Mall Customers Project

W. Kerney

Project Overview

This project focuses on analyzing customer data to understand demographics, spending behavior, and key financial metrics, as well as building a machine learning model for customer segmentation.

Step 1: Setting Up the Environment

Installing and loading the necessary libraries.

```r
install.packages("dplyr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
install.packages("cluster")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
install.packages("caret")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
install.packages("randomForest")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
install.packages("knitr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(cluster)
library(randomForest)
```

```
## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:randomForest':
##
##     margin
```

```r
library(caret)
```

```
## Loading required package: lattice
```

```r
library(knitr)
```

Step 2: Load the Dataset

Loading the data into R and inspecting the structure.

```r
# Load the dataset
data <- read.csv("Mall_Customers.csv")

# Preview the dataset
head(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
```

```r
str(data)
```

```
## '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 ...
```

Step 3: Data Cleaning and Exploration

Cleaning the dataset by removing duplicates and missing values. Exploring the dataset to understand customer demographics and spending behavior.

```r
# Data cleaning
data_cleaned <- data %>%
  distinct() %>%
  na.omit()

# Basic exploration
summary(data_cleaned)
```

```
##    CustomerID        Gender               Age        Annual.Income..k..
##  Min.   :  1.00   Length:200         Min.   :18.00   Min.   : 15.00
##  1st Qu.: 50.75   Class :character   1st Qu.:28.75   1st Qu.: 41.50
##  Median :100.50   Mode  :character   Median :36.00   Median : 61.50
##  Mean   :100.50                      Mean   :38.85   Mean   : 60.56
##  3rd Qu.:150.25                      3rd Qu.:49.00   3rd Qu.: 78.00
##  Max.   :200.00                      Max.   :70.00   Max.   :137.00
##  Spending.Score..1.100.
##  Min.   : 1.00
##  1st Qu.:34.75
##  Median :50.00
##  Mean   :50.20
##  3rd Qu.:73.00
##  Max.   :99.00
```

Step 4: Calculate Financial Metrics

Calculating key financial metrics like average spending and distribution by gender and age. Trying to understand customer spending trends and identify high-value customer segments.

```r
# Calculate average spending
avg_spending <- mean(data_cleaned$Spending.Score..1.100.)
avg_spending  # Average spending score
```

```
## [1] 50.2
```

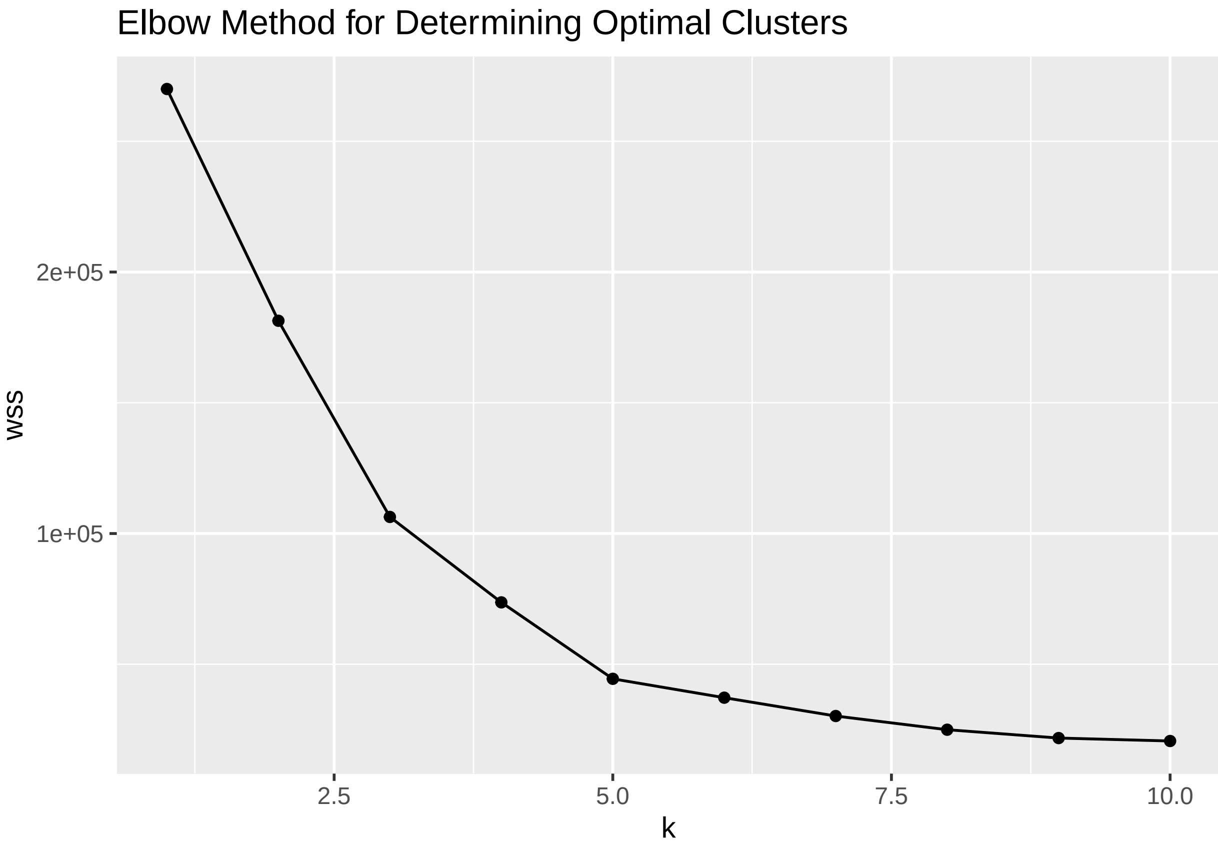
```r
# Spending by gender
gender_spending <- data_cleaned %>%
  group_by(Gender) %>%
  summarize(AverageSpending = mean(Spending.Score..1.100.))

# Spending by age group
age_spending <- data_cleaned %>%
  mutate(AgeGroup = cut(Age, breaks = c(0, 20, 30, 40, 50, 60, Inf), labels = c("0-20", "21-30", "31-40
  group_by(AgeGroup) %>%
  summarize(AverageSpending = mean(Spending.Score..1.100.))
```

Step 5: Building a Machine Learning Model

Creating a customer segmentation model using clustering techniques. This will allow us to identify distinct customer groups based on spending and other characteristics.

```r
# Determine the optimal number of clusters using the elbow method
set.seed(123)
wss <- sapply(1:10, function(k) {
  kmeans(data_cleaned[, c("Annual.Income..k..", "Spending.Score..1.100.")], centers = k, nstart = 10)$to
})
```

```r
# Plot the elbow curve to identify the optimal number of clusters
ggplot(data.frame(k = 1:10, wss = wss), aes(x = k, y = wss)) +
  geom_line() +
  geom_point() +
  labs(title = "Elbow Method for Determining Optimal Clusters")
```

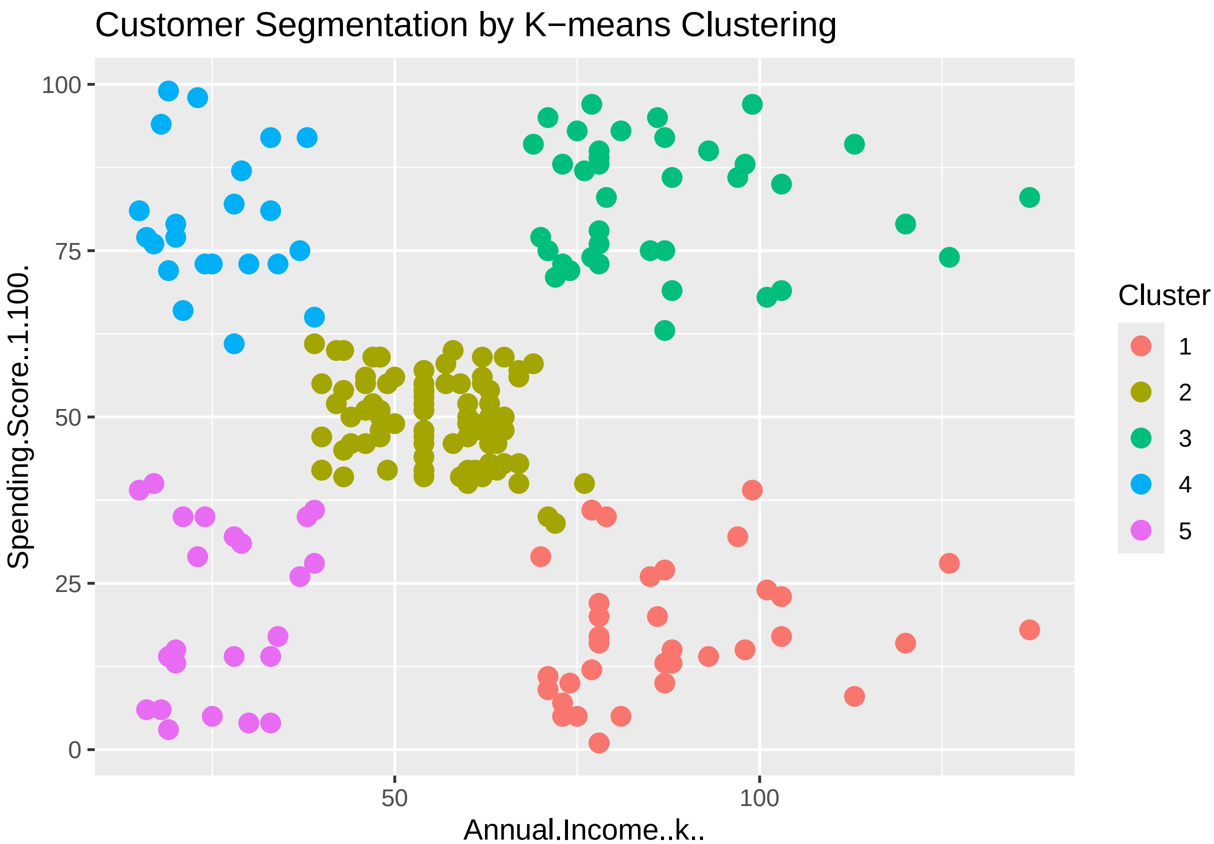



```r
# Assuming the optimal number of clusters is 5
kmeans_result <- kmeans(data_cleaned[, c("Annual.Income..k..", "Spending.Score..1.100.")], centers = 5,

# Assign cluster labels to the dataset
data_cleaned$Cluster <- kmeans_result$cluster
```

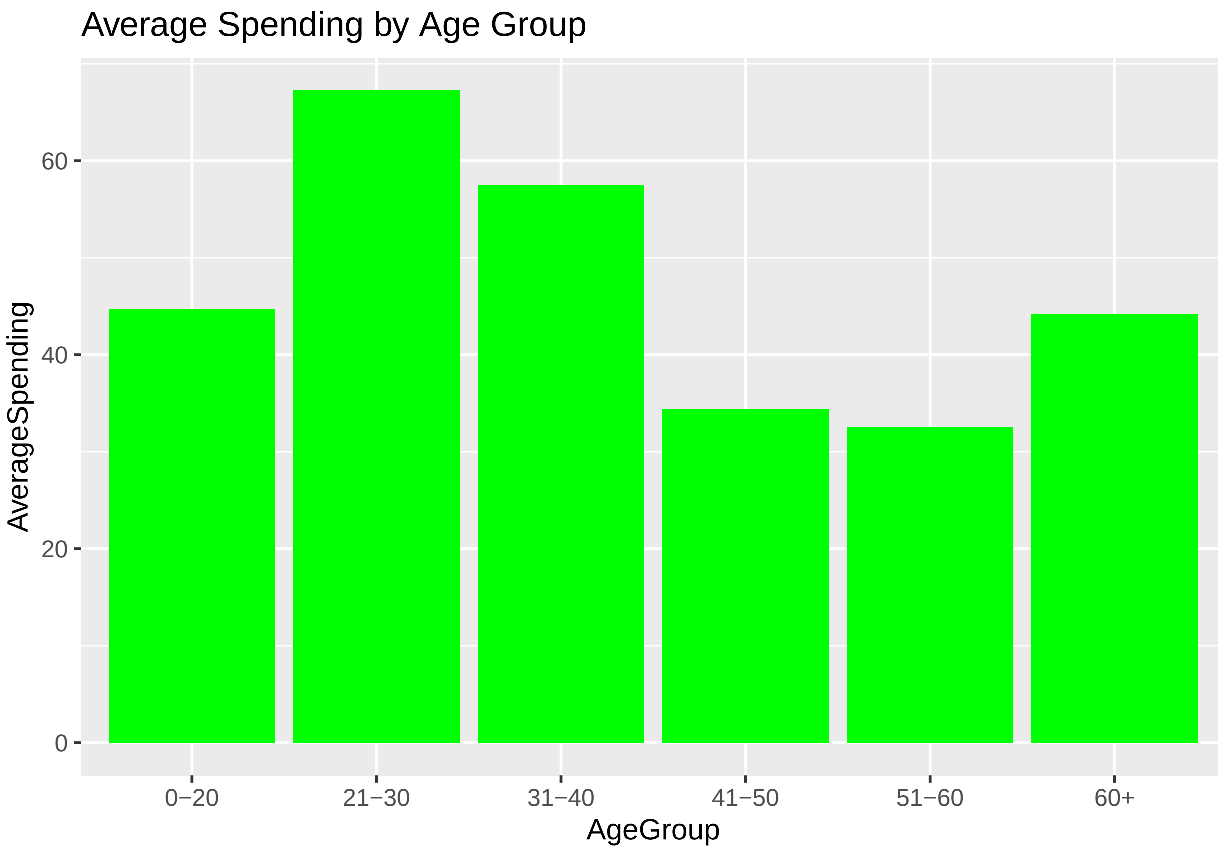
Step 6: Visualization and Presentation

I then visualize the customer clusters and key financial insights using ggplot2. Creating a scatter plot to show customer segmentation and bar plots for spending distribution to communicate our findings.

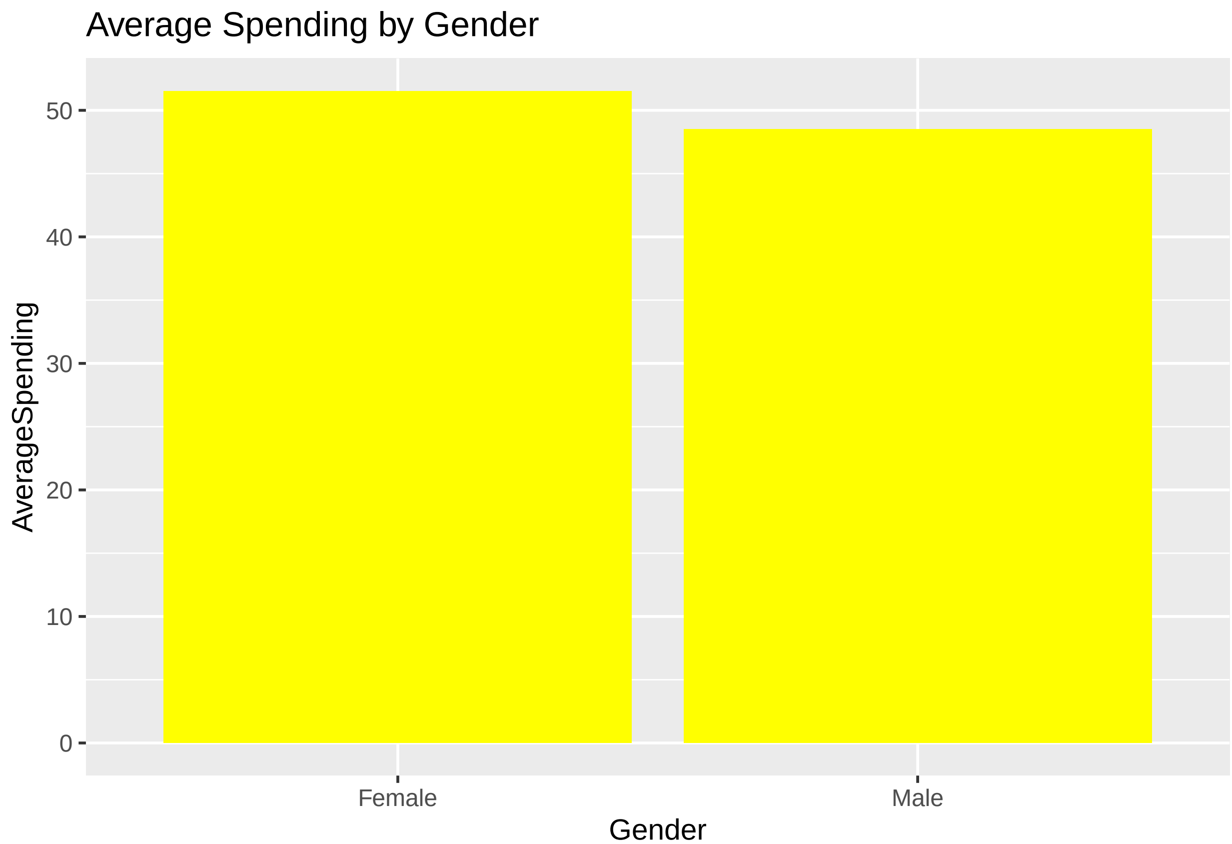
```r
# Scatter plot of clusters
ggplot(data_cleaned, aes(x = Annual.Income..k.., y = Spending.Score..1.100., color = as.factor(Cluster))
  geom_point(size = 3) +
  labs(title = "Customer Segmentation by K-means Clustering",
       color = "Cluster")
```




```
# Bar plot of average spending by age group
ggplot(age_spending, aes(x = AgeGroup, y = AverageSpending)) +
  geom_bar(stat = "identity", fill = "green") +
  labs(title = "Average Spending by Age Group")
```




```r
# Bar plot of average spending by gender
ggplot(gender_spending, aes(x = Gender, y = AverageSpending)) +
  geom_bar(stat = "identity", fill = "yellow") +
  labs(title = "Average Spending by Gender")
```


Average Spending by Gender
50
40
30
20
10
0
AverageSpending
Female
Male
Gender