**Airlines and Airports  
Punctuality Performance Analysis**

Parag Siddamsettiwar Abhishek Upadhayay

Data Science Data Science

Indiana University Indiana University

[psiddam@iu.edu](mailto:psiddam@iu.edu) [abupad@iu.edu](mailto:abupad@iu.edu)

**Abstract**

Flight data visual analysis of US air traffic to compare carriers and airports based on punctuality, delays, cancelations and diversions.

**Terms**

IATA – International Air Transport Association

BTS – Bureau of Transportation Statistics

DOT – Department of Transportation

**1 Introduction**

In today’s busy world, time is a very precious commodity. Whether you are travelling for work or leisure, mode of transportation plays a very important role in your trip’s success. Flights are the most convenient and fastest way to travel. In fact, frequent travelers prefer living in a city well connected by air transport. Airlines and airports are becoming critical necessity for every individual, as we are moving towards a more collaborative global economy.

**1.1 Motivation**

More and more people are taking flights and according to a report by IATA, around 8.2 billion people will be travelling by flights by 2037 (IATA, 2018). This trend will continue to grow due to growing economy and a need to save time by travelling using the fastest mode of Transportation.

Flight delays and cancellations are a major cost escalator for businesses, passengers and Airlines. In 2010, passengers lost as much as $16.7 billion due to schedule buffer, delayed flights, flight cancellations and missed connections (Airlines for America, 2010).

Whether you are going on a business trip or going on a vacation, planning and booking a flight that will make the entire trip as planned and pleasant is very important. The hassle to adjust the entire trip or adjust the bookings takes a toll on everyone including passengers, airlines and businesses. According to a research, an average Briton wastes 16 days of their holiday time waiting for their flight to take off (Andrew Hough, 2010).

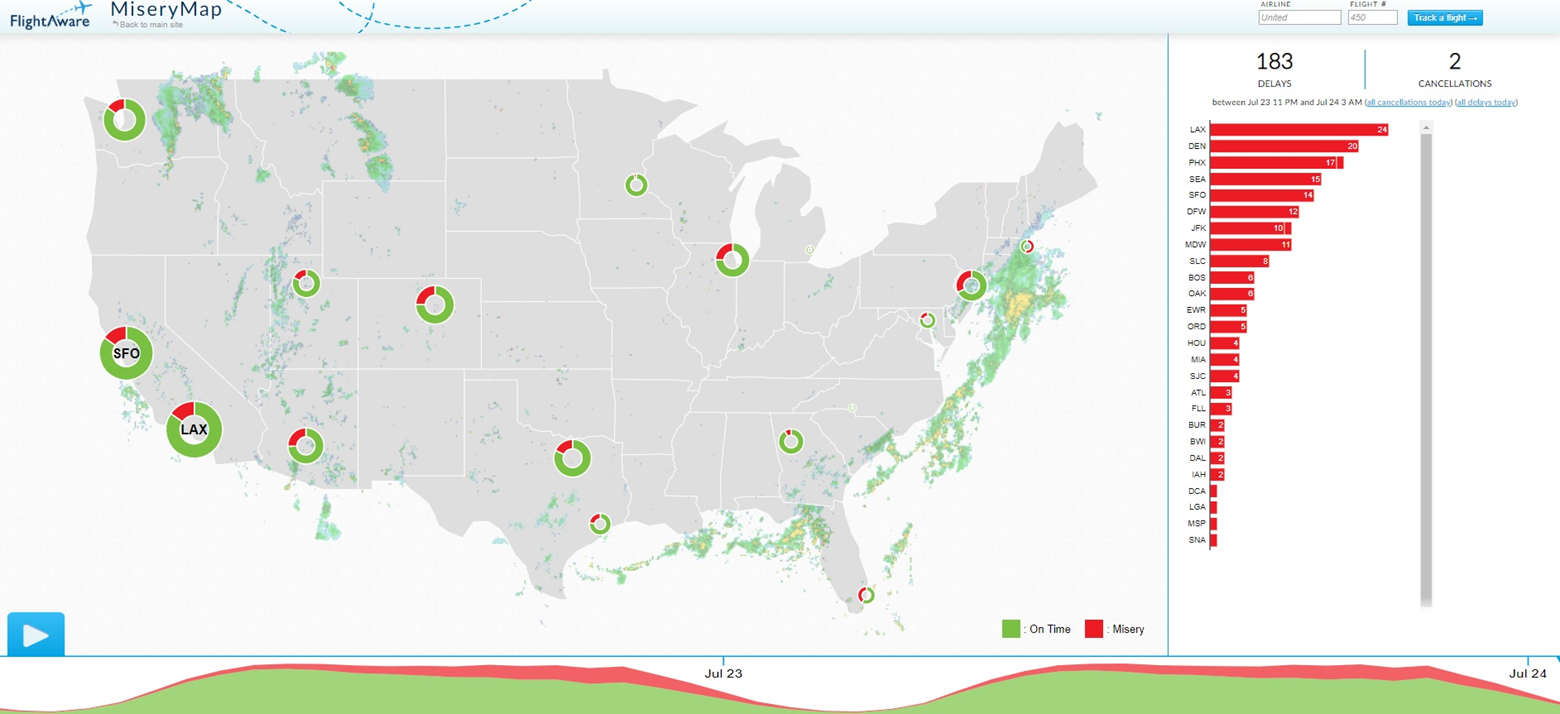
It will really help if we can analyze and visualize the flight data to get a better understanding of Airlines and Airports to make an informed choice before finalizing our travel plans

**1.2 Existing work**

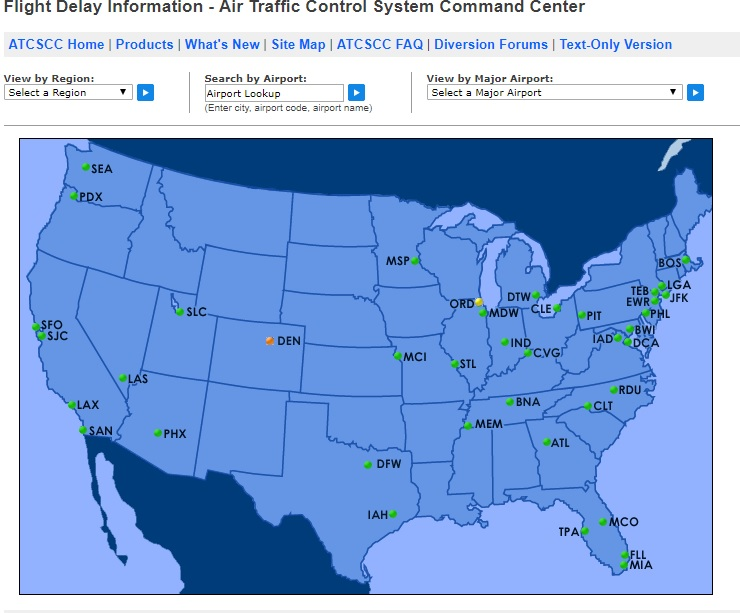
Even though there are tons of travel websites to book flights, there are only a handful websites that provide an analysis of punctuality of airlines and airports.

Even the ones that are available are either on current data (real time or past few hours) and not an analysis on historical data for a proper analysis of trends or behavior.

*FlightAware’s flight delay visualization* (FlightAware, 2019)  
This visualization shows some big Airports only, no information is available for Airlines and no option to choose other airports e.g. Indianapolis, Cincinnati etc. More importantly, this is based on current data (last few hours) and not an analysis on a historical dataset.



*Flight Delay Information - Air Traffic Control System Command Center* (FAA, 2019)  
This visualization also shows airports and it’s also based on current data (last few hours) and not an analysis on a historical dataset.  
Its interactive and lets user select airport or flights and provides current situation for that.



There is not enough detail available that can help us in making decision for future travels from airports and airlines based on past trends and airlines behavior.

Even though a flight may show on time on these current data visualizations, there is a possibility that the flights of an airline are regularly late due to unavailability of aircrafts or pilots. We may want to avoid taking that Airline

**1.3 Contribution (To be done)**

**2 Process**

Each country has an authority that controls airline traffic for them. Usually these authorities host the flight data including delay information.

For United States, U.S DOT maintains and make this data public. Thanks to Bureau of Transportation Statistics, which is part of DOT, we have found the source for a reliable, robust and continuous data of flights including delays, cancellation and diversion information.

These government agencies also regularly publish data via their open Data.gov platform

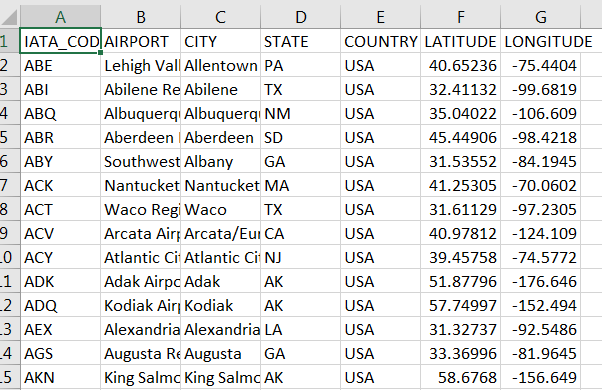
We have downloaded on time performance data of all the US domestic flights. Such airline dataset is generally associated to IATA\_CODE which is an unique code given to every airline and airport. It is used across databases for easy cross-reference.

Let us look at main data constituents:

**Airlines**



**Airports**



**On-time performance Dataset of USA domestic flights. (Bureau of Transportation Statistics, 2013)**

This dataset is regularly updated and published, it can be downloaded from following URL:

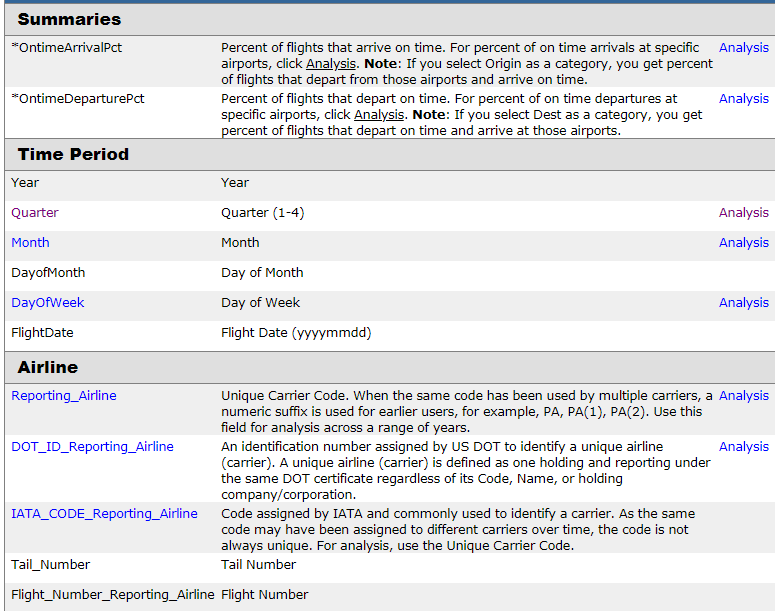
<https://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time>

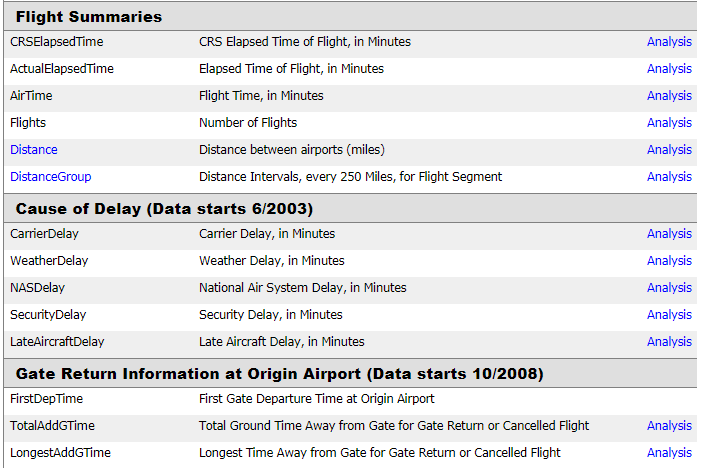
This is a large dataset with millions of records, depending on how much historical data we download.

Below are all the dataset columns downloaded by us:

Index(['YEAR', 'MONTH', 'DAY', 'DAY\_OF\_WEEK', 'AIRLINE\_CODE', 'FLIGHT\_NUMBER','TAIL\_NUMBER', 'ORIGIN\_AIRPORT', 'DESTINATION\_AIRPORT','SCHEDULED\_DEPARTURE', 'DEPARTURE\_TIME', 'DEPARTURE\_DELAY', 'TAXI\_OUT','WHEELS\_OFF', 'SCHEDULED\_TIME', 'ELAPSED\_TIME', 'AIR\_TIME', 'DISTANCE','WHEELS\_ON', 'TAXI\_IN', 'SCHEDULED\_ARRIVAL', 'ARRIVAL\_TIME','ARRIVAL\_DELAY', 'DIVERTED', 'CANCELLED', 'CANCELLATION\_REASON',’AIR\_SYSTEM\_DELAY', 'SECURITY\_DELAY', 'AIRLINE\_DELAY','LATE\_AIRCRAFT\_DELAY', 'WEATHER\_DELAY', 'DATE', 'IATA\_CODE', 'AIRLINE','AIRPORT', 'CITY', 'STATE', 'COUNTRY', 'LATITUDE', 'LONGITUDE'],dtype='object')

Additional details of these columns can be found on the TranStats website (Bureau of Transportation Statistics, 2013), e.g. below screenshots:





**3 Process and Results**

After carefully studying the data columns available to us, we decided to download the ones that will be useful in our analysis.

We downloaded 10 years of data and tried to load it into python for exploratory analysis. Obviously this data was very large and we have limited memory available to us on our local computers, we noticed extremely slow responsive behavior of python program. It was taking a lot of time to load data and process it.

Our main focus was to creat visualization and not how much data we can bring in, so we decided to restrict our dataset to 1 year. After using only 1 year of data, we were able to load and process it successfully.

Winter months generally have lots of disruption in air traffic due to unusual weather, so we decided to focus on the month of January.

**Here are the various visualizations we developed in the below sequence.**

***Top 20 delayed flights***

We wanted to begin by getting a list of flights with highest delays.

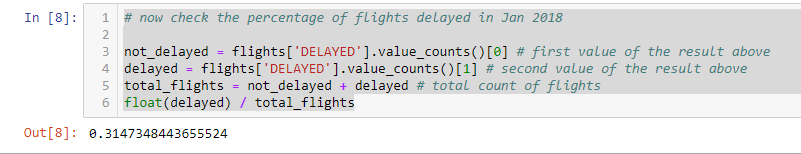
ARR\_DELAY column shows delay in minutes. Applying SORT\_VALUES method on this column gave the output we wanted. For display purpose, we restricted the results to top 20 delayed flights.



The most delayed flight was of American Airlines, which was more than 2000 minutes delayed, almost 1.5 days.

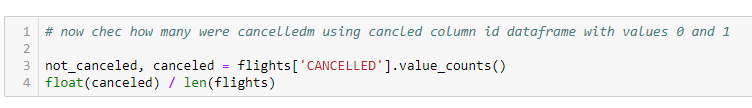
That made us wonder how many overall flights were delayed in January month irrespective of carrier.

So, we did some calculations:



More than 31 % of flights were delayed in January of 2018. That is a big chunk of flights being delayed.

So how about cancellations?

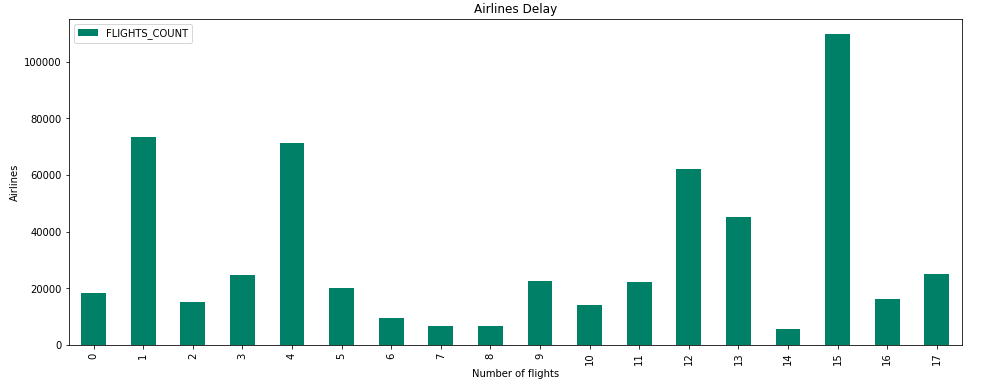


Flight cancellation in January is not as scary as delays. Around 3 % of flights were cancelled.

***On time vs delayed ratio of flights per airline***

Next, we were interested in understanding number of flights operated by each carrier in January. But we wanted to compare total number of flights for each carrier by ‘delayed’ and ‘on-time’.

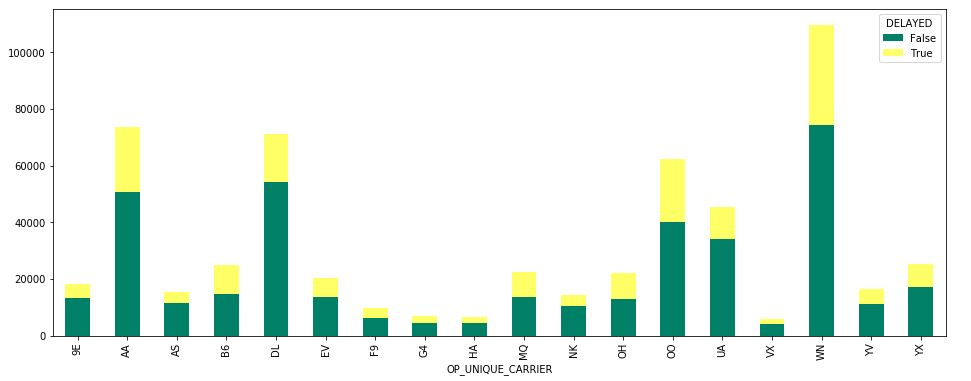
So, we decided to create a bar chart.



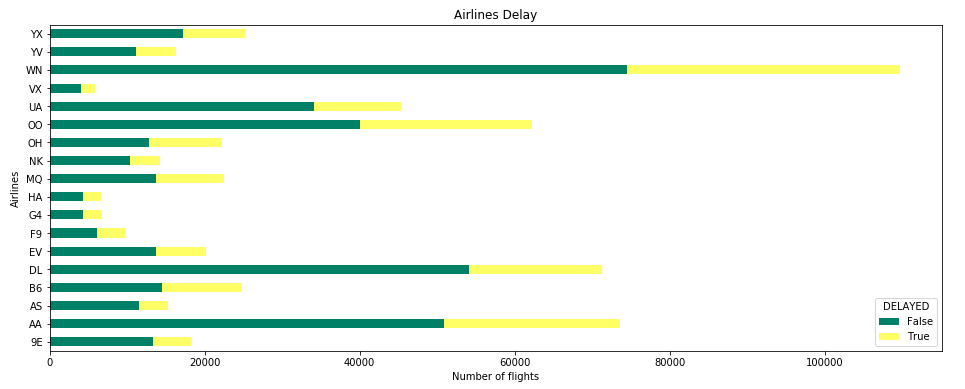
As we can see, above chart is not showing airline codes and its uninformative.

We had discussion and decided to come up with stacked bar chart for this.

The first version we created looked like this.

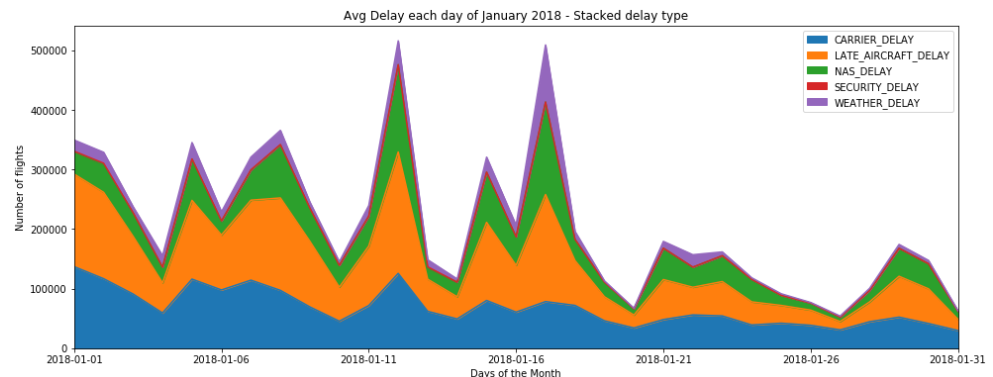


After looking at the chart, we realized rotating this to horizontal bar chart will make it better. Also we can read the carrier names horizontally instead of sideways.

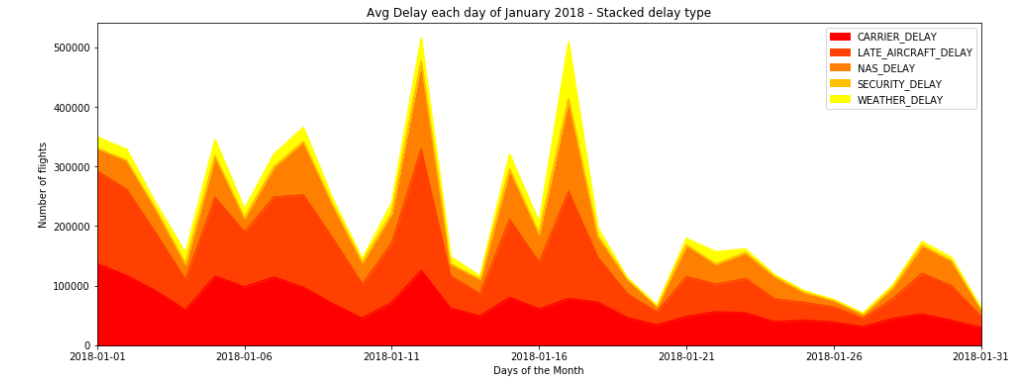


***Flight delay reason heat map***

It was time to dig deeper into reasons of delays. From the data set columns we figured out that there are 5 types of delays. Carrier delay, weather delay, late aircraft delay, NAS(National Air System) delay and Security delay. We wanted to compare different types of delays to better understand which delay type was more significant for the Month of January. Heat map to the rescue. We created the chart to do just that and in our first attempt, we were able to come up a wonderful visualization.



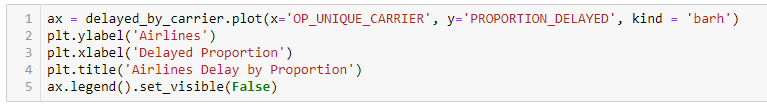
It would make it more appealing if we can somehow color code it differently. So we decided on using sequential color map. ‘autumn’ color map was yielding best results so we finalized on that for this chart. We can clearly see that the main reasons for delay are Carrier delay and Late Aircraft delay.

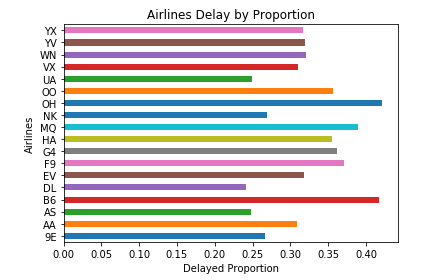


***Airline carrier delays proportion***

With motivation to further analyze the delay, we decided to come up with proportion of delays by comparing it against all the available airline carriers. For this purpose we created new column and used it to visualize the proportion of delayed flights across airlines.



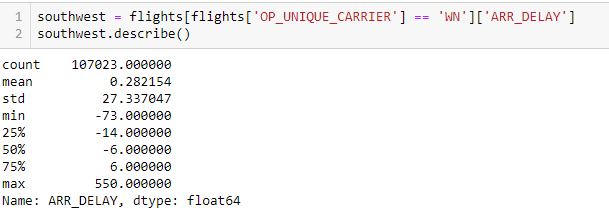




Clearly the ones that looked like had more delays are actually better when compared with proportions. Like Southwest (WIN) and Delta(DL)

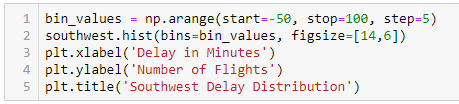
Let us pick southwest for some detailed analysis, as it is our most preferred airline.

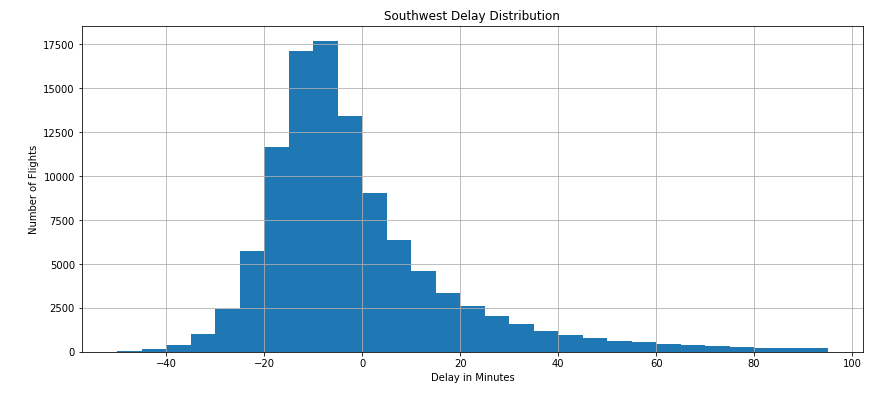
We wanted to understand the distribution of Southwest airline delay so we started with basic describe function



Average delay is just 0.28 minutes, while longest was 550 minutes. Interesting insights here is that IQR (Inter quartile range) for Southwest airlines delay is actually in minus. That means most of the time southwest airlines arrives before scheduled arrival time. Let us verify this statement by histogram.

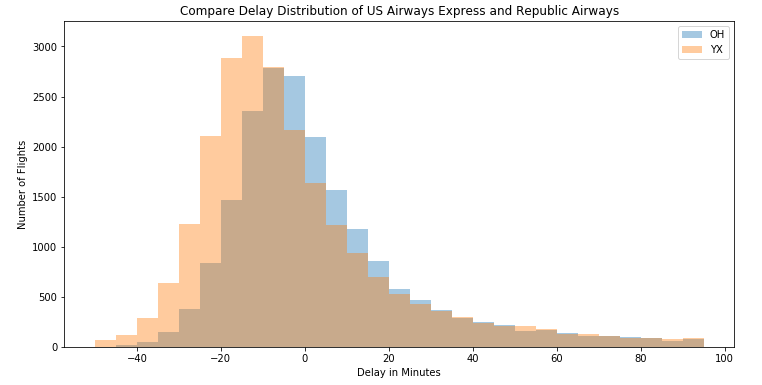
Since some delays are in negative, we need to make some adjustments for bin sizes and values. From above we can see that most of the data should be captured between -50 and 100 minutes delay, so we use that for bin values.





Looks like southwest is mostly before time or on time even during January's stormy winters.

Now let us compare two airlines with almost similar number of total flights. For example, US Airways Express 'OH' and Republic Airways 'YX'

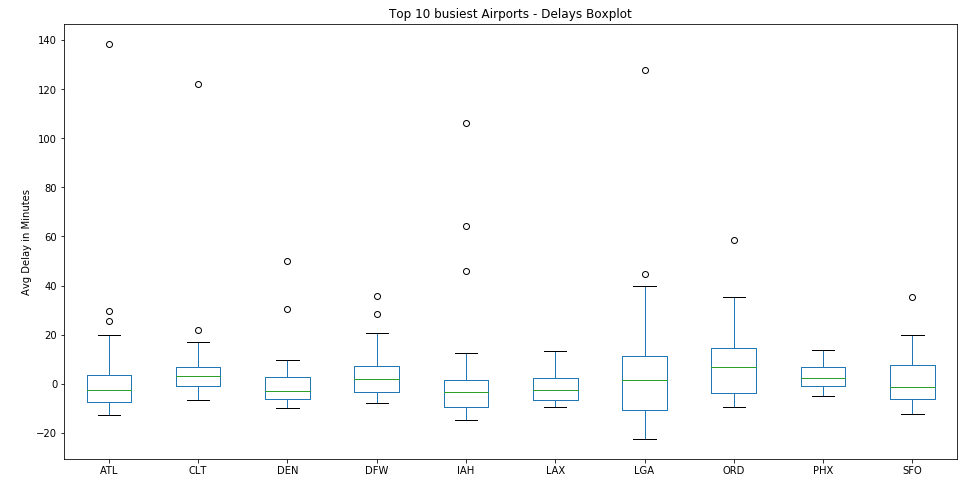


The darker area is where the distribution overlaps. The distribution looks similar but they are not the same as OH has more shifting to the left showing more delays, as it is evident from the earlier analysis as well.

***Top 10 Busiest airports and their delays***

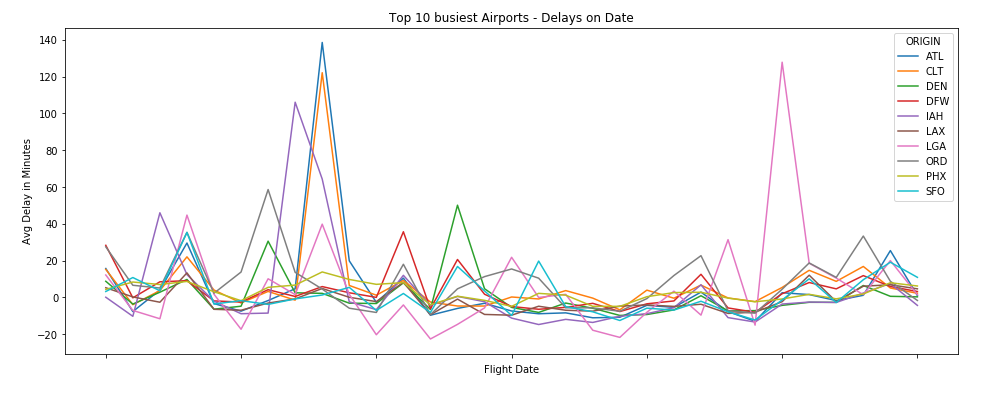
Now let us do some analysis from the airports perspective. Let us find out top busiest 10 airports and analyze where not go for vacation in January (based on delays etc.)

For such comparison we can use box plots very effectively as it can help us compare airports side by side.



We can see clear dispersion in data at Laguardia NYC and Ohare' Chicago. While other airports delays are evenly distributed around 0. Based on 2018 data, we can say that try to avoid these two airports during January.

Let us try to visualize this using line chart instead of boxplots.

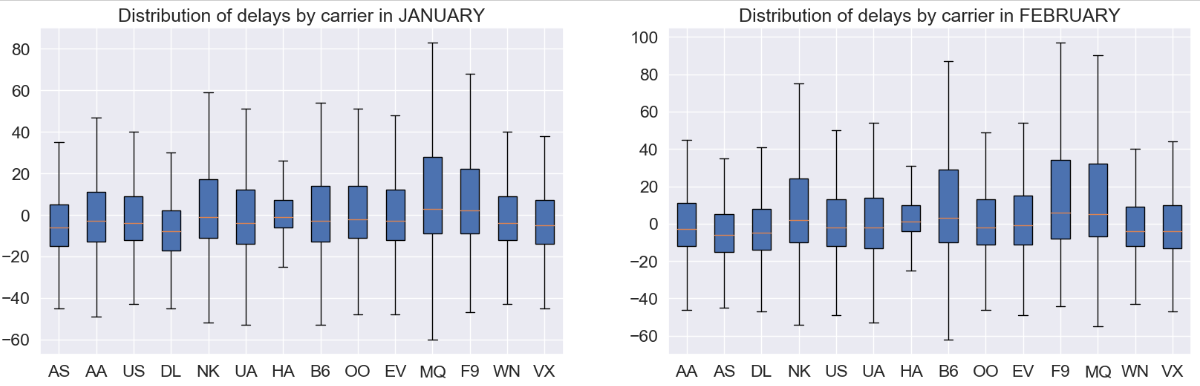


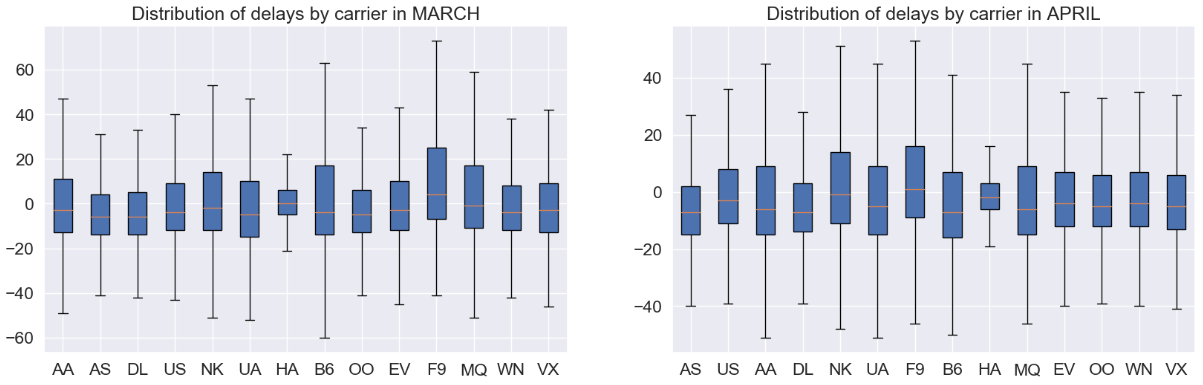
Clearly, some days have spikes for most of the carriers suggesting weather related delays on those dates. Suggesting delays at one large airport has a chain reaction effect on other airport delays.

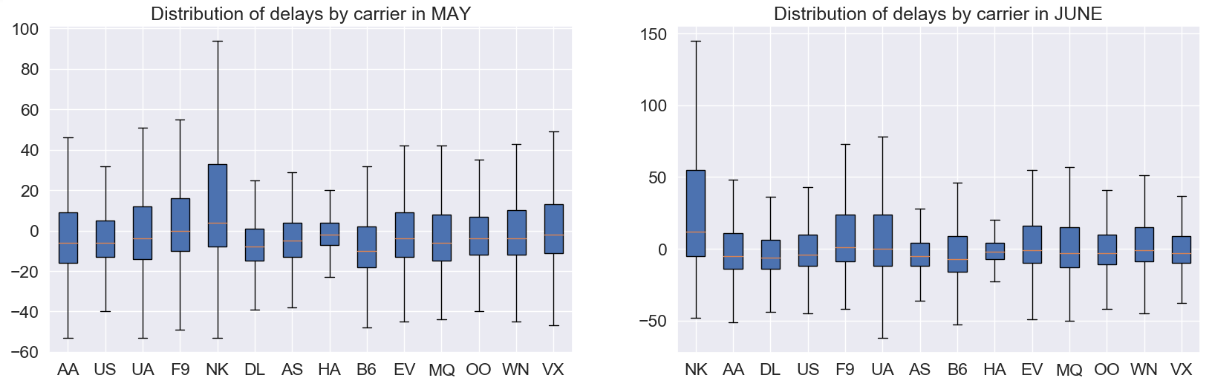
***Distribution of delays by month***

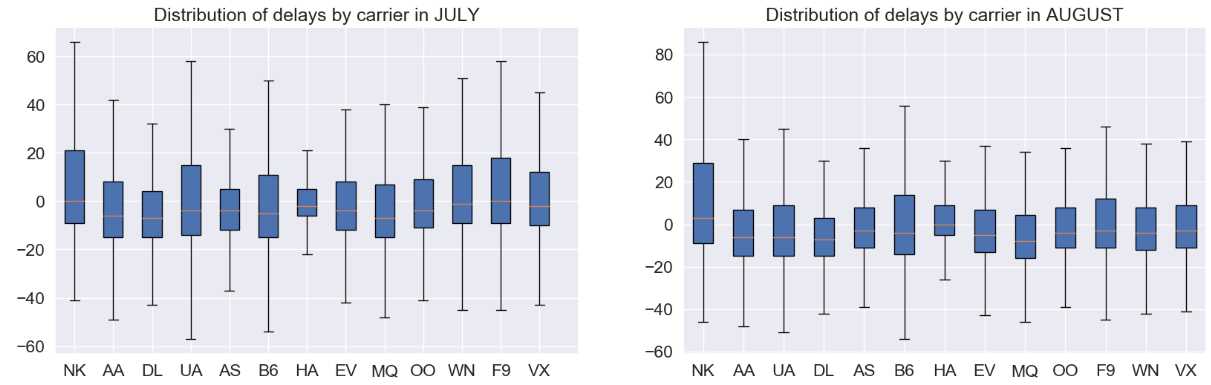
It is time we look at complete year data and try to find some interesting insights into flight data and delays.

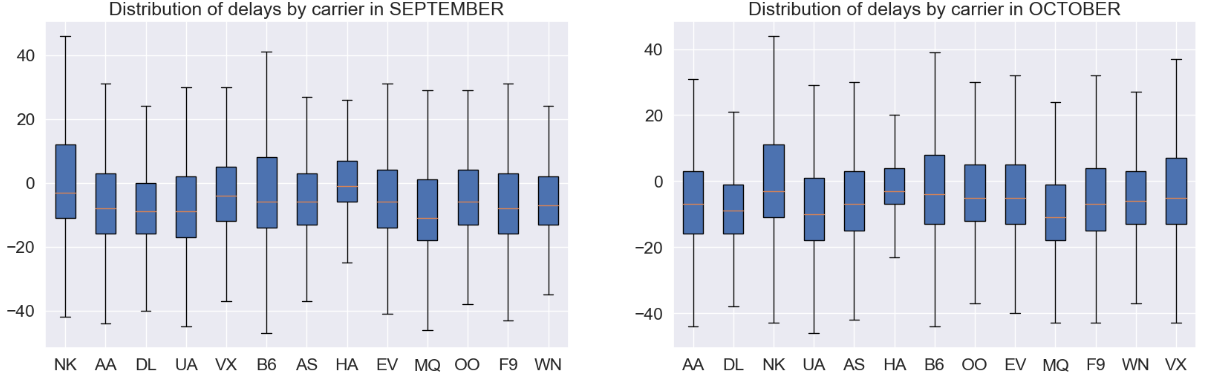
As we are looking at all 12 months data, first let us compare all the airlines and their delays in each month. We used sub-plotting to achieve this.

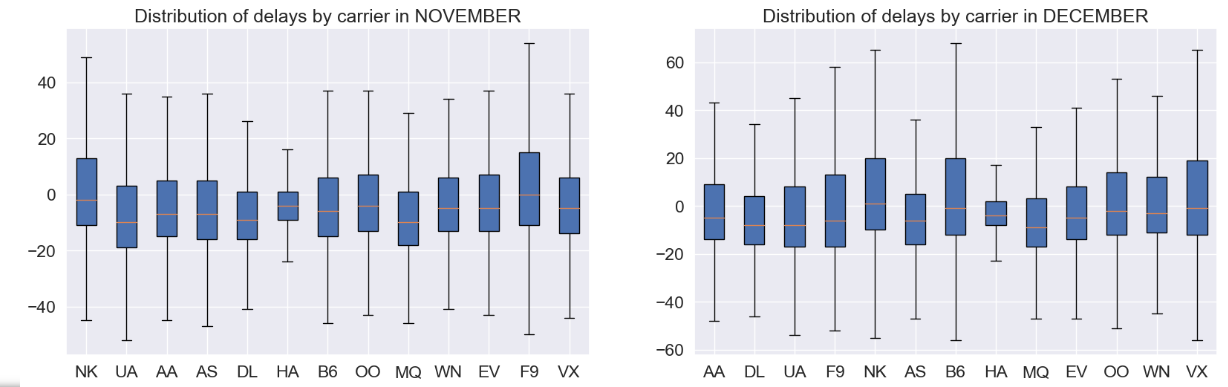










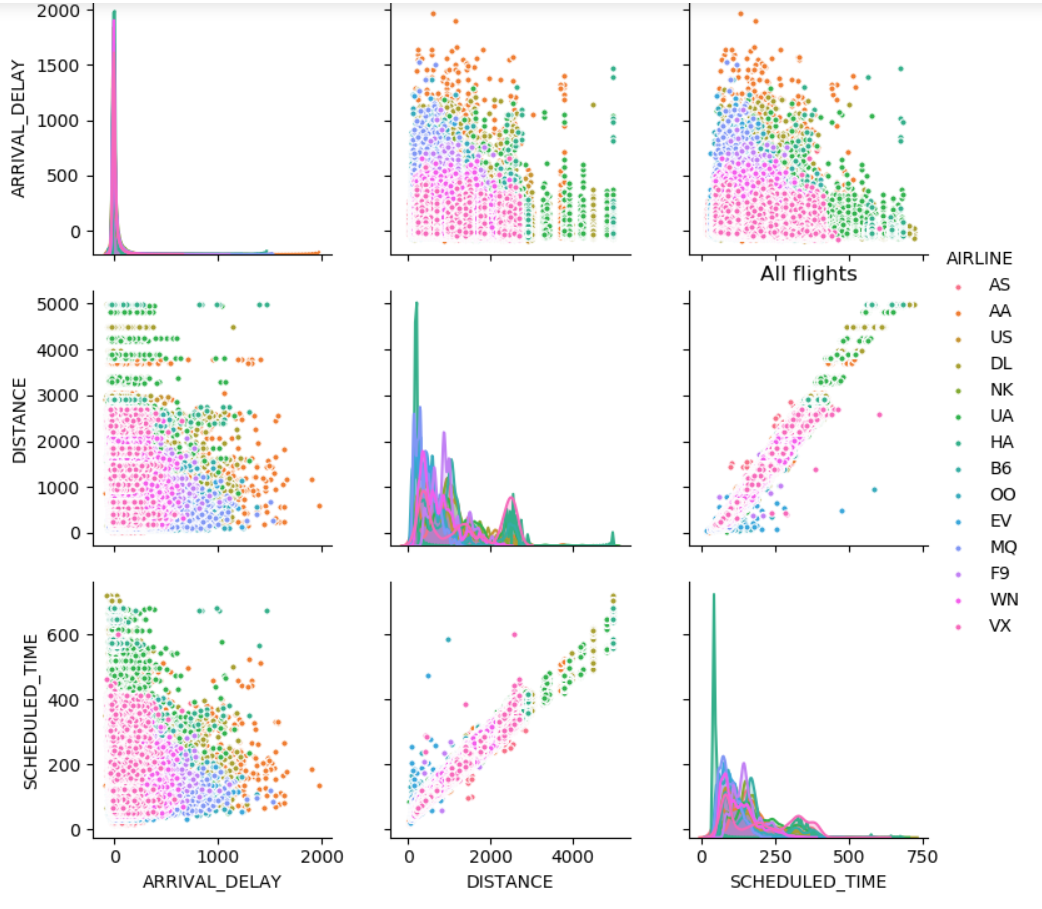


There are some insights that we can figure out by studying each month and comparing each airline carrier but it’s not user friendly. It will take a significant amount of time for any user to derive any insights from these boxplot subplots.

***Comparing airline, arrival delay, distance and scheduled time using pair-plot***

Let us try pair-plot and see if that can give us any useful and easy to perceive insights.





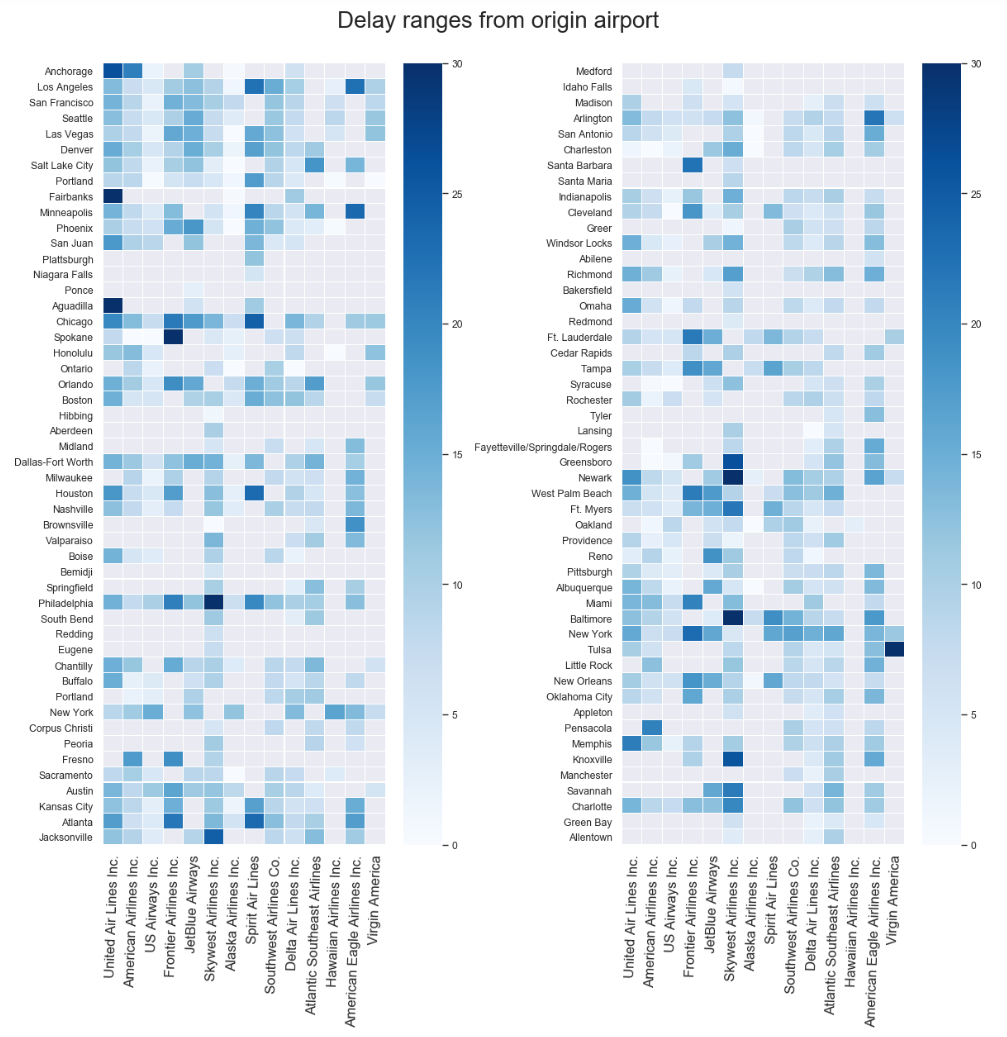
***Delay ranges from airports***

Again, we can come up with some sights like maximum number of flights were operated by Southwest just by observing pink dots but the color coding gets confusing and we have too much data being displayed on this chart which is making it very difficult to read. We needed something to the point and descriptive visualization to provide the data on delays at each airport probably divided by airlines. This will help passengers make decision on choosing airline very easily.

What we really need is some kind of heat map that can show us delays at each airport that too for each airline. This will be very helpful. After researching on internet we found some interesting heat maps like the one explained in this pager (Group56, 2013) . This heat map is great if apart from identifying the airport, it does no help in actually comparing airline carriers at each airport.

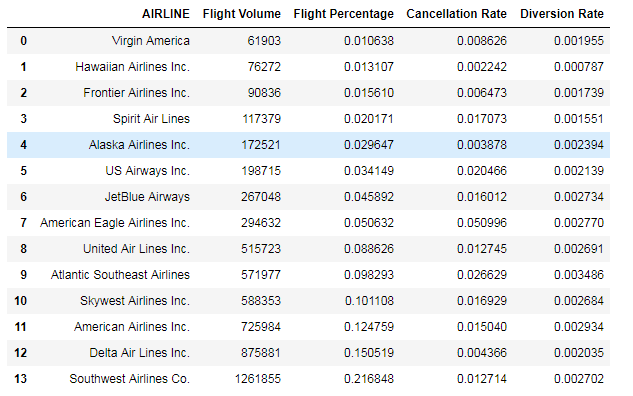
(quarbby's, 2015) Visualization actually shows each airline against airports and gives some insights but it is very difficult to read.

We started designing our own heat map by getting inspiration from (TutorialGateway, 2019) tableau heat map. At the end, we came up with this version. In this heat map, airports are plotted on y line and x line shows airlines. As the shade of blue, get darker, delay increases. Shade range and delay in minutes is listed to the right.



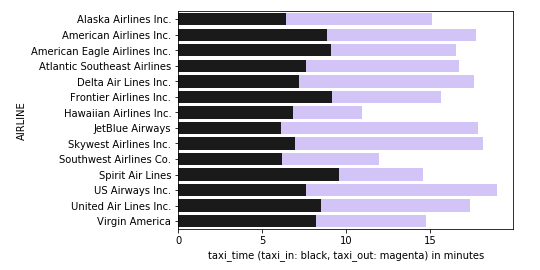
***Calculating cancellation, diversion rate and flight percentage for each airline carrier.***

We did some calculation on the data to derive cancellation and diversion rates and calculated percentage of flights for each airline across all the flights from the entire year of data.



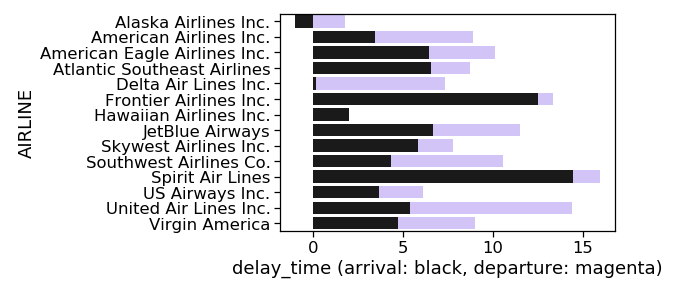
***Taxi-in vs taxi-out in stacked bar chart for every airline***

Another insight we wanted to highlight was taxi in vs taxi out by airline.



It is clearly visible that US airways has largest taxi-in and taxi-out time amongst all the airlines. Hawaiian Airline flights spends least time on taxi-in and taxi-out.

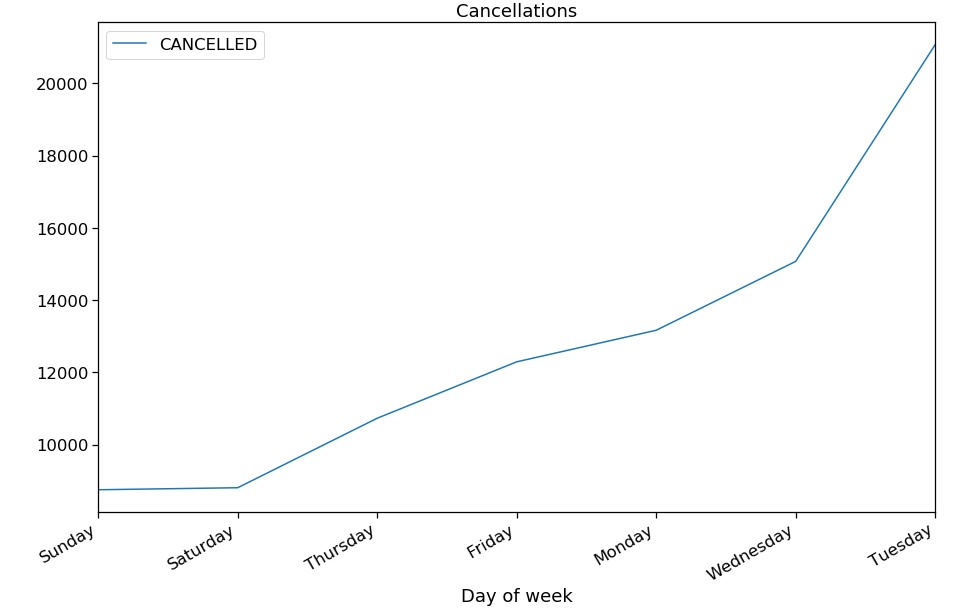
***Arrival vs departure delay for every airline***



Very quick insights we can derive from this visualization is that Delta airlines almost has no arrival delay time whereas Alaska airlines has negative arrival delay. That means Alaska airlines always reaches prior to arrival time. Spirit airlines performs worst when it comes to these delays having the highest arrival delay.

***Week by cancellation***

We wanted to see if we could come up with any interesting insights by looking at data of each day of week for the entire day irrespective of the month. We found that cancellation is highest on Tuesdays and lowest on Sunday & Saturday. This is mostly due to higher number of passengers travelling during weekends. However, this assumption of higher passenger travelling during weekends, we will have to download passenger data set from transportation website and then analyze that. This is something we will probably do in future.



***Choropleth state wise map visualization for cancelled flights***

Is there a way we can visualize cancellation, diversion and delay data by every state of USA? We have airport information like city, state, latitude and longitude available in our data-set. So let’s try to use the number of cancellation of flights for all the airports for a particular state and try to visualize that using map. We attempted it by referring the visualization shown on (Altair, 2019). The biggest problem we ran into while trying this is the state coding done in the vega\_dataset

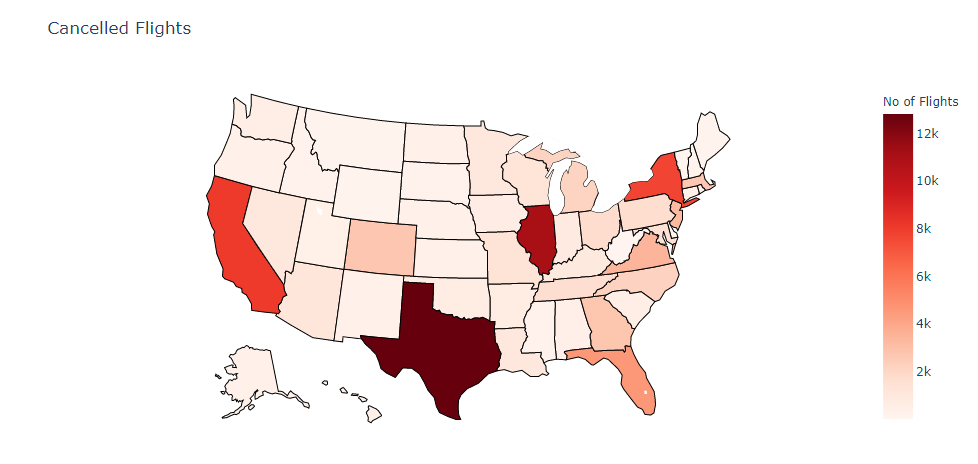
**import** **altair** **as** **alt**

**from** **vega\_datasets** **import** data

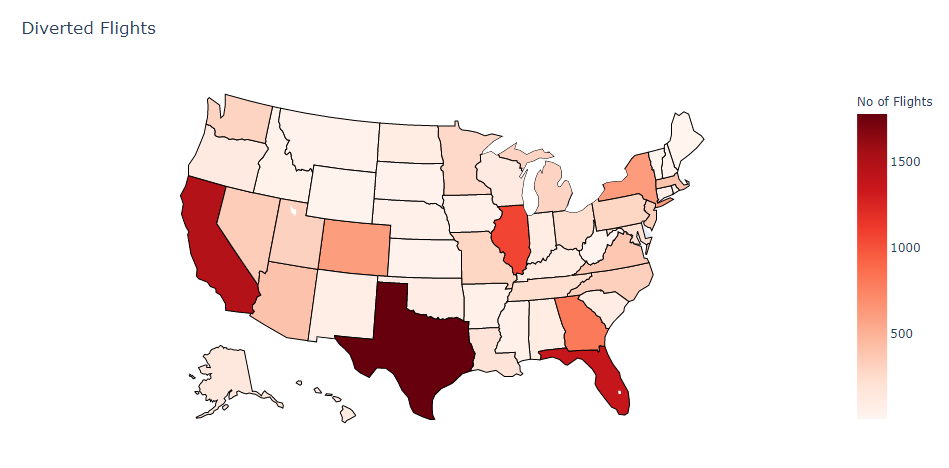
airports = data.airports.url

states = alt.topo\_feature(data.us\_10m.url, feature='states')

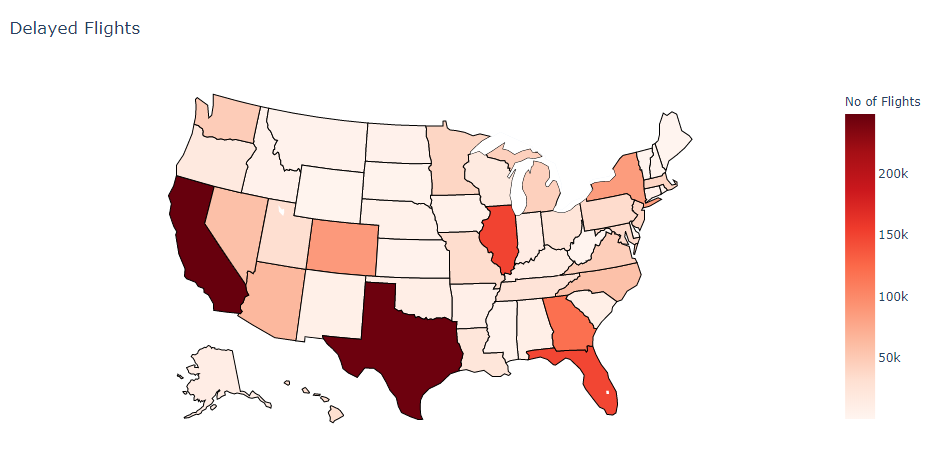
We were not able to map the state ID returning from states dataset and match them with state codes like TX, IN, CA that is present in our data set. We tried for several days but we were not successful. Finally, we dumped idea of using altair for this purpose and started looking into other python library which can help do this fairly quickly. We were able to plot Choropleth by referring to (Plotly, n.d.)



***Choropleth state wise map visualization for diverted flights***



***Choropleth state wise map visualization for delayed flights***



**4 Conclusion**

**5 Future Work**

# **6 References**

Airlines for America. (2010). *Annual U.S. Impact of Flight Delays (NEXTOR report)*. Retrieved from airlines.org: http://airlines.org/data/annual-u-s-impact-of-flight-delays-nextor-report/

Altair. (2019). *Locations of US Airports*. Retrieved from Altair: https://altair-viz.github.io/gallery/airports\_count.html

Andrew Hough. (2010). *Holidaymakers waste 16 days waiting in airports over lifetime*. Retrieved from telegraph.co.uk: https://www.telegraph.co.uk/travel/travelnews/7546209/Holidaymakers-waste-16-days-waiting-in-airports-over-lifetime.html

FAA. (2019). *Flight Delay Information*. Retrieved from fly.faa.gov: https://www.fly.faa.gov/flyfaa/usmap.jsp

FlightAware. (2019). *MiseryMap*. Retrieved from FlightAware: https://flightaware.com/miserymap/

Group56. (2013). *Heat Mapping and Predicting Flight Delays and Their Propagations in a Real-World Air Traffic Simulation.* Stanford. Retrieved from http://snap.stanford.edu/class/cs224w-2013/projects2013/cs224w-056-final.pdf

IATA. (2018). *IATA Forecast Predicts 8.2 billion Air Travelers in 2037*. Retrieved from IATA: https://www.iata.org/pressroom/pr/Pages/2018-10-24-02.aspx

Plotly. (n.d.). *Ploty*. Retrieved from Plotly Graphing Libraries: https://plot.ly/python/choropleth-maps/

quarbby's. (2015). *Heatmap.* NA: Plotly. Retrieved from Plotly: https://plot.ly/~quarbby/58

TutorialGateway. (2019, 07 25). *TutorialGateway*. Retrieved from TutorialGateway: https://www.tutorialgateway.org/tableau-heat-map/