Leitfaden für nachvollziehbare Schritte

1.Kurze Darstellung des Problembereichs / Aufriss des Themas

1.1Inhaltlich

Kern der Untersuchung: Descriptive and Explorartory Analysis for Customer Segmenttion and Clustering

Grobziele der Arbeit: Today, customers are more than ever at the centre of e-commerce. In times of high competition, long-term customer loyalty as well as the development and maintenance of customer relationships have top priority.

1.2Begründung desThemas

Darstellung der Relevanz des Themas?

Warum ist das Thema wichtig und interessant und daher bearbeitungs- und förderungswürdig?

Marketing segmentation or Customer Segmentation can be defined as the process of Assessing and classifying customer groups to facilitate targeted marketing.There are many reasons why ecommerce stores fail to target the deesired customer , and one of the reason is that there is no adiquate segmentation of existing customers .Also mass marketing will not bring them more sales and customers,which is costly and time consuming .Hence classifying customer based on various ingformation collected can help the owners for consumer understanding and customer satisfaction.

Darstellung eines persönlichen Erkenntnisinteresses.

Dieser Abschnitt soll ein prägnanter Einstieg in die Projektarbeit / Seminararbeit sein.

Er soll beim Leser Interesse für das Thema und die Bereitschaft wecken oder verstärken, die Arbeit zu betreuen bzw. zu fördern und dient der Eigenmotivation.

Customer Segmentation is one of the most important application of unsupervise learning . With the growth of the ecommerce and the compititions in buissness , it has become essential to study the Patterns of shopping and customer behaviour . Companies use the clustering process to foresee or map customer segments with similar behavior to identify and target potential user base.

2.Nachvollziehbare Schritte

2.1Der Stand der Forschung / Auswertung der vorhandenen Literatur / Tutorials ...

Wurde das Problem früher bereits untersucht?

Welche Aspekte wurden untersucht und welche nicht?

Welche Kontroversen gab es und welche Methoden standen bis jetzt im Vordergrund?

Yes the problem has been investigated perviously implimenting different clustering methods. The chanllenges involved in study of the data set is that many attibutes have been not collected which really could help in better segmentation of the customers .

Lösungswege strukturieren!

Wichtigste (verwendete) wissenschaftliche Positionen zum ausgewählten Thema?

(Z.B. **Tutorials …**)

Descriptive Analysis and Exploratory data Analysis is done using the pakages like readr, tidyverse, dplyr, tidyr, ggplot2, janitor and plotly. For Determing and Visualizing the optimal number of clusters pakage factoextra and NbClust are used and for clustering used Kmeans clustering from clusterR pakacge

2.2 Fragestellung

Can Clustering increase the buisseness and attract more customers

2.3Wissenslücke

The data set had very few arrribute to be considered. The factor like the purchase deatils were missing. The location of the customer, the website , the item details , and many more on the basis of which customer segmentation would have been much more effective .

2.4Methode

Detaillierte nachvollziehbare Beschreibung der Vorgehensweise !!

Vgl. MUSTER-PROJEKTE in den Tutorials !!

Used readr package to read the data.Descriptive Analysis and Explortary Analysis is done using dplyr, tidyr, ggplot2, janitor and plotly. Clustering is done using clusterR and Factoextra packages .

2.5 Ergebnisse

Installing Neccessary Packages

```
############### CUSTOMER SEGMENTATION ###########################

######## Importing important Libraries ##########

install.packages("sqldf")
install.packages("plotly")
install.packages("gcc")
install.packages("g++")
install.packages(c("ggplot2"))
install.packages("colorspace")
install.packages("mltools")
install.packages("ClusterR")
install.packages("factoextra")
install.packages("NbClust")
```

Importing Necessary Libraries

```r
#
##### Necessary Libraries ######

library(plotly)
library(tibble)
library(ggplot2)
library(tidyr)
library(tidyverse)
library(readr)
library(ggpubr)
library(ggmap)
#library(sqldf)
library(dplyr )
library(janitor)
library(ClusterR)
library("factoextra")
library(NbClust)
#
```

Reading and Exploring the Data

```r
#
###### Reading and Exploring the Data ########

Customer_data <- read_csv("C:/Users/Vaishu/Desktop/Work/Mall_Customers.csv")
View(Customer_data)
str(Customer_data)
names(Customer_data)
attach(Customer_data)
summary(Customer_data)
sapply(Customer_data,class)
```

Output Screenshot:

```
> Customer_data <- read_csv("C:/Users/Vaishu/Desktop/Work/Mall_Customers.csv")
Rows: 200 Columns: 5
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (1): Gender
dbl (4): CustomerID, Age, Annual Income (k$), Spending Score (1-100)
```

```
> str(Customer_data)
spec_tbl_df [200 × 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ CustomerID            : num [1:200] 1 2 3 4 5 6 7 8 9 10 ...
 $ Gender                : chr [1:200] "Male" "Male" "Female" "Female" ...
 $ Age                   : num [1:200] 19 21 20 23 31 22 35 23 64 30 ...
 $ Annual Income (k$)    : num [1:200] 15 15 16 16 17 17 18 18 19 19 ...
 $ Spending Score (1-100): num [1:200] 39 81 6 77 40 76 6 94 3 72 ...
 - attr(*, "spec")=
  .. cols(
  ..   CustomerID = col_double(),
  ..   Gender = col_character(),
  ..   Age = col_double(),
  ..   `Annual Income (k$)` = col_double(),
  ..   `Spending Score (1-100)` = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

Output Screenshot:

```
> summary(Customer_data)
   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
>
```

Identifying Missing Values

```
#########  Exploratory Data Analysis ##################

#Identifying the missing values

sum(is.na(Customer_data))
lapply(Customer_data,function(x) { length(which(is.na(x)))})
Customer_data1=clean_names(Customer_data)
Customer_data1
names(Customer_data1)
```

Output Screenshot:

```
> sum(is.na(Customer_data))
[1] 0
> lapply(Customer_data,function(x) { length(which(is.na(x)))})
$CustomerID
[1] 0

$Gender
[1] 0

$Age
[1] 0

$`Annual Income (k$)`
[1] 0

$`Spending Score (1-100)`
[1] 0
```

The data contains no missing Value

Exploroing Column

```
#
###### Exploring each column ########

summary(Customer_data1$age)
summary(Customer_data1$annual_income_k)
summary(Customer_data1$spending_score_1_100)

#
```

Output Screenshot:

```
> summary(Customer_data1$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.00   28.75   36.00   38.85   49.00   70.00 

> summary(Customer_data1$annual_income_k)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  15.00   41.50   61.50   60.56   78.00  137.00 
> 
> summary(Customer_data1$spending_score_1_100)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   34.75   50.00   50.20   73.00   99.00 
```

The minimum Age is 18 and the maximum age is 70. Customers Annual income ranges from 15k$ to 137K$. Spending score is between 1 to 100.

Calculating Gender Ratio

```
# ---------------------------------------------------------------------------
##### Gender ratio and Percentage of female to male #####

library(janitor)
gender_ratio <- tabyl(Customer_data1, gender)
gender_ratio

```

Output Screenshot:

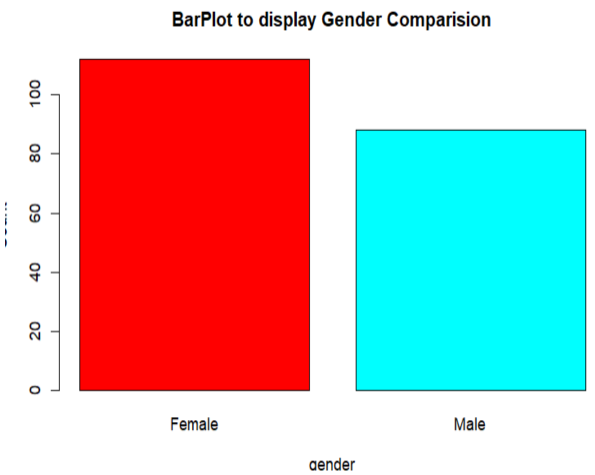
```
> gender_ratio
 gender   n percent
 Female 112    0.56
   Male  88    0.44
```

The number of female customers is 112 which is 56% and male customers is 88 which is 44%.

Female customers are arround 12% more than the male customers.

Visualization of Female vs Male Customer

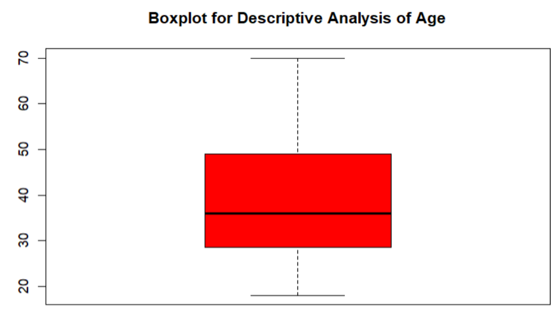
```
###### Data Visualization of Female and Male Customer #########

fig=table(Customer_data1$gender)
b<-barplot(fig,main="BarPlot to display Gender Comparision",
        ylab="Count",
        xlab="gender",
        col=rainbow(2))

```



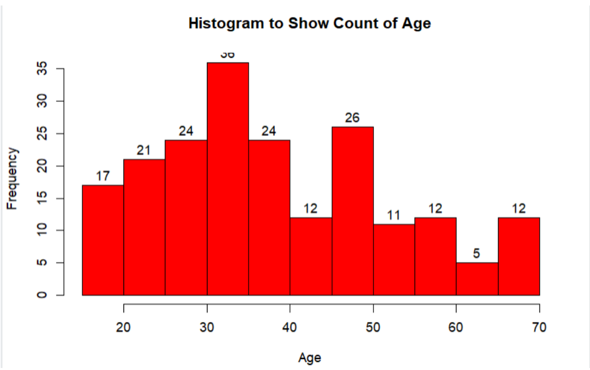
Clearly the graph shows that Female customers are more than male customers.

Exploratory Analysis of Age

```r
#### Descriptive Analysis of age #####

boxplot(Customer_data1$age,
          col="red",
          main="Boxplot for Descriptive Analysis of Age")
```




```r
########  Histogram go show count of Age ########

hist(Customer_data1$age,
     col="red",
     main="Histogram to Show Count of Age",
     xlab="Age",
     ylab="Frequency",
     labels=TRUE)
```

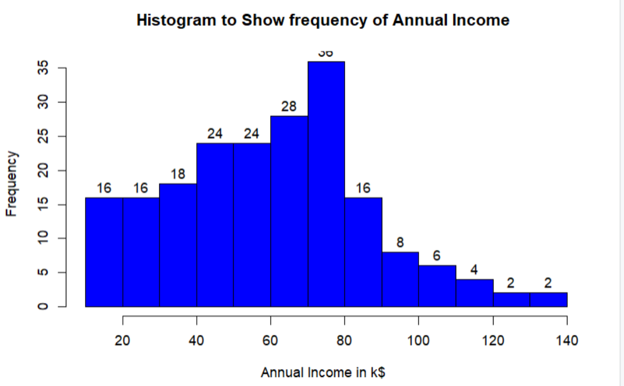


The customer between 30 to 35 age are more compare to other age group.

Also the number of customer are with age less than 20 and more than 50 are comparatively less than other age group.

Frequency of Annual Income

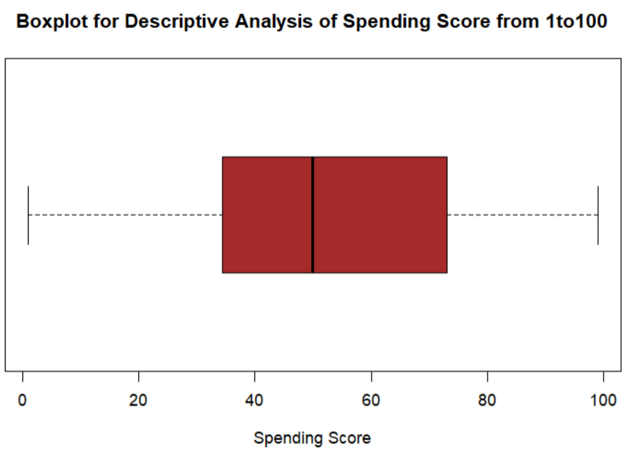
```
####### Visualization of  Annual Income of the Customers #######

hist(Customer_data1$annual_income_k ,
     col="blue",
     main="Histogram to Show Count of Age",
     xlab="Annual Income in k$ ",
     ylab="Frequency",
     labels=TRUE)
```



The graph clearly shows that the good number of customers income range is between 40k$ to 80k$. People earning an average income of 70 have the highest frequency count in #our distribution.

Range of Customers Spending Score

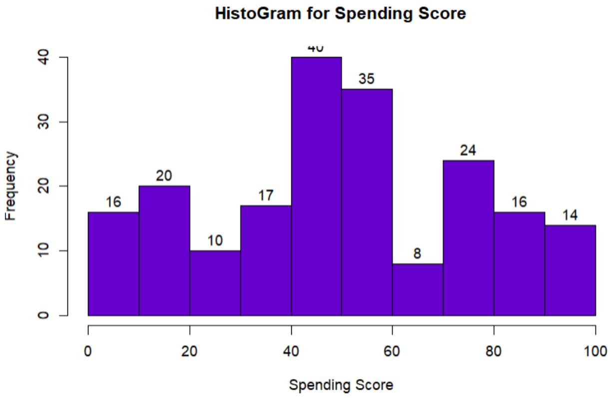
```
######## Visualization of customer Spending Score ########

boxplot(Customer_data1$spending_score_1_100,
        horizontal=TRUE,
        xlab = "Spending Score",
        col="brown",
        main="Boxplot for Descriptive Analysis of Spending Score from 1to100")
```



Most of the customers spending Score is between 40 to 70.

Frequency of Custmer Spending Score

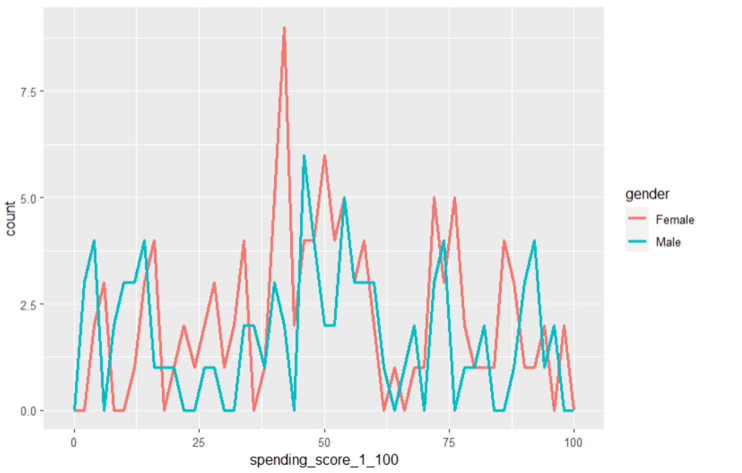
```
###### Visualization of Frequency of Customer Spending Score ######

hist(Customer_data1$spending_score_1_100,
     main="HistoGram for Spending Score",
     xlab="Spending Score",
     ylab="Frequency",
     col="#6600cc",
     labels=TRUE)
```



The minimum spending score is 1, maximum is 99 .Customers between class 40 and 50 have the highest spending score among all the classes.

Spending Score Male vs Female Customers

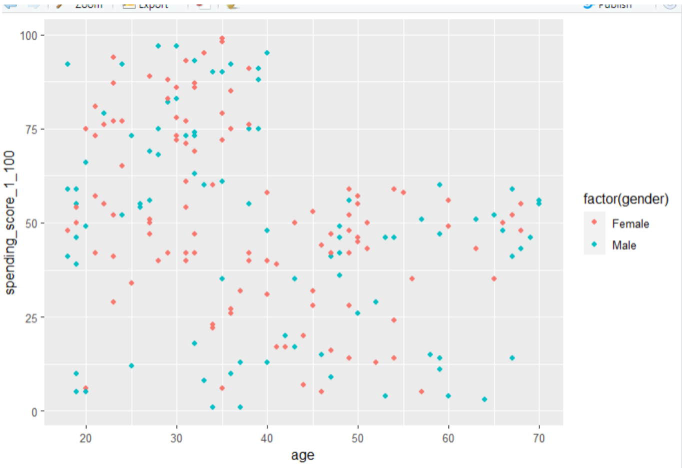
```
#########  Spending Score Male Vs Female with respect to Age #####

g<-ggplot(Customer_data1,aes(x= `spending_score_1_100`, col=gender)) +
      geom_freqpoly(bins=50, size=1)
g

p<-ggplot(Customer_data1, aes(age)) +geom_bar(aes(fill = gender))
p
```



The spending score of female customers is quite high than male customers.
Also in all age group the female customers are more than the male customers

Spending Score of customers wrt Age

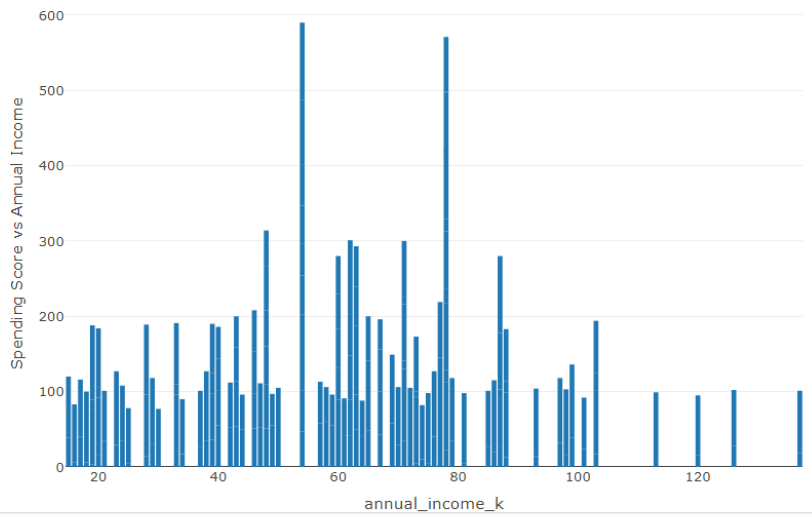
```
#### Spending Score vs Age ######

p <- ggplot(Customer_data1, aes(age,spending_score_1_100 ))+
     geom_point(aes(colour = factor(gender)))
p

```



From the plot we can say that spending score is more in the age group 20 to 40 than in the age group 45 to 70. But no pattern can be seen with respect to gender here.

Visualization of customers Annual Income vs Spending Score

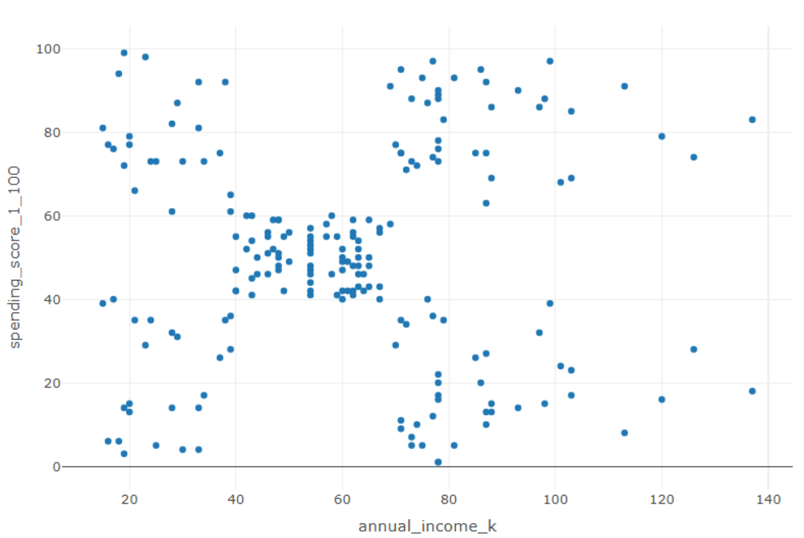
```
##### Descriptive Analysis of spending Score with respect to Annual Income #####


fig <- plot_ly(Customer_data1, x = ~annual_income_k, y = ~spending_score_1_100, type = 'bar')

fig <- fig %>% layout(yaxis = list(title = 'Spending Score vs Annual Income'))

fig
```



The customers with annual income 54k$ and 78k$ have the highest spending score than any other income group

Visualization of Distribution of Customers

```
######### Scatter plot to see the distribution of cutomers #########

fig <- plot_ly(data = Customer_data1, x = ~annual_income_k, y = ~spending_score_1_100)
fig
```



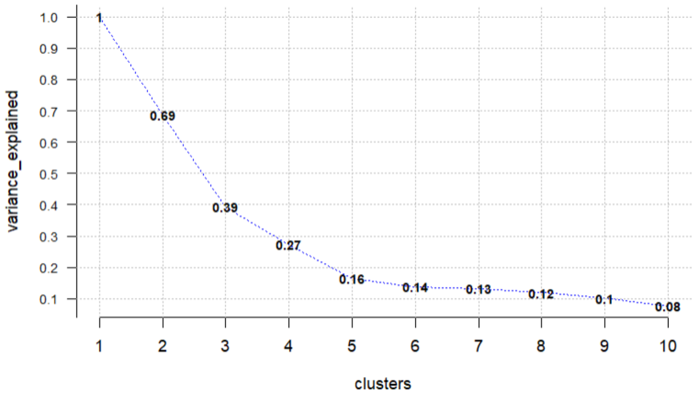
From the above scatter plot , the data seems to hold a pattern here. It seems there may have 5 category of customers .

```
#
##### One hot encoding for gender column ######

library(caret)

dummy <- dummyVars(" ~ .", data=Customer_data1)
new_data <- data.frame(predict(dummy, newdata = Customer_data1))
new_data
new_data <- subset (new_data, select = -customer_id )
new_data

new_data1=clean_names(new_data)
new_data1
```

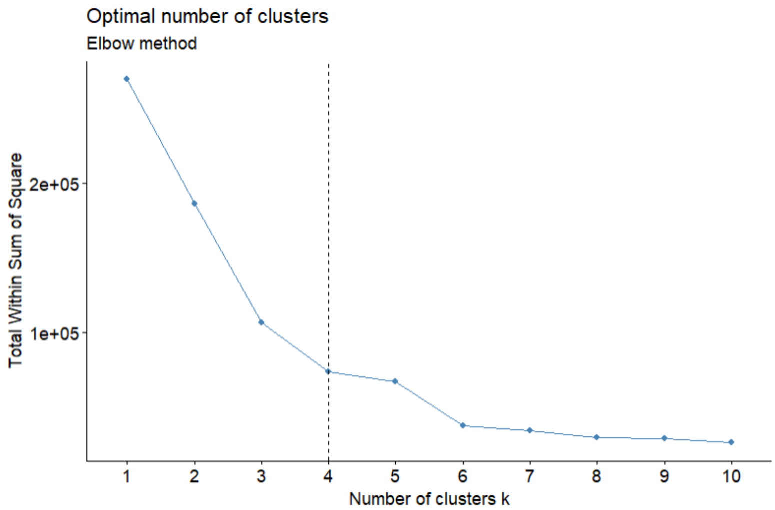
There was only one categorical column in our data set , that is Gender . Using caret package one hot encoding is been done.

Determining Optimal Clusters with Graphs

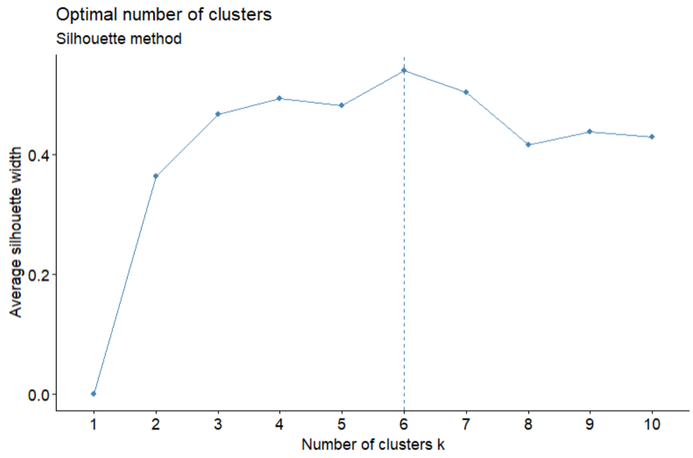
```r
######## To find the optimal number of cluster #########

###### Elebow Method  using ClusterR Library ########

opt <- Optimal_Clusters_KMeans(new_data1[,4:5], max_clusters = 10,max_iters=200,
                               plot_clusters = T)

opt
```



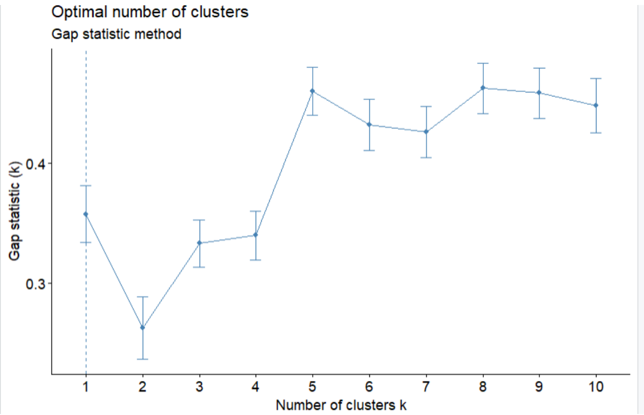

```r
##### Elbow method #######

fviz_nbclust(new_data1[,4:5], kmeans, method = "wss") +
  geom_vline(xintercept = 4, linetype = 2)+
  labs(subtitle = "Elbow method")
```




```r
####### Silhouette method #######
fviz_nbclust(new_data1[,4:5], kmeans, method = "silhouette")+
  labs(subtitle = "Silhouette method")
```




```
######## Gap statistic ############

# nboot = 50 to keep the function speedy.
# recommended value: nboot= 500 for your analysis.
# Use verbose = FALSE to hide computing progression.
set.seed(123)
fviz_nbclust(new_data1[,4:5], kmeans, nstart = 25,  method = "gap_stat", nboot = 50)+
  labs(subtitle = "Gap statistic method")
```



Elbow method: 4 clusters solution suggested

Silhouette method: 6 clusters solution suggested

Gap statistic method: 1 clusters solution suggested

From the scatterplots and with silhouette method it would be ideal to consider 6 clusters

```
# Taking k  as 6 due to all the above  analysis above

Clustered_data<-kmeans(new_data1[,4:5],centers=6, iter.max = 20,nstart = 1, algorithm ="Lloyd" )
print(Clustered_data)
```

Output Screenshot:

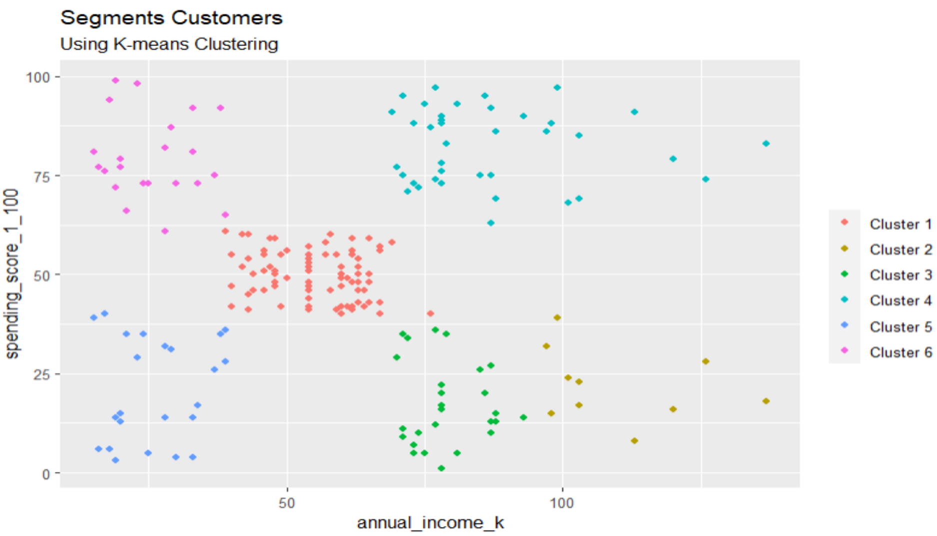
```
K-means clustering with 6 clusters of sizes 15, 19, 22, 4, 39, 101

Cluster means:
  annual_income_k spending_score_1_100
1        77.93333             9.00000
2        86.36842            26.47368
3        25.72727            79.36364
4       124.00000            17.50000
5        86.53846            82.12821
6        48.16832            43.39604
```

Output Screenshot:

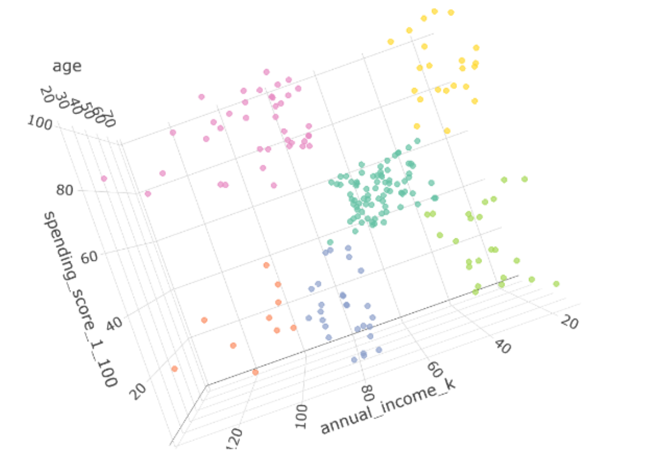
```
Within cluster sum of squares by cluster:
[1]   784.9333  3625.1579  3519.4545   513.0000 13444.0513 42712.2970
 (between_SS / total_SS =  76.1 %)
```

Visualizing the Clustering Results 2D and 3D

```
###### Ploting the Clusters ########

set.seed(1)
ggplot(new_data1[,4:5], aes(x =annual_income_k, y = spending_score_1_100)) +
 geom_point(stat = "identity", aes(color = as.factor(Clustered_data$cluster))) +
 scale_color_discrete(name=" ",
            breaks=c("1", "2", "3", "4", "5","6"),
            labels=c("Cluster 1","Cluster 2","Cluster 3","Cluster 4","Cluster 5","Cluster 6")) +
  ggtitle("Segments Customers", subtitle = "Using K-means Clustering")


```



```
# 3D Graph
p <- plot_ly(new_data1, x=~annual_income_k, y=~spending_score_1_100,
             z=~age, color=as.factor(Clustered_data$cluster)) %>%
  add_markers(size=1.5)
print(p)
```



Interpretation for the customer cluster/segment:

Cluster 1. Customers with medium annual income and medium spending score.

Cluster 2. Customers with high annual income but low spending score.

Cluster 3. Customers with 75k$ annual income and low spending score.

Cluster 4. Customers with arround 75k$ annual income and high spending score.

Cluster 5. Customers with low annual income and low spending score.

Cluster 6. Customers with low annual income but high spending score.

We could see from the EDA part that the female customers percentage (56%) is slightly higher than male customers (44%),with this information we could target the male customers more for marketing campaign or promotions though the percentage different is not too big. We can doing marketing campaigns/loyalty program to customer who has high spending score . For the customers with high annual income but low spending scores we cantry adding some more offers and more brands which are popular among that age group.

2.7 Ausblick

Whether its a small-scale, medium-scale, or a large-scale ecommerce business, understanding customers is ultimate key to success. Customer segmentation can help the buisseness owner's to learn more about particular set of customers . And at the same time taking care of customers satisfaction.