Coursework Report

PROGRAMMING FOR DATA SCIENCE – ST2195

STUDENT ID - 200699684

Contents

Introduction

This dataset is about commercial flights within USA explaining arrival and departure details of flights in different airports (Source - The 2009 ASA Statistical Computing and Graphics Data Expo). Datasets of 2006 and 2007 years were taken to this analysis. Answers to all the five questions asked were done in python and R languages, this report explains how each question was approached and analyzed. All the R and Jupyter notebooks have been submitted to the VLE portal.

Data Cleaning Process

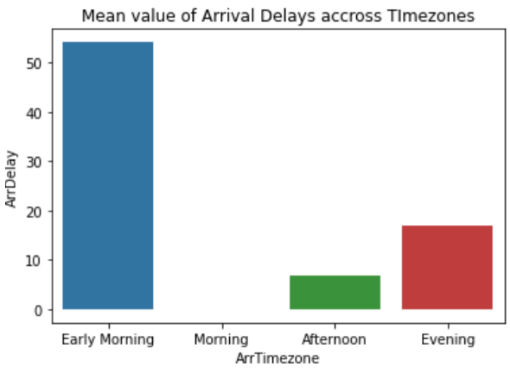
As an uncleaned data set would make inaccurate inferences and would provide wrong insights when analyzing, cleaning the data should be the first step in data analysis. Therefore the null values of the combined dataset were removed before starting answering each question. Also as there were some irrelevant data like ‘ArrTime’ has values like 2500, those were removed as well. Also as some data column values were not included in the meaningful way, such as ‘ArrTime’ values being 2200h,1900h, those were transformed as well. Not only that, but also data cleaning parts were done then and there when answering each question, when it was identified that a particular raw had null values.

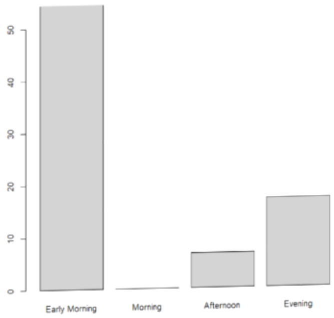
1. When is the best time of day, day of the week, and time of year to fly to minimize delays?

Best Time of Day

In order to answer this question, Arrival Time of flights (ArrTime) was categorized into four as Early Morning (00.00-06.00), Morning (06.00-12.00), Afternoon (12.00-18.00) and Evening (18.00-23.59). As the ‘ArrTime’ values were not in the time format, also as there were some entries like ‘2500’ which were irrelevant, those were converted into hour format and irrelevant entries were eliminated respectively.

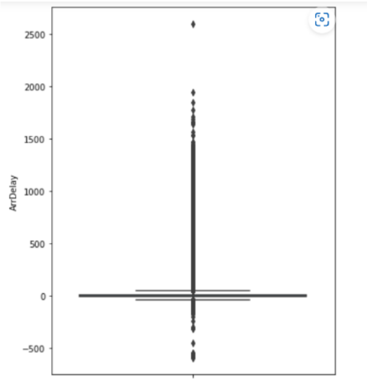
The mean Arrival Delay was calculated in each Time zone and findings were plotted in a bar graph. The following graphs were plotted.

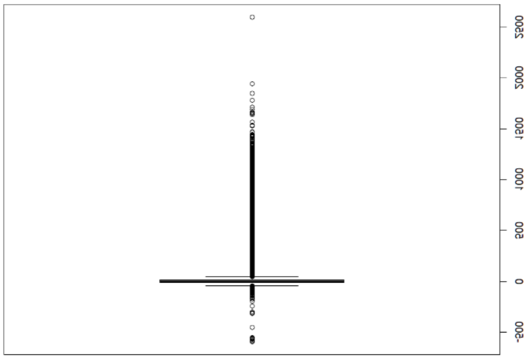


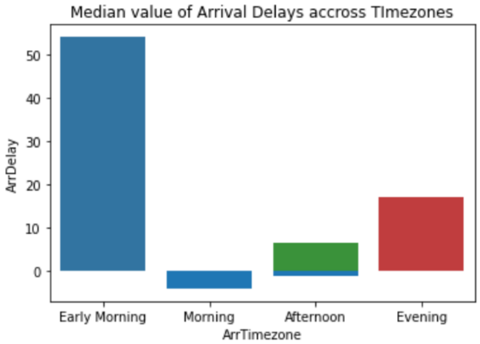


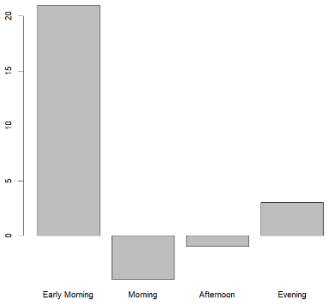
Python R

As outliers were obtained when checking the boxplot of the Arrival Delay entries, also as 'Mean' is sensitive for outliers, also as dropping outliers is not the best solution, Median of Arrival Delays was considered as well when answering this question. (Barbara Illowsky & OpenStax et al.) states that Median is the optimal central tendency when a few extreme scores in the distribution of the data, missing data exist.







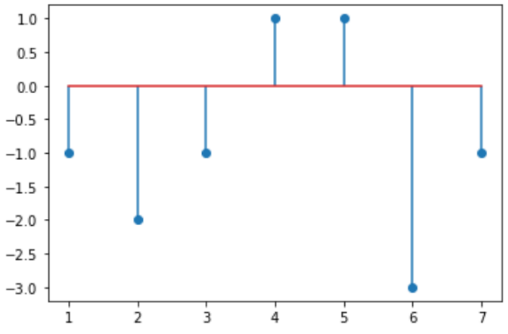


Python R

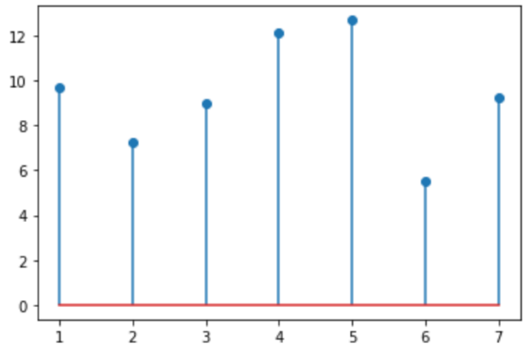
Therefore 'Morning' Timeslot, which means between 00.00 and 06.00 is the best time of day to minimize Arrival Delays.

Best Day of Week

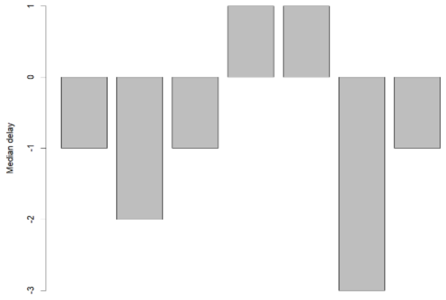
The approach to this question is almost similar as above, except ‘DayofWeek’ column was used instead of ‘ArrTime’ column. Bar graphs were plotted for mean Arrival delay and for its median value. The findings are attached below.

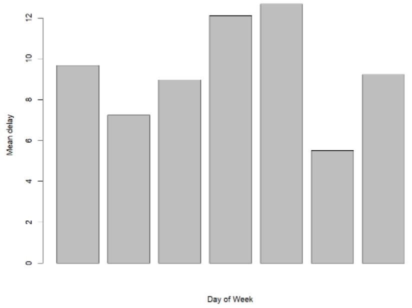


Median Values



Mean Values

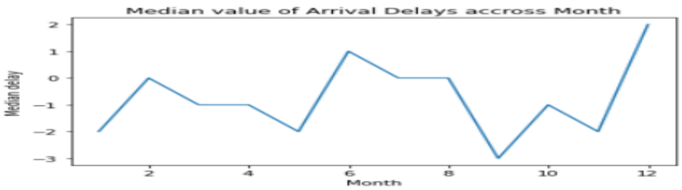


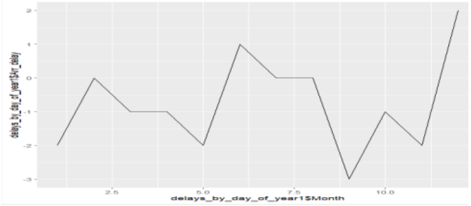


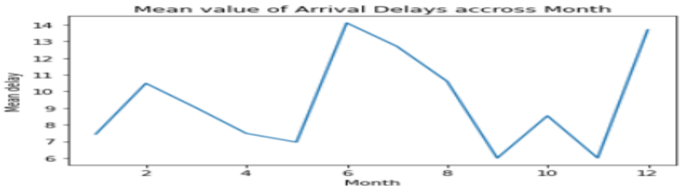
Therefore, Saturday (Mentioned by ‘6’ in graphs) is considered as the best day of week in order to minimize delays.

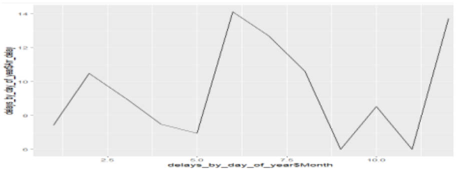
Best Month of Year

Approach to this question is similar to previous questions as well, ‘DayofWeek’ column was used instead of ‘ArrTime’ column. Line graphs were plotted for mean Arrival delay and for its median value. The findings are attached below







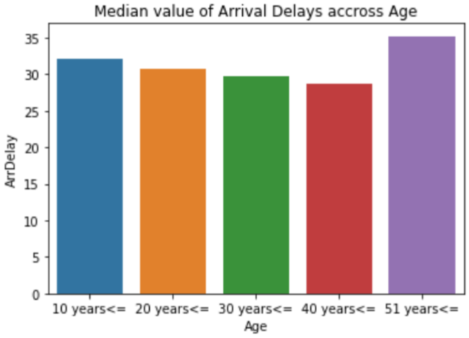


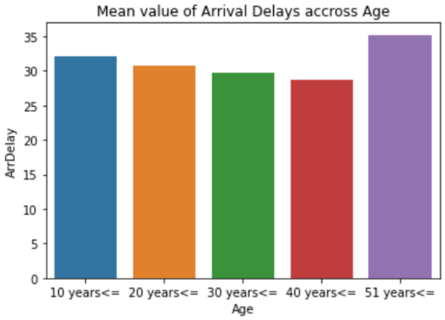
By considering the above graphs, we can identify that November and September have the lowest mean Arrival Delay among all other months, while September has a minor higher value compared to November. But when considering the median values of them, September's median value is lower than November's median value. As the data set has Null Values, and Outliers are present, Median value will be considered as its not sensitive for Outliers, therefore September is the best month of the year to minimize delays.

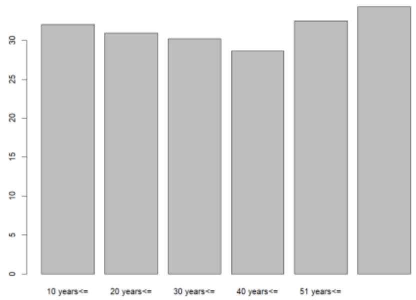
2. Do Older Planes Suffer More Delays?

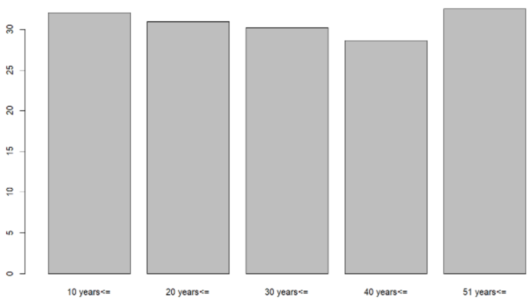
The dataset named 'plane-data.csv' was used along with the main dataset in order to get the year of manufacture of planes. Both the datasets have a common column which was the plane tail number, but as the column names of these two columns of were different, the header of one was changed as 'tailnum', then the two datasets were merged afterwards. After removing null values, getting the Arrival Delays which has value greater than zero, the age of each plane was calculated through the difference between the 'Year' and 'Manufactured Year', and a new column named 'Age' was then included.

Then Age values were categorized into 5 groups as '10 years<=', '20 years<=', '30 years<=', '40 years<=', '50 years<='. Mean, Median values of Arrival Delay of each category was then calculated in order to find whether there is a relationship between Age and Arrival Delays or not. Findings are mentioned below.

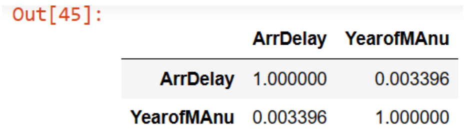








According to above graphs, we cannot identify that older planes do suffer Arrival Delays.(Both Mean and Median values), in order to clarify this doubt the correlation between ‘Year of Manufacture’ and ‘Arrival Delay was considered. As the correlation value is not significant,closer to zero, we can come up with a conclusion that older planes do not suffer Arrival Delays.


Out[45]:

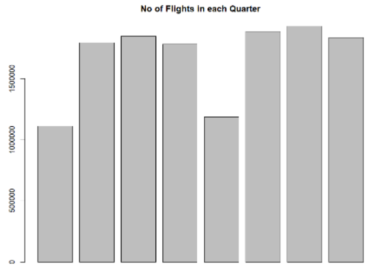
	ArrDelay	YearofMAnu
ArrDelay	1.000000	0.003396
YearofMAnu	0.003396	1.000000

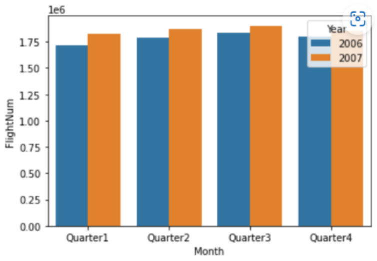
3. How does the number of people flying between different locations change over

Time?

Assumption – As there is no data about number of passengers, the number of flights (FlightNum) was considered as the number of passengers.

The 12 months of each year were categorized quarterly vise in order to find the relationship between time and number of flights, then the Flight Number count was taken in order to count the number of flights conceded in quarter. After removing all the null values and irrelevant entries for this question, bar graphs were drawn in order to find the pattern of number of flights in each quarter of the two years.

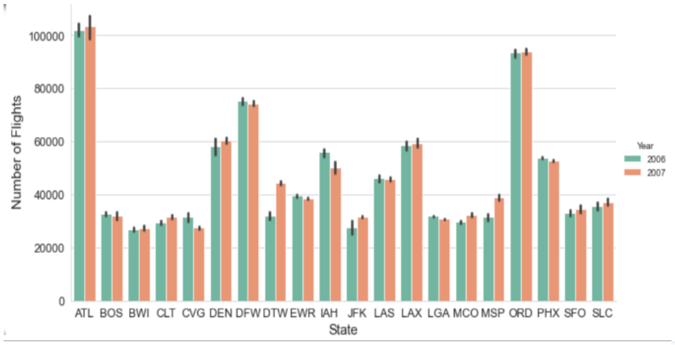


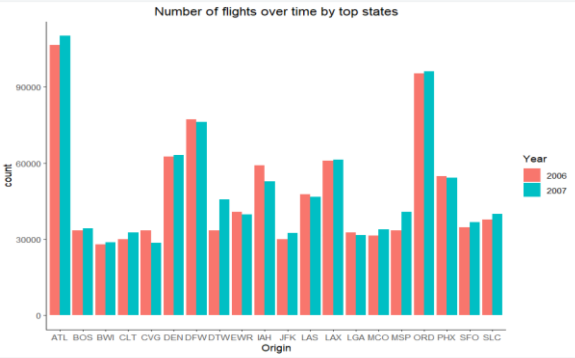


These bar graphs show that there is an increase of Fights from 2006 to 2007 in each year, which means more people have travelled over time. Quarter 3 seems to be the quarter with highest amount of Flights, this may due to the summer vacation as quarter 3 means from July to

September. In addition, Flights between different Airports (‘Origin’) was also considered in order to identify how the number of flights between different locations have been changed over time.

Assumption – As there can be flights between two Airports in the same state, Airport (Origin) has been considered instead of state. The top 20 Airports with most number of flights were calculated and then the relationship between flight counts with time was considered in order to identify whether people have travelled more in different airports over time or not. Findings are as follows:

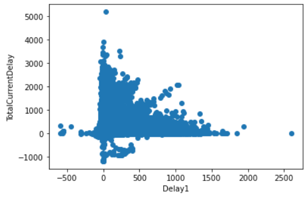




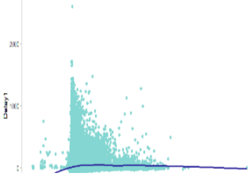
According to the graphs, we can say that there is a significant increase of number of flights in 2007 compared to 2006 in airports DTW, JFK, MSP, while most of other airports have similar amounts of flights in both years, some airports have a lower value in 2007 compared to 2006

As a conclusion, we can say that the number of flights have been consistent, slightly increased over time in between different locations over time.

4. Can you detect Cascading Delays in one airport create delays in others?



(Python plot)



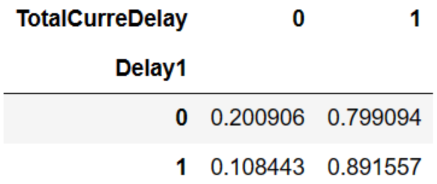
(R plot)

Firstly, the ‘scheduled departure times, CRSDepTime’ column was converted into Date-Time type. So then we can identify the flights occurred according to ascending order of time. Then the data set was grouped by tailnumber as it was easy to get in time order. Then a new variable named ‘TotalCurrentDelay’ was created adding the Arrival and Departure delays.

.In order to find the relationship of cascading delay and other delays (whether cascade delays do create delays in others) the Arrival delay variable was lagged by one period, named by ‘Delay1’. Then a scatter plot was drawn and correlation between the two variable was calculated.

The correlation coefficient value was 0.38 which is significant but not that strong enough to come up with a conclusion.

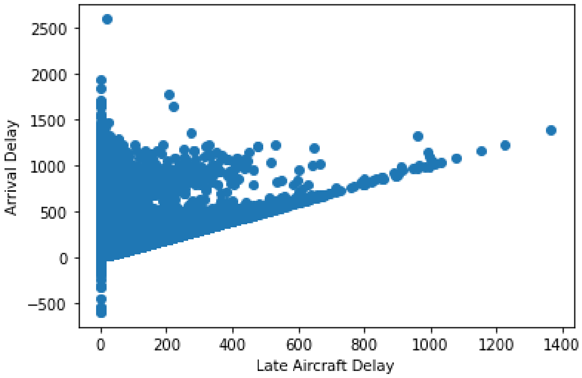
As the scatterplot also does not show a strong relationship a cross tabulation was done in order to clarify whether there is a real relationship between Totalcurrent delay and Delay1Delay.

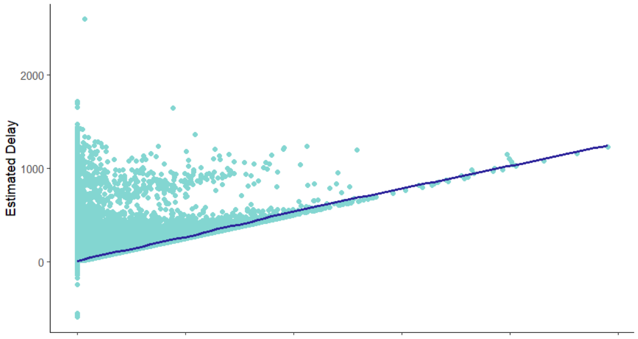


TotalCurreDelay	0	1
Delay1		
0	0.200906	0.799094
1	0.108443	0.891557

The cross tabulation indicates ‘Totalcuren’t delay as 0 and Delay1 as 1.This interprets conditional probabilities. Given that there is a delay in the previous airport there is a probability of 0.69 having a delay in the next airport which is a significant value.0.36 means that given that there is no delay in the previous airport there will be a delay in the next airport with a probability of that amount.

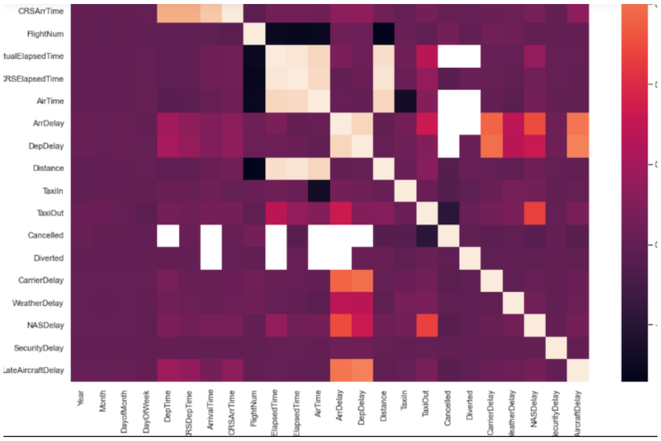
As the second approach to the question, another scatterplot was drawn using Lateaircraftdelay and ArrDelay being x axis and y axis. These two variables have a correlation coefficient value of 0.597 which is also significant. Therefore as per the second scatterplot and its correlation value and the cross tabulation value, we can say that there is a significant impact of cascading delays creating delays in others

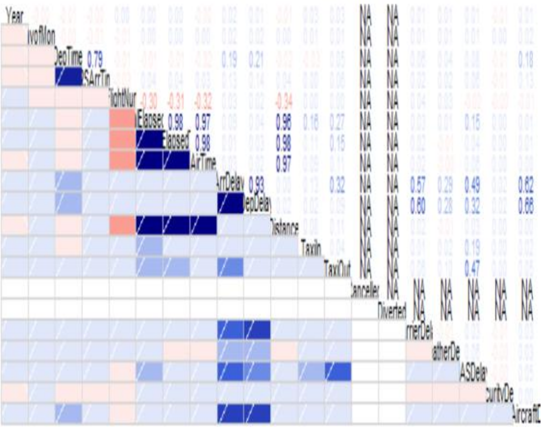




5. Use the available variables to construct a model that predicts delays.

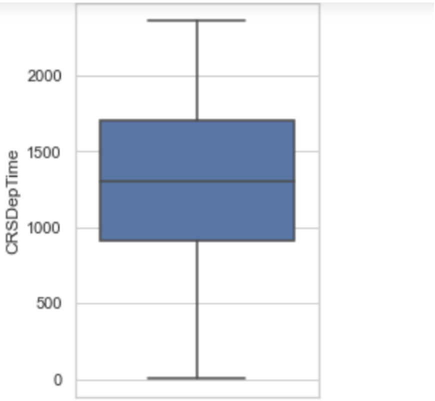
A multiple linear regression model, decision tree regressor and Random forrest regressor have been chosen in order to predict Arrival delays. After doing the necessary data cleaning, encoding the categorical variables as dummy variables, a sample of 65,000 was taken in order to run the model. The correlation matrix, heat map were created to identify the required variables for model.

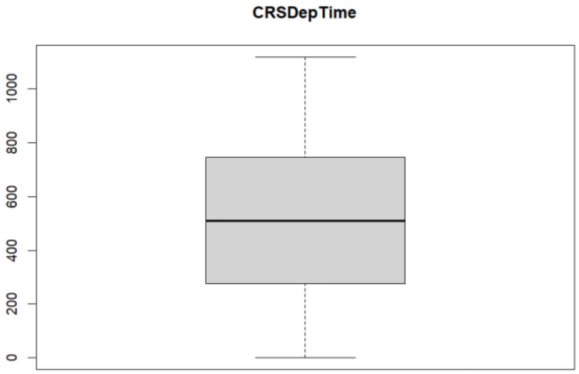




Python (Lower Triangle –R)

The independent variables that were highly correlated were removed from the model as if not multicollinearity would exist, which would violate all the assumptions of the linear regression problem as the residuals would not follow a normal distribution. Then the dataset was split into train and test data sets in order to run the model on the train set. Before running the model on train set values, Boxplots were created in order to identify whether outliers would be exist or not.





Outliers were found in DepDelay, CRSElapsedTime variables, therefore those outliers were removed from the dataset.
Therefore the expected equation of the model was,

Yi= B0 +B1*Month + B1*DayOfWeek +B2*CRSDepTime + B3*CRSArrTime+ B4*DepDelay +B5*CRSElapsedTime+ ei

Where Yi is the dependent variable, B0 is the intercept value and B1,, B5 are the coefficients of the variables and 'ei' is the error term.
Then the model was run on the test dataset independent variables subset (X_train) , model coefficient value, model intercept value were calculated. Therefore the equation of the final model is:

Yi= B0 +0.190427*Month -0.216432 *DayOfWeek -0.084965*CRSDepTime -0.336442 *CRSArrTime+ 23.362208 *DepDelay +1.526029*CRSElapsedTime+ ei.

The following assumptions were made prior to fit the model into the dataset.
1.There exists a linear relationship between the dependent variable and independent variables.
2.Error terms have constant variance

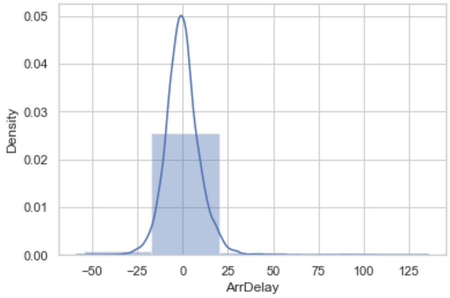
3.Error terms are normally distributed
4.There exist no linear relationship among independent variables

Output of the model were as follows

```
RMSE: 2.6515658580916015
R squared:  0.827
```

The R squared 0.872 means that 87.2% of the variation in the Arrival delay has been explained by the independent variables of the model.

RMSE measures the average difference between the predicted values and the original values.



The graph in the left explains the distribution of residuals. By looking at the graph we can realize that the residuals have followed a normal distribution which means the model is valid for prediction.

Also Decision Tree Regressor and Random Forrest Regressor were fitted to the dataset. The results of all three models were as follows

Model Name	R Squared Value	RMSE Value
Multiple Linear Regression	0.827	2.67
Random Forrest Regression	0.827	2.73
Decision Tree Regression	0.827	2.72

All the three models have the same level of performance, as the RMSE value of MLR model is the lowest, it suits this dataset well, compared to other two models.

Bibliography

Barbara Illowsky & OpenStax et al. (n.d.). Introduction to statistics. Retrieved April 3, 2023, from https://courses.lumenlearning.com/introstats1/chapter/when-to-use-each-measure-of-central-tendency/