**CCT College Dublin**

**Assessment Cover Page**

| **Module Title:** | Data Preparation & Visualisation, Machine Learning for Data Analytics, Statistics for Data Analytics, Programming for Data Analytics |
| --- | --- |
| **Assessment Title:** | Construction Industry in Ireland and Europe |
| **Lecturer Name:** | David McQuaid, Dr. Muhammad Iqbal, Marina Iantorno, Sam Weiss |
| **Student Full Name:** | Paul Ryan |
| **Student Number:** | sbs23013 |
| **Assessment Due Date:** | 26/05/2023 |
| **Date of Submission:** | 26/05/2023 |

**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**

*A comparison of the Irish construction industry with wider Europe including predicting future trends using time series analysis.*

**Keywords:**

**Introduction**

This is an analysis of the Irish construction industry, looking at different indicators, measured as index values, across a range of European countries. It involves loading a dataset from the Eurostat website, controlled by the European Union. After a small amount of preparation, the dataset was explored and visualised. As it was found to have quarterly values assigned to different countries and measurements, it was then tested using both parametric and non parametric statistical tests to compare the values of different countries, with a specific focus on Ireland.

Machine learning models will be created to attempt to predict future values using both time series and support vector regression models. These models will have their hyperparameters tuned, to deliver the best accuracy. The results of the two models will be compared. In addition, a sentiment analysis will be performed on text pulled from the r/Ireland and r/Europe subreddits from the Reddit discussion platform.

As part of the analysis, an interactive dashboard will be created to allow the user to see the Hours Worked Index for a selected Country, with a dynamic line chart showing both the actual past values as well as future predicted values.

There is also a discussion at the end of the report on testing and optimisation of the code and a comparison of two different data manipulation libraries in Python.

**Materials**

Overview

A .tsv file is downloaded from the Eurostat website, via API. The file is compressed using gzip so after loading it needs to be extracted using the gzip package. It is then loaded into a data frame for further modification.

In the format it is downloaded, the quarterly values are stored in separate columns. The melt function is used to combine the quarters and respective values into two columns.

There are records in the “Value” column with no value, but they are represented with a ‘:’. To enable calculations on this column, they are set to NaN, and the ‘Value’ column is then set to a data type of float, which will allow for results using decimal points such as 1.5.

As the data contains information around countries, it will likely be visualised via a map or choropleth plot. To allow for this there needs to be compatible values representing each country. The ISO3 country codes are one such format. The ‘pycountry’ package allows for the conversion of the existing two character country codes to the ISO3 three character country codes. These are mapped and assigned to a new column named ‘Country\_Codes’.

**Methods**

Exploratory Data Analysis

*Data Exploration*

M

*Visualisations*

I

Statistics

*T-Test*

T

*One - Way Anova*

T

*Two - Way Anova*

T

*Wilcoxon Signed-Rank Test*

T

*Kruskall Wallus*

T

Machine Learning - Sentiment Analysis

*Model Overview / Data Processing*

T

*Sentiment Analysis*

A

Machine Learning - Time Series Analysis

*Model Overview / Data Processing*

A

*Time Series Analysis*

A

*Hyperparameter Tuning*

A

Machine Learning Model - Support Vector Regression

*Model Overview / Data Processing*

A

*Support Vector Regression*

A

*Hyperparameter Tuning*

A

**Results**

Exploratory Data Analysis

Statistics

*T-Test*

T

*One - Way Anova*

T

*Two - Way Anova*

T

*Wilcoxon Signed-Rank Test*

T

*Kruskall Wallus*

T

Machine Learning

*Sentiment Analysis*

T

*Time Series Analysis*

T

*Support Vector Regression*

T

**Discussion**

Testing & Optimisation

Data Library Comparison

*Processing*

T

*Aggregation*

T

**Conclusion**

**References**

Al, E. (2005). *A modern introduction to probability and statistics : understanding why and how*. New York: Springer, Cop, p.64.

Central Statistics Office (n.d.). *Residential Property Price Index - CSO - Central Statistics Office*. [online] www.cso.ie. Available at: https://www.cso.ie/en/methods/surveybackgroundnotes/residentialpropertypriceindex/ [Accessed 31 Mar. 2023].

Department of Housing, Local Government, and Heritage (2022). *Tier 1 (Q1 2022 ) Sites where planning permission has been granted and the permission can be implemented immediately - data.gov.ie*. [online] data.gov.ie. Available at: https://data.gov.ie/dataset/tier-1-q1-2022-sites-where-planning-permission-has-been-granted-and-the-permission-can-be-imple?package\_type=dataset [Accessed 1 Apr. 2023].

Lander, J.P. (2021). *R For Everyone.* S.L.: Addison-Wesley.

Mckinney, W. (2017). *Python for Data Analysis, 2nd Edition*. 2nd ed. O’reilly Media, Inc, p.294.

Müller, A.C. and Guido, S. (2017). *Introduction to machine learning with Python : a guide for data scientists*. Beijing: O’reilly.

Myatt, G.J. and Johnson, W.P. (2009). *Making sense of data II : a practical guide to data visualization, advanced data mining methods, and applications*. Hoboken, N.J.: John Wiley & Sons.

Reitz, K. (n.d.). *PEP 8: The Style Guide for Python Code*. [online] pep8.org. Available at: https://pep8.org/#introduction [Accessed 8 Apr. 2023].

scikit-learn Developers (2019). *RBF SVM parameters — scikit-learn 0.21.3 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/auto\_examples/svm/plot\_rbf\_parameters.html [Accessed 2 Apr. 2023].

Severance, C.R. (2016). *Python for everybody : exploring data using Python 3*. Ann Arbor, Mi: Charles Severance, p.43.

Shai Vaingast (2009). *Beginning Python Visualization*. 2nd ed. Apress.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, [online] 5(4), p.13. Available at: https://www.academia.edu/42079490/CRISP\_DM\_The\_New\_Blueprint\_for\_Data\_Mining\_Colin\_Shearer\_Fall\_2000\_ [Accessed 7 Apr. 2023].

Stewart, W.J. (2009). *Probability, Markov Chains, Queues, and Simulation The Mathematical Basis of Performance Modeling*. Princeton University Press, p.105.

Suresh Kumar Mukhiya and Ahmed, U. (2020). *Hands-on exploratory data analysis with Python : perform EDA techniques to understand, summarize, and investigate your data smartly*. Birmingham: Packt Publishing, p.8.

The Editors of Encyclopaedia Britannica (2022). *normal distribution | Definition, Examples, Graph, & Facts*. [online] Encyclopedia Britannica. Available at: https://www.britannica.com/topic/normal-distribution [Accessed 1 Apr. 2023].

Toggerson, B. and Philibin, A. (n.d.). *Physics 132 Lab Manual*. [online] University of Massachusetts Amherst Libraries. Available at: http://openbooks.library.umass.edu/p132-lab-manual/ [Accessed 1 Apr. 2023].

Weiss, N.A. (2017). *Introductory statistics*. 10th ed. Harlow: Pearson Education Limited.

Yıldırım, S. (2020). *Hyperparameter Tuning for Support Vector Machines — C and Gamma Parameters*. [online] Medium. Available at: https://towardsdatascience.com/hyperparameter-tuning-for-support-vector-machines-c-and-gamma-parameters-6a5097416167 [Accessed 2 Apr. 2023].