BANA 6043-STATISTICAL COMPUTING

Project: Statistical Analysis to Reduce landing Overrun

Niharika Gupta- M13437287
Carl H Lindner School of Business: University of Cincinnati

Statistical Computing (BANA 6043 Project)

Gupta2na **Niharika Gupta** **M13437287**

Contents

Statistical Computing (BANA 6043 Project)

Gupta2na **Niharika Gupta** **M13437287**

CHAPTER 1: DATA EXPLORATION AND DATA CLEANING

***Goal**: Importing the given datasets, and exploring to check for outliers, missing values and duplicates and finally acting upon them accordingly.*

STEP 1: IMPORTING DATA FILES

Data set was imported into R studio using the below code.

Assignment 6, Landing overrun.R

Source on Save

```
flights=read.csv("FAA1.csv");
flights
```

```
> flights=read.csv("FAA1.csv");
> flights
   aircraft  duration no_pasg speed_ground speed_air    height    pitch  distance
1    boeing  98.47909      53    107.91568 109.32838 27.418924 4.043515 3369.8364
2    boeing 125.73330      69    101.65559 102.85141 27.804716 4.117432 2987.8039
3    boeing 112.01700      61     71.05196        NA 18.589386 4.434043 1144.9224
4    boeing 196.82569      56     85.81333        NA 30.744597 3.884236 1664.2182
5    boeing  90.09538      70     59.88853        NA 32.397688 4.026096 1050.2645
6    boeing 137.59582      55     75.01434        NA 41.214963 4.203853 1627.0682
7    boeing  73.02379      54     54.42980        NA 24.035322 3.837646  805.3040
8    boeing  52.90319      57     57.10166        NA 19.388838 4.643672  573.6218
9    boeing 155.51862      61     85.44362        NA 35.375390 4.228728 1698.9928
10   boeing 176.86203      56     61.79671        NA 36.748816 4.184399 1137.7458
11   boeing 158.46190      61     53.77813        NA 46.355833 5.556399 1075.3717
12   boeing 180.61656      54    141.21864 141.72494 23.575935 5.216802 6533.0477
13   boeing  72.28963      54     93.39176  92.86956 32.223489 3.818276 2128.7083
14   boeing 187.59955      58     94.03641  96.19646 33.661226 4.636185 2304.8576
15   boeing 154.36870      63     63.54061        NA 26.402992 3.856658 1089.9730
16   boeing 165.54195      69     48.77467        NA 31.228665 3.902046  943.0684
17   boeing 153.54634      61     83.55649        NA 29.897473 3.519784 1793.5628
18   boeing 107.11332      78     86.80796        NA 25.477015 4.414219 1910.8769
```

STEP 2: STRUCTURE OF DATASET

Will give the names of different variables and their data types.

```
#STRUCTURE OF THE DATASET#
str(flights)
```

```
> #STRUCTURE OF THE DATASET#
> str(flights)
'data.frame':   800 obs. of  8 variables:
 $ aircraft    : Factor w/ 2 levels "airbus","boeing": 2 2 2 2 2 2 2 2 2 2 ...
 $ duration    : num  98.5 125.7 112 196.8 90.1 ...
 $ no_pasg     : int  53 69 61 56 70 55 54 57 61 56 ...
 $ speed_ground: num  107.9 101.7 71.1 85.8 59.9 ...
 $ speed_air   : num  109 103 NA NA NA ...
 $ height      : num  27.4 27.8 18.6 30.7 32.4 ...
 $ pitch       : num  4.04 4.12 4.43 3.88 4.03 ...
 $ distance    : num  3370 2988 1145 1664 1050 ...
```

STEP 3: CHECKING FOR DUPLICATES

Observation:

There are no duplicate rows in the dataset.

```
sum(duplicated(flights[,-2]))
```

```
> sum(duplicated(flights[,-2]))
[1] 0
>
```

STEP 4: VERYFING THERE ARE NO DUPLICATES

Observation:

- The resulting dataset (after removing duplicates) has the same number of rows as the parent dataset, meaning no rows were found as duplicates.

```
#VERIFYING THE ABOVE RESULT BY USING COMMAND FOR REMOVAL OF DUPLICATES#
flightsnodup=flights[!duplicated(flights$height,flights$duration,flights$speed_air), ]
```

```
Environment   History   Connections
Import Dataset                                  List
Global Environment
Data
flights          800 obs. of 8 variables
flightsnodup     800 obs. of 8 variables
```

Gupta2na Niharika Gupta M13437287

STEP 5 : REMOVING ABNORMAL VALUES

Observation:

- 786 values remain which means that there were 14 abnormal values.

```
#GETTING RID OF ABNORMAL VALUES#
flightsclean<-flights[ which( (is.null(flights$duration) || flights$duration > 40)
                             & flights$speed_ground >= 30 & flights$speed_ground <=140
                             & (is.null(flights$speed_air) || (flights$speed_air >= 30
                                                              & flights$speed_air <=140))
                             & flights$height >=6 & flights$distance<6000) , ]
flightsclean
```

flights	800 obs. of 8 variables
flightsclean	786 obs. of 8 variables
flightsnodup	800 obs. of 8 variables

```
> flightsclean
   aircraft  duration no_pasg speed_ground speed_air    height    pitch  distance
1    boeing  98.47909      53    107.91568 109.32838 27.418924 4.043515 3369.8364
2    boeing 125.73330      69    101.65559 102.85141 27.804716 4.117432 2987.8039
3    boeing 112.01700      61     71.05196        NA 18.589386 4.434043 1144.9224
4    boeing 196.82569      56     85.81333        NA 30.744597 3.884236 1664.2182
5    boeing  90.09538      70     59.88853        NA 32.397688 4.026096 1050.2645
6    boeing 137.59582      55     75.01434        NA 41.214963 4.203853 1627.0682
7    boeing  73.02379      54     54.42980        NA 24.035322 3.837646  805.3040
8    boeing  52.90319      57     57.10166        NA 19.388838 4.643672  573.6218
9    boeing 155.51862      61     85.44362        NA 35.375390 4.228728 1698.9928
10   boeing 176.86203      56     61.79671        NA 36.748816 4.184399 1137.7458
11   boeing 158.46190      61     53.77813        NA 46.355833 5.556399 1075.3717
13   boeing  72.28963      54     93.39176  92.86956 32.223489 3.818276 2128.7083
14   boeing 187.59955      58     94.03641  96.19646 33.661226 4.636185 2304.8576
15   boeing 154.36870      63     63.54061        NA 26.402992 3.856658 1089.9730
16   boeing 165.54195      69     48.77467        NA 31.228665 3.902046  943.0684
17   boeing 153.54634      61     83.55649        NA 29.897473 3.519784 1793.5628
18   boeing 107.11332      78     86.80796        NA 25.477015 4.414219 1910.8769
19   boeing 233.80250      69    104.80843 103.86846 43.882732 3.245098 3213.9853
20   boeing 163.90650      55    119.38046 120.44471 38.558536 3.701449 4524.2789
21   boeing  97.47762      63     73.53398        NA 29.152465 4.014006 1332.0387
22   boeing 118.63054      55     79.99482        NA 29.366866 4.407181 1515.9653
23   boeing 126.54029      70     94.78123  91.14207 39.476299 3.594936 2182.2207
24   boeing 179.91592      66     63.67117        NA 19.574700 4.286734  873.4409
25   boeing 112.90010      53     98.18041  99.13583 28.152991 3.987471 2586.6651
26   boeing  56.64049      66     72.95366        NA 36.154157 4.387856 1205.1280
27   boeing  86.82891      62     91.71454  92.87485 28.773729 3.305888 2313.3357
28   boeing 157.35773      57     72.32713        NA 26.223285 4.223181 1105.3659
29   boeing 186.68141      49     66.41723        NA 44.692696 4.113544 1176.0277
30   boeing 140.23631      65    118.74200 119.40215 19.856192 4.646266 4217.1295
```

CHAPTER 2: DESCRIPTIVE STUDY OF VARIABLES

Goals: To study association of landing distance with different variables and try to find variables of significance.

STEP 1: SUMMARY OF ALL VARIABLES

Will give details of all variables including min, max, mean, median.

```
#SUMMARY OF ALL VARIABLES#

summary(flightsclean)
```

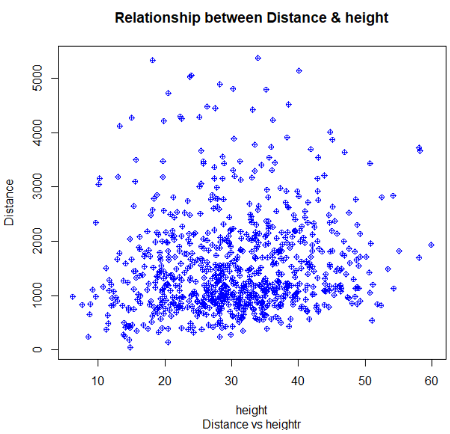
```
  aircraft      duration          no_pasg        speed_ground      speed_air          height           pitch            distance
 airbus:396   Min.   : 14.76   Min.   :29.00   Min.   : 33.57   Min.   : 90.00   Min.   : 6.228   Min.   :2.284   Min.   :  41.72
 boeing:390   1st Qu.:118.75   1st Qu.:55.00   1st Qu.: 66.20   1st Qu.: 96.14   1st Qu.:23.643   1st Qu.:3.654   1st Qu.: 920.39
              Median :154.13   Median :60.00   Median : 79.83   Median :100.88   Median :30.267   Median :4.015   Median :1277.47
              Mean   :153.93   Mean   :60.07   Mean   : 79.69   Mean   :103.47   Mean   :30.511   Mean   :4.014   Mean   :1544.88
              3rd Qu.:189.42   3rd Qu.:65.00   3rd Qu.: 92.37   3rd Qu.:109.37   3rd Qu.:37.009   3rd Qu.:4.382   3rd Qu.:1965.64
              Max.   :305.62   Max.   :87.00   Max.   :132.78   Max.   :132.91   Max.   :59.946   Max.   :5.927   Max.   :5381.96
                                                                NA's   :588
```

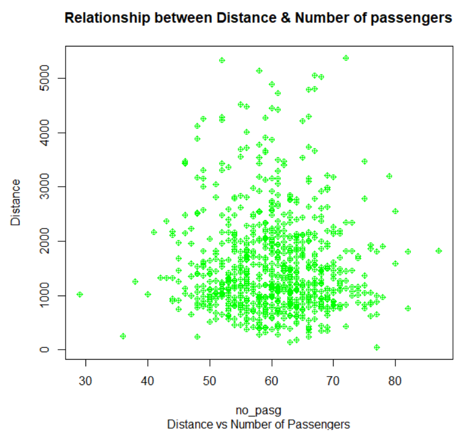
STEP 2: STUDYING LANDING DISTANCE WITH OTHER VARIABLES

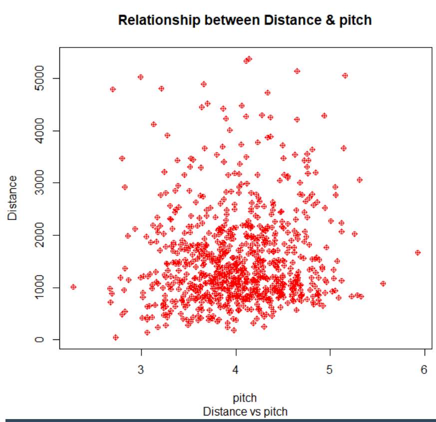
Each variable is plotted against distance to study their association with distance.

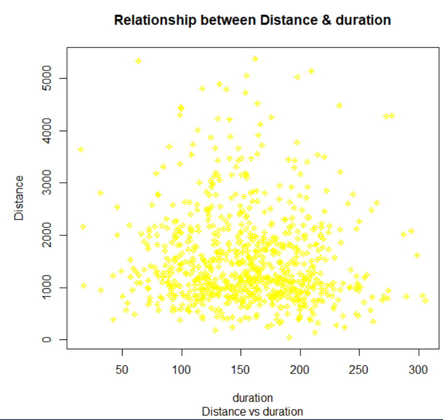
```
par(mfrow=c(1, 2))


plot ( flightsclean$distance~flightsclean$height,
       main="Relationship between Distance & height",
       sub="Distance vs heightr",
       xlab="height", ylab="Distance",
       pch=10,col="blue"
)

plot ( flightsclean$distance~flightsclean$no_pasg,
       main="Relationship between Distance & Number of passengers",
       sub="Distance vs Number of Passengers",
       xlab="no_pasg", ylab="Distance",
       pch=10,col="green"
)

par(mfrow=c(1, 2))
plot ( flightsclean$distance~flightsclean$pitch,
       main="Relationship between Distance & pitch",
       sub="Distance vs pitch",
       xlab="pitch", ylab="Distance",
       pch=10,col="red"
)

plot ( flightsclean$distance~flightsclean$duration,
       main="Relationship between Distance & duration",
       sub="Distance vs duration",
       xlab="duration", ylab="Distance",
       pch=10,col="yellow"
)

par(mfrow=c(1,1))

plot ( flightsclean$distance~flightsclean$aircraft,
       main="Relationship between Distance & aircraft",
       sub="Distance vs aircraft",
       xlab="aircraft", ylab="Distance",
       pch=10,col="purple"
)
```

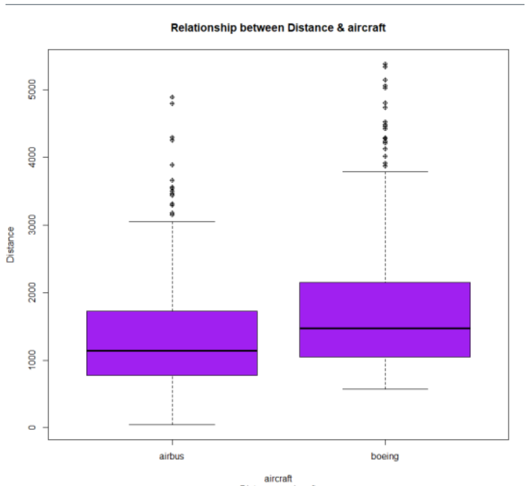
```
par(mfrow=c(1,2))

plot ( flightsclean$distance~flightsclean$speed_ground,
       main="Relationship between Distance & Speed_Ground",
       sub="Distance vs Speed of Ground",
       xlab="Speed of Ground", ylab="Distance",
       pch=10, col="green"
)


plot ( flightsclean$distance~flightsclean$speed_air,
       main="Relationship between Distance & Speed_Ground",
       sub="Distance vs Speed of Air",
       xlab="Speed of air", ylab="Distance",
       pch=10,col="yellow"
)
```

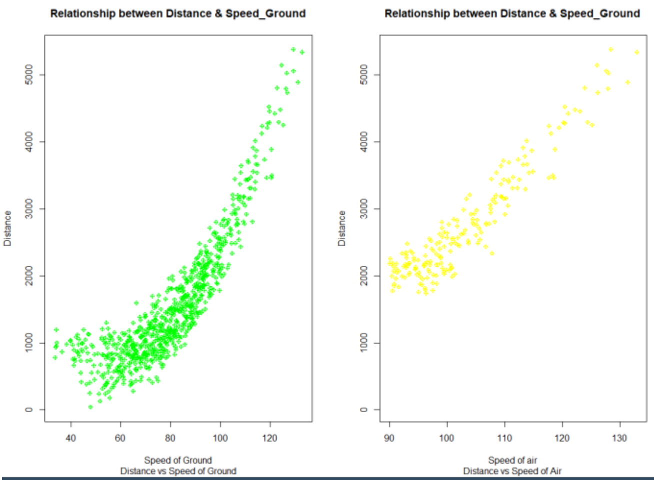

Relationship between Distance & height
Distance
0
1000
2000
3000
4000
5000
10
20
30
40
50
60
height
Distance vs heightr


Relationship between Distance & Number of passengers
Distance
0
1000
2000
3000
4000
5000
30
40
50
60
70
80
no_pasg
Distance vs Number of Passengers


Relationship between Distance & pitch
Distance
0
1000
2000
3000
4000
5000
3
4
5
6
pitch
Distance vs pitch


Relationship between Distance & duration
Distance
0
1000
2000
3000
4000
5000
50
100
150
200
250
300
duration
Distance vs duration


Relationship between Distance & aircraft
Distance
0
1000
2000
3000
4000
5000
airbus
boeing
aircraft
Distance vs aircraft



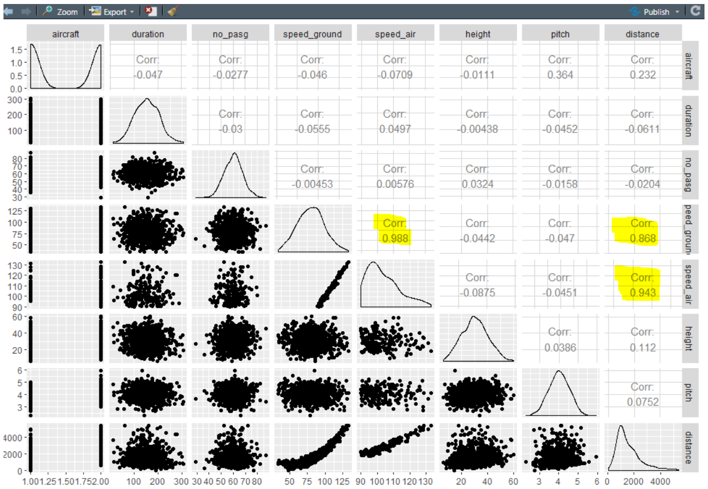
Observations:

- The distribution of all variables is random except speed air and ground speed.
- Speed air and speed ground show strong correlation when plotted against the variable distance.

STEP 3: STUDYING CO-RELATION BETWEEN DIFFERENT VARIABLES

I will study the co-relation between different coefficients using the 'ggpairs' functions. For this I will need to install package'Ggally' first. Aircraft type was then assigned numerical values. A value of '2' is assigned for boeing aircrafts and '1' for Airbus.

```
#ASSIGNINIG NUMERIC VALUE TO AIRCRAFT TYPE
flightsclean$aircraft<-as.numeric(factor(flightsclean$aircraft))
flightsclean$aircraft
flightscoded
#correlation


#Corrleation coefficients
install.packages("GGally")
library(GGally)

ggpairs(flightsclean)
```



From the results, I can draw the below table:

Variable	Correlation coefficient with distance	Direction of Correlation
aircraft	0.232	Positive
duration	0.0611	Negative
no_pasg	0.0204	Negative
speed_ground	0.868	Positive
speed_air	0.943	Positive
height	0.112	Positive
pitch	0.0752	Positive

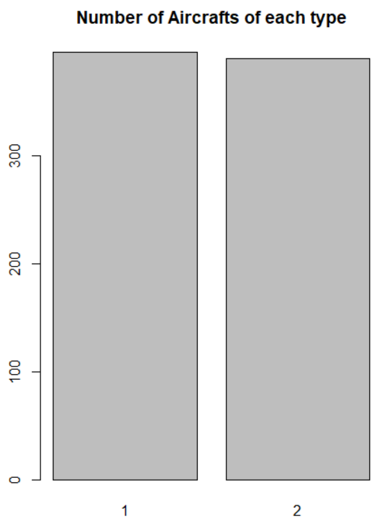
Observations:

1. High co-relation between speed air and distance.
2. High correlation between speed ground and distance.

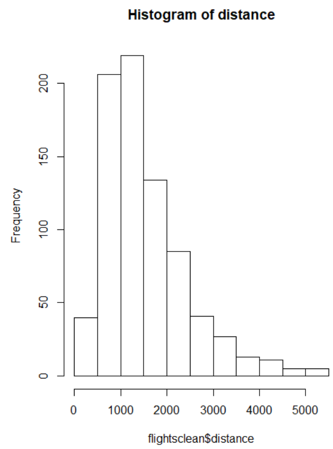
Step 4: STUDYING DISTRIBUTION OF ALL VARIABLES

I will use the Bar plot function to study the distribution of aircrafts and Histogram function to study other variables.

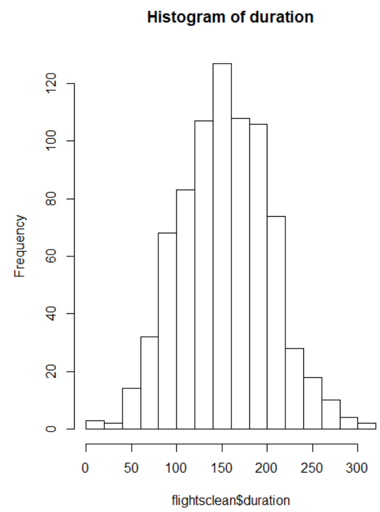
```
#histogram of all variables
barplot(table(flightsclean$aircraft), main = "Number of Aircrafts of each type")
hist(flightsclean$distance, main = "Histogram of distance")
hist(flightsclean$duration, main = "Histogram of duration")
hist(flightsclean$no_pasg, main = "Histogram of number of passengers")
hist(flightsclean$speed_air, main = "Histogram of speed of air")
hist(flightsclean$speed_ground, main = "Histogram of speed of ground")
hist(flightsclean$height, main = "Histogram of Height")
hist(flightsclean$pitch, main = "Histogram of Pitch")
```



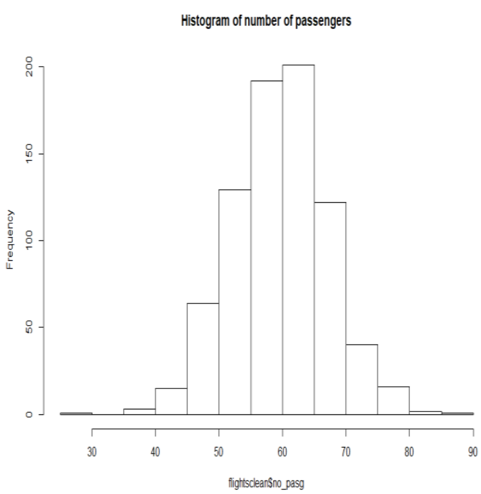
Number of Aircrafts of each type
0
100
200
300
1
2



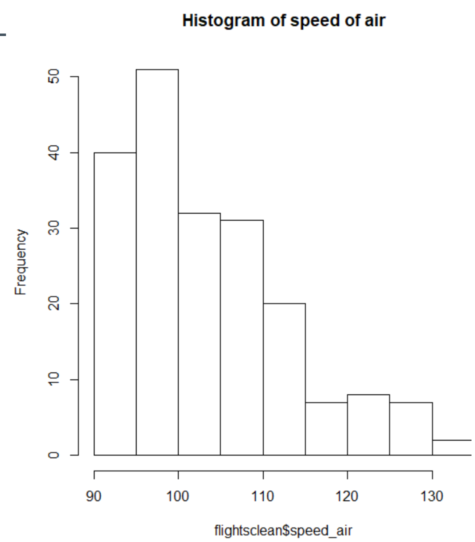
Histogram of distance
Frequency
flightsclean$distance



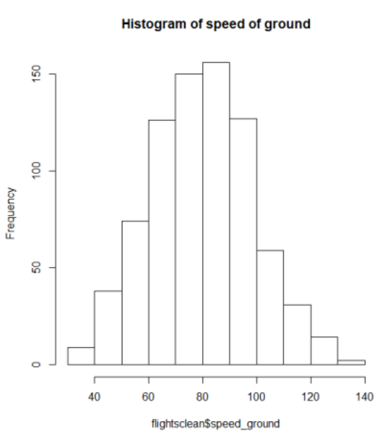
Histogram of duration
Frequency
flightsclean$duration



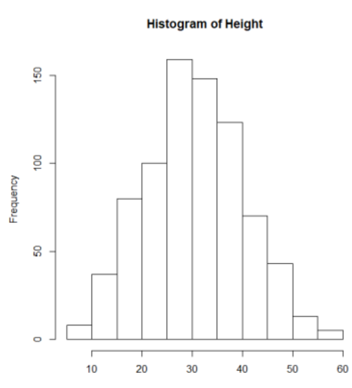
Histogram of number of passengers
Frequency
flightsclean$no_pasg



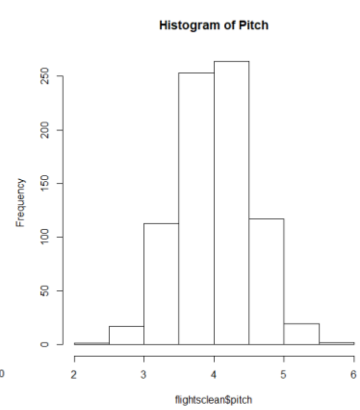
Histogram of speed of air
Frequency
flightsclean$speed_air



Histogram of speed of ground
Frequency
flightsclean$speed_ground



Histogram of Height
Frequency
flightsclean$height



Histogram of Pitch
Frequency
flightsclean$pitch

Observations:

1. Speed air and distance show right skewed distribution.
2. Speed air values are from 90-140, values below 90 are missing.
3. All other variables appear normally distributed.
4. Number of airbus aircrafts is slightly higher than Boeing.

CHAPTER 3: STATISTICAL MODELING

Goals: To use a linear regression model and study the relationship of the dependent variable (distance) with independent variables (aircraft, duration, no. of passengers, speed air, speed ground, pitch and height).

STEP 1: REGRESSION ANALYSIS OF EACH INDEPENDENT VARIABLE WITH DISTANCE

```
#Modelling each variable individually with distance
modelspeedground<-lm(distance ~ speed_ground,data=flightsclean)
summary(modelspeedground)

modelspeedair<-lm(distance ~ speed_air,data=flightsclean)
summary(modelspeedair)

modelduration<-lm(distance ~ duration,data=flightsclean)
summary(modelduration)

modelheight<-lm(distance ~ height,data=flightsclean)
summary(modelheight)

modelspeedground<-lm(distance ~ speed_ground,data=flightsclean)
summary(modelspeedground)

modelno_pasg<-lm(distance ~ no_pasg,data=flightsclean)
summary(modelno_pasg)

modelpitch<-lm(distance ~ pitch,data=flightsclean)
summary(modelpitch)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-912.18 -318.67  -74.54  216.85 1772.00

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1775.1067    69.6799  -25.48   <2e-16 ***
speed_ground    41.6591     0.8508   48.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 450.4 on 784 degrees of freedom
Multiple R-squared:  0.7536,    Adjusted R-squared:  0.7533
F-statistic:  2398 on 1 and 784 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-787.22 -189.21   -0.59  214.63  618.36

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5418.357    208.664  -25.97   <2e-16 ***
speed_air      79.288      2.008   39.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 277.1 on 196 degrees of freedom
  (588 observations deleted due to missingness)
Multiple R-squared:  0.8884,	Adjusted R-squared:  0.8878
F-statistic:  1560 on 1 and 196 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1461.9  -614.8  -278.9   418.6  3846.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1717.6364   105.8881  16.221   <2e-16 ***
duration      -1.1223     0.6551  -1.713   0.0871 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 905.5 on 784 degrees of freedom
Multiple R-squared:  0.003729,	Adjusted R-squared:  0.002459
F-statistic: 2.935 on 1 and 784 DF,  p-value: 0.08709
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1339.6  -613.7  -249.3   418.4  3927.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1226.602    105.585  11.617  < 2e-16 ***
height        10.432      3.296   3.165  0.00161 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 901.5 on 784 degrees of freedom
Multiple R-squared:  0.01261,	Adjusted R-squared:  0.01135
F-statistic: 10.02 on 1 and 784 DF,  p-value: 0.001612
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1294.5  -637.7  -232.4   392.0  3625.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   915.82      99.31   9.222  < 2e-16 ***
aircraft      420.44      62.96   6.678 4.57e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 882.5 on 784 degrees of freedom
Multiple R-squared:  0.05383,	Adjusted R-squared:  0.05262
F-statistic:  44.6 on 1 and 784 DF,  p-value: 4.574e-11
```

Gupta2na Niharika Gupta M13437287


```
Residuals:
    Min      1Q  Median      3Q     Max
-1461.5  -629.4  -265.8   415.4  3866.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1692.695    260.831   6.490 1.52e-10 ***
no_pasg       -2.461      4.309  -0.571    0.568
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 907 on 784 degrees of freedom
Multiple R-squared:  0.0004159,	Adjusted R-squared:  -0.0008591
F-statistic: 0.3262 on 1 and 784 DF,  p-value: 0.5681
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1359.8  -650.1  -252.3   400.4  3820.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1021.6      249.8   4.090 4.76e-05 ***
pitch          130.4       61.7   2.113   0.0349 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 904.7 on 784 degrees of freedom
Multiple R-squared:  0.005662,	Adjusted R-squared:  0.004393
F-statistic: 4.464 on 1 and 784 DF,  p-value: 0.03493
```

Observations:

- Considering a significance value of 0.05, 5 variables are significant i.e. Air_type, speed_ground, speed_air, pitch and height. I can drop duration and number of passengers from the model since they don't seem to have any impact on the distance variable.

STEP 2: REGRESSION ANALYSIS OF SIGNIFICANT VARIABLES TOGETHER WITH DISTANCE

```
#regression analysis of significant variables together with distance
model<-lm(distance ~ speed_ground+speed_air+pitch+height+aircraft,data=flightsclean)
summary(model)
```

```
Call:
lm(formula = distance ~ speed_ground + speed_air + pitch + height +
    aircraft, data = flightsclean)

Residuals:
    Min      1Q  Median      3Q     Max
-297.66  -93.98   13.16   91.25  339.23

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6763.290    132.732 -50.955   <2e-16 ***
speed_ground    -4.581      6.336  -0.723    0.471
speed_air       86.560      6.437  13.446   <2e-16 ***
pitch          -12.673     18.472  -0.686    0.494
height          13.659      1.011  13.505   <2e-16 ***
aircraft       435.544     21.043  20.698   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 134.3 on 192 degrees of freedom
  (588 observations deleted due to missingness)
Multiple R-squared:  0.9743,	Adjusted R-squared:  0.9737
F-statistic:  1457 on 5 and 192 DF,  p-value: < 2.2e-16
```

Observations:

- From the observations, I derived the below formula:

  Distance= -6763.290 + (-4.581* speed_ground) +(86.560*speed_air) +(13.659*height) +(-12.673*pitch) +(435.544*Aircraft)

- Considering a significance level of 0.05 (i.e. 5%) or less I can say that only variables speed_air, height, and Aircraft are significant.
- R_sq (0.9743) is a high value, that means that the model fits the data well.
- The significance value of the Pitch variable (0.494) suggests that it does not fit in my model and needs to be removed.
- Significance value for speed ground (0.471) no longer fits in my previous already inferred correlation between speed_ground and speed_air. Since speed_air has a stronger co-relation with distance compared to speed_ground and the direction of correlation is positive, only higher speed air values can cause landing over run, thus missing speed air values (less than 90) won't have any impact on the analysis. I will thus drop speed ground from my study.

STEP 3: REGRESSION ANALYSIS AFTER REMOVING SPEED GROUND AND PITCH

```
#regression analysis after removing speed ground and pitch
model.2<-lm(distance ~ speed_air+height+aircraft,data=flightsclean)
summary(model.2)
```

```
> summary(model.2)

Call:
lm(formula = distance ~ speed_air + height + aircraft, data = flightsclean)

Residuals:
    Min      1Q  Median      3Q     Max
-294.00  -93.88   11.22   89.71  335.67

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6807.4619   115.5793  -58.90   <2e-16 ***
speed_air      81.9725     0.9766   83.94   <2e-16 ***
height         13.7200     1.0054   13.65   <2e-16 ***
aircraft      430.6205    19.5442   22.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 133.9 on 194 degrees of freedom
  (588 observations deleted due to missingness)
Multiple R-squared:  0.9742,    Adjusted R-squared:  0.9738
F-statistic:  2442 on 3 and 194 DF,  p-value: < 2.2e-16
```

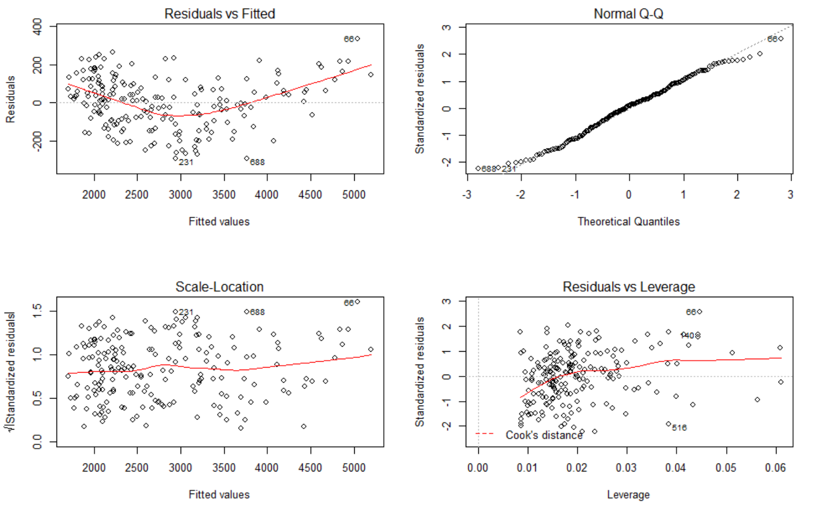
Observation:

- The model obtained using the above variables is the following:

  Distance= -6807.4619+ (81.9725*speed_air) + (13.7200*height) + (430.6205*aircraft)

- R square value of 0.9742 indicates my model is still fit.

STEP 4: MODEL DIAGNOSTICS



Observations:

QQ Plot shows that the residuals are normally distributed.

STEP 5: REGRESSION ANALYSIS by Air Craft

Next I will perform regression analysis for each type of aircraft separately to see if the results vary.

STEP 5.1: CREATING SEPARATE DATASETS FOR EACH AIRCRAFT TYPE

Two separate datasets are created for each aircraft type.

STEP 5.2: CREATING LINEAR MODELS FOR EACH AIRCRAFT TYPE

```
#dividing datasets by Aircarft Type for Induvidual Analysis

flightsboeing<-flightsclean[which(flights$aircraft=="boeing"), ]
flightsboeing

flightsairbus<-flightsclean[which(flights$aircraft=="airbus"), ]
flightsairbus
```

```
> model.airbus<-lm(distance ~ speed_air+pitch+height,data=flightsairbus)
> summary(model.airbus)

Call:
lm(formula = distance ~ speed_air + pitch + height, data = flightsairbus)

Residuals:
    Min      1Q  Median      3Q     Max 
-200.31  -75.98   -8.41   89.74  333.43 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6864.039    205.671 -33.374  < 2e-16 ***
speed_air      82.179      1.633  50.327  < 2e-16 ***
pitch         123.149     26.037   4.730 1.13e-05 ***
height         13.639      1.404   9.711 1.33e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 118.1 on 70 degrees of freedom
  (326 observations deleted due to missingness)
Multiple R-squared:  0.9739,	Adjusted R-squared:  0.9728 
F-statistic: 871.5 on 3 and 70 DF,  p-value: < 2.2e-16

> 
> 
> model.boeing<-lm(distance ~ speed_air+pitch+height,data=flightsboeing)
> summary(model.boeing)

Call:
lm(formula = distance ~ speed_air + pitch + height, data = flightsboeing)

Residuals:
    Min      1Q  Median      3Q     Max 
-491.04  -83.83   11.42   93.68  335.29 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5649.355    179.224 -31.521  < 2e-16 ***
speed_air      82.294      1.265  65.050  < 2e-16 ***
pitch         -81.305     25.489  -3.190  0.00182 ** 
height         13.685      1.422   9.622  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 146.8 on 120 degrees of freedom
  (276 observations deleted due to missingness)
Multiple R-squared:  0.9725,	Adjusted R-squared:  0.9718 
F-statistic:  1416 on 3 and 120 DF,  p-value: < 2.2e-16
```

Observations:

- Pitch is significant for both the aircrafts in contrast to what was observed in the dataset with both aircrafts. Parameter estimate for airbus is positive and for Boeing is negative. That could be the reason that it was non-significant in the dataset with both the aircrafts together.
- Based on this, we can derive the below formula:

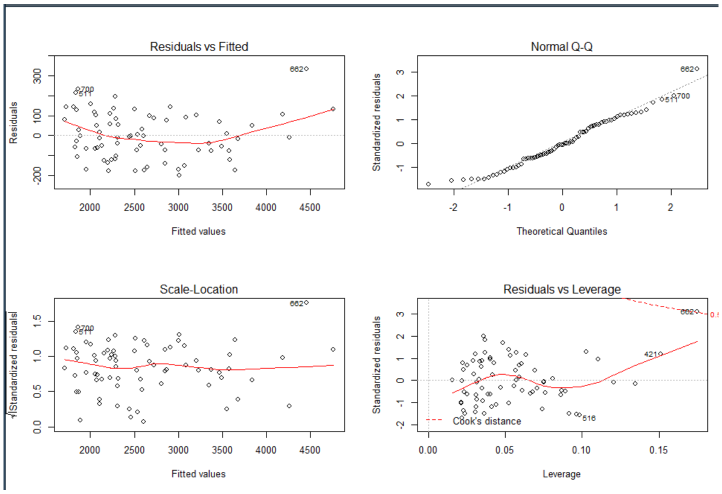
  Airbus: Distance= -6964.039 + (82.179* Speed_air) +(13.639*height) +(123.149*pitch)

- Boeing: Distance= -5649.355 + (82.294* Speed_air) +(13.685*height) +(-81.305*pitch)
- High R square values for airbus and Boeing (0.9739, 0.9725) means the model fits well.

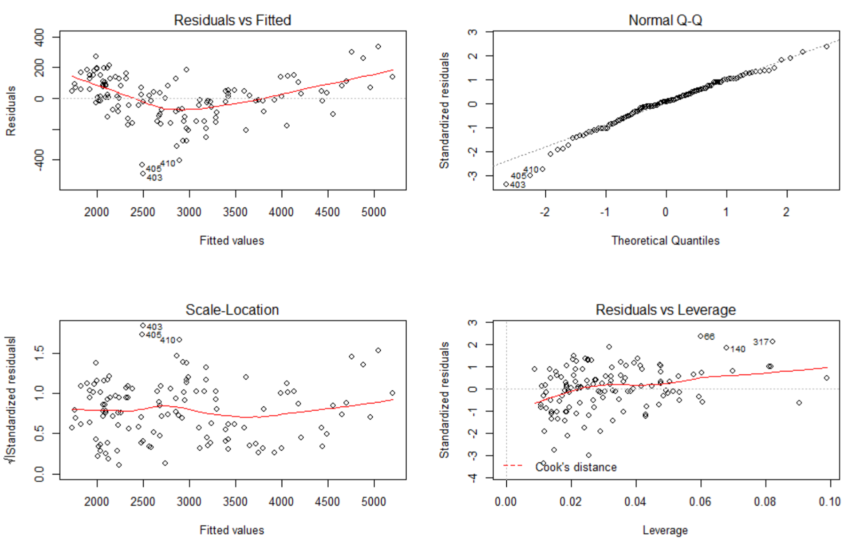
STEP 6: MODEL DIAGNOSTICS

```
#model diagnostics

par(mfrow=c(2,2))
plot(model.airbus)

par(mfrow=c(2,2))
plot(model.boeing)
```

AIRBUS



BOEING



Observations:

- Residuals are normally distributed for both the makes.