Flight Landing Prediction Project: Part 2

Import Clean data from Part 1 Practice

First, we import the clean data set from Part I. This data has 831 observations with 8 variables.

```
#import cleaned dataset
Data <- read.csv("clean_dataa.csv")
clean_data <- Data[,-1]
str(clean_data)
```

```
## 'data.frame':    831 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(clean_data)
```

```
##    aircraft      duration         no_pasg       speed_ground
##  airbus:444   Min.   : 41.95   Min.   :29.00   Min.   : 33.57
##  boeing:387   1st Qu.:119.63   1st Qu.:55.00   1st Qu.: 66.20
##               Median :154.28   Median :60.00   Median : 79.79
##               Mean   :154.78   Mean   :60.06   Mean   : 79.54
##               3rd Qu.:189.66   3rd Qu.:65.00   3rd Qu.: 91.91
##               Max.   :305.62   Max.   :87.00   Max.   :132.78
##               NA's   :50
##    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.: 893.28
##  Median :101.12   Median :30.167   Median :4.001   Median :1262.15
##  Mean   :103.48   Mean   :30.458   Mean   :4.005   Mean   :1522.48
##  3rd Qu.:109.36   3rd Qu.:37.004   3rd Qu.:4.370   3rd Qu.:1936.63
##  Max.   :132.91   Max.   :59.946   Max.   :5.927   Max.   :5381.96
##  NA's   :628
```

Step 1: Create binary responses

Two binary variables "long.landing" and "risky.landing" are created and added to the cleaned FAA dataset. The variables are created based on the distance variable and these rules.

long.landing = 1 if distance > 2500; =0 otherwise

risky.landing = 1 if distance > 3000; =0 otherwise

The continuous variable titled "distance" is discarded afterwards. The aircraft variable which is factor variable is recoded to a dummy variable, such that 1 = Boeing and 0 = Airbus.

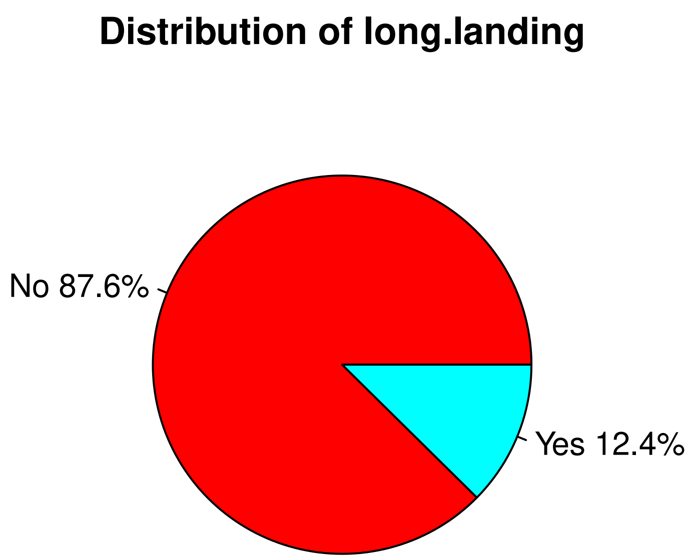
```r
#Create two binary variables
long.landing <- ifelse(clean_data$distance > 2500, 1, 0)
risky.landing <- ifelse(clean_data$distance > 3000, 1, 0)

#code aircraft into dummy variable
aircraft <- ifelse(clean_data$aircraft == "boeing", 1, 0)
clean_data$aircraft <- aircraft

Data_logit <- cbind(clean_data[-8],long.landing, risky.landing)
summary(Data_logit)
```

```
##     aircraft         duration         no_pasg       speed_ground
##  Min.   :0.0000   Min.   : 41.95   Min.   :29.00   Min.   : 33.57
##  1st Qu.:0.0000   1st Qu.:119.63   1st Qu.:55.00   1st Qu.: 66.20
##  Median :0.0000   Median :154.28   Median :60.00   Median : 79.79
##  Mean   :0.4657   Mean   :154.78   Mean   :60.06   Mean   : 79.54
##  3rd Qu.:1.0000   3rd Qu.:189.66   3rd Qu.:65.00   3rd Qu.: 91.91
##  Max.   :1.0000   Max.   :305.62   Max.   :87.00   Max.   :132.78
##                   NA's   :50
##    speed_air          height           pitch        long.landing
##  Min.   : 90.00   Min.   : 6.228   Min.   :2.284   Min.   :0.0000
##  1st Qu.: 96.23   1st Qu.:23.530   1st Qu.:3.640   1st Qu.:0.0000
##  Median :101.12   Median :30.167   Median :4.001   Median :0.0000
##  Mean   :103.48   Mean   :30.458   Mean   :4.005   Mean   :0.1239
##  3rd Qu.:109.36   3rd Qu.:37.004   3rd Qu.:4.370   3rd Qu.:0.0000
##  Max.   :132.91   Max.   :59.946   Max.   :5.927   Max.   :1.0000
##  NA's   :628
##  risky.landing
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.07341
##  3rd Qu.:0.00000
##  Max.   :1.00000
##
```

Step 2: Distribution of long.landing.

A piechart is created to show the distribution of long.landing. No represents "0", while Yes represents "1". The chart shows that 12.4 percent of the observations/flights are long landings (landing distance greater than 2500ft), while the remaining 87.6 percent of the flights are not long landings (that is their landing distances are less than or equal to 2500).

```
#piechart
x=sum (Data_logit$long.landing==0)
y=sum (Data_logit$long.landing==1)
slices <- c(x,y)
lbls <- c("No", "Yes")
pct <- round(slices/sum(slices)*100,1)
lbls <- paste(lbls, pct)
lbls <- paste(lbls,"%",sep="")
pie(slices,labels = lbls, col=rainbow(length(lbls)),
    main="Distribution of long.landing")
```



Step 3: Single-factor regression analysis for each of the potential risk factors.

First, the original variables are considered and the response variable "long.landing" is regressed on each of them individually. The odds ratio for each of the model is obtained using the **odds.ratio()** function in the "questionr" package.

```
library("questionr")
lmod1 <- glm(long.landing~aircraft, family=binomial, Data_logit)
summary(lmod1)
odds.ratio(lmod1)
lmod2 <- glm(long.landing~duration, family=binomial, Data_logit)
summary(lmod2)
odds.ratio(lmod2)
lmod3 <- glm(long.landing~no_pasg, family=binomial, Data_logit)
```

```r
summary(lmod3)
odds.ratio(lmod3)
lmod4 <- glm(long.landing~speed_ground, family=binomial, Data_logit)
summary(lmod4)
odds.ratio(lmod4)
lmod5 <- glm(long.landing~speed_air, family=binomial, Data_logit)
summary(lmod5)
odds.ratio(lmod5)
lmod6 <- glm(long.landing~height, family=binomial, Data_logit)
summary(lmod6)
odds.ratio(lmod6)
lmod7 <- glm(long.landing~pitch, family=binomial, Data_logit)
summary(lmod7)
odds.ratio(lmod7)
```

Then, the factors are ranked from the most important to least important as shown in the table below. The most important risk factors are speed_air, speed_ground, aircraft, and pitch. This is based on their significance as shown by the p values and the size of their respective regression coefficients.

```r
library(knitr)
library(kableExtra)
library(DT)
library(readxl)
Table1 <-read_excel("Data-BANA7042.xls",sheet = 1)
datatable(Table1, options = list(
  searching = TRUE,
  pageLength = 7,
  scrollX = FALSE,
  scrollCollapse = FALSE
))
```

Show 7 entries Search:

	Variables	Size of coefficient	Odds ratio	Direction of regression coefficient	P-value of coef.
1	Speed_air	0.5123	1.6692	Positive	4.33e-11
2	Speed_ground	0.4724	1.6038	Positive	3.94e-14
3	Aircraft	0.8641	2.3729	Positive	0.000084
4	Pitch	0.4005	1.4926	Positive	0.0466
5	Height	0.0086	1.0087	Positive	0.422
6	No_pasg	0.0073	0.9928	Negative	0.6059
7	Duration	0.0011	0.9989	Negative	0.631

Showing 1 to 7 of 7 entries Previous 1 Next

Regression of long.landing on Standardized Predictor variables

Then, each of the X variables are standardized such that X'= {X-mean(X)}/sd(X). The mean of X' is 0 and its standard deviation is 1. The **scale()** function is used for standardizing the X variables. The aircraft variable isn't standardized because it's a factor variable recoded into a dummy variable 0/1. The aircraft variable could lose it's interpretation if standardized.

```
#Standardize each X variable except "aircraft"
Data_logit.norm <- as.data.frame(scale(Data_logit[,-c(1,8,9)]))
Data_logit.n <- cbind(Data_logit$aircraft, Data_logit.norm,
                      Data_logit$long.landing,Data_logit$risky.landing)
summary(Data_logit.n)
```

```
##  Data_logit$aircraft    duration          no_pasg           speed_ground
##  Min.   :0.0000      Min.   :-2.33354   Min.   :-4.145514   Min.   :-2.45353
##  1st Qu.:0.0000      1st Qu.:-0.72687   1st Qu.:-0.674829   1st Qu.:-0.71220
##  Median :0.0000      Median :-0.01016   Median :-0.007389   Median : 0.01341
##  Mean   :0.4657      Mean   : 0.00000   Mean   : 0.000000   Mean   : 0.00000
##  3rd Qu.:1.0000      3rd Qu.: 0.72156   3rd Qu.: 0.660050   3rd Qu.: 0.65999
##  Max.   :1.0000      Max.   : 3.11988   Max.   : 3.596784   Max.   : 2.84174
##                      NA's   :50
##    speed_air           height              pitch
##  Min.   :-1.3847   Min.   :-2.47632   Min.   :-3.26772
##  1st Qu.:-0.7452   1st Qu.:-0.70804   1st Qu.:-0.69259
##  Median :-0.2430   Median :-0.02972   Median :-0.00783
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.6029   3rd Qu.: 0.66905   3rd Qu.: 0.69323
##  Max.   : 3.0223   Max.   : 3.01366   Max.   : 3.64933
##  NA's   :628
##  Data_logit$long.landing Data_logit$risky.landing
##  Min.   :0.0000          Min.   :0.00000
##  1st Qu.:0.0000          1st Qu.:0.00000
##  Median :0.0000          Median :0.00000
##  Mean   :0.1239          Mean   :0.07341
##  3rd Qu.:0.0000          3rd Qu.:0.00000
##  Max.   :1.0000          Max.   :1.00000
##
```

Then, long.landing is regressed on each of the scaled potential risk factors.

```
lmod1.n <- glm(long.landing~aircraft, family=binomial, Data_logit.n)
summary(lmod1.n)
odds.ratio(lmod1.n)
lmod2.n <- glm(long.landing~duration, family=binomial, Data_logit.n)
summary(lmod2.n)
odds.ratio(lmod2.n)
lmod3.n <- glm(long.landing~no_pasg, family=binomial, Data_logit.n)
summary(lmod3.n)
odds.ratio(lmod3.n)
lmod4.n <- glm(long.landing~speed_ground, family=binomial, Data_logit.n)
summary(lmod4.n)
odds.ratio(lmod4.n)
lmod5.n <- glm(long.landing~speed_air, family=binomial, Data_logit.n)
summary(lmod5.n)
odds.ratio(lmod5.n)
lmod6.n <- glm(long.landing~height, family=binomial, Data_logit.n)
```

```r
summary(lmod6.n)
odds.ratio(lmod6.n)
lmod7.n <- glm(long.landing~pitch, family=binomial, Data_logit.n)
summary(lmod7.n)
odds.ratio(lmod7.n)
```

The ranking of the factors based on the regression of the standardized predictor variables show that speed_ground, speed_air, aircraft and pitch are statistically significant and are important risk factors.

```r
Table2 <-read_excel("Data-BANA7042.xls",sheet = 2)
datatable(Table2, options = list(
  searching = TRUE,
  pageLength = 7,
  scrollX = FALSE,
  scrollCollapse = FALSE
))
```

Show 7 entries Search:

	Variables	Size of coefficient	Odds ratio	Direction of regression coefficient	P-value of coef.
1	Speed_ground	8.85	6972.4	Positive	3.94e-14
2	Speed_air	4.9881	146.6585	Positive	4.33e-11
3	Aircraft	0.8641	2.3729	Positive	0.000084
4	Pitch	0.2109	1.2348	Positive	0.0466
5	Height	0.08438	1.08805	Positive	0.422
6	No_pasg	0.05436	0.94709	Negative	0.606
7	Duration	0.05176	0.94956	Negative	0.631

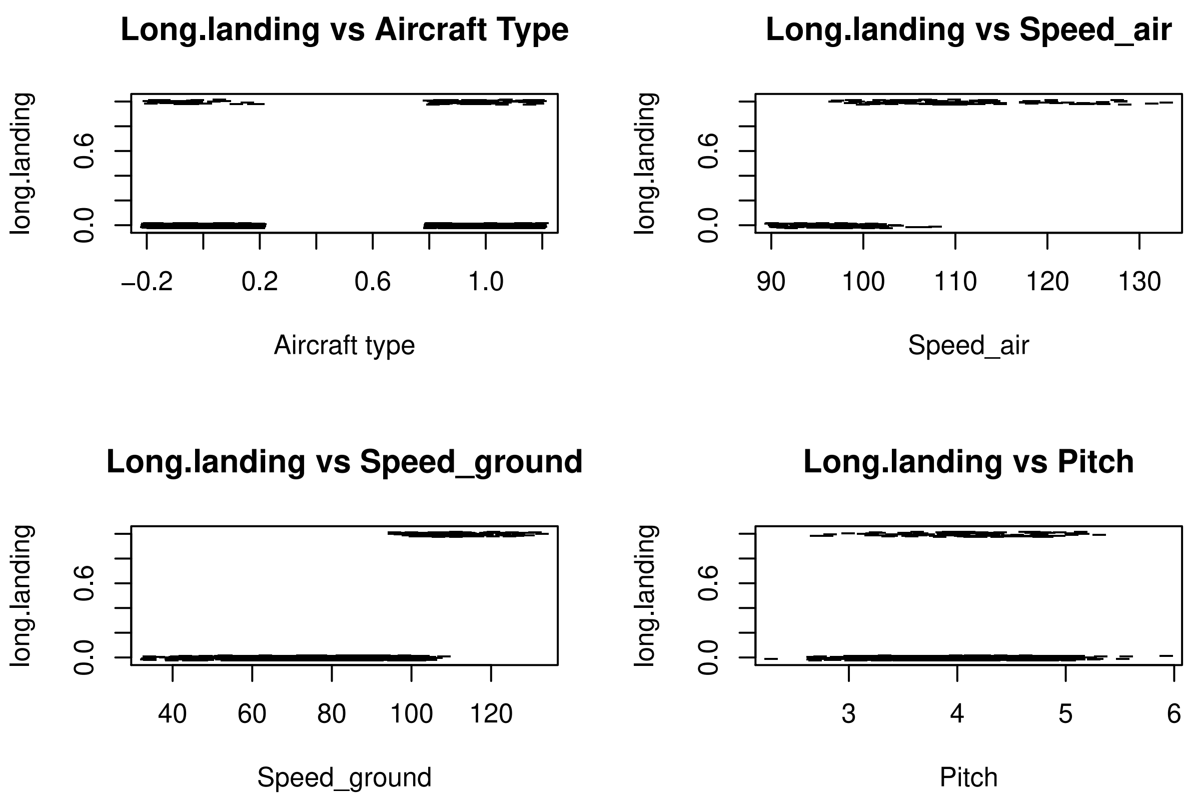
Showing 1 to 7 of 7 entries Previous 1 Next

Step 4: Visualize the association of the significant factors with "long.landing"

The associations are visualized using scatterplots.The plots show that there's a strong association between the following pairs: **long.landing and speed_air** ; **long.landing and speed_ground**. Although, there seem to be some association between **long.landing and pitch** and **long.landing aircraft type** , the plot suggests that the association may be moderate or weak. The plot of long.landing against aircraft shows that there are more Boeing aircrafts that are long landings compared to Airbus aircrafts.

```r
#significant factors : Aircraft, Speed_air, speed_ground,pitch
par(mfrow=c(2,2))
#long.landing vs Aircraft
plot(jitter(long.landing,0.1)~jitter(aircraft), Data_logit,
     xlab ="Aircraft type", ylab ="long.landing", pch="-",
```

```
     main="Long.landing vs Aircraft Type")

#long.landing vs Speed_air
plot(jitter(long.landing,0.1)~jitter(speed_air), Data_logit,
     xlab ="Speed_air", ylab ="long.landing", pch="-",
     main="Long.landing vs Speed_air")

#long.landing vs Speed_ground
plot(jitter(long.landing,0.1)~jitter(speed_ground), Data_logit,
     xlab ="Speed_ground", ylab ="long.landing", pch="-",
     main="Long.landing vs Speed_ground")

#long.landing vs Pitch
plot(jitter(long.landing,0.1)~jitter(pitch), Data_logit,
     xlab ="Pitch", ylab ="long.landing", pch="-",
     main="Long.landing vs Pitch")
```



Step 5: Full model

In part I, Step 16, it was indicated that there's a strong collinearity between speed_air and speed_ground. Though, there are both highly associated with long.landing. To select one of the variables to include in the full model, we look at the individual effect of both variables on long.landing as shown below.

```
#marginal model
summary(glm(long.landing~speed_ground, family=binomial, Data_logit))
```

```
##
## Call:
## glm(formula = long.landing ~ speed_ground, family = binomial,
##     data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.51572  -0.03478  -0.00180  -0.00004   2.37848
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -47.96055    6.28136  -7.635 2.25e-14 ***
## speed_ground   0.47235    0.06245   7.563 3.94e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 622.78  on 830  degrees of freedom
## Residual deviance: 115.47  on 829  degrees of freedom
## AIC: 119.47
##
## Number of Fisher Scoring iterations: 10
```

```r
summary(glm(long.landing~speed_air, family=binomial, Data_logit))
```

```
##
## Call:
## glm(formula = long.landing ~ speed_air, family = binomial, data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.55980  -0.35662   0.00067   0.20985   2.19936
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -51.98365    7.84323  -6.628 3.41e-11 ***
## speed_air     0.51232    0.07772   6.592 4.33e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 281.37  on 202  degrees of freedom
## Residual deviance: 102.33  on 201  degrees of freedom
##   (628 observations deleted due to missingness)
## AIC: 106.33
##
## Number of Fisher Scoring iterations: 7
```

Given that speed_air has a higher coefficient size and a lower AIC value for it's marginal regression model, we could consider including it in the full model. But, the variable has a lot of missing values. Hence, we would include speed_ground instead. First we look at the model with all of the variables, then the model without speed_air. Then, we create a model based on the significant factors identified in Steps 3 and 4. This is called the "full" model.

```r
#Model with all of the predictor variables
lmod.all <- glm (long.landing~aircraft+speed_air+speed_ground+
                  pitch+height+duration+no_pasg,family=binomial,Data_logit)
summary(lmod.all)
```

```
##
## Call:
## glm(formula = long.landing ~ aircraft + speed_air + speed_ground +
##     pitch + height + duration + no_pasg, family = binomial, data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.48853  -0.01367   0.00000   0.00047   1.56917
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.964e+02  5.607e+01  -3.502 0.000462 ***
## aircraft      8.784e+00  2.623e+00   3.349 0.000811 ***
## speed_air     1.985e+00  7.080e-01   2.804 0.005051 **
## speed_ground -2.255e-01  3.845e-01  -0.587 0.557471
## pitch         1.469e+00  1.055e+00   1.392 0.163818
## height        4.226e-01  1.429e-01   2.956 0.003116 **
## duration      3.031e-04  1.048e-02   0.029 0.976919
## no_pasg      -7.359e-02  7.009e-02  -1.050 0.293744
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 270.199  on 194  degrees of freedom
## Residual deviance:  32.909  on 187  degrees of freedom
##   (636 observations deleted due to missingness)
## AIC: 48.909
##
## Number of Fisher Scoring iterations: 10
```

```r
#Model with all predictor variables except speed_air
lmod.all2 <- glm (long.landing~aircraft+speed_ground+
                   pitch+height+duration+no_pasg,family=binomial,Data_logit)
summary(lmod.all2)
```

```
##
## Call:
## glm(formula = long.landing ~ aircraft + speed_ground + pitch +
##     height + duration + no_pasg, family = binomial, data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.12757  -0.00078   0.00000   0.00000   2.19551
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.149e+02  2.411e+01  -4.765 1.89e-06 ***
## aircraft      4.985e+00  1.181e+00   4.221 2.44e-05 ***
## speed_ground  9.795e-01  2.006e-01   4.883 1.05e-06 ***
```

```
## pitch         1.283e+00  8.423e-01   1.523   0.1278
## height        2.346e-01  7.188e-02   3.264   0.0011 **
## duration      5.361e-03  7.704e-03   0.696   0.4865
## no_pasg       7.420e-03  5.559e-02   0.133   0.8938
## ---
## 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:  51.084  on 774  degrees of freedom
##   (50 observations deleted due to missingness)
## AIC: 65.084
##
## Number of Fisher Scoring iterations: 12
```

```r
#Full Model with only significant variables
lmod.full <- glm (long.landing~aircraft+speed_ground+
                    pitch,family=binomial,Data_logit)
summary(lmod.full)
```

```
##
## Call:
## glm(formula = long.landing ~ aircraft + speed_ground + pitch,
##     family = binomial, data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.11589  -0.01116  -0.00026   0.00000   2.40741
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -67.92855   10.48408  -6.479 9.22e-11 ***
## aircraft       3.04348    0.73345   4.150 3.33e-05 ***
## speed_ground   0.61471    0.09184   6.694 2.18e-11 ***
## pitch          1.06599    0.60389   1.765   0.0775 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 622.778  on 830  degrees of freedom
## Residual deviance:  81.309  on 827  degrees of freedom
## AIC: 89.309
##
## Number of Fisher Scoring iterations: 10
```

The full model shows that aircraft and speed_ground are very significant factors that impact "long.landing". Pitch on the otherhand is significant at 10% level of significance, but not at 5%. This is different from the result of the marginal regression where Pitch was significant at 5% level of significance.

Step 6: Forward Variable Selection Using AIC

The model shows some consistency and inconsistencies with the marginal regression models in Step 3. The model and the table in Step 3 show that aircraft and speed_air are significant factors in long landings. However, the model is contrary to the table in Step 3 because height is significant in this step. Likewise, pitch is significant in Step 3, but not significant in the model below.

```
#Step using AIC
model.0 <- glm(long.landing ~ aircraft + duration + no_pasg + speed_ground +
                 speed_air + height + pitch,data = Data_logit,
               family = "binomial")

model.0_AIC <- step(model.0, trace = 0, direction = "forward")
summary(model.0_AIC)
```

```
##
## Call:
## glm(formula = long.landing ~ aircraft + duration + no_pasg +
##     speed_ground + speed_air + height + pitch, family = "binomial",
##     data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.48853  -0.01367   0.00000   0.00047   1.56917
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.964e+02  5.607e+01  -3.502 0.000462 ***
## aircraft      8.784e+00  2.623e+00   3.349 0.000811 ***
## duration      3.031e-04  1.048e-02   0.029 0.976919
## no_pasg      -7.359e-02  7.009e-02  -1.050 0.293744
## speed_ground -2.255e-01  3.845e-01  -0.587 0.557471
## speed_air     1.985e+00  7.080e-01   2.804 0.005051 **
## height        4.226e-01  1.429e-01   2.956 0.003116 **
## pitch         1.469e+00  1.055e+00   1.392 0.163818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 270.199  on 194  degrees of freedom
## Residual deviance:  32.909  on 187  degrees of freedom
##   (636 observations deleted due to missingness)
## AIC: 48.909
##
## Number of Fisher Scoring iterations: 10
```

Step 7: Forward Variable Selection Using BIC

The model is consistent with the model selected in Step 5. Both models show that aircraft, speed_air, and height are significant risk factors and influence long landings.

```
#Step using BIC
model.0_BIC <- step(model.0, trace = 0, direction = "forward", criterion = "BIC")
```

```r
summary(model.0_BIC)
```

```
##
## Call:
## glm(formula = long.landing ~ aircraft + duration + no_pasg +
##     speed_ground + speed_air + height + pitch, family = "binomial",
##     data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.48853  -0.01367   0.00000   0.00047   1.56917
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.964e+02  5.607e+01  -3.502 0.000462 ***
## aircraft      8.784e+00  2.623e+00   3.349 0.000811 ***
## duration      3.031e-04  1.048e-02   0.029 0.976919
## no_pasg      -7.359e-02  7.009e-02  -1.050 0.293744
## speed_ground -2.255e-01  3.845e-01  -0.587 0.557471
## speed_air     1.985e+00  7.080e-01   2.804 0.005051 **
## height        4.226e-01  1.429e-01   2.956 0.003116 **
## pitch         1.469e+00  1.055e+00   1.392 0.163818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 270.199  on 194  degrees of freedom
## Residual deviance:  32.909  on 187  degrees of freedom
##   (636 observations deleted due to missingness)
## AIC: 48.909
##
## Number of Fisher Scoring iterations: 10
```

Step 8: Risk factors for long landings and their influence

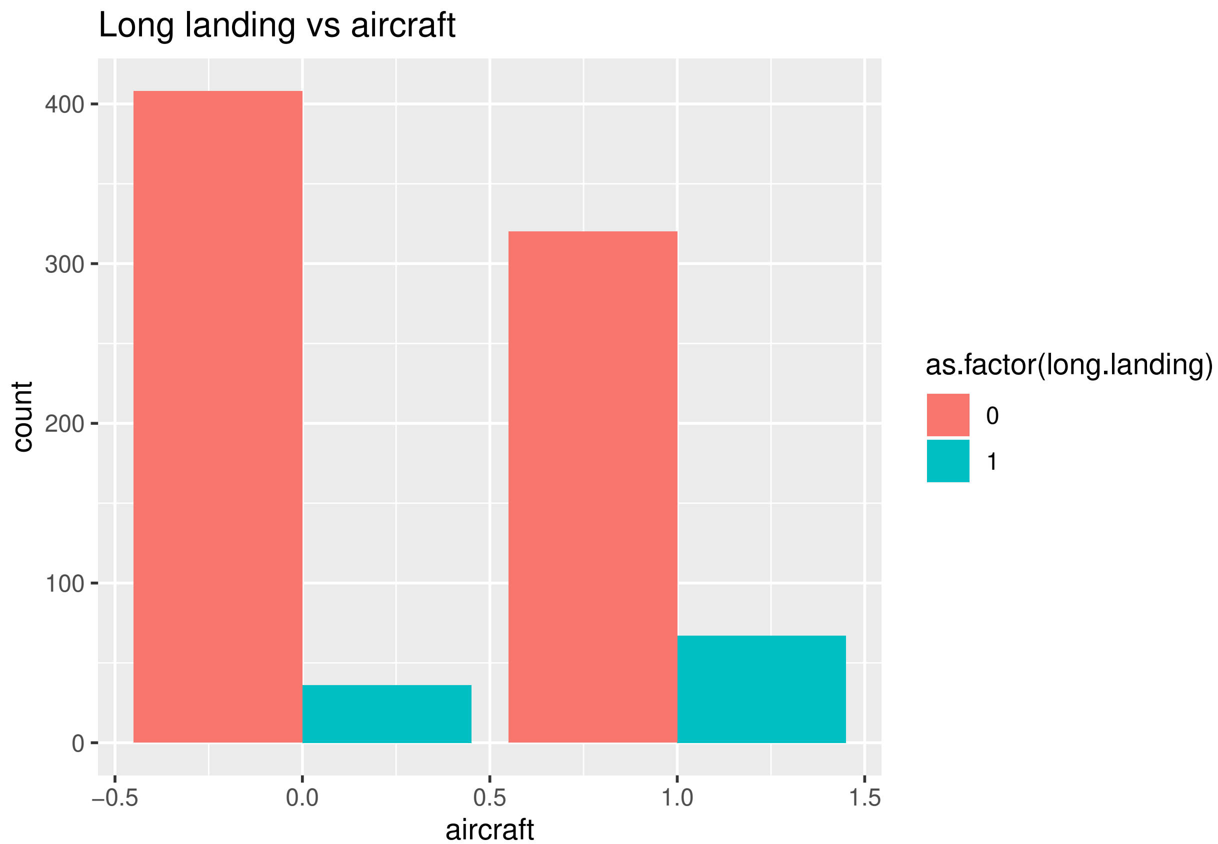
Executive summary

- Aircraft type, the air speed of the aircraft, and the height are the most important risk factors for long landings. An increase in these variables is associated with an increase in the probability of long landing.
- Boeing aircrafts have more long landings (>2500) than Airbus aircrafts. Compared to Boeing, more Airbus aircrafts have landing distance that's less than or equal to 2500ft.
- An increase in the air speed of an aircraft is associated with an increase in the probability of the aircraft being a long landing. For a one-unit increase in the air speed of the aircraft, we expect a 1.4315 increase in the log-odds of long landing.
- An increase in the height of an aircraft is associated with an increase in the probability of the aircraft being a long landing. For a one-unit increase in the height of the aircraft, we expect a 0.349 increase in the log-odds of long landing.
- The variable "speed_air" has 628 missing observations. More observations in this regard may further strengthen my analysis.

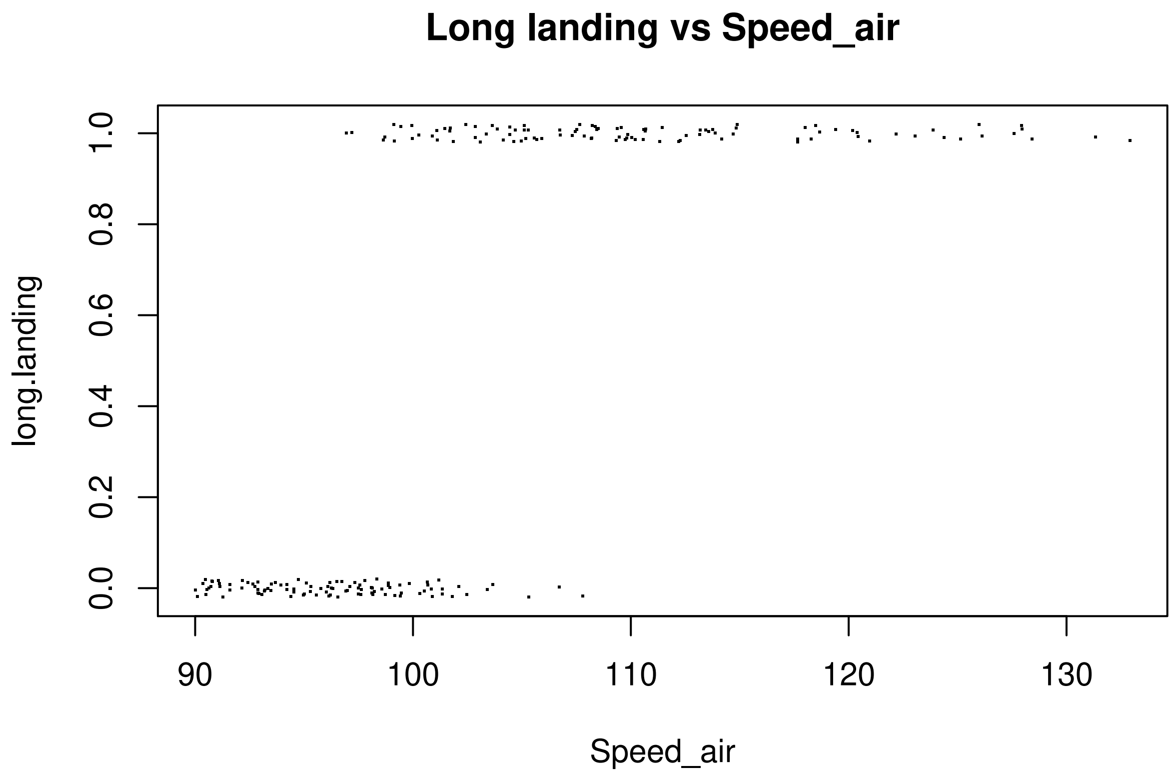
Association between the significant factors and long landings

The figures below show the association between the significant factors and the response variable "long landings".

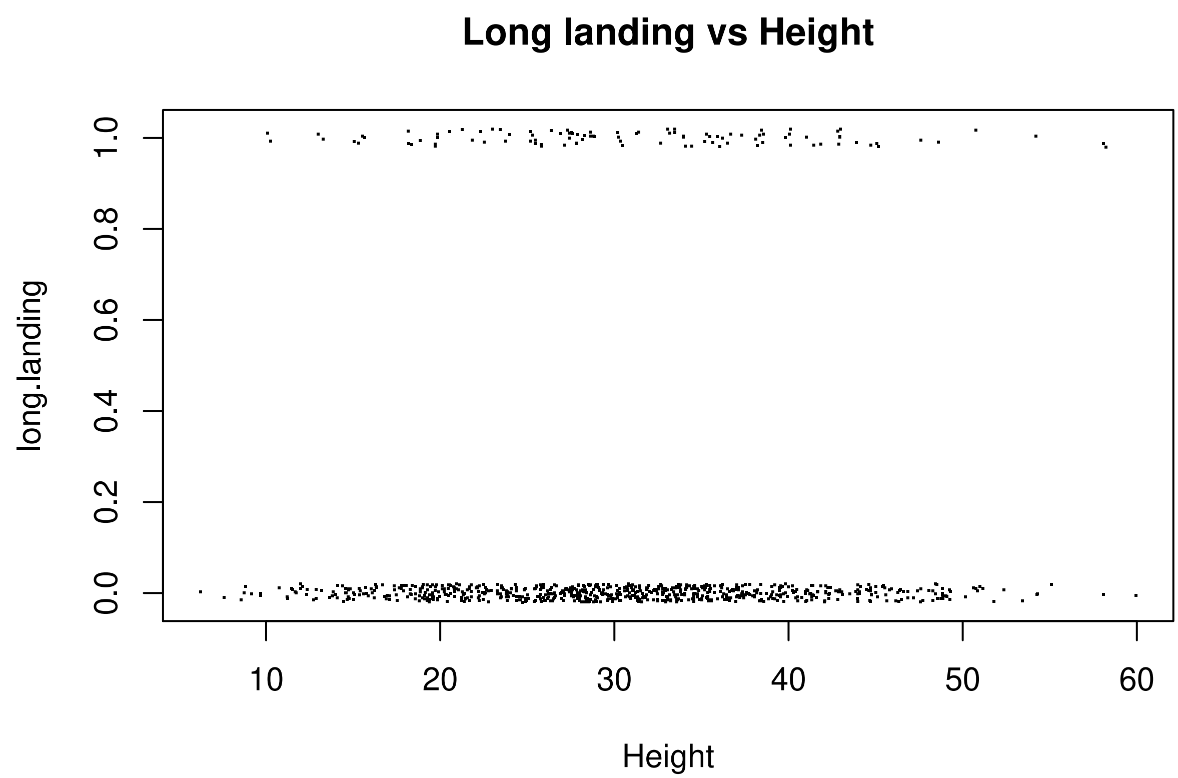
```
#long.landing vs Aircraft
library(ggplot2)
ggplot(Data_logit,aes(x=aircraft,fill=as.factor(long.landing)))+
  geom_bar(position="dodge")+ ggtitle("Long landing vs aircraft")
```




```
#long.landing vs Speed_air
plot(jitter(long.landing,0.1)~jitter(speed_air), Data_logit,
     xlab ="Speed_air", ylab ="long.landing", pch=".",
     main="Long landing vs Speed_air")
```




```
#long.landing vs Height
plot(jitter(long.landing,0.1)~jitter(height), Data_logit,
     xlab ="Height", ylab ="long.landing", pch = ".",
     main="Long landing vs Height")
```



Model for long landing

```
Ch_model <- glm(long.landing ~ aircraft + speed_air +
                  height,data = Data_logit,
                family = "binomial")
summary(Ch_model)
```

```
##
## Call:
## glm(formula = long.landing ~ aircraft + speed_air + height, family = "binomial",
##     data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.63624  -0.03742   0.00000   0.00237   2.21701
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -158.75148   37.50492  -4.233 2.31e-05 ***
## aircraft       7.66472    1.99774   3.837 0.000125 ***
## speed_air      1.43149    0.33639   4.255 2.09e-05 ***
## height         0.34900    0.09946   3.509 0.000450 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 281.373  on 202  degrees of freedom
## Residual deviance:  38.128  on 199  degrees of freedom
##   (628 observations deleted due to missingness)
## AIC: 46.128
##
## Number of Fisher Scoring iterations: 10
```

```r
odds.ratio(Ch_model)
```

```
##                     OR       2.5 %     97.5 %         p
## (Intercept)  1.1353e-69 1.7772e-109 0.0000e+00 2.308e-05 ***
## aircraft     2.1318e+03  8.9059e+01 2.7336e+05 0.0001247 ***
## speed_air    4.1849e+00  2.4735e+00 9.5231e+00 2.086e-05 ***
## height       1.4177e+00  1.2061e+00 1.7963e+00 0.0004501 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table showing the influence of the risk factors on long landing

```r
Table3 <-read_excel("Data-BANA7042.xls",sheet = 3)
datatable(Table3, options = list(
  searching = TRUE,
  pageLength = 3,
  scrollX = FALSE,
  scrollCollapse = FALSE
))
```

Show 3 entries   Search:

	Variables	Size of coefficient	Odds ratio	Direction of regression coefficient	P-value of coef.
1	Aircraft	7.6647	2131.8	Positive	0.000125
2	speed_air	1.4315	4.1849	Positive	0.0000209
3	Height	0.3409	1.4177	Positive	0.00045

Showing 1 to 3 of 3 entries   Previous 1 Next

Step 9: Identify important factors using "risky.landing"

Create binary responses

The binary variable "risky.landing" was created in Step 1 and is included in the clean dataset titled

"Data_logit".

risky.landing = 1 if distance > 3000; =0 otherwise

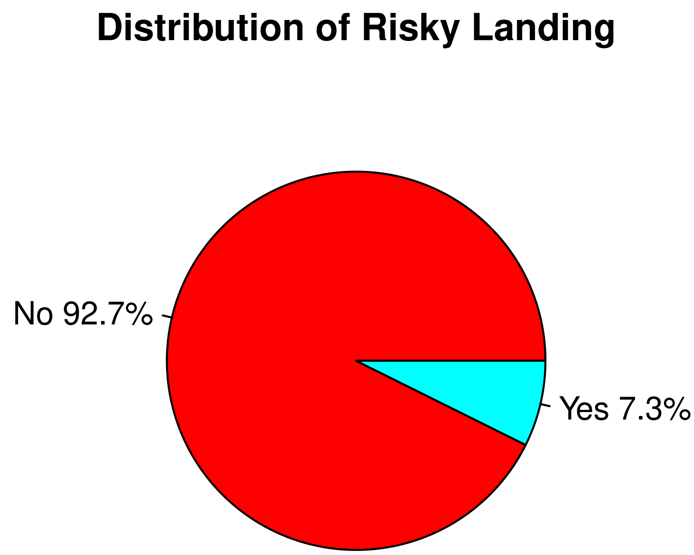
```
#Check the dataset
str(Data_logit)
```

```
## 'data.frame':    831 obs. of  9 variables:
##  $ aircraft     : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ 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 ...
##  $ long.landing : num  1 1 0 0 0 0 0 0 0 0 ...
##  $ risky.landing: num  1 0 0 0 0 0 0 0 0 0 ...
```

Distribution of risky.landing

A piechart is created to show the distribution of risky landings. No represents "0", while Yes represents "1". The chart shows that 7.3 percent of the observations/flights are risky landings (landing distance greater than 3000ft), while the remaining 92.7 percent of the flights are not risky landings (that is their landing distances are less than or equal to 3000 feet).

```
#piechart
xr=sum (Data_logit$risky.landing==0)
yr=sum (Data_logit$risky.landing==1)
slicesr <- c(xr,yr)
lblsr <- c("No", "Yes")
pctr <- round(slicesr/sum(slicesr)*100,1)
lblsr <- paste(lblsr, pctr)
lblsr <- paste(lblsr,"%",sep="")
pie(slicesr,labels = lblsr, col=rainbow(length(lblsr)),
    main="Distribution of Risky Landing")
```



Single-factor regression analysis for each of the potential risk factors

First, the original variables are considered and the response variable "risky.landing" is regressed on each of the X variables. The odds ratio of each of the model is obtained using the **odds.ratio()** function in the "questionr" package.

```
library("questionr")
lmod1r <- glm(risky.landing~aircraft, family=binomial, Data_logit)
summary(lmod1r)
odds.ratio(lmod1r)
lmod2r <- glm(risky.landing~duration, family=binomial, Data_logit)
summary(lmod2r)
odds.ratio(lmod2r)
lmod3r <- glm(risky.landing~no_pasg, family=binomial, Data_logit)
summary(lmod3r)
odds.ratio(lmod3r)
lmod4r <- glm(risky.landing~speed_ground, family=binomial, Data_logit)
summary(lmod4r)
odds.ratio(lmod4r)
lmod5r <- glm(risky.landing~speed_air, family=binomial, Data_logit)
summary(lmod5r)
odds.ratio(lmod5r)
lmod6r <- glm(risky.landing~height, family=binomial, Data_logit)
summary(lmod6r)
odds.ratio(lmod6r)
lmod7r <- glm(risky.landing~pitch, family=binomial, Data_logit)
summary(lmod7r)
```

```r
odds.ratio(lmod7r)
```

The factors are ranked from the most important to least important as shown in the table below. The most important potential risks factors are speed_air, speed_ground, and aircraft. This is based on their significance as shown by the p values and the size of their respective regression coefficients.

```r
Table4 <-read_excel("Data-BANA7042.xls",sheet = 4)
datatable(Table4, options = list(
  searching = TRUE,
  pageLength = 7,
  scrollX = FALSE,
  scrollCollapse = FALSE
))
```

Show 7 entries Search:

	Variables	Size of coefficient	Odds ratio	Direction of regression coefficient	P-value of coef.
1	Speed_air	0.8704	2.3879	Positive	0.00000373
2	Speed_ground	0.6142	1.8482	Positive	6.9e-8
3	Aircraft	1.0018	2.7231	Positive	0.000456
4	Pitch	0.3711	1.4493	Positive	0.1433
5	No_pasg	0.0254	0.9749	Negative	0.154
6	Duration	0.0012	0.9988	Negative	0.68
7	Height	0.0022	0.9978	Negative	0.871

Showing 1 to 7 of 7 entries Previous 1 Next

Regression of risky.landing on Standardized Predictor variables

Each of the X variables are standardized such that X'= {X-mean(X)}/sd(X).The mean of X' is 0 and its standard deviation is 1.The **scale()** function is used. The aircraft variable isn't standardized because it's a factor variable recoded into a dummy variable 0/1. The aircraft variable could lose it's interpretation if standardized.

```r
#Standardize each X variable except aircraft
Data_logit.norm <- as.data.frame(scale(Data_logit[,-c(1,8,9)]))
Data_logit.n <- cbind(Data_logit$aircraft, Data_logit.norm,
                      Data_logit$long.landing,Data_logit$risky.landing)
summary(Data_logit.n)
```

```
##  Data_logit$aircraft    duration             no_pasg              speed_ground
##  Min.   :0.0000      Min.   :-2.33354   Min.   :-4.145514   Min.   :-2.45353
##  1st Qu.:0.0000      1st Qu.:-0.72687   1st Qu.:-0.674829   1st Qu.:-0.71220
##  Median :0.0000      Median :-0.01016   Median :-0.007389   Median : 0.01341
##  Mean   :0.4657      Mean   : 0.00000   Mean   : 0.000000   Mean   : 0.00000
##  3rd Qu.:1.0000      3rd Qu.: 0.72156   3rd Qu.: 0.660050   3rd Qu.: 0.65999
```

```
##  Max.   :1.0000      Max.   : 3.11988   Max.   : 3.596784   Max.   : 2.84174
##                      NA's   :50
##    speed_air           height             pitch
##  Min.   :-1.3847   Min.   :-2.47632   Min.   :-3.26772
##  1st Qu.:-0.7452   1st Qu.:-0.70804   1st Qu.:-0.69259
##  Median :-0.2430   Median :-0.02972   Median :-0.00783
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.6029   3rd Qu.: 0.66905   3rd Qu.: 0.69323
##  Max.   : 3.0223   Max.   : 3.01366   Max.   : 3.64933
##  NA's   :628
##  Data_logit$long.landing Data_logit$risky.landing
##  Min.   :0.0000          Min.   :0.00000
##  1st Qu.:0.0000          1st Qu.:0.00000
##  Median :0.0000          Median :0.00000
##  Mean   :0.1239          Mean   :0.07341
##  3rd Qu.:0.0000          3rd Qu.:0.00000
##  Max.   :1.0000          Max.   :1.00000
##
```

Then, risky.landing is regressed on each of the scaled potential risk factors.

```
lmod1.nr <- glm(risky.landing~aircraft, family=binomial, Data_logit.n)
summary(lmod1.nr)
odds.ratio(lmod1.nr)
lmod2.nr <- glm(risky.landing~duration, family=binomial, Data_logit.n)
summary(lmod2.nr)
odds.ratio(lmod2.nr)
lmod3.nr <- glm(risky.landing~no_pasg, family=binomial, Data_logit.n)
summary(lmod3.nr)
odds.ratio(lmod3.nr)
lmod4.nr <- glm(risky.landing~speed_ground, family=binomial, Data_logit.n)
summary(lmod4.nr)
odds.ratio(lmod4.nr)
lmod5.nr <- glm(risky.landing~speed_air, family=binomial, Data_logit.n)
summary(lmod5.nr)
odds.ratio(lmod5.nr)
lmod6.nr <- glm(risky.landing~height, family=binomial, Data_logit.n)
summary(lmod6.nr)
odds.ratio(lmod6.nr)
lmod7.nr <- glm(risky.landing~pitch, family=binomial, Data_logit.n)
summary(lmod7.nr)
odds.ratio(lmod7.nr)
```

The ranking of the factors based on the regression of standardized predictor variables show that speed_ground, speed_air, and aircraft are statistically significant and are important factors for analysis.

```
Table5 <-read_excel("Data-BANA7042.xls",sheet = 5)
datatable(Table5, options = list(
  searching = TRUE,
  pageLength = 7,
  scrollX = FALSE,
  scrollCollapse = FALSE
))
```

Show 7 entries Search:

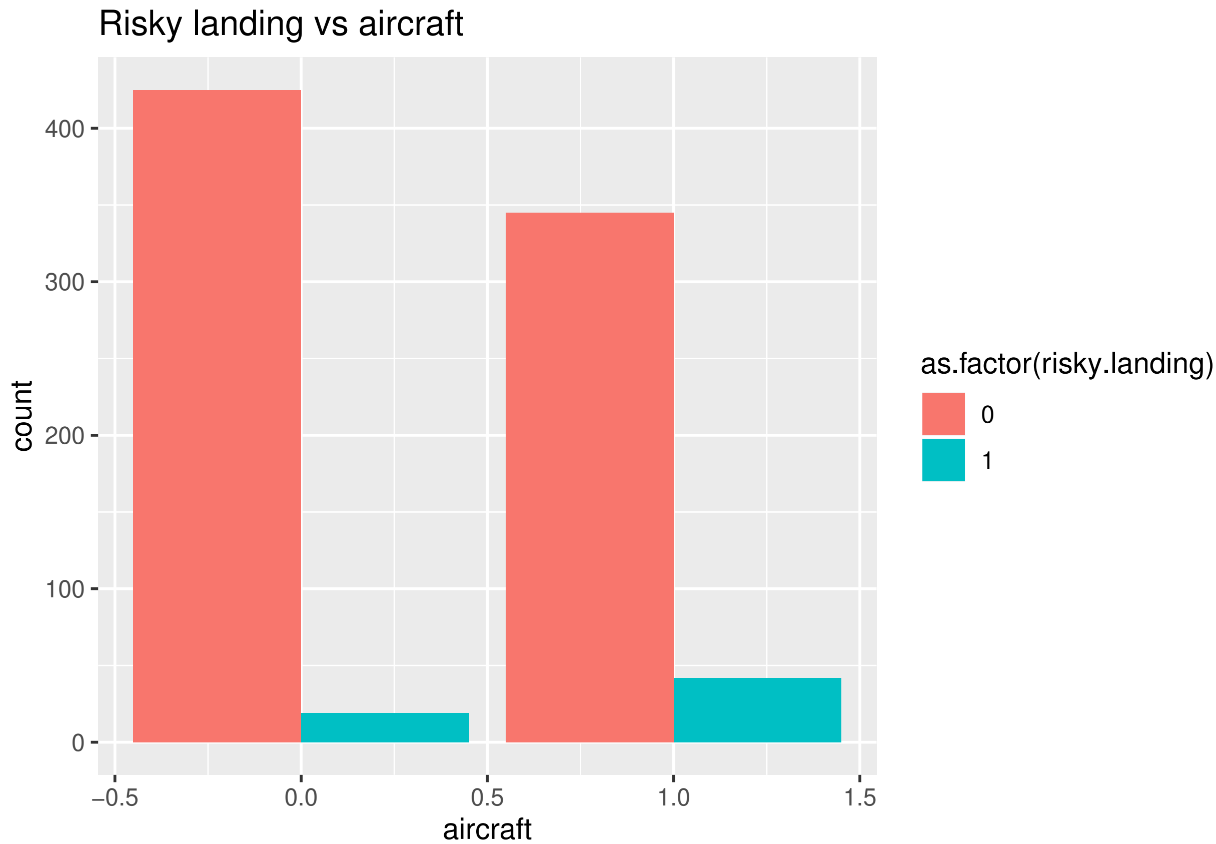
	Variables	Size of coefficient	Odds ratio	Direction of regression coefficient	P-value of coef.
1	Speed_ground	11.508	99489	Positive	6.9e-8
2	Speed_air	8.475	4790.9	Positive	0.00000373
3	Aircraft	1.0018	2.7231	Positive	0.000456
4	Pitch	0.1954	1.2158	Positive	0.1433
5	No_pasg	0.1901	0.8269	Negative	0.154
6	Duration	0.05569	0.945831	Negative	0.68
7	Height	0.02171	0.9785	Negative	0.871

Showing 1 to 7 of 7 entries Previous 1 Next

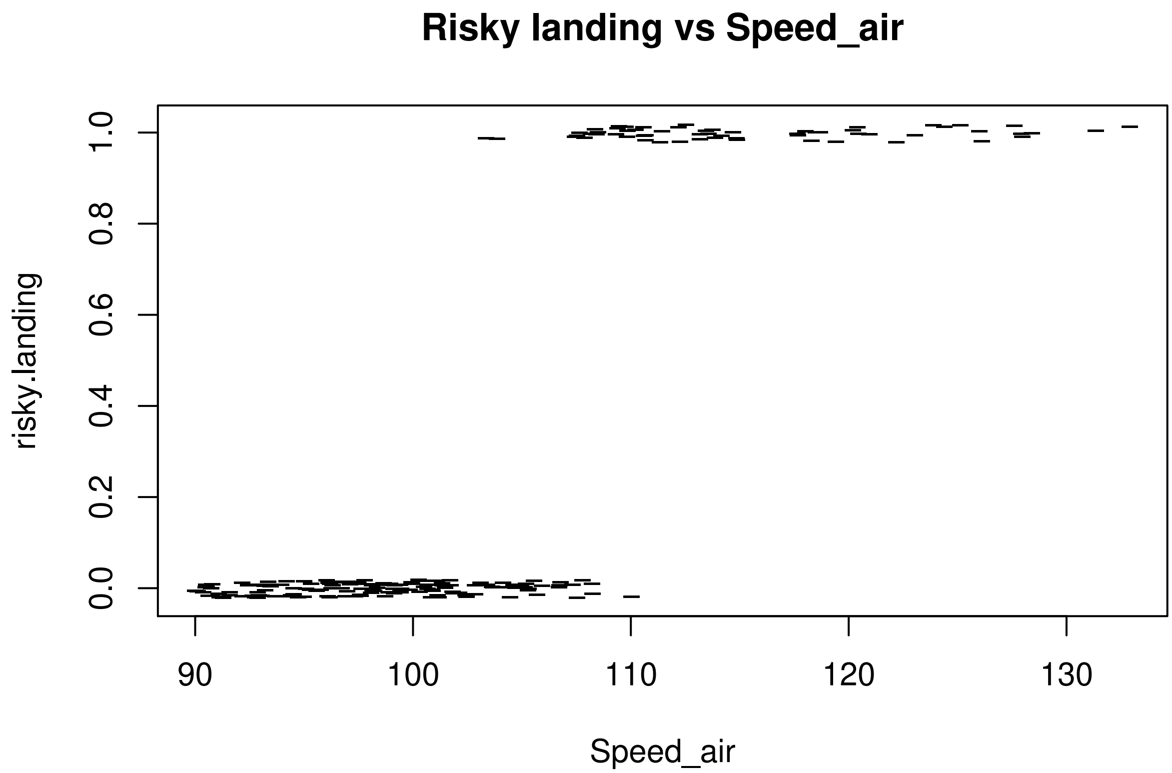
Visualize the association of the significant factors with "risky.landing"

The associations are visualized using scatterplot and barchart. The plot show that there's a strong association between the following pairs: **risky.landing and speed_air** ; **risky.landing and speed_ground** and **risky.landing aircraft type**. The plots also suggest that as the air speed and the ground speed of the aircraft increases, the probability of being a risky landing increases. In addition, there are more Boeing aircrafts that are risky landings compared to Airbus aircrafts.

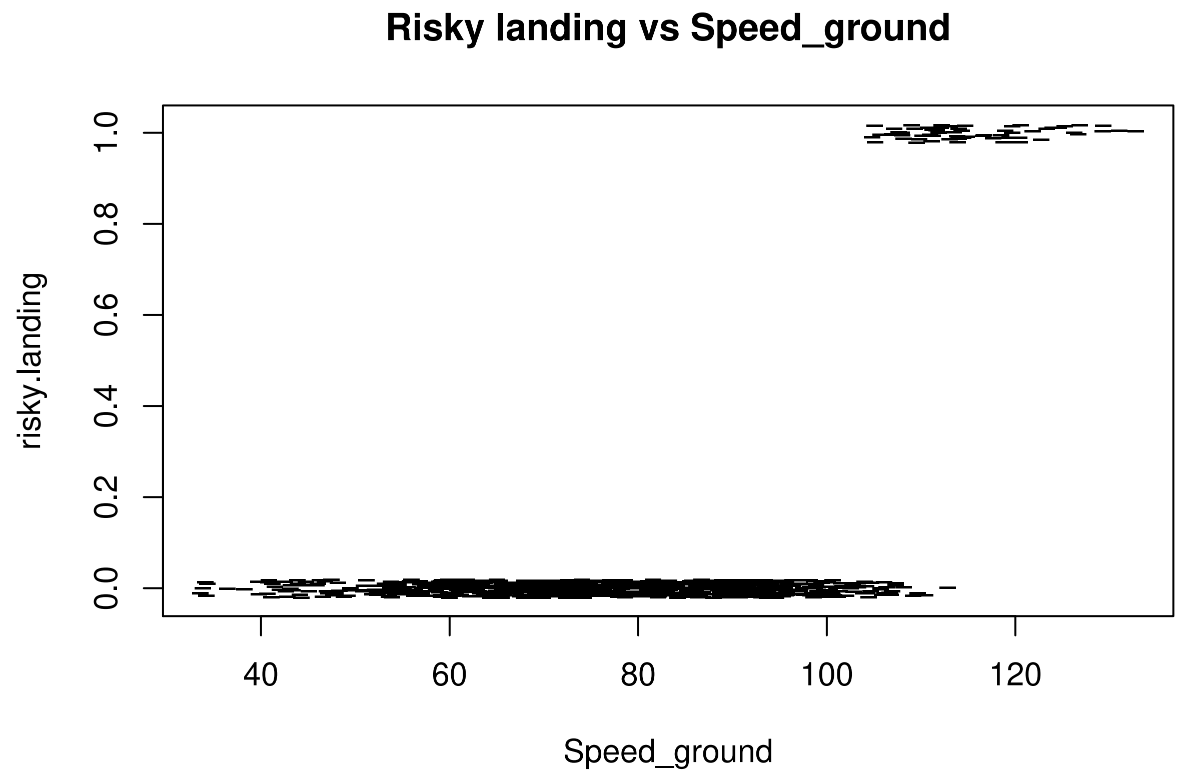
```
#significant factors : Aircraft, Speed_air, speed_ground
#risky.landing vs Aircraft
ggplot(Data_logit,aes(x=aircraft,fill=as.factor(risky.landing)))+
  geom_bar(position="dodge")+ ggtitle("Risky landing vs aircraft")
```




```
#risky.landing vs Speed_air
plot(jitter(risky.landing,0.1)~jitter(speed_air), Data_logit,
     xlab ="Speed_air", ylab ="risky.landing", pch="-",
     main="Risky landing vs Speed_air")
```




```
#risky.landing vs Speed_ground
plot(jitter(risky.landing,0.1)~jitter(speed_ground), Data_logit,
     xlab ="Speed_ground", ylab ="risky.landing", pch="-",
     main="Risky landing vs Speed_ground")
```



Full model

In part I, Step 16, it was indicated that there's a strong collinearity between speed_air and speed_ground. Though, there are both highly associated with risky.landing. To select one of the variables to include in the full model, we look at the individual effect of both variables on risky.landing as shown below.

```
#marginal model
summary(glm(risky.landing~speed_ground, family=binomial, Data_logit))
```

```
##
## Call:
## glm(formula = risky.landing ~ speed_ground, family = binomial,
##     data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.53709  -0.00383  -0.00009   0.00000   1.95417
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -66.1243    12.2025  -5.419  6.0e-08 ***
## speed_ground   0.6142     0.1139   5.394  6.9e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 436.043  on 830  degrees of freedom
## Residual deviance:  58.931  on 829  degrees of freedom
## AIC: 62.931
##
## Number of Fisher Scoring iterations: 11
```

```
summary(glm(risky.landing~speed_air, family=binomial, Data_logit))
```

```
##
## Call:
## glm(formula = risky.landing ~ speed_air, family = binomial, data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.15041  -0.05679  -0.00616   0.00102   2.69736
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -93.5700    20.2180  -4.628 3.69e-06 ***
## speed_air     0.8704     0.1882   4.626 3.73e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 248.18  on 202  degrees of freedom
## Residual deviance:  44.58  on 201  degrees of freedom
##   (628 observations deleted due to missingness)
## AIC: 48.58
##
## Number of Fisher Scoring iterations: 9
```

Given that Speed_air has a higher coefficient size and a lower AIC value for it's marginal regression model, we could consider including it in the full model. But, the variable has 628 missing values. Hence, we would include speed_ground instead. First we look at the model with all variables, then the model without speed_air. Then, we create a "full" model based on the significant factors.

```
#Model with all of the predictor variables
lmod.allr <- glm (risky.landing~aircraft+speed_air+speed_ground+
                   pitch+height+duration+no_pasg,family=binomial,Data_logit)
summary(lmod.allr)
```

```
##
## Call:
## glm(formula = risky.landing ~ aircraft + speed_air + speed_ground +
##     pitch + height + duration + no_pasg, family = binomial, data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.95653  -0.00291  -0.00017   0.00001   2.23576
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -149.41931   48.42462  -3.086  0.00203 **
## aircraft        7.33037    3.00197   2.442  0.01461 *
```

```
## speed_air       1.61745    0.65439   2.472  0.01345 *
## speed_ground   -0.16366    0.49825  -0.328  0.74256
## pitch          -1.31605    1.42985  -0.920  0.35736
## height          0.04535    0.05768   0.786  0.43179
## duration        0.00198    0.01587   0.125  0.90070
## no_pasg        -0.12011    0.09589  -1.253  0.21034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 240.724  on 194  degrees of freedom
## Residual deviance:  22.144  on 187  degrees of freedom
##   (636 observations deleted due to missingness)
## AIC: 38.144
##
## Number of Fisher Scoring iterations: 10
```

```
#Model with all predictor variables except speed_air
lmod.all2r <- glm (risky.landing~aircraft+speed_ground+
                    pitch+height+duration+no_pasg,family=binomial,Data_logit)
summary(lmod.all2r)
```

```
##
## Call:
## glm(formula = risky.landing ~ aircraft + speed_ground + pitch +
##     height + duration + no_pasg, family = binomial, data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.44763  -0.00011   0.00000   0.00000   1.85688
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.022e+02  2.811e+01  -3.635 0.000278 ***
## aircraft      4.406e+00  1.562e+00   2.821 0.004783 **
## speed_ground  9.366e-01  2.476e-01   3.782 0.000155 ***
## pitch         6.083e-01  8.000e-01   0.760 0.447089
## height        4.214e-02  4.618e-02   0.913 0.361502
## duration      7.386e-04  1.214e-02   0.061 0.951498
## no_pasg      -8.590e-02  6.011e-02  -1.429 0.152956
## ---
## 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:  36.372  on 774  degrees of freedom
##   (50 observations deleted due to missingness)
## AIC: 50.372
##
## Number of Fisher Scoring iterations: 12
```

```
#Full Model with only significant variables
lmod.fullr <- glm(risky.landing~aircraft+speed_ground
```

```
                  ,family=binomial,Data_logit)
summary(lmod.fullr)
```

```
##
## Call:
## glm(formula = risky.landing ~ aircraft + speed_ground, family = binomial,
##     data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.24398  -0.00011   0.00000   0.00000   1.61021
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -102.0772    24.7751  -4.120 3.79e-05 ***
## aircraft        4.0190     1.2494   3.217   0.0013 **
## speed_ground    0.9263     0.2248   4.121 3.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 436.043  on 830  degrees of freedom
## Residual deviance:  40.097  on 828  degrees of freedom
## AIC: 46.097
##
## Number of Fisher Scoring iterations: 12
```

The full model shows that aircraft and speed_ground are significant factors that impact risky landings.

Forward Variable Selection Using AIC

The model shows some consistency with the marginal regression models. The model shows that aircraft and speed_air are significant factors in risky landings. However, the model is contrary to the marginal regression model because speed_ground is not significant as shown below but it's marginal effect on risky landing is significant.

```
#Step using AIC
model.0r <- glm(risky.landing ~ aircraft + duration + no_pasg +
                  speed_ground + speed_air + height + pitch,data = Data_logit,
                family = "binomial")

model.0_AICr <- step(model.0r, trace = 0, direction = "forward")
summary(model.0_AICr)
```

```
##
## Call:
## glm(formula = risky.landing ~ aircraft + duration + no_pasg +
##     speed_ground + speed_air + height + pitch, family = "binomial",
##     data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.95653  -0.00291  -0.00017   0.00001   2.23576
```

```
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -149.41931   48.42462  -3.086  0.00203 **
## aircraft        7.33037    3.00197   2.442  0.01461 *
## duration        0.00198    0.01587   0.125  0.90070
## no_pasg        -0.12011    0.09589  -1.253  0.21034
## speed_ground   -0.16366    0.49825  -0.328  0.74256
## speed_air       1.61745    0.65439   2.472  0.01345 *
## height          0.04535    0.05768   0.786  0.43179
## pitch          -1.31605    1.42985  -0.920  0.35736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 240.724  on 194  degrees of freedom
## Residual deviance:  22.144  on 187  degrees of freedom
##   (636 observations deleted due to missingness)
## AIC: 38.144
##
## Number of Fisher Scoring iterations: 10
```

Forward Variable Selection Using BIC

The model is consistent with the model above. Both models show that aircraft and speed_air are significant risk factors and influence risky landings.

```r
#Step using BIC
model.0_BICr <- step(model.0r, trace = 0, direction = "forward", criterion = "BIC")
summary(model.0_BICr)
```

```
##
## Call:
## glm(formula = risky.landing ~ aircraft + duration + no_pasg +
##     speed_ground + speed_air + height + pitch, family = "binomial",
##     data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.95653  -0.00291  -0.00017   0.00001   2.23576
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -149.41931   48.42462  -3.086  0.00203 **
## aircraft        7.33037    3.00197   2.442  0.01461 *
## duration        0.00198    0.01587   0.125  0.90070
## no_pasg        -0.12011    0.09589  -1.253  0.21034
## speed_ground   -0.16366    0.49825  -0.328  0.74256
## speed_air       1.61745    0.65439   2.472  0.01345 *
## height          0.04535    0.05768   0.786  0.43179
## pitch          -1.31605    1.42985  -0.920  0.35736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 240.724  on 194  degrees of freedom
## Residual deviance:  22.144  on 187  degrees of freedom
##   (636 observations deleted due to missingness)
## AIC: 38.144
##
## Number of Fisher Scoring iterations: 10
```

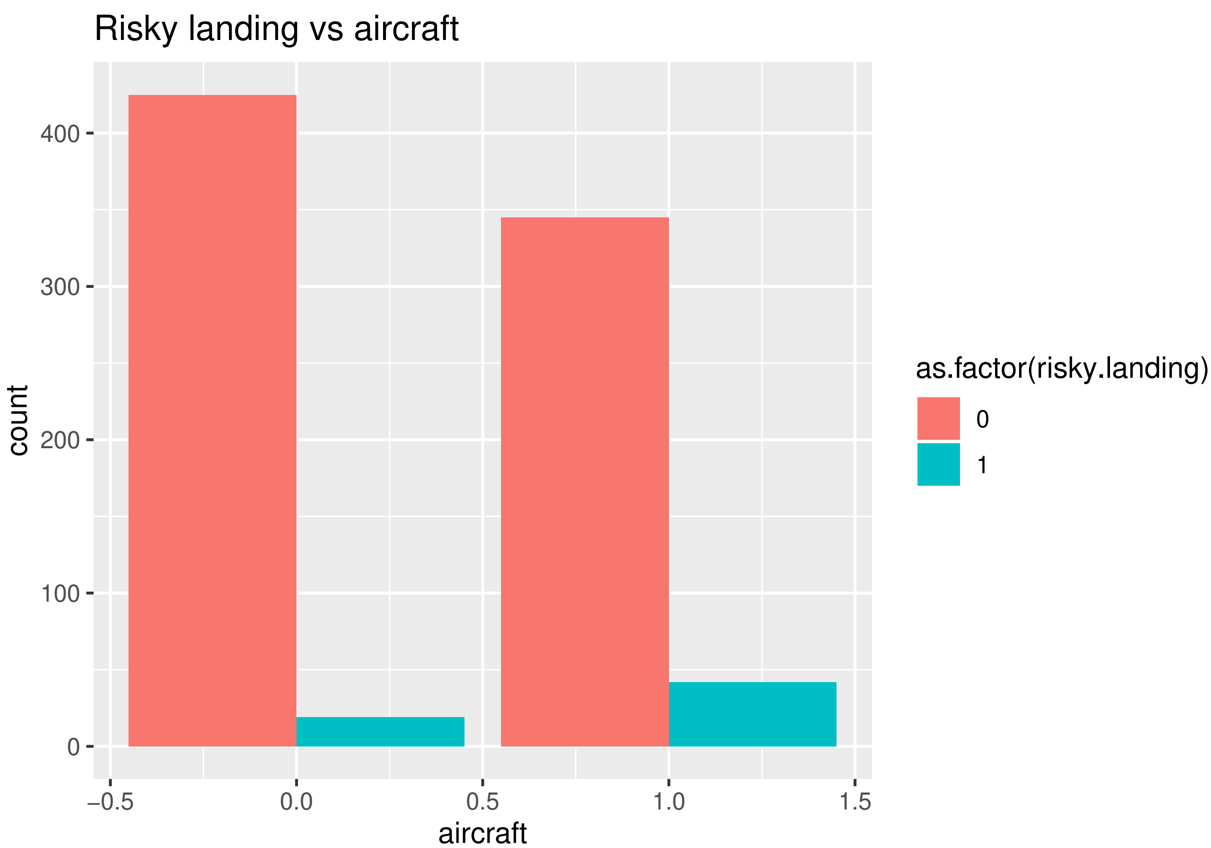
Step 10: Risk factors for risky landings and their influence

Executive summary

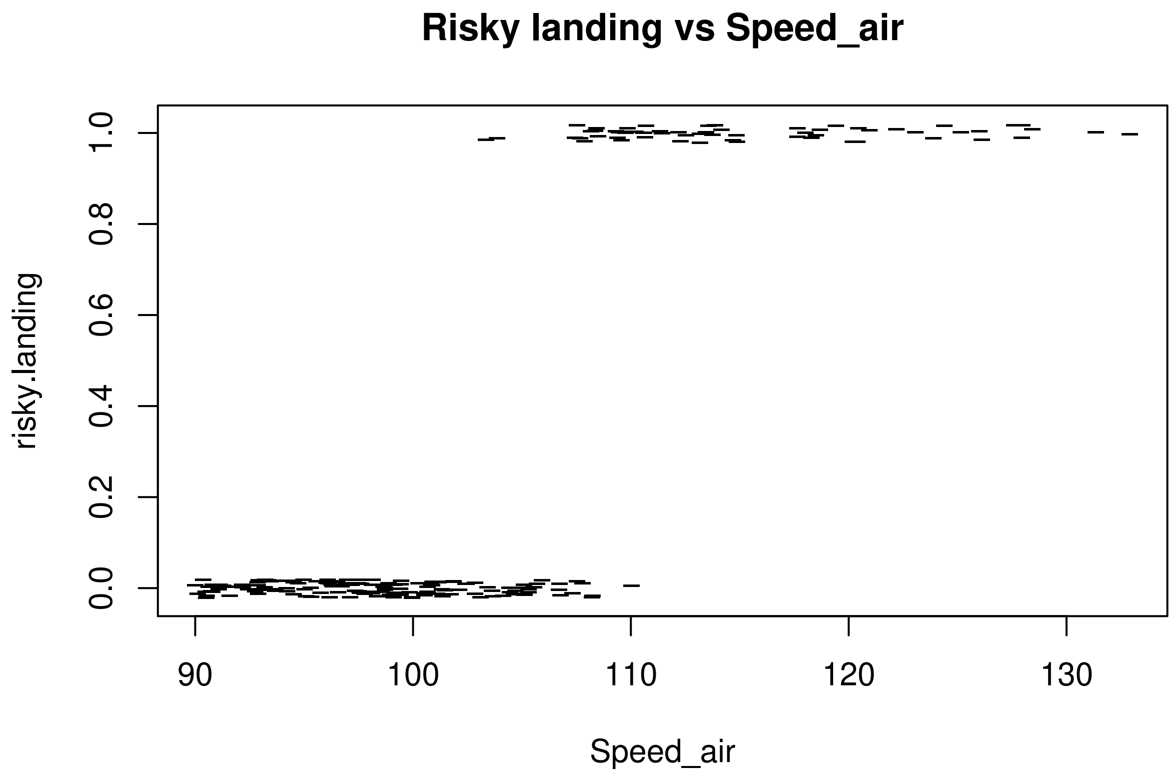
- Aircraft type and the speed of the flight in the air are the most important risk factors for risky landings. An increase in these variables is associated with an increase in the probability of risky landing.
- Boeing aircrafts have more risky landings than Airbus aircrafts.
- If the aircraft make is Boeing the chances of risky landing are higher than for Airbus aircraft.
- An increase in the air speed of an aircraft is associated with an increase in the probability of the aircraft being a risky landing. For a one-unit increase in the air speed of the aircraft, we expect a 1.224 increase in the log-odds of risky landing.
- The variable "speed_air" has 628 missing observations. More observations in this regard may further strengthen my analysis.

Association between the significant factors and risky landings

```r
#risky.landing vs Aircraft
ggplot(Data_logit,aes(x=aircraft,fill=as.factor(risky.landing)))+
  geom_bar(position="dodge")+ ggtitle("Risky landing vs aircraft")
```




```
#risky.landing vs Speed_air
plot(jitter(risky.landing,0.1)~jitter(speed_air), Data_logit,
     xlab ="Speed_air", ylab ="risky.landing", pch="-",
     main="Risky landing vs Speed_air")
```



Model for risky landing

```
best_modelr <- glm(risky.landing ~ aircraft + speed_air ,
                   data = Data_logit, family = "binomial")
summary(best_modelr)
```

```
##
## Call:
## glm(formula = risky.landing ~ aircraft + speed_air, family = "binomial",
##     data = Data_logit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.67290  -0.00812  -0.00068   0.00005   2.47739
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -134.0859    33.3811  -4.017 5.90e-05 ***
## aircraft       4.5648     1.5081   3.027  0.00247 **
## speed_air      1.2240     0.3052   4.010 6.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 248.180  on 202  degrees of freedom
## Residual deviance:  26.296  on 200  degrees of freedom
```

```
##   (628 observations deleted due to missingness)
## AIC: 32.296
##
## Number of Fisher Scoring iterations: 9
```

```r
#odds.ratio(Ch_modelr)
```

Table showing the influence of the risk factors on risky landing

```r
Table6 <-read_excel("Data-BANA7042.xls",sheet = 6)
datatable(Table6, options = list(
  searching = TRUE,
  pageLength = 2,
  scrollX = FALSE,
  scrollCollapse = FALSE
))
```

Show 2 entries Search:

	Variables	Size of coefficient	Odds ratio	Direction of regression coefficient	P-value of coef.
1	Aircraft	4.5648	96.044	Positive	0.00247
2	speed_air	1.224	3.4006	Positive	0.0000607

Showing 1 to 2 of 2 entries Previous 1 Next

Step 11: Difference between the two models built for "long.landing" and "risky.landing"

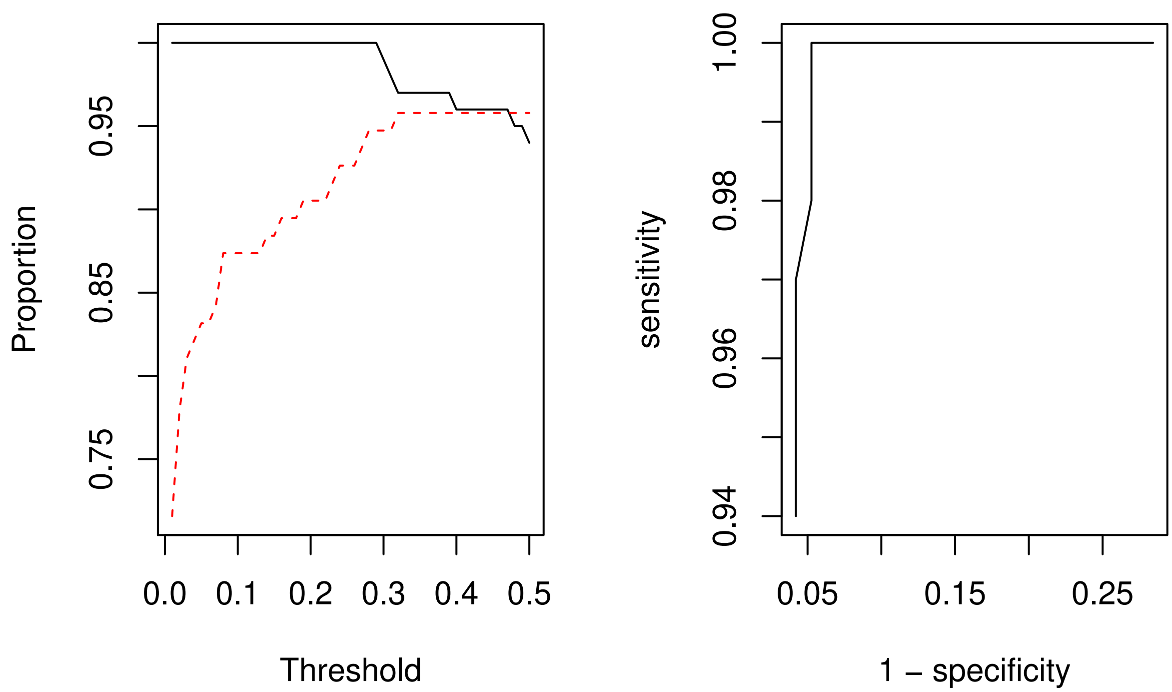
- Three significant risk factors - aircraft, speed_air and height- were highlighted in the model for long landing, while the model for risky landing has two significant risk factors - aircraft and speed_air. The variable height loses its significance in predicting risky landing.
- The effect of aircraft type is higher in the model for long landing, sugesting that the effect of aircraft type is higher in cases of long landings than risky landings. However, it may be difficult to delineate the difference in effect because the two binary variables "long.landing" and "risky.landing" overlap. Some flights could be both long landings and risky landings.
- The model for long landing has a lower AIC value compared to the model for risky landing.

Step 12: ROC Curve (sensitivity vs 1-specificity)

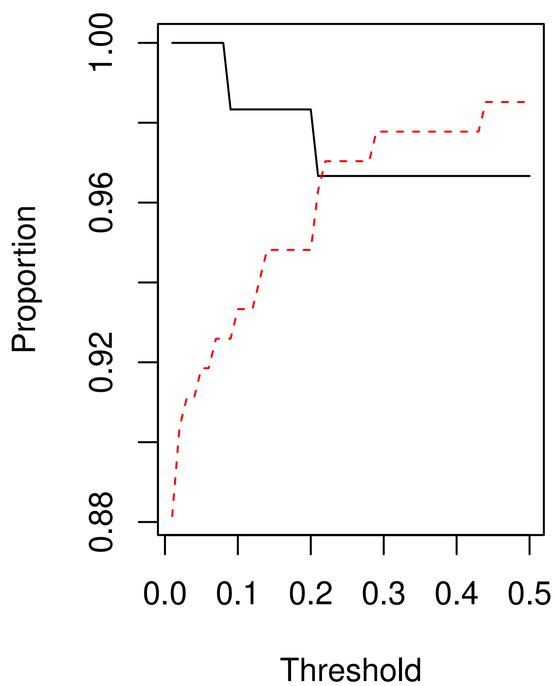
First, we draw the ROC curve for the model with each binary variable "long.landing" and "risky.landing". Then, the two curves are put in the same plot. The plots show the trade-off between specificity and sensitivity. As sensitivity increases, specificity decreases and vice versa. To evaluate the two models predictive power, we observed the area under the curve (AUC). The AUC for both models are high, suggesting good predictive power. However, the AUC for risky landing is slightly higher compared to long landing's AUC.This indicates that the model for risky landing has a slightly higher predictive power.

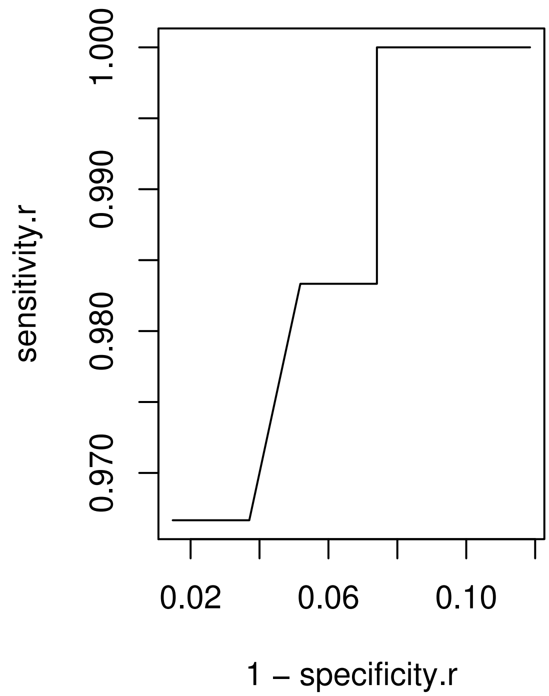
Note: In the plot with two curves, blue=risky landing model and red=long landing model.

```
#ROC curve for model for long landing
thresh <- seq(0.01,0.5,0.01)
predprob_l <- predict(lmod.all, type = "response")
predprob_r <- predict(lmod.allr, type = "response")
long.landing_l <- long.landing[!is.na(Data_logit$speed_air) & !is.na(Data_logit$duration)]

sensitivity <- specificity<-rep(NA,length(thresh))

for(j in seq(along=thresh)) {
  pp<-ifelse(predprob_l < thresh[j], "no", "yes")
  xx<-xtabs(~long.landing_l + pp, Data_logit)
  specificity[j] <- xx[1,1]/(xx[1,1] + xx[1,2])
  sensitivity[j] <- xx[2,2]/(xx[2,1] + xx[2,2])
}

par(mfrow= c(1,2))
matplot(thresh,cbind(sensitivity, specificity), type="l",
        xlab="Threshold", ylab="Proportion", lty=1:2)
plot(1-specificity, sensitivity, type="l");abline(0, 1, lty=2)
```

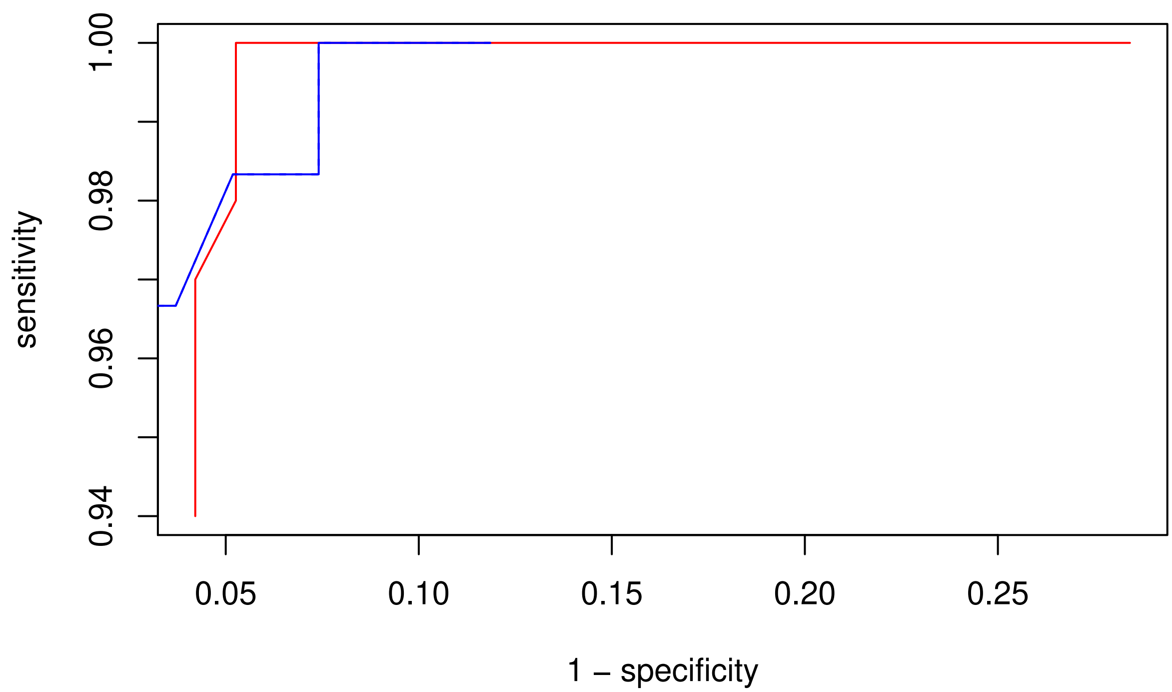



```
#ROC curve for model for risky landing
thresh <- seq(0.01, 0.5, 0.01)
predprob_l <- predict(lmod.all, type = "response")
predprob_r <- predict(lmod.allr, type = "response")
risky.landing_r <- risky.landing[!is.na(Data_logit$speed_air) &
                                   !is.na(Data_logit$duration)]

sensitivity.r <- specificity.r <- rep(NA,length(thresh))

for(j in seq(along=thresh)) {
  pp <- ifelse(predprob_r < thresh[j], "no", "yes")
  xx <- xtabs(~risky.landing_r + pp, Data_logit)
  specificity.r[j] <- xx[1, 1]/(xx[1, 1] + xx[1, 2])
  sensitivity.r[j] <- xx[2, 2]/(xx[2, 1] + xx[2, 2])
}
par(mfrow=c(1,2))
matplot(thresh,cbind(sensitivity.r,specificity.r),type="l",
        xlab="Threshold",ylab="Proportion",lty=1:2)
plot(1-specificity.r,sensitivity.r,type="l");abline(0,1,lty=2)
```






```
#Two curves on a plot
plot(1-specificity,sensitivity, type="l", col="red")
points(1-specificity.r,sensitivity.r,type="l",col="blue")
lines(1-specificity.r,sensitivity.r, col="blue",lty=2)
```



After plotting the curves, the area under the curves are examined using the the "ROCR" package. The curve in blue is that of the risky landing model, while long landing is in red.

```
#AUC
library(ROCR)
#risky landing
predl<-prediction(predprob_l,long.landing_l)
perfl<-performance(predl,"tpr", "fpr")
aucl<- performance(predl,"auc")

#risky landing
predr<-prediction(predprob_r,risky.landing_r)
perfr<-performance(predr,"tpr", "fpr")
aucr<- performance(predr,"auc")

paste("AUC for long landing is ", aucl@y.values)
```

```
## [1] "AUC for long landing is  0.996"
```

```
paste("AUC for risky landing is ", aucr@y.values)
```

```
## [1] "AUC for risky landing is  0.997283950617284"
```

Step 13: Probability Prediction

For a given set of information: (Boeing, duration = 200, no_pasg = 80, speed_ground = 115, speed_air = 120, height = 40, pitch = 4), we are asked to predict the probability that a commericial airplane passing over the threshold of the runway could be a long landing and a risky landing.The 95% confidence interval is also calculated.

Given the information on other variables, the predicted probability that the commercial airplane is either a long landing or risky landing is very high. The probability of long landing is 1, while for risky landing the probability is 0.999999998776229 which is very close to 1.

The 95% confidence interval for the predicted probability for both long landing and risky landing are very narrow, with both the upper and lower bounds equal to or approximately 1.

```
#probability of being a long landing
given_data <- data.frame(aircraft = character(), duration = numeric(),
                         no_pasg = numeric(), speed_ground = numeric(),
                         speed_air = numeric(), height = numeric(), pitch =
                           numeric(), stringsAsFactors = FALSE)

#recall for aircraft boeing =1
given_data <- rbind(given_data, list(1, 200, 80, 115, 120, 40, 4))
colnames(given_data) <- c("aircraft", "duration", "no_pasg",
                          "speed_ground", "speed_air", "height", "pitch")

library(faraway)
pred_l <- predict(model.0_AIC, given_data, type = "response", se.fit = T)
paste("Predicted probability for long landing is ", pred_l$fit[["1"]])
```

```
## [1] "Predicted probability for long landing is  1"
```

```
paste("The standard error for the predicted probability for long landing is ",
      pred_l$se.fit[["1"]])
```

```
## [1] "The standard error for the predicted probability for long landing is  2.71530502537882e-15"
```

```
#confidence interval for a long landing
conf_interval <- round(ilogit(c(pred_l$fit[["1"]] - 1.96*pred_l$se.fit[["1"]],
                                pred_l$fit[["1"]] + 1.96*pred_l$se.fit[["1"]])))
conf_interval
```

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

```
#probability of being a risky  landing
pred_r <- predict(lmod.allr, given_data, type = "response", se.fit = T)
paste("Predicted probability for risky landing is ", pred_r$fit[["1"]])
```

```
## [1] "Predicted probability for risky landing is  0.999999998776229"
```

```
paste("The standard error for the predicted probability for risky landing is ",
      pred_r$se.fit[["1"]])
```

```
## [1] "The standard error for the predicted probability for risky landing is  8.63858425577143e-09"
```

```
#confidence interval for a risky landing
conf_intervalr <- round(ilogit(c(pred_r$fit[["1"]] - 1.96*pred_r$se.fit[["1"]],
                                 pred_r$fit[["1"]] + 1.96*pred_r$se.fit[["1"]])))
conf_intervalr
```

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

Step 14: Compare models with different link functions

Here, the binary response "risky landing" is fitted on the identified risk factors in Steps 9-10 (speed_air and aircraft) using three models -probit model, hazard model with complementary log-log link, and the logit model. The performance of the models are then compared.

```
logit_model <- glm( risky.landing ~ aircraft + speed_air,
  data = Data_logit, family = binomial(link = logit))

probit_model <- glm(risky.landing ~ aircraft + speed_air,
  data = Data_logit, family = binomial(link = probit))

cloglog_model <- glm(risky.landing ~ aircraft + speed_air,
  data = Data_logit, family = binomial(link = cloglog))
```

The models are compared based on their coefficients, predicted values, AIC, and residual deviance. The hazard model with the clog-log link has the model with the least deviance and AIC. This model also has coefficients that are closer to the coefficeints of the logistic regression model than that of the probit model. However, the probit model gives predicted values, AIC value, and residual deviance that are closer to the logistic model compared to the hazard model.

Compare coefficients

```
round(coef(logit_model),3)
```

```
## (Intercept)    aircraft   speed_air
##    -134.086       4.565       1.224
```

```
round(coef(probit_model),3)
```

```
## (Intercept)    aircraft   speed_air
##     -74.225       2.645       0.677
```

```
round(coef(cloglog_model),3)
```

```
## (Intercept)    aircraft   speed_air
##    -103.681       3.261       0.942
```

Compare predicted values

```
predval <- sapply(list(logit_model,probit_model,cloglog_model),fitted)
colnames(predval) <- c("logit","probit","clog-log")
datatable(round(predval[1:10,],3), options = list(
  searching = TRUE,
  pageLength = 10,
  scrollX = FALSE,
  scrollCollapse = FALSE
))
```

Show 10 entries Search:

	logit	probit	clog-log
1	0.987	0.993	1
2	0.026	0.028	0.03
12	0	0	0
13	0	0	0
18	0.084	0.11	0.076
19	1	1	1
22	0	0	0
24	0	0	0.001
26	0	0	0
29	1	1	1

Showing 1 to 10 of 10 entries Previous 1 Next

Compare AIC

```
round(AIC(logit_model),3)
```

```
## [1] 32.296
```

```
round(AIC(probit_model),3)
```

```
## [1] 32.138
```

```
round(AIC(cloglog_model),3)
```

```
## [1] 30.363
```

Compare Residual Deviance

```
round(deviance(logit_model),3)
```

```
## [1] 26.296
```

```
round(deviance(probit_model),3)
```

```
## [1] 26.138
```

```
round(deviance(cloglog_model),3)
```

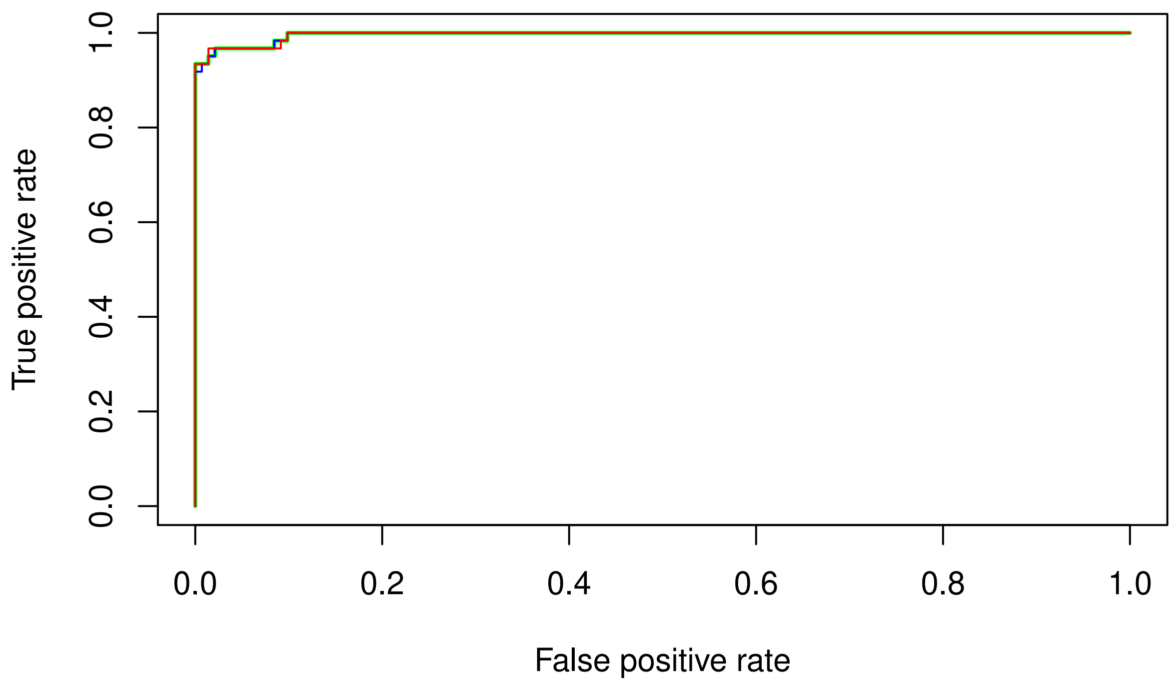
```
## [1] 24.363
```

Step 15: Compare the three models by showing their ROC curves

The three models that were fitted in Step 14 are compared using their ROC curves. The plot shows the trade-off between specificity and sensitivity. Note: In the plot with three ROC curves, the logit model curve is in green, probit is blue, and the hazard model with clog-log link is in color red.

It is seen that the curves overlap. To evaluate the three models predictive power, the area under the curve (AUC) is examined. The AUC for all the three models are high and very close, suggesting good predictive power. However, the AUC for the logit model and the hazard model with clog-log link are identical and slightly higher than the AUC of the probit model.

```
pred_logit <- predict(logit_model, type = "response")
pred_probit <- predict(probit_model, type = "response")
pred_cloglog <- predict(cloglog_model, type = "response")
risky.landing_r <- risky.landing[!is.na(Data_logit$speed_air)]

#logit model
predlogit<-prediction(pred_logit,risky.landing_r)
perflogit<-performance(predlogit,"tpr", "fpr")
aucllogit<- performance(predlogit,"auc")

#probit model
predprobit<-prediction(pred_probit,risky.landing_r)
perfprobit<-performance(predprobit,"tpr", "fpr")
auclprobit<- performance(predprobit,"auc")

#hazard model, with clog-log
predcloglog<-prediction(pred_cloglog,risky.landing_r)
perfcloglog<-performance(predcloglog,"tpr", "fpr")
auclcloglog<- performance(predcloglog,"auc")

#combine plots
plot(perflogit, lwd=2, col = "green")
plot(perfprobit, add = TRUE, col= 'blue', lwd=1)
plot(perfcloglog, add = TRUE, col= "red")
```




```r
#AUC
paste("AUC for the logit model is ", aucllogit@y.values)
```

```
## [1] "AUC for the logit model is  0.996421149849919"
```

```r
paste("AUC for the probit model is ", auclprobit@y.values)
```

```
## [1] "AUC for the probit model is  0.996305703070884"
```

```r
paste("AUC for the hazard model with clog-log link is ", auclcloglog@y.values)
```

```
## [1] "AUC for the hazard model with clog-log link is  0.996421149849919"
```

Step 16: Identify the top 5 risky landings

In this step, the three models are used to identify the top 5 risky landings. **Recall that aircraft is a factor/dummy variable such that: boeing = "1" and Airbus = "0"**

The logit model and the probit model only have one flight in common. They both identified the flight with index #408 which is an Airbus flight as one of the top 5 risky flights. Aside this, all the other four identified flights are different.

The hazard model and the probit model both predicted that the flights with index #751 and #769 are amongst the top 5 risky flights. Meanwhile, the hazard model with clog-log link had no predicted flights in common with the logit model.

```r
#logit model
pred_logit.indx <- sort(as.numeric(names(tail(sort(pred_logit), 5))))
pred_logit.m <- Data_logit[pred_logit.indx,1:7]
```

```r
datatable(pred_logit.m, options = list(
  searching = TRUE,
  pageLength = 5,
  scrollX = FALSE,
  scrollCollapse = FALSE
))
```

Show 5 entries Search:

	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch
64	1	161.8924678	72	129.2649183	128.417731	33.94899883	4.139951414
176	1	197.5463502	68	126.6691821	127.9641428	23.76423143	2.993151446
307	1	154.5246036	67	129.3071841	127.5933206	23.9784968	5.154698912
362	1	63.32952055	52	132.7846766	132.9114649	18.17703022	4.110664241
408	0	131.7310956	60	131.0351822	131.3379485	28.27796554	3.660193646

Showing 1 to 5 of 5 entries Previous 1 Next

```r
#probit model
pred_probit.indxx <- sort(as.numeric(names(tail(sort(pred_probit), 5))))
pred_probit.mp <- Data_logit[pred_probit.indxx,1:7]
datatable(pred_probit.mp, options = list(
  searching = TRUE,
  pageLength = 5,
  scrollX = FALSE,
  scrollCollapse = FALSE
))
```

Show 5 entries Search:

	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch
387	1	153.8344532	61	126.8392785	126.1186482	20.54783385	4.33455751
408	0	131.7310956	60	131.0351822	131.3379485	28.27796554	3.660193646
643	0	137.5857278	66	126.2443005	127.9371077	35.17570131	2.701923695
751	0	175.5144303	49	125.2123041	125.1385489	22.5247789	4.365772364
769	0	98.50030781	66	123.3105307	124.3907675	22.32717581	4.276710488

Showing 1 to 5 of 5 entries Previous 1 Next

```
#hazard model with clog-log link
pred_cloglog.indxx <- sort(as.numeric(names(tail(sort(pred_cloglog), 5))))
pred_cloglog.mp <- Data_logit[pred_cloglog.indxx,1:7]
datatable(pred_cloglog.mp, options = list(
  searching = TRUE,
  pageLength = 5,
  scrollX = FALSE,
  scrollCollapse = FALSE
))
```

Show 5 entries Search:

	aircraft	duration	no_pasg	speed_ground	speed_air	height	pitch
669	0	140.4531145	75	120.4189481	118.484704	31.26344553	2.796731407
751	0	175.5144303	49	125.2123041	125.1385489	22.5247789	4.365772364
765	0	220.0571274	61	120.557944	118.2881798	15.66565806	4.111265292
769	0	98.50030781	66	123.3105307	124.3907675	22.32717581	4.276710488
821	0		58	113.4273968	114.844449	40.10112197	4.248428264

Showing 1 to 5 of 5 entries Previous 1 Next

Step 17: Prediction using probit model and hazard model and compare with logit model

We refer back to Step 13 and use the probit and the hazard models to predict based on the given information: (Boeing, duration = 200, no_pasg = 80, speed_ground = 115, speed_air = 120, height = 40, pitch = 4). The results are then compared to the logistic model results in Step 13.

All the three models predict that the probability that the commercial airplane is a long landing is 1. The 95% confidence interval for the predicted probability for long landing using the logit, probit, or hazard model is identical and very narrow, with both the upper and lower bounds equal to 1.

The predicted probability that the commercial airplane is a risky landing using the logit model is 0.999999998776229 which is very close to 1. Meanwhile, the predicted probability using the probit and the hazard model is 1. The 95% confidence interval for the predicted probability for risky landing using the logit, probit, or hazard model is identical and very narrow, with both the upper and lower bounds equal to 1. Based on this, we can conclude that for the given information, there's a very high probability that the commercial airplane passing over the threshold of the runway would be a long and risky landing.

```r
#predicted probability for long landing
predl_probit <- predict(probit_model, given_data, type = "response", se.fit = T)
predl_cloglog <- predict(cloglog_model, given_data, type = "response", se.fit = T)

paste("Predicted probability for long landing using the logit model is",
      pred_l$fit[["1"]])
```

```
## [1] "Predicted probability for long landing using the logit model is 1"
```

```r
paste("Predicted probability for long landing using the probit model is ",
      predl_probit$fit[["1"]])
```

```
## [1] "Predicted probability for long landing using the probit model is  1"
```

```r
paste("Predicted probability for long landing using the hazard model is ",
      predl_cloglog$fit[["1"]])
```

```
## [1] "Predicted probability for long landing using the hazard model is  1"
```

```r
#Confidence interval for predicted probability for long landing
conf_intervallogit <- round(ilogit(c(pred_l$fit[["1"]] - 1.96*pred_l$se.fit[["1"]],
                                     pred_l$fit[["1"]] + 1.96*pred_l$se.fit[["1"]])))


conf_intervalprobit <- round(ilogit(c(predl_probit$fit[["1"]] - 1.96*predl_probit$se.fit[["1"]],
                                      predl_probit$fit[["1"]]
                                      +1.96*predl_probit$se.fit[["1"]])))
conf_intervalprobit
```

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

```r
conf_intervalcloglog <- round(ilogit(c(predl_cloglog$fit[["1"]] -
                                         1.96*predl_cloglog$se.fit[["1"]],
                                       predl_cloglog$fit[["1"]] +
                                         1.96*predl_cloglog$se.fit[["1"]])))

conf_intervallogit
```

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

```r
conf_intervalprobit
```

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

```r
conf_intervalcloglog
```

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

```r
#The models for risky landing- logit model is in Step 13
probit_modelr <- glm(risky.landing ~ aircraft + speed_air,
  data = Data_logit, family = binomial(link = probit))

cloglog_modelr <- glm(risky.landing ~ aircraft + speed_air,
  data = Data_logit, family = binomial(link = cloglog))
```

```r
#predicted probability for risky landing
predr_probit <- predict(probit_modelr, given_data, type = "response", se.fit = T)
predr_cloglog <- predict(cloglog_modelr, given_data, type = "response", se.fit = T)

paste("Predicted probability for long landing using the logit model is", pred_r$fit[["1"]])
```

```
## [1] "Predicted probability for long landing using the logit model is 0.999999998776229"
```

```r
paste("Predicted probability for long landing using the probit model is ",
      predr_probit$fit[["1"]])
```

```
## [1] "Predicted probability for long landing using the probit model is  1"
```

```r
paste("Predicted probability for long landing using the hazard model is ",
      predr_cloglog$fit[["1"]])
```

```
## [1] "Predicted probability for long landing using the hazard model is  1"
```

```r
#Confidence interval for predicted probability for risky landing
conf_interval_rlogit <- round(ilogit(c(pred_r$fit[["1"]] - 1.96*pred_r$se.fit[["1"]],
                                       pred_r$fit[["1"]] + 1.96*pred_r$se.fit[["1"]])))

conf_interval_rprobit <- round(ilogit(c(predr_probit$fit[["1"]] -
                                          1.96*predr_probit$se.fit[["1"]],
                                        predr_probit$fit[["1"]] +
                                          1.96*predr_probit$se.fit[["1"]])))

conf_interval_rcloglog <- round(ilogit(c(predr_cloglog$fit[["1"]] -
                                           1.96*predr_cloglog$se.fit[["1"]],
                                         predr_cloglog$fit[["1"]] +
                                           1.96*predr_cloglog$se.fit[["1"]])))

conf_interval_rlogit
```

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

```r
conf_interval_rprobit
```

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

```r
conf_interval_rcloglog
```

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