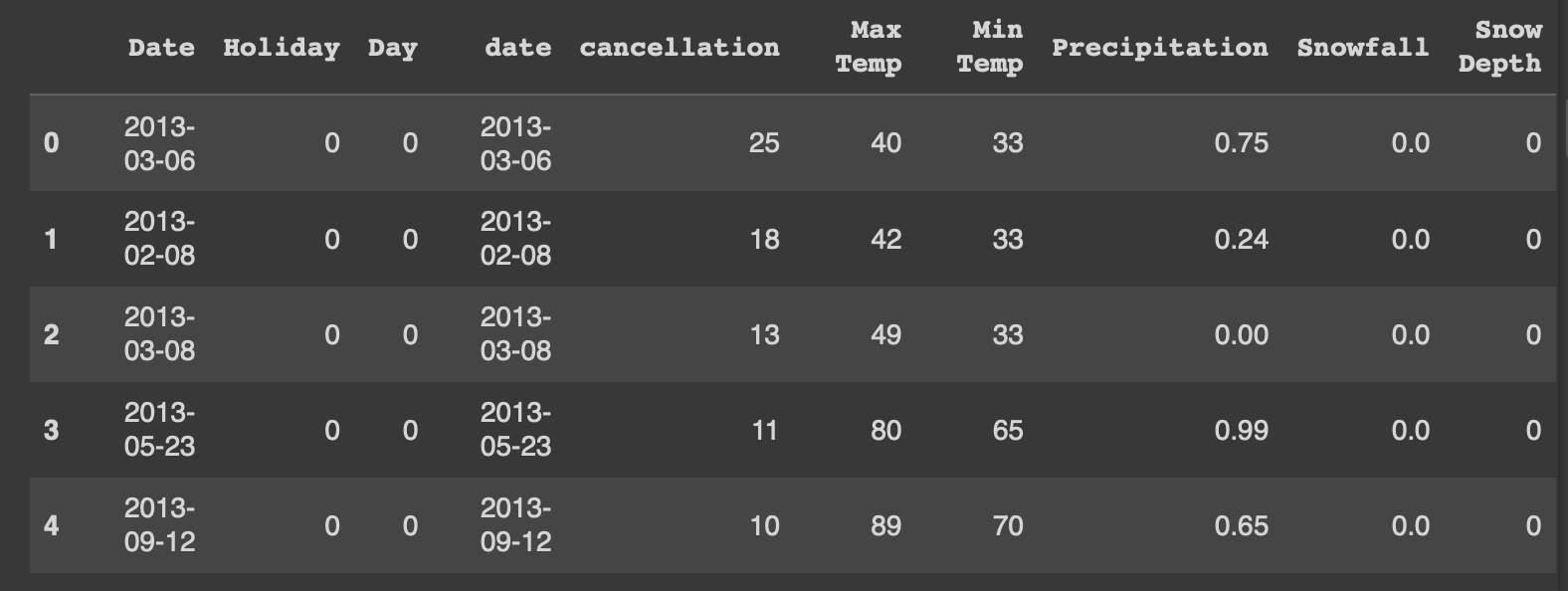
DATA 601 Final Project || Flights from New York to D.M.V Area || Fall 2022

Ruth, Shiva, Kapeesh, Anji

Introduction:

In this project, the flights from the state of New York to three local airports in the D.M.V area(DC, Maryland, Virginia) are analyzed to determine what factors influence flight schedules. The three destination airports interested are BWI: Baltimore/Washington International Thurgood Marshall Airport, DCA:Ronald Reagan Washington National Airport, and IAD: Dulles International Airport. This will be done by looking at relationships between weather/holidays and arrival delays/flight cancellations. Moreover, a hypothesis testing was used to determine which components affected flight delays and cancellation. After determining which factors lead to changes on the original flight schedule, a linear and logistic regression model were built to predict arrival delays and estimate what flights will be canceled.

First, the day with the highest and lowest number of flights was determined for each destination. The most popular destination among the three airports was DCA. And the most busy day in 2013 was on February 7 with 55 flights. Second, the day of the year with the most cancellations was on March 6. While the reason is unknown, it could be due to the weather. This leads us to question whether weather is a determinant in flight cancellations and delays. If it was, this meant that weather must be taken into consideration when planning and booking flights. To know whether a relation exists between weather and flight cancelation, we used Pearson correlation. The correlation between precipitation and cancellation was 0.25, while it was 0.12 with snowfall and cancellation. It was even lower for snow depth with .03. Since the Pearson correlation coefficient is so low, weather and flight cancellations do not have correlation.

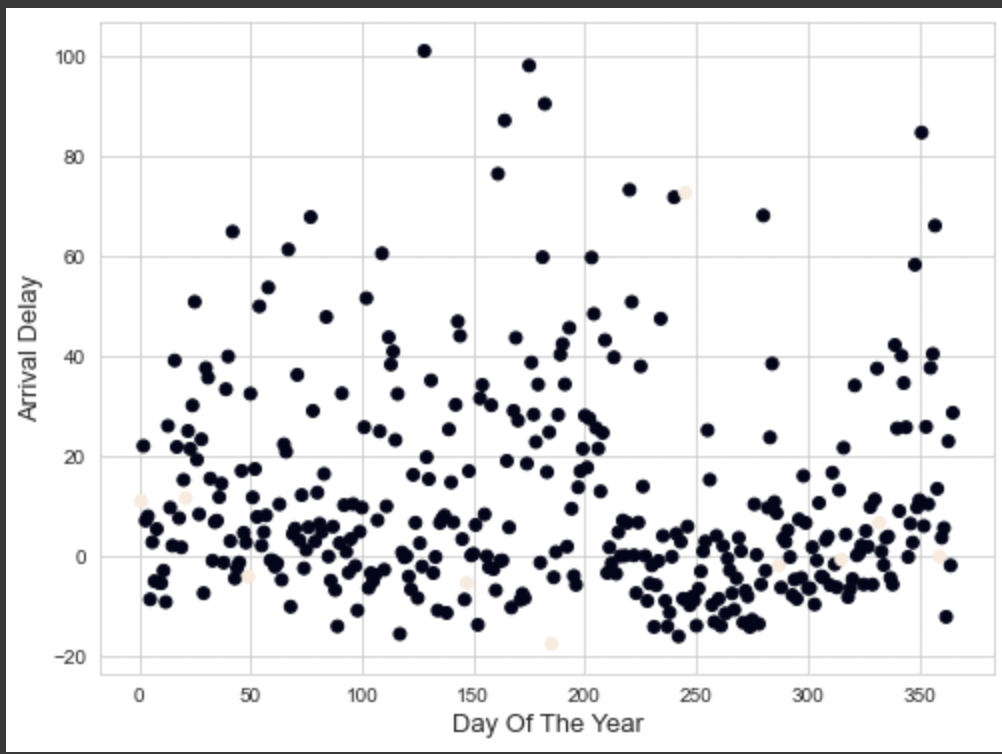
Moreover, holidays are expected to affect flight schedules because this is when people usually take vacations, so airports are usually busy during these times. As a result, we looked at whether there is a relationship between the Federal Holiday Schedule and cancellations. Once again, Pearson correlation showed there is no correlation because it is too low with -.06. Also, from the table below, one can see that cancellations do not necessarily happen on holidays. 

After calculating flight cancellations, the total canceled seats were 24,032 and this resulted in an economic loss of $1,201,600. This was calculated under the assumption that the average price of a flight ticket was $50. If ticket prices were higher than the assumption prices, which was likely, then the monetary loss of canceled flights would be even higher.

While holidays and weather determine flight schedules, it also depends on the airline company. Airlines provide different experiences, and the reliability is all different. Hence, the ratio of canceled flights and planned flights for each airline was calculated. Then, the ratio was used to decide the most and least reliable airline.

As seen on the table above, the most reliable airline was US Airways Inc. It has the most planned flights with zero cancellations. And the most unreliable airline was Envoy Air.

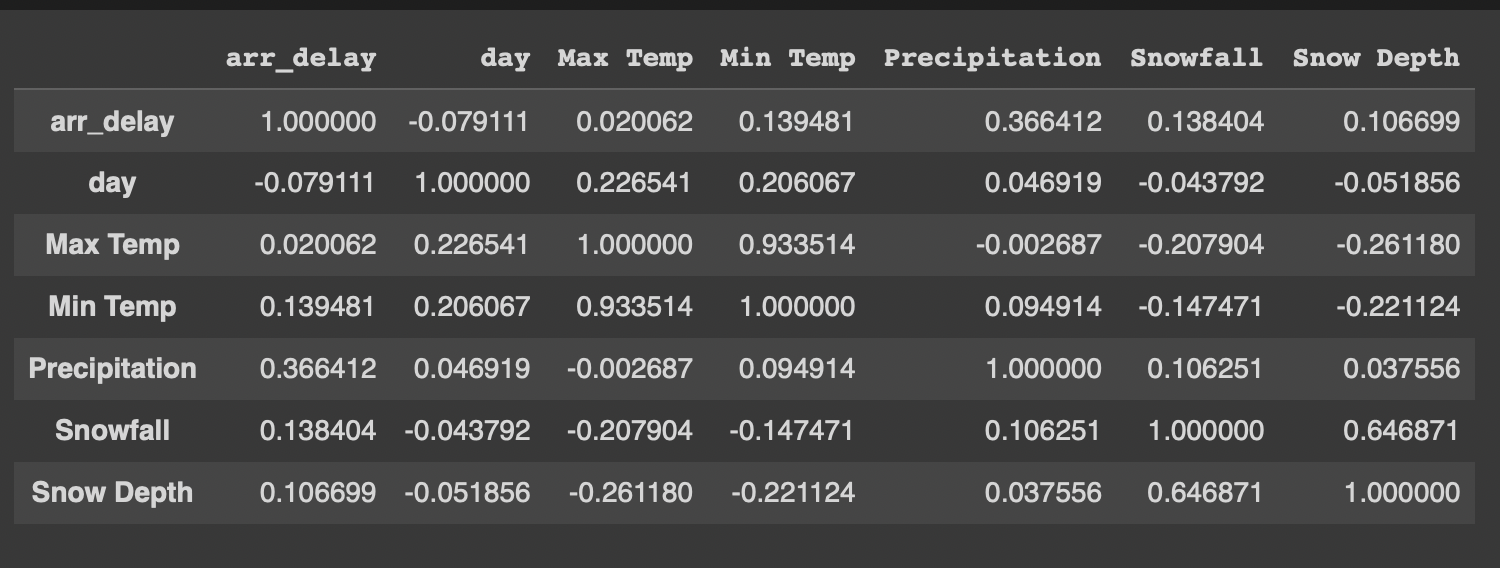
Additionally, we want to know whether there is a correlation between the Federal Holiday Schedule and daily average arrival delay. In order to know whether a relationship exists, the average arrival delay for all flights that took place in the same days was plotted. As seen below, Federal Holidays are marked with a lighter dot.



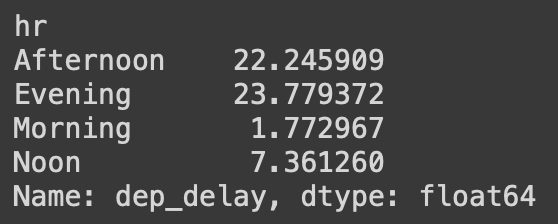
The plot shows Federal Holidays make up a small percentage of the arrival delays in the year. 2013. On top of the graph above, we did a correlation test to know whether a correlation exists between Federal Holiday Schedule and daily average arrival delay. As shown on the table below, there is a -1 correlation between arrival delay and precipitation. This is a perfect negative relationship. This shows that less Federal Holidays leads to more arrival delays.



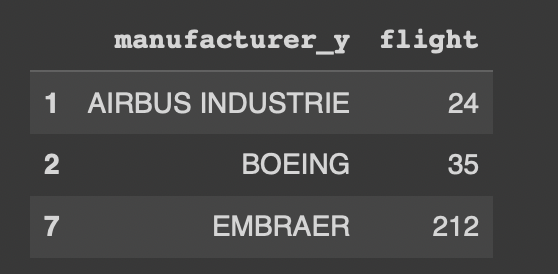
In addition, we also wanted to look at whether there is a relationship between weather conditions and the daily average arrival delay. As we can see on the table below, arrival delays and precipitation, snowfall, temperature or snow depth do not seem to have strong correlation with each other. As a result, a relationship does not exist between them.



The most reliable airport and airlines based on the average arrival delay was also of interest. The average arrival delay for all the flights to each destination airport was calculated. Ronald Reagan Washington National Airport had the shortest average delay of 8.57, so it is the most reliable airport. The least reliable airport was IAD:Dulles International Airport with an average delay of 13.87. Additionally, the most reliable airline is 00, SkyWest Airlines Inc with an average of 3 and the least reliable one is MQ, Envoy Air with 28.196. Also, the day of the week with the highest average delay was Sunday with 15.76. The highest average delay happened in the evenings, and the lowest average delays happened in the morning. Therefore, flights that took off from 6am to 10am had the shortest delays with the highest delays happening for flights that took off at 6pm to 10pm.



The most number of airlines used in the flights were manufactured by EMBRAER, followed by BOEING, and AIRBUS INDUSTRIE.



In order to estimate arrival delays of flights given, we built a linear regression model. Year, month, day, carrier, origin, destination, and distance were used to create the linear regression model needed to estimate the arrival delays. The data was split into a training and testing dataset, which is then used to create a linear regression model with a very low accuracy of 19%. This might be due to the inconsistencies in the dataset and the linear regression model isn’t perfect for predictions for this scenario.

Also, a logistic regression model was used to predict canceled flights with better accuracy of 53%. It can be noted though these models usually give high accuracy scores in real time scenarios, due to insufficient and inconsistent data, we are getting lower accuracy scores.

Conclusion:

After doing data analysis, there were several findings based on the datasets. First, we found that the most popular airport was DCA: Ronald Reagan National Airport in D.C. Calculating flight cancellations and flight arrival delays also showed that the most reliable airline was US Airways Inc because it has the most planned flights with zero cancellations. Then, the most unreliable airline was Envoy Air. Ronald Reagan Washington National Airport was the most reliable airport, while the least reliable airport was IAD:Dulles International Airport. Also, Federal Holidays and arrival delays had strong negative correlations. This means there are more arrival delays when there are fewer Federal Holidays. In addition, Sunday had the highest average delay, and the highest average delay happened in the evenings while the lowest average delays happened in the morning. Last but not least, the most popular manufacturer in 2013 was EMBRAER.