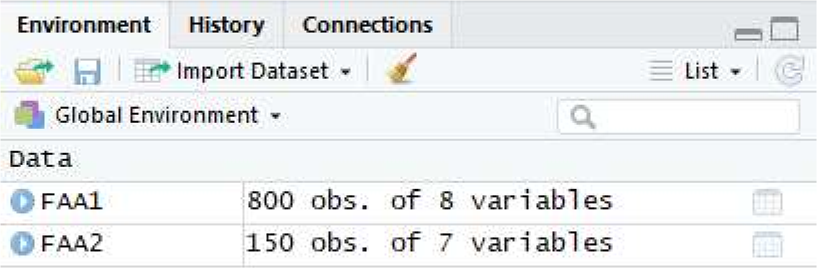
Statistical Modeling Project 1
Ashwita Saxena
M06119969

Background: Flight landing. Motivation: To reduce the risk of landing overrun.
Goal: To study what factors and how they would impact the landing distance of a commercial flight.
Data: Landing data (landing distance and other parameters) from 950 commercial flights (not real data set but simulated from statistical models). See two Excel files 'FAA-1.xls' (800 flights) and 'FAA-2.xls' (150 flights).

Step 1: Read the two files 'FAA-1.xls' (800 flights) and 'FAA-2.xls' into your R system. Please search "Read Excel files from R" in Google in case you do not know how to do that.

```
# importing data ----------------------------------------------------------

library(readxl)
FAA1 <- read_xlsx("D:/MSBA/Spring Sem/statistical modeling/FAA1(1).xlsx",)
FAA2 <- read_xlsx("D:/MSBA/Spring Sem/statistical modeling/FAA2(1).xlsx",)
```

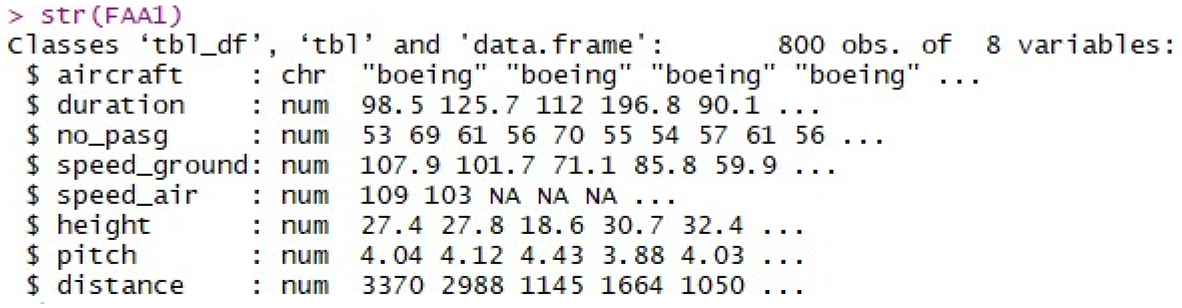


```
Data
FAA1    800 obs. of 8 variables
FAA2    150 obs. of 7 variables
```

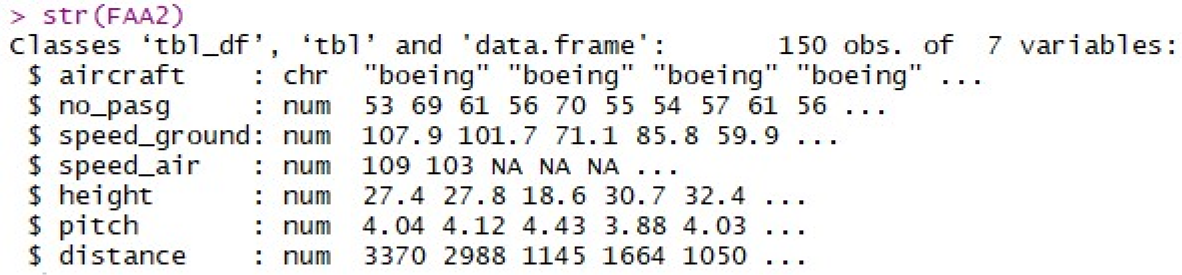
Observation: The two files have been imported into R.
Conclusion: FAA1 has 800 observations and 8 variables and FAA2 has 150 observations and 7 variables.

Step 2: Check the structure of each data set using the "str" function. For each data set, what is the sample size and how many variables? Is there any difference between the two data sets?

```
#### step 2: structure
str(FAA1)
str(FAA2)
```



```
> str(FAA1)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       800 obs. of  8 variables:
 $ aircraft    : chr  "boeing" "boeing" "boeing" "boeing" ...
 $ duration    : num  98.5 125.7 112 196.8 90.1 ...
 $ no_pasg     : num  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 ...
```



```
> str(FAA2)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       150 obs. of  7 variables:
 $ aircraft    : chr  "boeing" "boeing" "boeing" "boeing" ...
 $ no_pasg     : num  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 ...
```

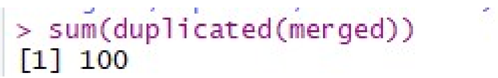
Observation: FAA1 has 800 observations and 8 variables and FAA2 has 150 observations and 7 variables. There are some differences between the two datasets.

1. FAA1 has a variable called duration which does not exist in the second dataset.
2. The number of observations in FAA1 is 800 whereas in FAA2 is 150.

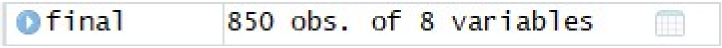
Conclusion: Both the datasets have 7 common variables and one of them has an extra variable.

Step 3. Merge the two data sets. Are there any duplications? Search “check duplicates in r” if you do not know how to check duplications. If the answer is “Yes”, what action you would take?

```
#### step 3: merging
library(dplyr)
FAA1_2 <- select(FAA1, aircraft, no_pasg, speed_ground, speed_air, height,
pitch, distance)

merged<- rbind(FAA1_2,FAA2)

sum(duplicated(merged))

new <- unique(merged)
summary(new)

# adding duration back in

final <- left_join(new, FAA1)
```



```
> sum(duplicated(merged))
[1] 100
```



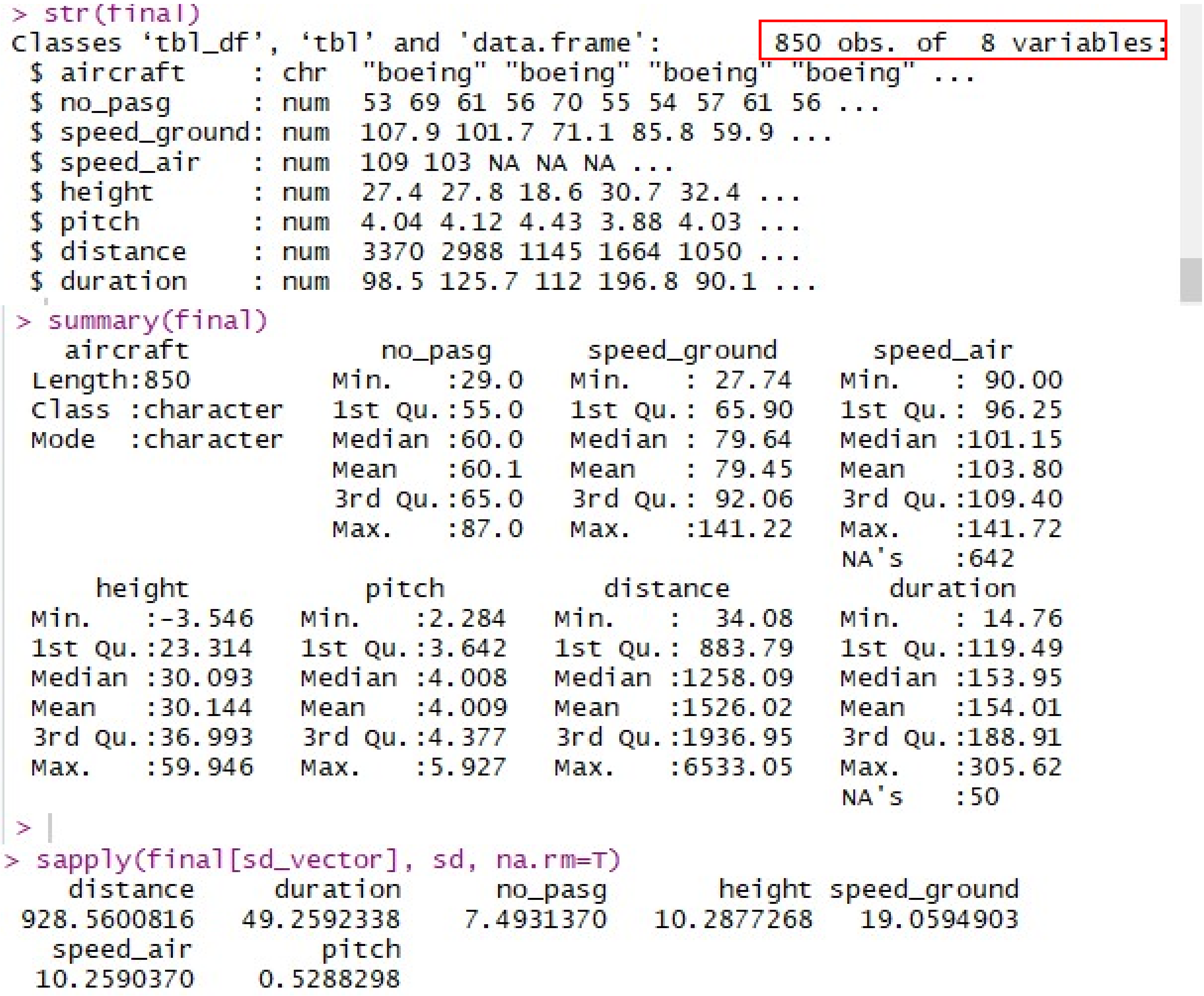
```
final          850 obs. of 8 variables
```

Observation: I removed duration from the first dataset to be able to combine the two datasets. There were 100 duplicate rows in the combined dataset. After removing the duplicates, duration was added back in to the final dataset. The final dataset has 850 observations and 8 variables.

Conclusion: 100 duplicate values were removed after merging the two datasets to come up with the combined dataset to perform further analysis.

Step 4: Check the structure of the combined data set. What is the sample size and how many variables? Provide summary statistics for each variable.

```
#### step 4: checking the final dataset
str(final)
summary(final)
#standard deviation
sd_vector <- c("distance","duration","no_pasg", "height", "speed_ground",
"speed_air", "pitch")
sapply(final[sd_vector], sd, na.rm=T)
```



```
> str(final)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':      850 obs. of  8 variables:
 $ aircraft    : chr  "boeing" "boeing" "boeing" "boeing" ...
 $ no_pasg     : num  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 ...
 $ duration    : num  98.5 125.7 112 196.8 90.1 ...
> summary(final)
   aircraft            no_pasg      speed_ground      speed_air     
 Length:850         Min.   :29.0   Min.   : 27.74   Min.   : 90.00  
 Class :character   1st Qu.:55.0   1st Qu.: 65.90   1st Qu.: 96.25  
 Mode  :character   Median :60.0   Median : 79.64   Median :101.15  
                    Mean   :60.1   Mean   : 79.45   Mean   :103.80  
                    3rd Qu.:65.0   3rd Qu.: 92.06   3rd Qu.:109.40  
                    Max.   :87.0   Max.   :141.22   Max.   :141.72  
                                                    NA's   :642     
     height           pitch          distance          duration     
 Min.   :-3.546   Min.   :2.284   Min.   :  34.08   Min.   : 14.76  
 1st Qu.:23.314   1st Qu.:3.642   1st Qu.: 883.79   1st Qu.:119.49  
 Median :30.093   Median :4.008   Median :1258.09   Median :153.95  
 Mean   :30.144   Mean   :4.009   Mean   :1526.02   Mean   :154.01  
 3rd Qu.:36.993   3rd Qu.:4.377   3rd Qu.:1936.95   3rd Qu.:188.91  
 Max.   :59.946   Max.   :5.927   Max.   :6533.05   Max.   :305.62  
                                                    NA's   :50      
> |
> sapply(final[sd_vector], sd, na.rm=T)
    distance     duration      no_pasg       height speed_ground 
 928.5600816   49.2592338    7.4931370   10.2877268   19.0594903 
   speed_air        pitch 
  10.2590370    0.5288298 
```

Observation: The final combined dataset has 850 observations and 8 variables. One variable is character and 7 variables are numeric. The summary statistics of all the variables are as follows:

Variable	Mean	Median	standard deviation	min	max	Missing values
Distance	1526.02	1258.090	928.560	34.080	6533.050	0
no_pasg	60.10	60.000	7.493	29.000	87.000	0
speed_ground	79.45	79.640	19.059	27.740	141.220	0
speed_air	103.80	101.150	10.259	90.000	141.720	642
height	30.14	30.093	10.288	-3.546	59.946	0
pitch	4.01	4.008	0.529	2.284	5.927	0
duration	154.01	153.950	49.259	14.760	305.620	50

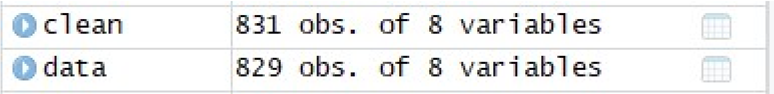
Conclusion: Final combined dataset has 850 observation and 8 variables. 2 of the variables have missing values.

Step 5. By now, if you are asked to prepare ONE presentation slide to summarize your findings, what observations will you bring to the attention of FAA agents?

- The data given had 100 duplicate observations that have been removed
- Speed_air and duration have some missing values. Are FAA agents able to provide us with the missing data?
- There are some abnormal values within the data. For example negative height, minimum duration of flight is 14 minutes, minimum distance is 34 feet. Any explanation for these values?

Step 6. Are there abnormal values in the data set? Please refer to the variable dictionary for criteria defining "normal/abnormal" values. Remove the rows that contain any "abnormal values" and report how many rows you have removed.

```
# data cleaning -------------------------------------------------------------

a <- filter(final, speed_ground >= 30, speed_ground <= 140, height >= 6,
distance <= 6000 )
#remove duration <40 but keep NAs
clean <- a[(a$duration >= 40 | is.na(a$duration)),]

summary(clean)

data <- filter (clean, distance > 140)
```



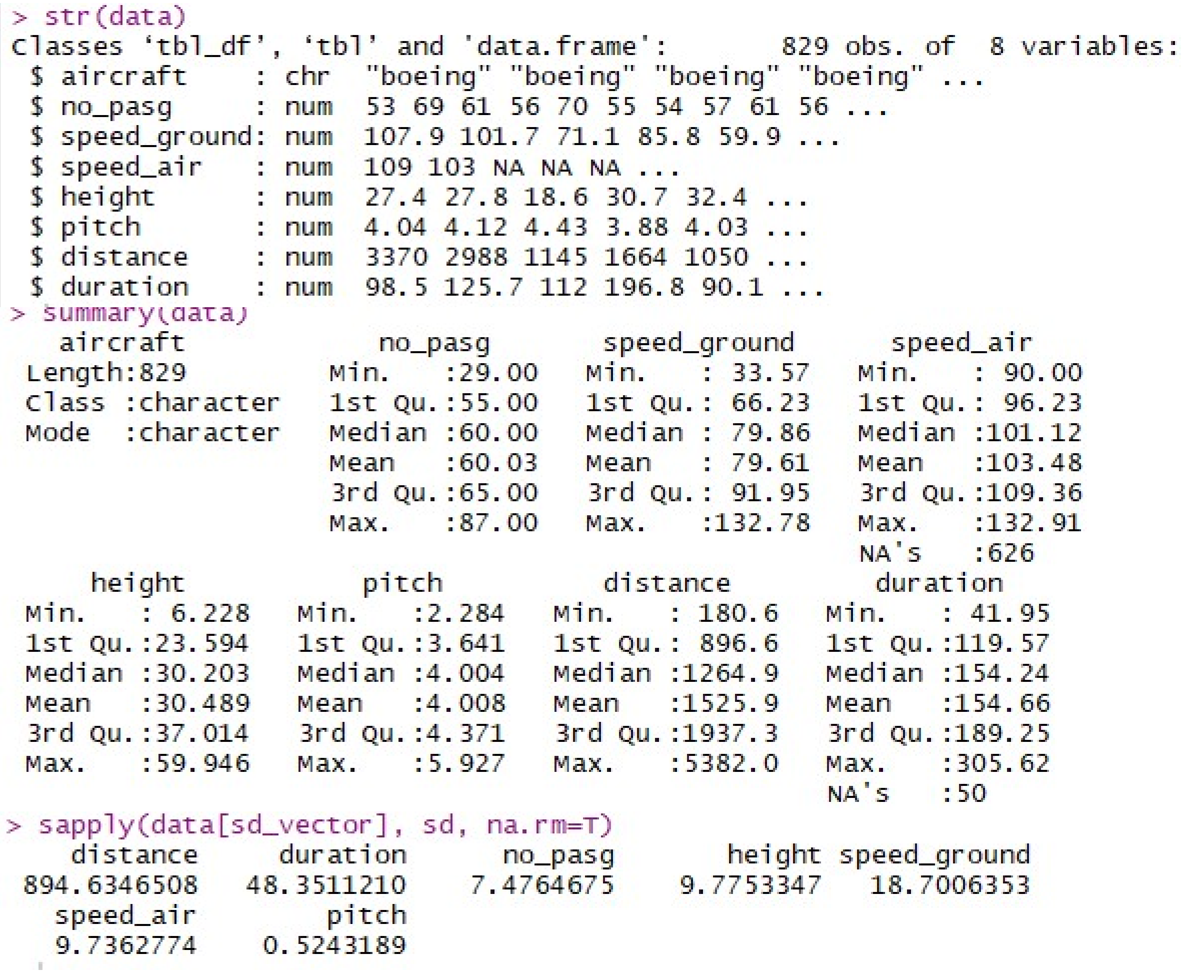
```
clean    831 obs. of 8 variables
data     829 obs. of 8 variables
```

Observation: I removed abnormal values according too the data dictionary. Upon looking at the minimum value of distance, I determined that I should remove the smallest two values from the data as the distance is too small for an aircraft to cover at landing. If the aircraft tries to stop after covering this distance, it would probably be similar to crashing. The final clean dataset has 829 observations and 8 variables.

Conclusion: I removed 21 observations while cleaning the dataset by removing abnormal values. Final dataset has 829 observations and 8 variables.

Step 7. Repeat Step 4.

```
####step 7
str(data)
summary(data)
#standard deviation
sd_vector <- c("distance","duration","no_pasg", "height", "speed_ground",
"speed_air", "pitch")
sapply(data[sd_vector], sd, na.rm=T)
```



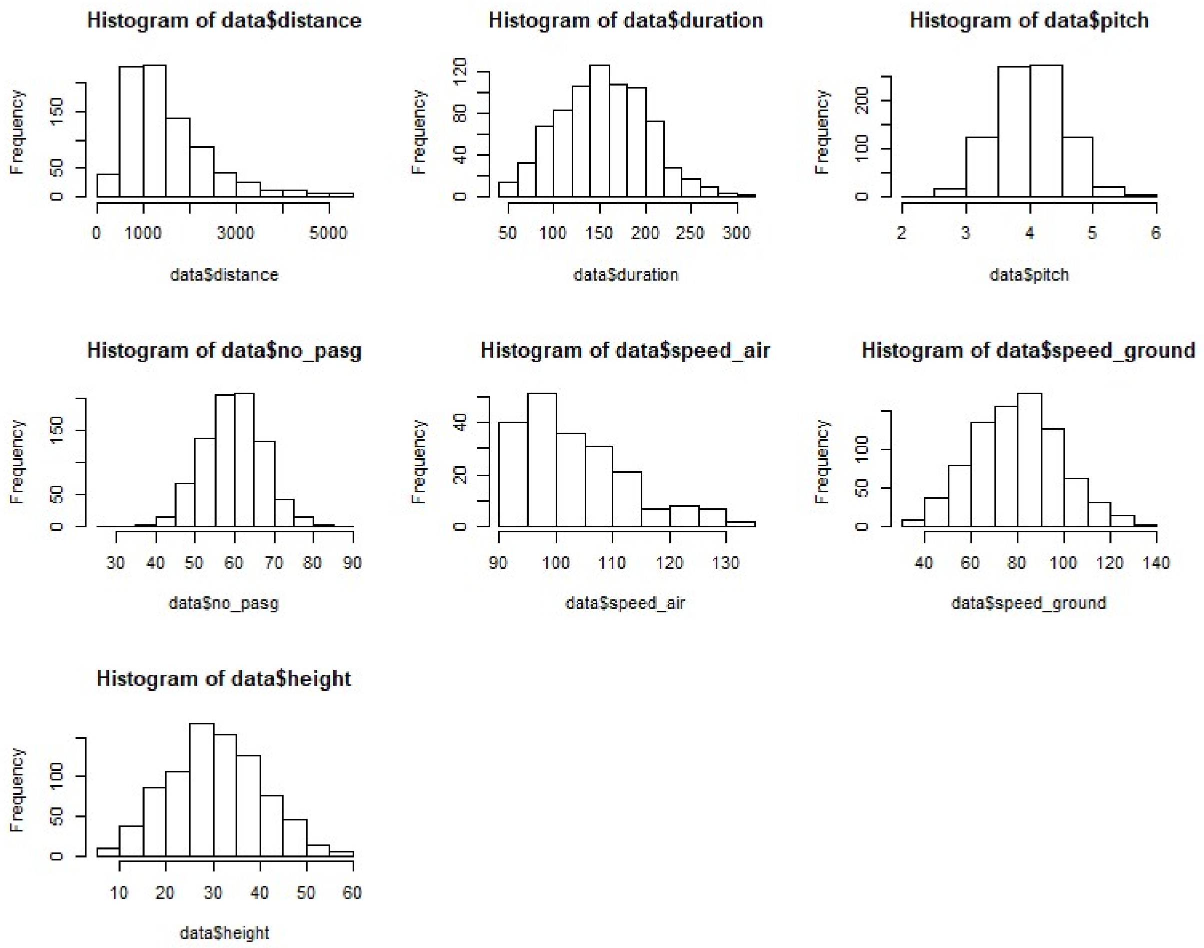
```
> str(data)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       829 obs. of  8 variables:
 $ aircraft    : chr  "boeing" "boeing" "boeing" "boeing" ...
 $ no_pasg     : num  53 69 61 56 70 55 54 57 61 56 ...
 $ speed_ground: num  107.9 101.7 71.1 85.8 59.9 ...
 $ speed_air   : num  109 103 NA NA NA ...
 $ 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 ...
 $ duration    : num  98.5 125.7 112 196.8 90.1 ...
> summary(data)
   aircraft            no_pasg       speed_ground      speed_air     
 Length:829         Min.   :29.00   Min.   : 33.57   Min.   : 90.00  
 Class :character   1st Qu.:55.00   1st Qu.: 66.23   1st Qu.: 96.23  
 Mode  :character   Median :60.00   Median : 79.86   Median :101.12  
                    Mean   :60.03   Mean   : 79.61   Mean   :103.48  
                    3rd Qu.:65.00   3rd Qu.: 91.95   3rd Qu.:109.36  
                    Max.   :87.00   Max.   :132.78   Max.   :132.91  
                                                     NA's   :626     
     height           pitch          distance          duration     
 Min.   : 6.228   Min.   :2.284   Min.   : 180.6   Min.   : 41.95  
 1st Qu.:23.594   1st Qu.:3.641   1st Qu.: 896.6   1st Qu.:119.57  
 Median :30.203   Median :4.004   Median :1264.9   Median :154.24  
 Mean   :30.489   Mean   :4.008   Mean   :1525.9   Mean   :154.66  
 3rd Qu.:37.014   3rd Qu.:4.371   3rd Qu.:1937.3   3rd Qu.:189.25  
 Max.   :59.946   Max.   :5.927   Max.   :5382.0   Max.   :305.62  
                                                   NA's   :50      
> sapply(data[sd_vector], sd, na.rm=T)
    distance     duration      no_pasg       height speed_ground 
 894.6346508   48.3511210    7.4764675    9.7753347   18.7006353 
   speed_air        pitch 
   9.7362774    0.5243189 
```

Observation: The summary statistics of the new dataset are shown in the output above. The min max values look good as per the data description. Duration variable still has 50 missing values and speed air has 626 missing values.

Conclusion: We don’t see any abnormal values in the summary statistics of the data anymore. This dataset is clean.

Step 8. Since you have a small set of variables, you may want to show histograms for all of them.

```
#### histogram
par(mfrow=c(3,3))
hist(data$distance)
hist(data$duration)
hist(data$pitch)
hist(data$no_pasg)
hist(data$speed_air)
hist(data$speed_ground)
hist(data$height)
```



<u>Observation</u>: Most of the plots look normally distributed. Distance variable is highly skewed to the right because it has outliers. I have not removed outliers for this analysis. Speed air is also skewed to the right, but we have a lot of missing values in it.
<u>Conclusion:</u> The histograms for the variables look good. They can be made better after outlier treatment, but for now they are fine.

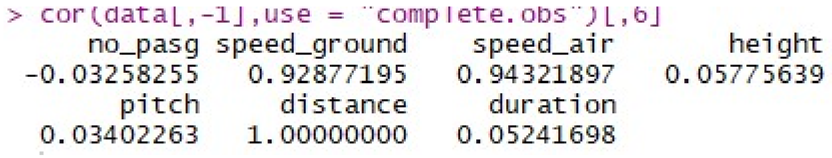
Step 9. Prepare another presentation slide to summarize your findings drawn from the cleaned data set, using no more than five "bullet statements".

- The dataset contained a few values which looked abnormal, such as negative height, very small distance etc. Those abnormal values have been removed to create a clean dataset.
- Final clean dataset has 829 observations and 8 variables. Variables duration and speed air have some missing values that have not been removed.
- The distribution of the variables in the new dataset look close to normally distributed.

- Distance and air speed have right skewed distributions because of outliers and missing values respectively.

Step 10. Compute the pairwise correlation between the landing distance and each factor X. Provide a table that ranks the factors based on the size (absolute value) of the correlation. This table contains three columns: the names of variables, the size of the correlation, the direction of the correlation (positive or negative). We call it Table 1, which will be used for comparison with our analysis later.

```
#### Correlation table
cor(data[,-1],use = "complete.obs")[,6]
```



```
> cor(data[,-1],use = "complete.obs")[,6]
      no_pasg speed_ground    speed_air       height 
  -0.03258255   0.92877195   0.94321897   0.05775639 
        pitch     distance     duration 
   0.03402263   1.00000000   0.05241698 
```

Variable	Size Correlation	Direction
Distance	1.0000	NA
speed_air	0.9432	Positive
speed_ground	0.9288	Positive
airbus	0.2355	Negative
height	0.0578	Positive
duration	0.0524	Positive
pitch	0.0340	Positive
no_pasg	0.0326	Negative

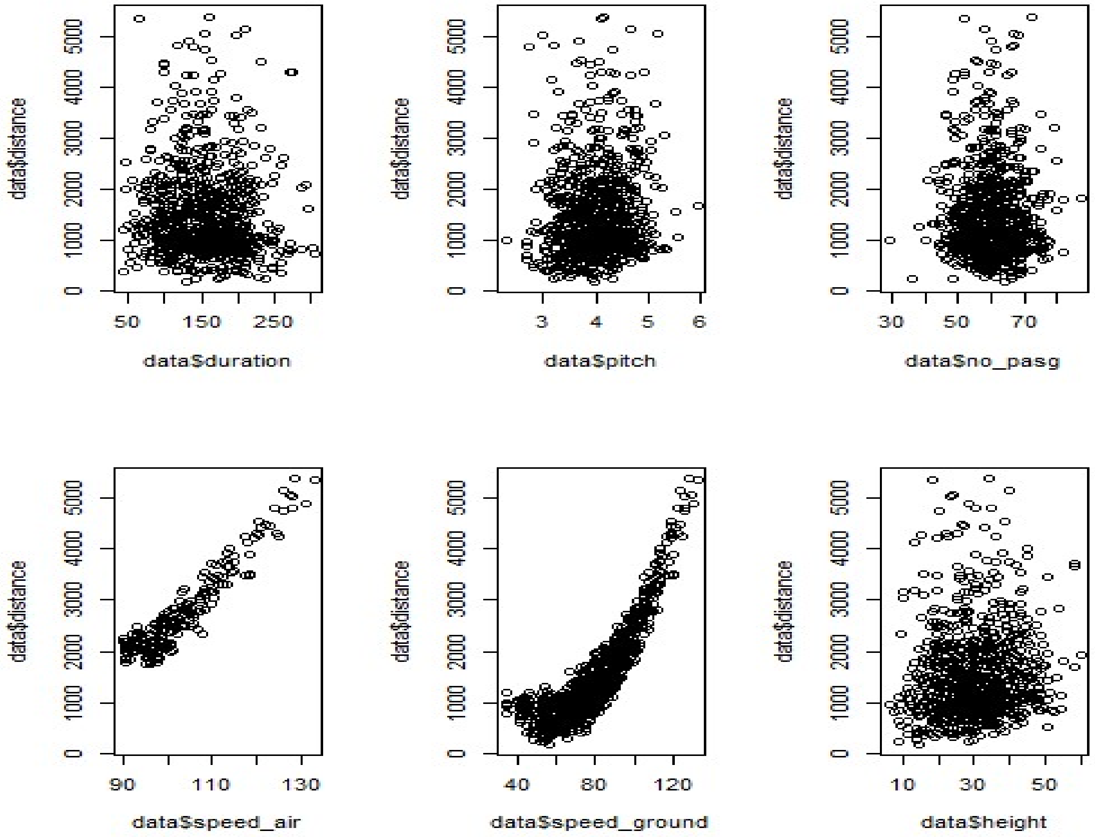
Table 1

Observation: We observe that distance is highly positively correlated with air speed and ground speed. It is also negatively correlated to number of passengers and aircraft type airbus.

Conclusion: number of passengers is slightly negatively correlated with distance. Other variables are positively correlated with distance with speed air and speed ground showing high correlation.

Step 11. Show X-Y scatter plots. Do you think the correlation strength observed in these plots is consistent with the values computed in Step 10?

```
#### scatter
par(mfrow=c(2,3))
plot(data$duration, data$distance)
plot(data$pitch, data$distance)
plot(data$no_pasg,data$distance)
plot(data$speed_air,data$distance)
plot(data$speed_ground,data$distance)
plot(data$height,data$distance)
```

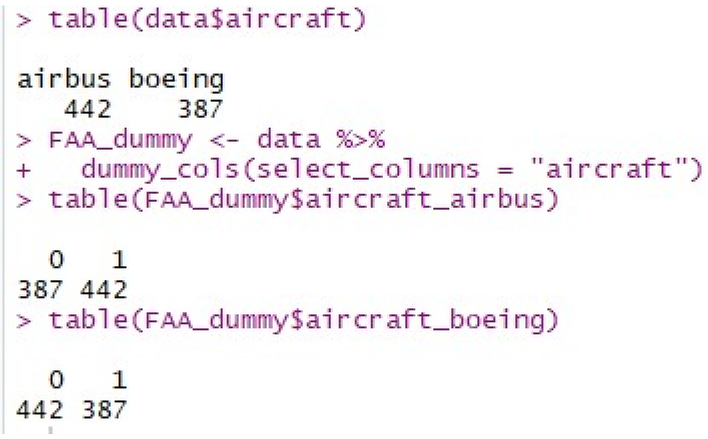


Observation: The scatterplots show correlation of distance with all the other variables. We do not see a clear correlation pattern between distance and duration, distance and number of passengers, distance and height and distance and pitch. However, we do observe high positive correlation between distance and speed air and distance and speed ground. Speed ground shows somewhat quadratic relationship with distance. It also appears that speed_air data is truncated.

Conclusion: These scatter plots show results which are consistent with our results in table 1 in step 10.

Step 12. Have you included the airplane make as a possible factor in Steps 10-11? You can code this character variable as 0/1.

```
### dummies
library(fastDummies)
table(data$aircraft)
FAA_dummy <- data %>%
  dummy_cols(select_columns = "aircraft")
table(FAA_dummy$aircraft_airbus)
table(FAA_dummy$aircraft_boeing)
```



```
> table(data$aircraft)

airbus boeing
   442    387
> FAA_dummy <- data %>%
+   dummy_cols(select_columns = "aircraft")
> table(FAA_dummy$aircraft_airbus)

  0   1
387 442
> table(FAA_dummy$aircraft_boeing)

  0   1
442 387
```

Observation: We created dummy variables for aircraft type. Aircraft_airbus shows Boeing as 0 and airbus as 1. We will use this variable for our analysis further. There are 387 boeing planes and 442 airbus planes.

Conclusion: A dummy variable aircraft_airbus will be used to understand the impact of airplane type on landing distance.

Step 13. Regress Y (landing distance) on each of the X variables. Provide a table that ranks the factors based on its significance. The smaller the p-value, the more significant the factor. This table contains three columns: the names of variables, the size of the p-value, the direction of the regression coefficient (positive or negative). We call it Table 2.

```
#### regression
model1 <- lm( distance ~ duration, data = FAA_dummy)
summary(model1)
model2 <- lm( distance ~ no_pasg, data=FAA_dummy)
summary(model2)
model3 <- lm( distance ~ speed_air, data=FAA_dummy)
summary(model3)
model4 <- lm( distance ~ speed_ground, data=FAA_dummy)
summary(model4)
model5 <- lm( distance ~ pitch, data=FAA_dummy)
summary(model5)
model6 <- lm( distance ~ height, data=FAA_dummy)
summary(model6)
model7 <- lm( distance ~ aircraft_airbus, data=FAA_dummy)
summary(model7)
```

Variable	P-Value	direction of regression coefficient
speed_air	<2e-16	Positive
speed_ground	<2e-16	Positive
airbus	6.53E-12	Negative
height	0.0063	Positive
pitch	0.022	Positive

duration	0.184	Negative
no_pasg	0.717	Negative

Table 2

Observation: The above table shows that speed ground and speed air are the most significant variables when regressed by distance variable. Airbus, duration and number of passengers show negative coefficients, however duration and number of passengers are not significant at relevant significance levels.

Conclusion: The variables that impact landing distance the most are speed ground, speed air, aircraft type, height and pitch in that order.

Step 14. Standardize each X variable. In other words, create a new variable X'= {X-mean(X)}/sd(X).
The mean of X' is 0 and its standard deviation is 1.
Regress Y (landing distance) on each of the X' variables. Provide a table that ranks the factors based on the size of the regression coefficient. The larger the size, the more important the factor. This table contains three columns: the names of variables, the size of the regression coefficient, the direction of the regression coefficient (positive or negative). We call it Table 3.

```
 #### creating standardized variables

no_pasg1 <- {(FAA_dummy$no_pasg-
mean(FAA_dummy$no_pasg))/sd(FAA_dummy$no_pasg)}
speed_ground1 <- {(FAA_dummy$speed_ground-
mean(FAA_dummy$speed_ground))/sd(FAA_dummy$speed_ground)}
speed_air1 <- {(FAA_dummy$speed_air-mean(FAA_dummy$speed_air,
na.rm=T))/sd(FAA_dummy$speed_air, na.rm=T)}
height1 <- {(FAA_dummy$height-mean(FAA_dummy$height))/sd(FAA_dummy$height)}
pitch1 <- {(FAA_dummy$pitch-mean(FAA_dummy$pitch))/sd(FAA_dummy$pitch)}
duration1 <- {(FAA_dummy$duration-mean(FAA_dummy$duration,
na.rm=T))/sd(FAA_dummy$duration, na.rm=T)}

together<- cbind(FAA_dummy,
no_pasg1,speed_ground1,speed_air1,height1,pitch1,duration1)
standard <- select(together,
aircraft,no_pasg1,speed_ground1,speed_air1,height1,pitch1,distance,
duration1, aircraft_boeing, aircraft_airbus)

summary(standard)

models1 <- lm( distance ~ duration1, data = standard)
summary(models1)
models2 <- lm( distance ~ no_pasg1, data=standard)
summary(models2)
models3 <- lm( distance ~ speed_air1, data=standard)
summary(models3)
models4 <- lm( distance ~ speed_ground1, data=standard)
summary(models4)
models5 <- lm( distance ~ pitch1, data=standard)
summary(models5)
```

```
models6 <- lm( distance ~ height1, data=standard)
summary(models6)
models7 <- lm( distance ~ aircraft_airbus, data=standard)
summary(models7)
```

Variable	size of regression coefficient	direction of regression coefficient
speed_air1	774.35	Positive
speed_ground1	774.23	Positive
airbus	422.07	Negative
height1	84.82	Positive
pitch1	71.16	Positive
duration1	43.01	Negative
no_pasg1	11.26	Negative

Table 3

Observation: The order of importance after standardizing is same as table 2, with speed air, speed ground, airbus, height and pitch being the most important factors in that order. We did not standardize the dummy variables as we don't need to.

Conclusion: The variables that impact landing distance the most are speed ground, speed air, aircraft type, height and pitch in that order.

Step 15. Compare Tables 1,2,3. Are the results consistent? At this point, you will meet with a FAA agent again. Please provide a single table than ranks all the factors based on their relative importance in determining the landing distance. We call it Table 0.

cor(FAA_dummy$distance, FAA_dummy$aircraft_airbus)

I used the above code to find correlation between distance and aircraft type and added the result in table 2.

Variable	Rank
speed_ground	1
speed_air	2
airbus	3
height	4
pitch	5
duration	6
no_pasg	7

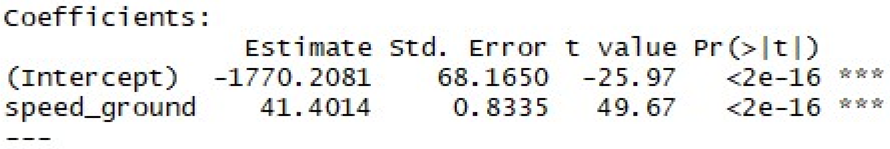
Table 0

Conclusion: Upon comparing results from table 1, 2, 3, we conclude that the most important factors that impact landing distance are speed ground, speed air and aircraft type.

Step 16. Compare the regression coefficients of the three models below:
Model 1: LD ~ Speed_ground
Model 2: LD ~ Speed_air
Model 3: LD ~ Speed_ground + Speed_air 5

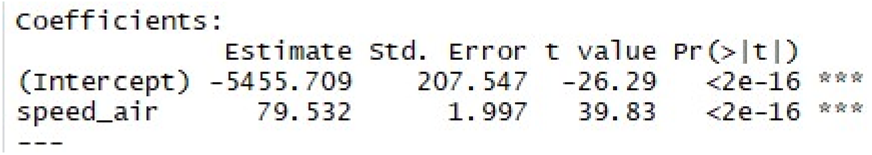
Do you observe any significance change and sign change? Check the correlation between Speed_ground and Speed_air. You may want to keep one of them in the model selection. Which one would you pick? Why?

```
####step 16

model_1 <- lm(distance~speed_ground, data=FAA_dummy)
summary(model_1)
```



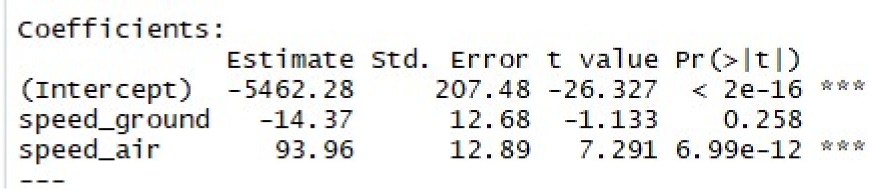
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1770.2081    68.1650  -25.97   <2e-16 ***
speed_ground    41.4014     0.8335   49.67   <2e-16 ***
---
```

```
model_2 <- lm(distance~speed_air, data=FAA_dummy)
summary(model_2)
```



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5455.709    207.547  -26.29   <2e-16 ***
speed_air      79.532      1.997   39.83   <2e-16 ***
---
```

```
model_3 <- lm(distance~speed_ground + speed_air, data=FAA_dummy)
summary(model_3)
```



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -5462.28     207.48 -26.327  < 2e-16 ***
speed_ground   -14.37      12.68  -1.133    0.258    
speed_air       93.96      12.89   7.291 6.99e-12 ***
---
```

<u>Observation</u>: When we run a linear model with speed air and speed ground, speed ground loses significance and its coefficient sign changes from positive to negative. We now check the correlation between speed air and speed ground.

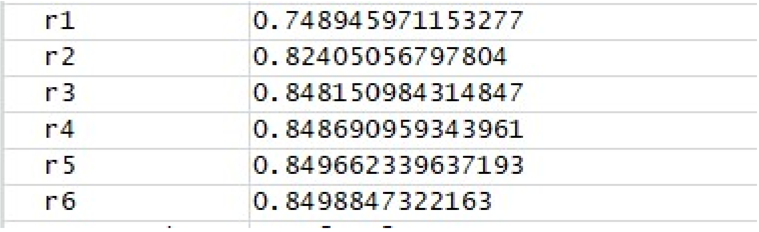
```
cor(FAA_dummy$speed_ground,FAA_dummy$speed_air,use = "complete.obs")
[1] 0.9879383
```

<u>Observation</u>: speed ground and speed air are highly correlated with each other with a correlation of 98.8%. This brings a problem of multicollinearity. Hence we cannot use both the variables in our linear model. I choose speed ground over speed air in my model as speed air has many null values and hence will not be able to predict as accurately as speed ground. We don’t want to lose all the important information.

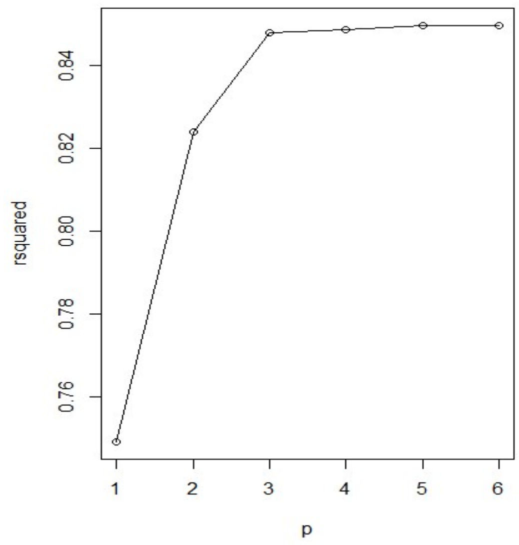
Conclusion: Speed air and speed ground are highly correlated and bring a problem of multicollinearity. Hence we will only use speed ground in our analysis.

Step 17. Suppose in Table 0, the variable ranking is as follows: X1, X2, X3….. Please fit the following six models:

```
#### step 17

mod1 <- lm(distance ~ speed_ground, data=FAA_dummy)
r1 <- summary(mod1)$r.squared
mod2 <- lm(distance ~ speed_ground + aircraft_airbus, data=FAA_dummy )
r2 <- summary(mod2)$r.squared
mod3 <- lm(distance ~ speed_ground + aircraft_airbus + height, data=FAA_dummy
)
r3 <- summary(mod3)$r.squared
mod4 <- lm(distance ~ speed_ground + aircraft_airbus + height + pitch,
data=FAA_dummy )
r4 <- summary(mod4)$r.squared
mod5 <- lm(distance ~ speed_ground + aircraft_airbus + height + pitch +
duration, data=FAA_dummy )
r5 <- summary(mod5)$r.squared
mod6 <- lm(distance ~ speed_ground + aircraft_airbus + height + pitch +
duration + no_pasg, data=FAA_dummy )
r6 <- summary(mod6)$r.squared

rsquared <- c(r1,r2,r3,r4,r5,r6)
p <- c(1,2,3,4,5,6)
par(mfrow=c(1,1))
plot(p, rsquared)
lines.default(x=p, y=rsquared)
```



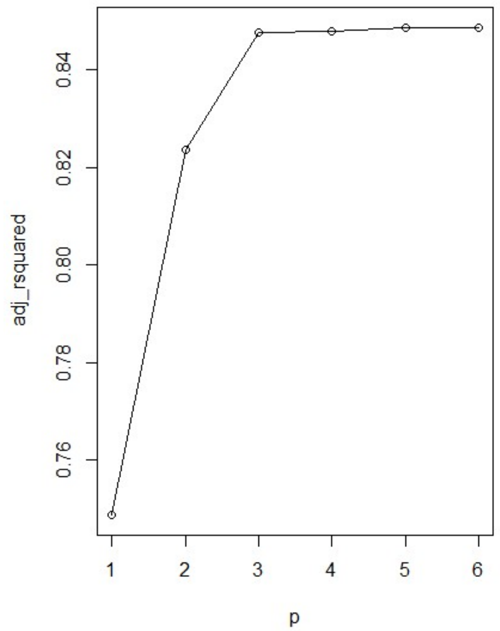
r1	0.748945971153277
r2	0.82405056797804
r3	0.848150984314847
r4	0.848690959343961
r5	0.849662339637193
r6	0.8498847322163

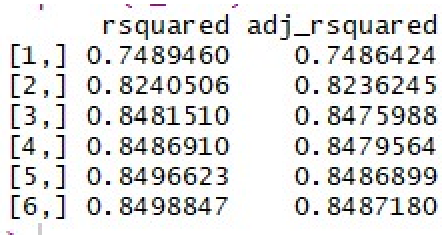


Observation: We calculated r squared of all the models as we kept adding important variables in descending order in the model. We observe that as the number of variables increase, r squared values increase too. We can also see this through the plot where we have rsquared vs number of variables.
Conclusion: As number of variables in the model increase, r squared value increases.

Step 18. Repeat Step 17 but use adjusted R-squared values instead.

```
##### adjusted r squared
adj1 <- summary(mod1)$adj.r.squared
adj2 <- summary(mod2)$adj.r.squared
adj3 <- summary(mod3)$adj.r.squared
adj4 <- summary(mod4)$adj.r.squared
adj5 <- summary(mod5)$adj.r.squared
adj6 <- summary(mod6)$adj.r.squared

adj_rsquared <- c(adj1,adj2,adj3,adj4,adj5,adj6)
plot(p,adj_rsquared)
lines.default(x=p, y=adj_rsquared)

r_diff <- cbind(rsquared,adj_rsquared)
print(r_diff)
```





```
      rsquared adj_rsquared
[1,] 0.7489460    0.7486424
[2,] 0.8240506    0.8236245
[3,] 0.8481510    0.8475988
[4,] 0.8486910    0.8479564
[5,] 0.8496623    0.8486899
[6,] 0.8498847    0.8487180
```

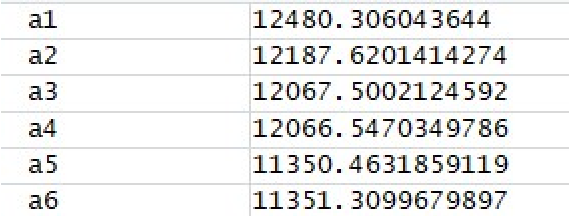
Observation: When we find the adjusted r squared values for each model and plot those versus the number of variables in each model, we see a very similar plot as in step 17. When we compare r squared values to the adjusted r squared values, we see that adjusted r squared values are slightly smaller than the r squared values for each model. This is because adjusted r-squared values account for model complexity.

Conclusion: adjusted r squared is a better method for model selection as it accounts for model complexity. However, in this scenario, they are both very similar.

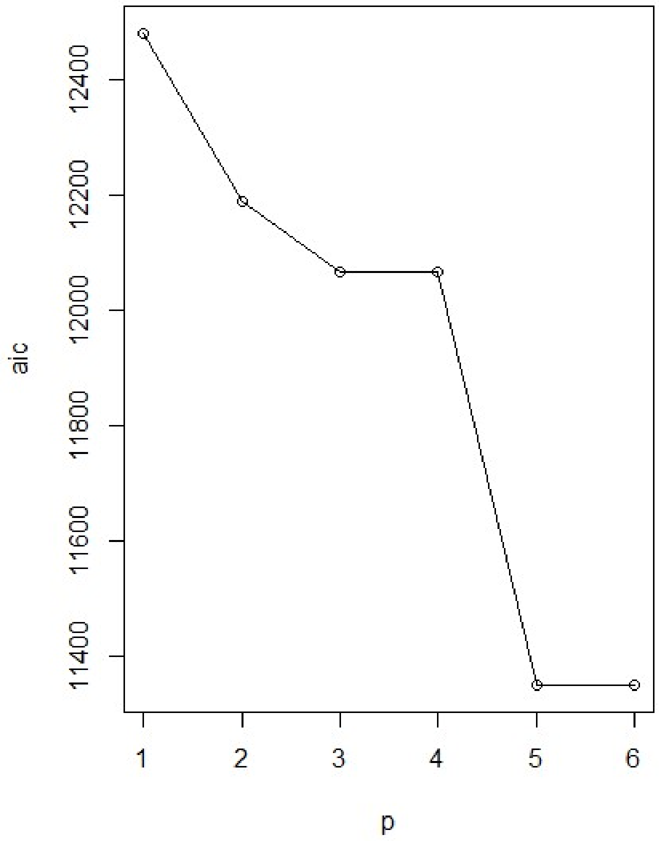
Step 19. Repeat Step 17 but use AIC values instead

```
#### AIC
a1 <- AIC(mod1)
a2 <- AIC(mod2)
a3 <- AIC(mod3)
a4 <- AIC(mod4)
a5 <- AIC(mod5)
a6 <- AIC(mod6)

aic <- c(a1,a2,a3,a4,a5,a6)
plot(p, aic)
```

```
lines.default(x=p, y=aic)
```



a1	12480.306043644
a2	12187.6201414274
a3	12067.5002124592
a4	12066.5470349786
a5	11350.4631859119
a6	11351.3099679897

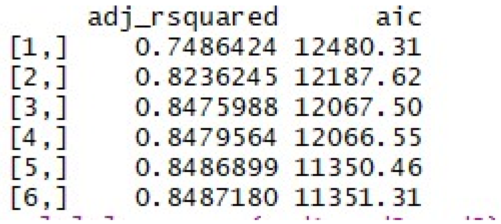


Observation: When we found the AIC values of all the models and plot them against the number of variables in each model, we see that as number of variables increase, the AIC value drops. The lower the AIC value the better.

Conclusion: As number of variables in the model increase, AIC value drops, thus giving a stronger, more accurate model.

Step 20. Compare the results in Steps 18-19, what variables would you select to build a predictive model for LD?

```
allmodels <- cbind(adj_rsquared,aic)
print(allmodels)
```



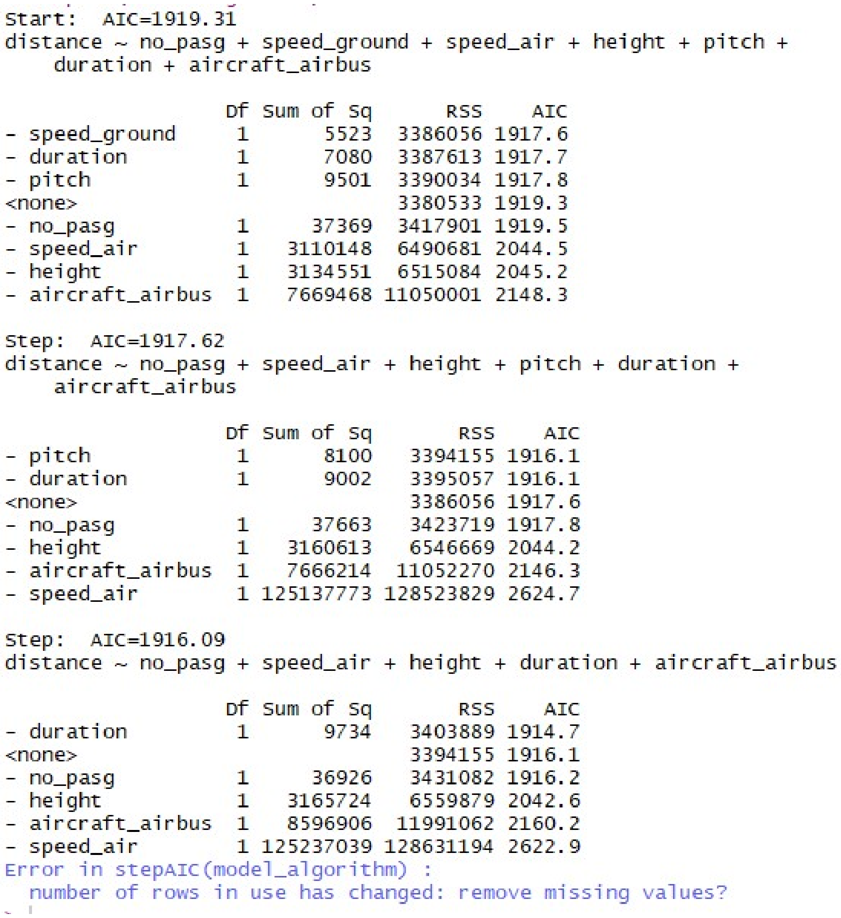
```
     adj_rsquared      aic
[1,]    0.7486424 12480.31
[2,]    0.8236245 12187.62
[3,]    0.8475988 12067.50
[4,]    0.8479564 12066.55
[5,]    0.8486899 11350.46
[6,]    0.8487180 11351.31
```

<u>Observation</u> : Based on the adjusted r squared and AIC results, we see that model 5 has the lowest AIC value as compared to other models. Model 5 and 6 have very similar adjusted r squared values (in the 10 thousandth place). Hence we can choose better AIC value for model selection purpose.

<u>Conclusion</u>: Based on my analysis, I would select model 5 with variables speed ground, aircraft type (airbus), height, pitch and duration for predicting landing distance.

Step 21. Use the R function "StepAIC" to perform forward variable selection. Compare the result with that in Step 19.

```
####21 step AIC
model_algorithm <- lm (distance ~.,data = FAA_dummy[,-c(1,9)])
stepAIC(model_algorithm)
```



```
Start:  AIC=1919.31
distance ~ no_pasg + speed_ground + speed_air + height + pitch + 
    duration + aircraft_airbus

                  Df Sum of Sq      RSS    AIC
- speed_ground     1      5523  3386056 1917.6
- duration         1      7080  3387613 1917.7
- pitch            1      9501  3390034 1917.8
<none>                          3380533 1919.3
- no_pasg          1     37369  3417901 1919.5
- speed_air        1   3110148  6490681 2044.5
- height           1   3134551  6515084 2045.2
- aircraft_airbus  1   7669468 11050001 2148.3

Step:  AIC=1917.62
distance ~ no_pasg + speed_air + height + pitch + duration + 
    aircraft_airbus

                  Df Sum of Sq       RSS    AIC
- pitch            1      8100   3394155 1916.1
- duration         1      9002   3395057 1916.1
<none>                           3386056 1917.6
- no_pasg          1     37663   3423719 1917.8
- height           1   3160613   6546669 2044.2
- aircraft_airbus  1   7666214  11052270 2146.3
- speed_air        1 125137773 128523829 2624.7

Step:  AIC=1916.09
distance ~ no_pasg + speed_air + height + duration + aircraft_airbus

                  Df Sum of Sq       RSS    AIC
- duration         1      9734   3403889 1914.7
<none>                           3394155 1916.1
- no_pasg          1     36926   3431082 1916.2
- height           1   3165724   6559879 2042.6
- aircraft_airbus  1   8596906  11991062 2160.2
- speed_air        1 125237039 128631194 2622.9
Error in stepAIC(model_algorithm) : 
  number of rows in use has changed: remove missing values?
```

Statistical Modeling Part1
Ashwita Saxena
M06119969

Observation: Based on step AIC algorithm, the best model should include variables number of passengers, speed air, height, duration and aircraft type airbus. The following table shows the comparison between my model and stepAIC model.

Variable	my model	stepAIC model
speed_ground	x	
speed_air		x
airbus	x	x
height	x	x
pitch	x	
duration	x	x
no_pasg		x
AIC	**11350.46**	**1916.09**

The final model that step AIC gives us has much better AIC than our own model, however, it is using a very small number of observations as it chose to select speed_air over speed_ground. Speed air has a lot of missing values and hence brings down the number of observations to a very small number which would lead to the model inaccuracy.

Conclusion: Even though StepAIC algorithm gives us a better AIC value, human mind is important to make certain decisions while performing an analysis. The algorithm purely loos at data with statistical eyes and selects the model that gives it a better AIC score by disregarding the fact that less number of observations would lead to loss of information. Hence I think it is very important for us to make sure the final model contains appropriate number of observations and hence more relevant variables.