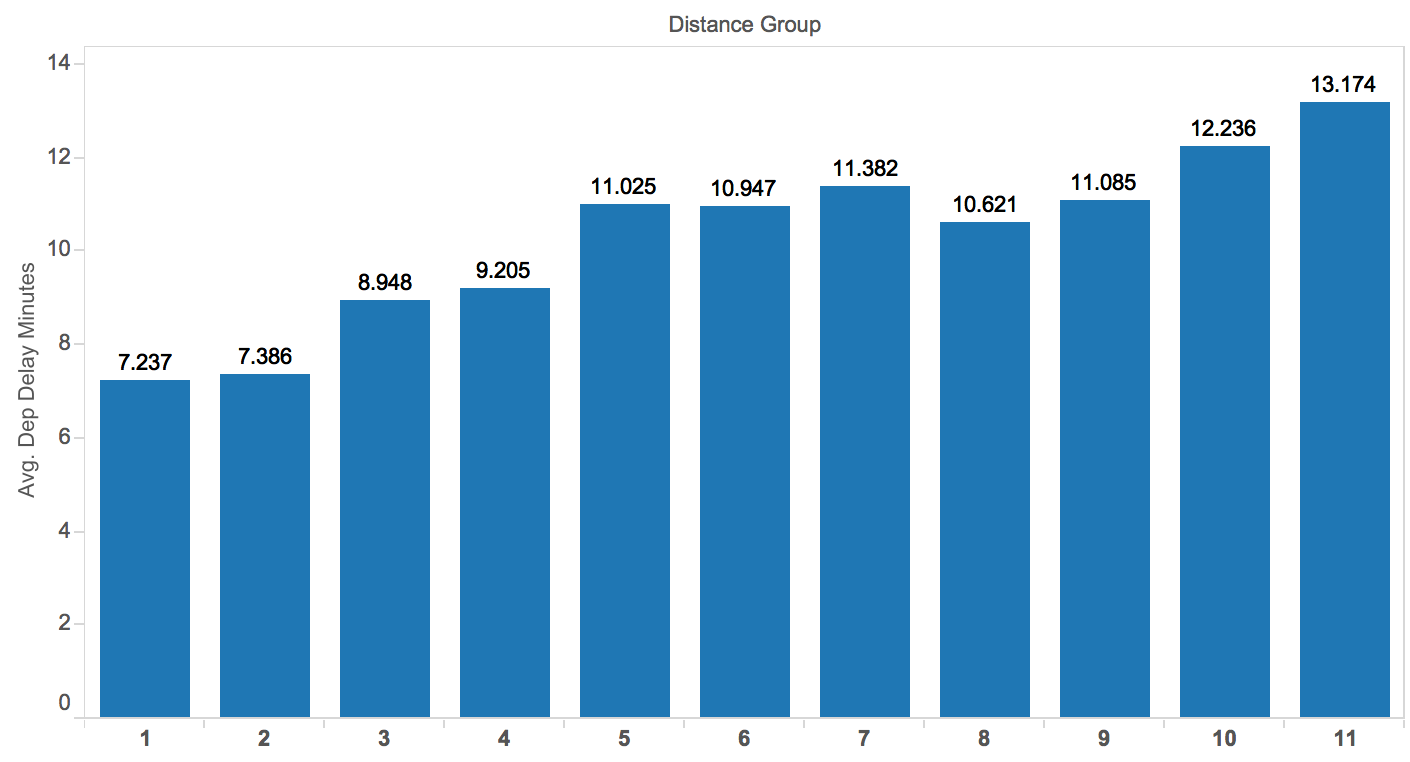
### A2. Exploratory Data Analysis (Template)

### Brandon Shurick

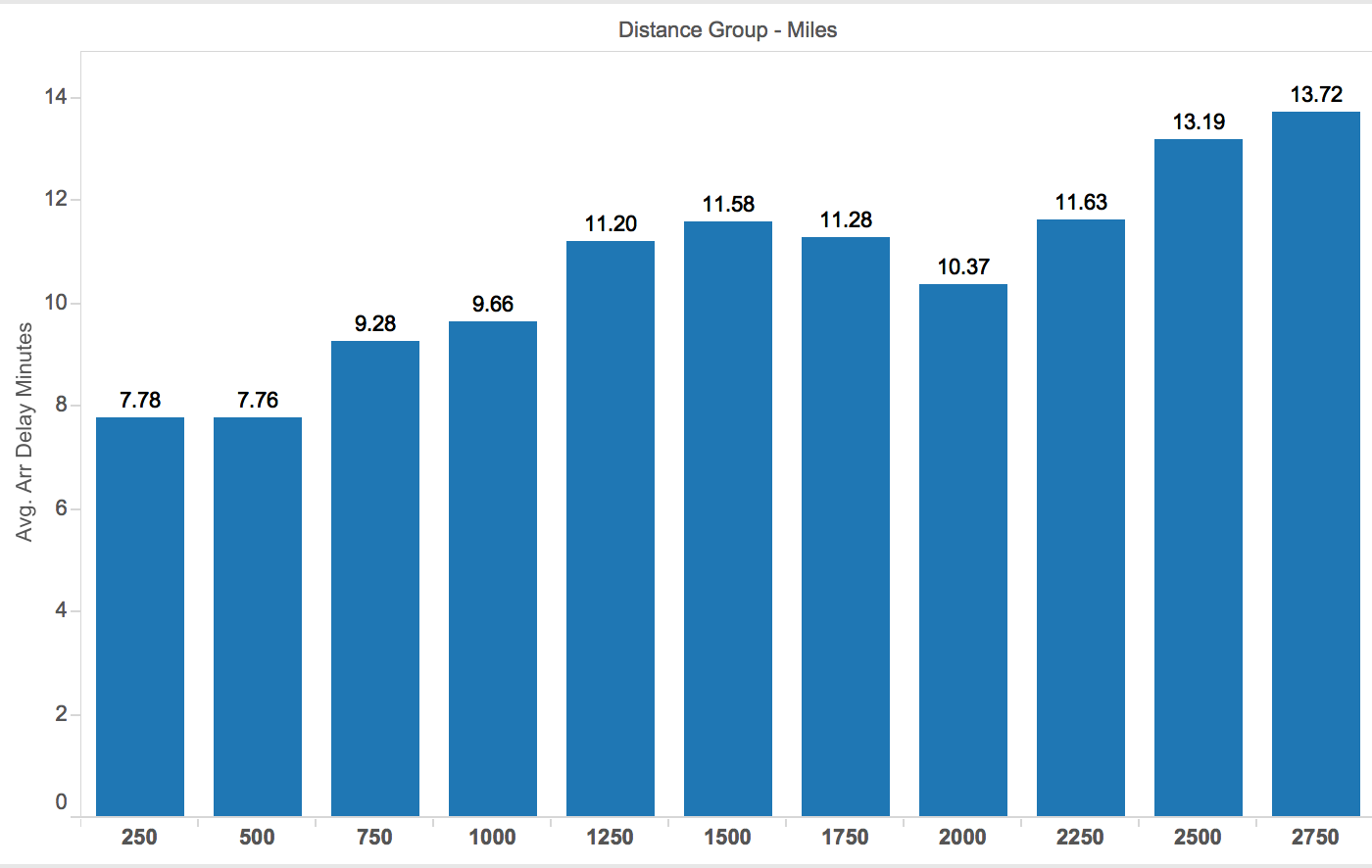
### W209 - 2

**Hypothesis 1:** Delays are more frequent when the aircraft originates from further distances.



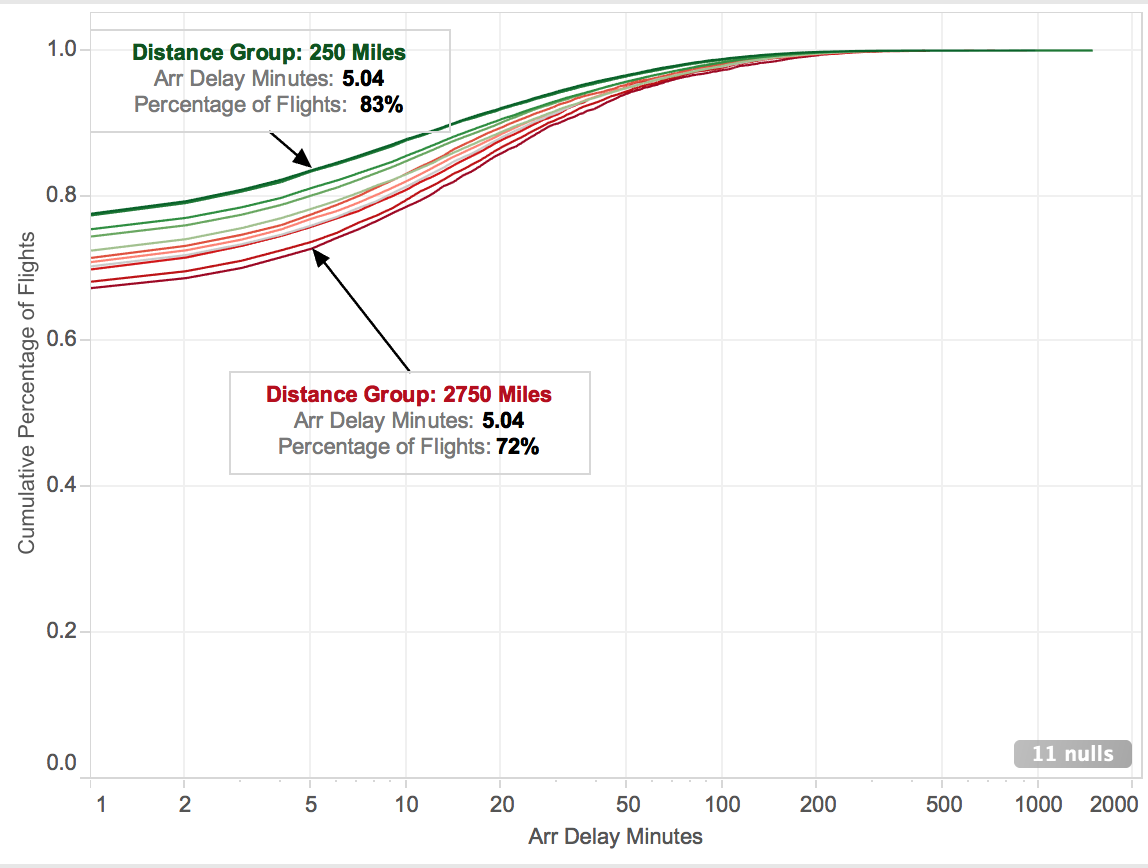
**What's informative about this view:** This view shows the average departure delay based on the distance of the flight. There is a clear correlation between delay and destination distance.

**What could be improved about this view:** This view does not quite get to the answer of the question that we want -- that delays are more frequent for long distances from the origination point. Also, it is simply showing the average delay, rather than that the frequency of occurrence. Further, it is unclear what Distance Group means in this chart.



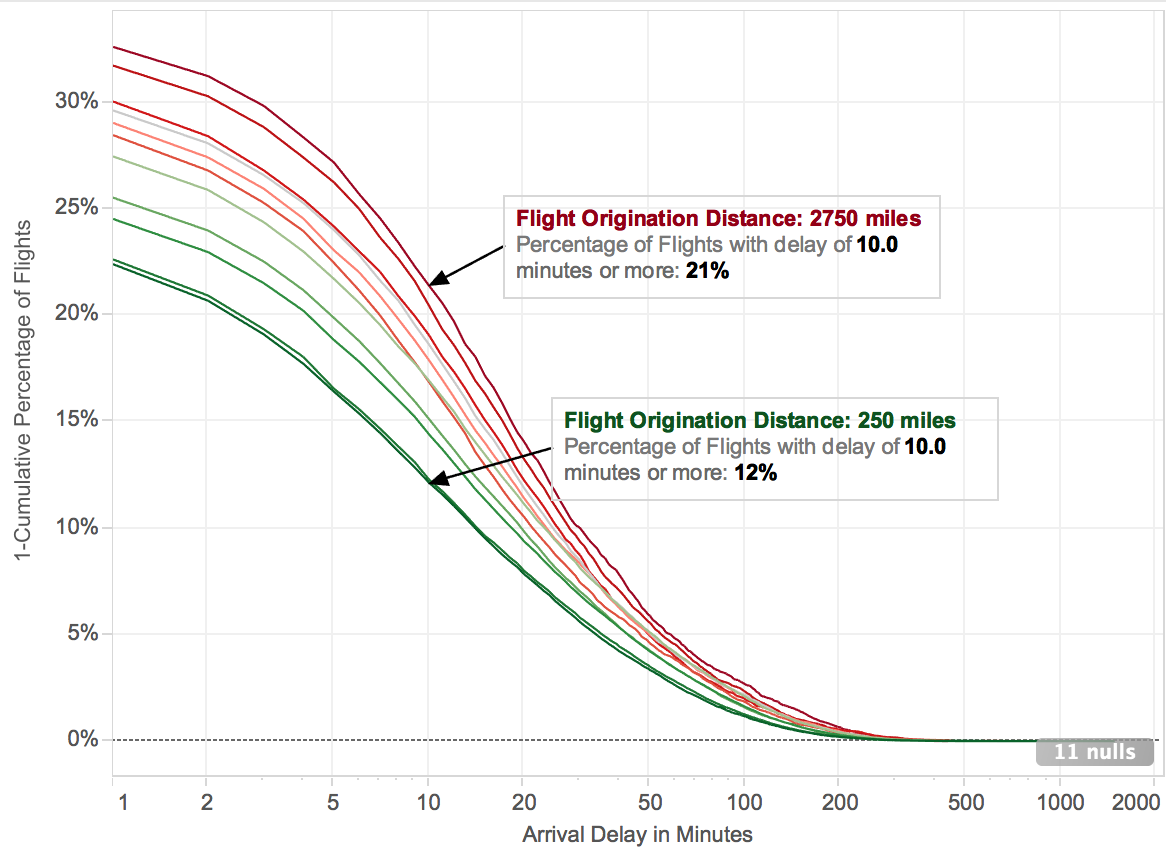
**What's informative about this view:** This view shows the average **arrival** delay by distance group, which is now a calculated field which is translated to miles. This view shows that there is a correlation between the distance of origination and delays, where it is inferred that the next departure delay for any flight will be directly related to the arrival delay of the prior flight.

**What could be improved about this view:**  This chart does not show the frequency of occurrence, as stated in the original hypothesis, but rather shows that the delay times on average are higher for flights that originate from larger distances. It might be necessary to show the distributions of arrival delay to somehow show that the probability of arrival delay is very high for flights that originate from long-distances.



**What's informative about this view:** This view shows the cumulative percentage of flights by distance group along the arrival delay axis. The arrival delay in minutes has a logarithmic axis, and instead of a legend there are labeled points that clarify the distance groups.

**What could be improved about this view:**  The chart is a bit misleading because it is showing the cumulative percentage of flights that have delay up until X point, for example the percentage of flights that have a delay of up to 5 minutes is higher for lower distance groups, where higher is actually better. But this would be easier to interpret if we showed the percentage of flights that have a delay of **more than 5 minutes** instead (so, 1 minus the cumulative percentage of flights shown in this chart).

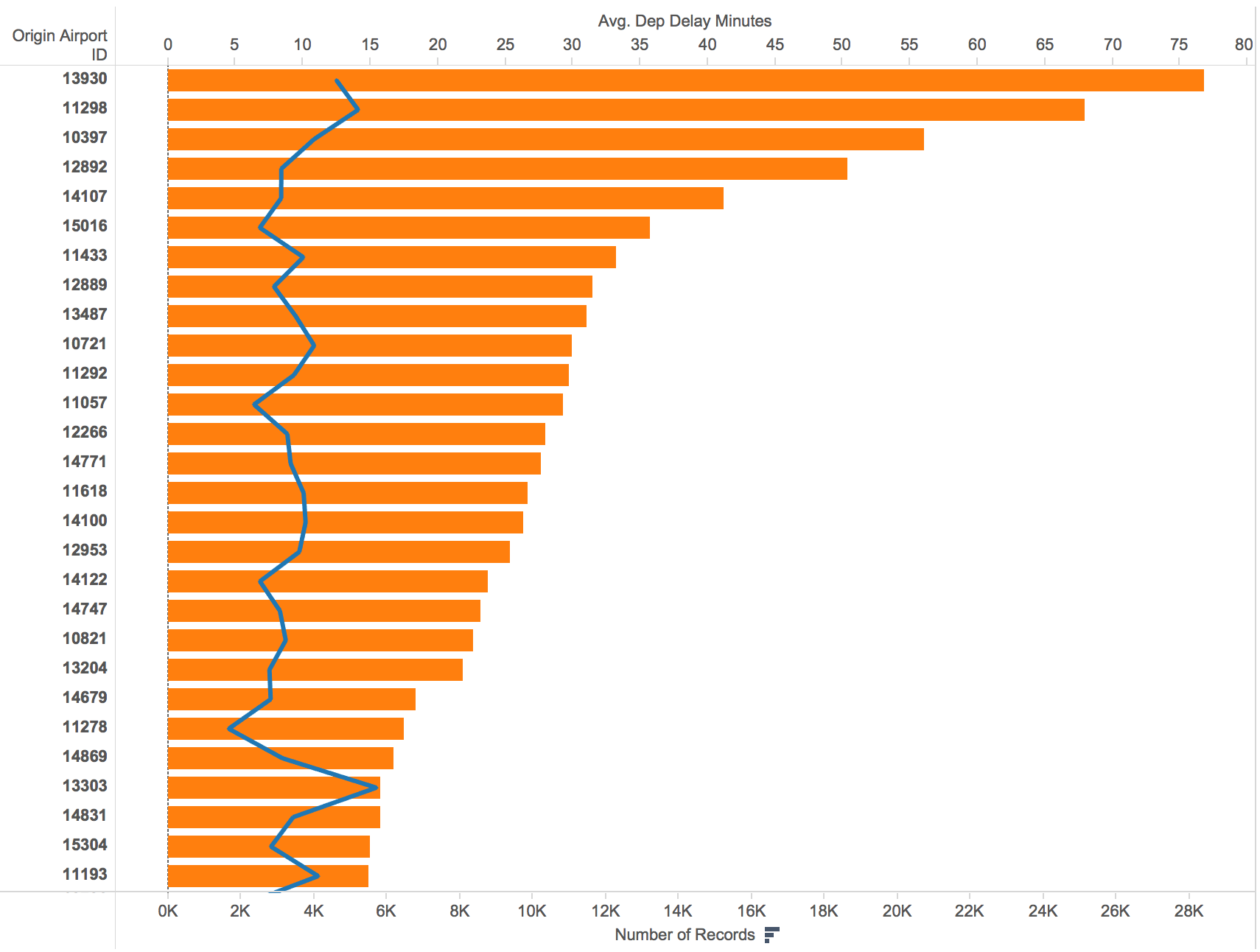


**What's informative about this view:** This view shows 1-cumulative percentage of flights by distance group along the arrival delay axis. The arrival delay in minutes has a logarithmic axis, and instead of a legend there are labeled points that clarify the distance groups.

**What could be improved about this view:**  My Y-axis description is not totally clear, but I can’t think of a better wording. Also I am only labeling two of the distance groups so it might be a bit misleading since the groups do not always have clear order inside of the chart. Also, I am still showing arrival delay which is a proxy for next flight departure delay, and assumes that those two are directly related, which may or may not be totally true.

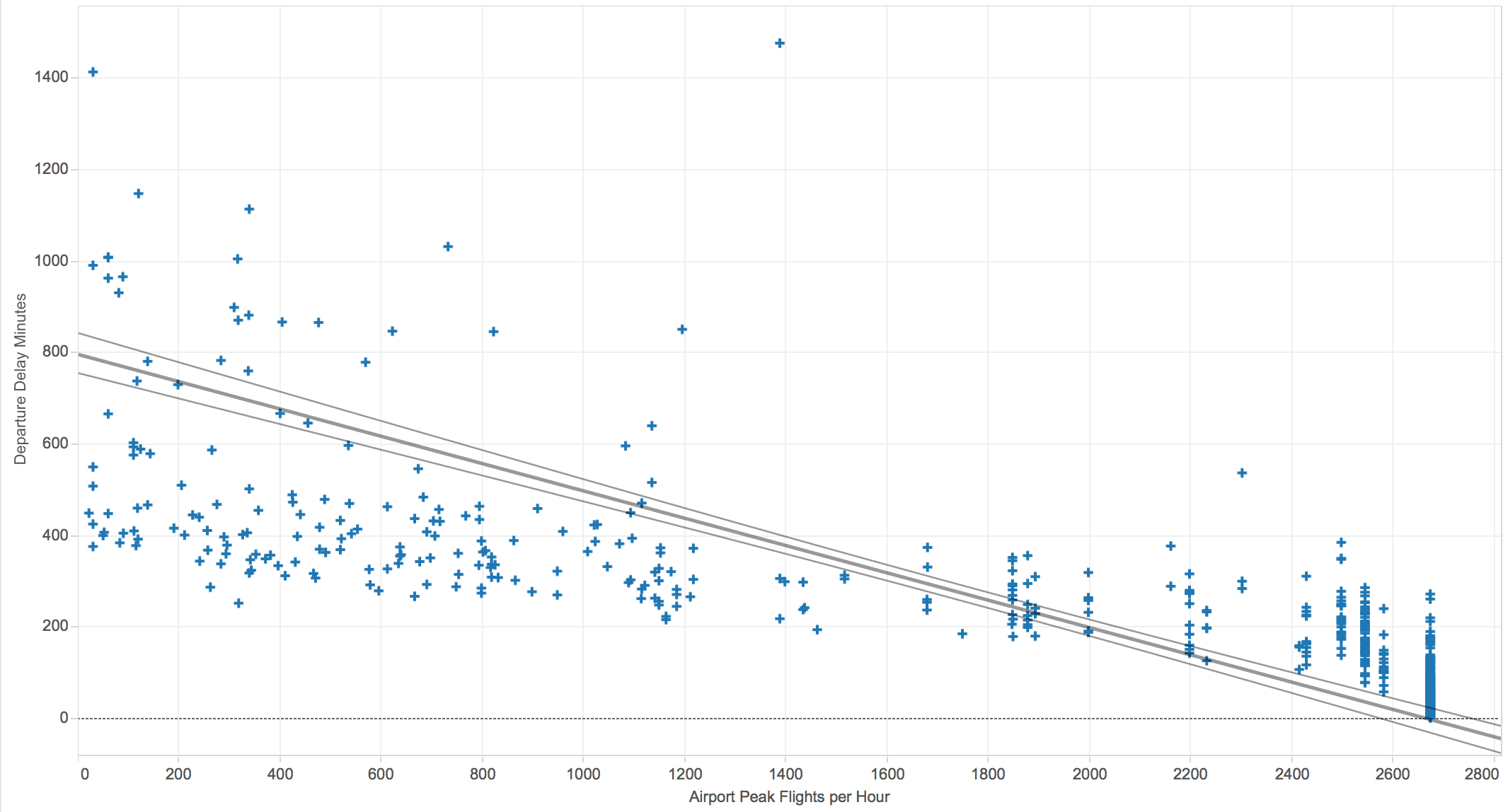
**Conclusion** (do the data appear to support the hypothesis, or not?): The final view does indeed show quite clearly that flights which originate from long distances are more likely to have a long delay. In the example shown, it is 75% more likely to experience a delay of more than 10 minutes if the flight before yours originates from across the country rather than a nearby city.

**Hypothesis 2:** Popular airports experience delays more frequently than less busy airports.



**What's informative about this view:** This view shows airports by ID code with average departure delay in minutes on one axis and number of flights on another axis. It shows us that there probably isn’t a correlation between delay minutes and airport popularity.

**What could be improved about this view:** A number of things could be improved -- the dual axis are probably not a good idea here due to how much of a difference there is in units between count of flights and average departure delay. Also, total number of flights is probably not be a great metric to show airport popularity.

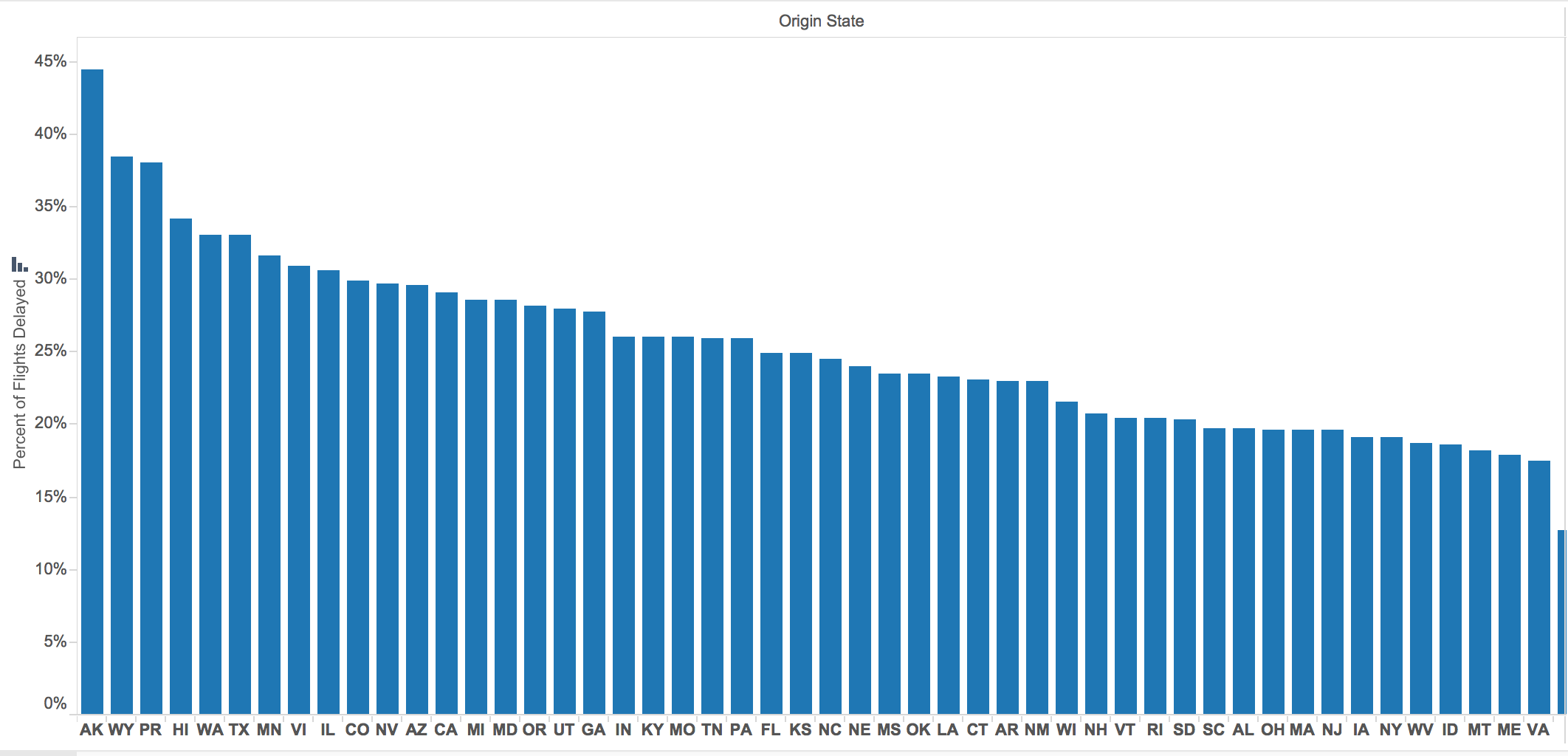


**What's informative about this view:** This view shows actually that there may be a **negative** correlation between the number of flights an airport can support at peak and the departure delay of flights. It shows also that variance in delay times decreases as the number of flights per hour increases.

**What could be improved about this view:**  The chart may be a little bit misleading because it is showing all flights for each airport on the Y-axis, but each point on the X-axis is the same for all flights for any given airport. Since the vast majority of flights occur at the largest airports, their data points are all stacked together on top of each other. It would be good also to display the linear model statistics on the chart, but I couldn’t immediately figure out how to do that.

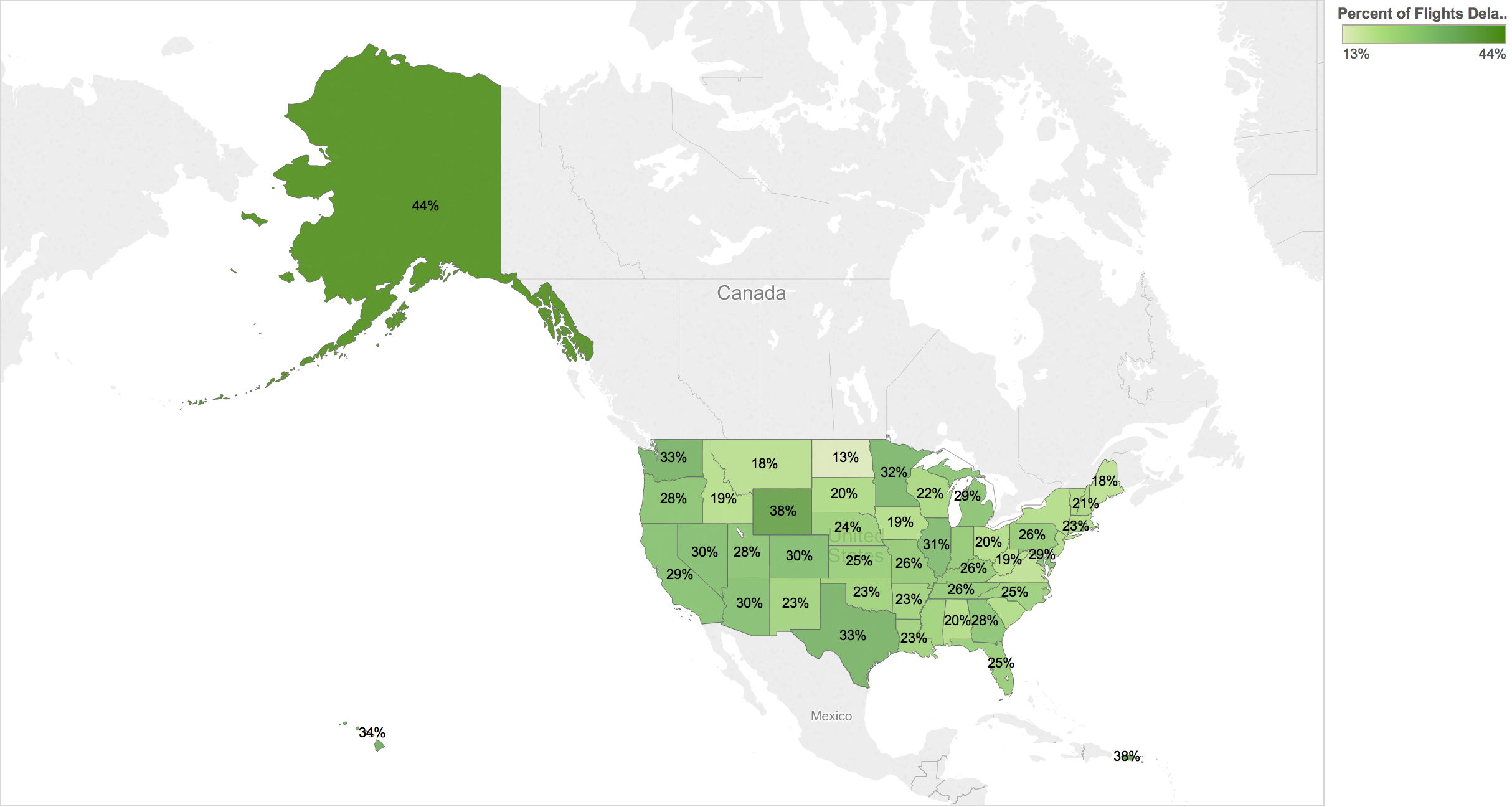
**Conclusion** (do the data appear to support the hypothesis, or not?): The last view seems to support actually the opposite of my original hypothesis -- flight delays as well as variance of delay times are higher for airports that don’t support very many flights, i.e. small airports.

**Hypothesis 3:** Flights originating from the East Coast are have delayed departures more often, and flights landing in the East Coast are more likely to be delayed (excluding Sept 11th).



**What's informative about this view:** It shows that the hypothesis is most likely incorrect, because the states which have the highest percentage of delayed flights appear to mostly be on the west coast or midwest.

**What could be improved about this view:** It is not easy to determine, based on this chart, whether or not the east coast comparison with other regions is accurate, since the states are not grouped by region. It would be best to create another field which groups states manually by region and show the results.



**What's informative about this view:** It pretty clearly shows that the hypothesis is incorrect while not removing any relevant data points from the prior view.

**What could be improved about this view:** Perhaps simply showing a comparison of percentage of delays by east coast versus “everything else” might better clarify that the hypothesis is incorrect; however, the chart would be much less useful overall.

**Conclusion** (do the data appear to support the hypothesis, or not?): The hypothesis is surely not correct, and it is likely that region is not a significant factor on flight delays.