**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | Data Preparation and Visualization |
| **Assessment Title:** | Industry( Construction) annual % growth |
| **Lecturer Name:** | David MCQuaid |
| **Student Full Name:** | Farhad Khankishiyev |
| **Student Number:** | 2023068 |
| **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. |

2023068 – MSc in Data Analytics

Author: Farhad Khankishiyev

e-mail: [2023068@student.cct.ie](mailto:20230%36%38@s%74u%64e%6e%74%2e%63%63%74%2e%69%65)

Student ID: 2023068

Data Access And Licensing

The World Bank Group makes data publicly available according to open data standards and licenses datasets under the Creative Commons Attribution 4.0 International license (CC-BY 4.0). Many datasets are available under other licenses. They are labeled accordingly, and when they are accessed by users, users agree to comply with all of the terms of the respective licenses, as explained below.

Creative Commons Attribution 4.0 (CC-BY 4.0)

The Creative Commons Attribution 4.0 International license allows users to copy, modify and distribute data in any format for any purpose, including commercial use. Users are only obligated to give appropriate credit (attribution) and indicate if they have made any changes, including translations. CC-BY 4.0, with the additional terms below, is the default license for all Datasets produced by the World Bank itself and distributed as open data.

All users of these Datasets under the CC-BY 4.0 License also agree to the following mandatory terms: Any and all disputes arising under this License that cannot be settled amicably shall be resolved in accordance with the following procedure:

1.Pursuant to a notice of mediation communicated by reasonable means by either You or the Licensor to the other, the dispute shall be submitted to non-binding mediation conducted in accordance with rules designated by the Licensor in the copyright notice published with the Work, or if none then in accordance with those communicated in the notice of mediation. The language used in the mediation proceedings shall be English unless otherwise agreed.

2.If any such dispute has not been settled within 45 days following the date on which the notice of mediation is provided, either You or the Licensor may, pursuant to a notice of arbitration communicated by reasonable means to the other, elect to have the dispute referred to and finally determined by arbitration. The arbitration shall be conducted in accordance with the rules designated by the Licensor in the copyright notice published with the Work, or if none then in accordance with the UNCITRAL Arbitration Rules as then in force. The arbitral tribunal shall consist of a sole arbitrator and the language of the proceedings shall be English unless otherwise agreed. The place of arbitration shall be where the Licensor has its headquarters. The arbitral proceedings shall be conducted remotely (e.g., via telephone conference or written submissions) whenever practicable

Open Database License (ODbL)

The Open Database License allows users to copy, modify and distribute data, including for commercial purposes. Users are obligated to give appropriate credit (attribution), and to provide any redistributions of data under the same license and in open formats. The World Bank distributes data under this license when required to do so by the original data provider or by the conditions of a data partnership.

Data Access And Licensing

The World Bank Group makes data publicly available according to open data standards and licenses datasets under the Creative Commons Attribution 4.0 International license (CC-BY 4.0). Many datasets are available under other licenses. They are labeled accordingly, and when they are accessed by users, users agree to comply with all of the terms of the respective licenses, as explained below.

Creative Commons Attribution 4.0 (CC-BY 4.0)

The Creative Commons Attribution 4.0 International license allows users to copy, modify and distribute data in any format for any purpose, including commercial use. Users are only obligated to give appropriate credit (attribution) and indicate if they have made any changes, including translations. CC-BY 4.0, with the additional terms below, is the default license for all Datasets produced by the World Bank itself and distributed as open data.

All users of these Datasets under the CC-BY 4.0 License also agree to the following mandatory terms: Any and all disputes arising under this License that cannot be settled amicably shall be resolved in accordance with the following procedure:

Pursuant to a notice of mediation communicated by reasonable means by either You or the Licensor to the other, the dispute shall be submitted to non-binding mediation conducted in accordance with rules designated by the Licensor in the copyright notice published with the Work, or if none then in accordance with those communicated in the notice of mediation. The language used in the mediation proceedings shall be English unless otherwise agreed.

If any such dispute has not been settled within 45 days following the date on which the notice of mediation is provided, either You or the Licensor may, pursuant to a notice of arbitration communicated by reasonable means to the other, elect to have the dispute referred to and finally determined by arbitration. The arbitration shall be conducted in accordance with the rules designated by the Licensor in the copyright notice published with the Work, or if none then in accordance with the UNCITRAL Arbitration Rules as then in force. The arbitral tribunal shall consist of a sole arbitrator and the language of the proceedings shall be English unless otherwise agreed. The place of arbitration shall be where the Licensor has its headquarters. The arbitral proceedings shall be conducted remotely (e.g., via telephone conference or written submissions) whenever practicable

Open Database License (ODbL)

The Open Database License allows users to copy, modify and distribute data, including for commercial purposes. Users are obligated to give appropriate credit (attribution), and to provide any redistributions of data under the same license and in open formats. The World Bank distributes data under this license when required to do so by the original data provider or by the conditions of a data partnership.

Summary

In today's dynamic construction industry, the power of data-driven insights cannot be ignored. This article dives into a detailed analysis of the annual growth percentages in Ireland's construction sector, offering a comparative perspective against countries worldwide. By examining year-on-year growth rates, we gain valuable insights into how Ireland's construction industry stacks up against its global counterparts.

To conduct this analysis, I meticulously gathered reliable data from trusted sources. By including data from various countries, I was able to provide a comprehensive evaluation. Analysing growth patterns over time helps identify emerging trends and enables informed predictions for the future trajectory of Ireland's construction sector and how it compares to other nations.

The findings shed light on Ireland's performance in terms of annual growth rates relative to other countries. Through this comparative analysis, we uncover commonalities, disparities, and potential avenues for improvement. Leveraging these insights, evidence-based recommendations are offered to enhance the performance of Ireland's construction sector. This may involve learning from successful strategies implemented in high-growth countries or advocating for policies that support sustainable growth.

Throughout the article, I provide a transparent account of the research methodology employed, including the sources of data and the analytical techniques employed to scrutinize annual growth percentages. It is important to acknowledge any limitations or biases that may have influenced the outcomes to ensure a balanced perspective.

By delving into the annual growth rates of Ireland's construction sector and comparing them globally, this article aims to empower industry stakeholders with actionable insights. It underscores the significance of data analysis in driving informed decision-making and fostering sustainable growth within the construction management domain.

Data Manipulation and Processing

In our quest to analyze the construction sector data, I refined my dataset through various data manipulation techniques. The primary focus was on eliminating irrelevant columns and optimizing the dataset's structure for more insightful analysis.

To begin, I identified and removed columns that held no relevance to my analysis objectives. By discarding columns such as 'Series Code' and 'Series Name', I narrowed my dataset down to the essential information required to address our research questions effectively.

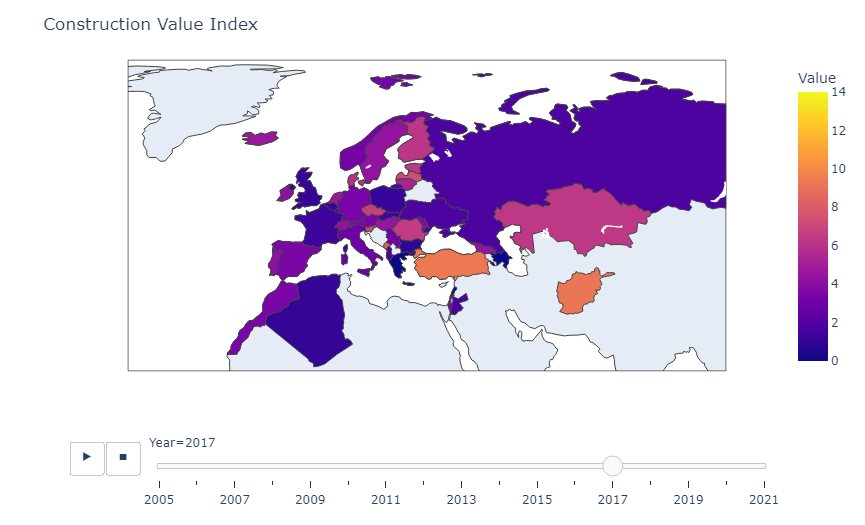
Additionally, I handled missing data within the dataset to ensure its integrity. Specific columns, potentially containing missing values were carefully examined. Any instances of missing values were addressed by applying suitable data handling techniques, ensuring that my analysis is based on reliable and complete data.

To streamline the dataset's structure, I made modifications that simplified its organization. This included restructuring the dataset by transforming certain columns and adjusting their data types. These modifications enabled to optimize the dataset for analysis purposes and facilitated further exploration of the construction sector data.

I refined the dataset by removing data that was not applicable to analysis timeframe. Specifically, data for the year '2022 [YR2022]' was excluded to ensure a focused and relevant analysis within the desired time period.

By performing these critical data manipulation steps, successfully prepared dataset for a more comprehensive analysis of the construction sector. My refined dataset now consists of pertinent columns, handles missing data effectively, and has been structured to facilitate a deeper understanding of the subject matter.

Statistics and Visualisation



Picture 1.1 <NaturalMap.html>

The colour scale and figure design were thoughtfully chosen to communicate the Construction Value Index. The colour scale, "Plasma," was selected to provide a clear representation of values, with lighter shades indicating lower values and darker shades representing higher values. This facilitates easy interpretation and comparison of construction values across countries.

The choropleth figure showcases the Construction Value Index on a map, with each country's shading reflecting its respective index value. The animation feature enables viewers to observe changes in the index over time, identifying trends and patterns. The natural earth projection enhances geographic accuracy, while the concise title, "Construction Value Index," ensures clarity.

Descriptive statistics:

From Choropleth map I choose three country with similar values.

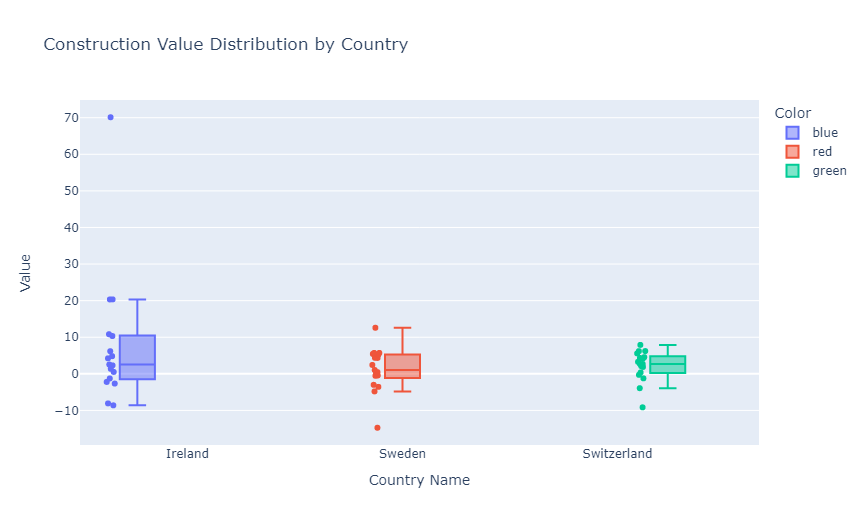
To compare the construction sector growth in Ireland, Sweden, and Switzerland, I analysed the dataset for these countries. The descriptive statistics reveal key insights:

A picture containing text, screenshot, font, number

Description automatically generated

Picture 1.2

These statistics provide an overview of construction sector growth in the selected countries, offering insights into average growth, variability, and extreme values. Such analysis aids decision-making and informs further exploration in the construction management industry.



Picture 1.3 <ValueDistribution.html>

Visualisation:

The colour choices were deliberately selected to enhance the distinction between countries and facilitate easy interpretation of the box plot. The use of different colours allows viewers to quickly identify and differentiate the construction value distributions for each country.

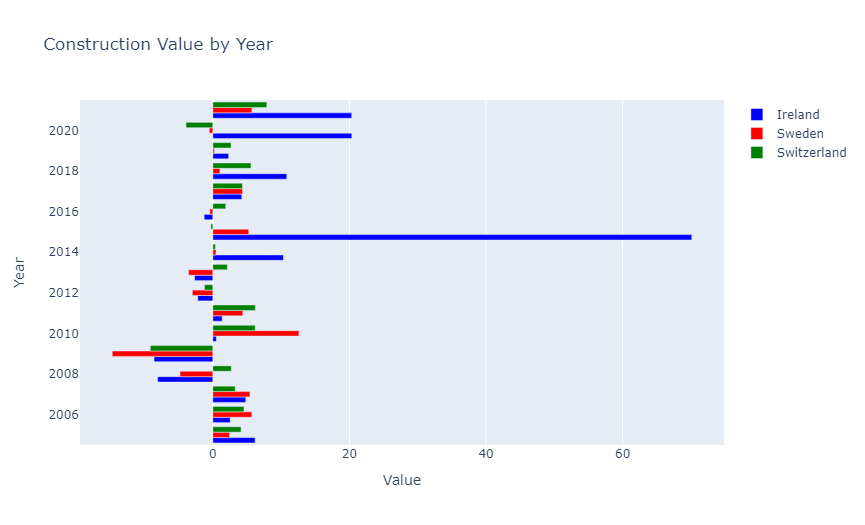
To compare the distribution of construction values among different countries, a box plot visualization was created using Plotly Express. The countries Ireland, Sweden, and Switzerland were assigned distinct colours for clarity.

The resulting box plot provides a concise overview of the construction value distribution for each country. It allows viewers to easily compare the distributions, identify any outliers, and gain insights into the overall performance of the construction sector in each country.

This visualization serves as a valuable tool for understanding the variation in construction values across the selected countries, facilitating data-driven decision-making and further analysis.

Statistics:

The construction value distributions for Ireland, Sweden, and Switzerland. They offer valuable insights into the central tendency, variability, and range of construction values, allowing stakeholders to gain a deeper understanding of the sector's performance within each country.



Picture 1.4 <BarChartByYear.html>

Visualisation:

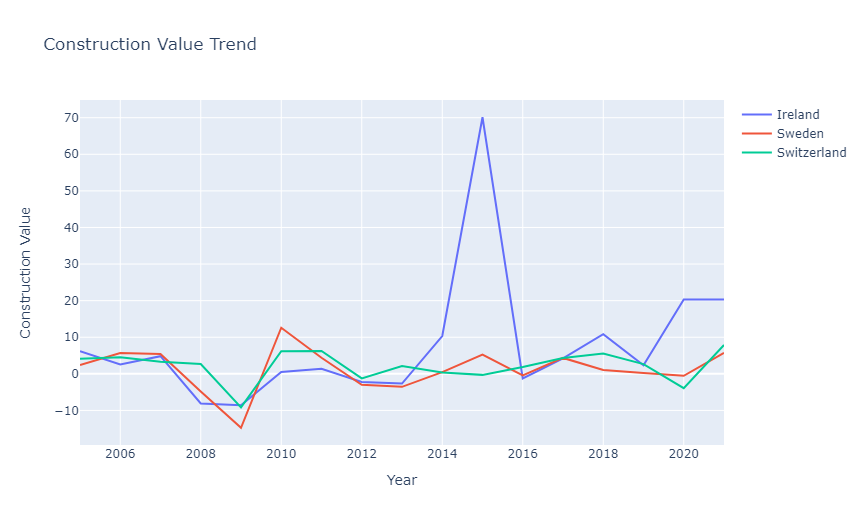
To explore the trends in construction values over the years for different countries, a visually engaging bar chart was created using Plotly Express. The chart showcases the construction value variations for Ireland, Sweden, and Switzerland, with each country represented by a unique color.

The bar chart offers an intuitive depiction of how construction values have changed over time for each country. The horizontal orientation of the chart allows for effortless comparison between the years, while the y-axis represents the respective years and the x-axis represents the construction values.

By analysing this visualization, viewers can easily identify noteworthy shifts in construction values and gain insights into the relative performance of each country. This visual exploration serves as a valuable tool for identifying potential growth opportunities or challenges within the construction industry.

Statistic:

The bar chart visualization of construction value trends for different countries provides valuable insights into the yearly variations. Alongside the visualization, statistical analysis was conducted to uncover key measures describing the construction values represented in the chart.



Picture 1.5<TrendByYear.html>

Visualisation:

The visualization created in the code is a line graph that displays the trend of construction values over the years for Ireland, Sweden, and Switzerland. The graph helps us observe the changes in construction values over time and detect any patterns or trends. The choice of a line graph is suitable for showing the continuity and progression of values. The different colours assigned to each country make it easy to distinguish and compare their construction value trends. Overall, the line graph provides a clear and visually appealing representation of the construction value trend for the selected countries.

The choice of the colours:

Blue: Blue is often associated with qualities such as stability, trust.

Red: Red is a bold and attention-grabbing colour that is often associated with energy, strength

Green: Green is commonly associated with growth, renewal, and sustainability

Statistic:

By examining the data for this specific year 2017, we gain insights into how closely aligned the construction values were among these countries. The statistics for the year 2017 enable us to evaluate the proximity of construction values among Ireland, Sweden, and Switzerland. A close mean, small standard deviation, and a narrow range between the minimum and maximum values indicate a higher degree of similarity in construction values.

Statistic Tests:

T-Test

ANOVA

Chi-Squared Test

Paired T-Test

Wilcoxon Signed Rank Test

A screenshot of a computer program

Description automatically generated with medium confidence

Picture 1.6

The T-Test results shows us the t-statistic and p-value for comparisons between Ireland, Sweden, and Switzerland. Based on the results, it appears that there are not significant differences in construction values between Ireland and Sweden, Ireland and Switzerland, and Sweden and Switzerland.

The ANOVA results indicate that there is a statistically significant variation in construction values among the three countries. This suggests that there are differences in construction values across the countries.

The Chi-Squared Test results demonstrate an association between country names and colours. The low p-value suggests a significant relationship between the variables, indicating that the chosen colours for visualization correspond well with the country names.

The Paired T-Test and Wilcoxon Signed-Rank Test results:

In this analysis, the results suggest that there may not be significant differences in construction values within the pairs of countries being compared. For example, when comparing Ireland and Sweden, the t-statistic and p-value indicate that the observed differences in construction values between the two countries may not be statistically significant. The same applies for other pair of countries.

Overall difficulties about Statistic:

With the limited data and small number of values, conducting comprehensive tests and statistical analyses becomes challenging. The results indicate that there may not be significant differences in construction values within the paired countries. However, it is important to interpret these findings cautiously due to the limitations of the analysis. Collecting more data and conducting further analysis would provide a more reliable understanding of construction values across different countries.

Machine Learning

Supervised Learning:

A picture containing text, font, screenshot, graphics

Description automatically generated

Picture 1.7

After we trained the model, we use it to make predictions on the test data. These predictions represent the estimated 'Value' based on the 'Country Name' and 'Year'.

To evaluate the performance of our model, we calculate the Mean Squared Error (MSE), which measures the average difference between the predicted values and the actual values. In our case, the calculated MSE is 38.923659550917236.

By examining the MSE, we can assess how well our model is performing. A lower MSE indicates that the model is making more accurate predictions.

A picture containing text, screenshot, font

Description automatically generated

Picture 1.8

For Picture 1.8, I used The GridSearchCV algorithm, which is explores different combinations of hyperparameters, trains multiple models, and selects the one with the best performance. The best parameters found during the search, the best score achieved during cross-validation, and the accuracy of the model on the test data are printed.

A screenshot of a computer

Description automatically generated

Picture 1.9

The output of the code provides the final data with predicted values. The code performs label encoding, splits the data into labeled and unlabeled sets, trains a LinearRegression model on the labeled data, predicts values for unlabeled data, and combines the data to create the final dataset with predicted values.

A picture containing text, font, screenshot, graphics

Description automatically generated

Picture 1.10

The mean score obtained from cross-validation of the Linear Regression model is -0.2393.

A picture containing text, font, screenshot, graphics

Description automatically generated

Picture 1.11

The mean score obtained from the cross-validation of the Linear Regression model is -0.2393



Picture 1.12

The mean score obtained from cross-validation of the Decision Tree Regressor model is -0.5414

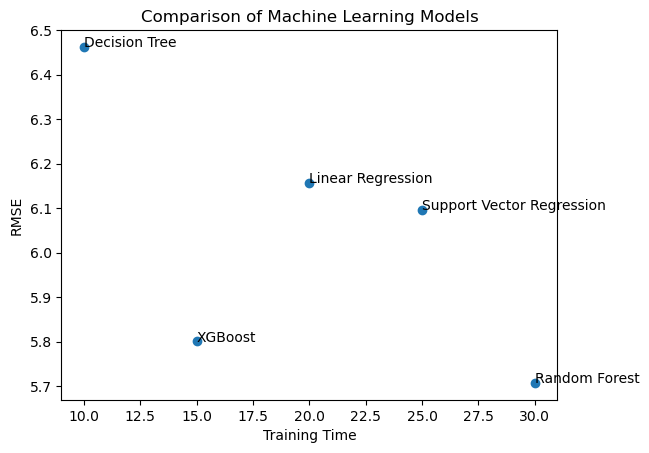
Overall, for these three models negative mean score says that the model's performance is not satisfactory. However, it is important to consider the specific problem and the range of possible scores when interpreting the result.

A picture containing text, font, screenshot

Description automatically generated

Picture 1.13

RMSE is a metric that helps us understand how well a model predicts the target variable. It calculates the average difference between the predicted values and the actual values. Lower RMSE value is better because the model's predictions are closer to the actual values. We trained and evaluated multiple regression models using different algorithms: Decision Tree, Linear Regression, Random Forest, XGBoost, and Support Vector Regression. We calculated the RMSE for each model to compare their performance.



Picture 1.14

The scatter plot offers a visual summary of the performance of different machine learning models in terms of their training times and RMSE values. It allows us to compare the models and understand the trade-offs between training time and prediction accuracy, providing insights into their relative strengths and weaknesses.

Visualisation:

White board and blue dots gives us clear visual to see difference between models.