Aviation Data Analysis

Leveraging the power of Business Intelligence Techniques

Nishant Gupta & Kartik Nagras 1/12/18 Maths for Data Analytics

Table of Contents:

Introduction:

"Errors using inadequate data are much less than those using no data at all."

-*Charles Babbage*

In 2013, Warren Buffett called the commercial aviation industry a "death trap for investors." Fast forward to 2016 and the legendary value investor, Warren Buffet spent more than US$1.3 billion buying the stock of four major U.S. commercial carriers: American Airlines, Delta, United Continental, and Southwest Airlines — and he has recently upped his stake to more than $8 billion (Zhang, 2017).

He may or may not be right to invest, but it is undeniable that airlines in the U.S. and in most other regions are enjoying a run of good results, buoyed by steadily rising demand and an extended drop in fuel costs. Industry-wide passenger traffic grew by 6.3 percent in 2016.

In our view, an essential driver of continued growth will be the use of business intelligence techniques to help aviation executives develop sharper, more nuanced competitive positioning. Being a frequent flyer and an early career data analyst myself, I was intrigued to work with aviation data and look for any possible patterns that could help me avoid being stranded at airports while I wait for my delayed flight. US commercial airline data is available to public and would form the basis for this report.

The U.S. commercial airline industry is one of the most diverse, dynamic and perplexing in the world. As more and more people travel by flights, on-time performance becomes a key factor that travellers are concerned about.
The development in the field of data analytics in combination with statistics has made it possible to investigate big airline datasets and gain useful insights out of it.
Each year in the United States, every 2 flights are delayed out of 10 and 1 flight out of every 10 is either cancelled or diverted. As the number of passengers and flights increases, it will cause proportionate increase in flight delay and flight cancellation.
Therefore, it is prudent for aviation executives to minimize the inconvenience caused to customers due to flight delay & cancellations. On the customer end, we wanted to develop a tool to help flyers in making an educated decision to avoid flight related delays. In this report we have combined different data manipulation and statistical techniques to derive meaningful inferences and conclude on the following business questions:

1.What is the best time to travel?
2.Which carrier is reliable to travel with?
3.Which airport should we choose while travelling?

These findings will assist the flyer is selecting flights appropriately.
We have created an online portal using 'Shiny R' for this project and the code is available under Aviation repository at https://github.com/ngupta10. With the help of this online portal travellers will be able to make smart decisions while choosing their flight carrier and respective airport hubs.

What the Data says?

The data was obtained from Bureau of Transportation Statistics US for the year 2008. The data gives the information regarding all flights flying in the US for the year 2008, representing 20 unique commercial airlines and 7,00,9728 (Approx. ~7 Million) observations with 20 variables.

Data Summary:

No. of Unique Carriers: 20

No. of Unique Flights: 7131

No. of Airports: 286

Data Source: http://stat-computing.org/dataexpo/2009/the-data.html

Below are the variables considered for this analysis (Full list of all variables can be found at http://stat-computing.org/dataexpo/2009/the-data.html):

Variables
Year
Month
Day of Month
DayOfWeek1 (Monday) - 7 (Sunday)
Unique Carrier unique carrier code
Arr Delay arrival delay, in minutes
Dep Delay departure delay, in minutes
Origin origin IATA airport code
Dest destination IATA airport code
Distance in miles
Cancelled was the flight cancelled?
Cancellation Code reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
Carrier Delay in minutes

Random Sampling of Data:

The dataset consists of Approx. ~ 7 Million Observations. To enhance the processing speed a random sampling was performed on the data set based on following conditions:

- Confidence Level : 95%
- Confidence Interval : 0.5
- Population : 1048575
- Sample Population : 38205

Descriptive Statistics:

1. Arrival Delay

This variable details the delay in minutes for a flight arriving to an airport.

Descriptive Statistics:

nbr.val	nbr.null	nbr.na	min	max
37323.0	1058.0	884.0	-74.0	1225.0
range	sum	median	mean	SE.mean
1299.0	302049.0	-2.0	8.1	0.2
CI.mean.0.95	var	std.dev	coef.var	
0.4	1521.3	39.0	4.8	

```
> kurtosis(bm$ArrDelay,na.rm=TRUE)
[1] 74
```

```
> skewness(bm$ArrDelay,na.rm= TRUE)
[1] 5.6
```

Outliers in Arrival Delay:

- In Arrival Delay Data, there are negative values which are due to early arrival of planes. These observations are redundant to our cause as they distort the mean Arrival Delay. These outliers give a false representation on our analysis of statistical descriptive as we want to investigate only the flights which arrived after scheduled time.
- Secondly, Outliers were removed by keeping the lower bound as 0 and the upper bound calculated using the quartile method.

Descriptive Statistics (Without Outlier):

```
> stat.desc(ArrDelay_NoOutlier)
     nbr.val     nbr.null       nbr.na          min          max
    14482.00         0.00         0.00         1.00        85.00
       range          sum       median         mean      SE.mean
       84.00    289212.00        13.00        19.97         0.16
CI.mean.0.95          var      std.dev     coef.var
        0.32       382.56        19.56         0.98
> kurtosis(ArrDelay_NoOutlier,na.rm=TRUE)
[1] 4.2
> skewness(ArrDelay_NoOutlier,na.rm= TRUE)
[1] 1.4
```

2. Departure Delay

This variable detail the delay in minutes for a flight in departing from an airport.

Descriptive Statistics:

```
> stat.desc(bm$DepDelay)
     nbr.val     nbr.null       nbr.na          min          max
    37429.00      2873.00       778.00       -35.00      1233.00
       range          sum       median         mean      SE.mean
     1268.00    369929.00        -1.00         9.88         0.18
CI.mean.0.95          var      std.dev     coef.var
        0.36      1257.77        35.47         3.59
>
> kurtosis(bm$DepDelay,na.rm=TRUE)
[1] 100
> skewness(bm$DepDelay,na.rm= TRUE)
[1] 6.6
```

Outliers in Departure Delay:

- In Departure Delay Data, there are negative values which are due to early departure of planes. These observations are redundant to our cause as they distort the mean Departure Delay. These outliers give a false representation on our analysis of statistical descriptive as we want to investigate only the flights which departed after scheduled time.
- Secondly, Outliers were removed by keeping the lower bound as 0 and the upper bound calculated using the quartile method.

Descriptive Statistics (Without Outlier):

```
> stat.desc(DepDelay_NoOutlier)
     nbr.val     nbr.null       nbr.na          min          max
    13250.00         0.00         0.00         1.00        84.00
       range          sum       median         mean      SE.mean
       83.00    252742.00        12.00        19.07         0.17
CI.mean.0.95          var      std.dev     coef.var
        0.34       396.45        19.91         1.04
> kurtosis(DepDelay_NoOutlier,na.rm=TRUE)
[1] 4.2
> skewness(DepDelay_NoOutlier,na.rm= TRUE)
[1] 1.4
```

Results:

1. **Best time to Travel.**
   To know Which is the best time to travel? we began our analysis by checking the trends of cancellations of flights by month and week. For this we created tree maps, that present the number of cancellations by month and by week.

- Month-Wise Cancellation of flights.

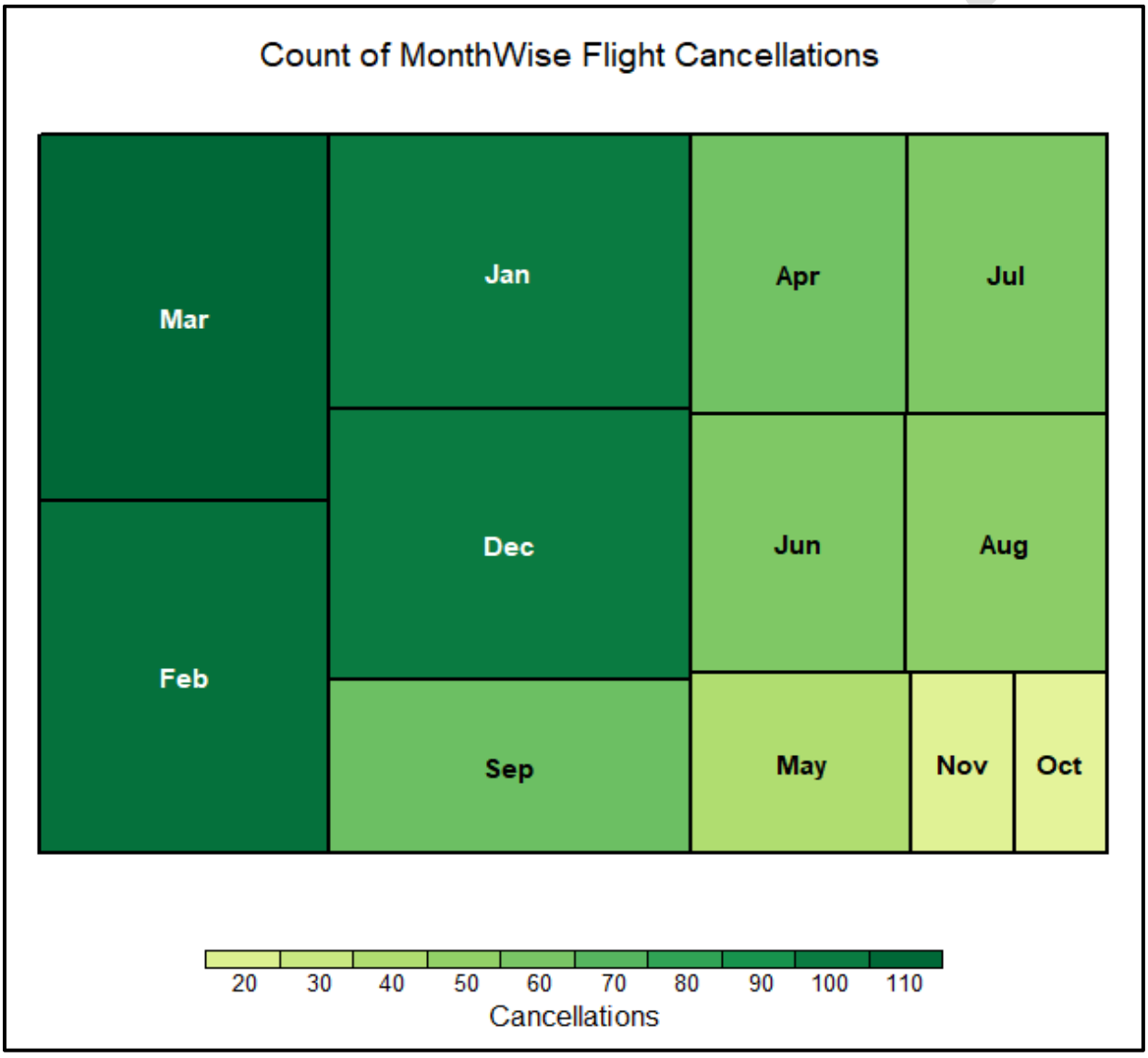

Figure 1: The treemap shows the number of flight cancellations for each month. As seen above March has maximum number of cancellations.

- Day of Week – Wise Cancellation of flights.

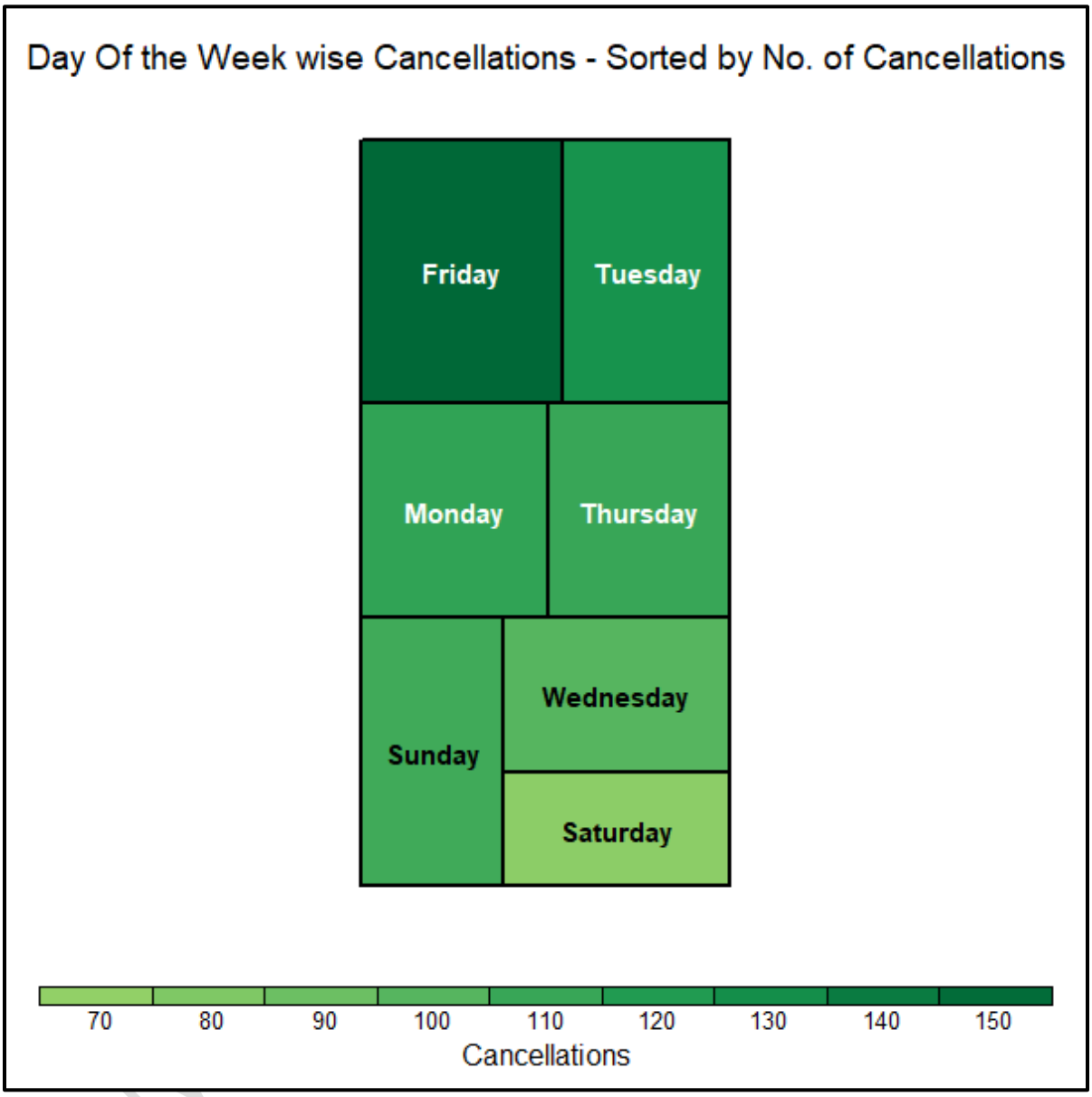

Figure 2: The treemap shows the number of flight cancellations for each day in a week. As seen above Friday has maximum number of cancellations.

From the above tree maps we can say that most number of cancellations happen in the month of March followed by the month of February and January.

From the above figures we can assert that least number of cancellations happen in the month of October and November. Further Saturday faces the least number of cancellations out of all the days in a week.

2. Reliable carrier.

Next piece of the puzzle is to select a carrier based on its performance. The first criterion was to analyze the arrival delay for each carrier.

- Boxplot of Unique Carriers with Arrival Delay

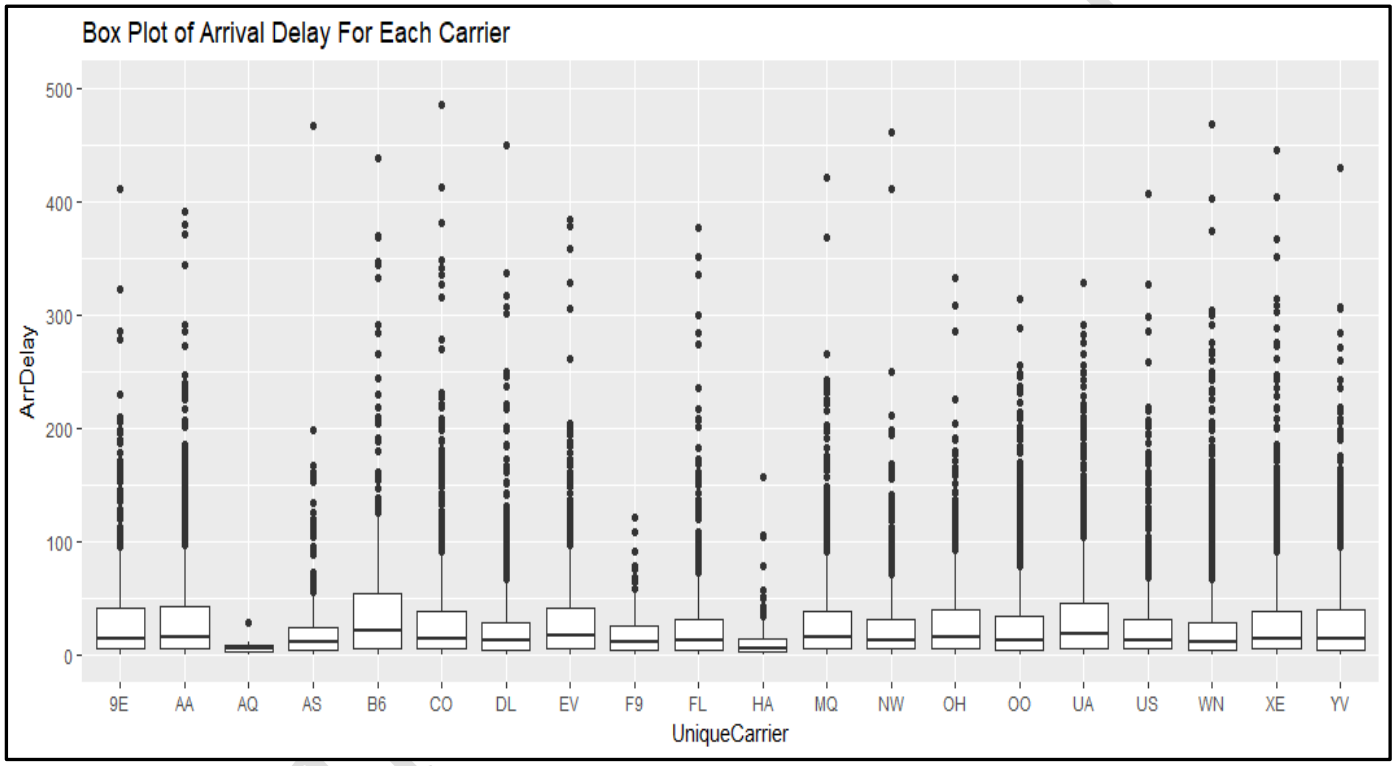

Figure 3: The above boxplot shows the spread of Arrival delay in minutes for each airplane carrier present in the dataset.

The above figure showcases the spread & range of arrival delay for each carrier. Several Carriers including AS & WN have the least variability in the arrival delay. While carriers like FL, EV, B6 have very variable arrival delay.

To further the findings, we plotted the average delay vales for each carrier.

- Average delay (in minutes) for each Unique carrier.

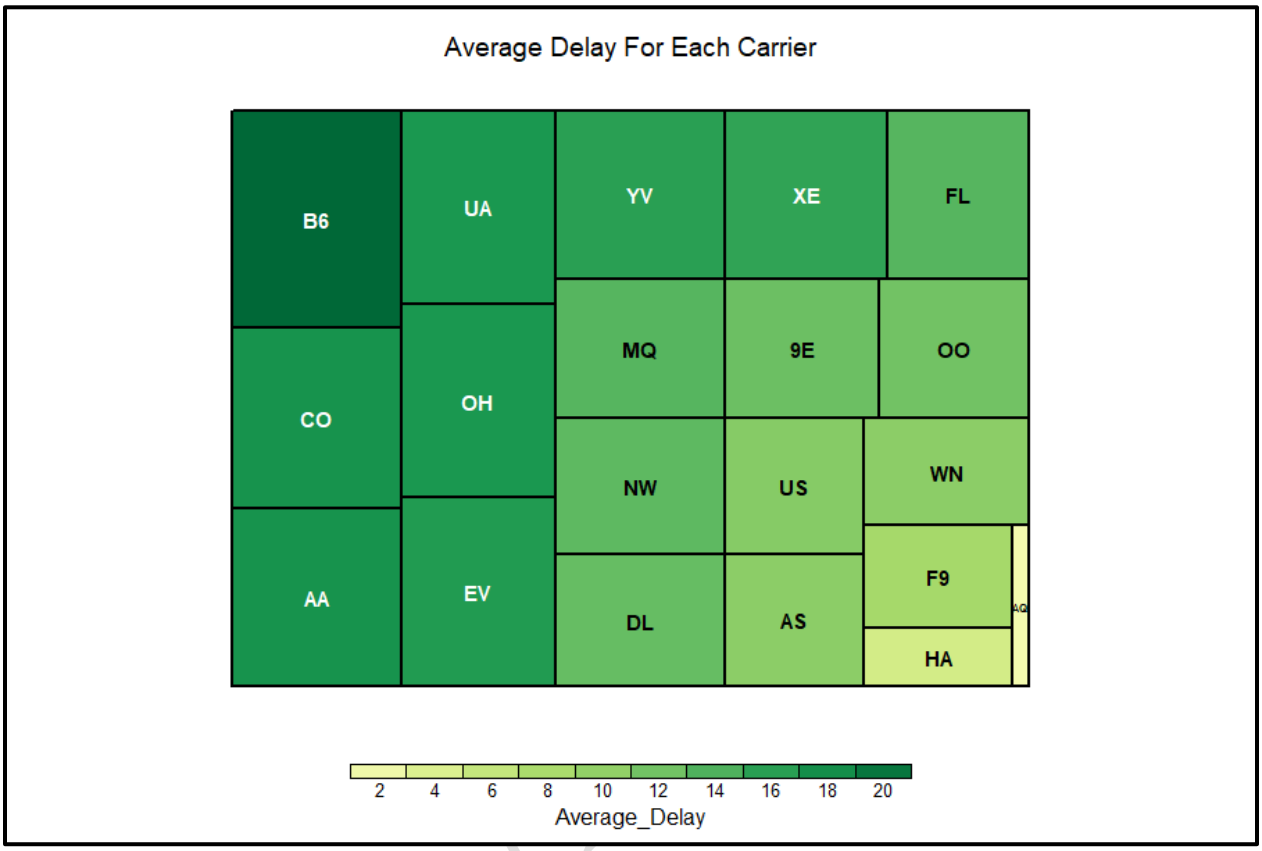

Figure 4: The average delay (in minutes) for each carrier.

The above tree map shows us the average airplane delay time by carrier. JetBlue Airline(B6) has the highest airplane delay time followed by Continental Airlines(CO), American Airlines(AA).

Aloha Airlines(AQ) has the least average Delay time followed by Hawaiian Airlines(HA), Frontier Airlines(F9) and Southwest Airlines (WN).

Next, we wanted to examine the relationship between a carrier and its respective arrival delay measures. For the same, we conducted a t-test to answer, "Is there a significant difference between the arrival delay for different carriers?". We selected two random carriers AA and AS to test our hypothesis.

```
        Welch Two Sample t-test

data:  AA.delay and AS.delay
t = 5.187, df = 1506, p-value = 0.0000002428
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.686439 10.386492
sample estimates:
mean of x mean of y
13.296908  5.760442
```

From the above t-test we can see that, the p value is < 0.05 which says that mean arrival delay for AA and AS did not happen by chance but instead that the arrival delay is significantly dependent on the carrier. This test details that the selection criterion for a reliable carrier will have average delay value as a factor.

Finally, to conclude upon the reliability of a carrier we looked at percentage of flights cancelled for each carrier.

- Cancellations Percentage for each Carrier.

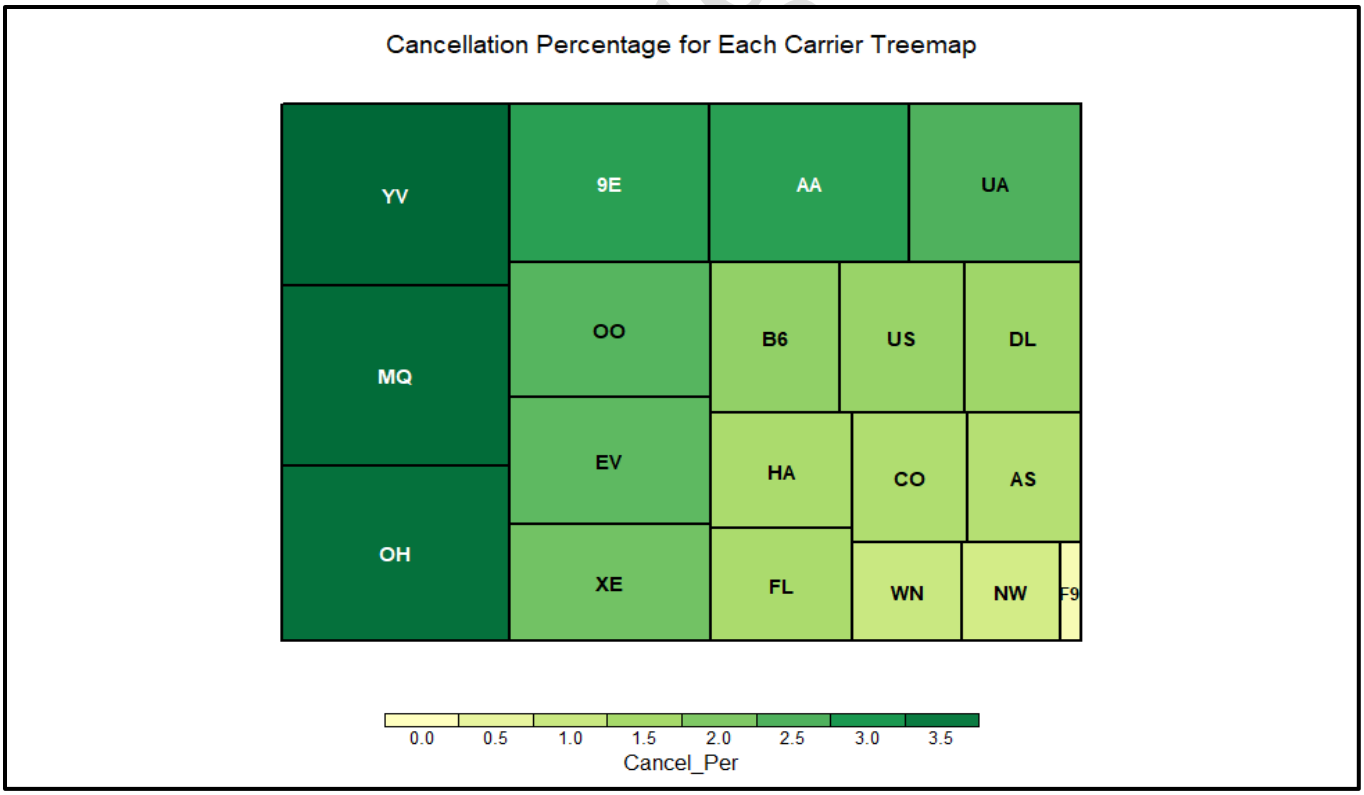

Figure 5: The above treemap shows the flight cancellation percentage for each carrier.

The above tree map shows cancellations by airplane carriers, where deeper color and bigger block size indicate higher cancellation rate. The cancel percentage is based on number of cancelled flights out of every 100 flights for each carrier.

The Highest number of flights were cancelled by Mesa Airlines(YV) and Envoy Air(MQ) followed by Comair Inc(OH) i.e. out of every 100 flights 3-4 flights are getting cancelled.

The least number of flights were cancelled by Frontier Flights (F9) followed by Northwest Airlines(NW) and Alaska Airlines(AS) i.e. out of every 100 flights only 1 flight is getting cancelled.

3. Reliable Airport

Whether the choice of airports influence the performance of a carrier was the next logical step in our analysis. First, we analyzed the percentage of flights getting delayed at each airport.

- Percentage of delayed flights at each airport

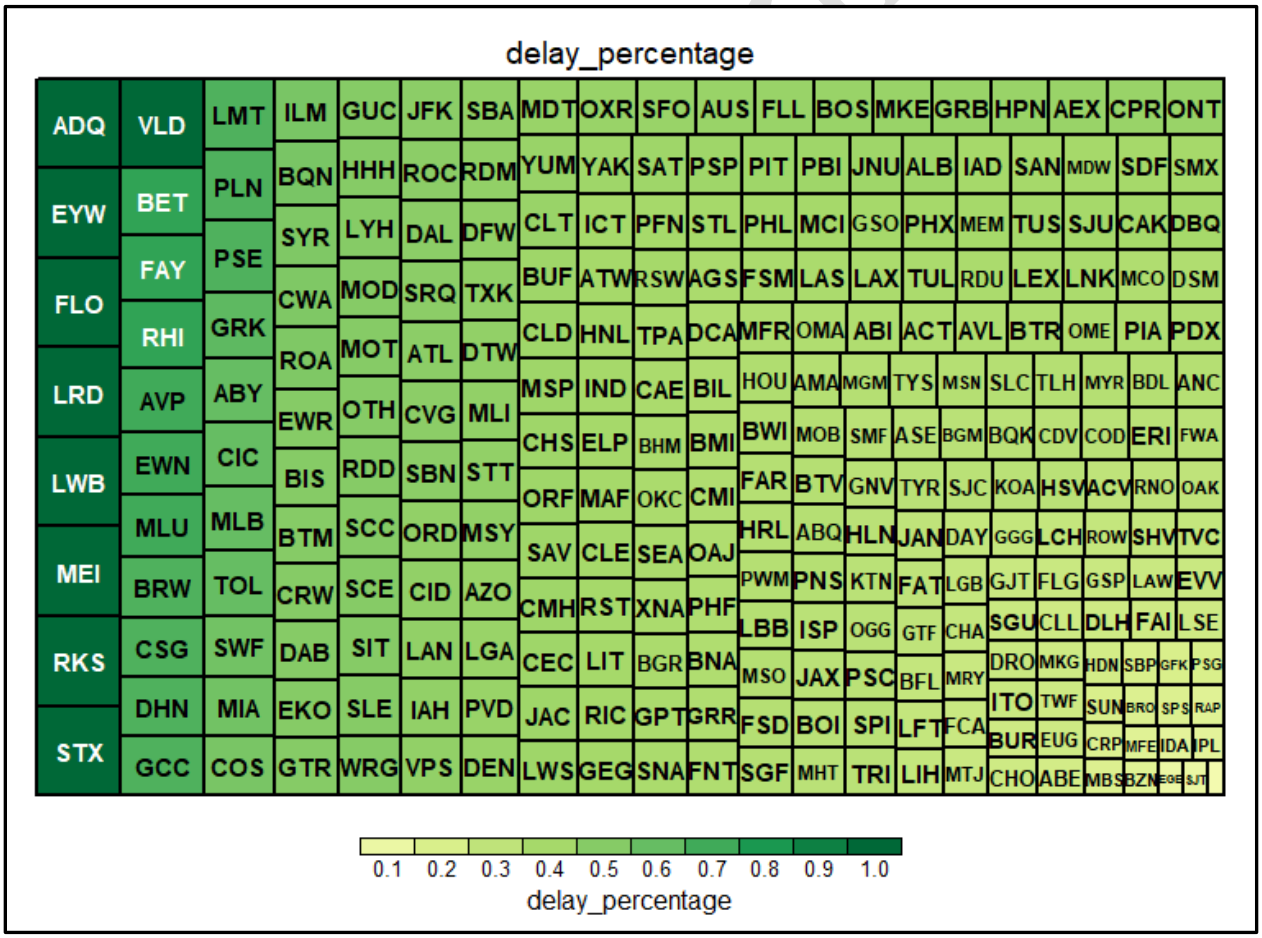

Figure 6: The treemap depicts the percentage of flights that are delayed at each airport

The above tree map showcases the percentage of delayed flight at each Airports. We can see that the many airports such as., Kodiak Airport in Alaska(ADQ), Key West International Airport in Florida(EYW) and Bethel Regional airport in Alaska(BET) etc. have high chance of flight delay.

Similarly, San Angelo Regional Airport (SJT), Idaho Falls Regional Airport (IDA) and Imperial County Airport(IPL) have the least delay percentage.

- Average delay time at Airports.

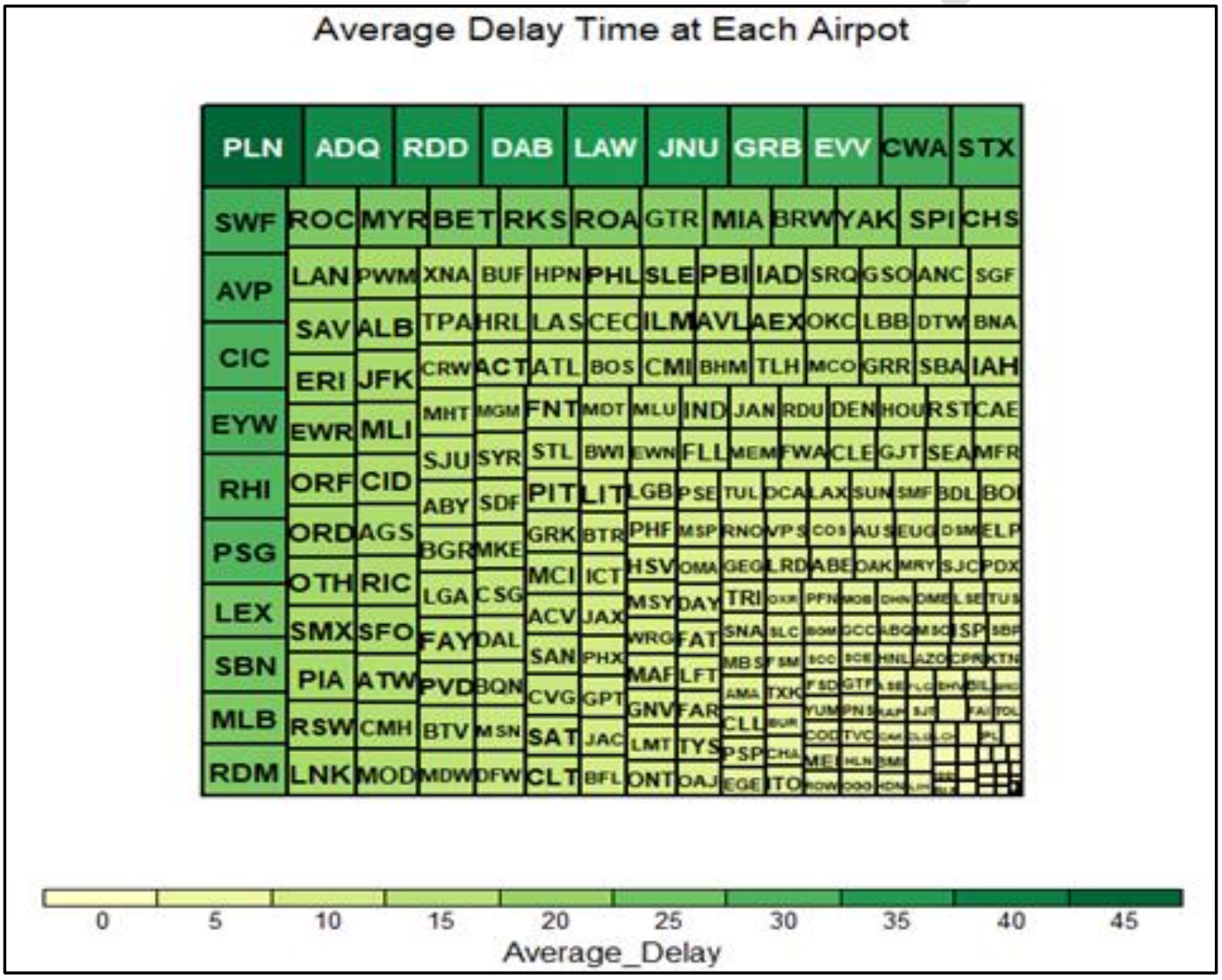

Figure 7: The treemap shows the average delay (in minutes) for each airport

As for the average delay time (minutes) for each airport, we can tell from the figure that Pellston Regional Airport (PLN) has highest average delay time of 45 mins followed by Kodiak Airport (ADQ), Redding Municipal Airport(RDD), with the average delay time of almost 40 minutes.

Lake Charles Airport (LCH), St. George Airport(SGU), Imperial County Airport(IPL) will be the lowest delay time airports with no more than 5 minutes' delay.

To analyse the relationship between delays in arrival time, given the interaction of Origin and Destination airport, we performed an analysis of variance (ANOVA). For this study, null hypothesis that is being tested states that the origin and the destination locations of a given flight route do not have a significant effect on the delay in arrival time that is observed in arrival time for each of the origin and destination airports were solely the result of randomization in this experiment.

```
Analysis of Variance Table

Response: ArrDelay
             Df   Sum Sq Mean Sq F value    Pr(>F)
Origin      286  1005942  3517.3  2.3525 < 2.2e-16 ***
Dest        290   833102  2872.8  1.9214 < 2.2e-16 ***
Residuals 36746 54940188  1495.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A p-value < 0.05 is returned, indicating that the interaction of these two factors does have a significant effect on the response variable. Therefore, based on this result, we would reject the null hypothesis, leading us to believe that the variation that is observed in the mean values of the delays in arrival time can be explained by the variation existent in the interaction of the different origin and destination airport locations being considered in this analysis and, as such, is likely not caused solely by randomization.

4. Do departure time, distance and departure delay predict arrival delay?

We created a model in which a traveller must enter his departure time, amount of distance to be travelled and the departure delay. Based on this information given by the traveller, we would be able to predict arrival delay for the traveller.

As a working example we placed fictitious values such that if the traveller is departing at 1400 and travelling 2000 miles with 15 mins of departure delay for his flight is 15 mins an arrival delay of 12 minutes is predicted.

```
# Testing the function
what.are.my.chances(dep = 1400, dist = 2000, dep_delay= 15)
```

```
> what.are.my.chances(dep = 1400, dist = 2000, dep_delay= 15)
       1
11.87706
```

Discussion:

The theme of this report is to help travellers make smart choices while planning their trips and for aviation executives to see through various factors that may be influential in forming future strategies.

From the above analysis we can say that the October & November are the best months for travelling followed by Wednesday or Sunday as the best day due to least probability of the flight cancellation. For an Airline to be reliable, factors such as cancellation percentage and average delay time play a very important role. Considering these 2 factors we can say that Frontier Airlines(F9), Alaskan Airways(AS) and Southwest Airline(WN) are the most reliable carriers which the travellers should consider while travelling. As for the airports, many a times traveller has an option of changing either destination or origin when travelling to or from a city with multiple airports. Lake Charles Airport (LCH), St. George Airport(SGU), Imperial County Airport(IPL) had around 5 minutes' delay time. While, airports like are San Angelo Regional Airport (SJT), Idaho Falls Regional Airport (IDA) had the least delay percentage. Therefore, these airports might be a better choice for travel.

Conclusion:

To conclude on the analysis, we have verified our findings from different sources ((Pinola, 2014), (Zhang, 2017), (M, 2017), (Medina, 2016)) and can say with a greater likelihood that the same techniques can be applied elsewhere on a flight dataset to reproduce effective insights. As for the aviation executive, we can further the analysis to look at the carrier manufacturer & carrier age as dimensions to identify any relation with delay times & cancellations.

A web portal is being developed for travellers with user-friendly interface to help travellers in choosing best possible carrier, time to travel & airports.

References:

- Zhang, B. (2017, February 15). Warren Buffett is investing $8 billion in an industry he once called a 'death trap'. Retrieved January 12, 2018, from http://www.businessinsider.com/warren-buffett-invests-in-airlines-american-delta-southwest-2017-2.
- Pinola, M. (2014, December 05). The Best (and Worst) Airlines in the US. Retrieved January 12, 2018, from https://lifehacker.com/the-best-and-worst-airlines-in-the-us-1667028259.
- Zhang, B. (2017, April 12). Here are the 12 best airlines in America. Retrieved January 12, 2018, from http://www.businessinsider.com/best-airlines-america-2017-4/#12-frontier-airlines-1.
- M. (2017, June 28). 10 Worst Airports in the U.S. for Flight Delays. Retrieved January 12, 2018, from https://www.cheatsheet.com/culture/worst-airports-us-delays.html/?a=viewall
- Medina, J. (2016, November 1). Best Days, Times to Fly to Avoid Flight Delays and Cancellations. Retrieved January 12, 2018, from http://www.flightbucks.com/blog/best-days-times-to-fly-to-avoid-flight-delays-cancellations