Holly Ghaemi, Johan Norvik, Andre Nguyen

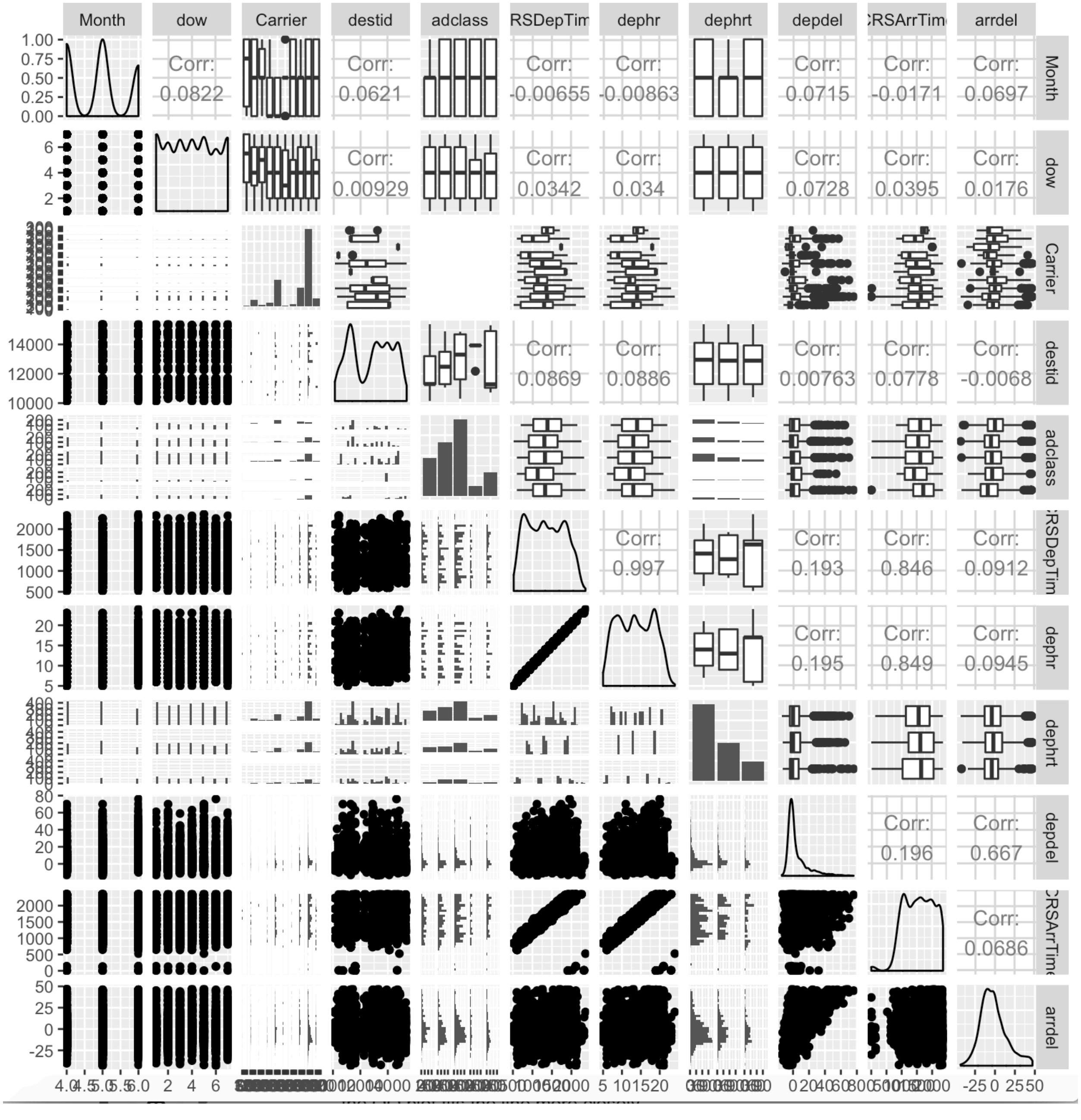
Building a Model for Airport Arrival Data Set

Section 1: Data Description

1. The original data set contains over 1 million data points with 21 variables related to airport operations. The response variable of interest is arrival delays and we were interested in building a model from the data that would be useful in predicting arrival delays based on the other variables, which have information on time and of departures and arrivals, airline carriers, origin and destination airports, and departure delays. We subset the data to only include flights departing from George Bush Intercontinental Airport (IAH) and this left us with 29,070 data points. We were looking to determine which variables had the most significant impact on arrival delays and also answering a few research **questions** such as which variables best explain the model? Which model best explains the data? How good is the fit of our data?
2. Data source: Microsoft Azure Machine Learning Studio—"Flight Delays Data." Link: https://1drv.ms/u/s!Aps4xVzIwpewlP8a_QB21WmsXy8thA
3. We simplified the data set by discretizing numeric data on departure time and day of week and groupings to reduce the levels of factor data on destination airports. The departure times were discretized into three levels based on relative volume of departures and the destination airports were classified into five groups based on average arrival delay. Based on this simplification and domain knowledge, we narrowed down the variables to the following variables of interest: Month, day of week (dow starting with 1 as Monday), Carrier (10 airlines), destination airport (destid), five levels of average arrival delay into airports (adclass with levels being very low, low, medium high and very high), flight volume at departure times (dephrt: trough, medium, peak), the departure time (CRSDepTime), departure hour (dephr), departure delay (depdel) and arrival time (CRSArrTime.)

Section 2: Model Building

First, we removed the outlier values in the response variable, arrival delay. This reduced the 29,070 data points to 26,430 data points. Then we created a correlation plot to visualize the relationships:



We found the following significant correlations with arrdel: month = 0.040, dow = -0.0204, dephr = 0.0569. We were going to choose the 3 variables with highest absolute correlation values to use in our model but we replace dephr with dephrt which is a factor variable assuming 3 values (trough, medium, peak based on volume of flights during that hour) because it's not possible to fit arrival delays linear function to the hour values ranging from 1-24. Then we added arrival airport class and carrier because they were variables of interest. **Our final variables were: month, day of week, departure hour type, airport delay class and carrier.**

Depdel, dephr, crsarrtime, dephrt had high correlations because they were variables based on departure delays and arrival times but since departure times and departure delays are not know in advance, they don't help in timely predictions. We also took a closer look at the distribution of arrival delays and noticed that it has an almost normal/Gaussian distribution except for a heavy right tail (see the graph above) with mean lying a bit lower than 0. There are no linear or quadratic relations between arrival delay and the other variables seen from the correlation plots so we did not move forward with establishing quadratic variables within our model.

We created 4 different linear models for comparison: full model (**lm.full**), an untransformed model with 5 variables (**lm.one**), a model with 5 variables where the response was square root transformed (**lm.square.root**) and an interaction model between month and day of week (**lm.interaction**).

The following are the equations for the four models:

Variable definitions:
m_i= indicator variable for month. For i= 0, 1, 2 (only three months in the data set)
a_i= indicator variable for airline. For i = 0, 1, …,9
d_i= indicator variable for day of week. For i = 0, 1, … 6
h_i= indicator variable for hour of day classified by volume of flights during that hour (medium, peak, trough). For i = 0, 1, 2
g_i = indicator variable for destination airport grouped by average delay (very low, low, medium, high, very high). For i = 0, 1, …, 4
t_i= departure time.
r_i= departure hour.
l_i= departure delay.
c_i= arrival time.
e_i= destination ID or i=0,1,..,59

lm.one: $Y = \beta_0 + \sum \rho_\iota m_i + \sum \varphi_\iota a_i + \sum \omega_\iota d_i + \sum \gamma_\iota h_i + \sum \rho_\iota g_i + \varepsilon_\iota$

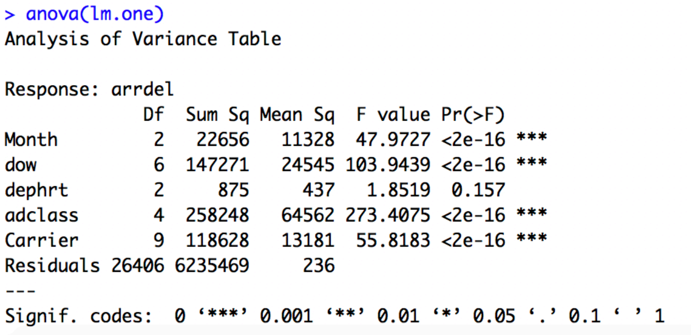
lm.square.root: $\sqrt{Y} = \beta_0 + \sum \rho_\iota m_i + \sum \varphi_\iota a_i + \sum \omega_\iota d_i + \sum \gamma_\iota h_i + \sum \rho_\iota g_i + \varepsilon_\iota$

lm.full: $Y = \beta_0 + \sum \rho_\iota m_i + \sum \varphi_\iota a_i + \sum \omega_\iota d_i + \sum \gamma_\iota h_i + \sum \rho_\iota g_i + \mu_\iota t_i + \tau_\iota r_i + \pi_\iota l_i + \theta_\iota c_i + \sum \alpha_\iota e_i + \varepsilon_\iota$

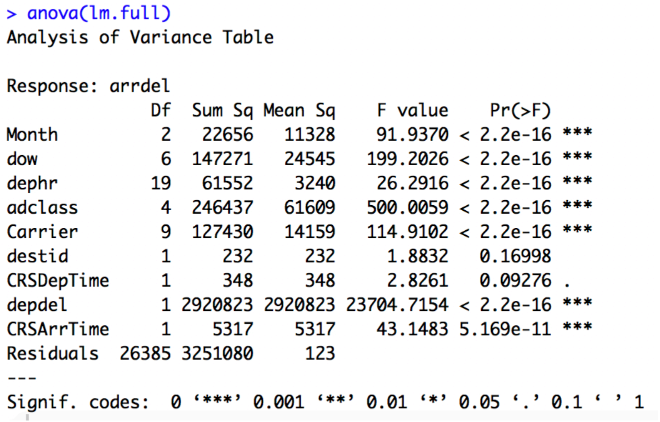
lm.interaction: $Y = \beta_0 + \sum \rho_\iota m_i + \sum \varphi_\iota a_i + \sum \omega_\iota d_i + \sum \gamma_\iota h_i + \sum \rho_\iota g_i + \sum\sum \mathrm{C}_{i,j} m_\iota d_i + \varepsilon_\iota$

The summaries of each model showed a significant result for the F-Test with p-values smaller than the alpha level of 0.05.

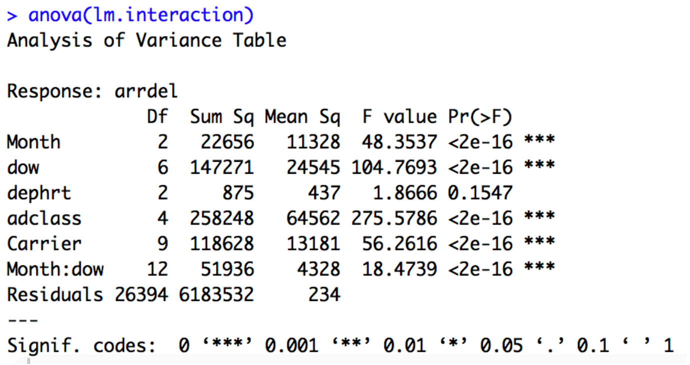
The following ANOVA tables for each model show that almost all the factor variables are statistically significant according to their p-values and F-statistics.



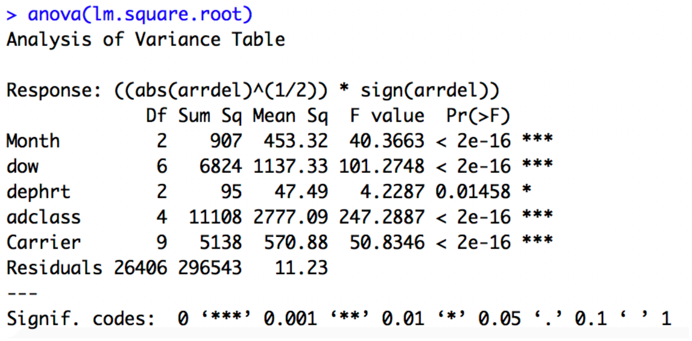
```
> anova(lm.one)
Analysis of Variance Table

Response: arrdel
             Df  Sum Sq Mean Sq  F value Pr(>F)    
Month         2   22656   11328  47.9727 <2e-16 ***
dow           6  147271   24545 103.9439 <2e-16 ***
dephrt        2     875     437   1.8519  0.157    
adclass       4  258248   64562 273.4075 <2e-16 ***
Carrier       9  118628   13181  55.8183 <2e-16 ***
Residuals 26406 6235469     236                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



```
> anova(lm.full)
Analysis of Variance Table

Response: arrdel
              Df  Sum Sq Mean Sq    F value    Pr(>F)    
Month          2   22656   11328    91.9370 < 2.2e-16 ***
dow            6  147271   24545   199.2026 < 2.2e-16 ***
dephr         19   61552    3240    26.2916 < 2.2e-16 ***
adclass        4  246437   61609   500.0059 < 2.2e-16 ***
Carrier        9  127430   14159   114.9102 < 2.2e-16 ***
destid         1     232     232     1.8832   0.16998    
CRSDepTime     1     348     348     2.8261   0.09276 .  
depdel         1 2920823 2920823 23704.7154 < 2.2e-16 ***
CRSArrTime     1    5317    5317    43.1483 5.169e-11 ***
Residuals  26385 3251080     123                         
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

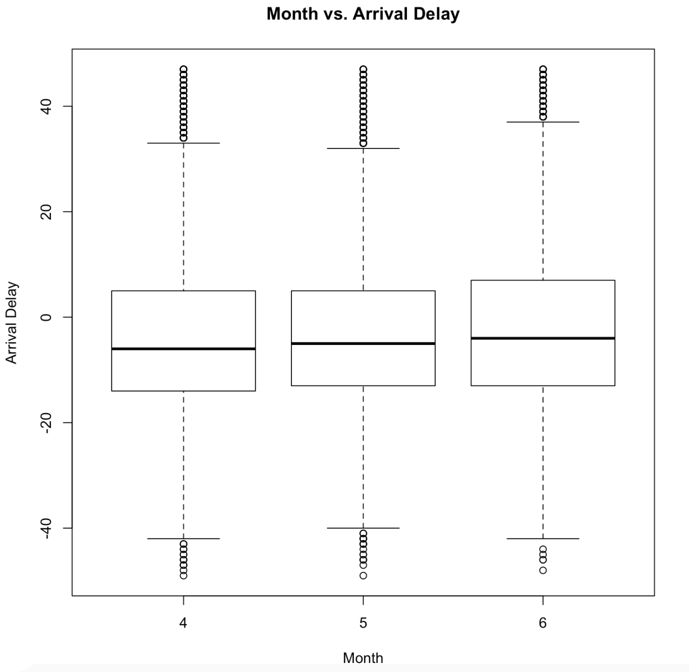


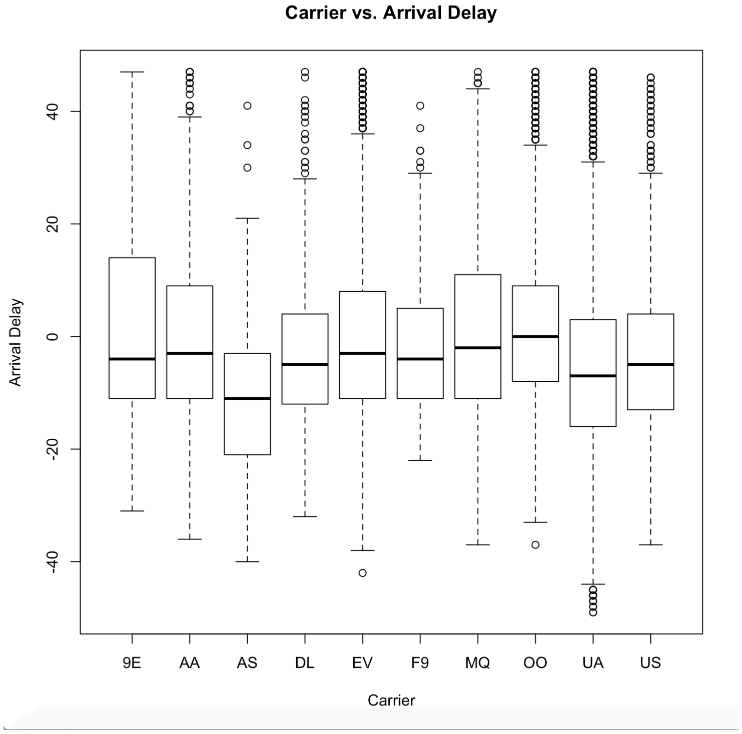
```
> anova(lm.interaction)
Analysis of Variance Table

Response: arrdel
             Df  Sum Sq Mean Sq  F value Pr(>F)    
Month         2   22656   11328  48.3537 <2e-16 ***
dow           6  147271   24545 104.7693 <2e-16 ***
dephrt        2     875     437   1.8666 0.1547    
adclass       4  258248   64562 275.5786 <2e-16 ***
Carrier       9  118628   13181  56.2616 <2e-16 ***
Month:dow    12   51936    4328  18.4739 <2e-16 ***
Residuals 26394 6183532     234                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

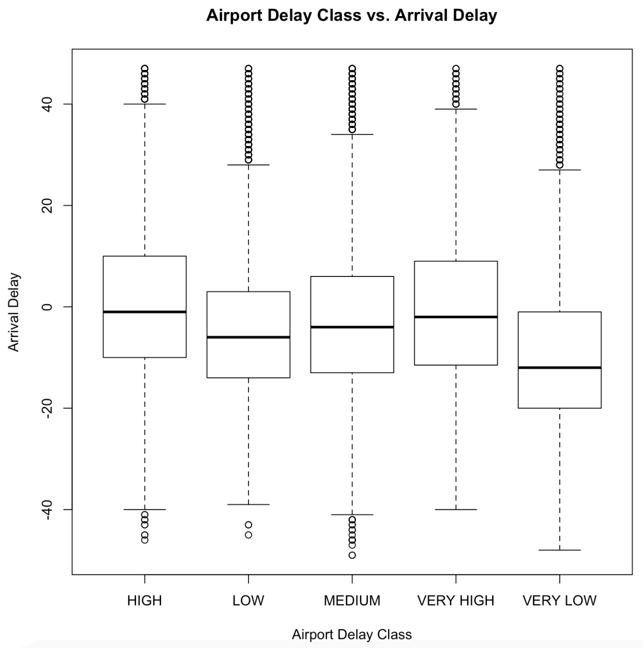


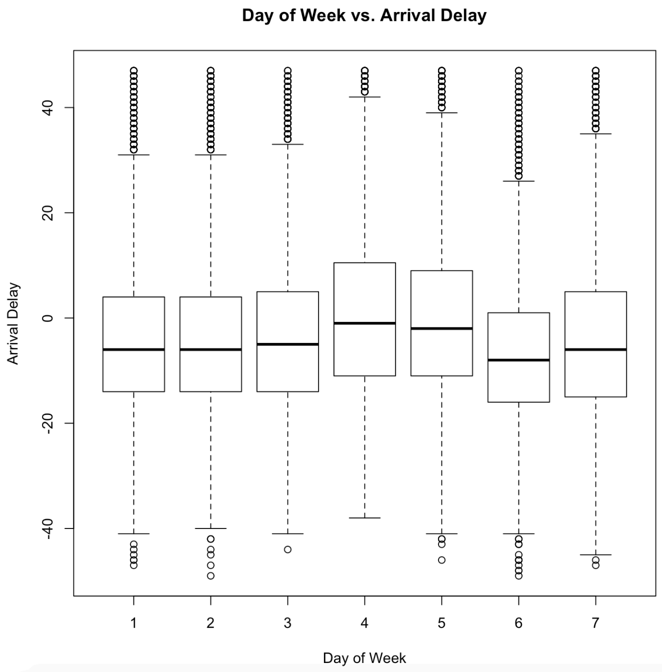
```
> anova(lm.square.root)
Analysis of Variance Table

Response: ((abs(arrdel)^(1/2)) * sign(arrdel))
             Df Sum Sq Mean Sq  F value  Pr(>F)    
Month         2    907  453.32  40.3663 < 2e-16 ***
dow           6   6824 1137.33 101.2748 < 2e-16 ***
dephrt        2     95   47.49   4.2287 0.01458 *  
adclass       4  11108 2777.09 247.2887 < 2e-16 ***
Carrier       9   5138  570.88  50.8346 < 2e-16 ***
Residuals 26406 296543   11.23                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

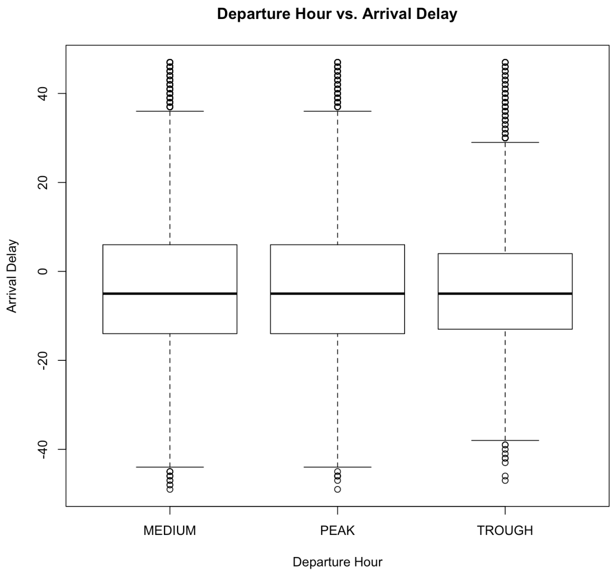
We further scrutinized the variables for quadratic relationships and normal boxplots as well by creating the following plots for each variable versus the response variable.





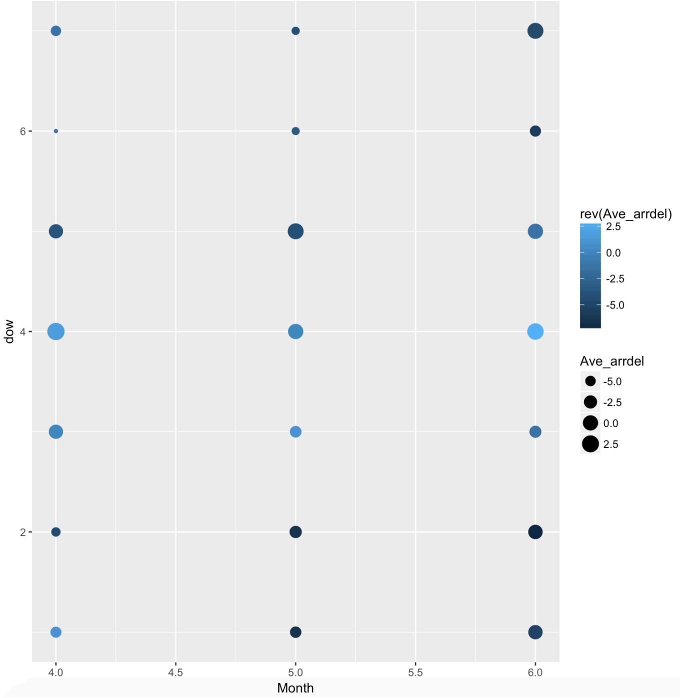

Airport Delay Class vs. Arrival Delay
Arrival Delay
40
20
0
-20
-40
HIGH
LOW
MEDIUM
VERY HIGH
VERY LOW
Airport Delay Class


Day of Week vs. Arrival Delay
Arrival Delay
40
20
0
-20
-40
1
2
3
4
5
6
7
Day of Week



Many outliers were seen in a majority of the boxplots, therefore we cannot assume normality within those factored groups.

We built the interaction model by first creating the following plot to gauge whether or not there is an interaction between the two variables month and day of week.



The size of the dots indicate the average of the arrival delay response variable as well as the color. The larger and darker the dot, the higher the average of the arrival delay for that specific month and day of week. We can see on (4,4), (5,5) and (6, 7) the dots are larger and darker therefore indicating a possible interaction between month and day of week. This is why we created the interaction model.

The following are the adjusted R^2 values for each of the models:

lm.full = 0.5199
lm.one = 0.07994
lm.square.root = 0.07427
lm.interaction = 0.08719

The full model has the largest adjusted R^2 value but its usefulness is limited because it includes departure delay which will not be known until the actual departure time Depdel, dephr, crsarrtime, dephrt had high correlations because they were variables based on departure delays and arrival times but since departure times and departure delays are not know in advance, they don't help in timely predictions. Removing that model, the second highest adjusted R^2 value is the best at explaining the variation of arrival delays and it belongs to the **interaction model**. Therefore, 8.719% of the variance in the response variable is accounted for by the predictor variables of this model. However, despite being the 2nd highest adjusted R^2 value, it is important to note that it is still a very low value. We note again that all models had very significant values for the F statistic, so we have a case where there is strong evidence for a statistical relationship between the response and predictor variables, but one that explains little of the variance of the response

We then used the bidirectional step() function for each of the models and the backwards step() function for the full model to decide whether or not we should drop variables from each model and also to cipher which model maintained the lowest AIC value. We received the following AIC values:

lm.full = 127,277.6
lm.one = 144,279.1
lm.square.root: = 63,947.61
lm.interaction = 144,081

Because the full model includes variables depdel, dephr, crsarrtime and dehprt which have innate insignificant correlations, we do not use this model and we can see how it has strongly affected the AIC value for the full model to keep these variables. We therefore use the lm.interaction model because it has the second lowest AIC out of the four models and of course the lm.square.root model because it has the lowest AIC.

Section 3: Model Diagnostics: Interaction Model and Square Root Transformed Model

Interaction Model

We ran a F-test for our final model, the interaction model. The assumptions for this test are the following 1.) normality 2.) independence 3.) constant variance

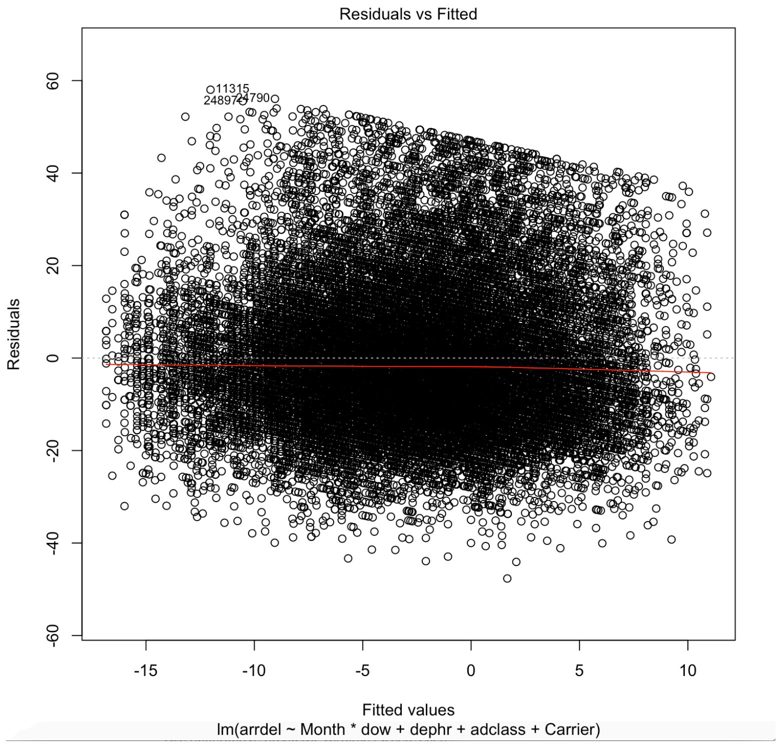
Hypotheses:

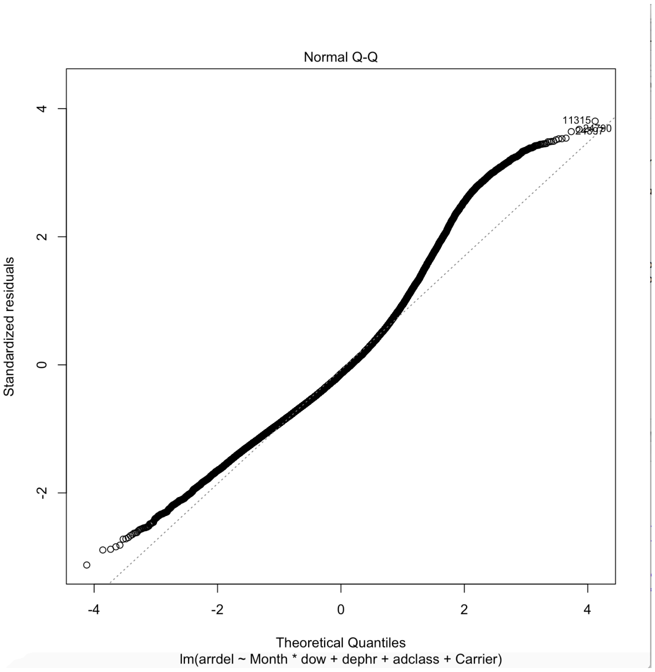
$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

H_a: Betas are not all zeros

The F-stat is 80.73 on 34 and 26395 degrees of freedom with a p-value that is < 2.2e-16. Therefore, we can conclude that we can reject H_0 in favor of H_a stating that the model is significant and is a good fit for the data.

The following residual plot and normal Q-Q plot was also obtained for the **interaction model**:





The residual plot shows a relatively random distribution and an almost horizontal line, but the variance may be decreasing as the fitted values increase. This indicates close to normality for the model but a power transformation may help stabilize the variance. Additionally, the Q-Q plot lies on the line in the middle but digresses from the reference line at the ends, which is in line with the comment made earlier about the heavy right tail of the arrival delays histogram. We cannot assume normality with confidence from the QQ plot but because of the residual plot, we will go ahead and assume relative normality.

Square Root Transformed Model

We ran a F-test for our second final model, the **square root transformed model**. The assumptions for this test are the following 1.) normality 2.) independence 3.) constant variance

Hypotheses:

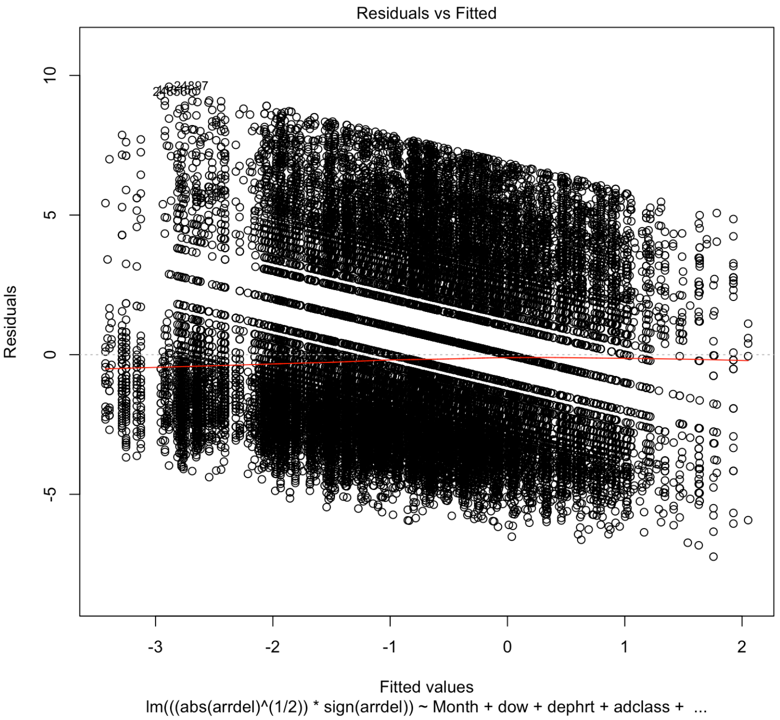
$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

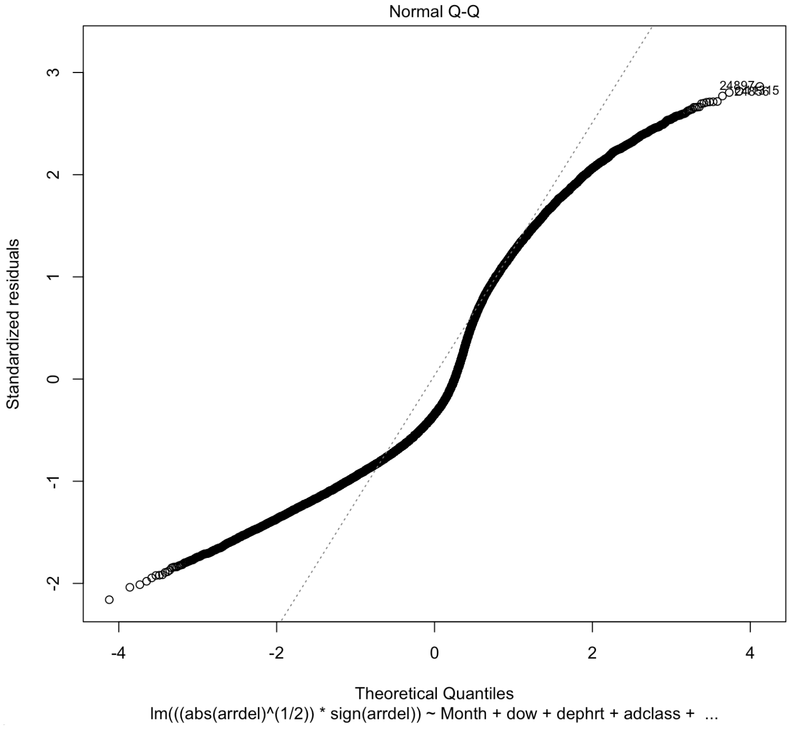
H_a: Betas are not all zeros

The F-stat is 93.2 on 34 and 26395 degrees of freedom with a p-value that is < 2.2e-16. Therefore, we can conclude that we can reject H_0 in favor of H_a stating that the model is significant and is a good fit for the data.

The following residual plot and normal Q-Q plot was also obtained for the **square root transformed model**:

The following residual plot and normal Q-Q plot was obtained for the **square root transformed model**:





The residual plot has a clear downwards trend but the unequal variance that we had in the interaction model has gone away with this power transformation . The interpretation would be that the model overestimates large arrival delays and underestimates early arrivals and short delays. A possible improvement is to try a different power transformation such x^(1/3) or something between the square and the cube root. Another conjecture is that the dynamics of delays are different for short and long delays and that fitting separate models to each of the two classes of delays may provide better results. The normal Q-Q plot deviates greatly from the reference line which stands for the CDF of the normal variable. Therefore, we cannot confirm normality from these plots.

In conclusion, we were not able to choose one model that explains the arrival delay response because despite the significance of the coefficients of the predictor variables, they were unable to explain a large amount of the variance in the model as shown by the small values of R^2. Also, one of our models had a better AIC (square root) while the other model (interaction) had residual and Q-Q plots that exuded a greater normality. And we are not using the full model despite the low AIC value and higher R^2 because it is not a useful predictor because it depends greatly on departure delay, which is never known in advance and to the insignificant correlation variables. Lastly, we conclude that the type of departure time (Trough, Medium, Peak) doesn't contribute significantly as a predictor variable and we could eliminate it.