STAT 480 Final Project

Spring 2018

Analysis of US Flight Data in 2000 and 2003



Ankush Agrawal

[1. Introduction 3](#_Toc536533758)

[2. Data Cleaning 4](#_Toc536533759)

[2.1 Downloading Data 4](#_Toc536533760)

[2.2 Airlines Data 4](#_Toc536533761)

[2.3 Airports Data 5](#_Toc536533762)

[2.4 Plane Data 5](#_Toc536533763)

[3. Preliminary Analysis 6](#_Toc536533764)

[3.1 Trend in Overall Number of Flights Across Years 6](#_Toc536533765)

[3.2 Trend in Overall Number of Flights Across Months 6](#_Toc536533766)

[3.3 Trend in Overall Number of Flights Across Days of the Week 7](#_Toc536533767)

[3.4 Major Characteristic Comparisons Among States 8](#_Toc536533768)

[3.5 Trends in Flight Movement Within and Outside Regions 9](#_Toc536533769)

[4. Analysis 10](#_Toc536533770)

[4.1 Cancellation Rates 11](#_Toc536533771)

[4.1.1 Preliminary Analysis : Comparison of Monthly Cancellation Rates (2000 vs. 2003) 11](#_Toc536533772)

[4.1.2 Comparison of Cancellation Rates by Carrier (2000 vs. 2003) 12](#_Toc536533773)

[4.1.3 Comparison of Monthly Cancellation Rates by Carrier (2000 vs. 2003) 14](#_Toc536533774)

[4.1.4 Comparison of Monthly Cancellation Rates by Region (2000) 16](#_Toc536533775)

[4.1.5 Comparison of Monthly Cancellation Rates by Region (2003) 18](#_Toc536533776)

[4.1.6 Cancellation Rate by Manufacturer (2000 vs. 2003) 18](#_Toc536533777)

[4.2 Flight Delays 20](#_Toc536533778)

[4.2.1 Trends of Arrival Delay 20](#_Toc536533779)

[4.3 Diverted Flights 24](#_Toc536533780)

[4.3.1 Trends in Number of Diverted Flights Across Time 24](#_Toc536533781)

[4.3.2 Trends in Diverted Flights by Destination 27](#_Toc536533782)

[4.3.3 Top 10 Routes with the Most Number of Diverted Flights: 28](#_Toc536533783)

[4.4 Flight Paths and Number of Flights 28](#_Toc536533784)

[4.4.1 Mapping Flight Paths 28](#_Toc536533785)

[4.4.2 Number of Flights By Airport and State 32](#_Toc536533786)

[4.5 Top 10 Airplane Models Used for Long and Short Duration Flights 35](#_Toc536533787)

[4.5.1 Long and Short Duration Airplane Models Overview 36](#_Toc536533788)

[4.5.2 Long and Short Duration Airplane Models in 2000 vs. 2003 38](#_Toc536533789)

# 1. Introduction

Our analysis is based upon the *airlines* data set from the US Department of Transportation’s Bureau of Transportation Statistics (BTS), more specifically the years 2000 and 2003. The dataset consists of flight data from major carriers in the United States, with variables like departure and arrival time of the flight (*Year, Month, DayofMonth, DayofWeek, DepTime, ArrTime*), information about the carrier (*UniqueCarrier*) and the actual plane used for the flight (*TailNum*), actual and scheduled elapsed time of the flight (*ActualElapsedTime, CRSElapsedTime*), flight delay length (*ArrDeley, DepDelay*), flight status (*Cancelled, Diverted*), and others.

Other relevant data sets are also utilized in the analysis to provide additional information for the flight:

The *airports* data set (*airports.csv*) from the Open Flights initiative (<http://openflights.org/data.html>) provides information about the location (*lat, long, city, state, country*) as well as iata code (*iata*) and name (*airport*) of the airports in the United States.

The *carriers* data set (*carriers.csv*) from the US Department of Transportations Bureau of Transportation Statistics (BTS) (<http://www.transtats.bts.gov/Download_Lookup.asp?Lookup=L_UNIQUE_CARRIERS>) provides information about the carrier code (*Code*) and name (*Description*) of the carrier.

The *plane-data* data set (*plane-data.csv*) from the FAA registration database (<http://www.faa.gov/licenses_certificates/aircraft_certification/aircraft_registry/releasable_aircraft_download/>) provides information about airplane models, including manufacturer (*manufacturer*) and model (*model*) of the plane, the plane’s tail number for identification purposes (*tailnum*), and other related variables.

First the data sets were cleaned of data errors before any analysis, followed by a preliminary analysis on the *airlines* data set, both of which are discussed in detail in the following sections. The preliminary analysis explores broad trends in some of the included variables, as well as general differences across the years 2000 and 2003. In the subsequent analysis and summary, variable relationships that are explored extensively include those between flight time and plane models, numbers of incoming flights and percentage delay in airtime, various data (cancellation trends, number of flights, etc.) grouped by airline company over years, number of outgoing flights by airport. The analysis also includes visualizations of all of these relationships and also of flight paths between and over both years.

Different tools learned in class were used for different parts of the data cleaning and analysis in the fashion of which tool is best suited for the situation. For example, Hive and big matrix objects were used in the data processing steps, as they are much faster than R to process large amounts of data. When it comes to analysis, R is mainly used for model building and visualization, as it has various packages and functions to provide a more thorough analysis.

# 2. Data Cleaning

## 2.1 Downloading Data

Data was downloaded for the year 2000 and 2003 and the two years appended together.

## 2.2 Airlines Data

Before proceeding with any data analysis project, a thorough understanding of the data set is required. An exploratory data analysis was conducted which helped in understanding how variables interact with each other, as well as any issues within each variable including missing values, erroneous data, etc. With that objective in mind, the *airlines* data was analyzed using Hive. It makes use of the Hadoop distributed file system and is a powerful tool to manipulate and wrangle data. Interesting insights into the data were discovered which formed the foundation of further analysis and visualizations in R. The procedure below was used to explore the *airlines* data set:

1. After loading the data, the first focus of analysis was finding the number of missing values in each variable of the dataset. As it turned out, the initial shell script to calculate it gave a result in stark contrast with the *Readme* file supplied with the *airlinesauxiliaryfiles.zip*. It underreported the number of missing values. Further inspection showed that missing values are represented via blanks (‘ ‘) and NA values as well. The code was adopted for these modifications and a new shell script was created to automate the calculation for each variable and output the result in a csv file.
2. The next focus was to understand why there are missing values in couple of variables. There are 288,959 flights in the overall data which are cancelled. This is roughly the same number of observations that are missing in *DepTime, ArrTime, ActualElapsedTime, AirTime, ArrDelay, DepDelay* (Please refer to the “Original Data” worksheet within the excel file). This was subsequently proved, and it makes intuitive sense that cancelled flights will not have values for all these variables. So, our data is comprised of cancelled flights and non-cancelled flights.
3. The shell script was run again to find the number of missing observations for each variable for the non-cancelled flights. The expectations was that there will be drastic reduction in the number of missing observations for the above mentioned list of variables. Our expectations were met; however we observed a common number of observations missing for non-cancelled flights. 25,635 number of observations are missing across *ArrTime, ActualElapsedTime, AirTime* and *ArrDelay*. Extracting all these cases in a CSV table and inspecting in Excel to hopefully find a common characteristic. What we found was that all these cases were for diverted flights. It again made intuitive sense as these values will not be populated for flights that are diverted (=1). So, our data comprises of 3 distinct set of flights:
   1. Diverted (representing a tiny share of overall data – 0.21%)
   2. Non-diverted, cancelled (representing 2.37% of overall data)
   3. Non-diverted, non-cancelled flights (representing majority of our data – 97.42%)
4. Now that we understood the reason behind missing values, the focus was on to find any erroneous data. To do this, a shell script to conduct univariate analysis on continuous variables was run. It calculated the minimum and maximum value along with percentile distributions. There were 3,957 flights which had negative values for *ActualElapsedTime, CRSElapsedTime* or *AirTime* between them. All such cases were pulled out in Excel to see if there is a common characteristic between them but unfortunately no such finding was observed. So we removed all such erroneous observations from subsequent analysis as they do not make sense and this comprised only 0.03% of overall data.
5. There are variables which captures arrival delay (*ArrDelay*) or departure delay (*DepDelay*) but there is no variable which captures the overall delay for a flight. We tested the hypothesis if the difference between actual elapsed time of a flight (*ActualElapsedTime*) and scheduled elapsed time of a flight (*CRSElapsedTime*) is equal to difference between arrival delay and departure delay. In all but 191 cases we see this to be true. Since the vast majority of the data confirms that overall delay is the difference between actual elapsed time and scheduled elapsed time, we created a variable *overall\_delay* to capture the same. The difference is then converted to percent difference and is further explored in detail in Section 4.2.1

## 2.3 Airports Data

In preparing the supplementary file *airports.csv*, several data errors were identified. The data was only needed for a specific join accomplished through Hive, and only the *state* and *iata* variables were used in our eventual analysis. The errors were limited to incorrect *state* values for a handful of airports (‘NA’s or a city name) which became apparent during the treemap analysis of outgoing flights by airport seen in the analysis section below. No other errors in the subset of *airports* were identified.

## 2.4 Plane Data

The *plane-data* set consist of airplanes characterized by their tail number, with other related information provided. While exploring the data set, we discovered that out of the total 5,029 entries, 549 of them only have a tail number without any other information about its manufacturer, model, etc., leaving only 4,480 lines of valid data. With the help of Hive, a new table is created with the valid 4,480 rows of data to be further joined with the *airlines* data set for further analysis.

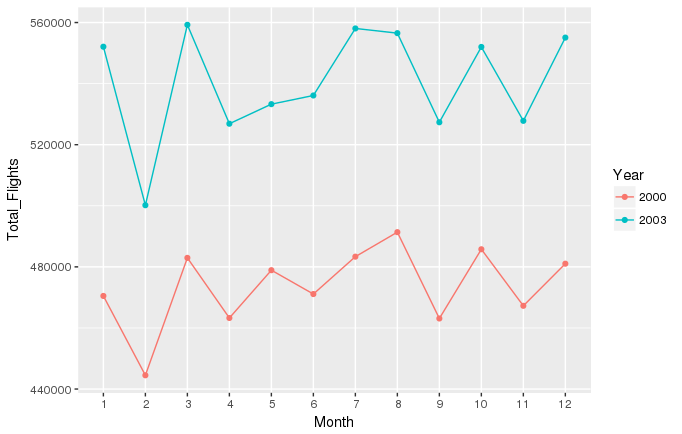
# 3. Preliminary Analysis

Following data cleaning, a preliminary data analysis was undertaken to become familiar with the data set and to explore general trends between certain variables, more specifically the number of flights. This leads into our more in-depth formal analysis in the section 4.

## 3.1 Trend in Overall Number of Flights Across Years

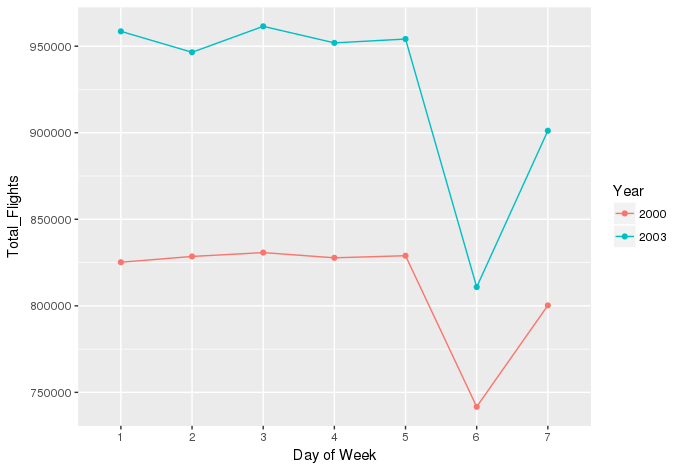
To get a general idea of the volume of flights in the data set, we identified 5,683,044 flights in 2000 and 6,484,586 flights in 2003. There is an increase in the number of flights by 14.1%. This may be explained by a general growth in the commercial airline industry and commuter base. In total, there are roughly 12 million flights between the 2 years.

## 3.2 Trend in Overall Number of Flights Across Months



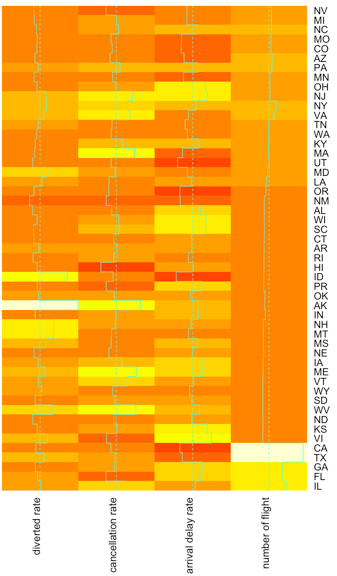
The same metric was explored but by month. It was observed that the total number of flights across different months for 2000 and 2003 follow a similar trend. The total number of flights peak during the summer months (June to August) with August being the highest. One possible reason for the increased number of flights is that more travelers travel out during these months on summer vacation. September and November are lean months in terms of overall number of flights whereas there is substantial increase in the winter months of January and December. Again, one reason for this is that December and January hosts Christmas and New Year, the two biggest events when people take time off and travel. One exception to this is March which is equally busy month as the summer and winter months. This could be due to Spring break when also a lot of people travel. In both the years, February has the least number of flights as it lacks popular traveling holidays or events. Also winter is at its peak with winter storms a common phenomenon which could also explain the decrease. It sees an average drop of 7.61% from the previous month (January) for the year 2000 and 2003. The drop is more pronounced for 2003 than 2000 where it is 9.4% compared to 5.5% in 2000.

## 3.3 Trend in Overall Number of Flights Across Days of the Week



We observe that there are roughly the same overall number of flights throughout the weekdays. However, there is a big drop over Saturday. This may be because people generally prefer to either travel on Friday or Sunday and spend their entire weekend on vacation or at home. This also explains why number of flights bounce back on Sunday when people want to fly back for the new week. Those who fly for business reasons also would fly more often during workdays.

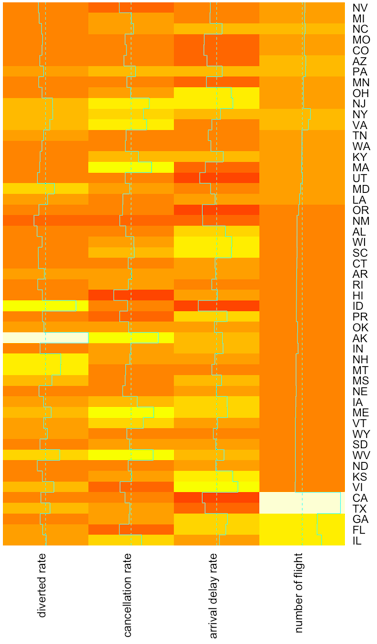
## 3.4 Major Characteristic Comparisons Among States

****

The heatmap above compared total number of flights, diverted rate, cancellation rate, and arrival delay rate among different states in the United States in year 2000. The color of the boxes represents the value of the variable, and deeper color means smaller value. Looking at the heatmap, we can see that the states with high value of total number of flights are CA and TX. There is no state with extreme small values of number of flight. The states with high arrival delay rates seems to be VT, WY, and ID, and states with small arrival delay rates seems to be MN, MI and TN. The states with high cancellation rate seems to be MA, VA, and IL, and a state with low cancellation rate seems to be NM. The states with high diverted rate seems to be WY and AK, and the states with low diverted rate appears to be NV and NM.

Just based on the heatmap above, we can't observe any obvious trends between number of flights, cancellation rates, arrival delay rates, and diverted rates. The main reason might be that there are major differences between air traffic condition of different states, and there might not be a general rule that applies to most of them.

States that are most similar with respect to these statistics are CA and TX as they all have large number of flight, and medium number of diverted rate and cancellation rate. Also, MN and TN are very similar, since they all have small number of arrival delay rate, and their diverted rate, cancellation rate and number of flight are relatively close to each other. States that are the most different with respect to these statistics are CA and WY. We can clearly observe from the heatmap that they have opposite colour among the four measured variables.



The heatmap above compared total number of flights, diverted rate, cancelled rate, and arrival delay rate among different states in US in year 2003. The color of the boxes represents the value of the variable, and darker color means smaller value. Looking at the heatmap, we can see that the states with high value of total number of flights are CA and TX. There is no state with extreme small values of number of flight. The states with high arrival delay rates seems to be OH and NJ, states with small arrival delay rates seems to be UT, OR and CA. The states with high cancellation rate seems to be OK and WV, and the states with low cancellation rate seems to be HI and NV. The states with high dicerted rate seems to be AK, and the states with low diverted rate seems to be NV.

As we mentioned before, we still can't observe any obvious trends between number of flights, cancellation rates, arrival delay rates, and diverted rates.

States that are most similar with respect to these statistics are CA and TX as they all have large number of flight, small number of arrival delay rate, and medium number of diverted rate and cancellation rate. Also, UT and OR are very similar, since they all have small number of arrival delay rate, and their diverted rate, cancellation rate, and number of flights are relatively close to each other. States that are the most different with respect to these statistics are CA and AK. We can clearly observe from the heatmap that they have opposite colors among the four measured variables.

By comparing heatmaps for those two years we can see that CA and TX have large decreases in arrival delay rate in 2003. However, most of the states stay at the same rate as 2000, or change slightly.

## 3.5 Trends in Flight Movement Within and Outside Regions

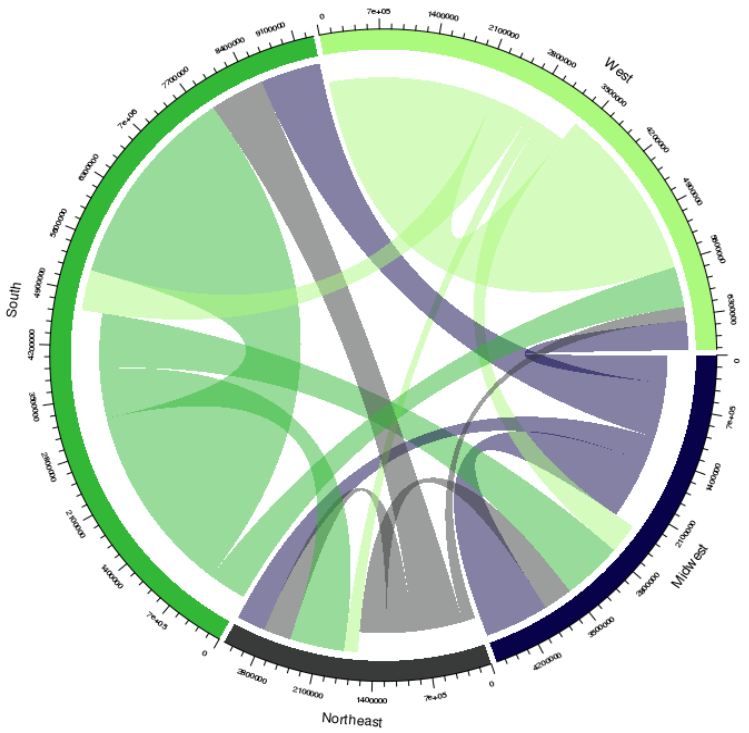
To perform this analysis, states were grouped into regions based on the definition provided by Census Bureau-designated regions and divisions.

(<https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States>)

They are as follows:

1. Northeast : Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, New Jersey, New York, and Pennsylvania
2. Midwest : Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota, Illinois, Indiana, Michigan, Ohio, and Wisconsin
3. South : Delaware, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, District of Columbia, West Virginia, Arkansas, Louisiana, Oklahoma, Texas, Alabama, Kentucky, Mississippi, and Tennessee
4. West : Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming, Alaska, California, Hawaii, Oregon, and Washington

The south region has the maximum number of flights within the region whereas northeast has the minimum number of flights within the region. Northeast has the highest proportion of flights flowing out of the region. Midwest has a fair share of within the region and outside the region flights.

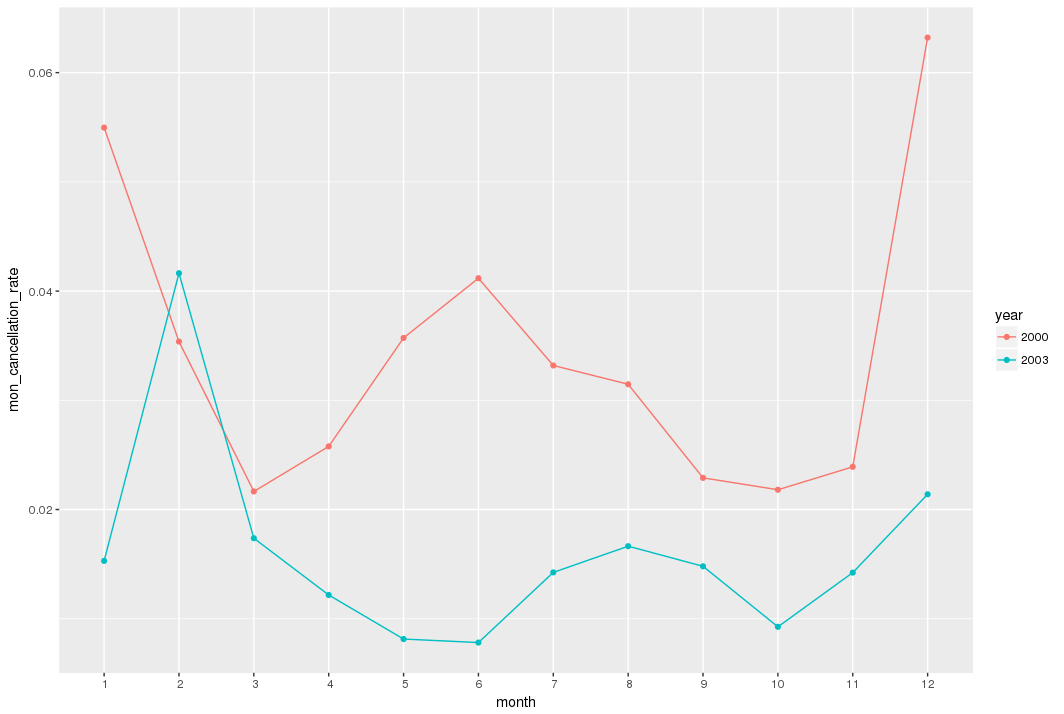


# 4. Analysis

This section includes our detailed observations and analysis of several variables and those related to them, which include: cancellation rates, measures of flight delay, diverted flights, number of outgoing flights, flight paths, arrival times of flights, and airplane models.

## 4.1 Cancellation Rates

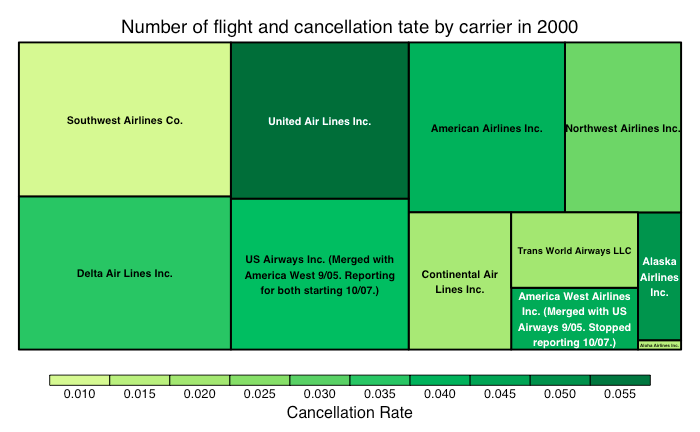
### 4.1.1 Preliminary Analysis : Comparison of Monthly Cancellation Rates (2000 vs. 2003)



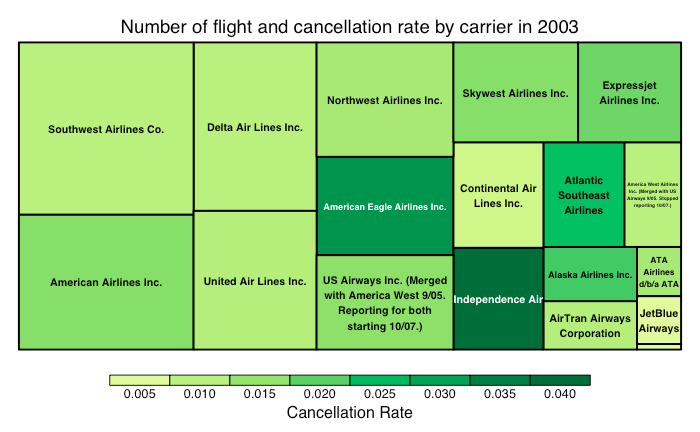
The monthly cancellation rates for 2000 and 2003 provide interesting insights into the 2 years for air travel. As seen above, the monthly cancellation rates for 2000 are higher than 2003 for all the months except February. February also happens to be the month when cancellation rate is the highest for any month in 2003 by almost a factor of 2.

Our research showed us that three powerful snowstorms happened in these two years which caused a hike in cancellation rate. They affected different parts of the country in January 2000, December 2000, and February 2003, which had a devastating effect on the airline cancellations during these times. This is evident from the data as shown in the plot above. More details will be provided in the subsequent section where we analyze cancellation rate by region.

### 4.1.2 Comparison of Cancellation Rates by Carrier (2000 vs. 2003)

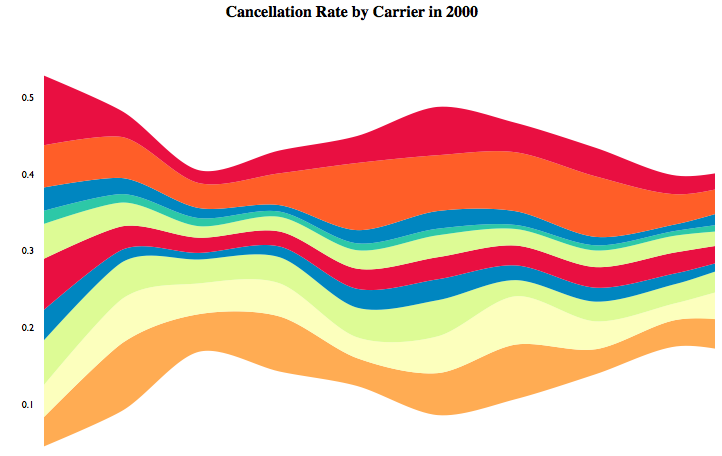
****

We analyzed the cancellations by carriers based on the cancellation rate and total number of flights to see which airlines seem to have the most issues in 2000. First thing we can see from this map is that the range of cancellation rate for all the airline companies is very small, ranging from 0.01 to 0.055. This means that most of the flights are not cancelled in year 2000 no matter what airline company it is. Furthermore, it’s clear that United Airlines Inc. had the highest cancellation rate. The airline with the lowest cancellation rate was Aloha Airlines Inc. Some reason for the cancellation rates can be attributed to the fact that United Airlines Inc. seems to be more popular and serves more patrons than Aloha Airlines Inc. This is reasonable, since the more people uses an airline, the more problems are likely to occur. However, this role doesn't apply to all the boxes in this graph. For example Southwest Airlines Co. is also a popular airline company, but the cancellation rate of it is relatively low. On the other hand, Alaska Airlines Inc. is not very popular but it has a relatively high cancellation rate.



We analyzed the cancellations by carriers based on the cancellation rate and total number of flight to see which airline seems to have the most issues in 2003. Again we see the cancellation rate is very small, between 0.005 and 0.04. This means that most of the flights are not cancelled in year 2003 regardless of airline, and most of the airline have lower cancellation rate than 2000. Furthermore, it’s clear that Independence Air had the highest cancellation rate. The airline with the lowest cancellation rate was JetBlue Airways. The most popular airline companies such as Southwest Airlines Co. and American Airlines Inc are doing relatively well, and are slightly different from the similar analysis in 2000. We can also see that there are more airline companies in 2003 than 2000, which might lead to the decreases of market shares of previous popular airlines such as United Air Lines Inc. The increases of the number of airline might also introduce more competition between airline companies, and that might be the reason that most of them do better in reducing cancellation rate than 2000.

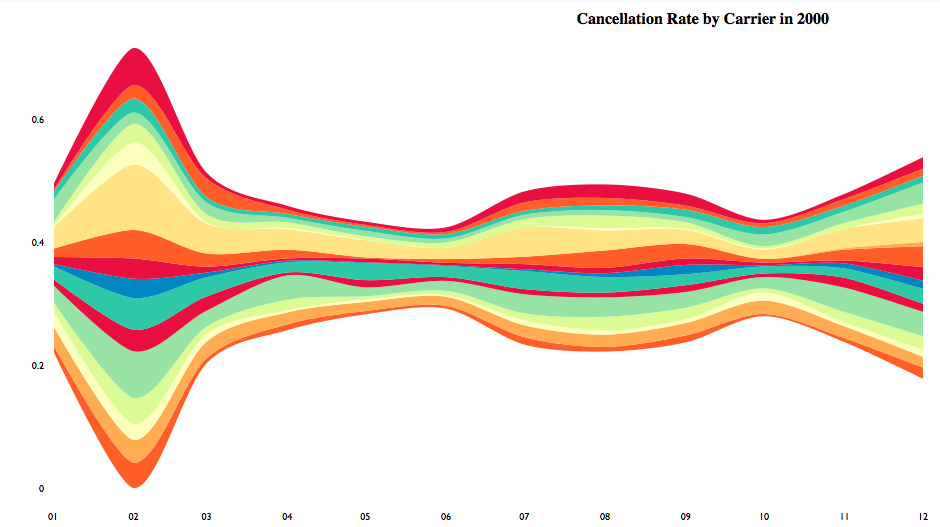
### 4.1.3 Comparison of Monthly Cancellation Rates by Carrier (2000 vs. 2003)

****

Top to bottom: US Airways Inc., United Airlines Inc, Trans World Airways LLC, Southwest Airlines Co, Northwest Airlines Inc, Delta Air Lines Inc, Continental Air Lines Inc, American Airlines Inc, America West Airlines Inc., Alaska Airlines Inc..

Note: Streamgraphs are interactive, but since we are using a Word document, these features aren’t available and we are sharing a still image.

The graph above represents the cancellation rates by month of different carriers in 2000. Looking at the graph, we can see that the cancellation rates increased in summer and winter. One reason could be that winter and summer are generally vacation time and therefore, more flights would be booked and it might cause some problems for air traffic. We can also see that US Airways, United Airlines, America West Airlines Inc, and Alaska Airlines took a major proportion of cancellation rate among all the airline companies, and all the other airlines companies have a relatively smaller cancellation rates.

Top to bottom: US Airways Inc., United Airlines Inc, Southwest Airlines Co, Skywest Airlines Inc, Northwest Airlines Inc, JetBlue Airways,Independence Air,Expressjet Airlines Inc, Delta Air Lines Inc, Continental Air Lines Inc, Atlantic Southeast Airlines,ATA Airlines, American Eagle Airlines Inc, American Airlines Inc, America West Airlines Inc., Alaska Airlines Inc., AirTran Airways Corporation.

The graph above represents the cancellation rates by month of different carriers in year 2003. We can still see similar trends compared to 2000 that cancellation rates in winter times and summer times tend to be higher, due to more air traffic during vacation times. We can also see that cancellation rates increased dramatically in February 2003. Our research showed that there was a severe blizzard from February 14 to February 19, 2003. It spread heavy snow across the major cities of the Northeastern and Mid-Atlantic states, and that may be a major factor that led to a huge increase of flight cancellation. When looking into each carrier, we can see that Independence Air and American Eagle Airlines Inc. had relatively higher values of cancellation rates. All other airlines are relatively homogenous.

### 4.1.4 Comparison of Monthly Cancellation Rates by Region (2000)

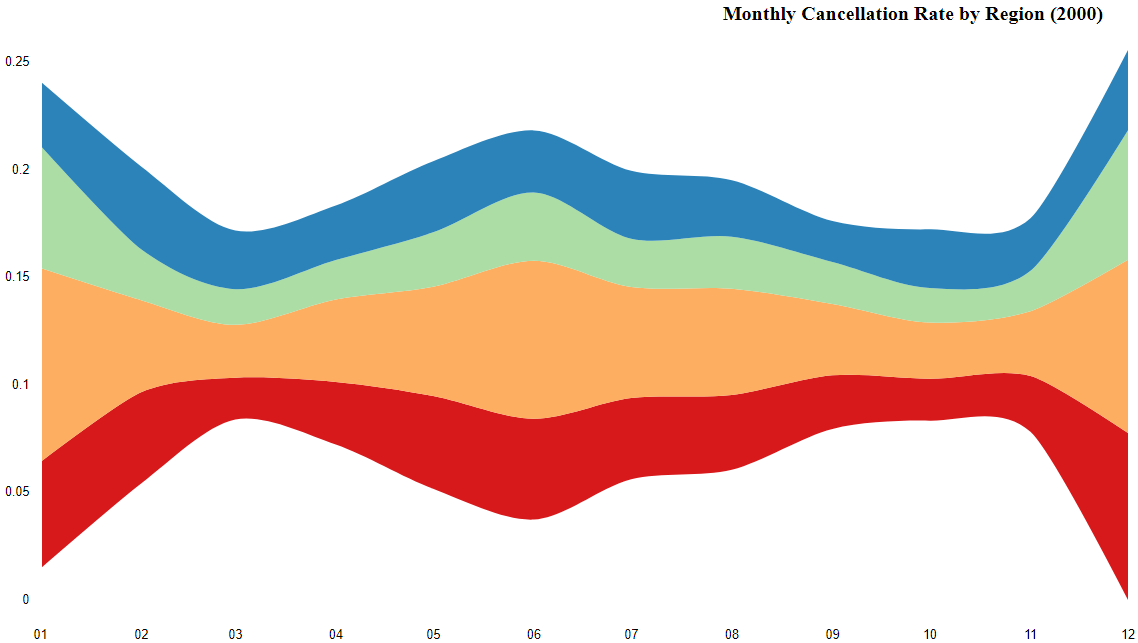
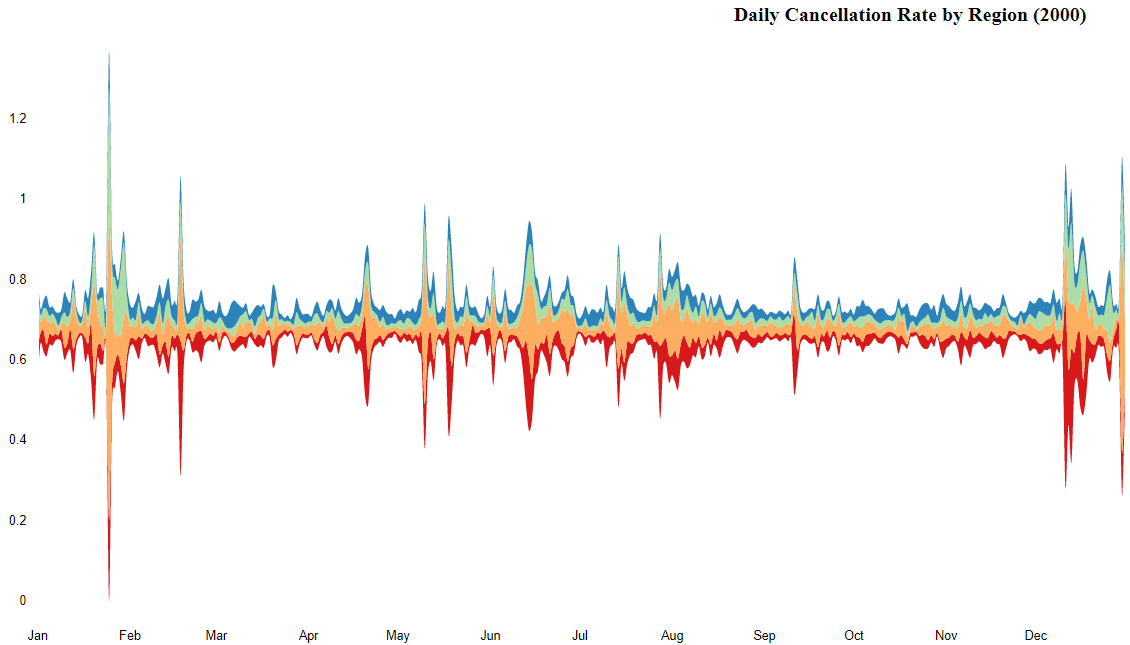
****

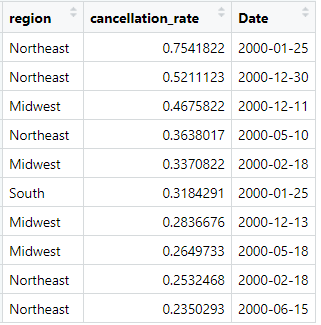
Figure 1: Monthly Cancellation Rate by Region: Blue represents West, green represents South, orange represents Northeast, red represents Midwest. Analysis is below.

Figure 2: Daily Cancellation Rate by Region: Blue represents West, green represents South, orange represents Northeast, red represents Midwest. Analysis is below.

Note: We decided to produce monthly as well as daily streamgraphs to analyze the cancellation rate by region. We did this because monthly cancellation rates smooth out, whereas the daily cancellation rates show the spikes in cancellation rates which lead to interesting insights about historic events in specific regions.

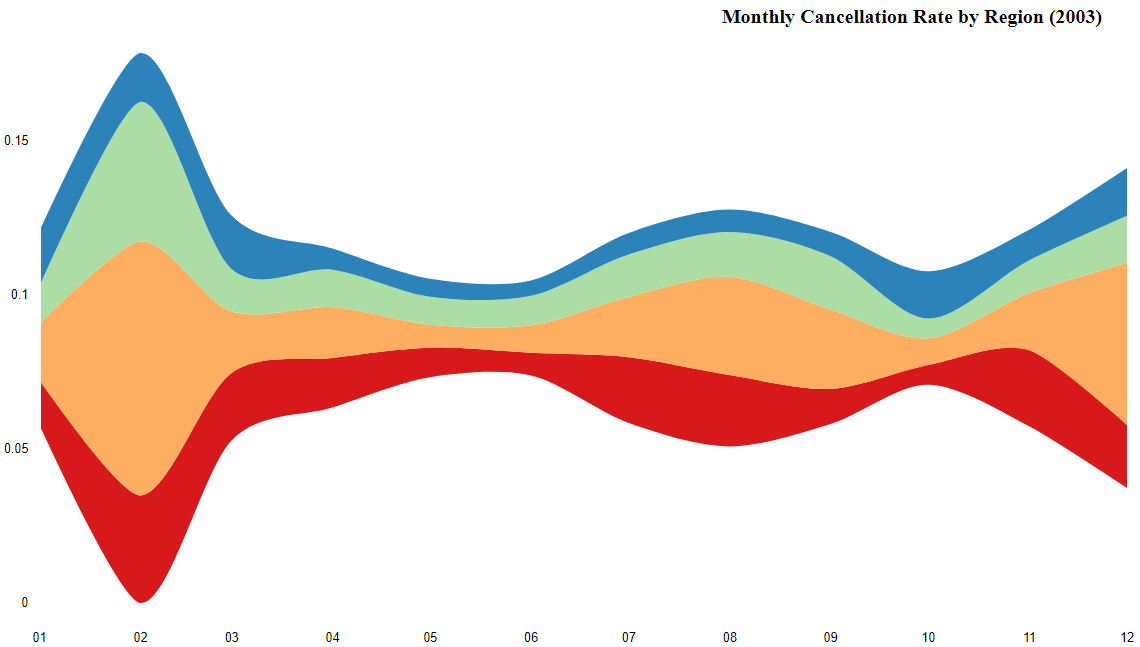
Monthly Cancellation Rates (Figure 1): The observations are in line with expectations. Winter months have the highest rate of cancellation due to snowstorms and blizzards. This is why the monthly cancellation rates are quite high at the start (January 2000) and the end (December 2000). There are some high rates of cancellation in the middle of the year. One possible reason could be the higher rush in the summer might have led to mismanagement and crowds and subsequent cancellation of flights. Another possibility is an active hurricane cycle, but our research did not show a very prominent one striking United States during this interval.

Daily Cancellation Rates (Figure 2): We observe that there is a very big spike in the cancellation rate towards the end of January and one towards the end of December. The top 10 days for the highest cancellation rates in 2000 are as follows:



Based on our research, the dates with high cancellation rates coincide with some of the worst snowstorms in the history of United States. In the early part of the year, a powerful blizzard struck the eastern part of the nation on January 25 which led to high cancellation rates (<https://en.wikipedia.org/wiki/January_2000_North_American_blizzard>). December 2000 witnessed another storm which struck the Northeast from the 27-30, causing high rates of cancellation (<https://en.wikipedia.org/wiki/December_2000_nor%27easter>).

### 4.1.5 Comparison of Monthly Cancellation Rates by Region (2003)

****

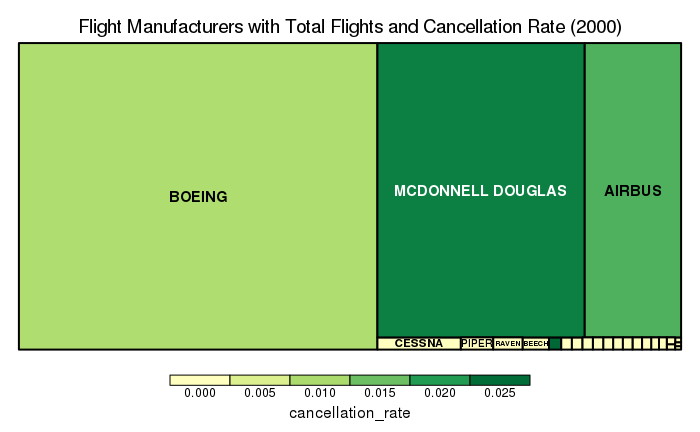
Monthly Cancellation Rate by Region: Blue represents West, green represents South, orange represents Northeast, red represents Midwest.

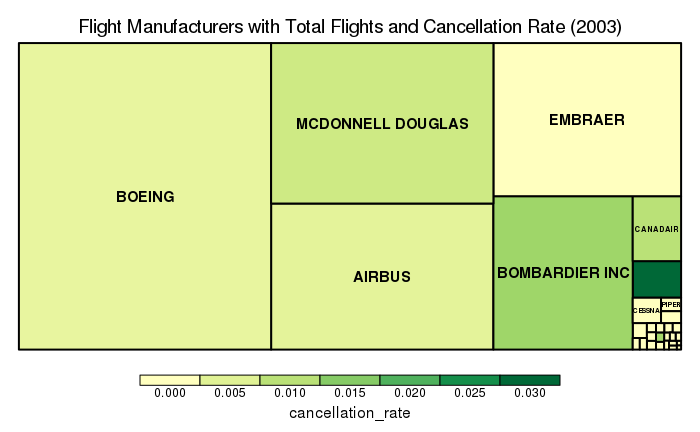
Unlike 2000, the monthly cancellation rate streamgraph for 2003 needed no further granularity at a daily level to see significant cancellation rate through the year. As we observe, February 2003 stands out from the rest of the months with the highest rates of cancellation for every region. As we discussed in the preliminary analysis and further in the cancellation rates by carrier section, this is because of the Blizzard of 2003 also known as Presidents’ Day Storm (<https://en.wikipedia.org/wiki/North_American_blizzard_of_2003>). It completely shut down Reagan National Airport, Baltimore-Washington International Airport, Philadelphia International Airport, and LaGuardia Airport leading to high cancellation rates in the Northeast.

### 4.1.6 Cancellation Rate by Manufacturer (2000 vs. 2003)

The join between the *airlines* dataset with *plane-data* dataset was quite disjoint. Of the roughly 12 million observations in the *airlines* dataset, only ~5.5 million observations could find corresponding information in the *plane-data* dataset. This was only 46% of the overall dataset. This is a limitation while analyzing the cancellation rate by manufacturer.

Also, we observed variations within the names of the manufacturers. For instance “McDonnell Douglas” has variations like “MCDONNELL DOUGLAS AIRCRAFT CO”, “MCDONNELL DOUGLAS CORPORATION”, and “DOUGLAS”. It was found that Boeing acquired McDonnell Douglas which explains the variations in the name of the planes. Since they represent the same entity and to keep consistency with the same manufacturer, these were combined under same manufacturer “McDonnell Douglas”. Similarly, there were two variations for Airbus and Aerospatiale both which were all renamed as Airbus and Aerospatiale respectively.





We see that the top 3 manufacturers control the major share of flights (98.17%) which are flown in the 2000. In 2003, that ratio decreases to 71.66% as other manufacturers enter as contending players in the industry like Embraer and Bombardier. The new entrants eat into the share of the Big 3 - Boeing, McDonnell Douglas, and Airbus and control a 24.69% share of the flights flown based on their planes.

With respect to cancellation rate, we observe that McDonnell Douglas had the highest cancellation rate for the flights flown by major planes manufacturers in 2000 at 2.29%, followed by Airbus at 1.68%. At an overall level, Diamond Aircraft Industries Inc. was the worst performer in terms of cancellation rate in 2000 (with 2.52% flights cancelled). However, it’s planes contributed a miniscule number of 1,630 flights, which is why we observe a small dark green box in the treemap.

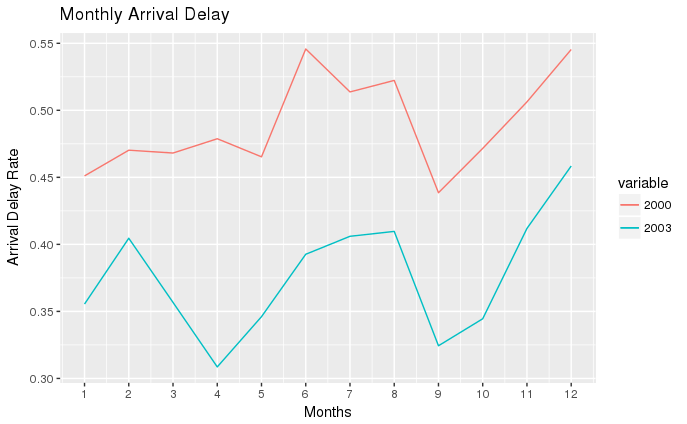
All the major plane manufacturers of 2000 improve their cancellation rate in 2003. Bombardier unseated McDonnell Douglas to have the highest cancellation rates for the flights flown by major plane manufacturers in 2003. At an overall level, Aerospatiale has the worst cancellation rate in 2003 with 2.97% of their flights cancelled.

## 4.2 Flight Delays

Delayed flights are one major concern for airline passengers. In order to predict the amount of delay to bring insight to what to avoid while travelling, we have analyzed the relationship between flight delays and the busyness of the airport, defined by number of incoming flights at the destination airport, and the relationship between flight delays and the airplane carrier.

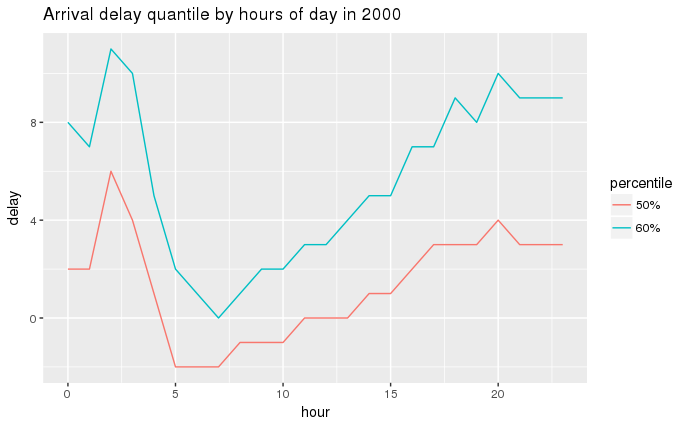
### 4.2.1 Trends of Arrival Delay

Besides using the percent difference between actual elapsed time and planned elapsed time, another measure for delay rate was calculated by the number of delay flights divided by the total number of flights

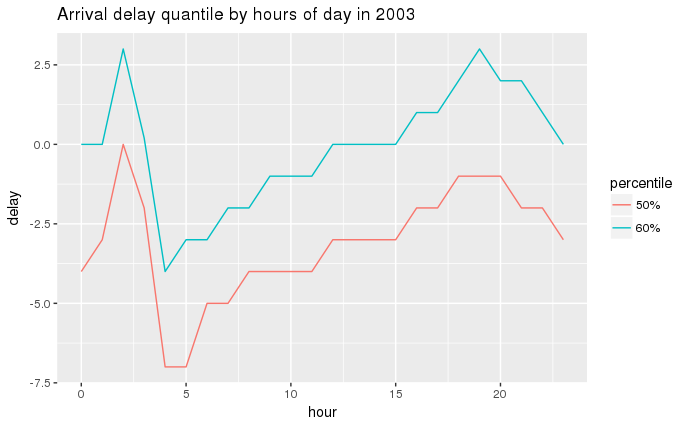
****

The graph above shows us the monthly arrival delay rate in year 2000 and 2003. As seen above, the monthly arrival delay rates for 2000 are higher than 2003 for all the months. Summer times and winter times generally have higher arrival delay rate. Our research shows us that summer months are peak times for vacations and thunderstorms. There is a chance that hotter and thinner air makes it more difficult for planes to get enough lift, which might cause delay (<https://www.nytimes.com/2017/08/07/business/summer-flight-delays.html>).

Flights in winter times are more likely to be affected by storms or blizzards, which makes it more difficult to take off or land. As discussed previously, there were heavy storms in January 2000, December 2000, and February 2003. This might be the reasons that we can observed higher arrival delay rates in those months.

****

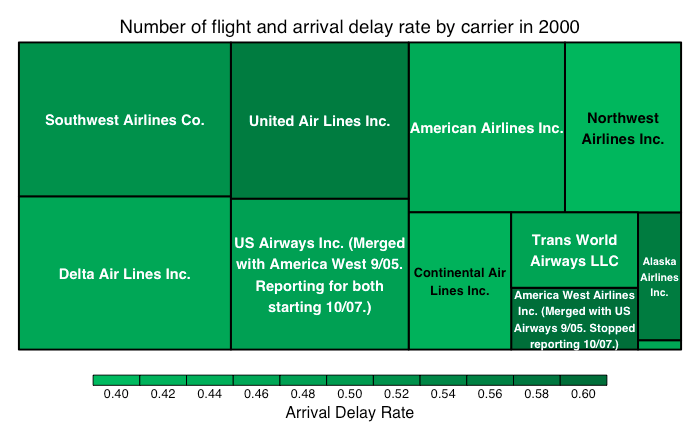
In order to see the amount of delay by hours in 2000, we can look at the delay quantiles of 50% and 60%. Based on the graph above, looking at 50% quantile, hour 2 has 6 minutes of delay, which means that at hour 2 at least 50% of the flight are 6 minutes late. The same rule applies to the 60% quantile. Therefore, we can say that the worst arrival delay occurs at 2:00 a.m. Similarly the best on time or early arrival happens between 5:00 a.m. to 7:00 a.m. We can also observe that after 2:00 a.m. the arrival delay time decreases until 6:00 a.m, and slowly increases again up till 11:00 p.m. or midnight, and then increases dramatically from 1:00 a.m. to 2:00 a.m.



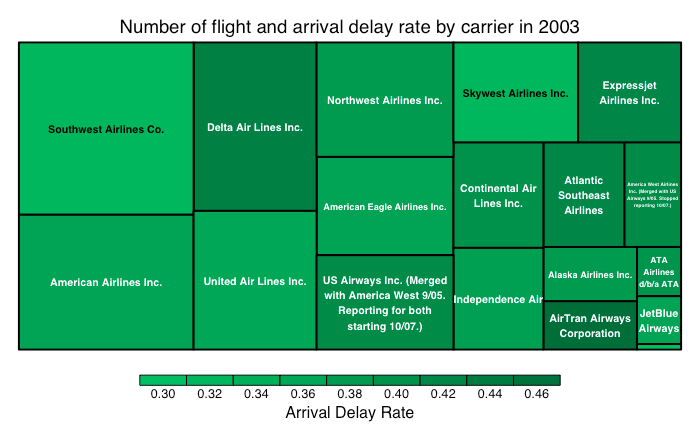
Similarly, in order to see the amount of delay by hours in 2003, we can look at the delay quantiles on 50% and 60%. Based on the graph above, if we are looking at 60% quantile, hour 2 has 3 minutes of delay. This means that at hour 2 at least 40% of the flight are 3 minutes late. The same rule applies to the 50% quantile. Therefore, we can say that the worst arrival delay happens at 2:00 a.m. Similarly the best ontime or early arrival happens between 4:00 a.m.to 5:00 a.m. We also observed a similar trends as seen in 2000 that after 2:00 a.m., the arrival delay time decreases until 4:00 a.m., and slowly increases again until 8:00 p.m., after that it decreased slowly from 8:00 p.m.to midnight, and significantly increased from midnight to 2:00 a.m.

Comparing year 2000 and 2003, we can see that the trends of the hourly delay rates are very similar. The worst arrival delay appears around 2:00 a.m. and the best ontime arrival appears around 5:00 a.m. in both years. Also, 2003 generally has lower hourly arrival delay than 2000.

### 

****

We analyzed the arrival delay by carriers based on the arrival delay rate and total number of flights to see which airline had the most issues in 2000. The first thing we can see from this map is that the range of arrival delay rate is between 0.4 and 0.6. This means that a significant amount of the flights are delayed regardless of airline. Furthermore, it’s clear that United Air Lines Inc. had the highest arrival delay rate. The airline with the lowest arrival delay rate was Aloha Airlines Inc (lower right box, the space is too small to display the name). There is a trend that bigger boxes tends to have darker colors, which means that popular airlines has higher rate of arrival delay rate. This is reasonable since popular carriers, with more planes to operate, are more likely to run into issues that could cause delay. However, America West Airlines Inc. also had a high arrival delay rate but with a relatively smaller number of flights. We can still see a relation between number of flights and arrival delay despite some exceptions.

****

We analyzed the arrival delay by carriers based on the arrival delay rate and total number of flights to see which airline has the most issues in 2003. The first thing we can see from this map is that the range of arrival delay rate is between 0.3 and 0.46. This means that the arrival delay rates are still high regardless of airline. Furthermore, it’s clear that Delta Air Lines Inc. and Airtran Airways Corporation had the highest arrival delay rates. The airline with the lowest arrival delay rate was JetBlue Airways. The relationship between the size and the color of boxes is not obvious here. High-traffic airline companies do not always have the higher arrival delay rate.

Comparing to the analysis we did for 2000, we can see that the range of arrival delay was smaller, which means that most of the airlines companies seemed to decrease their arrival delays. The increase of the number of airline companies might bring more competition in the industry, which might force airline companies to do better in reducing delay rates. Also, the relation between number of flights and arrival delay rate is not obvious. The reason could be that popular airline companies spent a lot more effort on decreasing their arrival delay rate when compared to the map from 2000.

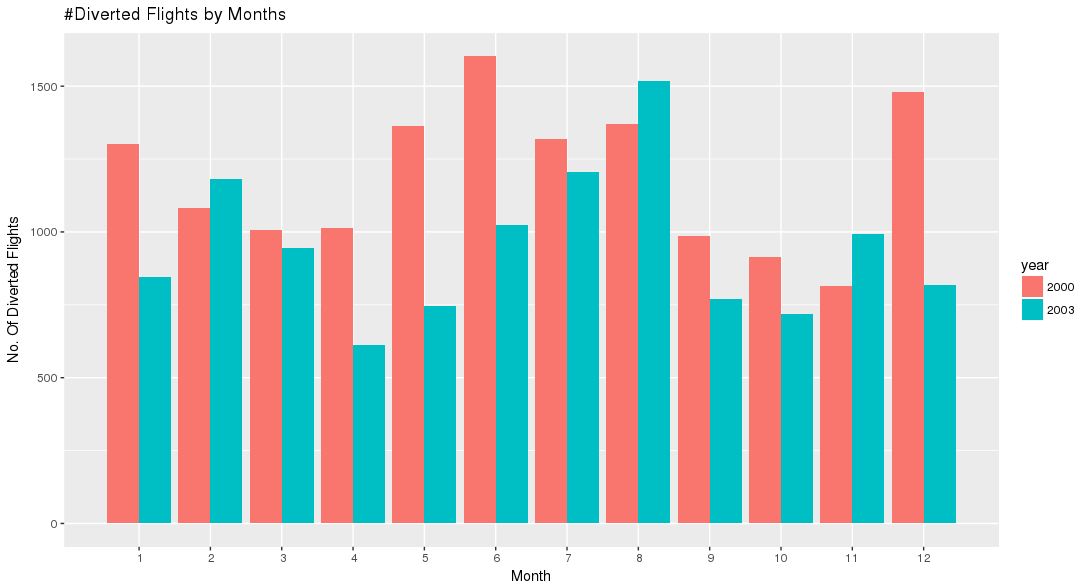
## 4.3 Diverted Flights

### 4.3.1 Trends in Number of Diverted Flights Across Time

So far we have analyzed delayed and cancelled flights, and now we want to focus on diverted flights in this section. The total number of diverted flights in both the years is included in the following table:

|  |  |
| --- | --- |
| Year | Number Diverted Flights |
| 2000 | 14,254 |
| 2003 | 11,378 |
| %YoY Change | -20.17 |

There is a 20% reduction in the number of diverted flights. So it looks like the commercial airline industry reduced the number of diverted flights overall.

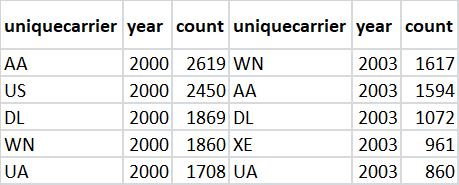


The number of diverted flights is the highest for the month of June in 2000 and for August in 2003. Our research did not find any particular event influencing large amount of flight diversion for June 2000. However, the high amount of flight diversion in December 2000 and January 2000 could be attributed to snowstorm discussed earlier in the cancellation rate section.

**Top Airline Carriers with the Most Diverted Flights**

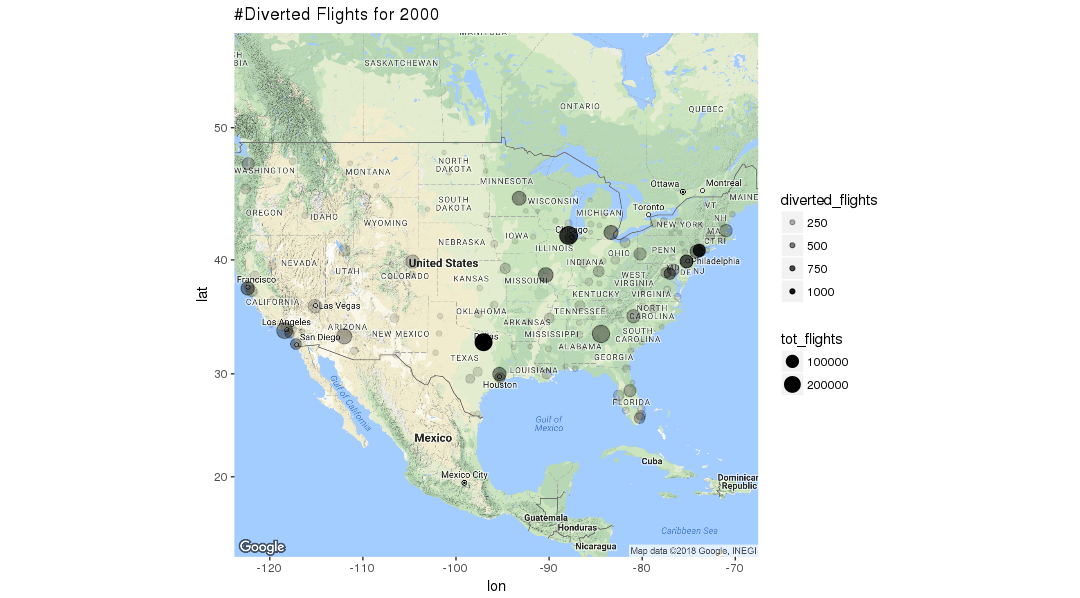
****

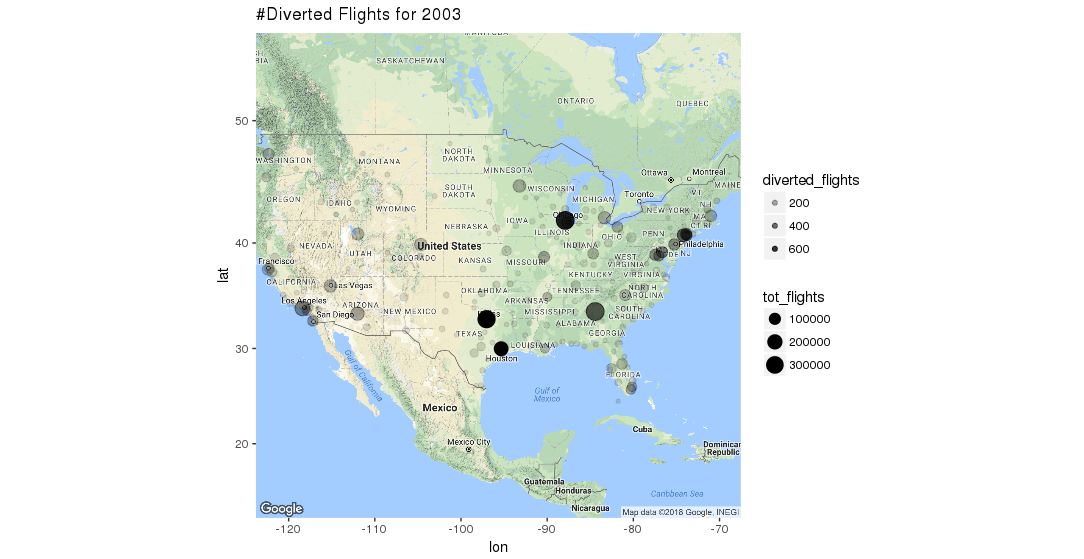
**Top 5 Carriers by Number of Diverted Flights**

****

US Airways (Carried Code: US) reduced the number of diverted flights by an extreme 75% to not figure in the top 5 airline carriers with the most diverted flights in 2003. The rest of the 4 airline carriers improve their tally but still continue to form the bulk of diverted flights in 2003. Express Jet made to the top 5 airline carriers with the most diversion for the year 2003, having no previous record in 2000. America West Airlines Inc. (Carrier Code: HP) happens to have only slightly improved its number of diverted flights (from 275 in 2000 to 270 in 2003).

### 4.3.2 Trends in Diverted Flights by Destination

****

****

Note: For this analysis, destination airports were analyzed assuming the diversion of a flight enroute is related to destination rather than origin airport. If there were any issues with origin airport, the flight would have been cancelled rather than being diverted. During in-air emergencies, flights can be diverted to the nearest landing strip, but due to lack of such data it was assumed that analyzing destination airport will make more sense.

The total number of diverted flights were plotted for transparency while the size reflected the traffic (overall volume) of non-cancelled and non-diverted flights in the year 2000 and 2003.

It is worth taking note that the busiest airports have the most amount of flight diversion. These also happens to be big business hubs and the pattern is consistent with the population distribution of the United States. Most of the flight diversion is concentrated in Northeast (New York, Philadelphia, Newark), South (Houston, Dallas-Fort Worth, Atlanta), Midwest (Chicago, St Louis) and West coast (San Francisco, Los Angeles). Historical events related to possible reasons for this phenomena were not found, so the reasons remain dubious.

### 4.3.3 Top 10 Routes with the Most Number of Diverted Flights:



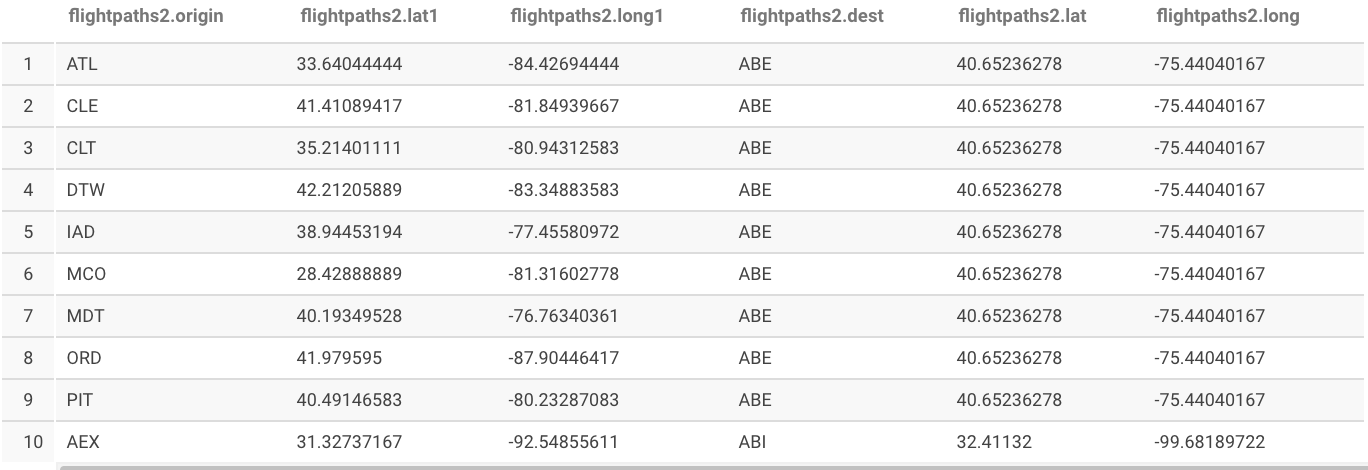
LaGuardia and Dallas/Forth Worth have the most number of diverted flights across the 2 years. When comparing the years, we found little difference between the rankings.

## 4.4 Flight Paths and Number of Flights

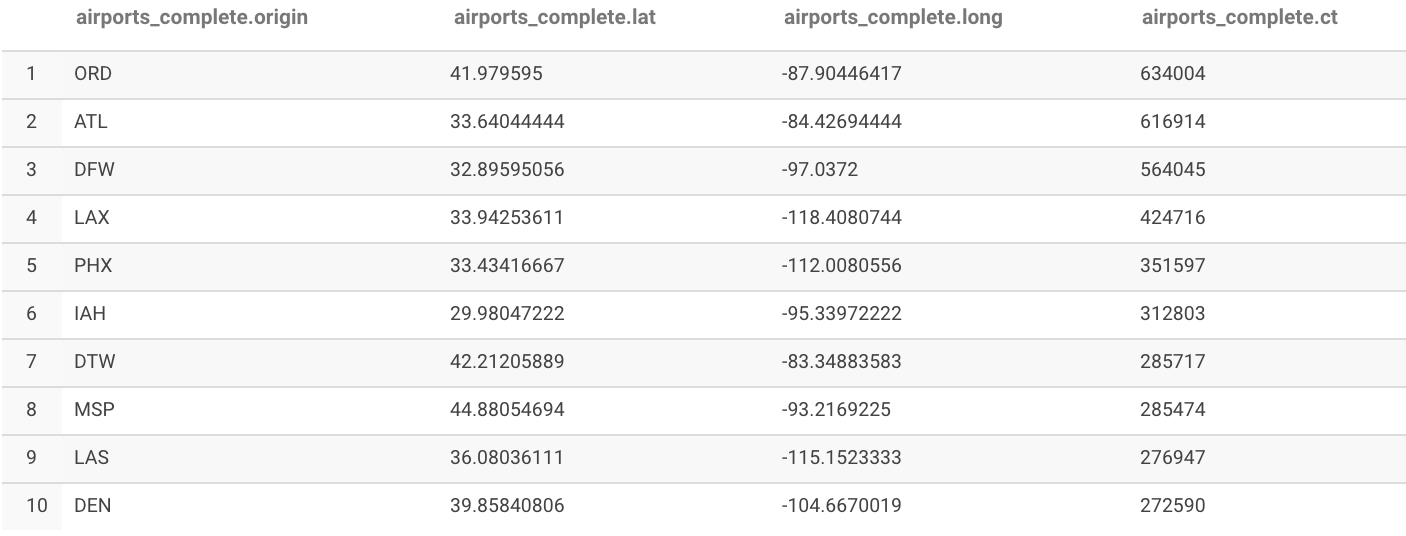
### 4.4.1 Mapping Flight Paths

In the section below we explore, compare, and contrast the total number of flights and their paths by year, state, and airport.

We constructed from the *airlines* data several Hive tables that were useful in visualizing the frequency and direction of individual flight paths. From the Hive table of the original cleaned *airlines* data, the unique combinations of the variables *origin* and *destination* were extracted and joined with *lat* and *long* variables from the *airports* data set in order to get the coordinates for each airport for each possible flight path. Two joins were necessary to get the coordinates for both airports. This was done individually for each year as there could be differences in the specific routes completed in a given year. The count of the number of outgoing flights per airport was also calculated in a separate table that is also utilised for the following flight paths maps. Outgoing flights were used as a measure of an airports’ busyness and level of air traffic, as it was assumed this number would be in proportion to the number of incoming flights by airport as well. The first new Hive table is the format below:

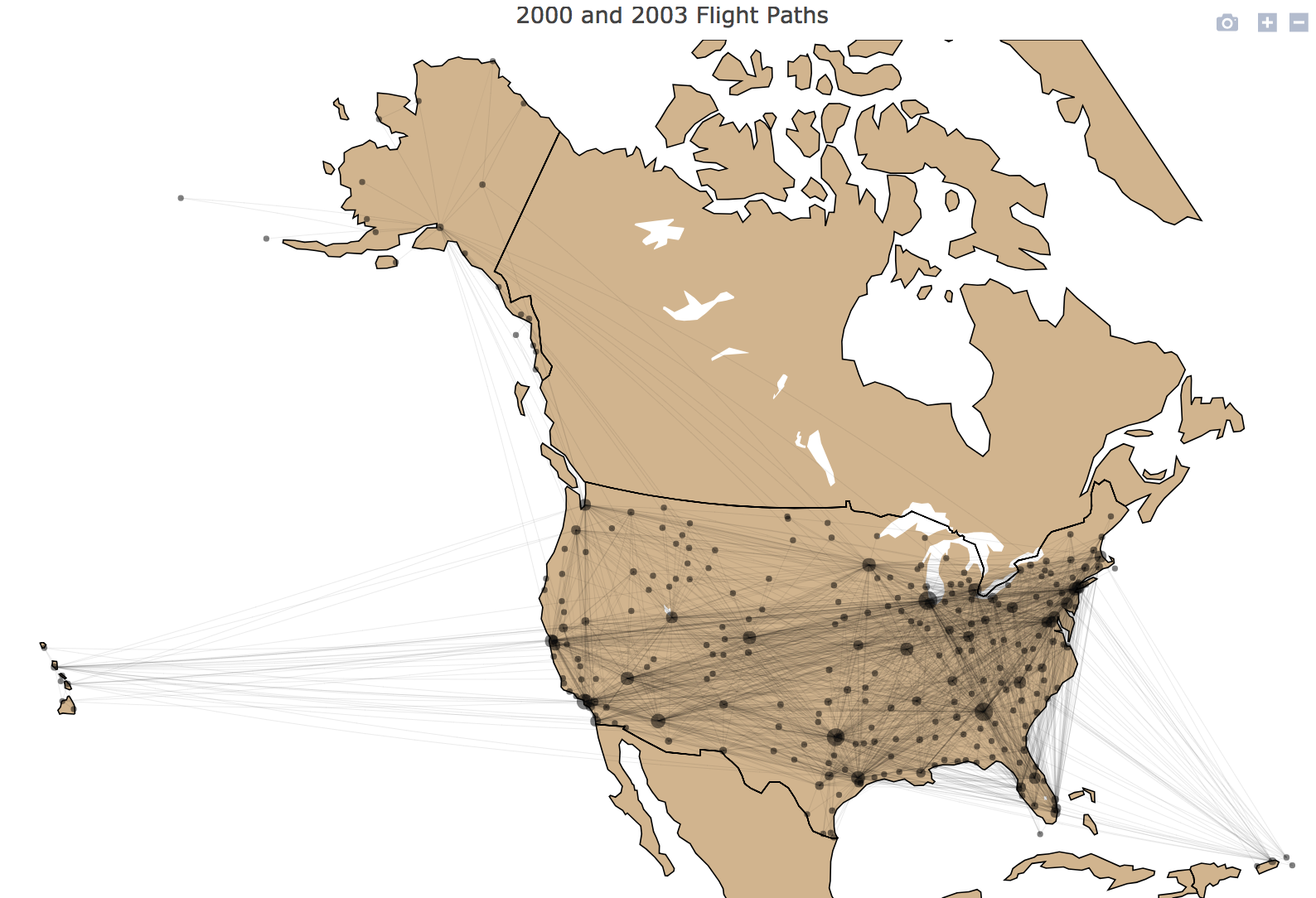


And the second appears as:



This was sorted by the number of flights to give an idea of the busiest airports. The major hubs appear to be Chicago’s O’Hare, Atlanta’s Hartsfield–Jackson, Dallas/Fort Worth International Airport, and Los Angeles International Airport.

Using these two tables, the data was imported into R and visualized with ggplot2. The following maps highlight the sheer volume in the number of flights across and between the years we are looking at. The sizes of the circles at the airport locations (not every single airport location was plotted to reduce map clutter) represent the number of outgoing flights from that airport, an indicator of the number of patrons using an airport and its general busyness. The map for both 2000 and 2003 is below:

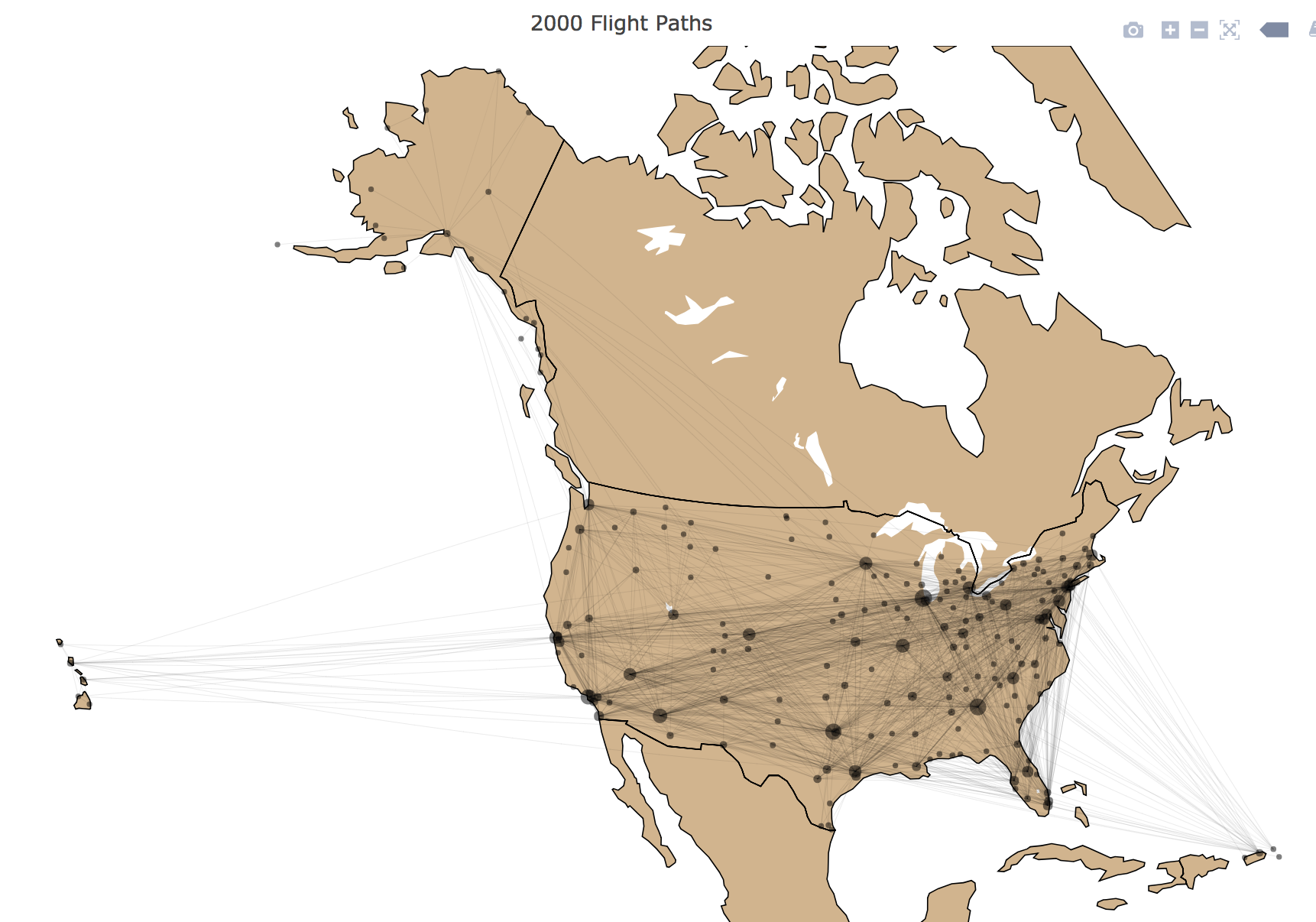


The alpha value of the flight path lines was altered to give a clearer view of the lanes of air travel receiving the most traffic. Looking at the map, it is evident that there are in general two major flyways: one east-west bound from the Northeast, which is densely populated by airports, to the Western of United States (Colorado, Arizona, Nevada, California, and by a lesser extent Hawaii), and another running north-south stretching from Chicago, Minnesota, and the cities of the east down to Florida, Atlanta, and the airports of Texas. This includes the area of the country that contains the highest density of flight paths: a triangle between Chicago, New York, and Atlanta. These were identified as hubs in the earlier Hive table, and others evident include DFW and IAH (Houston) in Texas, and several smaller circles in the west like LAX (LA), LAS (Las Vegas), and DEN (Denver), and multiple airports in the San Francisco bay area.

Also note the large number of flight paths running from the southwest to the northeast and vice versa, which appears to be as many if not more than those following a path from the northeast to the southeast. This could be explained by the large population centers in both of these corners of the country.

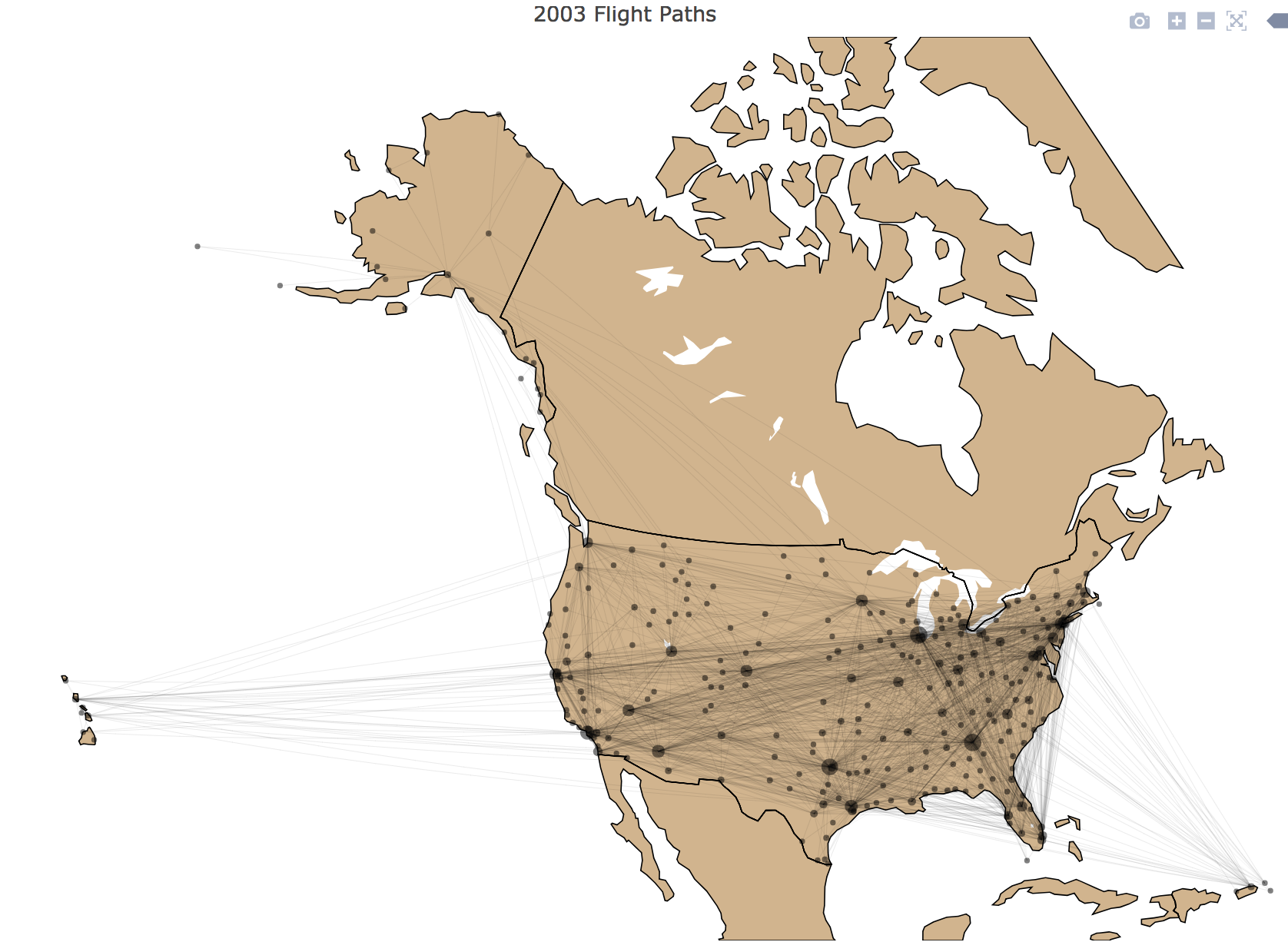
It is also interesting to note that the dataset included flights to US territories, including the US Virgin Islands and Puerto Rico. The longest flight paths are those between AK and HI and the contiguous US, with most of those to HI appearing to depart from the west coast in CA.

Specific flight path and airport data was extracted from these tables to plot flight paths of the individual years. The plot for flight paths in 2000 is below:



Compared to the overall flight paths map, the map for 2000 is obviously less dense with less flights and less airport destinations than the full map. This difference was queried and we found that there are 206 airports in 2000 and 282 in airports 2003 that appear either as a destination or origin. Long distance paths appear to be present in both. A noticeable missing airport in this map is EYW, or Key West International, that was present in the overall map. After investigating this, one possible explanation was that the airlines included in this dataset did not serve that airport during this year. This is somewhat doubtful, as we did not find available information on EYW that suggests flights were running abnormally during that year. This remains suspicious to us after some querying showed that there were 785 outgoing flights from EYW in 2003 but 0 in 2000 and research confirmed the airport was operational in that year.

The flight paths for 2003 is seen below:



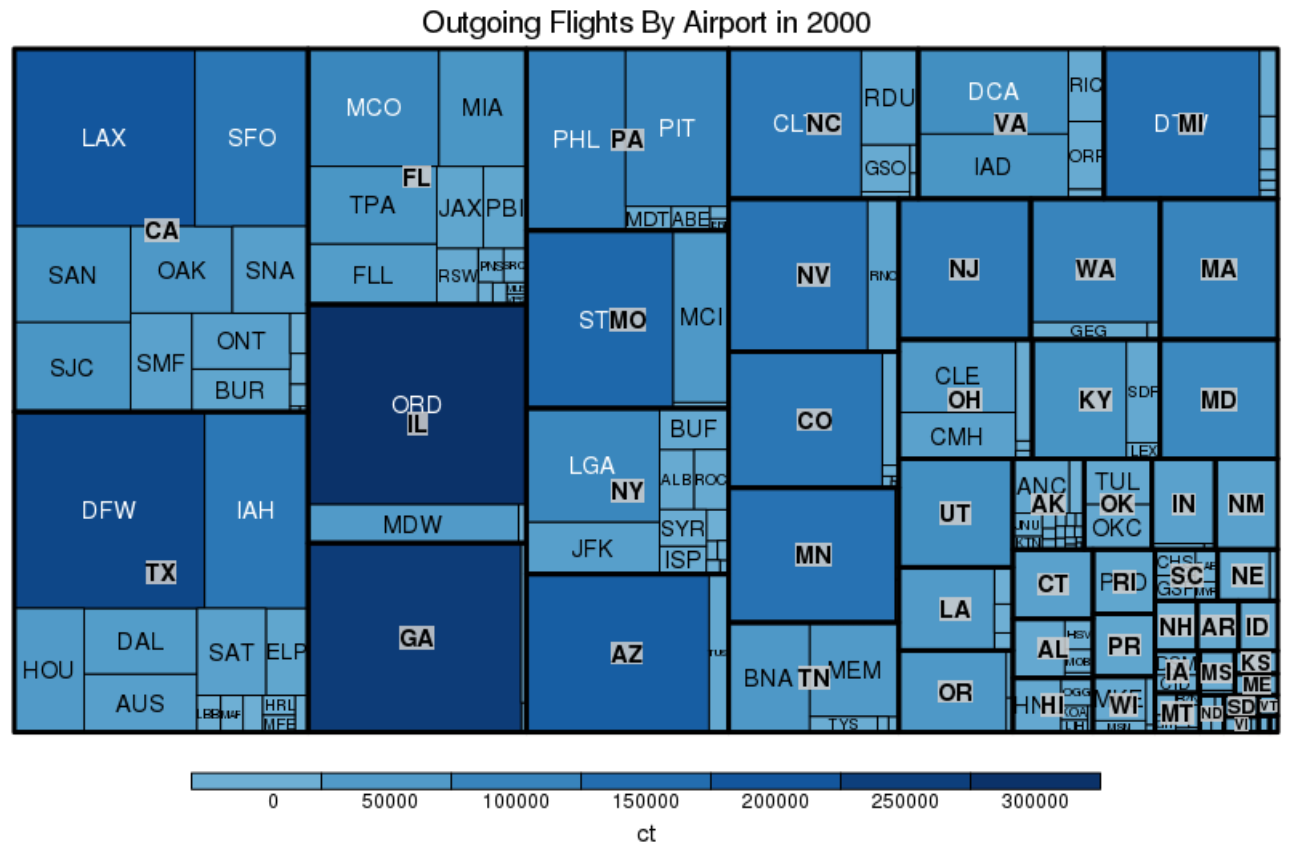
Comparing this map to the last one for 2000, this projection shows modest growth in the outgoing flights for most of the major hubs, like ORD, ATL, and DFW. Smaller airports by flight traffic also were identified as increasing, including Salt Lake City International Airport, Orlando’s major airport, and the string of airports in the cities of the Northeast (Washington, New York, Boston). In terms of flight paths, this year has all of the same discernable destinations as the full map. It does a good job of highlighting the straight, uninterrupted flights going between those major airports to the west and east of the Great Plains over the commonly termed “flyover” states.

It is important to remember external variables or data that may not be included in this data set during our analysis. The impact of international flight numbers on flight patterns and number of flights is not represented here, but can be assumed to account for a not insignificant portion of flights at the United States’ many international airports (ORD, SFO (San Francisco), etc.). Also airports nearest to the neighboring countries would intuitively harbor more international commuters. Thinking about these effects can give a more holistic view of the data that presented here.

### 4.4.2 Number of Flights By Airport and State

Next we continued to investigate the number of flights, now looking to organize the volume of flights by airport and state. To accomplish this, Hive tables for each year were constructed to include the number of outgoing flights for each airport, obtained from the last analysis, with an included *state* variable. To visualize these tables, tree maps were constructed that used number of outgoing flights as the block size and color. This is an easier way to compare the sizes of the circles in the last sections’ maps as well as gain any possible insights from grouping by *state*.

The tree map for the 2000 data is below:

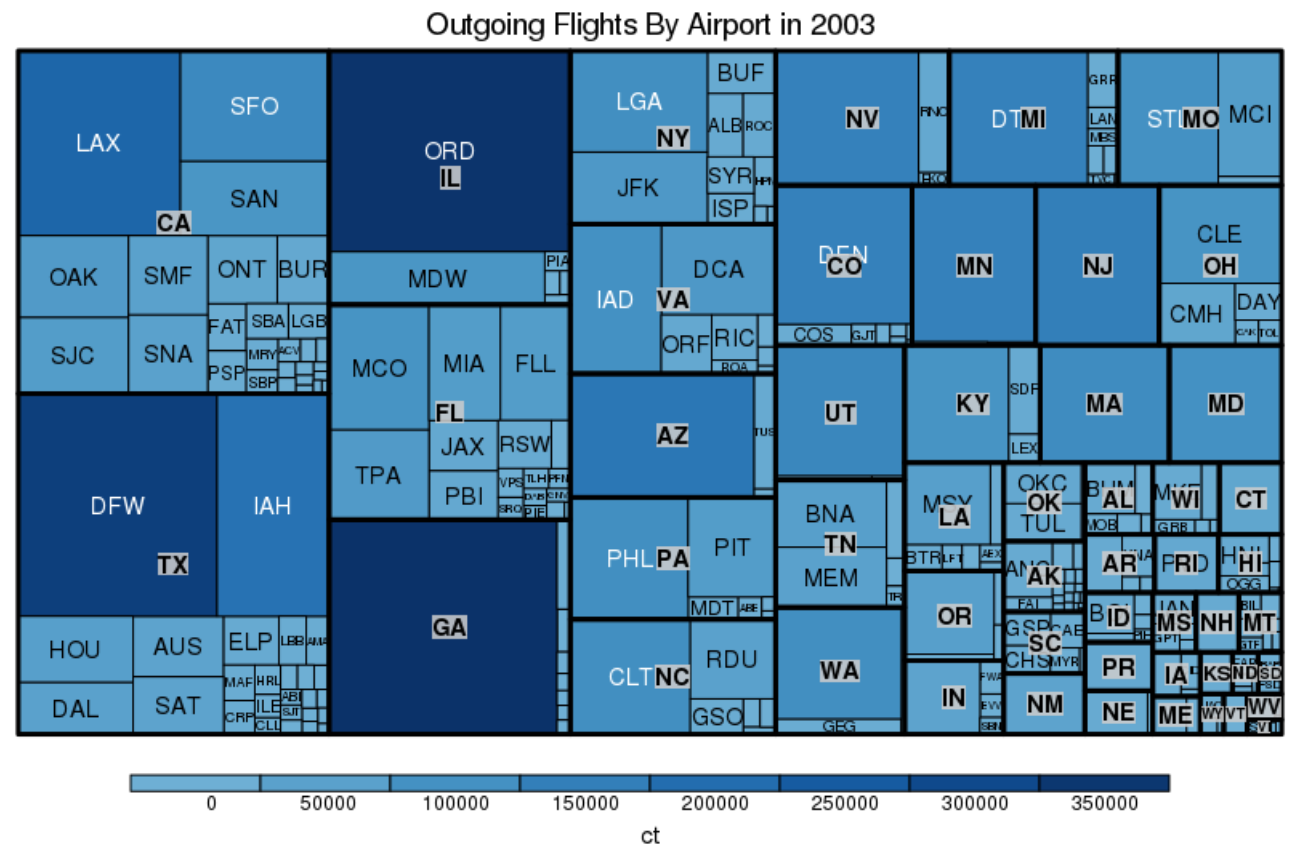


The hubs identified in the flight paths maps are noticeable here: ORD, ATL, DFW, and LAX dominate the count coloring which is not unexpected. CA and TX have the most number of outgoing flights, and combined they host more than all of the outgoing flights in FL, IL, and GA combined.

Looking at the blocks by state, we can identify several interesting characteristics. The top two airports for number of outgoing flights are not located in the states that lead in the number of outgoing flights. These are ORD in Chicago and ATL in Atlanta. This could be explained by the geography these airports share: they have central locations and are presumably airports commonly used during layovers in multi-leg flights.

Many states, no matter their total amount of outgoing flights, are singly dominated by a large airport usually in a metropolitan area. Examples include: CA (LAX), TX (DFW), GA (ATL), IL (ORD), AZ (PHX), and states with a medium amount of air traffic like NV (LAS), CO (DEN), and MN (MSP). It should be noted that many of these dominating airports are also international and not regional airports. Even the states with the least amount of outgoing flights share this characteristic, for example, CT, OR, NH, AR, etc. It seems as if airports within states rarely share the share of flight traffic evenly. The only state that comes close would be FL, which has Orlando’s airport with the most outgoing flights, followed by Miami’s and Tampa’s airports respectively. This could be explained by the state having several large population centers spread out across the state. Pennsylvania is also an exception, almost split evenly between Philadelphia and Pittsburgh, as well as Virginia, which has two major airports serving the Washington, D.C. Metropolitan area.

The tree map for the same data in 2003 is below:



Comparing this tree map to the last map, several differences are apparent. For overall state trends, the blocks for TX, CA, IL, FL, and GA are larger in general. This can be attributed to the increase in the number of flights and the rising populations of the areas in which these airports serve. An overall shrinking is observed for the state blocks for medium outgoing flight traffic, like MN, NJ, CO, UT, NV, MI, AZ, etc. I do not consider this as much as an effect of a lesser amount of outgoing flights in these states, but rather their proportion to the whole number of outgoing flights for all states has decreased, with the major hubs and states gaining more and more of a proportion.

One state in particular seemed to have shrunk in number of outgoing flights more than its neighbors had between years. Missouri’s STL airport was probably in the top ten highest number of outgoing flights in 2000 but in 2003 it appeared to have shrunk drastically, bringing the state down with it. After some research the following information was discovered:

“The September 11, 2001 terrorist attacks were a huge demand shock to air service nationwide, with total airline industry domestic revenue passenger miles dropping 20% in October 2001 and 17% in November 2001. Overnight, American no longer had the same need for a hub that bypassed its hubs at Chicago and Dallas, which suddenly became less congested. As a result of this and the ongoing economic recession, service at Lambert was subsequently reduced over the course of the next few years; to 207 flights by November 2003. Total passenger traffic dropped to 20.4 million that same year.”

Source: <https://en.wikipedia.org/wiki/St._Louis_Lambert_International_Airport>

The decreased utility in the location of the STL airport explains the higher than normal drop in the size of the MO block as well as the proportion that the STL airport has within it.

In the NY block, the JFK airport appears to have gained a proportion of the states’ traffic while LGA’s traffic decreased. This could be due to LGA’s problem with overcrowding delays, which was so bad in 2000 that Congress passed legislation to revoke the federal traffic limits on LaGuardia (<https://en.wikipedia.org/wiki/LaGuardia_Airport>). This could have resulted in more commuters using the nearby JFK airport.

Other trends included: CLE taking over a larger proportion of Ohio outgoing flights, and the same happened for PHL in Pennsylvania. The opposite occurred for Detroit’s airport in Michigan, and MCO in Florida. These patterns could have many explanations, with one being that the wake of the 9/11 terrorist attacks changed the dynamic of flight patterns in the country and thus changed the proportion of outgoing flights for many airports by state. Again the true causes are dubious.

## 4.5 Top 10 Airplane Models Used for Long and Short Duration Flights

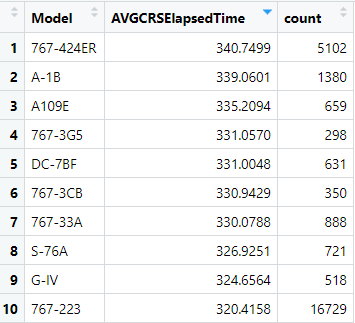
Towards the end, we have decided to do a fun analysis on which airplanes models were used for long duration flights, and which airplane models were used for short duration flights, and if the choice of airplane models have changed from year 2000 to 2003.

In this analysis, scheduled elapsed time (*CRSElapsedTime*) is used to measure the duration of flights, as airline carriers tend to choose various airplane models for different flights based on how long the flight is planned to be and it is more stabilized. Actual elapsed time (*ActualElapsedTime*) will vary from situation to situation, depending on the weather conditions, if the flight is diverted, or if the flight spent extra time circling the destination airport, waiting for its turn in the line of arriving planes.

Number of flights flown by each airplane model is also included in the analysis to see among airplane models used for similar level of flight duration, which is the most popular model.

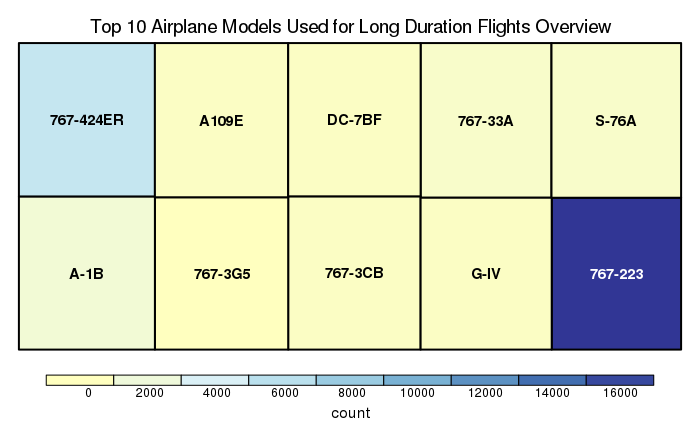
### 4.5.1 Long and Short Duration Airplane Models Overview

**Long Duration Airplane Models Overview**

****

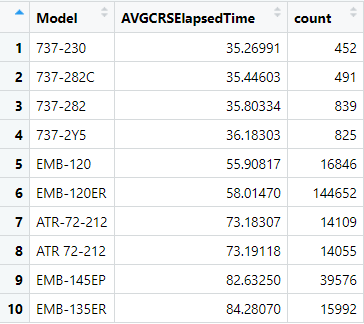
The above data shows the overview of 10 models with the longest average scheduled elapsed time with counts indicating the number of flights flown by the model in years 2000 and 2003.

As we can see from the data table, these models have average scheduled elapsed time from 320 minutes to 340 minutes, which is around 5 and a half hours, about the time needed to fly between west and east coasts.

The above treemap is plotted with the size of each block representing the duration of flights and colors representing the number of flights each model flown in years 2000 and 2003, ranging from about 300 to 16,700. 

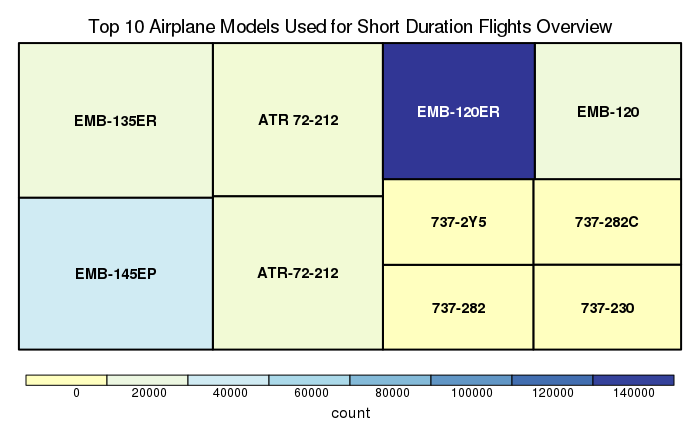
From the above treemap, it can be seen these models have similar average flight durations, but Boeing’s 767-223 is the most used airplane model for long duration flights. As a matter of fact, it accounts for over 60% of the flight in the above data table. The model following it is another version of Boeing 767, 767-424ER, with about ⅓ number of flights than 767-223. Other models are only used in flights less than 800 times in years 2000 and 2003.

**Short Duration Airplane Models Overview**



The above data shows the overview of 10 models with the shortest average scheduled elapsed time with counts indicating the number of flights flown by the model in years 2000 and 2003.

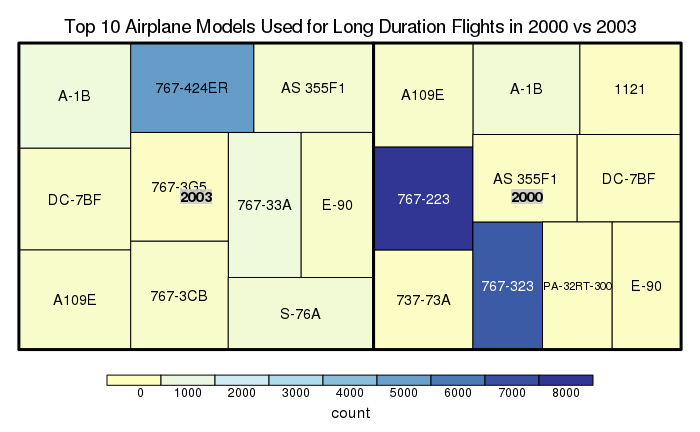
From the data table for short duration flights, the average scheduled elapsed time have a larger variation compared to long duration flights, from 35 minutes to 84 minutes, which indicates that these were probably mainly used to commuter flights in between large airports and the small airports surrounding them, like a flight from Champaign Airport to Chicago O’Hare International Airport.

The above treemap is plotted with the size of each block representing the duration of flights and colors representing the number of flights each model flown in years 2000 and 2003, ranging from about 450 to 145,000. 

The shape of the treemap shows confirms that there is more variation within short duration flights compared to long duration flights. Model EMB-120ER from Embraer is in middle of short flight duration and is also the most commonly used model for short duration flights. Following it is the model from the same company, EMB-145EP. Various version of Boeing 737 flew the shorts durations, and also have the least amount of flights in this map.

Then, we split the data to compare the conditions for long and short duration flights in 2000 and 2003.

### 4.5.2 Long and Short Duration Airplane Models in 2000 vs. 2003

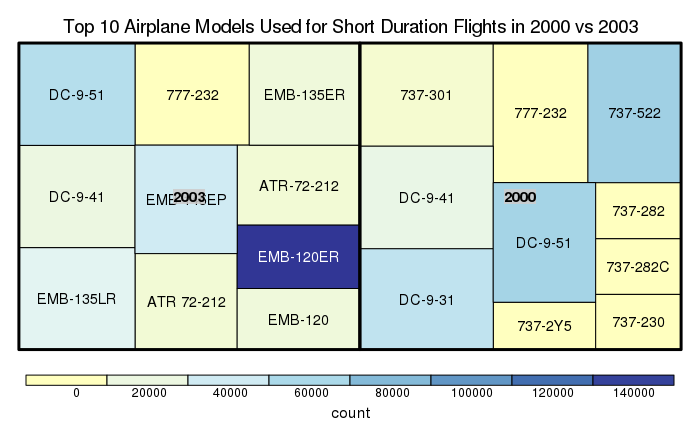
**Long Duration Airplane Models in 2000 vs. 2003**

The above treemap is plotted with the size of each block representing the duration of flights and colors representing the number of flights each model flown in years 2000 or 2003, ranging from about 200 to 7,500.

The trend of having a smaller variance in duration for long duration flight is continued when we split the data into year 2000 and 2003. The models with longest durations have switched to A-1B from Avia Aircraft Inc. in 2003 and A109E from Agusta Spa in 2000. However, the various versions of Boeing 767 still has the most number of flights.

Another observation from the treemap is that the colors for 2003 are generally lighter than colors in 2000, indicating that there are less long duration flights in 2003. This is quite counterintuitive, as you would think that 2003 should have more flights compared to 2000, as the aviation industry should grow over time. However, we shall not forget the influence of 9/11 on this industry, which led to less people flying, as shown in the treemap.

**Short Duration Airplane Models in 2000 vs. 2003**

The above treemap is plotted with the size of each block representing the duration of flights and colors representing the number of flights each model flown in years 2000 or 2003, ranging from about 450 to 145,000. 

In 2000, flight duration is more variable than in year 2003. This could indicate that over time, some extremely short flights are cancelled, probably due to the fact that people would rather choose to drive compared to taking a plane. This would be an example of how the industry modifies itself for only the most worthy flight routes.

Number of flights flown by each model is more evenly distributed in 2000 compared to 2003, where EMB-120ER from Embraer take over the majority of short duration flights.