Statistical Modeling_ProjectPart2

#FLIGHT LANDING DISTANCE ANALYSIS - Logistic Regression Following is the continuation of the Part 1 report on the FAA dataset. The response variable of the original FAA dataset has been updated and has been split into 2 binary variables(Long landing, Risky landing) to suit our purpose of using Logistic regression model. The cleaned FAA dataset file has been used as is.

```
library(tidyverse)
library(psych)
library(dplyr)
library(funModeling)
library(ggplot2)
```

```
FAA_uniq <- read.csv("FAA_uniq.csv", header=TRUE)
str(FAA_uniq)
```

```
## 'data.frame':    781 obs. of  8 variables:
##  $ aircraft    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_pasg     : int  36 38 40 41 43 44 45 45 45 45 ...
##  $ speed_ground: num  47.5 85.2 80.6 97.6 82.5 ...
##  $ speed_air   : num  NA NA NA 97 NA ...
##  $ height      : num  14 37 28.6 38.4 30.1 ...
##  $ pitch       : num  4.3 4.12 3.62 3.53 4.09 ...
##  $ distance    : num  251 1257 1021 2168 1321 ...
##  $ duration    : num  172 188 93.5 123.3 109.2 ...
```

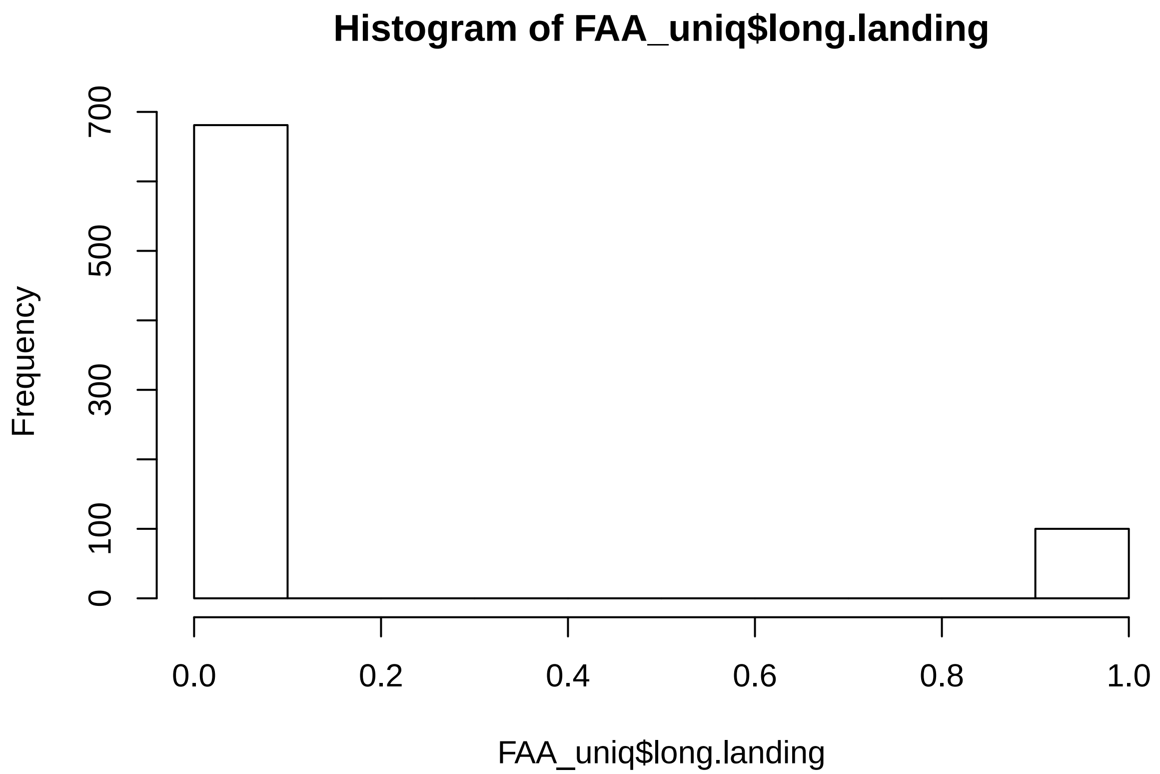
###Step 1

```
###Adding 2 new binary variables
FAA_uniq$long.landing <- ifelse(FAA_uniq$distance>2500,1,0)
FAA_uniq$risky.landing <- ifelse(FAA_uniq$distance>3000,1,0)

#Dropping the distance column
FAA_uniq$distance <- NULL
str(FAA_uniq)
```

```
## 'data.frame':    781 obs. of  9 variables:
##  $ aircraft     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ no_pasg      : int  36 38 40 41 43 44 45 45 45 45 ...
##  $ speed_ground : num  47.5 85.2 80.6 97.6 82.5 ...
##  $ speed_air    : num  NA NA NA 97 NA ...
##  $ height       : num  14 37 28.6 38.4 30.1 ...
##  $ pitch        : num  4.3 4.12 3.62 3.53 4.09 ...
##  $ duration     : num  172 188 93.5 123.3 109.2 ...
##  $ long.landing : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ risky.landing: num  0 0 0 0 0 0 0 0 0 0 ...
```

The two new variables have been added and distance variable has been dropped successfully. ###Step 2

```
hist(FAA_uniq$long.landing)
```




```
colnames(FAA_uniq)
```

```
## [1] "aircraft"      "no_pasg"       "speed_ground"  "speed_air"
## [5] "height"        "pitch"         "duration"      "long.landing"
## [9] "risky.landing"
```

As speed_air variable has close to 75% missing values, is capped and has a high collinearity with speed_ground, we would be ignoring this variable from further analysis .

```
#Dropping the Speed_Air column
FAA_uniq$speed_air <- NULL
```

#Step-3 Performing single-factor regression analysis

```
lr.longlanding.aircraft<- glm(long.landing~aircraft, family=binomial(link = "logit"), data=FAA_uniq)
lr.longlanding.np<- glm(long.landing~no_pasg, family=binomial(link = "logit"), data=FAA_uniq)
lr.longlanding.sg<- glm(long.landing~speed_ground, family=binomial(link = "logit"), data=FAA_uniq)
lr.longlanding.h<- glm(long.landing~height, family=binomial(link = "logit"), data=FAA_uniq)
lr.longlanding.pitch<- glm(long.landing~pitch, family=binomial(link = "logit"), data=FAA_uniq)
lr.longlanding.duration<- glm(long.landing~duration, family=binomial(link = "logit"), data=FAA_uniq)
```

###Table to rank the factors on the basis on regression summary statistics

```
sum.aircraft <- summary(lr.longlanding.aircraft)
sum.aircraft$coefficients[2,4]
```

```
## [1] 0.0002470517
```

```
Variable_Name<- rbind(variable.names(lr.longlanding.aircraft)[2],variable.names(lr.longlanding.np)[2],
variable.names(lr.longlanding.sg)[2],variable.names(lr.longlanding.h)[2],
variable.names(lr.longlanding.pitch)[2],variable.names(lr.longlanding.duration)[2])

Size_Reg_Coeff <- rbind((summary(lr.longlanding.aircraft))$coefficients[2],(summary(lr.longlanding.np))$
```

```r
                  ,(summary(lr.longlanding.sg))$coefficients[2],
                  (summary(lr.longlanding.h))$coefficients[2],
                  (summary(lr.longlanding.pitch))$coefficients[2],
                  (summary(lr.longlanding.duration))$coefficients[2])
odds_ratio <- rbind(exp(coef(lr.longlanding.aircraft)[2]),exp(coef(lr.longlanding.np)[2]),
                    exp(coef(lr.longlanding.sg)[2]),exp(coef(lr.longlanding.h)[2]),
                    exp(coef(lr.longlanding.pitch)[2]),exp(coef(lr.longlanding.duration)[2]))
direction_p <- rbind(ifelse(coef(lr.longlanding.aircraft)[2]>0,'Positive','Negative'),
                     ifelse(coef(lr.longlanding.np)[2]>0,'Positive','Negative'),
                     ifelse(coef(lr.longlanding.sg)[2]>0,'Positive','Negative'),
                     ifelse(coef(lr.longlanding.h)[2]>0,'Positive','Negative'),
                     ifelse(coef(lr.longlanding.pitch)[2]>0,'Positive','Negative'),
                     ifelse(coef(lr.longlanding.duration)[2]>0,'Positive','Negative'))
p_value <- rbind(summary(lr.longlanding.aircraft)$coefficients[2,4],
                 summary(lr.longlanding.np)$coefficients[2,4],
                 summary(lr.longlanding.sg)$coefficients[2,4],
                 summary(lr.longlanding.h)$coefficients[2,4],
                 summary(lr.longlanding.pitch)$coefficients[2,4],
                 summary(lr.longlanding.duration)$coefficients[2,4])
Table_1 <- cbind(Variable_Name,Size_Reg_Coeff,odds_ratio,direction_p,p_value)

colnames(Table_1) <- c("Names of Variables","Size of Regression Coefficient","Odds Ratio","Direction of

Table_1 <- as.data.frame(Table_1)

Table_1[,2] <- as.character(Table_1[,2])
Table_1[,5] <- as.character(Table_1[,5])
Table_1[,3] <- as.character(Table_1[,3])
Table_1[,1] <- as.character(Table_1[,1])
Table_1[,2] <- as.numeric(Table_1[,2])
Table_1[,5] <- as.numeric(Table_1[,5])
Table_1[,3] <- as.numeric(Table_1[,3])


Table_1 <- Table_1[order(Table_1[,5],-Table_1[,3],-abs(Table_1[,2])),]
Table_1
```

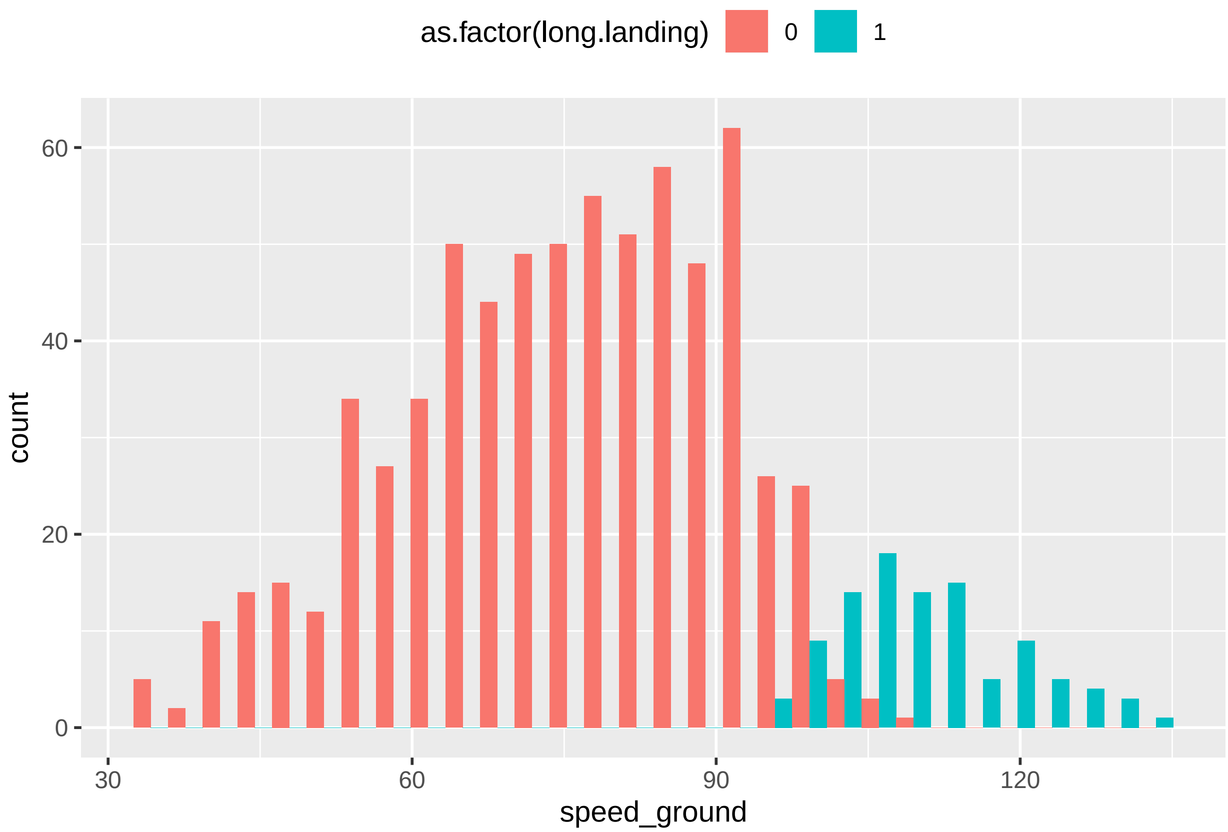
```
##   Names of Variables Size of Regression Coefficient Odds Ratio
## 3       speed_ground                    0.476544128  1.6104991
## 1           aircraft                    0.828742020  2.2904356
## 5              pitch                    0.332877729  1.3949767
## 4             height                    0.006978662  1.0070031
## 2            no_pasg                   -0.006931281  0.9930927
## 6           duration                   -0.001070492  0.9989301
##   Direction of coefficient      p-value
## 3                 Positive 1.296679e-13
## 1                 Positive 2.470517e-04
## 5                 Positive 1.067211e-01
## 4                 Positive 5.250220e-01
## 2                 Negative 6.263969e-01
## 6                 Negative 6.305122e-01
```

```r
str(Table_1)
```

```
## 'data.frame':    6 obs. of  5 variables:
##  $ Names of Variables            : chr  "speed_ground" "aircraft" "pitch" "height" ...
##  $ Size of Regression Coefficient: num  0.47654 0.82874 0.33288 0.00698 -0.00693 ...
##  $ Odds Ratio                    : num  1.61 2.29 1.395 1.007 0.993 ...
##  $ Direction of coefficient      : Factor w/ 2 levels "Negative","Positive": 2 2 2 2 1 1
##  $ p-value                       : num  1.30e-13 2.47e-04 1.07e-01 5.25e-01 6.26e-01 ...
```

###Step4 - Visualize the association

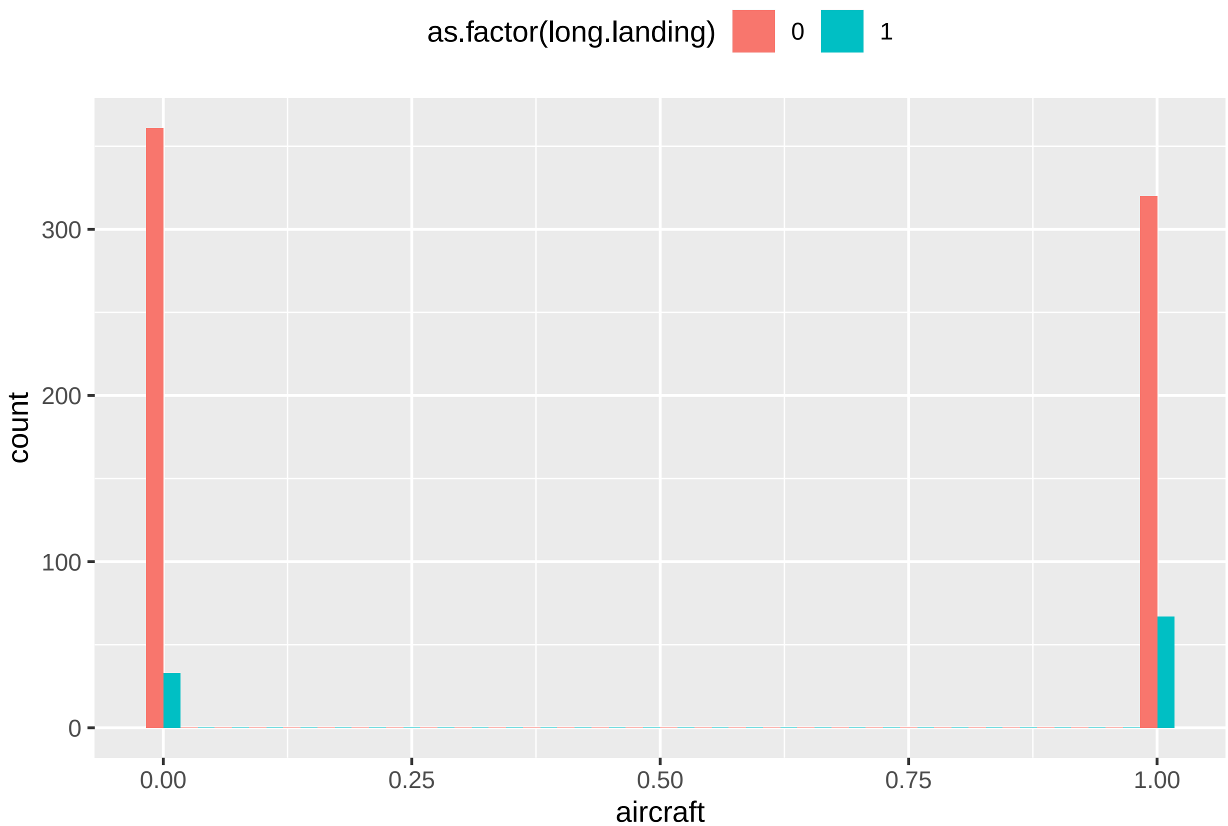
```r
ggplot(FAA_uniq, aes(x=speed_ground, fill=as.factor(long.landing)))+geom_histogram(position="dodge")+the
```



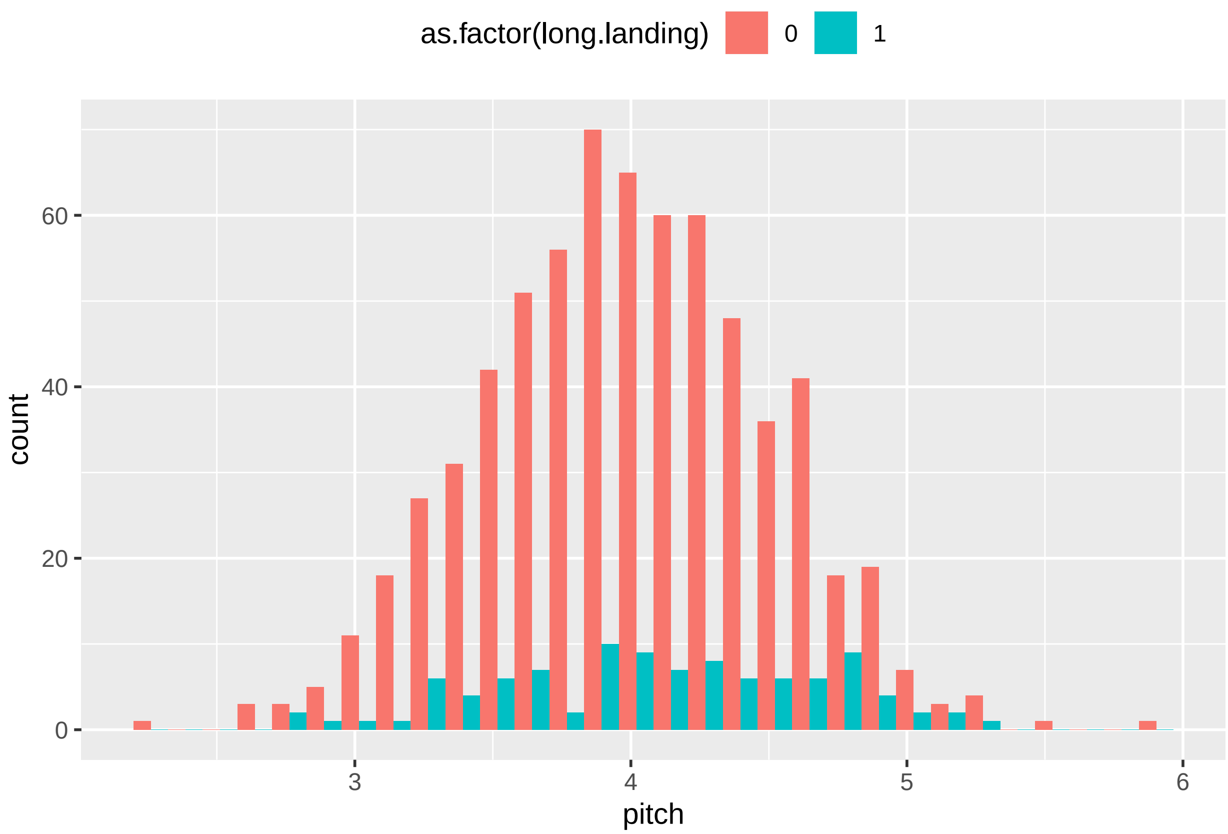
The data for speed_ground in respect to long landing = 0 is rightly skewed.

```r
ggplot(FAA_uniq, aes(x=aircraft, fill=as.factor(long.landing)))+geom_histogram(position="dodge")+theme(
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

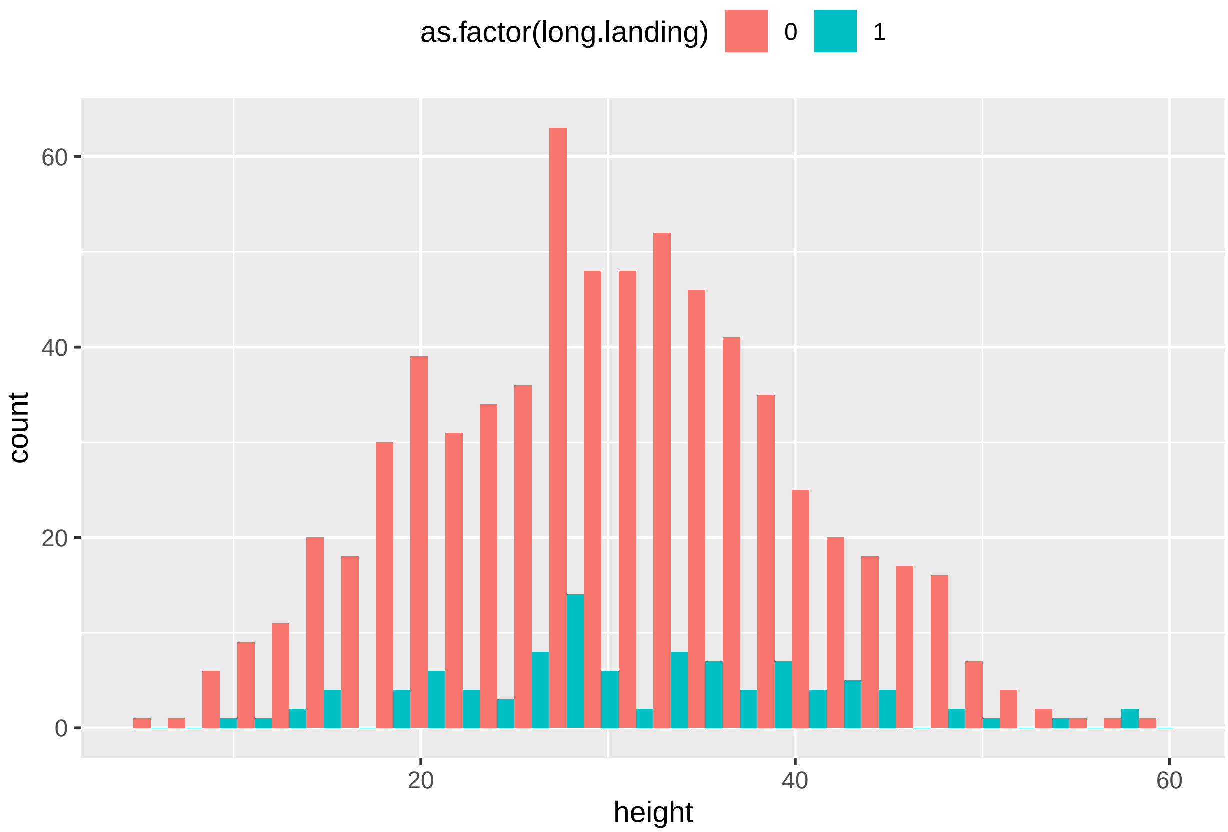



```
ggplot(FAA_uniq, aes(x=pitch, fill=as.factor(long.landing)))+geom_histogram(position="dodge")+theme(lege
```



The data for pitch in respect to long landing = 0 is normally distributed.

```r
ggplot(FAA_uniq, aes(x=height, fill=as.factor(long.landing)))+geom_histogram(position="dodge")+theme(leg
```



The data for height in respect to long landing = 0 is normally distributed.

###Step5 Building the full model

```r
full.lr<-glm(long.landing~., FAA_uniq, family=binomial)
summary(full.lr)
```

```
##
## Call:
## glm(formula = long.landing ~ ., family = binomial, data = FAA_uniq)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.10283  -0.00089   0.00000   0.00000   2.21181
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.131e+02  2.399e+01  -4.715 2.42e-06 ***
## aircraft        4.994e+00  1.189e+00   4.200 2.67e-05 ***
## no_pasg         9.929e-03  5.550e-02   0.179  0.85803
## speed_ground    9.632e-01  2.001e-01   4.815 1.47e-06 ***
## height          2.356e-01  7.174e-02   3.284  0.00102 **
## pitch           1.197e+00  8.521e-01   1.404  0.16019
## duration        5.393e-03  7.649e-03   0.705  0.48077
## risky.landing   1.522e+01  2.566e+03   0.006  0.99527
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 597.692  on 780  degrees of freedom
## Residual deviance:  50.718  on 773  degrees of freedom
## AIC: 66.718
##
## Number of Fisher Scoring iterations: 20
```

#Step6 Forward Variable selection using AIC

```
null_model<- glm(long.landing ~ 1,data=FAA_uniq,family=binomial)
full_model <- glm(long.landing ~ .,data=FAA_uniq,family=binomial)
mAIC<-step(null_model,scope=list(lower=null_model, upper=full_model),trace=0,direction = "forward")
summary(mAIC)
```

```
##
## Call:
## glm(formula = long.landing ~ speed_ground + aircraft + height +
##     pitch, family = binomial, data = FAA_uniq)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.17844  -0.00083   0.00000   0.00000   2.24356
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -113.30604   23.67268  -4.786 1.70e-06 ***
## speed_ground    0.97243    0.19660   4.946 7.57e-07 ***
## aircraft        4.92457    1.16265   4.236 2.28e-05 ***
## height          0.24201    0.06858   3.529 0.000417 ***
## pitch           1.33615    0.84078   1.589 0.112021
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 597.69  on 780  degrees of freedom
## Residual deviance:  51.58  on 776  degrees of freedom
## AIC: 61.58
##
## Number of Fisher Scoring iterations: 12
```

The forward selection method has selected the significant variables to be 1. Speed_Ground 2. Aircraft 3. Height 4. Pitch

#Step7 Forward Variable Selection using BIC

```
mBIC<-step(null_model,scope=list(lower=null_model, upper=full_model),k=log(nrow(FAA_uniq)))
```

```
## Start:  AIC=604.35
## long.landing ~ 1
##
##                 Df Deviance    AIC
## + speed_ground   1   107.40 120.72
## + risky.landing  1   309.08 322.40
```

```
## + aircraft       1   583.49 596.81
## <none>               597.69 604.35
## + pitch          1   595.08 608.40
## + height         1   597.29 610.61
## + no_pasg        1   597.46 610.78
## + duration       1   597.46 610.78
##
## Step:  AIC=120.72
## long.landing ~ speed_ground
##
##                Df Deviance    AIC
## + aircraft      1    78.16  98.15
## + height        1    95.06 115.04
## + pitch         1    97.01 116.99
## <none>              107.40 120.72
## + risky.landing 1   104.66 124.64
## + duration      1   107.30 127.28
## + no_pasg       1   107.37 127.36
## - speed_ground  1   597.69 604.35
##
## Step:  AIC=98.15
## long.landing ~ speed_ground + aircraft
##
##                Df Deviance    AIC
## + height        1    54.40  81.04
## <none>               78.16  98.15
## + pitch         1    75.18 101.82
## + duration      1    76.64 103.28
## + risky.landing 1    77.65 104.29
## + no_pasg       1    77.82 104.47
## - aircraft      1   107.40 120.72
## - speed_ground  1   583.49 596.81
##
## Step:  AIC=81.04
## long.landing ~ speed_ground + aircraft + height
##
##                Df Deviance    AIC
## <none>               54.40  81.04
## + pitch         1    51.58  84.88
## + risky.landing 1    53.63  86.94
## + duration      1    53.68  86.98
## + no_pasg       1    54.40  87.70
## - height        1    78.16  98.15
## - aircraft      1    95.06 115.04
## - speed_ground  1   583.00 602.99
```

```
summary(mBIC)
```

```
##
## Call:
## glm(formula = long.landing ~ speed_ground + aircraft + height,
##     family = binomial, data = FAA_uniq)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
```

```
## -2.35721  -0.00161  -0.00001   0.00000   2.55053
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -98.86483   18.64748  -5.302 1.15e-07 ***
## speed_ground   0.88939    0.16685   5.331 9.79e-08 ***
## aircraft       4.91354    1.10439   4.449 8.62e-06 ***
## height         0.22063    0.06028   3.660 0.000252 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 597.692  on 780  degrees of freedom
## Residual deviance:  54.401  on 777  degrees of freedom
## AIC: 62.401
##
## Number of Fisher Scoring iterations: 11
```

BIC penalizes free parameters more strongly, unlike the AIC. Hence, as per BIC, the best model is with 3 factors(speed_ground, aircraft, height) removing the additional and less significant factor(pitch) while AIC provided the good model with all the 4 factors.

#Step8 Important Inferences to present 1. Best Logistic Regression model: #long landing=-98.86 + 0.89*speed_ground + 4.91*aircraft + 0.22*height 2. We have ignored speed_air as a variable from our model as it has a lot of missing values and is highly collinear with speed_ground. 3. The analysis after performing forward selection using AIC and BIC has provided us with the best model having 3 predictor variables(speed_Ground, Aircraft and Height) having AIC=62.40. 4. p-values are very less than 0.05; and regression coefficients are positive depicting to be in same direction of the response variable.

##Identifying important factors using the binary data of "risky.landing" Step9 - Performing the same steps for risky.landing variable

```
#Fitting the logistic model - single factor analysis
lr.risklanding<-glm(risky.landing~., family=binomial, FAA_uniq)
summary(lr.risklanding)
```

```
##
## Call:
## glm(formula = risky.landing ~ ., family = binomial, data = FAA_uniq)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.423   0.000   0.000   0.000   1.850
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.135e+02  2.695e+03  -0.042  0.96641
## aircraft      4.271e+00  1.583e+00   2.698  0.00698 **
## no_pasg      -8.492e-02  5.999e-02  -1.416  0.15692
## speed_ground  9.093e-01  2.521e-01   3.607  0.00031 ***
## height        4.001e-02  4.618e-02   0.866  0.38634
## pitch         5.735e-01  8.002e-01   0.717  0.47359
## duration      3.038e-04  1.204e-02   0.025  0.97987
## long.landing  1.460e+01  2.695e+03   0.005  0.99568
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 423.215  on 780  degrees of freedom
## Residual deviance:  36.171  on 773  degrees of freedom
## AIC: 52.171
##
## Number of Fisher Scoring iterations: 23
```

```
coef(lr.risklanding)
```

```
##   (Intercept)      aircraft       no_pasg  speed_ground        height
## -1.135088e+02  4.270629e+00 -8.491912e-02  9.092891e-01  4.000837e-02
##         pitch      duration  long.landing
##  5.734726e-01  3.038288e-04  1.459737e+01
```

```
lr.risklanding.aircraft<- glm(risky.landing~aircraft, family=binomial(link = "logit"), data=FAA_uniq)
lr.risklanding.np<- glm(risky.landing~no_pasg, family=binomial(link = "logit"), data=FAA_uniq)
lr.risklanding.sg<- glm(risky.landing~speed_ground, family=binomial(link = "logit"), data=FAA_uniq)
lr.risklanding.h<- glm(risky.landing~height, family=binomial(link = "logit"), data=FAA_uniq)
lr.risklanding.pitch<- glm(risky.landing~pitch, family=binomial(link = "logit"), data=FAA_uniq)
lr.risklanding.duration<- glm(risky.landing~duration, family=binomial(link = "logit"), data=FAA_uniq)
```

###Table to rank the factors on the basis on regression summary statistics

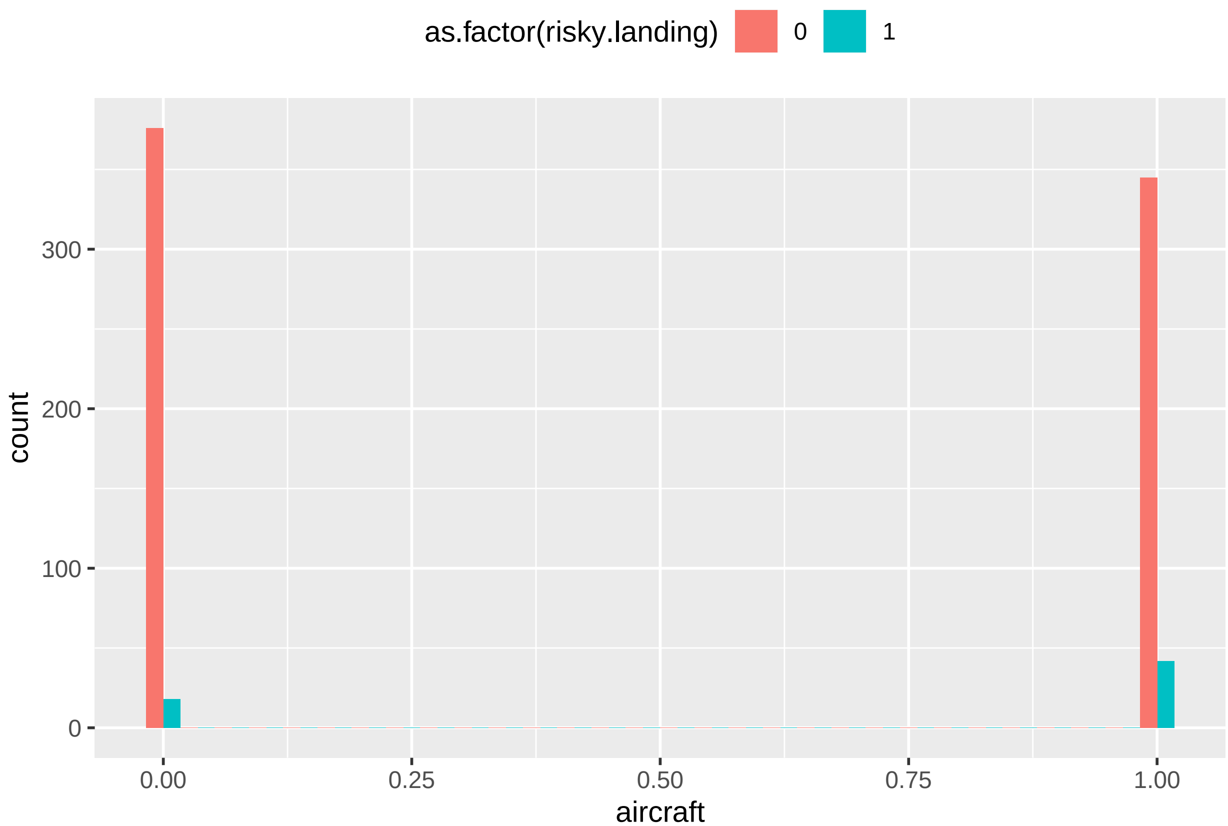
```
sum.aircraft.risk <- summary(lr.risklanding)

Variable_Name1<- rbind(variable.names(lr.risklanding.aircraft)[2],variable.names(lr.risklanding.np)[2],
variable.names(lr.risklanding.sg)[2],variable.names(lr.risklanding.h)[2],
variable.names(lr.risklanding.pitch)[2],variable.names(lr.risklanding.duration)[2])

Size_Reg_Coeff1 <- rbind((summary(lr.risklanding.aircraft))$coefficients[2],(summary(lr.risklanding.np))
                  ,(summary(lr.risklanding.sg))$coefficients[2],
                  (summary(lr.risklanding.h))$coefficients[2],
                  (summary(lr.risklanding.pitch))$coefficients[2],
                  (summary(lr.risklanding.duration))$coefficients[2])
odds_ratio <- rbind(exp(coef(lr.risklanding.aircraft)[2]),exp(coef(lr.risklanding.np)[2]),
                    exp(coef(lr.risklanding.sg)[2]),exp(coef(lr.risklanding.h)[2]),
                    exp(coef(lr.risklanding.pitch)[2]),exp(coef(lr.risklanding.duration)[2]))
direction_p <- rbind(ifelse(coef(lr.risklanding.aircraft)[2]>0,'Positive','Negative'),
                     ifelse(coef(lr.risklanding.np)[2]>0,'Positive','Negative'),
                     ifelse(coef(lr.risklanding.sg)[2]>0,'Positive','Negative'),
                     ifelse(coef(lr.risklanding.h)[2]>0,'Positive','Negative'),
                     ifelse(coef(lr.risklanding.pitch)[2]>0,'Positive','Negative'),
                     ifelse(coef(lr.risklanding.duration)[2]>0,'Positive','Negative'))
p_value <- rbind(summary(lr.risklanding.aircraft)$coefficients[2,4],
                 summary(lr.risklanding.np)$coefficients[2,4],
                 summary(lr.risklanding.sg)$coefficients[2,4],
                 summary(lr.risklanding.h)$coefficients[2,4],
                 summary(lr.risklanding.pitch)$coefficients[2,4],
                 summary(lr.risklanding.duration)$coefficients[2,4])
Table_3 <- cbind(Variable_Name,Size_Reg_Coeff,odds_ratio,direction_p,p_value)

colnames(Table_3) <- c("Names of Variables","Size of Regression Coefficient","Odds Ratio","Direction of
```

```
Table_3 <- as.data.frame(Table_3)

Table_3[,2] <- as.character(Table_3[,2])
Table_3[,5] <- as.character(Table_3[,5])
Table_3[,3] <- as.character(Table_3[,3])
Table_3[,1] <- as.character(Table_3[,1])
Table_3[,2] <- as.numeric(Table_3[,2])
Table_3[,5] <- as.numeric(Table_3[,5])
Table_3[,3] <- as.numeric(Table_3[,3])


Table_4 <- Table_3[order(Table_1[,5],-Table_1[,3],-abs(Table_1[,2])),]
Table_4
```

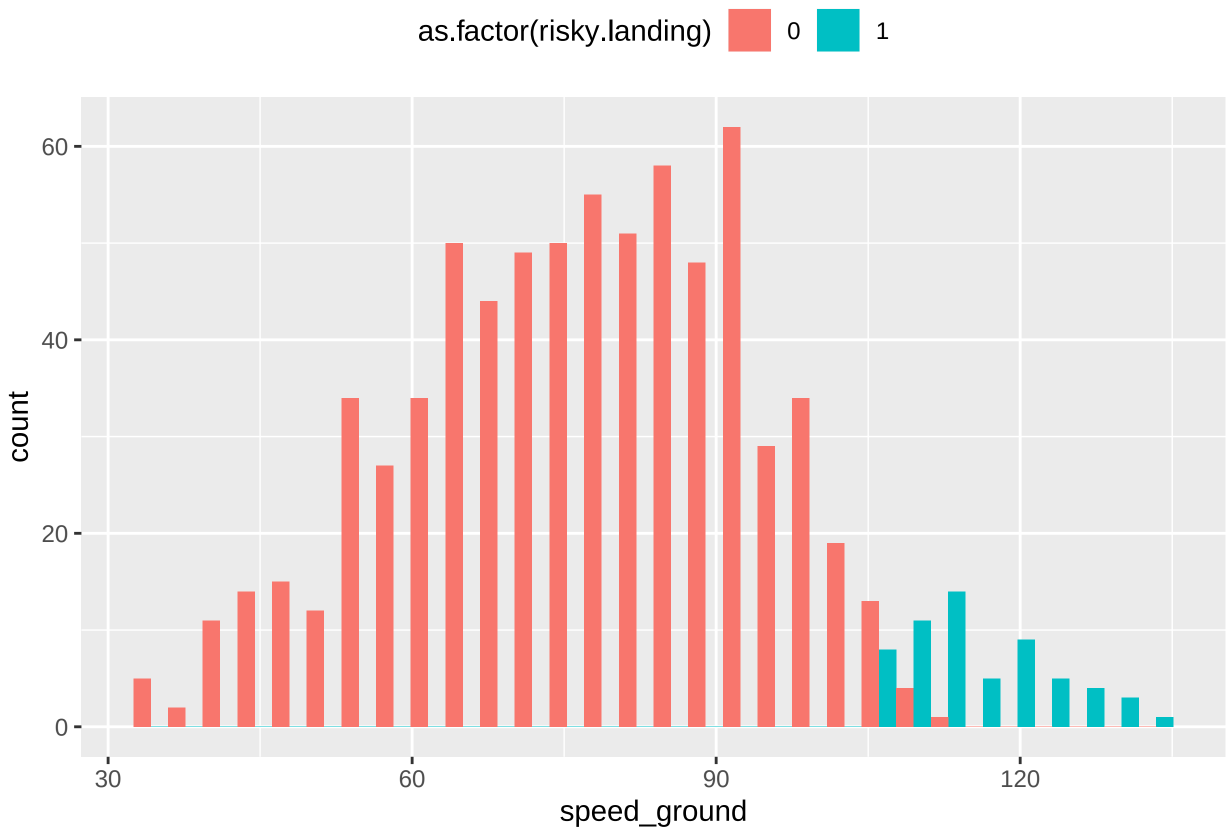
```
##   Names of Variables Size of Regression Coefficient Odds Ratio
## 1           aircraft                    0.828742020  2.5429951
## 2            no_pasg                   -0.006931281  0.9748192
## 3       speed_ground                    0.476544128  1.8194715
## 4             height                    0.006978662  0.9959128
## 5              pitch                    0.332877729  1.3957990
## 6           duration                   -0.001070492  0.9988488
##   Direction of coefficient      p-value
## 1                 Positive 1.359971e-03
## 2                 Negative 1.537335e-01
## 3                 Positive 1.085933e-07
## 4                 Negative 7.672182e-01
## 5                 Positive 1.970447e-01
## 6                 Negative 6.801987e-01
```

###Visualize the association

```
ggplot(FAA_uniq, aes(x=aircraft, fill=as.factor(risky.landing)))+geom_histogram(position="dodge")+theme
```



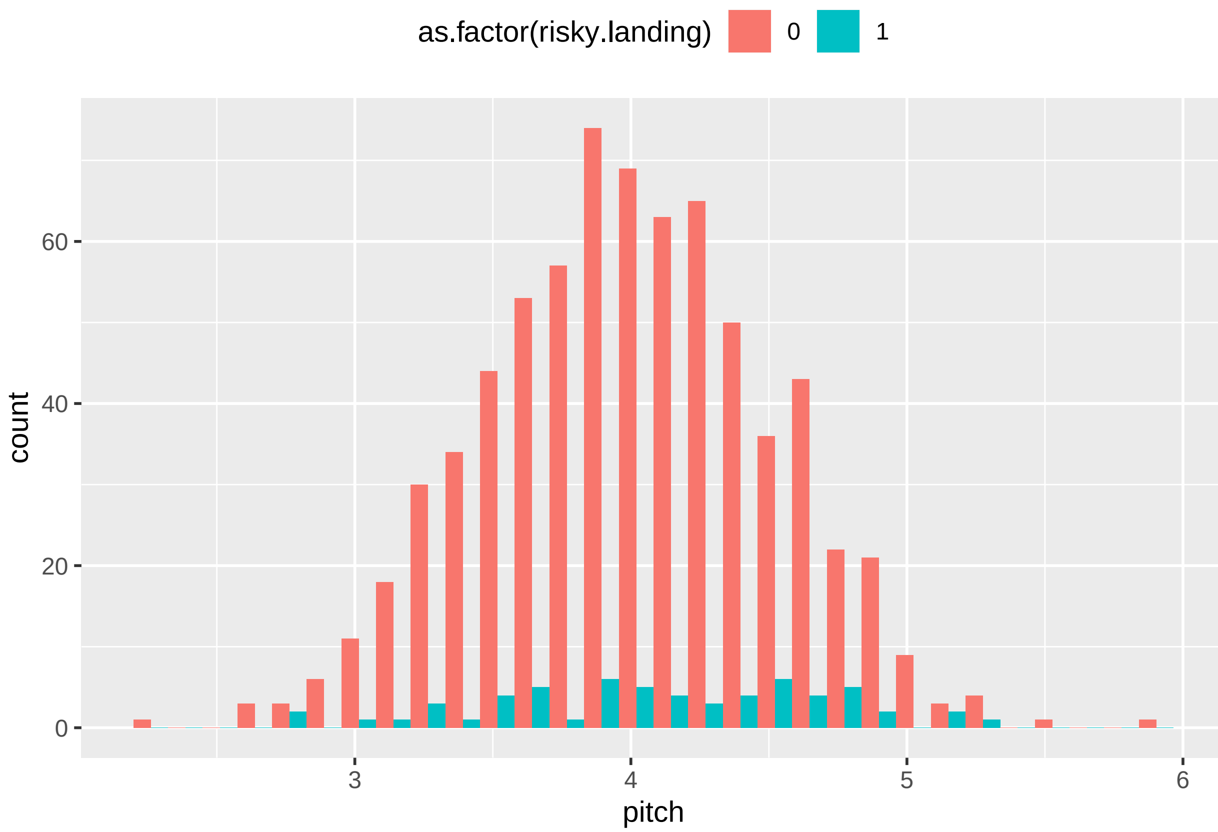

```
ggplot(FAA_uniq, aes(x=speed_ground, fill=as.factor(risky.landing)))+geom_histogram(position="dodge")+th
```



```
ggplot(FAA_uniq, aes(x=height, fill=as.factor(risky.landing)))+geom_histogram(position="dodge")+theme(le
```

The speed_ground variable is normally distributed for risky.landing = 0

```
ggplot(FAA_uniq, aes(x=pitch, fill=as.factor(risky.landing)))+geom_histogram(position="dodge")+theme(leg
```




```r
#fitting the logistic regression model
full2.lr <- glm(risky.landing~.,family=binomial, FAA_uniq)
summary(full2.lr)
```

```
##
## Call:
## glm(formula = risky.landing ~ ., family = binomial, data = FAA_uniq)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.423   0.000   0.000   0.000   1.850
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.135e+02  2.695e+03  -0.042  0.96641
## aircraft      4.271e+00  1.583e+00   2.698  0.00698 **
## no_pasg      -8.492e-02  5.999e-02  -1.416  0.15692
## speed_ground  9.093e-01  2.521e-01   3.607  0.00031 ***
## height        4.001e-02  4.618e-02   0.866  0.38634
## pitch         5.735e-01  8.002e-01   0.717  0.47359
## duration      3.038e-04  1.204e-02   0.025  0.97987
## long.landing  1.460e+01  2.695e+03   0.005  0.99568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 423.215  on 780  degrees of freedom
```

```
## Residual deviance:  36.171  on 773  degrees of freedom
## AIC: 52.171
##
## Number of Fisher Scoring iterations: 23
```

#Performing the forward selection using AIC and BIC criteria

```r
null_model2<- glm(risky.landing ~ 1,data=FAA_uniq,family=binomial)
full_model2 <- glm(risky.landing ~ .,data=FAA_uniq,family=binomial)
#AIC
AIC1<-step(null_model2,scope=list(lower=null_model2, upper=full_model2),trace=0,direction = "forward")
summary(AIC1)
```

```
##
## Call:
## glm(formula = risky.landing ~ speed_ground + aircraft + no_pasg,
##     family = binomial, data = FAA_uniq)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.31800  -0.00011   0.00000   0.00000   1.87101
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -98.34678   25.69889  -3.827  0.00013 ***
## speed_ground   0.93523    0.23663   3.952 7.74e-05 ***
## aircraft       4.59217    1.47920   3.104  0.00191 **
## no_pasg       -0.08442    0.05710  -1.478  0.13929
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 423.215  on 780  degrees of freedom
## Residual deviance:  37.559  on 777  degrees of freedom
## AIC: 45.559
##
## Number of Fisher Scoring iterations: 12
```

#AIC has resulted in a model with 3 most significant factors for the resposne variable, risky landing which are speed_ground, aircraft type and number of passengers.

Using BIC

```r
BIC1 <- step(null_model2,scope=list(lower=null_model2, upper=full_model2),k=log(nrow(FAA_uniq)))
```

```
## Start:  AIC=429.88
## risky.landing ~ 1
##
##                Df Deviance    AIC
## + speed_ground  1    57.99  71.31
## + long.landing  1   134.60 147.92
## + aircraft      1   412.07 425.39
## <none>              423.22 429.88
## + no_pasg       1   421.18 434.50
## + pitch         1   421.54 434.87
## + duration      1   423.04 436.37
```

```
## + height         1   423.13 436.45
##
## Step:  AIC=71.31
## risky.landing ~ speed_ground
##
##                Df Deviance    AIC
## + aircraft      1    39.96  59.94
## <none>               57.99  71.31
## + pitch         1    51.63  71.62
## + long.landing  1    53.53  73.51
## + no_pasg       1    57.18  77.16
## + height        1    57.79  77.77
## + duration      1    57.95  77.93
## - speed_ground  1   423.22 429.88
##
## Step:  AIC=59.94
## risky.landing ~ speed_ground + aircraft
##
##                Df Deviance    AIC
## <none>               39.96  59.94
## + no_pasg       1    37.56  64.20
## + height        1    39.30  65.94
## + long.landing  1    39.46  66.10
## + duration      1    39.76  66.40
## + pitch         1    39.78  66.43
## - aircraft      1    57.99  71.31
## - speed_ground  1   412.07 425.39
```

```
summary(BIC1)
```

```
##
## Call:
## glm(formula = risky.landing ~ speed_ground + aircraft, family = binomial,
##     data = FAA_uniq)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.22465  -0.00014   0.00000   0.00000   1.60326
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -100.6448    24.8834  -4.045 5.24e-05 ***
## speed_ground    0.9132     0.2258   4.044 5.26e-05 ***
## aircraft        3.9763     1.2520   3.176  0.00149 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 423.215  on 780  degrees of freedom
## Residual deviance:  39.955  on 778  degrees of freedom
## AIC: 45.955
##
## Number of Fisher Scoring iterations: 12
```

#BIC has resulted in a model with 2 most significant factors for the resposne variable, risky landing which are speed_ground and aircraft type. #I would be taking the more conservative model that is presented by the BIC variable selection method. Step10 - Important inferences to present for Risky landings variable 1. Best model: #risky landing=-100.64 + 0.91*speed_ground + 3.97*aircraft 2. speed_air contributing to a lot of missing values, so we based our final analysis using speed_ground as the most pertinent factor. 3. I would be taking the more conservative model that is presented by the BIC variable selection method with AIC value being 45.96 4. p-values are very less than 0.05; and regression coefficients are positive depicting to be in same direction of the response variable.

###Compare the two models built for "long.landing" and "risky.landing" #Step 11 1. Model to predict long landing response variable uses less predictors than risky landing model 2. AIC value for risky landing model is less than long landing. 3. Below is the table for the comparison

```
#Step11
data.table::data.table(
    check.names = FALSE,
       `Model:` = c("No. of variables", "AIC","Variable Selection Method", "Common parameters"),
   long.landing = c(3, 62.40,'Backward', 2),
  risky.landing = c(2, 45.96,'Backward', 2)
)
```

```
##                        Model: long.landing risky.landing
## 1:          No. of variables            3             2
## 2:                       AIC         62.4         45.96
## 3: Variable Selection Method     Backward      Backward
## 4:         Common parameters            2             2
```

#Step12 ROC Curve

```
#long landing model
long <-glm(long.landing~aircraft+speed_ground+height, family=binomial,FAA_uniq)

#risky landing model
risky <-glm(risky.landing~aircraft+speed_ground, family=binomial,FAA_uniq)
```

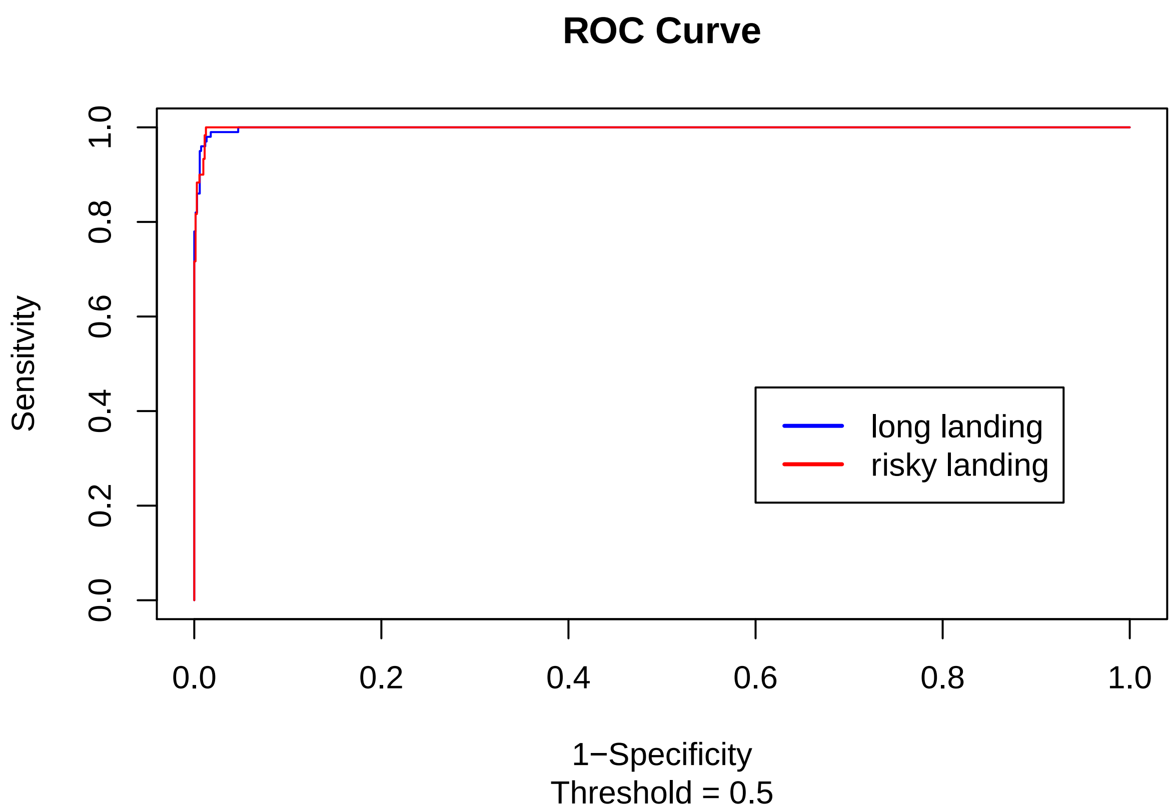
```
#Model evaluation based on predictive power
library(ROCR)
pred1 <- prediction(predict(long), FAA_uniq$long.landing)
perf1 <- performance(pred1,"tpr","fpr")

pred2 <- prediction(predict(risky),FAA_uniq$risky.landing)
perf2 <- performance(pred2,"tpr","fpr")

#AUC values
unlist(slot(performance(pred1 , "auc"), "y.values"))
```

```
## [1] 0.998326
```

```
unlist(slot(performance(pred2 , "auc"), "y.values"))
```

```
## [1] 0.9984975
```

```
plot(perf1,col="blue", main = 'ROC Curve', xlab = "1-Specificity", ylab = "Sensitvity", sub = "Threshol

plot(perf2,col="red",add=TRUE)

legend(0.6,0.45, c('long landing','risky landing'),lty=c(1,1),
lwd=c(2,2),col=c('blue','red'))
```



ROC curve for long landing and risky landing model have almost same AUC.

#Step13 Predicting the probabilities

```
#Boeing, duration=200, no_pasg=80, speed_ground=115, speed_air=120, height=40, pitch=4
#Aircraft is encoded 0/1, Boeing =1
values<-data.frame(aircraft=1, duration=200, no_pasg=80, speed_ground=115, speed_air=120, height=40, pi

#Long.landing(L_P) probability and 95% Confidence Interval
P1 <- predict(long, newdata=values, type="response", se=T)
CI_Long <- c((P1$fit-
                (1.96*P1$se.fit)),(P1$fit+(1.96*P1$se.fit)))
#Risky.landing(R_P) probability and 95% Confidence Interval
P2 <- predict(risky, newdata=values, type="response",se=T)
CI_Risky <- c((P2$fit-
                (1.96*P2$se.fit)),(P2$fit+(1.96*P2$se.fit)))
```

Below is the summary:

```
data.table::data.table(
                         check.names = FALSE,
                         `Model:` = c("Prob", "SE", "95% CI"),
  L_P = c(1, 1.194455e-07,"(0.999,1.000)"),
  R_P = c(0.99976, 0.0004983012, "(0.998,1.000)")
)
```

```
##    Model:           L_P           R_P
## 1:   Prob             1       0.99976
## 2:     SE  1.194455e-07  0.0004983012
## 3: 95% CI (0.999,1.000) (0.998,1.000)
```

#Compare models with different link functions

```
#Step14
risky.probit <- glm(risky.landing~aircraft+speed_ground, family=binomial(link = probit),FAA_uniq)
risky.cloglog <- glm(risky.landing~aircraft+speed_ground, family=binomial(link=cloglog),FAA_uniq)
risky.logit <- glm(risky.landing~aircraft+speed_ground, family=binomial,FAA_uniq)

#Table comparing AIC and Standard error for the 3 models
data.table::data.table(
  Parameters = c("AIC","Coefficients_Speed_ground", "SE_speed_ground"),
      Probit = c(45.32,0.52, 0.12),
     CLogLog = c(47.29,0.61, 0.13),
       Logit = c(45.96,0.91, 0.22)
)
```

```
##                    Parameters Probit CLogLog Logit
## 1:                        AIC  45.32   47.29 45.96
## 2: Coefficients_Speed_ground   0.52    0.61  0.91
## 3:           SE_speed_ground   0.12    0.13  0.22
```

Compared to logistic, the coefficients of probit complementary log-log model are smaller. This happens because probit has fat tails while cloglog is assymentric and has right skewed tail.

#Step15 Plotting ROC curve of different models for riskylanding variable in the same plot

```
#Probit model
pred_p <- prediction(predict(risky.probit), FAA_uniq$long.landing)
perfp1 <- performance(pred_p,"tpr","fpr")

#Hazard model
pred_c <- prediction(predict(risky.cloglog),FAA_uniq$risky.landing)
perfp2 <- performance(pred_c,"tpr","fpr")

#Logit model
pred_l <- prediction(predict(risky.logit),FAA_uniq$risky.landing)
perfp3 <- performance(pred_l,"tpr","fpr")



#AUC values
unlist(slot(performance(pred_p , "auc"), "y.values"))
```

```
## [1] 0.9966373
```

```
unlist(slot(performance(pred_c , "auc"), "y.values"))
```

```
## [1] 0.9984512
```

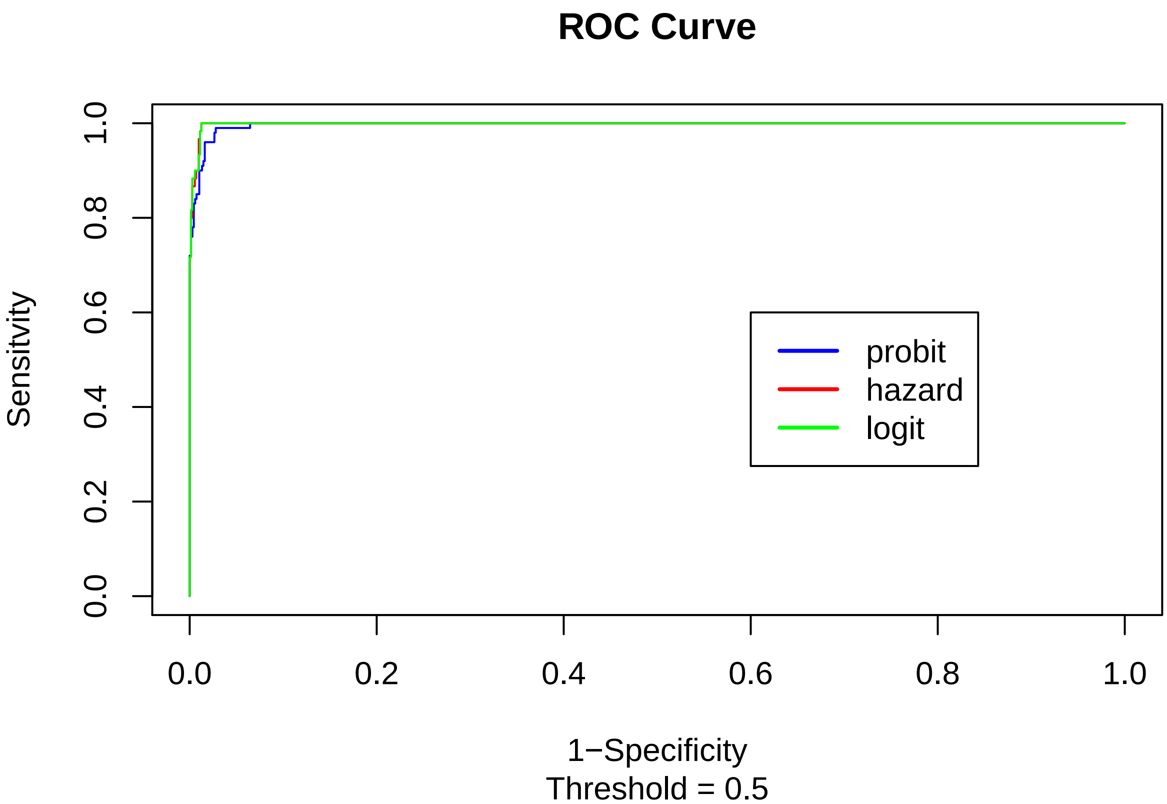
```
unlist(slot(performance(pred_l , "auc"), "y.values"))
```

```
## [1] 0.9984975
```

The Area under curve is almost same for all the different links' models. The Area is maximum for the Logistic model. #Plotting the ROC Curve

```
plot(perfp1,col="blue", main = 'ROC Curve', xlab = "1-Specificity", ylab = "Sensitvity", sub = "Threshol

plot(perfp2,col="red",add=TRUE)
plot(perfp3,col="green",add=TRUE)
```

```r
legend(0.6,0.6, c('probit','hazard','logit'),lty=c(1,1),
lwd=c(2,2),col=c('blue','red','green'))
```



#Step16 Identification of top 5 risky landings

```r
logit_p<- predict(risky.logit, type = "response")
probit_p<- predict(risky.probit, type = "response")
cloglog_p <- predict(risky.cloglog, type = "response") #Hazard

#Top 5 for logit link
head(sort(logit_p, decreasing = TRUE),5)
```

```
## 463 723 764 625 198
##   1   1   1   1   1
```

```r
#Top 5 for probit link
head(sort(probit_p, decreasing = TRUE),5)
```

```
## 198 318 463 520 551
##   1   1   1   1   1
```

```r
#Top 5 for Hazard link
head(sort(cloglog_p, decreasing = TRUE),5)
```

```
##  21  28 198 220 242
##   1   1   1   1   1
```

All the 3 links model have retured different top 5 flights with an overlap of Row 198 in all the three models.

#Step17 Using probit and hazard models to make prediction

```r
###Probit model

#Predicted Probability
probit_prob <- predict(risky.probit, newdata=values, type="response",se=T)
#95% Confidence interval
CI_probit <- c((probit_prob$fit-             (1.96*probit_prob$se.fit)),(probit_prob$fit+(1.96*probit_pr
```

```r
###Hazard model
#Predicted Probability
hazard_prob <- predict(risky.cloglog, newdata=values, type="response",se=T)
#95% Confidence interval
CI_hazard <- c((hazard_prob$fit-             (1.96*hazard_prob$se.fit)),(hazard_prob$fit+(1.96*hazard_pr
```

```r
#Comparison of 3 different link function models
data.table::data.table(
            check.names = FALSE,
                  Model = c("Logit", "Probit", "Hazard"),
  Pred_Prob = c(P2$fit, probit_prob$fit, hazard_prob$fit),
                     SE = c(P2$se.fit, probit_prob$se.fit, hazard_prob$se.fit),
               `95%.CI` = c("(0.9987858,1.0007392)", "(0.9999909,1.0000076)", "(1,1)")
)
```

```
##     Model Pred_Prob           SE                95%.CI
## 1:  Logit 0.9997625 4.983012e-04 (0.9987858,1.0007392)
## 2: Probit 0.9999993 4.258755e-06 (0.9999909,1.0000076)
## 3: Hazard 1.0000000 2.641028e-16                 (1,1)
```