# Group ID - MSc in Data Analytics

Author: Sonia Guzman

e-mail: 2023070@student.cct.ie

Student ID: 2023070

# Abstract

*This paper aims to compare the values of different variables related to new dwelling completions between Ireland and other three countries (Northern Ireland, England, and Scotland). The dataset consists of the number of new dwelling completions from 2011 to 2022 per country at a national level. The analysis focuses on comparing the total number of new dwelling completions across the years for each country. The objective is to identify any significant differences or trends in the housing construction industry between these regions. The findings from this comparative analysis can provide insights into the housing market dynamics and construction activities in the selected countries. This information may be valuable for policymakers, researchers, and industry professionals interested in understanding the variations and patterns in new dwelling completions across different regions. The study will utilize a few statistical techniques and data visualization tools such as to analyse and interpret the results.*

*Keywords: comparative analysis, variables, trends, statistical analysis, results.*

**Table of Contents**

[Abstract 2](#_Toc136006908)

[Introduction 5](#_Toc136006909)

[Datasets 6](#_Toc136006910)

[General Information About the Dataset – Country: Ireland. 6](#_Toc136006911)

[Data Preparation, EDA, Visualization and Statistics 6](#_Toc136006912)

[General Information About the Dataset – Country: Northern Ireland. 12](#_Toc136006913)

[Data Preparation, EDA, Visualization and Statistics 12](#_Toc136006914)

[General Information About the Dataset – Country: England. 16](#_Toc136006915)

[Data Preparation, EDA, Visualization and Statistics 16](#_Toc136006916)

[General Information About the Dataset – Country: Scotland. 18](#_Toc136006917)

[Data Preparation, EDA, Visualization and Statistics 19](#_Toc136006918)

[Combining Datasets – Countries: Ireland, Northern Ireland, England and Scotland. 21](#_Toc136006919)

[Data Preparation, EDA, Visualization and Statistics 22](#_Toc136006920)

[Parametric and Non-parametric Analyses 26](#_Toc136006921)

[Two Sample t-test 26](#_Toc136006922)

[Wilcoxon Signed-rank Test 27](#_Toc136006923)

[Chi-squared Test 29](#_Toc136006924)

[Anova Test 29](#_Toc136006925)

[Rank Test 30](#_Toc136006926)

[Import Database from MySQL to a CSV file through Python 31](#_Toc136006927)

[Machine Learning 32](#_Toc136006928)

[Linear Regression 35](#_Toc136006929)

[Sentiment Analysis on Construction Sector in Ireland 35](#_Toc136006930)

[Challenges 36](#_Toc136006931)

[Conclusion 37](#_Toc136006932)

[Acknowledgements 39](#_Toc136006933)

**List of Figures**

[Figure 1: Importing required libraries 6](#_Toc136005244)

[Figure 2: Loading dataset and displaying the first five (5) rows of the dataset 6](#_Toc136005245)

[Figure 3: Structure of the Dataset 7](#_Toc136005246)

[Figure 4: New Dwelling Completions 2011-2022 8](#_Toc136005247)

[Figure 5: Histogram of New Dwelling Completions 2011-2022 9](#_Toc136005248)

[Figure 6: Descriptive statistics for each type of house 10](#_Toc136005249)

[Figure 7: Box plot of Descriptive statistics for each type of house 10](#_Toc136005250)

[Figure 8: Normal Distribution of # of NDC 11](#_Toc136005251)

[Figure 9: KDE 12](#_Toc136005252)

[Figure 10: Loading dataset and displaying the first five (5) rows of the dataset 13](#_Toc136005253)

[Figure 11: Displaying the first five (5) rows of the dataset 13](#_Toc136005254)

[Figure 12: Structure of the Dataset 13](#_Toc136005255)

[Figure 13: Structure of the Dataset 14](#_Toc136005256)

[Figure 14: Displaying the first five (5) rows of the updated dataset 15](#_Toc136005257)

[Figure 15: New Dwelling Completions 2011-2022 15](#_Toc136005258)

[Figure 16: Loading dataset and displaying the first five (5) rows of the dataset 16](#_Toc136005259)

[Figure 17: Structure of the Dataset 17](#_Toc136005260)

[Figure 18: Structure of the Dataset 17](#_Toc136005261)

[Figure 19: New Dwelling Completions 2011-2022 18](#_Toc136005262)

[Figure 20: Loading dataset and displaying the first five (5) rows of the dataset 19](#_Toc136005263)

[Figure 21: Structure of the Dataset 20](#_Toc136005264)

[Figure 22: Structure of the Dataset 20](#_Toc136005265)

[Figure 23: New Dwelling Completions 2011-2022 21](#_Toc136005266)

[Figure 24: Loading dataset and displaying the first five (5) rows of the dataset 22](#_Toc136005267)

[Figure 25: Adding new columns to the dataset 23](#_Toc136005268)

[Figure 26: Structure of the Dataset 23](#_Toc136005269)

[Figure 27: New Dwelling Completions 2011-2022 24](#_Toc136005270)

[Figure 28: Descriptive statistics for each country 25](#_Toc136005271)

[Figure 29: Box plot of descriptive statistics for each country 26](#_Toc136005272)

[Figure 30: Two sample t-test results 26](#_Toc136005273)

[Figure 31: Wilcoxon signed-rank test results 28](#_Toc136005274)

[Figure 32: Chi-squared test results 29](#_Toc136005275)

[Figure 33: Anova test results 30](#_Toc136005276)

[Figure 34: Rank test results 30](#_Toc136005277)

[Figure 35: MySQL database creation 31](#_Toc136005278)

[Figure 36: MySQL data output setting index 31](#_Toc136005279)

[Figure 37: Importing required libraries and entering codes 32](#_Toc136005280)

[Figure 38: Checking the data imported 32](#_Toc136005281)

[Figure 39: Importing required libraries 33](#_Toc136005282)

[Figure 40: ARIMA, testing model and forecasting 33](#_Toc136005283)

[Figure 41: Coding to plot and forecast values 34](#_Toc136005284)

[Figure 42: Forecast of new dwelling completions 35](#_Toc136005285)

[Figure 43: Linear regression code 36](#_Toc136005286)

[Figure 44: Importing required libraries for sentiment analysis 37](#_Toc136005287)

# Introduction

Companies all over the world try to get the benefits from accessing to information that is available in social media

to improve their performance and increase their revenue, processing heterogeneous type of data to extract the valuable data is a

problem that many organizations try to solve. One of the most important trends is in general known as “Big Data”, technology

for Storing, Processing and analyzing data, companies are Managing data in order to use it in new levels and direct decision

makers to make agile decisions in real time, Big Data trend have the capability to guide a revolutionary transformation in

research, invention, and business marketing. In this research we highlight some aspects of Big Data and its importance on

organizations' business performance and how companies can use the famous open source platform Hadoop to process data to

gain the competitive advantage.

Companies all over the world try to get the benefits from accessing to information that is available in social media

to improve their performance and increase their revenue, processing heterogeneous type of data to extract the valuable data is a

problem that many organizations try to solve. One of the most important trends is in general known as “Big Data”, technology

for Storing, Processing and analyzing data, companies are Managing data in order to use it in new levels and direct decision

makers to make agile decisions in real time, Big Data trend have the capability to guide a revolutionary transformation in

research, invention, and business marketing. In this research we highlight some aspects of Big Data and its importance on

organizations' business performance and how companies can use the famous open source platform Hadoop to process data to

gain the competitive advantage.

Data analytics has become an integral part of various research fields due to the rapid advancements in the digital technologies available for dealing with data. The construction industry is no exception and has seen a spike in the data being generated due to the introduction of various digital disruptive technologies. However, despite the availability of data and the introduction of such technologies, the construction industry is lagging in harnessing big data. Algorithm development, machine learning (ML), statistical analysis, and computational model development are among the various techniques that depend on data that can be easily gathered by day-to-day usage gadgets. The presence of bulks of data makes it possible for researchers to make informed decisions and conduct relevant analyses for their ﬁeld of study.

The biggest challenge in data management is identifying which data is useful and vice versa through data reﬁnement. The immense amounts of data easily available make it hard to identify the datasets used for a particular purpose. Moreover, the available data format may not be ready for use or easily readable for the intended purpose. These barriers to accessing, understanding, and utilizing data make it important to develop systems for extracting key information and analysing it. In the case of construction, some barriers to data adoption include latency, data privacy, data availability, data governance, poor broadband connectivity at construction sites, and cost implication for long-term use. Furthermore, there is an increase in vulnerability in technology adoption due to the ﬂuidity of security parameters. Storing construction design and ﬁnancial information in shared resources concerns the construction industry. Construction is a data-intensive sector where the bulk of data is generated and not capitalized on adequately due to slow technology adoption.

This paper explores the use of data analytics techniques and visualisation tools to identify data trends and how the construction industry can benefit from data. Specifically, the paper will compare the total number of new dwelling completions in Northern Ireland, England and Scotland against Ireland over the period 2011-2022. This will allow us to identify any significant differences, trends, or patterns in the housing market of these regions. The findings from this analysis can shed light on the dynamics of new dwelling completions over the countries under study and gain some insight to improve the Irish market.

The project utilizes datasets containing few variables, which includes the total number of new dwelling completions in Ireland, Northern Ireland, England and Scotland from 2011 to 2022. The chosen tool to analyse the datasets is Jupyter Notebook and the programming language is Python, this will allow us to process, clean, and transform the data for meaningful analysis. Subsequently, we will employ statistical techniques and data visualizations tools to compare and present a comprehensive and detailed comparative analysis of new dwelling completions across Ireland, Northern Ireland, England and Scotland.

In the following sections, we will delve into the data analysis process, showcasing how Python and Jupyter Notebook are utilized to handle and analyse the dataset for comparative analysis.

# 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 : 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 : 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. 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 : 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 : 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. As the plot shows, most of the values are concentrated in the lower range of the histogram, especially in the earlier years. However, as the years progress, the distribution spreads out.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure : Histogram of New Dwelling Completions 2011-2022

* Descriptive statistics for type of house:

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 : 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 : 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 the variant represent the centre of the distribution, while the standard deviation shows the spread of the distribution. The plot in Figure 8 shows a bell-shaped curve, which is a characteristic shape of a normal distribution. The highest point of the curve indicates the most common value from the variable. 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 : 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 : KDE

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

Sector: Construction.

Topic: New Dwelling Completions.

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 : 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 : 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. From the figure below, it can be seen 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 : Structure of the Dataset

* The describe code ‘.describe’ provides purely descriptive information about the dataset, which 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 : Structure of the Dataset

* Converting index to a datetime object for plotting and visualization.
* 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 14 below.

A screenshot of a computer

Description automatically generated with low confidence

Figure : 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 : New Dwelling Completions 2011-2022

Figure 14 examines the total new dwelling completions in Northern Ireland. It is noticeable that over the years there have been both increases and decreases in the number of completions; However, the increase trend in the number of completions has been higher. 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 : 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. 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 : 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: *.drop.*
* 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 : 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 : New Dwelling Completions 2011-2022

Figure 18 shows the total new dwelling completions in England and its fluctuations over the years. It is noticeable that from 2011 to 2012 there is a slight increase in the total new dwelling completions going from 114,030 to 115,590 respectively; However, in 2013, there is a decline in the number of completions to 109,450, thereafter there was 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 : 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. 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 : 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: *.drop.*
* 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 : 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 : 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 : Loading dataset and displaying the first five (5) rows of the dataset

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

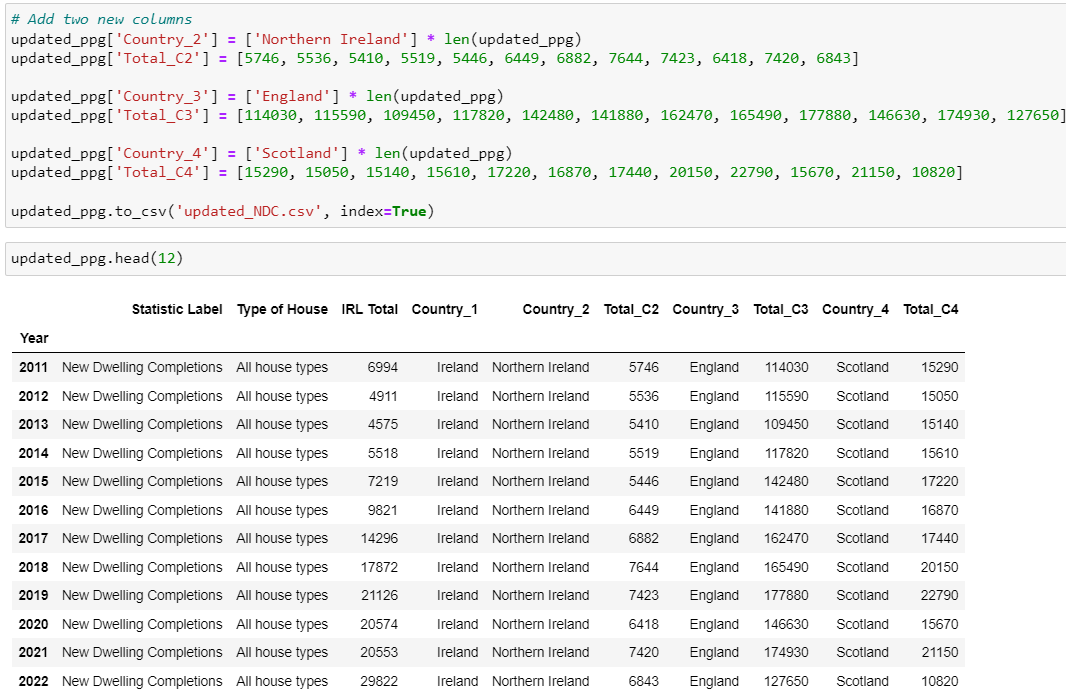


Figure : Adding new columns to the dataset

* 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. 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 : 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 : 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.

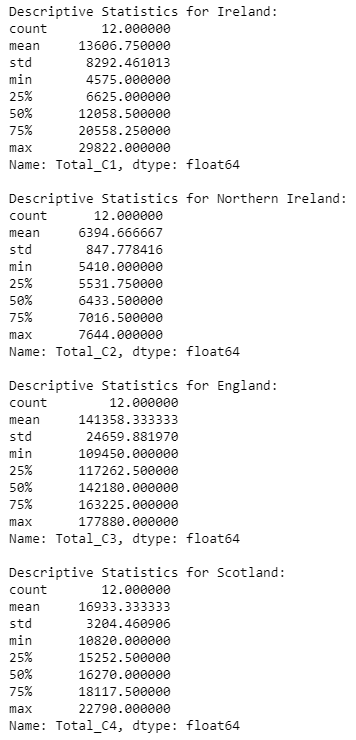


Figure : Descriptive statistics for each country

* Figure 29 below displays a boxplot that graphics the descriptive statistics of each country, namely Ireland, Northern Ireland, England, and Scotland. Looking at the graphic, we can be infer that England has both, the highest median and overall mean number of new dwelling completions, followed by Scotland and Ireland; However, Northern Ireland has the lowest rate of completions during the given period.

A picture containing text, screenshot, diagram, software

Description automatically generated

Figure : Box plot of descriptive statistics for each country

# Parametric and Non-parametric Analyses

## Two Sample t-test

Two-sample t-test is used to compare the values of different variables between three countries (Northern Ireland, England, and Scotland) and Ireland. The results can be seen in the Figure below.



Figure : Two sample t-test results

Based on the Figure above, here is a breakdown of the results:

1. Northern Ireland vs Ireland: The t-statistic is 3.00, indicating that there is a significant difference between the means of Northern Ireland in Ireland and Northern Ireland. The p-value is 0.0066, which is less than the commonly used significance level of 0.05. Therefore, the p-value suggests strong evidence to reject the null hypothesis and conclude that there is a significant difference between the means of Northern Ireland in Ireland and the other country.

2. England vs Ireland: The t-statistic is -17.01, indicating a significant difference between the means of England in Ireland and the other country. The p-value is 0.0000, which is less than 0.05. Therefore, there is strong evidence to reject the null hypothesis and conclude that there is a significant difference between the means of England in Ireland and the other country.

3. Scotland vs Ireland: The t-statistic is -1.30, suggesting a relatively smaller difference between the means of Scotland in Ireland and the other country compared to the previous comparisons. The p-value is 0.2083, which is greater than 0.05. Consequently, there is not enough evidence to reject the null hypothesis, indicating that the means of Scotland in Ireland and the other country are not significantly different.

In summary, based on the t-test results, there is a significant difference between the means of Northern Ireland and England in Ireland compared to the Scotland, while there is no significant difference in the means of Scotland.

## Wilcoxon Signed-rank Test

Wilcoxon signed-rank test to compare the distribution of values of the total number of new dwelling completions between three countries (Northern Ireland, England, and Scotland) and Ireland over the years. The results can be seen in the Figure below.

A screenshot of a computer program

Description automatically generated with medium confidence

Figure : Wilcoxon signed-rank test results

Here is a breakdown of the results shown in the figure above:

1. Ireland vs. Northern Ireland: The test statistic result for this comparison is 6.0, which suggests that there is a significant difference between the total number of new dwelling completions in Ireland and Northern Ireland. The p-value associated with this test is 0.0068, which means that the difference in the total number of new dwelling completions between the two countries is statistically significant. Therefore, the null hypothesis can be rejected.

2. Ireland vs. England: The test statistic for this comparison is 0.0, which suggests that there is no significant difference between the total number of new dwelling completions in Ireland and England. On the other hand, the p-value associated with this test is 0.00049. In this case the result suggests that there is a significant difference between the total number of new dwelling completions in the two countries, hence the null hypothesis can be rejected.

3. Ireland vs. Scotland: The test statistic for this comparison is 17.0, which suggests a larger difference between Ireland and Scotland. And The p-value associated with the test is 0.0922, which suggests that there is no significant difference between the total number of new dwelling completions in the two countries. Therefore, the null hypothesis cannot be rejected.

Based on the above, we can infer that there are significant differences in the total number of new dwelling completions between Ireland and Northern Ireland, as well as between Ireland and England. However, there is no significant difference observed in the total number of new dwelling completions between Ireland and Scotland.

## Chi-squared Test

The purpose of the chi-squared test in this context is to determine if there is a significant association or difference between the values of the three countries (Northern Ireland, England, and Scotland) and Ireland. The results can be seen in the Figure below.



Figure : Chi-squared test results

Based on the results, for each country compared to Ireland, the chi-squared statistic is 132.00, indicating a significant discrepancy between the observed and expected values. However, the p-value for each comparison is 0.2329, which is relatively high. This suggests that there is not enough evidence to reject the null hypothesis of independence between the variables. In other words, there may not be a significant association between the distribution of values in these countries and Ireland.

## Anova Test

The purpose if Anova test is to compare the means between three countries (Northern Ireland, England, and Scotland) and Ireland. The results can be seen in the Figure below.

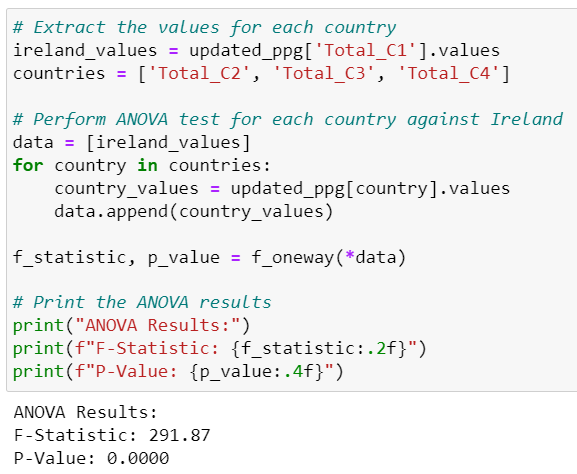


Figure : Anova test results

Based on the Figure above we can conclude that since the p-value is very low (rounded to 0.0000), it suggests that there is a significant difference between the means of the countries' values. In other words, Ireland values differ significantly across Northern Ireland, England, and Scotland. The high F-statistic value further supports this conclusion.

## Rank Test

This test is a non-parametric statistical test used to compare two independent groups when the dependent variable is ordinal or continuous. The results in the Figure below compare the rankings of three different pairs of groups: Northern Ireland vs. Ireland, England vs. Ireland, and Scotland vs. Ireland.



Figure : Rank test results

In summary, based on the rank test results show in the figure above, there is a significant difference between England and Ireland, with Ireland having higher ranks. However, no significant differences were found between Northern Ireland and Ireland or between Scotland and Ireland.

# 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 : MySQL database creation

A screenshot of a computer

Description automatically generated with medium confidence

Figure : MySQL data output setting index



Figure : Importing required libraries and entering codes

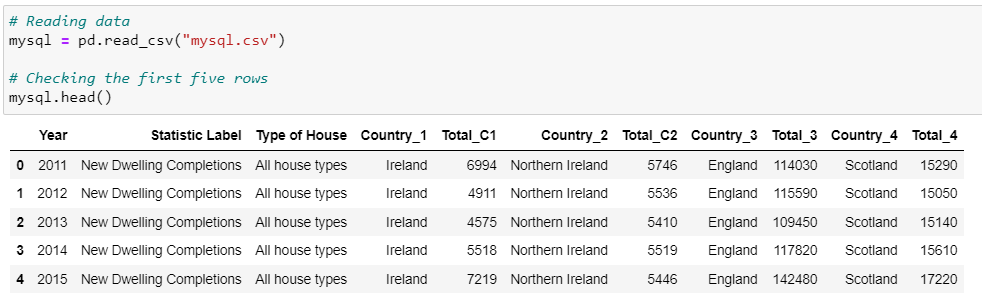


Figure : Checking the data imported

# Machine Learning

Since the task involves forecasting time series data, which is a supervised learning problem, we won't be using unsupervised or semi-supervised learning models for this particular scenario. The appropriate approach for forecasting is to use supervised learning models, such as ARIMA or other time series forecasting techniques.

To provide authenticity and evaluate the performance of the forecasting models, we can indeed use cross-validation and appropriate evaluation metrics. To start the process we initiate importing the required libraries, see figure below.

Importing required libraries

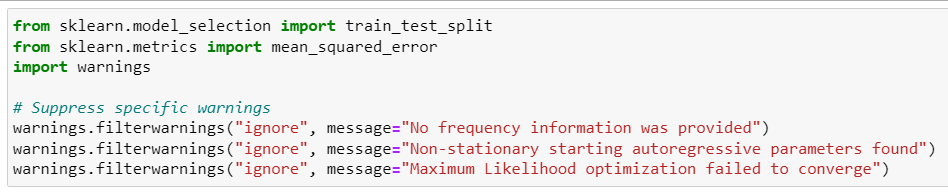


Figure : Importing required libraries

A screenshot of a computer program

Description automatically generated with medium confidence

Figure : ARIMA, testing model and forecasting

From the code in the figure above, we can infer that the data is split into training and testing sets using the train\_test\_split function from scikit-learn. The ARIMA models are then fitted to the training data, and forecasts are generated for the testing period. The Mean Squared Error (MSE) is calculated to evaluate the performance of the forecasts. Lower MSE values indicate better performance.

Interpreting the results:

Ireland: The MSE for Ireland is 11,300,217. This value represents the average squared difference between the predicted and actual values of new dwelling completions in Ireland. In this case, the MSE suggests that the model's predictions for Ireland have a relatively high level of error.

Northern Ireland: The MSE for Northern Ireland is 375,761.99. This value indicates a lower level of error compared to the Ireland MSE. The predictions for new dwelling completions in Northern Ireland are more accurate on average, as the MSE is smaller.

England: The MSE for England is 2,407,828,221.17. This value is significantly higher than the MSEs for both Ireland and Northern Ireland. The high MSE suggests that the model's predictions for England have a large amount of error and may not be accurate in capturing the patterns or trends in new dwelling completions for England.

Scotland: The MSE for Scotland is 96,927,909.99. This value is lower compared to the England MSE but higher than the Northern Ireland MSE. It indicates a moderate level of error in the model's predictions for new dwelling completions in Scotland.

In summary, the MSE values provide an indication of the accuracy of the ARIMA models' predictions for each region. A lower MSE implies better predictive performance. Therefore, based on the provided MSE values, the model's predictions for Northern Ireland are relatively accurate, while the predictions for Ireland, Scotland, and especially England have higher levels of error.

A screenshot of a computer program

Description automatically generated with medium confidence

Figure : Coding to plot and forecast values

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure : Forecast of new dwelling completions

## Linear Regression

A screenshot of a computer program

Description automatically generated with medium confidence

Figure : Linear regression code

Based on the results shown above, below is an interpretation of the results:

* Ireland Model Coefficient: This result: 2162.07 indicates that, on average, for each unit increase in the 'Year' variable, the 'Ireland' variable is expected to increase by approximately 2162.07. This coefficient represents the slope of the regression line.
* Ireland Model Intercept: The intercept for the Ireland model is approximately -4346214.31. It represents the predicted value of 'Total Ireland' when the 'Year' variable is zero. In this case, since the 'Year' variable likely represents years after a specific starting point, the intercept may not have a direct practical interpretation.
* Northern Ireland Model Coefficient: This result: 184.007, suggests that, on average, for each unit increase in the 'Year' variable, the ' NortIreland' variable is expected to increase by approximately 184.007.
* Northern Ireland Model Intercept: The intercept result is approximately -364655.434. Similar to the interpretation for the Ireland model, this value represents the predicted value of 'NortIreland' when the 'Year' variable is zero.
* England Model Coefficient: This result: 4664.545 implies that, on average, for each unit increase in the 'Year' variable, the 'England' variable is expected to increase by approximately 4664.545.
* England Model Intercept: The result is approximately -9264697.575, which represents the predicted value of 'England' when the 'Year' variable is zero.
* Scotland Model Coefficient: This result: 191.258, suggests that, on average, for each unit increase in the 'Year' variable, the 'Scotland' variable is expected to increase by approximately 191.258.
* Scotland Model Intercept: The intercept result of the model is approximately: 368739.918, which represents the predicted value of 'Scotland' when the 'Year' variable is zero.

## Sentiment Analysis on Construction Sector in Ireland

Sentiment analysis is based on understanding the subjects’ emotions from their text patterns to help in organising viewpoints into good or bad, positive or negative. This analysis helps firms by alerting them where customers are dissatisfied or seeking to shift to other products, allowing preventative actions to be taken.

To start the Sentiment Analysis on the construction sector in Ireland, we begin importing the required libraries.

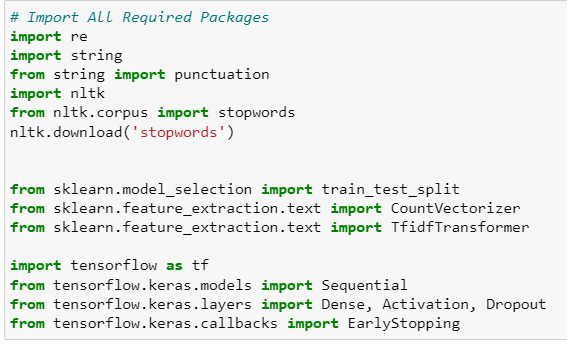


Figure : Importing required libraries for sentiment analysis

## Challenges

The main challenges encountered while creating this paper include the collection, integration, and storage of data with minimal requirements (hardware and software), finding datasets stored in the required formats, and finding datasets to perform sentiment analysis. When talking about data visualization, it can be said that one of the main goals of data visualization is to explain knowledge effectively through the use of diagrams; to easily transfer information to the user, knowledge hidden in large-scale and complex data sets becomes visible, therefore trying to achieve this goal is not as easy as using the right code, it deeply depends on quality of the data that is going to be displayed. Regarding the permissions, it was not possible to build the board due to lack of privileges since the machine that was used to carry out the project belongs to the company in which I am working.

# Conclusion

The construction industry is yet to reap the true beneﬁts of using data analytics aptly. Over the last two decades, the rapid growth of data analytics technologies has caused a spike in the number of models and platforms that have been developed for increasing digitalization across different ﬁelds. However, the same level of digitalization has not truly been harnessed or integrated by the construction industry. A critical overview of the existing literature points towards the bulk of existing resources and platforms that can easily be applied for construction management. However, the state of implantation of adoption in construction is below par. Therefore, the utilization and commercialization of data to beneﬁt the construction industry are crucial. The development of online tools and software which enable infrastructure modelling is a crucial step in the right direction for futuristic constructions. In this paper, we have discussed the existing tools used in data analytics, the use of statistics and ML. We also conducted a comparative analysis of new dwelling completions between four countries namely Ireland, Northern Ireland, England and Scotland. Leveraging the power of data analytics and utilizing Python and Jupyter Notebook, we processed, analysed, and visualized the dataset to extract valuable insights and understand the housing market dynamics across these regions.

Our analysis revealed several interesting findings. Firstly, we observed variations in the number of new dwelling completions between the countries under study over the mentioned period. While the total number of new dwelling completions in Ireland and England consistently showed higher numbers compared to the other regions, we identified fluctuations and trends specific to each country. These variations can be attributed to factors such as COVID-19, population growth, economic conditions, and housing policies. Furthermore, our analysis highlighted specific years where notable changes occurred in new dwelling completions. For instance, Ireland experienced a significant increase in new dwelling completions in 2019, while England showed a spike in 2018. These observations can indicate periods of increased construction activity and potentially reflect underlying market factors or policy interventions.

The visualizations, such as line graphs and box plots, allowed us to compare the values of different variables across the countries and understand the relative performance of each region. These visual representations facilitated the identification of patterns, trends, and potential outliers in the data. The outcomes of this comparative analysis can have valuable implications for various stakeholders. Policymakers can gain insights into the effectiveness of housing policies and identify areas that require attention or improvement. It is important to note that the analysis presented in this paper is based on the gathered datasets, and there may be additional factors and variables that could influence the topic. Future research could consider incorporating additional data sources and variables to gain a more comprehensive understanding of the topic and in each of the countries under study.

In conclusion, this project successfully utilized data analytics techniques, ML tools, Python programming language, and Jupyter Notebook to conduct a comparative analysis of new dwelling completions between Ireland, Northern Ireland, England and Scotland. The findings from this analysis provide valuable insights into the housing market dynamics, regional variations, and potential implications for policymakers, researchers, and industry stakeholders. By leveraging data analytics tools and methodologies, we can continue to explore and uncover valuable insights to support evidence-based decision-making in the housing sector.

# Acknowledgements

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