Caused of Delay in Airline

Abstract

Delays are a significant problem in air transport. Delays affect both the costs and the satisfaction of the passengers. In order to determine which factors best predict flight delays, this study examines and reports on the many properties and features of the Airline Delay Caused dataset. It also looks into the characteristics' complexity to see whether there is an underlying structure that affects how complicated the dataset is. This research will utilize analysis to show the strategies employed to help the airline reach out to its system in new ways. We will use six factors to help us achieve the goal of reducing the dataset. The methods utilized in this work include principle component analysis (PCA), multiple linear regression (MLR) and K-Nearest Neighbour (KNN). We may compare each of the outputs in respect to this dataset using these algorithms to decide which approach is the most effective in this case study. According to the PCA results, this investigation has also shown that *security_ct* has the least impact on dependent variable. We concluded that the independent variables are the common components that are impacting delay fighting based on the fact that multiple linear regression provided a 100% accuracy in prediction for this study. Thus, the organization in the USA may use this study to concentrate on improving the elements that affect the duration of their arriving flights.

Introduction

Travelling via aeroplane, jet aircraft, and other similar vehicles is known as air travel. In the decade between the middle of the 1980s and the year 2000, the global use of air travel doubled. Since air travel is faster and less accident-prone than other forms of transportation, this generation tends to favour it. Most domestic destinations may be reached by air travel in a few hours; foreign travel almost never takes longer than 24 hours. 4.3 million people work for airlines and airports worldwide, making up the aviation sector, which generates 5 million direct employment worldwide and adds over US$ 275 billion to the global GDP (Group, 2004).

Although the majority of aircraft arrive on time, unexpected delays do occur, which are both terrible and unavoidable. According to (Sheffield School of Aeronautics, 2022), any flight in the United States that departs or arrives at the airport after 15 minutes of its scheduled time is considered late. Misleading flight delay estimations will result in damages for aircraft sectors and travellers, since delays will damage the transportation channel's functionality and cause disruptions at other airport terminals (Wang, 2022). The objective of this study is to identify the common factors that affect the flight delay.

Data

This study was integrated with the synchronisation of several machine learning algorithms in order to establish an adaptable strategy that would provide more knowledge and insight into current and future airline delay status optimization. Machine learning software such as Python, R studio, and Excel are being applied to obtain data information across the raw data process with massive amounts of data acquired. This study employs the characteristics of three machine learning algorithms to filter, alter, organise, and manage data in order to provide insightful information about airline delay causes.

The date utilised in this study was obtained from the website Kaggle which was provided by the author RyanJT (2022) in August 2021. It was downloaded in "csv" format and contained the folder entitled Airline Delay Caused.

The dataset consists of 70,695 rows with 21 variables. The definitions of all the attributes will be illustrated in appendices for detailed knowledge.

Data cleaning is the process of identifying incomplete, inaccurate, and inappropriate data. It is a phase for fixing any errors and inconsistencies that are found to improve the quality (Fakhitah, Mohd, & Zainon, 2019). Hence, the na.omit method was used to remove from all rows with NA data.

Methods

RStudio Team (2022). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

In this study, RStudio is being used to process data, analyze statistics, and visualize data. It provides a toolkit for both simple and advanced statistical procedures in order to create the conditional expressions, iterative operations, insight directions, and built-in technicians for estimations utilizing arrays and matrices. The research utilises various alternative methodologies to determine the objective.

In a wide range of prediction-type situations, K-Nearest Neighbour (KNN), is acknowledged as being appealing straightforward, intuitive, and adaptable for use in a variety of application fields as it is a method that can achieve high accuracy. In contrast, Principal component analysis (PCA) is a technique for reducing the number of dimensions in big data sets by condensing a large collection of variables into a smaller set that retains the majority of the large set's information. Finally, multiple linear regression algorithms, which are frequently used to illuminate the association between a continuous dependent variable and two or more independent variables. For all methods and codes, it may be discovered in the appendices.

Results and Discussion

Innovative insights are a subset of predictive analytics, which uses historical data together with analytical modelling, data mining techniques, and machine learning to predict prospective outcomes (Tucci, 2021). In this section, the insights of this study will be illustrated by various machine learning approaches to identify trends and anticipate probable events.

```
##     classifier_knn
##        9E   AA   AS   B6   DL   EV   F9   FL   HA   MQ   NK   OO   UA   US   VX
##   9E    1    4   23    6   96  142   30    0    1   37    4  100    2    4    0
##   AA    0  297    7   34  294  406   61    0    2  207    2  151   72   51    0
##   AS    0    7  189   26  369   58   75    0    0   54    9  331    5   47    0
##   B6    0   45   21  100  176  342   57    0    2   53    8  163   20   99    0
##   DL    0   77   54    4 1795  212   65    0   17  177    1  294   58   28    0
##   EV    0   32   27   13  270 1904  106    0   12  143    0  635   24    9    0
##   F9    0    5   62    6  176  153  493    0    1   13    2  233   13    5    0
##   FL    0    6    9    0  108  119   31    0    4   11    4  141    3    2    0
##   HA    0    2    0    4  132   28    0    0  104    0    0   11    0   10    0
##   MQ    0   87    2   44  204  312   41    0    1  853    0  306   10   11    0
##   NK    0    5   22   18   22   69   45    0    0    9   39   42    0   41    0
##   OO    0   62   47   36  495  807  164    0    4  233    0 1368   27   26    0
##   UA    0  127   16    9  413  309   61    0    4  101    2  287  169   23    0
##   US    0   32   42   65  265  159   54    0    2   19    6  129   16  176    0
##   VX    0   18    6    5   73   12    4    0    0  113    0   98    7    3    1
##   WN    0   91    0   22   80  251    2    0    1   24    0  116   15   16    0
##   YV    0    4   11   11  115  111   33    0    3   36    0  165    3   12    0
##     classifier_knn
##        WN   YV
##   9E    4    0
##   AA  133    0
##   AS   27    0
##   B6  105    0
##   DL   50    0
##   EV  114    0
##   F9    3    0
##   FL   16    0
##   HA   41    0
##   MQ   29    0
##   NK    3    0
##   OO  181    1
##   UA  129    0
##   US   11    0
##   VX   31    0
##   WN 1105    0
##   YV   10    0
```

Figure 1.0: Confusion Matrix

K-Nearest Neighbour (KNN)

As known, KNN approach is the most frequently employed technique within data mining research. In Figure 1.0, after the scaling of the dataset's test and train, it can be observed that 1904 data points were properly identified as EV on the carrier name. Hence, the carrier variable has shown the greatest accuracy classification in this testing. However, by setting k = √(total number of samples), is discovered that kNN's

accuracy is relatively poor (34.7%), especially when compared to other algorithms. This is because the selection of new data classes is based on a simple vote majority approach that ignores the closeness of data. This is undesirable when the distances between each closest neighbour and the test data are very different.

Principal Component Analysis (PCA)

```
airline_delay.new.table <- round(cor(airline_delay.new[8:13]), 3)
head(airline_delay.new.table)
```

```
##                   arr_del15 carrier_ct weather_ct nas_ct security_ct
## arr_del15             1.000      0.950      0.782  0.930       0.499
## carrier_ct            0.950      1.000      0.749  0.826       0.515
## weather_ct            0.782      0.749      1.000  0.717       0.378
## nas_ct                0.930      0.826      0.717  1.000       0.418
## security_ct           0.499      0.515      0.378  0.418       1.000
## late_aircraft_ct      0.962      0.904      0.724  0.816       0.489
##                   late_aircraft_ct
## arr_del15                    0.962
## carrier_ct                   0.904
## weather_ct                   0.724
## nas_ct                       0.816
## security_ct                  0.489
## late_aircraft_ct             1.000
```

Table 1.0 : Correlation Matrix

By creating this correlation matrix, it can be observed that *security_ct* and the other variables have a poor correlation (below 0.51). Accordingly, *carrier_ct* and *late_aircraft_ct* have correlations of 0.950 and 0.962, respectively, suggesting that both variables have a very strong link with *arr_del15*.

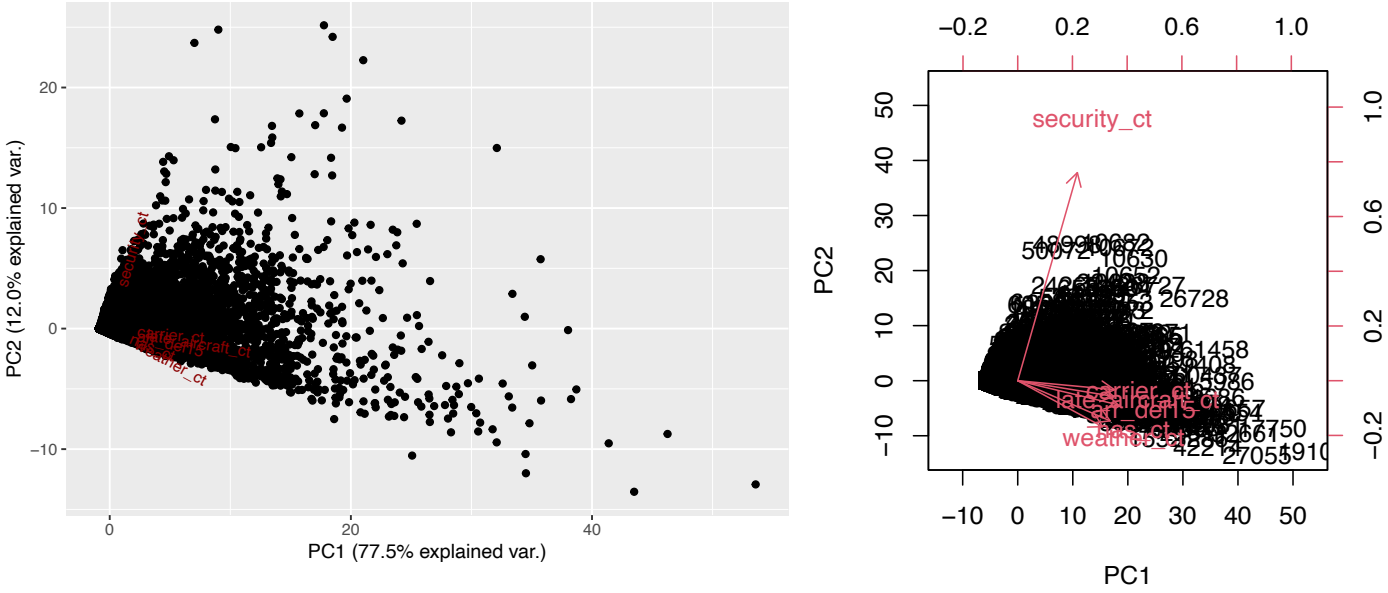


Figure 2.0: PCA ggbiplot (PCA bi-plot evaluating*arr_del15, carrier_ct,weather_ct, nas_ct, security_ct* and *late_aircraft_ct* for the first two principal components (PC2 vs PC1)

By plotting a graph, the link between variables becomes clearer. In Figure 2.0, PC1 corresponds for 77.5% of total variance, whereas PC2 accounts for 12%, with a combined proportion of 89.5% within both PCA. Comparing with the output above, the initial loading

vector shows almost identical weights for *arr_del15*, *carrier_ct*, *weather_ct*, *nas_ct*, and *late_aircraft_ct*, with much less weight for *security_ct.*

Multiple Linear Regression (MLR)

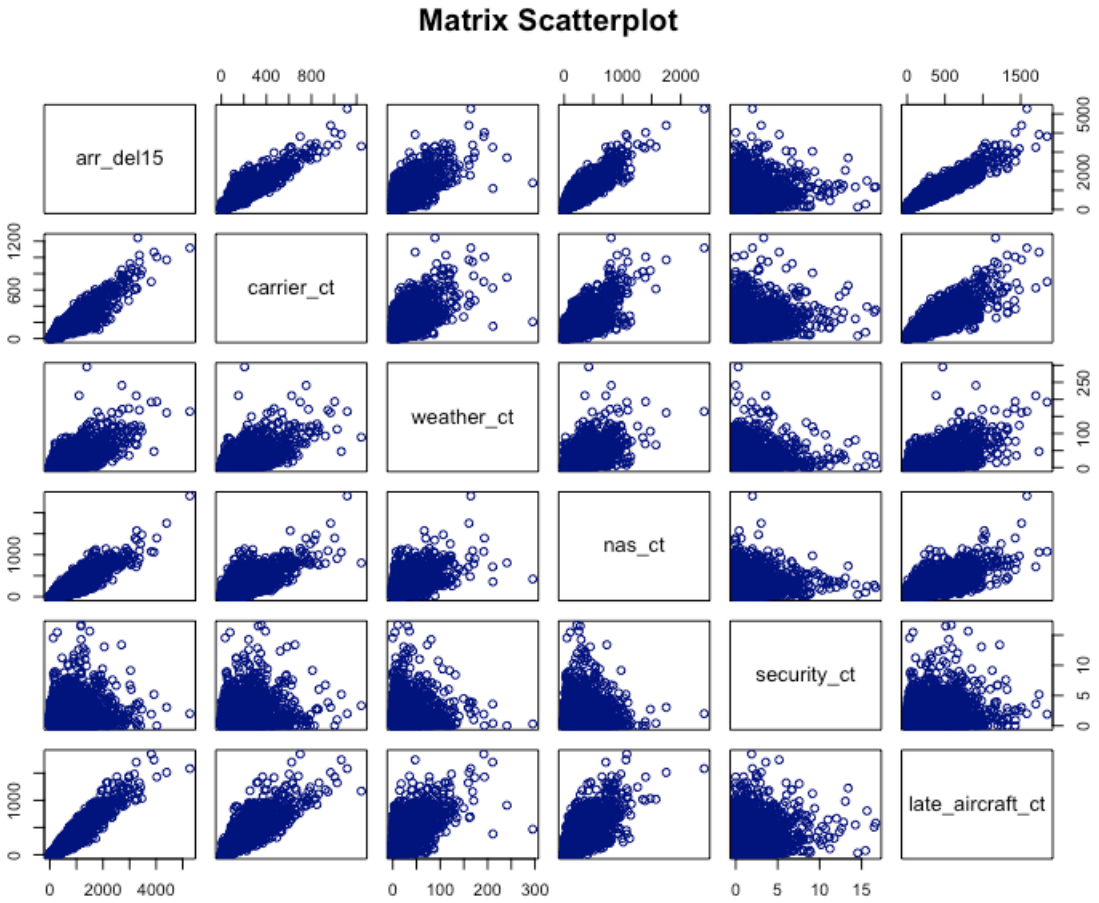


Figure 3.0: Matrix Scatterplot

As mentioned before, *arr_del15* is utilised as the dependent variable to assess the importance of the independent variables. Therefore, *arr_del15* is used as the y-axis in the first row of Figure 3.0, with the other variables on the x-axis. It can be seen that the variables of *carrier_ct*, *nas_ct*, and *late_aircraft_ct* indicate a perfect positive correlation; but, for *weather_ct*, the majority of the points are grouped at the bottom left and widely scattered, indicating outliers. However, *security_ct*'s results indicate that it has the least significant effect on the independent variable.

Looking at the output below, a regression model should be evaluated for a number of key factors, including R^2, significance F, variable coefficients, and p-value. As shown in Table 2.0, a regression model is used to determine the most frequent factor influencing the dependent variables, using the hypothesis of flight delays as the dependent variables. Since the median is nearly zero and the model is somewhat skewed to the left when looking at the min of -0.0202 and 1Q of 0.0000576, the result shows that the residuals produced a symmetrical output. Our null hypothesis states that *carrier_ct*, *weather_ct*, *nas_ct*, *late_aircraft_ct*, and *security_ct* all do not influence delayed flights and our alternative hypothesis states that *carrier_ct*, *weather_ct*, *nas_ct*, *late_aircraft_ct*, and *security_ct* all influence delayed flights.

```
##
## Call:
## lm(formula = arr_del15 ~ ., data = airline_delay.new[8:13])
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0202088  0.0000576  0.0000659  0.0000759  0.0206157
##
## Coefficients:
##                    Estimate Std. Error    t value Pr(>|t|)
## (Intercept)      -6.683e-05  2.168e-05 -3.083e+00  0.00205 **
## carrier_ct        1.000e+00  1.029e-06  9.717e+05  < 2e-16 ***
## weather_ct        1.000e+00  4.223e-06  2.368e+05  < 2e-16 ***
## nas_ct            1.000e+00  5.236e-07  1.910e+06  < 2e-16 ***
## security_ct       9.999e-01  3.939e-05  2.538e+04  < 2e-16 ***
## late_aircraft_ct  1.000e+00  5.664e-07  1.766e+06  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005155 on 70586 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.115e+13 on 5 and 70586 DF,  p-value: < 2.2e-16
```

Table 2.0: Multiple Linear Regression Output

The calculated coefficient shows that an average rise in arr_del15 is related to a 1.000e+00 unit increase in *carrier_ct*, *weather_ct*, *nas_ct*, *late_aircraft_c*t, and a 9.999e-01 unit increase in *security_ct*. In addition, since the R^2 values are 1, we may infer that the data's points can account for all of the variance in the dependent variable's values, and the model's relatively close residual standard error is indication of its high prediction performance for this dataset.

Finally, think about the concept that "The better the model, the greater the F value." By examining this result, we can see that the F-statistic gives a significant large value of 2.115e+13 and a p-value of 2.2e-16, which is almost zero. Based on this finding, it is able to reject the null hypothesis while accepting the alternative hypothesis.

Conclusions

The evaluation of the study gives a clearer perspective on numerous qualities and characteristics of the USA airline delay dataset, which may aid in understanding the best targeted group to choose for future activities with different goals. By optimizing the correlation graph, the organization should concentrate more on the 4 components, such as *carrier_ct*, *weather_ct*, *nas_ct*, and *late_aircraft_ct*, which have a strong association with the dependent variables. Through plotting the PCA output, we are able to visualise the relationship between variables. In MLR, it can be observed that it has 99% of impacting towards *arr_del15* as compared to the other variables who have 100% affectability. Moving on, it is extremely surprising to discover that *security_ct* has the least concern about this issue despite the fact that the five causes investigated in this study are the prevalent ones that are impacting flight delays, according to internet research. Lastly, this analysis has also resulted in the conclusion that MLR is the best method for accurately predicting the cause of an aircraft delay. Because there were a lot of observations, it was straightforward to quickly identify the variables affecting airline delays. However, there has been no prior research on this dataset, which makes it challenging to support the accuracy in this area.

References

Group, A. T. (2004). *The economic & social benefits of air transport.* Geneva: Air Transport Action Group. Retrieved from The economic & social benefits of air transport: https://www.icao.int/meetings/wrdss2011/documents/jointworkshop2005/atag_socialbenefitsairtransport.pdf

Coral. (n.d.). *These Gold Wings*. Retrieved from Why Do Flights Get Delayed? 10 Reasons You'd Never Thought Of : https://www.thesegoldwings.com/delayed-flights/

Sheffield School of Aeronautics. (2022). *Why Do Flights Get Delayed So Often?* Retrieved from https://www.sheffield.com/why-do-flights-get-delayed

Wang, F. &. (2022). Flight delay forecasting and analysis of direct and indirect factors. *IET Intelligent Transport Systems*, 891-895.

RYANJT. (2022, August 29). *USA Airline Delay* . Retrieved from Kaggle: https://www.kaggle.com/datasets/ryanjt/airline-delay-cause

Fakhitah, R., Mohd, N., & Zainon. (2019). A Review on Data Cleansing Methods for Big Data. *Procedia Computer Science*, 731-738.

Tucci, L. (2021, December). *What is predictive analytics? An enterprise guide*. Retrieved from Search Business Analytics: https://searchbusinessanalytics.techtarget.com/definition/predictive-analytics

Appendices

Capstone MA3405

Sally Pang Shue Yan

2022-09-25

Load Data

```
airline_delay <- read.csv("/Users/sallypang/Desktop/airline_delay causes.csv")
head(airline_delay) # view data
```

```
##   year month carrier           carrier_name airport
## 1 2012     4      AA American Airlines Inc.     ABQ
## 2 2012     4      AA American Airlines Inc.     ATL
## 3 2012     4      AA American Airlines Inc.     AUS
## 4 2012     4      AA American Airlines Inc.     BDL
## 5 2012     4      AA American Airlines Inc.     BHM
## 6 2012     4      AA American Airlines Inc.     BNA
##                                       airport_name arr_flights arr_del15
## 1     Albuquerque, NM: Albuquerque International Sunport         234        41
## 2  Atlanta, GA: Hartsfield-Jackson Atlanta International         404        61
## 3           Austin, TX: Austin - Bergstrom International         672        96
## 4                    Hartford, CT: Bradley International         119        20
## 5 Birmingham, AL: Birmingham-Shuttlesworth International          90        15
## 6                  Nashville, TN: Nashville International         291        45
##   carrier_ct weather_ct nas_ct security_ct late_aircraft_ct arr_cancelled
## 1      14.56       4.77   8.64           0            13.03            19
## 2      20.73       6.32  15.36           0            18.59            28
## 3      33.34       8.98  20.42           0            33.26            30
## 4       6.03       1.55   5.98           0             6.44             3
## 5       4.54       3.18   2.04           0             5.24             5
## 6      19.89       5.92   9.45           0             9.74            16
##   arr_diverted arr_delay carrier_delay weather_delay nas_delay security_delay
## 1            2      2445           712           495       260              0
## 2            0      3304           743           544       493              0
## 3            1      4917          1333           709       751              0
## 4            0      1080           284           254       166              0
## 5            0      1085           306           335       137              0
## 6            2      2530          1082           526       316              0
##   late_aircraft_delay
## 1                 978
## 2                1524
## 3                2124
## 4                 376
## 5                 307
## 6                 606
```

```
dim(airline_delay) # display number of observations and variables
```

```
## [1] 70695    21
```

Data Preparation & Transformation

Prior to processing, it is a crucial stage that frequently entails reformatting data, correcting data, and integrating datasets to enhance data.

Add a Categorical variables called month name

```
library(readxl)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 4.1.2
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(knitr)

month <- 1:12
month_nm <- c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul",
              "Aug", "Sep", "Oct", "Nov","Dec")
months <- data.frame(month, month_nm)
flights_raw <- left_join(airline_delay, months, by = "month")
```

Data descriptions

```
Field <- c("year", "month", "carrier",
           "carrier_name", "airport", "airport_name",
           "arr_flights", "arr_del15", "carrier_ct",
           "weather_ct", "nas_ct", "security_ct",
           "late_aircraft_ct", "arr_cancelled", "arr_diverted",
           "arr_delay", "carrier_delay", "weather_delay",
           "nas_delay", "security_delay", "late_aircraft_delay")

Description <- c("Year (yyyy)", "Month (mm)",
                 "Airline carrier abbreviation", "Airline carrier name",
                 "Airport Code", "Airport Name",
                 "Number of flights which arrived at the airport.", "Number of flights delayed (>= 15minutes lat
e).",
                 "Number of flights delayed due to air carrier (e.g. maintenance or crew problems, aircraft clean
ing, baggage loading, fueling, etc.).",
                 "Number of flights delayed due to weather.",
                 "Number of flights delayed due to National Aviation System (e.g. non-extreme weather conditions,
 airport operations, heavy traffic volume, and air traffic control).",
                 "Number of flights delayed due to security (e.g. evacuation of a terminal or concourse, re-board
ing of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 min
utes at screening areas).",
                 "Number of flights delayed due to a previous flight using the same aircraft being late.",
                 "Number of cancelled flights", "Number of flights diverted",
                 "Total time (minutes) of delayed flights.", "Total time (minutes) of delayed flights due to air
 carrier.",
                 "Total time (minutes) of delayed flights due to weather.",
                 "Total time (minutes) of delayed flights due to National Aviation System.",
                 "Total time (minutes) of delayed flights due to security.",
                 "Total time (minutes) of delayed flights due to a previous flight using the same aircraft being
 late.")

VariableType <- c("Qualitative",    "Qualitative",  "Qualitative",  "Qualitative",
                  "Qualitative",    "Qualitative",  "Quantitative", "Quantitative",
                  "Quantitative",   "Quantitative", "Quantitative", "Quantitative",
                  "Quantitative",   "Quantitative", "Quantitative", "Quantitative",
                  "Quantitative",   "Quantitative", "Quantitative", "Quantitative",
                  "Quantitative")

VariableMeasure <- c("Explanatory", "Explanatory",  "Explanatory",  "Explanatory",
                     "Explanatory", "Explanatory",  "Explanatory",  "Response",
                     "Independent", "Independent",  "Independent",  "Independent",
                     "Independent", "Explanatory",  "Explanatory",  "Explanatory",
                     "Explanatory", "Explanatory",  "Explanatory",  "Explanatory",
                     "Explanatory")
FieldDefinitions <- data.frame(Field, VariableType, VariableMeasure, Description)
FieldDefinitions %>% kable() %>% kable_styling()
```

Field	VariableType	VariableMeasure	Description
year	Qualitative	Explanatory	Year (yyyy)
month	Qualitative	Explanatory	Month (mm)
carrier	Qualitative	Explanatory	Airline carrier abbreviation
carrier_name	Qualitative	Explanatory	Airline carrier name
airport	Qualitative	Explanatory	Airport Code
airport_name	Qualitative	Explanatory	Airport Name
arr_flights	Quantitative	Explanatory	Number of flights which arrived at the airport.
arr_del15	Quantitative	Response	Number of flights delayed (>= 15minutes late).
carrier_ct	Quantitative	Independent	Number of flights delayed due to air carrier (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
weather_ct	Quantitative	Independent	Number of flights delayed due to weather.
nas_ct	Quantitative	Independent	Number of flights delayed due to National Aviation System (e.g. non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control).
security_ct	Quantitative	Independent	Number of flights delayed due to security (e.g. evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas).
late_aircraft_ct	Quantitative	Independent	Number of flights delayed due to a previous flight using the same aircraft being late.
arr_cancelled	Quantitative	Explanatory	Number of cancelled flights
arr_diverted	Quantitative	Explanatory	Number of flights diverted
arr_delay	Quantitative	Explanatory	Total time (minutes) of delayed flights.
carrier_delay	Quantitative	Explanatory	Total time (minutes) of delayed flights due to air carrier.
weather_delay	Quantitative	Explanatory	Total time (minutes) of delayed flights due to weather.
nas_delay	Quantitative	Explanatory	Total time (minutes) of delayed flights due to National Aviation System.
security_delay	Quantitative	Explanatory	Total time (minutes) of delayed flights due to security.
late_aircraft_delay	Quantitative	Explanatory	Total time (minutes) of delayed flights due to a previous flight using the same aircraft

Dependent Variable

```
FieldDefinitions %>% filter(VariableMeasure == "Response") %>% kable() %>% kable_styling()
```

Field	VariableType	VariableMeasure	Description
arr_del15	Quantitative	Response	Number of flights delayed (>= 15minutes late).

Independent Variable

```
FieldDefinitions %>% filter(VariableMeasure == "Independent") %>% kable() %>% kable_styling()
```

Field	VariableType	VariableMeasure	Description
carrier_ct	Quantitative	Independent	Number of flights delayed due to air carrier (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
weather_ct	Quantitative	Independent	Number of flights delayed due to weather.
nas_ct	Quantitative	Independent	Number of flights delayed due to National Aviation System (e.g. non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control).
security_ct	Quantitative	Independent	Number of flights delayed due to security (e.g. evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas).
late_aircraft_ct	Quantitative	Independent	Number of flights delayed due to a previous flight using the same aircraft being late.

Objectives

The purpose of this study is to identify the independent variables that are influencing delayed flights.

Ho: The independent variables do not affect the delayed flights.

Ha: he independent variables does affect the delayed flights.

Pre-process Data

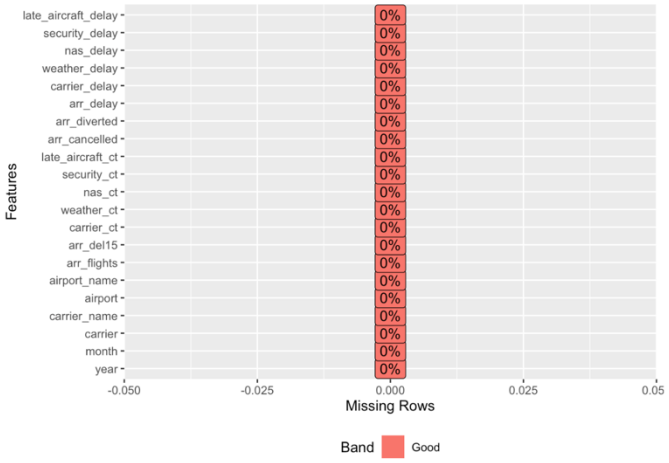
This process is to put raw data into a comprehensible format. Given that we cannot deal with raw data, it is also a crucial stage in data mining. Prior to using machine learning or data mining methods, the quality of the data should be evaluated.

Clean Data

```
airline_delay.new <- na.omit(airline_delay) # remove all rows having NA.
dim(airline_delay.new) # display number of observations and variables after cleaning process.
```

```
## [1] 70592    21
```

```
library(DataExplorer)
plot_missing(airline_delay.new)
```



The plot's results demonstrate that the dataset contains no missing data.

Split dataset into the Training set and Test set

Divide a dataset into train and test sets to see how effectively our machine learning model works.

```
library(caTools)
split <- sample.split(airline_delay.new, SplitRatio = 0.7)  # split data into ratio of 7:2
training_set <- subset(airline_delay.new, split == "TRUE")
test_set <- subset(airline_delay.new, split == "FALSE")
```

Scale data

Scaling the data is one of the pre-processing steps used in machine learning algorithms on the data set, which makes it easier for the model to understand and learn about the problem.

```
training_set.scale <- scale(training_set[, 8:13])
test_set.scale <- scale(test_set[, 8:13])
```

Exploratory Analysis

This section helps to explore the dataset to understand the dataset before making any assumptions.

Heatmap of Correlation & Demdrogram

Heatmap with the Correlation and Dendrogram grouping of the allocation on many key metrics that an analyst uses to identify possible links between variables and comprehend the strength of these associations.

```
library(heatmaply)
```

```
## Loading required package: plotly
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following object is masked from 'package:graphics':
##
##     layout
```

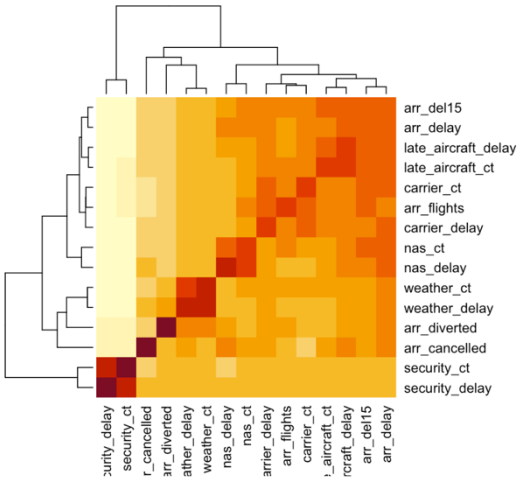
```
## Loading required package: viridis
```

```
## Loading required package: viridisLite
```

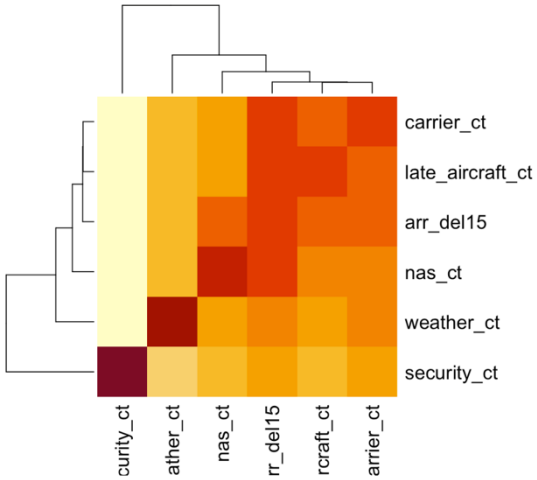
```
## Warning: package 'viridisLite' was built under R version 4.1.2
```

```
##
## ======================
## Welcome to heatmaply version 1.3.0
##
## Type citation('heatmaply') for how to cite the package.
## Type ?heatmaply for the main documentation.
##
## The github page is: https://github.com/talgalili/heatmaply/
## Please submit your suggestions and bug-reports at: https://github.com/talgalili/heatmaply/issues
## You may ask questions at stackoverflow, use the r and heatmaply tags:
##   https://stackoverflow.com/questions/tagged/heatmaply
## ======================
```

```
airline_delay.new.table <- round(cor(airline_delay.new[7:21]), 3)
dataMatNorm <- as.matrix(normalize(airline_delay.new.table))
heatmap(dataMatNorm) # whole dataset
```



The heatmap displays the correlation for each variable included in this dataset.



The correlation for the six factors used in this study is shown in the second heatmap. According to the intensity of the colors, the graphs clearly reveal that arr_del15 has the lowest correlation with security ct and higher connection with nas_ct, late_aircraft_ct, and carrier_ct. The dendrogram indicates that the variables carrier_ct and late_aircraft_ct are the most comparable, but as security_ct is fused later, it becomes clear that this variable is substantially distinct from the rest of the variables in terms of their relationships to and affectivity on delayed flights.

Bar chart of Flight delays by airport

Amount of delayed flights per airports.

```
library(ggplot2)
library(kableExtra)
library(scales)
```

```
## Warning: package 'scales' was built under R version 4.1.2
```

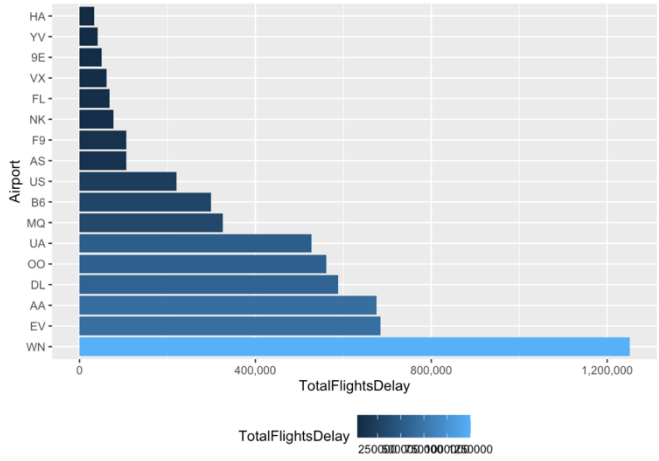
```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:viridis':
##
##     viridis_pal
```

```
library(dplyr)
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```
flights_raw %>% select(carrier, arr_del15) %>% drop_na() %>% group_by(carrier) %>%
  dplyr::summarise(TotalFlightsDelay = sum(arr_del15)) %>%
  ggplot( aes(x=reorder(carrier, -TotalFlightsDelay), y=TotalFlightsDelay, fill=TotalFlightsDelay))+
  geom_col() + coord_flip() + scale_y_continuous(labels = comma) + xlab("Airport") +
  scale_colour_continuous(labels = comma) + theme(legend.position="bottom")
```



The result shows airport of WN have the most total flight delay.

Descriptive Analysis

Using historical data, analytical modeling, data mining techniques, and machine learning, predictive analytics is a subset of new insights that forecasts probable outcomes.

1. Principal component analyse (PCA)

A type of unsupervised statistical learning is this algorithm. It offers data visualization, dimension reduction techniques, and, most importantly, it offers data pre-processing techniques before applying another methodology.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✔ tibble  3.1.8     ✔ stringr 1.4.0
## ✔ readr   2.1.2     ✔ forcats 0.5.2
## ✔ purrr   0.3.4
```

```
## Warning: package 'tibble' was built under R version 4.1.2
```

```
## Warning: package 'readr' was built under R version 4.1.2
```

```
## Warning: package 'forcats' was built under R version 4.1.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ readr::col_factor()      masks scales::col_factor()
## ✖ purrr::discard()         masks scales::discard()
## ✖ plotly::filter()         masks dplyr::filter(), stats::filter()
## ✖ kableExtra::group_rows() masks dplyr::group_rows()
## ✖ dplyr::lag()             masks stats::lag()
```

```
library(skimr)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(plm)
```

```
## Warning: package 'plm' was built under R version 4.1.2
```

```
##
## Attaching package: 'plm'
##
## The following objects are masked from 'package:dplyr':
##
##     between, lag, lead
```

```
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 4.1.2
```

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.1.2
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.1.2
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(dplyr)
library(devtools)
```

```
## Warning: package 'devtools' was built under R version 4.1.2
```

```
## Loading required package: usethis
```

```
## Warning: package 'usethis' was built under R version 4.1.2
```

```
library(ggbiplot)
```

```
## Loading required package: plyr
```

```
## Warning: package 'plyr' was built under R version 4.1.2
```

```
## ------------------------------------------------------------------------------
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## ------------------------------------------------------------------------------
##
## Attaching package: 'plyr'
##
## The following object is masked from 'package:purrr':
##
##     compact
##
## The following objects are masked from 'package:plotly':
##
##     arrange, mutate, rename, summarise
##
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
##
## Loading required package: grid
```

```
apply(is.na(airline_delay), MARGIN=2, FUN = sum)
```

```
##                year               month             carrier        carrier_name
##                   0                   0                   0                   0
##             airport        airport_name         arr_flights            arr_del15
##                   0                   0                  95                 103
##          carrier_ct          weather_ct              nas_ct          security_ct
##                  95                  95                  95                  95
##    late_aircraft_ct       arr_cancelled        arr_diverted           arr_delay
##                  95                  95                  95                  95
##       carrier_delay       weather_delay           nas_delay      security_delay
##                  95                  95                  95                  95
## late_aircraft_delay
##                  95
```

```
apply(is.na(airline_delay), MARGIN = 2, FUN = mean)
```

```
##                year               month             carrier        carrier_name
##         0.000000000         0.000000000         0.000000000         0.000000000
##             airport        airport_name         arr_flights            arr_del15
##         0.000000000         0.000000000         0.001343801         0.001456963
##          carrier_ct          weather_ct              nas_ct          security_ct
##         0.001343801         0.001343801         0.001343801         0.001343801
##    late_aircraft_ct       arr_cancelled        arr_diverted           arr_delay
##         0.001343801         0.001343801         0.001343801         0.001343801
##       carrier_delay       weather_delay           nas_delay      security_delay
##         0.001343801         0.001343801         0.001343801         0.001343801
## late_aircraft_delay
##         0.001343801
```

```
airline_delay.new.table <- round(cor(airline_delay.new[8:13]), 3)
head(airline_delay.new.table)
```

```
##                 arr_del15 carrier_ct weather_ct nas_ct security_ct
## arr_del15           1.000      0.950      0.782  0.930       0.499
## carrier_ct          0.950      1.000      0.749  0.826       0.515
## weather_ct          0.782      0.749      1.000  0.717       0.378
## nas_ct              0.930      0.826      0.717  1.000       0.418
## security_ct         0.499      0.515      0.378  0.418       1.000
## late_aircraft_ct    0.962      0.904      0.724  0.816       0.489
##                 late_aircraft_ct
## arr_del15                  0.962
## carrier_ct                 0.904
## weather_ct                 0.724
## nas_ct                     0.816
## security_ct                0.489
## late_aircraft_ct           1.000
```

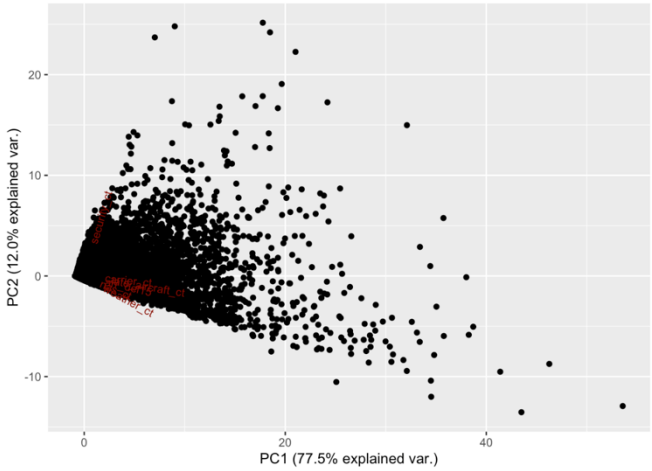
```
airline_delay.pca <- prcomp(airline_delay.new[8:13], center = TRUE,scale. = TRUE)

print(airline_delay.pca)
```

```
## Standard deviations (1, .., p=6):
## [1] 2.155731e+00 8.478455e-01 5.816408e-01 4.479691e-01 3.082197e-01
## [6] 2.210407e-05
##
## Rotation (n x k) = (6 x 6):
##                       PC1         PC2        PC3          PC4          PC5
## arr_del15        0.4585806 -0.11012875 -0.1994978  0.002081005 -0.074919488
## carrier_ct       0.4417387 -0.04109452 -0.1602587  0.360049191  0.776396363
## weather_ct       0.3872141 -0.21012878  0.8952689  0.006897705 -0.058299197
## nas_ct           0.4222893 -0.18345048 -0.2305215 -0.802078927 -0.008300451
## security_ct      0.2724094  0.95038967  0.1035732 -0.102408068 -0.036491917
## late_aircraft_ct 0.4382932 -0.07165333 -0.2629519  0.465289859 -0.621929581
##                         PC6
## arr_del15        -0.855662836
## carrier_ct        0.212293630
## weather_ct        0.030956091
## nas_ct            0.302453351
## security_ct       0.002471361
## late_aircraft_ct  0.361012616
```

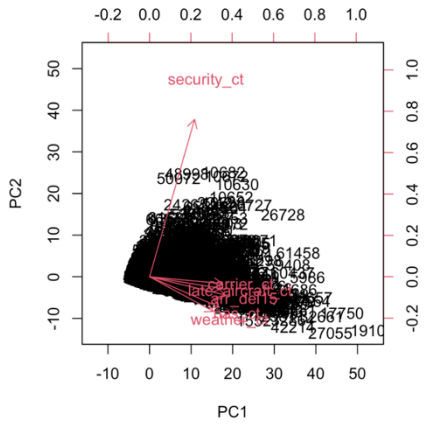
```
summary(airline_delay.pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5       PC6
## Standard deviation     2.1557 0.8478 0.58164 0.44797 0.30822 2.21e-05
## Proportion of Variance 0.7745 0.1198 0.05638 0.03345 0.01583 0.00e+00
## Cumulative Proportion  0.7745 0.8943 0.95072 0.98417 1.00000 1.00e+00
```

```
ggbiplot(airline_delay.pca, scale=0)
```




```
biplot(airline_delay.pca, scale = 0)
```



2. K-Nearest Neighbors (KNN)

The k-nearest neighbors (KNN) method computes the chance that a data point will belong to one group or another based on which group the data points closest to it do.

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.1.2
```

```
## Registered S3 methods overwritten by 'proxy':
##   method           from
##   print.registry_field registry
##   print.registry_entry registry
```

```
library(caTools)
library(class)



k = sqrt(70592)

classifier_knn <- knn(train = scale(training_set[, 7:14]),
                      test = scale(test_set[, 7:14]),
                      cl = training_set$carrier,
                      k = k)

cm <- table(test_set$carrier, classifier_knn)
cm
```

```
##     classifier_knn
##        9E   AA   AS   B6   DL   EV   F9   FL   HA   MQ   NK   OO   UA   US   VX
##   9E    1    4   23    6   96  142   30    0    1   37    4  100    2    4    0
##   AA    0  297    7   34  294  406   61    0    2  207    2  151   72   51    0
##   AS    0    7  189   26  369   58   75    0    0   54    9  331    5   47    0
##   B6    0   45   21  100  176  342   57    0    2   53    8  163   20   99    0
##   DL    0   77   54    4 1795  212   65    0   17  177    1  294   58   28    0
##   EV    0   32   27   13  270 1904  106    0   12  143    0  635   24    9    0
##   F9    0    5   62    6  176  153  493    0    1   13    2  233   13    5    0
##   FL    0    6    9    0  108  119   31    0    4   11    4  141    3    2    0
##   HA    0    2    0    4  132   28    0    0  104    0    0   11    0   10    0
##   MQ    0   87    2   44  204  312   41    0    1  853    0  306   10   11    0
##   NK    0    5   22   18   22   69   45    0    0    9   39   42    0   41    0
##   OO    0   62   47   36  495  807  164    0    4  233    0 1368   27   26    0
##   UA    0  127   16    9  413  309   61    0    4  101    2  287  169   23    0
##   US    0   32   42   65  265  159   54    0    2   19    6  129   16  176    0
##   VX    0   18    6    5   73   12    4    0    0  113    0   98    7    3    1
##   WN    0   91    0   22   80  251    2    0    1   24    0  116   15   16    0
##   YV    0    4   11   11  115  111   33    0    3   36    0  165    3   12    0
##     classifier_knn
##        WN   YV
##   9E    4    0
##   AA  133    0
##   AS   27    0
##   B6  105    0
##   DL   50    0
##   EV  114    0
##   F9    3    0
##   FL   16    0
##   HA   41    0
##   MQ   29    0
##   NK    3    0
##   OO  181    1
##   UA  129    0
##   US   11    0
##   VX   31    0
##   WN 1105    0
##   YV   10    0
```

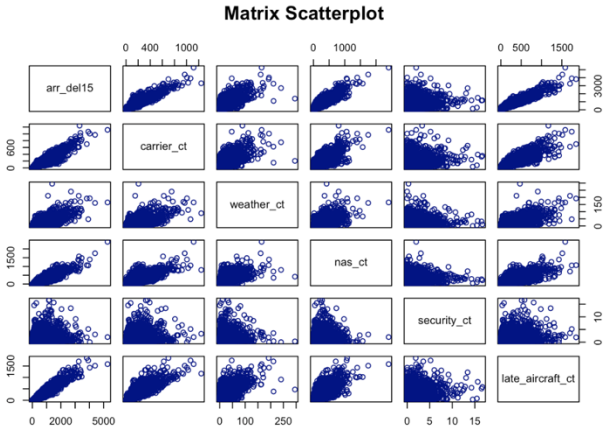
```
misClassError <- mean(classifier_knn != test_set$carrier)
print(paste('Accuracy =', 1- misClassError))
```

```
## [1] "Accuracy = 0.347003824904377"
```

3. Mutiple Linear Regression

This method is a statistical approach that forecasts the result of a dependent variable using two or more independent variables. Using this method, we may calculate the model's variance as well as the proportional contributions of each independent variable to the overall variance.

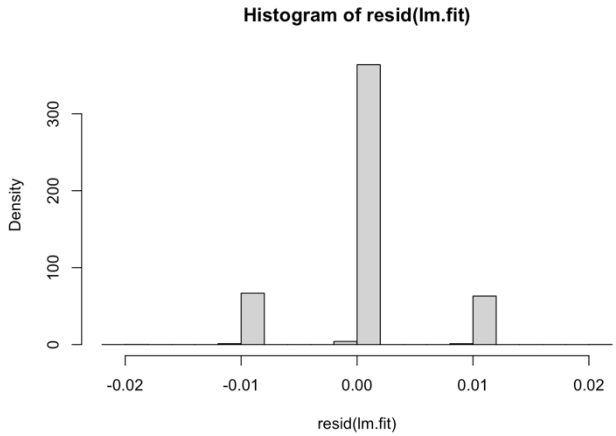
```
plot(airline_delay.new[8:13], col="navy", main="Matrix Scatterplot")
```



```
lm.fit =lm(arr_del15~.,data= airline_delay.new[8:13])
summary(lm.fit)
```

```
##
## Call:
## lm(formula = arr_del15 ~ ., data = airline_delay.new[8:13])
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0202088  0.0000576  0.0000659  0.0000759  0.0206157
##
## Coefficients:
##                     Estimate Std. Error    t value Pr(>|t|)
## (Intercept)      -6.683e-05  2.168e-05 -3.083e+00  0.00205 **
## carrier_ct        1.000e+00  1.029e-06  9.717e+05  < 2e-16 ***
## weather_ct        1.000e+00  4.223e-06  2.368e+05  < 2e-16 ***
## nas_ct            1.000e+00  5.236e-07  1.910e+06  < 2e-16 ***
## security_ct       9.999e-01  3.939e-05  2.538e+04  < 2e-16 ***
## late_aircraft_ct  1.000e+00  5.664e-07  1.766e+06  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005155 on 70586 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.115e+13 on 5 and 70586 DF,  p-value: < 2.2e-16
```

```
hist(resid(lm.fit), prob = TRUE)
```



The 0.01, 0.02 as 1 standard deviation, 2 standard deviation, can see that on the right side, the majority of the data points is within the 2nd deviation. Since the median is nearly zero and the model is somewhat skewed to the left when looking at the min of -0.0202 and 1Q of 0.0000576, the result shows that the residuals produced a symmetrical output.

Conclusion

By examining this result, we can see that the F-statistic gives a significant large value of 2.115e+13 and a p-value of 2.2e-16, which is almost zero. Based on this finding, it is able to reject the null hypothesis while accepting the alternative hypothesis.

Ho: The independent variables do not affect the delayed flights. (Reject)

Ha: The independent variables does affect the delayed flights.