Final Exam

David Slusser

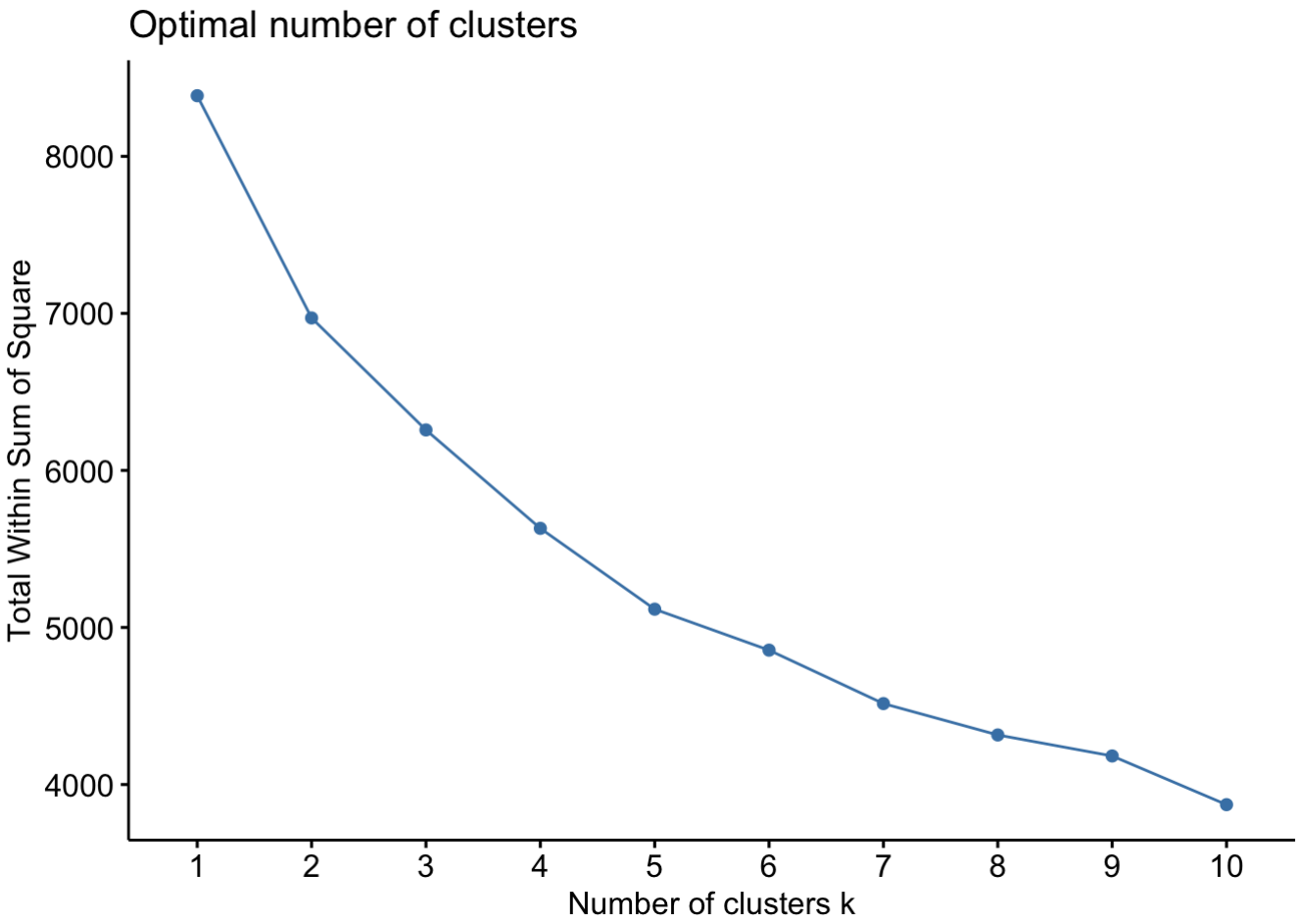
12/17/2020

This is the final exam Looking at consumer segmentation for consumers of bath soap Three questions CRISA has transaction and demographic data We are going to use K-Means 1. Purchase behavior 2. Basis for purchase 3. Both behavior and basis

```
# First load the packages that will be needed
library(readr) # Load in the csv of data from CRISA
library(caret) # For pre-processing the data
library(factoextra) # Clustering algorithms & visualization
library(Hmisc) # For summary stats
library(cluster) # For cluster validity

## Set the seed for consistency
set.seed(12345)

## Load in the bathsoaps data
BathSoap <- read_csv("~/Desktop/School/Graduate/Machine Learning/HW/Data/BathSoap.csv")

## Check to see if the data loaded
## There are 600 observations
## head(BathSoap) Commented out to suppress output

## We notice that columns 20-46 are characters and contain a % sign
## Need to convert into numeric
## This removes the "%" and makes the value numeric
BathSoap <- data.frame(sapply(BathSoap, function(x) as.numeric(gsub("%", "", x))))

## The brandwise purchasing since buying all A is similar to buying all B, we need to cr
eate a new variable, something like max brand, so create that before scaling
BathSoap$maxBrand <- apply(BathSoap[c(23:30)], 1, max) # Select the max value between co
lumns 23:30

## Get sum stats of columns
# describe(BathSoap) ## Not categories 5 and 14 have the highest mean
# This is commented out to suppress the output file


## We need to scale the data
## The economic class, eating habits, native language, gender, age,  education, presence
of children, and television availability are all dummy variables
## Thus, only scale columns 11-47
## This takes z-score of the variables in columns 11-47
BathSoap[c(11:47)] <- lapply(BathSoap[c(11:47)], function(x) c(scale(x)))
```

Now that the data is scaled, use k-means for purchase behavior

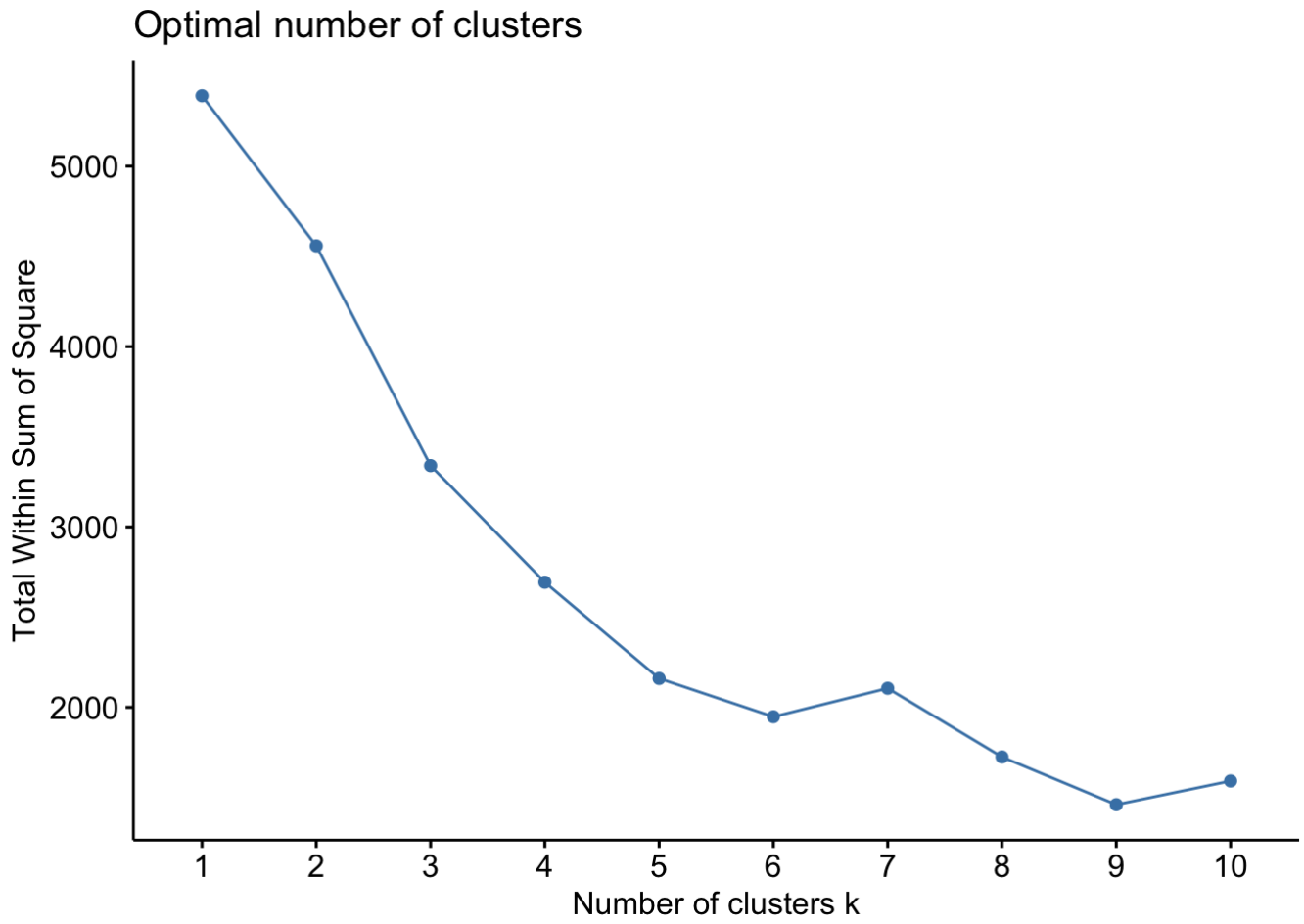
```
## Use k-means to cluster groups

## First, look at purchase behavior
## We will use the bathSoap, but only use variables related to purchase behavior
## This includes brand loyalty
## This includes number of brands purchased, brand runs, total volume, number of transac
tions, value, Trans/Brand Runs, Vol/Trans, average price, and max of one brand and other
## We also care about the sex and age, as older men will prefer different things compare
d to younger woman, and those who are more affluent will prefer difference things. More
 affluent individuals will buy more high priced things, so include thesevariables as wel
l
## Don't have domain knowledge so let's use "wss" to determine the number of K
## My theory is there are two types of people (those who care about brand and those who
 don't) so I expect K = 2, but there could be people who prefer certain brands (such as
 high end, medium level, and lower tier brands) instead of just 1 so do need the wss to
 determine k

## Determine the k
fviz_nbclust(BathSoap[c(11:22, 31, 47)], kmeans, method = "wss")
```



```
## Using the elbow method, we get k = 3 (which advertisement firm would choose between 2
-5)
## The three clusters in theory would then be those who care about a single brand, those
who like brand bundles, and those who do not care about the brand but we have to see the
means of the clusters

## Create the kmeans for brands
k.brand <- kmeans(BathSoap[c(11:22, 31, 47)], centers = 3, nstart = 25) # k = 3, number
 of restarts = 25
k.brand$centers
```

```
##   Affluence.Index No..of.Brands Brand.Runs Total.Volume No..of..Trans
## 1      -0.2390270    -0.4609266 -0.6846086    0.3903763    -0.3285947
## 2       0.4986536     0.8017672  0.9476303    0.2874630     0.8416093
## 3      -0.3151023    -0.4492764 -0.4263012   -0.5771744    -0.5878525
##        Value Trans...Brand.Runs   Vol.Tran Avg..Price Pur.Vol.No.Promo....
## 1  0.1224281          0.6786640  0.7474922 -0.5597747            0.2839561
## 2  0.4583424         -0.2837304 -0.3642194  0.2633391           -0.3834419
## 3 -0.5458379         -0.2270841 -0.1988727  0.1582578            0.1672972
##   Pur.Vol.Promo.6.. Pur.Vol.Other.Promo.. Others.999   maxBrand
## 1        -0.4164781            0.06606614 -1.0860904  1.2218536
## 2         0.4659933            0.03897392  0.3186390 -0.5020581
## 3        -0.1498545           -0.08809218  0.4975989 -0.4175223
```

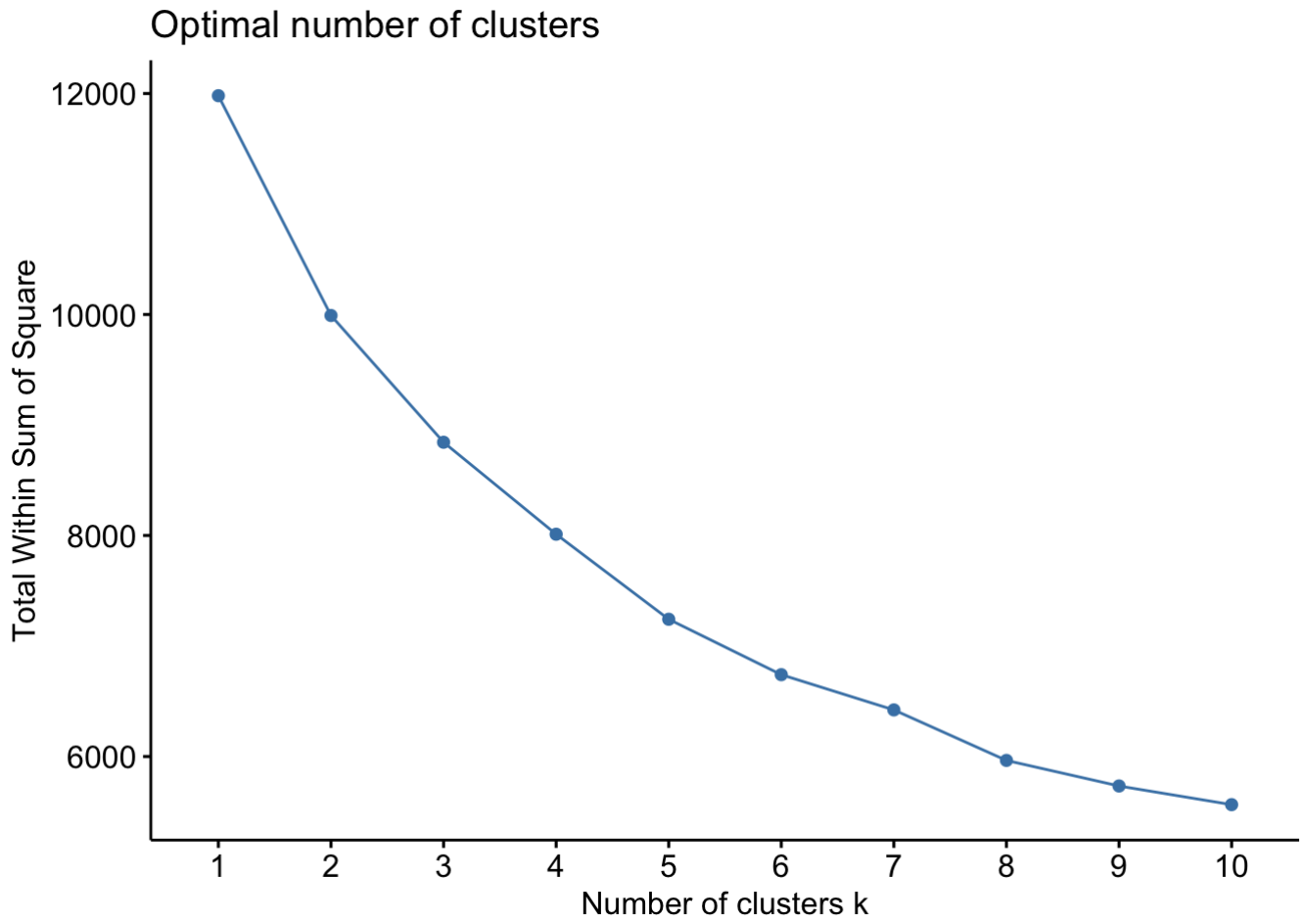
```
## Second, look at basis for purchase
## The variables we prefer are pur no promo, pur vol promo 6%, pur vol other promo %, pr
ice categories 1-4, and selling propositionwise 5 and 14 (the only means > 10, see code
 line 42)
## We need to determine k

fviz_nbclust(BathSoap[c(20:22, 32:36, 45)], kmeans, method = "wss")
```




```
# We get 2 clusters as the optimal

## Create the kmeans for brands
k.basis <- kmeans(BathSoap[c(20:22, 32:36, 45)], centers = 2, nstart = 25) # k = 2, numb
er of restarts = 25
k.basis$centers
```

```
##   Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..   Pr.Cat.1
## 1          -0.03029684        0.05928166           -0.02675504  0.1163032
## 2           0.20578245       -0.40265337            0.18172581 -0.7899554
##     Pr.Cat.2   Pr.Cat.3    Pr.Cat.4 PropCat.5 PropCat.14
## 1  0.1668112 -0.3515442  0.04920528  0.162797 -0.3521502
## 2 -1.1330160  2.3877611 -0.33421249 -1.105751  2.3918777
```

```
## Third, for behavior and basis
## Determine the value of K
fviz_nbclust(BathSoap[c(11:22, 31:36, 45, 47)], kmeans, method = "wss")
```




```
## Optimal number of clusters is 2

## Create the kmeans for both behavior and basis
k.bb <- kmeans(BathSoap[c(11:22, 31:36, 45, 47)], centers = 2, nstart = 25) # k = 2, num
ber of restarts = 25
k.bb$centers
```

```
##   Affluence.Index No..of.Brands Brand.Runs Total.Volume No..of..Trans
## 1      -0.7818884   -0.57242957 -0.7943754   0.14675880   -0.41531721
## 2       0.1049415    0.07682892  0.1066175  -0.01969731    0.05574201
##         Value Trans...Brand.Runs    Vol.Tran Avg..Price Pur.Vol.No.Promo....
## 1 -0.50773503          1.0757298  0.56666044  -1.317165           0.19235477
## 2  0.06814591         -0.1443796 -0.07605462   0.176784          -0.02581699
##   Pur.Vol.Promo.6.. Pur.Vol.Other.Promo.. Others.999   Pr.Cat.1   Pr.Cat.2
## 1       -0.42283516             0.2284974 -1.2614495 -0.8028521 -1.1863116
## 2        0.05675103            -0.0306679  0.1693061  0.1077552  0.1592214
##     Pr.Cat.3   Pr.Cat.4  PropCat.5 PropCat.14   maxBrand
## 1  2.4574051 -0.3263334 -1.1376353  2.4600492  1.4356025
## 2 -0.3298219  0.0437990  0.1526883 -0.3301767 -0.1926801
```

```
## Use k-means to cluster groups
BathSoap$segmentBehavior <- k.brand$cluster
BathSoap$segmentBasis <- k.basis$cluster
BathSoap$segmentBB <- k.bb$cluster

## Calculate the "success" (i.e. the stability)
sil <- silhouette(k.bb$cluster, dist(BathSoap))
si.sum <- summary(sil)
si.sum$clus.avg.widths # Get average widths
```

```
##          1          2
## 0.40299218 0.08942867
```

```
## Remember that silhouette is [-1, 1]
## What we find is that segment 1 is more stable (closer to 1) than segment 2
## Implies we should use segment 1 for our marketing campaign since we are more confiden
t in this
```

Percentages of total purchases by various brands should be classified as the max of the values as it would help illustrate if the individual prefers a certain brand (a higher max), a bundle (a max value of say 20%-80% where the individual prefers certain brands but isn't focused in one one specific brand), and one where the max is low as the individual has no preference of the brand. This is more helpful because somebody who prefers brand A has the same level of loyalty as somebody who prefers brand B, and thus we have a caculated value of the people who prefer a certain brand. Without creating this new variable, the optimal level of K would be large and the distances between the clusters would be larger (i.e. somebody who really prefers A to C would have a wider gap despite being equally as loyal as somebody who prefers C to A).

- For purchase behavior, we derived 3 optimal segments, which could be thought of as people preferring one brand, multiple brands, or no preference in brands. Segment 1 is more affluent and buys the most amount of brands, buying for a higher price (matching segment 3's willingness to pay) and has no strong preference for a particular brand. They generate the least amount of value. Segment 2 is not very affluent and has a strong preference for brands, making little transactions on the same brand in a row, but preferring a bundle of different brands. This segment buys lower priced goods and doesn't buy goods without a promotion offer. Segment 3 is the least affluent segment but they do care about the brands, preferring one singular brand over a bundle and will purchase these goods with promos, despite their low volume transactions.
- For purchase basis, there are 2 optimal segments Segment 1 prefers price category 3 and buys under proposition 14, and either no promo or other promos. Segment 2 prefers price categories 1 and 2, proposition 5, and promo 6.
- For behavior and basis, there are 2 optimal segments. Segment 1 is less affluent but prefers certain brands despute not consecutively buying the same brand. They don't buy a large amount of volume, but when they make a transaction, segment 1 has high volume. They prefer to purchase the lower priced products and prefers price category 2 and promotion five. Segment 2 is more affluent and buys a variety of different brands and purchases a high volume of transactions. They do have a preference for a singular brand but also have a bundle of brands in their market basket. They create value within their transactions despite not generating a lot of volume per their transaction, but these come with a higher price.

- The best segment is combining the behavior and basis as there are two distinct segments. Segment 1 generates less value than segment 2 despite their volume per transaction (they pay lower prices). Segment 2, on the other hand, is more affluent and do not have a strong preference for brand and are willing to pay more for goods than segment 1 and prefer price category 2 compared to segment 1 (who prefers price category 3)
- As a result, the marketing team should focus on segment 1.Segment 1 is very brand loyal and less affluent and buy products that are lower cost. We know them to be more stable (silhouette of 0.40 for segment 1 compared to 0.09 for segment 2). Since segment 1 is more stable, we want the marketing team to focus on them because there's more confidence on what this segment cares about

```
## Graph the segment Behavior and Basis clusters
## This is for the presentation that will be used to give to the marketing team
## View the clusters and show who we should target in our marketing campaign
fviz_cluster(k.bb, data = BathSoap,
             palette = c("#BF5700", "darkgreen"),
             geom = "point",
             ellipse.type = "convex",
             ggtheme = theme_bw()) +
  labs(title = "Bath Soaps Market Segmentation",
       subtitle = "CRISA Marketing Firm",
       caption = "David Slusser\nMIS 64060: Machine Learning") +
  theme(text = element_text(size = 15))
```