JANUARY 2, 2022

Data Science Project

Performed by: Group 10

TEAM MEMBERS

Names	*ID*	*Task*
Mark Ehab Farouk	**20221372981**	**Auditing**
Muaz Muhammed Sulaiman	**20221310296**	**Guidelines, Visualization, and Report**
Osama Elsayed Abdelsalam	**20221466241**	**Visualization**
Ramy Ahmed Eid	**20221445844**	**Report**
Ziad Mohamed Ahmed	**20221373065**	**Kmeans, and Apriori**

A) What will the program do?

- **The program will take a dataset path and use it to apply certain formulas and build insightful charts for deeper understanding of the data.**
- **Splitting the customers into groups (clusters).**
- **Generating association rules used to help for predictive analysis.**

B) What are the inputs?

The program will take the following inputs:

- **Dataset path is where the master data is located.**
- **Number of clusters will be used for the K-means, ranged between $2 \leq x \leq 4$**
- **Number of minimum support and minimum confidence will be used for the Apriori algorithm, ranged between $0.001 \leq x \leq 1$**

C) What the output from the program will be?

- **Visualization charts to describe the problems insightfully, through presenting the clustered customers based on the user input**
- **Association rules based on the user inputs of minimum support and minimum confidence by using the Apriori algorithm.**

We had categorized the project into the 6 phases of data analysis process: Ask, Prepare, Process, Analyze, Share, and Act.

Ask:

In this phase, we defined the problem to be solved and we made sure that we have fully understood stakeholder's expectations.

In our case are the data generators or imaginary stakeholders as the:

- Owner of the grocery.
- An executive buyer.
- OR any other in-charge who may help making decisions, influence actions and strategies, and have specific goals they want to meet.

Through a brainstorming process, we tried to define some problems to look at the current state (received from the dataset) and identify how it's different from the ideal state that we want to drive the business to.
3 obstacles we believe that it may help the grocery to improve the business revenue.

1. **Figuring out how many items are being sold in the grocery and the quantities sold in each period, to help the stakeholders to estimate the consumption forecasting by item and avoiding both stock shortage and unnecessary extra stock. This will be reflected on the cashflow financial management and will enhance the customer services as well.**

2. **Defining patterns of consumption (i.e., commonly conjoining items). This will help the grocery to arrange/present the items on the shelf to enhance the possibility of purchasing process. Also, it may support the proper placing of different offers advertised by the grocery.**

3. **Measuring the potentiality per category:**
    I. **Age group.**
    II. **City / Site / Branch.**
    III. **Payment Mode.**
    IV. **Customer name.**
    V. **Gender.**
    VI. **Etc.**

    This will help the marketing team to build a customized marketing plan of high efficiency and accurate approach.

Prepare:

Here we downloaded and stored the data received for the project (in CSV format) which we used for the upcoming analysis process. We made sure that the data and results are objective and unbiased (Unfortunately, the data were dummy). In other words, we wanted to confirm that any decisions will be made from our analysis should always be based on facts and be fair and impartial. In the beginning, we had used Excel to sort out the data and understand the context of fields within, through making some basic analysis using Pivot Table and some simple features as sort & filter.

```
#Downloading needed packages
install.packages("readr")
library(readr)
install.packages("dplyr")
library(dplyr)
install.packages("arules")
library(arules)
```

We also prepared our tools in R

- Readr: To read the CSV file
- Dplyr: To select and group data
- Arules: Used for the Apriori algorithm

```
#Importing Dataset to R
grc <- read_csv("grc.csv")
View(grc)
```

Importing the dataset to R

```
grc <- as_tibble(read.csv("grc.csv",stringsAsFactors = F))
grc_1 <- read.transactions("C:\\Users\\muazs\\Documents\\grc_1.txt", sep=",")
```

Preparing data for the Apriori algorithm

Process:

Here, we tried to find and eliminate any errors and inaccuracies that can get in the way of results. This usually means cleaning data, transforming it into a more useful format, combining two or more datasets to make information more complete and removing outliers, which are any data points that could skew the information.

```
#Removing column rnd for data cleaning
grc$rnd = NULL
```

We removed column rnd since it was meaningless.

97	cream cheese	24
10	cream cheese	366

146	roll products	15
141	roll products	86

```
#cream cheese
single_items$count[97] = 390
single_items = single_items[-c(10),]
```

```
#roll products
single_items$count[146] = 101
single_items = single_items[-c(141),]
```

After getting the unique item list and their count… We've found two duplicates "cream cheese" and "roll products". They were caused by a mistake from the data entry process. According to the "get occurrence" method I used; it doesn't remove the ending space in the words between commas but do remove extra spaces if it was the last item since there are no commas beyond. There were 24 "cream cheese" at the end of the item list which resulted to no extra space. While there were 366 "cream cheese" in between the comas which resulted to an extra space at the end. Same goes to "roll products". So, we removed the row with the extra space of the item list and edited the row with no extra space with the total item count.

Analyze:

Here we used tools to transform and organize that information so that we can draw some useful conclusions, make predictions, and drive informed decision-making. There are lots of powerful tools we used as excel and R, which are often.

```
#Grouping Ages and Sum by Total Spending for #2
ages = group_by(grc,age)
ages = summarise(ages,totalSpending=sum(total))
View(ages)
```

We created a new table with all ages and their total spending.

```
#Grouping Cities and Sum by Total Spending for #3
cities = group_by(grc,city)
cities = summarise(cities,totalSpending=sum(total))
cities = arrange(cities, desc(totalSpending))
View(cities)
```

We created a new table with all cities and their total spending by descending order.

```
#Number of clusters for Kmeans calculation
nclusters <- readline("Enter number of clusters")
if (nclusters>=2 & nclusters<=4){
  Kmeans<-kmeans(ages,center=nclusters)
  Kmeans
}else {print("Wrong input")}
```

We calculated the number of clusters using the R built in function. We also made data validation for the user input.

```
K-means clustering with 3 clusters of sizes 3, 5, 4

Cluster means:
    age totalSpending
1 38.00     1675852.7
2 40.40      816588.2
3 31.25      900931.2

Clustering vector:
 [1] 1 2 3 3 2 3 3 1 2 2 1 2
```

Assuming that the number of clusters we have are 3… The cluster points are {(38.00, 1675852.7), (40.40, 816588.2), (31.25, 900931.2)}
And the clustering vector is
{1, 2, 3, 3, 2, 3, 3, 1, 2, 2, 1, 2}
Based on the age group
{22, 23, 25, 29, 30, 35, 36, 39, 50, 55, 60}

Ages	22	23	25	29	30	35	36	37	39	50	55
Clusters	1	2	3	3	2	3	3	1	2	2	1

```
min_support<-as.numeric(readline("Enter minimum suport: "))
min_confidence<-as.numeric(readline("Enter minimum confidence: "))

if ((min_support>=0.001 & min_support<=1) &
    (min_confidence>=0.01 & min_confidence<=1)){
apriorirules <- apriori(grc_1,
parameter = list(supp = min_support, conf = min_confidence,minlen=2))
print(inspect(apriorirules))
}else {print("Wrong input")}
```

We made association rules between items based on the user minimum support and minimum confidence with data validation as well.

```
     lhs                    rhs                support    confidence coverage   lift
[1]  {hard cheese}       => {whole milk}       0.01006609 0.41078838 0.02450432 1.6076815
[2]  {whole milk}        => {hard cheese}      0.01006609 0.03939515 0.25551601 1.6076815
[3]  {butter milk}       => {other vegetables} 0.01037112 0.37090909 0.02796136 1.9169159
[4]  {other vegetables}  => {butter milk}      0.01037112 0.05359958 0.19349263 1.9169159
[5]  {butter milk}       => {whole milk}       0.01159126 0.41454545 0.02796136 1.6223854
[6]  {whole milk}        => {butter milk}      0.01159126 0.04536411 0.25551601 1.6223854
[7]  {ham}               => {whole milk}       0.01148958 0.44140625 0.02602949 1.7275091
[8]  {whole milk}        => {ham}              0.01148958 0.04496618 0.25551601 1.7275091
[9]  {sliced cheese}     => {whole milk}       0.01077783 0.43983402 0.02450432 1.7213560
[10] {whole milk}        => {sliced cheese}    0.01077783 0.04218066 0.25551601 1.7213560
```

Assuming that the minimum support is 0.01 and minimum confidence is 0.01… that is a sample of the first 10 association rules from the Apriori algorithm.

```r
customers = group_by(grc,customer)
customers = summarise(customers,totalSpending=sum(total))

gender = c("Male", "Male", "Female", "Female", "Female",
           "Female", "Male", "Male", "Male", "Female",
           "Male", "Male", "Male", "Female", "Female")

customers = cbind(customers, gender)
customers = arrange(customers, desc(totalSpending))
View(customers)
```

We created a new table with all customers by their total spending and their gender.

```r
genders = group_by(customers,gender)
genders = summarise(genders,totalSpending=sum(totalSpending))
View(genders)
```

We created a new table with genders and their total spending.

```r
transactions = strsplit(as.vector(items2$items), ',')

unique_items = unique(unlist(transactions))
```

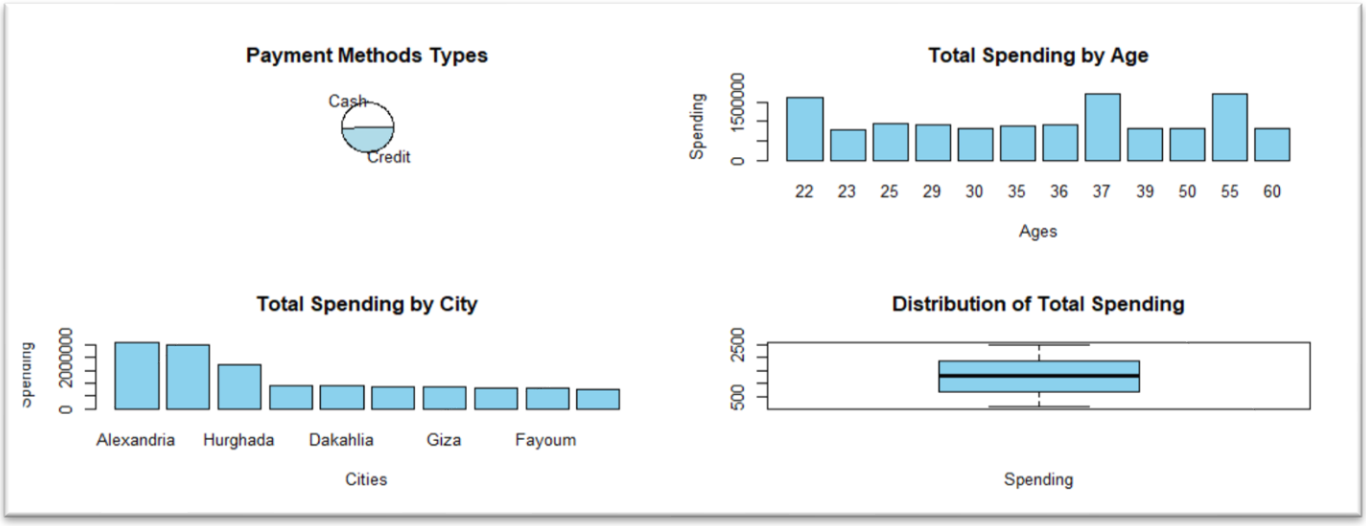
```r
single_items <- data.frame(item = unique_items, stringsAsFactors = FALSE )

single_items <- mutate(single_items,count = get_occurrences(item))
View(single_items)
```

We separated the items by “,” to get each unique item. The goal was to get the count of each item and add that new row to the item table.

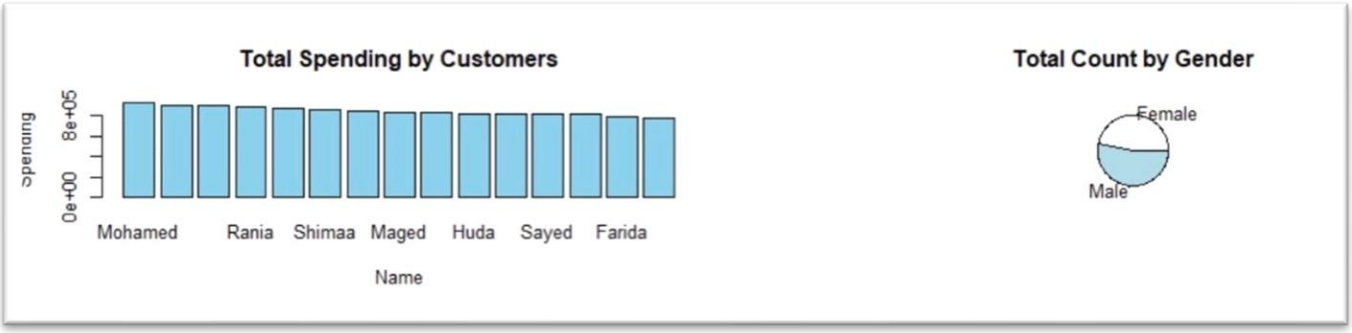
Share:

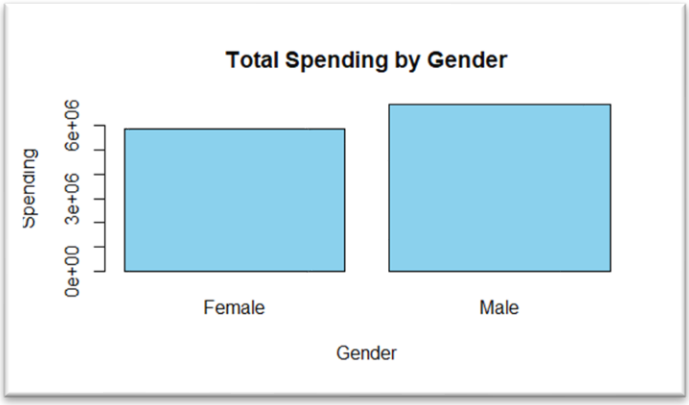
Here we had interpreted the results and share them with others to help stakeholders make effective data-driven decisions. In this phase, visualization was our best friend to understand what the data is telling us.



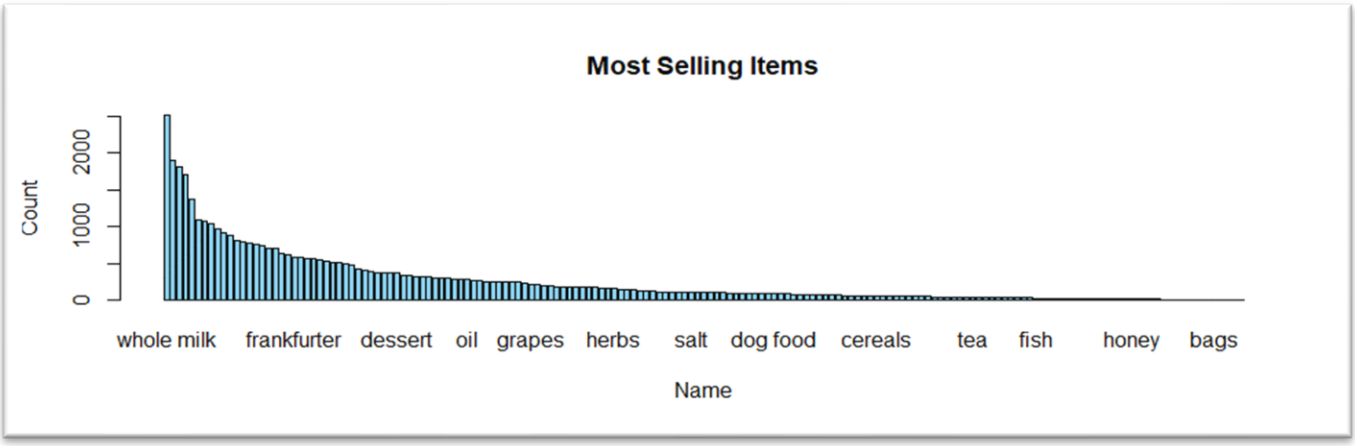
- **Cash spending are slightly more than Credit spending.**
- **Ages 22, 37, and 55 are the most spending.**
- **Alexandria, Cairo, and Hurghada are the most spending cities.**
- **Minimum of total spending is 100**
- **Maximum of total spending is 2500**
- **Median of total spending is approximately 1250**

Extra:





- **The most 3 spending customers are "Mohamed", "Magdy", and "Walaa"**
- **Male's count is slightly more than Female's**
- **Male's total spending is more than Female's**



- **We sorted out all items from "Z" to "A" as per repeated transactions (count)**
- **The most 3 selling items are "whole milk", "other vegetables", and "rolls/buns"**
- **The least 3 selling items are "baby food", "sound storage medium", and "preservation products"**

Act:

This is the exciting moment when the grocery's stakeholders took all the necessary insights our team have provided and puts them to work in order to solve the original business problem. We should be in on-going contact with the stakeholders and the updated dataset which will be received to monitor the impact of recommended actions and taking all necessary corrective actions -if needed-.

Notes:

.....

.....

.....

.....

.....

.....

.....

.....

.....