BIG DATA SCIENCE

MAHSA RAHIMIAN

UNIVERSITY OF COLORADO

DR. FARNOUSH KASHANI

SUBJECT OF PROJECT:

MALL CUSTOMERS SEGMENTATION — USING MACHINE LEARNING

GITHUB LINKS: HTTPS://GITHUB.COM/MAHSA7915/MAHSA/PROJECTS/1

HTTPS://GITHUB.COM/MAHSA7915?TAB=REPOSITORIES

OUTLINE

- Introduction
- technologies
- Problem statement
- Methods
- Tools
- challenges
- Results
- Plans to mitigate challenges
- References

INTRODUCTION

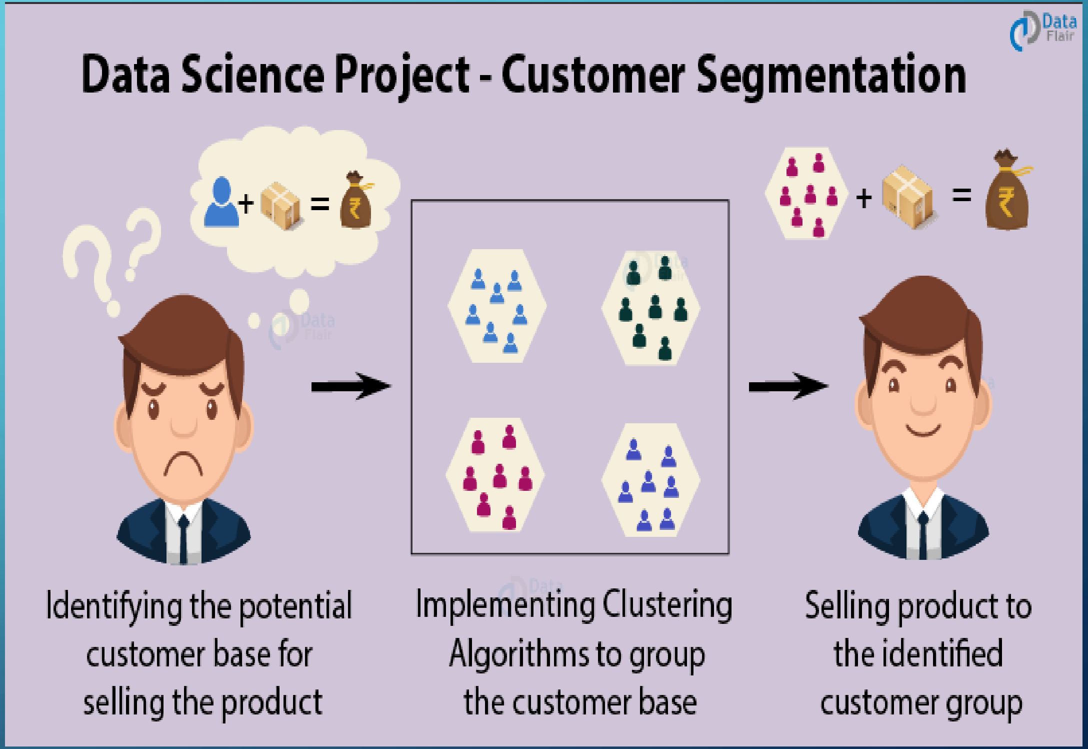
- In this Data Science R project, I will execute an interesting application of machine learning called customer segmentation. Customer Segmentation can play crucial role in commercial organization when they are looking for best customer.
- Customer Segmentation is a crucial application in unsupervised learning. Cluster techniques can allow commercial companies to identify the several segments of their customers which enable the organization to target the potential user base.
- The technologies that I am going to use is Python libraries to implement k-means algorithm.

TECHNOLOGIES

- I will use TensorFlow as an open source machine learning framework. I decided to use this technology because it is available in Python and C++.
- TensorFlow is one of the most well-maintained frameworks for machine learning projects.
- This framework has been created by Google to support research objectives but for a short time it widely used by huge companies such as intel, eBay, Twitter and more.
- My platform is my local server because I am working on a basic problem and local server can support my project.

PROBLEM STATEMENT

- Customer segmentation can divide customers into several groups which share a similarity in different ways that is based on gender, interests, miscellaneous spending habits, and age. This project can help organization to solve a big problem which is target the potential customer for a particular product.



Data Science Project - Customer Segmentation

Identifying the potential customer base for selling the product

Implementing Clustering Algorithms to group the customer base

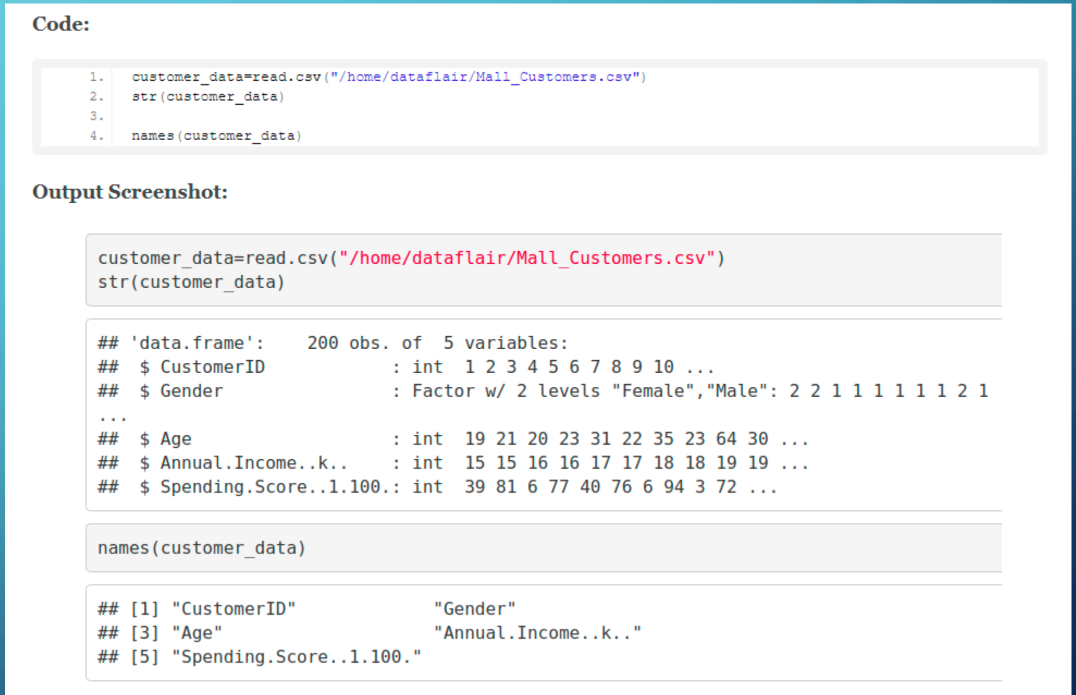
Selling product to the identified customer group

PROBLEM STATEMENT

- Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach of the customer they are targeting.
- Thus, they should target the requirements of each and every customer. Accordingly, commercial organizations can gain deeper knowledge about customer’s preferences which will cause maximum profit to the company.

METHODS

- In the first step of this data science project, we will perform data exploration. We will import the essential packages required for this role and then read our data. Finally, we will go through the input data to gain necessary insights about it.



Code:

```
customer_data=read.csv("/home/dataflair/Mall_Customers.csv")
str(customer_data)

names(customer_data)
```

Output Screenshot:

```
customer_data=read.csv("/home/dataflair/Mall_Customers.csv")
str(customer_data)
```

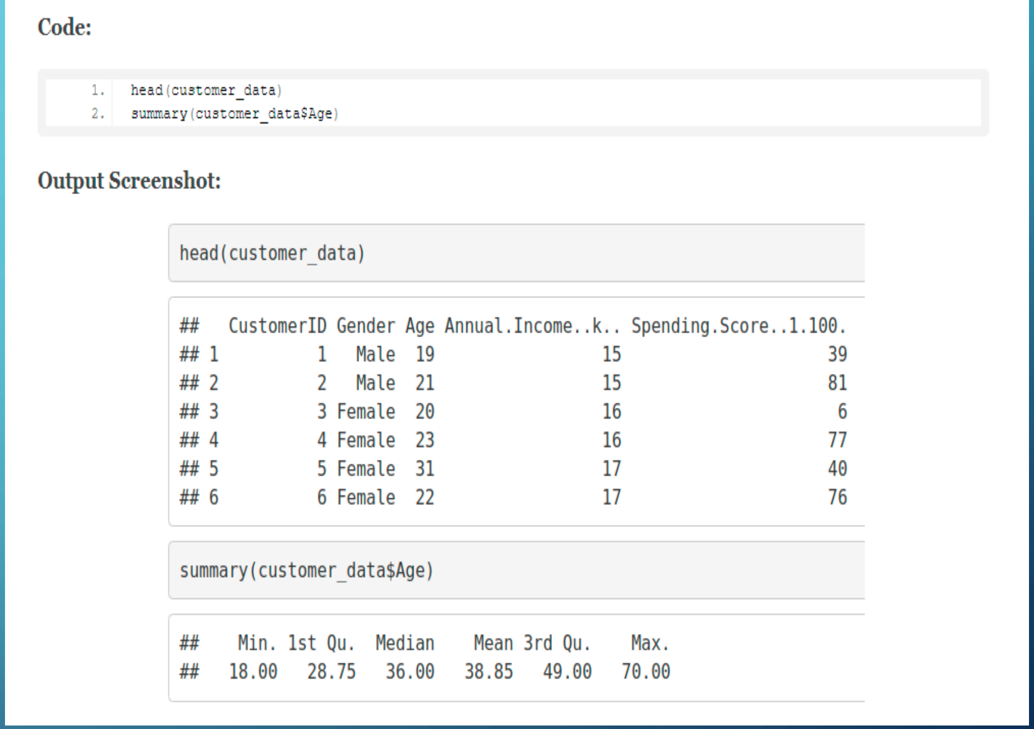
```
## 'data.frame':    200 obs. of  5 variables:
##  $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Gender                : Factor w/ 2 levels "Female","Male": 2 2 1 1 1 1 1 1 2 1
...
##  $ Age                   : int  19 21 20 23 31 22 35 23 64 30 ...
##  $ Annual.Income..k..    : int  15 15 16 16 17 17 18 18 19 19 ...
##  $ Spending.Score..1.100.: int  39 81 6 77 40 76 6 94 3 72 ...
```

```
names(customer_data)
```

```
## [1] "CustomerID"             "Gender"
## [3] "Age"                    "Annual.Income..k.."
## [5] "Spending.Score..1.100."
```

METHODS

- Now, we display the first six rows of our dataset using the head() function and use the summary() function to output summary of it.



Code:

```
head(customer_data)
summary(customer_data$Age)
```

Output Screenshot:

```
head(customer_data)
```

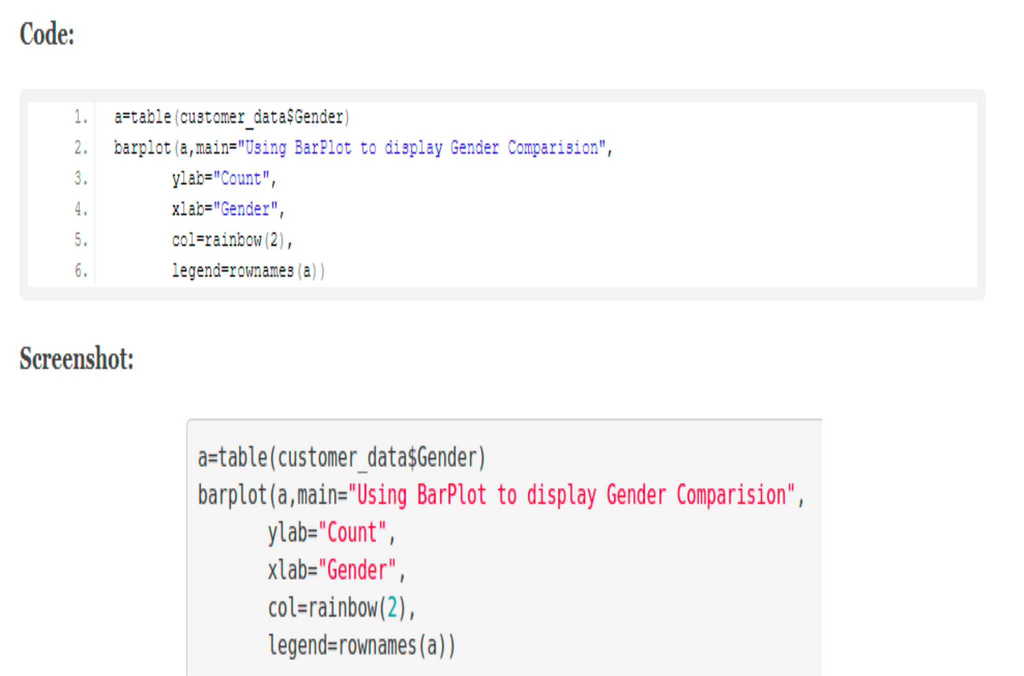
```
##   CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.
## 1          1   Male  19                 15                     39
## 2          2   Male  21                 15                     81
## 3          3 Female  20                 16                      6
## 4          4 Female  23                 16                     77
## 5          5 Female  31                 17                     40
## 6          6 Female  22                 17                     76
```

```
summary(customer_data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   28.75   36.00   38.85   49.00   70.00
```

METHODS
CUSTOMER GENDER VISUALIZATION

- In this, we will create a barplot and a piechart to show the gender distribution across our customer data dataset.



Code:

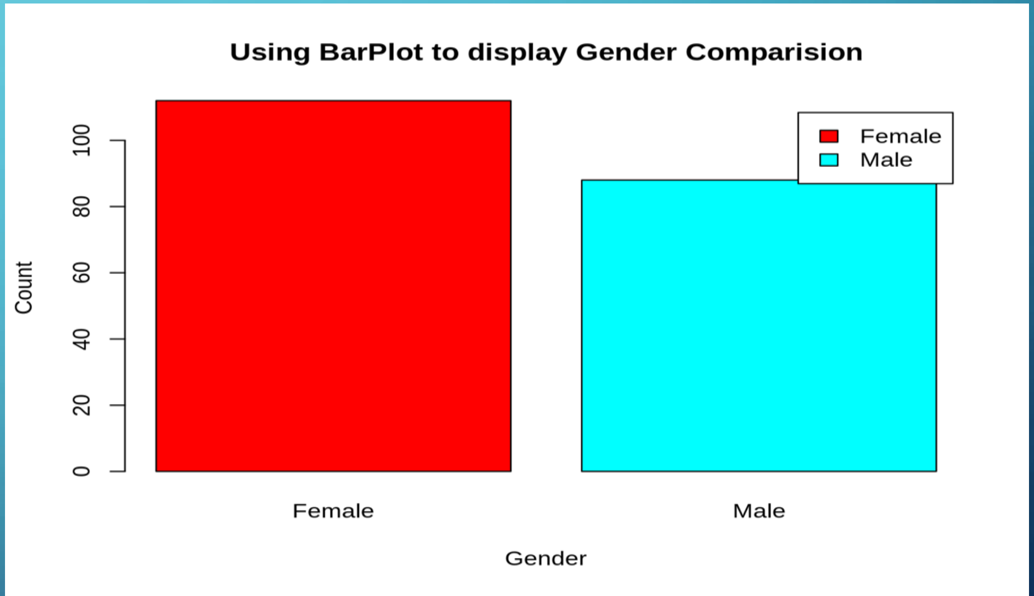
```
a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))
```

Screenshot:

```
a=table(customer_data$Gender)
barplot(a,main="Using BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="Gender",
        col=rainbow(2),
        legend=rownames(a))
```

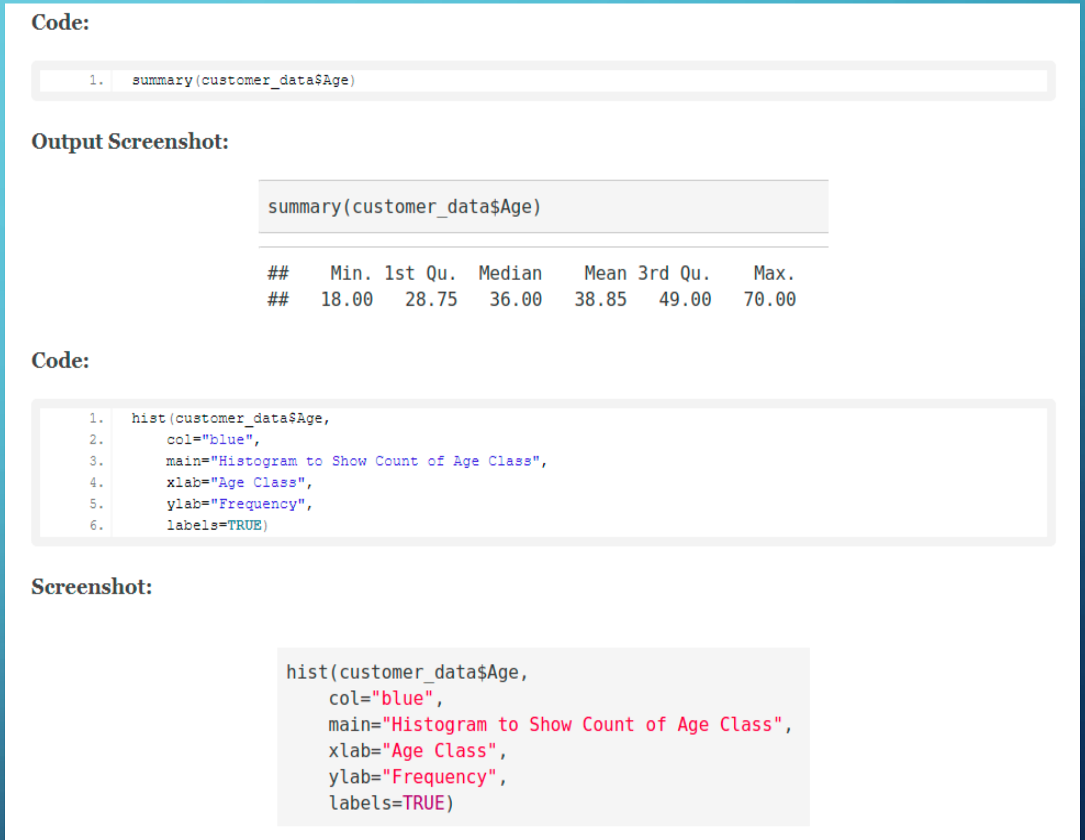
METHODS

- From this barplot, we observe that the number of females is higher than the males. Now, let us visualize a pie chart to observe the ratio of male and female distribution.



METHODS
VISUALIZATION OF AGE DISTRIBUTION

- Now lets plot a histogram to view the distribution to plot the frequency of customer ages. We will first proceed by taking summary of the Age variable.



Code:

```
summary(customer_data$Age)
```

Output Screenshot:

```
summary(customer_data$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   18.00   28.75   36.00   38.85   49.00   70.00
```

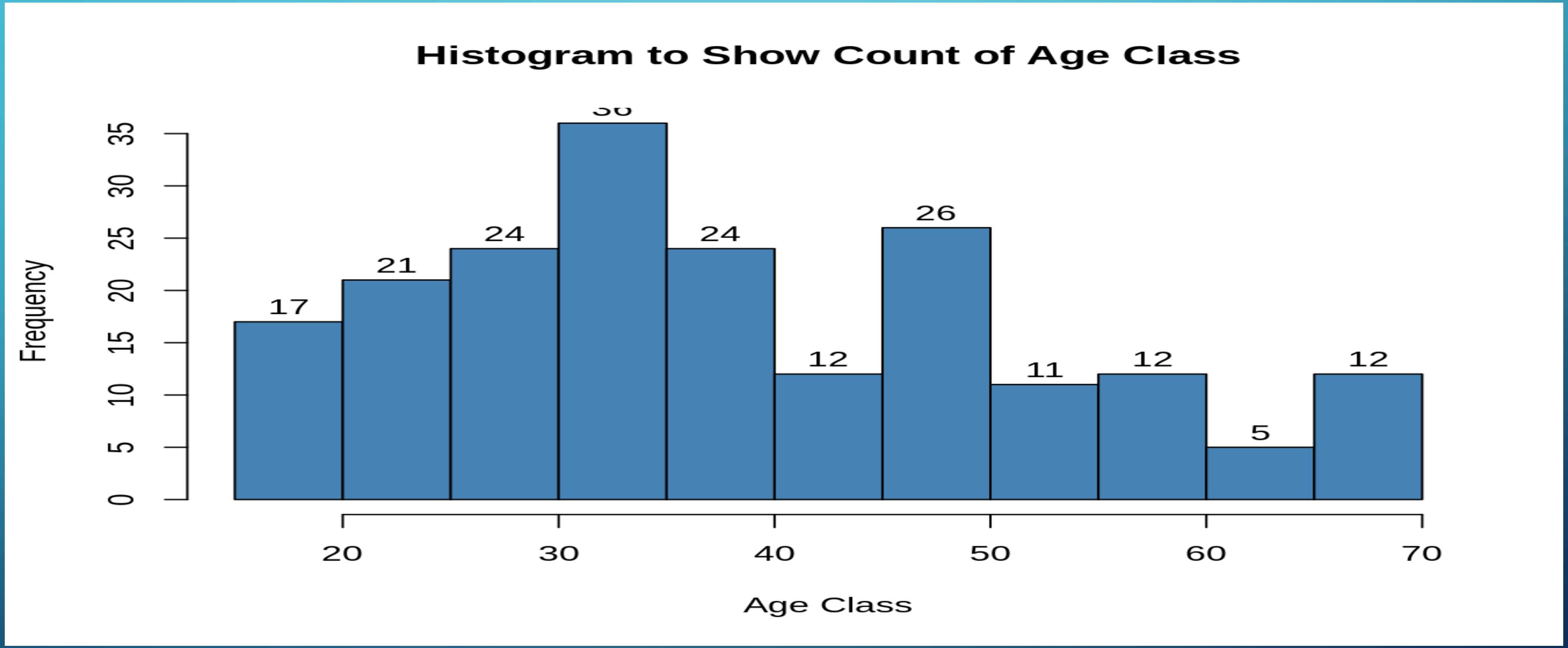
Code:

```
hist(customer_data$Age,
    col="blue",
    main="Histogram to Show Count of Age Class",
    xlab="Age Class",
    ylab="Frequency",
    labels=TRUE)
```

Screenshot:

```
hist(customer_data$Age,
    col="blue",
    main="Histogram to Show Count of Age Class",
    xlab="Age Class",
    ylab="Frequency",
    labels=TRUE)
```

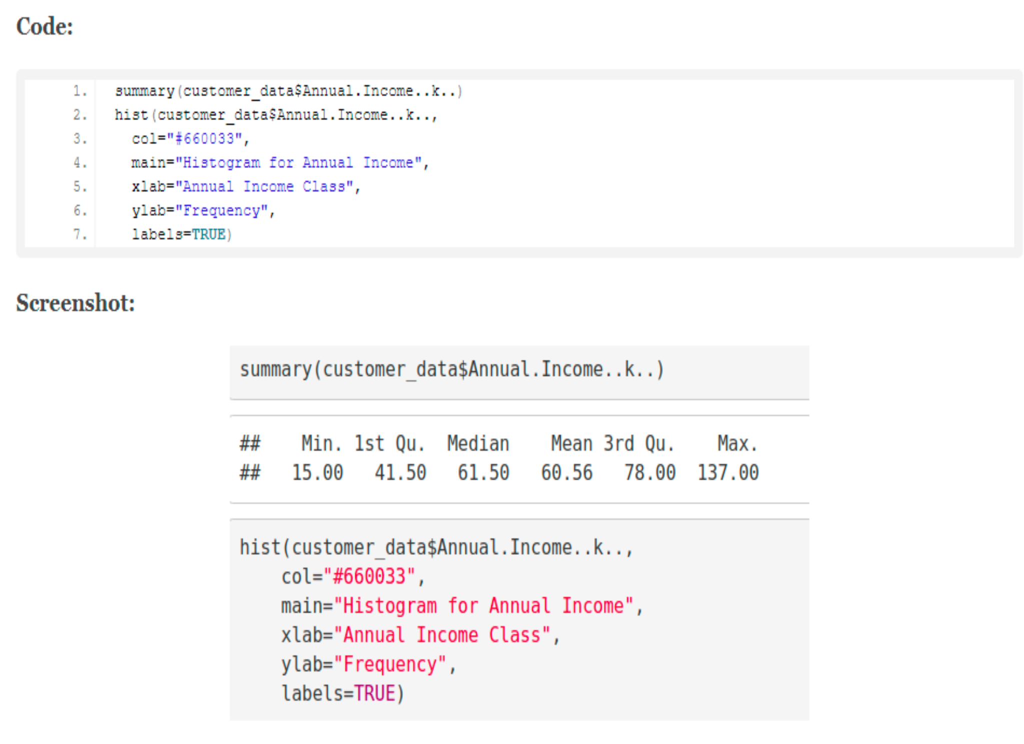
METHODS



Histogram to Show Count of Age Class
Frequency
0
5
10
15
20
25
30
35
17
21
24
36
24
12
26
11
12
5
12
20
30
40
50
60
70
Age Class

METHODS
ANALYSIS OF THE ANNUAL INCOME OF THE CUSTOMERS

- In this section of the project, we will create visualizations to analyze the annual income of the customers. We will plot a histogram and then we will proceed to examine this data using a density plot.



Code:

```
summary(customer_data$Annual.Income..k..)
hist(customer_data$Annual.Income..k..,
  col="#660033",
  main="Histogram for Annual Income",
  xlab="Annual Income Class",
  ylab="Frequency",
  labels=TRUE)
```

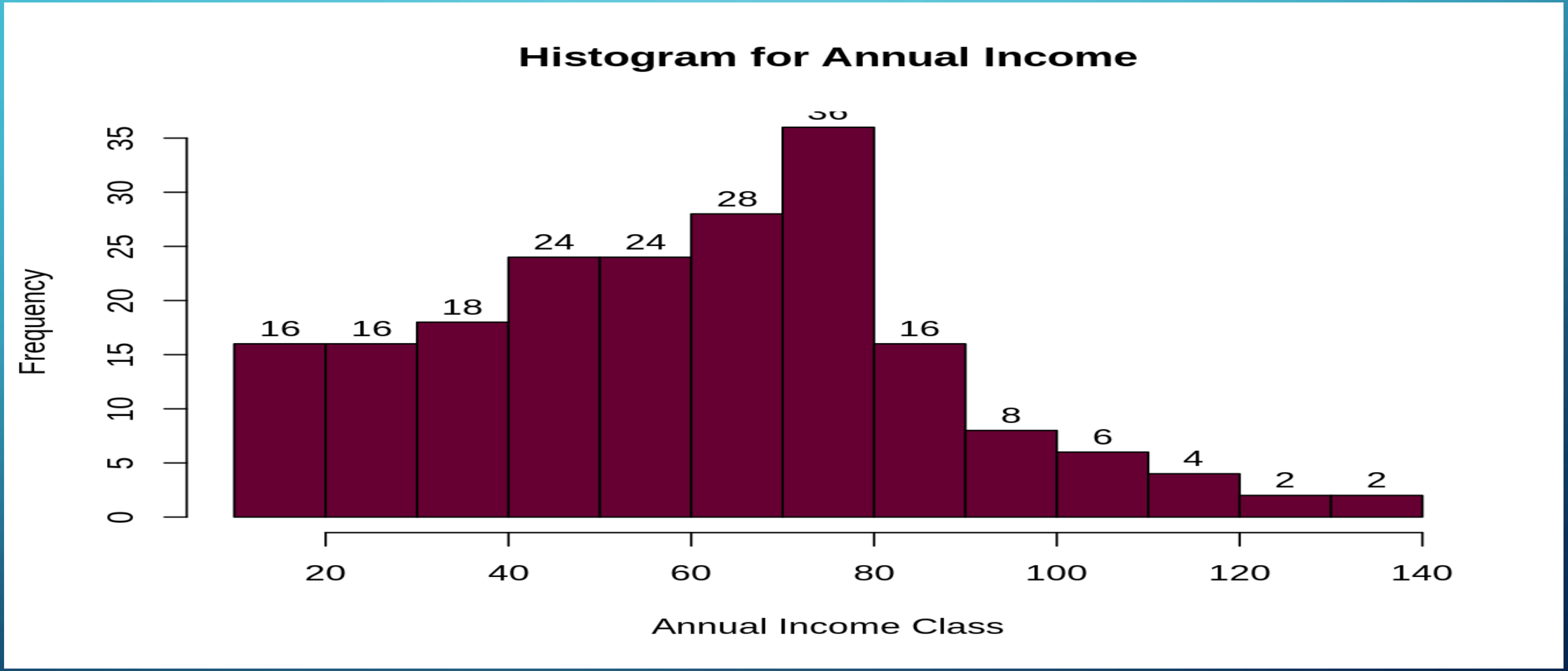
Screenshot:

```
summary(customer_data$Annual.Income..k..)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   41.50   61.50   60.56   78.00  137.00
```

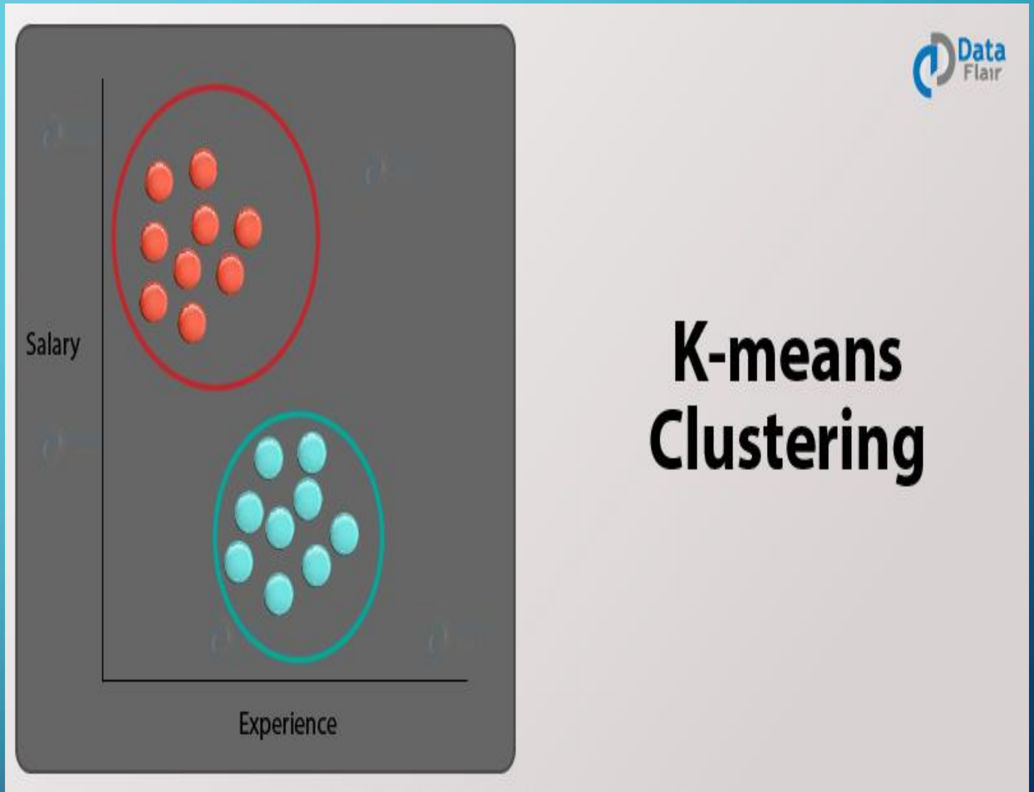
```
hist(customer_data$Annual.Income..k..,
    col="#660033",
    main="Histogram for Annual Income",
    xlab="Annual Income Class",
    ylab="Frequency",
    labels=TRUE)
```

OUTPUT


Histogram for Annual Income
Frequency
0
5
10
15
20
25
30
35
16
16
18
24
24
28
36
16
8
6
4
2
2
20
40
60
80
100
120
140
Annual Income Class

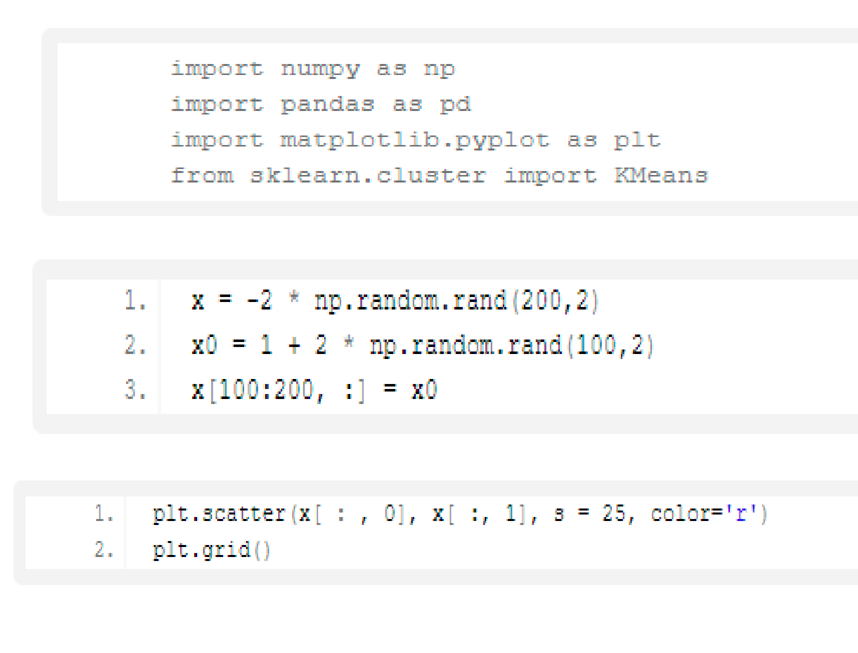
TOOLS

- One of the most popular Machine Learning algorithms is K-means clustering. It is an unsupervised learning algorithm, meaning that it is used for unlabeled datasets. Imagine that you have several points spread over an n-dimensional space.



TOOLS

- We can use K-means over random data using Python libraries.
- 1. First, we import the essential Python Libraries required for implementing our k-means algorithm
- 2. We then randomly generate 200 values divided in two clusters of 100 data points each.
- 3. We proceed to plot our generated random values and obtain the following graph.

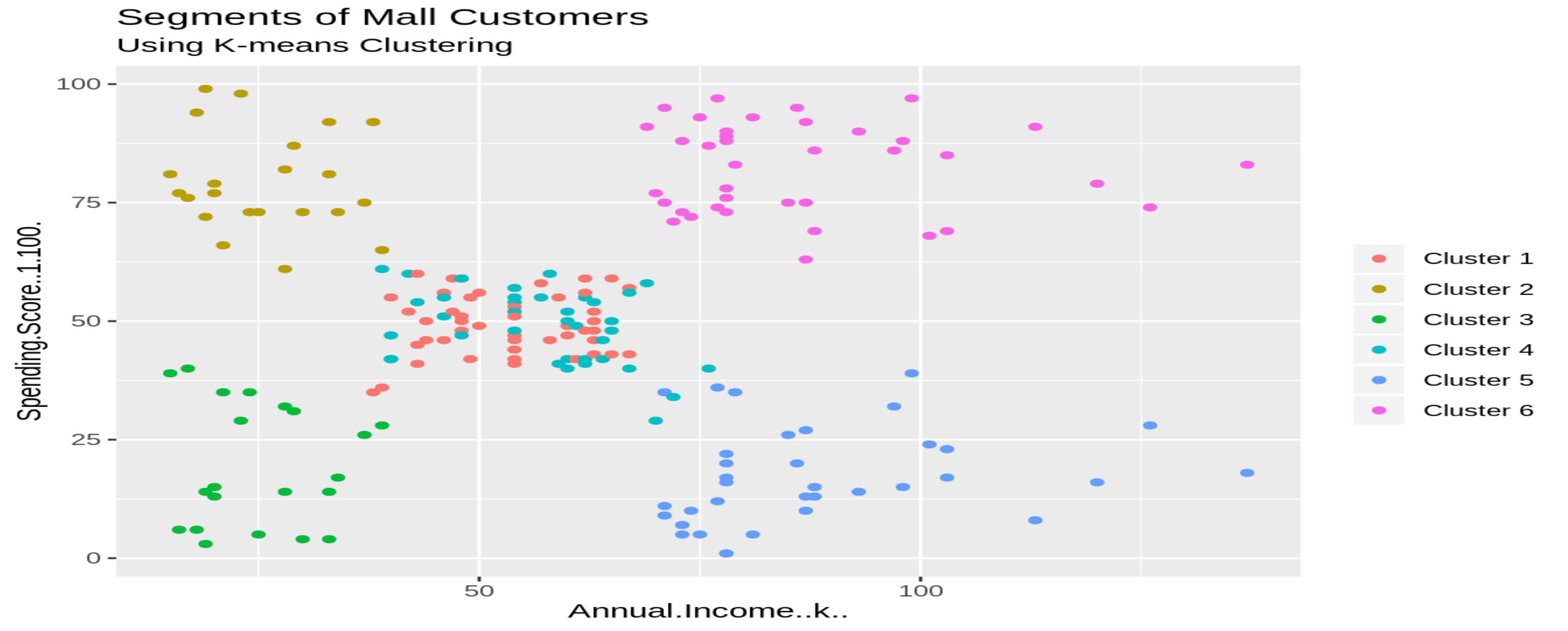


```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
```

```
x = -2 * np.random.rand(200,2)
x0 = 1 + 2 * np.random.rand(100,2)
x[100:200, :] = x0
```

```
plt.scatter(x[ : , 0], x[ :, 1], s = 25, color='r')
plt.grid()
```

RESULT
SEGMENTS OF MALL CUSTOMERS

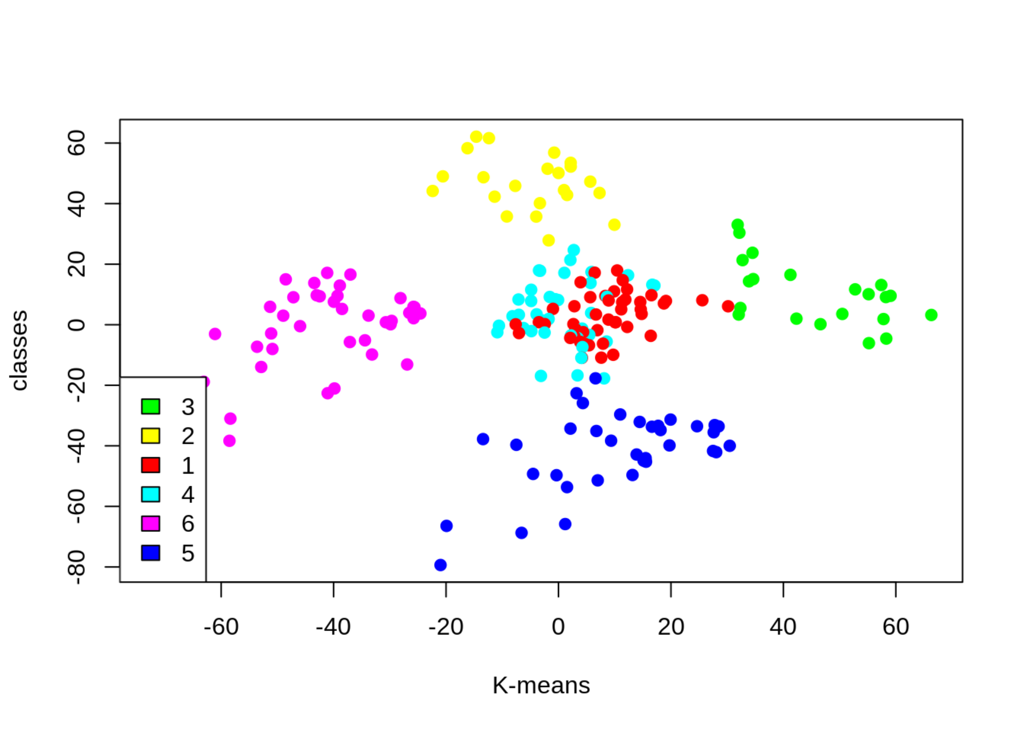


RESULT

- From the above visualization, we observe that there is a distribution of 6 clusters as follows:
- Cluster 6 and 4
- Cluster 1
- Cluster 3
- Cluster 2
- Cluster 5

RESULT

- Cluster 4 and 1
- Cluster 6
- Cluster 5
- Cluster 3
- Cluster 2

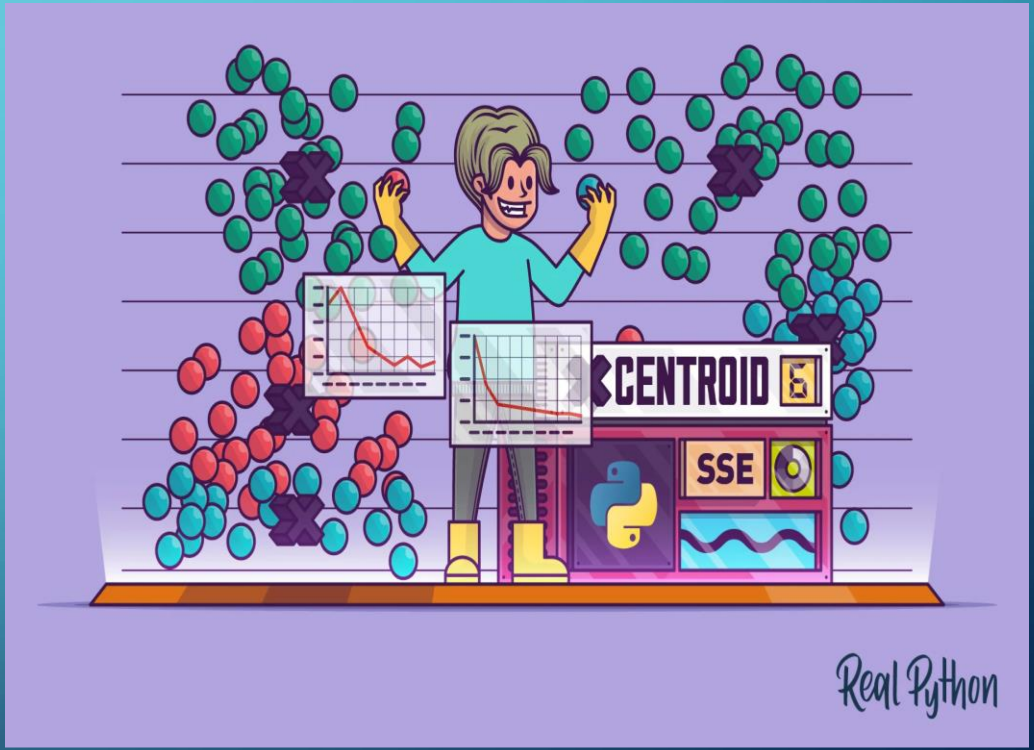


RESULTS

- In this data science project, as you know we went through customer segmentation model. We used a class of machine learning called unsupervised learning. As I mentioned in this report, we used a clustering algorithm called k-means clustering. In this project we learned how effective is this K-means clustering in this project. We analyzed and visualized the data and then proceeded to implement our algorithm.

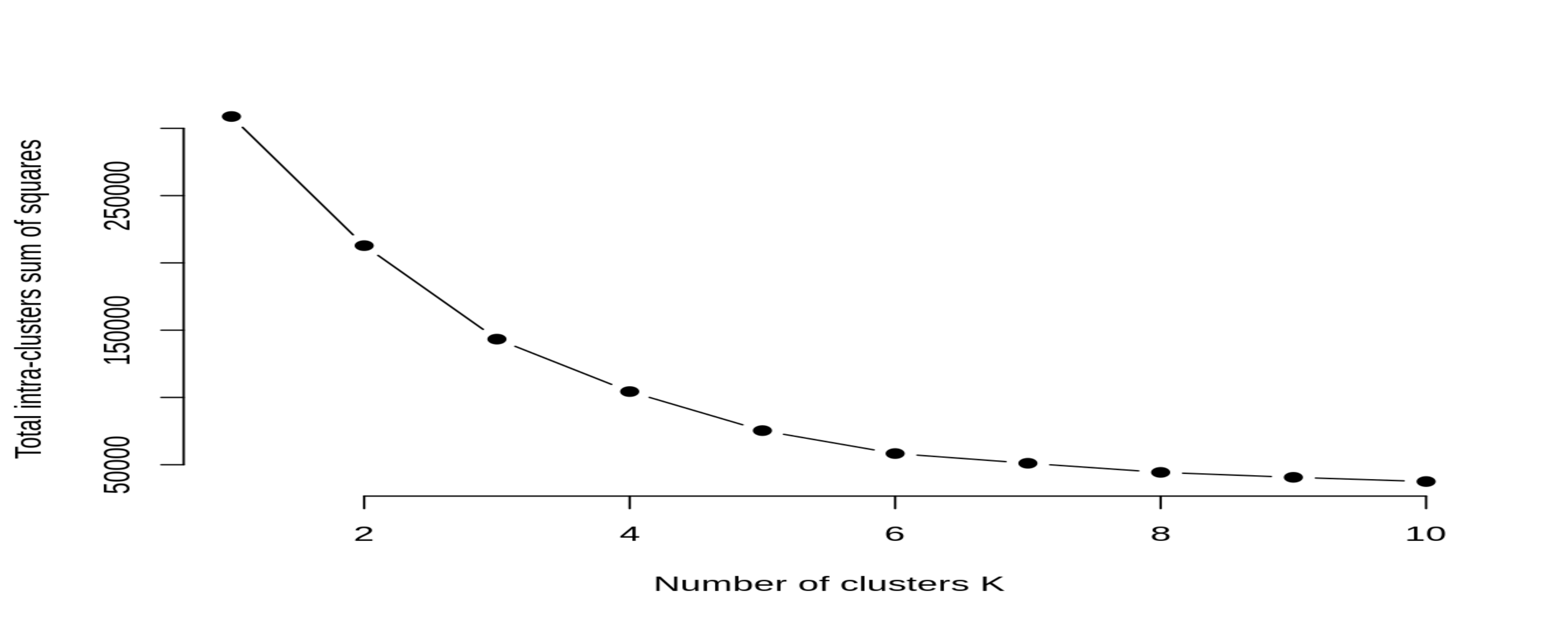
PLANS TO MITIGATE CHALLENGES

- the number of clusters (k) that we wish to produce in the final output is a challenge.
- On the other hand, While working with clusters, you need to specify the number of clusters to use.



PLANS TO MITIGATE CHALLENGES

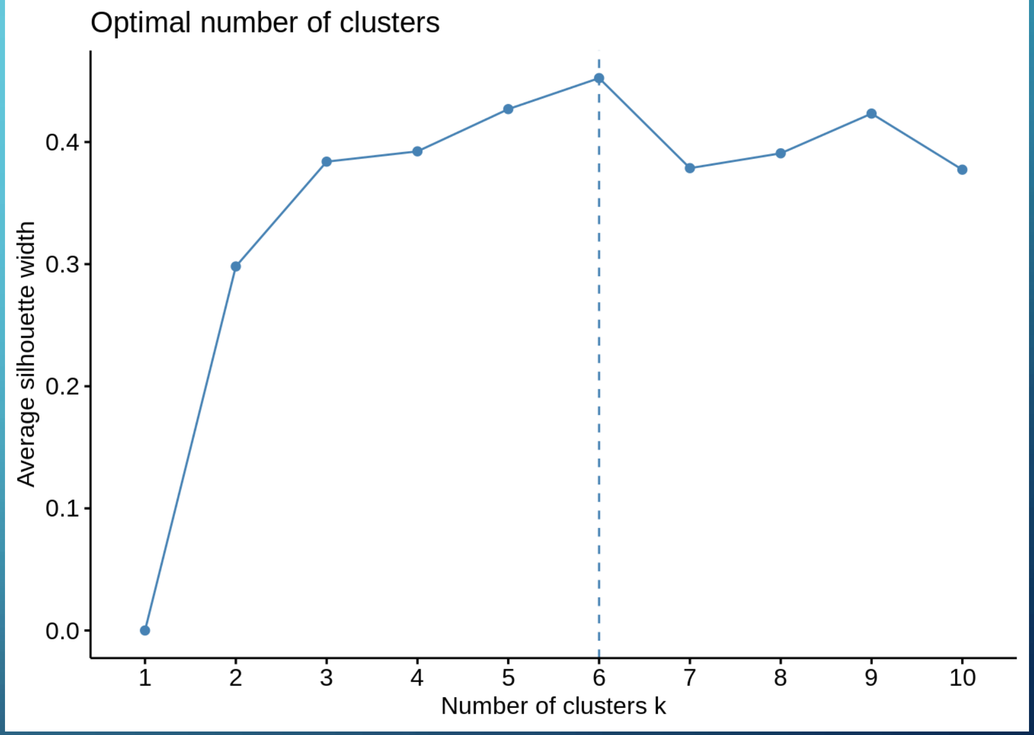
NUMBER OF CLUSTERS K



PLANS TO MITIGATE CHALLENGES

OPTIMAL NUMBER OF CLUSTERS

- Using the gap statistic, one can compare the total intracluster variation for different values of k along with their expected values under the null reference distribution of data. With the help of Monte Carlo simulations, one can produce the sample dataset.


Optimal number of clusters
Average silhouette width
Number of clusters k