Variability in Flight Delays

Sameer Korlahalli (ssk698), Dhaval Popat (dkp288), Ronak Thakkar (rht241 - Online)

Project description

Variability in flight delays can be divided into various aspects such as temporal, regional, airlines etc. We aim to explore the variation of delays in some of these aspects to find out if any patterns exist because if they do, future research might possibly help in considerable reduction of delays. For example, there could be delays pertaining to outbound flights from one or a cluster of regions. We can find out individually the regions or the times during the year when least number of delays occur through the components of the dataset. We hence aim to explore if such patterns exist in the given dataset and communicate these results to the users.

Data Set

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, cancelled, and diverted flights are published in DOT's monthly Air Travel Consumer Report and in this dataset of 2015 flight delays and cancellations.

The dataset comprises of 3 different CSV files - airlines, airports and flights. We plan on using several attributes from these files for our visualization storytelling such as:

YEAR, MONTH, DAY, DAY_OF_WEEK: dates of the flight
AIRLINE: An identification number assigned by the US DOT to identify a unique airline
ORIGIN_AIRPORT and DESTINATION_AIRPORT: code attributed by IATA to identify the airports
SCHEDULED_DEPARTURE and SCHEDULED_ARRIVAL: scheduled times of take-off and landing
DEPARTURE_TIME and ARRIVAL_TIME: real times at which take-off and landing took place
DEPARTURE_DELAY and ARRIVAL_DELAY: difference (in minutes) between planned and real times
DISTANCE: distance (in miles)

Analytical Questions

1) Which are the busiest airports?
Proxy Task: What is the total number of inbound and outbound flights at each airport?
Proxy Values: busiest -> total number of inbound and outbound flights

2) What is the effect of busiest airports on flight delay?
Proxy Task: Does the total number of inbound and outbound flights at an airport correlate with an average delay in flights at that airport?
Proxy Values: busiest -> total number of inbound and outbound flights; effect -> is there a correlation; flight delay -> average delay in flights each day

3) What is the temporal variability in flight delays?
Proxy Task: Is there any pattern in average flight delay every month over the years?
Proxy Values: temporal -> month of the year; variability -> any pattern over the years

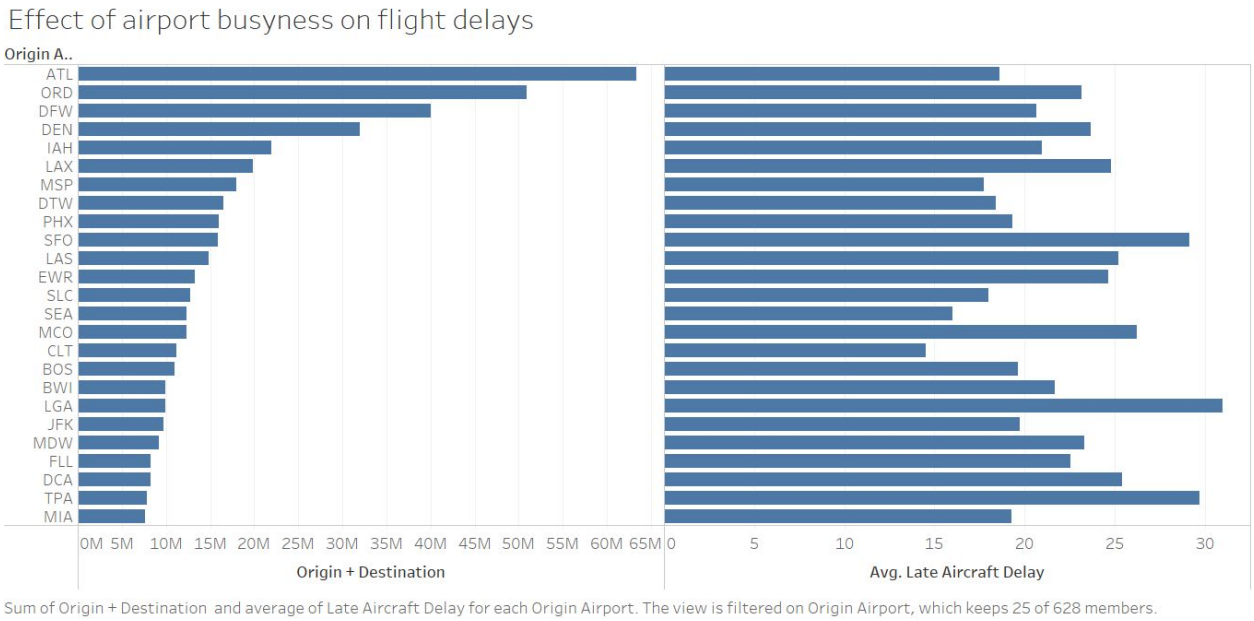
4) How does the top cancellation reason for flights vary over time?
Proxy Task: Is there a pattern in the most frequent cancellation reason for flights for each month over the years.
Proxy Values: top -> most frequent; vary -> any pattern or changes each month; over time -> over the years

5) What is the best combination of airline, region, and time?
Proxy Task: What combination of airline and airport, along with which month of the year results in minimum flight delay?
Proxy Values: best -> leads to minimum delay; region -> airport; time -> month of the year

Story Design

Data Analysis

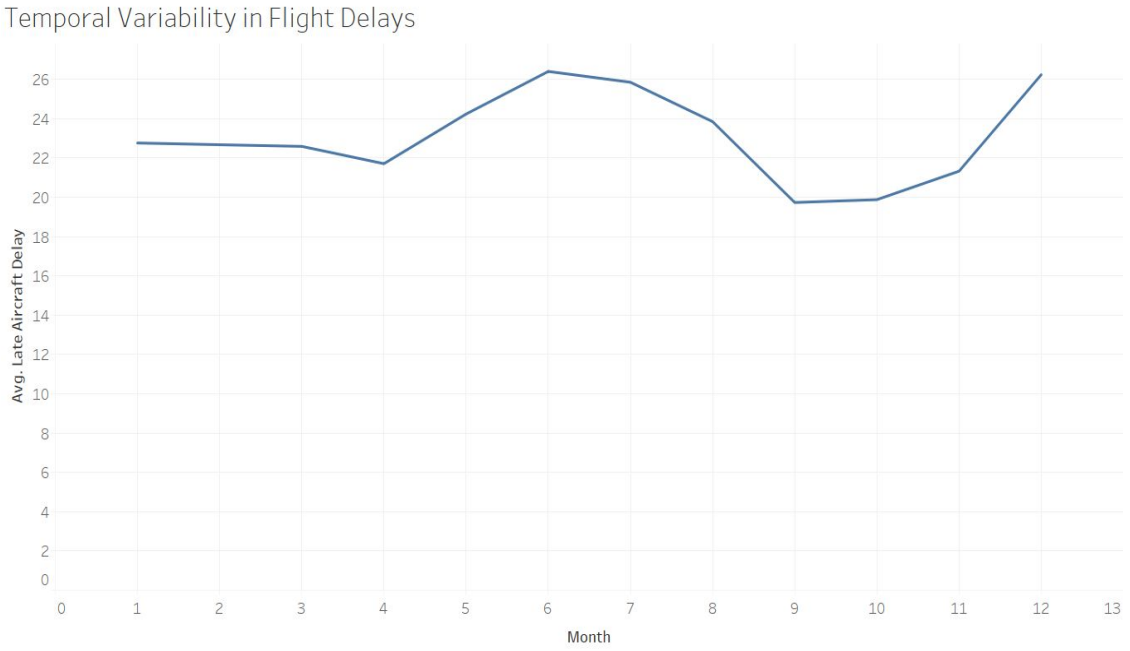
Which are the busiest airports?
What is the effect of busiest airports on flight delay?



From the above visualization, our aim was to determine whether there is a correlation between the busyness of airports and the average aircraft delays occurring at each airport.

We did not find any patterns and observed that the delays occurring at most airports are uncorrelated with the amount of traffic handled by the airport. Hence, we did not include this relationship in our story design.

What is the temporal variability in flight delays?



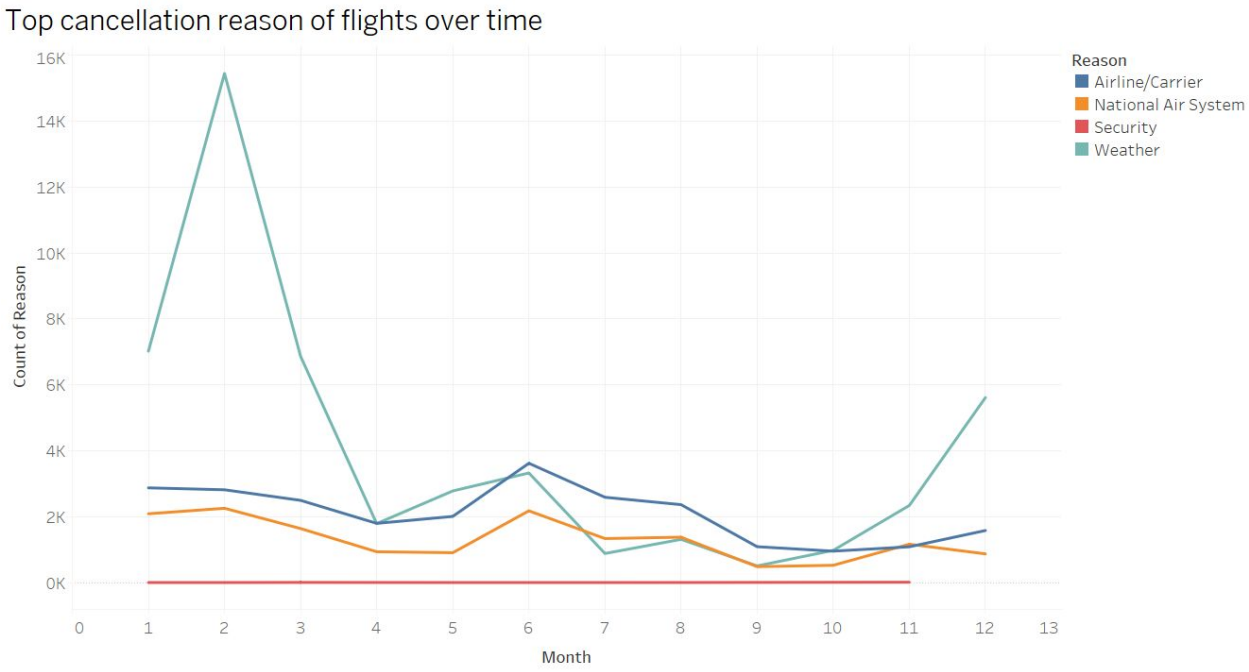
The trend of average of Late Aircraft Delay for Month.

Our aim was to determine whether there is any pattern in average flight delay over the year 2015. Our intuition was that there might be an increase in the average delay during the holiday seasons of the year.

It can be seen from the above visualization that there is a significant rise in airline delays during summer up to fall season (April to September) due to summer holidays.

We also noticed a sharp increase in delays during the Christmas holiday season, which records the highest average delay.

How does the top cancellation reason for flights vary over time?



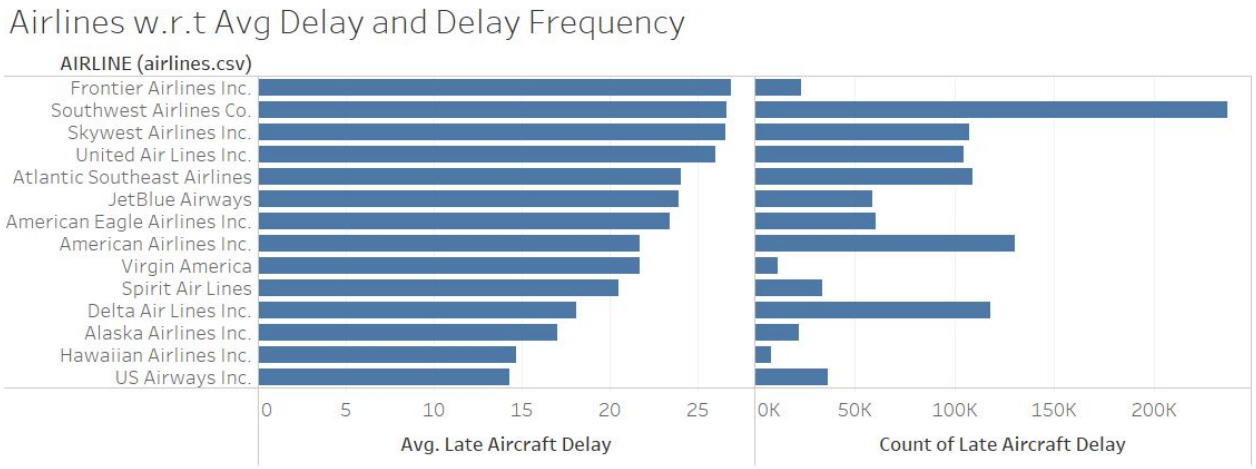
The trend of count of Reason for Month. Color shows details about Reason. The data is filtered on Cancellation Reason, which has multiple members selected.

From the above visualization, we wanted to determine the most frequent cancellation reason for flights for each month. We observe that weather is the primary cause for delays during the month of December to April. This is attributed to the weather conditions during those months.

We also found a loose correlation between all the plots except Security, which barely showed any frequency.

As suggested by Prof. Bertini, we did not include this question in the story design section because it was not aligning well with the rest of the story flow.

What is the best combination of airline, region, and time?



Average of Late Aircraft Delay and count of Late Aircraft Delay for each AIRLINE (airlines.csv).

Finally, we give a comparison of the airlines in terms of the average duration of delay and the number of delays that each airline makes.

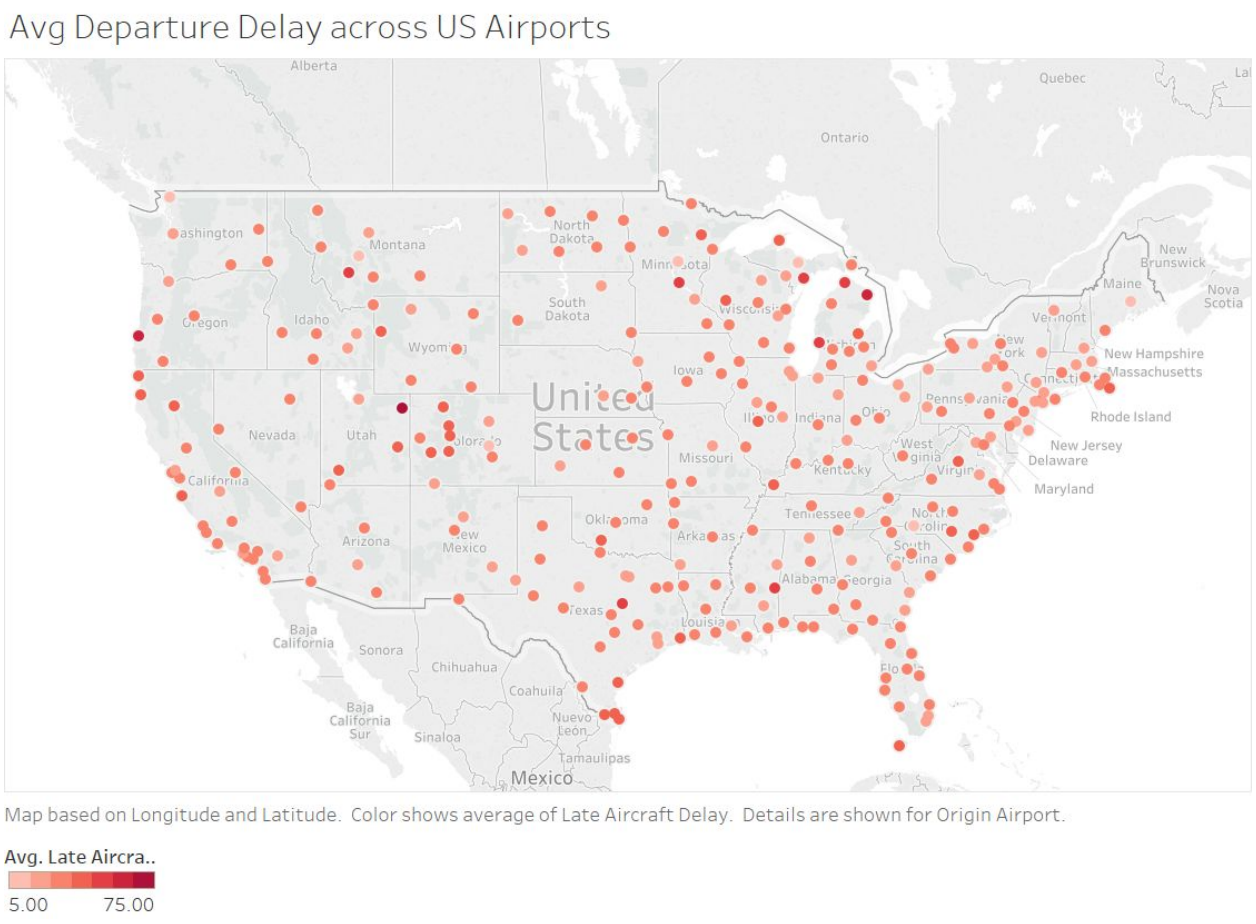
This gives us a perspective as to which airlines have an optimal combination of the number of delays as well as the average duration of all delays.

Storyboard (Sketches): Timely Flights For You!

Initially, before jumping into the insights about the flight delay, we have added an interesting fact on our website that would educate the user including the statistics regarding the total number of flights in a year.

Question: What is the average flight delay across all airports in the US?

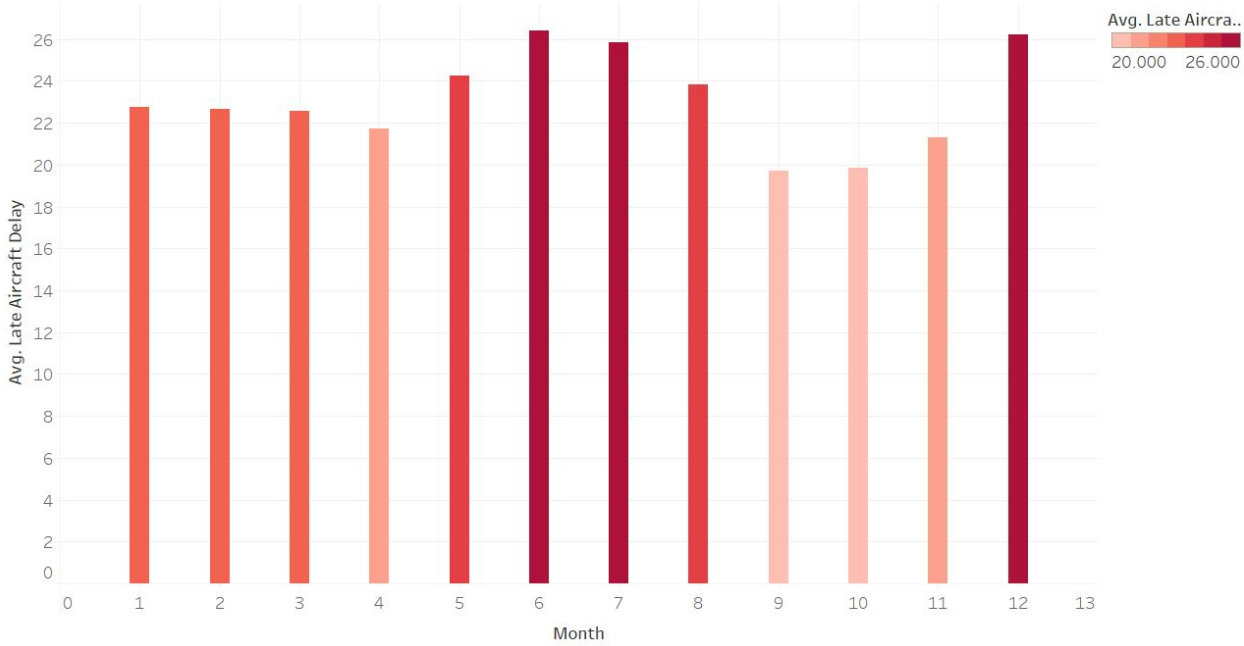
The delay section of the story starts with displaying the US map with average flight delay of each airport. We are using bubbles to represent each city and its colour intensity is proportional to the average flight delay. The flight delay signifies net delay that we had in the dataset (called as *late aircraft delay*).



Based on the feedback, we have modified the above visualization in our implementation. We have also added the total number of daily flights at each airport which is represented by the colour intensity, whereas delay is now indicated by the size of the circle.

Question: Does the time of the year affect flight delay?

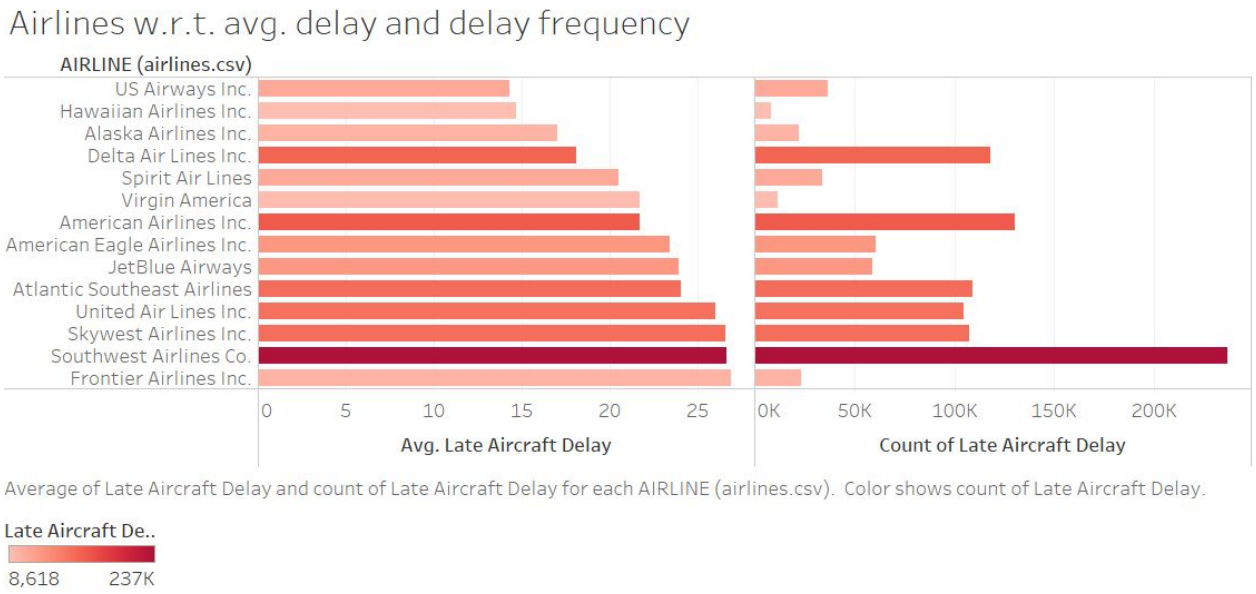
The below visualization gives insights about the average flight delay throughout the year in different months. This helped us in gaining information about delays for particular months, which might be helpful for users while planning future travels.



`

Question: How does the average flight delay vary for different airlines?

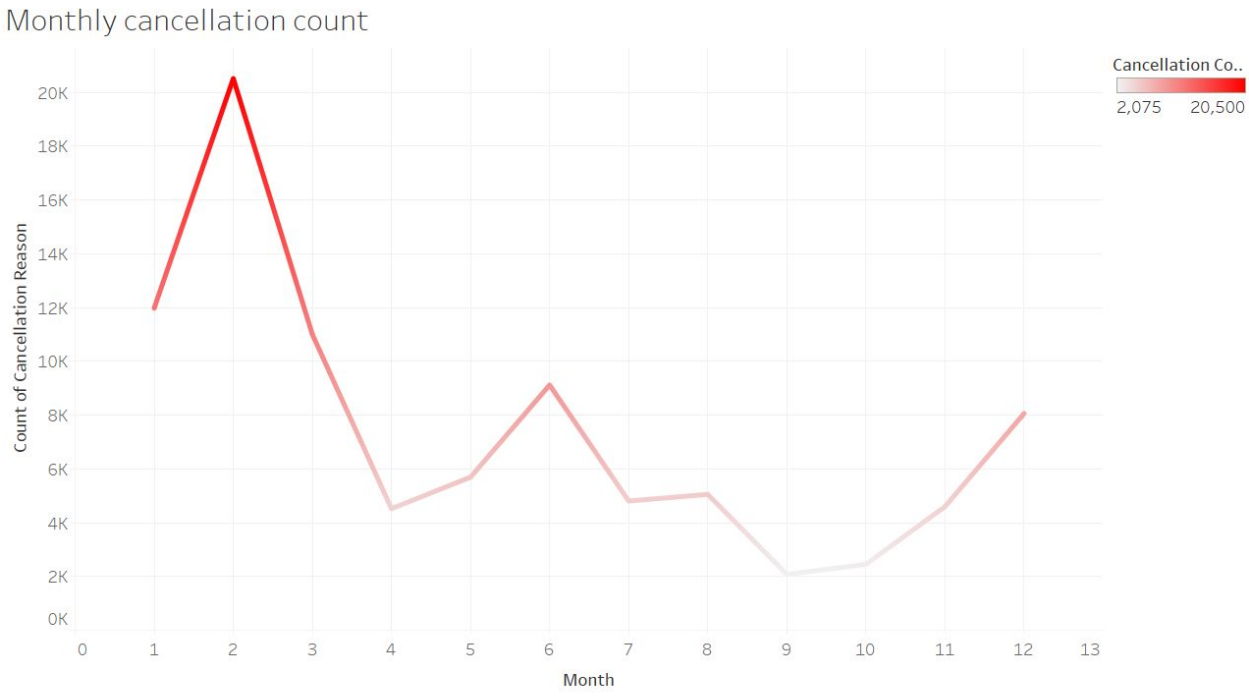
The next section of the story will give users an overview of the average delay for all the airlines across all regions in the U.S. This visualization will look something like the following graph.



In the above graph, the length of the bars represents the average delay while its colour intensity represents the frequency of the number of times the airline got delayed in the selected month. These two parameters help the users in gaining more insight and understanding the trade-offs which will eventually help in the decision-making process while selecting an airline.

Question: Is there a correlation between average flight cancellation and the time of the year?

The next section of the story covers cancellation of flights and gives insights about the monthly cancellation count of all the outbound flights. This gives a sense of the probability of cancellations taking place in a given month. As we can infer from the graph, there are peaks for the month of February, June and December.



The trend of count of Cancellation Reason for Month. Color shows count of Cancellation Reason.

Based on the feedback by Prof. Bertini, we have not included this graph in our implementation as cancellation doesn't align well with the overall story. We have instead tried to analyze the relationship between airline and delay in each month of the year by designing a multi-line graph.

Implementation

We have developed a website which showcases the required graphs generated using D3.

GitHub Repository:
https://github.com/NYU-VIS-FALL2018/timely-flights

Demo Page:
https://nyu-vis-fall2018.github.io/timely-flights

Changelog

Initially, the story was personalized for each user and the high-level project goal was to suggest airlines with minimum delay to the users based on airport and month of travel. The first graph would have allowed user to go into further inspection by selecting the airport of their choice while also giving insights about overall average delay in each region of the US. This would have led the user to the second visualization i.e. the next complementary part of the story. Here, the average delay (on y-axis) will be exclusively for the state selected by the user for each month. Here, the user will have an option to select any of the month as a parameter for the further story. In the final section, the user would have gained insights on the best airlines for their selected airport and month. We thought that the personalization would have helped the user in connecting with the story and aid the user in making a much more informed decision.

However, based on the feedback of phase 2, we realized that the story was very open and the storyboard become like an exploratory tool. Hence, we modified the flow of our storyboard to make it more connected and coherent, while also providing interesting insights and facts.

Changes made to individual graphs or visualizations, if any, have been described in their own sections.