Project\_Part1\_Stat\_Modeling

Kevin Gilmore

1/22/2020

# 

## Data Import and Cleaning

### Packages Required

library("dplyr")  
library(ggplot2)  
library(plyr)  
library("tidyverse")  
library(MASS)

### Structure and variables

### Step 1&2

FAA1 <- read.csv("C:/Users/14408/Desktop/Stat Modeling/FAA1.csv")  
FAA2 <- read.csv("C:/Users/14408/Desktop/Stat Modeling/FAA2\_New.csv")  
str(FAA1)

## '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 ...

str(FAA2)

## 'data.frame': 150 obs. of 7 variables:  
## $ aircraft : Factor w/ 2 levels "airbus","boeing": 2 2 2 2 2 2 2 2 2 2 ...  
## $ 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 ...

FAA1 contains 800 observations and 8 variables, while FAA2 contains only 150 observations and 7 variables. FAA also does not contain duration.

### Merge

### Step 3

New\_Combined\_FAA <- full\_join(FAA1,FAA2)

I have used the full\_join function in the dplyr package. I found the merge function to be difficult with the extra variable in the “FAA2” dataset. The full\_join function removes duplicates,thus showing there were 100 duplicates. There is no need to have duplicates, thus always eliminate.

### Step 4

New\_Combined\_FAA <- full\_join(FAA1,FAA2)  
str(New\_Combined\_FAA)

## 'data.frame': 850 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 ...

There are now 850 observation and 8 variables.

### Step 5

* There were 100 duplicate values between the two datasets when merged, which have now been eliminated
* In the summary statistics, the minimum duration is 14.76 minutes, however it should always be greater than 40m.
* There is a -3.546 value in height, however it is requested to be at least 6 meters high
* The length of the airport runway is typically less than 6000 feet, however the maxmimum distance is 6533.05 feet
* The speed\_air category contains majority NA’s, thus you might want to consider deleting

### Step 6: Further exploration

I first checked all the rows individually.

Duration <- New\_Combined\_FAA %>% filter(duration >=1&duration<=40)  
Speed\_Ground\_Less <- New\_Combined\_FAA %>% filter(speed\_ground < 30)  
Speed\_Ground\_Greater <- New\_Combined\_FAA %>% filter(speed\_ground > 140)  
Speed\_Air\_Less <- New\_Combined\_FAA %>% filter(speed\_air < 30)  
Speed\_Air\_Greater <- New\_Combined\_FAA %>% filter(speed\_air > 140)  
Height <- New\_Combined\_FAA %>% filter(height <6)  
Distance <- New\_Combined\_FAA %>% filter(distance > 6000)

I then filtered the datasets and merged them.

FAA1\_filtered <- FAA1 %>%  
 filter(duration >40,height>=6, speed\_ground >=30, speed\_ground <=140,  
 distance < 6000)  
FAA2\_filtered <- FAA2 %>%  
 filter(height>=6, speed\_ground >=30, speed\_ground <=140,  
 distance < 6000)  
filtered <- full\_join(FAA1\_filtered,FAA2\_filtered)

We now have 832 obsevations and 8 variables

### Step 7

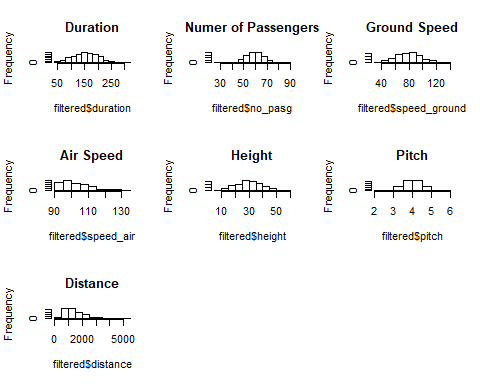
str(filtered)

## 'data.frame': 832 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 ...

summary(filtered)

## aircraft duration no\_pasg speed\_ground   
## airbus:444 Min. : 41.95 Min. :29.00 Min. : 33.57   
## boeing:388 1st Qu.:119.63 1st Qu.:55.00 1st Qu.: 66.16   
## Median :154.28 Median :60.00 Median : 79.77   
## Mean :154.78 Mean :60.06 Mean : 79.52   
## 3rd Qu.:189.66 3rd Qu.:65.00 3rd Qu.: 91.89   
## Max. :305.62 Max. :87.00 Max. :132.78   
## NA's :51   
## speed\_air height pitch distance   
## Min. : 90.00 Min. : 6.228 Min. :2.284 Min. : 41.72   
## 1st Qu.: 96.23 1st Qu.:23.530 1st Qu.:3.640 1st Qu.: 893.43   
## Median :101.12 Median :30.163 Median :4.001 Median :1261.38   
## Mean :103.48 Mean :30.455 Mean :4.005 Mean :1521.89   
## 3rd Qu.:109.36 3rd Qu.:37.000 3rd Qu.:4.370 3rd Qu.:1936.32   
## Max. :132.91 Max. :59.946 Max. :5.927 Max. :5381.96   
## NA's :629

### Step 8



### Step 9

* The minimum duration is no longer 14.76, instead it is 41.95 meeting the normal value
* There is no longer a -3 in height, the minimum is 6.228, thus being a normal value
* With the clean data the distance is no longer over the runway length of 6000 feet, instead it is 5,381.96
* The observation number dropped from 850 to 832

## Initial Analysis

cor(filtered$no\_pasg,filtered$distance)

## [1] -0.01801031

cor(filtered$speed\_ground,filtered$distance)

## [1] 0.8662701

cor(filtered$speed\_air,filtered$distance)

## [1] NA

cor(filtered$height,filtered$distance)

## [1] 0.09952859

cor(filtered$pitch,filtered$distance)

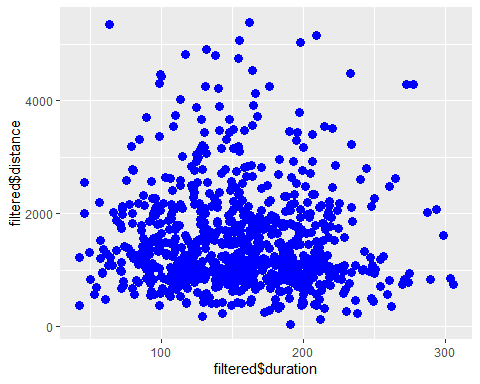
## [1] 0.08709602

table1 <- matrix(c('Speed\_Ground','0.08662701','Positive','Pitch','.08709602','Positive'  
 ,'Height','.09952859','Positive','No\_pasg','-.01801031','Negative'),ncol = 3,byrow = TRUE)  
colnames(table1) <- c('Name','Correlation Size','Direction')  
table1

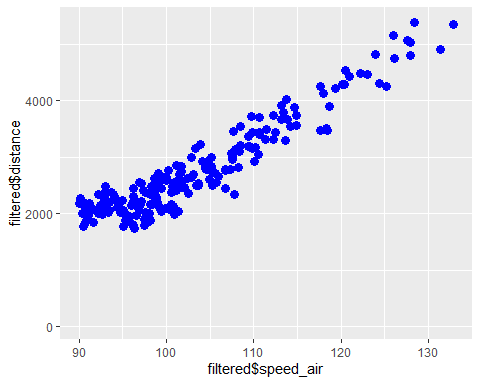
## Name Correlation Size Direction   
## [1,] "Speed\_Ground" "0.08662701" "Positive"  
## [2,] "Pitch" ".08709602" "Positive"  
## [3,] "Height" ".09952859" "Positive"  
## [4,] "No\_pasg" "-.01801031" "Negative"

### Step 11

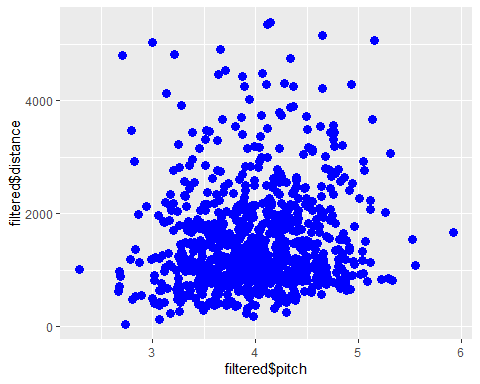
Passenger\_plot <- ggplot(data=filtered,aes(x=filtered$no\_pasg,y=filtered$distance))+  
 geom\_point()+  
 geom\_point(data=filtered,col='blue',size=3)  
  
Speed\_ground\_plot <- ggplot(data=filtered,aes(x=filtered$speed\_ground,y=filtered$distance))+  
 geom\_point()+  
 geom\_point(data=filtered,col='blue',size=3)  
  
Height\_plot <- ggplot(data=filtered,aes(x=filtered$height,y=filtered$distance))+  
 geom\_point()+  
 geom\_point(data=filtered,col='blue',size=3)  
  
Pitch\_plot <- ggplot(data=filtered,aes(x=filtered$pitch,y=filtered$distance))+  
 geom\_point()+  
 geom\_point(data=filtered,col='blue',size=3)  
  
air\_speed\_plot <- ggplot(data = filtered,aes(x=filtered$speed\_air,y=filtered$distance)) +  
 geom\_point()+  
 geom\_point(data = filtered,col='blue',size=3)  
  
duration\_plot <- ggplot(data = filtered,aes(x=filtered$duration,y=filtered$distance)) +  
 geom\_point()+  
 geom\_point(data = filtered,col='blue',size=3)  
  
duration\_plot



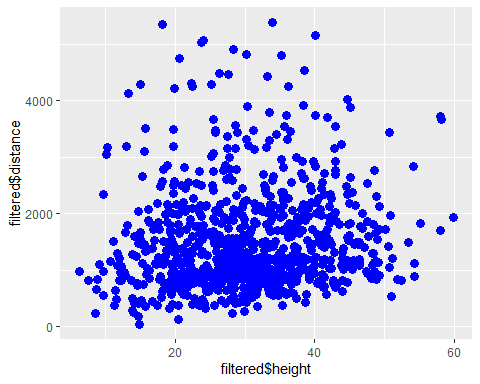
air\_speed\_plot



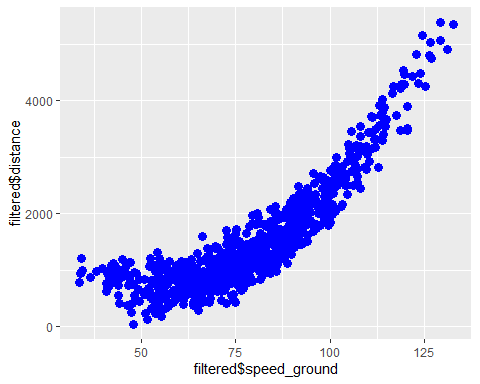
Pitch\_plot



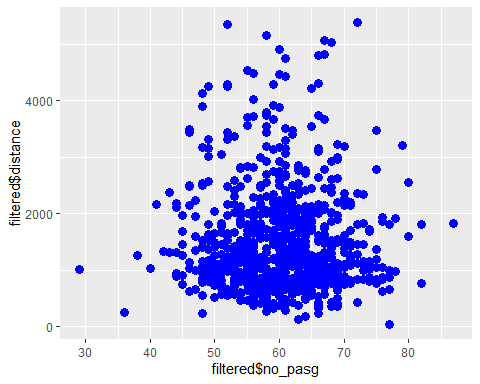
Height\_plot



Speed\_ground\_plot



Passenger\_plot



The speed is consistent.

### Step 12

filtered$aircraft <- revalue(filtered$aircraft,c("boeing"=1))  
filtered$aircraft <- revalue(filtered$aircraft,c ("airbus"=0))  
filtered$aircraft <- as.integer(filtered$aircraft)  
str(filtered$aircraft)

## int [1:832] 2 2 2 2 2 2 2 2 2 2 ...

cor(filtered$aircraft,filtered$distance)

## [1] 0.2372341

## Regression using a single factor each time

### Step 13

model1 <- lm(distance~speed\_ground, data = filtered)  
summary(model1)

##   
## Call:  
## lm(formula = distance ~ speed\_ground, data = filtered)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -897.34 -318.96 -70.73 210.37 1799.14   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1772.9862 67.7582 -26.17 <2e-16 \*\*\*  
## speed\_ground 41.4328 0.8294 49.96 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 447.9 on 830 degrees of freedom  
## Multiple R-squared: 0.7504, Adjusted R-squared: 0.7501   
## F-statistic: 2496 on 1 and 830 DF, p-value: < 2.2e-16

model2<- lm(distance~no\_pasg, data = filtered)  
summary(model2)

##   
## Call:  
## lm(formula = distance ~ no\_pasg, data = filtered)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1443.7 -622.1 -270.5 415.2 3885.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1651.328 251.342 6.570 8.87e-11 \*\*\*  
## no\_pasg -2.155 4.153 -0.519 0.604   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 896.4 on 830 degrees of freedom  
## Multiple R-squared: 0.0003244, Adjusted R-squared: -0.0008801   
## F-statistic: 0.2693 on 1 and 830 DF, p-value: 0.6039

model3 <- lm(distance~speed\_air, data = filtered)  
summary(model3)

##   
## Call:  
## lm(formula = distance ~ speed\_air, data = filtered)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -776.21 -196.39 8.72 209.17 624.34   
##   
## 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 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 276.3 on 201 degrees of freedom  
## (629 observations deleted due to missingness)  
## Multiple R-squared: 0.8875, Adjusted R-squared: 0.887   
## F-statistic: 1586 on 1 and 201 DF, p-value: < 2.2e-16

model4 <- lm(distance~height,data = filtered)  
summary(model4)

##   
## Call:  
## lm(formula = distance ~ height, data = filtered)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1337.7 -606.3 -253.3 388.8 3933.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1244.180 101.212 12.293 < 2e-16 \*\*\*  
## height 9.119 3.164 2.882 0.00406 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 892 on 830 degrees of freedom  
## Multiple R-squared: 0.009906, Adjusted R-squared: 0.008713   
## F-statistic: 8.304 on 1 and 830 DF, p-value: 0.004057

model5<- lm(distance~aircraft, data = filtered)  
summary(model5)

##   
## Call:  
## lm(formula = distance ~ aircraft, data = filtered)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1281.6 -631.8 -229.9 388.2 3632.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 897.50 93.74 9.574 < 2e-16 \*\*\*  
## aircraft 425.81 60.52 7.035 4.16e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 870.9 on 830 degrees of freedom  
## Multiple R-squared: 0.05628, Adjusted R-squared: 0.05514   
## F-statistic: 49.5 on 1 and 830 DF, p-value: 4.156e-12

model6<- lm(distance~duration,data = filtered)  
summary(model6)

##   
## Call:  
## lm(formula = distance ~ duration, data = filtered)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1464.9 -615.6 -274.7 408.5 3847.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1689.9942 108.5452 15.569 <2e-16 \*\*\*  
## duration -0.9613 0.6694 -1.436 0.151   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 904 on 779 degrees of freedom  
## (51 observations deleted due to missingness)  
## Multiple R-squared: 0.00264, Adjusted R-squared: 0.00136   
## F-statistic: 2.062 on 1 and 779 DF, p-value: 0.1514

model7 <- lm(distance~pitch,data=filtered)  
summary(model7)

##   
## Call:  
## lm(formula = distance ~ pitch, data = filtered)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1337.5 -643.7 -240.6 402.2 3840.1   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 928.01 237.81 3.902 0.000103 \*\*\*  
## pitch 148.28 58.87 2.519 0.011963 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 893.1 on 830 degrees of freedom  
## Multiple R-squared: 0.007586, Adjusted R-squared: 0.00639   
## F-statistic: 6.344 on 1 and 830 DF, p-value: 0.01196

table2 <- matrix(c('Speed\_Ground','<2.2e-16','Positive','Speed\_Air','<2.2e-16','Positive','Aircraft','4.156e-12','Positive',  
 'Height','.004057','Positive','Pitch','.01196','Positive','Duration','0.1514','Positive',  
 'No-pasg','.6039','Positive'),ncol = 3,byrow = TRUE)   
colnames(table2) <- c('Name','P-value Size','Direction')  
table2

## Name P-value Size Direction   
## [1,] "Speed\_Ground" "<2.2e-16" "Positive"  
## [2,] "Speed\_Air" "<2.2e-16" "Positive"  
## [3,] "Aircraft" "4.156e-12" "Positive"  
## [4,] "Height" ".004057" "Positive"  
## [5,] "Pitch" ".01196" "Positive"  
## [6,] "Duration" "0.1514" "Positive"  
## [7,] "No-pasg" ".6039" "Positive"

### Step 14

filtered$standardized\_height <- {filtered$height-mean(filtered$height)}/sd(filtered$height)  
  
filtered$standardized\_aircraft <-{filtered$aircraft-mean(filtered$aircraft)}/sd(filtered$aircraft)  
  
filtered$standardized\_no\_pasg <-{filtered$no\_pasg-mean(filtered$no\_pasg)}/sd(filtered$no\_pasg)  
  
filtered$standardized\_no\_pasg <-{filtered$no\_pasg-mean(filtered$no\_pasg)}/sd(filtered$no\_pasg)  
  
filtered$standardized\_speed\_ground<-{filtered$speed\_ground-mean(filtered$speed\_ground)}/sd(filtered$speed\_ground)  
  
filtered$standardized\_speed\_air<-{filtered$speed\_air-mean(filtered$speed\_air)}/sd(filtered$speed\_air)  
  
filtered$standardized\_pitch <- {filtered$pitch-mean(filtered$pitch)}/sd(filtered$pitch)  
  
lm(distance~standardized\_pitch,data=filtered)

##   
## Call:  
## lm(formula = distance ~ standardized\_pitch, data = filtered)  
##   
## Coefficients:  
## (Intercept) standardized\_pitch   
## 1521.89 78.03

lm(distance~standardized\_speed\_ground,data=filtered)

##   
## Call:  
## lm(formula = distance ~ standardized\_speed\_ground, data = filtered)  
##   
## Coefficients:  
## (Intercept) standardized\_speed\_ground   
## 1521.9 776.1

lm(distance~standardized\_no\_pasg,data=filtered)

##   
## Call:  
## lm(formula = distance ~ standardized\_no\_pasg, data = filtered)  
##   
## Coefficients:  
## (Intercept) standardized\_no\_pasg   
## 1521.89 -16.14

lm(distance~standardized\_no\_pasg,data=filtered)

##   
## Call:  
## lm(formula = distance ~ standardized\_no\_pasg, data = filtered)  
##   
## Coefficients:  
## (Intercept) standardized\_no\_pasg   
## 1521.89 -16.14

lm(distance~standardized\_aircraft,data=filtered)

##   
## Call:  
## lm(formula = distance ~ standardized\_aircraft, data = filtered)  
##   
## Coefficients:  
## (Intercept) standardized\_aircraft   
## 1521.9 212.6

lm(distance~standardized\_height,data=filtered)

##   
## Call:  
## lm(formula = distance ~ standardized\_height, data = filtered)  
##   
## Coefficients:  
## (Intercept) standardized\_height   
## 1521.89 89.17

table3 <- matrix(c('Speed\_Ground','776.1','Positive','Aircraft','212.6','Positive',  
 'Height','89.17','Positive','Pitch','78.03','Positive',  
 'No\_pasg','-16.14','Negative'),ncol = 3,byrow = TRUE)   
colnames(table3) <- c('Name','Regression coefficient Size','Direction')  
table3

## Name Regression coefficient Size Direction   
## [1,] "Speed\_Ground" "776.1" "Positive"  
## [2,] "Aircraft" "212.6" "Positive"  
## [3,] "Height" "89.17" "Positive"  
## [4,] "Pitch" "78.03" "Positive"  
## [5,] "No\_pasg" "-16.14" "Negative"

### Step 15

table1

## Name Correlation Size Direction   
## [1,] "Speed\_Ground" "0.08662701" "Positive"  
## [2,] "Pitch" ".08709602" "Positive"  
## [3,] "Height" ".09952859" "Positive"  
## [4,] "No\_pasg" "-.01801031" "Negative"

table2

## Name P-value Size Direction   
## [1,] "Speed\_Ground" "<2.2e-16" "Positive"  
## [2,] "Speed\_Air" "<2.2e-16" "Positive"  
## [3,] "Aircraft" "4.156e-12" "Positive"  
## [4,] "Height" ".004057" "Positive"  
## [5,] "Pitch" ".01196" "Positive"  
## [6,] "Duration" "0.1514" "Positive"  
## [7,] "No-pasg" ".6039" "Positive"

table3

## Name Regression coefficient Size Direction   
## [1,] "Speed\_Ground" "776.1" "Positive"  
## [2,] "Aircraft" "212.6" "Positive"  
## [3,] "Height" "89.17" "Positive"  
## [4,] "Pitch" "78.03" "Positive"  
## [5,] "No\_pasg" "-16.14" "Negative"

The results are consistent as ground speed is the most important factor in all 3, but air speed is also very important. Aircraft seems to have some correlation.

* ground speed
* air speed
* aircraft

## Check collinearity

### Step 16

collinearity1<- lm(distance~speed\_ground, data = filtered)   
collinearity2<- lm(distance~speed\_air,data = filtered)  
collinearity3<- lm(distance~speed\_ground+speed\_air,data=filtered)  
collinearity1

##   
## Call:  
## lm(formula = distance ~ speed\_ground, data = filtered)  
##   
## Coefficients:  
## (Intercept) speed\_ground   
## -1772.99 41.43

collinearity2

##   
## Call:  
## lm(formula = distance ~ speed\_air, data = filtered)  
##   
## Coefficients:  
## (Intercept) speed\_air   
## -5455.71 79.53

collinearity3

##   
## Call:  
## lm(formula = distance ~ speed\_ground + speed\_air, data = filtered)  
##   
## Coefficients:  
## (Intercept) speed\_ground speed\_air   
## -5462.28 -14.37 93.96

In model 3 air speed is added to ground speed and in this model ground speed decreases as air speed continues to increase. As ground speed is a more important factor in terms of correlation size, p-value, and regression coefficient size, I would keep this.

## Variable selection

### Step 17

ranking\_model1 <- lm(distance~speed\_ground,data=filtered)  
ranking\_model2<- lm(distance~speed\_ground+aircraft,data=filtered)  
ranking\_model3 <- lm(distance~speed\_ground+aircraft+height,data = filtered)  
ranking\_model4 <- lm(distance~speed\_ground+aircraft+height+pitch,data = filtered)  
ranking\_model5 <- lm(distance~speed\_ground+aircraft+height+pitch+duration,data = filtered)  
ranking\_model6 <- lm(distance~speed\_ground+aircraft+height+pitch+duration+no\_pasg,data = filtered)  
  
r.squared.1<- summary(ranking\_model1)$r.squared; print(r.squared.1)

## [1] 0.7504239

r.squared.2<- summary(ranking\_model2)$r.squared; print(r.squared.2)

## [1] 0.8251847

r.squared.3<- summary(ranking\_model3)$r.squared; print(r.squared.3)

## [1] 0.8489497

r.squared.4<- summary(ranking\_model4)$r.squared; print(r.squared.4)

## [1] 0.8494237

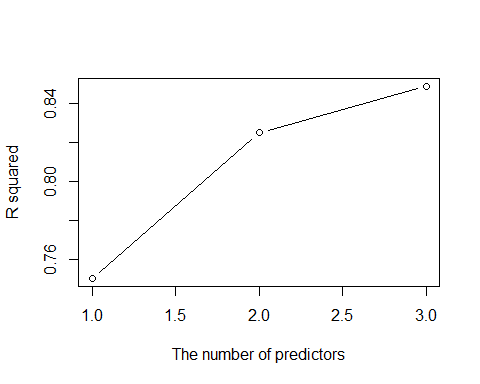
r.squared.5<- summary(ranking\_model5)$r.squared; print(r.squared.5)

## [1] 0.8504184

r.squared.6<- summary(ranking\_model6)$r.squared; print(r.squared.6)

## [1] 0.8506023

plot(c(1,2,3),c(r.squared.1,r.squared.2,r.squared.3),type="b",ylab="R squared",xlab="The number of predictors")



As the number of predictors increases, so does the R squared value.

### Step 18

adj.r.squared.1<-summary(ranking\_model1)$adj.r.squared; print(adj.r.squared.1)

## [1] 0.7501232

adj.r.squared.2<-summary(ranking\_model2)$adj.r.squared; print(adj.r.squared.2)

## [1] 0.8247629

adj.r.squared.3<-summary(ranking\_model3)$adj.r.squared; print(adj.r.squared.3)

## [1] 0.8484024

adj.r.squared.4<-summary(ranking\_model4)$adj.r.squared; print(adj.r.squared.4)

## [1] 0.8486954

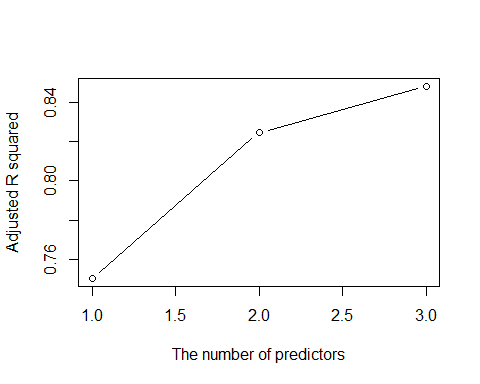
adj.r.squared.5<-summary(ranking\_model5)$adj.r.squared; print(adj.r.squared.5)

## [1] 0.8494534

adj.r.squared.6<-summary(ranking\_model6)$adj.r.squared; print(adj.r.squared.6)

## [1] 0.8494442

plot(c(1,2,3),c(adj.r.squared.1,adj.r.squared.2,adj.r.squared.3),type="b",ylab="Adjusted R squared",xlab="The number of predictors")



### Step 19

AIC(ranking\_model1)

## [1] 12523

AIC(ranking\_model2)

## [1] 12228.78

AIC(ranking\_model3)

## [1] 12109.21

AIC(ranking\_model4)

## [1] 12108.6

AIC(ranking\_model5)

## [1] 11378.84

AIC(ranking\_model6)

## [1] 11379.88

### Step 20

After comparing the results in step 18-19 I would select ground speed and aircraft to build a predictive model

### Step 21

stepAIC(ranking\_model1)

## Start: AIC=10159.89  
## distance ~ speed\_ground  
##   
## Df Sum of Sq RSS AIC  
## <none> 166487253 10160  
## - speed\_ground 1 500592905 667080159 11313

##   
## Call:  
## lm(formula = distance ~ speed\_ground, data = filtered)  
##   
## Coefficients:  
## (Intercept) speed\_ground   
## -1772.99 41.43

stepAIC(ranking\_model2)

## Start: AIC=9865.67  
## distance ~ speed\_ground + aircraft  
##   
## Df Sum of Sq RSS AIC  
## <none> 116615838 9865.7  
## - aircraft 1 49871415 166487253 10159.9  
## - speed\_ground 1 512921039 629536877 11266.5

##   
## Call:  
## lm(formula = distance ~ speed\_ground + aircraft, data = filtered)  
##   
## Coefficients:  
## (Intercept) speed\_ground aircraft   
## -2536.45 41.98 491.20

stepAIC(ranking\_model3)

## Start: AIC=9746.1  
## distance ~ speed\_ground + aircraft + height  
##   
## Df Sum of Sq RSS AIC  
## <none> 100762647 9746.1  
## - height 1 15853192 116615838 9865.7  
## - aircraft 1 50821351 151583998 10083.9  
## - speed\_ground 1 521695198 622457845 11259.1

##   
## Call:  
## lm(formula = distance ~ speed\_ground + aircraft + height, data = filtered)  
##   
## Coefficients:  
## (Intercept) speed\_ground aircraft height   
## -3008.42 42.40 495.92 14.15

stepAIC(ranking\_model4)

## Start: AIC=9745.48  
## distance ~ speed\_ground + aircraft + height + pitch  
##   
## Df Sum of Sq RSS AIC  
## <none> 100446449 9745.5  
## - pitch 1 316198 100762647 9746.1  
## - height 1 15711945 116158394 9864.4  
## - aircraft 1 41870629 142317078 10033.4  
## - speed\_ground 1 522011344 622457793 11261.1

##   
## Call:  
## lm(formula = distance ~ speed\_ground + aircraft + height + pitch,   
## data = filtered)  
##   
## Coefficients:  
## (Intercept) speed\_ground aircraft height pitch   
## -3145.87 42.43 481.15 14.09 39.66