**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | Programming for DA  Statistics for Data Analytics  Machine Learning for Data Analysis  Data Preparation & Visualisation |
| **Assessment Title:** | MSc\_data\_CA2 |
| **Lecturer Name:** | Sam Weiss  John O’Sullivan  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Fatma Akus |
| **Student Number:** | 2023009 |
| **Assessment Due Date:** | 26/05/2023 |
| **Date of Submission:** | 26/05/2023 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

**TABLE OF CONTENTS**

TABLE OF CONTENTS2

Introduction 3

Programming for Data Analysis 5

Data Preparation and Visualisation5

**Analytical Framework** 5

**Problem Identification** 6

**Data Wrangling**

Statistics for Data Analytics10

Descriptive Statistics10-11

Statistical Tests12-13

Correlation…………………………………………………………………………………………………………………………………14

Machine Learning Data Analytics15

One Hot Encoding16

Standard Scaler ………………………………………………………………………………………………………………………………..17

Standard Scaler ………………………………………………………………………………………………………………………………..17

Lineaar Regression 16-18

Gradient Boosting ………………………………………………………………………………………………………………………19-20

Conclusion21

Dashboard Overview22

References23

**INTRODUCTION**

This study is about new house prices in Ireland between Netherlands. New house prices in different states of Ireland and Netherlands are included in the dataset. In addition, the house prices in Europe were visualized with the map, and the houses with the highest house prices were compared by appearing darker in coloration. In this study, we will compare these states using data analysis techniques and make a price estimation. In this study, Python language and Jupyter Notebook IDE are used. Also, the dataset type is CSV. Before starting the data analysis, the dataset was uploaded to Jupyter Notebook and then analyzed. First, the necessary libraries were imported because it is necessary to make use of these libraries while analyzing, otherwise the analysis will not be performed. Then, general information about the data was learned using the Python programming language. In general, the information learned is the first 5 data of the data set, its type, information, the number of null values, functional data analysis, data cleaning and visualization. After this phase was completed, statistical analysis of the dataset was performed. These are mean, median, mode, quartiles, and describe statistical analysis. In addition, categorical and numerical variables were analyzed and visualized. The correlation matrix, which is the final stage of the statistical analysis, is applied and visualized. After this stage, I prepared the dataset to apply Machine Learning, so I had to make the categorical variables numeric, but all the variables were numeric and there was no need. But I used Standard Scaler because it is easier to analyze, all the data in the dataset have the same standard. Next, there are some Machine learning models that I use in my dataset. These are Linear Regression and Gradient Boosting. The purpose of choosing these algorithms is that my dataset has numerical variables, and these methods should be used.

**PROGRAMMING FOR DATA ANALYSIS**

In this study, Jupyter Notebook is used, and the dataset is in CSV type. First of all, I imported the libraries used in basic data analysis. Numpy and Pandas libraries are used for analysis Seaborn, Matplotlib is used for visualization. After importing the libraries, I loaded my dataset into my code file. While loading the CSV dataset, we must use the “read\_csv” method, otherwise the dataset script file will give an error. Then I output the data set with the "irl.head() ,hl.head() and eu.head()" method of the first 5 data. To have information about the dataset, I got more information about the data by using the "df.info" method. Even though I know that my dataset is not null, I still wanted to be sure, so I used the "irl.isnull().sum(),hl.isnull().sum()" method to print the null values in all columns in the dataset. There is no null value as a result, which is a good thing because we could clean up the null data and fill it with the mean, which would make analysis more difficult. After completing these stages, I applied exploratory data analysis, but it would take long processes to do this analysis with normal writing, but I did this analysis by writing a function, so I learned all the information about the data. I have successfully completed the first stage of the data analysis phase.

Data Preparation and Visualisation

**Exploratory Data Analytics (EDA)**

**Check Information Dataset**

To get general information about the dataset, I checked the data for Ireland, Netherlands and Europe using the irl.info (), hl.info () and eu.info () methods. The results are as follows:

A screenshot of a computer

Description automatically generated with low confidence

Figure 1: Check Information Ireland

A screenshot of a computer

Description automatically generated with low confidence

Figure 2: Check Information Netherlands

A screenshot of a computer

Description automatically generated

Figure 3: Check Information Europe

Especially in the Netherlands' dataset, some of the object type data is actually float, but when saving to csv, it seems to be object type because they save the dataset as object. I will convert this object type to float in the next steps.

House Price in Europe:

When we evaluate the new house prices according to Europe, the countries in Europe with the highest house prices are cities such as Turkey, Germany and Italy. In short, house prices are higher in places where red colors are dominant. New house prices vary according to each country. However, these high countries are generally based on the materials used in the construction sector. This comparison is between 2019 and 2023. If I had included the rates in the previous years, the results could have been different.

A map of the world

Description automatically generated with medium confidence

Figure 4: House Price in Europe

**Check Null Values**

It is necessary to detect null values in the dataset and eliminate these empty data. Therefore, I checked the null data with the "isnull (). sum ()" method and saw which ones were empty.

As a result, there are no null values in the Irish dataset, which is a good situation. However, the number of nulls in the Netherlands is high, but some of the nulls are in unnecessary rows for us, so we can make the comparison more accurate by deleting all the unnecessary rows, so I will delete the empty rows from the dataset in the next steps.

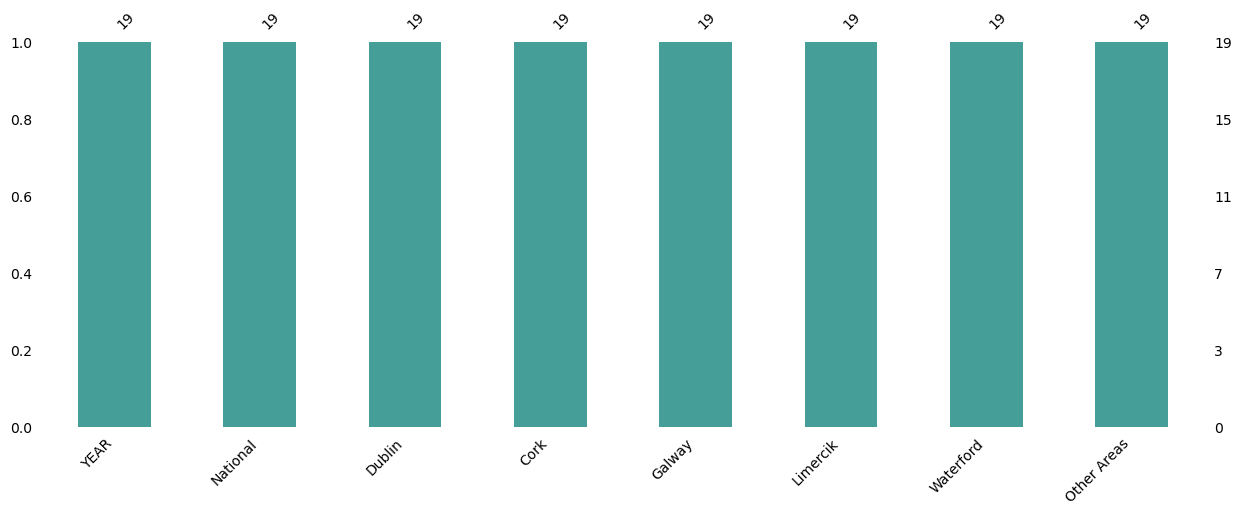


Figure 5: Check Null Values Ireland

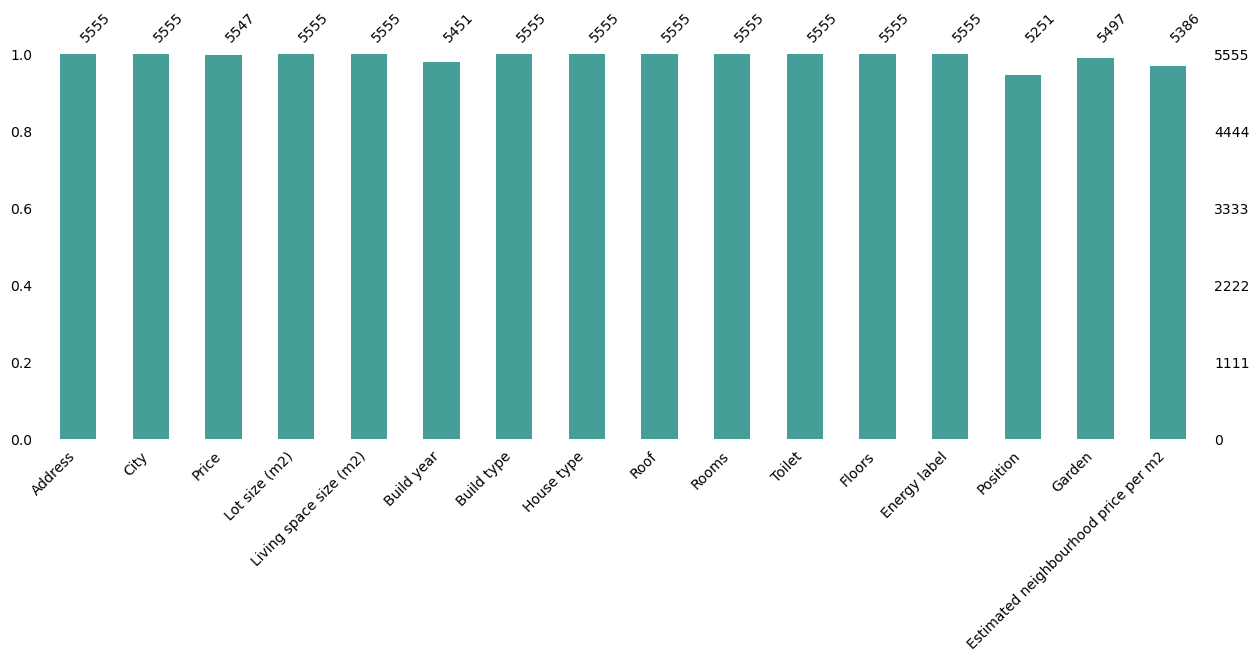


Figure 6: Check Null Values Netherlands

**Categorical and Numerical Values**

Functions suitable for the categorical and numerical data of Ireland and the Netherlands were written and visualized. The visualization results are shown below, you can examine them in detail. While there are no categorical variables in the Irish data, there are in the Dutch data, the exact number of which was determined by the conversion operations in the previous section and which variables were categorically finalized.

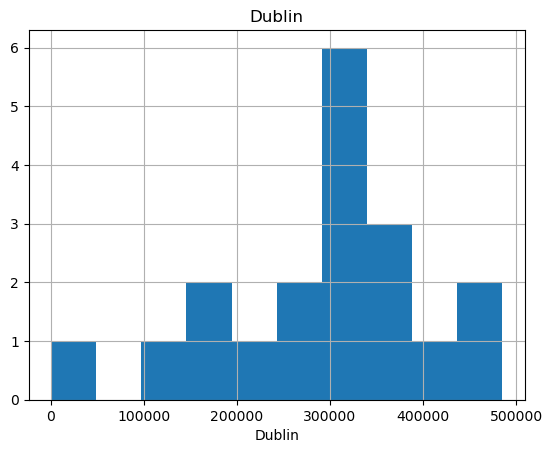
Ireland Categorical Values:   
  
[ ]

Netherlands Categorical Values:

['Address','City','Build type', 'House type','Roof', 'Toilet', 'Energy label', 'Position', 'Garden']

**Numerical Values Ireland**

A picture containing text, screenshot, line, square

Description automatically generated

A picture containing text, screenshot, line, diagram

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generated

A picture containing text, screenshot, line, diagram

Description automatically generatedA picture containing text, screenshot, diagram, square

Description automatically generated

Figure 6: Numerical Values Ireland

**Netherlands Numerical Values**

A picture containing text, screenshot, line, diagram

Description automatically generated A picture containing text, diagram, plot, screenshot

Description automatically generated

A picture containing text, screenshot, diagram, line

Description automatically generatedA picture containing text, diagram, screenshot, line

Description automatically generated

A picture containing text, screenshot, line, diagram

Description automatically generated

Figure 7: Numerical Values Netherlands

**Compare City Ireland**

I wrote a function to compare house prices within cities. I analyzed and visualized only Irish cities. Since the Netherlands data is complex and the number of cities is high, it has been a lot to analyze and compare the cities created. Of all the cities in Ireland, the city with the highest prices is Dublin.

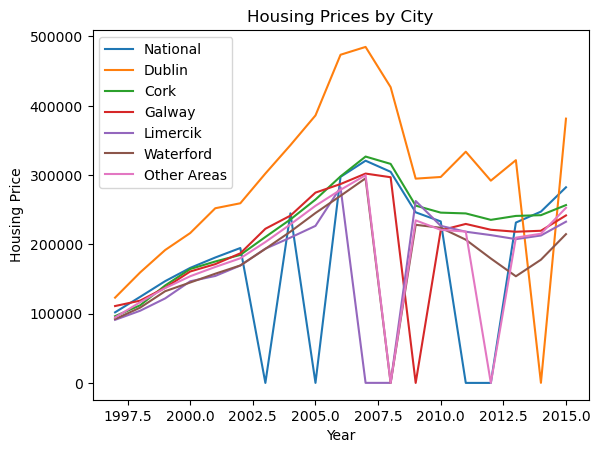


Figure 8: Housing Price by City Ireland

As I approached the end of the data preparation phase, I deleted the unnecessary data from the data set. There were many unnecessary columns, especially in the Dutch dataset, and I deleted them.

**Netherlands first 5 data**

A picture containing text, screenshot, font, number

Description automatically generated

Figure 9: Ireland Head

**Ireland first 5 data**

A picture containing text, screenshot, font, number

Description automatically generated

Figure 10: Netherlands Head

**STATISTICS FOR DATA ANALYTICS**

First, I printed an overall statistical analysis of the data for Ireland and the Netherlands using the "describe ()" method. As a result of the analysis, I learned count, mean, standard deviation, min,25%, 50%, 75%, max values for each column.

Values in statistical data analysis are as follows. Statistical analyzes were calculated separately for each city in the Irish data set and the results were shared. In the Netherlands, there is only the numeric variable Price with the deleted columns and the calculation was made accordingly.

**Ireland Statisticial Describe**

A picture containing text, screenshot, font, number

Description automatically generated

Figure 11: Ireland Describe Statistics

**Netherlands Statisticial Describe**

A screenshot of a cell phone

Description automatically generated with low confidence

Figure 12: Netherlands Describe Statistics

The mean, standard deviation and median values of the Irish and Netherlands data were calculated among themselves, and the details are shared in the table below.

According to the data in Ireland, the city with the highest average price is "Dublin". Compared to the Netherlands, this value is lower. The standard deviation values are highest in Other Areas in Ireland. When we compare the standard deviation values, the lowest standard deviation value is the Netherlands. The median value is higher for the Netherlands.

|  |  |  |  |
| --- | --- | --- | --- |
| City | Mean | Standard Deviation | Median |
| National | 175009.892 | 109988.884 | 194835 |
| Dublin | 291636.873 | 120208.583 | 297294 |
| Cork | 223630.263 | 63859.395 | 241127 |
| Galway | 203346.388 | 74547.758 | 220351 |
| Limercik | 172451.808 | 78989.016 | 207635 |
| Waterford | 179745.103 | 68055.605 | 179716 |
| Other Areas | 182606.067 | 83292.785 | 209772 |

|  |  |  |
| --- | --- | --- |
| Mean | Standard Deviation | Median |
| 558578.045 | 354104.369 | 469000 |

Figure 14: Netherlands Mean-Standard Deviation- Median

Figure 13: Ireland Mean-Standard Deviation-Median

**T-Test**

When T-test values are compared between both countries, the Netherlands has a higher T-test value and the Netherlands has a higher value when compared as a P-value. Ireland's values have low values, and having negative values is pretty bad.

|  |  |  |
| --- | --- | --- |
| Countries | T- statistics | P- value |
| Ireland | * 3.1200584198729735 | 0.003551877535326691 |
| Netherlands | 1.5298431552512572 | 0.17692772929597267 |

Figure 15: T-test

**Wilcoxon Test**

I wrote functions to compare Wilcoxon tests, but the value of Holland cannot be calculated, because it probably does not have much value. To explain Ireland's value, 0.5 p value is a value that can be considered successful, but the value here is 0.005 and not very successful.

|  |  |  |
| --- | --- | --- |
| Countries | Wilcoxon statistic | P- value |
| Ireland | 15.0 | 0.000522613525390625 |
| Netherlands | nan | nan |

Figure 16: Wilcoxon Test

**Chi-square Test**

When comparing the Chi-Square values, there is a huge difference between the values of both countries. The reason for this difference may be due to the size of the dataset and the size of the values in the dataset. The Netherlands has higher values than Ireland, but the P-value Netherlands has a value of 0. Ireland's value is also greater than 0.5 overfitting, which is not a very good situation for the data to show overfitting.

|  |  |  |  |
| --- | --- | --- | --- |
| Countries | Chi-square statistic | P-value | Expected frequencies |
| Ireland | 113.30459015923151 | 1.8504838210093205e-26 | [[100161.82739781 126243.17260219]  [125106.17260219 157682.82739781]] |
| Netherlands | 34957.949575798346 | 0.0 | [[585598.20678061 364401.79321939]  [514401.79321939 320098.20678061]] |

Figure 17: Chi-square Test

**Variance Analysis**

Ireland

Comparing the analysis of variance among Irish cities, Dublin has the highest value followed by the highest.

**F-Statistic:** 18.769944950919687

**p-value:** 1.2205277159936643e-17

|  |  |
| --- | --- |
| City | Variances |
| National | 12097554531.890 |
| Dublin | 14450103471.967 |
| Cork | 4078022354.871 |
| Galway | 5557368266.617 |
| Limercik | 6239264681.749 |
| Waterford | 4631565432.100 |
| Other Areas | 6937688046.578 |

**A picture containing text, screenshot, diagram, plot

Description automatically generated**

Figure 18: Variance Analysis

Netherlands

The Netherlands' overall analysis of variance is high, but I had a hard time calculating the f-statistic and p-value. Therefore, it is very difficult to predict.

Variances:

{'Price': 31950.66861135549}

F-Statistic: None

p-value: None

**Correlation**

The matrix is symmetric, which means that the correlation between variable A and variable B is the same as the correlation between variable B and variable A. The diagonal of the matrix shows the correlation of each variable with itself, which is always 1 since a variable is perfectly correlated with itself. The off-diagonal cells show the correlation between pairs of variables. . A positive correlation coefficient indicates that the variables are positively related, meaning that when one variable increases, the other variable tends to increase as well. A negative correlation coefficient indicates that the variables are negatively related, meaning that when one variable increases, the other variable tends to decrease. A correlation coefficient of zero indicates that there is no linear relationship between the variables. Overall, this correlation matrix can be useful in understanding how the different variables are related to each other. It can also be used to identify which variables are highly correlated, which can be important in certain types of data analysis, such as regression analysis.

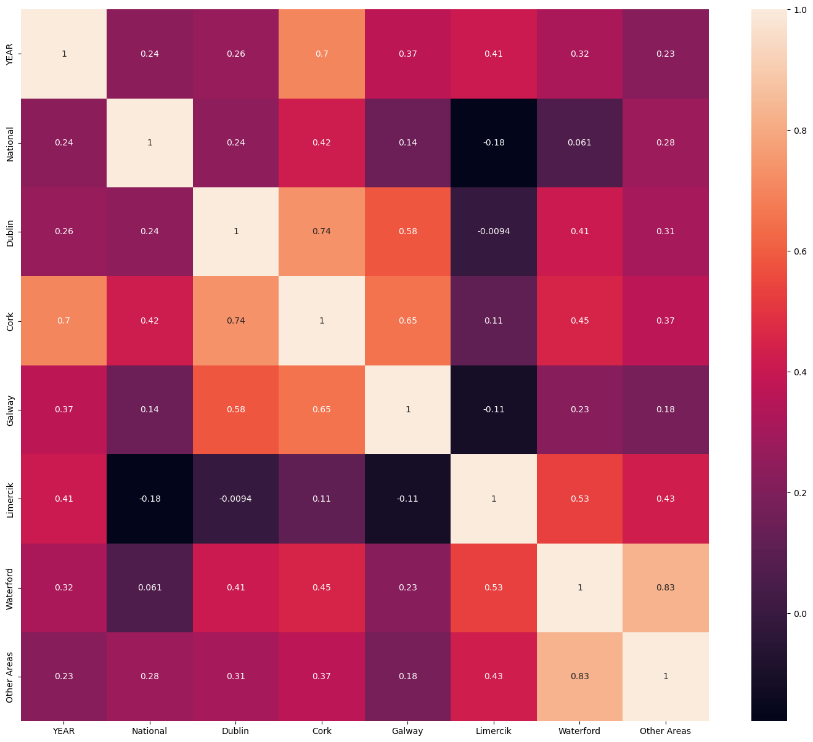
Ireland  


Figure 19: Correlation Analysis Ireland

Netherlands



Figure 20: Correlation Analysis Netherlands

**Machine Learning for Data Analytics**

After statistical analysis in detail, the next stage is the Machine Learning stage. At this stage, we model the dataset using various algorithms, so we make predictions about the data and develop strategies. Machine Learning algorithms are too many and if algorithms suitable for this study are not preferred, we cannot get successful results and make accurate predictions. This dataset consists of numerical variables, so Machine Learning algorithm suitable for numerical data should be used. The target variable of this data is "Price" and “Year” so "Supervised Learning" Machine Learning algorithms are used. I chose the 3 ML algorithm that is most suitable for this dataset and applied it in my work. The relationships between the results of these analyzes are explained by comparing the data of the Netherlands and Ireland.

These ML algorithms are:

* Linear Regression
* Gradient Boosting
* Support Vector Regression

I analyzed and modeled it in detail with these machine learning algorithms. Before applying these models, I applied the Standardscaler because I wanted to ensure equality in the dataset and the analysis.

**One Hot Encoding**

One Hot Encoding digitizes categorical variables, text type data and Machine Learning applications are started to be applied. Categorical variables are not present in the Irish data set, but there are categorical variables in the Netherlands data set, so I converted those values into numerical variables.

**Standard Scaler**

Before Modeling the Dataset, I standardized it so that it was easier to analyze and model. According to this modeling result, the following output.

**A picture containing text, screenshot, font, number

Description automatically generated**

Figure 21 : Standard Scaler

**Supervised Machine Learning**

Supervised machine learning is a type of machine learning where the model is trained on a labeled dataset. This means that the input data (or features) and the corresponding output data (or labels) are already known, and the goal is to learn a mapping between them so that the model can make predictions on new, unseen data. Supervised machine learning has many practical applications, Linear Regression, Gradient Boosting and Support Vector Regression were used in this study. It has been predicted using these algorithms.

**Linear Regression**

Linear regression is a statistical method used to model the relationship between two variables. The idea is to find a linear equation that best describes the relationship between the two variables. In this dataset, Linear Regression has been applied and visualized for all states.

R2 score Train and test RMSE CV Linear Regression RMSE: it is the mean square root error of a regression model that calculates the average distance between the train and test values so The lower the RMSE, the better a given model is able to “fit” a dataset.

R2 Score: The R2 score is a very important metric that is used to evaluate the performance of a regression-based machine learning model. In other words, it is the difference between the samples in the dataset and the predictions made by the model.

CV: model cross-validation is performed. The model is divided into 2 as train and test, while estimating data with train, testing with test data and procreation is performed.

Ireland

National

Size of X\_train: (15, 7)

Size of y\_train: (15, 1)

Size of X\_test: (4, 7)

Size of Y\_test: (4, 1)

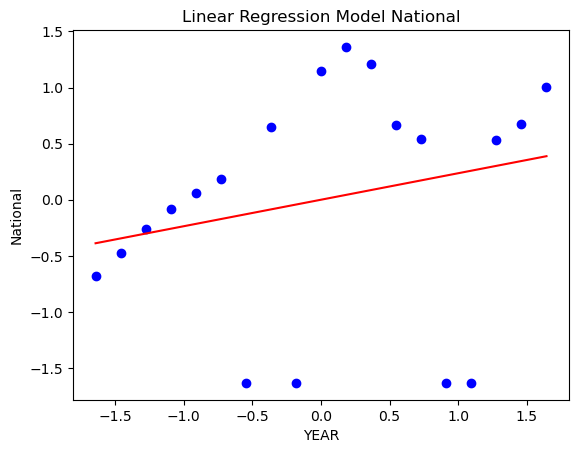


Figure 22: Linear Regression Ireland

CV Linear Regression: -3.445

R2\_score (train): 0.393

R2\_score (test): 0.462

RMSE: 0.541

I evaluated it according to the RMSE score according to the Linear Regression model in Ireland and it has a score of 0.54 according to this evaluation, this score shows that the model's success is insufficient and the model is not learning.

Netherlands

City\_Pannerden

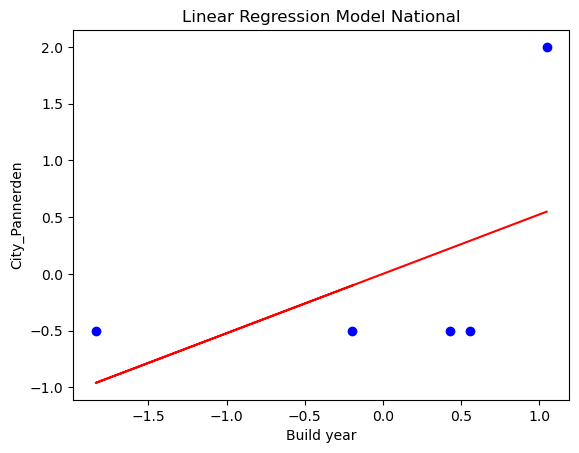


Figure 23: Linear Regression Netherlands

CV Linear Regression: nan

R2\_score (train):-0.386

R2\_score (test): nan

RMSE: 1.155

When I calculate the Linear Regression model of the Netherlands, it shows that the model underfits as seen in the graph, that is, it cannot ensure the success of the model and does not calculate the model correctly. More data must be processed in order for the model to learn. When I evaluate the success of the model, the RMSE score shows that it could not learn either. In order to ensure the success of the model, the data must be increased.

**Gradient Boosting**

Ireland

MSE: 0.7487852278052423

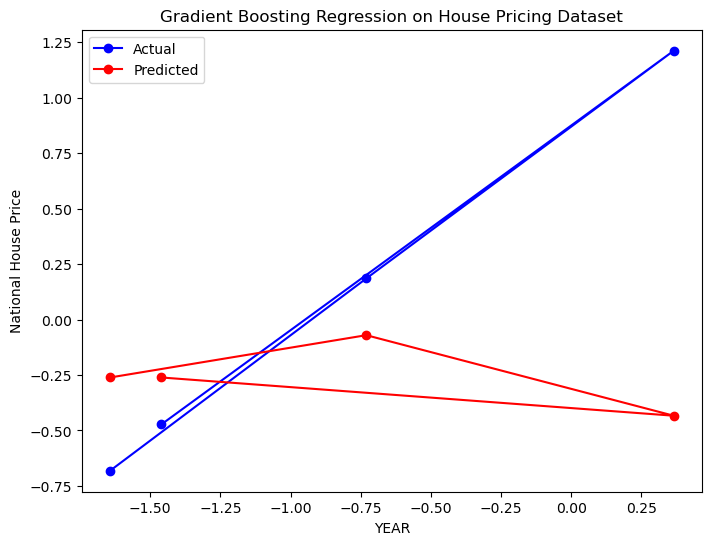


Figure 24: Gradient Boosting Regression Ireland

According to Gradient Boosting, when I analyze the data in Ireland, the MSE success is 0.74. This success value is low, but the size of the dataset we have is low, so this success rate may be successful.

Netherlands

MSE: 7.321926593100242’e-17

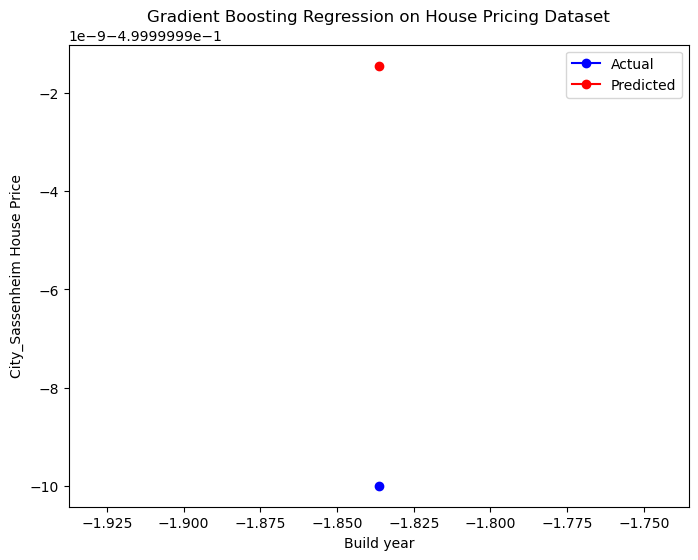


Figure 25: Gradient Boosting Regression Netherlands

According to the dataset of the Netherlands, it could not learn the data according to the Gradient Boosting method and as seen in the graph, a correct graph did not emerge. Also, MSE value, the value of 7.32 does not have a successful score. More data is required to evaluate this score.

**Support Vector Regression**

Ireland

MSE: 1.650823657029558

A picture containing text, screenshot, line, plot

Description automatically generated

Figure 26: Support Vector Regression Ireland

I calculated the Support Vector Regression model by writing a function on the Irish dataset. It did not calculate the model correctly enough in this dataset and outputs MSE:1.65. This shows that the score is not enough for success.

Netherlands

MSE: 0.6438474487839831

A picture containing text, screenshot, display, number

Description automatically generated

Figure 27: Support Vector Regression Netherlands

According to the Dutch data, according to this regression model, the Dutch dataset underfitted this model and the MSE value of the model is 0.64, which shows that this value is not sufficient for success.

**Conclusion**

As a result, many models were used during data analysis and compared among themselves. Gradient Boosting was the ML technique that gave the best estimation as a result of the analyzes made for the Irish and Dutch data sets. In the future, it can be used in different methods by adding data to the dataset. In addition to these methods, the Linear Regression model is also at a level that can be considered successful. While performing data analysis, all these stages are created to prepare for the Machine Learning model. After we have general information about the data, we clean it, visualize it, and make it ready for the model, respectively. In this study, based on similar data sets, the most suitable methods were decided for the model and it was aimed to make predictions by using the most accurate technique.

**Dashboard**

A screen shot of a graph

Description automatically generated with medium confidence A screenshot of a computer screen

Description automatically generated with medium confidence

Figure 28: Dashboard

**Resources:**

[1] <https://www.analyticsvidhya.com/blog/2021/07/basic-statistics-concepts-for-machine-learning-newbies/>

[2] <https://www.w3schools.com/ai/ai_statistics.asp>

[3] <https://machinelearningmastery.com/statistics_for_machine_learning/>

[4] <https://github.com/unpingco/Python-for-Probability-Statistics-and-Machine-Learning-2E>

[5] <https://data-flair.training/blogs/machine-learning-algorithms-in-python/>

[6] <https://medium.com/coders-camp/40-machine-learning-algorithms-with-python-3defd764b961>

[7] <https://wesmckinney.com/book/>

[8] <https://github.com/wesm/pydata-book>

[9] <https://towardsdatascience.com/the-easiest-way-to-create-an-interactive-dashboard-in-python-77440f2511d1>

[10] <https://realpython.com/tutorials/machine-learning/>

[11] <https://www.tutorialspoint.com/machine_learning_with_python/index.htm>

[12] <https://link.springer.com/book/10.1007/978-1-4842-3207-1>

[13] <https://cognitiveclass.ai/courses/machine-learning-with-python>

[14] <https://www.springboard.com/resources/learning-paths/machine-learning-python/>

[15] <https://ocw.mit.edu/courses/6-867-machine-learning-fall-2006/>

[16] <https://data.gov.ie/dataset/average-new-house-price?package_type=dataset>

[17] <https://data.gov.ie/dataset/bcms-commencement-notices-2017?package_type=dataset>

[18] <https://www.kaggle.com/datasets/bryan2k19/dutch-house-prices-dataset>