**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | MSc. In Data Analytics |
| **Assessment Title:** | * Programming for DA * Statistics for Data Analytics * Machine Learning for Data Analysis * Data Preparation & Visualisation |
| **Lecturer Name:** | * Sam Weiss * Taufique Ahmed * Muhammad Iqbal * David McQuaid |
| **Student Full Name:** | Maria Anany Casas Manosalva |
| **Student Number:** | 2023326 |
| **Assessment Due Date:** | 07-01-2024 |
| **Date of Submission:** | 07-01-2024 |

**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. |

|  |
| --- |
| Introduccion  The purpose of this paper is to collate information related to transport in Ireland and compare it with a cross section of countries from both the European Union or outside the EU. For this assignment, I worked with the official website: https://stats.oecd.org/ to obtain a dataset for both Ireland and for 10 other countries belonging to the European Union (Austria, Belgium, Denmark, Finland, Greece, Luxembourg, Netherlands, Norway, Sweden) as well as Switzerland. Four files with.csv extension were obtained and a series of graphs made to see relationships between investment, maintenance and capital value in terms of transport infrastructure in these countries.  The end goal of this project being to compare which countries invest the most in the different categories presented, including: Total inland transport infrastructure investment, road infrastructure maintenance, rail infrastructure maintenance, and total road spending among others.  Don’t forget to show with invest the most, maintenance  Las funciones definidas incluirlas en programming   1. **Data Preparation and Visualisation**     1. The row data was compiled from the website https://stats.oecd.org/ which belongs to the Organization for Economic Co-operation and Development (OECD) and is a public source of statistical data provided by this international organization.   The archives contain information on investment and maintenance used by different countries around the world regarding roads, rails, airports, maritime ports, among others.  The positive aspects found were that the data is public access and obtained without any problems, allowing wide availability and transparency. Also, the diversity of information provided a comprehensive view of transportation infrastructure in different countries.    Figure Nº 1   * 1. Exploratory Data Analysis (EDA):   The project started with an exploratory analysis to understand the general structure of the data and to analyse the existing variables.  The dataset contains detailed information on transportation infrastructure investment and maintenance in several countries, including variables such as country, variable, measure, year, and value.  Eleven countries (including Ireland) were chosen for the data analysis, from which the following were selected with some similarity between them in terms of population size and density: Austria, Belgium, Denmark, Finland, Greece, Irlanda, Luxembourg, Netherlands, Norway, Sweden and Switzerland.  Null values were identified in several sections and we chose to eliminate them, since the information obtained was from different countries and it was not feasible to replace them with the median or mean.  Box plots and scatter plots were used to detect the presence of outliers, some of which were found to be outside the acceptable range and appropriate techniques (selection of rows under a boolean condition) were used to eliminate them.  Variation of variables over time was examined to identify trends and temporal patterns, with fluctuations in infrastructure investment and maintenance over the years, which may be influenced by economic and political factors.   * 1. Outlier Treatment   Outliers identified during the EDA were addressed to mitigate the impact of these values as they can introduce distortions and noise to machine learning models. In this graph it was observed that after the elimination of the atypical values, the boxplot presents values not as far away as at the beginning.  Boxplot before eliminating outliers    Figure Nº 2  Boxplot after eliminating outliers    Figure Nº 3  Feature Engineering  A new column was created for the implementation of the machine learning models called Total\_Infraestructure, which was considered the target variable and it was important to add it since it shows the sum of the investment, maintenance and capital values for each year.    Figure Nº 4   * 1. Data Visualisation   Graphs were implemented for the visualization of transport infrastructure values for the countries studied over the years.    Figure Nº 5  Figure Nº 5 shows the Total Investment by country over the years. When we delve deeper into the data, it becomes evident that the country that contributes the most in investment is Switzerland while the one that invests the least is Luxembourg. Regarding Ireland, investment increases until 2007 and then there is a drop until 2014 followed by a peak of investment and then decreases again until 2020.    Figure Nº 6  As for Figure Nº 6, the total maintenance by country over the years is observed. In terms of maintenance, Austria is at the top of the list, while Luxembourg is once again the country that contributes the least in maintenance. Ireland is one of the countries that invests the least in maintenance with a fairly constant trend of low investment. Only 6 countries participate in capital value, with Sweden having the highest values compared to other countries, while Ireland has the lowest capital value.    Figure Nº 7  Figure Nº 7 presents a dynamic data visualization. In this it can be seen, with respect to investment, that very little has been invested in inland waterways infrastructure and maritime ports, while greater investment is observed with respect to road infrastructure. Ireland invests much more in airport infrastructure (not for nothing they are leaders with the airline Ryanair) investing very little in maritime ports.    Figure Nº 8  In Figure No. 8 it can be seen in general, with respect to maintenance, that not much has been invested in inland waterway infrastructure, while there is a higher investment priority with roads. Since 2014, in Ireland there has been a growth in investment in the maintenance of train rails, while on the other hand, maintenance with respect to airport infrastructure has been decreasing (the opposite occurred with investment in airports).   1. **Statistics for Data Analytics**    1. Descriptive statistics   These descriptive statistics provide a summary of the central tendency and variability in the dataset.    Figure Nº 9    Figure Nº 10  It can be generally observed that the results highlight the variability in transport infrastructure investments between countries.  Likewise, the high average investment in Norway, the Netherlands, Sweden and Switzerland indicates a significant focus on transportation infrastructure.   * 1. Histogram and Correlation     Two histograms were plotted, one before the elimination of the outliers and the other after the elimination, showing a notable difference in the graphs, which are presented below.    Figure Nº 11    Figure Nº 12  In figure Nº 13 it can be concluded that there is a strong positive correlation (0.92) between investment and maintenance of road infrastructure, which is understandable, since greater investment usually goes hand in hand with greater maintenance.  It is also observed that there is a negative correlation between the year and most of the variables related to infrastructure (-0.84: year and maritime port infrastructure investment) (-0.84: year and road infrastructure maintenance) (-0.81: year and road infrastructure investment), indicating that as time passes, investment and maintenance tend to decrease. This could be due to several economic and political factors.    Figure Nº 13   * 1. Parametric and non-parametric inferential statistical   To perform the comparison between variables we can use parametric and non-parametric statistical tests to assess similarities and differences. Here is a general approach using some statistical tests:  Chi-squared Test  Used to evaluate the independence between two categorical variables  Ho (null hypothesis): There is no significant relationship or association between 'Country' and Variable which contains the word 'investment'  H1 (alternative hypothesis): There is significant relationship or association between 'Country' and Variable which contains the word 'investment'    Figure Nº 13  The null hypothesis is rejected, meaning that there is a significant relationship or association between "Country" and the variable containing the word "investment".  T-student Test  Used to compare the means of two independent samples  Ho (null hypothesis): There is no significant difference in investments between Ireland and Austria in 2010  H1 (alternative hypothesis): There is significant difference in investments between Ireland and Belgium in 2010    Figure Nº 14  In this case the null hypothesis is not rejected, which means that there is no significant difference in investments between Ireland and Austria in 2010.  ANOVA  Used to compare the means of more than two groups  Ho (null hypothesis): There is no significant difference in maintenance among Ireland, Austria, Belgium, Finland, Luxembourg, Norway, Sweden and Switzerland in 2020  H1 (alternative hypothesis): There is significant difference in maintenance among Ireland, Austria, Belgium, Finland, Luxembourg, Norway, Sweden and Switzerland in 2020    Figure Nº 15  The null hypothesis means that it is rejected, which means that there is a significant difference in support between Ireland, Austria, Belgium, Finland, Luxembourg, Norway, Sweden and Switzerland in 2020.   * 1. Challenges faced   - The initial data was large but was considerably reduced because countries with certain similarities were chosen and null data were eliminated  - The machine learning model has a R^2 of almost one, which may indicate that the training of the data requires more data to perform better.  - Know what kind of statistical tests should be used because they depend on the initial nature of the data to be used.   1. **Machine Learning for Data Analysis**    1. Machine Learning Models   For this project, it was determined that the target variable will be total infrastructure (aggregate variable which contemplates the sum of the columns presented in the DataFrame) in which it is observed that all the columns present numerical values. That is why it was decided to use applied models for supervised regression learning which include: Linear Regression, Polynomial Regression, K-Nearest Neighbors Regression (KNN Regression), Decision Tree Regression, Random Forest Regression, Support Vector Regression (SVR), Ridge Regression, etc.  The following models were used for this occasion: Linear Regression, Ridge Regression and Ridge Regression with polynomial features.  Linear Regression    Figure Nº 16    Figure Nº 17  The machine learning model was applied and an R^2 of 0.9999 was obtained for the regression linear. In turn, GridSearchCV was applied and no hyperparameters were observed, which indicates that this model does not have hyperparameters to adjust, so an empty set was obtained  Ridge Regression    Figure Nº 18    Figure Nº 19    Figure Nº 20  For the Ridge Regression model, an R^2 of 0.999 was obtained. Likewise, when GridSearchCV was applied, an alpha hyperparameter of 0.005 was obtained, which was replaced in the model to obtain an R^2 value of 0.999.   * 1. Sentiment Analysis   For sentiment analysis, the applied technique of VADER (Valence Aware Dictionary and Sentiment Reasoner) was applied, since this is used to analyze opinions in texts found on social networks such as Twitter, Reddit, Facebook, Instagram, among others.    Figure Nº 21    Figure Nº 22  From the graphs shown it can be seen that there are more positive opinions among transport users in Ireland, but there are also neutral and negative opinions.  On the other hand, it can be seen that there are many more positive opinions than negative or neutral opinions among transport users in Switzerland.   1. **Programming for DA**    1. Programming   The pandas library was used to load and manipulate the files obtained in csv format. This allows us to easily collect, clean and manipulate the data.  Many other libraries were also used, such as those detailed below: Numpy, Matplotlib, Seaborn, Scipy, Plotly.express, Dash, Sklearn, Nltk.   * 1. Data structures   Four files with csv extension were used for this project.  On the other hand, comments obtained from Reddit API were extracted in JSON format and then converted into a DataFrame.   * 1. Documentation: |