Predicting Flight Delays

Aastha Jairath, Elizabeth Pfeiffer and Maizie Seidl

Purdue University

INTRODUCTION

It’s a bird, it’s a plane, it’s a… delayed flight?! Your nearest airport may be silent now, but on a typical day, it bustles nonstop, working to get nervous and experienced fliers alike to their destinations on time. In 2019 alone, the United States flight industry served approximately 811 million people, according to the U.S. Bureau of Transportation. Unfortunately, a system of such massive proportions is sure to include several flight delays and cancellations. Altered flight plans can be a nuisance, whether you fly rarely or frequently. Sometimes, we can see these delays coming, other times they are caused by unavoidable or unforeseen circumstances. What if we could use historical data to determine the signs and predict the duration of a potential delay before it occurs? Doing so would certainly help passengers better prepare themselves for a late departure and, consequently, a late arrival to their destination well ahead of their scheduled flight. Keeping time constraints and convenience in mind, the goal of our project is to predict delayed flight arrivals on the basis of several relevant factors.

The dataset upon which we have based our analysis was collected by the United States Bureau of Transportation Statistics and published by Federal Aviation Administration in a release entitled “On-Time Performance: U.S. Domestic Airlines,” (January, 2020). This dataset contains statistics pertaining to the performance of every domestic flight originating from and landing in the United States in January, 2020. After performing some exploratory analysis, removing irrelevant attributes, and making appropriate transformations, we succeeded in fitting a linear regression model on our response variable (delay in arrival time), taking several other variables (delay of departure, flight time, origin airport, distance, and others) as the predictors.

METHOD

Exploratory Analysis

The original dataset contained 607,346 rows (flights) and 12 attributes. The names and description of these attributes are listed as follows:

1. FLY_DATE: qualitative, date of flight in "m/d/y" format, including all days in the month of January, 1-31
2. OP_UNIQUE_CARRIER: qualitative, unique codes referring to the airline carrier name. Numeric suffix used to distinguish code shared by more than one carrier
3. ORIGIN: qualitative, IATA code of departure airport
4. DEST: qualitative, IATA code of arrival airport
5. ORIGIN_CITY_NAME: qualitative, city of flight departure
6. DEST_CITY_NAME: qualitative, city of flight arrival
7. DEP_DELAY: quantitative discrete variable, difference *in minutes* between scheduled and actual departure time. Early takeoff indicated by negative value.
8. ARR_DELAY: quantitative discrete, difference *in minutes* between scheduled and actual arrival time. Early arrival indicated by negative value. **This is our target variable.**
9. CANCELLED: binary indicator, cancelled flight (1) or non-cancelled flight (0)
10. DIVERTED: binary indicator, diverted flight (1) or non-diverted flight (0)
11. AIR_TIME: quantitative discrete, duration of flight *in minutes*
12. DISTANCE: quantitative continuous, distance traveled between airports *in miles*

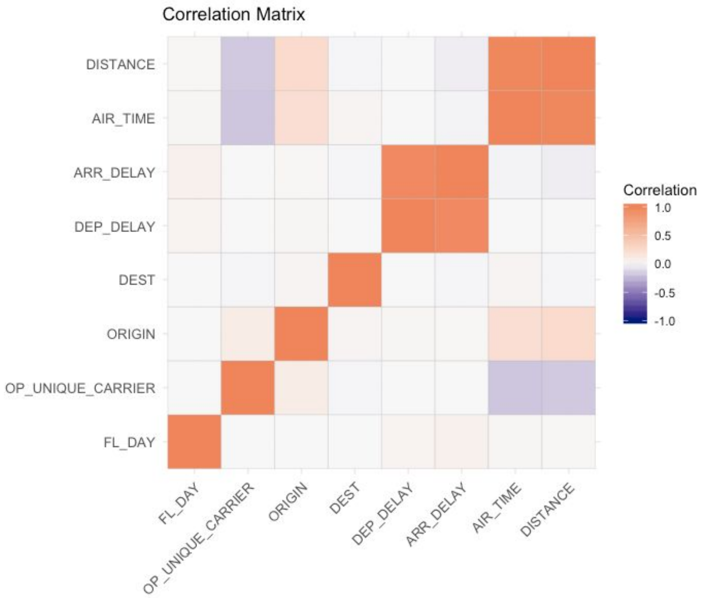
The next step was to preprocess the data in order to analyze it further:

- The original dataset was too large, but seeing as such a significant portion of American flights are accounted for by the busiest airports, we saw an opportunity to work with a smaller subset of the data. **We only chose to consider flights originating from the ten busiest airports in the United States.** These ten airports were determined by a 2019-20 United States passenger traffic report published by the Federal Aviation Administration. By removing all other instances, we reduced the size of our dataset to 190,147 rows (flights).
- The variables ORIGIN_CITY_NAME and DEST_CITY_NAME were insignificant to our analysis, so these columns were removed.

- Instances of flight diversion and cancellation do not have a departure and/or arrival time associated with them. In order to avoid the inclusion of such instances, we removed all rows containing missing delay data, as well as the columns CANCELLED and DIVERTED.
- FLY_DATE was transformed to create a new categorical variable FLY_DAY in order to account for discrepancies in travel between weekends and weekdays. Dates were altered to reflect the day of the week on which the flight took off, on a 1 to 7 scale corresponding to days from Sunday-Saturday.
- All remaining qualitative categorical variables OP_UNIQUE_CARRIER, ORIGIN and DEST were converted to quantitative discrete for further analysis.

After preprocessing, we were ready to analyze visual representations to help us further understand the trends in our data.

1. Our correlation matrix is as follows:



- As expected, we observe a strong, positive correlation between the delay in departure (DEP_DELAY) and our target variable, delay in arrival (ARR_DELAY). Their linear relationship is evident in their scatter plot (given in appendix B.)
- Again, as expected, flight time (AIR_TIME) and flight distance (DISTANCE) also have a strong, positive correlation. Flights travelling greater distances will have a longer flying time.

2. Next, we looked at the box plots of DEP_DELAY and ARR_DELAY (given in Appendix B.) We observed that both DEP_DELAY and ARR_DELAY contain a substantial number of extreme outliers. These are the flights that experienced a significant, longer-than-average delay. **We removed all flights that were delayed beyond 24 hours, or 1,440 minutes.** 12 months of prior FAA data reveal this amount of delay to be an extremely unlikely occurrence.
3. In order to apply the Box-Cox method to find the best transformation to apply on ARR_DELAY, and likewise, DEP_DELAY to induce a normal fit, we converted ARR_DELAY and DEP_DELAY to *positive* variables by applying to them the function $\boldsymbol{x = x - \min(x) + 1}$. After examining the corresponding plot (given in Appendix B), we deduced that $\lambda = 1$ is ideal. Therefore, ARR_DELAY, and likewise DEP_DELAY, **were left untransformed for modeling.**

Model Building

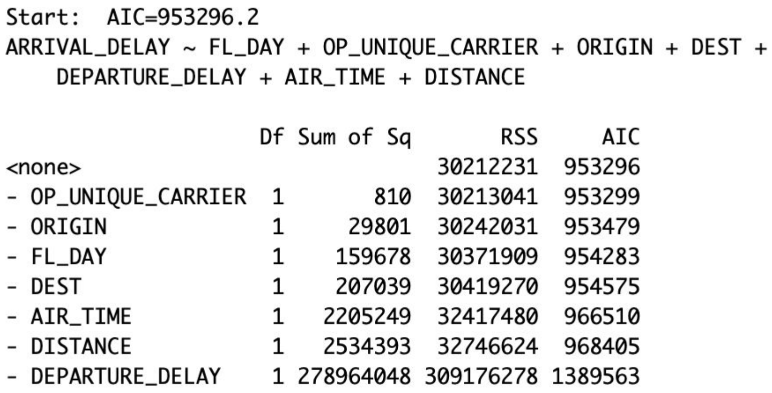
We began the model building process by fitting a linear model on the target variable ARR_DELAY with all the other variables as predictors, taking standard assumptions of multiple linear regression.

After fitting the initial model, we went on to validate our assumptions by observing all diagnostic plots (given in appendix B.)

- The residuals versus fitted plot produced a horizontal line running through zero. The distribution of points appears to be random, but it is difficult to tell with a high concentration of points between 0 and 500.
- The Q-Q plot that helps us assess normality is linear for the most part, with some expected skewing at the higher tail. Again, we expect such a trend given the presence of numerous and relatively scattered outliers at the extreme right end of the distribution.
- The scale-location plot also produced a roughly horizontal line. There appears to be an equal distribution of points, but, again, it is difficult to tell with a high concentration of points between 0 and 500.
- Finally, the residual versus leverage plot emphasizes the most influential cases. These cases correspond to flights with significant delays less than the 24-hour threshold. The model performs well despite these.

RESULTS

We used **backward stepwise regression using AIC** to eliminate variables from the regression model, evaluate the fit all subsets of the initial model, and select the most optimal one.

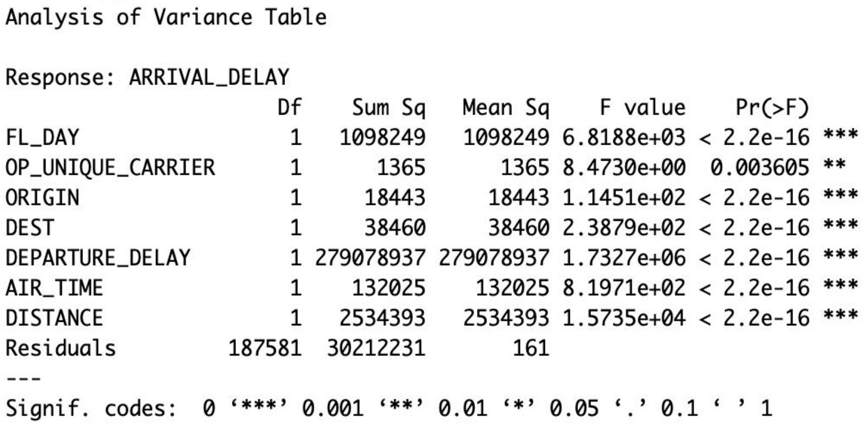


```
Start:  AIC=953296.2
ARRIVAL_DELAY ~ FL_DAY + OP_UNIQUE_CARRIER + ORIGIN + DEST +
    DEPARTURE_DELAY + AIR_TIME + DISTANCE

                    Df Sum of Sq       RSS     AIC
<none>                            30212231  953296
- OP_UNIQUE_CARRIER  1       810  30213041  953299
- ORIGIN             1     29801  30242031  953479
- FL_DAY             1    159678  30371909  954283
- DEST               1    207039  30419270  954575
- AIR_TIME           1   2205249  32417480  966510
- DISTANCE           1   2534393  32746624  968405
- DEPARTURE_DELAY    1 278964048 309176278 1389563
```

As is evident from the output, the lowest or preferred AIC value belongs to our original model. Removing any of these predictors results in an unfavorable model with a higher AIC value. Therefore, stepwise selection revealed that our original model is the best fit, and it allowed us to confirm that each of these attributes is, indeed, an important predictor of arrival delay.

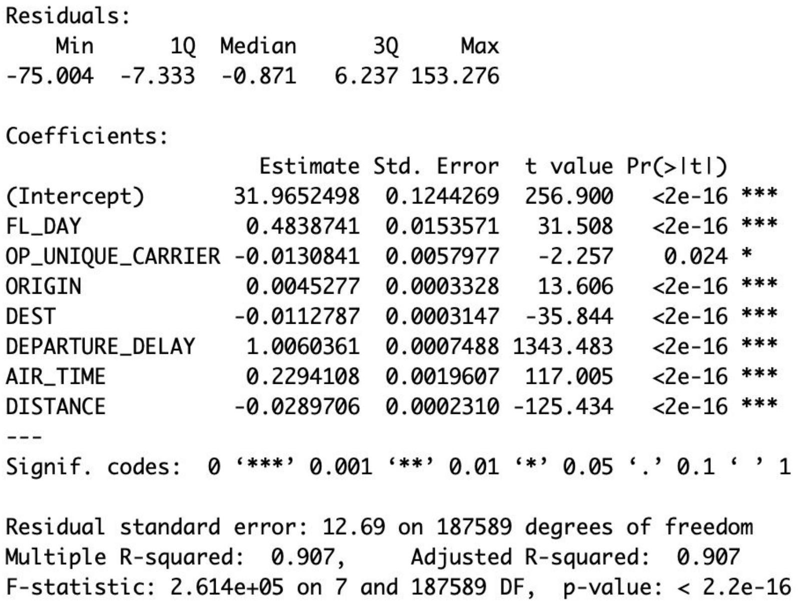
To further corroborate this, we examined the ANOVA table for our proposed model. This can be seen below. Again, it is evident that all the predictors are significant at the 5% level, OP_UNIQUE_CARRIER being relatively less significant than the others.



```
Analysis of Variance Table

Response: ARRIVAL_DELAY
                    Df    Sum Sq   Mean Sq    F value    Pr(>F)    
FL_DAY               1   1098249   1098249 6.8188e+03 < 2.2e-16 ***
OP_UNIQUE_CARRIER    1      1365      1365 8.4730e+00  0.003605 ** 
ORIGIN               1     18443     18443 1.1451e+02 < 2.2e-16 ***
DEST                 1     38460     38460 2.3879e+02 < 2.2e-16 ***
DEPARTURE_DELAY      1 279078937 279078937 1.7327e+06 < 2.2e-16 ***
AIR_TIME             1    132025    132025 8.1971e+02 < 2.2e-16 ***
DISTANCE             1   2534393   2534393 1.5735e+04 < 2.2e-16 ***
Residuals       187581  30212231       161                         
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Our final model is the same as our original or the complete model. The parameters estimates, standard errors, t-statistics and p-values are provided in the summary of the final model given below:

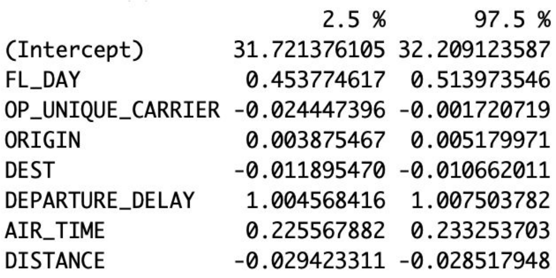


```
Residuals:
    Min      1Q  Median      3Q     Max 
-75.004  -7.333  -0.871   6.237 153.276 

Coefficients:
                    Estimate Std. Error  t value Pr(>|t|)    
(Intercept)       31.9652498  0.1244269  256.900   <2e-16 ***
FL_DAY             0.4838741  0.0153571   31.508   <2e-16 ***
OP_UNIQUE_CARRIER -0.0130841  0.0057977   -2.257    0.024 *  
ORIGIN             0.0045277  0.0003328   13.606   <2e-16 ***
DEST              -0.0112787  0.0003147  -35.844   <2e-16 ***
DEPARTURE_DELAY    1.0060361  0.0007488 1343.483   <2e-16 ***
AIR_TIME           0.2294108  0.0019607  117.005   <2e-16 ***
DISTANCE          -0.0289706  0.0002310 -125.434   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.69 on 187589 degrees of freedom
Multiple R-squared:  0.907,	Adjusted R-squared:  0.907 
F-statistic: 2.614e+05 on 7 and 187589 DF,  p-value: < 2.2e-16
```

The confidence intervals for all parameter estimates are as follows:



```
                         2.5 %       97.5 %
(Intercept)       31.721376105 32.209123587
FL_DAY             0.453774617  0.513973546
OP_UNIQUE_CARRIER -0.024447396 -0.001720719
ORIGIN             0.003875467  0.005179971
DEST              -0.011895470 -0.010662011
DEPARTURE_DELAY    1.004568416  1.007503782
AIR_TIME           0.225567882  0.233253703
DISTANCE          -0.029423311 -0.028517948
```

Critical observations:

- The p-values confirm that **the model and all the predictors are significant at the 5% level**.
- The airline carrier OP_UNIQUE_CARRIER is relatively less significant than the other predictors. This is an interesting observation, given that flyers generally estimate delay solely by assessing an airline's track record or their previous experience flying with it.
- The multiple R^2 value = 0.907, which means that the model succeeds in explaining **90.7%** of the variance in the data. This is an indicator of a reasonably strong and accurate model.

DISCUSSION

Before we book a plane ticket, there are many things to consider. Anticipating flight delays based on the day, time, airport and airline is one of them. The answers to these questions are readily available on the internet, but where is the evidence supporting these claims? Through this project, we were able to successfully examine the most influential factors that can cause mayhem come flight day. The first big takeaway from this project was that the airline carrier itself does not play as crucial a role as we think it does in getting us to our destination on time. There are several critical attributes at play here, including the day of the week, departure and arrival locations, duration of flight, departure delay, and the distance being flown. Our analysis concluded that all of these factors play an important role in determining arrival delay. Our backward stepwise analysis shows that the strongest model includes all of our selected predictors, and the ANOVA table corroborated the significance of all variables at the 5% level. The second big takeaway from this project was that after departure delay, the distance flown by the flight was found to be the second most influential predictor of arrival delay.

Limitations and Applications

While our findings appear to provide a statistically significant model for predicting arrival delays, we acknowledge there are some limitations to its applicability. Because we limited the data to the ten busiest U.S. airports (approximately 30% of domestic flights in the U.S.), we cannot comment on flights originating from other airports. Furthermore, choosing to analyze flights only in the month of January may have made our model too specific to make year-round predictions from. However, according to EU Refund.me reporting, February experiences similar instances of delay as January. It also appears, from this source, that the occurrence of delay is fairly consistent among these two months across recent years. Another limitation to our model is the duration of delay. Predictions are most accurate up to about 500 minutes in departure delay. According to the NAS, the average length of delay among the upper 5% of passenger delays is 150 minutes, with delays becoming increasingly unlikely for greater durations.

Our model can be used to build a proper ‘forecast’ for delays of January flights originating from the top ten busiest airports in the United States, perhaps something that can come in handy next winter! Another possibility includes extending this model to account for more, or other, months of the year. Adding other features such as the time of the day the flight is expected to take off, predicted weather conditions and expected airport traffic may improve the overall quality and accuracy of the model.

REFERENCES

"Final Full-Year 2019 Traffic Data for U.S. Airlines and Foreign Airlines U.S. Flights." *Bureau of Transportation Statistics,* United States Department of Transportation, 19 Mar. 2020, www.bts.gov/newsroom/final-full-year-2019-traffic-data-us-airlines-and-foreign-airlines-us-flights. (Data source)

"OST_R:BTS: Transtats." *Bureau of Transportation Statistics*, United States Department of Transportation, 2020, www.transtats.bts.gov/DatabaseInfo.asp?DB_ID=120&DB_URL=.

"OST_R: BTS: Title from h2." *Bureau of Transportation Statistics*, United States Department of Transportation, 2020, www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?pn=1. (Data supporting rarity of 24+ hour delays in the past year.)

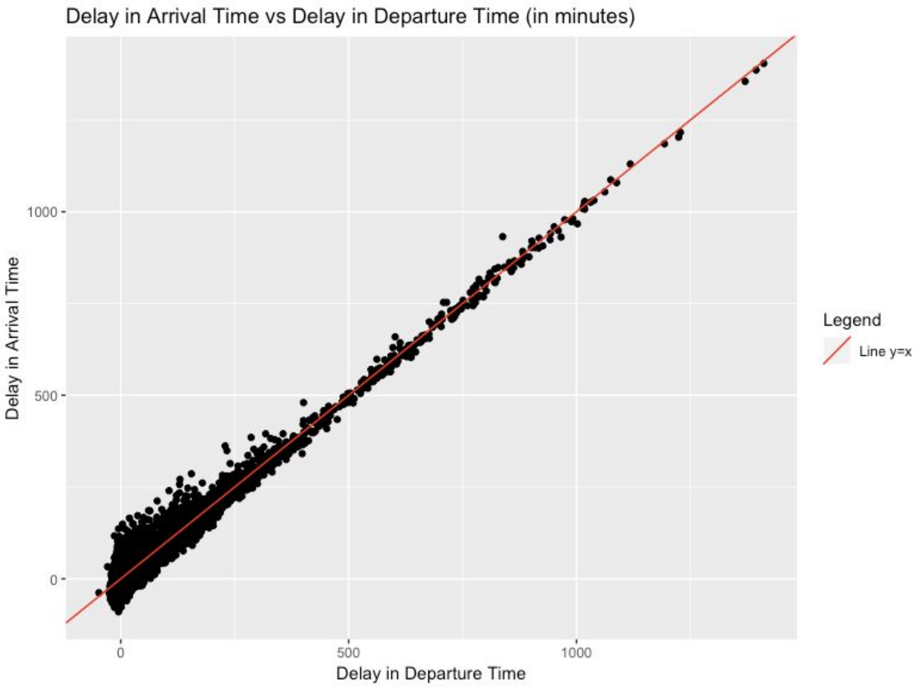
Sherry, Dr. Lance, et al. "Comparison of Flight Delays and Passenger Trip Delays In The National Airspace System (NAS)." *Center of Air Transportation and Systems Research*, 4 Dec. 2007, https://catsr.vse.gmu.edu/pubs/StatisticalComparisonMetrics%5b6%5d.pdf

Wang, Danyi & Xu, Ning & Larson, Melanie & Sherry, Lance. (2008). Statistical Comparison of Passenger Trip Delay and Flight Delay Metrics. Transportation Research Record. 2052. 10.3141/2052-09.

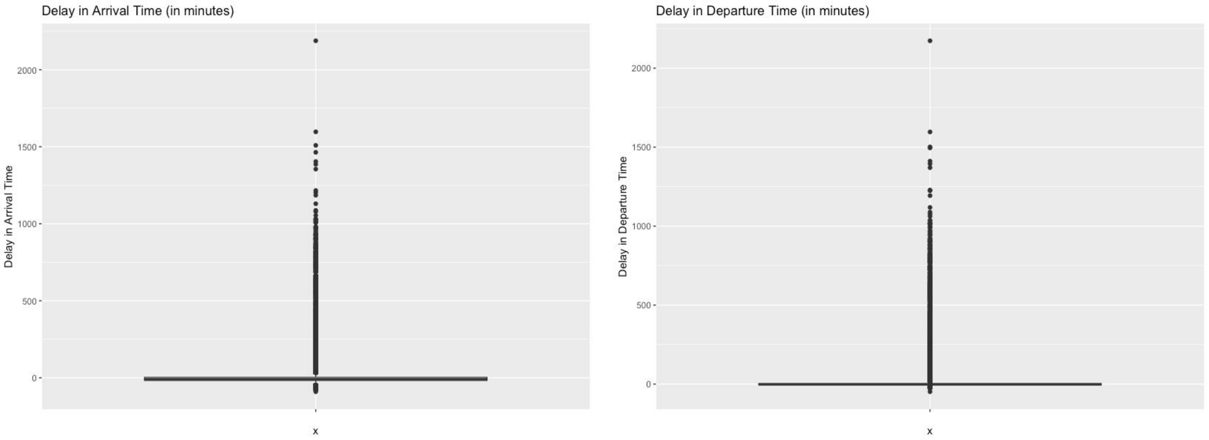
APPENDIX

Output

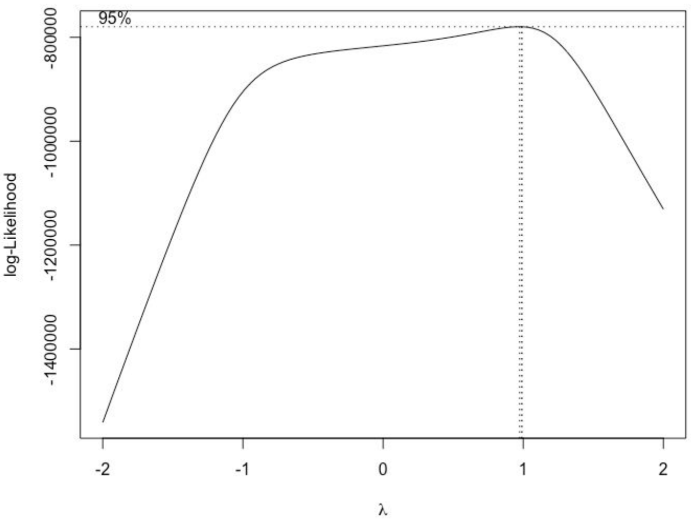
Plot of ARR_DELAY vs DEP_DELAY



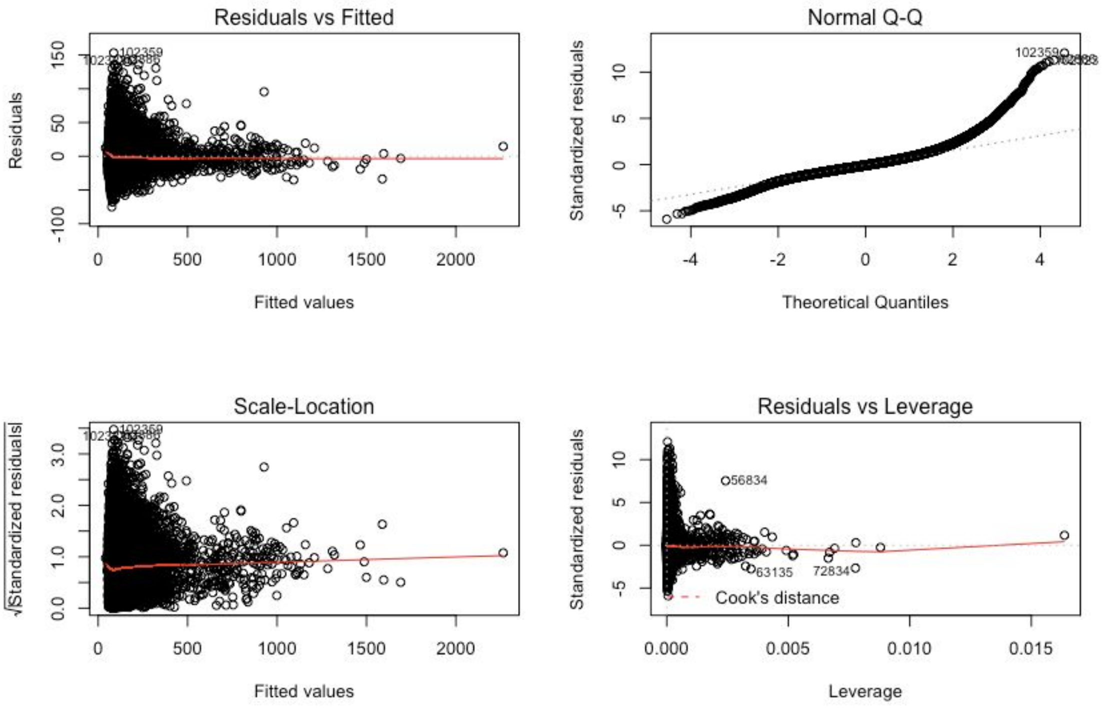
Boxplots of ARR_DELAY and DEP_DELAY



Box-Cox Transformation Analysis Plot



Diagnostic Plots


Residuals vs Fitted
Residuals
150
50
0
-100
102359
0
500
1000
1500
2000
Fitted values
Normal Q-Q
Standardized residuals
10
5
0
-5
102359
-4
-2
0
2
4
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
3.0
2.0
1.0
0.0
102359
0
500
1000
1500
2000
Fitted values
Residuals vs Leverage
Standardized residuals
10
5
0
-5
56834
63135
72834
Cook's distance
0.000
0.005
0.010
0.015
Leverage