

**MSc in Data Analytics**

Continuous Assessment 02

2023

Author: Fiona Clarke

e-mail: sba22177@student.cct.ie

Student ID: sba22177

Word Count: 3295

GitHub Repository: *https://github.com/sba22177/MSc\_S1\_CA02\_SBA22177.git*

**Table of Contents**

[Abstract 3](#_Toc137828229)

[1. Introduction 4](#_Toc137828230)

[2. Business Understanding 4](#_Toc137828231)

[3. Data Understanding and Preparation 5](#_Toc137828232)

[3.1. Initial Exploratory Data Analysis 5](#_Toc137828233)

[3.1.1. Pandas vs Numpy 6](#_Toc137828234)

[3.2. Descriptive Statistics 7](#_Toc137828235)

[3.2.1. Normality 8](#_Toc137828236)

[3.3. Inferential Statistics 9](#_Toc137828237)

[3.3.1. Comparison Between EU Countries and Ireland 10](#_Toc137828238)

[4. WebScraping and Sentiment Analysis 13](#_Toc137828239)

[5. Modelling and Evaluation 15](#_Toc137828240)

[5.1. Data Preparation 15](#_Toc137828241)

[5.2. Random Forest Regression 15](#_Toc137828242)

[5.3. Time Series Analysis with ARIMA 17](#_Toc137828243)

[5.4. Comparison of Models 18](#_Toc137828244)

[6. Dashboard 20](#_Toc137828245)

[7. Testing and Optimization Strategy 21](#_Toc137828246)

[8. Conclusion 21](#_Toc137828247)

**Table of Figures**

[Figure 1: Screenshot of the interactive chart for Production Volume Index from 1990 to 2023 for EU countries 5](#_Toc137828248)

[Figure 2: Screenshot of the interactive chart for Production Volume Index from 2000 to 2023 for Ireland 6](#_Toc137828249)

[Figure 3: Histogram for Production Volume Index Data in Ireland 7](#_Toc137828250)

[Figure 4: Production Volume Index for Ireland Grouped by Year and Quarter 7](#_Toc137828251)

[Figure 5: Production Volume vs GDP for Ireland by Year and Quarter, each line represents the quarters, Q1, Q2, Q3 and Q4 8](#_Toc137828252)

[Figure 6: Spearman Correlations for Variables in Data 8](#_Toc137828253)

[Figure 7: Quantile-Quantile Plot for Production Volume Index 9](#_Toc137828254)

[Figure 8: Image of Summary Table for Normality Test Results 9](#_Toc137828255)

[Figure 9: Boxplot for Output Price Index in Construction 13](#_Toc137828256)

[Figure 10: Sentiment Analysis using TextBlob 14](#_Toc137828257)

[Figure 11: Sentiment Analysis using Vader 14](#_Toc137828258)

[Figure 12: Wordcloud for News Articles on Construction Industry in Ireland 15](#_Toc137828259)

[Figure 13 Actual vs Predicted Values for Production Volume Index using Random Forest Regressor with Hyperparameter Tuning and Feature Importance 16](#_Toc137828260)

[Figure 14 Actual vs Predicted Values for Production Volume Index using Random Forest Regressor tuned Model for Irish Data 16](#_Toc137828261)

[Figure 15 Production Volume Index decomposition using Chart Studio Plotly 17](#_Toc137828262)

[Figure 16: Actual vs Predicted Values for Production Volume Index using Time Series Forecasting 18](#_Toc137828263)

[Figure 17 Forecasting for the next 6 years for Production Volume Index in Ireland 18](#_Toc137828264)

[Figure 18: Screenshot of Jupyter Plotly Dash Dashboard 20](#_Toc137828265)

[Figure 19 22](#_Toc137828266)

[Figure 20 22](#_Toc137828267)

[Figure 21 22](#_Toc137828268)

[Figure 22 22](#_Toc137828269)

[Figure 23 23](#_Toc137828270)

[Figure 24 23](#_Toc137828271)

[Figure 25 23](#_Toc137828272)

[Figure 26 23](#_Toc137828273)

[Figure 27 24](#_Toc137828274)

[Figure 28 24](#_Toc137828275)

[Figure 29 24](#_Toc137828276)

[Figure 30 24](#_Toc137828277)

## Abstract

*This project review the Construction Industry data in Ireland and EU Countries through the EuroStat database. The main objective was to establish if prediction models could be used to forecast information with the domain. The data was analysed and explored with descriptive and inferential statistics. Two supervised machine learning models, a random forest regression and time series ARIMA, were executed to predict the ‘Production Volume Index’ for Ireland and EU countries. The random forest regression performed the best due to the multivariate nature. A sentiment analysis was performed on a news outlet in Ireland which provide input into the current views on the industry. A dashboard was created in order to summarise the data within the construction industry.*

# Introduction

The construction industry is a major contributor to the economic and social growth of a country. This project focuses on exploring data around the construction industry in Ireland and its comparison to EU countries. Index values for production volumes, building permits, labour inputs and costs were collected through the EuroStat database. The data was explored, analysed, and modelled to provide insight into the industry. Sentiment analysis was performed to understand the perception of the industry today. Forecasting of future trends for production volume index was performed to support decision making. Visual dashboard was created to effectively communicate the trends within the industry to key stakeholders.

Project management frameworks are important for monitoring and controlling a project from start to finish. For this project, the CRISP-DM project management framework was chosen to provide a structured approach which allows business objectives to be defined upfront and drive the direction of the project. In addition, the framework provide flexibility to revise the approach at different stages if required. (Wirth and Hipp, 2000; Schröer, Kruse and Gómez, 2021)

# Business Understanding

The objective of this project is to collect, process, analyse and interpret data related to the construction industry in Ireland and compare to other countries. After review of online databases, index values on the construction industry were chosen from Eurostat. The index values are generated quarterly for EU countries and relate to production volumes, building permits, labour inputs and construction cost data.(Eurostat, no date a, no date d, no date f, no date b)

The process of acquiring the data began with a general search on the industry in Ireland to get an understanding data analysis performed. For Ireland, the main resources were the Irish Government or Central Statistics Office, which was a challenge when sourcing similar data from other countries due to differences in collation and formatting. Data analyses within the industry has become popular for risk management and cost analysis. As a result, many of the databases required subscriptions or licences if non-government owned e.g CIS Ireland. This reduced the amount of data available. Eurostat database was chosen as the primary database as it has comparable data for EU countries. The databased is managed to ensure compliance, comparability, and accuracy. It is also readily available for commercial and non-commercial purposes, and doesn’t require special licencing or permission to access.(EuroStat, no date b, no date a)

# Data Understanding and Preparation

The following datasets were chosen as they formed the main indexes available for the industry in EU: Production in Construction, Building Permits, Construction Producer Prices and Labour Input. The quarterly dataset was chosen as the monthly dataset had no data for Ireland and the quarterly provides more data than the annual dataset. During initial EDA, house price index and the gross domestic product (GDP) datasets where added to enhance the analysis.(Eurostat, no date d, no date a, no date b, no date c, no date e, no date f)

## Initial Exploratory Data Analysis

The csv files were read in to notebook, CA02\_sba22177\_Section\_01 and reviewed using .head(), .shape and .info() to understand the structure of the data. A function was created to generate the unique values in each column to see the categories available. Any columns that had only one category were removed including the dataflow (database name), last update date and frequency. The datasets were filtered for index ‘I15’ for consistency as it's the common index between datasets. Where required, the dataframes were pivoted on ‘year-quarter’ and ‘country’ to get the categories as columns. Column names were renamed to align and datasets were merged using pd.merge() with an outer join. The outer join was chosen to retain all the data as there may be categories for a country and time-period not available. (McQuaid, 2023c, 2023a)

Production volume index was graphed with an interactive plotly chart to visualise the data per country Figure 1. Plotly provides an easy selection method for country without requiring new dataframes. The main issue observed was that data pre-2000 was limited and not available for Ireland, therefore decision was made to remove. The main pattern observed was a seasonal up and down as a result of the quarterly values. Plotly graphs for each parameter were output to png files to review the data trends, to identify patterns and to visualise obvious outliers, Appendix. (Plotly, no date e, no date d)

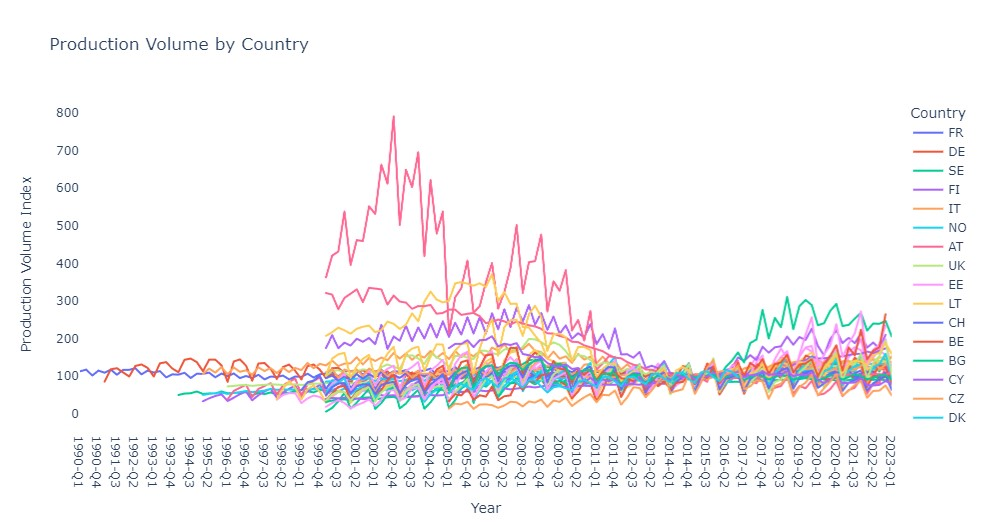


Figure : Screenshot of the interactive chart for Production Volume Index from 1990 to 2023 for EU countries

A picture containing text, screenshot, plot, line

Description automatically generated

Figure : Screenshot of the interactive chart for Production Volume Index from 2000 to 2023 for Ireland

### Pandas vs Numpy

For EDA of the CSV files, pandas was the primary library of choice. numpy library is another option for EDA on CSV files. A comparison of pandas and numpy was performed as per Table 1. While numpy outperforms pandas for speed, pandas was the choosen library for processing and aggregating due to its flexibility for data access, wrangling and structure. (GeeksforGeeks, no date a; Desktop, 2022)

Table Comparison of Pandas vs Numpy Libraries for Data Processing and Aggregating

|  |  |
| --- | --- |
| **Pandas** | **Numpy** |
| Works with Dataframes and Series | Works with Arrays and multi-dimensional Arrays |
| Consumes more memory but better with data >500K rows | Consumes less memory and better with <50K rows |
| Mainly used for tabular data | Mainly used with numerical data |
| Access data at index positions and labels | Access data at the index positions |

## Descriptive Statistics

The production volume index data for Ireland distribution was reviewed through a histogram, Figure 3. From this visualization, it was clear the data was a multimodal distribution with the three separate peaks.

A picture containing symbol, rectangle, graphics, screenshot

Description automatically generated

Figure : Histogram for Production Volume Index Data in Ireland

A bar chart was generated to show the index for each year and quarter, Figure 4. From this, there is a clear pattern relating to the economy. From 2000 to 2007, the high the country had high economic growth known as ‘Celtic Tiger’. In 2008, Ireland entered a recession, where a sudden decrease in the index is observed. In 2013, post-recession the economy improved relating to the gradual increase in the index value. (Whelan, 2014)

A picture containing screenshot, silhouette

Description automatically generated

Figure : Production Volume Index for Ireland Grouped by Year and Quarter

Gross Domestic Product (GDP) provides insight into the economy of a country. The relationship between production volume index and GDP was graphed to visualise the relationship, Figure 5. The data has the same pattern for highs and lows, however it does not appear to correlate fully.

A picture containing child art, sketch, art

Description automatically generated

Figure : Production Volume vs GDP for Ireland by Year and Quarter, each line represents the quarters, Q1, Q2, Q3 and Q4

Boxplots were created for each variable in the data and grouped by country in Jupyter Notebook Section 2.2.1.2. This was done to visualize the descriptive stats for mean, median and the quantiles and any outliers. However as most parameters don’t exhibit the normal distribution curves, section 3.2.1, the statistics for mean are non-value added. (Plotly, no date d, no date b; Weiss, 2017)

Correlations were also performed for the variables to identify any relationships that could be important for machine learning. Utilising pandas corr(), spearman method was chosen due to the non-normal data, Figure 6. (Pandas, no date) The seismic colour gradient was chosen as the contrast of the blue and red makes it easy to see strong vs weak correlations.

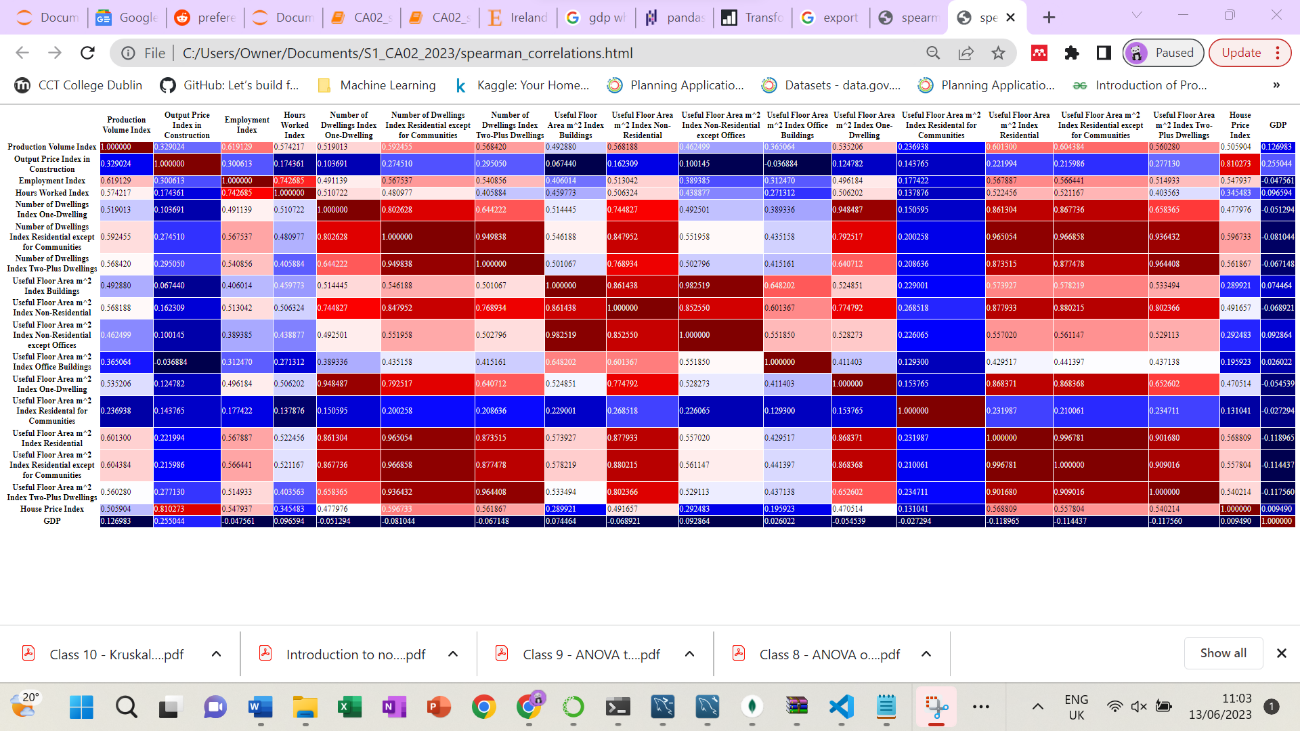


Figure : Spearman Correlations for Variables in Data

### Normality

To understand the shape of the data, a histogram was generated for each variable for all the data, Notebook Section 2.2.1.1. Except for three parameter, the variables don’t exhibit a normal distribution pattern. To examine this further, quantile-quantile (Q-Q) plots were generated which supported observation of non-normal, Figure 7. A summary table was created for Shapiro Wilk and Anderson Darling test results, Figure 8: Image of Summary Table for Normality Test Results. For assessing normality, it’s best practice to execute different test methods for a full overview of data so histograms, Q-Q plots and statistical tests were performed. Based on these, it can be concluded the data is non-normal. (Weiss, 2017; Mishra, 2020; GeeksforGeeks, 2022)

A picture containing screenshot, line, plot, rectangle

Description automatically generated

Figure : Quantile-Quantile Plot for Production Volume Index

A picture containing text, menu, document, number

Description automatically generated

Figure : Image of Summary Table for Normality Test Results

Normal distribution while not required, it’s favourable for some machine learning techniques and inferential statistics. The data was transformed using the three techniques: boxcox, square root and log. After each transformation, histograms, Q-Q plots and summary table of statistical test were generated to test for normality. While the majority of the histograms formed a normal distribution curve, the Q-Q plots and statistical tests concluded that the transformations for normality were unsuccessful. (Weiss, 2017; Dhingra, 2021)

## Inferential Statistics

Inferential statistics were performed on populations of the Production Volume Index in notebook section 2.3.1. As the data is non-normal, the non-parametric Kruskal Wallis test was chosen to determine if statistically significant difference between the medians of three or more independent groups. The results are summarised in Table 2. (GeeksforGeeks, no date d; SciPy, no date a; Zach, 2020a).

Table : Assumptions, Hypothesis and Result for Kruskal-Wallis Test on Production Volume Index Data

|  |  |  |
| --- | --- | --- |
| **Assumptions** | * independent variables with one or more groups. * values are ranked. * Doesn't require a specific distribution. | |
| **Hypothesis** | Hypothesis:   * H0: The medians for Production Volume Index for each year are equal * Ha: The medians for Production Volume Index for each year are not equal   Significance Level is 0.05 | Hypothesis:   * H0: The medians for Production Volume Index for last three years are equal * Ha: The medians for Production Volume Index for last three years are not equal   Significance Level is 0.05 |
| **Result** | * Statistic Value: 19.53730584536599 * p-value: 0.0006161540674819821 * Reject the Null Hypothesis, the medians for Production Volume Index every five years are not equal | * Statistic Value: 3.0120958207326103 * p-value: 0.22178476146606121 * Fail to Reject the Null Hypothesis, the medians for Production Volume Index for last three years are equal |

### Comparison Between EU Countries and Ireland

An analysis was performed by the Housing Agency Ireland on the construction costs between Ireland, UK, Germany, France and the Netherlands in 2018 due to similarities in climate, labour and economy. They concluded that Ireland, UK, Germany and France were comparable for construction costs. (The Housing Agency, 2018) Therefore, these countries will be assessed for similarities. The data was subset and normality was tested as discussed in section 3.2.1 and concluded non-normal. Therefore, non-parametric tests were chosen as these don’t require a desired distribution. The tests and results are summarised below. Each test has a set of assumptions that must be met for the test to be applied to the data. The applicability of the data to these tests were confirmed prior to selecting the variables. (Weiss, 2017)

The Mann-Whitney U Test, which compares the medians of two samples, is summarised in Table 3.(SciPy, no date b; Weiss, 2017; Pawangfg, 2020)

Table : Assumptions, Hypothesis and Result for Mann-Whitney U Test

|  |  |
| --- | --- |
| **Assumptions** | * observations are independent. * values are in an ordinal manner. * sample size of <20 recommended. * data does not need to be normally distributed but must be of similar shape. |
| **Hypothesis** | Hypothesis:   * H0: There is no significant difference between Ireland and UK for Employment Index * Ha: There is a significant difference between Ireland and UK for Employment Index   Significance Level is 0.05 |
| **Result** | * Statistic Value: 306.5 * p-value: 0.004138021115499273 * Reject the Null Hypothesis, there is a significent difference between Ireland and UK for Employment Index |

The Wilcoxon Signed-Rank Test, which compares the medians to determine if sample are identically distributed, are summarised in Table 4. (GeeksforGeeks, no date c; Python for Data Science, no date; SciPy, no date c; Weiss, 2017)

Table : Assumptions, Hypothesis and Result for Wilcoxon Signed Rank Test

|  |  |  |
| --- | --- | --- |
| **Assumptions** | * population distribution is symmetric. * randomly sampled. * observations are from the same population | |
| **Hypothesis** | Hypothesis:   * H0: Sample Distribution for Ireland/Germany Employment Index Before and After the Recession are Equal. * Ha: Sample Distribution for Ireland/Germany Employment Index Before and After the Recession are not Equal.   Significance Level is 0.05 | |
| **Result** | * Statistic Value: 8.0 * p-value: 4.76837158203125e-05 * Reject the Null Hypothesis, Sample Distribution for Ireland Employment Index Before and After the Recession are not Equal | * Statistic Value: 100.0 * p-value: 0.8694877624511719 * Fail to Reject the Null Hypothesis, Sample Distribution for German Employment Index Before and After the Recession are Equal |

The Kruskal-Wallis Test, determines if there’s a statistically significant difference between the medians of three or more independent groups, is summarised in Table 5.(GeeksforGeeks, no date d; SciPy, no date a; Zach, 2020a).

Table : Assumptions, Hypothesis and Result for Kruskal-Wallis Test

|  |  |
| --- | --- |
| **Assumptions** | * independent variables with one or more groups. * values are ranked. * Doesn't require a specific distribution. |
| **Hypothesis** | Hypothesis:   * H0: The medians of Production Volume Index for Ireland, Germany, France and the UK are equal * Ha: The medians of Production Volume Index for Ireland, Germany, France and the UK are not equal   Significance Level is 0.05 |
| **Result** | * Statistic Value: 79.72705061990443 * p-value: 3.512223437669616e-17 * Reject the Null Hypothesis, the medians of Production Volume Index for Ireland, Germany, France and the UK are not equal |

The Friedman tests determines if there is a statistically significant difference between the medians of three or more independent groups, is summarised in Table 6. (GeeksforGeeks, no date b; Hessing, no date; Zach, 2020b)

Table : Assumptions, Hypothesis and Result for Friedman Test

|  |  |
| --- | --- |
| **Assumptions** | * random sampling * one variable that is measured across at least three groups. * does not require normal distribution. * samples should be ordinal or continuous |
| **Hypothesis** | Hypothesis:   * H0: the means of Production Volume Index for Ireland, Germany, France and the UK are equal. * Ha: at least one mean of Production Volume Index for Ireland, Germany, France and the UK are not equal   Significance Level is 0.05 |
| **Result** | * Statistic Value: 60.888000000000034 * p-value: 3.797601950713565e-13 * Reject the Null Hypothesis, at least one mean of Production Volume Index for Ireland, Germany, France and the UK are not equal |

The spearman rank correlation, which measures the correlation between two ranked variables, is summarised in Table 7.

Table : Assumptions, Hypothesis and Result for Spearman Rank Correlation

|  |  |
| --- | --- |
| **Assumptions** | * no distribution required. * values are ranked |
| **Hypothesis** | Hypothesis:   * H0: there is a no significant correlation between Production Volume Index in Ireland and the UK * Ha: there is a significant correlation between Production Volume Index in Ireland and the UK   Significance Level is 0.05 |
| **Result** | * Rho: -0.013158210679231952 * p-value: 0.9277364114153783 * Fail to Reject the Null Hypothesis, there is not a significant correlation between Production Volume Index in Ireland and the UK |

These tests did not give back any similarities, so a function was defined for Kruskal Wallis to pass all parameters through. The variable ‘Output Price Index in Construction’ had a p-value > 0.05 so we fail to reject the null hypothesis that the medians for each country are equal. The alternative hypothesis was the medians for each country are not equal. A boxplot was created to visualise the similarities, Figure 9. The output price index measures the change in prices for the construction of buildings. (EuroStat, no date b) For research purposes, the question posed is why the construction cost indexes are similar across the four countries. From news articles, Ireland states that their cost are higher than EU, but UK also say their costs are higher than EU countries. (McQuinn, 2023; O’Carroll, 2023) As the variable is an index it’s measuring the change in costs each year, therefore where countries have similar economies and see similar fluctuations like Ireland, France, UK and Germany this may give rise to the similarities seen in the output price index.(The Housing Agency, 2018) It can be difficult to pinpoint exact reasons as there are many influencing factors including labour and material costs etc. There are resources for construction costs however limitations exist due to different interpretations, different methods of analysis or collation of data. One would need to get the input data used for calculating the output price index to really understand the similarities. (Arigoni, Kennedy and Killeen, no date)

A picture containing diagram, text, screenshot, rectangle

Description automatically generated

Figure : Boxplot for Output Price Index in Construction

# WebScraping and Sentiment Analysis

For the second format of data, webscraping was performed on the Independent website for news articles relating to the construction industry.

The ‘BeautifulSoup’ package was used to perform webscraping, notebook CA02\_sba22177\_Section\_03\_Section\_05. BeautifulSoup is primarily a data parsing tool that is well-known with many resources making it easy to learn. Another option for web scraping in Python is the Scrapy library, see Table 8 for comparison.(Karatas, 2023)

Table : Comparison of BeautifulSoup and Scrapy for Webscraping

|  |  |
| --- | --- |
| **BeautifulSoup** | **Scrapy** |
| Python parsin library | Open-Source framework |
| Requires additional libraries to work | Complicate installation |
| Fetches content of webpages | Extract data from APIs and Webscrapes |
| Required Python knowledge | Doesn’t required python knowledge |

The URL for the Independent website for construction topic was retrieved. The html.parser was used to extract the HTML CSS data out for the headlines. On the website, the ‘inspect’ window was utilised to find the elements and classes. The links of each article were retrieved and used to extract data from the articles. The final output had headline titles and paragraph contents.(Zafra, 2019) Limitations of this method was that the Independent required subscription to fully access the articles, so only partial scraping of paragraphs was performed. In addition, with beautifulsoup, only front page information scraped and would required scrolling through Selenium package to return more data. (Shivaji, 2021)

Sentiment analysis was performed on web scraped news articles. Polarity and subjectivity are two methods of determining the underlying emotion of a text. Polarity measures the strength of opinions whereas subjectivity measures the personal influence of a topic. The TextBlob package was used first to generate subjectivity and polarity results. The results were graphed, Figure 10, however the majority sentiment was positive which was unexpected. (Acharya, 2022; Omoniyi, 2022)

A picture containing screenshot, rectangle, graphics, square

Description automatically generated

Figure : Sentiment Analysis using TextBlob

A second analysis was performed with the Vader package. For sentiment analysis, multiple packages should be used and averaged due to different accuracy levels. Vader is more accurate for identifying negative sentiments so this was chosen.(Barai, 2021) Using the SentimentIntensityAnalyzer the compound score was extracted and graphed, Figure 11. The sentiment has been reversed compared to TextBlob. (Acharya, 2022; Selvaraj, 2022)

A picture containing screenshot, rectangle, graphics, square

Description automatically generated

Figure : Sentiment Analysis using Vader

A wordcloud was created to highlight the most used words in the news. Wordclouds can be strong data visualization tools to convey the sentiment for a domain. Data prep included removing unrequied columns, data tokenizing, stopword removal, character removal Max words was set to 30 so that it wasn’t overcrowded and a white background was chosen with colormap viridis as the colours had the best contrast so easier to read, Figure 12. (Acharya, 2022)

A close-up of words

Description automatically generated with low confidence

Figure : Wordcloud for News Articles on Construction Industry in Ireland

# Modelling and Evaluation

For this project, the question proposed was can we predict the production volume index for Ireland? Prediction and time series models were evaluated for this. As this data is labelled, supervised learning models were chosen.

## Data Preparation

The main data preparation required was the handling of missing data. Variables where there was too much data missing like House Price Index were removed from dataset as it wasn’t logical to fill missing data with >50% unavailable. For the remaining missing data, it was considered to fill with means/medians however the data distribution isn’t normal so was ruled out. Other options include interpolation, but the data gaps were clustered, or to predict the values with prediction modelling. Due to time constraints, the missing rows of data were dropped as impact on size of data was minimal. (Kumar, 2020)

For regression models, the categorical labels were transformed using one-hot encoding. This was done instead of removing variables as they could be important to the model. (Koehrsen, 2017; McQuaid, 2023b)

## Random Forest Regression

Once data was prepared for Random Forest, the features and labels were defined and data was split. An average value of the target was calculated to generate a baseline error. This was used to gauge the performance of the model later. If the model could not improve on the baseline error, then the model choice would be reconsidered following CRISP-DM structure. Mean squared error (MSE), mean absolute error (MAE), Mean Absolute Percentage Error and Root Mean Squared Error (RMSE) values were retired for the model without tuning. (Koehrsen, 2017)

GridSearchCV was utilised to determine the optimal parameters for the model. The cross validation score chosen was the neg\_mean\_absolute\_percentage\_error which could be used to compare to the model without tuning.(Koehrsen, 2017, 2018) Feature importance was performed for dimensionality reduction to understand the impact of each variable on the model. Features with an importance score of 0 were removed and the model was executed to determine the impact. A subset of the data for Ireland was prepped to be ran through the model, one-hot encoding and feature importance was applied.

The Production Volume Index was ran through the models and cross validation result, MAPE, was used to measure the performance.. The MAPE takes the difference of the actual and predicted value and divides by the actual. The absolute percentage is then applied. (M, 2021) The MAPE for models were summarized in Table 9. C performed the worst and A, with no optimization, performed better, indicating that the standard settings were suitable. B performed better than C therefore removing some features did impact the performance slightly. D performed the best predicting the Irish data, however this may be a result of the reduced quantity of data, Figure 13 and Figure 14.

Table : Performance of the three Random Forest Regression Models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A | B | C | D |
| **Baseline Model** | **Hyperparameter Tuned Model** | **Feature Importance and Hyperparameter Tuned Model** | **Model C with Irish Data** |
| Mean Absolute Percentage Error % | 10.63 | 11.91 | 12.95 | 9.10 |
| 11.97 | 13.25 | 11.34 |
| 13.16 | 13.42 | 9.71 |

A picture containing text, screenshot, plot, line

Description automatically generated

Figure Actual vs Predicted Values for Production Volume Index using Random Forest Regressor with Hyperparameter Tuning and Feature Importance

A picture containing text, screenshot, plot, line

Description automatically generated

Figure Actual vs Predicted Values for Production Volume Index using Random Forest Regressor tuned Model for Irish Data

## Time Series Analysis with ARIMA

As the index is measure over time, time series analysis is suitable for prediction/forecasting. A local univariate model was required so Autoregressive Integrated Moving Average (ARIMA) model was chosen. Prophet was also considered but can be considered less accurate.(Hariharan, 2020)

The data was prepped ensuring that a datetime dtype was the index. Decomposition was performed on the irish dataset using statsmodel to view the trend, seasonal and residual patterns of the data. A multiplicative model was chosen as data isn’t linear. The trend is the moving average, the seasonal component is the period average of the detrended data and the residual is the data when trend and seasonal components are removed. From this analysis, it can be concluded that the Irish dataset has a seasonal component. (Portilla, 2018; Lewinson, 2022)

A picture containing text, screenshot, plot, line

Description automatically generated

Figure Production Volume Index decomposition using Chart Studio Plotly

The data was split for train and test by taking 80% of the past data for training and 20% for testing. To determine the best hyperparameters for the model, the pmdarima auto\_arima model was utilized. Max and min values can be selected for p,q,d and P,Q,D and the model determines the best settings based on the quality indicator, Akaike information criterion (AIC). Models that fit the data with the fewest features have lower AIC scores which are favourable. The seasonal setting was applied and m was set to 4 to for quarters. The model was then trained and test data was used to generate prediction values. The test and prediction data were graphed to visualise the performance of the model. The model initially performed well for 2018 and 2019, however it deviates from 2020 onwards, Figure 16. This is potentially due to the impact of external factors, like the Covid-19 Pandemic. As a univariate model, these external factors are not taken into account. (Portilla, 2018)

A picture containing text, line, plot, diagram

Description automatically generated

Figure : Actual vs Predicted Values for Production Volume Index using Time Series Forecasting

Forecasting of the next 6 years was completed using the ARIMA model, Figure 17.

A picture containing text, screenshot, plot, line

Description automatically generated

Figure Forecasting for the next 6 years for Production Volume Index in Ireland

## Comparison of Models

The random forest regression and ARIMA models were established to predict the Production Volume Index for Ireland. The MSE, MAE and RMSE are performance measures for comparison. The MSE looks at the difference between the actual and predicted value and squares it, increased errors gives higher value. The RMSE takes the square root of the MSE, a value of 0 is a perfect model. The MAE is a simpler function that takes the difference between the actual and predicted values, a low result is desirable. (M, 2021)

The results are summarized in Table 10. The Random Forest is the best performing model with considerably lower values. The random forest used a multivariate approach with data all across Europe in order to understand the Production Volume Index where as the Time Series was univariate with only irish data. The difference in these input datas has had a big effect on the performance of the data. In order to optimize the time series, it would be recommended to utilize a multivariate approach in the future.

Table : Performance Metrics for the Random Forest and Time Series Models for Predicting Production Volume Index in Ireland

|  |  |  |  |
| --- | --- | --- | --- |
|  | Mean Squared Error | Mean Absolute Error | Root Mean Square Error |
| Random Forrest Regression | 128.67 | 9.04 | 11.34 |
| Time Series Analysis ARIMA | 863.78 | 21.82 | 29.39 |

# Dashboard

Dashboard was created in notebook CA02\_sba22177\_Section\_03\_Section\_05 under Section 5. Dashboard was created using JupyterDash application applying Tuftes Principles.(International Service Design Institute, no date) Jupyterdash was chosen as it s open source, works in the Jupyter notebook and as plotly was already utilised in the project it was continued into here. The dashboard contains four pieces of data. There is an interactive line plot for viewing the parameters by country. This was chosen as it provides good overview of all data. An interactive heatmap to view spearman correlations by country. This was included to give insight into the relationship of the parameters. Redblue colours were chosen to give a clear visual on what correlates and what does not. A snapshot of the top 10 headlines from webscraping to provide insight into the sentiment of the domain. The chart of the forecasting by ARIMA for Irish data to showcase prediction modelling. With more time, it would be beneficial to have a prediction modelling chart that changes based on country input.(Plotly, no date f, no date c, no date a, no date b; Aichara, 2019; Muthukrishnan, 2019)

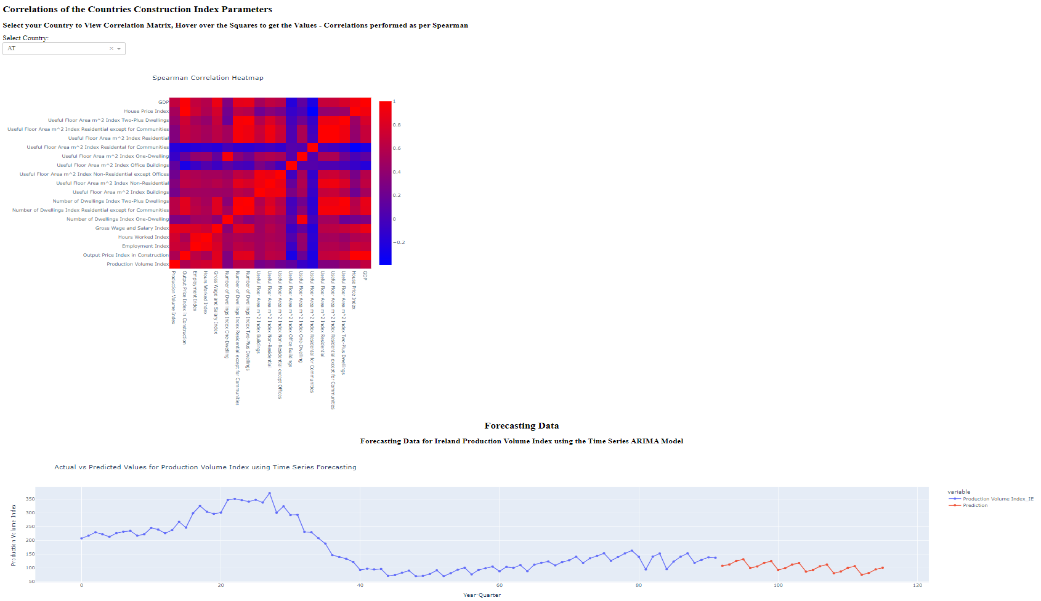
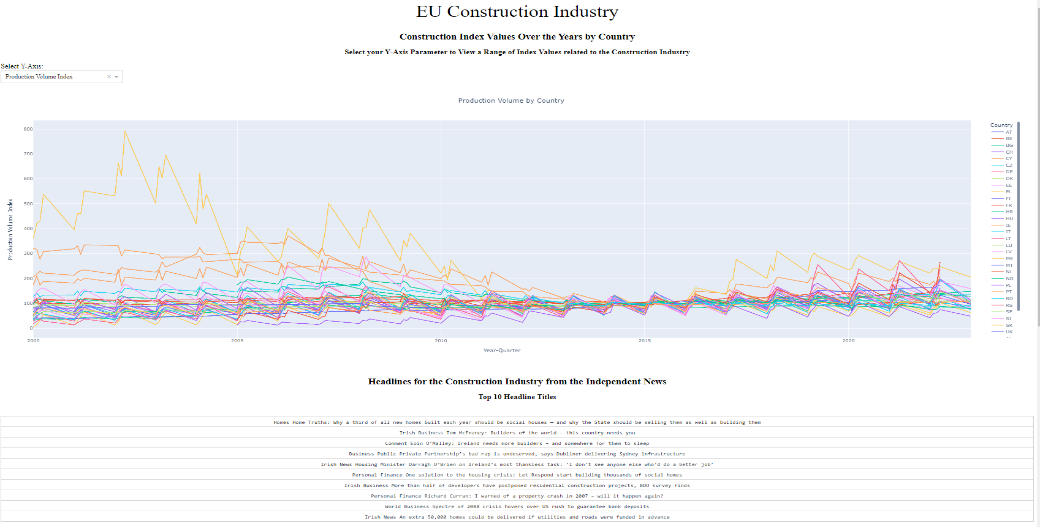


Figure : Screenshot of Jupyter Plotly Dash Dashboard

# Testing and Optimization Strategy

For this project, all code was manually tested using one of info(), head(), shape(), unique() to ensure changes executed correctly. Where required, graphs were utilsed to visualize changes. If the project was deployed in a production application automated testing such as unittest would be appropriate.(Shaw, no date)

For optimization, the dataframe memory usage was tracked at the end of each notebook. While updating dataframes, where possible, new dataframes were not created to reduce the stored memory. (Cmdlinetips, 2020) Functions were created if code was being ran multiple times to reduce the lines of code. In addition, profiling with the profile module in the notebook would be beneficial to understand the performance of the code.(Perrier, 2021; Tam, 2022) Trade offs between legibility, maintenance and speed of your code is required for optimal project work.(Orac, no date)

# Conclusion

The project goal was to analyse, evaluate and model data related to Construction Industry for Ireland and other EU countries. The data was explored and analysed statistically to provide insight into variables prior to modelling. A random forest regression and time series analysis was performed, where the random forest performed the best. This was due to the multivariate nature of the model. It was recommended that a multivariate time series be completed at a future state. Webscraping was performed on a news outlet and sentiment analysis was performed which demonstrated negative sentiments. The outputs of the data exploration and modelling were summarised in a plotly dashboard. From the analysis performed, it can be concluded the Construction Industry is ever changing, year to year and also country to country. There are many dependent factors for the Production Volume Index of a country including building permits available and economic factors that influence labour.

**Appendix**

A picture containing text, screenshot, diagram, line

Description automatically generated

Figure

A picture containing text, diagram, screenshot, plot

Description automatically generated

Figure

A picture containing text, screenshot, font, design

Description automatically generated

Figure

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure

A picture containing text, screenshot, font, design

Description automatically generated

Figure

A picture containing text, screenshot, diagram, design

Description automatically generated

Figure

A picture containing text, screenshot, design

Description automatically generated

Figure

A picture containing text, screenshot, design

Description automatically generated

Figure

A picture containing text, screenshot, diagram, design

Description automatically generated

Figure

A picture containing text, screenshot, design

Description automatically generated

Figure

A picture containing text, screenshot, line, diagram

Description automatically generated

Figure

**References**

Acharya, S. (2022) *Text Analytics 101 — Word Cloud and Sentiment Analysis*. Available at: https://towardsdatascience.com/text-analytics-101-word-cloud-and-sentiment-analysis-2c3ade81c7e8 (Accessed: 9 June 2023).

Aichara, D. C. (2019) *Building a Dashboard App using Plotly’s Dash: A Complete Guide from Beginner to Pro | by Dayal Chand Aichara | Analytics Vidhya | Medium*, *Medium*. Available at: https://medium.com/analytics-vidhya/building-a-dashboard-app-using-plotlys-dash-a-complete-guide-from-beginner-to-pro-e7657a4eb707 (Accessed: 16 June 2023).

Arigoni, F., Kennedy, G. and Killeen, N. (no date) ‘No.12 Rising construction costs and the residential real estate market in Ireland (Arigoni, Kennedy and Killeen)’, 2022(12).

Barai, M. K. (2021) *Sentiment Analysis with Textblob and Vader in Python*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/10/sentiment-analysis-with-textblob-and-vader/ (Accessed: 10 June 2023).

Cmdlinetips (2020) *How To Get The Memory Usage of Pandas Dataframe? - Python and R Tips*, *Python and R Tips*. Available at: https://cmdlinetips.com/2020/03/memory-usage-of-pandas-dataframe/ (Accessed: 16 June 2023).

Desktop, N. (2022) *Pandas vs. NumPy: Which Python Library is Better for Data Analytics?*, *Noble Desktop*. Available at: https://www.nobledesktop.com/classes-near-me/blog/pandas-vs-numpy-for-data-analytics.

Dhingra, D. (2021) *Gaussian Distribution Transform| Transformation of Normal Distribution*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/05/how-to-transform-features-into-normal-gaussian-distribution/ (Accessed: 13 June 2023).

Eurostat (no date a) *Statistics Building permits - quarterly data*, *Eurostat*. Available at: https://ec.europa.eu/eurostat/databrowser/view/STS\_COBP\_Q/default/table?lang=en (Accessed: 12 June 2023).

Eurostat (no date b) *Statistics Construction producer prices or costs, new residential buildings - quarterly data*, *Eurostat*. Available at: https://ec.europa.eu/eurostat/databrowser/view/STS\_COPI\_Q/default/table?lang=en (Accessed: 12 June 2023).

Eurostat (no date c) *Statistics House price index (2015 = 100) - quarterly data*. Available at: https://ec.europa.eu/eurostat/databrowser/view/EI\_HPPI\_Q/default/table?lang=en (Accessed: 12 June 2023).

Eurostat (no date d) *Statistics Labour input in construction - quarterly data*, *Eurostat*. Available at: https://ec.europa.eu/eurostat/databrowser/view/STS\_COLB\_Q/default/table?lang=en (Accessed: 12 June 2023).

Eurostat (no date e) *Statistics Main GDP aggregates per capita*, *Eurostat*. Available at: https://ec.europa.eu/eurostat/databrowser/view/NAMQ\_10\_PC/default/table?lang=en (Accessed: 12 June 2023).

Eurostat (no date f) *Statistics Production in construction - quarterly data*, *Eurostat*. Available at: https://ec.europa.eu/eurostat/databrowser/view/STS\_COPR\_Q/default/table?lang=en (Accessed: 12 June 2023).

EuroStat (no date a) *Copyright notice and free re-use of data - Eurostat*, *EuroStat*. Available at: https://ec.europa.eu/eurostat/about-us/policies/copyright (Accessed: 9 June 2023).

EuroStat (no date b) *Short-term business statistics (sts)*, *EuroStat*. Available at: https://ec.europa.eu/eurostat/cache/metadata/en/sts\_esms.htm#rel\_policy1677061311663 (Accessed: 9 June 2023).

GeeksforGeeks (2022) *How to Perform a Shapiro-Wilk Test in Python - GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/how-to-perform-a-shapiro-wilk-test-in-python/ (Accessed: 15 April 2023).

GeeksforGeeks (no date a) *Difference between Pandas vs Numpy*, *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/difference-between-pandas-vs-numpy/.

GeeksforGeeks (no date b) *Friedman Test*, *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/friedman-test/ (Accessed: 13 June 2023).

GeeksforGeeks (no date c) *How to Conduct a Wilcoxon Signed-Rank Test in Python? - GeeksforGeeks*, *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/how-to-conduct-a-wilcoxon-signed-rank-test-in-python/ (Accessed: 13 June 2023).

GeeksforGeeks (no date d) *How to Perform a Kruskal-Wallis Test in Python*, *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/how-to-perform-a-kruskal-wallis-test-in-python/ (Accessed: 13 June 2023).

Hariharan, K. (2020) *Time Series Forecasting — ARIMA vs Prophet* , *Medium*. Available at: https://medium.com/analytics-vidhya/time-series-forecasting-arima-vs-prophet-5015928e402a (Accessed: 14 June 2023).

Hessing, T. (no date) *Friedman Non Parametric Hypothesis Test*, *Six Sigma Study Guide*. Available at: https://sixsigmastudyguide.com/friedman-non-parametric-hypothesis-test/ (Accessed: 13 June 2023).

International Service Design Institute (no date) *Tufte’s 6 Principles for Graphical Integrity (Adopted for Service Design) - International Service Design Institute*, *International Service Design Institute*. Available at: https://internationalservicedesigninstitute.com/tuftes-6-principles-graphical-integrity-adopted-service-design/ (Accessed: 16 June 2023).

Karatas, G. (2023) *Beautiful Soup vs Scrapy: Which Should You Choose in 2023?*, *AI Multiple*. Available at: https://research.aimultiple.com/beautiful-soup-vs-scrapy/ (Accessed: 9 June 2023).

Koehrsen, W. (2018) *Hyperparameter Tuning the Random Forest in Python | by Will Koehrsen | Towards Data Science*, *Medium*. Available at: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74 (Accessed: 18 April 2023).

Koehrsen, Wi. (2017) *Random Forest in Python. A Practical End-to-End Machine Learning… | by Will Koehrsen | Towards Data Science*, *Medium*. Available at: https://towardsdatascience.com/random-forest-in-python-24d0893d51c0 (Accessed: 13 June 2023).

Kumar, S. (2020) *7 Ways to Handle Missing Values in Machine Learning*, *Towards Data Science*. Available at: https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e.

Lewinson, E. (2022) *Time Series DIY: Seasonal Decomposition*, *Towards Data Science*. Available at: https://towardsdatascience.com/time-series-diy-seasonal-decomposition-f0b469afed44 (Accessed: 14 June 2023).

M, P. (2021) *End-to-End Introduction to Evaluating Regression Models*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/.

McQuaid, D. (2023a) *Combining Data in Pandas With merge(), .join(), and concat() – Real Python*. Available at: https://realpython.com/pandas-merge-join-and-concat/.

McQuaid, D. (2023b) *Encoding Categorical Data*.

McQuaid, D. (2023c) *Handling Data with pandas DataFrame*.

McQuinn, C. (2023) *Building homes in Dublin costs up to 30% more than in Europe due to ‘higher spec’, Government told – The Irish Times*, *The Irish Times*. Available at: https://www.irishtimes.com/ireland/housing-planning/2023/05/04/building-homes-in-dublin-costs-30-more-than-in-europe-due-to-higher-spec-government-told/ (Accessed: 13 June 2023).

Mishra, S. (2020) *Methods for Normality Test with Application in Python*, *Towards Data Science*. Available at: https://towardsdatascience.com/methods-for-normality-test-with-application-in-python-bb91b49ed0f5.

Muthukrishnan, B. (2019) *Dashboard - interactive heatmap visualization using dash-plotly | by Balaji Muthukrishnan | Medium*, *Medium*. Available at: https://medium.com/@balajiciet/dashboard-interactive-heatmap-visualization-using-dash-plotly-6e69cf732e34 (Accessed: 16 June 2023).

O’Carroll, L. (2023) *Brexit: UK construction costs ‘have risen much more steeply than EU’ | Brexit | The Guardian*, *The Guardian*. Available at: https://www.theguardian.com/politics/2023/jan/24/brexit-uk-construction-costs-eu (Accessed: 13 June 2023).

Omoniyi, T. (2022) *Reddit Sentiment Analysis*, *Medium*. Available at: https://medium.com/@kiddojazz/reddit-sentiment-analysis-f8a1a790124a (Accessed: 10 June 2023).

Orac, R. (no date) *How to Speed Up and Save Memory in Your Python Code | by Roman Orac | Geek Culture | May, 2023 | Medium*, *Medium*. Available at: https://medium.com/geekculture/how-to-speed-up-and-save-memory-in-your-python-code-633325f009e9 (Accessed: 16 June 2023).

Pandas (no date) *pandas.DataFrame.corr — pandas 2.0.2 documentation*. Available at: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html (Accessed: 13 June 2023).

Pawangfg (2020) *Mann and Whitney U test*, *GeeksforGeeks*. Available at: https://www.geeksforgeeks.org/mann-and-whitney-u-test/ (Accessed: 13 June 2023).

Perrier, R. (2021) *Magic Commands for Profiling in Jupyter Notebook | by Remi Perrier | Towards Data Science*, *Medium*. Available at: https://towardsdatascience.com/magic-commands-for-profiling-in-jupyter-notebook-d2ef00e29a63 (Accessed: 16 June 2023).

Plotly (no date a) *A Minimal Dash App | Dash for Python Documentation | Plotly*. Available at: https://dash.plotly.com/minimal-app (Accessed: 16 June 2023).

Plotly (no date b) *Box plots in Python*. Available at: https://plotly.com/python/box-plots/ (Accessed: 12 April 2023).

Plotly (no date c) *Dash in 20 Minutes Tutorial | Dash for Python Documentation | Plotly*, *Plotly*. Available at: https://dash.plotly.com/tutorial (Accessed: 16 June 2023).

Plotly (no date d) *Plotly express in Python*, *Plotly*. Available at: https://plotly.com/python/plotly-express/ (Accessed: 16 June 2023).

Plotly (no date e) *Static image export in Python*, *Plotly*. Available at: https://plotly.com/python/static-image-export/ (Accessed: 16 June 2023).

Plotly (no date f) *Using Dash in Jupyter and Workspaces | Dash for Python Documentation | Plotly*, *Plotly*. Available at: https://dash.plotly.com/workspaces/using-dash-in-jupyter-and-workspaces (Accessed: 16 June 2023).

Portilla, J. M. (2018) *Using Python and Auto ARIMA to Forecast Seasonal Time Series | by Jose Marcial Portilla | Medium*, *Medium*. Available at: https://medium.com/@josemarcialportilla/using-python-and-auto-arima-to-forecast-seasonal-time-series-90877adff03c (Accessed: 14 June 2023).

Python for Data Science (no date) *Wilcoxon Sign-Ranked Test*. Available at: https://pythonfordatascienceorg.wordpress.com/wilcoxon-sign-ranked-test-python/ (Accessed: 13 June 2023).

Schröer, C., Kruse, F. and Gómez, J. M. (2021) ‘A systematic literature review on applying CRISP-DM process model’, *Procedia Computer Science*. Elsevier B.V., 181, pp. 526–534. doi: 10.1016/J.PROCS.2021.01.199.

SciPy (no date a) *scipy.stats.kruskal — SciPy v1.10.1 Manual*, *SciPy*. Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html (Accessed: 13 June 2023).

SciPy (no date b) *scipy.stats.mannwhitneyu — SciPy v1.10.1 Manual*. Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html (Accessed: 13 June 2023).

SciPy (no date c) *scipy.stats.wilcoxon — SciPy v1.10.1 Manual*, *SciPy*. Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.wilcoxon.html (Accessed: 13 June 2023).

Selvaraj, N. (2022) *How to Perform Sentiment Analysis with Python? | 365 Data Science*, *365DataScience*. Available at: https://365datascience.com/tutorials/python-tutorials/sentiment-analysis-with-python/ (Accessed: 4 June 2023).

Shaw, A. (no date) *Getting Started With Testing in Python – Real Python*, *Real Python*. Available at: https://realpython.com/python-testing/ (Accessed: 16 June 2023).

Shivaji, D. (2021) *How to Scroll using Selenium in Python*, *Medium*. Available at: https://davidshivaji.medium.com/how-to-scroll-using-selenium-in-python-ad1eba1e9bca (Accessed: 9 June 2023).

Tam, A. (2022) *Profiling Python Code* , *MachineLearningMastery*. Available at: https://machinelearningmastery.com/profiling-python-code/ (Accessed: 16 June 2023).

The Housing Agency (2018) *Comparison of Residential Construction Costs in Ireland to other European Countries*. Dublin. Available at: www.housingagency.ie (Accessed: 12 June 2023).

Weiss, N. A. (2017) *Introductory Statistics*. 10th edn. Pearson Education.

Whelan, K. (2014) ‘Ireland’s Economic Crisis: The Good, the Bad and the Ugly’, *Journal of Macroeconomics*. North-Holland, 39(PB), pp. 424–440. doi: 10.1016/J.JMACRO.2013.08.008.

Wirth, R. and Hipp, J. (2000) ‘CRISP-DM: Towards a Standard Process Model for Data Mining’, *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 1.

Zach (2020a) *How to Perform a Kruskal-Wallis Test in Python - Statology*, *Statology*. Available at: https://www.statology.org/kruskal-wallis-test-python/ (Accessed: 13 June 2023).

Zach (2020b) *How to Perform the Friedman Test in Python*, *Statology*. Available at: https://www.statology.org/friedman-test-python/ (Accessed: 13 June 2023).

Zafra, M. F. (2019) *Web Scraping news articles in Python*, *Medium*. Available at: https://towardsdatascience.com/web-scraping-news-articles-in-python-9dd605799558 (Accessed: 9 June 2023).