ML bpalazzo_FinalExam

Brandon Palazzo

12/11/2020

Libraries and Loading Data

```
library(dplyr)

library(factoextra)

library(ggplot2)
library(GGally)

library(gridExtra)

total_data <- read.csv("BathSoap.csv")
total_data <- na.omit(total_data)
```

Separate the data into two sets: Purchase Behavior and Basis of Purchase

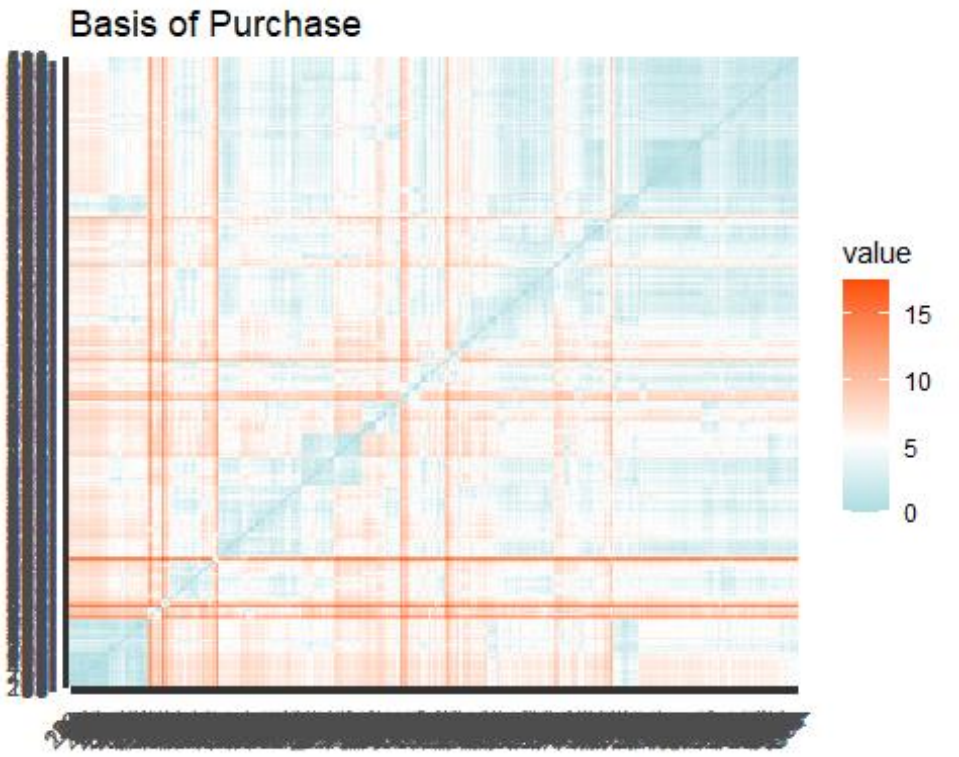
```
purchase_behavior_data <- total_data[, c(12:15, 17, 18, 20:31)]
purchase_behavior_data_scaled <- scale(purchase_behavior_data)
basis_of_purchase_data <- total_data[, c(16, 19, 32:46)]
basis_of_purchase_data_scaled <- scale(basis_of_purchase_data)
both_purchase_data <- total_data[, c(12:46)]
both_purchase_data_scaled <- scale(both_purchase_data)
```

Regardless of the brand, a customer who buys brand A is just as loyal as a customer who buys brand B. I believe that it is important to keep the brands as they are in order to get an idea of how loyal a customer is to a specific brand. However, one can derive this variable as a single variable as well.

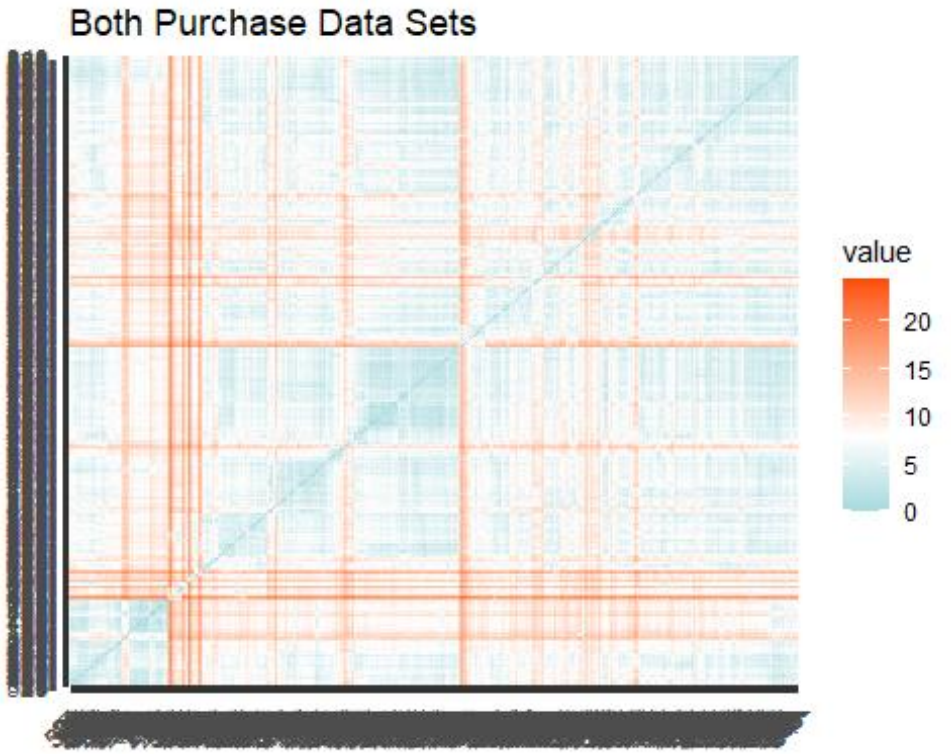
Distances and Distance Matrices

```
distance_pb <- get_dist(purchase_behavior_data_scaled)
fviz_dist(distance_pb, gradient = list(low = "#00AFBB", mid = "white",
                         high = "#FC4E07")) + ggtitle("Purchase Behavior")
```

```
distance_bop <- get_dist(basis_of_purchase_data_scaled)
fviz_dist(distance_bop, gradient = list(low = "#00AFBB", mid = "white",
                        high = "#FC4E07")) + ggtitle("Basis of Purchase")
```

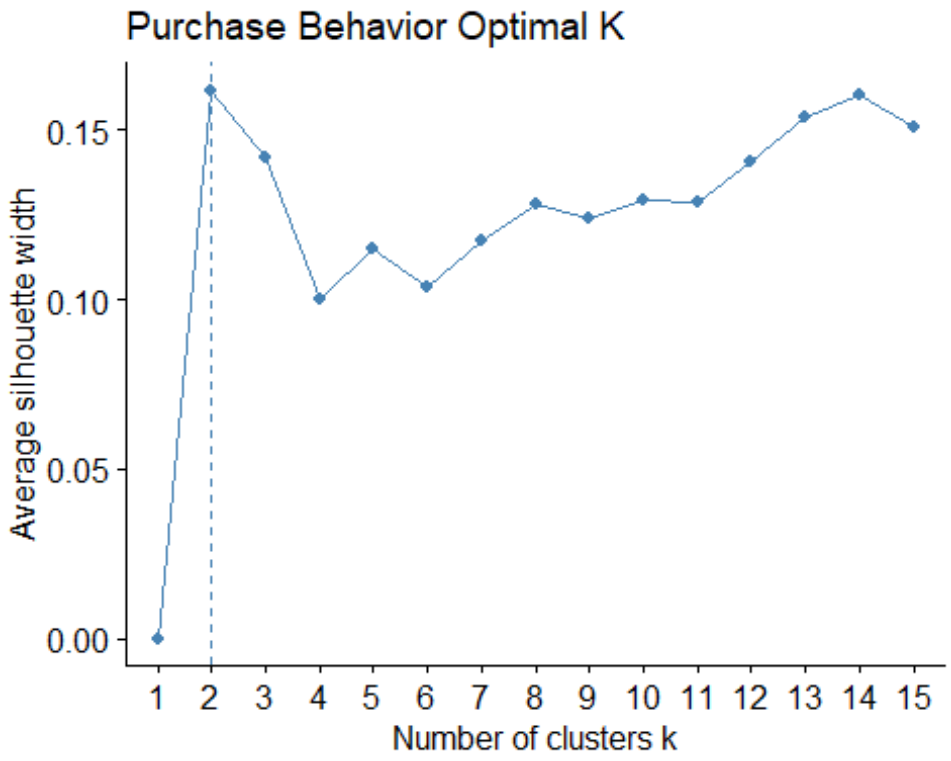


```
distance_bp <- get_dist(both_purchase_data_scaled)
fviz_dist(distance_bp, gradient = list(low = "#00AFBB", mid = "white",
                        high = "#FC4E07")) + ggtitle("Both Purchase Data
Sets")
```

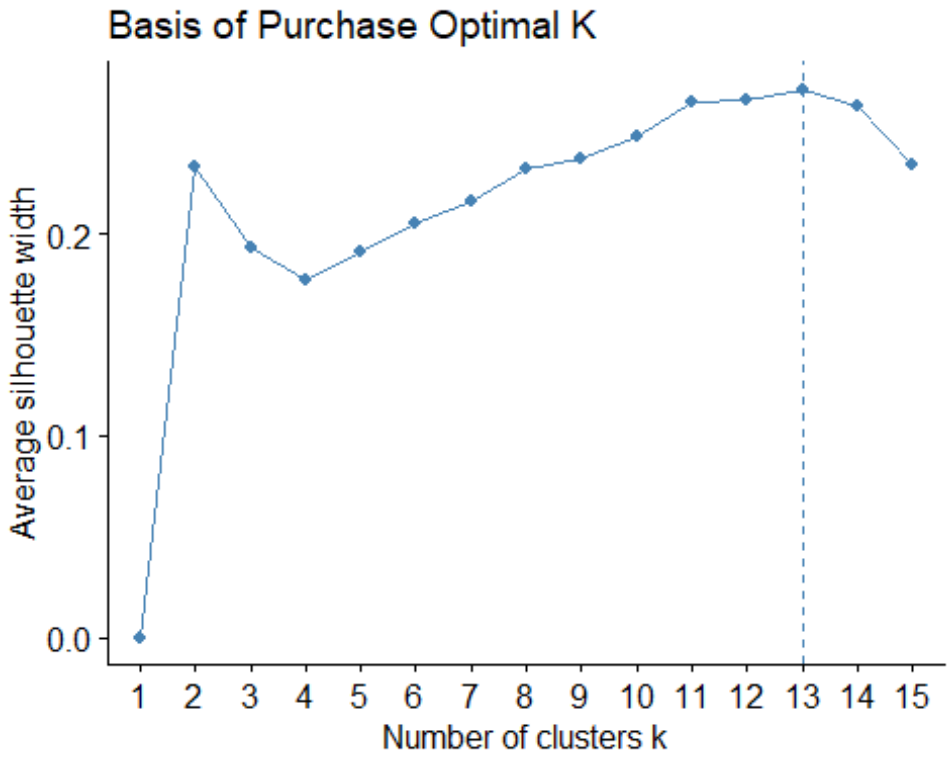


Determining the optimal K

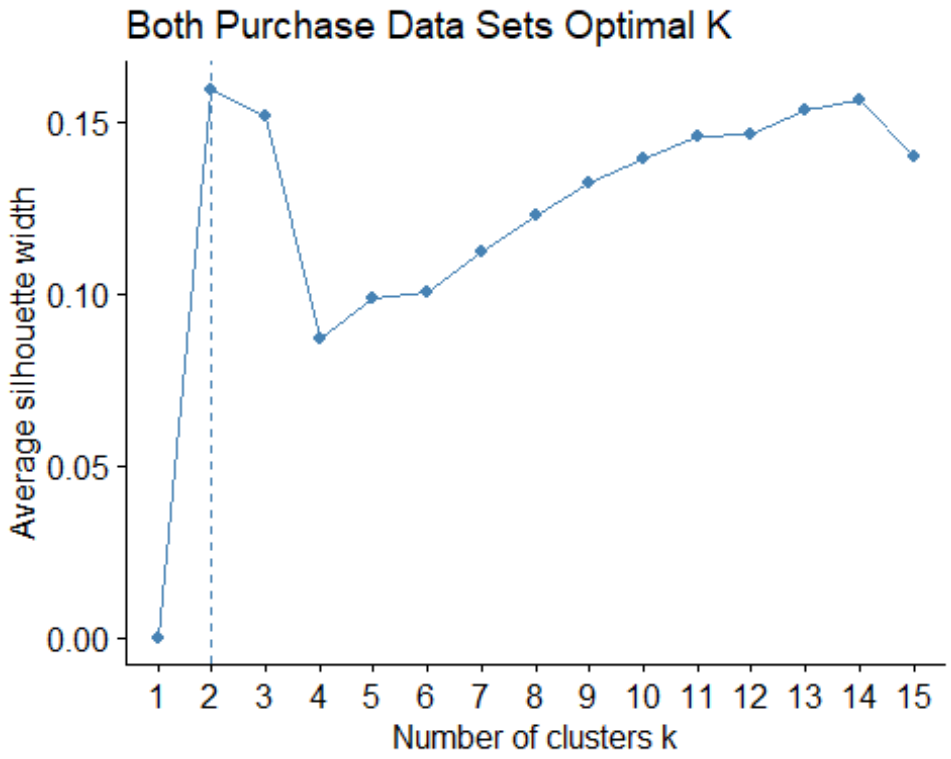
```
fviz_nbclust(purchase_behavior_data_scaled, FUN = hcut, method =
"silhouette",
             k.max = 15) + ggtitle("Purchase Behavior Optimal K")
```



```
fviz_nbclust(basis_of_purchase_data_scaled, FUN = hcut, method =
"silhouette",
             k.max = 15) + ggtitle("Basis of Purchase Optimal K")
```



```
fviz_nbclust(both_purchase_data_scaled, FUN = hcut, method = "silhouette",
             k.max = 15) + ggtitle("Both Purchase Data Sets Optimal K")
```

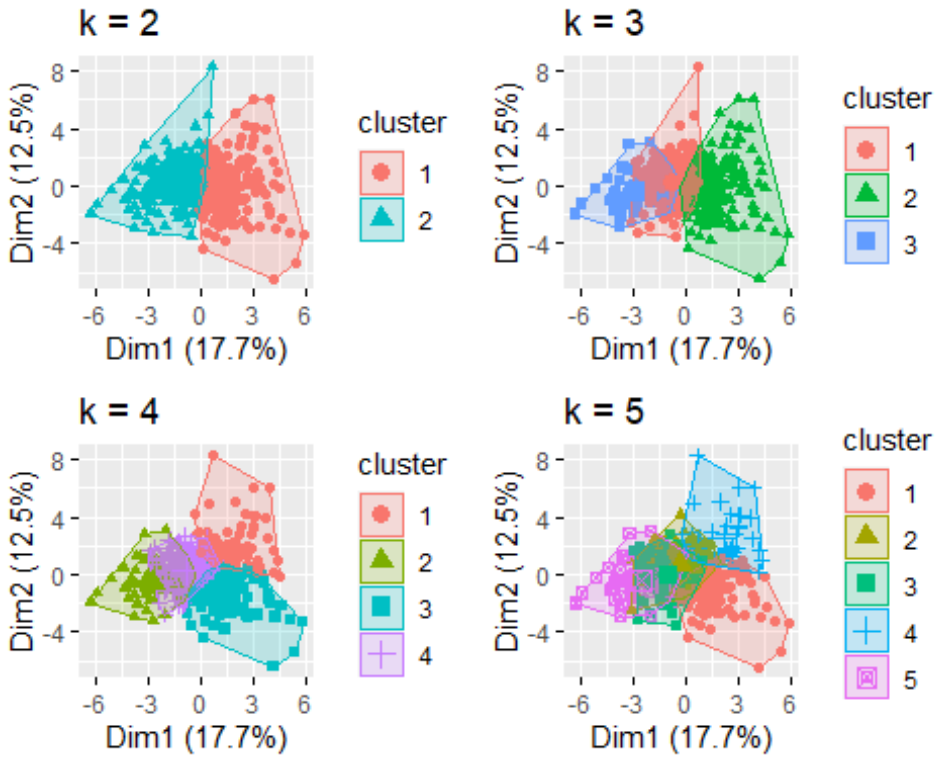


K Means Algorithms for Purchase Behavior Data set

```
k_pb2 <- kmeans(purchase_behavior_data_scaled, centers = 2, nstart = 25)
k_pb3 <- kmeans(purchase_behavior_data_scaled, centers = 3, nstart = 25)
k_pb4 <- kmeans(purchase_behavior_data_scaled, centers = 4, nstart = 25)
k_pb5 <- kmeans(purchase_behavior_data_scaled, centers = 5, nstart = 25)

# plots to compare
plot_pb2 <- fviz_cluster(k_pb2, geom = "point",
                         data = purchase_behavior_data_scaled) + ggtitle("k =
2")
plot_pb3 <- fviz_cluster(k_pb3, geom = "point",
                         data = purchase_behavior_data_scaled) + ggtitle("k =
3")
plot_pb4 <- fviz_cluster(k_pb4, geom = "point",
                         data = purchase_behavior_data_scaled) + ggtitle("k =
4")
plot_pb5 <- fviz_cluster(k_pb5, geom = "point",
                         data = purchase_behavior_data_scaled) + ggtitle("k =
5")
```

```
grid.arrange(plot_pb2, plot_pb3, plot_pb4, plot_pb5, nrow = 2)
```



After looking at both the silhouette and the four graphs above, the best k for purchase behavior is 2 because 2 clusters have the smallest overlap and matches the prediction of the silhouette. 2 clusters is very practical for a marketing team to focus on when considering purchase behavior.

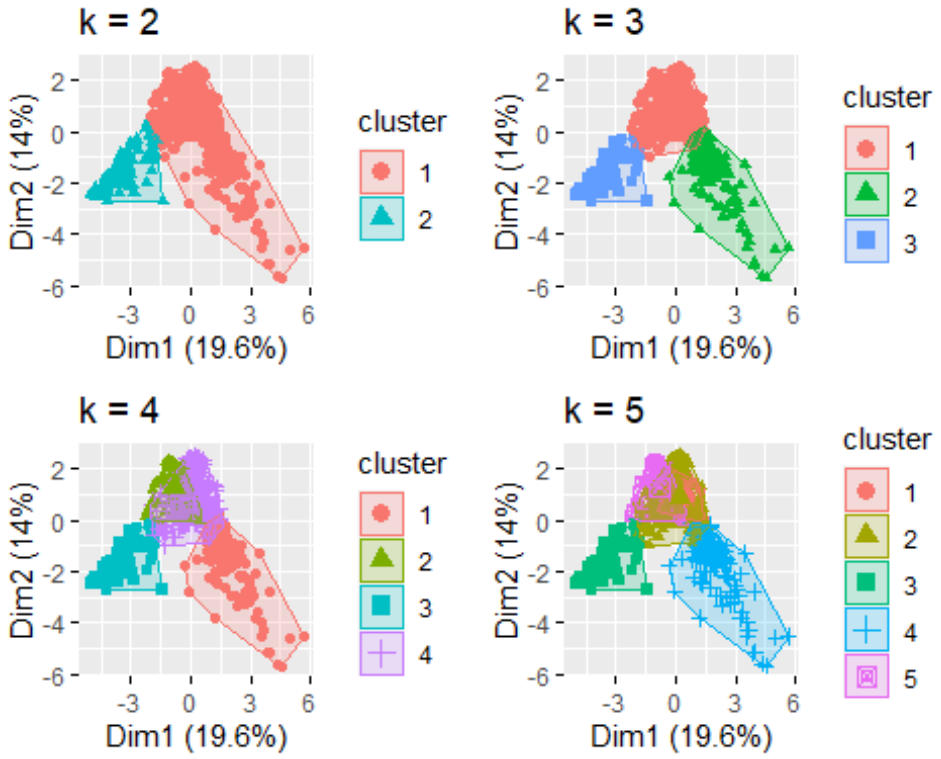
K Means Algorithms for Basis of Purchase Data set

```
k_bop2 <- kmeans(basis_of_purchase_data_scaled, centers = 2, nstart = 25)
k_bop3 <- kmeans(basis_of_purchase_data_scaled, centers = 3, nstart = 25)
k_bop4 <- kmeans(basis_of_purchase_data_scaled, centers = 4, nstart = 25)
k_bop5 <- kmeans(basis_of_purchase_data_scaled, centers = 5, nstart = 25)

# plots to compare
plot_bop2 <- fviz_cluster(k_bop2, geom = "point",
                          data = basis_of_purchase_data_scaled) + ggtitle("k
= 2")
plot_bop3 <- fviz_cluster(k_bop3, geom = "point",
                          data = basis_of_purchase_data_scaled) + ggtitle("k
= 3")
plot_bop4 <- fviz_cluster(k_bop4, geom = "point",
                          data = basis_of_purchase_data_scaled) + ggtitle("k
= 4")
```

```
plot_bop5 <- fviz_cluster(k_bop5, geom = "point",
                          data = basis_of_purchase_data_scaled) + ggtitle("k
= 5")


grid.arrange(plot_bop2, plot_bop3, plot_bop4, plot_bop5, nrow = 2)
```



After looking at both the silhouette and the four graphs above, the best k for basis of purchase is 3 because 3 clusters shows little overlap and fits the data better in the bottom right corner. 3 clusters is very practical for a marketing team to focus on when considering the basis of purchase.

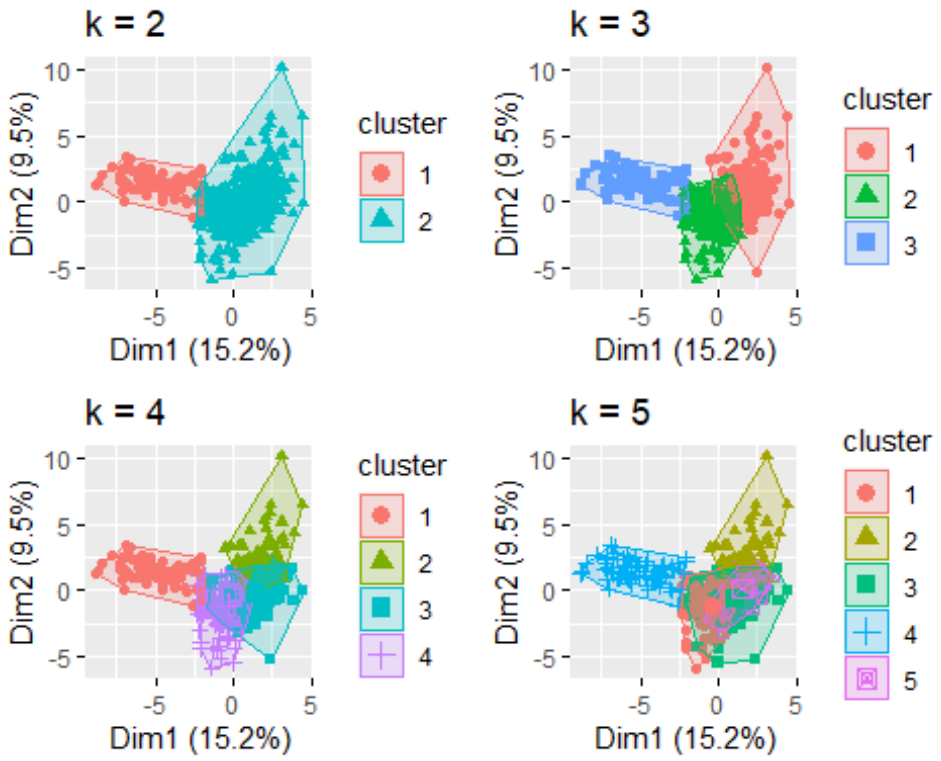
K Means Algorithms for Both Purchase Data sets

```
k_bp2 <- kmeans(both_purchase_data_scaled, centers = 2, nstart = 25)
k_bp3 <- kmeans(both_purchase_data_scaled, centers = 3, nstart = 25)
k_bp4 <- kmeans(both_purchase_data_scaled, centers = 4, nstart = 25)
k_bp5 <- kmeans(both_purchase_data_scaled, centers = 5, nstart = 25)

# plots to compare
plot_bp2 <- fviz_cluster(k_bp2, geom = "point",
                         data = both_purchase_data_scaled) + ggtitle("k = 2")
plot_bp3 <- fviz_cluster(k_bp3, geom = "point",
                         data = both_purchase_data_scaled) + ggtitle("k = 3")
plot_bp4 <- fviz_cluster(k_bp4, geom = "point",
```

```
                          data = both_purchase_data_scaled) + ggtitle("k = 4")
plot_bp5 <- fviz_cluster(k_bp5, geom = "point",
                          data = both_purchase_data_scaled) + ggtitle("k = 5")


grid.arrange(plot_bp2, plot_bp3, plot_bp4, plot_bp5, nrow = 2)
```



After looking at both the silhouette and the four graphs above, the best k for both purchase data sets is 2 because 2 clusters have the smallest overlap and matches the silhouette model. 2 clusters is very practical for a marketing team to focus on when considering both purchase behavior and the basis of purchase.

Determining which Households fall into each cluster based on Purchaser Behavior

```
table(total_data$SEC, k_pb2$cluster)

table(total_data$FEH, k_pb2$cluster)

table(total_data$MT, k_pb2$cluster)

table(total_data$SEX, k_pb2$cluster)

table(total_data$AGE, k_pb2$cluster)

table(total_data$EDU, k_pb2$cluster)
```

```r
table(total_data$HS, k_pb2$cluster)

table(total_data$CHILD, k_pb2$cluster)

table(total_data$CS, k_pb2$cluster)

table(total_data$Affluence.Index, k_pb2$cluster)
```

Determining which Households fall into each cluster based on Basis of Purchase

```r
table(total_data$SEC, k_bop3$cluster)

table(total_data$FEH, k_bop3$cluster)

table(total_data$MT, k_bop3$cluster)

table(total_data$SEX, k_bop3$cluster)

table(total_data$AGE, k_bop3$cluster)

table(total_data$EDU, k_bop3$cluster)

table(total_data$HS, k_bop3$cluster)

table(total_data$CHILD, k_bop3$cluster)

table(total_data$CS, k_bop3$cluster)

table(total_data$Affluence.Index, k_bop3$cluster)
```

Determining which Households fall into each cluster based on Both Purchase Data Sets

```r
table(total_data$SEC, k_bp2$cluster)

table(total_data$FEH, k_bp2$cluster)

table(total_data$MT, k_bp2$cluster)

table(total_data$SEX, k_bp2$cluster)

table(total_data$AGE, k_bp2$cluster)

table(total_data$EDU, k_bp2$cluster)

table(total_data$HS, k_bp2$cluster)

table(total_data$CHILD, k_bp2$cluster)

table(total_data$CS, k_bp2$cluster)

table(total_data$Affluence.Index, k_bp2$cluster)
```

After analyzing the data sets for purchase behavior and basis of purchase, I believe the best way to segment our data is to focus on the 2 clusters on purchase behavior. These clusters have very different demographics, where segmenting them will be crucial for marketing. The basis of purchase and both sets combined yield interesting results, however, the basis of purchase clusters had a third cluster that didn't single out specific demographics and both sets combined had a second cluster that did the same as well.

For purchase behavior, cluster 1 should be referred to as the high socioeconomic, high educated group and cluster 2 should be referred to as the low socioeconomic, less educated group. The majority of the demographics were very similar in both groups such as mostly non-vegetarians, with a decent amount of vegetarians, most native languages being 10 with a decent amount of 4's and the rest spread out, very high female ratio, ages being high in category 4, medium in 3, slightly less in 2, and low in 1. The amount of members in each household is the same for both where most have four to five members and a decent amount of three and six member households and the same categories for children, mostly category 4 and a decent amount in category 2. Most households have televisions in both clusters and both have an affluence index between 10-19. The only differences between clusters was socioeconomic class and education. Cluster 1 tends to have a higher socioeconomic class and a higher education level, whereas, cluster 2 has a lower socioeconomic class and a lower education level than cluster 1.

Model for Picking which Cluster each Household belongs too

```
#Chosen model
#k_pb2 <- kmeans(purchase_behavior_data_scaled, centers = 2, nstart = 25)

table(total_data$Member.id, k_bp2$cluster)
```