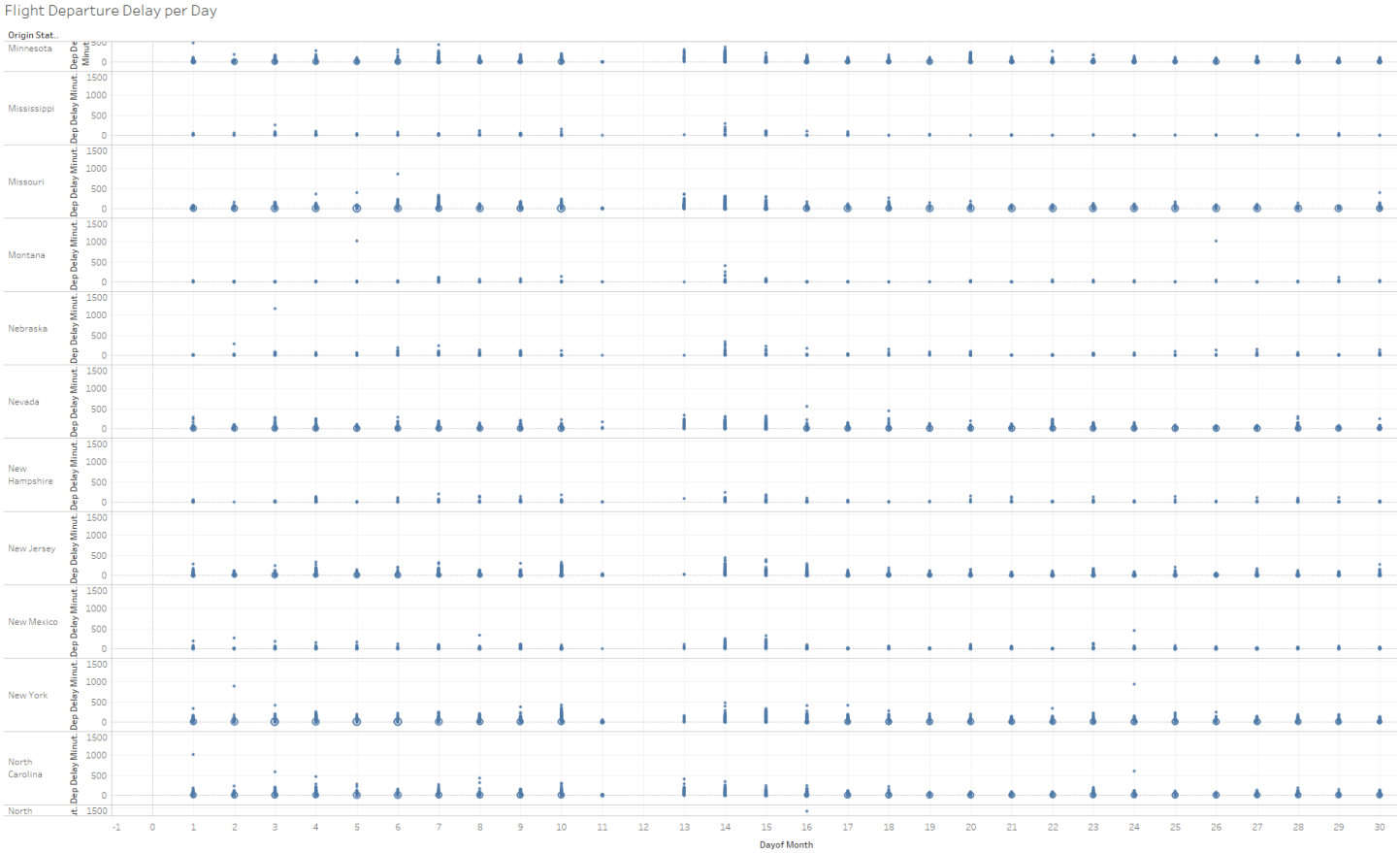
# A2. Exploratory Data Analysis (Template)

**Linfang Yang (yanglinfang@ischool.berkeley.edu)**

**W209 - 2**

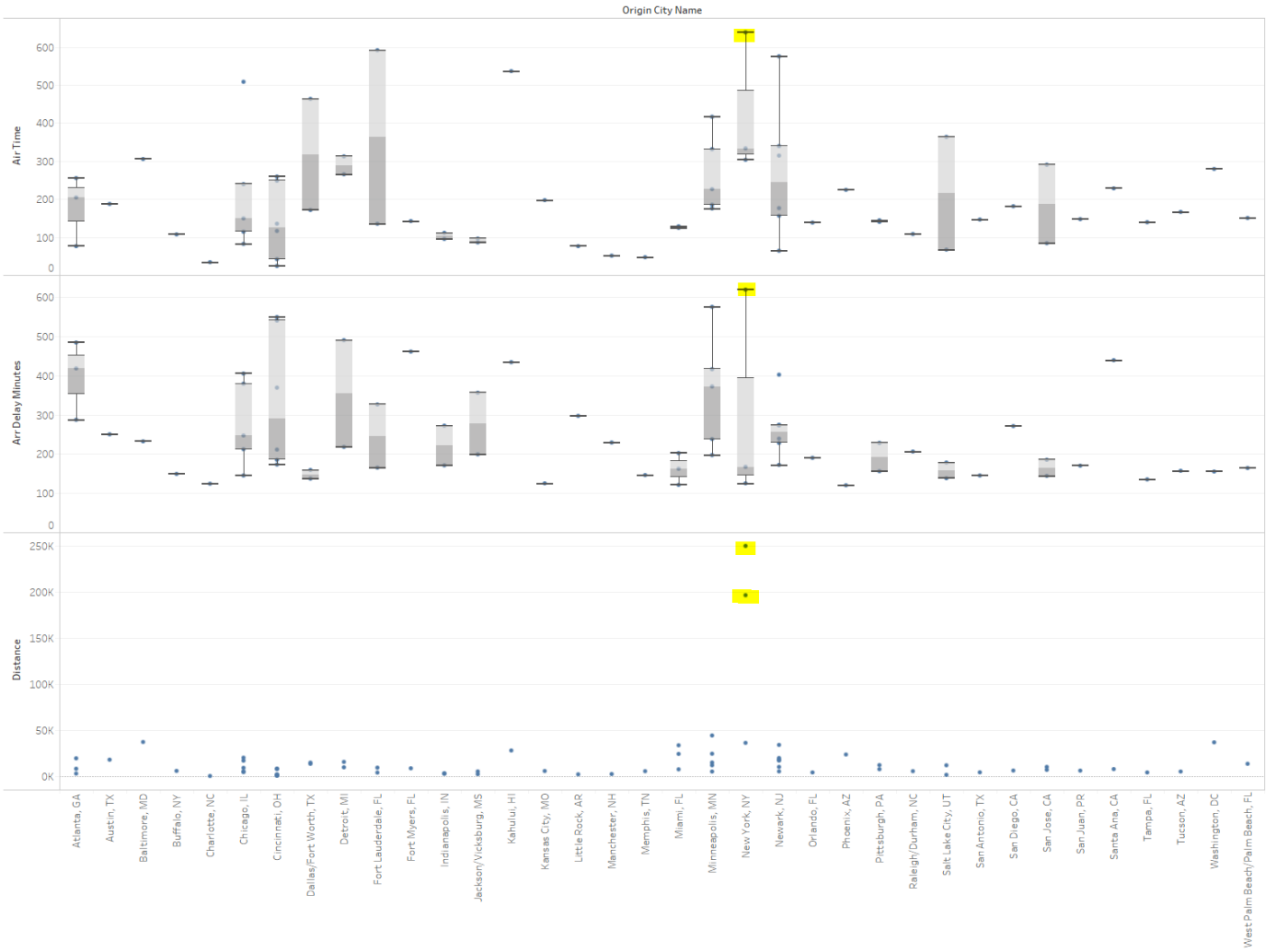
**Hypothesis 1:** There were abnormal activities happened between Sept 11 – 13 in the dataset for specify cities and flights.




**What's informative about this view:** This view above shows flight departure delay in minutes and total number of flights VS the flight origin state. We can see that all states have missing data points between 9.11 and 9.13 2001. This indicates abnormal activity during that time. Pretending we didn't know about 9.11 attack and trying to find out what happened during that time.

**What could be improved about this view:** Add more detail to this view will reveal more information, such as flight company, flight number, etc. Just showing state information isn't enough.

Arr Delay Between City

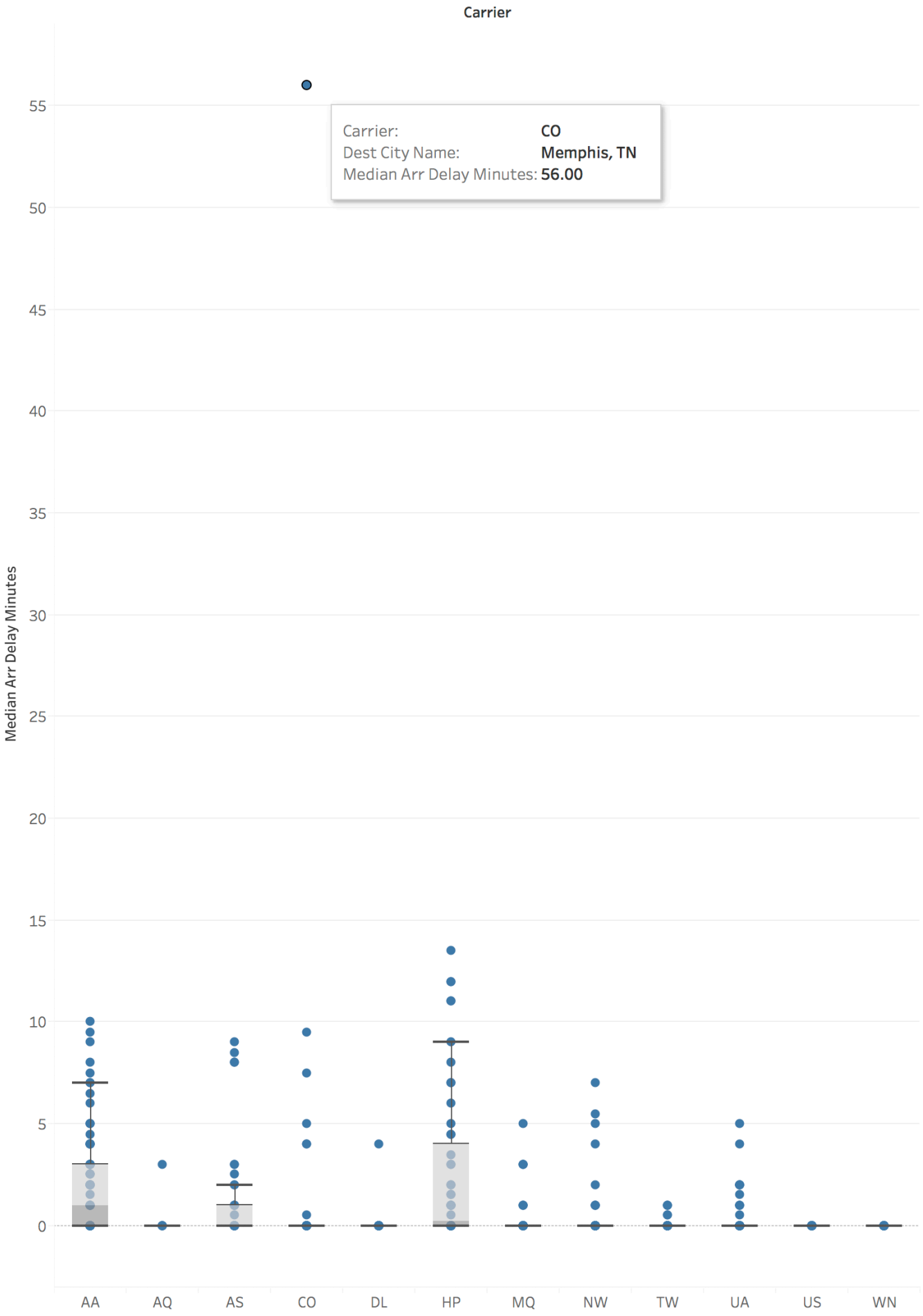


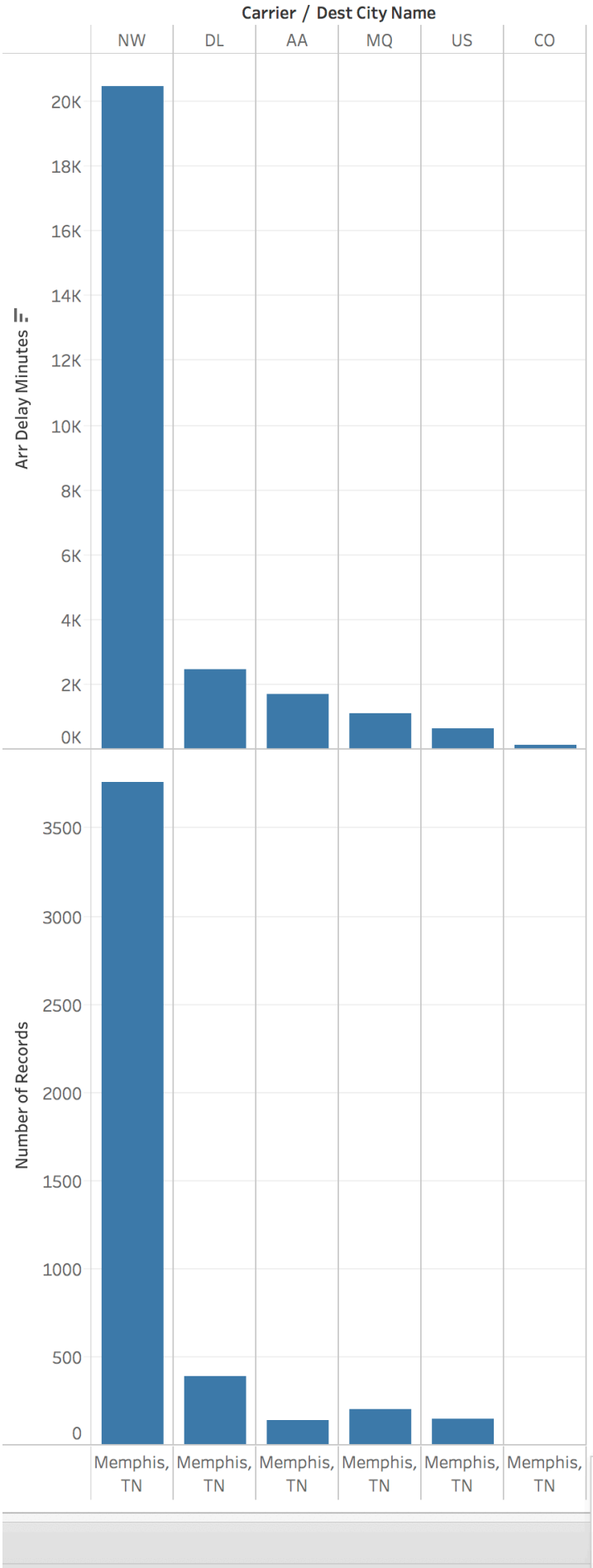

**What's informative about this view:** This view explores the possible air routes that were impacted by this abnormal activity the most. I didn't use airline id because it is not very intuitive to think in terms of id, name of city if much more intuitive. It shows arr delay (difference in minutes between scheduled and actual arrival time) and distance (distance between airports in miles)'s relationship and flight origin/departure city, and found that routes between New York and Los Angeles/San Francisco were experiencing the most delays.

**What could be improved about this view:** It is still difficult to locate the exact flight that were impacted during 911, its not event able to identify the exact city/cities that this event impacged. But data from the flight that were actually impacted appears to be missing. So I'm not sure at this point if this is even possible to find from just this dataset. It could be beneficial to compare data for the same airport across different time, that could reveal more info about impacted city.

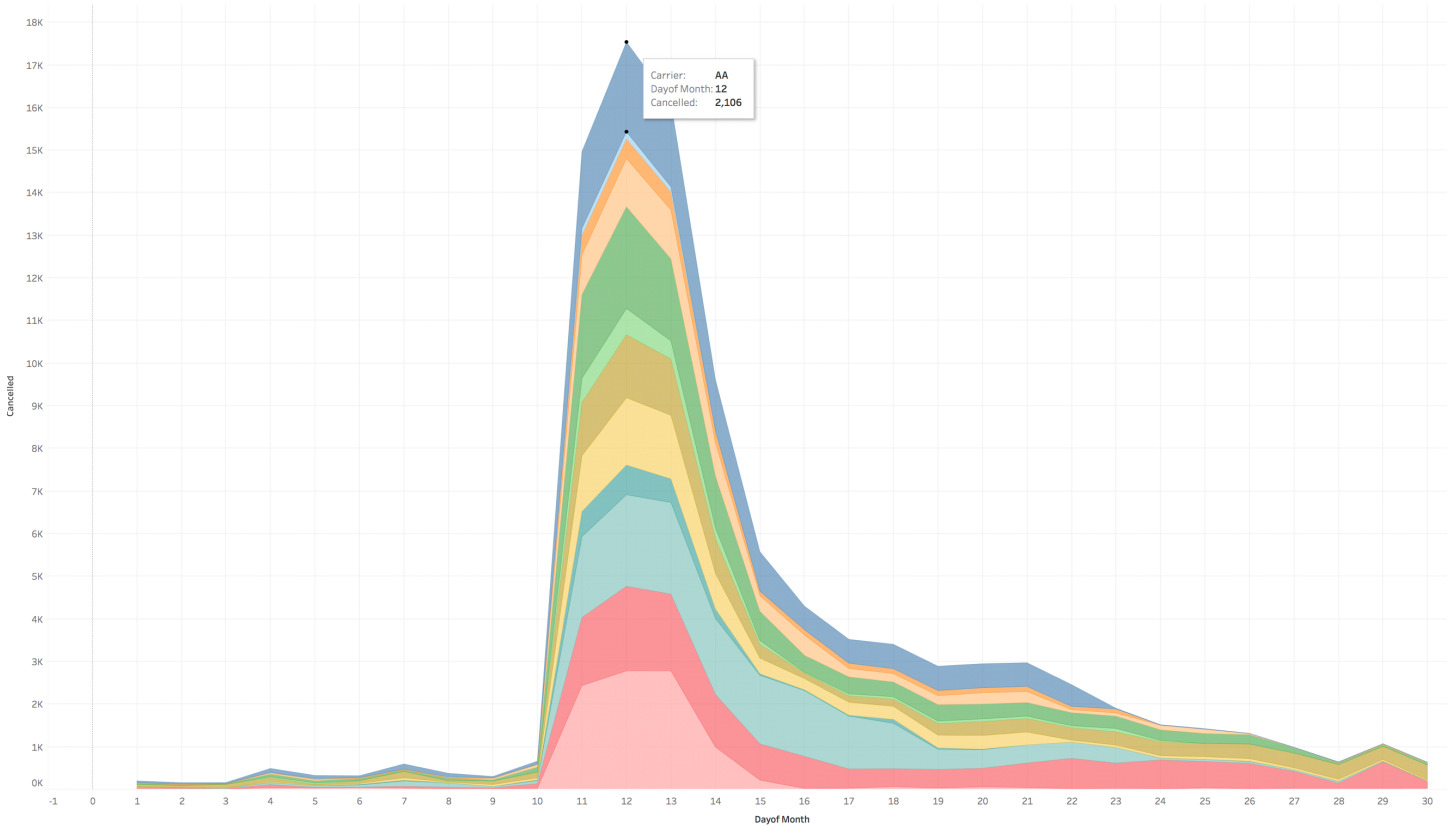
**Conclusion** (do the data appear to support the hypothesis, or not?): These data views suggest that there was clearly something going on during a time frame (9.11-9.13). The pattern of missing data is quite obvious when plotted against time. But in order to identity the exact locations and airlines affected, a better view to show outliers is needed. Different types of view make it more obvious to show different types of data, may it be trend, distribution, outlier etc., it is very important to choose the right type of visual to convey the right message.

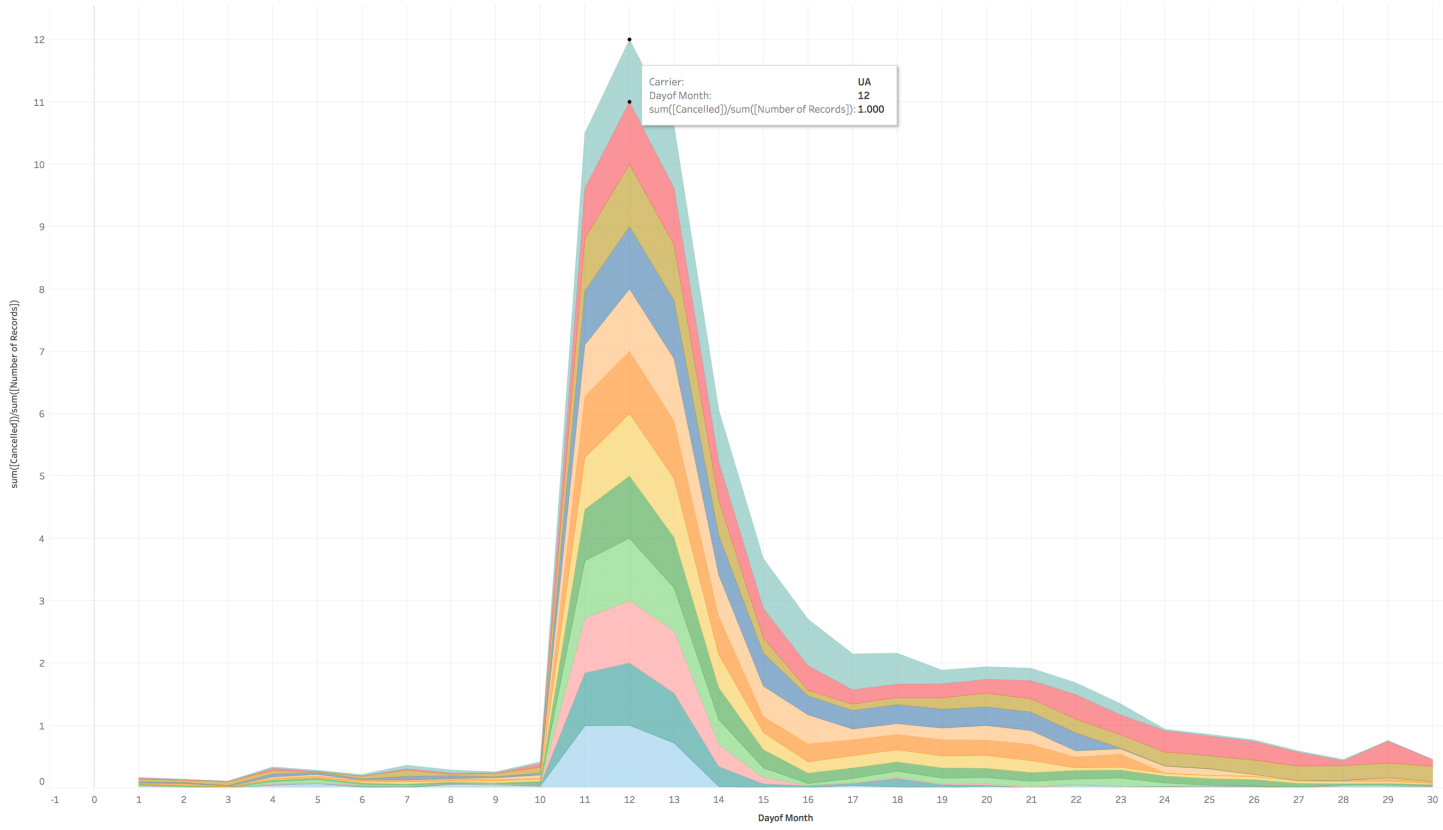
**Hypothesis 2:** Significant delay might due to data skewness or sparseness rather than actual difference.


Carrier
Carrier: CO
Dest City Name: Memphis, TN
Median Arr Delay Minutes: 56.00
Median Arr Delay Minutes
55
50
45
40
35
30
25
20
15
10
5
0
AA
AQ
AS
CO
DL
HP
MQ
NW
TW
UA
US
WN


Carrier / Dest City Name
NW
DL
AA
MQ
US
CO
Arr Delay Minutes
20K
18K
16K
14K
12K
10K
8K
6K
4K
2K
0K
Number of Records
3500
3000
2500
2000
1500
1000
500
0
Memphis, TN
Memphis, TN
Memphis, TN
Memphis, TN
Memphis, TN
Memphis, TN
Carrier: CO
Dest City Name: Memphis, TN
Arr Delay Minutes: 122
Number of Records: 3

**Hypothesis 3:** 911 flights, UA (United Airlines) and AA (American Airlines), got affected heavily in terms of number of flight got canceled and ratio of canceled flight against all flights operated, compare to flights that are not directly involved in 911.



***PLEASE NOTE THAT THE TEMPLATE IS NEEDED FOR ONLY THE FIRST HYPOTHESIS. THE OTHER TWO ONLY NEED TO INCLUDE THE HYPOTHESIS AND THE FIRST AND LAST VIEWS CREATED (NO ADDITIONAL TEXT NEEDED).***