DSC 465-910 — Course Project PD 2

Data Visualization – 2021 Spring

Submitted to:

Terriell Scrimager

Professor

College of CDM, DePaul University

Submitted by Silent Sonics (Team 2):

Ronaldlee Ejalu

Zhenyu Liu

Adil Shedbalkar

Nai Biao Zhou

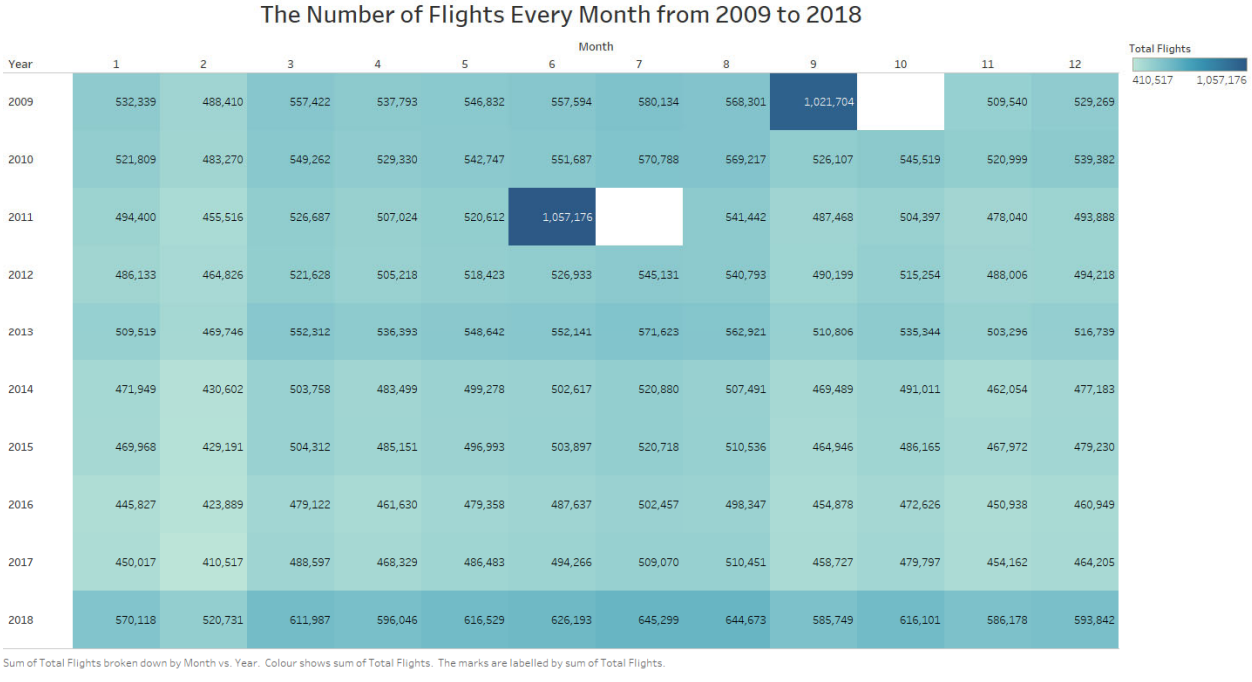
May 15, 2021

Executive Summary

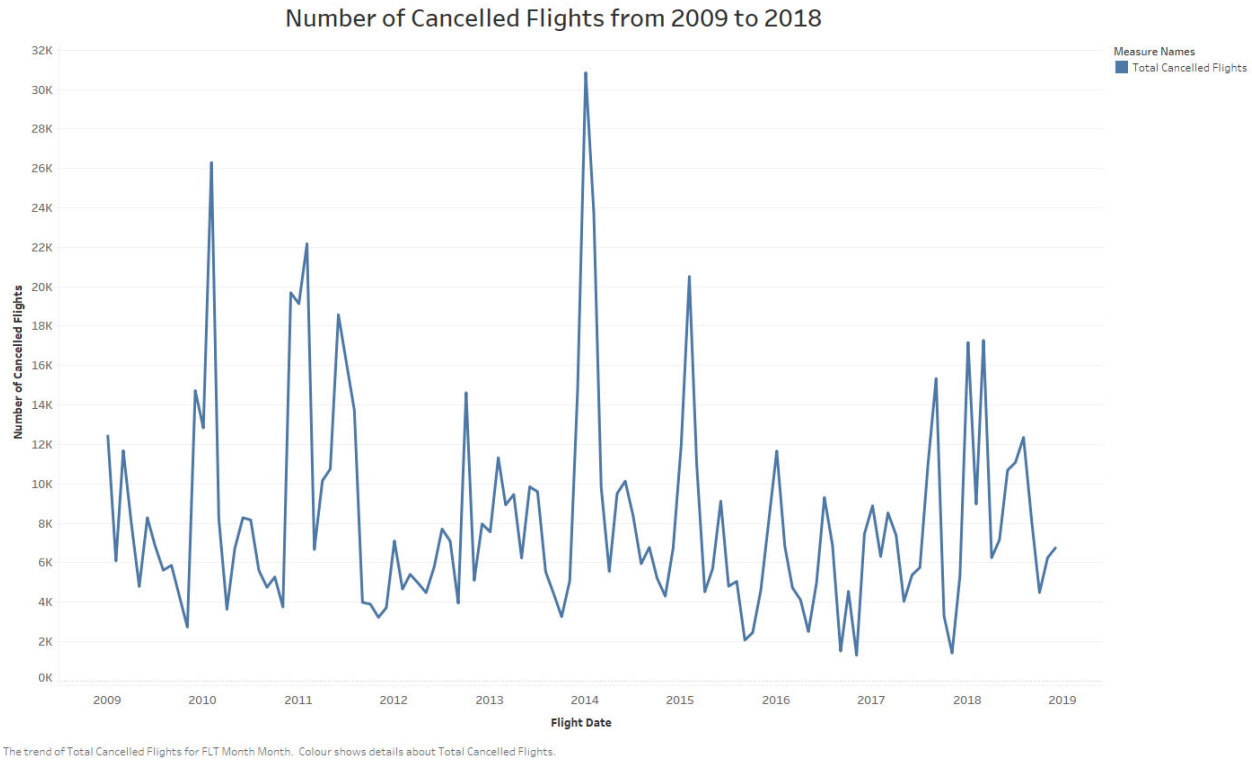
The dataset has multi-year US domestic flight data from 2009 to 2018. We explored ~~all of them~~the dataset and determined ~~to study the data in 2016~~that 2016 was the ideal candidate for this project. In this submission, we included a flight destination density map, which shows that most flight destinations ~~s come~~are located ~~from~~in ~~-~~California, the ~~e~~East ~~c~~Coast, the Midwest, ~~and southern coastlines~~Florida and the Gulf Coasts. We also employed two heatmaps to visualize flight delays by both days and months. We then performed exploratory and ~~explanation~~ explanatory analysis to demonstrate some of our ideas of ~~what the data now really looks like and how it breaks out.~~and set the groundwork for the hypothesis which we would wish to explore for this project.

Explore Flight Data Every Month from 2009 to 2018 (Nai Biao)

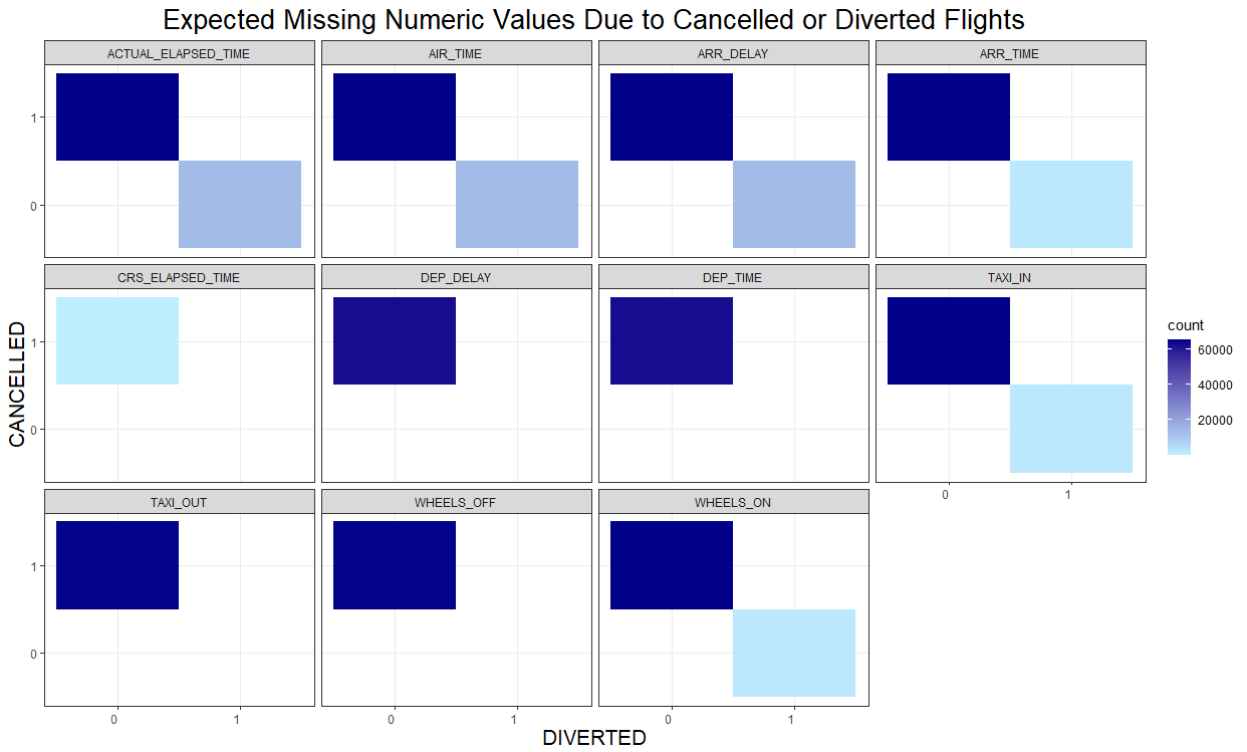
We used a highlighted table in Tableau to represent the number of flights every month from 2009 to 2018. We can observe that September in 2009 and June in 2011 seem to contain two ~~-~~months of flight data, respectively. The numbers of flights in the year 2014, 2016, and 2017 demonstrate a similar pattern, while the numbers in 2010, 2013, and 2018 show a different pattern ~~ways~~. The Fields, DEP_TIME, and ARR_TIME, in the 2016 flight data have the least missing values. Therefore, we decided to ~~study flight data in 2016.~~explore 2016 flight data for this project.



We also use the line graph to show the number of canceled flights every month from 2009 to 2010. The line graph demonstrates some degree of seasonality.

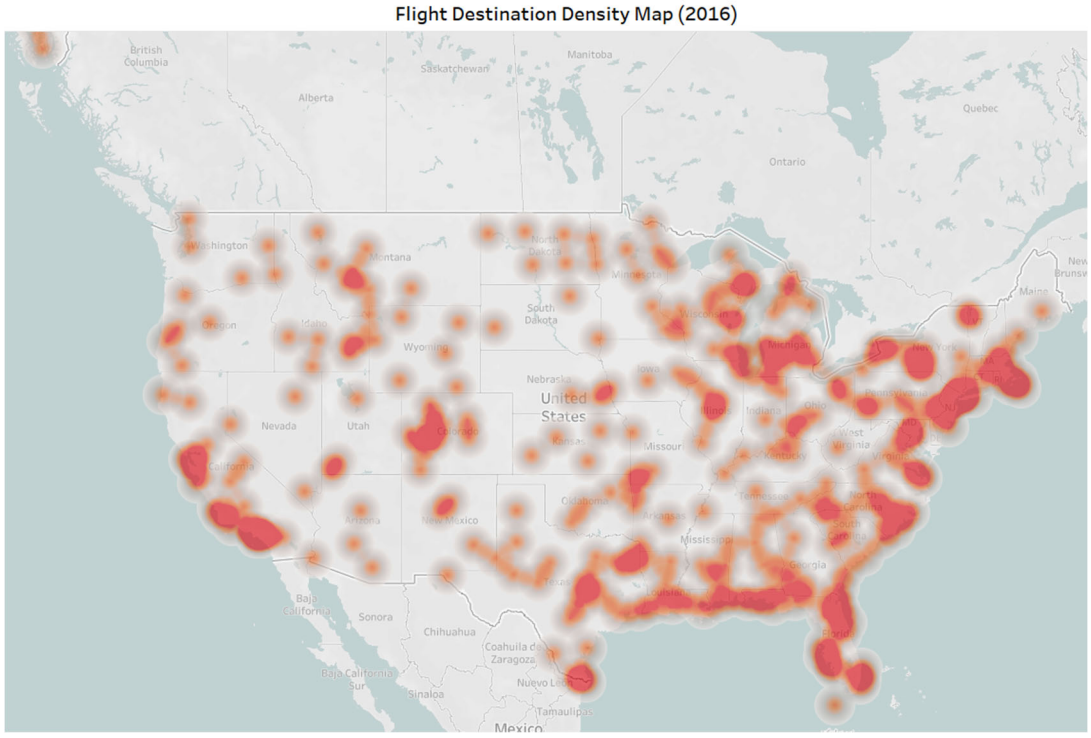


The dataset has many missing values. However, we found all of them are expected. For instance, w~~W~~hen a flight is canceled or diverted, it is reasonable that the flight does not have a corresponding ARR_TIME. The contingency plot explains these missed numeric values in more detail.

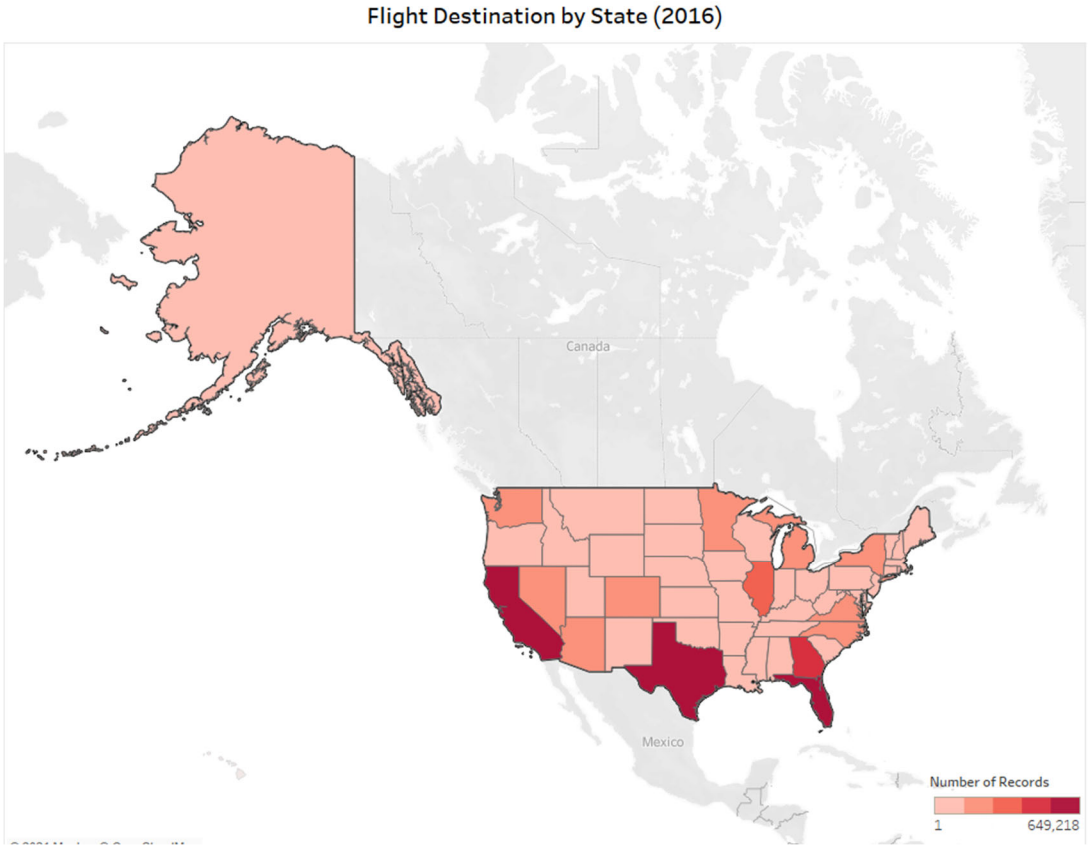


Flight Destination by State (Zhenyu)

The flight destination density map shows that most flights come from California, the Eeast cCoast, the Midwest, Florida and ~~and southern coastlines.~~the Gulf Coasts.

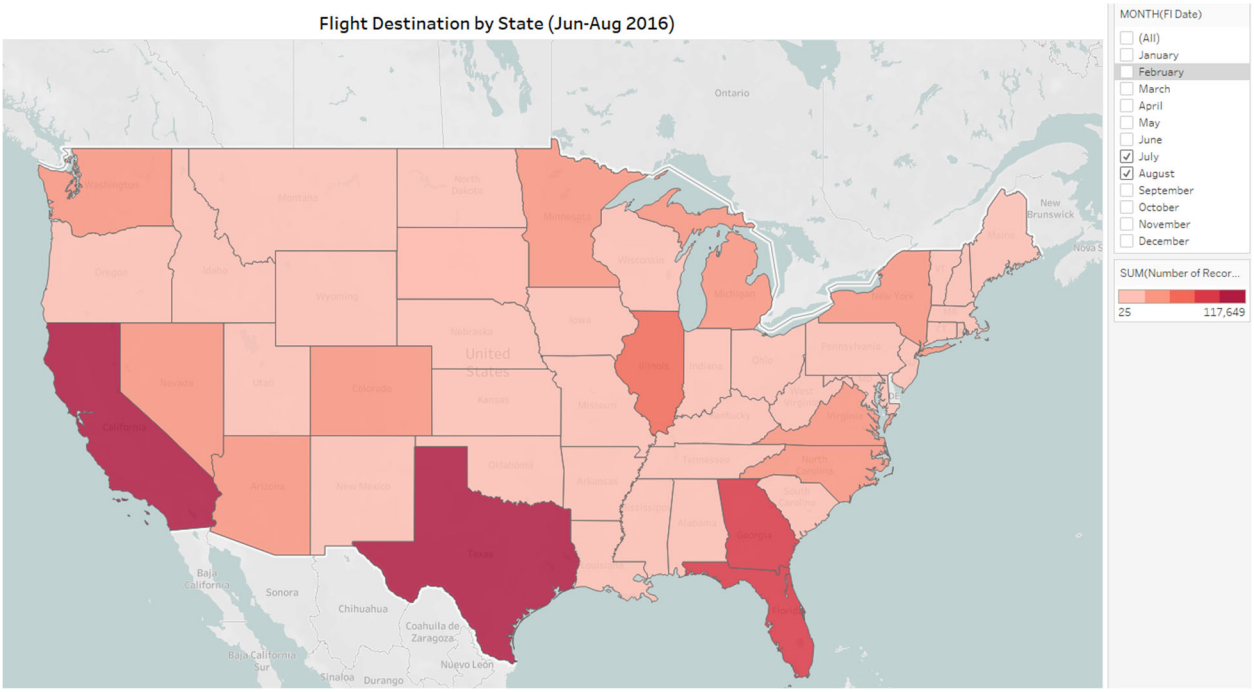


When we aggregate flights by state, we see that three of the most populous states actually had ~~the most~~the greatest number of flights for the year 2016.



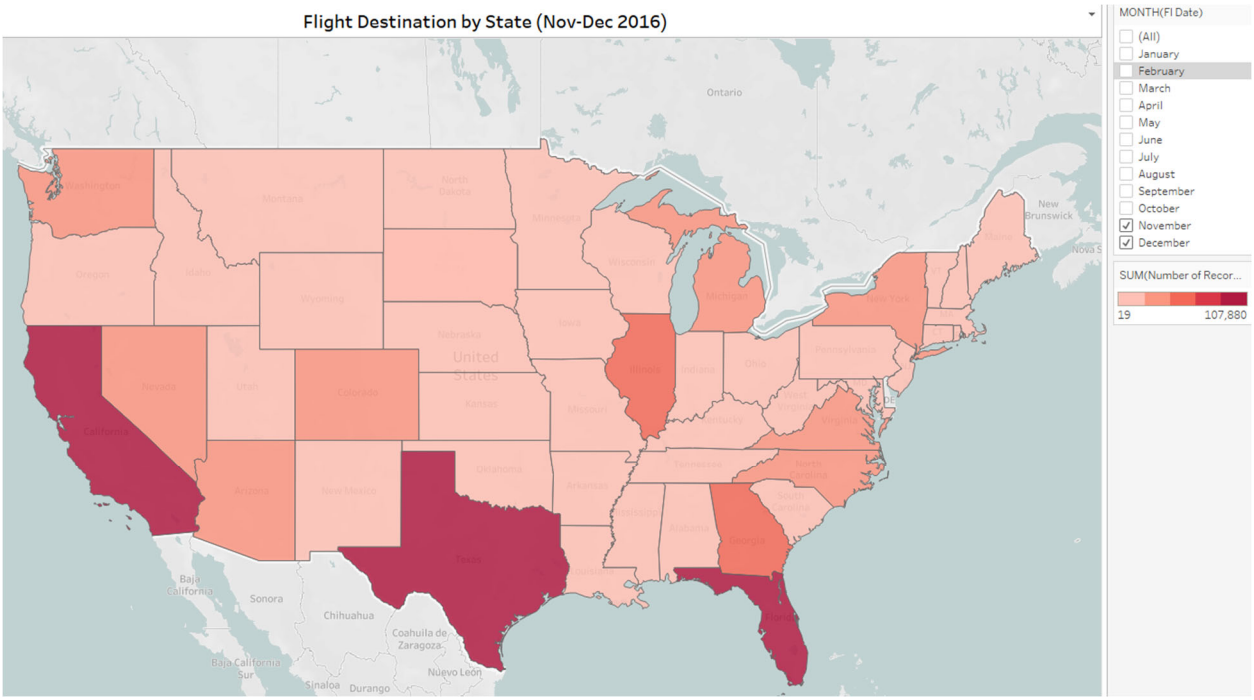
Summer Season

During the summer months (June to August), we see that there is a decline in the number of flights that have Florida as its destination.

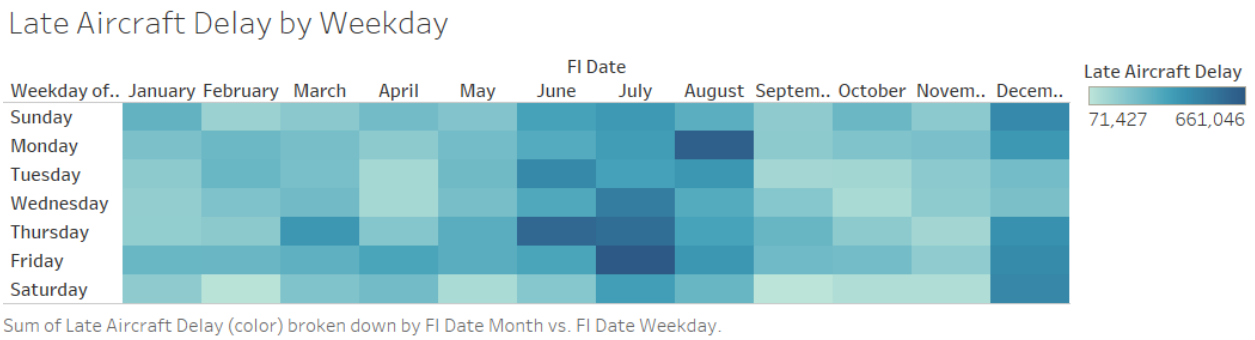


Holiday Season

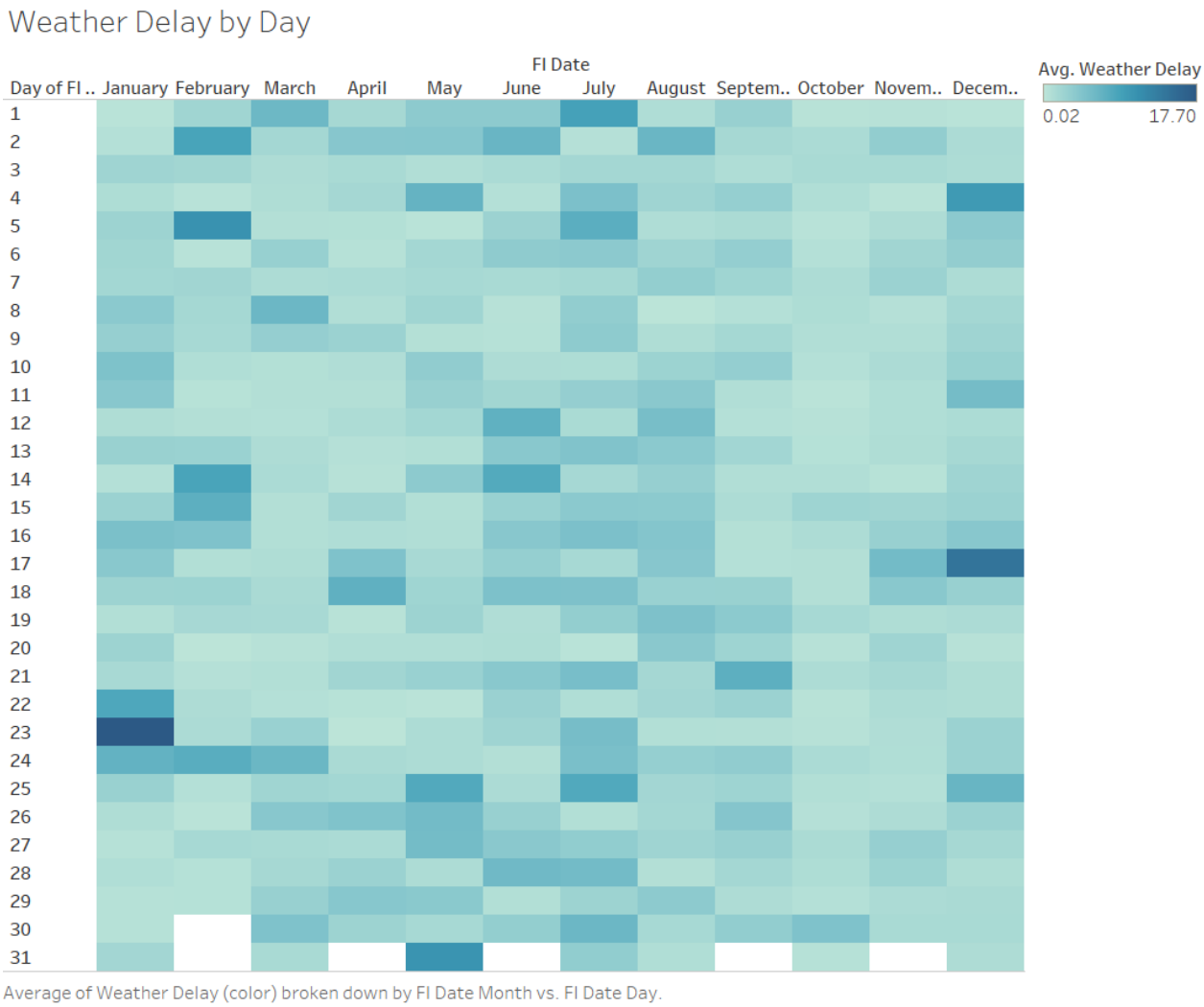
During the holiday season (November to December), there is an increase in the number of ~~people~~ flights that ~~go~~ have~~to~~ Florida as its destination.



Aircraft Delay by Days and Months (Adil)



December and January are not as high as expected compared to summer months such as June, July, and August. Is this due to the sheer quantity of flights for vacation?

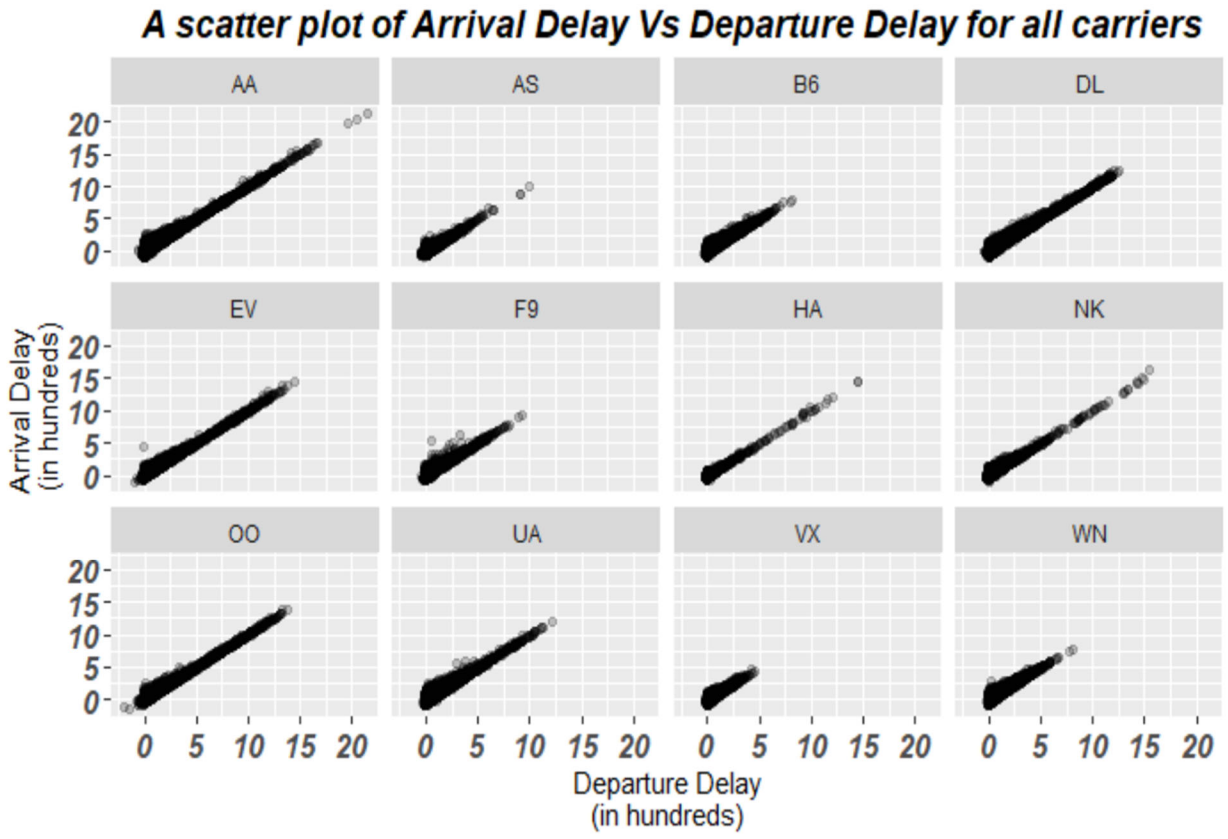


Jan 23: Huge snowstorm hits mid-Atlantic and northeastern states

Dec 17: Large snowstorm hits areas such as New York
These explain the outliers, overall, there seems to be an even spread of weather delays regardless of the month.

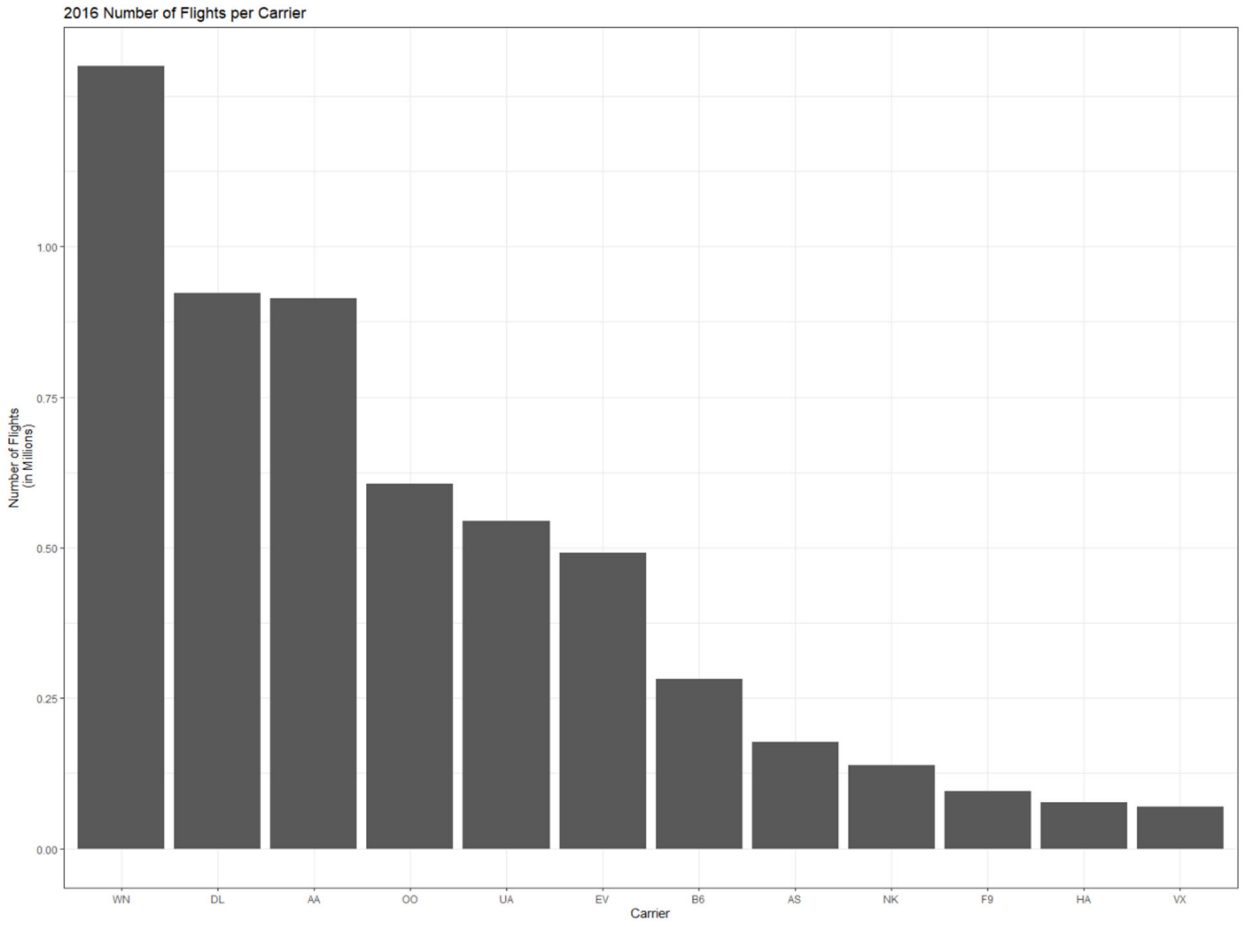
Perform Exploratory and Explanation Analysis (Ronaldlee)

(1), A scatter plot of Arrival Delay Vs Departure Delay for all Carriers

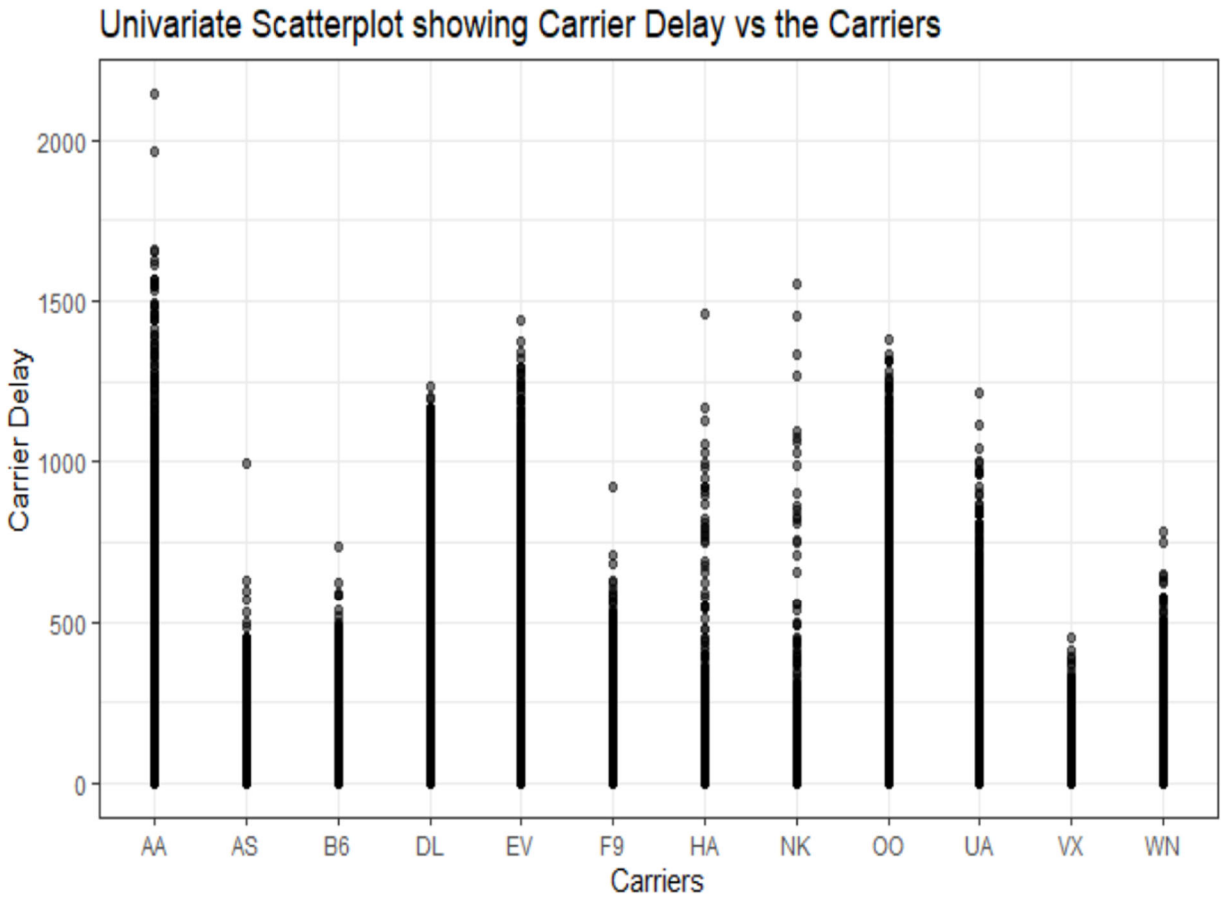


From the scatter plot above, we learned that there is a positive relationship between departure delay and arrival delay. ~~meaning~~ This is within reason since~~that~~ as departure delay increases, arrival delay tends to increase as well since the required time for a specific flight will be constant.

(2), 2016 Number of Flights per Carrier

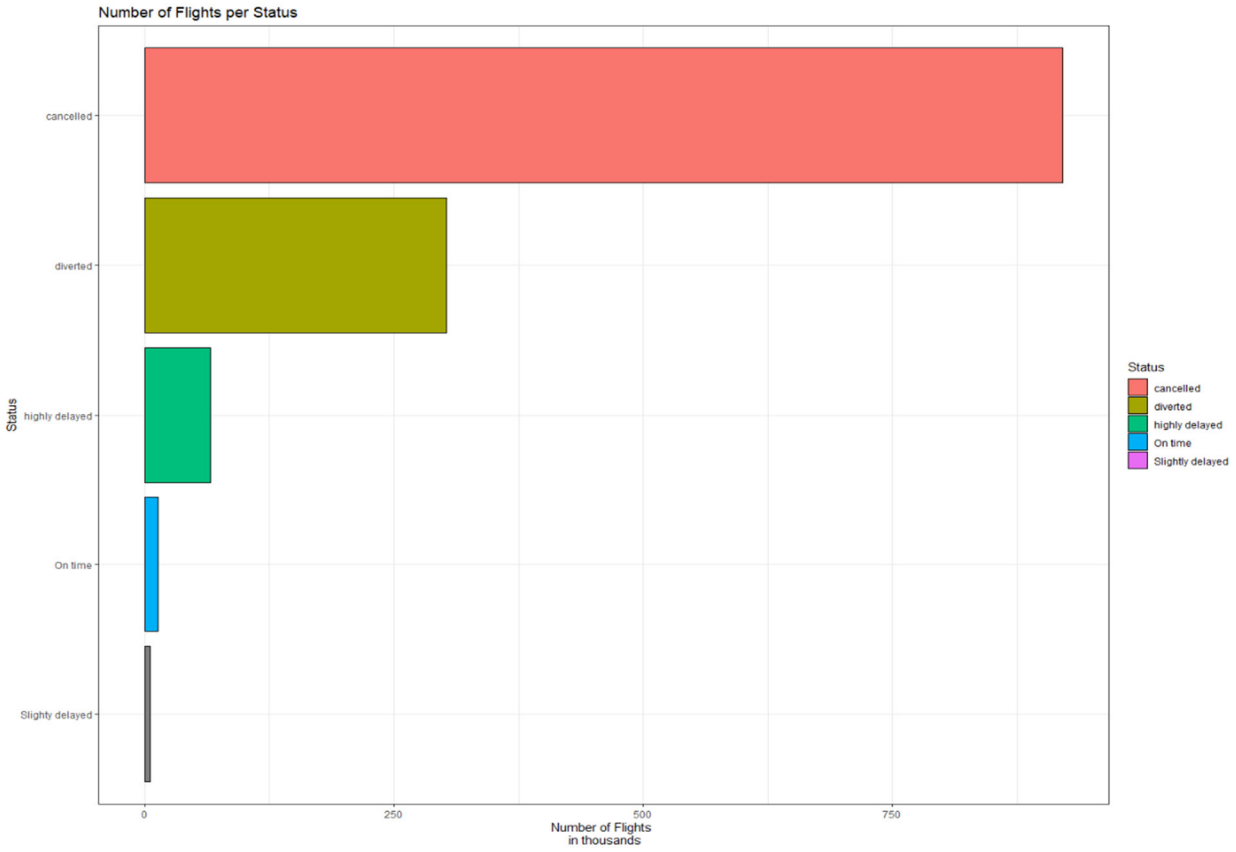


(3), Univariate Scatter plot of Carrier Delay Vs Carriers



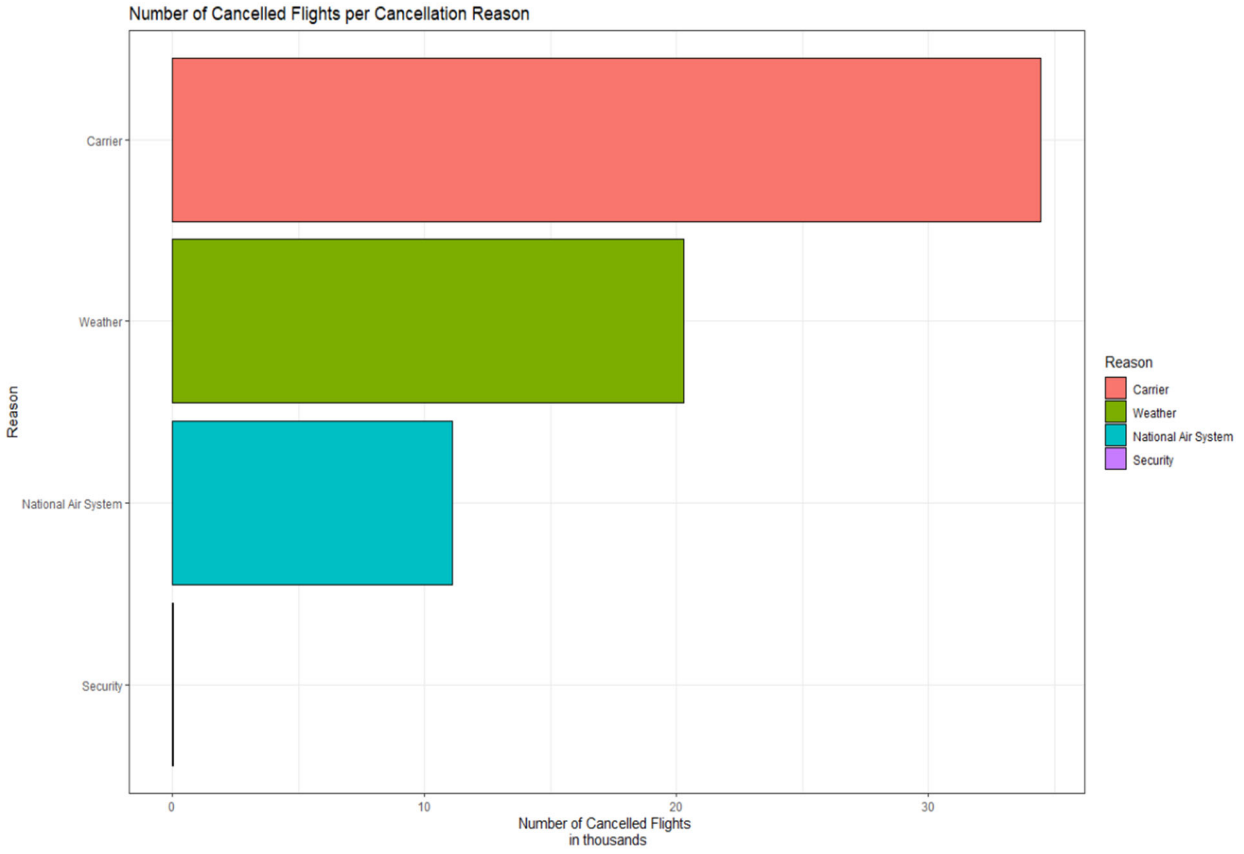
The above univariate scatter plot visualizes the distribution of Carrier delay against the different carriers that operated in 2016. From the scatter plot, we learned that American Airlines experienced ~~the most~~the greatest number of delays and Virgin America (VX) had the least amount of carrier delays.

(4), Number of Flights per Status



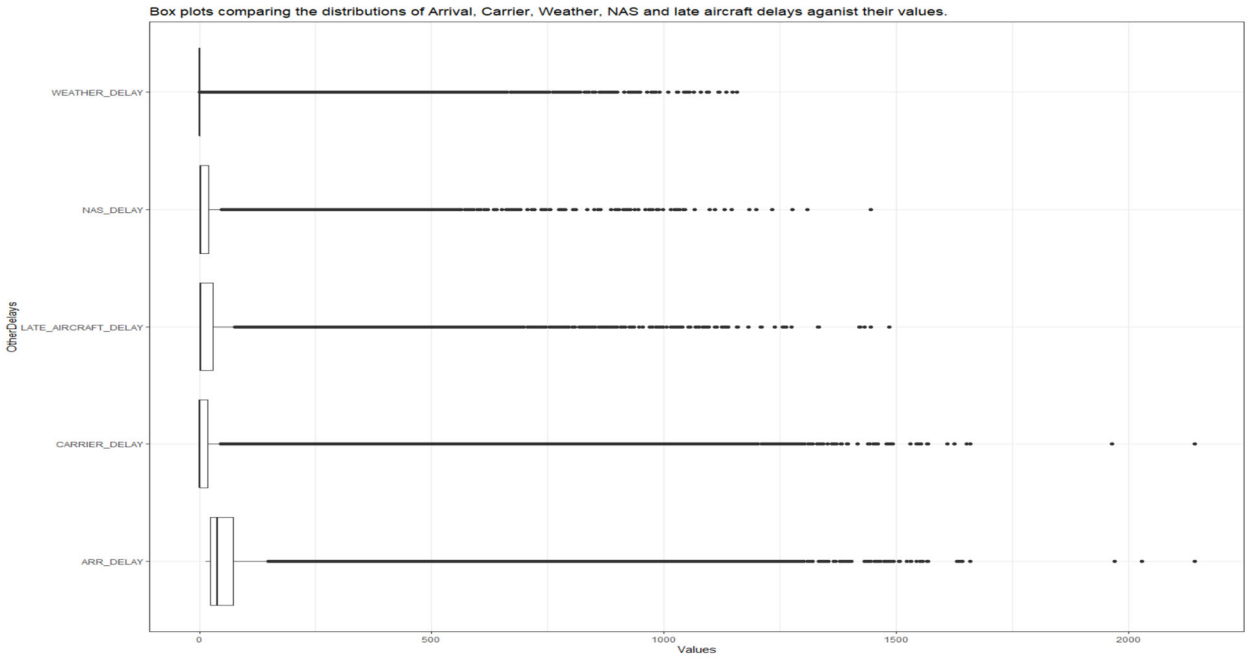
The bar graph indicates that a lot of flights were cancelled compared to those that were diverted.

(5), Number of Cancelled Flights per Cancellation Reason



Amongst the cancelled flights we learn that carrier reason dominates amongst all the other reasons especially when the planes are experiencing technical difficulties then weather comes in the next position as one of the other reasons why flights were cancelled in 2016. We see that security takes a small percentage as one of the reasons why the flights were cancelled.

Comparing the distributions of Arrival, Carrier, Weather, NAS and late aircraft delays against their values.



The above box plot shows the distribution of Arrival, Carrier, Weather, NAS and late aircraft delays against their values. We notice that they are potential outliers, which need to be investigated and removed amongst all the different delays.