Statistical Analysis of NYC Flight Delays

Lydia Niu, Alexis Castaneda, Aparna Petluri, Gracelynne Mohan, Jenny Zhang, Zoe Shum

Introduction

Handling the heavy air traffic above New York, the John F. Kennedy International Airport (JFK), LaGuardia Airport (LGA), and Newark Liberty International Airport (EWR) manage immense flight operations throughout the year. Given a dataset regarding the dynamic flight patterns between these three airports, we hope to explore the various factors that may affect flight delays and flight volume - amount of flights.

The overall question we want to answer is: **"*What factors influence flight volume and does this affect delay patterns across New York City's major airports (JFK, LGA, and EWR)?*"** We hypothesize that weather conditions, such as temperature or wind speed, play a significant role in the occurrence of delays among the three airports. By analyzing the nycflights13 dataset, we aim to provide valuable insight to travelers, as well as airline and airport administrators, as to flight frequency and delays in the New York metropolitan area.

Coherent Questions

To help answer our main question, we aim to answer the following:

1. Which **months and seasons** experience the highest **flight volumes** at each airport?
2. How do **average delays** vary by **month and season**?
3. Are **delays** more frequent/severe during **specific weather conditions**?
4. Are there significant differences in **average delays/ flight volume** across the 3 airports and across **different seasons**?
5. What relationship, if any, exists between **busy days (high flight volume)**, **weather conditions** (temperature and wind speed) and **flight delays**?
6. Which **airport** is most affected by **flight delays** due to **weather conditions** among JFK, LGA, and EWR?

Data Description

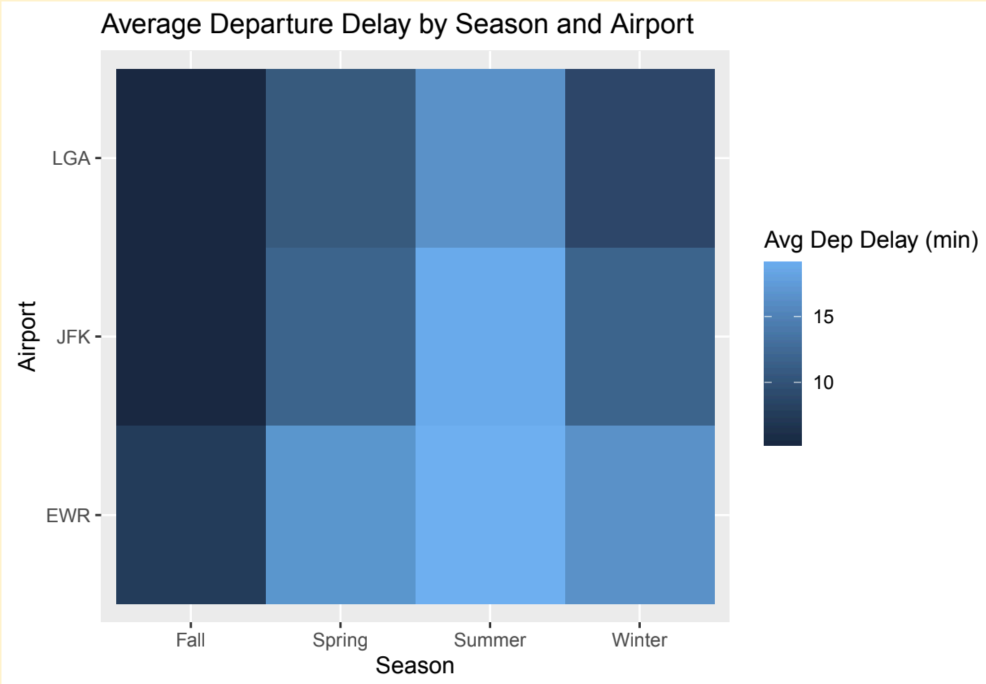
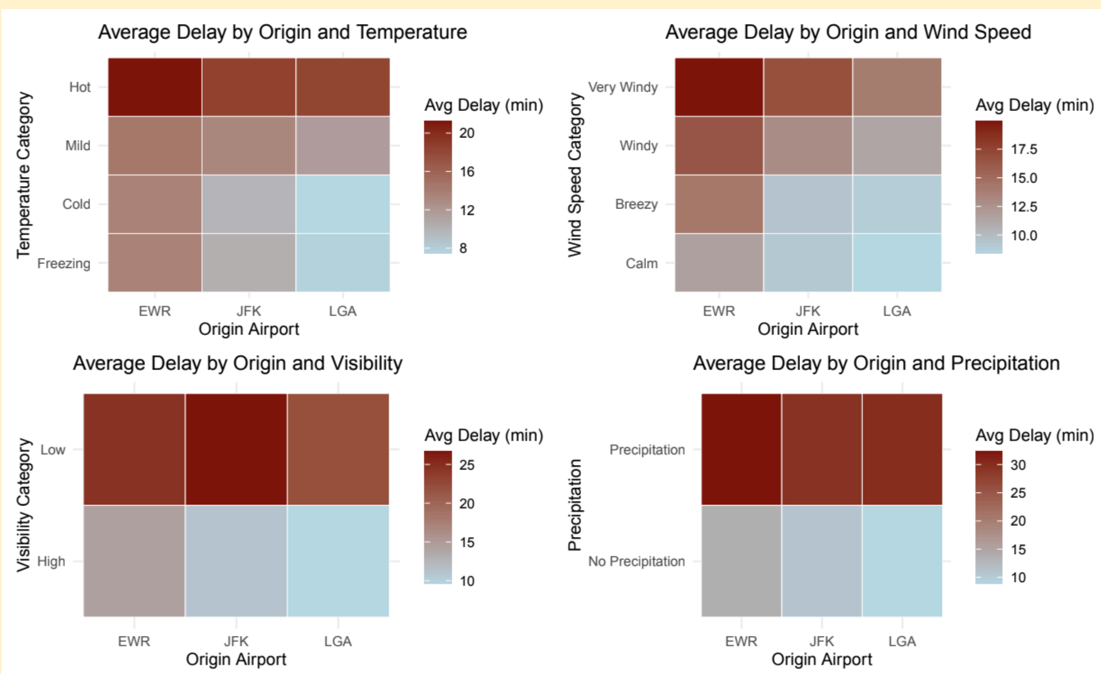
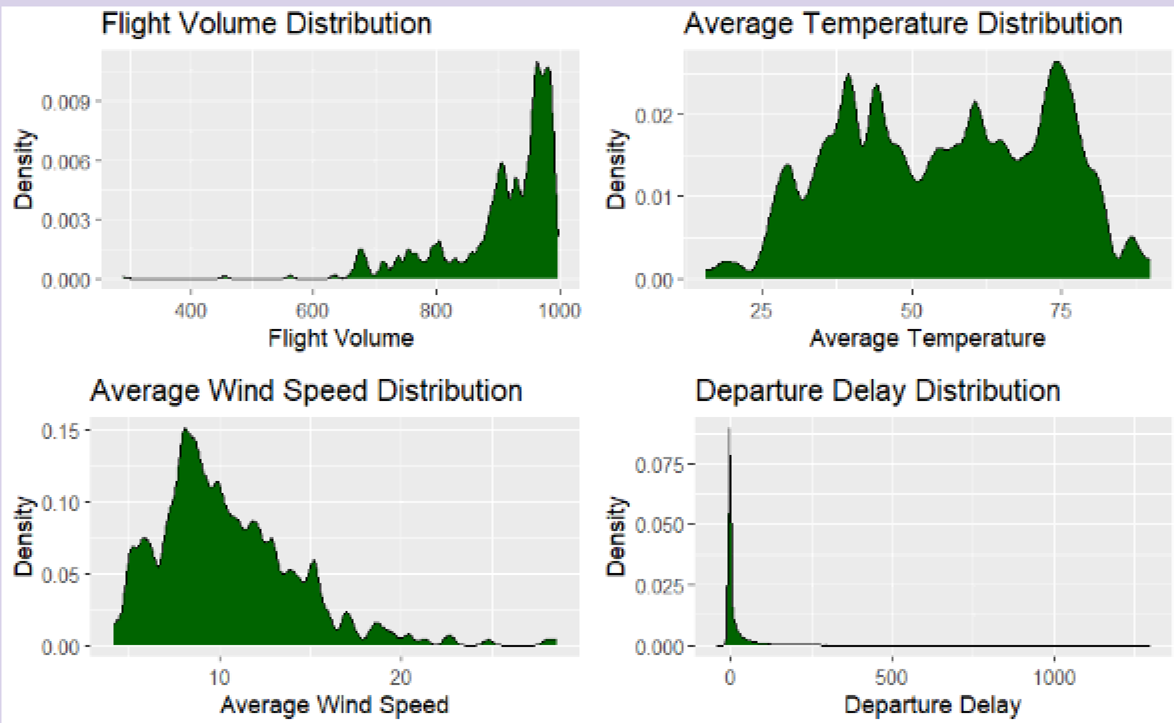
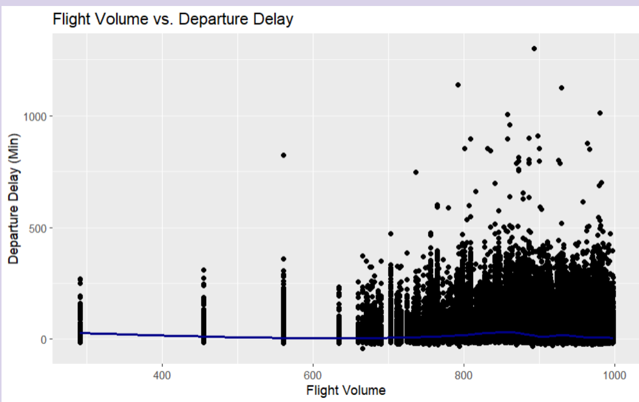
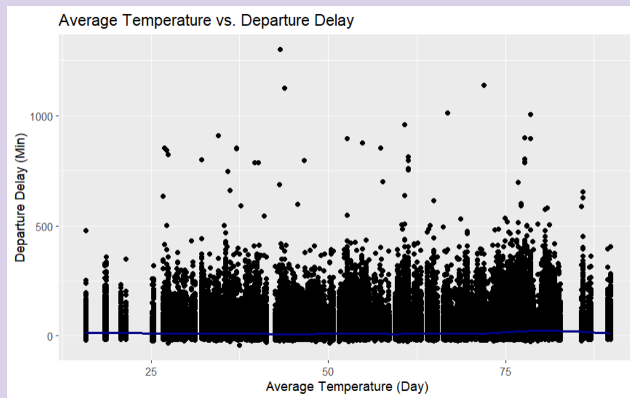
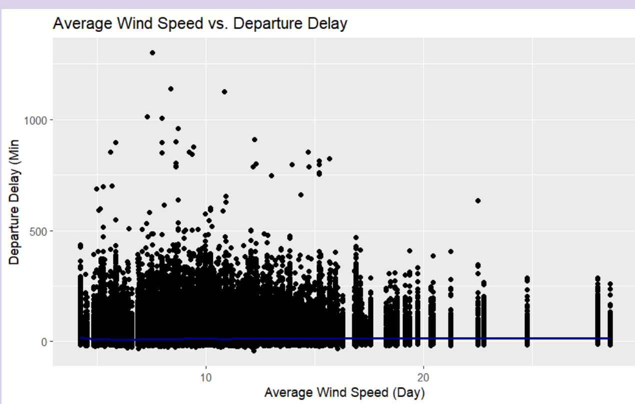
For this project, we will be focusing our analysis on the flights, airports, and weather datasets found in the nycflights13 tidyverse package.

Flights: all flights that departed from NYC in 2013. Size: 336,776 x 19
Weather: hourly meteorological data for each airport. Size: 26,115 x 15

As part of our data preprocessing, we removed all canceled flights and filtered out any entries with missing values related to weather variables. We also excluded data from December 31, which was missing wind speed and temperature information. After cleaning the data, we created new data frames containing calculated statistics to support a more structured analysis. These steps allowed us to explore the correlations between environmental factors and operational efficiency at New York City's airports.

<Provide a glimpse of the first few rows of datasets we will use>

Visualizations

````
```{r, zoe, echo=FALSE}
flights_cleaned <- flights %>%
  filter(origin %in% c("EWR", "JFK", "LGA")) %>%
  filter(!is.na(dep_delay), !is.na(arr_delay))

flights_weather <- flights_cleaned %>%
  left_join(weather, by = c("origin", "year", "month", "day", "hour")) %>%
  mutate(delay_status = ifelse(dep_delay > 15, "Delayed", "On-Time")) %>%
  filter(!is.na(precip), !is.na(wind_speed), !is.na(visib), !is.na(temp))
```
```{r, zoe, echo=FALSE}
seasonal_summary <- flights_cleaned %>%
  mutate(season = case_when(
    month %in% c(12, 1, 2) ~ "Winter",
    month %in% c(3, 4, 5) ~ "Spring",
    month %in% c(6, 7, 8) ~ "Summer",
    month %in% c(9, 10, 11) ~ "Fall"
  )) %>%
  group_by(origin, season) %>%
  summarise(
    flights_count = n(),
    delay_count = sum(dep_delay >= 15),
    avg_delay = mean(dep_delay),
    .groups = "drop"
  )
seasonal_summary

ggplot(seasonal_summary, aes(x = season, y = origin, fill = avg_delay)) +
  geom_tile() +
  labs(title = "Average Departure Delay by Season and Airport",
       x = "Season", y = "Airport", fill = "Avg Dep Delay (min)")
````
````

```
```




```
```{r, zoe, fig.width = 10, fig.height = 6, echo=FALSE}
# Summarize average delay by origin and temperature category
flights_temp <- flights_weather %>%
  mutate(temp_cat = factor(case_when(
    temp < 32 ~ "Freezing",
    temp < 60 ~ "Cold",
    temp < 80 ~ "Mild",
    TRUE ~ "Hot"
  ), levels = c("Freezing", "Cold", "Mild", "Hot")))

heatmap_data <- flights_temp %>%
  group_by(origin, temp_cat) %>%
  summarise(avg_delay = mean(dep_delay, na.rm = TRUE), .groups = "drop")

# Create the heatmap
hm1 <- ggplot(heatmap_data, aes(x = origin, y = temp_cat, fill = avg_delay)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightblue", high = "darkred", name = "Avg Delay (min)") +
  labs(title = "Average Delay by Origin and Temperature",
       x = "Origin Airport", y = "Temperature Category") +
```
```

```
  theme_minimal()

# Summarize average delay by origin and wind speed category
flights_wind_speed <- flights_weather %>%
  mutate(wind_cat = factor(case_when(
    wind_speed < 5 ~ "Calm",
    wind_speed < 10 ~ "Breezy",
    wind_speed < 20 ~ "Windy",
    TRUE ~ "Very Windy"
  ), levels = c("Calm", "Breezy", "Windy", "Very Windy")))

heatmap_data_wind <- flights_wind_speed %>%
  group_by(origin, wind_cat) %>%
  summarise(avg_delay = mean(dep_delay, na.rm = TRUE), .groups = "drop")

# Create the heatmap
hm2 <- ggplot(heatmap_data_wind, aes(x = origin, y = wind_cat, fill = avg_delay)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightblue", high = "darkred", name = "Avg Delay (min)") +
  labs(title = "Average Delay by Origin and Wind Speed",
       x = "Origin Airport", y = "Wind Speed Category") +
  theme_minimal()

# Summarize average delay by origin and visibility category
flights_visib <- flights_weather %>%
  mutate(visib_cat = case_when(
    visib < 5 ~ "Low",
    TRUE ~ "High"
  ))

heatmap_data_visib <- flights_visib %>%
  group_by(origin, visib_cat) %>%
  summarise(avg_delay = mean(dep_delay, na.rm = TRUE), .groups = "drop")

# Create the heatmap
hm3 <- ggplot(heatmap_data_visib, aes(x = origin, y = visib_cat, fill = avg_delay)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightblue", high = "darkred", name = "Avg Delay (min)") +
  labs(title = "Average Delay by Origin and Visibility",
       x = "Origin Airport", y = "Visibility Category") +
  theme_minimal()

# Summarize average delay by origin and precipitation category
```

```
flights_precip <- flights_weather %>%
  mutate(precip_cat = ifelse(precip > 0, "Precipitation", "No Precipitation"))

heatmap_data_precip <- flights_precip %>%
  group_by(origin, precip_cat) %>%
  summarise(avg_delay = mean(dep_delay, na.rm = TRUE), .groups = "drop")

# Create the heatmap
hm4 <- ggplot(heatmap_data_precip, aes(x = origin, y = precip_cat, fill = avg_delay)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightblue", high = "darkred", name = "Avg Delay (min)") +
  labs(title = "Average Delay by Origin and Precipitation",
       x = "Origin Airport", y = "Precipitation") +
  theme_minimal()

# Arrange in a grid (2 rows, 2 columns; 1 blank)
grid.arrange(hm1, hm2, hm3, hm4, ncol = 2)
```



Initial visualizations indicated seasonal trends, with summer showing the highest volume and severity of delays. Weather-based plots revealed that conditions like high precipitation, reduced visibility, and stronger winds cluster in specific months, potentially contributing to delay patterns. To explore this further, heatmaps were created showing the average departure delay by origin airport across various weather categories:

- **Temperature category:** Delays were highest during “Hot” conditions, especially at EWR.
- **Wind speed category:** Delays increased with wind intensity, peaking under “Very Windy” conditions.
- **Visibility category:** Lower visibility corresponded with longer average delays across all airports.
- **Precipitation presence:** Flights during precipitation consistently experienced higher delays.

````
```{r, alexis and aparna, echo=FALSE}
flights_cleaned <- flights |>
  filter(origin %in% c("EWR", "JFK", "LGA")) |>
  filter(!is.na(dep_delay), !is.na(arr_delay))

# Saves flights daily volume
flights_volume <- flights_cleaned |>
  group_by(month, day) |>
  summarise(flight_volume = n(),
            .groups = "drop")

# Saves flights daily weather
flights_weather <- weather |>
  filter(!is.na(temp), !is.na(wind_speed)) |>
  group_by(month, day) |>
  summarise(avg_temp = mean(temp),
            avg_wind_speed = mean(wind_speed),
            .groups = "drop")

# Combines them into the original flights_cleaned
flights_combined <- left_join(flights_cleaned, flights_volume, by = join_by(month, day))

flights_combined <- left_join(flights_combined, flights_weather, by = join_by(month, day))

# Weather doesn't have Dec 31 recorded
flights_combined_filtered <- flights_combined |> filter(!(month == 12 & day == 31),
                                                  !is.na(avg_wind_speed), !is.na(avg_temp),
                                                  !is.na(dep_delay), !is.na(flight_volume))
```
```{r, alexis and aparna, echo=FALSE}
Flight_Vol_Density <- ggplot(data = flights_combined_filtered,
       mapping = aes(x = flight_volume)) +
  geom_density(fill = "darkgreen") +
  labs(title = "Flight Volume Distribution", x = "Flight Volume", y = "Density")

Avg_Temp_Density <- ggplot(data = flights_combined_filtered,
       mapping = aes(x = avg_temp)) +
  geom_density(fill = "darkgreen") +
````
````

```
  labs(title = "Average Temperature Distribution", x = "Average Temperature", y = "Density")

Avg_Windspeed_Density<- ggplot(data = flights_combined_filtered,
       mapping = aes(x = avg_wind_speed)) +
  geom_density(fill = "darkgreen") +
  labs(title = "Average Wind Speed Distribution", x = "Average Wind Speed", y = "Density")

Dep_Delay_Density <- ggplot(data = flights_combined_filtered,
       mapping = aes(x = dep_delay)) +
  geom_density(fill = "darkgreen") +
  labs(title = "Departure Delay Distribution", x = "Departure Delay", y = "Density")

grid.arrange(Flight_Vol_Density, Avg_Temp_Density, Avg_Windspeed_Density, Dep_Delay_Density,
nrow = 2)
```




```
```{r, alexis and aparna, echo = FALSE}
ggplot(data = flights_combined_filtered,
       mapping = aes(x = flight_volume, y = dep_delay)) +
  geom_point() +
  geom_smooth(se = F, color = "darkblue") +
  labs(title = "Flight Volume vs. Departure Delay",
       x = "Flight Volume", y = "Departure Delay (Min)")

ggplot(data = flights_combined_filtered,
```
```

```
      mapping = aes(x = avg_temp, y = dep_delay)) +
  geom_point() +
  geom_smooth(se = F, color = "darkblue") +
  labs(title = "Average Temperature vs. Departure Delay", x = "Average Temperature (Day)", y =
"Departure Delay (Min)")

ggplot(data = flights_combined_filtered,
      mapping = aes(x = avg_wind_speed, y = dep_delay)) +
  geom_point() +
  geom_smooth(se = F, color = "darkblue") +
  labs(title = "Average Wind Speed vs. Departure Delay", x = "Average Wind Speed (Day)", y =
"Departure Delay (Min)")
```







Initial plotting of flight volume, average temperature, average wind speed, and departure delay show that the distribution of flight volume is strongly left skewed, the distribution of average wind speed is moderately right skewed, and the distribution of departure delay is extremely right skewed. The severity of departure delay's right skew could potentially impact our findings.

To examine departure delay's effects further, scatterplots were developed to plot each predictor variable (flight volume, average temperature, average wind speed) against departure delay. The scatterplots show that there is little to no correlation relationship in all 3 plots. However, we will determine whether there is minor or no correlation with further testing.

````
## Data Cleaning

```{r}
nyc_flights <- flights %>%
  filter(origin %in% c("EWR", "JFK", "LGA")) %>%
  filter(!is.na(dep_delay), !is.na(arr_delay)) # remove NA values

nyc_flights
```

## EDA

```{r}
flights_weather <- nyc_flights %>%
  filter(origin %in% c("JFK", "LGA", "EWR")) %>%
  left_join(weather, by = c("origin", "time_hour"))

flights_weather_filtered <- flights_weather %>%
  select(origin, arr_delay, dep_delay, precip, wind_speed, visib)

flights_weather_filtered
```

```{r}
flights_weather_filtered <- flights_weather_filtered %>%
  filter(!is.na(precip), !is.na(wind_speed), !is.na(visib)) %>%
  mutate(
    bad_weather = ifelse(precip > 0.5 | wind_speed > 15 | visib < 2, "Bad", "Good")
  )

delay_summary <- flights_weather_filtered %>%
  group_by(origin, bad_weather) %>%
  summarise(
    avg_arr_delay = mean(arr_delay, na.rm = TRUE),
    avg_dep_delay = mean(dep_delay, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  filter(!is.na(bad_weather))

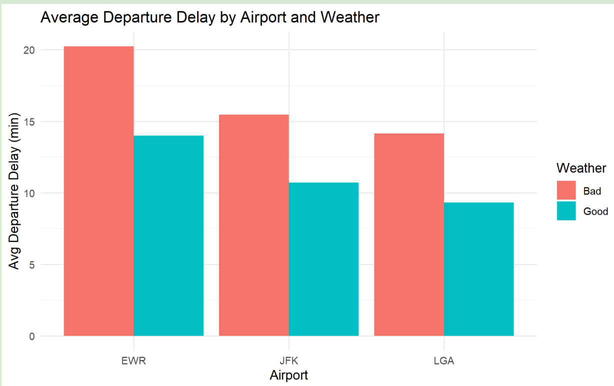
delay_summary

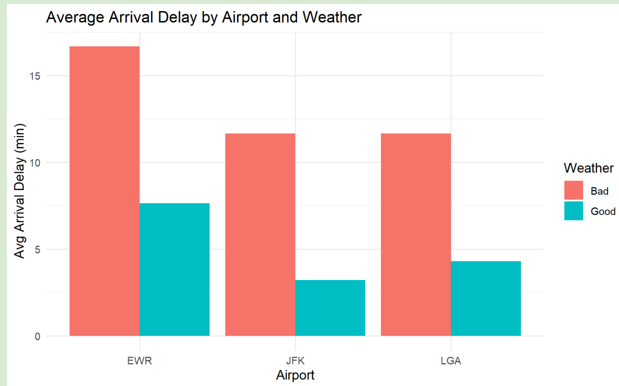
ggplot(delay_summary, aes(x = origin, y = avg_dep_delay, fill = bad_weather)) +
  geom_col(position = "dodge") +
  labs(title = "Average Departure Delay by Airport and Weather",
       x = "Airport", y = "Avg Departure Delay (min)", fill = "Weather") +
  theme_minimal()

ggplot(delay_summary, aes(x = origin, y = avg_arr_delay, fill = bad_weather)) +
  geom_col(position = "dodge") +
````
````

````
labs(title = "Average Arrival Delay by Airport and Weather",
     x = "Airport", y = "Avg Arrival Delay (min)", fill = "Weather") +
  theme_minimal()

```
````





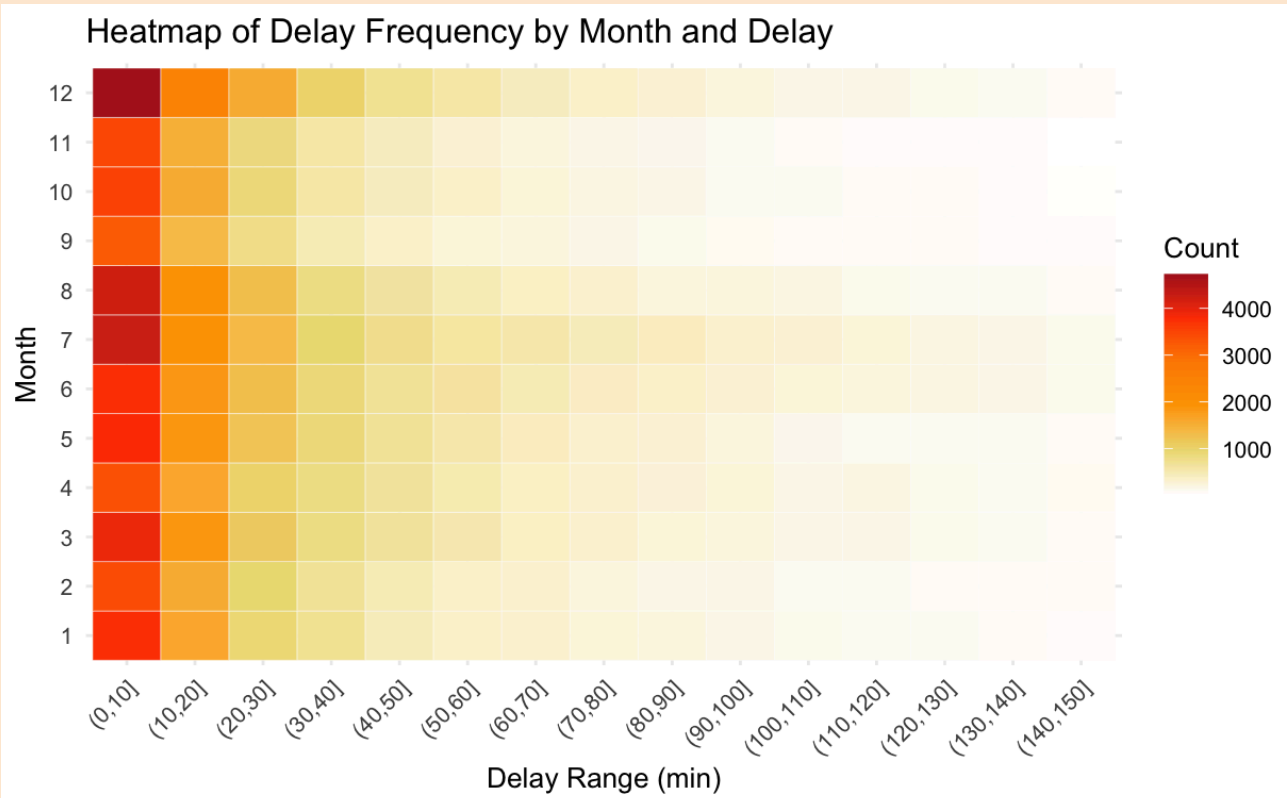
To analyze which of the three airports was most affected by flight delays due to weather conditions, we created two bar plots. In order to do so, we grouped the three weather conditions under one variable “Bad Weather”, which was defined as precipitation greater than 0.5 inches, wind speed exceeding 15 miles per hour, and visibility less than 2 miles. If the weather didn’t meet those conditions, it was considered “Good Weather”. Using this classification, we made two bar plots to compare the average departure and arrival delays at each airport. From the plots, we find that EWR experiences the highest level of delays overall compared to JFK and LGA, with an increase in delays in bad weather.

````
```{r, jenny, echo = FALSE}

flights_clean <- flights %>%
  filter(origin %in% c("EWR", "JFK", "LGA")) %>%
  filter(!is.na(dep_delay), !is.na(arr_delay))

flights_clean %>%
  filter(dep_delay >= 0, dep_delay <= 150) %>%
  mutate(delay_bin = cut(dep_delay, breaks = seq(0, 150, by = 10))) %>%
  count(month, delay_bin) %>%
  filter( !is.na(delay_bin)) %>%
  ggplot(aes(x = delay_bin, y = factor(month), fill = n)) +
  geom_tile(color = "white") +
  scale_fill_gradientn(
    colors = c("white","lightgoldenrod", "orange", "darkorange", "orangered", "firebrick"),
    name = "Count"
  ) +
  labs(
    title = "Heatmap of Delay Frequency by Month and Delay",
````

```
    x = "Delay Range (min)",
    y = "Month"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Months with the highest amount of delays ranging 1-10 minutes can be identified as the months of July, August and December.

As we move further down the heatmap, we see that June, July, and December still maintain colors between “white” and “lightgoldenrod” all the way down to the 90-100 interval, showing how these months may also have longer delays.

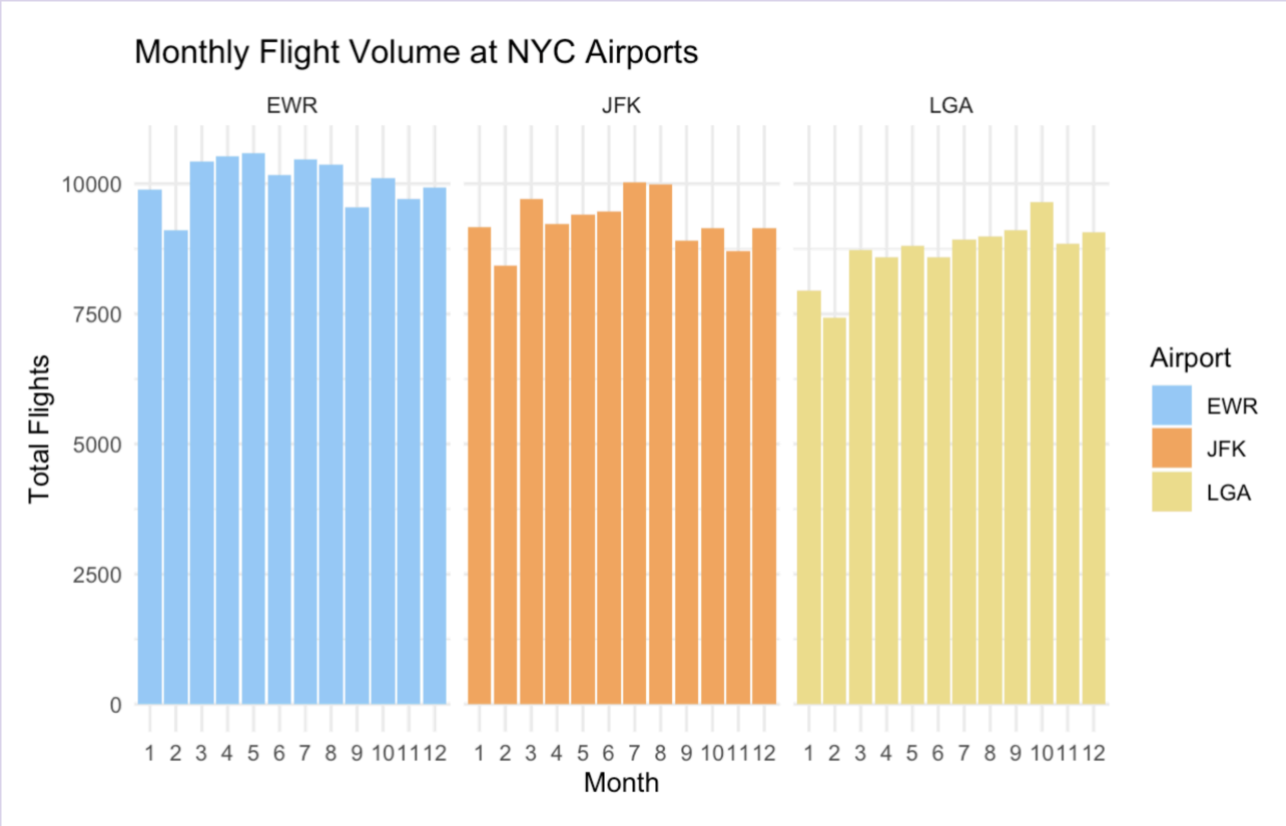
The fall months September, October, and November tend to have less of the longer delays seeing that the 90-100 end of their part of the graph is close to “white”.

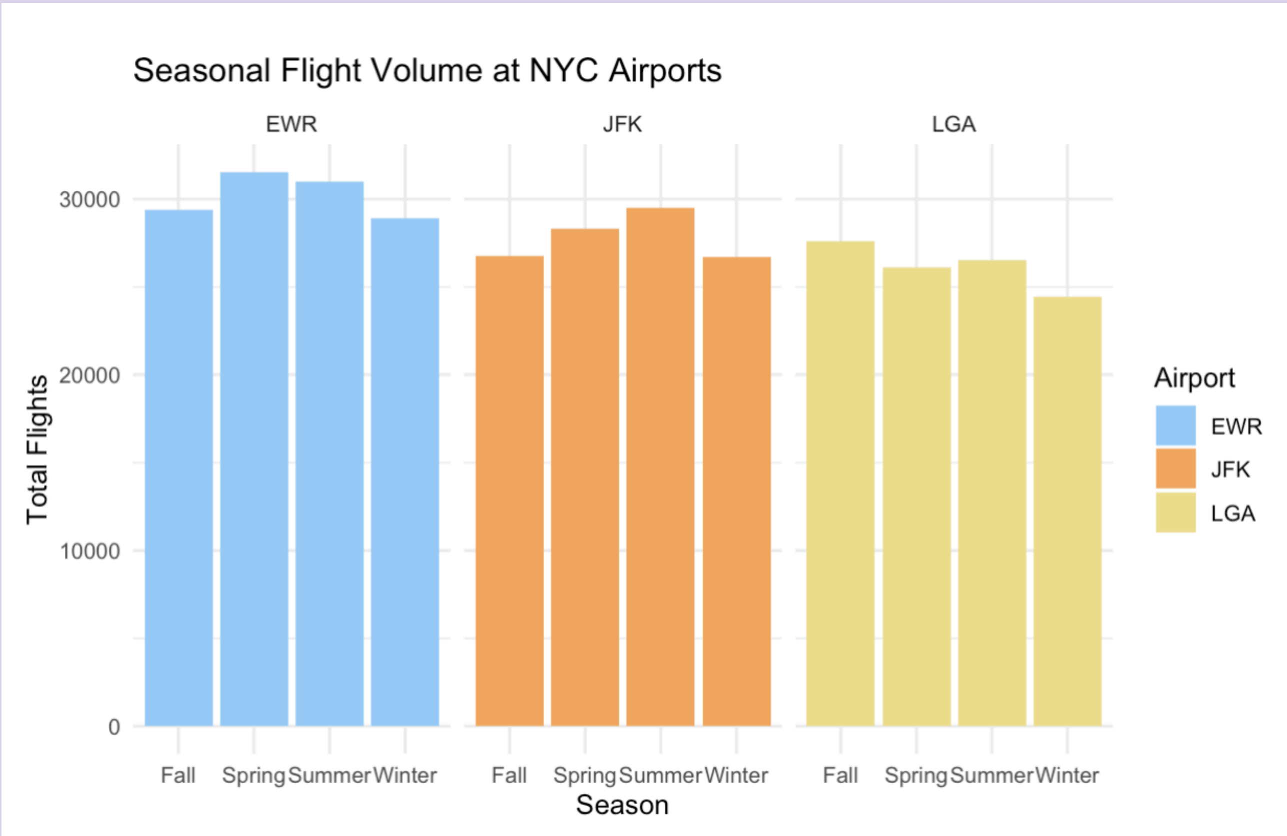
Data Analysis

Question 1

To get a general idea of the data and spread of flight volume over the months and seasons, bar charts were created to visualize the trends.

```
```{r jenny, echo=FALSE}
# classify seasons
flights_season <- flights %>%
  mutate(season = case_when(
    month %in% c(12, 1, 2) ~ "Winter",
    month %in% c(3, 4, 5) ~ "Spring",
    month %in% c(6, 7, 8) ~ "Summer",
    month %in% c(9, 10, 11) ~ "Fall"
  ))

# summarize by airport, month, and season
flight_counts <- flights_season %>%
  group_by(origin, month, season) %>%
  summarise(total_flights = n(), .groups = "drop")

# monthly volume
ggplot(flight_counts, aes(x = factor(month), y = total_flights, fill = origin)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = c("JFK" = "tan1", "LGA" = "lightgoldenrod", "EWR" = "lightskyblue")) +
  facet_wrap(~ origin) +
  labs(title = "Monthly Flight Volume at NYC Airports",
       x = "Month", y = "Total Flights", fill = "Airport") +
  theme_minimal()

# seasonal volume
seasonal_counts <- flights_season %>%
  group_by(origin, season) %>%
  summarise(total_flights = n(), .groups = "drop")

ggplot(seasonal_counts, aes(x = season, y = total_flights, fill = origin)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = c("JFK" = "tan1", "LGA" = "lightgoldenrod", "EWR" = "lightskyblue")) +
  facet_wrap(~ origin) +
  labs(title = "Seasonal Flight Volume at NYC Airports",
       x = "Season", y = "Total Flights", fill = "Airport") +
  theme_minimal()
```
```

```



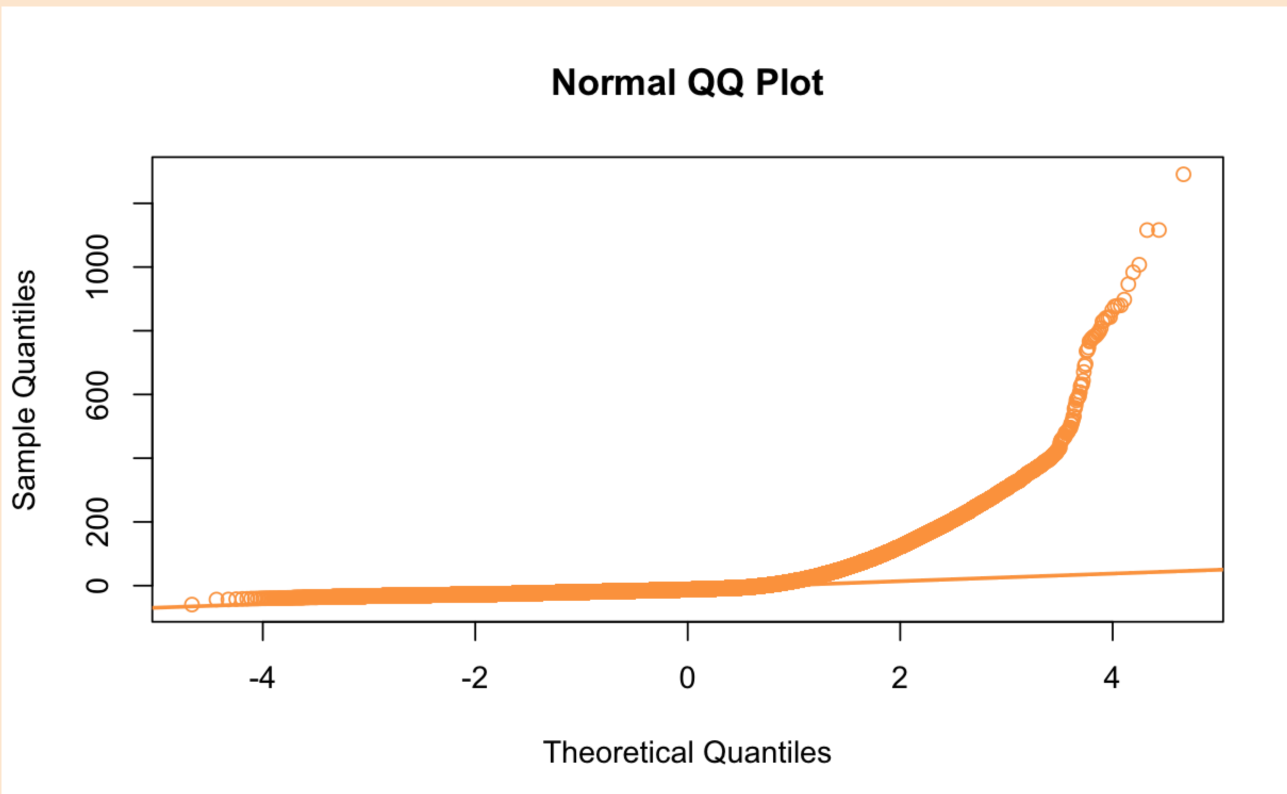
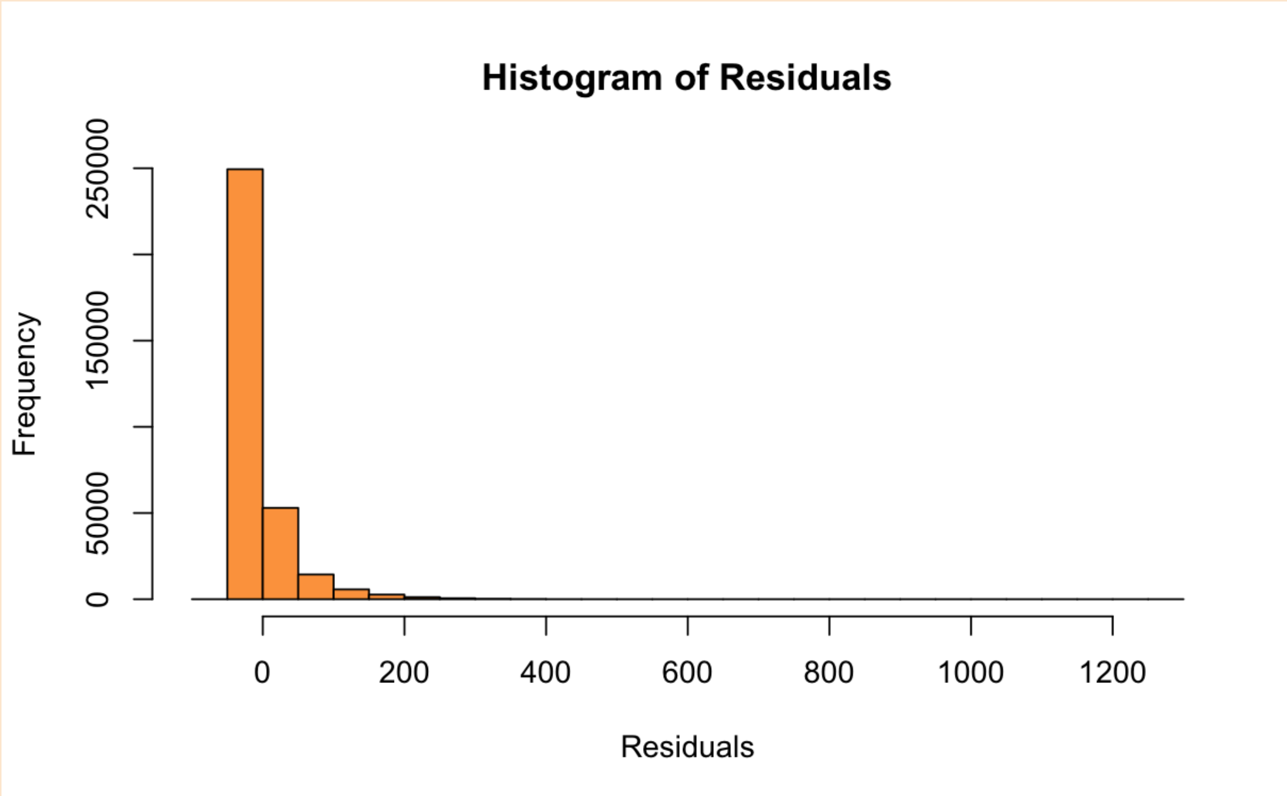
Monthly Flight Volume at NYC Airports
EWR
JFK
LGA
Total Flights
10000
7500
5000
2500
0
1 2 3 4 5 6 7 8 9 10 11 12
Month
Airport
EWR
JFK
LGA
```



In 2 out of the 3 examined airports, spring and summer tended to have the highest flight volumes, though this difference may not be significant enough to warrant any concrete conclusions as in LGA this trend does not reflect.

EWR had the highest average number of flights, while LGA had the lowest, though some of its season totals are comparable to that of JFK.

Question 2

Originally an ANOVA test to examine the relationship between the months and average flight delays was planned. However, after performing tests to determine whether or not the assumptions needed for ANOVA, it was found that the cleaned data violated the normality assumptions (as shown by below graphs), forcing a shift to using a non parametric test, Kruskal Wallis instead.

Assumptions:

```
```{r, jenny, echo=FALSE}

flights_clean <- flights %>%
```
```

```
  filter(origin %in% c("EWR", "JFK", "LGA")) %>%
  filter(!is.na(dep_delay), !is.na(arr_delay))

# assumption checks~~
model <- aov(dep_delay ~ factor(month), data = flights_clean)
residuals <- residuals(model)

leveneTest(dep_delay ~ factor(month), data = flights_clean)

# visual normality check
model <- aov(dep_delay ~ factor(month), data = flights_clean)
res <- residuals(model)

# normality check
hist(res, col = "tan1", main = "Histogram of Residuals", xlab = "Residuals")

qqnorm(res, col = "tan1", main = "Normal QQ Plot")
qqline(res, col = "tan1", lwd = 2)
```





Hypotheses:

- **Ho:** all months have the same average delay time
- **Ha:** at least one of the months differs from the rest of the group

```
```{r, jenny, echo=FALSE}

kruskal.test(dep_delay ~ factor(month), data = flights_clean)
```
```

```
        Kruskal-Wallis rank sum test

data:  dep_delay by factor(month)
Kruskal-Wallis chi-squared = 8668.9, df = 11, p-value < 2.2e-16
```

Results from Kruskal-Wallis test:

- **Chi-Squared:** 8668.9
- **df:** 11
- **p-value:** 2.2e^-16

Seeing that our p value is significant, we reject the null hypothesis. There is a difference between at least two of the months being compared.

A post-hoc test with Dunn's test can then be conducted to give us months that have the greatest average flight delay difference in comparison to other months after successfully rejecting the null hypothesis from Kruskal's.

```
```{r, jenny, echo=FALSE}

# dunn's
dunn_result <- dunnTest(dep_delay ~ factor(month), data = flights_clean, method = "bonferroni")

# result -> dataframe
dunn_df <- as.data.frame(dunn_result$res)

# get the months that were compared
dunn_df <- dunn_df %>%
  mutate(month1 = as.numeric(sub(" .*", "", Comparison)),
         month2 = as.numeric(sub(".*- ", "", Comparison)))

# get mean departure delay for each month
monthly_means <- flights_clean %>%
  mutate(month = as.numeric(as.character(month))) %>%
  group_by(month) %>%
  summarise(mean_delay = mean(dep_delay, na.rm = TRUE))

# delay info inputted into dunn's
dunn_df <- dunn_df %>%
  left_join(monthly_means, by = c("month1" = "month")) %>%
  rename(delay1 = mean_delay) %>%
  left_join(monthly_means, by = c("month2" = "month")) %>%
  rename(delay2 = mean_delay)

dunn_df <- dunn_df %>%
  mutate(higher_month = ifelse(delay1 > delay2, month1, month2),
         lower_month  = ifelse(delay1 > delay2, month2, month1),
         delay_diff   = abs(delay1 - delay2))

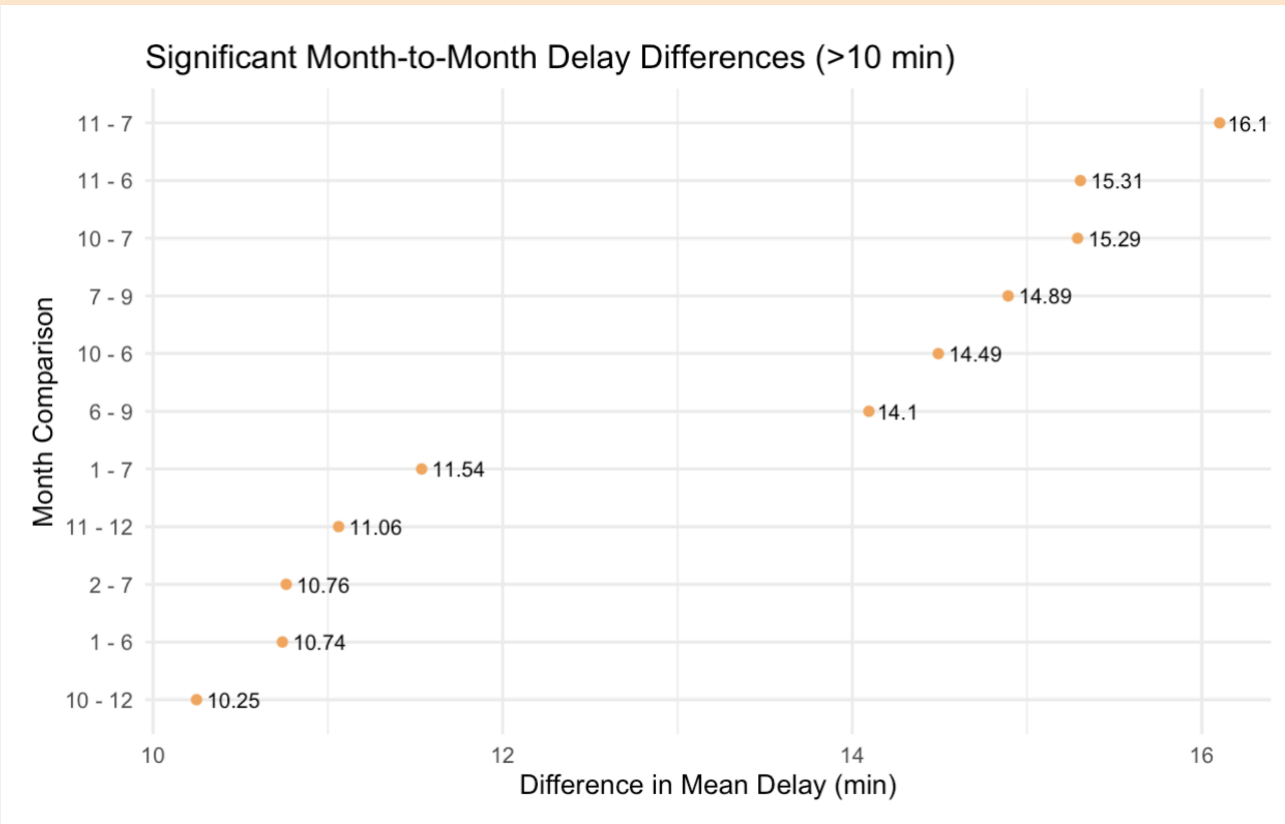
# sort by difference, filter for significance
significant_dunn <- dunn_df %>%
  filter(P.adj < 0.05) %>%
  arrange(desc(delay_diff))

significant_dunn %>%
  select(Comparison, Z, P.adj, higher_month, lower_month, delay1, delay2, delay_diff)

# filter for any interval greater than 10 minutes
```
```

```
significant_dunn_10plus <- significant_dunn %>%
  filter(delay_diff > 10)

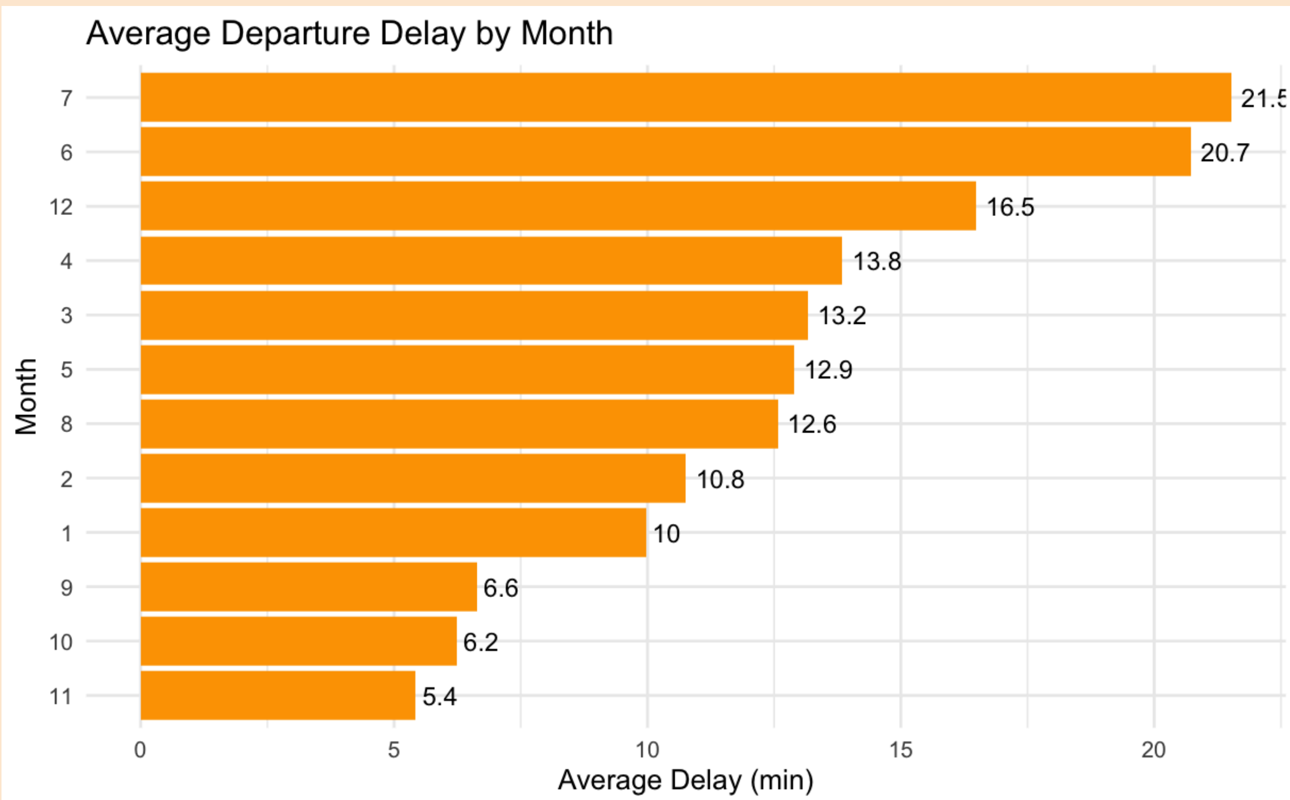
ggplot(significant_dunn_10plus, aes(x = reorder(Comparison, delay_diff), y = delay_diff)) +
  geom_point(color = "tan1") +
  geom_text(aes(label = round(delay_diff, 2)), hjust = -0.2, size = 3) +
  coord_flip() +
  labs(
    title = "Significant Month-to-Month Delay Differences (>10 min)",
    x = "Month Comparison", y = "Difference in Mean Delay (min)"
  ) +
  theme_minimal()
```



Having calculated the months that have the biggest delay differences, we can begin to rank the months by their average delay times based on their month to month delay differences and their monthly averages.

```{r, jenny, echo=FALSE}

avg_delay_by_month <- flights_clean %>%
  group_by(month) %>%
  summarise(avg_dep_delay = mean(dep_delay)) %>%
```

```
  arrange(month)

ggplot(avg_delay_by_month, aes(x = reorder(factor(month), avg_dep_delay), y = avg_dep_delay)) +
  geom_col(fill = "orange") +
  geom_text(aes(label = round(avg_dep_delay, 1)), hjust = -0.2, size = 3.5) +
  coord_flip() +
  labs(title = "Average Departure Delay by Month",
       x = "Month",
       y = "Average Delay (min)") +
  theme_minimal()
```



This confirms our Dunn's test results: that June and July have the highest average flight delays, while the fall months of September, October, and November have the lowest average delays of the year. Months with the biggest differences are these fall months and the two summer months have the greatest differences, reflected so in our graph of intervals.

Looking back on our data overview we can see that these findings are confirmed by the trends revealed in the heatmap of months.

Conclusion: There is a correlation between average delay times and the time of month.

Question 3

To formally assess whether flight delays are influenced by seasonal or weather conditions, Chi-Square Tests of Independence were conducted. For each analysis, the null hypothesis (Ho) assumed that delay occurrence or delay duration is independent of the tested factor, while the alternative hypothesis (Ha) asserted a statistically significant association.

Assumptions: The Chi-Square Test of Independence requires that data are categorical, observations are independent, and expected counts are sufficiently large in all cells; these were all satisfied.

````
```{r, zoe, echo=FALSE}
# Get contingency table
table_temp <- table(flights_temp$temp_cat, flights_temp$delay_status)
table_temp

# Run test
chisq.test(table_temp)
```
````

```
           Delayed On-Time
  Freezing    4991   19726
  Cold       29319  119361
  Mild       25388   90833
  Hot        10215   25891

        Pearson's Chi-squared test

data:  table_temp
X-squared = 1300.6, df = 3, p-value < 2.2e-16
```

````
```{r, zoe, echo=FALSE}
# Get contingency table
table_wind <- table(flights_wind_speed$wind_cat, flights_wind_speed$delay_status)
table_wind

# Run test
chisq.test(table_wind)
```
````

```
             Delayed On-Time
  Calm          7304   33248
  Breezy       21125   84127
  Windy        36177  123826
  Very Windy    5307   14610

        Pearson's Chi-squared test

data:  table_wind
X-squared = 849.87, df = 3, p-value < 2.2e-16
```

```{r, zoe, echo=FALSE}
# Get contingency table
table_visib <- table(flights_visib$visib_cat, flights_visib$delay_status)
table_visib

# Run test
chisq.test(table_visib)
```

```
       Delayed On-Time
  High   64172  244148
  Low     5741   11663

        Pearson's Chi-squared test with Yates' continuity correction

data:  table_visib
X-squared = 1447.5, df = 1, p-value < 2.2e-16
```

```{r, zoe, echo=FALSE}
# Get contingency table
table_precip <- table(flights_precip$precip_cat, flights_precip$delay_status)
table_precip

# Run test
chisq.test(table_precip)
```

```
                   Delayed On-Time
  No Precipitation   61742  243057
  Precipitation       8171   12754

        Pearson's Chi-squared test with Yates' continuity correction

data:  table_precip
X-squared = 4101.1, df = 1, p-value < 2.2e-16
```

Results:

- **Temperature:** χ^2 = 1300.6, df = 3, p < 2.2e-16

- **Wind Speed:** χ^2 = 849.9, df = 3, p < 2.2e-16

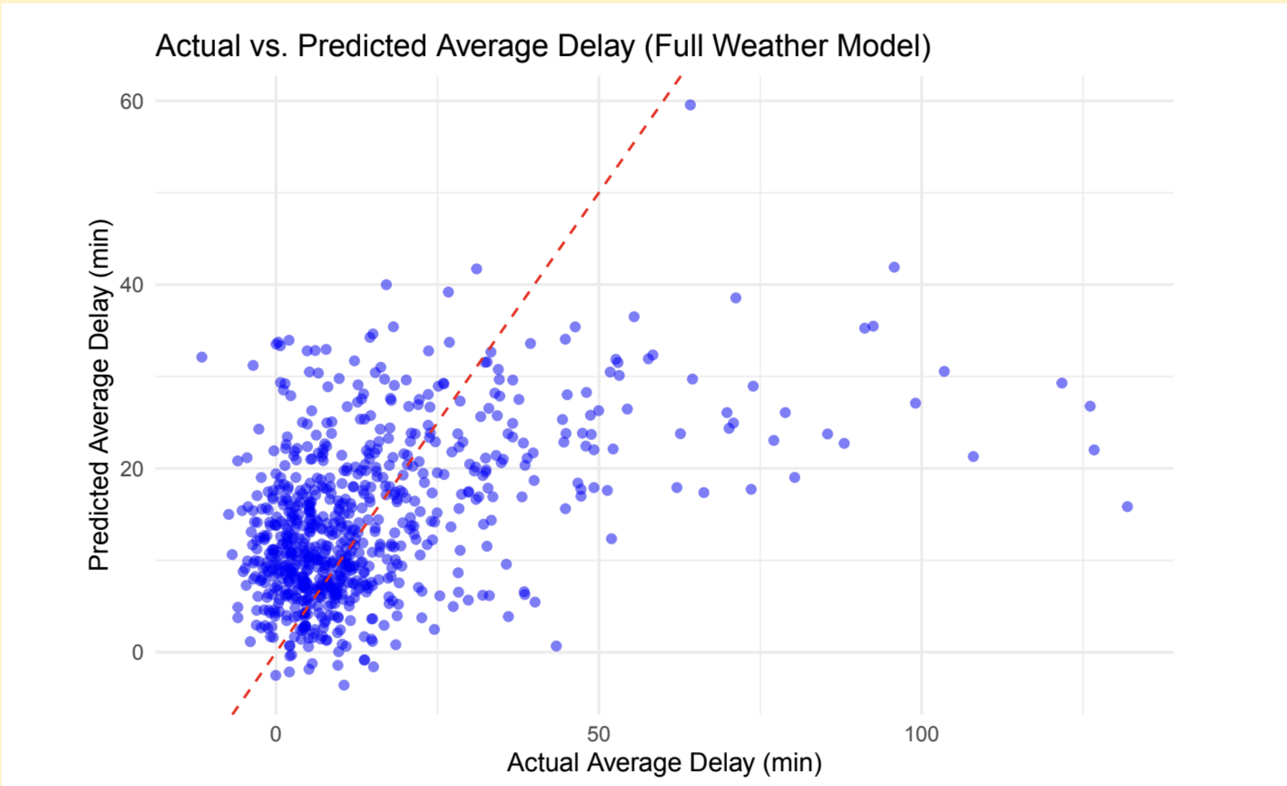
- **Visibility:** χ^2 =1447.5, df = 1, p < 2.2e-16

- **Precipitation:** χ^2 =4101.1, df = 1, p < 2.2e-16

Across all chosen variables – temperature, wind speed, visibility, and precipitation – the null hypothesis of independence was decisively rejected, with p-values < 2.2e-16, indicating strong associations between adverse weather and delay frequency.

In conclusion, there is robust statistical and visual evidence that both the frequency and severity of flight delays are significantly affected by seasonal and weather-related factors. Summer consistently yielded higher delays in both count and average duration. Poor weather conditions – most prominently high temperatures, very windy conditions, low visibility, and precipitation – are strongly linked to increased delays. These insights can guide proactive scheduling, staffing, and contingency planning, especially at New York's busiest airports.

Additionally, another model was made: a full weather model was made highlighting the relationship between predicted average delay and actual average delay.

````
```{r, zoe, echo=FALSE}
daily_weather_full <- flights_cleaned %>%
left_join(weather, by = c("origin", "year", "month", "day", "hour")) %>%
filter(!is.na(dep_delay), !is.na(temp), !is.na(dewp), !is.na(humid),
!is.na(wind_dir), !is.na(wind_speed), !is.na(wind_gust),
!is.na(precip), !is.na(pressure), !is.na(visib)) %>%
group_by(origin, year, month, day) %>%
summarise(
avg_delay = mean(dep_delay, na.rm = TRUE),
avg_temp = mean(temp),
avg_dewp = mean(dewp),
avg_humid = mean(humid),
avg_wind_dir = mean(wind_dir),
avg_wind_speed = mean(wind_speed),
avg_wind_gust = mean(wind_gust),
total_precip = sum(precip),
avg_pressure = mean(pressure),
````
````

```
avg_visib = mean(visib),
.groups = "drop"
)

# Fit full model with all weather predictors
full_model <- lm(avg_delay ~ avg_temp + avg_dewp + avg_humid + avg_wind_dir +

avg_wind_speed + avg_wind_gust + total_precip +
avg_pressure + avg_visib, data = daily_weather_full)

# View model summary
summary(full_model)
```

```
##
## Call:
## lm(formula = avg_delay ~ avg_temp + avg_dewp + avg_humid + avg_wind_dir +
##     avg_wind_speed + avg_wind_gust + total_precip + avg_pressure +
##     avg_visib, data = daily_weather_full)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.611  -9.627  -2.758   5.024 116.022
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    189.275381 106.651562   1.775  0.07635 .
## avg_temp         0.688799   0.424032   1.624  0.10471
## avg_dewp        -0.572664   0.462408  -1.238  0.21594
## avg_humid        0.707793   0.248482   2.848  0.00451 **
## avg_wind_dir     0.007057   0.008560   0.824  0.40996
## avg_wind_speed  -0.698931   0.373025  -1.874  0.06136 .
## avg_wind_gust    0.806466   0.316118   2.551  0.01093 *
## total_precip    -0.903550   0.923080  -0.979  0.32797
## avg_pressure    -0.231295   0.098191  -2.356  0.01875 *
## avg_visib       -0.361996   0.630260  -0.574  0.56590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.81 on 752 degrees of freedom
## Multiple R-squared:  0.2121, Adjusted R-squared:  0.2026
## F-statistic: 22.49 on 9 and 752 DF,  p-value: < 2.2e-16
```

```
# Add predicted values
daily_weather_full$predicted_delay <- predict(full_model)
# Plot
ggplot(daily_weather_full, aes(x = avg_delay, y = predicted_delay)) +
```

```
geom_point(alpha = 0.5, color = "blue") +
geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
labs(
title = "Actual vs. Predicted Average Delay (Full Weather Model)",
x = "Actual Average Delay (min)",
y = "Predicted Average Delay (min)"
) +
theme_minimal()
```



In order to truly understand if this model was appropriate, model diagnostics were run to prove the model was indeed valid. Additionally, the models were compared on a testing set where training occurred 80% of dates, and testing on the rest.

```
```{r, alexis and aparna, echo=FALSE}
set.seed(123)
dates <- unique(daily_weather_full$day)
train_dates <- sample(dates, 0.8 * length(dates)) #randomly selecting 80% of the dataset to be in the training model
#splitting the data into training and training sets
train_data <- semi_join(daily_weather_full, tibble(day = train_dates), by = "day")
test_data <- anti_join(daily_weather_full, tibble(day = train_dates), by = "day")
```
```

```
model_train <- lm(avg_delay ~ avg_temp + avg_dewp + avg_humid + avg_wind_dir +

avg_wind_speed + avg_wind_gust + total_precip +
avg_pressure + avg_visib, data = train_data)

#predicting on both training and test data
pred_train <- predict(model_train, newdata = train_data)
pred_test <- predict(model_train, newdata = test_data)
calculateMSE <- function(actual, predicted) {
mean((actual - predicted)^2)
}
train_mse <- calculateMSE(train_data$avg_delay, pred_train)
test_mse <- calculateMSE(test_data$avg_delay, pred_test)
cat("Training MSE:", train_mse, "\n")
```

```
## Training MSE: 266.9909
```

```
cat("Testing MSE:", test_mse, "\n")
```

```
## Testing MSE: 333.9324
```

The model performs well on the training data but shows decrease in performance on the test set, given the higher testing MSE. This suggests slight overfitting, however, the model is reliable given the overall scale of the model.

Question 4

Question 5

In order to analyze whether there was any relationship between high flight volume, weather conditions–specifically temperature and wind speed, and flight delays, we used the Kendall's Rank Correlation Coefficient test.

For this test, we needed to satisfy two assumptions:

1. Variables must be measured on an ordinal or continuous scale
2. (Non-strict) The relationship between the variables is monotonic.

As all the variables used in the three tests are numeric–with a test consisting of the dependent variable (departure delay) and each predictor variable (flight volume, average wind speed, average temperature), we pass the first assumption. We could not satisfy the second assumption that the relationship between the variables is monotonic, but we were still able to run the test as the assumption is not strict.

For each test:
Null Hypothesis (H0): There is no association between each respective predictor variable (flights per day, average wind speed, average temperature) and departure delay.

Alternative Hypothesis (H1): There is an association between each respective predictor variable (flights per day, average wind speed, average temperature) and departure delay.

Using a random sample of 2000 observations from the dataset, the following results were compiled:

````
```{r, alexis and aparna, echo = FALSE}
# Creates a sample of 2000 from the filtered flights combined dataset
sample_flights_c_f <- flights_combined_filtered |> slice_sample(n = 2000)

Flight_Volume_Dep_Delay <- cor.test(sample_flights_c_f$dep_delay,
sample_flights_c_f$flight_volume, method = "kendall")
Flight_Volume_Dep_Delay
```
````

````
```{r, alexis and aparna, echo = FALSE}
# Uses the same sample generated previously
Avg_Wind_Speed_Dep_Delay <- cor.test(sample_flights_c_f$dep_delay,
sample_flights_c_f$avg_wind_speed, method = "kendall")
Avg_Wind_Speed_Dep_Delay
```
````

````
```{r, alexis and aparna, echo = FALSE}
# Uses the same sample generated previously
Avg_Temp_Dep_Delay <- cor.test(sample_flights_c_f$dep_delay, sample_flights_c_f$avg_temp,
method = "kendall")
Avg_Temp_Dep_Delay
```
````

We then consolidated our results into a table, displaying the main values we tested for:

	z-score	**p-value**	**τ**
Flight Volume	-1.8691	0.0616	-0.0285
Average Wind Speed	2.3216	0.0203	0.0354
Average Temperature	2.8492	0.004	0.0434

As all 3 tau values are between 0 and 0.2, we can conclude that there is no correlation between each predictor variable and departure delay. However, we must also factor in each test's z-score and p-value.

- **Flight Volume**: As the p-value is 0.0616 > alpha = 0.05 and the z-score is -1.8691, which is less than 1.96 and greater than -1.96 (acceptable range for z-scores), we can conclude that flight volume shows no significant relationship with departure delay, meaning we fail to reject the null hypothesis for this test.
- **Average Wind Speed**: As the p-value is 0.0203 < alpha = 0.05 and the z-score is 2.3216, we can conclude that average flight volume does show a significant relationship with departure delay, but as the tau value is less than 0.2 (0.0354), the relationship is not impactful. Therefore we fail to reject the null hypothesis for this second test.
- **Average Temperature**: As the p-value is 0.004 < alpha = 0.05 and the z-score is 2.8492, we can conclude that average temperature does show a significant relationship with departure delay, but as the tau value is less than 0.2 (0.0434), the relationship is not impactful. Therefore we fail to reject the null hypothesis for this third test.

In conclusion, we fail to reject the null hypothesis for all three tests and can safely assume that there is no correlation relationship between flight volume, average wind speed, or average temperature to departure delay.

Question 6

````
```{r, gracelynne, echo = FALSE}
# Kruskal-Wallis Test
kruskal.test(arr_delay ~ origin, data = flights_weather_filtered)
kruskal.test(dep_delay ~ origin, data = flights_weather_filtered)

kruskal.test(arr_delay ~ bad_weather, data = flights_weather_filtered)
kruskal.test(dep_delay ~ bad_weather, data = flights_weather_filtered)

# Combine into group for pairwise comparison
flights_weather_filtered$group <- paste(flights_weather_filtered$origin,
flights_weather_filtered$bad_weather, sep = "_")

# Dunn Test
dunnTest(arr_delay ~ group, data = flights_weather_filtered, method = "bonferroni")
```
````

Due to assumptions for an ANOVA being violated, we chose to use the Kruskal-Wallis Test instead. The test revealed that there were significant differences in both arrival and departure delays between the three airports that we looked at, as all the p-values are less 2.2e-16. In the same way, when grouped by weather conditions, the p-values being less than 2.2e-16 indicates significant difference. These results show us that airport location and weather conditions have an impact on the durations of flight delays.

To specifically pinpoint which airport was affected the most, we conducted a post-hoc Dunn's Test to perform a pairwise comparison. Using this test, we found that JFK experiences a prominent increase in delays during bad weather, with arrival delays increasing by 13.48 minutes and departure delays by 9.5 minutes. Additionally, the data shows that EWR's arrival delays increase by 8.04 minutes when bad weather conditions are present. Therefore, we find that EWR has the highest overall delays, and JFK is the most affected by bad weather conditions.

Conclusions

Overall, our analysis shows the key patterns in flight delays throughout New York's airports. Firstly, it is clear that delays are more prevalent during the summer months of June and July. With this, the fall months of September, October, and November have the lowest amount of delays. Secondly, weather plays a major role in delays - high heat, strong winds, low visibility, and precipitation are strongly associated with increased delays across the airports.

Additionally, the locations of the airport and time of year of the flight do matter, suggesting the operational and environmental factors are important in delays. Lastly, while we expected flight volume, temperature, and wind speed to be strong predictors of delay, our tests showed no correlation.

Errors & Limitations

- **Question 3:** While the findings offer strong evidence that weather conditions are significantly associated with flight delay frequency and severity, several limitations and potential sources of error should be considered. One key limitation involves the grouping of weather variables into categories (e.g., "Windy", "Low Visibility") based on manually selected thresholds. Although this approach simplifies the analysis, it may introduce artificial boundaries and overlook more subtle variations within the data. Alternative grouping methods could potentially yield different outcomes.
- **Question 5:** Although the variables flight volume, average wind speed, and average temperature were confirmed to have no correlation with departure delay, there are key limitations that could have influenced our results. One such limitation and potentially even an error is that we needed to aggregate the wind speed and temperature variables, as taking the average of both for each day prevents our test from accurately assessing any potential impact of brief yet significant changes in wind speed or temperature. If our tests were to be run with non-aggregated values of these variables, there could potentially be a correlation between wind speed or temperature, with departure delay. Another limitation was that the distribution of data for departure delay was extremely right skewed, making it difficult to detect any correlation as a majority of the delays were either on time or slightly delayed. Narrowing down our departure delays to delays less

than 400 or 500 minutes or taking a smaller and carefully selected sample (equal distribution of different delay times) could potentially improve the accuracy of our results.

Author Contributions

- **Lydia Niu:** Question 4, Proposal alternative methods, Final report compilation
- **Alexis Castaneda:** Question 5, Distribution Plots, Correlation scatterplots, Model Diagnostics/Testing
- **Aparna Petluri:** Question 5, Distribution Plots, Correlation scatterplots, Model Diagnostics/Testing
- **Gracelynne Mohan:** Question 6, Bar Plots, Kruskal-Wallis Test & Post-hoc Dunn Test
- **Jenny Zhang:** Question 1-2,
- **Zoe Shum:** Question 3, Proposal methodology, Heatmap visualizations, Slideshow organization, Final report organization.