Trends in Flight Delays and Cancellations in 2003

In this project, we are interested in looking at delays and cancellations by time and by airlines for 2003. They are interesting because we fly a lot and the number of (domestic) airlines we usually use are limited to one or two; yet, we have experienced delays and/or cancellations for all sorts of reasons, whether it is the weather or the airline, when traveling. Because we are looking at year 2003 only, the results we get may not be applicable to today; however, it would still be interesting to see trends over the course of the year and also which airlines performed well or poorly. We feel that we have to look at both delays and cancellations because they are not necessarily the same. They both could be affected by the same factor like weather; however, cancellations should happen in more extreme cases.

By looking at time, we should be able to see trends in delays and cancellations; in addition, based on the month or season, we should be able to tell if the weather or season played a role in delays/cancellation trend without even using weather data. We are also interested in popular airlines in 2003; and, we want to see the relationship between popularity and reliability by looking at the delay and cancellation rates. The questions we want to answer are:

- What is the delay rate by month?
- What is the cancellation rate by month?
    - What is the cancellation rate by the day of week?
    - What is the cancellation rate by the day of month?
- What is the delay rate by airline?
- What is the cancellation rate by airline?
- How might geographical location affect cancellation rates?

Whenever we cannot explain a phenomenon based on just looking at our plots, we will look for events (using Google and citing the sources) in 2003 that could explain it.

The datasets we use are airlines.csv, carriers.csv, and airports.csv. Because we could not merge the datasets in Hive by joining (possibly due to " " around the values in the auxiliary files), we use R to merge them. In airlines.csv, variables with NA values are TaxiIn, TaxiOut, CancellationCode, CarrierDelay, WeatherDelay, and NASDelay. These values are deemed completely unnecessary, so they are dropped before merging. Then, airlines.csv and airports.csv are merged by UniqueCarrier from airlines.csv and Code from airports.csv.
After finishing merging in R, the file is uploaded to HDFS to be imported in Hive. When reading it in Hive, we terminate fields by comma and skip the first line because the header is unnecessary.

Now, we take a brief look at our variables of interest, delay time and cancellations. For delays, we need to look at arrival delay and departure delay separately. The maximum arrival delay time is 1612 minutes (or about 27 hours late), and the minimum arrival delay time is -937 minutes (or about 16 hours early). The maximum departure delay is 1582 minutes (or about 26 hours late) and the minimum is -1410 minutes (or about 23.5 hours early). These four numbers are huge; not only delays that are as long as 1 day are unexpected, but also 1-day early arrival/departure are problematic for people. The number of cancelled flights is 101469, and the number of non-cancelled flights is 6387071. Hence, the cancellation rate is 101469/638707, which is about 1.6% cancellation rate. This overall cancellation rate much lower than we expected.

In Hive, we create a table/dataset for each question we want to answer. First, we look at the number of flights by month, by day of month, and by airline, individually.

Number of Flights By Month

Month	Number of Flights
1=January	552109
February	500206
March	559342
April	527303
May	533782
June	536496
July	558568
August	556984
September	527714
October	552370
November	528171
12=December	555495

In this table of number of flights by month, we can easily see that February has the lowest number of flights. This makes sense easily because February has fewer days than other months. March has the most number of flights, which we believe is due to traveling for spring break. January and December, and July and August are also high in the number of flights due to holiday/vacation seasons.

Now, we look at the number of flights by day of month.

Number of Flights By Day of Month

Day	Number of Flights	Day	Number of Flights
1	207029	17	217670
2	215020	18	213995
3	217391	19	213782
4	209777	20	214552
5	212384	21	217167
6	213380	22	211179
7	215585	23	215253
8	209774	24	215075
9	213934	25	209550
10	216583	26	212426
11	212104	27	209352
12	212994	28	213318
13	213578	29	195619
14	216619	30	197442
15	210695	31	120583
16	214730		

We can see that Day 3 and Day 17 have the highest number of flights, and Day 31 has the lowest number of flights because not all months have Day 31. The range of the number of flights is roughly between 120,000 and 218,000.

Moving on, we now look at the number of flights by airline:

Number of Flights By Airline

Airlines	Number of Flights	Airlines	Number of Flights
“WN”	958566	“CO”	302742
“AA”	752241	“DH”	291600
“DL”	660617	“EV”	273712
“UA”	543957	“HP”	189519
“NW”	499160	“AS”	161594
“MQ”	429098	“FL”	144700
“US”	411956	“TZ”	69176
“OO”	396801	“B6”	67184
“XE”	328086	“HA”	7831

We can see that WN or Southwest Airlines Co. has the highest number of flights, and HA or Hawaiian Airlines Inc. has the lowest number of flights due to lacking data for the first ten months.

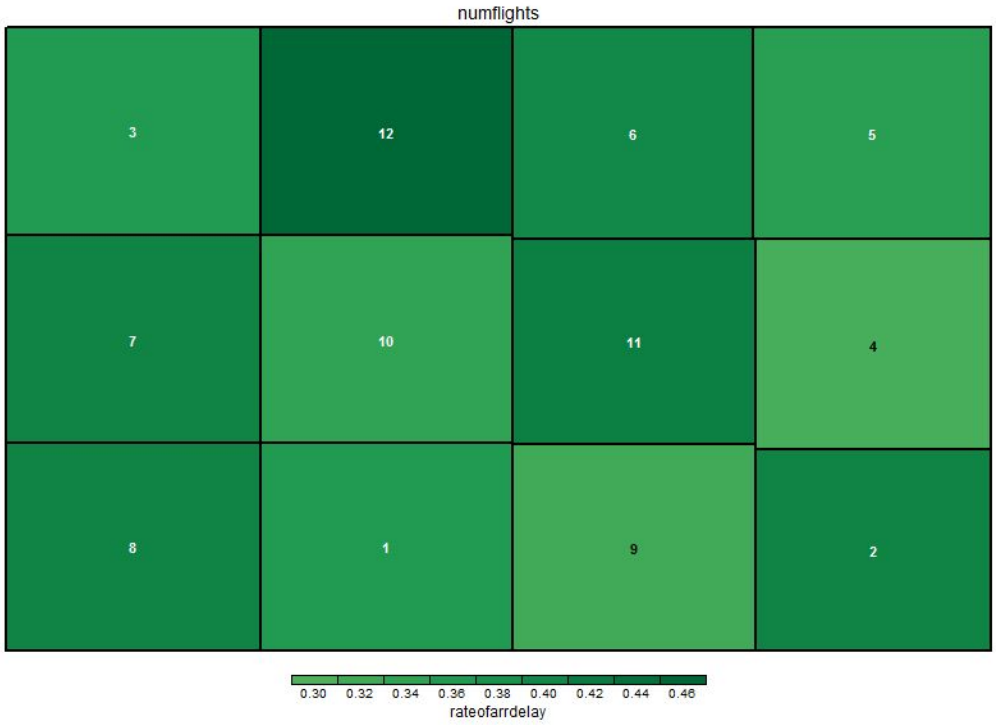
We will now move onto analysis using treemaps. In this analysis, we will look at treemaps of delay rates by month, cancellation rates by month, by day of week, and by day of month, and delay and cancellation rates by airline.

We first take a look at delay rates by month. Again, we are going to look at arrival delays and departure delays separately because there could be a difference, although they could be summed to total delays. Firstly, we look at arrival delay rates by month. The Hive table generated for this treemap consists of two fields, month and number of arrival delays. The number of delays is queried by counting the number of observations where ArrDelay and

DepDelay are greater than 0. The rate is then computed in R by dividing each month's number of delays by the corresponding month's total number of flights.

The treemap for the rates of arrival delays looks like this:

Arrival Delay Rates By Month

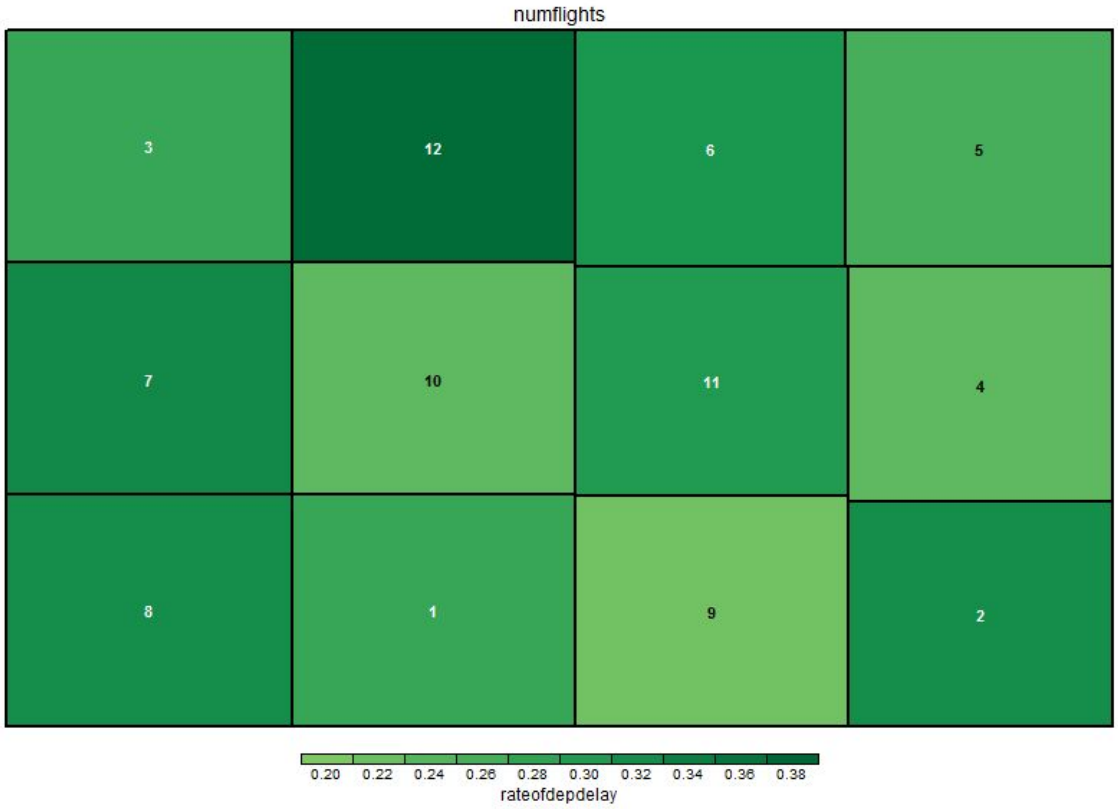


The size of the box is determined by the number of flights in each month, and we can see that number of flights is similar over 12 different months; and, the color is the rate of arrival delay. From this treemap, we can see that the top 4 months with most flights (March, July, August, and December) are high in arrival delay rate in addition to November, which includes Thanksgiving holidays, and February, possibly due to cold weather. The rate ranges from 30% to 46%. This minimum rate of arrival delay rate is surprisingly high.

The treemap for departure delay by month will have same components to the treemap for arrival delay, except that now the rate is departure delay rate. This rate is computed by dividing

the number of departure delays divided by the number of flights. The treemap for the rates of departure delays looks like this:

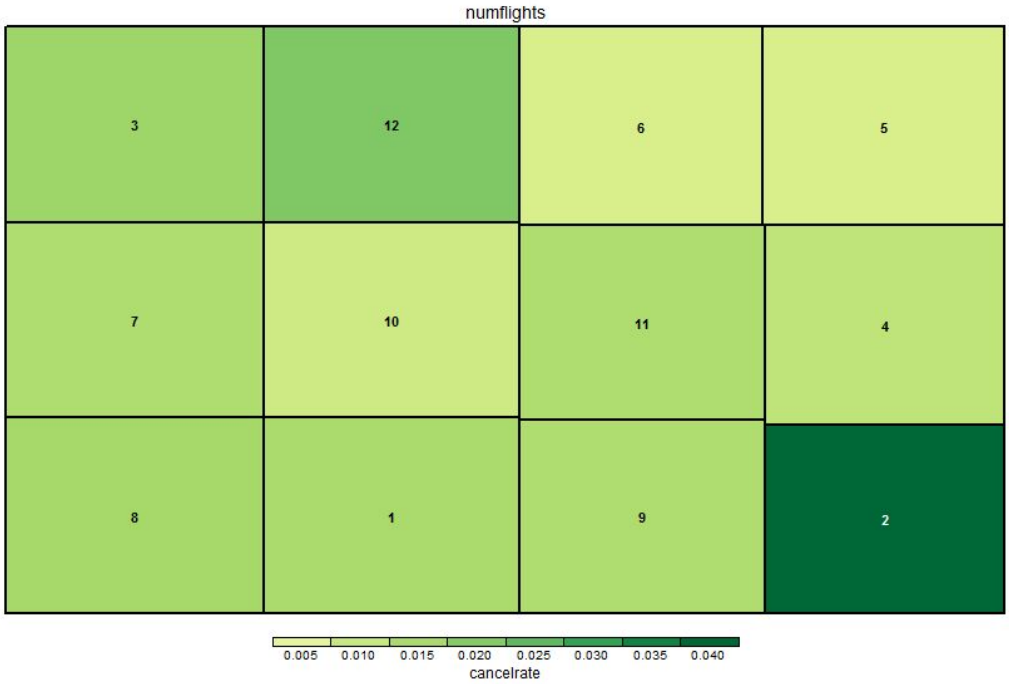
Departure Delay Rates By Month



We have calculated arrival delay and departure delay rates separately, knowing they are highly correlated, but not perfectly correlated. The treemap for departure delay rates looks very similar to the treemap for arrival delay rates. The range is 20%-38%, and this is shorter than that of the treemap above, meaning the two rates are not perfectly correlated. Departure delays are not as high as arrival delays. This makes sense because flights that both depart on time and depart late can end up arriving late. So, the arrival delay rate will likely to be higher.

Having looked at delay rates by month, we will now see if cancellation rates by month differs from delay rates.

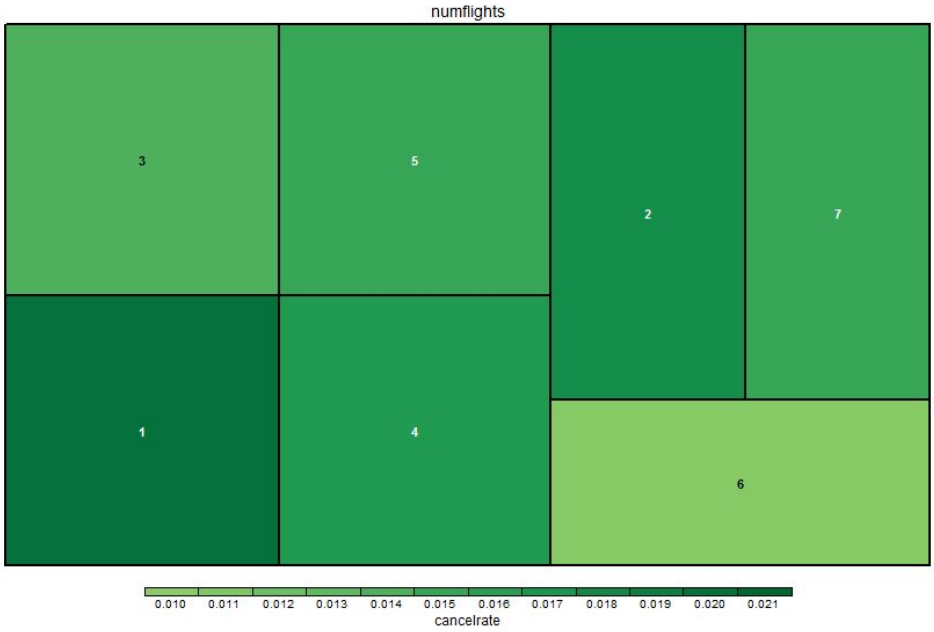
Cancellation Rates By Month



In this treemap here, the cancellation rate is computed by dividing the number of observations where "Cancelled" is 1 by the number of flights. The size of the box is again the number of flights in order to control for the number of flights. The month with the highest cancellation rate is February, even though it has the fewest number of flights. This is most likely due to the extreme winter weather. Also, the other months that are in the top 3 of the cancellation rate, December and November, are winter months. Although February is not the month with highest delay rate, it is the month with highest cancellation rate in 2003.

We now want to look at cancellation rate by day of the week. We expect more flights on the weekend simply because more people fly towards the weekend, and hence higher cancellation rates. For the treemap of cancellation by day of the week, each box will represent the day of the week and the color will represent the cancellation rate for the given day of the week. Here is the treemap of Monday (being 1) through Sunday (being 7):

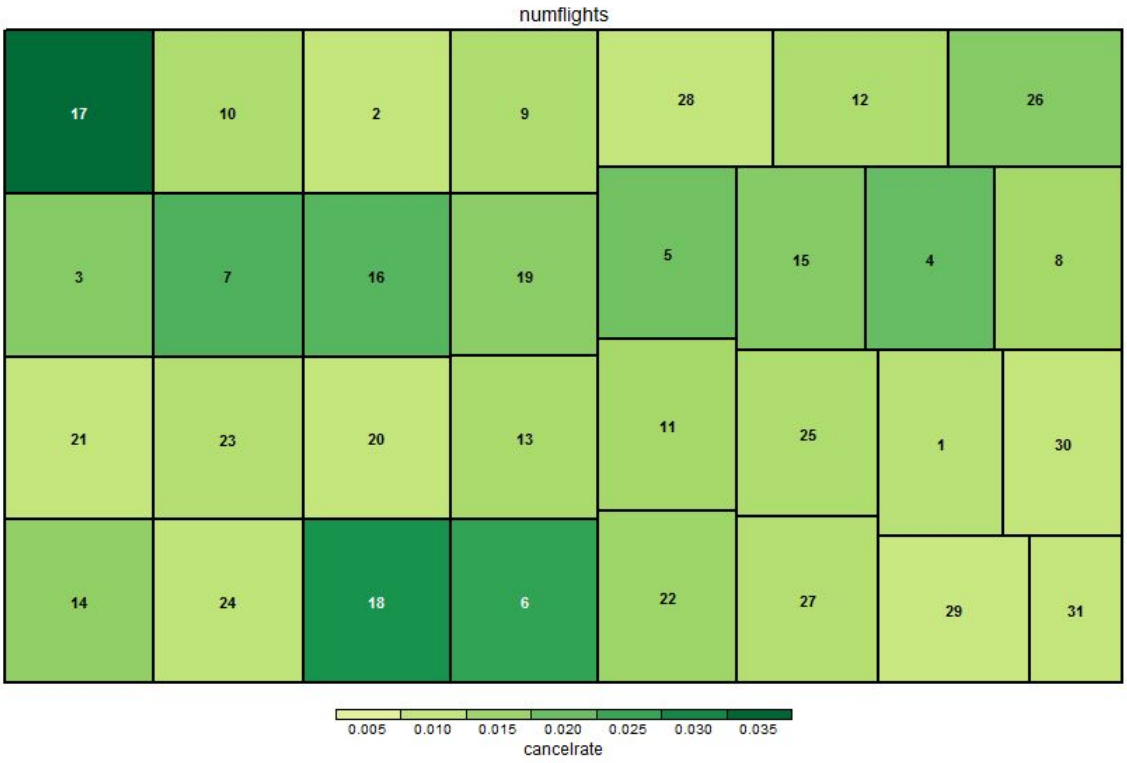
Cancellation Rates By Day of Week



We see that the number of flights is roughly equivalent per day, but Wednesday is the highest, with Monday in second and Friday third. As for the cancellation rate, Monday has the highest rate by far. Then, Tuesday comes in second. The smallest rate is, once again by a large amount, Saturday. The one possible reason for Monday being so high is that there is a large amount of national holidays that fall on Mondays (for example, MLK day, Labor Day, etc.) which creates 3-day weekends. This allows for a much larger amount of flights that occur on Mondays and thus creating more cancellations. In contrast to our expectation, the number of flights on Saturday and Sunday are relatively low. And, the cancellation rates are small, overall, around the global rate, 1.6%.

We now look at cancellation rate by day of the month. Again, the cancellation rate by day of the month is computed in a similar fashion to the rates we have already visited. We divide the number of cancellations where “Cancelled” is 0 by the number of flights by day of the month.

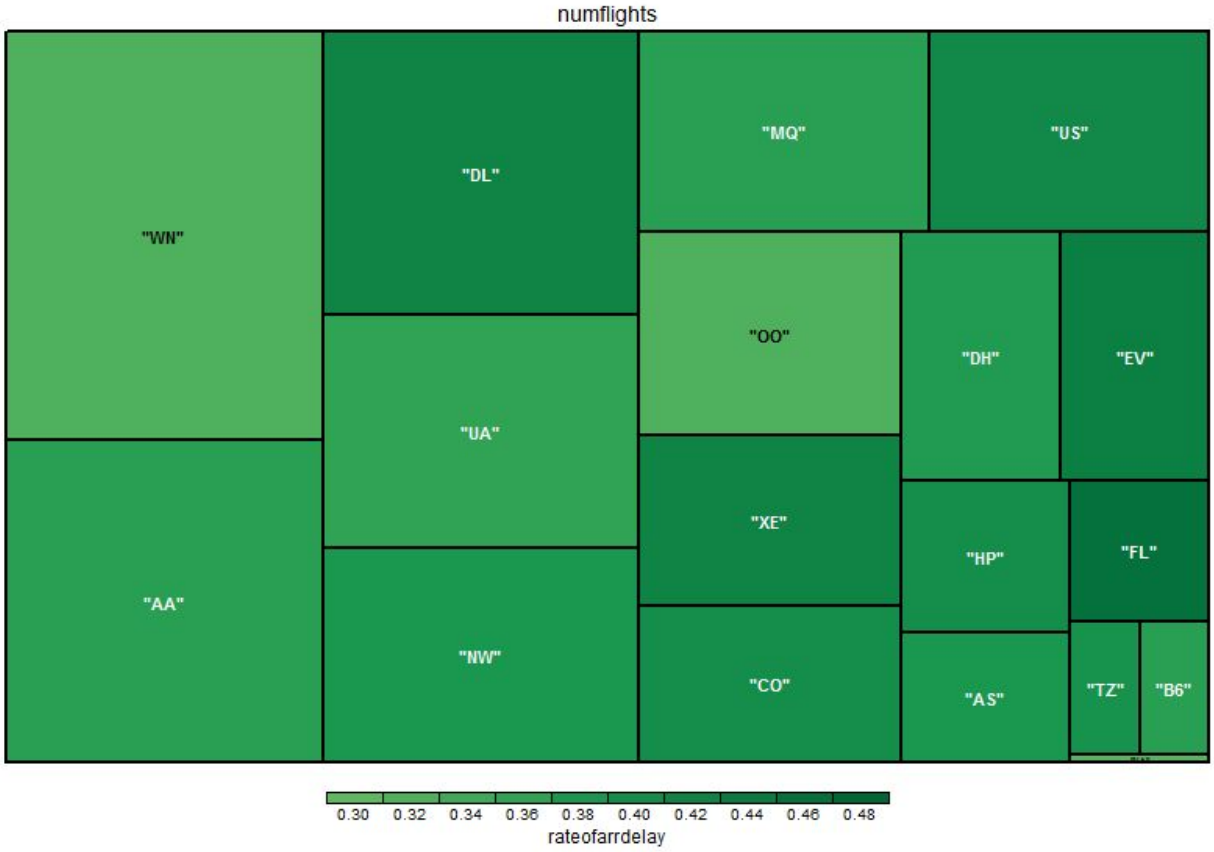
Cancellation Rates By Day of Month



Here is an effort to see if there is any significant differences in the cancellation rates by the day of the month. To start, the number of flights per day of the month is roughly equivalent amongst all days (with the only real noticeable exception being the 31st which is solely due to the fact that only a few months of the year have 31 days). Now, weirdly enough, there is a significant difference in cancellation rates for the 17th and 18th of the month. After doing some research (http://www.erh.noaa.gov/phi/archives.html) on extreme weather events in 2003, we discover that there was a giant snow storm that occurred on President's Day (the 17th of February). Certain places in New York and Pennsylvania recorded snow totals of more than 20 inches. This would clearly affect the cancellation totals of this year, if not all, of the air travel was suspended for that day and carried over a little bit into the 18th.

Having looked at time, we will now move onto airlines. The components to treemaps consist of three variables, again very similar to the rates we have seen: each box representing an airline; the size of the box again representing the number of flights by airline; and, the color of the box representing the delay or cancellation rate, computed by dividing the number of delays (where delay is greater than 0) or the number of cancellations (where "Cancelled" is 1) by the number of flights by airline. Again, we discuss the arrival delays and departure delays individually. Two fields are created in the Hive table for this treemap: the name of airline and the number of arrival/departure delays, respectively. The third component, the rate, is computed in R. The treemap for the rates of arrival delays by airline looks like this:

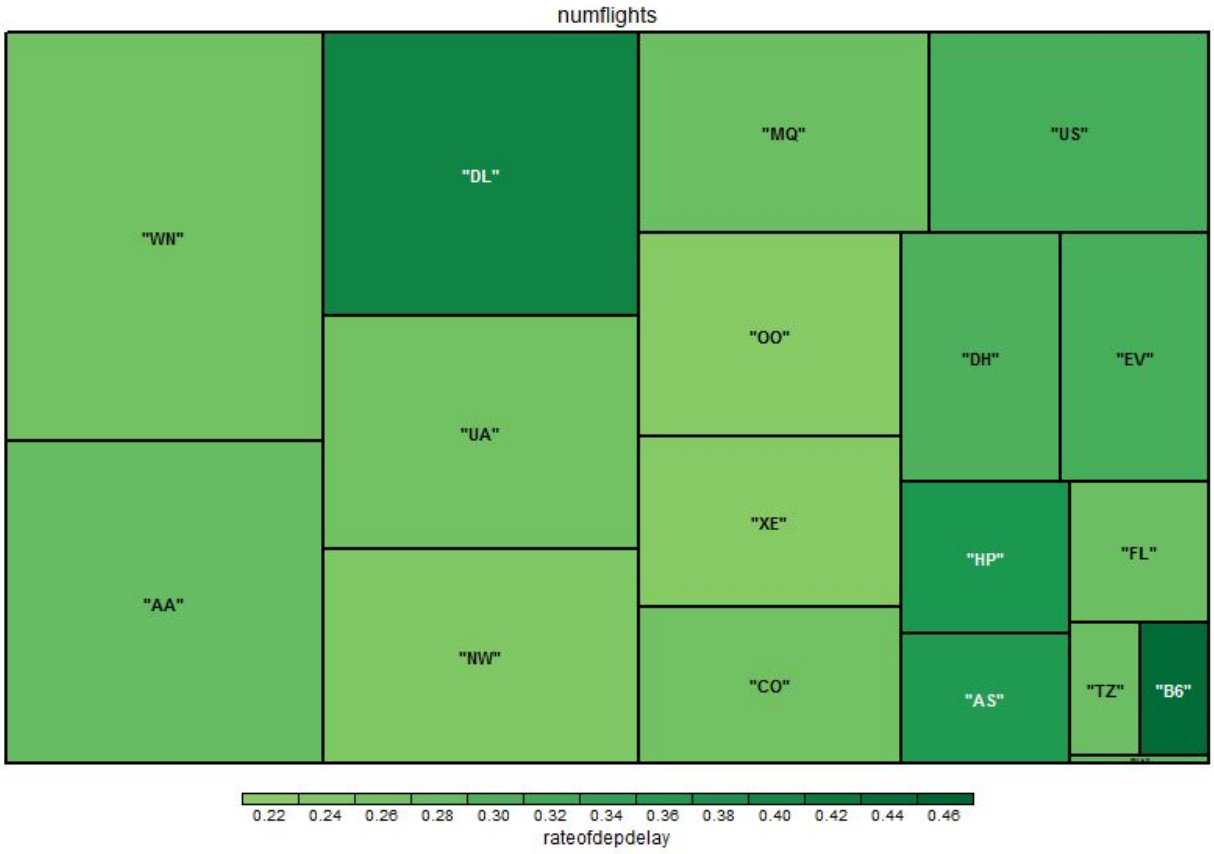
Arrival Delay Rates By Airline Carrier



Based on the treemap, we have that WN or Southwest Airlines Co. has the most number of flights since it has largest area; and, HA (Hawaiian Airlines, Inc.) has the fewest number of flights due to non-existing data from January to October. Meanwhile, FL or AirTran Airways Corporation has the highest rate of arrival delay despite not having a large number of flights, and WN (or Southwest Airlines Co.) has the lowest rate while being an airline with the most number of flights. There does not seem to be an obvious correlation between the number of flights of an airline and its arrival delay rate. The overall range is 30%-48%, which means the airlines listed above have a quite high arrival delay rate.

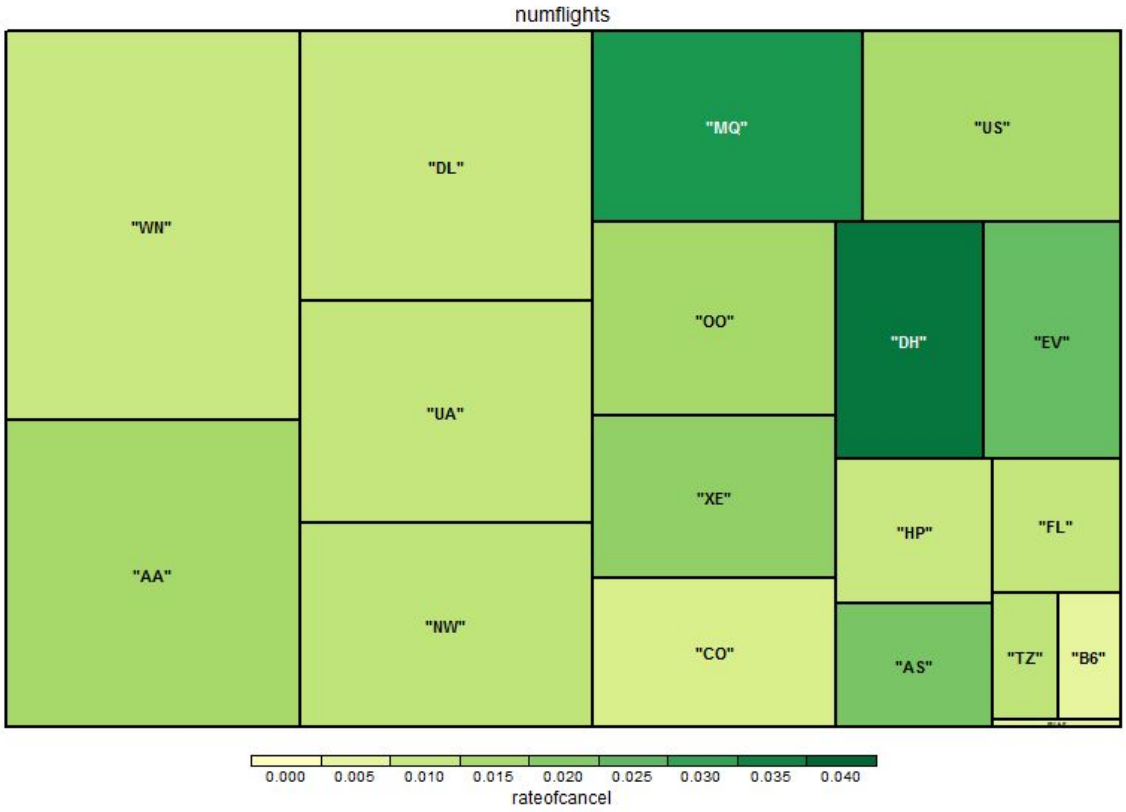
The treemap for the rates of departure delays looks like this:

Departure Delay Rates By Airline Carrier



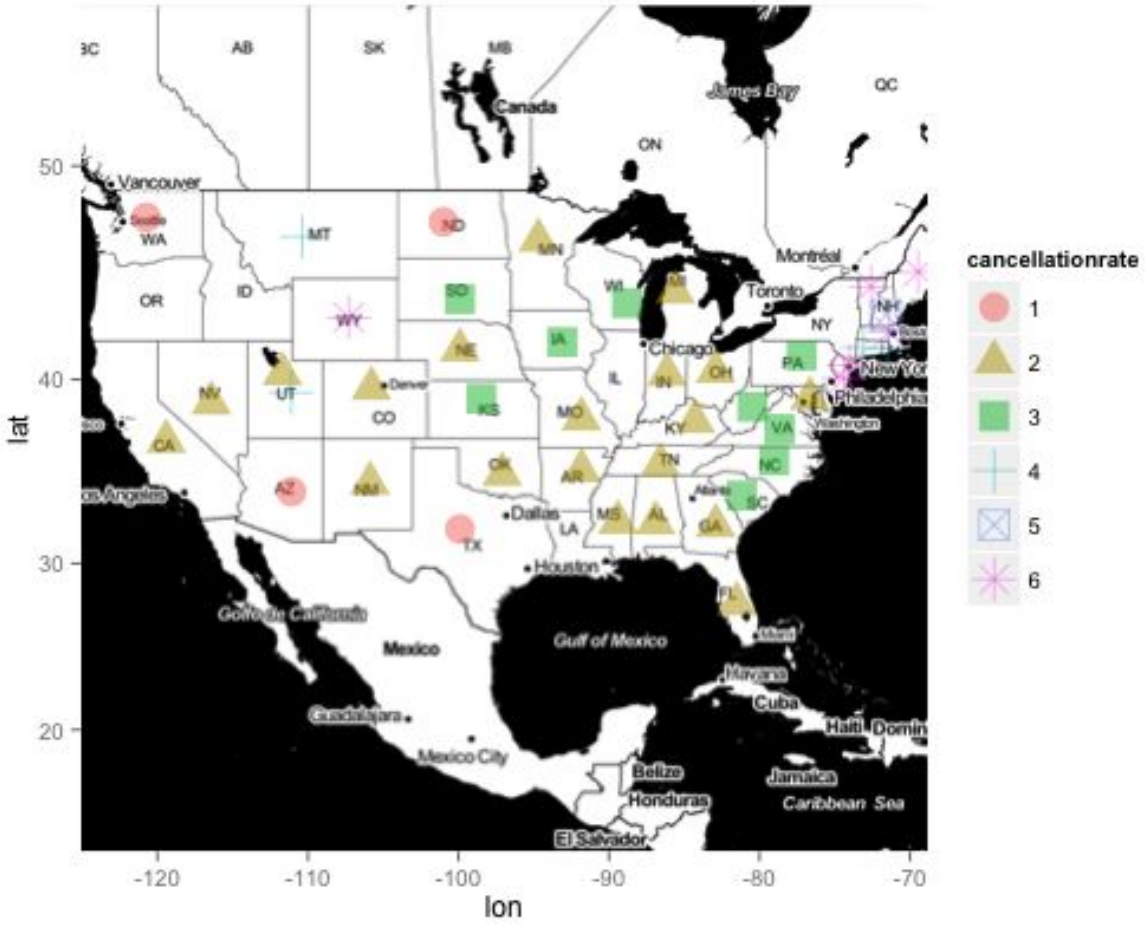
It can be easily seen that the departure delay rates are smaller than the arrival delay rates from the previous treemap since the colors are lighter. B6 or JetBlue Airways has the highest rate of departure delay, whereas it does not have the highest arrival delay in the previous treemap. In conclusion, B6 or JetBlue Airways is not an efficient airline because it has the least number of flights but the highest rate of departure delay; WN (Southwest Airlines Co.) is quite a good airline since it not only has the most number of flights but also has a low rate of departure delay.

Cancellation Rates By Airline



From this treemap, we can see that DH (Independence Air) and MQ (American Eagle Airlines Inc.) have exceptionally high cancellation rate, even though they do not have exceptionally high delay rates. And, there does not seem to be a relationship between the number of flights and cancellation rate.

Cancellation Rate By State in December



1: p<0.01
2: 0.01<p<0.02
3: 0.02<p<0.03
4: 0.03<p<0.04
5: 0.04<p<0.05
6: 0.05<p
where p is the cancellation rate

From the article we found at http://www.erh.noaa.gov/phi/archives.html, there were two severe snowstorm covering northeastern area where several states with exceptional high cancellation rate are located. This graph shows a good indication of where these snowstorms

affected the cancellation rates in some of the northeastern states and also in Wyoming. We also noticed that the cancellation rates were exceptionally higher for regions surrounded by mountains. So airlines like JetBlue has exceptionally high cancellation rate in December since it majorly covered the Northeastern area.

To conclude, the delay rates were correlated with the number of flights especially during holiday seasons, but the cancellation rates were more greatly affected by weather, such as February being affected by winter storms. Even for airlines, there is no significant correlation between delays and cancellations; thus, we can say that delays and cancellations are truly two different things. We can also see that geography added a layer of information to the weather conclusions we see within our dataset.

Simon Cho: Ran the merging R code and distributed the file to the group members using USB/Dropbox. Wrote Hive code for number of flights by day of month, which provided as an example that the group members can follow. Organized the Hive code and the R plot code. Organized the paper. Was the main leader of the group.

Yang Xu: Wrote and executed the Hive code to get arrival delays by month, departure delays by month, arrival delays by airline, departure delays by airline and described the data generated by these four csv files; Made the 4 treemaps in R; Searched for possible reasons related to odd data.

Dan Zielinski: Wrote the R code to merge the csv files. Wrote and executed the hive to code to get number of cancellations by month, cancellations by day of week and number of cancellations by day of month; made the corresponding treemaps in R.

Juntao Zhang: Wrote and executed the Hive code for: number of flights by month, by airlines, cancellation by states in Dec, total flights by states in Dec. Produced the ggmap with Stephen.

Steven Prince: Worked on Hive code for Cancellations by airlines and states. Made treemap of the rate of cancellation by airlines in R. Worked together with Juntao to create ggmap plot.