# Group ID - MSc in Data Analytics

Author: S. Clifford

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

Student ID: 23100

Contents

[Group ID - MSc in Data Analytics 1](#_Toc135085795)

[1 Abstract 3](#_Toc135085796)

[2 Introduction 3](#_Toc135085797)

[3 Data Preparation and Visualization 3](#_Toc135085798)

[3.1 Acquiring Raw Data 3](#_Toc135085799)

[3.2 Exploratory Data Analysis 3](#_Toc135085800)

[3.3 structure and enrich your data 3](#_Toc135085801)

[3.4 interactive dashboard 3](#_Toc135085802)

[4 Statistics 4](#_Toc135085803)

[4.1 summarise the dataset 4](#_Toc135085804)

[4.2 inferential statistics 4](#_Toc135085805)

[4.3 parametric and non-parametric inferential statistical techniques 4](#_Toc135085806)

[4.4 challenges 4](#_Toc135085807)

[5 Machine learning for data analytics 4](#_Toc135085808)

[5.1 Describe the rationale and justification for the choice of machine learning models 4](#_Toc135085809)

[5.2 Perform a sentimental analysis for an appropriate construction topic 5](#_Toc135085810)

[5.3 learning models 5](#_Toc135085811)

[5.4 similarities and contrast of the Machine Learning modelling 5](#_Toc135085812)

[6 Programming 5](#_Toc135085813)

[6.1 Programming 5](#_Toc135085814)

[6.2 Data structures 6](#_Toc135085815)

[6.3 Testing & Optimisation 6](#_Toc135085816)

[6.4 Data manipulation 6](#_Toc135085817)

[7 Conclusion 6](#_Toc135085818)

[8 Next Steps 6](#_Toc135085819)

[9 References 7](#_Toc135085820)

# Abstract

# Introduction

# Data Preparation and Visualization

## Acquiring Raw Data

As per the Scenario presented for this report 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 was inported in CSV format into Jupiter notebook ‘ConstructionData.ipynb’ (Clifford, 2023) section 2.1 and converted to pandas dataframe.

First look at the dataset ‘ConstructionData.ipynb’ 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 df information

From Figure 1 and from Jupiter notebook(Clifford, 2023) 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.

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 ConstructionData 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 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 constructionData 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 Missing last values in df

A check for null values was already done as part of section 3.2, 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, this library assignes 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 ot at 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.

## interactive dashboard

**Modern construction has a great dependence on technology and relies upon visualizations to communicate information, this includes web based, mobile based and many other digital transmission formats. Develop an interactive dashboard tailored to modern farmers, using tufts principles, to showcase the information/evidence gathered following your Machine Learning Analysis. Detail the rationale for approach and visualisation choices made during development. Note you may not use Powerbi, rapidminer, tableau or other such tools to accomplish this (at this stage).[0-30]**

# Statistics

## summarise the dataset

**Use descriptive statistics and appropriate visualisations in order to summarise the dataset(s) used, and to help justify the chosen models. [0-20 ]**

### 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 3 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 2 Processed df Information.

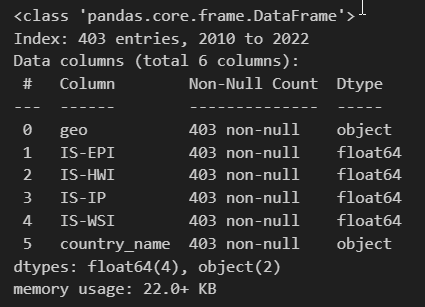


Figure Processed df Information

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 outliers. 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 Processed data information for numerical values

Looking at the distribution of the 4 indicator columns, in Figure 3 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.

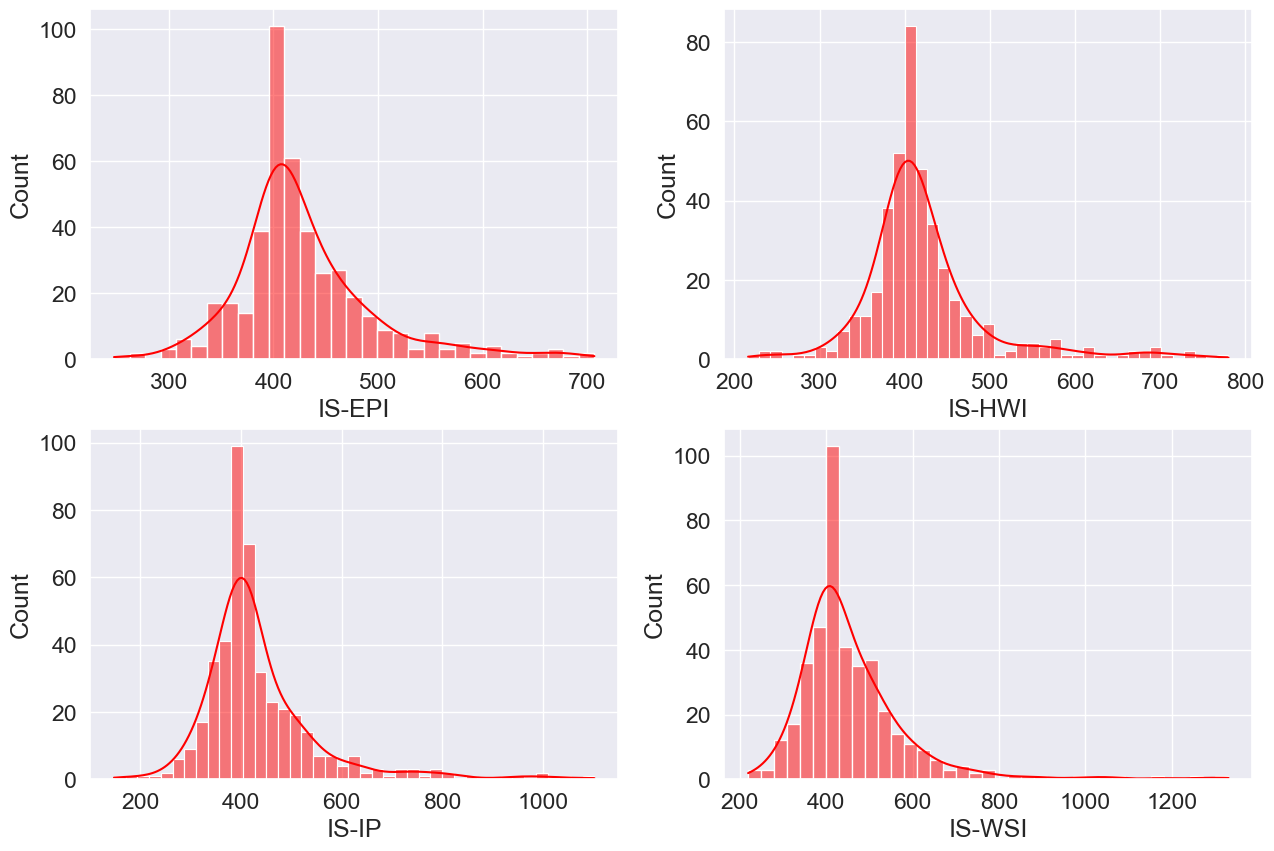


Figure Histograms displaying distribution of indicator values

Further to the observation as seen in Figure 3, looking at the boxplot plots for the 4 indicators in Figure 4, the presence of outliers predominantly to the right, confirms this tendency.

A picture containing text, screenshot, line, diagram

Description automatically generated

Figure Boxplot plot for indicator values

Quantifying the number of outliers for each indicator, see notebook section 7.2.1, results in Figure 5, we can see that the number of outliers are 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.

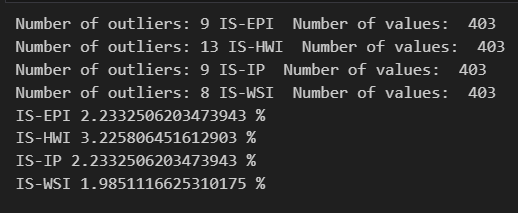


Figure Number of outliers for indicator values

Further testing of indicator data across all countries shows that none of the 4 indicators provided normal distribution, as was calculated by a 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 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 level of proportion per quarter in the construction 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 quarter
* 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 Production index data for Ireland for each quarter will fall with the confidence intervals (CI) for each quarter in Table 5. Also with 95% confidence we can say that Ireland has a production index which is not above the European average

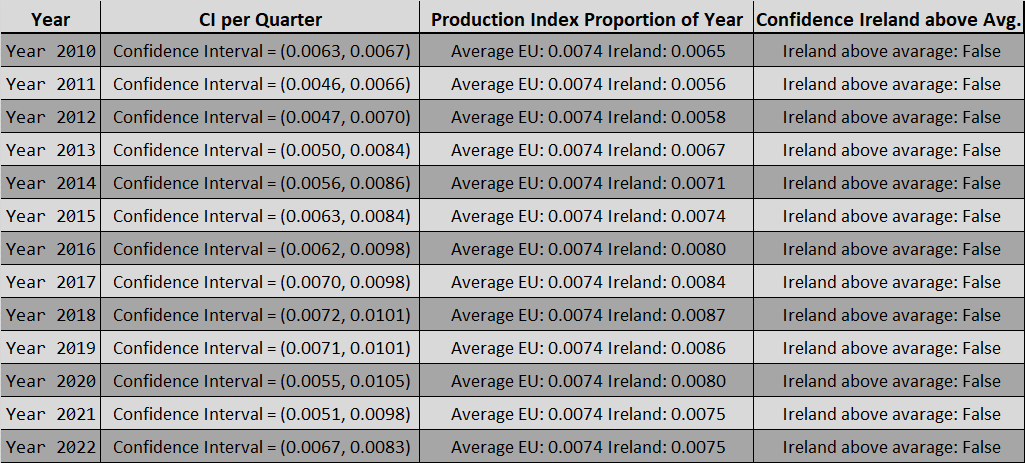


Table Population data results

## parametric and non-parametric inferential statistical techniques

**Undertake research to find similarities between some country(s) against Ireland and apply parametric and non-parametric inferential statistical techniques to compare them (e.g., t-test, analysis of variance, Wilcoxon test, chi-squared test, among others). You must justify your choices and verify the applicability of the tests. Hypotheses and conclusions must be clearly stated. You are expected to use at least 5 different inferential statistics tests. [0-40]**

Hypothesis Testing: 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. Commonly used tests include t-tests for comparing means, chi-square tests for testing independence or goodness-of-fit, and ANOVA tests for comparing means across multiple groups.

### Independent T-Test

## challenges

Use the outcome of your analysis to deepen your research. Indicate the challenges you faced in the process. [0-20]

# Machine learning for data analytics

## Describe the rationale and justification for the choice of machine learning models

Describe the rationale and justification for the choice of machine learning models for the above-mentioned scenario. Machine Learning models can be used for Prediction, Classification, Clustering, sentiment analysis, recommendation systems and Time series analysis. You should plan on trying multiple approaches (at least two) with proper selection of hyperparameters using GridSearchCV method. You can choose appropriate features from the datasets and a target feature to answer the question asked in the scenario in the case of supervised learning.

[0 - 30]

## Perform a sentimental analysis for an appropriate construction topic

Collect and develop a dataset based on the construction topic related to Ireland as well as other parts of the world. Perform a sentimental analysis for an appropriate construction topic (e.g., house price, availability of labour etc…) for producers and consumers point of view in Ireland.

[0 - 25]

## learning models

You should train and test for Supervised Learning and other appropriate metrics for unsupervised/ semi-supervised machine learning models that you have chosen. Use cross validation to provide authenticity of the modelling outcomes. You can apply dimensionality reduction methods to prepare the dataset based on your machine learning modelling requirements.

[0 - 30]

## similarities and contrast of the Machine Learning modelling

A Table or graphics should be provided to illustrate the similarities and contrast of the Machine Learning modelling outcomes based on the scoring metric used for the analysis of the above-mentioned scenario. Discuss and elaborate your understanding clearly.

[0 - 15]

# Programming

## Programming

Programming: The project must be explored programmatically: this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. [0-20]

## Data structures

Data structures: You are required to gather and process data that has been stored in at least two distinct formats. For example, this can be data in a CSV file, from a MySQL database or from a web API in JSON format. [0-20]

## Testing & Optimisation

Testing & Optimisation: You are required to document and evaluate a testing and optimisation strategy for your analysis. As part of this, you may want to plan and document how you ensured your code is doing what it is meant to, as well as ensuring that the code is making good use of your resources (eg computing, time etc). Note any trade-offs that you've made in these areas. [0-20]

## Data manipulation

Data manipulation: For each of the different data sources, compare and contrast at least two relevant libraries and techniques for a) processing and b) aggregating the respective data, in order to justify your chosen libraries/techniques. [0-20]

# Conclusion

# Next Steps

# References