ThesisSurveyCluster

07 March, 2021

https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

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)
```

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

Caret package

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

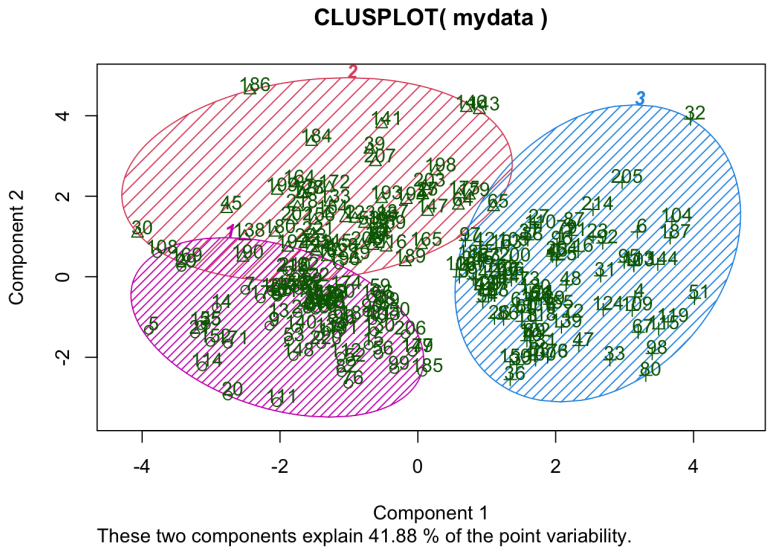
Clustering Numerical

```
# 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
```

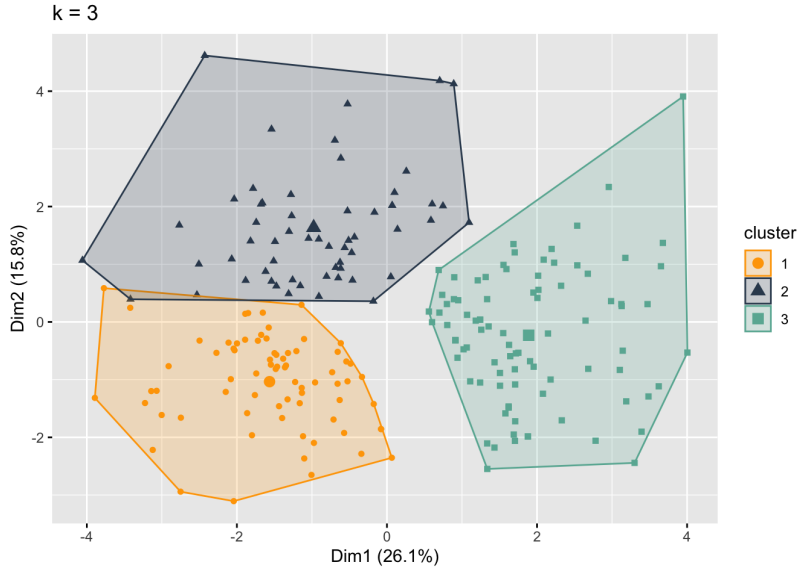
```
set.seed(123)

# 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)
```

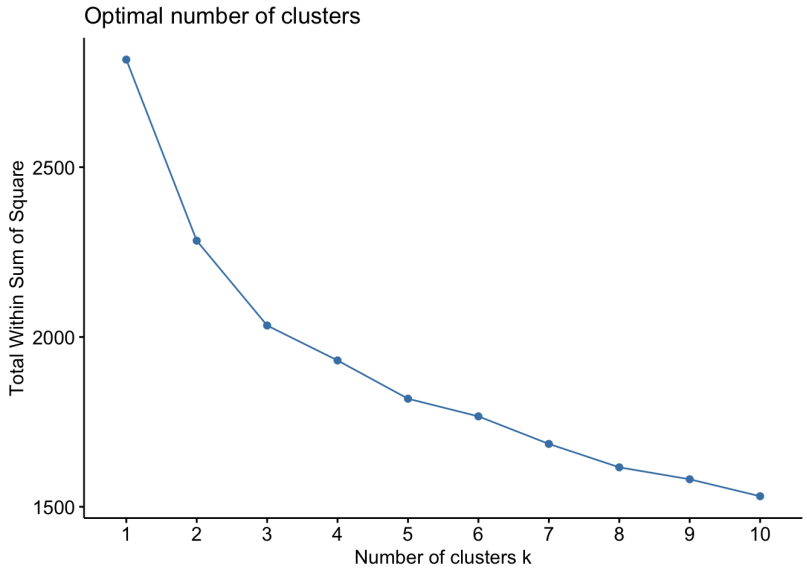



```
fviz_cluster(fit, geom = "point", data = mydata, outlier.color = "black", palette =
        dani) +
  ggtitle("k = 3")
```



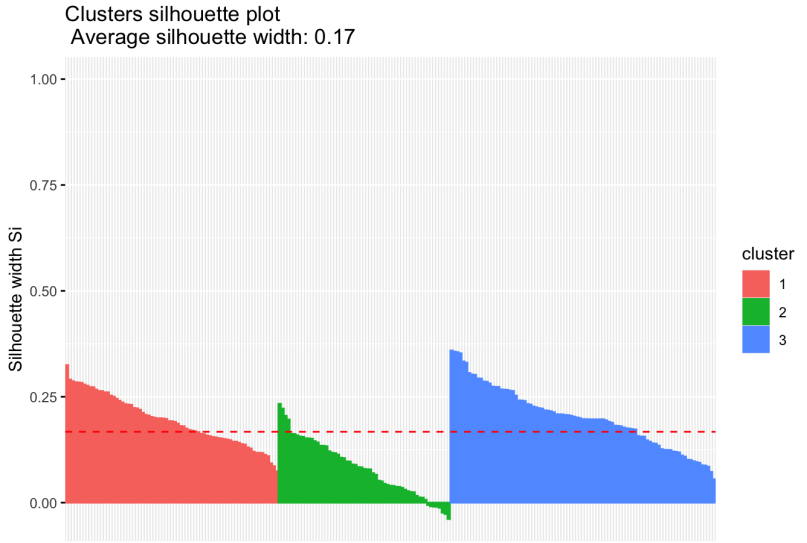
https://towardsdatascience.com/clustering-analysis-in-r-using-k-means-73eca4fb7967

```
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