**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:** | 10-11-2023 |
| **Date of Submission:** | 10-11-2023 |

**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. |

|  |
| --- |
| The present information of this project has been collected by The Central Statistics Office in Ireland, from the page <https://data.cso.ie> following the route: “Browse”, “Environment”, “Forestry”, “Afforestation Area”.  The main objective of this project is to analyse historical rates of afforestation in Ireland and accurately predict the area that will need to be afforested from 2023 to 2030.  Ireland is one of the least forested countries in the European Union, with current levels of 11.6% (808,848 hectares) of area covered by trees compared to the EU average of 38.3% covered by forest. There are several reasons for this, but the main one is the historical indifference towards forestry. When Ireland became independent from Great Britain in 1922, only 1.4% of the country was covered in forest and in the following decades very little effort was made to change this situation. It was not until the 1970s that a concerted afforestation campaign began. Over the last 40 years, the total area of land under forestry has increased to the aforementioned 11.6% that Ireland currently has. Unfortunately, this amount is nowhere near the amount needed for Ireland to meet its climate change targets. To meet its climate action commitments, Ireland will need to plant more than 8,000 hectares of forest each year until 2050. This will raise the country's land afforestation rate to more than 18% (although, according to the Environmental Protection Agency, this will need to be closer to 24% today - https://www.euroforestireland.ie/news-posts/forestry-planting-rates-must-exceed-8000ha-per-year-to-meet-climate-targets-epa). The problem with Ireland trying to achieve these targets is that current levels of forestry in Ireland are at the lowest they have been in 20 years. CSO data shows that the 8 000 hectares target, which is included in the Government's Climate Action Plan, is being missed by a significant margin. According to figures from the CSO's first forestry statistical report, Afforestation Area 2021, the rate of tree planting fell from 6,947 hectares in 2007 to just 2,016 hectares in 2021. Only 1,400 have been planted so far this year. hectares of forests. Significantly, the share of forest plantations by farmers in 2021 was only 18 percent, down from 97 percent in 2014, according to the report. County Cork had the largest forest area each year from 2007 to 2016 and from 2019 to 2021. Cork accounted for 17% of the total forest area in 2021, followed by Roscommon (9.4%), Clare (8.6%) and Cavan (7.9%). Despite government targets, the amount spent on forestry last year (€74 million) was less than that spent in 2002. To counter this worrying trend, the government has launched a new €1.3 billion forestry program that the European Commission has recently approved. |

|  |
| --- |
| Programming for DA  First of all, it is necessary to do Exploratory Data Analysis (EDA) which is a methodology used in data science and statistical analysis to explore, analyse and visualise datasets to understand their structure, patterns and main characteristics. To do so, we will follow next steps: loading the data, exploring the information, debugging the data, final data processing, analysing the data.   1. Loading the data   In [1]: Import Pandas library which is used for data manipulation and analysis. It works with DataFrames which allow you to efficiently load, clean, transform, and analyse data.  In [2]: Read the CSV file named 'Afforestation\_area.csv' and store the data in a DataFrame named 'afforestation\_df' (convert the CSV file in a DataFrame).   1. Exploring the information:   In [3]: Display the object type of the 'afforestation\_df' which should be a DataFrame.  In [4]: The shape method shows the number of rows and columns of the DataFrame (5184 rows, 7 columns).  In [5]: The info method describes the data type of each column and the number of non-null values of each one. We can see that there are 5184 entries and 3687 non-null in the ‘VALUE’ column, so we have to decide what we will do with this missing values.  In [6]: the head method shows, in this case, the first 10 rows of the DataFrame.   1. Debugging the data:   In [7]: Delete the rows that contain 'Total Afforestation' since I want to analyse the plantations carried out for each specie and forest owner instead of analysing them in their entirety, this is to avoid generating redundancy or noise.  Filter the DataFrame to remove rows where the columns 'Species' or 'Forest Owner' contain 'Total Afforestation' and update the DataFrame with the requested conditions.  In [8]: Remove columns 'Statistic Label' and 'UNIT' due to they both have the same value: 'Afforestation Area' and 'Hectares' respectively.  The drop method deletes rows or columns in a DataFrame, axis=1 (remove columns instead of rows), inplace=True (modify the DataFrame in place).  In [9]: Rename the columns 'VALUE' and 'Forest Owner' creating a dictionary for then use the method rename to rebrand columns named 'VALUE' to 'Value\_ha' and 'Forest Owner' to 'Forest\_Owner'.   1. Final data processing   It can be shown that there are many NaN or null values in the 'Values\_ha' column (in the csv file they appear as empty data) and most of this information comes from 'Non-Farmer' and 'Public Sector'. Initially, I decided to replace these values to 0 because the afforestation mainly comes from 'Farmer' who receive loans for this work, but when I did the normalization, I got a high frequency of 0 values (about 1000 values) so it is better to remove the rows of null values to avoid this interference.  In [10]: The dropna method drops rows or columns with NaN values in a DataFrame, axis=0 (remove rows instead of columns).  In [11]: Create a variable cero\_count to determine how many zero values there are in the 'Value\_ha' column.  In [12]: Create a dictionary to change some names shown in the 'County' and 'Forest\_Owner' columns.  In [13]: The replace method substitutes some values in the ‘County' and ‘Forest\_Owner' columns according to those found in the dictionary created.  In [14]: Call the info method again to see the new number of rows obtained and if there are still null values, observing that the DataFrame contains 1594 entries or rows but the index values start from 5 to 5182. It means that the indexes need to be reset since rows have been deleted.  It can also be seen that the data type of each column is according to its characteristics and we don't need to perform a conversion.  In [15]: The method reset reboots indexes to start from 0, drop=True (clear all indexes).  Verbose=False (it only will display range index, since it displays specific information)   1. Analysing the data   In [16]: Import Numpy library which is used for used for mathematical calculations, statistics and lineal algebra operations. It works with n-dimensional arrays that are suitable for numerical operations.  Import Matplotlib library that is a powerfull tool used for creating 2D visualizations, plots and for representing data visually.  Import Seaborn that is a data visualization library built on top of Matplotlib and it simplifies the creation of informative and visually appealing statistical graphics.  In [17]: Create a Bar Chart to visualize the “Total Afforestation per County from 2007 to 2022”, except for Ireland (it represents the sum of all counties and we do not want to generate redundancy), by following these steps:   * Make a graph and set its size to 12 (wide) x 10 (high) * ‘Filtered1’ variable is a boolean string that contains True if (afforestation\_df['County'] != 'Ireland') * Filter the data in the 'County' column except 'Ireland' and use the groupby method to calculate the sum of the values in the 'Value\_ha' column for each 'County' group * Generate a Bar Chart with the data contained in the 'afforestation\_county' variable, kind='bar' (create a Bar Chart), figsize=(12, 6) (Create Bar Chart dimensions: 12 (wide) x 6 (high)) * plt.title (set the graph title as "Total Afforestation by County from 2007 to 2022" with a font size of 20) * plt.xlabel (set the x-axis label to "County" with a font size of 15) * plt.ylabel (set the y-axis label to "Afforestation (ha)" with a font size of 15) * plt.xticks(rotation=90) (rotate the x-axes labels by 90 degrees to improve readability) * Prop = {'size': 10} (set the font size to 10) * Display the Bar Chart   In [18]: Create a Bar Chart to visualize the “Total Afforestation by Counties per Year”, except for Ireland, by following the same structure as the previous Bar Chart with the difference that the data is grouped by ‘Year’ instead of ‘County’.  In [19]: Generate a Histogram with Matplotlib library to view the distribution of values. The steps are shown below:   * Make a graph and set its size to 12 (wide) x 10 (high) * Create a Histogram with the name 'Value\_ha' on the x-axis, bins=20 (number of bars or bins in the graph), color='green' (set the color of the bars to green), edgecolor='black' (set the color of edges to black), linewidth=1 (set the linewidth of edges to 1)   In [20]: Create a Histogram with Seaborn library to display the distribution of values, by following a similar structure as the previous Histogram.  In [20]: We can observe that the data is completely right skewed (Asymmetric Distribution) and in order to see the outliers vert clearly, we can use Boxplot.   * Make a graph and set its size to 12 (wide) x 10 (high) * Create a Box Plot with the name'Value\_ha' on the x-axis, vert=False (create a horizontal boxplot because by default boxplot are vertical)   In [20]: |

|  |
| --- |
| Machine Learning for Data Analysis  Machine Learning (ML) is a branch of artificial intelligence that allows the progressive improvement of tasks through machine learning with computers, which through algorithms and models can learn patterns and make decisions without being explicitly programmed. ML can be classified into 3 categories:   1. Supervised learning: the algorithms work with labeled data trying to find a function that, given the input variables, assigns them the appropriate output label. 2. Unsupervised learning: occurs when labeled data is not available for training. We only know the input data, but there are no output data that correspond to a certain input. 3. Reinforcement learning: This type of learning is based on improving the response of the model using a feedback process that it obtains from the outside world in response to its actions.   For my project, I want to predict the total afforestation in Ireland by county and per year (Value\_ha will be the variable to predict).  The type of Machine Learning that I will use for this assignment is Supervised since I have past information where I can train the machine so that it can predict future values. I will also use the Regression Model since the results I want to obtain are numerical (hectares to aforest = Value\_ha). |