Analysis on Airlines Data in 1997 and 2002

Chendan Tang, Chunlei Liu, Conrad Manaugh,Yunan Shi

May 1, 2019

Abstract

This analysis focuses on the departure delay and cancellation rate in airline data. It includes four main parts: introduction, trend analysis and visualization, summary, and recommendation. The introduction part introduces the background and data preparation. The trend analysis and visualization part analyzes different segments of departure delay and cancellation with data visualization. The summary part briefly concludes what we found. The recommendation part gives some suggestions based on our analysis.

1 Introduction

This paper is a final project for STAT480 (Data Science Foundation) instructed by Prof. Glosemeyer during Spring 2019. We are given four data sets: 'airlines', 'airports', 'carriers' and 'plane -data'. 'airlines' and 'carriers' data sets are provided by the US Department of Transportations Bureau of Transportation Statistics (BTS). 'airports' data is from the Open Flights initiative and 'plane-data' is from the FAA registration database. The goal of this paper is to analyze causes of departure delay and cancallation.

There are total 29 variables in the 'airlines' data, and we selected important 9 variables out of them. We are focusing on analyzing departure delay and cancellation, so we have to drop all the missing values in 'DepDelay' and 'Cancelled' since there isn't enough information to fill them out and estimating them can cause large bias. In order to consider information from more than one table, it's necessary to combine data sets when needed. This analysis is conducted using R and Hive due to the large size of the data sets .

2 Trend Analysis and Visualization

2.1 Departure Delay Analysis

2.1.1 Summary Statistics for Two Years

According to Hive result and figure 1, the average departure delay in 2002 is 5.53 mins, lower than 8.24 mins in 1997. The median departure delay for the two years are both 0 min, most flights in those years have no delay. The minimum and maximum departure delay for year 1997 are -918 mins and 1618 mins. The

minimum and maximum departure delay for year 2002 are -1370 mins and 2119 mins. The range of the departure delay for year 2002 is wider than that for year 1997. The standard deviation for year 1997 and year 2002 is 28.47 mins and 26.10 mins. The departure delay for year 2002 is less variant than that for year 1997. The total flight amount for year 1997 and year 2002 is 5411843 and 5271359. The flight amount is similar between two years.

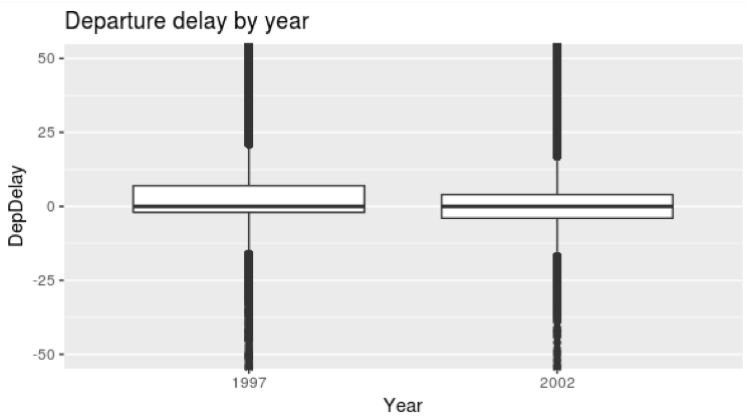

Figure 1: Box Plot of Departure Delay for Different Years Focusing on Departure Delay From -50 to 50

2.1.2 Different Months

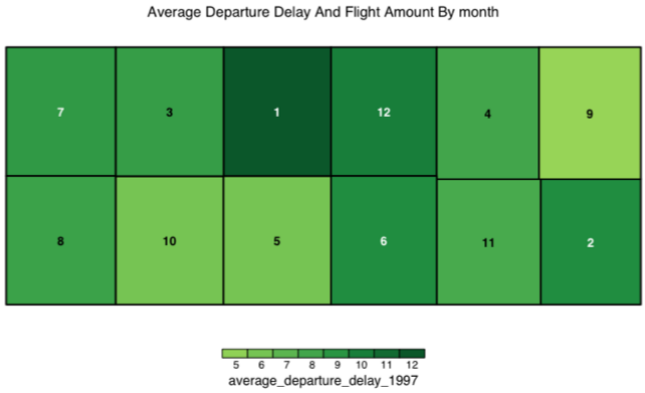

Figure 2: the size of the cell represents the flight amount, the color of the cell represents the average departure delay, the index represents different months in 1997

Looking at figure 2, there is no significant difference in flight amount between different months. So, the monthly flight amount is not strongly correlated with departure delay. In figure 3, there is also no significant difference in flight amount between different months. June and December have the highest average departure delay, different from year 1997. The flight amount is not the reason for departure delay.

From figure 4, the average departure delay in 2002 is lower than that in 1997. The general trend across months is very similar between 1997 and 2002, except for March and November. The highest and lowest

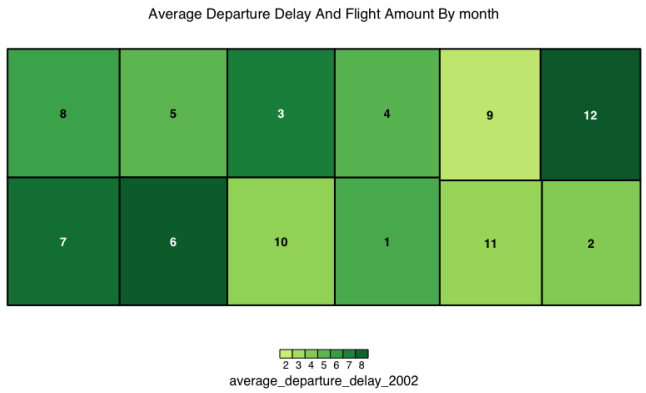

Figure 3: the size of the cell represents the flight amount, the color of the cell represents the average departure delay, the index represents different months in 1997

departure delay months in 1997 are January and September. The highest and lowest departure delay months in 2002 are December and September.

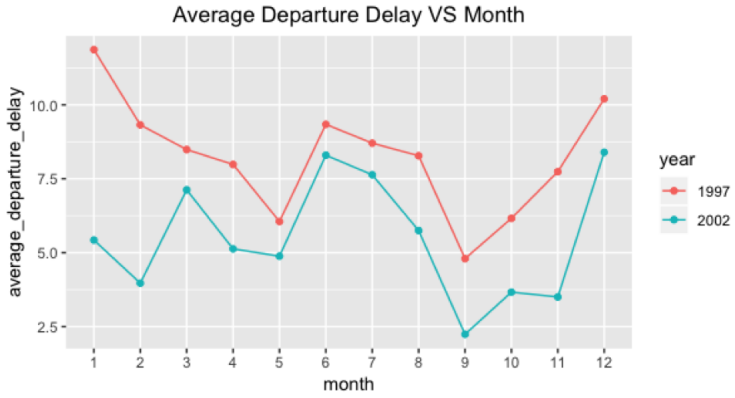

Figure 4

2.1.3 Different Hours

For figure 5, there is significant difference in flight amount between different hours of day. There is less flight and tends to have lower average departure delay between midnight and early morning. 7 PM and 8 PM have the highest average departure delay. We see similar trends in figure 6, which is a similar graph to figure 5 but for the year 2002. Once again early mornings have fewest flights and least delay, compared to the worst delay times of 7 or 8 PM.

The average departure delay in 2002 is lower than that in 1997. In figure 7, the general trend across hours of day is very similar between 1997 and 2002. The highest and lowest departure delay hour of day in 1997 are 8 PM and 3 AM. The highest and lowest departure delay hour of day in 2002 are 7 PM and 5 AM. One interesting fact is year 1997 doesn't have flights on 4 AM. Year 2002 doesn't have flights on 3 AM and 4 AM. This is probably due to cost control. A hypothetical flight that arrives in San Francisco at 2 am would connect

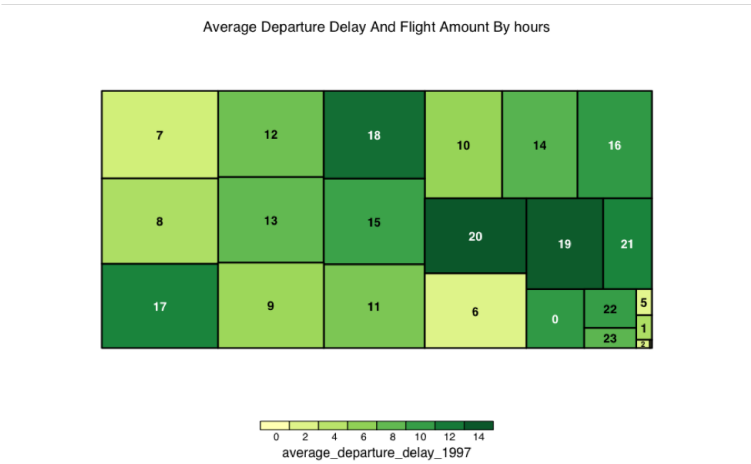

Figure 5: the size of the cell represents the flight amount, the color of the cell represents the average departure delay, the index represents different hours in 1997

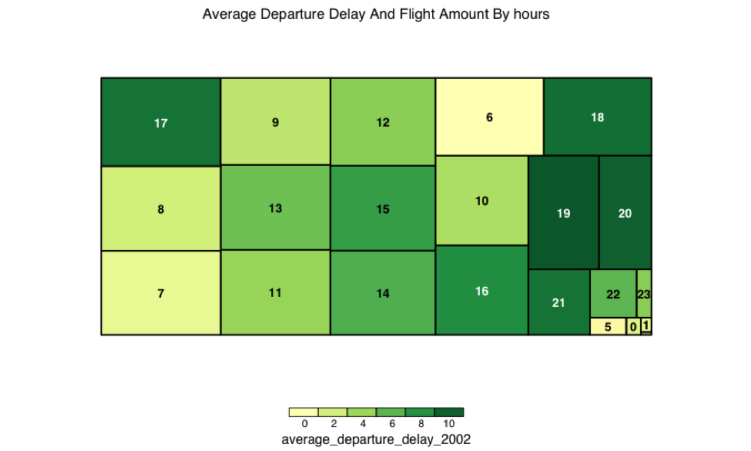

Figure 6: similar graph to Figure 5

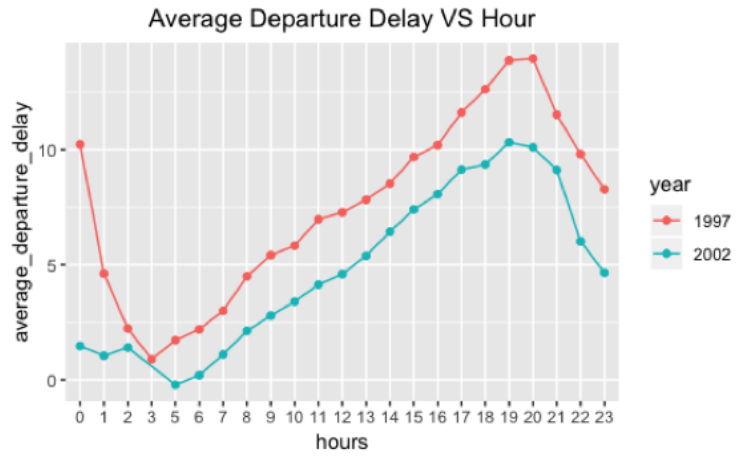

Figure 7

with flights departing to LA at 3:30 or 4 am, which few people would want to take. There would be almost no airport services open at that hour unless everyone put on another shift: restaurants, shops, for one or two flights max. Airport business owners would probably choose to remain shut down overnight anyway. No flights operates in a vacuum.

2.1.4 Different Carriers

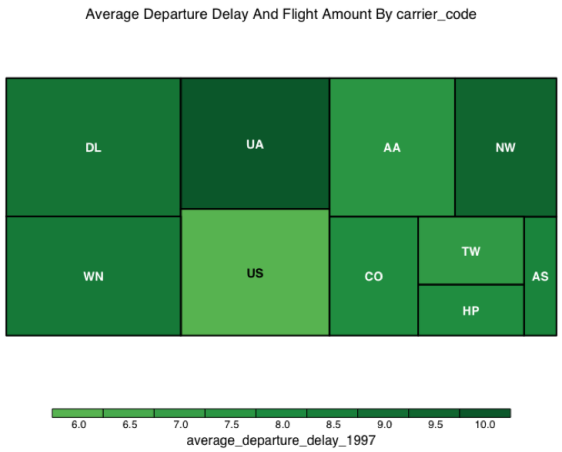

Figure 8: the size of the cell represents the flight amount, the color of the cell represents the average departure delay, the index represents different carriers in 1997

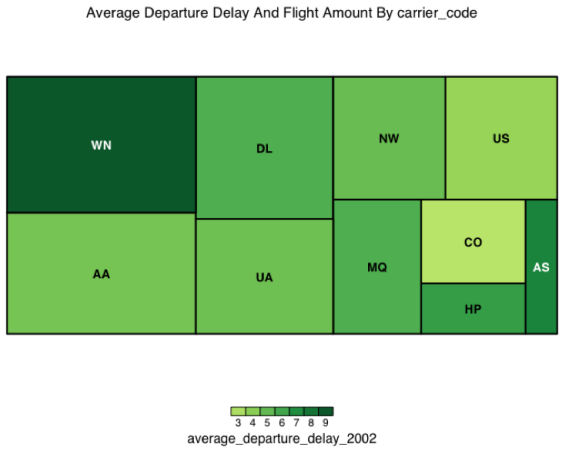

Figure 9: Similar to Figure 8

For figure 8, there is significant difference in flight amount between different carriers. CO, TW, HP and AS tend to have smaller flight amount and lower average departure delay. UA has the highest average departure delay.

Similar to figure 8, figure 9 also has significant difference in flight amount between different carriers. MQ, CO, HP and AS tend to have small flight amount, but AS has a high average departure delay. WN has

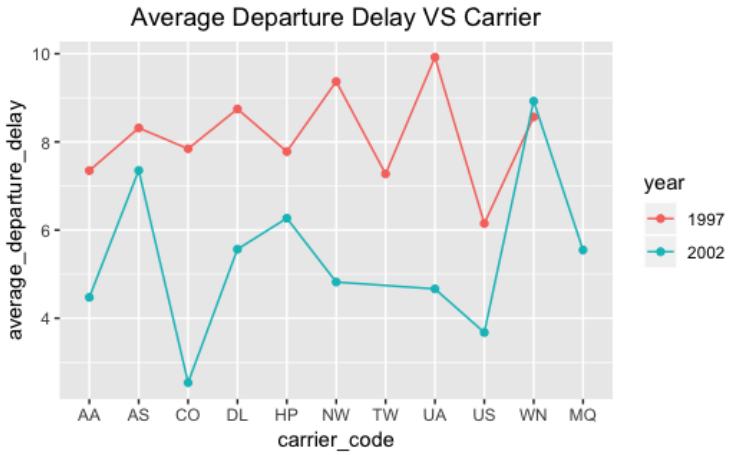

Figure 10: the size of the cell represents the flight amount, the color of the cell represents the average departure delay, the index represents different carriers in 2002

the highest average departure delay.

The average departure delay in 2002 is lower than that in 1997 in figure 10, but WN has a higher average departure delay then 1997. The highest and lowest departure delay carrier in 1997 are UA and TW. The highest and lowest departure delay hour of day in 2002 are WN and CO.

The average departure delay of United Airline(UA) and Continental Airline(CO) have decreased a lot during the several years. The average departure delay of each carrier is decreasing except Southwest Airlines is getting worse during the several years. American Eagle(MQ) appears in 2002 while not in 1997.

2.1.5 Different Manufacturers

We have already known that the overall delay in 2002 and in 1997, but what about for each manufacturer? Does different manufacturer have a different delay rate?

According to Table 1 and 2, Boeing, Mcdonnell Douglas [1] , Airbus [2], and Cessna take up most of the market in the world in 1997 and 2002. I will focus on these manufacturers in the following analysis.

The stacked density allows to plot several densities in one graph to make it easier to compare. In figure 11, all manufacturers have highest density at around 0 min. And the two graphs are skewed right, which means it's more likely for these manufactures to not have delay than having delay. When it comes to having delay, Airbus has the biggest probability of having delay in 0 to 20 min, then follows Boeing, Cessna and McDonnell Douglas. The pattern stays the same in 2002.

[1]According to Wikipedia, McDonnell Douglas was formed by the merger of McDonnell Aircraft and the Douglas Aircraft Company in 1967. So, I renamed 'MCDONNELL DOUGLAS AIRCRAFT CO', 'MCDONNELL DOUGLAS CORPORATION' and 'DOUGLAS' to 'MCDONNELL DOUGLAS', they are the same company.

[2]According to Wiki, in 2001, Airbus Industrie GIE was reorganised as Airbus SAS, a simplified joint-stock company. So we consider them the same compnay. here we unify them user 'AIRBUS' name

Table 1: top 4 manufacturers producing the biggest number of planes in 1997

Manufacturer	Number of Plane
BOEING	908694
MCDONNELL DOUGLAS	694928
AIRBUS INDUSTRIE	128712
CESSNA	11403

Table 2: top 4 manufacturers producing the biggest number of planes in 2002

Manufacturer	Number of Plane
BOEING	1067163
MCDONNELL DOUGLAS	512411
AIRBUS INDUSTRIE	409609
CESSNA	9405

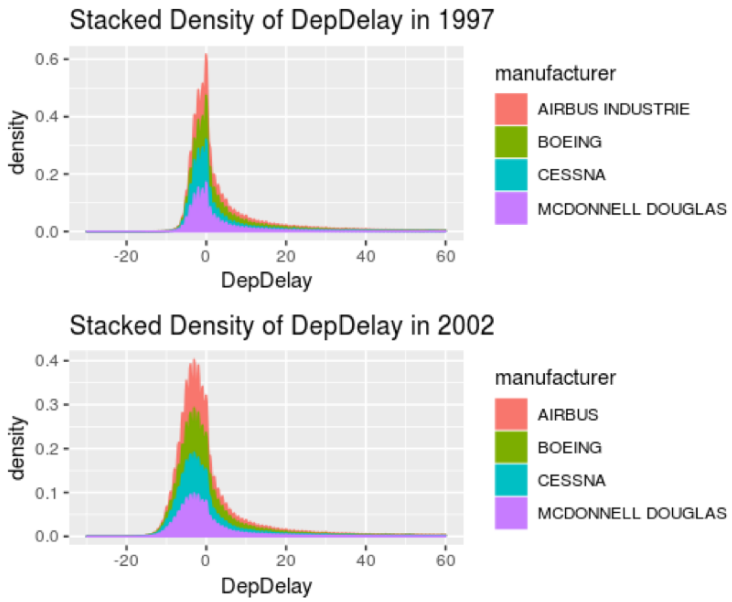

Figure 11: The stacked density of departure delay in 1997 and 2002 by different manufacturer. Here I only chose the top 4 manufactures based on the number of planes each manufacturer owns. Graph by Chendan Tang

2.1.6 Different Plane Age

If we add plane age into consideration, older planes may require more maintenance and check before flights, which can cause more delay than newly-built planes, but it may take more time for a pilot to get familiar with new planes since they may contain some updated features, and it can also cause the delay.

Figure 12 shows the relationship between age and delay, and it varies for different manufacturers. In 1997, old planes owned by Airbus, Boeing and Mcdonnell Douglas are more likely to have longer delay. But for

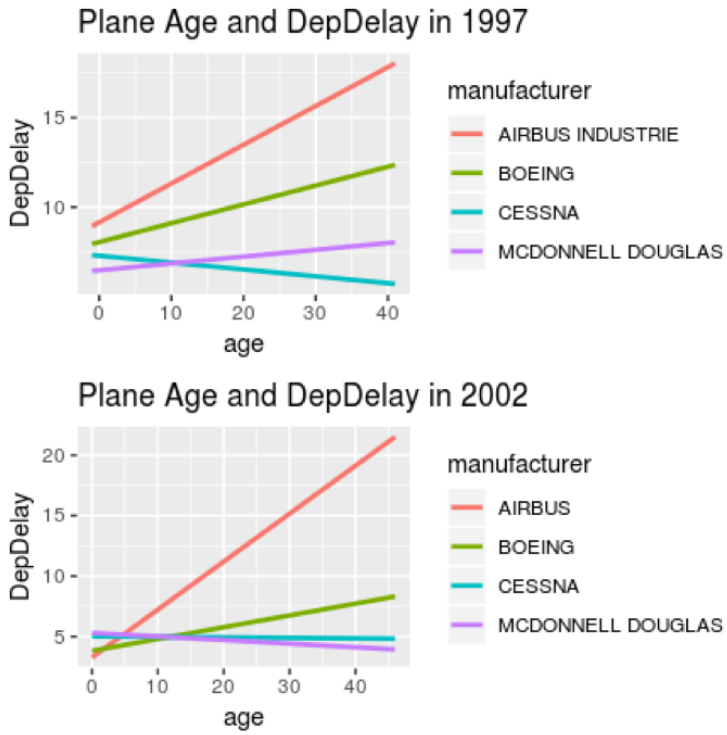

Figure 12: Smoothed linear relationship in full range between delay and plane age by top 4 manufacturer. Age is the difference between Year Manufactured and Year of the data set. Graph by Chendan Tang

Cessna, new planes tend to have more delay. And Airbus has the biggest increasing rate of delay as plane age increased. In 2002, the relationship for Mcdonnell Douglas changed, new planes are more likely to have longer delay.

The line for Airbus becomes steeper in 2002, which means the delay time is more sensitive to plane age than in 1997. However, the lines are flatter for Boeing, Mcdonnell Douglas and Cessna, which means the plane age has less impact on the delay.

2.1.7 Different Airports

In 1997 the U.S had 206 total airports that had a flight either depart or arrive. Compared to in 2002 the U.S has 13 more airports at a total of 2001. Looking at both years for the airports with the most departures we see a slight reduction in average daily flights for most airports. The biggest two airports in 2002 showed increases from 1997. One being O'Hare which has the most flights and has shown gain over the past 5 years.

For both years the airports with few departures are the ones that are going to sparsely populated areas, often these areas relate to native tribes and or military bases which have little to no public usage but are occasionally used for military flights. Examples of this are the airport in Guam from 1997 and the Sioux airport from 2002.

Looking at both years for the airports with the most departures we see a slight reduction in average daily

Table 3: Comparing the airports with the least daily departures, showing the 4 least for both 1997 and 2002

Airport 1997	Average Daily Departures	Airport 2002	Average Daily Departures
Guam International	0	Atlantic City International	0
Greenbrier Valley	0.23	Bush	0
Gustavus	0.23	Sioux Gateway	.0027
Nantucket Memorial	0.041	Erie Intl	.0547

Table 4: Comparing the airports with the least daily departures, showing the 4 least for both 1997 and 2002

Airport 1997	Average Daily Departures	Airport 2002	Average Daily Departures
Los Angeles International	503	Los Angeles International	494
Dallas-Fort Worth International	674	William B Hartsfield-Atlanta Int	628
William B Hartsfield-Atlanta Intl	676	Dallas-Fort Worth International	761
Chicago O'Hare International	793	Chicago O'Hare International	879

flights for most airports. The biggest two airports in 2002 showed increases from 1997. One being O'Hare which has the most flights and has shown gain over the past 5 years. When comparing the average airports average daily departures, we see that the average of 1997 is around 70, and 2002 is only around 65. This means that on average airports have less flights in 2002 then 1997.

Table 5: Comparing the airports with the least average departure delay, showing the 4 least for both 1997 and 2002

Airport 1997	Average Daily Departures	Airport 2002	Average Daily Departures
Quad City	-0.117	Great Falls Intl	-0.049
Lafayette Regional	-0.011	McGhee-Tyson	-0.043
Guam International	0	Glacier Park Intl	-0.034
Tompkins Cty	0.0029	Texarkana Regional-Webb	-0.033

The main commonality comparing the airports with the least average departure delay is that they are all not well-known airports, likely with few departures. This makes sense, as airports with less departures find it easier to stay on top of flight schedules.

Table 6: Comparing the airports with the most average departure delay, showing the 4 least for both 1997 and 2002

Airport 1997	Average Daily Departures	Airport 2002	Average Daily Departures
Chicago O'Hare International	5	William B Hartsfield-Atlanta Intl	3.33
William B Hartsfield-Atlanta Intl	5.27	Detroit Metropolitan-Wayne County	3.41
Detroit Metropolitan-Wayne County	5.26	McCarran International	3.52
Minneapolis-St Paul Intl	5.48	Chicago O'Hare International	3.76

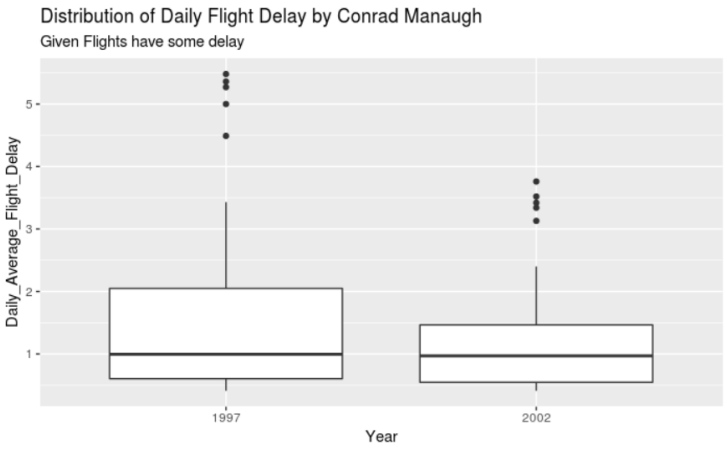

Figure 13

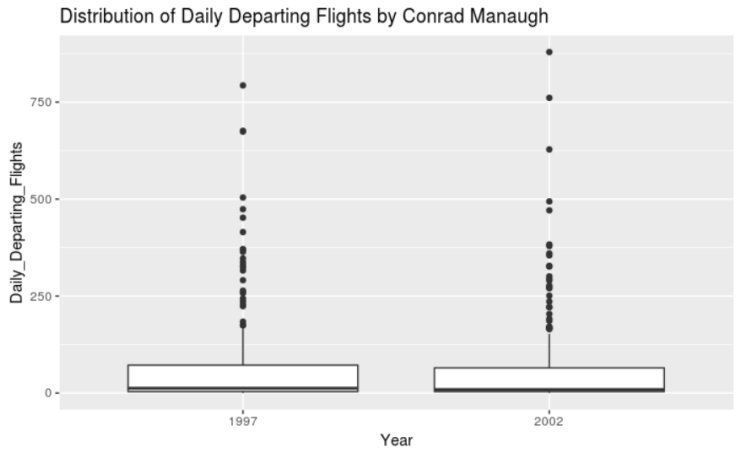

Figure 14

Looking at the airports with the most average departure delay in Table 6 we see a large difference between the 1997 and 2002 maximum delays. The highest in 1997 have delay around 5 while no airports in 2002 have delay above 4. The only commonality between having many departures and having high departure delay, comparing Table 6 and 4, we see that William B Hartsfield-Atlanta Int and O'Hare appears in both tables for both years. This may indicate a relationship between an airport's daily flights and its average delay. Regardless 2002 has become more efficient at reducing departure delay when compared to 1997, as the 10 average delay has reduced from .7 minutes to .4 minutes.

In figure 13 we see a boxplot of the average flight delay. This graph originally has many delays at zero, causing the boxplot to converge at zero. Therefore, delay's close to zero are removed to help us understand the distribution of delays when they happen. Delay tends to be only around 1 minute for both years, but in year 1997 it was also likely for delay to be up to 2 years. In 2002 airports seem to have improved, as average delay is only likely to be up to 1.5. 1997 also has many more outliers' way above the boxplot, compared to 2002 who's outliers are 1 or 2 minutes less than 1997's. figure 14 shows the likely range of average airport departures on the right. This figure has a mean, and quantiles, that stay similar. Most airports have less than

75 departures, on average at 0. Looking at the outliers in 1997 we see they tend to have more average flights. The only exception is the airports that have the most flights. So, over the 5 years not much change happened in average departures, a small reduction in average flights unless the airport is one of the most widely departed from airport

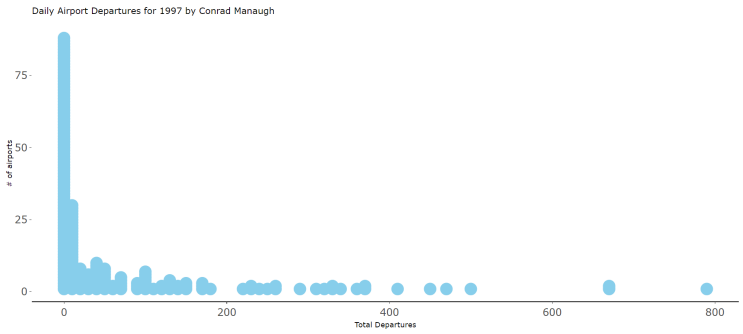

Figure 15

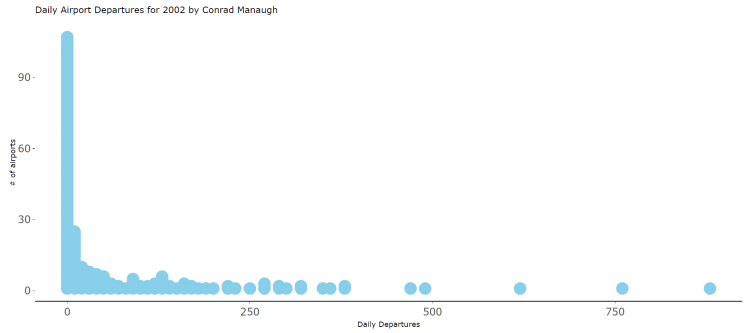

Figure 16

Looking at figure 15 and 16 we see to interactive histograms for year 1997 and 2002 that can both be accessed from the html documents of the same name submitted with the paper. For both years many airports have a close to zero daily flight, meaning that most airports are not very busy, but 2002 has many more airports near zero, so in 2002 we have more airports with little air traffic. We also see airports that are outliers, having hundreds of flights more than average daily. These outliers seem to be predominantly the same airports, showing that big airports tend to stay big. Looking at these outliers using the html we see these outlier airports generally got larger, so as time goes on big airports are getting busier and busier. Most airports are between 0 and 500 flights, this holds true for bother years. This portion of airports does not seem to have decreased on increased much.

The interactive histograms in figure 17 and 18 show a distribution of airports average delay for year 1997 and 2002, these histograms are also interactive on the source html. In 1997 our most delayed airports have around 6 minutes of average delay per flight, compared to only 4 minutes on average for 2002. This illustrates that in these 5 years airports have acted to reduce the worst cases of delay. Looking at most airports that have an average delay near or around zero we see that in 2002 we have more flights that leave on time. On average

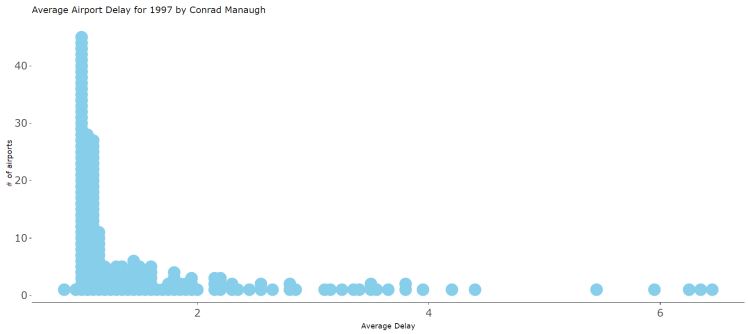

Figure 17

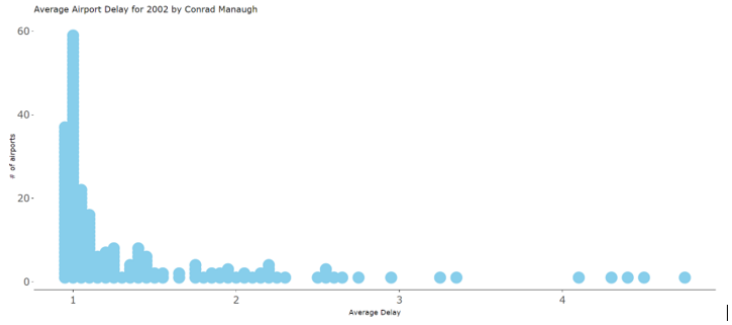

Figure 18

in 1997 flights are between 0 and 4 minutes of delay, this is reduced in 2002 to a range of 0 to 3. This indicates that generally all airports are getting more efficient at reducing departure delay over the 5-year period.

2.1.8.1 Airport Network Network Analysis is very useful when visualizing the connection and relation between a collection of entities. (Qu & Zhu)

LAX and SFO, LAS and LAX, and PHX and LAX are the busiest routes with more than 15000 flights between each pair in 1997. From the networks, ORD, LAX and PHX are the main transportation hubs, as there are a lot of edges connected to them. And among the airports shown in the figure 19, DTW and EWR have the largest departure delay in terms of the size of nodes. In 2002, PHX & LAX & LAS, ORD & MSP, ORD & EWR and so on seem to have more airline traffic, and LAX, PHX and ORDwere the main hubs then. According to the node size in figure 20, DTW, HOU and DAL have higher average delay.

If we compare two figures together, the scaled delay time in 2002 is generally shorter than in 1997. The number of airports shown in 2002 is fewer than in 1997, in other words, there are fewer airports with more than 6000 flights a year. In my opinion, the new construction of airports or optimization in the airline routes management contribute to this change, so that the traffic between each airport is reduced in 2002. For example, in figure 19, the weight between LAS and LAX is about 15000, but in figure 20 the weight is decreased to about 12000. Another thing I'd like to mention is the airports located on the edge of the network (or have fewer connection to other airports) tend to have less delay time, while the airports located in the center (or have more connections) tend to have more delay. It indeed manifests one of our common senses: It's more

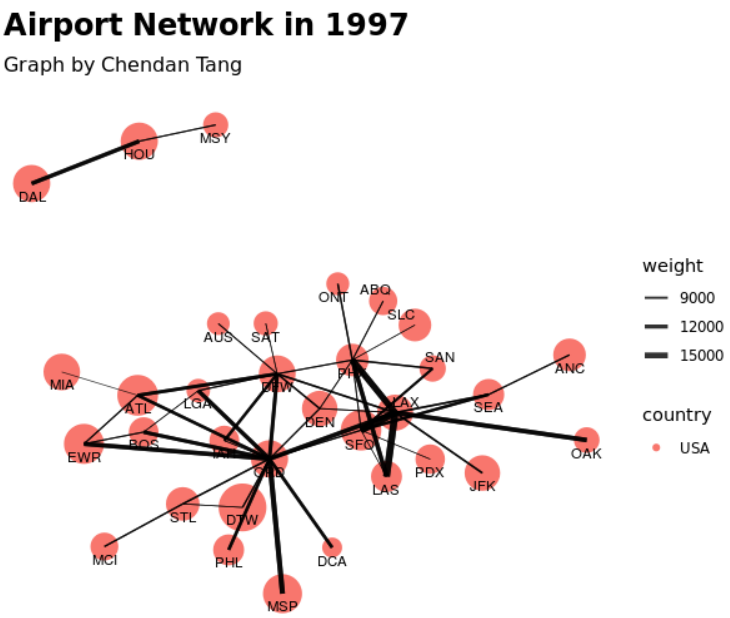

Figure 19: The network of airports with at least 6000 flights in 1997. Weight represents the number of flights between each pair of airports, the color shows the variety of countries the airports locate in, and the size of the node measures the average of departure delay. All values in 1997 and 2002 are calculated on the same scale. Graph by Chendan Tang.

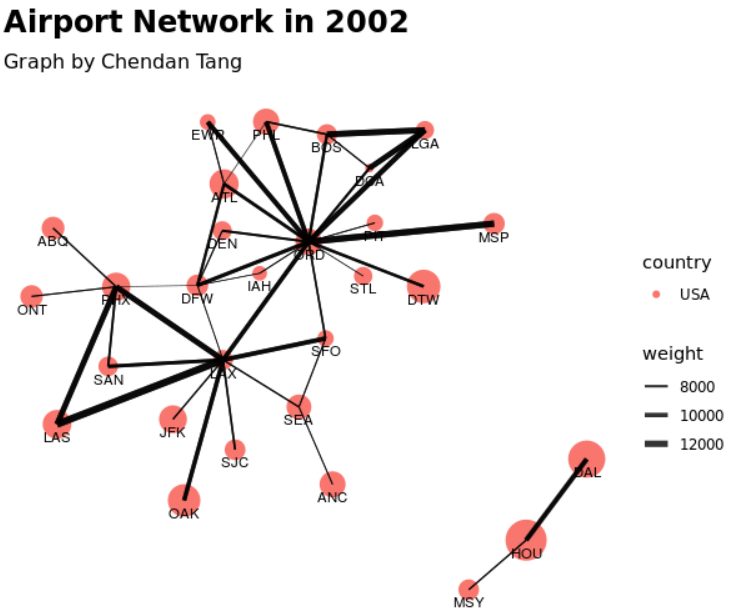

Figure 20: The network of airports with at least 6000 flights in 2002. Similar to figure 19. Graph by Chendan Tang.

likely to encounter delay at some busy airports.

2.1.8 Different States

2.1.9.1 State Networks We can group the flights by the state of origin and state of destination to get an idea of how state location might impact delay. We see that in 1997 we have 52 states, while in 2002 we have 51 states. The normal 50 states of the US are in both years, plus the territories of Guam and Puerto Rico in 1997. The year 2002 has an NA observation, so it only uses the original 50 stats.

Table 7: Comparing the states with the least daily departures, showing the 4 least for both 1997 and 2002

State 1997	Average Daily Departures	State 2002	Average Daily Departures
Guam	0	Wyoming	2.9
Wyoming	4.3	Vermont	8.2
West Virginia	4.7	South Dakota	9.1
Vermont	7.8	Virginia	10.6

We see that the states with the least departures in 1997 are states with less population generally, Guam, Virginia, Vermont, and West Virginia. all have low populations. We also see that in 2002 the states with the least departures had more departures on average, so a general trend of increasing daily departures may exist, at least for the sparsely populated states.

Table 8: Comparing the states with the most daily departures, showing the 4 least for both 1997 and 2002

State 1997	Average Daily Departures	State 2002	Average Daily Departures
Florida	894	Florida	901
Illinois	916	Illinois	1036
Texas	1641	California	1737
California	1748	Texas	1755

The states with the highest daily departures in 1997 tend to be more urban and populous states. Good examples are Illinois, California, and Texas, which have some of the highest populations of all the states. We also see a general rise in daily departures, like in Table 7. It is interesting how the four states with most average daily departure is the same.

Table 9: Comparing the states with the least average departure delay, showing the 4 least for both 1997 and 2002

State 1997	Average Departure Delay	State 2002	Average Departure Delay
West Virginia	-0.006	North Dakota	-0.022
Guam	0	South Dakota	0.0122
North Dakota	0.079	Montana	0.0695
New Hamsphire	0.32	Iowa	0.0775

The states with the least delay in both years all have around zero average departure delay. This could

be because they are often on time, but more likely it is because they have very few flights. Since all states in Table 9 seem to be sparsely populated states, it is likely they have few flights, and that is the reason they have such low average delay. Comparing the two years we see that they did have different states with lowest average delay in the 5-year period, but they are still sparsely populated states, so it becomes more likely that rural states have less departure delay.

Table 10: Comparing the states with the most average departure delay, showing the 4 least for both 1997 and 2002

State 1997	Average Departure Delay	State 2002	Average Departure Delay
Minnesota	8.7	Connecticut	6.0
Georgia	8.9	Georgia	6.2
Michigan	9.7	Michigan	6.9
Illinois	9.7	Illinois	7.4

We see a clear decrease in average departure delay when looking at the states with the largest departure delay in Table 10. Also, the states for both years are similar, indicating that something about these states gives them higher delay. For Illinois it is clearly because of the high amount of departing flights from O'Hare, but an explanation for the other states needs to be researched. We can use an interactive histogram like we did before to see why these states might have high departure delay, the most likely explanation is that the states have a high amount of departing flights. The state histograms are attached but are not printed to this reports text.

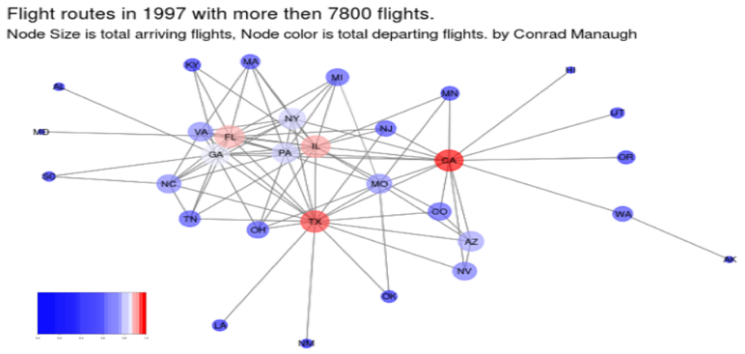

Figure 21: A network analyses of State flight flow in 1997. The color shows the number of departing flights(blue to red with shrunk flight from 0 to 1). The nodes size is dictated by total arriving flights. Since this only includes flights above 7800 big nodes that are blue are nodes that have many outgoing flights to small airports, as an airport typically has similar departing and arriving flights.

Looking at the air traffic network for high density in 1997 from figure 21 we see many medium sized nodes that are a variety of blue shades. We also see 4 larger nodes that are colored red shades. Lastly, we see many small dark blue nodes. The red ones are red because they have the highest total departing flights, they are large because they have the highest arriving flights. They are also the nodes with the most edges. The bluer the nodes get the smaller they tend to get and the less edges they tend to have.

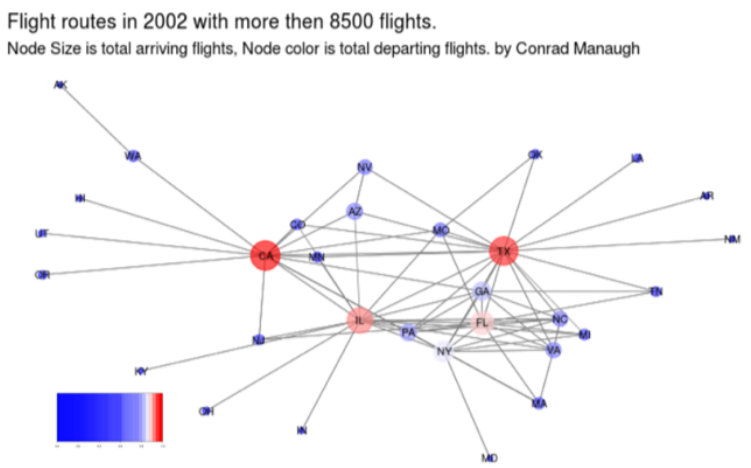

Figure 22: A network analyses of State flight flow in 2002. Similar to Figure 21

Compared to the 2002 air traffic network with high density we see that the medium sized states are gone. Now air traffic flows through the states with the most traffic. From these large airports traffic will flow to the smaller ones as needed, like a highway. It seems like over the 5 years between 1997 and 2002 air traffic has shifted from a more connected network where a flight may go from any airport to any destination, to a less connected network where flights travel from high traffic port to high traffic port. Then when a flight needs to go to a less high traffic port they send one flight from a high traffic airport to a low traffic one. Perhaps it is this highway like method that decreased delay on average for most states and most airports.

While looking at the airport and state data it has become reasonable to assume that airports, and to some extent states, that have more departing flights have a higher flight delay on average.

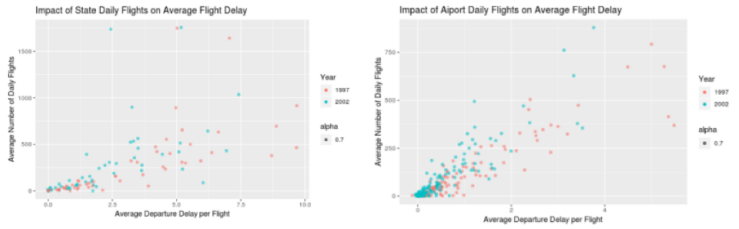

Figure 23: A set of scatterplots for states and airports, the x axis is the average delay per flight, the y axis is the number of departure from the state or airport, color is the year.

It is clear to some extent the idea that departures and departure delay are related is correct. The relationship is less strong with state, likely indicating the relationship is specific to airports themselves. States will share any relationship the airports of that state averaged out, so the relationship is weakened. In the right graph of Figure 23 we see the positive relationship. Also, worth note is it doesn't look like between the 5 years airports changed the daily departure amount very much, a small decrease if anything. It looks more like they mostly stagnated in number of departures, but drastically decreased departures, as the blue 2002 points are to the left of the red 1997 points.

2.2 Cancellation Trend Analysis

2.2.1 Average number of cancellations within 1997 & 2002

We create a Hadoop to calculate the average cancellation in both 1997 & 2002. According to the calculation, the overall cancellation rate for 1997 is 1.81%. Furthermore, the overall cancellation rate for 2002 is 1.24% which is smaller than the rate in 1997. Therefore, the result might indicate that 2002 has better performance in general. However, we still need further analysis on weather and other factors to see if this is the accurate conclusion.

2.2.2 Cancellation trend by month

We start by analyzing the trend based on months. We use line plots that indexed the number of cancellations by month. According to the figure 24 (left chart), most cancellation happen in the month of January. The result is kind of make sense since major snowstorms and extreme weather happen during January. The overall trend in 1997 based on month is kind of normal based on common sense of weather situation in each month. According to the figure 24 (right chart), most cancellation happen in the month of June in 2002 followed by December and January. The result is also reasonable since major snowstorms and extreme weather happen during January and December. June has great cancellation may because of heavy rain and major flooding. I will analyze the month of January in 1997 and the month of June in 2002 as special case based on the weather data in further section.

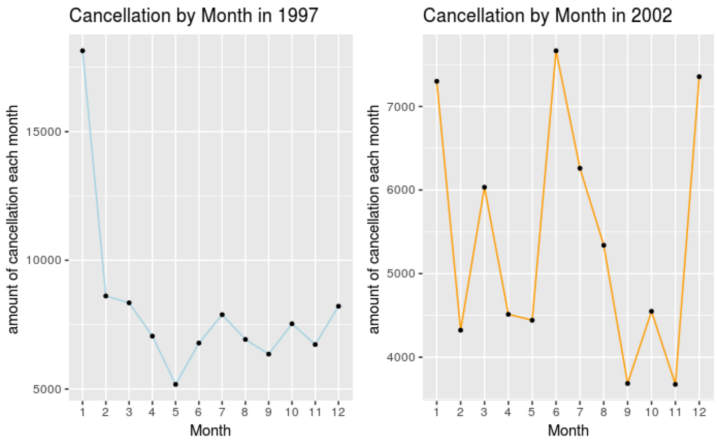

Figure 24

2.2.3 Cancellation trend by Day of Week

Next step we focus on the cancellation trend based on day of week. According to the figure 25 (left one), most cancellation happened on Wednesday through Monday while Friday through Sunday have relatively small amount of cancellation in 1997. The result might because there are more flights on weekdays compared

to the amount of flights on weekends. In 2002, the cancellation most happened on Thursday through Monday while Friday through Sunday have relatively small amount of cancellation (figure 25 right one) which is similar with the result we got in 1997.

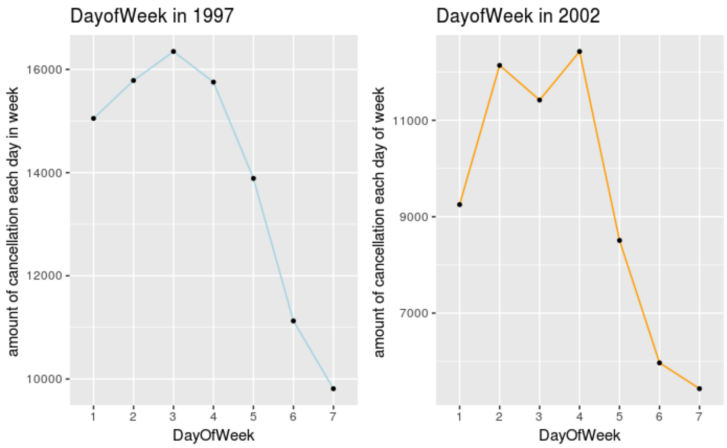

Figure 25

2.2.4 Cancellation trend by State

State is also a significant factor in the cancellation analysis. Our team use streamgraph to show analysis based on state (taking off from an airport in the state). In general, the states of Illinois, Taxes, California has relatively high amount of cancellation followed by New York, Miami and Pennsylvania in 1997(see figure 26). The result is also reasonable since these states hold top big airports in states which has large amount of flights going in and going out. Obviously, north and central part (like New York, Illinois, Texas) has larger cancellation in winter than other parts in January since there is more snowstorms and extreme weather in north and central part [3].

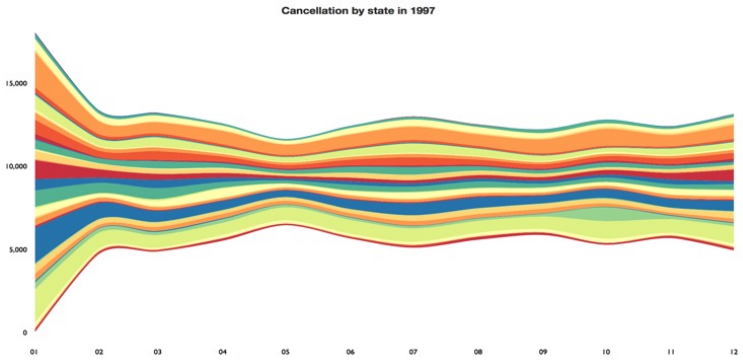

Figure 26

[3]US Department of Commerce,& Noaa. (2015, March 25). 1997 Sioux Falls Area Climate Summary.

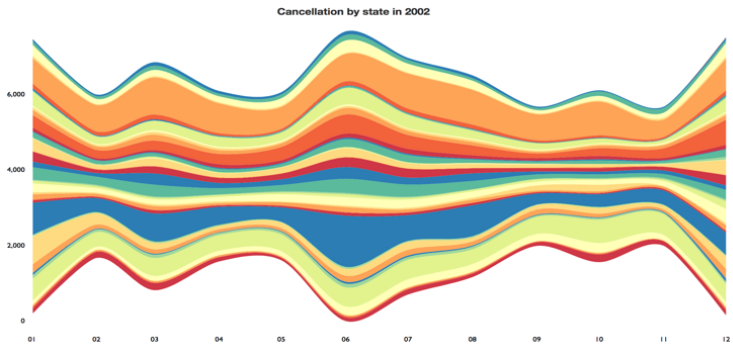

Figure 27

We also use streamgraph for data in 2002. In general, the states of Illinois, Taxes, California has relatively high amount of cancellation followed by New York, Miami and Pennsylvania(see figure 27). The result is very similar with the result in 1997. It is also reasonable since these states hold top big airports in states which has large amount of flights going in and going out. But differently, Texas, Illinois and New York also have great amount of cancellation in June. Based on the historical weather data in June 2002 [4], major flooding occurred in most parts of Texas. Airport in Texas was paralyzed which also affected the other states of major transaction airports like Illinois and New York.

2.2.5 Cancellation by Carrier

At last, we focusing on cancellation trend based on carrier. We display the result that contains 11 main carries using the following plot. We can see from the chart that United Airlines(UA) has the greatest cancellation followed by American Airlines(AA) and Delta Airlines(DL) in 1997(see figure 28 left one). The result is also reasonable since all three of them are major airline companies in United States. They contain more airlines each year than smaller airline companies. In year 2002, Envoy Airlines (MQ) has the greatest cancellation followed by Southwest Airlines (WN) and American Airlines (AA)(see figure 28 right one). The result is a little abnormal since Envoy Airlines is not kind of major airline in United States. However, Envoy Airlines is headquartered in Texas. The great amount of cancellation might occur because of the major flooding in Texas in 2002.

It is also necessary for us to analyze the carrier proportion to see which airline perform better since different airline companies holds different amount of flights each year. In 1997, the result is indeed different with the result we have in the above. Northwestern Airlines(NW)has the highest frequency of cancellation compared to other airline companies. United Airlines(UA)has relatively high cancellation rate too(see figure 29 left one). Therefore, the amount of cancellation by carrier might not be the comprehensive way to conclude. We may need to calculate the frequency of cancellation of the carrier to see which airline has better performance. As we can see, Hawaiian Pacific Airlines(HP) and Southwest Airlines(WN) has relatively low cancellation frequency than other airlines in 1997. In 2002, Envoy Airlines (MQ) still has the highest

[4]US Department of Commerce, & Noaa. (2018, August 14). A Review of Weather in 2002.

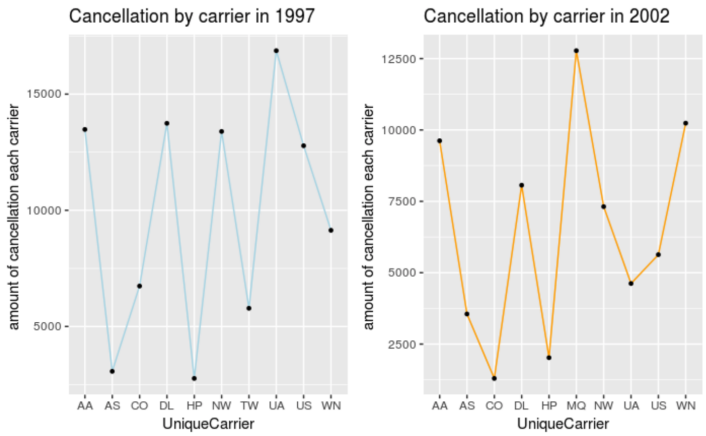

Figure 28

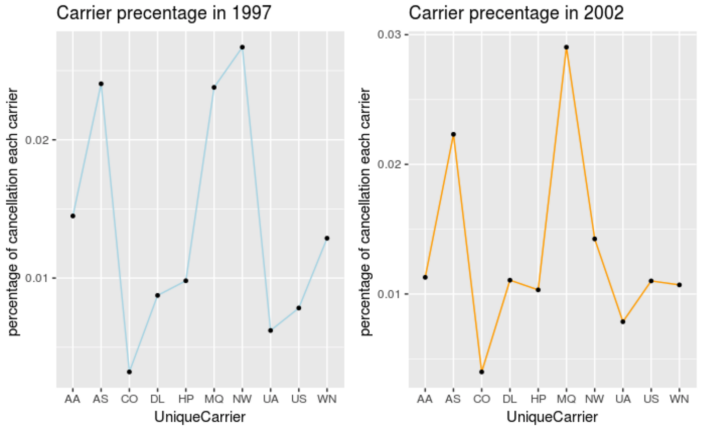

Figure 29

frequency of cancellation compared to other airline companies which confirm our analysis above. Alaska Airlines (AS) has relatively high cancellation rate too. As we can see, Hawaiian Pacific Airlines (HP) and Continental Airlines(CO) has relatively low cancellation frequency than other airlines(see figure 29 right one).

2.2.6 Comparison between two year (Year 1997 & Year 2002)

We create a streamgraph to get a better understanding of the change of cancellation trend over time. According to the streamgraph(see Figure 30 & Figure 31), amount of cancellation is definitely higher in 2002. However, based on our calculation, overall cancellation rate is lower in 2002 than in 1997. Therefore, we can consequently conclude that the overall performance of the major airline is better over time.

Like the analysis we did above, higher cancellation always happens on January, December and in some special circumstances in summer (like June in 2002). There are two obvious high peaks. Specifically, January in 1997 and June in 2002. To conclude, cancellation trend in 1997 is smooth and steady while trend in 2002 has both big and small fluctuations. It is also interesting that March and December in 2002 forms small peaks

in cancellation. The small peaks formed because of major snowstorm in March from Texas to Michigan and high winds and lightning in August. Weather is pretty terrible in 2002 especially in Texas region which confirms the dramatic increase of cancellation in Envoy Airlines (MQ) from 1997 to 2002.

In general, Hawaii Pacific Airlines (HP), and Continental Airlines (CO) are two airline companies whose cancellation did not fluctuate heavily according to the weather change.

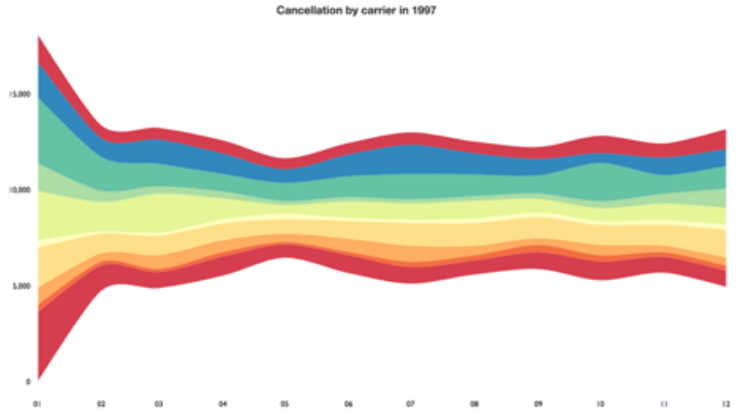

Figure 30

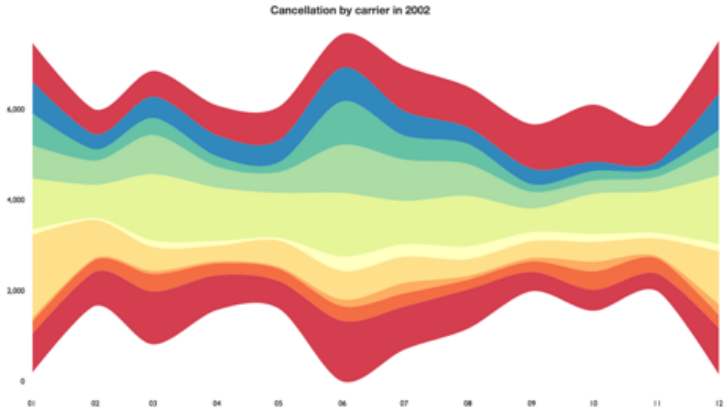

Figure 31

2.2.7.1 Special case studies I (January,1997) According to the streamgraph for days in January of 1997 (see figure 32), we can see that there are two peaks in January 9th and January 15th. American Airlines (AA) and United Airlines (UA) has relatively high cancellation in these two days. According to the weather report [5], there was strong blizzard and dangerous wind in northeast region (South Dakota) and central region (Minnesota) in 9th. The situation is quite similar in January 15th. Most of the airlines cancel the flight from 15th to 18th since the serious blizzard in central region. Visibility was near zero. People cannot even go outside. This caused high cancelations.

[5]US Department of Commerce, & Noaa. (2015, March 25). 1997 Sioux Falls Area Climate Summary.

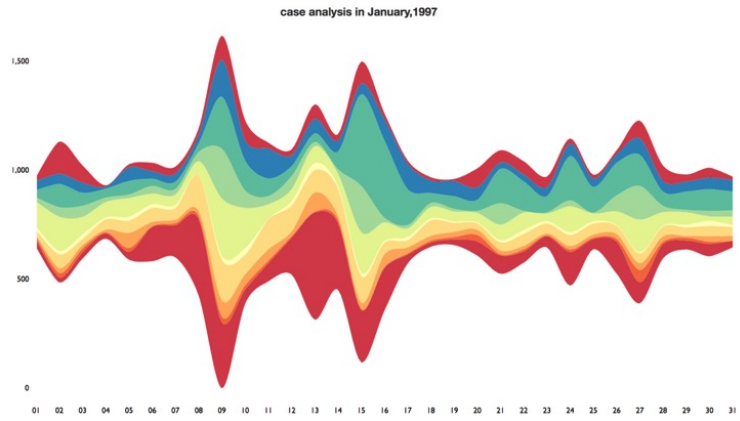

Figure 32

2.2.7.2 Special case studies II (June, 2002) According to the streamgraph(see figure 33) for days in June of 2002, we can see that there are several peaks in June 4th, June 11th and June 27th. American Airlines (AA), United Airlines (UA) and Envoy Airlines (MQ) has relatively high cancellation in these three days.

According to the weather report [6], there was heavy rain, thunderstorm and flooding occurred in most part of the US from northeast region through southeast region in June 3rd to 4th. The situation is quite similar in January 10th to 11th and January 26th to 27th. Most of region in US especially Texas suffered from the serious weather disaster. Therefore, the cancellation is quite high through the whole month.

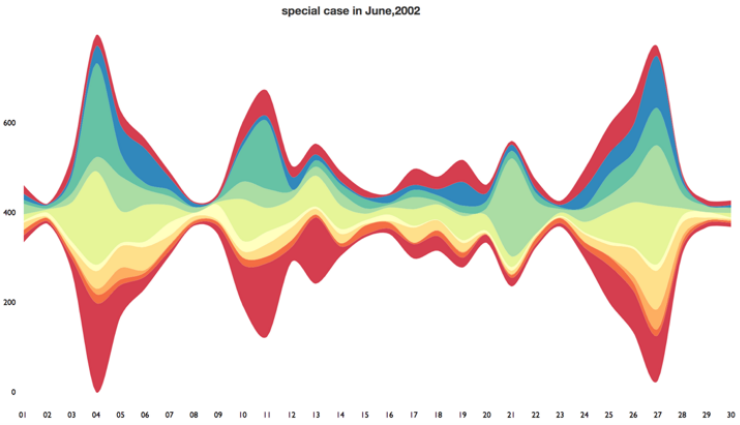

Figure 33

3 Summary

Assuming the trend of decreasing departure delay and cancellation of the flight has remained somewhat consistent over the past years, and matches the trend seen between 1997 and 2002 then it has continued to decrease on average. If correct our flights now would leave on time at a higher percentage and leave earlier on

[6]US Department of Commerce, & Noaa. (2018, August 14). A Review of Weather in 2002.

delayed flights.

As the total airports are increasing flights are becoming slightly less connected, flights tend to be between a large airport and another airport, rarely between two small airports. If this trend between 1997 and 2002 has continued, then the biggest airports will see increases in the amount of flights as well as the number of airports that fly to them. Opposingly the smaller airports may have less flights and will likely only fly to airports nearby or the biggest airports.

The trend between an airport departing flights and that airports average delay has been clearly established. We also know that air traffic is becoming more like a highway, with few very large airports that service many smaller local airports. The result of such a change is reduced departure time caused by excess flights for all airports that are not the largest. So, it may follow that the increasing density of flight travel is an optimization pattern that in part decreases delay which is also applied to cancellation trend. The main exception is that even the biggest airports saw a large drop in average delay and cancellation despite the increasing density, while under our assumptions they should have increased. It is likely that if our hypothesis about dense air traffic is true, and it does increase delay and cancellation for most airports, then something else at play must also decrease the delay and cancellation for the biggest airports.

Boeing, Mcdonnell Douglas, Airbus, and Cessna are the top 4 manufacturers, and Airbus has the largest probability of having 0 to 20 minutes delay in both 1997 and 2002. For these 4 manufacturers, the departure delay is no longer sensitive to plane age in 2002, except Airbus.

4 Reccomendation

June and December may have a high departure delay and cancellation due to high traffic or extreme weather like snowstorms and flooding. 7 PM and 8 PM have a high average departure delay because the time range is popular. For carrier, Southwest Airlines (WN) has a high departure delay. So, if your flight is WN carrier, during 7 PM and 8 PM in December or June, you should book an earlier flight, so you can get the destination on time. Envoy Airlines (MQ) has relatively high cancellation frequency. September always has a low average departure delay. 5 AM has a low departure delay because there are less flights. Continental Airlines (CO) has a low departure delay and cancellation frequency which is pretty reliable. So, if your flight is CO carrier, during 5 AM in September. You don't need to worry about departure delay.

5 Contribution

The contribution of each group member is shown in Table 11.

Table 11

Name	Contribution
Chendan Tang	Introduction, Different Manufactures, Different Plane Age, Airport Network, LaTeX Editing
Chunlei Liu	Summary statistics for two years, different segments of departure delay, Abstract, Recommendation
Conrad	Different Airports, Different States, Summary
Yunan Shi	Cancellation trend, special cases analysis, Recommendation, Summary

6 Reference

Airbus. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Airbus

McDonnell Douglas. (n.d.). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/McDonnell_Douglas

Qu, A., & Zhu, J. (n.d.). ST578: Statistical Learning in Data Science Chapter 4: Statistical Network and Graphical Models.

US Department of Commerce, & Noaa. (2015, March 25). 1997 Sioux Falls Area Climate Summary. Retrieved from https://www.weather.gov/fsd/fsd1997

US Department of Commerce, & Noaa. (2018, August 14). A Review of Weather in 2002. Retrieved from https://www.weather.gov/tae/climate_2002review