Big Data Analysis for Airlines Dataset

Chenzi Zhang, Shuoqi Zhang, Ziqin Xiong, Shuhui Guo (Group6)

Contents

Introduction **2**

Data Preparation **2**

1. Time **2**

1.1 Cancellation Rate 3
1.2 Delay 4
1.3 Number of Flights 9

2. Location **11**

2.1 Cancellation Rate 11
2.2 Departure Delay Rate 13

3. Carrier **17**

3.1 Cancellation Rate 17
3.2 Delay 19
3.3 Number of Flights 23
3.4 Information 24

4. Plane **26**

4.1 Cancellation rate vs Manufacturer of plane 26
4.2 Cancellation rate vs age of plane 27

5. Regression Model **34**

Conclusion **34**

Contributions **35**

Appendix **36**

Data preparation and Hadoop code 36
Pig code used in Carrier section 36
Hive code used in Plane section 36

References **37**

Introduction

Knowing about the flight conditions is a key part of planning trips. On occasion, delaying or canceling a flight is the only way we can maintain our high safety standards under bad situations. A flight delay is when a flight depart or arrive later than its scheduled time. We only focus on departure delay in this project. A flight cancellation is when a flight does not be planned to depart. Either condition is very challenging for passengers, carriers, and airports.

In this study, the cancellation and delay trends are first explored in four aspects, which are time, location, carrier, and plane type. Then some interesting information will be extracted. Finally, a model is fitted to predict if the flight will delay or not. The programming techniques used in this study include R, Hive, and Pig. Hopefully this study will give some suggestions to this field.

Data Preparation

Data analysis in this report is based on the Airlines Dataset from the US Department of Transportations Bureau of Transportation Statistics (BTS). The years studied in this project are the year of 2000 and the year of 2007. Dataset **airlines** contains more than 13 million samples and 30 attributes. Dataset **airports.csv** containes detailed information of airports in the US. Dataset **carriers.csv** containes detailed information of the carriers in the US. Dataset **plane-data.csv** contains detailed information of the plane type and manufacturers.

To prepare for the creation of a big.matrix object in R, the factors in the attributes in the **airlines** dataset are converted to their corresponding number of rows in other three datasets. Then a big.matrix object can be created using the processed data for further analysis.

In this project, the cancellation rate is calculated as below:

$$Cancellation \quad Rate = \frac{number \quad of \quad cancellations}{number \quad of \quad flights}$$

And the departure delay rate is calculated as below:

$$Departure \quad Delay \quad Rate = \frac{number \quad of \quad departure \quad delays}{number \quad of \quad flights}$$

where the missing values are excluded while calculating cancellation rate and delay rate.

1. Time

In this section, we analyze cancellation rate, departure delay (average and rate), and number of flights between year 2007 and 2000 by using R. To be more detailed, we compare data from these two years by month, and day of week in order to find special trend for the three variables. Overall, year 2007 data set has lower cancellation rate, departure delay and larger number of flights than year 2000. However, some results in departure delay are out of our expectation. We will analyze the reason why they appear in the following part.

1.1 Cancellation Rate

total cancel rate

```
## [1] 0.02650967
```

cancel rate by year

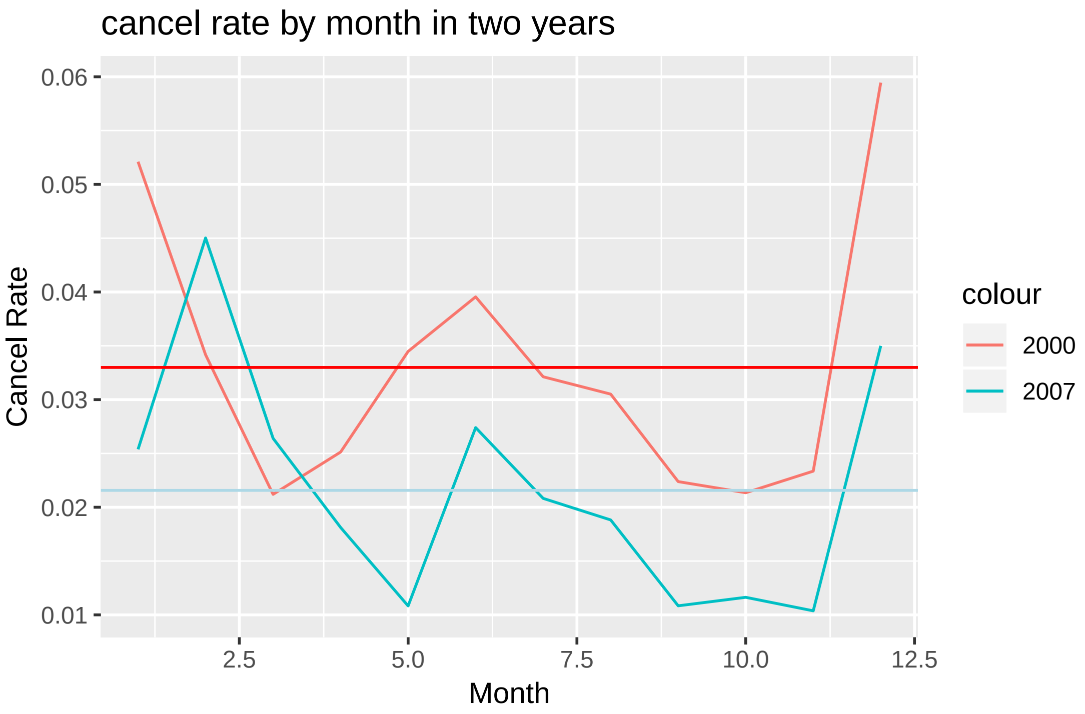
```
## [1] "2000" "2007"
```

```
## [1] 0.03299110 0.02156761
```

In general, cancellation rate in 2007 is lower than 2000, meaning there is a decreasing trend for cancel rate by year.

cancel rate by month in two years

```
##               1          2          3          4          5          6
## 2000 0.05210669 0.03416881 0.02119707 0.02513043 0.03448045 0.03954999
## 2007 0.02538295 0.04502267 0.02640294 0.01812094 0.01083107 0.02740116
##               7          8          9         10         11         12
## 2000 0.03212218 0.03050883 0.02238192 0.02134589 0.02335362 0.05945476
## 2007 0.02082460 0.01882044 0.01084162 0.01163031 0.01037596 0.03499696
```

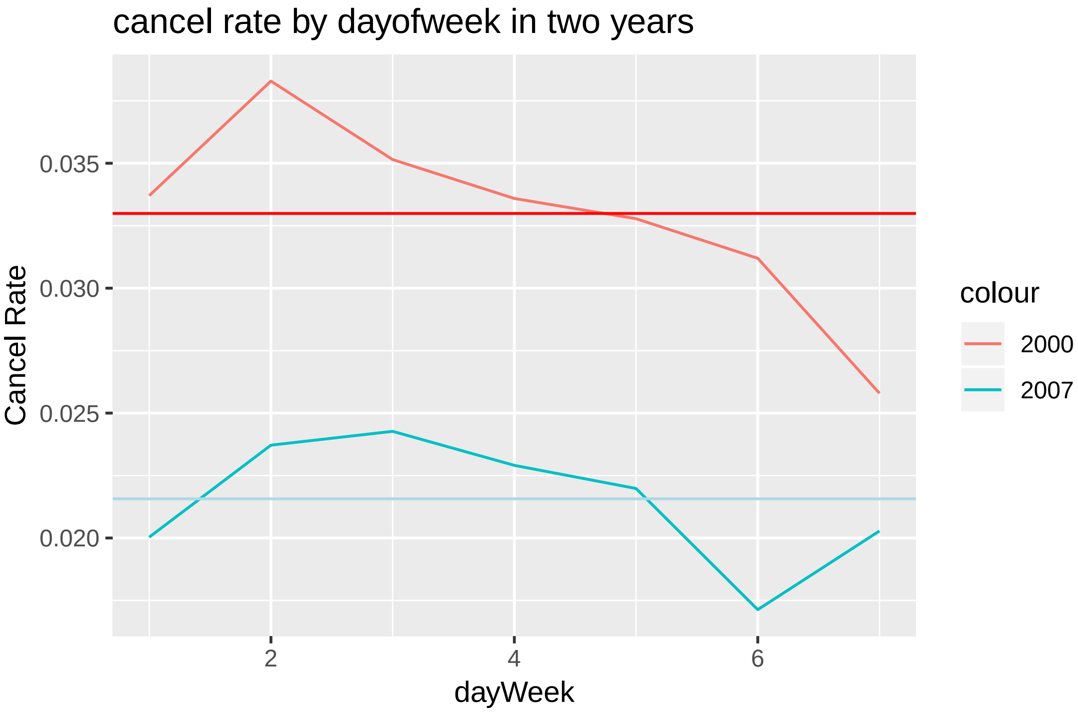


In the above plot, we can see that year 2000 and year 2007 almost have the same trend. Cancellation rate decreases in the first quarter of year, increases before June and decreases after June, increases again in November. Since weather in spring and fall is relatively better than winter and summer, high cancellation rates tend to appear in winter and summer.

However, the cancellation rate in 2007 has an increasing during first two month but cancellation rate in 2000 start decreasing from the beginning of the year. This is because the winter for 2007 ended later than that for 2000. Due to La Niña impact, some extrame low temparature appeared winter solstice, put off the decreasing for cancellation rate in 2007 .

cancel rate by dayofweek in two years

```
##               1          2          3          4          5          6
## 2000 0.03370392 0.03828982 0.03514772 0.03359026 0.03278026 0.03119615
## 2007 0.02003373 0.02371584 0.02426946 0.02290711 0.02198170 0.01713206
##               7
## 2000 0.02579348
## 2007 0.02028050
```



In the above plot, we can see that year 2000 and year 2007 almost have the same trend. Cancellation rate decreases from Tuesday or Wednesday, increases from Saturday or Sunday. Refering to the corresponding figure for number of flights, we find that this is because there are more flights in the weekdays and less flights in the weekends. During weekdays, the airplanes are busy with larger numeber of flights so that the airplanes have heavier stress for scheduling flights, leading to higher cancellation rate.

However, the trends for year 2000 and year 2007 are slightly different. Cancellation rate in year 2000 has a peak on Tuesday and decreases through Tuesday to Sunday, then increase from Sunday to Tuesday. Cancellation rate in year 2007 has it hightest point on Wednesday and decreases through Wednesday to Saturday, then increases from Saturday to Wednesday. In 2007, the airplane obviously has better management for airlines, resulting in extreme low cancellation rate on Saturday.

1.2 Delay

1.2.1 Yearly

Depature Delay Average

```
##                    [,1]               [,2]
## year               "2000"             "2007"
## yearly_dep_average "11.2806818307953" "11.3991417444878"
```

The yearly depature delay average for 2000 is 11.281 and for 2007 is 11.399 and there is no big difference between these two average values.

Depaature Delay Rate

```
##                  [,1]                [,2]
## year             "2000"              "2007"
## yearly_dep_rate "0.442081303132694" "0.431241444081955"
```

The yearly depature delay rate for 2000 is 0.442 and for 2007 is 0.431 and the rates for 2000 and 2007 are close.

1.2.2 Monthly

Depature Delay Average

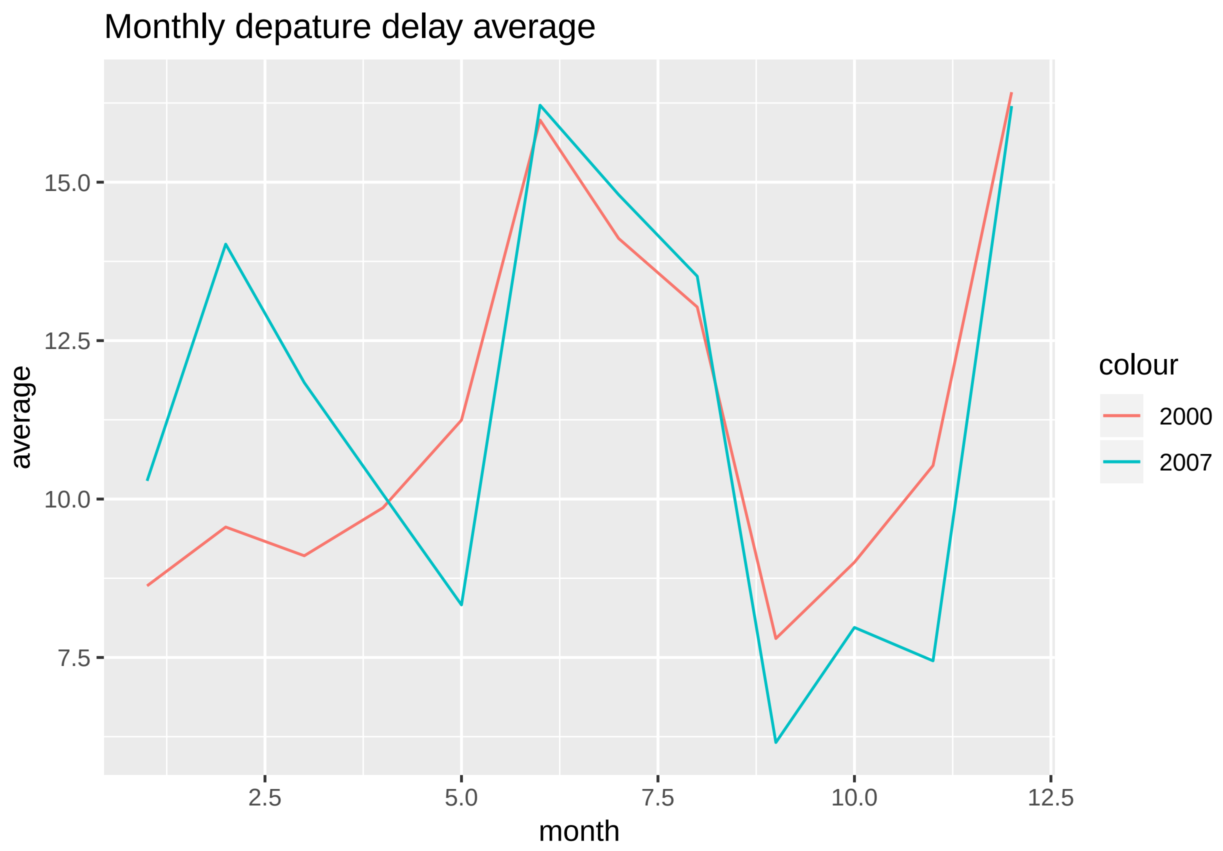
Firstly, we generate the monthly average depature delay value for different years in 2000 and 2007. Here is the result table.

Table 1: Depature Delay average (min) (Jan - June)

	1	2	3	4	5	6
2000	8.630	9.558	9.105	9.861	11.245	15.980
2007	10.286	14.022	11.837	10.078	8.329	16.215

Table 2: Depature Delay average (min) (July - Dec)

	7	8	9	10	11	12
2000	14.111	13.030	7.799	9.003	10.528	16.421
2007	14.803	13.516	6.158	7.974	7.447	16.201



Then we used R to generate the trend line graph between the two years to see a more clear trend. As we can see from the graph above, the overall trends of 2000 and 2007 in monthly average depature delay are similar. In the months May, and September to December, year 2000 have a relative higher value in depature delay than year 2007. In the months from January to March, year 2007 have a relative higher value of average depature delay value than 2000. In the article we found (Rebecca Lindsey 2008) (*La Niña* 2019), in year 2007 in north America the winter ended late which ends in the early spring. This can be a reason that from January to March, the average depature delay values are higher than the values in 2000.

We have two peaks in the trend graph, which is June and December. These two months all have relative higher value of depature delay average than other monthes. In the month of December, the weather condition is usually hard and the traffic of the planes is high since it has the Chrismas holiday. Then it can cause the higher depature delay average. In the month of June, lot of students in the school have the summer vacation then they need to travel. This can cause the higher value of average depature delay because of the heavy traffic.

Depature Delay Rate

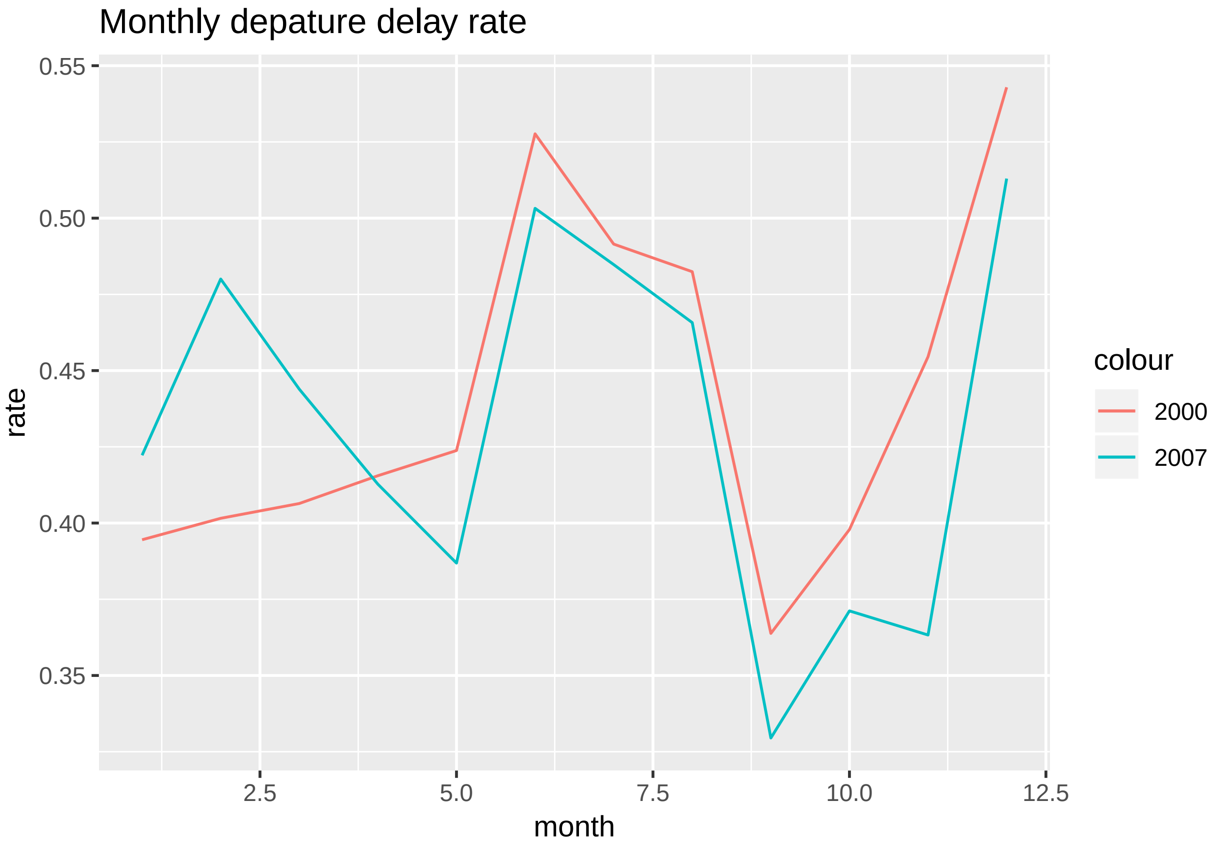
Here is the monthly data table for the depature delay rate in 2000 and 2007.

Table 3: Depature Delay Rate (Jan - June)

	1	2	3	4	5	6
2000	0.395	0.402	0.406	0.416	0.424	0.528
2007	0.422	0.480	0.444	0.413	0.387	0.503

Table 4: Depature Delay Rate (July - Dec)

	7	8	9	10	11	12
2000	0.491	0.482	0.364	0.398	0.455	0.543
2007	0.485	0.466	0.329	0.371	0.363	0.513



As we can see from the graph above, similar to the trend in the average depature delay, the trend for the depature delay rate in 2000 and 2007 is similar for most of the months. The only difference is from Janurary to May. Most of the values of depature delay rate in 2000 are higher than in 2007 except the months from Janurary to March. As mentioned before, north America suffered the long time winter in 2007 which influences the higher value of departure delay rate in 2007 for the months from Janurary to March. The trend of the delay rate in 2007 kept increasing sharply from Janurary to Feburary, but in 2000 starting from Janurary, the depature delay rate graduately increasing. The sharper increase maybe caused by the late end of winter in 2007.

Similar to the monthly depature delay average, there are two peaks in this graph, June and December. The reason for this is still caused by the fact that the plane traffic for those two months are usually busier than before.

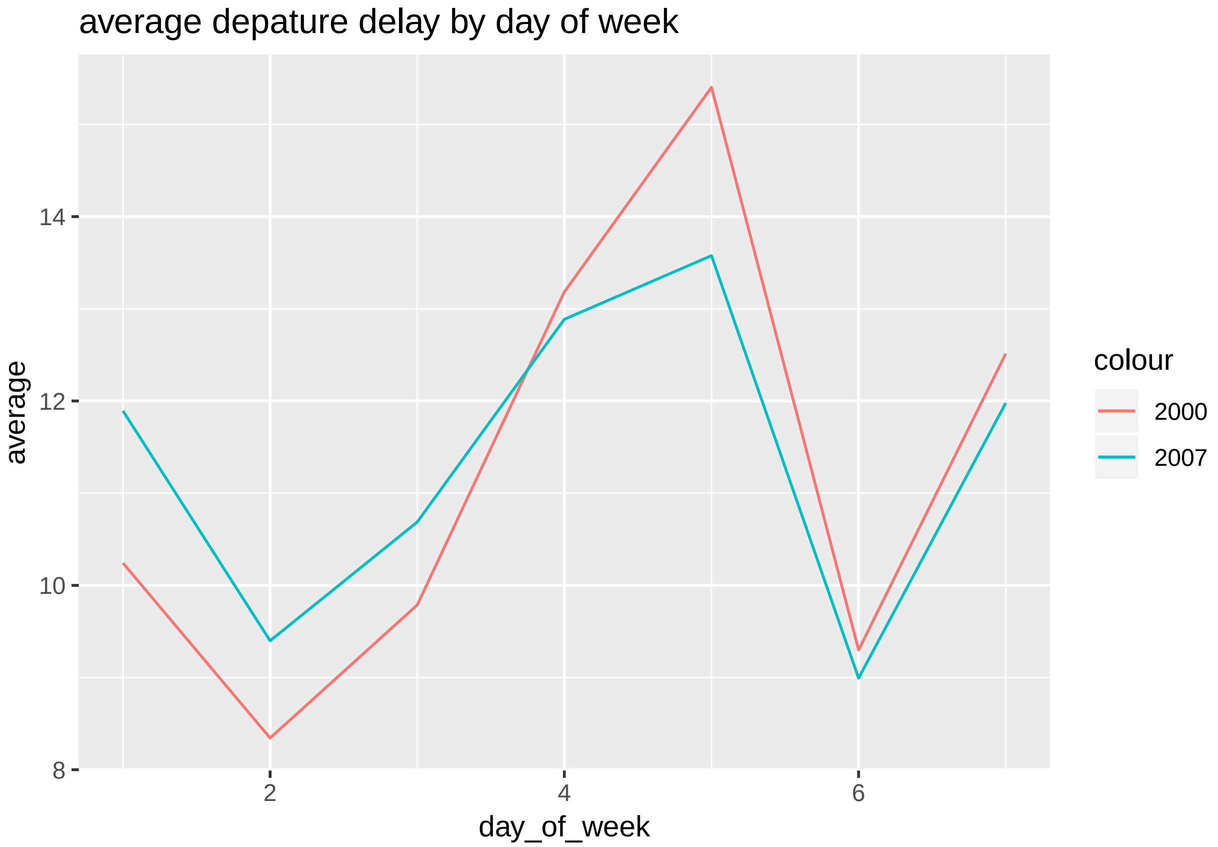
1.2.3 Weekly

Depature Delay Average

Here is the table for the depature delay average by day of the week.

Table 5: Depature Delay average (min)

	1	2	3	4	5	6	7
2000	10.242	8.344	9.787	13.184	15.403	9.299	12.514
2007	11.892	9.399	10.687	12.886	13.577	8.994	11.978



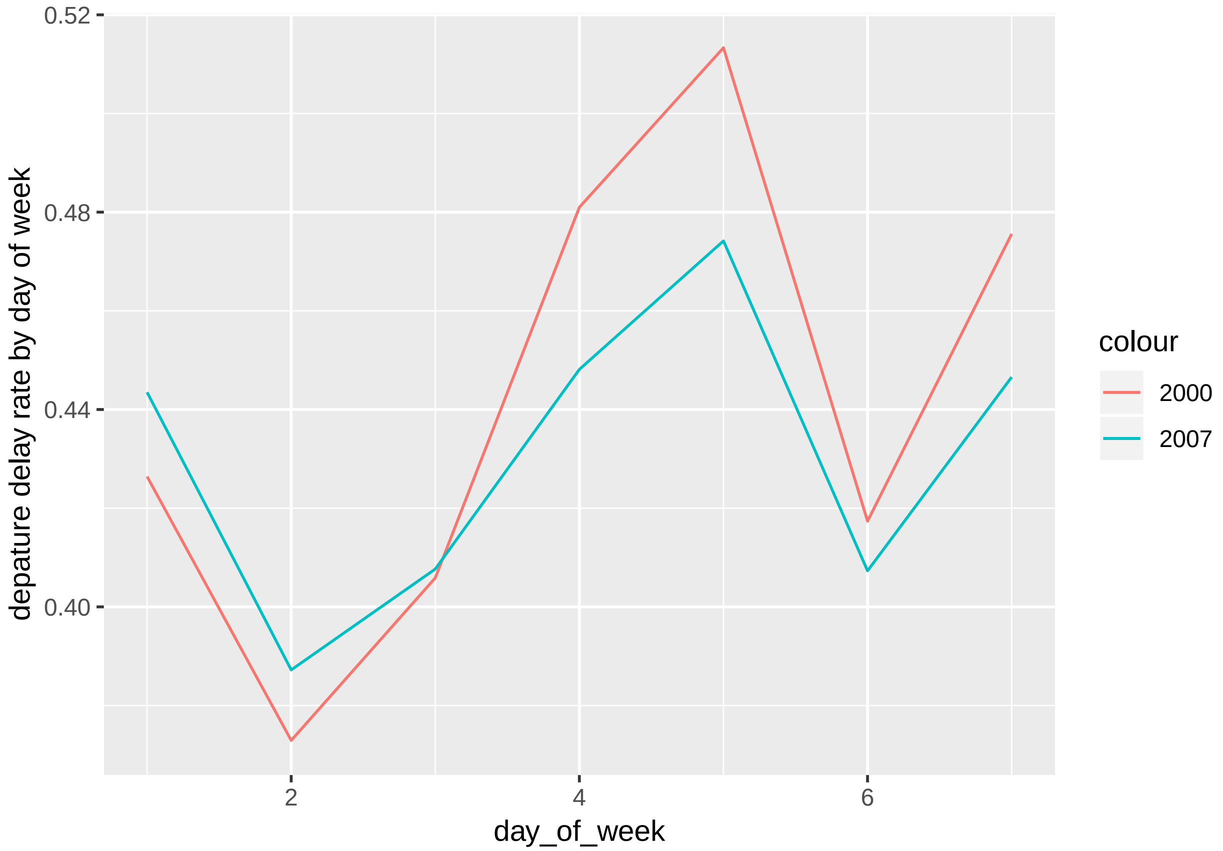
Above is the graph for the average depature delay for day of week. In the graph we can see that the trend of average depature delay in 2000 and 2007 for all the day of the week is the same. From Tuesday to Friday, the value of average depature delay is increasing. The peak of the graph is Friday for both 2000 and 2007. The peak maybe caused by the fact that many people leave at Friday to have a short vocation so the traffic of the plane is heavy then causes the higher depature delay average.The value decreases on Saturday but increases on Sunday since most people come back from other places on Sunday then the traffic of the plane increases again. The average depature delay in 2000 from Thursday to Sunday is higher than year 2007 and the average depature delay in 2000 from Monday to Wendesday is lower than year 2007.

Depature Delay Rate

Here is the table of the depature delay rate for each day of the week in 2000 and 2007

Table 6: Depature Delay rate

	1	2	3	4	5	6	7
2000	0.426	0.373	0.406	0.481	0.513	0.417	0.476
2007	0.443	0.387	0.408	0.448	0.474	0.407	0.447



Similar to the result of the depature delay average for each day of week, the trend for depature delay rate in 2000 and 2007 is quite similar. Increase from Tuesday to Friday, then decrease on Saturday then increase on Sunday, and last decreases from Monday to Tuesday. The high value of depature delay rate maybe due to the high traffic of the planes on Friday since peopel prefer to spend the weekend from Friday. Year 2000 has the higher depature delay rate from Thursday to Sunday than year 2007 and lower value on Monday, Tuesday, and Wednesday.

1.3 Number of Flights

total flight number

```
## [1] 13136262
```

flight numbers by year

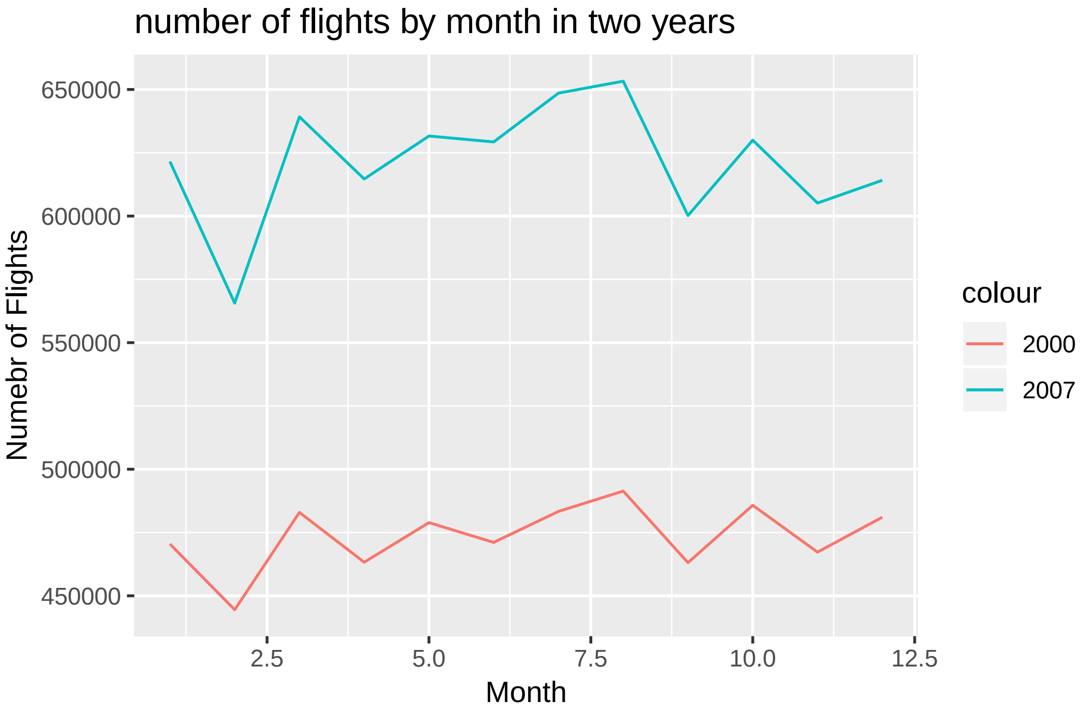
```
## [1] "2000" "2007"
```

```
## [1] 5683047 7453215
```

In general, number of flights in 2007 is larger than 2000, meaning there is a increasing trend for cancel rate by year.

flight numbers by month in two years

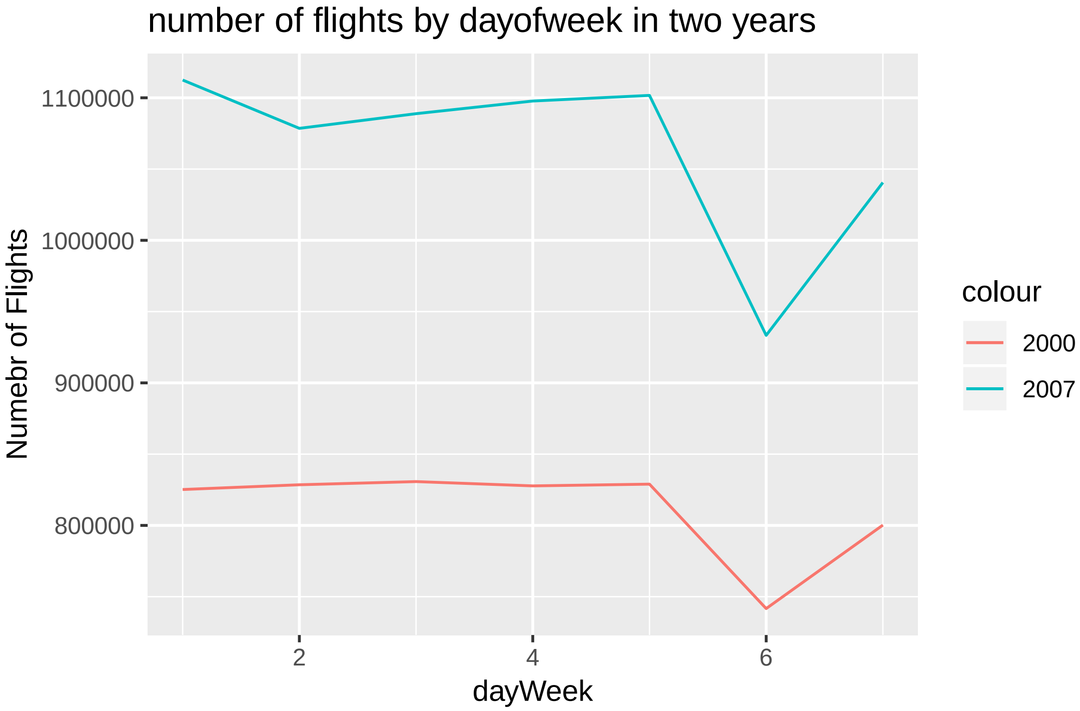
```
##           1      2      3      4      5      6      7      8      9     10
## 2000 470477 444499 482944 463263 478909 471100 483342 491366 463097 485761
## 2007 621559 565604 639209 614648 631609 629280 648560 653279 600187 629992
##          11     12
## 2000 467251 481038
## 2007 605149 614139
```



As it shown in the above plot, number of lights in 2007 and 2000 almost have the same trend by month. Extramely small number appears in February and the peak for number of flights appears in August. It can be concluded that the trend for number of flights in these two years has not changed.

flight numbers by dayofweek in two years

```
##            1       2       3       4       5      6       7
## 2000  825186  828523  830751  827740  828944 741662  800241
## 2007 1112474 1078562 1088858 1097738 1101689 933338 1040556
```



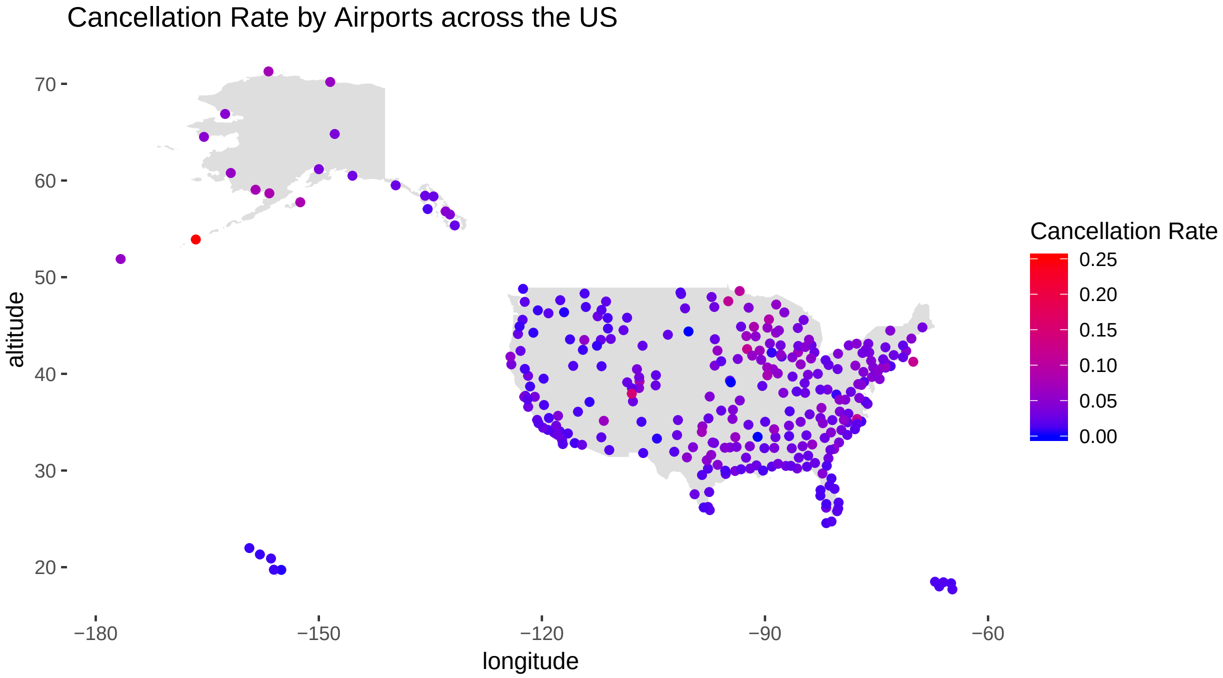
As it shown in the above plot, numebr of flights by day of week in 2007 and 2000 has the same trend though year 2000 has much smaller number of flights. There is a crcle that on Saturday, number of flights begins with its lowest point and increase during the week, decreases after Friday.

2. Location

In this section, we analyze cancellation rate, departure delay rate, and number of flights by airports in the year of 2000 and 2007 by R. Overall, the flights in the airports in the south and west regions have higher cancellation rate and departure delay rate. And there are some interesting results. We will analyze the reasons for these performance.

2.1 Cancellation Rate

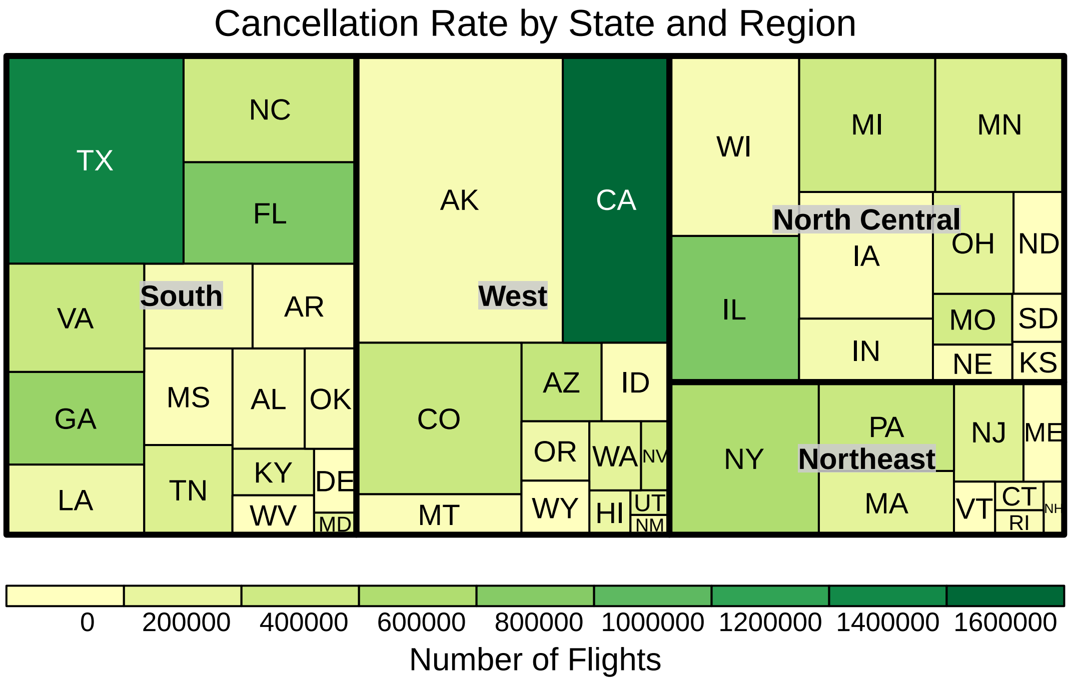
After data processing, the number of unique origin airports is 307. The cancellation rates for airports across the US are plotted as below:



Based on this plot, the cancellation rates are different across the airports in the US. For the airports in the northeast region, the cancellation rates are mostly ranging from 0.10 to 0.15. For the airports in the north central region, the cancellation rates are mostly ranging from 0.10 to 0.20. For the airports in the south region, the cancellation rates are mostly ranging from 0.05 to 0.10. And for the airports in the west region, the cancellation rates are mostly around 0.05. For the airports in Alaska, the cancellation rates are mostly ranging from 0.10 to 0.15. There is also an extremely high cancellation rate in Alaska, which is 0.25. While for the airports in Hawaii, the cancellation rates are not more than 0.05.

The differences of the cancellation rates across the airports make sense when related to the climate and number of flights. The northeast, north central regions and Alaska have long cold and snowy winters. They can have storms in winters. Therefore, it is reasonable that the cancellation rates become higher in these two regions. While the south, west regions, and Hawaii have short winters and dry summers, so it is reasonable that the cancellation rates in the airports in these two regions are in the lower level.

Beside the climate differences across regions, the difference among number of flights across all airports is an important feature affecting the cancellation rates because it can affect the airport workflows. To explore this, the treemap comparing the cancellation rate and number of flights, grouping by regions is plotted as below:



In general, the south and west region have higher value in cancellation rate than other regions, while the northeast region has lower value in cancellation rate. The south and west region have higher value in number of flights, while the northeast region has lower value in number of flights.

In south region, the cancellation rate is higher in TX, FL, while it is lower in DE, MD. The number of flights is higher in TX, FL, GA, while it is lower in DE, WV, MD.

In west region, the cancellation rate is higher in AK, CA, CO, while it is lower in NM, UT. The number of flights is higher in CA, CO, while it is lower in AK, NM.

In north central region, the cancellation rate is higher in WI, IL, MI, while it is lower in KS, SD, NE. The number of flights is higher in IL, while it is lower in KS, SD.

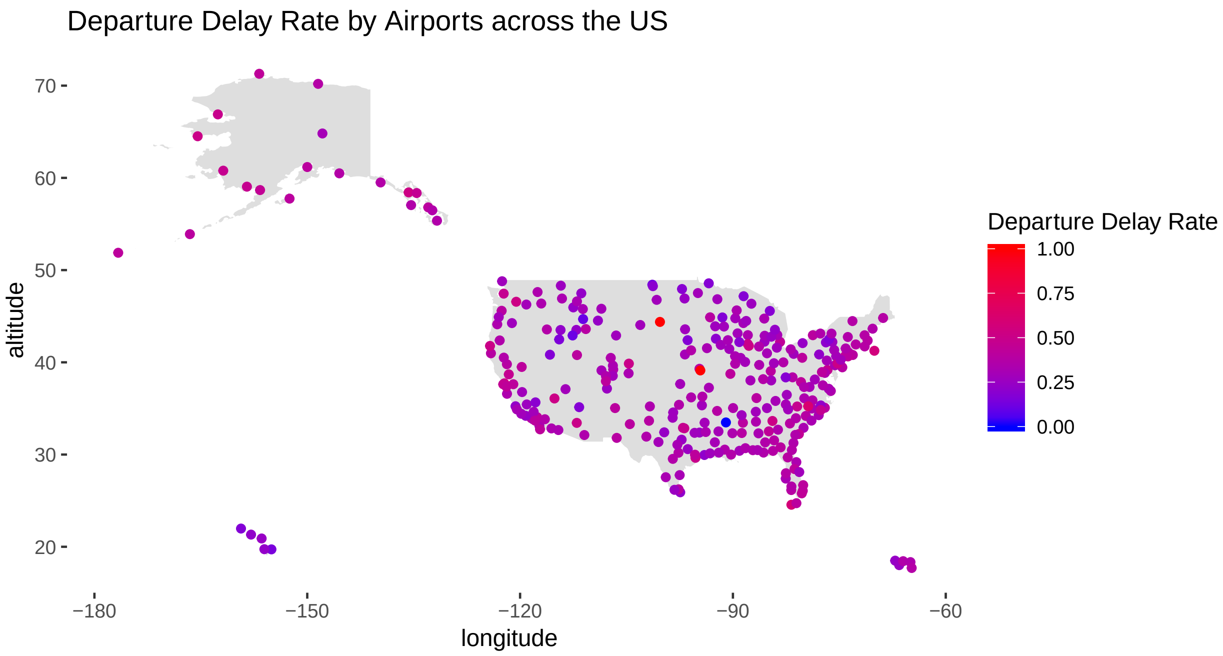
In northeast region, the cancellation rate is higher in NY, while it is lower in NH, RI, CT. The number of flights is higher in NY, PA, while it is lower in ME, VT, NH.

Therefore, the cancellation rate generally tends to be higher in the airports locating in the regions that have bad climate and large number of flights, except for the Alaska state.

It could be noticed that the cancellation rate is pretty high in Alaska, while the number of flights is very low. The airport with the highest cancellation rate is the Unalaska airport located in Unalaska City, Alaska. The cancellation rate of this airport reaches 0.25. It means that the cancellation occurs in about one in four flights. The reasons may include a few flights and extreme weather conditions (Jamie Gonzales 2016). Most small airports in Alaska only have a few scheduled flights. And there are often winds or rain or snow or fog for days. Usually there are not alternative plans for weather conditions but to cancel the flights, so the cancellation rate could be very high.

2.2 Departure Delay Rate

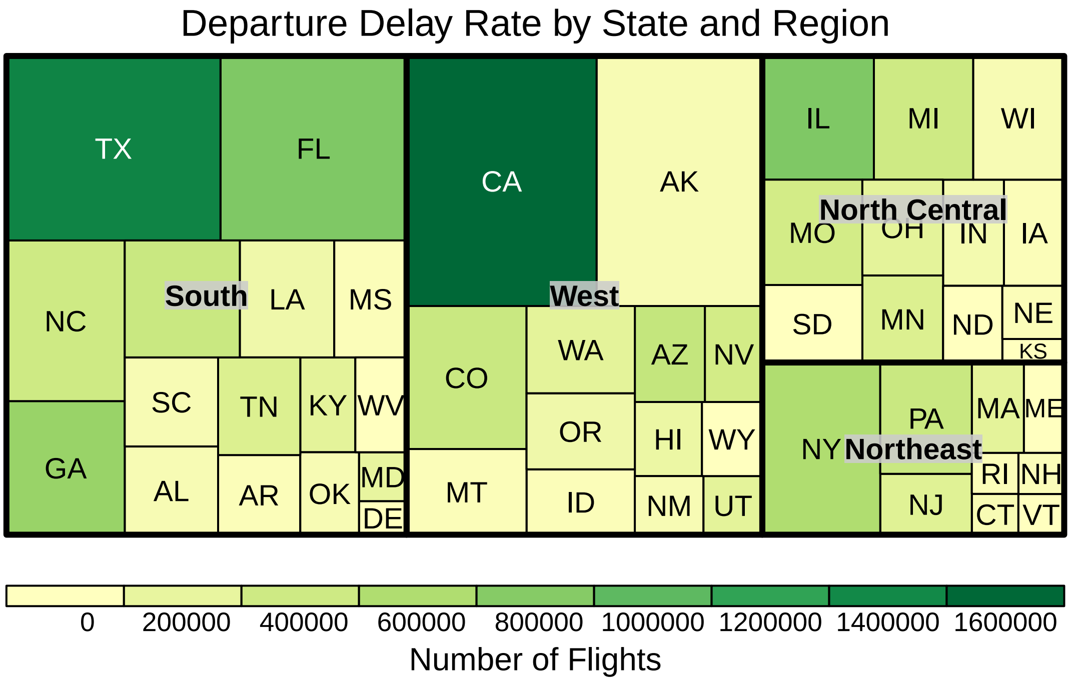
The departure delay rates for airports across the US are plotted as below:



Based on this plot, the departure delay rates are different across the airports in the US. For the airports in the northeast region, the departure delay rates are mostly ranging from 0.25 to 0.50. For the airports in the north central region, the departure delay rates are mostly ranging from 0.50 to 0.75. For the airports in the south region, the departure delay rates are mostly ranging from 0 to 0.50. And for the airports in the west region, the departure delay rates are mostly ranging from 0.25 to 0.50. For the airports in Alaska, the departure delay rates are mostly ranging from 0.25 to 0.50. While for the airports in Hawaii, the departure delay rates are mostly around 0.25.

Similar to the cancellation rate, the differences of the departure delay rates across the airports make sense when related to the climate and number of flights. The northeast, north central regions and Alaska have long cold and snowy winters. They can have storms in winters. Therefore, it is reasonable that the departure delay rates become higher in these two regions. While the south, west regions, and Hawaii have short winters and dry summers, so it is reasonable that the departure delay rates in the airports in these two regions are in the lower level.

To explore the effects of number of flights on the departure delay rate, the treemap is plotted as below:



In general, the south and west region have higher value in departure delay rate than other regions, while the northeast region has lower value in departure delay rate. The south and west region have higher value in number of flights, while the northeast region has lower value in number of flights.

In south region, the departure delay rate is higher in TX, FL, while it is lower in DE, WV. The number of flights is higher in TX, FL, GA, while it is lower in DE, WV, OK.

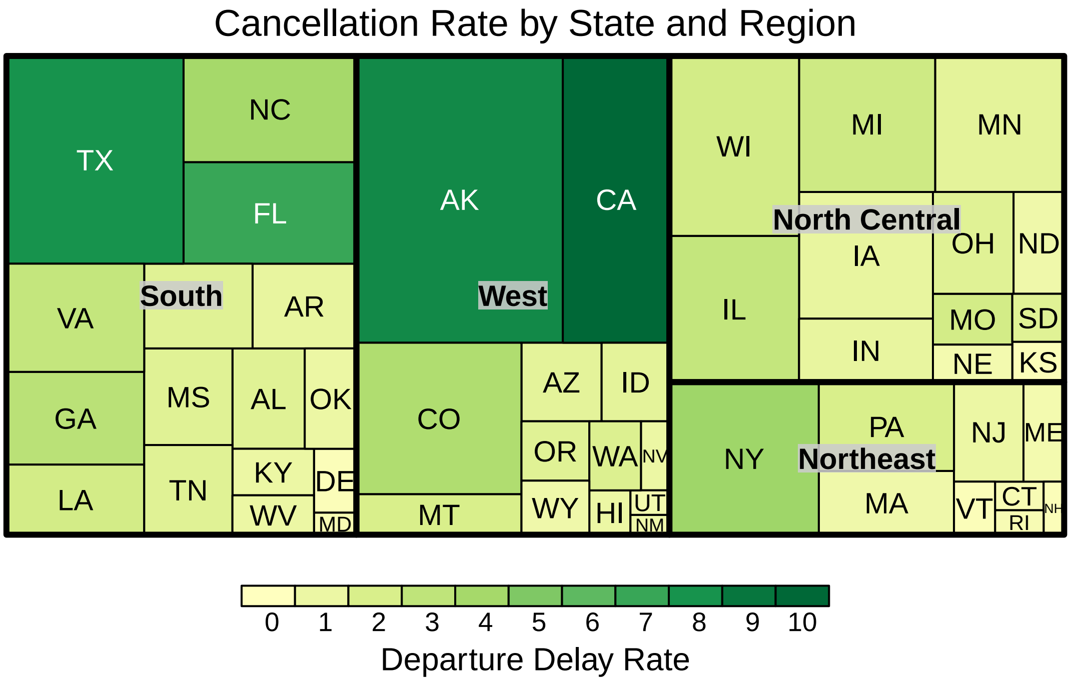
In west region, the departure delay rate is higher in CA, AK, while it is lower in NM, WY. The number of flights is higher in CA, CO, while it is lower in AK, NM.

In north central region, the departure delay rate is higher in IL, while it is lower in KS. The number of flights is higher in IL, while it is lower in KS, SD.

In northeast region, the departure delay rate is higher in NY, while it is lower in CT, VT. The number of flights is higher in NY, PA, while it is lower in ME, VT, NH.

Therefore, the departure delay rate generally tends to be higher in the airports locating in the regions that have bad climate and large number of flights.

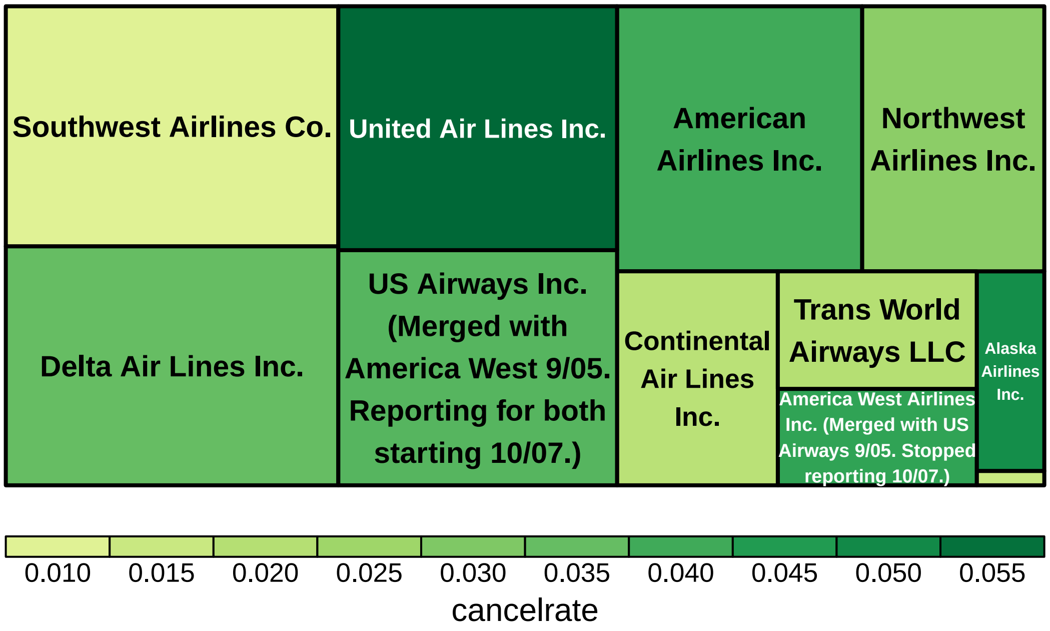
Moreover, to explore the relationship between cancellation rate and departure delay rate by locations, the treemap is plotted as below:


Cancellation Rate by State and Region

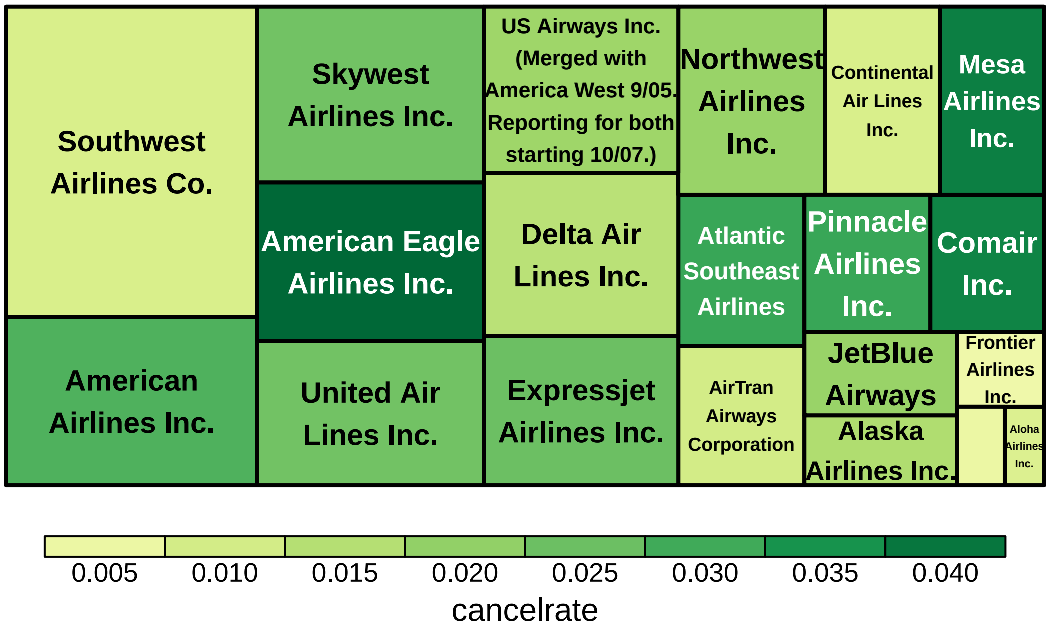
Based on the above plot, we can see for the states having higher cancellation rate, they also have higher departure delay rate. Therefore, the relationship between cancellation rate and departure delay rate is positive.

3. Carrier

3.1 Cancellation Rate



Cancellation Rate in Year 2000 (Size: Num of Flights)



Cancellation Rate in Year 2007 (Size: Num of Flights)

The treemap above showns cancellation rate as color and the number of flights for specific carrier as size. Overall, the scale for cancellation rate shrinked which shows a decrease in this rate from 2000 to 2007. Further exploring the cancellation rate, we find with some carriers, the airline cancellation rate tend to stay high through these two years. At the same time, with some carriers, the airline cancellation rate sharply decreased. In a word, carrier can be taken as a categorical feature while building a model to predict airline cancellation rate.

The distributions of cancellation rate for carriers different between 2000 and 2007. The carriers with dark green holds highest cancellation rate.

In 2000, `United Air Lines Inc.`, `Alaska Airlines Inc.`, `Delta Air Lines Inc.`, `US Airwats Inc.` and `American Airlines Inc.` tend to have higher cancellation rate.
In 2007, `American Eagle Airlines Inc.`, `American Airlines Inc.`, `Mesa Airlines Inc.`, `Comair Inc.` tend to have higher cancellation rate.

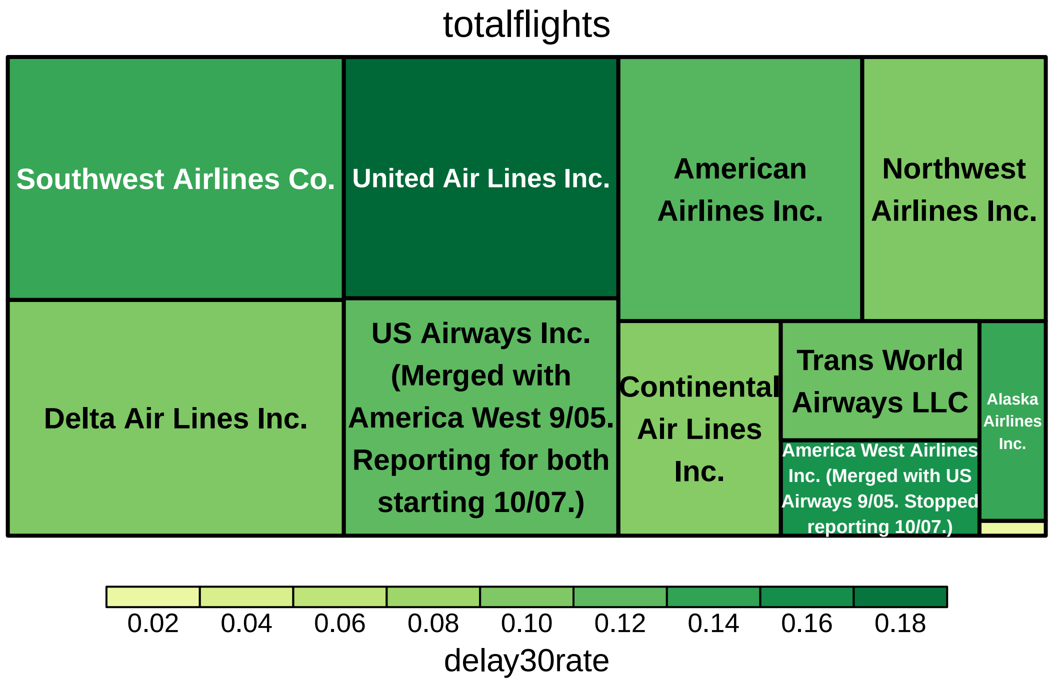
From these differences, `United Air Lines Inc.`, `Alaska Airlines Inc.`, `Delta Air Line Inc.` have improved their cancellation condition significantly compared to other carriers. On the contrary, `American Airlines Inc.` stays at the same level among carriers.

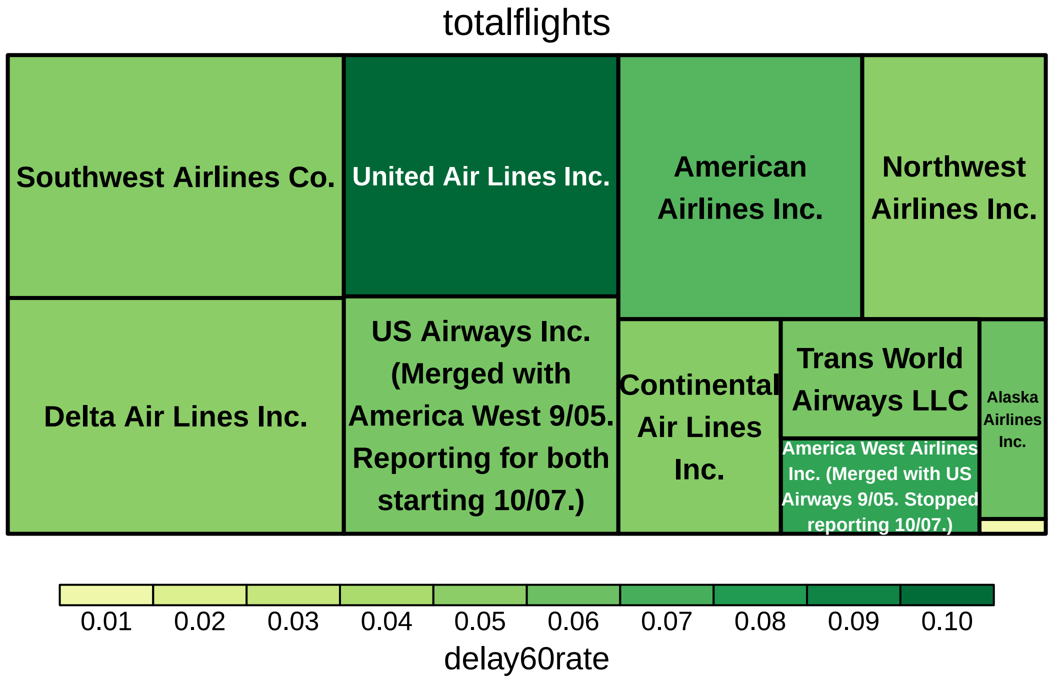
Since the datasets are very large here, merge datasets in R is very expensive and limited by the size of RAM. So we built connection between the airline, airport, carriers and plane-data by preprocessing data to match the index number (as introduced before). But in Hadoop we can easily merge those datasets together. To validate that our data conversion does not have error, we calculated the cancellation rate of each carrier in Pig. And below is the results in 2000:

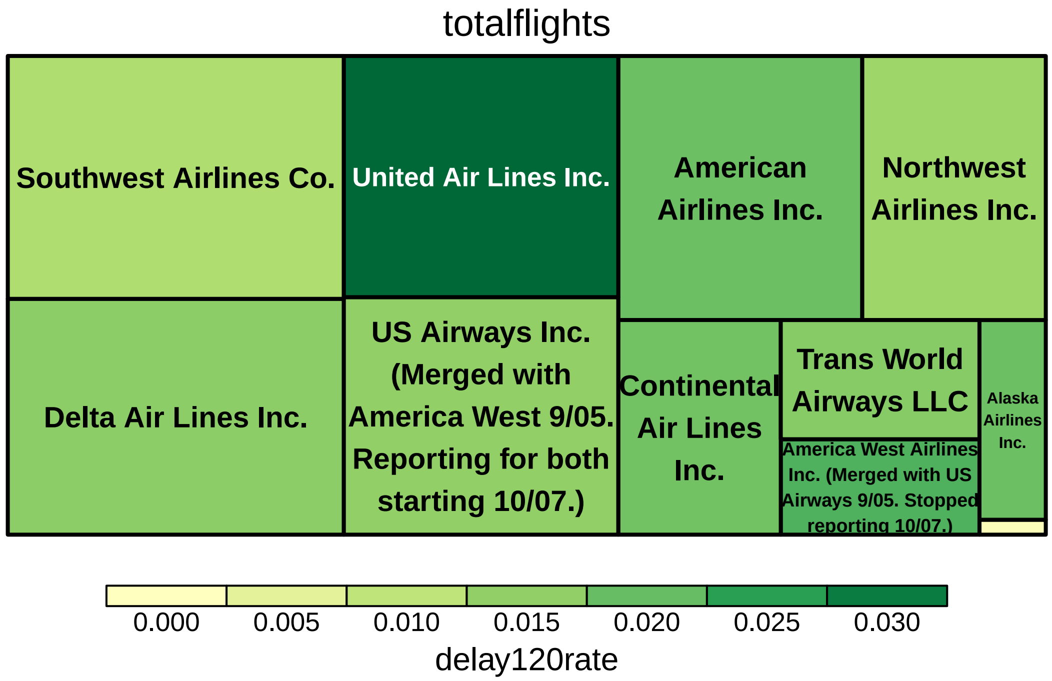
Carrier	Total flights	Cancelled flights	Cancellation rate
Aloha Airlines Inc.	11036	173	0.01567597
Alaska Airlines Inc.	154171	7506	0.0486862
Delta Air Lines Inc.	908029	31569	0.03476651
United Air Lines Inc.	776559	44159	0.056864966
American Airlines Inc.	742265	29677	0.039981678
Southwest Airlines Co.	911699	9039	0.009914457
Northwest Airlines Inc.	551337	15340	0.027823273
Trans World Airways LLC	267131	5254	0.019668253
Continental Air Lines Inc.	393036	7296	0.018563185
America West Airlines Inc. (Merged with US Airways 9/05. Stopped reporting 10/07.)	219160	9422	0.042991422
US Airways Inc. (Merged with America West 9/05. Reporting for both starting 10/07.)	748624	28055	0.037475422

Results generated from Pig is same with what we got in R. We do not show the raw results in 2007 for clarity of report. Please refer to the treemap above.

3.2 Delay


totalflights
Southwest Airlines Co.
United Air Lines Inc.
American Airlines Inc.
Northwest Airlines Inc.
Delta Air Lines Inc.
US Airways Inc. (Merged with America West 9/05. Reporting for both starting 10/07.)
Continental Air Lines Inc.
Trans World Airways LLC
America West Airlines Inc. (Merged with US Airways 9/05. Stopped reporting 10/07.)
Alaska Airlines Inc.
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
delay30rate

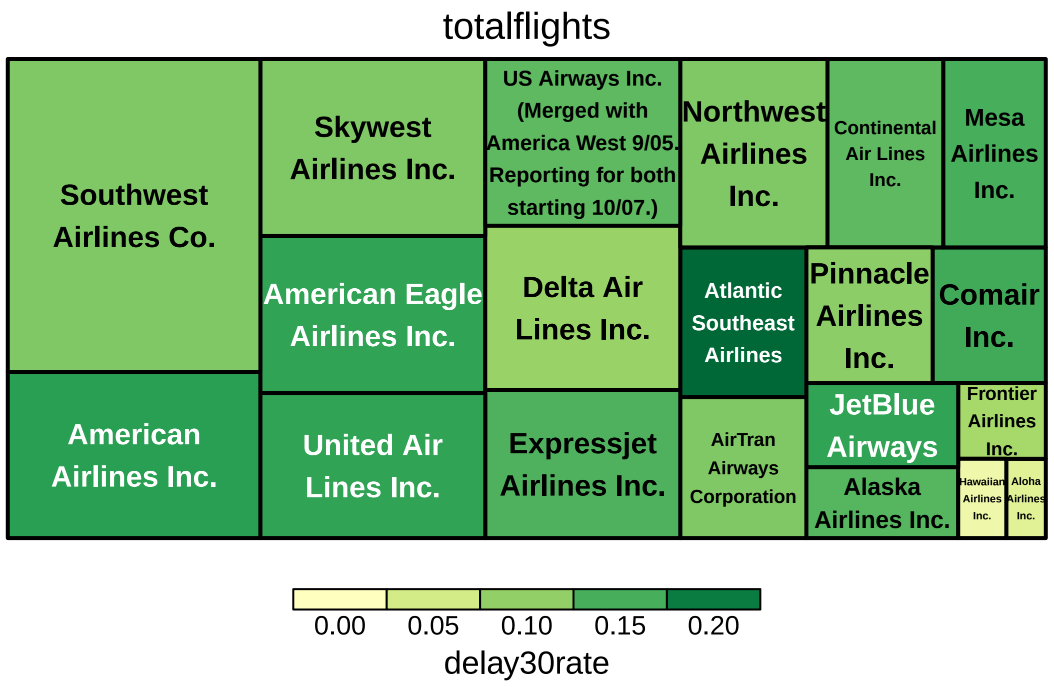

totalflights
Southwest Airlines Co.
United Air Lines Inc.
American Airlines Inc.
Northwest Airlines Inc.
Delta Air Lines Inc.
US Airways Inc. (Merged with America West 9/05. Reporting for both starting 10/07.)
Continental Air Lines Inc.
Trans World Airways LLC
America West Airlines Inc. (Merged with US Airways 9/05. Stopped reporting 10/07.)
Alaska Airlines Inc.
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10
delay60rate

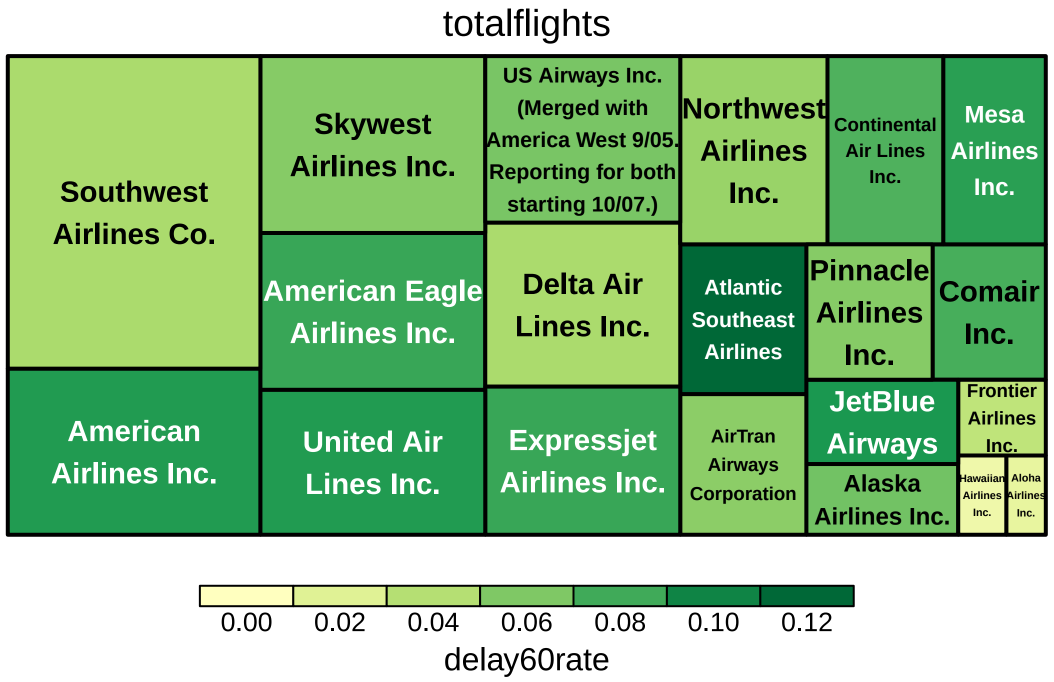


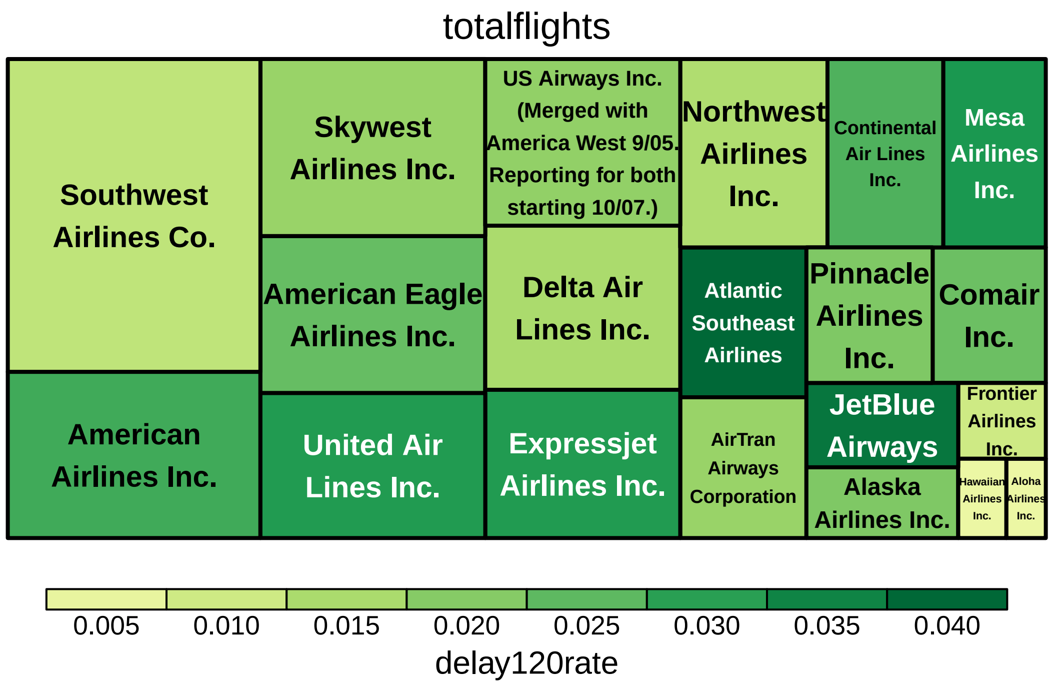
These three treemaps shows delay rate in year 2000 with three standard: 30 mins, 60 mins and 120 mins. Based on the color of the treemaps, `United Air Line Inc.` holds extramely large maginitude of departure delay time with dark green color in three treemaps.

Besides, carriers holds relatively high delay rate but with smaller maginitude of departure delay time are:

- Southwest Airlines Co
- US Airways Inc
- America West Airlines
- Alaska Airlines Inc


totalflights
Southwest Airlines Co.
Skywest Airlines Inc.
US Airways Inc. (Merged with America West 9/05. Reporting for both starting 10/07.)
Northwest Airlines Inc.
Continental Air Lines Inc.
Mesa Airlines Inc.
American Eagle Airlines Inc.
Delta Air Lines Inc.
Atlantic Southeast Airlines
Pinnacle Airlines Inc.
Comair Inc.
American Airlines Inc.
United Air Lines Inc.
Expressjet Airlines Inc.
AirTran Airways Corporation
JetBlue Airways
Frontier Airlines Inc.
Alaska Airlines Inc.
Hawaiian Airlines Inc.
Aloha Airlines Inc.
0.00
0.05
0.10
0.15
0.20
delay30rate


totalflights
Southwest Airlines Co.
Skywest Airlines Inc.
US Airways Inc. (Merged with America West 9/05. Reporting for both starting 10/07.)
Northwest Airlines Inc.
Continental Air Lines Inc.
Mesa Airlines Inc.
American Eagle Airlines Inc.
Delta Air Lines Inc.
Atlantic Southeast Airlines
Pinnacle Airlines Inc.
Comair Inc.
American Airlines Inc.
United Air Lines Inc.
Expressjet Airlines Inc.
AirTran Airways Corporation
JetBlue Airways
Frontier Airlines Inc.
Alaska Airlines Inc.
Hawaiian Airlines Inc.
Aloha Airlines Inc.
0.00
0.02
0.04
0.06
0.08
0.10
0.12
delay60rate



These three treemaps shows delay rate in year 2007 with three standard: 30 mins, 60 mins and 120 mins. Based on the color of the treemaps, `Atlantic Southeast Airlines` holds extramely large maginitude of departure delay time with dark green color in three treemaps. `American Airelines Inc.`, `United Air Lines Inc.`, `Expressjet Airlines Inc.` share the same characteristic.

Interesting things happened to `JetBlue Airways`, it holds moderately high delay rate for 30 mins standard, but holds extramely high delay rate for 120 mins standard compared to other carriers. This means that `JetBlue Airways` has a regular pattern that if airline delays, customers will have high possibility to be faced with a long delay.

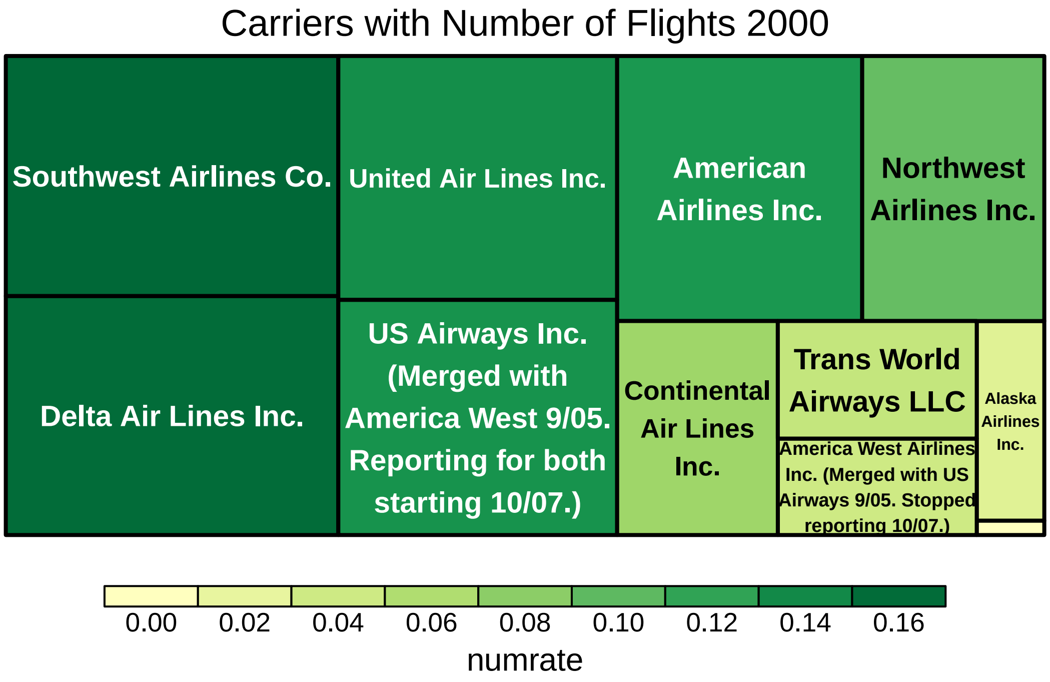
Besides, carriers holds relatively high delay rate but with smaller maginitude of departure delay time are:

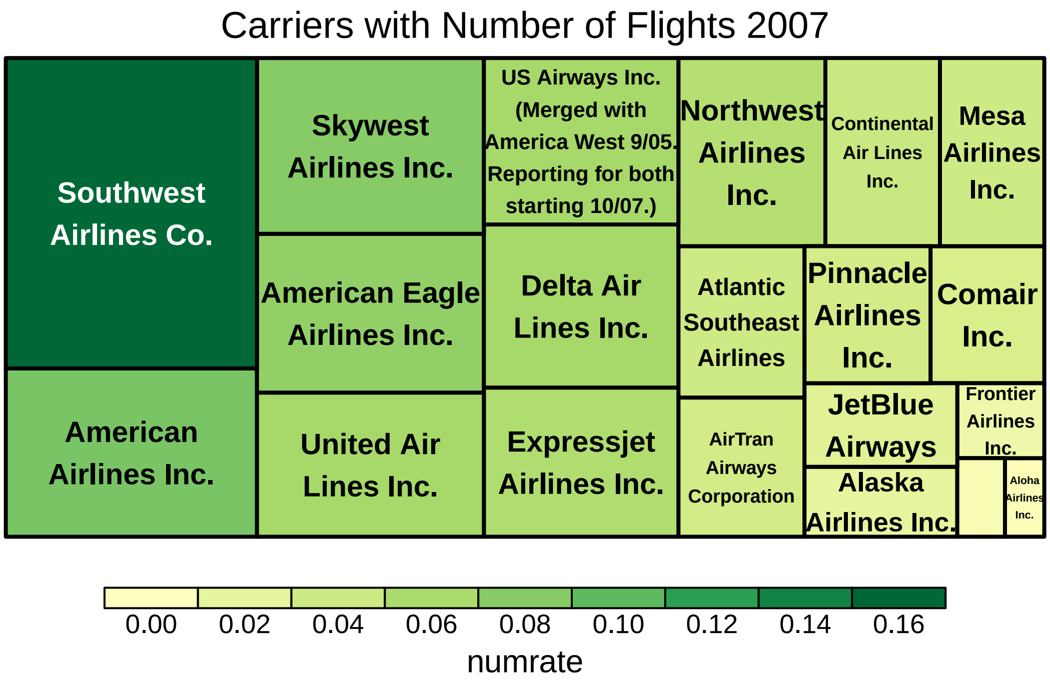
- Southwest Airlines Co
- Northwest Airline Inc

This means that when considering to predict the magnitude of depature delay time, carriers should be taken in account as an efficient categorical feature.

3.3 Number of Flights

flight numbers by carriers treemap 2000





The treemap above shows number of flights by carriers in 2000. The size shows number and the color shows percentage. Assuming that there is some relationship between cancellation and number of flights, we find evidence to overthrew this assumption. Evidence is that number of flights `Southwest Airlines Co.` takes the largest percentage with the lowest cancellation rate in year 2000 and 2007. But in general, carriers with low cancellation rate are accompanied with smaller number of flights.

Conclusion

Carriers can be divided to different groups with their special characteristics which is how to effect the cancellation rate or depature delay rate or magnitude.

Carriers with high quality airlines (low cancellation rate and delay rate/magnitude): `Southwest Airlines Co.`, `Continental Air Lines Inc.`, `American Airlines Inc`, `Mesa Airlines Inc.`.

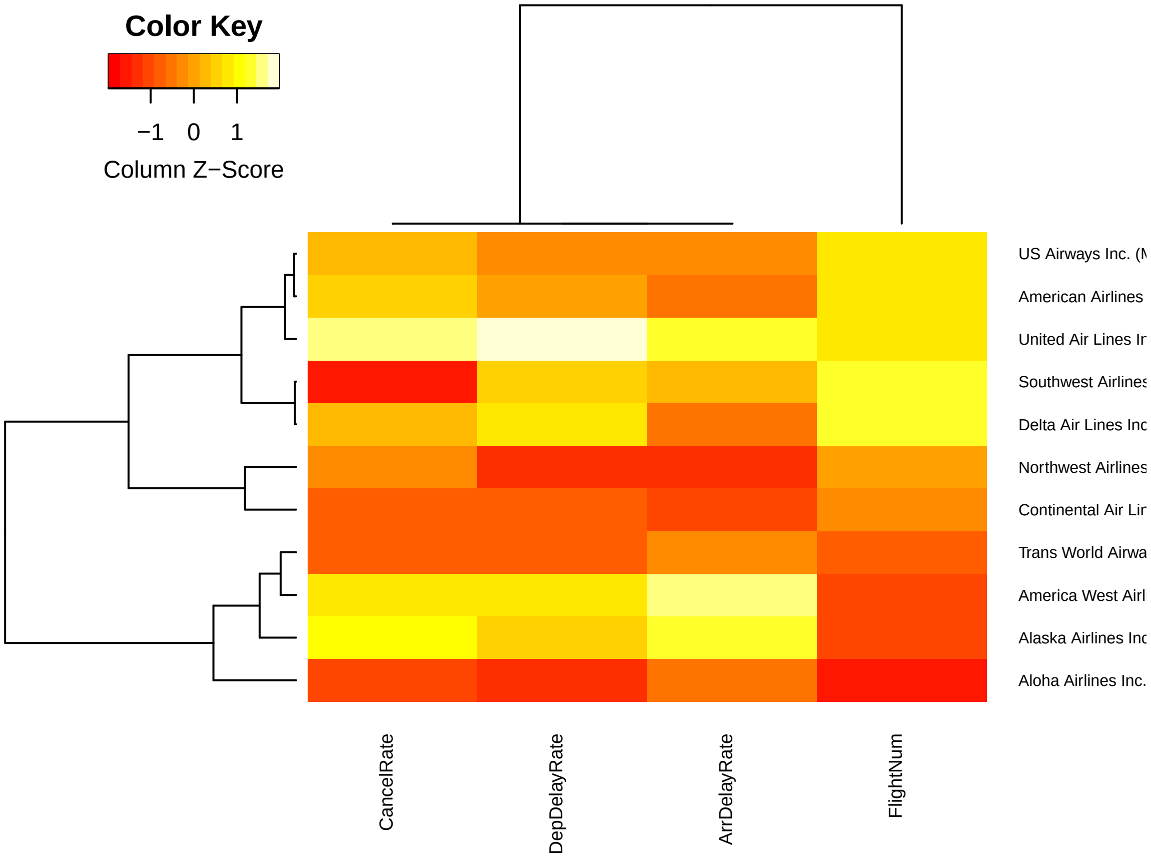
Carriers with low quality airlines (high cancellation rate and delay rate/magnitude):`United Air Lines Inc.` a large company, `American Eagle Airlines Inc.` a small company.

From 2000 to 2007, number of flights increased and number of carriers increase. But cancellation rate and delay decreased. The performace of carriers changed, too. For example, `Delta Air Lines Inc.` has an obvious airline quality improvement during this period.

3.4 Information

Below are the two heatmaps based on day of month for different carriers in the relationship with cancel rate, departure delay rate, arrival delay rate, and the number of flights.

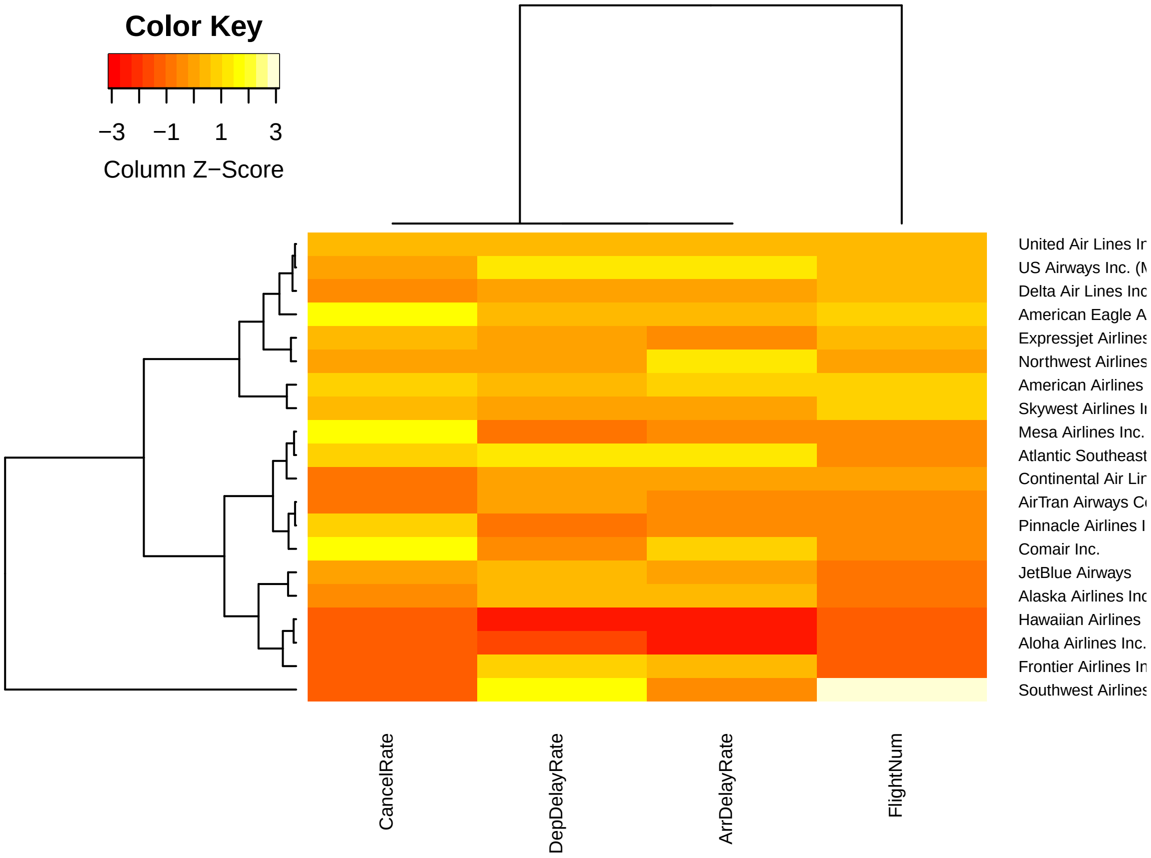
cancel rate, delay, flight numbers by carriers heatmap 2000



In the heatmap of 2000, we can see that Southwest Airlines and Aloha Airlines have a relative lower cancel rate and United Airlines has the highest cancel rate; United Airlines also has the highest departure delay rate and Aloha Airlines and Northwest Airlines have relative lower departure delay rate; America west Airlines and Alaska Airlines have relative higher arrival delay rate and Northwest Airlines has relative lower arrival delay rate; Aloha Airlines has relative lower flight numbers and Southwest Airlines and Delta Airlines have relative higher number of flights.

In the heatmap of 2000, most of the airlines with the relative high number of flights tend to have a relative higher rate in cancellation, departure delay and arrival delay too, but the airlines with small flight numbers like America West Airlines and Alaska Airlines also have relative high rate of cancellation, departure delay and arrival delay. So, we can't conclude a clear relationship pattern between flight numbers and the cancellation, departure delay, and arrival delay rate. But for most of the carriers, as we can see from the dendrogram of rates and flight numbers, we can see that different rates are grouped together. Higher cancellation rate often comes with relative high departure delay rate and arrival delay rate. For the carrier's group, we can group the carriers into three different groups, the first group is from US Airways to Delta Airlines; the second group is from Northwest Airlines and Continental Airlines; the third group is from Trans World Airways to Aloha Airlines. As we can see from the graph, the first group tends to have a relative higher value of cancellation rate, and number of flights. Group three tends to have a lower value of flight numbers. Group two tends to have a relative lower value in departure delay rate and arrival delay rate.

cancel rate, delay, flight numbers by carriers heatmap 2007



In the heatmap of 2007, we can see that the number of carriers increased. Besides, American Eagle Airlines, Mesa Airlines, and Comair Airlines have relative high value of cancellation rate, and Hawaiian Airlines, Aloha Airlines, Frontier Airlines, and Southwest Airlines have relative low value of cancellation rate; Hawaiian Airlines has relative low departure delay rate and Southwest Airlines has relative high departure delay rates; Hawaiian Airlines and Aloha airlines have relative low arrival delay rates and Northwest Airlines and Atlantic Southeast have relative high arrival delay rate; the flight number value doesn't differ much between different carriers except Southwest Airlines has a significant high flight number than other carriers.

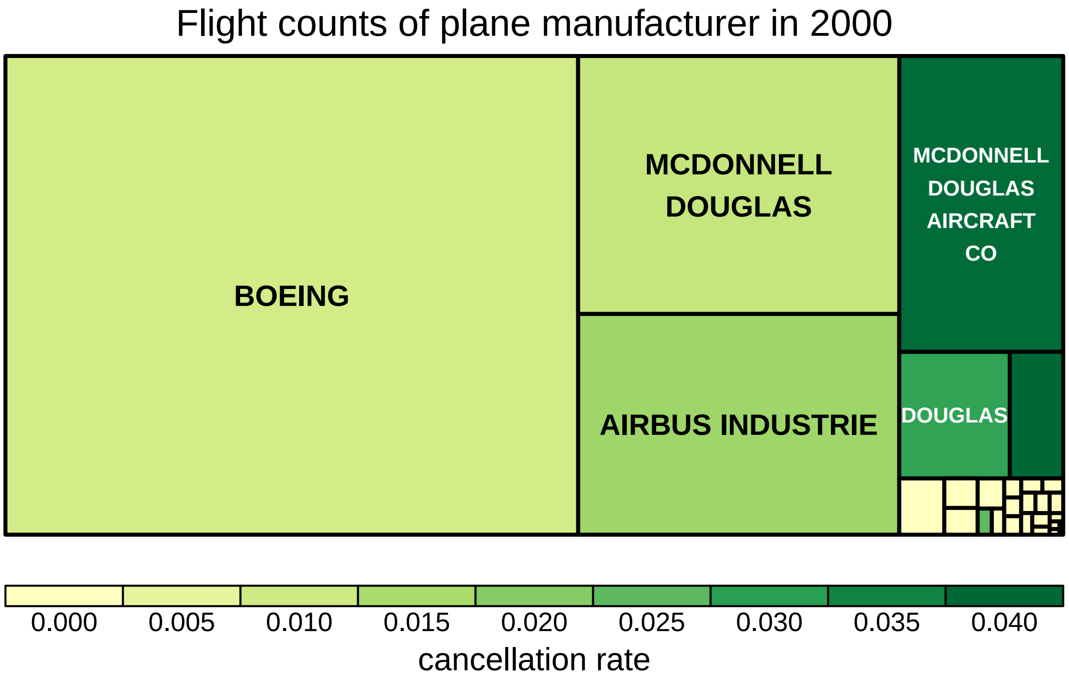
In the heatmap of 2007, still, there is no clear trend of the relationship between flight numbers and the rates of cancellation, departure delay, and arrival delay. But for most of the carriers, as we can see from the dendrogram of rates and flight numbers, we can see that different rates are grouped together. Higher cancellation rate often comes with relative high departure delay rate and arrival delay rate. We can group the carriers into two. The first group is from United Airlines to Skywest Airlines, and the second group is from Mesa Airlines to Southwest Airlines. The first group tends to have a relative higher value in all the rates and the flight numbers but the difference between these two groups is not very big.

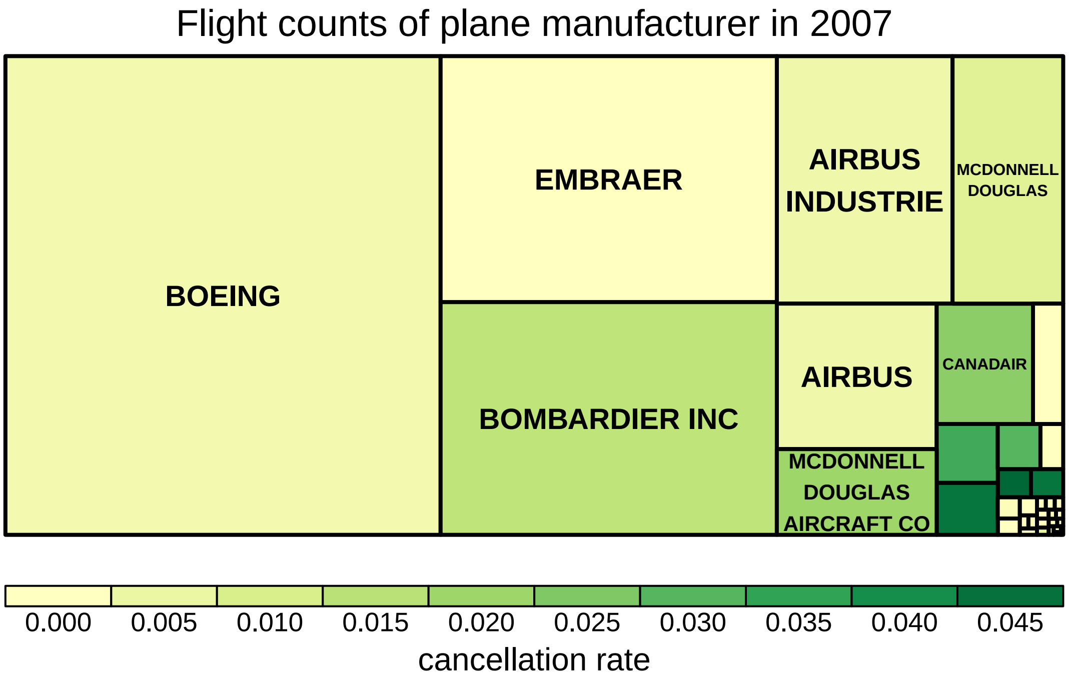
The difference of flight numbers between different carriers in 2007 decreased comparing to year 2000. The overall rate situation between different carriers in 2007 seems to have a lower difference too than 2000 which is understandable since the market is more opening to more carriers and the ability of different carriers doesn't differ much.

4. Plane

4.1 Cancellation rate vs Manufacturer of plane

In below treemaps, size of squares represents the total number of flights flew by planes manufactured by this company. The color represents the cancellation rate, which is calculated by `number of cancelled flights flew by planes manufactured by this company/total number of flights flew by planes manufactured by this company`. In 2000, planes manufactured by Boeing, Airbus and Mcdonnell Douglas flew the most passenger flights in us. planes manufactured by mcdonnell douglas aircraft co had the highest cancellation rate. One interesting point is that only 10 companys were still manufacturing new aircrafts after 1998. From the treemaps below we can also see that planes manufactured by Douglas, Mcdonnell Douglas and Mcdonnell Douglas Aircraft Co. still have a large share in 2000. But in 2007, only a very small amount of their planes were still flying. This is because Douglas merged with McDonnell in 1967 and then McDonnell Douglas later merged with Boeing in 1997 (*Douglas Aircraft Company* 2019). So planes with their names were pretty old in 2000/2007. This might also be reason that their cancel rates were relatively high.
In 2007, new manufactures gained large shares: Embraer is now No.2 only behind Boeing; Bombardier and Canadair also have a very large size. Overall cancellation rate is lower compared to 2000.

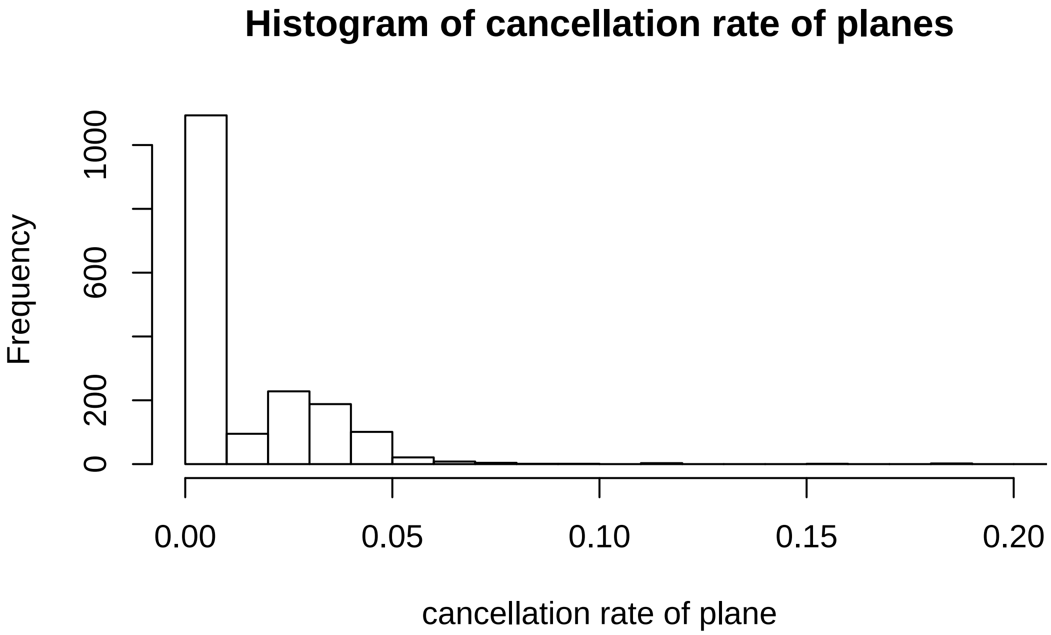


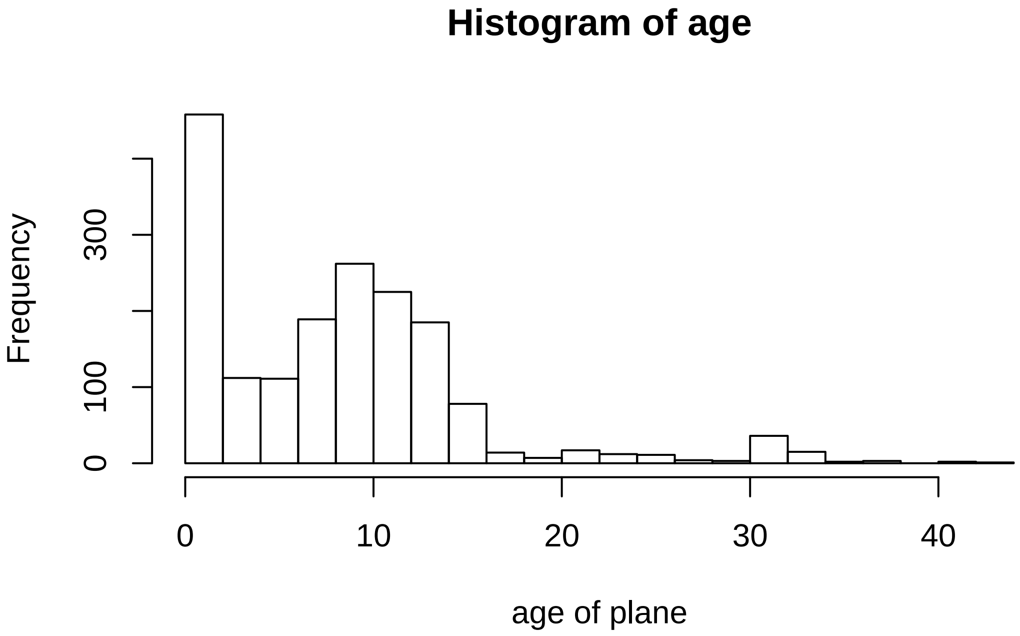


4.2 Cancellation rate vs age of plane

2000

We have 1747 planes flying in U.S in 2000. 48.3% were not cancelled once in the whole year. Average age of planes is 8.6468231, and 57.4% of planes were manufactured within 10 years.





Linear model (cancelrate ~ age of plane)

We then tried to fit a linear model on planes, using age to predict cancellation rate. We can see from the summary below that coefficients of age is very small but significant. R^2 is very small. A possible reason is that the chance for cancellation is very rare, about 1 in 100, and many planes do not any cancellation in the whole year.

```
## 
## Call:
## lm(formula = cancelrate ~ age, data = newplane2)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.02145 -0.01288 -0.01122  0.01184  0.98902 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.098e-02  1.123e-03   9.778   <2e-16 ***
## age         2.381e-04  9.999e-05   2.381   0.0174 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.02994 on 1745 degrees of freedom
## Multiple R-squared:  0.003238,   Adjusted R-squared:  0.002667 
## F-statistic: 5.668 on 1 and 1745 DF,  p-value: 0.01738
```

Binomial model (cancelled or not ~ age of plane)

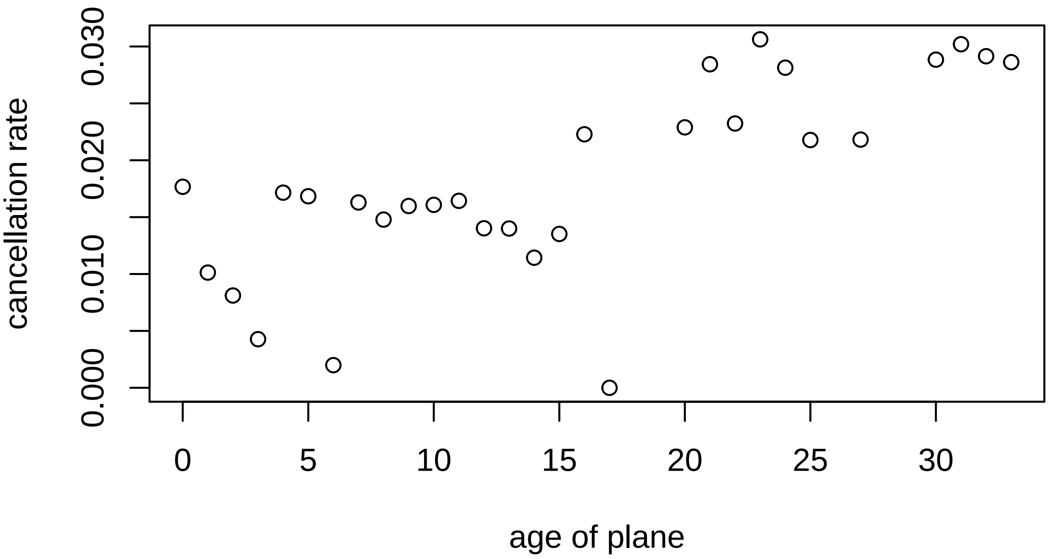
As explained above, since many plane has cancellation rate of 0, we then try to use binomial regression to fit model on age to predict whether or not the plane has one cancelled flight or not.

```
## 
## Call:
## glm(formula = cancelled ~ age, family = binomial)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.7334  -1.1848   0.8435   1.1419   1.2703  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -0.215853   0.076326  -2.828  0.00468 ** 
## age          0.033325   0.006993   4.766 1.88e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 2419.7  on 1746  degrees of freedom
## Residual deviance: 2396.0  on 1745  degrees of freedom
## AIC: 2400
## 
## Number of Fisher Scoring iterations: 4
```

We can see from the summary above that, though still very small, coefficient of `age` in this model is much more significant than in last model. `age` has coefficient of 0.0333249, which means that in this model when plane get 1 year older, odds ($\frac{P(\text{cancelled at least once})}{1-P(\text{cancelled at least once})}$) of being cancelled at least once in 2000 is multiplied by 0.0333249.

Scatter plot: Cancellation rate ~ age

We then sum all the flights flew by planes with same age (by year) and try to compare the cancellation rate. The scatter plot below seems to have 3 stages - planes within 5 years old had lowest cancellation rate around 0.005; planes aged from 5 to 15 had cancellation rate around 0.015; planes with age over 15 had higher cancellation rate around 0.025 to 0.03. This might represent the generations of passenger planes - newer planes have lower chance of cancellation due to mechanical failure. But further study is needed to draw the conclusion.



A possible data error

Following 4 planes were manufactured after 2000 (1 in 2001 and 3 in 2007), but they were flying in 2000. They are only planes with this kind of error.

```
##      tailnum        type       manufacturer issue_date             model
## 1325  N365AA Corporation         AGUSTA SPA 10/21/2003             A109E
## 1530  N394AA Corporation AVIAT AIRCRAFT INC       None              A-1B
## 1535  N395AA Corporation AVIAT AIRCRAFT INC 06/05/2008              A-1B
## 2370  N544AA  Individual     FRIEDEMANN JON 01/26/2007 VANS AIRCRAFT RV6
##      year flightcount.plane
## 1325 2001               325
## 1530 2007               322
## 1535 2007               299
## 2370 2007              1418
```

```
## [[1]]
##      Year Month DayofMonth DayOfWeek UniqueCarrier FlightNum TailNum
## [1,] 2000     1          3         1           101        27    1325
##
## [[2]]
##      Year Month DayofMonth DayOfWeek UniqueCarrier FlightNum TailNum
## [1,] 2000     1          5         3            101      1539    1530
##
## [[3]]
##      Year Month DayofMonth DayOfWeek UniqueCarrier FlightNum TailNum
## [1,] 2000     1          8         6            101        27    1535
##
## [[4]]
##      Year Month DayofMonth DayOfWeek UniqueCarrier FlightNum TailNum
## [1,] 2000     1         28         5            101        67    2370
```

To validate that results above were not due to mistakes in our data processing, we check if original datasets produce the same results in Hive. We join the original airline dataset (not the one with string columns converted) with plane dataset on tailnum. Then select rows where plane was manufactured after 2007 but were flying in 2000.

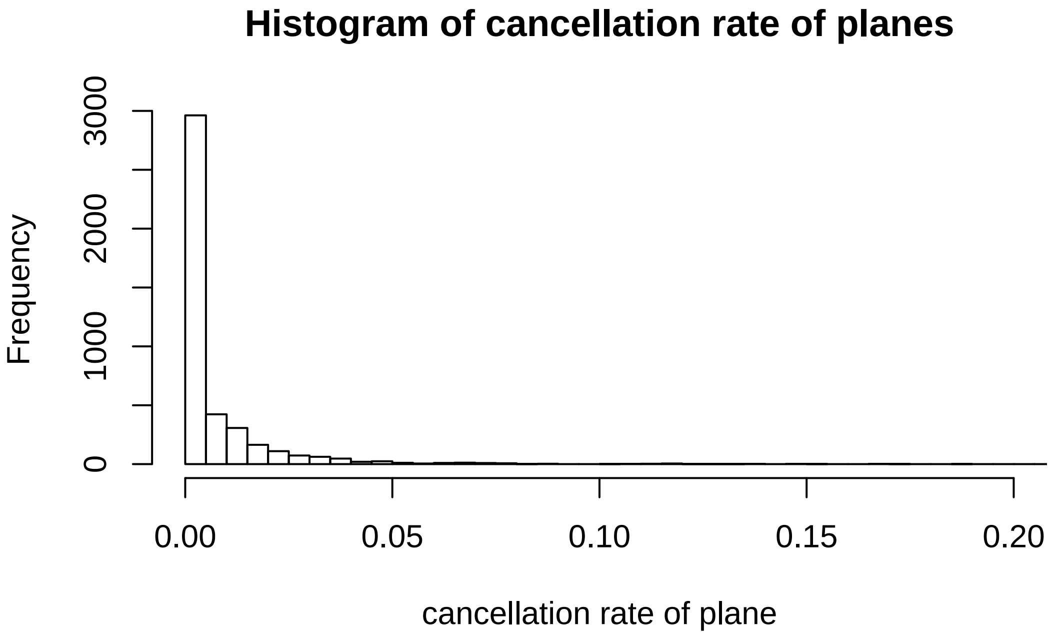
Results are shown below and it's same as the result we got by R.

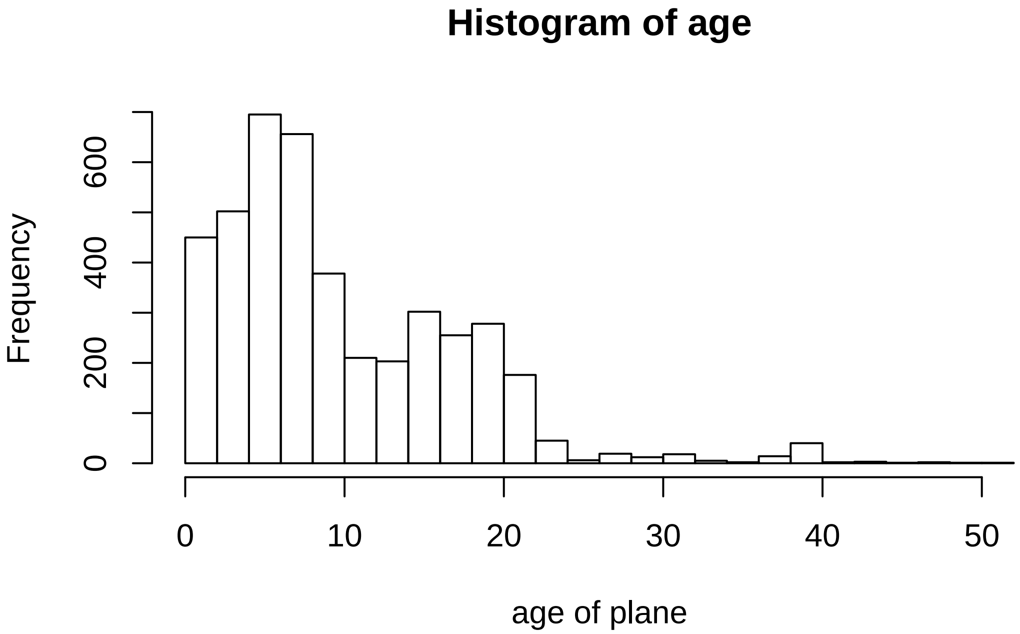
Year	Tail Number	Carrier	Plane Model	Issue Date	Year Manufactured
2000	N365AA	AA	A109E	10/21/2003	2001
2000	N395AA	AA	A-1B	06/05/2008	2007
2000	N544AA	AA	VANS AIRCRAFT RV6	01/26/2007	2007
2000	N394AA	AA	A-1B	None	2007

2007

We do similar analysis in 2007 as in 2000.

In 2007, We have 4276 planes flying in U.S, which is about 2.5 times more than 1747 in 2000. This means a lot of new planes were put into operation after 2000. We can see in 2007, overall cancellation rate is lower than in 2000. 56.9% of planes were not cancelled once in the whole year. Average age of planes is 10.1810103, this is higher than 8.6468231 in 2000. But the percentage of planes less than 10 years old is 59.7%, slighter higher than 57.4% in 2000.





Binomial model (cancelled or not ~ age of plane)

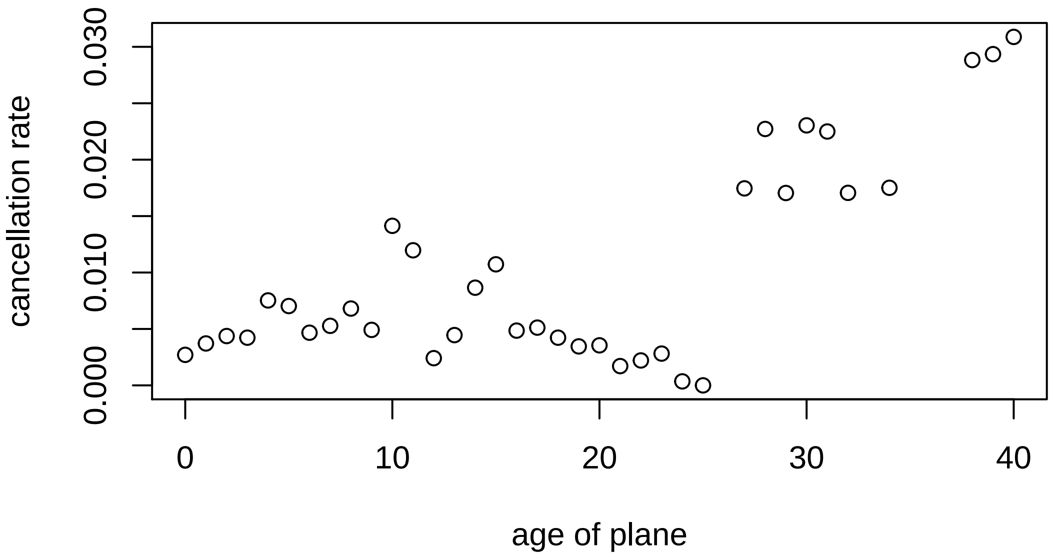
```
##
## Call:
## glm(formula = cancelled ~ age, family = binomial)
##
```

```
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.330  -1.055  -1.018   1.299   1.367
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.434748   0.052388  -8.299  < 2e-16 ***
## age          0.015442   0.004118   3.750 0.000177 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5846.7  on 4275  degrees of freedom
## Residual deviance: 5832.6  on 4274  degrees of freedom
## AIC: 5836.6
##
## Number of Fisher Scoring iterations: 4
```

In 2007, case is much different from 2000. Coefficient of `age` in the binomial model is no longer significant with a *p*-value of 0.371.

Scatter plot: Cancellation rate ~ age

In 2007, we can no longer see the 3-stages like plot in 2000. Plot in 2007 is more likely to have 2 stages. Planes with age from 0 to 25 all have a similar low cancellation rate around 0.005. An exception is in plane aged 10, 11, 14, 15 have higher cancellation rate around 0.01. We can see a sudden leap at around 25, planes older than that had a cancellation rate ranged from 0.015 ro 0.03.



5. Regression Model

Based on the above analysis, there are some interesting trends between departure delay and time, location, plane, carriers. To deeply explore their relationships, a regression model is fitted in this section with the relavant features for each year. The features used in this model include Month, DayofMonth, DayOfWeek, CRSArrTime, UniqueCarrier, Origin, Dest, Distance, and Diverted.

For the year of 2000, all of the features show significant effects on the departure delay time in 0.05 significant level. The regression model is fitted as below:

$$\begin{aligned} DepDelay = &-13.5073 + 0.3003 \times Month + 0.1253 \times DayofMonth + 0.5140 \times DayOfWeek \\ &+ 0.0087 \times CRSArrTime + 0.0023 \times UniqueCarrier + 0.0003 \times Origin \\ &+ 0.0010 \times Dest + 0.0013 \times Distance + 28.0787 \times Diverted \end{aligned}$$

And the R^2 value is 0.0244.

For the year of 2007, all of the features show significant effects on the departure delay time in 0.05 significant level. The regression model is fitted as below:

Based on the above results, all predictors show significant effects on the departure delay time in 0.05 significant level, because their p-values are both less than 0.05. The model can be written as

$$\begin{aligned} DepDelay = &-4.1314 - 0.0841 \times Month + 0.0665 \times DayofMonth + 0.0912 \times DayOfWeek \\ &+ 0.0108 \times CRSArrTime + 0.0001 \times UniqueCarrier - 0.0009 \times Origin \\ &- 0.0002 \times Dest + 0.0008 \times Distance + 14.7549 \times Diverted \end{aligned}$$

And the R^2 value is 0.0222.

There are some differences in the results between the year of 2000 and 2007. The parameter of Month is positive in 2000, while it is negative in 2007. This indicates that with other features unchanged, the delay time increases with month in 2000, while it decreases with month in 2007. The parameter of Origin is positive in 2000, while it is negative in 2007. And the parameter of Dest is positive in 2000, while it is negative in 2007. Also, the absolute values of Origin and Dest are very lower in 2007 than those in 2000.

The R^2 values are both extremely low in the year of 2000 and 2007, which indicates both of the models poorly fit DepDelay.

Conclusion

Time

- Overally, Year 2007 has a lower cancellation by both monthly analysis and day of week analysis than 2000.
- Also, year 2007 has higher flight numbers by both monthly analysis and day of week analysis. But for depature delay,there is no consistent magnitude difference between the two years for both the rates and the average in the analysis by month and day of week.
- The trend of the grpahs for month analysis is affected by the late ended winter in year 2007. The late ended winter also might be the reason of higger cancellation magnitude, depature delay rate and average in Feuburay of 2007 than 2000.

Location

- The flights in the airports in south and west regionn have relative higher cancellation rate and depature delay rate while northeast has relative lower value of depature delay rate. This might be caused by the climate difference across different regions.
- The most interesting feature is that Alaska have high cancellation rate but low number of flights. This might due to the location reason and the extreme weather conditions.
- Generally, states with higher cancellation rate tends to have higher depature delay rate too.

Carrier

- Generally the cancellation rate decreased in year 2007 but some carriers stays high cancellation rate in year 2007 and 2000.
- The distrubution of cancellation rate changed for carriers since the number of carriers increased in year 2007.
- It's worth noticing that United Air Line Inc has extreme large value of departure delay both in 2000 and 2007.
- For generally information of the heatmap, we can see that the difference of flight numbers between carriers in year 2007 decreased than the difference in year 2000.

Plane

- Planes manufactured by Douglas, Mcdonnell Douglas and Mcdonnell Douglas Aircraft have relative higher cancellation rate in 2007 and 2000 that might due to the fact that thier planes were pretty old in thoese two years.
- In the fitted binomial model of age and cancellation in 2000 we can conclude that in the model, when plane gets 1 year older, odds of being cancelled at least once in 2000 is multiploed by 0.033. But age is not significant in the binomial value in year 2007.
- The percentage of planes which are more than 10 years old in 2007 is slight higher than year 2000.

Regression Model

- In the linear regression model we created for depature delay in year 2000 and 2007, features of month, day of month, day of week, scheduled arrival time, unique carrier, origin, destination, distance and diverted flight indicator are all significant in both models. The R square of these two models are all very low which indicates the low fit of depature delay in these models.

Contributions

All members in the group make their best contributions to this project. The detailed responsibilities are as below:

Chenzi Zhang: Part 1 - Time and Part 3 - Carrier

Shuoqi Zhang: Part 1 - Time and Part 5 - Model

Ziqin Xiong: Data processing and Part 4 - Plane

Shuhui Guo: Part 2 - Location and Part 5 - Model

Appendix

Data preparation and Hadoop code

All hadoop code (including Pig and Hive), and results generated by Hadoop are stored in `extra code and results.zip`. Please unzip the files and put them in the same folder where airline data and data in `airlinesauxiliaryfiles.zip` (`plane.csv`, `carriers.csv` and `plane-data.csv`) are placed.
Airline data can be obtained by `airlines_data.sh` introduced in lecture by following commands:

```
#downlad airline data files
./airlines_data.sh 2000 2000
mv airlines.csv 2000.csv
./airlines_data.sh 2007 2007
mv airlines.csv 2007.csv
#concatenate two files to one
cp 2000.csv 0007.csv
tail -n+2 2007.csv >> 0007.csv
```

And here the final airline csv file we use is `0007.csv`

Pig code used in Carrier section

Pig is used once in Carrier section to calculate cancellation rate of each carrier. The executable data file is `carrier_cancelrate.pig`. We first preprocess the data so it can be read easily in pig and put the files into hdfs

```
#remove header of csv file
tail -n+2 0007.csv >> 0007_pig.csv
tail -n+2 carriers.csv >> carriers_pig.csv
#remove double quotes in carrier.csv file
sed -i 's/"//g' carriers_pig.csv
#move files into hdfs
hadoop fs -mkdir final
hadoop fs -put 0007_pig.csv final
hadoop fs -put carriers_pig.csv final
```

Then we run pig code use:

```
pig -f carrier_cancelrate.pig
```

The results are generated in hdfs. we can move them in local filesystem:

```
#copy results generated from pig in HDFS to local filesystem
hadoop fs -get 2000pig/part* pig2000result.txt
hadoop fs -get 2007pig/part* pig2007result.txt
```

Results are then stored in `pig2000result.txt` and `pig2007result.txt`.

Hive code used in Plane section

Hive is used once in Plane section to validate the error in data. The executable data file is `plane_dataerror.hive`. To run the file, please run: (-S option is for silent mode, to get cleaner results.))

```
hive -S-f plane_dataerror.hive
```

And results in printed in terminal in tab-delimited format:

```
2000    N365AA  AA  A109E   10/21/2003  2001
2000    N395AA  AA  A-1B    06/05/2008  2007
2000    N544AA  AA  VANS AIRCRAFT RV6   01/26/2007  2007
2000    N394AA  AA  A-1B    None    2007
```

References

The codes for this report are based on

Chapter 5 in [Chapman & Hall_CRC The R Series] Deborah Nolan, Duncan Temple Lang - Data Science in R_ A Case Studies Approach to Computational Reasoning and Problem Solving (2015, Chapman and Hall_CRC)

Chapter 16, 17 in Tom White - Hadoop_ The Definitive Guide, 4th Edition_ Storage and Analysis at Internet Scale (2015, O'Reilly Media)

Course materials by Darren Glosemeyer for Stat 480: Data Science Foundations at the University of Illinois at Urbana-Champaign, including heatmaptreemapdensityplot.r, Chapter5.R, Chapter16CodeSegments.sh, Chapter17CodeSegments.sh

The comments for this report are based on

Douglas Aircraft Company. 2019. Wikipedia. https://en.wikipedia.org/wiki/Douglas_Aircraft_Company.

Jamie Gonzales. 2016. *Try Again Tomorrow*. Anchorage Daily News. https://www.adn.com/special-sections/61degnorth/2016/06/27/try-again-tomorrow/.

La Niña. 2019. Wikipedia. https://en.wikipedia.org/wiki/La_Ni%C3%B1a#North_America.

Rebecca Lindsey. 2008. *Global Temperature Anomalies: 2007*. NASA Earth Observatory. https://earthobservatory.nasa.gov/images/8423/global-temperature-anomalies-2007.