Project Proposal for nycflights13 Dataset Analysis

Red Squadron

2025-02-07

Introduction

The objective of this project is to analyze the `nycflights13` dataset to gain insights into flight patterns, delays, and related phenomena. The overarching goal is to answer key questions about the factors influencing flight delays and performance. By the end of the project, we aim to provide actionable insights for stakeholders, including airlines, passengers, and airport authorities.

Research Questions and Objectives

We aim to answer the following research questions:

1. Which departure airports experience the longest delays, and what factors contribute to these delays?
2. How does the time of departure (morning, afternoon, evening) impact the likelihood of delays?
3. Are there significant differences in on-time performance between different airlines?
4. How does weather (e.g., temperature, precipitation) affect flight delays?
5. What is the relationship between flight distance and delay duration?
6. How do delays vary across different seasons and holidays?
7. Can predictive models help estimate delays based on factors such as flight schedule and carrier?

To achieve these goals, we have formulated the following specific research aims:

- Conduct exploratory data analysis to understand the distribution of delays.
- Identify key variables impacting flight performance.
- Perform time-series analysis to detect patterns over months and days.
- Develop predictive models to forecast delays.

Dataset Overview

The primary dataset for this analysis is `nycflights13`. It contains detailed information on all flights departing from New York City airports in 2013.

```
library(nycflights13)
## Warning: package 'nycflights13' was built under R version 4.4.2
library(dplyr)
## Warning: package 'dplyr' was built under R version 4.4.2
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
# Summary of the flights dataset
flights_summary <- flights %>% summarise(rows = n(), columns = ncol(flights))
print(flights_summary)
## # A tibble: 1 x 2
##     rows columns
##    <int>   <int>
## 1 336776      19
```

The dataset contains 336776 rows and 19 columns. Key variables include:

- `year`, `month`, `day`: Date of the flight.
- `dep_time`, `arr_time`: Departure and arrival times.
- `carrier`: Airline carrier code.
- `flight`: Flight number.
- `origin`, `dest`: Origin and destination airports.
- `air_time`, `distance`: Flight duration and distance.
- `dep_delay`, `arr_delay`: Departure and arrival delays.

Exploratory Data Analysis

We performed initial exploratory analysis to understand the distribution and relationships between key variables.

Summary Statistics

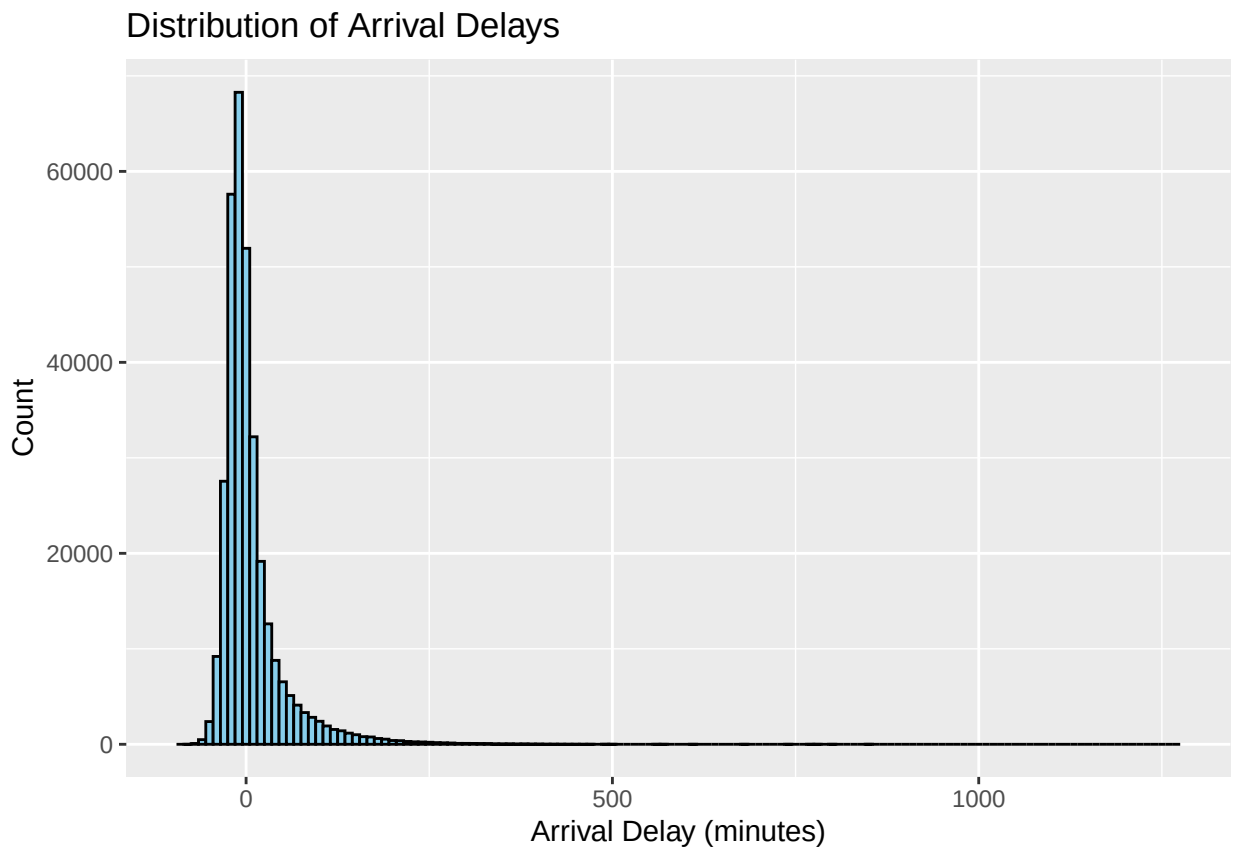
```
summary(flights)
##       year          month             day           dep_time    sched_dep_time
##  Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   :   1   Min.   : 106
##  1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907   1st Qu.: 906
##  Median :2013   Median : 7.000   Median :16.00   Median :1401   Median :1359
##  Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349   Mean   :1344
##  3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744   3rd Qu.:1729
##  Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400   Max.   :2359
##                                                  NA's   :8255
##    dep_delay          arr_time    sched_arr_time   arr_delay
##  Min.   : -43.00   Min.   :   1   Min.   :   1   Min.   : -86.000
##  1st Qu.:  -5.00   1st Qu.:1104   1st Qu.:1124   1st Qu.: -17.000
##  Median :  -2.00   Median :1535   Median :1556   Median :  -5.000
##  Mean   :  12.64   Mean   :1502   Mean   :1536   Mean   :   6.895
##  3rd Qu.:  11.00   3rd Qu.:1940   3rd Qu.:1945   3rd Qu.:  14.000
##  Max.   :1301.00   Max.   :2400   Max.   :2359   Max.   :1272.000
##  NA's   :8255      NA's   :8713                  NA's   :9430
##    carrier              flight       tailnum             origin
##  Length:336776      Min.   :   1   Length:336776      Length:336776
```

```
##  Class :character   1st Qu.: 553   Class :character   Class :character
##  Mode  :character   Median :1496   Mode  :character   Mode  :character
##                     Mean   :1972
##                     3rd Qu.:3465
##                     Max.   :8500
##
##      dest              air_time        distance         hour
##  Length:336776      Min.   : 20.0   Min.   :  17   Min.   : 1.00
##  Class :character   1st Qu.: 82.0   1st Qu.: 502   1st Qu.: 9.00
##  Mode  :character   Median :129.0   Median : 872   Median :13.00
##                     Mean   :150.7   Mean   :1040   Mean   :13.18
##                     3rd Qu.:192.0   3rd Qu.:1389   3rd Qu.:17.00
##                     Max.   :695.0   Max.   :4983   Max.   :23.00
##                     NA's   :9430
##      minute        time_hour
##  Min.   : 0.00   Min.   :2013-01-01 05:00:00.00
##  1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00.00
##  Median :29.00   Median :2013-07-03 10:00:00.00
##  Mean   :26.23   Mean   :2013-07-03 05:22:54.64
##  3rd Qu.:44.00   3rd Qu.:2013-10-01 07:00:00.00
##  Max.   :59.00   Max.   :2013-12-31 23:00:00.00
##
```

Distribution of Arrival Delays

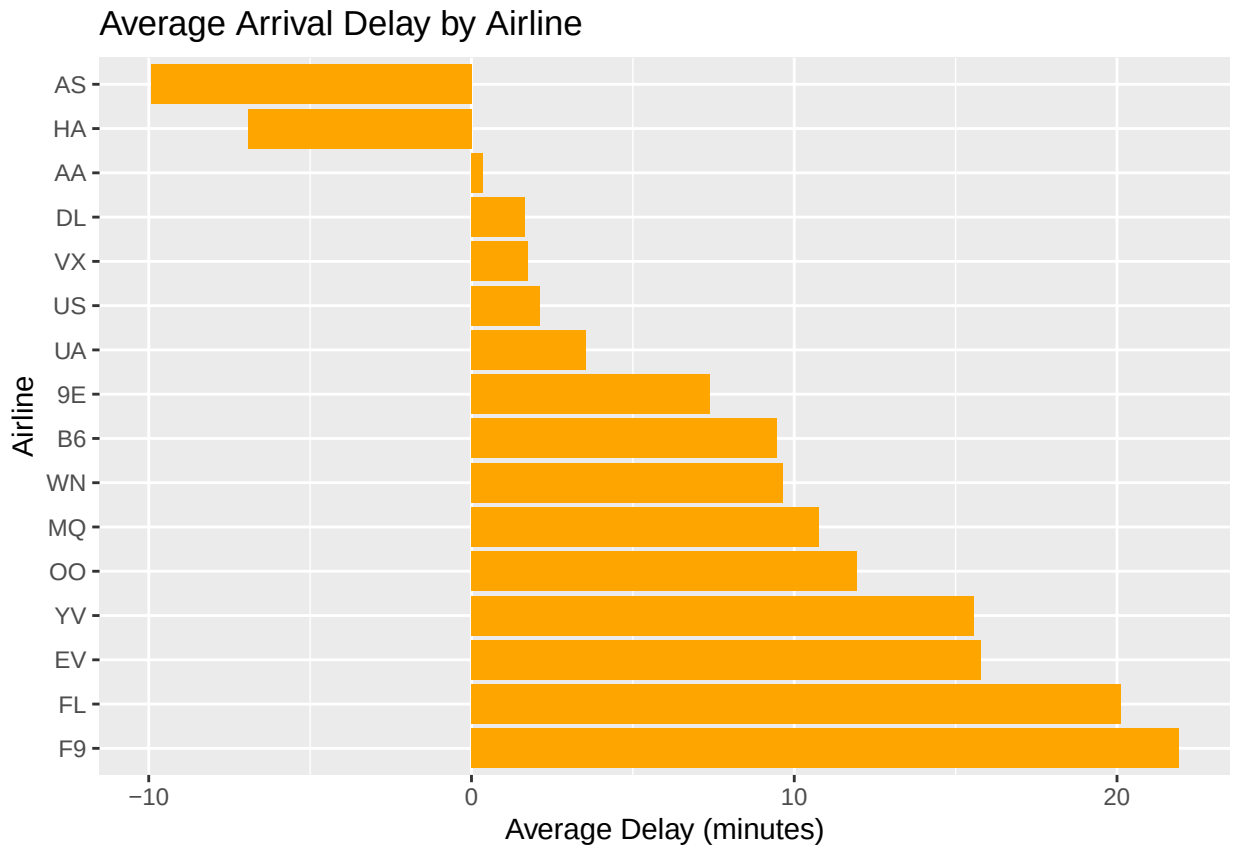
```
library(ggplot2)
## Warning: package 'ggplot2' was built under R version 4.4.2

# Plotting the distribution of arrival delays
flights %>%
  filter(!is.na(arr_delay)) %>%
  ggplot(aes(x = arr_delay)) +
  geom_histogram(binwidth = 10, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Arrival Delays", x = "Arrival Delay (minutes)", y = "Count")
```



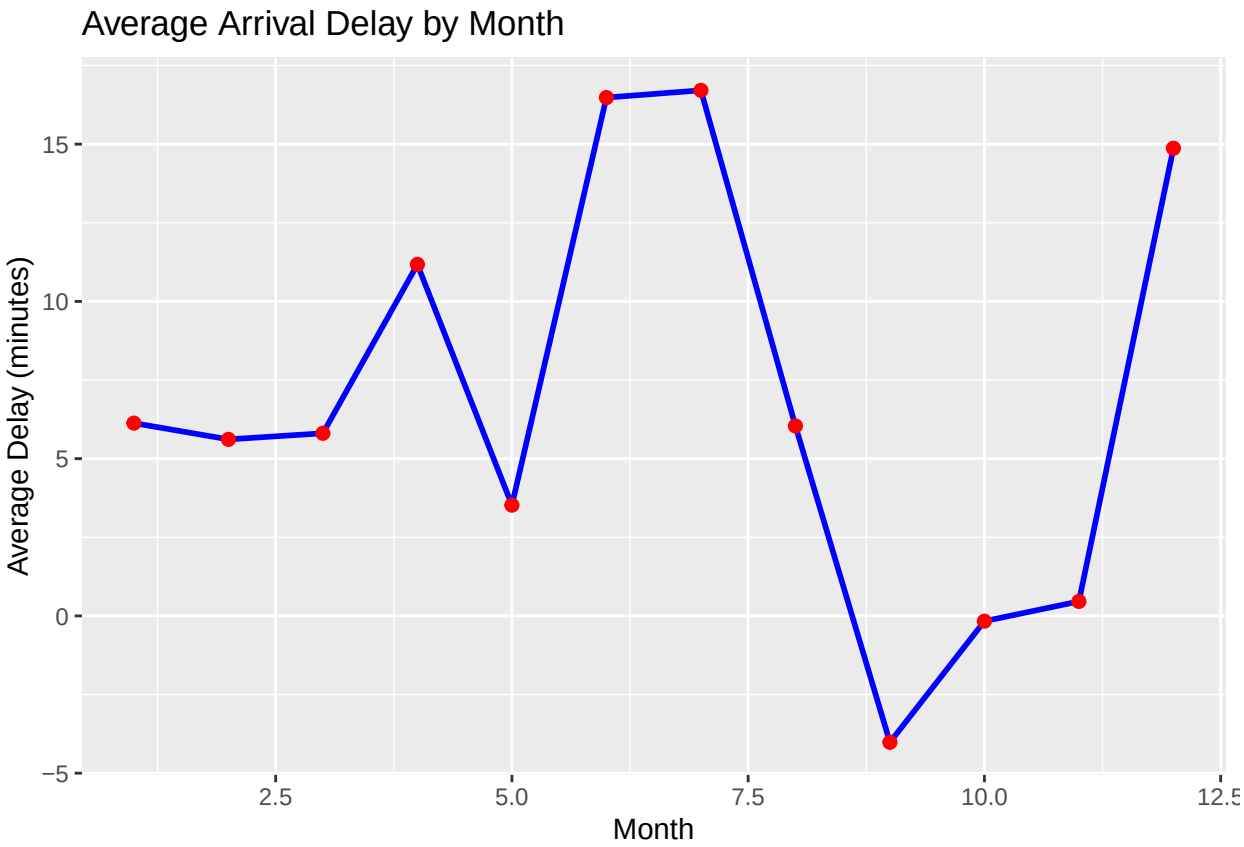
Comparison of Delays by Airline

```r
# Bar plot showing average arrival delay by airline
flights %>%
  filter(!is.na(arr_delay)) %>%
  group_by(carrier) %>%
  summarise(mean_arr_delay = mean(arr_delay, na.rm = TRUE)) %>%
  ggplot(aes(x = reorder(carrier, -mean_arr_delay), y = mean_arr_delay)) +
  geom_bar(stat = "identity", fill = "orange") +
  labs(title = "Average Arrival Delay by Airline", x = "Airline", y = "Average Delay (minutes)") +
  coord_flip()
```



Delays by Month

```
# Time series plot of average delays by month
flights %>%
  filter(!is.na(arr_delay)) %>%
  group_by(month) %>%
  summarise(avg_arr_delay = mean(arr_delay, na.rm = TRUE)) %>%
  ggplot(aes(x = month, y = avg_arr_delay)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "red", size = 2) +
  labs(title = "Average Arrival Delay by Month", x = "Month", y = "Average Delay (minutes)")
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Data Cleaning

The dataset contains some missing and erroneous values, particularly in columns related to delays and times. We plan to clean the data by:

- Removing rows with missing `dep_delay` or `arr_delay` values.
- Filtering out flights with extreme outliers in delay times.

Methodology

To answer the research questions, we will employ the following methods:

1. **Descriptive Analysis:** Summary statistics and visualizations to understand trends and distributions.
2. **Correlation Analysis:** Identifying relationships between variables such as distance, delay, and carrier.
3. **Time Series Analysis:** Examining seasonal and daily patterns in flight delays.
4. **Predictive Modeling:**

- We plan to build regression and classification models to predict delays.
- Models may include linear regression, decision trees, and random forests.

Example Correlation Analysis

```r
# Calculate correlation between distance and arrival delay
correlation <- cor(flights$distance, flights$arr_delay, use = "complete.obs")
cat("Correlation between flight distance and arrival delay: ", correlation)
## Correlation between flight distance and arrival delay:  -0.06186776
```

Potential Visualizations

- Scatter plots of delays vs. flight distance.
- Time series plots showing average delays by month.
- Bar plots comparing delays by carrier and destination.

Alternative Strategies / Backup Plans

If our initial analysis does not yield conclusive results, we will consider the following alternative strategies:

1. **Focus on Weather Data:** Integrate external weather data to explore its impact on delays.

- Collect historical weather data for NYC in 2013.
- Analyze how weather conditions (e.g., storms, visibility) affect delays.

2. **Airline Performance Comparison:**

- Concentrate on ranking airlines based on delay metrics.
- Perform detailed case studies on the top-performing and worst-performing airlines.

3. **Airport-Specific Analysis:**

- Analyze each NYC airport separately to identify unique patterns.
- Examine how traffic congestion and infrastructure influence delays.

Expected Outcomes

We expect to uncover significant patterns in flight delays that could inform decision-making for airlines and airports. Predictive models should help stakeholders anticipate delays and improve operational efficiency.

Conclusion

This project will provide a comprehensive analysis of the `nycflights13` dataset. By answering the research questions through a combination of exploratory data analysis, visualization, and modeling, we aim to generate valuable insights into flight performance and delays. The findings will be shared in the final project report and presentation.

```r
# Final steps: Displaying key results summary
cat("Proposal complete. Next steps involve deeper analysis and model implementation.")
## Proposal complete. Next steps involve deeper analysis and model implementation.
```