The impact of socioeconomics on property price valuation

- Machine learning and deep learning -

Social Data Science (SDS) M4

Group number 3

Authors:
Troels Ryberg Andersen
Simon Søndergaard
Kristian Thorndahl

Supervisor:
Hamid Bekamiri

Aalborg University Business School

Fibigerstræde 2,
9220 Aalborg Øst

This project was written in **Share**LaTeX.
ShareLaTeX's The source code is a part of an *open source* project *The LaTeX Project*

AAU EXECUTIVE
BUSINESS & SOCIAL SCIENCES
AALBORG UNIVERSITY

Aalborg University Business School
Fibigerstræde 2
9220 Aalborg Øst

Title:
The impact of socioeconomics on property price valuation

Theme:
Machine learning and deep learning

Project period:
Autumn 2021

Project group number :
3

Participants:
Troels Ryberg Andersen
Simon Søndergaard
Kristian Thorndahl

Supervisor:
Hamid Bekamiri

Number of pages: 38

Submission date:
10. december, 2022

Number of characters:
69.462

Abstract:

This project seeks to provide insight into how socioeconomics can explain some differences in property pricing as a proof of concept. The analysis is built as a two-parted analysis wheres the first part uses a definition of socioeconomics as a point of departure for choosing socioeconomic variables. These socioeconomical variables are then applied to machine learning models for predicting property pricing with the addition of housing features. Part two of the analysis is a matter of explaining the machine learning setup and provide an analytical understanding the findings. Lastly, the project aims at providing a reasonable result of which, if any, socioeconomic variables has an influence on property prices and to what purpose this may benefit.

The projects findings ends up not being satisfactory, and results in reflecting the limitations when using data that is under strict data protection.

Aalborg Universitet, January 9, 2022

Kristian Thorndahl
Kthorn17@student.aau.dk

Troels Ryberg Andersen
tran17@student.aau.dk

Simon Søndergaard
ssand17@student.aau.dk

Contents

Chapter 1

Introduction

In all societies, there is inequality. This shows itself in the form of social stratification, which is defined as:

> *Social stratification refers to the unequal distribution of valued resources across social groups. The resources that underlie stratification systems are both tangible and intangible: economic, political, social, civil, but also cultural and honorific* (McLeod, 2013, p.229).

For this reason, social stratification is essentially in all layers of societies.

This inequality and social stratification can impact whether or not people purchase an apartment or a house because, the social setup individuals exist in, influences their purchasing power and opportunities. Therefore, it is interesting to investigate how social indicators can be observed in housing prices, where it could be expected that negative social indicators make housing less attractive the to average consumer, and thus the general market.

The fact that physical surroundings and resources might impact socioeconomical factors is an indication that socioeconomically environments indirectly influence the housing prices in neighbourhoods and therefore is an essential element to be aware of, when investing in property. (Bowen & Quintiliani, 2019).

One of the ways to observe this relationship is through machine learning, where the algorithm will attempt to find relationships between variables, and as such can attempt to forecast which variables influence the outcome (Foote, 2021). In this case in particular, the study utilizes these algorithms in an attempt to predict the prices of properties, based on location and similar factors. Machine learning has proved to be a successful inclusion in many projects, and as such companies and organisations can take advantages of this as a decision-supporting tool (Haviv, 2019). Although, machine learning algorithms are not always easy to apply to all types of problems.

By applying machine learning in a socioeconomic context, complications may occur since there are, according to Haghdoost (2012), *"no gold standard method to measure it"*, when working with socioeconomics. Therefore, it might create a clash between the two fields since coding is based on the assumption that any result is final since it is based on mathematics, which is not the case when working with socioeconomics. As such, there is a need to understand both sides of the two, very different, scientific fields.

Chapter 2

Problem area

The conflict of these two scientific fields lead to considerations of their interaction. As such, interest in discussing how socioeconomics influence economics in a specific scenario, became apparent. To better explore this phenomena, housing prices came to mind.

Housing valuation is a tricky concept that can be effected by a large amount of variables, the obvious variables being square meters size, housing type, location, and/or build year etc. The trend of housing prices has according to "Global House Price Index Q3 2021" (2021), shown that the global housing market has steadily risen since the market crash in the year of 2008/09. This was further increased by the pandemic, which also increased the demand for housing in many countries around the world (Fund, n.d.). With that being said, many speculate that this surge in the market might be halting in the coming years. As an example, the United kingdom, who in the past year has experienced a 10% surge in house prices, predict that these numbers will fall to about 3.5% by 2026 (Peachey, 2021).

What this means for societies is, that dependent on the individuals who are interested in acquiring property, the housing value may shift depending on the demographic of the area, as well as the socioeconomic factors at play. The social value of these areas may shift over time for better or worse dependent on housing prices Social Value UK (2021). Thereby creating a possibility that some areas within specific cities may have varying demand which potentially influences property prices. This does of course include the tangible housing attributes, such as housing types (e.g. cement housing complexes or suburbs of detached houses).

Regarding factors for property valuation, intangible attributes, that is not clearly defined, can be hard to asses. Regarding the societal issue some cities, areas, or neighborhoods may be preferable to acquire housing, regardless of the tangible housing features. To explain these intangible attributes, observing the socioeconomic factors of the area or demographic could be a solution.

Chapter 3

Thesis statement

The presented problem area, leads the project to the following thesis statement:

How can intangible societal variables generate a meaningful reflection on property prices, and are there benefits by connecting socioeconomic variables to property prices?

3.1 Explanation of the thesis statement

The thesis statement designed to be exploratory and problem-solving, which is a needed process as it is based upon societal assumptions.

The resolution of the problem could be relevant for organisations or individuals. These could be the individual residents who may use the projects information to estimate an existing housing value or find a place that matches their socioeconomic status with a matching price point, based upon a consumers cost benefit perspective. Alternatively, findings may interest people who work in the real estate business, because the results can be used as a supporting housing valuation tool. Lastly, the knowledge gained by the project may be utilized by political parties, allowing for evidence based arguments upon influencing socioeconomics in societies as to change cities reputation and societal class distribution.

3.2 The limitations of the project

As certain boundaries typically appear in project work; it is essential to note that the projects' result has to be seen as a *'proof of concept'*. A lack of available data is the primary reason that several boundaries have to be set for the scope of the project. This means the results will not be able to be generalized for other demographics as the study is performed on a narrow dataset. General property price listings and valuations may be available on different levels of aggregation, though individual publicly available observations seem to be lacking. This has led to the creation of this project to be a proof of concept based upon German housing data. This was the most descriptive property pricing dataset the project could acquire.

The data contains features of which some are highly valuable, however the lack of observations over time causes the development of a property price index over time, to be impossible to display. Therefore, the project chooses to focus on one unique year of available property listings.

Thus, this project's findings will not predict a grand market prediction and will instead contain a stand still image of selected German housing prices. Therefore, this project may create a distinctive view of certain parts of Germany, drastically reducing the generalization, limited by the number of single observations in the specific areas of Germany. Though the extent of this project's scope, and thereby findings, should be limited only by data availability. There needs to be a greater amount of observations if the results of the project is to be considered fit for generalising beyond the examined areas, as the current dataset only contains around 6000 observations. Additionally, as the housing data used for the analysis is based upon sold listings garnered from a German real estate site, it can be unrepresentative of an actual value. This is because the data is sourced from a real estate company, accessed through kaggle Seref (2020), who may value housing differently than banks as mortgage lenders or consumers. The heterogeneity of housing may also be undermined regarding the public market value. Some housing options may have unique architectural properties or sentimental/historical value that may be downplayed. This may be information that even societal information like socioeconomics on a neighbourhood dis-aggregated level may not portray.

To combine socioeconomic factors to the housing data, the project acquired a dataset consisting of questionnaire results, gathered by the German research infrastructure "Socio-Economic Panel" [SOEP]. This consistent of socially aimed questions gathered throughout the latest decades (Panel, 2021). Because of General Data Protection Regulation [GDPR] the group was unable to access many of the personal identifying answers, that would assist in better formatting the social data to specific locations. As such, much of the data had to be aggregated to a state level, which was the most disaggrigated level that both datasets included. This means that several of the housing datapoints had to be further filtered, as several states did not include enough observations to provide meaningful analysis.

By combining these datasets, the objective of the analysis is to give sufficient insight into how the different socioeconomic factors, impacts the economics of property pricing. While this proof of concept might not provide an objective answer to which socioeconomic factors are the most important when determining property prices, the goal is to at least provide a general understanding of whether socioeconomics and housing valuations interact.

Chapter 4

Philosophy of science

The purpose of this chapter is to illustrate the projects' philosophy of science standpoint and thereby describe our methods that creates the possibility to answer our thesis statement. Throughout the following, there is an argumentation of the consequences when researchers take a standpoint.

4.1 Operative paradigm

The sources and inspiration for the following chapter is from Buch-Hansen and Nielsen (2005) and Arbnor and Bjarke (2009)

This project is written from a critical realism point of view and thereby assumes a part of the ontology with independent surroundings of knowledge that researchers cannot reach. This project's ontological point of view is housing pricing valuations because the project aims to clarify which indicators affect housing prices. Therefore, the goal throughout the project is to be as neutral as possible because of the inability to assess an objective reality, based on the intransitive domain in critical realism. This is of the fundamental assumptions throughout the project because the thesis states that there might be a type of correlation of socioeconomic variables which affects property prices. There is a possibility that there are some indicators of which the researchers are not aware of, that affects the property prices but are not included in the project. Therefore, it is inevitable that the project will be limited by bounded rationality throughout the process. This is apparent in the data gathering, as gaining access and manipulating the raw data for the project was more complicated than initially expected.

In regards to the critical realism domains, the housing prices and social variables can be explained in the empirical domain. In contrast, the empirical domain is affected by the actual domain, where it contains phenomena and occurrences that exist regardless of whether those will be discovered. The actual domain does not contain any straightforward observable structures or mechanisms. This project hopes to provoke some of those hidden mechanisms and structures throughout the analysis because those underlying structures and mechanisms that support and cause the phenomena in the real domain, could explain why some housing prices might be higher than others.

The use of secondary literature from SOEP, scientific literature, etc. creates the possibility to use a system analysis where the researchers do not influence the subject area of investigation and thereby investigate the ontology in an unaffected state,and by using this approach the researchers strive for objectivity. In addition to that, the secondary literature grants the possibility in the transitive domain to recognise property pricing. The argumentation for this, in a epistemological sense, is that it seems socially constructed and follows some of the ideas from constructivism, because people have individual perceptions of reality, which is why the literature is socially made according to critical realism.

The purpose of analysing this subject from a system analytical point of view is to recognise and understand the actual domain. Still, through the analysis, some of the results are based on assumptions, which potentially creates an ontological fallacy because it creates a system construction where the researchers create a partially closed system. Results derived from societal variables, is based on assumptions whereas this enables uncertainty of whether those included are the *"right ones"*. Thus, there is a possibility that the analysis results are oversimplified. As a result of working with machine learning/deep learning, the results of this project might be seen as absolute which is not the reality. The results should be seen as a way to recognise part of the independent surroundings and thereby investigate whether the housing prices as a system which might be connected to social indicators and potentially create synergistic effects.

4.2 Methodology

The project is build on deductive reasoning, primarily because its written from a theoretical point of view. This allowed defining socioeconomics as a concept initially, and let the connections between the this definition and property pricing to evolve from that. Consequently, it is crucial to take a methodical standpoint to analyse the particular phenomena. The project is written from the perspective of having an inside-out structure because there has not been collected any primary empirical research, which creates the possibility to be neutral in the following analysis. Though this is based upon the secondary literature being assumed to represent the truth of the domain.

The utilisation of secondary qualitative and quantitative literature in the project impacts the researchers using the quantitative and qualitative methods. This is shown in the analysis of which variables should be included. This qualitative method creates the possibility to analyse and understand the ontology.

4.3 Reliability and validity

The following section has the purpose of describing the reliability and validity of the project. There will also be an explanation of the potential caveats that occur through house prices and socioeconomic data. It gives transparency for the reader of how the sources have been collected, and why there is both internal and external validity throughout the project's results, which allows the reader to judge obtained results.

The literature that has been used in this section is based on the book Jacobsen, Nedergaard, and Rasmussen (2015).

When intending to develop a machine learning model it is crucial to be aware of external validity, to ensure a understanding of the analysis process. The internal validity is received by striving to obtain credible sources, which have been a struggle because there is a limited amount of research papers that describe the specific topic as the investigated in this research. Still, the included sources are carefully selected by considering the citation amount.

The property prices and socioeconomical datasets are context correlated since they both include key identifiers for merging, in this case, the German state in which they are observed. Although, it is still noteworthy that the generalizability of the project findings are only viable in the context of the data that is present herein. Therefore, the external validity should be improvable based solely upon the utilized data. This is not necessarily negative, as the models, and in extension the results as well, are developed as a proof of concept, that has the ability to be up scaled and changed.

The reliability through the project is important to be aware of, since a degree of external reliability has an influence of how some of the results are possible to recreate. This philosophy of science chapters purpose is to increase the transparency for readers, by presenting the chosen methods and theories.

Chapter 5

Theory

The following chapter describes the two essential scientific areas utilized throughout the project analysis; Socioeconomic definition and a machine learning introduction.

5.1 Discussing the definition of socioeconomics

When approaching the field of socioeconomics, it is necessary to consider that the field covers a wide range of scientific possibilities. One of the earlier definitions of the term was *"The use of economics in the study of society"* (Eatwell, Millgate, & Newman, 1989). Over the following decades the term evolved to focus on other types of sciences in relation to economics, such as sociology, psychology and ethics. Here the focus becomes to study the reciprocal relationship between these sciences as they appear in the world (Lutz, 2009).

By looking into earlier works in the space of housing and social science found, that social influences does have a sizeable effect on the housing market (Reed, 2001). But while the overall connection has been made, there is no clarified studies that investigate the direct correlation between individual socioeconomic factors and housing pricing. This gap in the science is what this project will be trying to uncover. It is also important to acknowledge that earlier work in the field has concluded that demographics plays a sizeable role when analysing this economic area (Berson & Berson, 1997). As such, this study will not be representative for other demographics than the selected ones, and will most likely have to be redone on a case-by-case level to adapt to different demographics.

So in the context of this project of identifying if social factors have an impact on house pricing, it is important to have a clear definition of what is considered to be socioeconomic factors. This could be factors such as perceived safety (crime rate in the area), perceived culture (Is the area "high class" or more of a lower class area), or environmental factors (such as which energy class a house is). Additionally, urban sciences would preferably be considered, by including more specific physical factors such as access to parking, public transport or parks. These factors were unfortunately unavailable due to the state level aggregation of the socioeconomic data.

In short, the study views socioeconomics as the dynamic relationship between the sciences of economics and social / urban sciences. By viewing the social and urban sciences from a economic perspective, the study hopes to gain a better understanding of how these influence the pricing and sales of property throughout the demographics in question, which in this case is the larger German cities.

5.2 Machine learning theory

This following section is primarily referenced from VanderPlas (2016).

The general purpose of a machine learning model is, simply put, to make predictions based upon data. The models can be of different types and have features that they are good at, which is why it is important to choose a model that fits the purpose and/or structure of the data it is fed with, and can thereby be a great forecasting tool.

Methods for creating a suitable machine learning model, is often reliant upon trial and error. It is commonplace to try out different model types, model settings and compositions of data, as to create a model that is optimal for the problem at hand.

With this, there are multiple theoretical procedures and preliminary parts to consider for creating a 'good' machine learning model.

First off, it is necessary to determine the prediction problem type. As for this project, this will be a regression problem, because the goal of the prediction is a numerical price, and will therefore be the dependent variable. The independent variables will be the features that are assumed to influence the dependent variable. Based upon the need of a numerical prediction, a simple regression model can be used as to test the impact of one independent variable and create a linear regression for the dependent variable. Though this does not allow for much depth, it can quickly show whether or not a correlation can be found between these variables. But as more independent variables are added, prediction outcomes may not be able to be predicted with a simple linear regression model.

The correlation of variables is paramount when creating a machine learning model, as this is the information of which the model finds patterns and systems to create predictions. Though herein lies the everlasting caveat of correlation not being equal to causation, and a healthy amount of skepticism is certainly needed when observing correlations. An example is the infamous correlation between Nicolas Cage starring in movies and accounts of people drowning in swimming pools which, in a matter of common sense, has no common causation and is therefore by random chance TylerVigen.com (2021).

Correlation testing can be achieved with different methods. Analysis of variance [ANOVA] is beneficial for testing both the correlation between dependent- and independent variables as well as the interactions between the independent variables. Furthermore, ANOVA allows for the correlation comparison of categorical variables as well as numerical. This provides a good baseline for understanding of which variables are suitable for the model to predict upon.

When the independent variables has been selected, the choice of machine learning model relies a bit upon trial and error. When feeding data to a machine learning model, this is usually done by splitting the datasets in two. One, and typically the largest portion, for training upon, as well as a test set where the models knowledge of the former data split is tested. The test set is like an answer sheet for the model to compare its predictions. With the results of how well the model made predictions compared to the test data, it is possible to create measurements of 'success' for the model such as Mean Squared Error, Accuracy, R2, Recall, Precision. As previously mentioned, simple regression models can be unreliable as not every problem can be defined as a linear function. Therefore, models like random forest models, may provide increased prediction scores. These function as a decision tree, where the model creates multiple outcomes for variables and decides a fitting prediction upon going their "decision tree". Though models like these typically leads to some caveats, such as overfitting.

Overfitting can be defined as when the model "memorizes" the data and correlations, instead of attempting to predict. An overfitted model would perform very well in a controlled environment, but could be off by a large margin when applied to new and unseen data. This entails a very low generalizability of the model. This phenomena is especially prominent for random forest decision trees, where the setup of the decision trees can account for each and every variable composition and would therefore lack the "knowledge" when handling new data.

Issues like these can be counteracted with hyperparameter tuning which, in its simplicity, is the tweaking of settings for the model. This is largely a question of trail and error and can take multiple tries to find a suitable model tuning, unless you have a large amount of experience.

For more complex multivariate problems, where the previously mentioned models won't suffice, neural networks can provide more powerful tools for a data scientist. A neural network can be seen as trying to replicate a brain with the sole purpose of understanding the data it is provided. A neural network can be a superordinate term for a very versatile model building method, which consists of neurons in multiple layers of which the 'brain' passes data from one neuron to the next. Each neuron can activate a varying amount, and this information is passed to the next neuron, in the next layer which influences that neurons outcome and so on. This continues until the last layer of the neural network is reached. Here an output is created, which can be translated to a result that can be compared with complementary test data.

The benefit of using a neural network compared to the previously mentioned machine learning models, is that we can tell the neural network how well it performed on the prediction outputs, and thereby force it to improve the result by automatically changing how the neurons activate in the network. At first this iterative process happens randomly, but as the model gets better scores on the test data, it "learns" and can improve over time. Usually the result is based on a "loss value", which a selected optimizer will seek to minimize.

Comparativly, the setup of hyperparameter tuning and the setup of a neural network are quite similar, as they can require a lot of trial and error. Typical decisions to make when starting out is choosing the amount of neurons of each layer, how many layers to add, what type of layer, what activation function to chose for each layer, what loss function to use, and what optimizer is optimal for the neural network. In addition to this model setup, it is up to the data scientist as well to decide how the test and training data distribution should be split as well as the amount of times the neural network are going to train on the data (Epochs).

Neural network dictionary

This sections purpose is to simplify and demystify some of the terms used in this chapter as well as the analysis of machine learning results. The intended purpose is to add some additional explanation to these terms as to help the reader. Some terms has a few examples of what could be the input, but this is far from every possible option within each step of creating a neural network.

Neurons: The linked nodes in a neural network where data influences to what degree the neuron activates. The amount of neurons can be chosen on a layer basis, throughout the neural network.

Layers: The layers in which the neurons are placed. Layers can be of different types and thereby add features and customizability to the neural network. These layers can e.g. work with different data types. (Deep learning is a term used when a multitude of layers is used in a neural network)

- A 'Dense' layer is a simple layer that connects the neurons. Its often used as the output layer when the output consist of a numeric value
- A 'Long Short-Term Memory layer' can process and remember time series data; data where the time or date of observations are paramount for the model.
- A 'Dropout layer' can add 'noise' to a neural network by randomly removing a percentage of neurons. This can be beneficial to reduce the tendency of overfitting, by making every epoch different.

Activation functions: Activation functions are applied to a whole layer and takes the summary of the previous neuron layer, taking input, weight and bias which is what defines the degree of activation of the neuron layer.

- 'Relu' activation function is sensible to small changes in input. Reduces output to 0 and up as a linear function.
- 'Sigmoid' activation function scales the between 0 to 1, which is useful for binary prediction outcomes for probabilities.
- 'Tanh' activation function is much like the Sigmoid function but scales from -1 to +1, which makes it good at explaining classification problems where there are two class outcome.(Sharma, 2017)

Loss function: Loss functions are the measure of precision of the neural networks predictions, and works as a numerical value that the neural network should aim at reducing. Mean Squared Error [MSE] or Mean Absolute Error [MAE] is usually used for linear regression problems. Whereas classification problems can use cross-entropy loss as measurement.

Optimizer (Backpropagation): An optimizer takes the loss score and update neuron layers weights in an attempt at reducing the loss functions value which is also called Gradient: Reduction of loss by adjusting weights to reach local minimum. A local minimum in regards to neural network loss reductions, is finding a composition of neuron weights which has the most effect on loss function reduction.

- Stochastic gradient descent: Applies to mini batches of data. An optimizer method that can be slow and a bit random, hence stochastic, but has the ability to evade a bad local minimum.
- Adadelta + momentum (Adam): An optimizer which is stated as a good optimizer overall, it should be good at reaching a local minimum quickly, though this may not be the best local minimum.

Chapter 6

Analysis

The first part of the analysis explores and analyses the property pricing- and Socioeconomic data, and are analysed separately.The housing data will be explored using various exploratory data analysis [EDA] methods, meanwhile the analysis of the socioeconomic data aims to characterise which variables need to be merged and included for machine learning. Part two of the analysis shows the results and interpretation of the different machine learning models and which models are sufficient to answer the thesis statement.

The figures presented in this chapter can be found in a larger format in appendix B.

6.1 Part 1: EDA & Feature engineering

The following section will contain a exploratory analysis of the property pricing dataset. The purpose is to be exploratory in the initial analysis, as to provide an understanding of the implications in the dataset.

6.1.1 EDA

The location data within the dataset is spread into three categories: State, City and Places within cities. State being the most aggregated form of location. Observations regarding the specific city are, e.g. in the state *Bayern*, relatively small, where there is 1338 observations shared among 95 cities, which gives an average of 3.5 observations pr. city. To get a better understanding of the distribution in the states, we chose to observe Bayern on a closer basis.

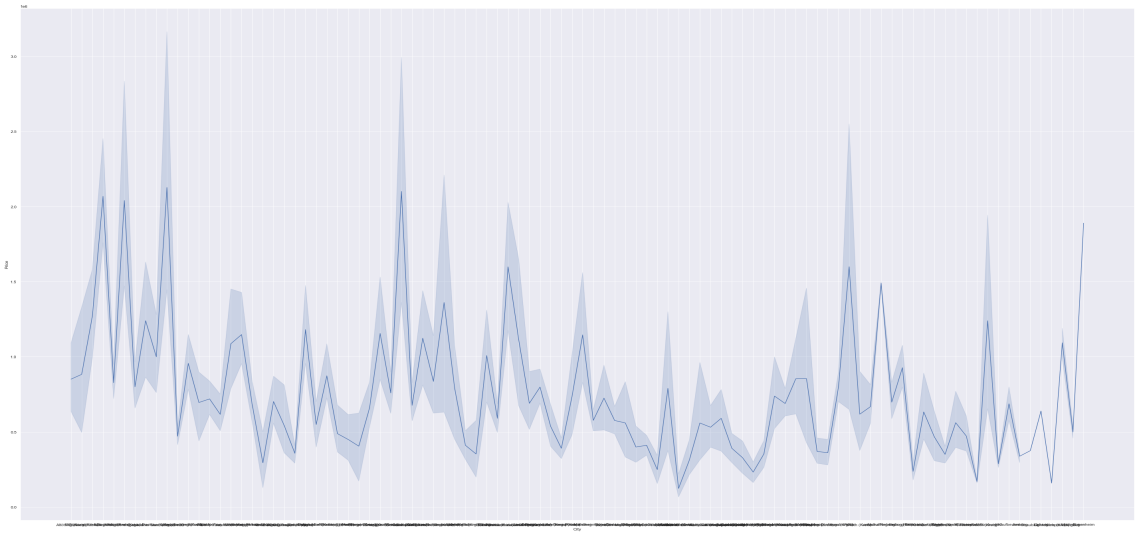


Figure 6.1: The distribution of houseprices based on cities in Bayern. The x-axis represents the different cities while the y-axis shows the house prices.(Appendix B)

The figure 6.1 presents the distribution of house prices in *Bayern*. Although the x-axis is unreadable since there are (95) different cities, it still visualises how the house prices differentiate. The blue shaded area shows the variation of house prices, while the highlighted blue line represents the mean house prices in a specific city. It could indicate that some of the cities in the state have relatively higher house prices than others. In some cities, the blue shaded area is broader, showing that housing prices are more spread between maximum and minimum. For that reason, this might indicate outliers within specific cities, which may not only be unique for Bayern and will have to be tested for the remaining states.

As such, outliers have been filtered out throughout all states since it will cause bias in the accuracy estimation for machine learning models, which is not preferable. In the filtering process, 428 observations were removed, and it did not negatively impact the distribution of observations. This was done by using the standard 1.5 interquartile for filtering. (Ismail, Kalid, & Khamis, 2005, p.1397) Following this, the analysis focused on comparing the five states with the highest observation count. The following states have been chosen: Nordrhein-Westfalen, Bayern, Baden-Wüttemberg, Neidersachsen & Rheinland-Pfalz. To gain a further insight into these states, the following plot shows how new versus old property prices is, distributed upon the selected states.

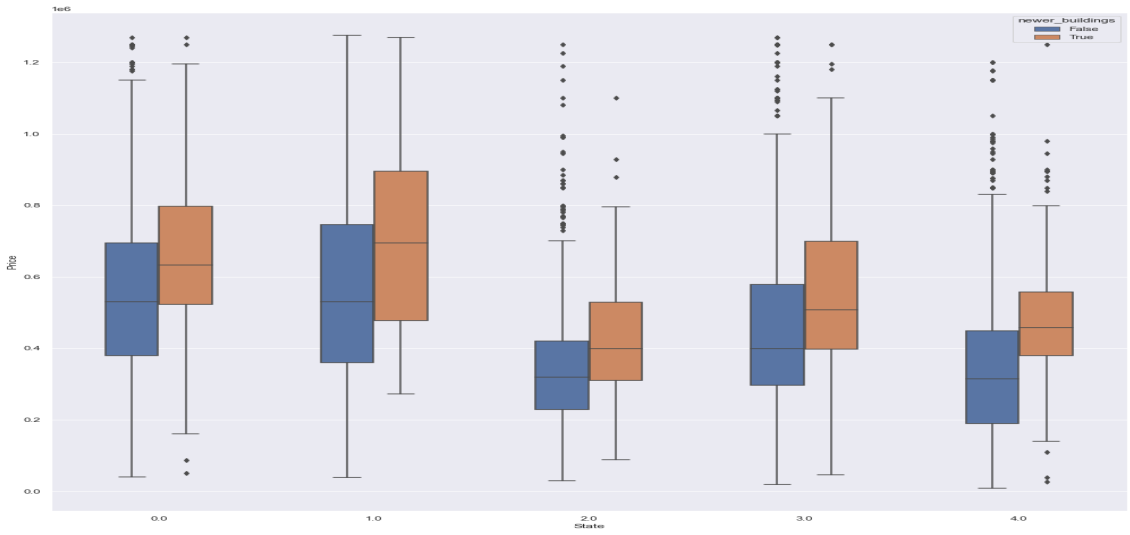


Figure 6.2: The boxplots display which houses that are built after 2005. The x-axis represents the different states while the y-axis shows the house price distribution. (Appendix B)

The figure 6.2 represents the impact on property pricing, of whether or not a house is built after the year 2005. The figure shows a tendency through all of the states that the *True* buildings constructed after 2005 is higher than the *false* ones build before. The boxplot explains how the data is distributed around the median, referring to the dataset's middle value(Galarnyk, 2018). An example of what can be observed in the plot, is illustrated in that the property prices of "state 4" vary more if a buyer purchases a house built before 2005 which is indicated by the length of the boxplots. The other plots have a similar correlation between the True and False plots and demonstrate either an increase or decrease in the length of the interquartile range. Therefore it can be observed that the differences in the price spread are similar to each other. The influence of the shifted boxplots results in a change of the median.

To further examine these plots, they can be transformed into swarmplots that better show the individual distributions across the states.

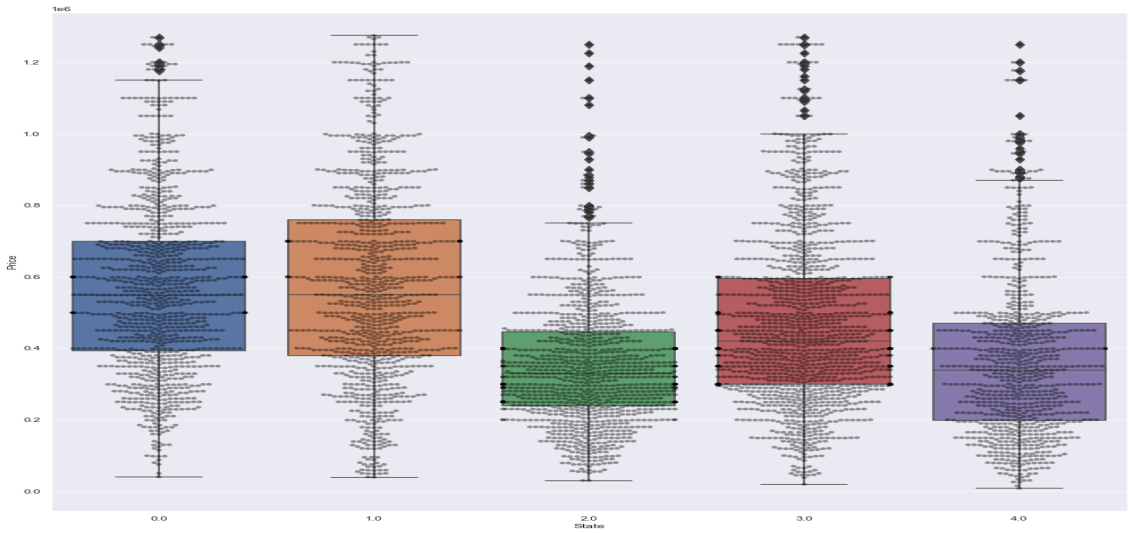


Figure 6.3: The boxplots and swarmplots show the differences in the different states. The x-axis displays the different states while the y-axis represents the house prices. (Appendix B)

Figure 6.3 contains swarmplots layered on top of a general boxplot to visualise the distribution. In state 1, the box plot is taller than the other, which testifies a greater spread of the housing prices, whereas the more flat boxplots display a more compact price distribution around the median. In state 3 the distribution is positively skewed. That is because half of the box and the whisker are placed towards the bottom of the median. On the other hand, the remaining states are more or less concentrated around the distribution. It is essential to be aware of the distribution because linear models assume that the dependent- and independent variables are similar(Jaggia & Kelly, 2019).

It is essential to take notice of the different medians in figure 6.3 where two of the five states have a similar median of around 575000 Euros, which is also why whiskers are higher compared to states 2-4. The different medians indicate that some states are more expensive to buy a property in. The mean and median could differ, but using the median instead of the mean creates a better understanding of the data since there are extreme values within the data set(Jaggia & Kelly, 2019).

6.1.2 Feature engineering for socioeconomic variables

When starting to analyze the socioeconomic dataset, it must be acknowledged that because of the vastness of the data, some feature engineering will be necessary. This will go through the selection of data and the reasoning for combining various variables.

Socioeconomic youth data

Following the definition mentioned in the section, 5.1 it is essential to state that children can indicate how a specific family performs from a socioeconomical point of view. The data contained in the *youth data* refers to questionnaire answers from persons 17 and down. These questions are similar to *"does your mother help you with homework?"*. Parents in general influence how well their children will be successful in life (Mclaunghlin & Sheridan, 2016). According to Conger, Conger, and Martin (2010) there is a relationship between socioeconomical factors and family life, which is the reason that the youth data is used, because it allows the project to examine this relationship. The youth data will, therefore, function as a proxy for the mental/psychological well being of households.

Some variables are merged in the youth data because the questions circulate the same category. This resulted in an aggregated parental score, dependent on the childs answers to some of the following questions:

- Mother/Father Talks About Things That Worry You
- Mother/Father Asks You Prior To Making Decision
- Mother/Father Has Impression Of Trusting You
- Mother/Father Shows That She Loves You

Additionally, some features as future aspirations and trust/distrust of strangers were included as features.

After filtering and merging to the property pricing data, the youth data left very few observations on the specific states, which would not suffice in machine learning models. Even though the youth data would be beneficial for machine learning, the data will not be used because of the above mentioned lack of observations. If the data contained adequate amount of observations, the variables would have been added as features for machine learning (Jaggia & Kelly, 2019).

Socioeconomic Household data

The socioeconomic dataset for households contained variables general for the whole household. This included variables as:

- Capacity to afford paying for one week annual holiday away from home [Affording holiday]
- Capacity to afford a meal with meat, chicken, fish (or vegetarian equivalent) every second day [Affording diverse food]

- Capacity to face unexpected financial expenses [Affording unexpected]
- Do you have a car? [Car]
- Total disposable household income
- Household size
- Dwelling type

These variables inclusion as features in machine learning will be discussed in their respective sections.

Relative poverty

Some of the aforementioned variables in the household dataset has some comparative quality, which suites them for being merged on a common term. According to a study from Braubach and Savelsberg (2009), conditions of a house have an impact on social inequalities and thereby the health inequalities. This is also related to the concept of *relative poverty*, which is defined as the following:

> *Relative poverty describes circumstances in which people cannot afford actively to participate in society and benefit from the activities and experiences that most people take for granted.*(Lichao & Walker, 2020, p.1).

The decision to merge *Affording diverse food* and *Affording unexpected* is based upon this definition of relative poverty. The survey questions are similar since both consider whether or not people can afford some minimum required necessities. The definition of *Affording unexpected* is as following *"surgery, funeral, major repair in the house, replacement of durables like washing machine, car"*(Bartels, Göth, & Nachtigall, 2019, p.223). Whereas *Affording diverse food* is described as *"whether, according to the household respondent, the household can afford a meal with meat, chicken or fish (or equivalent vegatarian) every second day regardless if the household wants it "*(Bartels et al., 2019, p.221). Therefore it is assumed that the two variables are similar. Furthermore, the *Car* variable is merged as the survey question specifies if the household has a car, cannot afford a car or any other reason for not owning a car. This allows for combining the car variable based upon observations where the individual cannot afford a vehicle. The variables could explain how much people can afford to partake in a variety of societal activities, which prevents the household for being categorized as within the relative poverty terminology.

Affording holiday

The *Affording holiday* variable will be the sole indicator of households being able to afford luxury activities or goods and would optimally be combined into a variable called *Affording luxury*. This could be combined with other luxuries as multiple vehicles or vehicle brand, eating habits/restaurant visits and quality of life goods within the household etc. Which can be elaborated upon addition of more data in this category.

In house necessities

The included variables from the socioeconomic data contains observations about whether or not the households have access to *phones, washing machine and computers*. In combination, access to these goods would be commonly referred to as *"in house necessities"*. As these types of items are expected to be present in a normally functioning household. Through the combination of the socioeconomic and property datasets, the variables washing machine and computer did not contain any observations, which caused them to be excluded. The only variable with a sufficient amount of observations was the *Phone* variable. Based on lacking observations in the variables, the independent variable called *"In house necessities"* will not be created, and will be represented as the *Phone* variable.

Social service income

This variable explains whether or not a household has social services income as part of their disposable income. Therefore, two variables *disposable income with social services* and *disposable income without social services* are subtracted as to provide a variables containing the social service income amount. This social service income variables will, on a state aggregated level, provide information of the average social service payout for households.

6.2 Analysis part 2: Machine Learning

Firstly, as both the housing- and socioeconomic data has been merged into a primary dataframe for use in machine learning, it is beneficial to test correlation between dependent and independent variables. Correlation testing is done through a Pearson correlation test and ANOVA.

This is done as to make further justified modifications to the data as it is prepared to be used in machine learning models. As the socioeconomic variables are aggregated to a state level, it could be an issue as these features lack explanation compared to housing prices. This could, of cause be improved upon with less aggregate data of socioeconomics. By removing uncorrelated variables, the probability of success of the machine learning models should increase, as the noise generated by uncorrelated variables can reduce the model's prediction accuracy. A slight caveat regarding noise is, that adding noise can reduce overfitting. This can primarily be tested by running the data through the models and noise can later be added artificially.

	parameter_one	parameter_two	correlation	Do_correlate?
0	Price	Living_space	0.405045	yes
1	Price	HH_gross_income_mean	0.349621	yes
2	Price	HH_disposable_income_mean	0.345542	yes
3	Price	Phone_std	0.302892	yes
4	Price	Phone_mean	0.301412	yes
5	Price	Bathrooms	0.295349	yes
6	Price	Year_built	0.282400	yes
7	Price	Rooms	0.266642	yes
8	Price	Bedrooms	0.246157	yes
9	Price	HH_gross_income_std	0.245698	yes
10	Price	Dwelling_type_mean	0.241202	yes
11	Price	Floors	0.227775	yes
12	Price	HH_disposable_income_std	0.224764	yes
13	Price	Dwelling_type_std	0.208359	yes
14	Price	Garages	0.171165	yes
15	Price	Usable_area	0.163777	yes
16	Price	newer_buildings	0.161396	yes
17	Price	City	0.104831	yes
18	Price	Lot	0.081408	yes
19	Price	Place	0.070100	yes
20	Price	Energy_source	0.061736	yes
21	Price	Type	0.026038	no
22	Price	HH_size_mean	0.015123	no
23	Price	HH_social_service_dis_income_std	-0.006097	no
24	Price	Heating	-0.019852	no
25	Price	Garagetype	-0.059173	yes
26	Price	Energy_efficiency_class	-0.135330	yes
27	Price	HH_size_std	-0.149200	yes
28	Price	Condition	-0.164862	yes
29	Price	HH_social_service_dis_income_mean	-0.195904	yes
30	Price	crime_rate_mean	-0.254205	yes
31	Price	State	-0.265194	yes
32	Price	Relative_pov_cat_mean	-0.341566	yes
33	Price	Affording_holiday_std	-0.348558	yes
34	Price	Affording_holiday_mean	-0.351118	yes
35	Price	Relative_pov_cat_std	-0.360207	yes

Figure 6.4: The table shows the Pearson correlation of different variables based on the prices. (Appendix B)

The results of the Pearson correlation test shown in figure 6.4 shows that some variables have very low correlation with the dependent variable of Price. These variables have then been arbitrarily sorted to only include variables with a correlation of 0.1 and higher as well as -0.1 and lower. – Note: a negative correlation can still be a valuable correlation. This means, that these independent variables can be chosen as the ones for machine learning models. As seen above, not a significant amount of variables seem to have low correlation with the dependent variable 'Price'. As the Pearson correlation test-results provides an output based on a linear relationship, the actual correlation may yet be unknown, as the assumption is that there may be unknown subjective measures which also has an influence on the price (Statkat, n.d.).

Though the Pearson correlation can give valuable information about independent variables linear relationship, it is still preferred to also create an ANOVA correlation (Jaggia & Kelly, 2019). This can give some information of interaction between the variables, especially the categorical variables, which can be used, in theory, to confirm whether some variables are dependent on other variables. This can be used in practice as an indicator of which variables are tied together and how some variables can explain each other (Jaggia & Kelly, 2019, p.480). This could also suffice as an argument of how feature engineering could be done differently, as these interactions between variables can show which variables are substitutable.

	PR(>F)	Do_correlate?
Garages	0.600569	no
Usable_area	0.527713	no
Bedrooms	0.312652	no
Place	0.180309	no
Heating	0.032318	yes
Energy_efficiency_class	0.003256	yes
Rooms	0.003077	yes
City	0.002767	yes
HH_social_service_dis_income_std	0.000273	yes
Type	0.000016	yes
HH_size_std	0.000012	yes
Garagetype	0.000012	yes
Bathrooms	0.000001	yes
Lot	0.000000	yes
Floors	0.000000	yes
HH_social_service_dis_income_mean	0.000000	yes
Energy_source	0.000000	yes
Condition	0.000000	yes
HH_gross_income_std	0.000000	yes
Phone_std	0.000000	yes
Dwelling_type_mean	0.000000	yes
HH_disposable_income_std	0.000000	yes
HH_gross_income_mean	0.000000	yes
Dwelling_type_std	0.000000	yes
Phone_mean	0.000000	yes
State	0.000000	yes
Relative_pov_cat_std	0.000000	yes
Affording_holiday_mean	0.000000	yes
Affording_holiday_std	0.000000	yes
Relative_pov_cat_mean	0.000000	yes
HH_disposable_income_mean	0.000000	yes
Year_built	0.000000	yes
Living_space	0.000000	yes
crime_rate_mean	0.000000	yes
HH_size_mean	0.000000	yes
Residual	nan	no

Figure 6.5: The table shows a two sided ANOVA test where it is tested whether there is a siginificance. (Appendix B)

As shown in figure 6.5 there are some variables that are classified as insignificant, with less than 5% significance based upon price. This indicates that the machine learning models performance may improve with the removal of the following variables; *Garages, Usable area, Bedrooms* and *Place*.

The dataset after EDA, preprocessing, feature engineering, and variable correlation filtering should in the theoretically preferred form for optimal performance in the machine learning models. Though, this is only an expectation as the results of the machine learning models will justify whether or not there actually has been an improvement when manipulating the data. Therefore, the optimal way of testing this would be to have comparisons of these models performance on different compositions of the data as machine learning models only are as good as the data input.

For the comparison of the different models, model setup and data input is paramount to create any meaningful measure of improvement to these three. The data will be split up into 3 separate iterations as to visualise any improvement when changes are made.

- The first dataset contains only the original housing price data. [House]
- The second dataset contains the full data, before variable correlation filtering. [Combined]
- The third dataset contains the finalised data with all changes. [Anova]

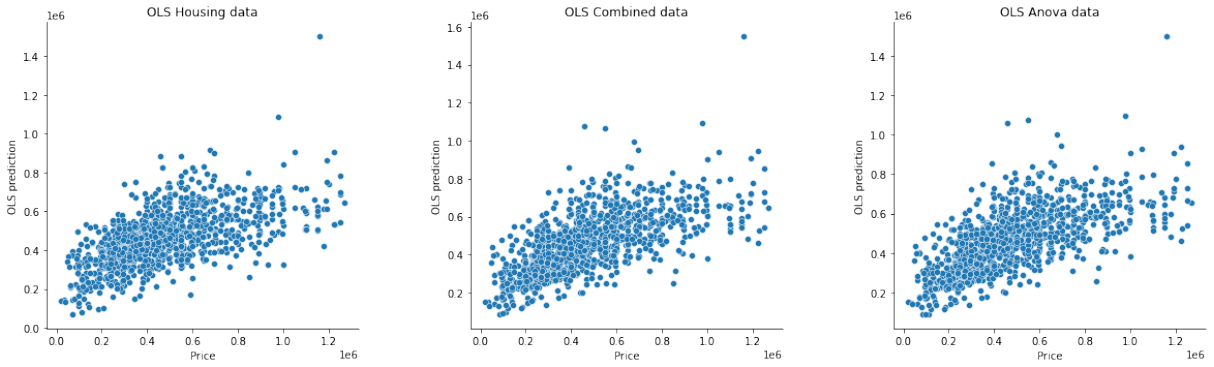
The comparison between these three iterations of the datasets should provide comparative results as to observe whether the model predictions has improved or decreased, which will be important for determining the outcome of the thesis statement.

6.2.1 Supervised machine learning

Supervised machine learning [SML] is the first machine learning models used to predict upon the dataset. These relatively simple SML models are quick to run and easy to setup (Kotsiantis & Pintelas, 2007). Seen from a cost/benefit perspective, this is the low cost method for machine learning, and should the prediction accuracy of SML models suffice, then going further with more advanced machine learning methods can be excessive.

Ordinary Least Squares [OLS] Model

Predictions made by an OLS model is provided the assumptions of a linear function. Therefore, the amount of variables provided and the complexity of the regression problem can cause some uncertainty in the OLS model. Though as the problem in this project is a regression problem, an OLS model should be attempted in order to test if it will suffice with predictions. The OLS model is used upon all previously mentioned iterations of the datasets as to compare the datasets success rate. Hindman (2015)



(a) Property price predicting on the House dataset using OLS

(b) Property price predicting on the Combined dataset using OLS

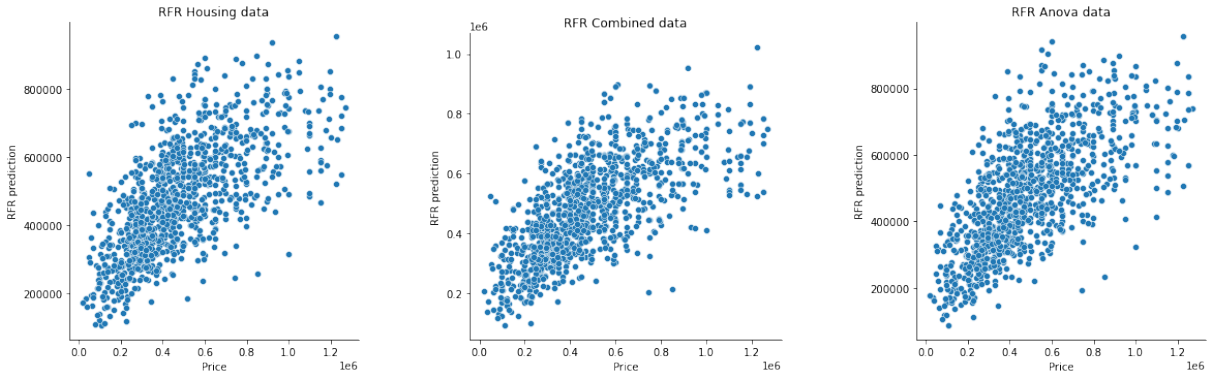
(c) Property price predicting on the Anova dataset using OLS

Figure 6.6: The figure displays the distribution of property prices by using OLS. The y-axis displays the predicted property prices while the x-axis shows the real observations. (Appendix B)

As shown above in figure 6.6 there are some similarities of the distribution for the prediction of property prices. None of the graphs have a linear distribution which indicates that the OLS performance among the different datasets is insufficient at making precise predictions regardless of the input data.

Random Forest Regression [RFR] Model

Contrary to the OLS model, the Random forest regressor algorithm keeps splitting the data into different decision trees where it is based on conditions. The RFR keeps reducing the data until there are no observations left. The algorithm selects the best current plot that maximizes the information gain. The algorithm does not backtrack and change the previous split. Therefore, are all the following splits depended on the previous and thereby does not guarantee that it gets the most optimal set of splits. The amount of mistakes in the different states decreases through the splits. As the complexity of the prediction problem increases, the efficiency of the RFR model decreases. This means that our data has a possibility to not be a good fit for the model. Segal (2004)



(a) Property price predicting on the House dataset using RFR

(b) Property price predicting on the Combined dataset using RFR

(c) Property price predicting on the Anova dataset using RFR

Figure 6.7: The figure displays the distribution of property prices by using RFR. The y-axis displays the predicted proverty prices while the x-axis shows the real observations. (Appendix B)

The figure 6.7 shows the same tendency as in figure 6.6 where there is no clear linear distribution.

6.2.2 SML results

After running all 3 datasets through both models, the following results were obtained:

R2 Supervised machine learning

	House	Combined	Anova
OLS	0.362646	0.415778	0.416261
RFG	0.471588	0.481586	0.471927

Figure 6.8: The table displays the $R\hat{2}$ for the OLS and RFG. (Appendix B)

As shown in figure 6.8 most of the $R\hat{2}$ scores are both quite low with the initial dataset, with scores of 0.36 and 0,47 for the OLS and RFR model respectively. This indicates that the supervised machine learning models algorithms are not very suited for this problem. With that being said, both models improve with changing the dataset. While the RFR model only sees a small improvement Combined dataset, the OLS model gets a sizeable increase to 0.41. Finally the Anova dataset included at the end affects them differently, albeit marginally.

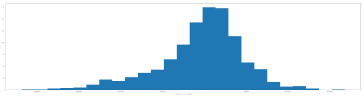
Because of the poor results of the SML models, neural network models might improve the ability to predict the property prices. At the same time, the neural network allows for more model tuning based on the input data.

6.2.3 Neural network

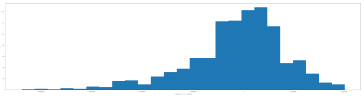
Neural network modelling can take some trail and error, which causes it to usually take more time and tweaking before a preferable model setup can be found. The method for making and testing models is starting out with a relatively simple model, and add layers, neurons and changing activation functions as the models are updated. These models are then tested upon using the three dataset iterations as to compare both the data compositions and model iterations.

1st model iteration [Dense]

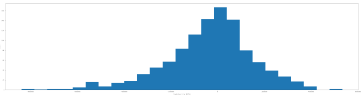
The first neural network model is based upon 3 dense layers where the first two dense layers have the 'relu' activation function and the 'adam' optimizer. The adam optimizer is used on all the next iterations of the model as it is good at quickly finding a local minimum. These layers have 35, 10 and 1 neurons respectively and provides the following results:



(a) Distribution of errors by using iteration 1 on the House data.



(b) Distribution of errors by using iteration 1 on the Combined data.



(c) Distribution of errors by using iteration 1 on the Anova data.

(a) Predicted vs True observations for model iteration 1. on the House data.

(b) Predicted vs True observations for model iteration 1. on the Combined data.

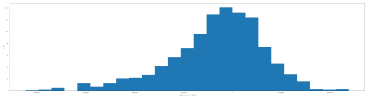
(c) Predicted vs True observations for model iteration 1. on the Anova data.

Figure 6.10: The figure shows the distribution of results among the different dataset, by using the first iteration of the model. (Appendix B)

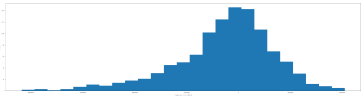
The figure displays iteration 1 (Dense) neural network to show the difference among the datasets "House, Combined and Anova". The first row from (a)-(c) represents the distribution of price prediction errors where the y-axis counts the observations while the x-axis shows the price difference. The second row from (a)-(c) shows the prediction of property prices. The blue line shows the true property prices while the orange line shows the predicted property prices. The y-axis is the price while the x-axis is the house ID.

2nd model iteration [Gaussian]

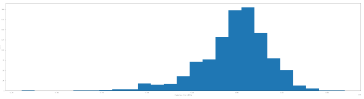
The second neural network model is based upon 3 dense layers and a gaussian dropout layer, as to provide some noise to the data. This is primarily done as so create some variance as well to counteract any artificially produced reduction of standard deviation when imputation of missing values were made. Additionally, the dropout part of this layer reduces the neuron count, which can reduce overfitting of the dataset. The first and the third dense layer have the 'relu' activation function and the last dense layer has the default linear activation function. These layers have 35, 35, 10 and 1 neurons respectively and provides the following results:



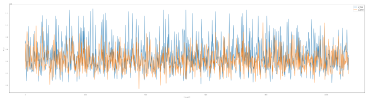
(a) Distribution of errors by using iteration 2 on the House data.



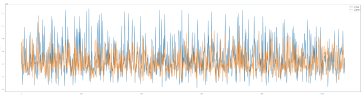
(b) Distribution of errors by using iteration 2 on the Combined data.



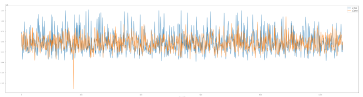
(c) Distribution of errors by using iteration 2 on the Anova data.



(a) Predicted vs True observations for model iteration 2. on the House data.



(b) Predicted vs True observations for model iteration 2. on the Combined data.



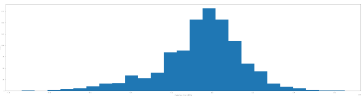
(c) Predicted vs True observations for model iteration 2. on the Anova data.

Figure 6.12: The figure shows the distribution of results among the different dataset, by using the second iteration of the model. (Appendix B)

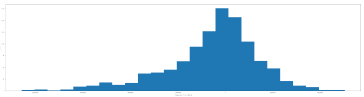
The figure displays iteration 2 (Gaussian) neural network to show the difference among the datasets "House, Combined and Anova".

3rd model iteration [High Neuron Count]

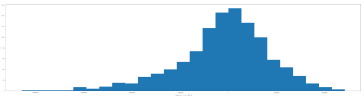
The third neural network model is based upon 4 dense layers and a gaussian dropout layer, as to provide some noise to the data. This is primarily done as so create some variance as well to counteract any artificially produced reduction of standard deviation when imputation of missing values were made. Additionally, the dropout part of this layer reduces the neuron count, which can reduce overfitting of the dataset. The first, third and fourth dense layer have the 'relu' activation function and the last dense layer has the default linear activation function. These layers have 64, 64, 32, 16 and 1 neurons respectively and provides the following results:



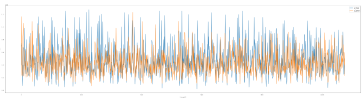
(a) Distribution of errors by using iteration 3 on the House data.



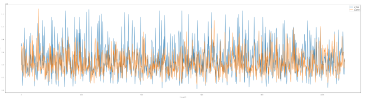
(b) Distribution of errors by using iteration 3 on the Combined data.



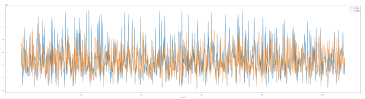
(c) Distribution of errors by using iteration 3 on the Anova data.



(a) Predicted vs True observations for model iteration 3. on the House data.



(b) Predicted vs True observations for model iteration 3. on the Combined data.



(c) Predicted vs True observations for model iteration 3. on the Anova data.

Figure 6.14: The figure shows the distribution of results among the different dataset, by using the third iteration of the model. (Appendix B)

The figure displays iteration 3 (High Neuron Count) neural network to show the difference among the datasets "House, Combined and Anova".

6.3 Comparative analysis

The figure below shows different tables of the numerical model metrics:

MAE neural network

	House	Combined	Anova
Simple dense model	0.102931	0.103657	0.107136
Gaussian dropout model	0.106840	0.105222	0.107142
Higher neuron count model	0.096188	0.101559	0.106418

MAE supervised machine learning

	House	Combined	Anova
OLS	0.119415	0.111982	0.111932
RFR	0.105826	0.104394	0.105219

MSE Neural network

	House	Combined	Anova
Simple dense model	0.020990	0.020687	0.021431
Gaussian dropout model	0.021934	0.020926	0.021473
Higher neuron count model	0.018596	0.020028	0.021456

MSE Supervised machine learning

	House	Combined	Anova
OLS	0.024147	0.022134	0.022116
RFR	0.020176	0.019520	0.019973

R2 Supervised machine learning

	House	Combined	Anova
OLS	0.362646	0.415778	0.416261
RFG	0.467453	0.484779	0.472819

Figure 6.15: The figure represents the results of the neural network and supervised machine learning models. (Appendix B)

Comparing the final results achieved with the various neural network models, shown in figure 6.15, the models did not significantly improve. The untuned supervised machine learning algorithms achieved slightly higher than 0.1 MAE on the scaled data (which equates to the prediction missing the actual price by about 10% as compared to the entire scale of prices). Improvements to this, by any meaningful amount, only happened on a single neural network model iteration (Higher Neuron Count model on House data). The reason for this can be manyfold, such as lack of knowledge to properly construct an advanced enough neural network model, or the chosen data not being specific enough to allow the model to utilize it.

Observing the progress through the dataset iterations, the neural network generally performed worse as the input data was altered, albeit marginally. The biggest variation comes in the final model setup with high neuron counts, where the MAE increased by almost 0.05 after adding the socioeconomic data to the input (Combined data). With that being said though, since many of the other iterations have minor changes in results, it can be hard to justify why the changes occur. Because of the short time span of the project there was not the opportunity to go in depth with more iterations of both the model and data. It is possibility that more trial and error, would have allowed the researchers to better test construction of both the model setup, and the input data to improve the results.

As it currently stands the study determines that the best way for the model to predict property pricing would be without the social data included. But this could simply be due to the way the data was accessed and utilized. The researchers preferred that the socioeconomic data had been a lot more disaggregated, to include identifiers of unique neighborhoods. The transition from the combined dataset to the filtered version utilizing ANOVA, also did not necessarily provide improvements. It can be observed that while this filtering was minor (4 variables removed), it generally made most of the models perform slightly worse. On the contrary, the model iteration changes brought some form of improvement to the models. It can be assumed this is because the initial model was very simplistic.

In conclusion the group has managed to vaguely improve the results from the initial supervised machine learning results. On top of this the group managed to slightly improve the results again, by attempting to tune the model with complexity, such as gaussian dropout layers and adding neurons. At the same time the SML models and NN models both reacted differently to the addition of the socialeconomic dataset. The SML models generally saw a small improvement in their performance, the NN models reacted varyingly.

Chapter 7

Reflection

The results through the analysis show that a more complex neural network causes a reduction in MAE, but there of course is more to consider. Bebis and Georgiopoulos (1994, p.30) writes the following *... to keep the depth of the network small and to minimize the number of connections to the hidden and output node.*

On one hand, this matches the approach that has been used since the construction of our neural networks, was based on keeping the models rather simple. On the other hand, the group may lack the knowledge of creating an optimized neural network, and may be limited by bounded rationality as well. Thereby this will create some uncertainty about the obtained results. The performance of the neural networks could be verified through comparison by applying the models on another dataset.

The project's overall weak point is the number of observations and the aggregation level. The analysis in section 6.1.1 demonstrated that the mean of observations in cities in the state of Bayern was 3.5. The initial idea behind the project was to get socioeconomical data based on cities and thereby potentially creating a neural network where the features were disaggregated.

The project scope may have been too large for the group to handle, as the previously mentioned limitations has cause the prerequisites for the projects success to be poor.

Chapter 8

Conclusion

The purpose of the conclusion is to provide an answer to the following thesis statement:

How can societal intangible variables generate a meaningful reflection on property prices, and are there benefits by connecting socioeconomic variables to property prices?

The underlying assumption of the thesis statement is, that socioeconomic variables add explanatory value to the prediction and assessment of property prices. This assumption had to be tested as to provide a comparative explanation of, whether the hypothesis could be confirmed or dismissed. The results of the different machine learning model- and data compositions, does not indicate that this is the case. However, contrary to the initial assumption of socioeconomics variables having an effect on property pricing, the immediate answer is that it decreases the performance of the tested models when including these variables. At the same time, increases in model complexity, transitioning from supervised machine learning- to neural network models, does not increase prediction precision substantially.

Limitations present in this project setting, is regarded as having hindered the probability of success of the model predictions, throughout the course of the project development. Therefore, even as a proof of concept, the project findings have little conceptual value in its current state. Though this study was unable to confirm the thesis statement, it is equally reluctant to deny the concept of there actually existing explanatory value in the combination of socioeconomics and housing prices as referenced by empirical evidence.

As the primary project limitations lies within the collection and validation of data, the group acknowledges that this project was better suited for testing by and/or with authorities within a 'General Data Protection Regulation' (GDPR) safe environment. As such, further investigation into the matter would enable the researchers to utilize fully disaggregated data as well as gain access to previously mentioned urban science parameters for the relevant areas.

As to what the benefits of connecting socioeconomic variables to property prices is, the previous limitations could be considered a boundary for enabling its use in the private sector. The monetary value of evaluating societal development and changes, could prove a valuable tool for real estate traders and investors etc. The more realistic approach could be in the public or even political sector, as these may have an interest in observing and changing e.g. socially vulnerable neighborhoods.

Bibliography

Arbnor, I., & Bjarke, B. (2009). *Methodology for creating business knowlegde* (3. ed.). Nota.

Bartels, C., Göth, A.-M., & Nachtigall, H. (2019). *Soep-core v34: Codebook for the eu-silc-like panel for germany based on the soep.*

Bebis, G., & Georgiopoulos, M. (1994, October-November). Feed-forward neural networks. *IEEE Potentials (Volume: 13, Issue: 4)*, pp. 27-31. (DOI: 10.1109/45.329294)

Berson, D. W., & Berson, D. L. (1997). The importance of demographics in economic analysis: The unusual suspects. *Business economics, Vol. 32*, pp. 12-16.

Bowen, D. J., & Quintiliani, L. (2019). *Socioeconomic influences on affordable housing residents: Problem definition and possible solutions.*

Braubach, M., & Savelsberg, J. (2009). *Social inequalities and their influence on housing risk factors and health.*

Buch-Hansen, H., & Nielsen, P. (2005). *Kritisk realisme.* Roskilde Universitetsforlag.

Conger, R. D., Conger, K. J., & Martin, M. J. (2010, June 18). Socioeconomic status, family processes, and individual development. *J Marriage Fam 72(3)*, 685-704. (`doi.org/10.1111/j.1741-3737.2010.00725.x` d. 3 january)

Eatwell, J., Millgate, M., & Newman, P. (1989). *Social economics* (1. ed.). The Macmillan Press Limited.

Foote, K. D. (2021, December 3). A brief history of machine learning. *dataversity.* (hentet fra `https://www.dataversity.net/a-brief-history-of-machine-learning/#` d. 29 December 2021)

Fund, I. M. (n.d.). *Housing prices continue to soar in many countries around the world.* (Hentet fra `https://blogs.imf.org/2021/10/18/housing-prices-continue-to-soar-in-many-countries-around-the-world/` d. 8. January 2022)

Galarnyk, N. (2018, September 12). Understanding boxplots. *Towards data science.* (`https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51` d. 5 january)

Global house price index q3 2021. (2021). *content.knightfrank.* (https://content.knightfrank.com/research/84/documents/en/global-house-price-index-q3-2021-8652.pdf)

Haghdoost, A. A. (2012, February 3). Complexity of the socioeconomic status and its disparity as a determinant of health. *Int J prev Med*, 75-76. (journal number3(2))

Haviv, Y. (2019, July 8). Why is it so hard to itegrate machine learning into real business applications. *Towards data science.* (hentet fra `https://towardsdatascience.com/why-is-it-so-hard-to-integrate-machine-learning-into-real-business-applications-69603402116a` d. 29 December 2021)

Hindman, M. (2015, April 9). Building better models: Prediction, replication, and machine learning in the social sciences. *Volume: 659 issue: 1*, pp. 48-62. (https://doi.org/10.1177/0002716215570279d. January 7)

Ismail, Z., Kalid, H., & Khamis, A. (2005, January). The effects of outliers data on neural network performance. *Journal of Applied Sciences 5(8)*, pp. 1394-1398.

Jacobsen, M. H., Nedergaard, P., & Rasmussen, K. L. (2015). *Videnskabsteori i statskundskab, sociologi og forvaltning* (2. ed.). Hans reitzels forlag.

Jaggia, S., & Kelly, A. (2019). *Business statistics and communicating with numbers* (3. ed.). McGraw-Hill education.

Kotsiantis, S., & Pintelas, P. (2007). Supervised machine learning: A review of classication techniques. *Emerging artificial intelligence applications in computer engineering, 160(1)*, pp. 3-24.

Lichao, Y., & Walker, R. (2020). *Beyond the goal of eradicating absolute poverty in china: relative poverty indicators and social security policies.*

Lutz, M. A. (2009). Chapter 67: Social economics. In J. Peil & I. van Staveren (Eds.), *Handbook of economics and ethics* (p. pp. 516-522). Cambridge: Cambridge University Press.

Mclaunghlin, K. A., & Sheridan, M. A. (2016, June 2). Neurobiological models of the impact of adversity on education. *Current Opinion in Behavioral sciences*, 108-113. (x)

McLeod, J. D. (2013). Chapter 67: Social economics. In C. S. Aneshensel, A. Bierman, & J. C. Phelan (Eds.), *Handbook of the sociology of mental health* (p. pp. 229-255). Springer, Dordrecht.

Panel, S.-E. (2021). *Research data center soep.* (Hentet fra https://www.diw.de/en/diw_01.c.678568.en/research_data_center_soep.html d. 26. dec 2021)

Peachey, B. K. (2021, October 28). Budget 2021: Surge in house prices predicted to slow. (Hentet fra https://www.bbc.com/news/business-59078455 d. 8. January 2022)

Reed, R. (2001, February). The significance of social influences and established housing values. *The Significance of Social Influences and Established Housing Values, University of Queensland, Australia Pacific Rim Real Estate Society Conference 2001, Adelaide.*

Segal, M. R. (2004, April). Machine learning benchmarks and random forest regression. *Center for bioinformatics and Molecular Biostatistics.*

Seref, E. (2020). *German house prices.* (Hentet fra https://www.kaggle.com/scriptsultan/german-house-prices d. 26. dec 2021)

Sharma, S. (2017, September 6). Activation functions in neural networks. (hentet fra https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6 d. 27 December 2021)

Statkat. (n.d.). *Pearson correlation - overview.* (Hentet fra https://statkat.com/stattest.php?t=19&t2=12 d. 7. January 2022)

Social Value UK. (2021). *What is social value?* (Hentet fra https://socialvalueuk.org/what-is-social-value/ d. 21. dec 2021)

TylerVigen.com. (2021). *Spurious correlations.* (Hentet fra http://www.tylervigen.com/spurious-correlations d. 25. dec 2021)

VanderPlas, J. (2016). *Python data science handbook*. O'Reilly Media, Inc.