Homework 3. Clustering Practice (80 Points)

Nikhil Ambati

2023-10-14

Part 1. USArrests Dataset and Hierarchical Clustering (20 Points)

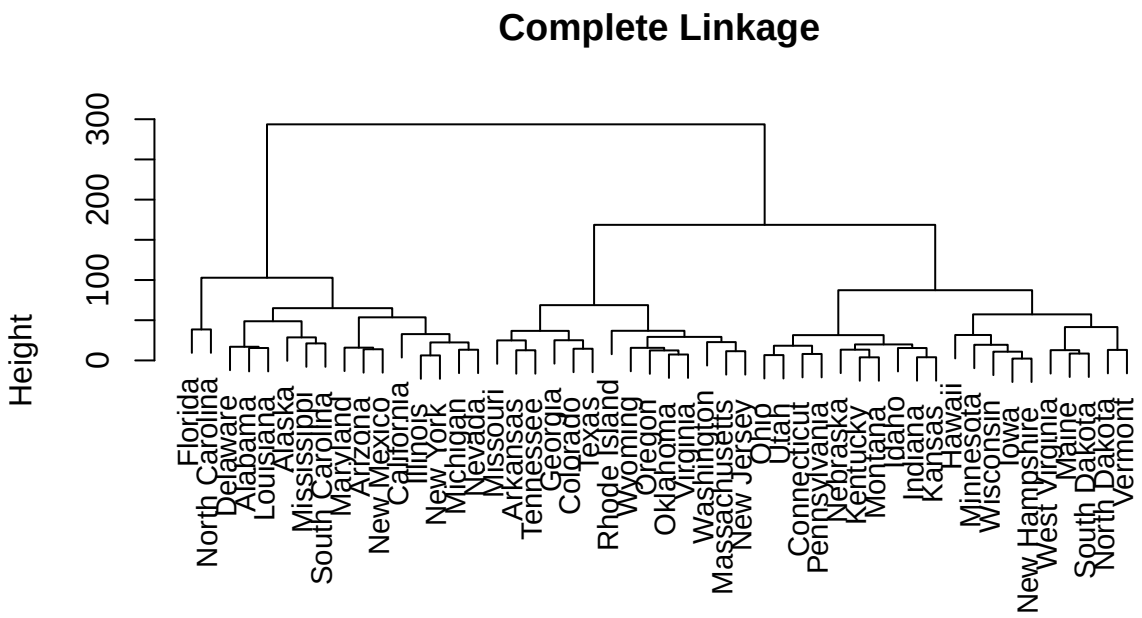
Consider the "USArrests" data. It is a built-in dataset you may directly get in RStudio. Perform hierarchical clustering on the observations (states) and answer the following questions.

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

Q1.1. Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states. (5 points)

```
set.seed(200)
clust_USArrests <- hclust(dist(USArrests),method='complete')
plot(clust_USArrests, main = "Complete Linkage",
    xlab = "", sub = "", cex = .9)
```

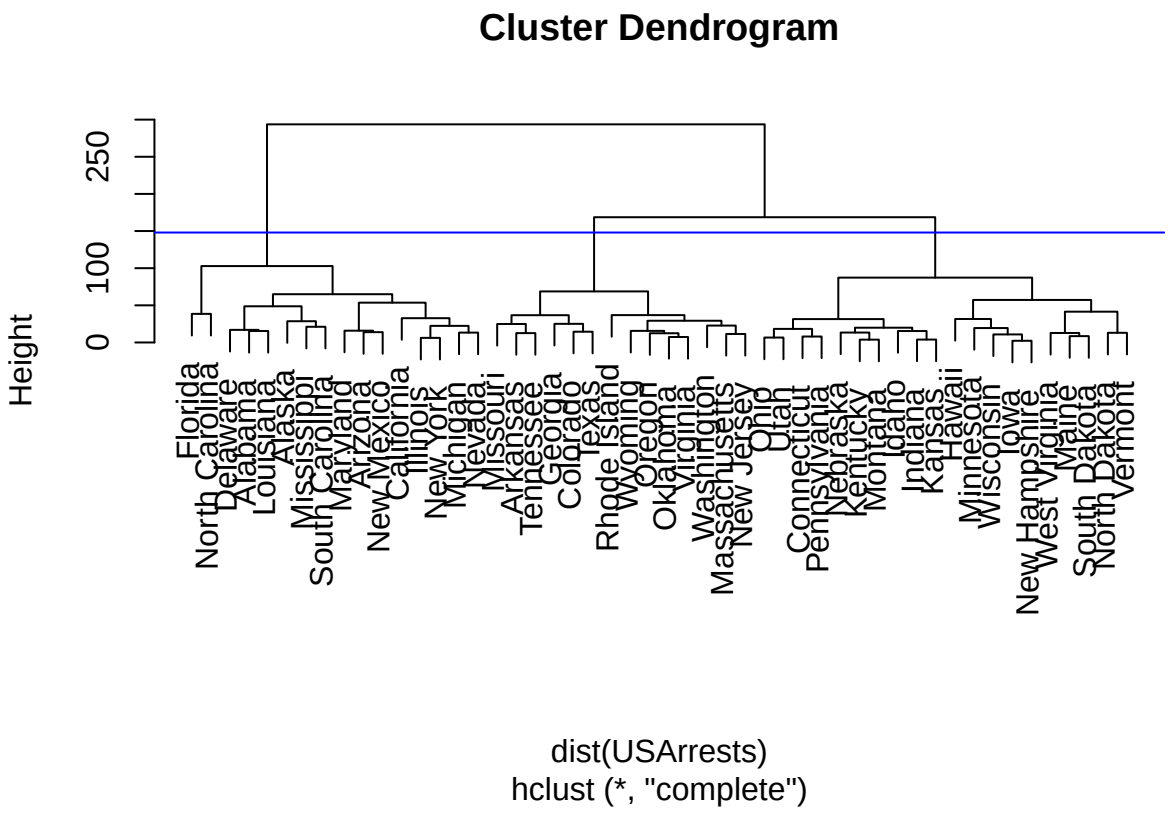


Q1.2. Cut the dendrogram at a height that results in three distinct clusters. Interpret the clusters. Which states belong to which clusters? (5 points)

```
set.seed(200)
cut <- cutree(clust_USArrests, 3)
table(cut)
```

```
## cut
##  1  2  3
## 16 14 20
```

```
plot(clust_USArrests)
abline(h=148, col="blue")
```

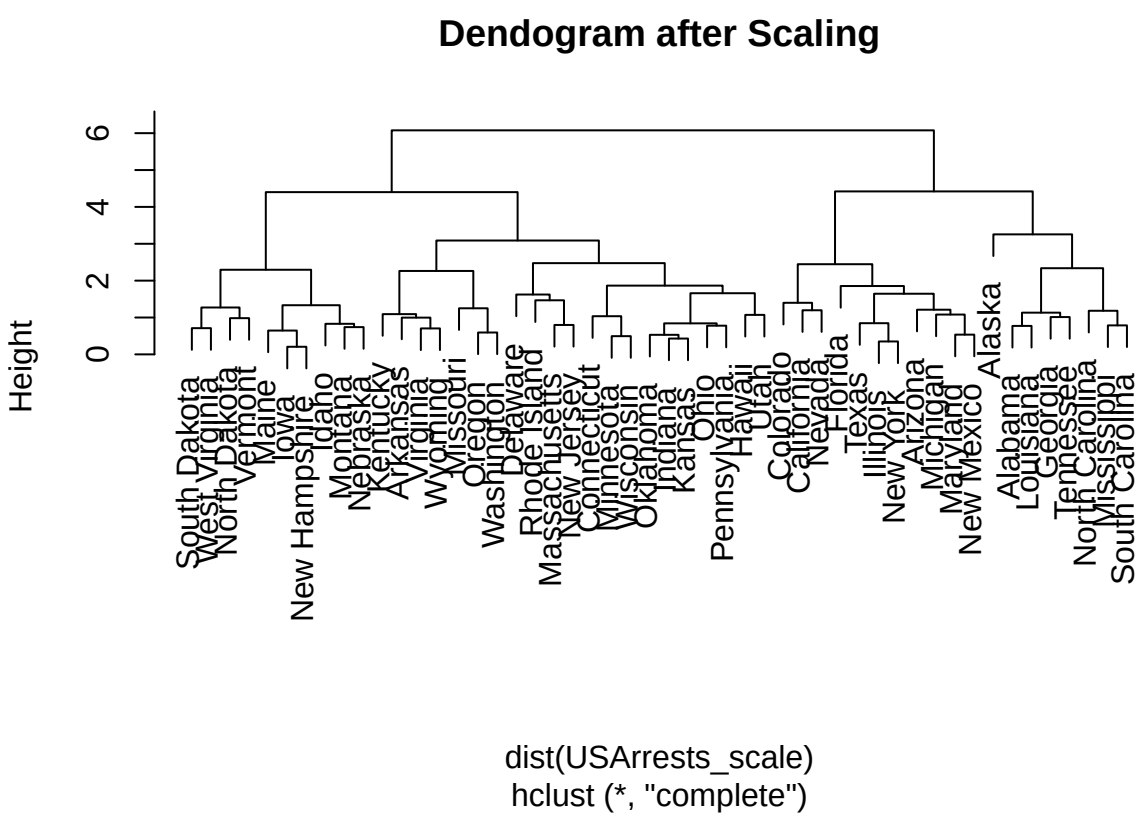



```
table(cut)
```

```
## cut
##  1  2  3
## 16 14 20
```

Q1.3 Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one. Obtain three clusters. Which states belong to which clusters?(5 points)

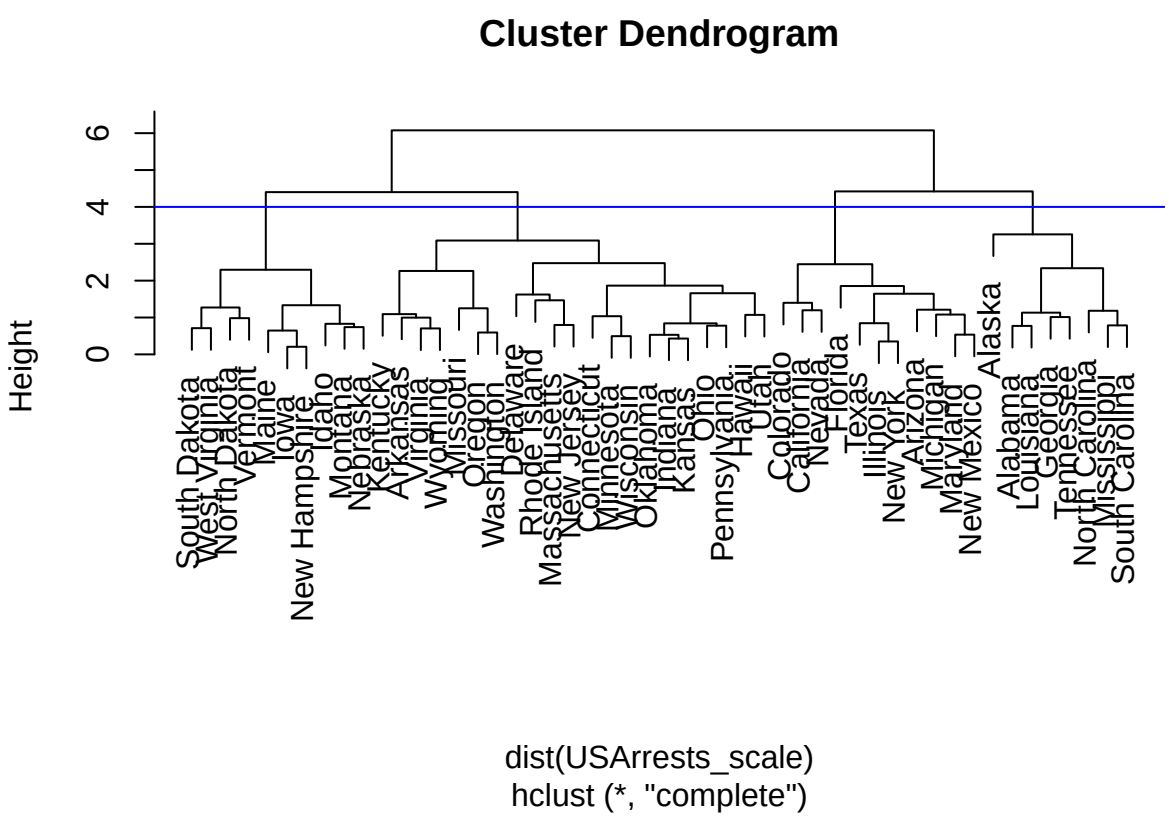
```
set.seed(200)
USArrests_scale <- scale(USArrests)
cl_USArrests_scale <- hclust(dist(USArrests_scale),method='complete')
plot(cl_USArrests_scale,main='Dendogram after Scaling')
```




```
set.seed(200)
scale_cut <- cutree(cl_USArrests_scale, 3)
table(scale_cut)
```

```
## scale_cut
##  1  2  3
##  8 11 31
```

```
plot(cl_USArrests_scale)
abline(h=4,col="BLUE")
```




```
scale_cut
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              2              3              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              3              2              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              2              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              2              3              1              3
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              2              3              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              3              3              3              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              1              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              3              3              3              3              3
```

Q1.4 What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer. *(5 points)*

Answer: Scaling the variables before performing hierarchical clustering is a important step . Failure to scale the variables can lead to skewed clustering results, where certain variables with higher variance may impact the influence the clustering process. Scaling the variables is essential for a more balanced and reliable hierarchical clustering analysis. Scaling also cover smaller and and entire set clusters.

Part 2. Market Segmentation (60 Points)

An advertisement division of large club store needs to perform customer analysis the store customers in order to create a segmentation for more targeted marketing campaign

You task is to identify similar customers and characterize them (at least some of them). In other word perform clustering and identify customers segmentation.

This data-set is derived from https://www.kaggle.com/imakash3011/customer-personality-analysis

```
Colomns description:
People
  ID: Customer's unique identifier
  Year_Birth: Customer's birth year
  Education: Customer's education level
  Marital_Status: Customer's marital status
  Income: Customer's yearly household income
  Kidhome: Number of children in customer's household
  Teenhome: Number of teenagers in customer's household
  Dt_Customer: Date of customer's enrollment with the company
  Recency: Number of days since customer's last purchase
  Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Products

  MntWines: Amount spent on wine in last 2 years
  MntFruits: Amount spent on fruits in last 2 years
  MntMeatProducts: Amount spent on meat in last 2 years
  MntFishProducts: Amount spent on fish in last 2 years
  MntSweetProducts: Amount spent on sweets in last 2 years
  MntGoldProds: Amount spent on gold in last 2 years

Place
  NumWebPurchases: Number of purchases made through the company's website
  NumStorePurchases: Number of purchases made directly in stores
```

Assume that data was current on 2014-07-01

Q2.1. Read Dataset and Data Conversion to Proper Data Format *(12 points)*

Read "m_marketing_campaign.csv" using `data.table::fread` command, examine the data.

```
# fread m_marketing_campaign.csv and save it as df (2 points)
data_market <- fread("C:/Users/Nikhil/Sem 2/SDM 2/HomeWork3/m_marketing_campaign.csv")
df_market <- as.data.frame(data_market)
head(df_market)
```

```
##     ID Year_Birth Education Marital_Status Income Kidhome Teenhome Dt_Customer
```

```
## 1 5524       1957  Bachelor         Single  58138       0        0  04-09-2012
## 2 2174       1954  Bachelor         Single  46344       1        1  08-03-2014
## 3 4141       1965  Bachelor       Together  71613       0        0  21-08-2013
## 4 6182       1984  Bachelor       Together  26646       1        0  10-02-2014
## 5 5324       1981       PhD        Married  58293       1        0  19-01-2014
## 6 7446       1967    Master       Together  62513       0        1  09-09-2013
##   Recency MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1      58      635        88             546             172               88
## 2      38       11         1               6               2                1
## 3      26      426        49             127             111               21
## 4      26       11         4              20              10                3
## 5      94      173        43             118              46               27
## 6      16      520        42              98               0               42
##   MntGoldProds NumWebPurchases NumStorePurchases Complain
## 1           88               8                 4        0
## 2            6               1                 2        0
## 3           42               8                10        0
## 4            5               2                 4        0
## 5           15               5                 6        0
## 6           14               6                10        0
```

```
summary(df_market)
```

```
##        ID          Year_Birth    Education         Marital_Status
##  Min.   :    0   Min.   :1893   Length:2209        Length:2209
##  1st Qu.: 2826   1st Qu.:1959   Class :character   Class :character
##  Median : 5462   Median :1970   Mode  :character   Mode  :character
##  Mean   : 5592   Mean   :1969
##  3rd Qu.: 8427   3rd Qu.:1977
##  Max.   :11191   Max.   :1996
##      Income          Kidhome          Teenhome      Dt_Customer
##  Min.   :  1730   Min.   :0.0000   Min.   :0.0000   Length:2209
##  1st Qu.: 35246   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median : 51390   Median :0.0000   Median :0.0000   Mode  :character
##  Mean   : 52244   Mean   :0.4418   Mean   :0.5052
##  3rd Qu.: 68627   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :666666   Max.   :2.0000   Max.   :2.0000
##     Recency         MntWines        MntFruits      MntMeatProducts
##  Min.   : 0.00   Min.   :   0.0   Min.   :  0.00   Min.   :   0.0
##  1st Qu.:24.00   1st Qu.:  24.0   1st Qu.:  2.00   1st Qu.:  16.0
##  Median :49.00   Median : 174.0   Median :  8.00   Median :  68.0
##  Mean   :49.08   Mean   : 305.2   Mean   : 26.35   Mean   : 167.2
##  3rd Qu.:74.00   3rd Qu.: 505.0   3rd Qu.: 33.00   3rd Qu.: 233.0
##  Max.   :99.00   Max.   :1493.0   Max.   :199.00   Max.   :1725.0
##  MntFishProducts  MntSweetProducts  MntGoldProds    NumWebPurchases
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 0.000
##  1st Qu.:  3.00   1st Qu.:  1.00   1st Qu.:  9.00   1st Qu.: 2.000
##  Median : 12.00   Median :  8.00   Median : 24.00   Median : 4.000
##  Mean   : 37.56   Mean   : 27.07   Mean   : 43.85   Mean   : 4.082
##  3rd Qu.: 50.00   3rd Qu.: 33.00   3rd Qu.: 56.00   3rd Qu.: 6.000
##  Max.   :259.00   Max.   :262.00   Max.   :321.00   Max.   :27.000
##  NumStorePurchases    Complain
##  Min.   : 0.000    Min.   :0.000000
##  1st Qu.: 3.000    1st Qu.:0.000000
```

```
##  Median : 5.000    Median :0.000000
##  Mean   : 5.803    Mean   :0.009507
##  3rd Qu.: 8.000    3rd Qu.:0.000000
##  Max.   :13.000    Max.   :1.000000
```

```r
# Convert Year_Birth to Age (assume that current date is 2014-07-01) (2 points)
current_date <- as.Date("2014-07-01")
current_year <- as.numeric(format(current_date, format = "%Y"))
df_market$Age <- ((current_year)-df_market$Year_Birth)

df_market$Dt_Customer<-as.Date(as.Date(df_market$Dt_Customer,"%d-%m-%Y"),"%Y-%m-%d")

df_market$MembershipDays <- as.numeric(floor( difftime(current_date,df_market$Dt_Customer,units="days")))
head(df_market)
```

```
##     ID Year_Birth Education Marital_Status Income Kidhome Teenhome Dt_Customer
## 1 5524       1957  Bachelor         Single  58138       0        0  2012-09-04
## 2 2174       1954  Bachelor         Single  46344       1        1  2014-03-08
## 3 4141       1965  Bachelor       Together  71613       0        0  2013-08-21
## 4 6182       1984  Bachelor       Together  26646       1        0  2014-02-10
## 5 5324       1981       PhD        Married  58293       1        0  2014-01-19
## 6 7446       1967    Master       Together  62513       0        1  2013-09-09
##   Recency MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1      58      635        88             546             172               88
## 2      38       11         1               6               2                1
## 3      26      426        49             127             111               21
## 4      26       11         4              20              10                3
## 5      94      173        43             118              46               27
## 6      16      520        42              98               0               42
##   MntGoldProds NumWebPurchases NumStorePurchases Complain Age MembershipDays
## 1           88               8                 4        0  57            665
## 2            6               1                 2        0  60            115
## 3           42               8                10        0  49            314
## 4            5               2                 4        0  30            141
## 5           15               5                 6        0  33            163
## 6           14               6                10        0  47            295
```

```r
# Summarize Education column (use table function) (2 points)

table(df_market$Education)
```

```
##
##  Associate   Bachelor HighSchool     Master        PhD
##        200       1114         54        363        478
```

```r
# Lets create a new column EducationLevel from Education
# Lets treat Education column as ordinal categories and use years in education as a levels
# for distance calculations (2 points)
# Assuming following order and years spend for education:
#    HighSchool (13 years), Associate(15 years), Bachelor(17 years), Master(19 years), PhD(22 years)
# create EducationLevel from Education
# hint: use recode function (in mutate statement)
```

```r
df_market <- df_market %>%
  mutate(EducationLevel = recode(Education,"HighSchool"=13,"Associate"=15,"Bachelor"=17,"PhD"=22,"Maste
head(df_market)
```

```
##     ID Year_Birth Education Marital_Status Income Kidhome Teenhome Dt_Customer
## 1 5524       1957  Bachelor         Single  58138       0        0  2012-09-04
## 2 2174       1954  Bachelor         Single  46344       1        1  2014-03-08
## 3 4141       1965  Bachelor       Together  71613       0        0  2013-08-21
## 4 6182       1984  Bachelor       Together  26646       1        0  2014-02-10
## 5 5324       1981       PhD        Married  58293       1        0  2014-01-19
## 6 7446       1967    Master       Together  62513       0        1  2013-09-09
##   Recency MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1      58      635        88             546             172               88
## 2      38       11         1               6               2                1
## 3      26      426        49             127             111               21
## 4      26       11         4              20              10                3
## 5      94      173        43             118              46               27
## 6      16      520        42              98               0               42
##   MntGoldProds NumWebPurchases NumStorePurchases Complain Age MembershipDays
## 1           88               8                 4        0  57           665
## 2            6               1                 2        0  60           115
## 3           42               8                10        0  49           314
## 4            5               2                 4        0  30           141
## 5           15               5                 6        0  33           163
## 6           14               6                10        0  47           295
##   EducationLevel
## 1             17
## 2             17
## 3             17
## 4             17
## 5             22
## 6             19
```

```r
# Summarize Marital_Status column (use table function)


# Lets convert single Marital_Status categories for 5 separate binary categories  (2 points)
# Divorced, Married, Single, Together and Widow, the value will be 1 if customer
# is in that category and 0 if customer is not
# hint: use dummy_cols from fastDummies or dummyVars from caret package, model.matrix
# or simple comparison (there are only 5 groups)
# Keep Marital_Status for later use
df_market<-fastDummies::dummy_cols(df_market,select_columns="Marital_Status")
head(df_market)
```

```
##     ID Year_Birth Education Marital_Status Income Kidhome Teenhome Dt_Customer
## 1 5524       1957  Bachelor         Single  58138       0        0  2012-09-04
## 2 2174       1954  Bachelor         Single  46344       1        1  2014-03-08
## 3 4141       1965  Bachelor       Together  71613       0        0  2013-08-21
## 4 6182       1984  Bachelor       Together  26646       1        0  2014-02-10
## 5 5324       1981       PhD        Married  58293       1        0  2014-01-19
## 6 7446       1967    Master       Together  62513       0        1  2013-09-09
##   Recency MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
```

```
## 1      58      635        88              546              172                88
## 2      38       11         1                6                2                 1
## 3      26      426        49              127              111                21
## 4      26       11         4               20               10                 3
## 5      94      173        43              118               46                27
## 6      16      520        42               98                0                42
##   MntGoldProds NumWebPurchases NumStorePurchases Complain Age MembershipDays
## 1           88               8                 4        0  57            665
## 2            6               1                 2        0  60            115
## 3           42               8                10        0  49            314
## 4            5               2                 4        0  30            141
## 5           15               5                 6        0  33            163
## 6           14               6                10        0  47            295
##   EducationLevel Marital_Status_Divorced Marital_Status_Married
## 1             17                       0                      0
## 2             17                       0                      0
## 3             17                       0                      0
## 4             17                       0                      0
## 5             22                       0                      1
## 6             19                       0                      0
##   Marital_Status_Single Marital_Status_Together Marital_Status_Widow
## 1                     1                       0                   0
## 2                     1                       0                   0
## 3                     0                       1                   0
## 4                     0                       1                   0
## 5                     0                       0                   0
## 6                     0                       1                   0
```

```
# lets remove columns which we will no longer use:
# remove ID, Year_Birth, Dt_Customer, Education, Marital_Status
# and save it as df_sel
df_sel<-select(df_market,-c("ID","Year_Birth", "Dt_Customer","Education","Marital_Status"))
head(df_sel)
```

```
##   Income Kidhome Teenhome Recency MntWines MntFruits MntMeatProducts
## 1  58138       0        0      58      635        88             546
## 2  46344       1        1      38       11         1               6
## 3  71613       0        0      26      426        49             127
## 4  26646       1        0      26       11         4              20
## 5  58293       1        0      94      173        43             118
## 6  62513       0        1      16      520        42              98
##   MntFishProducts MntSweetProducts MntGoldProds NumWebPurchases
## 1             172               88          88                8
## 2               2                1           6                1
## 3             111               21          42                8
## 4              10                3           5                2
## 5              46               27          15                5
## 6               0               42          14                6
##   NumStorePurchases Complain Age MembershipDays EducationLevel
## 1                 4        0  57            665             17
## 2                 2        0  60            115             17
## 3                10        0  49            314             17
## 4                 4        0  30            141             17
## 5                 6        0  33            163             22
```

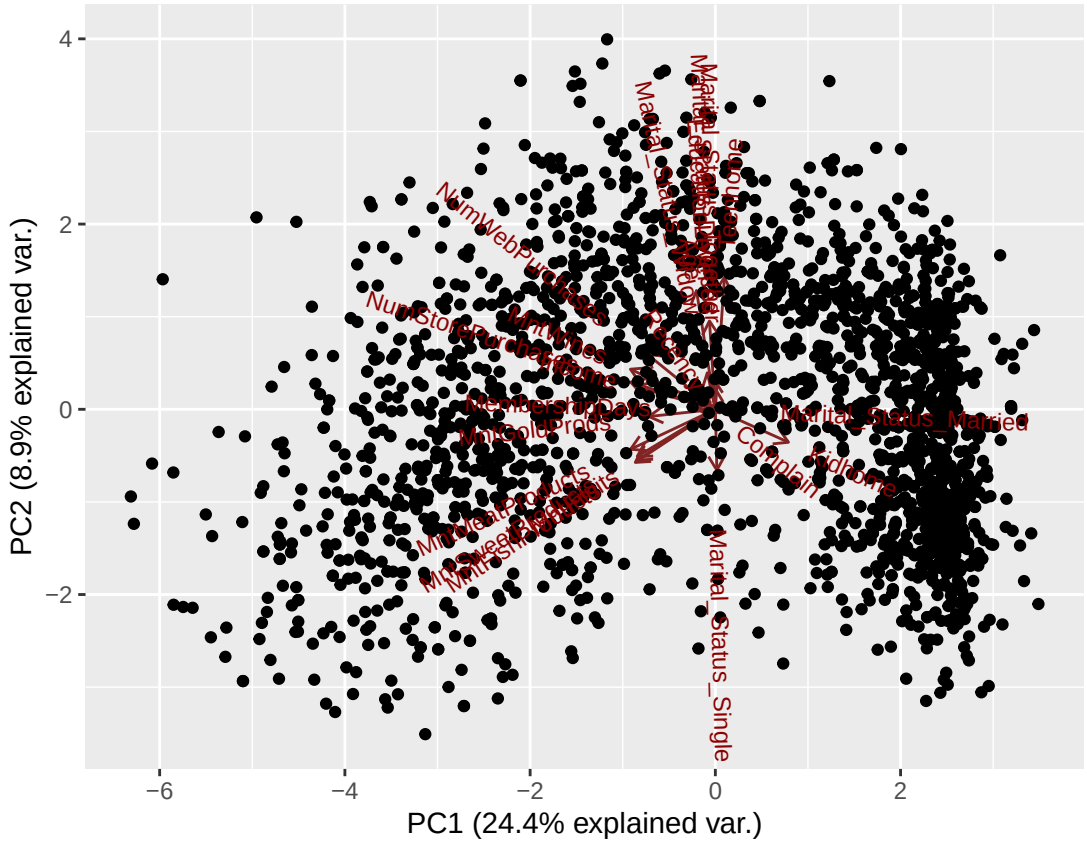
```
## 6                10        0  47            295             19
##   Marital_Status_Divorced Marital_Status_Married Marital_Status_Single
## 1                       0                     0                     1
## 2                       0                     0                     1
## 3                       0                     0                     0
## 4                       0                     0                     0
## 5                       0                     1                     0
## 6                       0                     0                     0
##   Marital_Status_Together Marital_Status_Widow
## 1                       0                    0
## 2                       0                    0
## 3                       1                    0
## 4                       1                    0
## 5                       0                    0
## 6                       1                    0
```

```
# lets scale (2 points)
# run scale function on df_sel and save it as df_scale
# that will be our scaled values which we will use for analysis
df_scale<-scale(df_sel)
write.csv(df_scale,file="dsfjndsf.csv")
```

PCA

Q2.2. Run PCA, make biplot and scree plot *(6 points)*

```
# Run PCA on df_scale, make biplot and scree plot/percentage variance explained plot
# save as pc_out, we will use pc_out$x[,1] and pc_out$x[,2] later for plotting

pc_out<-prcomp(df_scale)
ggbiplot(pc_out,scale=0)
```

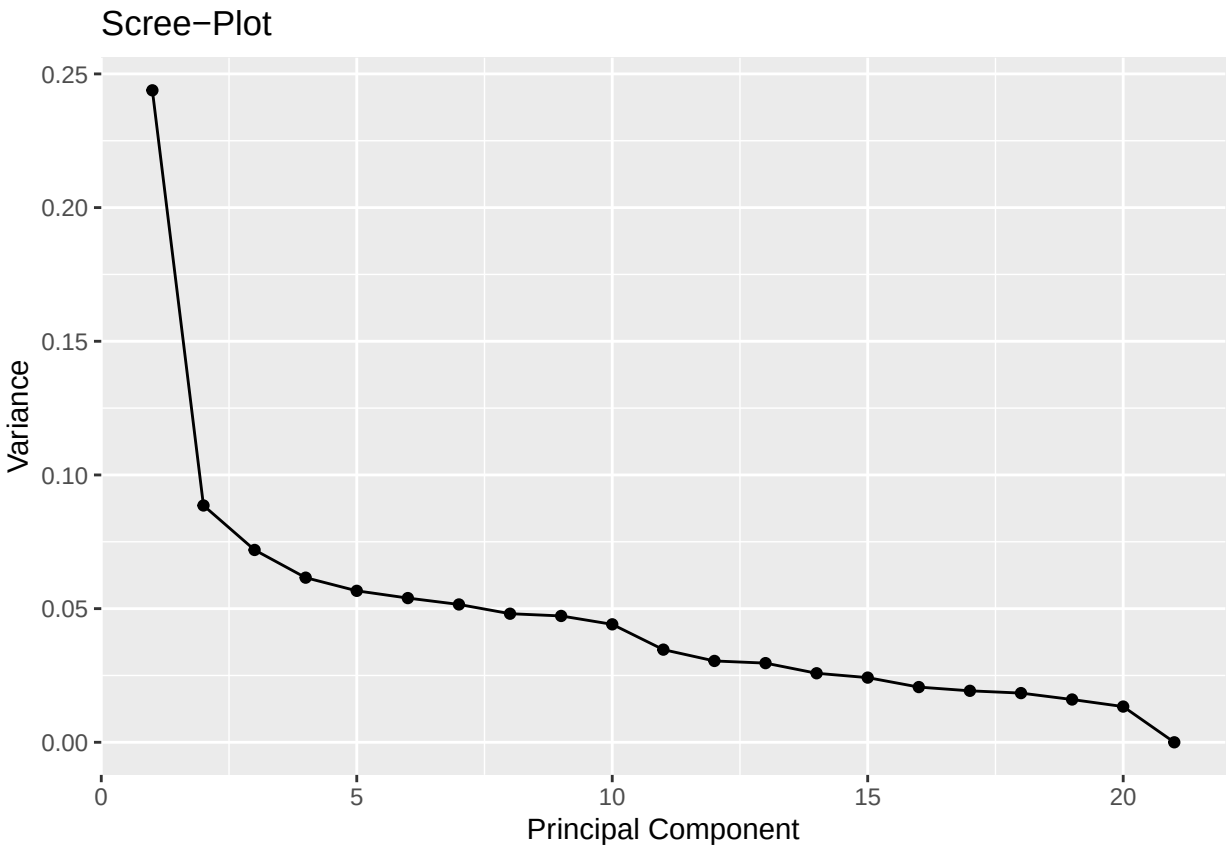



```
var <- (pc_out$sdev)^2 / sum(pc_out$sdev^2)
round(var,3)
```

```
##  [1] 0.244 0.089 0.072 0.062 0.057 0.054 0.052 0.048 0.047 0.044 0.035 0.030
## [13] 0.030 0.026 0.024 0.021 0.019 0.018 0.016 0.013 0.000
```

```
qplot(c(1:21), var) +
  geom_line() +
  xlab("Principal Component") +
  ylab("Variance") +
  ggtitle("Scree-Plot") +
  scale_y_continuous(breaks = seq(0, 0.30, 0.05))
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Q2.3 Comment on observation (any visible distinct clusters?) *(2 points)*

I could see we already have 2 clusters from the above plot.

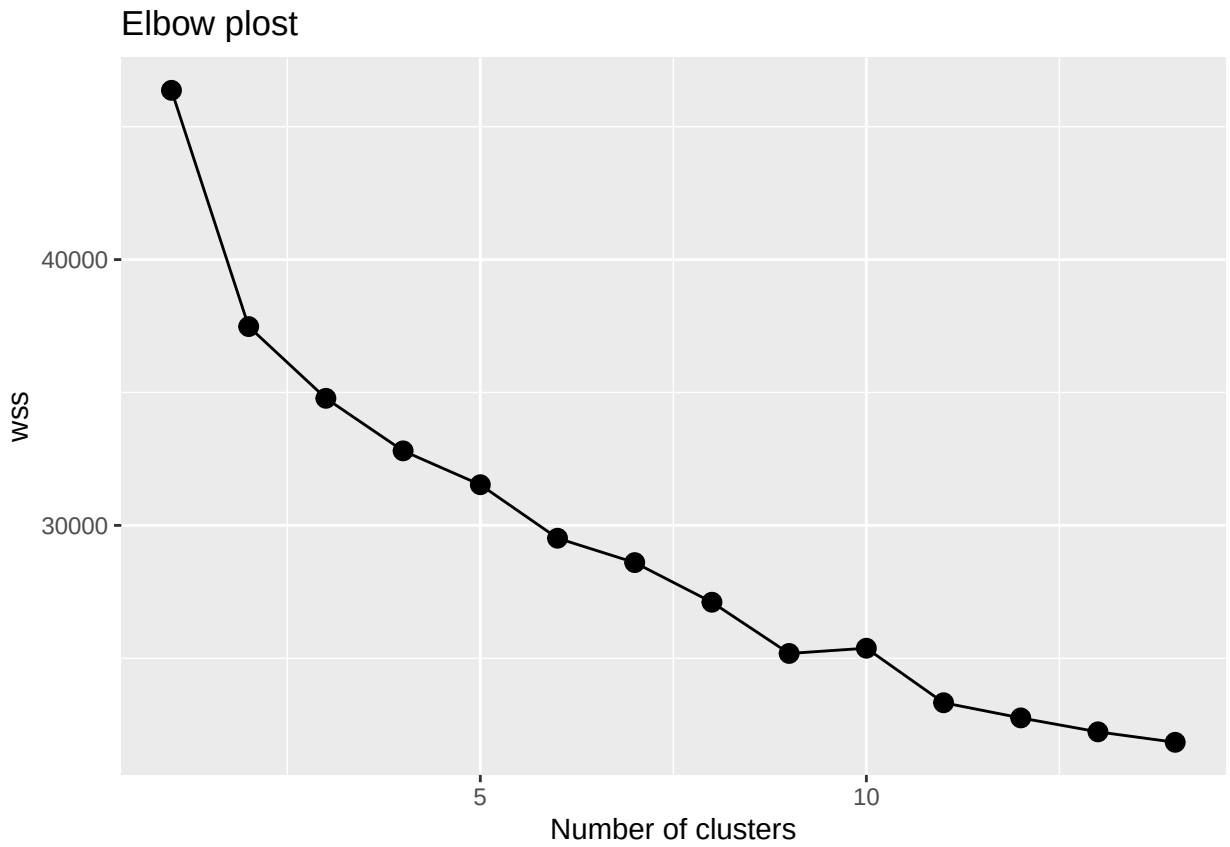
Cluster with K-Means

In questions Q2.4 to Q2.9 use K-Means method for clustering

Selecting Number of Clusters

Q2.4 Select optimal number of clusters using elbow method. *(4 points)*

```
set.seed(2)
ws <- sapply(1:14,function(k){kmeans(df_scale,k,nstart=10)$tot.withinss})
  ggplot(data.frame(k=1:14,WSS=ws), aes(x=k, y=WSS)) + geom_point(size=3) + geom_line() +
  labs(title="Elbow plost", x="Number of clusters", y="wss")
```

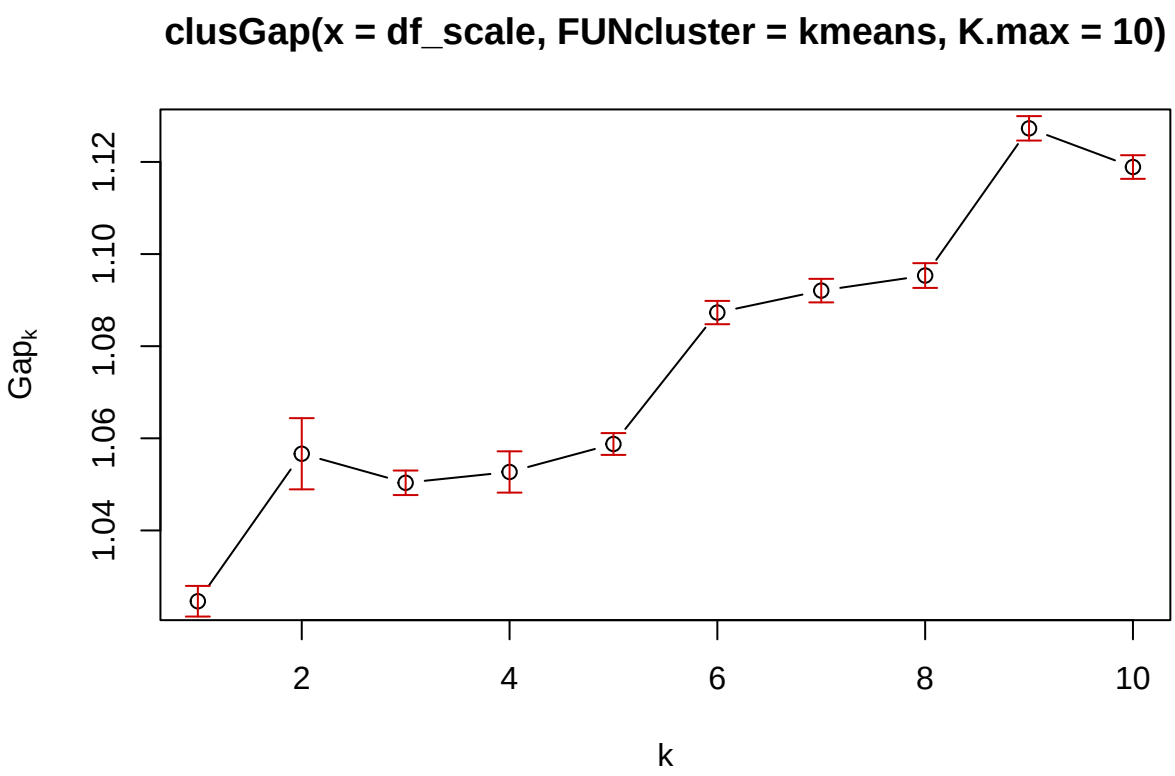


Here we select 2,10 as optimal clusters for elbow method, as we can see the elbow at those points. **Q2.5** Select optimal number of clusters using Gap Statistic.*(4 points)*

```
set.seed((200))
gap_stat <- clusGap(df_scale, FUNcluster = kmeans, K.max = 10,)
```

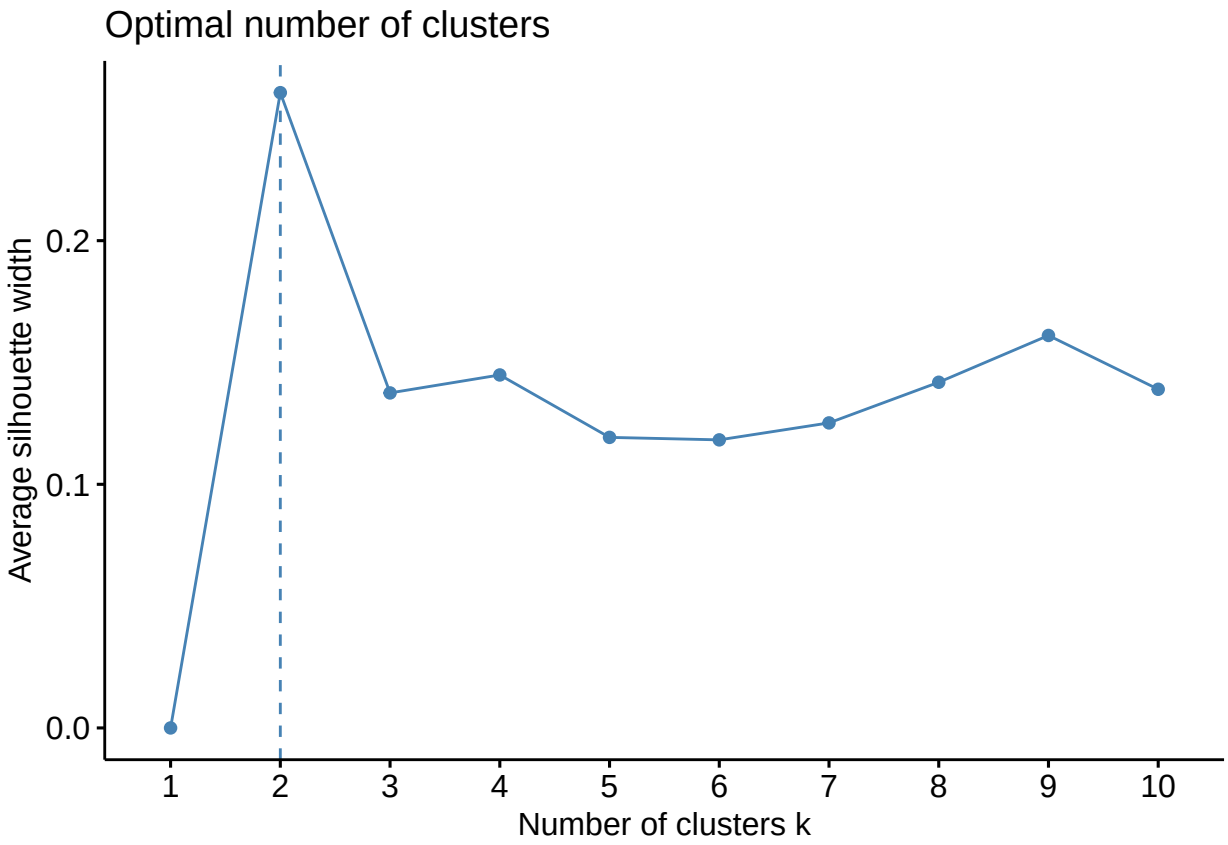
```
## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations

## Warning: did not converge in 10 iterations
```

```
plot(gap_stat)
```



#I feel we have optimal clusters at 2 and 9 for Gap Statistic. **Q2.6** Select optimal number of clusters using Silhouette method.*(4 points)*

```
fviz_nbclust(df_scale,kmeans,method=c("silhouette"),print.summary = TRUE,barfill = "blue",barcolor = "bl
```



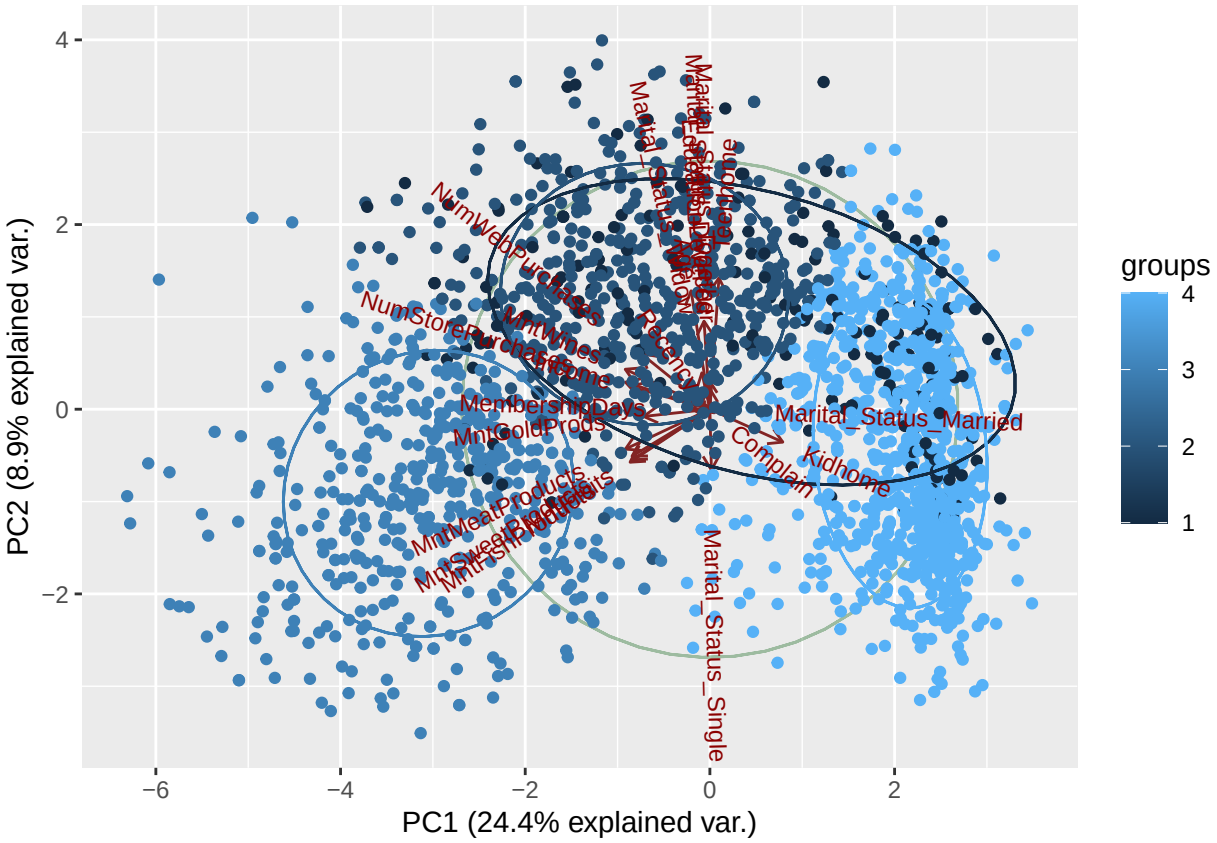
#I feel that 2 and 9 are optimal clusters for silhuettes. **Q2.7** Which k will you choose based on elbow, gap statistics and Silhouette as well as clustering task (market segmentation for advertisement purposes, that is two groups don't provide sufficient benefit over a single groups)? *(4 points)*

I will choose 2 & 5 cluster for elbow,gap and Silhouette. So Initally I have chosen eblow method and later refined my choice of clusters using and gap & Silhouettes. As 2 doesn't choose market segmentation I will got for the next value which will be 5.

Clusters Visulalization

Q2.8 Make k-Means clusters with selected k_kmeans (store result as km_out). Plot your k_kmeans clusters on biplot (just PC1 vs PC2) by coloring points by their cluster id. *(4 points)*

```
km_out <- kmeans(df_scale, 4)
custom_colors <- c("red", "blue", "green", "black")
ggbiplot(pc_out, groups = km_out$cluster, scale = 0,ellipse = TRUE, circle = TRUE)
```



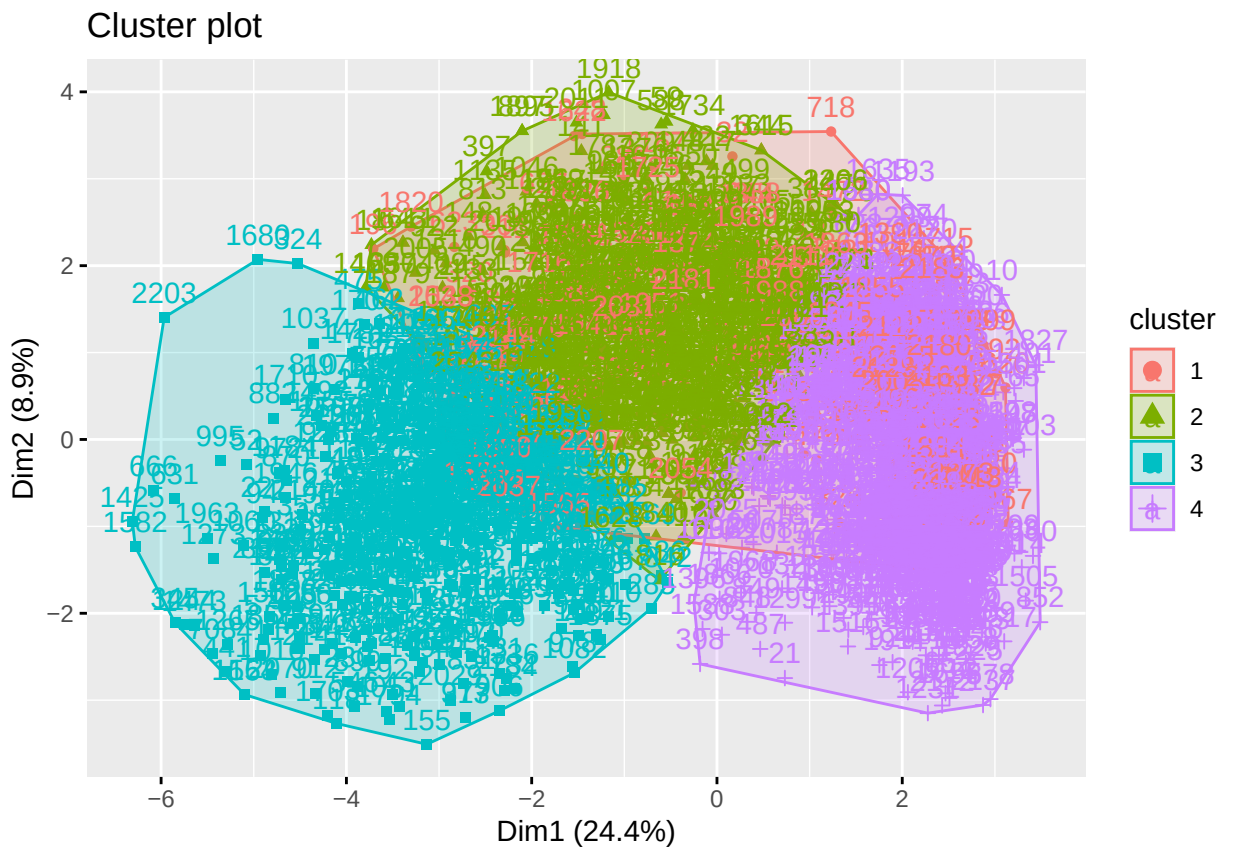
Q2.9 Do you see any grouping? Comment on you observation.*(2 points)*

Answer... I would see we have 4 grouping considering the above plot

Characterizing Cluster

Q2.10 Perform descriptive statistics analysis on obtained cluster. Based on that does one or more group have a distinct characteristics? *(8 points)* Hint: add cluster column to original df dataframe

```
df_market$cluster <- km_out$cluster
cl_summary <- fviz_cluster(km_out, data=df_scale, ellipse = TRUE, ellipse.type = "convex", xlab = NULL,
  ylab = NULL,outlier.color = "black",ggtheme = theme_grey())
cl_summary
```



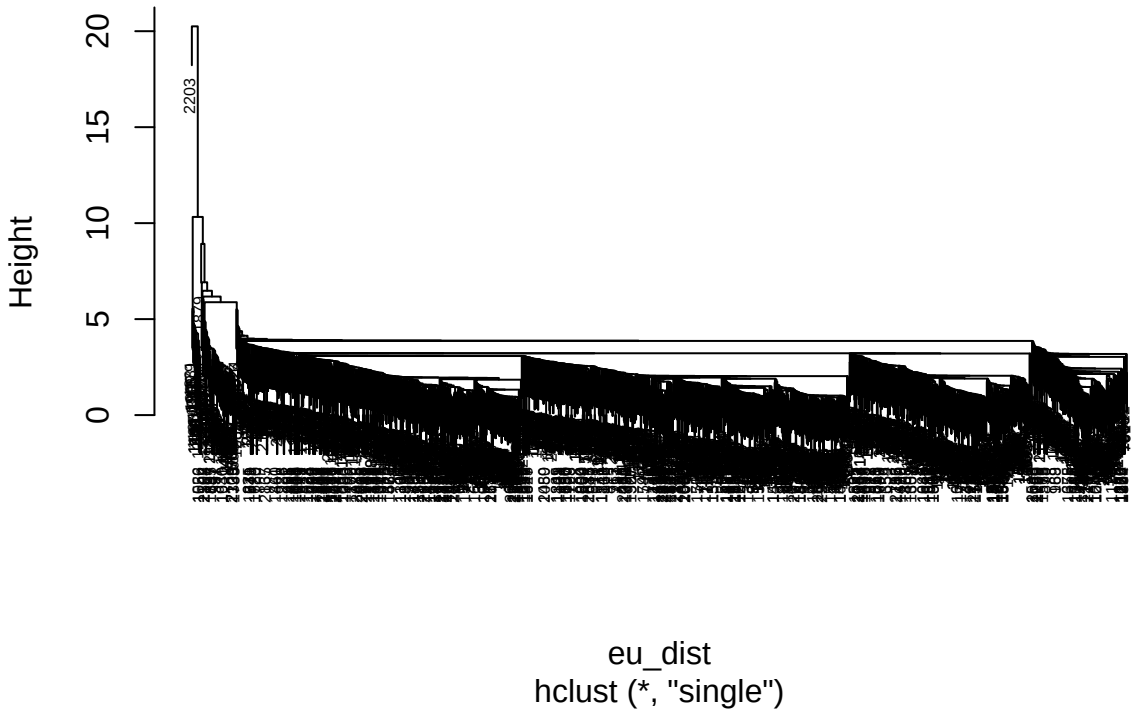

```
eu_dist <- dist(df_scale,method='euclidean')
```

Cluster with Hierarchical Clustering

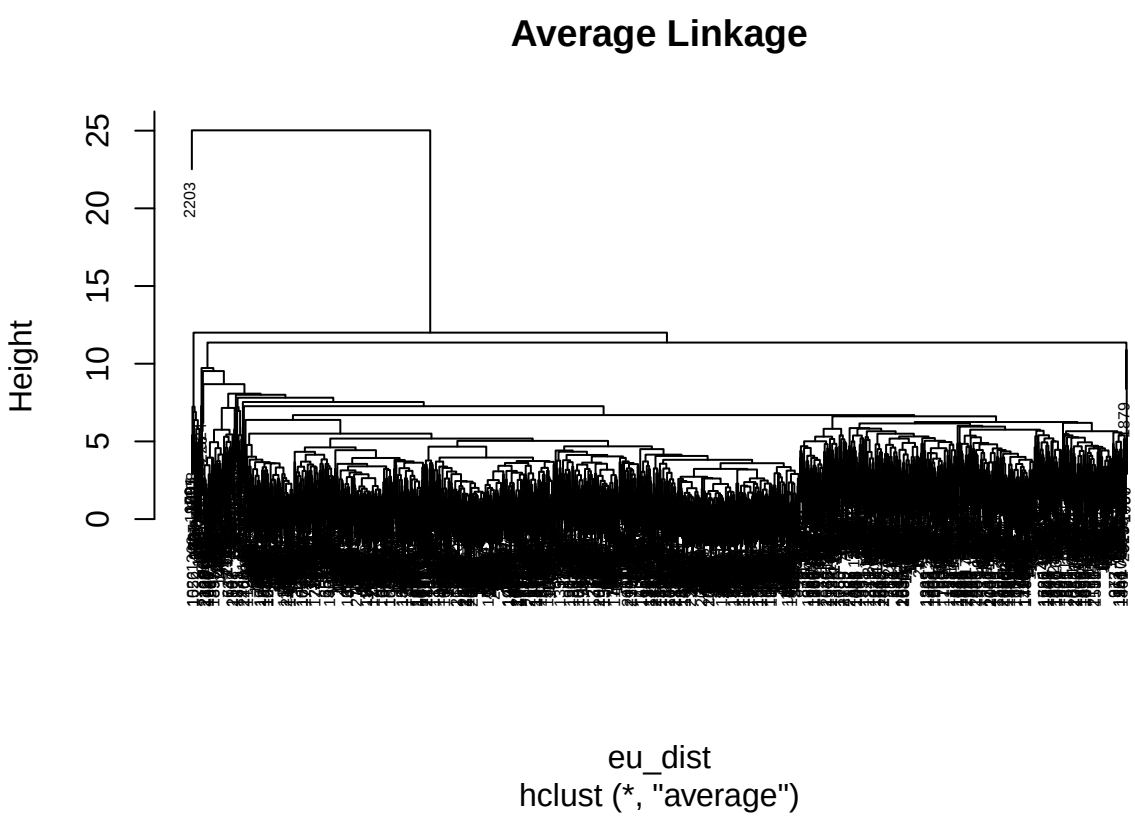
Q2.11 Perform clustering with Hierarchical method (Do you need to use scaling here?). Try complete, single and average linkage. Plot dendagram, based on it choose linkage and number of clusters, if possible, explain your choice. *(8 points)*

```
single<-hclust(eu_dist,method='single')
average<-hclust(eu_dist,method='average')
complete<-hclust(eu_dist,method='complete')
```

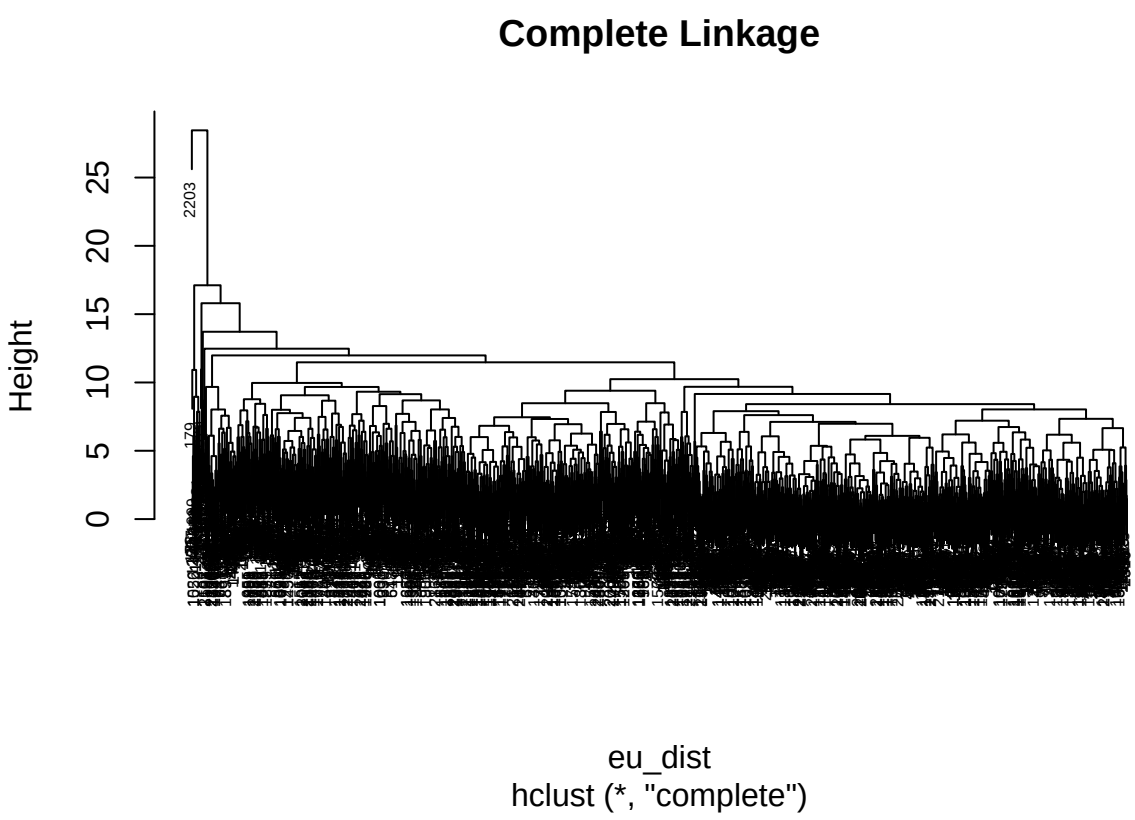
```
plot(single,main="Single Linkage", cex =  .5)
```




```
plot(average,main="Average Linkage",cex=.5)
```




```
plot(complete,main="Complete Linkage",cex=.5)
```




```
table(cutree(single, 11))
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11
## 2102    1   76   20    3    1    1    2    1    1    1
```

```
table(cutree(average,11))
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11
## 2086    4   76   21    1   15    1    1    2    1    1
```

```
table(cutree(complete,11))
```

```
##
##    1    2    3    4    5    6    7    8    9   10   11
##  541 1054  507    4   76   20    1    2    2    1    1
```

Additional grading criteria:

i feel that complete linkage look good and then average and then single. **G3.1** Was all random methods properly seeded? *(2 points)* yes all random methods are random seeded.