Final.R

Jason

2021-05-09

```
# AFinal Assignment Fundamentals of Machine Learning
# Data comes from bathsoap.csv


library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
WD<-setwd("C:/Users/Jason/Documents/MSBA/Fundamentals for Machine Learning/Final")
Soap<-read.csv("bathsoap.csv", header = TRUE)


row.names(Soap) <- Soap[,1]
Soap1 <- Soap[,-1]

summary(Soap1)
```

```
##       SEC            FEH              MT              SEX       
##  Min.   :1.00   Min.   :0.000   Min.   : 0.000   Min.   :0.000  
##  1st Qu.:1.75   1st Qu.:1.000   1st Qu.: 4.000   1st Qu.:2.000  
##  Median :2.50   Median :3.000   Median :10.000   Median :2.000  
##  Mean   :2.50   Mean   :2.048   Mean   : 8.178   Mean   :1.738  
##  3rd Qu.:3.25   3rd Qu.:3.000   3rd Qu.:10.000   3rd Qu.:2.000  
##  Max.   :4.00   Max.   :3.000   Max.   :19.000   Max.   :2.000  
##       AGE             EDU              HS             CHILD      
##  Min.   :1.000   Min.   :0.000   Min.   : 0.000   Min.   :1.000  
##  1st Qu.:3.000   1st Qu.:3.000   1st Qu.: 3.000   1st Qu.:2.000  
##  Median :3.000   Median :4.500   Median : 4.000   Median :4.000  
##  Mean   :3.213   Mean   :4.043   Mean   : 4.192   Mean   :3.233  
##  3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.: 5.000   3rd Qu.:4.000  
##  Max.   :4.000   Max.   :9.000   Max.   :15.000   Max.   :5.000  
##        CS         Affluence.Index No..of.Brands     Brand.Runs
##  Min.   :0.0000   Min.   : 0.00   Min.   :1.000   Min.   : 1.00
##  1st Qu.:1.0000   1st Qu.:10.00   1st Qu.:2.000   1st Qu.: 8.00
##  Median :1.0000   Median :15.00   Median :3.000   Median :15.00
##  Mean   :0.9317   Mean   :17.02   Mean   :3.637   Mean   :15.75
##  3rd Qu.:1.0000   3rd Qu.:24.00   3rd Qu.:5.000   3rd Qu.:21.00
##  Max.   :2.0000   Max.   :53.00   Max.   :9.000   Max.   :74.00
##   Total.Volume   No..of..Trans        Value        Trans...Brand.Runs
##  Min.   :  150   Min.   :  1.00   Min.   :  20.0   Min.   : 1.000
##  1st Qu.: 6825   1st Qu.: 22.00   1st Qu.: 789.6   1st Qu.: 1.420
##  Median :10360   Median : 28.00   Median :1216.0   Median : 1.845
##  Mean   :11915   Mean   : 31.15   Mean   :1337.4   Mean   : 2.618
##  3rd Qu.:15344   3rd Qu.: 40.00   3rd Qu.:1675.8   3rd Qu.: 2.690
##  Max.   :50895   Max.   :138.00   Max.   :6371.9   Max.   :23.000
##     Vol.Tran         Avg..Price    Pur.Vol.No.Promo.... Pur.Vol.Promo.6..
##  Min.   :  94.43   Min.   : 5.62   Length:600           Length:600
##  1st Qu.: 250.51   1st Qu.: 9.76   Class :character     Class :character
##  Median : 361.52   Median :11.25   Mode  :character     Mode  :character
##  Mean   : 415.05   Mean   :11.83
##  3rd Qu.: 490.89   3rd Qu.:13.42
##  Max.   :2525.00   Max.   :33.33
##  Pur.Vol.Other.Promo.. Br..Cd..57..144     Br..Cd..55        Br..Cd..272
##  Length:600            Length:600         Length:600         Length:600
##  Class :character      Class :character   Class :character   Class :character
##  Mode  :character      Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Br..Cd..286         Br..Cd..24        Br..Cd..481        Br..Cd..352
##  Length:600         Length:600         Length:600         Length:600
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   Br..Cd..5          Others.999          Pr.Cat.1           Pr.Cat.2
##  Length:600         Length:600         Length:600         Length:600
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##    Pr.Cat.3           Pr.Cat.4          PropCat.5          PropCat.6
##  Length:600         Length:600         Length:600         Length:600
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   PropCat.7          PropCat.8          PropCat.9          PropCat.10
##  Length:600         Length:600         Length:600         Length:600
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##   PropCat.11         PropCat.12         PropCat.13         PropCat.14
##  Length:600         Length:600         Length:600         Length:600
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   PropCat.15
##  Length:600
##  Class :character
##  Mode  :character
##
##
##
```

```
#dataset is skewed towards females since mean of sex is skewed towards 2
#removing non-gender identified since hygiene products are gender specific

Soap2 <- Soap1[Soap1$SEX != 0,]

#splitting males and females for the analysis
#again, this is because hygiene products are gender specific
#if we're looking to increase effectiveness of promotions for hygiene products
#then we don't want to spend promotion dollars
#advertising male products to females and vice versa

SoapMale <- Soap2[Soap2$SEX == 1,]
SoapFemale <- Soap2[Soap2$SEX == 2,]

511/(511+21)
```

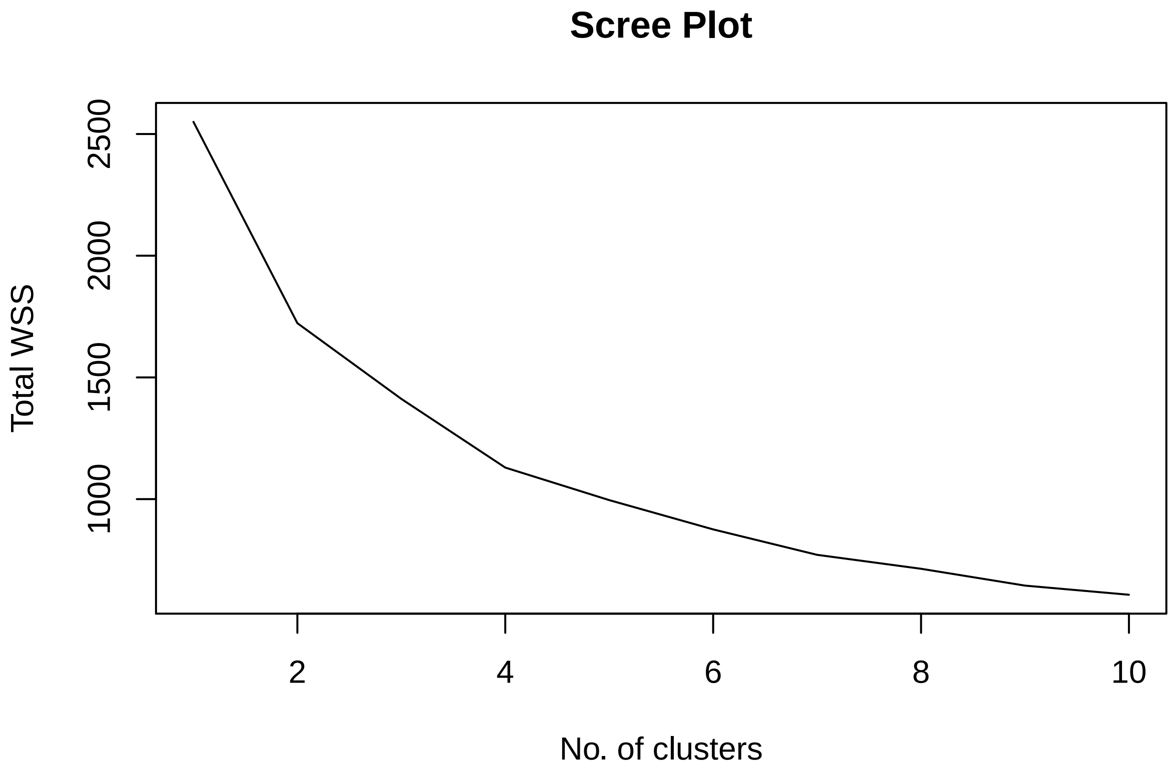
```
## [1] 0.9605263
```

```
#dataset is 96% female, so the rest of the analysis will focus on female data
#male adoption rates of the products are too low to meaningfully segment

SoapF <- SoapFemale[,-4] #dropping sex since it's no longer relevant

#making percentage variables numeric
for (i in 18:44) {

  SoapF[,i] <- as.numeric(sub("%", "", SoapF[,i]))/100

}

#using the max of volume purchased of 1 brand as brand loyalty
#since this is the most loyal the consumer would be
SoapF$BLoyalty <- pmax(SoapF[,21], SoapF[,22], SoapF[,23], SoapF[,24], SoapF[,25], SoapF[,26], SoapF[,2

#dropping brand codes and other
#dropping other since that could be multiple brands and purchases
```

```
SoapF1 <- SoapF[,-c(21:29)]

#Calculating Max volume per brand
SoapF1$MaxBrandVolume <- SoapF1[,36]*SoapF1[,12]

#calculating Max value per brand
SoapF1$MaxBrandValue <- SoapF1[,36]*SoapF1[,14]

#calculating promotion susceptibility
SoapF1$PromoWorks <- 1-SoapF1[,18]

#normalizing numeric data
SoapFNorm <- scale(SoapF1[,9:39])

#creating matrix for purchase behavior
SoapPbehavior <- SoapFNorm[,c(7:8,28:29,31)]
#creating matrix for purchase basis
SoapPbasis <- SoapFNorm[,c(9,13:27,30)]
#creating matrix for both
SoapBoth <- SoapFNorm[,c(7:9,13:31)]

#k-means clustering for purchase behavior
BehavClus. <- sapply(1:10, function(i){return(kmeans(SoapPbehavior, centers = i)$tot.withinss)})
cbind(No.of.Cluters=1:10, BehavClus.)
```

```
##       No.of.Cluters BehavClus.
##  [1,]             1  2550.0000
##  [2,]             2  1722.8119
##  [3,]             3  1411.7686
##  [4,]             4  1129.5936
##  [5,]             5   995.9092
##  [6,]             6   875.6732
##  [7,]             7   771.1464
##  [8,]             8   713.8088
##  [9,]             9   644.6484
## [10,]            10   607.1398
```

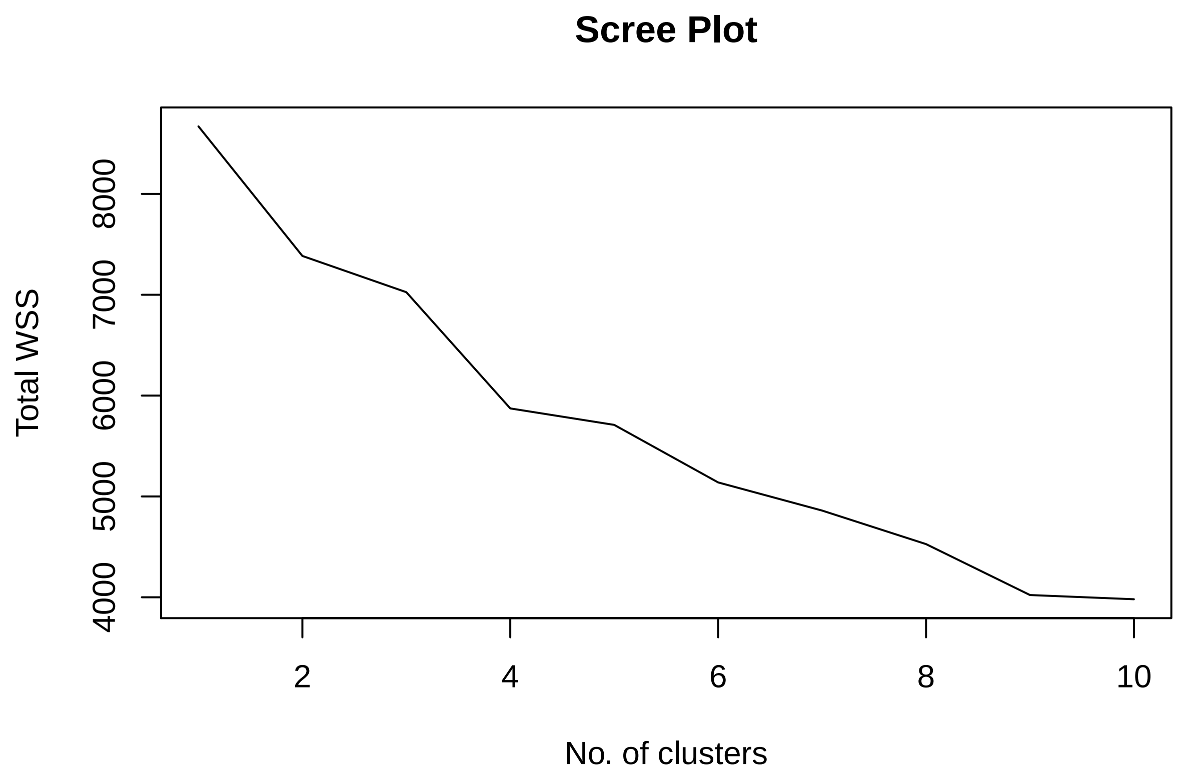
```
plot(1:10, BehavClus., type="l", xlab = "No. of clusters", ylab = "Total WSS", main = "Scree Plot")
```



```r
#5 is ideal k based on domain, purpose, and results of scree plot

#k-means clustering for purchase basis
BasisClus. <- sapply(1:10, function(i){return(kmeans(SoapPbasis, centers = i)$tot.withinss)})
cbind(No.of.Cluters=1:10, BasisClus.)
```

```
##       No.of.Cluters BasisClus.
##  [1,]             1   8670.000
##  [2,]             2   7384.866
##  [3,]             3   7025.533
##  [4,]             4   5872.763
##  [5,]             5   5709.574
##  [6,]             6   5139.107
##  [7,]             7   4859.936
##  [8,]             8   4527.278
##  [9,]             9   4022.390
## [10,]            10   3980.629
```

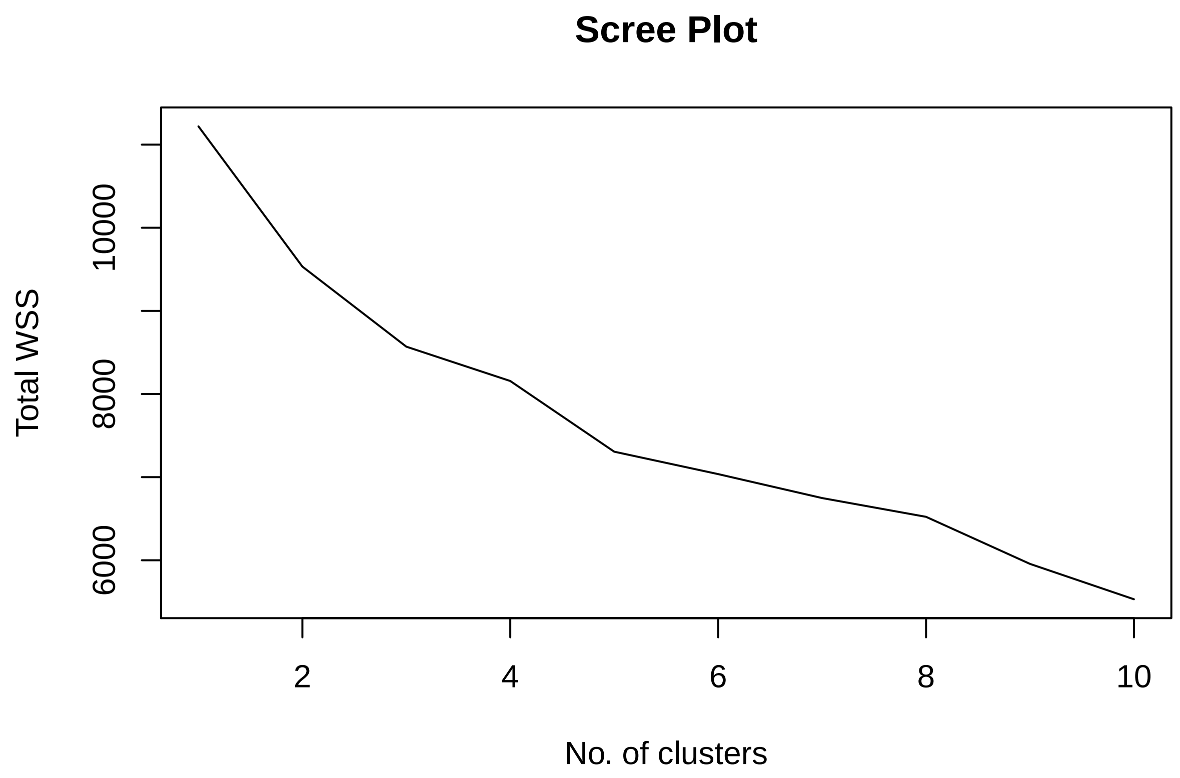
```r
plot(1:10, BasisClus., type="l", xlab = "No. of clusters", ylab = "Total WSS", main = "Scree Plot")
```



```
#5 is ideal k based on domain, purpose, and results of scree plot

#k-means clustering for both
BothClus. <- sapply(1:10, function(i){return(kmeans(SoapBoth, centers = i)$tot.withinss)})
cbind(No.of.Cluters=1:10, BothClus.)
```

```
##       No.of.Cluters BothClus.
##  [1,]             1 11220.000
##  [2,]             2  9533.073
##  [3,]             3  8569.611
##  [4,]             4  8156.469
##  [5,]             5  7306.333
##  [6,]             6  7036.743
##  [7,]             7  6748.433
##  [8,]             8  6522.582
##  [9,]             9  5956.463
## [10,]            10  5531.212
```

```
plot(1:10, BothClus., type="l", xlab = "No. of clusters", ylab = "Total WSS", main = "Scree Plot")
```



```
#ideal isn't clear from scree plot
#based on domain, purpose, and results from previous two, k = 5

BehavClus. <- kmeans(SoapPbehavior, centers = 5)
BasisClus. <- kmeans(SoapPbasis, centers = 5)
BothClus. <- kmeans(SoapBoth, centers = 5)

SoapF1$BehavClus <- BehavClus.$cluster
SoapF1$BasisClus <- BasisClus.$cluster
SoapF1$BothClus <- BothClus.$cluster

#removing periods in names to run dplyr
names(SoapF1) <- gsub("\\.", "", names(SoapF1))

#calculate the average value and grouping by cluster for each cluster method
#Average value = sum(total value)/sum(total transactions)
AvgValueBehav <- SoapF1 %>% group_by(BehavClus) %>% summarise(Value = sum(Value)/sum(NoofTrans))
AvgValueBehav
```

```
## # A tibble: 5 x 2
##   BehavClus Value
## *     <int> <dbl>
## 1         1  37.0
## 2         2  44.1
## 3         3  46.9
## 4          4  39.2
## 5          5  92.6
```

```
ClusterMixBehav <- SoapF1 %>% group_by(BehavClus) %>% summarise(Percentage = n()) %>% mutate(Percentage=
ClusterMixBehav
```

```
## # A tibble: 5 x 2
##   BehavClus Percentage
## *     <int>      <dbl>
## 1         1      15.9
## 2         2      25.4
## 3         3       2.74
## 4         4      49.3
## 5         5       6.65
```

```
AvgValueBasis <- SoapF1 %>% group_by(BasisClus) %>% summarise(Value = sum(Value)/sum(NoofTrans))
AvgValueBasis
```

```
## # A tibble: 5 x 2
##   BasisClus Value
## *     <int> <dbl>
## 1         1  43.8
## 2         2  40.3
## 3         3  61.8
## 4         4  43.3
## 5         5  37.1
```

```
ClusterMixBasis <- SoapF1 %>% group_by(BasisClus) %>% summarise(Percentage = n()) %>% mutate(Percentage=
ClusterMixBasis
```

```
## # A tibble: 5 x 2
##   BasisClus Percentage
## *     <int>      <dbl>
## 1         1      52.6
## 2         2       9.39
## 3         3       3.13
## 4         4      23.1
## 5         5      11.7
```

```
AvgValueBoth <- SoapF1 %>% group_by(BothClus) %>% summarise(Value = sum(Value)/sum(NoofTrans))
AvgValueBoth
```

```
## # A tibble: 5 x 2
##   BothClus Value
## *    <int> <dbl>
## 1        1  37.7
## 2        2  37.7
## 3        3  58.5
## 4        4  45.0
## 5        5  30.3
```

```
ClusterMixBoth <- SoapF1 %>% group_by(BothClus) %>% summarise(Percentage = n()) %>% mutate(Percentage=Pe
ClusterMixBoth
```

```
## # A tibble: 5 x 2
##   BothClus Percentage
## *    <int>      <dbl>
## 1        1     55.4
## 2        2     11.0
## 3        3     29.2
## 4        4      3.72
## 5        5      0.783
```

```
#Clustering by both is the best
#It allows us to identify 2 highest value clusters
#then target that cluster with promotions
#Behavior clustering identifies highest value cluster
#However, the addressable market of that customer is much lower (only 6.07%)
#clustering by both identifies 2 high value clusters and the addressable market
#is much higher (21.94%)
```