FINAL PROJECT- QUESTION 1

SNIGDHA PEDDI

INTRODUCTION

Microtus data from *Flury* package consists of data of two different species of Microtus, M.multiplex and M.subterraneus which are difficult to distinguish morphologically. The Microtus data consists of eight morphometric variables measured using Nikon measure-scope and dial calipers. The data set has records of 288 specimens out of which 89 were analyzed and their species was identified. Remaining 199 specimenwere grouped as unknown and are to be distinguished into respective species based on the morphometric variables. The 9 variables include Group(a factor with levels multiplex, subterraneus, unknown), MlLeft(width of upper left molar 1-0.001mm),M2Left (width of upper left molar 2-0.0001mm),M3Left(width of upper left molar 3-0.001mm),Foramen(Length of incisive foramen-0.001mm),Pbone(Length of palatal bone-0.001mm),Length(condylo incisive length or skull length-0.01mm),Height(skull height ablove bullae-0.01mm),Rostrum(skull width across rostrum-0.01mm).Generalized linear model will be fit and used to identify these unknown species.

ANALYSIS

Exploratory Data Analysis: Microtus data is subset to Training and Test data sets. 89 specimen that were previously identified and grouped into miltiplex and subterraneus were subset into Training data set and that were grouped as unknown species were subset to Test data set.Exploratory data analysis is done to verify the dimensions of the datasets and if there were any missing values.And the summary of the datasets give a basic idea of values in the datasets (mean ,median values etc.).

```
## Dimensions of Training Set: 89 9
```

```
## Number of missing values in Training Set: 0
```

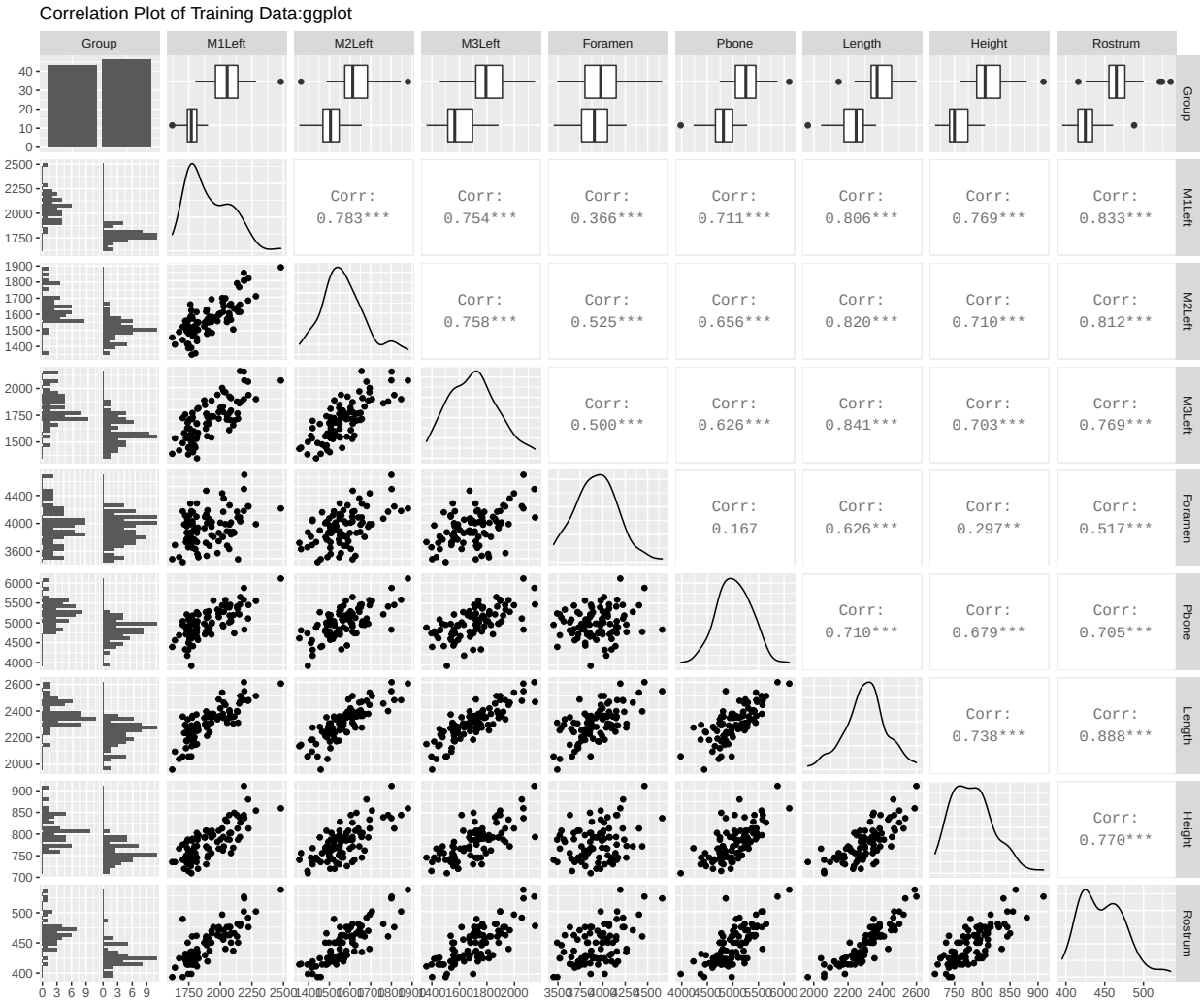
```
##           Group        M1Left          M2Left          M3Left        Foramen
##  multiplex   :43   Min.   :1619   Min.   :1355   Min.   :1361   Min.   :3451
##  subterraneus:46   1st Qu.:1770   1st Qu.:1504   1st Qu.:1561   1st Qu.:3764
##                    Median :1885   Median :1551   Median :1712   Median :3941
##                    Mean   :1909   Mean   :1568   Mean   :1705   Mean   :3932
##                    3rd Qu.:2052   3rd Qu.:1621   3rd Qu.:1815   3rd Qu.:4078
##                    Max.   :2479   Max.   :1880   Max.   :2150   Max.   :4662
##      Pbone          Length         Height         Rostrum
##  Min.   :3980   Min.   :1965   Min.   :715.0   Min.   :395.0
##  1st Qu.:4773   1st Qu.:2237   1st Qu.:750.0   1st Qu.:425.0
##  Median :5004   Median :2300   Median :776.0   Median :450.0
##  Mean   :5025   Mean   :2304   Mean   :782.9   Mean   :447.2
##  3rd Qu.:5254   3rd Qu.:2370   3rd Qu.:805.0   3rd Qu.:465.0
##  Max.   :6104   Max.   :2600   Max.   :910.0   Max.   :535.0
```

```
## Dimensions of Test Set: 199 8
```

```
## Number of missing values in Test Set: 0
```

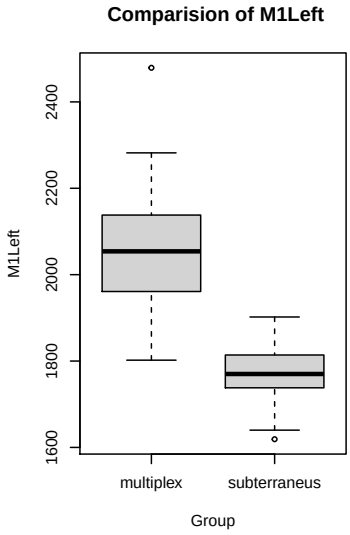
```
##      M1Left          M2Left          M3Left        Foramen          Pbone
##  Min.   :1534   Min.   :1362   Min.   :1416   Min.   :3155   Min.   :3928
##  1st Qu.:1804   1st Qu.:1502   1st Qu.:1614   1st Qu.:3746   1st Qu.:4844
##  Median :1950   Median :1576   Median :1739   Median :3930   Median :5100
##  Mean   :1947   Mean   :1598   Mean   :1737   Mean   :3904   Mean   :5108
##  3rd Qu.:2092   3rd Qu.:1672   3rd Qu.:1870   3rd Qu.:4082   3rd Qu.:5384
##  Max.   :2434   Max.   :1865   Max.   :2187   Max.   :4500   Max.   :6020
##      Length          Height          Rostrum
##  Min.   :1908   Min.   :700.0   Min.   :375.0
##  1st Qu.:2222   1st Qu.:760.0   1st Qu.:428.0
##  Median :2320   Median :790.0   Median :453.0
##  Mean   :2311   Mean   :794.4   Mean   :452.9
##  3rd Qu.:2406   3rd Qu.:825.0   3rd Qu.:475.0
##  Max.   :2605   Max.   :912.0   Max.   :545.0
```

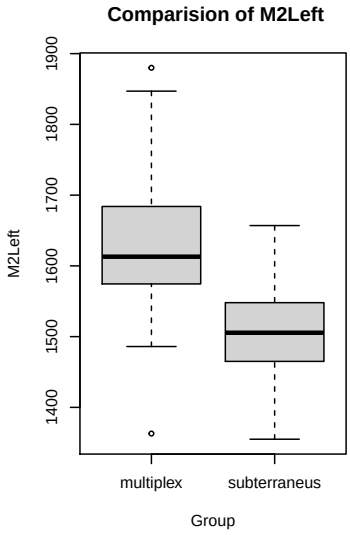
The correlation between the variables and the Group of the training data is reviewed by plotting the correlation using *ggpairs*.

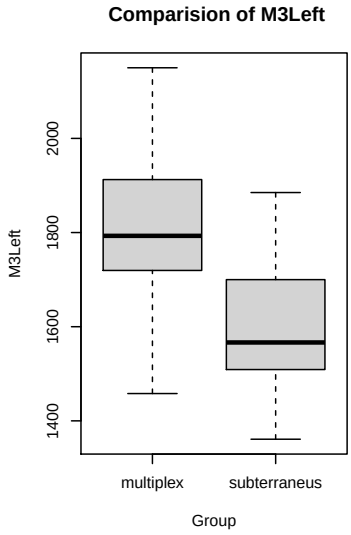


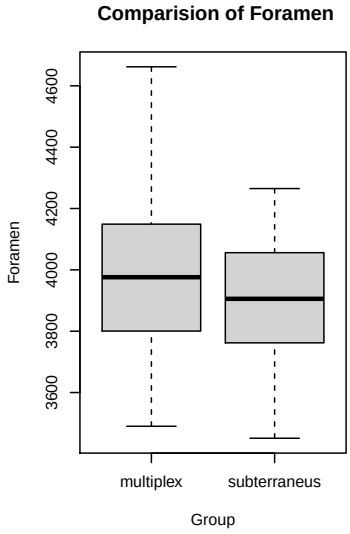
From the plot it is clear that there is a non-linear relationship between the Group variable and the rest of the variables.To further understand the relationship between the two different groups and their characteristics,

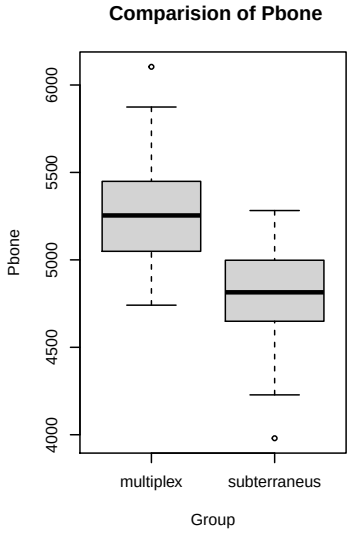
box plot of of all the variables is reviewed. These plot shows that the mean values of all the variables are higher for multiplex species compared to subterraneus species indicating that which a right model the unknown specimenn can be identified and grouped into multiplex and subterraneus species.

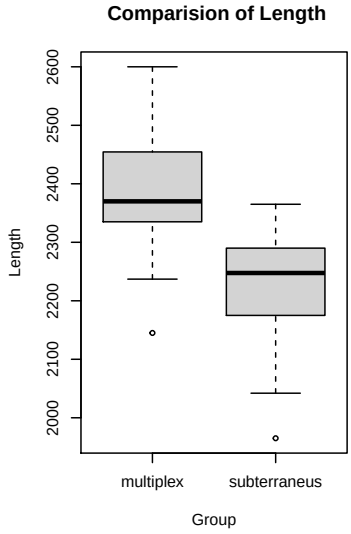

Comparision of M1Left
M1Left
1600
1800
2000
2200
2400
multiplex
subterraneus
Group

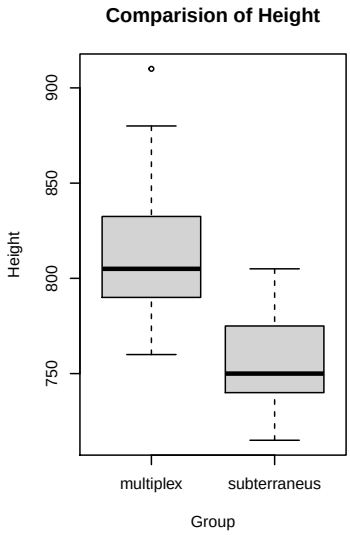

Comparision of M2Left
M2Left
1400
1500
1600
1700
1800
1900
multiplex
subterraneus
Group

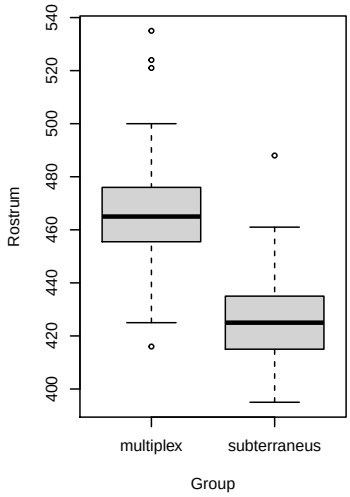

Comparision of M3Left
M3Left
1400
1600
1800
2000
multiplex
subterraneus
Group

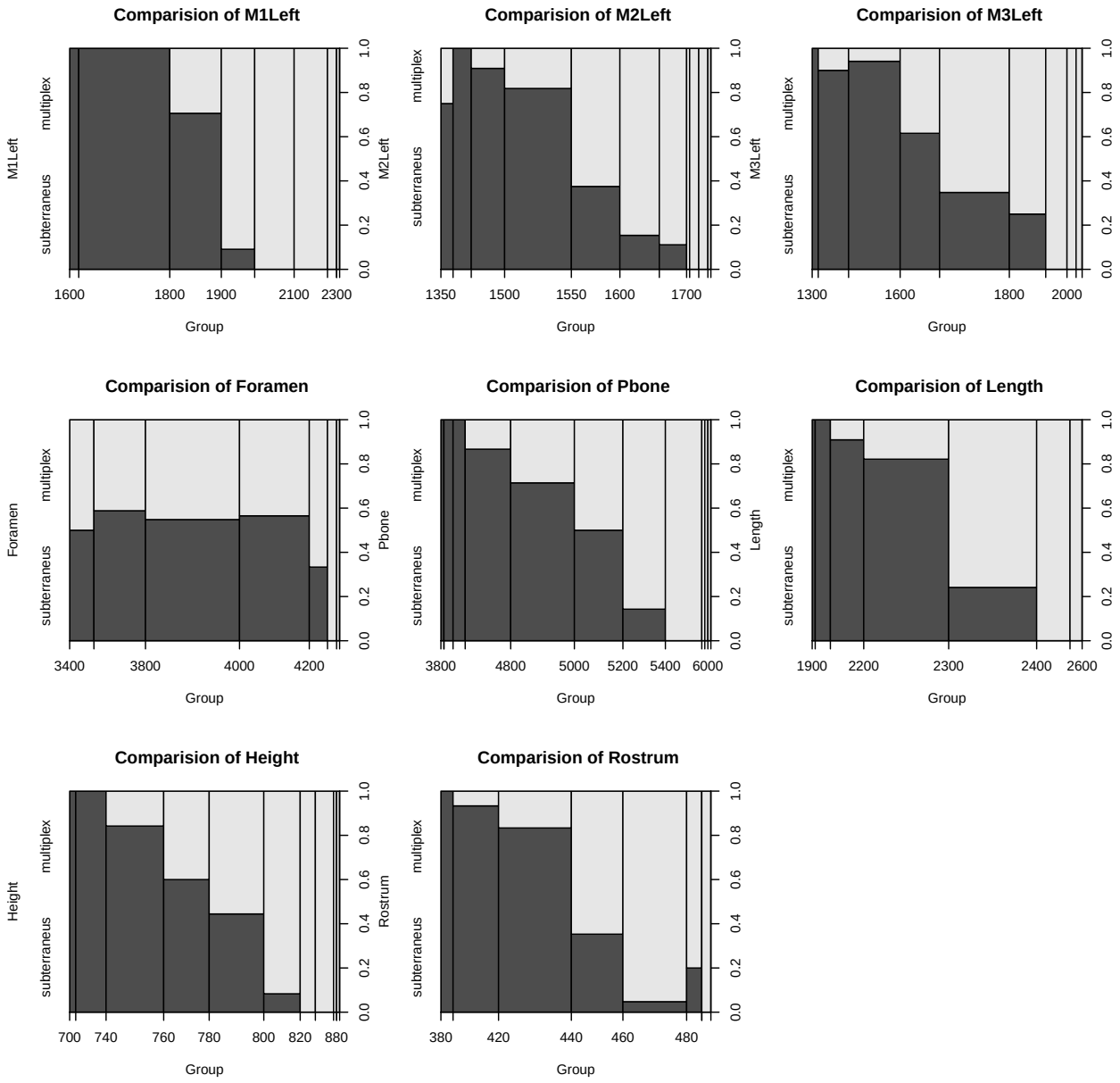

Comparision of Foramen
Foramen
3600
3800
4000
4200
4400
4600
multiplex
subterraneus
Group


Comparision of Pbone
Pbone
4000
4500
5000
5500
6000
multiplex
subterraneus
Group


Comparision of Length
Length
2000
2100
2200
2300
2400
2500
2600
multiplex
subterraneus
Group


Comparision of Height
Height
750
800
850
900
multiplex
subterraneus
Group


Comparision of Rostrum
Rostrum
400
420
440
460
480
500
520
540
multiplex
subterraneus
Group



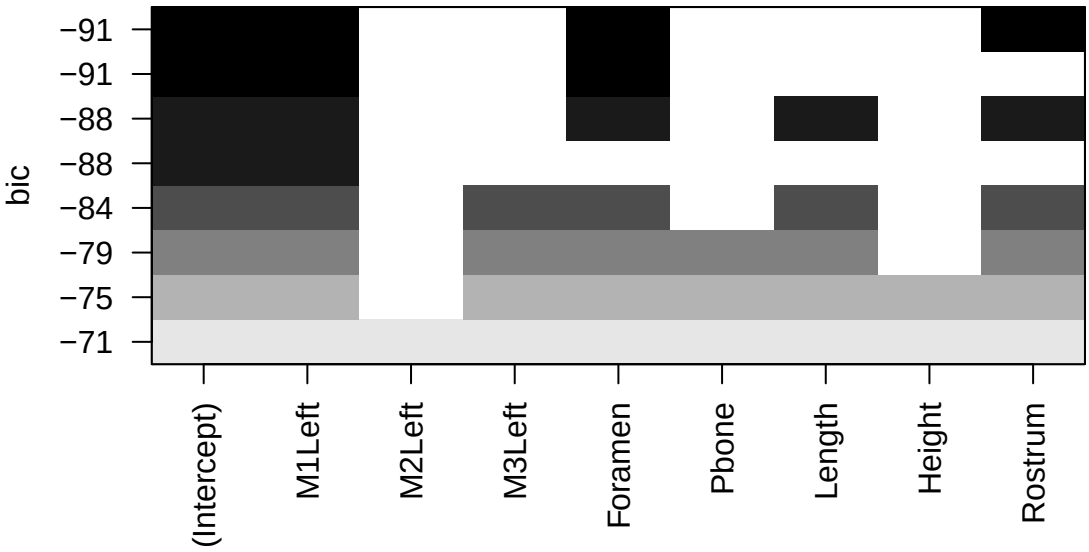
Feature Selection: Various methods were considered for the feature selection.Feature selection using *regsubsets()* function (from leaps library) helps in selecting the best model among the models with increasing number of predictor variables.For instance, a best model with two predictor variables contain M1Left and Foramen variables.By default, the *regsubsets()* function only outputs the results from the best fit models.The '*' indicate the variables selected in each best model.The adjusted R^2 of the selected models show that including the number of variables in the model gives the best performance. However, an optimal model is selected from the BIC plot of the fit. The BIC plot indicates that the optimal model can be fit using intercept, M1Left, Foramen and Rostrum variables.The top row of the plot has a black square indicating the best variables to be used in the model.A Generalized Linear Model(GLM) is fit using these variables.

```
## Subset selection:

## regsubsets.formula(Group ~ ., micro1)
```

```
##          M1Left M2Left M3Left Foramen Pbone Length Height Rostrum
## 1  ( 1 ) "*"    " "    " "    " "     " "   " "    " "    " "
## 2  ( 1 ) "*"    " "    " "    "*"     " "   " "    " "    " "
## 3  ( 1 ) "*"    " "    " "    "*"     " "   " "    " "    "*"
## 4  ( 1 ) "*"    " "    " "    "*"     " "   "*"    " "    "*"
## 5  ( 1 ) "*"    " "    "*"    "*"     " "   "*"    " "    "*"
## 6  ( 1 ) "*"    " "    "*"    "*"     "*"   "*"    " "    "*"
## 7  ( 1 ) "*"    " "    "*"    "*"     "*"   "*"    "*"    "*"
## 8  ( 1 ) "*"    "*"    "*"    "*"     "*"   "*"    "*"    "*"

##
##  Adjusted R squares of the selected models:
##
##  0.6584852 0.6830869 0.6973049 0.6967569 0.6942833 0.6912626 0.687801 0.6840614
```




```
## Model:

## glm(formula = Group ~ M1Left + Foramen + Rostrum, family = binomial(),
##     data = micro1)

## AIC values of the regsubsets model:  29.55

## P values of the intercept and features:
```

```
##             pvalues_regsubsets
## (Intercept)              0.002
## M1Left                   0.018
## Foramen                  0.030
## Rostrum                  0.490
```

A GLM model is fit using Group as the dependent variable and using all other features. Another GLM model is fit using only intercept.More GLM models were fit using higher order polynomials.

```
## Model:

## glm(formula = Group ~ ., family = binomial(), data = micro1)

## AIC values of the model with all variables:  32.96

## P values of the intercept and features:

##             pvalues_all.variables
## (Intercept)                 0.072
## M1Left                      0.099
## M2Left                      0.956
## M3Left                      0.103
## Foramen                     0.116
## Pbone                       0.578
## Length                      0.208
## Height                      0.156
## Rostrum                     0.450

## Model:

## glm(formula = Group ~ 1, family = binomial(), data = micro1)

## AIC values of the model with Intercept:  125.28

## P values of the intercept:

##   pvalues_Intercept
## 1             0.751

## Model:

## glm(formula = Group ~ M1Left + poly(M2Left, degree = 2) + poly(Pbone,
##     degree = 1), family = binomial(), data = micro1)

## AIC values of the Quadratic Model 1:  26.06

## P values of the intercept and Features:
```

```
##                            pvalues_Quadratic.Model.1
## (Intercept)                                    0.009
## M1Left                                         0.009
## poly(M2Left, degree = 2)1                      0.032
## poly(M2Left, degree = 2)2                      0.023
## poly(Pbone, degree = 1)                        0.071
```

```
## Model:
```

```
## glm(formula = Group ~ M1Left + poly(M2Left, degree = 2) + poly(M3Left,
##     degree = 2) + poly(Pbone, degree = 1) + poly(Height, degree = 2) +
##     poly(Length, degree = 2) + poly(Rostrum, degree = 2) + poly(Foramen,
##     degree = 2), family = binomial(), data = micro1)
```

```
## AIC values of the Quadratic Model 2:  30
```

```
## P values of the intercept and Features:
```

```
##                             pvalues_Quadratic.Model.2
## (Intercept)                                     1.000
## M1Left                                          1.000
## poly(M2Left, degree = 2)1                       1.000
## poly(M2Left, degree = 2)2                       0.999
## poly(M3Left, degree = 2)1                       0.999
## poly(M3Left, degree = 2)2                       1.000
## poly(Pbone, degree = 1)                         1.000
## poly(Height, degree = 2)1                       1.000
## poly(Height, degree = 2)2                       1.000
## poly(Length, degree = 2)1                       1.000
## poly(Length, degree = 2)2                       1.000
## poly(Rostrum, degree = 2)1                      1.000
## poly(Rostrum, degree = 2)2                      1.000
## poly(Foramen, degree = 2)1                      1.000
## poly(Foramen, degree = 2)2                      1.000
```

Further, Forward, Backward and Stepwise selection process is used to fit the models with best features. In Backward selection process the models are fit by subtracting one variable each time from given variables and picks the one that predicts the most on the dependent measure. From the summary of the model it is clear that the first model has a AIC of 32.96 and the final optimal model has only few variables and a AIC value of 27.7.

```
## Start:  AIC=32.96
## Group ~ M1Left + M2Left + M3Left + Foramen + Pbone + Length +
##     Height + Rostrum
##
##           Df Deviance    AIC
## - M2Left   1   14.965 30.965
## - Pbone    1   15.288 31.288
## - Rostrum  1   15.627 31.627
## <none>         14.962 32.962
## - Length   1   17.330 33.330
## - Height   1   18.744 34.744
```

```
## - Foramen  1   19.434 35.434
## - M3Left   1   20.654 36.654
## - M1Left   1   40.753 56.753
##
## Step:  AIC=30.97
## Group ~ M1Left + M3Left + Foramen + Pbone + Length + Height +
##     Rostrum
##
##           Df Deviance    AIC
## - Pbone    1   15.306 29.306
## - Rostrum  1   15.627 29.627
## <none>         14.965 30.965
## - Length   1   18.268 32.268
## - Height   1   18.945 32.945
## - Foramen  1   19.965 33.965
## - M3Left   1   20.763 34.763
## - M1Left   1   42.436 56.436
##
## Step:  AIC=29.31
## Group ~ M1Left + M3Left + Foramen + Length + Height + Rostrum
##
##           Df Deviance    AIC
## - Rostrum  1   15.703 27.703
## <none>         15.306 29.306
## - Length   1   18.625 30.625
## - Height   1   18.951 30.951
## - M3Left   1   20.855 32.855
## - Foramen  1   21.418 33.418
## - M1Left   1   42.970 54.970
##
## Step:  AIC=27.7
## Group ~ M1Left + M3Left + Foramen + Length + Height
##
##           Df Deviance    AIC
## <none>         15.703 27.703
## - Length   1   18.960 28.960
## - Height   1   19.019 29.019
## - M3Left   1   21.039 31.039
## - Foramen  1   21.463 31.463
## - M1Left   1   46.843 56.843

##
##
##  Model:

## glm(formula = Group ~ M1Left + M3Left + Foramen + Length + Height,
##     family = binomial(), data = micro1)

## AIC values of the Backward Selection:  27.7

## P values of the intercept and Features:

##             pvalues_Backward.Selection
```

```
## (Intercept)                    0.065
## M1Left                         0.029
## M3Left                         0.135
## Foramen                        0.097
## Length                         0.160
## Height                         0.191
```

Forward selection process the models are fit by adding one variable each time from given variables and picks the one that predicts the most on the dependent measure.Similar to the Backward selection process an optimal model with best variables is fit.

```
## Start:  AIC=125.28
## Group ~ 1
##
##           Df Deviance     AIC
## + M1Left   1   28.517  32.517
## + Rostrum  1   62.179  66.179
## + Height   1   67.588  71.588
## + Length   1   69.468  73.468
## + M2Left   1   75.984  79.984
## + Pbone    1   76.099  80.099
## + M3Left   1   79.294  83.294
## <none>        123.279 125.279
## + Foramen  1  121.465 125.465
##
## Step:  AIC=32.52
## Group ~ M1Left
##
##           Df Deviance    AIC
## + Foramen  1   22.049 28.049
## + Height   1   24.063 30.063
## + Pbone    1   26.025 32.025
## <none>         28.517 32.517
## + Length   1   28.286 34.286
## + M3Left   1   28.295 34.295
## + Rostrum  1   28.379 34.379
## + M2Left   1   28.430 34.430
##
## Step:  AIC=28.05
## Group ~ M1Left + Foramen
##
##           Df Deviance    AIC
## <none>         22.049 28.049
## + Height   1   21.100 29.100
## + Rostrum  1   21.553 29.553
## + M3Left   1   21.578 29.578
## + Pbone    1   21.738 29.738
## + Length   1   21.750 29.750
## + M2Left   1   21.758 29.758

## Model:

## glm(formula = Group ~ M1Left + Foramen, family = binomial(),
##     data = micro1)
```

```
## AIC values of the Forward Selection:  28.05
```

```
## P values of the intercept and Features:
```

```
##             pvalues_Forward.Selection
## (Intercept)                     0.065
## M1Left                          0.029
## M3Left                          0.135
## Foramen                         0.097
## Length                          0.160
## Height                          0.191
```

Stepwise selection is similar to Forward selection but a variable is removed if it is non significant. The Final optimal model is similar to the model obtained from Forward selection and has an AIC of 28.05.

```
## Start:  AIC=125.28
## Group ~ 1
##
##           Df Deviance     AIC
## + M1Left   1   28.517  32.517
## + Rostrum  1   62.179  66.179
## + Height   1   67.588  71.588
## + Length   1   69.468  73.468
## + M2Left   1   75.984  79.984
## + Pbone    1   76.099  80.099
## + M3Left   1   79.294  83.294
## <none>        123.279 125.279
## + Foramen  1  121.465 125.465
##
## Step:  AIC=32.52
## Group ~ M1Left
##
##           Df Deviance     AIC
## + Foramen  1   22.049  28.049
## + Height   1   24.063  30.063
## + Pbone    1   26.025  32.025
## <none>         28.517  32.517
## + Length   1   28.286  34.286
## + M3Left   1   28.295  34.295
## + Rostrum  1   28.379  34.379
## + M2Left   1   28.430  34.430
## - M1Left   1  123.279 125.279
##
## Step:  AIC=28.05
## Group ~ M1Left + Foramen
##
##           Df Deviance     AIC
## <none>         22.049  28.049
## + Height   1   21.100  29.100
## + Rostrum  1   21.553  29.553
## + M3Left   1   21.578  29.578
## + Pbone    1   21.738  29.738
## + Length   1   21.750  29.750
```

```
## + M2Left   1   21.758  29.758
## - Foramen  1   28.517  32.517
## - M1Left   1  121.465 125.465

##
##
##  Model:

## glm(formula = Group ~ M1Left + Foramen, family = binomial(),
##     data = micro1)

## AIC values of the Stepwise Selection:  28.05

## P values of the intercept and Features:

##             pvalues_Stepwise.Selection
## (Intercept)                      0.002
## M1Left                           0.001
## Foramen                          0.038
```

RESULTS AND DISCUSSION

AIC of all the models were compared.The lower the value of AIC better the model.The models obtained from Forward selection,Backward selection, Stepwise selection and Quadratic model(with variables M1Left, M2Left and pbone) have low values of AIC and were considered for further analysis.

```
##
##  AIC of all models:

##   Subset All_var Intercept Forward Backward Stepwise Quadretic.Mod.1
## 1   29.6      33     125.3      28     27.7       28            26.1
##   Quadretic.Mod.2
## 1              30
```

Both Forward and Stepwise selection models have same variables and same AIC. Model from Forward selection is used hereafter. Below table shows the p values of the three models. It is clear that the variables of the Backward model are not significant at 95% confidance interval except for M1Left variable though its AIC is similar to Forward selection model.Hence, Forward and Quadratic model (with variables M1Left, M2Left and pbone) is considered for further analysis.

```
##                           Forward Backward Quad.mod1
## (Intercept)...1             0.002       NA        NA
## M1Left...2                  0.001       NA        NA
## Foramen...3                 0.038       NA        NA
## (Intercept)...4                NA    0.065        NA
## M1Left...5                     NA    0.029        NA
## M3Left                         NA    0.135        NA
## Foramen...7                    NA    0.097        NA
## Length                         NA    0.160        NA
## Height                         NA    0.191        NA
## (Intercept)...10               NA       NA     0.009
```

```
## M1Left...11                    NA       NA     0.009
## poly(M2Left, degree = 2)1      NA       NA     0.032
## poly(M2Left, degree = 2)2      NA       NA     0.023
## poly(Pbone, degree = 1)        NA       NA     0.071
```

The AIC of Quadratic model is lower than Forward selection model but it is complex compared to the other model. Analysis of variace of these models show that they are marginally significant. Further, 10 fold cross validation is performed. The Error rate of the Forward selection model is 4.49% and is lower than the Quadratic model which had an error rate of 8.99%. Considering the facts that Forward selection model is simple and has a lower error rate ,this model is is used to analyze the Test data set.

```
## Analysis of Deviance Table
##
## Model 1: Group ~ M1Left + Foramen
## Model 2: Group ~ M1Left + poly(M2Left, degree = 2) + poly(Pbone, degree = 1)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        86     22.049
## 2        84     16.059  2   5.9905  0.05002 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Error Rate in % of model with Student and balance variables using Kfold approach : 4.49
```

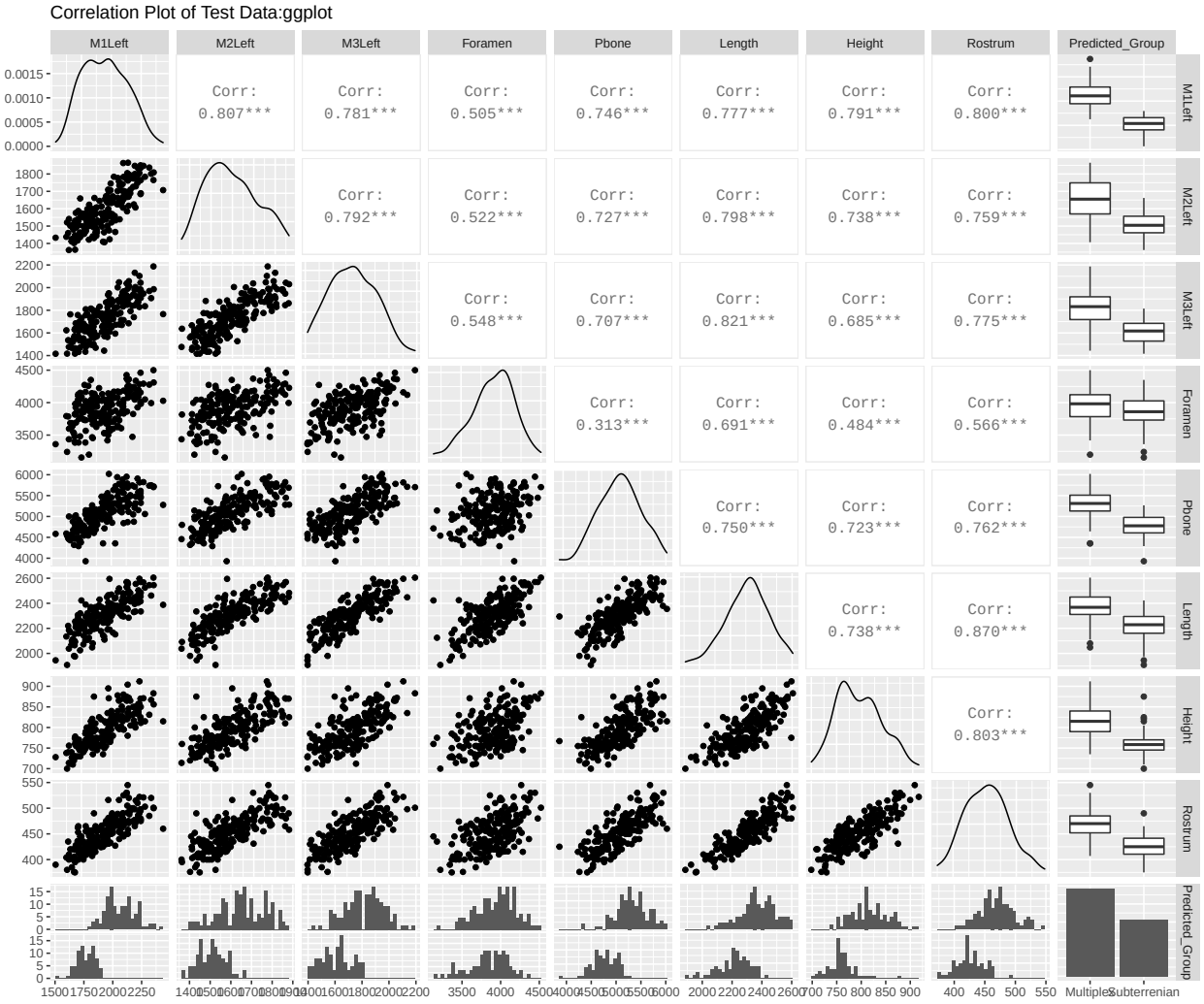
```
## Error Rate in % of model with only balance variables using Kfold approach : 8.99
```

Forward model is used to predict the specimen of test data. The predicted values are comibined to the test data. Dimension of the data is verified. After prediction is done 121 specimen were grouped as multiplex and 78 specimen were grouped into subterranean species.Pairs plot confirms similar trend in the values of all variable in relation to Group variable. The mean values of Multiplex species are on the higher side compared to the other species.Then the Test data with the predictions are exported as a Comma Separated File.

```
## Dimensions of the final dataset:  199 9
```

```
## Count of the Species falling into different groups:
```

```
##
##    Multiplex Subterrenian
##          121           78
```



CONCLUSION

Microtus data with records of Microtus specimen is used for the analysis. Data set with 89 specimen grouped into multiplex and subterranean is used as Training set and remaining Specimen of unknown origin is used as test data. Exploratory data analysis show that there are no missing values, and a non linear correlation between the Group variable and other variables. Feature selection is done using different methods like subset selection using *regsubsets()* function, Step selection using Forward, Backward and Stepwise directions, Quadratic models using polynomial terms, linear models with all variables and only intercept.The Forward selection, backward selection and Quadratic model with 3 variables have lower AIC values of 28,27,7 and 26.1 and were further analyzed. The p values of Backward model were not significant at 95% confidence interval except for the M1Left feature and was rejected. Analysis of Variance of the remaining two models are marginally significant. However, the 10 fold cross validation of these models showed a lower error rate of 4.49% for model obtained from Forward selection compared to the Quadratic model that had 8.99% error rate.The simple linear model obtained from Forward selection process is used to predict the test data.121 specimen were classified as multiplex species and 78 specimen were classified as subterranean species.

REFERENCES

- Snigdha Peddi,*Stat 601 Homework Assignment 3*

- CRAN,*microtus:Microtus classification(more vole data)*, (https://rdrr.io/cran/Flury/man/microtus.html)
- Lecture from Big Edu Youtube Channel,*Feature Selection in R programming|stepwise Regression|Machine Learning|Data Science*,April 20,2020 ,(https://www.youtube.com/watch?v=QKIsRYBkNCc)
- Lecture from Dragonfly Statistics Youtube Channel,*Backward Elimination-stepwise Regression with R*, October 18,2017,(https://www.youtube.com/watch?v=0aTtMJO-pE4)
- Lecture from Dragonfly Statistics Youtube Channel,*Stepwise Regression in R-Combining Forward and Backward Selection)*,October 18,2017,(https://www.youtube.com/watch?v=ejR8LnQziPY)
- stackoverflow blogpost,*Extract pvalue from glm*,(https://stackoverflow.com/questions/23838937/extract-pvalue-from-glm)