**MSc in Data Analytics – “Comparative Analysis of Construction Sectors in Ireland, Portugal and Sentiment Analysis on Redditt API Comments- A Data Mining Approach.”**

Omprakash Vootla

7 Parklands Row, Parklands

Saggart, Dublin 24 Ireland.

Email: [sbs23020@student.cct.ie](mailto:sbs23020@student.cct.ie)

Student Number: Sbs23020

***Abstract***

*Objective of this Data Mining project is to compare the Construction sector in Ireland and other countries. This study conducted on multiple datasets,*

*first dataset contains the information of average earnings per Quarter in construction industry of various employment types in Ireland,*

*Second dataset contains the information of Number of employees with different skill level of employees per quarter in Portugal construction industry.*

*Third and fourth datasets are web scraped comments from Reddit API Posts on Ireland and Australia construction industries.*

*For programming requirements implemented Python programming tools and libraries used on Jupyter notebooks. For storage and easy retrieval of data, data frame is exported to MYSQL database table.*

*Data sets are loaded on to pandas data frames and performed data cleaning and data preparation steps to ensure data quality and suitability for further analysis.*

*Data preparation involving validating data structures, renamed columns, removed unwanted columns. And converting data types to integers, handling string formatting issues like removing commas, numerical values in the data. identifying missing values and fixing them with imputation methods.*

*EDA (Exploratory Data Analysis), data processing, and using pandas and seaborn visualisation libraries Implemented Scatter plots to understand the trends and patterns in earnings, employment, over the time. Using Pearson correlation method and lasso regression methods identified important Features. Also implemented Principal component analysis to identify principal component of Ireland data frame.*

*Descriptive Statistics gain the insights from data as measure of central tendency, skewness, and kurtosis, Using IQR method outliers are identified and removed from the data frame.*

*Inferential Statistics, Confidence interval, Parametric and non-Parametric tests applied on various Hypothesis tests.*

*Machine Learning models applied on both data frames, Regression, Classification, and Clustering algorithms applied and measured model accuracies. Also, time series analysis applied to capture patterns and trends in the data. ACF and PACF plots generated to identify the presence of the autocorrelation and appropriate lag for time series forecasting.*

*Data visualisation addressed throughout the analysis where ever visual representation is required. Included Interactive Scatter plots, bar plots, boxplots, and Created scatter plot using Plotly express.*

***Introduction***

Ireland is facing a housing crisis at this moment of time, there has been lot of construction going on all over the country to meet the housing requirement. This study focused on finding insights from datasets of employee’s average earnings per quarter in Ireland construction Industry over the period, compared to Portugal number of employment per quarter with different skill levels over the period. This analysis includes various data mining approaches like Data Cleaning, Data Preparation, Exploratory Data Analysis and Machine learning Models including Time series and sentiment Analysis on Reddit API post comments related to construction industry.

Ireland dataset is derived from Data.Gov.ie, CSO in the Construction domain.

Portugal filtered dataset is derived from <https://ilostat.ilo.org/data/> in Labour force statistics.

Both datasets have the labelled data, with one time series columns year and quarter. Regression models will be best fit for labelled data. We explore regression models like Linear Regression, Decision Tree, and Random Forest and Classification models like Logistic Regression, Decision Tree, Random Forest Classifier and Support vector machine models to compare the results and find best model.

To apply Machine learning models, we split the data set as train and test sets with 80%-20%. We evaluate the results by various methods like, MSE, R-Square, F1-Score, Accuracy and Confusion Matrix etc.

Table of Contents

[**1.** **Scope and Methodology:** 5](#_Toc136509904)

[1.1 Business Understanding 5](#_Toc136509905)

[1.2 Data Understanding and Data Preparation 6](#_Toc136509906)

[1.3 Data Import to MySQL Database Table using Python. 7](#_Toc136509907)

[1.4 EDA, Features Selection and Outliers detection 8](#_Toc136509908)

[1.4.1 Pearson Correlation heat map 9](#_Toc136509909)

[1.4.2 Outliers detection using IQR Method. 9](#_Toc136509910)

[1.5 Descriptive Statistics 10](#_Toc136509911)

[1.5.1 Distribution of data between “Year” with other variables. 10](#_Toc136509912)

[1.6 Inferential Statistics 12](#_Toc136509913)

[1.6.1 Parametric Tests on Quarterly earnings in Construction Industry in Ireland 13](#_Toc136509914)

[1.6.2 Non-Parametric Tests on Quarterly earnings in Construction Industry in Ireland 14](#_Toc136509915)

[1.6.3 Parametric Tests on Number of Employees per Quarter in Construction Industry of Portugal. 15](#_Toc136509916)

[1.6.4 Non-Parametric Tests on number of employees per Quarter in Construction Industry in Portugal*.* 17](#_Toc136509917)

[1.7 Challenges faced in this process. 18](#_Toc136509918)

[1.8 Machine Learning Modelling 19](#_Toc136509919)

[1.9 Time Series Analysis 21](#_Toc136509920)

[1.10 Sentiment Analysis on Construction Crisis in Ireland and Australia. 23](#_Toc136509921)

[1.10.1 Reddit API Data Collection Process. 23](#_Toc136509922)

[1.10.2 Summary of Sentiment analysis. 23](#_Toc136509923)

[1.11 Programming Tasks 25](#_Toc136509924)

[1.11.1 Programming: 25](#_Toc136509925)

[1.11.2 Data Structures: 25](#_Toc136509926)

[1.11.3 Documentation: 25](#_Toc136509927)

[1.11.4 Testing and Optimisation: 25](#_Toc136509928)

[1.11.5 Data Manipulation: 25](#_Toc136509929)

[2.0 GitHub: 26](#_Toc136509930)

[3. References 27](#_Toc136509931)

# **Scope and Methodology:**

For this study we are going to implement CRISP-DM Methodology. CRISP-DM is well-established and widely used for data mining and Machine learning projects. This framework consists of six phases.

Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, Deployment.

Diagram

Description automatically generated

## 1.1 Business Understanding

To answer the give task our first challenge was to find the right data sets. This task takes longer time than expected. The challenge is to find data sets in construction industry of Ireland and compare the insights with any other country of the world. Spent number of days in finding the suitable data sets for the study, finally come up with two data sets related Ireland and Portugal construction employees’ earnings and number of employees per quarters from years 1969-2008. this dataset derived from publicly available on <https://www.cso.ie/en/databases> Central Statistics Office (CSO), Ireland. Second dataset is about the Portugal Quarterly Employment in construction industry by gender. This dataset is derived from International Labour Organisation <https://ilostat.ilo.org/data/> .

## 1.2 Data Understanding and Data Preparation

**Jupyter Notebook 1:** “1. Ireland Avg Earnings per Quarter and Portugal Avg Number of employees per Quarter.”

**Jupyter Notebook Sections 1-3:** we explore data collection, data cleaning and data preparation part.

As first step we imported various libraries to perform data analysis on this data set like Pandas, NumPy, Seaborn, Matplotlib etc.

To understand the data we performed head, tail, shape, info, Isnull.sum, describe, values count, duplicated and memory usage, etc functions were applied.

For data cleaning and preparation, we implemented as type, rename, drop, ffill, backfill, dropna, etc functions were applied.

This data frame has 160 rows, and 10 columns.

Graphical user interface

Description automatically generated  
This data frame shows each variable data type as below, we have two float types and rest of them are object types.

A picture containing graphical user interface

Description automatically generated

we renamed columns as python friendly. For example, renamed 'Unnamed: 0': 'Year\_Quarter' etc. than split the “Year\_Quarter” in to two new columns “Year” and “Quarter”. And dropped the original “Year Quarter”.

A picture containing graphical user interface

Description automatically generated

Found 39 missing values in few columns, which are less than 20% the total data, so we imputed them with function “ffill” and “backffill”.

Next step replace “,” “” , and converting all variable data type as integers.

We export cleaned data frame to MySQL data base table “avg\_earnings\_week”. First, we convert all rows into a list of tuples and save as .txt file.

After data cleaning and preparation sample data shown as below:

Graphical user interface

Description automatically generated with medium confidence

Next step, using pandas MySQL connector we connect to the database. And created a database and table with columns as exactly as in our data frame. Now log in to MySQL and connect to data base table and using insert statement. We load entire .txt file into the table. Now we can use this database table to user for our EDA and ML modelling.

We have implemented same steps to prepare the Portugal Quarterly Number of Employees in Construction Industry, datasets apply Machine Learning Models and Sentiment Analysis.

## 1.3 Data Import to MySQL Database Table using Python.

**Jupyter Notebook Section 4:**

we export cleaned and prepared data to MySQL data base table. Before loading the data into the table, we converted data frame as a list of tuples and save the list as a .txt file.

We have implemented python code to create database and table with columns same as in our data frame. after creating the database and the table in MySQL go to .txt file and copy all the data and go to MySQL data base table and create an Insert INTO table values query and paste all the data and run it. This process is simple and easy to load and retrieve data.

## 1.4 EDA, Features Selection and Outliers detection

**Jupyter Notebook Section 5:**

EDA - Ireland-Avg Earnings per Week in Construction Industry.

1. Connecting to MySQL data base allow us to access the cleaned data stored in “avg\_earnings\_week” table created in data extraction step.
2. Verifying the number of rows and columns, data types, allowed to identify whether any data conversions or handling is required.
3. Descriptive statistics provide key insights in to central tendencies dispersions, and range of the variables.
4. Scatter plots observe trends and identify any patterns or correlations between the variables and years. This step helped in understanding the earnings of different employees over the period.
5. Correlation matrix allowed to understand the relationship between different variables in the data frame. with Heat map visualization we can identify highly correlated variables. This step allows us to identify any multicollinearity issues and understand the interdependencies between the variables.
6. To address the multicollinearity and reduce dimensionality we perform Feature selection. We can identify highly correlated features and drop from data frame. By doing this we ensured that the selected features were not redundant. We identified key features in predicting manual employee grades.

Scatter plot indicates that there is spike in the average earning of “Manual\_employee\_grades” and “Clerical\_and\_Operatives” in years 1980-1985. Fallowed by “Semi skilled adults” and “Skilled operatives”. Over all high earnings for “Manual\_employee\_grades” and “Clerical\_and\_Operatives” for entire time periods.

Chart, histogram

Description automatically generated

### 1.4.1 Pearson Correlation heat map

Graphical user interface, chart

Description automatically generated with medium confidence

### 1.4.2 Outliers detection using IQR Method.

We identified outliers IQR method and visualise them as box plot as below.

Chart, box and whisker chart

Description automatically generated

There are few outliers in Manual employee grades, Skilled Operatives, Apprentices, Semi-skilled adults and Clerical and Operatives. No outliers found others. We can understand as construction industry is more adults and male workers domain. Some time they need to work overtime, so they get paid overtime payment as double. We found them 14 data point overall, after removing those outliers. Store the dataset with new name “df1\_no\_outlier”.

## 1.5 Descriptive Statistics

**Jupyter Notebook Section 6:**

In this section descriptive statistics provided valuable insights into the central tendencies, dispersions, and distribution of the variables.

1. We drop year and quarter variables to focus only on variables related to employee grades and earnings.
2. Descriptive statistics provide statistical summary of variable metrics such as mean, median, mode, standard deviation, min, max, range and distribution of data.
3. Comparing mean and median allowed to assess the skewness of data distribution. It is identified there is significant differences in the mean and median. This indicates skewness in the data distribution.
4. Standard deviation measured how spread out the values from mean. “Apprentices”, “Clerical employees female”, “Clerical employees male” and “Foremen and supervisor” means are closer to the standard deviation compared to other variables.
5. Skewness and Kurtosis provide insights into shape of the data distribution, “Manual employee grades”, “Skilled operatives” shows positive skewness and ”, “Clerical employees female” and “Foremen and supervisor” shows negative skewness.
6. Data distribution plots allow to observe trends, variations, potential outliers and any significant changes or patterns in the data.

### 1.5.1 Distribution of data between “Year” with other variables.

#### 1.5.1.1 Year Vs Manual employee grades

Chart, bar chart, histogram

Description automatically generated

#### 1.5.1.2 Year Vs Skilled operatives

Chart, bar chart, histogram

Description automatically generated

#### 1.5.1.3 Year Vs Apprentices

Chart, bar chart, histogram

Description automatically generated

#### 1.5.1.4 Year Vs Clerical\_employees\_female

Chart, bar chart, histogram

Description automatically generated

#### 1.5.1.5 Year Vs Foremen\_and\_supervisors

Chart, bar chart, histogram

Description automatically generated

Descriptive Stats of each variable:

Table

Description automatically generated

Table

Description automatically generated

Observations: There is a big difference in earnings of Clerical employee’s male and Clerical and Operatives, Manual employee grades, Semi-skilled-adult’s earnings are high in construction industry. Over all data distribution is Moderate to slightly moderate positive skewed.

## 1.6 Inferential Statistics

**Jupyter Notebook Section 7:** We explored Ireland dataset inferential statistics and

**Jupyter Notebook Section 12: We** explored Portugal data set inferential statistics.

“Inferential statistics are statistical tests for testing statistical significance of results.

“Statistical significance means the likelihood that a given set of results being obtained if the null hypotheses were true ” (bps.com)

This analysis focuses on key metrics from Ireland and Portugal data sets such as average earnings of employee type, number of employees, the association between gender and clerical support worker, difference in number of employees across quarters and skill levels.

### 1.6.1 Parametric Tests on Quarterly earnings in Construction Industry in Ireland

Below table display 5 Parametric tests carried on different variables of dataset to find the Hypotheses.

Table

Description automatically generated

Table

Description automatically generated

### 1.6.2 Non-Parametric Tests on Quarterly earnings in Construction Industry in Ireland

A screenshot of a computer

Description automatically generated with low confidence

Table

Description automatically generated

### 1.6.3 Parametric Tests on Number of Employees per Quarter in Construction Industry of Portugal.

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

### 1.6.4 Non-Parametric Tests on number of employees per Quarter in Construction Industry in Portugal*.*

Table

Description automatically generated

Table

Description automatically generated

Observations:

We can find some insights from Ireland and Portugal Construction Industry Earnings and Number of employees per quarter over the years.

* Average Earnings of Apprentices is 1588
* Earning of “Clerical\_employees\_female”. Is equal in Q1 & Q4.
* Female employees Earnings per Quarter is between 717-792.
* There are significant differences between the earnings per quarter.
* There is significant association between number of female and Clerical\_support\_workers employees.
* the average Number\_of\_employees 463, Quarterly.
* Number\_of\_employees per Quarter is between 432 and 494.
* Q1 Number\_of\_employees and Q4 Number\_of\_employees are equal.
* The median Number\_of\_employees per Quarter in construction industry are the same for the years between 2011Q1 and 2020Q4.
* There is no significant difference between number of employees per quarters.
* There is no difference Number of “Skill\_levels\_high” employees and Number of “Skill\_levels\_low”.

## 1.7 Challenges faced in this process.

1. Data Collection:

Bigger challenges in acquiring the data sets on construction data sets, as very limited data sets are publicly available. More time spend on finding the data set with reasonable number of labelled data and some time series and a feature for predictions.

Some of the tasks we need to compare insights from Ireland with other countries. Identifying other countries data sets become more challenging.

1. We need to address data quality issues, identify inconsistencies, missing values, outliers, requires data preparation on both data sets.
2. Sample sizes, similar labelled data is challenging to find target variables, requiring feature selection techniques to identify important features for further analysis.
3. Data sets has only few common columns like Year and Quarter and employee type. But one data is earnings, and one data set is number of employees per quarter. There is no common point to compare. Different employee columns were created from Occupation column in Portugal data set. Gender column also split in to male female columns.
4. Data sets has no specific target variable, so each test has been used different variables, interpreting the results from Hypothesis is become Challenging in inferential statistics implementation.

## 1.8 Machine Learning Modelling

**Part 1: Ireland Data set**

**Jupyter Notebook Section 8:**

We explored Selection of Machine learning Algorithms:

There are Supervised, unsupervised, Reinforcement learning and semi supervised learning methods in Machine learning algorithms.

* Supervised learning approach is where a model is trained on labelled data. In supervised learning model learn from the data and make predictions.

For Ireland Avg Earnings per Quarter dataset is a labelled dataset, for prediction of a specific employee for example “Skilled\_operatives” earnings. We have input features and output features are labelled data, so we apply supervised learning algorithms like Linear Regression, Logistic Regression, Decision Trees, Random Forest, Support vector machines (SVM) and K-Nearest Neighbours (KNN) etc.

Linear Regression model predicts the target variable using input variables. Captures trend and patterns in time series data. Measured with Root mean squared error value of 95% indicating an average deviation in the predicted values from the actual values.

Decision Trees model obtained an RMSE of 330.121 represents the average deviation in the decision tree predictions from actual values. The R-Squared value indicates the goodness of fit of the model. Explain the percentage of variance in the target variable.

Random forest model is an ensemble method using multiple decision trees for predictions.

Comparing the model performance, we observe that the random forest achieved lowest RMSE and highest R-Squared values. Indicating best performance model. based on the evaluation metrics random forest model predict the target variable “Skilled Operative ” in the data set.

Chart, bar chart

Description automatically generated

**Part 2: Portugal data set**

**Jupyter Notebook Section 14:**

In this section we explore classification models on Portugal Data set, this dataset has number of employees per Quarter in Construction industry.

We have applied Logistic Regression, K-Nearest Neighbours, Decision Tree Classifier, and Random Forest algorithms with different target variables for each model the accuracy is shown as bar plot as below.

Chart, bar chart

Description automatically generated

Logistic Regression model confusion matrix shows that there were 37 true negatives (TN), 30 false positives (FP), 30 false negatives (FN), and 38 true positives (TP). With the accuracy of 56%.

Graphical user interface

Description automatically generated with low confidence

Random forest and decision trees classifier models give the best accuracy with 100% and 99%, and KNN accuracy at 92% in predicting the number of employees per quarter in Portugal construction industry.

More models also evaluated in Jupyter note book like SVM, PCA and Market Basket analysis but those models did not perform well on this data set. It is identified from Based on Market Basket analysis that Clerical\_support\_workers male is more frequent employees and male skill level high and skill level low employees.

## 1.9 Time Series Analysis

**Jupyter Notebook Section 9:**

For implementing time series analysis, we converted the “Unnamed: 0” column in the initial data set with year and Quarter(1969Q1) to datetime format.

We apply Time series analysis on “Clerical\_employees\_female” Quarterly earnings in Ireland construction industry.

We find the female employee’s earnings trends as below. From year 1969-1974 growth in quarterly earnings, then drop from 1975-1980, and sudden raise of earnings in 1980-1982, from 1988 again growth in earnings till 2004.

Chart, line chart

Description automatically generated

We apply moving average to smoothing the time series, below plot shows the difference.

Chart, line chart

Description automatically generated

We Perform Augmented Dickey-Fuller (ADF) test to check the time series stationary, and with P-Value : 0.43283936290703573 > 0.05; Data is not stationary.

We create Naïve Base model be shifting time series by one step and created “Clerical\_employees\_female\_Forecast” column and calculated MSE and RMSE bet ween the actual and forecast columns. Root mean squared Error (RMSE) = 67.28.

We apply ARIMA Model and plot “acf” is for evaluate Auto Regressive and “pacf” is for moving average.

Chart, bar chart, histogram

Description automatically generated

Graphical user interface, table, Excel

Description automatically generated

ARIMA model and calculated RMSE = 0.48, when compared to Naïve Base Model RMSE = 67.28. ARIMA model performs better in forecasting “Clerical\_employees\_female” earnings.

Chart, line chart

Description automatically generated

## 1.10 Sentiment Analysis on Construction Crisis in Ireland and Australia.

**Jupyter Notebook 2: Sentiment Analysis on Reddit API Post comments.**

### 1.10.1 Reddit API Data Collection Process.

For data collection process we used Reddit API post, we did a random search on construction crisis in Ireland and Australia, to access the Reddit API we import “praw” library and supply reddit API credentials to access using python code. An instance of Reddit has created. Used a python function to take the list of comments and iterated over each comment and copy comment body, author, score, upvotes, and downvotes to a csv file.

### 1.10.2 Summary of Sentiment analysis.

Sentiment Analysis was performed on Reddit API posts related to construction crisis in Ireland and Australia. The analysis aimed to understand the sentiment expressed in comments and compare the sentiment between the two countries. Data was scraped from Reddit and stored as CSV files. TfidfVectorizer and VADER method applied in this analysis.

Data was pre processed by removing missing values, text data was cleaned by removing punctuations, special characters, numbers, and converted text in to lowercase. TfidfVectorizer applied to break down the text in to words or tokens. Cleaned data split as train and test sets.

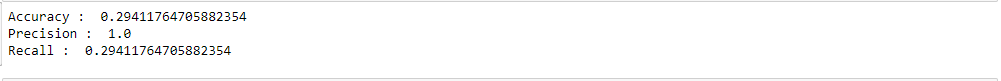
Sentiment analysis with TfidfVectorizer:

Logistic Regression was used as classification algorithm. Model was trained on train set and evaluated using accuracy, precision, and recall methods. Trained model was used to predict sentiment on test set. Accuracy of the model was calculated to assess the model performance.

**Ireland Model accuracy (47%), Precision (1.0) and Recall (0.47).**



**Australia Model accuracy (29%), precision (1.0) and recall (0.29).**



**Sentiment analysis with VADER:**

VADER method used to determine the sentiment of the comments. Polarity scores (Positive, neutral, negative, and compound) were calculated for each comment. Scores were visualised using bar plots to show the distribution of sentiment.

**Ireland Comments Sentiment scores:**

Graphical user interface, application

Description automatically generated

**Ireland Comments Sentiment scores visualisation:**

Chart, bar chart, histogram

Description automatically generated

**Australia Comments Sentiment scores:**

Graphical user interface, application

Description automatically generated

**Australia Comments Sentiment scores visualisation:**

Chart, bar chart, histogram

Description automatically generated

**Conclusion:**

The analysis revealed differences in sentiment between two countries, In Ireland, there is higher proportion of negative comments compared to Australia, indicating a more critical sentiment towards the construction crisis. This resembles the current situation in Ireland. However, this analysis is subjective and influenced by various factors such as data collection and context of comments.

## 1.11 Programming Tasks

### 1.11.1 Programming:

Analysis is presented programmatically using python tools and imported necessary libraries, to load and process data set, pre-processing, splitting the data into target and independent variables. Splitting data in train and test data sets, scaling the features for machine learning modelling and for data visualisations.

### 1.11.2 Data Structures:

Analysis is presented with two data gathering and storage and retrieval methods, raw data is derived from public domains available in csv format. Ireland data set is in csv file format and this csv files read as pandas data frame in Jupyter Notebook and applied data processing methods. The clean data frame is converted as list of tuples to load in to MySQL data base table for easy storage and retrieval for further analysis. MySQL is used because of usability, scalability, schemas, and ensuring data integrity.

### 1.11.3 Documentation:

Code standards applied on python code in Jupyter Notebook by commenting each cell the purpose of the code and interpreted results in statistics and machine learning models,

Various libraries are used in coding, NumPy is used for numerical computing and array operations, mostly used for statical and machine learning algorithms.

Seaborn is used in statistical graphics creates scatter plots, line plots, bar plots histograms etc.

Matplotlib is used in static animated interactive visualisations, like pie charts, heat maps and more.

Plotly is used to create interactive visualisations with extra features like zooming, panning, hoover tooltips, and animations.

Dash is framework for building web applications and Dashboards for visualisations. Dash allows to create custom web interfaces with controls.

### 1.11.4 Testing and Optimisation:

Throughout the Analysis code testing has been performed via, code validations, for example when we impute missing values with “ffill”, we validation again with performing “info. Sum” function. likewise for each action, we revalidate in data preparation or Machine learning Phase. We validate models with different features in data set for better performance of the model.

### 1.11.5 Data Manipulation:

We explore pandas, praw, text blob, and re libraries in this analysis.

# 2.0 GitHub:

Please find the assignment repository here.

APPENDIX A

Please refer to Jupyter Notebook (.ipynb) code attached for 1. Ireland Avg Earnings per Quarter and Portugal Avg Number of employees per Quarter.

APPENDIX B

Please refer to Jupyter Notebook (.ipynb) code attached for 2. Sentiment Analysis on Reddit API Post comments.

# 3. References