Predictive Modeling of Property Valuations in Denmark: The Role of Socioeconomic Factors

Mert Cetinkaya[1], Dilek Coskun[1], and Jonathan Hoch[1]

[1]*Department of Mathematical Sciences, University of Copenhagen,
Copenhagen, Denmark*

August 2023

Contents

Group Member	Largest Contributor to Section	Largest Contributor to Code
Mert	Theory, Empirical Results	Machine Learning, Home web scraper
Dilek	Literature Review, Empirical Results	Boligsiden web scraper
Jonathan	Introduction, Background	API-Integrations, Boliga web scraper, Machine Learning

1 Introduction

The valuation of properties, particularly residential real estate, has long been a subject of intense scrutiny and concern for a multitude of stakeholders—ranging from governmental bodies to real estate investors and the Danish (especially young) population. The reason for this widespread interest is straightforward: fluctuations in property values have far-reaching consequences on Denmark's economic landscape [1]. Since the Financial Crisis of 2008, the Danish real estate market has witnessed a consistent uptick in both the volume of property sales and their associated prices [2].

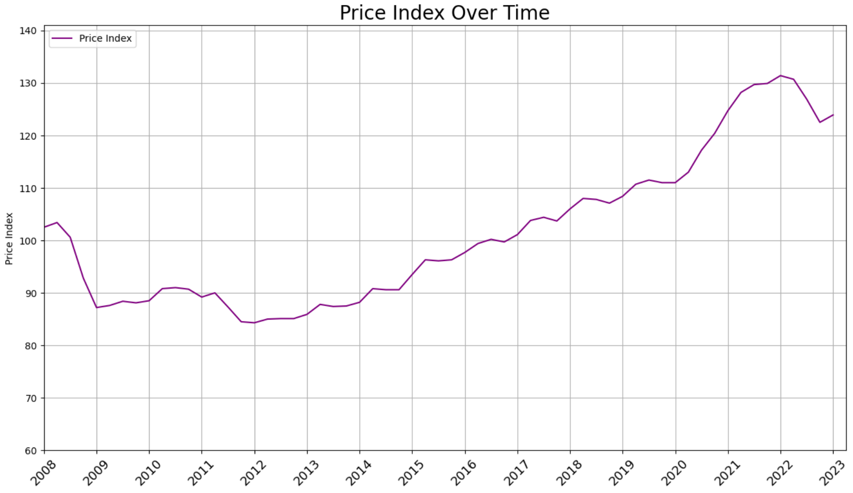


Figure 1: Historical price index for family homes from Danmarks Statistik

While this trend suggests a thriving market, it also underscores the need for a sophisticated understanding of the factors that shape property valuations [5, 6]. This paper aims to delve into the complexities of property valuations in Denmark, focusing on the conditions on a municipal level that influence these valuations.

This research contributes to the scientific community by employing a multidisciplinary approach that integrates economics, urban planning, and data science to understand property valuations. By utilizing machine learning algorithms, this study aims to address limitations of traditional valuation models, offering a more dynamic and adaptive model for property valuation [6]. Furthermore, the comprehensive dataset used in this research, encompassing real-time property listings, historical records, and socioeconomic indicators, sets a new standard for data richness and granularity in the field. These methodological advancements not only provide a more nuanced understanding of the Danish real estate market but also offer a framework that can be adapted for studying property markets in different geographical and economic contexts.

In summary, this paper tackles an important gap in existing research by using data-driven methods to analyze property values in Denmark. By using machine learning and a comprehensive data set, we aim to answer important questions about prediction property valuations. While this study is certainly relevant for people interested in the Danish property market, it also contributes to the global scientific discourse about the future of social data science.

1.1 Theoretical Framework and Methodology

The following section will briefly explain the methodological and scientific-theoretical reflections that underpin this study.

Based on the scientific philosophy of positivism, which argues that reliable knowledge can be obtained through observable and numerical data, our study adopts a robust data-centric approach. This approach is particularly beneficial in today's data-abundant environment and it is crucial for the machine learning models we use, as they generally perform better with more data.

Our research draws from Big Data theory, advocating the use of large datasets for better insights. To collect mass amounts of property- and socio-economic data, we use data aggregation techniques such as web scraping and API integration. This approach also aligns with the Contextualism theory, which asserts that a property's value isn't merely dictated by its inherent qualities, but is also substantially impacted by the surrounding socio-economic and geographical conditions. By building a comprehensive dataset that combines property specific attributes with wider contextual elements, our study aims not to overlook this aspect.

It's important to acknowledge the limitations of our methodological approach. One issue is the potential for inaccuracies or incomplete data, especially when consolidating information from various sources through web scraping and APIs. Moreover, our dependence on "Big Data" opens the door for inheriting biases present in these larger datasets, which could, in turn, affect the generalizability of our results. These limitations will be discussed in greater detail in a subsequent section of this paper.

2 Background & Literature Review

Before delving into modeling for property valuations in Denmark, it is essential to establish the academic context within which this study is situated. The following Background and Literature Review sections provide an overview of and prior studies that have influenced our grasp of property valuation.

2.1 Post Financial Crisis Market & Model Selection

The 2008 Financial Crisis serves as a pivotal moment in the history of property valuations, fundamentally altering the economic landscape and influencing a multitude of variables that affect property prices. Research on the Greek property market by Kallergis, Kavvathas, and Kounetas (2019) provides valuable insights into how structural and financial factors have been impacted post-crisis. Their study reveals significant changes in property valuations, with variables like quality of construction and financial factors such as unemployment and inflation becoming increasingly important [3].

Given these profound shifts, focusing on the period after 2008 offers a more relevant and timely analysis for the current study on the Danish property market. This approach allows us to capture the complexities and nuances introduced by the crisis, which are still reverberating in property valuations today. Moreover, the Danish real estate market has seen a consistent increase in property sales prices since 2008 [2]. By limiting our study to the post-2008 period, we aim to provide a more accurate and contextually relevant predictive model that accounts for the long-term effects of the crisis on property valuations in Denmark. This focus aligns well with our multidisciplinary approach, which seeks to integrate economics, urban planning, and data science to offer a comprehensive understanding of property valuations in a post-2008-financial-crisis world.

The choice of machine learning model is critical for the success of any predictive analysis. In the context of our study on the Danish property market, the decision to employ a linear regression model is informed by several factors. Most notably, data from DST shows that the development in property prices has followed a somewhat linear trend since 2012:

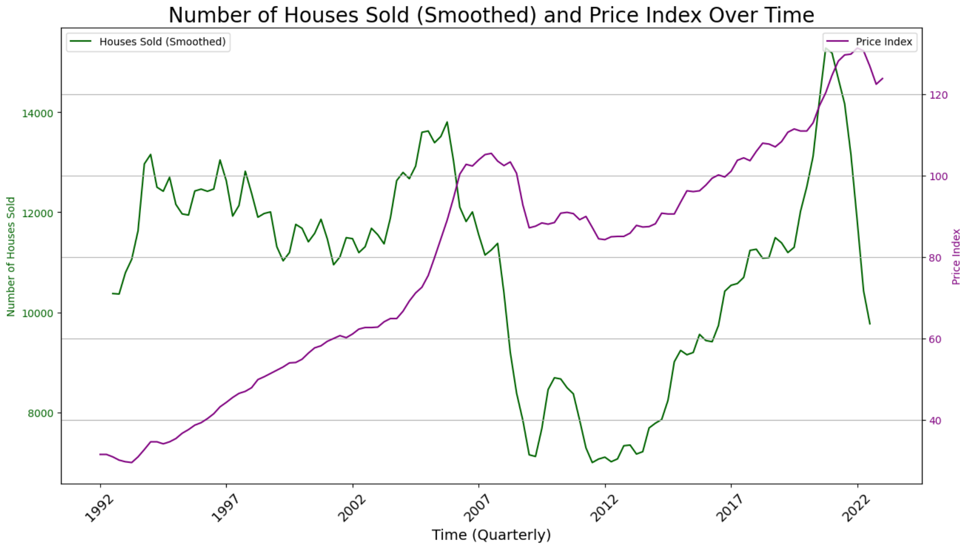


Figure 2: - Houses sold and Price Index over time

This observed linearity in the data provides a strong rationale for using linear regression, as it suggests that the relationship between the dependent variable (property prices) and the independent variables (socioeconomic factors) can be adequately captured through a linear equation.

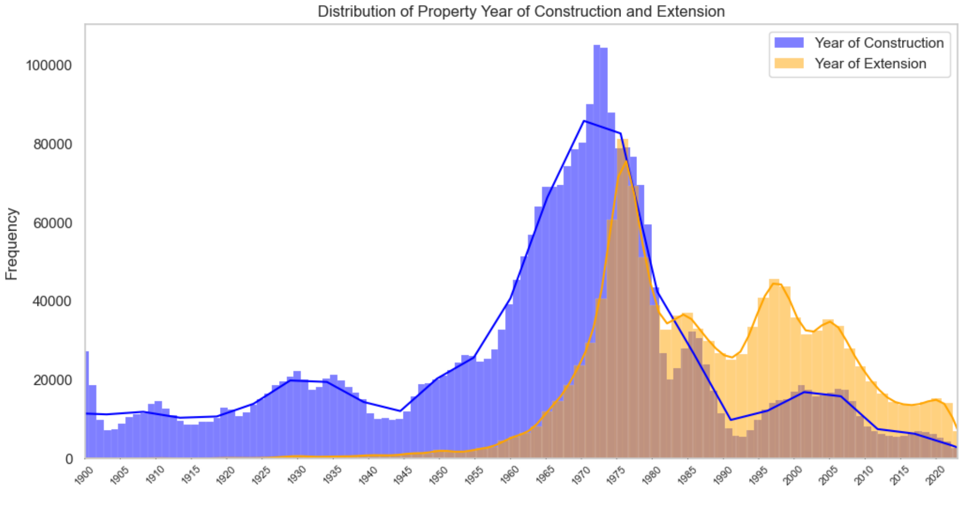


Figure 3: - Property Construction and Extensions

Linear regression models are particularly effective for analyzing and forecasting trends that follow a linear pattern, such as those seen in the Danish property market after 2012, and they have the benefit of being easy to interpret meaning we can easily understand how each factor impacts property values. However, the linear regression makes assumptions such as linearity, independence, and homoscedasticity. While our data suggests a linear trend in property prices, the model may not capture more complex, non-linear relationships between variables. One of these variables could for example be a sudden influx in houses being built (see fig. 3). Additionally, the assumptions of independence among predictors and constant variance (homoscedasticity) are not guaranteed.

2.2 Gaps in Existing Research

In this study, we focus on the interplay between property valuations and a range of socioeconomic indicators. As previously mentioned, our aim is to understand and model fluctuations in real estate prices by considering factors such as income levels, population density, unemployment.

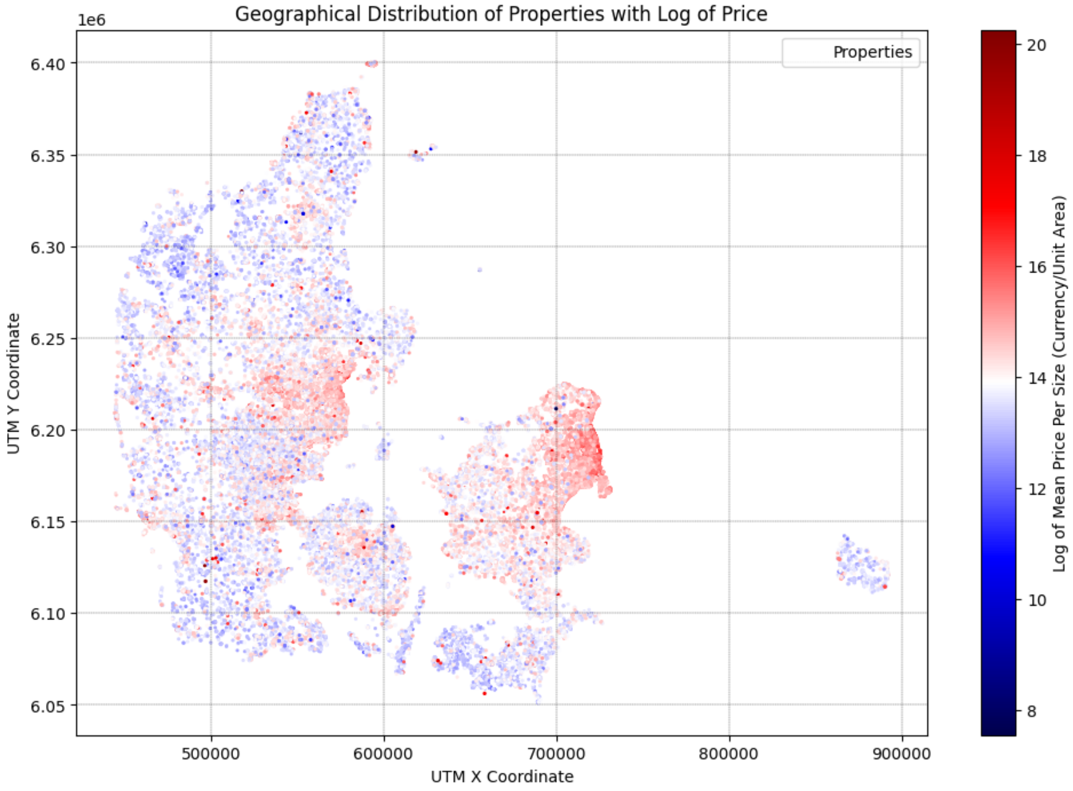


Figure 4: - Geographical scatterplot of the data

The decision to focus on a municipality-level analysis is not arbitrary but is inspired by existing literature. For instance, a study on the Apulia region in Southern Italy emphasized the importance of examining property prices at the municipal level to capture the nuanced interplay between locational and socioeconomic variables [4]. This level of detail helps us see that distinct real estate pricing patterns can emerge within broader geographic regions. This is especially relevant for Danish municipalities, where, as we can we see above, there is a significant disparity in real estate prices between urban and rural areas.

Taking a look at the scatterplot (generated from our data), we see that real estate prices in different regions differ, for example, real estate prices are higher in big cities, while lower prices in rural areas come to the fore. This contributes to our understanding of how the real estate market is affected by variables such as geographical factors and socioeconomic conditions.

There is indeed a multitude of existing studies and models used to predict property demand. These approaches typically blend historical data, economic indicators, demographic factors, and advanced analytical techniques to forecast trends in the real estate sector. Here, we outline the main features of some common approaches and model.

The article “Analyzing the local geography of the relationship between residential property prices and its determinants” which examines the relationship between house prices in local geography and the factors affecting these prices. Sales prices and structural and spatial features obtained from the Malaysian Valuation and Services Department database were used for the analysis. This research was obtained by using semi-parametric geographic weighted regression (S-GWR) technique to determine house prices. Reveals a strong geographically varying relationship between housing prices and their determinants. While the determinants of housing prices have a positive effect on prices in some regions, they have a negative or no effect on others. The magnitude of the effect was also found to vary geographically; Capitalization in housing prices is greater in some areas, while in others it has less or no impact. [1]

The article “House price modelling of Denmark’s municipalities using vector autoregression and gradient descent” represents a thesis that explores the use of a house price evolution equation to model the real estate market in municipalities across Denmark. The author proposes two methods to solve the equation: a recursive optimization algorithm and a closed-form solution using Vector Autoregression (VAR). The study examines and compares the quality of these solutions, focusing on the accuracy of price predictions and the relationship between model parameters and expected properties. The research finds that the closed-form solution method performs poorly in system identification, while the recursive method produces much better models. Depending on the initial conditions, the recursive method captures expected properties more accurately, with price predictions within 5% over a four-year period. It is observed that the distance between municipalities has a relatively significant impact on price relationships, while the population size does not exhibit any noteworthy mutual effects. The study also highlights the importance of price inflation, a significant factor that should be applied more precisely in future research.
The model has many parameters, and the presence of many local minima is expected. To address this, the study employs substantial actions, such as out-of-sample validation and utilizing parameter starting values to examine the model’s learning process. The study concludes that this research could be highly beneficial for governments and society. [2]

3 Theory

3.1 Linear regression

When predicting housing prices, we have to account for numerous factors or predictors, that can range from the number of toilets to its age, location, and many more. And since we work with multiple variables, it then dictates that our predictive model operates in a multi-dimensional space and is given by the following linear regression model:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

[1] *Analyzing the local geography of the relationship between residential property prices and its determinants* by Mohd Faris Dziauddin, Kamarul Ismail and Zainudin Othman

[2] *House price modelling of Denmark’s municipalities using vector autoregression and gradient descent* by Love Gillberg

Put in simpler terms, a linear regression is a prediction based on the weighted sum of its features and the intercept term.[3]

3.2 OLS & Error

The method used to estimate the weighted coefficients (θ_i), is called OLS, Ordinary Least Squares. OLS seeks to find the line or linear function that minimizes the sum of squared deviations from the observed values.[4] The linear regression model is given by:

$$y = \theta_0 + \sum_{i=1}^{n} \theta_i x_i + \varepsilon$$

$$\Leftrightarrow y = \hat{y}(\theta) + \varepsilon$$

where the actual housing price is equal to the predicted housing price, plus the error or residual, that is the difference:

$$\Leftrightarrow \varepsilon = y - \hat{y}(\theta)$$

Hence, OLS determines estimations of the weighted coefficients that minimizes the sum of the squared deviations:

$$OLS : \min_{\theta} \sum_{j=1}^{m} \left(y_j - \hat{y}_j(\theta)\right)^2$$

3.3 Bias-Variance Trade-Off

Having introduced OLS, we will henceforth denote $\hat{\theta}_{OLS}$ as $\hat{\theta}$. In statistics, especially in parameter estimation, there are two important characteristics to consider are bias and variance. Bias measures how much the average/expected estimation deviates from the true parameter:[5]

$$Bias\left(\hat{\theta}\right) = E\left(\hat{\theta}\right) - \theta$$

Variance measures how much the parameter estimates vary around their expected values. In the world of regressions we have to find the optimal model complexity, since an error too small will eliminate variance and cause overfitting, while an error too big will increase bias and cause underfitting. To understand the trade-off between bias and variance, and error, look at the illustration below:

[3] *Hands-On Machine Learning with Scikit-Learn and TensorFlow* by Géron, Aurélien. P. 106

[4] *Ordinary Least Squares* by Wikipedia

[5] *Regularization in R Tutorial: Ridge, Lasso and Elastic Net* by Datacamp

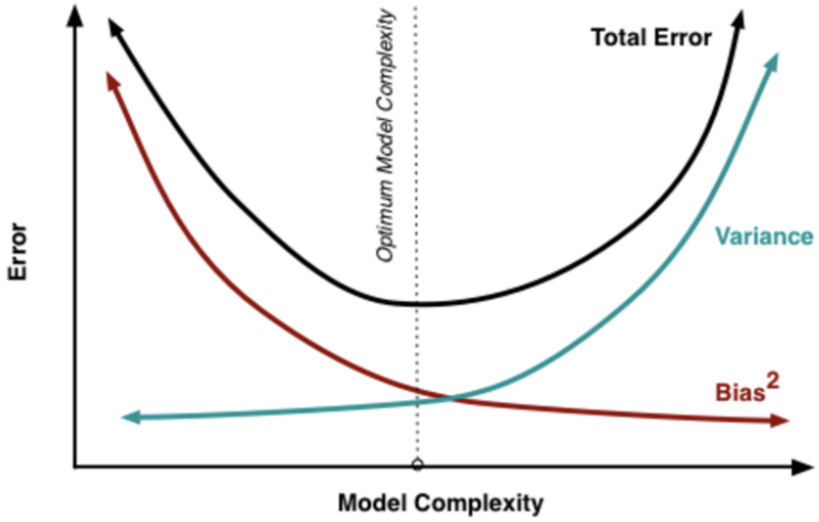


Figure 5: Illustration showing the theoretical purpose of optimum model complexity from Datacamp

To find the optimal model complexity, we have to solve the minimization problem in the following equation:[6]

$$E(e) = (E(\hat{\theta}x) - \theta x)^2 + E(\hat{\theta}x - E(\hat{\theta}x))^2 + \sigma^2$$
$$\Leftrightarrow E\left((y - \hat{y})^2\right) = Bias^2 + Variance + \sigma^2$$

MSE

We will now introduce a method that is used to asses the quality of our model's predictions compared to the actual data. The method is called Mean Squared Error, MSE, and is a statistical tool that measures the average of the squared differences between the predicted values and actual values.[7] We have actually already derived the MSE-function:

$$MSE = E\left((y - \hat{y})^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Minimizing MSE

Now that we have laid the groundwork for the theory behind our actions, we can finally focus on the practical part. We now want to minimize MSE by finding the optimal regularizing hyperparameter λ.

Lasso regression

The Lasso regression augments the OLS approach by adding a penalty on the coefficients:[8]

[6] *Regularization in R Tutorial: Ridge, Lasso and Elastic Net* by Datacamp
[7] *Mean squared error* by Wikipedia
[8] *Regularization in R Tutorial: Ridge, Lasso and Elastic Net* by Datacamp

$$L_{lasso}(\hat{\theta}) = \min_{\theta} \left(\sum_{j=1}^{m} (y_j - \hat{y}_j(\theta))^2 + \lambda \sum_{i=1}^{n} |\hat{\theta}_i| \right)$$

The L1 penalty in Lasso regression is controlled by the hyperparameter λ, which directly influences the degree of shrinkage applied to the coefficients. As λ increases, the coefficients are driven towards zero, and some may be exactly zero. This property of the L1 penalty allows Lasso regression to perform feature selection, effectively eliminating non-important features by setting their corresponding coefficients to zero.

Ridge regression

Just like the Lasso regression, the Ridge regression also adds a penalty on the coefficients:[9]

$$L_{ridge}(\hat{\theta}) = \min_{\theta} \left(\sum_{j=1}^{m} (y_j - \hat{y}_j(\theta))^2 + \lambda \sum_{i=1}^{n} \hat{\theta}_i^2 \right)$$

The L2 penalty in Ridge Regression is controlled by the hyperparameter λ, which directly affects the extent of shrinkage applied to the coefficients. As λ increases, the coefficients are driven towards zero, but unlike L1 penalty, they are not set to exactly zero. This property of the L2 penalty leads to a reduction in the model's complexity by constraining the coefficients, but it does not perform feature selection by eliminating them. The Ridge Regression thus includes all features in the model but with reduced influence when λ is high, helping to mitigate overfitting, especially in cases of numerous features.

Computational method to determine hyperparameters

The optimal hyperparameter is the one that minimizes MSE, and so we set out to estimate this parameter through machine learning. Our first step is using K-Fold Cross-Validation to determine our hyperparameter. The process starts by splitting our data set into 5 equal sized folds, where we then train the model on 4 folds and test it on the fifth. We then repeated this process 5 times, where we for each iteration used the hyperparameter on the test fold to determine MSE. We then computed the average MSE across all iterations for each chose hyperparameter and chose the λ that resulted in the lowest average MSE.

4 Data Acquisition and Ethical Considerations

The following section provides a comprehensive overview of the data attributes and the sources from which data is collected, the methods employed for data collection and preprocessing, as well as the ethical guidelines adhered to throughout the study.

The scripts and code mentioned throughout this section have been shared (excluding of course secrets such as API tokens and authentication details).

4.1 API integration

The emergence of big data has necessitated robust pipelines capable of collecting, transforming, and integrating data from a variety of sources. To address this challenge, we've developed a pipeline

[9] *Regularization in R Tutorial: Ridge, Lasso and Elastic Net* by Datacamp

that utilizes Python's asyncio library. The architecture utilizes asynchronous API calls, incorporate rate limiting, caching, and queuing mechanisms to ensuring optimal performance and resource utilization. This approach is particularly crucial given the time-sensitive nature of our study. With a limited timeframe for data collection, the ability to gather data quickly without compromising on quality is of paramount importance.

Using this architechture, we successfully mapped and collected data from multiple registries, including:

- **BBR Grund Data**: "Building and Housing Register - Base Data"
- **BBR Bygning Data**: "Building and Housing Register - Building Data"
- **BBR Enhed Data**: "Building and Housing Register - Unit Data"
- **Ejendomsbeliggenhedsregistret (EBR) Data**: "Property Location Register Data"
- **Ejendomsvurdering (VUR) Data**: "Danish Property Assesment Agency - Valuation Data"
- **Matrikel Data**: "Land Register Data"
- **Danmarks Statistik (DST) Tabel Data**: "Statistics Denmark Table Data"

The asynchronous nature of the pipeline allows for concurrent data collection from the data sources, significantly reducing the time required to amass a comprehensive dataset. When we stopped data collection, we had effectively amassed over 4.000.000 rows of historical property evaluations, with data attributes including address, property size, valuation, size of garage, geographic coordinates, number of rooms, number of toilets and more. With each unique entity requiring mapping data from at least four databases, this illustrates the demands set by the data aggregating task.

The ethics surrounding data gathering, especially in the realm of big data, cannot be overstated. We were granted access to the data via API from "Styrelsen for Dataforsyning og Infrastruktur"'s department "Datafordeleren". This access was not only a testament to the trust placed in our research but also a reminder of the ethical obligations we have towards the data and its stakeholders. Our access to the data was transparent and consensual and "Styrelsen for Dataforsyning og Infrastruktur" was fully aware of the nature and objectives of our research. Having data resources like these freely accessible is a privilege, and as researchers, it is spececially important that we we adhere to stringent ethical standards. This includes adhereing to General Data Protection Regulation (GDPR) and, for example, only sharing aggregate data.

4.2 Web Scraping

In the following paragraphs, we will describe our web scraping efforts.

4.2.1 Boliga

One of the methods used in this study to obtain data is employing a web scraping technique using Selenium and Python. The target website for data extraction was boliga.dk, a leading property listing site in Denmark.

From this platform, we aggregated data related to current property listings and their respective prices. While the site offers in-depth details such as property size, number of rooms, and more, obtaining this data demands individual URL access for each listing. By instead utilizing the API data pipeline, we can obtain this data by mapping the address to our dataset, thus also reducing potential strain on boliga's servers.

Before conducting the web scraping, we ensured that our actions were in compliance with the terms of service of boliga.dk. This involved reviewing the url the robots.txt document:

```
User-agent: *

Disallow: /boligrapporter?*
Disallow: /vurderingsrapporten?*
Disallow: /projektsalg/resultat*
Disallow: /projektsalg/kortsoegning*
Disallow: /indeks/arkiv*
Disallow: /indeks/bbr*
Disallow: /indeks/solgt*
Disallow: /bolig/statistik*
Disallow: /bbr/search*
Disallow: /salg/info*
```

Accessing and using data without permission can be both unethical and illegal, so we reached out to boliga and got their explicit permission to web scrape the site. After explaining the purpose of our webscraping they got back to us, allowing us to web scrape the site.

Naturally all ethical responsibilities don't cese here. We must also make sure that all the we data extracted was anonymized, ensuring no personal or identifiable information about property owners or listers was retained. Luckily, the nature of the attributes we were looking for made this very easy. However this is an important consideration since that it not only protects the privacy of individuals but also aligns with GDPR and other data protection regulations.

While being an effective tool in automating repetitive data aggregation tasks, the web scraping technique does have a couple drawbacks worth mentioning.

Accuracy and Completeness: Due to the dynamic nature of property listing sites, the data scraped from the site is rarely a complete representation of the entire dataset. In fact, sorting on oldest postings first turned out to be a reliable technique to asstist with manual debugging.

Robustness and reliability: By using XPATH and CSS element references, each web scraping script is tailored specifically for the site that they're meant to be scraping. The web-scrapers sensitivity towards the site that it is built on also means that it is sensitive towards any changes in the layout.

4.2.2 Boligsiden

First, we started by importing the requests and BeautifulSoup libraries. These libraries were very useful for extracting data from the website and analyzing this data. We did not encounter a big problem because there was not much data on the boligsiden page, there was a total of 10.000 data, and we tried to extract all 200 pages of data. One of the biggest problems we faced was when we were transferring our data to the dataframe, the number of our data was changing. This may be because the web site has changed the structure and layout of a particular page. In this case, BeautifulSoup or other data extraction methods may not be able to pull the correct data, or it is pulling incompletely. But this only prevented us from accessing very little data so we ignored that and visualized our data.

4.2.3 Home

Home's infinite scroll feature made it quite challenging to web scrape large quantities of data. Moreover, the necessity to open each house in a new tab to access all its information required substantial time - a resource we didn't have - so we had to find a workaround.

What we ended up doing was scraping the information from the infinite scroll page and then merging this data with API data. By doing so, we obtained the market price for each house, combined with the information we couldn't access through Home.

Our scraping process consisted of two loops. The outer loop was our solution to the problem of being unable to web scrape the infinite scroll website for 38,000+ properties, as this would completely overload our computer's memory and cause a crash. Our solution involved web scraping another site for all the postal codes in Denmark, from which we compiled a list using BeautifulSoup. We then encountered a new problem: only the page for the first postal code was fully scraped, with elements missing from the subsequent pages. Our workaround was to add a final step to this loop, returning to the infinite scroll page, and then accessing the next postal code. This approach eliminated the problem.

During our project, we faced an ethical dilemma. Before beginning our work, we reached out to Home and explained our intention to web scrape their site and the purpose behind doing so and with regard to the data itself, we only scraped anonymous and publicly available data, thereby maintaining ethical integrity.

5 Empirical Results

5.1 Analysis of our models

The table below presents the prediction errors for the three linear regression models we evaluated:

Model	MAE	MSE
OLS regression	558082.947832	7.135689e+12
Lasso regression	558087.330556	7.135690e+12
Ridge regression	558083.204459	7.135689e+12

The observed models have almost identical error values, with OLS regression having the smallest MAE value, and the smallest MSE value, tied with the Ridge regression. The data values being this similar suggests that the data may not benefit from regularization, since both the Lasso- and Ridge regression are providing very similar results to the non-regularized OLS regression. The fact that the OLS regression performed approximately as well as the Lasso regression can indicate that there are not many irrelevant features in the dataset, since Lasso excels in scenarios where feature selection is needed.

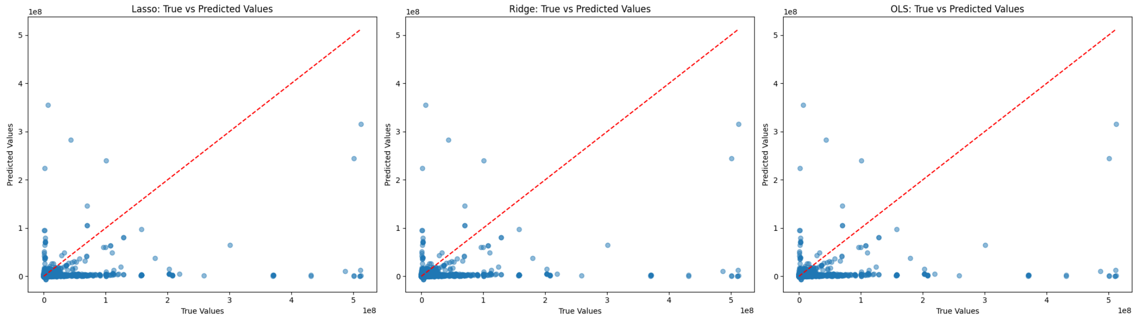
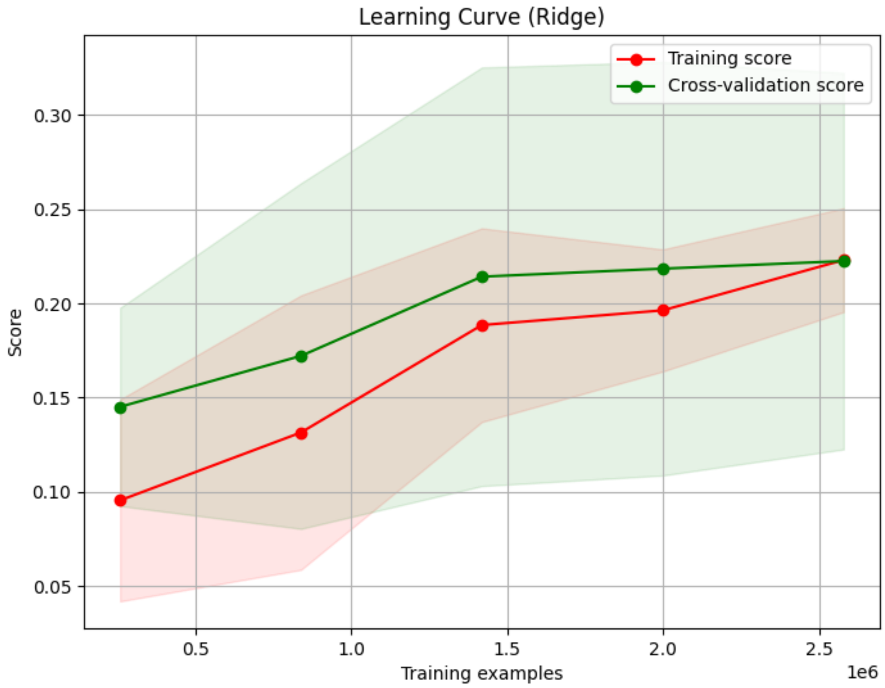


Figure 6: Illustration showing the identicality between the three regressions

That the MSE and MAE of our three regression are this identical, is the reason why we cannot confidently pick a best-performing model. Since our dataset has numerous features, it then stand to reason to pick the Ridge model in order to mitigate overfitting. As mentioned, the fact that the differences between the OLS and the Ridge is barely existent benefits our decision to proceed with this regression. Again, the error metrics do not favour any model, but due to the reasons listed above, we will further analyse the Ridge regression.

5.1.1 Learning Curve

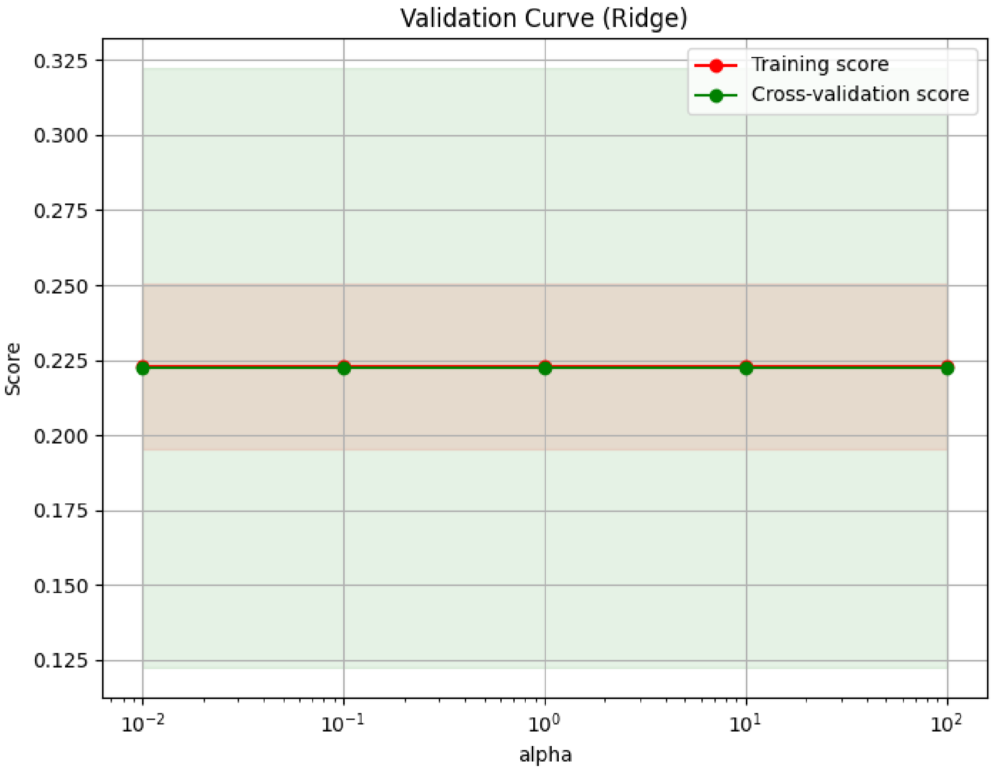


Our y-axis represents the coefficient of determination, R^2, that is the statistical measure that represents proportion of the variance for a dependent variable that is predictable from the independent variables. [10] If we focus on the convergence point at 0.23, then low convergence score could indicate

[10] *R-Squared: Definition, Calculation Formula, Uses, and Limitations* by Investopedia

high bias, that is, an underfitting problem. And by adding more training examples it will unlikely improve the model's performance. On the other hand, if the so-called convergence point is just a passing point, for which the training score will keep increasing, while the CV-score will become constant, then there is reason to believe that we are dealing with a overfitting and thereby high variance.

5.1.2 Validation Curve



Given that both curves are constant along the same score-value indicates that they are not effective by the hyperparameter, by which it means that that adjusting the Ridge hyperparameter wouldn't affect the model's performance. It has again converged in a low-score point, which indicates the model being underfitting. Due to our model being so biased, we have failed to train a model that could properly predict present housing prices and let alone future housing prices.

5.2 Socioeconomic Factors & Property Valuations: Present Analysis

```
>>> target_correlation
PropertyValuationAmount      1.000000
rolling_std_2_years          0.385005
MunicipalityAverageIncome    0.148718
MunicipalityPersonsCount     0.120596
numberToilets                0.106681
propertyHouseSize            0.106354
propertySizeTotal            0.084723
sizeGarage                   0.017774
MunicipalityUnemployment     0.002596
PropertyTaxRate             -0.036610
MunicipalitySize            -0.120485
Name: PropertyValuationAmount, dtype: float64
```

```
>>> print(lasso_coefficients)
numberToilets                102381.809610
MunicipalityUnemployment      61031.800531
sizeGarage                     5388.863251
propertyHouseSize              4747.047460
PropertyTaxRate                 945.741066
propertySizeTotal                18.176664
MunicipalityAverageIncome        13.420942
MunicipalityPersonsCount          3.376753
rolling_std_2_years               0.848163
MunicipalitySize               -675.632692
dtype: float64
```

```
>>> print(ridge_coefficients)
numberToilets                102554.199290
MunicipalityUnemployment      61087.860725
sizeGarage                     5389.581214
propertyHouseSize              4745.533072
PropertyTaxRate                 947.296189
propertySizeTotal                18.176881
MunicipalityAverageIncome        13.421894
MunicipalityPersonsCount          3.376486
rolling_std_2_years               0.848162
MunicipalitySize               -675.575520
```

```
>>> print(metrics_df)
   Model            MAE           MSE
0    OLS  558082.947832  7.135689e+12
1  Lasso  558087.330556  7.135690e+12
2  Ridge  558083.204459  7.135689e+12
```

First, let's look at the correlation between the variables in the output above, that is, whether there is a positive or negative relationship. We see a positive correlation between our variable PropertyValuationAmountand rolling_std_2_years (0.385). It indicates the direction and strength of the linear relationship between two variables, showing that the increase of one tends to increase the other. In addition, other factors such as Municipality Average Income and Number of Persons in the Municipality also show positive correlations. Moreover, we used two different regression methods used for regression analysis, Lasso and Ridge. These methods are property valuations of a dependent variable called PropertyValuationAmount in our output and independent variables are socioeconomic factors, we will try to estimate the relationship between them. We found the RMSE value of MSE for Lasso, Lasso regression model to be 2,671,271.2553649493. This shows how far the Lasso model's predictions are usually from the actual data. In short, the predictions of the model have deviated by this much from the real values.

The other model, RMSE for Ridge, shows the Ridge regression model. We found the RMSE value to be 2,671,271.106871828, this model also shows us how far the model's predictions are usually from the real data.
We examined the coefficients of the Lasso and Ridge regression models. In particular, factors such as numberToilets, MunicipalityUnemploymentand sizeGarage seem to have significant effects on property valuations because they have larger coefficients.
Moreover, the Mean Absolute Error (MAE) and Mean Square Error (MSE) we found are about 558.083 to 558.087. These metrics show how close the models' predictions are to the actual data.
As a result, We can deduce that there seems to be a complex relationship between property valuations and socioeconomic factors. Some factors may affect property values more, while others may affect them less. The analysis results can be used to better understand these factors and predict future property valuations.

6 Discussion

The RMSE values for both models are similar, meaning the accuracy of their predictions is similar. A lower RMSE value represents better forecast accuracy. In the values we found, there is no significant difference between the Lasso and Ridge models, making it unimportant which model we use to estimate property valuations.

MunicipalityAverageIncomeand MunicipalityUnemploymenthave been found to have an impact on property valuations. This means that individuals and local governments can make property investments by taking these factors into account. The results of the research can guide local governments' infrastructure investments, zoning policies and economic development strategies. This can open new doors for more balanced and sustainable urban development.

Change in model

Given that our 3 regressions basically were the same, and the model we further analysed was biased/underfitted, we have reason to reuse our machine learning code with a polynomial regression, to test how such a regression would work with our dataset. This is due to polynomials being nonlinear, which can add complexity to our model, for which we could potentially fix our underfitting problem.

7 Conclusion

In the series analysis when we visualize, policy makers can predict the future from trend changes in the market, thanks to the distribution of property prices. This will guide their policies and taxation. Another building year of the property we visualized. shows us the age of the property's structure. Newer buildings are generally more modern, require less maintenance, and receive more attention. Of course, the expansion year is also important. Knowing when your property has been extended will inform potential buyers or tenants what innovations have been made and how functional the property is.

By gathering a big dataset through various methods, one being webscraping, we applied machine learning to build a model that could predict the price valuation of any house in the country both in the present and in the future. By searching the internet for articles and based on our own dataset, we respectively found and made graphs that showed the relationship between time and property prices. In both scenarios we had a pretty linear increase in property prices. So by such assessments we decided to train a linear regression, and a Lasso and Ridge regression. By using the K-Fold Cross Validation we estimated hyperparameters that optimised the MSE for both regularisation models. Once we calculated the errors we came to realize that all 3 regression models basically gave the same errors, for which we understood that were wasn't a definitive answer to pick one model over the next. We went on to examine the the training- and validation curve, for which we concluded that our model suffers from underfitting, and thereby had to acknowledge it wasn't fit to predict housing prices.

8 References

References

[1] K. E. Case and J. M. Quigley, *How housing booms unwind: Income effects, wealth effects, and feedbacks through financial markets*, European Journal of Housing Policy, 8(2), 161–180, 2008.

[2] Danmarks Statistik, *Annual Report on Danish Real Estate Market*, 2023.

[3] E. Kallergis, C. Kavvathas, and K. Kounetas, *The Impact of the Financial Crisis on the Property Valuation Market in Greece*, 2019.

[4] Pierluigi Morano, Francesco Tajani, and Marco Locurcio, *Land Use, Economic Welfare and Property Values: An Analysis of the Interdependencies of the Real-Estate Market with Zonal and Socio-Economic Variables in the Municipalities of Apulia Region (Italy)*, International Journal of Agricultural and Environmental Information Systems, 2015.

[5] Carl Folke, *Resilience: The emergence of a perspective for social–ecological systems analyses*, Global Environmental Change, 2006.

[6] Mohd Faris Dziauddin, Kamarul Ismail, and Zainudin Othman, *Analysing the Local Geography of the Relationship Between Residential Property Prices and Its Determinants*, Bulletin of Geography Socio-economic series, 2015.