# Group ID - MSc in Data Analytics

Author: S. Clifford

e-mail: [sbs23100@student.cct.ie](mailto:sbs23100@student.cct.ie)

Student ID: 23100

Github link: <https://github.com/sbs23100/MScDataAnalyticsCA2.git>

Contents

[Group ID - MSc in Data Analytics 1](#_Toc136037152)

[1 Introduction 4](#_Toc136037153)

[2 Data Preparation and Visualization 4](#_Toc136037154)

[2.1 Acquiring Raw Data 4](#_Toc136037155)

[2.2 Exploratory Data Analysis 5](#_Toc136037156)

[2.3 structure and enrich your data. 6](#_Toc136037157)

[2.4 interactive dashboard 8](#_Toc136037158)

[3 Statistics 8](#_Toc136037159)

[3.1 Overall Dataset Summary 8](#_Toc136037160)

[3.2 inferential statistics 12](#_Toc136037161)

[3.3 parametric and non-parametric inferential statistical techniques 13](#_Toc136037162)

[3.3.1 Independent T-Test (Parametric) 14](#_Toc136037163)

[3.3.2 One-Way ANOVA (Parametric) 14](#_Toc136037164)

[3.3.3 Wilcoxon Rank-Sum Test (Non-parametric) 14](#_Toc136037165)

[3.3.4 Chi-Squared Test (Non-parametric) 15](#_Toc136037166)

[3.3.5 Mann-Whitney U Test (Non-parametric) 15](#_Toc136037167)

[3.4 Challenges 16](#_Toc136037168)

[4 Machine learning for data analytics 16](#_Toc136037169)

[4.1 Prediction 16](#_Toc136037170)

[4.1.1 Random Forest regression 16](#_Toc136037171)

[4.1.2 Support Vector Regression (SVR) 17](#_Toc136037172)

[4.2 sentimental analysis 19](#_Toc136037173)

[4.2.1 Sentiment on home building outlook from a buyer’s perspective 19](#_Toc136037174)

[4.2.2 Cost of home building from a builder’s perspective 22](#_Toc136037175)

[5 Conclusion 24](#_Toc136037176)

[6 Bibliography 25](#_Toc136037177)

# Introduction

The following analytical report is based on construction industry data in Ireland, compared to similar data in the European perspective, to quantify how the Irish construction sector compares to its European peers.

# Data Preparation and Visualization

## Acquiring Raw Data

As per the Scenario presented for this report in the introduction, Data would have to be chosen that has a common quantifiable and comparable basis for Ireland but also across European countries, which is the chosen geographical area of comparison for this report.

Comparing data within the European union allows for the use of NACE Rev. 2 data. NACE Rev. 2 (Statistical Classification of Economic Activities in the European Community) is a European standard for classifying economic activities. It provides a unified and standardized classification system that is used for statistical and analytical purposes, particularly in the European Union (EU) and European Free Trade Association (EFTA) member countries. Statistics produced on the basis of NACE are comparable at European and, in general, at world level. The use of NACE is mandatory within the European Statistical System. (Eurostat, 2008)

Basing the analysis produced in this report from data acquired from Eurostat (statistical office of the European Union) (Eurostat, n.d.) offers reliable, comparable and standardised data.

Additionally, from a legal and copyright perspective Eurostat has a policy of encouraging free re-use of its data, both for non-commercial and commercial purposes. All statistical data, metadata, content of web pages or other dissemination tools, official publications and other documents published on its website, with the exceptions listed in, [Exceptions](https://ec.europa.eu/eurostat/web/main/about-us/policies/copyright) (Eurostat, n.d.), can be reused without any payment or written licence provided that:

* the source is indicated as Eurostat.
* when re-use involves modifications to the data or text, this must be stated clearly to the end user of the information. (Eurostat, n.d.)

Based on the information given a custom dataset was created filtered to the following criteria:

* 37 EU member states, including Ireland were selected for initial data comparison.
* Data between the years of 2010 and 2023
* The following indicators were used:
  + Production Index
  + Number of Persons Employees Index
  + Hours worked Index.
  + Gross Wages and Salaries Index

The link to the above mentioned filtered data is [ei\_isbu\_q\_\_custom\_6200052\_linear](https://ec.europa.eu/eurostat/databrowser/view/EI_ISBU_Q__custom_6200052/settings_1/table?lang=en). (Eurostat, 2023). This data was download and is presented separately as a link to this report (Eurostat, 2023).

## Exploratory Data Analysis

Initially data from (Eurostat, 2023) was imported in CSV format into Jupiter notebook ‘ConstructionData.ipynb’ (Clifford, 2023) section 2.1(to be referred as notebook from this point forward) and converted to pandas data frame.

First look at the dataset, notebook section 2.2 we can see there is 11 columns, 2 of them as we can see from Figure 1 have a different quantity of entries from the remaining, this indicates missing or null data. Rows with Null values in the ‘OBS\_VALUE’ column came to 4, these rows were dropped as they represent a small percentage of the overall data, and were Q4 of 2022, which would indicate possibly values were not recorded for the final year of the dataset time span.

A screenshot of a computer

Description automatically generated

Figure 1 df information

From Figure 1 and from notebook section 2.3 OBS\_FLAG has 634 entries consisting of 351 for ‘p’ and 279 for ‘e’. These flag values refer to data being provisional and estimated respectively. For the purpose of this report, we will include the values that are flagged as estimated and provisional and delete the OBS\_FLAG column. This was done as estimated and provisional data can act as an early warning to a shift in trends or the emergence of a new pattern in the data.

6 other columns ‘DATAFLOW’, ‘LAST UPDATE’, ‘freq’, ‘unit’,s\_adj’ and ‘nace\_r2’ contain single recurring values which are characteristics which are common to all values, therefore these columns will be dropped before proceeding, see notebook section 2.4.

## structure and enrich your data.

Looking at the data there is a column ‘indic’ column, which denotes the indicator, which is present for each country, the indicator references are shown in Table 1 Indicator Reference, below.

|  |  |
| --- | --- |
| IS\_IP | Production Index |
| IS-EPI | Numbers of persons employed Index |
| IS-HWI | Hours worked index |
| IS-WSI | Gross wages and salaries Index |

Table 1 Indicator Reference

To aid further analysis in machine learning, these indicator values were separated out into individual column. The separation of these values can be seen in notebook section 2.4.1.

In notebook section 5 an interactive bar chart for each indicator was created which visualises the indicators across the full range of the time period. Observing the sequence, it was visually noted that bars dropped off of Turkey (Country Indicator – TR). A further check for last data entry for each Country is shown in Table 2 Missing last values in df. We can see the last entry for Turkey was 2016, As a considerable section of the dataset for Turkey is not present, it was decided to drop Turkey from the analysis completely, as seen in notebook section 2.5.



Table 2 Missing last values in df

A check for null values was already done as part of, section 2.2, to ensure we have all non-zero data a check was carried out in notebook section 2.4.2 which showed no ‘0’ values.

To aid the graphical display further in the report an additional ‘country\_name’ column was added, using the ‘pycountry’ library (Theune, n.d.), this library assigns a readable country name for each 2-letter country code in the ‘geo’ Column’. See notebook section 5.1. to confirm this was implemented for all countries a check was carried out in notebook 5.2 which shows one country did not have a country name assigned to it, country code ‘EL’. An online search showed this country to be ‘Greece’, the data was then filtered to check if Greece existed, when confirmed not the correct country name was assigned, notebook 5.4.

This change was applied to animated charts in notebook section 6.1.

A screenshot of a computer

Description automatically generated with low confidence

Figure 2 Interactive Bar chart

## interactive dashboard

# Statistics

## Overall Dataset Summary

To summarise the initial data from (Eurostat, 2023) we can see from Figure 1 that the dataset contains:

* 7252 entries
* Column count: 11

The data was then pre-processed as described in section 2 page 3. The resulting dataset has 4 numerical columns representing data values and two categorical both of which represent country names, one is the country code (‘geo’ column) the second is the readable country name (‘country\_name’ column), see Figure 3 Processed df Information.

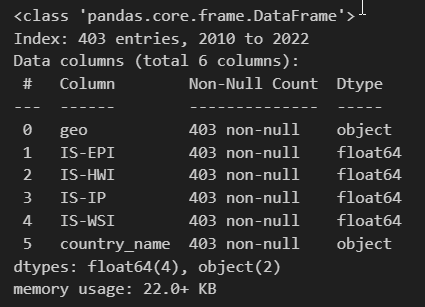


Figure 3 Processed df Information

Based on the provided information in Table 3, from notebook section 7.1, we can see that IS-WSI has the highest mean, standard deviation, and maximum value among the variables, indicating a wider spread and potential outlier. IS-HWI has the lowest mean and lower standard deviation, suggesting a tighter distribution.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **IS-EPI** | **IS-HWI** | **IS-IP** | **IS-WSI** |
| **count** | 403 | 403 | 403 | 403 |
| **mean** | 430.0789 | 424.594 | 438.2898 | 459.3511 |
| **std** | 65.55036 | 77.15776 | 120.6349 | 130.8457 |
| **min** | 247.6 | 216 | 148 | 219.7 |
| **25%** | 399 | 389.25 | 375 | 390.25 |
| **50%** | 414.6 | 408.8 | 409.1 | 422.6 |
| **75%** | 454.45 | 439.55 | 472.35 | 500.85 |
| **max** | 706.7 | 780.6 | 1101.7 | 1330.7 |

Table 3 Processed data information for numerical values

Looking at the distribution of the 4 indicator columns, in Figure 4 it can be seen that the 4 indicators are displaying skewed distribution, this suggests that the data for each is not evenly distributed around the mean, as is depicted by the long tail to right of each indicator plots.

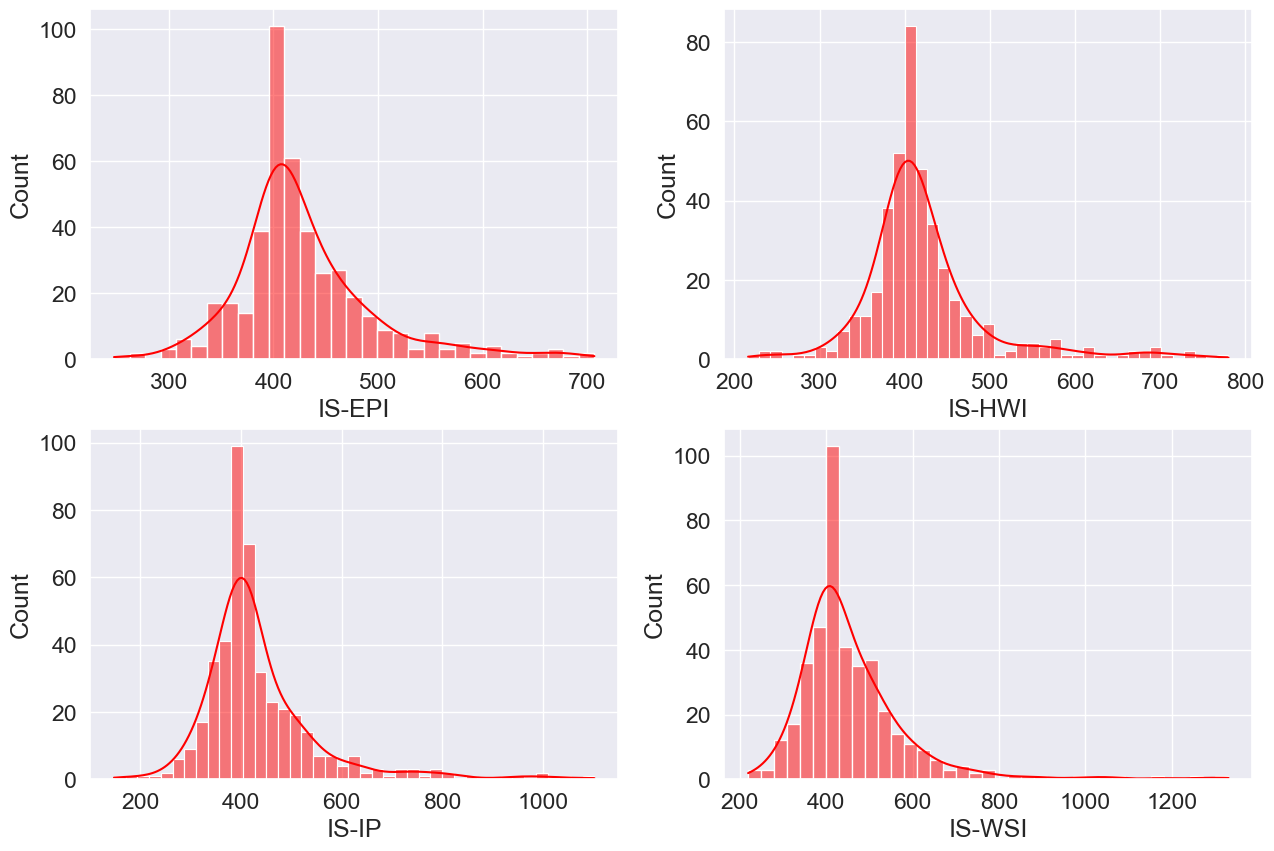


Figure 4 Histograms displaying distribution of indicator values

Further to the observation as seen in Figure 4, looking at the boxplot plots for the 4 indicators in Figure 5, the presence of outliers predominantly to the right, confirms this tendency.

A picture containing text, screenshot, line, diagram

Description automatically generated

Figure 5 Boxplot plot for indicator values

Quantifying the number of outliers for each indicator, see notebook section 7.2.1, results in Figure 6, we can see that the number of outliers is a small percentage of the overall values. Taking into consideration the economic factors which can affect the construction industry over the time period in question 2010 to 2022, we know there existed great swings in activity in the sector. For this reason, it has been decided not to remove outlier values and to continue with the dataset as it is for further analysis.

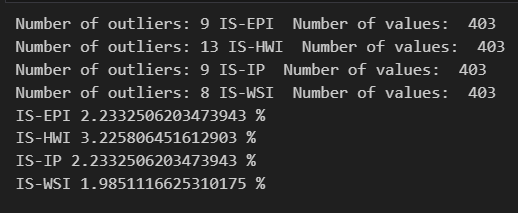


Figure 6 Number of outliers for indicator values

Further testing of indicator data across all countries shows that none of the 4 indicators provided normal distribution, as was calculated by an Anderson-Darling test in notebook section 7.3. Additionally to the Anderson-Darling test, a Shapiro-wilk test and a scripy.stats normaltest, which is based on D'Agostino and Pearson's test, which combines skewness and kurtosis to assess the departure from normality, were applied as confirmation. Table 4 Tests for assertaining normality

|  |
| --- |
| Anderson-Darling |
| Shapiro-wilk |
| D'Agostino and Pearson's |

Table 4 Tests for assertaining normality

All 3 tests showed that the indicator values were not normally distributed.

## inferential statistics

As infernal statistics is best applied to normally distributed data samples, firstly a deeper understanding of the distribution of country data was required. The Production Index indicator (IS-IP) was chosen and an analysis of which countries within the dataset had normally distributed IS-IP values over the time period was assessed, notebook section 7.3.1 shows 17 of the countries contain normally distributed IS-IP data, including Ireland. Have a sample of normally distributed for several countries including Ireland allows us to propose the following question:

To gain insights into the population values:

*Calculate the confidence intervals that Ireland has an above average proportion of Production Index compared to its European counterparts.*

In order to address this question, firstly a column to Calculate the proportion of Construction Index values for each country relative to the total Construction Index values across all European countries for each quarter was created, see notebook section 7.5.1. To confirm the values were correct a test was created, notebook section 7.5.2, to confirm the sum of the proportions per year, sum to 1.

To calculate the confidence interval for Ireland having an above-average proportion per quarter in the Production index data, the following steps were applied, and can be seen in notebook section 7.5.3.

* Create a dataset with data for Ireland only.
* Calculate the average proportion across all countries for each quarter.
* Calculate the sample size for Ireland for each quarter.
* Calculate the mean and standard deviation of the proportions for Ireland for each year.
* Calculate the t-value for the desired confidence level.
* Calculate the margin of error.
* Calculate the confidence intervals.
* Check if the lower bound is above the average proportion.
* Check if the lower bound is above the average proportion, indicating that Ireland has an above-average level of proportion with the specified confidence level.

The results of the above noted sequence of calculations can be seen in Table 5 below. From this data we can state with 95% confidence that the proportion of Production index data for Ireland for each year will fall with the confidence intervals (CI) for each year in Table 5. Also, with 95% confidence for each year except 2017, we can say that Ireland has a proportion of the production index which is not above the European average.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **TIME\_PERIOD** | **Lower Bound** | **European Average Proportion** | **Upper Bound** | **Above Average** | **Ireland** |
| 2010 | 0.0062 | 0.0075 | 0.0098 | FALSE | 0.0080 |
| 2011 | 0.0065 | 0.0075 | 0.0080 | FALSE | 0.0072 |
| 2012 | 0.0035 | 0.0073 | 0.0097 | FALSE | 0.0066 |
| 2013 | 0.0069 | 0.0075 | 0.0081 | FALSE | 0.0075 |
| 2014 | 0.0031 | 0.0070 | 0.0090 | FALSE | 0.0061 |
| 2015 | 0.0065 | 0.0075 | 0.0127 | FALSE | 0.0096 |
| 2016 | 0.0058 | 0.0077 | 0.0098 | FALSE | 0.0078 |
| 2017 | 0.0077 | 0.0070 | 0.0080 | TRUE | 0.0079 |
| 2018 | 0.0072 | 0.0076 | 0.0086 | FALSE | 0.0079 |
| 2019 | 0.0066 | 0.0071 | 0.0085 | FALSE | 0.0076 |
| 2020 | 0.0035 | 0.0072 | 0.0090 | FALSE | 0.0062 |
| 2021 | 0.0068 | 0.0078 | 0.0070 | FALSE | 0.0069 |
| 2022 | 0.0038 | 0.0069 | 0.0102 | FALSE | 0.0070 |

Table 5 Population data results

## parametric and non-parametric inferential statistical techniques

Hypothesis testing involves making a claim or hypothesis about a population parameter and then using sample data to evaluate the evidence for or against that claim.

Using only the countries identified in notebook section 7.3.1 who have all 4 indicators normally distributed the following parametric and non-parametric tests were carried out:

### Independent T-Test (Parametric)

This test is used to compare the mean production index (IS\_IP) between Ireland and other countries.

#### Hypotheses

Null Hypothesis (H0): The mean production index of Ireland is equal to the mean production index of another country.

Alternative Hypothesis (H1): The mean production index of Ireland is greater than the mean production index of another country.

#### Conclusion

If the p-value is less than the significance level, reject the null hypothesis and conclude that Ireland has a significantly higher mean production index compared to the other country.

### One-Way ANOVA (Parametric)

Use this test to compare the mean production index (IS\_IP) among multiple countries (including Ireland).

#### Hypotheses

Null Hypothesis (H0): The mean production index is the same across all countries.

Alternative Hypothesis (H1): At least one country (including Ireland) has a higher mean production index compared to the others.

#### Conclusion

If the p-value is less than the significance level, reject the null hypothesis and conclude that there is a significant difference in the mean production index among the countries.

### Wilcoxon Rank-Sum Test (Non-parametric)

Use this test to compare the distribution of gross wages and salaries (IS-WSI) between Ireland and another country.

#### Hypotheses

Null Hypothesis (H0): The distribution of gross wages and salaries is the same between Ireland and another country.

Alternative Hypothesis (H1): The distribution of gross wages and salaries is greater in Ireland compared to another country.

#### Conclusion

If the p-value is less than the significance level, reject the null hypothesis and conclude that Ireland has a significantly higher distribution of gross wages and salaries compared to the other country.

### Chi-Squared Test (Non-parametric)

Use this test to examine the association between the country (Ireland vs. another country) and the categorical variable, such as the number of persons employed index (IS-EPI).

#### Hypotheses

Null Hypothesis (H0): There is no association between the country and the number of persons employed index.

Alternative Hypothesis (H1): Ireland has a higher proportion of a larger number of employed persons compared to the other country.

#### Conclusion

If the p-value is less than the significance level, reject the null hypothesis and conclude that there is a significant association between the country and the number of persons employed index, indicating that Ireland has a higher proportion of a larger number of employed persons compared to the other country.

### Mann-Whitney U Test (Non-parametric)

Use this test to compare the distribution of hours worked index (IS-HWI) between Ireland and another country.

#### Hypotheses

Null Hypothesis (H0): The distribution of hours worked index is the same between Ireland and the other country.

Alternative Hypothesis (H1): The distribution of hours worked index is greater in Ireland compared to the other country.

#### Conclusion

If the p-value is less than the significance level, reject the null hypothesis and conclude that Ireland has a significantly higher distribution of work hours worked index between Ireland and other countries.

## Challenges

The main challenge was filtering to data that was usable for infernal statistics i.e., preferably normally distributed.

# Machine learning for data analytics

## Prediction

The following question has been proposed for the application of a machine learning algorithm:

*Can we predict the proportion of production Index for a European country in a given year based on the other construction indices (IS-EPI, IS-HWI, IS-WSI) and country information?*

The following two models have been applied:

### Random Forest regression

Random forest regression was applied to the data in order to address the question posed in section 4.1 above.

The analysis can be seen in notebook section 8.1

The results of this analysis can be seen in Figure 7, as can be seen the actual v the predicted values are located reasonably close to the trendline, the closer the markers lie to the predictive trend, indicates the more accurate the model is performing.

A picture containing text, screenshot, line, plot

Description automatically generated

Figure 7 Actual v Predicted for Random Forest Regressor

### Support Vector Regression (SVR)

Support Vector Regression was the second approach to addressing the question posed in section 4.1. The application of this model can be seen in notebook section 8.2.

The results of the modelling are depicted in Figure 8.

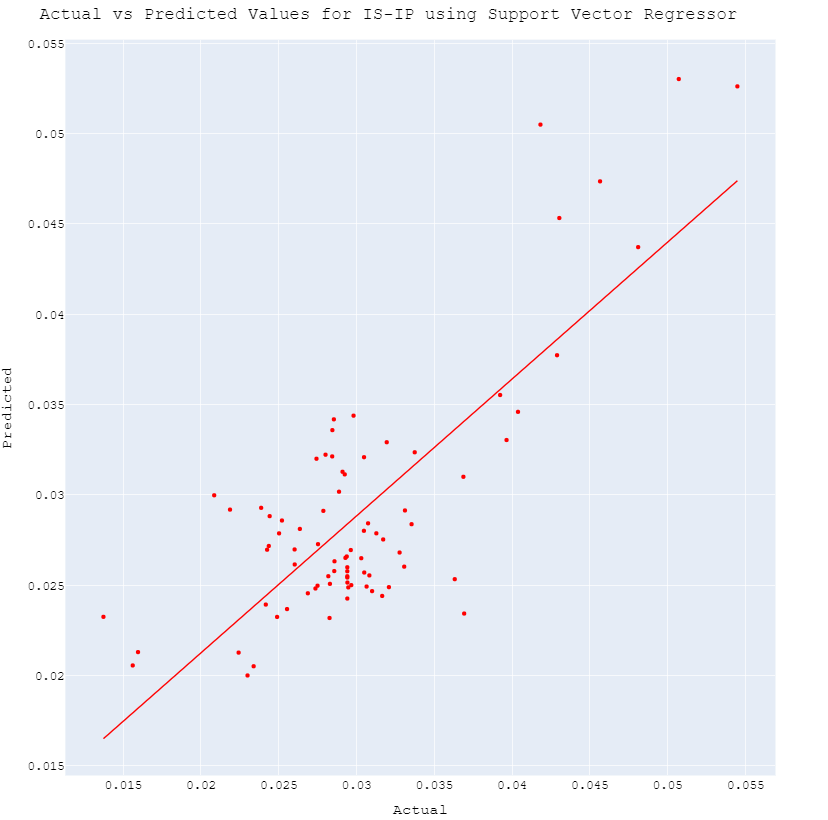


Figure 8 Actual v Predicted for IS-IP using SVR

Comparing the outcomes of both learning models in Figure 9, it can be seen that both models seemed to perform very similarly, we can see however from Table 6 that the Random forest model performed better. As both root mean square errors are quiet low, we can state that both models performed very well based on this matric.

|  |  |
| --- | --- |
| SVR Mean Square Error | 2.17e-05 |
| Random Forest Mean Square Error | 1.59e-5 |

Table 6 Root Mean Square comparison of RFR and SVR

A graph with red and blue dots

Description automatically generated with low confidence

Figure 9 Comparison of RFR and SVR

## sentimental analysis

Sentimental analysis was carried on a number of reports produced on the state of various aspects of the domestic home building sector of the construction industry in Ireland and in Europe. Two notebooks have been attached with this report, sentimentAnlNotebook.ipynb and MLsentiment.ipynb. (PDF’s are also attached to report in file) The analysis was carried out under the following headings.

### Sentiment on home building outlook from a buyer’s perspective

4 PDF reports were analysed for sentiment analysis in the Irish market, with the results for each document shown below, Figure 10. Analysis can be seen in notebook sentimentAnlNotebook.ipynb .As can be seen the sentiment lies just marginally on the positive side which indicates a neutral sentiment.

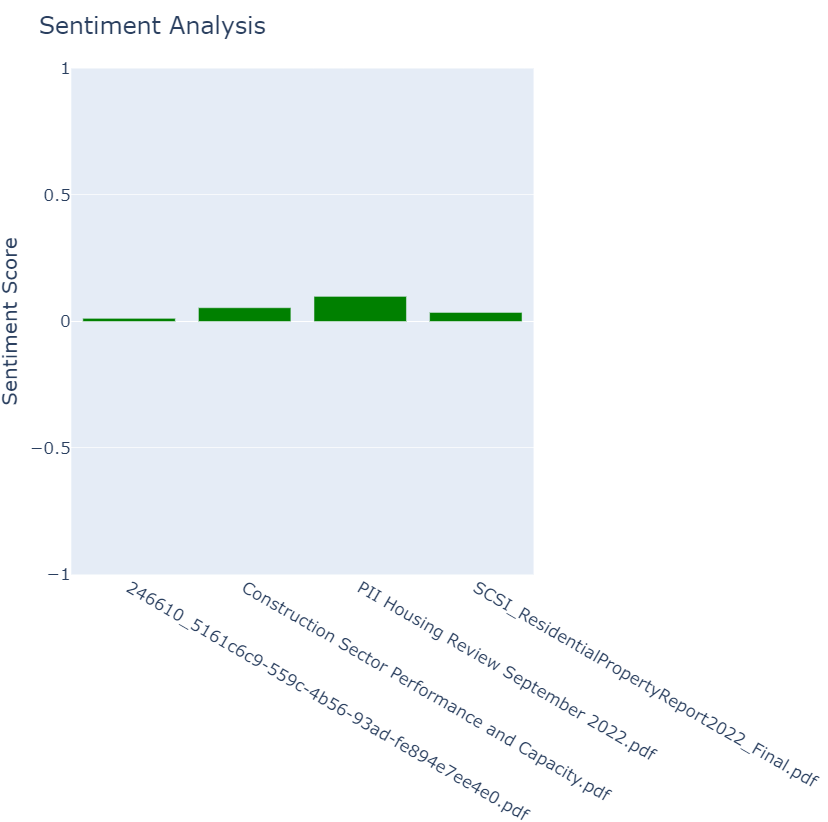


Figure 10 Sentiment analysis for Home builds Ireland

Comparing the this to similar documents based on the current state in Europe, Figure 11 show the result of analysis again located in notebook sentimentAnlNotebook.ipynb.

A picture containing text, screenshot, line, diagram

Description automatically generated

Figure 11 Sentiment analysis for Home builds Europe

Based on sentiment analysis shown in Figure 10 and Figure 11 and based on figures shown in Table 7 , we can state that the sentiment in Europe while also classified as neutral, appears slightly more positive.

|  |  |
| --- | --- |
| **File** | **sentiment** |
| **Ireland** |  |
| 246610\_5161c6c9-559c-4b56-93ad-fe894e7ee4e0.pdf | 0.012418 |
| Construction Sector Performance and Capacity.pdf | 0.054362 |
| PII Housing Review September 2022.pdf | 0.099269 |
| SCSI\_ResidentialPropertyReport2022\_Final.pdf | 0.035755 |
|  |  |
| **Europe** |  |
| Emerging Trends in Real Estate Europe 2023 Rep... | 0.115881 |
| at-property-index-2022-final.pdf | 0.095851 |
| dp171\_en.pdf | 0.091793 |

Table 7 Home build sentiment figures for Ireland and Europe

### Cost of home building from a builder’s perspective

As with 4.2.1 above 3 separate documents were analysed for sentiment on building costs (notebook: sentimentAnlNotebook.ipynb.) both in Ireland, Figure 12 and also in Europe Figure 13. Both sets of analysis are classified as being neutral as they are marginally in positive figures, as can be seen in Table 8 also. The figures for Europe, as was the case in section 4.2.1, here also are slightly more positive in sentiment compared to Ireland.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 12 Sentiment of construction costs Ireland

A screenshot of a computer

Description automatically generated with low confidence

Figure 13 Sentiment of construction costs Ireland

|  |  |
| --- | --- |
| **File** | **sentiment** |
| **Ireland** |  |
| 2022-05-24\_opening-statement-kevin-james-vice-... | 0.075669 |
| 256082\_afbe94c3-ebf1-4201-9a4a-a6ac9cddc69a.pdf | 0.039334 |
| Rising-construction-costs-and-the-residential-... | 0.066723 |
|  |  |
| **Europe** |  |
| 22-cs11-\_Construction\_Building\_Materials\_-\_Com... | 0.07358 |
| ING-Think-eu-construction-outlook-optimism-amo... | 0.076489 |
| International Construction Costs 2022-2.pdf | 0.111894 |

Table 8 Construction cost sentiment figures for Ireland and Europe

# Conclusion

Not all aspects of the assignment were implemented, Time management from my perspective was poor. I went down rabbit holes in my learning and development which were not relevant to this report, ultimately costing me time which could have been put to better use focused on the task at hand.

# Bibliography

Birchwood University, 2023. *birchwoodu.org.* [Online]   
Available at: https://www.birchwoodu.org/top-programming-languages-for-data-scientists/  
[Accessed 24 March 2023].

Clifford, S., 2023. *ConstructionData.ipynb,* Killarney: s.n.

Clifford, S., 2023. *CPI.ipynb,* Killarney: s.n.

Clifford, S., 2023. *house\_registrations\_by\_area.ipynb,* Killarney: s.n.

Clifford, S., 2023. *MIR.ipynb,* Killarney: s.n.

Clifford, S., 2023. *population.ipynb,* Killarney: s.n.

Coursera , 2023. *Python or R for Data Analysis: Which Should I Learn?.* [Online]   
Available at: https://www.coursera.org/articles/python-or-r-for-data-analysis  
[Accessed 14 April 2023].

Data Science Process Alliance, n.d. *What is CRISP DM.* [Online]   
Available at: https://www.datascience-pm.com/crisp-dm-2/  
[Accessed 13 April 2023].

Data Science Process Alliance, n.d. *www.datascience-pm.com.* [Online]   
Available at: https://www.datascience-pm.com/semma/  
[Accessed 23 April 2023].

DATA.GOV.IE, 2021. *HSA10 - New House Registrations.* [Online]   
Available at: https://data.gov.ie/dataset/hsa10-new-house-registrations?package\_type=dataset  
[Accessed 02 March 2023].

datacamp.com, 2022. *datacamp.com.* [Online]   
Available at: https://www.datacamp.com/blog/top-programming-languages-for-data-scientists-in-2022  
[Accessed 24 March 2023].

datadriveninvestor.com, n.d. *datadriveninvestor.com.* [Online]   
Available at: https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb  
[Accessed 13 April 2023].

Dhingra, D., n.d. *How to transform features into Normal/Gaussian Distribution.* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2021/05/how-to-transform-features-into-normal-gaussian-distribution/  
[Accessed 1 April 2023].

Dissanayake, A., 2021. *Python vs R for Data Science.* [Online]   
Available at: https://medium.com/octave-john-keells-group/python-vs-r-for-data-science-320e167ffe90  
[Accessed 14 April 2023].

Eurostat, 2008. *NACE Rev. 2 – Statistical classification of economic activites in the European Community,* Luxembourg: Office for Official Publications of the European Communities,.

Eurostat, 2023. *Construction - quarterly data - index (2015 = 100) (NACE Rev. 2).* [Online]   
Available at: https://ec.europa.eu/eurostat/databrowser/view/EI\_ISBU\_Q\_\_custom\_6200052/settings\_1/table?lang=en  
[Accessed 15 May 2023].

Eurostat, 2023. *ei\_isbu\_q\_\_custom\_6200052\_linear.csv,* Luxembourg: European Commission, Eurostat.

Eurostat, n.d. *About Us/Who we are.* [Online]   
Available at: https://ec.europa.eu/eurostat/web/main/about-us/who-we-are  
[Accessed 15 May 2023].

Eurostat, n.d. *Copyright notice and free re-use of data.* [Online]   
Available at: https://ec.europa.eu/eurostat/web/main/about-us/policies/copyright  
[Accessed 15 May 2023].

Iantorno, M., 2023. *Lecturer.* Dublin: s.n.

IBM.com, n.d. *CRISP-DM Help Overview.* [Online]   
Available at: https://www.ibm.com/docs/it/spss-modeler/saas?topic=dm-crisp-help-overview  
[Accessed 13 April 2023].

James, G., Witten, D., Hastie, T. & Tibshirani, R., 2013. An introduction to statistical learning. In: *An introduction to statistical learning.* New York: Springer.

jetbrains.com, 2023. *The State of Developer Ecosystem 2022.* [Online]   
Available at: https://www.jetbrains.com/lp/devecosystem-2022/?source=google&medium=cpc&campaign=19867226536&term=best%20programming%20languages&content=651996376797&gclid=Cj0KCQjwlPWgBhDHARIsAH2xdNdnHHpx-c5mgllSOCY5\_IjYN6Mnu6j8Q4UmWJPl38747STLRfDTPQsaAq5rEALw\_wcB  
[Accessed 24 March 2023].

Khete, T., 2022. *R vs Python for Data Science and visualization: The language debate.* [Online]   
Available at: https://medium.com/@tusharkhete118/r-vs-python-for-data-science-and-visualization-the-language-debate-1aac453e7e29  
[Accessed 14 April 2023].

Korstanje, J., 2020. *Is Python faster than R?.* [Online]   
Available at: https://towardsdatascience.com/is-python-faster-than-r-db06c5be5ce8  
[Accessed 14 April 2023].

numpy.org, n.d. *numpy.org.* [Online]   
Available at: https://numpy.org/doc/stable/reference/random/generated/numpy.random.normal.html  
[Accessed 1 April 2023].

Reichenbach , M., 2021. *R vs Python: Which Is Best for Data?.* [Online]   
Available at: https://blog.boot.dev/python/r-vs-python/  
[Accessed 14 April 2023].

sas.com, n.d. *SAS Analytics Software & Solutions.* [Online]   
Available at: http://documentation.sas.com/doc/en/emref/14.3/n061bzurmej4j3n1jnj8bbjjm1a2.htm  
[Accessed 13 April 2023].

scikit-learn.org, n.d. *https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.GridSearchCV.html.* [Online]   
Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.GridSearchCV.html  
[Accessed 8 04 2023].

scipy.org, n.d. *docs.scipy.org.* [Online]   
Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.probplot.html  
[Accessed 1 April 2023].

scripy.org, n.d. *https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html.* [Online]   
Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html  
[Accessed 8 April 2023].

seaborn.pydata.org, n.d. *www.seaborn.pydata.org.* [Online]   
Available at: https://seaborn.pydata.org/generated/seaborn.pairplot.html  
[Accessed 31 March 2023].

seaborn.pydata.org, n.d. *www.seaborn.pydata.org.* [Online]   
Available at: https://seaborn.pydata.org/generated/seaborn.regplot.html  
[Accessed 31 March 2023].

Theune, C., n.d. *pycountry 22.3.5,* s.l.: pypi.