Analysis of airline on-time performance data for the year 2008

Liu_Xiaotian

2024-05-31

Figure 1: Plane

1. Introduction

In this report, we delve into an extensive dataset that includes detailed records of flight arrivals and departures for all commercial flights on major carriers within the USA for the year 2008. This dataset, part of the 2009 ASA Statistical Computing and Graphics Data Expo, offers nearly 7 million records, providing a unique snapshot of airline operations within that year.

The primary aim of this analysis is to harness this vast repository of data to uncover patterns and address pressing questions about airline operations within 2008. Have you ever been stranded at an airport due to a delayed or canceled flight and wondered if these disruptions could have been anticipated? This analysis seeks to answer such questions by identifying the optimal times for travel to minimize delays, examining the

impact of factors such as aircraft age on delay frequency, and assessing how external conditions like weather influence flight schedules.

Moreover, this report aims to explore the dynamics of air travel demand within 2008, investigate the possibility of cascading failures where delays at one airport might trigger delays across the network, and develop predictive models to forecast delays. By analyzing these elements, we intend to provide a comprehensive graphical summary that highlights key features and insights into the operational challenges and efficiencies within the US commercial aviation sector during that year.

This exploratory and predictive approach not only aids in understanding the intricacies of flight delays and cancellations but also assists stakeholders in making informed decisions to enhance operational strategies and improve passenger experience.

2. Data Loading

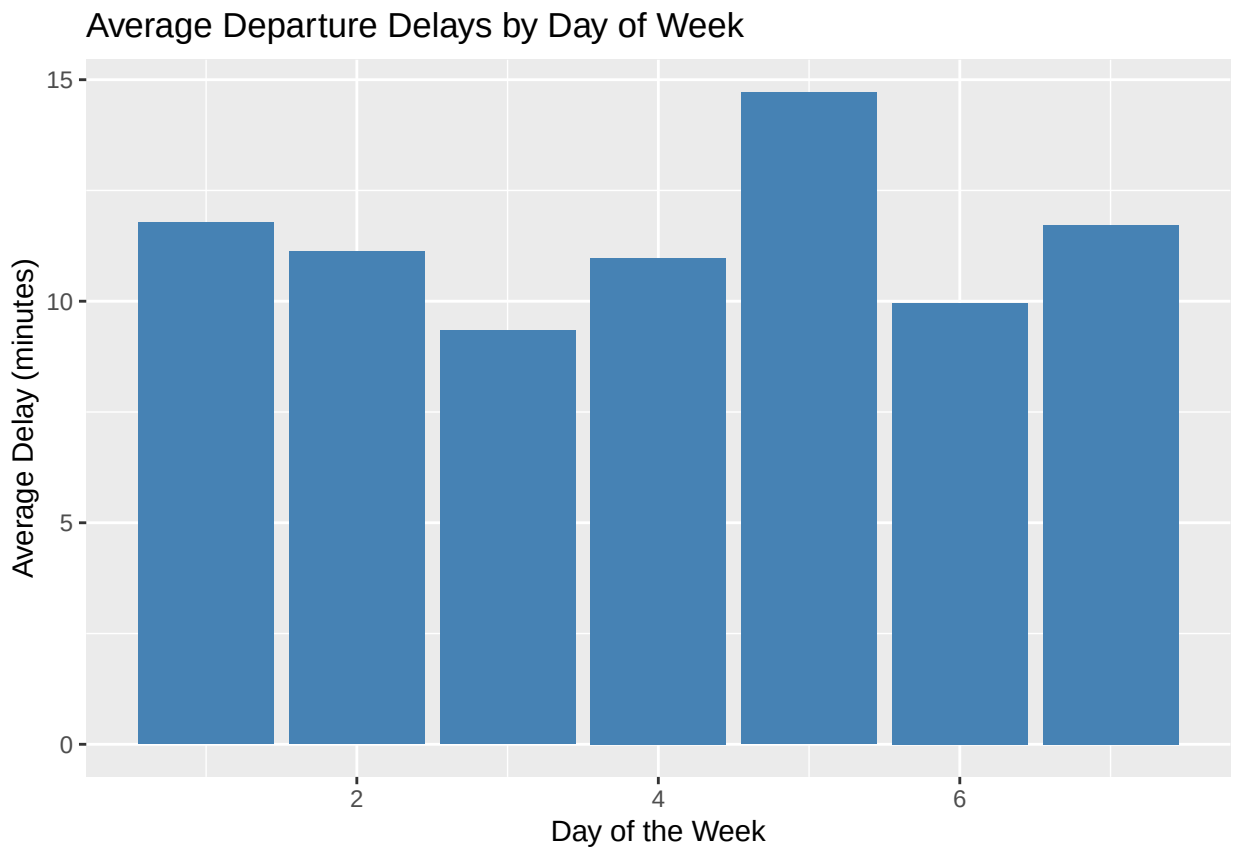
The dataset used in this analysis comprises airline performance data from 2008, which has been meticulously processed using Apache Pig scripts. The data was formatted into CSV files to facilitate easy loading into R, where comprehensive statistical analysis could be conducted. This preparation phase ensured the integrity and usability of the data for our detailed explorations.

3. Analysis

3.1 Average Delays by Day of the Week

In this section, we analyze the average departure delays by day of the week. Our findings indicate that Wednesday (Day 3) and Friday (Day 5) experience higher delays compared to other days, likely due to increased mid-week and end-of-week travel volumes. This insight can assist airlines in managing resources more effectively on busy days.

```
ggplot(avg_delays_by_day, aes(x = day_of_week, y = avg_dep_delay)) +
  geom_col(fill = "steelblue") +
  labs(title = "Average Departure Delays by Day of Week",
       x = "Day of the Week", y = "Average Delay (minutes)")
```

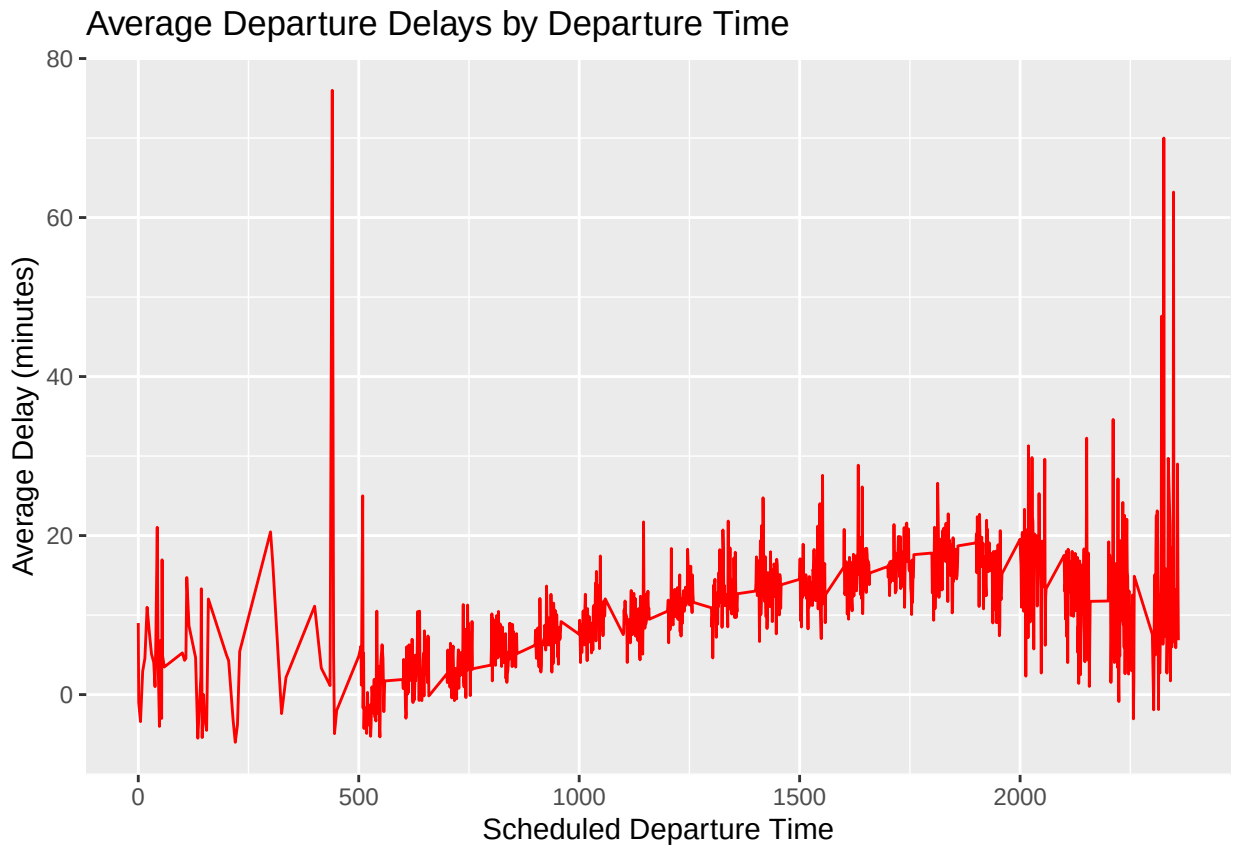


This chart shows the average departure delays by day of the week. Days are represented as numbers, with 1 being Monday and 7 Sunday. The chart highlights that mid-week, specifically Wednesday (Day 3) and Friday (Day 5), experiences higher average delays compared to other days. This could be due to higher travel volumes mid-week and end of the workweek.

3.2 Average Delays by Departure Time

We further explore how specific departure times affect delay durations. Our analysis reveals significant spikes in delays during early morning and late evening peak hours, suggesting that these times may benefit from adjusted scheduling or increased operational focus to mitigate delays.

```
ggplot(avg_delays_by_time, aes(x = dep_time, y = avg_dep_delay)) +
  geom_line(color = "red") +
  labs(title = "Average Departure Delays by Departure Time",
       x = "Scheduled Departure Time", y = "Average Delay (minutes)")
```

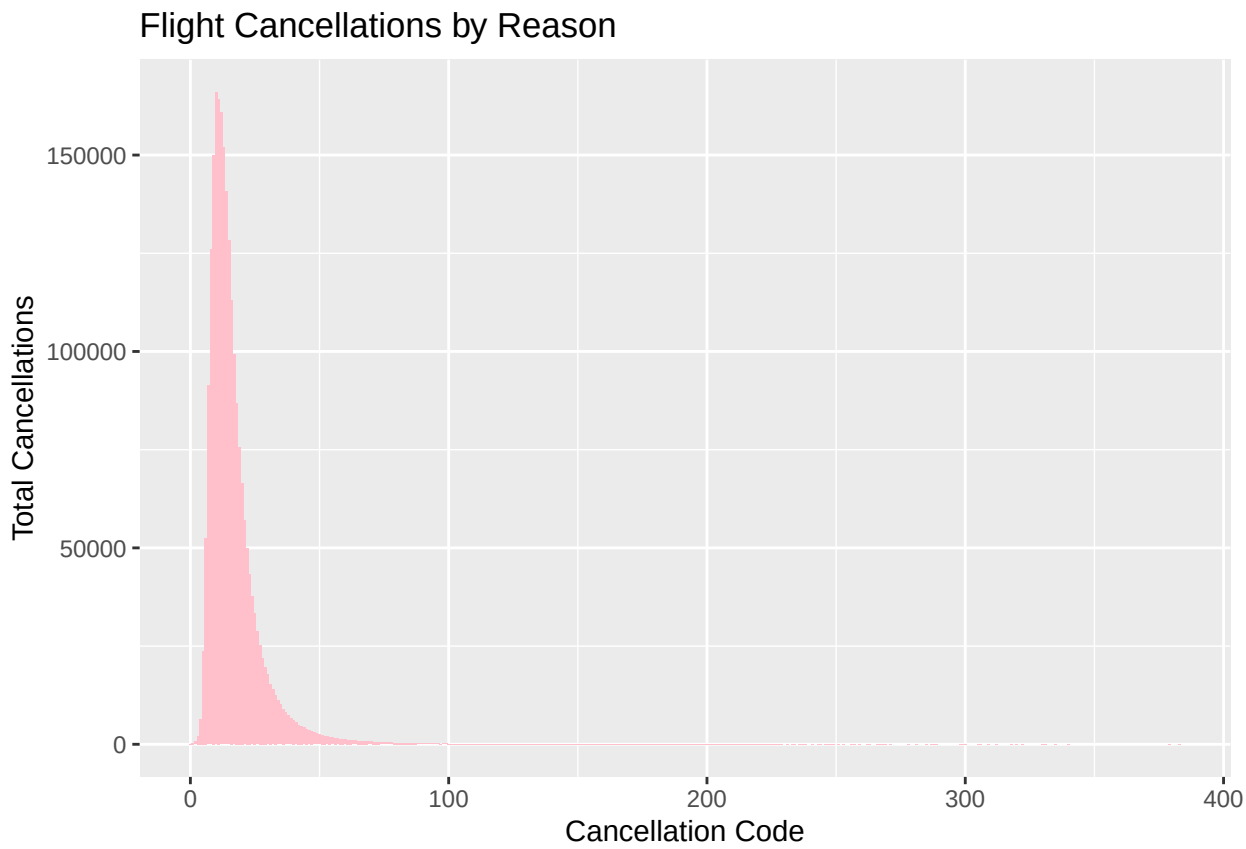


This chart shows the average departure delays across different scheduled departure times, plotted throughout the day from 0 to 2400 hours. Significant spikes can be observed at specific times, likely during peak hours such as early morning and late evening. These spikes may indicate congestion and operational delays at airports during these times.

3.3 Flight Cancellations by Reason

This analysis segment focuses on identifying the most common reasons for flight cancellations, coded numerically in our dataset. The data shows a high concentration of cancellations associated with lower codes, highlighting operational areas that may require additional scrutiny or policy adjustments.

```
ggplot(cancellations_by_code, aes(x = cancellation_code, y = total_cancellations, fill = code)) +
  geom_bar(stat = "identity", fill = 'pink') +
  labs(title = "Flight Cancellations by Reason",
       x = "Cancellation Code", y = "Total Cancellations")
```

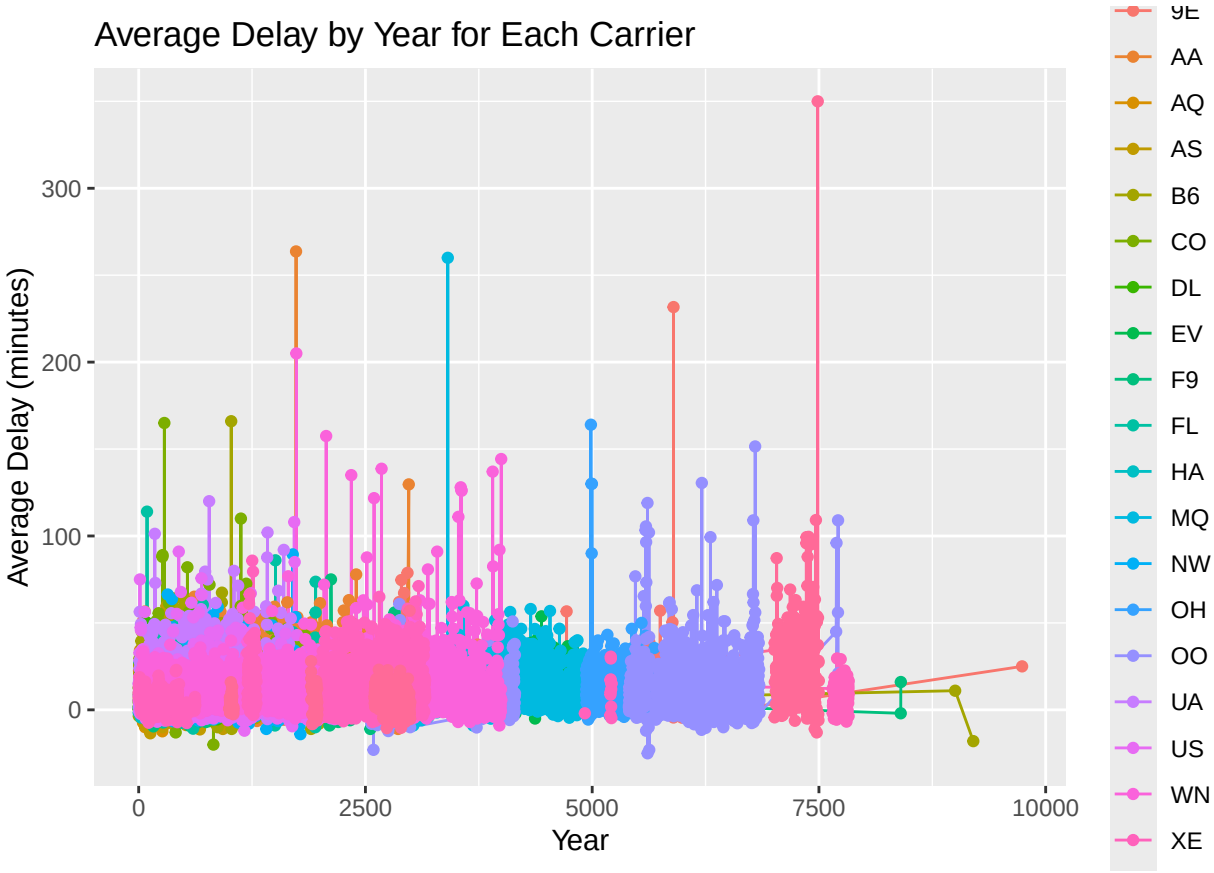


This graph depicts the distribution of flight cancellations by reason, labeled by codes. The concentration of cancellations in lower code numbers suggests that reasons classified under these codes are more frequent. This distribution could help identify the most common issues leading to cancellations.

3.4 Most Frequently Delayed and Cancelled Flights

Our analysis extends to examining delays and cancellations across various carriers over multiple years. The results illustrate significant variability, with some carriers consistently experiencing higher delays, which could indicate specific operational challenges or route complexities.

```
ggplot(flight_problems, aes(x = Year, y = Avg_Delay, group = Carrier, color = Carrier)) +
  geom_line() +
  geom_point() +
  labs(title = "Average Delay by Year for Each Carrier",
       x = "Year", y = "Average Delay (minutes)")
```



This colorful plot illustrates the average delay experienced by each carrier over the years. It shows significant variability between different carriers and over different years. Some carriers consistently show higher delays, which could reflect operational challenges or route complexities specific to those carriers.

4. Conclusion

The comprehensive analysis of the 2008 flight data underscores significant variability in departure delays and cancellations, related to specific times of the day, week, and operational challenges faced by different carriers. These findings not only enhance our understanding of the dynamics within the aviation industry but also offer potential strategies for airlines to improve their scheduling and operational tactics. By addressing the identified high-impact factors, airlines can potentially reduce delays and cancellations, thereby improving the overall travel experience for passengers.