**Airline Seat Utilization Capstone Project Report**

**Problem to solve**

**Summary: Predicting the number of empty seats by month ahead of time can help airlines to take actions for maximizing reservation / reducing the planned number of flights.**

In the airline industry where the profit margins are usually razor-thin, ensuring that there are no empty seats on every flight is very important. Flights where there are many empty seats hurt the airline's bottom line as the costs to operate the flight is pretty high. Predicting the percentage of empty seats of a flight is critical as the airlines could then act on the data to maximize profits.

This can depend on a variety of factors like brand reputation, pricing of the different classes of seats (first, business, economy etc.) and month of the year as demand would be higher in certain months due to holidays or other factors. The goal of this project would be to focus on predicting the demand based on the month of the year. We will also look at potential impacts of the revenue of an airline on the seat utilization.

**Data**

The data for this project is sourced from Bureau of Transportation Statistics (BTS). This has data of the number of seats , passengers and departures for hundreds of airlines on a monthly basis. So we can calculate the total number of empty seats for an airline for that month. This data set also has other details like the number of departures, so we can find out the average seat utilization by flight as well. Also this data set has data for every month starting October 0f 2018 to January of 2008 (10 years).

Reference: <https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=293&DB_Short_Name=Air%20Carriers>

We also sourced the financial information from the BTS which was available for a subset of the airlines. This data was based on every quarter in a year and one file is generated per year. We will be matching the financial data with the main data set.

Reference: <https://www.transtats.bts.gov/DL_SelectFields.asp>

**Approach**

The first step of this project would be to gather all the source data from the website. The Bureau of Transportation Services generates one file of data for every year. The financial data is also generated in a separate file for every year. All the files of both data sets needs to be imported and cleaned and linked to form one master dataset.

**Information about the dataset:**

Details of the different variables of both datasets are listed and described below.

**Seat Utilization Dataset:**

|  |  |  |
| --- | --- | --- |
| **Field** | **Data Type** | **Description** |
| DEPARTURES\_SCHEDULED | Number | Number of planned departures |
| DEPARTURES\_PERFORMED | Number | Number of actual departures |
| SEATS | Number | Total number of seats across all the departed flights |
| PASSENGERS | Number | Total number of passengers across all the departed flights |
| UNIQUE\_CARRIER | String | Unique Carrier Code for each airline |
| UNIQUE\_CARRIER\_NAME | String | Name of the airline |
| ORIGIN\_COUNTRY | String | Origin Country of the flights |
| DEST\_COUNTRY | String | Destination Country of the flights |
| YEAR | Number | Year in which the flight departed |
| MONTH | Number | Month in which the flight departed |

**Financial Dataset:**

|  |  |  |
| --- | --- | --- |
| **Field** | **Data Type** | **Description** |
| NET\_INCOME | Number | Net Income for the Year |
| OP\_PROFIT\_LOSS | Number | Total Profit/Loss for the Year |
| OP\_REVENUES | Number | Total Revenue for the Year |
| OP\_EXPENSES | Number | Total Expenses for the Year |
| UNIQUE\_CARRIER | String | Unique Carrier Code for each airline |
| UNIQUE\_CARRIER\_NAME | String | Name of the airline |
| YEAR | Number | Year for the financial information |

**Importing Libraries:**

We import the dplyr and the formattable libraries that we will be using with the following import statements.



**Importing Data Files and Data Cleansing:**

Every financial file is first imported and cleaned with a custom function below.

Details of the custom financial function:

* Each year’s financial file is first read with the read.csv function as the file is in a CSV format.
  + Setting the na.strings to some other value as we had data in the file referring to North American Airlines that had NA has the value.
* As the financial data is by quarter, we sum up each of the financial columns by grouping on the carrier and year.
* As we need the numbers on a monthly basis, we create new columns for the monthly numbers by computing the average of the yearly numbers.



This function is called for every finance dataset file:



Details of the custom function to import and clean the Seat Utilization dataset:

* The data set files is read using the read.csv
* The focus is on only Domestic US airlines, so all the International flights are removed.
* Group\_by and Summarise are used to get the total number of seats , passengers and departures grouped by airline and month.
* Any rows with zero Seats are deleted as these would outliers to the data.
* All the carriers that do not have data atleast for 12 months need to be removed as this can skew the dataset if there are quite a number of smaller airlines with very little data.
* Both data sets for a particular year are then merged with the Unique Carrier Key as the backup



The custom function for each year’s files are called as follows along with the financial dataFrame that we covered yearlier:



Then we can combine all these resulting dataFrames to form 1 master dataFrame,



We then calculate the seat utilization for all the carriers



To calculate the seat utilization of all airlines by month and year, we create a new dataframe that computes the mean of seat utilization across all carriers by month and year

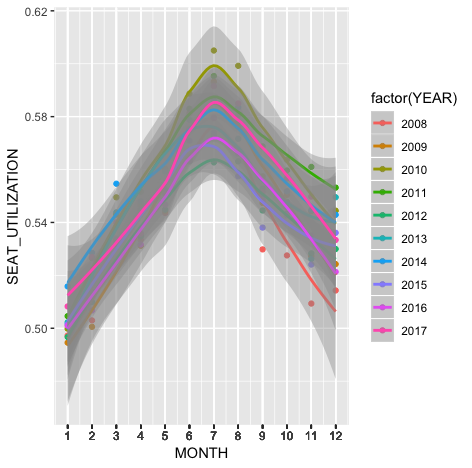


**Exploratory Data Analysis (EDA)**

**Overall Industry Trends:**

To observe the monthly trends across the entire US domestic airline industry (151 airlines), I aggregated the seat utilization by month and year and plotted it by month against the seat utilization.





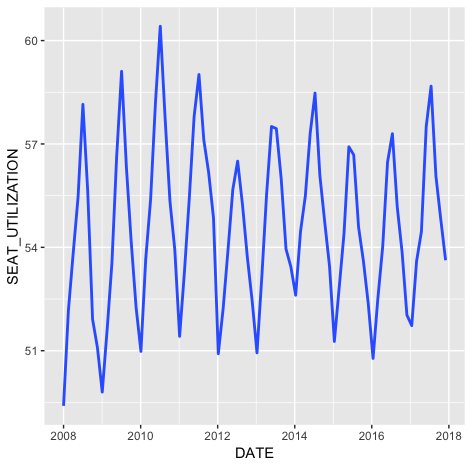
As we can infer from this plot, across all years on average, the months of June, July and August had the best seat utilization and hence the lowest number of empty seats.

January had the most number of empty seats. Also we see that the trends have been pretty consistent over the years with the peaks being in the same months.

Next, a Time Series plot of the data with date on the X axis and Seat Utilization on the Y axis shows us the same trends as well.

The beginning and end of each year comparatively had the most number of empty seats (lowest seat utilization), whereas the middle of the year had the least number of empty seats (maximum seat utilization)

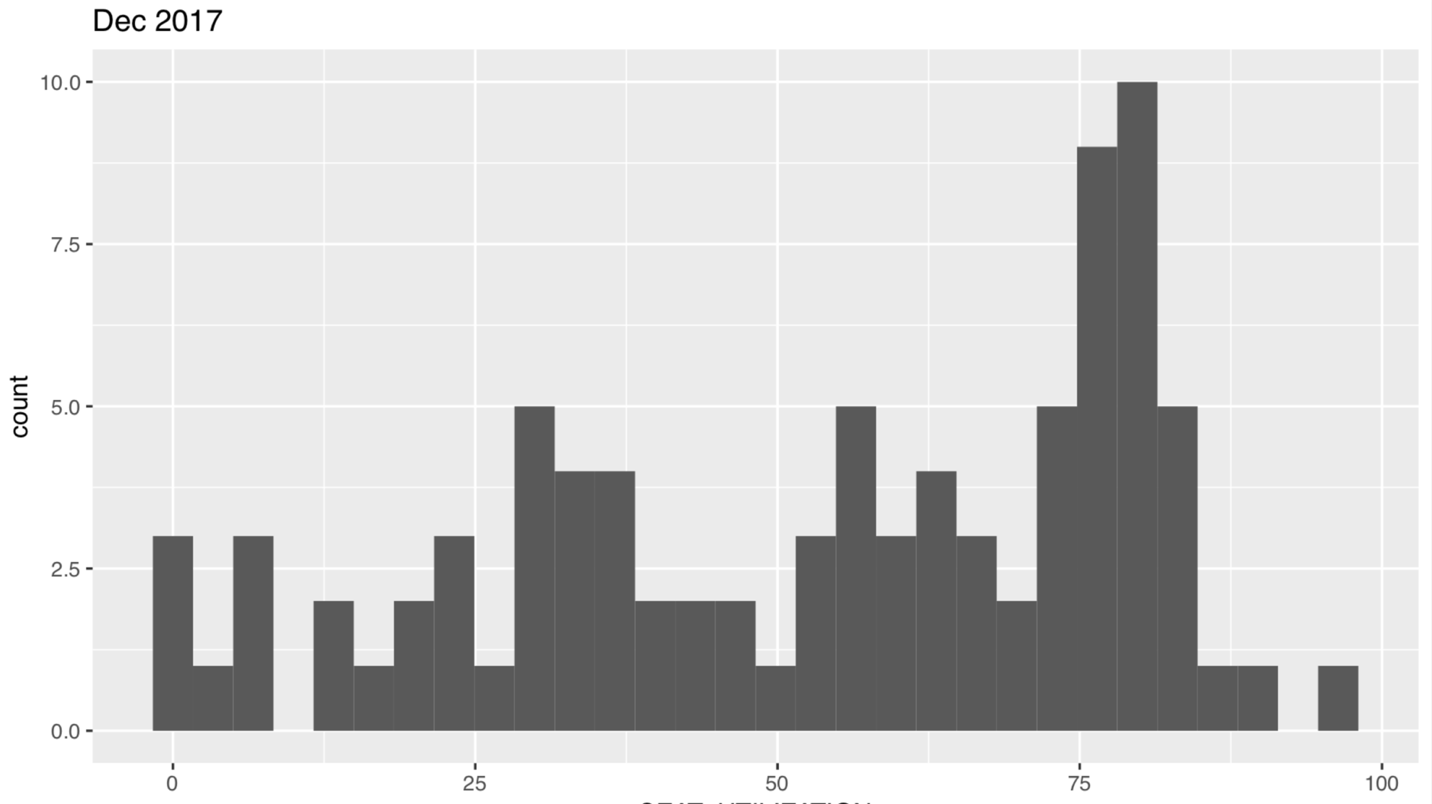


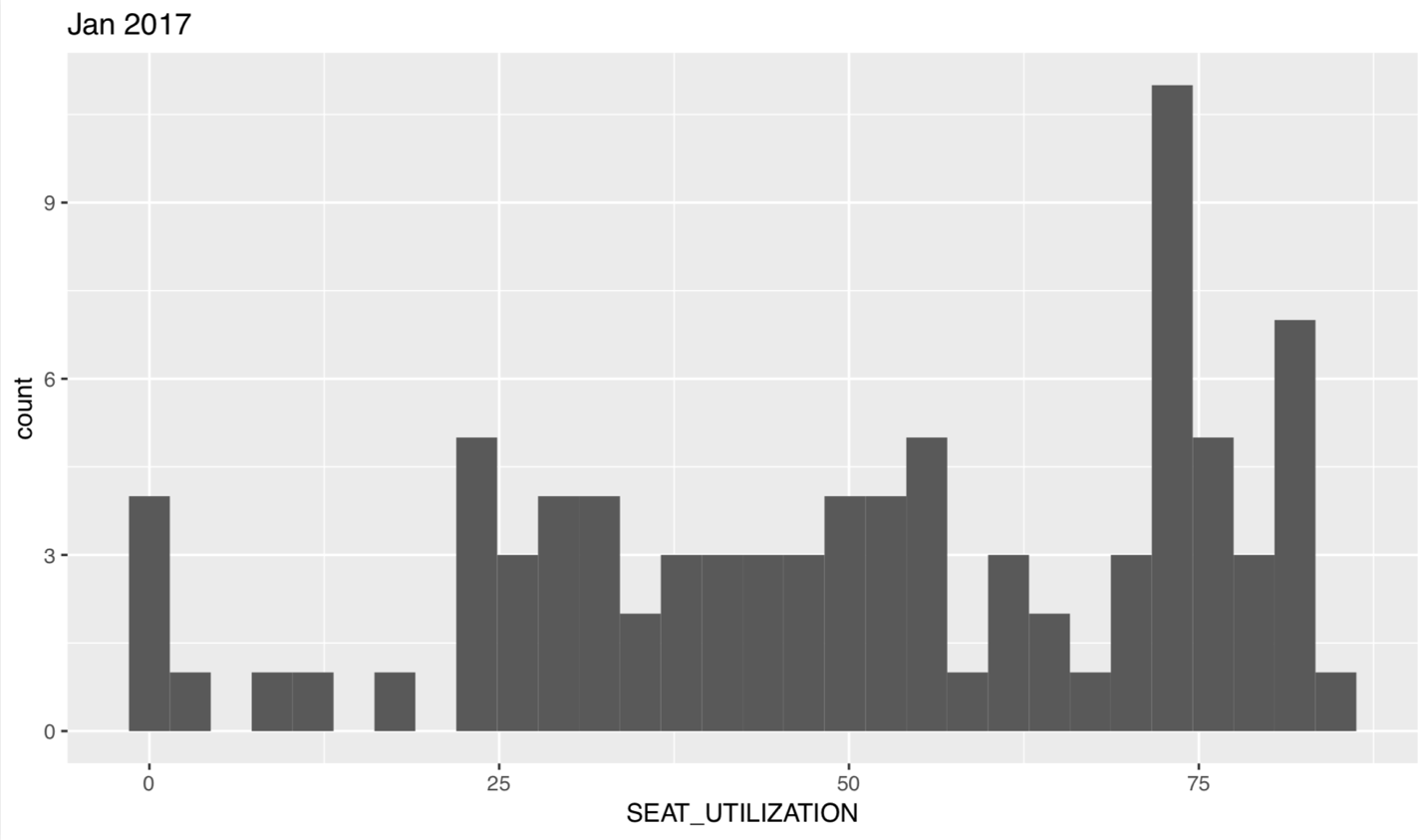
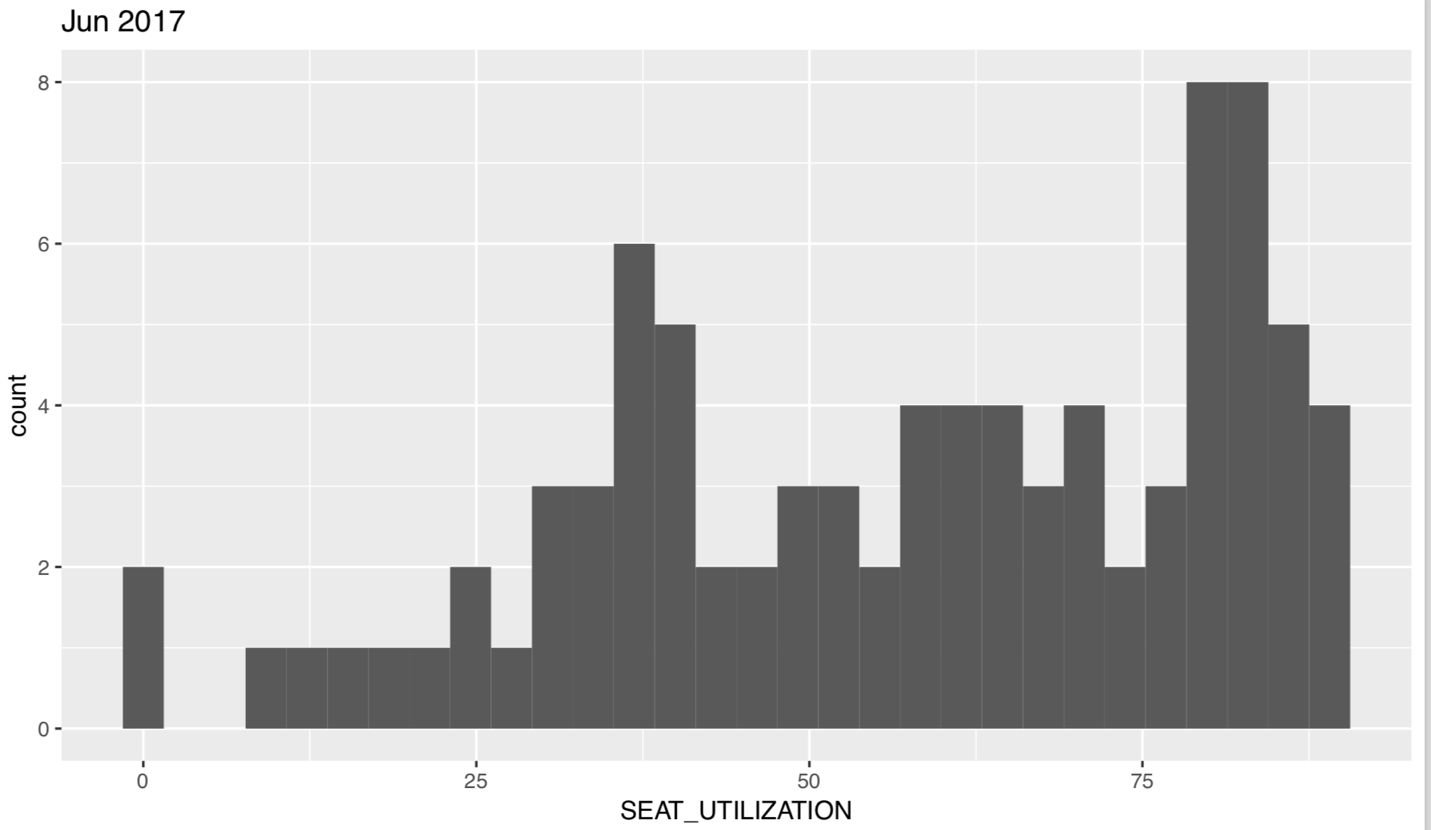


**Monthly Trends By Airline:**

A Histogram plot by month and year would be a good way of representing the number of airlines in each range of seat utilization in every month across every year. These plots are very descriptive of the trends we saw at the overall industry level. Below is the code and a few sample plots of different months.



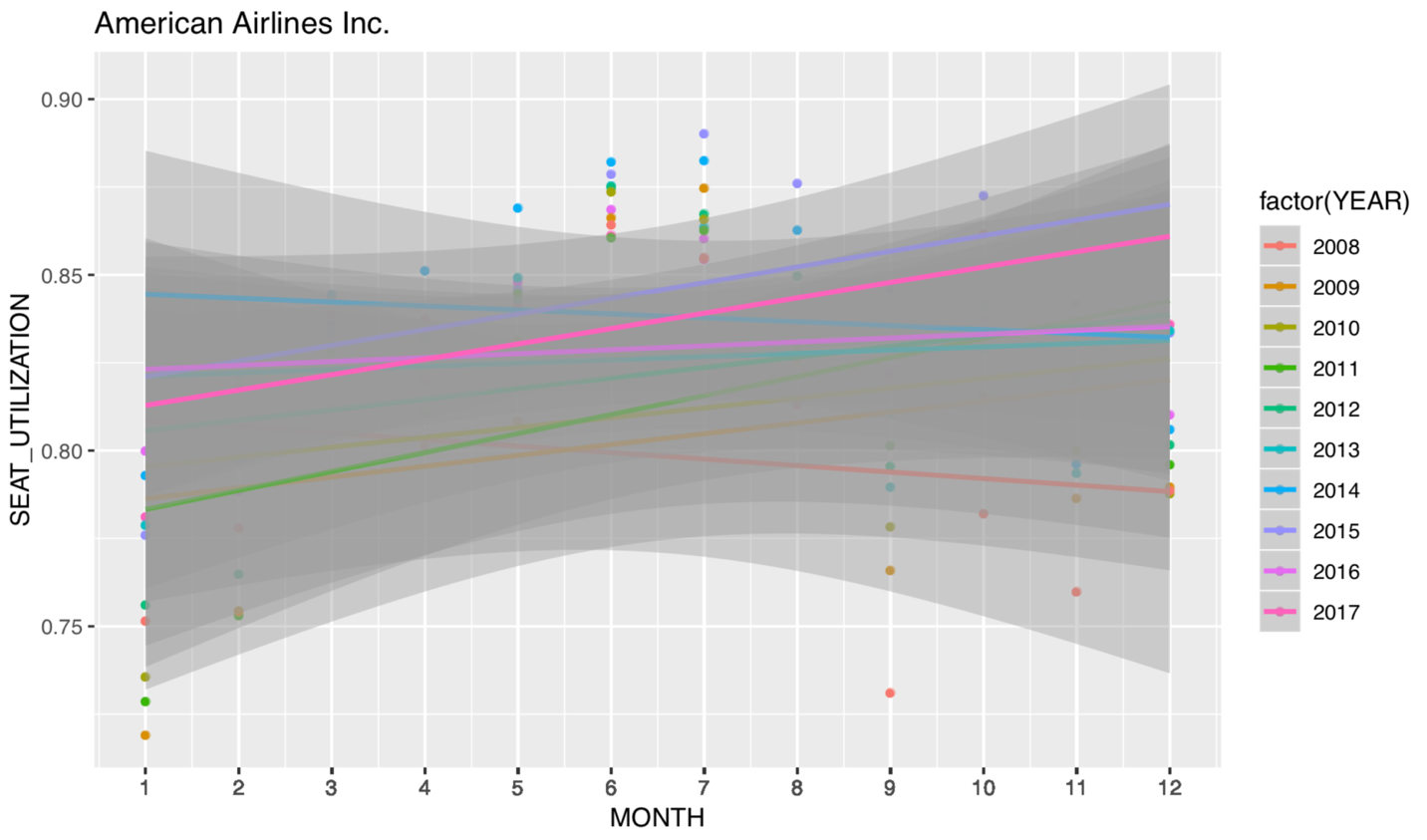




For 2017, Looking at the number of airlines that had greater than 75% seat utilization, June was the best month, while January and December had lower seat utilization.

Every airline was also individually plotted by month and seat utilization, with geom\_smooth, an upward slope was noticed throughout the year with a lot of high data points in the middle of the year.





There were also airlines with straight or downward slopes as well, majority of those airlines were lesser known airlines with comparatively lesser number of flights.

Overall, similar monthly trends were observed across the different approaches.

**Modelling Techniques**

The goal for us is to predict the number of empty seats for a given airline which can be inferred by predicting the seat utilization.

As the input data has all the required variables for the computation and also has the output variable (Seat Utilization), We classified this problem as a Supervised regression problem as the output variable (Seat Utilization) is not a category as well.

**Features:**

In this case, Seat Utilization would be the dependent variable.

The independent feature variables are:

* Month and Year
* Date
* Departures
* Monthly Revenues

**Linear Regression:**

We are using the Linear Regression machine learning technique as this is best suited for the problem and the data at hand. Multiple linear regression allows us to use all of our variables to improve our predictive ability.

We will measure the performance of the model by using the R-squared value where a R-squared value close to 1 indicates a good model.

**Approach:**

We use the recipes package to build up the data frame that can be used in any modelling technique.

We created dummy variable columns for each airline (151 airlines), so that each of these columns can be treated as independent variables.

To be able to predict airlines that are not in the data set, we introduce an “OTHER” airline:



We now use the lm function to perform linear regression analysis.

First, we use the lm function with seat utilization against all the dummy variable columns except the seats, date and Monthly revenue columns.



Which gets us an R squared value of 0.8694



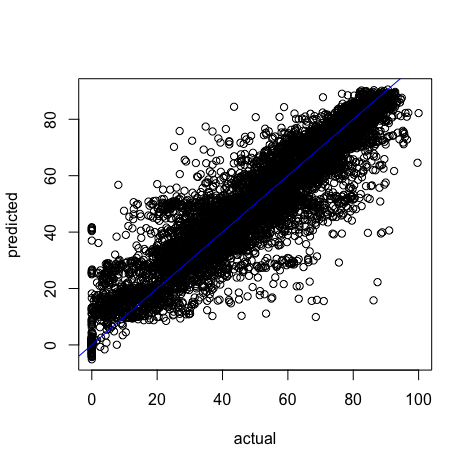
lm function with all the columns including revenue



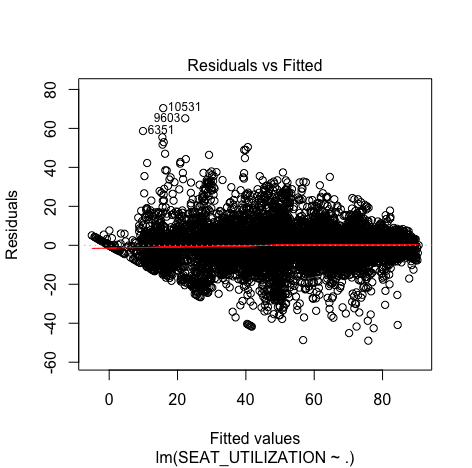
R squared value is 0.8726



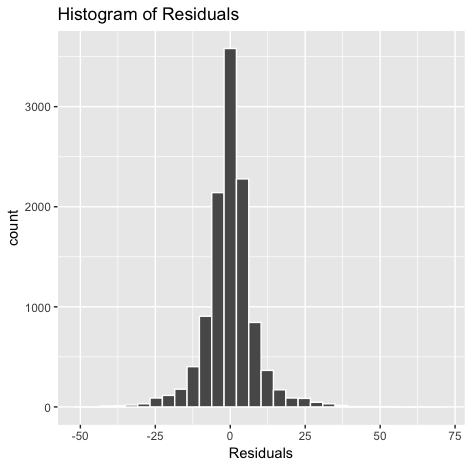




**Scatter plot** depicting the predicted values against the actual values.

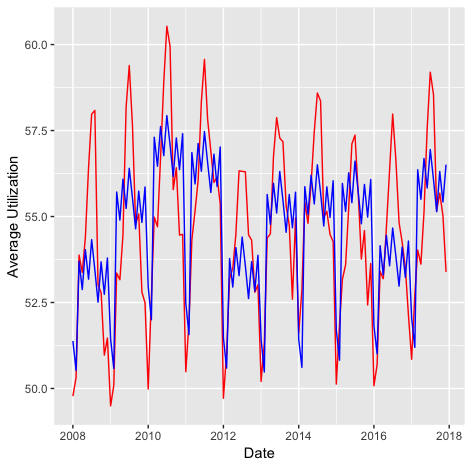


**Scatter plot** of the model depicting the fitted values against the residuals. This shows that most of the predicted values have minimal residual values supporting the accuracy of the model.



**Histogram of Residuals** also shows that majority of the data has zero residual values.





Comparison of the actual (blue) vs predicted (red) average utilization over time.

**Summary**

**Further Research Ideas:**

* We can look at other features like
  + departure times of the flights (early morning, red eye etc.)
  + Brand recognition of the airlines
  + Size of the aircraft
  + Transit times
  + Ticket price
* We can extend our research to international airlines as well
* We can also look at the different passenger classes (Economy, Business etc.)

**Recommendations:**

Based on the number of empty seats predicted, the airline can:

* Reduce the number of flights in the months where the demand is low.
* If the number of empty seats predicted is higher than the industry average then the airline can take measures by running promotions etc.