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

Flight delays are a constant challenge in the air travel industry, impacting efficiency and passenger satisfaction. This project aims to investigate the underlying causes of flight delays in New York City and how these patterns have evolved over time. By analyzing both recent and historical flight data, we seek to identify the major contributors to delays and provide actionable insights for improving airline performance and the overall passenger experience.

Project Description

This analysis will utilize the `nycflights13` and `nycflights23` datasets, which contain records of flights departing from NYC airports. The project will involve exploratory data analysis (EDA), statistical testing, and comparative analysis, using tools such as `dplyr`, `ggplot2`, and many more to assess the significance of delay-related factors. Through this project, we intend to discover trends and patterns in flight delays, to provide a deeper insight into the aspects we can improve in air travel.

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
```

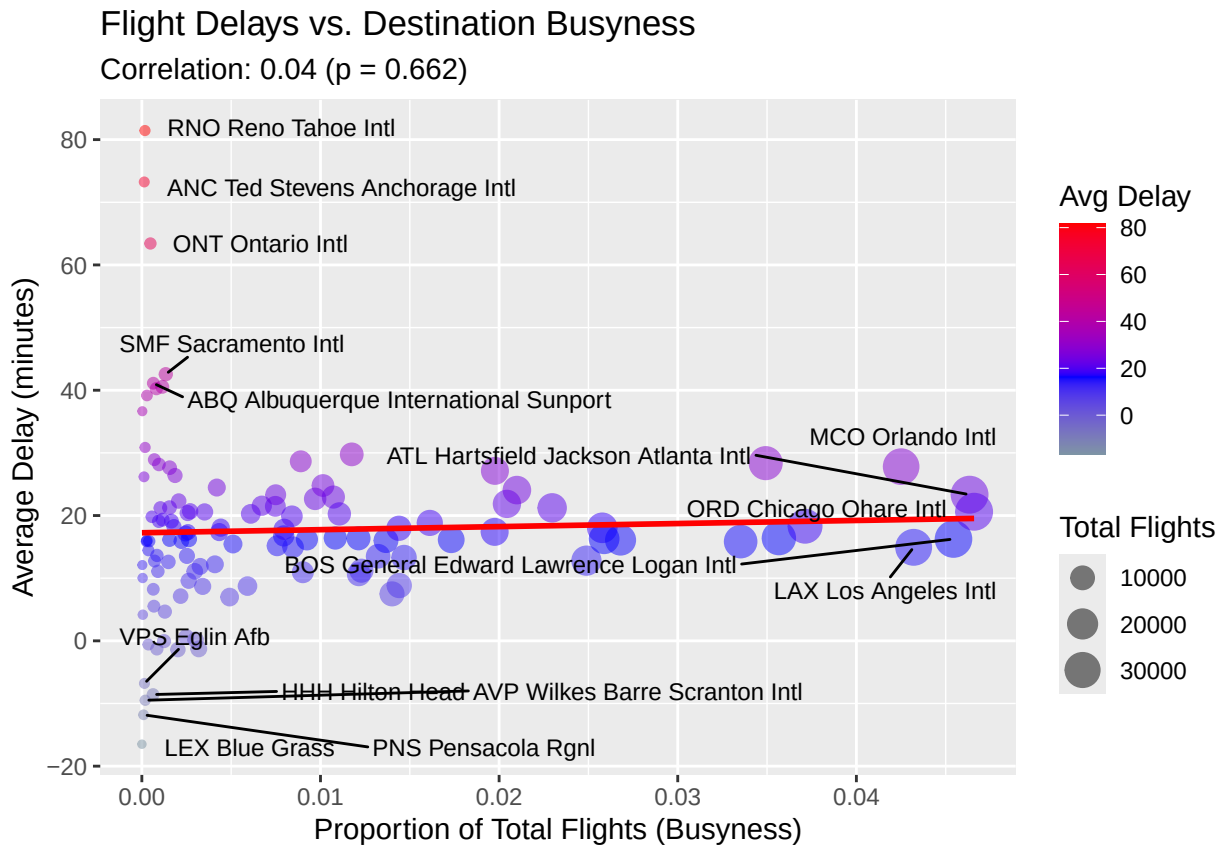
[Discuss your results. Don't forget that no results is still an important conclusion, with plenty to discuss! What are some important takeaways? Any possible explanations for these takeaways? How can we apply this new found knowledge?]

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

```
# 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

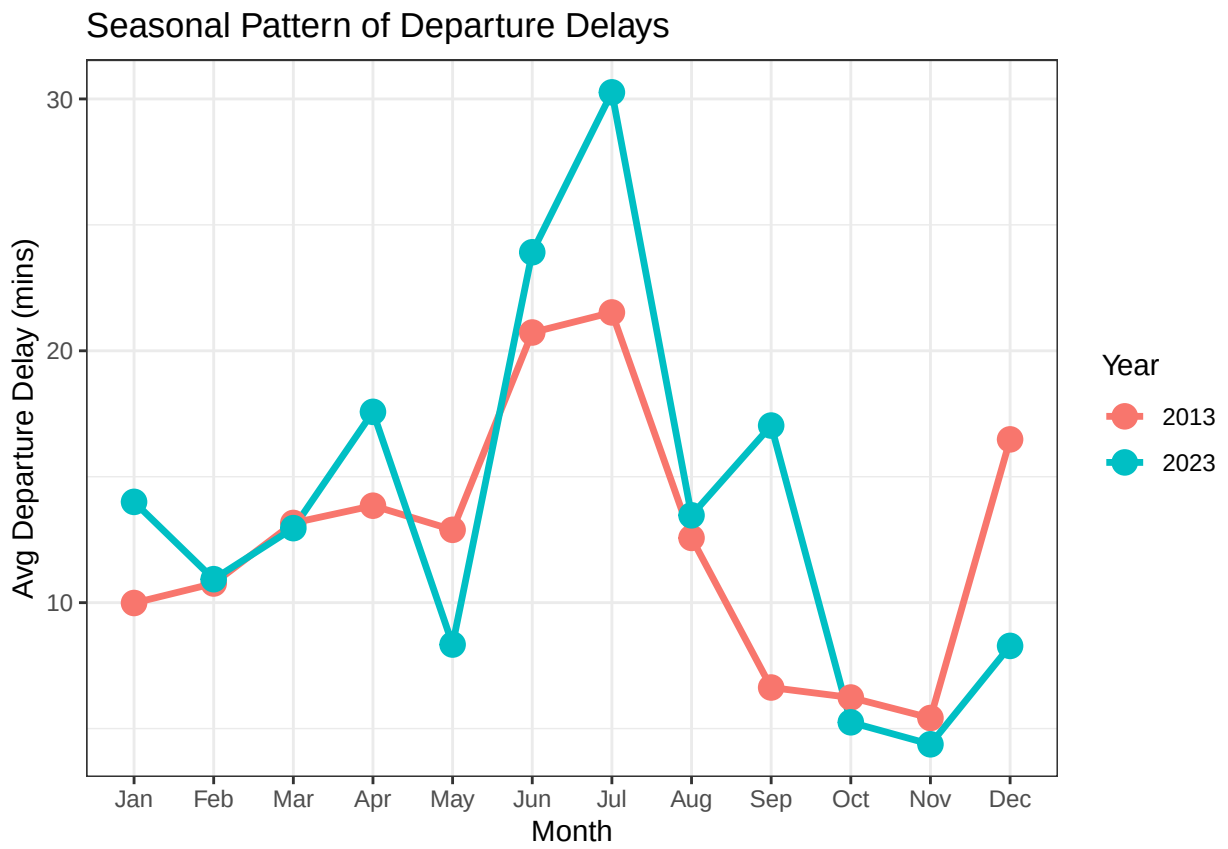
```
# code for testing your hypotheses/models

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

```
# 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]

```
# 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]

[REPLACE WITH QUESTION #5]

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 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?]

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?

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

Authors' Contributions

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