DS 861- Data Mining and Advanced Statistical methods for Business Analytics

Data Analysis Report

Name: Poonam Patil
Student ID: 920123591

❖ Summary

Flight delays are frustrating to air travelers and costly to airlines. Therefore on-time performance of airline schedule is key factor in maintaining customer satisfaction and attracting new one. In this research, analysis for flights arrival delay is performed using data visualization and statistical method. The model uses the data collected and published by the Department of Transportation's (DOT) Bureau of Transportation Statistics. Models developed in this research are tested at significance level of 5%.

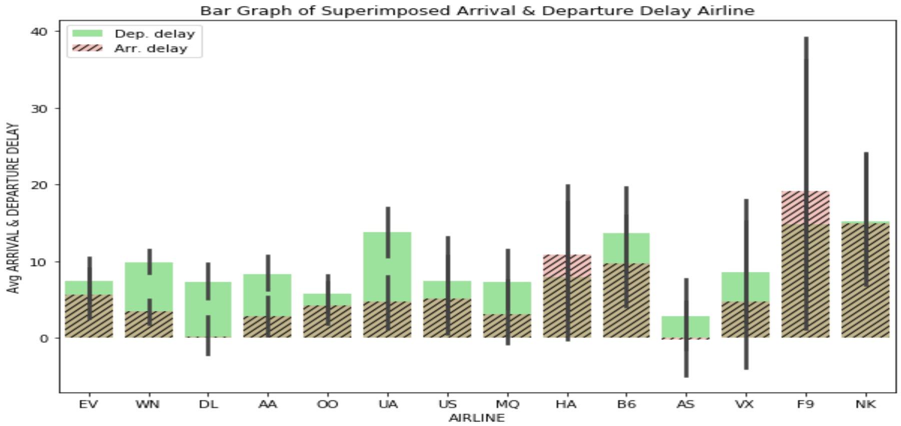
❖ Data description

The dataset contains arrival delay and departure delay for flights by major airlines and also provides additional details as origin airport, destination airport, flight number, scheduled and actual departure and arrival time, cancelled or diverted flights, taxi in and taxi out time, and distance. The data contains 31 attributes and 5821 observations of 14 different airlines. The interest of this research is analyzing factors impacting arrival time delay.

❖ Data analysis and Data visualization

Median values of departure and arrival delay reflect that flights depart 2 minutes early and arrive approximately 5 minutes early than scheduled arrival time. However, average departure delay for flights is around 9 minutes and average arrival delay is around 4 minutes. Which shows data is right skewed and has some observations with high delay.

Superimposed Bar graph below clear picture of comparison of departure delay and arrival delay for different airlines and it is seen that arrival delay is less than departure delay for 12 airlines which means flights manage to make up with arrival time even after delay in departure.

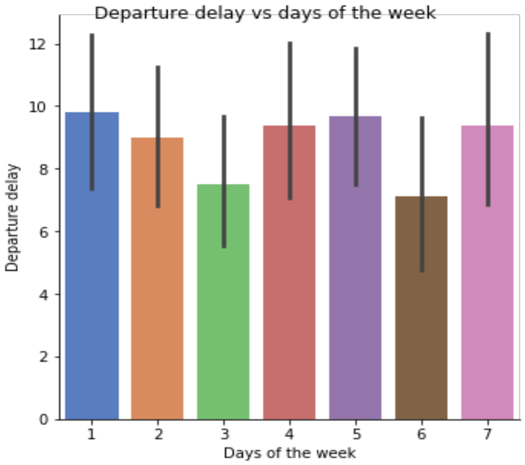


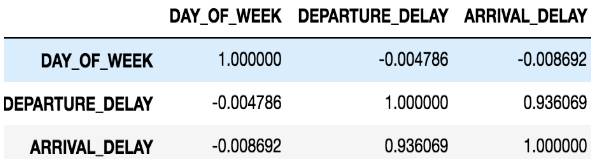
On further analysis about the origin airport causing highest delay, it is found that 'FAR' airport reported highest average departure delay of 161 min. Prime reason for delay is air system and aircraft delay.

Correlation of predictor variables with Arrival delay

1. Day of week:

The correlation between Day of week with delay is very less. The bar graph of Day of week with departure delay shows no particular pattern. Also, it is not clear from data which day number is for which day of week.

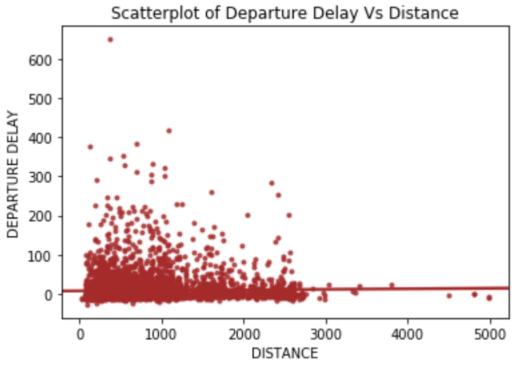


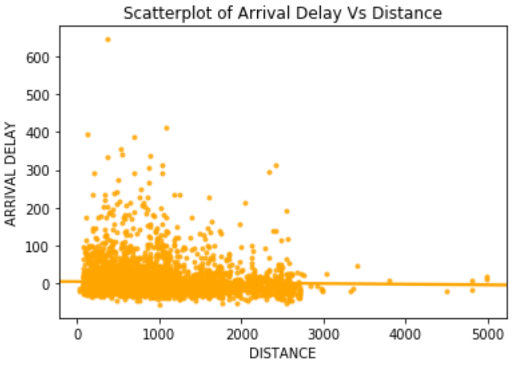


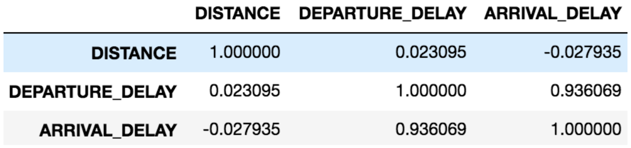
	DAY_OF_WEEK	DEPARTURE_DELAY	ARRIVAL_DELAY
DAY_OF_WEEK	1.000000	-0.004786	-0.008692
DEPARTURE_DELAY	-0.004786	1.000000	0.936069
ARRIVAL_DELAY	-0.008692	0.936069	1.000000

2. Distance:

To consider relation between distance and delay, as seen in below scatterplot, the regression line of departure delay and arrival delay with distance has slope zero which means there is no linear relation between departure delay and arrival delay with distance. Correlation matrix also proves low correlation between distance and delay.

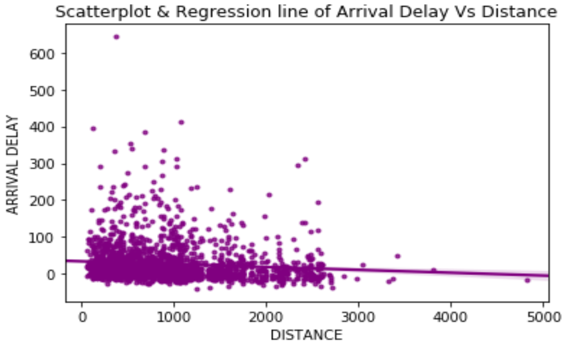


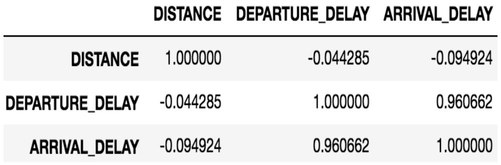




	DISTANCE	DEPARTURE_DELAY	ARRIVAL_DELAY
DISTANCE	1.000000	0.023095	-0.027935
DEPARTURE_DELAY	0.023095	1.000000	0.936069
ARRIVAL_DELAY	-0.027935	0.936069	1.000000

Since the data includes both positive and negative values for departure delay, further analysis for only positively delayed flights we selected only observations with positive arrival delay and the scatterplot and correlation matrix proves that there is still no linear relation between delay and distance. So, the distance can't be good predictor of delay.

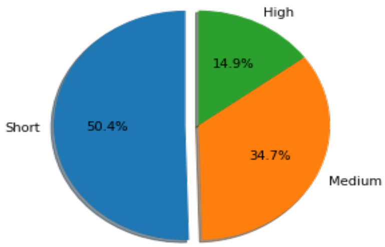




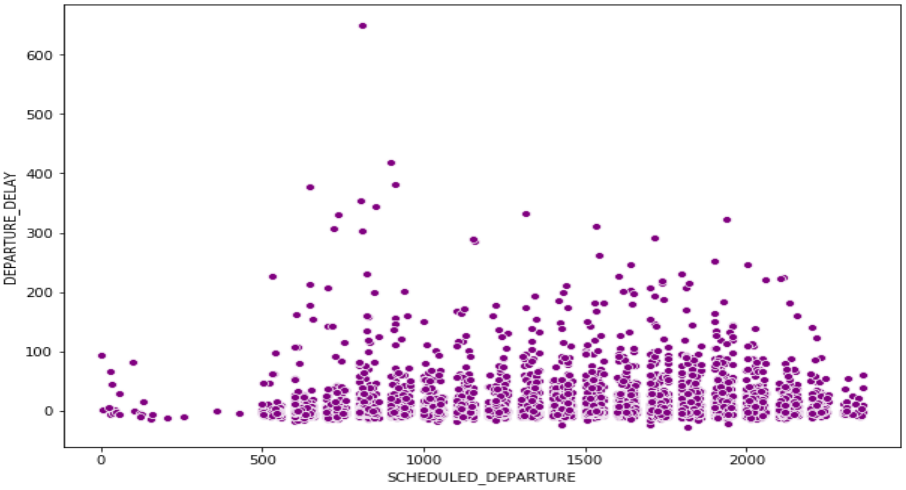
	DISTANCE	DEPARTURE_DELAY	ARRIVAL_DELAY
DISTANCE	1.000000	-0.044285	-0.094924
DEPARTURE_DELAY	-0.044285	1.000000	0.960662
ARRIVAL_DELAY	-0.094924	0.960662	1.000000

3. Departure Time:

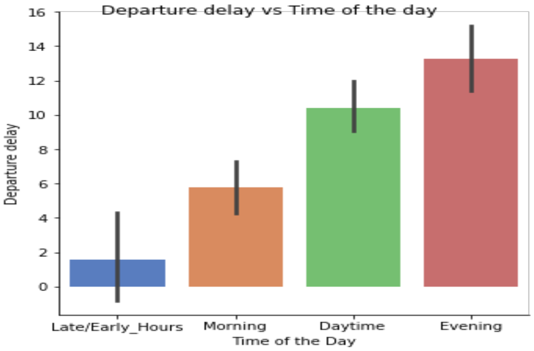
Customer satisfaction is very much dependent on departure time as higher departure delay can cause cost customer satisfaction and customer retention. To simplify visualization, we categoried departure delay into three categories – 'Short', 'Medium', and 'High' delays. Pie Chart shows for 50% of the flight's departure delay is less than 15 minutes, 34.7% flights departed 15-60 min later than scheduled time and around 15% flights incurred higher departure delay of more than one hour. Trend as seen; high delays are one third of short delays.



An interesting pattern identified in time of day (24 hrs) and departure delay which shows departure delay is more during daytime and evening than late night/ early morning as seen in below plot.



To better understand departure pattern, we categorized scheduled departure time into categories as Morning (06:00 – 12:00), Daytime (12:00 – 18:00), Evening (18:00 – 00:00), Late/ Early hours (00:00 – 06:00). And its pretty clear that departure is delayed more during daytime and evening than early morning.



❖ Data cleaning

91 observations have missing value in departure delay and 108 observations have missing value in arrival delay. After dropping the observations with null values in departure delay and arrival delay the dataset contains 5731 observation. To clean data further, observations with null value in weather delay are dropped, we get 1072 records. Regression analysis is done to check the model fit, results of the regression are included in data modelling section.

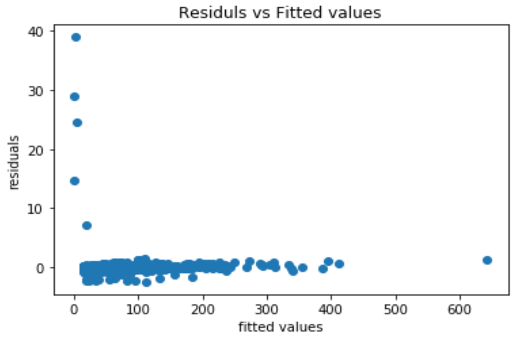
However, there are 86 outliers in Arrival delay which impacts the model. The dataset is cleaned by removing the 86 outliers for statistical modeling. So, the final dataset includes 986 observations.

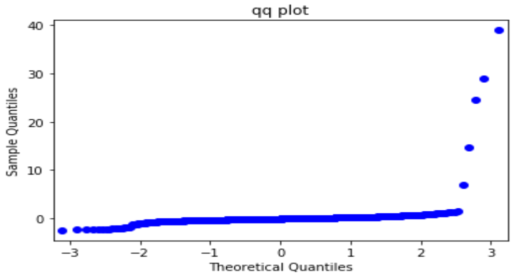
❖ Data modeling

Regression Analysis:

Regression analysis is done with 1072 records using all delay variables, day of week, departure time, distance and Airline categorical variable. Regression result p-values (including arrival delay outlier) and plot for residuals vs fitted values are shown in table below.

```
const                    0.055180
LATE_AIRCRAFT_DELAY      0.000000
AIRLINE_DELAY            0.000000
AIR_SYSTEM_DELAY         0.000000
WEATHER_DELAY            0.000000
DAY_OF_WEEK              0.260204
DEPARTURE_TIME           0.363757
DEPARTURE_DELAY          0.000003
DISTANCE                 0.281405
AS                       0.000020
B6                       0.976316
DL                       0.275878
EV                       0.562644
F9                       0.942622
HA                       0.772627
MQ                       0.647808
NK                       0.161248
OO                       0.602843
UA                       0.114207
US                       0.548723
VX                       0.877168
WN                       0.337717
```





The result shows that at 5% significance level; aircraft delay, airline delay, air system delay, weather delay, departure delay, and airline 'AS' are the only significant variables as their p value is less than 0.05. Rest of the variables prove insignificant. R-square value of the model is 0.999. However, the residual vs fitted values plot shows that error terms does not follow homoskedasticity assumption of linearity which makes the model unfit for prediction.

The box plot of arrival delay showed that there were 86 outliers records. Therefore, Data is cleaned by removing these outlier records and regression model is again built using logarithm scale of arrival delay (log ARRIVAL_DELAY).

Regression model with response variable transformation:

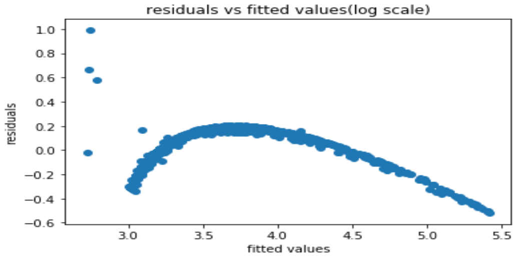
Regression model is built by taking log of arrival delay as response variable and independent variables which found significant in previous model i.e. aircraft delay, airline delay, air system delay, weather delay, departure delay. Regression results shows the R-square value of 0.917 and p-value of all variables less than 0.05 making them significant for prediction. However, residual vs fitted plot shows error terms making curve pattern spread mostly below zero line. This shows that the model deviates from the homoskedasticity assumption of linearity. Also, qq plot shows that it is improved as compared to earlier model but still there is deviation from normality assumption of error terms.

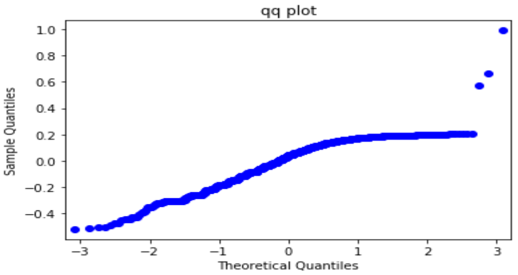
Regression Coefficients

```
const                    2.707571
LATE_AIRCRAFT_DELAY      0.019111
AIRLINE_DELAY            0.019317
AIR_SYSTEM_DELAY         0.020272
WEATHER_DELAY            0.019478
DEPARTURE_DELAY          0.000854
```

P Values

```
const                     0.000000e+00
LATE_AIRCRAFT_DELAY      1.781863e-209
AIRLINE_DELAY            4.990839e-204
AIR_SYSTEM_DELAY          0.000000e+00
WEATHER_DELAY            1.746004e-127
DEPARTURE_DELAY           2.570338e-02
```





❖ **Conclusion**

Data analysis and visualization, Regression model and analysis of variance (ANOVA) are used to study effect of one or more independent variables to analyze flight delays. Still there is a lot to do to make better model for prediction of arrival delay more accurately. To improve model further, stepwise regression should be done to add significant variables and remove non-significant variables from the model. Also, from the data visualization we have seen that arrival delay has some pattern with time of day. It can be included in model to check if the model performance improves further. Day of week variable is an ordinal variable hence it should be converted to categorical variable and its dummy variable can be added in model to study the effect on model further. Indicator variables 'Cancelled' and 'Diverted' can also be used to check their impact on arrival delay.