Visualisation Project Component 1

Alok Dhar Dubey

2023-09-29

Introduction

In the past few decades, we had major technological advancements which changed our lifestyle, our ways of thinking and even our future goals. Among all these, the aviation industry is one such sector. It has become an important part of today's world due to its numerous uses, some of which has now become necessities. It is a boon to our civilization for enabling efficient transportation of goods and people. Apart from that, it fosters international trade, tourism and cultural exchange. It has now become a priority in cases of emergencies by providing humanitarian reliefs efficiently and much faster than ever. Due to these reasons, it is important to devise necessary strategies to keep the aviation industry running efficiently and without much hurdles. This project is an attempt in that direction.

Data Description

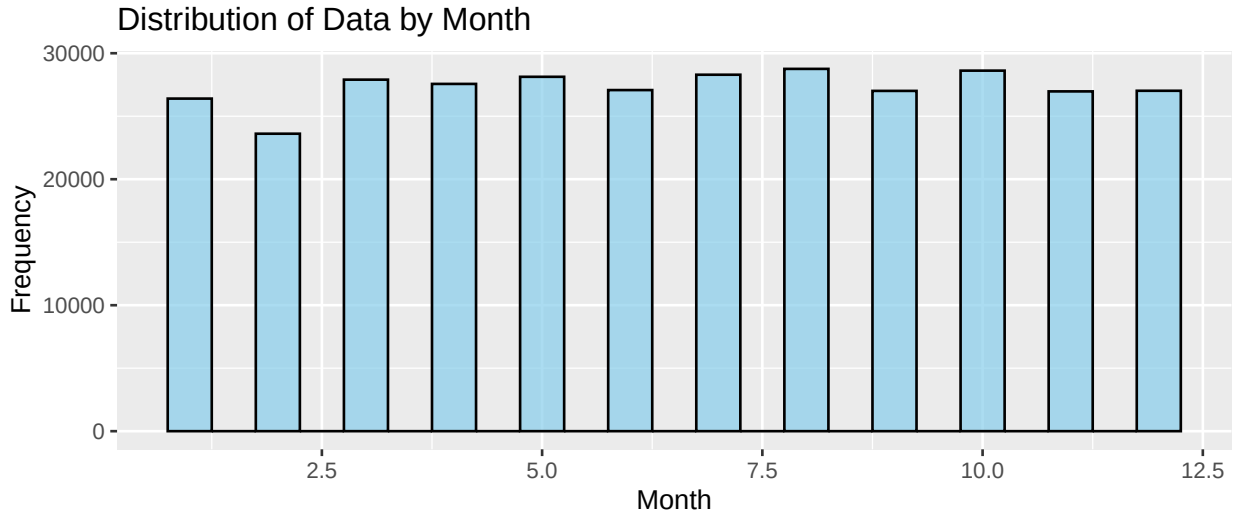
Table 1: The Flight Dataset

month	day	dep_time	sched_dep_time	dep_delay	name	origin	dest
1	1	517	515	2	United Air Lines Inc.	EWR	IAH
1	1	533	529	4	United Air Lines Inc.	LGA	IAH
1	1	542	540	2	American Airlines Inc.	JFK	MIA
1	1	544	545	-1	JetBlue Airways	JFK	BQN
1	1	554	600	-6	Delta Air Lines Inc.	LGA	ATL
1	1	554	558	-4	United Air Lines Inc.	EWR	ORD

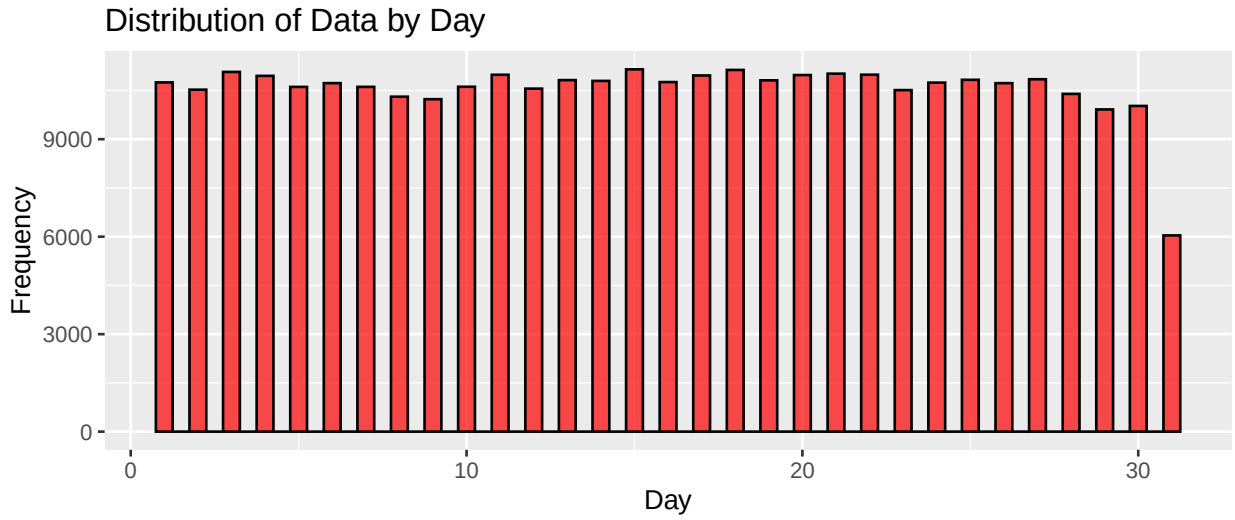
We will be working with the flight dataset available on Kaggle. The link to the Kaggle page containing the dataset can found here. The dataset contains information about the flights of an airport for the year 2013. It contains 3,36,776 rows and 21 columns. It includes information such as departure and arrival time, delays, flight company, flight number, flight origin and destination, flight duration, distance, hour and minute of flight, and exact date and time of flight. Due to the large size of dataset and restrictions on the size of this project, we will be using only some of its columns for our analysis.

Exploratory Data Analysis

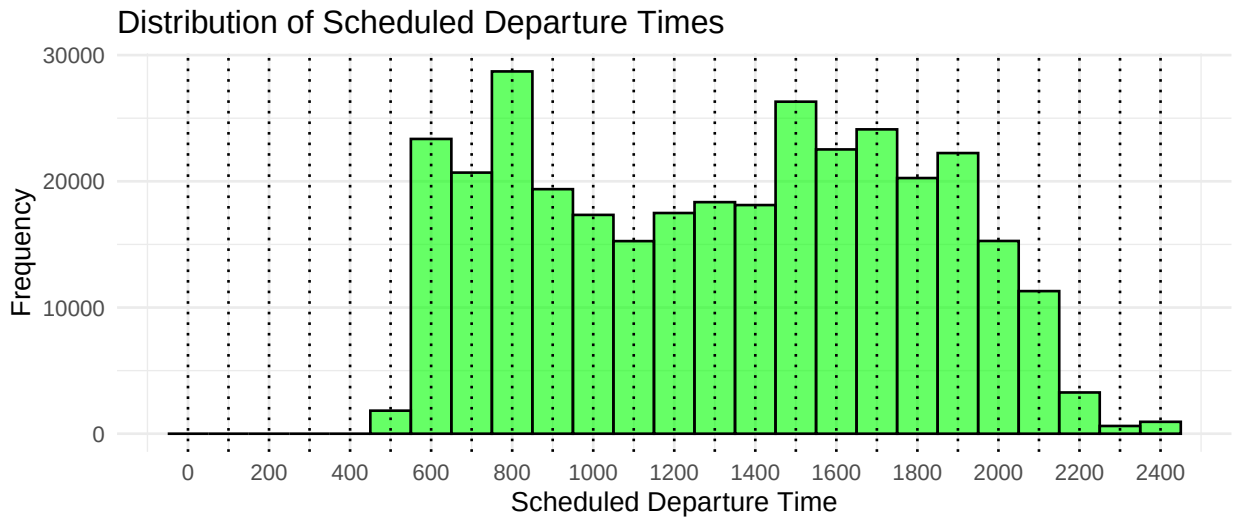
Let's start by studying some of the columns of the dataset.



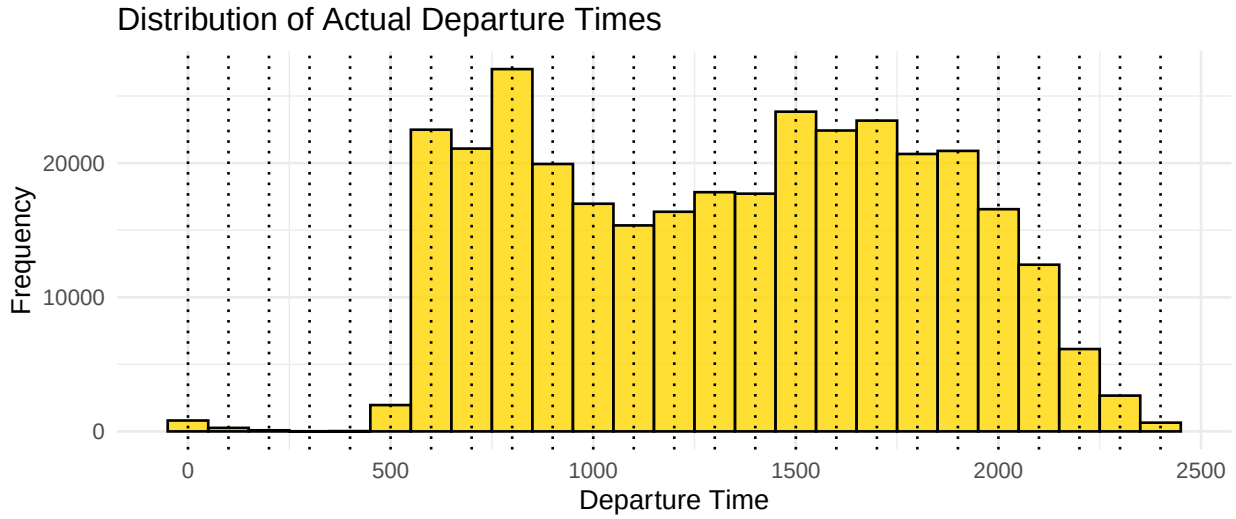
From the graph above, it seems that there are approximately equal number of flights accross months.



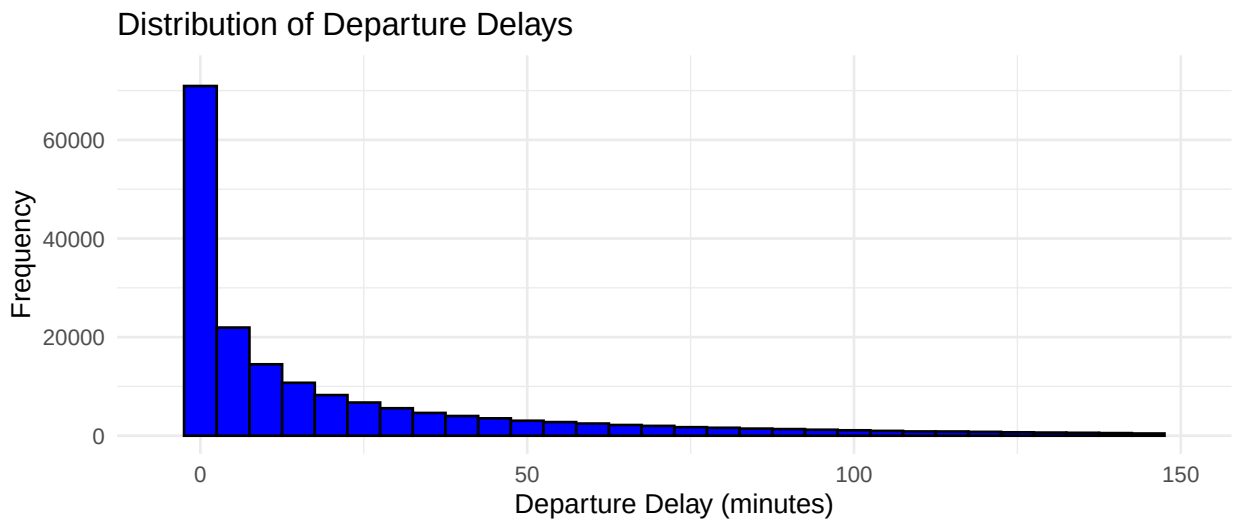
The distribution of number of flights still remains approximately the same per day. Hence our dataset is well scattered across all the days and months almost equally.



Now this is an interesting graph. It shows the distribution of scheduled departure times of flights by their frequency. It seems that around 8 am and 3 pm are the times when the number of flights are at their peak. Also, there are no (or maybe some, but not visible on the graph) flights between 12am and 5 am.

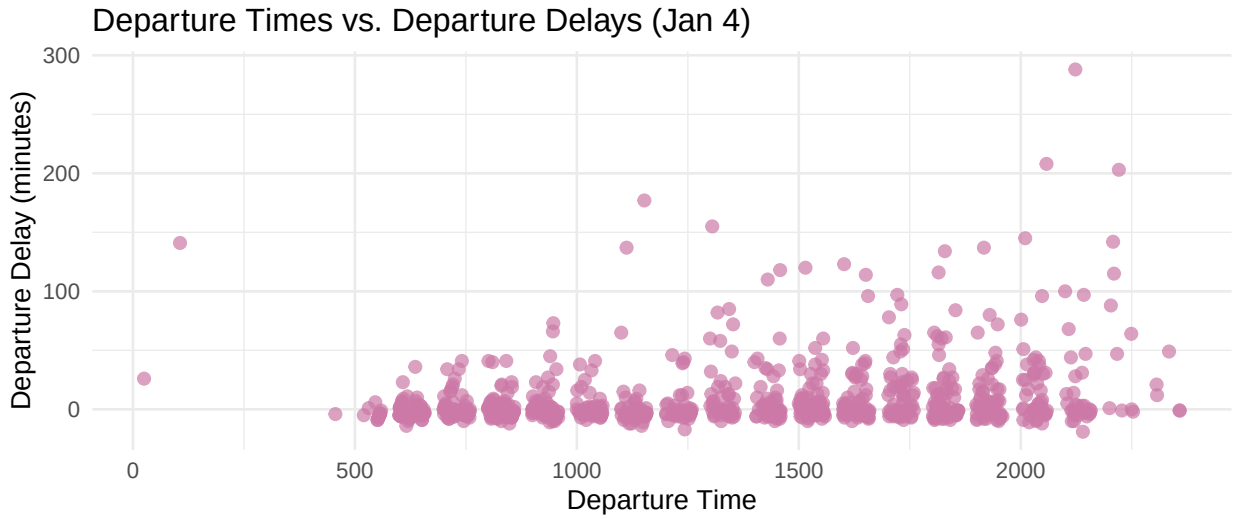


Same inference is also given by the actual departure times of flights. Also, both the above graphs are very similar.



The above graph gives us the distribution of amount of delay in departure of flights. As expected, flights with very long delay in departure will be very few as compared to the ones with relatively small delay. In fact, the graph suggests that the relation would be a decreasing exponential.

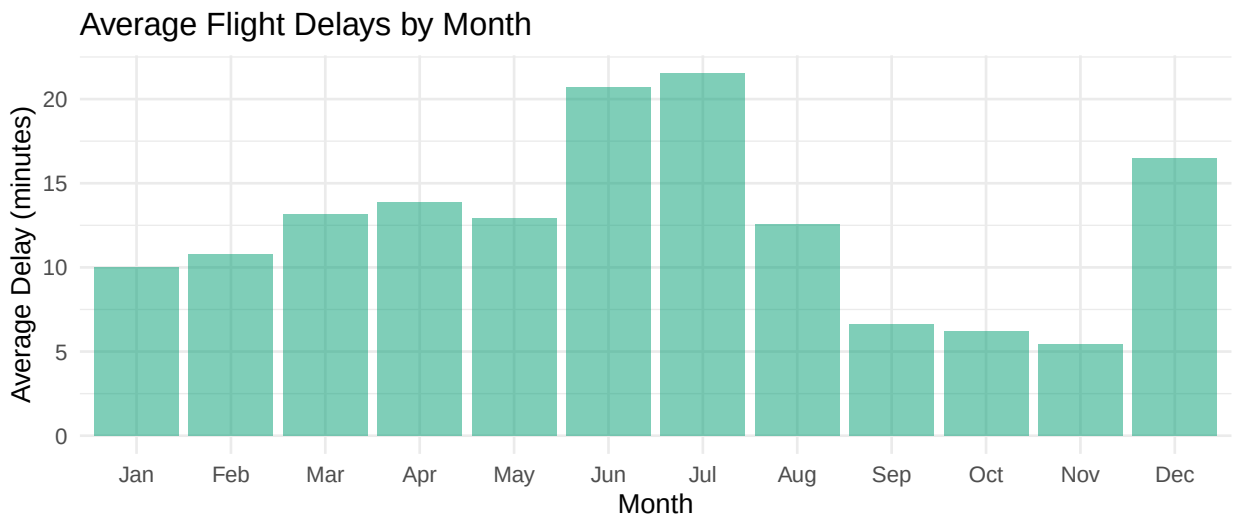
Now let's look at some graphs for a specific month and day. Let's say January 4th.



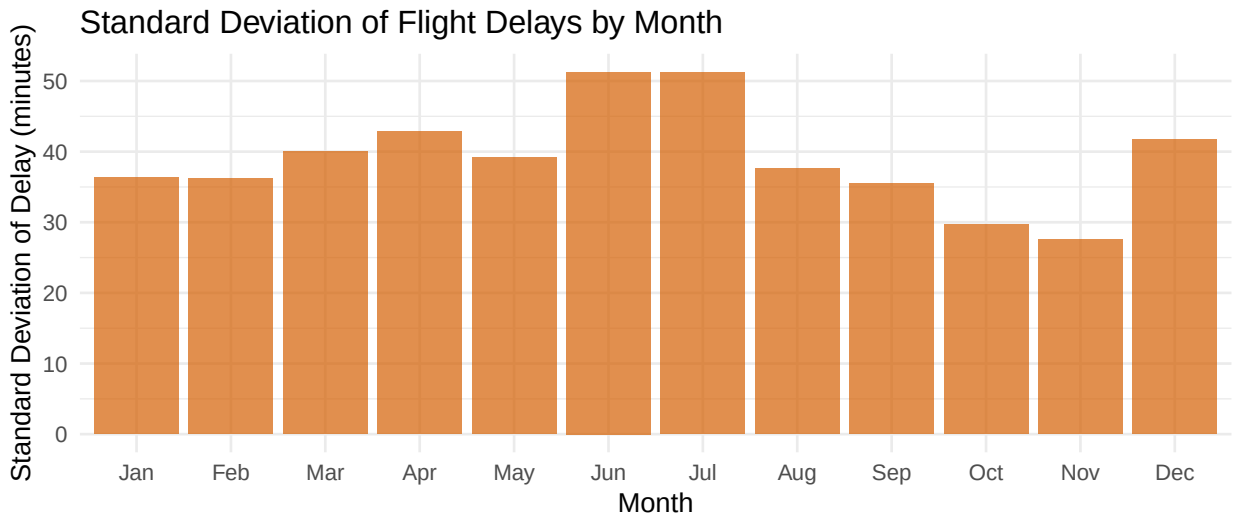
The delays are first concentrated near the x axis around 5 am, when most of the flights start becoming operational. Then the departure in flights keeps on getting scattered as we move to the right, depicting that there are higher flight delays in the evening than there are in the morning. Highest delays are around the midnight. The interpretation of this would be, once there is a delay in a flight, its amount of delay gets added to the previously delayed flights, since the airline can only operate a small number of takeoffs and landings concurrently. Hence any delay in earlier flights affects the delays in future scheduled flights.

Although this is not clear from one graph, but we can plot all the 365 graphs for all the days of every month of this dataset, and still we would get the same kind of distribution almost every time.

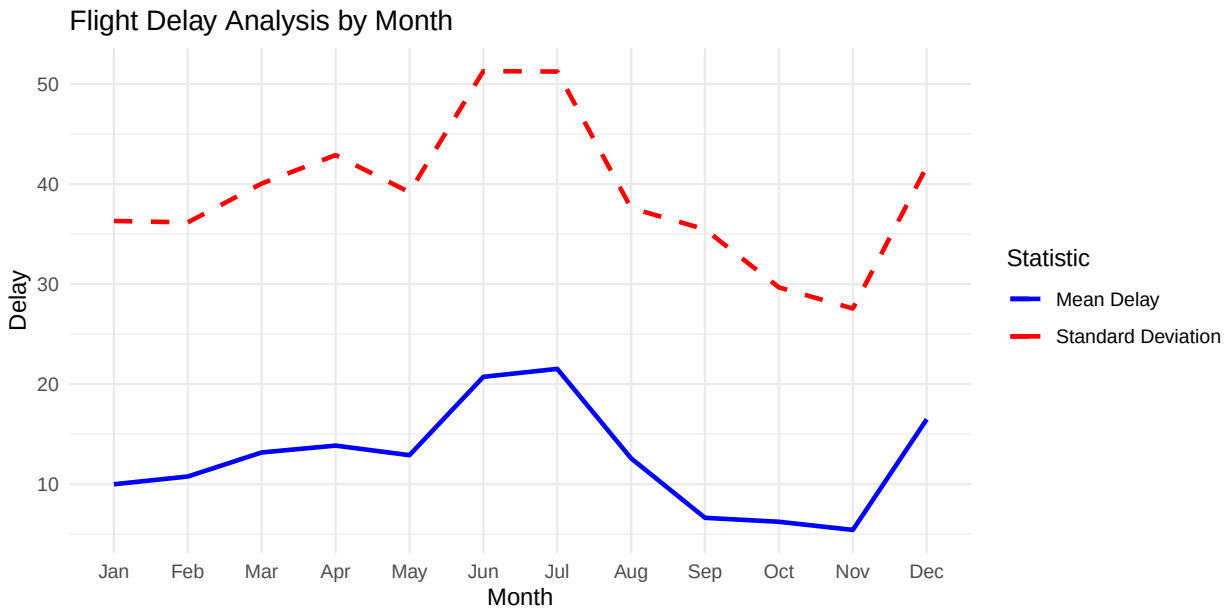
Now let's move to bivariate distributions.



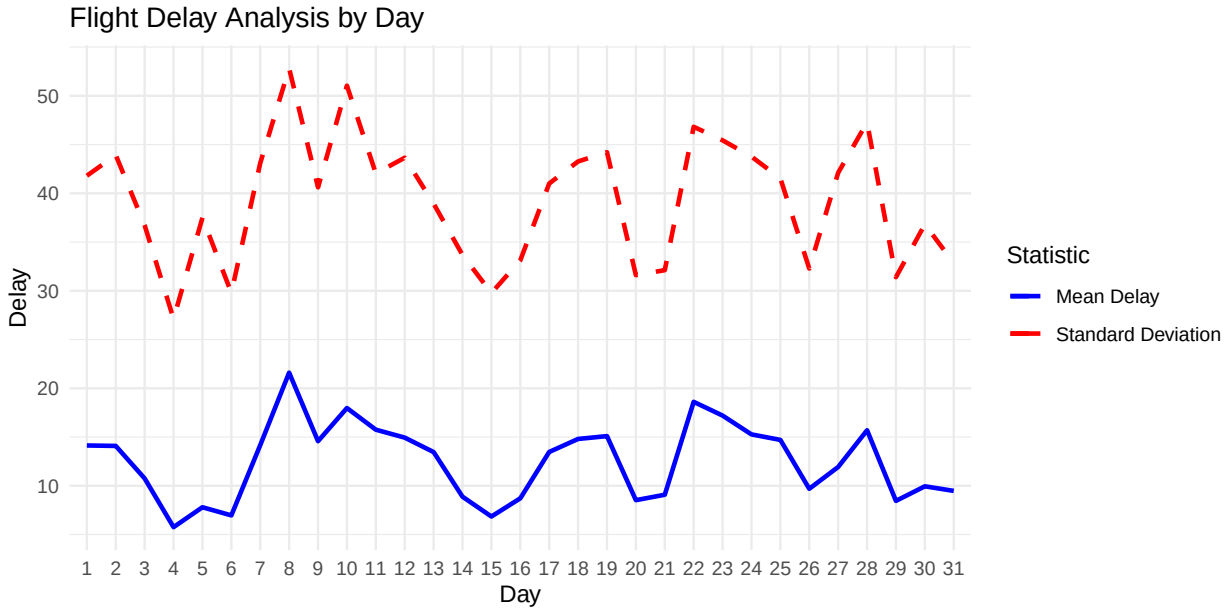
This graph shows us the average delay of flights for each month. The average seems to be the well suited metric here, since there are so many flight records that the outlier flight delays also gets even out. The graph shows a suddden rise in flight delays in the months of June, July and December. There can be many reasons for this, such as holidays in July and December (4th July is independence day for USA and Christmas and New Year in December), so more number of passengers booking and cancelling the flights.



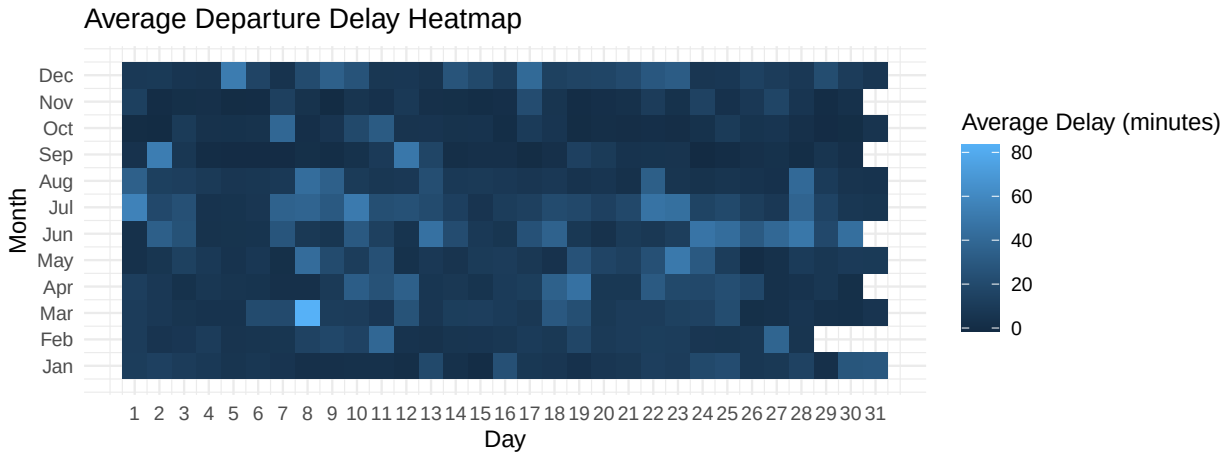
This is the standard deviation of flight delays for every month. It has almost the same distribution as that for mean of flight delays.



The above graph confirms our claim, perhaps too nicely. The mean and standard deviation of flight delays seems to be almost very identical. They are even plotted on the same set of co-ordinate axis. This means that higher the delay of a particular flight, more will be the spread of delay for that flight, i.e., a flight that delays too much also has a high spread in its magnitude of delay.

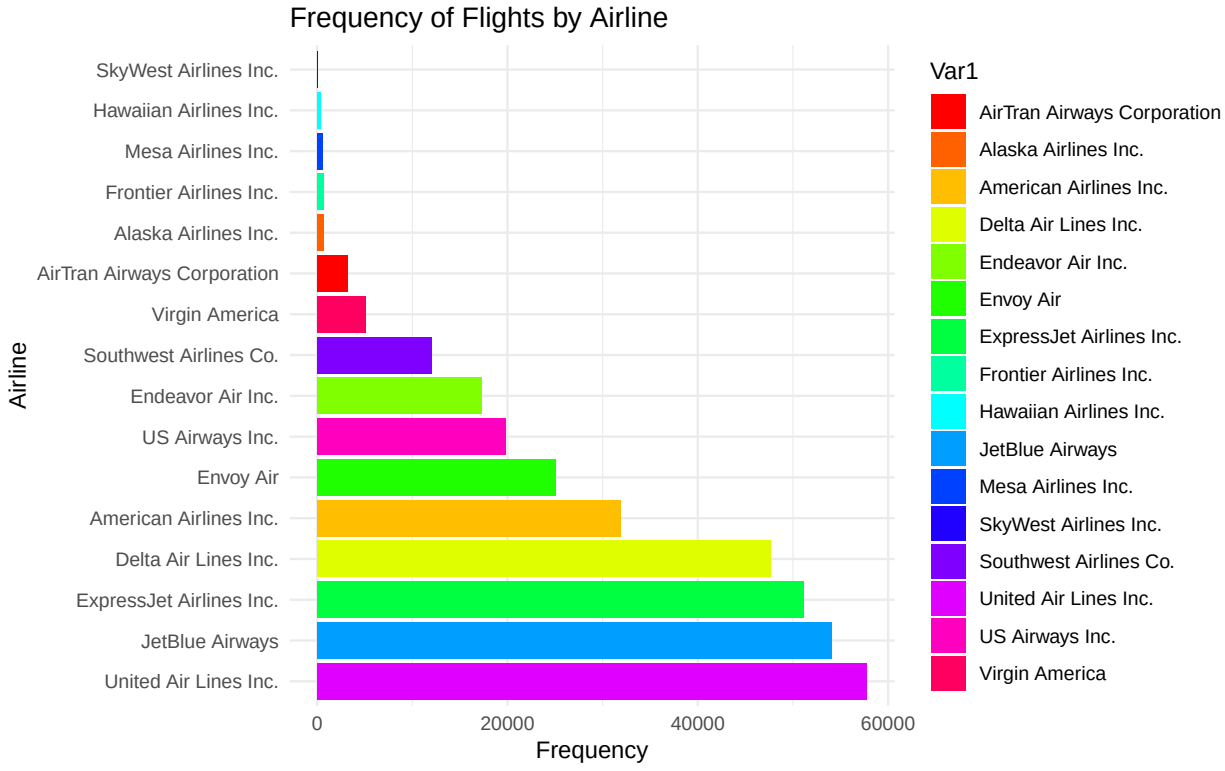


The above graph shows that this trend also prevails across days throughout the year. Hence on an average this trend seems to be the same for every flight of the year.

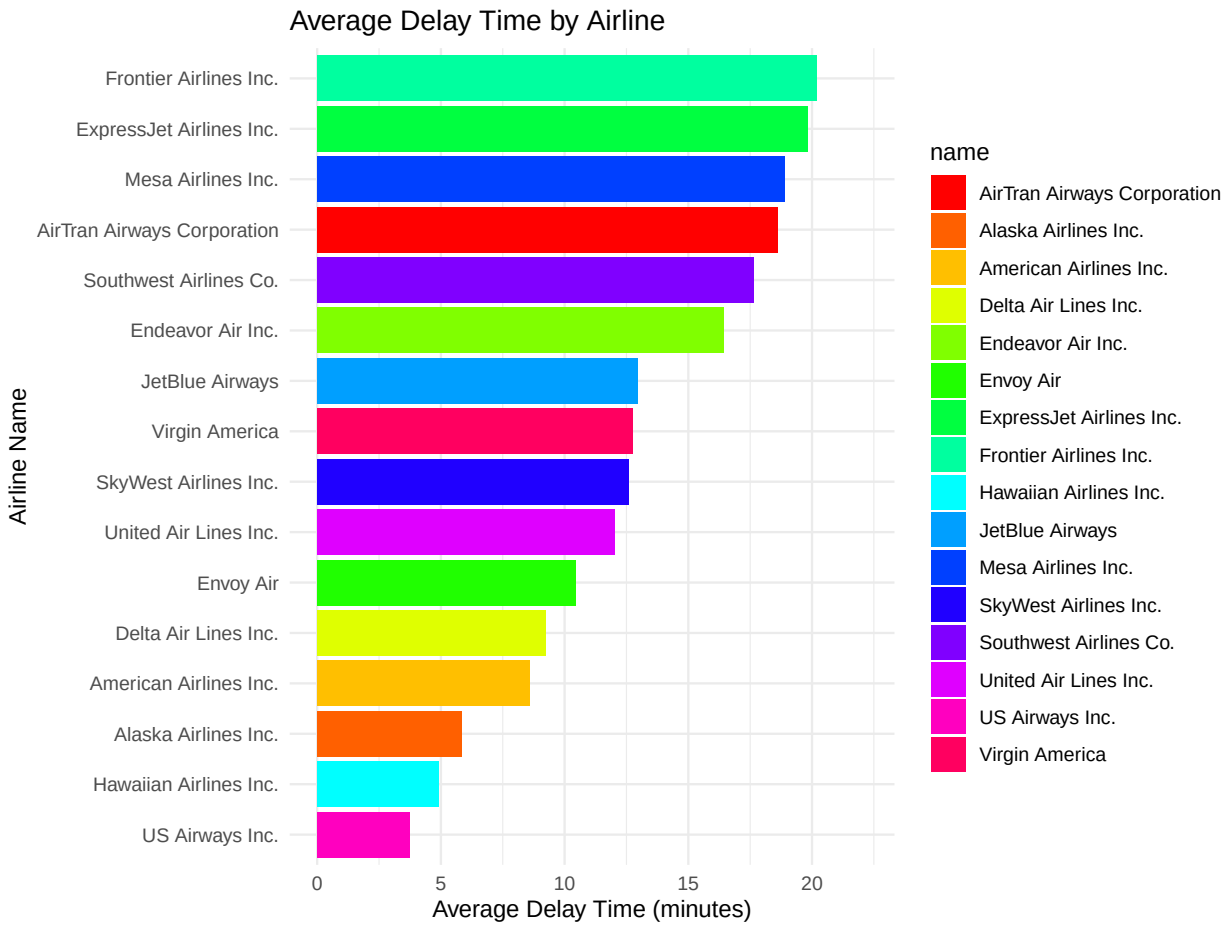


This gives us all the information about average delay of flights accross days and months. Perhaps maximum average delays are around the center, i.e. around June and July. Also there are some noticeable delays in December.

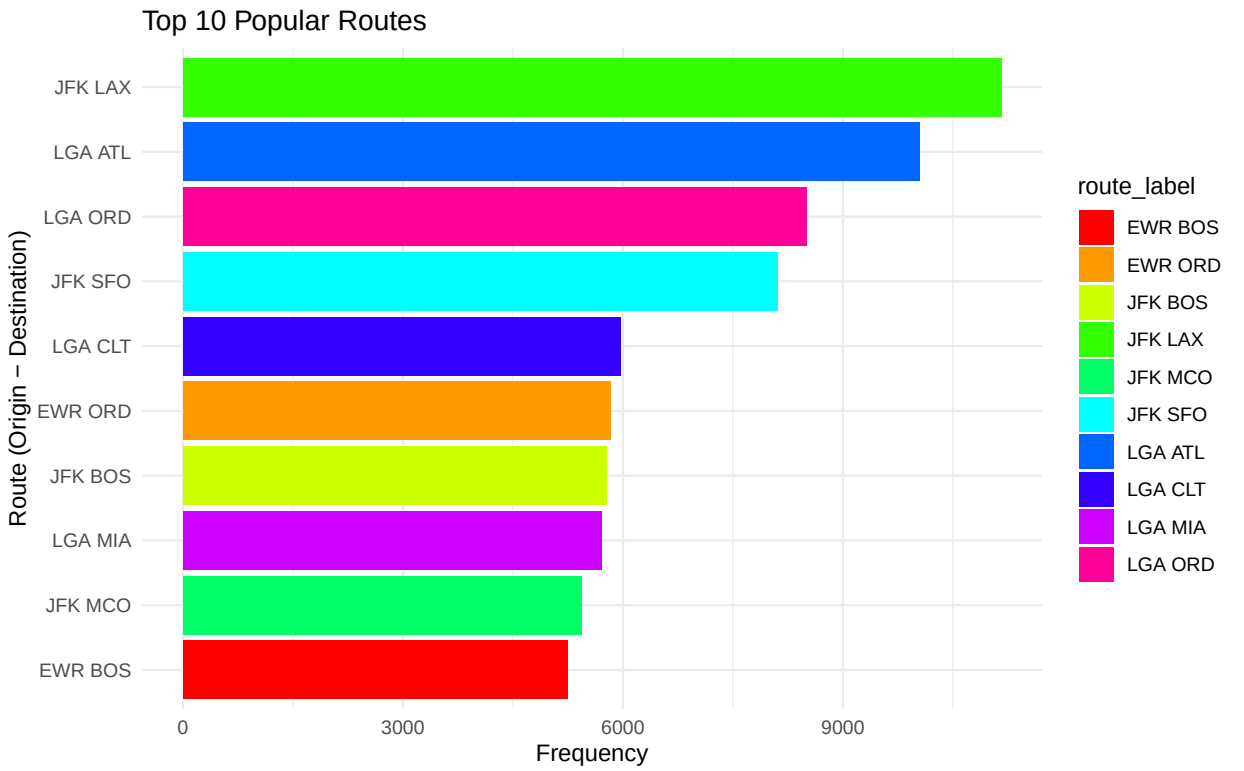
There is also an unexpextedly high average delay on 8th of March. This is not that unexpected. On further research, it turns out that on 8th March 2013, there was a wintery storm that prevailed in the Northeast, which caused many flights to get delayed on that day.



This map gives us the frequencies of flights for each airline. Thus maximum number of flights are from the United Air Lines while lowest are from the SkyWest Airlines. JetBlue Airways is known for its low cost flights, making it more frequent than most.



This graph gives us the average delay of flights by their airlines. Hence Frontier Airline flights has the maximum average delay, while US Airways has the lowest. Although, note that US Airways was operational upto 2015, after which it merged with American Airlines.



Finally, this graph maps the top 10 most popular Flight Routes. Hence the route New York (JFK) to Los Angeles (LAX) is the most popular flight route, followed by New York (LGA) to Atlanta (ATL) and then by New York (LGA) to Chicago (ORD).

Results

Many results can be drawn from our EDA.

- Rise in delays of flights throughout the day.
- Maximum average delays of flights occur around June, July and December.
- The mean and standard deviation of flight delays are almost identical. So a flight with higher average delay will tend to deviate more from that average.
- Unexpectedly high average delay on 8th of March, which turned out to be due to a wintery storm in the Northwest on that day.
- United Airlines is the most frequent flight in the USA, and probably the most favorable flight by people.
- Frontier Airlines with the highest average delay time.

Conclusion

Many implications can be found from these results. First, note that a delay in flight can have many reasons, such as passenger being late for boarding a flight, pilot or staff not being in their positions in time, environmental factors, suspicious activities detected by custom officers, etc. All these reasons seems to be unrelated. So a delay in flight can mean any of the above listed things, or maybe something out of it.

High average delays in June, July and December would probably mean a higher number of such occurrences from our list provided. Hence the Airline industry can take special measures for staff check, custom service efficiency with proper screening and a thorough check on whether conditions around these months.

There were rise in the flight delays throughout the day. For this the airline can install a program which keeps monitoring the flights and starts operating for making everything function efficiently once the flight delay for that day reaches a threshold. This would ensure the program to get proper breaks and to itself function efficiently (since it would not be working continuously).

There can be many more measures that can be taken for efficient running of the airline industry, which might be found from this dataset by a more elaborate EDA and including other columns that we left in our project.