**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Data Preparation and Visualization  Statistics for Data Analytics  Programming for Data Analytics  Machine Learning for Data Analytics |
| **Assessment Title:** | Machine Learning-Based Prediction and Comparative Analysis of Passenger Numbers in Air Transport between the UK and Ireland Considering Flight Frequency, Distance, Flight Type, and Travel Coverage |
| **Lecturer Name:** | 1. David McQuaid 2. Sam Weiss 3. Mohammed Iqbal 4. Taufique Ahmed |
| **Student Full Name:** | DIANA FLORA NAMAEMBA |
| **Student Number:** | 20233856 |
| **Assessment Due Date:** | 7th January 2024 |
| **Date of Submission:** | 6th 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 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. |

**Machine Learning-Based Prediction and Comparative Analysis of Passenger Numbers in Air Transport between the UK and Ireland Considering Flight Frequency, Distance, Flight Type, and Travel Coverage**

**Contents**

[**List of Acronyms** 4](#_Toc155465697)

[**1** **INTRODUCTION** 5](#_Toc155465698)

[**2** **METHOD** 5](#_Toc155465699)

[2.1 DATA PREPARATION & VISUALIZATION 5](#_Toc155465700)

[**2.1.1** **DATA ACQUISITION** 5](#_Toc155465701)

[**2.1.2** **EXPLORATORY DATA ANALYSIS (EDA)** 9](#_Toc155465702)

[**2.1.3** **DATA PREPARATION AND CLEANING** 9](#_Toc155465703)

[**2.1.4** **VISUALIZATION** 11](#_Toc155465704)

[2.2 STATISTICAL ANALYSIS 11](#_Toc155465705)

[**2.2.1** **Data Visualization** 11](#_Toc155465706)

[**2.2.2** **Descriptive Statistics** 12](#_Toc155465707)

[**2.2.3** **Inferential Statistics** 13](#_Toc155465708)

[2.3 MACHINE LEARNING 17](#_Toc155465709)

[**2.3.2** **Sentiment analysis** 21](#_Toc155465710)

[2.4 PROGRAMMING 25](#_Toc155465711)

[2.5 Reference 27](#_Toc155465712)

# **List of Acronyms**

* MAE Mean absolute error
* MAPE Mean absolute percentage error
* MSE Mean square error
* R2 R-squared, coefficient of determination
* RMSE root mean square error
* SMAPE symmetric mean absolute percentage error
* EDA Early Data Analysis
* MCAR missing completely at random
* CV coefficient of variation
* SD Standard Deviation
* IQR Interquartile ranges
* Min Minimum
* Max Maximum
* CDF cumulative density function
* KDD Knowledge discovery in databases
* SEMMA The sample, explore, modify, model, and assess
* CRISPM-DM The Cross-industry standard process for data mining
* ML Machine Learning
* UK United Kingdom
* CSV comma-separated values file
* TSV Tab-separated values
* API Application Programming Interface
* SRS Simple Random Sample
* SML Supervised Machine Learning
* VADER Valence Aware Dictionary and sEntiment Reasoner
* MNB Multinomial Naive Bayes algorithm
* TF-IDF Term Frequency - Inverse Document Frequency

**List of Figures**

[**Figure 1: Q-Q plot for Number of Passengers in Ireland** 14](#_Toc155455595)

[**Figure 2: Q-Q plot for Number of Passengers in UK** 14](#_Toc155455596)

[**Figure 3: Q-Q plot for Number of Flights in Ireland** 14](#_Toc155455597)

[**Figure 4:Q-Q plot for Number of Passengers in UK** 15](#_Toc155455598)

[**Figure 5: Data Science Cycle used for this project** 18](file:///C:\Users\Diana\Documents\Msc%20Data%20Analytics%20course%20work\CA%202%20Materials\CA%202%20Materials\Diana%20Flora%20Namaemba%20CA%202%20REPORT.docx#_Toc155455599)

**List of Tables**

[**Table 1: A total of 15 datasets were explored. Some of the databases explored included the following:** - 6](#_Toc155455586)

[**Table 2: Measures of Central Tendency and Dispersion of Number of Passengers and Number of Flights in UK and Ireland** 13](#_Toc155455587)

[**Table 3: The Aircrafts with the Most and Least Number of Passengers and Flights** 13](#_Toc155455588)

[**Table 4: Kolmogrov-Sminorv Test of Normality** 15](#_Toc155455589)

[**Table 5: Krus Wallis Test of UK and Ireland** 16](#_Toc155455590)

[**Table 6: Wilcoxon's Rank Sum Test of UK and Ireland** 16](#_Toc155455591)

[**Table 7: Pearson Correlation between Number of Passengers and Number of Flights.** 16](#_Toc155455592)

[**Table 8: Supervised ML Regression Algorithms for Ireland Data** 20](#_Toc155455593)

[Table 9: **Supervised ML Regression Algorithms for UK Data** 21](#_Toc155455594)

# **INTRODUCTION**

Over the years there has been massive progress in the aviation industry. Since the first engine flight, there have been numerous passengers flown, the establishment of thousands of airports, and aviation businesses. Countries with close connections to other countries have improved at international levels. Civil aviation has become a necessary mode of public transport in many countries because of its efficiency and simplicity.(Tolga and GÖKMEN, 2021) .Several factors determine the number of passengers in air transportation. These factors are important to manufacturers, airports, airlines, and the industry at large. (Kluge *et al.*, 2017). A lot of research has been done to understand the mobility behavior of European passengers.(Kluge *et al.*, 2017) Machine learning has been applied in various areas of air passenger research. It has been applied in making predictions about air passenger traffic(Xiong *et al.*, 2022). This study aims to develop a predictive model utilizing machine learning techniques to estimate and compare passenger numbers in air transport between the UK and Ireland, integrating features such as flight frequency, distance, aircraft, year, and travel coverage. The specific objectives included: - to evaluate the impact of flight frequency, distance, flight type, year, and travel coverage on predicting passenger numbers in both countries, analyzing variations in passenger numbers between the UK and Ireland based on flight frequency, distance, aircraft type, and year, distinguishing between national and international coverage, and to perform sentiment analysis on mode of public transport from the perspectives of producers and consumers in Ireland.

# **METHOD**

## DATA PREPARATION & VISUALIZATION

### **DATA ACQUISITION**

1. **Searching for Machine learning Transport Data (Dataset 1)**

The process of acquiring raw data was divided into two. One, reading various publications made on transport and Machine learning to get an understanding of what has been done. This made the formulation of research questions and research topic/area easy. Two, looking for datasets on transport in Ireland and other countries. This involved exploring various open-source government databases and websites that collect transport data that were easily accessible/ with manageable restrictions. Open government databases were preferred because they encouraged transparency, verification, and reliability of the data(de Juana-Espinosa and Luján-Mora, 2019).

**Table 1: A total of 15 datasets were explored. Some of the databases explored included the following:** -

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Database (URLs)** | **Dataset Name** | **Country** | **Positive** | **Negative (Limitation)** | **Licensing and permission for the Data** |
| [Eurostat](https://data.europa.eu/en) | vehicle traffic performance registered in the reporting country by transport coverage | UK had missing data so opted to use Finland. | The data was relevant to the research questions.  Data points/ variables were collected for both countries.  The units of measure were the same for all the variables between the two countries.  The time of evaluation was the same for both countries, making comparing them easy and reliable.  Data was stored in both JSON, CSV, and TSV, thus data could be manipulated using different formats | The dataset was very small, with a total of 44 observations for each country risking overfitting issues.  The number of variables was too few for any proper comparison between the countries to be done.  A lot of missing values that required dropping them as the choice of handling the missing values.  missing data points in one country making it hard to do a comparison with the other. | Licensing was available, one could use the data, provided citing the source.  [License](http://creativecommons.org/licenses/by/4.0/)  **You are free to:**  **Share**— copy and redistribute the material in any medium or format for any purpose, even commercially.  **Adapt**— remix, transform, and build upon the material for any purpose, even commercially. |
| [Eurostat](https://data.europa.eu/en) | Buses and coaches traffic performance registered in the reporting country by transport coverage | for UK data was missing so opted to use Poland | The data was relevant to the research questions.  Data points/ variables were collected for both countries.  The units of measure were the same for all the variables between the two countries.  The time of evaluation was the same for both countries, making comparing them easy and reliable.  Data was stored in both JSON, CSV, and TSV, thus data could be manipulated using different formats | The dataset was very small, with a total of 84 observations for each country risking overfitting issues.  The number of variables was too few for any proper comparison between the countries to be done.  A lot of missing values that required dropping them as the choice of handling the missing values.  missing data points in one country making it hard to do a comparison with the other. | [License](https://creativecommons.org/licenses/by/4.0/)  **You are free to:**  **Share**— copy and redistribute the material in any medium or format for any purpose, even commercially.  **Adapt**— remix, transform, and build upon the material for any purpose, even commercially. |
| [Eurostat](https://data.europa.eu/en)  [Transport for London](https://data.london.gov.uk/dataset/number-bicycle-hires) | DCC Dublinbikes  and the number of bikes | Dublin  London | Dublin bikes had relevant data on transport.  London bike had relevant transport data.  Data was stored in both JSON, CSV, and TSV, thus data could be manipulated using different formats | The two datasets didn’t have similar variables. Comparing them was difficult. | [License](https://opendefinition.org/licenses/cc-by/)  **You are free to:**  **Share**— copy and redistribute the material in any medium or format for any purpose, even commercially.  **Adapt**— remix, transform, and build upon the material for any purpose, even commercially.  [License](https://www.nationalarchives.gov.uk/doc/open-government-licence/version/2/)  **You are free to:**  copy, publish, distribute, and transmit the Information;  adapt the Information;  exploit the Information commercially and non-commercially for example, by combining it with other Information, or by including it in your product or application. |
| [Eurostat](https://ec.europa.eu/eurostat/databrowser/view/avia_paodis/default/table?lang=en) | Air passenger transport by aircraft model, distance bands, and transport coverage | Dublin and UK | Had relevant data for both the UK and Ireland. Comparing them would be easy.  -Data was stored in both JSON, CSV, and TSV, thus data could be manipulated using different formats |  | [License](https://creativecommons.org/licenses/by/4.0/)  **You are free to:**  **Share**— copy and redistribute the material in any medium or format for any purpose, even commercially.  **Adapt**— remix, transform, and build upon the material for any purpose, even commercially. |

The countries explored included Poland and Finland but due to a lack of uniformity of the variable between them and Ireland they were not utilized. The dataset “Air passenger transport by aircraft model, distance bands and transport coverage”, highlighted above was used for all the analyses in this report except for the sentiment analyses because it had data for Ireland and other countries, collected for similar variables. The UK was chosen due to its similarities in transport with Ireland.

1. **Searching for text data on transport in Ireland for sentiment analysis (Dataset 2)**

Sentiment analysis is processing and analyzing sentiments from text data.(Liaqat *et al.*, 2022).Sources of text data include; social media posts, newspaper articles, and marketing materials.(Bae, Yu Hung and van Lent, 2023). Social media provides real-time data in high volumes, rich in information, embedded in short text content, and videos and images.(Hou *et al.*, 2021). Reddit was used for this project because it is a popular social media platform where users can ask questions, share their views and opinions, and experiences.(Janchevski and Gievska, 2019). Several subreddits were chosen due to their relevance to transport in Ireland. They included: -

* r/irishtourism
* r/Dublin.
* r/ireland

The subreddit r/irishtourism was used because it had more comments on transport in Ireland that captured both consumer and producer points of view. The keywords used to scrape the comments are shown as follows:

**Consumers point of view keywords:** bus, buses, coach, rail, train, tram, Luas, bikes, bus Eireann, public transport.

**Producers point of view keywords:** driving in Ireland, I am a driver, I drive, my car, my vehicle, my company, driving my car, drive my car.

**Positive Aspects:** Reddit provided a vast pool of user-generated content providing diverse opinions and experiences related to modes of transport. The Reddit licence: allows one to use Reddit data API for research purposes provided you use it exclusively for academics. [Reddit Data API License](https://support.reddithelp.com/hc/en-us/articles/14945211791892)

**Negative Aspects:** It was challenging to come up with keywords that represented consumers' opinions about modes of transport. Due to a wide range of possible keywords.

### **EXPLORATORY DATA ANALYSIS (EDA)**

The primary aim of the EDA was to examine the data’s distribution, outliers, and any anomalies that would be used to generate specific hypotheses for testing and to assist in pattern recognition. (*Secondary Analysis of Electronic Health Records*, 2016) The steps involved in EDA were: -

**2.1.3.1 Checking the data dimensions:** High-dimension data is where the number of features or variables is larger than the number of observations. (Narisetty, 2020). Higher dimensional data pose statistical challenges, hence the need to check the data for dimensionality.(Johnstone and Titterington, 2009) Dataset 1 was a low-dimension data (12 features and 799008 observations). Dataset 2 had 730 observations and 3 variables. Both datasets were low-dimensional, making them easy to display and explore efficiently.

**2.1.3.2 Checking the data types:** understanding the data types helps determine the type of statistical analyses to be done and the best way to visualize the data. (Dettori and Norvell, 2018). Dataset 1 had (9 categorical, 2 continuous, and 1 date-time). Dataset 2 is comprised of text data.

**2.1.3.3 Checking for any missing data:** Missing data have major effects on conclusions made from the data. Therefore, identifying them is crucial for handling problems they cause.(Dettori and Norvell, 2018). Both datasets had missing data points.

**2.1.3.4 checking for duplicates:** Some impacts of duplicates include; the generation of erroneous observations, generation of more repeated observations, loss of observations, and incorrect statistics. (Cheng, no date) Both Datasets had duplicates.

**2.1.3.5 Checking for outliers**:Outliers are observations that are different from most of the observations. They change the results of the data. Identifying them is significant to maintain the results of the data.(Cousineau and Chartier, 2010). Dataset 1 had outliers. see Jupyter Notebook line 13, 14, 47,48

### **DATA PREPARATION AND CLEANING**

Data cleaning organizes data, making it ready for analysis. It helps identify and remove inconsistencies and errors in data, improving the data quality.(Ridzuan and Wan Zainon, 2019).

The Data cleaning steps included:

**Step 1: Handling missing data:** Handling missing data ensured the data was reliable, meaningful in analysis, and unbiased(Kang, 2013). The listwise deletion method was used to handle missing data in both datasets. Other techniques would alter the shape of the distribution.(Kang, 2013). Since dataset 2 was text, any other technique would not be applicable.

**Step 2: Removing features that were not used:** Removing irrelevant features helps overcome the curse of dimensionality and reduce overfitting problems.(Afshar and Usefi, 2022). In dataset 1, 5 features were removed. They were labels for other variables.

**Step 3: Removing duplicate Observations:** Duplicate observations were dropped because they could result in incorrect statistics. All the categorical variables had a category called Total which was equal to the summation of each category. This would cause multicollinearity hence the need to drop them.

**Step 4: Transforming the data using Merge**. The dataset was restructured using merge to separate the number of flights and number of passengers data, which were stored together.

**Step 5: Encoding Data:**There are several categorical data encoding techniques like; one hot encoding, ordinal encoding, label encoding, Helmert coding, polynomial coding, binary coding, and backward difference coding, etc.(Potdar, S. and D., 2017). Label and one hot encoding were explored and evaluated for their effects on the models. One hot encoding transformed individual variables with n categories into n new variables that are binary. Label encoder on the other hand assigned an integer to the categories as a label and didn’t add new categories.(Potdar, S. and D., 2017).

**Step 6: Data Splitting:** All datasets were split into independent variables (X) and the dependent variable (y). The X and y variables were split into Training and test sets as shown below: -

Training set: X\_train and y\_train included 80% of the X data and y data respectively.

Test set: X\_test and y\_test included 20% of the X data and y data respectively.

Dataset 1 had 624 observations and 12 features for the X\_train, 156 observations and 12 features for the X\_test, 624 observations for the y\_train, and 156 observations for the y\_test.

Splitting the datasets was very important because it helped find the most efficient set of model parameters that had the correct balance between the model complexity and the model’s generalization capabilities.(Eliane Birba, 2020).

**Step 7: Data Transformation*:*** is a technique that ensures data is in the best possible manner for machine learning algorithms.

(*OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059*, 2020).

It can be achieved through the following steps: -

1. ***Rescaling*** is rescaling the scale of all attributes if they are not of the same scale to the same scale.
2. ***Standardization*** is transforming attributes into a standard normal distribution.
3. ***Normalization*** is rescaling the observations to have a length of one.

Normalization was used because it is sensitive to outliers and it retains the shape of the original distribution. While standardization technique was used because the distribution of the data was unknown and it preserved the relationship between the data points.(Bhandari, 2023) The data splitting was applied to standardized and normalized data. Both techniques provided different machine-learning results.

**Step 8: Handling Outliers:**

Outliers affect significantly small sample size outcomes when one performs any robust statistical procedures such as coefficient of variation and variance, hence the need to handle them. (Cousineau and Chartier, 2010). There are over 20 techniques for handling outliers. The keep method (*which is a technique of handling outliers that acknowledges the presence of outliers but doing nothing about the outlier values considering any prior analysis*.) was chosen because the outlier was an influential outlier, i.e., accurate data points that are at a distance from the other points and are neither error nor interesting outliers.(Aguinis, Gottfredson and Joo, 2013)

### **VISUALIZATION**

#### **Heatmap for Correlation**

Heatmaps were plotted to check for correlated continuous variables in the UK and Ireland Data. They showed:

* A strong positive correlation between the number of flights and the number of passengers for both countries.
* A weak negative correlation between the time period and the number of flights for both countries.
* A weak positive correlation between time period and the number of passengers for both countries. Ref JupyterNotebook line 45.

#### **Box Plot**

Box plots were plotted for both UK and Ireland data for the number of passengers and number of flights over the years for each transport coverage. They both showed the presence of Outliers. Ref JupyterNotebook Line 47-48

#### **Use Tufte Principle**

According to Edward Tufte, there are 6 principles a visualization should strive toward, that is comparison rather than description, high resolution and utilization of classic designs, Content focus, concepts proven by time, and integrity. (Globus, 2014). From machine learning, UK untransformed Label encoded data was the best in making predictions about the number of passengers for some models not all while untransformed Label encoded data was the best in making predictions about the number of passengers for Ireland data.

**Dashboard**

The tufte principle was utilized to create a dashboard that communicates findings to different Air transportation stakeholders that is passengers, the Data Science team, and aircraft companies.

Information on the comparison of the best models was plotted using a bar graph and the most and least common aircraft were also plotted for national and International Transport. Passengers are required to know the most common and least common aircraft to help in managing their travels. A bar plot was used because the data was categorical. Findings from Machine learning were relevant to the data science team since they could know which model is better when making predictions about air passengers’ numbers. Aircraft companies need to know which aircraft is common to ensure targeted Ref Jupyter Notebook

## STATISTICAL ANALYSIS

Descriptive and inferential statistics were used in this study. Descriptive was used to understand the past data while inferential analysis was used to make to make inferences about dataset 1 using a sample drawn using simple random sampling.

### **Data Visualization**

**Histograms**

were plotted for the number of passengers by transport coverage for the UK and Ireland. All 8 plots showed the data for each was rightly skewed, meaning that while over the years different aircraft and distances in various transport coverage had smaller numbers of passengers there were occasional instances with more passengers for both countries. Ref Jupyter Notebook line 49-50.

### **Descriptive Statistics**

Descriptive statistics included calculating the measures of central tendencies: (mean, median, and frequencies) and the measures of dispersion (Variance, standard deviation) for the number of passengers.

**Measures of central tendency and dispersion**

These are measures of statistics that use a single value as a representative of the entire distribution. (Manikandan, 2011). When data is normally distributed, mean and standard deviation are the best measures of central tendency and variability of data. When data is non-parametric the best measures are median and interquartile ranges.(McCluskey and Lalkhen, 2007) The mean, median, mode, variance, interquartile range, and standard deviation were calculated for the number of passengers, number of flights, and number of passengers vs transport coverage.

**Table 2: Measures of Central Tendency and Dispersion of Number of Passengers and Number of Flights in UK and Ireland**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Country** | **Variable** | **n (Mean)** | **Median (SD)** | **IQR** | **Min-Max** |
| Ireland | Number of Passengers | 780(454638.1) | 311.5(1280187.6) | 0-63760.25 | 0-7587947 |
| Number of Flights | 780(3507.8) | 4.0(9017.1) | 0-506.25 | 0-60662 |
| UK | Number of Passengers | 780(3570163.8) | 139672.0(9049416.0) | 0-1668302.0 | 0-57817776 |
| Number of Flights | 780(27100.1) | 1259.5(58120.5) | 0-21247.5 | 0-310181 |

**Table 3: The Aircrafts with the Most and Least Number of Passengers and Flights**

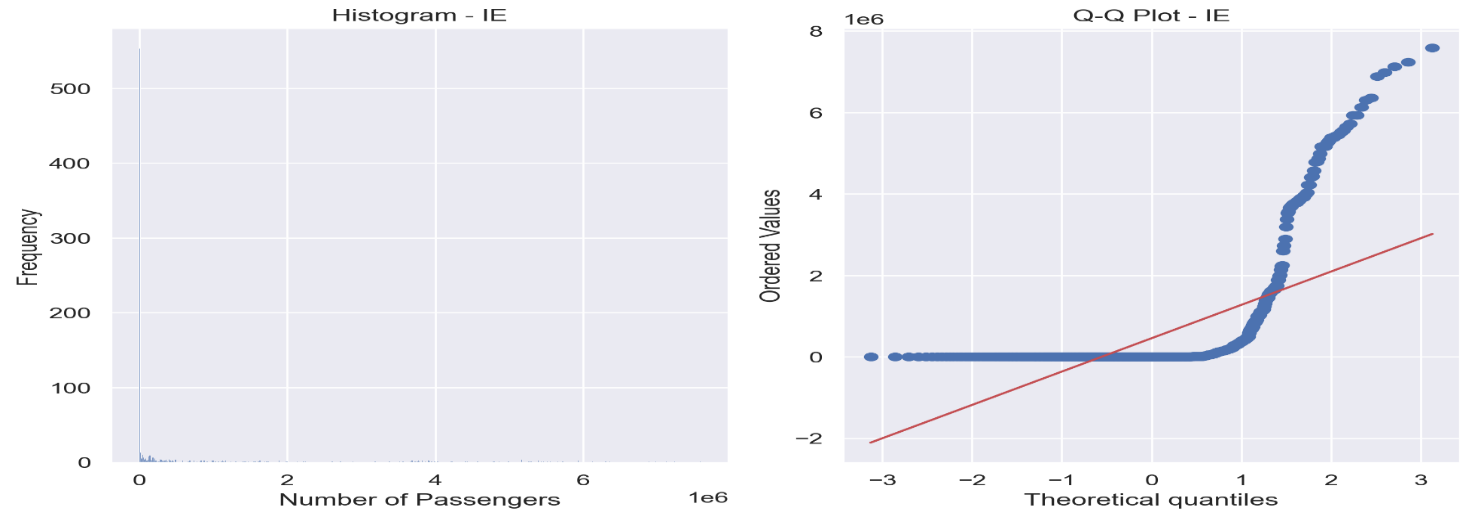
|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Country** | **Transport Coverage** | **Summary** |
| Number of Passengers | Ireland | National Transport | The Aircraft with the most and least number of passengers were AC\_RT and AC\_JJ respectively. |
| International Transport | The Aircraft with the most and least number of passengers were AC\_NJ and AC\_RJ respectively. |
| UK | National Transport | The Aircraft with the most and least number of passengers were AC\_NJ and AC\_JJ respectively. |
| International Transport | The Aircraft with the most and least number of passengers were AC\_NJ and AC\_RT respectively. |
| Number of Flights | Ireland | National Transport | The Aircraft with the most and least number of flights were AC\_RT and AC\_JJ respectively. |
| International Transport | The Aircraft with the most and least number of flights were AC\_NJ and AC\_JJ respectively. |
| UK | National Transport | The Aircraft with the most and least number of flights were AC\_RT and AC\_JJ respectively. |
| International Transport | The Aircraft with the most and least number of flights were AC\_NJ and AC\_JJ respectively. |

### **Inferential Statistics**

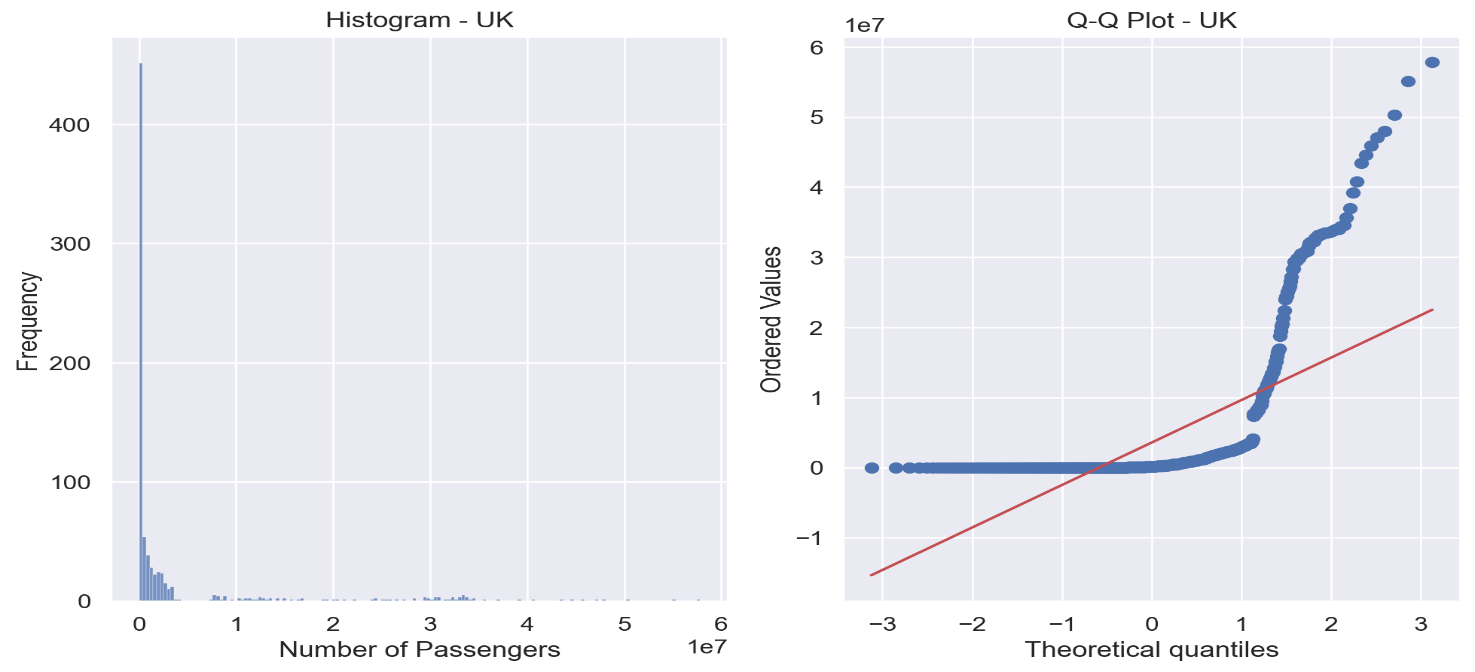
#### **Normality Tests (Normal Distribution)**

This was done using Q-Q plots. The number of passengers and the number of flight variables were assessed. They were not normally distributed for both countries. Ref JupyterNotebook line 51-52. Non-parametric tests were therefore applied to draw inferences from them (McCluskey and Lalkhen, 2007).

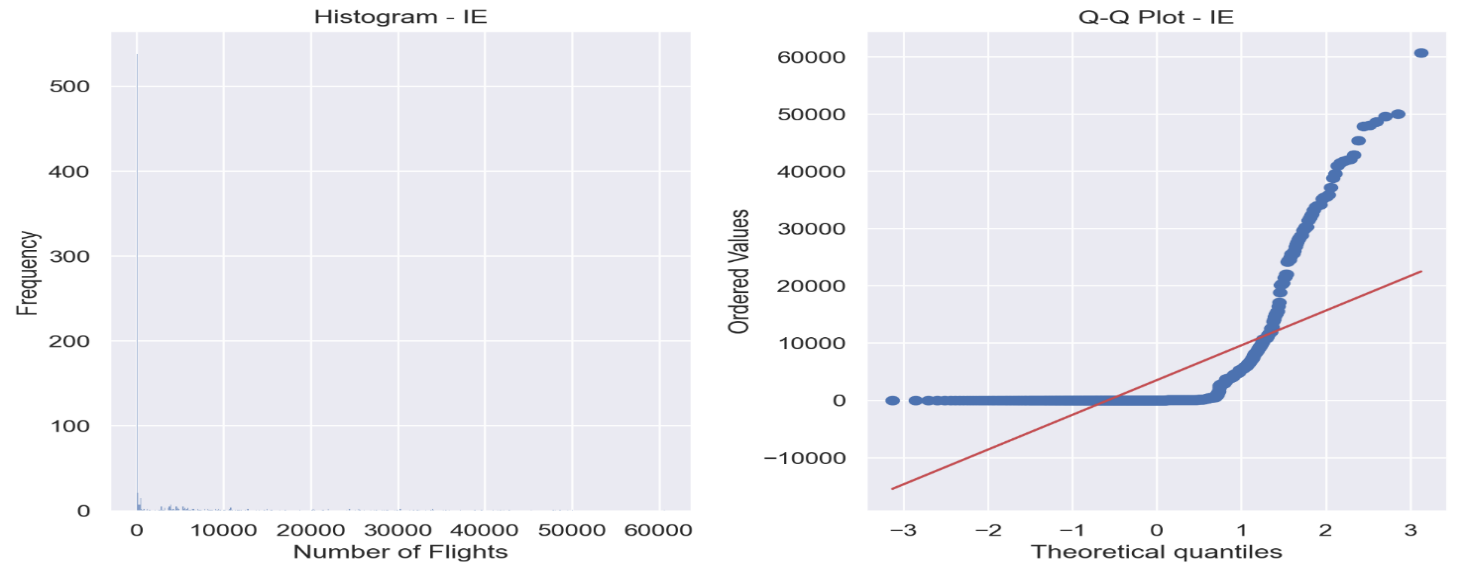
**Figure 1: Q-Q plot for Number of Passengers in Ireland**



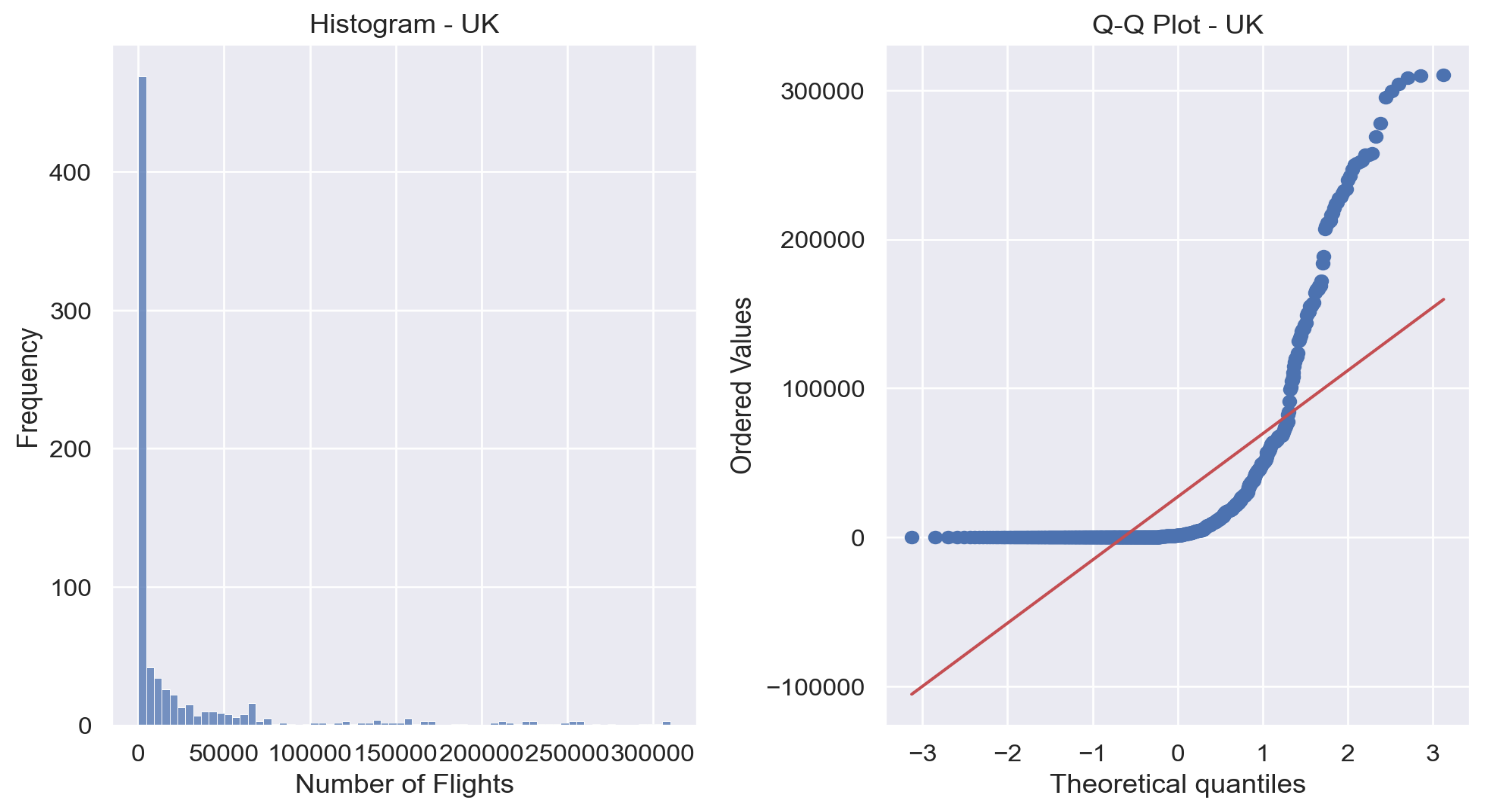
**Figure 2: Q-Q plot for Number of Passengers in UK**



**Figure 3: Q-Q plot for Number of Flights in Ireland**



**Figure 4:Q-Q plot for the Number of Passengers in the UK**



#### **Sampling**

Simple Random sampling (SRS) was used to draw a sample from the data. SRS was used because it ensures unbiased, representative, and equal probability of the population.(Tajik and Golzar, 2022) A sample size of 200 was drawn, a total of 400 (203 Ireland and 197 UK) observations were sampled and the sample mean was calculated. Inference about the population was made using the sample.

**Ireland:** The sample mean was calculated for the number of passengers. The value was 451415.4. The sampling proportion for travel coverage (National transport) was determined to be 52%.

**UK:** The sample mean was calculated for the number of passengers. The value was 3459238.6. The sampling proportion for travel coverage (National transport) was determined to be 48%.

#### **Hypothesis Testing**

This is the method of determining the probability of an event that is observed to occur by chance. (Allua and Thompson, 2009). It helps to draw inferences about a population using a sample. Various hypotheses were formulated to draw inferences about the population.

##### **Normality test for the sample**

Kolmogorov-Smirnov test was used to test for normality using p-values. It was used because it is better for sample sizes greater than or equal to 50 (Mishra *et al.*, 2019).

Set the hypothesis to test for normality:

*H0: The number of passengers for each country follows a normal distribution*

*H1: The number of passengers for each country does not follow a normal distribution*

**Table 4: Kolmogrov-Sminorv Test of Normality**

|  |  |  |  |
| --- | --- | --- | --- |
| **Country** | **Test-statistic** | **P-value** | **Conclusion** |
| Ireland | 0.55 | 2.366e-56 | The p-value is < 0.05, therefore reject Ho and conclude that the Number of Passengers is not normally distributed |
| UK | 0.77 | 9.045e-124 | The p-value is < 0.05, therefore reject Ho and conclude that the Number of Passengers is not normally distributed |

##### **Non-Parametric Tests**

These are tests applied when data is not normally distributed or data doesn’t adhere to the assumptions of normally distributed data(Nahm, 2016).

###### **Krus Wallis Test**

Analyses variances. It analyses the difference between the median values of independent samples.(Nahm, 2016)

The hypothesis was: -

*Null Hypothesis (H0): The median between the UK and Ireland is equal for the Number of Passengers*

*Alternative Hypothesis (H1): The median between the UK and Ireland is not the same for Number of Passengers*

**Table 5: Krus Wallis Test of UK and Ireland**

|  |  |  |
| --- | --- | --- |
| **Test-statistic** | **P-value** | **Conclusion** |
| 50.88 | 9.808e-13 | The p-value is < 0.05, reject the null hypothesis and conclude the median between the UK and Ireland is not the same for the Number of Passengers Variable |

###### **Wilcoxon’s rank sum test**

Ranks all data points in order and calculates the rank sum of each sample then compares the difference in the rank’s sums. Groups with similar scores have similar rank sums, otherwise, the ranks are different.(Nahm, 2016)

It was applied to test the hypothesis below: -

*Null Hypothesis (H0): UK and Ireland Numbers of Passengers have the same or are drawn from the same distribution*

*Alternative Hypothesis (H1): The UK and Ireland Numbers of Passengers do not have the same or are not drawn from the same distribution*

**Table 6: Wilcoxon's Rank Sum Test of UK and Ireland**

|  |  |  |
| --- | --- | --- |
| **Test-statistic** | **P-value** | **Conclusion** |
| 6.99 | 2.658e-12 | The p-value is < 0.05, therefore reject Ho and conclude that the UK and Ireland numbers of passengers do not have the same or are not drawn from the same distribution |

###### **Pearson Correlation**

Measures the strength of the linear relationship between two continuous variables. Its value ranges from -1 to 1. -1 is a negative linear correlation while 1 is a positive one.(Williams *et al.*, 2020)

*Null Hypothesis (H0): There is no significant linear relationship between the Number of Flights and the Number of Passengers.*

*Alternative Hypothesis (H1): There is a significant linear relationship between the Number of Flights and the Number of Passengers.*

**Table 7: Pearson Correlation between Number of Passengers and Number of Flights.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Country** | **Category** | **Test-statistic** | **P-value** | **Conclusion** |
| Ireland | International Transport | 0.97 | 4.45e-246 | the p-value is < 0.05, reject Ho, conclude that there is a significant linear relationship between the number of flights and passengers |
| National Transport | 0.84 | 9.75e-106 | the p-value is < 0.05, reject Ho, conclude that there is a significant linear relationship between the number of flights and passengers |
| UK | International Transport | 0.97 | 5.115e-64 | the p-value is < 0.05, reject Ho, conclude that there is a significant linear relationship between the number of flights and passengers |
| National Transport | 0.87 | 4.25e-30 | the p-value is < 0.05, reject Ho, conclude that there is a significant linear relationship between the number of flights and passengers |

###### **Linear regression**

is a statistical analysis technique that studies the linear relationship between a continuous dependent variable and one independent variable**.**(Schneider, Hommel and Blettner, 2010). A linear regression model was fit for the number of passengers (y) vs other independent variables

The hypothesis was: -

*Null Hypothesis (H0): There is no significant linear relationship between the independent variable and the dependent variable.*

*Alternative Hypothesis (H1): There is a significant linear relationship between the independent variable and the dependent variable*

**Ref Jupyter Notebook 81-87**

**Ireland**

* In the Distance Variable, no category had a linear relationship with the dependent variable (The p-values were greater than 0.05).
* In the Aircraft Variable, only aircraft AC\_NJ had a linear relationship with Y (the p-value was less than 0.05).
* In transport coverage, National Transport has a linear relationship with Y. (the p-value is less than 0.05)

**UK**

* In the Distance Variable, distance\_KN\_GE2000 and distance\_KM\_LT300 had a linear relationship with the dependent variable (The p-values were less than 0.05).
* In the Aircraft Variable, only aircraft AC\_NJ had a linear relationship with Y (the p-value was less than 0.05).
* In transport coverage, National Transport has a linear relationship with Y. (the p-value is less than 0.05)

## MACHINE LEARNING

Machine learning involved the consideration of three project management frameworks: CRISP-DM, KDD, and SEMMA. Knowledge discovery in databases (KDD) selects the target data, pre-processes, transforms, data mines, and interprets it. The sample, explore, modify, model, and assess (SEMMA) samples, explores, modifies, models, and assesses the data by evaluating the results. The Cross-industry standard process for data mining (CRISP-DM) uses business understanding, data understanding, data preparation, data modeling, evaluation, and deployment of the results. (Martins, Pesado and García-Martínez, 2016).

For this study the project management framework used was the one below, it is close to CRISP-DM.

**Figure 5: Data Science Cycle used for this project**

Supervised Machine Learning (SML) was chosen for transport\_data. SML is machine learning that makes predictions by forecasting or classifying specific outcomes of interest.(Jiang, Gradus, and Rosellini, 2020) Regression is a supervised machine learning algorithm that makes predictions of a continuous dependent variable (number of passengers) based on one or more independent variables.(Sarker, 2021) was used.

For dataset 2, classification a supervised learning technique that makes predictions about the dependent variables(sentiments) that are class variables using one or more independent variables was used. Sentiments of transport in Ireland from both the producers' and consumers' point of view were predicted.

**The machine learning process involved the following steps: -**

1. ***Encoding data with both Label and One hot encoding***
2. ***Splitting the dataset into training and test data***
3. ***Data Transformation (using standardization and normalization).***
4. ***Building Regression Machine Learning Models.***
5. **Regression analysis**

#### **Regression Analysis**

regression analyses involved testing 7 regression algorithms. The models were fit for both training and test data and evaluated for label vs one hot encoding and when data is transformed using (Normalization vs standardization) vs when data is not transformed. The regression models were: -

1. **Linear regression:** creates a relationship between a dependent variable and an independent variable by using the best fit straight line.(Sarker, 2021). Multiple Linear regression was performed to determine if there was a relationship between the number of passengers and the following independent variables: -
   * + Number of flights
     + Distance covered by each aircraft
     + Type of aircraft.
     + Year
     + Transport coverage.
2. **Decision Tree Regression:** makes predictions of a continuous variable by forming decision trees by asking a series of questions and creating decision rules according to the dataset structure that constitutes the problem.(KOCARIK GACAR and DEVECİ KOCAKOÇ, 2020). A decision tree was plotted for the best-performing algorithm.
3. **Ridge and Lasso regression:** are used when building models with large number of features because of their capabilities of preventing overfitting and reducing model complexities. Lasso which uses the L1 regularization technique finds subsets of the independent variables that minimize the error of prediction for a continuous variable. Ridge on the other hand uses the L2 regularization i.e. is the squared magnitude of coefficients and forces the weights to be small but the coefficient value is never set to zero.(Sarker, 2021) Both ridge and lasso regressions were fit with and without gridsearchCV and results were compared for both countries. Their parameters were adjusted for different alpha values. The alpha value for the ridge was alpha = (10, 0.1, 1) and alpha= (0.01,1, 0.0001) for lasso.
4. **ElasticNet Regression:** combines the two penalized regression techniques, i.e. lasso and ridge to the advantage of both. It was fit because it is superior to Lasso and ridge since it combines the shrinkage effects of the ridge and feature selections of Lasso(Saleh, Layous and Republic, 2022). The alpha value was 0.01.
5. **Random forest regression*:*** uses a collection of tree predictors to make predictions about a continuous target variable.(Segal, 2003).
6. **Support vector regression:** maximizes the distance between the separating hyperplane and then trains the samples that are close to that hyperplane.

(*OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059*, 2020).

1. **K-NN regression:**makes predictions about a continuous target variable by identifying the K observations nearest to the new point we want to predict.(*Chapter 7 Regression I: K-nearest neighbors | Data Science*, 2023). KNN was fit for transport\_data for 5 Neighbors.
2. **gridsearchCV**is a hyperparameter tuning that picks out a grid of hyperparameter values evaluates them, then returns the one that is the best. It was applied to Ridge, Lasso, and support vector machine. The GridsearchCV on ridge regression used 5 folds for each of the 3 alpha (10, 0.1,1) parameters. A total of 15 fits were evaluated.On lasso regression, it was Fit on 5 folds for each of the 3 alpha (0.01, 0.0001,1) parameters. A total of 15 fits were evaluated. On Support Vector Machine Regression, the parameters were C (0.01,0.1,1,10,100,1000) and gamma (0.01,0.1,0.01,0.001,0.001), verbose =4.

#### **Model Evaluation**

The coefficient of determination R2 was used to evaluate the model’s performance. It takes the range of values from (-infinity, 1], according to the mutual relation between the prediction model and the ground truth. (Chicco, Warrens and Jurman, 2021). It was preferred to SMAPE, MAPE, MAE, MSE, and RMSE, because it was the most informative rate in many model evaluation cases.(Chicco, Warrens and Jurman, 2021). A negative R-squared indicated that the model performed poorly.

**IRELAND**

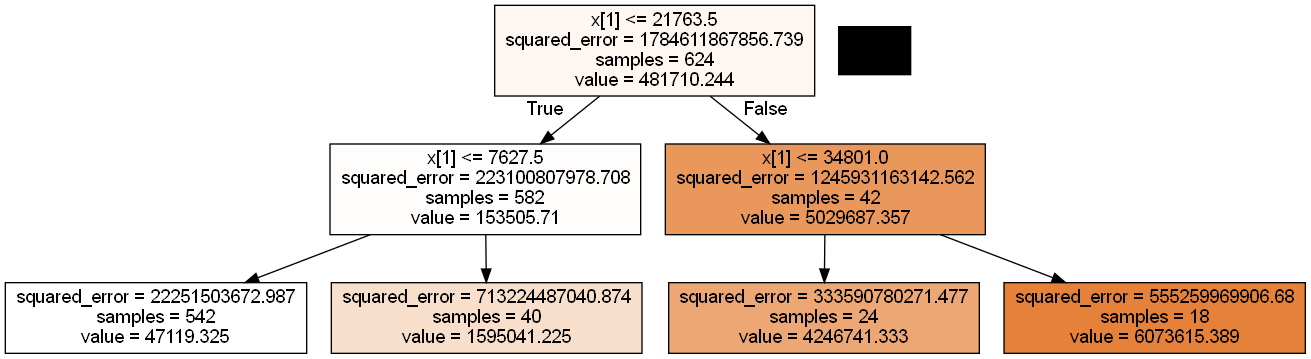
**Table 8: Supervised ML Regression Algorithms for Ireland Data**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Untransformed Data** | | **Normalized Data** | | **standardized Data** | |
|  | **ONE-HOT ENCODED** | **LABEL ENCODED** | **ONE-HOT ENCODED** | **LABEL ENCODED** | **ONE-HOT ENCODED** | **LABEL ENCODED** |
| Multiple Linear Regression | 0.95 | 0.97 | 0.95 | 0.94 | 0.95 | 0.93 |
| K-Nearest Neighbors regression  (neighbors = 5) | 0.96 | 0.97 | 0.75 | 0.74 | 0.89 | 0.97 |
| Decision Tree Regression | 0.91 | 0.96 | 0.92 | 0.91 | 0.93 | 0.83 |
| Random Forest Regression | 0.98 | 0.995 | 0.99 | 0.99 | 0.98 | 0.96 |
| Ridge Regression (alpha=1) | 0.95 | 0.97 | 0.95 | 0.94 | 0.95 | 0.93 |
| Ridge Regression (alpha=10) | 0.95 | 0.97 | 0.80 | 0.77 | 0.95 | 0.93 |
| Ridge Regression(alpha=0.1) | 0.95 | 0.97 | 0.95 | 0.94 | 0.95 | 0.93 |
| Ridge Regression(gridSearchCV)  Best Parameter | 0.95 | 0.97  Alpha=10 | 0.95 | 0.94  Alpha=0.1 | 0.95  Alpha =1 | 0.93  Alpha=0.1 |
| Lasso Regression | 0.95 | 0.97 | 0.95 | 0.94 | 0.95 | 0.93 |
| Lasso Regression(alpha=0.01) | 0.95 | 0.97 | 0.95 | 0.94 | 0.95 | 0.93 |
| Lasso Regression (alpha=0.0001) | 0.95 | 0.97 | 0.95 | 0.94 | 0.95 | 0.93 |
| Lasso Regression(gridSearchCV)  Best Parameter | 0.95 | 0.97  Alpha = 1.0 | 0.95 | 0.94  Alpha=1.0 | 0.95  Alpha = 1 | 0.93  Alpha=1.0 |
| Support vector Machine | -0.12 | -0.14 | -0.11 | -0.10 | -0.15 | -0.11 |
| Support Vector Machine (gridsearchCV)  Best Parameter | -0.13  C=1000  Gamma=0.001 | -0.11  C=1000  Gamma=0.001 | -0.10 | -0.10  C=1000  Gamma=0.1 | -0.14  C=1000  Gamma=0.1 | -0.09  C=1000  Gamma=0.1 |
| Elastic Net Regression (alpha=0.01) | 0.95 | 0.97 | 0.92 | 0.91 | 0.95 | 0.93 |

From the table above all the models with r-squared values above 80% were considered to be good models for making predictions about the number of passengers. This meant that 80% of the variability observed in the target variable could be explained by the regression models. The support vector machine was not a good model for predicting the number of passengers, because it performed poorly.

Label-encoded data that didn’t undergo any standardization or normalization was the best in making predictions about the number of passengers.

**Figure 6: Visualization of The Decision Tree of the Label Encoded Data that did not undergo any feature scaling**

****

**UNITED KINGDOM**

Table 9: **Supervised ML Regression Algorithms for UK Data**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Untransformed Data** | | **Normalized Data** | | **standardized Data** | |
|  | **ONE-HOT ENCODED** | **LABEL ENCODED** | **ONE-HOT ENCODED** | **LABEL ENCODED** | **ONE-HOT ENCODED** | **LABEL ENCODED** |
| Multiple Linear Regression | 0.95 | 0.82 | 0.96 | 0.93 | 0.95 | 0.92 |
| K-Nearest Neighbors regression  (neighbors = 5) | 0.96 | 0.92 | 0.84 | 0.66 | 0.98 | 0.99 |
| Decision Tree Regression | 0.91 | 0.85 | 0.92 | 0.92 | 0.91 | 0.94 |
| Random Forest Regression | 0.99 | 0.98 | 0.99 | 0.98 | 0.99 | 0.99 |
| Ridge Regression (alpha=1) | 0.95 | 0.82 | 0.95 | 0.93 | 0.95 | 0.92 |
| Ridge Regression (alpha=10) | 0.95 | 0.82 | 0.95 | 0.83 | 0.95 | 0.92 |
| Ridge Regression(alpha=0.1) | 0.95 | 0.82 | 0.96 | 0.93 | 0.95 | 0.92 |
| Ridge Regression(gridSearchCV)  Best Parameter | 0.95 | 0.82  Alpha=10 | 0.96 | 0.93  Alpha=0.1 | 0.95  Alpha=1.0 | 0.92  Alpha=1.0 |
| Lasso Regression | 0.95 | 0.82 | 0.96 | 0.93 | 0.95 | 0.92 |
| Lasso Regression(alpha=0.01) | 0.95 | 0.82 | 0.96 | 0.93 | 0.95 | 0.92 |
| Lasso Regression (alpha=0.0001) | 0.95 | 0.82 | 0.96 | 0.93 | 0.95 | 0.92 |
| Lasso Regression(gridSearchCV)  Best Parameter | 0.95 | 0.82  Alpha=1.0 | 0.96 | 0.93  Alpha= 1.0 | 0.95  Alpha= 1.0 | 0.92  Alpha=1.0 |
| Support vector Machine | -0.17 |  | -0.16 | -0.16 | -0.14 | -014 |
| Support Vector Machine (gridsearchCV)  Best Parameter | -0.17 | -0.13  C=1000  Gamma=0.001 | -0.16 | -0.16  C= 1000  Gamma=0.1 | -0.14  C=1000  Gamma=0.1 | -0.14  C=1000  Gamma=0.1 |
| Elastic Net Regression (alpha=0.01) | 0.95 | 0.82 | 0.93 | 0.91 | 0.95 | 0.92 |

From the table above, all the models with r-squared values above 80% were considered to be good models for making predictions about the number of passengers. This reveals that 80% of the variability observed in the target variable is explained by the regression models. Support vector machine is not a good model for predicting the number of passengers, because it performed poorly.

Unlike Ireland, the UK label-encoded data that didn’t undergo any standardization or normalization was the best in making predictions about the number of passengers for some models not all.

### **Sentiment analysis**

Sentiment analysis is processing and analyzing opinions, and sentiments of people towards issues or topics.(Liaqat *et al.*, 2022). Reddit platform provided text data that was utilized for sentiment analysis of the modes of public transport in Ireland from the consumer and producer points of view.

#### **Data Collection**

The Irish tourism subreddit was utilized. Titles, posts, and comments were searched using the search query “Transport in Ireland”. Further, the comments were searched for keywords related to each group.

***Consumer of modes of public transport was*** defined as people who use public transport or passengers. To extract modes of transport, Titles related to transport usage were extracted and keywords related to modes of transport were used to further extract comments related to transport usage by passengers. The keywords used were: - bus, buses, coach, train, bike, rail, bus Eireann, public transport, tram, Luas, I use the bus, I use the dart.

***Producers of modes of public transport*** were defined as drivers of any public transport. For sentiments on modes of transport from producers’ point of view, the keywords used to extract comments from the posts related to the usage of public transport in Ireland were: - driving in Ireland, I am a driver, I drive, my car, my vehicle, my company, drive my car, driving my car.

#### **Early Data Analysis**

After extracting the data from Reddit the data was stored in CSV and JSON format. Early Data analysis was conducted on the transport Ireland CSV dataset. The transport dataset for consumers had 3 variables (Post Title, Post Body, and comments) and 764 comments. The data was checked for any duplicates. A total of 3 duplicates were found from the comment variable. The duplicates were dropped. The data was also checked for any missing data points and there were no missing data points. After data cleaning the data had 761 observations and 3 variables. The JSON transport Ireland data was also viewed and the number of dictionaries was identified. The producer dataset had 23 comments and 3 variables.

#### **Sentiment Analysis Process**

Sentiment analysis for transport\_ireland CSV data was conducted using VADER sentiment analysis and TF-IDF for both consumers and producers. The Transport Ireland Json dataset was used to conduct VADER sentiment analysis for consumers only.

##### **Using VADER for Sentiment Analysis**

VADER is a rule and lexicon-based sentiment analysis tool that handles words, slang, emojis, and abbreviations that are normally found in social media. When compared to machine learning algorithms it is much faster and training of the data is not required. It was used for sentiment analysis of both consumer and producer data because it can handle various characters that social media data has. (Pano and Kashef, 2020).

**CSV Data Format**

Each body of the comment produced a vector of sentiment scores that has polarities: positive, negative, neutral, and compound. The compound polarity obtained for both producer and consumer data was used as the aggregate measure of all the sentiments in a comment. After obtaining the sentiment of each comment from the compound the sentiments were then used to make predictions. The comments were preprocessed by using a regular expression tokenizer that tokenized the text based on the regular expression [a-zA-Z0-9]+ followed bytheUse of CountVectorizer to convert the comments into a matrix of token counts by removing stop words and considering only single words, therefore, creating features that were based on single words in the comments.

**JSON Format**

The data extracted from Reddit was saved as a JSON File. The keys of the data were checked. The data had 3 keys, Post Title, Post Body, and Comments. The polarities of the texts were obtained for each comment using VADER and a compound value was also obtained. The sentiments were classified into neutral, positive, or negative by utilizing the compound value. If the compound ≥0.05 the sentiment is positive, if the compound is ≤0.05 the sentiment is negative otherwise it is neutral. The comments were tokenized using RegexpTokenizer followed by converting the text into a matrix using CountVectorizer.

**Multinomial Naïve Bayes classification**

After preprocessing the resulting matrix was split into X and Y, then split into training and test data. The matrix was high-dimensional, meaning had more features than observations. Multinomial Naïve Bayes (MNB) classification model was fit. The multinomial naïve Bayes works with the assumption that the document is a bag of words and takes into account the word frequency and information.(Abbas *et al.*, 2019).

The accuracy of the model was 76% for the Vader model on consumer data. This means that 76% of the predictions are true/correct.

The accuracy of the model was 50% for the Vader model on producer data. This means that 50% of the predictions are true/correct.

##### **Using TF-IDF for Sentiment Analysis**

Term frequency-inverse document frequency (TF-IDF) was used for sentiment analysis because it is preferred for natural language processing. TF-IDF is a scheme that assigns weights to token frequencies in the form of matrices. (Dogra *et al.*, 2022)

1. **Text Pre-processing**

The text data extracted from Reddit was stored in CSV format and imported for preprocessing. Preprocessing text is important since it helps to remove noise from text and reduce inconsistencies to ensure the data can be used for sentiment analysis of mining text.(Samuels and Mcgonical, 2019). Preprocessing involved the comments being tokenized. Tokenization is breaking several sentences into tokens. Tokens could be either words, symbols, phrases, or even the whole sentence.

The pre-processing includes: -

1. **EDA:** The number of words, characters, upper cases, special characters, and stop words in each comment were counted.
2. **Convert uppercase to lowercase**
3. **Stop words:** The number of stop words was counted for each comment and then removed from the comments.

**iii.) Obtaining unique words.** For consumer data: Unique words were obtained. The first 10 unique words were obtained. They were Dublin, bus, get, Galway, day, public, train, Ireland, transport and care. For producer data: Unique words were obtained. The first 10 unique words were obtained. They were driving, Ireland, drive, get, Dublin, don’t, car, like, day and 2.

iv.) **Lemmatization** is the process of finding the root of a word rather than the stem. (S *et al.*, 2020) . Lemmatization was applied to the comment variable to obtain the root of the word. Stemming was done but the output did not convey any meaningful information, so lemmatization was done because the root of the words was more meaningful. Below Is the process of text processing: -

**v) Obtaining sentiment:** After lemmatization the sentiment of the comments was determined for each comment. Using the polarity scores the sentiment was classified as either positive, negative, or neutral.

**vi.) TfidfVectorizer:** The tokens were converted into a numerical format using TfidfVectorizer. This is a vectorizer that uses the term frequency-inverse document frequency by calculating two matrices and representing the document as vectors for analysis.(Das Sarit Chakraborty Student Member and Member, 2018)

**vi.) Text Classification:** The resulting matrix from TfidVectorization was classified using Multinomial Naïve Bayes classification and the results were evaluated using classification evaluation metric accuracy

**Model Evaluation**

**Consumer:** The accuracy was 76%, meaning using TfidfVectorizer and naïve Bayes classification in making predictions about sentiments of Ireland consumers on the mode of transport in Ireland, 77% of predictions were made correctly.

**Producer:** The accuracy was 80%, meaning using TfidfVectorizer and Multinomial naïve Bayes classification making predictions about sentiments of Ireland producers on the mode of transport in Ireland, 80%prediction were made correctly.

|  |  |  |
| --- | --- | --- |
| **Metric** | **Accuracy (%)** | |
|  | **VADER sentiment Analysis** | **TfidVectorizer** |
| **Consumer** | 76 | 76 |
| **Producer** | 50 | 80 |

From the table above TF-IDF is better for making predictions on consumer and producer sentiments on modes of transport in Ireland than VADER.

## PROGRAMMING

#### **1. Programming**

Programming for data visualization, statistical analysis, and machine learning were all executed in the jupyter notebook

#### **Data structures**

Data gathered for machine learning and statistics was data stored in CSV format from the Eurostat open data source. All the data was processed for EDA, statistics, and machine learning analyses

The sentiment analysis data was extracted from the Reddit app using Reddit API. The data was saved as a JSON file for consumer data analysis using VADER sentiment analysis. The other datasets were saved in CSV format. Text processing was done.

#### **Code Choices**

1. **Data Preparation and Visualization**

Several Python libraries were used for Data Preparation and visualization for this study. The pandas, NumPy, matplotlib, and seaborn were used. The pandas were used for EDA, data preprocessing, and data cleaning processes. Matplotlib, plotly. express and seaborn libraries were used for data visualization for both EDA, visualization after data cleaning, and Dashboard creation.

This project employed various Python programming data structures. Lists, Dictionaries, and Data frames to perform various data preparation and visualizations. Loops were used to create functions for the execution of various data preparation and visualization processes.

1. **statistical analysis**

in statistical analysis, various statistical libraries were used. The selection of statistical methods to be executed depended on the type of statistical analysis to be done. When sampling data the statistics library was preferred over numpy because the mean and variance calculated are from a sample. Statistics library was used to perform various inferential statistics since inference was being drawn from sample data about the population

1. **Machine Learning**

specific libraries were used such as NLKT, and VADER. These libraries were chosen because they are used for sentiment analysis and tokenization. NLKT was used because of its extensive collection of documents. Vader Library was used because of its ease of use and its suitability for sentiment analysis.The preprocessing library was utilized more for machine learning problems due to the library having many options for various machine learning steps eg preprocessing, feature engineering, model development, and evaluation.

#### **Testing & Optimization**

1. **Data Preparation and Visualization**

Conducted initial data checks for missing, duplicates and outliers. Performing various data cleaning processes and checking if they have been executed.Dropping the missing data and checking if they had been dropped was another testing technique employed to ensure the code did what it was supposed to do. Plotting various graphs for the same variable to check if the conclusions are the same for all the graphs. The memory was optimized by only selecting relevant columns for initial analysis and splitting the analysis of Ireland and the UK in the different notebooks.

1. **statistical analysis**

Test and optimize the statistical process involved executing a particular command aimed at doing a particular task, then check if it worked. For eg conducted test of normality using more than one test of normality i.e. Q-Q plot and Kolmogrov-Sminorv Test and checking if te conclusions are the same for both. Utilized simple random sampling to explore the dataset, ensuring representative insights without processing the entire dataset. Leveraged summary statistics (e.g., mean, median) to quickly grasp data distributions.

1. **Machine Learning**

Compared two encoding techniques to check if the results will be the same. Trade off made in machine learning included not performing ANN modelling due to longer processing time. ANN was also relevant for the analysis of this ML problem.

#### **Data manipulation**

1. **Data Preparation and Visualization:**

Data manipulation techniques included: -

* + - Loading data from various sources such as CSV and Json.
    - Inspecting the data to understand their structure and content.
    - Merging two data frames, filtering variables.
    - Cleaning and preprocessing techniques.

1. **statistical analysis**

When calculating the measures of central tendency, two libraries were used. The import statistics and pandas to calculate the means. The numpy library. Numpy results differed from those of statistics, especially for the variance because statistics uses the formula for sample where variance is calculated using n-1 while in numpy it uses n (sample).

Other data manipulation techniques were: -

* + - Aggregating data using grouby ( ) function

1. **Machine Learning**

Data manipulation techniques included: -

For the machine learning regression problems when using data encoded by one hot encoding the 7 regression models were built separately when data was normalized and when data was untransformed, but when dealing with standardized data all models were handled in a pipeline.

In sentiment analysis, two different libraries were used, VADER and TF-IDF vectorizer were used and the findings compared. TF-IDF was better.

## Reference

Abbas, M. *et al.* (2019) ‘Multinomial Naive Bayes Classification Model for Sentiment Analysis’, *IJCSNS International Journal of Computer Science and Network Security*, 19(3), p. 62. Available at: https://doi.org/10.13140/RG.2.2.30021.40169.

Afshar, M. and Usefi, H. (2022) ‘Optimizing feature selection methods by removing irrelevant features using sparse least squares’, *Expert Systems with Applications*, 200, p. 116928. Available at: https://doi.org/https://doi.org/10.1016/j.eswa.2022.116928.

Aguinis, H., Gottfredson, R.K. and Joo, H. (2013) ‘Best-Practice Recommendations for Defining, Identifying, and Handling Outliers’, *Organizational Research Methods*. SAGE Publications Inc., pp. 270–301. Available at: https://doi.org/10.1177/1094428112470848.

Allua, S. and Thompson, C.B. (2009) ‘Hypothesis Testing’, *Air Medical Journal*, 28(3). Available at: https://doi.org/10.1016/j.amj.2009.03.002.

Bae, J., Yu Hung, C. and van Lent, L. (2023) ‘Mobilizing Text As Data’, *European Accounting Review* [Preprint]. Routledge. Available at: https://doi.org/10.1080/09638180.2023.2218423.

Bhandari, A. (2023) *Feature Scaling for Machine Learning: Understanding the Difference Between Normalization vs. Standardization Introduction to Feature Scaling*.

*Chapter 7 Regression I: K-nearest neighbors | Data Science* (2023). Available at: https://datasciencebook.ca/regression1.html#multivariable-knn-regression (Accessed: 11 November 2023).

Cheng, A.M. (no date) *The Causes, Impact and Detection of Duplicate Observations*.

Chicco, D., Warrens, M.J. and Jurman, G. (2021) ‘The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation’, *PeerJ Computer Science*, 7, pp. 1–24. Available at: https://doi.org/10.7717/PEERJ-CS.623.

Cousineau, D. and Chartier, S. (2010a) ‘Outliers detection and treatment: a review.’, *International Journal of Psychological Research*, 3(1), pp. 58–67. Available at: https://doi.org/10.21500/20112084.844.

Cousineau, D. and Chartier, S. (2010b) ‘Outliers detection and treatment: a review.’, *International Journal of Psychological Research*, 3(1), pp. 58–67. Available at: https://doi.org/10.21500/20112084.844.

Dettori, J.R. and Norvell, D.C. (2018) ‘The Anatomy of Data’, *Global Spine Journal*. SAGE Publications Ltd, pp. 311–313. Available at: https://doi.org/10.1177/2192568217746998.

Dogra, V. *et al.* (2022) ‘A Complete Process of Text Classification System Using State-of-the-Art NLP Models’, *Computational Intelligence and Neuroscience*. Hindawi Limited. Available at: https://doi.org/10.1155/2022/1883698.

Eliane Birba, D. (2020) *A Comparative study of data splitting algorithms for machine learning model selection*, *DEGREE PROJECT IN COMPUTER SCIENCE AND ENGINEERING*.

Globus, A. (2014) *Principles of Information Display for Visualization Practitioners Principles of Information Display for Visualization Practitioners Principles of Information Display for Visualization Practitioners*. Available at: https://www.researchgate.net/publication/24285628.

Hou, Q. *et al.* (2021) ‘Understanding social media beyond text: A reliable practice on Twitter’, *Computational Social Networks*, 8(1). Available at: https://doi.org/10.1186/s40649-021-00088-x.

Janchevski, A. and Gievska, S. (2019) ‘A Study of Different Models for Subreddit Recommendation Based on User-Community Interaction’, in *Communications in Computer and Information Science*. Springer, pp. 96–108. Available at: https://doi.org/10.1007/978-3-030-33110-8\_9.

Jiang, T., Gradus, J.L. and Rosellini, A.J. (2020) ‘Supervised Machine Learning: A Brief Primer’, *Behavior Therapy*, 51(5), pp. 675–687. Available at: https://doi.org/10.1016/j.beth.2020.05.002.

Johnstone, I. and Titterington, D. (2009) ‘Statistical challenges of high-dimensional data’, *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 367, pp. 4237–4253. Available at: https://doi.org/10.1098/rsta.2009.0159.

de Juana-Espinosa, S. and Luján-Mora, S. (2019) ‘Open government data portals in the European Union: Considerations, development, and expectations’, *Technological Forecasting and Social Change*, 149. Available at: https://doi.org/10.1016/j.techfore.2019.119769.

Kang, H. (2013a) ‘The prevention and handling of the missing data’, *Korean Journal of Anesthesiology*, pp. 402–406. Available at: https://doi.org/10.4097/kjae.2013.64.5.402.

Kang, H. (2013b) ‘The prevention and handling of the missing data’, *Korean Journal of Anesthesiology*, pp. 402–406. Available at: https://doi.org/10.4097/kjae.2013.64.5.402.

Kluge, U. *et al.* (2017) *Factors influencing European passenger demand for air transport*. Available at: http://www.dataset2050.com.

KOCARIK GACAR, B. and DEVECİ KOCAKOÇ, İ. (2020) ‘Regresyon Analizleri mi Karar Ağaçları mı?’, *Celal Bayar Üniversitesi Sosyal Bilimler Dergisi*, pp. 251–260. Available at: https://doi.org/10.18026/cbayarsos.796172.

Liaqat, M.I. *et al.* (2022) ‘Sentiment analysis techniques, challenges, and opportunities: Urdu language-based analytical study’, *PeerJ Computer Science*, 8. Available at: https://doi.org/10.7717/PEERJ-CS.1032.

Manikandan, S. (2011) ‘Measures of central tendency: The mean’, *Journal of Pharmacology and Pharmacotherapeutics*, pp. 140–142. Available at: https://doi.org/10.4103/0976-500X.81920.

Martins, S., Pesado, P. and García-Martínez, R. (2016) ‘Information mining projects management process’, in *Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE*. Knowledge Systems Institute Graduate School, pp. 504–509. Available at: https://doi.org/10.18293/SEKE2016-009.

McCluskey, A. and Lalkhen, A.G. (2007) ‘Statistics II: Central tendency and spread of data’, *Continuing Education in Anaesthesia, Critical Care and Pain*, 7(4), pp. 127–130. Available at: https://doi.org/10.1093/bjaceaccp/mkm020.

Mishra, P. *et al.* (2019) ‘Descriptive statistics and normality tests for statistical data’, *Annals of Cardiac Anaesthesia*, 22(1), pp. 67–72. Available at: https://doi.org/10.4103/aca.ACA\_157\_18.

Nahm, F.S. (2016) ‘Nonparametric statistical tests for the continuous data: The basic concept and the practical use’, *Korean Journal of Anesthesiology*, 69(1), pp. 8–14. Available at: https://doi.org/10.4097/kjae.2016.69.1.8.

Narisetty, N.N. (2020) ‘Bayesian model selection for high-dimensional data’, *Handbook of Statistics*, 43, pp. 207–248. Available at: https://doi.org/10.1016/bs.host.2019.08.001.

*OReilly.Media.Machine.Learning.and.Data.Science.Blueprints.for.Finance.1492073059* (2020). Available at: https://studylib.net/doc/25722275/oreilly.media.machine.learning.and.data.science.blueprint... (Accessed: 7 November 2023).

Pano, T. and Kashef, R. (2020) ‘A complete vader-based sentiment analysis of bitcoin (BTC) tweets during the ERA of COVID-19’, *Big Data and Cognitive Computing*, 4(4), pp. 1–17. Available at: https://doi.org/10.3390/bdcc4040033.

Potdar, K., S., T. and D., C. (2017) ‘A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers’, *International Journal of Computer Applications*, 175(4), pp. 7–9. Available at: https://doi.org/10.5120/ijca2017915495.

Ridzuan, F. and Wan Zainon, W.M.N. (2019) ‘A review on data cleansing methods for big data’, in *Procedia Computer Science*. Elsevier B.V., pp. 731–738. Available at: https://doi.org/10.1016/j.procs.2019.11.177.

S, S.B. *et al.* (2020) *An Interpretation of Lemmatization and Stemming in Natural Language Processing*. Available at: https://www.researchgate.net/publication/348306833.

Saleh, H., Layous, J.A. and Republic, S.A. (2022) ‘Machine Learning-Regression’. Available at: https://doi.org/10.13140/RG.2.2.35768.67842.

Samuels, A. and Mcgonical, J. (2019) *Sentiment Analysis of News Articles: A Lexicon based Approach*. Available at: http://mlg.ucd.ie/datasets/bbc.html.

Das Sarit Chakraborty Student Member, B. and Member, I. (2018) *An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation*.

Sarker, I.H. (2021) ‘Machine Learning: Algorithms, Real-World Applications and Research Directions’, *SN Computer Science*. Springer. Available at: https://doi.org/10.1007/s42979-021-00592-x.

Schneider, A., Hommel, G. and Blettner, M. (2010) ‘Lineare regressionsanalyse - Teil 14 der serie zur bewertung wissenschaftlicher publikationen’, *Deutsches Arzteblatt*, pp. 776–782. Available at: https://doi.org/10.3238/arztebl.2010.0776.

*Secondary Analysis of Electronic Health Records* (2016) *Secondary Analysis of Electronic Health Records*. Springer International Publishing. Available at: https://doi.org/10.1007/978-3-319-43742-2.

Segal, M.R. (2003) *UCSF Recent Work Title Machine Learning Benchmarks and Random Forest Regression Publication Date Machine Learning Benchmarks and Random Forest Regression*.

Tajik, O. and Golzar, J. (2022) ‘Simple Random Sampling’. Available at: https://doi.org/10.22034/ijels.2022.162982.

Tolga, T. and GÖKMEN, N. (2021) ‘The Determination of the Factors Affecting Air Transportation Passenger Numbers’, *International Journal of Aviation, Aeronautics, and Aerospace*, 8(1), pp. 1–20. Available at: https://doi.org/10.15394/ijaaa.2021.1553.

Williams, B. *et al.* (2020) ‘Data-Driven Model Development for Cardiomyocyte Production Experimental Failure Prediction’, in S. Pierucci et al. (eds) *Computer Aided Chemical Engineering*. Elsevier, pp. 1639–1644. Available at: https://doi.org/https://doi.org/10.1016/B978-0-12-823377-1.50274-3.

Xiong, H. *et al.* (2022) ‘A Novel Approach to Air Passenger Index Prediction: Based on Mutual Information Principle and Support Vector Regression Blended Model’, *SAGE Open*, 12(1). Available at: https://doi.org/10.1177/21582440211071102.