Stat Modelling Assignment 1

Digvijay Kawale

1/15/2020

# Loading the required packages

library(tidyverse)

## ── Attaching packages ────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.2.1 ✔ purrr 0.3.2  
## ✔ tibble 2.1.3 ✔ dplyr 0.8.3  
## ✔ tidyr 0.8.3 ✔ stringr 1.4.0  
## ✔ readr 1.3.1 ✔ forcats 0.4.0

## ── Conflicts ───────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(readxl)  
library(dplyr)  
library(corrr)  
library(MASS)

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

# 

# Introduction

* **Backgroud**: 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)

# Initial exploration of the data

## Step 1: Importing Data sets into R Environment

* The R code for importing excel files into R environment was found through google searh. Here is its [link](https://readxl.tidyverse.org)

Flights\_800 <- read\_xls("~/Desktop/Subjects/Flex 3/Statistical Modelling/WeeK 1/FAA1-1.xls")  
Flights\_150 <- read\_xls("~/Desktop/Subjects/Flex 3/Statistical Modelling/WeeK 1/FAA2-1.xls")

## Step 2: Study of imported data sets

* The structure of the data sets is checked using the function “Str” (Structure)
* **Data Set Flights\_800:** The class of the data set is data frame with 800 observations of 8 variables. The variable aircraft is of the class ‘character variable’ where as the remmaining 7 variables are of the class ‘numeric variables’.

str(Flights\_800)

## 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 ...

* **Data Set Flights\_150:** The class of the data set is data frame with 150 observations of 7 variables. The variable aircraft is of the class ‘character variable’ where as the remmaining 6 variables are of the class ‘numeric variables’.
* **Difference in Data sets:** We observe that there is a difference between two data sets ‘Flights\_800’ and ‘Flights\_150’. The data set ‘Flights\_800’ has a variable ‘duration’ which is not present in ‘Flights\_150’.

str(Flights\_150)

## 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 ...

## Step 3: Merging the two data sets

* The data sets are merged using the function ‘rbind’ (row bind). This function is usually used when we want two append two data sets of similar kind.
* Before merging the data sets, a column ‘duration’ has been created in the data set ‘Flights\_150’ to make both the data sets similar.
* After merging the data sets, the combined data set was checked for duplicates without using the column ‘duration’ as it was not present in the data set ‘Flights\_150’. The combined data set had 100 duplicates. Those duplicates are removed from the combined data set. The usage of ‘Duplicated’ function was found using the google search. This is the [link](https://stackoverflow.com/questions/13742446/duplicates-in-multiple-columns) where it shows the usage of the function.

Flights\_150$duration <- NA  
  
flights\_final <- rbind(Flights\_800, Flights\_150)  
  
flights\_columns <- flights\_final[c("aircraft" , "no\_pasg" , "speed\_ground" ,"speed\_air" , "height" , "pitch" , "distance" )]  
  
flights\_final <- flights\_final[!duplicated(flights\_columns),]

## Step 4: Structure and summary of combined data set

* The sample size of combined data set is 850 with 8 variables. The variable ‘aircraft’ belongs to the class of ‘character varibles’ and the remaining 7 variables belong the the class of ‘Numeric variables’. Below is the structure of data set using the function ‘str’ in R:

str(flights\_final)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 850 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 ...

* Below are the summary statictics for all the variables in the combined data set using the function ‘summary’ in R.

summary(flights\_final)

## aircraft duration no\_pasg speed\_ground   
## Length:850 Min. : 14.76 Min. :29.0 Min. : 27.74   
## Class :character 1st Qu.:119.49 1st Qu.:55.0 1st Qu.: 65.90   
## Mode :character Median :153.95 Median :60.0 Median : 79.64   
## Mean :154.01 Mean :60.1 Mean : 79.45   
## 3rd Qu.:188.91 3rd Qu.:65.0 3rd Qu.: 92.06   
## Max. :305.62 Max. :87.0 Max. :141.22   
## NA's :50   
## speed\_air height pitch distance   
## Min. : 90.00 Min. :-3.546 Min. :2.284 Min. : 34.08   
## 1st Qu.: 96.25 1st Qu.:23.314 1st Qu.:3.642 1st Qu.: 883.79   
## Median :101.15 Median :30.093 Median :4.008 Median :1258.09   
## Mean :103.80 Mean :30.144 Mean :4.009 Mean :1526.02   
## 3rd Qu.:109.40 3rd Qu.:36.993 3rd Qu.:4.377 3rd Qu.:1936.95   
## Max. :141.72 Max. :59.946 Max. :5.927 Max. :6533.05   
## NA's :642

## Step 5: Summary findings for FAA agents:

* The landing aircraft is required to be at least 6 meters high at the threshold of the runway. It is observed that min Height is negative. This must a data recording error as it is not possibe to have negative Height.
* If the value of Ground Speed is less than 30MPH or greater than 140MPH, then the landing is considered as abnormal. It is observed that the minimum and maximum values of Ground Speed are 27.74 and 141.22 respectively. It shows that there are abnormal landings in our data set.
* If the value of Air Speed is less than 30MPH or greater than 140MPH, then the landing is considered as abnormal. It is observed that th maximum value of Air Speed is 141.72 which is abnormal. Also, there are 642 missing values in the data.
* The distribution of Distance is not normal as there a huge difference between its Median and Mean.
* We observe that the minimum value of duration is 14.76 mins which should not be the case as the duration of a normal flight should always be greater than 40 mins.

# Data Cleaning and further exploration

## Step 6: Removing abnormal observation from the data

* As per the abnormalities in the data, we have removed 18 observations from the data set. We now have 832 final observations with 8 variables in our data set. The abnormal values are removed using the ‘filter’ funtion in the ‘dplyr’ library. NA values are also taken care of using the function ‘is.na’ and they are retained in the data set.

flights\_final <- filter(flights\_final, ifelse(is.na(height), TRUE, height >= 6))  
  
flights\_final <- filter(flights\_final, ifelse(is.na(speed\_ground), TRUE, (speed\_ground >= 30 & speed\_ground <= 140)))  
  
flights\_final <- filter(flights\_final, ifelse(is.na(speed\_air), TRUE, (speed\_air >= 30 & speed\_air <= 140)))  
  
flights\_final <- filter(flights\_final, ifelse(is.na(duration), TRUE, duration >= 40 ))  
  
dim(flights\_final)

## [1] 832 8

## Step 7: Observing the structure summary of final data set.

* There are 832 observations wih 8 variables. The variable ‘aircraft’ is of the class character variables while the remaining 7 variables are of the class numeric variables.

str(flights\_final)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 832 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 ...

* After filtering the data, we observe that there are no abnormal observations in our final data set. But there are 50 NA’s in the column ‘duration’ and 628 NA’s in the column ‘speed\_air’

summary(flights\_final)

## aircraft duration no\_pasg speed\_ground   
## Length:832 Min. : 41.95 Min. :29.00 Min. : 33.57   
## Class :character 1st Qu.:119.65 1st Qu.:55.00 1st Qu.: 66.20   
## Mode :character Median :154.26 Median :60.00 Median : 79.83   
## Mean :154.73 Mean :60.06 Mean : 79.61   
## 3rd Qu.:189.64 3rd Qu.:65.00 3rd Qu.: 91.99   
## Max. :305.62 Max. :87.00 Max. :136.66   
## NA's :50   
## speed\_air height pitch distance   
## Min. : 90.00 Min. : 6.228 Min. :2.284 Min. : 41.72   
## 1st Qu.: 96.25 1st Qu.:23.530 1st Qu.:3.640 1st Qu.: 893.43   
## Median :101.15 Median :30.185 Median :4.002 Median :1263.54   
## Mean :103.65 Mean :30.474 Mean :4.005 Mean :1528.24   
## 3rd Qu.:109.40 3rd Qu.:37.018 3rd Qu.:4.370 3rd Qu.:1937.58   
## Max. :136.42 Max. :59.946 Max. :5.927 Max. :6309.95   
## NA's :628

* We checked for the number of observations by aircraft type using the summarize and group\_by functions. We found that 444 of the observations are for the aircraft type ‘airbus’ and the remaining 388 observations are of the type ‘boeing’

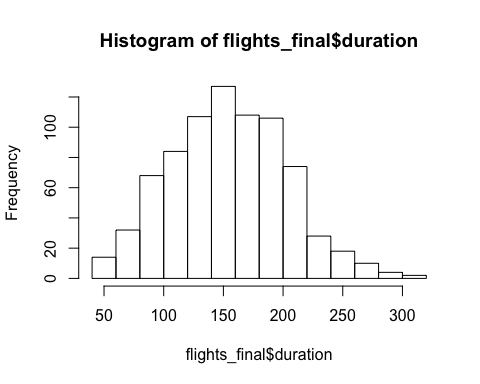
flights\_final %>% group\_by(aircraft) %>% summarize(n())

## # A tibble: 2 x 2  
## aircraft `n()`  
## <chr> <int>  
## 1 airbus 444  
## 2 boeing 388

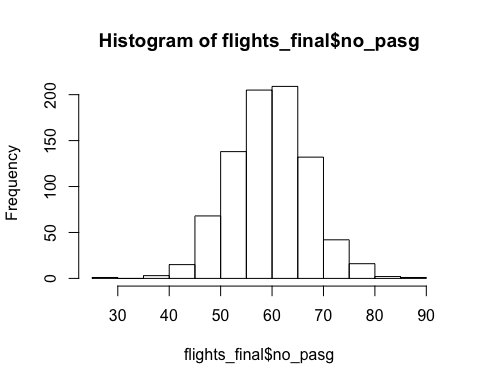
## Step 8:Histograms of all the numeric variables.

* We observe that the variables duration, no\_pasg, height and pitch follow distribution closer to the normal distribution. Rest of the varibles i.e. speed\_ground, speed\_air and distance have curves skewed towards right.

hist(flights\_final$duration)



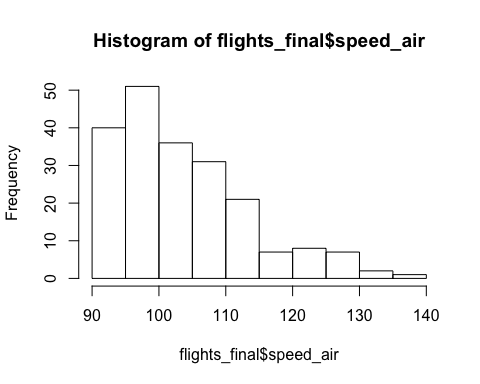
hist(flights\_final$no\_pasg)



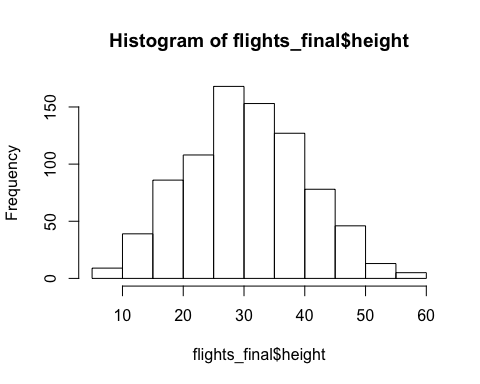
hist(flights\_final$speed\_ground) A screenshot of a cell phone

Description automatically generated

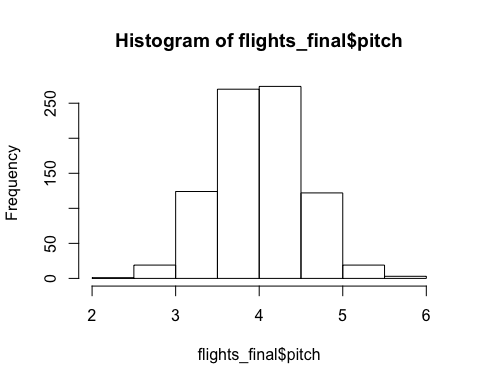
hist(flights\_final$speed\_air)



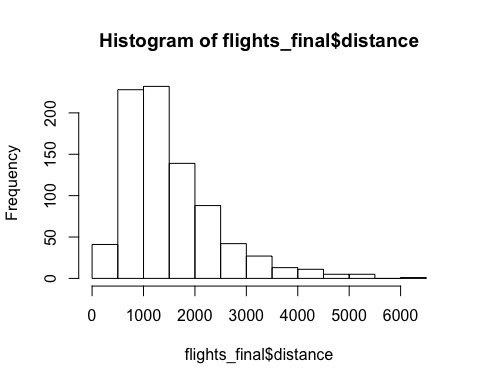
hist(flights\_final$height)



hist(flights\_final$pitch)



hist(flights\_final$distance)



## Step 9: Summary of the cleaned data:

* We observe that there are missing values in the variables ‘speed\_air’ and ‘duration’ in the cleaned data set.
* In our cleaned data set, there are no abnormal observations in the variables duration, speed\_ground, speed\_air, and height.
* We observe that the variables speed\_ground, duration, no\_pasg, height and pitch follow normal distribution.
* The variables speed\_air and distance do not follow normal distribution and have their curves skewed towards right. It means that the minimum values in these variables are closer to mean and median while the maximum values are far from the mean and median. As we also observe from the summary that for all the variable except distance, the means and medians are close to each other.
* We found that 444 of the total observations are of the type ‘airbus’ and the remaining 388 observations are of the type ‘boeing’

# Initial analysis for identifying important factors that impact the response variable “landing distance”

## Step 10: Finding the correlation of “Landing Distance” with all other variables

* The Table\_1 shows the corrlation of “Landing Distance” with all the other variables present in the data set. It was created using the library “corrr”. This was found by using the google and this is its [link](https://www.r-bloggers.com/focus-on-correlations-of-some-variables-with-many-others/)
* Before creating Table\_1, a variable “aircraft\_num” was introduced to accomodate the character variable “aircraft”
* The table is sorted in the descending order of absolute values of correlation. We see that the factor that affect most to the “Landing Distance” are “speed\_air” and “speed\_ground”.

flights\_final$aircraft\_num <- ifelse(flights\_final$aircraft == "airbus", 1, 0)  
  
Table\_1 <- flights\_final[,2:9] %>% correlate() %>% focus(distance)

##   
## Correlation method: 'pearson'  
## Missing treated using: 'pairwise.complete.obs'

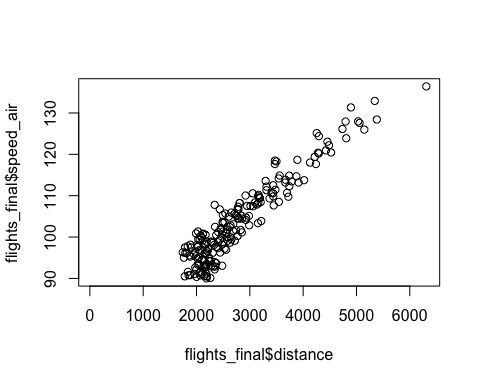
Table\_1$direction\_of\_correlation <- ifelse(Table\_1$distance < 0, "Negative", "Positive")  
  
Table\_1 <- rename(Table\_1, size\_of\_correlation = distance)  
  
Table\_1 <- rename(Table\_1, Variable\_Names = rowname)  
  
Table\_1 <- arrange(Table\_1, desc(abs(Table\_1$size\_of\_correlation)))  
  
Table\_1

## # A tibble: 7 x 3  
## Variable\_Names size\_of\_correlation direction\_of\_correlation  
## <chr> <dbl> <chr>   
## 1 speed\_air 0.944 Positive   
## 2 speed\_ground 0.866 Positive   
## 3 aircraft\_num -0.241 Negative   
## 4 height 0.107 Positive   
## 5 pitch 0.0875 Positive   
## 6 duration -0.0553 Negative   
## 7 no\_pasg -0.0141 Negative

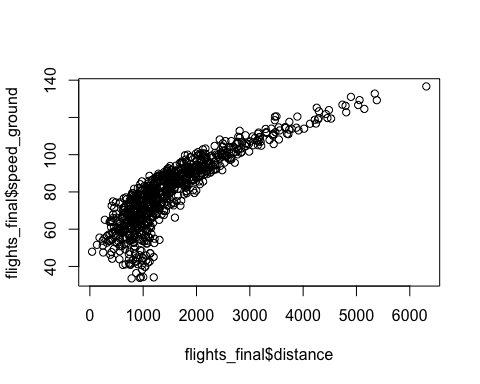
## Step 11: Confirming results using Scatter Plots

* It is evident from the scatter plots that there is a stong positive correlation between the “Landing distance” with the variables “Speed\_air” and “speed\_ground”. It has very less correlation with rest of the variables.

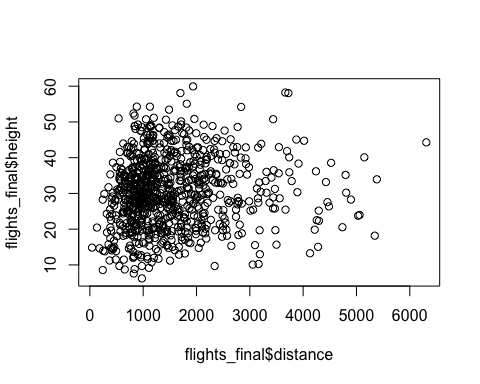
plot(flights\_final$distance, flights\_final$speed\_air)



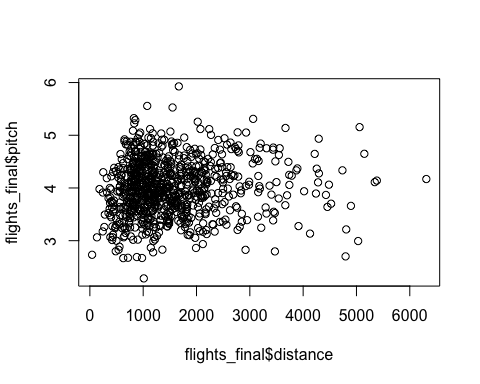
plot(flights\_final$distance, flights\_final$speed\_ground)



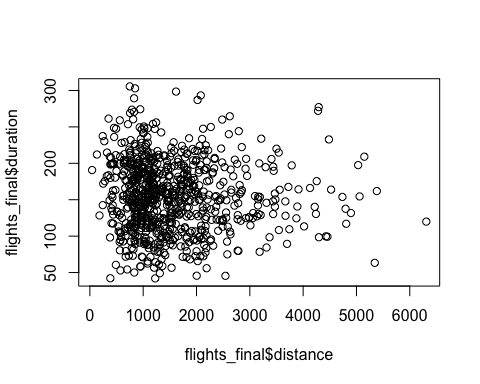
plot(flights\_final$distance, flights\_final$height)



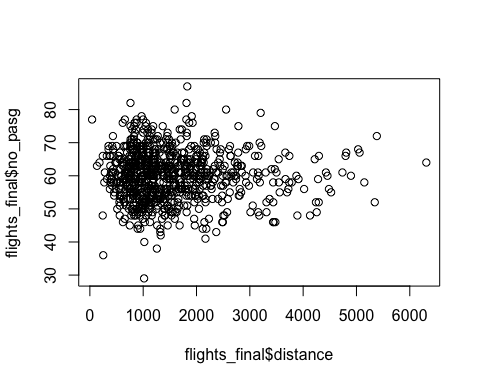
plot(flights\_final$distance, flights\_final$pitch)



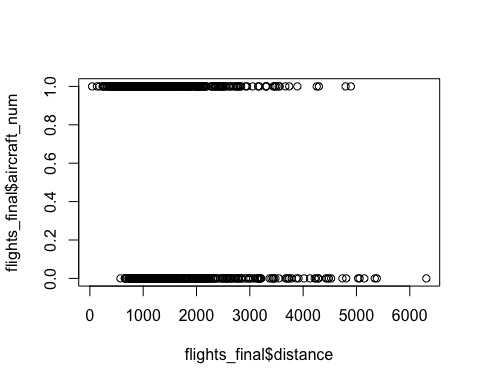
plot(flights\_final$distance, flights\_final$duration)



plot(flights\_final$distance, flights\_final$no\_pasg)



plot(flights\_final$distance, flights\_final$aircraft\_num)



## Step 12: The aiplane make has already been considered as a factor in steps 11-12.

# Regression using a single factor each time

## Step 13:

speed\_air <- lm(distance ~ speed\_air, data = flights\_final)  
speed\_ground <- lm(distance ~ speed\_ground, data = flights\_final)  
aircraft\_num <- lm(distance ~ aircraft\_num, data = flights\_final)  
height <- lm(distance ~ height, data = flights\_final)  
pitch <- lm(distance ~ pitch, data = flights\_final)  
duration <- lm(distance ~ duration, data = flights\_final)  
no\_pasg <- lm(distance ~ no\_pasg, data = flights\_final)  
  
  
variable\_names <- c("speed\_air", "speed\_ground", "aircraft\_num", "height", "pitch", "duration", "no\_pasg")  
  
  
p\_values <- c(  
summary(speed\_air)$coefficients[2,4],  
summary(speed\_ground)$coefficients[2,4],  
summary(aircraft\_num)$coefficients[2,4],  
summary(height)$coefficients[2,4],  
summary(pitch)$coefficients[2,4],  
summary(duration)$coefficients[2,4],  
summary(no\_pasg)$coefficients[2,4])  
  
regression\_coefs <- c(  
summary(speed\_air)$coefficients[2,1],  
summary(speed\_ground)$coefficients[2,1],  
summary(aircraft\_num)$coefficients[2,1],  
summary(height)$coefficients[2,1],  
summary(pitch)$coefficients[2,1],  
summary(duration)$coefficients[2,1],  
summary(no\_pasg)$coefficients[2,1])  
  
  
Table\_2 <- data.frame(variable\_names, p\_values, regression\_coefs)  
  
Table\_2$direction\_of\_coefficient <- ifelse(Table\_2$regression\_coefs < 0, "Negative", "Positive")  
  
Table\_2 <- dplyr::select(Table\_2, variable\_names, p\_values, direction\_of\_coefficient)  
  
Table\_2 <- arrange(Table\_2, p\_values)  
  
Table\_2

## variable\_names p\_values direction\_of\_coefficient  
## 1 speed\_ground 2.833344e-252 Positive  
## 2 speed\_air 1.980682e-99 Positive  
## 3 aircraft\_num 1.952008e-12 Negative  
## 4 height 2.087142e-03 Positive  
## 5 pitch 1.153563e-02 Positive  
## 6 duration 1.226206e-01 Negative  
## 7 no\_pasg 6.840372e-01 Negative

## Step 14: Standardising all X variables

#Calculating Means   
speed\_air\_bar <- mean(flights\_final$speed\_air, na.rm = TRUE)  
speed\_ground\_bar <- mean(flights\_final$speed\_ground)  
aircraft\_num\_bar <- mean(flights\_final$aircraft\_num)  
height\_bar <- mean(flights\_final$height)  
pitch\_bar <- mean(flights\_final$pitch)  
duration\_bar <- mean(flights\_final$duration, na.rm = TRUE)  
no\_pasg\_bar <- mean(flights\_final$no\_pasg)  
  
  
  
  
#Calculting Standard Deviations  
speed\_air\_sd <- sd(flights\_final$speed\_air, na.rm = TRUE)  
speed\_ground\_sd <- sd(flights\_final$speed\_ground)  
aircraft\_num\_sd <- sd(flights\_final$aircraft\_num)  
height\_sd <- sd(flights\_final$height)  
pitch\_sd <- sd(flights\_final$pitch)  
duration\_sd <- sd(flights\_final$duration, na.rm = TRUE)  
no\_pasg\_sd <- sd(flights\_final$no\_pasg)  
  
#Converting to standardised variables   
  
flights\_final$speed\_air\_std <- (flights\_final$speed\_air- speed\_air\_bar) \* (1/speed\_air\_sd)  
  
flights\_final$speed\_ground\_std <- (flights\_final$speed\_ground - speed\_ground\_bar) \* (1/speed\_ground\_sd )  
  
flights\_final$aircraft\_num\_std <- (flights\_final$aircraft\_num - aircraft\_num\_bar) \* (1/aircraft\_num\_sd)  
  
flights\_final$height\_std <- (flights\_final$height - height\_bar) \* (1/height\_sd)  
  
flights\_final$pitch\_std <- (flights\_final$pitch - pitch\_bar) \* (1/pitch\_sd)  
  
flights\_final$duration\_std <- (flights\_final$duration - duration\_bar) \* (1/duration\_sd)  
  
flights\_final$no\_pasg\_std <- (flights\_final$no\_pasg - no\_pasg\_bar) \* (1/no\_pasg\_sd)  
  
  
# Regressing "Landing Distance" on the standardised X variables   
  
speed\_air\_std <- lm(distance ~ speed\_air\_std, data = flights\_final)  
speed\_ground\_std <- lm(distance ~ speed\_ground\_std, data = flights\_final)  
aircraft\_num\_std <- lm(distance ~ aircraft\_num\_std, data = flights\_final)  
height\_std <- lm(distance ~ height\_std, data = flights\_final)  
pitch\_std <- lm(distance ~ pitch\_std, data = flights\_final)  
duration\_std <- lm(distance ~ duration\_std, data = flights\_final)  
no\_pasg\_std <- lm(distance ~ no\_pasg\_std, data = flights\_final)  
  
  
variable\_names\_std <- c("speed\_air\_std", "speed\_ground\_std", "aircraft\_num\_std", "height\_std", "pitch\_std", "duration\_std", "no\_pasg\_std")  
  
  
regression\_coefs\_std <- c(  
summary(speed\_air\_std)$coefficients[2,1],  
summary(speed\_ground\_std)$coefficients[2,1],  
summary(aircraft\_num\_std)$coefficients[2,1],  
summary(height\_std)$coefficients[2,1],  
summary(pitch\_std)$coefficients[2,1],  
summary(duration\_std)$coefficients[2,1],  
summary(no\_pasg\_std)$coefficients[2,1])  
  
Table\_3 <- data.frame(variable\_names\_std, regression\_coefs\_std)  
  
Table\_3$direction\_of\_coefs <- ifelse(Table\_3$regression\_coefs\_std < 0, "Negative", "Positive")  
  
Table\_3

## variable\_names\_std regression\_coefs\_std direction\_of\_coefs  
## 1 speed\_air\_std 808.72372 Positive  
## 2 speed\_ground\_std 789.13202 Positive  
## 3 aircraft\_num\_std -219.34174 Negative  
## 4 height\_std 97.07273 Positive  
## 5 pitch\_std 79.75118 Positive  
## 6 duration\_std -50.83188 Negative  
## 7 no\_pasg\_std -12.87247 Negative

## Step 15: Final Results Table

* It is observed that our results in Table 1 and Table 3 are consistent, but not consistent with the Table 2. As we observe that there are a lot of NA’s in the predictor variable speed\_air, the p-value when regressing speed\_air on Landing distance is greater than the p-value when regressing speed\_ground on Landing distance. Since there are lot of NA’s in the speed\_air, we should give more importance to the variable speed\_ground.

Importance\_Ranks <- c(2,1,3,4,5,6,7)  
  
Table\_0 <- data.frame(variable\_names, Importance\_Ranks)  
  
Table\_0 <- arrange(Table\_0, Importance\_Ranks)  
  
Table\_0

## variable\_names Importance\_Ranks  
## 1 speed\_ground 1  
## 2 speed\_air 2  
## 3 aircraft\_num 3  
## 4 height 4  
## 5 pitch 5  
## 6 duration 6  
## 7 no\_pasg 7

# Checking Collinearity

## Step 16:

* It is observed that when “Landing Distance” is regressed on “speed\_air” and speed\_ground" both the predictor variables, there is an erractic change in the values of the regression coefficients. We also observe a sign change (negative) for the regression coefficient of speed ground which was positive when “Landing Distance” was only regressed on it.
* The correlation of speed\_air and speed\_ground is 0.99, which suggests that both these predictor variables are strongly associated with each other.
* If we want to select any one of them, we would select the speed\_ground. The reason being we have seen the result Tables 1, 2, 3 that it is the most important predictor variable.

model\_3 <- lm(distance ~ speed\_air + speed\_ground, data = flights\_final)  
  
summary(speed\_air)$coefficients[2,1]

## [1] 81.01572

summary(speed\_ground)$coefficients[2,1]

## [1] 41.91088

summary(model\_3)$coefficients[2:3,1]

## speed\_air speed\_ground   
## 95.10311 -14.03113

cor(flights\_final$speed\_ground, flights\_final$speed\_air)

## [1] NA

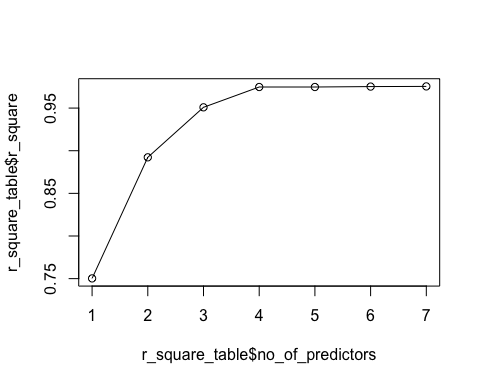
corr.test(flights\_final[,4:5], use = 'pairwise')

## Call:corr.test(x = flights\_final[, 4:5], use = "pairwise")  
## Correlation matrix   
## speed\_ground speed\_air  
## speed\_ground 1.00 0.99  
## speed\_air 0.99 1.00  
## Sample Size   
## speed\_ground speed\_air  
## speed\_ground 832 204  
## speed\_air 204 204  
## Probability values (Entries above the diagonal are adjusted for multiple tests.)   
## speed\_ground speed\_air  
## speed\_ground 0 0  
## speed\_air 0 0  
##   
## To see confidence intervals of the correlations, print with the short=FALSE option

# Variable selection based on our ranking in table 0.

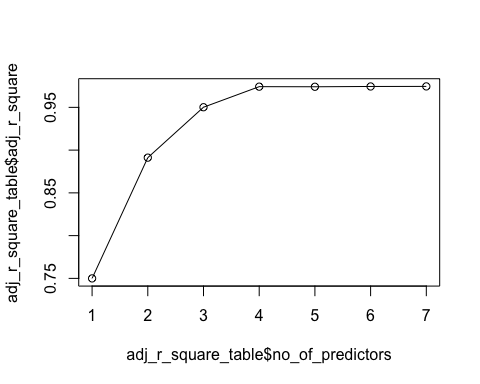
## Step 17: Plotting R square values

model\_01 <- lm(distance ~ speed\_ground, data = flights\_final)  
  
model\_02 <- lm(distance ~ speed\_air + speed\_ground, data = flights\_final)  
  
model\_03 <- lm(distance ~ speed\_air + speed\_ground + aircraft\_num, data = flights\_final)  
  
model\_04 <- lm(distance ~ speed\_air + speed\_ground + aircraft\_num + height, data = flights\_final)  
  
model\_05 <- lm(distance ~ speed\_air + speed\_ground + aircraft\_num + height + pitch, data = flights\_final)  
  
model\_06 <- lm(distance ~ speed\_air + speed\_ground + aircraft\_num + height + pitch + duration, data = flights\_final)  
  
model\_07 <- lm(distance ~ speed\_air + speed\_ground + aircraft\_num + height + pitch + duration + no\_pasg, data = flights\_final)  
  
  
r\_square <- c(  
summary(model\_01)$r.squared,  
summary(model\_02)$r.squared,  
summary(model\_03)$r.squared,  
summary(model\_04)$r.squared,  
summary(model\_05)$r.squared,  
summary(model\_06)$r.squared,  
summary(model\_07)$r.squared)  
  
no\_of\_predictors <- c(1,2,3,4,5,6,7)  
  
r\_square\_table <- data.frame(r\_square, no\_of\_predictors)  
  
plot(r\_square\_table$no\_of\_predictors, r\_square\_table$r\_square)  
lines(r\_square\_table$no\_of\_predictors, r\_square\_table$r\_square)



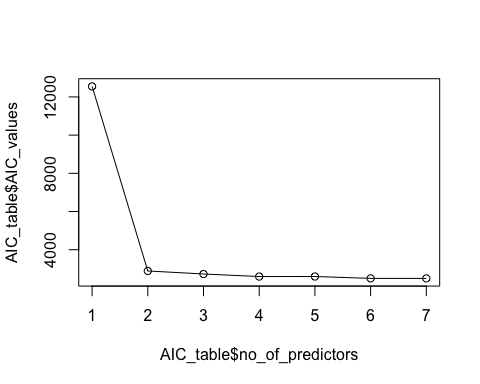
## Step 18: Plotting adjusted R square values.

adj\_r\_square <- c(  
summary(model\_01)$adj.r.squared,  
summary(model\_02)$adj.r.squared,  
summary(model\_03)$adj.r.squared,  
summary(model\_04)$adj.r.squared,  
summary(model\_05)$adj.r.squared,  
summary(model\_06)$adj.r.squared,  
summary(model\_07)$adj.r.squared)  
  
adj\_r\_square\_table <- data.frame(adj\_r\_square, no\_of\_predictors)  
  
plot(adj\_r\_square\_table$no\_of\_predictors, adj\_r\_square\_table$adj\_r\_square)  
lines(adj\_r\_square\_table$no\_of\_predictors, adj\_r\_square\_table$adj\_r\_square)



## Step 19: Plotting AIC values

AIC\_values <- c(AIC(model\_01),  
 AIC(model\_02),  
 AIC(model\_03),  
 AIC(model\_04),  
 AIC(model\_05),  
 AIC(model\_06),  
 AIC(model\_07))  
  
  
AIC\_table <- data.frame(AIC\_values, no\_of\_predictors)  
  
plot(AIC\_table$no\_of\_predictors, AIC\_table$AIC\_values)  
lines(AIC\_table$no\_of\_predictors, AIC\_table$AIC\_values)



## Step 20:

* Comparing the results from steps 18-19 we will only predict the response variable “Landing Distance” using the predictor variable “speed\_ground”.

## Variable selection based on automate algorithm.

# Step 21:

* We will be using the stepAIC fundtion in R to check our results in the previous steps. Since, duration and speed\_air has NA’s we will be excluding the, from our analysis.
* We have used the bi-directional (forward and backward) stepwise regression. The results show that the AIC value is minimum when none of the variables are removed. When we compare results with the one in Step 19, the values are different as stepAIC does not work when there are NA’s in the data. We have to either remove the records with NA values or remove the variables that have NA’s. In this case we have excluded the variables with NA’s and hence the results are varying when compared with step 19.

flights\_2 <- dplyr::select(flights\_final, speed\_ground, aircraft\_num, height, pitch, no\_pasg, distance)  
  
LM <- lm(distance ~ ., data = flights\_2)  
  
fit1\_LM <- stepAIC(LM, direction = 'both')

## Start: AIC=9774.95  
## distance ~ speed\_ground + aircraft\_num + height + pitch + no\_pasg  
##   
## Df Sum of Sq RSS AIC  
## - no\_pasg 1 196930 104014783 9774.5  
## <none> 103817853 9775.0  
## - pitch 1 306686 104124539 9775.4  
## - height 1 16737232 120555085 9897.3  
## - aircraft\_num 1 42765976 146583829 10060.0  
## - speed\_ground 1 537515402 641333255 11287.9  
##   
## Step: AIC=9774.53  
## distance ~ speed\_ground + aircraft\_num + height + pitch  
##   
## Df Sum of Sq RSS AIC  
## <none> 104014783 9774.5  
## + no\_pasg 1 196930 103817853 9775.0  
## - pitch 1 312651 104327433 9775.0  
## - height 1 16601650 120616433 9895.7  
## - aircraft\_num 1 42867444 146882227 10059.6  
## - speed\_ground 1 537455111 641469893 11286.1