Name: Prashant Pandey

ID: 16200112

STAT 40340 Assignment 2

Q1.

```
library(MASS)

hotelling_test <- function(homeData){
homesData.PA <- subset(homeData, homeData$Area =="PA")  #find the data with A
rea PA
homesData.MP <- subset(homeData, homeData$Area =="MP")    #find the data with 
Area MP
covPA <- cov(homesData.PA[,-1])    # covariance for PA
covMP <- cov(homesData.MP[,-1])    # covariance for MP
estimate_cov <- (((nrow(homesData.PA)-1)*covPA) + ((nrow(homesData.MP)-1)*cov
MP))/((nrow(homesData.PA)+nrow(homesData.MP))-2)
meanPA <- colMeans(homesData.PA[,-1])
meanMP <- colMeans(homesData.MP[,-1])
mahalanobis_dist<- (t(meanPA-meanMP)%*%solve(estimate_cov)%*%(meanPA-meanMP))
numeratr <- nrow(homesData.PA) + nrow(homesData.MP) - ncol(homesData.PA[,-1])
- 1
denom <- (nrow(homesData.PA) + nrow(homesData.MP) - 2) * ncol(homesData.PA[,-
1])
hotelling_tsq <- (nrow(homesData.PA) * nrow(homesData.MP))/ (nrow(homesData.P
A) + nrow(homesData.MP))
hotelling_tsq <- hotelling_tsq * mahalanobis_dist
f_stat <- (numeratr / denom) * hotelling_tsq
p_val<-pf(f_stat,df1=ncol(homesData.PA[,-1]), df2=numeratr,lower.tail=FALSE)
print(p_val)
if(p_val < 0.05) {  # 0.05 significance level
print("The two communities are significantly different with respect to the ch
aracteristics of the properties available for sale")
} else
  {
  print("The two communities are NOT significantly different with respect to 
the characteristics of the properties available for sale")
  }
}
data <- read.csv("prices.csv")    # read prices data
hotelling_test(data)  # test the function
```

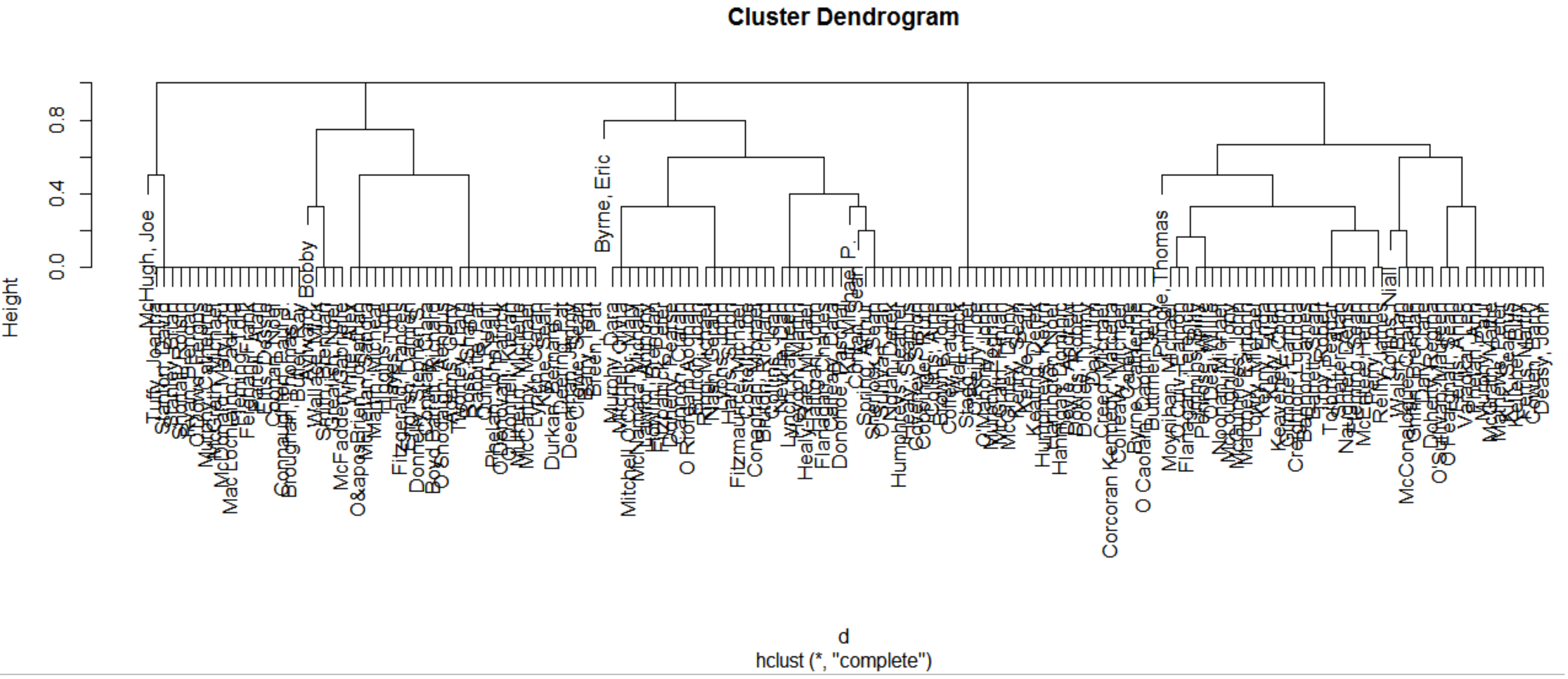
```
          [,1]
[1,] 0.2272253
[1] "The two communities are NOT significantly different with respect to the 
characteristics of the properties available for sale"
```

The p-value, 0.2272253 is greater than 0.05 significance level, hence we **fail to reject the null hypothesis** and the two communities are not significantly different.

Q2(a).

```
load("2016_First6Votes_YesNoAbsent.Rdata")
bin_data <- (votes==1)*1   #convert data in binary values
d<- dist(bin_data,method = "binary")  # binary dissimilarity
c<- hclust(d,method = "complete") #hierarchical clustering
plot(c)
hcl = cutree(c, k = 5)
print(table(hcl))
```

Hierarchical clustering can be used. We can find dissimilarity between binary data vectors using Jaccard dissimilarity measure.



```
1  2  3  4  5
24 46 42 35 19   #cluster assignment
```

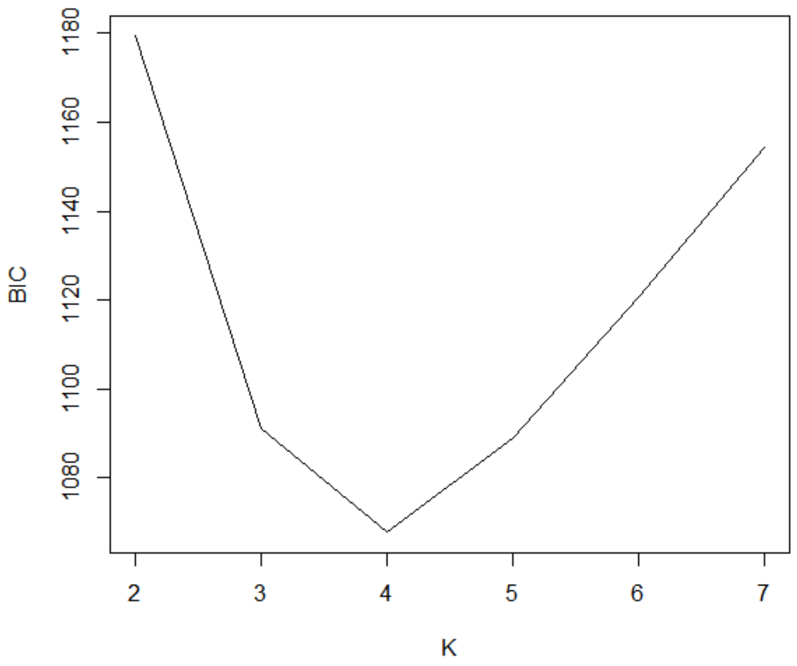
K=5 seems to be a good option from the above dendrogram.

Q2(b)i) binary presence/absence data:

Code for BIC (binary data):

```
library(MASS)
library(poLCA)
load("2016_First6Votes_YesNoAbsent.Rdata")
bin_data <- ((votes==1)*1) +1    # 1 absent 2 present
bin_dataframe<-as.data.frame(bin_data)
rownames(bin_dataframe)<-NULL
f<- cbind(ED1,ED2,Credit,Confidence1,Confidence2,Trade)~1   #formula for poLCA
bic_array=vector()
aic_array=vector()
for(k in 2:7){             # check for cluster k=2 to 7
  min_bic=100000
  min_aic=100000
  for(j in 1:500){        # try to avoid local maxima by running multiple times
    res1<-poLCA(f, bin_dataframe, nclass = k, maxiter = 10000)
    if(res1$bic < min_bic)
    {
      min_bic = res1$bic
    }
    if(res1$aic < min_aic)
    {
      min_aic = res1$aic
    }
  }
  bic_array<-c(bic_array,c(min_bic))
  aic_array<-c(aic_array,c(min_aic))
}
plot(bic_array,x=c(2:7),t='l',xlab = "K", ylab="BIC")
```

Bayesian Information Criteria(BIC) is calculated for different values of K (number of clusters) and plotted against K using poLCA function. Looking at the plot, for K=4, the BIC is smallest. Hence **K=4** seems to be a good choice for **binary presence/absence data**. I ran poLCA for 10000 iterations. Also for each cluster, I ran PCA 500 times to avoid local maxima.



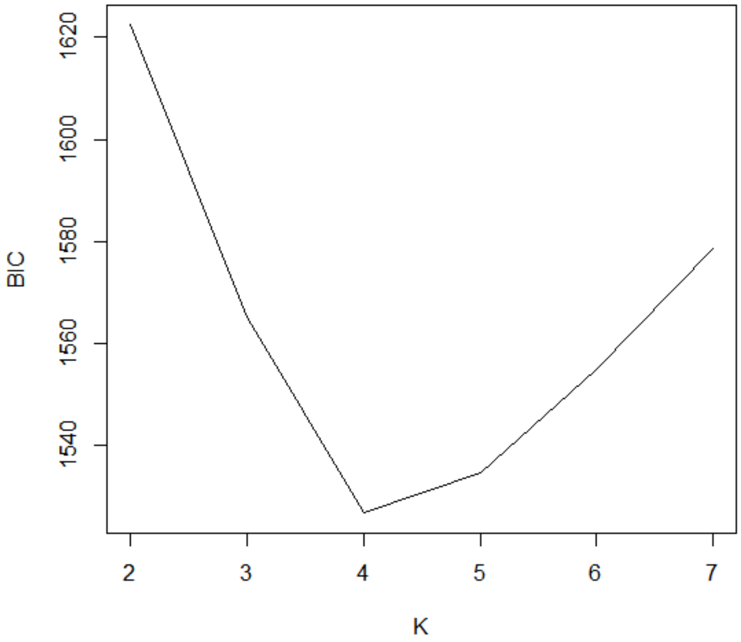
ii) **polytomous voting data:**

Code for BIC (polytomous data):

```
load("2016_First6Votes_YesNoAbsent.Rdata")

cat_dataframe<-as.data.frame(votes)
rownames(cat_dataframe)<-NULL
f<- cbind(ED1,ED2,Credit,Confidence1,Confidence2,Trade)~1   #formula for poLCA
bic_array=vector()
aic_array=vector()
for(k in 2:7){             # check for cluster k=2 to 7
  min_bic=100000
  min_aic=100000
  for(j in 1:500){        # try to avoid local maxima by running multiple times
    res<-poLCA(f, cat_dataframe, nclass = k, maxiter = 10000)
    if(res$bic < min_bic)
    {
      min_bic = res$bic
    }
    if(res$aic < min_aic)
    {
      min_aic = res$aic
    }
  }
  bic_array<-c(bic_array,c(min_bic))
  aic_array<-c(aic_array,c(min_aic))
}
plot(bic_array,x=c(2:7),t='l',xlab = "K", ylab="BIC")
```

Bayesian Information Criteria(BIC) is calculated for different values of K (number of clusters) and plotted against K using poLCA function. Looking at the plot, for K=4, the BIC is smallest. Hence **K=4** seems to be a good choice for **polytomous voting data**. I ran poLCA for 10000 iterations. Also for each cluster, I ran PCA 500 times to avoid local maxima.



Q2(c)

<u>Comparison</u>:

Between hierarchical clustering and LCA using binary presence/absent data:

```
adjustedRandIndex(hcl,res1$predclass)
[1] 0.4834326
```

Adjusted rand index is 0.4834 which is low so there is significant disagreement between two clustering's. There should be, as hierarchical clustering suggested five clusters while LCA suggested four based on BIC.

Partition of 166 TD’s using hierarchical clustering (with 5 clusters):

```
 1  2  3  4  5
24 46 42 35 19
```

Between LCA using binary presence/absent data and polytomous voting data:

Mixing proportion for polytomous voting data

```
> res$P
[1] 0.1927711 0.3433735 0.2351865 0.2286689
```

Partition of 166 TD’s using LCA for polytomous data (with 4 clusters):

```
> table(res1$predclass)

1  2  3  4
32 57 39 38
```

Mixing proportion for binary present/absent data:

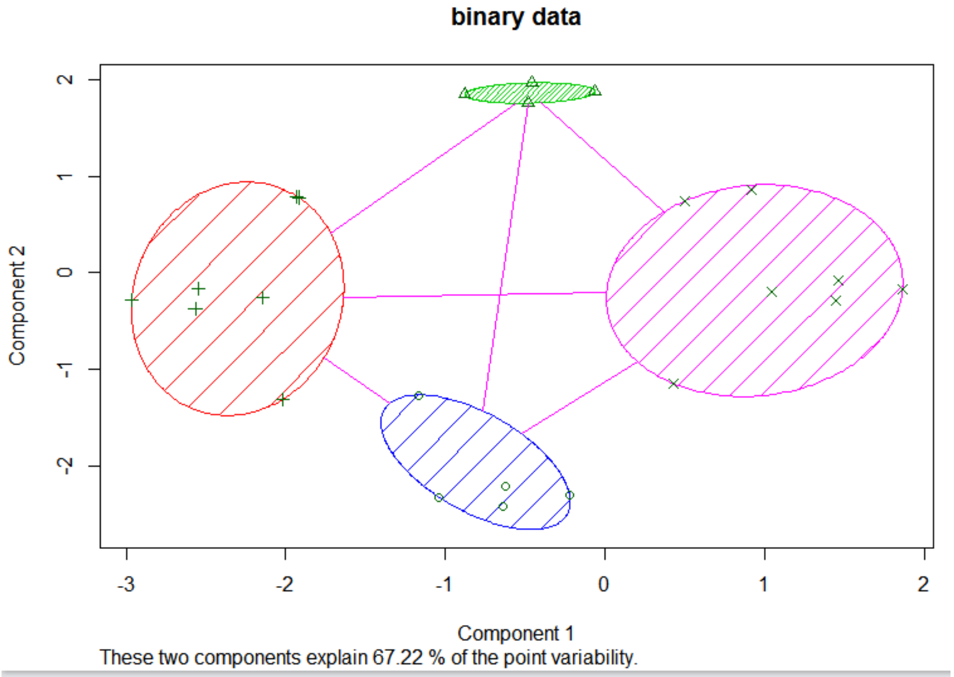
```
> res1$P
[1] 0.1385542 0.2347241 0.1749145 0.4518072
```

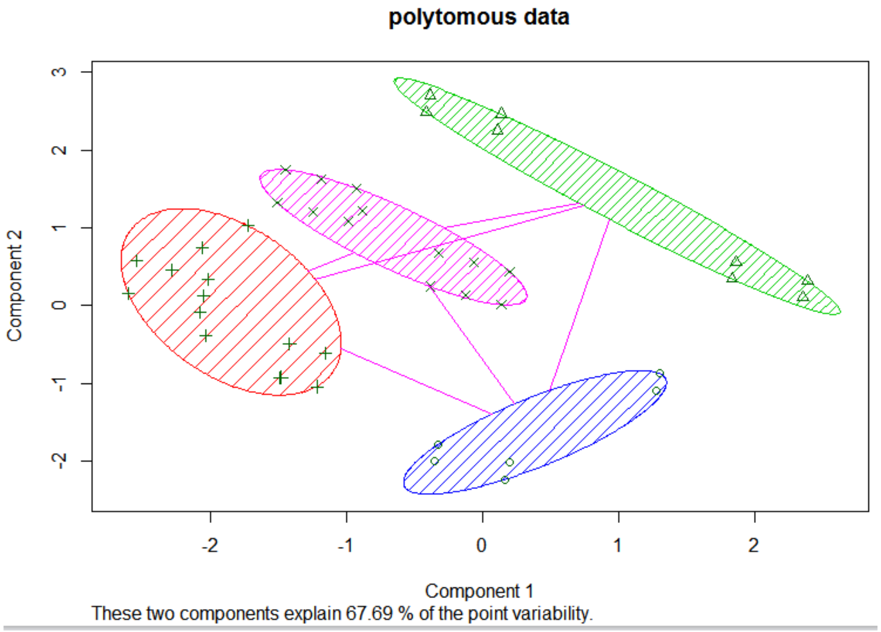
Partition of 166 TD’s using LCA for binary data (with 4 clusters):

```
> table(res1$predclass)

 1  2  3  4
23 39 29 75

> adjustedRandIndex(res$predclass,res1$predclass)
[1] 0.4936148
```





With polytomous data, the mixing proportions are less skewed as compared to binary data.