Table of Contents

Abstract

As more and more people travel around the world by planes, whether the airlines have an on-time performance becomes a key factor that customers concerned about. In this paper, we applied multiple big data applications to analyze the flights cancellation, delay and amount trends as well as the network between cities and airports. We also built an elastic net regression to predict flights' delay and obtained an accuracy of 70.6% and 73.1%. With diverse visualization applied, we found some interesting historical events such as blizzards and terrorist attacks. In the result part, we draw some conclusions and made suggestions for the best time, cities, airports, carriers to choose in order to avoid flights delay or cancellations.

In our analysis, multiple tools were used to make a more diverse analysis as well as improve our understanding and skills to those big data applications. The following table is a summary of tools we used:

Actions	**Clean data**	**Merge datasets**	**Descriptive statistics**	**Data visualization**
tools	**Linux**	**Map join (Hive)**	**Hadoop, Pig(Macro), Hive**	**R**

1. Introduction

As more and more people travel around the world by planes, whether the airlines have an on-time performance becomes a key factor that customers concerned about. Many factors especially weather conditions would have a great influence on the performance of flights. The development of computer science and statistics makes it possible to store a lot of data information, conduct analysis to figure out the factors lead to flights delay or cancellation as well as dig some patterns shown by the data. Therefore, in this article, we will have a deep analysis for the factors that influence flights cancellation, flights delay and number of flights.

Our data is the airlines data set from the US Department of transportation's Bureau of Transportation Statistics (BTS)[1]. The years we are focusing on are 1999 and 2000 and each record represent one flight. The variables contain information about the time of the flights, the basic information about the flights, the delay and cancelled information. Three more small data sets are provided with information about the carrier, airports and the plane itself.

2. Data Processing

2.1 Data Set Cleaning

In order to join the distinct datasets in hive we had to process the data because some fields that we wanted to join contained quotes around the values in some datasets but not others. Trying to join the data sets without resolving this would result in no matching fields. To overcome this, we wrote a linux script that would remove all quotation marks from each data set. In addition, we removed the header of every data set so hive doesn't import the header as the first row in a table.

2.2 Data processing

The airline.csv was the largest at just over 1 GB while the other 3 were under 500kb, so we used mapjoin to combine all the data sets into 1 table because mapjoin is extremely efficient when one table fits in memory. Data Processing was also done in Pig and Hadoop to understand basic summary statistics of the data.

---

[1] Data and Statistics | Bureau of Transportation Statistics. (n.d.) Retrieved May 04,2016,from http://www.rita.dot.gov/bts/data_and_statistics/index.html

2.3 New Variables Generating

In order to run regression certain features were changed to one-hot vectors. What we did, say, for the month variable, is create 12 new columns (one for each month) and each data point would contain a 1 in the column that corresponds to the month value and 0 everywhere else. We did this for all features that should be treated categorically as opposed to numerically. The only feature that we treated as a continuous variable was distance.

In addition, we generated a new binary variable for arrival delay, where 1 indicates an arrival delay and 0 indicates on time. Here we used the definition given by FAA that "arrival delay is 15 minutes or more".

2.4 Descriptive Statistics Analysis

In this section, we calculate basic descriptive statistics from our data set and draw some interesting conclusions.

Best Time To Travel

We used hive query to generate the average delay time for 24 hours over two years. We found that the hour with largest delays was 12 am, and that the average delay is about 58 minutes. 12 am flights are usually referred to as "RedEye" flights. The average delay time from 5 PM - 8 PM was fairly large--it was about 35 minutes. 5AM - 8AM incur the smallest length of delay, which is about 20 minutes' delay on average. Therefore, we can give a recommended departure time between 5AM - 8AM to avoid the delay, and furthermore, be prepared for a delay if departure time is between 5 PM - 8 PM.

3. Trend Investigation and Analysis

3.1 Cancellation Trends Analysis

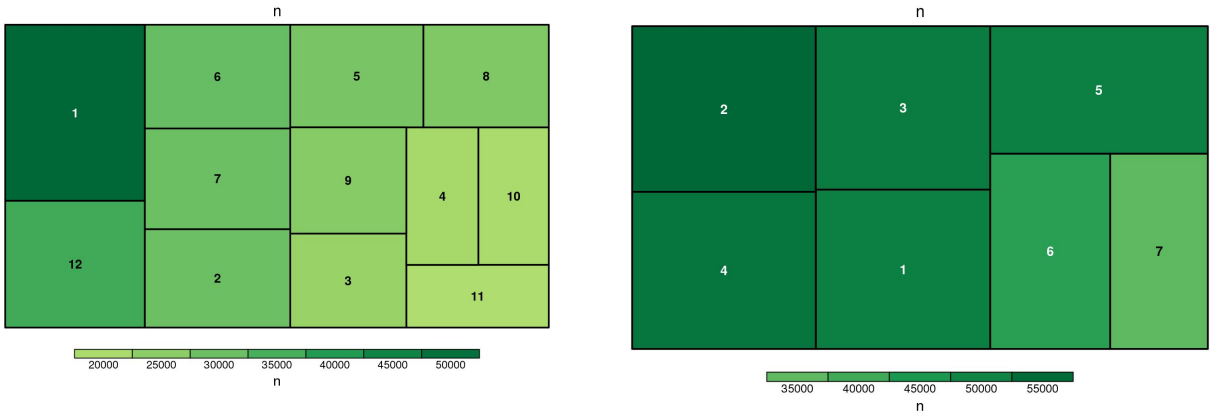
3.1.1 Average Number of Cancellations

To being the cancellation trend analysis, we created a Hadoop that calculated the average number of flights that are cancelled. We created map and reduce scripts that counted the number of flights that are cancelled and divided that by the total number of flights. These scripts can be seen in the code under the hadoop folder along with the command used to run these scripts. From this dataset, it appears that 3.0488% of all flights are cancelled. This value will be used as a baseline to compare other cancellation trends.

3.1.2 Cancellations by Month and Day of the Week

There are many cancellations trends to notice in this dataset, however, we began our analysis by understanding trends based on the month and day of the week. In order to accomplish this, we created treemaps that indexed the number of cancellations by month (left) and day of the week (right). It appears that most cancellations happen in the month of January followed by the month of December. This makes sense as many major snowstorms happen during these months. I will specifically further explore the months of January 1999, January 2000, and December 2000 as these months had major snowstorms according to historical weather data.

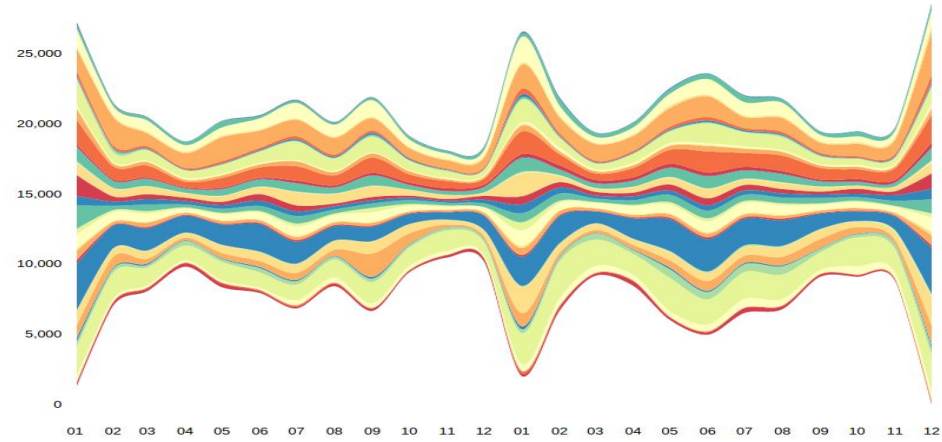
Cancellations by Month (Left) and Day of the Week (Right):



The least number of cancellations happen in the months of March, April, October, and November. From the day of the week treemap, it seems that most delays happen on Mondays through Thursday and the smallest amount of delays happen on Sunday. This trend seems to exist due to the fact that there are more flights on Mondays through Thursday compared to Sunday, so naturally there will be differences in the number of cancellations.

3.1.3 Cancellations by State

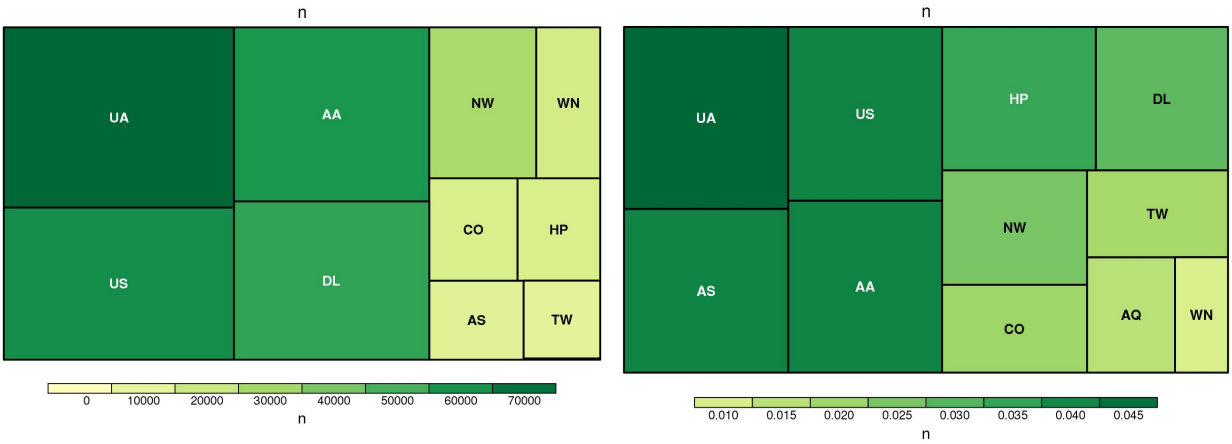
Cancellations by State:



This is a stream graph shows the number of cancellations by flights originating in each state (taking off from an airport in that state). In general the states of Illinois, Georgia, California, New York, and Texas had the largest number of delays. This makes sense as these states house the largest airports in the country so they will have the most cancellations because they have the most flights going in and out of them. It also seems like Northern states (such as Illinois, New York, Massachusetts, Pennsylvania, etc.) have more cancellations during the winter months (especially January and December) when compared to other states.

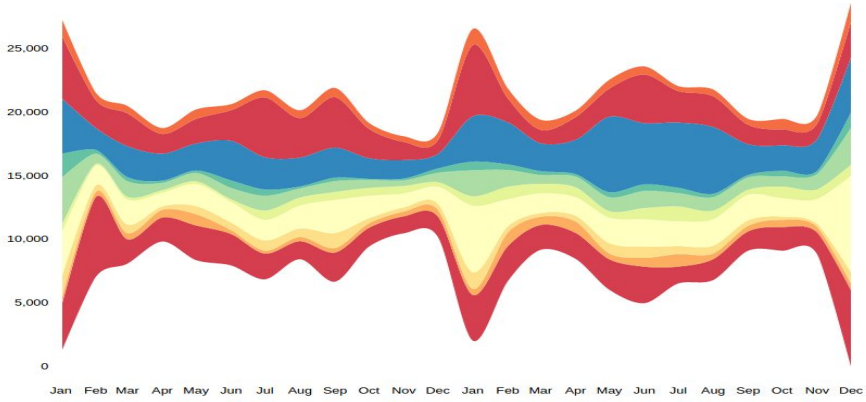
3.1.4 Cancellations by Carrier

Cancellations by Carrier (Left) and Carrier Proportions (Right):



We then shifted our attention and began
looking at cancellation trends for specific carriers. From a hadoop script, we noticed that an average of 3.0488% of total flights were cancelled from 1999 to 2000 in this dataset. We created a treemap that displays the total number of cancellations for 11 carriers (left). The cancellations treemap by Carrier demonstrates that United Airlines (UA) has the most cancellations in this

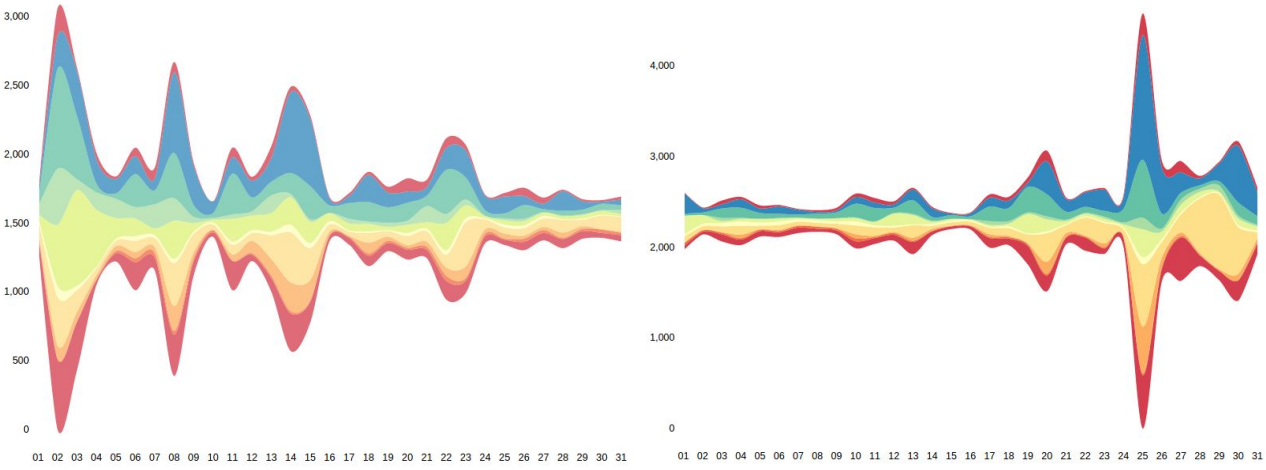
dataset with a total of 71,518 cancellations. Aloha Airlines (AQ) has the least number of cancellations with a total of 173 cancellations between 1999 and 2000. This seems to make sense as there are not many Aloha Airlines flights compared to many of the other major airlines. To gain more insight on a carrier's service and likelihood of cancellation, I also made a treemap on the proportion of cancellations by carrier. United Airlines has the highest proportion of cancelled flights at a rate around 4.086%. SouthWest(WN) has the lowest proportion of cancelled flights at a rate around 0.943% during the years of 1999 and 2000. This seems to suggest that Southwest is more reliable (less likely to have a cancellation) than United Airlines or most of the other airlines listed in this cancellation data set.



We also created a Stream Graph to understand the changes over time in the number of cancellations by each carrier. This Stream Graph seems to confirm many of the facts that we noticed from the months by cancellations treemap. Most cancellations happen in the month of January followed by the month of December as there are large spikes in January 1999, January 2000, and December 2000. I will specifically further explore the months of January 1999, January 2000, and December 2000 next as these month had major snowstorms according to historical weather data. The least number of cancellations happen in the months of April and November. In general, most cancellations are from US Airways, United Airlines,Delta, and American Airlines (huge spikes in cancellations tend to also result from these airlines). There is an interesting increase in the number of cancellations in United Airlines (UA) between May 2000 and August 2000 (much higher number of cancellations than other months). This may be an interesting time for the business as this trend is inexplicable. In general, Southwest (WN), T'Way Air (TW), Northwest Airlines (NW), Hawaii Pacific Airlines (HP), and Alaska Airlines (AS) all seem to have relatively the same proportion of cancellations throughout the entire data set with very few small spikes.

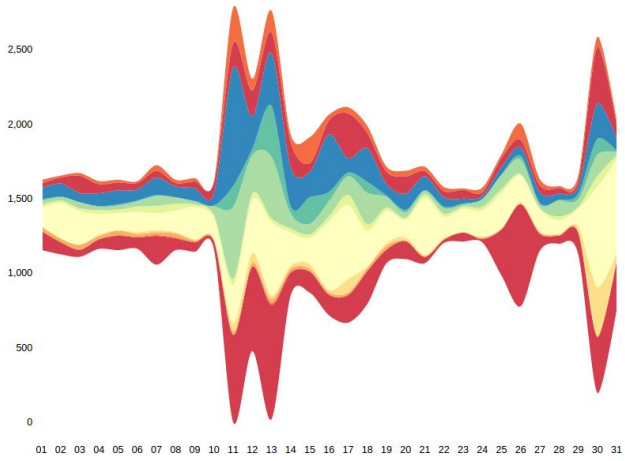
3.1.5 Case Studies (January 1999, January 2000, and December 2000)

After noticing significant spikes in the number of cancellations during January 1999, January 2000, and December 2000 from the previous Stream Graph, we decided to investigate these months more closely. We created Stream Graphs for each of these months which show the total number of cancellations per carrier for every day that month.



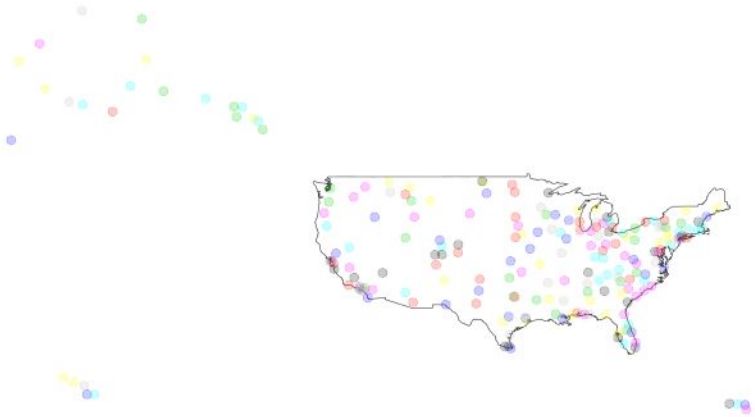
During the month of January 1999 (left), there is a huge spike in the number of cancellations on January 2nd. There were many cancellations from American Airlines (508), United Airlines (462), NorthWest Airlines (446), Delta (351), and US Airways (236) on this day. By looking at historical weather data, I notice that there was a major snowstorm that struck the American Midwest (Milwaukee and Chicago). There are also smaller spikes on January 8th, 14th, and 22nd. Historical data demonstrates that there was a major snowstorm that struck the New England area on January 14th (Toronto needed military assistance to remove snow).

The Stream Graph of January 2000 reveals that there is a huge spike in the number of delays on January 25th, 2000 (right). 1377 US Airways flights were cancelled, 689 Delta flights, and 586 American Airlines flights were cancelled on this particular day. There was probably a very large snow storm on this day. According to a Wikipedia article, there was a particular powerful and surprise snowstorm that struck North Carolina and Virginia on this day. There was over 20 inches of snow at the airport at Raleigh, NC alone probably causing many of the cancellations notes in this stream graph. We can also see this in the Stream graph that shows number of cancellations by state and month. During January of 2000 there were 1595 flights cancelled from North Carolina and 1839 flights were cancelled from Virginia, both of which are much higher than usual for those states during other months.



The Stream Graph of December 2000 reveals a huge spike in the number of cancellations on December 11th-13th, 2000 and an additional large spike on the 30th of December. There is not much of an explanation for the large spike from december 11th to the 13th, however, there is a significant increase in the number of cancellations from United Airlines on December 11th, with a total of 797 cancellations on that day (much larger than the rest of the month). American Airlines also had a large number of cancellations from December 11th to 13th, with 587, 567, and 767 cancellations on each of those days respectively. A large nor'easter struck the east coast of the United States on December 30th 2000 causing many cancellations at airports. Delta (DL) had a particularly large number of cancellations (672 cancellations) on December 30th.

3.1.6 US Map of Cancellations by Airport Origins



This map shows the cancellations for each airport on the map of the United States. The deeper colors demonstrate higher numbers of cancellations for that specific airport. It appears as though some areas of the country (the more populous areas), have higher numbers of cancellations compared to other airports. For example, Chicago and Atlanta are two airports with the largest number of cancellations.

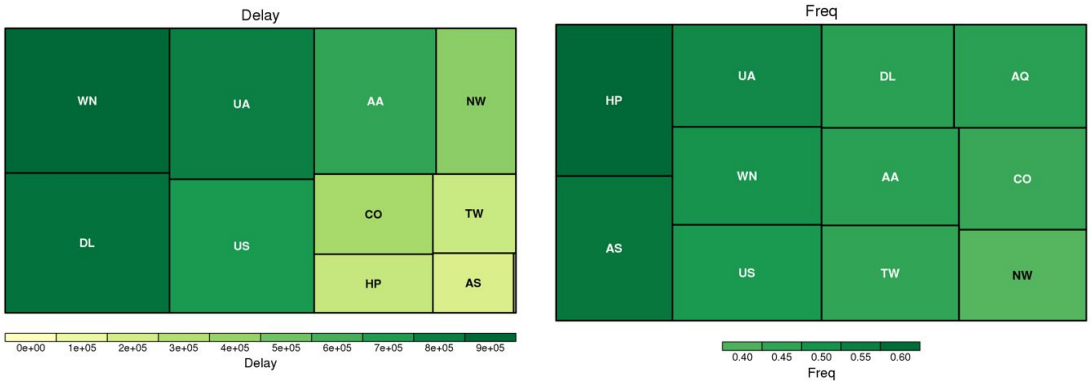
3.2 Delay Trends Analysis

3.2.1 Delay by Unique Carrier

Flights Delay Rate

For the two years in all, the left figure showed that Southwest Airlines(WN), Delta Air

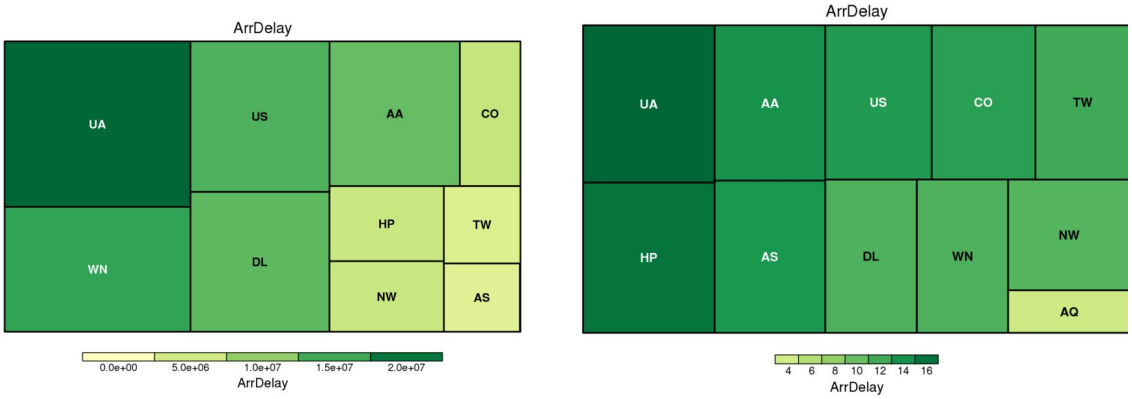
Lines (DL) and United Airlines (UA) ranked the top three delay numbers while Alaska Airlines (AS), Trans World Airways (TW), America West Airlines (HP) ranked the last three. But if divided by each carrier's total flights number, which means the delay rate for each carrier, as shown by the right figure America West Airlines (HP) and Alaska Airlines (AS) will be the most higher two, and Northwest Airlines (NW) will be the least one. And no big differences existed between others. The median of delay rate was 46.61% and the mean of delay rate was 48.66%. Therefore, the reason why Southwest Airlines(WN), Delta Air Lines (DL) and United Airlines (UA) had a higher delay numbers was that they had more flights than other carriers. From the basic statistics we generate from pig macro script we found that, Southwest Airlines(WN) had 1.7 million flights, Delta Air Lines (DL) had 1.7 million flights, United Airlines (UA) and US Airways (US) had 1.4 million flights.



As for 1999 and 2000, America West Airlines (HP) and Alaska Airlines (AS) were still the top two carriers of delay rate, and Northwest Airlines (NW) is still the last one. Therefore, America West Airlines (HP) and Alaska Airlines (AS) had a higher delay rate about 60%, and Southwest Airlines(WN) had a lower delay rate about 40%. And the other carriers' delay is between 45% to 50%.

Flights Delay Time

For two years in all, United Airlines (UA), Southwest Airlines(WN), US Airways (US) will be the most higher two, and Alaska Airlines (AS) will be the lowest one. But if the total delay time is divided by each carrier's total flights number, which means the average delay time (minutes) for each carrier, then we will get different patterns. The left figure below showed that United Airlines (UA), America West Airlines (HP) will be the most higher two, and Aloha Airlines (AQ) will be the least one. And no big differences between the others. The average delay time for all the carriers was 12.56 minutes and the median was 13.14 minutes. Therefore, this result showed again that United Airlines (UA), Southwest Airlines(WN), US Airways (US) were the biggest carriers with more flights for 1999-2000.



As for individual years, United Airlines (UA), America West Airlines (HP) and Alaska Airlines (AS) still ranked the top ones which were around 15 minutes, and Northwest Airlines (NW), Aloha Airlines (AQ) were still the last carriers which were around 10 minutes delay on average. But the differences were only no more than five minutes.

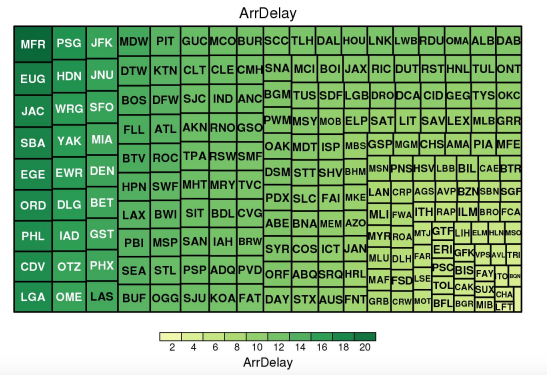
Therefore, for carriers, Southwest Airlines(WN), Delta Air Lines (DL), United Air Lines(UA) and US Airways (US) had more flights in 1999-2000. In addition, America West Airlines (HP), Alaska Airlines (AS) were more easy to have higher delay rate and delay time while Northwest Airlines (NW) tended to have both lower delay rates and delay time.

3.2.2 Delay By Origin Airport

As a passenger or tourist, sometimes, we are not able to choose the destination but we are more free to pick which city to departure. Therefore, our research put more emphasis on the origin airports instead of destination airports.

Flights Delay Rate By Airports

O'Hare International Airport (ORD), Hartsfield-Jackson Atlanta International Airport (ATL), Dallas/Fort Worth International Airport(DFW) ranked the first three. But if divided by each airport's total flights number, which means the delay rate for each airport, then we will get different patterns. From the left graph we can see that: Gustavus (GST), McCarran International Airport (LAS), Kerle K Smith Airport (CDV) will be the most higher three, and Lafayette (LFT), Minot Afb Airport (MIB), Gulfport-Biloxi International Airport (GPT) will be the least three. Gustavus (GST) had an extremely high delay rate – 41%, which was almost 10% higher than the second one, McCarran International Airport (LAS), with30%. The mean of origin airports' delay rate is 41.74% and the median is 43.07%.



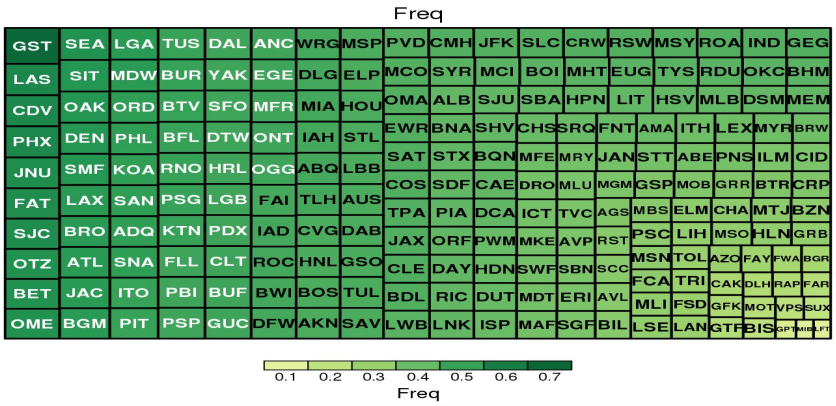
As for the comparison between two years, we found that Lafayette (LFT) is the one changed a lot. Lafayette (LFT) had about 50% delay rate in 1999, while no delays in year 2000. Gustavus (GST) had a delay rate of 41% in 1999 but 18.18% in 2000. There were some airports that have a robust performance during the two year: Long Island Macarthur Airport (ISP), Pittsburgh International Airport (PIT), Ted Stevens Anchorage International Airport (ANC), Miami International Airport (MIA), Bill And Hillary Clinton National/Adams Fi Airport (LIT) and Hector International Airport (FAR) had almost no changes in two years.

Furthermore, let us see the performance of some popular airports which people might care more about. The detailed information was listed below:

IATA	Airports Name	City	Delay in 1999	Delay in 2000	Average Delay Rate
ATL	Hartsfield-Jackson Atlanta International Airport	Atlanta	24.19%	23.68%	23.89%
LAX	Los Angeles International Airport	Los Angeles	22.14%	27.11%	24.63%
ORD	O'Hare International Airport	Chicago	26.03%	33.30%	29.67%
DFW	Dallas/Fort Worth International Airport	Dallas/Fort Worth	21.3%	22.96%	22.13%
JFK	John F. Kennedy International Airport	New York	21.84%	27.41%	24.63%
LGA	LaGuardia Airport	New York	27.63%	33.40%	30.51%

Average Delay Time By Origin Airports

As for the average delay time (minutes) for each airport,we can tell from the right figure that Rogue Valley International Medford Airport (MFR), Eugene Airport (EUG), Jackson Hole Airport (JAC), Santa Barbara Municipal Airport (SBA), Eagle County Regional Airport (EGE) will be the highest delay time airports with almost 20 minutes' delay; Gulfport-Biloxi International Airport (GPT), Lafayette (LFT), Chattanooga Metropolitan Airport (CHA), Aguadilla Airport, Rafael Hernandez Airport (BQN), Hilo International Airport (ITO) will be the lowest delay time airports with no more than 5 minutes' delay. The mean of origin airports' Arrival Delay time was 10.62 minutes and the median also was 10.62 minutes. As for the comparison between two years, we found that Lafayette (LFT) is the one changed a lot. Lafayette (LFT) had about 10 minutes delay in 1999, while more than 10 minutes early arrival for year 2000. There were some airports that have a robust performance during the two year: Merle K Smith Airport (CDV), Yakutat (YAK), Jackson Hole Airport (JAC) had around 15 minutes delay time for both year, Yampa Valley Airport (HDN), Miami International Airport (MIA), Detroit Metropolitan Wayne County Airport (DTW), Hartsfield-Jackson Atlanta International Airport (ATL), Newark Liberty International Airport (EWR) and Pittsburgh International Airport (PIT) had around 5 minutes delay time for two years.



In addition, airports like Gulfport-Biloxi International Airport (GPT), Minot International Airport (MOT), Sioux Falls Regional Airport (SUX), Great Falls International Airport (GTF), Grand Forks International Airport (GFK), Hector International Airport (FAR) and Bismarck Airport (BIS) had early arrival for both two years. Therefore, these airports might be a better choice for travel.

Furthermore, let us see the performance of some popular airports which people might care more about. The detailed information was listed below:

IATA	Airports Name	City	Delay in 1999	Delay in 2000	Average delay rate
ATL	Hartsfield-Jackson Atlanta International Airport	Atlanta	10.37	10.34	10.30

LAX	Los Angeles International Airport	Los Angeles	8.68	11.94	10.31
ORD	O'Hare International Airport	Chicago	12.29	17.67	14.98
DFW	Dallas/Fort Worth International Airport	Dallas/Fort Worth	8.32	9.98	9.15
JFK	John F. Kennedy International Airport	New York	6.03	11.17	8.6
LGA	LaGuardia Airport	New York	11.76	15.78	13.77

It easy to find out that Chicago O'hare International (ORD), Hartsfield-Jackson Atlanta International Airport (ATL) and LaGuardia Airport (LGA) had a higher average delay time while John F Kennedy International Airport (JFK) and Los Angeles International Airport (LAX) had a pretty good performance.

This is the map of the united state, and different points represent different airports. Here we plot the top 100 delay airports. The color represents the total delay time for each airport. And the nearer to red means a longer delay time while the nearer to blue means a shorter delay time. We can find Chicago O'hare International (ORD) in pink, LaGuardia Airport (LGA) in purple, Orlando International Airport (MCO) also in purple.

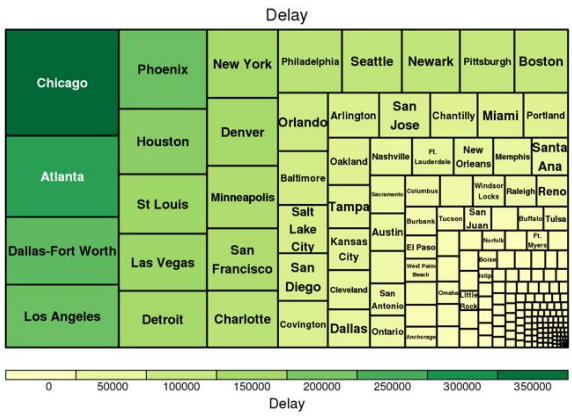
In summary, Lafayette (LFT) and Gustavus (GST) were airports that had big changes in two years which made it hard to predict the delay time and delay rate. There were some airports that have a robust performance during the two year: Long Island Macarthur Airport (ISP), Pittsburgh International Airport (PIT), Ted Stevens Anchorage International Airport (ANC), Miami International Airport (MIA), Bill And Hillary Clinton National/Adams Fi Airport (LIT) and Hector International Airport (FAR) had almost no changes in two years.

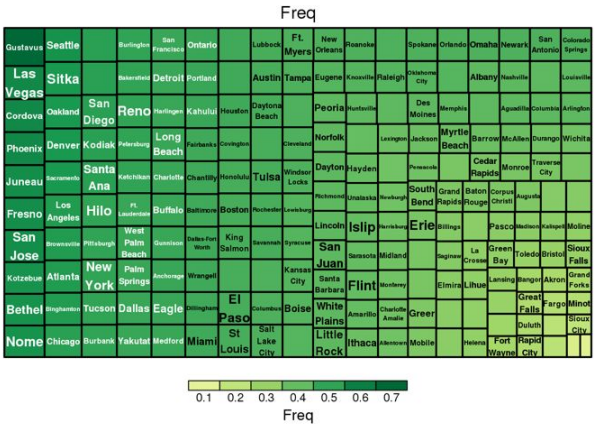
There were some airports that have robust performance during the two year: ISP, PIT, ANC, MIA, LIT and FAR had almost no changes in two years. Merle K Smith Airport (CDV), Yakutat (YAK), Jackson Hole Airport (JAC) had around 15 minutes delay time for both year, Yampa Valley Airport (HDN), Miami International Airport (MIA), Detroit Metropolitan Wayne County Airport (DTW), Hartsfield-Jackson Atlanta International Airport (ATL), Newark Liberty International Airport (EWR) and Pittsburgh International Airport (PIT) had around 5 minutes delay time for two years. In addition, airports like Gulfport-Biloxi International Airport (GPT), Minot International Airport (MOT), Sioux Falls Regional Airport (SUX), Great Falls International Airport (GTF), Grand Forks International Airport (GFK), Hector International Airport (FAR) and Bismarck Airport (BIS) had early arrival for both two years. Therefore, these airports might be a better choice for travel.

3.2.3 Delay by City

Fights Delay Rate

For the two years in all, Chicago, Atlanta, Dallas-Fort Worth, Los Angeles and Phoenix ranked the first. But if divided by each carrier's total flights number, which means the delay rate for each city, Gustavus, Las Vegas, Cordova, Phoenix and Juneau ranked first with more than 50 percent delay rates. Thereto, Gustavus had an extremely high delay rate of 66.14%. On the other hand, Lafayette, Gulfport-Biloxi, Sioux City, Valparaiso and Bismarck ranked the last with no more than 20 percent delay rates. And no big differences between the others. The median of delay rate is 42.82% and the mean of delay rate is 41.77%. Therefore, the reason why Chicago, Atlanta, Dallas-Fort Worth, Los Angeles and Phoenix had a higher delay numbers was that they had more flights as big airports.





As for individual year, Gulfport-Biloxi, Sioux City, Valparaiso and Bismarck had a pretty low delay rate, which is about 20%, for both 1999 and 2000. While in year 1999, Lafayette had an extremely high delay rate of 50% and no delay rate in 2000. There were 8 records of flights in Lafayette, in year 2000 and none of them delayed. For top delay cities, Gustavus, Juneau, Cordova, Phoenix and Las Vegas had a high delay rate for both 1999 and 2000. However, Gustavus had an extreme high delay rate in 2000, almost 80%. On average, year 1999 had higher delay rate than year 2000. 5 cities had delay rate over 60% in 1999 while no one happened in

2000.

Go back to the data set to see why there was such a high delay rate for Gustavus in 2000, we found that there were three delays greater than 1 hour and two greater than 2 hours. The three were the same flights, from Juneau to Gustavus, scheduled to departure at 17:50 from Carrier AS77 on June 13, June 30 and July 31.

Some big cities like seattle, Oakland, los angeles, Atlanta, Chicago and San Diego all had delay rate greater than 50 percent. While New York was about 49.66%,46.06% for Boston, 45.60% for st Louis, 44.16% for Orlando and 35.89% for Springfield.

Average Delay Time By Origin Airports

For the two years in all, Chicago, Atlanta, Dallas-Fort Worth, Los Angeles, Phoenix and New York ranked the first. But if divided by each carrier's total flights number, which means the average delay time for each city, Nashville, barrow, newburgh, Portland and chantilly ranked first relatively delay time. Thereto, Nashville had an extremely high delay time of nearly 5000 minutes.

On the other hand, Lafayette, Gulfport-Biloxi, Sioux City, Valparaiso and Bismarck ranked the last with no more than 20 percent delay rates. And no big differences between the others. The median of delay rate is 46.61% and the mean of delay rate is 48.66%. Therefore, the reason why Chicago, Atlanta, Dallas-Fort Worth, Los Angeles and Phoenix had a higher delay numbers was that they had more flights as big airports.

Case Study Of Some Cities

Nashville had the highest average delay time. There were 4559 flights that delayed for over 1 hour, 1404 flights delayed for over 2 hours and 6 flights delayed for more than 12 hours and 4 of them were to Dallas-Fort Worth, Flight AA 1715 / AA 1817, scheduled to departure at late afternoon. And the delay numbers were almost the same for the two years.

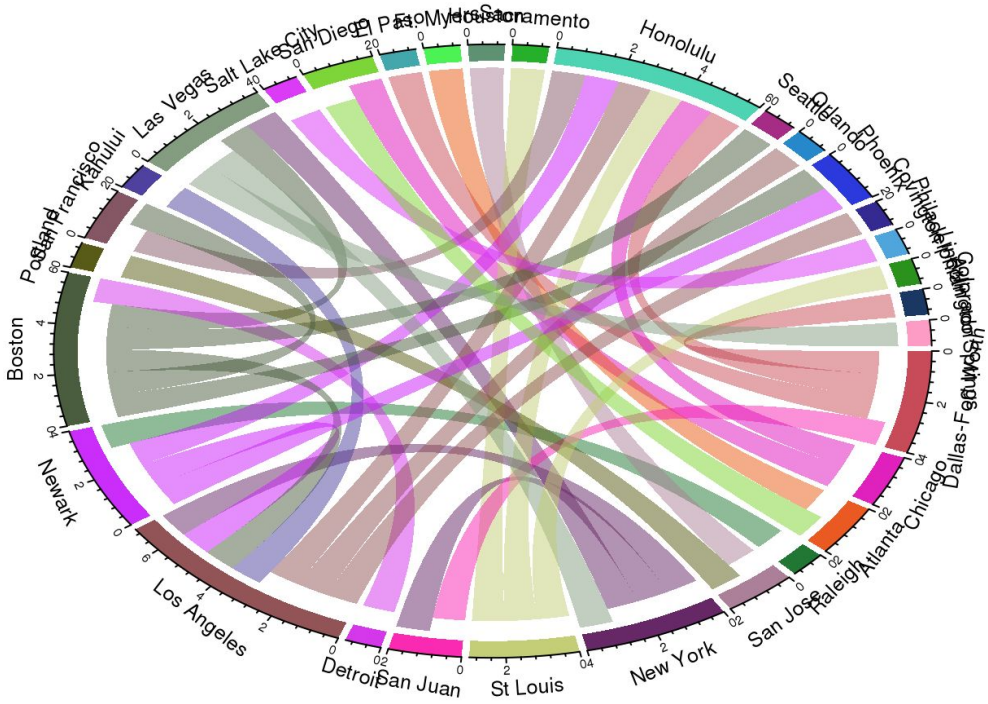
The second extremely high delayed city is Barrow. There were 109 flights that delayed for over 1 hour, 38 flights delayed for over two hours.

Christiansted is the city with lowest delay time. Another interesting conclusion is that, Christiansted has a high early arrival rate which is almost 50% and 5% earlier than 20 minutes. Therefore, other airports can try to learn some operation and management skills from Christiansted.

For 1999, New York had a high arrival delay time, this is almost 1000 minutes on average. There were almost 1000 flights that delay more than 1 hour, 3285 for 2 hours and 11 for 12 hours. For those delay for 12 hours, most were scheduled to departure between 5pm to 8pm and 8 were from the LGA. And four of the flights were in August. The carriers were AA and NW.

Then we got the top 50 delay for the two years for cities. It's easy to figure out that departure

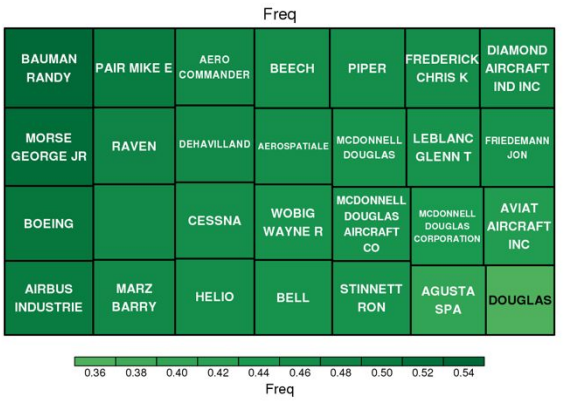
from New York, Dallas-Fort Worth, Newark, Los Angeles, St Louis, Boston and Las Vegas were easier to get a longer delay time while arrival at Honolulu, Los Angeles were easier to get a longer delay time. And the same trend happened to both individual years.

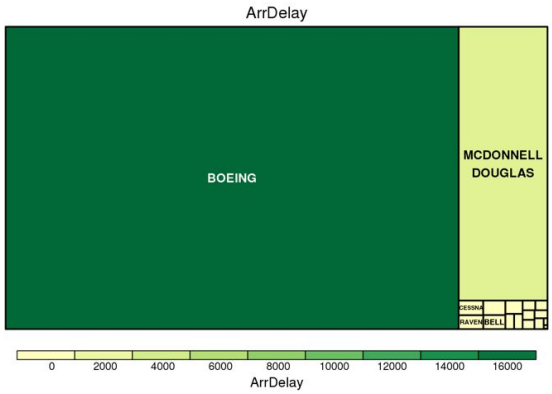


3.2.4 Delay by Manufacturer

Flights Delay Rate

For the two years in total, Boeing was the leading position for the total delay number, followed by Mcdonnell Douglas, Mcdonnell Douglas Aircraft Co and Airbus Industrie. But if divided by each manufacturer's total flights number, which means the delay rate for each manufacturer, bauman randy, morse George FR, Boeing and Airbus Industrie will be the most higher ones with more than 50 percent delay rate, and Douglas will be the least one. And no big differences between the others. The median of delay rate is 45.84% and the range is 17% from 36.56% to 53.98%. Therefore, Boeing had the largest amount of planes but it still had a high delay rate.





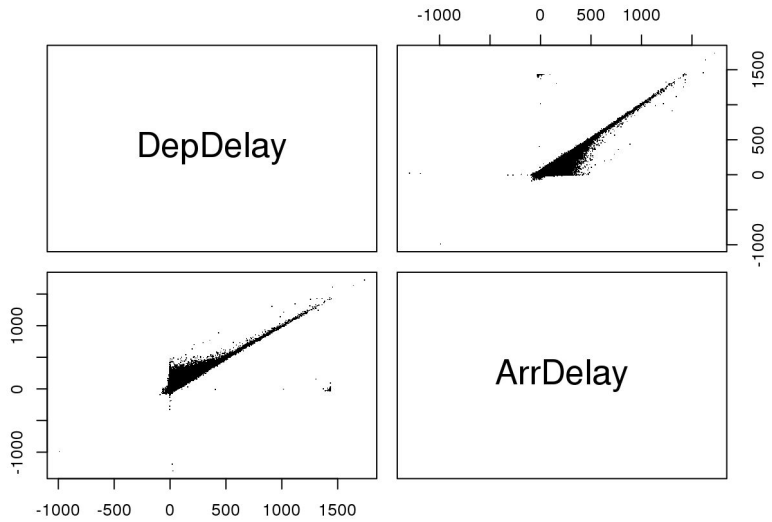
Average Delay Time By Manufacturer

As for average delay time, Boeing, MacDonnell Douglas, Cessna and Raven were at the top levels. And Boeing had a really high delay time, which is almost 5 times to the second manufacturer.
The patterns for both delay rate and delay time were the same for individual year.

3.2.5 Correlations of Numeric variables

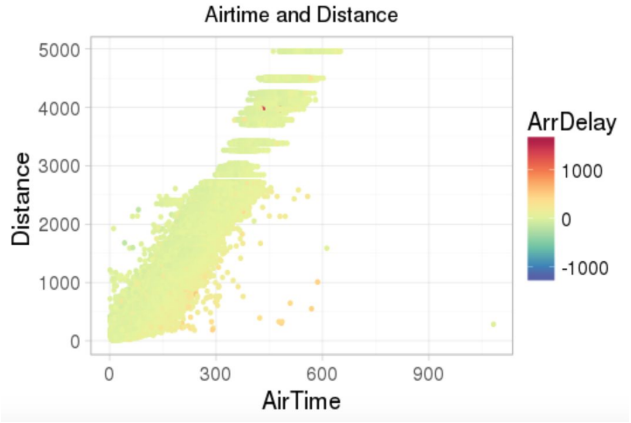
Departure Delay and Arrival Delay

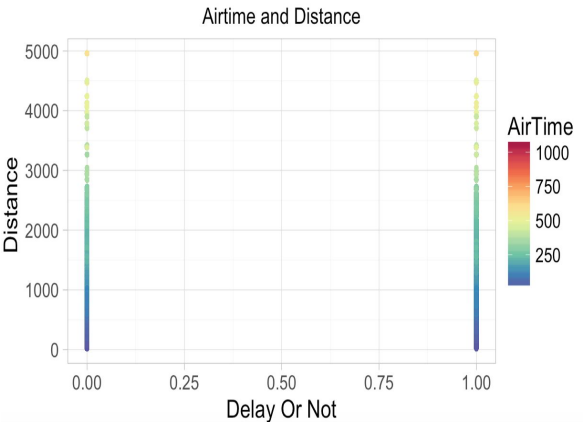
What is the relationship between departure delay and arrival delay? We can tell from the plot that when the departure delay is 0, there were a lot of arrival delay. As departure delay growing, the relationship between departure and arrival were almost linear. Therefore, we can conclude that, many reasons will lead to a short arrival delay but when the arrival delay is extremely high, more than 8 hours, departure delay will be the main reason.



Air time and distance

Then let us have a look at the relationship between distance, airtime as well as arrival delay. From the pairs plot we can tell that there was an obvious linear relationship between distance and airtime. While no obvious pattern for this two variables with arrival delay. This was shown more clear in the second graph. The delay time is almost 0 in all and some high arrival delay happened randomly with orange color. But no patterns show that high distance or air time will lead to a high arr delay time.

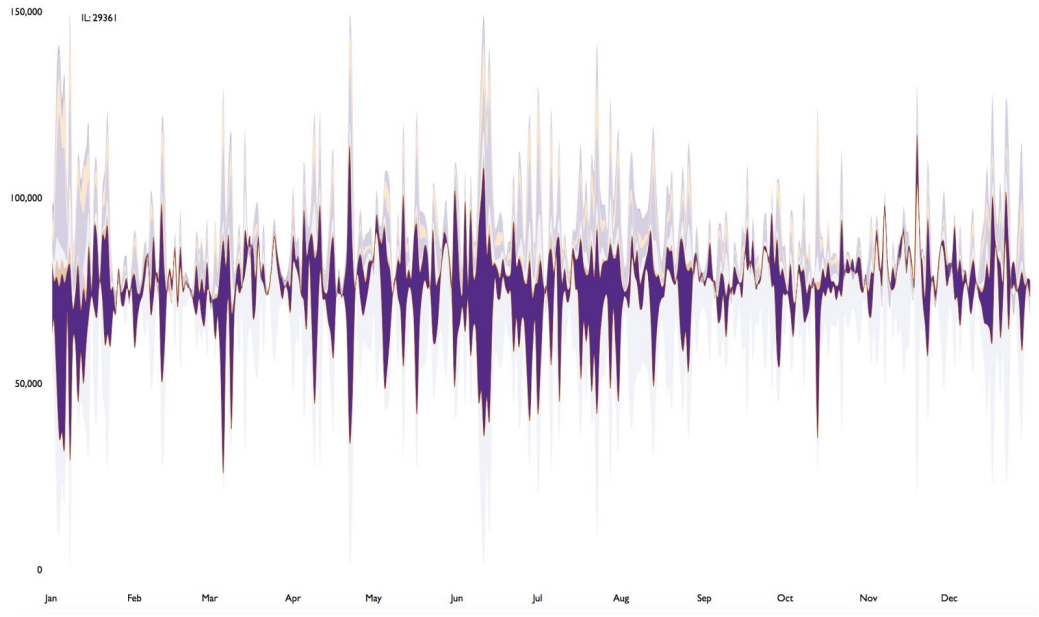




3.2.6 Timeline delay trend

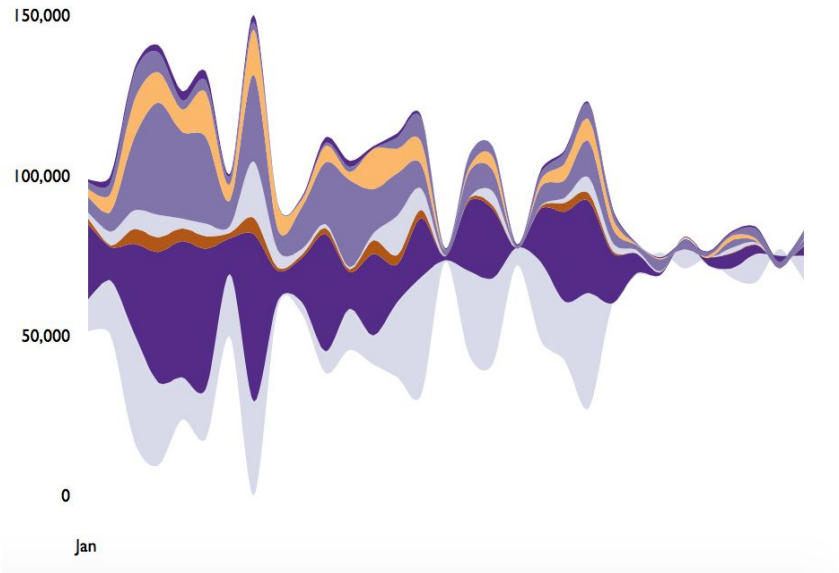
Year 1999

In 1999, we can see several huge spikes at the beginning of January, the beginning of March, late April, the beginning of June and late July. As for delay rate, there were two very obvious peaks for the beginning of January and the others were the same to the total numbers of delays. Then let us have a deep look at what happened in January 1999.



Case Study For January 1999

The graph shows that there was a significant peak at January 2nd - 4th for Midwest states such as Illinois, Michigan, Indiana, Idaho, Washington and Wisconsin. Therefore, we might guess that some big storms or other big events might happened to these states.



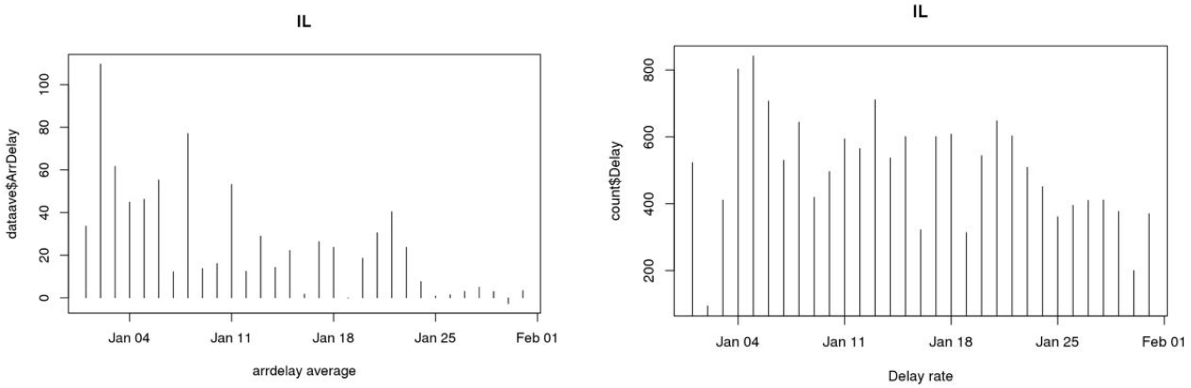
As we go back to the news we found that there was a blizzard called "the blizzard of 1999 Midwestern[2]" struck Iowa, Wisconsin, Illinois, Indiana and Michigan. "The storm produced 22

[2] January 1999 Blizzard. (n.d.). Retrieved May 04, 2016, from

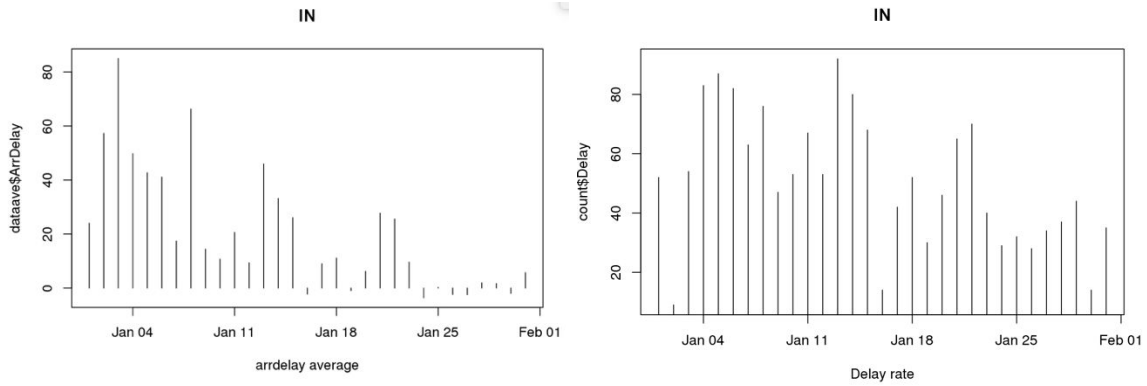
inches of snow in Chicago and was rated by the NWS as the second worst blizzard of the 20th century, ranking behind the blizzard in January 1967. Soon after the snow ended, record low temperatures occurred with values of -20 degrees or lower in parts of Illinois and surrounding states on January 3 and 4."

Then let's have a deeper look at states like Illinois, Indiana individually. The average delay time for Illinois on January 2rd was extremely high however, the delay rate on January 2rd was extremely low but increased quickly on January 3rd and 4th. Therefore, this indicated that on January 2nd, some serious delays happened but most of the flights were cancelled so the delay rate was low which was proved by the cancellation trend we found.

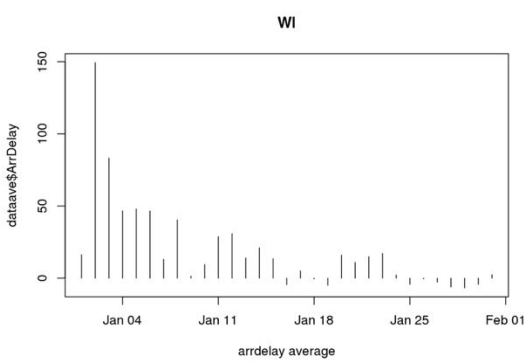
Lines plots for individual state

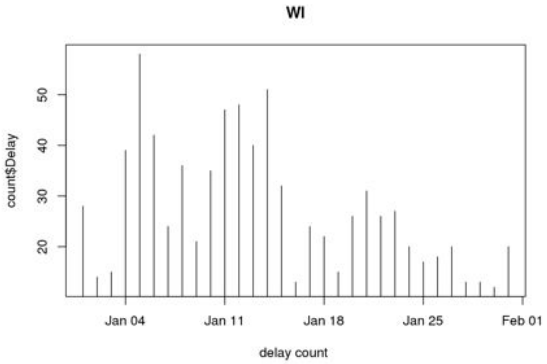


The same pattern happened to Indiana and Wisconsin and WI had a very clear pattern, there was a very high average delay time on January 2rd but a low delay rate and increased quickly on January 3rd - January 5th.

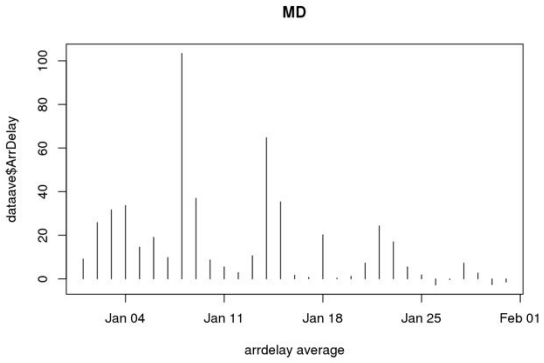


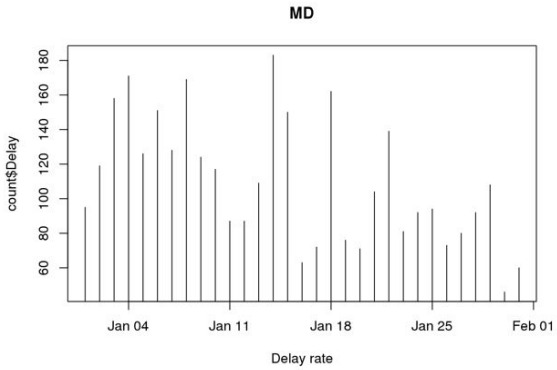
http://www.ncdc.noaa.gov/oa/climate/extremes/1999/january/blizzard





We can also find an interesting pattern in Maryland(MD). Except for the high delay rate during January 2- January 4, there was an extremely high average delay time as well as delay rate on January 8th. And there was also huge spike on January13th for both average delay and delay rate. By further information, we found that an ice storm happened on January 8th - 16th ,1999 Washington, which caused the first peak. And there was a severe ice storm called "The January 1999 North American Ice Storm" [3] struck Washington D.C. on January 14 and 15, 1999. Heavy ice accumulation bringing down power lines resulted in around 745,000 people in the area losing power. Lead to a high canceled rate which was also caught by us in the former section.



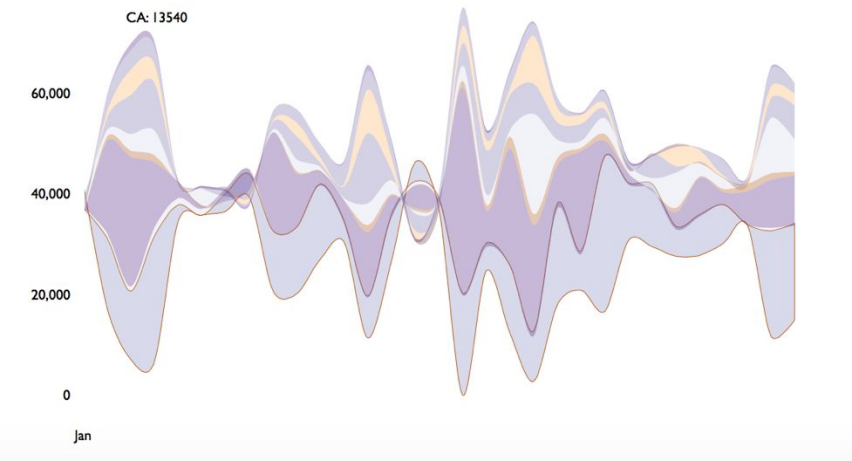


Year 2000

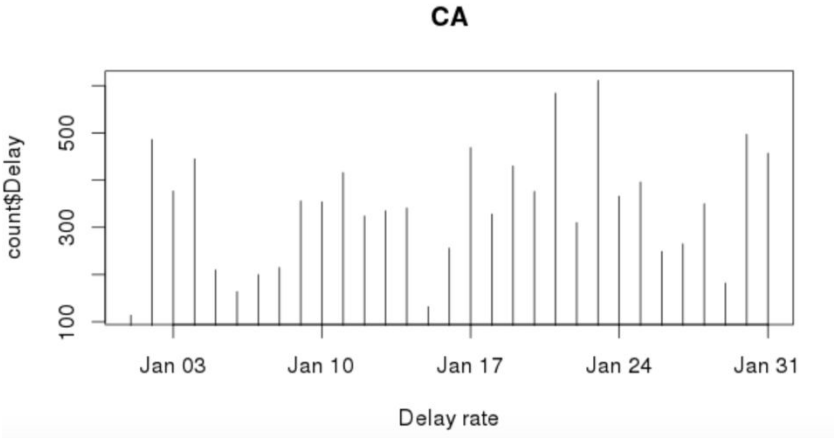
The following graph was average delay time drawn based on different states in 2000. As the streamgraph below showed that there were several spikes at late April, early August, and late December, which worth a deeper investigation.

[3] January 1999 North American ice storm. (n.d.). Retrieved May 04, 2016, from https://en.wikipedia.org/wiki/January_1999_North_American_ice_storm

Therefore, we first got a more detailed plot of January. We found that CA had several increasing points at the beginning of January and the late middle of January. By more background information, we found that the first peaks might be caught by a terrorist --The Year 2000 attack plots, or the Millennium Plot. That were a series of Islamist terrorist attacks that were planned to occur on or near January 1, 2000, with the bombing of four sites in Jordan, the bombing of Los Angeles International Airport (LAX).
2000 millennium attack plots. (n.d.). Retrieved May 04, 2016, from https://en.wikipedia.org/wiki/2000_millennium_attack_plots



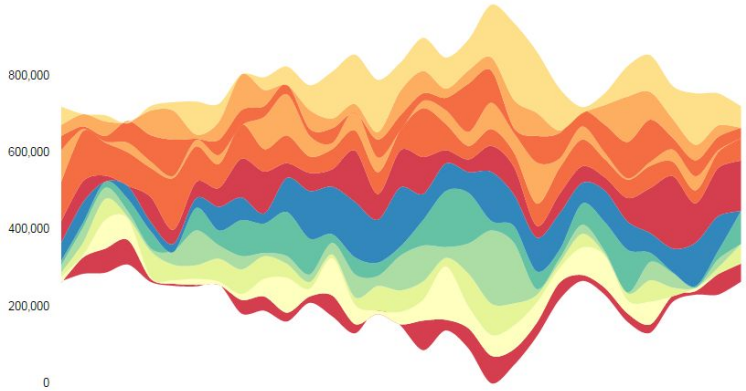
There was also a high peak around January 20th and last for about a week. By more information, we found out that the Winter storms impacted North Carolina on January 18th, 23th , 25th and 30th. The January 25, 2000 storm was preceded by a rather weak storm that dropped up to an inch of snow and some freezing rain across the North Carolina . That is a good explanation for the peaks lasting from January 17 to January 25.



3.2.7 Delay rate by month

From the analysis above, it seems that January trended to have extreme weather like blizzard or ice storm. Therefore, it might have different delay rates or average delay time for each month.

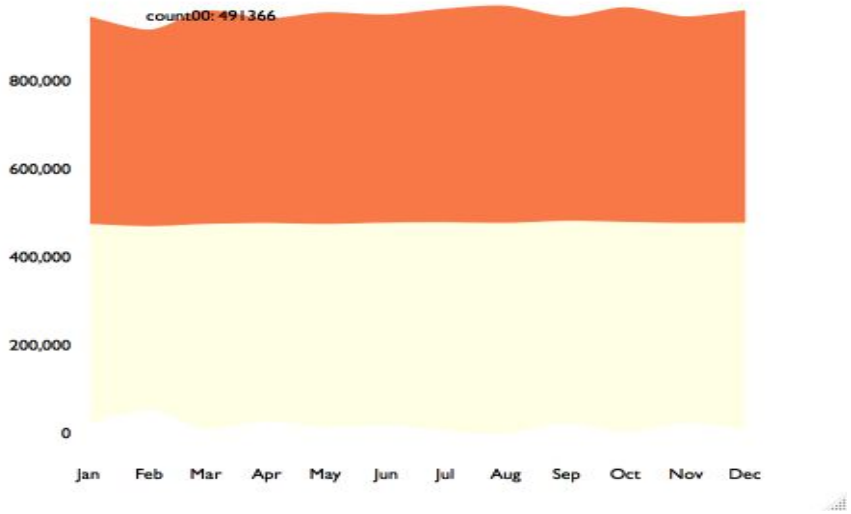
Therefore, we draw a streamplot by each month to see whether a significant difference exist between months.



In the graph above, different 12 colors represent 12 months and the x-axis represented 31 days in a month. We can have a clear look by this plot that,it seems June, december tend to have higher delay time in total. January and November seems to have the lowest delay time .

3.3 Number of flights

In looking for trends in the number of flights, the starting point was to look over the months. The data set was split up by months, and we looked at various flight counts over the months such as the difference in number of flights between 1999 and 2000 and the difference in number of flights per carrier.

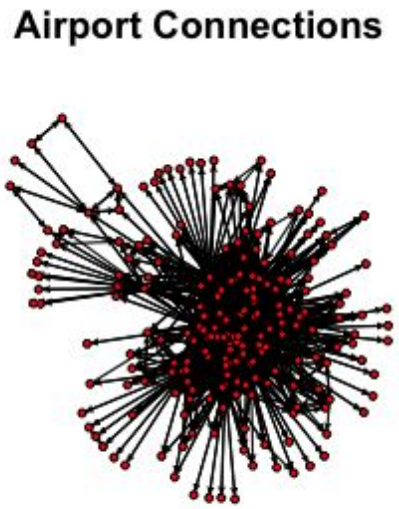


As we can see this stream graph for the total number of flights per month for 1999 and 2000 is not very interesting. It looks like 2000 has slightly more flights overall. One thing to note; however, is that for February, the flights in 1999 has a slightly larger dip than that of 2000. Now, at first glance this dip isn't that exciting since February has the least number of days compared to

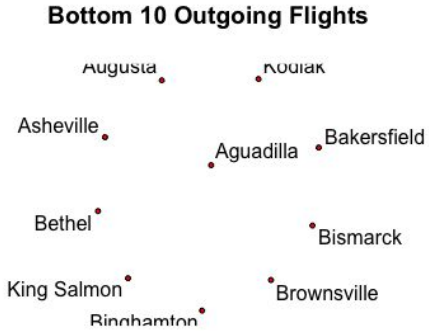
any other month, but if you look closely there is a slightly larger dip in 1999 than 2000. This is due to the fact that 2000 was a leap year, so there was an extra day in February.

3.4 Airport Connections Analysis

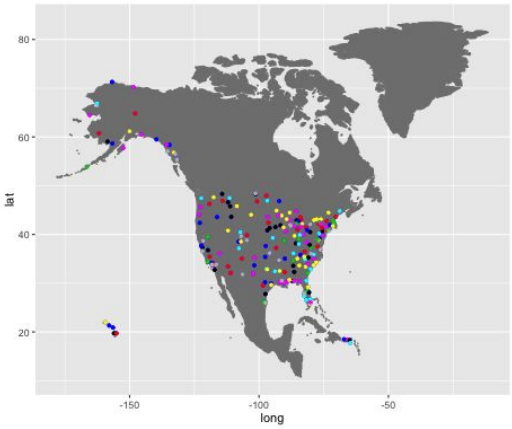
In order to investigate the connectedness of the airports we created a digraph where each vertex, u, is an airport and directed edge e = (u,v) that represents a flight going from airport u to airport v. There are 208 airports in the dataset, and many of them are connected so a full plot of the network is somewhat cluttered; however, we can see there are many vertices on the outside that only go to a single airport. These are likely small airports that only fly to the nearest major city (e.g. Champaign airport only connects to ORD).



We investigate further by looking at the connectedness of the vertices with the most edges. We look at both incoming and outgoing edges. We looked at the top ten incoming and outgoing edges and found that they were both complete graphs, and furthermore, they contained the same set of vertices (airports). That is, from any airport we can fly to any other airport directly in this subgraph. If we look at the bottom 10 vertices (i.e. the airports that fly to the least number of locations) we see that it is a disconnected graph. And plotting both the top ten and bottom 10 together shows that most make connections with the major airports.



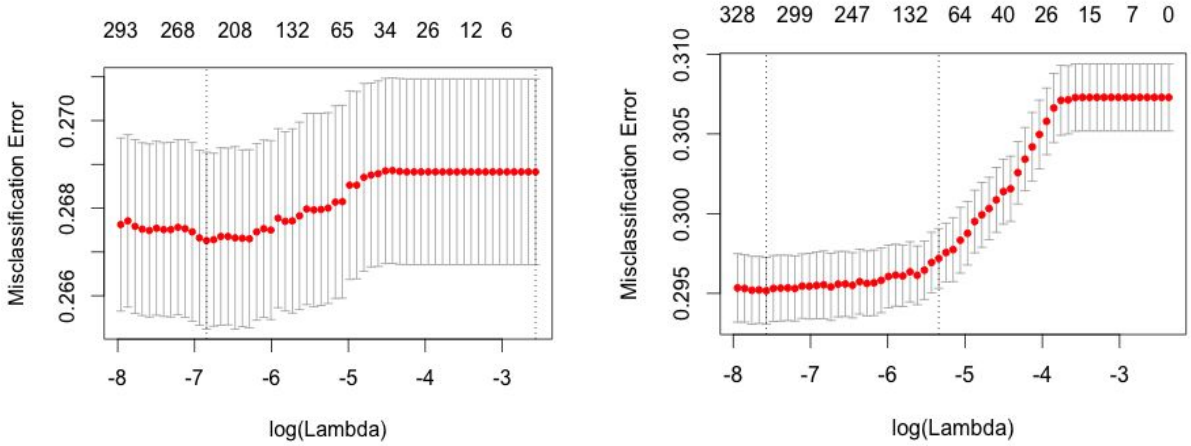
Below is a map of the North America along with the location of each airport. Deeper colored points represent larger connectivity among other airports. Chicago was the airport with the most connectivity, so we have displayed this connectivity on a map. As we can see Chicago (ORD) covers most regions where we have airport data.



3.5 Prediction of Flights Delay

We were interested in classifying delays. We decided to look at only one major airport and one airline. Since ORD had the most connections and United Airlines had the most flights in ORD we choose to look at only United Airlines flights at ORD. We made 2 different classifiers for flights leaving ORD and entering ORD. This reduced our dataset to about 137,000 flights for both cases. We split this into a 70/30 train/test split. In order to predict a delay we created a new binary feature indicating whether or not a delay occurred to the FAA definition of arrival delay which is considers a delay when expected arrival time is more than 15 minutes before the actual arrival time. We then built an elastic net to do the classification. The features that we chose

consisted of flights data: DayOfWeek, Month, CRSDepTime, destination state, and Distance as well as the plane data: aircraft_type, engine_type, plane_year, model, issue_date. All features were converted into one hot vectors except for distance. This greatly increased our number of features from 10 to 2736. Having such large number of features, we decided to use elastic net in order to reduce the dimensionality of our feature space. We used cv.glmnet for this task as it uses 100 different regularization constants and performs 10 fold cross validation so we can choose the best regularization constant for our model. The plot below is a log(lambda) vs misclassification error, where lambda is the 100 different regularization constants. The number on top represents the number of non zero features for each lambda. The best model for for our elastic net regression on incoming ORD flights (plot on the left) used only 237 of the 2736 features and obtained an accuracy of 73.1. We determined the best lambda to be the one with minimum misclassification error (indicated by the first vertical dashed line on the left). Similarly, for outgoing flights we found the model with the lowest misclassification error during cross validation used 314 features, and received an accuracy of 70.6 on the held out test data. Plot is shown on the right.



4. Results

4.1 Best time to travel

From the former analysis we can conclude that, January and December had the highest cancellation while March, April, October and November had the lowest cancellation. While June and December had higher delay time, January and November had the lowest delay time. The average delay time from 5 pm - 8pm was pretty high, it is about 35 minutes. The lowest delay time is 5am-8am, which is about 20 minutes' delay.

Therefore, we can give suggestions that the best travel time will be November 5am-8am, and the worst travel time will be December 5 pm - 8pm.

4.2 Reliable Carrier

United Airlines has the highest proportion of cancelled flights at a rate around 4.086%. SouthWest (WN) has the lowest proportion of cancelled flights at a rate around 0.943% during the years of 1999 and 2000. As for delay, In addition, America West Airlines (HP), Alaska Airlines (AS) were more easy to have higher delay rate and delay time while Northwest Airlines (NW) tended to have both lower delay rates and delay time.

Therefore, SouthWest (WN) and Northwest Airlines (NW) are the most reliable carriers while United Airlines(UA) and America West Airlines (HP), Alaska Airlines (AS) might be the worst choice.

4.3 Trends Of Cities

Illinois, Georgia, California, New York, and Texas had the largest number of delays. It also seems like Northern states (such as Illinois, New York, Massachusetts, Pennsylvania, etc.) have more cancellations during the winter months (especially January and December) when compared to other states. In addition, departure from New York, Dallas-Fort Worth, Newark, Los Angeles, St Louis, Boston and Las Vegas were easier to get a longer delay time while arrival at Honolulu, Los Angeles were easier to get a longer delay time.

4.4 Top Delay Manufacturer

As for average delay time, Boeing, MacDonnell Douglas, Cessna and Raven were at the top levels.

4.5 Choice Of Airports

Airports like Merle K Smith Airport (CDV), Yakutat (YAK), Jackson Hole Airport (JAC) had around 15 minutes' delay time for both year. While, airports like Gulfport-Biloxi International Airport (GPT), Minot International Airport (MOT), Sioux Falls Regional Airport (SUX), Great Falls International Airport (GTF), Grand Forks International Airport (GFK), Hector International Airport (FAR) and Bismarck Airport (BIS) had early arrival for both two years. Therefore, these airports might be a better choice for travel.

4.6 Big Historical Events

From our analysis, we identify big events happened during two years because there had some peaks or extremely low values that disturbed the normal patterns. Those events including:

- "The Blizzard of 1999 Midwestern" struck Iowa, Wisconsin, Illinois, Indiana and Michigan on January 2nd- January 4th,1999;
- A severe ice storm called 'the January 1999 north American ice storm' struck Washington D.C. on January 14 and 15, 1999;
- A series of Islamist terrorist attacks that were planned to occur on or near January 1, 2000, with the bombing of four sites in Jordan, the bombing of Los Angeles International Airport (LAX);
- The January 25,2000 winter storms impacted North Carolina;
- A large nor'easter struck the east coast of the United States on December 30th 2000 causing many cancellations at airports;

5. Contribution

Neeraj Asthana	The cancellation trends analysis and Hadoop
Wanxin Bai	The arrival and departure delay analysis
Goran Tomic	The analysis of number of flights, city network, delay classification
Syed Mohammad Ali Shah	

6. References

- http://www.airfarewatchdog.com/pages/3799702/airline-letter-codes/

- http://www.onthisday.com/

- https://en.wikipedia.org/wiki/January_2000_North_American_blizzard

- https://en.wikipedia.org/wiki/December_2000_nor%27easter