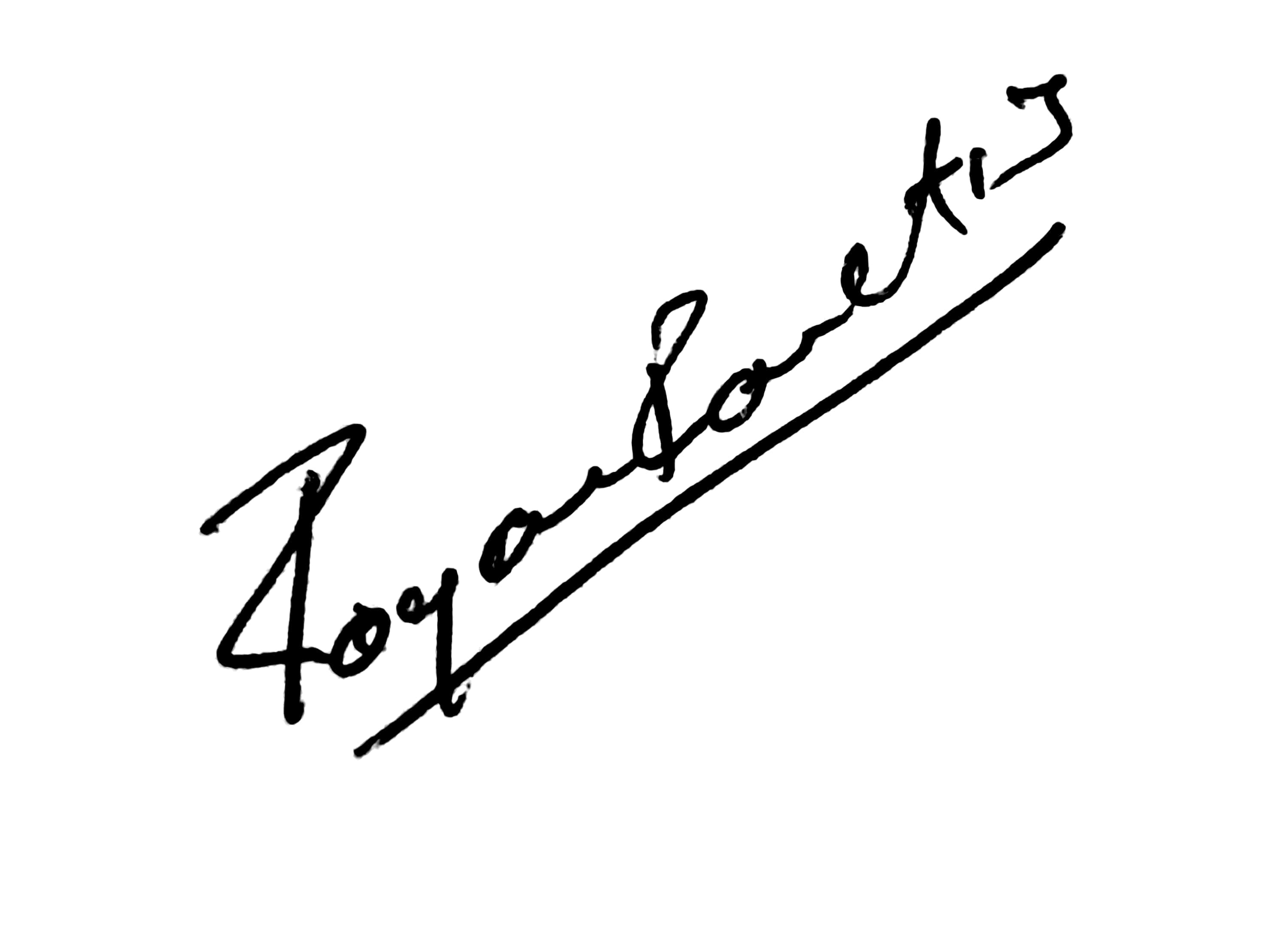
**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | *MSC\_DA\_Integr\_CA2\_Sem1* |
| **Lecturer Name:** | Sam Weiss/ David Gonzalez  Taufique Ahmed  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Ashish Royan Sanctis <https://github.com/AshishSanctis/CA_2-Trasnport-in-Ireland/upload/main> |
| **Student Number:** | 2023444 |
| **Assessment Due Date:** | 07th January 2024 |
| **Date of Submission:** | 12th January 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. |

**Transport by Luas in Ireland**

**Table of content**   **Page**

**1. Introduction................................................................................................. 4**

**2. Dataset Overview....................................................................................... 4**

**3. Exploratory Data Analysis (EDA)............................................................. 4**

- 3.1 Data Cleaning and Transformation...............................................................4

- 3.2 Outlier Identification and Handling............................................................. 5

- 3.3 Finding corelation using Heat Map............................................................. 5

- 3.3 Using Dash for visualisation..........................................................................6

1. **Statistical Analysis......................................................................................**6

- 4.1 Chi squared test.........................................................................................6

-4.2 Wilcoxon Test..............................................................................................7

-4.3 T\_Test...........................................................................................................7

**5. Data Preprocessing and Transformation................................................8**

- 5.1 Encoding Categorical Variables....................................................................8

- 5.2 Logarithmic Transformation and Standard Scaling..................................8

**6. Machine Learning Models.........................................................................8**

- 6.1 Linear Regression...........................................................................................8

- 6.2 Random Forest Regressor............................................................................9

- 6.3 Decision Tree Regressor..............................................................................10

- 6.4 Lasso Regression..........................................................................................10

- 6.5 Logistic Regression for Binary Classification...........................................11

1. **Conclusion...................................................................................................11**
2. **Breif report on Analysis of Transport in Dubai.....................................12**

**1. Introduction**

Transportation plays a significant role in the development and sustenance of urban areas, and efficient public transportation is a key component. This report focuses on the analysis of transport by Luas in Ireland from the year 2018 to 2022, aiming to derive the insights that can contribute to decision-making processes. The analysis involves a combination of exploratory data analysis (EDA), data preprocessing, and the implementation of various machine learning models. The goal is to understand passenger trends and patterns, identify outliers, build models for predicting passenger counts against the years.

**2. Dataset Overview**

The dataset used for this analysis is sourced from <https://data.gov.ie/organization/transport-infrastructure-ireland>. The data used includes columns such as Statistic, Statistic label,Tlist(A1),Year,C01885V02316,Month, unit, and Value. The initial exploration it is understood that data cleaning and transformation are required to enhance its usability.

**3. Exploratory Data Analysis (EDA)**

**3.1 Data Cleaning and Transformation.**

The initial steps involved making the data more understable/simplified by dropping unnecessary columns and renaming others.The renaming of columns was necessary since the existing name did not make clear about the data which was available in it. Each columns were checked for different data in them using .unique() which gives a brief idea and helps us to change the column name which seems to be revlevant.Following are the columns whose names were changed as per the info they showed- Statistic label is changed to Luas line as it only gave us an insight of the luas line used by the passenger for the following month.Column Value which gave us information about the number of passengers who travelled a specific luas line for the specific month was changed to Number of Passengers. The column TLIST(A1) was dropped as it had the same information as of column Year, the column Statistic was dropped as it didnot help to gain any information,the column C01885V02316 was dropped as it only named the months by numbers which information was available in column -Months in which the months were named in words which is easier to grasp, the last column which was dropped was unit that showed indicated the data in column value is only numbers. once the necessary columns are dropped and names of the columns have been changed has desired names the data is now checked for number of rows and columns using the function .shape.Further the dataset is checked for duplicate rows using the function .duplicated( ) this helps to undertand and know if any rows are repeated. if the rows are repeated with the same values they can be eliminated. The data is now checked for Null values(Nan) within the dataset using isnull().sum() which gives us the sum of all the null vlaues in each columns. The same is cross verified using .info() which gives us detail about the number of non null values in each of the columns. it can also be confirmed if the number of non null values are same as the number of rows which were seen using .shape().

**3.2 Outlier Identification and Handling.**

Outliers are the data points that differ from majority of other observations. These outliers have a significant effect on the statistical analysis. visualisation of the outliers makes it easy for us to understand them. The removal of outliers contributes to a more accurate understanding of passenger trends.Boxplot and displot from seaborn is used to visualise the outliers in the data from the column Number of passengers. The same can be done for rest of the columns.However,only Number of passengers was chosen as it is the target variable.

once the outliers have been visualised they are checked for skewness of the data. The skewness if the measure of asymetry of probability distribution. in this case the since the value is positive i.e skew value- **4.5614** the data is skewed towards the right. it is also checked by kurtosis value. the valued received **23.0310** shows the data is peaked and not normaly distributed as seen in the visualisation using displot.

The outliers are now removed following basic steps of Interquartile Range Method.(IQR) i.e IQR=Q3-Q1 where Q3 is third quartile where 75% data falls it is the median of the upper halfof the dataset and Q1 is first quartile which has 25% of the data it is the median of second half of data. The data is now visualised again to check for outliers and the removal is repeated untill desired. Once most of the outliers are removed the data is again checked for skewness which here is **0.3863** which is still skewed towards right but significantly lower than the initial stage. However, the Kurtosis value **-0.0747** indicates that the distribution is below normal distribution.

**3.3 Finding correlation using Heat Map.**

The Heat Map is used to visualise the correlationship between column Year and Number of passengers. The confusion matrix with correlation coefficient of -0.2692 suggests a weak negative coorelation between the two variables. As the Year increases we see the Number of Passengers tend to decrease slightly. The Correlation between Number of Passengers and Year seems to be same as the matrix is symmetric.

The data is now visualsed using pairplot which shows the relationship between the variables Number of Passengers and Year.

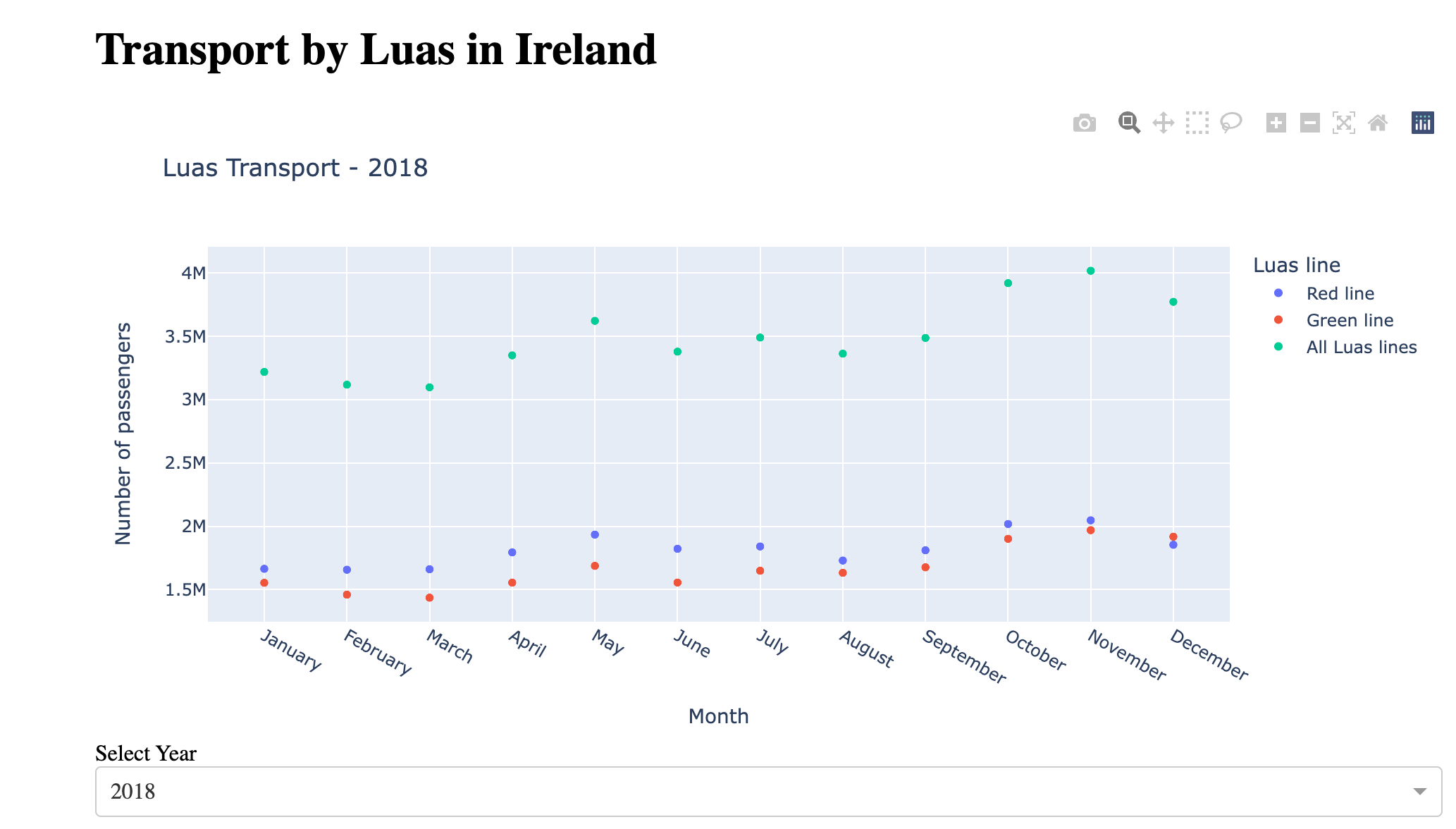
**3.4 Finding the outliers with Panel.**

The Panel is used to find the outliers for the variable Number of Passengers following the same code as reference from **Prof.David McQuaid** with changes in the varibale name and dataset.

**3.5 Using Dash for visualisation.**

The Data is now plotted using Dash with Months along the X-axis and Number of Passengers along Y axis. The unique values of the variable Year are taken as dropdown and unique values of varibale Luas line for color.

The Graph can be used to plot the Number of Passengers travelled in Red, Green and both luas lines for every month of the year. The Graph changes along with the Year selected for the months.



**4.Statistical Ananlysis**

Normal and Poisson Distribution Analysis

The analysis includes the exploration of the normal and Poisson distributions of the number of passengers. This provides insights into the distribution characteristics of the data.

**4.1 Chi squared test**

The chi-square test of independence was cross-tabulation of the 'Year' and 'Number of passengers' in the Luas1 dataset.

1. Chi-Square Statistic:The chi-square statistic is a measure of the association between the two categorical variables. In this case, the obtained chi-square value is 640.0.

2.Degrees of Freedom: The degrees of freedom (df) are determined by the size of the contingency table. In this test, df is 636.

3.p-value:The p-value associated with the chi-square statistic is 0.448. This p-value indicates the probability of observing such an extreme association between the variables if they were independent.

Conclusion:

With a p-value of 0.448, we fail to reject the null hypothesis at the 0.05 significance level. This suggests that, based on the available data, there is not enough evidence to conclude that the 'Year' and 'Number of passengers' variables are significantly associated.

**4.2 Wilcoxon Test**

The Wilcoxon signed-rank test was conducted to assess whether there is a significant difference between the 'Year' and 'Number of passengers' in the dataset and to compare the distribution of 'Year' and 'Number of passengers'.

The Wilcoxon statistic value was calculated to be 0.0. The extremely low value of the Wilcoxon statistic suggests a substantial difference between the paired samples.

Conclusion

p-value of 5.24e-28 at significance level 0.05 which rejects the null hypothesis.

Therefore, based on the Wilcoxon signed-rank test, there is a significant difference in the distribution of the number of passengers across different years. The rejection of the null hypothesis implies that the 'Year' variable has a statistically significant effect on the 'Number of passengers' variable.

**4.3 T\_Test**

A T\_Test was conducted determine the significant difference in passenger numbers between the two years 2019 and 2020. The null hypothesis assumes that there is no significant difference between the two years, while the alternative hypothesis suggests a significant difference.

With the t-test statistic value 9.6309, and the corresponding p-value was approximately 1.8375e-13. the p-value is less than the commonly used significance level of 0.05, we reject the null hypothesis.The positive t-test statistic indicates that the average number of passengers in 2020 is significantly higher than in 2019. The same test can be done to compare the number of passengers of different years.

Conclusion:

Based on the statistical analysis, we reject the null hypothesis and conclude there is a significant difference in the number of Luas passengers between 2019 and 2020. This information can be valuable for stakeholders, policymakers, and transportation authorities to understand the dynamics of public transportation usage and plan accordingly for future developments or improvements.

**5. Data Preprocessing and Transformation**

**5.1 Encoding Categorical Variables**

Encoding the data is necessary for statistical modeling and machine learning. Since the variables in object format can not be used they need to be tranformed into a numerical format for computational analysis.

**5.2 Logarithmic Transformation and Standard Scaling**

Logarithmic transformation is a mathematical operation applied to data, primarily used in statistical and mathematical modeling to address issues related to skewness or to stabilize variance. It involves taking the logarithm of each data point in a given set.Logarithmic transformations is used here as the dataset is skewed towards the right. By compressing large values and spreading out small values, logarithmic transformations can make the data more symmetrical, making it easier to meet the assumptions of certain statistical techniques.

**Standard Scaling**

Standard scaling, also known as z-score normalization, is a preprocessing technique commonly used in data analysis and machine learning to standardize the scale of numerical features. The process involves transforming the data so that it has a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean of the variable from each data point and then dividing the result by the standard deviation.

**6. Machine Learning Models**

The implementation of various machine learning models is a key aspect of this analysis, allowing for predictive analytics and understanding the factors influencing passenger counts.

**6.1 Linear Regression**

The linear regression model was trained and evaluated on the provided dataset, which includes the 'Year' variable as a predictor and 'Number of passengers' as the target variable.

1. R-squared (Coefficient of Determination):

-The R-squared of the model on the training set is approximately 0.0382, indicating that the model explains only a small proportion of the variance in the training data.

-The R-squared on the test set is similar, approximately 0.0373, suggesting limited predictive power when applied to new, unseen data.

2. Root Mean Squared Error (RMSE):

-The root mean squared error of the prediction on the test set is 0.9608. This metric provides a measure of the average deviation between the predicted and actual values, with lower values indicating better model performance.

3. Mean Absolute Percentage Error (MAPE):

- The mean absolute percentage error of the prediction on the test set is 89.89%. MAPE measures the percentage difference between predicted and actual values, and a lower MAPE is desirable.

The R-squared values suggest that the linear regression model has limited explanatory power in capturing the variability in the 'Number of passengers' based on the 'Year' feature. The relatively high RMSE and MAPE values indicate that the model's predictions deviate significantly from the actual values, highlighting potential limitations in its accuracy.

**6.2 Random Forest Regressor**

The Random Forest Regressor model was trained and evaluated on the provided dataset, using the 'Year' variable as a predictor and 'Number of passengers' as the target variable.

1.R-squared (Coefficient of Determination):

- The R-squared of the model on the training set is approximately 0.454, indicating that the model explains about 45.4% of the variance in the training data.

- The R-squared on the test set is slightly lower, approximately 0.344, suggesting a moderate ability to generalize to new, unseen data.

2. Root Mean Squared Error (RMSE):

- The root mean squared error of the prediction on the test set is 0.7933. This metric measures the average deviation between the predicted and actual values, with lower values indicating better model performance.

3. Mean Absolute Percentage Error (MAPE):

- The mean absolute percentage error of the prediction on the test set is 119.21%. MAPE measures the percentage difference between predicted and actual values, and a lower MAPE is desirable.

The Random Forest Regressor has shown improved performance compared to the linear regression model. The higher R-squared values on both training and test sets suggest that the Random Forest model captures more variability in the 'Number of passengers' based on the 'Year' feature. The lower RMSE and higher R-squared values indicate better predictive accuracy compared to the linear regression model.

**6.3 Decision Tree Regressor**

The Decision Tree Regressor model was trained and evaluated on the provided dataset, using the 'Year' variable as a predictor and 'Number of passengers' as the target variable.

1. R-squared (Coefficient of Determination): - The R-squared of the model on the training set is approximately 0.455, indicating that the model explains about 45.5% of the variance in the training data.

- The R-squared on the test set is slightly lower, approximately 0.354, suggesting a moderate ability to generalize to new, unseen data.

2. Root Mean Squared Error (RMSE):

- The root mean squared error of the prediction on the test set is 0.7871. This metric measures the average deviation between the predicted and actual values, with lower values indicating better model performance.

3. Mean Absolute Percentage Error (MAPE):

- The mean absolute percentage error of the prediction on the test set is 117.85%. MAPE measures the percentage difference between predicted and actual values, and a lower MAPE is desirable.

The Decision Tree Regressor has shown performance comparable to the Random Forest Regressor. The R-squared values on both training and test sets suggest that the Decision Tree model captures a substantial amount of variability in the 'Number of passengers' based on the 'Year' feature. The lower RMSE and higher R-squared values indicate good predictive accuracy.

**6.4 Lasso Regression**

The Lasso Regression model was trained and evaluated on the provided dataset, utilizing the 'Year' variable as a predictor and 'Number of passengers' as the target variable.

1. R-squared (Coefficient of Determination):

- The R-squared of the model on the training set is approximately 17.128%, indicating that the model explains a limited amount of the variance in the training data.

- Surprisingly, the R-squared on the test set is negative, approximately -13.82%. A negative R-squared suggests that the Lasso Regression model performs worse than a model predicting the mean of the target variable.

1. Root Mean Squared Error (RMSE):

- The root mean squared error of the prediction on the test set is exceptionally high, approximately 104.474%. This indicates a substantial deviation between the predicted and actual values.

3. Mean Absolute Percentage Error (MAPE):

- The mean absolute percentage error of the prediction on the test set is also high, approximately 137.176%. MAPE measures the percentage difference between predicted and actual values, and a lower MAPE is desirable.

The Lasso Regression model appears to struggle with capturing the underlying patterns in the data, as evidenced by the low R-squared values and high error metrics. The negative R-squared on the test set suggests that the model may not be suitable for this specific dataset.

**6.5 Logistic Regression for Binary Classification**

The Logistic Regression model was trained and evaluated on the provided dataset to predict whether the 'Number of passengers' on the Luas transportation system is above a given threshold.

1. Model Accuracy:

- The accuracy of the Logistic Regression model on the test set is approximately 96.88%. This metric measures the overall correctness of the model's predictions.

2. Confusion Matrix:

- The confusion matrix indicates that the model made 30 correct predictions for class 0 (below the threshold) and 1 correct prediction for class 1 (above the threshold).

- There is 1 false positive (predicting above the threshold when it's below), and there are no false negatives (predicting below the threshold when it's above).

3. Classification Report:

- The precision for class 0 is 1.00, indicating that all predictions labeled as below the threshold are indeed correct.

- The precision for class 1 is 0.50, suggesting that only half of the predictions labeled as above the threshold are correct.

- The recall for class 1 is 1.00, indicating that the model captures all instances of the actual positives.

The Logistic Regression model demonstrates high overall accuracy, correctly identifying instances below the threshold. However, there is room for improvement in predicting instances above the threshold, as reflected in the precision for class 1.

Below is the summary of the all the models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Approach | R^2 Training set | R^2 Test set | Root means square error for prediction | Mean absolute percentage error for prediction |
| Linear Regression | 0.03816 | 0.03725 | 0.96084 | 89.89 |
| Random Forest Regressor | 0.4539 | 0.3437 | 0.7933 | 119.21 |
| Decision Tree Regressor | 0.4552 | 0.3539 | 0.7871 | 117.85 |
| Lasso Regression | 17.128% | -13.82% | 104.474% | 137.176% |

**7. Conclusion**

In conclusion, the analysis of Transport by Luas data in Ireland provides valuable insights into passenger trends and facilitates predictive modeling. The use of various codes, ranging from data cleaning to machine learning model implementation, ensures a comprehensive exploration of the dataset. The integration of Dash for interactive visualization enhances the accessibility of the analysis. The combination of statistical and machine learning approaches contributes to a holistic understanding of Luas transport patterns, aiding stakeholders in making informed decisions for the improvement of public transportation services in Ireland.

8. **Breif report on Analysis of Transport in Dubai**

**Dataset Overview:**

The dataset is sourced [https://www.dsc.gov.ae/enus/Themes/Pages/Transport.aspx?Theme=31](https://www.dsc.gov.ae/en-us/Themes/Pages/Transport.aspx?Theme=31)

website, it gives a various data set for public transport for Dubai from year 2001 to 2022.since the dataset is to compare with dataset from Ireland it was necessary to have some similarities. however it was challenging to find the dataset for all the years combined hence 3 different datasets were taken from year 2017 to 2019 for number of passengers travelling by Metro and Tramp. The dataset included clomuns like Metro Red line. Metro Green line, Tramp and Month.

**Exploratory Data Analysis (EDA)**

**Data Cleaning and Transformation.**

Exploration of dataset gave a brief idea about the information about it. since the data included few words in Arabic it could not be opened using pandas. The data was later slightly altered by deleting any Arabic letters to open it. Year column was added to all three datasets which are now combined using melt for columns Month,Year and concate function to form a single dataset. the column Metro-red line, Metro-Green line and tramp are transposed under a new column Transport by and thier values under the column Number of Passengers.it is later checked for duoplicated rows and null values.

**Outlier Identification and Handling.**

The outliers are visualised using boxplot and displot for variable number of passengers. The skewness of the data is checked and the value is found to be **1.4758** with kurtosis value **0.7834.** The values show that the data is skewed towards right and xlightly peaked than the normal distribution.

The outliers are removed for 3 degrees and again plotted to visualise them.

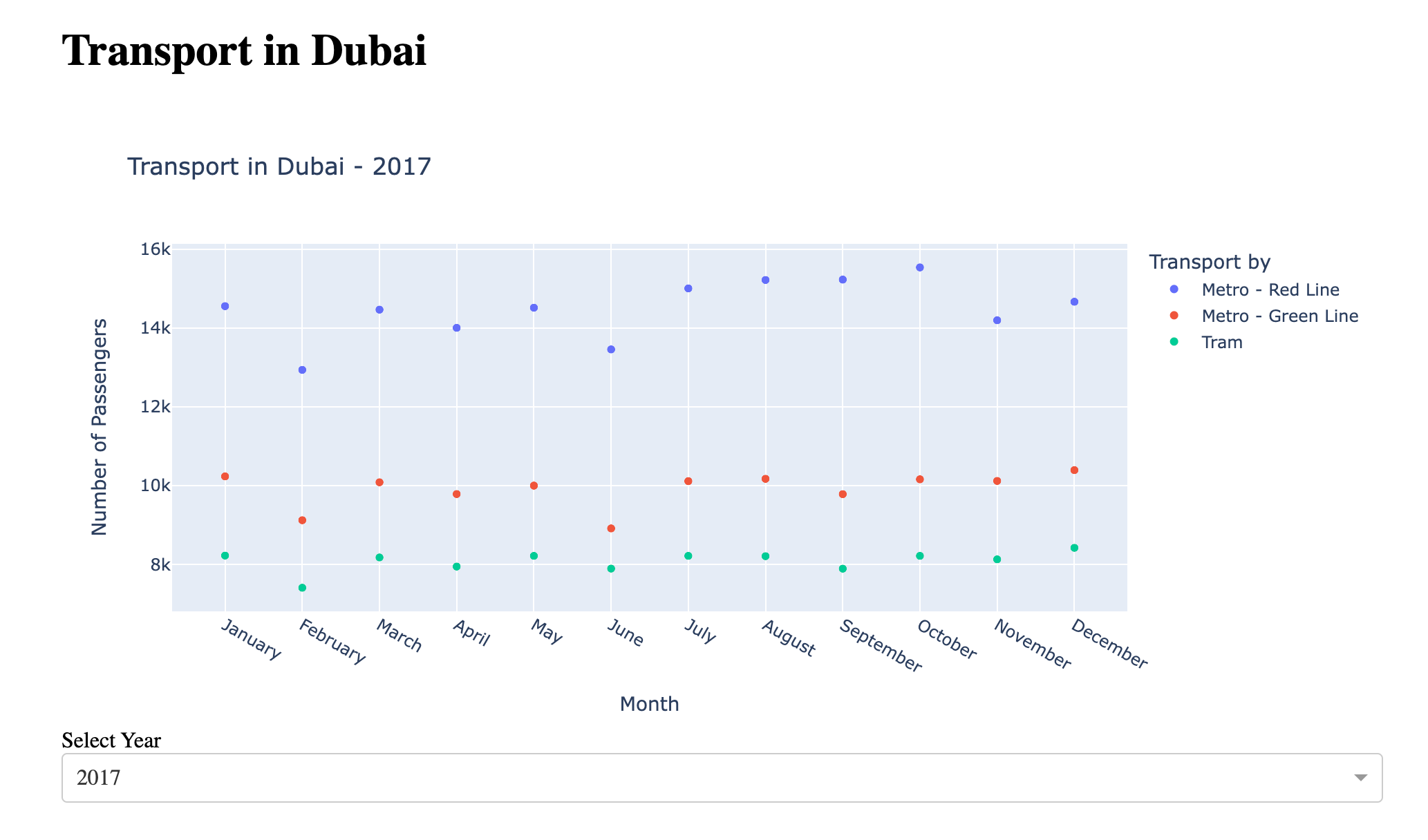
**Finding correlation using Heat Map.**

A correlation is visualised using heat map for Year and Number of passengers. it is found that the correlation coefficient is 0.03.since the value is close to zero it shows a vary weak linear correlation between two variable.

**Using Dash for visualisation.**

The Data is now visualised using Dash with Months along the X-axis and Number of Passengers along Y axis. The unique values of the variable Year are taken as dropdown and unique values of varibale Luas line for color.

The Graph can be used to plot the Number of Passengers travelled in Red, Green-Metro lines and Tramp for every month of the year. The Graph changes along with the Year selected for the months.



**Statistical Ananlysis**

Normal and Poisson Distribution Analysis

The analysis includes the exploration of the normal and Poisson distributions of the number of passengers. This provides insights into the distribution characteristics of the data.

**Chi squared test**

The chi-squared statistic and p-value obtained from the chi-squared test are as follows:

Chi-squared statistic: 0.0

P-value:1.0

A chi-squared statistic of 0.0 indicates perfect agreement between the observed and expected counts.The p-value associated with the chi-squared statistic is 1.0. A high p-value (close to 1.0) suggests that we fail to reject the null hypothesis of independence.

The variables 'Transport by' and 'Month' appear to be independent based on the chi-squared test.

**Wilcoxon Test**

The value obtained from the Wilcoxon test are has follows:

Wilcoxon Statistic: 0.0

P-value:1.8679594215042613e-19

A low p-value indicates that we can reject the null hypothesis.With a p-value significantly less than the typical significance level of 0.05, we reject the null hypothesis. Therefore, there is a significant difference between the two groups.

**T\_Test**

The independent samples t-test was conducted, and the results are as follows:

T-test Statistic:1.4819092969568393

P-value: 0.14357065210144251

With a p-value greater than 0.05, we fail to reject the null hypothesis. Therefore, there is no significant difference between the two groups.

**Data Preprocessing and Transformation**

**Encoding Categorical Variables:**

The encoding is done to change the obect type data into numerical values for easier computer analysis for machine learning analysis and modeling.

**Logarithmic Transformation and Standard Scaling:**

Logarithmic Transform is used in this case as the skewness of hte data is to the right. The data is now scaled for modeling puposes.

**Machine Learning Models**

**Linear Regression**

The evaluated values are as follows

1. R-squared (Coefficient of Determination):

- Training Set: 0.0178

- Test Set: -0.2068

2. Root Mean Squared Error (RMSE):

- Test Set: 1.035

3. Mean Absolute Percentage Error (MAPE):

- Test Set: 104.37%

1. R-squared:The low R-squared values for both the training and test sets (close to zero or negative) indicate that the model doesn't explain much of the variability in the target variable.

2.Root Mean Squared Error (RMSE):The value of 1.035 indicates that, on average, the model's predictions are off by approximately 1.035 units from the actual values in the test set.

3.Mean Absolute Percentage Error (MAPE):The MAPE value of 104.37% suggests that, on average, the model's predictions deviate by around 104.37% from the actual values in the test set.

**Random Forest Regressor**

The random forest regression model's performance on the training and test sets is evaluated are as follows

1.R-squared (Coefficient of Determination):

- Training Set: 0.9899

- Test Set: 0.8286

2.Root Mean Squared Error (RMSE):

- Test Set: 0.39

1. Mean Absolute Percentage Error (MAPE):

- Test Set: 34.30%

1. R-squared: The high R-squared values for both the training and test sets (close to 1) indicate that the model explains a significant portion of the variability in the target variable.

2.Root Mean Squared Error (RMSE):The low RMSE value of 0.39 indicates that, on average, the model's predictions are close to the actual values in the test set.

3.Mean Absolute Percentage Error (MAPE):The MAPE value of 34.30% suggests that, on average, the model's predictions deviate by around 34.30% from the actual values in the test set.

**Decision Tree Regressor**

The decision tree regression model's performance on the training and test sets is evaluated are listed below

1. R-squared (Coefficient of Determination):

- Training Set: 1.0

- Test Set: 0.8287

2. Root Mean Squared Error (RMSE):

- Test Set: 0.3899

3. Mean Absolute Percentage Error (MAPE):

- Test Set: 31.40%

1.R-squared:A perfect R-squared value of 1.0 on the training set indicates that the decision tree perfectly fits the training data.

2.Root Mean Squared Error (RMSE): The low RMSE value of 0.3899 indicates that, on average, the model's predictions are close to the actual values in the test set.

3.Mean Absolute Percentage Error (MAPE): The MAPE value of 31.40% suggests that, on average, the model's predictions deviate by around 31.40% from the actual values in the test set.

**Lasso Regression**

The Lasso regression model's performance on the training and test sets is evaluated are:

1. R-squared (Coefficient of Determination):

- Training Set: 0.0%

- Test Set: -21.619%

2. Root Mean Squared Error (RMSE):

- Test Set: 103.901%

3.Mean Absolute Percentage Error (MAPE):

- Test Set: 99.453%

1.R-squared: A very low R-squared value on both the training and test sets (especially the negative value on the test set) suggests that the Lasso regression model is not explaining the variance in the target variable well.

2.Root Mean Squared Error (RMSE):The high RMSE value of 103.901% indicates that, on average, the model's predictions deviate substantially from the actual values in the test set.

3.Mean Absolute Percentage Error (MAPE): The MAPE value of 99.453% suggests that the Lasso model's predictions have a high percentage error relative to the actual values in the test set.

The R-squared values indicate poor explanatory power, and the high RMSE and MAPE values suggest that the model's predictions are not accurate.

**Logistic Regression for Binary Classification**

1.Accuracy: The scikit-learn Logistic Regression model achieved an accuracy of approximately 36%. However, the confusion matrix and classification report suggest that the model's performance, especially in predicting the positive class (1), is poor.

2.Statsmodels GLM Results: The coefficient values and their corresponding p-values are provided. It seems that only the constant term has a statistically significant impact, while the other features (x1, x2, x3) do not show significant effects.

3.Pseudo R-squared:The pseudo R-squared is quite low (0.04401), indicating that the model does not explain a substantial proportion of the variance in the dependent variable.

4.Model Evaluation:It's crucial to evaluate the model's performance comprehensively, considering precision, recall, and F1-score, especially when dealing with imbalanced datasets.

Below is the summary of all the models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Approach | R^2 Training set | R^2 Test set | Root means square error for prediction | Mean absolute percentage error for prediction |
| Linear Regression | 0.01783 | 0.8286 | 1.0349 | 104.36 |
| Random Forest Regressor | 0.9899 | 0.3437 | 0.3900 | 34.29 |
| Decision Tree Regressor | 1.0 | 0.8287 | 0.3898 | 31.40 |
| Lasso Regression | 0.0% | -21.619% | 103.901% | 99.453% |