Final Report

Rachel Luo, Karen Kong, Carmen Huang, Xue Ling Ni

December 2, 2018

##Introduction This project focuses on the domestic flight statuses of various major American airlines in 2015. This is an important issue for both recurring customers as well as the airline companies themselves. The results of this dataset can point customers towards the right decision as they contemplate which airline to patron. Equally important, however, is for the airlines to recognize, understand, and solve any organizational or operational issues that is identified by this dataset. Airlines that experience low rates of delayed and/or cancelled flights can use that to their advantage to consolidate customer loyalty and reach out to other, less satisfied customers, while airlines that find themselves consistently failing to fly out on time will have the information they need to improve their operations and properly compete with their rival companies. The goal of this project, however, is to simply observe and take note of important patterns that appear in the dataset and its subsequent analysis to conclude which airline would be the most likely to deliver on its flight service through running summary statistics, data visualizations, statistical significance tests, and regression analyses.

##Data Description

There are three datasets: one that lists the names of airlines, one that lists the name and details of airports, and finally one dataset that includes all the flights in 2015. Without cleaning the dataset, we have 31 variables in our flights dataset, such as airline, air time, origin airport, month, etc.

We first looked at the variables that were clearly unnecessary from the onset. The columns “tail number” and “flight number” were identifiers of each flight and seems irrelevant to the punctuality of flights. Thus, we will not be considering these variables for flight delays.

flights$TAIL\_NUMBER = NULL  
flights$FLIGHT\_NUMBER = NULL

We added a date variable by using the year, month, and day. We changed the variables of “month,” “day of week,” and “airline” into factors because they are discrete data and hold no true numeric value.

dates = as.Date(with(flights, paste(flights$YEAR, flights$MONTH, flights$DAY,sep="-")),   
 "%Y-%m-%d")  
flights <- flights %>%  
 mutate(dates = dates)  
flights$YEAR = NULL  
flights$DAY = NULL  
  
flights$MONTH = as.factor(flights$MONTH)  
flights$DAY\_OF\_WEEK = as.factor(flights$DAY\_OF\_WEEK)  
flights$AIRLINE = as.factor(flights$AIRLINE)

We also changed “cancelled” and “diverted” into logical since we saw that the observations for these two columns returned 0 and 1, which is an indication of a boolean variable.

flights$CANCELLED = as.logical(flights$CANCELLED)  
flights$DIVERTED = as.logical(flights$DIVERTED)

Upon checking basic summary statistics, we realized that there were NA values in 11 columns: departure time, departure delay, taxi out, wheels off, scheduled time, elapsed time, air time, wheels on, taxi in, arrival time, and arrival delay.

We did tests on all of these variables to better understand why their observations were NA and we found that all these variables were NA because they were either cancelled or diverted.

|  | NA count | CANCELLED | DIVERTED |
| --- | --- | --- | --- |
| DEPARTURE\_TIME | 86153 | 86153 | 0 |
| DEPARTURE\_DELAY | 86153 | 86153 | 0 |
| TAXI\_OUT | 89047 | 89047 | 0 |
| WHEELS\_OFF | 89047 | 89047 | 0 |
| SCHEDULED\_TIME | 6 | 5 | 1 |
| ELAPSED\_TIME | 105071 | 89884 | 15187 |
| AIR\_TIME | 105071 | 89884 | 15187 |
| WHEELS\_ON | 92513 | 89884 | 2629 |
| TAXI\_IN | 92513 | 89884 | 2629 |
| ARRIVAL\_TIME | 92513 | 89884 | 2629 |
| ARRIVAL\_DELAY | 105071 | 89884 | 15187 |

Finally, to combat “cancelled” and “diverted” flights, we separated the cancelled and diverted flights into a different dataset. There are a total of 105,071 observations that we took out of the full dataset, leaving 5,714,008 observations in our cleaned dataset. Because we are only taking out 1.8% of our dataset, we do not believe that this will cause any biases. Furthermore, we believe that treating the cancelled/diverted flights as a completely separate category from the delayed flights will provide more accurate analysis on our current focus on delayed flights. We believe that the cancelled/diverted flights are a form of “missing” data at random. Their reasons for being cancelled/diverted are not necessarily arbitrary, as most are probably due to common issues like severe weather and mechanical issues as well as other variables that we will later examine in our regression. Thus, we do not believe we are creating biases by separating out these flights.

cancelledORdiverted = (flights[flights$CANCELLED == TRUE | flights$DIVERTED == TRUE,   
 c("dates", "MONTH", "DAY\_OF\_WEEK", "AIRLINE",   
 "ORIGIN\_AIRPORT", "DESTINATION\_AIRPORT",   
 "SCHEDULED\_DEPARTURE", "DEPARTURE\_TIME",   
 "DEPARTURE\_DELAY","TAXI\_OUT", "WHEELS\_OFF",   
 "SCHEDULED\_TIME", "ELAPSED\_TIME", "AIR\_TIME",   
 "DISTANCE", "WHEELS\_ON",  
 "TAXI\_IN", "SCHEDULED\_ARRIVAL", "ARRIVAL\_TIME",   
 "ARRIVAL\_DELAY", "DIVERTED", "CANCELLED",   
 "CANCELLATION\_REASON",   
 "AIR\_SYSTEM\_DELAY", "SECURITY\_DELAY", "AIRLINE\_DELAY",  
 "LATE\_AIRCRAFT\_DELAY", "WEATHER\_DELAY")])  
  
flightsClean = (flights[flights$CANCELLED == FALSE & flights$DIVERTED ==FALSE,   
 c("dates", "MONTH", "DAY\_OF\_WEEK", "AIRLINE",   
 "ORIGIN\_AIRPORT", "DESTINATION\_AIRPORT", "SCHEDULED\_DEPARTURE",  
 "DEPARTURE\_TIME", "DEPARTURE\_DELAY","TAXI\_OUT",   
 "WHEELS\_OFF", "SCHEDULED\_TIME", "ELAPSED\_TIME",   
 "AIR\_TIME", "DISTANCE", "WHEELS\_ON",  
 "TAXI\_IN", "SCHEDULED\_ARRIVAL", "ARRIVAL\_TIME", "ARRIVAL\_DELAY",   
 "DIVERTED", "CANCELLED", "CANCELLATION\_REASON",   
 "AIR\_SYSTEM\_DELAY", "SECURITY\_DELAY", "AIRLINE\_DELAY",  
 "LATE\_AIRCRAFT\_DELAY", "WEATHER\_DELAY")])

All of these measures were taken to ensure we would have a condensed and simplified dataset that was easier to understand, with only the most relevant numbers and data displayed.

##Summary Statistics and Visualizations

summary(flightsClean)

## dates MONTH DAY\_OF\_WEEK AIRLINE   
## Min. :2015-01-01 7 : 514384 1:841794 WN :1242403   
## 1st Qu.:2015-04-05 8 : 503956 2:827399 DL : 870275   
## Median :2015-07-03 6 : 492847 3:843242 AA : 712935   
## Mean :2015-07-02 3 : 492138 4:857886 OO : 576814   
## 3rd Qu.:2015-09-30 5 : 489641 5:851387 EV : 554752   
## Max. :2015-12-31 10 : 482878 6:689745 UA : 507762   
## (Other):2738164 7:802555 (Other):1249067   
## ORIGIN\_AIRPORT DESTINATION\_AIRPORT SCHEDULED\_DEPARTURE DEPARTURE\_TIME  
## Length:5714008 Length:5714008 Min. : 1 Min. : 1   
## Class :character Class :character 1st Qu.: 916 1st Qu.: 921   
## Mode :character Mode :character Median :1325 Median :1330   
## Mean :1329 Mean :1335   
## 3rd Qu.:1730 3rd Qu.:1740   
## Max. :2359 Max. :2400   
##   
## DEPARTURE\_DELAY TAXI\_OUT WHEELS\_OFF SCHEDULED\_TIME   
## Min. : -82.000 Min. : 1.00 Min. : 1 Min. : 18.0   
## 1st Qu.: -5.000 1st Qu.: 11.00 1st Qu.: 935 1st Qu.: 85.0   
## Median : -2.000 Median : 14.00 Median :1343 Median :123.0   
## Mean : 9.295 Mean : 16.07 Mean :1357 Mean :141.9   
## 3rd Qu.: 7.000 3rd Qu.: 19.00 3rd Qu.:1754 3rd Qu.:174.0   
## Max. :1988.000 Max. :225.00 Max. :2400 Max. :718.0   
##   
## ELAPSED\_TIME AIR\_TIME DISTANCE WHEELS\_ON   
## Min. : 14 Min. : 7.0 Min. : 31.0 Min. : 1   
## 1st Qu.: 82 1st Qu.: 60.0 1st Qu.: 373.0 1st Qu.:1054   
## Median :118 Median : 94.0 Median : 650.0 Median :1508   
## Mean :137 Mean :113.5 Mean : 824.5 Mean :1471   
## 3rd Qu.:168 3rd Qu.:144.0 3rd Qu.:1065.0 3rd Qu.:1911   
## Max. :766 Max. :690.0 Max. :4983.0 Max. :2400   
##   
## TAXI\_IN SCHEDULED\_ARRIVAL ARRIVAL\_TIME ARRIVAL\_DELAY   
## Min. : 1.000 Min. : 1 Min. : 1 Min. : -87.000   
## 1st Qu.: 4.000 1st Qu.:1110 1st Qu.:1058 1st Qu.: -13.000   
## Median : 6.000 Median :1520 Median :1512 Median : -5.000   
## Mean : 7.429 Mean :1493 Mean :1476 Mean : 4.407   
## 3rd Qu.: 9.000 3rd Qu.:1917 3rd Qu.:1916 3rd Qu.: 8.000   
## Max. :248.000 Max. :2400 Max. :2400 Max. :1971.000   
##   
## DIVERTED CANCELLED CANCELLATION\_REASON AIR\_SYSTEM\_DELAY   
## Mode :logical Mode :logical Length:5714008 Min. : 0   
## FALSE:5714008 FALSE:5714008 Class :character 1st Qu.: 0   
## Mode :character Median : 2   
## Mean : 13   
## 3rd Qu.: 18   
## Max. :1134   
## NA's :4650569   
## SECURITY\_DELAY AIRLINE\_DELAY LATE\_AIRCRAFT\_DELAY WEATHER\_DELAY   
## Min. : 0 Min. : 0 Min. : 0 Min. : 0   
## 1st Qu.: 0 1st Qu.: 0 1st Qu.: 0 1st Qu.: 0   
## Median : 0 Median : 2 Median : 3 Median : 0   
## Mean : 0 Mean : 19 Mean : 23 Mean : 3   
## 3rd Qu.: 0 3rd Qu.: 19 3rd Qu.: 29 3rd Qu.: 0   
## Max. :573 Max. :1971 Max. :1331 Max. :1211   
## NA's :4650569 NA's :4650569 NA's :4650569 NA's :4650569

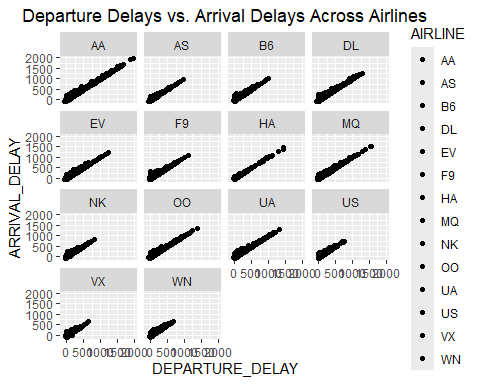
We selected the variables that we thought were the most relevant to answering our initial question as to which airline would be best to fly on in regards to flight cancellations and delays. We kept the month and day of the week variable to see if there would be a significant relationship between the time frame and the success of the flights; was there a certain day of the week, or a certain month, where the most amount of flights occurred? We simply looked at the number of flights that occurred on every day of the week and every month, and we ultimately found that the summer months had the highest number of flights, between June, July, and August. Furthermore, flights travelled the most frequently on Thursdays, while Saturday appeared to be the least popular day.

Next, we looked at the number of flights by airlines and found that Southwest delivered the most, followed by Delta and American Airlines. The departure delay variable was one of the most crucial parts of our dataset, as it provided information on what flights were delayed and by how long. After converting the results into minutes, we found that the flight with the smallest delay time had actually departed almost an hour and a half early from its scheduled time (min. = -82.000). On the other hand, the flight with the latest delay was well over 33 hours (max. = 1988.000). The average delay time for a flight is about 9 minutes (mean = 9.295).

Just as relevant to our initial question as the departure delay was the arrival delay variable. Flights ranged from arriving as early as 87 minutes from its estimated arrival time to as late as 33 hours (max. = 1971.000), with an average of four and a half minutes later than estimated (mean = 4.407).

To delve deeper into examining flight delays, we wanted to look at the taxiing and wheels variables that showed when the flight left the origin gate and runway, and then when it arrived at the destination runway and gate. We included scheduled time, elapsed time, and air time as well to see if that would provide any insight on the causes of the delays. The time it took for a flight to leave the gate and fly off from the runway ranged from a single minute to almost 4 hours (max. = 225.00). Likewise, at the destination airport, taxiing time ranged from a minute to 4 full hours (max = 248.000). The average taxi out and taxi in time were 16 minutes and 7 and a half minutes, respectively. These results point towards problems on the runway as a possible reason for some delays.

ggplot(flightsClean, aes(x=DEPARTURE\_DELAY, y=ARRIVAL\_DELAY, fill=AIRLINE)) +   
 geom\_point() + facet\_wrap(~AIRLINE) +   
 ggtitle("Departure Delays vs. Arrival Delays Across Airlines") +   
 theme(plot.title = element\_text(hjust = 0.5))

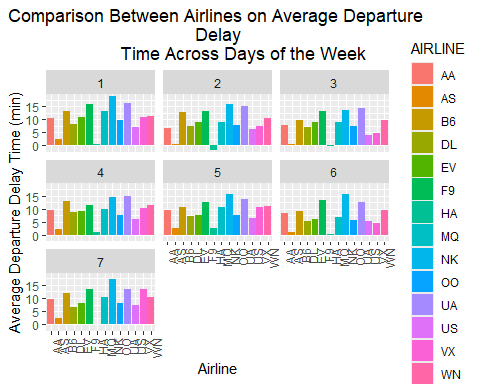


To synthesize our results, we made a scatter plot comparing departure delay time with arrival delay time, divided into subplots by airline. This visualization demonstrated that there was a positive, though not perfectly, linear relationship between departure and arrival delays across all of the airlines. Since there is a clear relationship between departure and arrival delays, we can infer that these two variables will act the same way when compared against other variables. The range of the results also pointed us closer towards answering our initial question, as the graphs depict Southwest Airlines and Virgin America with having the smallest levels of departure and arrival delays, whereas American Airlines has the widest-ranging delays.

#DayofWeek table   
DayOfWeekTable <-  
 flightsClean %>%  
 group\_by(AIRLINE, DAY\_OF\_WEEK) %>%  
 summarize(meanDepartDelays = mean(DEPARTURE\_DELAY))

## `summarise()` has grouped output by 'AIRLINE'. You can override using the  
## `.groups` argument.

#Mean Departure delay (Y) vs. Airline (X) facet\_wrap Day of Week  
ggplot(DayOfWeekTable, aes(x=AIRLINE,y=meanDepartDelays, fill=AIRLINE), color = "black") +   
 geom\_col() + facet\_wrap(~DAY\_OF\_WEEK) +   
 ggtitle("Comparison Between Airlines on Average Departure \n Delay   
 Time Across Days of the Week") +   
 xlab("Airline") + ylab("Average Departure Delay Time (min)") +   
 theme(plot.title = element\_text(hjust = 0.5),   
 axis.text.x = element\_text(angle=90, hjust=1))

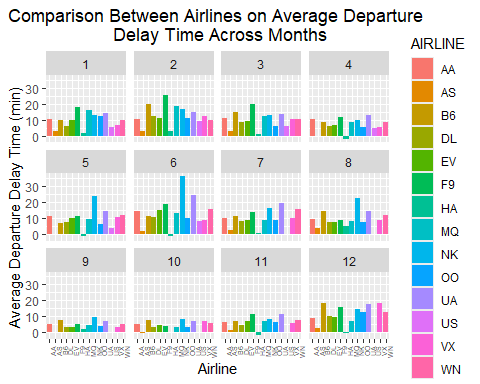


Next, we looked at the average departure delay by airline for each day of the week to see if there was a specific day that should be avoided for travelling, or a specific day that might be more favorable towards flights. This visualization showed a consistently low level of delays, if not early departures, for Hawaiian Airlines, ranging from two minutes early to three minutes late. Meanwhile Spirit Airlines has the highest average delay time, with Mondays and Sundays generally having the highest averages. Their delay times ranged from 14 to 18 minutes.

#Month Table   
MonthTable <-  
 flightsClean %>%  
 group\_by(AIRLINE, MONTH) %>%  
 summarize(meanDepartDelays = mean(DEPARTURE\_DELAY))

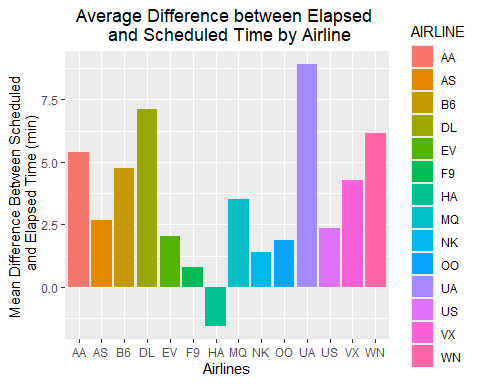
## `summarise()` has grouped output by 'AIRLINE'. You can override using the  
## `.groups` argument.

#Mean Departure delay (Y) vs. Airline (X) facet\_wrap Month  
ggplot(MonthTable, aes(x=AIRLINE,y=meanDepartDelays, fill=AIRLINE), color = "black") +   
 geom\_col() + facet\_wrap(~MONTH) +   
 ggtitle("Comparison Between Airlines on Average Departure \n Delay Time Across Months") +   
 xlab("Airline") + ylab("Average Departure Delay Time (min)") +   
 theme(plot.title = element\_text(hjust = 0.5),   
 axis.text.x = element\_text(angle=90, hjust=1, size=5))



We also compared average departure delay by airline across months to understand which months would have the highest delays. From this visualization, we can easily tell that September and October have the lowest departure delay since all of these delay times are under 10 minutes, while June and February have the highest delay times. December also seems to have high departure delay, which is also consistent across airlines. This could possibly be due to weather conditions and outside factors. Within these graphs, Hawaiian Airlines consistently has the lowest departure delay times, sometimes even ranging into a negative departure delay, which means that it departed early rather than late. Alaska Airlines is usually the next lowest. The departure delay is always under 5 minutes, but never negative. In contrast, Spirit Airlines often has a large departure delay, especially during the summer months.

#New column for difference between scheduled and elapsed  
flightsClean$difference = flightsClean$SCHEDULED\_TIME-flightsClean$ELAPSED\_TIME  
  
#Table for airlines vs average difference between scheduled and elapsed  
 meandifftable =   
 flightsClean %>%  
 group\_by(AIRLINE) %>%  
 summarize(meandifference = mean(difference))  
   
#Graph for meandifftable  
 ggplot(meandifftable, aes(x = AIRLINE, y = meandifference, fill = AIRLINE),  
 colour = "black") +  
 geom\_col()+  
 xlab("Airlines")+ylab("Mean Difference Between Scheduled \n and Elapsed Time (min)")+  
 ggtitle('Average Difference between Elapsed \n and Scheduled Time by Airline')+  
 theme(plot.title = element\_text(hjust = 0.5))

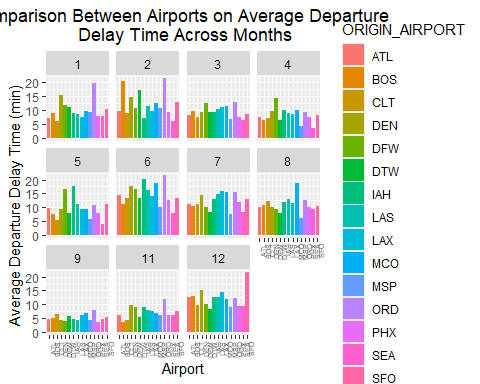


We were also curious about the difference between the scheduled and elapsed time for each airline. The elapsed time is the total time spanning from the moment the plane actually departed to the moment it arrived at its destination, including any delays and the time for taxiing ins/outs. When we subtracted the elapsed time from the scheduled time, this would mean that positive values indicate that the flights arrived earlier than expected while negatives values mean that the flights arrived later than expected. From the graph shown, it is clear that United Airlines has the largest difference between scheduled and elapsed time with approximately 9 minutes, followed by Delta Airlines at about 7 minutes. On the flip side, Frontier Airlines has the least with about 1 minute and Spirit Airlines next with 1.5 minutes. Interestingly enough, Hawaiian Airlines has a negative value of -1.5 minutes, which means that they tend to underestimate their flight times, while United and Delta Airlines tend to significantly overestimate their flight times. All of this measures how accurately airlines predict their flight times will be, knowing that there will be delays and other conditions involved.

#Removing October due to missing data in Airports  
flightsClean10 = flightsClean[!(flightsClean$MONTH=="10"),]  
   
#Month Table 2  
MonthTable2 <-  
 flightsClean %>%  
 filter(ORIGIN\_AIRPORT=="ATL" |ORIGIN\_AIRPORT=="ORD" | ORIGIN\_AIRPORT=="DFW" |  
 ORIGIN\_AIRPORT=="DEN" | ORIGIN\_AIRPORT=="LAX" | ORIGIN\_AIRPORT=="PHX" |  
 ORIGIN\_AIRPORT=="SFO" |ORIGIN\_AIRPORT=="IAH" |ORIGIN\_AIRPORT=="LAS" |  
 ORIGIN\_AIRPORT== "MSP" | ORIGIN\_AIRPORT=="SEA" | ORIGIN\_AIRPORT=="MCO" |  
 ORIGIN\_AIRPORT=="DTW" | ORIGIN\_AIRPORT=="BOS" | ORIGIN\_AIRPORT=="CLT")%>%  
 group\_by(ORIGIN\_AIRPORT, MONTH) %>%  
 summarize(meanDepartDelays = mean(DEPARTURE\_DELAY))

## `summarise()` has grouped output by 'ORIGIN\_AIRPORT'. You can override using  
## the `.groups` argument.

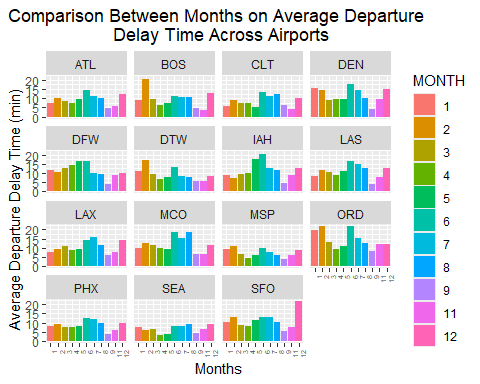
#Mean Departure delay (Y) vs. Airport (X) facet\_wrap Month  
ggplot(MonthTable2, aes(x=ORIGIN\_AIRPORT,y=meanDepartDelays, fill=ORIGIN\_AIRPORT), color = "black") +   
 geom\_col() + facet\_wrap(~MONTH) +   
 ggtitle("Comparison Between Airports on Average Departure \n Delay Time Across Months") +   
 xlab("Airport") + ylab("Average Departure Delay Time (min)") +   
 theme(plot.title = element\_text(hjust = 0.5),   
 axis.text.x = element\_text(angle=90, hjust=1, size=5))



We first calculated the top 15 airports by determining the origin airport with the most amount of flights. Of these 15 airports, 33% of them were from the Western US, 20% from Southwestern US, 20% from Southeast, 13.3% from the Midwest and 13.3% from the Northeast. In addition, we noticed that airport codes in October were missing, which led us to take out October before making this analysis. We are essentially taking out 8% of our data, which adds bias because October is the second month with the lowest departure delay time; however, most of the ones that we removed have a low number of flights (ranging from 4 to 32,509 flights).

We compared the average departure delay time across months between the top 15 airports. In general, we can see that the average departure delay time for all 15 airports are relatively high during the summer months such as June and winter months such as February. This makes sense because during the summer time, many people tend to be on vacation and travel more frequently than other months, thus increasing flight traffic. High departure delay during winter months could possibly resulted from severe weather conditions. On the other hand, September has the lowest average departure delay for all 15 airports. Chicago O’Hare International Airport has the highest average departure delay time of 20 minutes or more during January, February, and June. San Francisco International Airport has the highest average departure delay time during December in comparison to the other 14 airports.

#Mean Departure delay (Y) vs. Month (X) facet\_wrap Airport  
ggplot(MonthTable2, aes(x=MONTH,y=meanDepartDelays, fill=MONTH), color = "black") +   
 geom\_col() + facet\_wrap(~ORIGIN\_AIRPORT) +   
 ggtitle("Comparison Between Months on Average Departure \n Delay Time Across Airports") +   
 xlab("Months") + ylab("Average Departure Delay Time (min)") +   
 theme(plot.title = element\_text(hjust = 0.5),   
 axis.text.x = element\_text(angle=90, hjust=1, size=5))

 In this graph, instead of looking at average departure delay of airports across months, we are looking at the average departure delay of months across different airports. Specifically, the top 15 airports with the most number of flights in descending order are chosen to be the focus of this graph. If we look at Boston Logan, the top two months with the most mean departure delays are February (about 17 min) and December (about 12 min). There are other airports such as DEN (Denver) and LAX (Los Angeles) with similar delays across months and no particular months with drastically different delay time. Other airports such as SFO (San Francisco) have one month that is abnormally high compared to other ones. This graph narrows down to the airports rather than months in the previous visualization.

##Preliminary Statistical Analyses

Before we ran our baseline regression model, we wanted to run some statistical tests and correlations to see which variables to focus on.

We began by running an anova test to understand whether the differences between scheduled and elapsed time were significant between each airline. In the test we did below, we compared all of the airlines in pairs to evaluate their delays in relation to each other. We believe that the bigger the number, the earlier the flight had arrived.

Our null hypothesis is that the differences between each airline are the same and our alternate hypothesis is that they’re not. Our results indicated that all of these airlines had statistical differences in the differences between their predicted and actual flight time, with p-values adjusted for multiple comparisons all under 0.05. It is clear that all of these airlines calculate their estimations differently, based on what they know of their tendencies to be delayed.

We noticed that there was a significant difference of 10.45 between United Airlines and Hawaiian Airlines; this means that United Airlines on average has an additional difference of almost 11 minutes in their delays in comparison with Hawaiian Airlines. This is the greatest difference between any airline. This suggests that, when choosing between a flight with Hawaiian Airlines and a flight with United Airlines, a customer should patron the former.

It was interesting to look at the differences by each airline. In comparison with other companies, American Airlines had significantly higher differences in predicted and actual time for all other airlines with the exception of Delta, United, and Southwest. Likewise, Delta Airlines also had significantly higher differences in comparison to the other airlines, indicating that they will perform better than expected in terms of punctual flights. Meanwhile, Hawaiian Airlines, Frontier Airlines, Skywest, and Spirit Airlines all have significantly lower differences than the airlines they were compared with.

Large differences between scheduled time and elapsed time do not necessarily indicate higher quality flights; instead, we believe that much of this could be attributed to airlines purposely inflating the duration of their flights to minimize their level of delays and consequently customer dissatisfaction. Previous analysis has shown that Hawaiian Airlines tended to underestimate their flight times, while United and Delta significantly overestimated their flight times. However, in previous visualizations, Hawaiian had the lowest average delay across all the months and days of the week, whereas Spirit Airlines had the highest average of departure delays. Therefore, this anova test is more of an indication on how accurately airlines predicted their flights.

AirlineDiff\_aov = aov(difference ~ AIRLINE, data = flightsClean)  
TukeyHSD(AirlineDiff\_aov, ordered = FALSE)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = difference ~ AIRLINE, data = flightsClean)  
##   
## $AIRLINE  
## diff lwr upr p adj  
## AS-AA -2.6792447 -2.7934227 -2.56506666 0e+00  
## B6-AA -0.6101276 -0.7070962 -0.51315901 0e+00  
## DL-AA 1.7518122 1.6840074 1.81961696 0e+00  
## EV-AA -3.3445151 -3.4205085 -3.26852172 0e+00  
## F9-AA -4.5760660 -4.7261539 -4.42597818 0e+00  
## HA-AA -6.9279093 -7.0898393 -6.76597931 0e+00  
## MQ-AA -1.8654206 -1.9602353 -1.77060588 0e+00  
## NK-AA -3.9634330 -4.0982221 -3.82864383 0e+00  
## OO-AA -3.4843031 -3.5594748 -3.40913151 0e+00  
## UA-AA 3.5267277 3.4487818 3.60467356 0e+00  
## US-AA -2.9999432 -3.1085884 -2.89129802 0e+00  
## VX-AA -1.1189542 -1.2976833 -0.94022520 0e+00  
## WN-AA 0.7674855 0.7044190 0.83055206 0e+00  
## B6-AS 2.0691171 1.9372644 2.20096977 0e+00  
## DL-AS 4.4310568 4.3188975 4.54321618 0e+00  
## EV-AS -0.6652704 -0.7825616 -0.54797925 0e+00  
## F9-AS -1.8968214 -2.0714884 -1.72215428 0e+00  
## HA-AS -4.2486646 -4.4336066 -4.06372261 0e+00  
## MQ-AS 0.8138241 0.6835472 0.94410092 0e+00  
## NK-AS -1.2841883 -1.4458990 -1.12247764 0e+00  
## OO-AS -0.8050585 -0.9218189 -0.68829802 0e+00  
## UA-AS 6.2059724 6.0874068 6.32453789 0e+00  
## US-AS -0.3206985 -0.4613612 -0.18003581 0e+00  
## VX-AS 1.5602904 1.3604748 1.76010608 0e+00  
## WN-AS 3.4467302 3.3373702 3.55609022 0e+00  
## DL-B6 2.3619398 2.2673565 2.45652309 0e+00  
## EV-B6 -2.7343875 -2.8350032 -2.63377186 0e+00  
## F9-B6 -3.9659384 -4.1298740 -3.80200285 0e+00  
## HA-B6 -6.3177817 -6.4926240 -6.14293939 0e+00  
## MQ-B6 -1.2552930 -1.3707848 -1.13980125 0e+00  
## NK-B6 -3.3533054 -3.5033608 -3.20324997 0e+00  
## OO-B6 -2.8741756 -2.9741720 -2.77417910 0e+00  
## UA-B6 4.1368553 4.0347569 4.23895363 0e+00  
## US-B6 -2.3898156 -2.5169073 -2.26272395 0e+00  
## VX-B6 -0.5088267 -0.6993328 -0.31832052 0e+00  
## WN-B6 1.3776131 1.2863667 1.46885951 0e+00  
## EV-DL -5.0963273 -5.1692526 -5.02340202 0e+00  
## F9-DL -6.3278782 -6.4764361 -6.17932027 0e+00  
## HA-DL -8.6797215 -8.8402344 -8.51920849 0e+00  
## MQ-DL -3.6172328 -3.7096066 -3.52485892 0e+00  
## NK-DL -5.7152452 -5.8483287 -5.58216168 0e+00  
## OO-DL -5.2361153 -5.3081839 -5.16404679 0e+00  
## UA-DL 1.7749155 1.6999578 1.84987321 0e+00  
## US-DL -4.7517554 -4.8582771 -4.64523368 0e+00  
## VX-DL -2.8707664 -3.0482127 -2.69332018 0e+00  
## WN-DL -0.9843267 -1.0436603 -0.92499296 0e+00  
## F9-EV -1.2315509 -1.3840205 -1.07908134 0e+00  
## HA-EV -3.5833942 -3.7475341 -3.41925421 0e+00  
## MQ-EV 1.4790945 1.3805530 1.57763608 0e+00  
## NK-EV -0.6189179 -0.7563541 -0.48148162 0e+00  
## OO-EV -0.1397880 -0.2196090 -0.05996708 4e-07  
## UA-EV 6.8712428 6.7888039 6.95368167 0e+00  
## US-EV 0.3445719 0.2326595 0.45648430 0e+00  
## VX-EV 2.2255609 2.0448271 2.40629460 0e+00  
## WN-EV 4.1120006 4.0434587 4.18054257 0e+00  
## HA-F9 -2.3518433 -2.5608727 -2.14281386 0e+00  
## MQ-F9 2.7106454 2.5479746 2.87331627 0e+00  
## NK-F9 0.6126330 0.4238472 0.80141886 0e+00  
## OO-F9 1.0917629 0.9397012 1.24382454 0e+00  
## UA-F9 8.1027937 7.9493417 8.25624576 0e+00  
## US-F9 1.5761228 1.4050213 1.74722439 0e+00  
## VX-F9 3.4571118 3.2348146 3.67940898 0e+00  
## WN-F9 5.3435516 5.1970956 5.49000755 0e+00  
## MQ-HA 5.0624887 4.8888317 5.23614571 0e+00  
## NK-HA 2.9644763 2.7661456 3.16280694 0e+00  
## OO-HA 3.4436061 3.2798450 3.60736726 0e+00  
## UA-HA 10.4546370 10.2895840 10.61968996 0e+00  
## US-HA 3.9279661 3.7463877 4.10954446 0e+00  
## VX-HA 5.8089550 5.5784968 6.03941328 0e+00  
## WN-HA 7.6953948 7.5368252 7.85396441 0e+00  
## NK-MQ -2.0980124 -2.2466851 -1.94933974 0e+00  
## OO-MQ -1.6188825 -1.7167918 -1.52097332 0e+00  
## UA-MQ 5.3921483 5.2920933 5.49220327 0e+00  
## US-MQ -1.1345226 -1.2599786 -1.00906657 0e+00  
## VX-MQ 0.7464664 0.5570475 0.93588523 0e+00  
## WN-MQ 2.6329061 2.5439520 2.72186021 0e+00  
## OO-NK 0.4791299 0.3421463 0.61611344 0e+00  
## UA-NK 7.4901607 7.3516353 7.62868608 0e+00  
## US-NK 0.9634898 0.8056370 1.12134258 0e+00  
## VX-NK 2.8444787 2.6322105 3.05674697 0e+00  
## WN-NK 4.7309185 4.6001855 4.86165152 0e+00  
## UA-OO 7.0110308 6.9293488 7.09271280 0e+00  
## US-OO 0.4843599 0.3730040 0.59571595 0e+00  
## VX-OO 2.3653489 2.1849592 2.54573863 0e+00  
## WN-OO 4.2517887 4.1841590 4.31941835 0e+00  
## US-UA -6.5266709 -6.6399181 -6.41342361 0e+00  
## VX-UA -4.6456819 -4.8272453 -4.46411860 0e+00  
## WN-UA -2.7592422 -2.8299426 -2.68854167 0e+00  
## VX-US 1.8809889 1.6842824 2.07769547 0e+00  
## WN-US 3.7674287 3.6638586 3.87099885 0e+00  
## WN-VX 1.8864398 1.7107495 2.06213004 0e+00

Due to the discrepancy between Hawaiian Airlines and Spirit Airlines, we subsequently ran a correlation test to understand the relationship between a given flight’s departure delay and the difference between scheduled and elapsed time. This test indicates that as a flight’s departure delay increases by the minutes, the difference between the predicted flight time and the actual time goes down; there is a significantly negative, rather loose linear relationship (x = -0.01620321, p < 0.05).

While this is not a perfectly linear relationship, this shows that departure delays tighten the gap between the predicted and actual duration of a flight, indicating that airlines do overpredict their flight times. From these two tests, we conclude that Hawaiian Airlines does not overinflate their flight times simply because they do not face the same delay conditions as other airlines do, such as technical or weather issues. Meanwhile, Spirit Airlines do face the same conditions, and should therefore overestimate their flight durations to minimize their departure delays.

cor.test(flightsClean$difference, flightsClean$DEPARTURE\_DELAY)

##   
## Pearson's product-moment correlation  
##   
## data: flightsClean$difference and flightsClean$DEPARTURE\_DELAY  
## t = -38.737, df = 5714006, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.01702292 -0.01538348  
## sample estimates:  
## cor   
## -0.01620321

Next, we ran correlation tests to understand how continuous variables like delays interacted with each other. Previously, our graph showed that there was a clear relationship between departure delay and arrival delay. To better understand their relationship, we ran a correlation test between departure delay and arrival delay, which indicated a tight positive, almost perfectly linear relationship between the two variables where for every minute that a flight departs later than it was scheduled to, it will arrive nearly a minute later as well. The correlation value (r = 0.94) indicates a highly positive correlation given how close it is to 1.

cor.test(flightsClean$DEPARTURE\_DELAY, flightsClean$ARRIVAL\_DELAY)

##   
## Pearson's product-moment correlation  
##   
## data: flightsClean$DEPARTURE\_DELAY and flightsClean$ARRIVAL\_DELAY  
## t = 6884.2, df = 5714006, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9445832 0.9447597  
## sample estimates:  
## cor   
## 0.9446715

We also decided to see if there was correlation between distance and departure delay. Upon running the correlation test, we found that there was low correlation between the two variables. A correlation coefficient of 0.02 shows a nonlinear, not tight, but positive relationship between distance and departure delay. We can conclude that distance doesn’t have any significant impact on the amount of departure - they are independent of each other. This makes sense, as departure delay takes into account the time before the airplane actually leaves to embark on its journey.

cor.test(flightsClean$DISTANCE, flightsClean$DEPARTURE\_DELAY)

##   
## Pearson's product-moment correlation  
##   
## data: flightsClean$DISTANCE and flightsClean$DEPARTURE\_DELAY  
## t = 57.284, df = 5714006, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.02313772 0.02477664  
## sample estimates:  
## cor   
## 0.02395719

##Regression Analyses

Before building our main, baseline regression model, we wanted to run other regressions based on different data aggregation levels to further develop our ideas of what variables should be included and to get an idea of by exactly how much they could cause a change in the departure delay time.

We began with a regression setting departure delay as the dependent variable and month as the independent variable.

lm\_MonthDelay = lm(DEPARTURE\_DELAY ~ factor(MONTH), data = flightsClean)  
summary(lm\_MonthDelay)

##   
## Call:  
## lm(formula = DEPARTURE\_DELAY ~ factor(MONTH), data = flightsClean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -93.69 -14.36 -10.36 -1.87 1978.31   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.68971 0.05443 178.034 < 2e-16 \*\*\*  
## factor(MONTH)2 2.08600 0.07927 26.317 < 2e-16 \*\*\*  
## factor(MONTH)3 -0.11233 0.07558 -1.486 0.1372   
## factor(MONTH)4 -2.06002 0.07607 -27.080 < 2e-16 \*\*\*  
## factor(MONTH)5 -0.32611 0.07568 -4.309 1.64e-05 \*\*\*  
## factor(MONTH)6 4.18412 0.07556 55.376 < 2e-16 \*\*\*  
## factor(MONTH)7 1.65757 0.07479 22.162 < 2e-16 \*\*\*  
## factor(MONTH)8 0.18203 0.07516 2.422 0.0154 \*   
## factor(MONTH)9 -4.89298 0.07676 -63.747 < 2e-16 \*\*\*  
## factor(MONTH)10 -4.74591 0.07593 -62.502 < 2e-16 \*\*\*  
## factor(MONTH)11 -2.80391 0.07675 -36.534 < 2e-16 \*\*\*  
## factor(MONTH)12 1.99802 0.07645 26.136 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 36.79 on 5713996 degrees of freedom  
## Multiple R-squared: 0.005206, Adjusted R-squared: 0.005204   
## F-statistic: 2719 on 11 and 5713996 DF, p-value: < 2.2e-16

Our results show that, on average, January’s departure delay is 9.69 minutes, with all other months being compared against this average. From this, we could tell that June has the highest average delay with 13.87 minutes, while September has the lowest average delay of 4.80 minutes. Generally, months in the winter and summer had higher departure delay averages, whereas spring and autumn months had smaller averages. This adds on to our previous findings in which June, July, and August had the highest number of flights out of the other twelve months of the year, which we can attribute to the time frame of being a heavy-vacation time of the year. Meanwhile, winter months would have more delay due to increased likelihood of bad weather, particularly snow storms.

lm\_WeekDelay = lm(DEPARTURE\_DELAY ~ factor(DAY\_OF\_WEEK), data = flightsClean)  
summary(lm\_WeekDelay)

##   
## Call:  
## lm(formula = DEPARTURE\_DELAY ~ factor(DAY\_OF\_WEEK), data = flightsClean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -91.33 -14.10 -10.76 -2.10 1978.63   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.75553 0.04020 267.58 <2e-16 \*\*\*  
## factor(DAY\_OF\_WEEK)2 -1.65068 0.05709 -28.91 <2e-16 \*\*\*  
## factor(DAY\_OF\_WEEK)3 -2.15787 0.05682 -37.98 <2e-16 \*\*\*  
## factor(DAY\_OF\_WEEK)4 -0.88009 0.05658 -15.55 <2e-16 \*\*\*  
## factor(DAY\_OF\_WEEK)5 -1.38158 0.05669 -24.37 <2e-16 \*\*\*  
## factor(DAY\_OF\_WEEK)6 -3.02070 0.05990 -50.43 <2e-16 \*\*\*  
## factor(DAY\_OF\_WEEK)7 -1.42823 0.05754 -24.82 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 36.88 on 5714001 degrees of freedom  
## Multiple R-squared: 0.0005414, Adjusted R-squared: 0.0005403   
## F-statistic: 515.9 on 6 and 5714001 DF, p-value: < 2.2e-16

While we previously saw how airlines compared to one another regarding departure delays, we next wanted to further understand exactly how departure delays were affected depending on the day of the week. Our visualization in the previous section gave us an idea that Mondays and Sundays had the highest averages. This test gives us a better understanding of how the delay times compare with one another. Monday’s average departure delay was 10.76 minutes, and all other coefficients are compared against Monday. Difference in departure delays are all significant between each day of the week, with Monday having the highest departure delay and Saturday having the lowest departure delay of 7.73 minutes.

lm\_airlineDelay = lm(DEPARTURE\_DELAY ~ AIRLINE, data = flightsClean)  
summary(lm\_airlineDelay)

##   
## Call:  
## lm(formula = DEPARTURE\_DELAY ~ AIRLINE, data = flightsClean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -83.72 -13.97 -10.52 -1.83 1979.17   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.82611 0.04356 202.611 < 2e-16 \*\*\*  
## AIRLINEAS -7.10718 0.09894 -71.834 < 2e-16 \*\*\*  
## AIRLINEB6 2.61636 0.08403 31.137 < 2e-16 \*\*\*  
## AIRLINEDL -1.51281 0.05876 -25.748 < 2e-16 \*\*\*  
## AIRLINEEV -0.21051 0.06585 -3.197 0.00139 \*\*   
## AIRLINEF9 4.47725 0.13006 34.425 < 2e-16 \*\*\*  
## AIRLINEHA -8.35619 0.14032 -59.552 < 2e-16 \*\*\*  
## AIRLINEMQ 1.14108 0.08216 13.888 < 2e-16 \*\*\*  
## AIRLINENK 7.05699 0.11680 60.420 < 2e-16 \*\*\*  
## AIRLINEOO -1.09002 0.06514 -16.734 < 2e-16 \*\*\*  
## AIRLINEUA 5.50695 0.06754 81.533 < 2e-16 \*\*\*  
## AIRLINEUS -2.74511 0.09414 -29.158 < 2e-16 \*\*\*  
## AIRLINEVX 0.16738 0.15488 1.081 0.27981   
## AIRLINEWN 1.69108 0.05465 30.944 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 36.78 on 5713994 degrees of freedom  
## Multiple R-squared: 0.005855, Adjusted R-squared: 0.005853   
## F-statistic: 2589 on 13 and 5713994 DF, p-value: < 2.2e-16

The intercept in this test shows the departure delay for American Airlines, and all the other coefficients is the difference in departure delay for each airline. From this regression, we can tell that Spirit Airlines (noted as NK) has the highest departure delay at 15.88 minutes, while Hawaiian Airlines (noted as HA) has the lowest departure delay of 0.47 minutes. This is consistent with the visualizations that we’ve shown earlier. In previous visualizations, we saw that Spirit Airlines always had the highest departure delays and Hawaiian had the lowest across both months and days of the week.

We then constructed our baseline regression model based on the results of the above, smaller regression, especially in light of their particularly small r-squared values. Thus, our baseline regression used departure delay as the dependent variable and airline, day of the week, month, and distance as the independent variables.

baseline <- lm(DEPARTURE\_DELAY ~ AIRLINE + DAY\_OF\_WEEK + MONTH + TAXI\_OUT, data=flightsClean)  
summary(baseline)

##   
## Call:  
## lm(formula = DEPARTURE\_DELAY ~ AIRLINE + DAY\_OF\_WEEK + MONTH +   
## TAXI\_OUT, data = flightsClean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -86.31 -14.29 -9.64 -1.59 1979.55   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.328949 0.084896 74.550 < 2e-16 \*\*\*  
## AIRLINEAS -6.705703 0.098649 -67.976 < 2e-16 \*\*\*  
## AIRLINEB6 2.270693 0.083710 27.126 < 2e-16 \*\*\*  
## AIRLINEDL -1.786681 0.058601 -30.489 < 2e-16 \*\*\*  
## AIRLINEEV -0.309279 0.065740 -4.705 2.54e-06 \*\*\*  
## AIRLINEF9 4.788874 0.129520 36.974 < 2e-16 \*\*\*  
## AIRLINEHA -6.885173 0.140231 -49.099 < 2e-16 \*\*\*  
## AIRLINEMQ 1.053253 0.081984 12.847 < 2e-16 \*\*\*  
## AIRLINENK 7.680977 0.116420 65.977 < 2e-16 \*\*\*  
## AIRLINEOO -1.506979 0.064962 -23.198 < 2e-16 \*\*\*  
## AIRLINEUA 5.286169 0.067332 78.510 < 2e-16 \*\*\*  
## AIRLINEUS -4.562548 0.095836 -47.608 < 2e-16 \*\*\*  
## AIRLINEVX 0.688595 0.154252 4.464 8.04e-06 \*\*\*  
## AIRLINEWN 2.918229 0.055537 52.546 < 2e-16 \*\*\*  
## DAY\_OF\_WEEK2 -1.665880 0.056701 -29.380 < 2e-16 \*\*\*  
## DAY\_OF\_WEEK3 -2.121413 0.056470 -37.567 < 2e-16 \*\*\*  
## DAY\_OF\_WEEK4 -0.937268 0.056242 -16.665 < 2e-16 \*\*\*  
## DAY\_OF\_WEEK5 -1.338743 0.056335 -23.764 < 2e-16 \*\*\*  
## DAY\_OF\_WEEK6 -2.742460 0.059515 -46.080 < 2e-16 \*\*\*  
## DAY\_OF\_WEEK7 -1.231737 0.057122 -21.563 < 2e-16 \*\*\*  
## MONTH2 1.951347 0.078874 24.740 < 2e-16 \*\*\*  
## MONTH3 -0.042584 0.075273 -0.566 0.5716   
## MONTH4 -1.912944 0.075733 -25.259 < 2e-16 \*\*\*  
## MONTH5 -0.170380 0.075318 -2.262 0.0237 \*   
## MONTH6 4.150380 0.075233 55.167 < 2e-16 \*\*\*  
## MONTH7 1.370894 0.074735 18.343 < 2e-16 \*\*\*  
## MONTH8 -0.059484 0.075125 -0.792 0.4285   
## MONTH9 -5.076781 0.076736 -66.159 < 2e-16 \*\*\*  
## MONTH10 -4.950604 0.075858 -65.262 < 2e-16 \*\*\*  
## MONTH11 -3.124040 0.076738 -40.711 < 2e-16 \*\*\*  
## MONTH12 1.661870 0.076423 21.746 < 2e-16 \*\*\*  
## TAXI\_OUT 0.266260 0.001796 148.290 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 36.6 on 5713976 degrees of freedom  
## Multiple R-squared: 0.0156, Adjusted R-squared: 0.01559   
## F-statistic: 2921 on 31 and 5713976 DF, p-value: < 2.2e-16

The results of the regression analysis show how much a certain factor could affect the amount of departure delay of a given fact, assuming that all of the other variables are held constant. With an intercept of 6.33, a flight from American Airlines on a Monday in January is represented as the default, whereas the other levels are compared to their average delay of about 6 minutes. The factors that will result in the earliest flights include a flight being from Hawaiian Airlines or Alaska Airlines, or taking place in either September or October. Conversely, a flight from Spirit Airlines results in much later delay (x = 7.680977), or a flight that takes place in June (x = 4.150380) compared to the default flight. While almost all of these factors are significant, this model displays no significant difference in flights that take place in March and August. We also see that for every minute that taxiing out takes, departure delay also increases by 0.266260 of a minute. However, this model predicts just a little more than 1% of the variance (r-squared=0.0156), and so this model is not a good fit for our data.

##Predictive Analyses

For our predictive analysis, we are predicting departure delays times for Boston Logan airport in the month of February, since we previously saw that February had the highest departure delay across months. In order to do so, we filtered our data and split it 90-10 to create training and validation data sets to run our regressions on. We will then find the best model that would help us predict average delays by choosing the one that has the lowest predictive error.

logan2delay = flightsClean%>%  
 filter(ORIGIN\_AIRPORT == "BOS")%>%  
 filter(MONTH == 2)  
  
#Set the seed  
set.seed(12345678)  
  
#Splitting the data  
randOrder = order(runif(nrow(logan2delay)))  
training.data = subset(logan2delay, randOrder < 0.9 \* nrow(logan2delay))  
validation.data = subset(logan2delay, randOrder >= .9 \* nrow(logan2delay))  
  
prediction.error = function(lm\_model, validation.data){  
predicted.logan2delay = predict(lm\_model, validation.data)  
error = sqrt(mean((predicted.logan2delay-validation.data$DEPARTURE\_DELAY)^2))}  
  
#Prediction Models  
  
#REGRESSION 1 : baseline  
lm\_1 = lm(DEPARTURE\_DELAY ~ AIRLINE + factor(DAY\_OF\_WEEK) + TAXI\_OUT, data=logan2delay)  
error\_model1 = prediction.error(lm\_1, validation.data)  
  
#REGRESSION 2 : location  
lm\_2 = lm(DEPARTURE\_DELAY ~ factor(AIRLINE) + factor(DESTINATION\_AIRPORT) + factor(DAY\_OF\_WEEK), data=logan2delay)  
error\_model2 = prediction.error(lm\_2, validation.data)  
  
# Regression 3 : Interacting destination airport and day of week  
lm\_3 = lm(DEPARTURE\_DELAY ~ factor(AIRLINE) + AIR\_TIME\* factor(DAY\_OF\_WEEK), data=logan2delay)  
error\_model3 = prediction.error(lm\_3, validation.data)  
  
#REGRESSION 4 : Interacting Day of Week and Taxi out   
lm\_4 = lm(DEPARTURE\_DELAY ~ factor(AIRLINE) + factor(DAY\_OF\_WEEK)\*TAXI\_OUT, data=logan2delay)  
error\_model4 = prediction.error(lm\_4, validation.data)  
  
#REGRESSION 5 : Interacting Airline and Taxi out   
lm\_5 = lm(DEPARTURE\_DELAY ~ factor(AIRLINE)\*TAXI\_OUT + factor(DAY\_OF\_WEEK), data=logan2delay)  
error\_model5 = prediction.error(lm\_5, validation.data)  
  
#REGRESSION 6 :   
lm\_6 = lm(DEPARTURE\_DELAY ~ factor(AIRLINE) + factor(DAY\_OF\_WEEK) + TAXI\_OUT + DISTANCE, data=logan2delay)  
error\_model6 = prediction.error(lm\_6, validation.data)  
  
#REGRESSION 7:   
lm\_7 = lm(DEPARTURE\_DELAY ~ factor(AIRLINE) + factor(DAY\_OF\_WEEK) + TAXI\_OUT + DISTANCE + SCHEDULED\_TIME + ELAPSED\_TIME, data = logan2delay)  
error\_model7 = prediction.error(lm\_7, validation.data)  
  
#REGRESSION 8 :   
lm\_8 = lm(DEPARTURE\_DELAY ~ factor(AIRLINE) + SCHEDULED\_TIME + ELAPSED\_TIME, data = logan2delay)  
error\_model8 = prediction.error(lm\_8, validation.data)  
  
#REGRESSION 9:   
lm\_9 = lm(DEPARTURE\_DELAY ~ factor(AIRLINE) + SCHEDULED\_TIME+ELAPSED\_TIME+AIR\_TIME + TAXI\_OUT, data = logan2delay)  
error\_model9 = prediction.error(lm\_9, validation.data)  
  
#REGRESSION 10 :   
lm\_10 = lm(DEPARTURE\_DELAY ~ factor(AIRLINE) \* SCHEDULED\_TIME + factor(DAY\_OF\_WEEK), data = logan2delay)  
error\_model10 = prediction.error(lm\_10, validation.data)  
  
#REGRESSION 11:   
lm\_11 = lm(DEPARTURE\_DELAY ~ factor(AIRLINE) + DISTANCE + AIR\_TIME, data = logan2delay)  
error\_model11 = prediction.error(lm\_11, validation.data)  
  
#REGRESSION 12 :   
lm\_12 = lm(DEPARTURE\_DELAY ~ factor(AIRLINE) + factor(DAY\_OF\_WEEK) + TAXI\_OUT + SCHEDULED\_TIME + ELAPSED\_TIME + AIR\_TIME +DISTANCE, data = logan2delay)  
error\_model12 = prediction.error(lm\_12, validation.data)  
  
#REGRESSION 13:   
lm\_13 = lm(DEPARTURE\_DELAY ~ factor(AIRLINE) + DISTANCE + SCHEDULED\_TIME + ELAPSED\_TIME, data = logan2delay)  
error\_model13 = prediction.error(lm\_13, validation.data)  
  
#REGRESSION 14:   
lm\_14 = lm(DEPARTURE\_DELAY ~ factor(AIRLINE) + DISTANCE + factor(DAY\_OF\_WEEK) + AIR\_TIME, data = logan2delay)  
error\_model14 = prediction.error(lm\_14, validation.data)  
  
#REGRESSION 15:   
lm\_15 = lm(DEPARTURE\_DELAY ~ factor(DAY\_OF\_WEEK) \* ELAPSED\_TIME + DISTANCE, data = logan2delay)  
error\_model15= prediction.error(lm\_15, validation.data)  
  
#Prediction Error table  
data.frame(Names = c("Baseline", "DestinationAirport", "Airtime X weekday", "Weekday X taxiout", "Airline X taxiout", "Airlineday taxi distance", "Weekday X TimeandDistance", "Predicted and Actual Time", "Times", "Airline x Scheduled time and weekday", "Time and Distance", "All Regressors", "Distance Scheduled and Elapsed", "Day of Week Distance and Time", "Dayweek x elapsed and distance"), Error = c(error\_model1,error\_model2, error\_model3,error\_model4,error\_model5,error\_model6,error\_model7,error\_model8,error\_model9,error\_model10,error\_model11,error\_model12,error\_model13,error\_model14,error\_model15))

## Names Error  
## 1 Baseline 40.70397  
## 2 DestinationAirport 40.06130  
## 3 Airtime X weekday 40.65445  
## 4 Weekday X taxiout 40.19534  
## 5 Airline X taxiout 40.56437  
## 6 Airlineday taxi distance 40.70571  
## 7 Weekday X TimeandDistance 40.62283  
## 8 Predicted and Actual Time 41.80079  
## 9 Times 41.61190  
## 10 Airline x Scheduled time and weekday 40.57821  
## 11 Time and Distance 41.64409  
## 12 All Regressors 40.46455  
## 13 Distance Scheduled and Elapsed 41.79799  
## 14 Day of Week Distance and Time 40.55965  
## 15 Dayweek x elapsed and distance 41.11019

The second model regressing departure delay against airline, destination, and weekday has the lowest prediction error at 40.06130, therefore is the best predictive model.

summary(lm\_2)

##   
## Call:  
## lm(formula = DEPARTURE\_DELAY ~ factor(AIRLINE) + factor(DESTINATION\_AIRPORT) +   
## factor(DAY\_OF\_WEEK), data = logan2delay)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -80.95 -21.51 -11.61 5.65 1065.62   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 23.2835 4.8613 4.790 1.70e-06 \*\*\*  
## factor(AIRLINE)AS 14.0292 8.2537 1.700 0.089223 .   
## factor(AIRLINE)B6 5.1146 3.1648 1.616 0.106118   
## factor(AIRLINE)DL 19.4102 3.8402 5.055 4.42e-07 \*\*\*  
## factor(AIRLINE)EV 7.3918 6.8273 1.083 0.278983   
## factor(AIRLINE)NK 18.2961 6.3993 2.859 0.004261 \*\*   
## factor(AIRLINE)OO -1.1618 9.7217 -0.120 0.904880   
## factor(AIRLINE)UA 10.2574 3.7107 2.764 0.005720 \*\*   
## factor(AIRLINE)US -4.8308 3.9229 -1.231 0.218193   
## factor(AIRLINE)VX -3.2169 5.8202 -0.553 0.580483   
## factor(AIRLINE)WN 8.8888 4.9006 1.814 0.069750 .   
## factor(DESTINATION\_AIRPORT)AUS 0.4073 9.9347 0.041 0.967299   
## factor(DESTINATION\_AIRPORT)BNA 25.8768 10.5319 2.457 0.014034 \*   
## factor(DESTINATION\_AIRPORT)BUF 4.3455 6.6314 0.655 0.512303   
## factor(DESTINATION\_AIRPORT)BWI 12.1830 4.4871 2.715 0.006641 \*\*   
## factor(DESTINATION\_AIRPORT)CAK 1.3057 10.9482 0.119 0.905073   
## factor(DESTINATION\_AIRPORT)CHS 7.1073 9.9350 0.715 0.474399   
## factor(DESTINATION\_AIRPORT)CLE 7.0494 8.9878 0.784 0.432871   
## factor(DESTINATION\_AIRPORT)CLT 17.8530 4.7023 3.797 0.000148 \*\*\*  
## factor(DESTINATION\_AIRPORT)CVG -7.0374 10.2656 -0.686 0.493031   
## factor(DESTINATION\_AIRPORT)DCA 10.6349 4.0493 2.626 0.008648 \*\*   
## factor(DESTINATION\_AIRPORT)DEN 11.3649 5.4250 2.095 0.036214 \*   
## factor(DESTINATION\_AIRPORT)DFW 6.1517 5.4518 1.128 0.259198   
## factor(DESTINATION\_AIRPORT)DTW 7.6900 4.2026 1.830 0.067320 .   
## factor(DESTINATION\_AIRPORT)EWR 9.5596 4.8493 1.971 0.048723 \*   
## factor(DESTINATION\_AIRPORT)FLL 16.3853 4.9300 3.324 0.000893 \*\*\*  
## factor(DESTINATION\_AIRPORT)HOU 19.4629 6.5765 2.959 0.003092 \*\*   
## factor(DESTINATION\_AIRPORT)IAD 9.0507 5.8885 1.537 0.124336   
## factor(DESTINATION\_AIRPORT)IAH 3.5563 6.4152 0.554 0.579356   
## factor(DESTINATION\_AIRPORT)JAX 9.0477 7.2694 1.245 0.213313   
## factor(DESTINATION\_AIRPORT)JFK 17.3776 4.0419 4.299 1.74e-05 \*\*\*  
## factor(DESTINATION\_AIRPORT)LAS 27.9332 7.1653 3.898 9.77e-05 \*\*\*  
## factor(DESTINATION\_AIRPORT)LAX 4.8048 4.7785 1.005 0.314692   
## factor(DESTINATION\_AIRPORT)LGA 22.4089 3.5944 6.234 4.79e-10 \*\*\*  
## factor(DESTINATION\_AIRPORT)LGB 17.2428 13.3456 1.292 0.196391   
## factor(DESTINATION\_AIRPORT)MCI -1.6564 8.2934 -0.200 0.841701   
## factor(DESTINATION\_AIRPORT)MCO 9.4929 4.4308 2.142 0.032189 \*   
## factor(DESTINATION\_AIRPORT)MDW 5.8356 7.1174 0.820 0.412297   
## factor(DESTINATION\_AIRPORT)MIA 6.1823 5.7653 1.072 0.283609   
## factor(DESTINATION\_AIRPORT)MKE 22.1115 8.9894 2.460 0.013928 \*   
## factor(DESTINATION\_AIRPORT)MSP 4.1033 5.0267 0.816 0.414352   
## factor(DESTINATION\_AIRPORT)MSY 46.5473 9.9350 4.685 2.85e-06 \*\*\*  
## factor(DESTINATION\_AIRPORT)MYR -14.0735 11.8862 -1.184 0.236446   
## factor(DESTINATION\_AIRPORT)ORD 11.5513 4.7252 2.445 0.014525 \*   
## factor(DESTINATION\_AIRPORT)PBI 9.3453 5.1026 1.831 0.067071 .   
## factor(DESTINATION\_AIRPORT)PDX -15.5457 14.0494 -1.106 0.268548   
## factor(DESTINATION\_AIRPORT)PHL 11.7254 4.3315 2.707 0.006805 \*\*   
## factor(DESTINATION\_AIRPORT)PHX 23.8483 5.9037 4.040 5.41e-05 \*\*\*  
## factor(DESTINATION\_AIRPORT)PIT 9.4683 6.8119 1.390 0.164587   
## factor(DESTINATION\_AIRPORT)RDU 10.2635 5.9549 1.724 0.084836 .   
## factor(DESTINATION\_AIRPORT)RIC 8.0939 5.7627 1.405 0.160199   
## factor(DESTINATION\_AIRPORT)RSW 16.3113 4.9416 3.301 0.000969 \*\*\*  
## factor(DESTINATION\_AIRPORT)SAN 10.1905 7.1836 1.419 0.156065   
## factor(DESTINATION\_AIRPORT)SAV 6.5473 9.9350 0.659 0.509910   
## factor(DESTINATION\_AIRPORT)SEA 8.6780 7.3245 1.185 0.236137   
## factor(DESTINATION\_AIRPORT)SFO 15.6835 5.0337 3.116 0.001842 \*\*   
## factor(DESTINATION\_AIRPORT)SJU 13.6385 6.1506 2.217 0.026625 \*   
## factor(DESTINATION\_AIRPORT)SLC -12.4775 8.1125 -1.538 0.124080   
## factor(DESTINATION\_AIRPORT)SRQ 9.7873 9.9350 0.985 0.324593   
## factor(DESTINATION\_AIRPORT)STL 23.3766 8.3654 2.794 0.005213 \*\*   
## factor(DESTINATION\_AIRPORT)STT 6.5720 8.4199 0.781 0.435101   
## factor(DESTINATION\_AIRPORT)TPA 13.4466 6.0172 2.235 0.025468 \*   
## factor(DAY\_OF\_WEEK)2 -14.4996 2.2721 -6.382 1.86e-10 \*\*\*  
## factor(DAY\_OF\_WEEK)3 -28.4268 2.2161 -12.828 < 2e-16 \*\*\*  
## factor(DAY\_OF\_WEEK)4 -19.4267 2.2020 -8.822 < 2e-16 \*\*\*  
## factor(DAY\_OF\_WEEK)5 -28.4012 2.1963 -12.931 < 2e-16 \*\*\*  
## factor(DAY\_OF\_WEEK)6 -26.1495 2.4055 -10.871 < 2e-16 \*\*\*  
## factor(DAY\_OF\_WEEK)7 -15.0069 2.4533 -6.117 1.00e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 46.32 on 7130 degrees of freedom  
## Multiple R-squared: 0.06714, Adjusted R-squared: 0.05838   
## F-statistic: 7.659 on 67 and 7130 DF, p-value: < 2.2e-16

Using this regression, we predicted the average departure delay at Boston Logan Airport. After taking into account standard error of our predictive analysis, our confidence interval came out to be between 11.85 minutes and 33.23 minutes for predicted departure delay at Boston Logan.

test.data = logan2delay %>%  
group\_by(DAY\_OF\_WEEK, AIRLINE, DESTINATION\_AIRPORT) %>%  
summarize(DEPARTURE\_DELAY = mean(DEPARTURE\_DELAY))

## `summarise()` has grouped output by 'DAY\_OF\_WEEK', 'AIRLINE'. You can override  
## using the `.groups` argument.

predicted.y = predict(lm\_2, test.data, se.fit= TRUE)  
  
#standard error of predictions   
names(predicted.y)

## [1] "fit" "se.fit" "df" "residual.scale"

head(predicted.y$fit)

## 1 2 3 4 5 6   
## 29.43522 40.66115 28.08831 29.46588 34.83479 21.76701

head(predicted.y$se.fit)

## 1 2 3 4 5 6   
## 4.041011 3.785613 4.102935 3.920729 3.413759 11.359389

#predict departure delay  
test.data$predicted.y=predicted.y$fit  
test.data$se.fit=predicted.y$se.fit  
  
#upper limit of confidence interval for each predicted y  
test.data$upper.limit = test.data$predicted.y +qnorm(0.975)\*test.data$se.fit  
mean(test.data$upper.limit)

## [1] 33.22575

#lower limit of confidence interval for each predicted y   
test.data$lower.limit= test.data$predicted.y -qnorm(0.975)\*test.data$se.fit  
mean(test.data$lower.limit)

## [1] 11.85242

##Conclusion

Based on our data analysis, Spirit, United, Frontier and JetBlue have the highest average departure and arrival delay times. Spirit and United Airlines should take this information to reevaluate and reorganize their internal operations because they have significantly more delays than their other competitors. We speculate that since Spirit and Frontier are both ultra low-cost carriers with lower unit costs and revenues, they have fewer resources to optimize their services, which may be ultimately reflected in the delays they have. On the other hand, Alaska Airlines, Delta Airlines, Hawaiian Airlines, and US Airways have the lowest average departure and arrival delay times. These airlines should take advantage of this information and use it as a selling point to consolidate customer loyalty and enhance their brand image. Customers can also use this information to make better informed decisions.

From the average difference between scheduled and elapsed time by airline graph and the t-tests that we ran, we can see the importance of giving a correctly estimated scheduled time, or at least an overestimated scheduled time, so that consumers do not see the airline in negative light when elapsed time exceeds scheduled. From this, we also see the importance of not viewing only scheduled and elapsed but also other factors such arrival and departure delays as airlines can overestimate the scheduled time as a way to market themselves. Through each test that we ran, we saw that each airline significantly overestimates their scheduled time to account for delays, except for Hawaiian Airlines, which tends to underestimate their scheduled time. It would be interesting to further understand how customers perceive Hawaiian Airlines because of this factor.

Building off our statistical analyses, we conducted hypothesis testings and regression analyses to see if any of the relationships displayed on the graphs were significant. After running regressions on the impact of month on the departure delay, we found that summer and winter months experienced more delays than the other seasons. We predict that this is due to weather and increased flights during times of vacation. Our regression with departure delay against weekdays showed that all weekdays were statistically significant. Saturdays had the lowest average departure delay while Mondays had the highest, which seems to be consistent with the fact that Saturdays had the lowest amount of flights, which means less traffic at the airports. From a managerial perspective, managers should focus on improving flight delays during high flight frequencies in terms of months and days of week, possibly through better staffing or operational services during the summer and winter months, as well as on Mondays and Thursdays. From a customer’s perspective, it is better to avoid the summer and winter months when traveling, as well as on Mondays. Instead, customers can travel on low frequency months and days, such as September and on Saturdays.

From this, created a multiple regression where we analyzed the impact of airline, day of week, month, and taxi-out on departure delay. The longest delays are coming from Spirit Airlines and flights that take place in June, affirming the conclusions of previous regressions performed. As a result, even though the regressions point out specific airline, day of week, and month that have more delays, they explain very little of the variability of the data around the mean, which led us to find a better model that predicts departure delays well.

We also examined the relationship of departure delay across months between the top 15 origin airports with the highest amount of flights. In general, we found that summer and winter months tend to have higher flight traffic across airports than spring and fall months, which supported our previous findings.

We then chose to predict departure delays during the month of February in Boston Logan (BOS) because the airport is the most relevant out of the top 15. Moreover we chose February to predict because it was the month with the highest flight delays for BOS airport. In order to find the best model for prediction, we created 15 different regressions and determined that lm\_2, which regresses departure delay against airline, destination, and weekday, has the lowest predictive error of 40.06130. Using the model, we predicted the departure delay to be between 11.85 minutes and 33.23 minutes at Boston Logan in February.

Each visualization, statistical analysis, and regression led us to better understand the relationship between departure delay and other variables by allowing us to close in on what airlines a customer may want to avoid, what day, what month, and for what reason delays are prominent. Our predictive analysis gave us an idea on how long a typical customer would be delayed for as a result of all these factors that we’ve looked at in this analysis.