DS 861 HW1: 2015 FLIGHT DELAY DATA ANALYSIS

A data analysis report submitted to Prof. Jamie Eng of
San Francisco State University
for
DS 861 Homework 1

Master of Science

In

Business Analytics

by

Shailesh Krishna

920752388

San Francisco, California

March 2020

Introduction

The primary goal of this task is to analyze the truncated Kaggle flights delay dataset containing pertinent information about the flight delays for the year 2015 and extract insights. The below mentioned sections describe the analysis performed.

Exploratory Analysis

This dataset consists of a total of 5821 observations, and each observation consists of 31 features. As this is a flight delay dataset for the year 2015, it contains information for only 14 airlines, where each airline is associated with a unique airline code. Preliminary analysis shows that the airline denoted by code HA operated only 52 flights, which was the lowest, and the airline WN operated 1285 flights, which was the highest for the year 2015.

In the dataset, there are a lot of missing values. Upon further analysis, we can see that for feature DEPARTURE_DELAY, there are 91 missing values, and for feature ARRIVAL_DELAY, there are 108 missing values. The reason for the mismatch in the counts is primarily due to flight diversion. There are some flights which got diverted and hence never reached their original destination. These observations have departure information but do not have any information about arrival. Such observations containing missing values were dropped from the dataset for further analysis.

We calculated the average and median for both departure and arrival delays and found out that for both the delays, the average is higher than the median. The table below shows the calculated values.

	Median	Average (Mean)
Arrival Delay	-5.0	4.0
Departure Delay	-2.0	8.9

The reason for this discrepancy is due to the presence of extreme values in the data. Mean is susceptible to extreme values in the data. The boxplot of the departure and arrival delay shown below confirms our observation. We can see a lot of extreme values in the data. Also, upon checking the skewness coefficient, we found that the skewness coefficient is significantly higher than zero indicating the distribution is skewed towards the right.

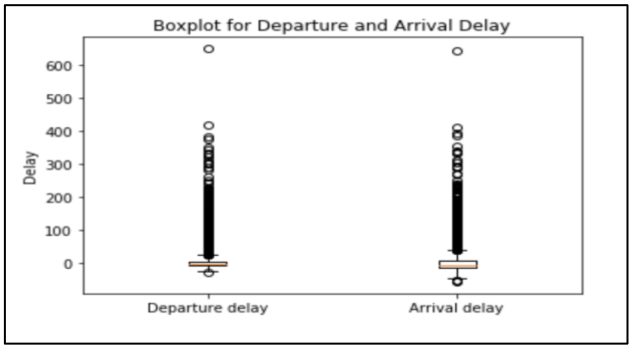

Figure 1. Boxplot for Departure and Arrival Delay

We also made a comparison of departure and arrival delay patterns and found that they both follow the same pattern closely. This similarity in the pattern shows a strong correlation between them and indicates that departure delays impact arrival delays. A correlation matrix with departure delay and arrival delay indicates the same. The below-mentioned graphs show the departure delay and arrival delays trends.

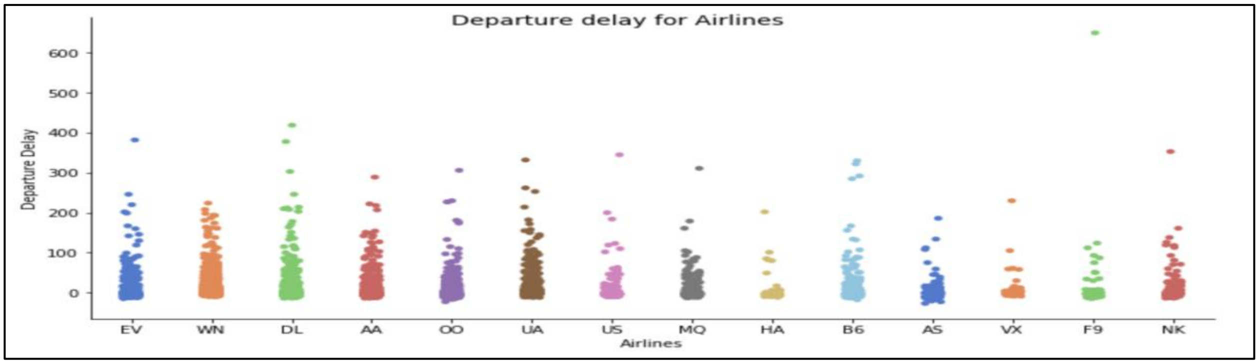

Figure 2. Departure delay for Airlines

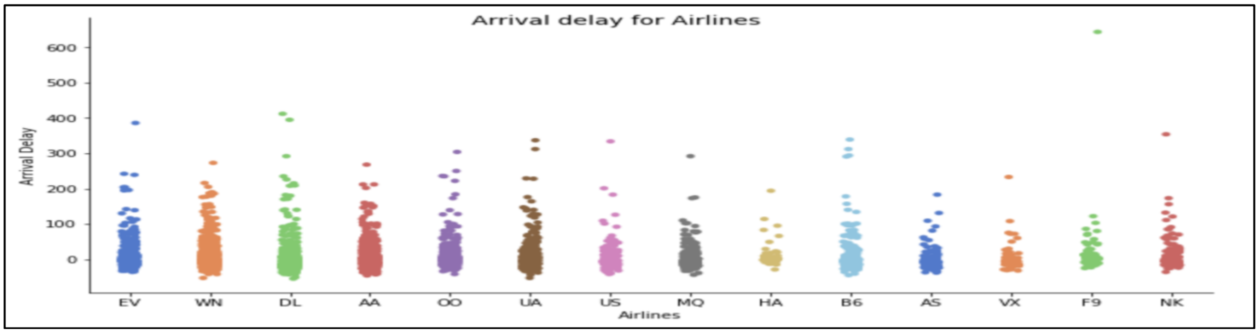

Figure 3. Arrival delay for Airlines

After calculating the 5-number summary for departure and arrival delays, we can easily say that the airline UA has the highest median value of 1.5 for departure delay, and airline F9 has the highest median value of 1.0 for arrival delay. Continuing the analysis further, we also identified the airport with the maximum average departure delay. The airport denoted by code FAR had the average departure delay of 161.0, which is the highest among other airports. Since there is only a single value in the dataset corresponding to this airport, this is the main reason for the resulting high average departure delay for the airport.

To determine the impact of distance on delays, we analyzed if there is a correlation between the flight distance and departure and arrival delays. After creating the correlation matrix, we observed that the flight distance does not impact delays as there is no correlation between the distance and delays. However, we can see a strong correlation between departure and arrival delays indicating that they depend on each other. To verify this further, we created a subset of data containing only positive departure delay values to determine if the distance can still be considered a factor as generally long-distance flights can make up for the lost time. We plotted a scatterplot of distance and arrival delay, which again confirmed that the delays do not depend on the distance and not all long-distance flights reach their destination on time.

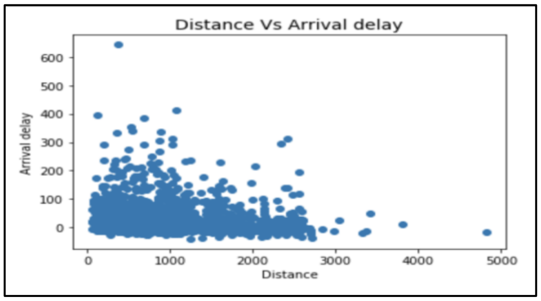

Figure 4. Distance vs Arrival delays

Similarly, we also analyzed if the day of the week has any impact on the departure delays and based on the graph below; we can deduce that day of the week has no impact on departure delays.

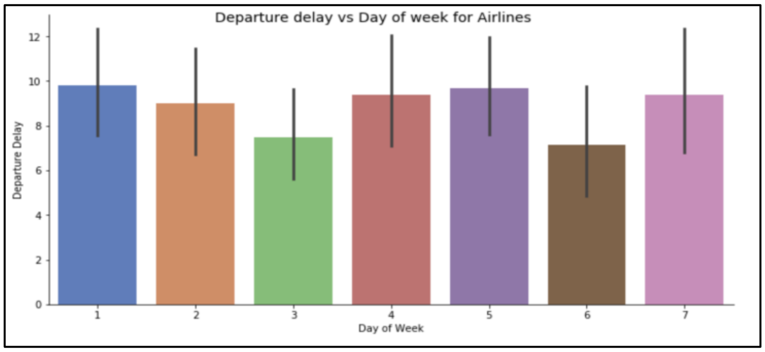

Figure 5. Departure delay vs Day of week

Additional Insights

As we dived more in-depth into the dataset analysis, many questions came across our minds. We selected a few and tried to dig deeper to get the required answers.

Based on the counts of the originating and departing flights for each airport, we can say that the ATL airport was the busiest in 2015, with a total of 691 flights, having about 326 as departing and rest 365 arriving flights. Both the charts below confirm our findings.

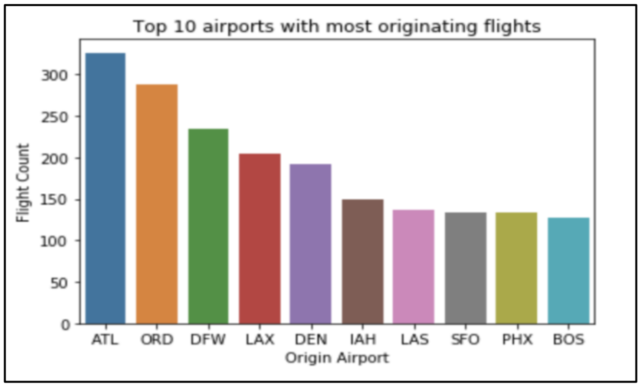

Figure 6. Top 10 airports with most originating flights

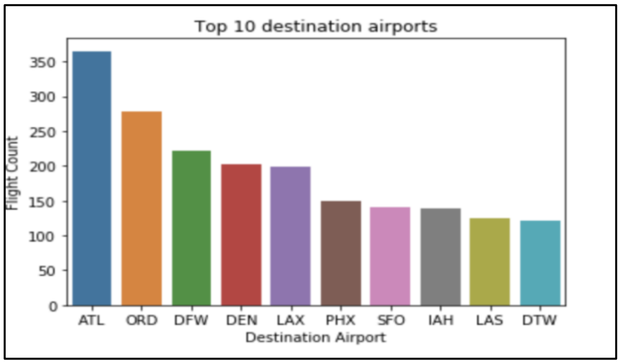

Figure 7. Top 10 destination airports

Since ATL was the busiest airport, we also tried to find the time of the year when people visit ATL the most. We calculated the monthly count of airlines arriving at ATL airport and found that it visited most during the spring and summer months. The below-mentioned bar plot denotes the same.

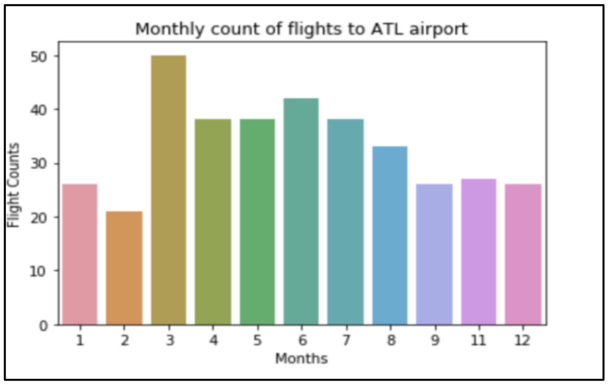

Figure 8. Monthly count of flights to ATL airport

We also tried to identify if the taxi in and out times for all the airlines is the same, or they are different. By doing a side by side comparison, we were able to determine that they are different for different airlines, but the taxi in time is always less than the taxi out time for all the airlines.

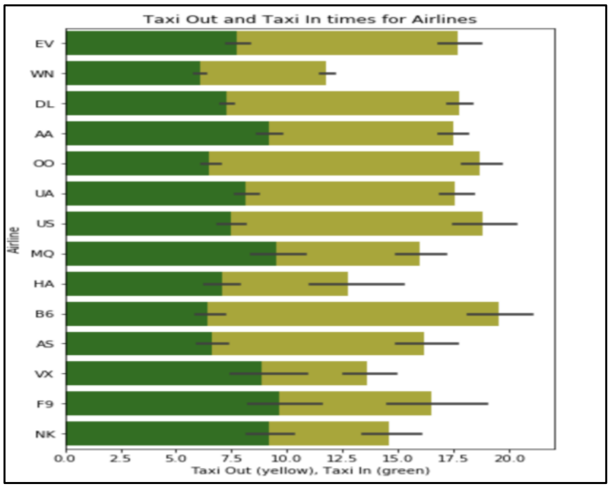

Figure 9. Taxi in/out times for airlines

The chart above shows the taxi in and out times for all the airlines.

Regression Analysis

This section describes the various approaches taken for model building to analyze the arrival delay. We utilized linear regression to do so.

Model 1

For building this model, as part of the data cleaning process, we removed all the data in the weather delay column. In the dataset, there were a couple of columns containing categorical variables such as airline and day for which their respective dummy variables were created. In the model, ARRIVAL_DELAY is the response variable and a total of nine predictors consisting of LATE_AIRCRAFT_DELAY, AIRLINE_DELAY, AIR_SYSTEM_DELAY, WEATHER_DELAY, DAY_OF_WEEK, DEPARTURE_TIME, DEPARTURE_DELAY, DISTANCE and, AIRLINE.

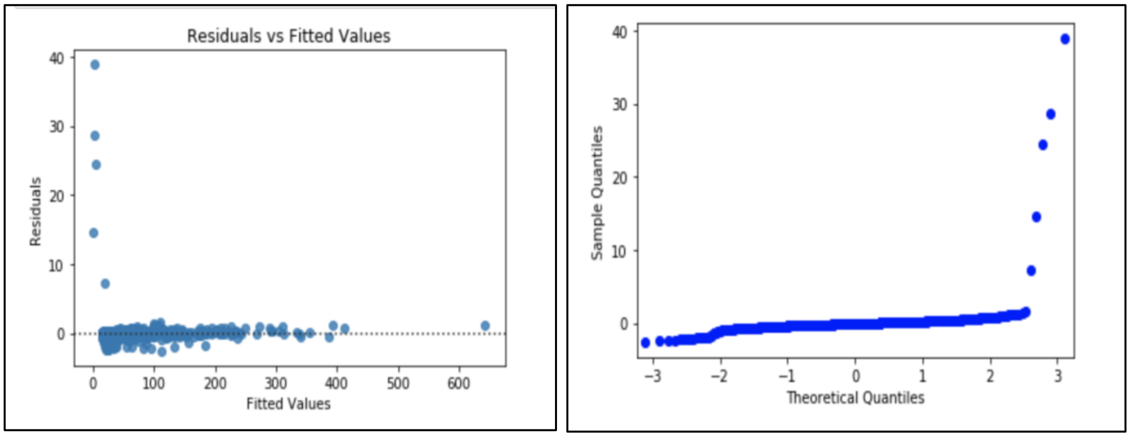

Figure 10. Residual Analysis for Model 1

After performing the above residual analysis, our model assessment says that this model does not exhibit linearity, constant variance, and normality. Though we observe a high R-square value, but it is primarily due to the large number of predictors in the model resulting in more variability. Also, many predictors present in the model are not significant and need removal.

Model 2

In this model, we first remove the extreme values in the ARRIVAL_DELAY response variable. We determined the IQR and calculated the threshold values, which we use to create the subset of data without the outliers. Also, we take a log of ARRIVAL_DELAY to use it as the response variable, and only the significant predictors are kept in the model, which are LATE_AIRCRAFT_DELAY, AIRLINE_DELAY, AIR_SYSTEM_DELAY, WEATHER_DELAY, and DEPARTURE_DELAY.

Our assessment of this model also shows that it also does not exhibit linearity, constant variance, and normality. The R-square value is also less as compared to the previous model. All the predictor variables in the model are significant. The distribution is skewed towards the right as denoted by the normality plot.

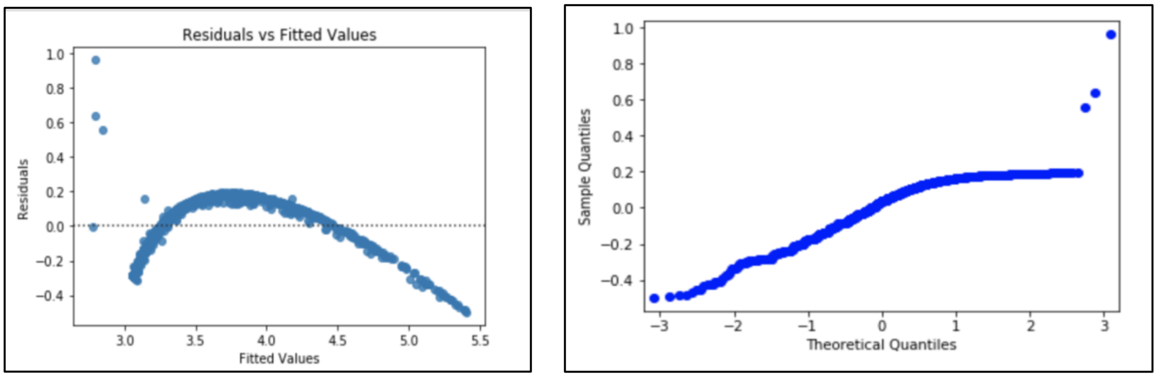

Figure 11. Residual Analysis for Model 2

Based on the analysis of both the models, we can see that still further improvements are needed to build a better model to analyze arrival delays. The below-mentioned additions can be done to the model:

- Introduce interaction variables in the model.
- Using Tukey's ladder transformation, we can either increase or decrease the power of independent variables and use them in the model.

Summary

From the data analysis performed on the flight delay dataset, we were able to extract useful insights and also determined the useful predictors required to analyze the arrival delay. The models created were not a great fit and required further improvements.