**CCT College Dublin**

**Assessment Cover Page**

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

|  |  |
| --- | --- |
| **Module Title:** | MSc in Data Analytics |
| **Assessment Title:** | MSC\_DA\_CA2 |
| **Lecturer Name:** | David McQuinn  David Gonzalez  Marina Iantorno  Muhammad Iqbal |
| **Student Full Name:** | Ana Carla Ferreira de Oliveira |
| **Student Number:** | sba23438 |
| **Assessment Due Date:** | 07th January 2024 |
| **Date of Submission:** | 07th 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. |

**CA2 | MSC in Data Analytics**

***1º S E M E S T E R***

*Ireland’s Aviation Trends in Focus Alongside European Counterparts*

***https://github.com/Oliveiranac/CA2\_MScDataAnalytics.git***

**S t u d e n t:**

|  |  |
| --- | --- |
| **sba23438** | Ana Carla Ferreira de Oliveira |



**L e c t u r e r s:**

|  |  |
| --- | --- |
| **David McQuaid** | Data Preparation & Visualization |
| **David Gonzalez** | Python Programming |
| **Marina Iantorno** | Statistics |
| **Muhammed Iqbal** | Machine Learning |

**J A N U A R Y 2 0 2 4**

Contents

[1. INTRODUCTION 4](#_Toc155560096)

[1.1 OBJECTIVES 4](#_Toc155560097)

[2. BUSINESS UNDERSTANDING 4](#_Toc155560098)

[2.1 RESEARCH QUESTIONS 4](#_Toc155560099)

[3. DATA UNDERSTANDING 6](#_Toc155560100)

[3.1 DATA 1 6](#_Toc155560101)

[3.1.1 Data Source 6](#_Toc155560102)

[3.1.2 Data Description 7](#_Toc155560103)

[3.1.3 Data Usage 8](#_Toc155560104)

[3.2 DATA 2 8](#_Toc155560105)

[3.2.1 Data Source 8](#_Toc155560106)

[3.2.2 Data Description 8](#_Toc155560107)

[3.2.3 Data Usage 9](#_Toc155560108)

[4. DATA PREPARATION 9](#_Toc155560109)

[4.1 DATA IMPORT OF CSV AND SQLITE FILES 9](#_Toc155560110)

[4.1.1 Data preprocessing 10](#_Toc155560111)

[5. DATA VISUALIZATION AND STATISTICAL ANALYSIS 11](#_Toc155560112)

[5.1 DATA VISUALIZATION 11](#_Toc155560113)

[5.2 STATISTICAL ANALYSIS 15](#_Toc155560114)

[6. MODELLING AND SENTIMENTAL ANALYSIS 18](#_Toc155560115)

[6.1 Further Data Preprocessing 18](#_Toc155560116)

[7. RESULTS AND DISCUSSIONS 19](#_Toc155560117)

[8. CONCLUSION 20](#_Toc155560118)

[9. REFERENCES 22](#_Toc155560119)

List of Figures

[Figure 1: Ireland and European countries' air transport data set. 9](#_Toc155560122)

[Figure 2: Twitter transport dataset. 10](#_Toc155560123)

[Figure 3: preprocessed Ireland and European countries data. 10](#_Toc155560124)

[Figure 4: Tweeter preprocessed data. 11](#_Toc155560125)

[Figure 5: Boxplot of Statistics labels in Ireland. 11](#_Toc155560126)

[Figure 6: Line plot of Airport trends in Ireland. 12](#_Toc155560127)

[Figure 7: Bar plot of Airport performance in Ireland by Month. 13](#_Toc155560128)

[Figure 8: Bar plot of Airport performance per country. 13](#_Toc155560129)

[Figure 9: Bar plot of airport statistics in all countries. 14](#_Toc155560130)

[Figure 10: Line plot of yearly performance per country. 14](#_Toc155560131)

[Figure 11: Stacked bar plot of flight type comparison across countries. 15](#_Toc155560132)

[Figure 12: Line plot of mean squared error for default regression models. 19](#_Toc155560133)

[Figure 13: Line plot of R-squared for default regression models. 19](#_Toc155560134)

[Figure 14: Line plot of mean squared error for hypertuned regression models. 20](#_Toc155560135)

[Figure 15: Table of classification models results. 20](#_Toc155560136)

# **INTRODUCTION**

Air travel is an integral component of global connectivity, linking nations and fostering economic and cultural exchange. Understanding the patterns of passenger movement at airports is crucial for various stakeholders, from aviation authorities to travel agencies.

This project embarks on a comprehensive analysis of airport transport in Europe, focusing on Ireland and three comparable European countries with proportional population sizes. The objective is to unravel the trends and relationships between key indicators across these nations. The datasets encompass a diverse range, including passenger statistics, freight movements, commercial flight details, and even the sentiments expressed by passengers during their journeys. The temporal scope spans from 2020 to 2023, providing a robust foundation for longitudinal insights.

The research employs a multistage process, including data collection, pre-processing, visualization, statistical analysis, machine learning, and even sentimental analysis. Python programming language is applied in the analysis, facilitating a robust and scalable examination of the datasets.

The significance of this project extends beyond the world of academics, aiming to illuminate industry stakeholders, policymakers, and the general public with practical insights. As we explore the data, our goal is to find useful patterns that can help make smart decisions. This information can be valuable for those dealing with the ever-changing world of the aviation industry.

## OBJECTIVES

The main goal is to conduct a comprehensive analysis of Ireland's air transport, comparing it with other European countries for the period 2020 to 2023. The study will specifically delve into passengers, freight, and commercial flights, contributing to a clearer picture of air travel trends in the selected countries.

# **BUSINESS UNDERSTANDING**

In this project, several research questions are taken into consideration to help in gaining insights into the data.

## RESEARCH QUESTIONS

1. Are there differences in mean values for airports in Ireland?

* Hypothesis 1:
* Null hypothesis(H0): there is no significant difference in the mean values of VALUE across different Airports in Ireland.
* Alternative hypothesis(H1): there is a significant difference in the mean values of VALUE across different Airports in Ireland.

1. Is there a difference between the distribution of mean VALUE between Ireland and Denmark?

* Hypothesis 2:
* H0: There is no significant difference in the distributions of VALUE between Ireland and Denmark.
* H1: There is a significant difference in the distributions of VALUE between Ireland and Denmark.

1. Does the Year and VALUE variables have a correlation between each other?

* Hypothesis 3:
* H0: There is no correlation between VALUE and Year (correlation coefficient equals 0).
* H1: There is a significant correlation between VALUE and Year (correlation coefficient is not equal to 0).

1. Is there a difference in distribution across different months of the year?

* Hypothesis 4:
* H0: There is no significant difference in the distribution of VALUE across different months.
* H1: There is a significant difference in the distribution of VALUE across different months.

1. Are mean values of scheduled flights different from unscheduled flights?

* Hypothesis 5:
* H0: There is no significant difference in the mean values of VALUE between Scheduled and Unscheduled flights.
* H1: There is a significant difference in the mean values of VALUE between Scheduled and Unscheduled flights.

1. Is there a correlation between flight type and direction?

* Hypothesis 6:
* H0: There is no correlation between direction and flight type.
* H1: Flight\_Type and Direction have a significant correlation.

1. Does the distribution of VALUE for scheduled flights alter according to arrival and departure?

* Hypothesis 7:
* H0: For scheduled flights, there is no discernible change in the VALUE distribution between the points of arrival and departure.
* H1: For scheduled flights, there is no discernible difference in the distribution of VALUE between the points of arrival and departure.

# **DATA UNDERSTANDING**

This section offers a thorough look at the data used to analyze air travel in Ireland and other European nations. The study leverages insights from two datasets, each providing a unique perspective on the dynamics of aviation.

The combination of these datasets enables a holistic examination of the aviation landscape, blending quantitative metrics with qualitative insights. The ensuing exploratory analysis aims to uncover patterns, trends, and relationships within the data, laying the foundation for subsequent statistical and machine-learning analyses.

## DATA 1

TAM07.20231217131259.CSV: This dataset provides a thorough overview of Ireland's aviation trends, enabling in-depth examination and research of passenger, freight, and commercial flight patterns throughout time.

### 3.1.1 Data Source

Data.gov is the source of Ireland's passenger, freight, and commercial flight data. The Pandas package in Python can be used to access the dataset, which offers insightful statistics about air travel.

### 3.1.2 Data Description

With 29,160 entries and 9 columns, the collection contains data on several topics. The following are the columns and the corresponding data types for them:

* Statistical\_Label: This column represents the type of statistic being recorded. Divided into 'Freight,' 'Passengers,' and 'Commercial Flights.'
* Month: The data is recorded per month.
* Airports in Ireland: A variety of airports are located around Ireland, such as 'Cork, 'Dublin', 'Kerry', 'Knock', and 'Shannon.'
* Country: The nations taking part in the transit.
* Direction: 'All directions', 'Arrival', or 'Departure.'

**Note:**

|  |  |
| --- | --- |
| Arrivals: | This direction indicates statistics related to passengers arriving at the specified airports during the given period. |
| Departures: | Indicates statistics related to passengers departing from the specified airports during the given period. |
| All directions: | There could be a category for transits, connecting flights, or other specific aspects of air travel. |
| Understanding the "Direction" column is crucial for analyzing air travel patterns and understanding the dynamics of passenger movements at airports. | |

* Flight\_Type: Divided into three categories: scheduled, unscheduled, and all flights.

**Note:**

|  |  |
| --- | --- |
| Scheduled Flights: | These are regular flights that operate according to a published schedule. Airlines plan and advertise these flights well in advance. |
| Unscheduled Flights: | Also known as charter flights, these are not part of a regular schedule and are often booked for a specific purpose or by a particular group of travelers. |
| Understanding the Flight Type can provide insights into the nature of air travel activities. Scheduled flights are usually associated with routine passenger travel, while unscheduled flights might involve special events, group travel, or cargo transport based on specific needs. | |

* UNIT: This column represents the unit of measurement for the passenger count. In this dataset, it is labelled as “Thousand”, indicating that the passenger counts are in thousands.
* VALUE: The number corresponding to the particular category, measured in thousands.
* Year: The year the data was first recorded.
* Float values that indicate different metrics pertaining to passengers, freight, and commercial flights are found in the 'VALUE' column.

### 3.1.3 Data Usage

This dataset holds significance as it carries the potential to offer valuable information, benefiting a diverse range of stakeholders including the general public, lawmakers, and business experts.

## DATA 2

Database.SQLite: It utilizes the Twitter Airline Sentiment dataset to perform an in-depth analysis of sentiments directed towards diverse air companies. By employing sentiment analysis, our goal is to reveal patterns, sentiments, and trends that provide insights into the ever-evolving landscape of public opinion within the airline industry.

### 3.2.1 Data Source

The Twitter Airline Sentiment dataset, available on Kaggle, provides a diverse collection of tweets related to airline experiences, making it a valuable resource for conducting sentiment analysis research.

### 3.2.2 Data Description

With 14,485 entries and 15 columns, this dataset delves into various aspects of tweets associated with airline sentiments. Key columns include:

* tweet\_id: A distinct number assigned to every tweet.
* airline\_sentiment: Labels for sentiment that can be classified as "positive," "neutral," or "negative."
* airline\_sentiment\_confidence: The sentiment label's confidence score.
* negativereason: The cause of the unfavorable attitude.
* airline: The airline that the tweet is connected to.
* retweet\_count: The tweet's total number of retweets.
* text: The tweet's contents.
* tweet\_created: The creation time stamp of a tweet.
* user\_timezone: The user's time zone.

### 3.2.3 Data Usage

This dataset is a useful tool for sentiment analysis, enabling an investigation of public sentiment towards different airlines on Twitter. The sentiment labels and associated features provide comprehensive insights into user perspectives and opinions regarding their airline experiences.

# **DATA PREPARATION**

The data preparation process for this analysis adheres to Tuftes' principles of clarity, simplicity, and accuracy, ensuring that the information presented is both insightful and easily comprehensible.

## DATA IMPORT OF CSV AND SQLITE FILES

Using the pandas function pd.read\_csv(), the first data set, which is in comma-separated values (CSV), is loaded into pandas as a data frame. The data can be seen in brief as follows:

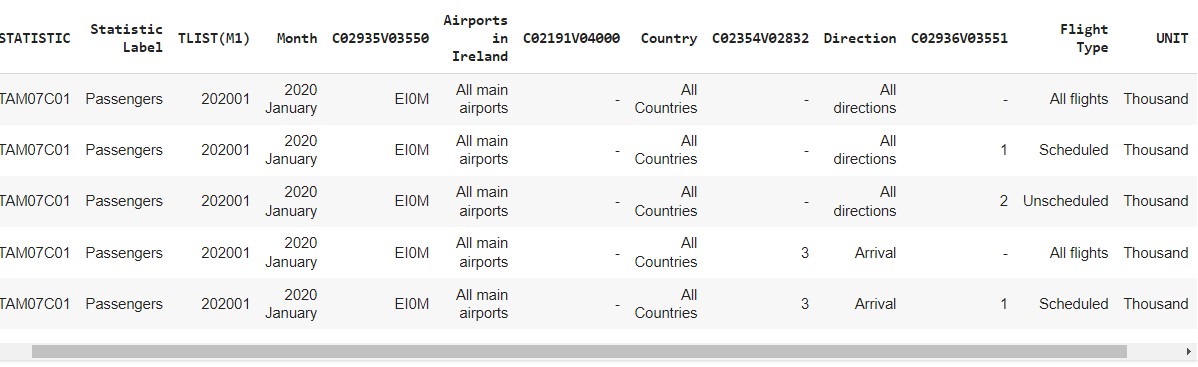


Figure 1: Ireland and European countries' air transport data set.

The second dataset, stored in a SQL database, is accessed using the Python pandas function pd.read\_sql\_query(). This function establishes a connection with the database and retrieves the data from the specified table. A brief overview of the loaded data is presented below:

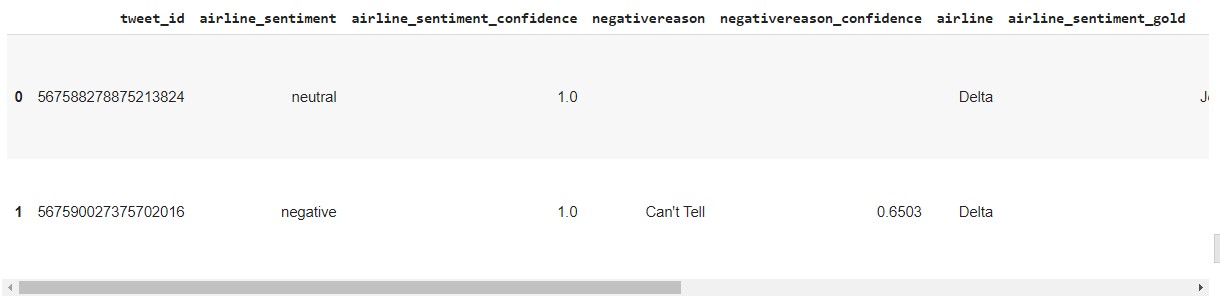


Figure 2: Twitter transport dataset.

Note: The figure is not showing all the columns in the data.

### 4.1.1 Data preprocessing

The first dataset is loaded using the pandas’ library and assigned the variable name 'df.' The data frame (df) includes several countries, but only those with populations proportional to Ireland are chosen—namely, Slovakia, Finland, and Denmark. Consequently, the data frame encompasses four countries. Subsequently, only pertinent columns essential for the analysis are retained, such as Statistic Label, Month, Airports in Ireland, Country, Direction, Flight type, Unit, and Value.

The next step involves renaming columns, where columns with spaces in their names are replaced with underscores for easier manipulation. A new 'Year' column is created from the 'Month' column by splitting it based on the space between its values. The 'Year' column is crucial for examining trends across individual years or over the entire period.

Finally, duplicates and nulls, which can lead to inconsistencies during analysis, are removed from the data. The final data looks as follows:



Figure 3: preprocessed Ireland and European countries data.

The second data frame is loaded using the pandas' library and named 'tweets\_df'. The data frame contains many columns that are not needed for the analysis, and thus, they are dropped. The resulting data frame is refined to include only two columns: 'text' and 'sentiment'. The 'text' column holds the information communicated by passengers during their travels. To ensure data consistency, any null values are removed. The final data frame is presented below:

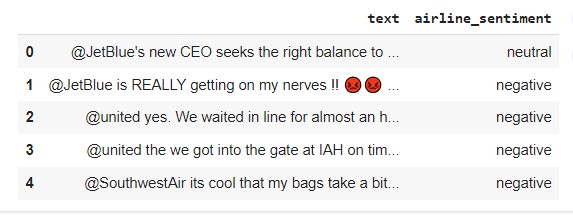


Figure 4: Tweeter preprocessed data.

# **DATA VISUALIZATION AND STATISTICAL ANALYSIS**

In this section, the objective is two-fold: first, to leverage the communicative power of visuals for a thorough understanding of datasets, and second, to employ robust statistical methods to tackle relevant research questions. By strategically combining charts, graphs, and interactive visualizations, the intricate narratives hidden within the data are brought to the forefront, making them accessible and interpretable to both researchers and a broader audience.

## DATA VISUALIZATION

The first visualization is designed to explore the distribution differences in Ireland with respect to statistic labels. Utilizing boxplots for each statistic label, this choice is deliberate as boxplots effectively present essential measures like the mean, interquartile range, quartiles, and outliers.

Uma imagem contendo Diagrama

Descrição gerada automaticamente

Figure 5: Boxplot of Statistics labels in Ireland.

Based on the plot, passengers exhibit the highest mean, interquartile range, and extreme outliers compared to other statistic labels. Freight follows as the second highest, while commercial flights show the lowest values in terms of mean, interquartile range, and outliers. This observation suggests that passenger travel dominates the air transport landscape in Ireland, with freight taking the second spot and commercial flights being the least prominent.

The second question aims to explore the trend of air transport across various airports in Ireland over the years. The trend is better shown by a time series plot, and using a line plot achieves this perfectly. The line plot distinctly illustrates the fluctuations in our data over time, providing a clear visualization of the evolving patterns in air transport across different Irish airports.

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Figure 6: Line plot of Airport trends in Ireland.

From the plot, almost all Airports are having a rising trend from 2020 to 2023.

Dublin consistently emerged as the leading airport in Ireland over the years, followed by Kerry. These two airports show a discernible trend in their performance. On the other end, Knock, Shannon, and Cork rank at the bottom. While Shannon displays a noticeable trend, the other two airports, Cork and Knock, show minimal or almost negligible trends.

The subsequent investigation focuses on understanding the performance of different months throughout the year in Ireland. This analysis aims to reveal whether specific months witness better performance than others. To effectively capture these variations, a bar plot is chosen for its ability to highlight differences between two or more data points. Since our interest lies in comparing each month against the others, a bar plot is ideal for visualizing this distribution.

The approach involves grouping the data from the selected countries by month and calculating the average value for each month. Subsequently, these average values for all months are arranged from greatest to smallest and presented in a bar plot, providing a clear visual representation of the distribution across different months.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Figure 7: Bar plot of Airport performance in Ireland by Month.

Observing the plot, a notable trend emerges where the 3rd quarter months contribute to higher performance, while the 1st quarter months exhibit lower performance. Specifically, August stands out as the peak month, slightly surpassing September, while January records the lowest performance, slightly trailing behind March and February.

The subsequent analysis delves into understanding the performance of different countries in comparison to each other. This examination aims to provide insights into how well Ireland's air transport fares when juxtaposed with other European countries. The approach involves calculating the mean for each country and arranging these mean values from the greatest to the smallest, facilitating a comparative assessment of air transport performance across different nations.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Figure 8: Bar plot of Airport performance per country.

Denmark is producing the best value performance when compared to the other countries but this value is slightly greater than Ireland's value. Slovakia and Finland are the lowest when compared to the top two countries.

To explore the comparative statistics across all countries more extensively, a detailed examination is carried out, and the findings are then compared with the results obtained from Ireland alone. While a boxplot could offer additional statistical insights, a bar plot is chosen for its simplicity in visualizing the mean value per statistic label for all countries in the dataset. This approach effectively addresses the question at hand.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Figure 9: Bar plot of airport statistics in all countries.

The passengers' statistic label is the highest among the four countries, mirroring the observation made in Ireland. The statistical patterns align closely with those observed in Ireland, with freight slightly surpassing commercial flights.

Furthermore, the trend over the years is examined for each country, and these trends are compared with one another. Each country's values for each year are visualized and compared to those of the other countries.

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Figure 10: Line plot of yearly performance per country.

Across all countries, there has been a consistent upward trend over the years. Ireland and Denmark have the highest values over the years, closely resembling each other. However, from 2022 to 2023, Denmark experiences a slightly greater increase than Ireland.

In the final analysis, flight types are compared across all countries to explore potential relationships. Given that the flight type column has more than two categories, a stacked bar chart is employed to provide a clear and straightforward visualization.

Gráfico, Gráfico de barras

Descrição gerada automaticamente

Figure 11: Stacked bar plot of flight type comparison across countries.

Across all countries, it is evident that the number of unscheduled flights is relatively low. Finland stands out with the highest number of unscheduled flights, followed by Ireland. In contrast, both Denmark and Slovakia report lower instances of unscheduled flights.

## STATISTICAL ANALYSIS

In this part, we explore statistical analysis to uncover key insights that go beyond raw data, revealing important patterns that help us understand how air transport works. Adhering to Tuftle's principles of data visualization, we employ techniques that prioritize simplicity and clarity, ensuring our analyses not only resonate with academic rigour but also cater to a broader audience.

Now, let's explore the statistical results:

1. Are there differences in mean values for airports in Ireland?

|  |  |
| --- | --- |
| **ANOVA results:** | |
| F-statistic = 147.60702169911534 | p-value = 1.0303840500953543e-149 |

The ANOVA results indicate a statistically significant difference in the mean values of 'VALUE' across different 'Airports\_in\_Ireland'.

The p-value is very close to zero which is less than our alpha level (0.05). This suggests that there is strong evidence to reject the null hypothesis, indicating that at least one of the means of 'VALUE' across the different airports is different.

1. Is there a difference between the distribution of mean VALUE between Ireland and Denmark?

|  |  |
| --- | --- |
| **Mann-Whitney U test results:** | |
| U-statistic = 32492343.0 | p-value = 2.405576424781503e-218 |

The test findings show that the distributions of 'VALUE' in 'Ireland' and 'Denmark' differ statistically significantly.

The p-value approaches zero quite closely. The null hypothesis, according to which there is no difference between the distributions of 'VALUE' in 'Ireland' and 'Denmark,' can be rejected.

Consequently, there's reason to believe that the distribution of 'VALUE' in these two nations differs significantly.

1. Do the Year and VALUE variables correlate with each other?

|  |  |
| --- | --- |
| **Correlation results:** | |
| Correlation coefficient = 0.11636968455525232 | p-value = 1.8843292503932015e-88 |

The correlation results indicate a statistically significant but relatively weak positive correlation between 'VALUE' and 'Year.'

The positive correlation coefficient suggests that as the 'Year' increases, there is a slight increase in the 'VALUE' variable.

However, the strength of the correlation is considered weak based on the magnitude of the correlation coefficient.

1. Is there a difference in distribution across different months of the year?

|  |  |
| --- | --- |
| **Kruskal-Wallis H-test results:** | |
| H-statistic = 57.291339564317646 | p-value = 2.940465305125592e-0 |

The findings show that there is a statistically significant variation in the 'VALUE' distributions among the various months.

The p-value approaches zero quite closely. The null hypothesis, according to which there is no variation in the 'VALUE' distributions throughout months, can be rejected. Consequently, there appears to be a notable variation in the 'VALUE' distribution among the months, based on our evidence.

1. Are mean values of scheduled flights different from unscheduled flights?

|  |  |
| --- | --- |
| **T-test results:** | |
| T-statistic = 24.187320137011557 | p-value = 2.2981116539205896e-127 |

The t-test results indicate a statistically significant difference between the mean values of 'VALUE' for 'Scheduled' and 'Unscheduled' flights.

The p-value is very close to zero, with such a low p-value, we can reject the null hypothesis that there is no difference in the mean values of 'VALUE' between 'Scheduled' and 'Unscheduled' flights.

Therefore, there is evidence to suggest that there is a significant difference in the average 'VALUE' for these two types of flights.

1. Do flight type and Direction have an association?

|  |  |
| --- | --- |
| **Chi-squared test results:** | |
| Chi2-statistic = 0.0 | p-value = 1.0 |

The results indicate that there is no significant association between 'Flight\_Type' and 'Direction'.

A p-value of 1.0 suggests that there is no evidence to reject the null hypothesis, which in this case would be the hypothesis that there is no association between 'Flight\_Type' and 'Direction'.

1. Is there a difference in the distribution of VALUE between Arrival and Departure for Scheduled Flights?

|  |  |
| --- | --- |
| **Wilcoxon Signed-Rank Test results:** | |
| W-statistic = 32215.5 | p-value = 6.810558894481764e-09 |

The test results indicate a statistically significant difference in the 'VALUE' between 'Arrival' and 'Departure' for 'Scheduled' flights.

The p-value is very close to zero. We can reject the null hypothesis that there is no difference in the 'VALUE' between 'Arrival' and 'Departure' for 'Scheduled' flights.

Therefore, we have evidence to suggest a significant difference in the 'VALUE' between these two directions for 'Scheduled' flights.

# **MODELLING AND SENTIMENTAL ANALYSIS**

In the domain of machine learning, predicting the continuous response variable, 'VALUE', involves deploying regression models due to its continuous nature. This section employs four distinct models – Random Forest, Decision Trees, Gradient Boosting, and K-Nearest Neighbors – to forecast the 'VALUE' variable.

Simultaneously, in the field of sentiment analysis, the prediction of sentiments (neutral, positive, or negative) is conducted using the text column. Four classification models are employed for this purpose: Multinomial Naive Bayes, Support Vector Machines, Random Forest, and Logistic Regression. This diverse set of models ensures a comprehensive exploration of predictive capabilities, contributing valuable insights to both regression and sentiment analysis tasks.

## Further Data Preprocessing

Before training the models to predict the response variables, several crucial data preprocessing steps are implemented to ensure the data is well-prepared for machine learning. Initially, when dealing with a dataset rich in categorical variables, a necessary transformation is applied as it can’t be used in machine learning models. The categorical columns are transformed into dummy variables using the one-hot encoding function in Python.

Following that, the process involves separating the response variable, 'VALUE', from the independent variables into distinct datasets. The data is then partitioned into an 80:20 ratio, allocating 80% for model training and 20% for testing. Default models are initially executed, and their results are systematically compared. After this, a hyperparameter tuning process is conducted using grid search to optimize the model parameters. The refined models are once again compared to identify the most effective one.

In the field of sentiment analysis, additional data preparation steps are undertaken. The initial step involves tokenization, breaking down the text column into a sequence of words. Following this, a predefined Python library is utilized to remove stop words, which are the most frequently occurring words in the data. The data is then split into an 80:20 ratio for training and testing. To facilitate machine learning algorithms, a bag of words is created, simplifying the conversion of textual data into a suitable format.

# **RESULTS AND DISCUSSIONS**

The mean squared error plot shows that both Random Forest and Decision Trees are leading to lower errors. The Gradient Boosting model has the highest mean squared error, and this indicates it’s performing poorly than any other model. Random Forest is achieving the highest R-squared, indicating exceptional performance compared to other models.

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Figure 12: Line plot of mean squared error for default regression models.

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Figure 13: Line plot of R-squared for default regression models.

When hyper-tuning the models using Grid Search in Python, it is observed that mean squared errors tend to increase for all models. The KNN model leads to a lower MSE as compared to other models. The Gradient-Boosting model improves when compared to Random Forest and Decision Trees. However, it can be noted that the hyper-tuned models of Random Forest and Decision Trees do not perform better than the default one.

Gráfico, Gráfico de linhas

Descrição gerada automaticamente

Figure 14: Line plot of mean squared error for hypertuned regression models.

Analyzing the results of sentimental analysis, Logistic Regression emerges as the best model, slightly surpassing Support Vector Machines in accuracy. Overall, all models perform well, except for Multinomial Naïve Bayes, which achieves an accuracy below 70%.

Tabela

Descrição gerada automaticamente com confiança baixa

Figure 15: Table of classification models results.

# **CONCLUSION**

In conclusion, this study unfolds crucial insights into the dynamics of air travel in Ireland, presenting valuable comparisons with selected European counterparts. The prominence of passenger travel, notably in Dublin, underscores the industry landscape. Monthly analysis reveals peak performance in the third quarter, especially in August. Ireland's competitive standing is evident in the comparative study, especially against Denmark.

In light of the extensive statistical analyses conducted, several crucial insights have emerged. The ANOVA results reveal a significant disparity in the mean values of 'VALUE' across different airports in Ireland, underscoring variations in air transport dynamics. Furthermore, the Mann-Whitney U test demonstrates a substantial distinction in the distribution of mean 'VALUE' between Ireland and Denmark, emphasizing the unique characteristics of these two nations in terms of air travel metrics.

The correlation analysis indicates a statistically significant but weak positive correlation between 'VALUE' and 'Year.' While the correlation is weak, this finding suggests a gradual increase in 'VALUE' over the years, prompting further exploration into the evolving trends within the specified timeframe.

Month-wise variations, as revealed by the Kruskal-Wallis H-test, showcase a notable difference in the distribution of 'VALUE' across different months. This insight into monthly variations contributes to a more nuanced understanding of air transport patterns, highlighting specific periods of heightened or subdued performance.

The T-test results affirm a significant difference in the mean values of 'VALUE' between scheduled and unscheduled flights, offering valuable insights for operational planning and resource allocation within the aviation sector.

The chi-squared test results indicate no significant association between 'Flight\_Type' and 'Direction.' This suggests that the choice of flight type does not significantly influence the direction of travel, providing clarity on the independent factors affecting these variables.

Lastly, the Wilcoxon Signed-Rank Test results underscore a substantial difference in the 'VALUE' between 'Arrival' and 'Departure' for 'Scheduled' flights. This finding sheds light on directional disparities, contributing to a more nuanced understanding of air travel dynamics within the scheduled flight category.

In essence, these statistical analyses contribute valuable nuances to our understanding of air transport in Ireland and provide a robust foundation for informed decision-making within the aviation industry.

Moreover, the study explores machine learning techniques, with Random Forest and Decision Trees proving effective in predicting air transport factors. On the sentiment analysis front, Logistic Regression emerges as the optimal choice from the Twitter dataset. These revelations carry practical implications for public awareness campaigns, policy enhancements, and industry adaptability.

While acknowledging these contributions, it's crucial to consider limitations such as the constrained timeline from 2020 to 2023. Future investigations could expand the dataset, incorporate real-time components, and explore additional dimensions. Ultimately, this study serves as a valuable tool, equipping stakeholders with practical knowledge to navigate the ever-evolving landscape of air travel.

# **REFERENCES**

1. Sojasingarayar, A. (2022) 'Data Preprocessing in Python: Data cleaning and transformation for Machine Learning.' Medium, Available at: https://medium.com/@abonia/data-preprocessing-in-python-1f90d95d44f4 (Accessed: 28 December 2023).
2. Chakure, A. (2022) 'How to Preprocess Data in Python: A step-by-step tutorial on data preprocessing implementation using Python, NumPy, and Pandas.' Built-In, Published on Jun. 10, 2022. Available at: https://builtin.com/machine-learning/how-to-preprocess-data-python(Accessed: 28 December 2023).
3. Ravikiran, A. S. (2023) 'A Complete Guide to Data Visualization in Python With Libraries & More.' Simplilearn, Last updated on May 18, 2023. Available at: https://www.simplilearn.com/tutorials/python-tutorial/data-visualization-in-python#:~:text=Matplotlib%20and%20Seaborn%20are%20python,primarily%20used%20for%20statistical%20graphs. (Accessed: 28 December 2023).
4. GeeksforGeeks 'Data Visualization with Python.' Available at: https://www.geeksforgeeks.org/data-visualization-with-python/
5. Sharma, G. (2020) 'Statistical Analysis using Python.' Analytics Vidhya, Published on Sep 16, 2020. Available at: https://medium.com/analytics-vidhya/statistical-analysis-using-python-e83f10ca3c82
6. GeeksforGeeks (2023) 'Machine Learning with Python Tutorial.' Updated on 7 Nov 2023. Available at: https://www.geeksforgeeks.org/machine-learning-with-python/
7. Arora, S. (2023) 'Sentiment Analysis Using Python.' Analytics Vidhya, Updated on July 24th, 2023. Available at: https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python/