



Data Science Project

Group 10 (Mark, Muaz, Osama, Rami, & Ziad)

Delivering an Actionable Insight

Project Framework



Project Majors

What will the Project 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.

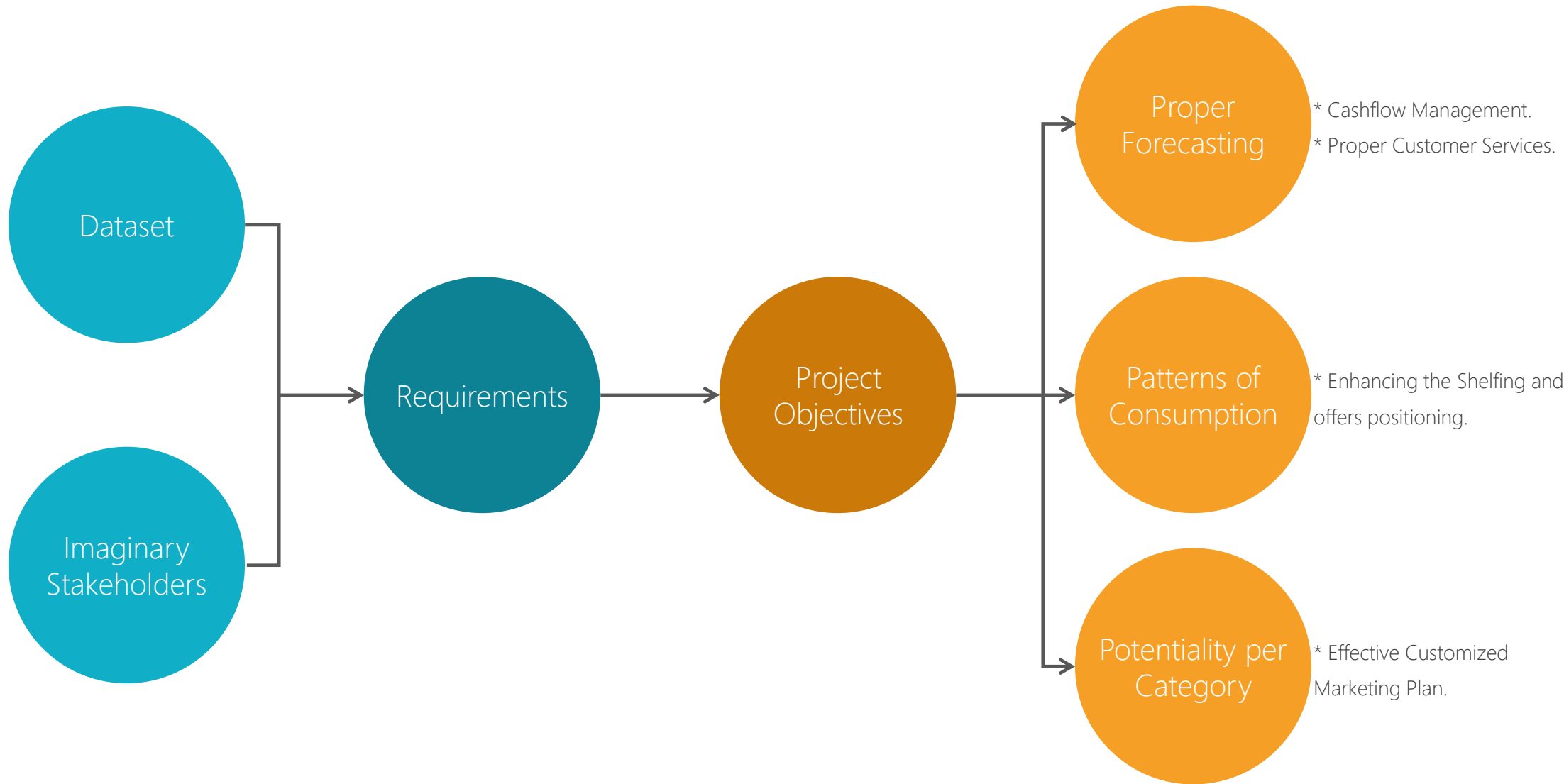
What are the Inputs?

- 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 $0.001 \leq x \leq 1$

What are the Outputs?

- 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.

Project Charter



Project Phases

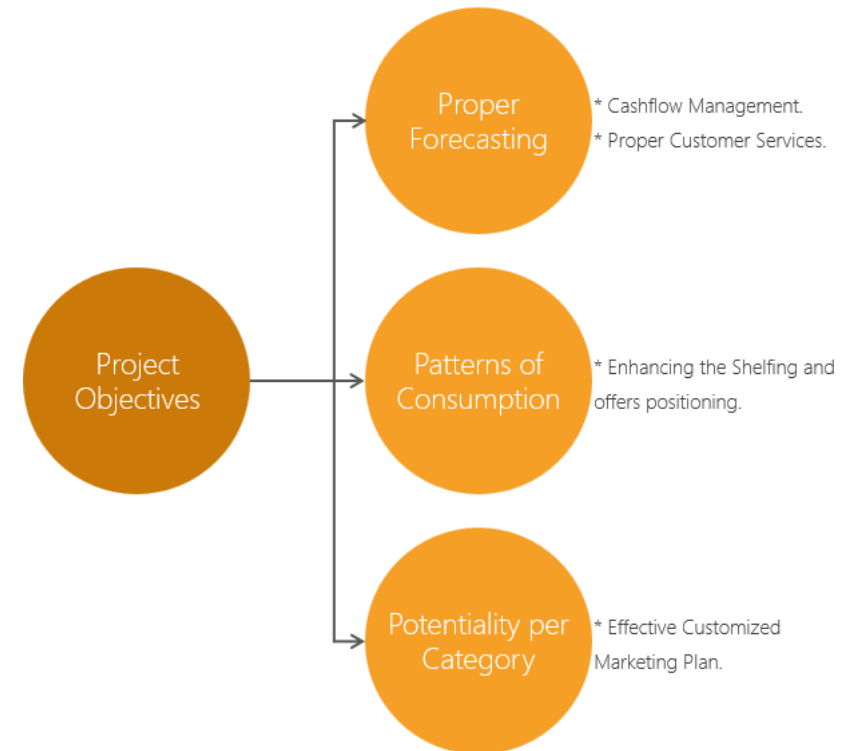


Ask

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.

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.



PREPARE

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
- Features as sort & filter

Importing the dataset to R

Preparing data for the Apriori algorithm

```
1 #Downloading needed packages
2 install.packages("readr")
3 library(readr)
4 install.packages("dplyr")
5 library(dplyr)
6 install.packages("arules")
7 library(arules)
```

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

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

PROCESS

We removed column `rnd` since it was meaningless.

- 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 commas 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.

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

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),]
```


ANALYZE 1

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

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

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

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}

```
16 #Grouping Ages and Sum by Total Spending for #2
17 ages = group_by(grc,age)
18 ages = summarise(ages,totalspending=sum(total))
19 View(ages)
```

```
21 #Grouping Cities and Sum by Total Spending for #3
22 cities = group_by(grc,city)
23 cities = summarise(cities,totalspending=sum(total))
24 cities = arrange(cities, desc(totalspending))
25 View(cities)
```

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

ANALYZE 2

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

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

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.

```
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")}
```

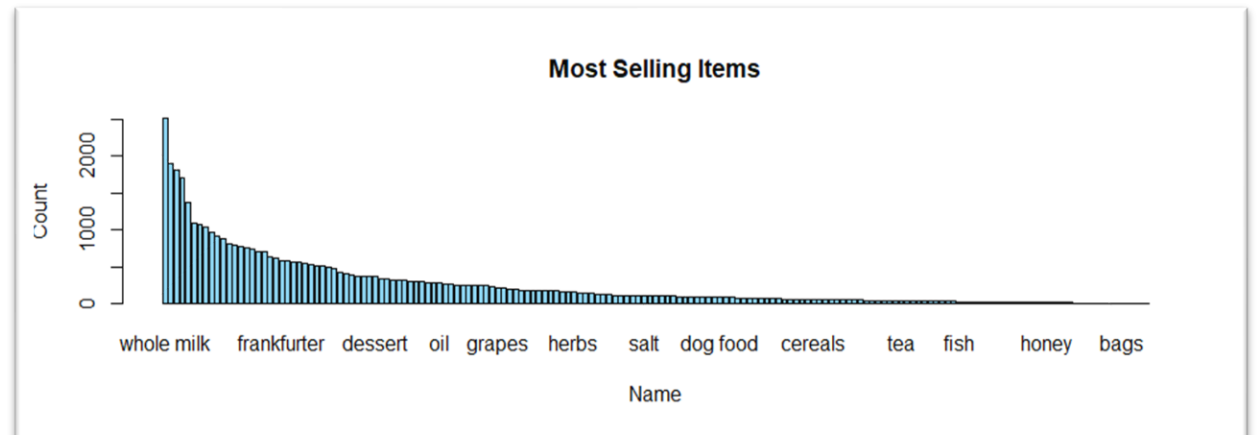
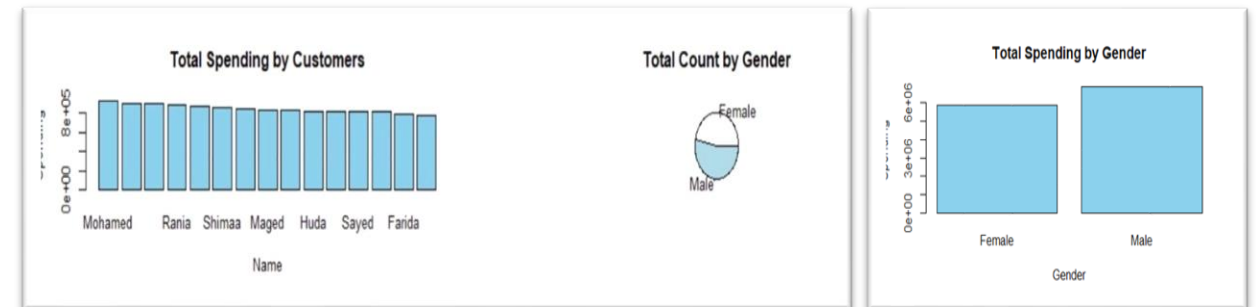
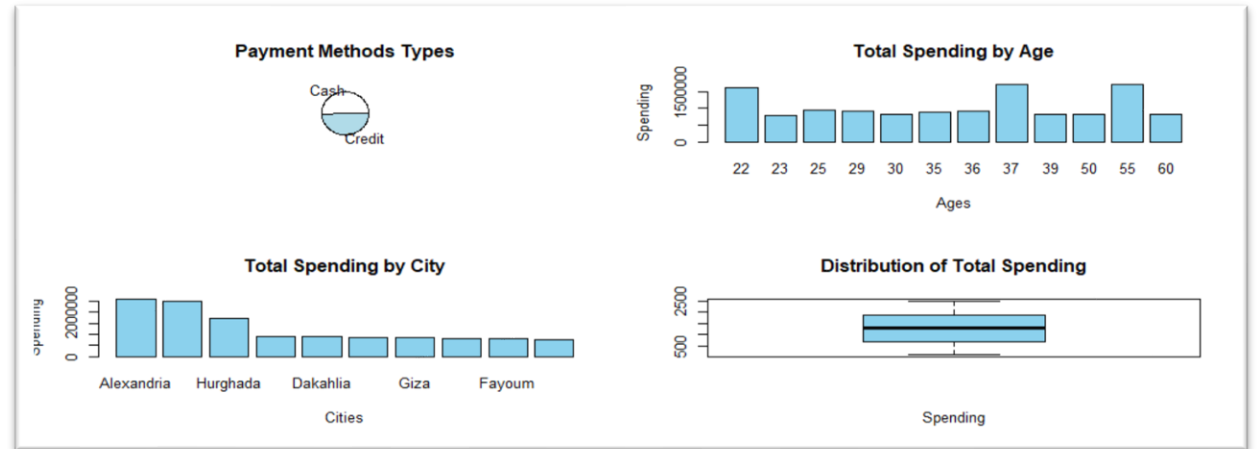
	lhs	rhs	support	confidence	coverage	lift
[1]	{hard cheese}	=> {whole milk}	0.01006609	0.41078838	0.02450432	1.6076815
[2]	{whole milk}	=> {hard cheese}	0.01006609	0.03939515	0.25551601	1.6076815
[3]	{butter milk}	=> {other vegetables}	0.01037112	0.37090909	0.02796136	1.9169159
[4]	{other vegetables}	=> {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

SHARE

- 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.

- 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

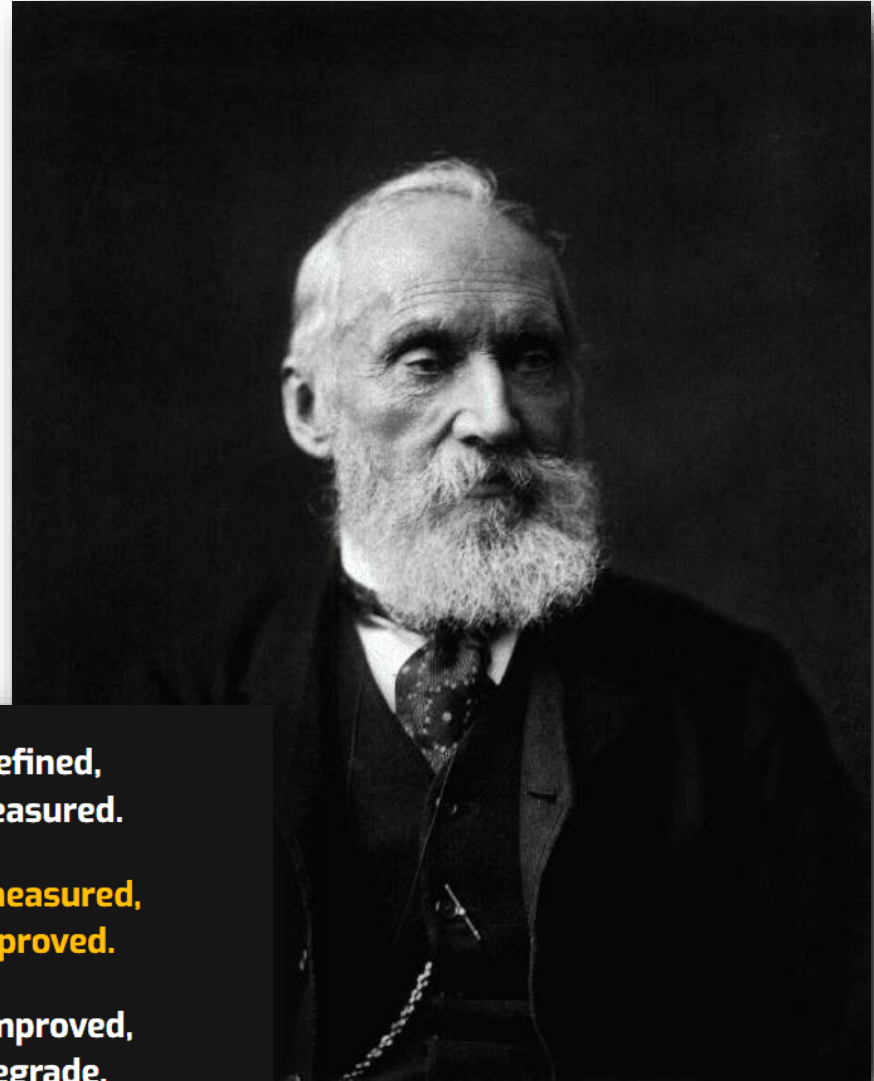
This is the exciting moment when the grocery's stakeholders took all the necessary insights our team have provided and puts them to work in order to solve the original business problem. We should be in on-going contact with the stakeholders and the updated dataset which will be received to monitor the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

**What is not defined,
can not be measured.**

**What is not measured,
can not be improved.**

**What is not improved,
will always degrade.**

William Thomson Kelvin (1824 – 1907)



Thank You



Group 10