BANA 7042 PROJECT
LONG AND RISKY LANDING PREDICTION

Name: Renganathan Lalgudi Venkatesan UCID: M12366827

Goal: To study what factors and how they would impact binary variables long landing and risky landing.
Data: Landing data (landing distance and other parameters) from 950 commercial flights (not real data set but simulated from statistical models). See two Excel files 'FAA-1.xls' (800 flights) and 'FAA-2.xls' (150 flights).

Variable dictionary:

Aircraft: The make of an aircraft (Boeing or Airbus).

Duration (in minutes): Flight duration between taking off and landing. The duration of a normal flight should always be greater than 40min.

No_pasg: The number of passengers in a flight.

Speed_ground (in miles per hour): The ground speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.

Speed_air (in miles per hour): The air speed of an aircraft when passing over the threshold of the runway. If its value is less than 30MPH or greater than 140MPH, then the landing would be considered as abnormal.

Height (in meters): The height of an aircraft when it is passing over the threshold of the runway. The landing aircraft is required to be at least 6 meters high at the threshold of the runway.

Pitch (in degrees): Pitch angle of an aircraft when it is passing over the threshold of the runway.

Distance (in feet): The landing distance of an aircraft. More specifically, it refers to the distance between the threshold of the runway and the point where the aircraft can be fully stopped. The length of the airport runway is typically less than 6000 feet.

long.landing: A binary flag that has a value of 1 if distance > 2500 and 0 otherwise

risky.landing: A binary flag that has a value of 1 if distance > 3000 and 0 otherwise

1. **Reading the xls files into R workspace and adding the Long/Risky landing variables:**

The xls files FAA1 and FAA2 from the path are read into the R workspace. Then the two variables of interest are created: Long Landing and Risky Landing

Code:

```
library(readxl)
d1<-read_excel("FAA1.xls")
d2<-read_excel("FAA2.xls")

#Dealing with Duplicates:
d2$duration <- NA
d <- rbind(d1,d2)
dup <-duplicated(d[,-2])
df <- d[!dup,]

#Removing Abnormal Values:
df1<- df[which((df$duration>40 | is.na(df$duration)) &
          (df$speed_ground >30 & df$speed_ground < 140) &
          (df$speed_air >30  & df$speed_air < 140 | is.na(df$speed_air)) &
          df$height >6 & df$distance < 6000),]

#Step-1: Adding Columns:
df1$long.landing[df1$distance > 2500] <- 1
df1$long.landing[df1$distance <= 2500] <- 0

df1$risky.landing[df1$distance > 3000] <- 1
df1$risky.landing[df1$distance <= 3000] <- 0

table(df1$long.landing)
table(df1$risky.landing)

#Removing the Distance column:
d <- df1[,-8]
```

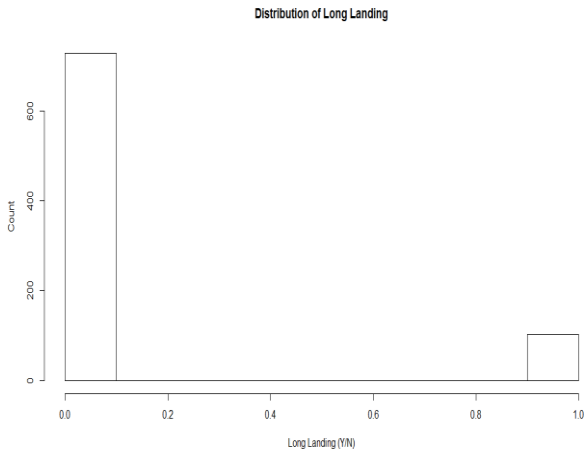
2. **Distribution of Long Landing variable:**

To understand the distribution of Long landing variable in the dataset, histograms and pie charts were used.

Code:

```
#Bar Plot
par(mfrow = c(1,1))
```

```
hist(d$long.landing, main = "Distribution of Long Landing", xlab = "Long Landing (Y/N)", ylab = "Count")

#Pie Chart
pct <- round(table(d$long.landing)/length(d$long.landing)*100,1)
labs<-c("No","Yes")
labs <- paste(labs,pct)
labs <- paste(labs,"%",sep = "")
pie(table(d$long.landing),labels = labs,col = rainbow(length(labs)),main = "Distribution of Long Landing")
```



[Figure -1]

[Figure-2]

We find that the data has 12.4% rows with long landing value set to 1 and 87.6% of rows with a 0 for long landing as shown in the Figures 1 and 2.

3. **Single Factor Regression Analysis:**

We then perform a single-factor regression analysis for each of the potential risk factors for the long landing variable.

Code:

```
m1<-glm(long.landing~aircraft, family = binomial, data = d)
m2<-glm(long.landing~duration, family = binomial, data = d)
m3<-glm(long.landing~no_pasg, family = binomial, data = d)
m4<-glm(long.landing~speed_ground, family = binomial, data = d)
m5<-glm(long.landing~speed_air, family = binomial, data = d)
m6<-glm(long.landing~height, family = binomial, data = d)
m7<-glm(long.landing~pitch, family = binomial, data = d)


summary(m1)$coef
summary(m2)$coef
```

```
summary(m3)$coef
summary(m4)$coef
summary(m5)$coef
summary(m6)$coef
summary(m7)$coef
```

We get the following observations by running the logistic regression analysis:

Name	size of Beta	Odds Ratio	Direction	P-value
speed_ground	0.4723	1.603678415	Positive	3.94E-14
speed_air	0.512	1.66862511	Positive	4.33E-11
Aircraft	0.8641	2.372869542	Positive	8.40E-05
Pitch	0.4	1.491824698	Positive	0.0466
Height	0.008	1.008032086	Positive	0.4
no_pasg	0.007	1.007024557	Negative	0.605
Duration	0.001	1.0010005	Negative	0.63

We find that four variables have a significant impact on the long landing variable as per the p-values obtained in the models. The variables height, no_pasg and duration doesn't seem to have a significant impact on the long landing variable.

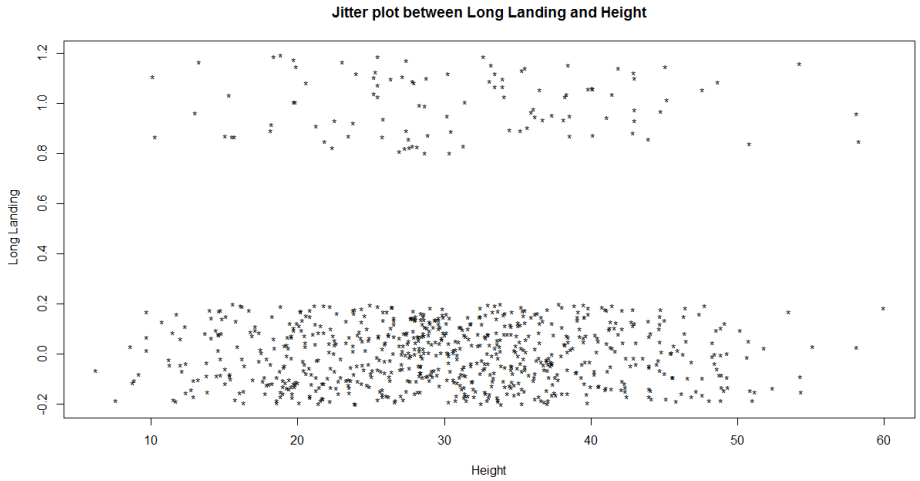
4. **Association of all variables to long landing:**

We find the association of each variable to the long landing variable by plotting the Jitter plots along with the histograms as shown below:

Code:

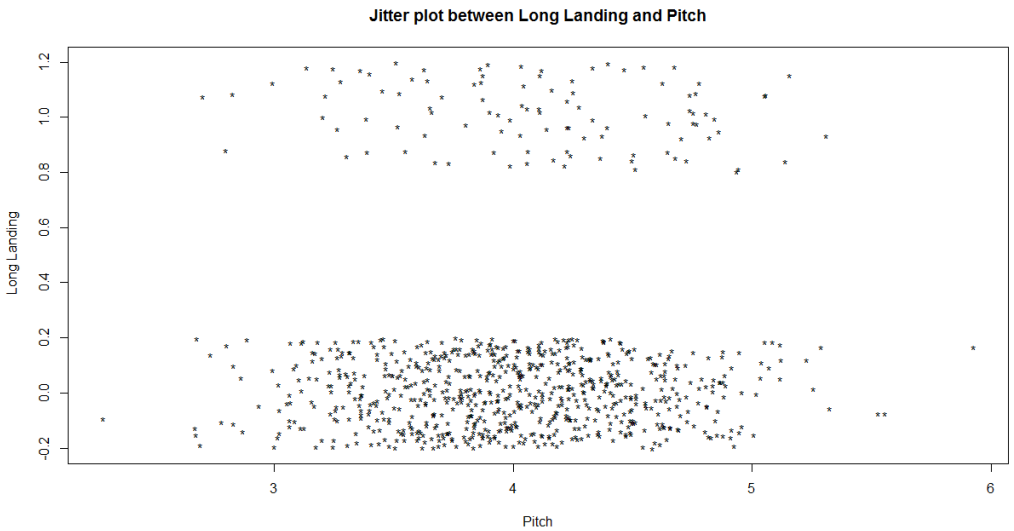
```
plot(jitter(long.landing)~jitter(height), data = d, pch = "*", main = "Jitter plot between Long Landing and Height", xlab = "Height", ylab = " Long Landing")
plot(jitter(long.landing)~jitter(speed_air), data = d, pch = "*", main = "Jitter plot between Long Landing and Speed Air", xlab = "Speed Air", ylab = " Long Landing")
plot(jitter(long.landing)~jitter(speed_ground), data = d, pch = "*", main = "Jitter plot between Long Landing and Speed Ground", xlab = "Speed Ground", ylab = " Long Landing")
plot(jitter(long.landing)~jitter(pitch), data = d, pch = "*", main = "Jitter plot between Long Landing and Pitch", xlab = "Pitch", ylab = " Long Landing")
plot(jitter(long.landing)~jitter(duration), data = d, pch = "*", main = "Jitter plot between Long Landing and Duration", xlab = "Duration", ylab = " Long Landing")

#plot(jitter(long.landing)~jitter(aircraft), data = d, pch = ".")
```

```
library(ggplot2)

#Histograms

  ggplot(d, aes(x=speed_air, fill =as.factor(long.landing))) + geom_histogram(position = "dodge",binwidth = 1)
  ggplot(d, aes(x=speed_ground, fill = as.factor(long.landing))) + geom_histogram(position = "dodge",binwidth = 1)
  ggplot(d, aes(x=height, fill = as.factor(long.landing))) + geom_histogram(position = "dodge",binwidth = 1)
  ggplot(d, aes(x=pitch, fill = as.factor(long.landing))) + geom_histogram(position = "dodge",binwidth = 1)
  ggplot(d, aes(x=duration, fill = as.factor(long.landing))) + geom_histogram(position = "dodge",binwidth = 1)
```



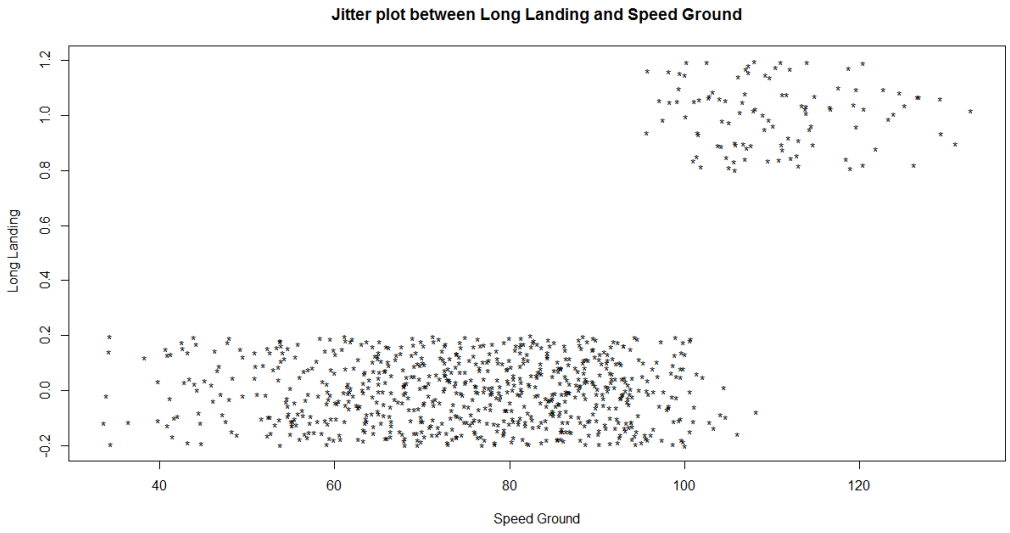
[Figure-3]

The above figure says that the Height might not be a good variable to differentiate between the 1's and 0's in the data.



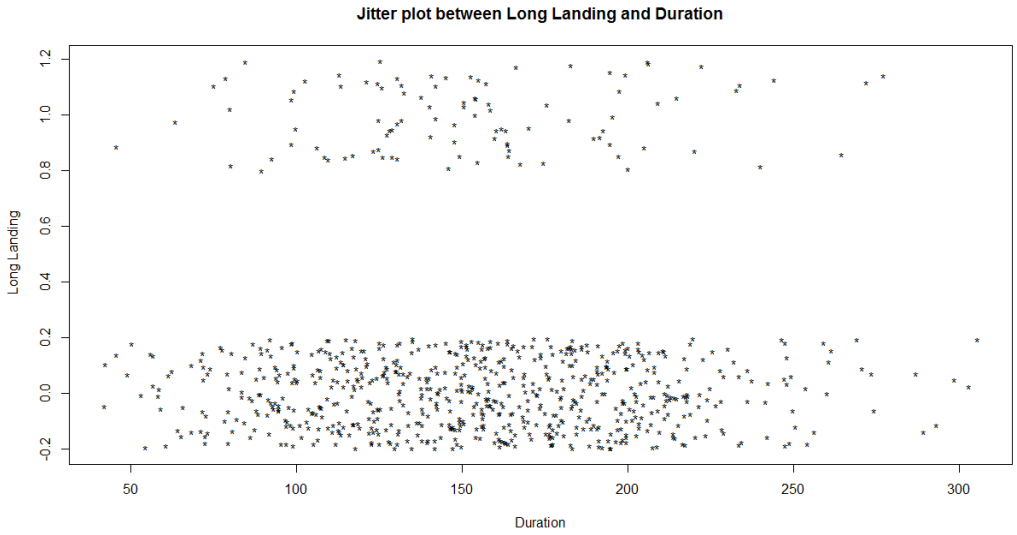
[Figure-4]

The above figure says that the Pitch might not be a good variable to differentiate between the 1's and 0's in the data.



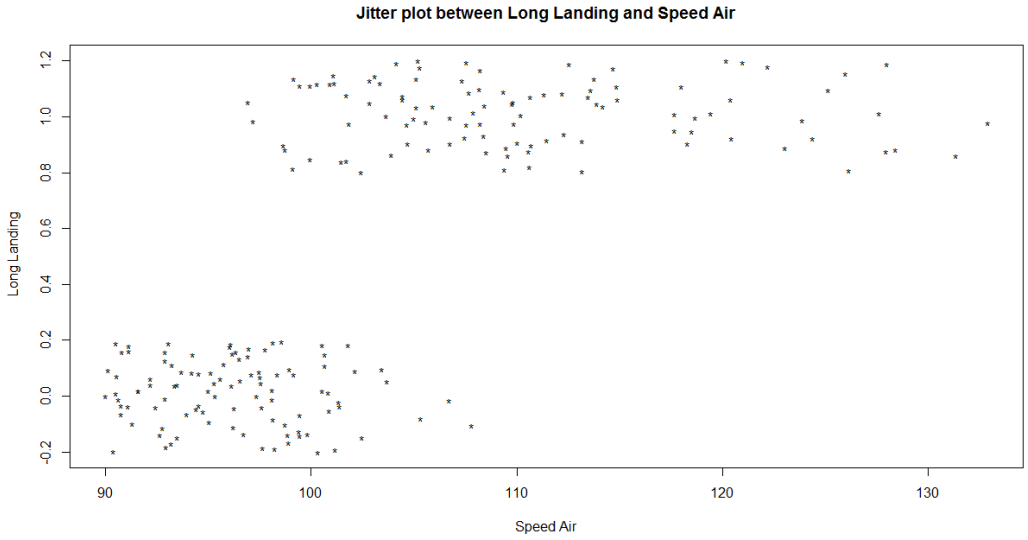
[Figure-5]

The above figure says that the Speed Ground might be a good variable to differentiate between the 1's and 0's in the data.



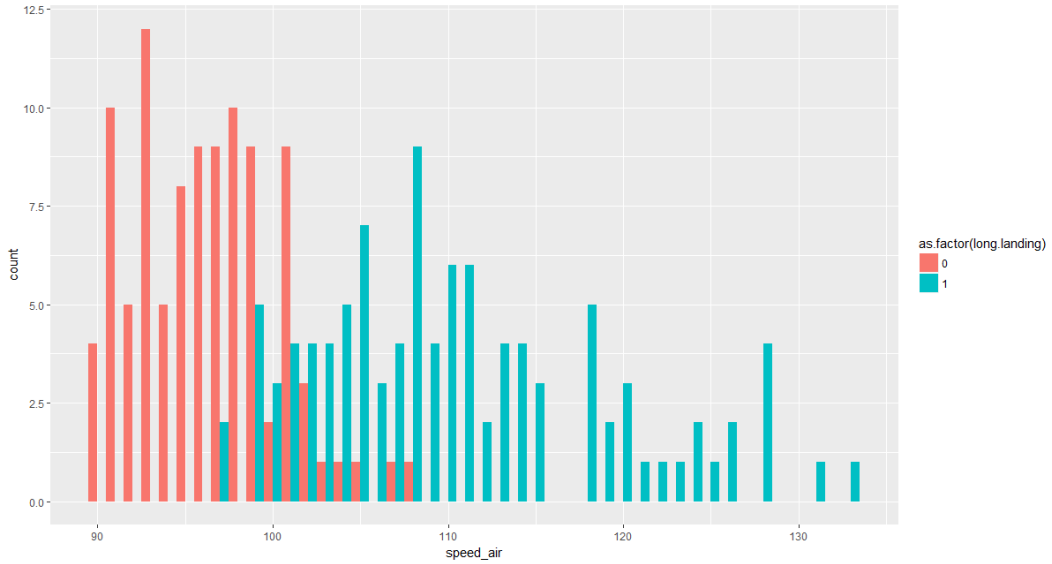
[Figure-6]

The above figure says that the Duration might not be a good variable to differentiate between the 1's and 0's in the data.



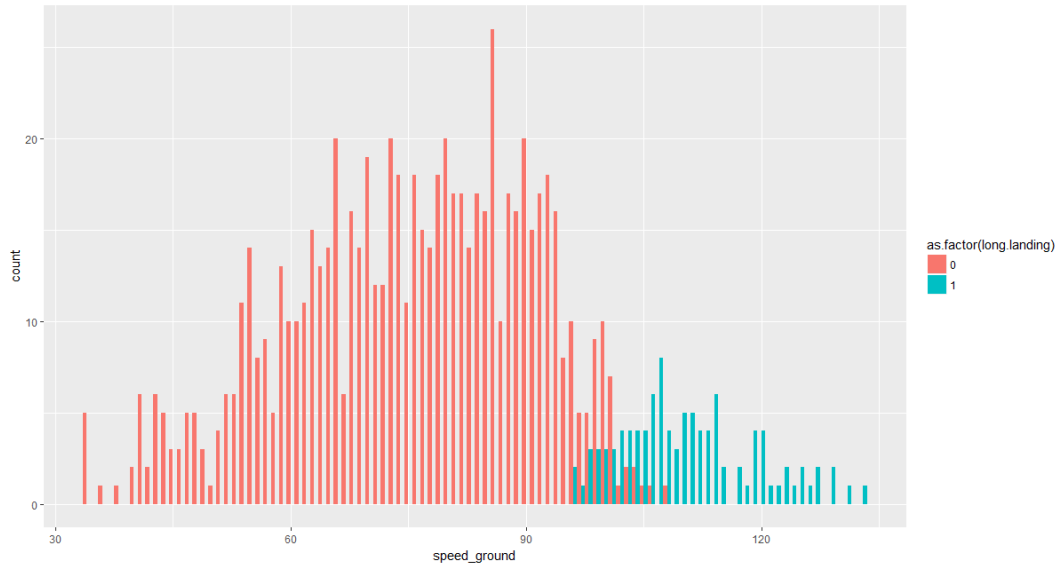
[Figure-7]

The above figure says that the Speed Air might be a good variable to differentiate between the 1's and 0's in the data.



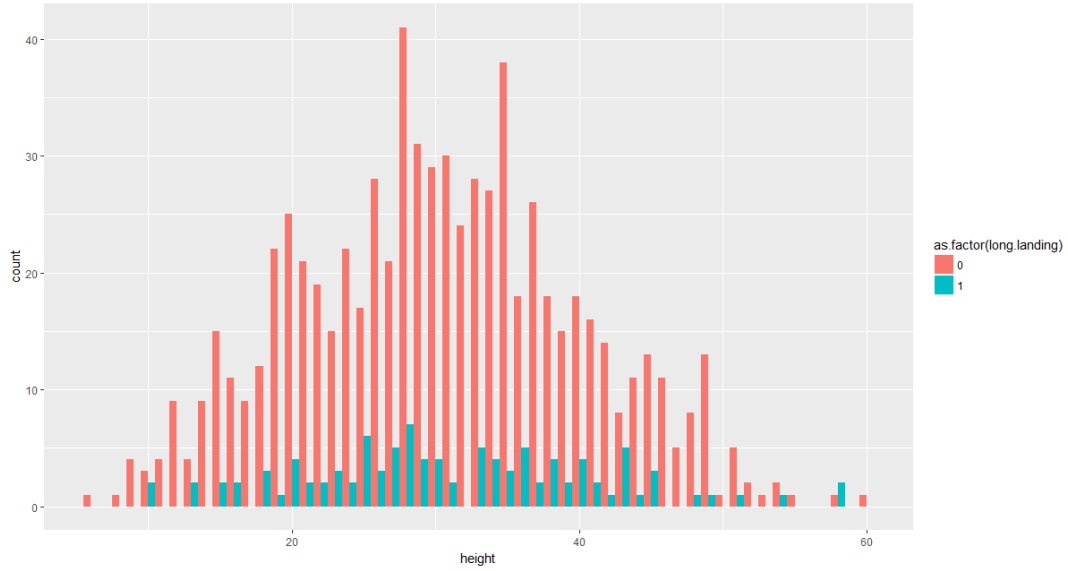
[Figure-8]

The above figure says that the Speed Air might be a good variable to differentiate between the 1's and 0's in the data.



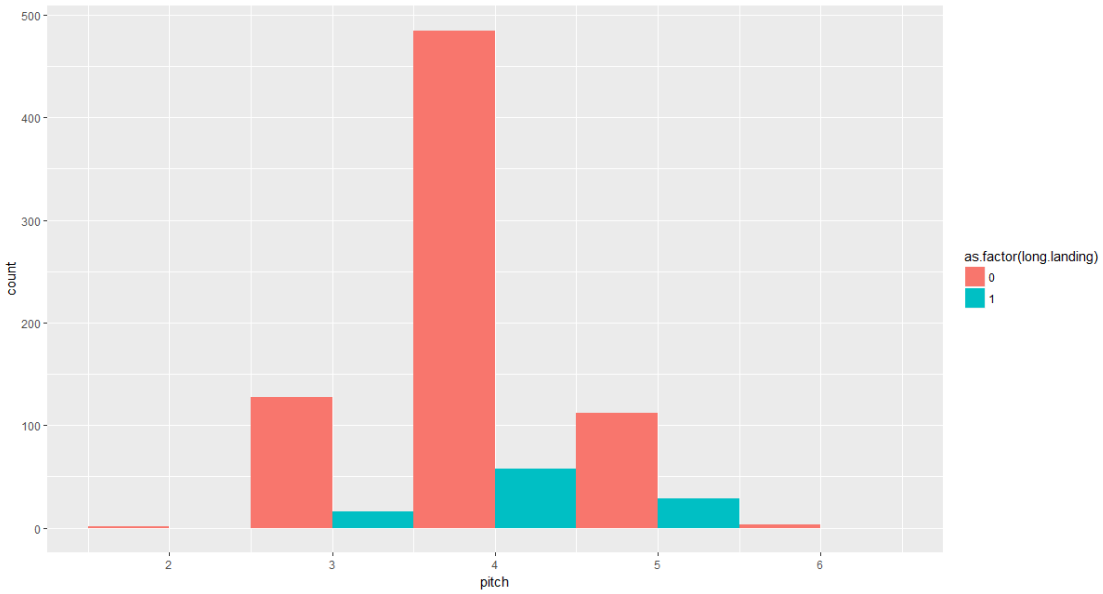
[Figure-9]

The above figure says that the Speed Ground might be a good variable to differentiate between the 1's and 0's in the data.



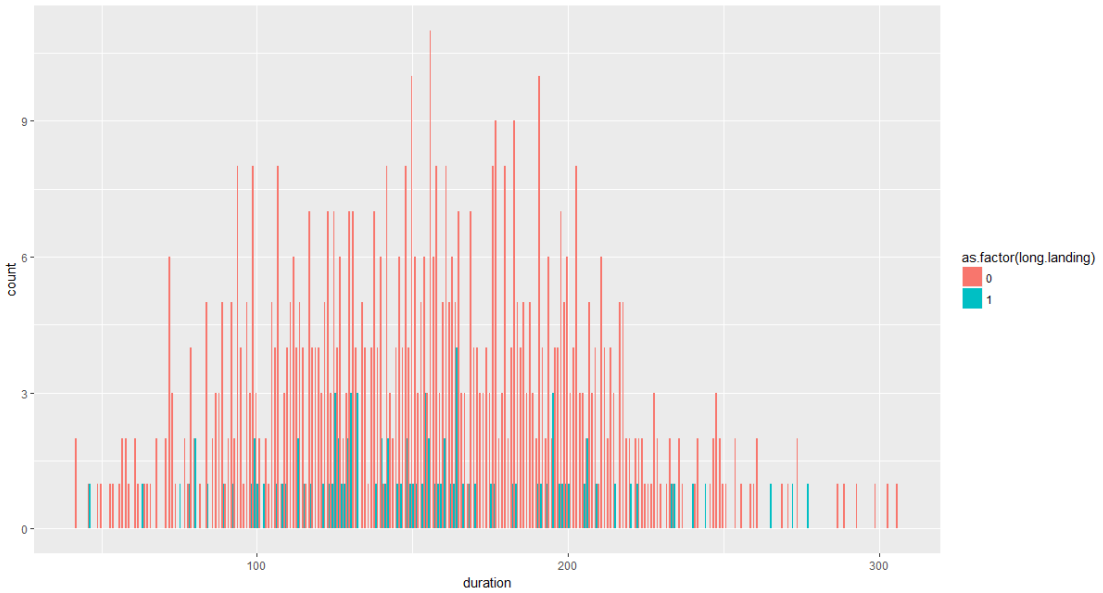
[Figure-10]

The above figure says that the Height might not be a good variable to differentiate between the 1's and 0's in the data.



[Figure-11]

The above figure says that the Pitch might not be a good variable to differentiate between the 1's and 0's in the data.



[Figure-12]

The above figure says that the Duration might not be a good variable to differentiate between the 1's and 0's in the data.

In all we find that Speed ground, Speed Air and Aircraft are the variables that might be used to differentiate the long landing significantly.

5. **Developing the Full model for predicting the long landing variable:**

   From our earlier analysis we know that speed ground and speed air are correlated and only one of those can be used for the prediction model. We are using speed ground because we find that there are lesser missing values in speed ground than in speed air and this might help in our model building.

   CODE:

```
fm <- glm(long.landing~ speed_ground + pitch + aircraft, family = binomial)
summary(fm)
```

```
> summary(fm)

Call:
glm(formula = long.landing ~ speed_ground + pitch + aircraft,
    family = binomial)

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 ***
speed_ground     0.61471    0.09184   6.694 2.18e-11 ***
pitch            1.06599    0.60389   1.765   0.0775 .
aircraftboeing   3.04348    0.73345   4.150 3.33e-05 ***
---
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 table above gives the model coefficients from the full model developed. We find that the speed ground and aircraft variables have a significant p-value but the estimate for pitch is not found to be that significant.

6. **Forward variable selection using AIC:**

Code:

```
#6 Stepwise: Forward Selection Using AIC
nullmodel<- glm(long.landing~1,data=d)
fullmodel<- fm

step1<- stepAIC(nullmodel, scope=list(lower=nullmodel, upper=fullmodel), direction='forward')
summary(step1)
```

From the table below we find that by forward selection we are almost getting the same model as in the full model with only speed ground and aircraft as significant variables but pitch doesn't seem to be significant.

```
> summary(step1)

Call:
glm(formula = long.landing ~ speed_ground + aircraft + pitch,
    data = d)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.42577  -0.18545  -0.04816   0.12155   0.70520

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.9046139  0.0797201 -11.347  < 2e-16 ***
speed_ground    0.0110714  0.0004687  23.621  < 2e-16 ***
aircraftboeing  0.0993756  0.0187997   5.286  1.6e-07 ***
pitch           0.0253749  0.0178189   1.424    0.155
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.06385945)

    Null deviance: 90.233  on 830  degrees of freedom
Residual deviance: 52.812  on 827  degrees of freedom
AIC: 78.126

Number of Fisher Scoring iterations: 2
```

7. Forward variable selection method using BIC:

Code:

```
#7 Backward Selection
n<-dim(d)[1]
step2<- step(fullmodel, direction='backward', k = log(n))
summary(step2)
```

```
> summary(step2)

Call:
glm(formula = long.landing ~ speed_ground + aircraft, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.28368  -0.01418  -0.00039   0.00000   2.56541

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -60.77049    8.67075  -7.009 2.41e-12 ***
speed_ground     0.58534    0.08441   6.934 4.08e-12 ***
aircraftboeing   3.23679    0.71189   4.547 5.45e-06 ***
---
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:  84.665  on 828  degrees of freedom
AIC: 90.665

Number of Fisher Scoring iterations: 10
```

Again, we find that the significant variables are Speed ground and aircraft in terms of predicting the long landing variable.

8. **What are risk factors for long landings and how do they influence its occurrence?**
When we built a model to predict the long landing, we get the following model after variable selection using forward stepwise methodology using both AIC and BIC.

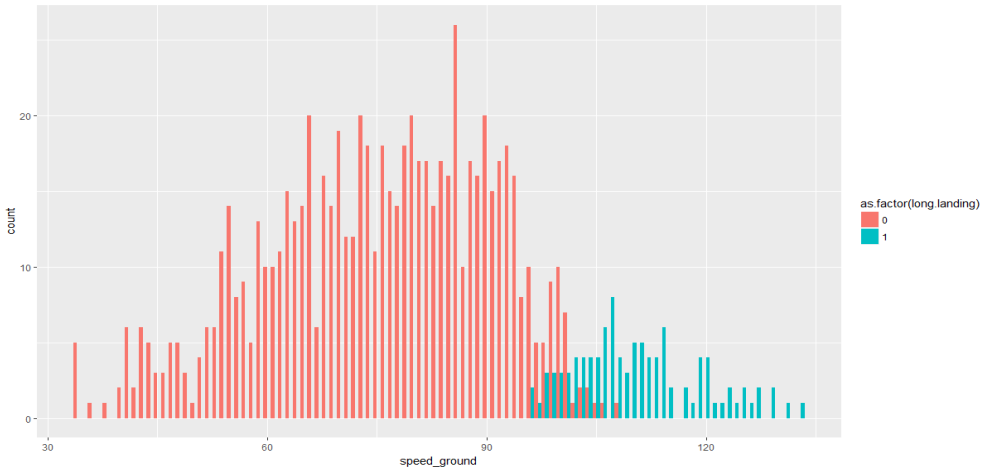
Long.landing ~ Speed_Ground + Aircraft

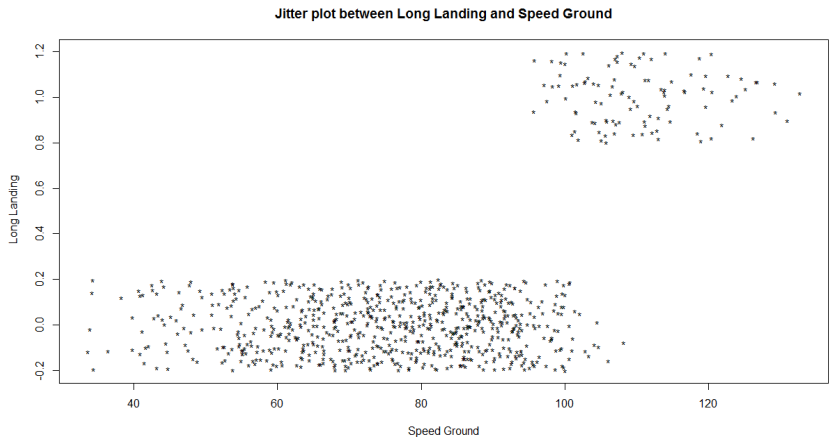
With the following beta estimates:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-60.77049	8.67075	-7.009	2.41E-12	***
speed_ground	0.58534	0.08441	6.934	4.08E-12	***
aircraftboeing	3.23679	0.71189	4.547	5.45E-06	***

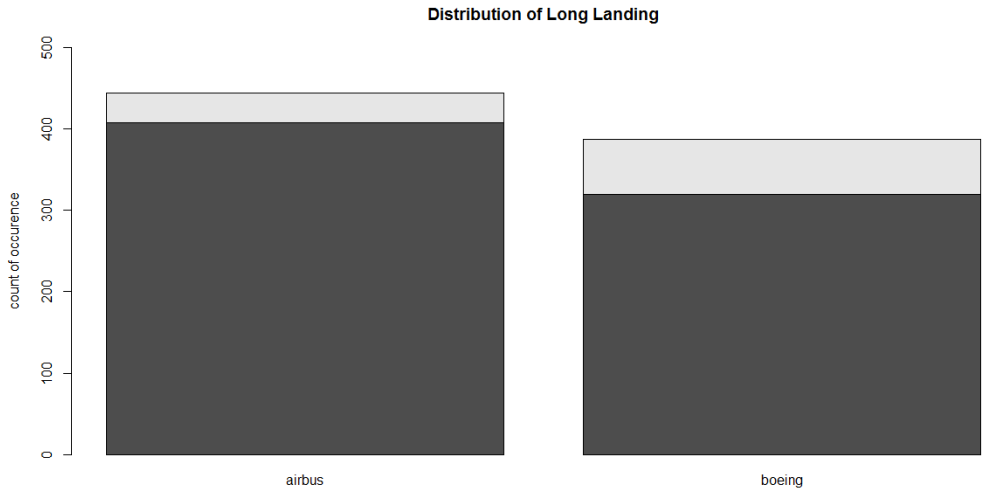
We have the following observation from the model:

- We find that the there is a marked difference in long landing variable across the speed ground variable as shown in the image below. This is also reflected in the speed ground variable being chosen for the prediction of long landing.



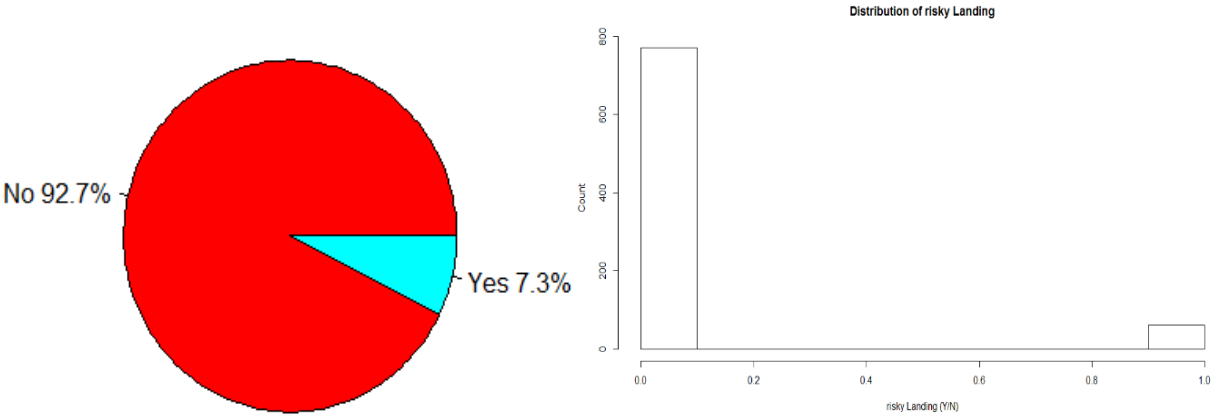


- We also find that the distribution of long landing variable is different across the different aircraft variables as shown in the image below. This explains why aircraft variable has a significant impact on predicting the long landing.



- From the beta estimates, we find that for every unit change in Speed ground, the odds ratio of long landing becoming 1 increases by e^0.5853 = 1.79553 times or 79.55%.
- From the beta estimates, we find that by changing the aircraft type from Boeing to Airbus, the odds of long landing increases by e^3.2367 = 25.45 times.

9. **Risky Landing Prediction:**
   We then did a similar analysis with the Risky Landing variable.
   Distribution of Data:



[Figure-14]

We find that there are 7.3% rows that have been labelled as 1's and 92.7% of rows marked as 0's. This seen in the Figure- 14 above.

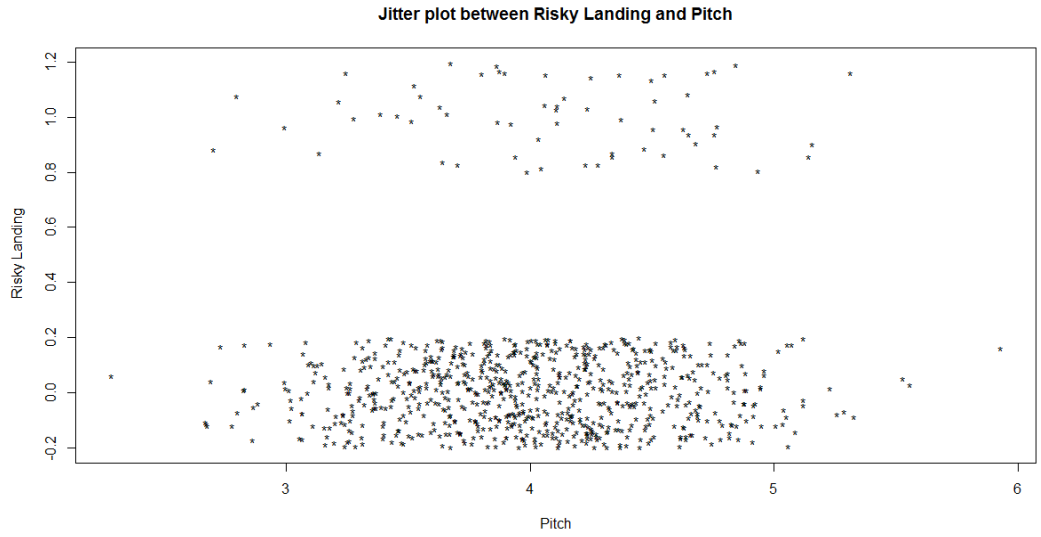
Single Factor Models

We then developed the single factor model which gave us the following estimates

Name	size of Beta	Odds Ratio	Direction	P-value
speed_ground	0.6142187	1.848212	Positive	6.90E-08
speed_air	0.8704019	2.3878703	Positive	3.73E-06
aircraft	1.001775	2.7231111	Positive	4.56E-04
height	0.0022186	1.0022211	Negative	8.71E-01
pitch	0.371072	1.4492874	Positive	0.143296135
no_pasg	0.0253793	1.0257041	Negative	0.1536237
duration	0.0011518	1.0011525	Negative	0.68

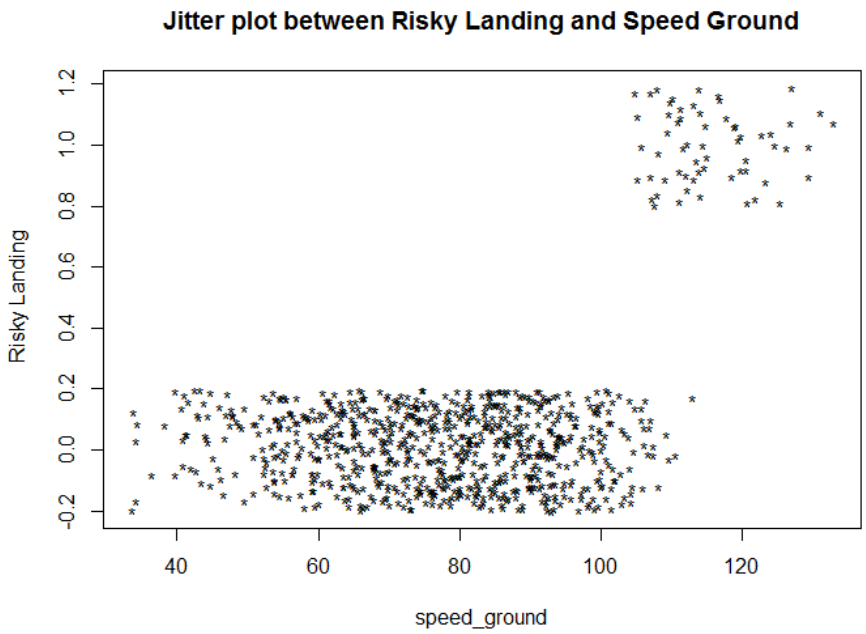
We find that only speed ground, speed air and aircraft seem to be significant variables in predicting the risky landing variable.

Association of variables to the Risky Landing Variable:



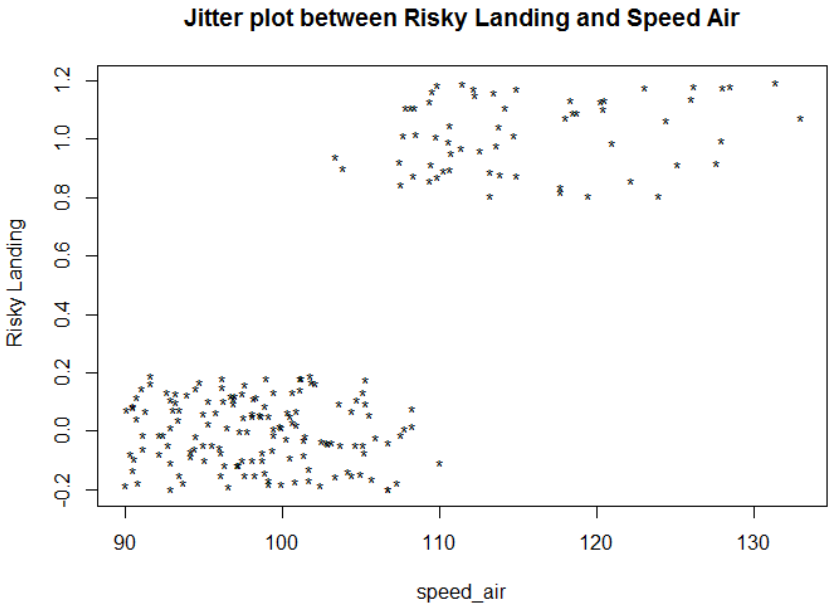
[Figure-15]

We find that there is no significant difference in pitch values for the Risky Landing variable



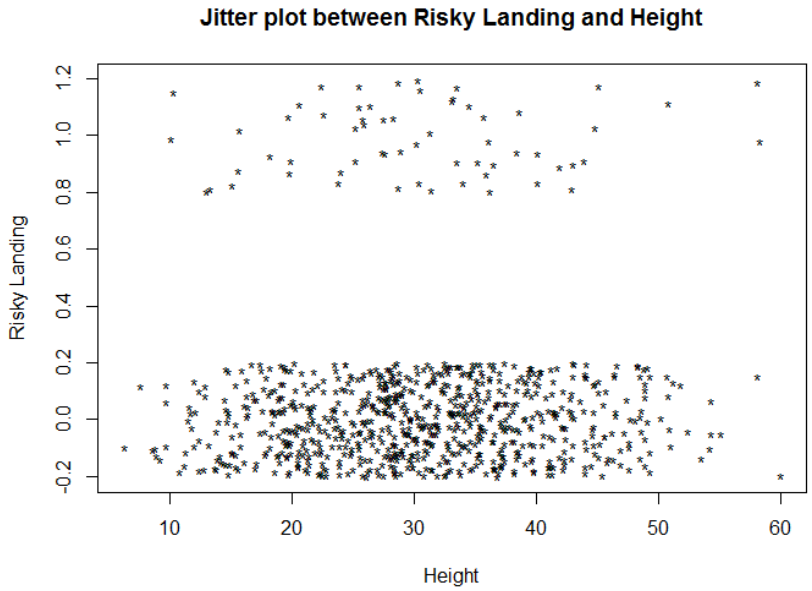
[Figure-16]

We find that there is a significant difference in Speed Ground values for the Risky Landing variable



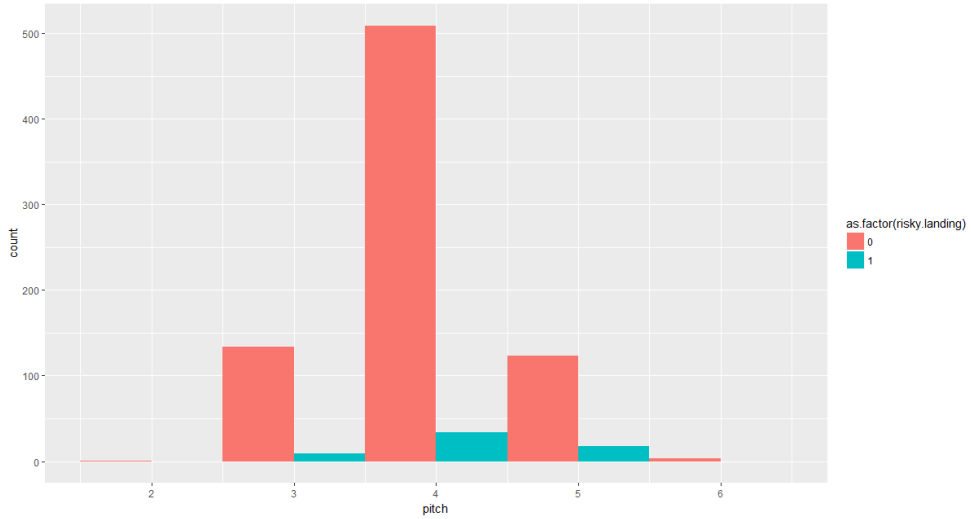
[Figure-17]

We find that there is a significant difference in speed air values for the Risky Landing variable



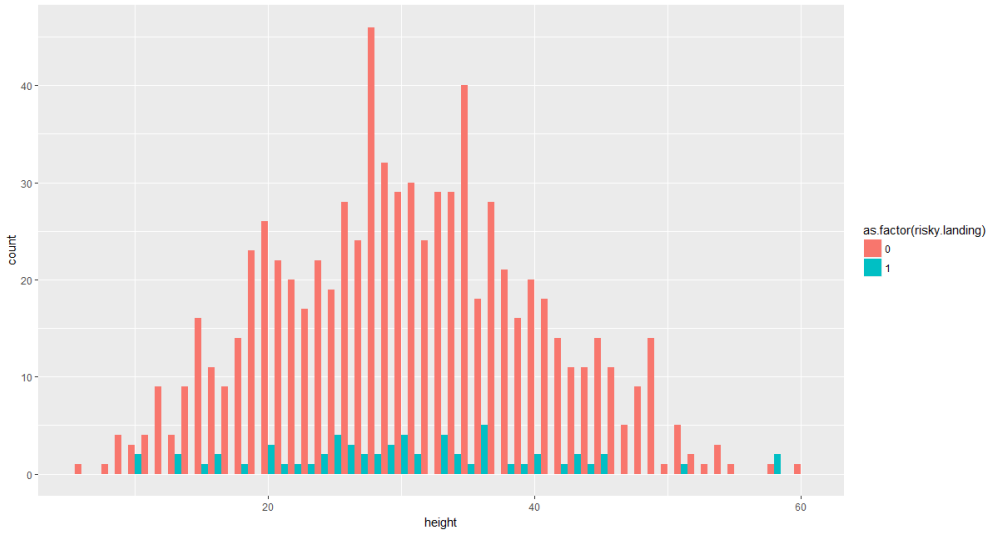
[Figure-18]

We find that there is no significant difference in height values for the Risky Landing variable



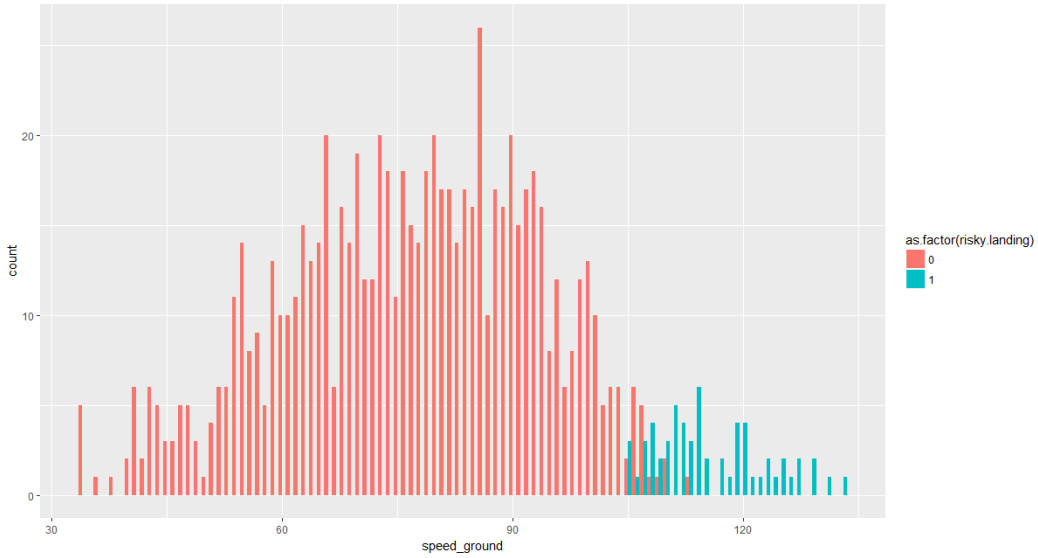
[Figure-19]

We find that there is no significant difference in pitch values for the Risky Landing variable



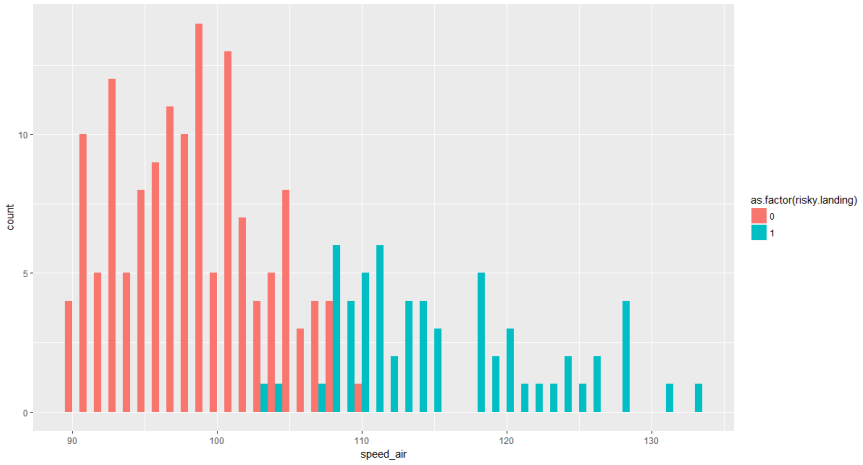
[Figure-20]

We find that there is no significant difference in height values for the Risky Landing variable



[Figure-21]

We find that there is significant difference in speed ground values for the Risky Landing variable



[Figure-22]

We find that there is significant difference in speed air values for the Risky Landing variable

Full Model Development:

```
fm1<- glm(risky.landing~ speed_ground + aircraft, family = binomial)

summary(fm1)
```

From the above analysis we can say that the significant variables that have an impact on the risky landing variable are speed air, speed ground and aircraft. But we know that speeds are correlated and thus we are using only speed ground variable in our prediction model because it has lesser missing values when compared to speed air variable.

```
> summary(fm1)

Call:
glm(formula = risky.landing ~ speed_ground + aircraft, family = binomial)

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 ***
speed_ground      0.9263     0.2248   4.121 3.78e-05 ***
aircraftboeing    4.0190     1.2494   3.217   0.0013 **
---
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
```

Stepwise Forward Selection using AIC:

```
> summary(step3)

Call:
glm(formula = risky.landing ~ speed_ground + aircraft, data = d)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.32632  -0.13664  -0.04278   0.06619   0.73574

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.568967   0.033776 -16.845  < 2e-16 ***
speed_ground   0.007623   0.000401  19.011  < 2e-16 ***
aircraftboeing 0.077310   0.015052   5.136  3.5e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.0467696)

    Null deviance: 56.522  on 830  degrees of freedom
Residual deviance: 38.725  on 828  degrees of freedom
AIC: -181.69

Number of Fisher Scoring iterations: 2
```

We find that we obtain similar results by using a stepwise forward selection as shown above

Stepwise forward selection using BIC:

```
> summary(step4)

Call:
glm(formula = risky.landing ~ speed_ground + aircraft, family = binomial)

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 ***
speed_ground     0.9263     0.2248   4.121 3.78e-05 ***
aircraftboeing   4.0190     1.2494   3.217   0.0013 **
---
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
```

We find that we obtain similar results by using a stepwise backward selection as shown above

10. Presentation of results for the Risky Landing model:

When we built a model to predict the long landing, we get the following model after variable selection using forward stepwise methodology using both AIC and BIC.

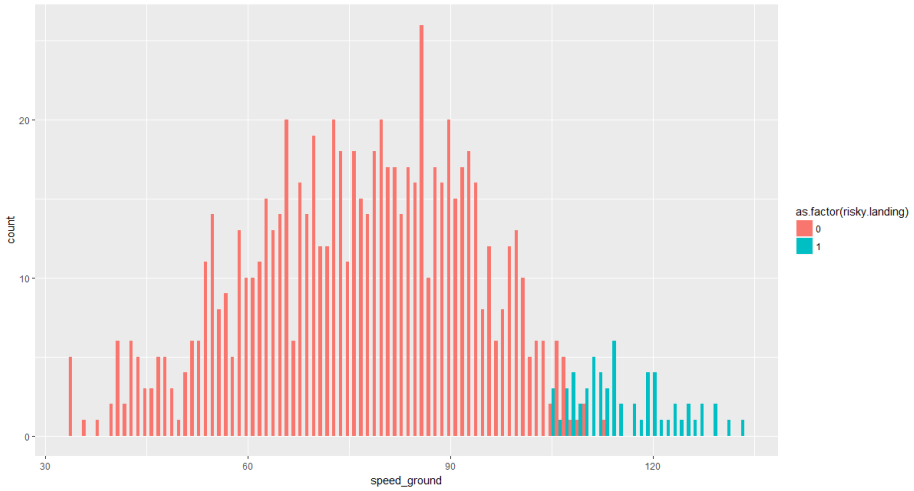
Risky.landing ~ Speed_Ground + Aircraft

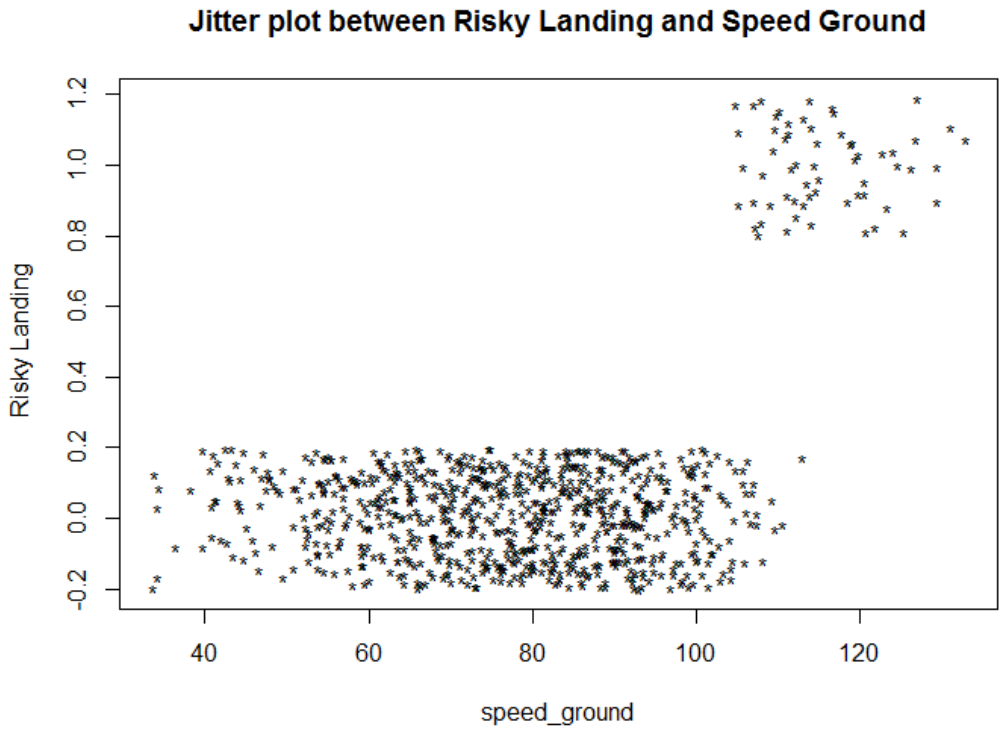
With the following beta estimates:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-102.0772	24.7751	-4.12	3.79E-05	***
speed_ground	0.9263	0.2248	4.121	3.78E-05	***
aircraftboeing	4.019	1.2494	3.217	0.0013	**

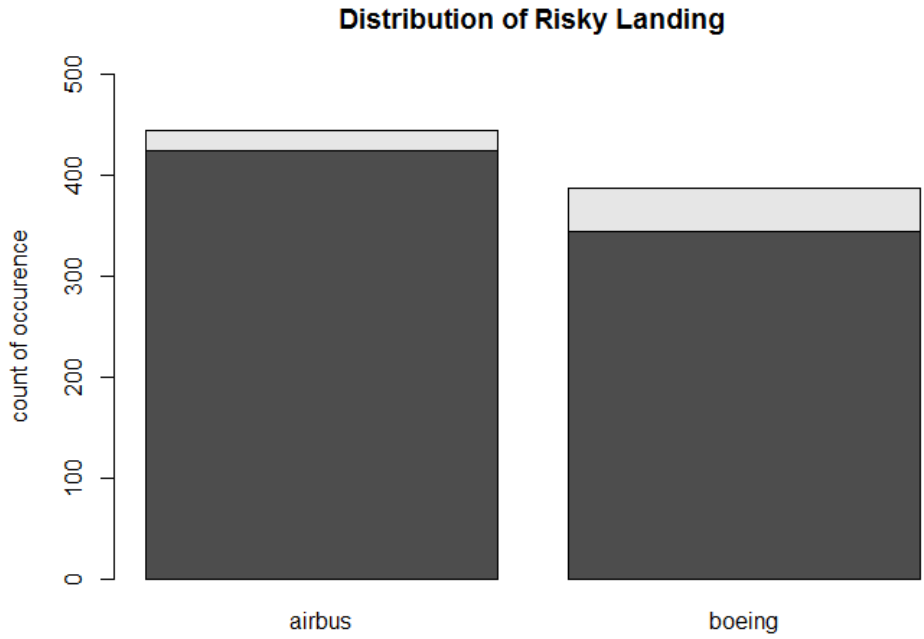
From the estimates we have the following model:

- We find that the there is a marked difference in risky landing variable across the speed ground variable as shown in the image below. This is also reflected in the speed ground variable being chosen for the prediction of risky landing.





- We also find that the distribution of risky landing variable is different across the different aircraft variables as shown in the image below. This explains why aircraft variable has a significant impact on predicting the risky landing.



-
  From the beta estimates, we find that for every unit change in Speed ground, the odds ratio of risky landing becoming 1 increases by e^0.9263 = 2.525149 times
- From the beta estimates, we find that by changing the aircraft type from Boeing to Airbus, the odds of long landing increases by e^4.01 = 58.73 times.

11. Difference between the two models:

By implementing both the models, we have the following inferences:

- From the data set we find that there are 103 long landing cases and there are 61 risky landings out of the 831 landings.
- We find that both the long landing as well as the risky landing depends majorly on the speed ground and aircraft type variables.
- But we find that the level at which both the variables affect the odds of long landing and risky landing are different. The odds of risky landing seem to be more for the same given unit change in speed ground and for a change of aircraft.

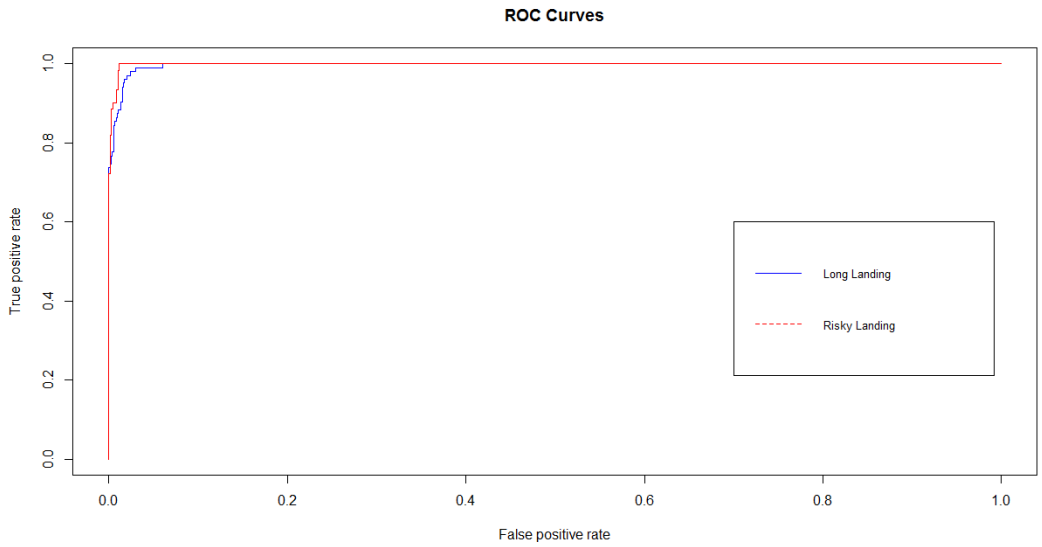
12. ROC Plots:

We then plotted the ROC curves for both the models used for the prediction of Long landing and the risky landing variables.

CODE

```
library(ROCR)
```

```
pred1 <- prediction(predprob1, d$long.landing)
perf1 <- performance(pred1, "tpr", "fpr")
plot(perf1, col = "blue", main = "ROC Curves")

pred2 <- prediction(predprob2, d$risky.landing)
perf2 <- performance(pred2, "tpr", "fpr")
plot(perf2, add=TRUE, col = "red")

legend(0.7,0.6, legend=c("Long Landing","Risky Landing"),
       col=c("blue", "red"), lty=1:2, cex=0.8)
```



We find that the ROC curves almost overlap each other. We also find that the AUC value is pretty high for both the variables meaning that there is little mis classification happening in our model.

13. Prediction for a given data:

We now want to predict the probability of long landing as well as risky landing for a given set of inputs.

CODE:

```
new_data<- data.frame(aircraft= "boeing", duration=200, no_pasg=80,
speed_ground=115, speed_air=120, height=40, pitch=4)
View(new_data)
as.data.frame(new_data)

library(faraway)
#Prediction + CI
out1 <- predict(fm,newdata = new_data, type = "link", se =T)
val1 <- ilogit(c(out1$fit-1.96*out1$se.fit,out1$fit+1.96*out1$se.fit))
#val1
```

```
out2 <- predict(fm1,newdata = new_data, type = "response", se = T)
val2 <- ilogit(c(out2$fit-1.96*out2$se.fit,out2$fit + 1.96*out2$se.fit))
```

We get the probability of long landing to be 0.9999577 with the CI as mentioned in the table below. We also get the probability of risky landing to be 0.999789 with a CI as shown in the table below.

Variable	Confidence Interval	
Long Landing	0.999	0.999998
Risky Landing	0.98748	0.999997

14. Compare models with different link functions

We now wanted to compare the results obtained by using the logit link function to the results we will obtain by using a probit or complimentary log-log link function.

CODE:

```
fm1_1<- glm(risky.landing~ speed_ground + aircraft, family = binomial)
fm1_2<- glm(risky.landing~ speed_ground + aircraft, family = binomial(link = "probit"))
fm1_3<- glm(risky.landing~ speed_ground + aircraft, family = binomial(link = "cloglog"))
summary(fm1_1)$coef
summary(fm1_2)$coef
summary(fm1_3)$coef
```

```
> summary(fm1_1)$coef
                  Estimate Std. Error   z value     Pr(>|z|)
(Intercept)    -102.0772449  24.775123 -4.120151 3.786244e-05
speed_ground      0.9262743   0.224791  4.120602 3.778843e-05
aircraftboeing    4.0190312   1.249390  3.216794 1.296316e-03
> summary(fm1_2)$coef
                  Estimate Std. Error   z value     Pr(>|z|)
(Intercept)    -58.6931140 13.3132792 -4.408614 1.040341e-05
speed_ground      0.5322474  0.1206597  4.411146 1.028250e-05
aircraftboeing    2.3567050  0.7015839  3.359120 7.819097e-04
> summary(fm1_3)$coef
                  Estimate Std. Error   z value     Pr(>|z|)
(Intercept)    -69.2654397 14.7396098 -4.699272 2.610902e-06
speed_ground      0.6220561  0.1326413  4.689763 2.735222e-06
aircraftboeing    2.8984289  0.8002036  3.622114 2.922049e-04
```

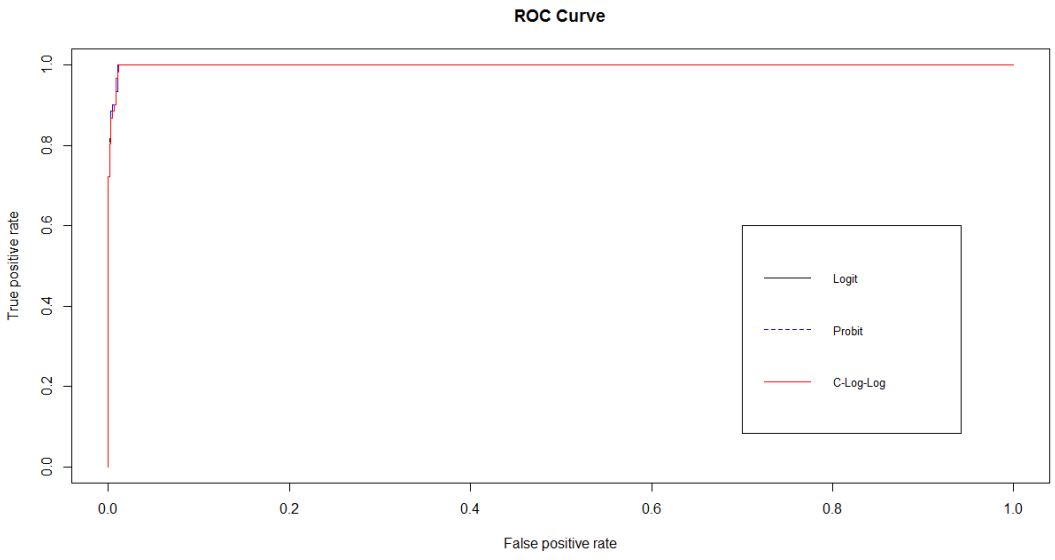
We find that all the logit link function gave almost twice the beta estimates for the varibles speed ground and aircraft. But both the probit as well as the complimentary log-log gave a conservative estimates for Betas in the model.

15. Compare the three models by showing their ROC curves in the same plot

We then wanted to compare the performance of the models using the ROC curves for all the three models in predicting the risky landing variable.

CODE:

```
predprob1_1<-predict(fm1_1,type="response")
predprob1_2<-predict(fm1_2,type="response")
predprob1_3<-predict(fm1_3,type="response")

pred1_1 <- prediction(predprob1_1, d$risky.landing)
perf1_1 <- performance(pred1_1, "tpr", "fpr")
plot(perf1_1, lty = 2, main = "ROC Curve")

pred1_2 <- prediction(predprob1_2, d$risky.landing)
perf1_2 <- performance(pred1_2, "tpr", "fpr")
plot(perf1_2, add=T, col = "green")

pred1_3 <- prediction(predprob1_3, d$risky.landing)
perf1_3 <- performance(pred1_3, "tpr", "fpr")
plot(perf1_3, col = "red", add = T)

legend(0.7,0.6, legend=c("Logit","Probit", "C-Log-Log"),
      col=c("black","blue", "red"), lty=1:2, cex=0.8)
```



We obtain the ROC curve as shown above. We find that the curves overlap for most part of the analysis. This might be because the variables used for prediction were already selected and were significant. So it was not making much of a difference by choosing a different link function.

16. Top 5 risky landings:

We wanted to compare the top 5 risky landings from all the three models. We can do that by implementing the codes below:

CODE:

```
risk1<-predict(fm1_1,type="response")
risk2<-predict(fm1_2,type="response")
risk3<-predict(fm1_3,type="response")

View(risk1)
class(risk1)
r1<-as.data.frame(risk1)
col2<-c(no = 1:831)
 r1<-cbind(risk1,col2)
r1<-as.data.frame(r1)

library(dplyr)
head(r1<-arrange(r1,desc(risk1)))

r2<-as.data.frame(risk2)
col2<-c(no = 1:831)
r2<-cbind(risk2,col2)
r2<-as.data.frame(r2)

head(r2<-arrange(r2,desc(risk2)))

r3<-as.data.frame(risk3)
col2<-c(no = 1:831)
r3<-cbind(risk3,col2)
r3<-as.data.frame(r3)

head(r3<-arrange(r3,desc(risk3)))
```

Risk1	Flight
1	362
1	307
1	64
1	387
1	408
1	176

Risk2	Flight
1	56
1	64
1	134
1	176
1	179
1	307

Risk3	Flight
1	19
1	29
1	30
1	56
1	64
1	90

- By ranking the probabilities in descending order for each of the models, we find that there are any flights that are predicted with a probability 1 to have a risky landing.
- But we actually cannot make any significant comment regarding if there is a difference in the flights that had a risky landing prediction based on the method used for prediction.
- The first 5 observations in the dataset that were categorised as risky landing is shown in the table above.

17. Comparison of results from prediction of different models:

We finally want to compare the results obtained in the prediction by using different link functions. We find that there is a bit of a difference in the CI's obtained, but the values are not different by a big margin.

CODE:

```
new_data<- data.frame(aircraft= "boeing", duration=200, no_pasg=80, speed_ground=115, speed_air=120, height=40, pitch=4)
View(new_data)
as.data.frame(new_data)
out3 <- predict(fm1_2,newdata = new_data, type = "link", se =T)
out3 <- predict(fm1_2,newdata = new_data, type = "link", se =T)
val3 <- pnorm(c(out3$fit-1.96*out3$se.fit,out3$fit+1.96*out3$se.fit))
val3
out4 <- predict(fm1_3,newdata = new_data, type = "link", se = T)
```

LINK	Confidence Intervals	
Logit	0.9989955	0.9999982
Probit	0.9961	1
C-LogLog	0.9952	1

Usually the idea behind comparing the values is to understand the concept that probit models have thinner tails and logit models have larger tails. Also, the complimentary log-log has a larger tail near the 0 end and has a thinner tail near the end of predicting 1.

We are able to see that since the new data lies in the region close to predicting 1, the slope of the link function is more for the probit and the complimentary log-log model and thus they give a wider confidence interval for the prediction as shown in the table above.