Overview and Motivation

Our motivation for doing this project began with a discussion our group members had about spring break and our travel plans. We all traveled to places that forced us to take a plane. Some of us had preferred carriers to fly from, others didn’t care as long as it was the cheapest flight. We also talked about other factors that could influence our decision like our preferred airports, if a day of the week was less crowded for traveling, what cities we should expect a delay for, etc…

Airline carriers compete for customers in a lot of different ways: advertisements, reward programs, word-of-mouth, claiming they have the safest or cleanest flights, always arriving and departing on time, etc… By looking into data about carriers and these variables, we thought it would be interesting to analyze how the carriers and airports and other factors actually compare to each other.

We found a great dataset to work with on The Bureau of Transportation Statistics website that we knew we could trust and would be a reliable source to cite. In addition, this website allowed us to interactively grab which variables we wanted to download along with the month and year we wanted to focus on. It also included a lot of variables to choose from so we were able to expand upon what we originally thought we wanted to explore. There was an overwhelming amount of observations, since there are an overwhelming amount of flights, so we knew we had a big enough dataset to work with to create valuable insights from.

Related work

While we didn’t get motivation for this project from seeing other websites or blogs about flight data, we did get motivation from the types of visuals we have learned about and seen in class. Besides bar graphs, histograms, scatterplots and other basic plots the visual that stood out to use the most that we thought would be a great fit to our project, was putting data into a US map and creating a visual utilizing location and another variable, like the number of flights from that particular city. Having taken other statistics and analytics classes, we’ve seen visuals like this that allowed us to envision how we could show insights on a graph like this. Also, we coded departure graph as red and arrival graph as blue which would be easy to compare. Besides, we put the same type of graph horizontally so that we can compare the related data. Such as average departure delay by day of the week VS average arrival delay by day of the week. Then we put median of each group of data below the average graph so that we can assessing both average and median at the same time.

Questions

We are interested in 4 questions related to this dataset.

1. *The delay time by cities*

We aimed to see the average arrival/departure time of a flight with a certain airline in a certain month. We used departure city, arrival city and month as the filters for the user to select the airlines and the days they are interested in. In this case, the customers are able to schedule better in order to plan their travel.

1. *The average/median arrival/departure delay time by day of the week*

We want to know which day of the week is more likely to have a highest average flight delay time. Customers will be able to avoid the most busy day and improve efficiency. The reason we include median is because that we want to eliminate the effect caused by outliers.

1. *The average arrival/departure delay time by carriers*

Picking an airline company is important for the travelers. When the airline companies are offering similar trip plans, the time of delay for each carrier becomes an important factor of how the customers will make a decision.

1. *The average delay by reasons*

We want to know what seems to be the most frequent reason in causing a flight delay. Airline companies need to provide a reasonable explanation to the customers whenever there is a delay and customers are concerned why their flight got delayed. For example, if a certain airline company has most of their flight delay due to security reason, then the customer is less likely to choose this airline company.

Data

We obtained this data from The Bureau of Transportation Statistics provided by the U.S. Department of Transportation. With millions of flights flying in and out of the United States, we thought it would be interesting to dig deep into the data to see what kinds of trends we might be able to find.

The data allowed us to easily download csv files based on what month and year we chose. Our decision to use all months from 2016 was based on the fact that we wanted the most recent data to see what’s been going on and what might go on in the future. We wanted to see trends over one year and if there were any patterns we could pick out. With over 100 variables to choose from we realized we had to narrow down which variables to pick. We decided to pick variables based on delays and cancellations, airports, arrival/destination cities, flights and carriers. We picked variables to explore that are relevant to the questions we wanted to answer and ones we thought would be interesting to those looking at our dashboard.

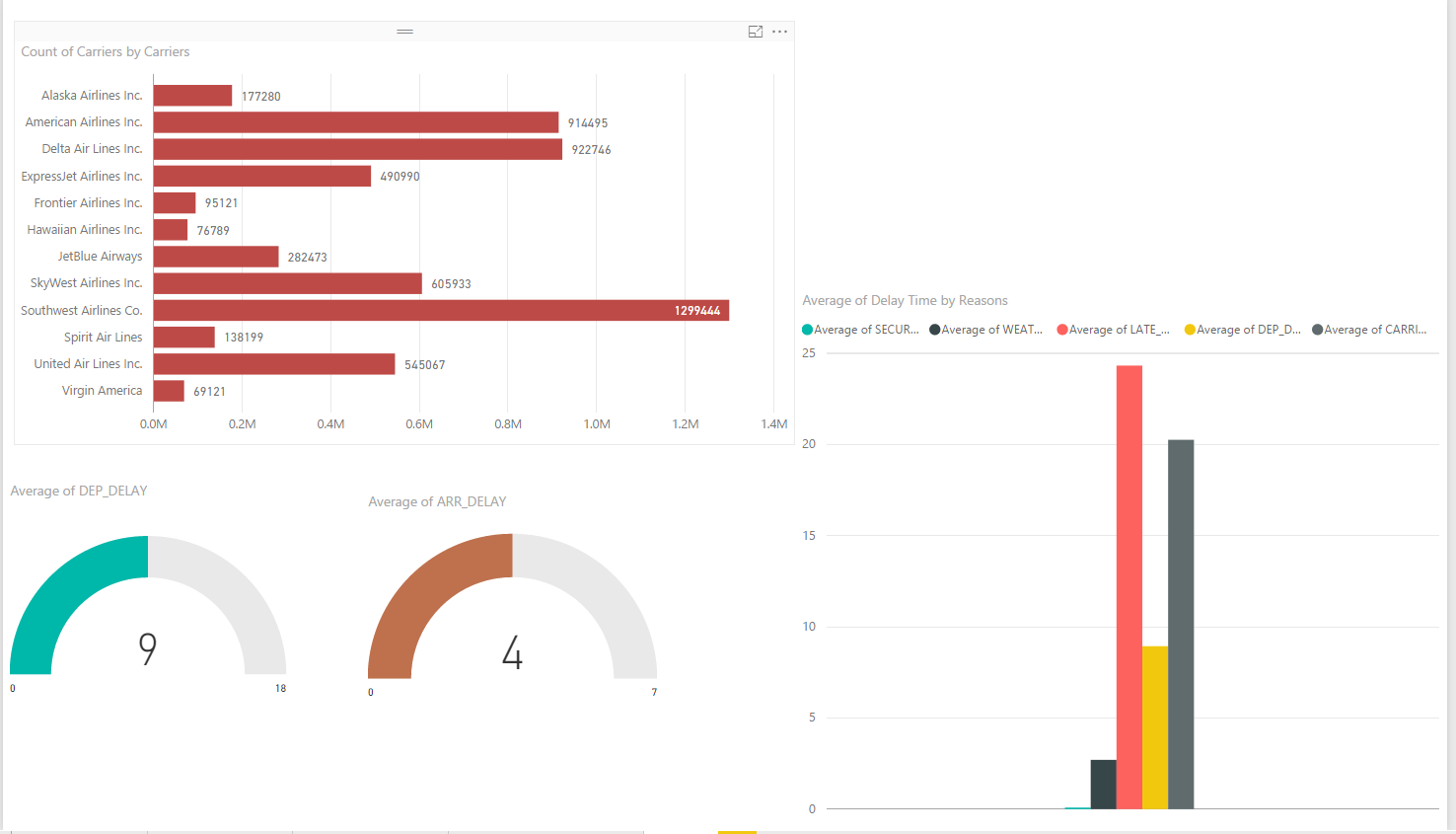
Below is the link to the website we used:

<https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time>

Exploratory Data Analysis

At the beginning, we were interested in a slightly different questions. However, with developing of our data visualization design, we found that some other questions are more intriguing and useful. So we end up doing couple bar charts and one map to illustrate our story.

The following graph is an overview of the data we use:



Design Evolution

From the beginning, we knew we wanted to make this dashboard personable so that anyone looking at it could relate it to themselves based on their specific travel plans. The way we went about doing this was creating multiple filters so that you could look at the visuals that showed data relevant to your own experience and see how things would change if you were to travel elsewhere.

The original sketch of our dashboard is presented below:



Implementation

Our first step was exploring the data and decide what variables we are going to focus on. And since there was data related to couple years, we have to decide which months, seasons, or years we are going to analyze. After settling down our business questions, we decide to use 2016 data set which is the newest whole year data we can get. Then we picked arrival, delay, cancel, day of a week, carriers etc. as our focus variables. To get the data ready for analyzing, we first did data cleaning. We combined all months data into one year data file. Then we changed month from number into words and the abbreviation of carrier were changed into carrier full name, which we think can enhance the understandability of our visualization. Also we changed some data types which were mismarked. After that, we put data into Power BI and built couple graphs and charts.

Evaluation

Our objective was to determine where, which carrier, and why an airline (not carrier) delay the most. The result of our graph not only answers those questions but also interactively helping user compare different airline delay in what city and for what reason by implementing filters in the graph.

For example, if the user chose the to leave from Oxford to Chicago in May. We will know that if we leave from Cincinnati, we will have a relatively shorter departure delay compared to leaving from Dayton. Also, Wednesday has both the highest average departure and arrival time and Saturday has the lowest average departure and arrival time. Carrier wide, SkyWest airlines Inc has the highest average departure delay time and arrival delay time. This most likely delay reason for this flight will be national security delay.

There are some limitations for our system:

* Our system is unable to check two or more departure cities at the same time, since it will automatically filter out the airlines that do not fly from both cities. We will have to check them separately and then compare them.
* When we download the data, we did not include the “flight time” data, which are datas that record the specific time of the flight. In this case, we are unable to analyse any patterns based on the time of the fight.
* There is no international flight information. All the flight data that we analyzed are all within the US. So if the users want to see the flights that fly outside the US, we are able to provide any recommendations.
* The dataset that we are using does not include transfer flight informations. They are all non-stop flights. In this case, customer who are looking to switching flights will have to check for the airlines individually.

Video