ThesisSurveyCluster

21 March, 2021

Reference to Methodology applied.

One-hot encoding

Primarily several multi-option variables are removed from the dataset. Next dummies are created for the categorical variables.

If the option is not selected, the answer corresponds with 0.
If the option has been selected, the answer corresponds with 1.

```
data <- data[,-match(c("Criteria_Type_Coffee", "Subscription_Not_Likely",
        "Supermarket_Negative_ Reasons", "Supermarket_Positive_ Reasons", "Language",
        "Participant"),names(data))]

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

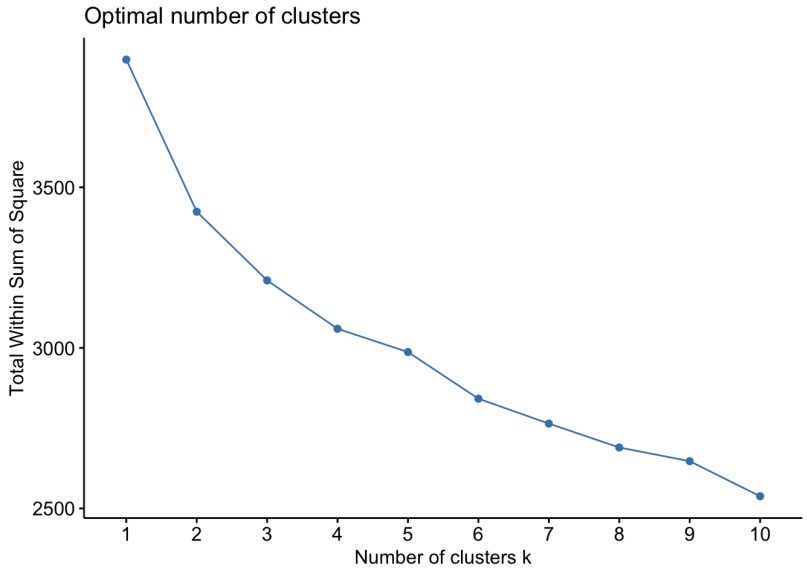
Clustering including categorical variables

The next step is to remove this missing variables (NA), standardize the numerical variables and combine to make the new data table to be used for the cluster analysis.

```
# Prepare Data
Orgdata <- na.omit(dataf) # listwise deletion of missing
stdata <- scale(Orgdata[,c(1:12)]) # standardize variables
NewData <- cbind(stdata, Orgdata[,c(13:50)])
```

Based on the graph below, I have decided to use 4 numbers of cluster.

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

distance <- get_dist(NewData)
graph <- fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high =
        "#FC4E07"))
fviz_nbclust(NewData, kmeans, method = "wss")
```



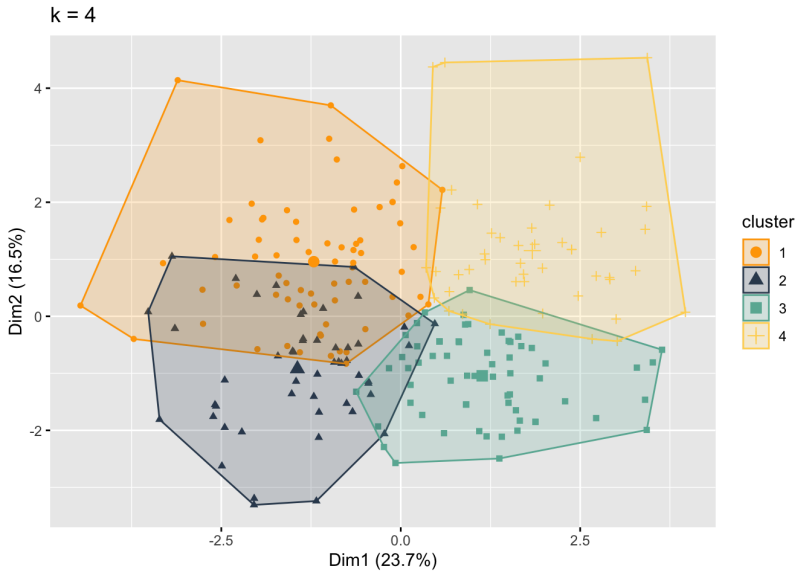

```
set.seed(224)

# K-Means Cluster Analysis
fit <- kmeans(na.omit(NewData), 4, nstart = 25) #4 cluster solution
# get cluster means
aggregate(na.omit(NewData),by=list(fit$cluster),FUN=mean)
```

```
  Group.1 AmountWeek AmountOutMonth MoneyCoffee MoneyGroceries KnowledgeCoffee
1       1 -0.1312931      0.1967348   0.1929305    0.001207041       0.4140374
2       2 -0.4704610     -0.1011181  -0.3398370    0.104713747      -0.2201409
3       3 -0.1603911     -0.2691517  -0.5087628   -0.481006016      -0.3483145
4       4  0.9902463      0.2175337   0.8613835    0.603973363       0.1366689
  Purchase_Price Purchase_Sustainability Purchase_Certificate
1     0.20123654               0.3719934            0.3303161
2     0.01224639               0.9668525            0.9289028
3     0.19926130              -0.5120427           -0.5396210
4    -0.62829196              -0.9221548           -0.7717966
  Purchase_Fairtrade Purchase_Packaging Subscription_Likely  App_Likely
1          0.4390226          0.4002786           1.0517009  0.90925026
2          0.9223396          0.5311954          -0.4183465 -0.47605807
3         -0.6118793         -0.4540626          -0.6099851 -0.54006201
4         -0.8241210         -0.5503423          -0.2308013 -0.04818404
  Machine_Aeropress Machine_CupMachine Machine_Espresso machine
1        0.01428571          0.2857143                0.4857143
2        0.00000000          0.1730769                0.1923077
3        0.00000000          0.3970588                0.1764706
4        0.00000000          0.4000000                0.4222222
  Machine_Filter machine Machine_French press Machine_Instant coffee
1             0.07142857           0.01428571             0.01428571
2             0.32692308           0.11538462             0.01923077
3             0.29411765           0.01470588             0.04411765
4             0.13333333           0.02222222             0.00000000
  Machine_Moka pot Machine_Percolator Machine_V60 BrandChange_Every time
1       0.05714286         0.00000000  0.05714286             0.04285714
2       0.15384615         0.01923077  0.00000000             0.00000000
3       0.07352941         0.00000000  0.00000000             0.00000000
4       0.02222222         0.00000000  0.00000000             0.00000000
  BrandChange_Never BrandChange_Sometimes BrandChange_Very often
```

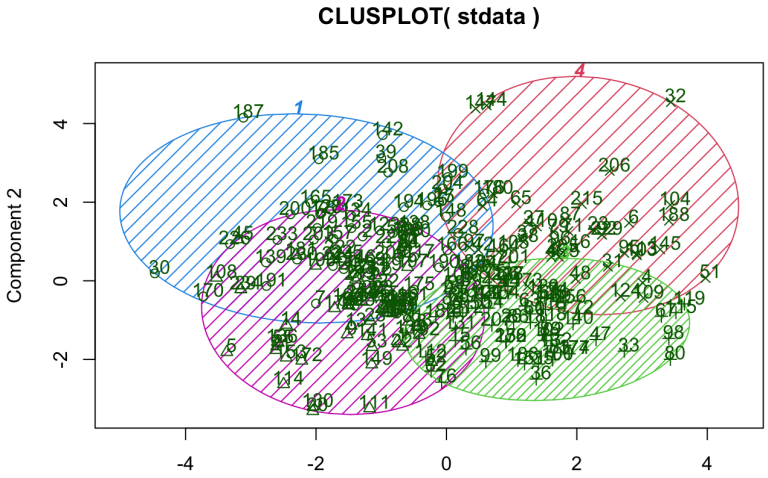
```
1         0.2285714             0.5857143             0.14285714
2         0.3269231             0.5000000             0.17307692
3         0.3970588             0.5735294             0.02941176
4         0.3777778             0.5777778             0.04444444
  PurchaseLocation_E-commerce PurchaseLocation_Online subscription
1                  0.17142857                            0.14285714
2                  0.15384615                            0.00000000
3                  0.07352941                            0.00000000
4                  0.33333333                            0.08888889
  PurchaseLocation_Specialty stores or cafés PurchaseLocation_The supermarket
1                                 0.24285714                         0.4428571
2                                 0.11538462                         0.7307692
3                                 0.04411765                         0.8823529
4                                 0.06666667                         0.5111111
  Education_Associate degree Education_Bachelor's degree
1                 0.07142857                    0.6142857
2                 0.07692308                    0.3846154
3                 0.10294118                    0.6764706
4                 0.06666667                    0.4222222
  Education_Elementary school Education_High school Education_Master
1                  0.00000000            0.10000000        0.2000000
2                  0.03846154            0.07692308        0.4038462
3                  0.01470588            0.07352941        0.1323529
4                  0.00000000            0.13333333        0.3333333
  Education_Phd AgeCategory_< 18 AgeCategory_> 60 AgeCategory_18-25
1    0.01428571       0.00000000       0.01428571        0.47142857
2    0.01923077       0.03846154       0.01923077        0.21153846
3    0.00000000       0.00000000       0.07352941        0.36764706
4    0.04444444       0.00000000       0.08888889        0.06666667
  AgeCategory_25-45 AgeCategory_45-60 Frequency_Specialty_Always
1         0.3857143         0.1285714                 0.27142857
2         0.5769231         0.1538462                 0.11538462
3         0.3382353         0.2205882                 0.01470588
4         0.4666667         0.3777778                 0.06666667
  Frequency_Specialty_I do (did) not know what this is
1                                           0.08571429
2                                           0.28846154
3                                           0.33823529
4                                           0.24444444
  Frequency_Specialty_Never Frequency_Specialty_Only in cafes
1                0.08571429                          0.1857143
2                0.13461538                          0.2307692
3                0.27941176                          0.1470588
4                0.20000000                          0.2666667
  Frequency_Specialty_Sometimes Home_Rural (Town) Home_Suburbs
1                     0.3714286        0.04285714   0.05714286
2                     0.2307692        0.11538462   0.05769231
3                     0.2205882        0.11764706   0.07352941
4                     0.2222222        0.15555556   0.13333333
  Home_Urban (City) Occupation_Employed (Full time)
1         0.9000000                        0.4428571
2         0.8269231                        0.6538462
3         0.8088235                        0.6029412
4         0.7111111                        0.6000000
  Occupation_Employed (Part time)
1                       0.2000000
2                       0.1346154
3                       0.1764706
4                       0.1777778
```

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

```
fviz_cluster(fit, geom = "point", data = stdata, outlier.color = "black", palette =
        dani) +
  ggtitle("k = 4")
```



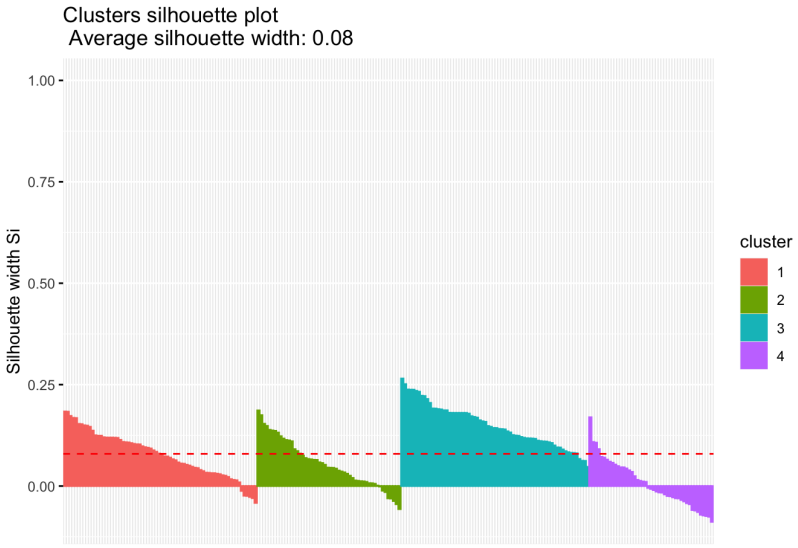

```
clusplot(stdata, fit$cluster, color=TRUE, shade=TRUE,
   labels=2, lines=0)
```




```
clustereddata <- cbind(data, Cluster = fit$cluster)

cluster1 <- subset(clustereddata, Cluster=='1')
cluster2 <- subset(clustereddata, Cluster=='2')
cluster3 <- subset(clustereddata, Cluster=='3')
cluster4 <- subset(clustereddata, Cluster=='4')
```

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

```
  cluster size ave.sil.width
1       1   70          0.08
2       2   52          0.05
3       3   68          0.15
4       4   45          0.01
```



The clusters individual results

```
Results <- as.data.table(aggregate(na.omit(Orgdata[,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	17	10	29	248	7
2	13	7	19	263	5
3	17	6	15	180	5
4	30	10	42	333	6

```
Results <- as.data.table(aggregate(na.omit(Orgdata[,6:10]),
        by=list(cluster=fit$cluster), mean), by = round)

Results_Round <- round(Results,1)

my_table(Results_Round)
```

cluster	Purchase_Price	Purchase_Sustainability	Purchase_Certificate	Purchase_Fairtrade	Purchase_Packaging
1	3.4	3.7	3.0	3.7	

2.9					
2	3.2	4.3	3.7	4.3	3.1
3	3.4	2.7	2.0	2.5	1.9
4	2.4	2.2	1.8	2.3	1.7

```
Results <- as.data.table(aggregate(na.omit(Orgdata[,11:12]),
        by=list(cluster=fit$cluster), mean), by = round)

Results_Round <- round(Results,1)

my_table(Results_Round)
```

cluster	Subscription_Likely	App_Likely
1	6.8	7.1
2	2.7	2.9
3	2.2	2.7
4	3.2	4.2

```
Results <- as.data.table(aggregate(na.omit(Orgdata[,13:20]),
        by=list(cluster=fit$cluster), mean), by = round)

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

cluster	Machine_Aeropress	Machine_CupMachine	Machine_Espresso machine	Machine_Filter machine	Machine_French press	Ma-

chine_Instant coffee	Machine_Moka pot	Machine_Percolator							
1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0

```
agetable1 <- as.data.table(table(cluster1$AgeCategory))
colnames(agetable1) <- c("Age", "Frequency")

agetable2 <- as.data.table(table(cluster2$AgeCategory) )
colnames(agetable2) <- c("Age", "Frequency")

agetable3 <- as.data.table(table(cluster3$AgeCategory) )
colnames(agetable3) <- c("Age", "Frequency")

agetable4 <- as.data.table(table(cluster4$AgeCategory) )
colnames(agetable4) <- c("Age", "Frequency")

my_table(agetable1)
```

Age	Frequency
> 60	1
18-25	33
25-45	27
45-60	9

```
my_table(agetable2)
```

Age	Frequency
< 18	2
> 60	1
18-25	11
25-45	30
45-60	8

```
my_table(agetable3)
```

Age	Frequency
> 60	5
18-25	25
25-45	23
45-60	15

```
my_table(agetable4)
```

Age	Frequency
> 60	4
18-25	3
25-45	21

45-60	17

```
table1 <- as.data.table(table(cluster1$Machine))
colnames(table1) <- c("Machine", "Frequency")

table2 <- as.data.table(table(cluster2$Machine) )
colnames(table2) <- c("Machine", "Frequency")

table3 <- as.data.table(table(cluster3$Machine) )
colnames(table3) <- c("Machine", "Frequency")

table4 <- as.data.table(table(cluster4$Machine) )
colnames(table4) <- c("Machine", "Frequency")

my_table(table1)
```

Machine	Frequency
Aeropress	1
CupMachine	20
Espresso machine	34
Filter machine	5
French press	1
Instant coffee	1
Moka pot	4
V60	4

```
my_table(table2)
```

Machine	Frequency
CupMachine	9
Espresso machine	10
Filter machine	17
French press	6
Instant coffee	1
Moka pot	8
Percolator	1

```
my_table(table3)
```

Machine	Frequency
CupMachine	27
Espresso machine	12
Filter machine	20
French press	1
Instant coffee	3
Moka pot	5

```
my_table(table4)
```

Machine	Frequency
CupMachine	18
Espresso machine	19
Filter machine	6
French press	1
Moka pot	1

```
table1 <- as.data.table(table(cluster1$PurchaseLocation))
colnames(table1) <- c("PurchaseLocation", "Frequency")

table2 <- as.data.table(table(cluster2$PurchaseLocation) )
colnames(table2) <- c("PurchaseLocation", "Frequency")

table3 <- as.data.table(table(cluster3$PurchaseLocation) )
colnames(table3) <- c("PurchaseLocation", "Frequency")

table4 <- as.data.table(table(cluster4$PurchaseLocation) )
colnames(table4) <- c("PurchaseLocation", "Frequency")

my_table(table1)
```

PurchaseLocation	Frequency
E-commerce	12
Online subscription	10
Specialty stores or cafés	17
The supermarket	31

```
my_table(table2)
```

PurchaseLocation	Frequency
E-commerce	8
Specialty stores or cafés	6
The supermarket	38

```
my_table(table3)
```

PurchaseLocation	Frequency
E-commerce	5
Specialty stores or cafés	3
The supermarket	60

```
my_table(table4)
```

PurchaseLocation	Frequency
E-commerce	15
Online subscription	4
Specialty stores or cafés	3
The supermarket	23

```
table1 <- as.data.table(table(cluster1$Frequency_Specialty))
colnames(table1) <- c("PurchaseLocation", "Frequency")

table2 <- as.data.table(table(cluster2$Frequency_Specialty) )
colnames(table2) <- c("Frequency_Specialty", "Frequency")

table3 <- as.data.table(table(cluster3$Frequency_Specialty) )
colnames(table3) <- c("Frequency_Specialty", "Frequency")

table4 <- as.data.table(table(cluster4$Frequency_Specialty) )
colnames(table4) <- c("Frequency_Specialty", "Frequency")

my_table(table1)
```

PurchaseLocation	Frequency
Always	19
I do (did) not know what this is	6
Never	6
Only in cafes	13
Sometimes	26

```
my_table(table2)
```

Frequency_Specialty	Frequency
Always	6
I do (did) not know what this is	15
Never	7
Only in cafes	12
Sometimes	12

```
my_table(table3)
```

Frequency_Specialty	Frequency
Always	1
I do (did) not know what this is	23
Never	19
Only in cafes	10
Sometimes	15

```
my_table(table4)
```

Frequency_Specialty	Frequency
Always	3
I do (did) not know what this is	11
Never	9
Only in cafes	12
Sometimes	10

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

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

cluster	Education_Phd	AgeCategory_< 18	AgeCategory_> 60	AgeCategory_18-25	AgeCategory_25-45	AgeCategory_45-60
1	0	0	0	0	0	0
2	0	0	0	0	1	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0

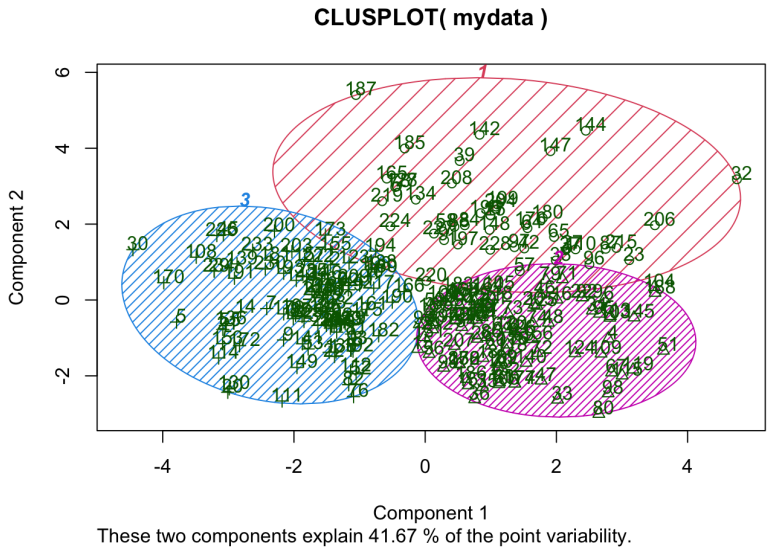
Clustering only the numerical

```
NumMydata <- na.omit(Orgdata[,c(1:12)]) # listwise deletion of missing
Mydata <- scale(Orgdata[,c(1:12)]) # standardize variables
```

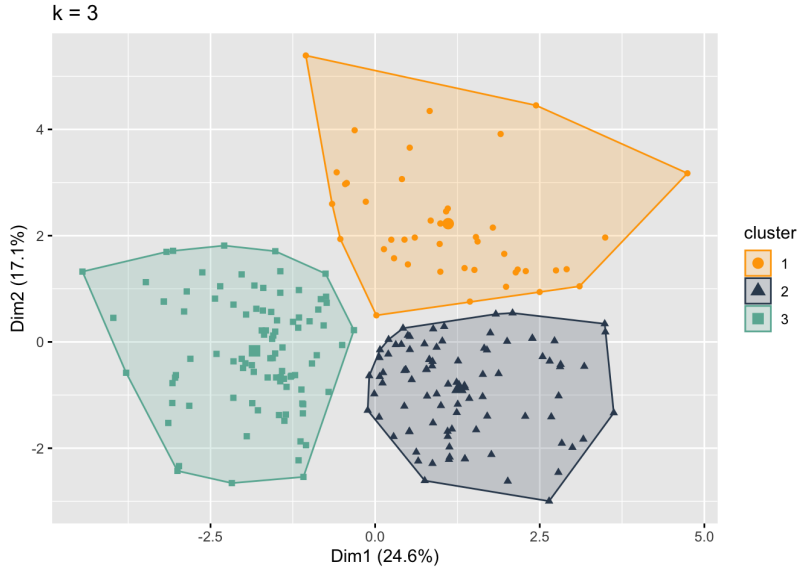
```
set.seed(123)

# 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  0.6839767      1.2416994   1.0664408     0.44953912      0.56191391
2       2  0.1411089     -0.2775268  -0.2129085    -0.15376891     -0.20359707
3       3 -0.4715468     -0.3109990  -0.2930917    -0.05816431     -0.06107242
  Purchase_Price Purchase_Sustainability Purchase_Certificate
1     -0.2800641             -0.03401391            0.1440235
2     -0.1886543             -0.70196986           -0.6181230
3      0.3267415              0.73318864            0.5623271
  Purchase_Fairtrade Purchase_Packaging Subscription_Likely App_Likely
1         0.06958705         -0.2104698           0.5480884  0.3929815
2        -0.74033587         -0.5117713          -0.4987982 -0.4137540
3         0.72277475          0.6234168           0.2470282  0.2344279
```

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

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

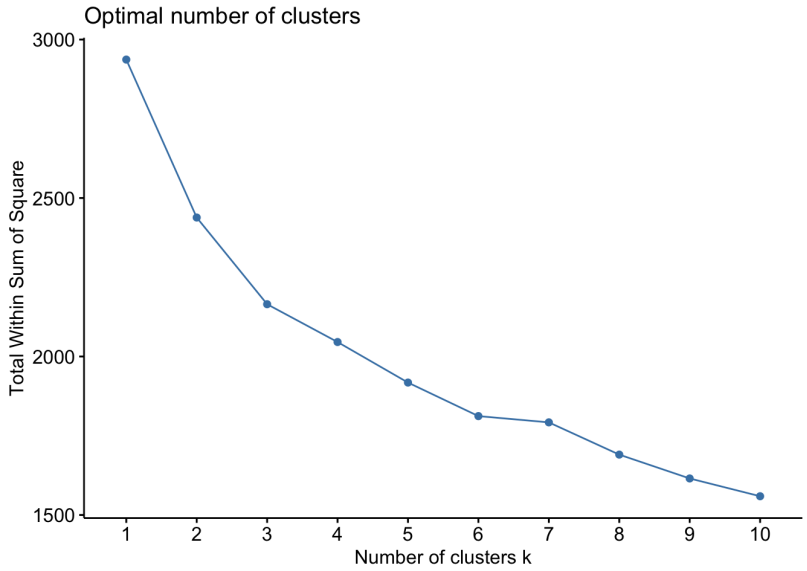



```
fviz_cluster(fit, geom = "point", data = mydata, outlier.color = "black", palette =
        dani) +
  ggtitle("k = 3")
```



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   45          0.03
2       2   96          0.20
3       3   94          0.18
```