**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| Module Title(s): | Programming for DA  Statistics for Data Analytics  Machine Learning for Data Analysis  Data Preparation & Visualisation |
| Assessment Title: | MSC\_DA\_CA2 |
| Lecturer(s) Names: | Sam Weiss  John O’Sullivan  Muhammad Iqbal  David McQuaid |
| Student Full Name: | Clelia Caetano |
| Student Number: | 2023060 |
| 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 another source. I declare it to be my work and that all material from third parties has been appropriately referenced. I further confirm that this work has yet to be submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Author: Clelia Caetano

E-mail: [2023060@student.cct.ie](mailto:2023060@student.cct.ie)

Student ID:2023060

Constructions:

Workplace Safety

CCT College Dublin

April (2023).

# Abstract

*This report results from a critical analysis of data written and programmed to find the Area which retained the highest increase in average house values ​​in Ireland between 1976 and 2016. After that, a comparison of the average increase was made on the price of new and used homes in the Area identified as having the most expensive average in the first evaluation. Finally, to complete the investigation, a prediction was made to find out the estimate of the price of houses in that Area would cost by the year 2026.*

*For this, the concepts of Statistics, Programming, Data Visualization and Machine Learning acquired during classes within the Python program were applied. Therefore, the choice of models and codes has been explained and justified as part of this report.*

*In addition, more comprehensive research was needed to deepen the concepts and improve the quality of the programming codes utilised in the Jupiter Notebook.*

***Keywords:*** *Statistics, Programming, Data Visualization, Machine Learning, Python.*

Table of Contents

[Abstract 3](#_Toc132581044)

[List of Tables 4](#_Toc132581045)

[Introduction 6](#_Toc132581046)

[Chapter I - Materials and Methods 7](#_Toc132581047)

[1.1 - Statistics for Data Analytics 7](#_Toc132581048)

[1.1.1 – Dataset Overview 7](#_Toc132581049)

[1.1.2 – Descriptive Statistics 8](#_Toc132581050)

[1.1.3 – Distribution 14](#_Toc132581051)

[1.1.4 – Probability 18](#_Toc132581052)

[1.2 - Data Preparation & Visualisation 20](#_Toc132581053)

[1.3 - Programming for Data Analytics 28](#_Toc132581054)

[1.3.1 - Justifications and Explanations of the chosen codes 28](#_Toc132581055)

[1.3.2 – Java versus Python: Overview - Another Programming Language 31](#_Toc132581056)

[1.4 - Machine Learning for Data Analysis 32](#_Toc132581057)

[1.4.1 - Project Management Framework 32](#_Toc132581058)

[1.4.2 - Machine Learning Models 33](#_Toc132581059)

[1.4.3 - Machine Learning Modelling Comparisons: Similarities and differences 34](#_Toc132581060)

[Chapter II - Results and Discussion 36](#_Toc132581061)

[Conclusions 38](#_Toc132581062)

[References 39](#_Toc132581063)

# List of Tables

[Figure 1: Displaying the size of the original dataset 7](#_Toc132570796)

[Figure 2: Part II: Sample size 8](#_Toc132570797)

[Figure 3: Finding the Outliers 9](#_Toc132570798)

[Figure 4: Applying Winzorization for Outliers 9](#_Toc132570799)

[Figure 5: Determine the main statistical measures 10](#_Toc132570800)

[Figure 6: Distribution of values by Area. 11](#_Toc132570801)

[Figure 7: Average price of the areas by types: New Houses and Second-hand. 12](#_Toc132570802)

[Figure 8: Average Price of Houses by Area. 13](#_Toc132570803)

[Figure 9: Filter out only the most expensive Area. 13](#_Toc132570804)

[Figure 10: Distribution of Price variable 15](#_Toc132570805)

[Figure 11: Applying the logarithm function. 16](#_Toc132570806)

[Figure 12: Shapiro-Wilk test 16](#_Toc132570807)

[Figure 13: Sample estimates of Mean & Standard deviation. 17](#_Toc132570808)

[Figure 14: Probability density function (PDF). 18](#_Toc132570809)

[Figure 15: Probability calculation. 19](#_Toc132570810)

[Figure 16: Analysing DataSet. 20](#_Toc132570811)

[Figure 17: EDA: Cleaning Dataset. 21](#_Toc132570812)

[Figure 18: EDA: Checking missing values. 21](#_Toc132570813)

[Figure 19: EDA: Cleaning missing values. 22](#_Toc132570814)

[Figure 20: Recognising a valid label. 23](#_Toc132570815)

[Figure 21: Sort Area by descending order. 23](#_Toc132570816)

[Figure 22: Correlation Heatmap 24](#_Toc132570817)

[Figure 23: Pairplot: Correlation Matrix. 25](#_Toc132570818)

[Figure 24: Lineplot: Evolution Value vs Year by Area. 26](#_Toc132570819)

[Figure 25: Lineplot: Average Price of houses in Dublin by Statistic label. 27](#_Toc132570820)

[Figure 26: Importing libraries. 28](#_Toc132570821)

[Figure 27: Basics Codes used for (EDA). 30](#_Toc132570822)

[Figure 28: Modelling Comparisons 34](#_Toc132570823)

[Figure 29: Actual Values versus Prediction Values. 34](#_Toc132570824)

[Figure 30: Predicted Values for 'Statistics labels' for the year 2026. 36](#_Toc132570825)

[Figure 31: Prediction of average of houses in Dublin for the year 2026. 37](#_Toc132570826)

# Introduction

Buying your own home can be a dream for many people. However, if these people have lived in Ireland for the past few years, that dream may fade. It is because it is common to watch the news or read in the newspapers about the real estate crisis that the country has been facing over the years. There are some hypotheses and speculations that only some properties are available on the market, making demand more significant and, consequently, the value becomes increasingly expensive. This event is popularly known as the law of supply and demand in the economy. Nevertheless, that is not the objective of this report.

This report will present the evolution of house values in the main areas of Ireland (National, Dublin, Cork, Galway, Limerick, Waterford, Other areas) between 1976 and 2016, taking which Area is likely the most expensive average between them; also comparing the price ratio between new and second-hand houses. After all these analyses, a value forecast for subsequent years also will be presented.

# Chapter I - Materials and Methods

## 1.1 - Statistics for Data Analytics

## 1.1.1 – Dataset Overview

The HSA06 dataset, which served as the basis for exploring this study, was chosen from the official website of the Irish Government. It consists of average house prices nationally and in the main areas of Ireland over the years 1975 and 2016. (Department of Housing, Local Government, and Heritage, n.d.)

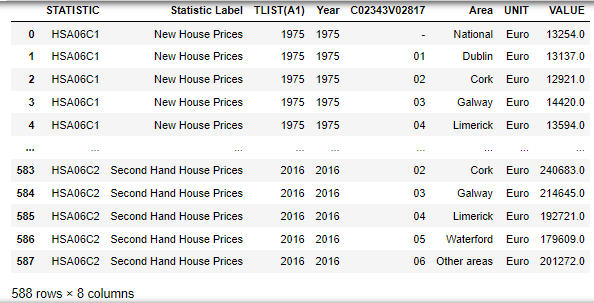
The dataset initially had a population size of 588 numerical and categorical data types spread over eight columns. It covers seven areas of Ireland (National, Dublin, Cork, Galway, Limerick, Waterford and Other areas). However, only a sample of 574 was used for the primary analyses, where 1975 was no longer considered essential for the study due to the missing values.

Figure 1: Displaying the size of the original dataset

For the second part of this assay, the sample was reduced to 82, where only the Dublin area was considered by the proposed objective.

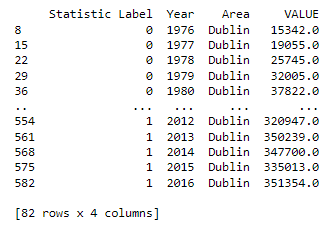


Figure 2: Part II: Sample size

## 1.1.2 – Descriptive Statistics

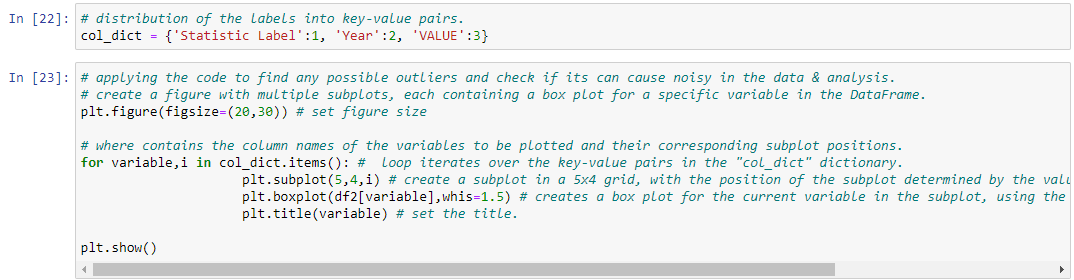
According to Grolemund, G., & Wickham, H., statistical measures are the fundamental part of any data analysis because, through the summary and descriptions, one can show the patterns and the relationship between the variables in the data. In addition, it can determine trends and possible outliers that hide behind the information and lead to errors in tests and results. (Grolemund, 2017).

Besides this introduction of statistical measures, the conduction of this assessment, the application of codes and the results will be shown.

After demonstrating this dataset's population and sample size, described at the beginning of this chapter, an investigation was applied to find the possible presence of outliers.

"Outliers can be problematic in statistical analysis because they can skew the distribution of data and impact the results. Therefore, it is important to detect and reduce outliers to improve the quality and accuracy of statistical analysis." (Lawrence S. Meyers, 2012).

With the detection of these outliers, which can interfere with the accuracy of the results, one examination was made to check where exactly they were and what percentage of noise elimination was performed.



A picture containing diagram, parallel, rectangle, text

Description automatically generated

Figure 3: Finding the Outliers

Once checked and found the outliers in the Value label, which were impacting the testing results or determining unrealistic values, the next step was applying Winzorization for Outliers to reduce the noise in the dataset. Here was established that 1% of the highest value fall in the 99% percentile.

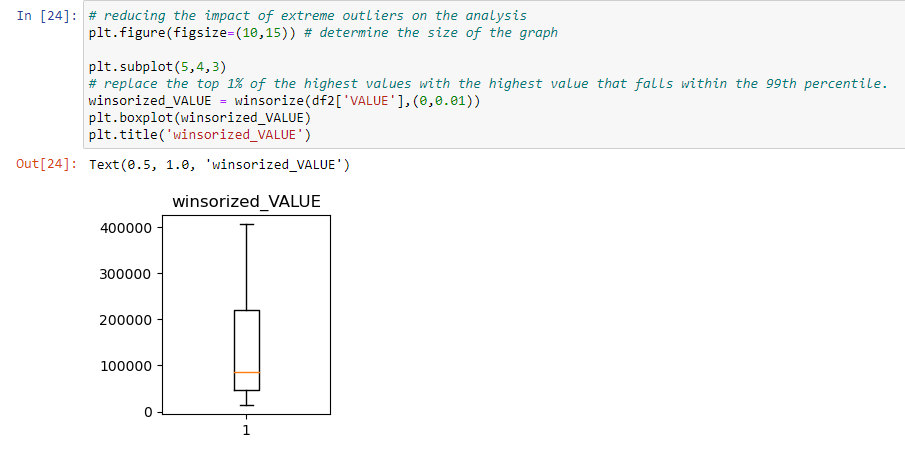


Figure 4: Applying Winzorization for Outliers

Following this, it was necessary to determine the main statistical measures for this study, such as mean, median, mode, range (difference between the maximum and minimum values), standard deviation, and variance.

For this, first, the 'describe( )' function was used to find out the average price of houses in Ireland between 1976 and 2016 year, which returned the values ​​in the VALUE label:

Mean = 135535.111498, std = 104381.88, min = 13900.00, the first quartile (25%) = 46406.75, the second quartile (50%) = 85490.00, the third quartile (75%) = 219907.75 and max = 405957.00,

as can be seen, illustrated in the table below.

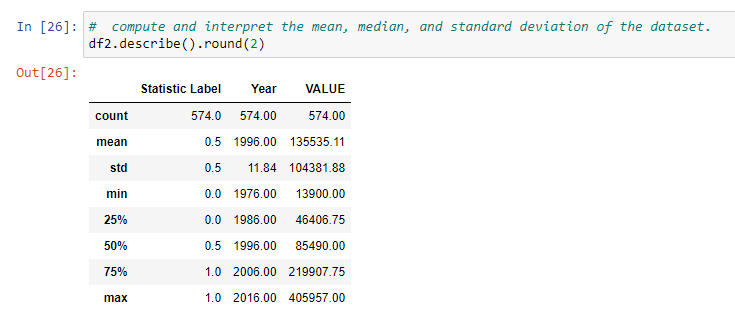


Figure 5: Determine the main statistical measures

So, it can be concluded that over the years (1976 and 2016), the average price was €135,535.11, being a standard deviation of €104,381.88. With these results, it is established that prices were not consistent during these years and that an average of €104,381.88 dispersed the mean.

With this, a graph was plotted to show the distribution of values ​​by region:

A picture containing text, screenshot, font

Description automatically generated

A picture containing diagram, rectangle, square, screenshot

Description automatically generated

Figure 6: Distribution of values by Area.

The graph above shows the comparison of values ​​distributed in each Area. It is noted that Dublin has the highest maximum value and is in its third quartile, surpassing even the national average.

Waterford appears in last place among the list of expensive areas in Ireland, having the cheapest region for buying new and second-hand homes. Subsequently, the second cheapest Area comes to Limerick, although the difference between these two regions is relatively small. Another way to see the result:

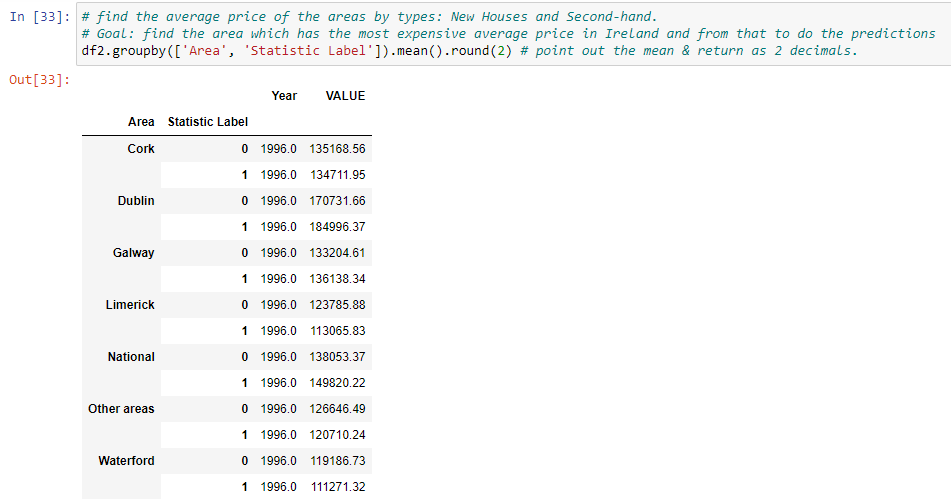


Figure 7: Average price of the areas by types: New Houses and Second-hand.

And then, a box graph was plotted to give a better overview and evaluate this result.

A picture containing text, font, screenshot

Description automatically generated

A picture containing text, screenshot, diagram, design

Description automatically generated

Figure 8: Average Price of Houses by Area.

As mentioned earlier, the first part of the study ends here at this point, where the most expensive Area to buy a home was discovered and from now on, only Dublin will be considered an inspection item. For this reason, the function ‘filtered’ was used to take out only the desired region.

A screenshot of a computer

Description automatically generated

Figure 9: Filter out only the most expensive Area.

## 1.1.3 – Distribution

Before justifying the chosen model and explaining the techniques developed, it is necessary to understand the reason for investigating the distribution within the exploration process. For that, a concept is given:

"The distribution of a dataset is one of the most important characteristics to consider when analysing data. Understanding the distribution allows us to better describe the data, make informed decisions about statistical techniques, and draw meaningful conclusions from our analyses". (Sullivan, 2018).

As the author above said, for a fair conclusion of the results, it is necessary to understand how the information is distributed and if they are uniform.

This report contains continuous numerical data, so the Normal Distribution model was used for the calculation.

For the first stage, a histogram was plotted to visualise the bell-shaped curve or Gaussian distribution and calculate whether it was evenly distributed.

A picture containing text, screenshot, font

Description automatically generated

A picture containing diagram, text, screenshot, plot

Description automatically generated

Figure 10: Distribution of Price variable

It can be noted in the graph that the algorithms were distributed in the form of waves between the bins, indicating there is no normal distribution. So, the next step was to apply the log function; thus, the algorithms were distributed on a more normal scale.

A picture containing text, screenshot, font, line

Description automatically generated

A picture containing diagram, line, plot, rectangle

Description automatically generated

Figure 11: Applying the logarithm function.

It can also be seen that the curve continues in a wave and not a bell shape. With this return, the Shapiro-Wilk test was applied to statistically calculate the sample and compare the result with the already known distribution. This test was chosen to be utilised because it is a dataset with little information and, consequently, a small sample.

A picture containing text, screenshot, font, line

Description automatically generated

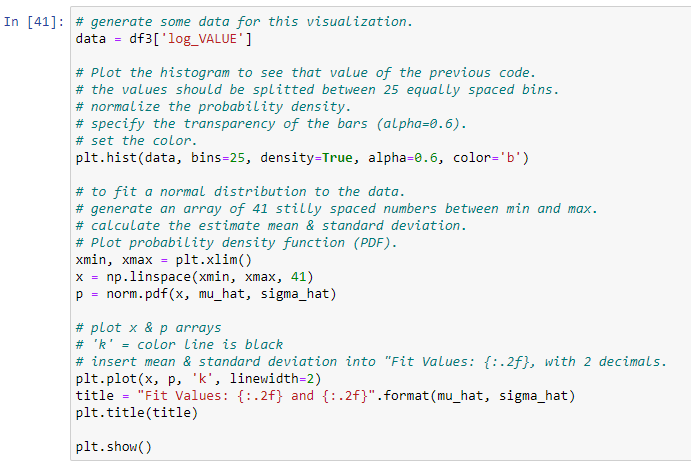
Figure 12: Shapiro-Wilk test

The value returned with the Shapiro-Wilk test is very small, which shows that the data is not normally distributed. However, as a last attempt to find a more consistent result and close this conclusion, the mean and standard deviation values ​​were calculated within the 'VALUE' column, which is the focus of this investigation. For this calculation, the code mu\_hat (keeps the value of the mean) and sigma\_hat (calculates the standard deviation values ​​in 'VALUE') and then a histogram is plotted to verify the curve.

A picture containing text, font, line, screenshot

Description automatically generated

Figure 13: Sample estimates of Mean & Standard deviation.



A picture containing screenshot, text, plot, diagram

Description automatically generated

Figure 14: Probability density function (PDF).

Regarding the delivered calculation and result, there is no proof that the value variable follows a normal distribution.

The Shapiro-Wilk test's p-value, which is substantially lower than the usual significance level of 0.05 and is very small (1.4846e-29), provides strong evidence against the null hypothesis and suggests that the data are not regularly distributed. The data is also non-stationary, meaning the price variable's mean and standard deviation increase over time. It could also be an indication of potential trends and economic issues.

## 1.1.4 – Probability

Another critical tool within the statistical descriptive is the precise calculation of the probability that a hypothetical situation or real-life factors will happen, determining the percentage of occurrence of these facts. This topic will not go into much depth, as it is not the focus of this essay, but as a way of demonstrating its function, it was used in *Jupyter Notebook* as part of the analysis and will be shared below.

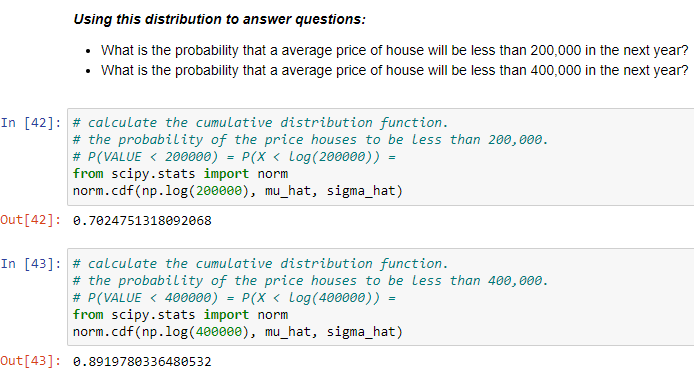


Figure 15: Probability calculation.

Then, he was asked about the possibility of house prices being lower than the values ​​of 200,000 and 400,000, being returned as a percentage that there is a high probability (70% & 89% consequently) of this happening.

## 1.2 - Data Preparation & Visualisation

As an essential component of any data examination, DataFrame preparation was performed as part of exploratory data analysis (EDA). First, check the shape (size of the dataset) and all the composite elements in it and divide them into types: numerical and categorical.

A screenshot of a computer

Description automatically generated

Figure 16: Analysing DataSet.

After learning about the information contained in the dataset, it was investigated which were the unique variables and how they were relevant to the analysis and whether these variables contained repeated evidence. Given this, it was determined that these columns would be dropped, and so it was done. This point is commonly named data cleaning.

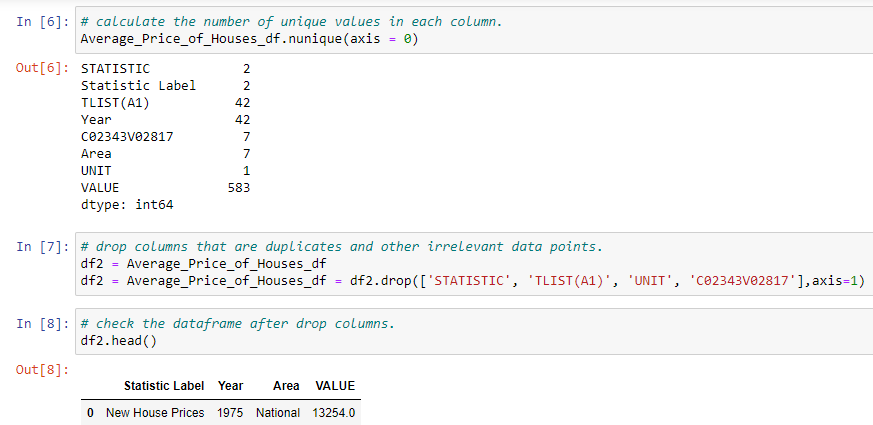


Figure 17: EDA: Cleaning Dataset.

The next step was to locate the missing values ​​and verify whether their position could interfere with the analysis results or whether they would be irrelevant to the study.

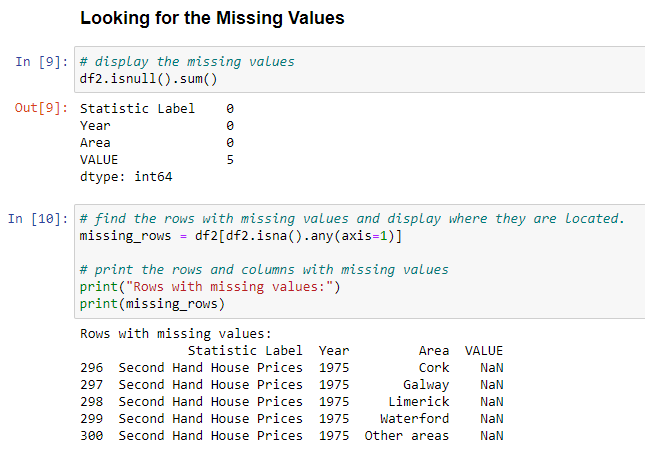


Figure 18: EDA: Checking missing values.

It was found that the missing values ​​referred to the lack of price corresponding to the average value of the houses in the year 1975. Since it is the first year of measurement, it was decided to drop these rows and start the analysis for the following year, 1976.

In other words, the decision to drop the missing values was made because it was the initial year. In addition, it was unnecessary to apply the mean instead. So, that was the code and the output:

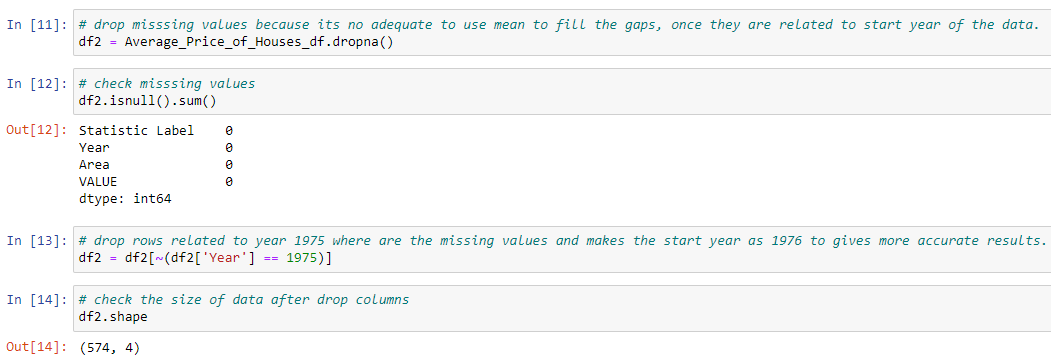
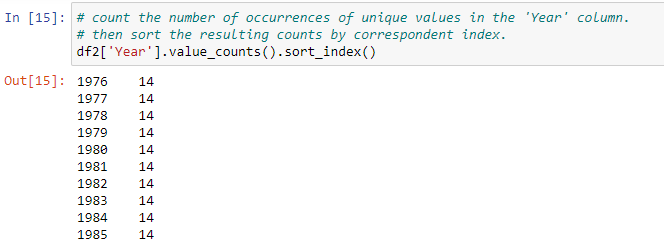


Figure 19: EDA: Cleaning missing values.

Then it was checked the unique values for each year and counted their occurrence.



Thus, the datatype of column 'Year' was changed, starting to be recognised as a valid index label (coding as .astype(‘datetime64[ns]’).

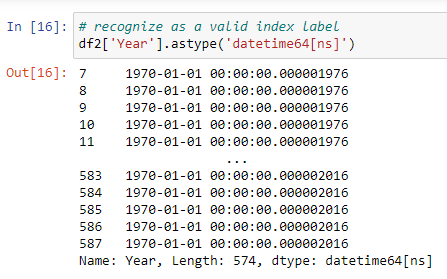


Figure 20: Recognising a valid label.

Finally, an overview of the dataset after cleaning and preparation will be shown. It is sorted by Area in descending order (coding as ascending=False).

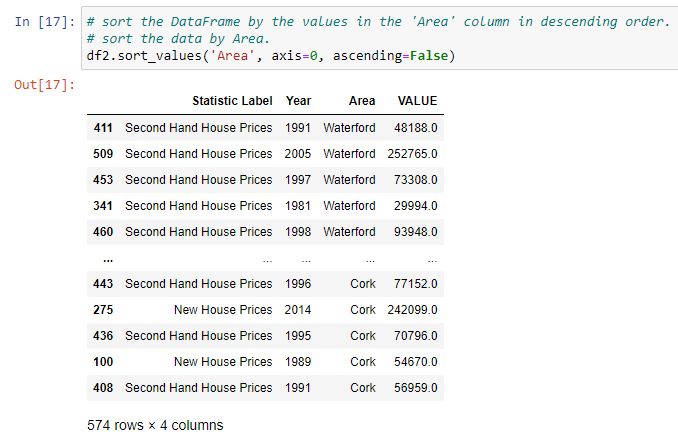


Figure 21: Sort Area by descending order.

Furthermore, it was necessary to understand and point out the correlation between the feature coefficient. Thus, it was applied the correlation matrix to find these values and plot the heatmap to answer the question: Are they correlated or not?

A picture containing text, font, line, screenshot

Description automatically generated

A picture containing text, screenshot, font

Description automatically generated

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

Figure 22: Correlation Heatmap

The heatmap shows the correlation between 'Year' & 'VALUE' is 0.860988, which means they are strongly positively correlated. In other words, it means when the value of 'Year' increases, the value of 'VALUE' also tends to increase. And, of course, they have perfectly correlated with it compares itself.

Using the *pair plot* to check the correlation between the ‘Statistic label’ and ‘Year’ & ‘VALUE’.

A picture containing text, font, line, screenshot

Description automatically generated

A picture containing text, diagram, screenshot, map

Description automatically generated

Figure 23: Pairplot: Correlation Matrix.

Note that the new houses and second-hand are correlated with themselves, where both increase and decrease in the same years. It demonstrates that was a fluctuation during the years. They were highly increasing from 1976 to earlier 2000 and then were a massive decreasing over the 2000 years. Now it can be seen that the average value is constantly growing, which can be explained by the current real estate crisis, where few properties are available in the residential market.

Afterwards, examining the correlation, an investigation was carried out on possible outliers, and the result of this analysis can be found in the above chapter.

At this point in the study, there was a curiosity to examine how this same oscillation between areas occurred. Thus, a line graph was chosen to compare the evolution of value *versus* time (year) in all different areas.



A picture containing text, screenshot, diagram, plot

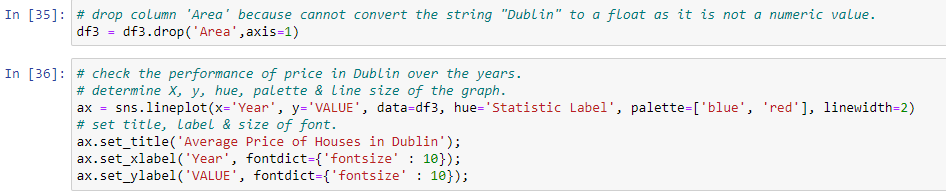
Description automatically generated

Figure 24: Lineplot: Evolution Value vs Year by Area.

As expected, the same fluctuation as in the previous chart was found, showing that a trend occurred during these years. Prices gradually grew faster, reaching values ​​250% higher in the 1990s and early 2000s than in the 1970s; values ​​plummeted in response to accelerated and emerging growth.

In short, it was a summary of the Exploratory Data Analysis that covered the first part of this assay.

For the second part, Dublin will be the central point of examination. Therefore, the region was filtered from the list and checked its average house value separately between new and second-hand. Once again, the line graph was chosen to compare the growth of the value *versus* time (year) of Dublin's 'Statistic label'.



A picture containing text, screenshot, plot, diagram

Description automatically generated

Figure 25: Lineplot: Average Price of houses in Dublin by Statistic label.

It is worth mentioning that both Data Preparation and Visualization were integrated throughout the entire analysis process. That is, both played a vital and indispensable role throughout the study and were present from the beginning to the end of the analysis.

In the next topic, some of the codes used, the libraries chosen, and the reason for these choices will be indicated.

## 1.3 - Programming for Data Analytics

## 1.3.1 - Justifications and Explanations of the chosen codes

In this chapter, you will find all the libraries used, most of the codes and the justifications for their use.

As the previous chapters were already very illustrative and with the intention of not being repetitive, the study will be more succinct and direct.

So, to begin with, when choosing the dataset, it was necessary to import some libraries to start the DataFrame and all its analysis. The library was increased in size throughout the examination and tests to adapt to what was being proposed.

Without further ado, the screenshot will be shown below and explained in sequence.

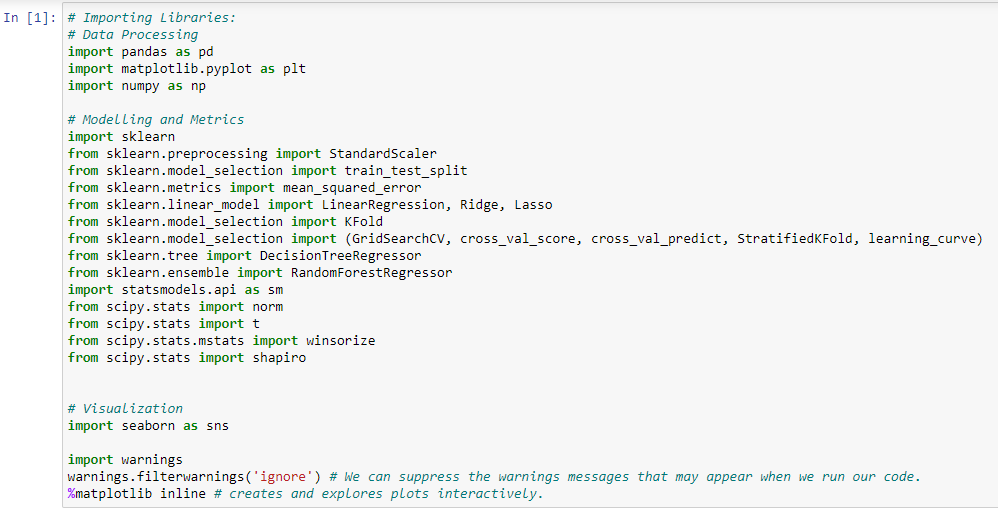


Figure 26: Importing libraries.

The first imported libraries are essential to any data analysis because they provide the complete Overview of your dataset and the most basic numerical functions. (Team, 2022). & (Foundation, 2021).

For example:

* Pandas: has the tools to clean, reshape and merge.
* Matplotlib: allows graphs of all complexities to be plotted, facilitating and helping the visualisation part.
* Numpy: permits the numerical data of the dataset to be worked on.

Other libraries were also imported to cover more specific subjects, such as:

* Scikit-learn: a library that composes machine learning tools for both supervised and unsupervised algorithms, being the basis of analysis in classification, clustering and regression. This library, in addition, provides sources for model analysis; it is essential to point out that it also contains elements for evaluation and validation.
* Statsmodels: provides metrics for statistical models in both numbers and visualisation.
* Scipy: is also a tool library for statistical norms. It can be used for Normal Distribution or other distribution models, as well as hypothesis testing models.
* Seaborn: It is another library that covers all visualisation parts but provides more complex plots such as categorical plots, heatmaps, and pair plots.

A table was set up as a more simplified way to demonstrate the basic codes used during the Exploratory Data Analysis and briefly explain what that code was referring to or doing.



Figure : Basics Codes used for (EDA).

From this point on, the codes became more complex and more specific according to the proposal of each subject such as statistical data, mathematics and Machine Learning models. Some codes had to be adapted according to demand during this assay. The most complex code that took the longest to reach a satisfactory result was the prediction, in which the author almost believed that she would not be able to find the expected result, but in the end, she achieved her objective.

All these codes can be seen in the *Jupyter Notebook*, along with their comments.

## 1.3.2 – Java versus Python: Overview - Another Programming Language

According to Hartman, Java has a more complex language, and its codes are also typed in a complex way. For example, while Python uses two lines of coding, Java will be structured in 10. Another great advantage of Java is its multi-platform, allowing it to run on any computer and cell phone. It also has notable performance in terms of connection. Java also has an easy-to-browse library. It is known among programmers for its excellent performance. (Hartman, 2023).

From another point of view, Pedamkar, likewise the previous author, states about the complexity of Java over Python. According to the table of the main characteristics of Python and Java, which the author provides on her page, in Java, numerical values ​​are not differentiated, being all read as floats. It also does not allow the use of hashtags, so comments cannot be made. (Pedamkar, 2023).

In summary, the two authors suggest that Python has a more straightforward and easier-to-understand language, but that Java performs better in connection and functioning.

It is my first-time having contact with the programming language. Thus, I need more technical and broader knowledge to describe any programming language than Python. During the research, I could check that the Java codes were long and at first contact, it did not look effortless.

Once again, due to the lack of intimacy with this world of programming, the codes in Jupyter Notebook would have to be highly adapted to the Java language if I must use it.

## 1.4 - Machine Learning for Data Analysis

## 1.4.1 - Project Management Framework

CRISP-DM was the chosen framework to assist in this project because its structure makes it possible to develop analysis in steps. It permits always reach to the previous point stages and improving the process. Check below how it fits into this study.

Business understanding: Ireland has been going through a terrible property crisis, both in its inflated value and the lack of available properties. Therefore, this study aims explicitly to verify the fluctuation of house prices over the years and determine possible increases or stagnation of these values.

Data understanding: To better understand what is happening with the average house values ​​in Ireland, a dataset was taken from the official government website that supplied all the analyses carried out. First, all its categorical and numerical elements and size dimensions were checked to start the next step.

Data preparation: After analysing the dataset, exploratory data analysis was initiated to clean and reduce outliers that could alter the results, leading to interpretation errors and unrealistic values.

Therefore, a search for missing values ​​was done, and it was decided that they should be dropped because they related to the dataset's initial year and would not interfere in the future. Some outliers were also found in the value column, and these were already causing noise and compromising the result. Then, the Winzorization technique was applied, which reduced 1% of the maximum value. With that, the dataset became more reliable. Other statistics metrics were utilised to check the mean and standard deviation. Furthermore, various graphics and metrics ​​were produced to find correlations and comparisons between categorical and numerical data. For example, how was the correlation of values ​​during the years and the Area with the highest expensive average price over the years.

Modelling: Because it was a dataset in which all information was known and the objective of this study was to return an expected output, the most suitable branch was Supervised Learning algorithms. In addition, the Regression model was chosen to analyse since the goal was to find between the continuous and numerical values of the average house prices to predict a possible increase or stagnation in future years. It is worth remembering that the intention was to arrive at output values ​​that are accurate and closer to the current ones. Additionally, to arrive at these results, Linear, Lasso and Ridge Regression were applied to determine which was better suited.

Evaluation: Evaluating the results of the regression models, was used the decision tree, which consists of nodes that represent the possible returns and, consequently, the decisions. These decisions are based on the internal attributes of the nodes, and each piece represents a prediction. The purpose of the tree is to separate similar returns into groups and thus make predictions for a new date. At this point, it is possible to check whether the study's primary objective has been achieved and if the result was satisfactory. Furthermore, an examination of the random forest was performed to verify and calculate the accuracy of the decision tree. Moreover, finally, K-fold cross-validation was applied to measure the accuracy of the prediction given by the linear regression model.

Deployment Phase: After all this process, the final stage is reached, where a development plan is made with the result found, and it is also the phase in which a report is developed explaining in detail each part of the tests, justifications and conclusions.

## 1.4.2 - Machine Learning Models

Having defined the objectives of predicting the average prices ​​of houses in Dublin and knowing that the dataset was based on variable and continuous values ​​in which the outputs would also be numerical and continuous variables, Regression was the model chosen for its analysis. The models, Linear Regression, Lasso Regression and Ridge Regression, were operated in a trial-and-error way to make the study more accurate. All these techniques had excellent performance indicators. It should be mentioned that to make the analyses more accurate, the GridSearchCV hyperparameter was used, which by definition:

"GridSearchCV is a method to search the candidate best parameters exhaustively from the grid of given parameters. It is often used in machine learning to tune the hyperparameters of a model. The method performs an exhaustive search over a specified parameter grid, and for each combination of hyperparameters, it trains a model and evaluates its performance using cross-validation." (Raschka, 2019).

This hyperparameter was used to cross-validate the Lasso Regression and Ridge Regression metrics. It was determined as the 'alpha' parameter to regulate the control of the applied force in the model and verbose = 1 to control the amount of output.

The K-fold cross-validation method with five folders was used to validate the Linear Regression outputs, and the results were excellent. The R-squared scores range from 0.7378 to 0.8691, which shows that the linear regression model can account for a sizeable percentage of the prediction in housing prices and that its performance is generally consistent between folds.

Furthermore, two different methods were added to the study, Decision Tree and Random Forest, to test and examine for better solutions.

As Guido (2016) said, “Gradient boosted decision trees are among the most powerful and widely used models for supervised learning”.

And, “Random forests for regression and classification are currently among the most widely used machine learning methods. They are very powerful, often work well without heavy tuning of the parameters, and don’t require scaling of the data”. (Guido, 2016).

Due to their strengths and importance to this material, they were integrated into the analyses.

## 1.4.3 - Machine Learning Modelling Comparisons: Similarities and differences

Among the investigations, there were many similarities between the models, especially between their results. All of them behaved in such a way as to achieve the objectives and thereby bring the tangible values ​​that were being sought. Of course, on the other hand, there were the techniques that stood out: Decision Tree and Random Forest for having their returns with better fits. As can be checked in the table below.

A picture containing text, screenshot, font, number

Description automatically generated

Figure 28: Modelling Comparisons

According to the table, the Decision Tree and Random Forest models had the best fit, as they had higher R-squared (test) and lower RMSE (test) values than Linear Regression, Lasso Regression and Ridge Regression. Yet, these last three models had the same R-squared value (train) and similarity in R-squared (test) & RMSE (test), indicating that they all accurately represent the training dataset.

Complementing this, a table was created comparing the current and predicted values ​​between the two methods that had high performance.

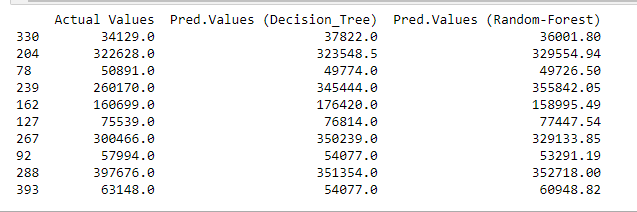


Figure 29: Actual Values versus Prediction Values.

It can be noted that the values ​​did not have a large discrepancy and were mainly similar to the outputs of random numbers. In other words, both printed good results and values ​​very similar to the current ones, maintaining a favourable response to the chosen metrics.

# Chapter II - Results and Discussion

To answer the central question of this project, how much will the average price of new and second-hand homes in the most expensive area of ​​Ireland be in 2026? After cleaning and knowing the dataset, the first step was to determine each area's average and compare them. Then to find the object of research and what would require further analysis. The seven regions (National, Dublin, Cork, Galway, Limerick, Waterford, and Other areas) were analysed. Their statistics metrics (mean, standard deviation, minimum & maximum value sand quartiles) and averages were displayed in a graph to better understand their dimensions and differences, as they can be seen in chapter I, 1.1.2 – Descriptive Statistics.

Given this, it was concluded that Dublin was the most expensive area. With that, the study advanced to the second part, which was to make a correlation between the variable's values ​​and years, where it was realized that both were utterly correlated insofar as the Years went on increasing, the values ​​also increased, although, here making an addition, that in the years 2008 and 2009 the values ​​suffered an abrupt drop due to the real estate crisis, where the sector had been heating up to the point that prices were exaggeratedly high. However, after this crisis, the market has been recovering since then. Consequently, prices have been increasing, generating another type of crisis where the low real estate supply is increasingly present, causing a direct impact on prices and affecting the relationship between offers and demand.

Subsequently comprehension this correlation of value and year, it was time to identify their correlation concerning new and second-hand houses. Once again, it was concluded that they were correlated, going through the same oscillation in the market in which they were inserted. After understanding the correlation between these variables, another metric utilised was the normal distribution to find out if, at any point, the value was evenly distributed, and here the result was reached that there was no normal distribution since the 'value' was not was stationary and so were the 'years'. That is, they were in continuous growth. The probability method was applied to determine the percentage of values ​​less than 200,000 and 400,000. Both had a percentage above 70 and 80, which can be judged that this is consistent with the current price of the data time.

Moving on to the next concept, it was time to apply Machine Learning methods, which, as already mentioned, had a branch of the Supervised learning algorithm, as its possible outputs were known. In addition, Regression was the model employed to find the prediction of values ​​over the years.

This entire method and how each technique were operated can be found in chapter I, 1.4 - Machine Learning for Data Analysis.

Afterwards applying all these techniques and knowing how each one performed, it was time to code for tangible value. For this, the Linear Regression model was chosen, which gave the following prices as shown in the table below:

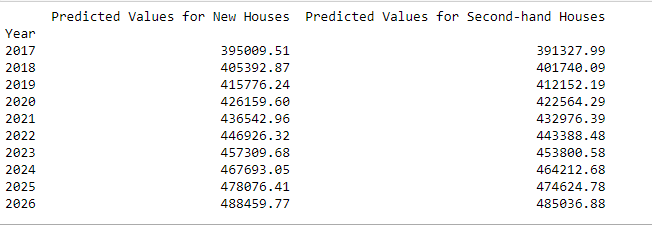


Figure 30: Predicted Values for 'Statistics labels' for the year 2026.

Additionally, a graph is displayed to compare the prediction of average price of new and second-hand houses in Dublin over 2017 e 2026.

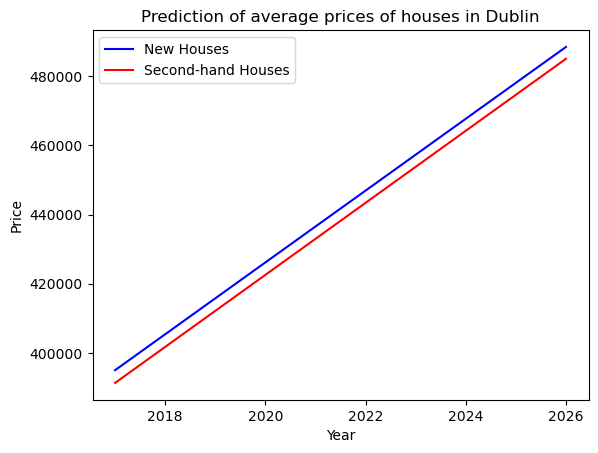


Figure 31: Prediction of average of houses in Dublin for the year 2026.

To close this chapter, it seems that the objective was reached and that the results were positive and satisfactory.

# Conclusions

Through the dataset chosen and all analyses completed, it can be concluded that this project was able to achieve its objectives.

Nevertheless, it is important to emphasise that there were consistent boundaries directly linked to the analysis of predictions that the database presented 2016 as the last base year, which may or may not represent different values ​​from the current reality, considering that an unforeseen and devastating pandemic hit the world and changed the price of products and services, including those subject to analysis in this report, the average value of homes in Ireland. It could have significantly impacted any prediction with the presented data.

# References

Department of Housing, Local Government, and Heritage, n.d.. *Data.Gov.IE.* [Online]   
Available at: https://data.gov.ie/dataset/hsa06-average-price-of-houses  
[Accessed 20 March 2023].

Foundation, P. S., 2021. *Python 3.10.1 documentation.* [Online]   
Available at: https://docs.python.org/3/library/index.html  
[Accessed 12 April 2023].

Grolemund, G. &. W. H., 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* 1st ed. Sebastopol, CA, USA.: O'Reilly Media, Inc. .

Guido, A. C. M. &. S., 2016. *Introduction to Machine Learning with Python.* 1st ed. Sebastopol, CA, USA.: O’Reilly Media, Inc..

Hartman, J., 2023. *Guru 99.* [Online]   
Available at: https://www.guru99.com/java-vs-python.html  
[Accessed 14 April 2023].

Lawrence S. Meyers, G. C. G. A. J. G., 2012. *Applied Multivariate Research: Design and Interpretation.* 2nd ed. California, USA: Sage Publications Inc..

Pedamkar, P., 2023. *EDUCBA.* [Online]   
Available at: https://www.educba.com/python-vs-javascript/  
[Accessed 14 April 2023].

Raschka, S. &. M. V., 2019. *Python Machine Learning.* 3rd ed. Birmingham, UK.: Packt Publishing.

Sullivan, M., 2018. *Fundamentals of statistics.* 5th ed. s.l.:Pearson Education, Inc..

Team, G. L., 2022. *Great Learning.* [Online]   
Available at: https://www.mygreatlearning.com/blog/open-source-python-libraries/  
[Accessed 12 April 2023].

**https://ec.europa.eu/eurostat/databrowser/view/HSW\_N2\_01\_\_custom\_6154887/default/table?lang=en**

<https://ec.europa.eu/eurostat/databrowser/view/HSW_N2_02__custom_6154914/default/table?lang=en>

<https://ec.europa.eu/eurostat/databrowser/view/SBS_SC_CON_R2__custom_6155496/default/table?lang=en>