**Exploratory Analysis of Flight Delay Dataset**

**1. Introduction: Business Problem and Data Identification**

In today’s aviation industry, flight delays represent a significant challenge for both airlines and passengers. These delays cause inconvenience and economic losses, leading airlines to seek data-driven insights to mitigate and manage these disruptions. This report focuses on analyzing a flight dataset to uncover patterns and trends in departure and arrival delays. The goal is to identify key factors contributing to delays and provide actionable insights that airlines can use to improve operational efficiency and enhance customer experience.

The dataset under analysis includes various fields such as the scheduled day of the week, departure and arrival times, actual departure and arrival times, and delay information. Specifically, key columns include:

* **DAY\_OF\_WEEK**: Coded as integers, where 1 = Monday, 2 = Tuesday, and so on.
* **IATA\_CODE\_AIRLINE**: IATA code for the airline operating the flight.
* **ORIGIN\_AIRPORT** and **DESTINATION\_AIRPORT**: The IATA codes of the origin and destination airports.
* **SCHEDULED DEPARTURE\_TIME** and **SCHEDULED ARRIVAL\_TIME**: The planned departure and arrival times.
* **ACTUAL ARRIVAL\_TIME**: The actual time of flight arrival.
* **ARRIVAL\_DELAY In Minutes**: The delay is calculated as the difference between actual and scheduled arrival times.

Our goal in this exploratory analysis is to assess the quality of the data, perform necessary transformations, and generate meaningful insights through visualizations and summary statistics.

**2. Data Cleaning and Transformation**

**2.1. Identifying and Extracting Relevant Fields**

The first step in the analysis involves identifying relevant columns from the dataset. The dataset contains 26 columns, some of which are directly related to flight timings and delays. Fields such as **YEAR**, **MONTH**, **IATA\_CODE\_AIRLINE**, **ORIGIN\_AIRPORT**, **DESTINATION\_AIRPORT**, and **ARRIVAL\_DELAY In Minutes** are essential for the analysis.

**2.2. Handling Missing or Incorrect Data**

Before performing any meaningful analysis, it is crucial to clean the data. Initial inspection of the dataset revealed several issues:

* **DAY\_OF\_WEEK** contained invalid entries such as 'WEDS', 12, and 10, which do not fit within the valid range (1-7).
* Some fields contained missing or null values, especially in delay-related columns.

The following cleaning actions were performed:

1. **DAY\_OF\_WEEK correction**: Invalid entries in DAY\_OF\_WEEK were mapped to appropriate values. For example, 'WEDS' was mapped to 3 (Wednesday), 12 was assumed to be Monday (1), and 10 was mapped to Tuesday (2). Additionally, any rows with missing values in this field were dropped.
2. **Date and time formatting**: Columns representing time data (e.g., SCHEDULED DEPARTURE\_TIME, SCHEDULED ARRIVAL\_TIME, and ACTUAL ARRIVAL\_TIME) were converted to datetime format. This conversion ensured that calculations involving differences in time (for delay analysis) could be performed accurately.
3. **Outliers**: Flight delay data often contain outliers, such as flights with extreme delays that might skew the results. Outliers in delay columns (e.g., extremely high or negative delay values) were identified using descriptive statistics such as mean and standard deviation. A threshold was set to filter out any delays greater than 500 minutes, which are highly improbable for most commercial flights.

**3. Exploratory Visualizations and Summary Statistics**

Once the data was cleaned, the next step involved visualizing key patterns and calculating summary statistics to gain insights into flight delays.

**3.1. Visualization of Delay Patterns by Day of the Week**

One of the core objectives was to identify if certain days of the week experienced more delays than others. The cleaned dataset was grouped by DAY\_OF\_WEEK, and the average arrival delay was calculated for each day.

A bar chart was generated to display the average arrival delay by day of the week. The results showed that:

* Flights on Monday (DAY\_OF\_WEEK = 1) experienced the least delays on average.
* Friday and Saturday (DAY\_OF\_WEEK = 5 and 6) saw the highest average delays, potentially due to increased flight volumes at the end of the week.

This insight can help airlines allocate additional resources or adjust schedules on high-traffic days to reduce delays.

**3.2. Airline-Specific Delays**

Next, we explored whether certain airlines were more prone to delays. The dataset was grouped by IATA\_CODE\_AIRLINE to calculate the average delay for each airline.

A boxplot was generated to visualize the spread of delays for the top five airlines. Interestingly, while most airlines had similar median delays, some airlines exhibited a higher frequency of extreme delays (outliers). This could be due to operational inefficiencies or the geographic regions where they predominantly operate. Identifying these trends can help airlines review and optimize their schedules and resource allocation.

**3.3. Airport-Specific Delays**

The analysis was extended to identify the airports where delays were most common. Both the ORIGIN\_AIRPORT and DESTINATION\_AIRPORT fields were analyzed to understand how delays were distributed across different airports.

The top five origin and destination airports with the highest average delays were identified. Airports with high flight traffic, such as Los Angeles International (LAX) and Chicago O'Hare (ORD), predictably showed higher delays, possibly due to congestion. This insight can prompt airports to evaluate runway usage, ground crew efficiency, and scheduling practices to reduce delays.

**3.4. Delay Distribution Across the Year**

The dataset also contained YEAR and MONTH columns, which allowed us to analyze delay trends throughout the year. A line chart was created showing the monthly average delays across all years in the dataset. The results highlighted a seasonal trend:

* Delays increased during the winter months (December to February), likely due to adverse weather conditions.
* Summer months (July and August) also experienced spikes in delays, which could be attributed to higher travel demand.

This seasonal pattern can be leveraged by airlines to prepare for peak periods and implement strategies such as optimizing staff schedules or adjusting flight times to mitigate expected delays.

**4. Diagnostic Techniques and Descriptive Measures**

In this exploratory analysis, several descriptive and diagnostic techniques were used to uncover insights from the data:

1. **Mean and Median**: The mean and median of delay times were calculated to understand the central tendency of the delay distribution. The difference between mean and median (in some cases) indicated the presence of outliers or skewness in the delay data.
2. **Boxplots**: Boxplots were used to visualize the spread and distribution of delays, helping to identify not only the typical delay but also outliers. These plots were instrumental in understanding variability across different airlines and airports.
3. **Correlation Analysis**: To investigate whether certain factors were associated with delays, a correlation matrix was generated. It was found that departure delays were strongly correlated with arrival delays (as expected), but no strong correlations existed between delays and the scheduled departure or arrival times.
4. **Filtering and Aggregation**: Techniques such as filtering out extreme delays and grouping data by relevant fields (e.g., airline, day of the week) provided a clearer picture of the trends and patterns in the dataset.

**5. Initial Findings and Insights**

Based on the exploratory analysis, several key insights were derived:

* **Day of the Week**: Flights on Fridays and Saturdays experience the highest delays, suggesting that these are peak days for travel. Airlines could focus on optimizing their schedules and resources on these days to minimize delays.
* **Airlines**: While most airlines showed similar average delays, some were more prone to extreme delays. Investigating these airlines' operations might provide actionable steps to improve their on-time performance.
* **Airports**: Major airports with high traffic volumes consistently showed higher delays. Infrastructure improvements or better traffic management could reduce congestion-related delays.
* **Seasonal Trends**: Delays spike during winter months due to weather conditions and in summer due to travel demand. Airlines should prepare accordingly for these peak periods.

**6. Conclusion**

This exploratory analysis of flight delays provided valuable insights into when and where delays are most likely to occur. By cleaning the data, identifying patterns, and applying diagnostic techniques, we uncovered trends that can inform decision-making for airlines and airport authorities. Future analysis could include predictive modeling to forecast delays or expand the dataset to include additional features such as weather data to further refine the analysis.

These findings can help airlines focus on critical areas where they can reduce delays, improve efficiency, and ultimately enhance the passenger experience.