KJ Morgan

Data Science Internship

Research Project Introduction

July 3, 2019

**Introduction**

*Why Are Flights Cancelled or Delayed?*

In the world today, researchers have stated that approximately 8-15% percent of the people that book a flight somewhere in the world, misses their intended predicted flight. Approximately, 15% are late to their flight and out of all the flights flown in one complete year, 32% of the total flights are cancelled or delayed. The dataset, shows that between the initiations of the airplane to the landing of the airplane, there are numerous variables that take effect to the timing of the flight. There are numerous variables are included into the original dataset. The variables consist of weather, late aircraft, and security delay. An incremental change in variables affect the punctuality and take-off of aircrafts at airports. With the abundance of airports in the world, the problem with passengers and cancelled and delayed flights, is common. However, the variables in the dataset show the variables factored into airports in the US. The passenger that has more time between arrival and departure is usually the person that experiences cancelled or delayed flights.

I would like to look at a model of an airplane’s arrival variable dataset and analyze the variables and information in the dataset. This will lead me to create a probability analysis that will show a decrease in the probability of a flight being cancelled or delayed. I will use probability and statistics in my experiment on the variables to create graphs and models that displays the variation in probability change with the manipulation of different delay variables. I predict that shortening the time of the late aircraft delay will have the biggest impact on the probability of a flight being cancelled. Even from the other factors, weather may have the largest impact to create a delay in time for an airliner. To know exactly how they affect the different variables, that is where the models and statistics favor. Furthermore, the overall experiment will use data analysis and eventually lead to a solution being brought upon the problem of airline delay in society.

**Data Wrangling**

Remove Insignificant Variables-

As one scans through the dataset, you will see that there was an abundance of variables that were not going to give you an accurate approach to my research question. Therefore, I had to use a simple logic approach to each variable. That approach was to simply ask myself do I see this variable impacting airlines. Going down the list, I saw that the delay times for Late Aircraft, Security, and Weather Delay were going to impact enough to provide a statistical analysis. I went into RStudio to use a code to remove 26 variables of the starting 30, which leaves me with the 4 variables in the final dataset.

Rename Variables-

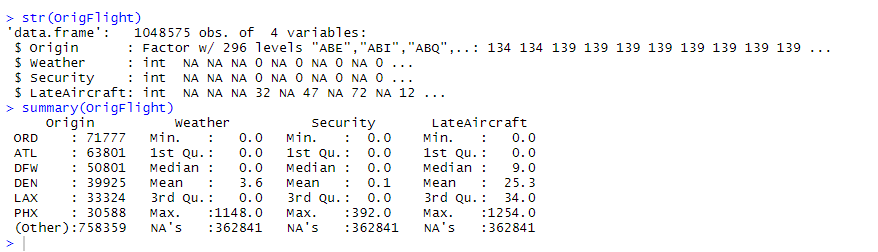
The variables in the dataset were given a name that came with the dataset. However, for my preference I would like for the variables to be named to where I understand them fully. Therefore, I used a code that renames all the variables in the dataset. I renamed the variables to Weather, Security, and Late Aircraft.

Remove N/A-

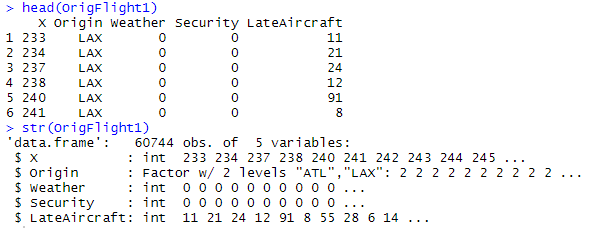
With respect to having a dataset with more than a million observations. Some of the observations in the dataset were filled with N/A values. This is information that I do not need in the dataset. From this point, the N/A removal code was very beneficial because the actual values in the dataset will be the values being analyzed. The code deleted the N/A observations, which results me with my final dataset.

Create a Subset of Data-

The final subset of data comes from the final product of removing the redundant data in the dataset. The information that was not needed such as the N/A and extra observations were removed by the code. I found that to actually analyze my data to the standard was that I had to mine the data down to two major airports on different sides of the US. Those two airports are places where I was going to analyze data from, so that there will be more of an accurate approach to the delay rate from major airports.



(Above) is the original dataset that was given from the open educated source. The original dataset contained about 1.1 million observations, with values ranging from 0-875. Some of the values in the dataset were insignificant. Therefore, by data mining/ wrangling, those values were removed.

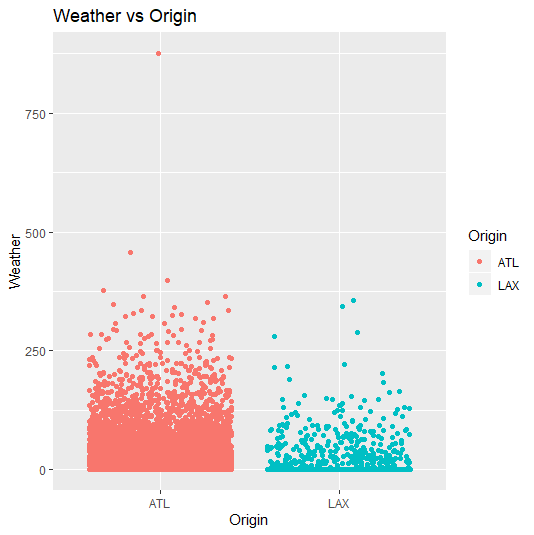


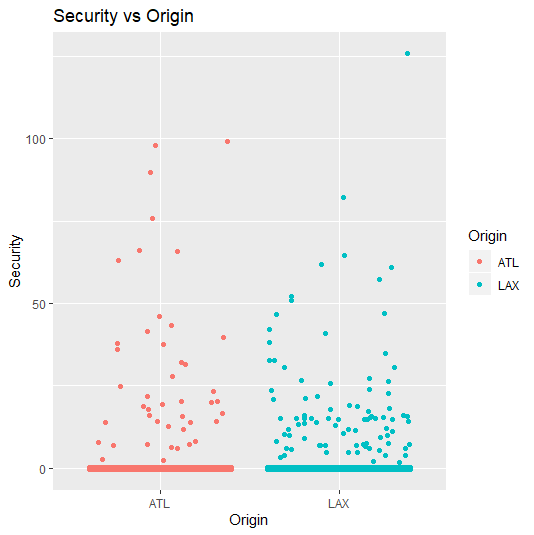
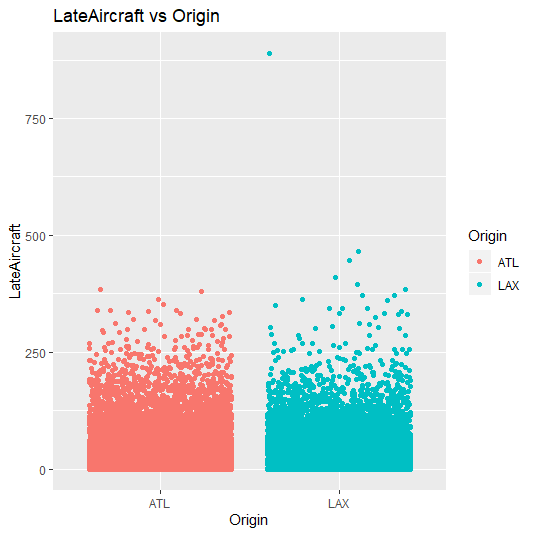
(Above) This is the final product after the data wrangling process has been applied. The dataset contains a significant difference in observations from the original dataset.

**Statistical Analysis**

By analyzing data from this dataset, the best way to catch trends in the data values is a scatter plot. A scatter plot is analogous to the line graph in which values will be will be measured over time. I can translate my data into a scatter plot and watch the variations in the delay times. The two airports will be measured from a month span of delayed flights. Each airport, ATL and LAX will have variations in delay times. The airport with the higher frequency in delay flights will be one that I feel needs to work on airport services and efficiency techniques, so that the high frequency obtained, can be lowered. From this, one can find build a model using the different variables.

A B C





1. Looking into the Weather vs Origin graph you can see on that ATL has a higher delay rate than LAX. The mean of the delay times in ATL was significantly higher than the mean of LAX.
2. Displayed in the Security vs Origin scatter plot. You can see that both mean averages look similarly close. The values of delay time on this scatter plot range close to the same value. However, the times of ATL, makes it out to have a higher mean. This indicates a higher probability of delay.
3. Late Aircraft has a large amount of observations detailing that the Late Aircraft variable is highly affecting the delay times of both airports. In the graph the averages of LAX were higher than ATL.

**Data Analysis**

The dataset contains a lot of information. The original dataset contained 30 different variables. Each of the variables affecting a flight, causing a time delay. However, the problem is which variable can be manipulated the most, so that there will be a decrease in the delay time. The dataset has variables that include; scheduled arrival and departure time, actual arrival and departure times, weather, security, and many more variables. Important information formulates around the importance of the values in the dataset. The variables are also very important because the purpose of the research is to figure the variable or variables that are causing the highest probability in flight delay.

In the dataset is a great quantity of delay times. There are only so many questions that can be answered with so many time variations. A question that I cannot answer from the data is, “Will the decrease in airplane delay result in bigger or smaller airplanes?” Or, “Can the delay times during this month help a researcher on an annual analysis on delay times? The data in the dataset is from the year 2019. This indicates that one cannot analyze annual data because the dataset does not contain annual data.

I used data wrangling to clean the original dataset. I eliminated 26 variables by identifying their importance and significance. The data wrangling techniques helped shave insufficient data. The values in the clean dataset were the values that remained analyzed. Those remaining leave about 60000 observations. However, with the validation of the models, the validation dataset contain approximately 13000 observations. As you can see this is a smaller dataset with more significance of the values.

In this experiment, I am going to analyze 3 dependent variables against an independent variable. From there I am going to plot the variations of delayed time over a month into a scatter plot. The scatter plots will all show the delay times over a month period. The scatter plots will have the independent variable which is origin, and the dependent variables will be measured against origin. The origin variable contains ATL and LAX, which are the major airports that I received from data wrangling. Then, I will use Linear Regression models on the different variables to find which variable is significant towards the research. After that, I would like to figure out how to manipulate a process or service at the airports, to lessen the probability of a flight getting cancelled in ATL and LAX.