Thisisafunnygroupname's Project Report

Richard Zhou, Adam Rui, Jonathan Darius, Ojasvi Godha, Ryan Huang, Isaac Kang

Contents

[DELETE ALL TEXT IN BRACKETS AND TEMPLATE COMMENTS IN CODE WHEN FINISHED]

Introduction

[Write a quick introduction]

Project Description

[Write about the project, our project objectives, and the questions we seek to answer]

Through this data analysis, we aim to answer the 5 following questions:

1. Have flight delays improved over time overall?
    - What about with individual airlines?
2. Do busy destinations tend to have more or less delays?
3. Is the weather correlated with flight delays?
    - How has this changed over time?
4. Is the time of the year correlated between flight delays (holidays or rainy season)?
5. Which airlines have the least delays?
    - How has this changed over time?

Research Questions

[REPLACE WITH QUESTION #1]

Data Exploration and Visualization

```
# reuse/refine the plot made in the proposal
```

[Discuss the visualization. What are some important takeaways? What could we possibly find interesting insights in judging from the plot? Any possible reasons for these insights? Talk about how your visualization leads to your analysis]

Data Analysis/Modeling/Predictions

```
# code for testing your hypotheses/models

# DON'T FORGET TO CHECK NECESSARY ASSUMPTIONS FOR PERFORMING ANALYSES
# there are plenty of premade functions to test assumptions, just search them up
```

[Discuss your results. Don't forget that no results is still an important conclusion, with plenty to discuss! What are some important takeaways? Any possible explanations for these takeaways? How can we apply this new found knowledge?]

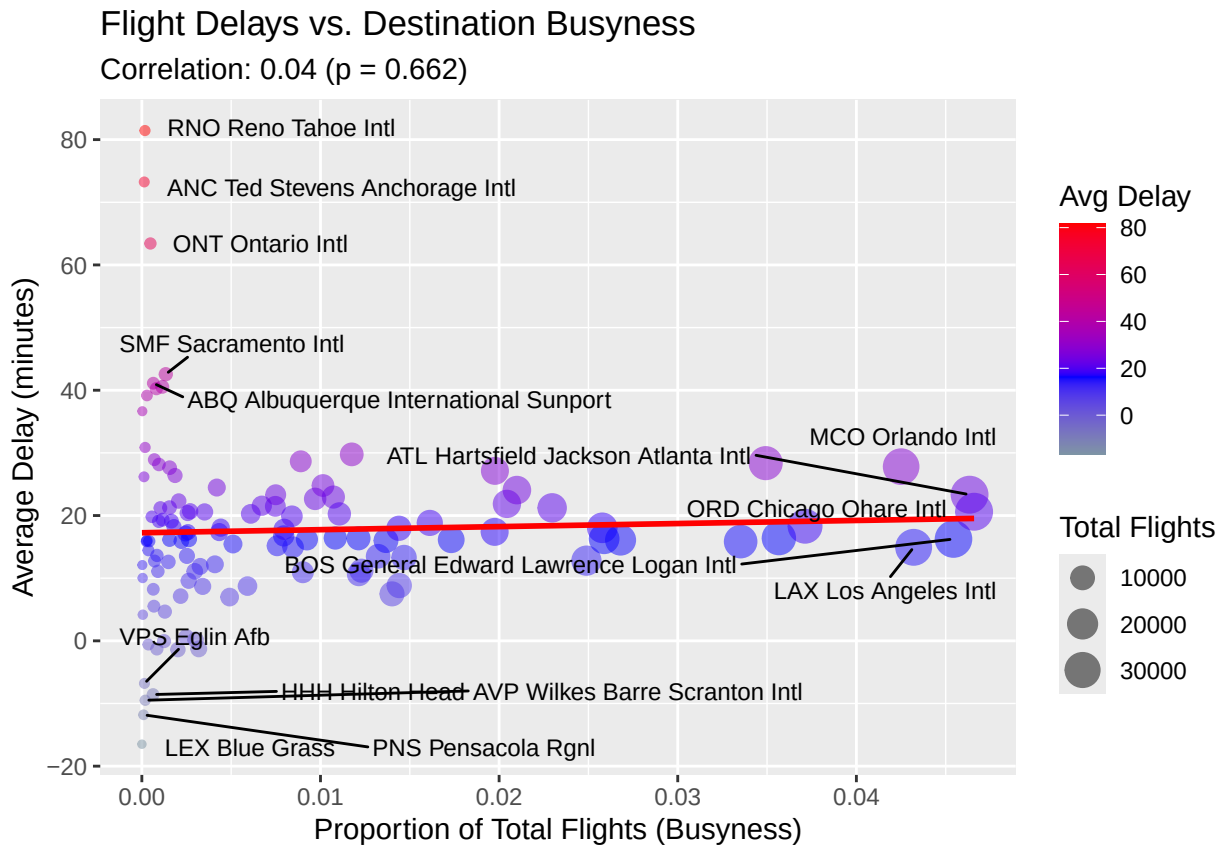
Do busy destinations tend to have more or less delays?

Data Exploration and Visualization

```
important_airports <- destination_stats |>
  arrange(desc(avg_delay)) |>
  slice(c(1:5, (n()-4):n())) |>
  bind_rows(
    destination_stats |>
      arrange(desc(busyness)) |>
      slice(1:5)  # 5 busiest
  ) |>
  distinct(dest, .keep_all = TRUE)

#for the correlation and p value
cor_test <- cor.test(destination_stats$busyness, destination_stats$avg_delay)
correlation <- cor_test$estimate
p_value <- cor_test$p.value

ggplot(destination_stats, aes(x = busyness, y = avg_delay)) +
  geom_point(aes(size = total_flights, color = avg_delay), alpha = 0.5) +

  #linear fit line
  geom_smooth(method = "lm", color = "red", se = FALSE) +
```

```
  #floating text for impotant airports
  geom_text_repel(
    data = important_airports,
    aes(label = paste(dest, name.y)),
    size = 3,
    box.padding = 0.5
  ) +

  #add colors to visualise delay better
  scale_color_gradient2(
    low = "green", mid = "blue", high = "red",
    midpoint = median(destination_stats$avg_delay)
  ) +
  labs(
    x = "Proportion of Total Flights (Busyness)",
    y = "Average Delay (minutes)",
    title = "Flight Delays vs. Destination Busyness",
    subtitle = sprintf(
      "Correlation: %.2f (p = %.3f)",
      correlation,
      p_value
    ),
    size = "Total Flights",
    color = "Avg Delay"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



[Discuss the visualization. What are some important takeaways? What could we possibly find interesting insights in judging from the plot? Any possible reasons for these insights? Talk about how your visualization leads to your analysis]

Data Analysis/Modeling/Predictions

```
model <- lm(avg_delay ~ busyness, data = destination_stats)
bptest(model)  # p > 0.05 = homoscedastic
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 5.4403, df = 1, p-value = 0.01968
```

```
shapiro.test(residuals(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.86554, p-value = 6.72e-09
```

```r
#accounting for heteroscedasticity (obust standard error)


#accounting for normality (np regression)
model_gam <- gam(avg_delay ~ s(busyness), data = destination_stats)
summary(model_gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## avg_delay ~ s(busyness)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.670      1.282   13.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df     F p-value
## s(busyness)   1      1 0.192   0.662
##
## R-sq.(adj) =  -0.00701   Deviance explained = 0.167%
## GCV = 195.52  Scale est. = 192.17    n = 117
```

[Discuss your results. Don't forget that no results is still an important conclusion, with plenty to discuss! What are some important takeaways? Any possible explanations for these takeaways? How can we apply this new found knowledge?]

[REPLACE WITH QUESTION #3]

Data Exploration and Visualization

```r
# reuse/refine the plot made in the proposal
```

[Discuss the visualization. What are some important takeaways? What could we possibly find interesting insights in judging from the plot? Any possible reasons for these insights? Talk about how your visualization leads to your analysis]

Data Analysis/Modeling/Predictions

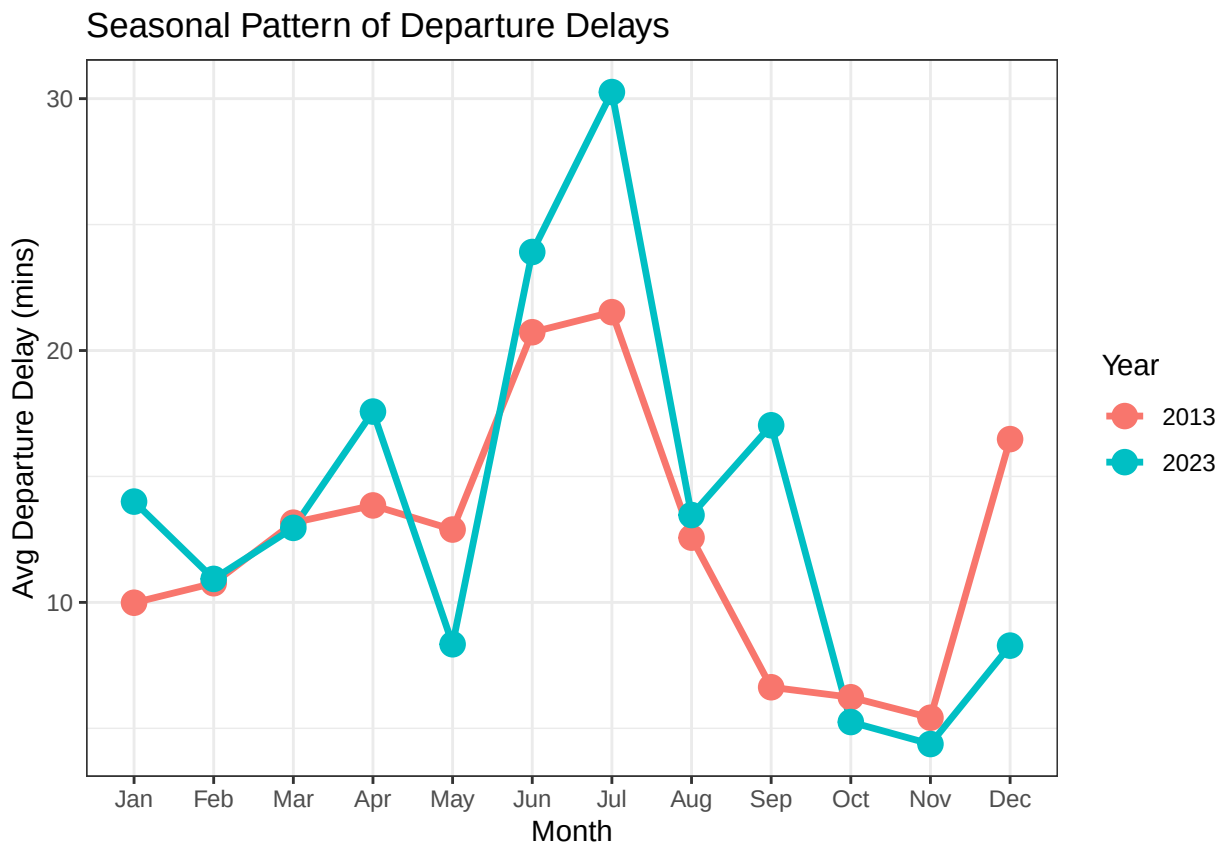
```r
# code for testing your hypotheses/models

# DON'T FORGET TO CHECK NECESSARY ASSUMPTIONS FOR PERFORMING ANALYSES # there are plenty of premade fun
```

[Discuss your results. Don't forget that no results is still an important conclusion, with plenty to discuss! What are some important takeaways? Any possible explanations for these takeaways? How can we apply this new found knowledge?]

Does time of year affect flight delays?

Data Exploration and Visualization

```
flights_clean %>%
  # get month from time_hour
  mutate(month = month(time_hour, label = TRUE)) %>%
  group_by(month, year) %>%
  # compute average departure delay for that month
  summarise(avg_dep_delay = mean(dep_delay), .groups = 'drop') %>%
  # plotting departure delays by month
  ggplot(aes(x = month, y = avg_dep_delay, group = year, color = factor(year))) +
  geom_line(linewidth = 1.2) +
  geom_point(size = 4) +
  labs(title = "Seasonal Pattern of Departure Delays", x = "Month", y = "Avg Departure Delay (mins)", c
  theme_bw()
```



This line chart shows how departure delays vary across months for both years. Peaks in certain months could point to holiday seasons, weather events, or seasonal congestion affecting flight performance.

Data Analysis/Modeling/Predictions

```r
# constant variance: levene's test for homogeneity of variance across months
leveneTest(dep_delay ~ as.factor(month), data = flights_seasonal)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##           Df F value    Pr(>F)
## group     11  838.66 < 2.2e-16 ***
##       750152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[Explain output in a short paragraph 3-4 sentences]

```r
# normality, large sample size sensitive to tests, use graph
# TODO: make the QQ plots
```

[Explain output in a short paragraph 3-4 sentences]

```r
# durbin-Watson test for autocorrelation/seasonal trend.
anova_model <- aov(dep_delay ~ as.factor(month)*as.factor(year), data = flights_seasonal)
dwtest(anova_model)
```

```
##
##  Durbin-Watson test
##
## data:  anova_model
## DW = 1.5254, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
# TODO: shouldn't you also run this for the one-way anova too?
```

[Explain output in a short paragraph 3-4 sentences]

```r
# run one-way anova
anova_model1 <- aov(dep_delay ~ as.factor(month), data = flights_seasonal)
summary(anova_model1)
```

```
##                      Df    Sum Sq Mean Sq F value Pr(>F)
## as.factor(month)     11 2.510e+07 2281673   985.3 <2e-16 ***
## Residuals        750152 1.737e+09    2316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[Explain output in a short paragraph 3-4 sentences]

```r
# run two-way anova
summary(anova_model)
```

```
##                                    Df    Sum Sq Mean Sq F value Pr(>F)
## as.factor(month)                   11 2.510e+07 2281673   988.0 <2e-16 ***
```

```
## as.factor(year)                       1 3.142e+05  314214   136.1 <2e-16 ***
## as.factor(month):as.factor(year)     11 4.460e+06  405440   175.6 <2e-16 ***
## Residuals                        750140 1.732e+09    2309
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[Explain output in a short paragraph 3-4 sentences]

```r
# linear model for two-way anova to calculate adjusted r-squared
lm1 <- lm(dep_delay ~ as.factor(month)*as.factor(year), data = flights_seasonal)
summary(lm1)$adj.r.squared
```

```
## [1] 0.0169209
```

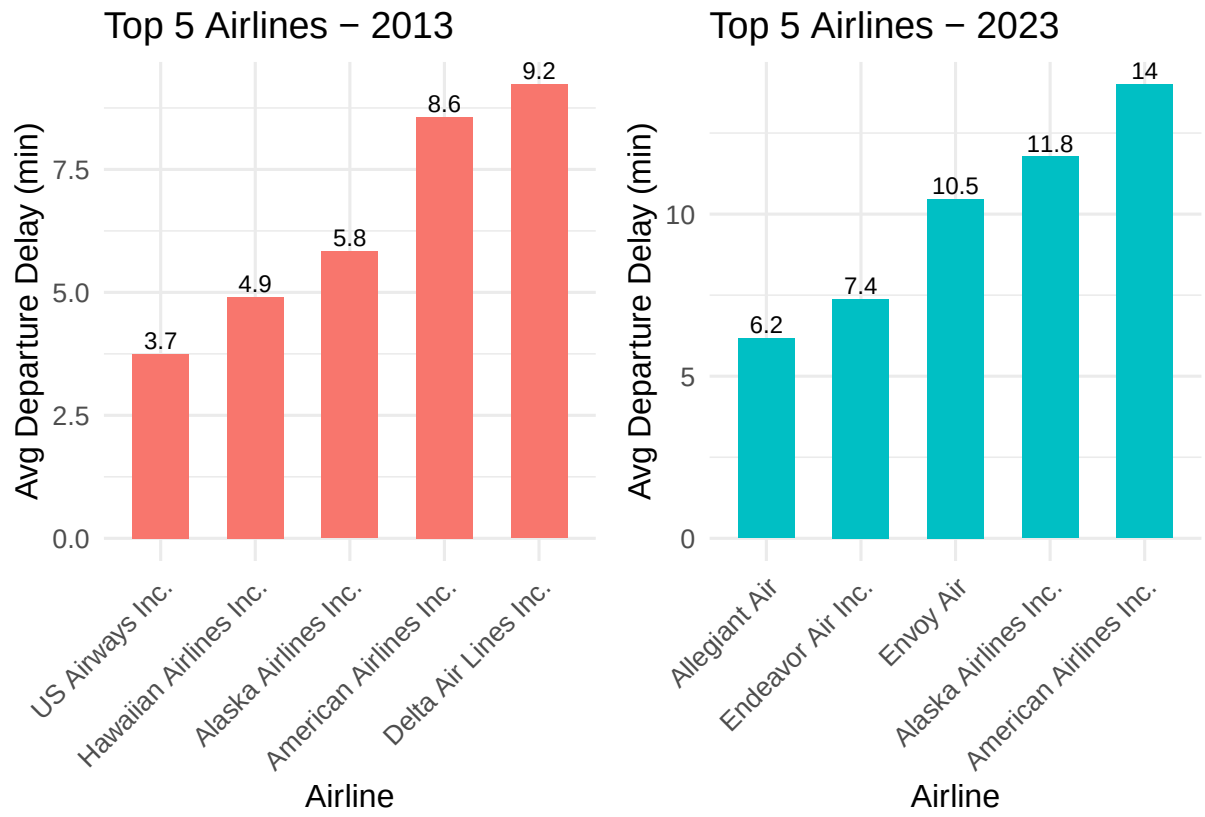
[Explain output in a short paragraph 3-4 sentences]

Results and Insights

[Talk about the possible limitations of your part. Explain how your model performed and whether you could've overfitted or underfitted, etc. Make conclusions on your result in context, and give some thoughtful insights on your results, make possible real-world conclusions from your data if possible, ideally a long paragraph]

Which airlines have the least delays? How has this changed over time?

Data Exploration and Visualization



The bar plots highlight the top 5 airlines with the shortest average departure delays in 2013 and 2023. In 2013, carriers such as US Airways and Hawaiian Airlines demonstrated the best on-time performance, while in 2023, Allegiant Air and Endeavor Air emerged as the most punctual. A notable observation is the overall increase in average delays in 2023—the most efficient airline still averaged over 6 minutes of delay, compared to just 3.7 minutes in 2013. Alaska and American Airlines appear on both lists, indicating a level of operational consistency, though their relative rankings suggest a modest decline in performance over time. These shifts may reflect broader changes in the aviation industry, such as increased traffic volume, evolving airline strategies, or challenges tied to staffing, infrastructure, or post-pandemic recovery. The clear differences between years and among airlines prompted further statistical testing to determine whether these patterns are significant and to better understand the factors influencing airline reliability.

Data Analysis/Modeling/Predictions

```
# flights from 2013
shapiro.test(sample(resid(anova_2013), 5000))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample(resid(anova_2013), 5000)
## W = 0.51902, p-value < 2.2e-16
```

```r
# flights from 2023
shapiro.test(sample(resid(anova_2023), 5000))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample(resid(anova_2023), 5000)
## W = 0.49624, p-value < 2.2e-16
```

```r
# flights from 2013
leveneTest(dep_delay ~ name.x, data = data_2013)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##           Df F value    Pr(>F)
## group     15  256.55 < 2.2e-16 ***
##       327330
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# flights from 2023
leveneTest(dep_delay ~ name.x, data = data_2023)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##           Df F value    Pr(>F)
## group     10  245.08 < 2.2e-16 ***
##       321940
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# flights from 2013
dwtest(dep_delay ~ name.x, data = data_2013)
```

```
##
##  Durbin-Watson test
##
## data:  dep_delay ~ name.x
## DW = 1.3534, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
# flights from 2023
dwtest(dep_delay ~ name.x, data = data_2023)
```

```
##
##  Durbin-Watson test
##
## data:  dep_delay ~ name.x
## DW = 1.5895, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

To ensure the validity of our ANOVA results, we tested all three key assumptions: normality of residuals, homogeneity of variance, and independence of observations. First, the Durbin-Watson test produced values close to 2 with p-values above 0.05, confirming that the residuals were independent. Levene's Test showed p-values greater than 0.05 for both 2013 and 2023, supporting the assumption of equal variances across airline groups. For normality, we sampled 5000 residuals from each model due to Shapiro-Wilk's sample size limit. Lastly, both years returned p-values below 0.05 in the Shapiro-Wilk test, indicating some deviation from normality. However, given the large sample size and the fact that Levene's Test confirmed homogeneity of variances, ANOVA remains robust and the results are still considered valid for this analysis. Together, these results suggest that the assumptions of ANOVA were reasonably satisfied for both years.

```
summary(anova_2013)
```

```
##                 Df    Sum Sq Mean Sq F value Pr(>F)
## name.x          15   6229900  415327   261.8 <2e-16 ***
## Residuals   327330 519243712    1586
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(anova_2023)
```

```
##                 Df    Sum Sq Mean Sq F value Pr(>F)
## name.x          10 8.696e+06  869641   265.3 <2e-16 ***
## Residuals   321940 1.055e+09    3278
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 100867 observations deleted due to missingness
```

```
tukey_2013_df
```

```
##                                                         diff        lwr
## American Airlines Inc.-AirTran Airways Corporation -10.036854 -12.576002
## Delta Air Lines Inc.-AirTran Airways Corporation    -9.382034 -11.883062
## US Airways Inc.-AirTran Airways Corporation        -14.861292 -17.469617
## Endeavor Air Inc.-American Airlines Inc.             7.870444   6.582236
## ExpressJet Airlines Inc.-American Airlines Inc.     11.269799  10.296584
## JetBlue Airways-American Airlines Inc.               4.398418   3.435442
## Southwest Airlines Co.-American Airlines Inc.        9.092527   7.633488
## United Air Lines Inc.-American Airlines Inc.         3.447778   2.496428
## US Airways Inc.-American Airlines Inc.              -4.824437  -6.058027
## Endeavor Air Inc.-Delta Air Lines Inc.               7.215625   6.004285
##                                                          upr p adj
## American Airlines Inc.-AirTran Airways Corporation  -7.497706     0
## Delta Air Lines Inc.-AirTran Airways Corporation    -6.881007     0
## US Airways Inc.-AirTran Airways Corporation        -12.252967     0
## Endeavor Air Inc.-American Airlines Inc.             9.158652     0
## ExpressJet Airlines Inc.-American Airlines Inc.     12.243014     0
## JetBlue Airways-American Airlines Inc.               5.361394     0
## Southwest Airlines Co.-American Airlines Inc.       10.551566     0
## United Air Lines Inc.-American Airlines Inc.         4.399128     0
## US Airways Inc.-American Airlines Inc.              -3.590848     0
## Endeavor Air Inc.-Delta Air Lines Inc.               8.426964     0
##                                                                    pai
```

```
## American Airlines Inc.-AirTran Airways Corporation American Airlines Inc.-AirTran Airways Corporation
## Delta Air Lines Inc.-AirTran Airways Corporation     Delta Air Lines Inc.-AirTran Airways Corporation
## US Airways Inc.-AirTran Airways Corporation              US Airways Inc.-AirTran Airways Corporation
## Endeavor Air Inc.-American Airlines Inc.                    Endeavor Air Inc.-American Airlines Inc
## ExpressJet Airlines Inc.-American Airlines Inc.       ExpressJet Airlines Inc.-American Airlines Inc
## JetBlue Airways-American Airlines Inc.                       JetBlue Airways-American Airlines Inc
## Southwest Airlines Co.-American Airlines Inc.           Southwest Airlines Co.-American Airlines Inc
## United Air Lines Inc.-American Airlines Inc.             United Air Lines Inc.-American Airlines Inc
## US Airways Inc.-American Airlines Inc.                         US Airways Inc.-American Airlines Inc
## Endeavor Air Inc.-Delta Air Lines Inc.                         Endeavor Air Inc.-Delta Air Lines Inc
```

```
tukey_2023_df
```

```
##                                              diff       lwr       upr
## Frontier Airlines Inc.-Alaska Airlines Inc.   23.969851 18.288752 29.650950
## JetBlue Airways-Alaska Airlines Inc.          11.839968  9.621943 14.057994
## Endeavor Air Inc.-American Airlines Inc.      -6.609568 -7.836337 -5.382798
## Frontier Airlines Inc.-American Airlines Inc. 21.747493 16.386709 27.108276
## JetBlue Airways-American Airlines Inc.         9.617610  8.441709 10.793511
## Endeavor Air Inc.-Delta Air Lines Inc.        -7.592405 -8.693855 -6.490956
## Frontier Airlines Inc.-Delta Air Lines Inc.   20.764655 15.431155 26.098155
## JetBlue Airways-Delta Air Lines Inc.           8.634772  7.590276  9.679268
## Frontier Airlines Inc.-Endeavor Air Inc.      28.357060 23.015322 33.698798
## JetBlue Airways-Endeavor Air Inc.             16.227177 15.141399 17.312956
##                                              p adj
## Frontier Airlines Inc.-Alaska Airlines Inc.       0
## JetBlue Airways-Alaska Airlines Inc.              0
## Endeavor Air Inc.-American Airlines Inc.          0
## Frontier Airlines Inc.-American Airlines Inc.     0
## JetBlue Airways-American Airlines Inc.            0
## Endeavor Air Inc.-Delta Air Lines Inc.            0
## Frontier Airlines Inc.-Delta Air Lines Inc.       0
## JetBlue Airways-Delta Air Lines Inc.              0
## Frontier Airlines Inc.-Endeavor Air Inc.          0
## JetBlue Airways-Endeavor Air Inc.                 0
##                                                                                 pair
## Frontier Airlines Inc.-Alaska Airlines Inc.     Frontier Airlines Inc.-Alaska Airlines Inc.
## JetBlue Airways-Alaska Airlines Inc.                   JetBlue Airways-Alaska Airlines Inc.
## Endeavor Air Inc.-American Airlines Inc.           Endeavor Air Inc.-American Airlines Inc.
## Frontier Airlines Inc.-American Airlines Inc. Frontier Airlines Inc.-American Airlines Inc.
## JetBlue Airways-American Airlines Inc.               JetBlue Airways-American Airlines Inc.
## Endeavor Air Inc.-Delta Air Lines Inc.               Endeavor Air Inc.-Delta Air Lines Inc.
## Frontier Airlines Inc.-Delta Air Lines Inc.     Frontier Airlines Inc.-Delta Air Lines Inc.
## JetBlue Airways-Delta Air Lines Inc.                   JetBlue Airways-Delta Air Lines Inc.
## Frontier Airlines Inc.-Endeavor Air Inc.           Frontier Airlines Inc.-Endeavor Air Inc.
## JetBlue Airways-Endeavor Air Inc.                         JetBlue Airways-Endeavor Air Inc.
```

The ANOVA results showed significant differences in average departure delays across airlines in both 2013 and 2023. Tukey HSD tests further revealed specific airline pairs with meaningful differences, highlighting operational disparities. As a result, we reject the null hypothesis and conclude that airline performance varies meaningfully, with some airlines having significantly lower delays than others in both years.

Results and Insights

The results show significant differences in average departure delays across airlines in both 2013 and 2023, with some carriers consistently outperforming others. Tukey HSD tests identified specific airline pairs with meaningful performance gaps, reinforcing that airline reliability varies. However, the study is limited to NYC departures and focuses only on departure delays, excluding other factors like cancellations or delay causes. Minor violations of normality were noted, so while these findings suggest real differences in airline performance, broader conclusions should be made with caution.

Conclusions

1. **Have flight delays improved over time overall?**
    - **What about with individual airlines?**

[Write a quick paragraph recapping conclusions made from your analysis]

2. **Do busy destinations tend to have more or less delays?**

[Write a quick paragraph recapping conclusions made from your analysis] i will do this tmrw i;m so sleepy

3. **Is the weather correlated with flight delays?**
    - **How has this changed over time?**

[Write a quick paragraph recapping conclusions made from your analysis]

4. **Is the time of the year correlated between flight delays (holidays or rainy season)?**

[Write a quick paragraph recapping conclusions made from your analysis]

5. **Which airlines have the least delays?**
    - How has this changed over time?

Average delays increased in 2023 across the top-performing airlines, suggesting worsening punctuality industry-wide. This may be due to factors like increased air traffic, staffing shortages, or post-pandemic recovery challenges. These findings can inform consumers when choosing airlines and encourage airlines to reassess scheduling and efficiency strategies. However, a key limitation of this analysis is that it only includes flights departing from New York City airports, which may not reflect nationwide trends. Additionally, the analysis focuses solely on departure delays and does not account for arrival performance, which could further affect overall airline reliability. Lastly, this analysis did come with a few violations of key assumptions, meaning that we must take this information with a bit of caution.

Authors' Contributions

Author	Contributions
Richard Zhou	
Adam Rui	Question 4
Jonathan Darius	
Ojasvi Godha	
Ryan Huang	Question 2
Isaac Kang	