Stat 167: Project Proposal | G14 - ICE CREAM

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

2025-04-28

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 different **seasons or a specific weather**?
4. Are there significant differences in **average delays/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

Airports: airport names and locations. Size: 1,458 x 8

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

Some of the variables important to our research are:

- flights$month: A variable that provides the month of departure of each flight.
- weather$month: A variable that provides the time of recording (which month) of weather conditions.
- flights$dep_delay: A variable that provides the departure delay of each flight in minutes.
- flights$origin: A variable that provides the origin of each flight.
- airports$faa: A variable that provides each FAA airport code.
- weather$precip: A variable that provides precipitation recordings (in inches).

There are some missing values, however they fall under variables that we will not be looking into. A possible data limitation could be outliers in the dataset. In order to deal with this, we will choose variables that will help us answer our questions the most accurately and remove data that could give us a misrepresentation. Our plan for cleaning the data is just to condense the information down to the three airports we will be looking at. This dataset is sufficient for the objectives of our project because it has the right variables we need to look into to answer our questions.

Exploratory Data Analysis

Under this section, we will explore the data with some visualizations:

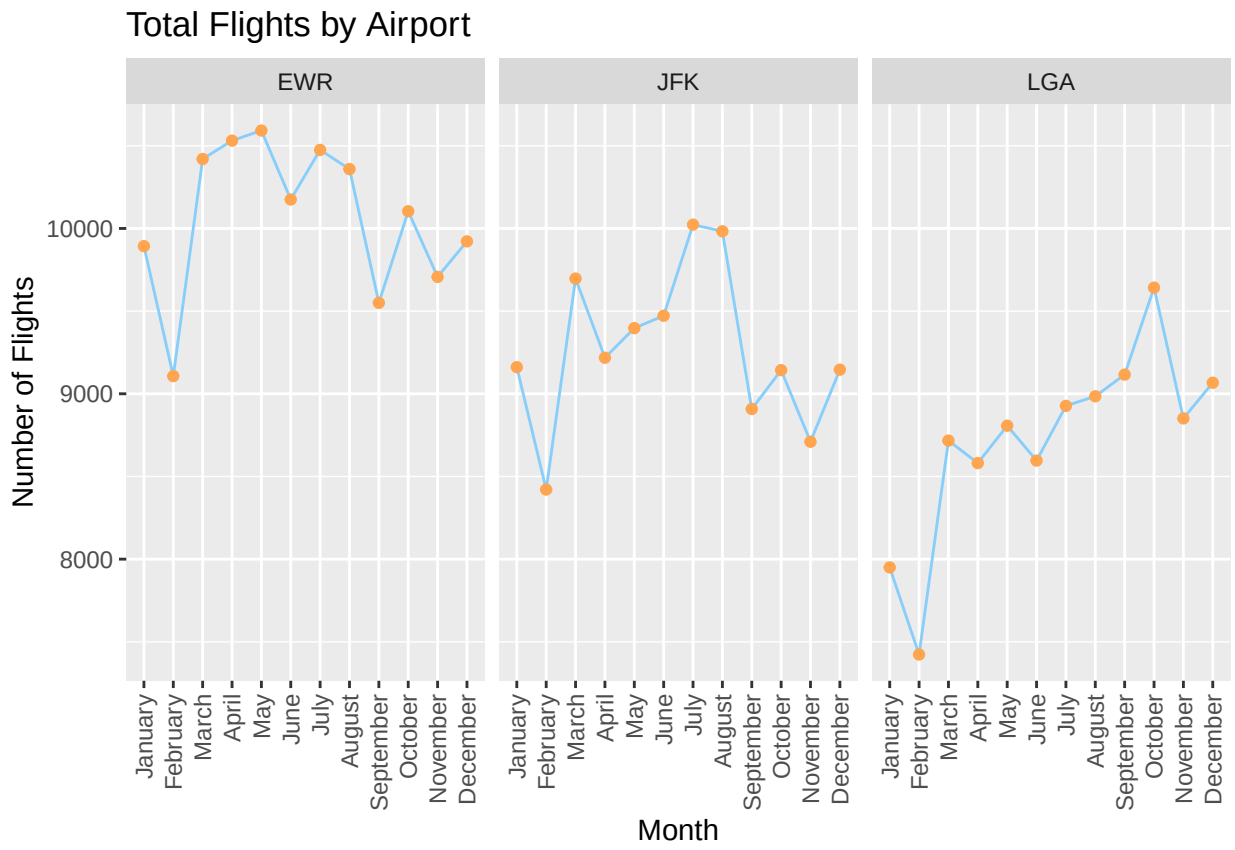
1. Trend Analysis with Time Series Visualization
    1. Visualize flight volume / delay changes month-by-month to detect trends (e.g., peaks in summer or holidays) for each airport
    2. To visually explore seasonal trends (x-axis = Month, y-axis = Flight count / delay count)
    3. Not a formal test, but it can help us gain some insightful insights
2. Use facet plots to compare monthly flight volumes and delays for 3 airports
3. Delay boxplots by airport (**dep_delay >= 15)
4. The first graph shows us the distribution of flights departing from each airport across the 12 months. We can see that EWR has the highest volume of departing flights.

```
three_airports <- flights %>%
  filter(origin %in% c("JFK", "LGA", "EWR"))

# summary
monthly_summary <- three_airports %>%
  group_by(origin, month) %>%
  summarise(
    flight_count = n(),
    delay_count = sum(dep_delay >= 15, na.rm = TRUE)
  )%>%
  mutate(month = factor(month, levels = 1:12, labels = month.name))
```

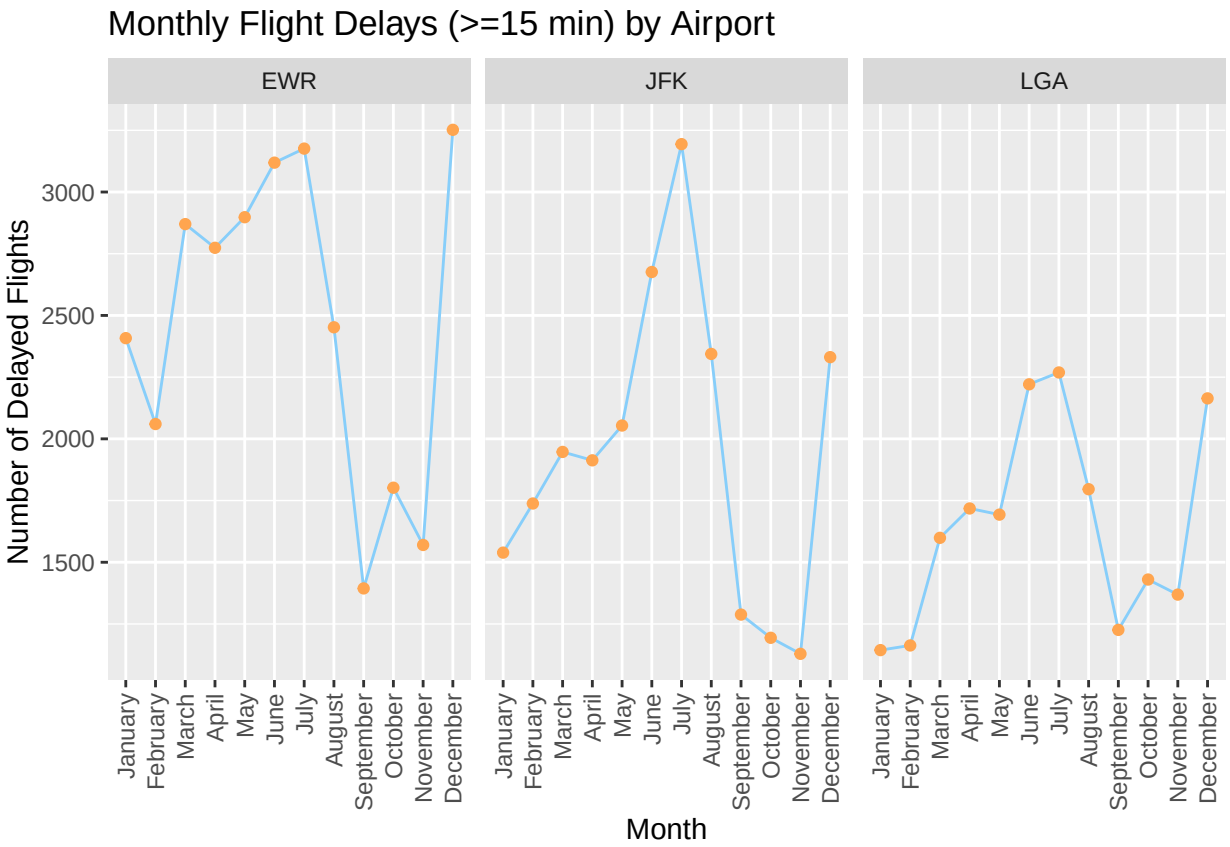
```
## `summarise()` has grouped output by 'origin'. You can override using the
## `.groups` argument.
```

```r
# flight counts
ggplot(monthly_summary, aes(x = month, y = flight_count, group = 1)) +
  geom_line(color = "lightskyblue") +
  geom_point(color = "tan1") +
  facet_wrap(~origin) +
  labs(title = "Total Flights by Airport",
       x = "Month", y = "Number of Flights") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



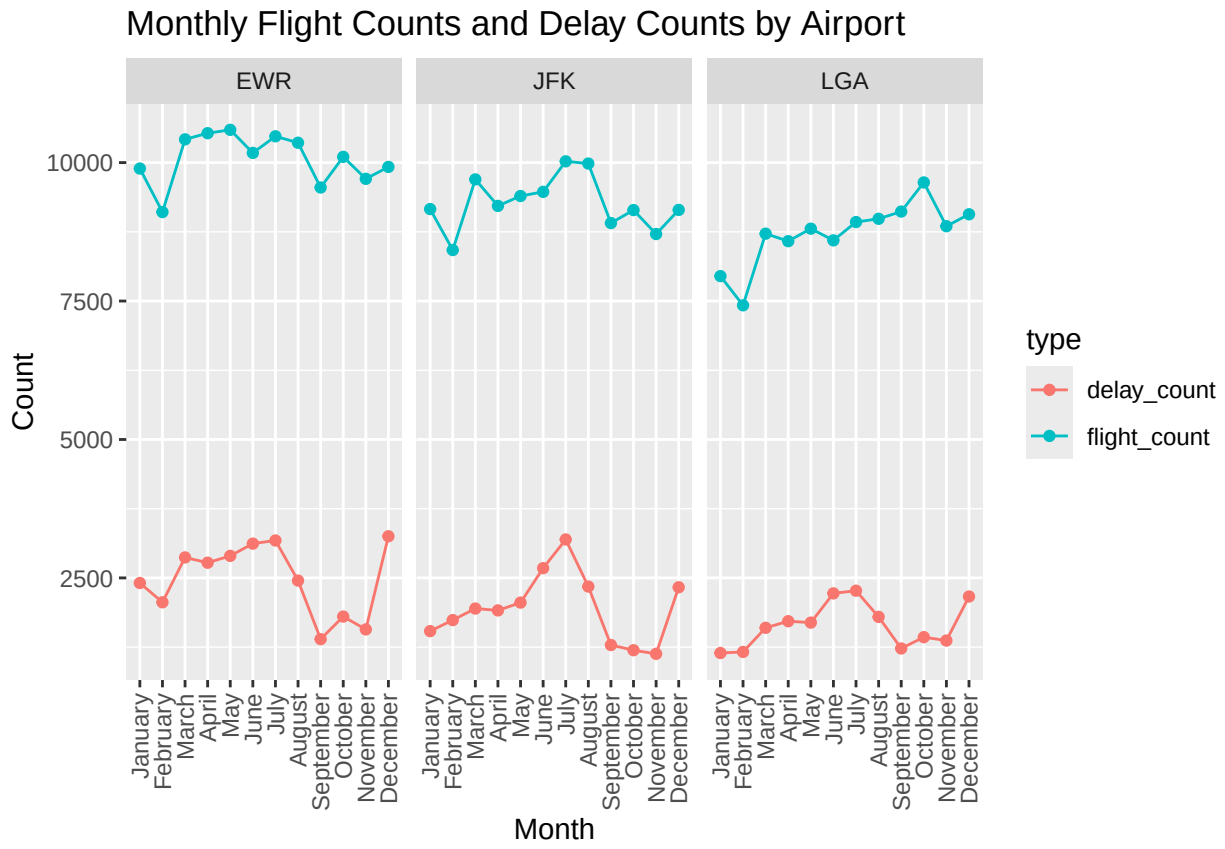
2. This graph gives us the distribution of the total amount of delayed flights for each airport. Airports with a smaller amount of departing flights will also have a proportionally lower amount of delayed flights.

```r
ggplot(monthly_summary, aes(x = month, y = delay_count, group = 1)) +
  geom_line(color = "lightskyblue") +
  geom_point(color = "tan1") +
  facet_wrap(~origin) +
  labs(title = "Monthly Flight Delays (>=15 min) by Airport",
       x = "Month",
       y = "Number of Delayed Flights") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



3. This graph compares the total number of flights departing from each airport and total number of delayed flights. Delayed numbers average out to roughly below one fourth of the total number of departing flights from each airport.

```
library(tidyr)
monthly_long <- monthly_summary %>%
  pivot_longer(
    cols = c(flight_count, delay_count),
    names_to = "type",
    values_to = "count"
  )
ggplot(monthly_long, aes(x = month, y = count, color = type, group = type)) +
  geom_line() +
  geom_point() +
  facet_wrap(~origin) +
  labs(title = "Monthly Flight Counts and Delay Counts by Airport",
       x = "Month", y = "Count") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```



Methodology

In this study, each flight will be treated as an independent observation, and delays are defined as departures delayed by 15 minutes or more, following FAA standards.

To address our main question – what factors influence flight volume and how this affects delay patterns across JFK, LGA, and EWR – we will use a combination of **visualizations, ANOVA tests, Chi-Square tests of independence, and linear regression modeling.**

First, to answer Question 1 (Which months and seasons experience the highest flight volumes?) we will create time-series plots of flight volume by month and by season for each airport. Month-level plots will provide detailed insights, while season-level plots will offer a broader overview of trends.

For the **ANOVA tests**, we will answer questions 2 and 4.

For Question 2 (How do average delays vary by month and season?), we will test:

- Null Hypothesis (H0): Mean departure delays are equal across all groups (months/seasons).
- Alternative Hypothesis (Ha): At least one group differs.

For Question 4 (Are there significant differences in average delays/volume across the three airports and seasons?), we will similarly test:

- Null Hypothesis (H0): Mean delays/volume are the same across airports or seasons.

- Alternative Hypothesis (Ha): At least one group differs.

If significant, we will follow up with Tukey's HSD post-hoc testing to identify which groups differ.

A **Chi-Square Test** of Independence will be used to answer Question 3 (Are delays more frequent during specific seasons or weather conditions?). Delay status will be categorized (dep_delay > 15), and contingency tables will be created to compare delay occurrence across seasons and weather conditions. Our hypotheses are:

- Null Hypothesis (H0): Delay occurrence is independent of season/weather.
- Alternative Hypothesis (Ha): Delay occurrence depends on season/weather.

Finally, for the **linear regression analyses**, we will address questions 5 and 6.

For Question 5 (What relationship exists between busy days, weather, and delays?), we will model departure delay as a function of daily flight volume, wind speed, and temperature.

- Null Hypothesis (H0): Flights per day, wind speed, and temperature are not significant predictors of departure delay.
- Alternative Hypothesis (Ha): They are significant predictors.

For Question 6 (Which airport is most affected by weather-related delays?), we will extend the model by including airport of origin as a predictor.

- Null Hypothesis (H0): Weather conditions and airport of origin are not significant predictors of delay.
- Alternative Hypothesis (Ha): They are significant predictors.

Work Plan

To ensure steady progress and deliver high-quality results, our project timeline is as follows:

- **Week 4:** Finalize and submit the formal written project proposal.
- **Week 5:** Clean and organize datasets to prepare for analysis.
- **Week 6:** Develop and refine exploratory visualizations to identify initial patterns.
- **Week 7:** Deliver an oral progress presentation summarizing early findings and next steps.
- **Week 8:** Conduct formal statistical analyses, including ANOVA, Chi-Square test of independence, and linear regression tests.
- **Week 9:** Begin compiling analysis results into a cohesive written report and draft final presentation materials.
- **Week 10:** Deliver the final oral presentation and submit the complete written report.

To better manage the project amongst 6 people, we plan on creating a GitHub repository for our R files. We will be dividing our tasks every week by creating and assigning issues on GitHub, working on individual branches, and pushing our changes to the main branch when done. We also will make sure to meet as a group at least once a week, to discuss our progress, any issues, and to hold each other accountable. If possible, we would like to look over finished branches at our weekly meetings to guarantee that the work is accurate and completed before adding to/changing our main branch.

Backup Plans & Alternative Strategies

Potential Pitfalls: Data Limitations and Variable Skewness

We found out that a couple of variables in the weather data table are completely skewed to either right or left, such as precip, wind_gust, wind_speed, and visib. This may cause the analysis to be misleading or having potential pitfalls.

Alternative idea

Main idea 1 - Comparing Flight Delays Over Time and Their Correlation with Weather Patterns

Introduction The nycflights13 dataset provides detailed flight information for all departures from NYC airports (LGA, JFK, EWR) in 2013, including weather, plane, airport, and airline data, while nycflights23 is a potential dataset assumed to cover similar data for 2023, allowing for a decade-long comparison. The main objective of this project is to compare flight delays over time and investigate their correlation with weather patterns, aiming to understand how weather impacts have evolved and whether airlines have adapted to these challenges. I want to analyze this because understanding delay trends and their weather-related causes can reveal operational improvements, highlight climate change effects on aviation, and provide actionable insights for airlines to mitigate delays, which is especially relevant given increasing weather variability.

Variables In the nycflights13 and nycflights23 packages, the datasets that we will use are weather, flight and airlines. Under the datasets, we list down the variables that will be used for analysis.

1. **Weather**: Hourly meteorological data for each airport, providing context for delay causes.
    1. **visib**: Visibility in miles, crucial for assessing its impact on arrival delays.
    2. **wind_speed**: Wind speed in knots, a key factor in flight disruptions.
    3. **precip**: Rainfall amount in inches/hour, indicating potential weather-related delays.
    4. **pressure**: Atmospheric pressure in hPa, which may influence flight operations.
2. **Flight**: Core dataset of flight records, central to delay analysis.
    1. **Dep_delay**: Departure delay in minutes, measuring how late flights left.
    2. **Arr_delay**: Arrival delay in minutes, capturing how late flights arrived.
    3. **Sched_dep_time**: Scheduled departure time, used to track delay patterns across the day.
    4. **Sched_arr_time**: Scheduled arrival time, aiding in analyzing arrival delay trends.
    5. **Time_hour**: Rounded timestamp, essential for matching flights with weather data.
3. **Airlines**: Translation of carrier codes to names, enabling airline-specific analysis.

Coherent Questions/Possible Direction This idea will explore several directions to deepen the understanding of flight delays and weather impacts, such as assessing the role of airlines in delay times, examining visibility's effect on arrivals, and identifying key weather variables driving delays, etc. By analyzing these aspects, we can uncover whether specific weather conditions disproportionately affect certain airlines, how operational strategies have evolved, and if airports play a significant role in delay variations. The following questions will guide our analysis:

1. How relevant are the departure and arrival airports to the overall flight delay time?

2. How does the correlation between visibility and arrival delays differ between 2013 and 2023?
3. Which weather variables show the strongest correlation with departure delays in 2013 versus 2023?
4. Do certain airlines show more resilience to weather-related delays in 2023 compared to 2013, and what might explain this (e.g., operational improvements)?
5. Which airlines are most affected by flight delays due to adverse weather conditions?
6. Have airlines improved in dealing with weather-related delays over the past decade, especially through 2023?

Main idea 2 - Seasonal Trends in Aircraft Usage Across New York Airports

We are investigating the factors that influence the choice of plane models based on seasonal usage patterns. Our study will explore which plane types are most common in each season and compare these patterns across New York's three major airports: JFK, EWR, and LGA. We are unsure whether the choices for plane models are driven by weather conditions, differences in travel demand across seasons, or other operational considerations. To answer these questions, we will analyze the data table, weather, airplanes, and flights to uncover potential relationships between aircraft selection and seasonal conditions.

Main idea 3 - The most common flight time amongst JFK, EWR, and LGA airports

We are studying the most common flight departure times across New York's three major airports: JFK, EWR, and LGA. Our goal is to identify whether certain times of day—such as early morning, midday, or evening—are preferred for departures at each airport. By analyzing scheduled departure times, we aim to uncover patterns in airport operations, passenger behavior, or airline scheduling strategies. This analysis will help us better understand how time-of-day preferences vary across airports and may reflect broader trends in air travel demand or airport management.