**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| Module Title: | MSc in Data Analytics |
| Assessment Title: | CA2 |
| Lecturer Name: | Sam Weiss  Marina Iantorno  Muhammad Iqbal  David McQuaid |
| Student Full Name: | Glenn Tarpey |
| Student Number: | SBS23066 |
| Assessment Due Date: | 09th June 2023 |
| Date of Submission: | 09/06/2023 |
| Word Count (excluding Abstract, Figures and References) : |  |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

## Abstract

## Sections

[Abstract 2](#_Toc137129528)

[Sections 3](#_Toc137129529)

[Figures 4](#_Toc137129530)

[Introduction 5](#_Toc137129531)

[Data Sources 6](#_Toc137129532)

[Data Explanation 6](#_Toc137129533)

[ETL Process – Construction Data 8](#_Toc137129534)

[ETL Process – Sentiment Analysis 9](#_Toc137129535)

[Output of analysis 9](#_Toc137129536)

[Example of Sentiment Analysis Process 10](#_Toc137129537)

[Analysis of time efficiency 11](#_Toc137129538)

[Statistical Analysis 12](#_Toc137129539)

[Construction Trends 12](#_Toc137129540)

[Construction/Dwelling Value 12](#_Toc137129541)

[Statistical Tests 14](#_Toc137129542)

[Note 14](#_Toc137129543)

[Parametric vs Non Parametric 14](#_Toc137129544)

[T-Test 15](#_Toc137129545)

[Spearman's rank correlation test 16](#_Toc137129546)

[Machine Learning 17](#_Toc137129547)

[Issues with the datasets 17](#_Toc137129548)

[Linear Regression 18](#_Toc137129549)

[Polynomial Regression 19](#_Toc137129550)

[K Means Clustering 21](#_Toc137129551)

## Figures

[Figure 1- Sentiment by topic 10](#_Toc137129516)

[Figure 2- Sentiment Example Before 10](#_Toc137129517)

[Figure 3- Sentiment Example Translated 10](#_Toc137129518)

[Figure 4- Sentiment Example with score 10](#_Toc137129519)

[Figure 5 - Time efficiency Reddit API 11](#_Toc137129520)

[Figure 6 - Construction/Dwelling Value by country 12](#_Toc137129521)

[Figure 7- Global Data Report - Construction Output 13](#_Toc137129522)

[Figure 8 - Spearmans Rank Correlation Test 16](#_Toc137129523)

[Figure 9- Linear Regression Metrics 18](#_Toc137129524)

[Figure 10- Linear Regression Prediction 18](#_Toc137129525)

[Figure 11- Polynomial Regression Metrics 19](#_Toc137129526)

[Figure 12 - Polynomial and Forecast – Ireland 20](#_Toc137129527)

## Introduction

In 2021, the size of the building market in Ireland was $31.8 Billion. In the coming years 2023–2026 the market is expected to grow at an Average Annual Growth Rate of at least 4%. This will be supported by investments in energy, transportation, and housing. The government intends to invest in projects related to energy and transportation infrastructure through the National Development Plan. (Global Data, 2022)

As compared to the overall European construction market which is currently valued at approximately $3.01 Trillion and expected to increase to 3.73 Trillion by 2028 it is clear that the construction industry is a rapidly expanding market with increasing scopes of work. In recent years the construction industry have started adopting business intelligence practices to further identify opportunities, cost savings and to maintain a strategic advantage over competitors. (researchandmarkets.com, 2023) (datashapa.com, 2022)

AI and Machine Learning are increasingly utilized in large scale construction projects for tasks such as project design, installation, data collection, and analysis, all with the aim of optimizing construction operations.

This report aims to explore this market, both nationally in Ireland and also in Europe in the goal of identifying trends, predictions and recommendations through the use of data analytics.

## Data Sources

The study and analysis of the construction sector, primarily focusing on Ireland as the primary case study alongside other countries in Europe, involves easily collecting data from government data sources and market research data. However, it is important to justify the choice of data and understand how crucial it is to comprehend the construction sector, observe patterns, trends, and make accurate forecasts.

In this report data from Eurostat will be used to analyse various aspects of the industry, including construction output, labour input, and gross value added from construction. These data sources will help to provide a good overview of the European construction industry. Eurostat is well regarded as a provider of trusted construction data as well as other industries. In addition the data is available via an open licence for commercial and non-commercial use as per the Eurostat [Data Policy](https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjgy5iWvbP_AhX6SkEAHVXrByMQFnoECAoQAw&url=https%3A%2F%2Fec.europa.eu%2Feurostat%2Fabout-us%2Fpolicies%2Fcopyright&usg=AOvVaw0syPEbrggORi5CWi0TcoZH) attached in the link. (Eurostat, 2021)

The data types being used in this report are:

* Construction Output (Building and Civil Engineering) indicates the value of production.
* Building Permit.
* Labor Input (Wages and Salary, Hours worked).
* Structural business statistics, including the Number of enterprises involved in construction, Multi-year Subcontracting, and Turnover.
* Gross Value Added from Construction and GDP from Construction, mainly encompassing Gross Fixed Capital Formation (to comprehend investments and returns).

## Data Explanation

Data was primarily sourced from Eurostat, categorized into two main categories:

1. Short-term business statistics (STS) provide information on various economic activities, including production, building permits indicators, producer prices in construction, and labour input indicators.
2. Structural business statistics (SBS) offer detailed insights into the structure, conduct, and performance of economic activities across several sectors. Key data include business demographics, output-related and input-related variables.

This data was collected in excel/csv format. It's used to analyse several aspects of the construction sector such as GDP share, producing values, turnover, challenges, costs, and exporting services.

Gross fixed capital formation (GFCF) data was considered to understand the economic impact of the construction sector. It refers to the investments made in fixed assets by resident producers over a specific period, minus disposals. This data, in Million Euros, is sourced from Eurostat's National Accounts.

The production value index, which represents the output and activity of the construction sector, was taken from Eurostat's short-term business statistics. The index measures changes in the volume of output on a monthly basis and includes building construction and civil engineering.

Building permits and dwelling data were also considered. A building permit authorizes the start of a building project, and an index based on these permits provides an indication of future workload for the building industry. Dwelling data refers to the rooms or suite of rooms intended for private habitation.

Labour inputs include the total number of people working in the unit (employment size), total hours worked, and total wages and salaries. These aim to show the volume of work done and the development of the wage and salaries bill.

Finally, Structural Business Statistics provided information on turnover value added, the number of enterprises, employment size, and subcontracting.

## ETL Process – Construction Data

A total of 12 construction datasets were used in this assignment. A variety of functions and techniques were performed to extract, manipulate and save these datasets to a new folder. To summarize this process an overall summary of the data cleaning process will be given and a summary of the functions used will be provided.

Overall steps:

1. Extract the data into variables to store them
2. Data Summary to check for missing value, info, shape, and summary statistics.
3. Transformation of the extracted data through a function to remove unnecessary values, change datatypes, and convert country codes to the country name.
4. Load the data into a folder named Transformed Data.

Functions:

* Data\_summary: This function accepts a file name, it reads the data from this file into a data frame and prints a variety of information about the data such as columns, shape and null values. This function was used to understand the datasets and identify any issues with them early on.
* Get\_country\_name: This function takes the alpha numeric shorthand country name used in Eurostats datasets and converts it into its full English spelling e.g. IE to Ireland. This allowed the dataset to be easier to read and understand.
* Clean\_and\_transform: This function was the most complex in this process. It takes custom parameters such as columns to drop (irrelevant columns) and a value name to replace the common “OBS\_Value” in the Eurostat datasets with a more meaningful name, making the datasets easier to work with and understand.

## ETL Process – Sentiment Analysis

As part of this report data was scraped using the Reddit API to gain insights into positive and negative viewpoints of construction topics. Like the construction data, this data also needed to be transformed.

Overall steps:

1. Connect to the Reddit API
2. Select the subreddit of choice e.g. Ireland
3. Search the subreddit for a pre-defined list of keywords
4. Take the resulting titles and translate to English
5. Tokenize the title, remove stop words and recreate the title
6. Repeat steps 1-5 for the next subreddit e.g. Germany

Once these steps were completed we had a data frame containing a list of titles to perform sentiment analysis on. The sentiment analysis method chosen was Lexicon based, more specifically using the VADER lexicon. This method was chosen as it applies a simple metric for sentiment based on dictionary matches in the VADER lexicon. It doesn’t require machine learning models to compute a sentiment value.

A variety of functions were written to assist with the ETL of the data and sentiment analysis steps:

* translateText : Uses the google translate module from the deep translator library to auto convert text to English
* cleanText : Removes quotes and dashes, then tokenizes the sentence and removes stop words before reconstructing the cleaned sentence
* dataFrameSentiment: applies a lambda function to a data frame to get sentiment scores
* dataFrameSentimentLabel : uses a defined threshold value to label a sentiment score as either negative, positive or neutral
* sentimentAnalysis : used to call the dataFrameSentiment and dataFrameSentimentLabel functions above

### Output of analysis

In the figure below you can see a breakdown of the count of sentiment scores by topic. We can see from the chart that posts containing the word **construction** have a substantially more negative sentiment. Using this sentiment data a construction company could benefit from understanding what public perception of their key topics are, in order to address the negative sentiment more effectively. This could be used in combination with a marketing or PR campaign to further improve public perception and opinion. Of course more data would be needed to support this endeavour but none the less the sentiment data could help them target their audience more effectively.

Note more charts on specific subreddits and trends are available in the accompanying notebook.

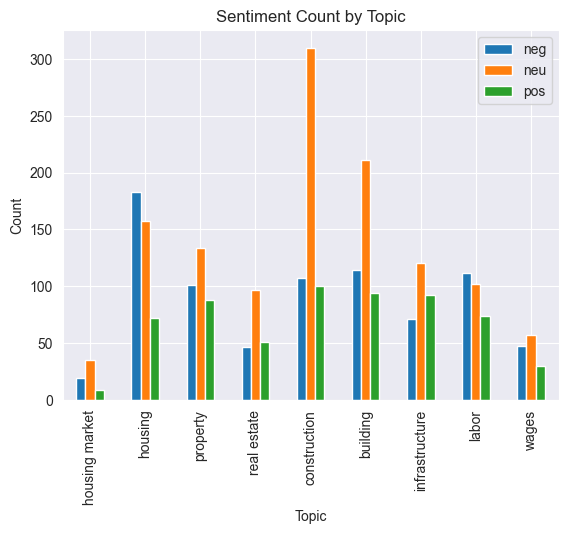


Figure 1- Sentiment by topic

### Example of Sentiment Analysis Process

In the 3 figures below you can see an example of the full process for the sentence for item number 1 in French. The sentence was translated from French to English and given a sentiment score (in this example positive).

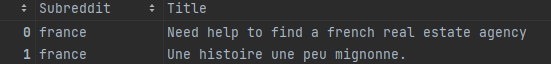


Figure 2- Sentiment Example Before

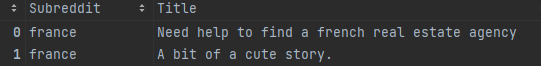


Figure 3- Sentiment Example Translated

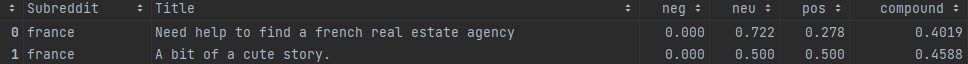


Figure 4- Sentiment Example with score

### Analysis of time efficiency

One potential issue with connecting to live data sources via an API is that depending on the amount of data requested it can take a long time to process. The Reddit API has a limit of 1000 results per query. We managed to get around this by disconnecting a reconnecting in-between searches.

Also we took measures of the start and finish time of the query in order to inform decisions on trade-off between data quantity and data speed. If you look at the figure below you will see that the time per post follows a somewhat linear relationship. Using a simple Linear Regression we found that we could obtain **3.68 posts per second**. This would be useful for future work on gathering more data. Ultimately we used the dataset with approximately 6000 posts.

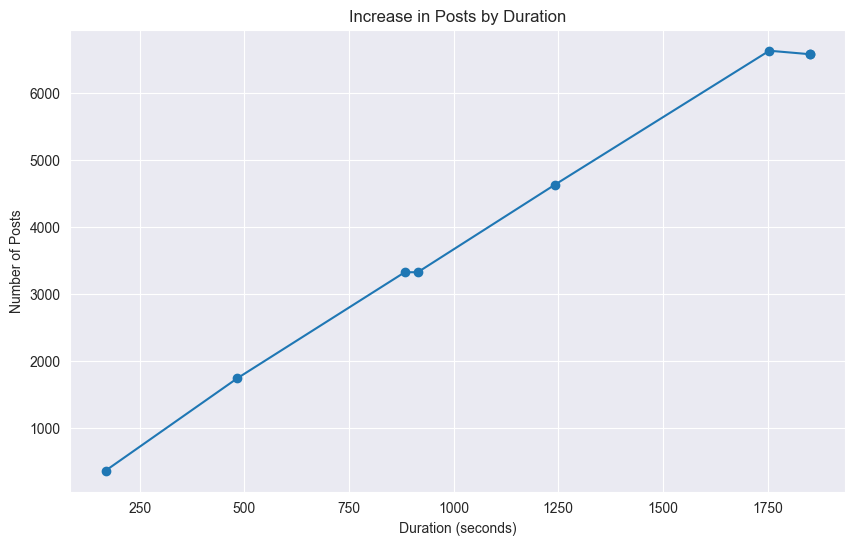


Figure 5 - Time efficiency Reddit API

## Statistical Analysis

A number of statistical analysis were performed on the datasets in this research, they include:

* Construction Trends
* Statistical Tests
* Compound Annual Growth Rate (CAGR)

### Construction Trends

Note that additional charts and analysis can be viewed in the statistics notebook.

#### Construction/Dwelling Value

In the figure below we can see a breakdown of Construction/Dwelling value by country in Millions of Euro. Interestingly we can clearly see that Spain and Greece have the highest Construction and Dwelling value. For Ireland we can see that our Construction value is almost twice as much as our Dwelling value, this could indicate more value being provided in non-dwelling properties such as factories and offices. This lines up with a report by Global Data featuring the chart in figure 7 below where we can see an excerpt from the report showing a chart with a decrease in Residential construction in the years 2020 to 2022. (Global Data, 2022)

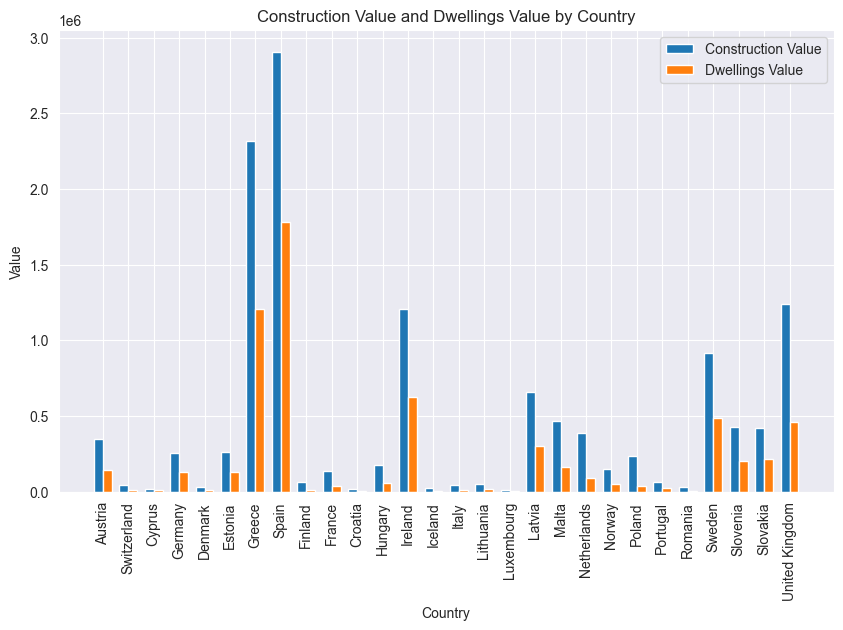


Figure 6 - Construction/Dwelling Value by country

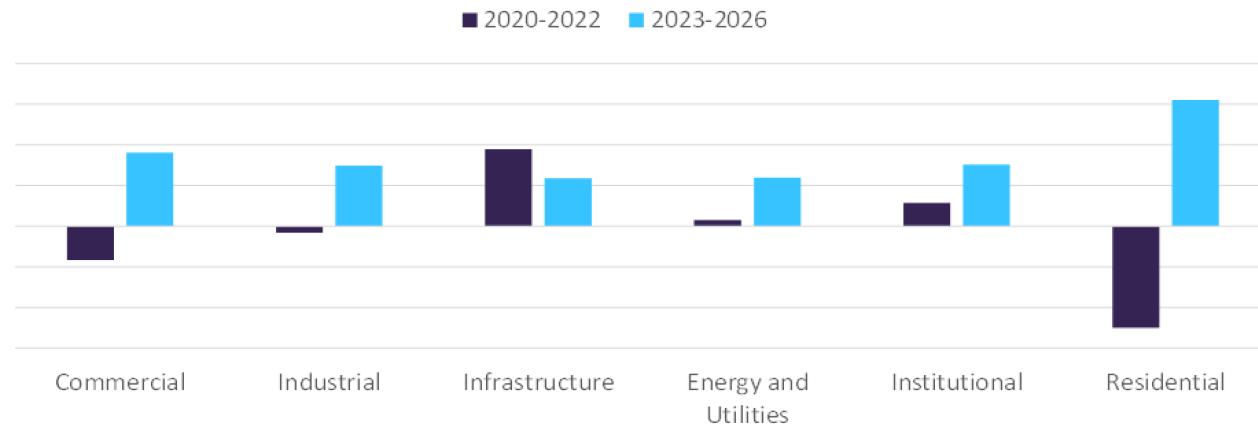
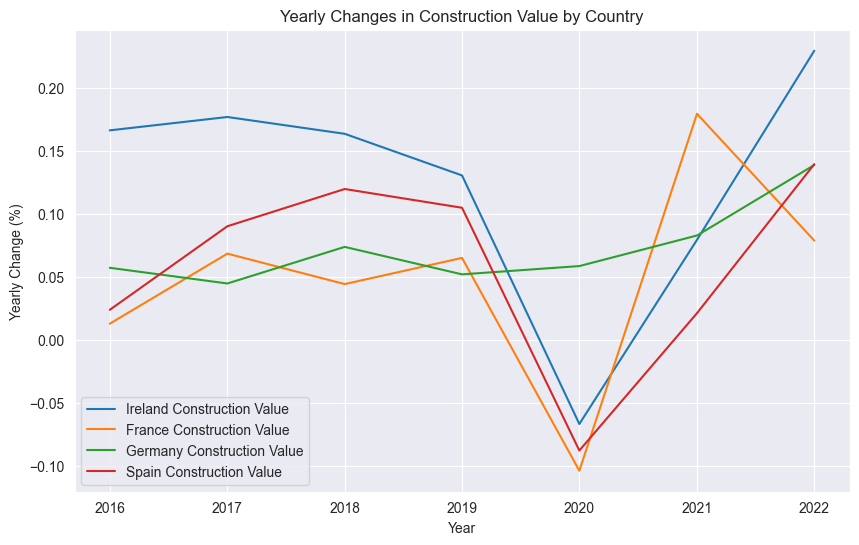


Figure 7- Global Data Report - Construction Output

Further to this analysis in figure 8 below you can see a chart for percent change in construction value for Ireland, France, Germany and Spain. Interestingly enough all countries experienced a decline in the year 2020, possible due to the Covid 19 pandemic. It’s clear to see impact the pandemic had on construction value across these countries.



### Statistical Tests

Dozens of statistical tests were done throughout this assignment. These included the below tests:

* T-Test
* Spearman's rank correlation test
* Kruskal-Wallis test
* One way ANOVA
* Wilcoxon signed-rank test

#### Note

For a full review of all statistical tests please the stats notebook.

#### Parametric vs Non Parametric

In order to distinguish which tests were statistically valid prior to any statistical test 3 assumptions were defined, and 2 were calculated using functions. The 3 assumptions were:

* All variables are independent – assumed from data source
* Data follows a normal distribution – used Shapiro Wilk test in the check\_normality function to confirm
* Data has an equal variance – used Levenes Test in the check\_variances function to confirm

Only if the data passed these 2 tests (the independence is assumed) then a parametric test was possible e.g. a T Test, otherwise a non-parametric test was used e.g. Kruskal Wallis test. It is noted that transforming the data to use a parametric test is possible but was not pursued due to time constraints.

#### T-Test

Several T-Tests (approximately 15) were done on the variables from the STS Dataset: Total Employment and Total Enterprise Production.

Here is the T-Test for Ireland and Sweden on the Total Enterprise Production variable:

**Null Hypothesis (H0):** No difference in mean between Ireland and Sweden.

**Alternative Hypothesis (H1):** A significant difference exists in mean between Ireland and Sweden.

**P-Value:** The obtained p-value is 2.31e-10

**Alpha:** 0.05, meaning we tolerate a 5% chance of wrongly rejecting H0

**Result:** The t-statistic is -12.59, p-value is 2.31e-10. The negative t-value indicates that Ireland mean is lower than Sweden.

**Conclusion:** We reject H0. A significant difference in the variable exists between Ireland and Sweden, with Ireland having a lower mean.

#### Spearman's rank correlation test

A correlation tests was performed on the variables Total Employment and Total Enterprise Production.

**Null Hypothesis (H0):** There is no correlation between the variables.

**Alternative Hypothesis (H1):** There is a correlation between the variables.

**P-Value:** The calculated p-value is 1.66e-78

**Alpha**: 0.05, meaning we tolerate a 5% chance of wrongly rejecting H0

**Result**: The calculated Spearman's rho is 0.86, which is less than the critical value of 1.97. The p-value is 1.66e-78.

**Conclusion**: The rho value is less than the critical value, which means we cannot reject the H0. Therefore, there isn't enough evidence to support a correlation between the variables. This can be seen in figure 8 below that they don’t seem to have a complete linear relationship.

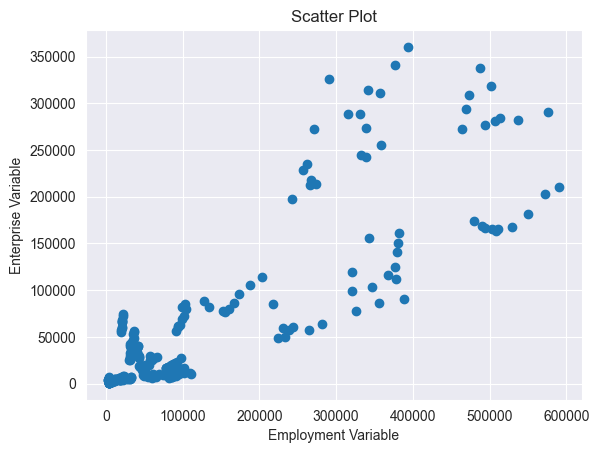


Figure 8 - Spearmans Rank Correlation Test

### CAGR

Compound Annual Growth Rate was calculated for each country on the Construction Value variable using the following formula:

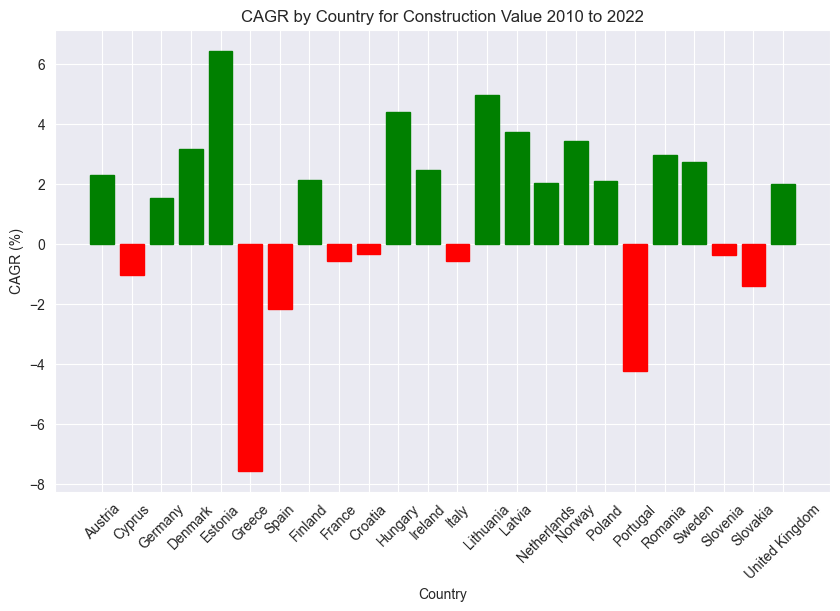


Figure 9 - CAGR for each country 2010-2022

In figure 9 above you can see that Estonia has had the largest increase meanwhile Greece has had the worst.

This analysis can help a construction company determine which countries have the strongest markets and potential growth area.

## Machine Learning

A number of machine learning models were tested on the construction datasets, these included:

* ARIMA
* SARIMA
* Linear Regression
* Lasso - Least Absolute Shrinkage and Selection Operator
* Polynomial Regression
* K Means Clustering

### Issues with the datasets

While attempting to use the ARIMA and SARIMA models for predictions on the Construction Value from the GFCF dataset, a number of issues were identified:

1. The dataset was small, with annual data we didn’t have a lot of rows in the data
2. The dataset was not stationary
3. No values of p, d or q were suitable (Autoregressive, differencing and moving average)

In summary it was found that the dataset we had was not suitable for ARIMA or SARIMA models as well as Lasso. In the future, we should gather more data at a monthly frequency and try out different machine learning models to get better predictions.

### Linear Regression

A simple linear regression model was made to predict the Gross Value Added variable using the independent variables of Year and Wages.

A GridSearchCV was used to tune the hyper parameter fit intercept and evaluate the model.

Ultimately the model provided a good result as seen in the table below, however when using the model to make predictions it was clear that a linear model is not appropriate for predicting GVA. This can be seen in Figure 11 below which shows a flat linear prediction.



Figure 10- Linear Regression Metrics

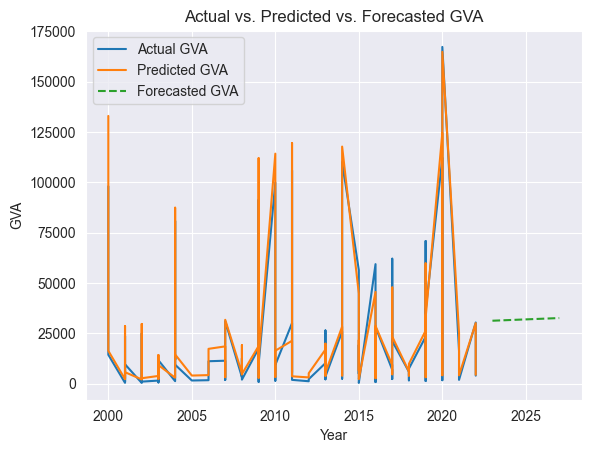


Figure 11- Linear Regression Prediction

### Polynomial Regression

After the poor prediction of the GVA using Linear Regression it was decided to try a non-linear approach using Polynomial Regression.

After using some GridSearchCV to optimize parameters degrees and include bias we had a model that provided a good prediction of the GVA with a forecast being provided assuming a growth rate of a moving percent range based on the previous data.

The resulting Polynomial metrics can be seen in figure 12 below for each country. All though it does on average provide worse predictions than the Linear Regression it is less likely to be overfitted.



Figure 12- Polynomial Regression Metrics

If you look at an example below in figure 13 below you can see the polynomial prediction as well as the forecast for Ireland, providing a decent prediction and forecast.

The forecasts have been included in the dashboard further down.

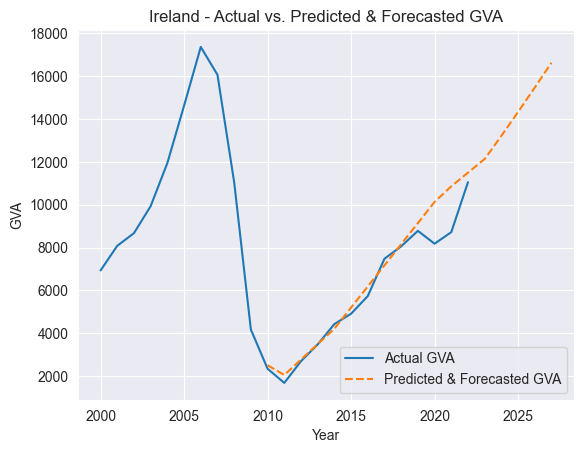


Figure 13 - Polynomial and Forecast – Ireland

### K Means Clustering

A clustering analysis was completed on the Employment and Wages indexes in the construction datasets. In order to find the optimum number of clusters an elbow plot was used which shows the WCSS (Within Cluster Sum of Squares) values. To find a balance between minimizing the WCSS (indicating tightly packed clusters) and having a manageable number of clusters for interpretation we look at the elbow chart below for a point where you get diminishing returns on the WCSS for N number of clusters.

In this example we can see that at 4 clusters the rate of return diminishes.

Therefore we used 4 as our K cluster value.

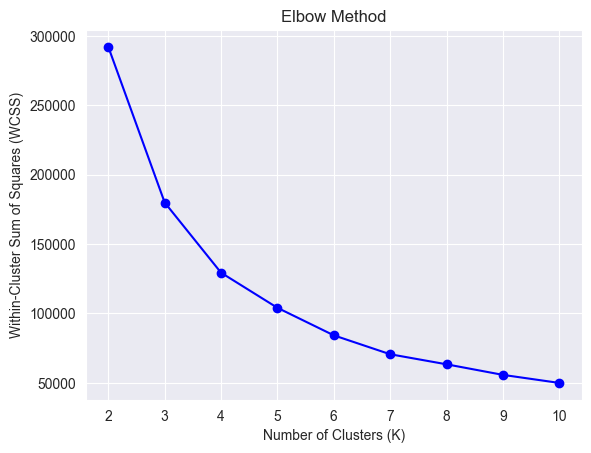


Figure 14- Elbow Chart

The resulting cluster chart can be seen in figure 15 below showing the 4 clusters in different colours. Cluster 0, the purple cluster, seems to have the lowest values for both Total Labor Employment Index and Total Wages Index indicating countries with lower employment and wages. Cluster 3, the yellow one appears to be the cluster with the highest values but seems to be relatively small and less frequent compared to the other clusters, suggesting that these high values might be outliers.

A construction company could use this analysis in a few ways, to maximize their influence in high wage and employment areas, they should target areas in clusters 2 and 3. If the goal is to expand in areas with potential for growth, targeting areas in clusters 0 and 1 might be a better approach.

Of course more analysis and data would be needed to inform this strategy, but the application is clearly advantageous to apply.



Figure 15 - Cluster Chart K = 4

## Dashboard

For this assignment a dashboard was designed to capture some useful metrics for a construction company to use. The chosen method for generating this was the dash module.

Dash is a framework in python used for building web apps designed for analytics. It's written on top of Flask and utilizes JavaScript libraries for additional functionality but is coded in Python. It's useful for creating interactive and visual dashboards that can be quite complex. It’s very flexible and allows a big range of visual charts and filters.

### Layout

The dashboard is laid out in the following way:

* Title
* KPIs
* Charts
* Filters

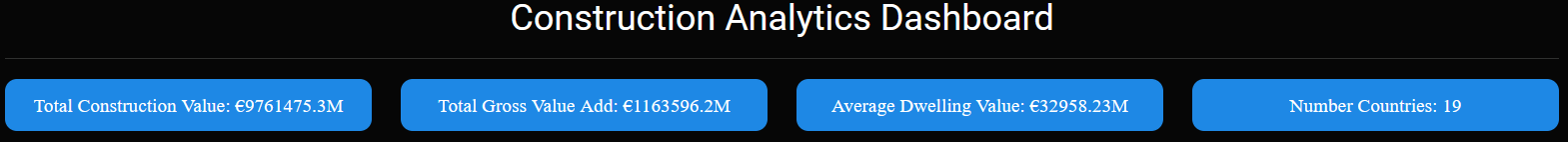


Figure 16 - Title and KPIs

Figure 16 above shows the title of the dashboard, as well as 4 different KPIs for the overall report for all countries.

* Total Construction Value
* Total Gross Value Add
* Average Dwelling Value
* Number of Countries

The KPIs were given a cool blue colour to stand out from the dark themed background and white text to clearly display the values.

A dark theme was chosen as it is easier on the eyes to view the data making it more accessible.

4 different charts were selected to be shown in the dashboard:

1. GVA and Predicted GVA – a line chart showing the total and predicted GVA by year was used as line charts are useful for showing time series.
2. Construction Value by Country – a bar chart was chosen as the countries are categorical values
3. Sentiment Analysis – a bar chart was chosen as sentiment is categorical, in addition the colours for positive, negative and neutral were given as green, red and grey respectively as they align with public understanding of colour.
4. Labor and Wages Clustering – A scatter chart here was used as it clearly highlights the spread of values of a cluster. A bright colour scheme was used to distinguish the clusters.

4 different filters are available on the chart:

* Country filter – allows the user to select/deselect countries. This works for the GVA and Construction Value charts as well as the KPIs.
* Year filter – allows the user to alter the year range for the KPIs and the Construction Value
* Top 5 – this filter allows the user to view the top 5 countries construction value
* Sentiment filter – allows the user to turn on/off sentiment topics