King Saud University

College of Computer and Information Sciences

Information Technology department

IT 326: Data Mining

Course Project

McDonald menu

Project Report

Group #: 8 LAB Day-Time: Monday (8-10)

Group members:

Name	ID	Section
Nada Almutairi		

[Pick the date]

1 Problem

The availability of fast foods, fatty food, and unhealthy snacks in the living environment of children is assumed to give a very big reason of obesity among children.

In particular, it is very clear that food retailers are focusing on kid's using different strategies to dazzle them. We are using McDonald's menu to investigate :

(1) the clustering of food facts in the whole menu .

(2) the effects of particular food item in mac's big menu on children.

McDonald menu contains breakfasts and fast food like burgers and sandwiches and we would like to analyze it with respect to the content of food and quality. Also, we want to know all the good and bad sides of McDonald's most famous menu items such as Mc Nuggets, Big Mac and McChicken.

2 Data Mining Task

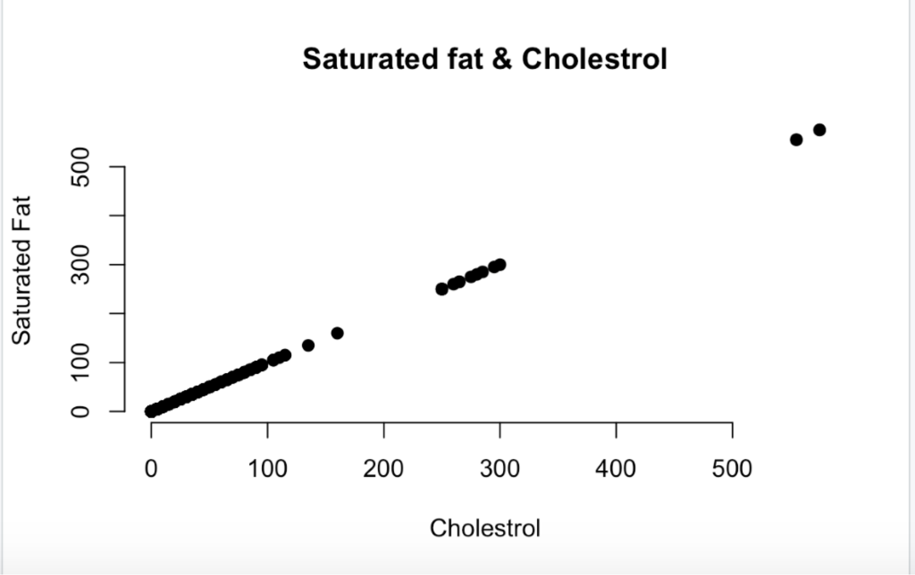
For this problem we will use clustering on food facts in the overall menu of US McDonald's to predict reasons of obesity among children.

3 Data

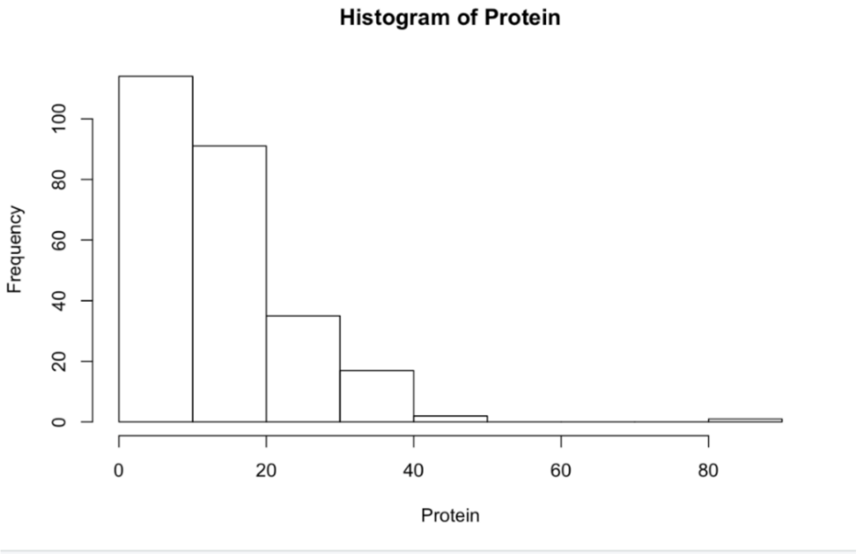
Dats set information	
This data set contains nutrition facts for mcDonalds' menu items	
Source	Kaggle https://www.kaggle.com/mcdonalds/nutrition-facts/data
Number of observations	260
Number of attributes	24
Missing values	No missing values

Attribute	**Attribute information**	**Data type**
Category	The types of meals	String
Item	Items of food sold from the McDonalds' US menu	String
Serving size	Recommended serving size for each item	String
Calories	The amount of calories in each item	Integer
Calories from fat	The amount of calories in each item that comes from fat	Integer
Total fat	The total fat for each item	Integer
Total fat (% daily value)	How much nutrition value in % from fat in an item	Integer
Saturated fat	How much saturated fat in an item	Integer
Saturated fat(% daily value)	How much nutrition value in % from saturated fat in an item	Integer
Trans fat	How much trans fat in an item	Integer
Cholesterol	How much choleterol in an item	Integer
Cholesterol (% daily value)	How much nutrition value in % from cholesterol in an item	Integer
Sodium	How much sodium in an item	Integer
Sodium (% daily value)	How much nutrition value in % from sodium in an item	Integer
Carbohydrates	How much carbohydrates in an item	Integer
Carbohydrates (% daily value)	How much nutrition value in % from carbohydrates in an item	Integer
Dietary fiber	How much dietary fiber in an item	Integer
Dietary fiber (% daily value)	How much nutrition value in % from dietary fiber in an item	Integer
Sugars	How much sugars in an item	Integer
Protein	How much protein in an item	Integer
Vitamin A (% daily value)	How much nutrition value in % from vitamin A in an item	Integer
Vitamin C (% daily value)	How much nutrition value in % from vitamin C in an item	Integer
Calcium (% daily value)	How much nutrition value in % from calcium in an item	Integer
Iron (% daily value)	How much nutrition value in % from iron in an item	Integer

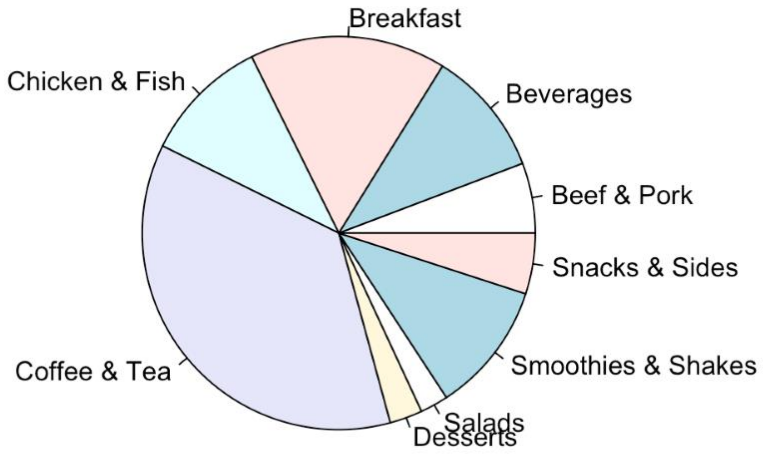
4 Data preprocessing



The graph shows that the more number you have of saturated fat in a meal the more the cholesterol in the product as increases , which also might increase the risk of getting heart disease and stroke.



This histogram made us find out that the data is skewed on the right-hand side having highest frequency of items with low protein while protein is an essential element to build muscles especially for children

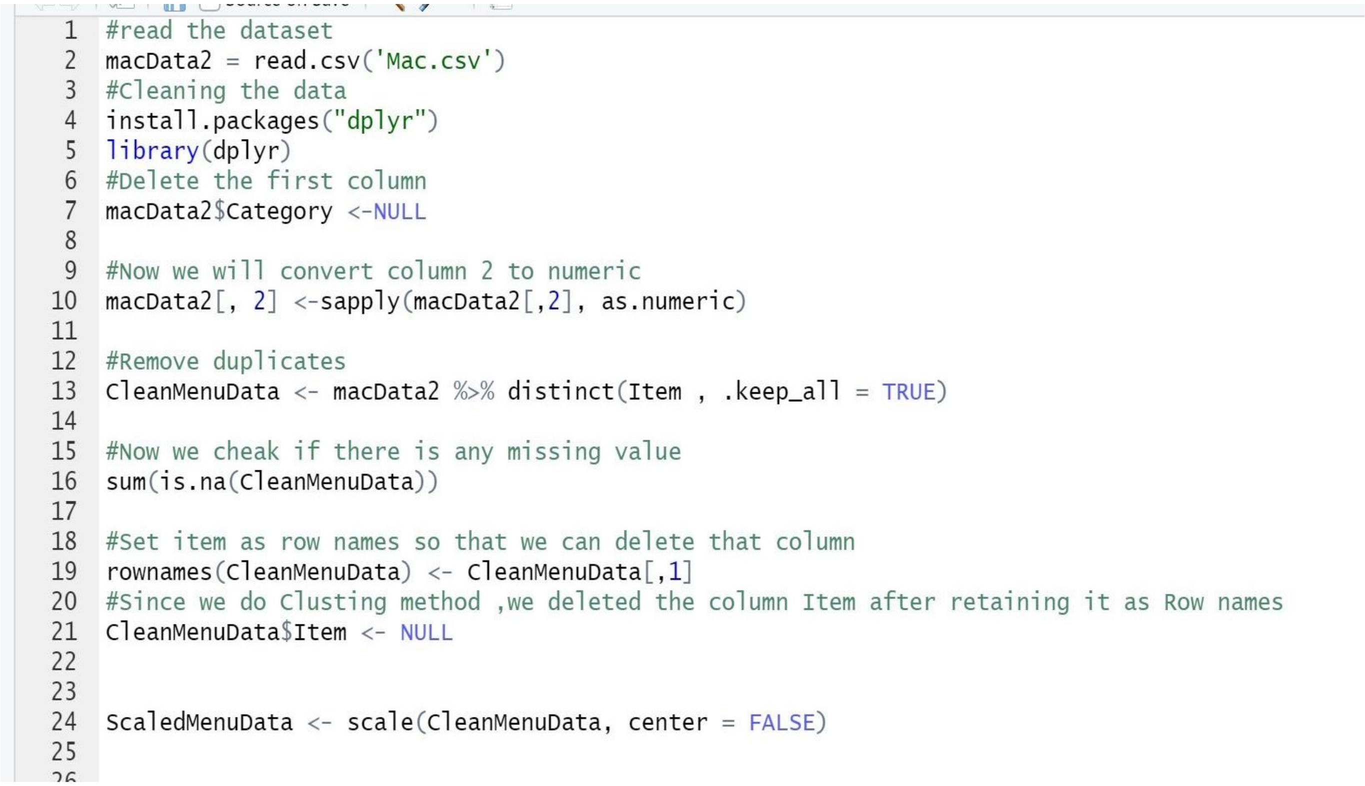


From this pie chart we can tell that the most category that its items varies the most is coffee&tea, which is not the children’s first choice anyway, and the salads has the least amount of menu items which is less options to choose from. and it is one of the healthiest categories.

To analysis our dataset, we had to do some of preprocessing.
so, we start of deleting the first column which is the category. Because it is just a meal information that we will not be needing it in our study.
Also, we converted the first two columns to numeric , Since we still have two more character attributes Items and serving size.
finally, we want to make sure if there were any missing value or not.

- Important note :

  we can not transform the column Item to numeric because we need the items' value but not the category itself ,so we transformed to row names to perform data mining task.

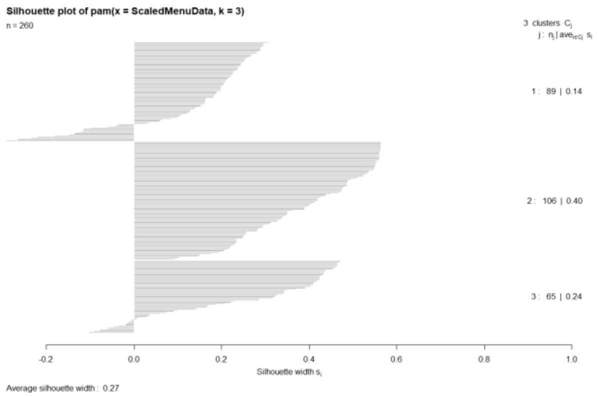
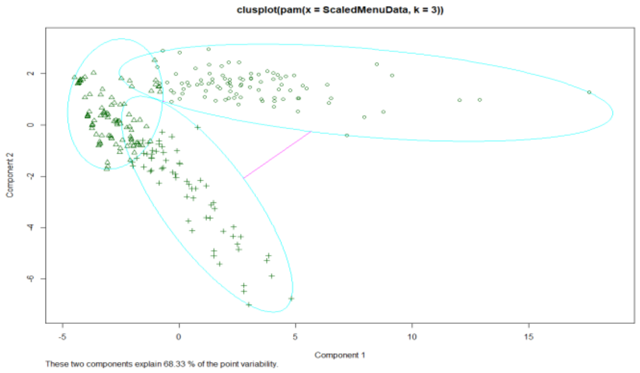
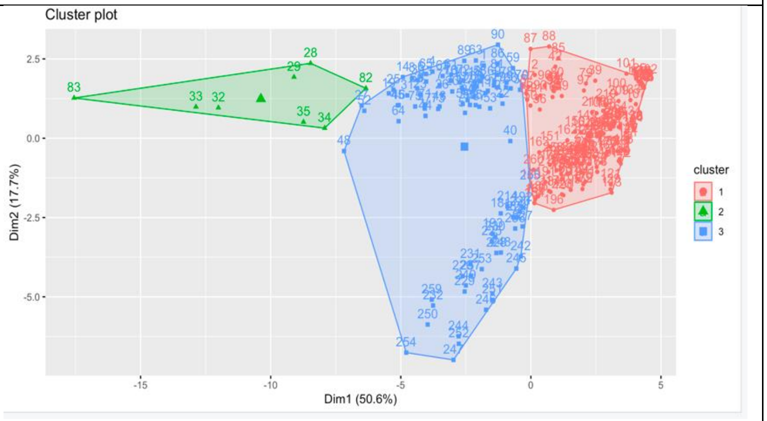


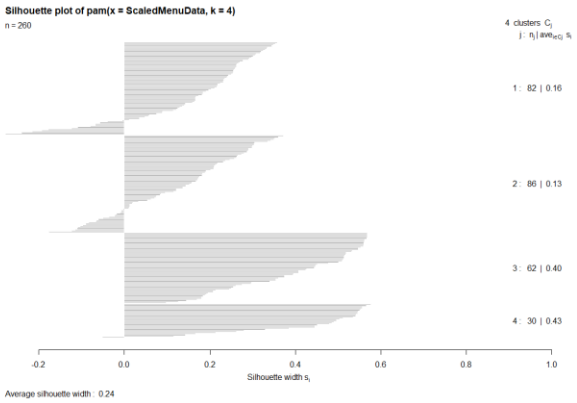
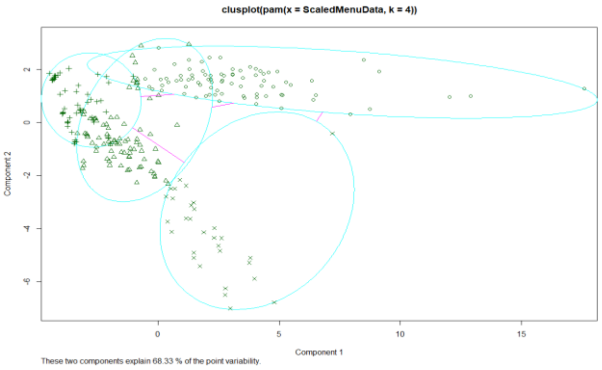
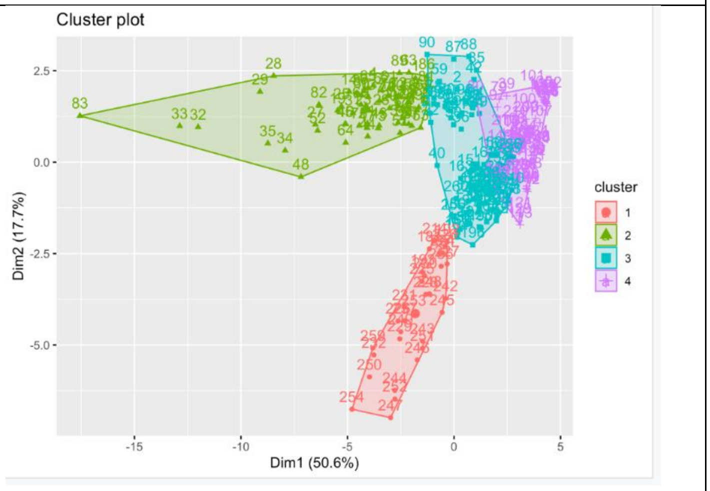
```
#read the dataset
macData2 = read.csv('Mac.csv')
#Cleaning the data
install.packages("dplyr")
library(dplyr)
#Delete the first column
macData2$Category <-NULL

#Now we will convert column 2 to numeric
macData2[, 2] <-sapply(macData2[,2], as.numeric)

#Remove duplicates
CleanMenuData <- macData2 %>% distinct(Item , .keep_all = TRUE)

#Now we cheak if there is any missing value
sum(is.na(CleanMenuData))

#Set item as row names so that we can delete that column
rownames(CleanMenuData) <- CleanMenuData[,1]
#Since we do Clusting method ,we deleted the column Item after retaining it as Row names
CleanMenuData$Item <- NULL


ScaledMenuData <- scale(CleanMenuData, center = FALSE)

```

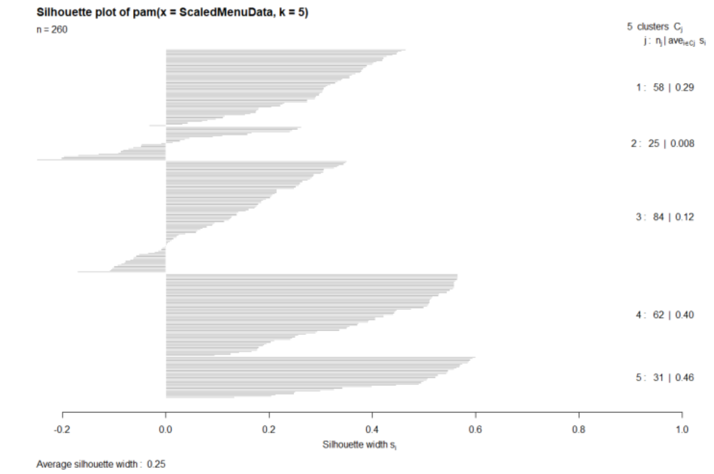
5 Data Mining Technique

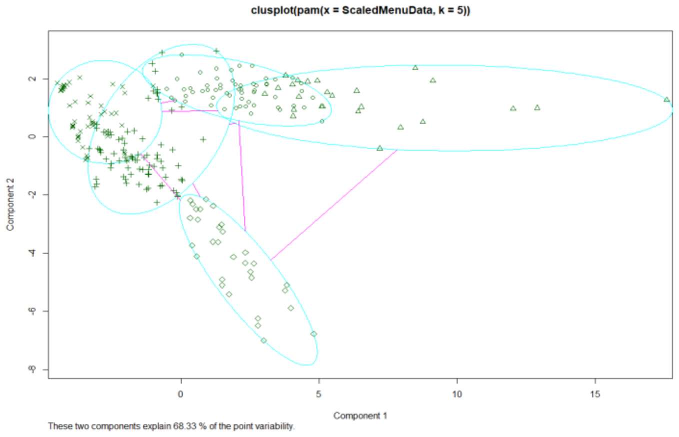
We used clustering technique which is used to find the structure of unlabeled data. Since that the type of data we have , and in order to discover groups of objects where the average distances between the members of each cluster shall be closer than to members in other clusters. All of that is for better understanding of the data by categorizing them.

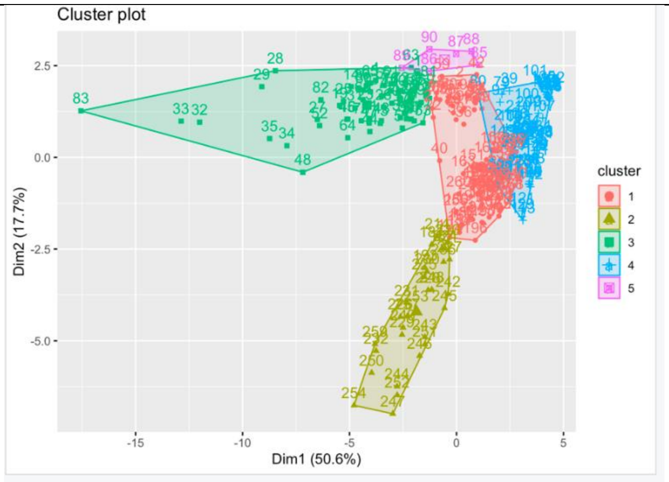
we used dplyr package for preprocessing , and factoextra , cluster , NbClust and ggplot2 for clustering .

	3 clusters
Silhouette width for 3 cluster	 
Visualization	 clusters are way close to each other which is not what we want, only cluster 2,3 have a good distance between them

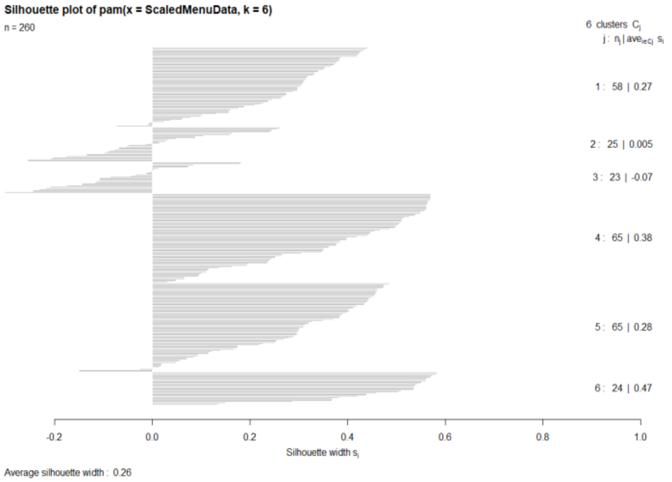
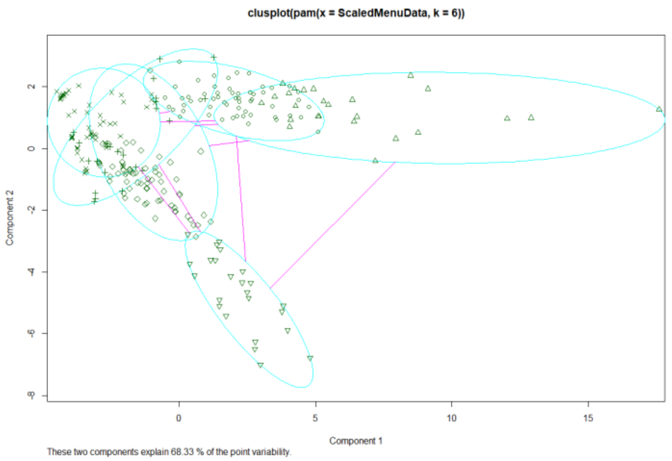
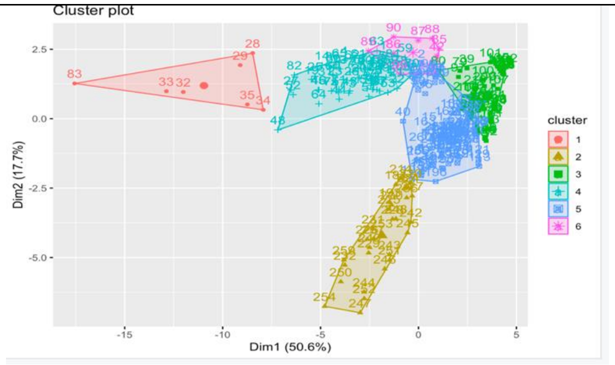
	4 clusters
Silhouette width for 4 cluster	 
Visualization	 we now can tell that the more the number of clusters increases the better the distance between clusters get

<table>
<tr><th></th><th>5 clusters</th></tr>
<tr><td><b>Silhouette width for 5 cluster</b></td><td>


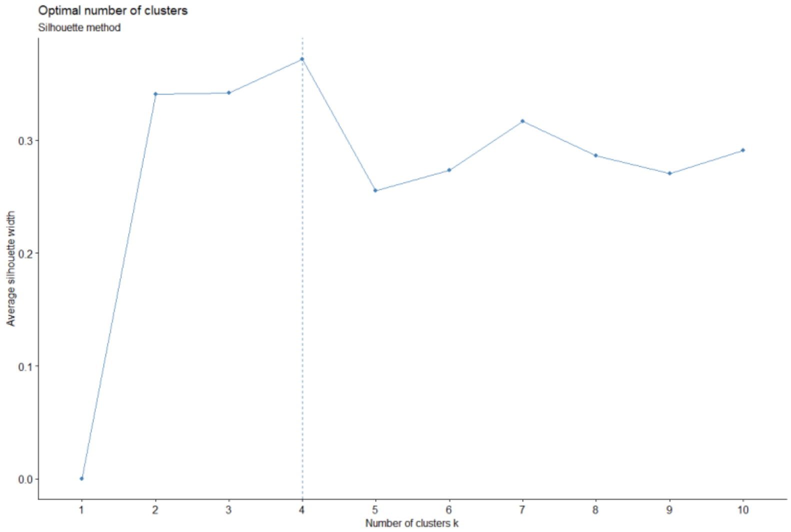


</td></tr>
<tr><td><b>Visualization</b></td><td>


clusters are getting more spaced out
</td></tr>
</table>

	6 clusters
Silhouette width for 6 cluster	 
Visualization	

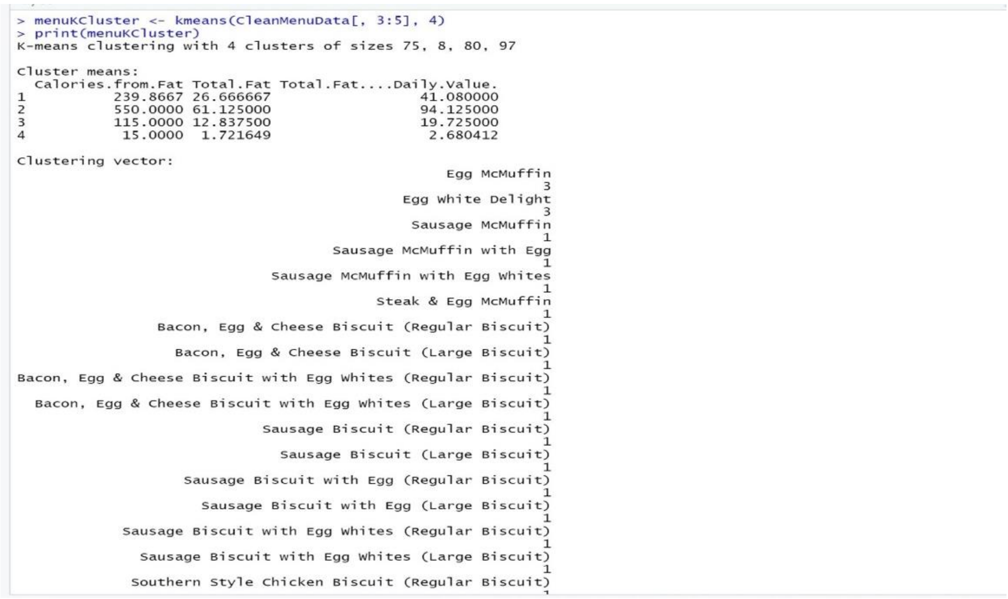
- **Silhouette width for all cluster**



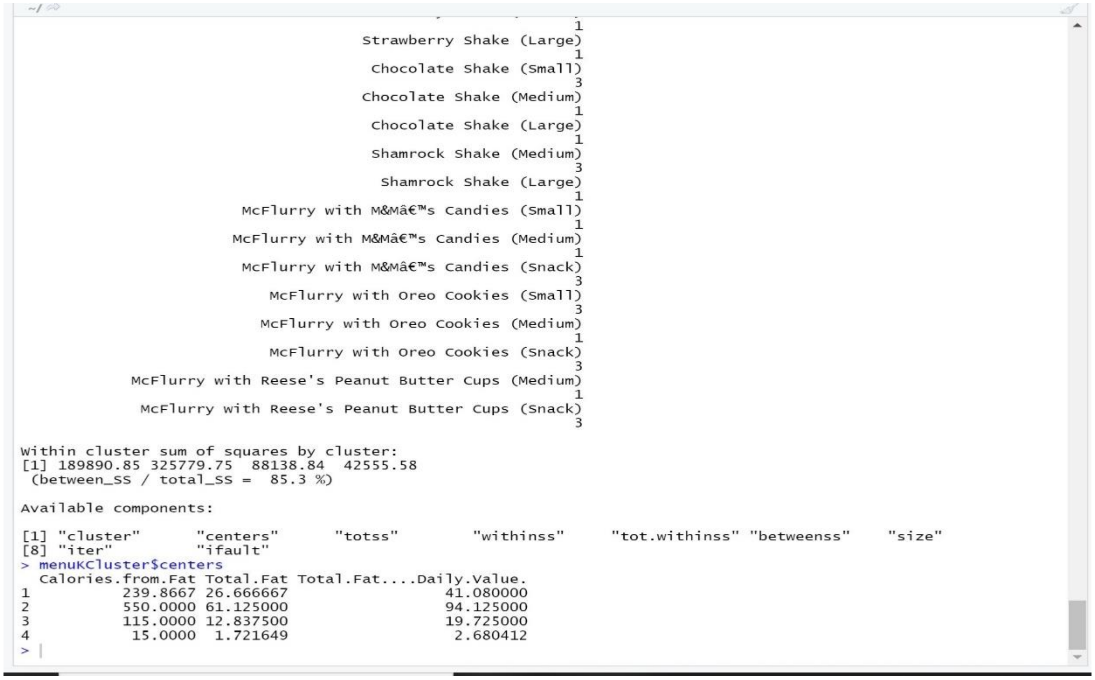
6 Findings

As we see the optimal number of cluster is 4. Now we will performs k-means clustering. Since the data has more than two attributes , we have to choose two attributes and studying them based on our problem.

So, our focus is mainly on the calories and total fat, which is the worst things that we can give to the children. And we will try avoid them.

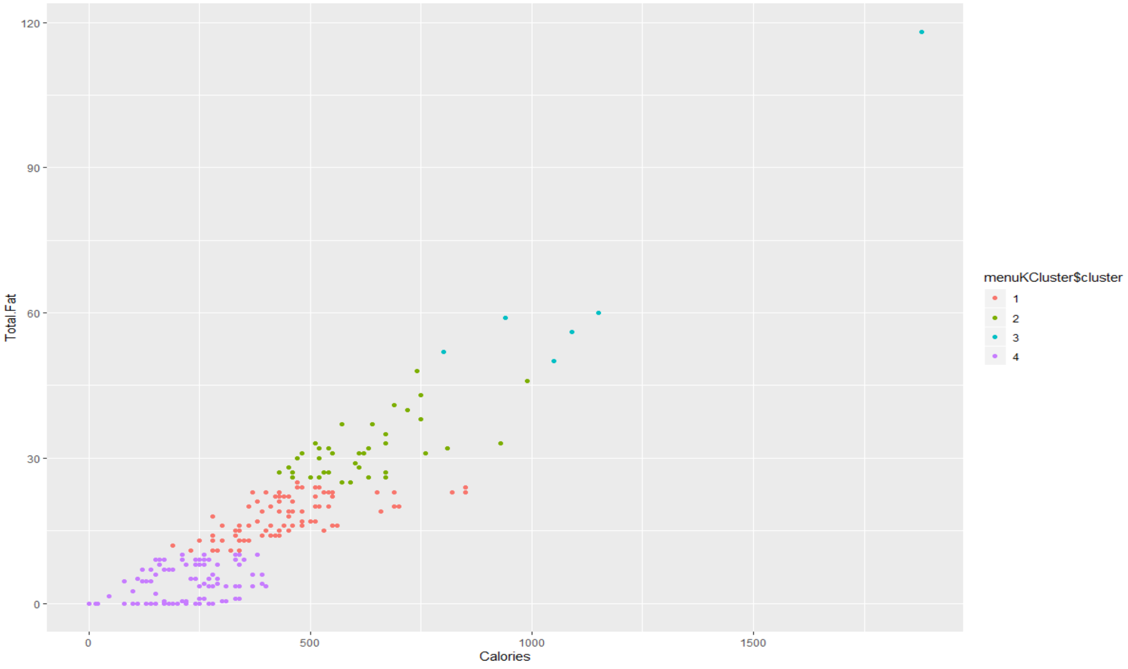


```
> menuKCluster <- kmeans(CleanMenuData[, 3:5], 4)
> print(menuKCluster)
K-means clustering with 4 clusters of sizes 75, 8, 80, 97

Cluster means:
  Calories.from.Fat Total.Fat Total.Fat....Daily.Value.
1          239.8667 26.666667                 41.080000
2          550.0000 61.125000                 94.125000
3          115.0000 12.837500                 19.725000
4           15.0000  1.721649                  2.680412

Clustering vector:
                                                 Egg McMuffin
                                                            3
                                            Egg White Delight
                                                            3
                                             Sausage McMuffin
                                                            1
                                    Sausage McMuffin with Egg
                                                            1
                             Sausage McMuffin with Egg Whites
                                                            1
                                         Steak & Egg McMuffin
                                                            1
                Bacon, Egg & Cheese Biscuit (Regular Biscuit)
                                                            1
                  Bacon, Egg & Cheese Biscuit (Large Biscuit)
                                                            1
Bacon, Egg & Cheese Biscuit with Egg Whites (Regular Biscuit)
                                                            1
  Bacon, Egg & Cheese Biscuit with Egg Whites (Large Biscuit)
                                                            1
                            Sausage Biscuit (Regular Biscuit)
                                                            1
                              Sausage Biscuit (Large Biscuit)
                                                            1
                   Sausage Biscuit with Egg (Regular Biscuit)
                                                            1
                     Sausage Biscuit with Egg (Large Biscuit)
                                                            1
            Sausage Biscuit with Egg Whites (Regular Biscuit)
                                                            1
              Sausage Biscuit with Egg Whites (Large Biscuit)
                                                            1
             Southern Style Chicken Biscuit (Regular Biscuit)
```



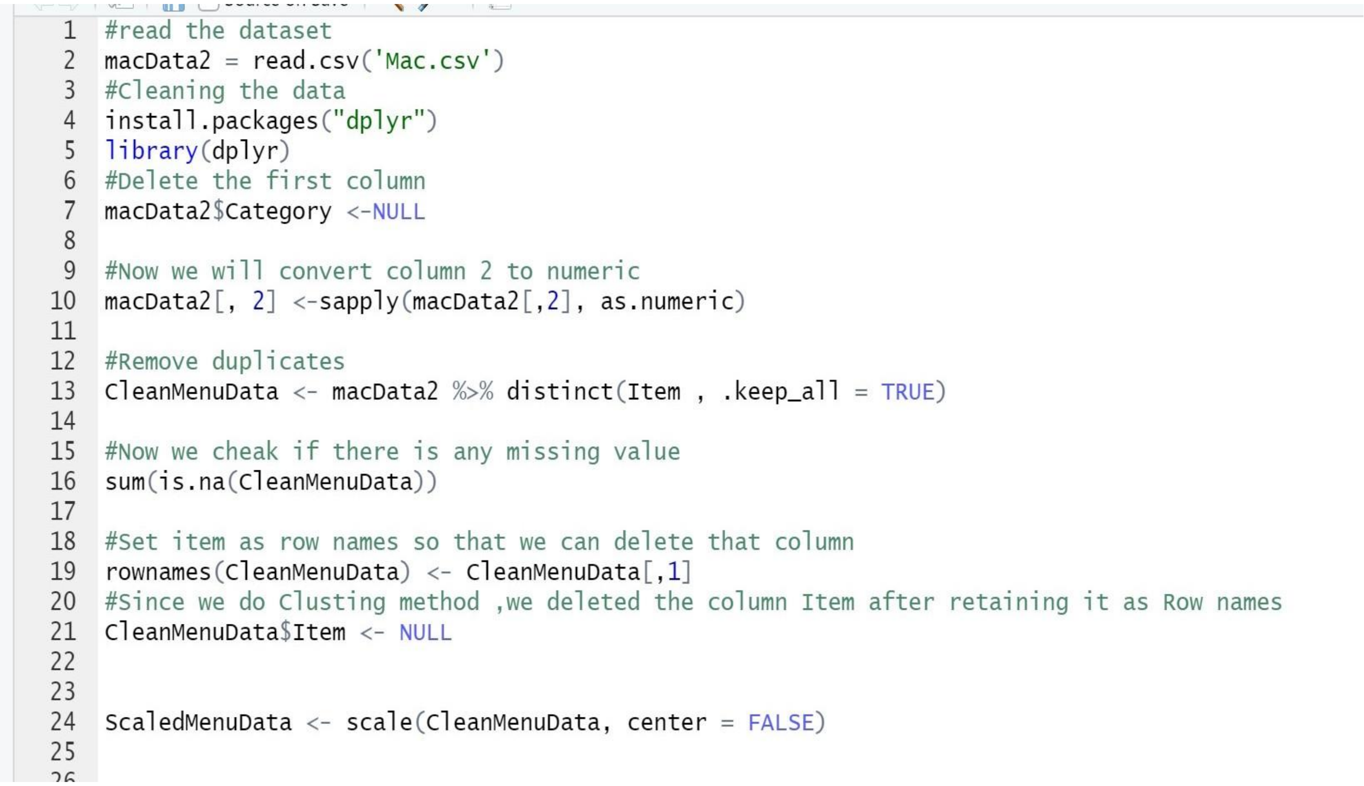
```
                                                            1
                                     Strawberry Shake (Large)
                                                            1
                                      Chocolate Shake (Small)
                                                            3
                                     Chocolate Shake (Medium)
                                                            1
                                      Chocolate Shake (Large)
                                                            1
                                      Shamrock Shake (Medium)
                                                            3
                                       Shamrock Shake (Large)
                                                            1
                        McFlurry with M&Mâ€™s Candies (Small)
                                                            1
                       McFlurry with M&Mâ€™s Candies (Medium)
                                                            1
                        McFlurry with M&Mâ€™s Candies (Snack)
                                                            3
                           McFlurry with Oreo Cookies (Small)
                                                            3
                          McFlurry with Oreo Cookies (Medium)
                                                            1
                           McFlurry with Oreo Cookies (Snack)
                                                            3
            McFlurry with Reese's Peanut Butter Cups (Medium)
                                                            1
             McFlurry with Reese's Peanut Butter Cups (Snack)
                                                            3

Within cluster sum of squares by cluster:
[1] 189890.85 325779.75  88138.84  42555.58
 (between_SS / total_SS =  85.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
[8] "iter"         "ifault"
> menuKCluster$centers
  Calories.from.Fat Total.Fat Total.Fat....Daily.Value.
1          239.8667 26.666667                 41.080000
2          550.0000 61.125000                 94.125000
3          115.0000 12.837500                 19.725000
4           15.0000  1.721649                  2.680412
> |
```

Then by plotting the points, each color represents the cluster.
As we can see there is a positive correlation between calories and total fat. As the calories and total fat increase, another cluster is formatted.



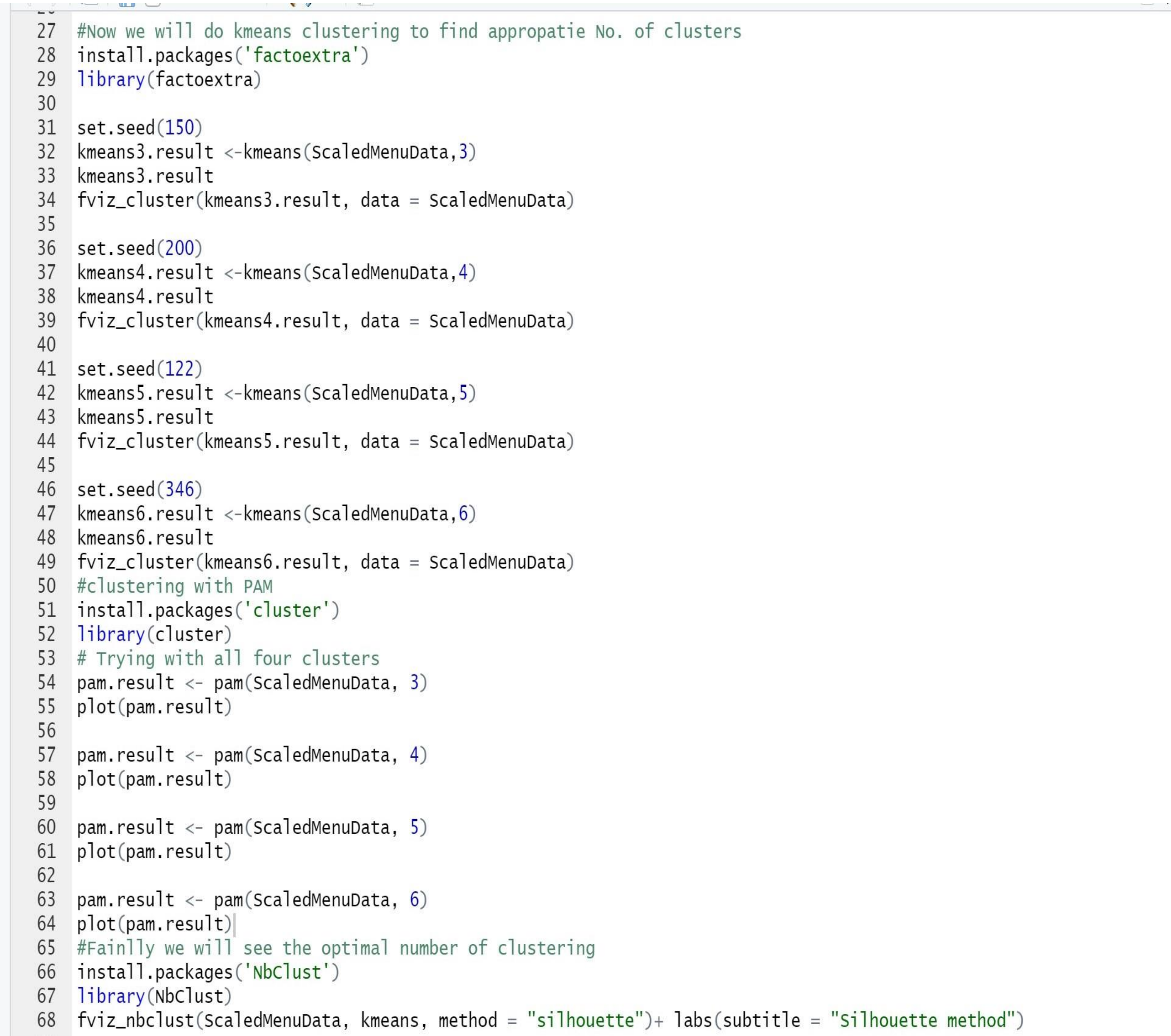
7 Code

Pre processing Code :

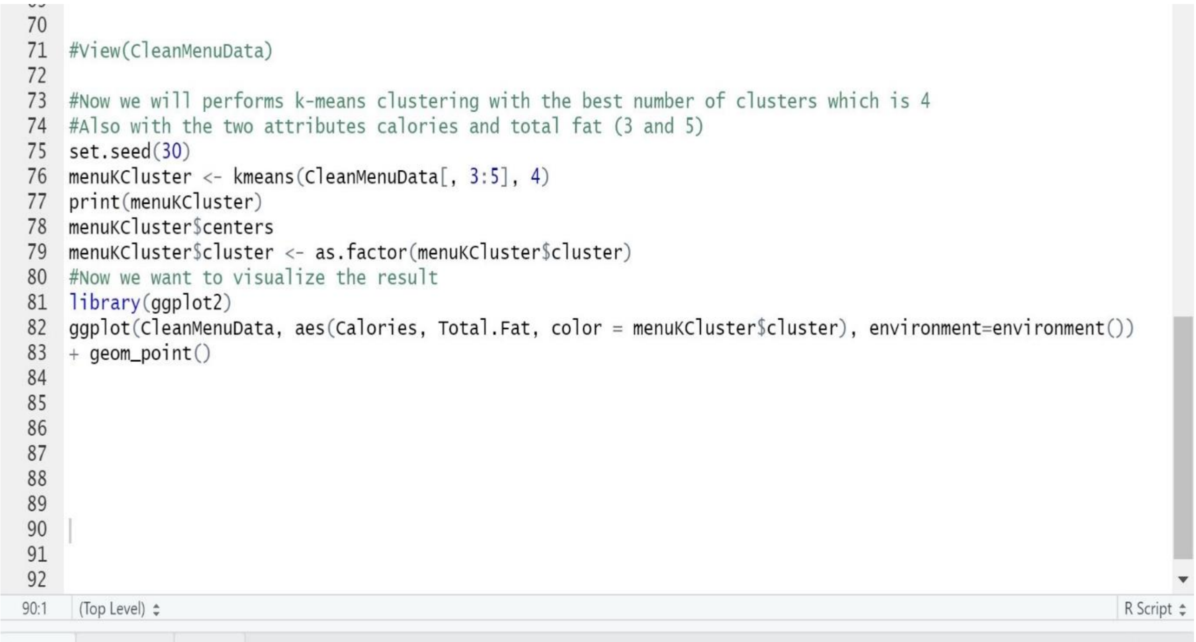


```r
#read the dataset
macData2 = read.csv('Mac.csv')
#Cleaning the data
install.packages("dplyr")
library(dplyr)
#Delete the first column
macData2$Category <-NULL

#Now we will convert column 2 to numeric
macData2[, 2] <-sapply(macData2[,2], as.numeric)

#Remove duplicates
CleanMenuData <- macData2 %>% distinct(Item , .keep_all = TRUE)

#Now we cheak if there is any missing value
sum(is.na(CleanMenuData))

#Set item as row names so that we can delete that column
rownames(CleanMenuData) <- CleanMenuData[,1]
#Since we do Clusting method ,we deleted the column Item after retaining it as Row names
CleanMenuData$Item <- NULL


ScaledMenuData <- scale(CleanMenuData, center = FALSE)

```

Clustering Code :



```
#Now we will do kmeans clustering to find appropatie No. of clusters
install.packages('factoextra')
library(factoextra)

set.seed(150)
kmeans3.result <-kmeans(ScaledMenuData,3)
kmeans3.result
fviz_cluster(kmeans3.result, data = ScaledMenuData)

set.seed(200)
kmeans4.result <-kmeans(ScaledMenuData,4)
kmeans4.result
fviz_cluster(kmeans4.result, data = ScaledMenuData)

set.seed(122)
kmeans5.result <-kmeans(ScaledMenuData,5)
kmeans5.result
fviz_cluster(kmeans5.result, data = ScaledMenuData)

set.seed(346)
kmeans6.result <-kmeans(ScaledMenuData,6)
kmeans6.result
fviz_cluster(kmeans6.result, data = ScaledMenuData)
#clustering with PAM
install.packages('cluster')
library(cluster)
# Trying with all four clusters
pam.result <- pam(ScaledMenuData, 3)
plot(pam.result)

pam.result <- pam(ScaledMenuData, 4)
plot(pam.result)

pam.result <- pam(ScaledMenuData, 5)
plot(pam.result)

pam.result <- pam(ScaledMenuData, 6)
plot(pam.result)
#Fainlly we will see the optimal number of clustering
install.packages('NbClust')
library(NbClust)
fviz_nbclust(ScaledMenuData, kmeans, method = "silhouette")+ labs(subtitle = "Silhouette method")
```



```
#View(CleanMenuData)

#Now we will performs k-means clustering with the best number of clusters which is 4
#Also with the two attributes calories and total fat (3 and 5)
set.seed(30)
menuKCluster <- kmeans(CleanMenuData[, 3:5], 4)
print(menuKCluster)
menuKCluster$centers
menuKCluster$cluster <- as.factor(menuKCluster$cluster)
#Now we want to visualize the result
library(ggplot2)
ggplot(CleanMenuData, aes(Calories, Total.Fat, color = menuKCluster$cluster), environment=environment())
+ geom_point()
```

8 References

E-books

[1] Practical guide to cluster analysis in R (Edition 1)

General Internet Site

[2] http://www.rdocumentation.org/

[3] https://www.datanovia.com/