# Group ID - MSc in Data Analytics

Author: Sonia Guzman

e-mail: 2023070@student.cct.ie

Student ID: 2023070

## Abstract

*This paper provides an overview of the average price of a new house in Ireland over the period 1997 to 2015. It reviews the factors that drove the fluctuations and the factors that influenced those fluctuations in the main cities of the country during the mentioned period, with a particular focus on Dublin. The analysis is based on the index Average New House Price, a dataset the Department of Housing Local Government and Heritage in the domain of Housing constructs using data supplied by the mortgage lending agencies on loans approved by them rather than loans paid. It is also important to highlight that the values in the dataset are a result of a mix of houses (including apartments). The study will utilize a few statistical techniques and data visualization tools such as to analyse and interpret the data, and we will try to build a model using Linear Regression method of Machine Learning for predicting property prices.*

*Keywords: fluctuations, factors, average, new house price, Dublin*

**Table of Contents**

[Group ID - MSc in Data Analytics 1](#_Toc132310958)

[Abstract 2](#_Toc132310959)

[Introduction 5](#_Toc132310960)

[Materials and Methods 7](#_Toc132310961)

[Project Implementation Tools and Programming Language 7](#_Toc132310962)

[Project Management Methodology: CRISP-DM Model 7](#_Toc132310963)

[Real-life Scenario 9](#_Toc132310964)

[The CRISP-DM Methodology Applied 10](#_Toc132310965)

[Business Understanding 10](#_Toc132310966)

[Data Understanding 10](#_Toc132310967)

[Collect the Data 11](#_Toc132310968)

[EDA and Data Preparation 11](#_Toc132310969)

[Data Exploration/Visualization 14](#_Toc132310970)

[Statistics: Distribution 18](#_Toc132310971)

[Model Building and Results 20](#_Toc132310972)

[Model Building 20](#_Toc132310973)

[Evaluation 23](#_Toc132310974)

[Deployment 23](#_Toc132310975)

[Conclusion 25](#_Toc132310976)

[Acknowledgements 26](#_Toc132310977)

[References 27](#_Toc132310978)

**List of Figures**

[Figure 1: Six Phases of CRISP-DM. Image taken from Project Management Methodologies Week 4 PPT. CCT College Dublin 2022 9](#_Toc132310583)

[Figure 2: First Five Rows of the Dataset after Dropping the original Index 12](#_Toc132310584)

[Figure 3: Field Types of the Dataset 12](#_Toc132310585)

[Figure 4: Updated Dataset 13](#_Toc132310586)

[Figure 5: Structure of the Dataset 13](#_Toc132310587)

[Figure 6: Null Value Detection 14](#_Toc132310588)

[Figure 7: Heatmap of Features Correlation 15](#_Toc132310589)

[Figure 8: Average House Price in Ireland (1997-2015) 16](#_Toc132310590)

[Figure 9: Percentage of difference between one year and the next 17](#_Toc132310591)

[Figure 10: Box plot of Mean House Prices by City 17](#_Toc132310592)

[Figure 11: Box plot of Mean House Prices by City 18](#_Toc132310593)

[Figure 12: KDE Plot of Dublin 19](#_Toc132310594)

[Figure 13: Calculates the probability that a new house in Dublin with a price of €400000 or less will be sold 19](#_Toc132310595)

[Figure 14: Calculates the probability that a new house in Dublin being sold for more than €400000 20](#_Toc132310596)

[Figure 15: Machine Learning 22](#_Toc132310597)

[Figure 16: GridSearchCV 22](#_Toc132310598)

## Introduction

We have used python Jupiter Notebook for a simple visualization of the used dataset.

## 

## Datasets

### General Information About the Dataset – Country: Ireland.

Sector: Construction.

Topic: New Dwelling Completions.

Period: 2011 – 2022.

Source: New Dwelling Completions, [Central Statistics Office](https://data.gov.ie/organization/central-statistics-office) (CSO). Link: <https://data.gov.ie/dataset/nda02-new-dwelling-completions>

Licensed under: Creative Commons Attribution 4.0.

Data set name file: NDA02.20230515111519.csv

Data Structure: The data is structured in a time-series format and each observation is recorded annually by type of house at a national level.

Variables: The original dataset is composed by eight (8) variables, as follows: STATISTIC, STATISTIC Label, TLIST(A1), Year, C02342V02816, Type of House, UNIT and VALUE.

Type of Variables:

(a) Quantitative variables: TLIST(A1), Year (discrete independent variable), C02342V02816 and VALUE.

(b) Qualitative variables: STATISTIC, STATISTIC Label, Type of House and UNIT.

#### Data Preparation, EDA, Visualization and Statistics

The most relevant codes used in the Jupiter Notebook for the data preparation of the dataset mentioned above are described below:

* Importing the required libraries using the code ‘import as’ and ‘from import’ to gain access to pre-existing codes and materials that can be used to perform arithmetic operations, visualize data, machine learning, etc.

A picture containing text, screenshot, font

Description automatically generated

Figure 1: Importing required libraries

* Loading the csv file and checking the first five (5) rows using the code ‘pd.read\_csv("mycsvfile.csv")’ and ‘.head()’ respectively.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 2: Loading dataset and displaying the first five (5) rows of the dataset

* Using the code: *‘.shape’* to return the number of rows and columns.
* Using the code: ‘*.info()’* to display a shorter summary of the dataset. This method helps to quickly inspect the structure of a dataset and identify any missing or null values in the data. From the figure below, it can be seen that that the dataset is composed by 49 rows and 8 columns with no NULL values, it can also be seen the type of data per column and/or variant.

A screenshot of a computer code

Description automatically generated with medium confidence

Figure 3: Structure of the Dataset

* The describe code ‘.describe’ provides purely descriptive information about the dataset. This function is extremely important since it will allow us to visualize all the data contained in the dataset.
* The code ‘Print(ppg.columns)’ provides a list of all the columns in the dataset, which can be useful if you need to drop, modify and/or add columns.
* Dropping unnecessary columns: After analysing the data displayed using the ‘.describe’ and ‘Print(ppg.columns)’ codes, it can be appreciated that there are some columns that are not needed for the purpose of the analysis, therefore, they were removed from the dataset using the following code: ‘ppg.drop(['STATISTIC', 'TLIST(A1)', 'C02342V02816', 'UNIT',], axis=1, inplace=True’).
* Renaming the remaining columns to normalize the data using the code: ‘.rename(columns={'STATISTIC Label': 'Statistic Label', 'VALUE': 'IRL Total'})’.
* Adding new columns using the code: ‘ppg['Country\_1'] = 'Ireland'’.
* Setting the column ‘Year’ as the new index of the dataset using the code: ‘ppg = ppg.set\_index('Year')’ so it can be easy to plot the data if needed.
* Checking the index using the code: ‘print(ppg.index)’.
* Checking for missing values in the data using the code: ‘ppg.isna().sum()’.
* Creating a new csv file with the index in it using the code: ‘ppg.to\_csv('/Users/AP139GE/MSc\_CA2/Dataset/NDC.csv', index=True)’.
* Plotting the data with a line graph using the go.scatter function for a more dynamic and descriptive graph as shown in the figure below.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 4: New Dwelling Completions 2011-2022

In the figure above it can be seen that although in 2011 the number of completions of scheme houses was below variant single house completions, 1,358 and 4,814 respectively, in 2015 the trend of scheme house completions started to improve significantly, going from 1,795 in 2014 to 3,294 in 2015, which represents an increase of 83,51% from one year to the next.

From the three types of new houses completions shown in Figure 4, the variant apartment was the lowest through 2020. It wasn’t until that year that the mentioned variant started to see a change in the number of completions, jumping from 3,909 units up to 5,126 in 2021, which represents an improvement of 31,13%. Compared with the number of single houses completed in 2021, the variant apartment grew 8.19%. This improvement placed this type of construction above the single house trend in the following year 2022.

On the other hand, the variant scheme house began to experience an exponential growth since 2015, going from 3,294 to 15,160 in 2022, which represents a growth of more than 400%. So far, the scheme house is the one that leads the number of completions in the market since 2015 and compared with the other two type of houses, it seems to be the type of house that is most in demand.

It is important to note that despite the COVID restrictions in 2020, the total number of new dwelling completions was 20,574, down -2,61% from the 21,126 in 2019. The impact of the COVID-19 pandemic and associated restrictions continued to affect completions in 2021, where the total number of new dwelling completions was 20,553, down -0,10% from the 20,574 in 2020. It wasn’t until 2022 that the total number of new dwelling completions began to raise, 29,822 up 45,10% from the 20,553 in 2021. This level of completions shows a great recovery from the impact of the COVID-19 pandemic and the associated restrictions.

* Histogram of number of New Dwelling Completions. The histogram shown in the figure below represents the distribution of the 'IRL Total' values. The x-axis of the histogram represents the range of the 'IRL Total' values, divided into 20 equally spaced bins. The y-axis represents the frequency or count of occurrences of 'IRL Total' values falling within each bin. The 'IRL Total' values increase gradually from 2011 to 2016, with some fluctuations. However, from 2017 onwards, there is a significant increase in the values. This trend is indicated by the increasing height of the bars in the histogram. The majority of the 'IRL Total' values seem to be concentrated in the lower range of the histogram, especially in the earlier years. As the years progress, the distribution spreads out.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 5: Histogram of New Dwelling Completions 2011-2022

* Descriptive statistics for each type of house. By analysing the statistics in Figure 6, we can infer the following:

1. All house types: The std of 8292.46 indicates a relatively wide spread of values around the mean. This suggests that the number of completions for all house types is varied. The mean of 13606.75 represents the average number of completions for all house types. The min and max values of 4575.0 and 29822.0, respectively, indicate that the range of completions for all house types spans from the lowest to the highest values observed.
2. Apartment: The std of 2586.04 suggests a moderate spread of values around the mean. The mean of 2533.25 represents the average number of completions for apartments. The min and max values of 446.0 and 9147.0, respectively, indicate the range of completions for the variant apartments.
3. Scheme house: The std of 5188.63 suggests a relatively wide spread of values around the mean. The mean of 6877.50 represents the average number of completions for scheme houses. The min and max values of 964.0 and 15160.0, respectively, indicate the range of completions for the variant.
4. Single house: The std of 896.243473 suggests a relatively narrow spread of values around the mean. The mean of 4196.00 represents the average number of completions for single houses. The min and max values of 2947.0 and 5515.0, respectively, indicate the range of completions for the variant.

Overall, descriptive statistics provide information about the distribution of completion numbers, the range of values observed and the average or central tendency for each type of house. They indicate the variation and spread of completion numbers within each variant.

A picture containing text, font, screenshot, algebra

Description automatically generated

Figure 6: Descriptive statistics for each type of house

* Looking at Figure 7, the boxplot shows the descriptive statistics measures of the three different types of houses "Apartment," "Scheme house," and "Single house". Additionally, there is a row called "All house types" that represents the combined statistics for all types of houses. From the boxplot can be infer that (excluding the variant ‘all type of house’ the variant ‘scheme house’ has the highest median and highest overall mean number of completions, followed by the variant ‘single house’ and the variant apartment has the lowest rate of completions. Overall, the box plot shows a clear and concise summary of the distribution of mean for each variant, allowing us to identify which type of house has the highest or lowest number of completions, as well as any significant outliers in the data.

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

Figure 7: Box plot of Descriptive statistics for each type of house

* Normal Distribution of new dwelling completions. This function could be used to explain/identify some information about the dataset. The variant used to define the parameters of the normal distribution is ‘all type of house’. The mean values of this variant represent the center of the distribution, while the standard deviation indicates the spread of the distribution. The resulting plot below shows a bell-shaped curve, which is a characteristic shape of a normal distribution. The x-axis represents the range of the number of completions, and the y-axis represents the probability density. The highest point of the curve represents the most common value in the dataset. The curve tapers off to the left and right, indicating the decreasing probability of getting a value that is further away from the mean value.

A picture containing plot, diagram, line, screenshot

Description automatically generated

Figure 8: Normal Distribution of # of NDC

* KDE. The mean and std values from the variant ‘all type of house’ were used to plot the histogram that shows the random data generated according to a normal distribution. Figure 9 shows a histogram of randomly generated data that follows a normal distribution.

A picture containing plot, diagram, text, screenshot

Description automatically generated

Figure 9: KDE

### General Information About the Dataset – Country: Northern Ireland.

Sector: Construction.

Topic: New Dwelling Completions. New dwelling completions by property type

Period: 2005 – 2022.

Source: data.gov.uk. Link: <https://www.data.gov.uk/dataset/613a03f5-a7f7-4ed3-97e4-c9c0ca0486c9/northern-ireland-new-dwelling-completions/datafile/b4b36fd1-6482-40c3-8a15-3e75a404756b/preview>

Licensed under: Open Government Licence v3.0.

Data set name file: Northern Ireland New Dwelling Completions.csv

Data Structure: The data is structured in a time-series format and each observation is recorded quarterly by property type at a national level.

Variables: The original dataset is composed by five (5) variables, as follows: Quarter Year, Apartments, Houses, Total New Dwelling Completions.

Type of Variables: All the variables in the dataset are quantitative variables.

#### Data Preparation, EDA, Visualization and Statistics

The most relevant codes used in the Jupiter Notebook for the data preparation of the dataset mentioned above are described below:

* Loading the csv file and checking the first five (5) rows using the code ‘pd.read\_csv("mycsvfile.csv")’ and ‘.head()’ respectively.

A screenshot of a computer

Description automatically generated with low confidence

Figure 9: Loading dataset and displaying the first five (5) rows of the dataset

* Splitting the column ‘Quarter Year’ into two columns to have the quarter and the year separately. The code used was the following: *ppg1[['Quarter', 'Year']] = ppg1['Quarter Year'].str.split(' ', expand=True).*
* Droping the original column that had the combined values ‘Quarter Year’ using the code: *ppg1.drop(columns=['Quarter Year'], inplace=True).*
* Printing the resulting dataframe using the code: *print(ppg1),* and *ppg1.head()* to visualize the result of the changes made. See figure 10 below.

A screenshot of a computer

Description automatically generated with low confidence

Figure 10: Displaying the first five (5) rows of the dataset

* Using the code: *‘.shape’* to return the number of rows and columns.
* Using the code: ‘*.info()’* to display a shorter summary of the dataset. This method helps to quickly inspect the structure of a dataset and identify any missing or null values in the data. From the figure below, it can be seen that that the dataset is composed by 72 rows and 5 columns and includes NULL values.

A screenshot of a computer code

Description automatically generated with low confidence

Figure 11: Structure of the Dataset

* The describe code ‘.describe’ provides purely descriptive information about the dataset. This function is extremely important since it will allow us to visualize all the data contained in the dataset.
* Setting the column ‘Year’ as the new index of the dataset using the code: ‘ppg1 = ppg1.set\_index('Year')’ so it can be easy to plot the data if needed.
* Checking the index using the code: ‘print(ppg1.index)’.
* Dropping rows to keep only the rows with years between 2011 and 2022. The code used was the following: ppg1 = ppg1.loc['2011':'2022'].
* The code ‘Print(ppg1.columns)’ provides a list of all the columns in the dataset, which can be useful if you need to drop, modify and/or add columns.
* Dropping unnecessary columns: After analysing the data displayed using the ‘.describe’ and ‘Print(ppg1.columns)’ codes, it can be appreciated that there are some columns that are not needed for the purpose of the analysis, therefore, they were removed from the dataset using the following code: ppg1.drop(['Quarter'], axis=1, inplace=True).
* Renaming the remaining columns to normalize the data using the code: ppg1 = ppg1.rename(columns={'Total \nNew Dwelling \nCompletions': 'NorthIRL Total'}).
* Checking for missing values in the data using the code: ‘ppg1.isna().sum()’. From Figure 12 it can be seen that after dropping some columns there are no more missing values in the dataset.

A screenshot of a computer

Description automatically generated with low confidence

Figure 12: Structure of the Dataset

* Converting index to a datetime object. To do so, the code used was the following: ppg1.index = pd.to\_datetime(ppg1.index), ppg1.index = pd.to\_datetime(ppg1.index).strftime('%Y'), print(ppg1.head()).
* Grouping the data by year and sum the values for each column using the code: ppg1\_grouped = ppg1.groupby('Year').sum(), # Print the grouped data, print(ppg1\_grouped).
* Saving the grouped data to a new csv file using the code: ppg1\_grouped.to\_csv('ppg1.csv').
* Adding new columns using the code: ppg1\_grouped['Country\_2'] = 'Northern Ireland', ppg1\_grouped['Statistic Label'] = 'New Dwelling Completions'.
* Updating the DataFrame to a new csv file using the code: ppg1\_grouped.to\_csv('ppg\_country.csv', index=False).
* Printing the resulting dataframe using the code: *ppg1\_grouped.head().*See figure 13 below.

A screenshot of a computer

Description automatically generated with low confidence

Figure 13: Displaying the first five (5) rows of the updated dataset

* Plotting the data with a line graph using the go.scatter function for a more dynamic and descriptive graph as shown in the figure below.

A picture containing diagram, plot, line, text

Description automatically generated

Figure 14: New Dwelling Completions 2011-2022

Figure 14 shows the total new dwelling completions in Northern Ireland and its fluctuations over the years. Over the years there have been both increases and decreases in the number of completions; However, the dominant trend has been an increase in the number of completions. It can be appreciated that the values generally increase from 2011 to 2018. The highest number of completions occurred in 2018 with a count of 7644 dwellings completed, representing the highest spike in the data.

Following the peak in 2018, there is a decline in the number of completions in subsequent years. The years 2019 and 2020 have lower completion count compared to 2018. It can be inferred that the trend in 2020 continued to be low due to the effects of COVID-19. In 2021, can be seen that there is a nearly 16% recovery in the number of completions, up to 7,420 from the lowest count in 2020. In 2021 and 2022 the completions seem to stabilize at a relatively high level compared to the earlier years.

Overall, the graphic suggests that the total new dwelling completions in Northern Ireland, excluding the effects of the COVID-19 in 2020, has been experiencing an underlying trend of exponential growth since 2016.

### General Information About the Dataset – Country: England.

Sector: Construction.

Topic: New Dwelling Completions.

Period: 1978 – 2022.

Source: data.gov.uk. Link: <https://www.gov.uk/government/statistical-data-sets/live-tables-on-house-building>

Licensed under: Open Government Licence v3.0.

Data set name file: England Dwelling.csv

Data Structure: The data is structured in a time-series format and each observation is recorded quarterly by property type at a national level.

Variables: The original dataset is composed by ten (10) variables, as follows: Revised, Period, Started – All Dwellings, Started - Private Enterprise, Started - Housing Associations, Started - Local Authorities, England Completed - All Dwellings, Completed - Private Enterprise, Completed - Housing Associations, Completed - Local Authorities.

Type of Variables: All the variables in the dataset are quantitative variables.

#### Data Preparation, EDA, Visualization and Statistics

The most relevant codes used in the Jupiter Notebook for the data preparation of the dataset mentioned above are described below:

* Loading the csv file and checking the first five (5) rows using the code ‘pd.read\_csv("mycsvfile.csv")’ and ‘.head()’ respectively.



Figure 15: Loading dataset and displaying the first five (5) rows of the dataset

* Using the code: *‘.shape’* to return the number of rows and columns.
* Using the code: ‘*.info()’* to display a shorter summary of the dataset. This method helps to quickly inspect the structure of a dataset and identify any missing or null values in the data. From the figure below, it can be seen that that the dataset is composed by 179 rows and 5 columns and includes NULL values.

A screenshot of a computer

Description automatically generated

Figure 16: Structure of the Dataset

* The code ‘Print(ppg2.columns)’ provides a list of all the columns in the dataset, which can be useful if you need to drop, modify and/or add columns.
* Dropping columns that are not needed. The code used is: *ppg2.drop(['Revised', 'Started - All Dwellings', 'Started - Private Enterprise', 'Started - Housing Associations', 'Started - Local Authorities', 'Completed - Private Enterprise', 'Completed - Housing Associations', 'Completed - Local Authorities'], axis=1, inplace=True).*
* Renaming the remaining columns to normalize the data using the code: ppg2 = ppg2.rename(columns={'England Completed - All Dwellings': 'Total\_C3'}).
* Extract the year from the "Period" column using the code: ppg2['Year'] = ppg2['Period'].str[-4:].
* Converting the 'Year' column to datetime format using the code: ppg2['Year'] = pd.to\_datetime(ppg2['Year'], format='%Y').dt.year.
* Setting the "Year" column as the index using the code: ppg2.set\_index('Year', inplace=True).
* Dropping rows to keep only the rows with years between 2011 and 2022. The code used was the following: ppg2 = ppg2.loc['2011':'2022'].
* Using the code: ‘*.info()’* to review a summary of the dataset.

A screenshot of a computer code

Description automatically generated with low confidence

Figure 17: Structure of the Dataset

* Converting the values in column 1 "Total C\_3" to numeric data using the code: *ppg2['Total\_C3']=ppg2['Total\_C3'].str.replace(',','').astype(float).*
* Dropping the column ‘Period’ using the code: ppg2.drop(['Period'], axis=1, inplace=True).
* Adding new columns using the code: ppg2\_g['Country\_3'] = 'England'.
* Plotting the data with a line graph using the go.scatter function for a more dynamic and descriptive graph as shown in the figure below.

A picture containing text, diagram, plot, line

Description automatically generated

Figure 18: New Dwelling Completions 2011-2022

Figure 18 shows the total new dwelling completions in England and its fluctuations over the years. From 2011 to 2012, there is a slight increase in the total new dwelling completions from 114,030 to 115,590; However, in 2013, there is a decline in the number of completions to 109,450, experiencing an increase of 8% in 2014 when the completions went up to 117,820.

The year 2015 marks a notable increase in total new dwelling completions with a substantial jump to 142,480. This growth continues into 2016, where the number remains high at 141,880. Furthermore, from 2017 to 2019, there is a consistent growth trend, with the completions rising from 162,470 to 177,880.

In 2020, there is a significant decrease in completions, dropping to 146,630. This decrease could be attributed to the negative effects of the COVID-19. However, the following year, 2021, sees a notable recovery with an increase in total new dwelling completions to 174,930.

Overall, the graphic suggests that the total new dwelling completions in England over the reporting period have been very mixed, with periods of growth, decline and recovery. The significant drop from 2019 to 2020 suggests that the decline was due to the effects of the COVID-19 but may also be due to a combination of the pandemic mixed with other factors such as economic conditions, government policies, demand-supply dynamics, and other market factors.

### General Information About the Dataset – Country: Scotland.

Sector: Construction.

Topic: New Dwelling Completions.

Period: 1978 – 2022.

Source: data.gov.uk. Link: <https://www.gov.uk/government/statistical-data-sets/live-tables-on-house-building>

Licensed under: Open Government Licence v3.0.

Data set name file: Scotland Permanent Dwelling.csv

Data Structure: The data is structured in a time-series format and each observation is recorded quarterly by property type at a national level.

Variables: The original dataset is composed by ten (10) variables, as follows: Revised, Period, Started – All Dwellings, Started - Private Enterprise, Started - Housing Associations, Started - Local Authorities, England Completed - All Dwellings, Completed - Private Enterprise, Completed - Housing Associations, Completed - Local Authorities.

Type of Variables: All the variables in the dataset are quantitative variables.

#### Data Preparation, EDA, Visualization and Statistics

The most relevant codes used in the Jupiter Notebook for the data preparation of the dataset mentioned above are described below:

* Loading the csv file and checking the first five (5) rows using the code ‘pd.read\_csv("mycsvfile.csv")’ and ‘.head()’ respectively.

A screenshot of a computer

Description automatically generated with low confidence

Figure 19: Loading dataset and displaying the first five (5) rows of the dataset

* Using the code: *‘.shape’* to return the number of rows and columns.
* Using the code: ‘*.info()’* to display a shorter summary of the dataset. This method helps to quickly inspect the structure of a dataset and identify any missing or null values in the data. From the figure below, it can be seen that that the dataset is composed by 179 rows and 5 columns and includes NULL values.

A picture containing text, receipt, font, algebra

Description automatically generated

Figure 20: Structure of the Dataset

* The code ‘Print(ppg3.columns)’ provides a list of all the columns in the dataset, which can be useful if you need to drop, modify and/or add columns.
* Dropping columns that are not needed. The code used is: *ppg3.drop(['Revised', 'Started - All Dwellings', 'Started - Private Enterprise', 'Started - Housing Associations', 'Started - Local Authorities', 'Completed - Private Enterprise', 'Completed - Housing Associations', 'Completed - Local Authorities'], axis=1, inplace=True).*
* Renaming the remaining columns to normalize the data using the code: ppg2 = ppg3.rename(columns={'Scotland Completed - All Dwellings': 'Total\_C4'}).
* Extract the year from the "Period" column using the code: ppg3['Year'] = ppg3['Period'].str[-4:].
* Converting the 'Year' column to datetime format using the code: ppg3['Year'] = pd.to\_datetime(ppg3['Year'], format='%Y').dt.year.
* Setting the "Year" column as the index using the code: ppg3.set\_index('Year', inplace=True).
* Dropping rows to keep only the rows with years between 2011 and 2022. The code used was the following: ppg3 = ppg3.loc['2011':'2022'].
* Using the code: ‘*.info()’* to review a summary of the updated dataset.

A screenshot of a computer code

Description automatically generated with low confidence

Figure 21: Structure of the Dataset

* Converting the values in column 1 "Total C\_4" to numeric data using the code: *ppg3['Total\_C4']=ppg3['Total\_C3'].str.replace(',','').astype(float).*
* Dropping the column ‘Period’ using the code: ppg3.drop(['Period'], axis=1, inplace=True).
* Adding new columns using the code: ppg3\_g['Country\_4'] = 'Scotland'.
* Plotting the data with a line graph using the go.scatter function for a more dynamic and descriptive graph as shown in the figure below.

A picture containing plot, line, diagram, text

Description automatically generated

Figure 22: New Dwelling Completions 2011-2022

Figure 22 shows the total new dwelling completions in Scotland and its fluctuations over the years. During 2011-2014 the total new dwelling completions in Scotland remained relatively stable. However, in 2014 and till 2019, there was a consistent upward trend in new dwelling completions, the numbers increased from 15610 in 2014 to 22790 in 2019, indicating a significant growth in the construction of new dwellings in Scotland. In 2021 there was a substantial increase in new dwelling completions, with the number rising to 21150. This indicates a recovery from the dip in 2020 and suggests a renewed growth in the construction sector.

Overall, the graphic suggests that there was a fluctuation trend in the total new dwelling completions in Scotland in early years, reaching its maximum peak in 2019. The significant drop from 2019 to 2020 suggests that the decline was due to the effects of the COVID-19.

### Combining Datasets – Countries: Ireland, Northern Ireland, England and Scotland.

Sector: Construction.

Topic: New Dwelling Completions.

Period: 2011 – 2022.

Source: Python.

Licensed under: N/A.

Data set name file: updated\_NDC.csv

Data Structure: The data is structured in a time-series format and each observation is recorded annually at a national level per country.

Variables: The original dataset is composed by eleven (11) variables, as follows: Year, Statistic Label, Type of House, Total\_C1, Country\_1, Country\_2, Total\_C2, Country\_3, Total\_C3, Country\_4, Total\_C4.

Type of Variables:

1. Quantitative variables: Year (discrete independent variable), Total\_C1, Total\_C2, Total\_C3, Total\_C4.
2. (b) Qualitative variables: Statistic Label, Type of House, Country\_1, Country\_2, Country\_3, Country\_4,

#### Data Preparation, EDA, Visualization and Statistics

The most relevant codes used in the Jupiter Notebook for the data preparation of the dataset mentioned above are described below:

* Loading the csv file and checking the first five (5) rows using the code ‘pd.read\_csv("mycsvfile.csv")’ and ‘.head()’ respectively.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 23: Loading dataset and displaying the first five (5) rows of the dataset

* Adding new columns to the dataset as shown in the Figure below.

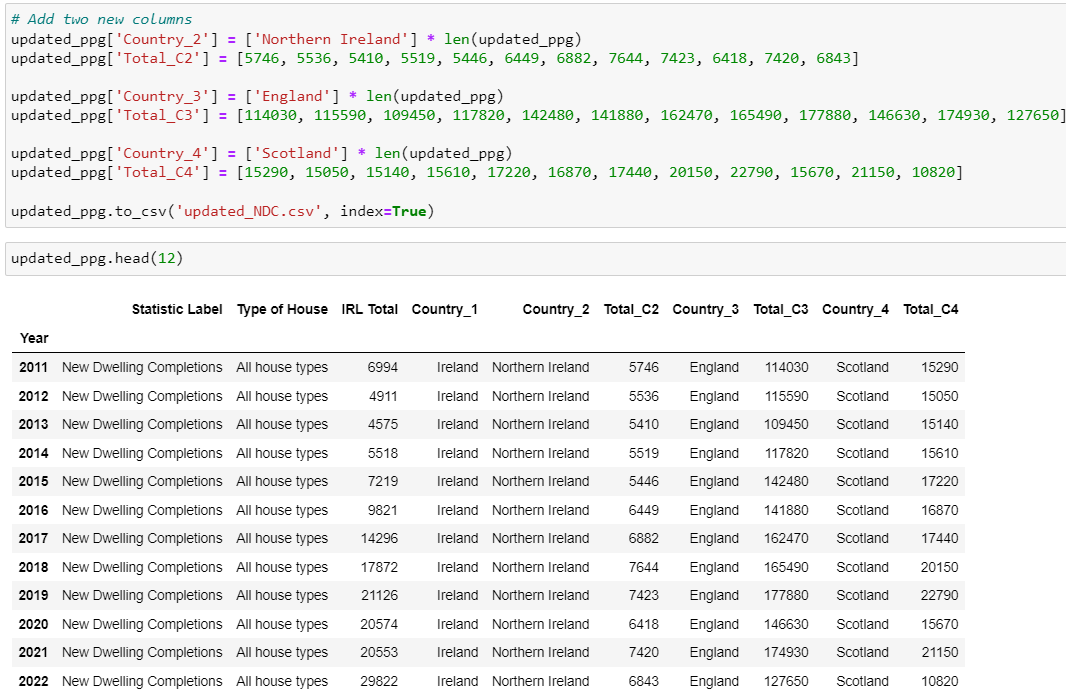


Figure 24: Adding new columns to the dataset

* Renaming the remaining columns to normalize the data using the code: updated\_ppg = updated\_ppg.rename(columns={'IRL Total': 'Total\_C1'})
* Updating csv file using the code: updated\_ppg.to\_csv('updated\_NDC.csv', index=True)
* Using the code: *‘.shape’* to return the number of rows and columns.
* Using the code: ‘*.info()’* to display a shorter summary of the dataset. This method helps to quickly inspect the structure of a dataset and identify any missing or null values in the data. From the figure below, it can be seen that that the dataset is composed by 12 rows and 10 columns with no NULL values.

A screenshot of a computer

Description automatically generated

Figure 25: Structure of the Dataset

* Plotting the data with a line graph using the go.scatter function for a more dynamic and descriptive graph as shown in the figure below.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 26: New Dwelling Completions 2011-2022

Figure 26 showcase the total new dwelling completions per country from 2011 to 2022. England had the highest number of new dwelling completions among the countries and experienced consistent growth until 2018, followed by significant fluctuations. In contrast the other three (3) countries (Ireland, Northern Ireland, and Scotland) are more or less similar in terms of number of new dwelling completions, with some fluctuations but no substantial growth or decline patterns observed.

* Descriptive statistics per country (Ireland, Northern Ireland, England and Scotland). By analysing the statistics in Figure 26, we can infer the following:

1. Ireland: The std of 8292.46 indicates a relatively wide spread of values around the mean. This suggests that the number of completions for all house types is varied. The mean of 13606.75 represents the average number of completions for all house types. The min and max values of 4575.0 and 29822.0, respectively, indicate that the range of completions for all house types spans from the lowest to the highest values observed.
2. Northern Ireland: The std of 847.77a suggests a moderate spread of values around the mean. The mean of 6394.66 represents the average number of completions for all house types. The min and max values of 5410.00 and 7644.00, respectively, indicate the range of completions for the variant apartments.
3. England: The std of 24659.88 suggests a relatively wide spread of values around the mean. The mean of 141358.33 represents the average number of completions for all house types. The min and max values of 109450.00 and 177880.00, respectively, indicate the range of completions for the variant.
4. Scotland: The std of 3204.46 suggests a relatively narrow spread of values around the mean. The mean of 16933.33 represents the average number of completions for all house types. The min and max values of 10820.00 and 22790.00, respectively, indicate the range of completions for the variant.

Overall, descriptive statistics provide information about the distribution of completion numbers, the range of values observed and the average or central tendency for each country. They indicate the variation and spread of completion numbers within each variant/country.

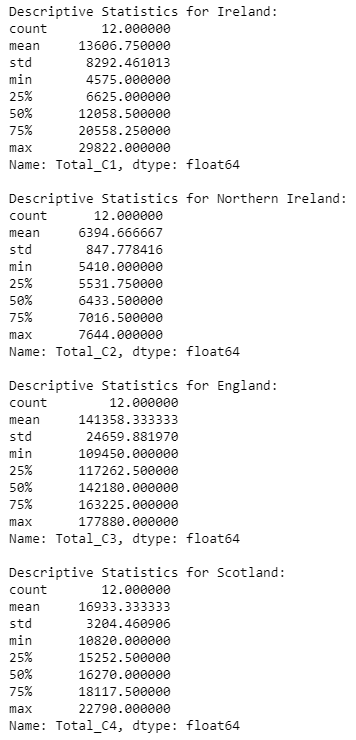


Figure 26: Descriptive statistics for each country

* Looking at Figure 27, the boxplot shows the descriptive statistics measures of the four countries (Ireland, Northern Ireland, England, and Scotland). From the boxplot can be infer that England has the highest median and highest overall mean number of completions, followed by Scotland and Ireland, and last but not least Northern Ireland has the lowest rate of completions during the given period. Overall, the box plot shows a clear and concise summary of the distribution of mean for each country, allowing us to identify what country has the highest or lowest number of completions, as well as any significant outliers in the data.

A picture containing text, screenshot, diagram, software

Description automatically generated

Figure 27: Box plot of descriptive statistics for each country

* Looking at Figure 27, the boxplot shows the descriptive statistics measures of the four countries (Ireland, Northern Ireland, England, and Scotland). From the boxplot can be infer that England has the highest median and highest overall mean number of completions, followed by Scotland and Ireland, and last but not least Northern Ireland has the lowest rate of completions during the given period. Overall, the box plot shows a clear and concise summary of the distribution of mean for each country, allowing us to identify what country has the highest or lowest number of completions, as well as any significant outliers in the data.

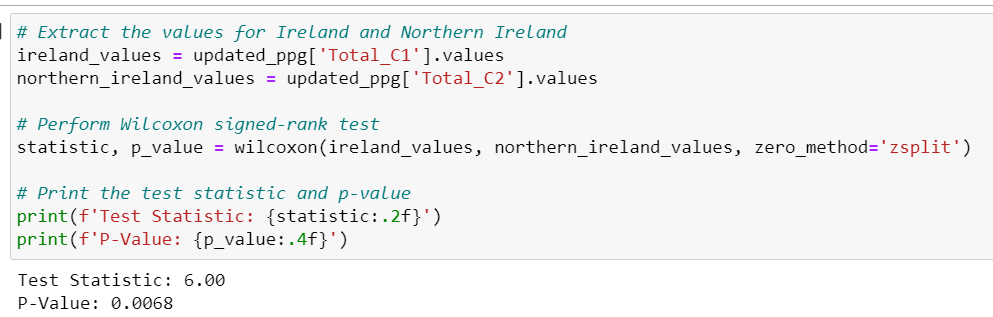
### Parametric analyses and non-parametric analyses

#### Two sample t-test



Figure 28: Two sample t-test results

#### Wilcoxon Test



Northern Ireland

Test Statistic: 6.00

P-Value: 0.0068

### England:

Test Statistic: 0.00

P-Value: 0.0005

Scotland:

Test Statistic: 17.00

P-Value: 0.0923

Figure 29: Wilcoxon Test results

#### Chi-squared test

### 

Figure 30: Chi-squared test results

#### Anova Test

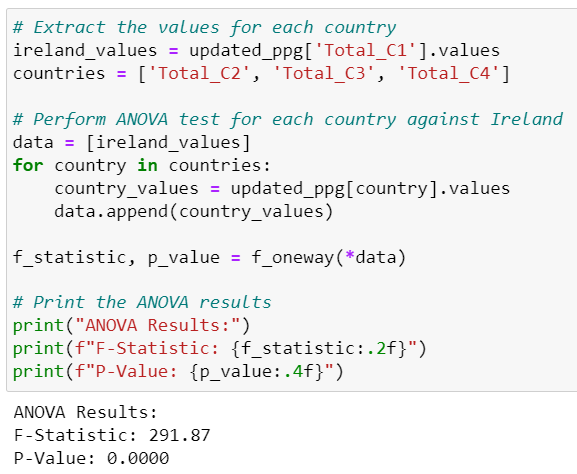


Figure 31: Anova test results

#### Rank test



Figure 32: Rank test results

### Import Database from MySQL to a CSV file through Python

As part of the assessment, we manually entered data using MySQL Workbench to create a database that would then be imported into Python as a CSV file. See figures below.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 23: MySQL database action output

A screenshot of a computer

Description automatically generated with medium confidence

Figure 23: MySQL data output setting index



Figure 23: Importing required libraries and entering codes

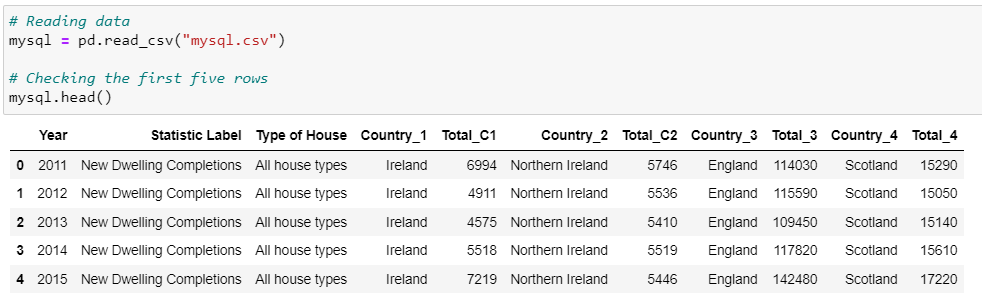


Figure 23: Checking the data imported

### Model Building and Results

#### Model Building

There are three types of machine learning algorithms we were asked to choose from to use for our project: supervised, semi-supervised and unsupervised.

1. Supervised machine learning is defined as machine learning done with labelled data. Labelled data could for instance be a dataset containing attributes describing a set of cakes and a label for each cake indicating whether the cake tastes good or not.
2. Unsupervised machine learning is learning without a label set. The learner does not know specifically what prediction is “correct” and cannot correct itself according to labels as supervised learning can.
3. Semi-supervised learning is similar to supervised learning, but instead uses both labelled and unlabelled data. Labelled data is essentially information that has meaningful tags so that the algorithm can understand the data, whilst unlabelled data lacks that information. By using this combination, machine learning algorithms can learn to label unlabelled data.

The type of machine learning we will be operating within is supervised machine learning. In the case of real estate price prediction, the label is tied to price of the houses and the rest of the data can be attributes describing the year.

When creating a machine learning model, the input given to the learner can be divided into different types of data (Shalev-Shwartz and Ben-David, 2014). The learner is given a training set of data, which contains a domain set of data and a label set of data. The domain set is the data describing attributes believed to be predictive of the label set. The label entries are tied to specific entries in the domain set.

RESEARCH QUESTIONS

1. Create a model to predict property pricing.

2. What are the features that influence the property pricing?

METHODOLOGY

The CRISP-DM process is used to achieve the project's goal. CRISP-DM stands for Cross

Industry Standard Process for Data Mining. The CRISP-DM methodology is used for

structured data mining project planning. CRISP-DM is divided into six stages.

RESEARCH QUESTIONS

1. Create a model to predict property pricing.

2. What are the features that influence the property pricing?

METHODOLOGY

The CRISP-DM process is used to achieve the project's goal. CRISP-DM stands for Cross

Industry Standard Process for Data Mining. The CRISP-DM methodology is used for

structured data mining project planning. CRISP-DM is divided into six stages.

RESEARCH QUESTIONS

1. Create a model to predict property pricing.

2. What are the features that influence the property pricing?

METHODOLOGY

The CRISP-DM process is used to achieve the project's goal. CRISP-DM stands for Cross

Industry Standard Process for Data Mining. The CRISP-DM methodology is used for

structured data mining project planning. CRISP-DM is divided into six stages.

RESEARCH QUESTIONS

1. Create a model to predict property pricing.

2. What are the features that influence the property pricing?

METHODOLOGY

The CRISP-DM process is used to achieve the project's goal. CRISP-DM stands for Cross

Industry Standard Process for Data Mining. The CRISP-DM methodology is used for

structured data mining project planning. CRISP-DM is divided into six stages.

RESEARCH QUESTIONS

1. Create a model to predict property pricing.

2. What are the features that influence the property pricing?

METHODOLOGY

The CRISP-DM process is used to achieve the project's goal. CRISP-DM stands for Cross

Industry Standard Process for Data Mining. The CRISP-DM methodology is used for

structured data mining project planning. CRISP-DM is divided into six stages.

The output of the learner is a prediction rule that can be used to label new data. The success of a trained model can be measured as the error of predictions made to new data. The less the error, the better the model.

Under the umbrella of supervised learning fall: Classification, Regression and Forecasting.

* Classification: In classification tasks, the machine learning program must draw a conclusion from observed values and determine to what category new observations belong. For example, when filtering emails as ‘spam’ or ‘not spam’, the program must look at existing observational data and filter the emails accordingly.
* Regression: In regression tasks, the machine learning program must estimate – and understand – the relationships among variables. Regression analysis focuses on one dependent variable and a series of other changing variables – making it particularly useful for prediction and forecasting.
* Forecasting: Forecasting is the process of making predictions about the future based on the past and present data and is commonly used to analyse trends. A method of forecasting and now-casting of prices in each market is by extrapolation using a price index. The price index works by using the average market trend to extrapolate a previous sales price to current time. The index can be described as:

Based on the above, the approach that we will take is Regression and Forecast since they are the ones that best fit the task of predicting the price of a house with an error as small as possible.

The mean squared error (MSE) is a measure of the average squared difference between the predicted and actual values. In this case, the MSE value of 11204648207.808912 suggests that the model's predictions are not very accurate, as there is a large difference between the predicted and actual values.

The predicted prices for the years 2016 to 2020 are shown as an array of five values, each representing the predicted price for the corresponding year. The predicted prices are based on the trained model, which uses the relationship between the year and house prices in Dublin to make these predictions.

For example, the predicted price for 2016 is 392399.80 euros, which means that the model predicts that the average price of a house in Dublin in 2016 would be around 392399.80 euros. Similarly, the predicted prices for 2017, 2018, 2019, and 2020 are 401788.56, 411177.32, 420566.08, and 429954.85 euros, respectively.

Graphical user interface, text, application, email

Description automatically generated

Figure 15: Machine Learning

Regarding hyperparameter tuning, for the Linear Regression model, there is only one hyperparameter to tune, which is the regularization parameter alpha. We can use GridSearchCV or RandomizedSearchCV to find the optimal value for alpha. Using GridSearchCV, GridSearchCV searches for the best value of alpha from a list of possible values. The cv parameter specifies the number of cross-validation folds to use. After fitting the GridSearchCV object, you can access the best hyperparameters using the best\_params\_ attribute. You can then use the best hyperparameters to train the final model on the full training set and evaluate its performance on the testing set.

Graphical user interface, text, application

Description automatically generated

Figure 16: GridSearchCV

The output "Best hyperparameters: {'alpha': 100}" suggests that the best value for the hyperparameter 'alpha' was found to be 100 using the hyperparameter tuning method (either GridSearchCV or RandomizedSearchCV).

The hyperparameter 'alpha' is used in the Ridge regression algorithm, which is a variant of linear regression that adds a regularization term to the loss function to prevent overfitting. The 'alpha' parameter controls the strength of the regularization, with higher values of 'alpha' resulting in stronger regularization and lower values resulting in weaker regularization.

The fact that the best value of 'alpha' found was 100 indicates that a relatively strong regularization was needed to achieve the best performance on the data. This could suggest that the data was prone to overfitting and that stronger regularization was needed to prevent this.

Overall, the best hyperparameter value of 'alpha' found using hyperparameter tuning suggests that a Ridge regression model with strong regularization would perform well on the data. However, it's important to note that this is just one possible hyperparameter combination and other combinations could also result in good performance.

## Conclusion

The widespread acknowledgement that actual housing completions in the Irish housing

## Acknowledgements

I would like to thank my family for being support during this stressful time.

## References

The department of [housing](https://moodle.cct.ie/mod/resource/view.php?id=129150) local government and heritage in the domain of [HOUSING](https://moodle.cct.ie/mod/resource/view.php?id=129150), This data is available at: <https://data.gov.ie/dataset?q=houses&organization=department-of-housing-local-government-and-heritage&sort=score+desc%2C+metadata_created+desc&theme=Housing>