McDonald’s Nutritional Facts

By Thomas Kwok

A study of McDonald’s Nutritional Facts using clustering to determine similarities between categories.

Section 1: Introduction:

McDonald's is one of the largest American hamburger and fast food restaurant chain in the world today, with approximately 38,000 restaurants. In this project, we analyze the nutritional facts of McDonald's data, trying to see if there are any patterns between the various items on the menu to create groupings different than the ones set by the "Category" section of the menu. Our question of interest is if we can make subgroups of McDonald items which combines several categories of the menu, to show that there are patterns between different categories in the McDonald's United States menu. This will be achieved with both clustering and component analysis.

In our data we see 260 items on the McDonald's menu and 24 variables. The data set provides a nutritional analysis of every menu item on the United States McDonald's menu, split into nine categories. The data is taken from Kaggle, using the nutritional facts listed on the McDonald's website in 2017. Listed below, is the table describing the variables and descriptions of them.

Table 1 – McDonald's variables and description of the variables.

Variables	Description of Variables
Category	Factor with 9 levels: e.g. "Beef & Pork", "Breakfast"
Item	Factor with 260 levels e.g. "Low Fat Milk", "McNuggets"
Serving Size	Factor with 107 levels e.g. "1 carton", "260 g"
Calories	Integer
Calories from Fat	Integer (Part of Calories)
Total Fat	Number
Total Fat Daily Value	Integer (out of 100%)
Saturated Fat	Number
Saturated Fat Daily Value	Integer (out of 100%)
Trans Fat	Number
Cholesterol	Integer
Cholesterol Daily Value	Integer (out of 100%)
Sodium	Integer
Sodium Daily Value	Integer (out of 100%)
Carbohydrates	Integer
Carbohydrates Daily Value	Integer (out of 100%)
Dietary Fiber	Integer
Dietary Fiber Daily Value	Integer (out of 100%)
Sugars	Integer
Protein	Integer
Vitamin A Daily Value	Integer (out of 100%)
Vitamin C Daily Value	Integer (out of 100%)
Calcium Daily Value	Integer (out of 100%)
Iron Daily Value	Integer (out of 100%)

Our main focus in this data set will be on the category variable, to see if we can create clusters between the nine categories and what nutritional facts go into determining these clusters. Looking at the number of items for each category, we notice that there are two initial groups we can break the data set into, McDonald's foods and McDonald's drinks.

Table 2 – Summary of McDonald's Categories

Beef & Pork	Beverages	Breakfast	Chicken & Fish	Coffee & Tea
15	27	42	27	95

Desserts	Salads	Smoothies & Shakes	Snacks & Sides
7	6	28	13

From the McDonald's food and drink data sets, we determine the maximal number of clusters using clustering and component analysis to figure out what groupings can come from the nutritional values and variables of the data set.

Section 2: Data

The data set we use is from the 2017 McDonald's United States menu. This data set has no missing values and one notable outlier, which will be shown in figure 2. Next, we looked at all of the variables listed and noticed that there are definitely heavy correlations between several, which was the first step we took in reducing the number of variables in our study. We noticed that several variables are split into total value and daily value, such as sodium or carbohydrate. We decided to keep the daily value variables instead of total value, which split our variables from 24 to 15. We also took out items and serving size, as we were not as interested in the actual item names or the serving size of the item. The variables used in our analysis are listed below.

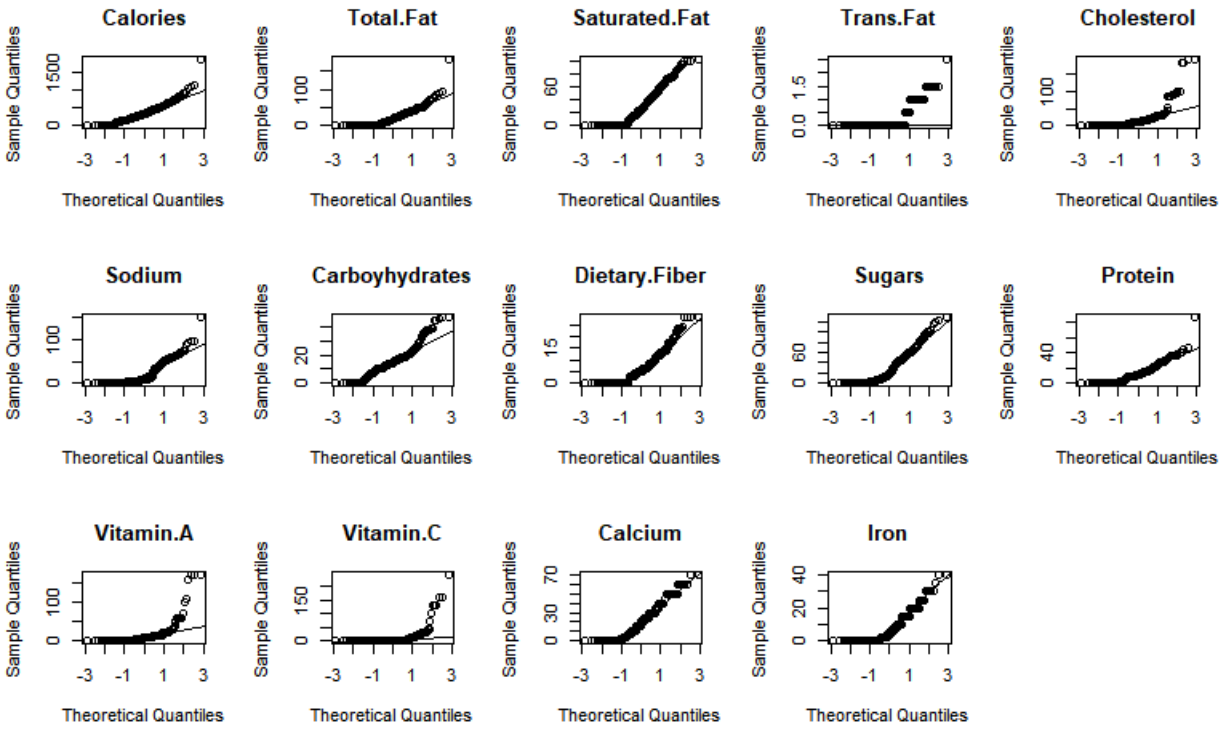
Table 3: Variables of interest from our data set

Category
Total Fat Daily Value
Saturated Fat Daily Value
Trans Fat
Cholesterol Daily Value
Sodium Daily Value
Carbohydrates Daily Value
Dietary Fiber Daily Value
Sugars
Protein
Vitamin A Daily Value
Vitamin C Daily Value
Calcium Daily Value
Iron Daily Value

Next, we decided to plot qqplots of our variables to test for normality for our entire data set (before split). We notice that most of the data is normally distributed for McDonald food items, although trans

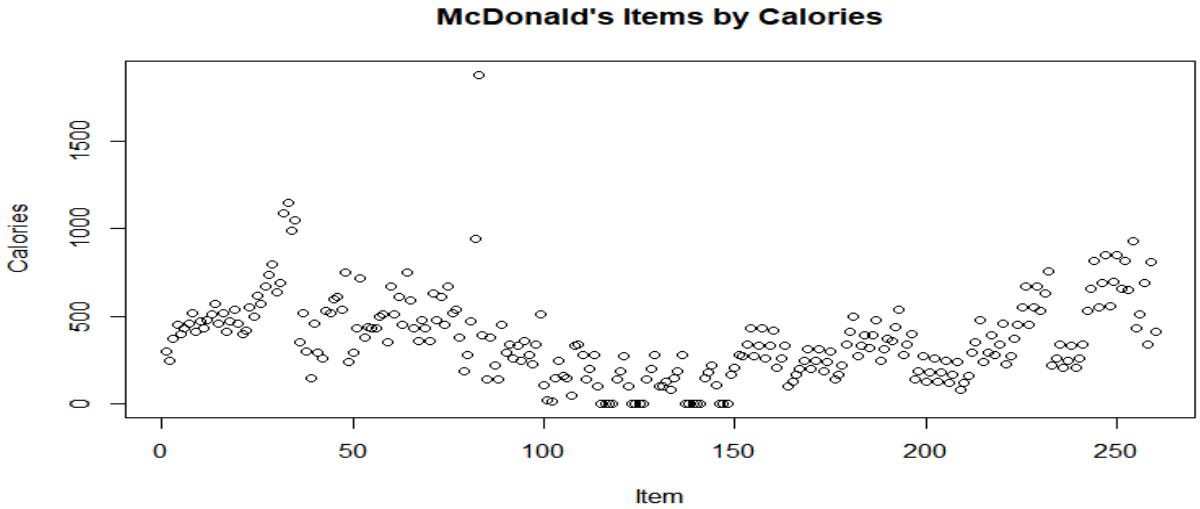
fat is one of note that does not. Also a few are heavily skewed, such as Cholesterol and Dietary Fiber Daily Value, Vitamin A Daily Value, and Vitamin C Daily Value.

Figure 1: QQPlot of variables in McDonald's Data Set



Trans fat is a hot topic in most nutritional value discussions, so the fact that it is not normally distributed makes sense as most items have no trans fat and others have almost no trans fat. Next, we decided to focus on McDonald's items by calories to see if we can find out patterns or trends, from the data we notice that most items stay around the same calories, but there are a few outliers such as one item with 1880 calories (40 piece Chicken McNuggets) and several items with no calories (Diet Soda, Water, Tea, and Coffee).

Figure 2: Calories Graph of McDonald Items



Section 3: Analysis

Since our research topic is to see if we can make subgroups of McDonald items which combines several categories of the menu, the best analysis would require finding variables that relate to one another. The two analytic strategies for this is Principal Component Analysis, to perform variable reduction, and clustering to find which observations can be grouped.

The first step we did for our data was to split it into two groups, McDonald's food and McDonald's drinks. The reason for this is that there are clear discrepancies between food and drink nutritional facts, so that was the easiest split. In doing so, we created a McDonald food data set with the categories: Beef & Pork, Breakfast, Chicken & Fish, Desserts, Salads, and Snacks & Sides. Then we created a McDonald drink data set with Beverages, Smoothies & Shakes, and Coffee & Tea. We also removed all items that are zero calories, as we felt that they had no nutritional value to our research. There was 134 items that are drink items, and 110 items that are food items.

For clustering, we decided to use k-means clustering which utilizes the Euclidean distances between items and creates a within-cluster variation based off of this. The standard algorithm for this formula is listed below.

$$W(Ck)=\sum(xi-\mu k)^2$$

The k-means analysis minimizes the distance of the observation to their assigned cluster center and we can define this total within-cluster variation as listed below.

$$tot.withiness=\sum W(Ck)=\sum\sum(xi-\mu k)^2$$

The reason why k-means clustering is effective is that it uses nearest means of the variables to create the k clusters. Since we are using nutritional facts of McDonald's items, the average of each nutritional variable is a good estimate to create clusters. In order to choose the best k for our data, we use kl, the maximal value of each index. We then compare this to a model-based clustering method from the library McLust to compare the clusters created and the nutritional data for the clusters.

Our other method of analysis is using principal component analysis to convert a set of observations with correlated variables into linearly uncorrelated variables called principal components. Since our data set involves all nutritional values, we know for sure that they would be correlated, and we use Principal Component Analysis to see what groups can be created from these variables. These two analyses are effective as clustering deals with finding groups of observations and PCA deals with finding groups of variables, so both combined can help us determine patterns for McDonald items and nutritional facts of the items.

Optimizing the number of k, using the NbCluster library, we found out that the best k using kl is 2 clusters for food is 3 clusters for drinks., after scaling the data as some are total number and others are percentages. Utilizing the McLust library we found the optimized number of k is 2 for food and

optimized number of k is 4 for drink in terms of maximizing the BIC value. This is interesting, especially for drink as there are only 3 categories, but the optimized number of clusters is 4.

Using Principal Component analysis, we found out that 85% of the variance can be explained by five principal components for McDonald's food and three principal components for McDonald's drinks explain approximately 82% of the variance.

Lastly, we decided to perform k-means and model based clustering with both cluster optimizations to see the pattern for each, and then analyze our principal component results to see which variables are significant for each data. The output for each will be produced in the next section.

Section 4: Tables and Graphs

Our first study was with McDonald's food items, which we used k-means clustering with the optimized cluster of 2. In our data, we notice that there are 88 food items that fit into the first cluster and only 22 items that fit into the second cluster, as shown in our table below.

Table: 4: Number of items per cluster for McDonald's food using k-means with k=2

Cluster 1	Cluster 2
88	22

We then looked at the number of items per category for each cluster and noticed that there are no desserts, salads, and snacks & sides for cluster 2 results.

Table 5: The number of items per Category for each cluster for McDonald's food using K-means

Cluster	Beef & Pork	Breakfast	Chicken & Fish	Desserts	Salads	Snacks & Sides
1	11	32	19	7	6	13
2	4	10	8	0	0	0

A breakdown of the cluster means show us the reason why there are two clusters, as it seems that cluster 2 items are much higher in calories, fats, cholesterol, etc. than cluster 1 items. This means that the clusters created are likely unhealthy food (cluster 1) and even more unhealthy food (cluster 2). The table of this data is shown below. One item that would fit into cluster 2 would be the 40 piece Chicken McNuggets, the 1880 calories outlier from our data set.

Table 6: Cluster means of each variable based off nutritional facts for McDonald's foods

Cluster	Calories	Total Fat	Sat. Fat	Trans Fat	Cholesterol	Sodium	Carbs	Fiber	Sugars	Protein	Vitamin A	Vitamin C	Calcium	Iron
1	374.43	27.17	31.89	0.14	22.98	33.88	12.34	9.94	8.48	16.94	18.02	13.12	14	13.04
2	812.73	67.95	67.82	0.55	66	72.09	22.82	16.27	9.45	36.14	15.45	10.68	20.78	24.55

Next, when we observe McDonald's drinks with the 3 clusters, we notice that there are 73 items in cluster 1, 45 items in cluster 2, and only 16 items in cluster 3. One thing to note here is that cluster 1

contains all of the beverages such as soft drinks. We also notice that cluster 3 is mostly made of up smoothies and shakes. Listed below is a table of the category breakdown for each cluster.

Table 7: The number of items per Category for each cluster for McDonald's drinks using K-means

Cluster	Beverages	Coffee & Tea	Smoothies & Shakes
1	18	49	6
2	0	35	10
3	0	4	12

Next, when we look at the actual values, we notice that cluster 3 is created because it has much higher calories and fats than the other two clusters. Part of this seems to be that cluster 3 has a much higher sugar content than the other two also. In addition to that, the difference between cluster 1 and cluster 2 seems to be that cluster 1 are items high on Vitamin C items while cluster 2 are items high on Vitamin A items. Also cluster 1 has no trans fat at all and cluster 2 seems to be highest on fiber.

Table 8 Cluster means of each variable based off nutritional facts for McDonald's drinks

Cluster	Calories	Total Fat	Sat. Fat	Trans Fat	Cholesterol	Sodium	Carbs	Fiber	Sugars	Protein	Vitamin A	Vitamin C	Calcium	Iron
1	200.68	5.10	9.86	0	4.42	3.66	12.78	2.40	34.59	5.37	6.98	10.27	17.04	0.90
2	409.11	18.04	34.78	0.27	12.67	7.82	21.49	5.22	57.33	11.93	15.11	1.76	38	4.04
3	741.25	37.63	75.50	1.03	25.94	10.69	38.44	3.44	99.44	15.19	23.44	0	50.63	4.69

If we use the model based modeling, we get a very different breakdown as opposed to using k-means clustering. In this data set, we see that desserts is evenly split between the two clusters and cluster 2 contains most of the snacks and sides compared to cluster 1. Also, breakfast and chicken and fish are both more evenly split between the two data sets. And there are 50 items in cluster 1 and 60 in cluster 2, so there is a much more even distribution when not accounting for means.

Table 9 The number of items per Category for each cluster for McDonald's food using McLust

Cluster	Beef & Pork	Breakfast	Chicken & Fish	Desserts	Salads	Snacks & Sides
1	13	25	11	3	6	2
2	2	17	16	4	0	11

The breakdown of the nutritional facts also seem much more evenly split in this case as opposed to the one using k-means, as shown below. When using McLust, it seems the clusters are much more similar than when using k-means.

Table 10: Cluster means of each variable based off nutritional facts for McDonald's foods

Cluster	Calories	Total Fat	Sat. Fat	Trans Fat	Cholesterol	Sodium	Carbs	Fiber	Sugars	Protein	Vitamin A	Vitamin C	Calcium	Iron
1	462.10	35.33	39.07	0.22	31.58	41.52	14.44	11.21	8.67	20.78	17.51	12.64	15.35	15.35
2	249.33	24.15	24.58	0.48	40.01	24.36	6.97	6.24	8.69	12.75	35.55	24.40	8.60	8.22

For McDonald’s drinks, there are four clusters created with 16 items in cluster 1, 61 items in cluster 2, 45 items in cluster 3, and 12 items in cluster 4. In this breakdown it seems that smoothies and shakes are predominantly seen in cluster 3, coffees and teas are predominantly seen in cluster 2, beverages are predominantly seen in cluster 1, and cluster 4 only contains coffees and teas. A table of this entire breakdown is shown below.

Table 11 The number of items per Category for each cluster for McDonald’s drinks using McLust

Cluster	Beverages	Coffee & Tea	Smoothies & Shakes
1	12	4	0
2	2	56	3
3	4	16	25
4	0	12	0

There does not seem to be a way to pull the variable breakdown from all four clusters in McLust so there is no more that can be said about this clustering for nutritional value wise, so without there there is little analysis that can be done for this clustering outside of figuring out the categories alone.

Using Principal Component Analysis, we get a breakdown of the cumulative proportion of variance for each components in the food and drink data set and then look at the variables per data set to make our conclusion. Using a type of biplot, we graph the most important variables for McDonald’s food and drink using PCA and find out that Calories, Total Fat, Saturated Fat, and Sodium are high contributing variables for both McDonald’s food and drinks.

Figure 3: Most important variables for McDonald’s food according to PCA

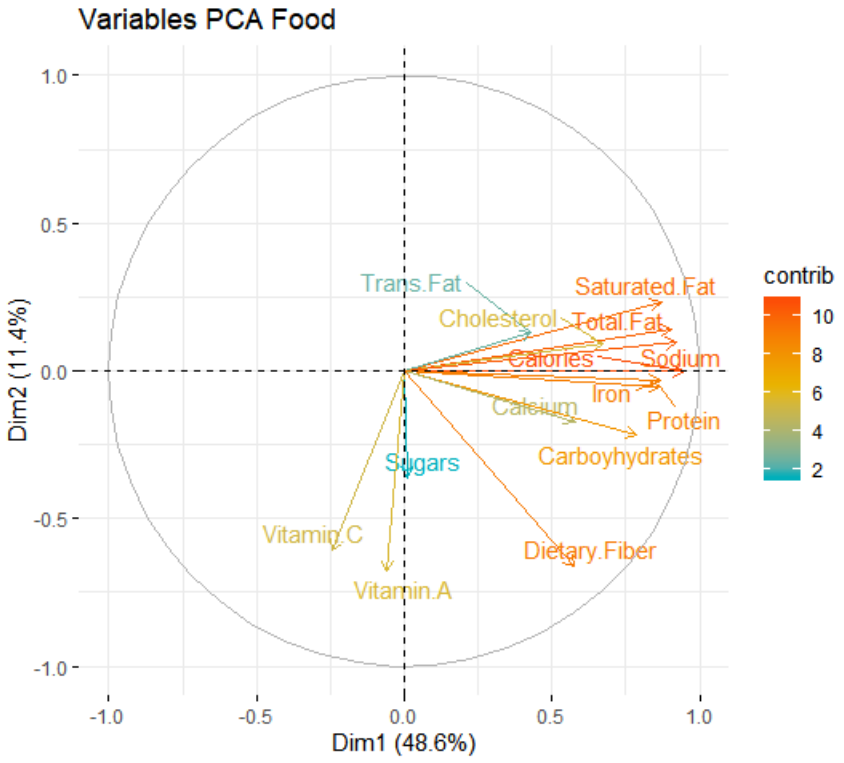
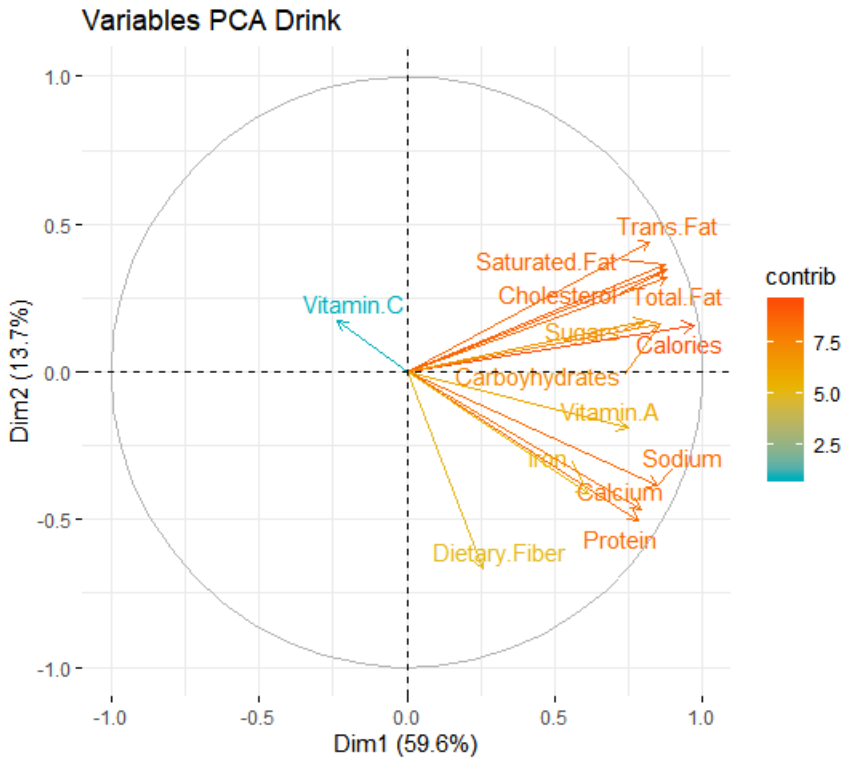


Figure 4: Most important variables for McDonald's drink according to PCA



Section 5: Conclusion

In conclusion, using clustering and PCA helped us make some interesting discoveries about the McDonald's food and drink menus. We learned from the cluster groups that there seems to be items with a lot of calories and items with less calories in each menu, which means that there are unhealthy fast foods in McDonald's and very unhealthy fast foods in McDonalds.

We also noticed from using PCA that there are definitely some variables in both cases that are not as important, such as sugars and trans fat for food and Vitamin C for drinks. We also noticed that Calories, Fat content, and sodium are very important variables for both data, which makes sense as these are important variables for fast food options in general.

This answers our question if there are patterns between the categories in the McDonald's menu and it seems there definitely are, as the clusters created have a combination of several category items instead of just one category item. It seems that the McDonald's menu is much more complex than simply what the nutritional facts say, as the categories shown can be grouped together, so it is possible to create different categories among the categories to group the items by nutritional facts.

In the future, one extra step to study after this is to figure a way to combine both PCA and cluster analysis together and come up with another analysis that uses both the variables and clusters to create different menus in McDonald's, especially if we can come up with healthy items, super unhealthy items, and regular fast food items. Another thing that can be done from this study is to look at customer satisfaction, either through ratings of food or restaurants and see if there is a connection between the

menu item clusters in general and rating, or nutritional facts and ratings in general. There are definitely other ways we can take this study in the future.

Section 6: Code

```
setwd("C:/Users/thoma/Desktop/Stat 717/Project/")
mcd <- read.csv("McDonalds_Menu.csv")

library(mclust)
library(lattice)
library(dplyr)
library(cluster)
library(ggplot2)
library(corrplot)
library(NbClust)
library(MASS)
library(class)
library(ISLR)
library(factoextra)

############################## Introduction ##########################
#data
str(mcd)
summary(mcd)
summary(mcd$Category)

############################## Data ###################################

#using daily value instead of total
mcd_menu <- mcd[,c(1,4,7,9,10,12,14,16,18,19,20,21,22,23,24)]
colnames(mcd_menu) <- c("Category", "Calories", "Total.Fat", "Saturated.Fat", "Trans.Fat",
"Cholesterol", "Sodium", "Carboyhydrates", "Dietary.Fiber", "Sugars", "Protein", "Vitamin.A",
"Vitamin.C", "Calcium", "Iron")
str(mcd_menu)

#split data into food and drink
mcd_food <- mcd_menu %>%
  filter(Category == "Beef & Pork" | Category == "Breakfast" | Category == "Chicken & Fish" |
Category == "Salads" | Category == "Snacks & Sides" | Category == "Desserts")
mcd_drink <- mcd_menu %>%
  filter(Category == "Beverages" | Category == "Coffee & Tea" | Category == "Smoothies & Shakes")
%>%
  filter(Calories != 0)

#split mcd_food and mcd_drink to numeric variables
mcd_menu <- mcd_menu[,-1]

#qqplot for normality
```

```
par(mfrow=c(3,5))
for (i in 1:14) {
  qqnorm(mcd_menu[,i],main=names(mcd_menu)[i])
  qqline(mcd_menu[,i])
}

#McDonalds Calories
par(mfrow=c(1,1))
plot(mcd_menu$Calories, main="McDonald's Items by Calories", xlab = "Item", ylab="Calories")

#correlation plot of data
corr <- cor(mcd_menu)
corrplot(corr, method="shade")

############################## Analysis ##################################

summary(mcd_food$Category)
summary(mcd_drink$Category)

#find the best k for mcd_food
bestK <- NbClust(scale(mcd_food[,-1]), min.nc=2, max.nc=5,index = "kl", method="kmeans")
bestK$Best.nc

bestK2 <- NbClust(scale(mcd_drink[,-1]), min.nc=2, max.nc=5,index = "kl", method="kmeans")
bestK2$Best.nc

#best k using mclust
mc1 <- Mclust(scale(mcd_food[,-1]))
plot(mc1, mcd_food[,-1], what = "BIC", col = "black")

mc2 <- Mclust(scale(mcd_drink[,-1]))
plot(mc2, mcd_drink[,-1], what = "BIC", col = "black")

#pca_food
pca_food <- prcomp(mcd_food[,-1], scale=TRUE)
summary(pca_food)

#pca_drink
pca_drink <- prcomp(mcd_drink[,-1], scale=TRUE)
summary(pca_drink)

############################## Tables & Graphs ##################################

#kmeans_food
set.seed(27)
mcd.kmeans_food <- kmeans(mcd_food[,-1],2)
cluster1 <- mcd.kmeans_food$cluster
table(cluster1)
xtabs(~cluster1+Category, data=mcd_food)
```

```
mcd.kmeans_food

#kmeans_drink
set.seed(27)
mcd.kmeans_drink <- kmeans(mcd_drink[,-1],3)
cluster2 <- mcd.kmeans_drink$cluster
table(cluster2)
xtabs(~cluster2+Category, data=mcd_drink)
mcd.kmeans_drink

#McLust Food
table(mc1$classification, mcd_food$Category)
mc1$data

#McLust Drink
table(mc2$classification, mcd_drink$Category)
mc2$data

#pca food
plot(pca_food$sdev^2, xlab = "Component Number", ylab = "Component Variance", type="l",
main="Covariance Scree diagram")
summary(pca_food)
fviz_pca_var(pca_food,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE,     # Avoid text overlapping
             title = "Variables PCA Food"
)

#pca drink
plot(pca_drink$sdev^2, xlab = "Component Number", ylab = "Component Variance", type="l",
main="Covariance Scree diagram")
summary(pca_drink)
fviz_pca_var(pca_drink,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE,     # Avoid text overlapping
             title = "Variables PCA Drink"
)
```