Thisisafunnygroupname's Project Report

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

Contents

Introduction **2**

Project Description **2**

Research Questions **3**

Have flight delays improved over time overall? What about with individual airlines? 3

Do busy destinations tend to have more or less delays? 15

[REPLACE WITH QUESTION #3] 18

Does time of year affect flight delays? 19

[REPLACE WITH QUESTION #5] 21

Conclusions **22**

Authors' Contributions **23**

[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?

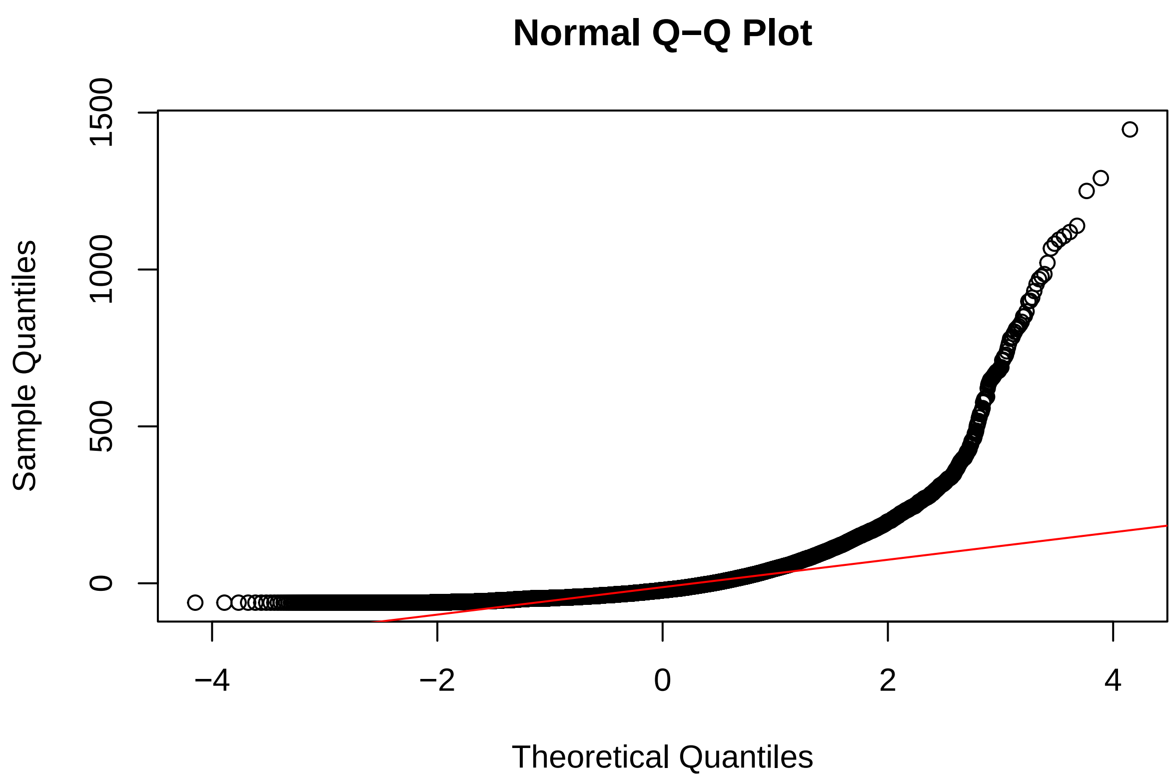
Research Questions

Have flight delays improved over time overall? What about with individual airlines?

Part 1: Have flight delays improved or gotten worse between 2013 and 2023?

First we test for normality.

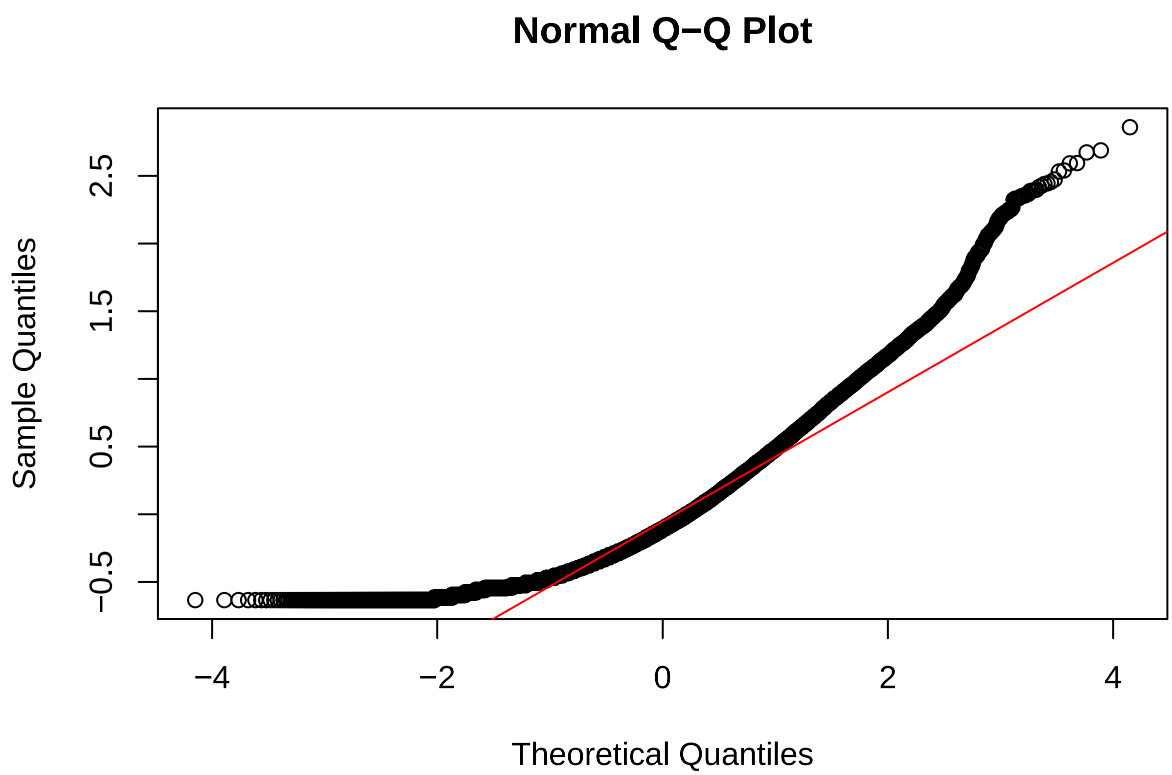
```
dep_delay_resids <- sample(residuals(dep_delay_model), size = 30000)

# Now plot the Q-Q plot with the sample for easier loading
qqnorm(dep_delay_resids)
qqline(dep_delay_resids, col = "red")
```



We don't seem to meet this assumption so let's do a shifted log transformation to help. This type of log transformation ensures all values are positive before the logging takes place and makes sure to include all values possible. After, we are going to check the normality assumption again.

```
min_dep_delay <- min(flights_clean$dep_delay, na.rm = TRUE)
flights_clean_log$log_dep_delay <- log(flights_clean_log$dep_delay - min_dep_delay + 1)
dep_delay_model_log <- lm(log_dep_delay ~ factor(year), data = flights_clean_log)

dep_delay_log_resids <- sample(residuals(dep_delay_model_log), size = 30000)
```

```r
# Now plot the Q-Q plot with the sample for easier loading
qqnorm(dep_delay_log_resids)
qqline(dep_delay_log_resids, col = "red")
```



And since the points are close to the red line, we can claim that yes the data is from a normal population.

```r
leveneTest(log_dep_delay ~ factor(year), data = flights_clean_log)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##            Df F value    Pr(>F)
## group       1  608.81 < 2.2e-16 ***
##        210624
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can also check for homoscedasticity using the Levene's Test which is not met. The small p-value here shows signs of heteroscedasticity which means we cannot use regular 'summary()' to get results of the model. Let's look into the model itself using robust standard errors which assumes unequal error variances.

```r
coeftest(dep_delay_model_log, vcov = vcovHC)
```

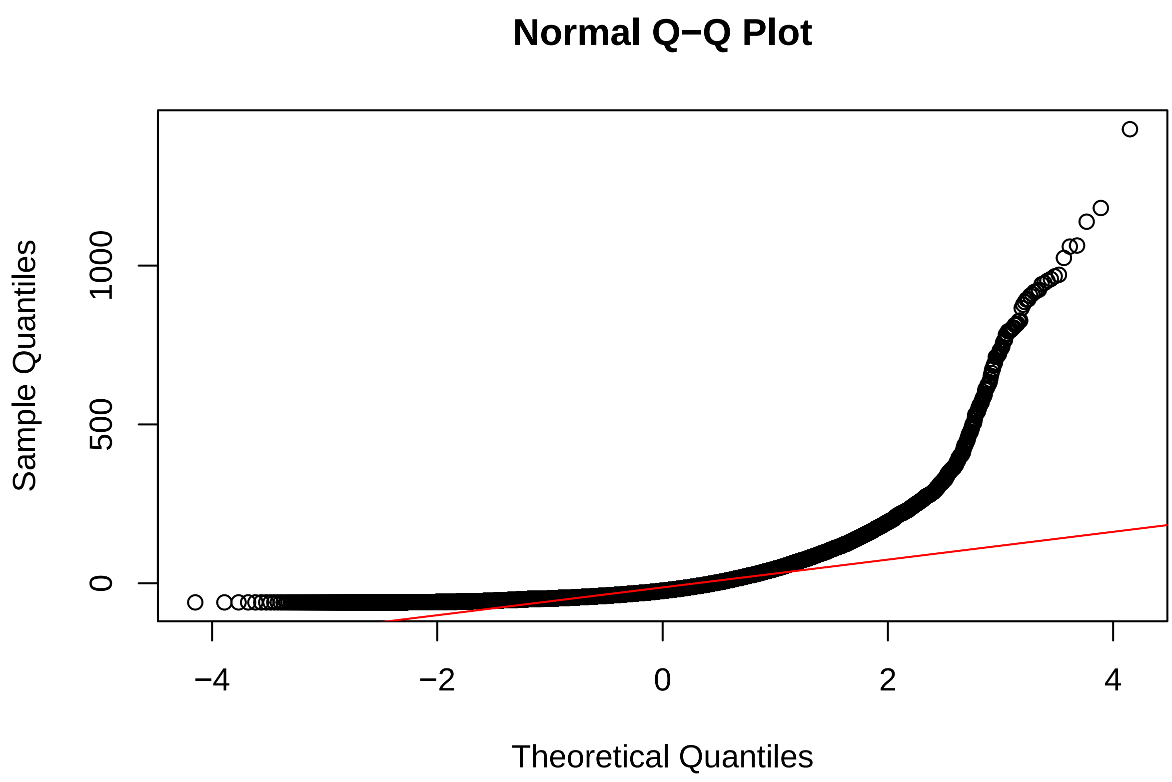
```
##
## t test of coefficients:
##
```

```
##                    Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)      4.4764619  0.0014273 3136.424 < 2.2e-16 ***
## factor(year)2023 0.0894405  0.0020859   42.879 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This model helps us see the overall difference in departure delays between 2013 and 2023. The model shows a statistically significant increase in departure delays from 2013 to 2023, indicated by a positive coefficient with a very small p-value. In this context, we can conclude that that flight departure delays have gotten worse over time.

Let's do the whole process again, with arrival delays.

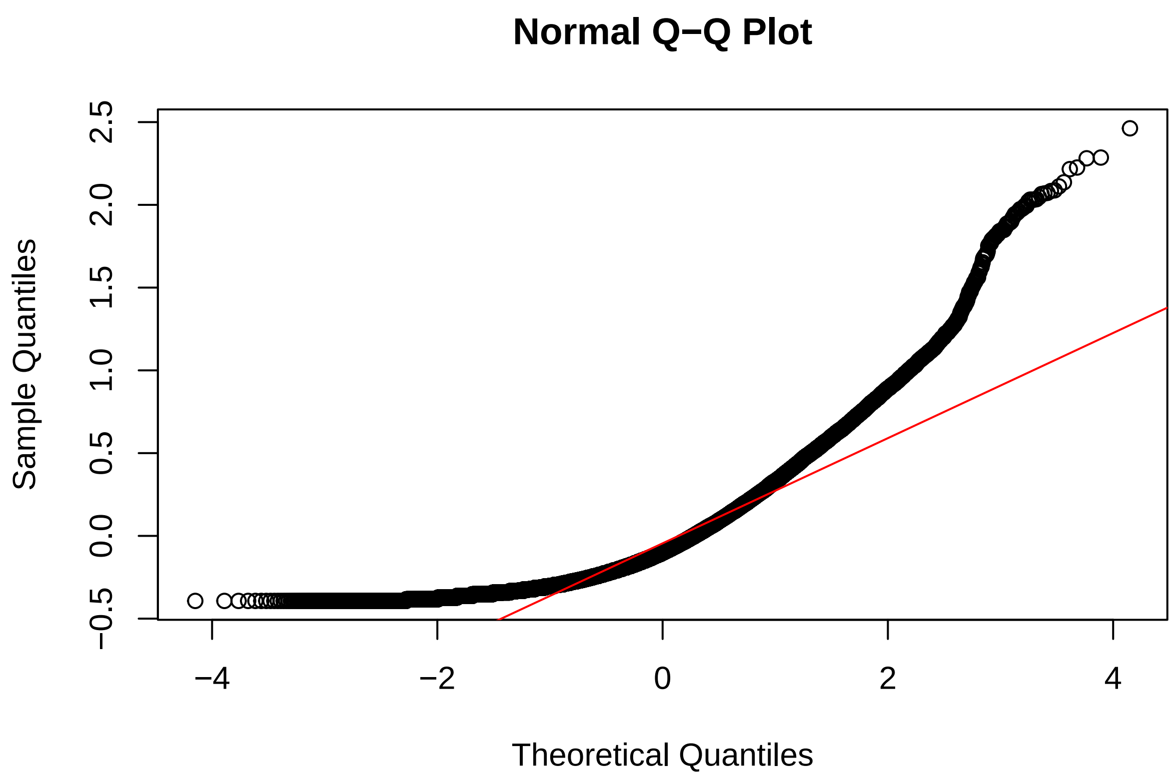
```r
arr_delay_model <- lm(data=flights_clean_log, arr_delay~factor(year))

arr_delay_resids <- sample(residuals(arr_delay_model), size = 30000)

# Now plot the Q-Q plot with the sample for easier loading
qqnorm(arr_delay_resids)
qqline(arr_delay_resids, col = "red")
```



We seem to have the same issue as the departure delays, so let's do another shifted log transformation and test again.

```r
min_arr_delay <- min(flights_clean$arr_delay, na.rm = TRUE)
flights_clean_log$log_arr_delay <- log(flights_clean_log$arr_delay - min_arr_delay + 1)
```

```r
arr_delay_model_log <- lm(log_arr_delay ~ factor(year), data = flights_clean_log)

arr_delay_log_resids <- sample(residuals(arr_delay_model_log), size = 30000)

# Now plot the Q-Q plot with the sample for easier loading
qqnorm(arr_delay_log_resids)
qqline(arr_delay_log_resids, col = "red")
```




```r
leveneTest(log_arr_delay ~ factor(year), data = flights_clean_log)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##            Df F value    Pr(>F)
## group       1  935.59 < 2.2e-16 ***
##        210624
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can, again, check for homoscedasticity using the Levene's Test which is, again, not met. The small p-value here shows signs of heteroscedasticity which means we cannot use regular 'summary()' to get results of the model. Let's look into the model itself using robust standard errors which assumes unequal error variances.

```r
coeftest(arr_delay_model_log, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)      4.93679104 0.00098648 5004.466 < 2.2e-16 ***
## factor(year)2023 0.04099626 0.00148077   27.686 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When we compare the arrival delays, we have a statistically significant increase as well, with a positive coefficient and small p-value, indicating that arrival delays have also gotten worse between 2013 and 2023, however slightly.

Let's take a quick look at the performance of both these arrival and delay models.

```r
print(paste('Arrival Adj R^2: ', summary(arr_delay_model_log)$adj.r.squared))
## [1] "Arrival Adj R^2:  0.0035573193253432"

print(paste("Departure Adj R^2: ", summary(dep_delay_model_log)$adj.r.squared))
## [1] "Departure Adj R^2:  0.00854136653743243"
```

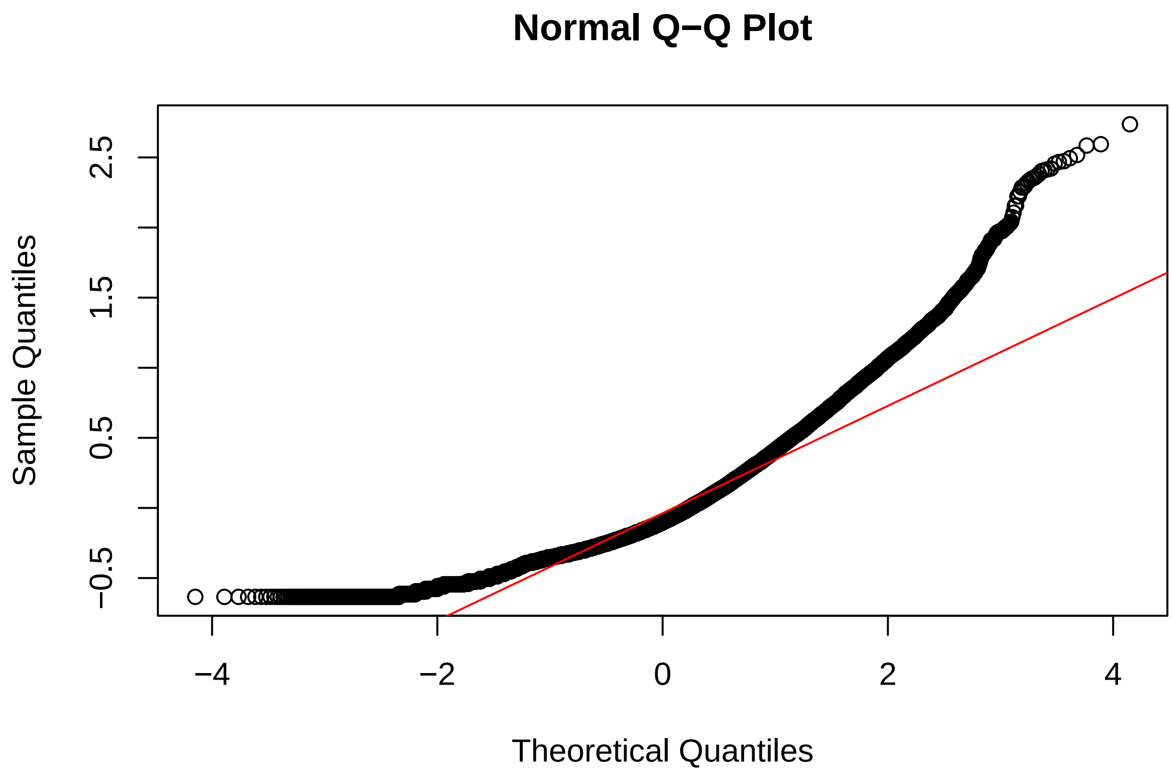
The Adjusted R^2 value helps us measure the quality of the model. These are really small values which means these models are not great. But in this case the small values are fine as the dataset is large, and the focus of this question was to identify average differences in delays over time, so the models still provided meaningful insights for this question.

For some additional comparison, we can try to see how much more or less departure delays have changed versus arrival delays from 2013 to 2023.

```r
both_delay_flights <- flights_clean_log %>%
  select(year, log_dep_delay, log_arr_delay) %>%
  pivot_longer(cols = c(log_dep_delay, log_arr_delay),
               names_to = "delay_type",
               values_to = "delay_value")

both_delay_model <- lm(data=both_delay_flights, delay_value~factor(year)*delay_type)
```

Let's first check for the normality assumption.

```r
both_delay_resids <- sample(residuals(both_delay_model), size = 30000)

# Now plot the Q-Q plot with the sample for easier loading
qqnorm(both_delay_resids)
qqline(both_delay_resids, col = "red")
```



Based on this plot, we can say we meet the normality assumption. Now let's check for homoscedasticity with the Levene's Test.

```
leveneTest(delay_value ~ factor(year)*delay_type,  data = both_delay_flights)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##            Df F value    Pr(>F)
## group       3    6481 < 2.2e-16 ***
##        421248
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can check for homoscedasticity using the Levene's Test which is, again, not met. The small p-value here shows signs of heteroscedasticity which means we cannot use regular 'summary()' to get results of the model. Let's look into the model itself using robust standard errors which assumes unequal error variances.

```
coeftest(both_delay_model, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                                          Estimate  Std. Error  t value
## (Intercept)                            4.93679104  0.00098648 5004.466
## factor(year)2023                       0.04099626  0.00148077   27.686
## delay_typelog_dep_delay               -0.46032914  0.00173499 -265.321
```

```
## factor(year)2023:delay_typelog_dep_delay  0.04844426  0.00255802   18.938
##                                            Pr(>|t|)
## (Intercept)                               < 2.2e-16 ***
## factor(year)2023                          < 2.2e-16 ***
## delay_typelog_dep_delay                   < 2.2e-16 ***
## factor(year)2023:delay_typelog_dep_delay < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The key term in the model is 'factor(year)2023:delay__typedep__delay', which represents the additional change in departure delays from 2013 to 2023 relative to arrival delays. From the two previous models, we know arrival delays increased over the years by about 10% and departure delays increased by about 28%. In this combined model, we can see that as the coefficient is positive and, based on the p-value, is statistically significant. With this, we can conclude that the departure delays not only increased over time, but did so to a significantly greater extent than arrival delays did.

Part 2: Have individual airlines gotten better or worse with delays over time?

```
delay_model_airline <- lm(data=flights_clean_log, log_dep_delay~factor(year) * name.x)
summary(delay_model_airline)
```

```
##
## Call:
## lm(formula = log_dep_delay ~ factor(year) * name.x, data = flights_clean_log)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7388 -0.3684 -0.1064  0.2664  2.9494
##
## Coefficients: (5 not defined because of singularities)
##                                                 Estimate Std. Error t value
## (Intercept)                                     4.412264   0.012409 355.577
## factor(year)2023                                0.139575   0.004658  29.962
## name.xAlaska Airlines Inc.                      0.070165   0.043542   1.611
## name.xAmerican Airlines Inc.                    0.060822   0.013616   4.467
## name.xDelta Air Lines Inc.                      0.040264   0.013206   3.049
## name.xEndeavor Air Inc.                         0.182796   0.014031  13.028
## name.xEnvoy Air                                 0.050477   0.013550   3.725
## name.xExpressJet Airlines Inc.                  0.153855   0.012859  11.965
## name.xFrontier Airlines Inc.                    0.057058   0.030807   1.852
## name.xHawaiian Airlines Inc.                    0.001792   0.075629   0.024
## name.xJetBlue Airways                           0.053297   0.012916   4.126
## name.xMesa Airlines Inc.                        0.159252   0.035445   4.493
## name.xSkyWest Airlines Inc.                     0.236242   0.169346   1.395
## name.xSouthwest Airlines Co.                    0.038806   0.014257   2.722
## name.xUnited Air Lines Inc.                    -0.002563   0.012907  -0.199
## name.xUS Airways Inc.                          -0.033307   0.014476  -2.301
## name.xVirgin America                            0.037862   0.017907   2.114
## factor(year)2023:name.xAlaska Airlines Inc.    -0.066792   0.043460  -1.537
## factor(year)2023:name.xAmerican Airlines Inc.  -0.031584   0.008813  -3.584
## factor(year)2023:name.xDelta Air Lines Inc.    -0.022266   0.007576  -2.939
## factor(year)2023:name.xEndeavor Air Inc.       -0.196119   0.009324 -21.034
```

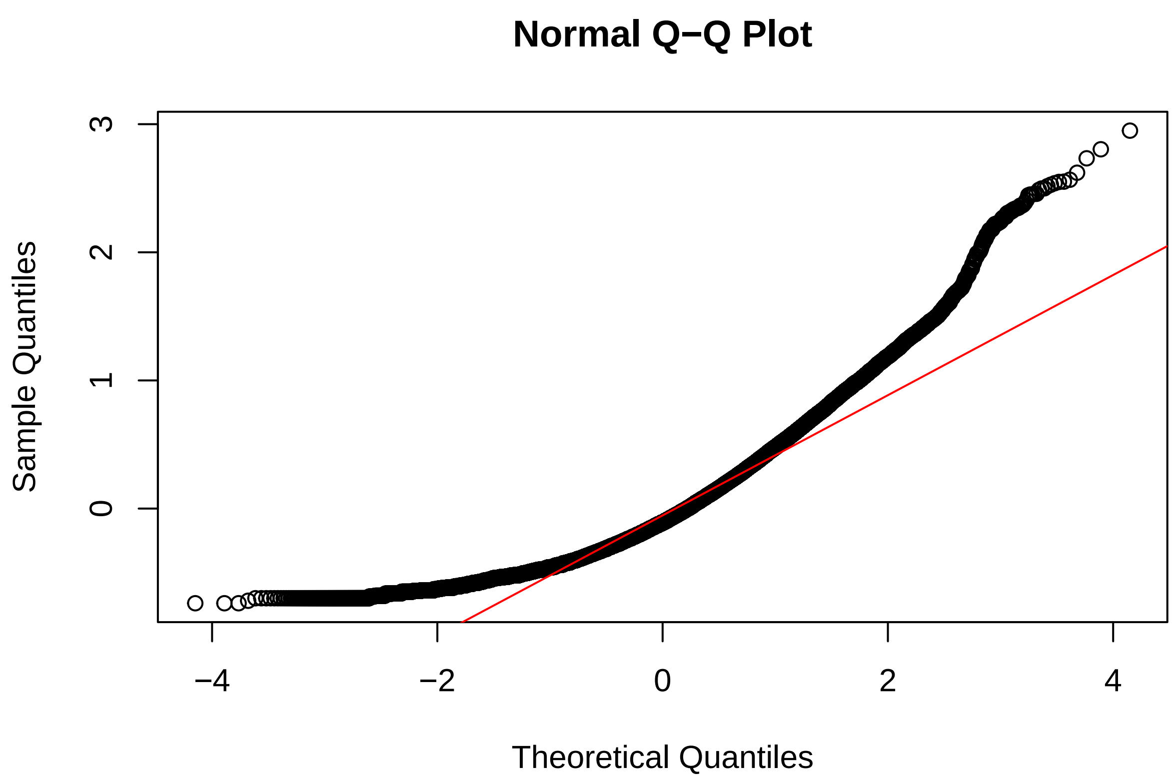
```
## factor(year)2023:name.xEnvoy Air              -0.179610   0.046106  -3.896
## factor(year)2023:name.xExpressJet Airlines Inc.        NA         NA      NA
## factor(year)2023:name.xFrontier Airlines Inc.    0.061739   0.034918   1.768
## factor(year)2023:name.xHawaiian Airlines Inc.   -0.160427   0.082629  -1.942
## factor(year)2023:name.xJetBlue Airways           0.026472   0.006633   3.991
## factor(year)2023:name.xMesa Airlines Inc.              NA         NA      NA
## factor(year)2023:name.xSkyWest Airlines Inc.    -0.170036   0.169349  -1.004
## factor(year)2023:name.xSouthwest Airlines Co.   -0.174678   0.011056 -15.799
## factor(year)2023:name.xUnited Air Lines Inc.           NA         NA      NA
## factor(year)2023:name.xUS Airways Inc.                 NA         NA      NA
## factor(year)2023:name.xVirgin America                  NA         NA      NA
##                                                 Pr(>|t|)
## (Intercept)                                      < 2e-16 ***
## factor(year)2023                                 < 2e-16 ***
## name.xAlaska Airlines Inc.                      0.107088
## name.xAmerican Airlines Inc.                    7.94e-06 ***
## name.xDelta Air Lines Inc.                      0.002297 **
## name.xEndeavor Air Inc.                          < 2e-16 ***
## name.xEnvoy Air                                 0.000195 ***
## name.xExpressJet Airlines Inc.                   < 2e-16 ***
## name.xFrontier Airlines Inc.                    0.064014 .
## name.xHawaiian Airlines Inc.                    0.981099
## name.xJetBlue Airways                           3.69e-05 ***
## name.xMesa Airlines Inc.                        7.03e-06 ***
## name.xSkyWest Airlines Inc.                     0.163011
## name.xSouthwest Airlines Co.                    0.006494 **
## name.xUnited Air Lines Inc.                     0.842615
## name.xUS Airways Inc.                           0.021402 *
## name.xVirgin America                            0.034485 *
## factor(year)2023:name.xAlaska Airlines Inc.     0.124328
## factor(year)2023:name.xAmerican Airlines Inc.   0.000339 ***
## factor(year)2023:name.xDelta Air Lines Inc.     0.003292 **
## factor(year)2023:name.xEndeavor Air Inc.         < 2e-16 ***
## factor(year)2023:name.xEnvoy Air                9.80e-05 ***
## factor(year)2023:name.xExpressJet Airlines Inc.       NA
## factor(year)2023:name.xFrontier Airlines Inc.   0.077041 .
## factor(year)2023:name.xHawaiian Airlines Inc.   0.052195 .
## factor(year)2023:name.xJetBlue Airways          6.59e-05 ***
## factor(year)2023:name.xMesa Airlines Inc.             NA
## factor(year)2023:name.xSkyWest Airlines Inc.    0.315354
## factor(year)2023:name.xSouthwest Airlines Co.    < 2e-16 ***
## factor(year)2023:name.xUnited Air Lines Inc.          NA
## factor(year)2023:name.xUS Airways Inc.                NA
## factor(year)2023:name.xVirgin America                 NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4777 on 192181 degrees of freedom
##   (18418 observations deleted due to missingness)
## Multiple R-squared:  0.02332,    Adjusted R-squared:  0.02319
## F-statistic: 176.5 on 26 and 192181 DF,  p-value: < 2.2e-16
```

Based on the F-statistic, this model is significant. However, there are a few airlines that seem to be discontinued in 2023, so lets remove them and create a new model.

```r
# getting airlines that only appear in both years
active_airlines <- flights_clean_log %>%
  group_by(name.x, year) %>%
  summarise(n = n(), .groups = "drop") %>%
  count(name.x) %>%
  filter(n == 2) %>%
  pull(name.x)

flights_filtered <- flights_clean_log %>%
  filter(name.x %in% active_airlines)


dep_delay_model_airline_filtered <- lm(data=flights_filtered, log_dep_delay~factor(year) * name.x)
```

Now with our new model, let's check for the normality assumption.

```r
dep_delay_airline_resids <- sample(residuals(dep_delay_model_airline_filtered), size = 30000)

# Now plot the Q-Q plot with the sample for easier loading
qqnorm(dep_delay_airline_resids)
qqline(dep_delay_airline_resids, col = "red")
```



Based on this plot, we can say we meet the normality assumption. Now let's check for homoscedasticity with the Levene's Test.

```r
leveneTest(log_dep_delay ~ factor(year) * name.x, data = flights_filtered)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##           Df F value    Pr(>F)
## group     21  78.152 < 2.2e-16 ***
##       164960
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the small p-value, we can see signs of heteroscedasticity which means we cannot use regular 'summary()' to get results of the model. Let's look into the model itself using robust standard errors which assumes unequal error variances.

```r
dep_delay_robust_se <- vcovHC(dep_delay_model_airline_filtered, type = "HC1")
dep_delay_tidy_robust_model <- tidy(coeftest(dep_delay_model_airline_filtered, vcov. = dep_delay_robust_

# only getting interaction terms, what we need
dep_interaction_terms <- dep_delay_tidy_robust_model[grep(":", dep_delay_tidy_robust_model$term),]
print(dep_interaction_terms)
```

```
## # A tibble: 10 x 5
##    term                                     estimate std.error statistic p.value
##    <chr>                                       <dbl>     <dbl>     <dbl>   <dbl>
##  1 factor(year)2023:name.xAmerican Airline~   0.0352    0.0432     0.816 0.415
##  2 factor(year)2023:name.xDelta Air Lines ~   0.0445    0.0428     1.04  0.299
##  3 factor(year)2023:name.xEndeavor Air Inc.  -0.129     0.0431    -3.00  0.00271
##  4 factor(year)2023:name.xEnvoy Air          -0.113     0.0525    -2.15  0.0317
##  5 factor(year)2023:name.xFrontier Airline~   0.129     0.0563     2.28  0.0225
##  6 factor(year)2023:name.xHawaiian Airline~  -0.0936    0.117     -0.802 0.422
##  7 factor(year)2023:name.xJetBlue Airways     0.0933    0.0426     2.19  0.0287
##  8 factor(year)2023:name.xSkyWest Airlines~  -0.103     0.168     -0.614 0.539
##  9 factor(year)2023:name.xSouthwest Airlin~  -0.108     0.0434    -2.49  0.0129
## 10 factor(year)2023:name.xUnited Air Lines~   0.0668    0.0426     1.57  0.117
```

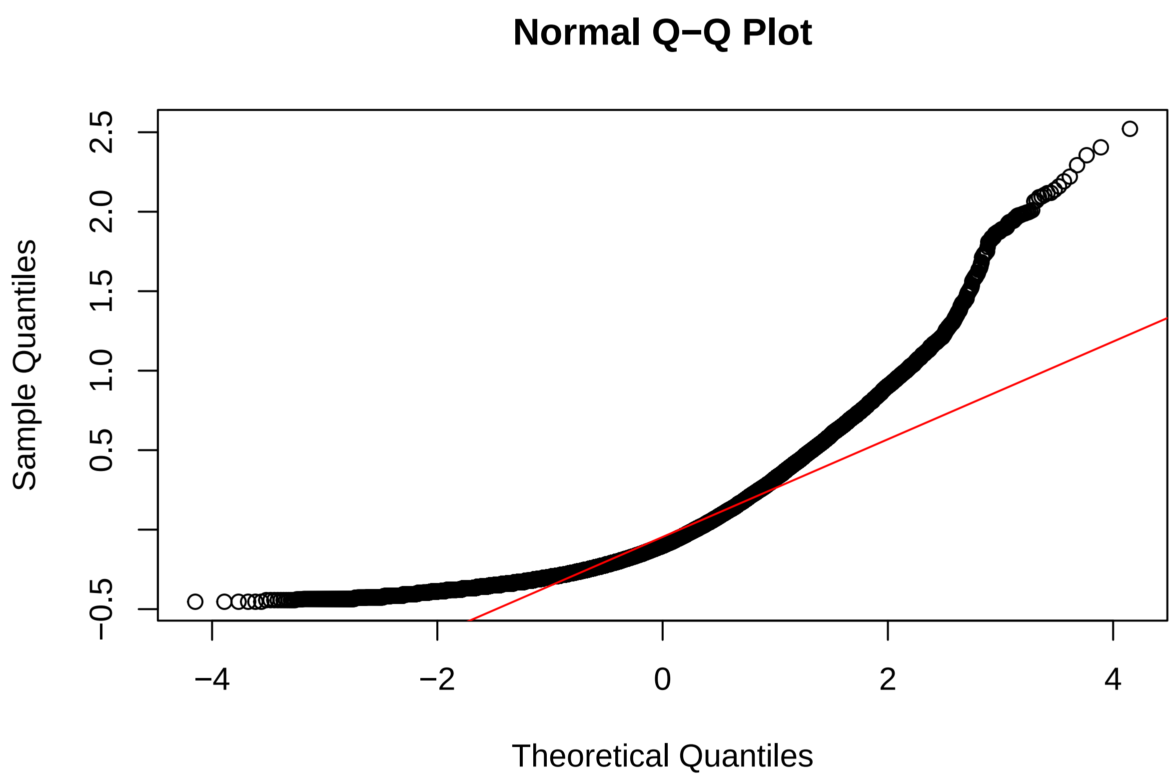
Based on the p-values, only 3 of these 10 airlines are statistically significant. Yet despite that we can still gain an idea of a general trend by looking at the coefficients. 6 out of 10 of these coefficients are positive, which means the majority of the airlines have gotten worse in 2023 compared to 2013.

Let's do this again with arrival delays.

```r
arr_delay_model_airline_filtered <- lm(data=flights_filtered, log_arr_delay~factor(year) * name.x)
```

Let's check for the normality assumption.

```r
arr_delay_airline_resids <- sample(residuals(arr_delay_model_airline_filtered), size = 30000)

# Now plot the Q-Q plot with the sample for easier loading
qqnorm(arr_delay_airline_resids)
qqline(arr_delay_airline_resids, col = "red")
```



Based on this plot, we can say we meet the normality assumption. Now let's check for homoscedasticity with the Levene's Test.

```
leveneTest(log_arr_delay ~ factor(year) * name.x, data = flights_filtered)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##            Df F value    Pr(>F)
## group      21  89.975 < 2.2e-16 ***
##        164960
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the small p-value, we can see signs of heteroscedasticity which means we cannot use regular 'summary()' to get results of the model. Let's look into the model itself using robust standard errors which assumes unequal error variances.

```
arr_delay_robust_se <- vcovHC(arr_delay_model_airline_filtered, type = "HC1")
arr_delay_tidy_robust_model <- tidy(coeftest(arr_delay_model_airline_filtered, vcov. = arr_delay_robust_

# only getting interaction terms, what we need
arr_interaction_terms <- arr_delay_tidy_robust_model[grep(":", arr_delay_tidy_robust_model$term),]
print(arr_interaction_terms)
```

```
## # A tibble: 10 x 5
##    term                                       estimate std.error statistic p.value
```

```
##    <chr>                                      <dbl>     <dbl>     <dbl>   <dbl>
##  1 factor(year)2023:name.xAmerican Airline~  0.00770    0.0271     0.284 7.76e-1
##  2 factor(year)2023:name.xDelta Air Lines ~ -0.0189     0.0268    -0.706 4.80e-1
##  3 factor(year)2023:name.xEndeavor Air Inc. -0.102      0.0271    -3.78  1.58e-4
##  4 factor(year)2023:name.xEnvoy Air         -0.181      0.0332    -5.46  4.73e-8
##  5 factor(year)2023:name.xFrontier Airline~ -0.0161     0.0380    -0.425 6.71e-1
##  6 factor(year)2023:name.xHawaiian Airline~ -0.0663     0.0771    -0.860 3.90e-1
##  7 factor(year)2023:name.xJetBlue Airways    0.0172     0.0267     0.645 5.19e-1
##  8 factor(year)2023:name.xSkyWest Airlines~ -0.136      0.110     -1.24  2.17e-1
##  9 factor(year)2023:name.xSouthwest Airlin~ -0.108      0.0272    -3.95  7.69e-5
## 10 factor(year)2023:name.xUnited Air Lines~  0.00695    0.0267     0.261 7.94e-1
```

Based on the p-values here, again only 3 of these 10 airlines are statistically significant. We can still gain an idea of a general trend by looking at the coefficients. 6 out of 10 of these coefficients are negative, opposite of the departure trends. This means the majority of the airlines have actually gotten better in 2023 compared to 2013.

Graphs

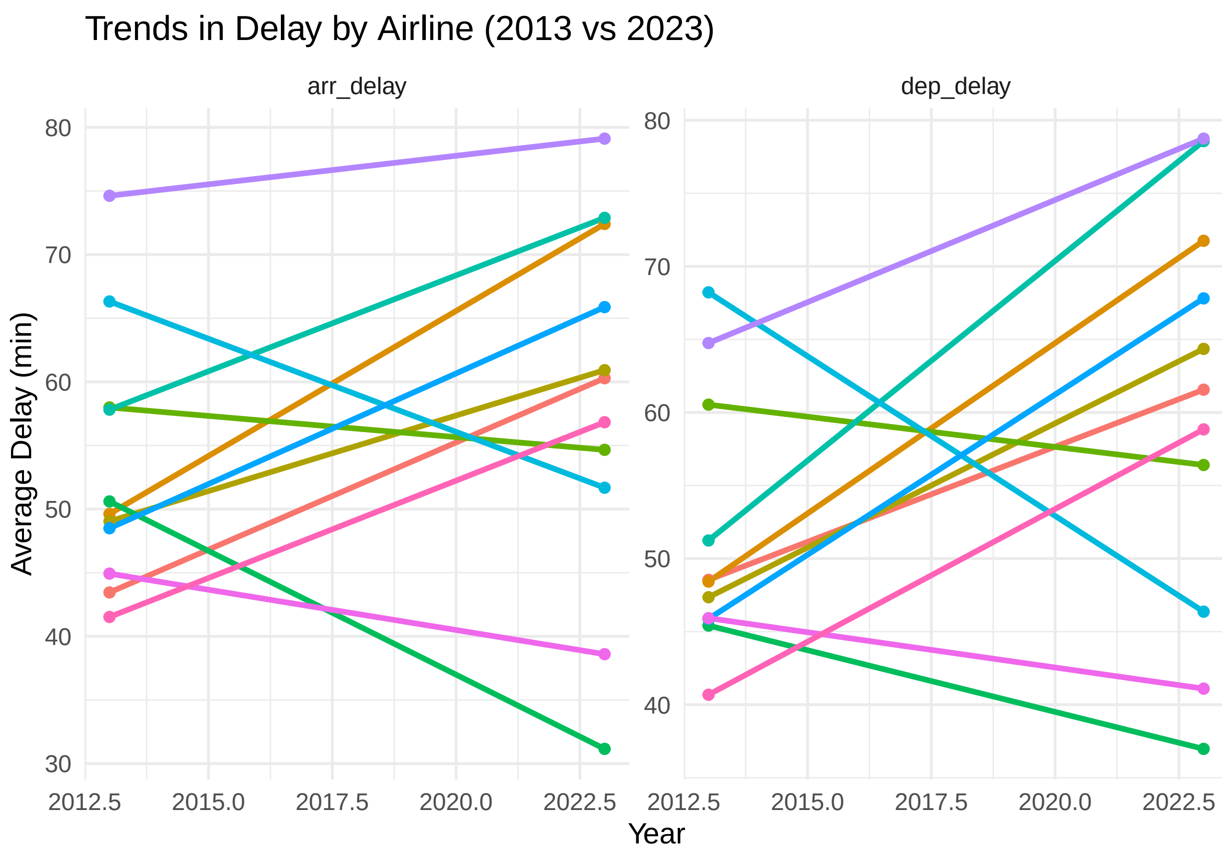
```r
both_delay_filtered <- flights_filtered %>%
  select(year, name.x, dep_delay, arr_delay) %>%
  pivot_longer(cols = c(dep_delay, arr_delay),
               names_to = "delay_type",
               values_to = "delay_value")
```

Here, we can easily how the average arrival and departure delays have changed between 2013 and 2023. It seems that both arrival and departure delays have gotten worse.

```r
both_delay_filtered %>%
  group_by(name.x, year, delay_type) %>%
  summarise(mean_delay = mean(delay_value, na.rm = TRUE)) %>%
  ggplot(aes(x = year, y = mean_delay, color = name.x, group = name.x)) +
  geom_line(size = 1) +
  geom_point() +
  facet_wrap(~delay_type, scales = "free_y") +
  labs(title = "Trends in Delay by Airline (2013 vs 2023)",
       x = "Year", y = "Average Delay (min)") +
  theme_minimal() +
  theme(legend.position = "none")
```

```
## `summarise()` has grouped output by 'name.x', 'year'. You can override using
## the `.groups` argument.
```

```
## 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.
```



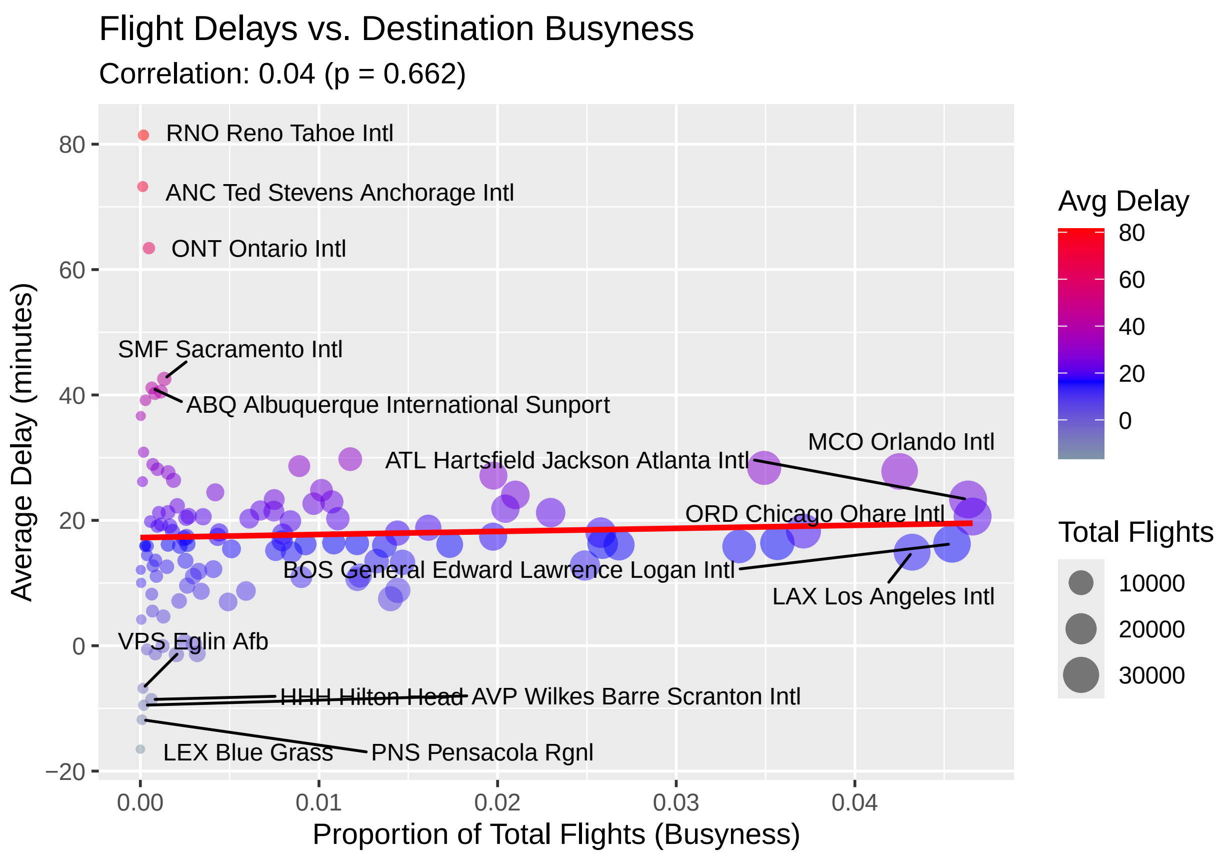
Do busy destinations tend to have more or less delays?

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) +
```

```r
  #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

```r
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
```

```r
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

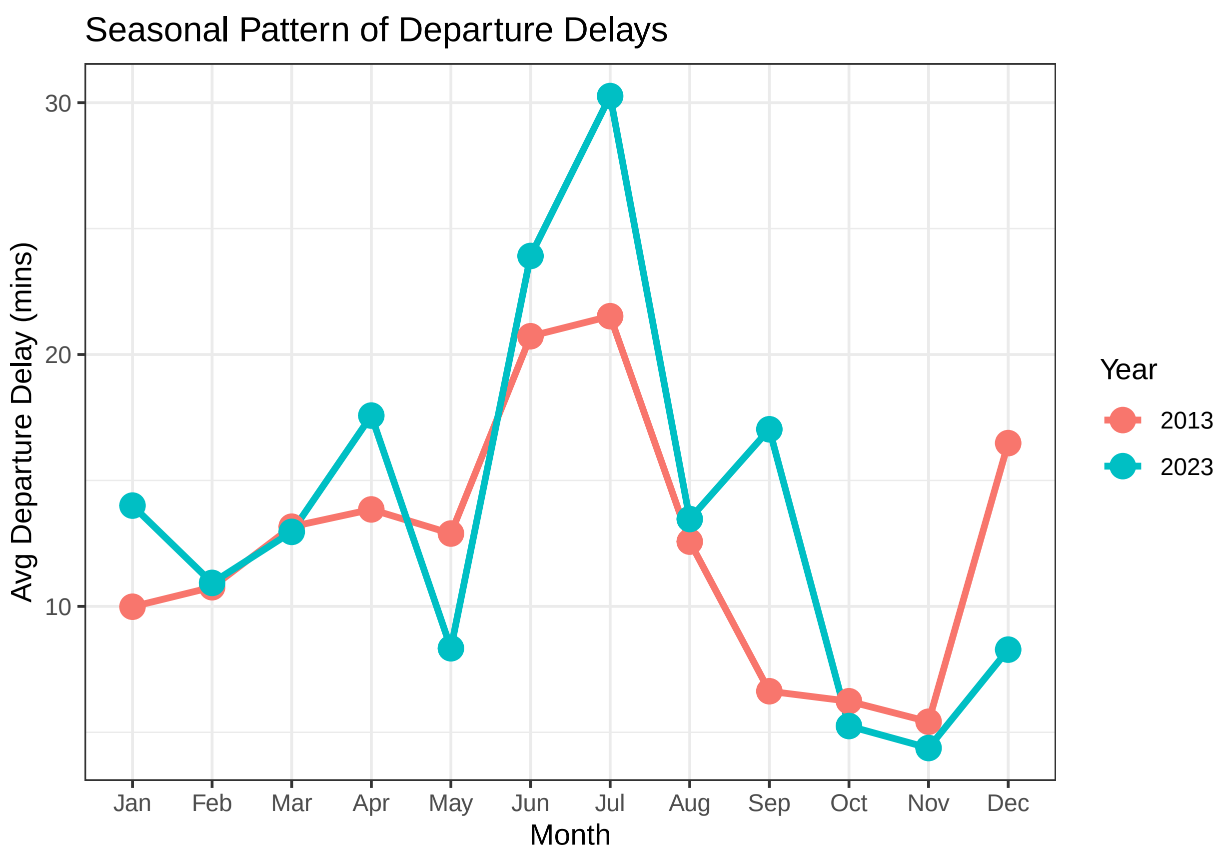
```r
# code for testing your hypotheses/models

# 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)", co
  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]

[REPLACE WITH QUESTION #5]

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?]

Conclusions

1. **Have flight delays improved over time overall?**

    - **What about with individual airlines?**

From 2013 to 2023, both departure and arrival delays generally worsened, with departure delays showing a more noticeable increase. When looking at individual airlines, SkyWest, American, and Delta showed no significant change in either type of delay, suggesting stable performance over time. Frontier and JetBlue experienced significant increases in departure delays, while United showed a smaller, non-significant increase. Southwest Airlines significantly improved arrival delays, with possible improvement in departures as well. Envoy Air saw no change in departure delays but did show significant improvement in arrivals. Notably, Endeavor Air was the only airline to significantly improve in both departure and arrival delays.

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	Question 1
Ryan Huang	Question 2
Isaac Kang	