STAE03: Business Analytics

Assignment 5

Diana (990820T222)

1. Introduction

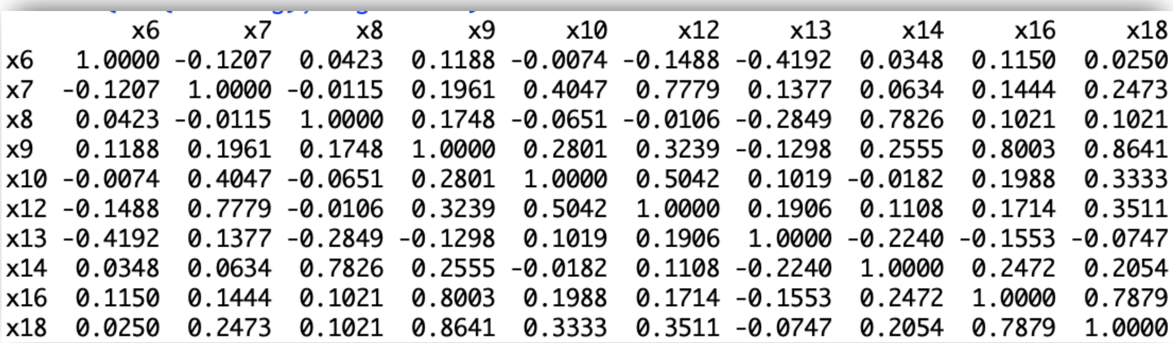
This assignment is divided into two parts. The first part attempted to conduct a Principal Component Analysis (PCA) on HBATred dataset. The dataset is a web-survey result that was conducted among 100 companies' customers. It contains information about 10 variables, each on a scale from 0-10, with 10 being "excellent" and 0 being "poor". One extracted 80 observations randomly and performed a Principal Component Analysis.

The second part utilized Store dataset to perform a cluster analysis. The dataset is a survey of purchasing behaviour of 275 women which comprises of 18 questions about their attitudes regarding choice of clothing store, each on a scale from 1 to 5 with 1 being "Not at all important" and 5 being "Very important". One extracted 200 observations randomly and attempted to create clusters to define typical customers in a clothing store for women.

>> set.seed(2008)

2. Part 1

Before moving on to perform a PCA on HBATred dataset, one analysed the correlation between variables first.



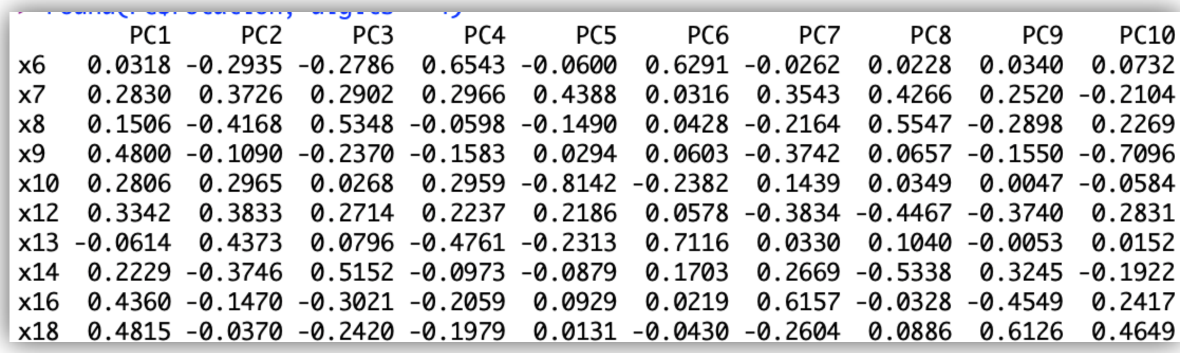
```
         x6      x7      x8      x9     x10     x12     x13     x14     x16     x18
x6   1.0000 -0.1207  0.0423  0.1188 -0.0074 -0.1488 -0.4192  0.0348  0.1150  0.0250
x7  -0.1207  1.0000 -0.0115  0.1961  0.4047  0.7779  0.1377  0.0634  0.1444  0.2473
x8   0.0423 -0.0115  1.0000  0.1748 -0.0651 -0.0106 -0.2849  0.7826  0.1021  0.1021
x9   0.1188  0.1961  0.1748  1.0000  0.2801  0.3239 -0.1298  0.2555  0.8003  0.8641
x10 -0.0074  0.4047 -0.0651  0.2801  1.0000  0.5042  0.1019 -0.0182  0.1988  0.3333
x12 -0.1488  0.7779 -0.0106  0.3239  0.5042  1.0000  0.1906  0.1108  0.1714  0.3511
x13 -0.4192  0.1377 -0.2849 -0.1298  0.1019  0.1906  1.0000 -0.2240 -0.1553 -0.0747
x14  0.0348  0.0634  0.7826  0.2555 -0.0182  0.1108 -0.2240  1.0000  0.2472  0.2054
x16  0.1150  0.1444  0.1021  0.8003  0.1988  0.1714 -0.1553  0.2472  1.0000  0.7879
x18  0.0250  0.2473  0.1021  0.8641  0.3333  0.3511 -0.0747  0.2054  0.7879  1.0000
```

Figure 2.1: Correlation between Variables of HBATred

From the correlation matrix above, one observed that there are some variables that are highly correlated and one divided them into four different groups:

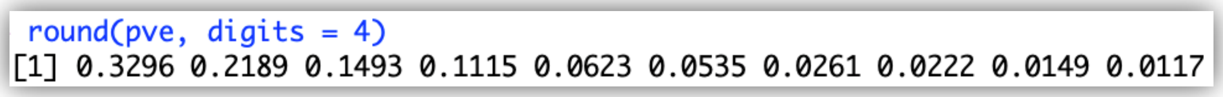
- Group 1: x9, x16, and x18. They are customer service activities.
- Group 2: x7, x10, and x12. They are marketing activities.
- Group 3: x8 and x14. They are about the quality of technical support and warranty.
- Group 4: x6 and x13. They are about the quality of the product.

This division of variables into four groups suggests that one could reduce the dimensionality of the dataset by utilizing PCA. Thus, after performing PCA on the dataset, one obtained the following result:



```
> round(pc$rotation, digits = 4)
        PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
x6   0.0318 -0.2935 -0.2786  0.6543 -0.0600  0.6291 -0.0262  0.0228  0.0340  0.0732
x7   0.2830  0.3726  0.2902  0.2966  0.4388  0.0316  0.3543  0.4266  0.2520 -0.2104
x8   0.1506 -0.4168  0.5348 -0.0598 -0.1490  0.0428 -0.2164  0.5547 -0.2898  0.2269
x9   0.4800 -0.1090 -0.2370 -0.1583  0.0294  0.0603 -0.3742  0.0657 -0.1550 -0.7096
x10  0.2806  0.2965  0.0268  0.2959 -0.8142 -0.2382  0.1439  0.0349  0.0047 -0.0584
x12  0.3342  0.3833  0.2714  0.2237  0.2186  0.0578 -0.3834 -0.4467 -0.3740  0.2831
x13 -0.0614  0.4373  0.0796 -0.4761 -0.2313  0.7116  0.0330  0.1040 -0.0053  0.0152
x14  0.2229 -0.3746  0.5152 -0.0973 -0.0879  0.1703  0.2669 -0.5338  0.3245 -0.1922
x16  0.4360 -0.1470 -0.3021 -0.2059  0.0929  0.0219  0.6157 -0.0328 -0.4549  0.2417
x18  0.4815 -0.0370 -0.2420 -0.1979  0.0131 -0.0430 -0.2604  0.0886  0.6126  0.4649
```

Figure 2.2: Principal Component (PC) loadings



```
round(pve, digits = 4)
[1] 0.3296 0.2189 0.1493 0.1115 0.0623 0.0535 0.0261 0.0222 0.0149 0.0117
```

Figure 2.3: Proportion of variance explained

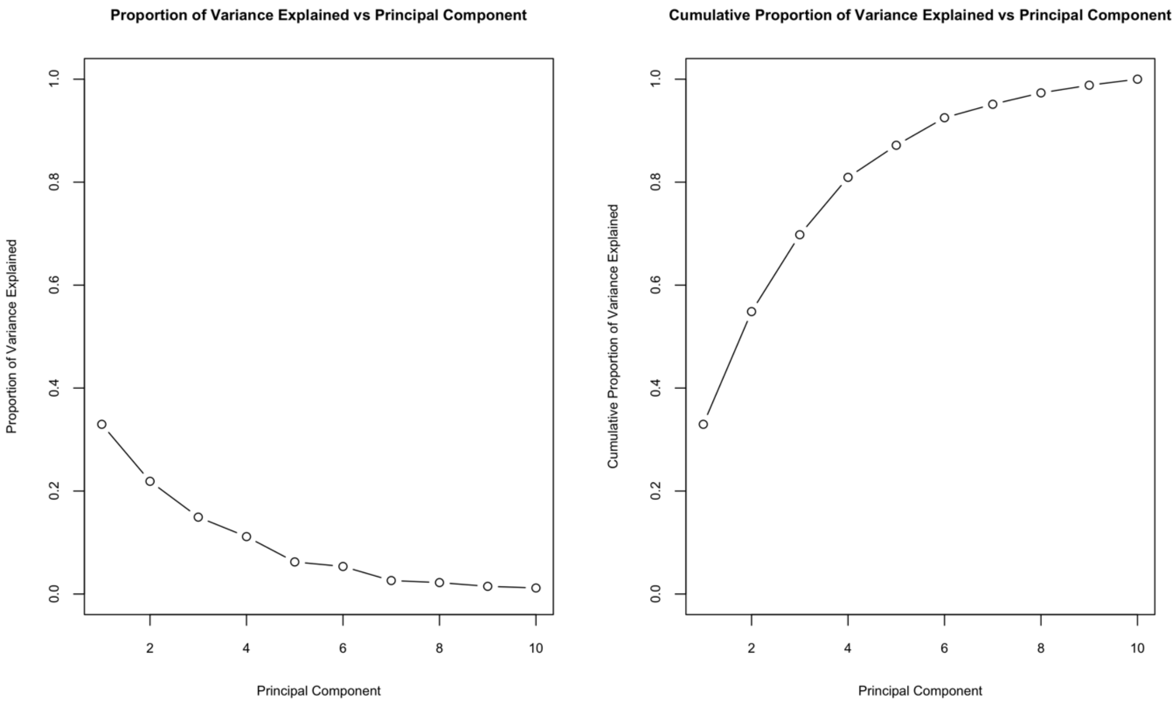


Figure 2.4 (left to right): Proportion of variance explained, Cumulative Proportion of Variance explained

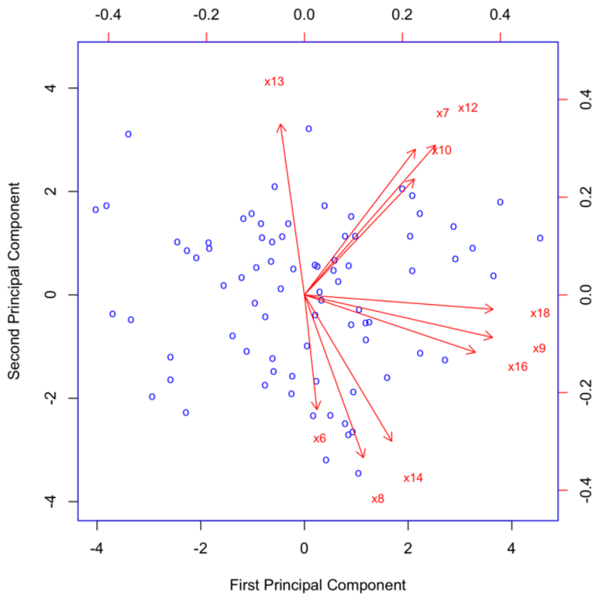


Figure 2.5: Biplot of HBATred dataset

Figure 2.3 and 2.4 suggest that one should use four principal components. One arrived at this suggestion after calculating the cumulative proportion of variance explained. One believes that once the cumulative proportion of variance explained reaches somewhere above 0.75, one should stop and choose the number of PC that is at the stop point. In this case, at PC = 4, one has reached cumulative proportion of variance explained of 0.8093. Thus, this supports the suggestion that was made earlier in this paragraph.

By observing figure 2.2 and 2.5, one interpreted the meaning behind four different principal components:

PC	Variables that load more in this PC	Interpretation
1	x9, x16, x18	Customer service activities.
2	x7, x8, x10, x12, x13, x14	Marketing activities and support and warranty.
3	x8, x14	Support and warranty
4	x6, x13	The quality of the product.

In the beginning of this part, one attempted to divide the variables into four different groups. The groups were similar in terms of interpretation to the groups in the table above except for PC2. Group 2 comprises of x7, x10, and x12 while PC2 adds in x8, x13, and x14 to this group of variables. The interpretation of this group also changes slightly due to these additional variables. This group initially describes variables that are related to marketing activities. The insertion of x8 and x14 to this group gives this group a broader interpretation. Now, it describes the combination of marketing activities and support and warranty. Thus, after performing PCA on this dataset, one concluded that PCA enables the possibility of dimensionality reduction from 10 variables to 4 variables (that are linear combination of those 10 initial variables) while maintaining high variability.

3. Part 2

Since this part is aimed at characterizing customers in a clothing store for women, one attempted to define 3 groups from 10 chosen variables in the Store dataset. Variables that are kept for further analyses are: x1, x2, x3, x4, x5, x7, x9, x11, x12, x15. From these 10 variables, one formed 3 groups as follows:

- Group 1: Economical customers. Variables that are associated with this group: x11, x12, x15.
- Group 2: High-end customers (those who value high quality product with high price) Variables that are associated with this group: x3, x7, x9.
- Group 3: Customers that concern about the quality of the stores. Variables that are associated with this group: x1, x2, x4, x5.

To validate these assumptions regarding types of customers, one performed a hierarchical cluster analysis with Euclidean measure as the dissimilarity measure and four different methods: Single Linkage, Complete Linkage, Average Linkage and Ward Linkage. These methods give clearer view of how many clusters that actually exist in this dataset.

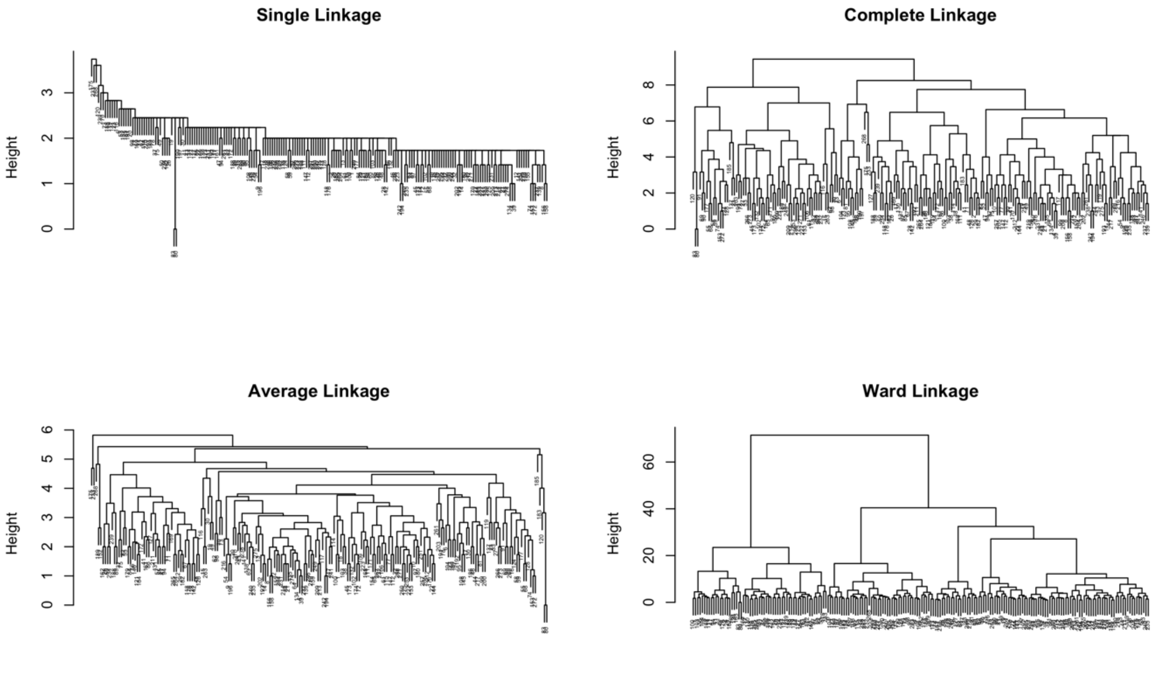


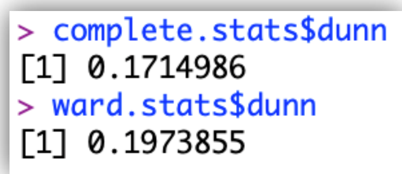
Figure 3.1 (from top-left to bottom-right): hierarchical cluster analysis with single linkage, complete linkage, average linkage, and ward linkage

The results of Single Linkage and Average Linkage methods are pretty difficult to observe. It seems like there is only one cluster defined with these methods. However, one believes that there should be more than one cluster in this dataset. Thus, one omitted the results of these methods.

The results of Complete Linkage and Ward Linkage methods are more obvious. Complete Linkage method suggests that there are two clusters found in this dataset and Ward Linkage method proposes an addition of one cluster, which results in total of three clusters. In order to choose which method provides the most optimal cluster for the Store dataset, one utilized Dunn index as a clustering validation measure. Dunn index is computed as follows:

$$D = \frac{min.separation}{max.diameter}$$

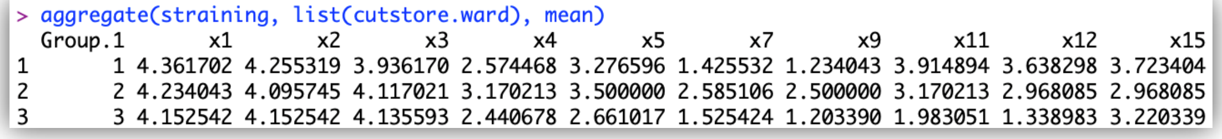
, where D = Dunn Index, min.separation = minimum of pairwise distance (inter-cluster separation), and max.diameter = maximum intra-cluster distance.



```
> complete.stats$dunn
[1] 0.1714986
> ward.stats$dunn
[1] 0.1973855
```

Figure 3.5: Dunn index for complete and ward methods

Dunn index suggests that one should choose the method which maximizing the minimum inter-cluster separation whilst minimizing the maximum intra-cluster distance, or in other words the higher Dunn index. One obtained Dunn index for Complete method = 0.1715 and Dunn index for Ward method = 0.1974. Since Ward method yields higher Dunn index, one opted this method as the final model to cluster the store dataset. One thus proceed with the interpretation of the clusters.

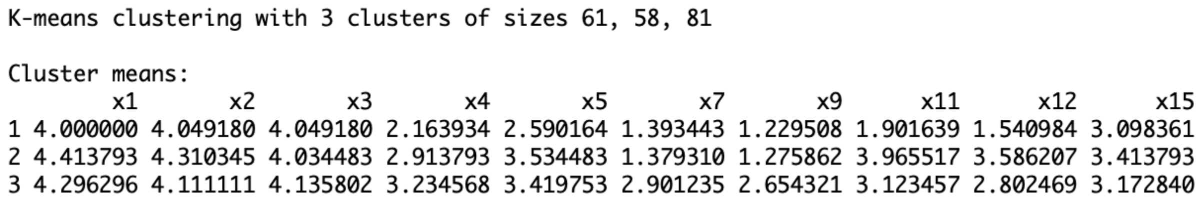


```
> aggregate(straining, list(cutstore.ward), mean)
  Group.1       x1       x2       x3       x4       x5       x7       x9      x11      x12      x15
1       1 4.361702 4.255319 3.936170 2.574468 3.276596 1.425532 1.234043 3.914894 3.638298 3.723404
2       2 4.234043 4.095745 4.117021 3.170213 3.500000 2.585106 2.500000 3.170213 2.968085 2.968085
3       3 4.152542 4.152542 4.135593 2.440678 2.661017 1.525424 1.203390 1.983051 1.338983 3.220339
```

Figure 3.3: Cluster profiling with ward linkage method

Variables x1, x2, and x3 have relatively high mean in all clusters. These are variables related to the quality of staff and product in the store. One inferred that these variables are very important regardless of the type of customers. Furthermore, group 1 could be defined as economical shoppers, who are aware of the price of the products and thus always seek for the best price. Group 2 has relatively similar means across all variables (exclude: x1, x2, x3). This suggests that group 2 is those high-end shoppers who concern about the price as well as the quality of the products, staff, and store. In fact, it is natural for shoppers who regularly purchase high-end products to concern about almost everything. Group 3 covers shoppers who are interested in the visual aspect of the store (such as interior design) and stores which offer reasonable prices. After performing cluster analysis with Ward Linkage method, one observed that there is a major change in interpretation of the groups. However, this new interpretation provides a more representative and realistic clustering for typical customers.

It follows from Ward Linkage method that there are 3 clusters. Thus, to further the analysis on the Store dataset, one performed a K-means clustering with 3 clusters and obtained the following result:

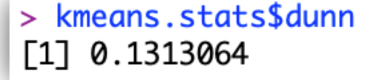


```
K-means clustering with 3 clusters of sizes 61, 58, 81

Cluster means:
        x1       x2       x3       x4       x5       x7       x9      x11      x12      x15
1 4.000000 4.049180 4.049180 2.163934 2.590164 1.393443 1.229508 1.901639 1.540984 3.098361
2 4.413793 4.310345 4.034483 2.913793 3.534483 1.379310 1.275862 3.965517 3.586207 3.413793
3 4.296296 4.111111 4.135802 3.234568 3.419753 2.901235 2.654321 3.123457 2.802469 3.172840
```

Figure 3.4: Cluster profiling with K-means clustering method

One observed that K-means clustering gives a roughly similar result as Ward Linkage method. Nevertheless, the groups are in different order. With K-means clustering, group 1 is group 3 in Ward Linkage method, group 2 is group 1 in Ward Linkage method, and group 3 is group 2 in Ward Linkage method.

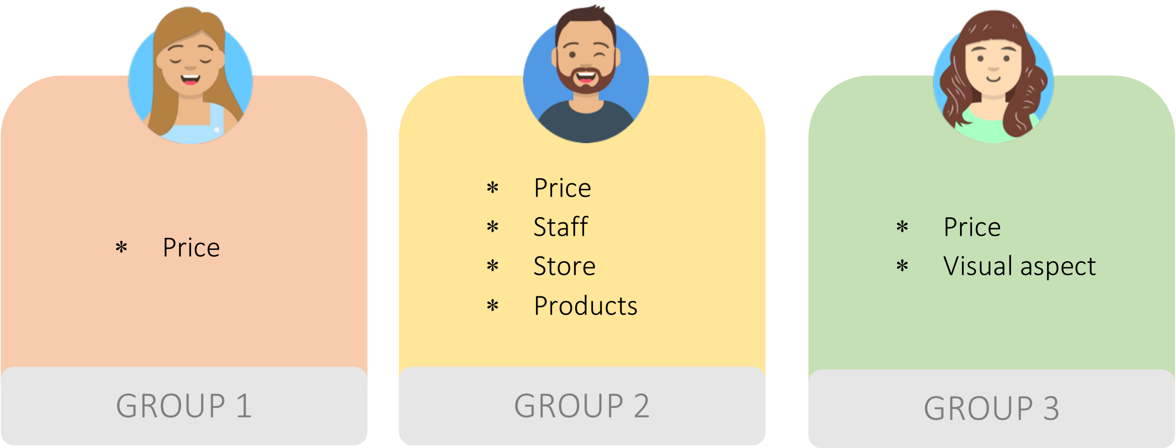


```
> kmeans.stats$dunn
[1] 0.1313064
```

Figure 3.5: Dunn index for K-means clustering

Similar as before, one utilized Dunn index as a clustering validation measure to compare K-means clustering with Ward Linkage method and Single Linkage method. Dunn index for K-means clustering is the lowest among other methods in hierarchical clustering. This indicates that K-means clustering probably has either lower minimum inter-cluster separation or higher intra-cluster diameter or both.

There is no single definitive answer to which clustering method performs the best with this dataset. Dunn Index has certainly given a clearer view, however, one must note that there are still other tools or validation techniques that could possibly lead to a better conclusion but no consensus has been made on which approach is the best. Despite that fact, after considering the profiling result and Dunn index, one concluded that Ward Linkage method gives the most representative clusters to describe typical customers in a clothing store for women.



- Price

GROUP 1

- Price
- Staff
- Store
- Products

GROUP 2

- Price
- Visual aspect

GROUP 3

Figure 3.6: Final clustering of customers

4. Appendix

```
## PART 1

setwd("/Users/angeladianas/Desktop/Lund/r_files")

library(dplyr)

HBATred <- read.csv("HBATred.csv", header = TRUE, sep = ",", dec = ".")

head(HBATred)

dim(HBATred)

HBATred <- select(HBATred, -X)


set.seed(2008)

n = length(HBATred$x6)

train <- sample(1:n, 80)

training <- HBATred[train,]

test <- HBATred[-train,]


round(cor(training), digits = 4)


PC <- prcomp(training, scale = TRUE)

biplot(PC, scale = 0, cex = 0.8, xlabs = rep("o", nrow(training)), col = c("blue", "red"), xlab = "First Principal Component", ylab = "Second Principal Component")


round(PC$rotation, digits = 4)

PCvar <- PC$sdev ^ 2

pve <- PCvar/sum(PCvar)

round(pve, digits = 4)

par(mfrow = c(1,2))

plot(pve, xlab="Principal Component", ylab="Proportion of Variance Explained", main = "Proportion of Variance Explained vs Principal Component", cex.main = 0.85, cex.axis = 0.75, cex.lab = 0.75, ylim=c(0,1), type = "b")
plot(cumsum(pve), xlab="Principal Component", ylab="Cumulative Proportion of Variance Explained",
main = "Cumulative Proportion of Variance Explained vs Principal Component", cex.main = 0.85,
cex.axis = 0.75, cex.lab = 0.75, ylim=c(0,1), type = "b")


## PART 2

store <- read.csv("Store.csv", sep = ";", header = TRUE)

head(store)

store <- select(store, -Individual)

round(cor(store), digits = 2)


sel_store <- select(store, -c(x14, x13, x17, x18, x16, x6, x10, x8))

round(cor(sel_store), digits = 4)


set.seed(2008)


ns = nrow(sel_store)

strain <- sample(1:ns, 200)

straining <- sel_store[strain,]

stest <- sel_store[-strain,]

dim(straining)


store.single <- hclust(dist(straining), method = "single")

store.complete <- hclust(dist(straining), method = "complete")

store.average <- hclust(dist(straining), method = "average")

store.ward <- hclust(dist(straining), method = "ward.D")


par(mfrow = c(2,2))

plot(store.single, main ="Single Linkage", xlab ="", sub ="", cex = 0.4)

plot(store.complete, main ="Complete Linkage", xlab ="", sub ="", cex = 0.4)          # tree = 2?

plot(store.average, main ="Average Linkage", xlab ="", sub ="", cex = 0.4)
```

```r
plot(store.ward, main ="Ward Linkage", xlab ="", sub ="", cex = 0.4)                    # tree = 3?
## current interpretation to determine the number of trees: look at heights, if the fuse at approximately similar height, they're one tree. Otherwise, they're different.


cutstore.complete <- cutree(store.complete, 2)
cutstore.ward <- cutree(store.ward, 3)
aggregate(straining, list(cutstore.complete), mean)
aggregate(straining, list(cutstore.ward), mean)


set.seed(2008)
store.kmeans <- kmeans(straining, 3, nstart = 20)
store.kmeans


## validation statistics
library(fpc)
library(NbClust)
kmeans.stats <- cluster.stats(dist(straining), store.kmeans$cluster)
complete.stats <- cluster.stats(dist(straining), cutstore.complete)
ward.stats <- cluster.stats(dist(straining), cutstore.ward)
kmeans.stats$dunn
complete.stats$dunn
ward.stats$dunn
```

5. Image References

https://www.google.com/url?sa=i&url=https%3A%2F%2Fya-webdesign.com%2Fexplore%2Ffunny-png-avatar%2F&psig=AOvVaw0gQzcX5Oe9b81WG8AudS5o&ust=1590204723232000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCLDn4bbExukCFQAAAAAdAAAAABAY

https://www.google.com/url?sa=i&url=http%3A%2F%2Fhappyfacesparty.com%2Ftestimonial%2Fbrittany-nicole%2Favataaars-brittany%2F&psig=AOvVaw0gQzcX5Oe9b81WG8AudS5o&ust=1590204723232000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCLDn4bbExukCFQAAAAAdAAAAABAd

https://www.google.com/url?sa=i&url=http%3A%2F%2Fhappyfacesparty.com%2Ftestimonial%2F1672%2Favataaars-frances%2F&psig=AOvVaw0gQzcX5Oe9b81WG8AudS5o&ust=1590204723232000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCLDn4bbExukCFQAAAAAdAAAAABAi