2008 Arrival Delay Analysis

American Airlines vs. Everyone Else

Report Date: 06/27/18
Brandon Croarkin [bcroarki@syr.edu]
Michelle Mak [mmak01@syr.edu]
Teng Siong (TS) Yeap [teyeap@syr.edu]

Table of Contents

1. Project Abstract
2. Contribution Statement
3. Final Report
    a. Introduction
    b. The Dataset
    c. Methodology
    d. Analysis
    e. Results
4. Conclusions
5. References
6. Appendix

Project Abstract

In this project, we analyzed all recorded airline delays in the year of 2018. During our analysis, we decided to limit the scope to only arrival delays because departure delays did not seem to impact scheduled arrival times (late departures would still arrive on time). Due to time constraints, we also ignored cancellations and just focused on the delays. Our goal was to find out specifically which airlines and locations exhibited a pattern of heavy delays in order to look for solutions to avoid arrival delays in the future. We framed our perspective from the carrier American Airlines and sought improvement within that context.

In order to reach these objectives we used various exploratory data analysis techniques to analyze the trends in the data. This involved a large amount of ggplot2 and visualizations ranging from line charts, histograms, bar charts, and maps. Additionally, we used linear regression, SVM models, and logistic regression to predict the length of a delay and to understand the causes.

We were unable to achieve a high classification rate or adjusted R-squared in these models. This is largely due to some of the inherent randomness in flight delays and the fact that we did not have data on weather, which likely would have increased our predictive abilities.

Contribution Statement

Brandon Croarkin: For the Type 1 Questions, I answered which airline experiences the most/longest delays and which airport has the most/longest delays. I also visualized air travel data for American Airlines. For Type 2 Questions, I worked to predict the length of a delay via a linear regression. I also attempted to predict whether there would be a delay or not (defined as over 24 minutes) using a logistic regression model."

Michelle Mak: For this project, I set up the initial project management aspects by way of Trello and Google docs. We worked as a team to fill in the necessary steps and came up with the questions to complete the analysis for this topic. In terms of analysis, I answered the following Type 1 Questions "How frequently do delays occur?" and "What is the average length of a delay?" Additionally, I investigated the Type 2 questions "What are the impacts of the delays?" by looking at the 4 given causes in the data set.

TS Yeap: I found the dataset for our project and started off by cleaning the data for the group. Secondly, my assigned Type 1 Question investigated which month, day of the month, and day of the week have the most arrival delays.

Introduction

We decided to do our project on the Airlines Delay dataset because it seemed interesting and is a topic that is relevant to nearly everyone. Through our analysis, we hoped to better understand the causes of airline delays so we could help better pick when, where, and with whom we would fly in the future. The dataset also seemed especially well-suited for data science methods, in particular for exploratory data analysis, and we thought we would be able to find a lot of interesting things out about the data.

The main questions we wanted to answer were when, where, and who has the most/longest airline delays. Based on these questions, we wanted to see if and how well we could predict airline delays.

In order to give our analysis some context and a goal, we framed our questions through the perspective of American Airlines(AA). We used AA's data and compared it to the industry average in order to find out where and how a specific airline could improve.

Ultimately, we were unable to predict airline delays. Airline delays have a large degree of uncertainty applied to them and without additional data such as weather or diagnostic reports of the plane, it is hard to make accurate predictions -- especially since human behavior is such a large component of delays as well. However, we did gain a lot of insights into what some airlines do better than others and which airports are more prone to delays. We hope to apply this knowledge in our future flight purchases!

Type 1 Findings

1. How many delays did each airline experience in 2008?/ Which airlines experience the most delays?

Airline Code	**Total # of Delays**
AQ	654
HA	7199
F9	25708
AS	34179
9E	46896
B6	48177
OH	49104
YV	63289
FL	65008

NW	72395
EV	75170
US	83262
CO	83646
XE	94313
DL	100923
OO	121942
UA	123989
MQ	130647
AA	172197
WN	324717

2. Which airline experiences the least delays?
    - Top 5
        1. AQ (9 Air Co Ltd) - 21 minute average arrival delay
        2. F9 (Frontier Airlines, Inc.) - 27 minute average arrival delay
        3. WN (Southwest Airlines Co.) - 30 minute average arrival delay
        4. HA (Hawaiian Airlines, Inc.) - 34 minute average arrival delay
        5. AS (Alaska Airlines Inc.) - 36 minute average arrival delay
    - Bottom 5
        1. YV (Mesa Airlines, Inc.) - 55 minute average arrival delay
        2. B6 (Jetblue Airways Corporation) - 55 minute average arrival delay
        3. OH (PSA Airlines, Inc.) - 51 minute average arrival delay
        4. XE (Delux Public Charter LLC) - 48 minute average arrival delay
        5. UA (United Airlines, Inc.) - 48 minute average arrival delay

3. What is the average length of a delay?
    - American Airlines:
        - Mean: about 42 minutes
        - Median: about 29 minutes
        - Mode: about 10 minutes
    - Industry Average:
        - Mean: about 42 minutes
        - Median: about 24 minutes
        - Mode: about 10 minutes
4. When do arrival delays most frequently occur?
    - Which day of the week that has the most arrival delays?

        - American Airlines: Friday
        - Industry Average: Friday
    - Which Month has the most delays?
        - American Airlines: June
        - Industry Average: December
5. Which airport has the most delays?
    - American Airlines:
        - Worst Airports
            1. GUC (Gunnison, Colorado) - 126 minute average arrival delay, 53 trips
            2. HDN (Hayden, Colorado) - 117 minute average arrival delay, 99 trips
            3. EGE (Vail, Colorado) - 103 minute average arrival delay, 300 trips
            4. KOA (Kailua, Hawaii) - 96 minute average arrival delay, 122 trips
            5. JAC (Jackson, Wyoming) - 82 minute average arrival delay, 121 trips
        - Best Airports
            1. OAK (Oakland, California) - 20 minute average arrival delay, 199 trips
            2. FAT (Fresno, California) - 36 minute average arrival delay, 118 trips
            3. BWI (Baltimore, Maryland) - 38 minute average arrival delay, 648 trips
            4. CMH (Columbus, Ohio) - 38 minute average arrival delay, 259 trips
            5. BUR (Burbank, California) - 39 minute average arrival delay, 343 trips
    - Industry:
        - Worst Airports
            1. CMX (Hancock, Michigan) - 123 minute average arrival delay, 34 trips
            2. PLN (Pellston, Michigan) - 95 minute average arrival delay, 21 trips
            3. SPI (Springfield, Illinois) - 87 minute average arrival delay, 356 trips
            4. MQT (Marquette, Michigan) - 80 minute average arrival delay, 190 trips
            5. ALO (Waterloo, Iowa) - 80 minute average arrival delay, 31 trips
        - Best Airports
            1. TUP (Tupelo, Mississippi) - 6 minute average arrival delay, 1 trip
            2. INL (International Falls, Minnesota) - 15 minute average arrival delay, 1 trip
            3. SLE (Salem, Oregon) - 15 minute average arrival delay, 85 trips

4. WYS (West Yellowstone, Montana) - 18 minute average arrival delay, 10 trips
5. ADK (Adak Island, Alaska) - 20 minute average arrival delay, 53 trips

Type 2 Findings (Exploratory)

1. What are the causes of airline delays?
    - The data set came with five predetermined causes: carrier delay, weather delay, National Air Security delay, security delay, late arrival delay

Cause	Avg. Delay (min)
Carrier (Industry)	19.18
Carrier (AA)	21.34
Weather (Industry)	3.70
Weather (AA)	3.17
NAS (Industry)	15.02
NAS (AA)	16
Security (Industry)	0.09
Security (AA)	0.05
Late Arr (Industry)	25.3
Late Arr (AA)	25.21

    - Based on these numbers, it is clear that late arrivals were those most negatively impacting type of delay. So it is fitting that we were trying to mitigate future late arrivals.

2. What factors can we use to predict delays?
    - American Airlines:
        i. Model Used: aa.model2 <- lm(data = aa, formula = ArrDelay ~ Distance + Weekend + Month + DepHour)
        ii. Relative Importance
            1. DepHour - 67%
            2. Month - 32%

3. Distance - 0.5%
4. Weekend - 0.006%

- Industry Average:
  i. Model Used: model2 <- lm(data = df_clean, formula = ArrDelay ~ UniqueCarrier + Distance + Weekend + Month + DepHour)
     1. Note: I did not use Origin in this model because R Studio was freezing when I tried to run the model with it included.
  ii. Relative Importance
     1. UniqueCarrier - 42%
     2. DepHour - 39%
     3. Month - 16%
     4. Distance - 2%
     5. Weekend - 0.02%
- Top 3 Competitors:
  i. Model Used: comp.model2 <- lm(data = df_clean, formula = ArrDelay ~ Distance + Weekend + Month + DepHour)
     1. Note: I did not use Origin in this model because R Studio was freezing when I tried to run the model with it included.
  ii. Relative Importance
     1. DepHour - 67%
     2. Month - 32%
     3. Distance - 0.5%%
     4. Weekend - 0.006%

3. Can we predict the length of a delay?
   - Linear Regression
     i. Model: model4 <- lm(data = train, formula = ArrDelay ~ UniqueCarrier + Distance + Weekend + DepHour + Origin)
        1. Adjusted R-Squared = .04586
   - Linear Regression with LogArrivalDelay
     i. Model: model5 <- lm(data = df_clean_log, formula = logArrDelay ~ UniqueCarrier + Origin + Distance + DayOfWeek + DepTime + Month)
        1. Adjusted R-Squared = .07034
   - Logistic Regression
     i. Note: made variable for OverMedian that states whether the flight's arrival delay is over or under the median time of 24 minutes
     ii. Model: logmodel1 <- glm(OverMedian ~ Weekend + Origin + DepHour + Distance + Airline, data = test, family = binomial)
        1. Prediction Percent = 59%

Based on this we recommend….

There is not much we can recommend based on our analysis since we do not have other dataset to support the prediction of a delay. We, however, do have some insights to share. For example, Southwest Airline is one of the airlines that has the least delays compared to other like airlines. Additionally, the data shows that a delay is very likely to happen when traveling on Friday, so maybe plan accordingly for your next long weekend. Furthermore, plan for more travel time in December as delays often occurs due to holiday/vacation season. You do not want to miss celebrating a holiday with your family due to your flight delays!

The Dataset

1. Airline Delay Dataset:
   Link: https://www.kaggle.com/giovamata/airlinedelaycauses/data
   Main features that we use in this dataset

Month	DayOfWeek
UniqueCarrier	ArrDelay
Origin	Dest
Distance	DepTime
CarrierDelay	WeatherDelay
NASDelay	SecurityDelay
LateAirCraftDelay	DayOfMonth

Table 1. Features in the dataset.

   Additional features engineered from the dataset.
   - DepHour = DepTime / 100
   - OverMedian = whether the ArrDelay is over the median ArrDelay
   - Weekend = Friday, Saturday, and Sunday

2. Airline Information:
   Link: https://raw.githubusercontent.com/jpatokal/openflights/master/data/airlines.dat
   Features used from dataset:
   - Airline name
3. Airport Information
   Link: https://raw.githubusercontent.com/jpatokal/openflights/master/data/airports.dat
   Features used from this dataset:
   - Lat
   - Long

Data Analysis Methods

The data analysis methods we used were linear and logistic regression. Both of these are forms of supervised statistical learning that are useful in predicting an outcome, but they differ in the data type of the outcome variable they are trying to predict.

Linear regression is useful for predicting a quantitative response Y. It can take a single predictor variable (simple regression) or multiple predictor variables (multiple linear regression). Linear regression is one of the best approaches available as a starting point for modeling and predicting data. The linear regression model attempts to estimate the relationship of your data and fit a line that explains this relationship. This line of best fit attempts to minimize the residuals (the difference between the predicted outcome and actual outcome). More specifically, it tries to minimize the sum of squared residuals.

The output of this model is an intercept (B0) and slopes (B1, B2, etc.) for our predictor variables. In our example, since we have X that never equal 0, there is no intrinsic meaning to the intercept. The slopes/coefficients tell us how our output differs at different values of the input. For categorical variables, such as airline, this coefficient is saying how a specific airline increases or decreases the predicted arrival delay. For continuous variables, like distance, we can see how a one-unit increase in the distance of a flight affects the average arrival delay.

In evaluating a linear regression model, the most critical metrics to check when evaluating how well it does in predicting the target variable are the F-statistic, adjusted R-squared, residuals, and the p-value of the coefficients. The F-statistic gives the statistical significance of the equation. If the p-value for the F-statistic is over .05, we can not have confidence in the model. The adjusted R-squared tells us how much of the variation in Y can be explained by our X variables. This lets us know how well our model actually fits the data. This value ranges from 0 to 1, where we want to be as close to 1 as possible. Next, we want to look at the residuals which give a measure of the model fit. Lastly, we want to look at the p-value/significance of our coefficients. We want the p-value to be less than .05 so we know the coefficient is significant so we can have confidence in its value.

Logistic regression is similar to linear regression, but it is used to model dichotomous outcome variables. In our case, we tried to predict whether there is a delay or not, which is a yes or no answer. Many of the metrics we look at to evaluate a logistic regression are the same as the ones we use to evaluate linear regression, but there is a big difference between the two in how we evaluate the coefficients. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

Analysis Summary

The first phase of our project was exploratory data analysis. This was done to understand the data and to gain some intuition about what to factor into our model. When working to determine what to explore, we thought some important variables to analyze would be

the actual arrival delays, the origin of delays, date data (day of the week, month, day of the month, etc.), and airline. We decided to do this all from the perspective of American Airlines in order to see where we stacked up against other competitors and the industry as a whole. From this analysis we would hope to take the information learned in order to decrease our average arrival delay by gaining insight into what causes arrival delays.

One of the first things we realized when analyzing the arrival delays was the large spread in the arrival delays, in particular some large outliers. This is evident from the summary statistics of arrival delay and from a histogram of the data. Additionally, we noticed some NA's in the data that we eventually removed.

ArrDelay	**Minutes**
Min.	-109.0
1st Qu.	9.0
Median	24.0
Mean	42.2
3rd Qu.	56.0
Max.	2461.0

Figure 1: Summary statistics for ArrDelay

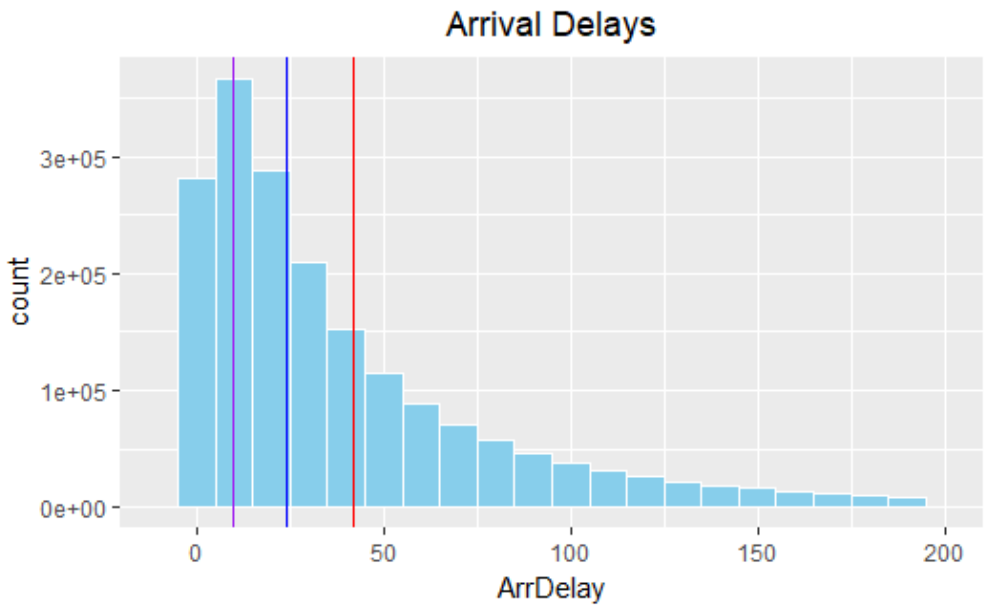

Figure 2: Histogram of Arrival Delays (cut off values over 200 minutes from graph). Mode (purple), median (blue), mean (red).

Once we had this data on the arrival delays as an industry, we wanted to dive into this information airline. One of the first things we noticed when we tried to compare delays by airline,

was the discrepancy in flights between airlines. To normalize this fact in our analyses, we decided to look at the percent of flights delayed. We defined a delayed flight as over the median arrival delay of 24 minutes.

Airline	**Flights**	**percentTotal**
Southwest Airlines	376201	18.55
American Airlines	190910	9.41
American Eagle Airlines	141223	6.96
United Airlines	140904	6.95
SkyWest	131780	6.5
Delta Air Lines	113728	5.61
ExpressJet	103147	5.09
Continental Airlines	99731	4.92
Continental Express	99731	4.92
US Airways	98007	4.83
Atlantic Southeast Airlines	81762	4.03
Northwest Airlines	78843	3.89
AirTran Airways	70969	3.5
Mesa Airlines	66769	3.29
JetBlue Airways	54925	2.71
Comair	52453	2.59
Pinnacle Airlines	51569	2.54
Alaska Airlines	39010	1.92
Frontier Airlines	28224	1.39
Hawaiian Airlines	7472	0.37
Aloha Airlines	744	0.04

Figure 3: Flights by Airline

In our analysis of American Airlines we also wanted to compare ourselves to similar companies. Since American Airlines is a large airline with a lot of flights, we decided to compare ourselves to three like companies: United Airlines, Delta Airlines, and Southwest Airlines. By framing ourselves to like companies, we hope to understand how we relate to our main competition so we can more accurately understand where we stand. From the analysis below, we sadly see that the American Airlines is one of the worst offenders for the percent of flights delayed and is the worst of the competitors as well.

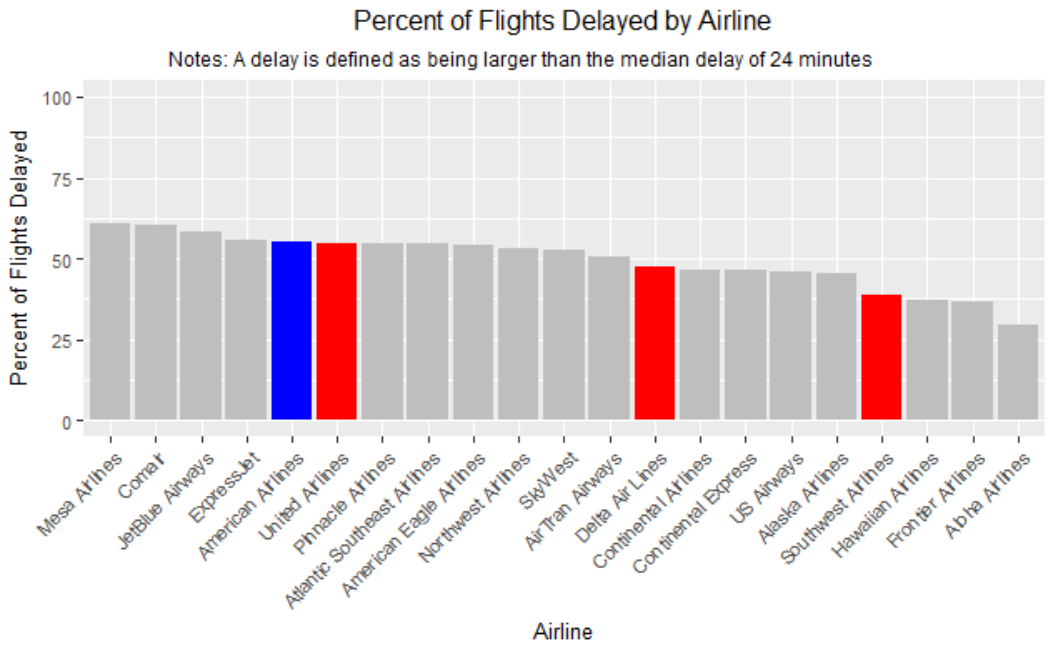

Figure 4: Percent of Flights Delayed by Airline

The above analysis gives information on the quantity of flight delays, but it is also important to understand the length of the delay. To do this we looked at the average arrival delay airline and compared it to the mean. Again, we can see that we are above the average for the length of our arrival delays.

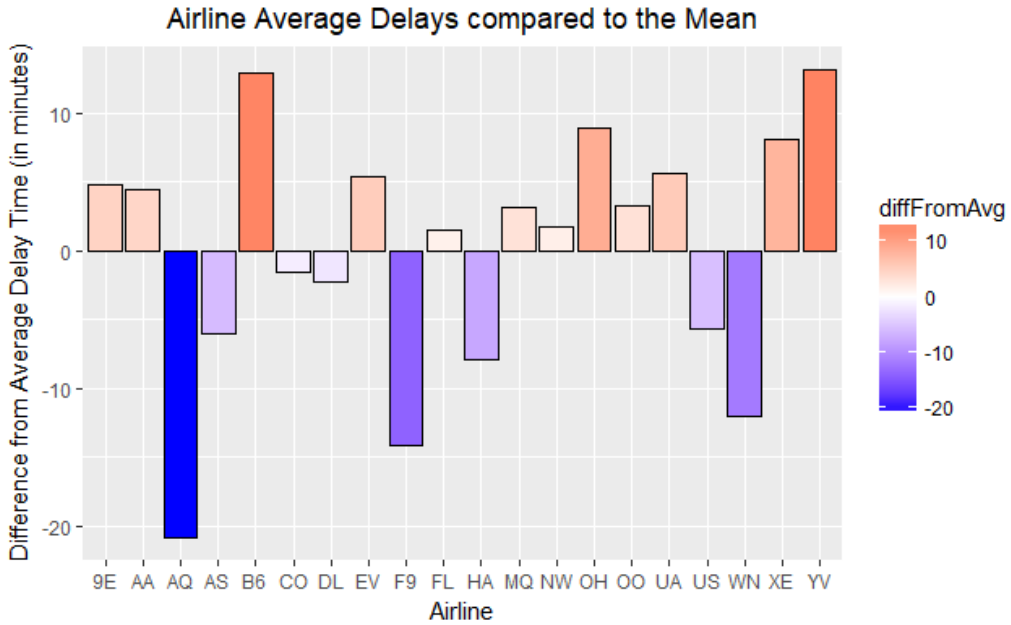

Figure 5: Average airline delays compared to the mean arrival delay.

Upon knowing where we stand as a company, the next step was to understand what factors go into causing airline delays. We started by looking by looking at the date variables

(Month, DayOfWeek, and DayOfMonth). Our initial intuition was that delays would occur most in the winter months where the weather is the worst, during the weekends when more flights are probably occurring, and towards the end of the month. We were partially correct, but not completely!

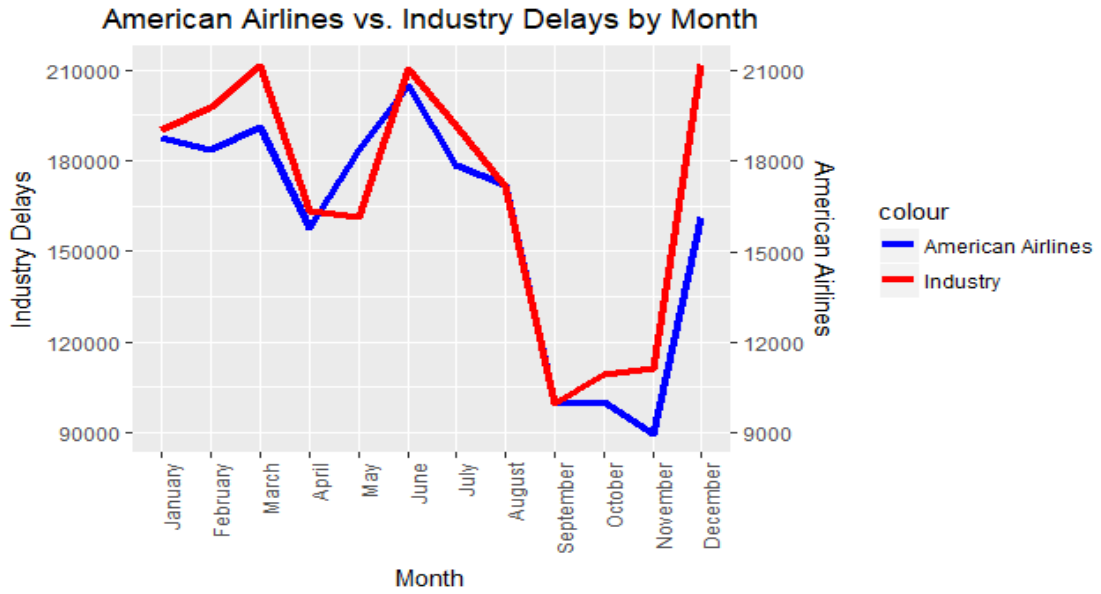

Figure 6: Delays by Month. American Airlines (blue) vs. all flights (red).

From the above chart we can see that we were partially right, but not completely. December is the month with the most delays, but March and June were the next highest. From this trend, it appears that most of the delays occur during these periods that are likely correlated with more overall flights. Each of these months is a month with a lot of holiday/vacation travel.

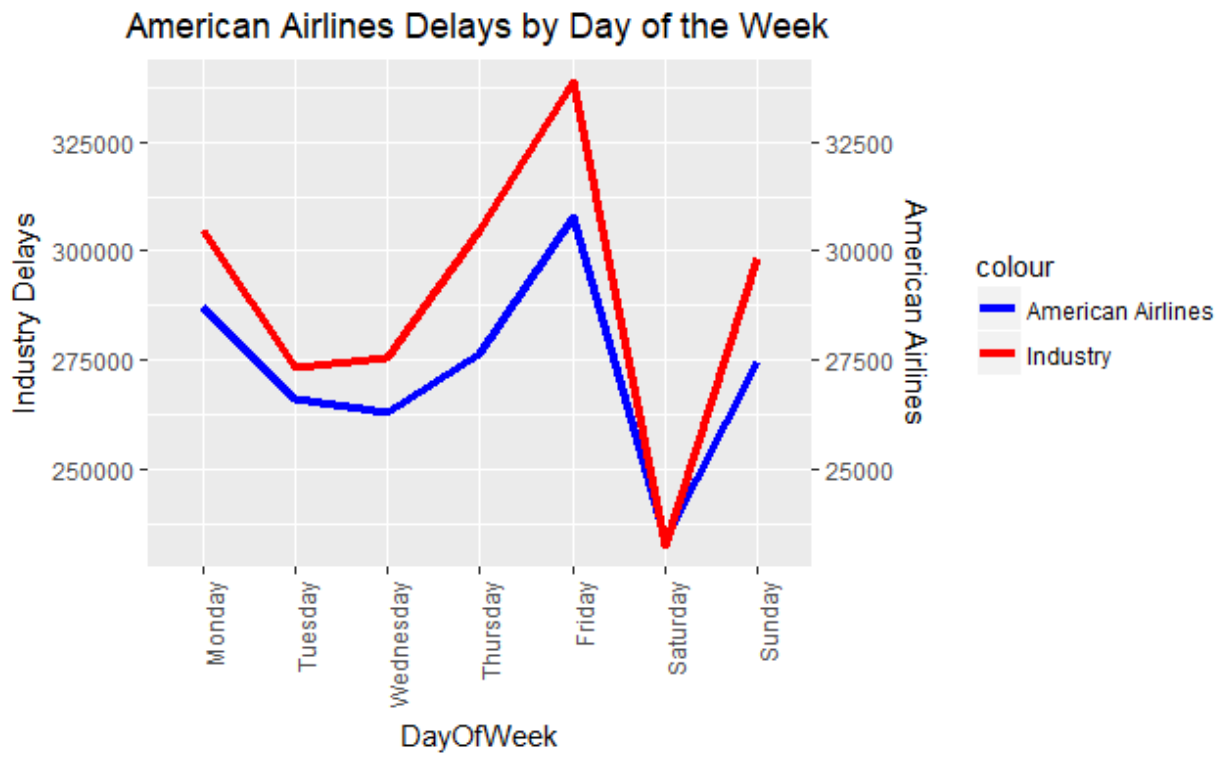

Figure 7: Delays by day of the week. American Airlines (blue) vs. all flights (red).

Again, similar to with the month analysis from above, we see a now predictable pattern of delays occurring on Friday, Monday, and Sunday, which are three days that make sense for getting a lot of air traffic.

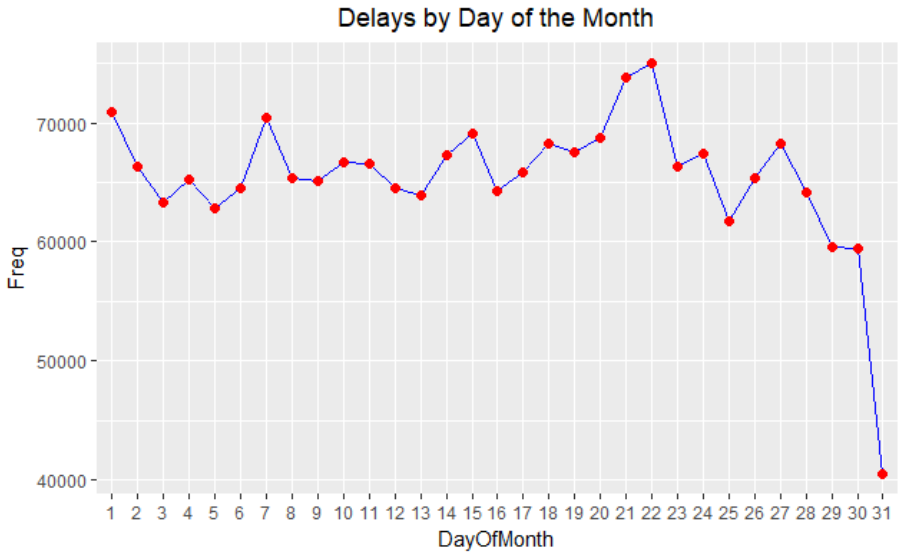

Figure 8: Delays by Day of the Month

We do not notice any noticeable pattern for delays by day of the month, except for the drop on the 31st. This drop makes sense since not all months have the 31st, so this is merely just a day with less flights. Interestingly, the 21st and 22nd are the days with the most delays. It is interesting that the top two days occur right next to each other, but there does not seem to be any obvious reason for this trend.

After exploring this date variable in-depth, the next step was to analyze the origin of the flight to see if this has any impact. An initial plotting of the data on a map, does not give any immediate insight. There appears to be longer delays on the east coast though, as indicated by the darker shade of red. We can also quickly see some of the larger airport hubs.

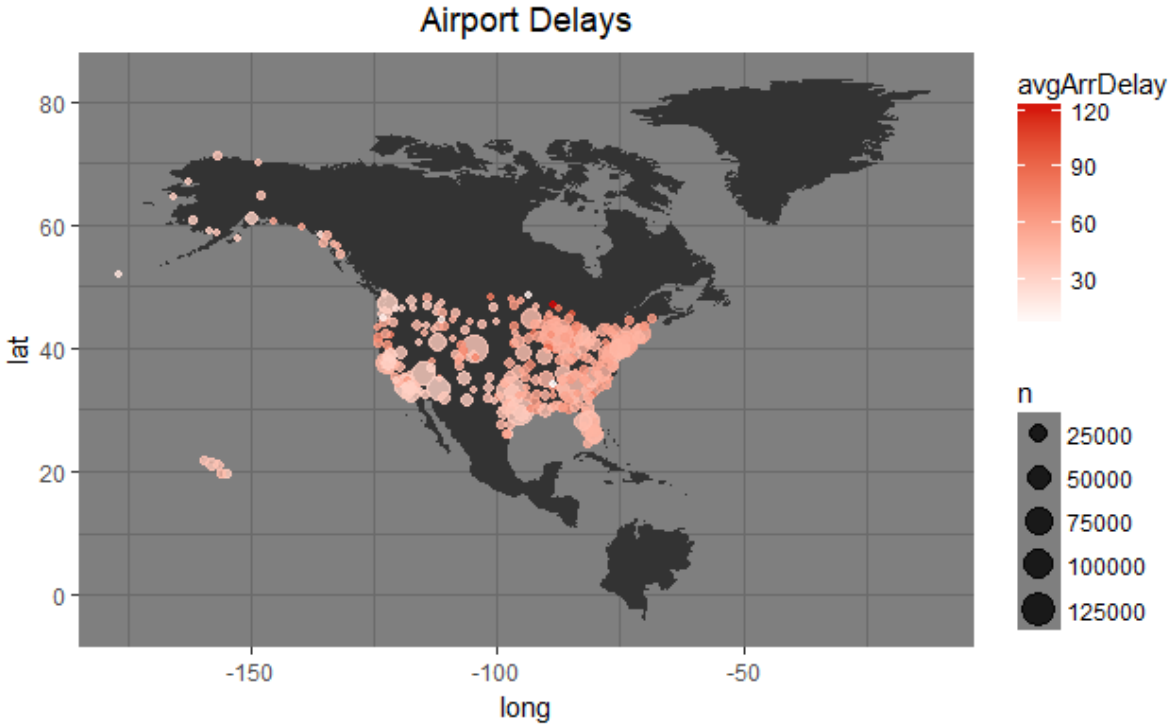

Figure 9: Airport delays map by origin airport

Although, the above map looks nice, it does not do a great job of quickly giving insight on the delays by origin. To help achieve this, a different visual tool is needed. The bar charts below give a quick visual on what airports are the best and worst for airport delays. The color of the bar also gives some additional information on the number of flights that occur at each airport.

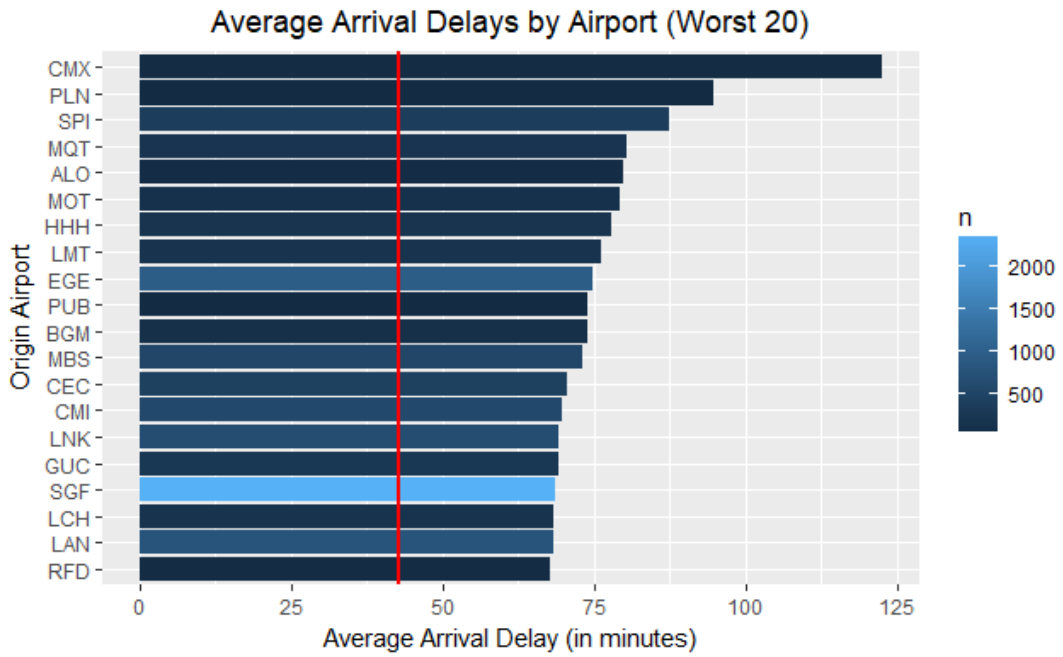

Figure 10: Average arrival delay by airport (worst 20). All airlines included. Red line is the mean arrival delay.

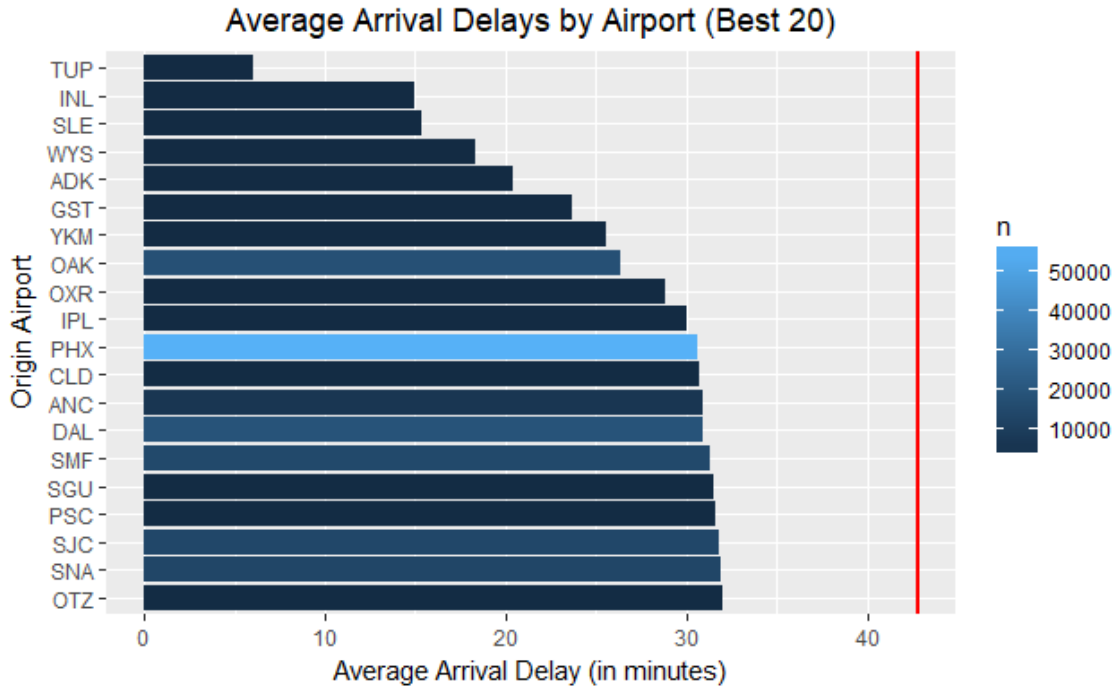

Figure 11: Average arrival delay by airport (best 20). All airlines included. Red line is the mean arrival delay.

Although this is useful information, in order to gain additional insights for American Airlines it is important to look at the airports specific to American Airlines flights so we can understand where we experience the delays.

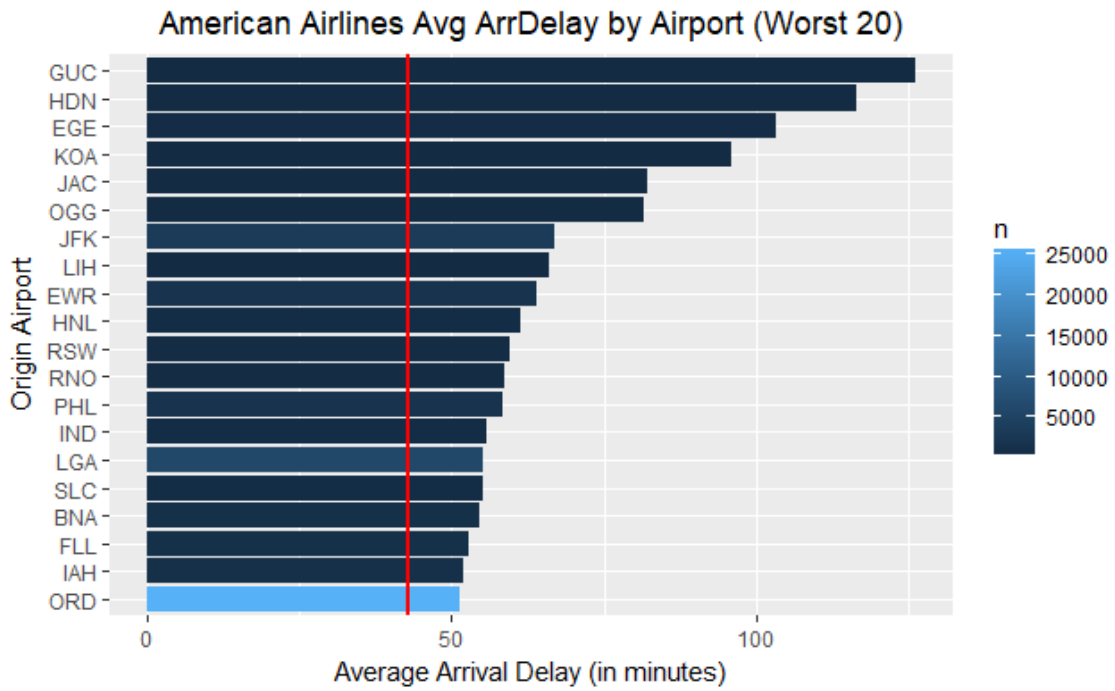

Figure 12: Average arrival delay by airport (worst 20) for American Airlines. Red line is the mean arrival delay.

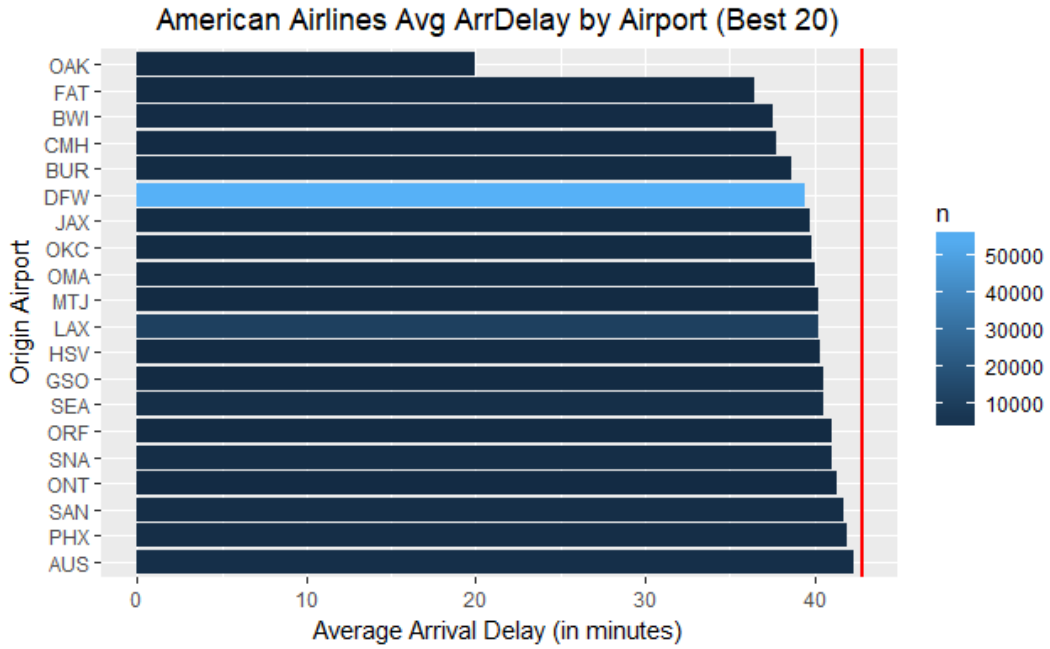

Figure 13: Average arrival delay by airport (best 20) for American Airlines. Red line is the mean arrival delay.

After this exploratory data analysis, we can understand from a quick glance at each chart a different aspect of what causes delays. This leads us to the question that this is all about, predicting actual delays. The first model used was a multiple linear regression. We used

the variables that we found to have some impact from the EDA above and a couple additional ones that were included based on intuition. Some of the first models used were more basic and we used to find the relative importance of different variables. Next, we iterated through different models with different variables to find the one with the best adjusted R-squared.

```{r}
#find the relative importance of the variables
relImportance <- calc.relimp(model2, type = "lmg", rela = TRUE)
sort(relImportance$lmg, decreasing = TRUE)
```

```
 UniqueCarrier        DepHour          Month       Distance        Weekend
   0.422078591   0.394731577   0.160721723   0.022207263   0.000260846
```

Figure 14: Relative importance of the variables in predicting arrival delay. Model used was ArrDelay ~ UniqueCarrier + Distance + Weekend + Month + DepHour

RStudio: Notebook Output

```
Call:
lm(formula = ArrDelay ~ UniqueCarrier + Distance + Weekend +
    DepHour + Origin, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-135.36  -31.46  -14.83   13.16 2436.69

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.602e+01  2.233e+00  16.131  < 2e-16 ***
UniqueCarrierAA 5.241e+00  3.929e-01  13.339  < 2e-16 ***
UniqueCarrierAQ -1.553e+01  2.568e+00  -6.049 1.46e-09 ***
UniqueCarrierAS -4.022e+00  5.608e-01  -7.171 7.43e-13 ***
UniqueCarrierB6  8.052e+00  4.781e-01  16.842  < 2e-16 ***
UniqueCarrierCO  1.345e+00  4.009e-01   3.355 0.000794 ***
UniqueCarrierDL -2.565e+00  4.080e-01  -6.285 3.27e-10 ***
UniqueCarrierEV  6.897e-01  4.364e-01   1.580 0.114026
UniqueCarrierF9 -1.207e+01  5.542e-01 -21.784  < 2e-16 ***
UniqueCarrierFL -2.223e+00  4.363e-01  -5.094 3.51e-07 ***
UniqueCarrierHA -1.354e+00  1.038e+00  -1.304 0.192228
UniqueCarrierMQ -1.239e+00  3.929e-01  -3.154 0.001613 **
UniqueCarrierNW  9.357e-01  4.014e-01   2.331 0.019752 *
UniqueCarrierOH  5.181e+00  4.692e-01  11.041  < 2e-16 ***
UniqueCarrierOO -2.122e-01  4.053e-01  -0.524 0.600563
UniqueCarrierUA  4.255e+00  4.008e-01  10.617  < 2e-16 ***
UniqueCarrierUS -6.773e+00  4.259e-01 -15.903  < 2e-16 ***
UniqueCarrierWN -1.509e+01  3.752e-01 -40.220  < 2e-16 ***
UniqueCarrierXE  6.615e+00  4.102e-01  16.126  < 2e-16 ***
UniqueCarrierYV  8.854e+00  4.442e-01  19.932  < 2e-16 ***
Distance        -2.211e-03  9.598e-05 -23.039  < 2e-16 ***
Weekend         -8.335e-02  1.058e-01  -0.788 0.430876
DepHour          1.709e+00  1.072e-02 159.525  < 2e-16 ***
OriginABI        7.384e+00  4.030e+00   1.832 0.066919 .
OriginABQ       -1.676e+01  2.306e+00  -7.268 3.66e-13 ***
OriginABY       -7.362e+00  4.485e+00  -1.641 0.100699
OriginACK        3.952e-01  5.171e+00   0.076 0.939080
OriginACT       -1.507e+01  4.133e+00  -3.645 0.000267 ***
OriginACV        9.912e-01  3.002e+00   0.330 0.741255
```

```
OriginLAX       -2.143e+01  2.228e+00  -9.620  < 2e-16 ***
OriginLBB       -1.580e+01  2.643e+00  -5.978 2.26e-09 ***
OriginLCH        3.909e+00  5.918e+00   0.660 0.508983
OriginLEX       -7.848e-02  2.676e+00  -0.029 0.976602
OriginLFT        8.695e-01  3.086e+00   0.282 0.778146
OriginLGA       -9.920e+00  2.243e+00  -4.423 9.73e-06 ***
OriginLGB       -2.613e+01  2.501e+00 -10.447  < 2e-16 ***
OriginLIH       -2.409e+01  2.807e+00  -8.580  < 2e-16 ***
OriginLIT       -7.725e+00  2.433e+00  -3.176 0.001495 **
OriginLMT        2.264e+01  5.775e+00   3.920 8.86e-05 ***
OriginLNK        8.965e+00  3.457e+00   2.594 0.009497 **
OriginLRD       -1.203e+01  3.857e+00  -3.118 0.001818 **
OriginLSE       -2.987e+00  3.849e+00  -0.776 0.437806
OriginLWB       -7.139e+00  9.540e+00  -0.748 0.454274
OriginLWS       -1.607e+01  7.522e+00  -2.136 0.032683 *
OriginLYH       -8.461e-01  8.257e+00  -0.102 0.918383
OriginMAF       -1.210e+01  2.744e+00  -4.410 1.03e-05 ***
OriginMBS        7.904e+00  3.686e+00   2.144 0.031997 *
OriginMCI       -1.407e+01  2.274e+00  -6.185 6.21e-10 ***
OriginMCN        1.265e+01  5.893e+00   2.147 0.031786 *
OriginMCO       -1.540e+01  2.238e+00  -6.879 6.04e-12 ***
 [ reached getOption("max.print") -- omitted 125 rows ]
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.67 on 1371626 degrees of freedom
Multiple R-squared:  0.04608,   Adjusted R-squared:  0.04586
F-statistic: 204.5 on 324 and 1371626 DF,  p-value: < 2.2e-16
```

Figure 15: Multiple linear regression with UniqueCarrier, Distance, Weekend, DepHour, and Origin as the independent variables. Some of the coefficients are left out above to due to quantity. Best performing linear regression model. Adjusted R-squared of 0.04586.

We were not able to achieve a great R-squared in this model. An analysis of the residuals gives us some insight on why this is. We have such a large spread in residuals. It appears that there is a lot of inherent error in the model that makes predicting airline delays hard.

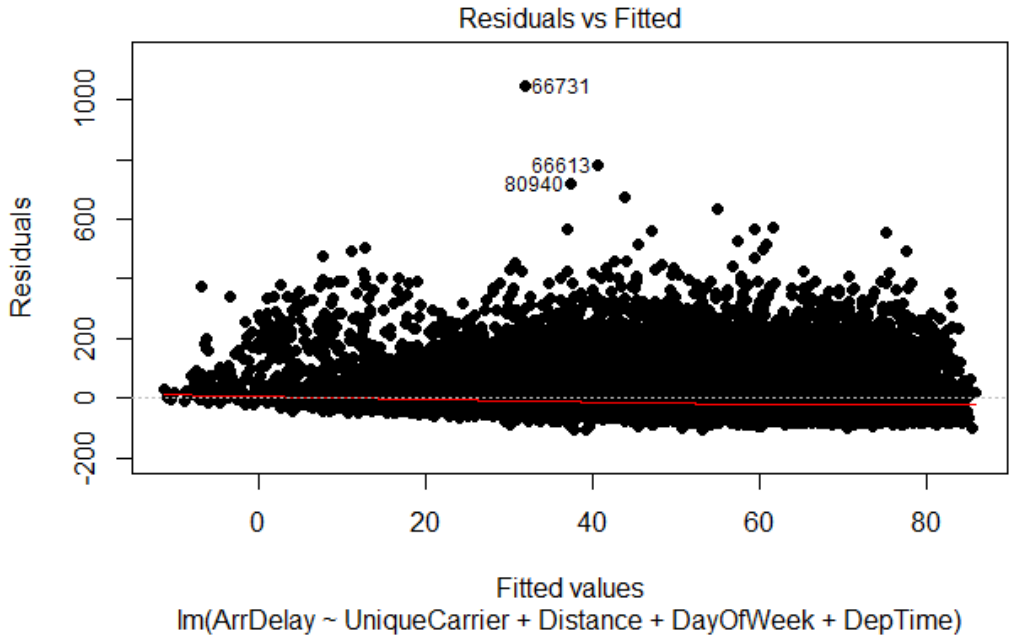

Figure 16: Residual plot

In an attempt to improve the model and factor in the large spread in arrival delays, we took the log of the arrival delay and re-ran the model. Doing so did allow us the increase the adjusted R-Squared to 0.07034. Although this is an improvement, it is still a pretty small result.

From this it appears that with the data we have available currently, we are only able to account for 7% of the variation in arrival delay.

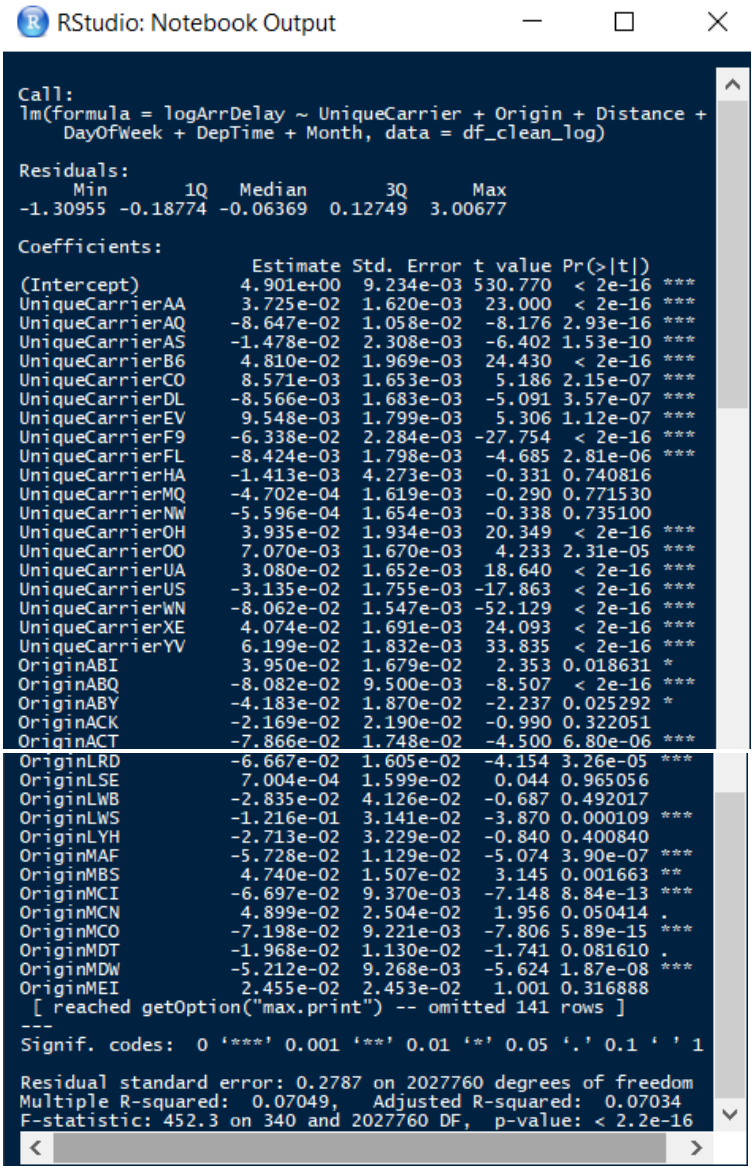

Figure 17: Linear regression with log arrival delay as the dependent variable and UniqueCarrier, Distance, DayOfWeek, DepTime, and Origin as the independent variables. Adjusted R-Squared is 0.07034.

Since it appears that predicting the length of the arrival delay is pretty hard to do with the data available, the next attempt was to try to predict whether there would be a delay or not. We again defined a delay as being over the median arrival delay. We then performed a logistic regression model to predict whether it was going to be a delay or not. After training the model on a test data set comprised of a random 70% of the data and trying it on our test data, which was the remaining 30% of the data, we were able to get a successful classification rate 58%. Similar to our linear regression models, we see that we are able to have only a minor success in

predicting whether there is a delay or not as we only have an 8% increase in predictive ability than we would have had from random guesses.

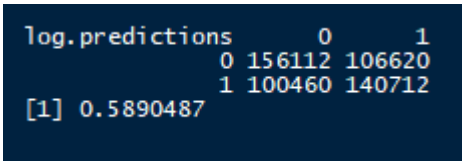

Figure 18: Classification matrix for our logistic regression model. Model was: glm(OverMedian ~ Weekend + Origin + DepHour + Distance + Airline, data = test, family = binomial).

Conclusions

Although we were able to gather a lot of valuable and interesting information, some additional would have helped to shed additional insights. One of the top things we could have added would have been weather information. If we could have had some hour by hour weather information merged into this dataset for both the origin and destination, we likely could have had much better success in predicting arrival delays since weather is such a large factor in this. Furthermore, it would have been good to have some free-form text fields from a pilot or flight attendant on the reason for the delay. This could have led to some good opportunities to incorporate text mining to add information on what causes delays.

References

- http://flowingdata.com/2011/05/11/how-to-map-connections-with-great-circles/
- https://hortonworks.com/tutorial/predicting-airline-delays-using-sparkr/
- https://dplyr.tidyverse.org/
- https://stats.idre.ucla.edu/r/dae/logit-regression/
- https://rpubs.com/FelipeRego/SimpleLinearRegression

Appendix

GitHub Link:
https://github.com/IST659Group2/AirlineDelays/blob/master/IST687_FinalProjectAnalysis.Rmd

Trello Link:
https://trello.com/b/qsjyhv2F/final-project