Flight Delay Analysis

Akshata Abhay Bhende

22: 544: 646: 40 Data Analysis & Visualization

May 08, 2020

Rutgers State University, Newark

Flight Delay Analysis

One of the most common yet unpleasant encounters people dread to have is a flight delay. A lot of flights are delayed each year, which involves some cost in different ways both for the airline as well as the passenger. The time and money of the passenger get affected, and at the same time, the credibility of the airline is at stake. Delay is regarded as one of the airline's most recognized performance benchmarks. There may be some unavoidable factors, such as environmental conditions, air trafficking, or some unexpected events like the ongoing lockdowns because of COVID 19 Pandemic. Still, there may also be some ground that can be solved by improving the process. Hence flight delay statistics play a vital role in understanding the efficiency of a flight.

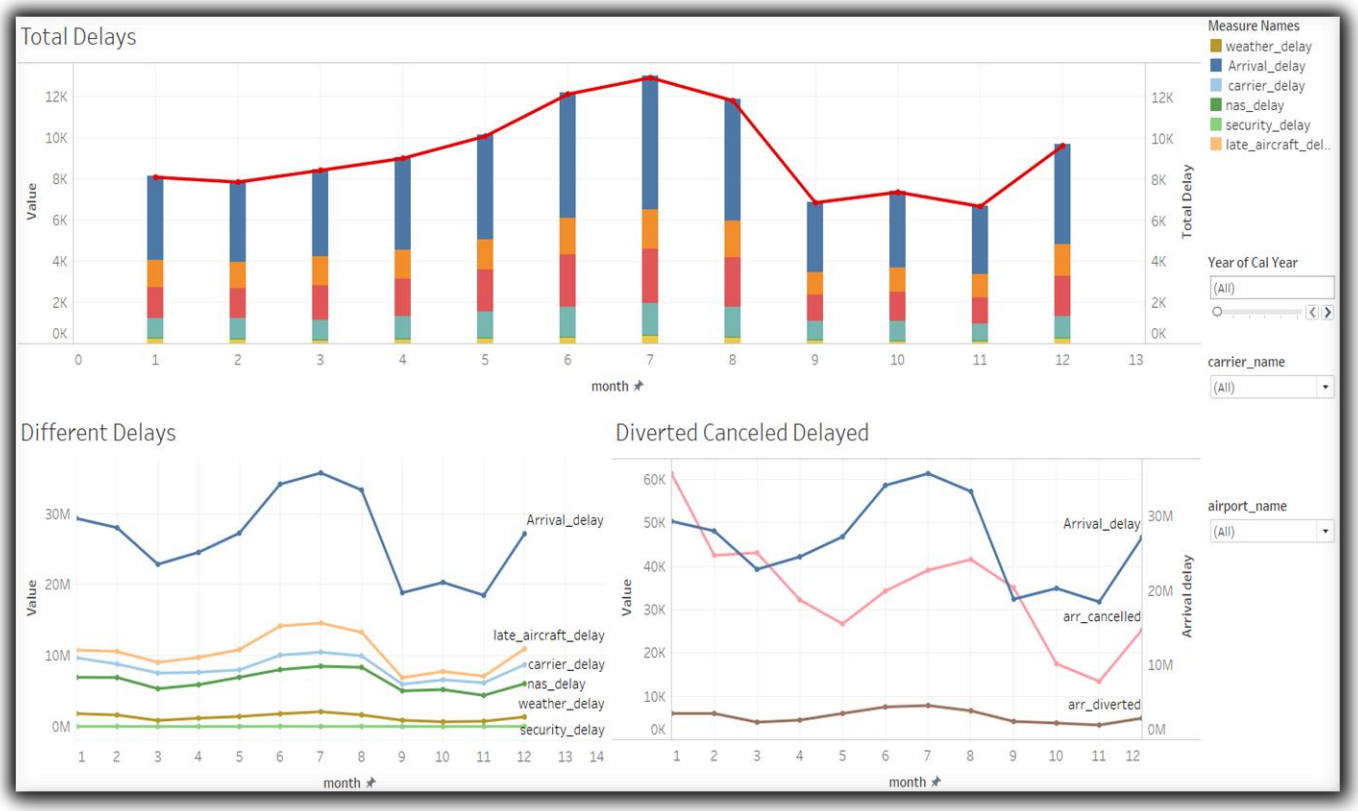
This study presents the analysis driven from flight delay data for the United States for the year 2016-2020. This study also analyses the variety of factors responsible for and associated with flight delays for different airlines.

The dataset for this project is obtained from RITA website, which contains information about flight delays and performance. The dataset I used ranges from 2016-2019.

The core objective of this project is to answer the following questions:

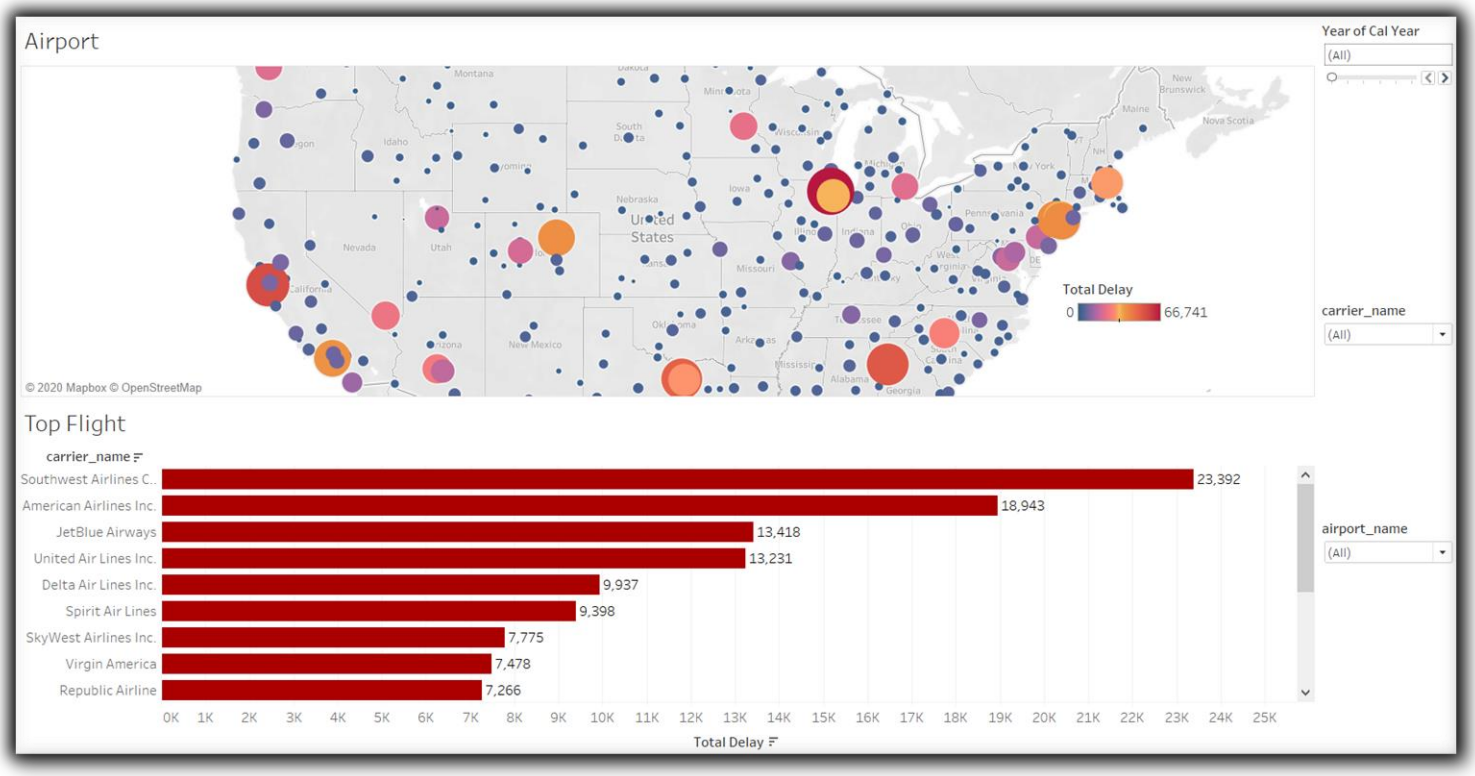
- What are the various types of flight delays or canceled/diverted arrivals, and how they are spread over months, types of carriers and airports?
- Who are the top 10 airports and airlines liable for delays, and how do they affect each other when combined?
- Which are the carriers that contributed towards most delays when all the airports are considered?

The first visualization was to breakdown various kinds of delay and saw the contrast between that and how much each contributed to the overall delay. We also need to break down further by months and see if the delays have a high and low point within a given time of a year. For this reason, I created a calculated field called 'Total Delay', where I added up all the different delays. And also used 'Total Delay' as a point of reference in my visualization. Here we can also filter by airport, carrier, and year to drill down even further.



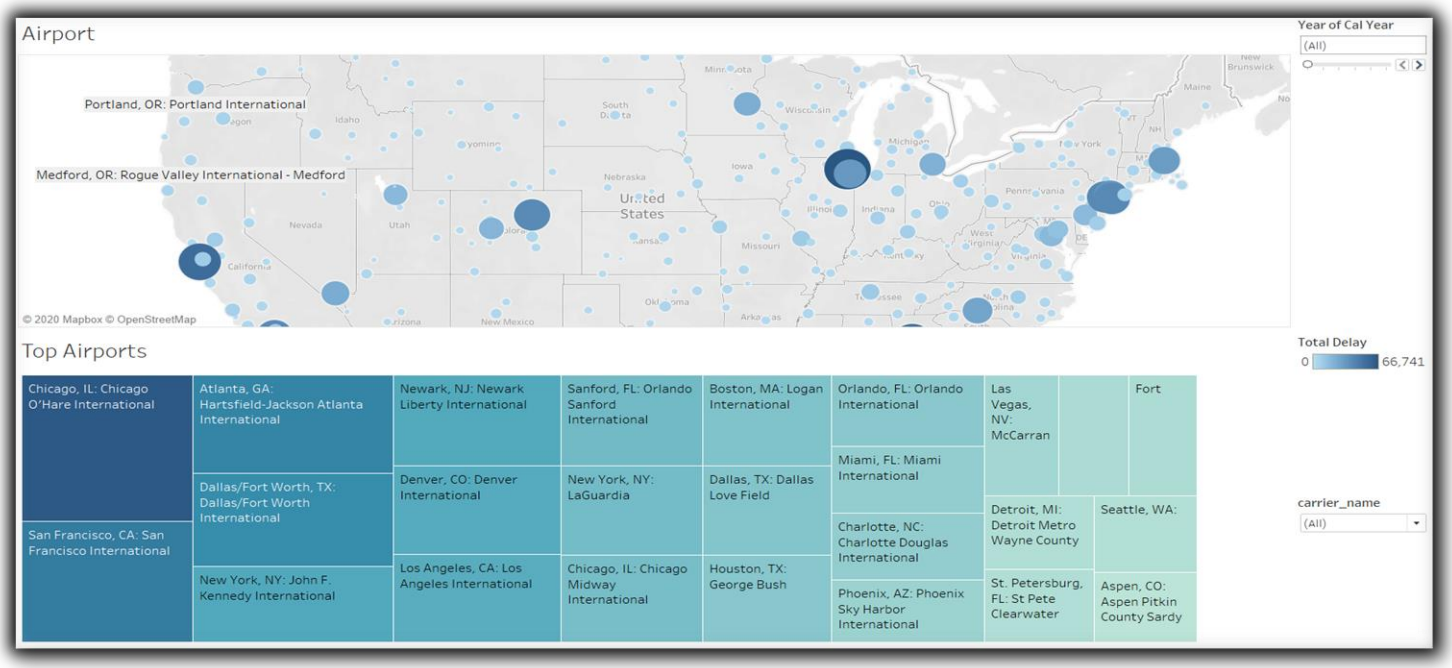
From the above visualization, we can observe that flight delays due to various reasons peaks at around June-July and also around December. We see that the security check delay contributes the least, and the arrival delay contributes the most towards the overall delay. We also understand that most flights get canceled during January, and most get delayed around the May-July period. This may be due to extreme weather or holiday seasons as we see it peaks around summer and winter.

The next visualization is to analyze data in a geolocation way, so for that, we need the airport coordinates. But in the original dataset, the longitude and latitude coordinates were not provided, but instead, there was an airport column which contained the abbreviation of airports. So we changed the datatype of this from string to geolocation and obtained the coordinates by choosing the airport option for the same. In this visualization, we find out which carriers were most delayed when taking into account all the airports and all the years. Also, to see that for a particular airline, what are all the airports that contribute towards most of its delay. To do that, I created a horizontal bar graph and mapped it to the above-mentioned map so that we could see for a year what were the carriers most delayed and what were the airports most contributing to the delay. Similar to the above visualization here, we can also filter by airport, airline and year to drill down even further.

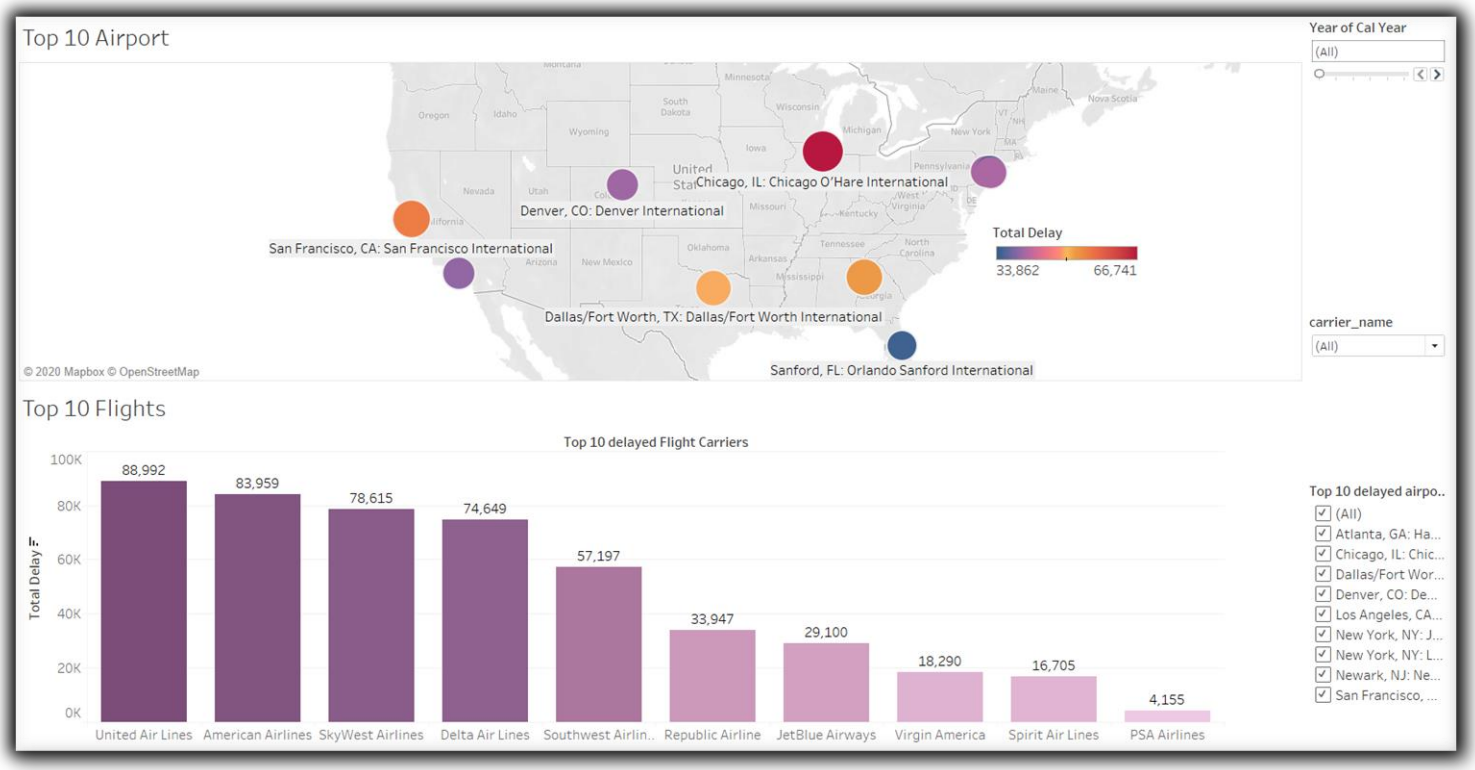


Here we can see the average delay distributed through multiple airports in the USA dependent on different carriers. We observe that when all airports all taken into account across the map, Southwest Airlines gets delayed the most, followed by American Airlines and then JetBlue Airways. When we click on Southwest Airlines in the bar plot, we get to see that Chicago, Denver, and Las Vegas airports are contributing most towards its delay. Similarly, we can see which flights were most delayed for a particular airport.

Similarly, the visualization below shows which airport contributed the most towards delay. We can observe that the airport which contributed most towards the delay in Chicago O'Hare Airport, followed by San Francisco International Airport then Hartsfield-Jackson Atlanta International Airport. We can also see for numerous carriers, which airport was responsible for most delay by selecting carrier name from the carrier legend beside the plot.



The last Visualization shows the top 10 airports and carriers that led to delay. To show that I created two sets that are 'Top 10 delayed airports' and 'Top to delayed carriers'. These two sets only contained the top 10 airports and carriers, respectively.



Overall, In this project, we have answered all the questions that stated earlier and discovered all the reason behind the delay of various flights. We also found out which airport and carriers are most responsible for flight delay for the United States in the year 2016-2018.