# **Flight data visualization using Tableau**

Author: Ashish Sharma

Date: 04/10/2018

Link to Workbooks:

**Initial design** : <https://public.tableau.com/profile/ashish.sharma4329#!/vizhome/AirlineOn-TimeStatisticsandDelayCauses_012013_012018_v1_0/Dashboard22?publish=yes>

**Final** : <https://public.tableau.com/profile/ashish.sharma4329#!/vizhome/AirlineOn-TimeStatisticsandDelayCauses_012013_012018_v4_0/Story1>

## Summary

This dataset is obtained from the RITA website which contains information about flight delays and performance. The dataset I used ranges from 2013-2018. The dataset can be found in the following website:

<https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?pn=1>

I have created the visualizations in Tableau that answer the following questions:

* What are the percentages of different types of flight delays or delayed/cancelled/diverted arrivals and how they are distributed over months/year, carrier types and airports?
* What are the carriers that cause most delays when all the airports are considered and what are the airports that contribute for these delays the most for different carriers?
* What are the airports that cause the most delays when all the carriers are considered and what are the airports that cause most delays for a specific carrier?
* What are the top 10 airports and carriers that are responsible for the delays and when combined how do they affect each other?

## Design

My initial design choice for the first point of the story was to break down different delay types to see a comparison between them and how much each was contributing towards the total delay. Also, in parallel I wanted to break down even further by months to see if the delays had a high and low point during a particular time of the year. For this reason, I created a derived column called **'Total Delay**' where I took every delay and added them up. And then in this first plot – ‘Different Delays’ worksheet, I placed it as “All Delay” to make a point of reference. Here we can also filter by airport, carrier and year to drill down even further.

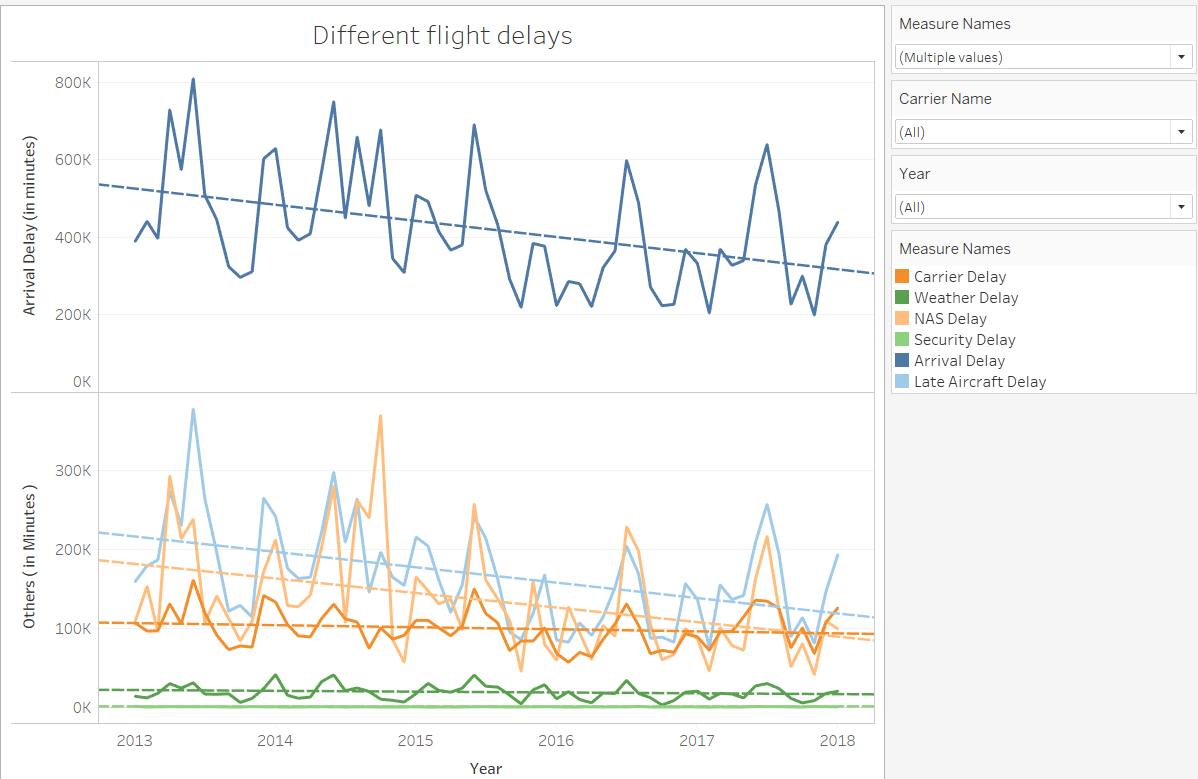
A screenshot of a social media post

Description generated with very high confidence

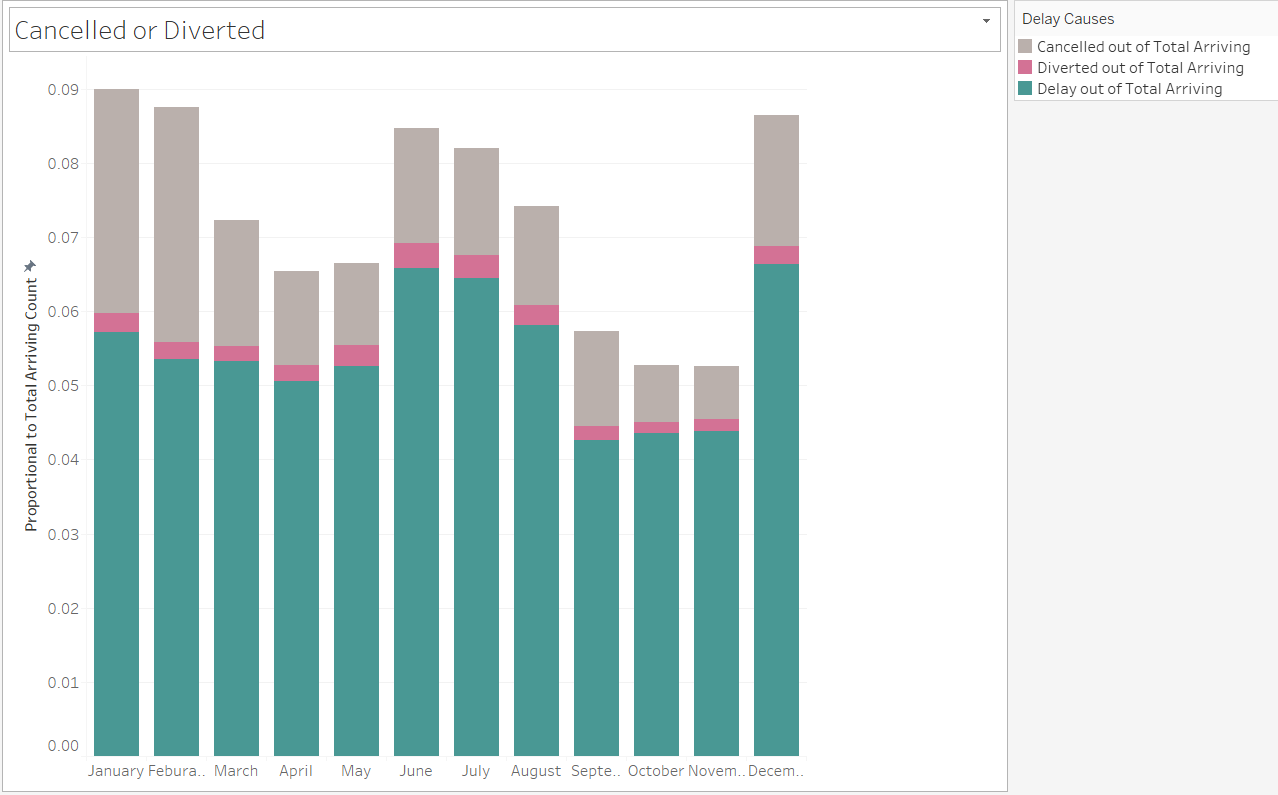
For each of the delays in the top plot to a proportion which tells us what percentage of the total delay the different delays contributed.

The next step was to understand the different types of delays distributed across several years and different months. During my exploration, I found that ‘Arrival Delay’ has the most significant impact and should be evaluated separately than other delays. Hence, I plotted ‘Arrival Delay’ and ‘Other Delays’ separately in the same plot. I created new calculated variables ‘MonthYear’ and ‘Month’ to create timeline in my plot across months and years. When I plotted the delay in minutes across the timelines, an interesting fact was unveiled -

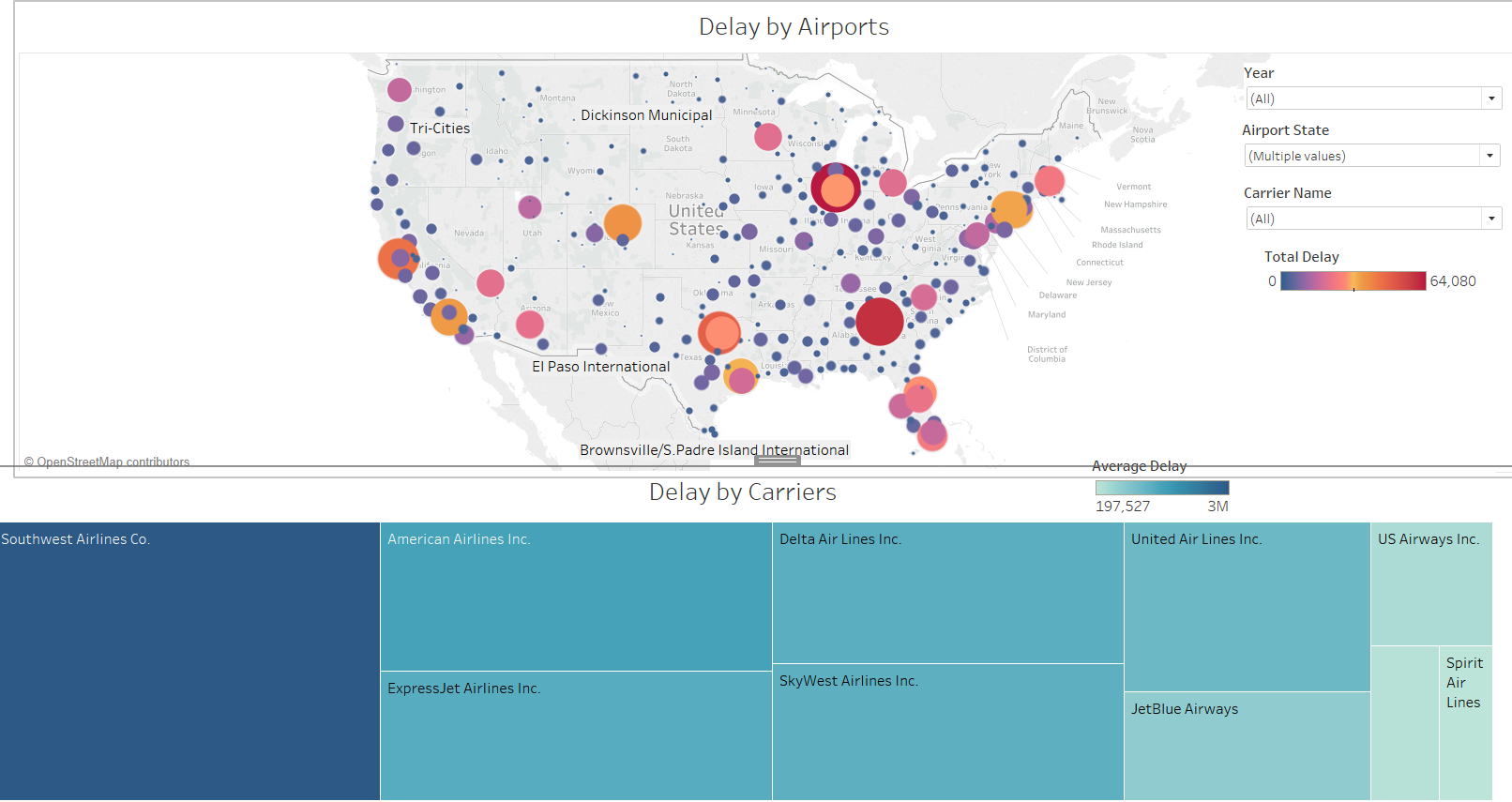
“Flight Delays due to various reasons peaks at around June-July and around December. We see that the security check delay contributes the least and the arrival delay contributes the most towards the overall delay. We also see that most flights get cancelled during January-February and most get delayed around May-July period. This may be due to extreme weather or holiday seasons as we see it peaks around summer and winter.”



In addition to that, I think it would make more sense to understand what proportion of total arrived flights were cancelled or diverted. It confirms the analysis from above graphs where June-July and December-January are most impacted months due to delays.

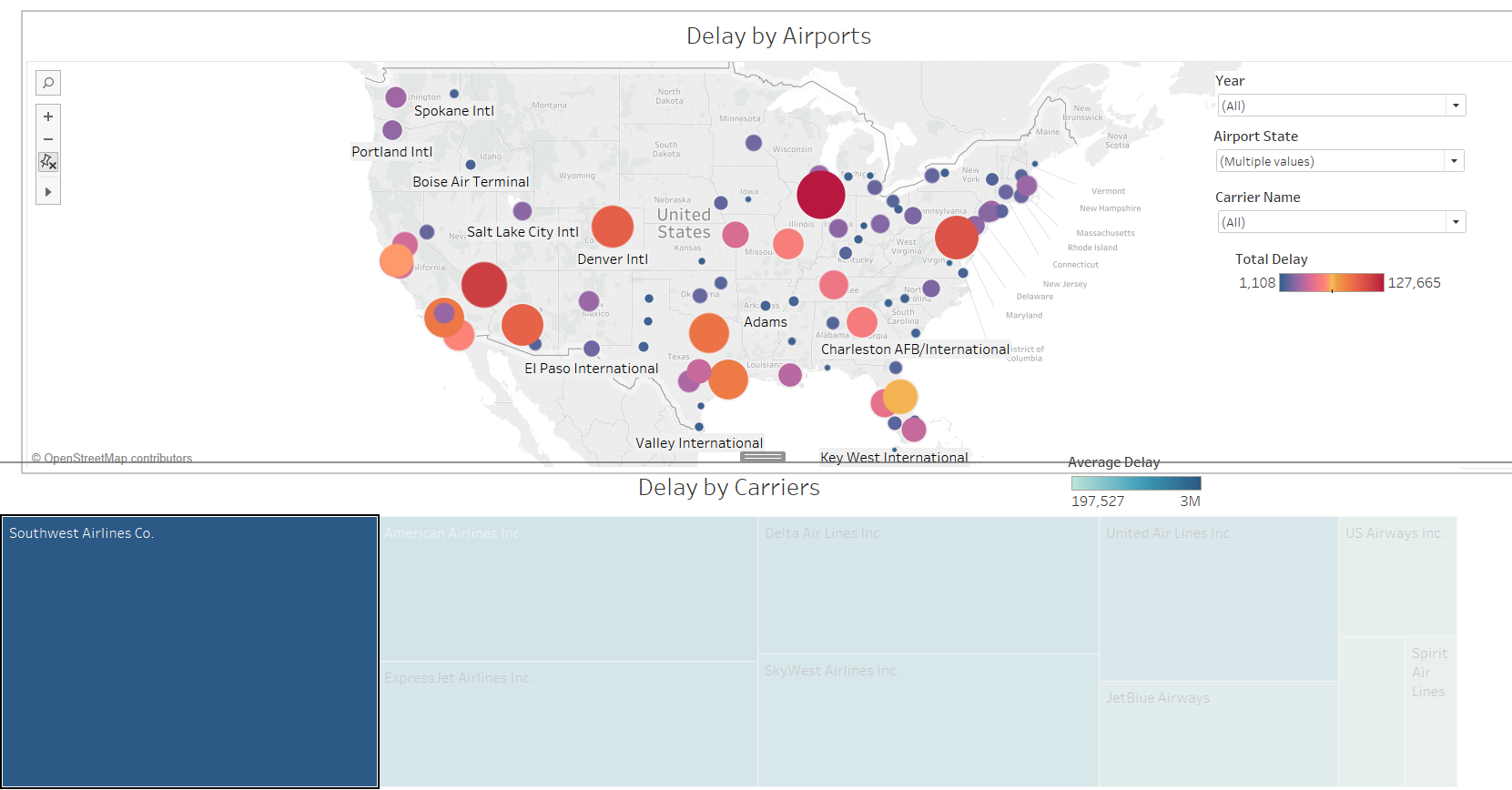


Now my next target was to see if there's a way to find out which were the carriers that got delayed the most when all the airports and all the years were considered. Also, in the same pair of plot, I wanted to see that for a carrier, what are all the airports that contributes the most for its delay. So, I made a heatmap for the carriers and mapped it to the above geolocation plot so that we can see for a year, which were the carriers that got delayed the most and what were the airports that contributed most for the delay. I also created an action here such that if only one flight was clicked in the heatmap, the corresponding airports from which it got delayed were shown on the map.



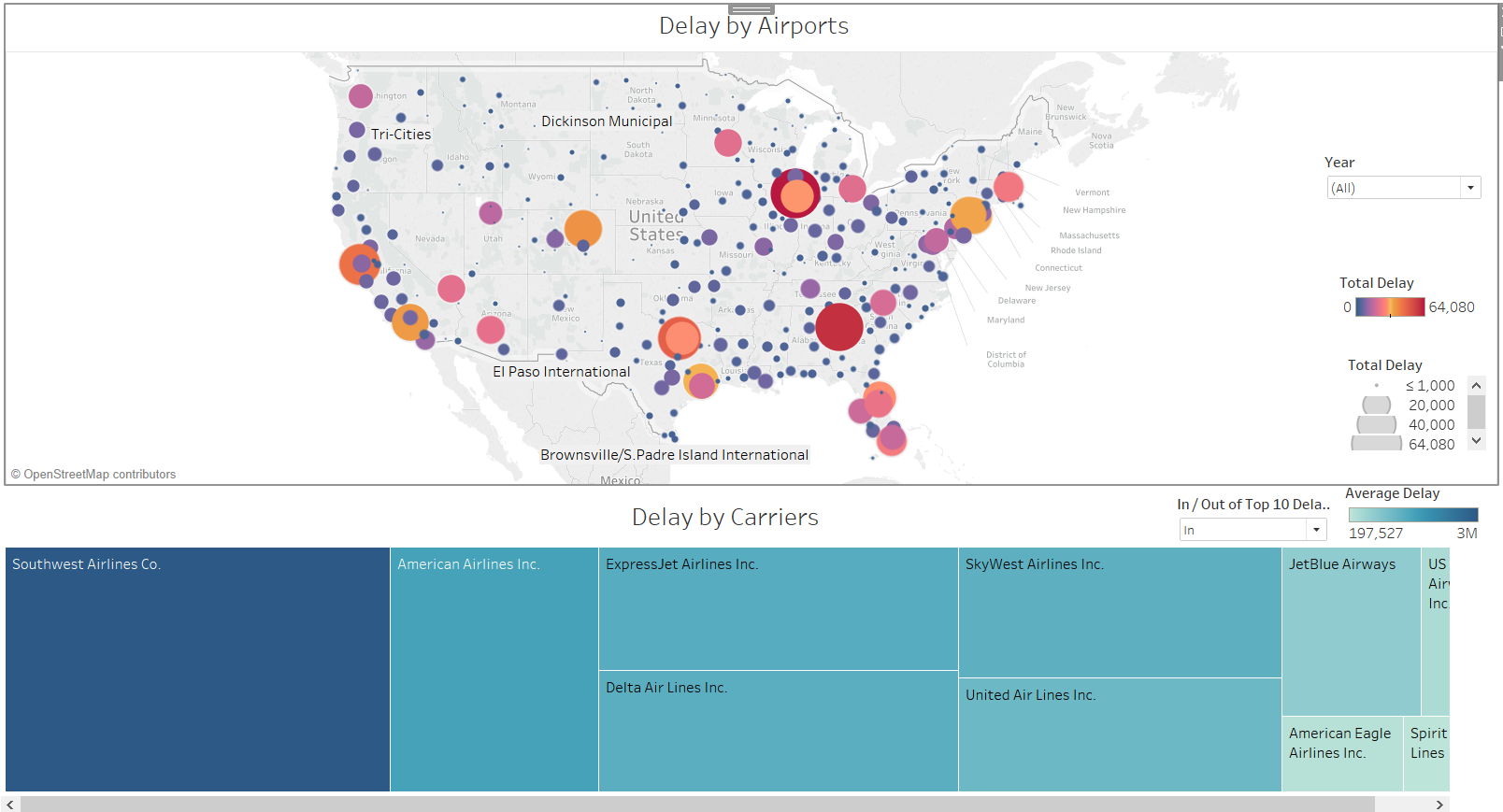
Here we see the average delay based on different carriers spread across different airports in USA. We see that when all airports are taken into considerations across the map, Southwest Airlines gets delayed the most followed by United Air and American Airlines.

When a carrier is clicked:



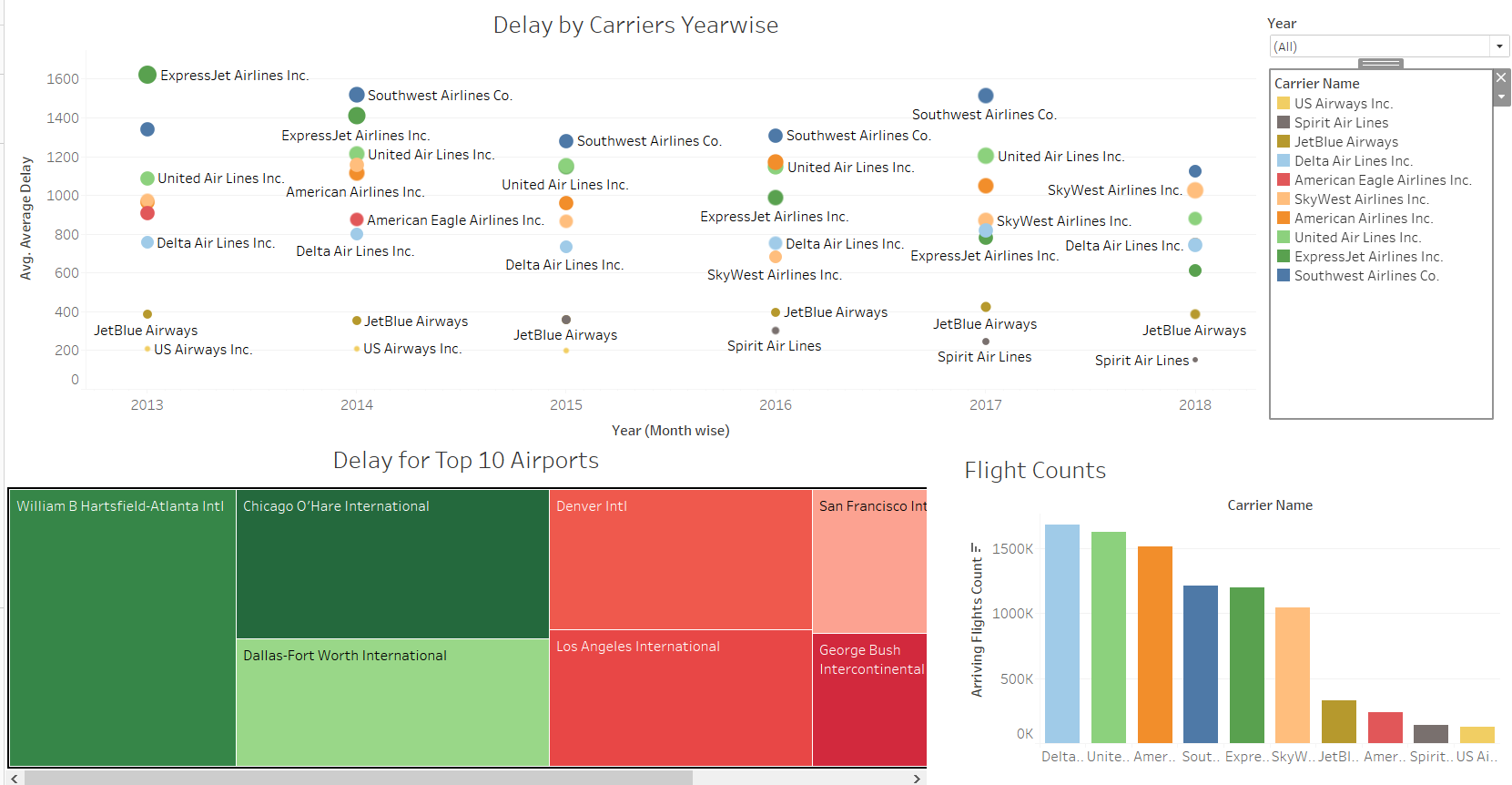
When we click Southwest Airlines in the bottom plot, we see the top 3 airports contributing towards its delay are Chicago, Las Vegas and Baltimore. Similarly, we can see by combining different airport names with year that which flights were delayed the most.

Further, I added another dimension to this plot, using new set I created – “Top 10 Delayed Carriers” where I filtered top 10 carriers using ‘Total Delay’ was feature.

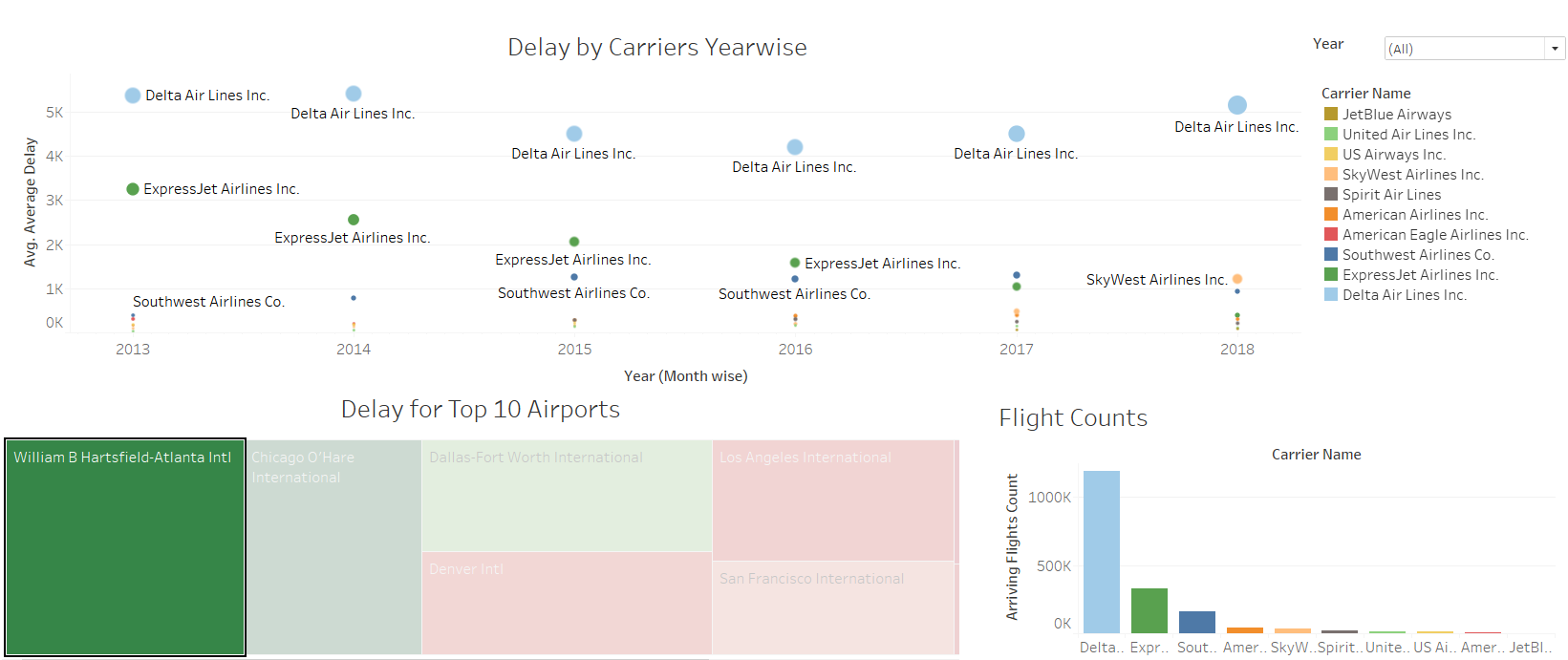


We see if we take all the years and all the carriers, the airport which caused the most delay is Chicago O'Hare airport followed by Atlanta and Texas. When the carrier is Delta airlines, we see that Atlanta caused the most delay. For Chicago airport, Southwest airlines is not even in top 10 delayed carriers. We can also hover our mouse on the bottom plot to see the exact geolocation of that airport.

Finally, I used ‘Top 10 airport’ set to explore which carriers were most impacted at a specific airport.



We see that the results change a bit. Now when we plot all the 10 airports on the map and plot the corresponding airlines which got most delayed (within top 10 overall).



We see though Southwest is still leading the delay charts, but when we select individual airport, the situation changes.eg. For Atlanta airport, Delta and ExpressJet Airlines contribute to most of delay. Similarly, for Chicago, Southwest is not even in top 10 delayed carriers. This means the most delay of Southwest is not from the most delayed airports but rather spread over the all the airports.

Feedbacks:

I received these set of feedback from my peers at work:

1. In my initial design, my team mate pointed out that displaying numeric value for month in ineffective and it should be transformed into categorical value. He suggested me to use calculated fields for this cause.

Hence, I added ‘Month’ as a calculated column.

1. He also pointed out that since data is distributed over years, collective dimension of month and year would be helpful to observe the trend. Hence, I added ‘MonthYear’ as additional column.
2. Also, for Delay for Top 10 Airports, initially I used bubble map, but he suggested to add treemaps as they show variance between different airports much more vividly.
3. For ‘Cancelled or Diverted’ I was using line graph which was not helpful for my thesis statement, when I transformed to bar graphs, the month and cause became apparent.
4. The most important feedback I received was to use filters in my dashboards. It was immediate impact, as I can see reflection of selected values in another graph. This helped a lot in creating true picture.

6. Add filters appropriately to see which airlines have delays and their causes.

7. Added various tooltip to mention percentage of whole that each bar represents

References:

* https://www.transtats.bts.gov/OT\_Delay/OT\_DelayCause1.asp?pn=1
* https://www.bts.dot.gov/explore-topics-and-geography/topics/airline-time-performance-and-causes-flight-delays
* https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations
* https://www.transtats.bts.gov/glossary.asp
* Tableau training videos: https://www.tableau.com/learn/training