Group 14 Ice Cream - STAT 167 Final Project Report

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

2025-06-05

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.

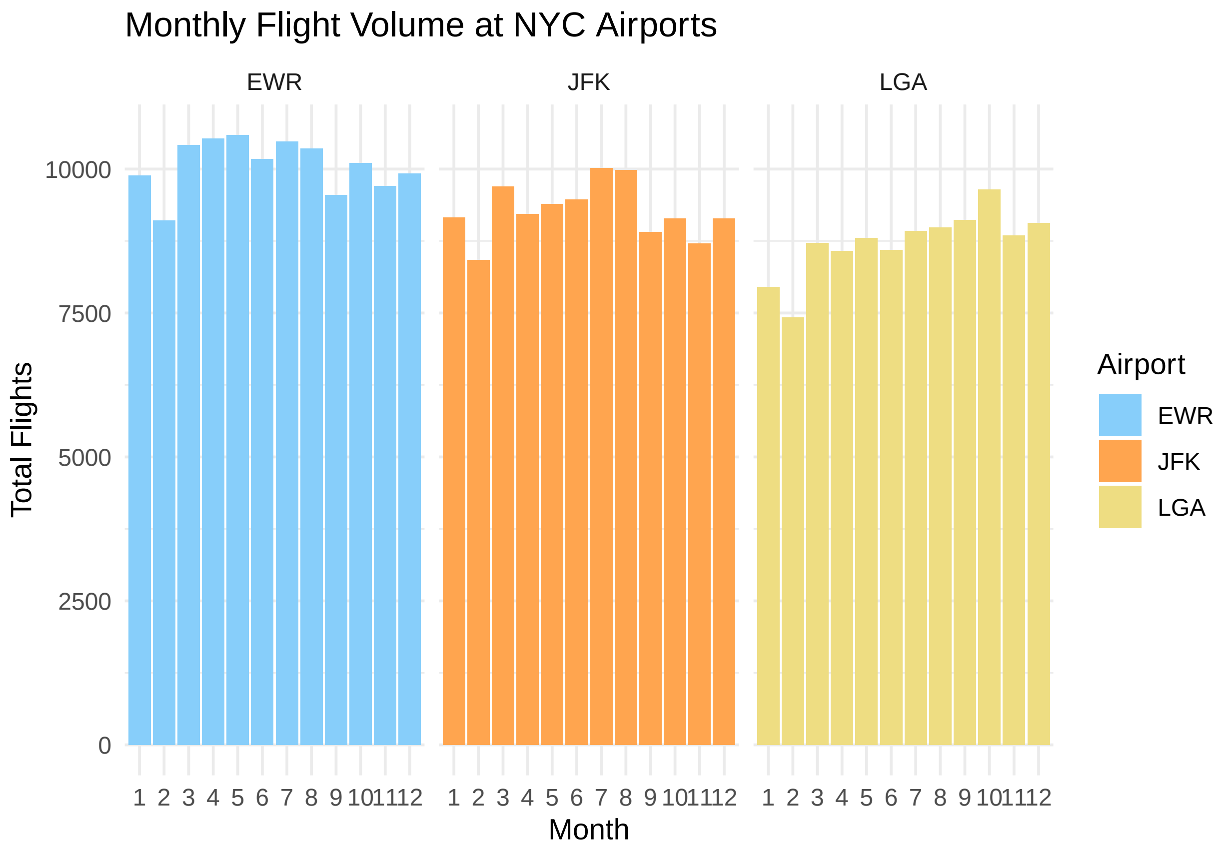
Flights: all flights that departed from NYC in 2013. Size: 336,776 x 19

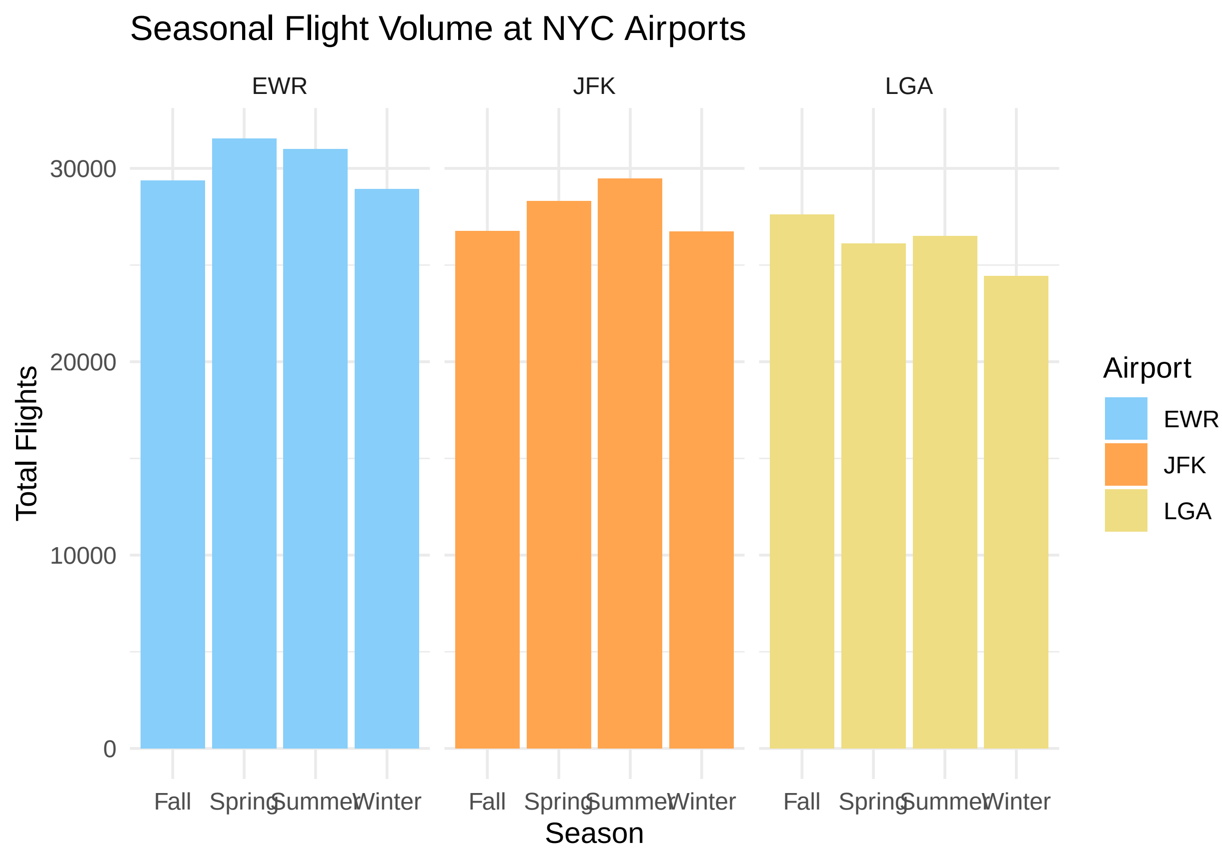
Weather: hourly meteorological data for each airport. Size: 26,115 x 15

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.

Data Exploration and Visualization

Question 1: Which months and seasons experience the highest flight volumes at each airport?
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.





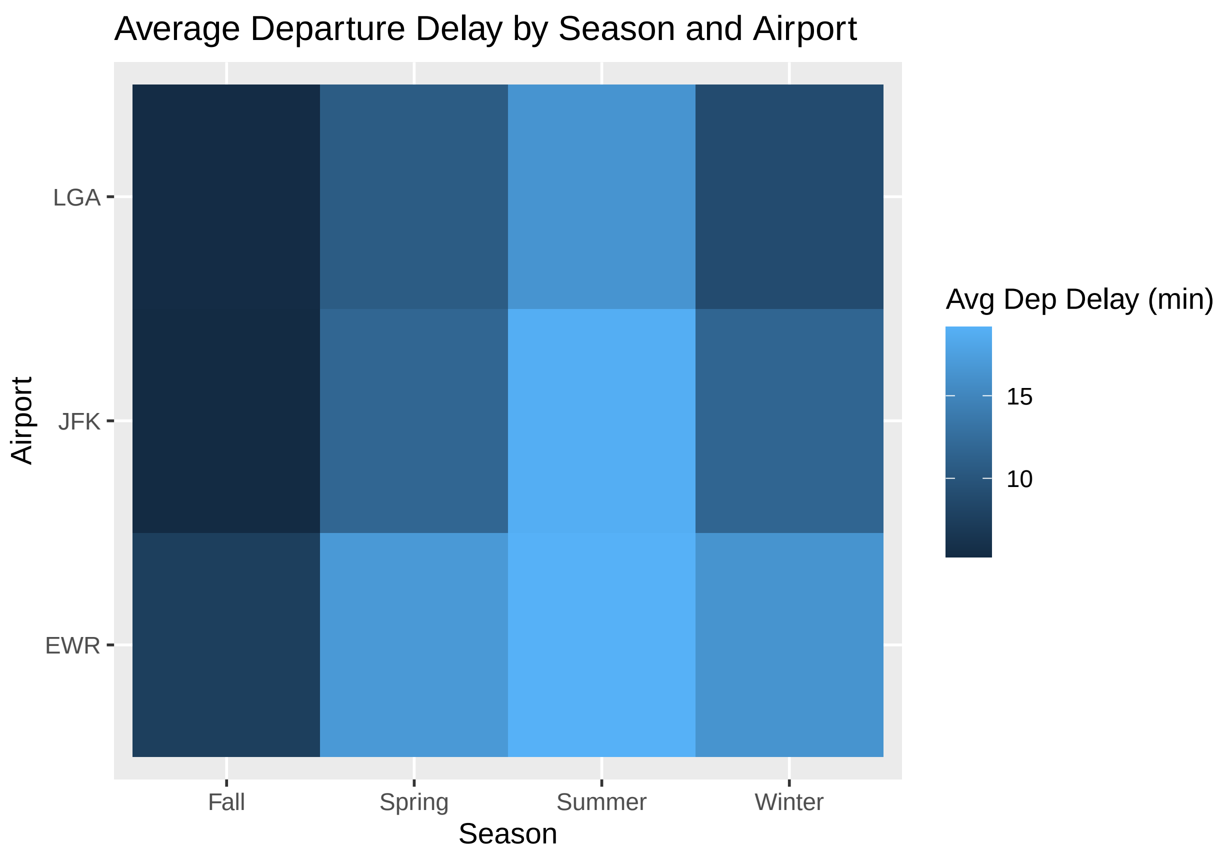
[needed] Analysis:

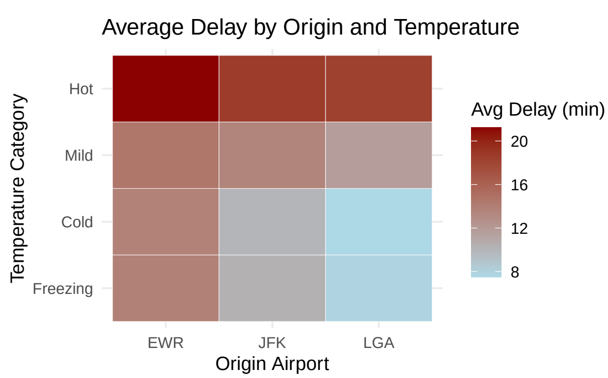
Question 2: How do average delays vary by month and season?

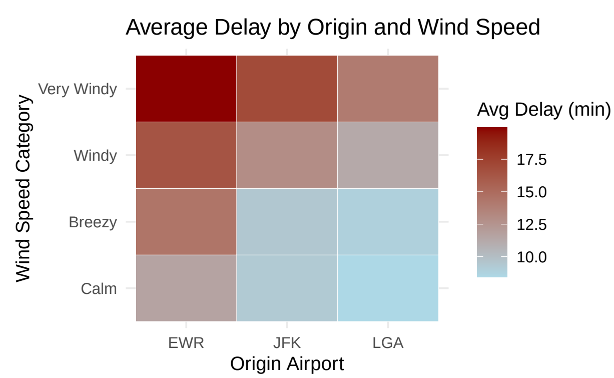
[needed] Analysis:

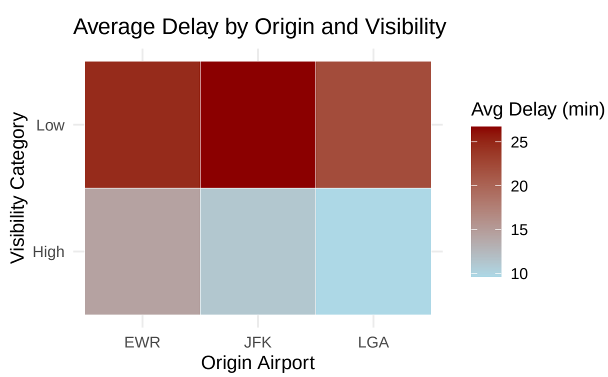
Question 3: Are delays more frequent/severe during specific weather conditions?

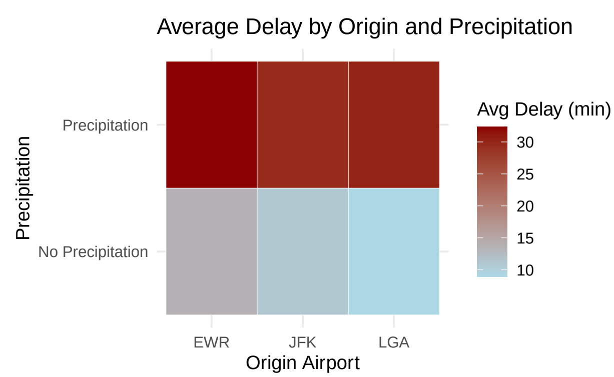
```
## # A tibble: 12 x 5
##    origin season flights_count delay_count avg_delay
##    <chr>  <chr>          <int>       <int>     <dbl>
##  1 EWR    Fall           28971        4739      7.52
##  2 EWR    Spring         30549        8477     16.8
##  3 EWR    Summer         30006        8662     19.2
##  4 EWR    Winter         27601        7669     16.3
##  5 JFK    Fall           26529        3589      5.25
##  6 JFK    Spring         27780        5890     11.8
##  7 JFK    Summer         28809        8153     18.8
##  8 JFK    Winter         25961        5577     11.6
##  9 LGA    Fall           27099        4002      5.40
## 10 LGA    Spring         25265        4972     10.7
## 11 LGA    Summer         25309        6231     16.3
## 12 LGA    Winter         23467        4459      8.95
```


Average Departure Delay by Season and Airport
Airport
LGA
JFK
EWR
Fall
Spring
Summer
Winter
Season
Avg Dep Delay (min)
15
10


Average Delay by Origin and Temperature
Temperature Category
Hot
Mild
Cold
Freezing
EWR
JFK
LGA
Origin Airport
Avg Delay (min)
20
16
12
8


Average Delay by Origin and Wind Speed
Wind Speed Category
Very Windy
Windy
Breezy
Calm
EWR
JFK
LGA
Origin Airport
Avg Delay (min)
17.5
15.0
12.5
10.0

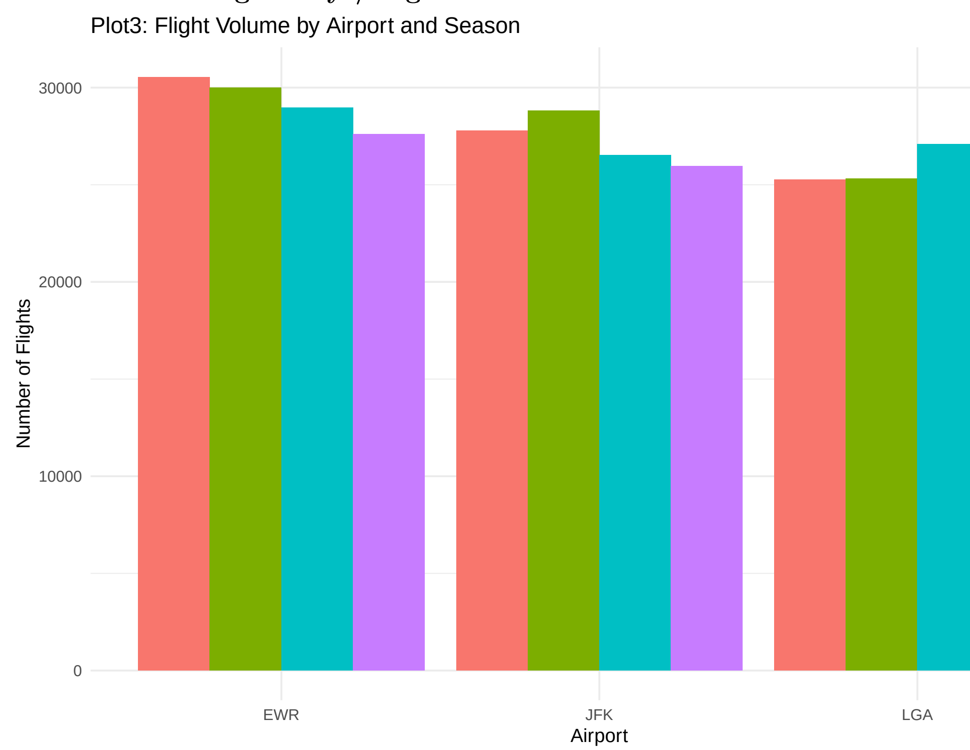

Average Delay by Origin and Visibility
Visibility Category
Low
High
EWR
JFK
LGA
Origin Airport
Avg Delay (min)
25
20
15
10

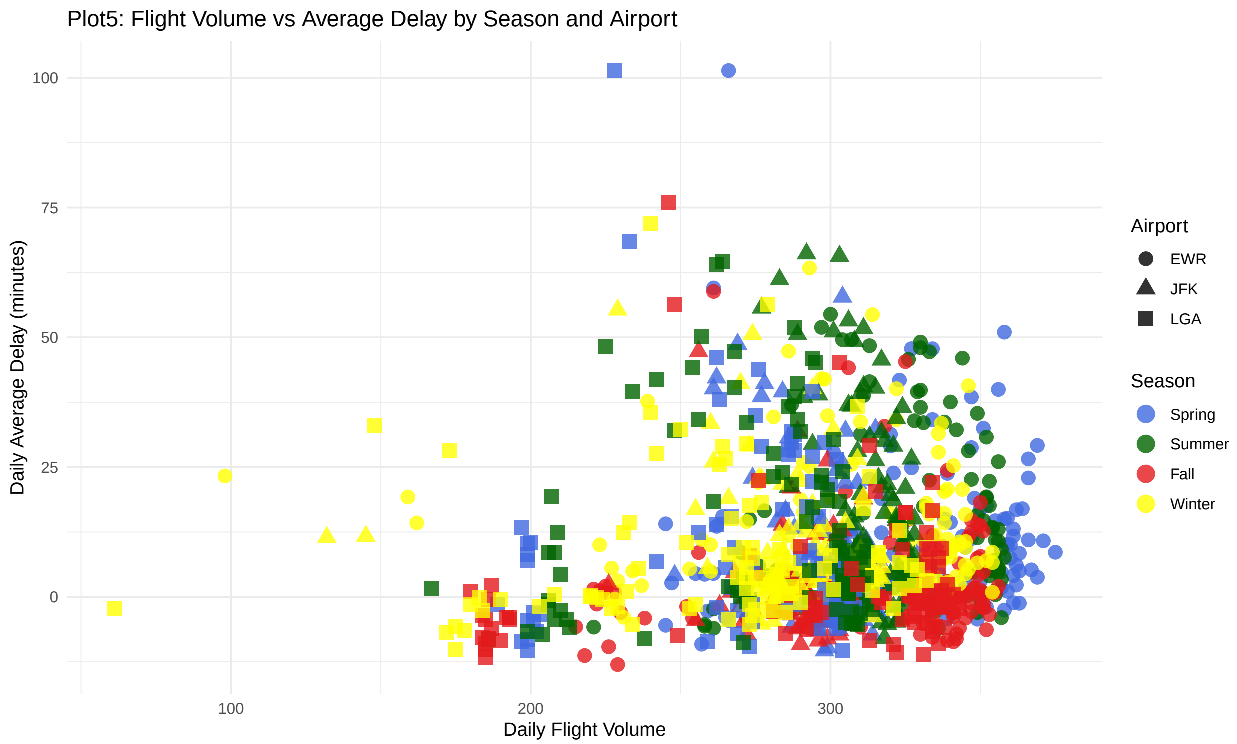

Average Delay by Origin and Precipitation
Precipitation
Precipitation
No Precipitation
EWR
JFK
LGA
Origin Airport
Avg Delay (min)
30
25
20
15
10

Analysis: 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.

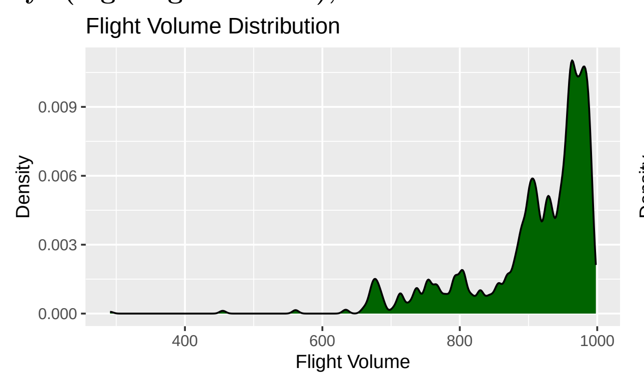
Question 4: Are there significant differences in average delays/ flight volume across the 3 airports and across different seasons?

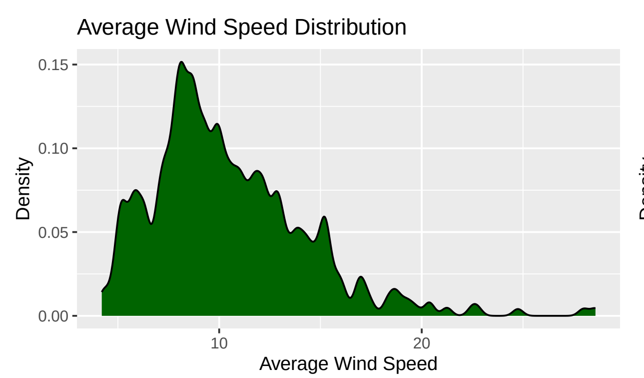




[needed] **Analysis:**

Question 5: What relationship, if any, exists between busy days (high flight volume), weather conditions (temperature and wind speed) and flight delays?





Analysis: 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.

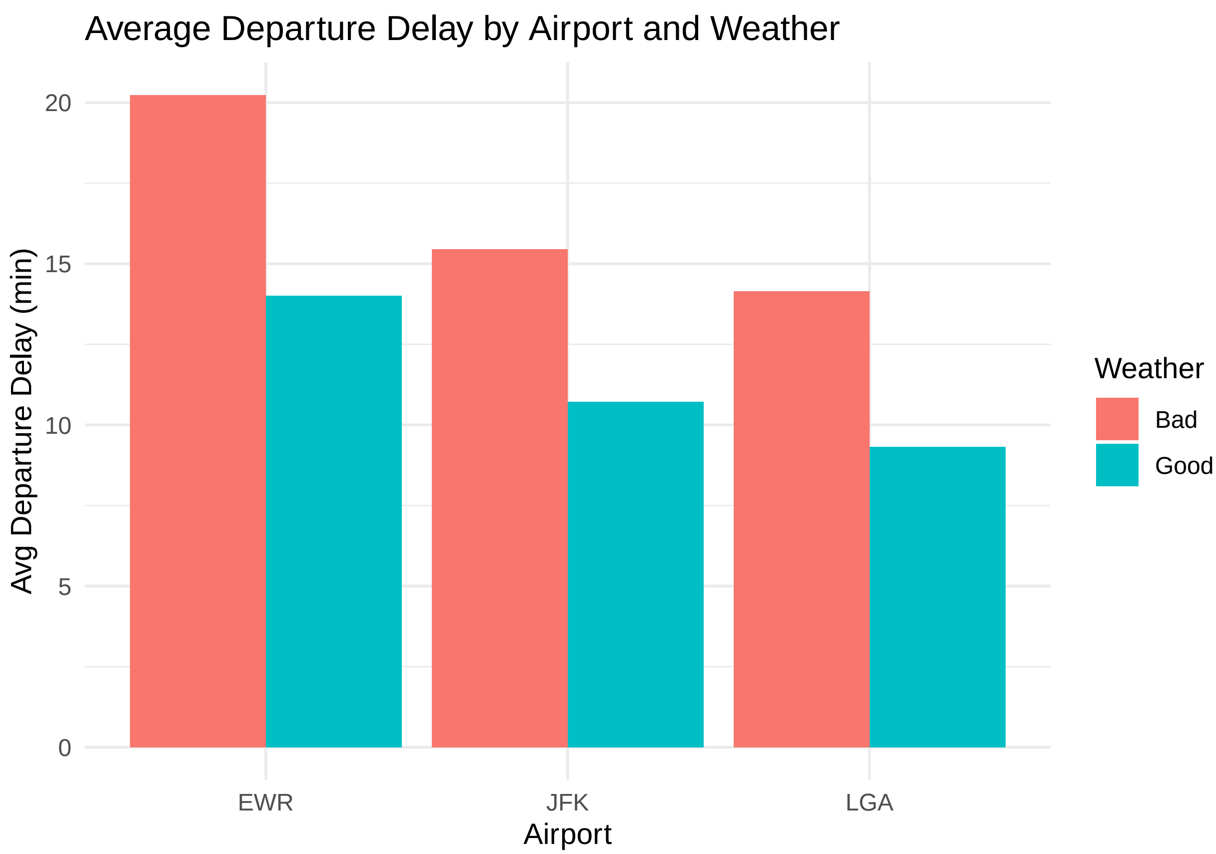
To examine departure delay's effects further, scatterplots were developed to plot each predictor variable (flight volume, average temperature, average wind speed) against departure delay. The scatterplots show that there is little to no correlation relationship in all 3 plots. However, we will determine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

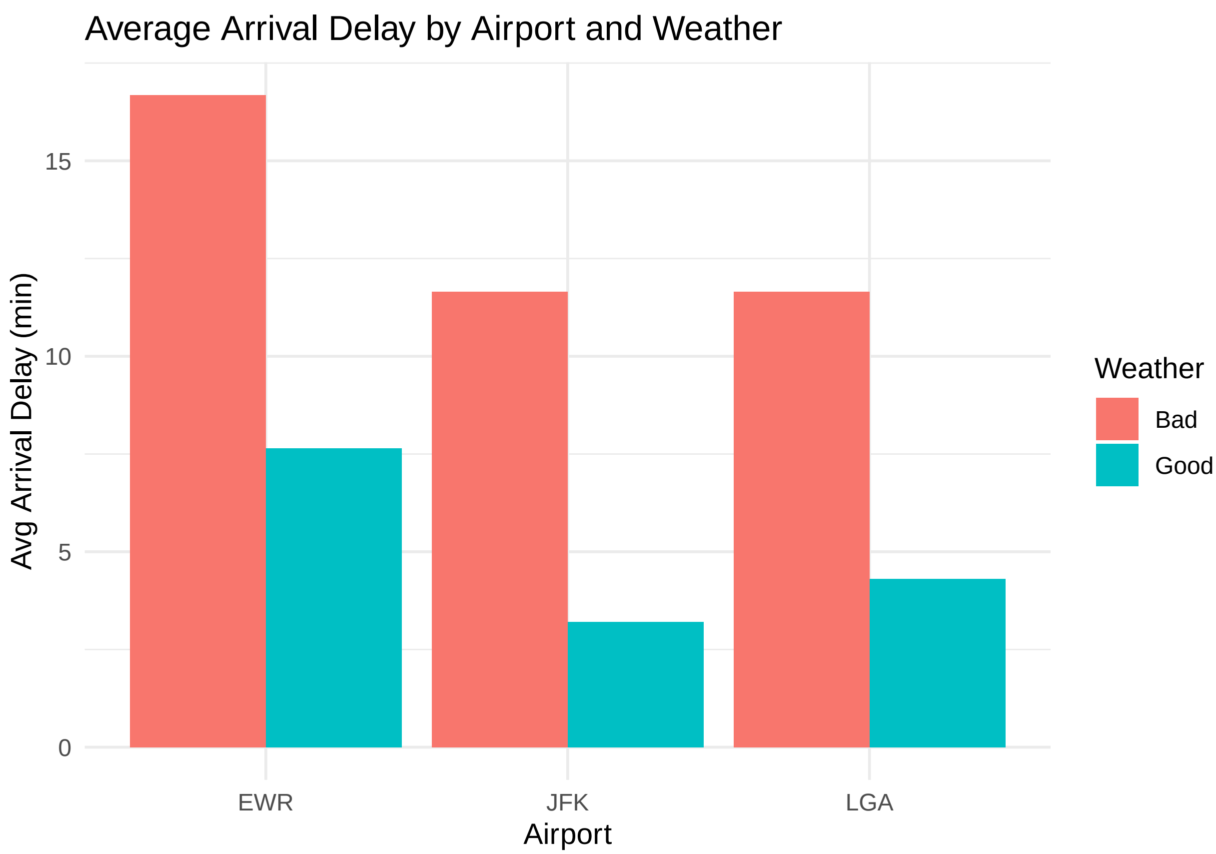
Question 6: Which airport is most affected by flight delays due to weather conditions among JFK, LGA, and EWR?

```
## Rows: 327,346
## Columns: 19
## $ year           <int> 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2~
## $ month          <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ day            <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ dep_time       <int> 517, 533, 542, 544, 554, 554, 555, 557, 557, 558, 558, ~
## $ sched_dep_time <int> 515, 529, 540, 545, 600, 558, 600, 600, 600, 600, 600, ~
## $ dep_delay      <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2, -1~
## $ arr_time       <int> 830, 850, 923, 1004, 812, 740, 913, 709, 838, 753, 849,~
## $ sched_arr_time <int> 819, 830, 850, 1022, 837, 728, 854, 723, 846, 745, 851,~
## $ arr_delay      <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7, -1~
## $ carrier        <chr> "UA", "UA", "AA", "B6", "DL", "UA", "B6", "EV", "B6", "~
## $ flight         <int> 1545, 1714, 1141, 725, 461, 1696, 507, 5708, 79, 301, 4~
## $ tailnum        <chr> "N14228", "N24211", "N619AA", "N804JB", "N668DN", "N394~
## $ origin         <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LGA",~
## $ dest           <chr> "IAH", "IAH", "MIA", "BQN", "ATL", "ORD", "FLL", "IAD",~
## $ air_time       <dbl> 227, 227, 160, 183, 116, 150, 158, 53, 140, 138, 149, 1~
## $ distance       <dbl> 1400, 1416, 1089, 1576, 762, 719, 1065, 229, 944, 733, ~
## $ hour           <dbl> 5, 5, 5, 5, 6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5, 6, 6, 6~
## $ minute         <dbl> 15, 29, 40, 45, 0, 58, 0, 0, 0, 0, 0, 0, 0, 0, 0, 59, 0~
## $ time_hour      <dttm> 2013-01-01 05:00:00, 2013-01-01 05:00:00, 2013-01-01 0~
```

```
## Rows: 327,346
## Columns: 6
## $ origin     <chr> "EWR", "LGA", "JFK", "JFK", "LGA", "EWR", "EWR", "LGA", "JF~
## $ arr_delay  <dbl> 11, 20, 33, -18, -25, 12, 19, -14, -8, 8, -2, -3, 7, -14, 3~
## $ dep_delay  <dbl> 2, 4, 2, -1, -6, -4, -5, -3, -3, -2, -2, -2, -2, -2, -1, 0,~
## $ precip     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ wind_speed <dbl> 12.65858, 14.96014, 14.96014, 14.96014, 16.11092, 12.65858,~
## $ visib      <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10,~
```

```
## # A tibble: 6 x 4
##   origin bad_weather avg_arr_delay avg_dep_delay
##   <chr>  <chr>               <dbl>         <dbl>
## 1 EWR    Bad                 16.7          20.2
## 2 EWR    Good                 7.64         14.0
## 3 JFK    Bad                 11.7          15.5
## 4 JFK    Good                 3.21         10.7
## 5 LGA    Bad                 11.7          14.1
## 6 LGA    Good                 4.31          9.32
```


Average Departure Delay by Airport and Weather
Avg Departure Delay (min)
20
15
10
5
0
EWR
JFK
LGA
Airport
Weather
Bad
Good



[needed]Analysis:

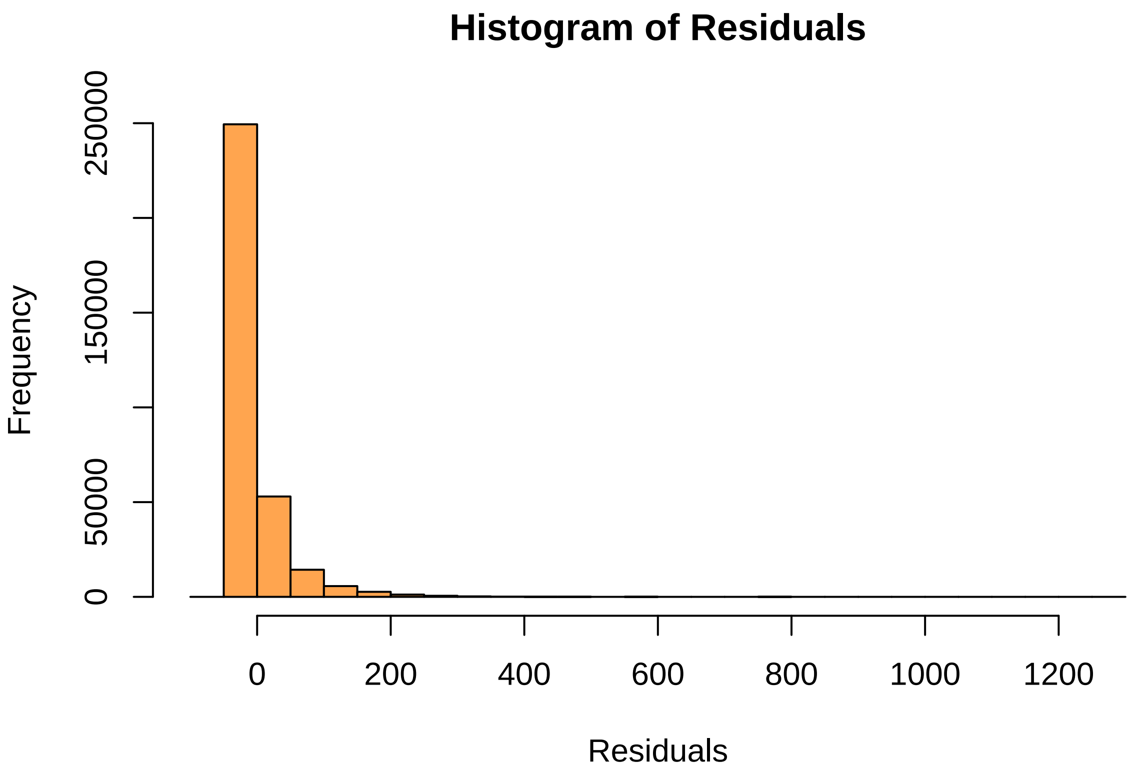
Data Analysis

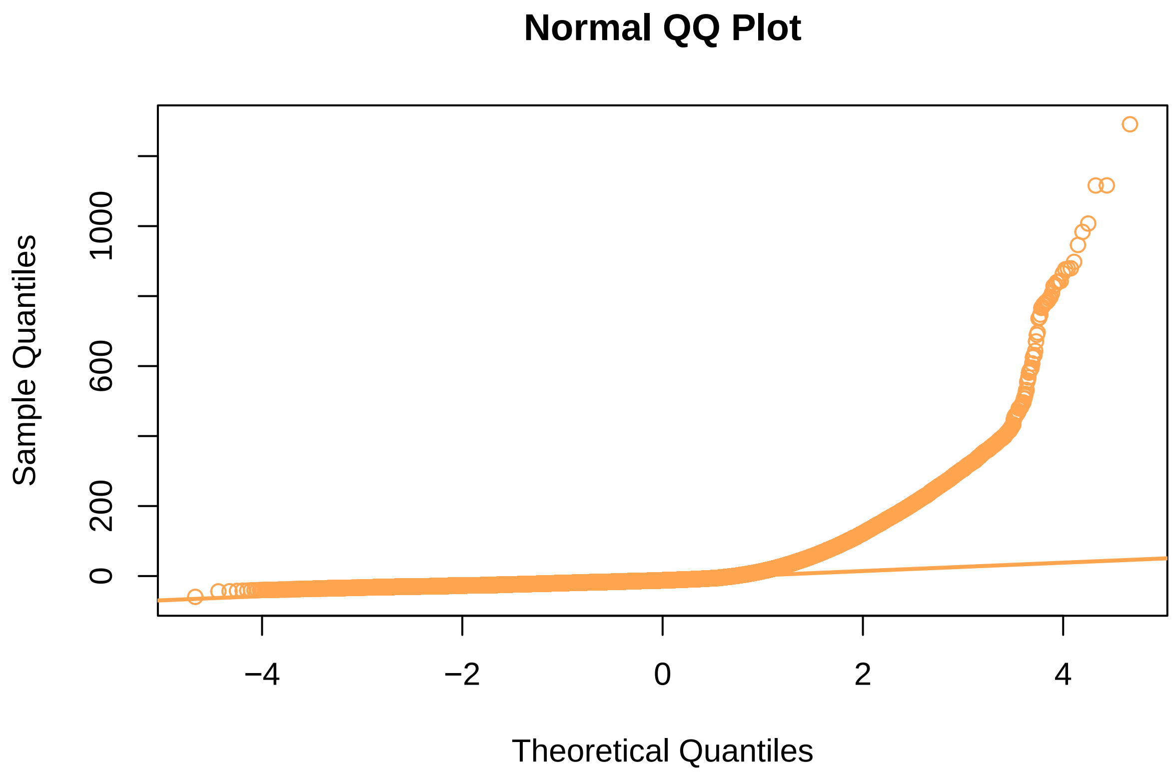
Question 1: Which months and seasons experience the highest flight volumes at each airport? This question is already being analyzed through data visualization.

Question 2: How do average delays vary by month and season? 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:

```
## Levene's Test for Homogeneity of Variance (center = median)
##           Df F value    Pr(>F)
## group     11  392.27 < 2.2e-16 ***
##       327334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Histogram of Residuals
Frequency
0
50000
150000
250000
0
200
400
600
800
1000
1200
Residuals



Analysis: Hypotheses:

Ho: all months have the same average delay time

Ha: at least one of the months differs from the rest of the group

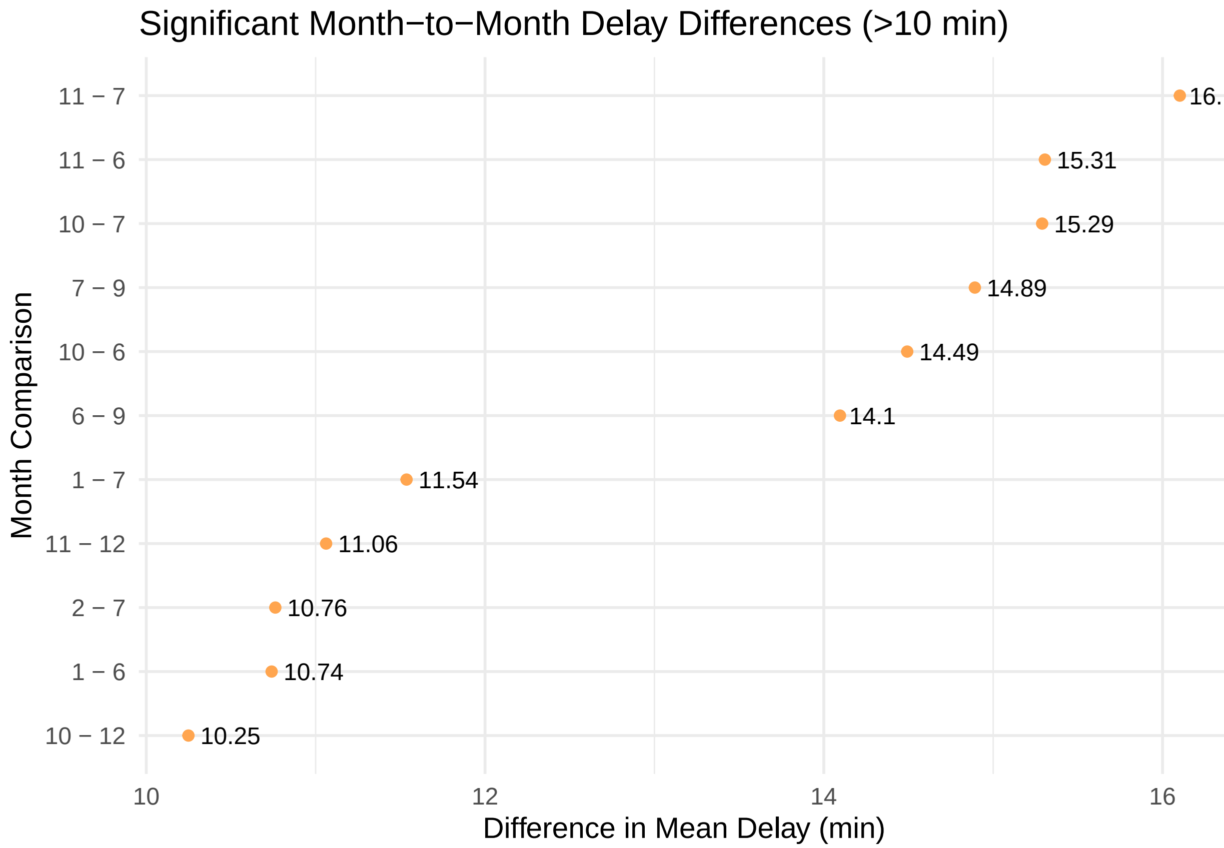
```
##
##  Kruskal-Wallis rank sum test
##
## data:  dep_delay by factor(month)
## Kruskal-Wallis chi-squared = 8603, df = 11, p-value < 2.2e-16
```

Results from Kruskal-Wallis test:

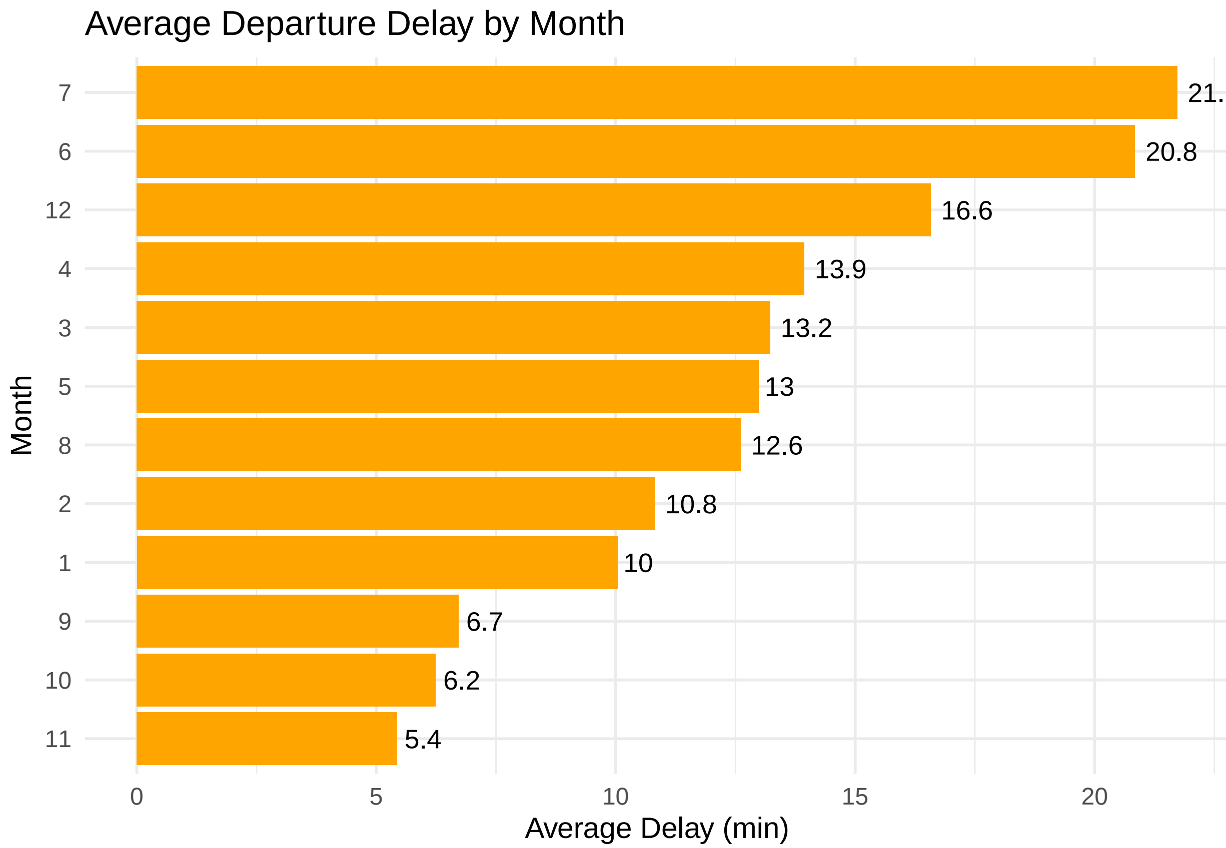
- **Chi-Squared: 8668.9**
- **df: 11**
- **p-value: 2.2e^-16**

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.



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.



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.

- July and November have the highest difference of 16.1
  - This makes sense; both months are either in the highest or lowest groups
- June and November also have one of the highest differences, and both are in, respectively, the highest and lowest groups

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

Question 3: Are delays more frequent/severe during different seasons or a specific weather? 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.

Analysis:

```
##
##             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
```

```
##
##               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
```

```
##
##         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
```

```
##
##                     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:** \$\chi^2\$ = 1300.6, df = 3, p < 2.2e-16
- **Wind Speed:** \$\chi^2\$ = 849.9, df = 3, p < 2.2e-16

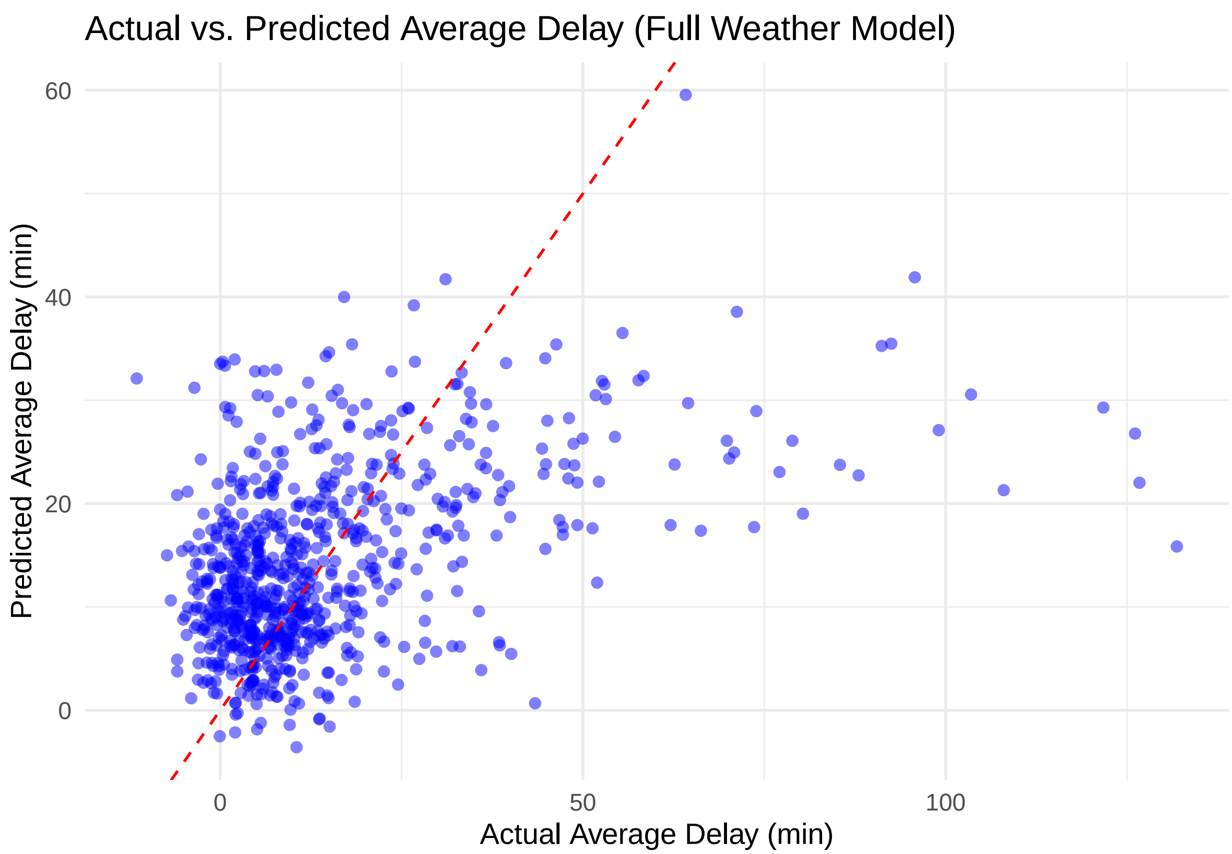
- **Visibility:** χ^2 =1447.5, df = 1, p < 2.2e-16
- **Precipitation:** χ^2 =4101.1, df = 1, p < 2.2e-16

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.

Conclusion: 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.

```
##
## 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
```



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.

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

```
## 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: Are there significant differences in average delays/volume across the 3 airports and across different seasons?

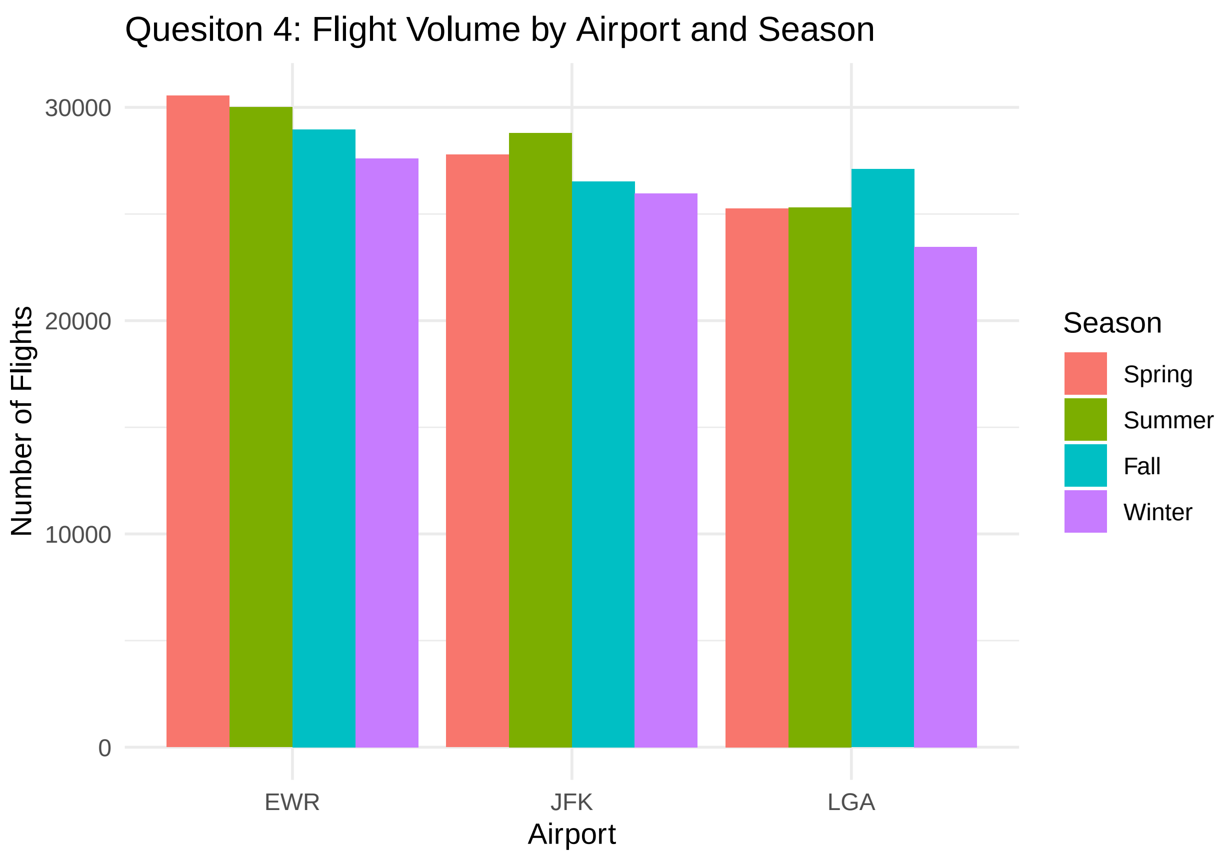
Assumptions: Initially, we considered using ANOVA, but ran into two issues: the raw dataset had over 5000 observations, making assumption testing challenging, and the summary table had only one observation per group, which is insufficient for ANOVA. It also failed to meet normality and constant variance assumptions, so a Kruskal-Wallis Test was used instead.

Kruskal-Wallis Test does not require normality and equal variances and requires two assumptions:

1. Continuous Dependent vVriables: average delay and flight volume
2. Categorical Factors: origin and season.

This test was appropriate for average delays, but could not be applied to flight volume since flight volume is an aggregated variable and Kruskal Wallis test required a sample size of 5 per group while summary table only has 1.

Analysis: For flight volume, the bar chart shows that EWR consistently has the highest flight volume across all seasons, with LGA having the lowest. Seasonally, Spring and Summer show higher flight volumes at JFK and EWR airports, while fall shows the highest flight volume in LGA.



For the average delay, a p-value of 2.2x10ˆ-16 from the Kruskal-Wallis test indicates that there is a significant difference in average delays between at least one group among season and origin. This result suggests that we need further questions to analyze which group is influencing the average delay.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  avg_delay by group
## Kruskal-Wallis chi-squared = 8144.9, df = 11, p-value < 2.2e-16
```

Conclusion: There are significant differences in both **flight volume** and **average delays** across the three NYC airports and across different **seasons**.

- **Flight volume** is highest at **EWR**, especially in **spring and summer**, while **LGA** consistently has the lowest volume except in **fall**.
- For **average delays**, the extremely small p-value (2.2e-16) from the **Kruskal-Wallis test** confirms that delays differ significantly across at least one airport or season.

Further analysis is needed to identify which specific groups (e.g. certain seasons or airports) contribute most to these differences.

Question 5: What relationship, if any, exists between busy days (high flight volume), weather conditions (temperature and wind speed) and flight delays? 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.

Assumptions: 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.

Analysis: 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:

```
##
##  Kendall's rank correlation tau
##
## data:  sample_flights_c_f$dep_delay and sample_flights_c_f$flight_volume
## z = -1.592, p-value = 0.1114
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##         tau
## -0.02431322
```

```
##
##  Kendall's rank correlation tau
##
## data:  sample_flights_c_f$dep_delay and sample_flights_c_f$avg_wind_speed
## z = 3.0334, p-value = 0.002418
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## 0.04621343
```

```
##
##  Kendall's rank correlation tau
##
## data:  sample_flights_c_f$dep_delay and sample_flights_c_f$avg_temp
```

```
## z = 1.1606, p-value = 0.2458
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## 0.01767159
```

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 τ 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 $0.0616 > \text{alpha} = 0.05$ 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 $0.0203 < \text{alpha} = 0.05$ 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 $0.004 < \text{alpha} = 0.05$ 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.

Conclusion: 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: Which airport is most affected by flight delays due to weather conditions among JFK, LGA, and EWR?

[needed] **Assumptions:**

Analysis:

```
##
##  Kruskal-Wallis rank sum test
##
## data:  arr_delay by origin
## Kruskal-Wallis chi-squared = 757.59, df = 2, p-value < 2.2e-16

##
##  Kruskal-Wallis rank sum test
##
## data:  dep_delay by origin
## Kruskal-Wallis chi-squared = 5308.2, df = 2, p-value < 2.2e-16
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  arr_delay by bad_weather
## Kruskal-Wallis chi-squared = 2258.8, df = 1, p-value < 2.2e-16
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  dep_delay by bad_weather
## Kruskal-Wallis chi-squared = 1027.6, df = 1, p-value < 2.2e-16
```

```
## Warning: group was coerced to a factor.
```

```
## Dunn (1964) Kruskal-Wallis multiple comparison
```

```
##   p-values adjusted with the Bonferroni method.
```

```
##             Comparison          Z       P.unadj         P.adj
## 1   EWR_Bad - EWR_Good  26.468805 2.216895e-154 3.325342e-153
## 2    EWR_Bad - JFK_Bad  12.730312  4.012335e-37  6.018502e-36
## 3   EWR_Good - JFK_Bad -13.990294  1.786707e-44  2.680061e-43
## 4   EWR_Bad - JFK_Good  42.647361  0.000000e+00  0.000000e+00
## 5  EWR_Good - JFK_Good  28.339379 1.131440e-176 1.697160e-175
## 6   JFK_Bad - JFK_Good  33.569533 4.671347e-247 7.007021e-246
## 7    EWR_Bad - LGA_Bad   9.772649  1.475429e-22  2.213143e-21
## 8   EWR_Good - LGA_Bad -14.560449  5.012760e-48  7.519140e-47
## 9    JFK_Bad - LGA_Bad  -2.167508  3.019616e-02  4.529424e-01
## 10  JFK_Good - LGA_Bad -31.513397 5.693103e-218 8.539655e-217
## 11  EWR_Bad - LGA_Good  38.531264  0.000000e+00  0.000000e+00
## 12 EWR_Good - LGA_Good  21.338197 5.019298e-101 7.528947e-100
## 13  JFK_Bad - LGA_Good  28.633065 2.605318e-180 3.907977e-179
## 14 JFK_Good - LGA_Good  -6.813855  9.501777e-12  1.425267e-10
## 15  LGA_Bad - LGA_Good  27.234156 2.560348e-163 3.840523e-162
```

[needed] Conclusion:

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 4:** Flight volume is an aggregated variable, meaning it is calculated by summarizing the raw dataset into grouped totals. In this case, the data is grouped by **3 airports** and **4 seasons**, resulting in only **12 observations**—one for each airport-season combination. This extreme reduction in data size significantly limits the statistical power of any analysis, as it no longer leverages the richness of the original raw dataset. Aggregation also reduces variability, making it difficult to detect meaningful differences or trends. Therefore, the flight volume was being analyzed with a bar chart without any supports from statistical test results.

- **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 (Captain): Question 4, Bar charts, Scatter plots, 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,**

- **Jenny Zhang: Question 1-2, Bar charts,**

- **Zoe Shum: Question 3, Proposal methodology, Heatmap visualizations, Slideshow organization, Final report organization.**