In this project hierarchical clustering technique is applied on Abalone dataset.

Accessing Dataset

```
df<-read.csv("abalone.csv") #accessing dataset from csv file
head(df,10)#printing top 10 data points
```

```
##    Sex Length Diameter Height Whole.weight Shucked.weight Viscera.weight
## 1    M  0.455    0.365  0.095       0.5140         0.2245         0.1010
## 2    M  0.350    0.265  0.090       0.2255         0.0995         0.0485
## 3    F  0.530    0.420  0.135       0.6770         0.2565         0.1415
## 4    M  0.440    0.365  0.125       0.5160         0.2155         0.1140
## 5    I  0.330    0.255  0.080       0.2050         0.0895         0.0395
## 6    I  0.425    0.300  0.095       0.3515         0.1410         0.0775
## 7    F  0.530    0.415  0.150       0.7775         0.2370         0.1415
## 8    F  0.545    0.425  0.125       0.7680         0.2940         0.1495
## 9    M  0.475    0.370  0.125       0.5095         0.2165         0.1125
## 10   F  0.550    0.440  0.150       0.8945         0.3145         0.1510
##    Shell.weight Rings
## 1         0.150    15
## 2         0.070     7
## 3         0.210     9
## 4         0.155    10
## 5         0.055     7
## 6         0.120     8
## 7         0.330    20
## 8         0.260    16
## 9         0.165     9
## 10        0.320    19
```

Data Cleaning and converting categorical value of ‘Sex’ feature to numeric value

```
df<-na.omit(df) #droping NA values if any in dataset
nrow(df)
```

```
## [1] 4177
```

```
df$Sex[df$Sex=="M"]<-0 # male=0
df$Sex[df$Sex=="F"]<-1 # female=1
df$Sex[df$Sex=="I"]<-2 # Infant=2
df$Sex<-as.integer(df$Sex)
head(df,10)
```

```
##    Sex Length Diameter Height Whole.weight Shucked.weight Viscera.weight
## 1    0  0.455    0.365  0.095       0.5140         0.2245         0.1010
## 2    0  0.350    0.265  0.090       0.2255         0.0995         0.0485
## 3    1  0.530    0.420  0.135       0.6770         0.2565         0.1415
## 4    0  0.440    0.365  0.125       0.5160         0.2155         0.1140
## 5    2  0.330    0.255  0.080       0.2050         0.0895         0.0395
## 6    2  0.425    0.300  0.095       0.3515         0.1410         0.0775
## 7    1  0.530    0.415  0.150       0.7775         0.2370         0.1415
## 8    1  0.545    0.425  0.125       0.7680         0.2940         0.1495
## 9    0  0.475    0.370  0.125       0.5095         0.2165         0.1125
## 10   1  0.550    0.440  0.150       0.8945         0.3145         0.1510
##    Shell.weight Rings
## 1         0.150    15
## 2         0.070     7
## 3         0.210     9
## 4         0.155    10
## 5         0.055     7
## 6         0.120     8
## 7         0.330    20
## 8         0.260    16
## 9         0.165     9
## 10        0.320    19
```

Data Scaling

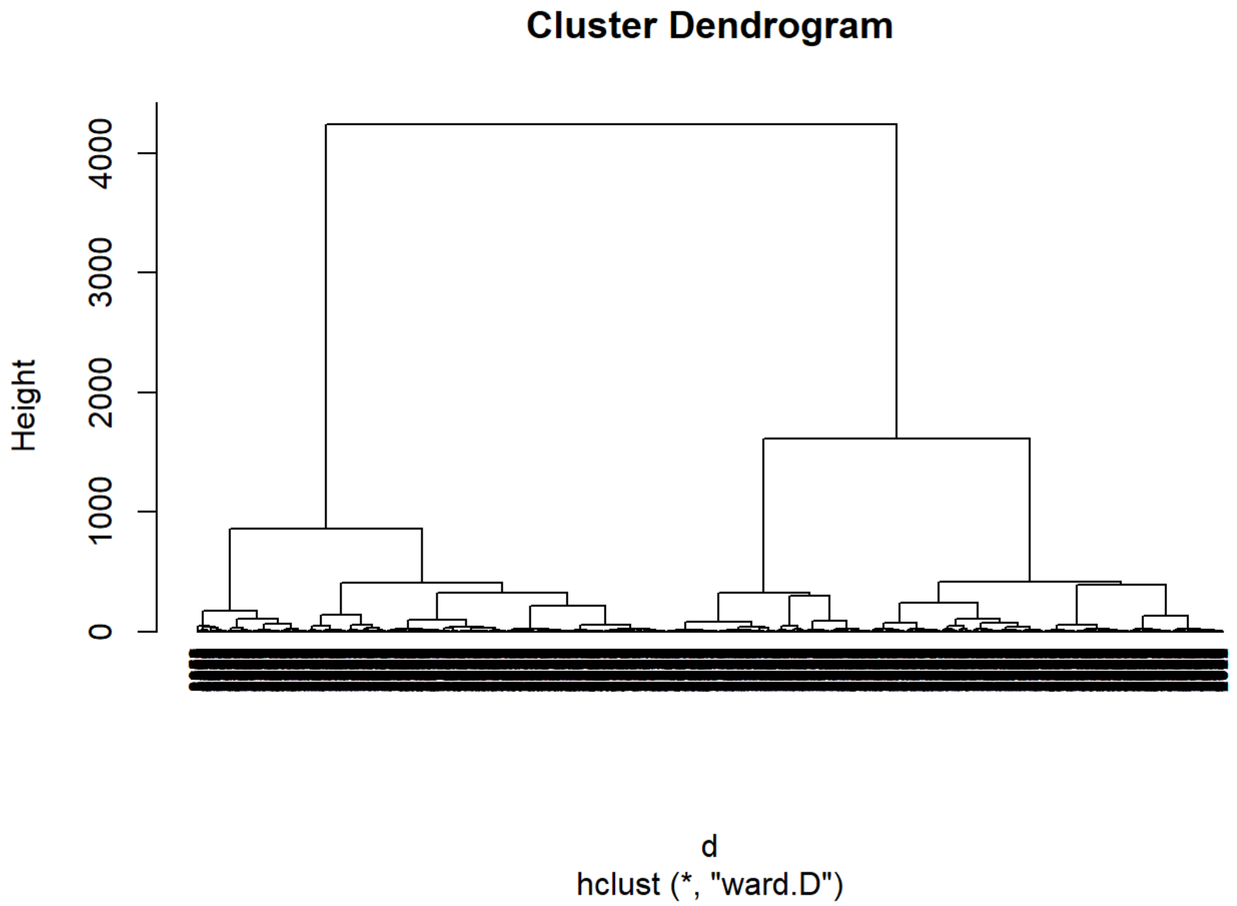
```
scaled_df<-scale(df) #scaling
scaled_df<-data.frame(scaled_df) #Converting dataset from array to dataframe
```

finding Euclidean distance

```
d<-dist(scaled_df,method="euclidean") #Calcualting distance between points
```

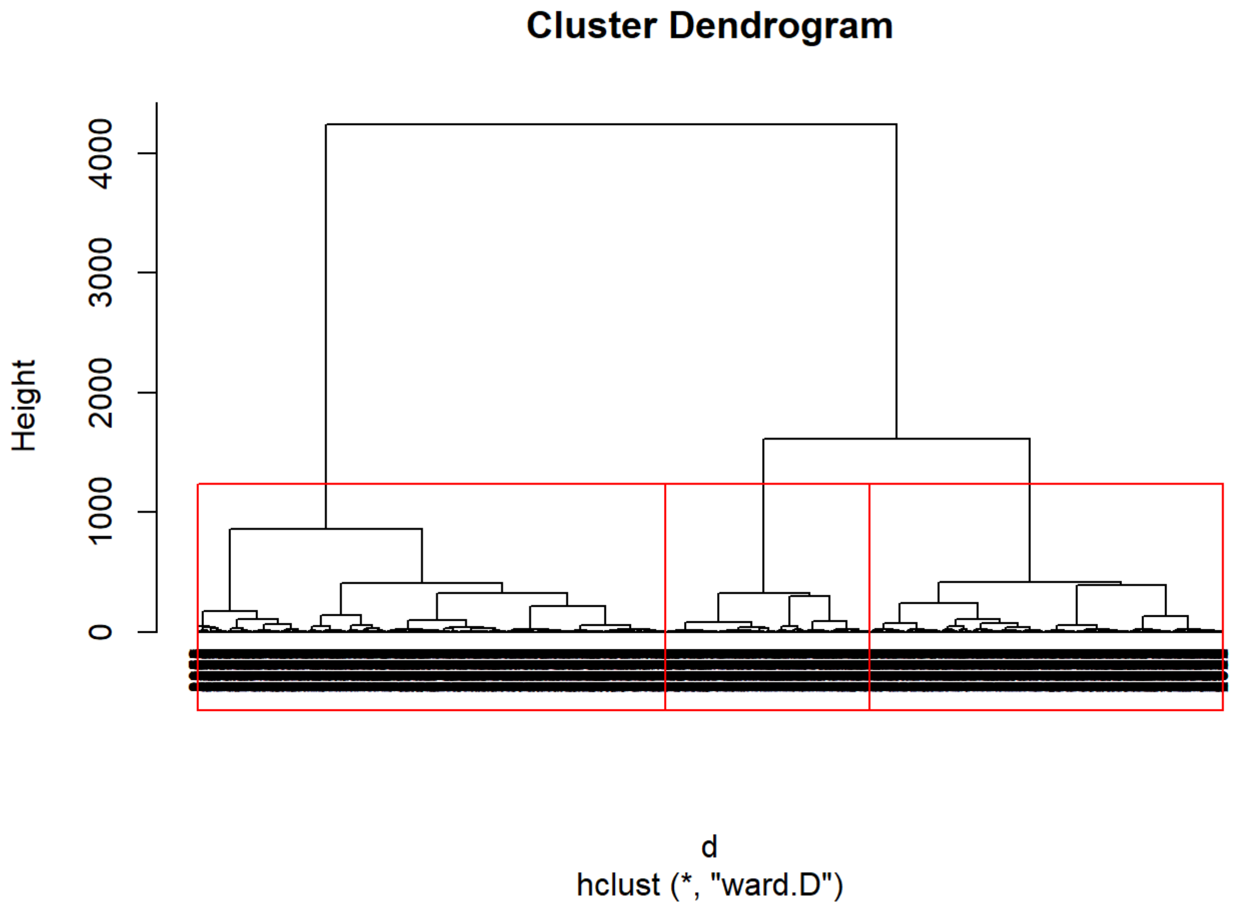
Generating dendrogram

```
cl<-hclust(d,method="ward.D") #creating cluster using ward’s method
plot(cl,cex=0.6,hang=-1) #ploting dendrogram
```



Creating Splitting boundary for 3 clusters

```
plot(cl,cex=0.6,hang=-1) #ploting dendrogram
rect.hclust(cl,k=3,border='red') #create rectangular boundary for cluster on dendrogram
```



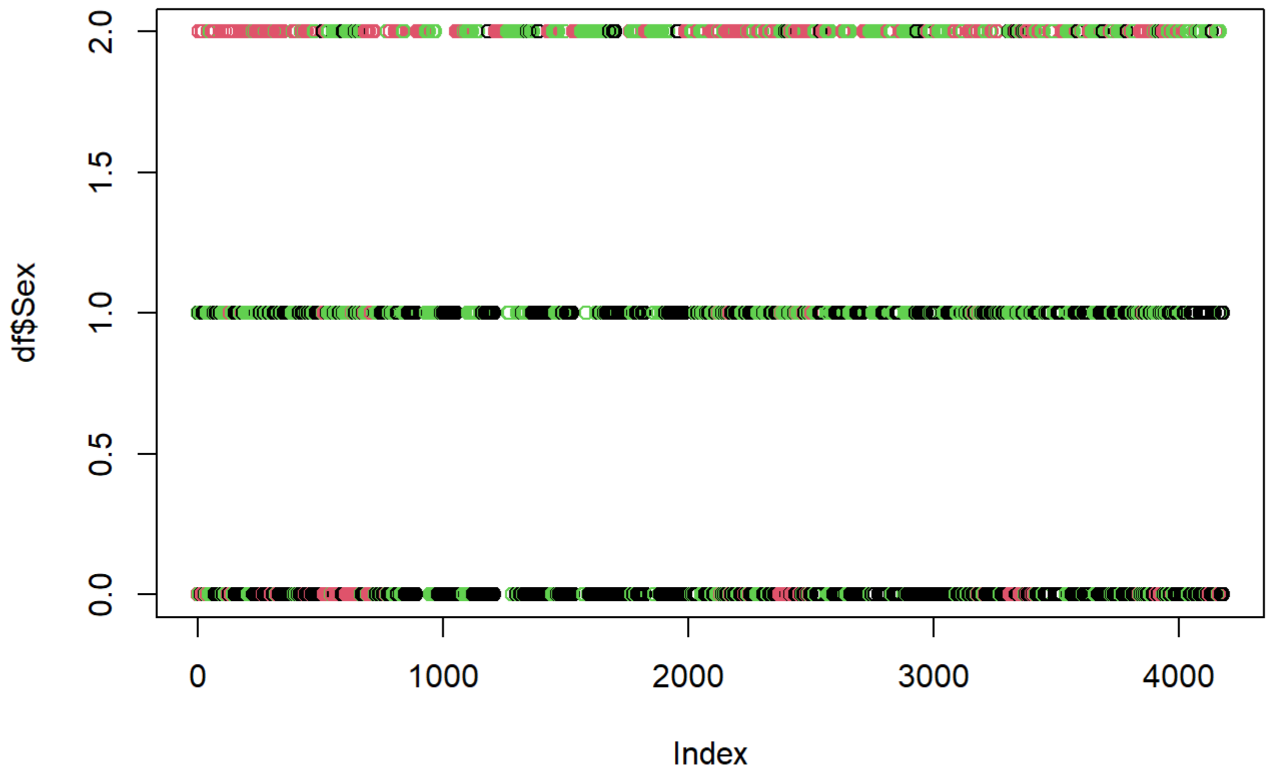
fetching data for 3 cluster

```
clusters<- cutree(cl,3) #fetching 3 clusters
table(clusters) #number of data points after division into clusters
```

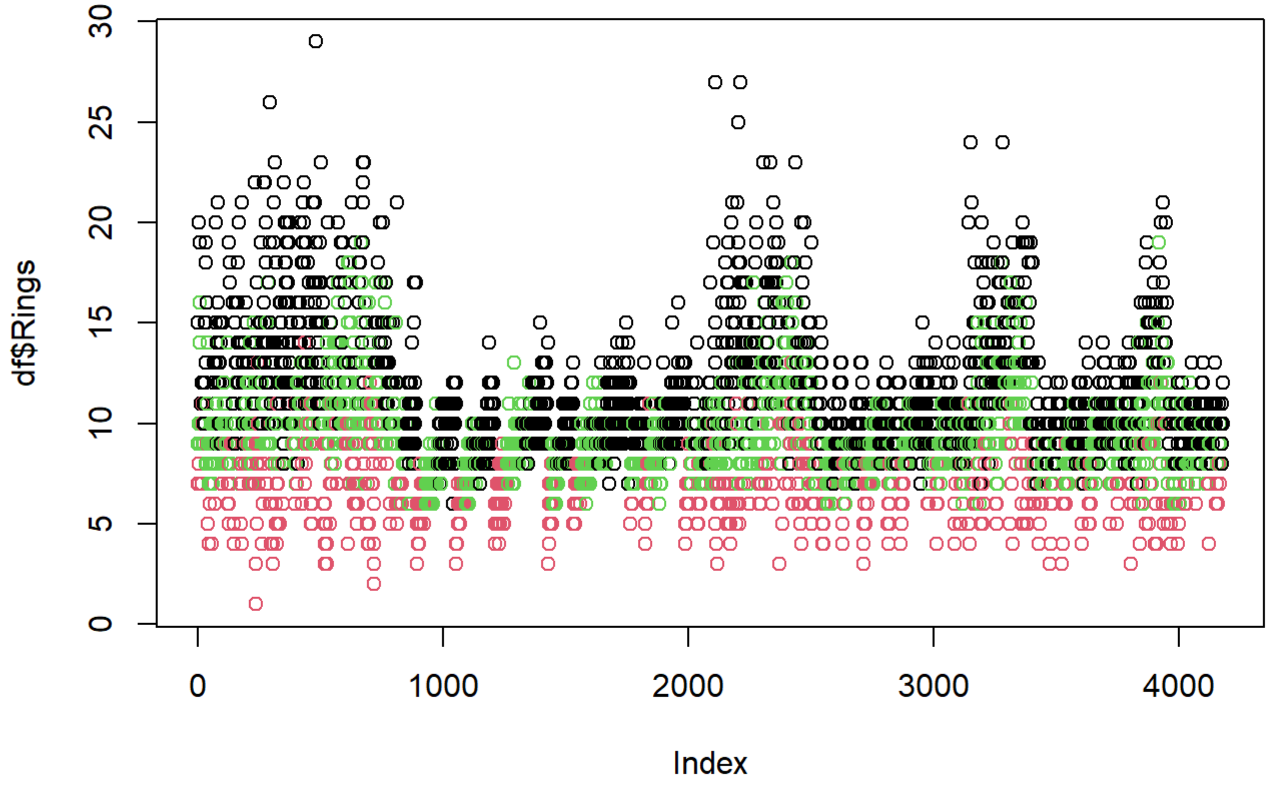
```
## clusters
##    1    2    3 
## 1905  833 1439
```

Plotting clusters

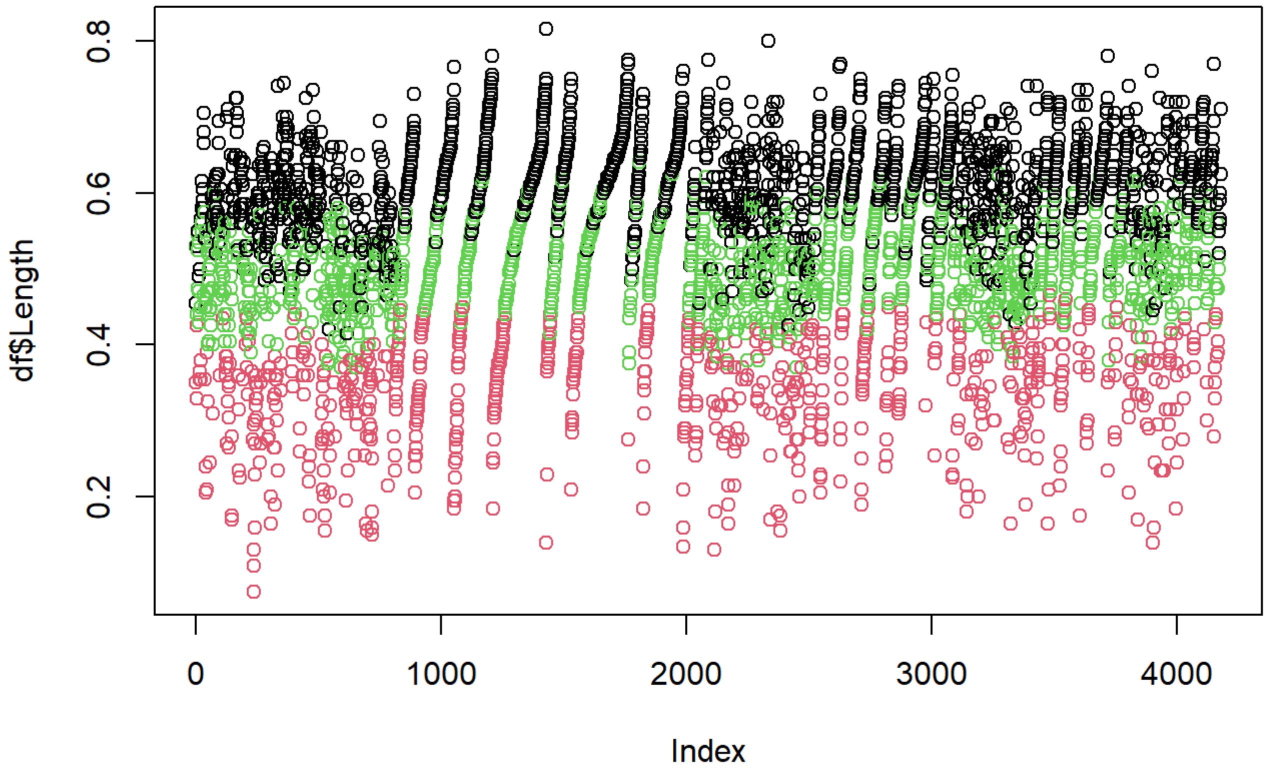
```
plot(df$Sex ,col=clusters) #plotting division of Sex feature into clusters
```



```
plot(df$Rings ,col=clusters)  #plotting division of Rings feature into clusters
```



```
plot(df$Length ,col=clusters) #plotting division of Length feature data points in clusters
```



OBSERVATION

Cluster 1 :- Abalones with length less than 0.4mm having less number of rings 5 to 9 are mostly infants. They are less in number than male and female abalone.

Cluster 2 :- Abalones with length 0.4mm to 0.5mm having number of rings 6 to 16 are mostly females.

Cluster 3 :- Abalones with length greater than 0.5mm having number of rings 9 to 29 are mostly males.

CONCLUSION

Using Hierarchical clustering, this model differentiates the categorical variable and also the numerical variables in abalone dataset. It recognizes pattern and based on the similarity and groups data accordingly to the clusters. Data is grouped into 3 clusters and observations are done based on features Rings, Sex and Length of abalone. After analyzing it is seen that there is a lack of young abalone this can be a statistical anomaly or a byproduct of the data collection process. However, young abalones have a number of predators, including crabs, lobsters, starfish and octopus. It would be useful to know whether or not any of these invasive species has infiltrated the abalone beds. Abalone size is also a very important feature because larger females can produce many more eggs possibly a million or more. Hence, a certain stock of larger female abalone may be necessary to maintain larger populations and helps saving abalone species from becoming extinct.