardless of the location or type based on our forecast. The growth will be strongest in the capital region, Tampere and Turku whereas the total sales volume for old apartments will only be moderate in Oulu and Jyväskylä. In Lahti and Kuopio, the total sales volume of old apartments will remain more or less the same.

According to our forecast, the sales volume of two-room flats will be growing most during the next 12 months. Surprisingly the sales volume of smaller studios will not be increasing as much. However, we have to remember that we only used the forecasts for Helsinki and that the overall sales volume of studios could increase when taking into account also the forecasts for other cities.

6 Managerial Implications and Limitations

6.1 Limitations and Future Studies

There are a lot of details that could be improved in this study. For instance, this study only focused on the regression as the primary predicting method. Future studies could emphasize machine learning algorithms such as XG Boost or combinations such as Random Forest instead when predicting the number of houses sold. Furthermore, when predicting the number of houses sold and real estate market it is crucial to be able to predict the economy as well. This has been somehow taken into account also in this research by including the predicted GDP and other economic variables. However, future studies should first focus more on forecasting the economy before forecasting the real estate market as the two are so correlated. This was mostly ignored in the scope of this research because of the lack of time, resources and relevant skills. Another issue for development is the selection of variables. Although this study addressed to this problem quite well by performing analytical variable selection such as building Lasso model and calculating VIF scores, there are still a lot of statistical methods, tests and procedures that could be used. Moreover, more sophisticated ways could have been used to compute the predictor importance such as dominance analysis that is currently trending.

In this study the predicted variable was the absolute number of properties sold for given cities and property types. An alternative way to define the question would be to predict the relative number of properties sold. For example, dividing the number of properties sold by the number of properties sold during the previous full year (four quarters) Furthermore, the features used were mostly absolute time series values. Sometimes it could be beneficial to use differencing for the features, subtracting or dividing the value by the previous year's value.

6.2 Managerial Implications

From the managerial perspective the results of this study could be used to improve decision making based on the data. All of the four models had good fit and hence all of them could be used for prediction. However, I would suggest to use the results as a basis

for building an even better and robust model with more statistical test and variable analysis behind it before implementing it to the business use. Some suggestions for similar future studies or projects can still be given.

First of all, when predicting the sales volume of apartments, it is recommended to choose the variables from the main categories discussed before. Moreover, the focus should be on predictors related to household demographics and finances or the whole economy. For this, it is important that the managers are gathering relevant and up-to-date data for example to a data warehouse, easy to access, when building the model or forecasting. The data for this study was from multiple sources and was even updated once during the process. Second, shrinkage method is recommended in order to reduce the number of predictors when dealing with data with many features but few observations. Moreover, other statistical tests and correlation analysis is highly suggested in order to take into account possible issues with the data. When it comes to sales volume, the data is often non-negative integers thus count data. Based on our results, it is recommended to use either Poisson regression or Negative Binomial regression or another generalized linear model when predicting similar target variable. Finally, it is important to constantly improve the model whether it is based on new information on possible predictors or other factors affecting the results. Hence it is crucial to know the industry and be aware of new disruptions and innovations.

References

Aaron, Ng. (2015) Machine Learning for a London Housing Price Prediction Mobile Application. Available from: http://www.doc.ic.ac.uk/~mpd37/theses/2015_beng_aaron-ng.pdf [Accessed on 14 July 2017]

Agresti, A. (2013) *Categorical data analysis*. 3rd ed. Hoboken, NJ: Wiley.

Arnott, R. (1987) Economic Theory and Housing. Handbook of Regional and Urban Economics vol. II. Elsevier Science Publishers. s. 1322. ISBN: 978-0-444-87970-7
Askitas, N & Zimmermann, K. (2011) Detecting Mortgage Delinquencies. *Discussion paper series,* 5895: 16. Available from: http://ftp.iza.org/dp5895.pdf [Accessed on 14 July 2017]

Bartlett, J. (2014) Deviance goodness of fit test for Poisson regression. Available from: http://thestatsgeek.com/2014/04/26/deviance-goodness-of-fit-test-for-poisson-regression/ [Accessed on 14 July 2017]

Bartlett, R. (2017) *Nowcasting: Economic Forecasting Meets Just-In-Time Delivery.* Available from: http://ifsd.ca/en/blog/last-page-blog/nowcasting-just-in-time-delivery [Accessed on 14 July 2017]

Bewick, V., Cheek, L. & Ball, J. (2003) 'Statistics review 7 Correlation and regression'. *Critical Care*, 7(6), 451. Available from: https://link.springer.com/content/pdf/10.1186/cc2401.pdf [Accessed on 14 July 2017]

Brooks, C. (2008) *Introductory Econometrics for Finance.* 2nd ed. Cambridge.

Brotherus, J. (2011), Kauppakorkeakoulu, Economics, S. o., Economics, D. o., laitos, T., University, A. & Aalto-yliopisto. 2011. *Asuntojen hinnanmuodostus hedonisella menetelmällä*.

Brotherus, J. (2017) *Hypon Asuntomarkkinakatsaus marraskuu 2017.* Hypo 4(4) http://www.hypo.fi/wp-content/uploads/2017/11/Hypon_Asuntomarkkinakatsaus_marraskuu2017.pdf

Brown, J., Song, H. & McGillivray, A. (1997) Forecasting UK house prices: A time varying coefficient approach. *Economic Modelling*, 14(4): 529-548. Available from: http://www.sciencedirect.com/science/article/pii/S0264999397000060 [Accessed on 14 July 2017]

Brownlee, J. (2016) '*Supervised and Unsupervised Machine Learning Algorithms'.* Available from: https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/ [Accessed on 14 July 2017]

Case, K.E., Quigley, J. M. & Shiller, R. 2005, Comparing wealth effects: The stock market versus the housing market, *Advances in Macroeconomics* 5, 1.

Campbell, J. & Cocco, J. (2007) How do house price affect consumption? Evidence from micro data. *Journal of Monetary Economics*, 54(3): 591-621. Available from:

http://www.sciencedirect.com/science/article/pii/S0304393206001279 [Accessed on 14 July 2017]

Cameron, C. (2013) 'Count Data Regression Made Simple'. Available from: http://cameron.econ.ucdavis.edu/racd/simplepoisson.pdf [Accessed on 14 July 2017]

Cameron, C.A. & Windmeijer, F. R-Squared Measures for Count Data Regression Models with Applications to Health Care Utilization. *Journal of Business & Economic Statistics* Vol. 14, No. 2 (Apr., 1996), pp. 209-220

Chamberlin, G. (2010) Googling the present. 37 s. *Economic & Labour Market Review*, 4(12): 59-95. Available from: http://www.ons.gov.uk/ons/rel/elmr/economic-and-labour-market-review/no--12-- december-2010/googling-the-present.pdf [Accessed on 14 July 2017]

Chen, G. (2008) *Encyclopedia of Statistics in Quality and Reliability.* John Wiley & Sons.

Chica-Olmo, J., Cano-Guervos, r. & Chica-Olmo, M. (2013) A Coregionalized Model to Predict Housing Prices. Urban Geography 34:3, pages 395-412.

Choi, H. & Varian, H. (2009) *Predicting the Present with Google Trends.* Available from: https://static.googleusercontent.com/media/www.google.com/en//googleblogs/pdfs/google_predicting_the_present.pdf [Accessed on 14 July 2017]

Coxe, S., West, S.G. & Aiken, L.S. *The Analysis of Count Data: A Gentle Introduction to Poisson Regression and its alternatives.* J Pers Assess, 91(2):121-36. [Accessed on 14 July 2017]

Dipasquale, D. The Journal of Real Estate Finance and Economics (1999) 18: 9. doi:10.1023/A:1007729227419

Durrant, G. (2016) *Poisson Regression Modes for Count Data.* Available from: https://www.slideshare.net/synchrony/poisson-regression-models-for-count-data-63688148 [Accessed on 14 July 2017]

Fan, Z. (2016) Lecture 27-Poisson Regression. *Statistics 200: Introduction to Statistical Inference.* Avilable from: https://stats200.stanford.edu/Lecture27.pdf [Accessed on 14 July 2017]

Friedman, J. (2010) Package "glmnet". Available from: http://cran.revolutionanalytics.com/web/packages/glmnet/glmnet.pdf [Accessed on 14 July 2017]

Frost, J. (2015) *The Danger of Overfitting Regression Models.* Available from: http://blog.minitab.com/blog/adventures-in-statistics-2/the-danger-of-overfitting-regression-models [Accessed on 14 July 2017]

Glen, S. (2016) '*ADF- Augmented Dickey Fuller Test'*. Available from: http://www.statisticshowto.com/adf-augmented-dickey-fuller-test/ [Accessed on 14 July 2017]

Gottlieb, M. (1976) Long Swings in Urban Development, New York: Columbia University Press for NBER.

Hair, J.F.Jr., Anderson, R.E., Tatham, R.L. & Black, W.C. (1995) Multivariate Data Analysis (3rd ed). New York: Macmillan.

Hannonen, M. (2015) A Field Theory of House Prices: An Empirical Study of the Helsinki Submarket

Hastie, T. & Qian, J. (2014) 'Glmnet Vignette'. Available from: https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html [Accessed on 14 July 2017]

Heiskanen, V. (2008) *Asuntomarkkinoiden suhtanteet ja epätasapainotekijät.* Espoo: Teknillinen korkeakoulu.

Hohenstatt, R. & Kaesbauer, M. (2014) GECO's Weather Forecast for the U.K. Housing Market: To What Extent Can We Rely on Google Econometrics? *Journal of Real Estate Research* 36 (2): 253-282. [Accessed on 14 July 2017]

Hurlin, C. (2013) 'Chapter 2: The Multiple Linear Regression Model.' *Advanced Econmterics-Lecture Notes.* [Accessed on 14 July 2017]

Hyndman, R. (2005) *'Another look at measures of forecast accuracy'.* Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.154.9771&rep=rep1&type=pdf [Accessed on 14 July 2017]

Jain, A. (2016) *A Complete Tutorial on Ridge and Lasso Regression in Python.* Available from:https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/#four [Accessed on 14 July 2017]

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) *An Introduction to Statistical Learning with Applications in R.* Springer.

Jansen, S., Coolen, H. & Goetgeluk, R. (2011) *The Measurement and Analysis of Housing Preference and Choice.* London: Springer

Jokinen, E., Heinonen, M., korkeakoulu, S., Lähdesmäki, H., Aalto-yliopisto & University, A. *Modeling protein stability with Gaussian processes*.

Jones, C. & Watkins, C. (2009). Housing market and planning policy (1st ed.). United Kingdom: John Wiley & Sons Ltd.

Kain, J. F. & Quigley, J.M. (1970). Measuring the value of house quality. *Journal of the American Statistical Association* 65(330): 532-548. Available from: http://urbanpolicy.berkeley.edu/pdf/kq_jasa70.pdf [Accessed on 14 July 2017]

Karazsia, B. & Van Dulmen, M. (2008) *Regression Models for Count Data: illustrations using longitudinal predictors of childhood injury,* 33(10):1076-84 Available from: https://www.ncbi.nlm.nih.gov/pubmed/18522994. [Accessed on 14 July 2017]

KTI (2017). The Finnish Property Market. Available from: https: //kti.fi/wp-content/uploads/page/The-Finnish-Property-Market-2017.pdf [Accessed on 14 July 2017]

Kuosmanen, P. (1997) Asuntojen hintojen ennustaminen: Ekonometriset mallits vs aikasarjamallit. Discussion Papers 231 Vaasa 1997.

Kusan, H., Aytekin, O & Özdemir, I. (2010) The use of fuzzy logic in predicting house selling price. *Expert Systems with Applications*, 37(3): 1808-1813. Available from: http://www.sciencedirect.com/science/article/pii/S0957417409006885 [Accessed on 14 July 2017]

Kekäläinen, A., Tähtinen, T. & Vuori, L. PTT-ennuste: Asuntomarkkinat 2017. ISSN 1799-9340. Helsinki 2017.

Laine, A., Toivonen, S., Holm, M., korkeakoulu, I., Viitanen, K., (2015) Aalto-yliopisto & University, A. *Joukkoliikenteen palvelutason merkitys asuntomarkkinoilla*.

Levitt, S. D., & Syverson, C. (2008). Market distortions when agents are better informed: The value of information in real estate transactions. The Review of Economics and Statistics, 90(4), 599-611.

Limsombunchao, V. (2004) House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. Available from: https://researcharchive.lincoln.ac.nz/bitstream/handle/10182/5198/House_%20price_%20prediction.pdf?sequence=1&isAllowed=y [Accessed on 14 July 2017]

Lockhart, R. (1997) 'STAT 350: Linear Models in Applied Statistics II'. *Heteroscedastic Errors.* Available from: http://people.stat.sfu.ca/~lockhart/richard/350/08_2/lectures/Heteroscedasticity/web.pdf [Accessed on 14 July 2017]

Lof, M. (2016) *Econometrics for Finance.* Course Slides.

Lönnqvist, H. & Vaattovaara, M. *Asuntomarkkinoiden vuoristorata – ovatko kaikki alueet samalla radalla? Helsingin kaupungin tietokeskuksen tutkimusjulkaisu.* 2004. 68 s. ISBN 952-473-339-0, ISSN 1455-724X

Maddala, G. S. & Lahiri, K. (2009) *Introduction to econometrics*. 4th ed. Hoboken, NJ: Wiley.

Mourouzi-Sivitanidou, R. (2011) *Market Analysis for Real Estate.* University of Southern California

Norros, P. (2014) *Google Trendsin hyödyntäminen asuntomarkkinoiden ennustamisessa.* Espoo: Teknillinen korkeakoulu.

Norros, P. (2017) *Internethakujen hyödyntäminen Suomen asuntomarkkinoiden nykytilan arvioinnissa.* Espoo: Teknillinen korkeakoulu.

Ostamo, T. (1997). *Asuntomarkkinat vuokraamisen näkökulmasta.* Espoo: Teknillinen korkeakoulu.

Ottensmann, J., Payton, S. & Man, J. (2008) Urban Location and Housing Prices within a Hedonic Model. *The Journal of regional Analysis & Policy,* 38(1): 19-35. Available from: http://ageconsearch.umn.edu/bitstream/132338/2/08-1-2.pdf [Accessed on 14 July 2017]

Pirounakis, N.G. 2013. Real Estate Economics: A Point-to-Point Handbook. Routledge S. Verkkokirja (pdf). ISSN 9780203094648

Procházka, D. (2017) *New trends in finance and accounting.* Springer

Reese, R. (2016) 'Poisson versus Negative Binomial Regression'. Available from: http://www.math.usu.edu/jrstevens/biostat/PoissonNB.pdf [Accessed on 14 July 2017]

Reinhart, C. & Rogoff, K. 2009, *This Time is Different: Eight Centuries of Financial Folly* (Princeton University Press: Princeton, New Jersey).

Ringle, C.M., Wende, S. & Becker, J. (2015). SmartPLS 3. Bönningstedt: SmartPLS.

Rodríguez, G. (2009) Lecture Notes on Generalized Linear Models. Available from: http://data.princeton.edu/wws509/notes/ [Accessed on 14 July 2017]

Rodríguez, G. (2013) Models for Count Data with Overdispersion. Available from: http://data.princeton.edu/wws509/notes/c4a.pdf [Accessed on 14 July 2017]

Rosen, S. 1974. Hedonic Prices and Implicit Markets: Product differentiation in Pure Competition. The University of Chicago Press. The Journal of Political Economy nro 82:1, 34-55 s.

Salo, P. (2009) Asuntojen hintojen muutosten vaikutus kotitalouksien kulutukseen. *Maisterin tutkinnon tutkielma.* Available from: https://aaltodoc.aalto.fi/bitstream/handle/123456789/188/hse_ethesis_12056.pdf?sequence=1&isAllowed=y [Accessed on 14 July 2017]

Shaikh, F. (2017) *Simple Beginner's guide to Reinforcement Learning & its implementation.* Available from: https://www.analyticsvidhya.com/blog/2017/01/introduction-to-reinforcement-learning-implementation/ [Accessed on 14 July 2017]

Shmueli, G. (2010) 'To Explain or to Predict?'. *Statistical Science,* 25(3): 289-310. Available from: https://www.stat.berkeley.edu/~aldous/157/Papers/shmueli.pdf [Accessed on 14 July 2017]

Siikanen, A. (1992) Asuntojen kysyntä, tarjonta ja alueellinen erilaistuneisuus. Asunto-hallitus, tutkimus- ja suunnitteluosaston asuntotutkimuksia 4/1992. Helsinki 1992. s. 169.

Siikanen, A. & Tyrkkö, A. (1993). *Koti, talous, asuntomarkkinat.* Helsinki: Tilastokeskus.

Skurnik, S. & Summa, H. (1978). *Asuntomarkkinoiden kokonaiskysyntä ja sen vaihtelut: ongelmakentän kartoitus* (Vuosik. Rakennus- ja yhdyskuntatekniikan julkaisu no 20). Espoo: Valtion teknillinen tutkimuskeskus.

Straszheim, M. (1975) *An Econometric Analysis of the Urban Housing Market.* Available from: *http://econpapers.repec.org/bookchap/nbrnberbk/stra75-1.htm* [Accessed on 14 July 2017]

Takala, N., Salenius, M. (2016), korkeakoulu, I., Falkenbach, H., Aalto-yliopisto & University, A. *Ostajan paikkakunnan vaikutus kerrostaloasunnon kauppahintaan*.

Tibshirani, R. (1996) *Regression Shrinkage and Selection via the Lasso.* Available from: https://statweb.stanford.edu/~tibs/lasso/lasso.pdf [Accessed on 14 July 2017]

Tuhkuri, J. (2014) Big Data: Google-haut ennustavat työttömyyttä Suomessa. *Elinkeinoelämän tutkimuslaitos*, 36 Available from: http://www.etla.fi/wp-content/uploads/ETLA-Raportit-Reports-31.pdf [Accessed on 14 July 2017]

Vatcheva, K., Lee, M., McCormick, J. & Rahbar, M. (2016) 'Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies'. *Epidemiology (Sunnyvale),* 6(2): 227. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4888898/ [Accessed on 14 July 2017]

Vehviläinen, T. (2016) *Suomen asuntomarkkinoiden kehitys vuosina 2000-2016.* Available from: http://www.doria.fi/bitstream/handle/10024/123507/Kandidaatintutkielma_Vehvilainen_Tero.pdf?sequence=2 [Accessed on 14 July 2017]

Zhang, Y., Kauppakorkeakoulu, Business, S. o., laitos, T., Economics, D. o., Aalto-yliopisto & University, A. 2015. *Analysis of China's current account: Evidence based on inter-temporal current account model*.

Wang, Z., Wang, C. & Zhang, Q. (2015) 'Population Ageing, Urbanization and Housing Demand'. *Journal of Service Science and Management*, 8, 516-525. Available from: https://file.scirp.org/pdf/JSSM_2015080615224683.pdf [Accessed on 14 July 2017]

Wu, L & Brynjolfsson E. (2009) *The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities*, 44 s. Massachusetts Institute of Technology Available from: http://www.nber.org/chapters/c12994.pdf [Accessed on 14 July 2017]

Wu, L & Brynjolfsson E. (2014) *The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales*, The economics of digitization. [Accessed on 14 July 2017]

Appendix A: Interview Questions

Haastattelukysymykset

Haastateltava

Haastateltavan tausta

Markkinat
-Minkälainen tilanne on tällä hetkellä asuntomarkkinoilla ja se on kehittynyt viimeisen 5 vuoden aikana?

1. Sijainti ja tyyppi
- Missä päin Suomea myydään enemmän asuntoja suhteessa väkilukuun?

-Onko joitakin trendejä tullut esille?

-Läheisyys julkisiin yhteyksiin, puistoihin, kauppoihin tai muihin vastaaviin? Vaikuttaako joku näistä enemmän kuin toinen vai yhtä paljon?

-Mitä asuntotyyppiä myydään eniten ja osaatko sanoa miksi?

-Onko eroavaisuuksia tekijöillä, jotka vaikuttavat rivitalon tai kerrostalon asuntokauppaan?

2. Talous ja rahoitus
-Miten asuntojen hinnan kasvaminen vaikuttaa asuntokauppaan? Mikä merkitys ja korrelaatio asunnon hinnalla on asuntokauppaan?

-Miten korkotaso vaikuttaa asuntokauppaan?

-Miten inflaatio vaikuttaa asuntokauppaan?

-Miten verot vaikuttavat asuntokauppaan? Varainsiirtoverot?

3. Suomi vs. Muu maailma
-Onko olemassa tiettyjä muuttujia/tekijöitä jotka vaikuttavat erityisesti Suomen asuntokauppaan verrattuna muihin maihin?

-Onko olemassa vuodenaikoja jolloin asuntoja ostetaan enemmän tai vähemmän?

4. Ostaja
-Miten ostajan ikä vaikuttaa asuntokauppaan?

-Miten suurien ikäluokkien eläköityminen vaikuttaa asuntokauppaan?

-Kuinka usein ihmiset keskimäärin vaihtavat asuntoa?

-Miten perheen koko vaikuttaa asuntokauppaan?

-Miten asuntokunnan tulot vaikuttavat asuntokauppaan?

5. Muut asiat
-Onko viimeisten 5 vuoden aikana tullut esille joitain uusia asuntokauppaan vaikuttavia tekijöitä tai onko joidenkin tekijöiden vaikutus/merkitys kasvanut tai vähentynyt.

-Miten tai vaikuttaako ekologisuus asuntokauppaan?

-Miten media vaikuttaa asuntokauppaan?

6. Tärkeimmät tekijät
-Mitkä tekijät vaikuttavat eniten asuntokauppaan? Positiivisesti? Negatiivisesti?

-Tärkeysjärjestys? Sekä positiiviset että negatiiviset.