Customer Segmentation using Machine Learning

Briana Moses

July 2021

Introduction

This project is a market basket analysis or customer segmentation of mall customers. I explored the mall customer dataset in this project by developing box plots and histograms and created a k-means clustering algorithm to view customer segments.

Data Source

For this project, I used a mall customer dataset provided on Kaggle. This data set contains mall customers' ID, gender, age, annual income, and spending score.

Methodology:

1.Installing packages:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
```

2. Load data set

```
mall_customers <- read_csv("Mall Customers Data/Mall_Customers.csv")
```

```
## Rows: 200 Columns: 5

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): CustomerID, Genre
## dbl (3): Age, Annual Income (k$), Spending Score (1-100)

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

3. Data Structure

```
str(mall_customers)
```

```
## spec_tbl_df [200 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ CustomerID            : chr [1:200] "0001" "0002" "0003" "0004" ...
##  $ Genre                 : 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_character(),
##   ..   Genre = col_character(),
##   ..   Age = col_double(),
##   ..   `Annual Income (k$)` = col_double(),
##   ..   `Spending Score (1-100)` = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
head(mall_customers)
```

```
## # A tibble: 6 x 5
##   CustomerID Genre    Age `Annual Income (k$)` `Spending Score (1-100)`
##   <chr>      <chr>  <dbl>                <dbl>                    <dbl>
## 1 0001       Male      19                   15                       39
## 2 0002       Male      21                   15                       81
## 3 0003       Female    20                   16                        6
## 4 0004       Female    23                   16                       77
## 5 0005       Female    31                   17                       40
## 6 0006       Female    22                   17                       76
```

```
sd(mall_customers$Age)
```

```
## [1] 13.96901
```

```
sd(mall_customers$`Annual Income (k$)`)
```

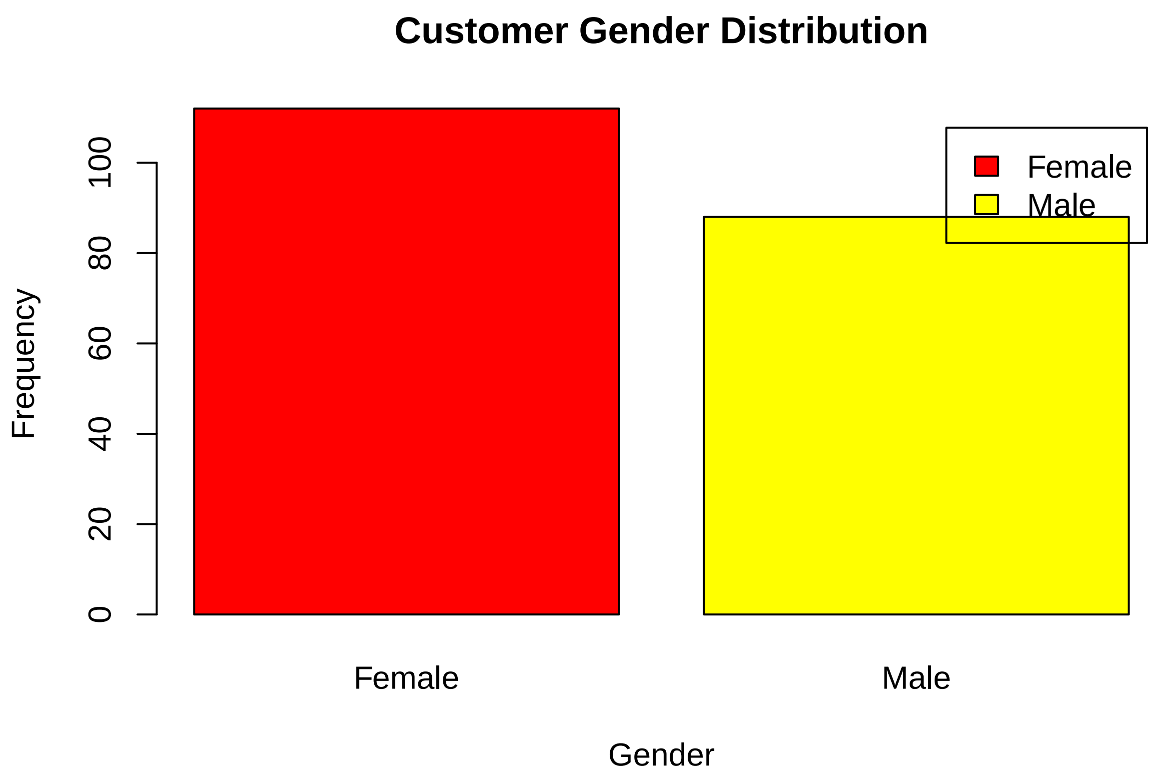
```
## [1] 26.26472
```

```
sd(mall_customers$`Spending Score (1-100)`)
```

```
## [1] 25.82352
```

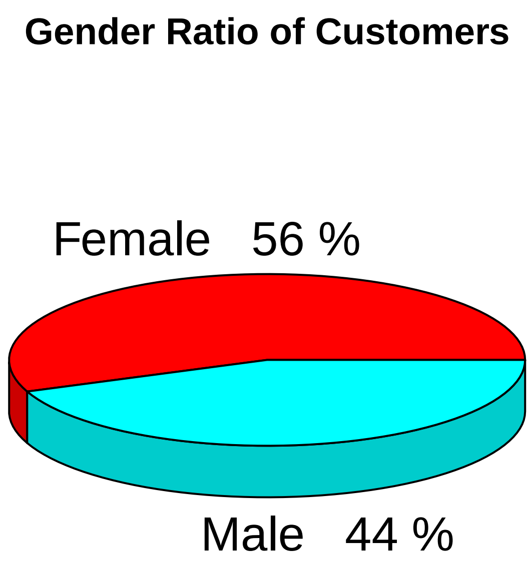
4. Descriptive Analysis

```
counts <- table(mall_customers$Genre)
barplot(counts, main = "Customer Gender Distribution", xlab = 'Gender',
        ylab ='Frequency', col = rainbow(6), legend = rownames(counts))
```




```
library(plotrix)
percentage = round(counts/sum(counts)*100)
name_lbs = paste(c("Female", "Male"), " ", percentage, "%", sep=" ")

pie3D (counts,labels = name_lbs, main = "Gender Ratio of Customers")
```

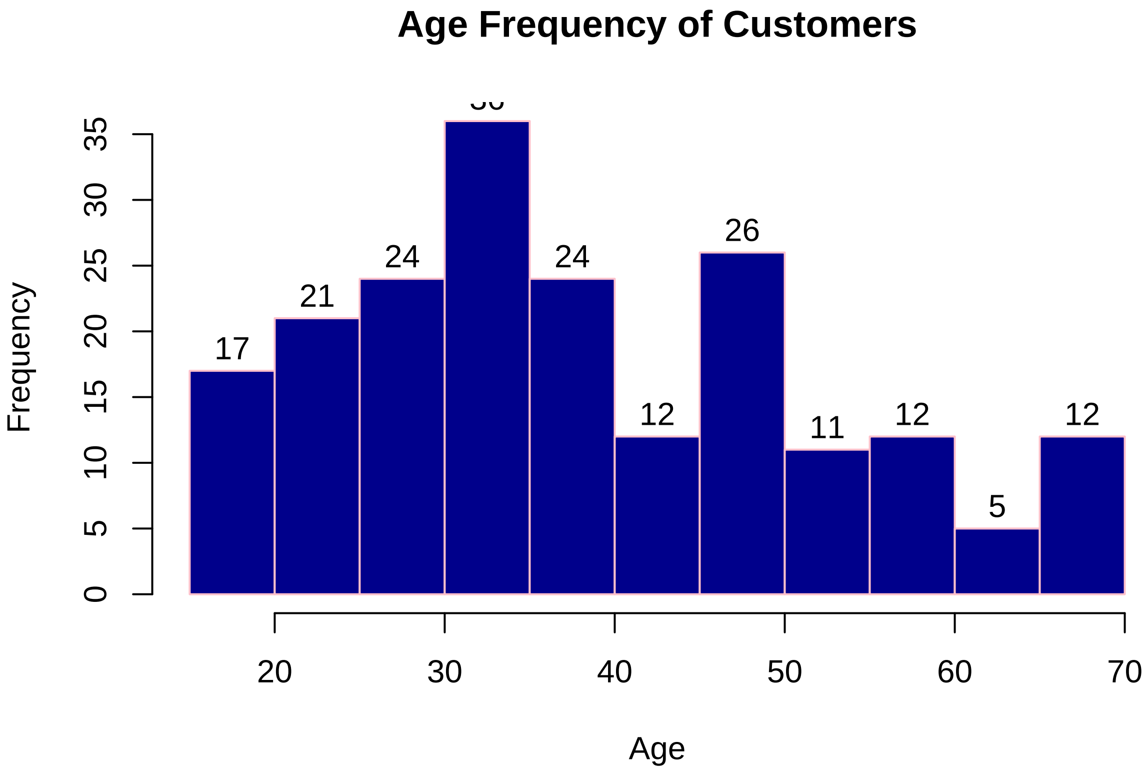


Results: There are more female customers within the data set than males.

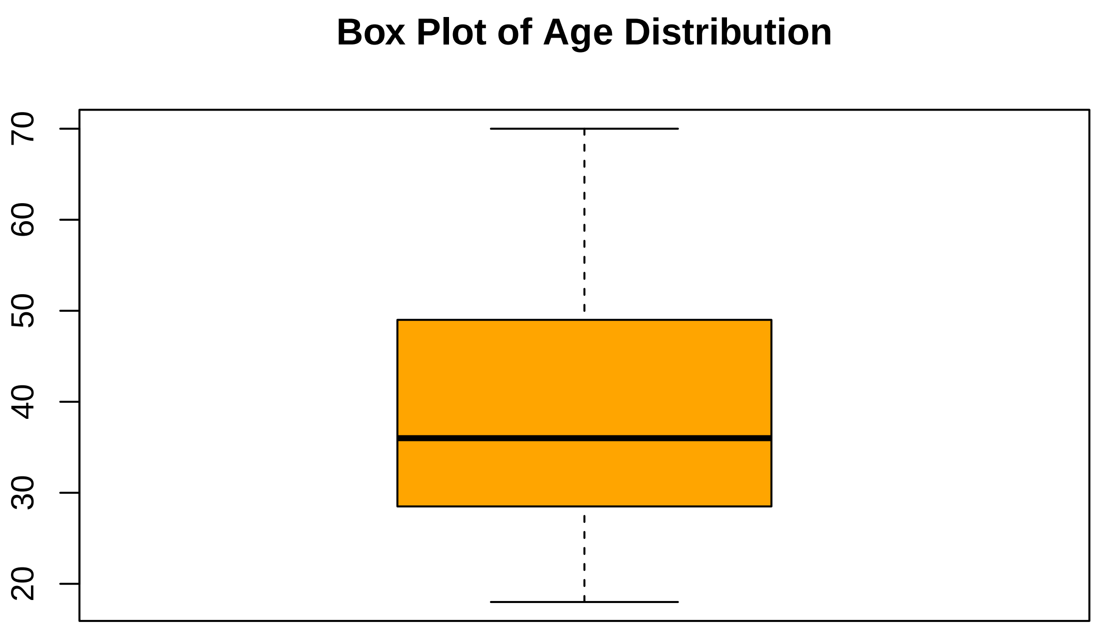
```
summary(mall_customers$Age)
```

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

```r
hist(mall_customers$Age,
                 col = "darkblue",
                 border = "pink",
                 main = "Age Frequency of Customers",
                 xlab = "Age",
                 ylab = "Frequency",
                 labels = TRUE)
```




```r
boxplot(mall_customers$Age,col = "orange", main = "Box Plot of Age Distribution")
```

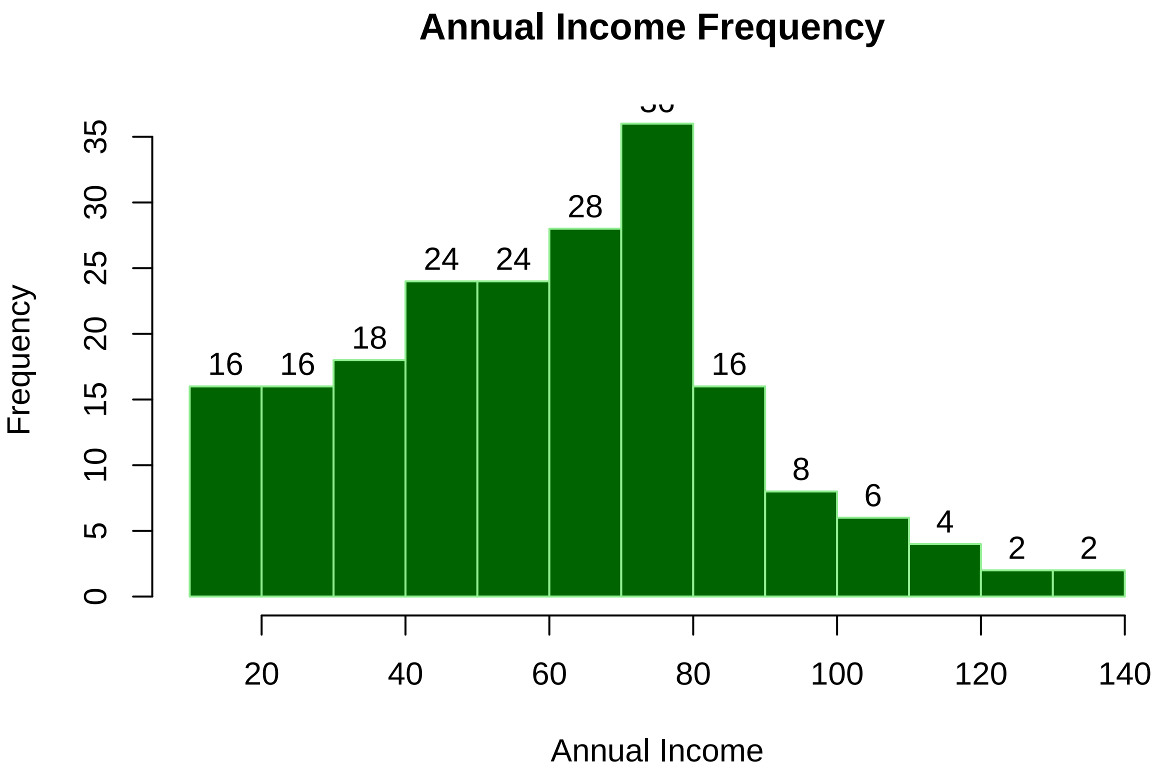


Results: The maximum age is 70 while the minimum age is 18. The maximum number of customers in the data set are between the ages of 30 and 35.

```
summary(mall_customers$`Annual Income (k$)`)
```

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

```
hist(mall_customers$`Annual Income (k$)`,
                    col = 'darkgreen',
                    border = "lightgreen",
                    main = "Annual Income Frequency ",
                    xlab = "Annual Income",
                    ylab = "Frequency",
                    labels = TRUE)
```

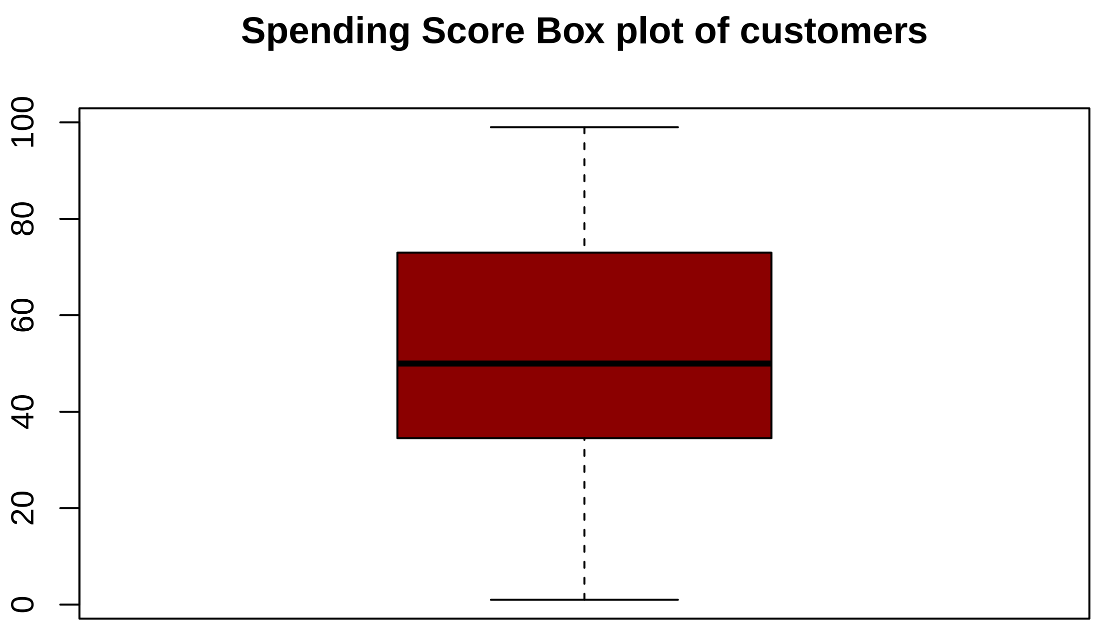


Results: The maximum annual income is approx. 140, while the minimum is 15. The maximum annual income frequency of customers is approx. 70.

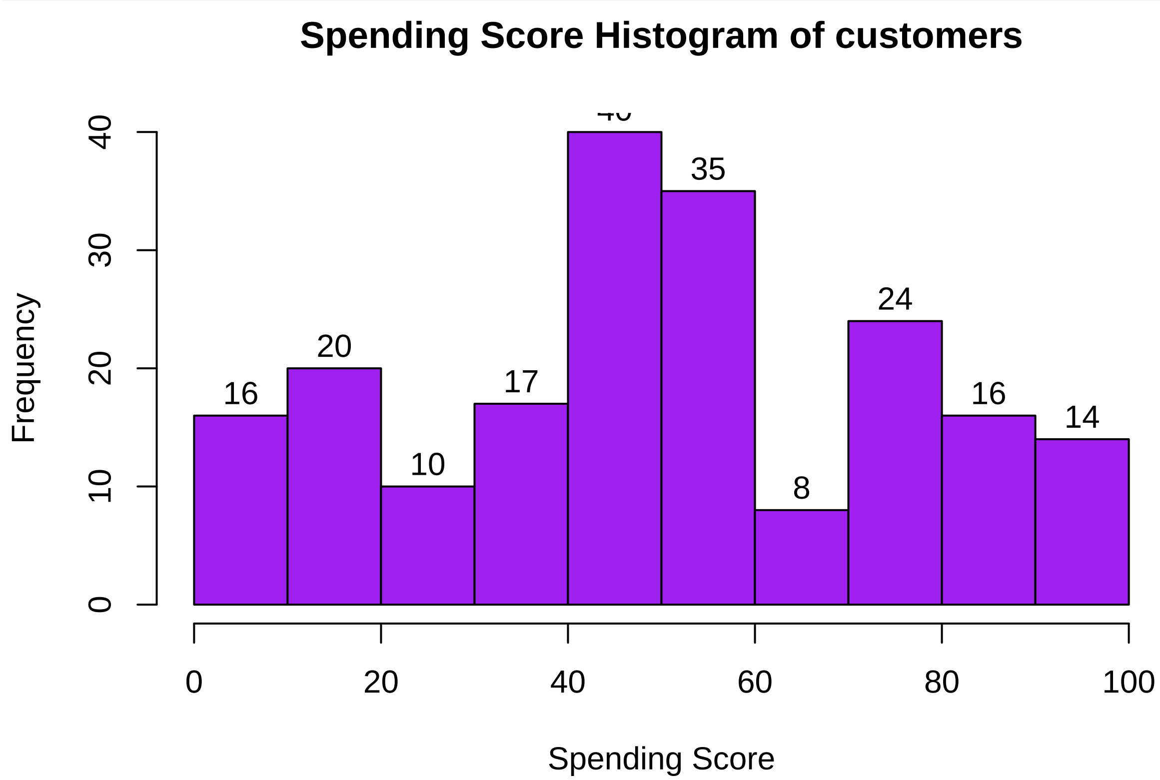
```
summary(mall_customers$`Spending Score (1-100)`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   34.75   50.00   50.20   73.00   99.00
```

```
boxplot(mall_customers$`Spending Score (1-100)`,
                            col = "darkred",
                            main = "Spending Score Box plot of customers")
```




```r
hist(mall_customers$`Spending Score (1-100)`,
                          col = "purple",
                          border = "black",
                          main = "Spending Score Histogram of customers",
                          xlab = "Spending Score",
                          ylab = "Frequency",
                          labels = TRUE)
```



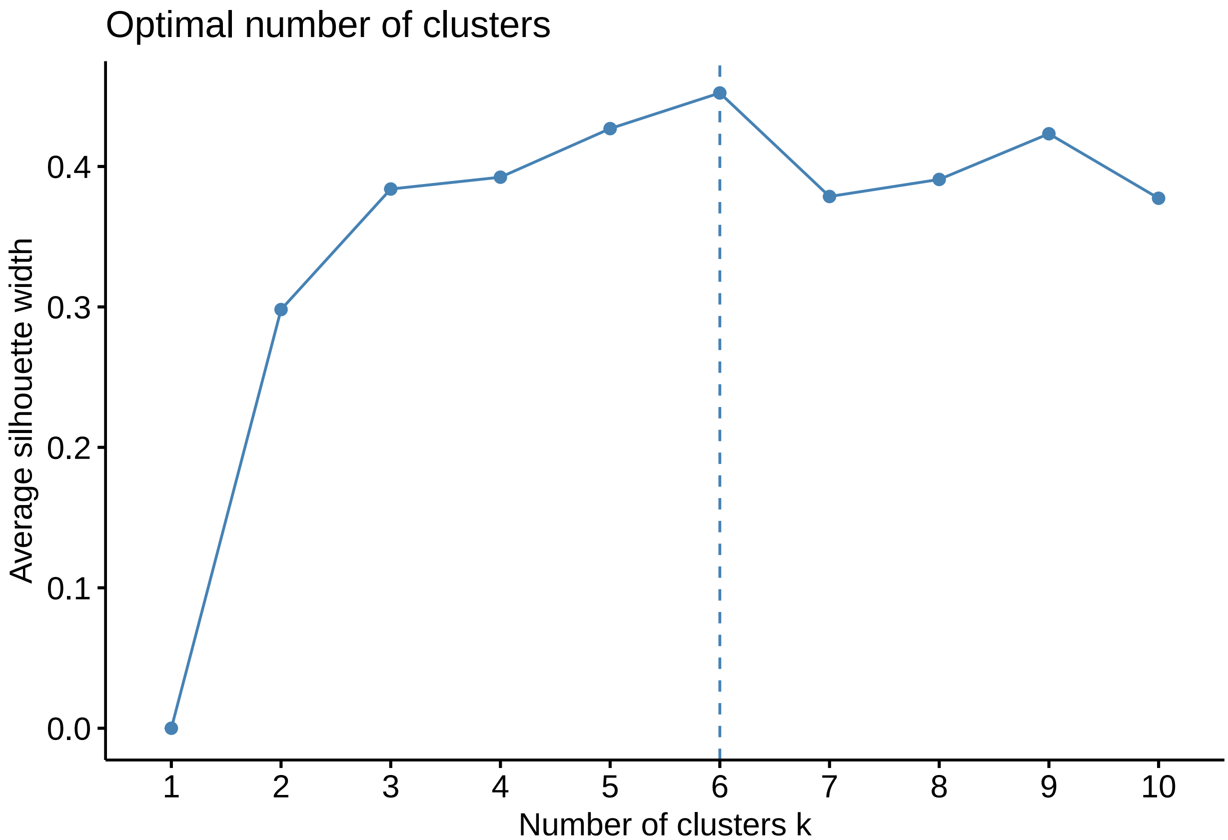
Results: The spending score of customers varies from 1 - 100. The average spending score is approx. 50. Mall customers with a spending score of 40 - 50 have the highest frequency.

5. K-means Clustering Algorithm Using the `silhouette` method to determine the optimal number of clusters for the k-mean algorithm. Then, performing principal component analysis to visualize the customer clusters.

```
library(NbClust)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_nbclust(mall_customers[,3:5], kmeans, method = "silhouette")
```



Results: By using the silhouette method, the optimal value of clusters is 6 for the k-means algorithm.

```
k6<-kmeans(mall_customers[,3:5],6,iter.max=100,nstart=50,algorithm="Lloyd")
k6
```

```
## K-means clustering with 6 clusters of sizes 45, 21, 35, 39, 38, 22
##
## Cluster means:
##        Age Annual Income (k$) Spending Score (1-100)
## 1 56.15556           53.37778               49.08889
## 2 44.14286           25.14286               19.52381
## 3 41.68571           88.22857               17.28571
## 4 32.69231           86.53846               82.12821
## 5 27.00000           56.65789               49.13158
## 6 25.27273           25.72727               79.36364
##
## Clustering vector:
##   [1] 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2 6 2
```

```
##  [38] 6 2 6 1 6 1 5 2 6 1 5 5 5 1 5 5 1 1 1 1 1 5 1 1 5 1 1 1 5 1 1 5 5 1 1 1 1
##  [75] 1 5 1 5 5 1 1 5 1 1 5 1 1 5 5 1 1 5 1 5 5 5 1 5 1 5 5 1 1 5 1 5 1 1 1 1 1
## [112] 5 5 5 5 5 1 1 1 1 5 5 5 4 5 4 3 4 3 4 3 4 5 4 3 4 3 4 3 4 3 4 5 4 3 4 3 4
## [149] 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4 3
## [186] 4 3 4 3 4 3 4 3 4 3 4 3 4 3 4
##
## Within cluster sum of squares by cluster:
## [1]  8062.133  7732.381 16690.857 13972.359  7742.895  4099.818
##  (between_SS / total_SS =  81.1 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
pcclust=prcomp(mall_customers[,3:5],scale=FALSE)
summary(pcclust)
```

```
## Importance of components:
##                            PC1     PC2     PC3
## Standard deviation     26.4625 26.1597 12.9317
## Proportion of Variance  0.4512  0.4410  0.1078
## Cumulative Proportion   0.4512  0.8922  1.0000
```

```
pcclust$rotation[,1:2]
```

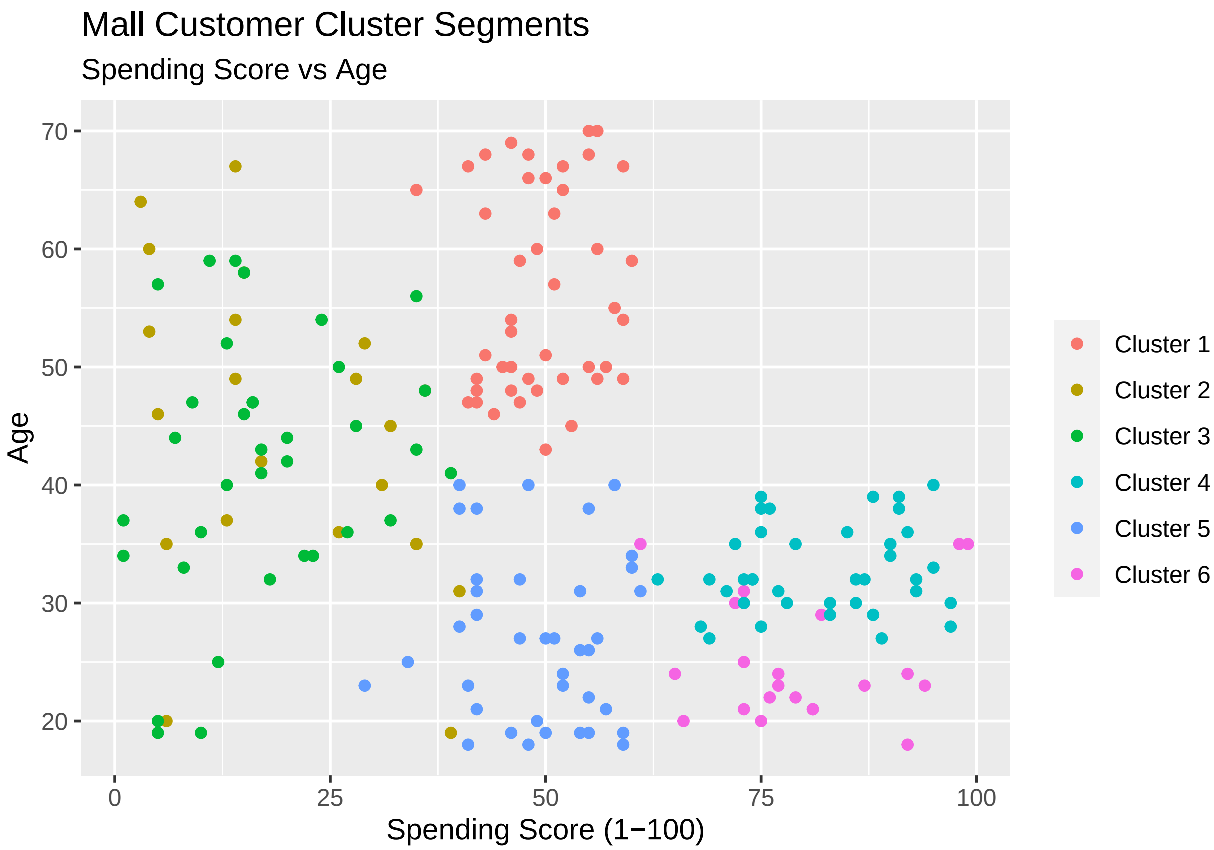
```
##                               PC1        PC2
## Age                     0.1889742 -0.1309652
## Annual Income (k$)     -0.5886410 -0.8083757
## Spending Score (1-100) -0.7859965  0.5739136
```

K-Means Clustering Plot: Annual Income vs Spending Score

```
set.seed(1)
ggplot(mall_customers, aes(x = `Annual Income (k$)`, `Spending Score (1-100)`)) +
  geom_point(stat = "identity", aes(color = as.factor(k6$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","Cluste
  ggtitle("Mall Customer Cluster Segments", subtitle = "Annual Income vs Spending Score")
```

K-Means Clustering Plot: Spending Score vs Age

```
ggplot(mall_customers, aes(x = `Spending Score (1-100)`, y =Age)) +
  geom_point(stat = "identity", aes(color = as.factor(k6$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","Cluste
  ggtitle("Mall Customer Cluster Segments", subtitle = "Spending Score vs Age")
```



Summary:

By using market basket analysis or customer segmenting of the mall customers, companies can identify patterns and develop efficient marketing strategies that target various customer groups, improve customer communication, and increase revenue. From the k-means clustering algorithm, mall customers are categorized into six clusters:

- Cluster 1: average spending score; average annual income; age range 40-70
- Cluster 2: low spending score; low annual income; age range 20-65
- Cluster 3: low spending score; high annual income; age range 20-60
- Cluster 4: high spending score; high annual income; age range 25 - 40
- Cluster 5: average spending score; average annual income; age range 18-40
- Cluster 6: high spending score; low annual income; age range 18-35