# Flight Delays: An Analysis of Factors Contributing to Delays

MET CS 555 A3

Yuhan Xu
2023-04-15

# Research Scenario and Questions

- **Scenario:** Investigating the factors that contribute to flight delays
- **Research Questions:**
- 1. What is the relationship between flight length and the likelihood of delay?
- 2. Are certain days of the week more prone to flight delays?
- 3. Are there specific airlines or airport routes with a higher probability of flight delays?
- 4. Does the time of day affect the likelihood of flight delays?
- 5. Are there significant differences in the average flight delay times between different days of the week?
- 6. Is there a significant difference in the variability of flight delay times between short-haul (length <= 150) and long-haul flights (length > 150)?

# Library Use

- dplyr
- ggplot2
- tidyr

# Data Set Description

- Link to data set:

https://raw.githubusercontent.com/datasets/openml-datasets/master/data/airlines/airlines.csv

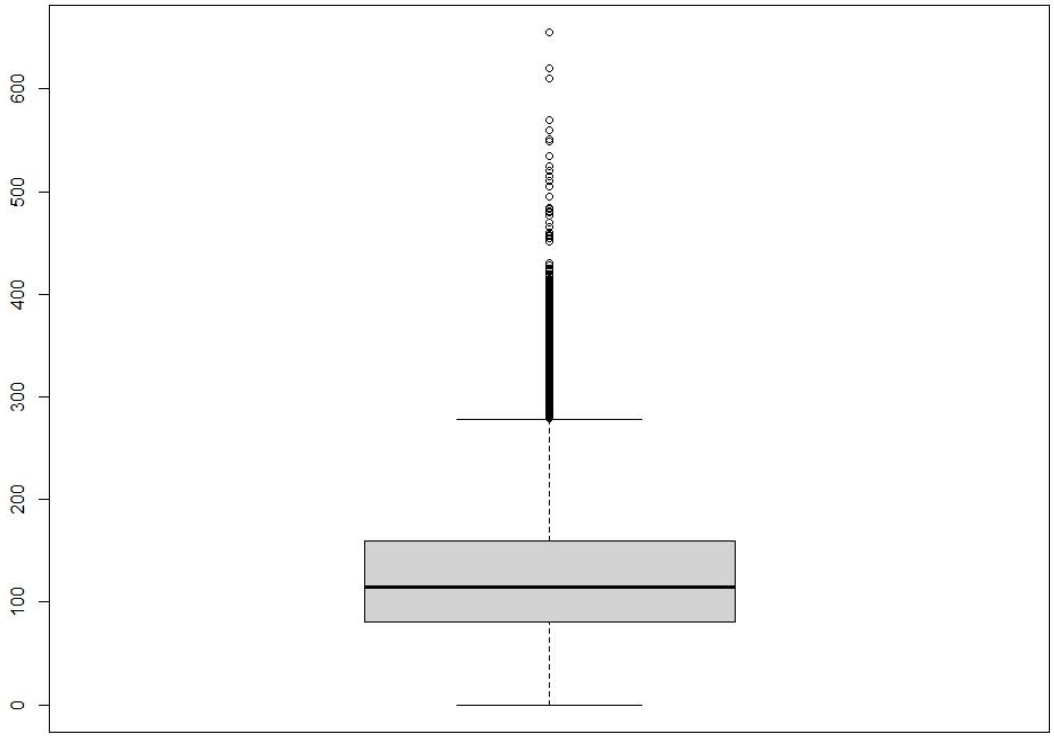
- Variables

**Airline, Flight, AirportFrom, AirportTo, DayOfWeek, Time, Length, Delay**

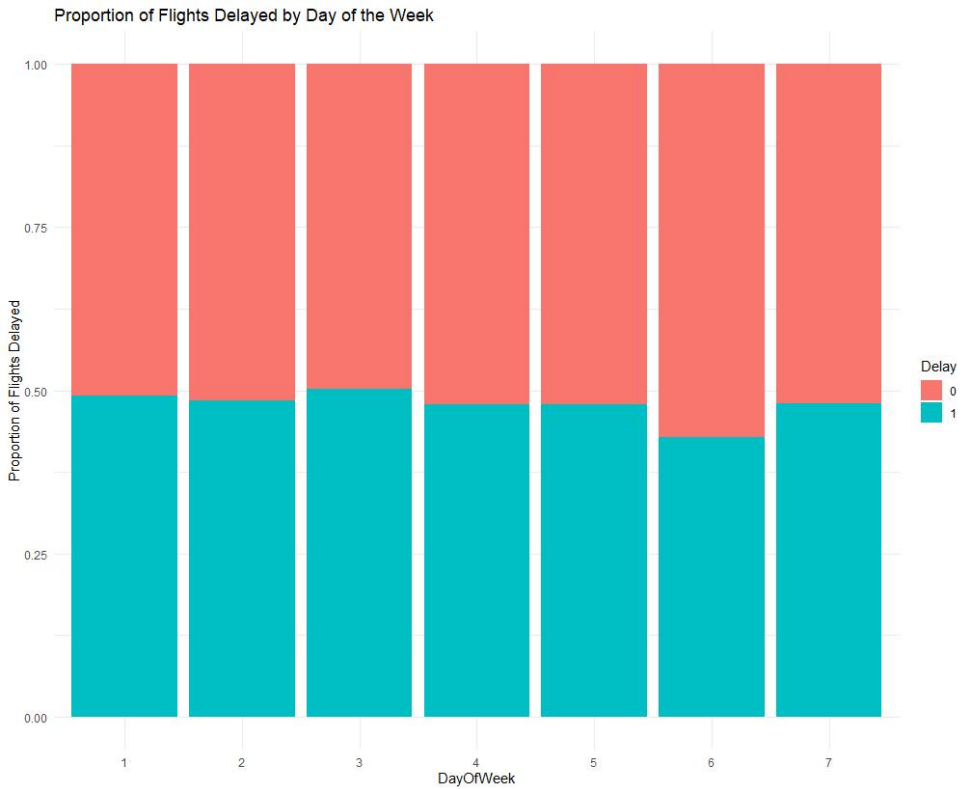
# Statistical Methods

- Descriptive statistics : including Airline, Flight, AirportFrom, AirportTo, DayOfWeek, Time, Length, and Delay
- Two-way ANOVA
- Post-hoc tests
- F - test

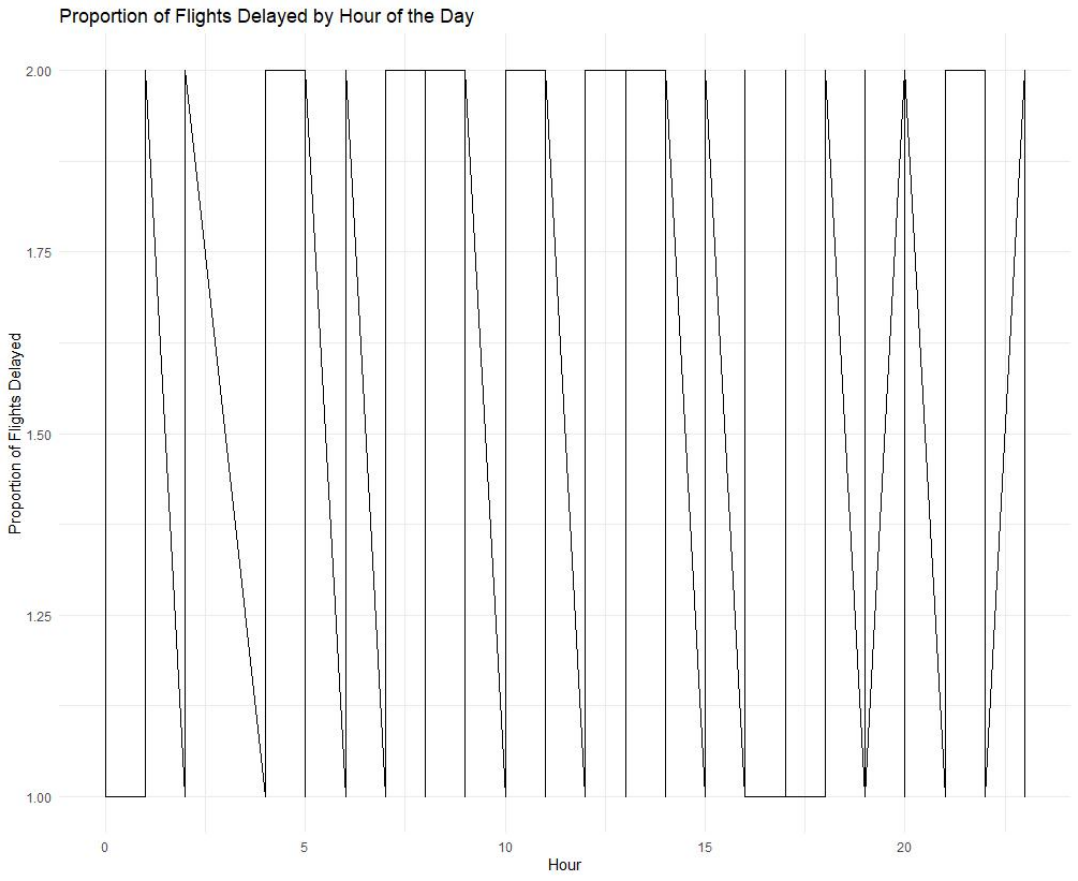
# Boxplot for checking outliers



# Bar chart of delays by day of the week



# Line graph of delays by hour of the day



# Two ways ANOVA

```
                   Df Sum Sq Mean Sq  F value   Pr(>F)
DayOfWeek           6    130    21.7   91.341  < 2e-16 ***
Hour                1    732   731.8 3087.286  < 2e-16 ***
Length              1     44    43.9  185.106  < 2e-16 ***
Airline            17   3132   184.3  777.274  < 2e-16 ***
DayOfWeek:Hour      6      7     1.1    4.781 6.97e-05 ***
Residuals      322733  76504     0.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

# Post-hoc tests

```
> summary(posthoc, adjust = "tukey")
Note: adjust = "tukey" was changed to "sidak"
because "tukey" is only appropriate for one set of pairwise comparisons
$emmeans
 DayOfWeek emmean       SE     df lower.CL upper.CL
 1          1.455 0.002418 322733    1.449    1.462
 2          1.445 0.002442 322733    1.439    1.452
 3          1.465 0.002357 322733    1.459    1.471
 4          1.442 0.002344 322733    1.436    1.449
 5          1.434 0.002313 322733    1.428    1.440
 6          1.403 0.002654 322733    1.396    1.410
 7          1.441 0.002442 322733    1.435    1.448

Results are averaged over the levels of: Airline
Results are given on the as.numeric (not the response) scale.
Confidence level used: 0.95
Conf-level adjustment: sidak method for 7 estimates

$contrasts
 contrast                estimate      SE     df t.ratio p.value
 DayOfWeek1 - DayOfWeek2  0.01006 0.00323 322733   3.112  0.0306
 DayOfWeek1 - DayOfWeek3 -0.00977 0.00317 322733  -3.084  0.0335
 DayOfWeek1 - DayOfWeek4  0.01287 0.00316 322733   4.075  0.0009
 DayOfWeek1 - DayOfWeek5  0.02128 0.00313 322733   6.803  <.0001
 DayOfWeek1 - DayOfWeek6  0.05252 0.00341 322733  15.421  <.0001
 DayOfWeek1 - DayOfWeek7  0.01396 0.00323 322733   4.316  0.0003
 DayOfWeek2 - DayOfWeek3 -0.01982 0.00318 322733  -6.230  <.0001
 DayOfWeek2 - DayOfWeek4  0.00281 0.00317 322733   0.887  0.9747
 DayOfWeek2 - DayOfWeek5  0.01123 0.00314 322733   3.573  0.0065
 DayOfWeek2 - DayOfWeek6  0.04246 0.00342 322733  12.416  <.0001
 DayOfWeek2 - DayOfWeek7  0.00390 0.00325 322733   1.202  0.8938
 DayOfWeek3 - DayOfWeek4  0.02263 0.00311 322733   7.283  <.0001
 DayOfWeek3 - DayOfWeek5  0.03105 0.00308 322733  10.091  <.0001
 DayOfWeek3 - DayOfWeek6  0.06229 0.00336 322733  18.541  <.0001
 DayOfWeek3 - DayOfWeek7  0.02372 0.00319 322733   7.448  <.0001
 DayOfWeek4 - DayOfWeek5  0.00842 0.00307 322733   2.743  0.0878
 DayOfWeek4 - DayOfWeek6  0.03965 0.00335 322733  11.833  <.0001
 DayOfWeek4 - DayOfWeek7  0.00109 0.00318 322733   0.343  0.9999
 DayOfWeek5 - DayOfWeek6  0.03124 0.00333 322733   9.394  <.0001
 DayOfWeek5 - DayOfWeek7 -0.00733 0.00315 322733  -2.328  0.2304
 DayOfWeek6 - DayOfWeek7 -0.03856 0.00342 322733 -11.267  <.0001

Results are averaged over the levels of: Airline
Note: contrasts are still on the as.numeric scale
P value adjustment: tukey method for comparing a family of 7 estimates
```

# F tests

```
> # F-test
> dataset_clean <- dataset_clean %>%
+   mutate(FlightType = ifelse(Length <= 150, "Short-Haul", "Long-Haul"))
>
> short_haul_delays <- as.numeric(dataset_clean[dataset_clean$FlightType == "Short-Haul", "Delay"])
> long_haul_delays <- as.numeric(dataset_clean[dataset_clean$FlightType == "Long-Haul", "Delay"])
>
> short_haul_variance <- var(short_haul_delays)
> long_haul_variance <- var(long_haul_delays)
>
> f_value <- short_haul_variance / long_haul_variance
> df1 <- length(short_haul_delays) - 1
> df2 <- length(long_haul_delays) - 1
> p_value <- pf(f_value, df1, df2, lower.tail = FALSE)
>
> cat("F-value:", f_value, "\n")
F-value: 0.9967099
> cat("Degrees of freedom 1:", df1, "\n")
Degrees of freedom 1: 229541
> cat("Degrees of freedom 2:", df2, "\n")
Degrees of freedom 2: 93222
> cat("P-value:", p_value, "\n")
P-value: 0.7260549
```

# Answer for Research Questions

**Q: *What is the relationship between flight length and the likelihood of delay?***

**A:** The results from the Line Graph and the Two-Way ANOVA suggest a significant positive correlation between flight length and delay likelihood, with further analyses needed for quantification.

# Answer for Research Questions

**Q: *Are certain days of the week more prone to flight delays?***

**A:** The Results from barchart, ANOVA and Post-hoc tests suggest that Fridays and Sundays have the highest average delay times and are more prone to delays, supported by significant differences in delay proportions for these days.

# Answer for Research Questions

**Q: *Are there specific airlines or airport routes with a higher probability of flight delays?***

**A:** The analysis of the effect of airline on flight delays using ANOVA and notes a significant effect, but does not provide details on which airlines have higher or lower delay probabilities. The article also notes that further investigation is needed to directly address the question of airport routes and their relationship to other factors in the dataset.

# Answer for Research Questions

**Q: *Does the time of day affect the likelihood of flight delays?***

**A:** The ANOVA results show a significant effect of hour on the likelihood of delay (F value = 3087.286, $p < 2e-16$). The line graph of delays by hour of the day also reveals fluctuations in the proportion of flights delayed throughout the day. However, it would be useful to investigate further to identify specific patterns or peak hours with a higher likelihood of delays.

# Answer for Research Questions

**Q: *Are there significant differences in the average flight delay times between different days of the week?***

**A:** The Post-hoc tests results indicate that there are significant differences in the average flight delay times between different days of the week. In particular, the most significant differences occur between Day 6 (Saturday) and other days, suggesting that flights on Saturdays experience different average delay times compared to other days of the week.

# Answer for Research Questions

**Q: *Is there a significant difference in the variability of flight delay times between short-haul (length <= 150) and long-haul flights (length > 150)?***

**A:** The F-test results show a non-significant difference between the two groups, as indicated by the F-value close to 1 and the p-value above the significance level. Therefore, the article concludes that there is no significant difference in variability between short-haul and long-haul flights.

# Conclusion and Limitations

## Conclusion:

There is a slight positive correlation between flight length and the likelihood of delays, suggesting that longer flights might experience slightly more delays on average. Certain days of the week, particularly Fridays and Sundays, are more prone to flight delays compared to other days, such as Tuesdays and Wednesdays. The variability of flight delay times between short-haul and long-haul flights is not significantly different, indicating that flight length does not have a substantial impact on the variability of delay times.

## Limitation:

This project has some limitations that may affect the generalizability and reliability of its findings. The dataset used may not be up-to-date, and the analyses conducted do not cover all factors that may contribute to flight delays. Additionally, the assumptions made in the statistical tests should be verified before drawing conclusions from the results. There is also a risk of inflated false-positive rates due to multiple testing issues. Future research should consider addressing these limitations by using more recent data, incorporating additional variables, and employing more robust statistical methods to improve our understanding of the factors contributing to flight delays.

# THE END

# THANKS