Mini Project

Computer Programming with R

Clustering Analysis

Dataset for Clustering Analysis

Mall Customer Segmentation Dataset

This dataset contains basic data about customers like Customer ID, age, gender, annual income and spending score.

Spending Score is assigned to each customer based on certain defined parameters like customer behaviour.

Purpose: Based on the above data we have to find a group of customer, which are unique in their own way. So that proper marketing strategy can be defined for each group and targeted separately.

What is Cluster Analysis?

Cluster: a collection of data objects

Similar to one another within the same cluster

Dissimilar to the objects in other clusters

Cluster analysis

Grouping a set of data objects into clusters

Clustering is unsupervised classification: no predefined classes

Clustering is used:

As a stand-alone tool to get insight into data distribution

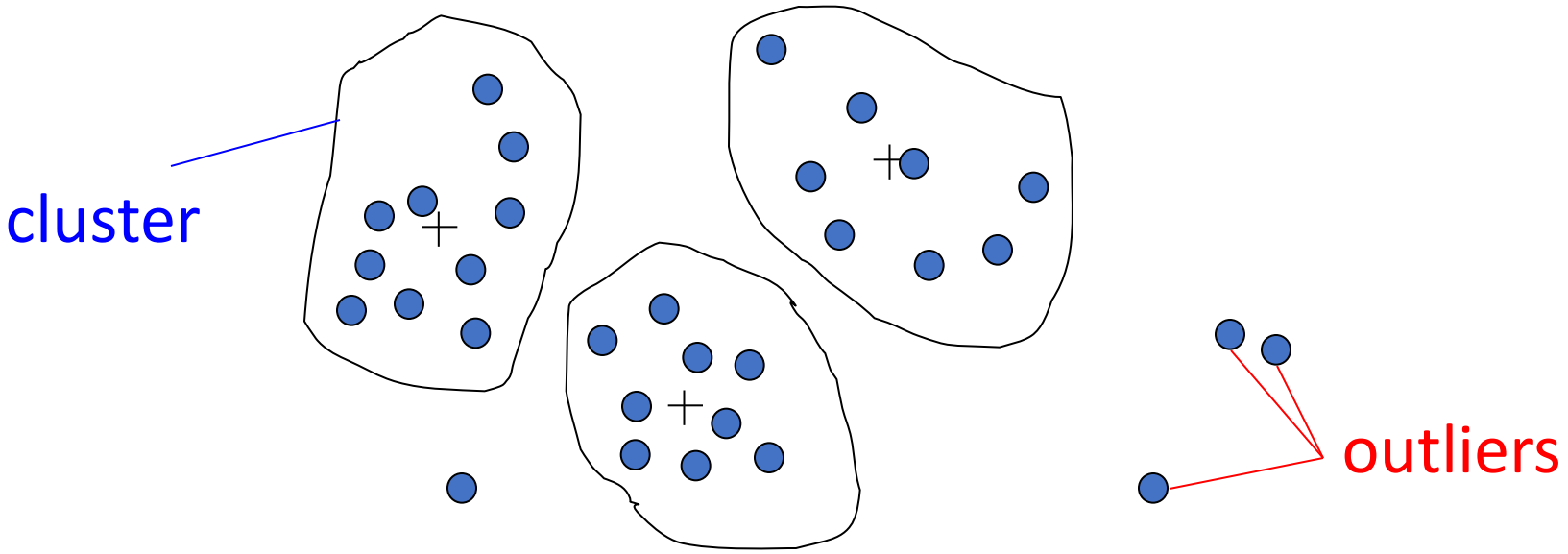
- Visualization of clusters may unveil important information

As a preprocessing step for other algorithms

- Efficient indexing or compression often relies on clustering

Outliers

- Outliers are objects that do not belong to any cluster or form clusters of very small cardinality



- In some applications we are interested in discovering outliers, not clusters (outlier analysis)

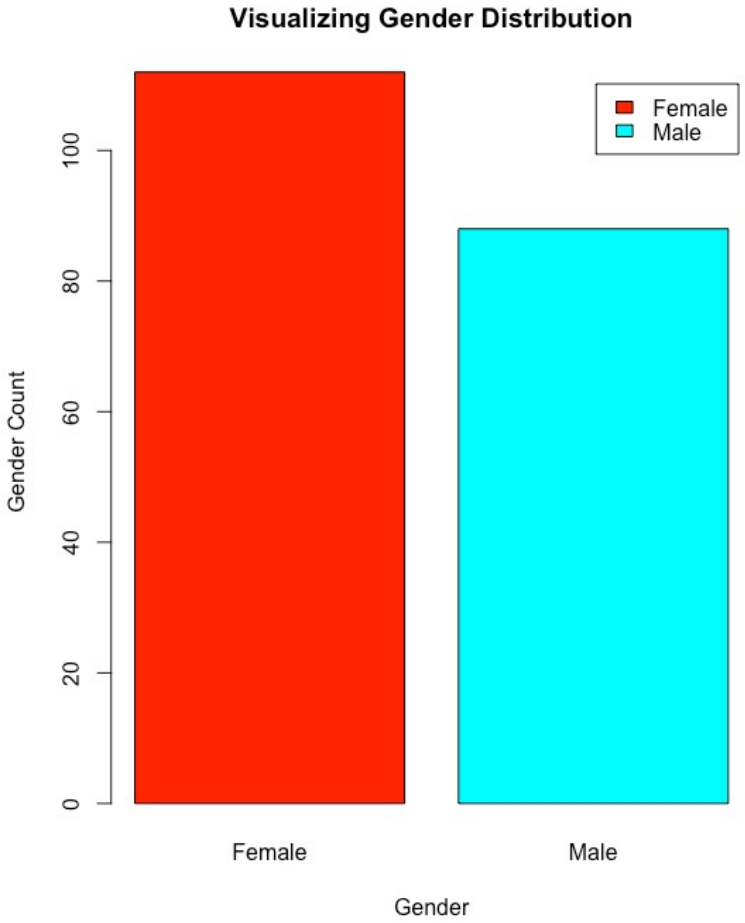
EPLORATORY DATA ANALYSIS

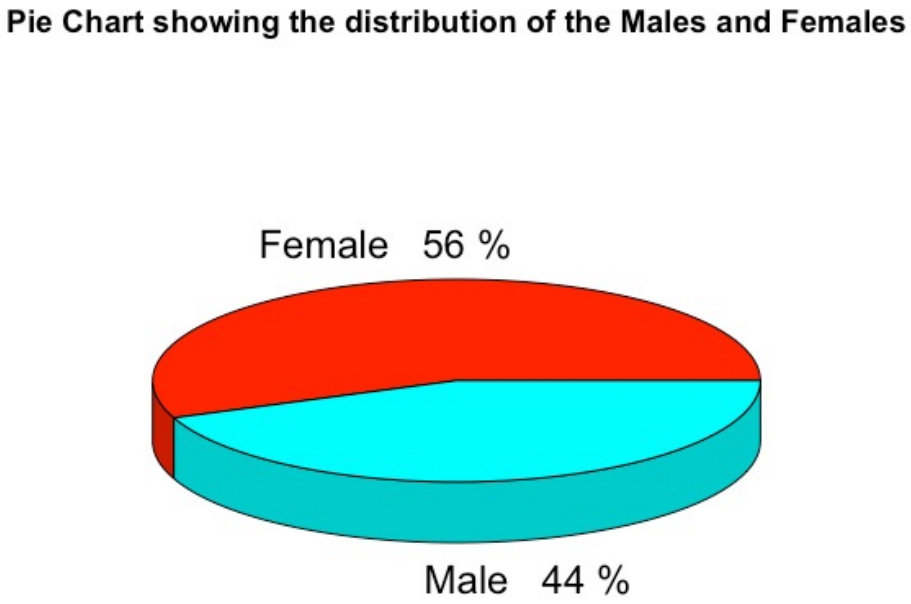
- Let us explore the data further and see the variations in the columns. It is also a good practice to look at the first and last few rows of the data.
- str(customer)
- summary(customer)
- head(customer)
- tail(customer)

```
> str(customer)
'data.frame':   200 obs. of  5 variables:
 $ CustomerID            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Gender                : chr  "Male" "Male" "Female" "Female" ...
 $ 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 ...
> summary(customer)
   CustomerID        Gender               Age        Annual.Income..k.. Spending.Score..1.100.
 Min.   :  1.00   Length:200         Min.   :18.00   Min.   : 15.00     Min.   : 1.00
 1st Qu.: 50.75   Class :character   1st Qu.:28.75   1st Qu.: 41.50     1st Qu.:34.75
 Median :100.50   Mode  :character   Median :36.00   Median : 61.50     Median :50.00
 Mean   :100.50                      Mean   :38.85   Mean   : 60.56     Mean   :50.20
 3rd Qu.:150.25                      3rd Qu.:49.00   3rd Qu.: 78.00     3rd Qu.:73.00
 Max.   :200.00                      Max.   :70.00   Max.   :137.00     Max.   :99.00
> head(customer)
  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
> tail(customer)
    CustomerID Gender Age Annual.Income..k.. Spending.Score..1.100.
195        195 Female  47                120                     16
196        196 Female  35                120                     79
197        197 Female  45                126                     28
198        198   Male  32                126                     74
199        199   Male  32                137                     18
200        200   Male  30                137                     83
>
```

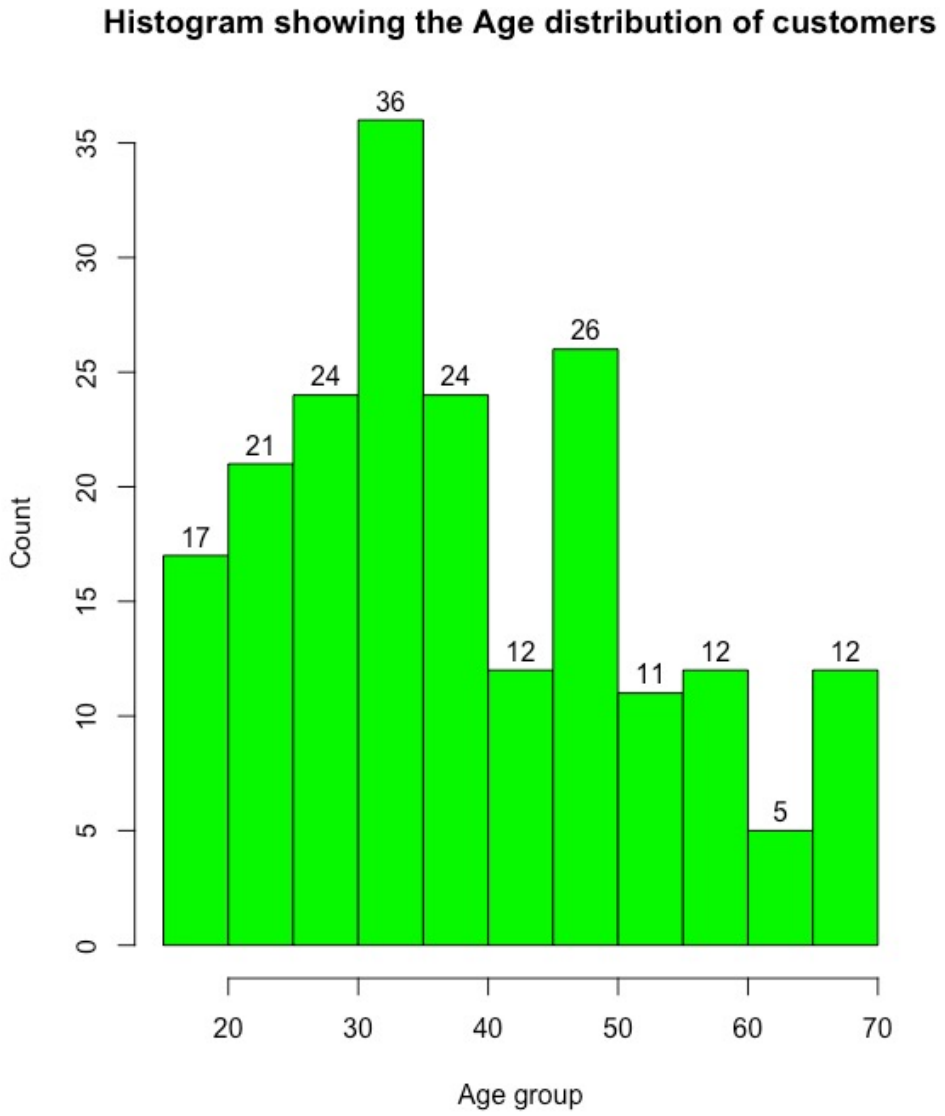
Exploring the **Gender** column.

- We will visualize the column using bar chart and pie chart

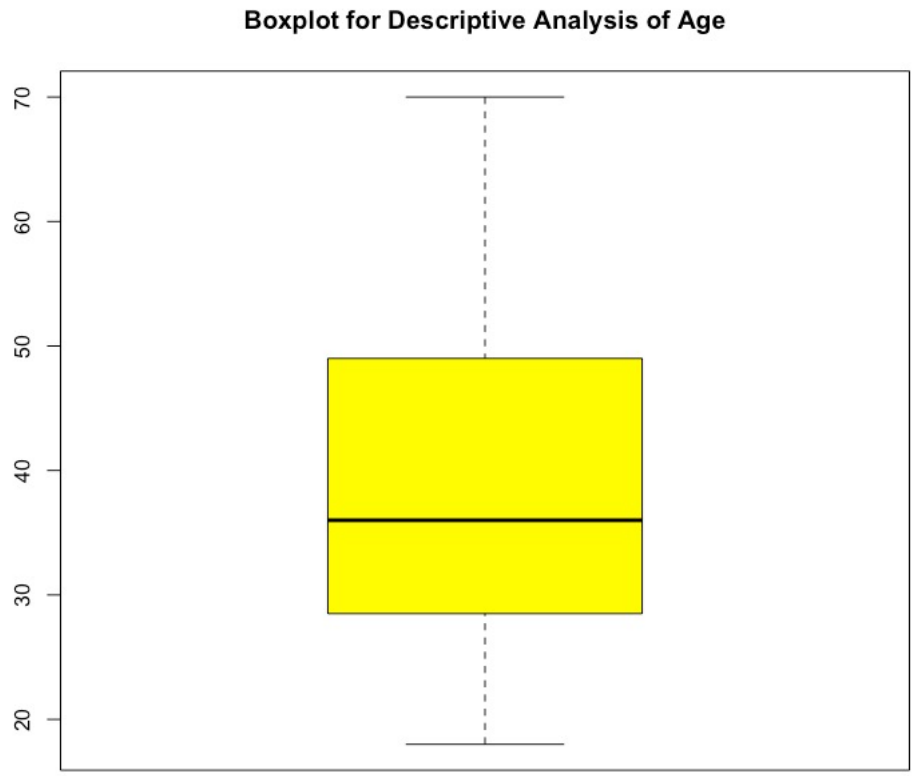


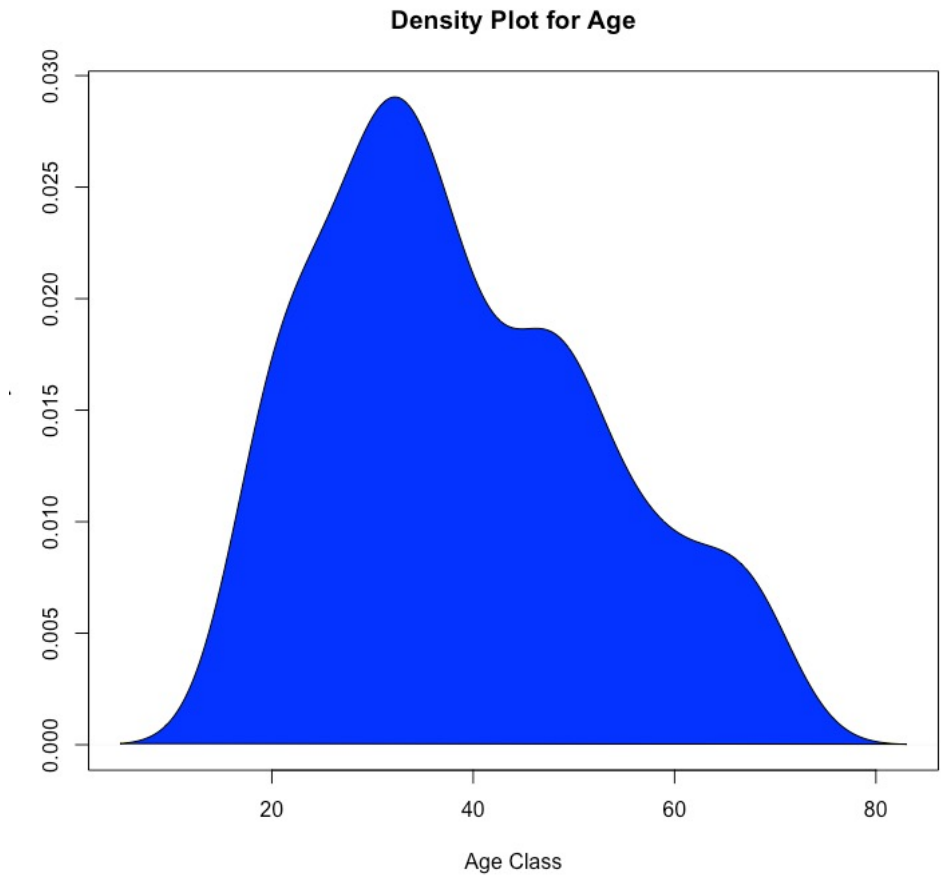


Explore the **Age** column to see its distribution.

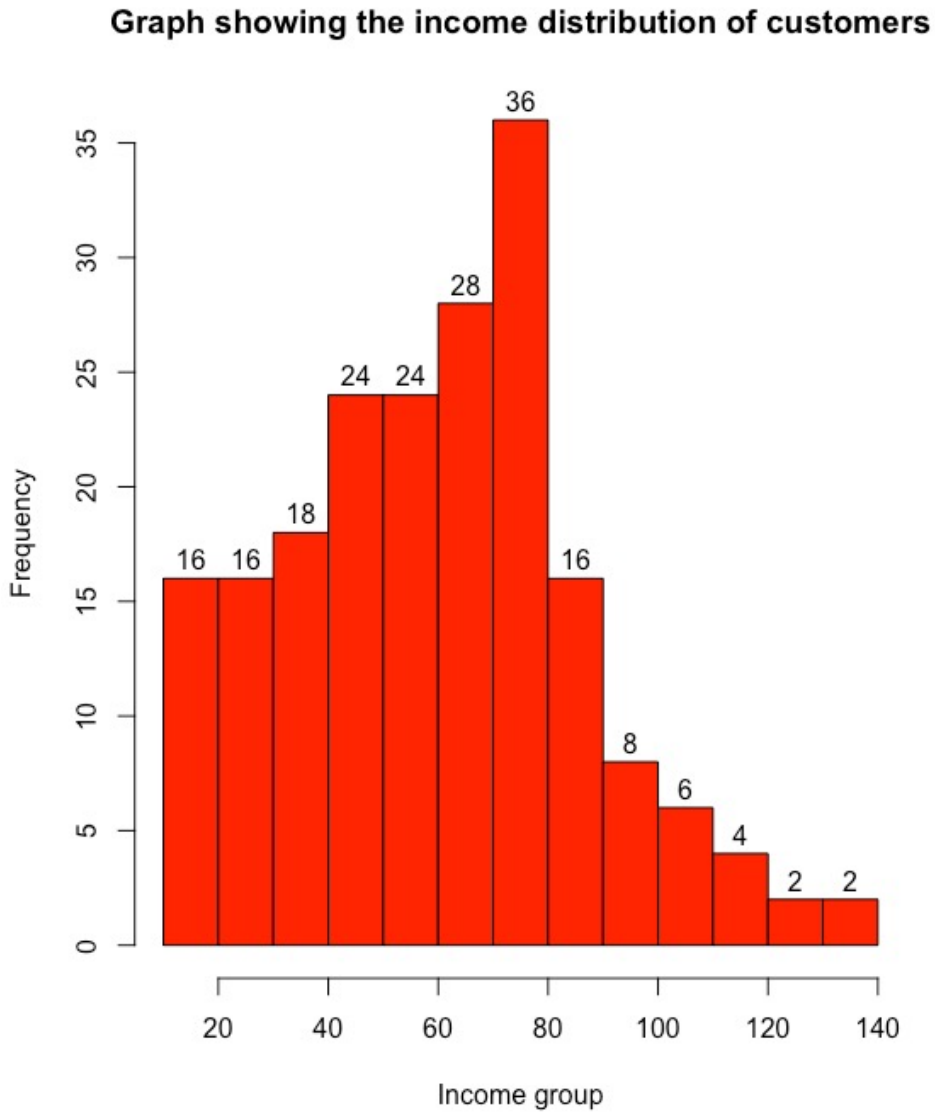


Additional Diagrams of Age Column

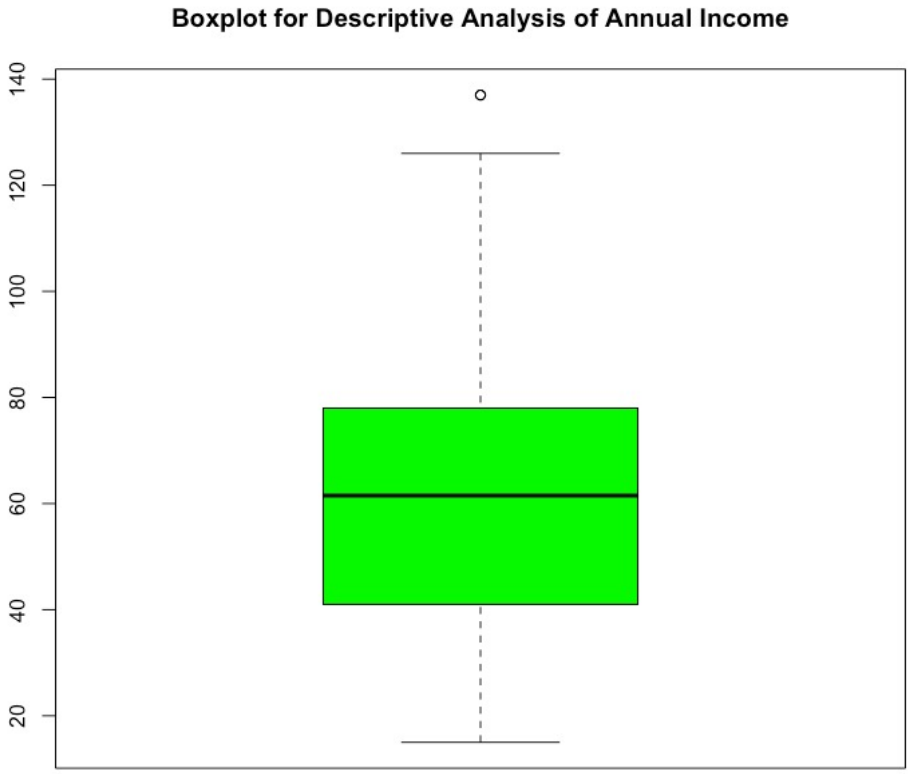


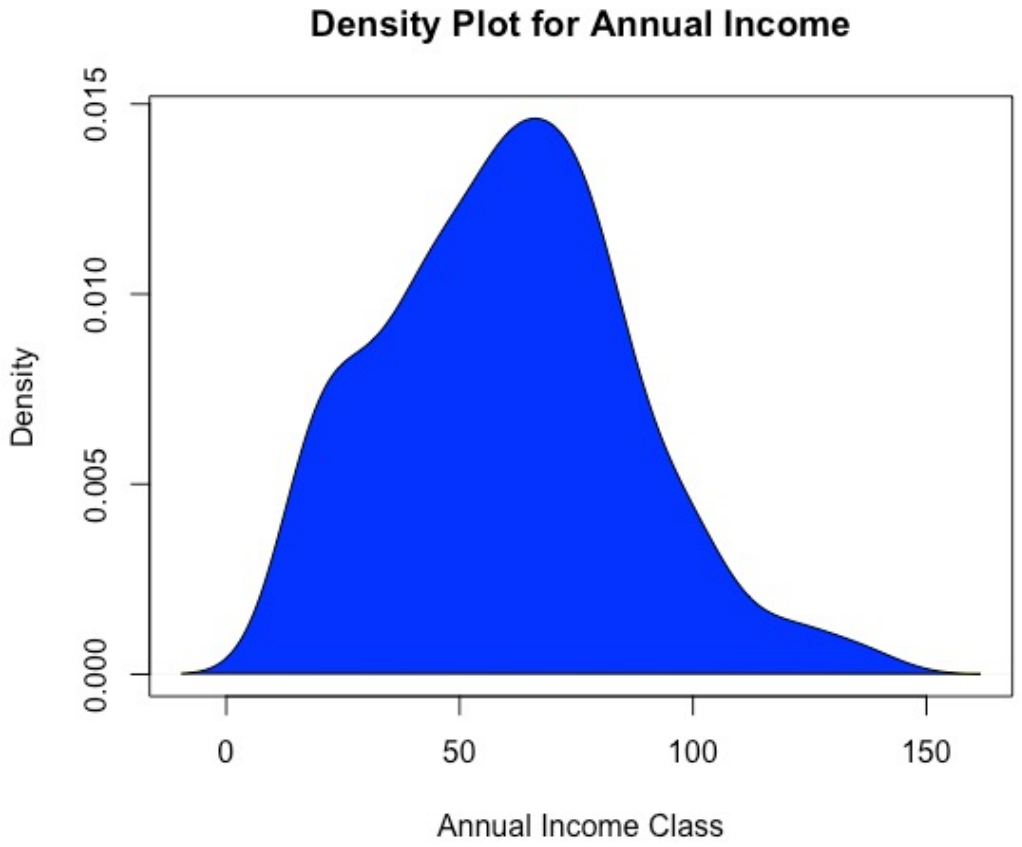


Explore the **Income** column to see its distribution.

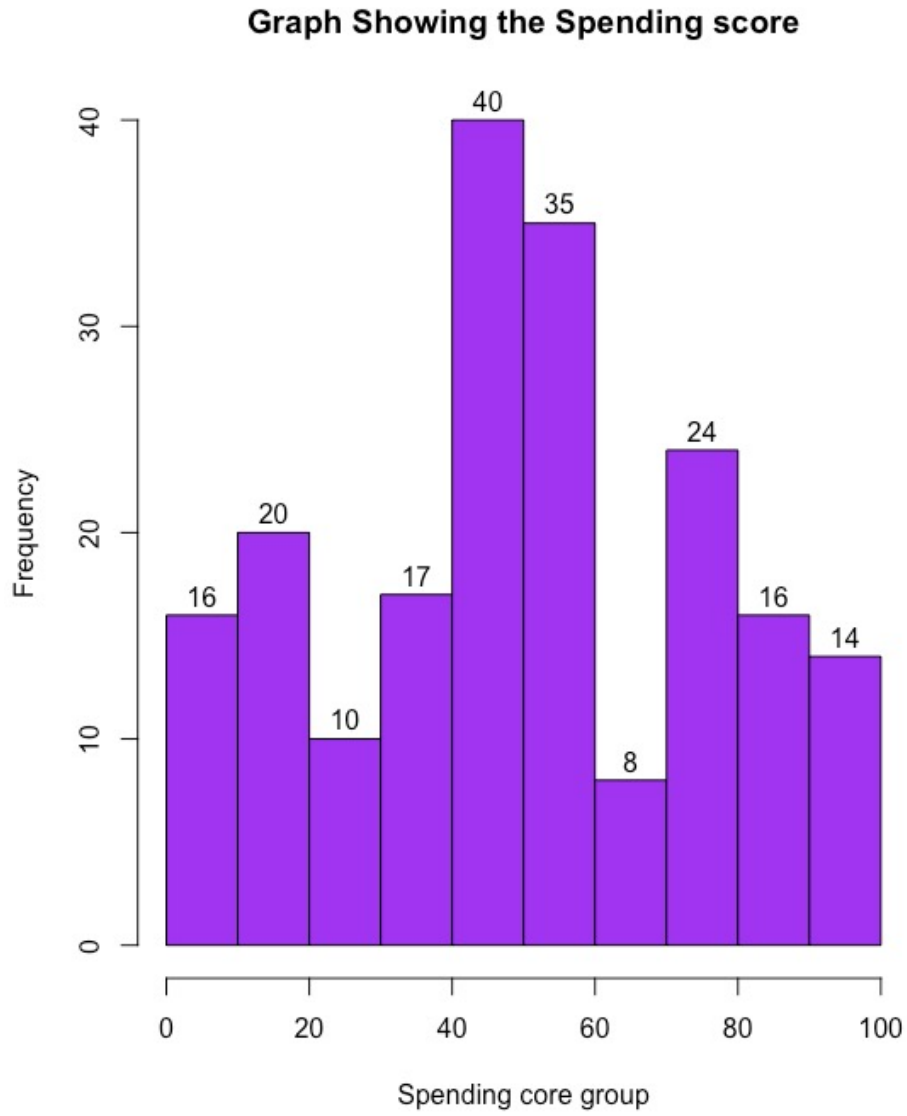


Additional Diagrams of Income Column

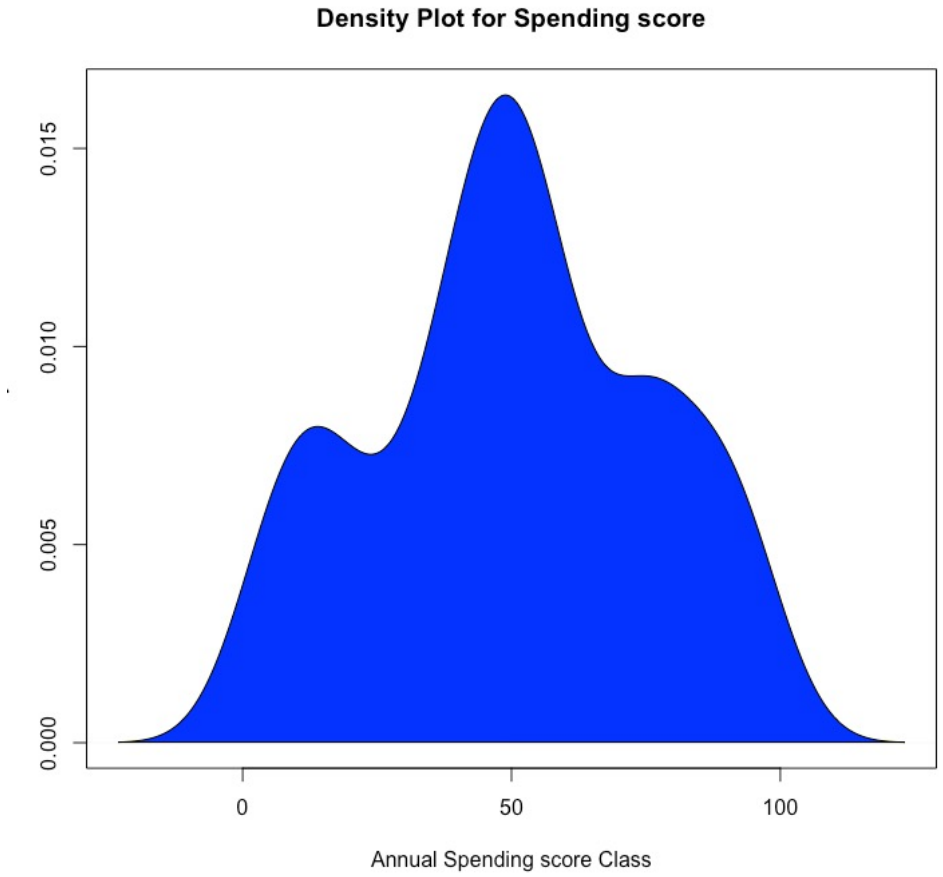


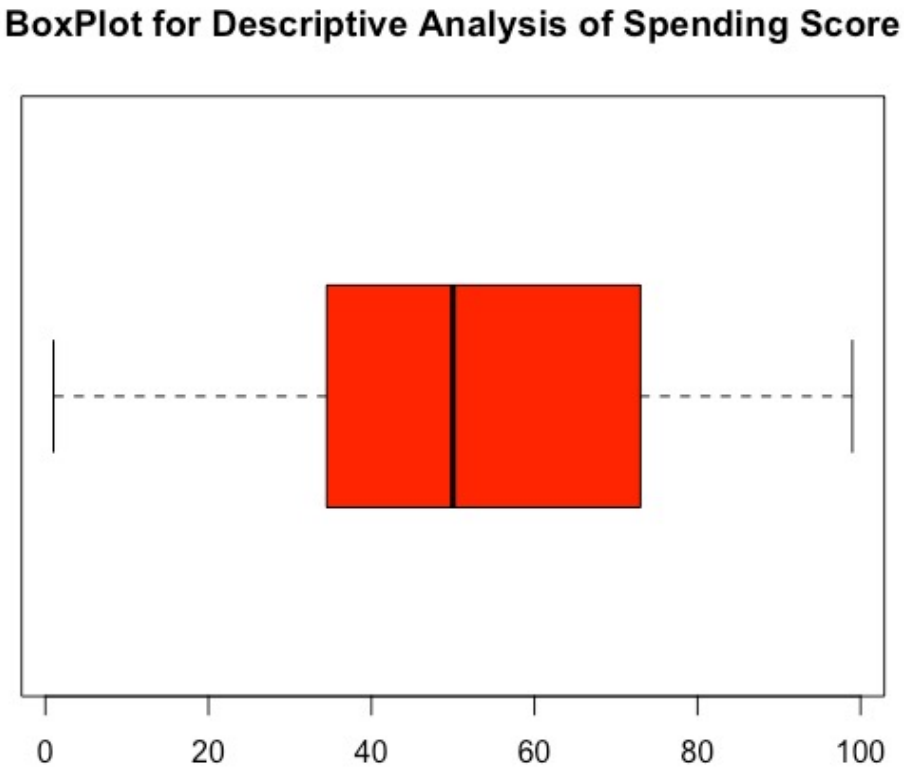


Analyzing the **Spending score** of customers

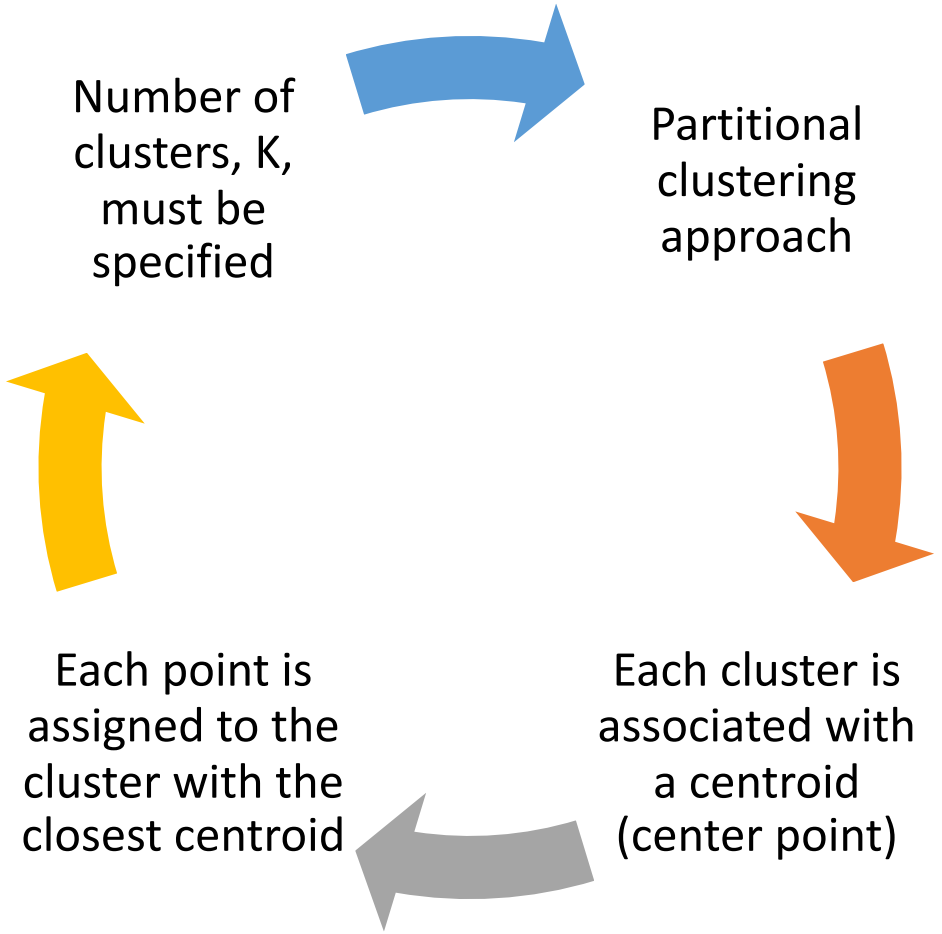


Additional Diagrams of Spending Score Column





K-Means Clustering Model



1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Binary Variable Conversation

Male -> 0

Female -> 1

Clustering Method to find optimal clusters:

1. Elbow Method
2. K Substitution Method
3. Average Silhouette Method

Elbow Method to determine the optimal number of clusters

We can construct the elbow graph and find the optimal k as follow:

Step 1: Construct a function to compute the total within clusters sum of squares

Step 2: Run the algorithm n times

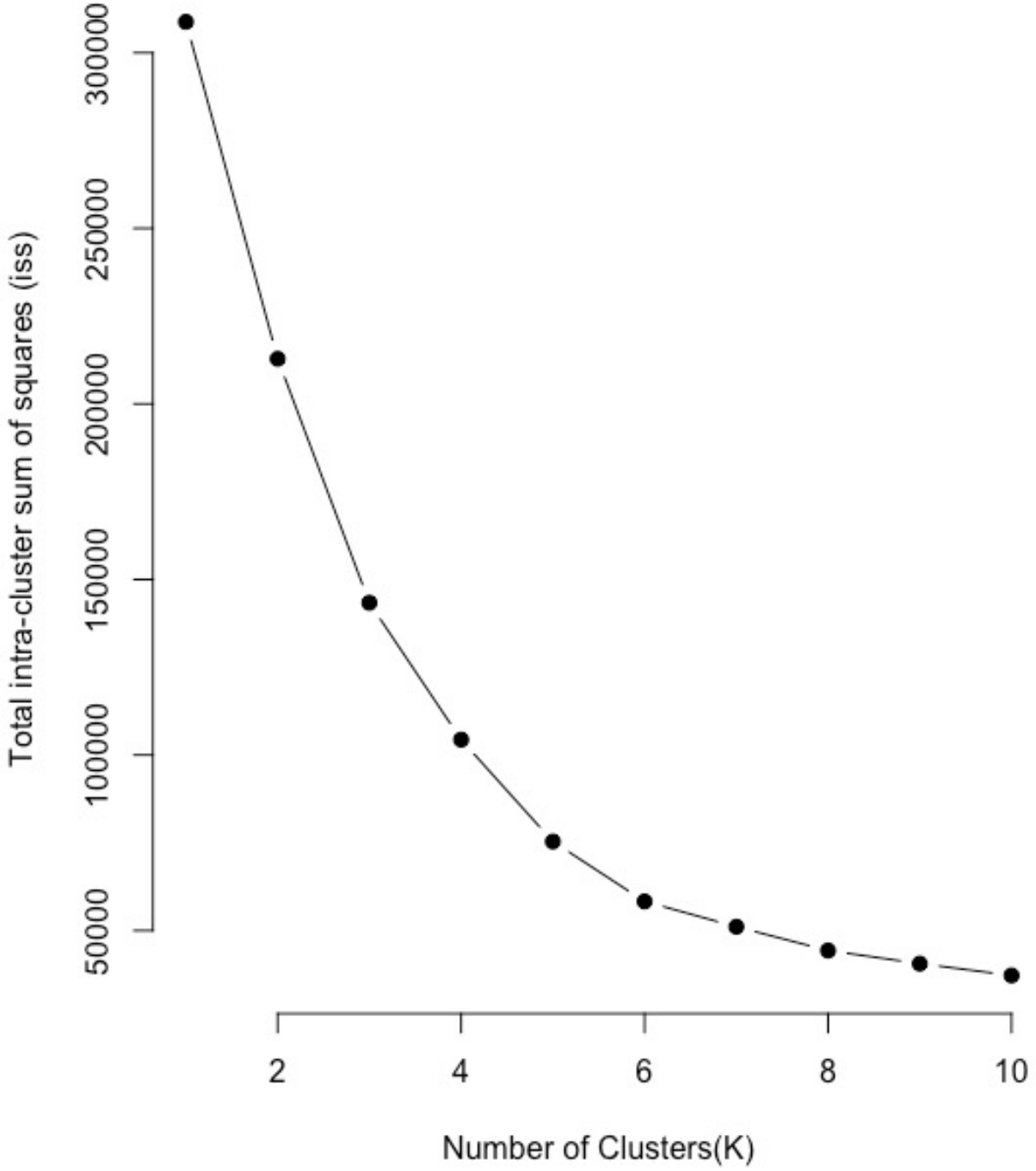
Step 3: Create a data frame with the results of the algorithm

Step 4: Plot the results

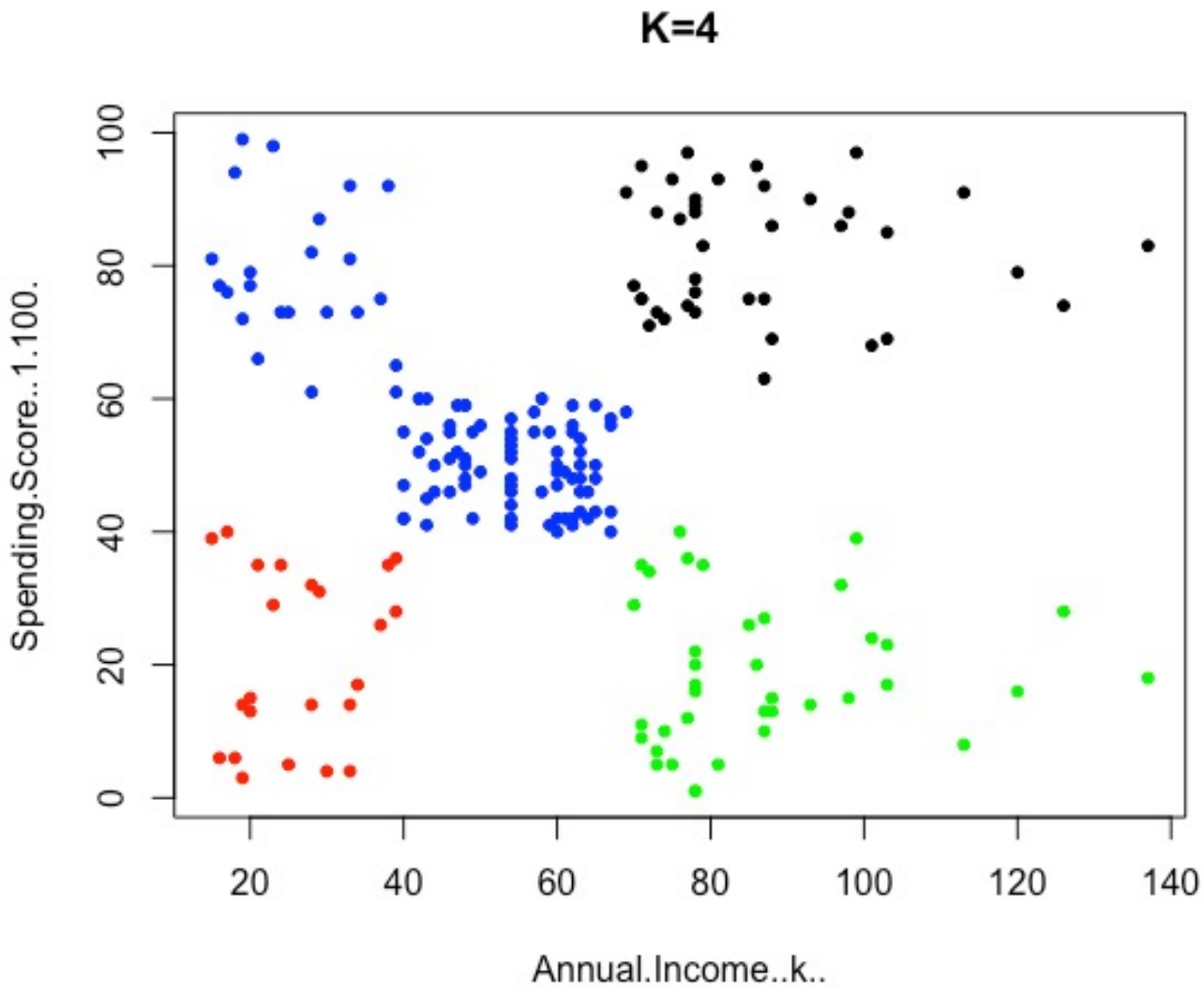
Elbow Method Pseudo Code

```
Using the Elbow Method to determine the optimal number of clusters

#function to calculate the total intra cluster sum of squares

wss = function(k) {

kmeans(customer[,3:6], k, iter.max = 100, nstart = 100, algorithm =
'Lloyd')$tot.withinss

}

k_values = 1:10

wss_values = map_dbl(k_values, wss)
```

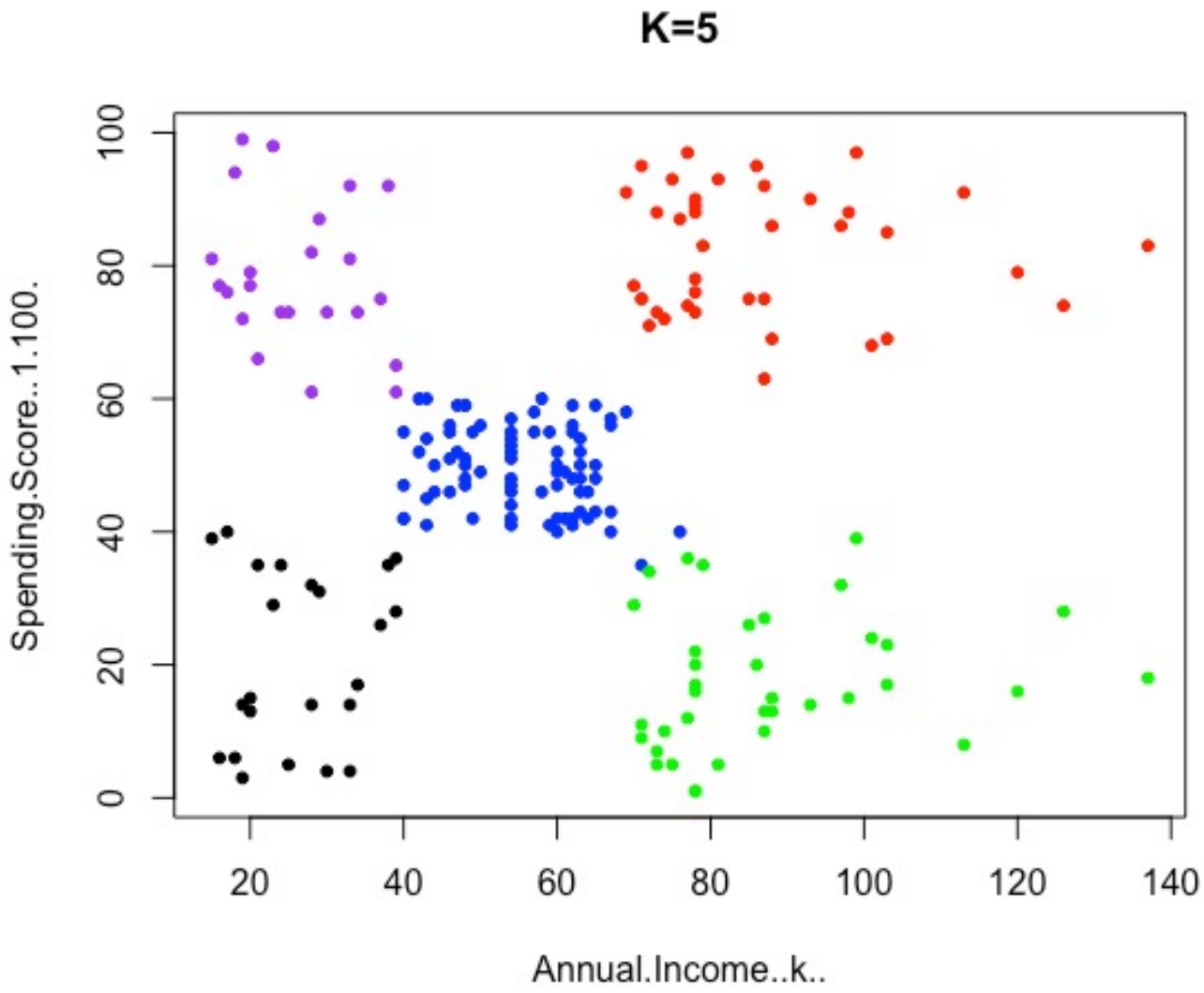
From the graph, it can be seen that the plot has a bend at 4, so we can choose that as the number of cluster.



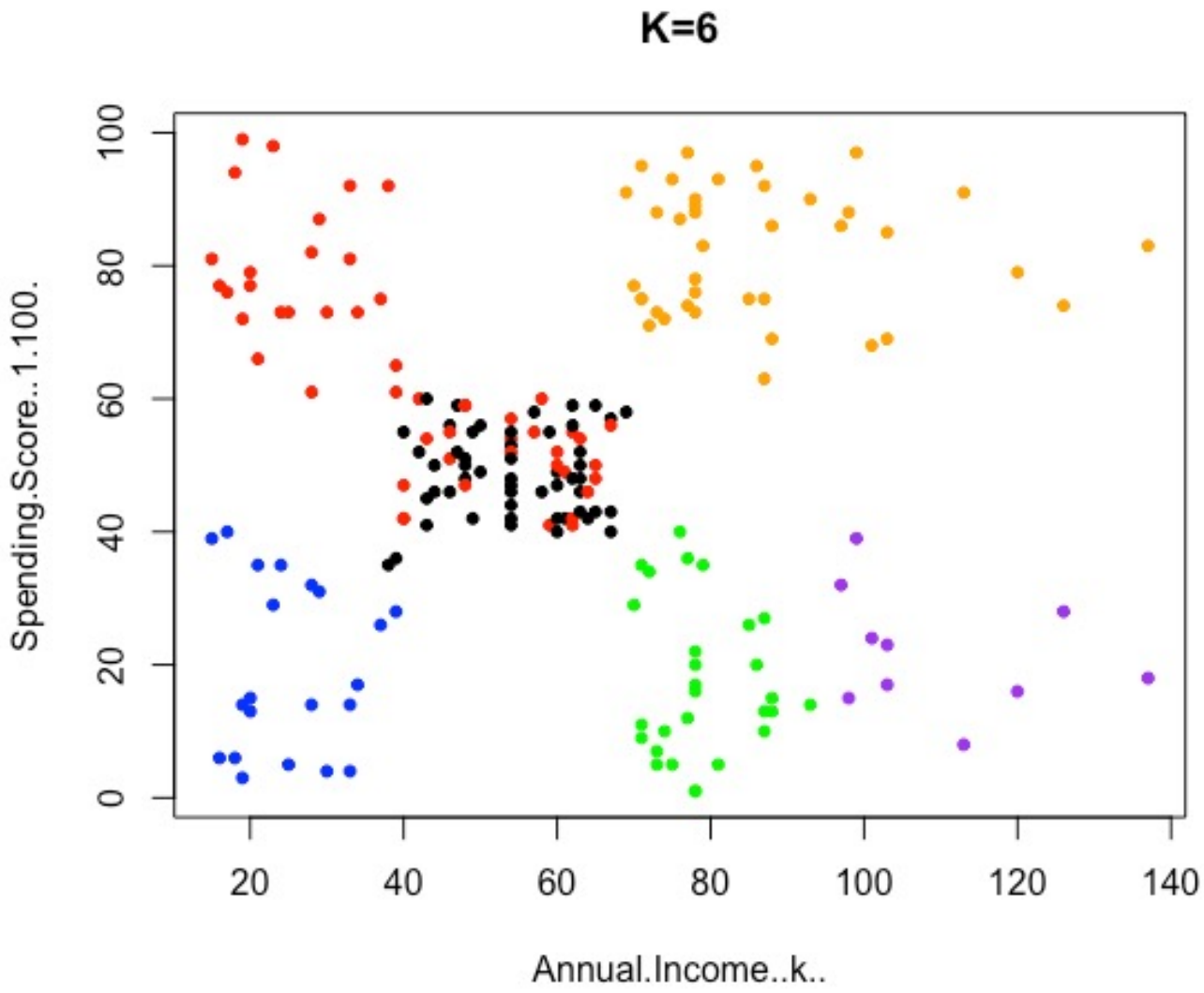
K Substitution Method k=4



K Substitution Method k=5



K Substitution Method k=6



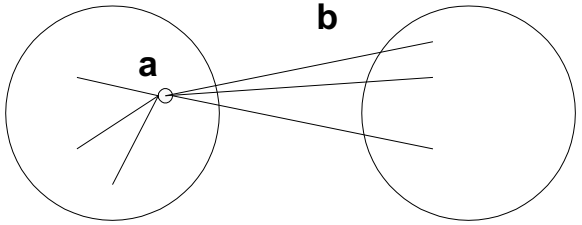
Silhouette Method

Silhouette Coefficient combines ideas of both cohesion and separation

For an individual point, *i*

- Calculate ***a*** = average distance of *i* to the points in its cluster
- Calculate ***b*** = min (average distance of *i* to points in another cluster)
- The silhouette coefficient for a point is then given by
  - Typically, between 0 and 1.
  - The closer to 1 the better.

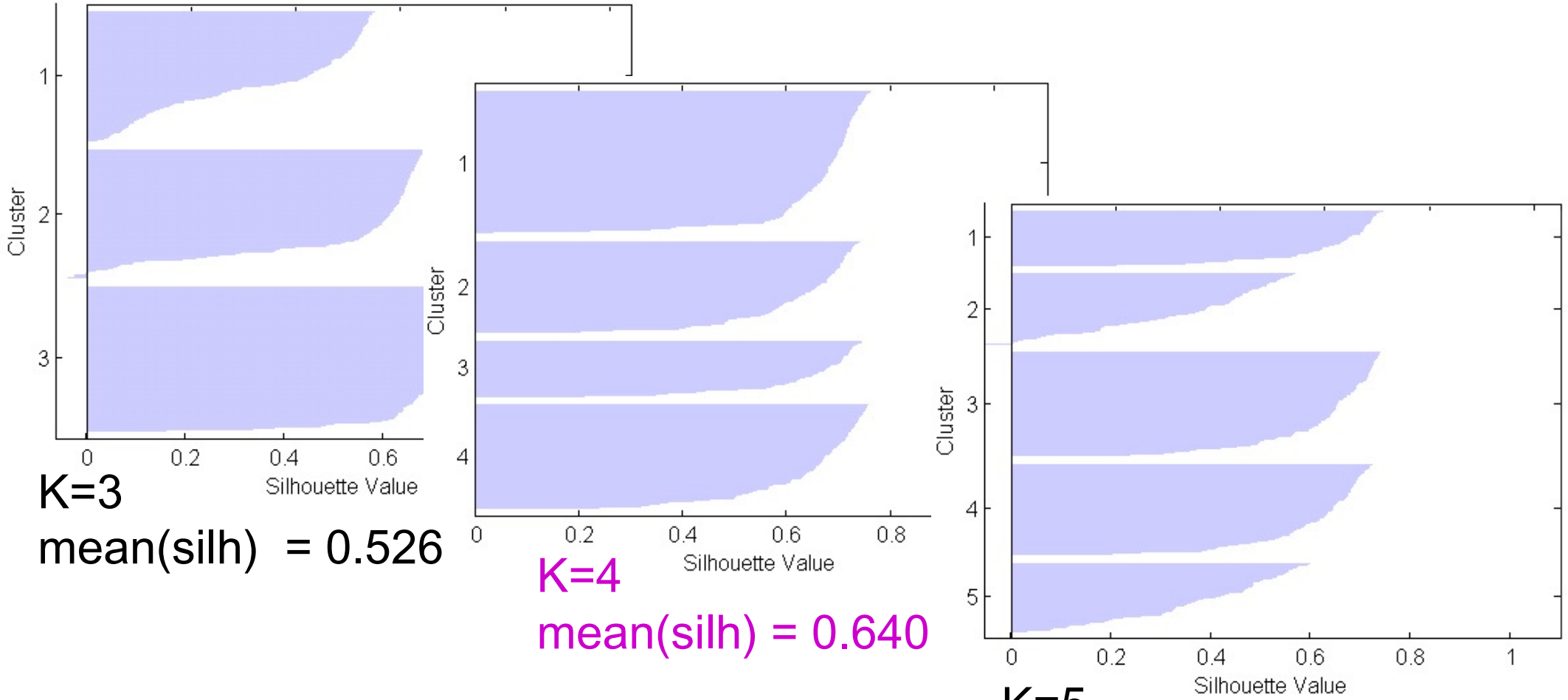
Can calculate the Average Silhouette width for a cluster or a clustering



$$S_i = 1 - \frac{a_i}{b_i} \qquad S_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}$$

Determine number of clusters by Silhouette Coefficient

- compare different clustering's by the average silhouette values



K=3

mean(silh) = 0.526

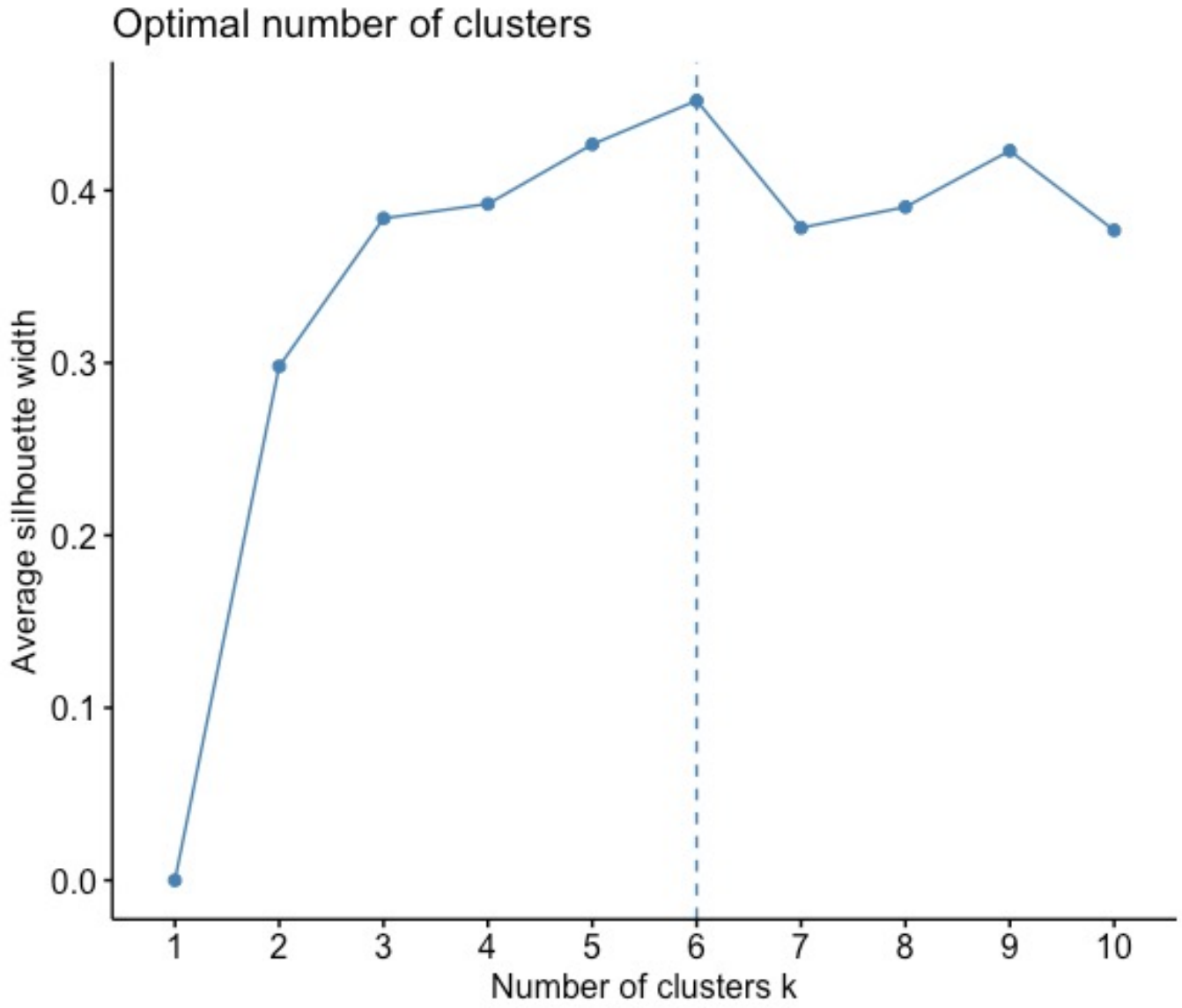
K=4

mean(silh) = 0.640

K=5

mean(silh) = 0.527

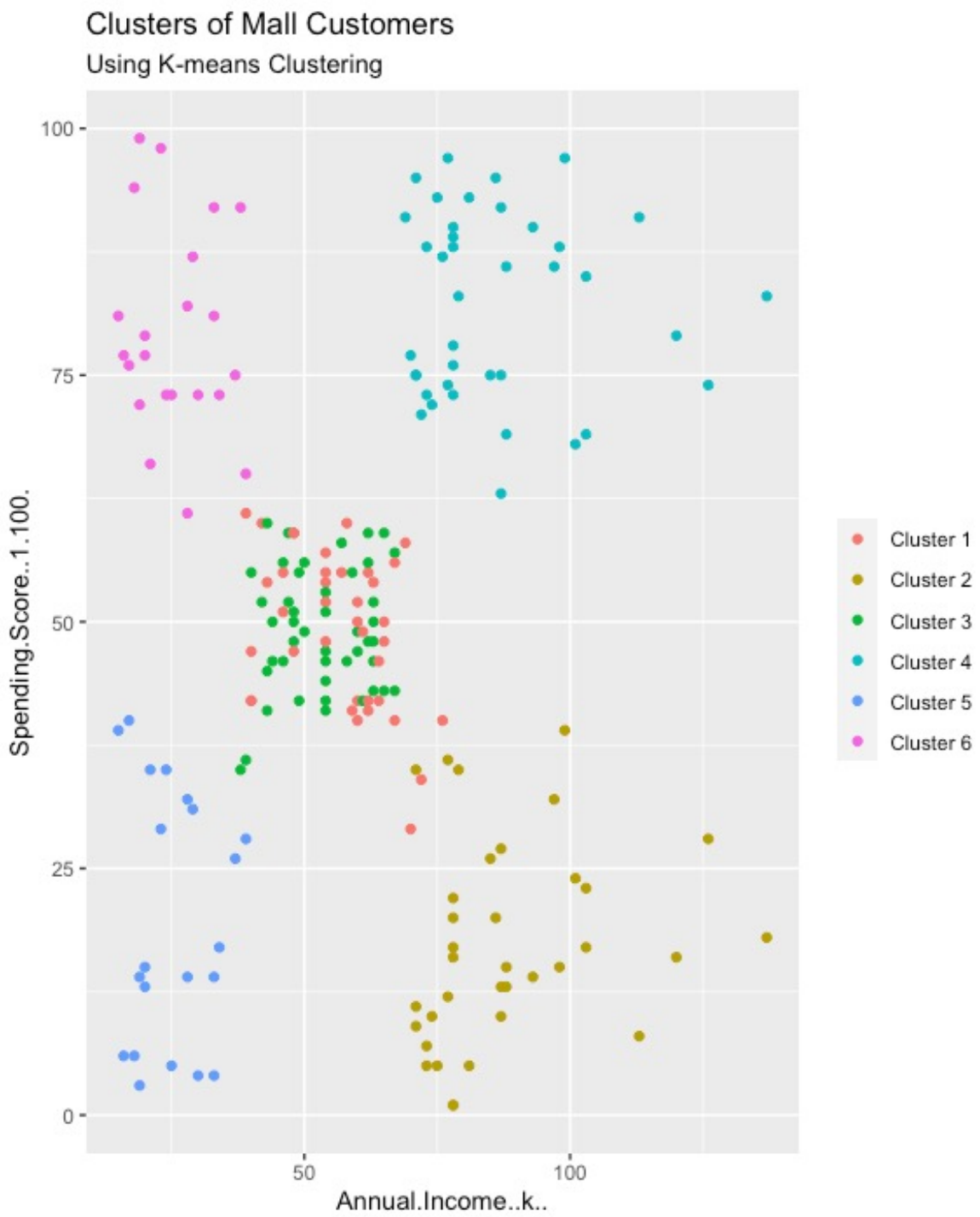
In our case,
For
K=4,S=0.41
K=5,S=0.44
K=6,S=0.45
K=7,S= 0.44
K=9,S=0.39



Average Silhouette Width : 0.45

The plot shows the distribution of the 6 clusters.

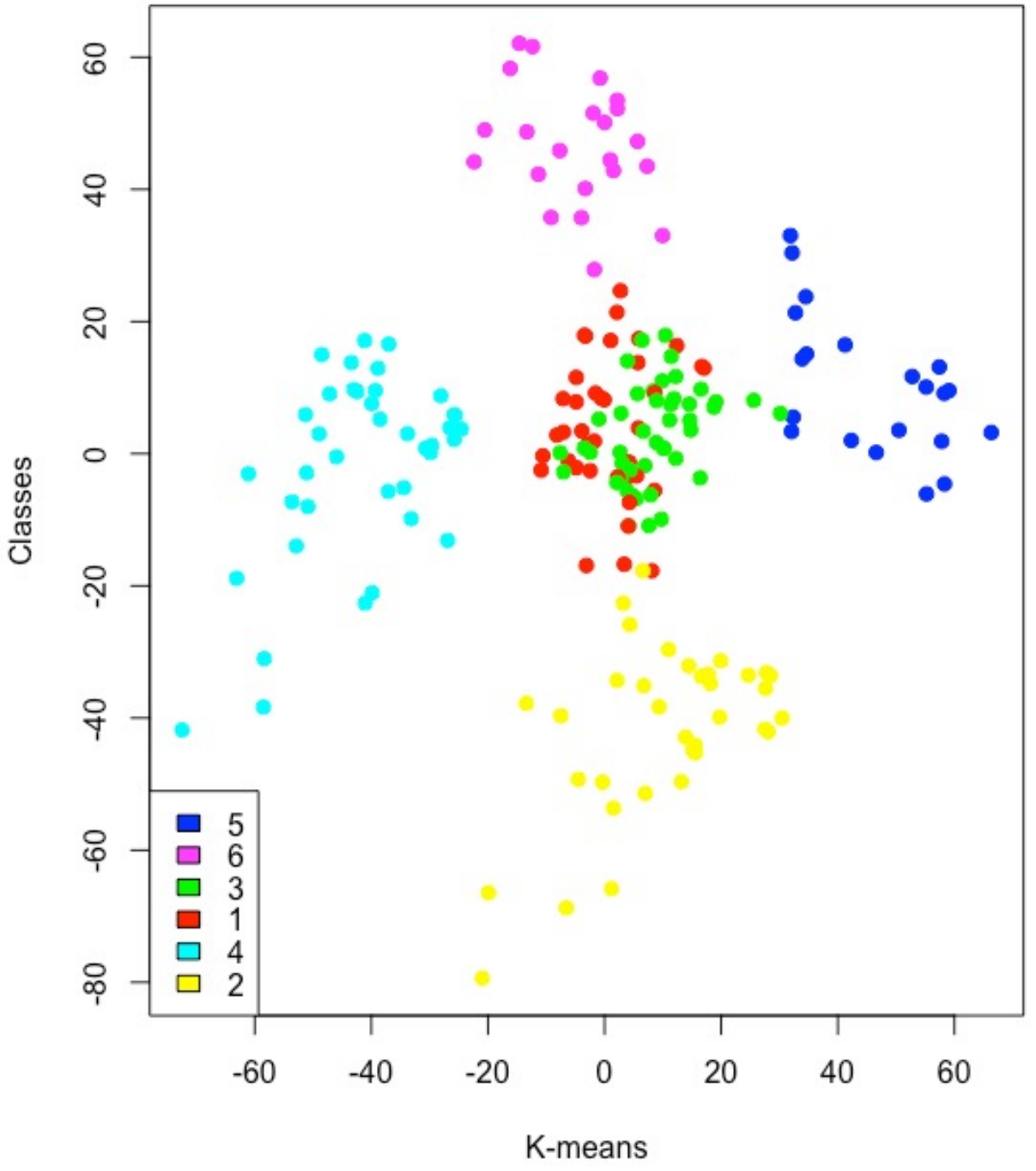
- Cluster 1. Customers with high annual income and high annual spend
- Cluster 2. Customers with high annual income but low annual spend
- Cluster 3. Customers with low annual income and low annual spend
- Cluster 5. Customers low annual income but high annual spend
- Cluster 6 and 4. Customers with medium annual income and medium annual spend

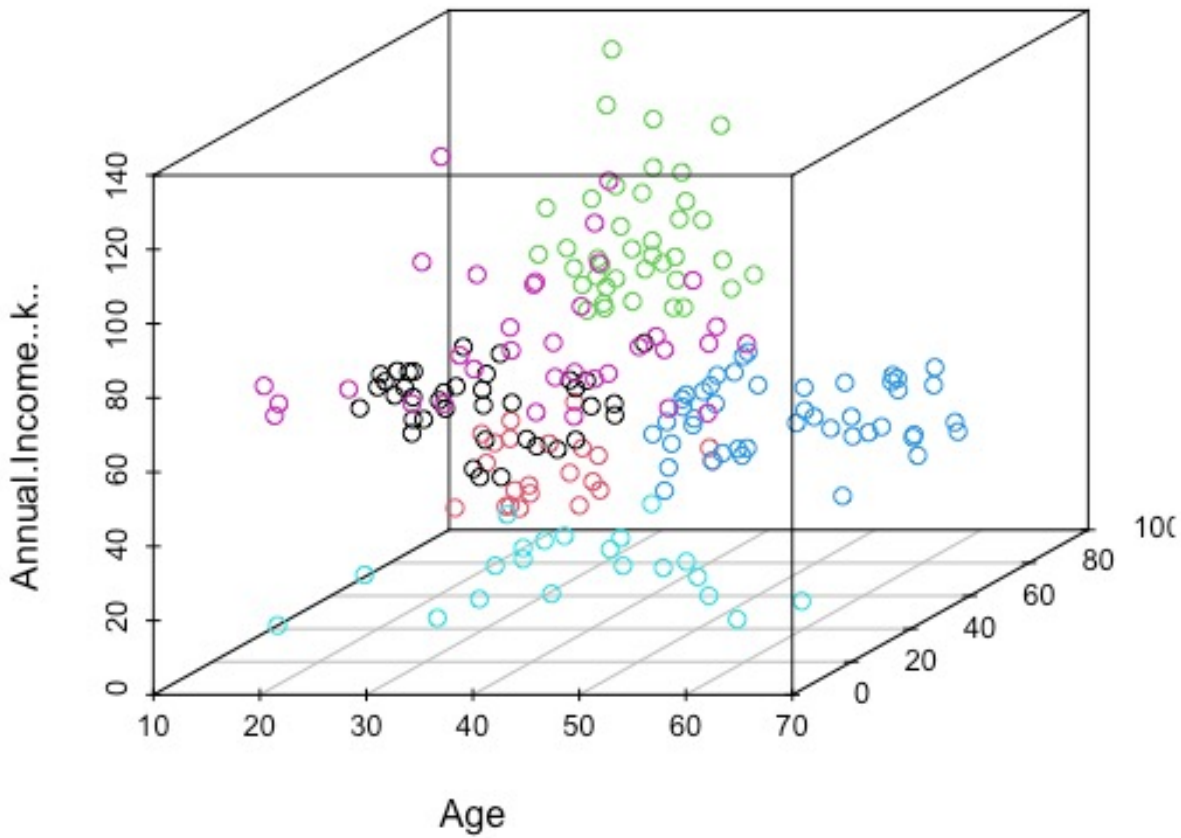


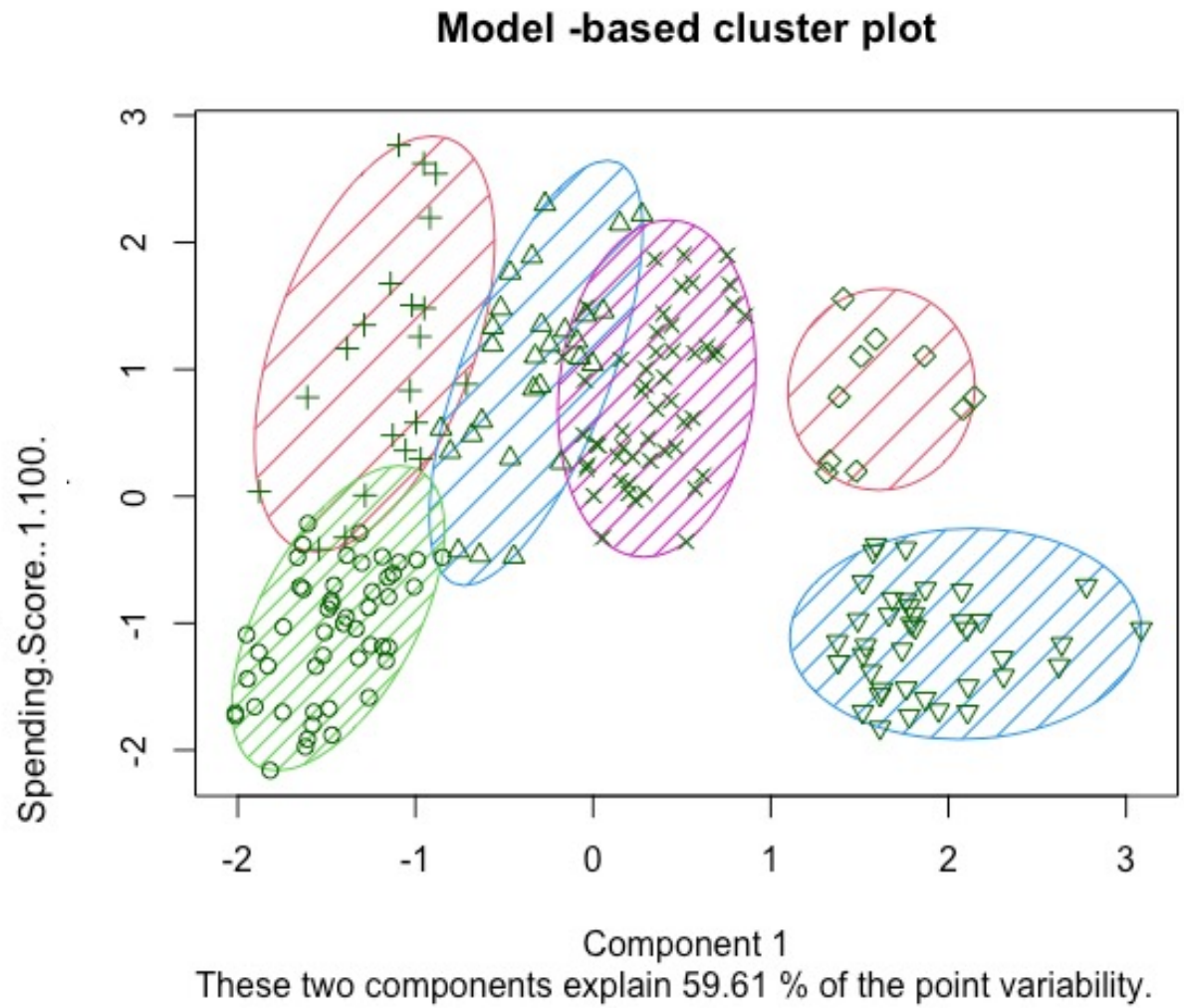
Clustering using PCA

PCA1 Annual Income and PCA2 Spending score

- Cluster 3 and 1. Customers with medium PCA1 and medium PCA2 score
- Cluster 6. Customers with high PCA2 and medium PCA1
- Cluster 5. Customers with high PCA1 and high PCA2 score
- Cluster 3. Customers with high PCA1 income and high PCA2
- Cluster 2. Customers with medium PCA1 and low annual spend PC2




Annual.Income..k..
0
20
40
60
80
100
120
140
10
20
30
40
50
60
70
Age
0
20
40
60
80
100


Model -based cluster plot
Spending.Score..1.100.
3
2
1
0
-1
-2
-2
-1
0
1
2
3
Component 1
These two components explain 59.61 % of the point variability.

Binding K-means cluster to dataset

```
cluster_6 <- kmeans(customer[,3:6], 6)
```

```
results <- cbind(customer,cluster_6$cluster)
```

```
results
```

	CustomerID	Gender	Age	Annual.Income..k..	Spending.Score..1.100.	Gender01	cluster_6$cluster
1	1	Male	19	15	39	0	6
2	2	Male	21	15	81	0	5
3	3	Female	20	16	6	1	6
4	4	Female	23	16	77	1	5
5	5	Female	31	17	40	1	6
6	6	Female	22	17	76	1	5
7	7	Female	35	18	6	1	6
8	8	Female	23	18	94	1	5
9	9	Male	64	19	3	0	6
10	10	Female	30	19	72	1	5
11	11	Male	67	19	14	0	6
12	12	Female	35	19	99	1	5
13	13	Female	58	20	15	1	6
14	14	Female	24	20	77	1	5
15	15	Male	37	20	13	0	6
16	16	Male	22	20	79	0	5
17	17	Female	35	21	35	1	6
18	18	Male	20	21	66	0	5
19	19	Male	52	23	29	0	6
20	20	Female	35	23	98	1	5
21	21	Male	35	24	35	0	6
22	22	Male	25	24	73	0	5
23	23	Female	46	25	5	1	6
24	24	Male	31	25	73	0	5
25	25	Female	54	28	14	1	6
26	26	Male	29	28	82	0	5
27	27	Female	45	28	32	1	6
28	28	Male	35	28	61	0	5
29	29	Female	40	29	31	1	6
30	30	Female	23	29	87	1	5
31	31	Male	60	30	4	0	6
32	32	Female	21	30	73	1	5
33	33	Male	53	33	4	0	6
34	34	Male	18	33	92	0	5
35	35	Female	49	33	14	1	6
36	36	Female	21	33	81	1	5

Conclusion

Clustering helps to better understand the variables to make better decisions.

In our case, there are high levels of income. A more strategic and targeted marketing approach could lift their interest and make them become higher spenders.

The focus should also be on the "loyal" customers and maintain their satisfaction.

References:

- Dataset :https://www.kaggle.com/kandij/mall-customers

- Clustering Method to find optimal clusters: ”Practical Guide To Cluster Analysis in R”, Published by STHDA (http://www.sthda.com), Alboukadel Kassambara,