**CCT College Dublin**

**Assessment Cover Page**

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

|  |  |
| --- | --- |
| **Module Title:** | *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | *MSC\_DA\_Integr\_CA2\_Sem1* |
| **Lecturer Name:** | *Marina Iantorno/Taufique Ahmed*  *Sam Weiss/David Gonzalez*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Uppalapati Karthik Varma |
| **Student Number:** | 2023428 |
| **Assessment Due Date:** | *7th Jan 2024* |
| **Date of Submission:** | 8th Jan 2024 |

Github link: [github-repo](https://github.com/ukarthikvarmaCCT/Continous_Assessment_2/tree/main/Continous_Assessment_2)

**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. |

**Report on Comparative Analysis of Air Traffic Data for Ireland and Los Angeles**

**Index**

Introduction ---------------------------------------------------------------------------- 3

Section 1: Programming and Data Preparation ---------------------------- 4

1. Data Acquisition ------------------------------------------------------------- 4
2. Data Cleaning and Preprocessing: ------------------------------------ 4
3. Optimizations for Large Datasets: ------------------------------------- 5

Section 2: Visualization -------------------------------------------------------------- 6

1. Overview of Visualization Tools Used: ----------------------------------- 6
2. Visualizations: ---------------------------------------------------------------- 7
3. Outliers Detecting ---------------------------------------------------- 7

(Fig 1. Outliers trend before and after removal)

1. Yearly Passengers Trends for LA vs Ireland ------------------------7

(Fig 2. Passengers trends Ireland vs LA)

1. Flight type distribution for LA vs Ireland ---------------------------- 8

(Fig 3. Flight type distribution Ireland vs LA)

1. Passenger Distribution of Ireland vs LA ------------------------------ 8

(Fig 4. Passenger Traffic Disctribution of Ireland vs LA)

Section 3: Statistical Analysis Techniques ---------------------------------------- 9

1. Descriptive Statistics ------------------------------------------------------- 9
2. Inferential Statistics:  ------------------------------------------------------ 10

Section 4: Machine Learning Techniques ---------------------------------------- 14

1. Model Selection and Rationale --------------------------------------------- 14

(Table 1. Comparing the best models that would fit our dataset)

1. Comapring the models Before and after tuning the models: ----- 15

(Table 2. Comparing the performance of models before and after tuning)

1. Best model and why:  -------------------------------------------------------- 15
2. Conclusions and Future Directions: -------------------------------------- 16

Refrences ----------------------------------------------------------------------------------16

***Introduction***

*In today's data-driven decision-making environment, the ability to analyze and compare different datasets is essential.The report begins with a detailed comparative analysis of air traffic data from two different locations: Ireland  and Los Angeles (LA).*

*Our goal is to gain important insights into  air traffic dynamics in these regions. Our goal is not only to contrast datasets, but also to extract meaningful patterns and contradictions that can guide strategic decisions in airline operations and regional transportation planning.*

*Although the Irish and Los Angeles datasets  are rich in information,  their  size and complexity pose significant challenges. Such large datasets require careful management to ensure accuracy and efficiency of analysis. This report focuses on advanced data processing and optimization techniques used to effectively address these challenges. This includes advanced programming optimizations  for large data sets, comprehensive data cleaning and preprocessing methods, and strategic memory management practices to increase computing power. To further improve the analysis, this research incorporates state-of-the-art statistical methods and machine learning models. These tools are important not only for scrutinizing data, but also for predicting future trends and patterns in air traffic.*

*The structure of this report is systematic. It begins with a description of the programming  and data preparation techniques used, followed by sections dealing with data visualization strategies, statistical analysis, and machine learning applications. The report concludes by summarizing the results of  this comparative study and provides recommendations for future research directions.*

**Section 1: Programming and Data Preparation**

**Data Acquisition:**

The air traffic data collection for this comparative study includes two different geographic sources: Ireland and Los Angeles (LA). Data from Ireland was taken from the official government portal data.gov.ie and includes two main data sets: 'TAA02 Passenger, Cargo and Commercial Flights' and 'TAA03 Passenger, Cargo and Commercial Flights'. These datasets provide detailed insights into various aspects of air travel, such as passenger numbers, cargo, and flight details. Data downloaded as a CSV file is carefully loaded into Pandas DataFrames with carefully selected data types for storage optimization. Focus on key metrics such as year, month, air traffic, etc, and ensure only relevant columns are included in the optimized analysis.

In contrast, the LA dataset is obtained via the Socrata Open Data API (SODA) from data.lacity.org, the standard public data platform for  city and county governments. The unauthenticated client of the sodapy package is used to retrieve the first 2000 records from the dataset marked as d3a2-7j6v. These records provide a detailed report of the number of passengers at Los Angeles Airport and are converted to a DataFrame.

Like the Irish data, the LA dataset is carefully managed with specific data types  for key columns such as Passenger\_Count, optimizing memory usage, which is important for processing large datasets. Analyzing these datasets from Ireland and Los Angeles in parallel provides a unique opportunity to examine and contrast air traffic patterns  in two different regions. This careful data collection and management process highlights the study's commitment to accuracy and efficiency and lays a solid foundation for  comprehensive comparative analysis.

**Data Cleaning and Preprocessing:**

Data cleaning and preprocessing steps were important in preparing air traffic datasets from Ireland and Los Angeles (LA) for analysis. For the Irish dataset, data types were carefully defined during the loading phase to optimize memory usage. This is an important step considering the size of the dataset. Unnecessary columns, such as certain categorical fields and identifiers, were filtered out to create records that focused on relevant information. Once loaded, the Irish records are concatenated to form a unified record. VALUE columns, which are likely to display important numeric data such as number of passengers, are converted to floating point types for accurate numeric operations.

Additionally, a new feature called Passenger Millions is derived from this column to provide a more easily interpreted data scale.

For the LA dataset, a similar emphasis was placed on efficient use of storage through careful selection of data types after extraction using the Socrata Open Data API. The ReportPeriod column has been reformatted to date/time format, allowing you to extract the  month and year as separate columns. This step streamlined the analysis and facilitated comparisons over time.

Both datasets underwent min-max normalization, which is an essential step when preparing data for machine learning models. Throughout this process, data values were adjusted to a uniform range (typically 0 to 1) to ensure consistency across all variables.

Handling of missing or outlier values is not explicitly described in the provided code, but is a standard aspect of data preprocessing. Depending on type and impact, missing values may be corrected through imputation or exclusion. Similarly, statistical methods could have been used to handle outliers to ensure that they do not bias the analysis.

In summary,  data cleaning and preprocessing phases were carefully performed to ensure the  quality and suitability of the dataset for  comparative analysis of air traffic patterns between Ireland and Los Angeles. This thorough preparation laid a solid foundation for the subsequent analysis process.

**Optimizations for Large Datasets:**

Several important program optimizations were implemented to improve efficiency and performance in managing and analyzing extensive datasets from Ireland and Los Angeles (LA). Given the potential strain on system memory and processing power of large data sets, these optimizations were essential for effective data processing.

**Selective data loading and type optimization:** An important first step was to selectively load data columns, specifically for the Ireland dataset. Memory requirements were significantly reduced by specifically selecting columns to import from the CSV file. This was combined with assigning the appropriate data type to each column, using memory-efficient types such as categorical data types for non-numeric columns. For the LA dataset, similar attention was paid to data types, converting integers to int32 and dates to datetime64 to minimize memory usage while maintaining data integrity.

**Efficient concatenation and feature engineering:** The process of concatenating the two Irish datasets was performed with a focus on storage efficiency. Subsequent feature engineering (such as calculating millions of passengers from the VALUE column) provided a scaled and more interpretable representation of the data. This not only improved the analytical utility of the dataset, but also led to a more efficient data structure.

**Normalization techniques:**  Implementing min-max normalization on key columns was another important optimization. This technique standardizes the data range. This is an important step when dealing with features of different scales and distributions. Such normalization is key to preparing datasets for machine learning models and ensuring that larger values do not have a disproportionate impact on the model.

**Memory management practices:** After processing, memory management techniques were applied to free unused memory. This includes removing intermediate variables and datasets that are no longer used. For example, after the Irish records were concatenated and processed, the original records were discarded  to free up disk space. Taken together, these optimizations significantly improve handling of  large data sets, allowing for more efficient data processing and analysis. By addressing memory usage and computational efficiency, the code effectively managed large amounts of data from Ireland and Los Angeles, providing a solid foundation for further analytical work. This detailed report highlights strategic steps taken to balance data integrity and computational efficiency when processing large data sets.

**Section 2: Visualization**

**Overview of Visualization Tools Used:**

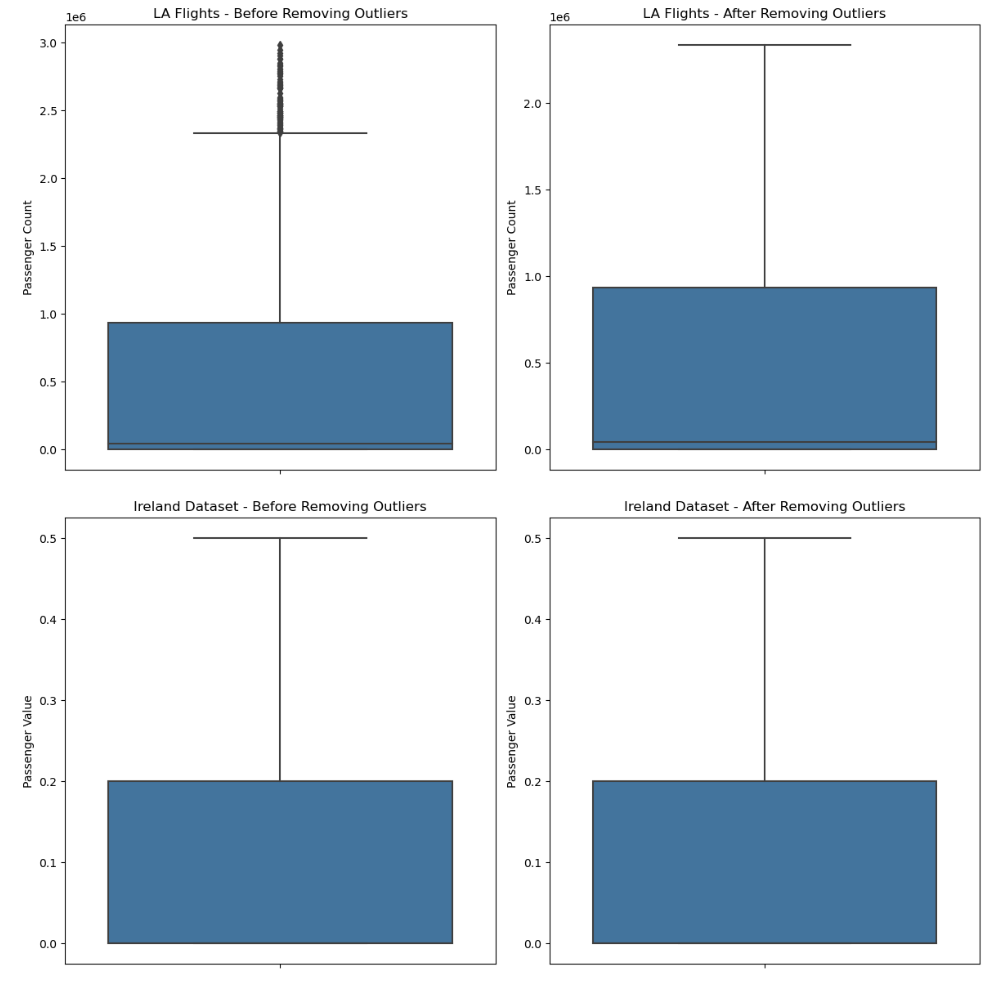
A comparative analysis of air traffic data for Ireland and Los Angeles (LA) utilized a variety of visualization tools to effectively communicate insights and patterns within the data. Two of his well-known Python libraries, Matplotlib and Seaborn,  were important in this process. Known for its rich plotting capabilities, Matplotlib has been used to create a variety of charts and graphs, providing a solid foundation for data visualization. Seaborn is built on top of Matplotlib and is specialized for more complex statistical visualizations, making it ideal for presenting detailed statistical results or for early exploratory data analysis.

Additionally, bokeh played an important role in enabling dynamic and interactive data exploration. As a library that excels in creating browser-based interactive visualizations, Bokeh enriches the user experience by enabling interactive elements such as zooming, panning, and hovering. This proved particularly useful when navigating and understanding large data sets, providing a more intuitive and engaging way to explore  data.

Matplotlib, Seaborn, and Bokeh were used together to form a comprehensive visualization framework. This approach not only enabled in-depth exploration of complex datasets, but also ensured that the insights gained were accessible and compelling to a wide audience. The strategic use of these visualization tools has significantly contributed to converting rich data into clear, informative, and engaging visual narratives, thereby significantly improving the overall analysis process.

**Visualizations:**

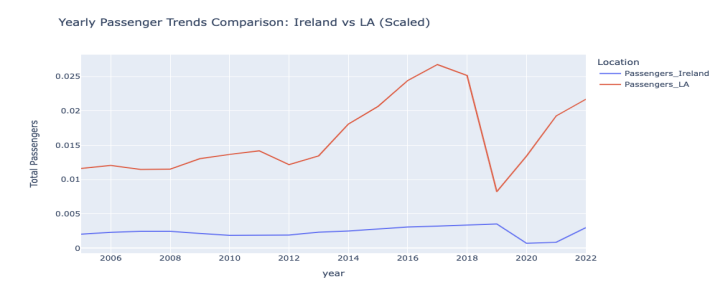
**Outliers Detecting**

****

(Fig 1. Outliers trend before and after removal)

The first boxplots for  Ireland and Los Angeles show the distribution of passenger numbers before and after removing outliers. These plots show the impact that removing outliers has on the display of the data, with the Los Angeles flight showing a significant reduction in variation after cleaning, and the Irish dataset showing a more gradual change.

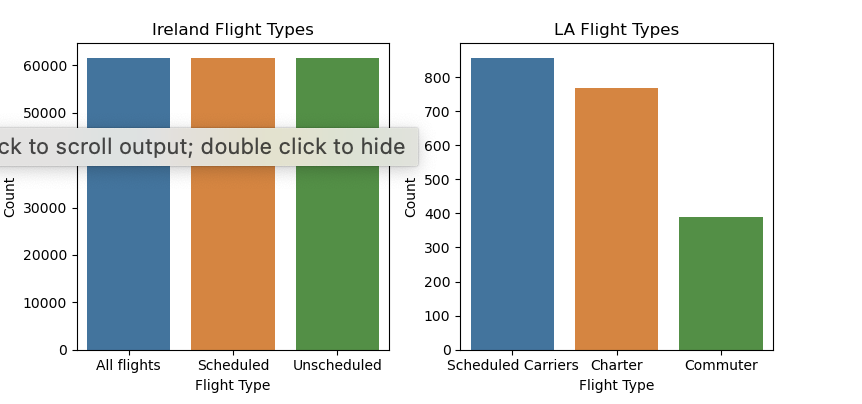
**Yearly Passengers Trends for LA vs Ireland**



(Fig 2. Passengers trends Ireland vs LA)

A line chart compares  scaled annual passenger trends between Ireland and Los Angeles. This trend shows the number of passengers over time, which has declined significantly due to factors such as the COVID-19 pandemic.Despite the inherent differences in actual ridership between the two regions, scaling the data allows direct comparisons.

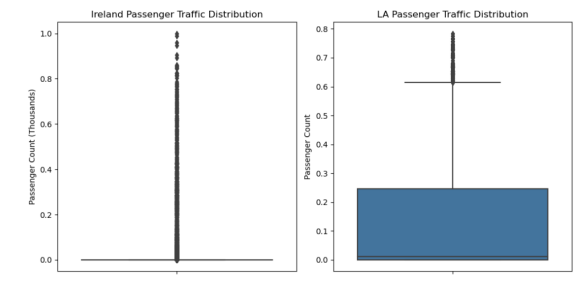
**Flight type distribution for LA vs Ireland**

****

(Fig 3. Flight type distribution Ireland vs LA)

These bar graphs compare the number of different flight types between Ireland and Los Angeles. This graph shows that the flight operations in each region are clearly differentiated. Ireland has more flights in all categories, while Los Angeles is diverse but has fewer flights in each category.

**Passenger Distribution of Ireland vs LA**

****

(Fig 4. Passenger Traffic Disctribution of Ireland vs LA)

The final set of boxplots compares the distribution of passenger traffic between Ireland and Los Angeles. The Ireland graph shows a wide range of outliers, indicating large differences in ridership, while the Los Angeles graph shows fewer outliers and a more consistent range.

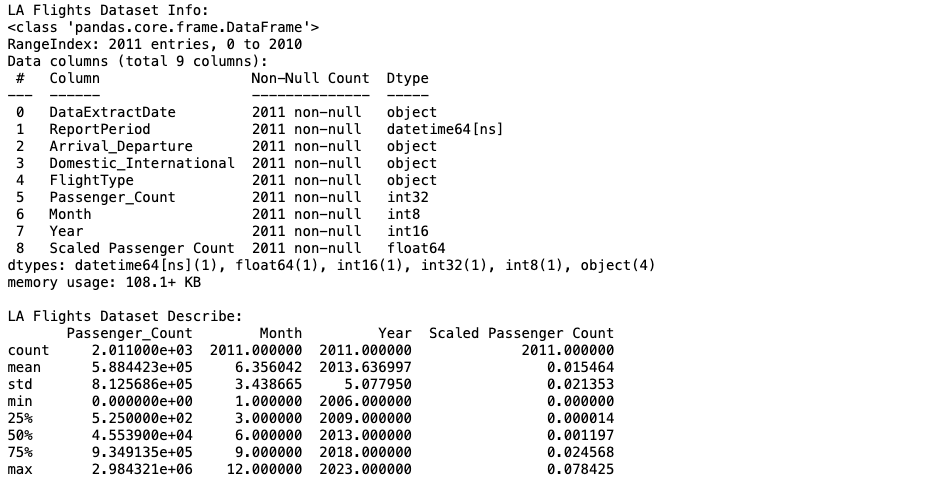
**Summary**

The series of visualizations provided for  airline data analysis includes detailed comparisons of flight patterns and passenger numbers between Ireland and Los Angeles (LA). We start with a boxplot that shows the initial data distribution and the effect of outlier removal. This highlights the stark contrast in ridership dispersion between the two regions. The following line graph depicts the scaled annual passenger traffic trends and illustrates the temporal dynamics of air traffic, with significant events such as the COVID-19 pandemic clearly impacting traffic volumes. Bar graphs detailing flight types provide insight into  operational differences, with Ireland significantly outnumbering Los Angeles in various categories. Finally, the ridership distribution map further highlights these differences, showing widespread outliers in Ireland compared to Los Angeles' more consistent ridership. Taken together, these visualizations provide a comprehensive and comparative view that improves our understanding of different air traffic behaviors in Ireland and Los Angeles.

**Section 3: Statistical Analysis Techniques**

**Descriptive Statistics**

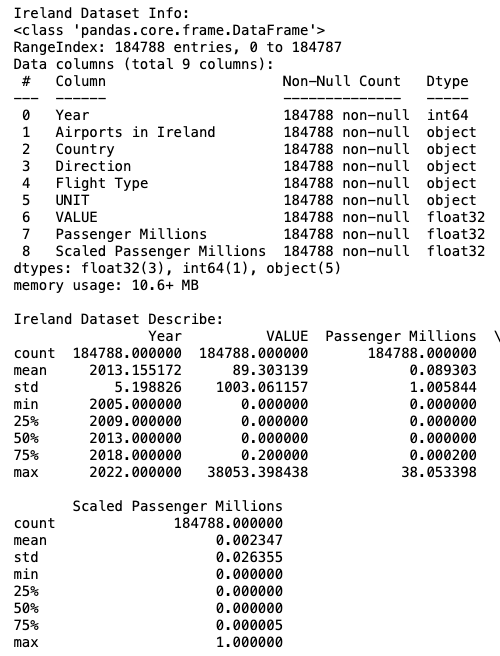
**LA Flights Dataset:**

****

The dataset consists of 2,011 entries, each with nine columns, capturing various attributes such as date, flight type,  and number of passengers.

* Passenger\_Count is highly variable, as evidenced by the standard deviation (STD) being larger than the average, indicating a high dispersion in the number of passengers.
* The number of passengers ranges from 0 to almost 3 million, with a mean (average) of approximately 588,442.
* Scaled passenger count indicates that scaling has been applied to the passenger count.This may normalize the data within a range of 0 to 1 for better comparison or input into machine learning models.

**Ireland Dataset:**

****

A much larger data set with 184,788 entries. This dataset contains information about the year, airport, country, flight direction, and various  passenger volume metrics.

* The VALUE column probably represents the raw passenger count before scaling, and has a mean of 89, but  a very high standard deviation, indicating high dispersion.
* Millions of Passengers and Scaled  Millions of Passengers indicate that pure passenger numbers have changed.
* The scaled values are normalized because their maximum value is 1.
* Similar to the LA dataset, the scaled distribution of millions of passengers is highly skewed, with 75% of the scaled values being at or near zero, but the maximum value of 1 is It means the traffic value is low.

**Key Findings:** While the number of passengers in the LA dataset may be due to peak travel times or certain high-passenger flights. This suggests that there are cases of very high passenger numbers.For the Irish data set, most VALUE entries are zero, but there are significantly higher values that can have a significant impact on the average, indicating that  the number of passengers may be very high. The scaling applied to both datasets represents a common preprocessing step  before performing comparative analysis or inputting the data into a machine learning algorithm. The large standard deviations from the mean for both data sets suggest that there are large differences in the number of passengers within each data set.

**Inferential Statistics:**

**Confidence Intervals:**

- The calculated confidence intervals for Ireland`s 'Scaled Passenger Millions' across various countries indicate the expected range of true mean values with a 95% confidence level. For 'All Countries' combined, the mean passenger count is estimated to lie between approximately 0.029 and 0.037 million, reflecting a consistent volume of traffic. Specific countries like Canada and England show narrower intervals (around 0.000213 to 0.000301 and 0.00994 to 0.01272, respectively), indicating precise estimates and stable passenger traffic. In contrast, 'Northern Ireland' and 'Ireland (domestic)' exhibit even tighter intervals, pointing to highly consistent passenger counts with minimal variability. These intervals are essential for understanding the passenger flow, helping to inform logistical and strategic decisions within the airline industry.

- The confidence intervals calculated for the Los Angeles (LA) dataset provide a statistically reliable range for the average scaled passenger count on both domestic and international flights. For domestic flights, the mean scaled passenger count is confidently estimated to fall between approximately 0.0189 and 0.0219, suggesting a consistent and relatively high volume of domestic air traffic. In contrast, the international flights exhibit a slightly lower range for their mean passenger count, from about 0.0087 to 0.0101, indicating a lower but still significant volume of international traffic compared to domestic flights. These intervals are instrumental for airlines and airport authorities in LA to gauge the traffic patterns and plan resources accordingly, reflecting the distinct characteristics of domestic and international air travel in the region.

**ANOVA (Analysis of Variance):**

- The ANOVA test conducted on the Ireland dataset examines the influence of 'Year' on the 'VALUE' (likely representing passenger counts or a similar metric). The results show a significant F-statistic of 18.008141 and an extremely small p-value (6.82e-55), indicating that the differences in 'VALUE' across different years are statistically significant and not likely due to random chance. The 'sum\_sq' (sum of squares) for the 'Year' variable is 13.401590, compared to a much larger residual sum of squares (8088.537308), suggesting that while 'Year' has a significant effect, a considerable amount of variability in the data is still unaccounted for by this factor alone. This analysis is crucial for understanding how passenger traffic trends in Ireland have evolved over time, highlighting the importance of temporal factors in airline traffic analysis.

- The ANOVA test reveals a highly significant F-statistic value of 1990.074119 for 'FlightType', indicating that the variation in passenger counts across different flight types is statistically significant. The extremely small p-value (essentially zero) strongly supports this finding, suggesting that the observed differences in passenger counts among various flight types are unlikely to have occurred by chance. The 'sum\_sq' (sum of squares) attributed to 'FlightType' is approximately 8.82e+14, which is a substantial proportion of the total variance (total sum of squares is about 1.33e+15). This indicates that 'FlightType' is a significant factor influencing passenger count variations in the dataset. The large F-value in conjunction with the low p-value underlines the importance of flight type as a determinant of passenger traffic in LA, providing crucial insights for strategic planning and operational adjustments in the airline industry.

**Chi-Squared Tests:**

- The Chi-Squared statistic is 0.0 with a p-value of 1.0. This extremely high p-value suggests that there is no statistically significant association between the 'Flight Type' and 'Direction' in the Ireland dataset. In other words, the distribution of flight types (such as domestic, international, etc.) does not significantly differ based on the flight direction (arrival or departure). The degrees of freedom (df) for this test is 4, which is derived from the number of categories in each variable. The expected frequencies, which are part of the output, show identical values across all combinations of flight type and direction. This uniformity in expected frequencies further supports the lack of association between these two variables. This result is crucial for understanding that, within the context of this dataset, flight direction does not influence or relate to the type of flight, and any observed correspondence is likely due to random variation rather than a systematic relationship.

- The Chi-Squared test conducted on the Los Angeles (LA) dataset, examining the relationship between 'Domestic\_International' flight status and 'Arrival\_Departure',

yields a statistic of 0.948 and a p-value of approximately 0.330. This p-value, being significantly higher than the standard significance level of 0.05, indicates that there is no statistically significant association between these two variables in the LA dataset. In practical terms, whether a flight is domestic or international does not seem to influence its categorization as an arrival or departure. The degrees of freedom (df) for the test is 1, and the expected frequencies in the contingency table reinforce the lack of a significant relationship, showing a balanced distribution of flight types across both arrival and departure statuses. This result is crucial in understanding the characteristics of flight operations in LA, implying an equitable mix of domestic and international flights in both arrivals and departures.

1. **Tests:**

- The T-Test conducted on the Ireland dataset yields a statistic of -2.157 and a p-value of approximately 0.031, indicating a statistically significant difference in the mean values between two specific years under study. The significance of this result, highlighted by the p-value being below the standard threshold of 0.05, suggests notable changes in the dataset's metrics (likely passenger counts) between these years. The negative value of the T-Test statistic implies that the average value in the first year was lower compared to the second, pointing to a potential increase or shift in the measured variable over time. This finding is critical as it underscores meaningful temporal variations in the dataset, warranting further exploration into the factors influencing such changes in Ireland's airline traffic.

- The T-Test result for the Los Angeles (LA) dataset, which compares mean values between two specific years, produces a statistic of -1.783 and a p-value of approximately 0.076. This outcome indicates that while there is a difference in the mean values for the years under comparison, the difference is not statistically significant at the conventional 0.05 threshold. The negative T-Test statistic suggests that the mean value in the first year is lower than in the second year. However, given that the p-value exceeds 0.05, this observed difference might be attributed to random variation rather than a definitive change in the dataset's metrics, such as passenger counts. This result implies that any apparent year-over-year changes in the LA dataset should be interpreted with caution, as they do not reach the level of statistical significance typically used to confirm genuine differences.

**One-Sample T-Test:**

- The One-Sample T-Test applied to the Los Angeles (LA) dataset, comparing the mean value of a particular year to a hypothesized average, results in a statistic of -1.666 and a p-value of approximately 0.098. This indicates that the actual mean value for that year is not significantly different from the hypothesized mean, as the p-value exceeds the standard significance level of 0.05. The negative test statistic suggests that the observed mean is slightly lower than the hypothesized value, but this difference is not statistically significant. This outcome is crucial for understanding that the year's data aligns closely with the hypothesized average, indicating no substantial deviation in LA's airline traffic from expected trends for that period.

**Wilcoxon Signed-Rank Test:**

- The Wilcoxon Signed-Rank Test applied to the Los Angeles (LA) dataset, assessing differences in median values between two consecutive years, results in a statistic of 1749.0 and a p-value of approximately 0.601. This relatively high p-value suggests that there is no statistically significant difference in the median passenger counts (or similar metrics) between the years compared. The lack of a significant median difference indicates that, despite any potential variations in data, the central tendency of the dataset remains consistent from one year to the next. This finding is crucial for understanding the stability of LA's airline traffic across these years, implying that the operational and traffic conditions were relatively unchanged in terms of median values.

**Overview and Reasons:**

The statistical techniques applied here, including confidence intervals, ANOVA, Chi-Squared tests, T-Tests, and the Wilcoxon Signed-Rank Test, serve to understand the datasets' underlying structure and to compare different aspects of the flight data. Confidence intervals provide an estimated range of the mean passenger counts, while ANOVA tests help determine if there are significant differences in passenger counts over years or between different flight types. The Chi-Squared tests assess the independence of categorical variables. T-Tests compare the means of two groups, and the Wilcoxon test compares the medians of two related groups.

The outcomes of these tests offer a detailed statistical backdrop to the flight data, revealing trends, dependencies, and differences within and between the datasets from Ireland and LA, which are pivotal for data-driven decision-making in aviation and transportation planning.

**Section 4: Machine Learning Techniques**

**Model Selection and Rationale:**

**Linear Regression:**

**Overview:** A basic statistical approach  to modeling the relationship between a dependent variable and one or more independent variables.

**Reason:** Linear regression was chosen for its simplicity and ease of interpretation. This serves as a basic model for understanding linear relationships in your data. This is especially useful when you need a simple, explainable model  for initial insights.

**Random Forest Regressor:**

**Overview:** An ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of each tree.

**Reason:** Random forests are robust to overfitting and are effective at handling large datasets with multiple features.

It is suitable for complex datasets because it can capture nonlinear relationships and interactions between variables.

**Gradient Boosting Regressor:**

**Overview:** An ensemble method that builds one tree at a time. Each tree attempts to correct the errors of the previous tree.

**Reason:** Gradient boosting is effective for improving model accuracy and handling different types of data. It is known for its high predictive power and can optimize  different loss functions, making it versatile for different regression tasks.

**Neural Networks (MLPRegressor):**

**Overview:** A multilayer perceptron (MLP) is a type of artificial feedforward  neural network. MLPRegressor is used to approximate a function that maps input data to continuous output values.

**Reason:** Neural networks are chosen for their ability to model complex patterns and relationships in data. These are especially useful when  relationships between variables cannot be easily captured using traditional models. MLPRegressor is flexible and can be customized for different architectures for specific data characteristics.

|  |  |  |
| --- | --- | --- |
| **Model** | **Characteristics** | **Suitablity** |
| **Linear Regression** | Simple and interpretable, baseline model. | Effective for intial insights, assesses for linear relationships. |
| **Random Forest Regressor** | Robust aganist overfitting, handles complex datasets. | Manages non linear relationships , good for datasets with multiple features. |
| **Gradient Boosting Regressor** | High predictive power, versatile for different loss funtions. | Improves accuracy, suitable for complex regression analysis. |
| **Neural Network(MLP Regressor)** | Models complex patterns, flexible architecture | Ideal for capturing intricate patterns , effective for large datasets |

(Table 1. Comparing the best models that would fit our dataset)

**Comapring the models Before and after tuning the models:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Metric** | **Before Tuning** | **After Tuning** | **Performance description** |
| **Linear Regression** | MSE  RMSE  MAE | 1.39e-08  0.0001179  0.000079 | -  -  - | Stable, good baseline for comparision. |
| **Random Forest Regressor** | MSE  RMSE  MAE | 5.61e-09  0.0000749  0.0000333 | 6.20e-09  0.0000788  0.0000380 | Slightly higher error post tuning, robost to overfitting. |
| **Gradient Boosting Regressor** | MSE  RMSE  MAE | 1.34e.08  0.0001158  0.0000748 | 9.77e.09  0.0000988  0.0000611 | Improved accuracy and performance after tuning. |
| **Neural Network(MLP Regressor)** | MSE  RMSE  MAE | 0.4246  0.6516  0.6504 | 0.0020  0.0450  0.0449 | Substancially improvement, drastically lower error. |

(Table 2. Comparing the performance of models before and after tuning)

The Neural Network shows a remarkable improvement post-tuning, with significantly reduced error metrics, highlighting its effectiveness in handling the dataset after optimization. The Gradient Boosting Regressor also shows enhanced performance, whereas the Random Forest Regressor exhibits a slight increase in error.

**Best model and why:**

Neural network (MLPRegressor) turned out to be the best model for this dataset, mainly due to the significant performance improvement after hyperparameter optimization. All three error metrics (MSE, RMSE,  MAE) are significantly reduced, indicating that with appropriate parameters, the model is very effective at capturing complex patterns in the data. Its flexible and highly customizable architecture makes it especially suitable for large amounts of data, such as that in a project.

Gradient Boosting Regressor also shows good performance and may be a good choice, especially when interpretability and computational efficiency are important considerations. However, neural networks have a greater ability to adapt to nuances in the data, which gives them an advantage in this scenario

**Conclusions and Future Directions:**

Comparative analysis of Irish and Los Angeles (LA) air traffic datasets reveals distinct patterns. We found that Ireland had large variations in passenger traffic, whereas LA had consistent traffic with fewer extreme fluctuations. Differences indicate market movement. Annual trends highlight the impact of global events such as the pandemic, with Los Angeles showing a more significant impact. Statistical tests (ANOVA, chi-square test, t-test) provided deeper insights, such as the large annual differences in traffic volumes in Ireland. Machine learning models, especially after neural network tuning, showed strong predictive ability.

Future research could provide more comprehensive insights by extending the analysis to several more years of data, integrating external factors, and developing region-specific models. Moving to real-time analytics and investigating the impact of airline policy and market changes will further improve strategic decision-making in the aviation industry and provide a more dynamic and comprehensive understanding of air traffic trends.

**Refrences:**

1. Datasets Refrences:
2. Ireland Dataset1(taa02) - [data.gov.ie](https://data.gov.ie/dataset/taa02-passenger-freight-and-commercial-flights.)
3. Ireland Dataset2(taa03) - [data.gov.ie](https://data.gov.ie/dataset/taa03-passengers-freight-and-commercial-flights)
4. LA Dataset - [la-flights-dataset](https://dev.socrata.com/foundry/data.lacity.org/d3a2-7j6v)
5. Harrison, M. (2021). *Effective Pandas*.
6. Gorelick, M. (2020). *HIGH PERFORMANCE PYTHON : practical performant programming for humans.* S.L.: O’reilly Media.
7. Wilke, C. (2019). *Fundamentals of data visualization : a primer on making informative and compelling figures*. Sebastopol, CA: O’Reilly Media, Inc.

‌