Airline Flight Delay Prediction and Analysis

Steffi Chern (steffic)

October 14, 2022

Introduction

Airline companies seek to reduce cost and provide good services to their customers at the same time. To achieve their goals, minimizing the times of delayed arrival of a flight to its destination is a critical problem that needs to be addressed. To understand the most important features that contribute to the flight delays, the Bureau of Transportation Statistics (BTS) has kept track the details of every flight for the past twenty years, some of which includes the actual flight time, security problems, late aircraft, and the time taxiing to the runway. **(1)** With these information provided, we intend to answer the following research questions: whether there is a linear (or some other) relationship between flight arrival delay and departure delay, and whether or not having weather delay is dependent on the relationship between flight arrival delay and departure delay.

After our extensive analysis on the variables, models, and assumptions used, we found out that there is evidence of a statistically significant linear relationship between the response variable flight Arrival Delay and the predictor Departure Delay. Furthermore, this relationship does not depend on whether or not there are weather delays, thus the predictor Weather does not seem to be a confounding variable in this scenario.

Exploratory Data Analysis/Initial Modeling

2.1 Data

We utilized the dataset that contains the flight details in 2008 (collected from the Bureau of Transportation Statistics) to do our analysis in the rest of the report. There are 4887 flights

in total in this dataset. While the dataset consists of 17 variables, we only explored the three most important variables that are to the best of our interests. Our response variable is the flight Arrival Delay, which indicates the number of minutes that the flight arrived early (negative) or late (positive). The quantitative predictor is flight Departure Delay, which indicates the number of minutes that the flight departed early (negative) or late (positive). The categorical predictor is Weather, where 1 means if delay is due to weather and 0 if delay is not related to weather.

2.2 Univariate EDA

(3) From figure 1, we see that the histogram for our response variable (Arrival Delay) is heavily right skewed with potential outliers at its tail. From our summary (not shown here), we see that the median is -2, which verifies what we observe about the histogram – most data are located around the negative values or close to 0. **(2)** We also observe that the histogram of our quantitative predictor (Departure Delay) is heavily right skewed (long right tails) with potential outliers. A transformation for both variables seems to be appropriate in this case to better fit our models later.

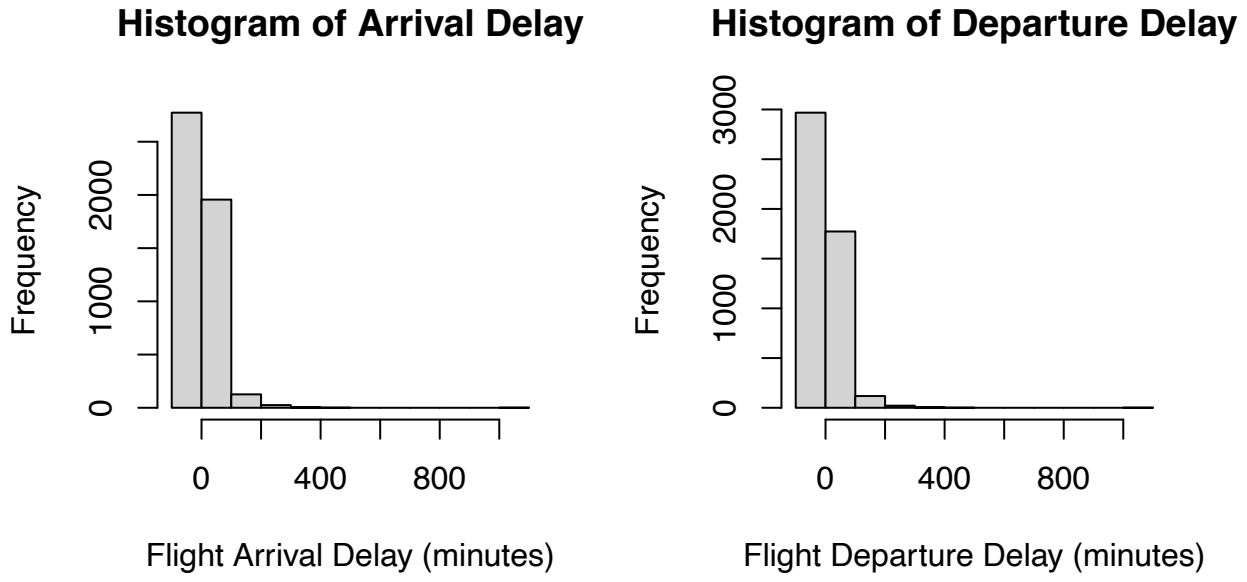

Figure 1: Histograms of Flight Arrival Delay and Departure Delay

From the frequency table for the Weather predictor (not shown), we see a very unbalanced number of samples between whether or not the flight delays were related to weather.

There are 4813 samples that are 0's and 74 samples that are 1's. It may be harder for us to draw conclusions on weather delays with such few flights (1.5% of the entire dataset).

2.3 Bivariate EDA

(4) From the figure 2 below, we notice a positive linear relationship between the flight Arrival Delay and Departure Delay. As the flight departs late, the flight tends to arrive late at the destination. There seems to be an outlier point when the departure delays more than 1000 minutes, but it doesn't affect much about the overall relationship between the two variables.

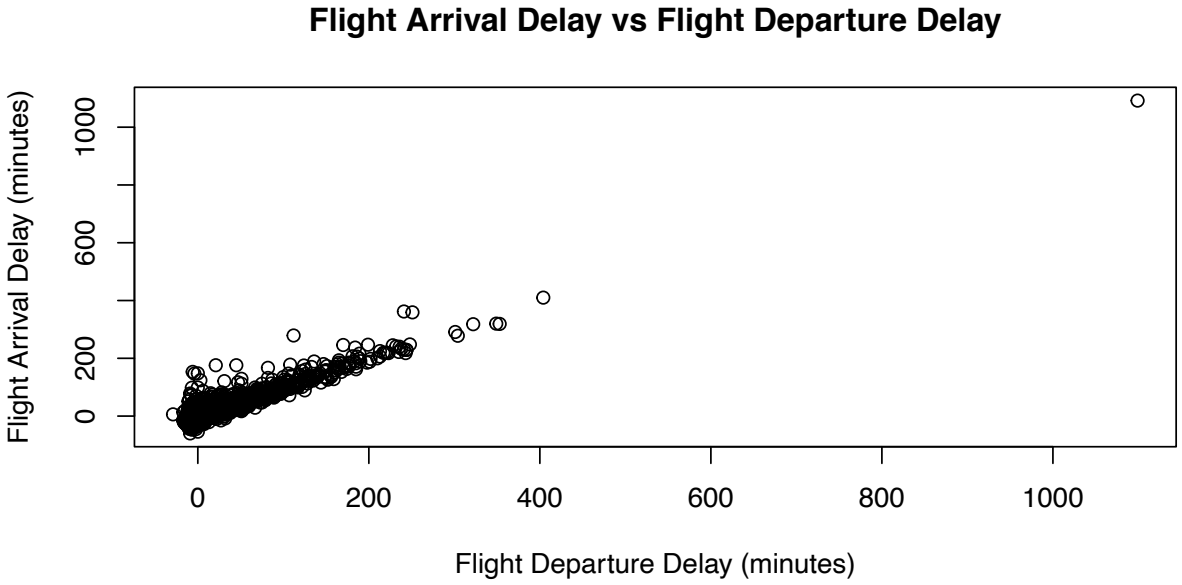

Figure 2: Scatter Plot of Flight Arrival Delay vs Flight Departure Delay

Since Weather is a categorical predictor, we decided to plot a boxplot to examine the relationship between the response variable (Arrival Delay) and Weather. From the boxplot below, we observe a slightly higher median for Weather = 1 (weather delay) than Weather = 0 (no weather delay). However, the two boxplots overlap each other partially, which may indicate a less association between Arrival Delay and Weather. Moreover, we mentioned previously that there are a lot fewer cases for Weather = 1 than Weather = 0, thus we need to be careful when making assumptions about the relationship between Arrival Delay and Weather.

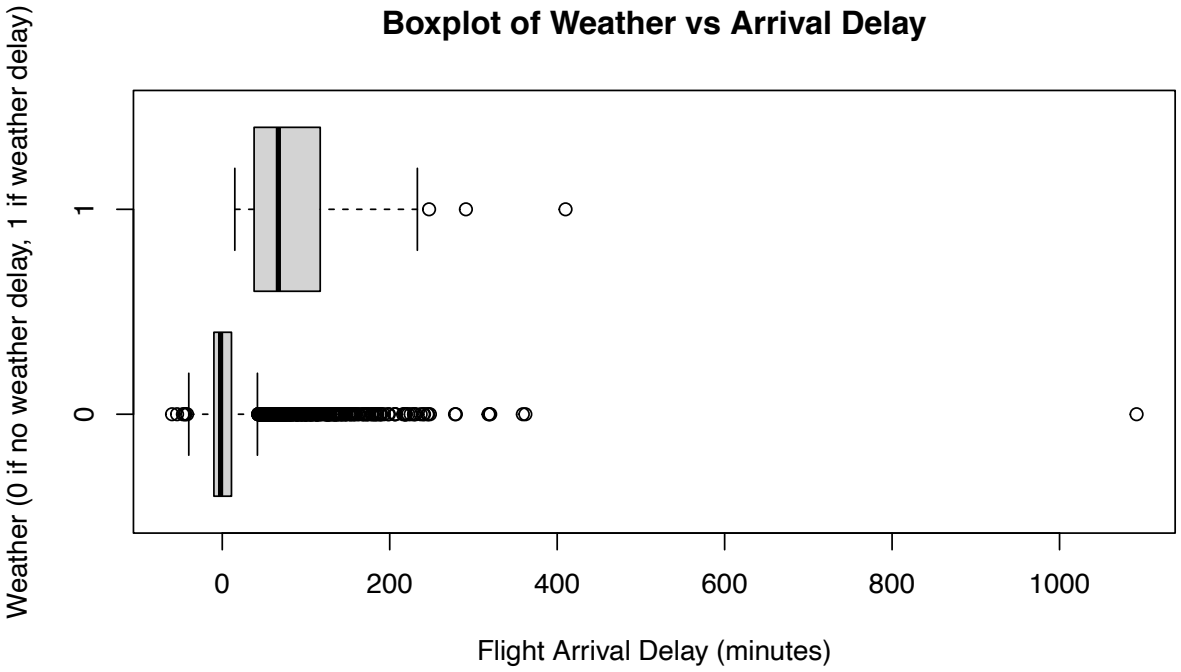

Figure 3: Boxplot of Transformed Flight Arrival Delay vs Weather

Diagnostics

Based on the results in figure 1, we decided to take a log transformation on the variables Arrival Delay and Departure Delay. However, since there are negative values in both variables, we shifted our data by adding the absolute of the minimum value in each variable plus one before taking log, such that we are only taking log on values strictly greater than 0 to avoid having undefined values.

From figure 4, we observe a few potential outliers when the flight Departure Delays for 0 or 3 minutes, but they do not affect much about the overall trend we see. The data points are more spread out along the Departure Delay values unlike the results in figure 2, where most data points are clustered in the bottom left corner. We can still see a clear, positive linear relationship between the transformed flight Arrival Delay and the transformed departure delay. To verify whether the fitted model after variable transformation is better than the original fitted model, we conduct a few diagnostics tests below.

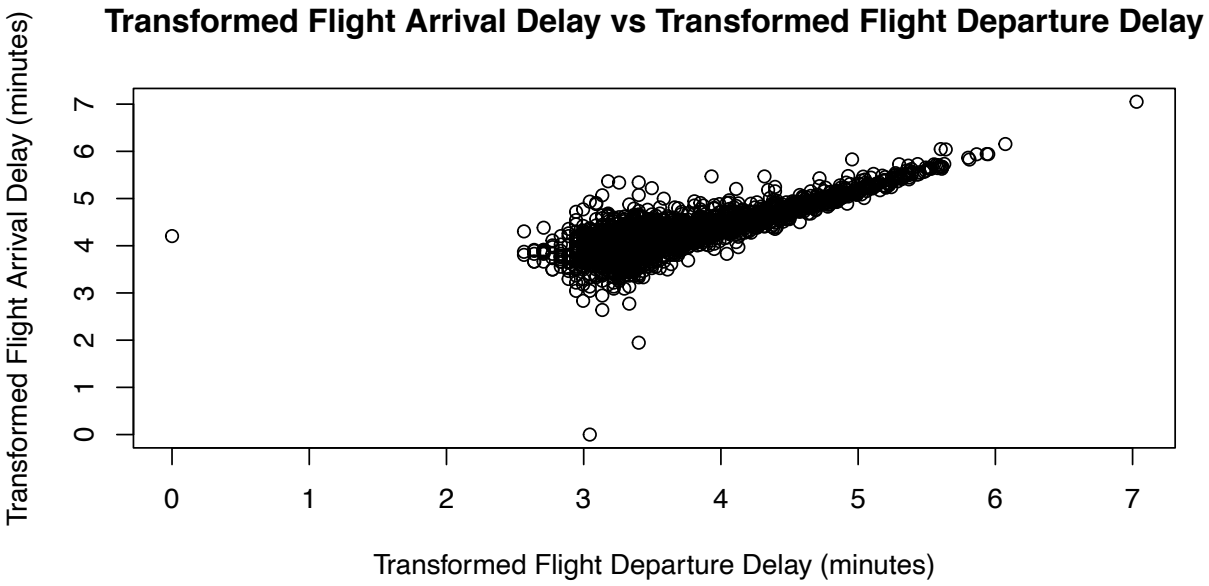

Figure 4: Scatter Plot of Transformed Flight Arrival Delay vs Transformed Flight Departure Delay

```
##        2108        4856        4239        4778        2114
## 0.37190996 0.24893126 0.21789074 0.11363919 0.06688058

##         340        1490        2108        4856        4778
## 0.47550968 0.06141466 0.04453661 0.01249141 0.01164191
```

We have calculated the Cook's distance above to check for unduly influential observations. The first block showcases the most influential observations with their associated cook's distance when original variables were used to fit the model, whereas the second block displays the same information when transformed variables were used. The five most influential points (highest Cook's distance) in each cases have been identified. We compare the values to the quantiles of the F distribution with 2 and n-2 (4885 in this case) degrees of freedom. Since all the observations in both cases did not exceed the 50th percentile of the F distribution (0.69), we conclude that none of the observations are overly influential.

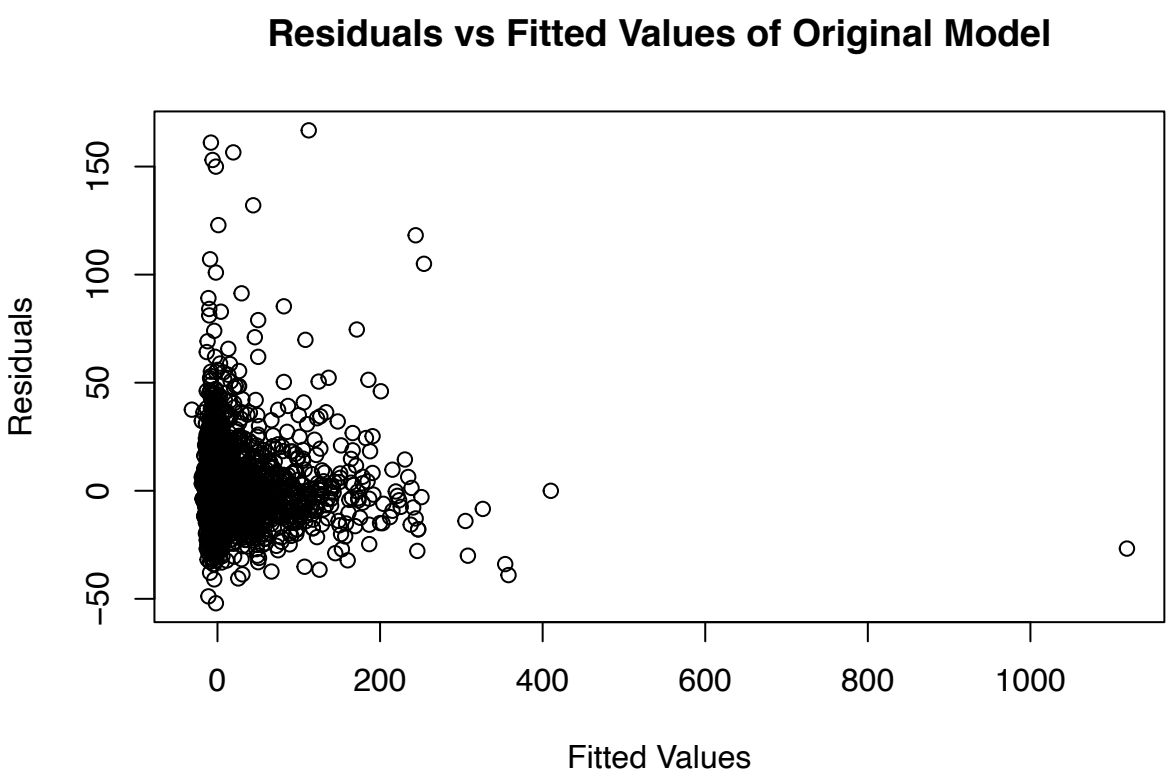

Figure 5: Residuals vs Fitted Values for Original Model

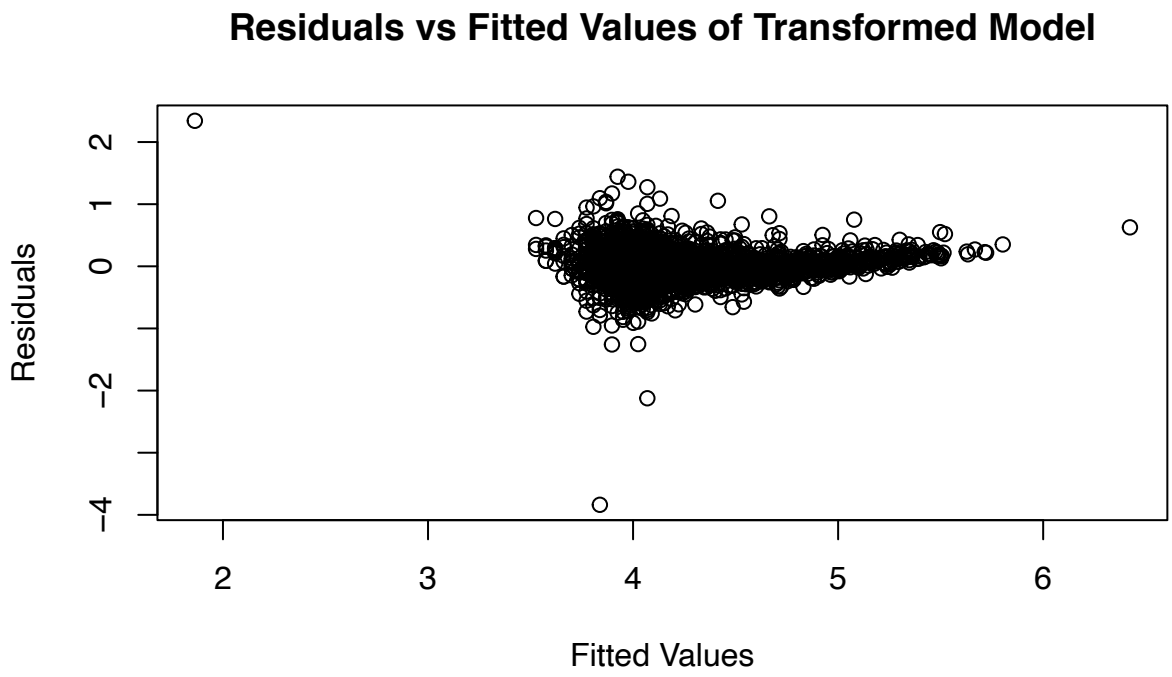

Figure 6: Residuals vs Fitted Values for Transformed Model

When we visually fit a horizontal zero line in both figures above, we notice the variance of the residuals in figure 6 is relatively more constant (residuals more evenly spread above and below the line) than what we have in figure 5 (thus figure 5 suffers from heteroskedasticity). The mean residual value is approximately 0 in figure 6 but not for figure 5 since it contains more outlier points. Moreover, the residuals in figure 6 are more patternlessly scattered above and below the horizontal zero line since the residuals in figure 5 seem to decrease as the fitted values increase (residuals might be correlated across observations).

Based on Normal Probability Plots below, we can assess whether or not the assumption that irreducible errors are normally distributed is valid. From figure 7, we observe the upper end of the Q-Q plot deviating a lot from the straight line, indicating a heavily right-skewed distribution. In figure 8, we see that the ends slightly curved away from the straight line, but the majority of the residuals lie along the straight line, and the deviation is certainly not as significant as figure 7. Figure 8 seems to indicate the residuals are close to normally distributed with slightly heavier tails.

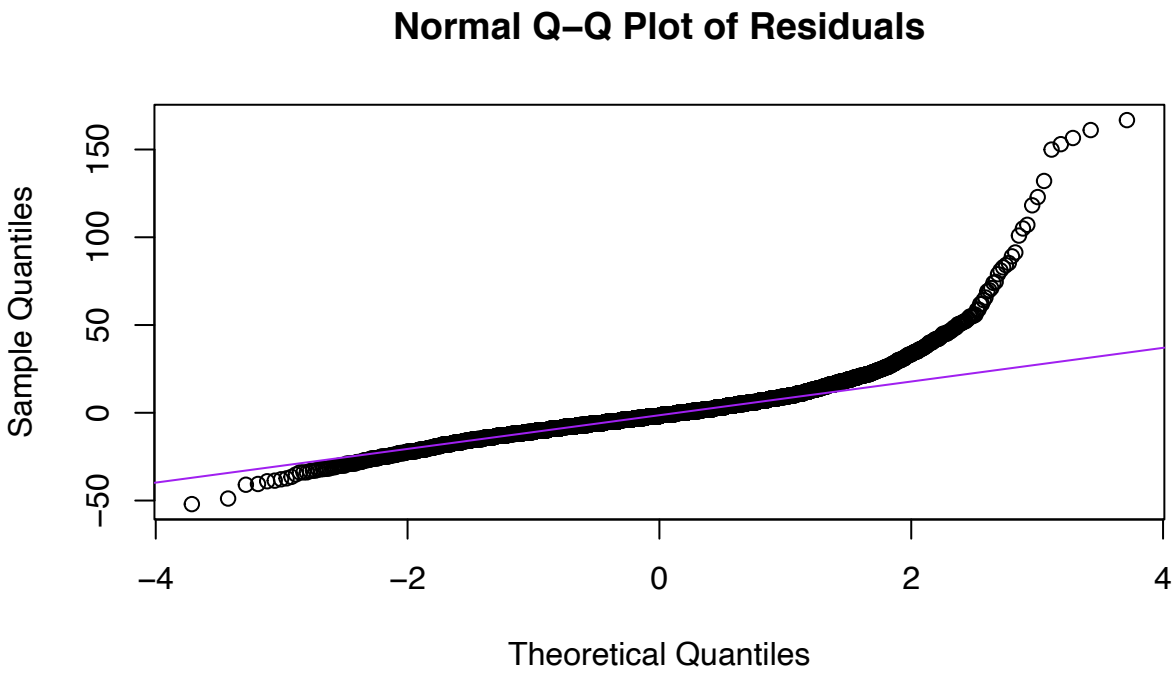

Figure 7: Normal Probability Plot of the Residuals for Original Model

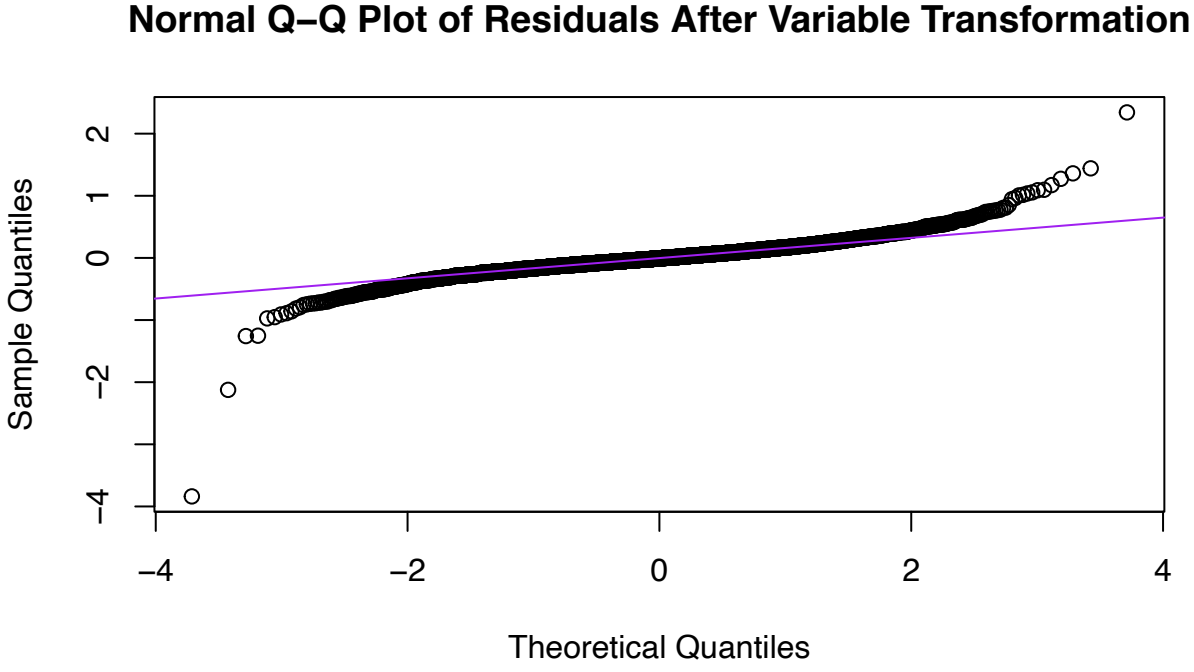

Figure 8: Normal Probability Plot of the Residuals for Transformed Model

(5) After evaluating the diagnostics tests for the fitted model using original variables vs the fitted model using transformed variables, we will be using figure 3 (transformed fitted model) as our final model to describe the relationship between the variables Arrival Delay and Departure Delay. **(6)** Since it seems reasonable to fit a simple linear regression model in this case, we can make the following assumptions about our model: the distribution of the predictor variable (Departure Delay) is arbitrary, the expected mean error is 0, the variance of the errors is constant for all predictor values, the errors are uncorrelated, and we can use the equation $Y = \beta_0 + \beta_1 x + \epsilon$ to predict the value of Arrival Delay. **(7)** We can ensure these assumptions are reasonable for our final model by looking at figure 6 and figure 8, where we have previously done a detailed analysis on the diagnostics tests to verify the simple linear regression model assumptions.

Model Inference and Results

To further explore the relationship between flight Arrival Delay and Departure Delay, we conduct a summary of their fitted linear model below.

```
##
## Call:
## lm(formula = logArrDelay ~ logDepDelay, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8381 -0.1124 -0.0069  0.1072  2.3418
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.862848   0.022532   82.68   <2e-16 ***
## logDepDelay 0.648784   0.006326  102.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2167 on 4885 degrees of freedom
## Multiple R-squared:  0.6828, Adjusted R-squared:  0.6828
## F-statistic: 1.052e+04 on 1 and 4885 DF,  p-value: < 2.2e-16
```

To test if there is a statistically significant linear relationship between the variables Arrival Delay and Departure Delay, we perform a hypothesis test:
H_0: There is no significant linear relationship between Arrival Delay and Departure Delay.
H_a: There is a significant linear relationship between Arrival Delay and Departure Delay.

When we test at the 95% confidence level, we get a critical value around 1.96. **(8)** From the summary above, we observe that the t-value is 102.55, which is greater than 1.96, thus we reject the null hypothesis in this case and conclude that there is a statistically significant linear relationship between Arrival Delay and Departure Delay. We can also verify this by inspecting the p-value. We have a p-value of 2e-16, which is less than 0.05, proving that we should reject the null hypothesis.

(9) To make estimations, we calculated the results using the *predict* function. We estimate that the mean arrival delay for a flight that has a departure delay of 200 minutes is 131.62 minutes, and we are 90% confident that the true mean arrival delay for a flight that has a departure delay of 200 minutes is between 129.57 and 133.66 minutes.

To find the confidence interval, we use the function $confint$ with the fitted models to get our results. The 95% confidence interval when no weather delay is included to fit the model between Arrival Delay and Departure Delay is between -0.70 to 0.81 minutes. On the other hand, the 95% confidence interval when weather delay is included to fit the model between Arrival Delay and Departure Delay is between -0.81 to 0.70 minutes. **(10)** Based on the confidence intervals, we do not see a significant difference in the fitted models of the relationship between Arrival Delay and Departure Delay when the variable Weather is taken into consideration, since the intervals overlap with each other.

Conclusion and Discussion

The issue of flight delays contributes to a large portion of how satisfied customers are with the airline services. As a result, airline companies have made efforts to identify the possible reasons that are associated with flight delays, and seek to find out the most problematic ones. We look forward for the airline industry to take advantage from our analysis and make improvements in the future. **(11)** In our study, we discovered a statistically significant linear relationship between the flight Arrival Delay and Departure Delay, which intuitively makes sense because the flight times would roughly be similar every time from location A to B, thus if a flight departs late it will likely arrive late at

its destination as well. We also discovered that the predictor Weather is not statistically dependent on the relationship between flight Arrival Delay and Departure Delay. It is important to be aware that there are many other predictors that weren't taken into consideration in predicting flight Arrival Delay this time. Variables such as the time taxiing to the gate and taxiing to the runway could also be statistically related to flight Arrival Delay. It could be the case that the more time spent on taxiing, the higher chance there is of having a flight Arrival Delay. When inspecting multiple predictors at the same time in the future, we would also need to take note of the possibility of collinearity between predictors, as it may undermine the statistical significance of what we intend to predict. In addition, a larger sample size for weather delay may help us get a better understanding of its role in predicting Arrival Delay.