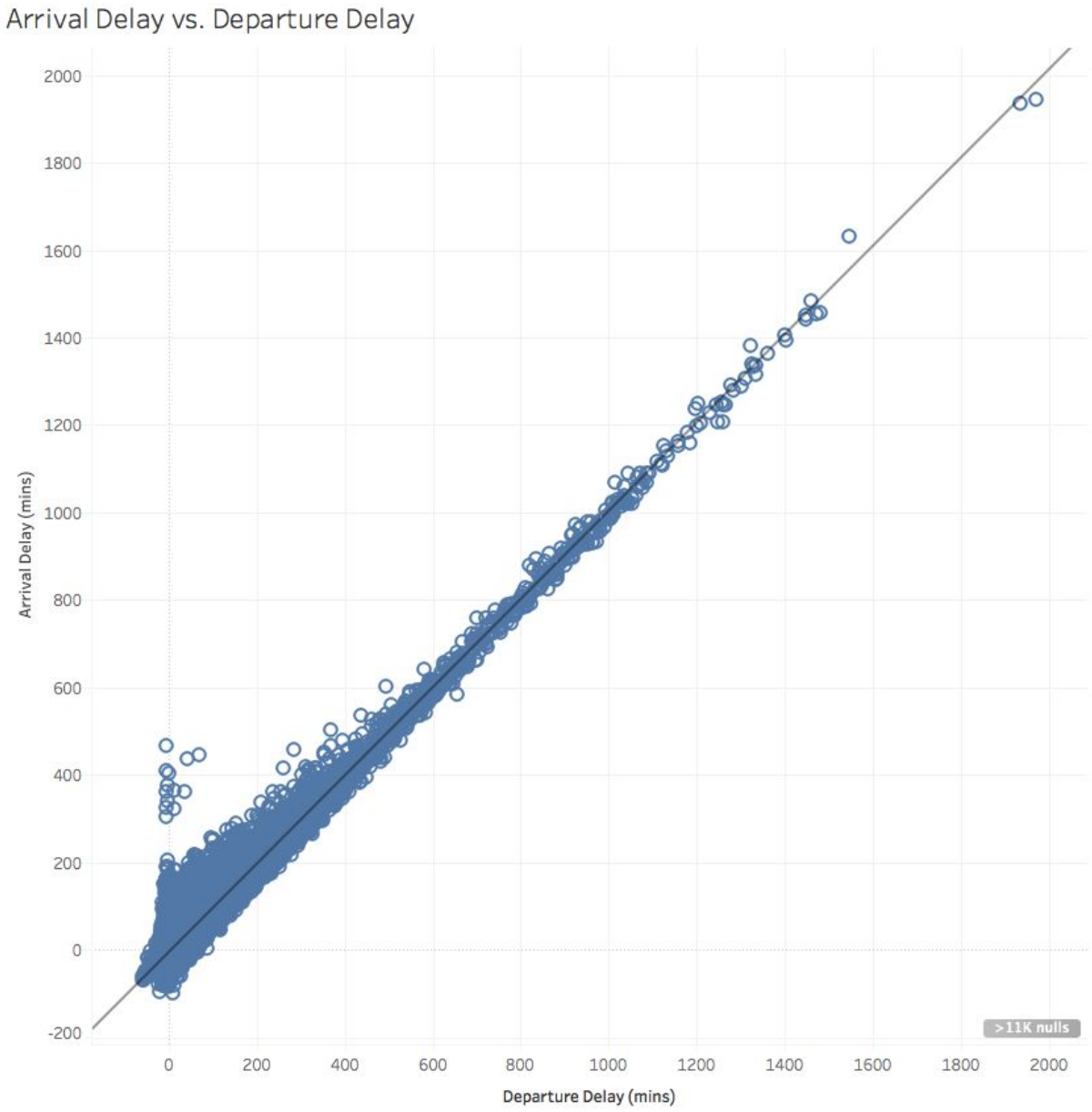
A2. Exploratory Data Analysis

Mohammad Habib

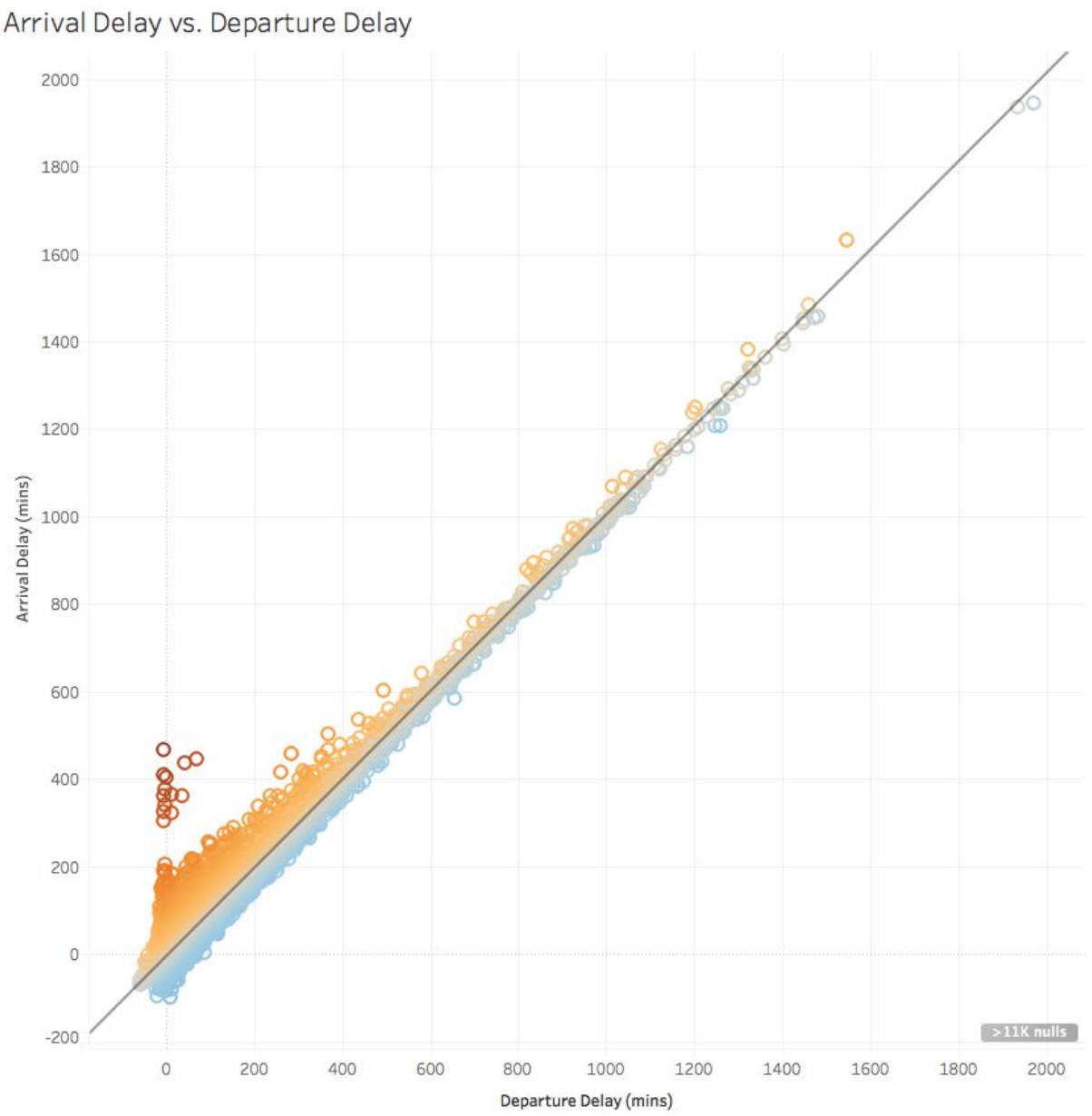
W209 - 4

Hypothesis 1: A Departure delay of X-minutes will cause an Arrival delay of at least X-minutes.

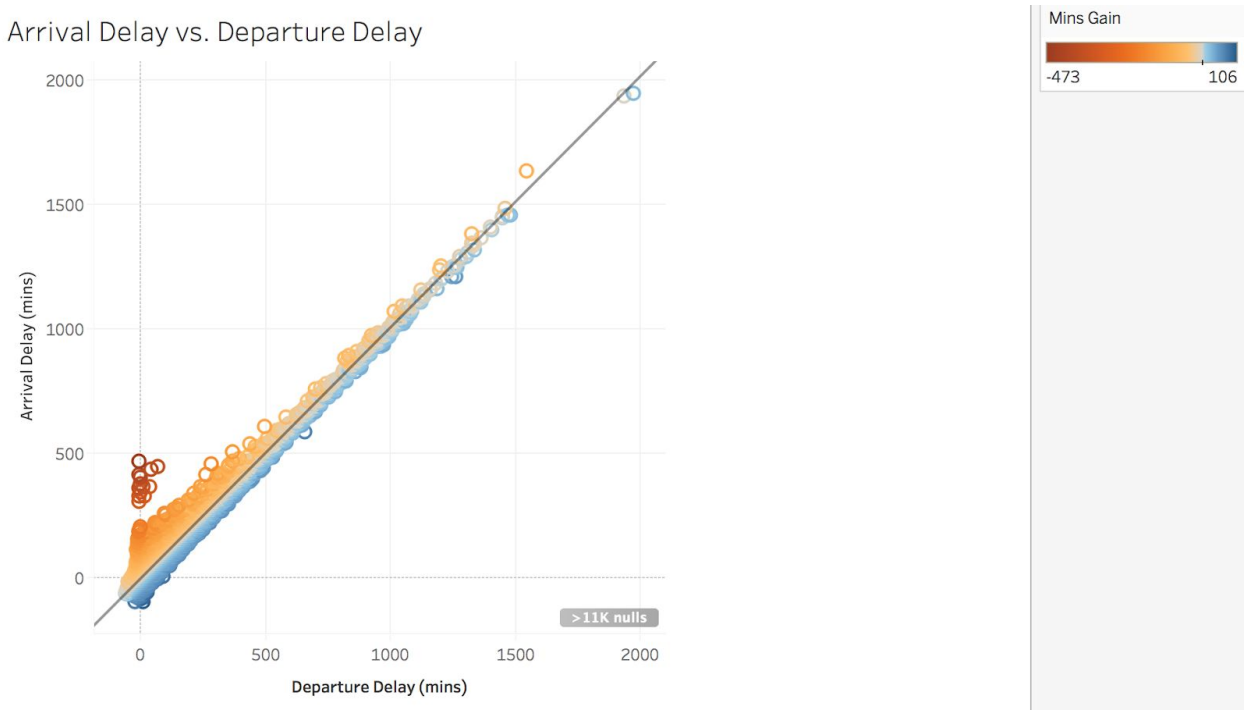


What's informative about this view: This view shows a scatterplot with a regression line where Departure Delay is the independent variable and Arrival Delay is the dependent variable for flights data in 2017. The dots below the diagonal line represent flights that Departed later than their scheduled time but Arrived sooner than their scheduled time i.e. they made up time in the air. It is interesting to note, however, that most of the flights were unable to gain time in the air.

What could be improved about this view: It would be interesting to add color as a visual indicator of minutes gained or minutes lost in the air i.e. flights that gained minutes in the air could be colored blue and those that lost time could be colored red -- on a diverging color scale that uses “minutes gained” for coloring. Copied below is an updated chart using color as an indicator of delay. It is easier to see that the red flights lost time in the air whereas the blue flights gained time in the air.



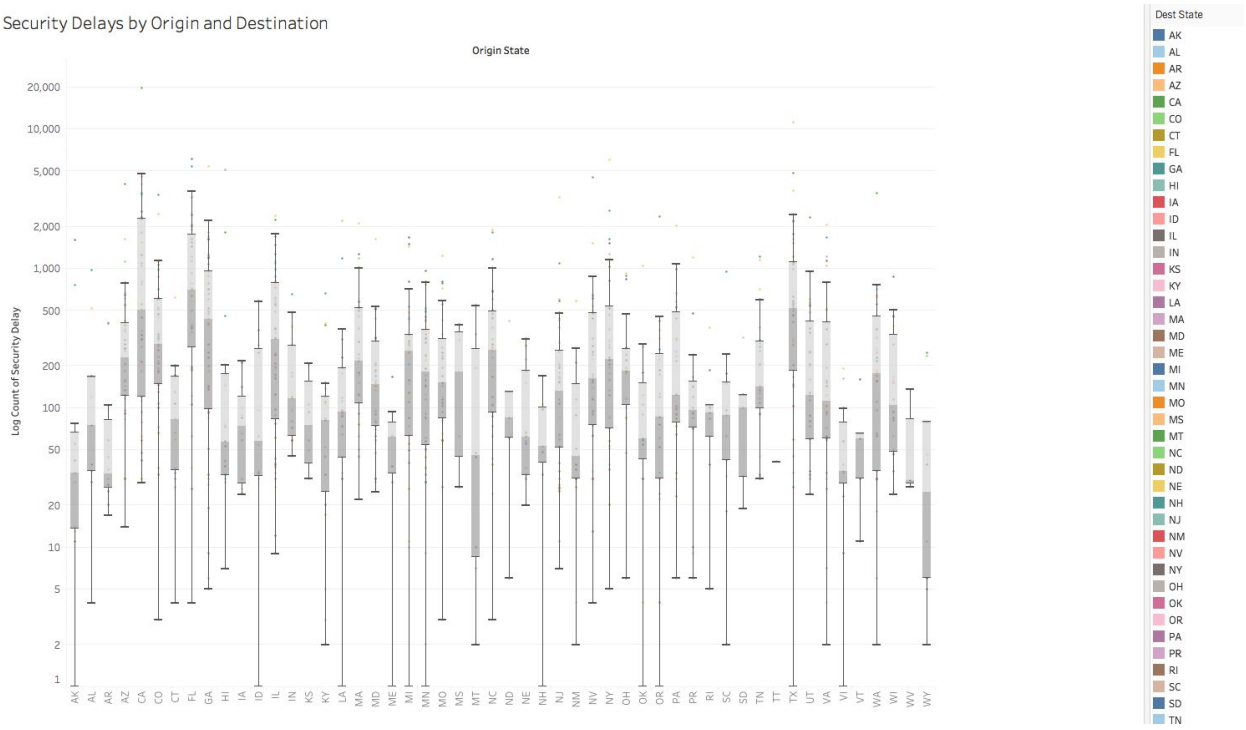
What could be improved about this view: It would be useful to add a color-scale legend, and the range of minutes gained.



Conclusion: These data suggest that some flights are able to make up time in the air, even when the departure is delayed.

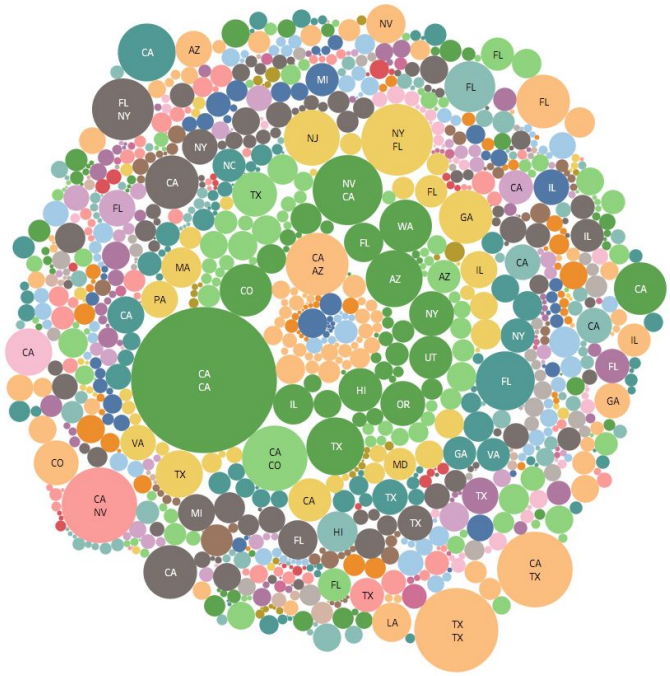
Hypothesis 2: Security delays affect more flights that travel from one State to a different State than flights within the same State.

Initial Chart: This is a boxplot that shows the origin State on the x-axis and the quartiles of log count of Security delays on the y-axis. It attempts to represent the destination State by Color which doesn't do a good job. I would find it more helpful if I saw origin and destination more clearly.



Final Chart: This bubble chart uses size to represent the count of Security delays in an airport, and uses color to represent the origin State. It is clear to see that the flights inside CA were delayed most times because of security reasons. It is also clear to see that flights that originate in CA are more likely to be delayed because of security reasons. Perhaps, this applies to air ports that serve international flights -- as can be seen in the graph for flights originating in NY and TX.

Security Delays by Origin and Destination

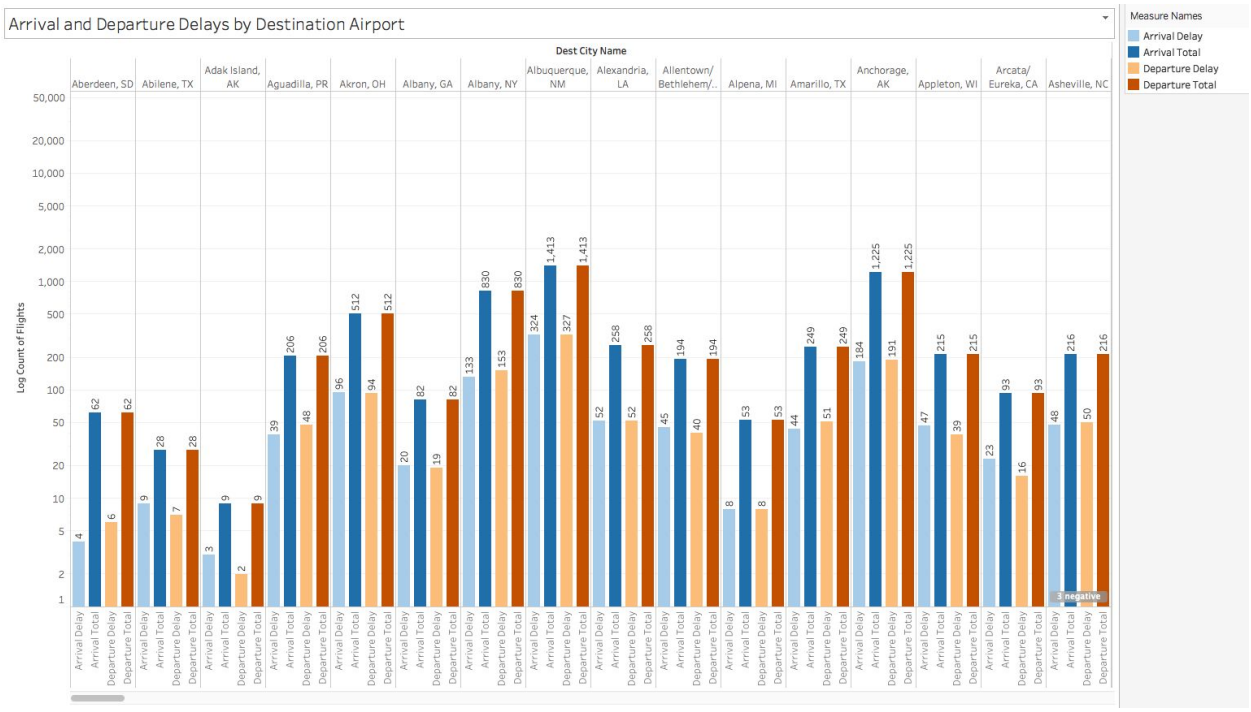


Origin State

- IA
- ID
- IL
- IN
- KS
- KY
- LA
- MA
- MD
- ME
- MI
- MN
- MO
- MS
- MT
- NC
- ND
- NE
- NH
- NJ
- NM
- NV
- NY
- OH
- OK
- OR
- PA
- PR
- RI
- SC
- SD
- TN
- TT
- TX
- UT
- VA
- VI
- VT
- WA
- WI
- WV
- WY

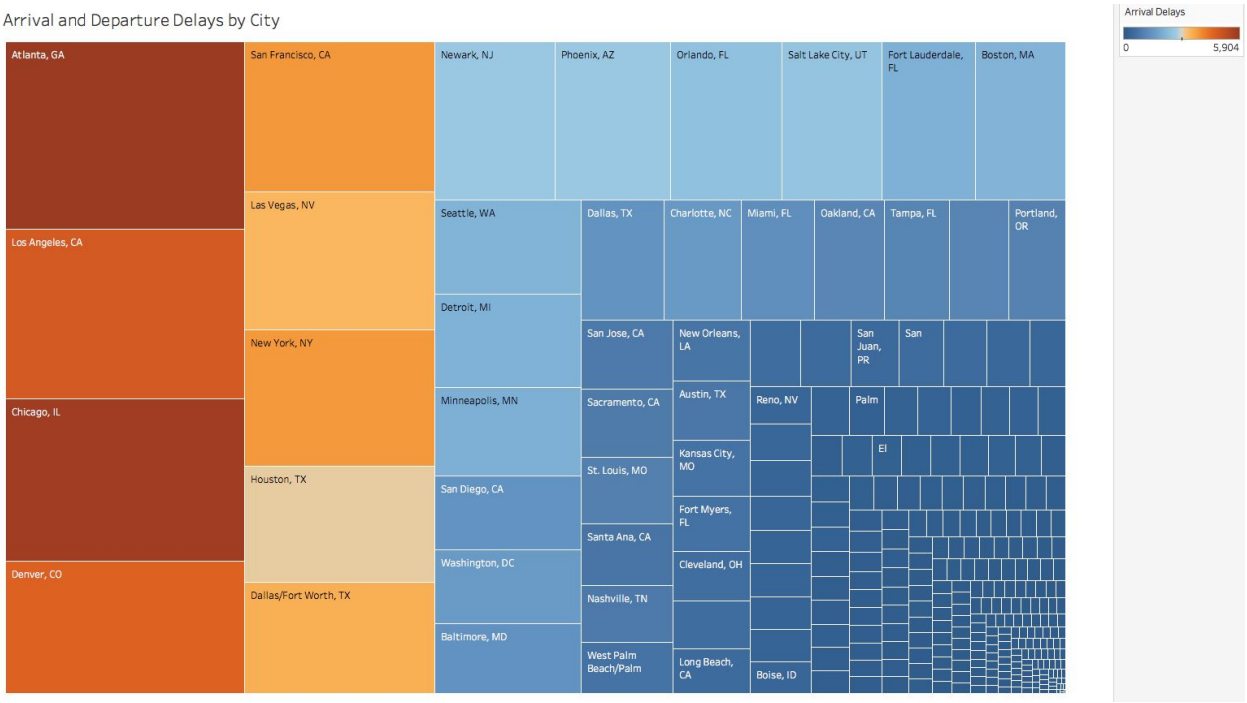
Hypothesis 3: Airports with a higher volume of flights are more affected by both departure and arrival delays than airports with a lower volume of flights.

Initial Chart:



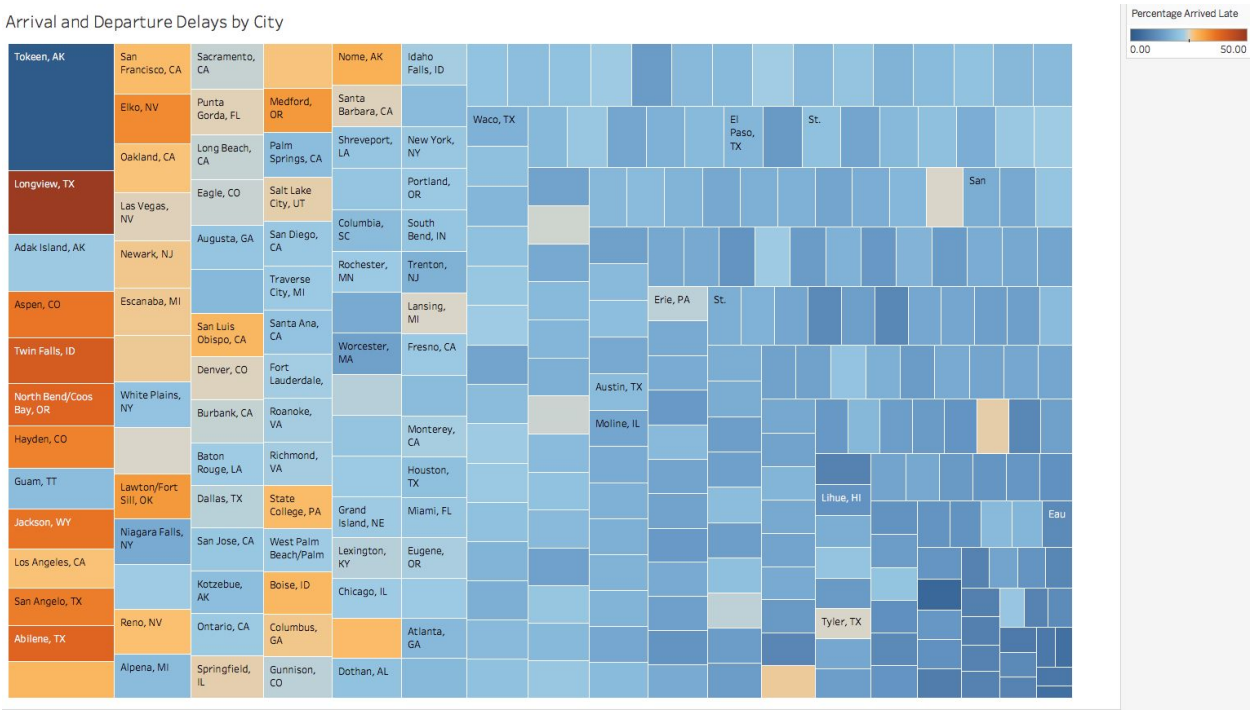
What's informative about this view: This chart is helpful because it shows the arrival and departure volume and arrival and departure delay for each city.

What could be improved about this view: This view does not provide a way to compare the arrival and departure delay of all cities side by side. So, we can improve that by using the total number of Departure Delays to represent Size in our graph, and the total number of Arrival Delays to represent Color in our graph.



What's informative about this view: This chart is helpful because it shows all cities in one picture. The orangest cities are those with the highest count of Arrival delays and the bluest cities are those with the lowest count of arrival delays. The biggest boxes represent those cities with the highest count of departure delays and the smallest boxes represent those cities with the lowest count of departure delays. We can conclude that the cities with the highest arrival delays also had the highest departure delays.

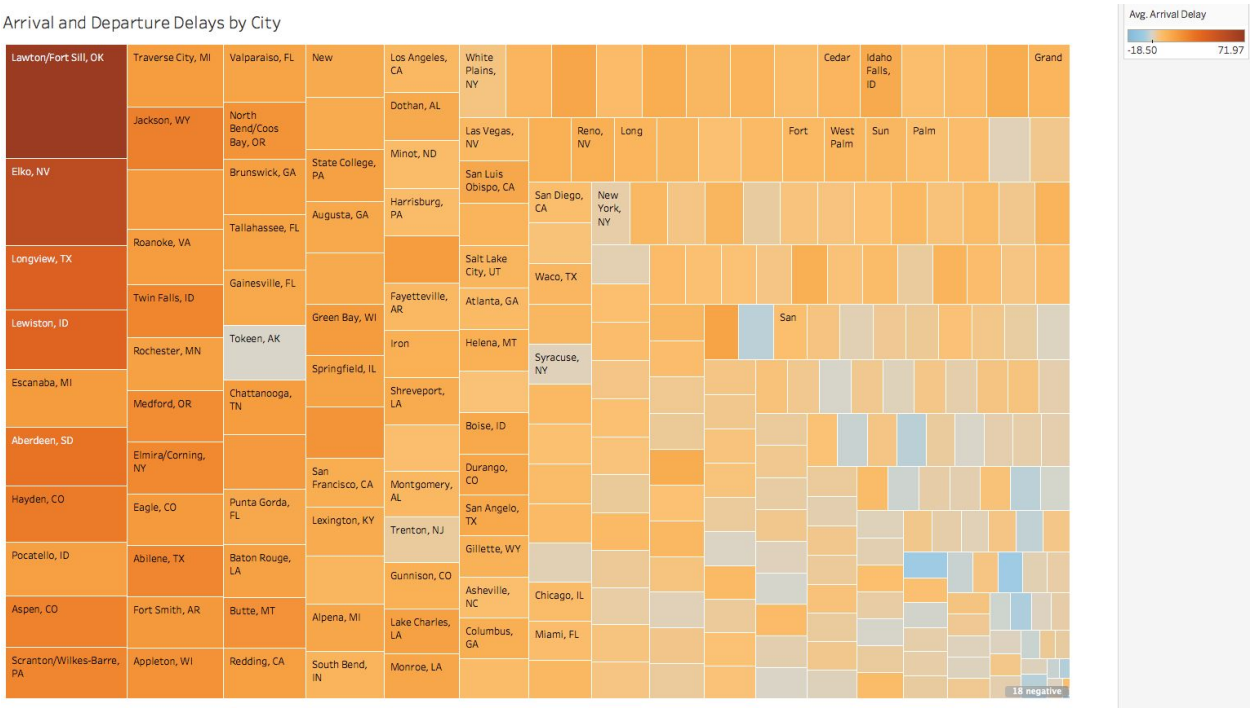
What could be improved about this view: This chart uses the total number of Arrival and Departure delays. We can improve it by using the Percentage of Arrival and Departure delays to represent color and size respectively.



The graph using percentages doesn't represent as good a picture as the one before because percentages take out the relative size comparison of volume. Because of this, the chart is dominated by smaller airports.

We can take into account the average time of delays and see if that gives us a better measure.

Final Chart:



Conclusion: I think using Avg. Arrival Delay and Departure Delay times makes for the best graphs. The cities that have the worst avg. arrival delay and departure delay are the orangest and biggest.