Exploratory Data Analysis

Airline Flight Delay Analysis & Prediction

Assignment by: Deepti Joshi

Mentor: Maimuneesa Kasi

July 6, 2025

Contents

1. Overview

In the highly interconnected world of air travel, flight delays and cancellations not only disrupt schedules but also impact airline operations, passenger satisfaction, and overall travel efficiency. Among these disruptions, flight cancellations pose a significant challenge, often resulting from a variety of causes such as weather conditions, air traffic control delays, technical problems, crew availability, or security concerns. Understanding the underlying patterns behind these cancellations is essential for airlines, airport authorities, and policymakers to make data-driven decisions that improve reliability and minimize inconvenience for travelers.

The primary purpose of the dataset considered is to provide comprehensive insights into the operational performance of U.S. air carriers across various airports from August 2013 to August 2023, with a particular focus on flight arrivals and associated delays. By detailing metrics such as the total number of arriving flights, delays exceeding 15 minutes, cancellations, and diversions, the dataset allows for in-depth examination of the underlying causes of disruptions. These causes are categorized into factors such as carrier-related problems, weather conditions, National Airspace System (NAS), security concerns, and delays from late arrival aircrafts. The dataset is designed to support researchers, data analysts, and aviation professionals in identifying patterns, understanding trends, and drawing actionable

conclusions that enhance the broader understanding of operational challenges within the U.S. aviation industry.

In conclusion, this analysis offers a clearer understanding of the patterns and root causes behind flight delays and cancellations in the US aviation sector over the past decade. By examining various delay factors and their frequency between carriers and airports, we gain valuable insights into operational challenges and areas for improvement.

2. Project Introduction

Air travel plays a vital role in connecting people and businesses across vast distances, serving as the backbone of domestic and international transportation. However, the efficiency of air travel is often challenged by delays, cancellations, and diversions, which not only inconvenience travelers but also have significant economic implications for airlines and airports. Understanding the underlying causes and patterns of these disruptions is essential to improve the performance and reliability of the aviation sector.

This project focuses on the analysis of a comprehensive dataset detailing flight arrival statistics for US carriers from August 2013 to August 2023. The data set includes rich information on the number of arriving flights, the frequency of delays exceeding 15 minutes, and various causes of delays—ranging from airline-related issues to weather conditions, security concerns, and late-arriving aircraft. It also includes data on flight cancellations and diversions across multiple airports.

By leveraging exploratory data analysis (EDA) and data visualization techniques, this project aims to uncover trends, identify high-risk factors, and highlight seasonal or operational patterns in flight delays and disruptions. The ultimate goal is to generate insights that can inform airline strategy, airport operations, and public policy, contributing to a more efficient and passenger-friendly air travel system.

3. Project Objectives

The primary objective of this project is to analyse and interpret flight arrival data of U.S. carriers from August 2013 to August 2023, with a focus on understanding the frequency, causes, and patterns of flight delays, cancellations, and diversions. By performing comprehensive exploratory data analysis (EDA) and creating insightful visualizations, the project seeks to:

- Identify trends in arrival delays across different carriers, airports, and time periods.
- Examine the major contributing factors to delays, such as carrier-related issues, weather conditions, NAS (National Airspace System) constraints, security incidents, and late-arriving aircraft.
- Analyse cancellation and diversion patterns to understand their distribution over time and geography.
- Compare carrier performance in terms of delay frequency and management.
- Provide actionable insights that can help stakeholders in the aviation industry—airlines, airport authorities, and policy-makers make data-driven decisions to improve operational efficiency and passenger experience.

This analysis aims to bridge the gap between raw flight data and

meaningful interpretations, ultimately contributing to the ongoing efforts to enhance the reliability and punctuality of air travel in the United States.

4. Dataset Description

The data set used in this project provides a comprehensive record of US domestic flight arrival statistics from August 2013 to August 2023, compiled by the U.S. Department of Transportation (DOT). It captures detailed monthly metrics for various airlines and airports throughout the United States, allowing in-depth analysis of flight delays, cancellations, and overall airline performance. Each row in the data set represents flight arrival information for a specific carrier-airport-month-year combination. The data set includes the following key attributes:

- **year:** The year of the data. This data set has records from year 2013 till 2023
- **month:** The month of the data. This data set has records for months 1–12.
- **carrier:** Carrier code (e.g., AA, DL, UA)
- **carrier_name:** Carrier name
- **airport:** Airport code (e.g., JFK, LAX)
- **airport_name:** Airport Name
- **arr_flights:** Total number of flights arrived.
- **arr_del15:** Number of flights delayed by 15 minutes or more.

- **carrier_ct:** Carrier count (delay due to the carrier).
- **weather_ct:** Weather count (delay due to weather).
- **nas_ct:** NAS (National Airspace System) count (delay due to the NAS).
- **security_ct:** Security count (delay due to security).
- **late_aircraft_ct:**Late aircraft count (delay due to late aircraft arrival). Delay due to late arrival of aircraft from a previous flight.
- **arr_cancelled:** Number of flights canceled.
- **arr_diverted:** Number of flights diverted.
- **arr_delay:** Total arrival delay in minutes

  Breakdown of total arrival delay is the sum of all the delays below:

  - **carrier_delay:** Delay minutes caused by airline-related issues such as maintenance or crew problems.
  - **weather_delay:** Delay minutes due to adverse weather conditions impacting flights.
  - **nas_delay:** Delay minutes caused by National Airspace System congestion or restrictions.
  - **security_delay:** Delay minutes resulting from security related checks or incidents.
  - **late_aircraft_delay:** Delay minutes caused by the late arrival of the aircraft from a previous flight.

The dataset is clean, structured, and suitable for both quantitative and visual exploratory analysis. It provides valuable insights into the operational performance of airlines and airports over a 10-year period, enabling the identification of delay trends and contributing factors.

5. Data Preprocessing and Cleaning

Before conducting any meaningful analysis, it is essential to ensure the data is clean, accurate, and consistent. Real-world data sets, especially those span multiple years and sources as in the case of airline performance data often come with various issues such as missing values, duplicates, inconsistent formatting, and incorrect data types.

In this project, the data set covers US flight arrival statistics over a ten-year period (August 2013 – August 2023), sourced from multiple carriers and airports. The primary goal of the cleaning process was to prepare the data for Exploratory Data Analysis (EDA) by:

- **Handling missing values:** Handling missing values is a crucial step in preparing data for analysis. The dataset was examined for missing or null values. Some null values were found in columns like arr_flights, arr_del15, carrier_ct, weather_ct, nas_ct, arr_cancelled. These values were dropped as they were no so significant in terms of count of missing values.
- **Checking for duplicate values:** No duplicates were found after deleting the null values so no action taken.

After completing the data cleaning and pre-processing steps, the refined dataset was stored in a MySQL database. The data which was initially available as a .CSV file and is stored into a table after defining primary key constraints.

6. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the process of examining and visualizing data sets to:

- Understand the structure and quality of the data
- Detect patterns, trends, and anomalies
- Form hypotheses or guide further analysis/modelling
- Validate assumptions using summary statistics and visual methods.

EDA can be categorized into three main forms:

1. **Univariate Analysis**

   Focus: One variable at a time.

   Goal: Understand the distribution, central tendency, spread, and presence of outliers.

2. **Bivariate Analysis**

   Focus: Relationship between two variables

   Goal: Explore correlations or associations.

3. **Multivariate Analysis**

   Focus: Interactions among three or more variables

Goal: Detect complex relationships, interactions, or groupings in the data.

6.1 Univariate Analysis

As the initial step in Exploratory Data Analysis (EDA), we examine individual variables to understand their overall distribution and key characteristics.

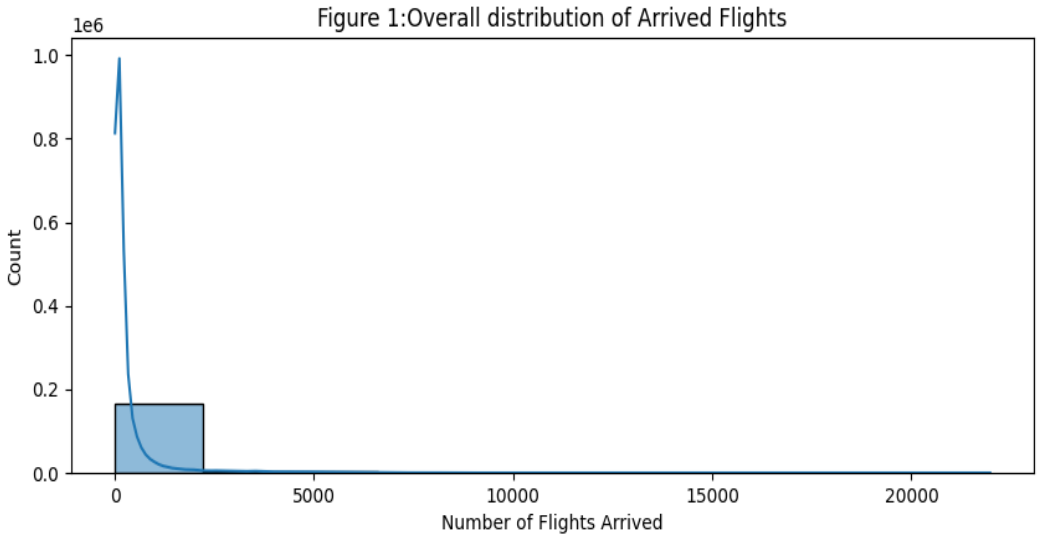

Figure 6.1: Over all Distribution of Arrived Flights

The x-axis representing the number of arrived flights and the y-axis representing the counts. The distribution here shows a highly skewed one. Most of the airports have the range of 0-2,500 arrived flights. A long tail at the end shows that some airports have higher flight arrival count or there are some outliers.

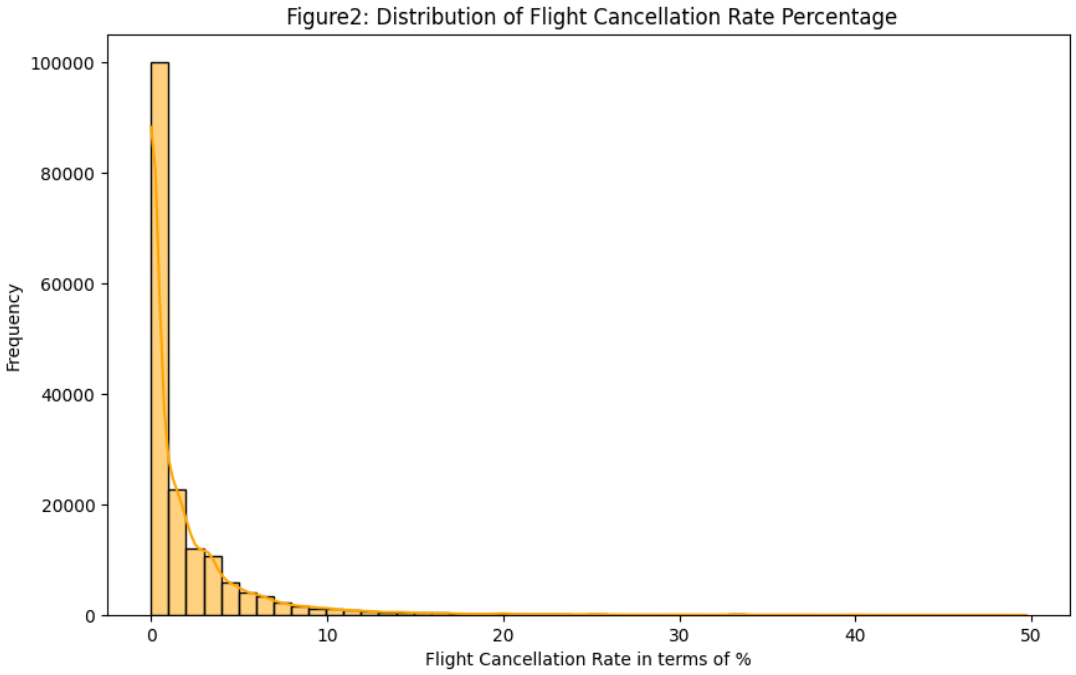

Figure 6.2: Distribution of Flight cancellation rates in %

The distributions for arrival canceled flights was a very tightly skewed plot so calculating the cancellation rate of the flights. Figure2 shows the distribution of flight cancellation rates which describes that many airports have canceled arrivals between 0 and 10%. The long tail shows that some airports have had up to 50% cancellations in that month.

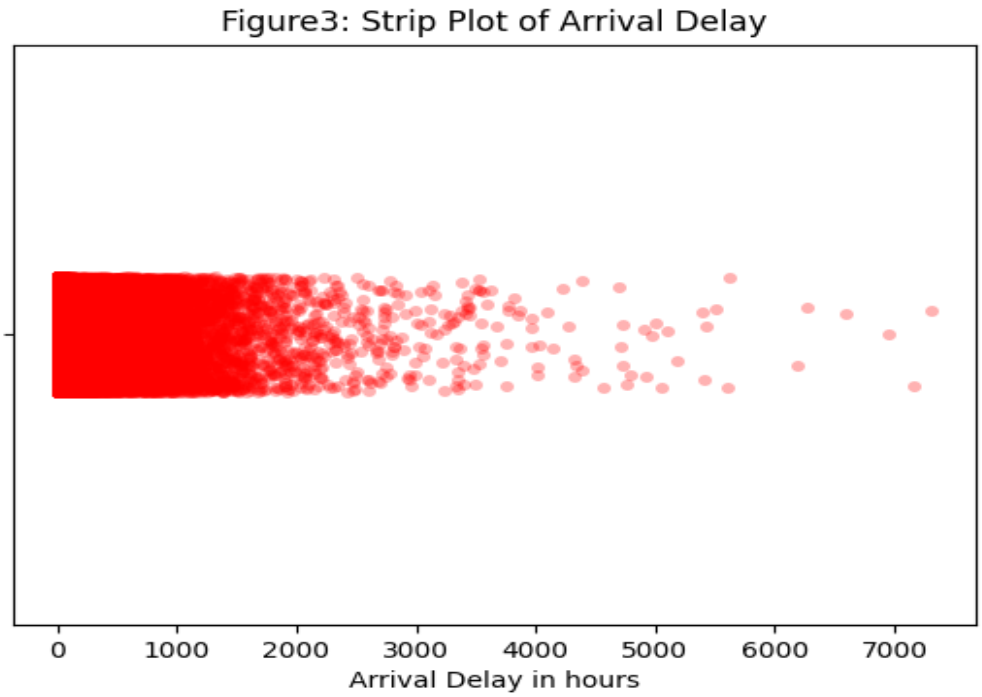

Figure 6.3: Strip Plot of Arrival Delay

As shown in Figure 3 most of the arrival delays are concentrated between 0 and 1,000 hours. After that the data becomes scattered. This can indicate the presence of outliers. So initial understanding of the arrival delay suggests that many flights have smaller arrival delays.

6.2 Bivariate Analysis

Bivariate analysis explores the relationship between two variables to identify potential correlations, trends, or patterns.

Before diving deeper into the analysis, the first step is to calculate the total number of delayed flights by summing all delay causes for each record. This gives us a comprehensive count of all delayed flights regardless of reason. Once calculated, we can visualize the relationship between the total delayed flights and the total number of flights

that arrived.

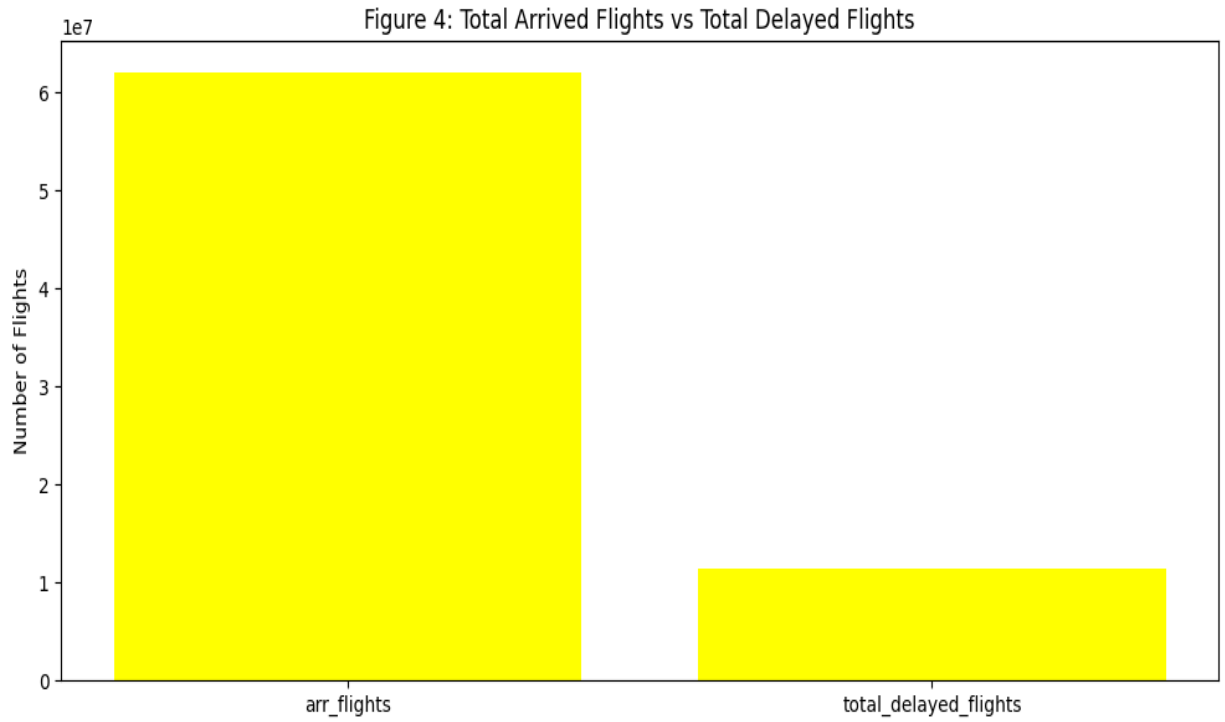

Figure 6.4: Total arrived flights vs total delayed flights

The bar representing total arrived flights is taller because it accounts for every flight that landed, while the bar for total delayed flights represents only the subset of those flights that experienced delays. Since delays affect only a portion of the total flights, the count of delayed flights is naturally lower than the total arrivals. This difference confirms the data's consistency and reflects real-world flight operations, where not all flights are delayed.

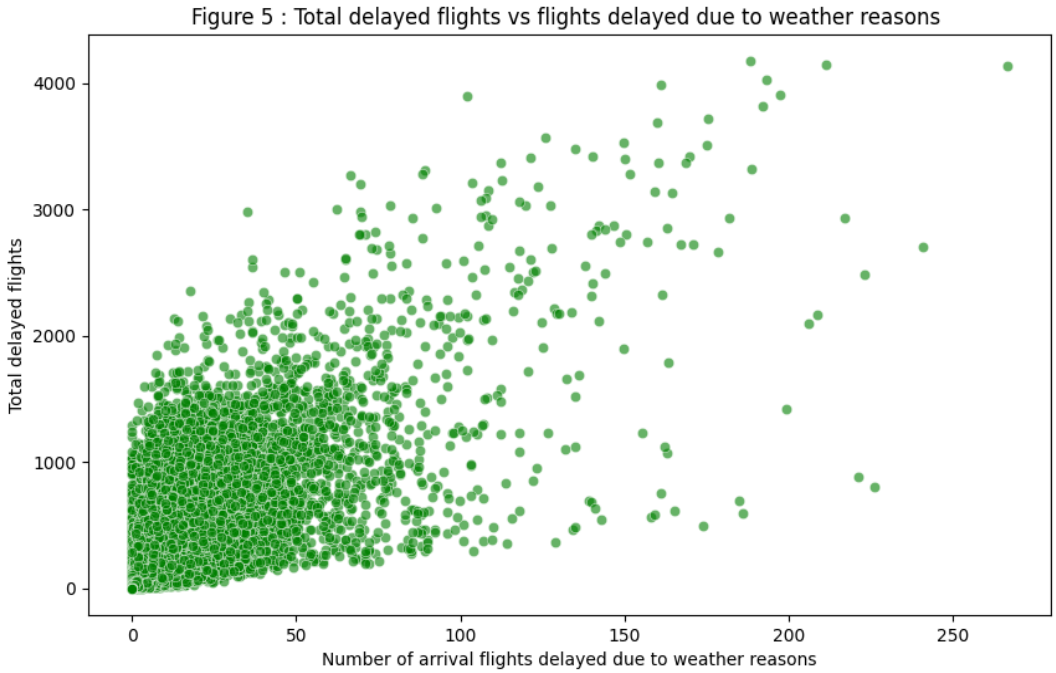

Figure 6.5: Arrival Delay vs Number of flights delayed due to weather

Figure 5 suggests that most data points are concentrated towards left corner between 0 and 50 on the x-axis and 0-1,000 on the y-axis.The graph is in increasing trend but also shows variability saying that not all weather related delays are contributing to total flight delay. So, the number of flights delayed due to bad weather conditions are low in most of the cases.

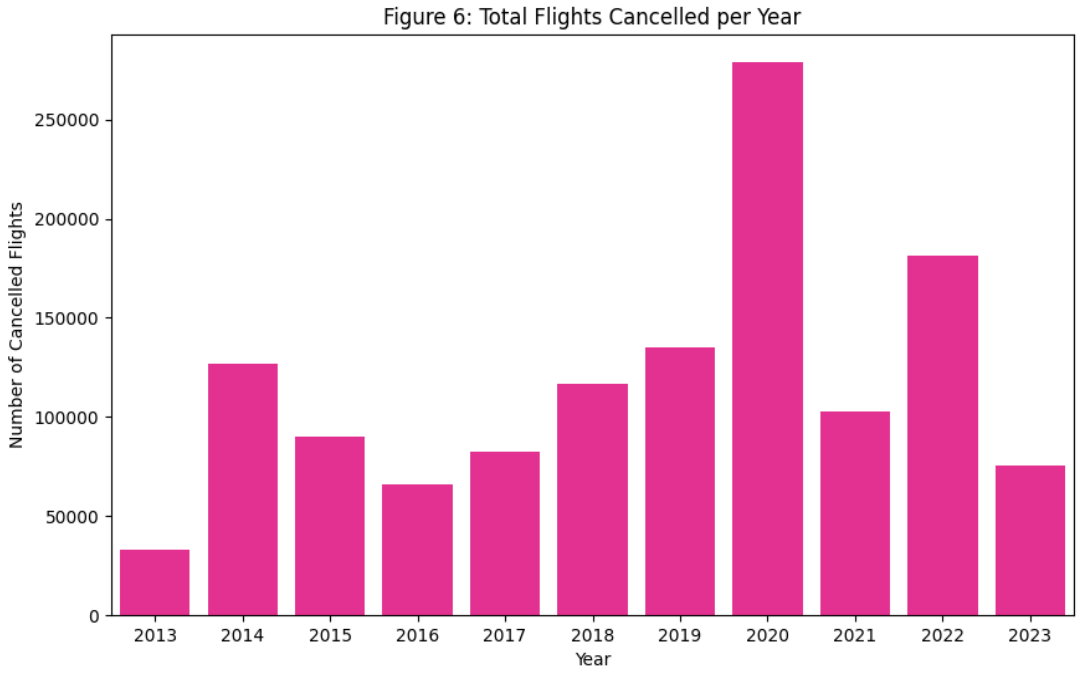

Figure 6.6: Total Flights canceled per year

As per Figure 6 the highest flight cancellations have happened in the year 2020. This aligns with the onset of the COVID-19 pandemic. During this period, widespread travel restrictions, reduced passenger demand, and public health concerns led to an unprecedented number of flight cancellations across the aviation industry.

6.3 Multivariate Analysis

Multivariate analysis involves examining three or more variables simultaneously to understand complex relationships and interactions in the data. Unlike univariate or bivariate analysis, multivariate analysis helps uncover how multiple factors together influence outcomes.This approach provides deeper insights into patterns and dependencies.

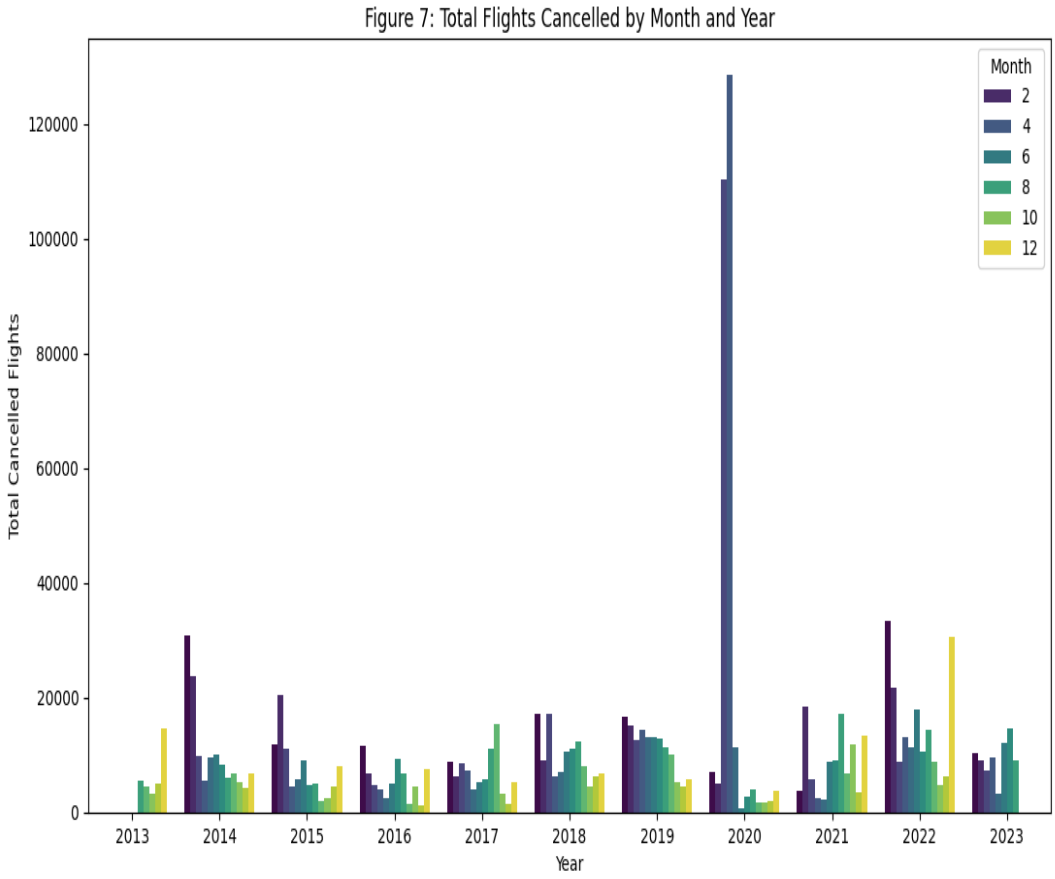

Figure 6.7: Total Flights canceled per year per month

As per Figure 7,the data reveals that the highest number of flight cancellations occurred in the year 2020, with a pronounced peak in April of that year. April 2020, in particular, stands out as the most affected month, reflecting the height of early pandemic-related impacts on flight operations. So this adds to more details to the results obtained from Figure 6.

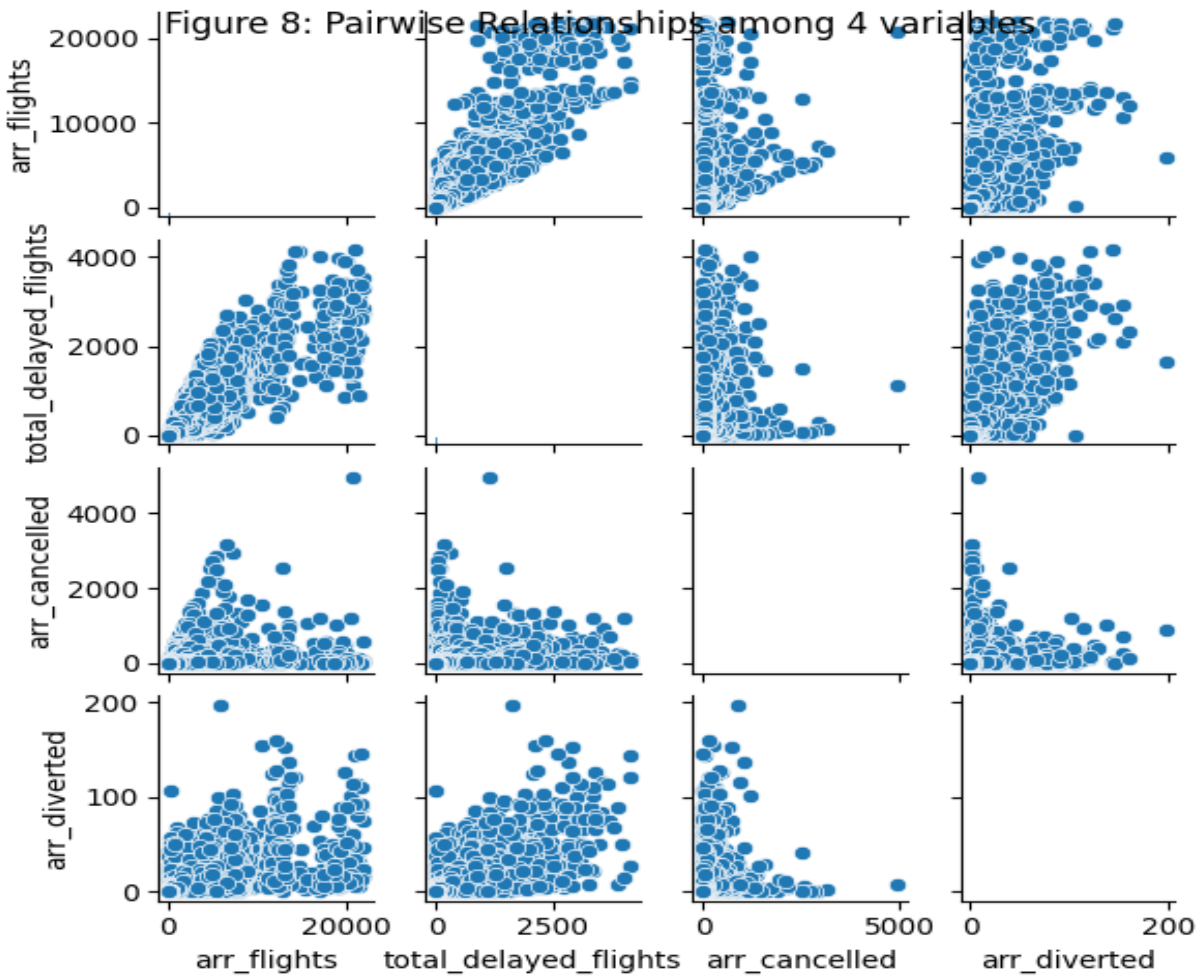

Figure 6.8: Pairwise Relationship among 4 attributes

The pair plot offers a clear visual overview of the relationships between metrics such as total flights arrived, total delayed flights, cancellations, and diversions. By examining these four variables together, we can observe general trends and potential correlations—for instance, as the number of arriving flights increases, there may also be a rise in delays, cancellations, or diversions.

The pair plot between flights diverted and flights canceled shows a moderate or weak positive correlation.Because these two variables are not necessarily interlinked.

The pair plot between flights canceled and total flights delayed is

expected to show a moderate correlation. Some flights can be delayed without being canceled, and some flights might be canceled without being significantly delayed.

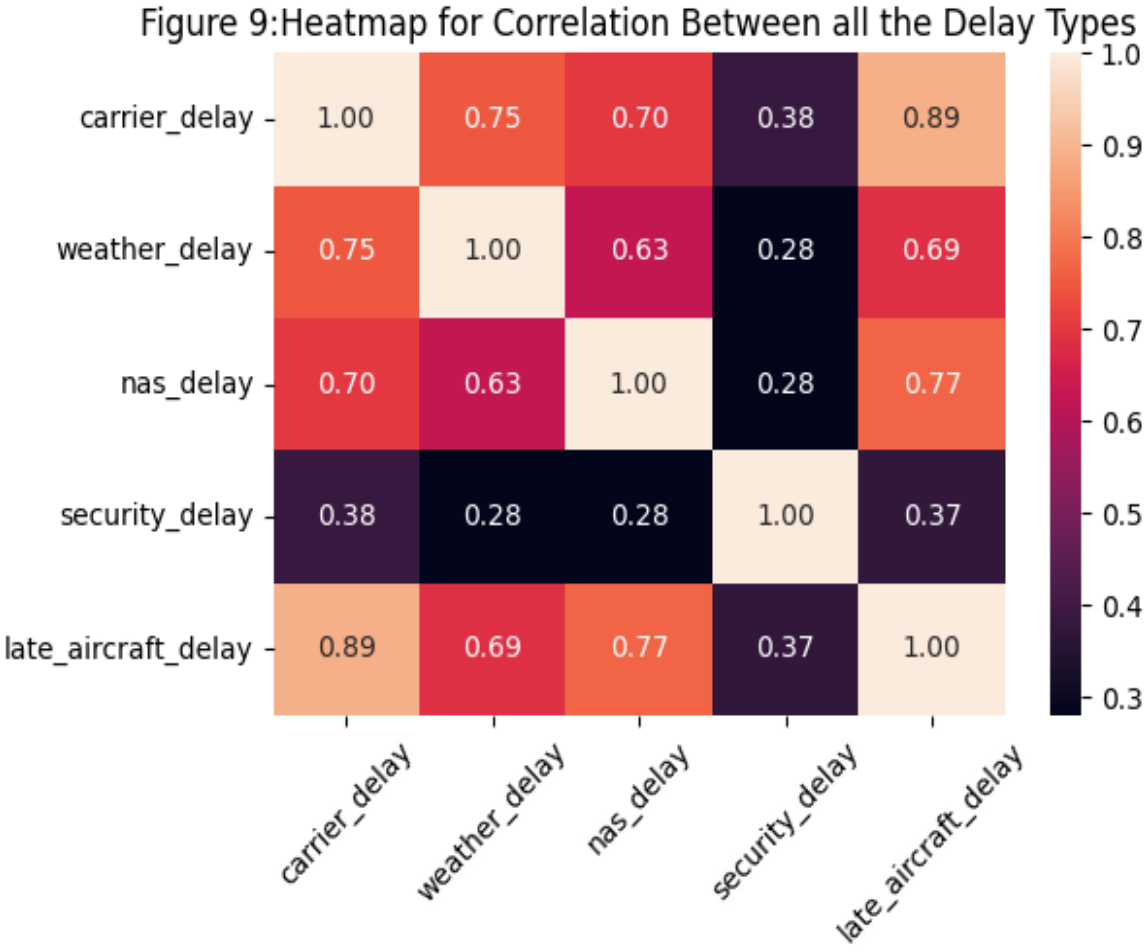

Figure 6.9: Heatmap representation

The heatmap illustrates the correlation between five types of flight delays: carrier delay, weather delay, NAS delay, security delay, and late aircraft delay. A value close to 1.0 indicates a strong positive correlation. For example, late aircraft delay shows a strong positive correlation with carrier delay, suggesting that delays caused by late arrivals are often linked to carrier-related operational issues. On the other hand, security delay exhibits very low correlation with other

delay types like weather related delays or NAS delays, indicating it occurs independently.

6.4 Meaningful insights from above analysis

By starting with univariate, bivariate, and multivariate analyses, deeper insights from the data can be gradually discovered. Looking at single variables helps spot important patterns, comparing two variables shows how they influence each other, and exploring multiple variables together reveals complex relationships. Deriving more meaningful insights from the above analysis as follows:

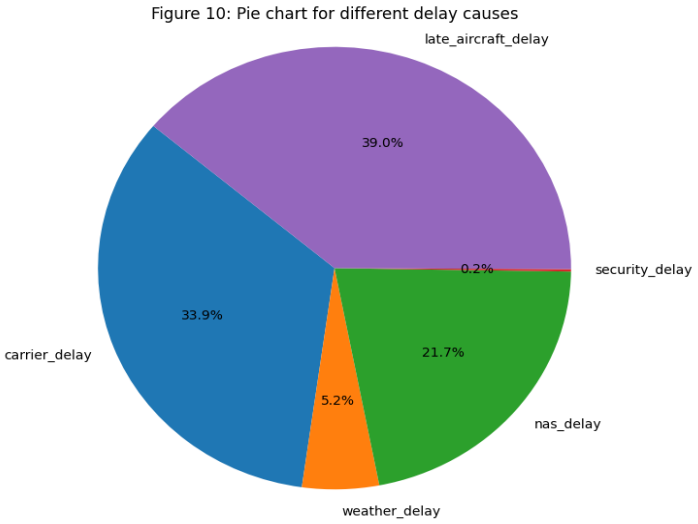

Figure 6.10: Pie chart for different delay causes

As per figure 10, The pie chart visualizes the breakdown of the flight delay causes by showing the proportion each factor contributes to total delays. Each slice represents a delay category like Carrier, Weather, NAS, Security, or Late Aircraft. Based on the graph, late aircraft arrivals are the largest cause of flight delays, taking up the

biggest portion of the chart and security-related delays make up the smallest slice, indicating they contribute the least to delays compared to other factors

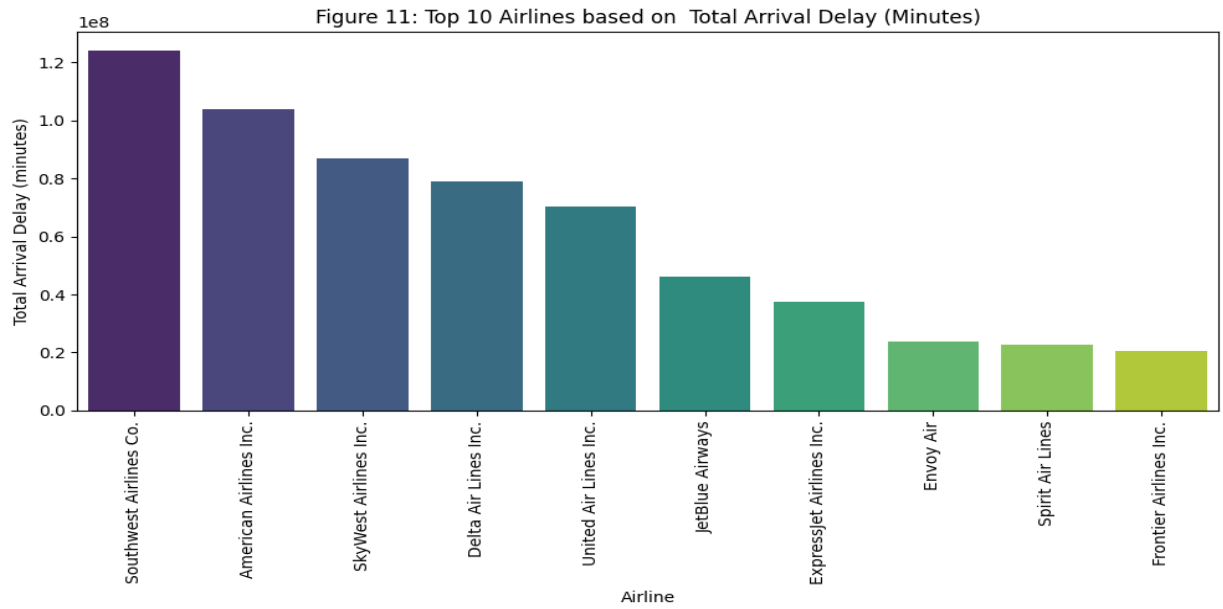

Figure 6.11: Top 10 Airlines based on Total Arrival Delay (Minutes)

As per Figure 11: The bar chart displays the top 10 airlines with the highest total arrival delay in minutes, allowing for a clear comparison of delay performance across carriers. Each bar represents the cumulative arrival delay for an airline. The visualization shows that Southwest Airlines Co has highest delays than others, highlighting potential areas for operational improvement or further investigation into delay causes.

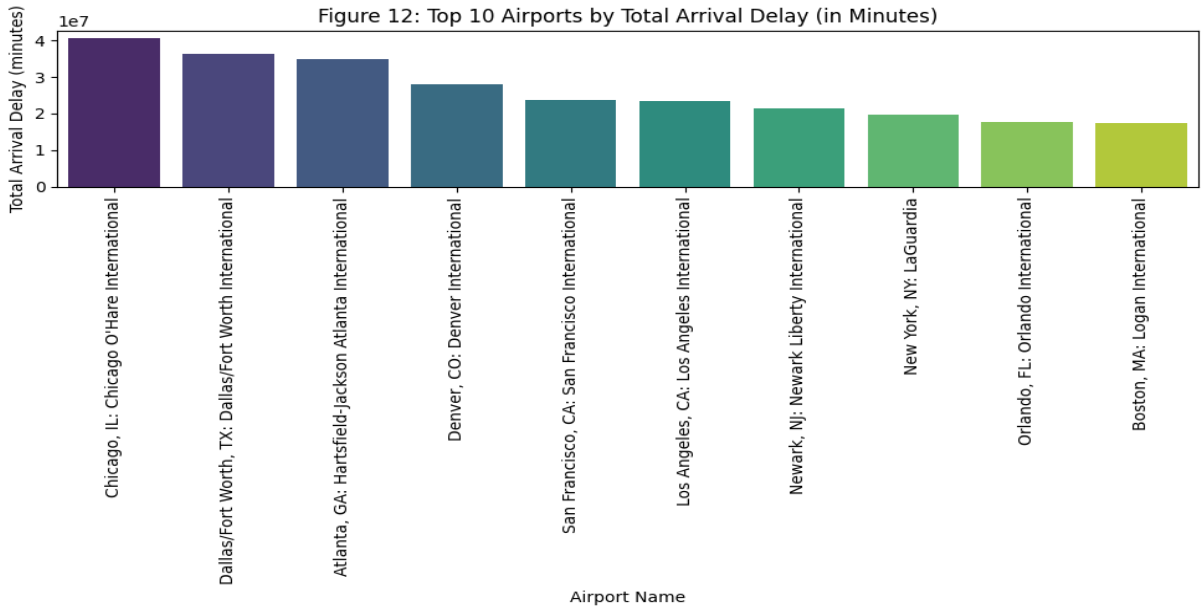

Figure 6.12: Top 10 Airports based on Total Arrival Delay (Minutes)

Figure 12: The bar chart shows the top 10 airports with the highest total arrival delays measured in minutes. This visualization helps pinpoint which airports contribute most to the delays. In this data set the airport at Chicago IL shows the highest arrival delays.

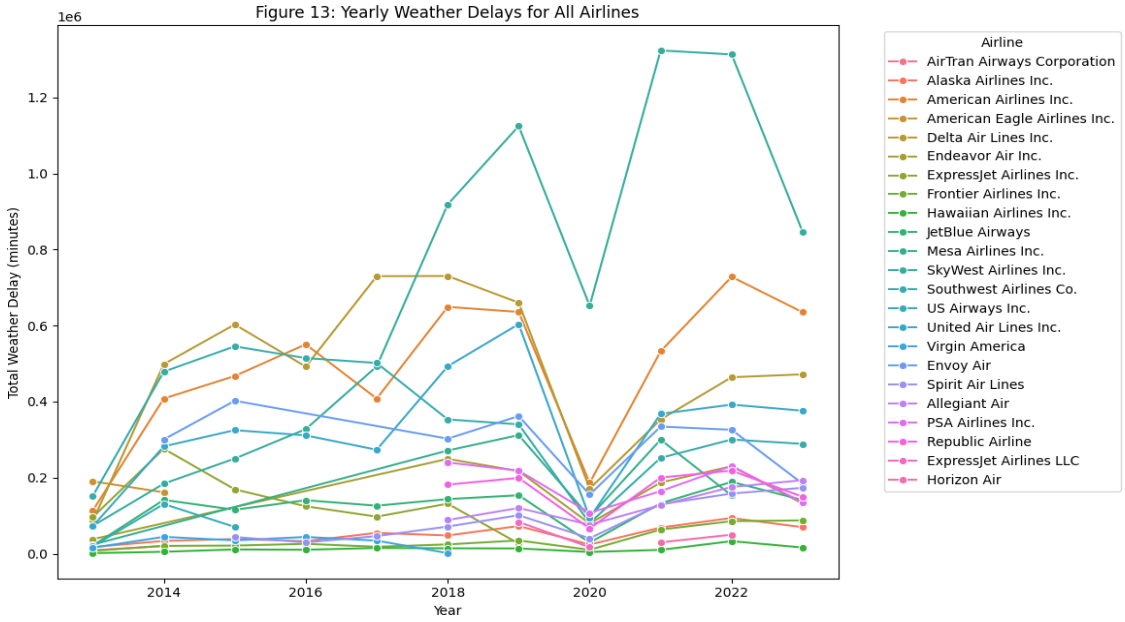

Figure 6.13: Yearly Weather Delays for All Airlines

As per Figure 13: This line graph illustrates the total weather-related delays experienced by each airline across different years. Each colored line represents a distinct airline, showing how their weather delay minutes have varied over time.

7. Predictive Modelling

Predictive modeling involves using historical data to create mathematical models that can forecast future outcomes—in this case, flight arrival delays. By analyzing the highest contributing delay factors such as carrier-related delays and delays caused by late arriving aircraft, the model uncovers underlying patterns and relationships affecting flight punctuality.

Metrics such as Mean Squared Error and R-squared assess how accurately the model predicts delays and how well the chosen features explain the variations in flight punctuality. The current data set has R-squared value: 0.96 indicating a strong fit and reliable predictive power.

Overall, these results suggest the model does a very good job capturing the main factors affecting flight delays, though there is still some variability because not all factors are considered.

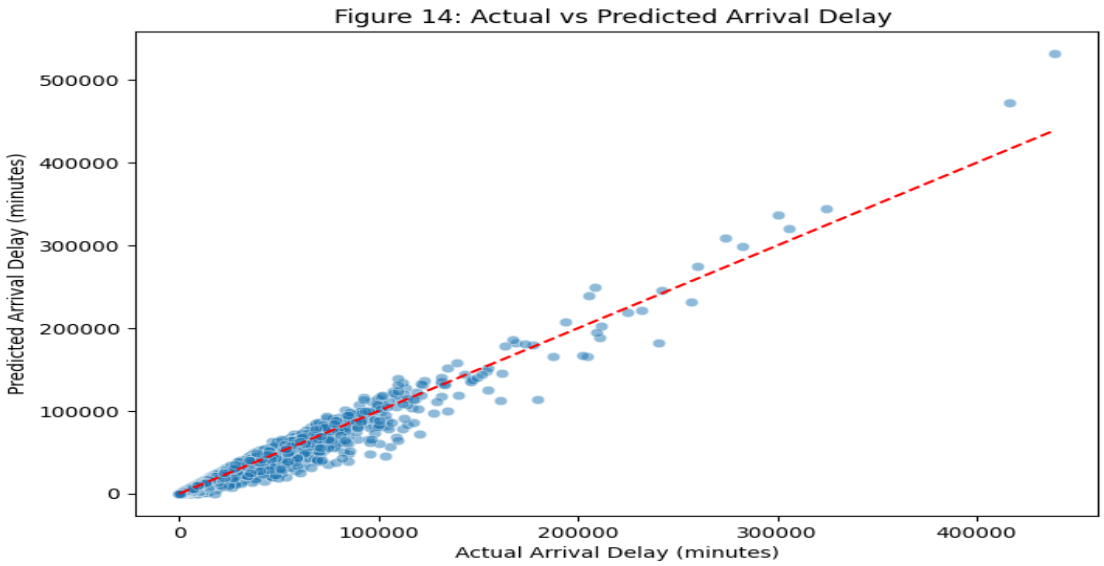

Figure 7.1: Linear Regression, Actual vs Predicted Arrival Delay

As per Figure 14: The scatter points being close to the red dotted diagonal line indicate that the predicted arrival delays closely match the actual delays. This means the model is accurately estimating flight delays for most of the data.

8. Visualization and Dashboard Development with Results Analysis

Visualization and dashboard development involves transforming raw data into clear, interactive visuals to understand key metrics and trends. Data from MySQL tables is imported into Power BI for analysis.

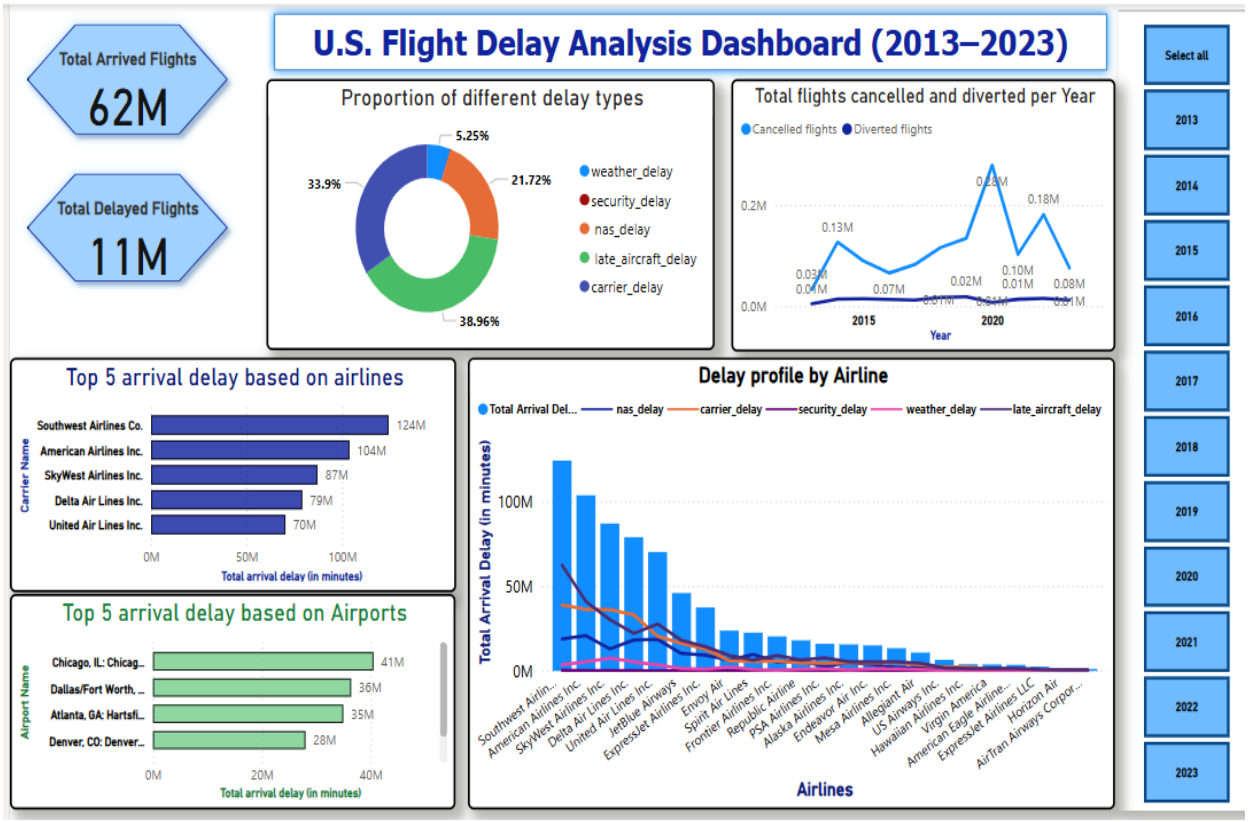

Figure 8.1: US Flight delay analysis dashboard

Exploring the dashboard to understand the underlying patterns

and metrics.

1. The dashboard features two key performance indicator (KPI) cards:

   - Total Arrived Flights: This card displays the overall number of flights arrived.
   - Total Delayed Flights: This card indicates the count of flights that arrived later than their scheduled time, highlighting potential areas for operational improvement.

     Generally it has been seen that the flights are on time.

2. The pie chart illustrates the distribution of various delay types contributing to overall flight delays. Each segment represents a specific category, such as Carrier Delay, Late Aircraft Delay, NAS Delay, Security Delay, and Weather Delay. The size of each segment correlates with the percentage of total delays attributed to that particular cause.

   In analyzing flight delays, late-arriving aircraft is consistently the highest contributor, accounting to about 39% of total delay minutes. This category encompasses delays where a previous flight with the same aircraft arrived late, hence causing the subsequent flight to depart behind schedule. Consequently, addressing the root causes of late arrivals is crucial to improve overall punctuality and operational efficiency.

   Following closely is the carrier delay, which typically accounts for around 34% of total delay minutes. This category includes delays attributable to the airline's internal operations, such as maintenance issues, crew shortages, fueling delays, and baggage handling problems. Addressing these issues requires airlines to

enhance internal processes, invest in staff training, and improve maintenance scheduling to minimize delays and improve overall service reliability.

3. The bar charts rank the top five Airlines and Airports based on delayed arrivals. According to the data set considered, Southwest Airline Co airline experienced the longest arrival delay. The Airlines experienced significant delays due to previous late aircraft and carrier delay.

   Chicago O'Hare International Airport experienced the highest delay in arrivals. The reasons being the same as above. Addressing these factors is crucial for improving punctuality and passenger satisfaction.

4. The line chart illustrates the trends in flight cancellations and diversions over time. In 2020, the aviation industry experienced an unprecedented surge in cancellations, primarily due to the COVID-19 pandemic. While occasional spikes may occur due to unforeseen circumstances such as weather events or technical issues. The diversions are mostly consistent.

5. The Line and clustered column Chart illustrating the delay profile by airline provides a comparative visualization of total arrival delays and the contributions from various delay factors for each airline. This chart typically includes categories such as Carrier Delay, Late Aircraft Delay, National Aviation System (NAS) Delay, Weather Delay, and Security Delay.

   Across most airlines, Late Aircraft Delay and Carrier Delay are the predominant contributors to overall arrival delays. These factors often stem from operational inefficiencies and scheduling issues within the airline's control.

The chart highlights variations between airlines. For example, Southwest Airlines Inc Co has the highest arrival delay, followed by American Airlines and so on. By analyzing the contributions of each delay factor, airlines can implement targeted strategies to enhance punctuality and overall service quality.

6. The Year Slicer helps to filter and analyze the data across a specific time frame, from 2013 to 2023. By interacting with this slicer, one or multiple years can be selected to dynamically adjust the visuals on the dashboard, providing a focused view of the data for the chosen period.

9. Areas for Improvement and Future Outlook

1. Enhancing Operational Efficiency: the late arriving aircraft and carrier delays are the primary contributors to overall flight delays. Implementing strategic processes, such as Fast-tracked boarding, refueling, and baggage handling, can significantly reduce ground time and improve on-time performance. Additionally, real-time data analytics can help identify and mitigate potential delays before they impact flight schedules.

2. Improving Weather Forecasting and Response: Weather-related delays remain a significant challenge, accounting for a substantial portion of flight disruptions. Advancements in weather forecasting for more accurate and timely weather predictions, enabling airlines to anticipate and mitigate weather-related delays more effectively.

3. Optimizing Crew and Resource Management: Effective crew scheduling and resource allocation are crucial in minimizing delays. Implementing automated systems for crew management can ensure compliance with regulations, reduce human error, and enhance operational efficiency.

4. Enhancing Customer Communication: Although the data did not have any information about customer experience, delays can

cause significant dissatisfaction among customers. Transparent and timely communication with passengers can alleviate frustration during delays. Utilizing mobile applications and SMS notifications to provide real-time updates on flight status, gate changes, and estimated departure times can improve the overall passenger experience.

5. Fostering Collaborative Decision-Making: Collaboration between airlines, airports, and air traffic control is essential in managing delays effectively. Sharing real-time information and coordinating actions can help mitigate bottlenecks and reduce the impact of disruptions.

10. Conclusion

This project is a detailed analysis of US flight arrival data spanning ten years 2013-2023, focusing on understanding and predicting flight delays. The process began with data pre processing and cleaning. The cleaned dataset, originally in CSV format, was efficiently stored in a MySQL database.

The exploratory analysis was structured in three phases:

- Univariate analysis: provided insights into individual variables by examining their distributions, central tendencies, and outliers, enabling a foundational understanding of flight arrivals and delays.
- Bivariate analysis: explored relationships between pairs of variables, revealing key correlations.
- Multivariate analysis: got deeper into interactions among multiple variables, using visualizations like pairwise plots and heatmaps to uncover complex patterns influencing delays, cancellations, and airline performance.

A series of informative visualizations — including the distribution of flight cancellations, annual and monthly trends in cancellations and delays, and a pie chart illustrating delay causes helped to the understand the operational challenges. The analysis also highlighted the top 10 airlines and airports with the highest total arrival delays, providing a targeted view of carrier performance.

The predictive modeling phase employed linear regression techniques, with results validated by comparing actual versus predicted arrival delays, showcasing the model's ability to capture significant delay factors.

Finally, an interactive dashboard built using Power BI consolidated these findings into a user-friendly interface.

Overall, this comprehensive approach from pre-processing to multivariate exploration, predictive modeling, and dashboard visualization demonstrates the power of data analytics to identify operational inefficiencies and support strategic decisions aimed at reducing flight delays and improving airline performance.