****

**DSC 441: Fundamentals of Data Science**

**Project Final Report**

**Group Name:** Speed Demons​

**Members:** Mashall Jahangir **Email Address:** ​ mashaljahangir@gmail.com​

Chris Noreikis chris.nore@gmail.com

Annie Nguyen A.Tho.Nguyen@gmail.com​

Hima Barla spandanabarla26@gmail.com

Tianhao Tan tthtantianhao@gmail.com​

**Appendix:**

**Group Technical Report**

**1. Abstract**

**2. Introduction**

**3. Literature Review**

**4. Data**

a. Variable Summary

b. Statistical Summary

c. Data Snapshot

d. Data Pipeline

e. New Feature Creation

**5. Exploratory Analysis and Visualizations**

**6. Model Building**

1. Model Building Process
2. Baseline Model
3. Linear Regression
4. Regression Decision Tree

**6. Results and Conclusions**

**7. References**

## 

## **Group Technical Report**

**1. Abstract:**

Big Data is such a popular and in-demand term in today’s world. Many companies and organizations are looking for data scientists to use the voluminous information to settle on sound business choices so as to amplify list potential and limit misfortune. The Bureau of Transportation Statistics (BTS) is constantly analyzing real-time data based on several airlines. We want to determine obvious trends and patterns to understand what factors and variables cause airline delays. With the use of advanced visualizations, we will attempt to make beneficial observations. Each individual chose a different technique to analyze the dataset and answer the research questions. We will use our understanding of past causes of flight delay to create data science models that predict the delay time of future flights.

**2. Introduction:**

On the air travel passenger complaints by the US today, the top five problems are baggage, cancellation, reservations and ticketing, customer service and Delay. Our goal is to analyze an aggregated Airline Dataset from the year 2019 to see what airline has experienced maximum delays and what caused those delays. And to do this, we plan to produce visualizations from the variables including the month of the year, UniqueCareer Code, Tail Num, Origin, Destination, Arrival time, Departure time and etc. Our dataset includes 7,422,037 observations with 37 variables. Table 1.1 shows important variables that were included in the dataset. Table 1.2 shows variables that were merged in from external sources or derived from existing data for better queries and prediction models.

**3. Data**

**a. Variable Summary**

Dataset Name: Reporting Carrier On-Time Performance (1987-present)

Airport Longitude and Latitude for each Airport State

The **outcome variable** is DEP\_DELAY\_NEW. All other variables are **independent variables**.

**Table 1.1, Variables from BTS**

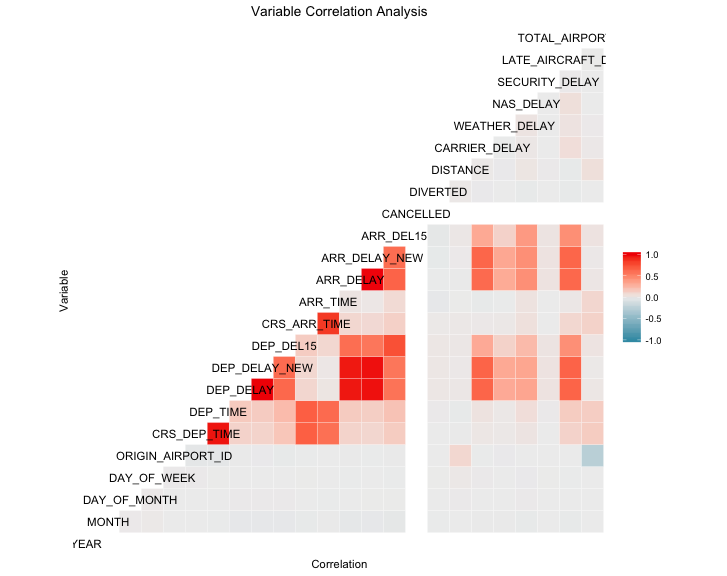
| **Name** | **Description** | **Examples** |
| --- | --- | --- |
| **Month** | The month of 2018. Values from 1-12 to indicate. | 1, 2, 3 |
| **OpCarrierAirlineID** | An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation. | DL, OO, AA |
| **Origin** | Origin Airport | LAX, ORD, MDW |
| **Dest** | Destination Airport | LAX, ORD, MDW |
| **DestCityName** | Destination Airport, City Name | Chicago, IL |
| **DestStateAbr** | Destination Airport, State Code | IL |
| **CRSDepTime** | CRS Departure Time (local time: hhmm) | 1230 |
| **DepTime** | Actual Departure Time (local time: hhmm) | 1420 |
| **DepDelay** | The Difference in minutes between scheduled and actual departure time. Early departures show negative numbers. | -17, 5, 30 |
| **DepDelayNew** | The Difference in minutes between scheduled and actual departure time. Early departures set to 0. | 0, 5, 30 |
| **DepDelay15** | Departure Delay Indicator, 15 Minutes or More (1=Yes), (0=No) | 0, 1 |
| **CRSArrTime** | CRS Arrival Time (local time: hhmm) | 2330 |
| **ArrTime** | Actual Arrival Time (local time: hhmm) | 2340 |
| **ArrDelay** | The difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers. | -15, 20 |
| **ArrDelayNew** | The difference in minutes between scheduled and actual arrival time. Early arrivals set to 0. | 15 |
| **ArrDelay15** | Arrival Delay Indicator, 15 Minutes or More (1=Yes) | 1 |
| **Cancelled** | Cancelled Flight Indicator (1=Yes) | 0, 1 |
| **Diverted** | Diverted Flight Indicator (1=Yes) | 0, 1 |
| **Distance** | Distance between airports (miles) | 2300 |
| **CarrierDelay** | Carrier Delay, in Minutes | 15 |
| **WeatherDelay** | Weather Delay, in Minutes | 20 |
| **NASDelay** | National Air System Delay, in Minutes | 25 |
| **SecurityDelay** | Security Delay, in Minutes | 30 |
| **LateAircraftDelay** | Late Aircraft Delay, in Minutes | 45 |

**b.** **Statistical Summary**

Here is a statistical summary of our dependent variable, **DEP\_DELAY\_NEW.** The units are in minutes.

| **Min** | 0.00 |
| --- | --- |
| **Max** | 959.00 |
| **1st Quartile** | 0.00 |
| **3rd Quartile** | 7.00 |
| **Mean** | 0.00 |
| **Median** | 0.00 |

Correlation analysis of our dataset revealed the following trends:

****

Investigation into variables with high correlation resulted in these insights:

**DEP\_DELAY and DEP\_DELAY\_NEW:** These variables are correlated because DEP\_DELAY\_NEW is simply DEP\_DELAY, with flights arriving early normalized to 0. This variable artificially inflates our models accuracy, and is not known prior to flight time, so we removed it from our models.

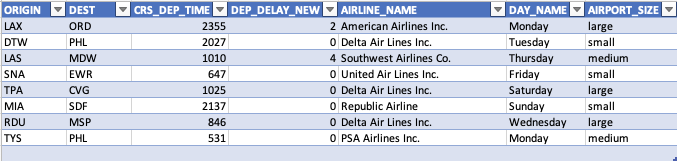
**ARR\_DELAY and ARR\_DELAY\_NEW:** These variables follow the same convention as DEP\_DELAY. We chose to remove ARR\_DELAY, because it is very similar and used to compute ARR\_DELAY\_NEW.

**CRS\_DEP\_TIME and DEP\_TIME:** These variables represent the scheduled departure time and actual departure time. They are both important, so we chose to keep both of these variables in our dataset.

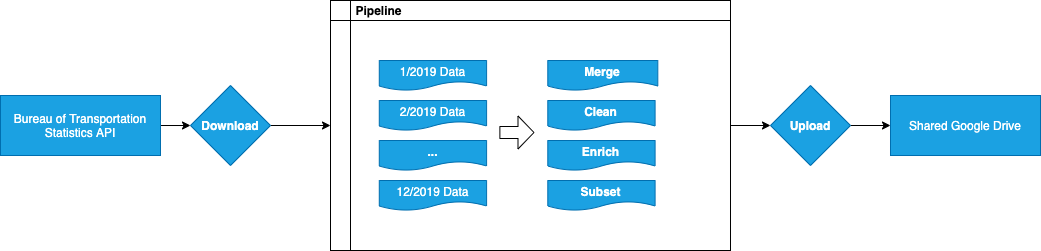
**CRS\_ARR\_TIME and ARR\_TIME:** These variables represent the scheduled and actual arrival times. We chose to keep both variables in our dataset.

**c. Data Snapshot**

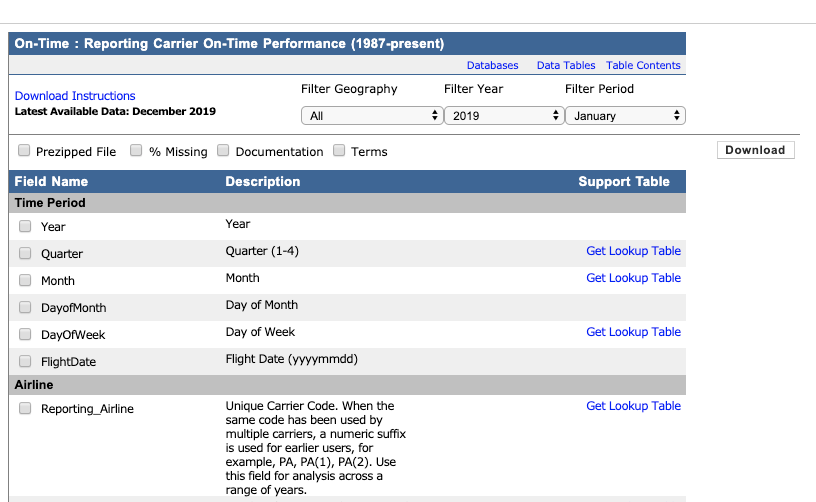
The following is a subset of our data and variables.

****

**d. Data Pipeline [0]**

****

[0] Some images did not fill well in this report due to scale. Lare versions of all images can be found in our code submission under the ‘pictures’ directory.

****

Our data source: the Bureau of Transportation Statistics Website

A problem we encountered early in our project was that it was very hard to add new variables to our data set. The BTS database had 107 unique data points which had to be checked one by one, and only allowed downloading a single month at a time. This meant that every time we wanted to test a new variable for a query or model, we had to:

1. Go to the BTS website and re-check all of the above boxes we were using previously
2. Select new variables
3. Download 12 months worth of data files, one by one
4. Merge these files into one CSV

This process was error prone, tedious, and made us hesitant to add new variables because of how long it took.

To solve this problem, we automated the entire process. The full code can be found under the **data\_pipeline** folder in our code submission. The general steps were as follows:

1. **Download**

Use cURL command line tool to download & unzip a set of BTS data files. The variables were parameterized, which allowed us to add new fields and re-generate on demand with new variables as we progressed through our project.

1. **Merge**

Using R, merge the data files resulting from step one into one large CSV. This script, and all remaining steps, were run automatically.

1. **Clean**

After running some sample queries, we found that there were many cancelled or diverted flights with blank values. There were flights with extreme delay times, sometimes over a day. We filtered these extreme outliers from our dataset in this step.

1. **Enrich**

This is where we added on features that we crafted as a part of data visualization and model building. The features that were added as a part of the enrich step can be seen in the ‘New Feature Creation’ section below. In general, this step merged external data sources to the BTS data, or derived new features from the existing data set.

1. **Subset**

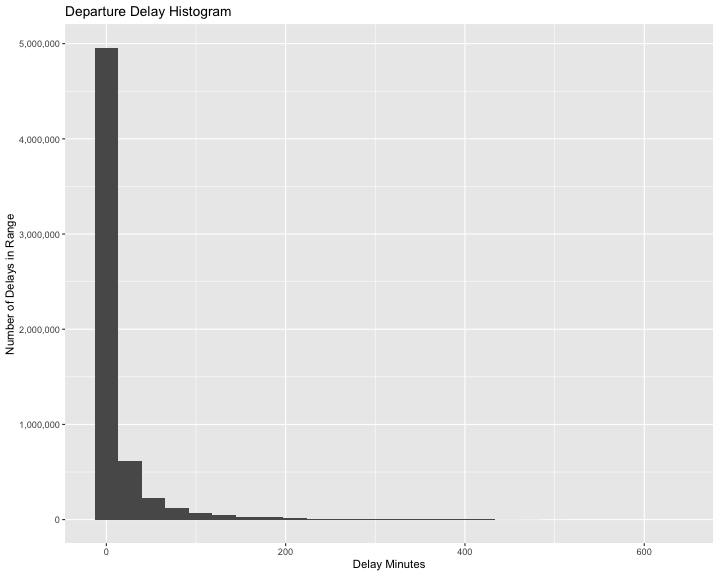
Our complete dataset included 6 million rows. To make development easier, the pipeline output multiple files: the complete dataset, and a development version. The development version included a random sampling of 1,000 rows from each month. We used a development version to quickly run queries and verify our coding syntax, then switched to the complete dataset once we verified our code was correct. This step also handled creating training / testing datasets for model building purposes.

**Table 1.2: Enriched Variables**

| **Name** | **Description** | **Example** |
| --- | --- | --- |
| **AIRLINE\_NAME** | Airline name in human readable format. Used in place of Airline ID for easier query readability | Southwest Airlines |
| **AIRPORT\_NAME** | Airport name in human readable format. Used in place of Airport ID for easier query readability. | O’Hare Airport |
| **MONTH\_NAME** | Month name in human readable format. BTS’s month field was in numerical format. This caused bugs in our prediction models as Month was incorrectly being interpreted as a continuous variable. | June, July |
| **DAY\_NAME** | Day name in human readable format. The reason for adding this variable is the same as the month name above. | Monday, Tuesday |
| **AIRPORT\_SIZE** | The size of the departing airport. Small, medium, or large, depending on how many flights the airport serviced.  Large means the total flights were above the IQR.  Medium means the total flights were within the IQR.  Small means the total flights were below the IQR. | Large, Medium, Small |
| **TOTAL\_AIRPORT\_FLIGHTS** | The total number of flights the airport had in our dataset. Used to test a theory that the busier an airport is, the higher the likelihood of delay. We found this to be an insignificant variable. | 12000 |
| **NORMALIZED\_ARR\_TIME** | If the scheduled arrival time of the aircraft is between 12am and 3am, this is 2400 + arrival time.  If not, it is scheduled arrival time | 1200, 2300, 2500 |
| **NORMALIZED\_DEP\_TIME** | The same as NORMALIZED\_ARR\_TIME, but for scheduled departure time | 1200, 2300, 2500 |

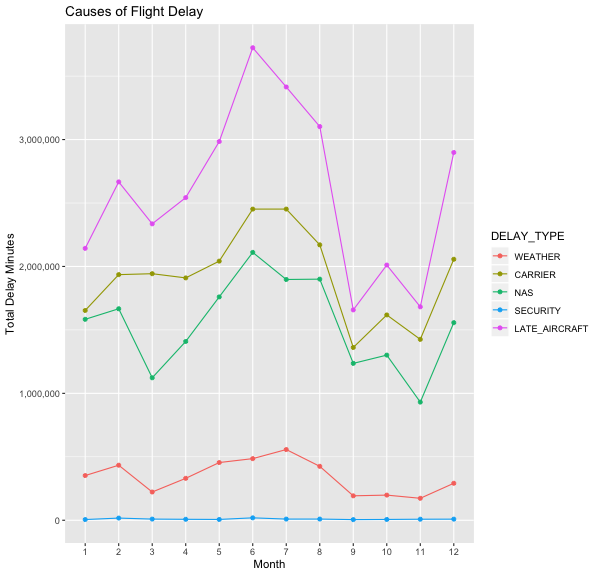
**4. Exploratory Analysis and Visualizations**

1. **Departure Delay Histogram**

****

To begin understanding our data, we created a histogram of departure delays. We expected that the data would show a strong right skew, with many delays at zero or slightly above zero. This histogram confirmed our hypothesis.

1. **Causes of Flight Delay Line Graph**

****

This line graph breaks down our dependent variable, departure delay, by two factors: delay type, and month. We wanted to answer two questions. What are the reasons flights are delayed? Is flight delay seasonal?

**Reasons for Delay**

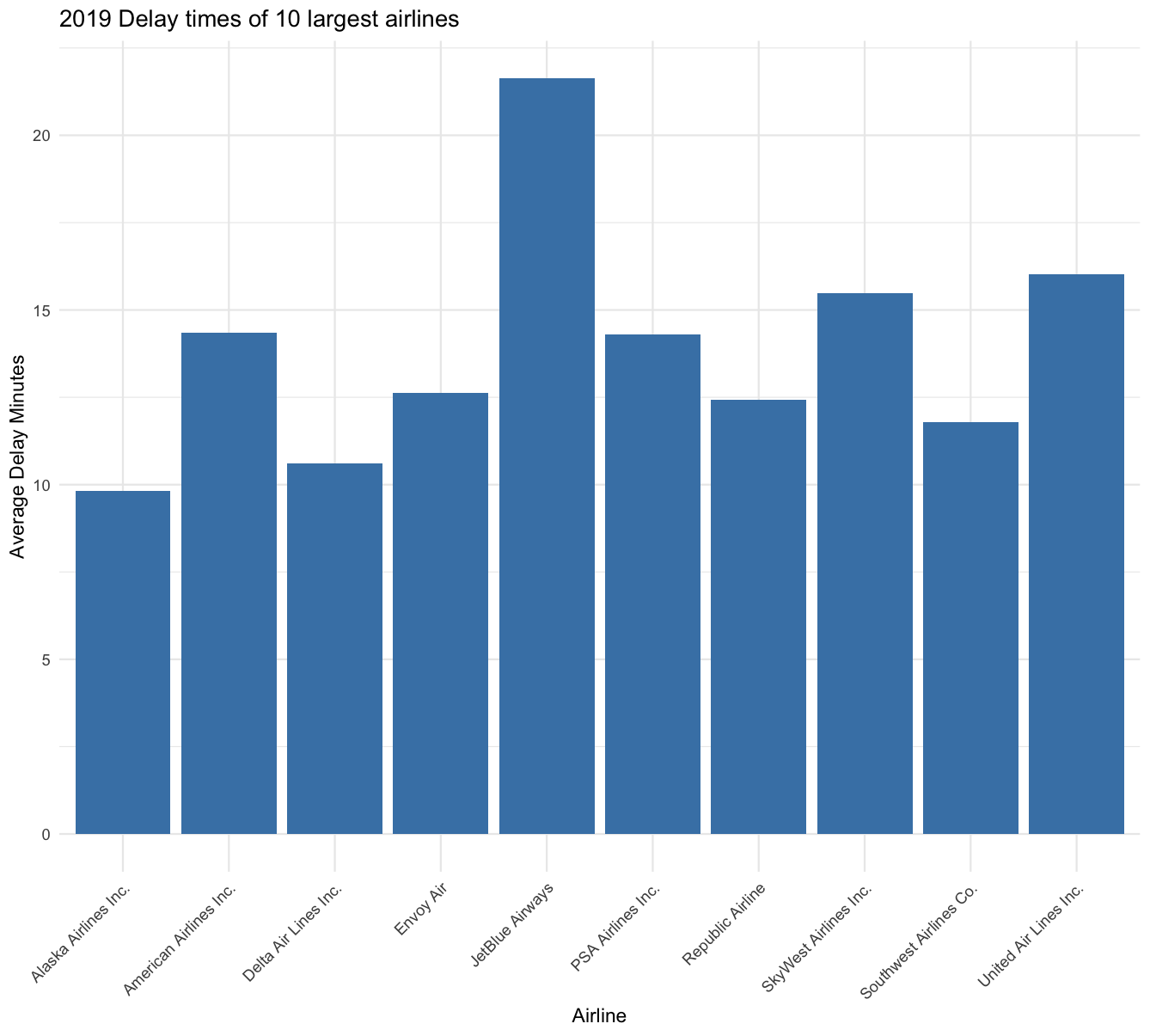
The BTS mandates that airlines attribute flight delay time to one or more categories: weather, carrier, NAS, security, and late aircraft. When we first created this graph, we were surprised to learn that late aircraft was the leading cause of delay, not weather, as we initially assumed. After seeing the visualization, we researched what the different variables represented on the BTS website. We learned that the “weather” delay type is only to be used for extreme, un-preventable cases such as a blizzard or hurricane. Delays caused by typical weather conditions, such as light rain or cloudy skies are classified under the NAS variable.

By far, the leading cause of delay was Late Aircraft. This raised an important question for our project: should we be concerned with the cause of the late aircraft? Ultimately we decided that we would not investigate, as this would require a complicated sequence of steps to merge the aircraft’s daily flight route into our dataset. It does leave room for future work.

**Seasonality**

We discovered flight delay displays some seasonality, but not in the way that we expected. Delay time decreased in the winter, and became higher in summer months. After further analysis, we discovered that the amount of flights increased in summer months, and decreased in winter months. This decreased flight volume is a likely cause of seasonality in flight delay.

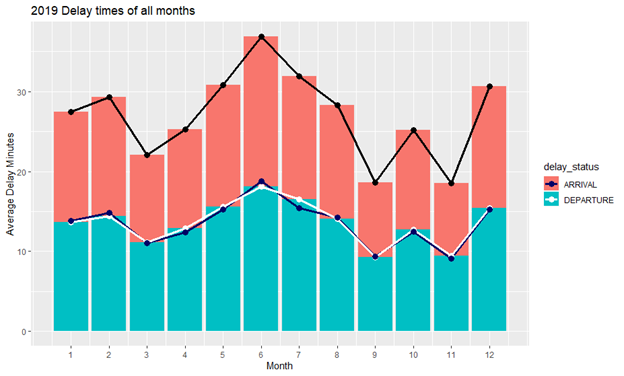
1. **Delay by Airline Bar Chart**

****

This barchart visualizes the average delay time by airline. We can see that JetBlue airlines has the worst delay time, Alaska Airlines has the best, and all other airlines fall somewhere in the middle.

One thing we found interesting about this graph is that while it appears that airline has a significant distance on flight delay time, our linear regression models found Airline to be an insignificant variable.

1. **Delay by Month Bar Chart and Line Graph**



The graph above consists of a bar chart which represents the total average of arrival and departure delays. The bar chart in blue represented the departure delay while the red represented the arrival delay. For the line graph, the white graph visualizes the mean point of departure delays while the dark blue line graph shows the mean point of arrival delays followed by black line graph which presents the mean of total delay.

As we notice, the month of June is the worst month to fly as it has the most delays for both arrival and departure delays. Whereas for the best month to fly, it is hard to distinguish between September and November. Thus, the calculation figure is taken into consideration and the result obtained is November to be the best month to fly.

**Model Building**

**Baseline Model**

We built our baseline model by the following process:

1. Compute the mean DEP\_DELAY\_NEW of the training dataset
2. Apply the mean to each row in the test set
3. Compute the RMSE using the resulting actual vs. predicted values

Once we had a baseline model established, we moved onto more advanced modelling techniques. The baseline model was important to us because we were able to use it to evaluate how our new models performed: if they had a higher RMSE than the baseline model, we knew it was performing poorly.

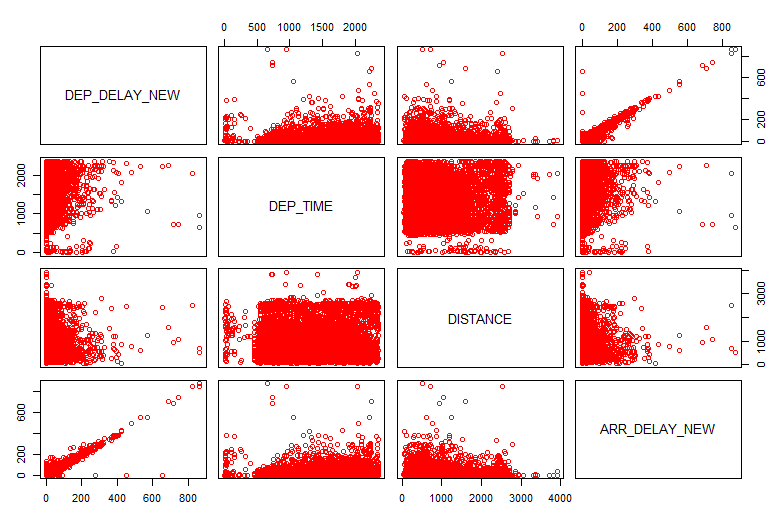
The resulting RMSE of our baseline model was 42.40 minutes.

**Linear Regression**

We followed this process for building linear regression models:

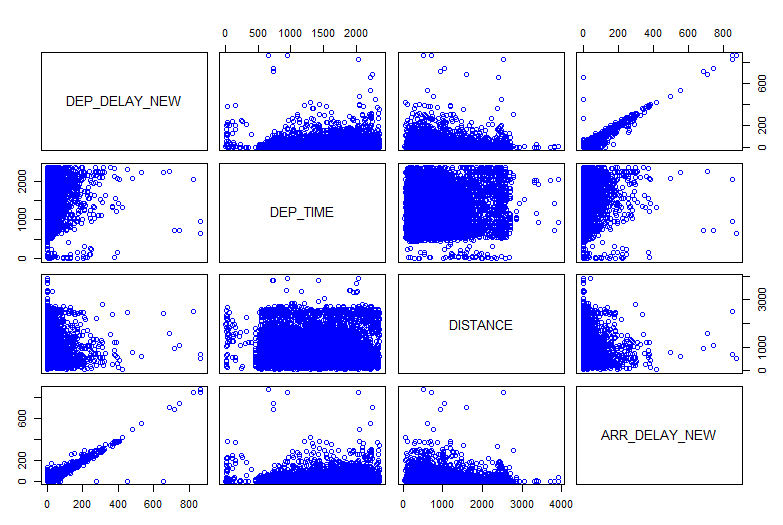
1. Perform correlation analysis on the input data set and make sure to only use one for regression analysis. We found correlated variables in our enriched feature set:
   1. Airline code to Airline name
   2. Airport code to Airport name
   3. Day code to Day name
   4. Month code to month name
2. Run the linear regression creation function on the training data
3. Inspect the resulting coefficients. Remove variables that are not statistically significant, and repeat step 3 after each removal
4. Once all of the remaining variables are statistically significant, run the linear regression model against the test set to create predictions
5. Compute the RMSE of the predictions against the test set
6. Attempt to improve RMSE by adding additional features

**Multiple Linear Regression Model :** R2 = 90%, RMSE = 12.65



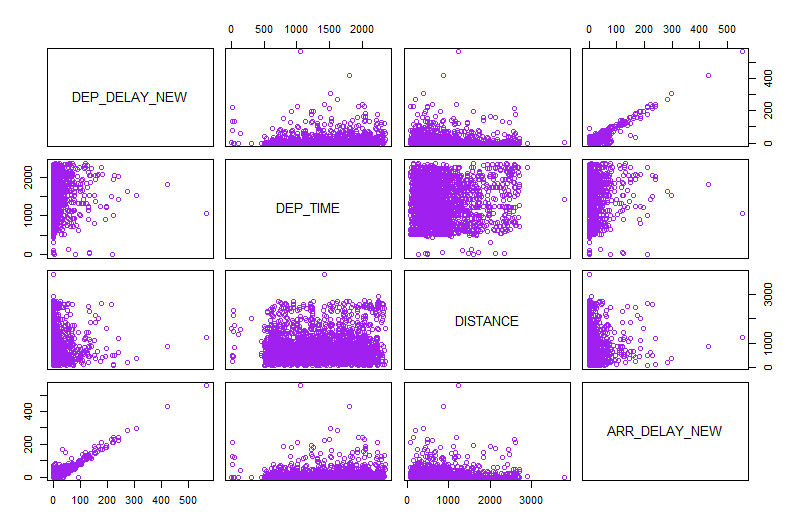
**Figure 1**

**Train Model** R2 = 90%, RMSE 13.33



**Figure 2**

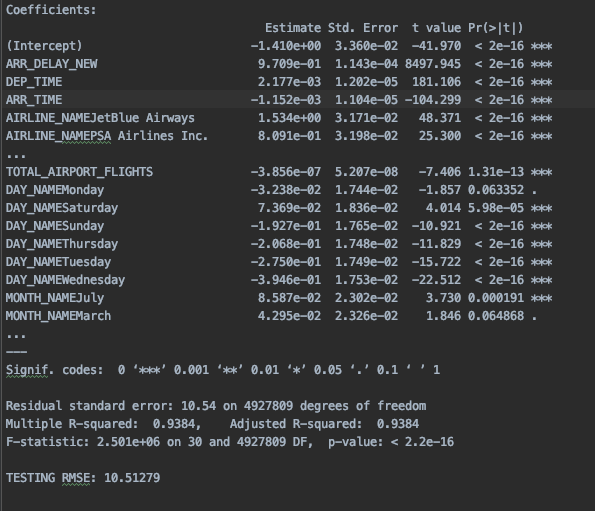
**Test Model** : R2 = 90%, RMSE = 50.56488



**Figure 3**

**Model: Linear Regression with Known Arrival Delay**

By following the above process, we found these results for our first linear regression model using ARR\_DELAY\_NEW, DEP\_TIME, ARR\_TIME, AIRLINE\_NAME, MONTH\_NAME, DAY\_NAME, and TOTAL\_AIRPORT\_FLIGHTS.



**Figure 4**

From this model, we can make some interesting observations about what leads to airline delay time. Arrival delay is predicted to be higher for March and July. Total airport flights shows a negative relationship with departure delay; the more flights an airport serves, the less delay time we can expect. If we want to avoid a delay, we should also avoid JetBlue and PSA airlines.

This model performed well, but it had a problem. The ARR\_DELAY variable represents how late the departing flight is to its destination. For our model to be useful in real world applications, the arrival delay of the flight is not something we can predict before a flight happens. To make our model more useful in a real world context, we attempted to create a model using only variables that are known prior to a flight.

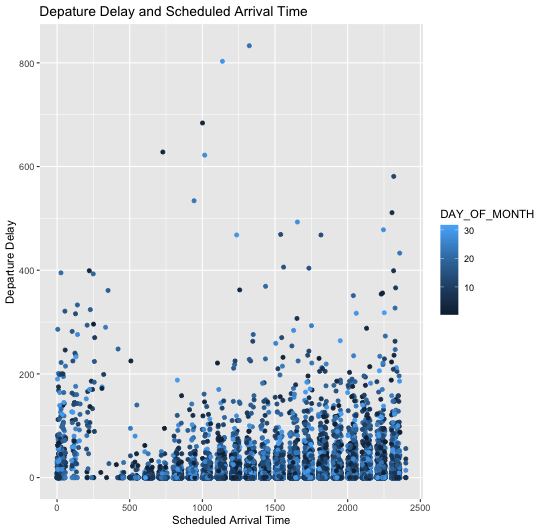
**Linear Regression with unknown arrival delay**

**Figure 5**

When we first tried creating a linear regression model with unknown arrival delay, we were unable to achieve an RMSE that was better than the baseline model.

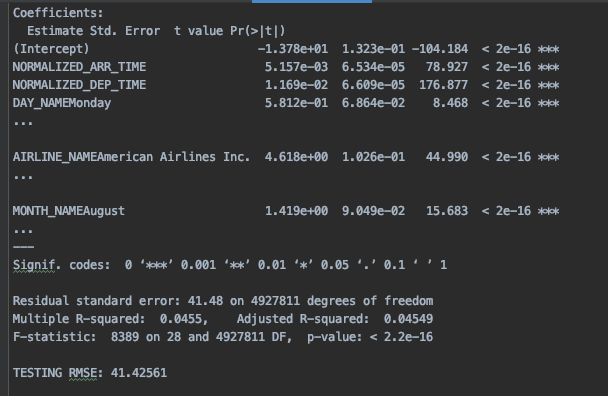
After multiple iteration attempts, we realized that scheduled ARR\_TIME and scheduled DEP\_TIME were our best predictor variables. These variables both showed a positive correlation with DEP\_DELAY\_NEW.

ARR\_TIME and DEP\_TIME were both correlated with an increased DEP\_DELAY\_NEW as the day went on. However, there was a large cluster of flights between 12am and 3am that had a large departure delay, as seen in figure 6:



**Figure 6**

To solve this problem, we created a new feature. We took flights that were scheduled to arrive between 12am and 3am, and added their arrival time to 2400. For example, a flight scheduled to land at 1am would become 2500. We did this to make our data more linear, and useful to our model. This improved our results and beat the baseline model, as shown in figure 7. Input variables to this model are NORMALIZED\_ARR\_TIME, NORMALIZED\_DEP\_TIME, DAY\_NAME, AIRLINE\_NAME, and MONTH\_NAME.



**Figure 7**

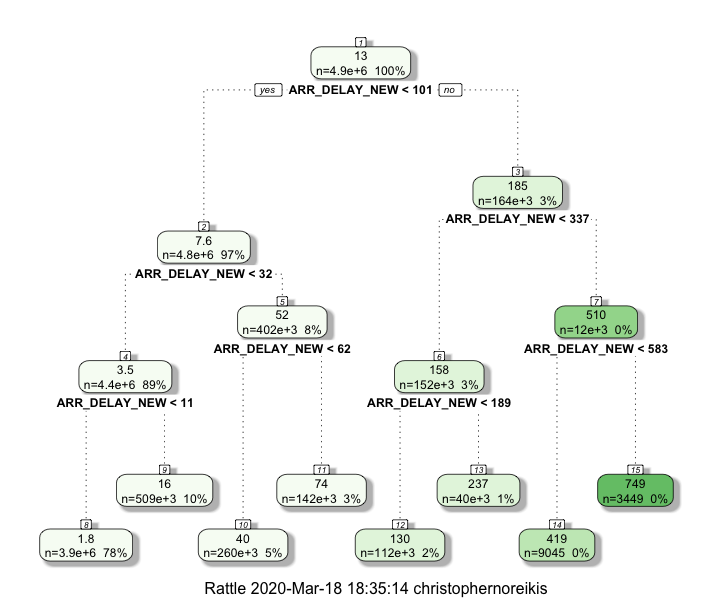
**Regression Decision Trees**

After learning about classification decision trees in class, we attempted to apply it to our project. Classification decision trees work by creating a sequence of nodes and leaves. Each node is a ‘split’, or decision to traverse left or right in the tree. This process is repeated until a leaf is reached. The leaf is a final prediction to make, which in the case of classification, is a categorical variable.

The performance of decision tree nodes are measured by purity and information gain. Nodes that are more pure are stronger predictors of the final value and give more information gain. Nodes with more impurity result in less information gain.

Regression decision trees are very similar to classification trees. The only difference is that the final prediction is not a categorical variable; it is a number. This resulting number is the prediction for observations in the testing set. Our first attempt at building a model using all variables and rpart’s ‘anova’ method can be seen in figure 1:

**Regression Tree with known arrival delay:**

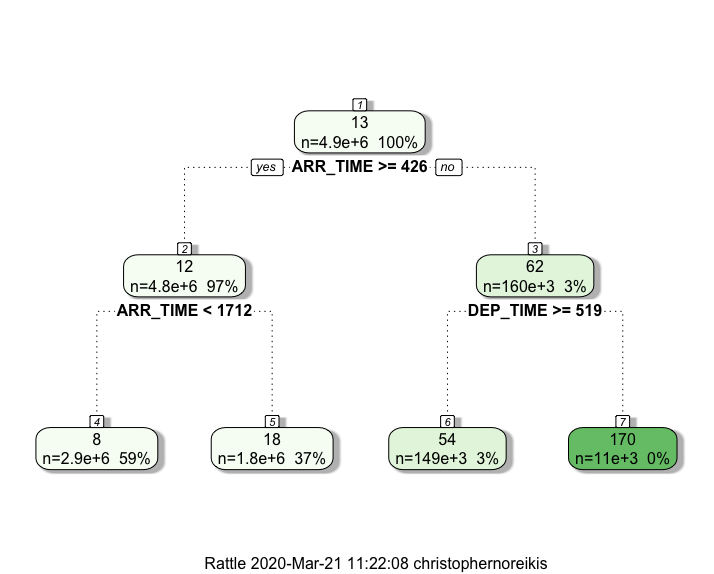
****

**Figure 1**

**Resulting RMSE: 12.87**

The rpart library chose only ARR\_DELAY\_NEW as a predictor variable. This is a very accurate predictor variable, however, it suffers from the same problem discussed in our linear regression findings. A flight’s arrival delay is a strong predictor, but we want to attempt to predict flight delay using variables known prior to the flight.

**Regression tree with unknown arrival delay:**



**Figure 2**

**Resulting RMSE: 40.79**

With variables that would be unknown in a real world situation removed, our model performed much worse than the model with known arrival delay. The updated model shows scheduled ARR\_TIME and scheduled DEP\_TIME significant variables. This tree can be interpreted to mean that if the flight is scheduled to arrive after 4:26 and land before 17:12, our model predicts 18 minutes of delay, otherwise, we predict 8 minutes of delay.

**Model results**

| **Model** | **Root Mean Squared Error** |
| --- | --- |
| Baseline | 42.40 |
| Decision Tree with unknown arrival delay | 40.79 |
| Decision Tree with known arrival delay | 12.87 |
| Linear Regression with unknown arrival delay | 41.42 |
| Linear Regression with known arrival delay | 10.51 |

Our modelling showed that linear regression performed best when arrival delay is known, and the regression decision tree performed best with unknown arrival times. Our linear regression model showed many variables that were statistically significant in evaluating departure delay. However, when applied to a large enough dataset, they had a very small coefficient, which made future predictions using these variables in-effective.

We found linear regression to be the harder of the two models to implement, as it involved correlation analysis and many iterations to see how variables interacted with each other. The rpart ‘anova’ method did most of this work for us.

The main conclusion drawn from our models is that to avoid flight delay, fly out as early as you can. Even if the flight is at 12pm, expected delay increases linearly as the day goes on.

**Results and Conclusions**

Our results confirmed some of our expectations regarding flight delay, and also offered some surprising insights. The first graph which shows the departure delay histogram did confirm our hypothesis delay minutes show a strong right skew where most of the delays are zero or under 15 minutes.

The graph that represents the delays by month surprised us. We expected winter months to show the highest delay due to weather conditions and holidays; however, summer months historically show the highest delay. This is likely due to increased flight volume.

The graph comparing airlines had expected and unexpected results. Alaska airlines, which had very good reviews online, also was a leader in low delay time. However, the airline with the highest delay time, JetBlue, was unexpected, as they are one of the highest rated airlines in the industry.

We learned many lessons when creating models to predict departure delay. Always inspect your input variables closely, as an independent variable may be derived from your dependent variable and make your model seem better than it is. Ensure you have an easy way to add new features to your data set. Ensure you have a training / testing set to detect overfitting. Finally, ensure you have a baseline model to establish benchmarks on your models and determine how well they are actually performing.

Lastly, if we were given more time to work on this data, we would investigate the cause of late aircraft delay which appears to have a large impact on departure delay based on visualization techniques. We believe drilling down to the root cause of incoming late aircraft might help in handling delays effectively. Besides that, we hope our project is helpful for consumers and airlines in understanding and drawing insights about the cause of flight delay.