Stat 167: Project Proposal | G14 - ICE CREAM

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)

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

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