Flight Landing Prediction Project: Part 3

Modeling Multinomial Data and Count Data

Q1: Modeling Multinomial Data

Data Import and Variable Creation

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

Create a multinomial response variable

A multinomial response variable is created based on the distance variable and these rules:

Y = 1 if distance < 1000

Y = 2 if 1000 <= distance < 2500

Y = 3 otherwise

The new multinomial response variable is named **multi_dist**. The continuous variable titled "distance" is discarded afterwards. We also assume that we do not know the order of the multinomial response variable.

For easier interpretation, we refer to Y=1 as "low distance", Y=2 as "medium distance", and Y=3 as "high distance".

```
#Create a multinomial response variable
clean_data$distance[clean_data$distance < 1000] <- 1
clean_data$distance[clean_data$distance >= 1000 &
                      clean_data$distance < 2500] <- 2
clean_data$distance[clean_data$distance > 2500] <- 3
multi_dist <- as.factor(clean_data$distance)

#Aircraft as a factor variable
clean_data$aircraft <- as.factor(clean_data$aircraft)

Data_logit <- cbind(clean_data[-8],multi_dist)
summary(Data_logit)
```

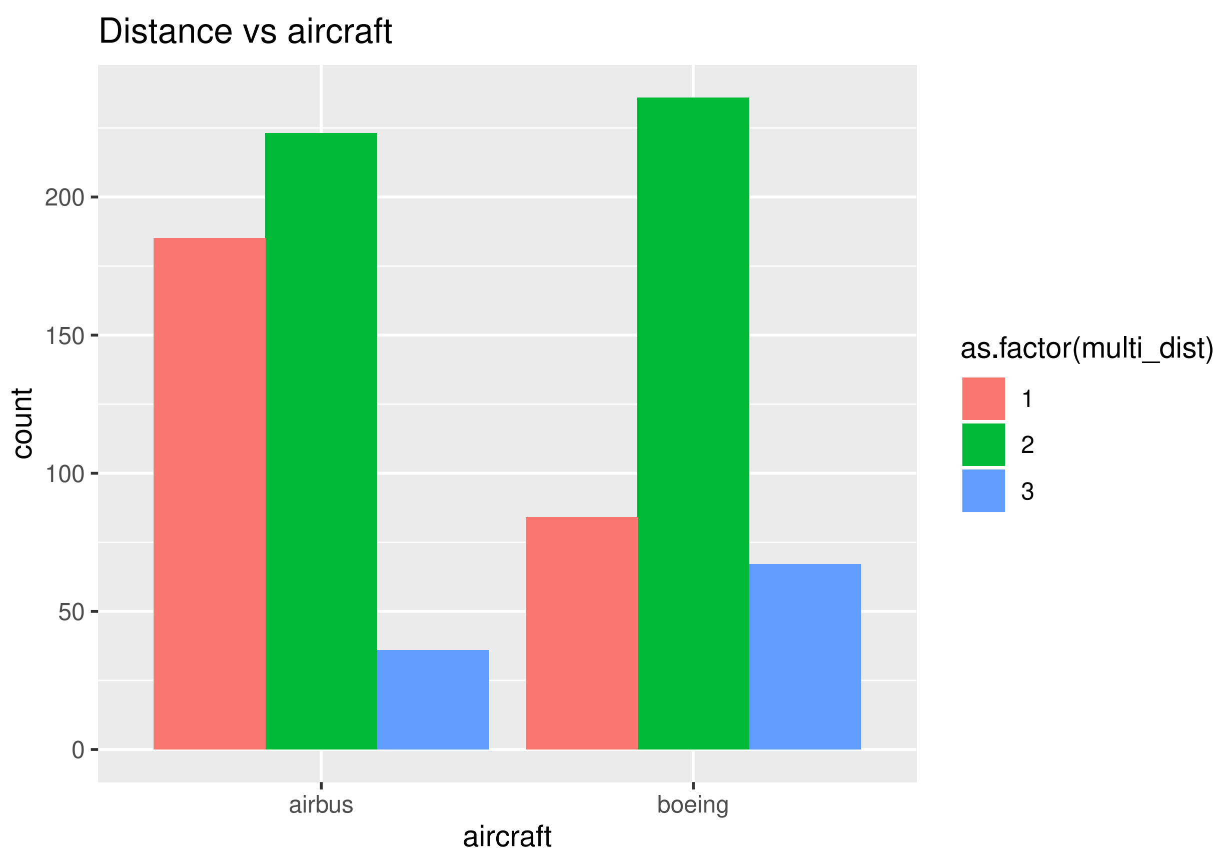
```
##    aircraft      duration         no_pasg       speed_ground
##  airbus:444   Min.   : 41.95   Min.   :29.00   Min.   : 33.57
##  boeing:387   1st Qu.:119.63   1st Qu.:55.00   1st Qu.: 66.20
##               Median :154.28   Median :60.00   Median : 79.79
##               Mean   :154.78   Mean   :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        multi_dist
##  Min.   : 90.00   Min.   : 6.228   Min.   :2.284   1:269
##  1st Qu.: 96.23   1st Qu.:23.530   1st Qu.:3.640   2:459
##  Median :101.12   Median :30.167   Median :4.001   3:103
##  Mean   :103.48   Mean   :30.458   Mean   :4.005
##  3rd Qu.:109.36   3rd Qu.:37.004   3rd Qu.:4.370
##  Max.   :132.91   Max.   :59.946   Max.   :5.927
##  NA's   :628
```

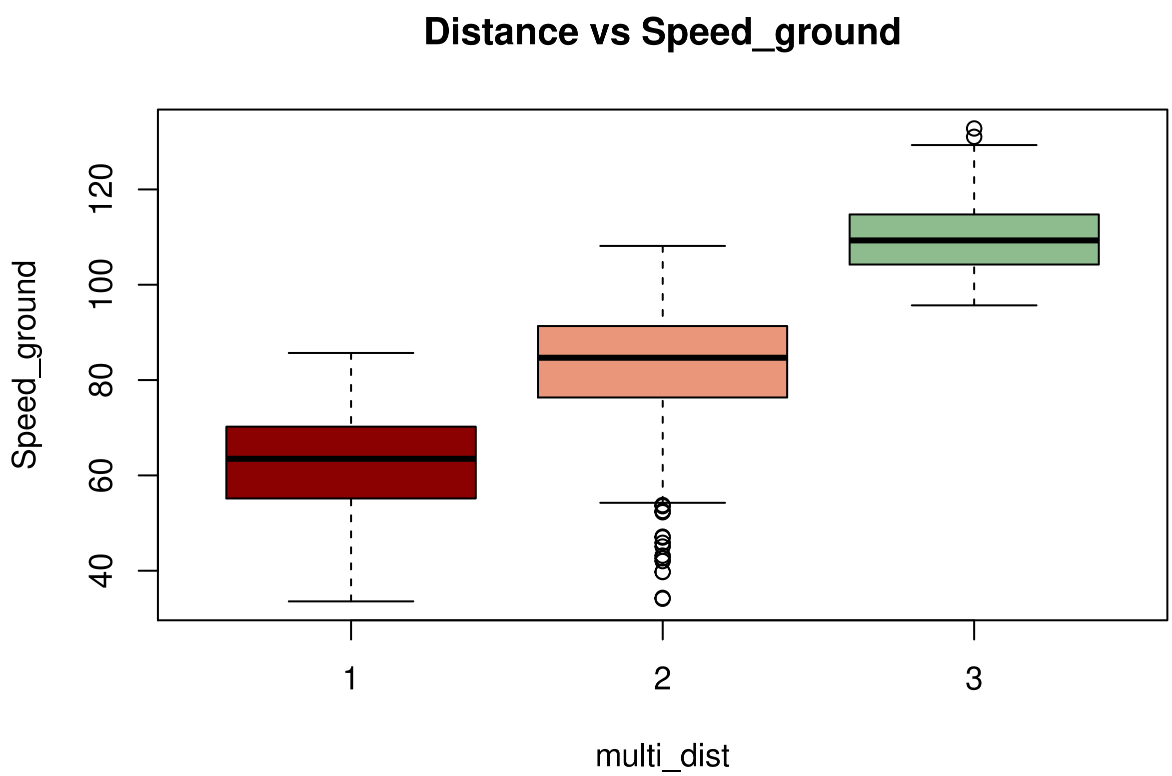
Data Visualization

The plot shows that Boeing aircrafts have higher landing distance and fall more in the "medium distance" and "high distance" categories compared to Airbus aircrafts. Meanwhile, more Airbus aircrafts have landing distance that's less than 1,000 ft. In addition, ground speed and the type of aircraft seem to be more significant differentiating factors across the distance categories compared to height. Higher ground speed seems to be associated with higher landing distance. But, the spread of height seems to be only slightly different across the three levels of landing distance, with an almost identical spread for medium and high distance.

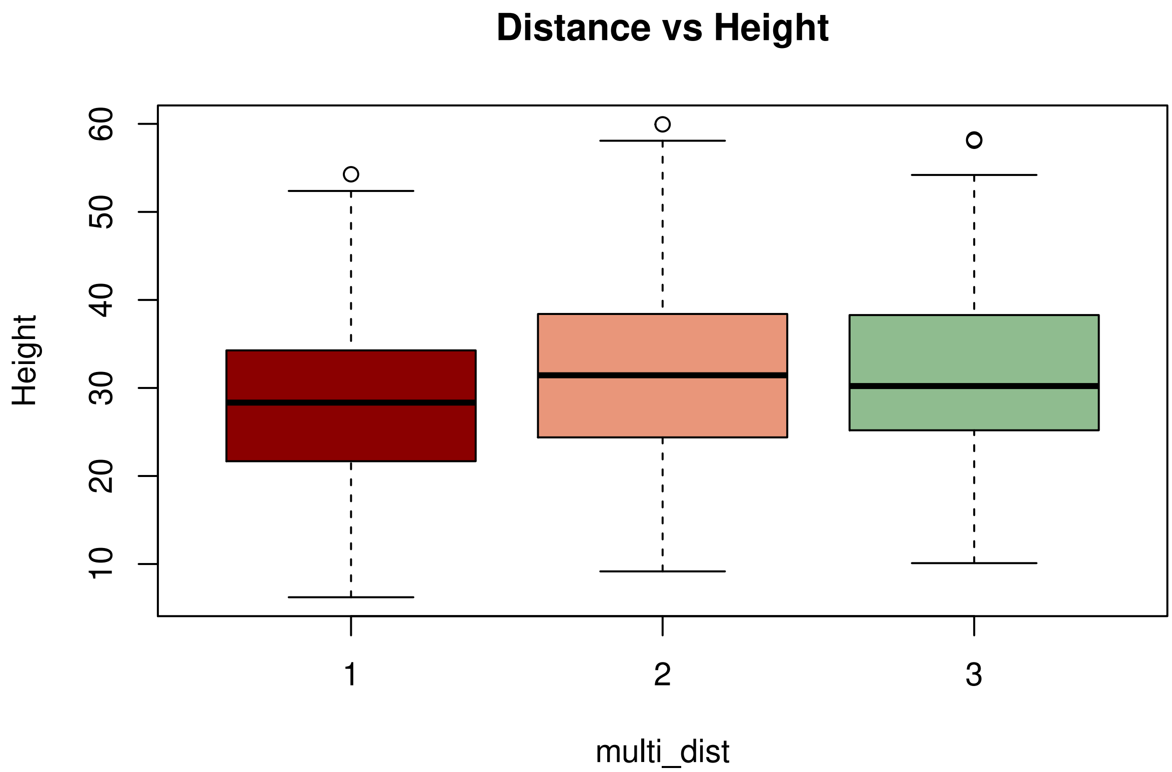
```r
library("dplyr")
library("ggplot2")

#Distance vs aircraft
ggplot(Data_logit,aes(x=aircraft,fill= as.factor(multi_dist)))+
  geom_bar(position="dodge")+ ggtitle("Distance vs aircraft")
```



```r
#Distance vs speed_ground
plot(speed_ground~multi_dist,data = Data_logit,col=colors()[100:102],
     ylab="Speed_ground",
     main="Distance vs Speed_ground")
```




```r
#Distance vs height
plot(height~multi_dist,data = Data_logit,col=colors()[100:102],
     ylab="Height",
     main="Distance vs Height")
```



Full Model

First, we fit a multinomial model with all of the variables. However, we observe that the coefficients and standard errors aren't based on the comparison of the two levels of the multinomial response variable with the reference category.

After some investigation, I realized that speed_air is the problematic variable. There seems to be a problem of perfect prediction. That is, speed_air seems to be only associated with category 2 and 3 of the response variable. Hence, speed_air is excluded from the model.

We aren't losing a lot of information by the exclusion because the model still includes speed_ground which is highly correlated with speed_air.

```
library("nnet")
mmod1a <- multinom(multi_dist~aircraft+duration+no_pasg+speed_ground
                   +speed_air+height+pitch, Data_logit)
```

```
## # weights:  9 (8 variable)
## initial  value 135.163700
## iter  10 value 67.318153
## iter  20 value 17.089456
## iter  30 value 16.593403
## iter  40 value 16.457169
## iter  50 value 16.454905
## iter  60 value 16.454730
## iter  60 value 16.454730
## iter  60 value 16.454730
```

```
## final  value 16.454730
## converged
```

```r
summary(mmod1a)
```

```
## Call:
## multinom(formula = multi_dist ~ aircraft + duration + no_pasg +
##     speed_ground + speed_air + height + pitch, data = Data_logit)
##
## Coefficients:
##                        Values  Std. Err.
## (Intercept)    -1.962720e+02 0.04010950
## aircraftboeing  8.780687e+00 0.96455561
## duration        3.035686e-04 0.01047097
## no_pasg        -7.356207e-02 0.06707488
## speed_ground   -2.253815e-01 0.37344471
## speed_air       1.984274e+00 0.39729688
## height          4.223513e-01 0.05903968
## pitch           1.468367e+00 0.89109555
##
## Residual Deviance: 32.90946
## AIC: 48.90946
```

```r
mmod1b <- multinom(multi_dist~ aircraft+duration+no_pasg+speed_ground
                   +height+pitch,Data_logit)
```

```
## # weights:  24 (14 variable)
## initial  value 858.016197
## iter  10 value 526.458578
## iter  20 value 215.771472
## iter  30 value 199.809707
## iter  40 value 199.420892
## iter  50 value 199.069171
## final  value 198.748963
## converged
```

```r
summary(mmod1b)
```

```
## Call:
## multinom(formula = multi_dist ~ aircraft + duration + no_pasg +
##     speed_ground + height + pitch, data = Data_logit)
##
## Coefficients:
##   (Intercept) aircraftboeing     duration     no_pasg speed_ground    height
## 2    -20.0985       4.084558 -0.003528410 -0.01864576    0.2444092 0.1564499
## 3   -134.9445       9.066761  0.001835174 -0.01119463    1.2236524 0.3909273
##        pitch
## 2 -0.4055170
## 3  0.8773354
##
## Std. Errors:
##   (Intercept) aircraftboeing    duration    no_pasg speed_ground     height
## 2  2.33164518      0.4370675 0.002795548 0.01790935   0.02045739 0.01860185
## 3  0.04012959      0.8816880 0.008101225 0.05827170   0.04073362 0.04903549
##       pitch
## 2 0.2798748
```

```
## 3 0.7670509
##
## Residual Deviance: 397.4979
## AIC: 425.4979
```

Model Selection

Best subset variable selection

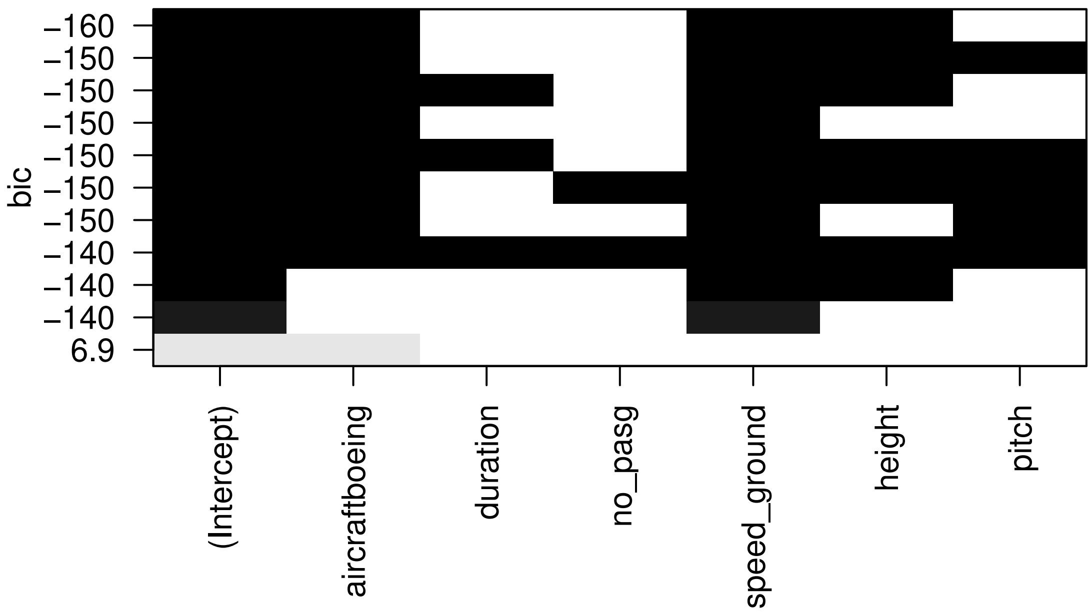
Next, we consider the best subset variable selection technique to select the best model using the BIC criterion. This is considered because BIC gives a simpler model, with less complexity.

Based on the BIC criterion, the reduced model with the lowest BIC value includes the following variables: aircraft type, speed_ground,and height as shown in the figure below.

```
#Best subset variable selection
require(leaps)
Model_best <- regsubsets(multi_dist~.-speed_air,data = Data_logit[,-8],
                         nbest=2, nvmax=14)
summary(Model_best)
```

```
## Subset selection object
## Call: regsubsets.formula(multi_dist ~ . - speed_air, data = Data_logit[,
##     -8], nbest = 2, nvmax = 14)
## 6 Variables  (and intercept)
##                Forced in Forced out
## aircraftboeing     FALSE      FALSE
## duration           FALSE      FALSE
## no_pasg            FALSE      FALSE
## speed_ground       FALSE      FALSE
## height             FALSE      FALSE
## pitch              FALSE      FALSE
## 2 subsets of each size up to 6
## Selection Algorithm: exhaustive
##          aircraftboeing duration no_pasg speed_ground height pitch
## 1  ( 1 ) " "            " "      " "     "*"          " "    " "
## 1  ( 2 ) "*"            " "      " "     " "          " "    " "
## 2  ( 1 ) "*"            " "      " "     "*"          " "    " "
## 2  ( 2 ) " "            " "      " "     "*"          "*"    " "
## 3  ( 1 ) "*"            " "      " "     "*"          "*"    " "
## 3  ( 2 ) "*"            " "      " "     "*"          " "    "*"
## 4  ( 1 ) "*"            " "      " "     "*"          "*"    "*"
## 4  ( 2 ) "*"            "*"      " "     "*"          "*"    " "
## 5  ( 1 ) "*"            "*"      " "     "*"          "*"    "*"
## 5  ( 2 ) "*"            " "      "*"     "*"          "*"    "*"
## 6  ( 1 ) "*"            "*"      "*"     "*"          "*"    "*"
```

```
plot(Model_best,scale="bic")
```



Reduced Model: Selected model using Best subset variable selection

Next, we create a model with the variables selected by the best subset technique. This is called the **Reduced Model**

```
#Reduced model
reduced_model <- multinom(multi_dist~ aircraft+speed_ground+height,
                          data = Data_logit)
```

```
## # weights:  15 (8 variable)
## initial  value 912.946812
## iter  10 value 358.336150
## iter  20 value 230.468155
## iter  30 value 219.198168
## iter  40 value 215.476362
## iter  40 value 215.476360
## iter  40 value 215.476360
## final  value 215.476360
## converged
```

```
summary(reduced_model)
```

```
## Call:
## multinom(formula = multi_dist ~ aircraft + speed_ground + height,
##     data = Data_logit)
##
## Coefficients:
##   (Intercept) aircraftboeing speed_ground    height
```

```
## 2   -23.28484       3.982905    0.2472743 0.1467859
## 3  -126.43265       9.040905    1.1756019 0.3782799
##
## Std. Errors:
##   (Intercept) aircraftboeing speed_ground     height
## 2  1.88720542      0.4027433   0.01980816 0.01714538
## 3  0.04519312      0.7502719   0.01276020 0.03604886
##
## Residual Deviance: 430.9527
## AIC: 446.9527
```

```r
BIC(reduced_model)
```

```
## [1] 484.7338
```

Model Comparison

We compare the full model and the reduced model (selected using the best subset variable selection procedure). The reduced model has a lower BIC compared to the full model. The chi-square test is also employed in selecting the best model between the full and the reduced model.

Null hypothesis: No difference between the two models

Alternative hypothesis: The reduced/smaller model is sufficient

Based on the p-value of the chi square test (=8.570841e-06), we reject the null hypothesis and conclude that the reduced model is sufficient. This suggests that the significant risk factors in predicting the landing distance category are aircraft type, speed_ground, and height.

```r
#model comparison based on BIC
BIC(mmod1b); BIC(reduced_model)
```

```
## [1] 490.746
```

```
## [1] 484.7338
```

```r
#model comparison based on the significance
#Difference between degrees of freedom
deviance(reduced_model) - deviance(mmod1b)
```

```
## [1] 33.45479
```

```r
#difference between degrees of freedom
mmod1b$edf - reduced_model$edf
```

```
## [1] 6
```

```r
###Chi^2 test
pchisq(deviance(reduced_model) - deviance(mmod1b), mmod1b$edf -
         reduced_model$edf, lower=F)
```

```
## [1] 8.570841e-06
```

Best Model & Model Performance

The selected model does a good job with in-sample classification and prediction given that the misclassification rate is 0.0975, which is very low.

Best Model

```r
#model
best_model <- multinom(multi_dist~ aircraft+speed_ground+height,
                       data = Data_logit)
```

```
## # weights:  15 (8 variable)
## initial  value 912.946812
## iter  10 value 358.336150
## iter  20 value 230.468155
## iter  30 value 219.198168
## iter  40 value 215.476362
## iter  40 value 215.476360
## iter  40 value 215.476360
## final  value 215.476360
## converged
```

```r
summary(best_model)
```

```
## Call:
## multinom(formula = multi_dist ~ aircraft + speed_ground + height,
##     data = Data_logit)
##
## Coefficients:
##   (Intercept) aircraftboeing speed_ground    height
## 2   -23.28484       3.982905    0.2472743 0.1467859
## 3  -126.43265       9.040905    1.1756019 0.3782799
##
## Std. Errors:
##   (Intercept) aircraftboeing speed_ground     height
## 2  1.88720542      0.4027433   0.01980816 0.01714538
## 3  0.04519312      0.7502719   0.01276020 0.03604886
##
## Residual Deviance: 430.9527
## AIC: 446.9527
```

Predicted probabilities

```r
#predicted probabilities
pred_lprob <- predict(reduced_model,Data_logit, type = "probs")
head(pred_lprob, 10)
```

```
##               1            2            3
## 1  2.403400e-09 0.0002163023 9.997837e-01
## 2  3.061082e-06 0.0620059283 9.379910e-01
## 3  2.697197e-01 0.7302803474 6.003268e-13
## 4  1.609247e-03 0.9983785196 1.223322e-05
## 5  4.347542e-01 0.5652457974 3.586254e-16
## 6  4.981863e-03 0.9950181312 6.094404e-09
## 7  9.100954e-01 0.0899046010 5.184192e-20
## 8  9.118276e-01 0.0881724148 2.071558e-19
## 9  8.941804e-04 0.9990804474 2.537216e-05
## 10 2.021342e-01 0.7978658138 8.148886e-15
```

Confusion Matrix

```r
#confusion matrix
predicted_class <- predict (reduced_model,Data_logit)
table(predicted_class, Data_logit$multi_dist, dnn = c("True", "Predicted"))
```

```
##     Predicted
## True   1   2   3
##    1 232  34   0
##    2  37 421   6
##    3   0   4  97
```

Misclassification Rate

```r
#misclassification rate
mean(as.character(predicted_class) != as.character(Data_logit$multi_dist),
     na.rm = TRUE)
```

```
## [1] 0.09747292
```

Q1: Presentation to the FAA Agent

Note: For easier interpretation, the 3 levels of the multinomial response variable "landing distance" are refered to as follows: Y=1 as "low distance", Y=2 as "medium distance", and Y=3 as "high distance".

Executive summary

- Aircraft type, ground speed of the aircraft, and height are the most important risk factors in the landing process. Although, ground speed and the type of aircraft seem to be more significant differentiating factors across the distance categories compared to height. For instance, higher speed_ground seems to be associated with higher landing distance. But, the spread of height seems to be only slightly different across the three levels of landing distance, with an almost identical spread for medium and high distance.
- Boeing aircrafts have higher distance and fall in the "medium" and "high" category more than Airbus aircrafts. Compared to Boeing, more Airbus aircrafts have landing distance that's less than 1,000ft.
- If ground speed increases by one unit, the aircraft is 0.25 times more likely to be in the medium landing distance category (1000<=distance<2500) as compared to low distance category. In the same vein, if ground speed increases by one unit, the aircraft is 1.18 times more likely to be in the high landing distance category (>=2500ft) as compared to low distance category, keeping all other variables constant.
- If height increases by one unit, the aircraft is 0.15 times more likely to be in the medium landing distance category (1000<=distance<2500) as compared to low distance category. In the same vein, if height increases by one unit, the aircraft is 0.38 times more likely to be in the high landing distance category (>=2500ft) as compared to low distance category, keeping all other variables constant.
- The selected model does a good job with classification and prediction given that the misclassification rate is 0.0975, which is very low.

Model

```r
#model
best_model <- multinom(multi_dist~ aircraft+speed_ground+height,
                       data = Data_logit)
```

```
## # weights:  15 (8 variable)
## initial  value 912.946812
## iter  10 value 358.336150
## iter  20 value 230.468155
## iter  30 value 219.198168
## iter  40 value 215.476362
## iter  40 value 215.476360
## iter  40 value 215.476360
## final  value 215.476360
```

```
## converged
```

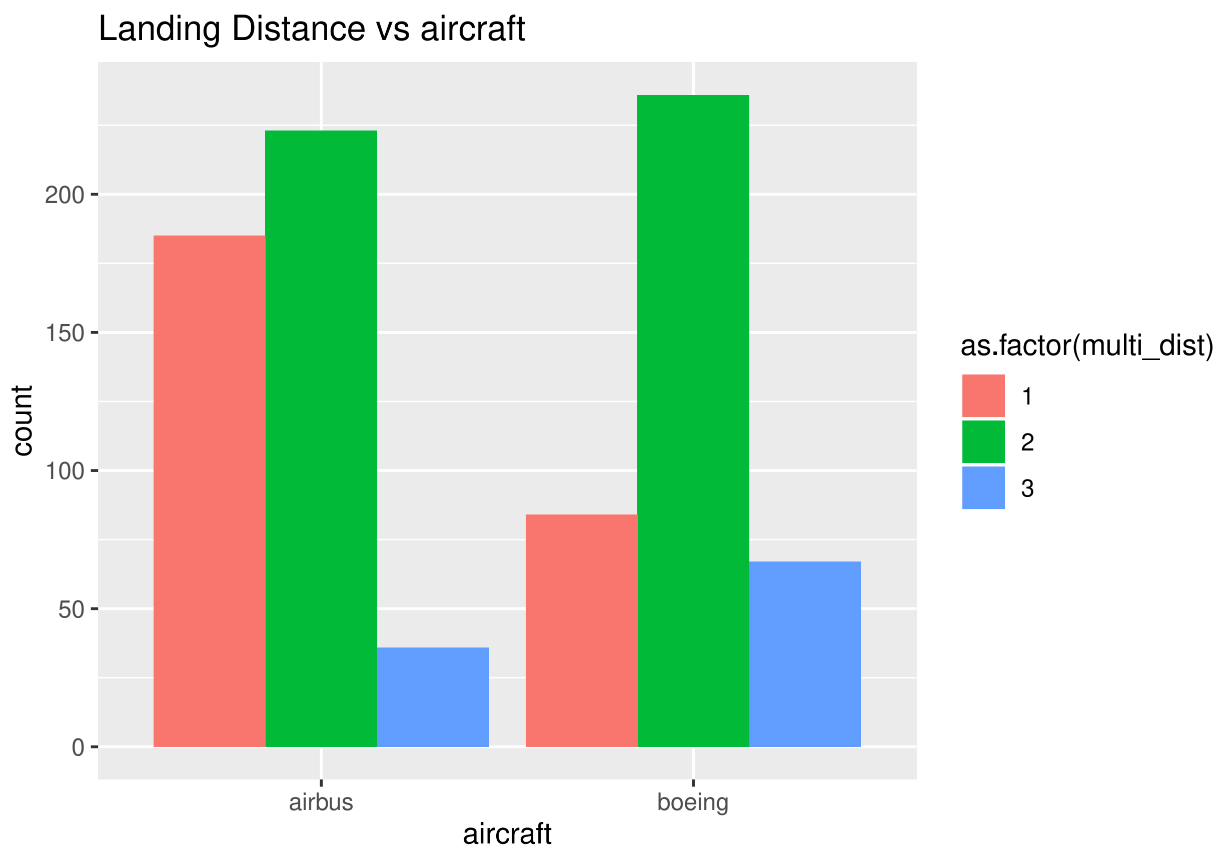
```r
summary(best_model)
```

```
## Call:
## multinom(formula = multi_dist ~ aircraft + speed_ground + height,
##     data = Data_logit)
##
## Coefficients:
##   (Intercept) aircraftboeing speed_ground    height
## 2   -23.28484       3.982905    0.2472743 0.1467859
## 3  -126.43265       9.040905    1.1756019 0.3782799
##
## Std. Errors:
##   (Intercept) aircraftboeing speed_ground     height
## 2  1.88720542      0.4027433   0.01980816 0.01714538
## 3  0.04519312      0.7502719   0.01276020 0.03604886
##
## Residual Deviance: 430.9527
## AIC: 446.9527
```

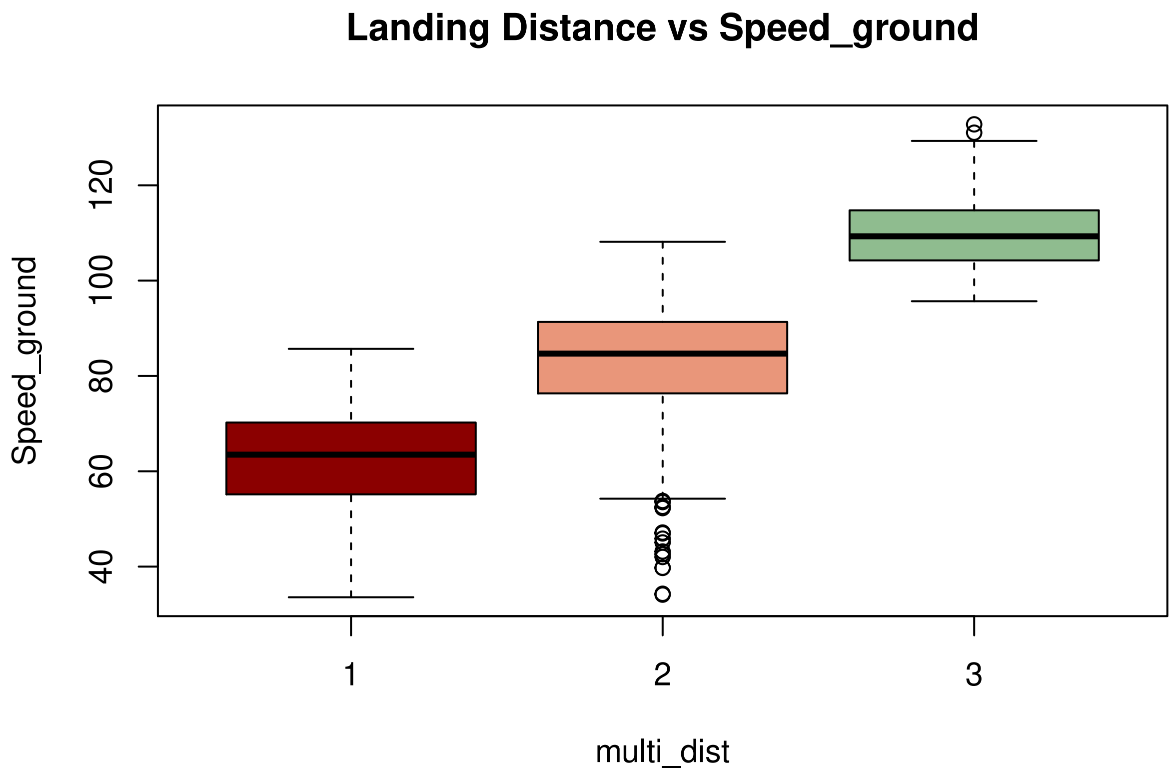
Association between the significant factors and the landing process

The figures below show the association between the significant factors and the 3 levels of the multinomial response variable "landing distance".

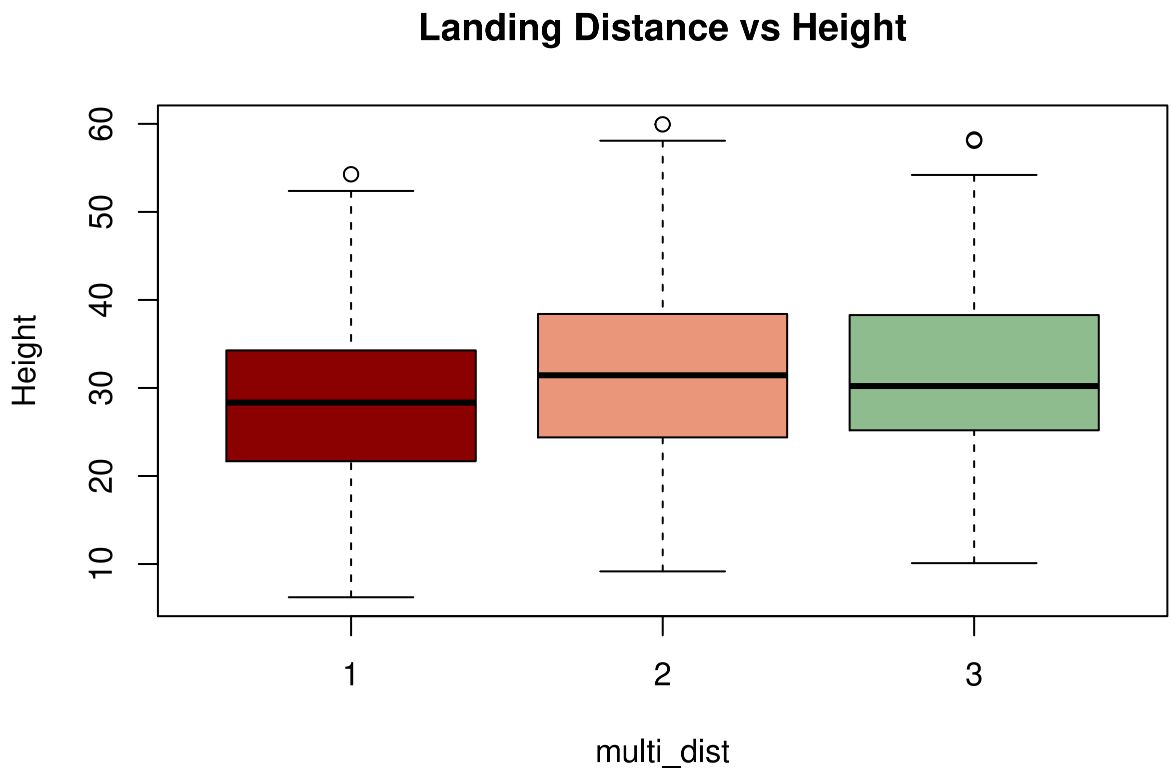
```r
#Landing Distance Categories vs aircraft
ggplot(Data_logit,aes(x=aircraft,fill= as.factor(multi_dist)))+
  geom_bar(position="dodge")+ ggtitle("Landing Distance vs aircraft")
```




```
#Landing Distance Categories vs speed_ground
plot(speed_ground~multi_dist,data = Data_logit,col=colors()[100:102],
     ylab="Speed_ground",
     main="Landing Distance vs Speed_ground")
```




```
#Landing Distance categories vs height
plot(height~multi_dist,data = Data_logit,col=colors()[100:102],
     ylab="Height",
     main="Landing Distance vs Height")
```



Confusion Matrix and Misclassification rate of the model

```
#confusion matrix
predicted_class <- predict (best_model,Data_logit)
table(predicted_class, Data_logit$multi_dist, dnn = c("True", "Predicted"))
```

```
##      Predicted
## True   1   2   3
##    1 232  34   0
##    2  37 421   6
##    3   0   4  97
```

```
#misclassification rate
mean(as.character(predicted_class) != as.character(Data_logit$multi_dist),
     na.rm = TRUE)
```

```
## [1] 0.09747292
```

Q2: Modeling count Data (No of Passengers)

Given that the number of passenger is a count data, it can be modeled using the poisson distribution

Fit a GLM

The model indicates that none of the variables in the dataset have significant impact on the number of passengers.

We consider further exploration by observing the correlation between the predictor variables and the response variable.

```
Data <- read.csv("clean_data.csv")
mmod_pass <- glm(no_pasg~., family=poisson, Data[,-1])
summary(mmod_pass)
```

```
##
## Call:
## glm(formula = no_pasg ~ ., family = poisson, data = Data[, -1])
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.63995  -0.54433   0.06167   0.55223   2.49267
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      3.563e+00  4.631e-01   7.694 1.43e-14 ***
## aircraftboeing   2.814e-02  3.687e-02   0.763    0.445
## duration        -1.607e-04  1.963e-04  -0.819    0.413
## speed_ground     4.506e-04  6.170e-03   0.073    0.942
## speed_air        7.074e-03  8.668e-03   0.816    0.414
## height           1.206e-03  1.381e-03   0.873    0.383
## pitch           -5.763e-03  1.793e-02  -0.321    0.748
## distance        -9.215e-05  6.988e-05  -1.319    0.187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 162.74  on 194  degrees of freedom
## Residual deviance: 159.57  on 187  degrees of freedom
##   (636 observations deleted due to missingness)
## AIC: 1330.9
##
## Number of Fisher Scoring iterations: 4
```

Correlation

We observe the presence of high correlation between the following pairs: speed_ground and speed_air; speed_ground and distance; speed_air and distance. However, speed_ground and speed_air have no relationship with the number of passengers on board. In addition, all of the other variables have very weak relationship with the number of passenger. This confirms why none of the predictor variables were significant in the model.

Hence, I conclude that with the number of passenger being modeled using a poisson distribution, none of the variables in the FAA data set are useful in predicting the number of passengers on board.

```
Data1 <- na.omit(Data[,-1])
round(cor(Data1[sapply(Data1, is.numeric)]),2)
```

```
##              duration no_pasg speed_ground speed_air height pitch distance
## duration         1.00   -0.07         0.02      0.04   0.07 -0.06     0.05
## no_pasg         -0.07    1.00         0.00      0.00  -0.01 -0.04    -0.03
## speed_ground     0.02    0.00         1.00      0.99  -0.10 -0.06     0.93
## speed_air        0.04    0.00         0.99      1.00  -0.09 -0.05     0.94
## height           0.07   -0.01        -0.10     -0.09   1.00 -0.03     0.06
## pitch           -0.06   -0.04        -0.06     -0.05  -0.03  1.00     0.03
## distance         0.05   -0.03         0.93      0.94   0.06  0.03     1.00
```

Conclusion and Recommendation

Given the conclusion that none of the variables in the FAA data set are useful in predicting the number of passengers on board, I would recommend that the FAA agent take a look at other variables such as ticket price, season, take-off location, destination of the aircraft, weather, discounts, and other events that may influence the number of passengers on board.