**1.**

We identify three different clusters of households using K-Means Clustering.

*library(readxl)*

*BathSoap\_Data <- read\_excel("~/OneDrive/UIC Documents/Studies/Data Mining - IDS 572/Assignments/Homework 5/BathSoap\_Data.xls",*

*sheet = "DM\_Sheet")*

*View(BathSoap\_Data)*

*normalize <- function(x) {*

*num <- x - min(x)*

*denom <- max(x) - min(x)*

*return (num/denom)*

*}*

1. **Variables that describe purchase behavior:**

Since the scales of the different variables used for clustering are different, we normalize the variables:

*BathSoap\_Purc\_Behav\_Data <- as.data.frame(lapply(BathSoap\_Data[c(12, 13, 14, 15, 16, 19)], normalize))*

*BathSoap\_Purc\_Behav\_Data[, "Share.To.Other.Brands"] <- normalize(BathSoap\_Data$`Total Volume`\*BathSoap\_Data$`Others 999`)*

*BathSoap\_Purc\_Behav\_Data[, "Max.To.One.Brand"] <- normalize(apply(BathSoap\_Data$`Total Volume`\*BathSoap\_Data[, 23:30], 1, max))*

*plot(BathSoap\_Purc\_Behav\_Data, pch =20, cex =2)*

*set.seed(7)*

To find the perfect K Value, we do the elbow method to obtain the graph of the sum of squares for different number of clusters.

*mydata = BathSoap\_Purc\_Behav\_Data*

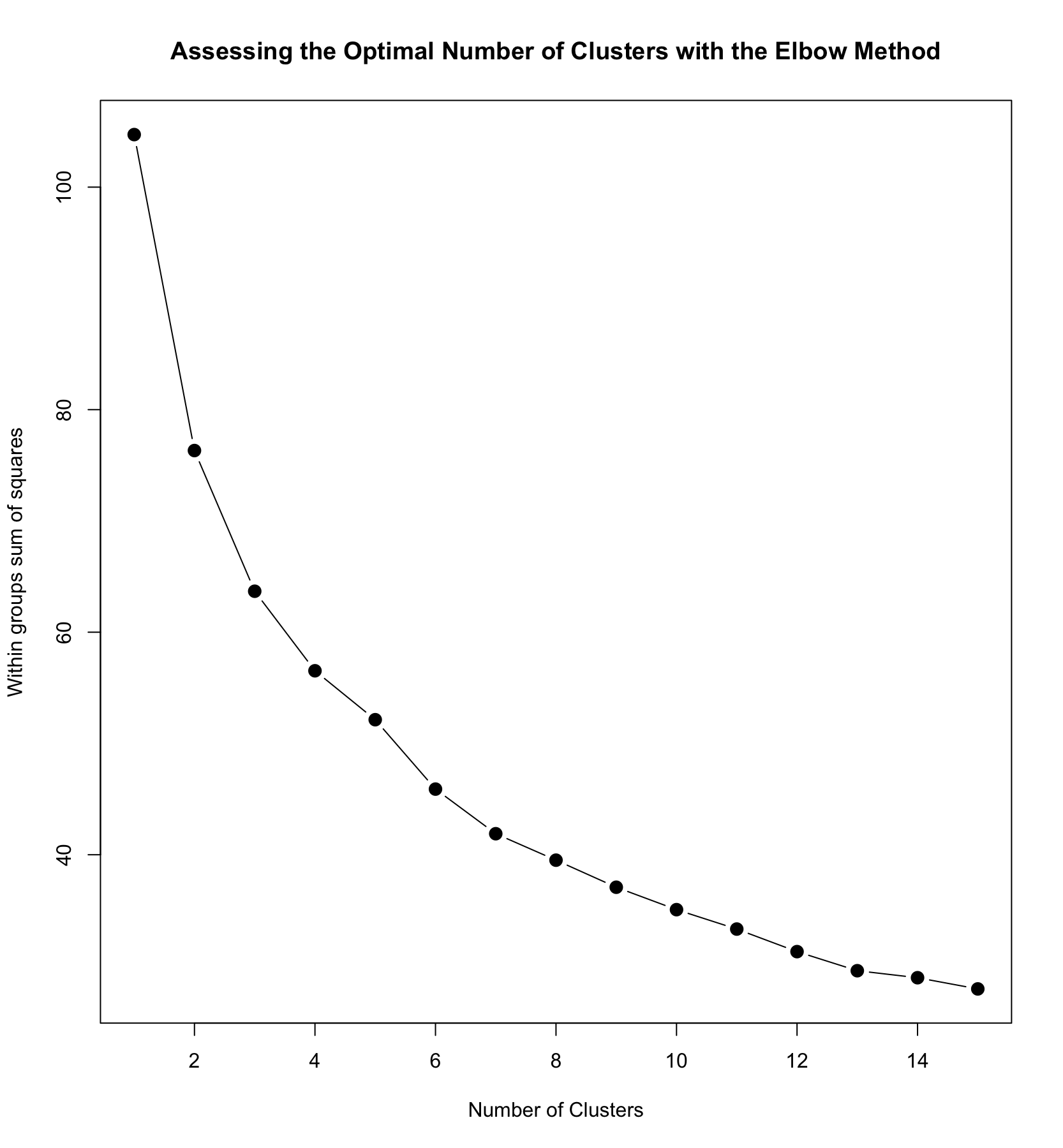
*wss = (nrow(mydata)-1)\*sum(apply(mydata,2,var))*

*wss*

*for (i in 2:15) wss[i] = sum(kmeans(mydata, centers=i)$withinss)*

*plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares",*

*main="Assessing the Optimal Number of Clusters with the Elbow Method", pch=20, cex=2)*



*km1 = kmeans(BathSoap\_Purc\_Behav\_Data, 6, nstart=100)*

*km1*

*plot(BathSoap\_Purc\_Behav\_Data, col =(km1$cluster +1) , main="K-Means result with 6 clusters", pch=20, cex=2)*

*km1*

The km1 variable gives us the summary of the clusters made on basis of purchase behavior.

1. **The variables that describe basis-for-purchase:**

After looking at the selling proposition for the different products, we have eliminated the Prop Cat’s 10, 11, 12, 13 as not significant. So we try to cluster with the remaining variables for basis-for-purchase.

*BathSoap\_Purc\_Basis\_purch <- BathSoap\_Data[c(20, 21, 22, 32:40, 45)]\*BathSoap\_Data$`Total Volume`*

*BathSoap\_Purc\_Basis\_purch <- as.data.frame(lapply(BathSoap\_Purc\_Basis\_purch, normalize))*

*plot(BathSoap\_Purc\_Basis\_purch, col =(km1$cluster +1) , main="K-Means result with 6 clusters", pch=20, cex=2)*

*mydata = BathSoap\_Purc\_Basis\_purch*

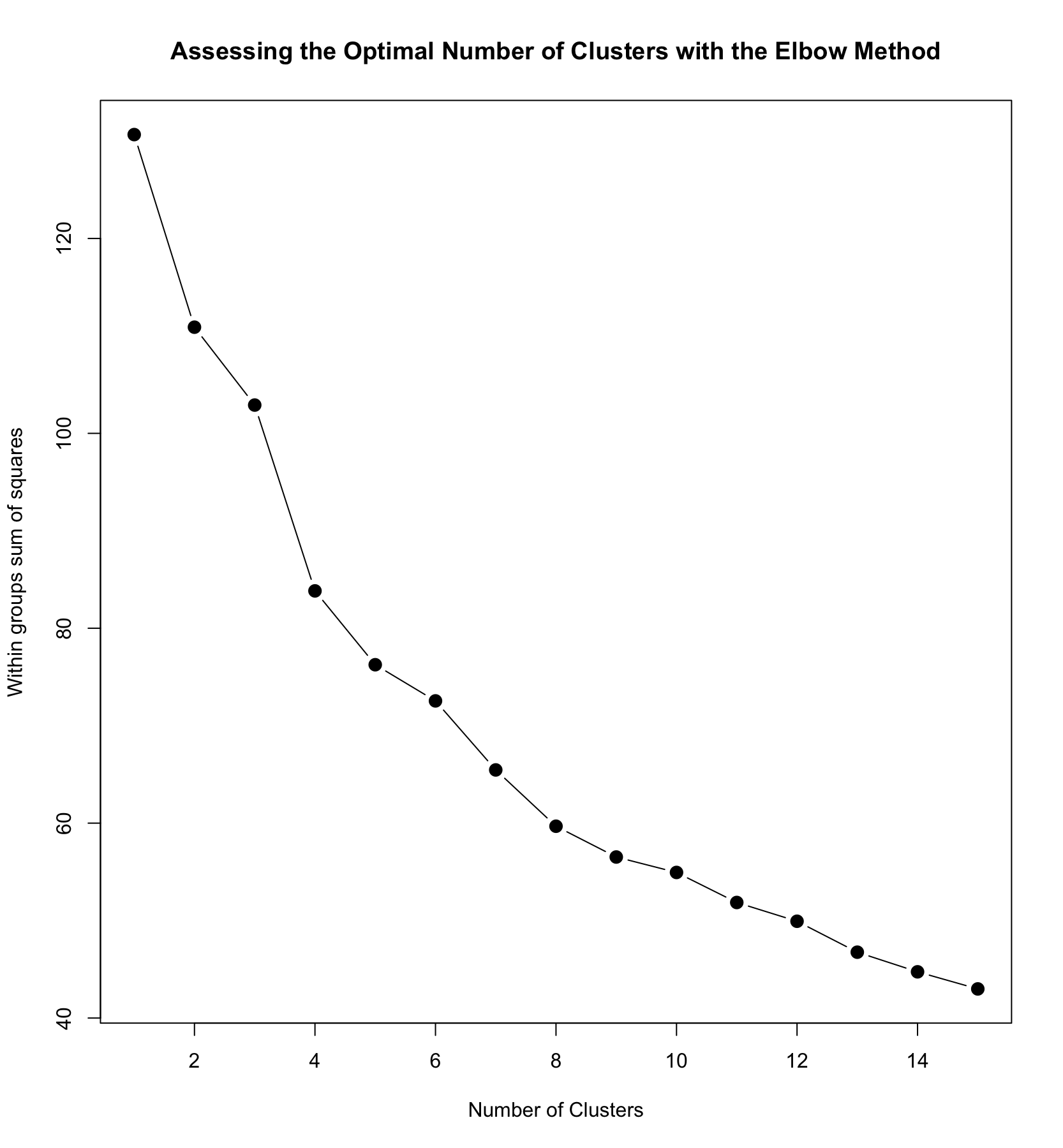
*wss = (nrow(mydata)-1)\*sum(apply(mydata,2,var))*

*wss*

*for (i in 2:15) wss[i] = sum(kmeans(mydata, centers=i)$withinss)*

*plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares",*

*main="Assessing the Optimal Number of Clusters with the Elbow Method", pch=20, cex=2)*



*km2 = kmeans(BathSoap\_Purc\_Basis\_purch, 6, nstart=100)*

*km2*

1. **The variables that describe both purchase behavior and basis of purchase:**

*BathSoap\_Combined <- cbind(BathSoap\_Purc\_Behav\_Data,BathSoap\_Purc\_Basis\_purch)*

*mydata = BathSoap\_Combined*

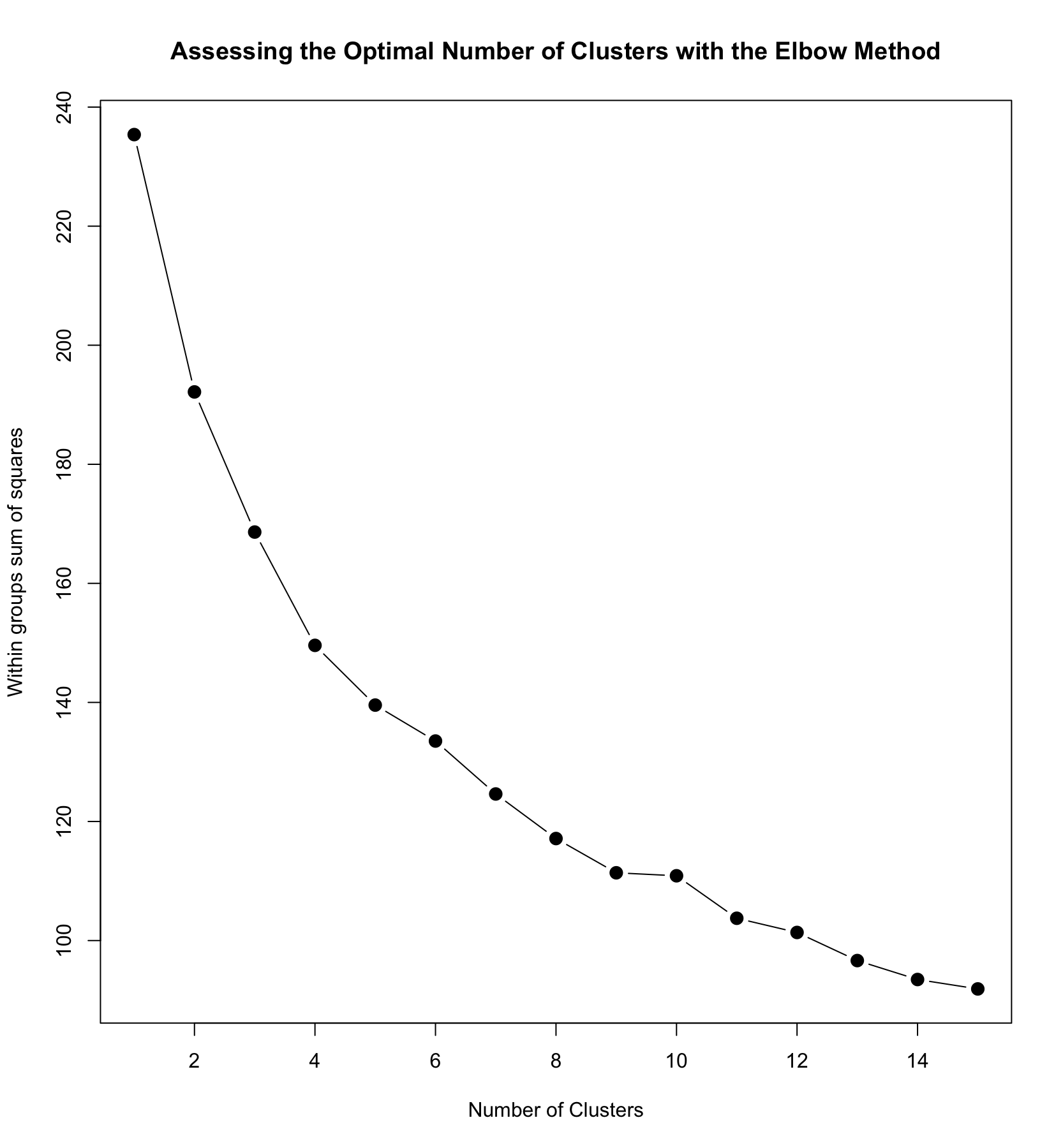
*wss = (nrow(mydata)-1)\*sum(apply(mydata,2,var))*

*wss*

*for (i in 2:15) wss[i] = sum(kmeans(mydata, centers=i)$withinss)*

*plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares",*

*main="Assessing the Optimal Number of Clusters with the Elbow Method", pch=20, cex=2)*



*km3 = kmeans(BathSoap\_Combined, 6, nstart=100)*

*km3*

**2.**

**A.**

We use the ratio of between SS and total SS to measure cluster effectiveness.

*km1$betweenss/km1$totss*

* 69.90%

*km2$betweenss/km2$totss*

* 63.37%

*km3$betweenss/km3$totss*

* 45.96%

Looking at the ratio of wss/tss of these three clusters we find that the WSS value of clusters in A (Variables that describe purchase behaviors) is the highest.

Hence clustering based on the purchase behaviors gives us the best segmentation.

**B. Characteristics of these clusters:**

We get the characteristics of these clusters by obtaining the instances and analyzing them individually.

*BathSoap\_Data\_With\_Clusters<-cbind(BathSoap\_Data,km1$cluster)*

*BathSoap\_Data\_With\_Clusters$`km1$cluster` <- as.factor(BathSoap\_Data\_With\_Clusters$`km1$cluster`)*

*plot(BathSoap\_Data\_With\_Clusters$`km1$cluster`, as.factor(BathSoap\_Data$SEX))*

*plot(BathSoap\_Data\_With\_Clusters$`km1$cluster`, as.factor(BathSoap\_Data$SEC))*

*plot(BathSoap\_Data\_With\_Clusters$`km1$cluster`, as.factor(BathSoap\_Data$FEH))*

*plot(BathSoap\_Data\_With\_Clusters$`km1$cluster`, as.factor(BathSoap\_Data$MT))*

*plot(BathSoap\_Data\_With\_Clusters$`km1$cluster`, as.factor(BathSoap\_Data$AGE))*

*plot(BathSoap\_Data\_With\_Clusters$`km1$cluster`, as.factor(BathSoap\_Data$EDU))*

*plot(BathSoap\_Data\_With\_Clusters$`km1$cluster`, as.factor(BathSoap\_Data$CS))*

Analyzing these clusters based on the different demographics available , we can deduce that

Cluster 1 consists of more observations in the lower SEC,

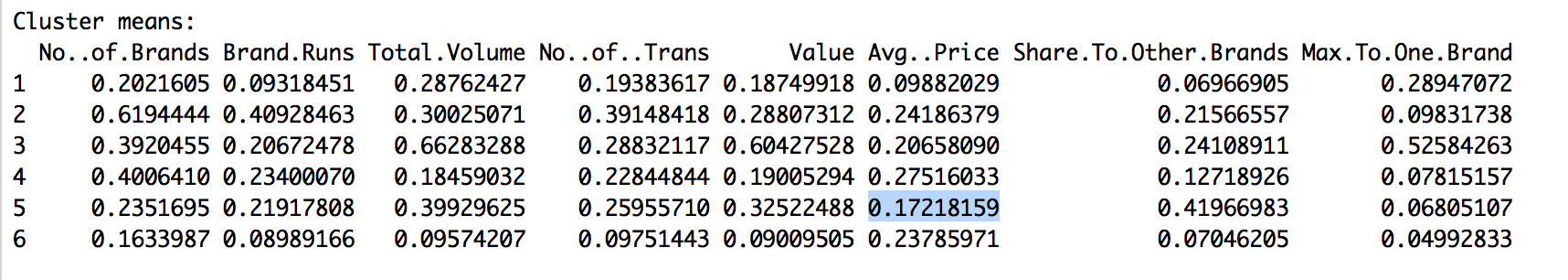
Cluster 3 constitutes almost entirely female and more than average Non Vegetarian’s,

Cluster 4 has a high number of people who have studied between 5-9 years in school

Cluster 5 has the highest percentage of non-vegetarians

Cluster 6 has the highest number of people who did not mention their educational qualifications,

Based on the brand loyalty:



We can see that,

Cluster 1 constitutes of people who have the lowest average price and very low share to other brands

Cluster2 has a higher number of people who buy from more no of brands and lower max to one brand (Indicating brand loyalty).

Cluster 3 has the highest max to one brand (Indicating brand loyalty), highest total volume and value indicating the highest buyers.

Cluster 4 has low brand loyalty and the second lowest total volume.

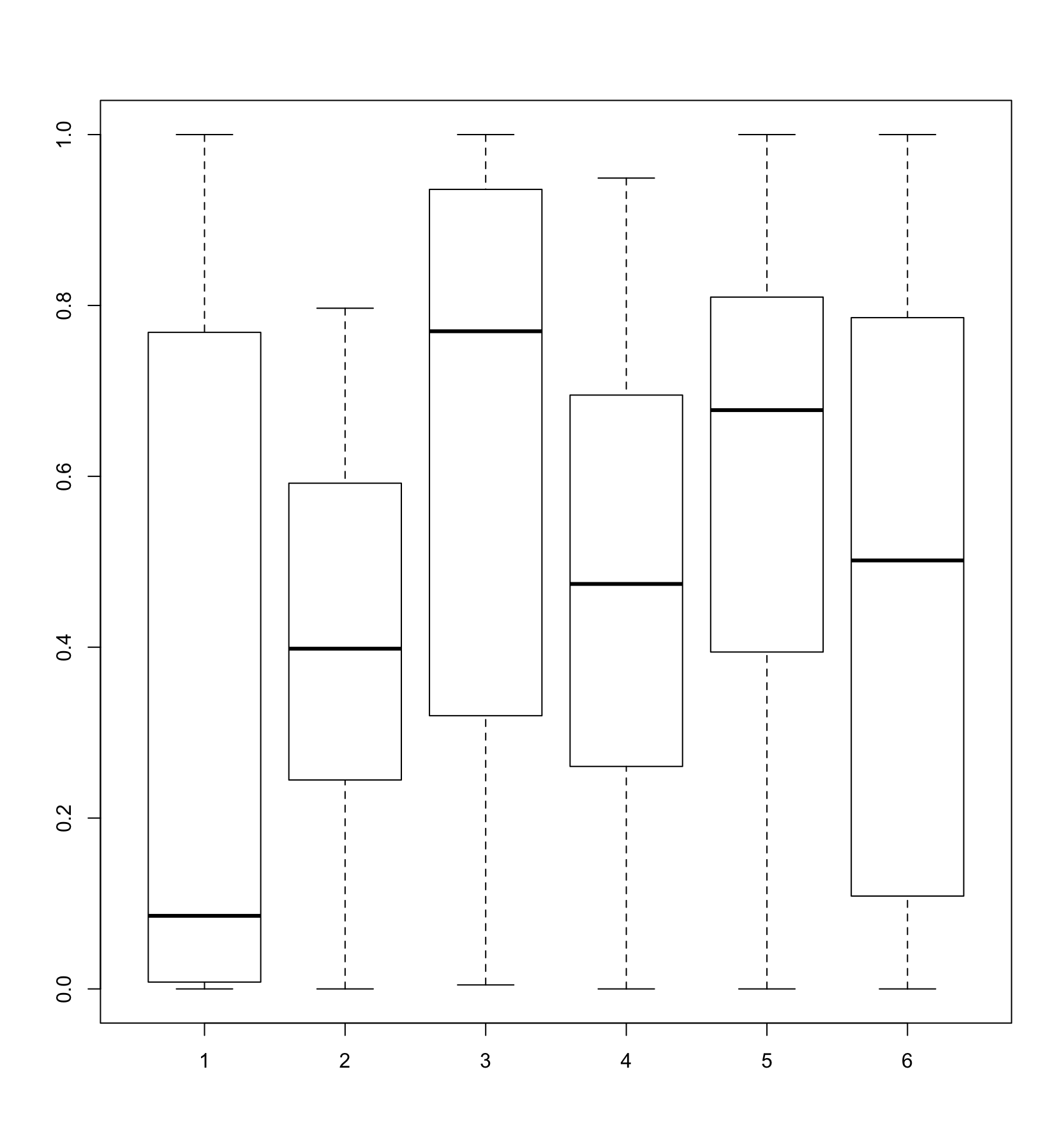
Cluster 6 constitutes the lowest value buyers, low max to one brand value and low total volume people.

Cluster 5 has the highest number of people who have share to other brands.

Characteristics based on the basis of purchase:

We can see that prop cat 5 has the most significance when compared to the other prop cat’s and the characteristics of this variable on different clusters can be given by a box plot.

plot(BathSoap\_Data\_With\_Clusters$`km1$cluster`, BathSoap\_Data$`PropCat 5`)



We can see that for basis of purchase, cluster 1 has the lowest median basis of purchase, cluster 3 has the highest. All the other clusters have average basis of purchase values.