```
rm(list=ls())
library(tidyverse)
library(readxl)
library(dplyr)
require(gdata)
```

###INITIAL EXPLORATION OF THE DATA

STEP 1:

```
FAA1 = read.xls ("FAA1.xls", sheet = 1, header = TRUE)
FAA2 = read.xls ("FAA2.xls", sheet = 1, header = TRUE)
```

STEP 2:

```
head(FAA1) # There are 8 variables: aircraft, duration, no_psg, speed_ground
           #speed_air, height, pitch, distance
str(FAA1) # 1. There are 800 observations
          #2. Two kinds of aircraft:airbus and boeing.Aircraft is factor variable
          #3. no_psg is integer variable and rest are numeric.
head(FAA2) # Same variables are FAA1 except duration which is absent in
FAA2
str(FAA2)  # 1. 150 Observations.
           # 2. The data type of variable same as FAA1

#Merging the two data frames row-wise
```

STEP 3:

```
FAA <- bind_rows(FAA1,FAA2)
# Removing Duplicates
FAA <-
FAA[!duplicated(FAA[c('aircraft','no_pasg','speed_ground','speed_air','height',
         'pitch','distance')]),]  # There were 100 elements
          # where duplicates were found and removed
```

STEP 4 and STEP 5:

```
summary(FAA)

# 1. There are 450 airbus and 400 boeing flights
# 2. The duration column has 50 NA's. The minimum duration is 14.76 and
maximum
#is 305.62. The mean is 154.01 and median is 154.01
```

```
# 3. The no_pasg is minimum 29 and maximum 87. The mean and median are
60 and 60.1
#  4. The speed_ground has minimum of 27.74, maximum of 141.22 and mean
and
#median of 79.64 and 79.45.
# 5.The speed_air has minimum 90, maximum of 141.72. The mean and median
values
#are 101.15 and 103.80. There is some difference between the mean and median
value
#The number of NAs is 642, which is very high. Therefore the data sample is
#less and hence the mean and median may not be accurate. Also the summary
statistic
#between speed_air and speed_ground is differing a lot.
#6.The height variable has mimium of -3.546, which is not possible. The
maximum is
#59.946. The mean and median are 30.144 and 30.093.
#7.The pitch variable has minimum 2.284, max value of 5.927. The mean and
median are
#4.009 and 4.008.
#8.The distance variable has minimum of 34.08, maximum of 6533.05 and
mean and
#median of 1526.02 and 1258.09. The minimum value is too small. The mean
and median
#are differing a lot showing the presence of outliers in the data set.
```

###DATA CLEANING AND FURTHER EXPLORATION

STEP 6:

```
FAA <- subset(FAA, (FAA$duration>=40 | is.na(FAA$duration))) # Five rows
were removed
FAA <- subset(FAA,((FAA['speed_air']<=140 & FAA['speed_air']>=30)|
                is.na(FAA$speed_air))) # one observation removed
FAA <- subset(FAA,((FAA['speed_ground']<=140 &
FAA['speed_ground']>=30)|
                is.na(FAA$speed_ground))) # two observation removed

FAA <- subset(FAA, (FAA['height']>=6 | is.na(FAA$height)))  #  10
observations removed
```

In the above steps only abnormal observations were removed and rows with NA’s
Where retained.

STEP 7:

```
str(FAA)
```

There are 831 rows of 8 variables.

```
summary(FAA$aircraft)
airbus boeing
   444    388

summary(FAA$duration)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  41.95  119.70  154.30  154.70  189.60  305.60      50

summary(FAA$no_pasg)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29.00   55.00   60.00   60.06   65.00   87.00

summary(FAA$speed_ground)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  33.57   66.20   79.83   79.61   91.99  136.70

summary(FAA$speed_air)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  90.00   96.25  101.10  103.60  109.40  136.40     628
summary(FAA$height)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.228  23.530  30.180  30.470  37.020  59.950
summary(FAA$pitch)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.284   3.641   4.002   4.005   4.370   5.927
summary(FAA$distance)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  41.72  893.30 1262.00 1522.00 1937.00 5382.00
```

STEP 8:

```
ggplot(data=FAA, aes(FAA$speed_ground)) + geom_histogram(binwidth = 5)
ggplot(data=FAA, aes(FAA$speed_air)) + geom_histogram(binwidth = 5)
ggplot(data=FAA, aes(FAA$height)) + geom_histogram(binwidth = 5)
ggplot(data=FAA, aes(FAA$pitch)) + geom_histogram(binwidth = 0.2)
ggplot(data=FAA, aes(FAA$distance)) + geom_histogram(binwidth = 200)
```

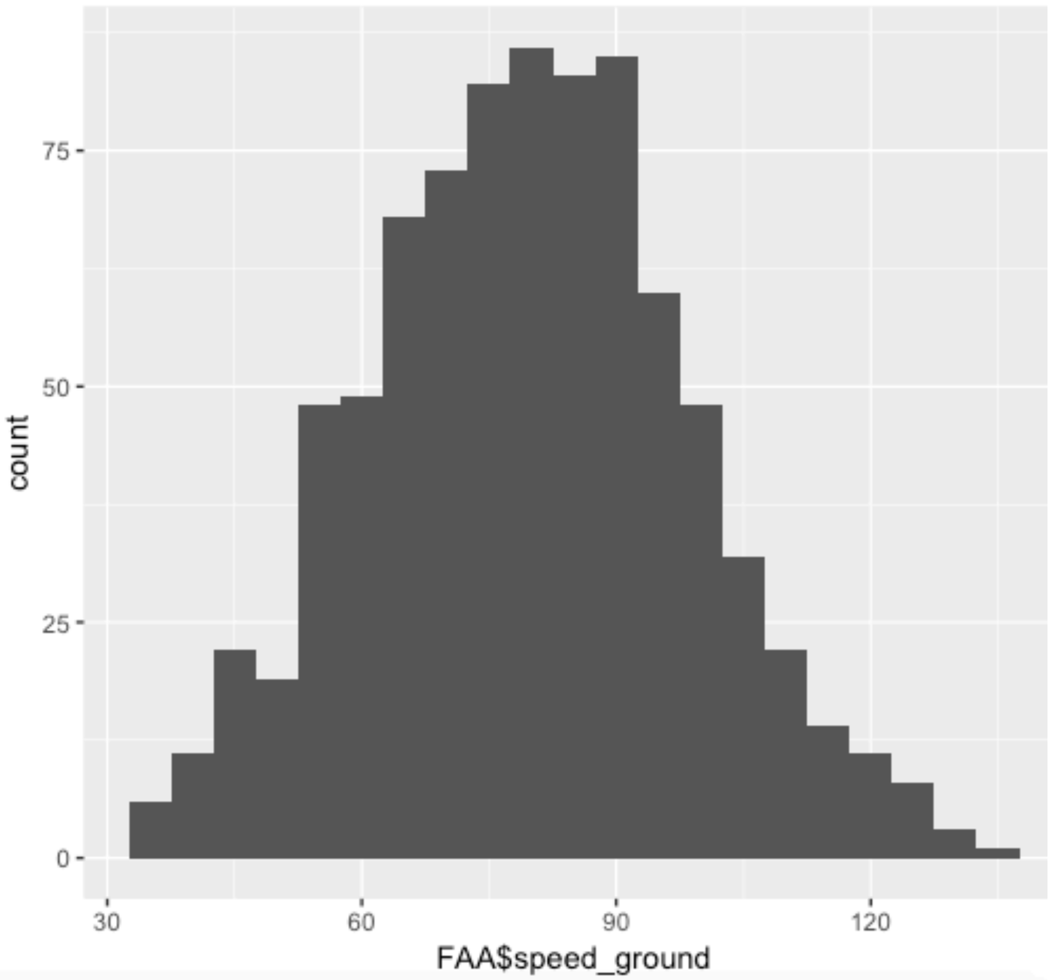

Fig 1a. Histogram of Speed_ground parameter.

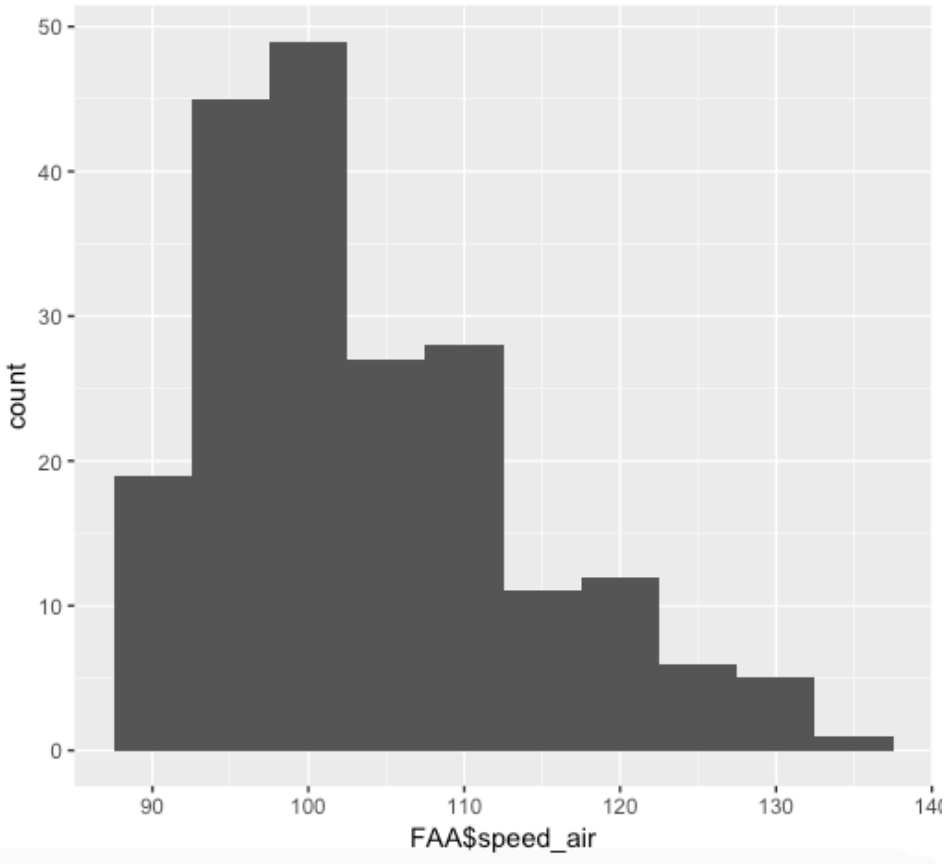

Fig 1b. Histogram of Speed_air parameter

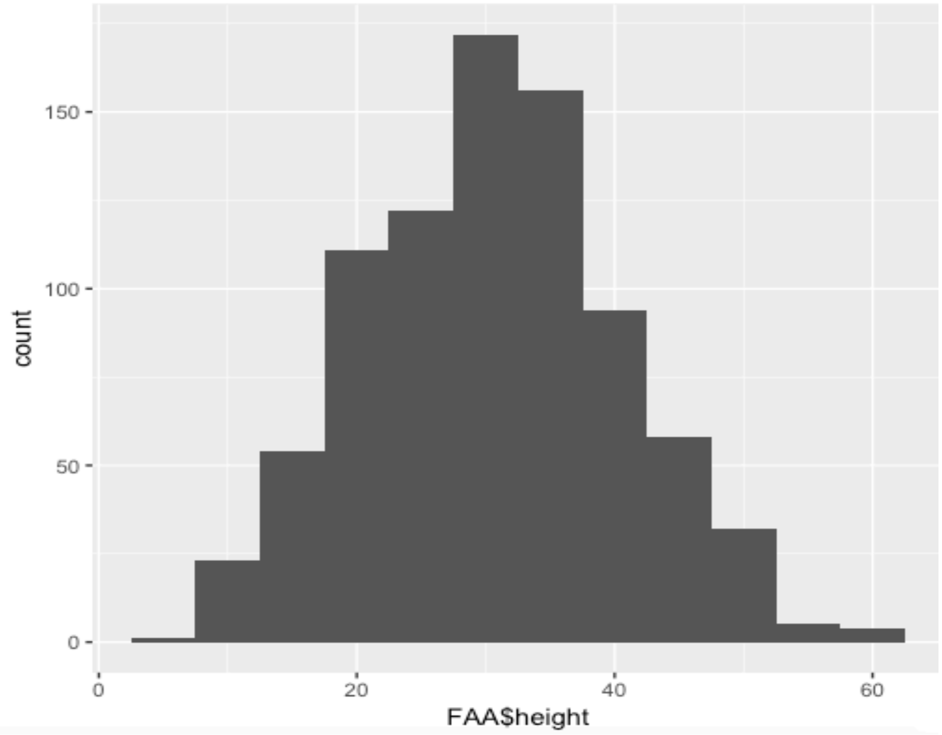

Fig 1c. Histogram of Height Parameter

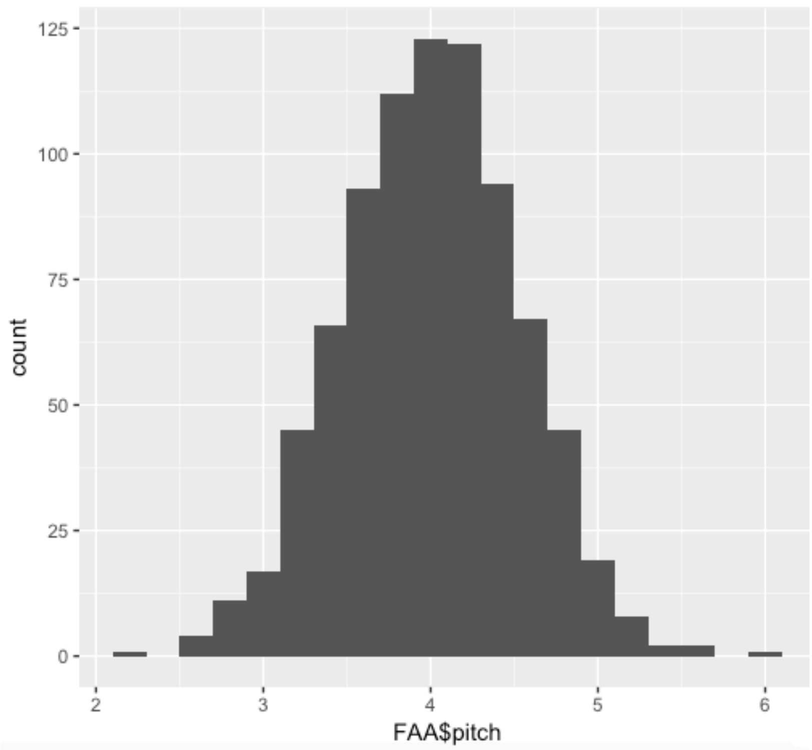

Fig 1d. Histogram of Pitch Parameter

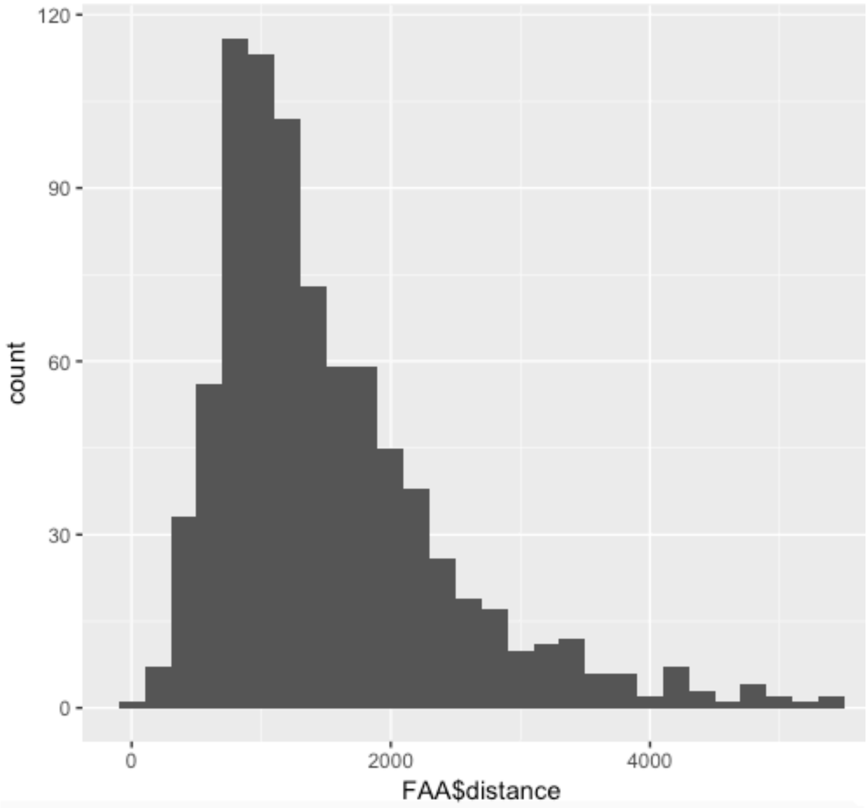

Fig 1e. Histogram of Distance Parameter

STEP 9:

1. From the summary statistics it is clear that the data has no abnormal values now.
2. The histogram plot shows that the Speed_air is skewed and is very much different from Speed_ground. This is because it contains 628 NA values and hence the sampling data for it is very less compared to Speed_air.
3. The distance variable is also skewed to the left.
4. The pitch, height and speed_ground variables are normally distributed.
5. The mean value of the speed_ground is around 79, height is around 30, pitch is around 4 and distance is around 1522.
6. The number of passengers and duration variables have not been included as they do not affect landing distance by common knowledge.

Initial analysis for identifying important factors that impact the response variable "landing distance"

STEP 10:

```
> cor(FAA$distance, FAA$speed_air, use = "pairwise.complete.obs")
[1] 0.9420971
> cor(FAA$distance, FAA$speed_ground, use = "pairwise.complete.obs")
[1] 0.8662438
> cor(FAA$distance, FAA$height, use = "pairwise.complete.obs")
[1] 0.09941121
> cor(FAA$distance, FAA$pitch, use = "pairwise.complete.obs")
[1] 0.08702846
> cor(FAA$distance, FAA$no_pasg, use = "pairwise.complete.obs")
[1] -0.01775663
> cor(FAA$distance, FAA$duration, use = "pairwise.complete.obs")
[1] -0.05138252
```

Variables	Size of Correlation	Direction of Correlation
Distance, Speed_air	0.942	Positive
Distance, Speed_ground	0.866	Positive
Distance, Height	0.1	Positive
Distance, Pitch	0.09	Positive
Distance, no_pasg	-0.02	negative
Distance, duration	-0.05	negative

Table1: Correlation between distance and other variables

STEP 11:

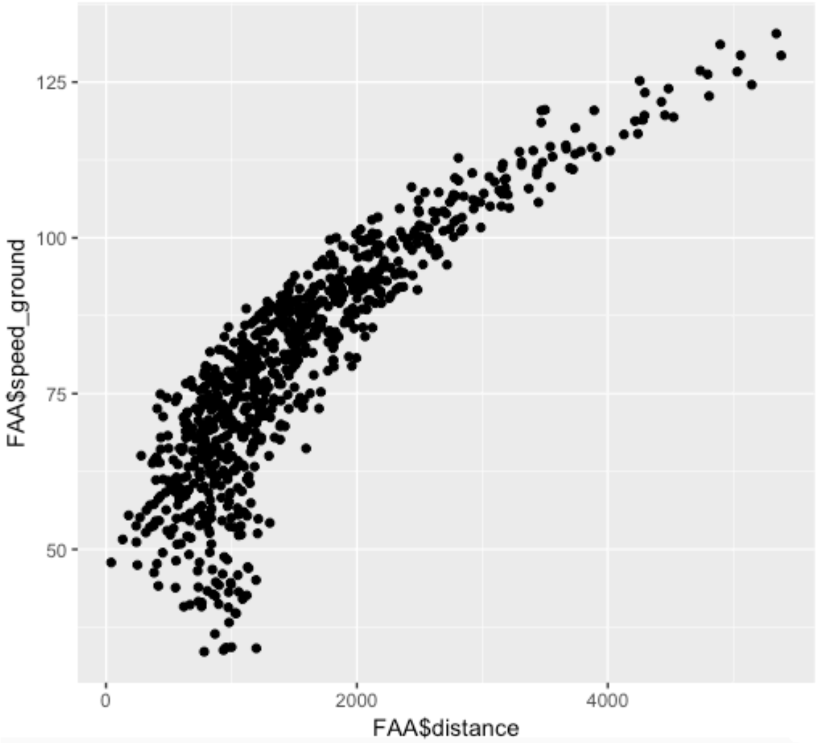

Fig 2a. Scatter plot of Distance vs Speed_ground

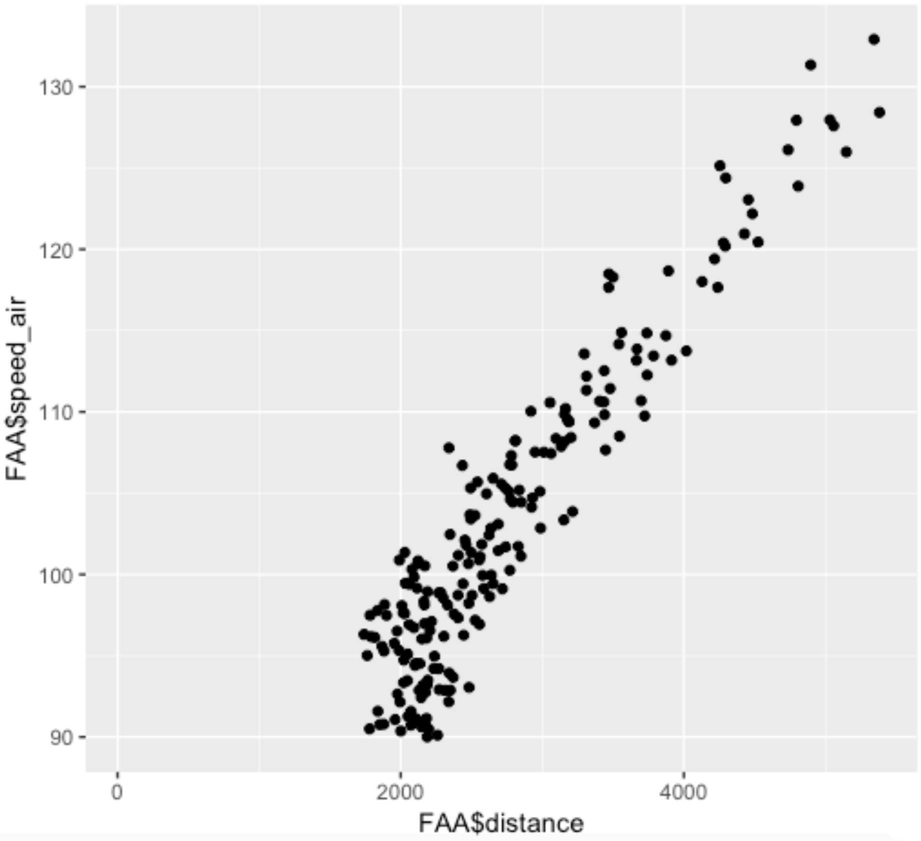

Fig 2b. Scatter plot of Distance vs Speed_air

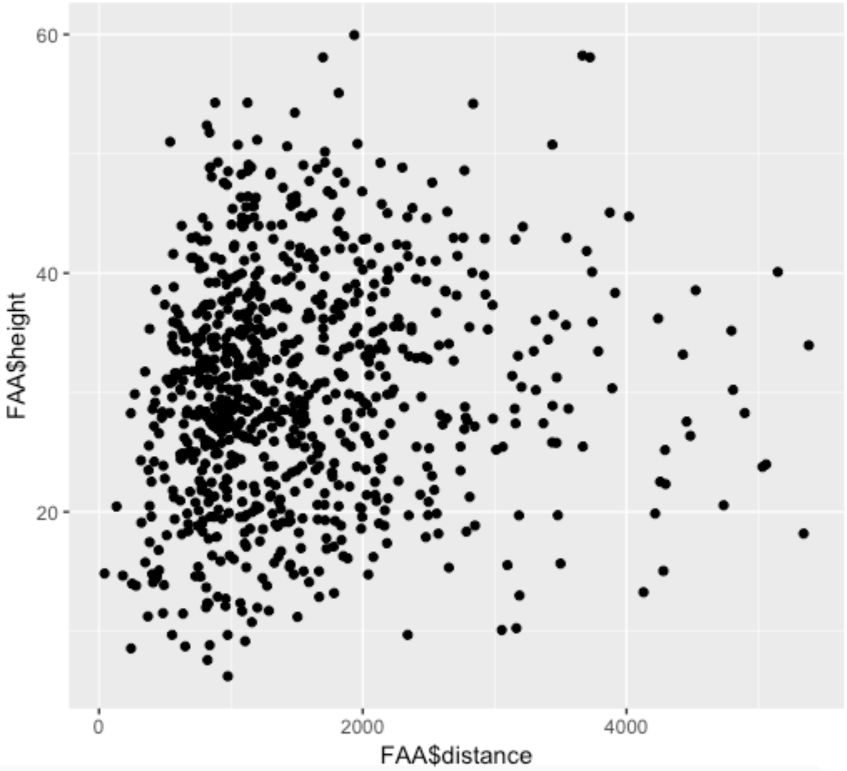

Fig 2c. Scatter plot of Distance vs Height

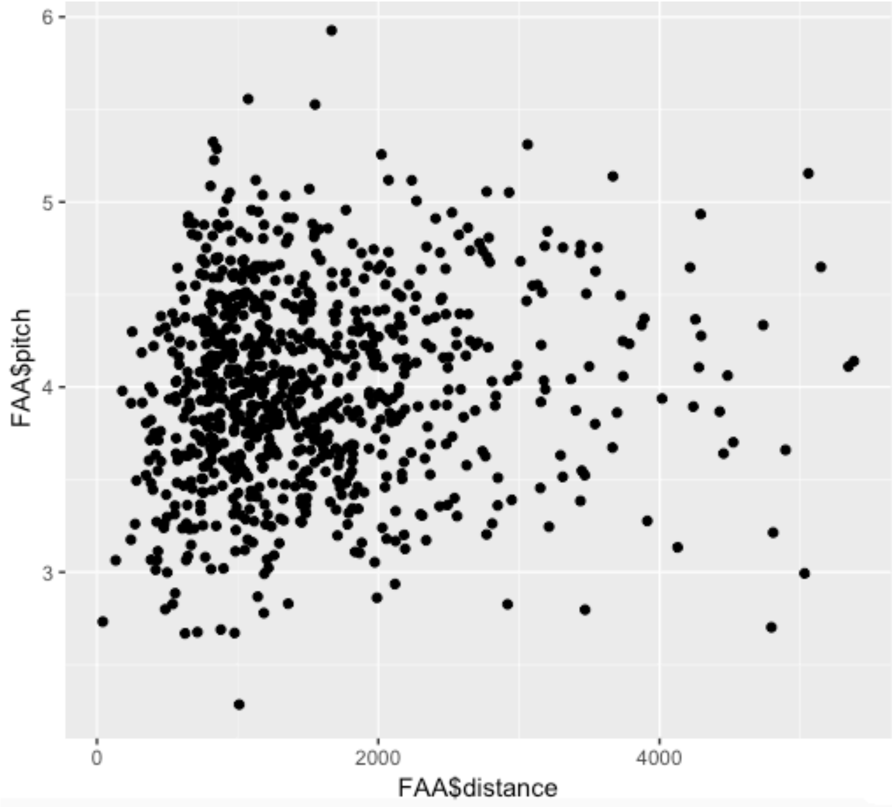

Fig 2d. Scatter plot of Distance vs Pitch

From the above plots it is evident that the correlation coefficients found in Table 1 is consistent.

Regression using a single factor each time

STEP 13:

Variables	p-Value	Direction
Speed_air	<2e-16	Positive
Height	<2e-16	Positive
Aircraft Boeing	<2e-16	Positive
No_pasg	0.152	Negative
Pitch	0.469	Negative
duration	0.532	Positive
Speed_ground	0.581	Negative

Table2: Linear Regression model with all the variables

STEP 14:

```
FAA$speed_ground <- (FAA$speed_ground-mean(FAA$speed_ground, na.rm
= TRUE))/sd(FAA$speed_ground, na.rm = True)

FAA$speed_air <- (FAA$speed_air-mean(FAA$speed_air, na.rm =
TRUE))/sd(FAA$speed_air, na.rm = TRUE)

FAA$height <- (FAA$height-mean(FAA$height, na.rm =
TRUE))/sd(FAA$height, na.rm = TRUE)

FAA$pitch <- (FAA$pitch-mean(FAA$pitch, na.rm = TRUE))/sd(FAA$pitch,
na.rm = TRUE)

FAA$no_pasg <- (FAA$no_pasg-mean(FAA$no_pasg, na.rm =
TRUE))/sd(FAA$no_pasg, na.rm = TRUE)

FAA$duration <- (FAA$duration-mean(FAA$duration, na.rm =
TRUE))/sd(FAA$duration, na.rm=TRUE)
```

Variables	**Coefficient value**	**Direction**
Speed_air	832.908	Positive
Aircraft Boeing	437.943	Positive
Height	133.813	Positive
No_pasg	-14.842	Negative
Pitch	-7.103	Negative
duration	6.171	Positive
Speed_ground	-3.546	Negative

TABLE 3: Coefficients after standardizing of variables

STEP 15:

Variables	**Ranking**
Speed_air	1
Aircraft Boeing	2
Height	3
No_pasg	4
Pitch	5
duration	6
Speed_ground	7

Table 0: Ranking of coefficients

STEP 16:

```
Model1 <- lm(distance ~ speed_ground, data=FAA)
Model2 <- lm(distance ~ speed_air, data=FAA)
Model3 <- lm(distance ~ speed_ground+speed_air, data=FAA)
```

```
summary(Model1)

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1773.9407    67.8388  -26.15   <2e-16 ***
speed_ground    41.4422     0.8302   49.92   <2e-16 ***
```

```
summary(Model2)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2774.67      19.39  143.07   <2e-16 ***
speed_air     774.35      19.44   39.83   <2e-16 ***
```

```
summary(Model3)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4261.05    1311.78   3.248  0.00136 **
speed_ground   -14.37      12.68  -1.133  0.25848
speed_air      914.81     125.46   7.291 6.99e-12 ***
```

The coefficient of speed_ground changes when speed_air is added into the model and also its p-value is increased when speed_air is added into the model.

```
> cor(FAA$speed_air, FAA$speed_ground, use = "pairwise.complete.obs")
[1] 0.9879383
```

The two variables are highly correlated. I would choose speed_ground because it has more data points and is a full normal distribution.

STEP 17:

```
M1 <- lm(distance ~ speed_ground, data=FAA)
M2 <- lm(distance ~ speed_ground+aircraft, data=FAA)
M3 <- lm(distance ~ speed_ground+aircraft+height, data=FAA)
M4 <- lm(distance ~ speed_ground+aircraft+height+no_pasg,
data=FAA)
```

```
M5 <- lm(distance ~ speed_ground+aircraft+height+no_pasg+pitch, data=FAA)
M6 <- lm(distance ~ speed_ground+aircraft+height+no_pasg+pitch+duration, data=FAA)
```

Variables	R-squared
M1	0.7504
M2	0.8251
M3	0.8489
M4	0.8492
M5	0.8497
M6	0.8506

STEP18:

Variables	Adjusted R-squared
M1	0.7501
M2	0.8247
M3	0.8484
M4	0.8485
M5	0.8488
M6	0.8494

STEP 19:

Variables	AIC
M1	12508.81
M2	12215.05
M3	12095.65
M4	12095.73
M5	12095.18
M6	11379.88

STEP 20:

From the above tables I would choose speed_ground, aircraft and height parameters in making the model as that results in the lowest AIC and highest Adjusted R-squared values.

STEP 21:

```
AIC <- stepAIC(M6, direction = 'forward')
summary(AIC)
```

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2091.092     56.698 -36.881   <2e-16 ***
speed_ground      42.567      0.668  63.719   <2e-16 ***
aircraftboeing   488.763     26.995  18.106   <2e-16 ***
height           139.791     12.665  11.038   <2e-16 ***
no_pasg          -12.231     12.532  -0.976    0.329
pitch             10.346     13.620   0.760    0.448
duration           2.261     12.614   0.179    0.858
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 351 on 774 degrees of freedom
  (50 observations deleted due to missingness)
Multiple R-squared:  0.8506,Adjusted R-squared:  0.8494
F-statistic: 734.5 on 6 and 774 DF,  p-value: < 2.2e-16
```

The p-values of the parameters confirms our previous choices of parameters. Hence, speed_ground, aircraft and height parameters are chosen for building the model.