PROJECT PART3:

QN1:

Considering Y as non-ordinal:

The new variable Y is added to the data set:

```
FAA$Y <- ifelse(FAA$distance < 1000,1,ifelse((FAA$distance>=1000 &
FAA$distance<2500),2,3))
```

The distance variable is removed from the data set:

```
FAA <- subset( FAA, select = -distance )
```

As was seen the previous two projects the variables speed_air and speed_ground are highly correlated. Hence, the variable speed_air is removed from the model:

```
FAA <- subset(FAA, select = -speed_air)
```

The NA values are omited and a multinomial model is fitted using the Y variable as the predicted variable and the rest as predictor variables:

```
FAA <- na.omit(FAA)
modl <- multinom(Y~.,FAA)
```

The step wise model selection algorithm based on AIC is used for the model Selection:

```
modl_reduced <- step(modl)
```

	Df	AIC
<none>	10	420.6864
- pitch	8	421.5598
- height	8	544.6126
- aircraft	8	599.5388
- speed_ground	8	1435.0477

It can be seen from the step wise model that the variables pitch, height, aircraft and speed_ground are useful in characterizing the multinomial Y variable.

The model is:

```
modl3 <- multinom(Y ~ pitch+height+aircraft+speed_ground, FAA_new)
summary(modl3)
```

```
Coefficients:
  (Intercept)   aircraftboeing   speed_ground    height        pitch
2   -21.67806      4.065194        0.2431976    0.1557486   -0.3987472
3  -135.02763      8.988756        1.2159693    0.3977880    0.9396833

Std. Errors:
  (Intercept)     aircraftboeing   speed_ground    height        pitch
2  2.09113433      0.4340363        0.02026387.  0.01856171   0.2793028
3  0.03719281      0.8697689        0.02874032   0.04079031   0.7484298

Residual Deviance: 400.6864
AIC: 420.6864
```

A new data frame was created without the output variable:

```
FAA_test <- subset(FAA_new, select = -Y)
```

A misclassification table is created with the predicted values and the actual values:

```
xtabs(~predict(modl_reduced) + FAA_new$Y)
```

```
                          FAA_new$Y
predict(modl_reduced)      1    2    3
                      1  207   31    0
                      2   38  400    6
                      3    0    5   94
```

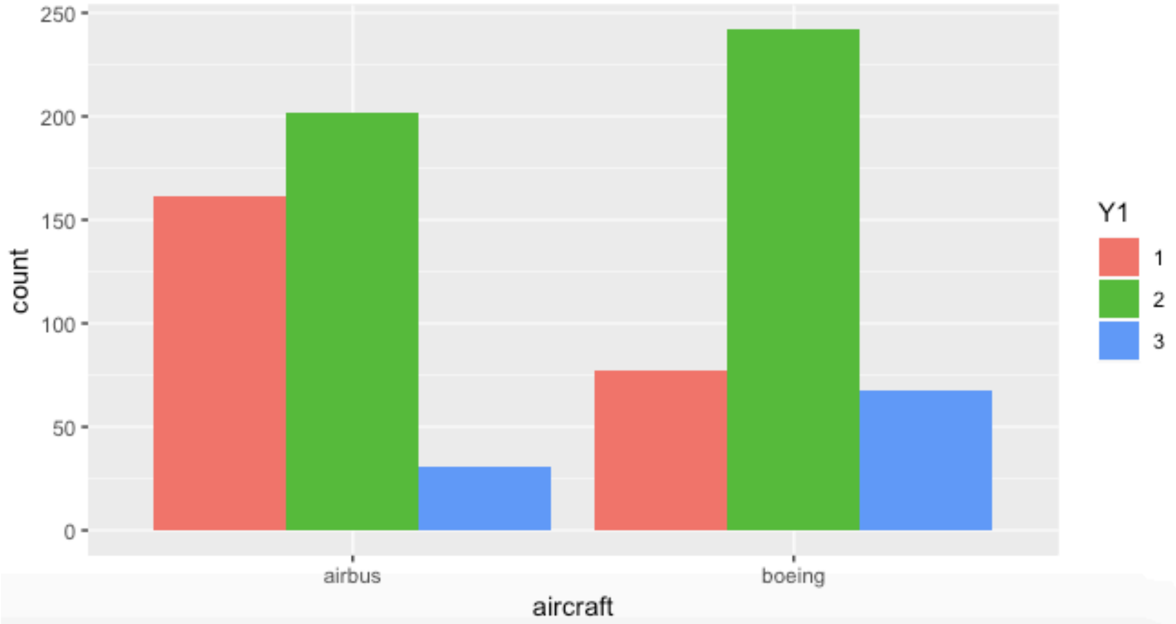
Table 1: The misclassification table

The total misclassification rate is:
(31+38+6+5)/ (207+31+38+400+6+5+94) = 0.102

Log-odds	Aircraft = boeing	speed_ground	Height	Pitch
Log(Y=2/Y=1)	58.26	1.27	1.17	0.67
Log(Y=3/Y=1)	8006	3.37	1.49	2.56

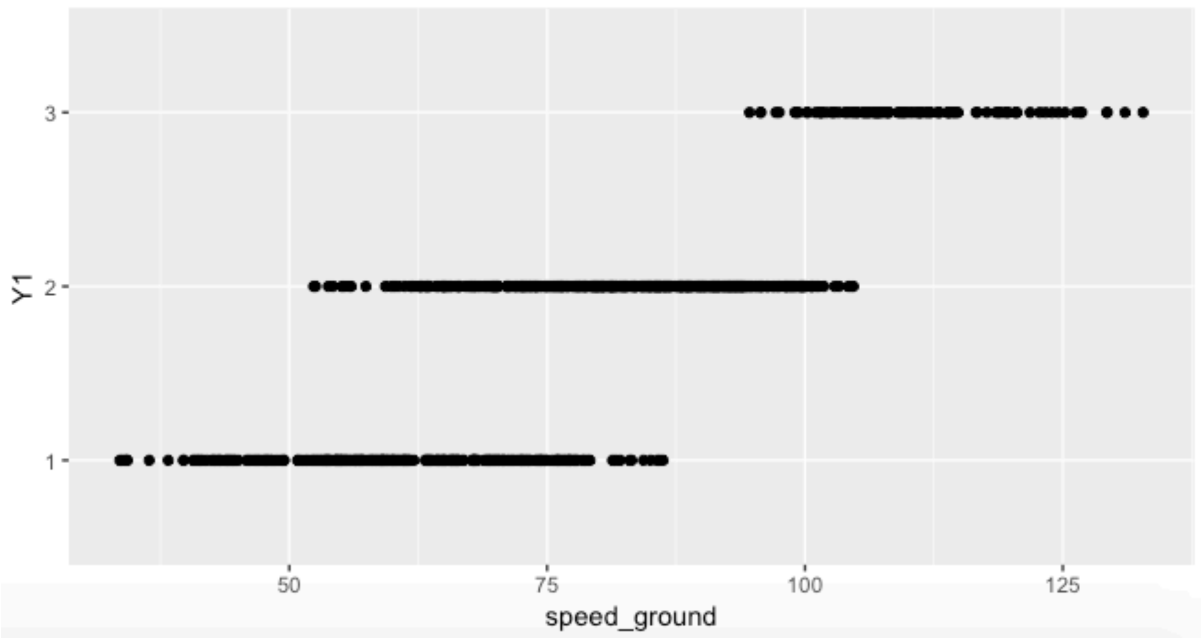
Table 2: The log odds ratio

```
ggplot(FAA_new, aes(x=aircraft, fill=Y)) + geom_bar(position = 'dodge')
```



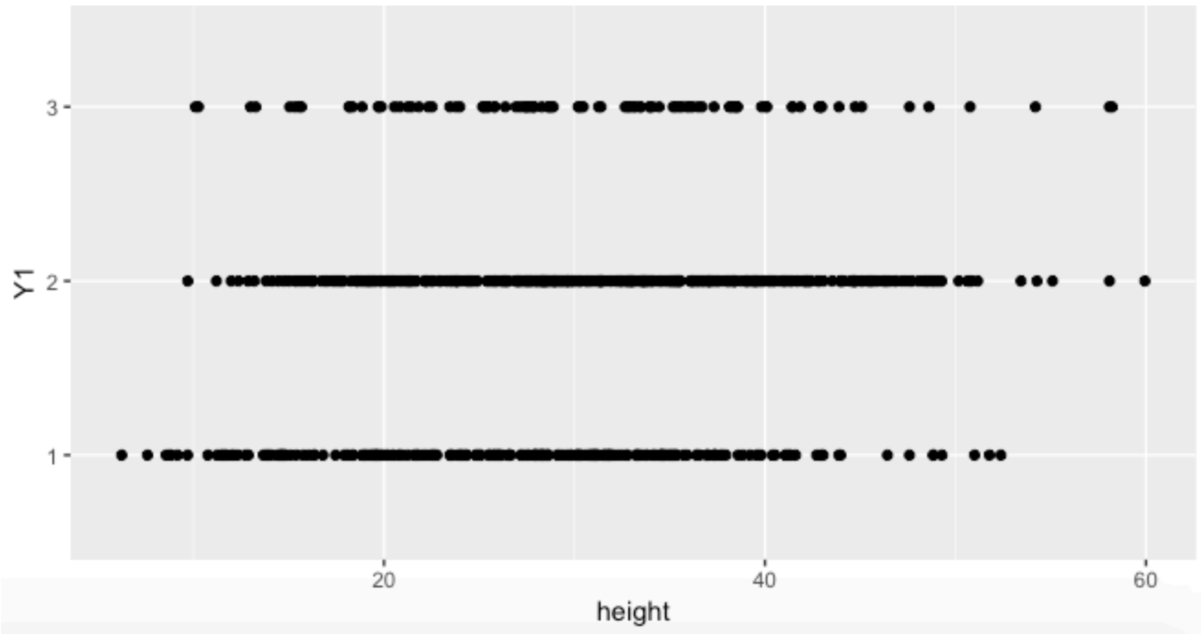
The plot shows the predicted Y value as a function of aircraft. The boeing aircraft has more proportion of Y=2,3 values compared to airbus.

```
ggplot(FAA_test, aes(x=speed_ground, y=Y1)) + geom_point()
```



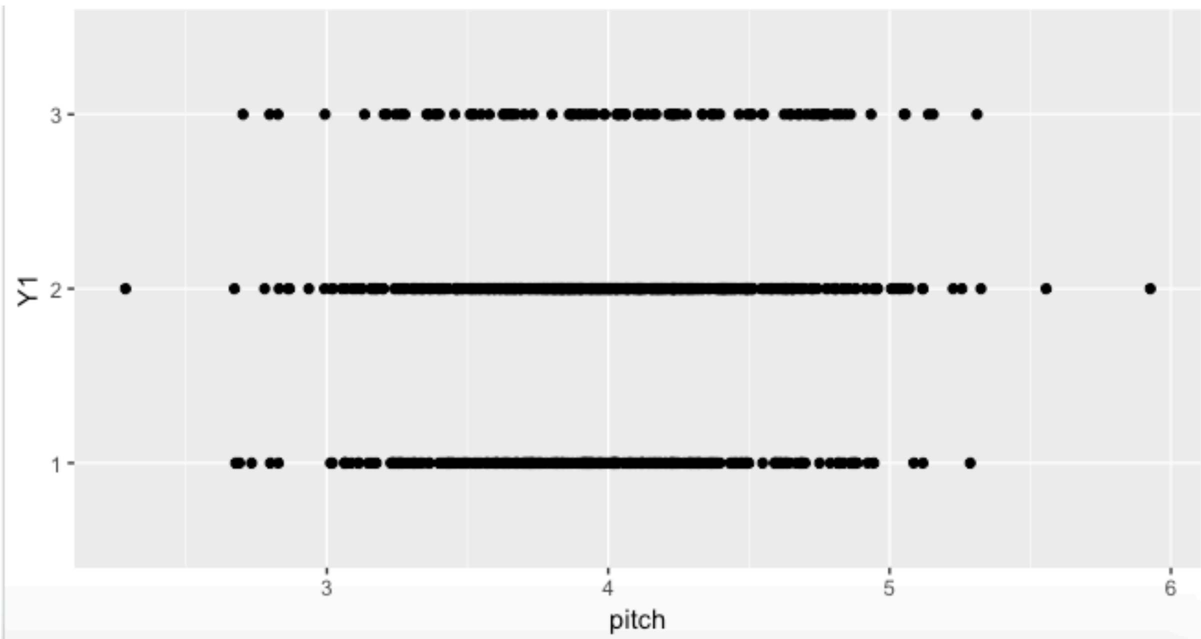
From the above plot it is clear that as the speed_ground increases the proportion of Y=2,3 increases compared to Y=1.

ggplot(FAA_test, aes(x=height, y=Y1)) + geom_point()



There seems to be very little variation in Y values as the height changes.

ggplot(FAA_test, aes(x=pitch, y=Y1)) + geom_point()



The pitch does not seem to affect the landing distance much.

Conclusion:

1. The main variables that seem to affect the multinominal variable Y are:

- Aircraft
- Speed Ground
- Height
- Pitch

2. The aircraft and speed_ground have a greater influence on the multi-nominal variable Y than height and pitch.
3. The aircraft boeing has a higher proportion of the landing distance in the Y=2,3 category compared to airbus.
4. Speed ground seems to directly influence the landing distance. As the speed ground increases the proportional of the Y=2,3 landing increases.
5. The height and pitch seem to have marginal influence on the landing distance.
6. The log odds ratio is presented in the Table 2. It clearly follows the pattern which we predicted by the plots.

Considering Y as ordinal:

The following model was built using the same data set:

model_ordinal<-vglm(Y ~ .,family=cumulative(parallel=TRUE), FAA_new2)

Pearson residuals:

	Min	1Q	Median	3Q	Max
logitlink(P[Y<=1])	-53.129	-0.155085	-0.01871	0.09456	3.685
logitlink(P[Y<=2])	-5.282	0.001495	0.01161	0.07565	1.510

Coefficients:

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept):1	**19.669758**	**1.979850**	**9.935**	**<2e-16 *****
(Intercept):2	**29.462780**	**2.436218**	**12.094**	**<2e-16 *****
aircraft	**3.597667**	**0.345562**	**10.411**	**<2e-16 *****
duration	0.002570	0.002364	1.087	0.277
no_pasg	0.012722	0.015317	0.831	0.406
speed_ground	**-0.276852**	**0.018940**	**-14.618**	**<2e-16 *****
height	**-0.136033**	**0.015078**	**-9.022**	**<2e-16 *****
pitch	0.116946	0.236571	0.494	0.621

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])

Residual deviance: 492.6454 on 1554 degrees of freedom

Log-likelihood: -246.3227 on 1554 degrees of freedom

Number of Fisher scoring iterations: 8

The summary of the model was caculated:

Summary(model_ordinal)

Exponentiated coefficients:

aircraft	duration	no_pasg	**speed_ground**	**height**	pitch
36.5129359	1.0025731	1.0128038	**0.7581663**	**0.8728141**	1.1240585

From the above table it is clear that speed_ground, height and aircraft are the three significant variables used for building the model.

CONCLUSION:

The model is:

Pr(Y<=1) = F(19.669758+3.597667*aircraft-0.27685*speed_ground-0.136033*height)

Pr(Y<=2)= F(29.462780+3.597667*aircraft-0.27685*speed_ground-0.136033*height)

Variable	Estimate	Std-Error
(Intercept):1	19.669758	1.979850
(Intercept):2	29.462780	2.436218
aircraft	3.597667	0.345562
speed_ground	-0.276852	0.018940
height	-0.136033	0.015078

The following conclusions can be drawn:

1. We have assumed the slopes to be equal meaning the predictor variables causes similar effects on the probabilities.

2. The variables that influence the probability of the ordinal multinomial variable Y are aircraft, speed_ground and height

3. Since the output model calculates the probability estimate, for a given observation we can calculate the calculate the probability if Y is 1,2 or 3.

QN2:

We can use poisson distribution to predict the number of passengers on board.

```
FAA_new <- subset(FAA, select = -speed_air)
FAA_new$aircraft <- ifelse(FAA_new$aircraft =="boeing",0,1)
FAA_new <- na.omit(FAA_new)
```

A modle was built using the glm function using family as poisson the cleaned data set

```
mdl_no_pasg <- glm(no_pasg ~ ., family=poisson, FAA_new)
```

The step function was used to get a simplified model:

```
modl_simp <- step(mdl_no_pasg)
summary(modl_simp)
```

```
Coefficients:
             Estimate    Std. Error    z value    Pr(>|z|)
(Intercept)  4.095709    0.004616       887.2     <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 742.75  on 780  degrees of freedom
Residual deviance: 742.75  on 780  degrees of freedom
AIC: 5374.8
```

The goodness of fit and dispersion factor was found for the model:

```
gof<-sum(residuals(modl_simp,type="pearson")^2)
dp<-gof/modl_simp$df.res
```

The dispersion factor was found to be 0.94. The revised summary using the dispersion factor is found below:

```
summary(modl_simp,dispersion=dp)
```

```
Coefficients:
              Estimate    Std. Error   z value    Pr(>|z|)
(Intercept)   4.095709    0.004482     913.7      <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 0.9427883)

    Null deviance: 742.75  on 780  degrees of freedom
Residual deviance: 742.75  on 780  degrees of freedom
AIC: 5374.8

Number of Fisher Scoring iterations: 4
```

Conclusion:

It is found that No variable is predicting the number of passengers in the aircraft as is expected. It is solely dependent on the intercept.