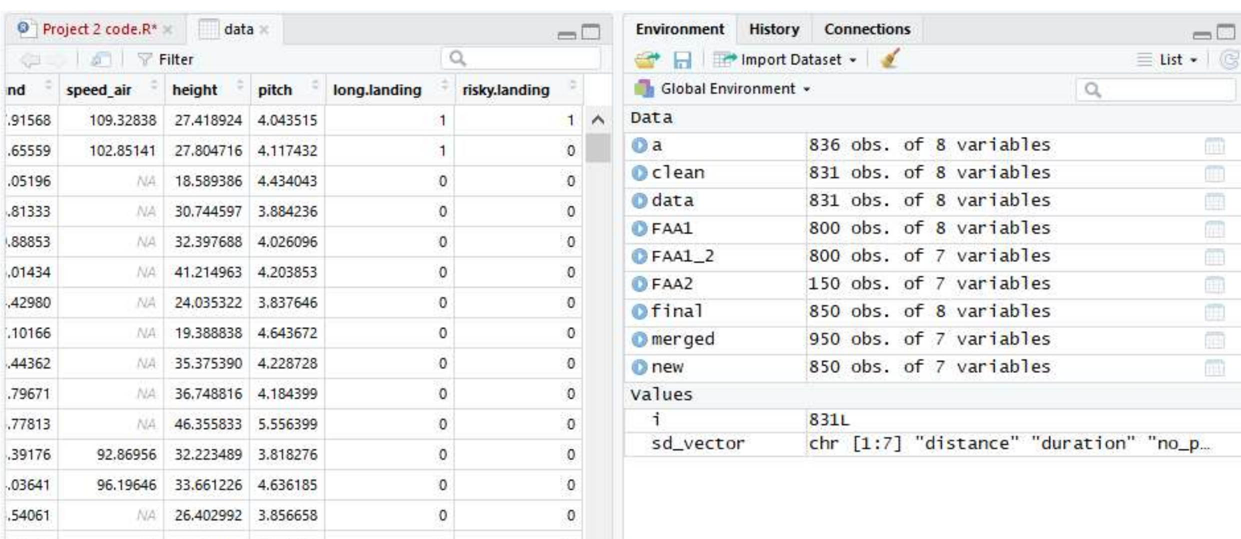
Statistical Modeling Project 2
Ashwita Saxena
M06119969

Step 1. From now on, please work on the cleaned FAA data set you prepared by carrying out Steps 1-9 in Part 1 of the project. Create two binary variables below and attach them to your data set. long.landing = 1 if distance > 2500; =0 otherwise risky.landing = 1 if distance > 3000; =0 otherwise. Discard the continuous data you have for "distance", and assume we are given the binary data of "long.landing" and "risky.landing" only.

```
##########
# step 1 #
##########

data$long.landing = 0

for (i in 1:831) {
  if (data$distance[i] > 2500) data$long.landing[i] = 1
}
sum(data$long.landing)

data$risky.landing = 0

for (i in 1:831) {
  if (data$distance[i] > 3000) data$risky.landing[i] = 1
}
sum(data$risky.landing)
```



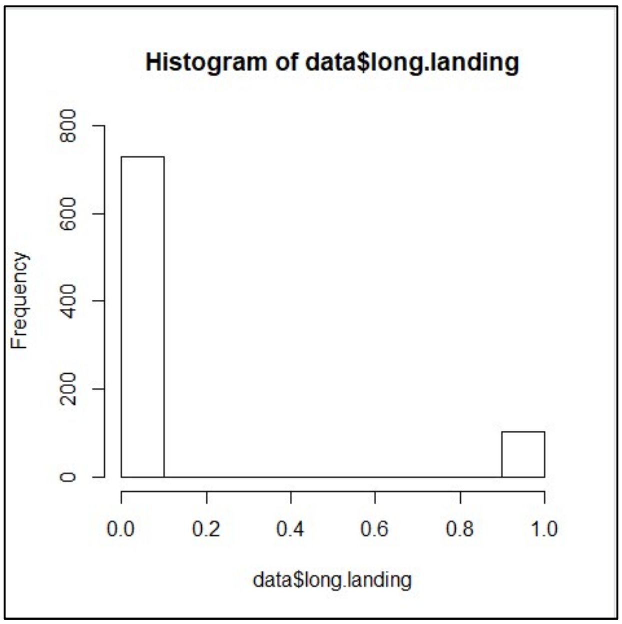
nd	speed_air	height	pitch	long.landing	risky.landing
.91568	109.32838	27.418924	4.043515	1	1
.65559	102.85141	27.804716	4.117432	1	0
.05196	NA	18.589386	4.434043	0	0
.81333	NA	30.744597	3.884236	0	0
.88853	NA	32.397688	4.026096	0	0
.01434	NA	41.214963	4.203853	0	0
.42980	NA	24.035322	3.837646	0	0
.10166	NA	19.388838	4.643672	0	0
.44362	NA	35.375390	4.228728	0	0
.79671	NA	36.748816	4.184399	0	0
.77813	NA	46.355833	5.556399	0	0
.39176	92.86956	32.223489	3.818276	0	0
.03641	96.19646	33.661226	4.636185	0	0
.54061	NA	26.402992	3.856658	0	0
.77467	NA	31.228665	3.902046	0	0

Observation: The new dataset 831 observations and 9 variables. We have added two binary variables long.landing where distance > 2500 and risky landing where distance > 3000.

Conclusion: There are 103 long landing observations and 61 risky landing observations

Step 2. Use a pie chart or a histogram to show the distribution of "long.landing"

```
##########
# step 2 #
##########

par(mfrow=c(1,1))
hist(data$long.landing,ylim = range(0,800))
```



Observation: According to the histogram, we see that there are more zeroes as compared to ones for long landing variable. That means very few flights have distance greater than 2500.

Conclusion: The histogram of the binary variable looks like the one above.

Step 3. Perform single-factor regression analysis for each of the potential risk factors, in a similar way to what you did in Steps 13-15 of Part 1. But here the response "long.landing" is binary. You may consider using logistic regression. Provide a table that ranks the factors from the most important to the least. This table contains 5 columns: the names of variables, the size of the regression coefficient, the odds ratio, the direction of the regression coefficient (positive or negative), and the p-value.

```
##########
# step 3 #
##########

data1 <- as.data.frame(data)
for (i in c(1:7)){
  model <- glm (long.landing ~ data1[,i],
                family = binomial(link='logit'),
                data=data1)
```

```
  print(summary(model))
}
```

Rank	Variable	Coefficient	Odds Ratio	Direction	P Value
1	Speed_ground	0.47235	1.603759	Positive	3.94E-14
2	speed_air	0.51232	1.669159	Positive	4.33E-11
3	Boeing	0.8641	2.372870	Positive	8.40E-05
4	pitch	0.4005	1.492571	Positive	0.0466
5	height	0.008624	1.008661	Positive	0.422
6	no_pasg	0.007256	1.007282	Negative	0.6059
7	duration	0.00107	1.001071	Negative	0.631

Observation: We can see that according to the p-values, the most significant factors are speed ground, speed air, aircraft Boeing and pitch, at 95% significance level. Height, number of passenger and duration are not significant.

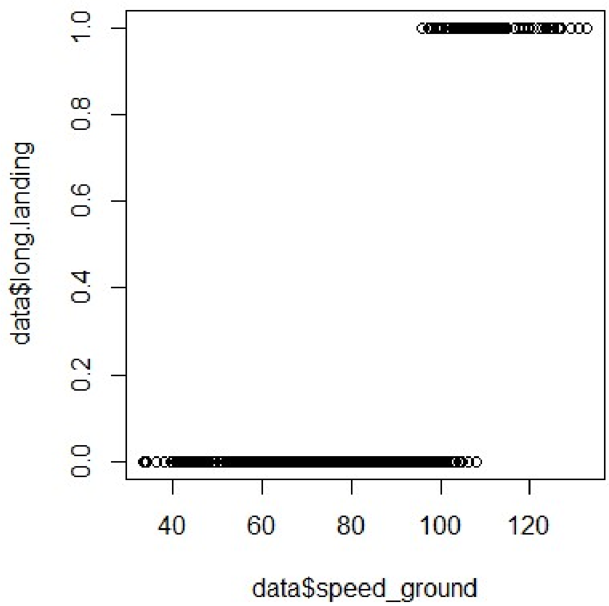
Conclusion: In order of importance, Speed ground, speed air, aircraft type Boeing and pitch are the most important variables

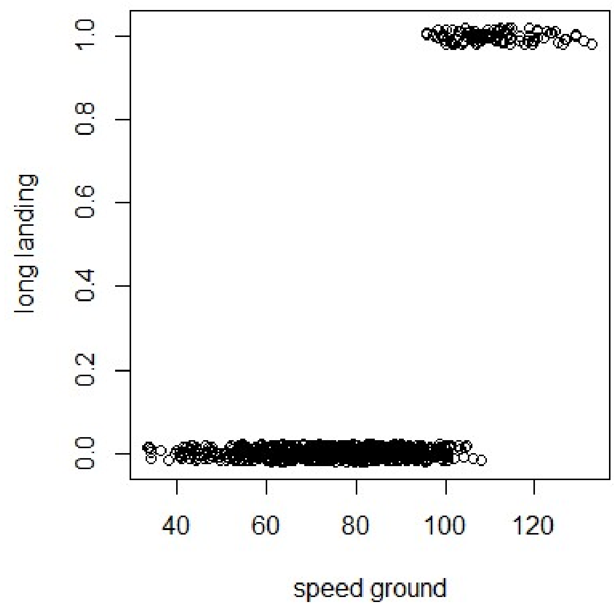
Step 4. For those significant factors identified in Step 3, visualize its association with "long.landing". See the slides (pp. 12-21) for Lecture 3.

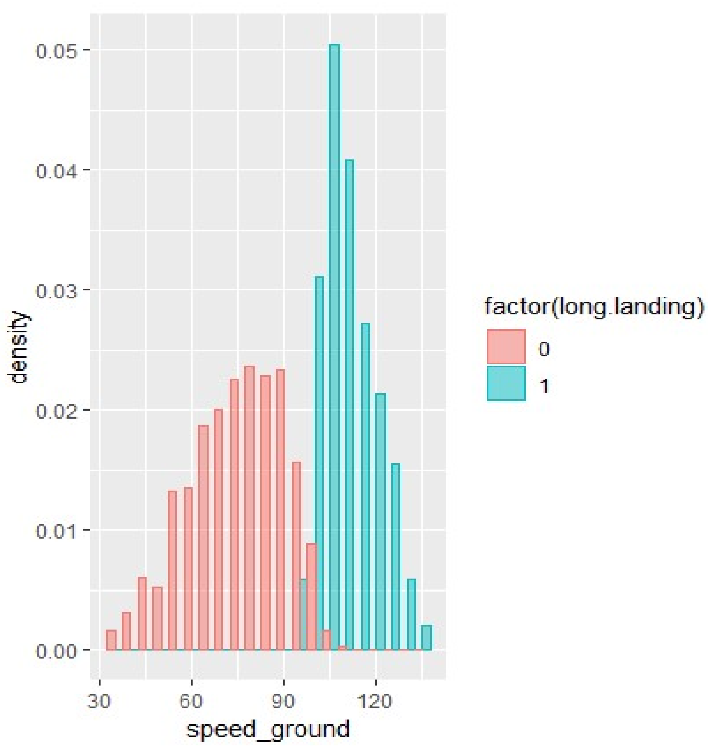
```
##########
# step 4 #
##########

plot(data$long.landing ~ data$speed_ground)
plot(jitter(long.landing,0.1)~jitter(speed_ground),data,xlab="speed ground",ylab="long
landing")
ggplot(data <- data,aes(x=speed_ground,fill=factor(long.landing)))+
  geom_histogram(position="dodge",binwidth=5,aes(y=..density..,
                                               colour=factor(long.landing)),alpha = 0.5)

plot(data$long.landing ~ data$speed_air)
plot(jitter(long.landing,0.1)~jitter(speed_air),data,xlab="speed air",ylab="long
landing")
ggplot(data <- data,aes(x=speed_air,fill=factor(long.landing)))+
  geom_histogram(position="dodge",binwidth=5,aes(y=..density..,
                                                 colour=factor(long.landing)),alpha =
0.5)

plot(jitter(long.landing,0.1)~jitter(as.numeric(aircraft)),data,xlab="aircraft
Boeing",ylab="long landing")

plot(data$long.landing ~ data$pitch)
plot(jitter(long.landing,0.1)~jitter(pitch),data,xlab="pitch",ylab="long landing")
ggplot(data <- data,aes(x=pitch,fill=factor(long.landing)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
                                                 colour=factor(long.landing)),alpha =
0.5)
```

Statistical Modeling Part2
Ashwita Saxena
M06119969

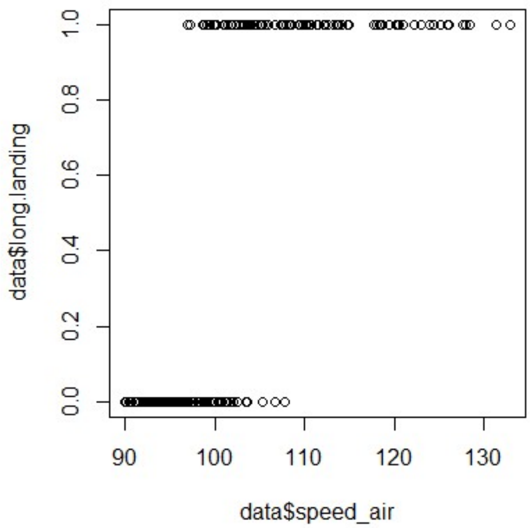
1. Speed Ground

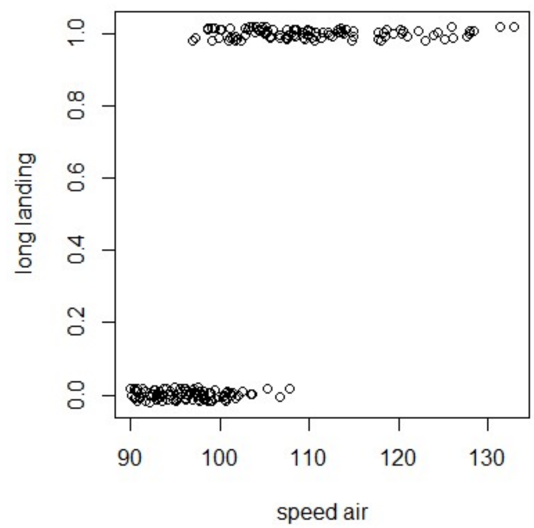


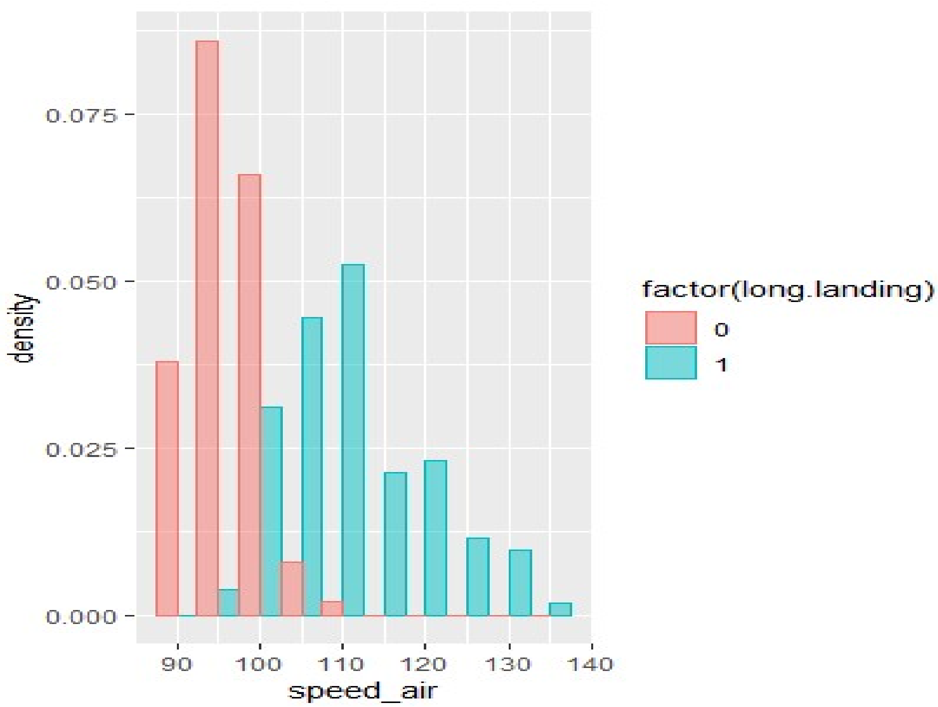




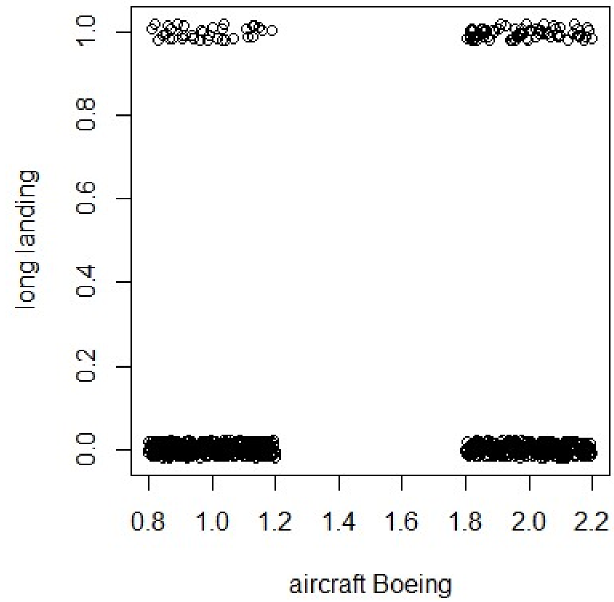
2. Speed Air



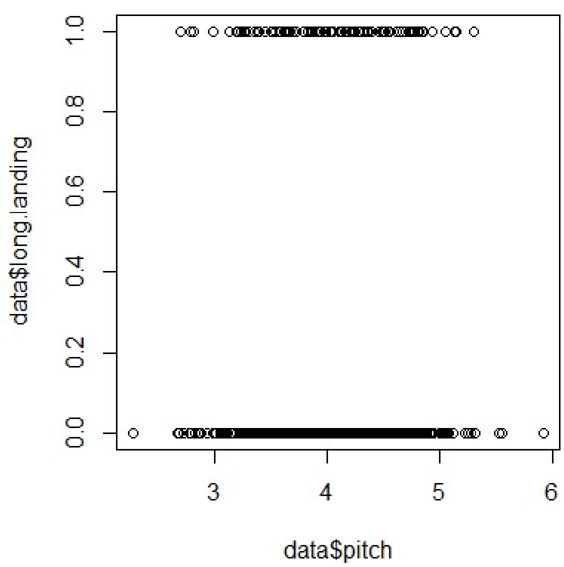


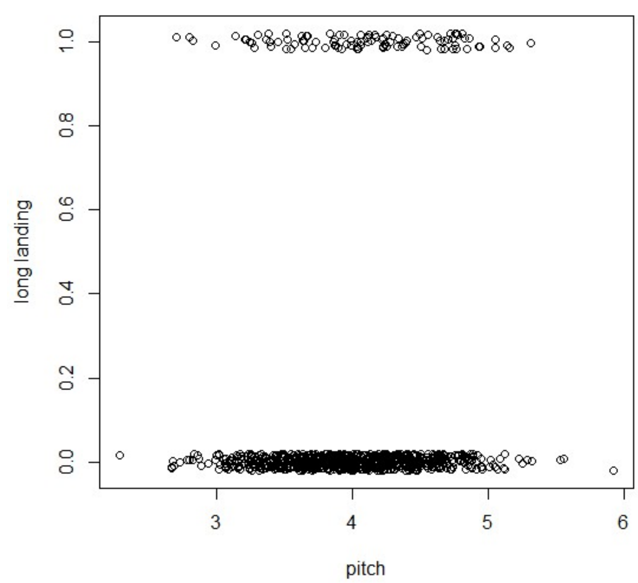


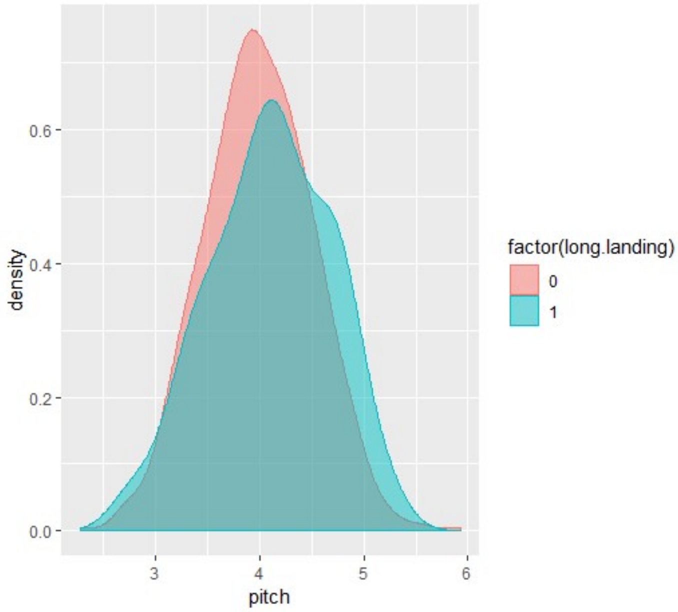
3. Aircraft Boeing


long landing
aircraft Boeing
0.0
0.2
0.4
0.6
0.8
1.0
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2

4. Pitch


data$long.landing
data$pitch
0.0
0.2
0.4
0.6
0.8
1.0
3
4
5
6


long landing
pitch
0.0
0.2
0.4
0.6
0.8
1.0
3
4
5
6



Observation: Based on the plots above, we can infer that as speed ground and speed air exceeds 90 mph, it is more likely for a flight to have a longer landing. In case of aircraft, Boeing has more long landing flights as compared to airbus. In case of pitch, there is not much difference in the pitch when looked at long landing. If the pitch is high, or low, both can result in a longer distance.

Conclusion: We can see significant effect of speed ground, speed air and pitch on the longer landing flights. Pitch however doesn't show much difference.

Step 5. Based on the analysis results in Steps 3-4 and the collinearity result seen in Step 16 of Part 1, initiate a "full" model. Fit your model to the data and present your result.

```
##########
# step 5 #
##########

step5 <- glm(long.landing ~ speed_ground+ aircraft+pitch, data = data, family =
binomial(link='logit'))
summary(step5)
```

```
Call:
glm(formula = long.landing ~ speed_ground + aircraft + pitch,
    family = binomial(link = "logit"), data = data)

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 ***
aircraftboeing   3.04348    0.73345   4.150 3.33e-05 ***
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
```

Observation: we formulated a model based on significant variables found in step 3-4. In those steps, pitch was significant at 95% significance level, however, in this logit model, pitch is significant only at 90 significance level.

Conclusion: Single variable linear model gives us pitch as a highly significant variable. However, a multi-variable logistic regression model shows that pitch is not significant at 95% significance level, but only at 90% significance level.

Step 6. Use the R function "Step" to perform forward variable selection using AIC. Compare the result with the table obtained in Step 3. Are the results consistent?

```
##########
# step 6 #
##########

nullmodel <- glm(long.landing ~ 1, data = data, family=binomial(link='logit'))
fullmodel <- glm(long.landing ~ speed_ground + pitch+ height + no_pasg+ duration +
aircraft, data=data, family= binomial(link = 'logit'))

#forward selection
model_step_f <- step(nullmodel, scope=list(lower=nullmodel, upper=fullmodel),
direction='forward')
summary(model_step_f)
```

```
Call:
glm(formula = long.landing ~ speed_ground + aircraft + height +
    pitch, family = binomial(link = "logit"), data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.20284  -0.00054   0.00000   0.00000   2.35719

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -119.77598   24.41821  -4.905 9.33e-07 ***
speed_ground       1.02266    0.20290   5.040 4.65e-07 ***
aircraftboeing     5.13443    1.18091   4.348 1.37e-05 ***
height             0.25795    0.06861   3.760  0.00017 ***
pitch              1.53751    0.84109   1.828  0.06755 .
---
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:  53.204  on 826  degrees of freedom
AIC: 63.204

Number of Fisher Scoring iterations: 12
```

Observation: According to Step AIC, the best model has speed ground, aircraft type boeing, height and pitch as the most important variables. We also see that height is highly significant in this model, whereas it is not significant in the single variable model in step3.Also, pitch is significant at 90% significance level whereas in the dingle variable model, it was significant at 95% level of significance.

Conclusion: AIC forward selection gives us height as a highly important variable to determine long landing along with speed ground and aircraft type boeing. Pitch is significant only at 90% significance level

Step 7. Use the R function “Step” to perform forward variable selection using BIC. Compare the result with that from the previous step.

```
##########
# step 7 #
##########

model_bic <- step(nullmodel, scope=list(lower=nullmodel, upper=fullmodel),
direction='forward', k=log(nrow(data)))
summary(model_bic)
```

```
Call:
glm(formula = long.landing ~ speed_ground + aircraft + height,
    family = binomial(link = "logit"), data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.43442  -0.00117   0.00000   0.00000   2.57435

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -102.95437   19.22882  -5.354 8.59e-08 ***
speed_ground     0.92657    0.17242   5.374 7.70e-08 ***
aircraftboeing   5.04813    1.11520   4.527 5.99e-06 ***
height           0.23106    0.05959   3.877 0.000106 ***
---
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:  57.047  on 827  degrees of freedom
AIC: 65.047

Number of Fisher Scoring iterations: 11
```

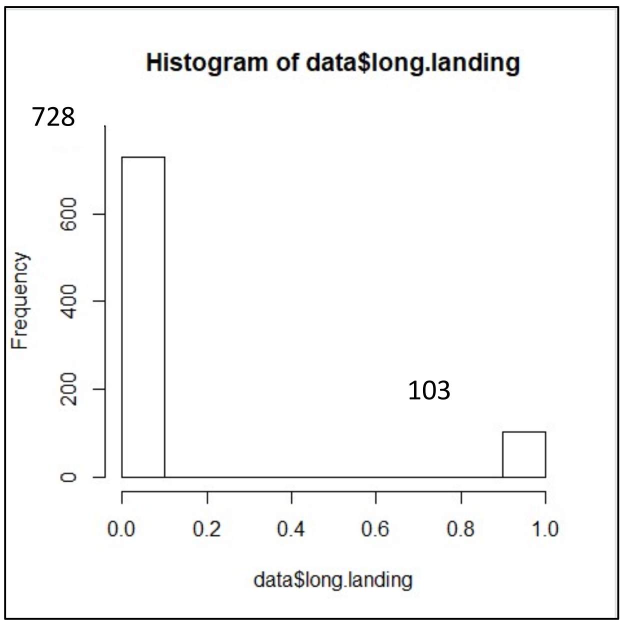
Observation: When we did BIC model selection, we can see that speed ground, aircraft type boeing and height are very significant and important variables in determining long landing. We observe that pitch is no longer included in the model and hence is not important in determining the long landing

Conclusion: Below is a comparison of both the models. Pitch has been excluded from the final model. It is no longer important. Speed ground, height and aircraft type Boeing are all highly significant variables. For our final model, we will choose the BIC step model with speed ground, aircraft type boeing and height. It has a lower BIC value as compared to the AIC step model and we choose BIC as a selection criteria as it gives simpler model.

Long Landing	AIC Step Model					BIC Step Model				
Deviance Residuals	Min	1Q	Median	3Q	Max	Min	1Q	Median	3Q	Max
	-2.20284	-0.00054	0	0	2.35719	-2.43442	-0.00117	0	0	2.57435
Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	Significance	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	-119.776	24.41821	-4.905	9.33E-07	***	-102.954	19.22882	-5.354	8.59E-08	***
speed_ground	1.02266	0.2029	5.04	4.65E-07	***	0.92657	0.17242	5.374	7.70E-08	***
aircraftboeing	5.13443	1.18091	4.348	1.37E-05	***	5.04813	1.1152	4.527	5.99E-06	***
height	0.25795	0.06861	3.76	0.00017	***	0.23106	0.05959	3.877	0.000106	***
pitch	1.53751	0.84109	1.828	0.06755	.	NA	NA	NA	NA	NA
Null deviance on degrees of freedom	622.778					622.778				
Residual deviance on degrees of freedom	53.204					57.047				
AIC	63.204					65.047				
BIC	86.81731					83.9372				

Step 8. You are scheduled to meet with an FAA agent who wants to know "what are risk factors for long landings and how do they influence its occurrence?". For your presentation, you are only allowed to show: One model • One table • No more than three figures • No more than five bullet statements. Please use statements that she can understand. The question is: what model/table/figures/statements would you include in your presentation. Be selective!

- We are calculating the impact of variables on long landing. Out of all the flights, 103 flights have a long landing (distance > 2500) and 728 do not.

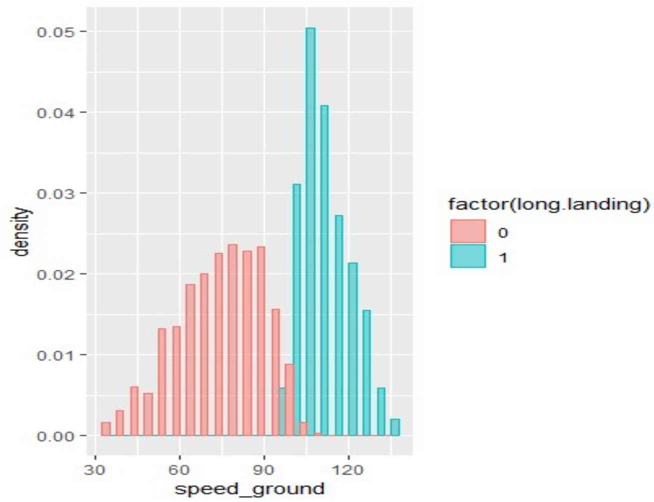


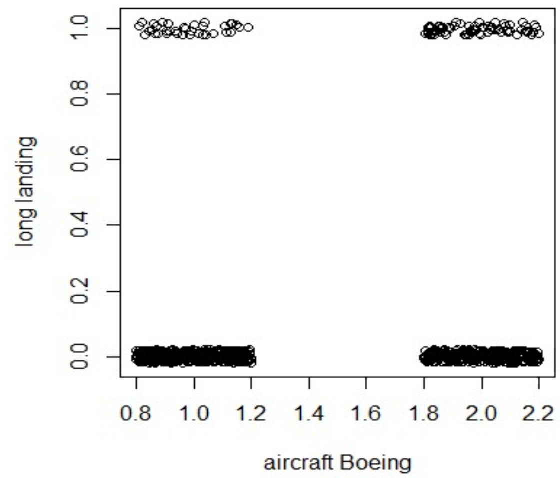
- Based on our analysis, we infer that ground speed, height of aircraft while passing over runway, and aircraft type boeing are most important variables in predicting the long landing of a plane.

Variable	exponential of coefficient
speed_ground	2.525830703
aircraftboeing	155.7309751
height	1.259934833

One mile per hour increase in ground speed increases the odds of long landing by 152.56%
One meter increase in the height increases the odds of long landing by 25%
When aircraft type changes from airbus to boeing, the odds ratio is 155.73.

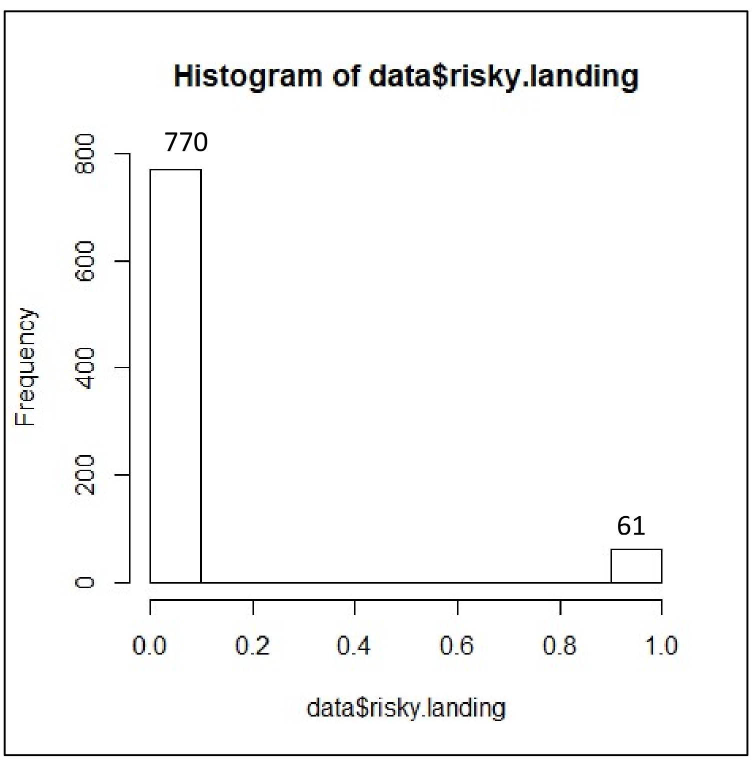
- This relationship can also be seen in the following visualizations





Step 9. Repeat Steps 1-7 but using "risky.landing" as the binary response.

```
par(mfrow=c(1,1))
hist(data$risky.landing,ylim = range(0,800))
```



According to this histogram, there are 770 non risky (coded 0) and 61 risky (coded1) flights.

```
data1 <- as.data.frame(data)
for (i in c(1:7)){
  model <- glm (risky.landing ~ data1[,i],
                family = binomial(link='logit'),
                data=data1)
```

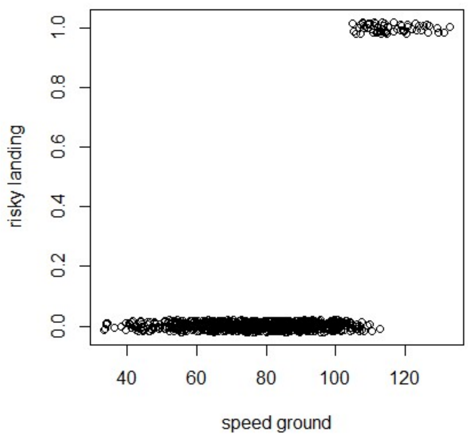
```
  print(summary(model))
}
```

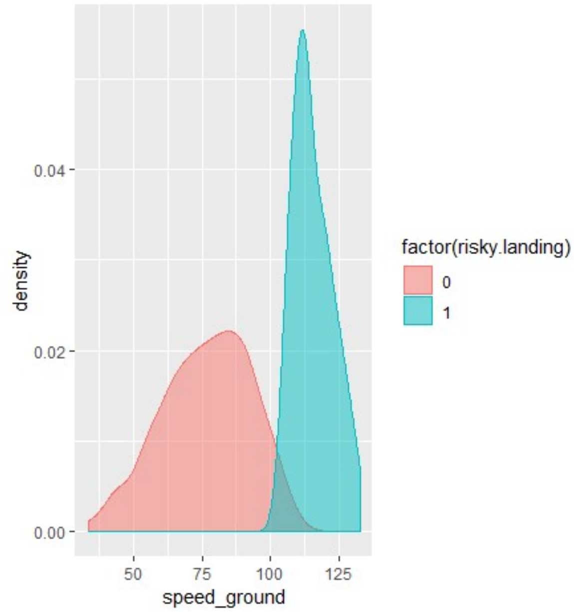
Rank	Variable	coefficient	Odds Ratio	Direction of coefficient	P-Value
1	speed_ground	0.6142187	1.8482121	positive	6.90E-08
2	speed_air	0.8704019	2.3878703	positive	3.73E-06
3	Aircraft (boeing)	1.0017753	2.723112	positive	0.000456056
4	pitch	0.371072	1.4492874	positive	0.143296135
5	no_pasg	0.0253793	0.97494	negative	0.153623692
6	duration	0.0011518	0.9988488	negative	0.680198706
7	height	0.0022186	0.9977839	negative	0.870591704

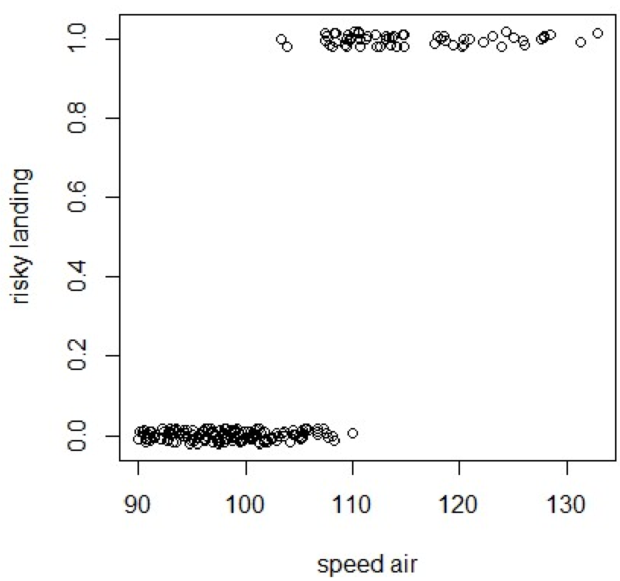
According to single variable logistic models (individual models), we can see that speed ground is the most significant, followed by speed_air abd aircraft type boeing. Pitch, number of passengers, duration and height are not significant.

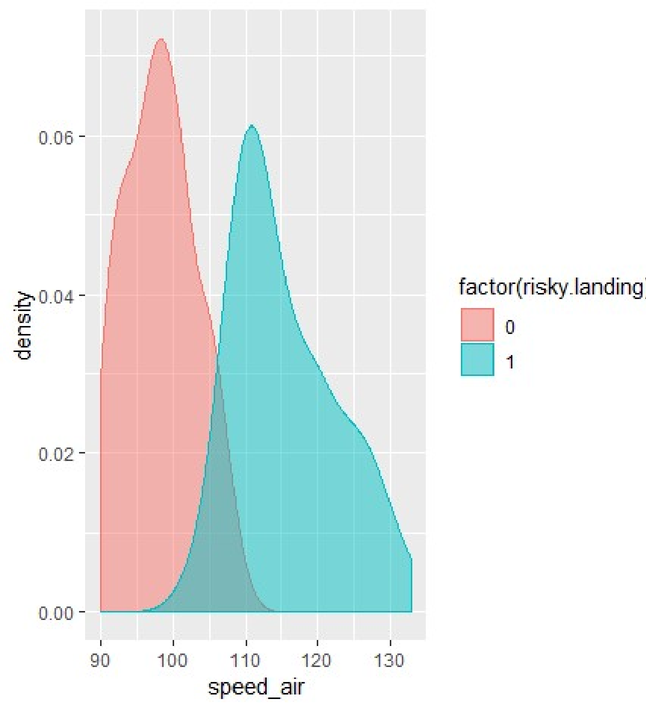
```
plot(jitter(risky.landing,0.1)~jitter(speed_ground),data,xlab="speed ground",ylab="risky
landing")
ggplot(data <- data,aes(x=speed_ground,fill=factor(risky.landing)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
                                                         colour=factor(risky.landing)),alpha =
0.5)

plot(jitter(risky.landing,0.1)~jitter(speed_air),data,xlab="speed air",ylab="risky
landing")
ggplot(data <- data,aes(x=speed_air,fill=factor(risky.landing)))+
  geom_density(position="dodge",binwidth=5,aes(y=..density..,
                                                         colour=factor(risky.landing)),alpha =
0.5)

plot(jitter(risky.landing,0.1)~jitter(as.numeric(aircraft)),data,xlab="aircraft
Boeing",ylab="risky landing")
```

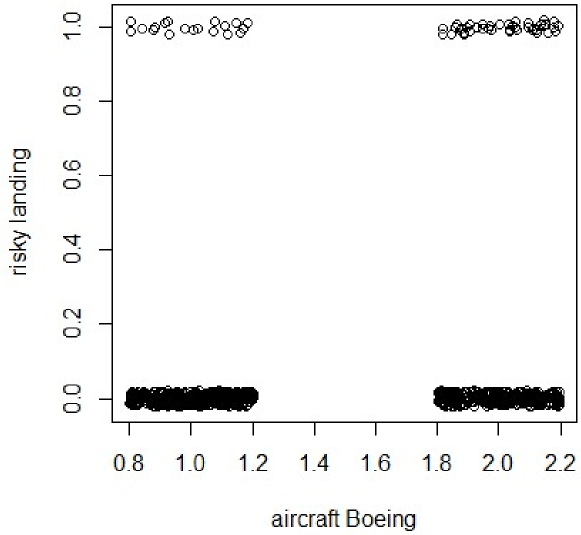
Statistical Modeling Part2
Ashwita Saxena
M06119969












```
step5_r <- glm(risky.landing ~ speed_ground + aircraft, data = data,
family=binomial(link='logit'))
summary(step5_r)
Call:
glm(formula = risky.landing ~ speed_ground + aircraft, family
= binomial(link = "logit"),
    data = data)

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
```

In the model we did not include speed air because of multicollinearity between speed air and speed ground. We can see that both speed ground and aircraft type boeing are significant variables when predicting factors that impact risky landing.

```
# important variable model
step5_r <- glm(risky.landing ~ speed_ground + aircraft, data = data,
family=binomial(link='logit'))
summary(step5_r)

nullmodel_r <- glm(risky.landing ~ 1, data = data, family=binomial(link='logit'))
fullmodel_r <- glm(risky.landing ~ speed_ground + pitch+ height + no_pasg+ duration +
aircraft, data=data, family= binomial(link = 'logit'))
```

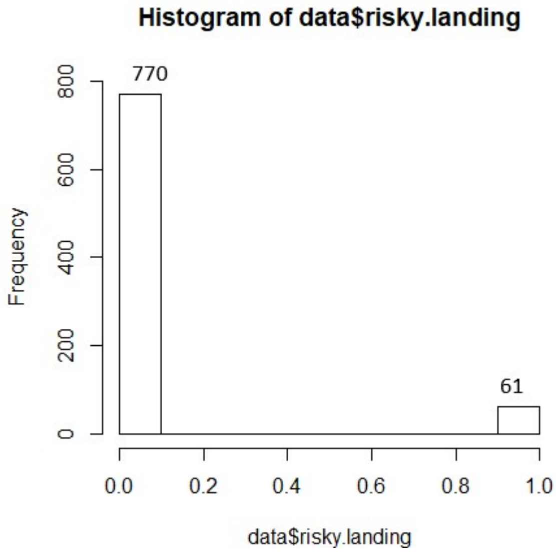
```
#forward AIC
model_step_f_r <- step(nullmodel_r, scope=list(lower=nullmodel_r, upper=fullmodel_r),
direction='forward')
summary(model_step_f_r)
BIC(model_step_f_r)

# forward BIC
model_bic_r <- step(nullmodel_r, scope=list(lower=nullmodel_r, upper=fullmodel_r),
direction='forward', k=log(nrow(data1)))
summary(model_bic_r)
BIC(model_bic_r)
```

Risky Landing	Forward AIC					Step Forward BIC for Risky Landing				
Deviance Residuals	Min	1Q	Median	3Q	Max	Min	1Q	Median	3Q	Max
	-2.33913	-0.00009	0	0	1.8781	-2.24398	-0.00011	0	0	1.61021
Coefficients	Estimate	Std. Error	z value	Pr(>|z|)	Significance	Estimate	Std. Error	z value	Pr(>|z|)	Significance
(Intercept)	-99.9078	25.57993	-3.906	9.39E-05	***	-102.077	24.7751	-4.12	3.79E-05	***
speed_ground	0.94963	0.23559	4.031	5.56E-05	***	0.9263	0.2248	4.121	3.78E-05	***
aircraftboeing	4.64188	1.4752	3.147	1.65E-03	**	4.019	1.2494	3.217	1.30E-03	**
no_pasg	-0.08462	0.05732	-1.476	0.13987		NA	NA	NA	NA	NA
Null deviance on degrees of freedom	436.043 on 830					436.043 on 830				
Residual deviance on degrees of freedom	37.707 on 827					40.097 on 828				
AIC	45.707					46.097				
BIC	64.59746					60.26449				

According to the AIC and BIC forward selection models, we see that the AIC model includes number of passengers in the final model however it is not significant at any significance level. The BIC model has speed ground and aircraft type as the most important predictors of risky landing. It also gives a smaller BIC value and a simpler model.

Step 10. You are scheduled to meet with an FAA agent who wants to know "what are risk factors for risky landings and how do they influence its occurrence?". For your presentation, you are only allowed to show: • One model • One table 8 • No more than three figures • No more than five bullet statements. Please use statements that she can understand.

- We are calculating the impact of variables on long landing. Out of all the flights, 61 flights have a long landing (distance > 3000) and 770 do not.

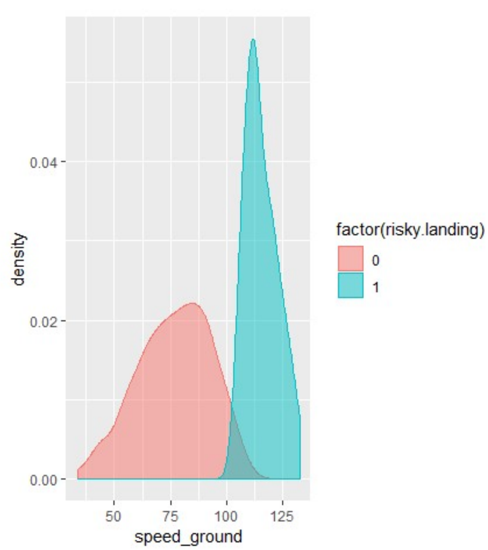


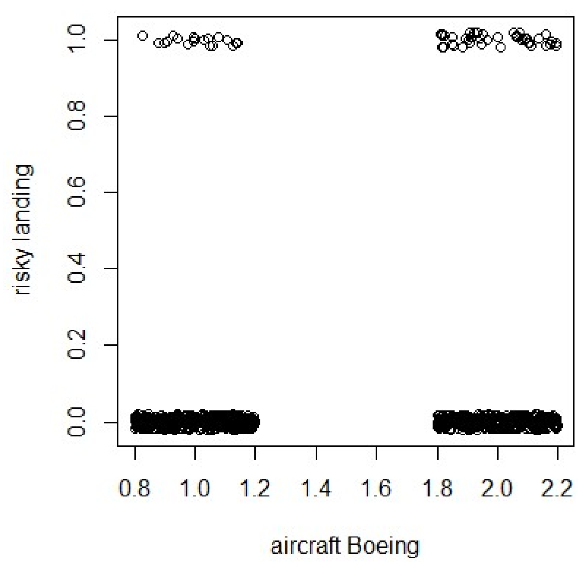
- Based on our analysis, we infer that ground speed and aircraft type boeing are most important variables in predicting the risky landing of a plane.

Variable	exponential of coefficient
speed_ground	2.525148821
aircraftboeing	55.64543256

One mile per hour increase in ground speed increases the odds of risky landing by 152.51%
When aircraft type changes from airbus to boeing, the odds ratio for risky landing is 55.64.

- This relationship can also be seen in the following visualizations





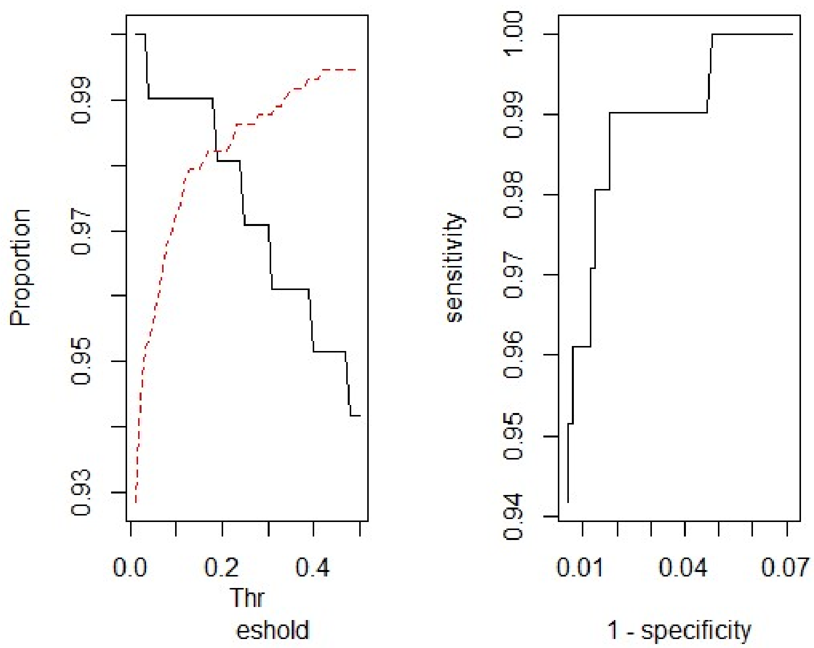
Step 11. Use no more than three bullet statements to summarize the difference between the two models.

- Ground speed and aircraft type Boeing are both the most important factors to determine both long landing and risky landing of an airplane.
- Height of an aircraft over the runway is an important predictor of long landing however it is not important in determining risky landing
- Ground speed has the same impact on risky landing as well as long landing. One mile per hour increase in ground speed increases the odds of risky landing by 152.5%

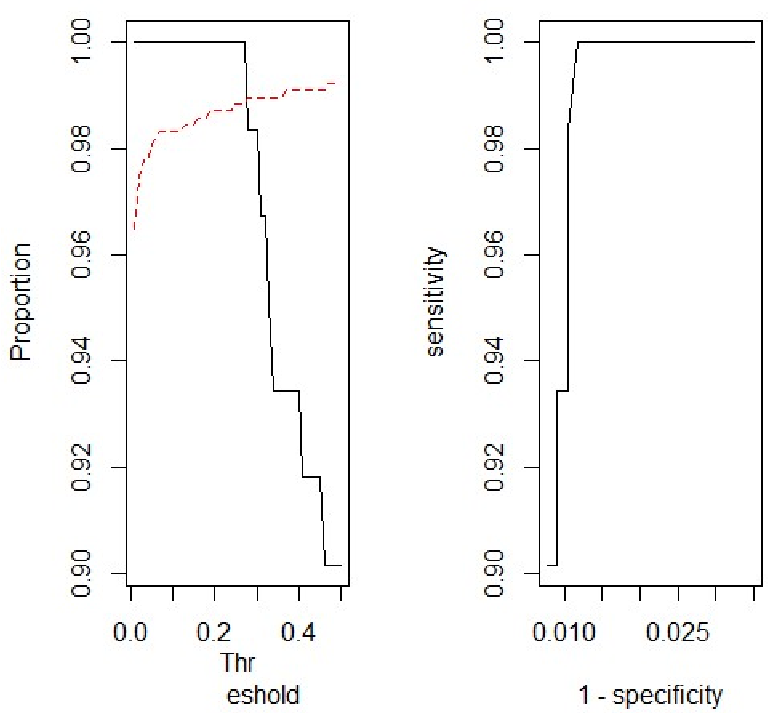
Step 12. Plot the ROC curve (sensitivity versus 1-specificity) for each model (see pp.32-33 in Lecture 4 slides). Draw the two curves in the same plot. Do you have any comment?

```
###########
# step 12 #
###########

## Long landing model
pred <- ifelse(predict(model_bic,type = 'response') < 0.5,0,1)
pred_r <- ifelse(predict(model_bic_r,type = 'response') < 0.5,0,1)

thresh <- seq(0.01,0.5,0.01)
sensitivity <- specificity <- rep(NA,length(thresh))
for( j in seq(along=thresh)) {
  pp<- ifelse(predict(model_bic,type = 'response') < thresh[j],0,1)
  xx<-xtabs(~data1$long.landing+pp)
  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="Thr
        eshold",ylab="Proportion",lty=1:2)
plot(1-specificity,sensitivity,type="l");abline(0,1,lty=2)

### risky landing model

pred <- ifelse(predict(model_bic,type = 'response') < 0.5,0,1)
pred_r <- ifelse(predict(model_bic_r,type = 'response') < 0.5,0,1)

thresh <- seq(0.01,0.5,0.01)
sensitivity <- specificity <- rep(NA,length(thresh))
for( j in seq(along=thresh)) {
  pp<- ifelse(predict(model_bic_r,type = 'response') < thresh[j],0,1)
  xx<-xtabs(~data1$risky.landing+pp)
  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="Thr
        eshold",ylab="Proportion",lty=1:2)
plot(1-specificity,sensitivity,type="l");abline(0,1,lty=2)
```

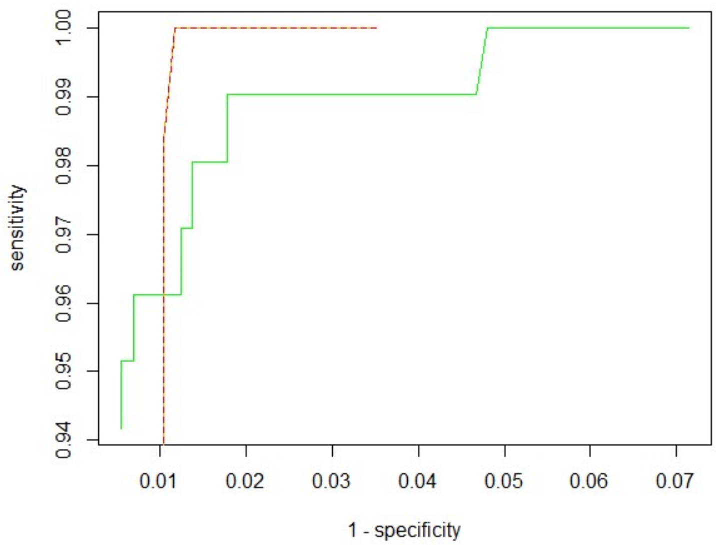
Statistical Modeling Part2
Ashwita Saxena
M06119969



long landing model ROC Curve



risky landing model



The orange line depicts ROC curve of risky landing, whereas ROC curve of long landing is depicted by the green line.

Observation: Both the long landing and risky landing ROC curves are very close to the left and top borders, thus meaning that Area under the curve is large. This means our test is quite accurate. We can say that risky landing test is more accurate than the long landing test as from the graph shows it has a greater area under the curve as compared to long landing.

Conclusion: Risky landing model has a better prediction accuracy as compared to long landing model

Step 13. A commercial airplane is passing over the threshold of the runway, at this moment we have its basic information and measures of its airborne performance (Boeing, duration=200, no_pasg=80, speed_ground=115, speed_air=120, height=40, pitch=4). Predict its probability of being a long landing and a risky landing, respectively. Report the predicted probability as well as its 95% confidence interval.

```
###########
# step 13 #
###########


airplane <- data.frame(aircraft="boeing",duration=200,no_pasg=80,
                      speed_ground=115,speed_air=120,height=40,pitch=4)
pred1 <- predict(model_bic,newdata=airplane,type = 'response',se.fit = T)
pred2 <- predict(model_bic_r,newdata=airplane,type='response' ,se.fit = T)

c(pred1$fit,pred1$fit-1.96*pred1$se.fit[1],pred1$fit+1.96*pred1$se.fit[1])

c(pred2$fit,pred2$fit-1.96*pred2$se.fit[1],pred2$fit+1.96*pred2$se.fit[1])
```

	Probability	Lower limit of CI	Upper limit of CI
long landing	1	0.9999999	1.0000001
risky landing	0.999789	0.998925	1.000653

Observation: This new data point acts as a out of sample data point which we use to test our model. We can see that the probability of this aircraft having a long landing is almost definite (100%). And it is also very likely to have a risky landing.

Conclusion. This aircraft has very high speed ground as well as is of the type Boeing. Hence according to our model's prediction, it is very likely to have a risky and long landing. We have statistical evidence to prove that this is true by using this flight's data in out model.

Step 14. For the binary response "risky landing", fit the following models using the risk factors identified in Steps 9-10: • Probit model • Hazard model with complementary log-log link Compare these two models with the logistic model. Do you have any comments?

```
###########
# step 14 #
###########

model_logit <- glm(risky.landing ~ speed_ground + aircraft, data = data,
family=binomial(link='logit'))
summary(model_logit)

model_probit <- glm(risky.landing ~ speed_ground + aircraft, data = data,
family=binomial(link='probit'))
summary(model_probit)
BIC(model_probit)

model_loglog <- glm(risky.landing ~ speed_ground + aircraft, data = data,
family=binomial(link='cloglog'))
summary(model_loglog)
BIC(model_loglog)
```

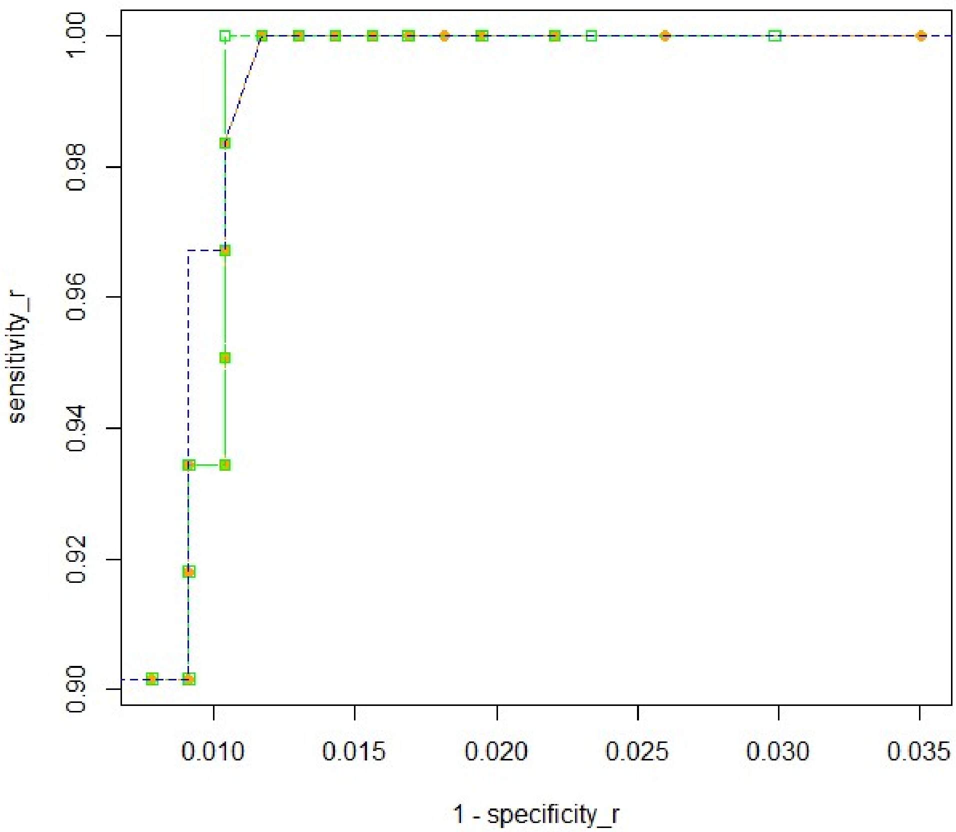
Risky Landing	Logit model			Probit model			cloglog model		
Coefficients	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)	Estimate	Std. Error	Pr(>|z|)
(Intercept)	-102.0772	24.7751	3.79E-05	-58.6931	13.3133	1.04E-05	-69.2654	14.7396	2.61E-06
speed_ground	0.9263	0.2248	3.78E-05	0.5322	0.1207	1.03E-05	0.6221	0.1326	2.74E-06
Aircraftboeing	4.019	1.2494	1.30E-03	2.3567	0.7016	7.82E-04	2.8984	0.8002	2.92E-04
Null deviance on df	436.043 on 830			436.043 on 830			436.043 on 830		
Residual deviance on df	40.097 on 828			39.436 on 828			41.443 on 828		
AIC	46.097			45.436			47.443		
BIC	60.26449			59.60437			61.6113		

Observation: When comparing logit, probit and cloglog models, we see that probit model has the smallest AIC and BIC values as compared to the other two. It also has the smallest standard errors. Given these three, we can say that probit model is the best model. We also see that cloglog model gives

Conclusion: Probit model is the best model for this dataset as compared to logit and cloglog models

Step 15. Compare the three models by showing their ROC curves in the same plot (see Step 12).

```
###########
# step 15 #
###########

thresh <- seq(0.01,0.5,0.01)
sensitivity_r <- specificity_r <- rep(NA,length(thresh))
for( j in seq(along=thresh)) {
  pp<- ifelse(predict(model_logit,type = 'response') < thresh[j],0,1)
  xx<-xtabs(~data1$risky.landing+pp)
  specificity_r[j]<-xx[1,1]/(xx[1,1]+xx[1,2])
  sensitivity_r[j]<-xx[2,2]/(xx[2,1]+xx[2,2])
}

thresh <- seq(0.01,0.5,0.01)
sensitivity_probit <- specificity_probit <- rep(NA,length(thresh))
for( j in seq(along=thresh)) {
  pp<- ifelse(predict(model_probit,type = 'response') < thresh[j],0,1)
  xx<-xtabs(~data1$risky.landing+pp)
  specificity_probit[j]<-xx[1,1]/(xx[1,1]+xx[1,2])
  sensitivity_probit[j]<-xx[2,2]/(xx[2,1]+xx[2,2])
}

thresh <- seq(0.01,0.5,0.01)
sensitivity_loglog <- specificity_loglog <- rep(NA,length(thresh))
for( j in seq(along=thresh)) {
  pp<- ifelse(predict(model_loglog,type = 'response') < thresh[j],0,1)
  xx<-xtabs(~data1$risky.landing+pp)
  specificity_loglog[j]<-xx[1,1]/(xx[1,1]+xx[1,2])
  sensitivity_loglog[j]<-xx[2,2]/(xx[2,1]+xx[2,2])
}

par(mfrow=c(1,1))
plot(1-specificity_r,sensitivity_r, type ="p", col="orange")
points(1-specificity_r,sensitivity_r,type="p",col="orange", pch =19)
lines(1-specificity_r,sensitivity_r, col="orange",lty=2)
points(1-specificity_probit,sensitivity_probit,type="b",col="green", pch = 22)
lines(1-specificity_probit,sensitivity_probit, col="green",lty=2)
lines(1-specificity_loglog,sensitivity_loglog, col="blue",lty=2)
```



Observation: The orange line and points show ROC curve of the logit model. The green points and line show the ROC curve for probit model and the blue dotted line shows the ROC curve of cloglog model. When we compare the three, we see some part of each ROC curve overlapping. The logit model AUC is the smallest as it gets overlapped by probit and cloglog curve. When we compare probit and cloglog curves, wee see that cloglog curve has additional area under the curve (left rectangle) which is bigger than the additional area that probit has (top triange).

Conclusion:The cloglog ROC curve has the most AUC. Hence it most accurately predicts risky landing.

Step 16. Use each model to identify the top 5 risky landings. Do they point to the same flights?

```
###########
# step 16 #
###########

pred_logit <- predict(model_logit,type = 'response')
summary(model_logit)
pred_probit <- predict(model_probit,type = 'response')
summary(model_probit)
pred_loglog <- predict(model_loglog,type = 'response')
summary(model_loglog)
```

```
data1[as.numeric(names(tail(sort(pred_logit),5))),]
data1[as.numeric(names(tail(sort(pred_probit),5))),]
data1[as.numeric(names(tail(sort(pred_loglog),5))),]
```

```
> data1[as.numeric(names(tail(sort(pred_logit),5))),]
    aircraft no_pasg speed_ground speed_air   height    pitch  duration long.landing risky.landing
408   airbus      60     131.0352  131.3379 28.27797 3.660194 131.73110            1             1
387   boeing      61     126.8393  126.1186 20.54783 4.334558 153.83445            1             1
64    boeing      72     129.2649  128.4177 33.94900 4.139951 161.89247            1             1
307   boeing      67     129.3072  127.5933 23.97850 5.154699 154.52460            1             1
362   boeing      52     132.7847  132.9115 18.17703 4.110664  63.32952            1             1
> data1[as.numeric(names(tail(sort(pred_probit),5))),]
    aircraft no_pasg speed_ground speed_air   height    pitch  duration long.landing risky.landing
362   boeing      52     132.7847  132.9115 18.17703 4.110664  63.32952            1             1
383   boeing      61     121.8371  120.9534 33.18460 3.867476  99.68150            1             1
387   boeing      61     126.8393  126.1186 20.54783 4.334558 153.83445            1             1
408   airbus      60     131.0352  131.3379 28.27797 3.660194 131.73110            1             1
643   airbus      66     126.2443  127.9371 35.17570 2.701924 137.58573            1             1
> data1[as.numeric(names(tail(sort(pred_loglog),5))),]
    aircraft no_pasg speed_ground speed_air   height    pitch  duration long.landing risky.landing
643   airbus      66     126.2443  127.9371 35.17570 2.701924 137.58573            1             1
669   airbus      75     120.4189  118.4847 31.26345 2.796731 140.45311            1             1
751   airbus      49     125.2123  125.1385 22.52478 4.365772 175.51443            1             1
765   airbus      61     120.5579  118.2882 15.66566 4.111265 220.05713            1             1
769   airbus      66     123.3105  124.3908 22.32718 4.276710  98.50031            1             1
```

Risky flights	**logit**	**probit**	**cloglog**
1	408	362	643
2	387	383	669
3	64	387	751
4	307	408	765
5	362	643	769

Observation: when we compare the top 5 risky landings from logit, probit and cloglog model, we see that observation number 408, 387 are common in logit and probit model. Also, observation number 643 is common between probit and cloglog model. There are no common observations in the top 5 risky landings between logit and cloglog models.

Conclusion: 2 flights are common in logit and probit top 5 risky flights and one flight is common between probit and cloglog top 5 risky flights.

Step 17. Use the probit model and hazard model to make prediction for the flight described in Step 13. Report the predicted probability as well as its 95% confidence interval. Compare the results with that from Step 13.

```
###########
# step 17 #
###########
model_logit_l <- glm(long.landing ~ speed_ground + aircraft + height, data = data1,
family=binomial(link='logit'))
summary(model_logit_l)
BIC(model_logit_l)
```

```
model_probit_l <- glm(long.landing ~ speed_ground + aircraft + height, data = data1,
family=binomial(link='probit'))
summary(model_probit_l)
BIC(model_probit_l)

model_loglog_l <- glm(long.landing ~ speed_ground + aircraft + height, data = data1,
family=binomial(link='cloglog'))
summary(model_loglog_l)
BIC(model_loglog_l)

model_logit_r <- model_logit
model_probit_r <- model_probit
model_loglog_r <- model_loglog


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

## logit for long and risky
pred_logit_l <- predict(model_logit_l,newdata=airplane,type = 'response',se.fit = T)
pred_logit_r <- predict(model_logit_r,newdata=airplane,type='response' ,se.fit = T)

c(pred_logit_l$fit,pred_logit_l$fit-
1.96*pred_logit_l$se.fit[1],pred_logit_l$fit+1.96*pred_logit_l$se.fit[1])

c(pred_logit_r$fit,pred_logit_r$fit-
1.96*pred_logit_r$se.fit[1],pred_logit_r$fit+1.96*pred_logit_r$se.fit[1])

## probit for long and risky
pred_probit_l <- predict(model_probit_l,newdata=airplane,type = 'response',se.fit = T)
pred_probit_r <- predict(model_probit_r,newdata=airplane,type='response' ,se.fit = T)

c(pred_probit_l$fit,pred_probit_l$fit-
1.96*pred_probit_l$se.fit[1],pred_probit_l$fit+1.96*pred_probit_l$se.fit[1])

c(pred_probit_r$fit,pred_probit_r$fit-
1.96*pred_probit_r$se.fit[1],pred_probit_r$fit+1.96*pred_probit_r$se.fit[1])

## Hazard for long and risky

pred_loglog_l <- predict(model_loglog_l,newdata=airplane,type = 'response',se.fit = T)
pred_loglog_r <- predict(model_loglog_r,newdata=airplane,type='response' ,se.fit = T)

c(pred_loglog_l$fit,pred_loglog_l$fit-
1.96*pred_loglog_l$se.fit[1],pred_loglog_l$fit+1.96*pred_loglog_l$se.fit[1])

c(pred_loglog_r$fit,pred_loglog_r$fit-
1.96*pred_loglog_r$se.fit[1],pred_loglog_r$fit+1.96*pred_loglog_r$se.fit[1])
```

	Model	Probability	Lower limit of CI	Upper limit of CI
Long landing	**logit**	1	0.9999999	1.0000001
	probit	1	1	1
	cloglog	1	1	1
	Model	**Probability**	**Lower limit of CI**	**Upper limit of CI**
Risky Landing	**logit**	0.999789	0.998925	1.000653
	probit	0.9999994	0.9999933	1.0000056
	cloglog	1	1	1

Observation:

As we can see from the table above, probit model and cloglog model have similar prediction power for the new data point as compared to logit model for risky as well as long landing model. The difference in all the predictions is very small, so we cannot conclude whether one way is better than the other.

Conclusion: All the links give us a similar probability and confidence interval to predict the new data point.