ThesisSurveyCluster

08 March, 2021

Reference to Methodology applied.

One-hot encoding

```
dataf <- dummy_cols(data, select_columns = c("Machine", "BrandChange",
        "PurchaseLocation", "Education", "AgeCategory", "Frequency_Specialty", "Home",
        "Occupation", "Gender", "Subscription_Not_Likely", "Criteria_Type_Coffee",
        "Supermarket_Positive_ Reasons", "Supermarket_Negative_ Reasons", "Language"),
        remove_selected_columns = TRUE, ignore_na = TRUE)
```

Including categorical variables

```
# Prepare Data
mydata <- na.omit(dataf) # listwise deletion of missing
NumData <- scale(dataf[,c(2:6, 7:13)]) # standardize variables
NewData <- cbind(NumData, dataf[,c(14:50)])
```

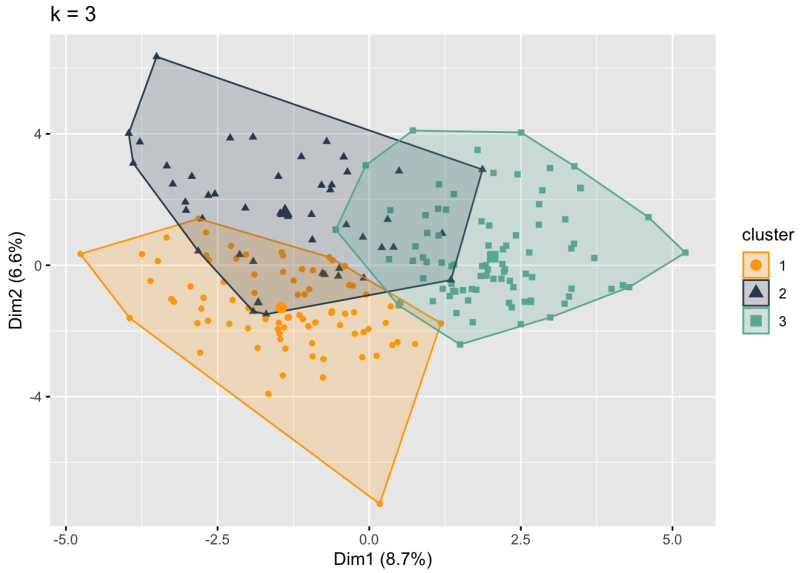
```
set.seed(123)

# K-Means Cluster Analysis
fit <- kmeans(na.omit(NewData), 3, nstart = 25) #3 cluster solution
# get cluster means
aggregate(na.omit(NewData),by=list(fit$cluster),FUN=mean)
```

```
  Group.1 AmountWeek AmountOutMonth MoneyCoffee MoneyGroceries KnowledgeCoffee
1       1 -0.5185918     -0.2422381 -0.37479743    -0.13300035      -0.1605922
2       2  0.3253951      0.7606476  0.53978428     0.38390600       0.6812677
3       3  0.3186313     -0.2033604  0.02260646    -0.09440055      -0.1942204
  Purchase_Price Purchase_Sustainability Purchase_Certificate
1      0.4907187               0.6656930            0.6168074
2     -0.1402655               0.2364084            0.1324155
3     -0.3421044              -0.7088341           -0.5972815
  Purchase_Fairtrade Purchase_Packaging Subscription_Likely  App_Likely
1          0.6520526         0.68111564          -0.0428773 -0.01373208
2          0.2860368        -0.08173388           1.1130149  0.85640419
3         -0.7053639        -0.56088632          -0.5600639 -0.47855612
  Machine_Aeropress Machine_CupMachine Machine_Espresso machine
1              0.00          0.2750000                0.3125000
2              0.02          0.2000000                0.4600000
3              0.00          0.4395604                0.2417582
  Machine_Filter machine Machine_French press Machine_Instant coffee
1              0.2250000           0.07500000             0.01250000
2              0.1400000           0.02000000             0.02000000
3              0.2307692           0.01098901             0.02197802
  Machine_Moka pot Machine_V60 BrandChange_Every time BrandChange_Never
1       0.10000000        0.00                  0.00         0.2875000
2       0.06000000        0.08                  0.06         0.2200000
3       0.05494505        0.00                  0.00         0.4065934
  BrandChange_Sometimes BrandChange_Very often PurchaseLocation_E-commerce
1             0.5875000             0.12500000                   0.1250000
2             0.5800000             0.14000000                   0.2400000
3             0.5604396             0.03296703                   0.1758242
  PurchaseLocation_Online subscription
1                      0.02500000
```

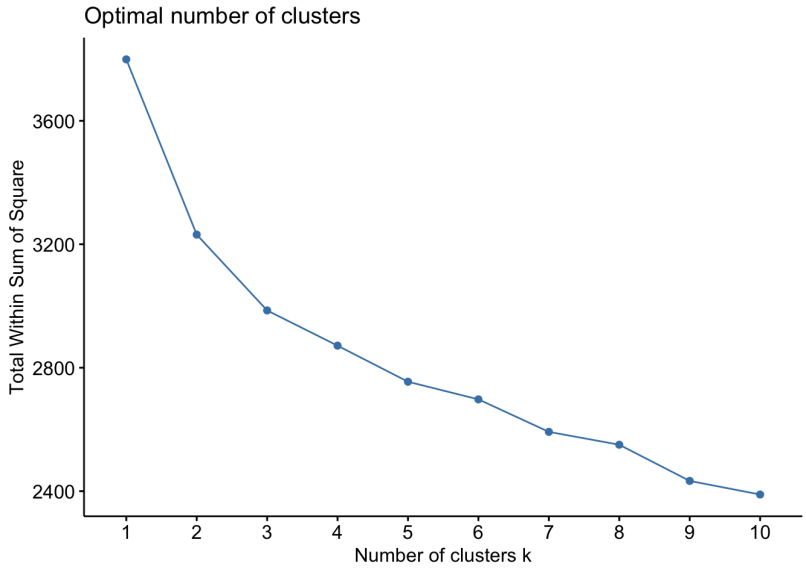
```
2                           0.12000000
3                           0.01098901
  PurchaseLocation_Specialty stores or cafés PurchaseLocation_The supermarket
1                                 0.11250000                         0.7375000
2                                 0.24000000                         0.4000000
3                                 0.04395604                         0.7692308
  Education_Associate degree Education_Bachelor's degree
1                  0.0625000                   0.4750000
2                  0.0600000                   0.6400000
3                  0.1098901                   0.5714286
  Education_Elementary school Education_High school Education_Master
1                  0.01250000            0.10000000        0.3500000
2                  0.00000000            0.10000000        0.1600000
3                  0.01098901            0.07692308        0.2087912
  Education_Phd AgeCategory_< 18 AgeCategory_> 60 AgeCategory_18-25
1    0.00000000           0.0125        0.0125000         0.3625000
2    0.04000000           0.0000        0.0200000         0.2800000
3    0.02197802           0.0000        0.0989011         0.2197802
  AgeCategory_25-45 AgeCategory_45-60 Frequency_Specialty_Always
1         0.5125000         0.1000000                 0.08750000
2         0.4400000         0.2600000                 0.28000000
3         0.3846154         0.2967033                 0.03296703
  Frequency_Specialty_I do (did) not know what this is
1                                            0.2125000
2                                            0.0800000
3                                            0.3296703
  Frequency_Specialty_Never Frequency_Specialty_Only in cafes
1                 0.1625000                          0.2750000
2                 0.0800000                          0.1200000
3                 0.2417582                          0.1978022
  Frequency_Specialty_Sometimes Home_Rural (Town) Home_Suburbs
1                     0.2625000         0.0875000    0.0500000
2                     0.4400000         0.0600000    0.0800000
3                     0.1978022         0.1318681    0.1098901
  Home_Urban (City) Occupation_Employed (Full time)
1         0.8625000                        0.5500000
2         0.8600000                        0.5000000
3         0.7582418                        0.6263736
  Occupation_Employed (Part time)
1                       0.1125000
2                       0.2400000
3                       0.2087912
```

```
# append cluster assignment
mydata <- data.frame(na.omit(NewData), fit$cluster)
```

```
fviz_cluster(fit, geom = "point", data = mydata, outlier.color = "black", palette =
        dani) +
  ggtitle("k = 3")
```

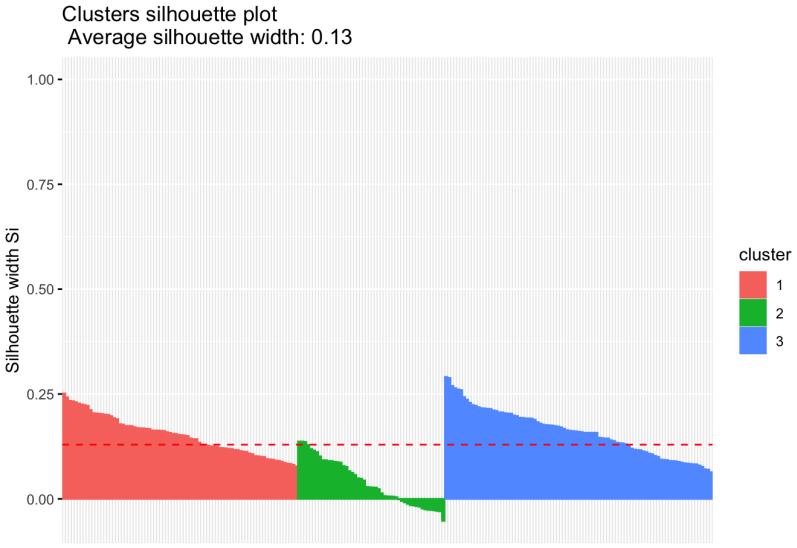



```
mydata <- na.omit(mydata)

distance <- get_dist(mydata)
graph <- fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high =
        "#FC4E07"))

set.seed(224)

fviz_nbclust(mydata, kmeans, method = "wss")
```




```
sil <- silhouette(fit$cluster, dist(mydata))
fviz_silhouette(sil)
```

```
  cluster size ave.sil.width
1       1   80          0.15
2       2   50          0.04
3       3   91          0.16
```


Clusters silhouette plot
Average silhouette width: 0.13
Silhouette width Si
1.00
0.75
0.50
0.25
0.00
cluster
1
2
3

```
Results <- as.data.table(aggregate(na.omit(mydata[,1:5]),
        by=list(cluster=fit$cluster), mean), by = round)

Results_Round <- round(Results)
my_table(Results_Round)
```

cluster	AmountWeek	AmountOutMonth	MoneyCoffee	MoneyGroceries	KnowledgeCoffee
1	-1	0	0	0	0
2	0	1	1	0	1
3	0	0	0	0	0

```
Results <- as.data.table(aggregate(na.omit(mydata[,19:25]),
        by=list(cluster=fit$cluster), mean), by = round)

Results_Round <- round(Results)
my_table(Results_Round)
```

cluster	Machine_Moka.pot	Machine_V60	BrandChange_Every.time	BrandChange_Never	BrandChange_Sometimes

BrandChange_Very.often	PurchaseLocation_E.commerce
1	0 0 0 0 1 0 0
2	0 0 0 0 1 0 0
3	0 0 0 0 1 0 0

```
Results <- as.data.table(aggregate(na.omit(mydata[,6:9]),
        by=list(cluster=fit$cluster), mean), by = round)

Results_Round <- round(Results)
my_table(Results_Round)
```

cluster	Purchase_Price	Purchase_Sustainability	Purchase_Certificate	Purchase_Fairtrade
1	0	1	1	1
2	0	0	0	0
3	0	-1	-1	-1

```
Results <- as.data.table(aggregate(na.omit(mydata[,35:40]),
        by=list(cluster=fit$cluster), mean), by = round)

Results_Round <- round(Results)
my_table(Results_Round)
```

cluster	AgeCategory_..18	AgeCategory_..60	AgeCategory_18.25	AgeCategory_25.45	AgeCategory_45.60	Frequency_Speci
1	0	0	0	1	0	

0						
2	0	0	0	0	0	0
3	0	0	0	0	0	0

Caret package

```
# Prepare Data
mydata <- na.omit(dat_transformed) # listwise deletion of missing
mydata <- scale(dat_transformed) # standardize variables
```

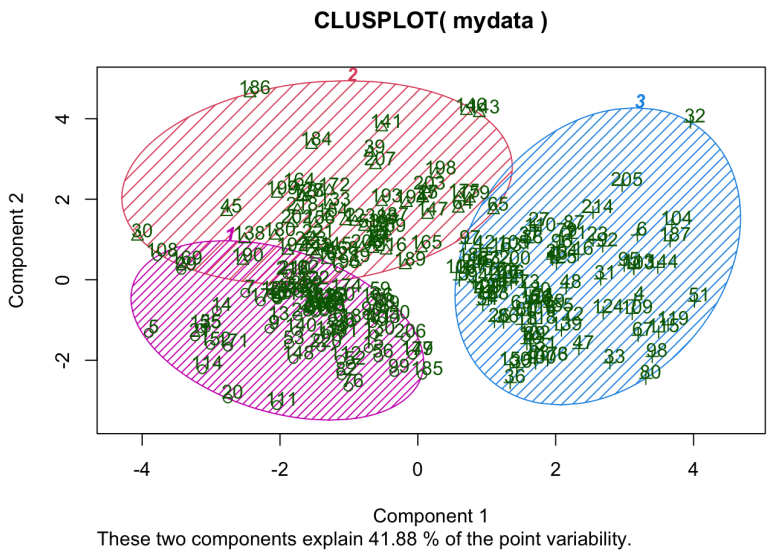
Clustering Numerical

```
# Prepare Data
NumMydata <- na.omit(data[,c(2:5, 12:17, 19, 21)]) # listwise deletion of missing
Mydata <- scale(data[,c(2:5, 12:17, 19, 21)]) # standardize variables
```

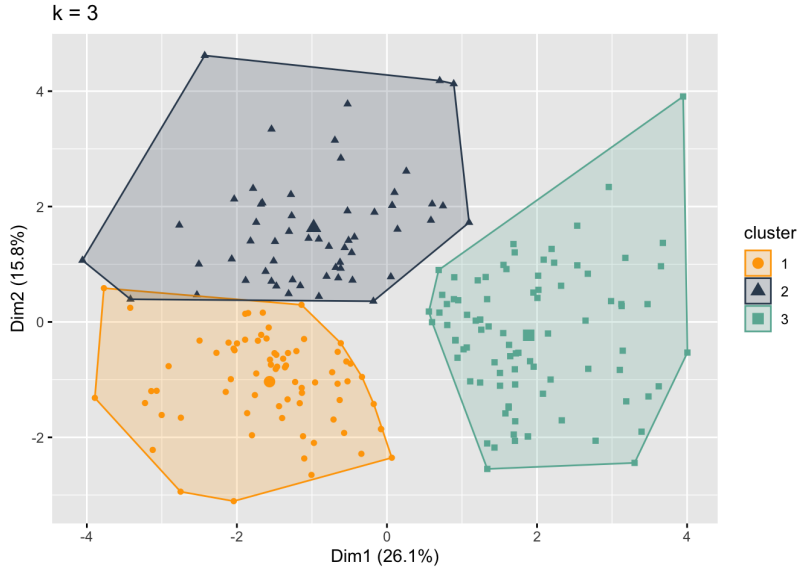
```
set.seed(123)

# K-Means Cluster Analysis
fit <- kmeans(na.omit(Mydata), 3, nstart = 1) #3 cluster solution
# get cluster means
aggregate(na.omit(Mydata),by=list(fit$cluster),FUN=mean)
```

```
  Group.1  AmountWeek AmountOutMonth MoneyCoffee MoneyGroceries KnowledgeCoffee
1       1 -0.54073357     -0.2107891 -0.48191624    -0.24962426      -0.1084058
2       2  0.08654604      0.4989730  0.46146858     0.34524071       0.4395870
3       3  0.39150131     -0.1444687  0.09853028    -0.02358935      -0.1722591
  Purchase_Price Purchase_Sustainability Purchase_Certificate
1     0.50512472               0.6624937            0.4918490
2     0.03777871               0.4125406            0.3678209
3    -0.42850807              -0.8163814           -0.6370795
  Purchase_Fairtrade Purchase_Packaging Subscription_Likely App_Likely
1          0.4234596          0.4033658          -0.3604525 -0.3390900
2          0.5230238          0.3306540           1.1507465  1.0113369
3         -0.6958440         -0.5608863          -0.4455199 -0.3720062
```

```
# append cluster assignment
mydata <- data.frame(na.omit(Mydata), fit$cluster)

clusplot(mydata, fit$cluster, color=TRUE, shade=TRUE,
   labels=2, lines=0)
```

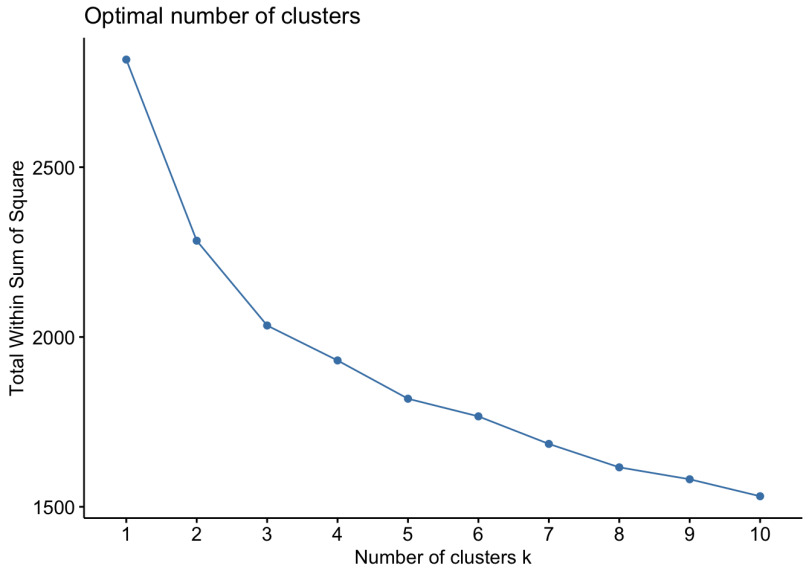



```
fviz_cluster(fit, geom = "point", data = mydata, outlier.color = "black", palette =
        dani) +
  ggtitle("k = 3")
```



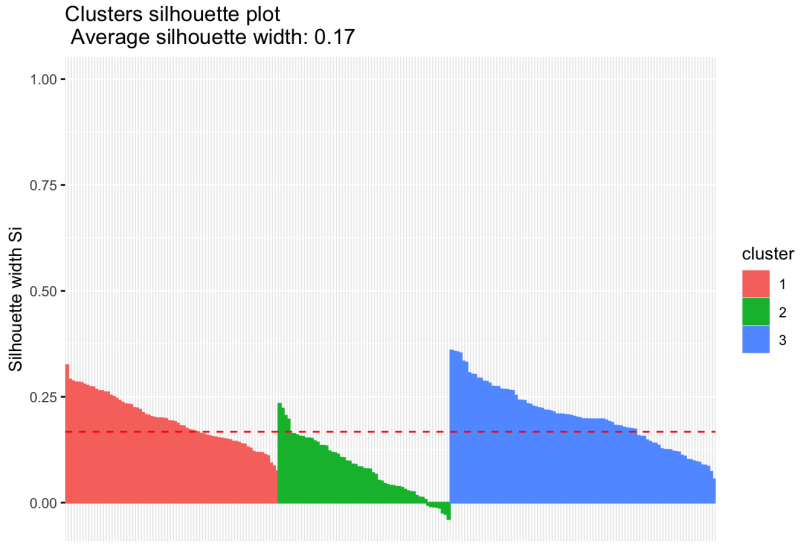
https://towardsdatascience.com/clustering-analysis-in-r-using-k-means-73eca4fb7967

```
mydata <- na.omit(mydata)

distance <- get_dist(mydata)
graph <- fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high =
        "#FC4E07"))

set.seed(224)

fviz_nbclust(mydata, kmeans, method = "wss")
```




```
sil <- silhouette(fit$cluster, dist(mydata))
fviz_silhouette(sil)
```

```
  cluster size ave.sil.width
1       1   73          0.19
2       2   59          0.08
3       3   91          0.20
```



```
Results <- as.data.table(aggregate(na.omit(NumMydata[,1:5]),
        by=list(cluster=fit$cluster), mean), by = round)

Results_Round <- round(Results)
my_table(Results_Round)
```

cluster	AmountWeek	AmountOutMonth	MoneyCoffee	MoneyGroceries	KnowledgeCoffee
1	12	6	16	212	5
2	20	13	35	297	7
3	23	7	27	244	5

```
Results <- as.data.table(aggregate(na.omit(NumMydata[,6:9]),
        by=list(cluster=fit$cluster), mean), by = round)

Results_Round <- round(Results)
my_table(Results_Round)
```

cluster	Purchase_Price	Purchase_Sustainability	Purchase_Certificate	Purchase_Fairtrade
1	4	4	3	4
2	3	4	3	4
3	3	2	2	2

```
Results <- as.data.table(aggregate(na.omit(NumMydata[,10:12]),
        by=list(cluster=fit$cluster), mean), by = round)

Results_Round <- round(Results)
my_table(Results_Round)
```

cluster	Purchase_Packaging	Subscription_Likely	App_Likely
1	3	3	3
2	3	7	7
3	2	3	3