Flight Delay Analysis

Yaritza Rendon

2025-02-26

Jetblue Flight Delay Analysis Using R

This project analyzes Jetblue flight delays from November 2023 to November 2024. The goal is to clean the data set, perform an exploratory data analysis, and visualize delay trends to identify the main causes of delays. By understanding these causes, we can provide insights that may help avoid future delays.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
knitr::opts_chunk$set(echo = TRUE)
```

Step 1: Data Cleaning

Load the data set.

```
# Load necessary libraries
library(tidyverse)
# Import the csv file
df <- read.csv("Jetblue_Delay_Cause_Nov2023-Nov2024.csv")
```

Inspect the data.

```
# View structure and summary
str(df)
```

```
## 'data.frame':    13 obs. of  21 variables:
##  $ year               : int  2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 ...
##  $ month              : int  11 10 9 8 7 6 5 4 3 2 ...
##  $ carrier            : chr  "B6" "B6" "B6" "B6" ...
##  $ carrier_name       : chr  "JetBlue Airways" "JetBlue Airways" "JetBlue Airways" "JetBlue Airways"
##  $ airport            : chr  "JFK" "JFK" "JFK" "JFK" ...
##  $ airport_name       : chr  "New York, NY: John F. Kennedy International" "New York, NY: John F. Ken
##  $ arr_flights        : num  2884 2981 2890 3218 3198 ...
##  $ arr_del15          : num  437 427 513 1067 1016 ...
##  $ carrier_ct         : num  202 204 225 316 306 ...
##  $ weather_ct         : num  1.7 2.39 4.49 25.18 18.31 ...
##  $ nas_ct             : num  100.6 95.4 120.8 262.2 285.4 ...
##  $ security_ct        : num  1.83 0.87 1.46 1.99 0.08 1.52 1.67 1 0 1.6 ...
##  $ late_aircraft_ct   : num  131 124 162 462 406 ...
##  $ arr_cancelled      : num  0 44 13 216 73 112 30 51 60 57 ...
##  $ arr_diverted       : num  3 1 7 18 17 27 13 11 11 5 ...
##  $ arr_delay          : num  28765 27781 39411 113727 94990 ...
##  $ carrier_delay      : num  14724 15806 21791 42530 35325 ...
##  $ weather_delay      : num  56 252 251 2120 1522 ...
##  $ nas_delay          : num  3596 2950 4224 21859 19613 ...
##  $ security_delay     : num  29 76 56 131 5 76 53 88 0 38 ...
##  $ late_aircraft_delay: num  10360 8697 13089 47087 38525 ...
```

```r
summary(df)
```

```
##       year          month          carrier          carrier_name
##  Min.   :2023   Min.   : 1.000   Length:13          Length:13
##  1st Qu.:2024   1st Qu.: 4.000   Class :character   Class :character
##  Median :2024   Median : 7.000   Mode  :character   Mode  :character
##  Mean   :2024   Mean   : 6.846
##  3rd Qu.:2024   3rd Qu.:10.000
##  Max.   :2024   Max.   :12.000
##    airport          airport_name        arr_flights     arr_del15
##  Length:13          Length:13          Min.   :2675   Min.   : 427.0
##  Class :character   Class :character   1st Qu.:2890   1st Qu.: 660.0
##  Mode  :character   Mode  :character   Median :3074   Median : 790.0
##                                        Mean   :3033   Mean   : 768.2
##                                        3rd Qu.:3198   3rd Qu.: 963.0
##                                        Max.   :3277   Max.   :1067.0
##    carrier_ct      weather_ct         nas_ct       security_ct
##  Min.   :202.3   Min.   : 1.700   Min.   : 95.44   Min.   :0.000
##  1st Qu.:224.6   1st Qu.: 4.490   1st Qu.:142.16   1st Qu.:0.870
##  Median :269.3   Median : 6.460   Median :187.99   Median :1.460
##  Mean   :266.3   Mean   : 9.143   Mean   :195.85   Mean   :1.245
##  3rd Qu.:305.8   3rd Qu.:11.780   3rd Qu.:262.17   3rd Qu.:1.670
##  Max.   :344.6   Max.   :25.180   Max.   :301.18   Max.   :3.220
##  late_aircraft_ct arr_cancelled     arr_diverted     arr_delay
##  Min.   :124.4    Min.   :  0.00   Min.   : 1.00   Min.   : 27781
##  1st Qu.:276.8    1st Qu.: 14.00   1st Qu.: 5.00   1st Qu.: 53042
##  Median :297.0    Median : 48.00   Median : 9.00   Median : 64287
##  Mean   :295.7    Mean   : 55.69   Mean   :10.15   Mean   : 67231
##  3rd Qu.:339.2    3rd Qu.: 60.00   3rd Qu.:13.00   3rd Qu.: 92800
##  Max.   :462.1    Max.   :216.00   Max.   :27.00   Max.   :113727
##  carrier_delay   weather_delay      nas_delay     security_delay
##  Min.   :14724   Min.   :  56.0   Min.   : 2950   Min.   :  0.00
##  1st Qu.:21791   1st Qu.: 252.0   1st Qu.: 5343   1st Qu.: 29.00
##  Median :24736   Median : 609.0   Median : 9543   Median : 53.00
##  Mean   :26862   Mean   : 964.4   Mean   :11355   Mean   : 50.85
##  3rd Qu.:32829   3rd Qu.:1522.0   3rd Qu.:14256   3rd Qu.: 76.00
##  Max.   :42530   Max.   :2181.0   Max.   :23776   Max.   :131.00
##  late_aircraft_delay
##  Min.   : 8697
##  1st Qu.:22510
##  Median :28702
##  Mean   :27999
##  3rd Qu.:33503
##  Max.   :50337
```

Clean the data.

```
# Check for missing values
colSums(is.na(df))
```

```
##                year               month             carrier        carrier_name
##                   0                   0                   0                   0
##             airport        airport_name         arr_flights           arr_del15
##                   0                   0                   0                   0
##          carrier_ct          weather_ct              nas_ct         security_ct
##                   0                   0                   0                   0
##    late_aircraft_ct       arr_cancelled        arr_diverted           arr_delay
##                   0                   0                   0                   0
##       carrier_delay       weather_delay           nas_delay      security_delay
##                   0                   0                   0                   0
## late_aircraft_delay
##                   0
```

```
# Fills or removes rows with NA values if necessary
df <- df %>% drop_na()
```

Convert data types

```
# Convert month and year to a date format
df <- df %>% mutate(date = as.Date(paste(year, month, "01", sep = "-")))
```

Step 2: Exploratory Data Analysis

Summary Statistics

```
# Summary of key delay-related columns
summary(df %>% select(arr_del15, carrier_ct, weather_ct, nas_ct, security_ct, late_aircraft_ct))
```

```
##    arr_del15        carrier_ct      weather_ct         nas_ct      
##  Min.   : 427.0   Min.   :202.3   Min.   : 1.700   Min.   : 95.44  
##  1st Qu.: 660.0   1st Qu.:224.6   1st Qu.: 4.490   1st Qu.:142.16  
##  Median : 790.0   Median :269.3   Median : 6.460   Median :187.99  
##  Mean   : 768.2   Mean   :266.3   Mean   : 9.143   Mean   :195.85  
##  3rd Qu.: 963.0   3rd Qu.:305.8   3rd Qu.:11.780   3rd Qu.:262.17  
##  Max.   :1067.0   Max.   :344.6   Max.   :25.180   Max.   :301.18  
##   security_ct    late_aircraft_ct
##  Min.   :0.000   Min.   :124.4   
##  1st Qu.:0.870   1st Qu.:276.8   
##  Median :1.460   Median :297.0   
##  Mean   :1.245   Mean   :295.7   
##  3rd Qu.:1.670   3rd Qu.:339.2   
##  Max.   :3.220   Max.   :462.1   
```

Identify top delay causes

```
# Calculate total delays by category
df %>% summarise(
  Carrier = sum(carrier_ct),
  Weather = sum(weather_ct),
  NAS = sum(nas_ct),
  Security = sum(security_ct),
  Late_Aircraft = sum(late_aircraft_ct)
) %>% pivot_longer(everything(), names_to = "Cause", values_to = "Total_Delays") %>%
  arrange(desc(Total_Delays))
```

```
## # A tibble: 5 x 2
##   Cause         Total_Delays
##   <chr>                <dbl>
## 1 Late_Aircraft       3844.
## 2 Carrier             3462.
## 3 NAS                 2546.
## 4 Weather              119.
## 5 Security              16.2
```

Correlation between delay causes

```
# Compute correlation between delay causes
cor_matrix <- cor(df %>% select(carrier_ct, weather_ct, nas_ct, security_ct, late_aircraft_ct))
cor_matrix
```

```
##                  carrier_ct  weather_ct     nas_ct security_ct late_aircraft_ct
## carrier_ct       1.00000000  0.56570678  0.8338186  0.09390217        0.7660426
## weather_ct       0.56570678  1.00000000  0.7310166 -0.05551154        0.7441039
## nas_ct           0.83381862  0.73101663  1.0000000 -0.24050495        0.8520618
## security_ct      0.09390217 -0.05551154 -0.2405049  1.00000000       -0.2155053
## late_aircraft_ct 0.76604264  0.74410393  0.8520618 -0.21550531        1.0000000
```

Step 3: Data Visualization

Install and run packages for data visualization.

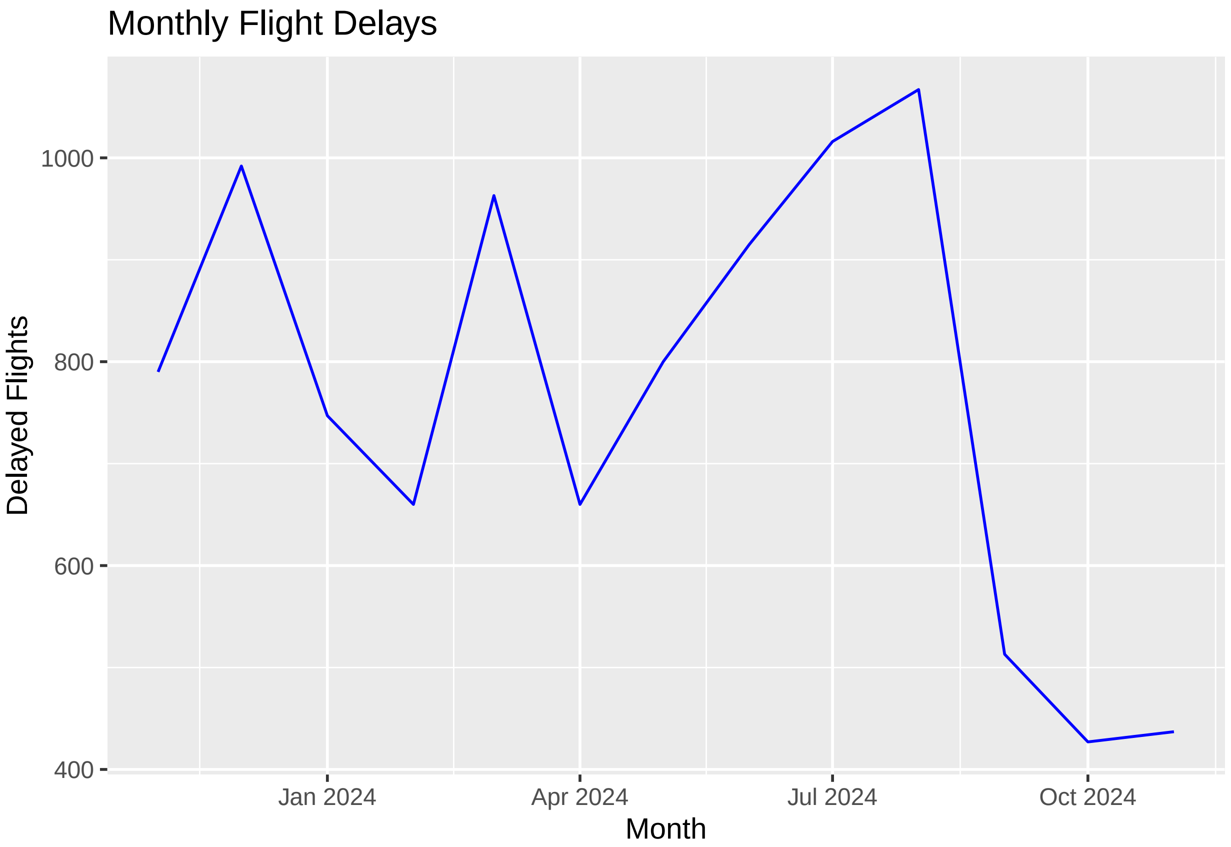
```
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(ggplot2)
```

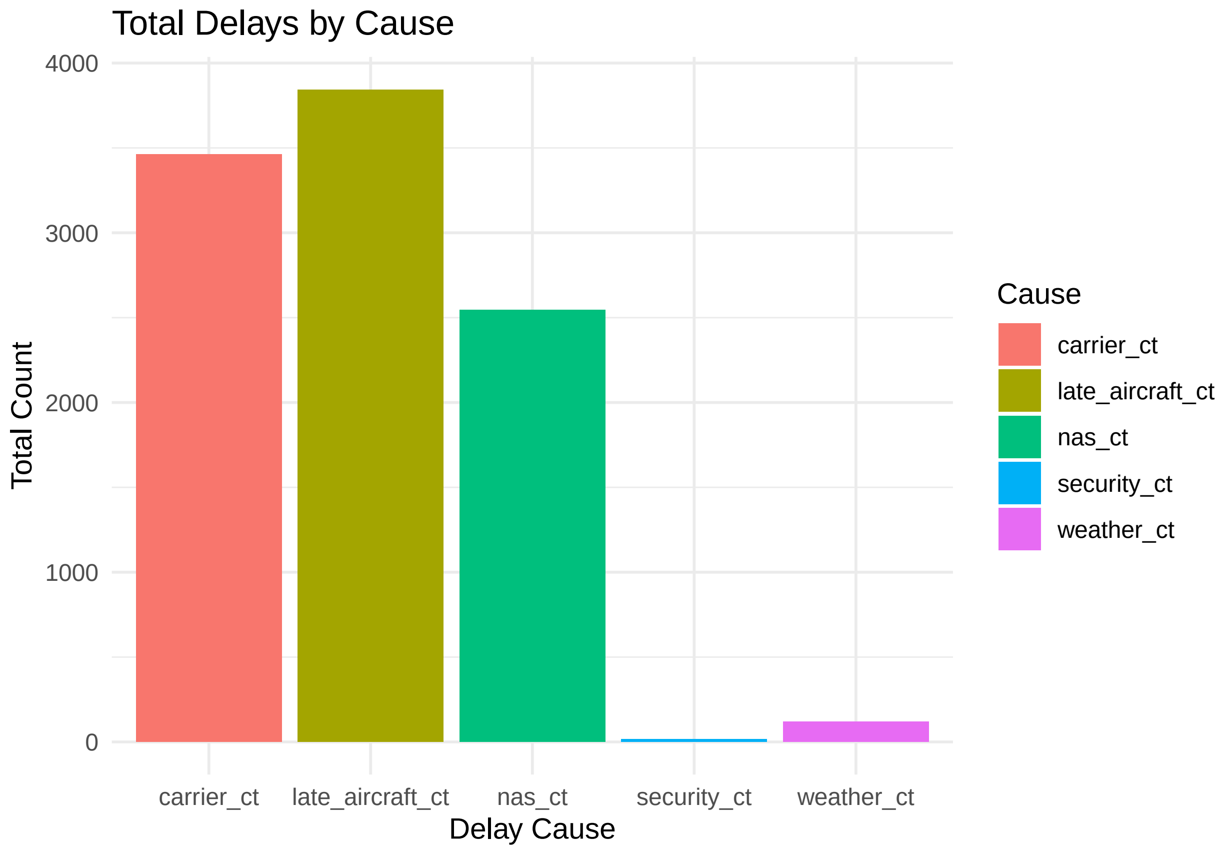
Let's visualize some data.

```
ggplot(df, aes(x = date, y = arr_del15)) +
  geom_line(color = "blue") +
  labs(title = "Monthly Flight Delays", x = "Month", y = "Delayed Flights") # line chart of delays over
```



Let’s do one more.

```
ggplot(df %>% pivot_longer(cols = carrier_ct:late_aircraft_ct, names_to = "Cause", values_to = "Count")
       aes(x = Cause, y = Count, fill = Cause)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Total Delays by Cause", x = "Delay Cause", y = "Total Count") # bar chart of total dela
```



```r
install.packages("reshape2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
library(reshape2)
```

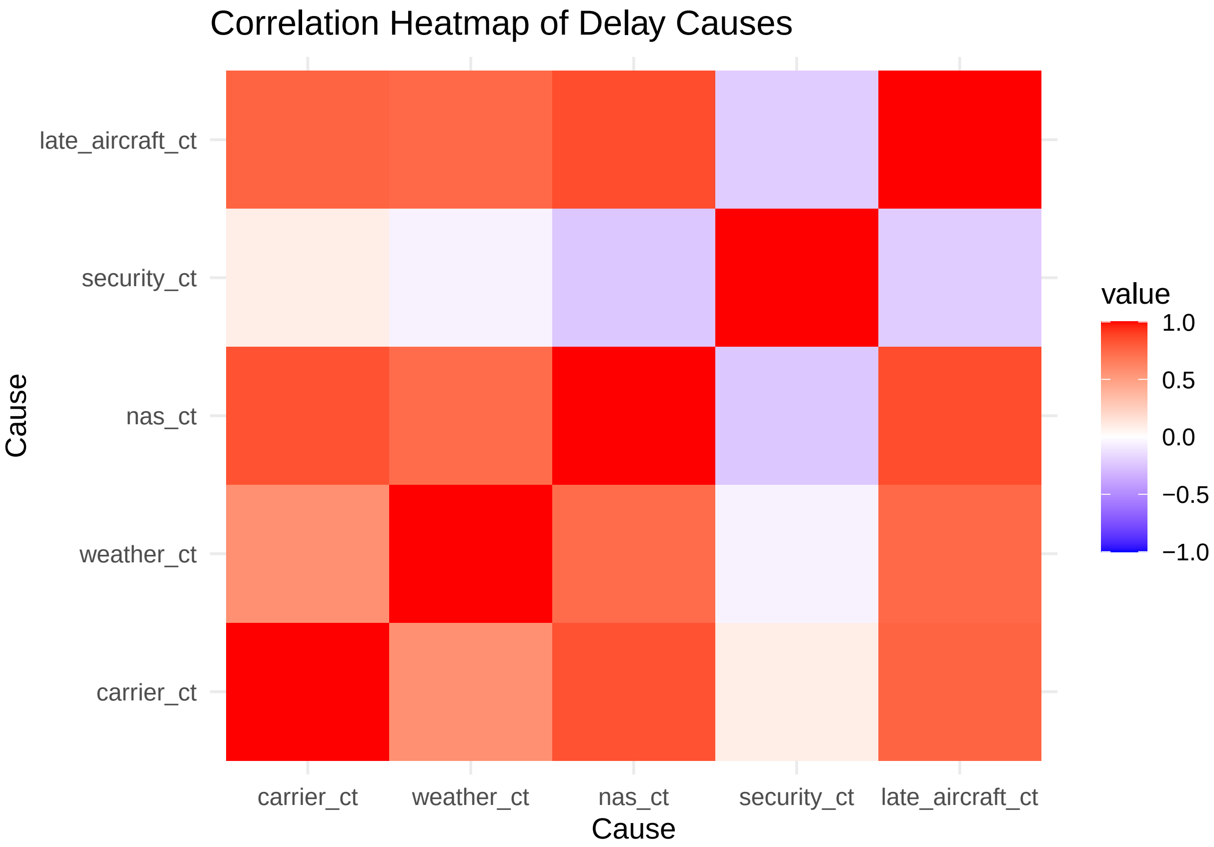
```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```r
library(ggplot2)
```

```r
# Convert correlation matrix to long format for heatmap
cor_melt <- melt(cor_matrix)

ggplot(cor_melt, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1,1)) +
  theme_minimal() +
  labs(title = "Correlation Heatmap of Delay Causes", x = "Cause", y = "Cause")
```



#

Conclusion

This analysis provides some insights into the major causes of JetBlue flight delays. These findings suggest that late aircraft and carrier delays are the most significant contributors to overall delays. The correlation analysis helps identify relationships between different types of delays, which can be useful for operational planning.

Recommendations:

Investigate Carrier Delays – Further analysis can help identify the root causes of operational delays within JetBlue.

Monitor Late Aircraft Impact – Since late aircraft delays trickle down throughout the day, improved scheduling could help minimize delays.

Weather-Based Predictions – Integrating weather data might help anticipate and mitigate weather-related disruptions.