Market Segmentation Analysis

Sirius Ife

2024-03-11

Introduction

In this project, we will conduct an unsupervised learning analysis on marketing data to uncover patterns and segments within the customer base. The goal is to gain insights that can inform targeted marketing strategies and enhance business decision-making.

Objective

- Perform exploratory data analysis (EDA) to understand the dataset.
- Pre-process the data by handling missing values, outliers, and encoding categorical variables.
- Conduct dimensionality reduction using PCA (Principal Component Analysis) to identify important features.
- Apply k-means clustering to segment the customer base.
- Interpret the clusters and derive actionable insights for marketing strategies.

Load Packages

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2

## Warning: package 'tidyr' was built under R version 4.3.2

## Warning: package 'readr' was built under R version 4.3.2

## Warning: package 'dplyr' was built under R version 4.3.2

## Warning: package 'stringr' was built under R version 4.3.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.2
```

```
## corrplot 0.92 loaded
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.2
```

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

```
## Warning: package 'lattice' was built under R version 4.3.2
```

```
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.3.3
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.3.2
```

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

```
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 4.3.2
```

Importing Data

```
customers <- read.delim("marketing_campaign.csv", stringsAsFactors = FALSE)
head(customers)
```

```
##     ID Year_Birth  Education Marital_Status Income Kidhome Teenhome Dt_Customer
## 1 5524       1957 Graduation         Single  58138       0        0  04-09-2012
## 2 2174       1954 Graduation         Single  46344       1        1  08-03-2014
## 3 4141       1965 Graduation       Together  71613       0        0  21-08-2013
## 4 6182       1984 Graduation       Together  26646       1        0  10-02-2014
## 5 5324       1981        PhD        Married  58293       1        0  19-01-2014
## 6 7446       1967     Master       Together  62513       0        1  09-09-2013
##   Recency MntWines MntFruits MntMeatProducts MntFishProducts MntSweetProducts
## 1      58      635        88             546             172               88
## 2      38       11         1               6               2                1
## 3      26      426        49             127             111               21
## 4      26       11         4              20              10                3
## 5      94      173        43             118              46               27
## 6      16      520        42              98               0               42
##   MntGoldProds NumDealsPurchases NumWebPurchases NumCatalogPurchases
## 1           88                 3               8                  10
## 2            6                 2               1                   1
## 3           42                 1               8                   2
## 4            5                 2               2                   0
## 5           15                 5               5                   3
## 6           14                 2               6                   4
##   NumStorePurchases NumWebVisitsMonth AcceptedCmp3 AcceptedCmp4 AcceptedCmp5
## 1                 4                 7            0            0            0
## 2                 2                 5            0            0            0
## 3                10                 4            0            0            0
## 4                 4                 6            0            0            0
## 5                 6                 5            0            0            0
## 6                10                 6            0            0            0
##   AcceptedCmp1 AcceptedCmp2 Complain Z_CostContact Z_Revenue Response
## 1            0            0        0             3        11        1
## 2            0            0        0             3        11        0
## 3            0            0        0             3        11        0
## 4            0            0        0             3        11        0
## 5            0            0        0             3        11        0
## 6            0            0        0             3        11        0
```

Exploratory Data Analysis (EDA)

```
# Check dimensions and summary statistics
dim(customers)
```

```
## [1] 2240   29
```

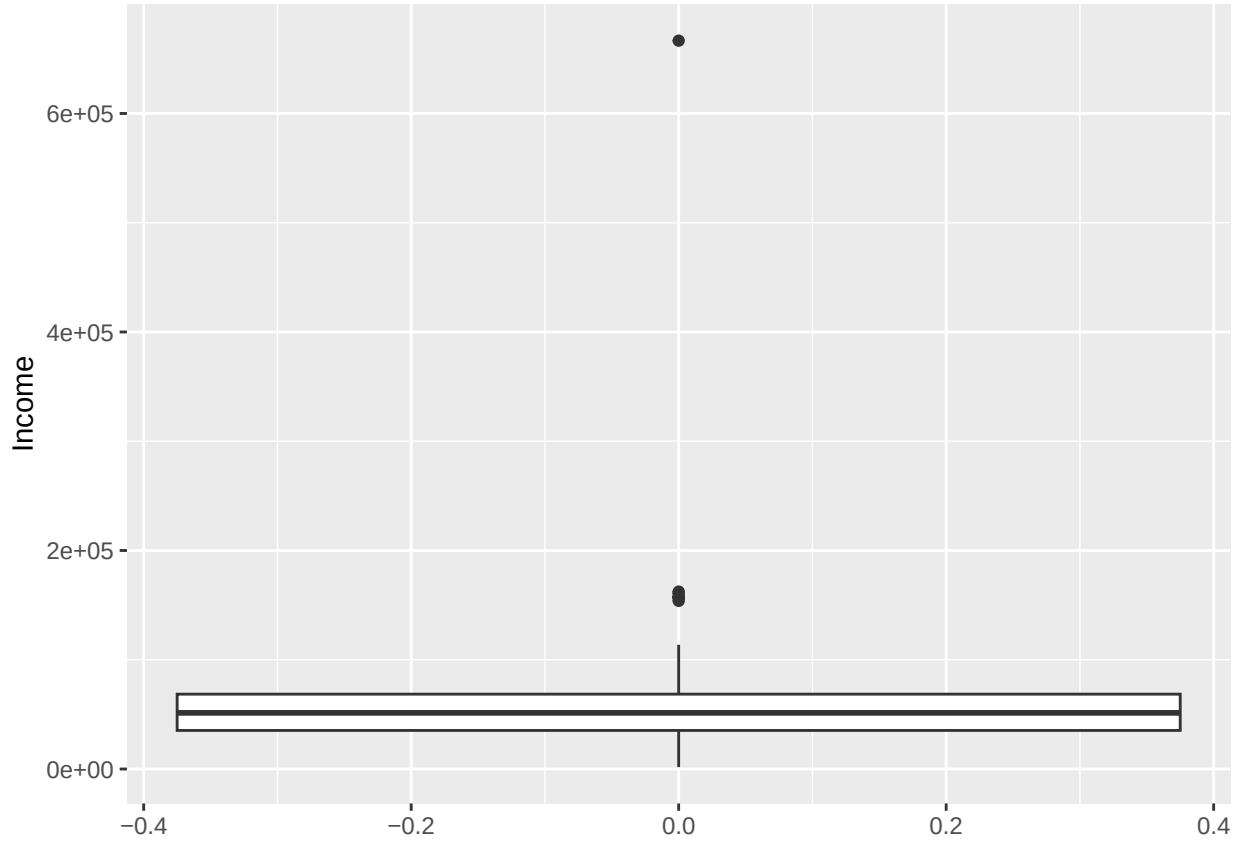
```
summary(customers)
```

```
##        ID          Year_Birth    Education         Marital_Status
##  Min.   :    0   Min.   :1893   Length:2240        Length:2240
##  1st Qu.: 2828   1st Qu.:1959   Class :character   Class :character
##  Median : 5458   Median :1970   Mode  :character   Mode  :character
##  Mean   : 5592   Mean   :1969
```

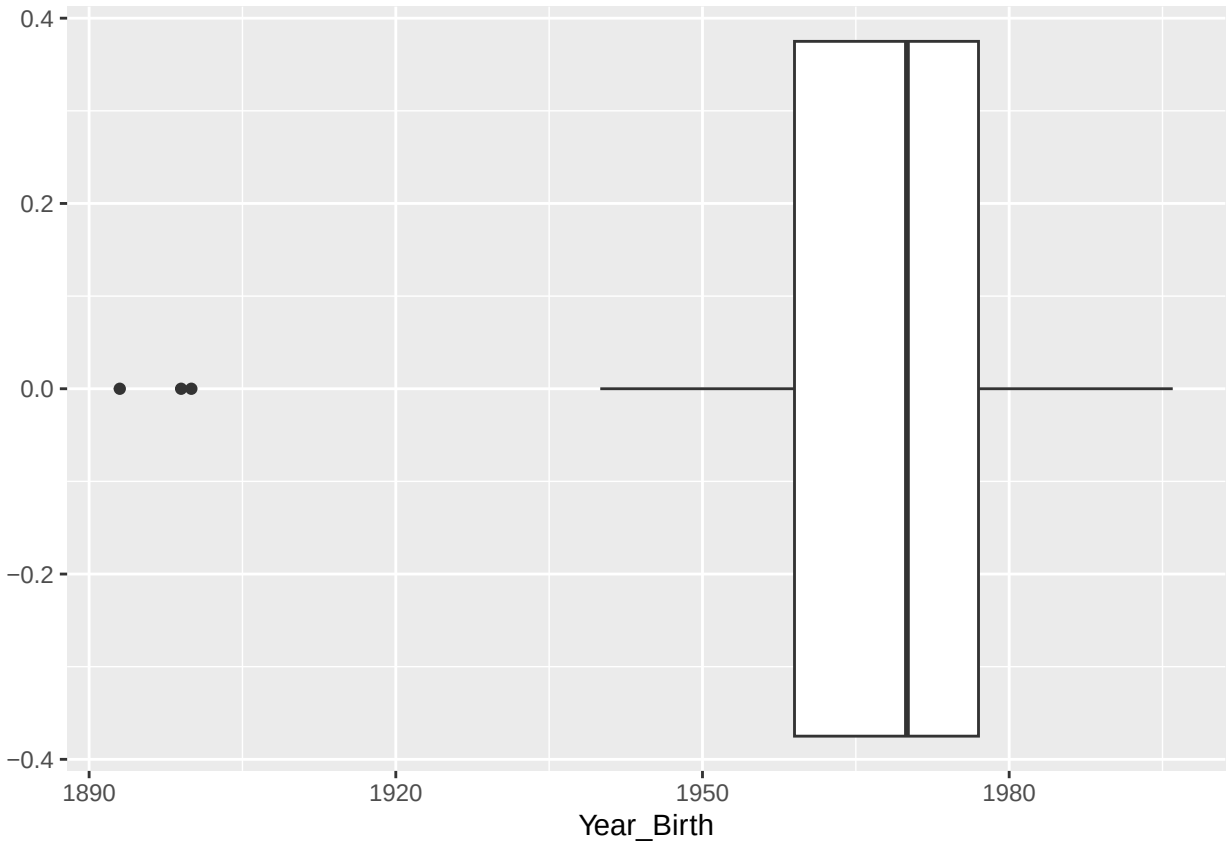
```
##  3rd Qu.: 8428   3rd Qu.:1977
##  Max.   :11191   Max.   :1996
##
##      Income          Kidhome          Teenhome      Dt_Customer
##  Min.   :  1730   Min.   :0.0000   Min.   :0.0000   Length:2240
##  1st Qu.: 35303   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median : 51382   Median :0.0000   Median :0.0000   Mode  :character
##  Mean   : 52247   Mean   :0.4442   Mean   :0.5062
##  3rd Qu.: 68522   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :666666   Max.   :2.0000   Max.   :2.0000
##  NA's   :24
##     Recency         MntWines         MntFruits     MntMeatProducts
##  Min.   : 0.00   Min.   :   0.00   Min.   :  0.0   Min.   :   0.0
##  1st Qu.:24.00   1st Qu.:  23.75   1st Qu.:  1.0   1st Qu.:  16.0
##  Median :49.00   Median : 173.50   Median :  8.0   Median :  67.0
##  Mean   :49.11   Mean   : 303.94   Mean   : 26.3   Mean   : 166.9
##  3rd Qu.:74.00   3rd Qu.: 504.25   3rd Qu.: 33.0   3rd Qu.: 232.0
##  Max.   :99.00   Max.   :1493.00   Max.   :199.0   Max.   :1725.0
##
##  MntFishProducts  MntSweetProducts  MntGoldProds    NumDealsPurchases
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   : 0.000
##  1st Qu.:  3.00   1st Qu.:  1.00   1st Qu.:  9.00   1st Qu.: 1.000
##  Median : 12.00   Median :  8.00   Median : 24.00   Median : 2.000
##  Mean   : 37.53   Mean   : 27.06   Mean   : 44.02   Mean   : 2.325
##  3rd Qu.: 50.00   3rd Qu.: 33.00   3rd Qu.: 56.00   3rd Qu.: 3.000
##  Max.   :259.00   Max.   :263.00   Max.   :362.00   Max.   :15.000
##
##  NumWebPurchases  NumCatalogPurchases NumStorePurchases NumWebVisitsMonth
##  Min.   : 0.000   Min.   : 0.000      Min.   : 0.00     Min.   : 0.000
##  1st Qu.: 2.000   1st Qu.: 0.000      1st Qu.: 3.00     1st Qu.: 3.000
##  Median : 4.000   Median : 2.000      Median : 5.00     Median : 6.000
##  Mean   : 4.085   Mean   : 2.662      Mean   : 5.79     Mean   : 5.317
##  3rd Qu.: 6.000   3rd Qu.: 4.000      3rd Qu.: 8.00     3rd Qu.: 7.000
##  Max.   :27.000   Max.   :28.000      Max.   :13.00     Max.   :20.000
##
##   AcceptedCmp3      AcceptedCmp4      AcceptedCmp5      AcceptedCmp1
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.07277   Mean   :0.07455   Mean   :0.07277   Mean   :0.06429
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##
##   AcceptedCmp2        Complain        Z_CostContact   Z_Revenue
##  Min.   :0.00000   Min.   :0.000000   Min.   :3     Min.   :11
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:3     1st Qu.:11
##  Median :0.00000   Median :0.000000   Median :3     Median :11
##  Mean   :0.01339   Mean   :0.009375   Mean   :3     Mean   :11
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:3     3rd Qu.:11
##  Max.   :1.00000   Max.   :1.000000   Max.   :3     Max.   :11
##
##     Response
##  Min.   :0.0000
##  1st Qu.:0.0000
```

```
##  Median :0.0000
##  Mean   :0.1491
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
```

```r
# Remove rows with missing income values
customers <- customers %>% filter(!is.na(Income))

# Visualize distribution of income and year of birth
ggplot(customers, aes(y = Income)) + geom_boxplot()
```




```r
ggplot(customers, aes(Year_Birth)) + geom_boxplot()
```



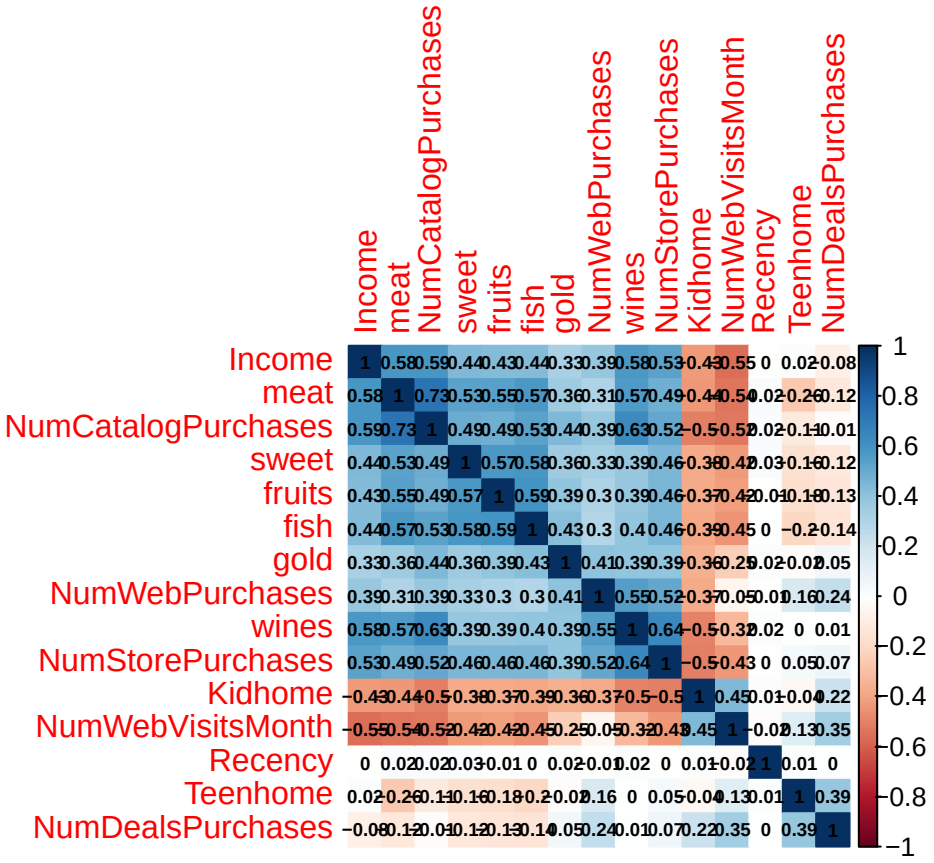
Pre-Processing Data

```
# Convert Dt_Customer to date format
customers <- customers %>% mutate(Dt_Customer = as.Date(dmy(Dt_Customer)))

# Create age variable from year of birth
customers <- customers %>% mutate(Age = 2024 - Year_Birth)
customers <- customers %>% filter(Age < 90)  # Remove outliers

# Collapse marital status into two categories: Single & Taken
customers <- customers %>% mutate(Marital_Status = ifelse(Marital_Status %in% c("Divorced", "Widow", "A

# Collapse education into two categories: Graduate & Non-graduate
customers <- customers %>% mutate(Education = ifelse(Education %in% c("Graduation", "PhD", "Master"), "

# Convert categorical variables to factors
customers <- customers %>% mutate(Marital_Status = as.factor(Marital_Status), Education = as.factor(Edu

# Rename features and create Total_spent variable
customers <- customers %>% rename(wines = MntWines, fruits = MntFruits, meat = MntMeatProducts, fish = M
customers <- customers %>% mutate(Total_spent = wines + fruits + meat + fish + sweet + gold)

# Remove redundant features
```

```r
customers <- customers %>% select(- ID, - Year_Birth, - Dt_Customer, - Z_CostContact, - Z_Revenue)
```

Correlation Analysis

```r
# Calculate correlation matrix
cust_cor <- cor(customers[,3:17])
corrplot(cust_cor, method = "color", order = "hclust", addCoef.col = "black", number.cex = .6)
```



Dimensionality Reduction: PCA

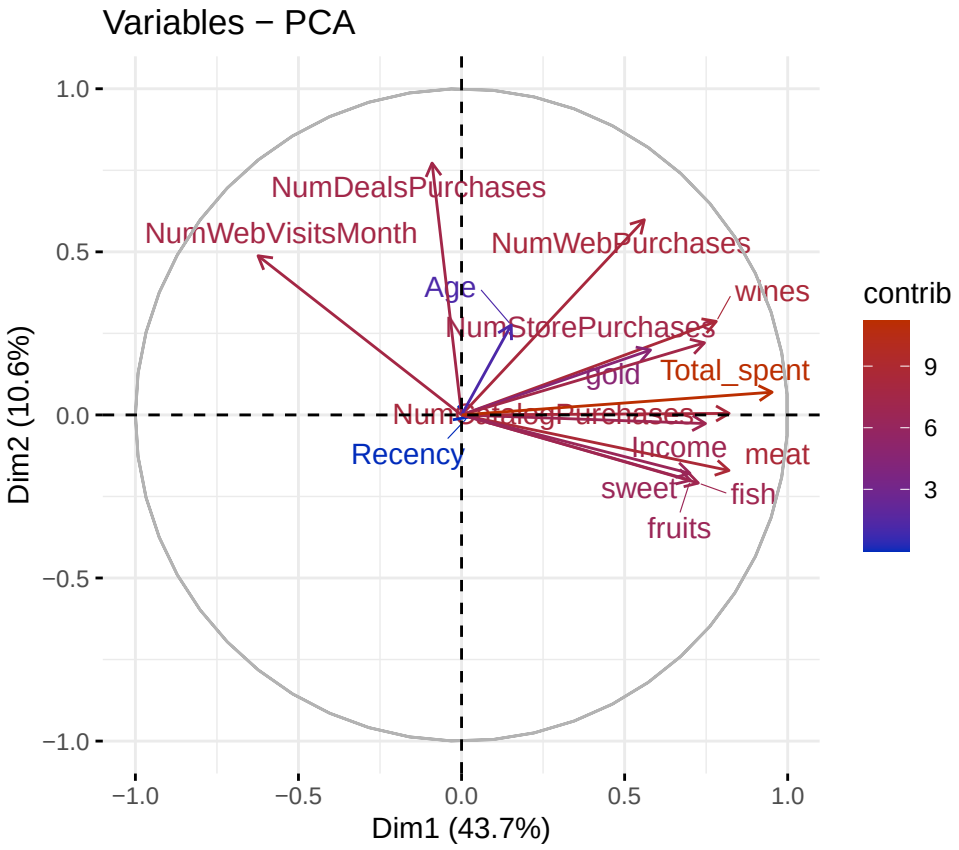
```r
# Running PCA
customers_pca <- PCA(customers[, c(3, 6:17, 25:26)], graph = FALSE)

# Summary of PCA
summary(customers_pca)
```

```
##
## Call:
## PCA(X = customers[, c(3, 6:17, 25:26)], graph = FALSE)
```

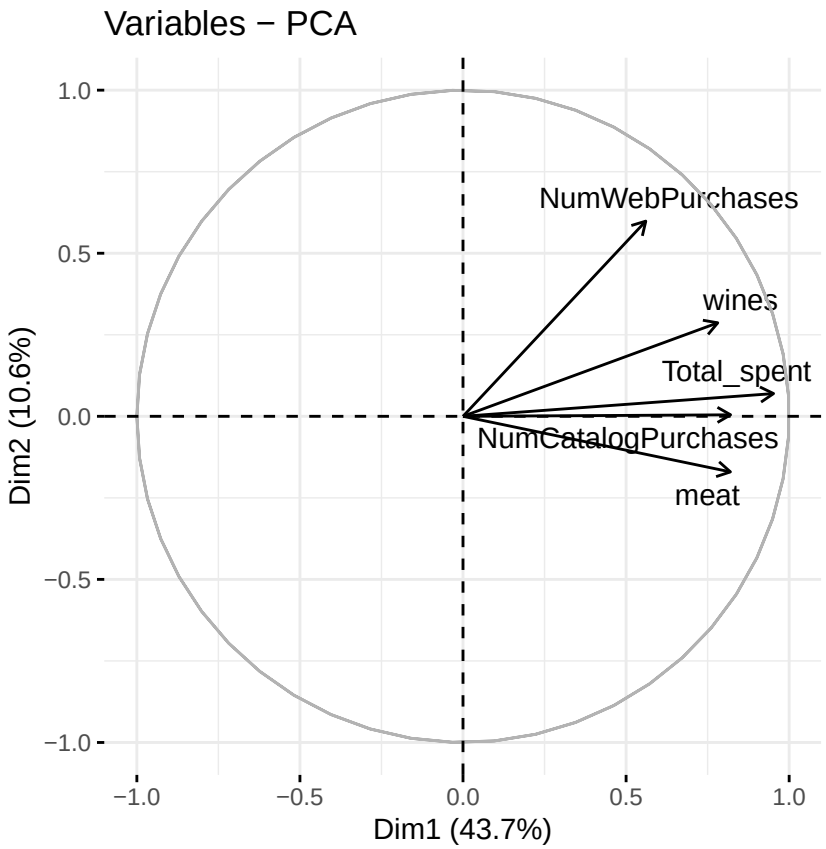
```
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance               6.557   1.590   1.074   1.001   0.826   0.666   0.636
## % of var.             43.715  10.598   7.157   6.676   5.504   4.438   4.242
## Cumulative % of var.  43.715  54.313  61.470  68.146  73.650  78.088  82.330
##                        Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14
## Variance               0.564   0.449   0.418   0.390   0.339   0.254   0.236
## % of var.              3.763   2.993   2.788   2.597   2.263   1.694   1.571
## Cumulative % of var.  86.093  89.086  91.874  94.471  96.734  98.429 100.000
##                       Dim.15
## Variance               0.000
## % of var.              0.000
## Cumulative % of var. 100.000
##
## Individuals (the 10 first)
##                       Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2
## 1                   |  5.329 |  4.087  0.115  0.588 |  0.645  0.012  0.015 |
## 2                   |  2.968 | -2.280  0.036  0.590 | -0.582  0.010  0.038 |
## 3                   |  2.872 |  1.655  0.019  0.332 |  0.155  0.001  0.003 |
## 4                   |  2.945 | -2.508  0.043  0.725 | -0.737  0.015  0.063 |
## 5                   |  2.535 | -0.142  0.000  0.003 |  0.473  0.006  0.035 |
## 6                   |  2.364 |  0.777  0.004  0.108 |  0.702  0.014  0.088 |
## 7                   |  1.967 |  0.626  0.003  0.101 |  0.805  0.018  0.168 |
## 8                   |  2.748 | -2.221  0.034  0.653 | -0.005  0.000  0.000 |
## 9                   |  3.319 | -2.871  0.057  0.748 | -0.299  0.003  0.008 |
## 10                  |  7.252 | -4.454  0.137  0.377 |  1.529  0.066  0.044 |
##                      Dim.3    ctr   cos2
## 1                    0.714  0.021  0.018 |
## 2                   -1.424  0.085  0.230 |
## 3                   -0.162  0.001  0.003 |
## 4                    0.940  0.037  0.102 |
## 5                    0.734  0.023  0.084 |
## 6                   -0.217  0.002  0.008 |
## 7                    0.836  0.029  0.181 |
## 8                    1.279  0.069  0.217 |
## 9                    0.572  0.014  0.030 |
## 10                   0.329  0.005  0.002 |
##
## Variables (the 10 first)
##                         Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
## Income                |  0.748  8.532  0.559 | -0.026  0.044  0.001 | -0.249
## Recency               |  0.019  0.006  0.000 | -0.007  0.003  0.000 | -0.160
## wines                 |  0.781  9.305  0.610 |  0.287  5.170  0.082 | -0.135
## fruits                |  0.702  7.505  0.492 | -0.200  2.523  0.040 |  0.260
## meat                  |  0.820 10.254  0.672 | -0.170  1.827  0.029 |  0.006
## fish                  |  0.726  8.038  0.527 | -0.209  2.745  0.044 |  0.218
## sweet                 |  0.700  7.479  0.490 | -0.179  2.018  0.032 |  0.225
## gold                  |  0.579  5.121  0.336 |  0.199  2.492  0.040 |  0.263
## NumDealsPurchases     | -0.090  0.125  0.008 |  0.772 37.509  0.596 |  0.176
## NumWebPurchases       |  0.560  4.790  0.314 |  0.598 22.529  0.358 |  0.097
##                         ctr   cos2
## Income                5.795  0.062 |
```

```
## Recency              2.372  0.025 |
## wines                1.702  0.018 |
## fruits               6.305  0.068 |
## meat                 0.003  0.000 |
## fish                 4.430  0.048 |
## sweet                4.722  0.051 |
## gold                 6.460  0.069 |
## NumDealsPurchases    2.895  0.031 |
## NumWebPurchases      0.876  0.009 |
```

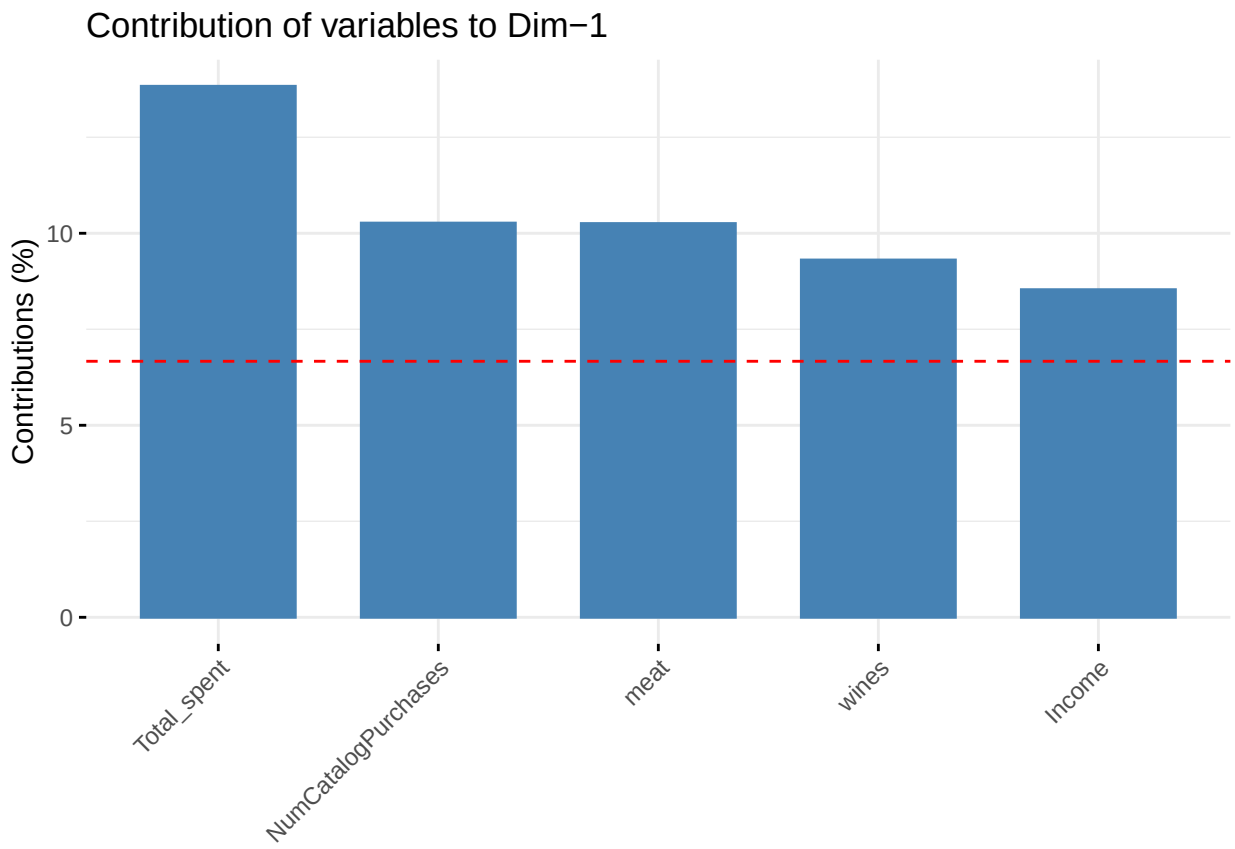
```r
# Plotting contributions of variables
fviz_pca_var(customers_pca, col.var = "contrib", gradient.cols = c("#002bbb", "#bb2e00"), repel = TRUE)
```




```r
# Plotting top 5 variables with highest contributions
fviz_pca_var(customers_pca, select.var = list(contrib = 5), repel = TRUE)
```

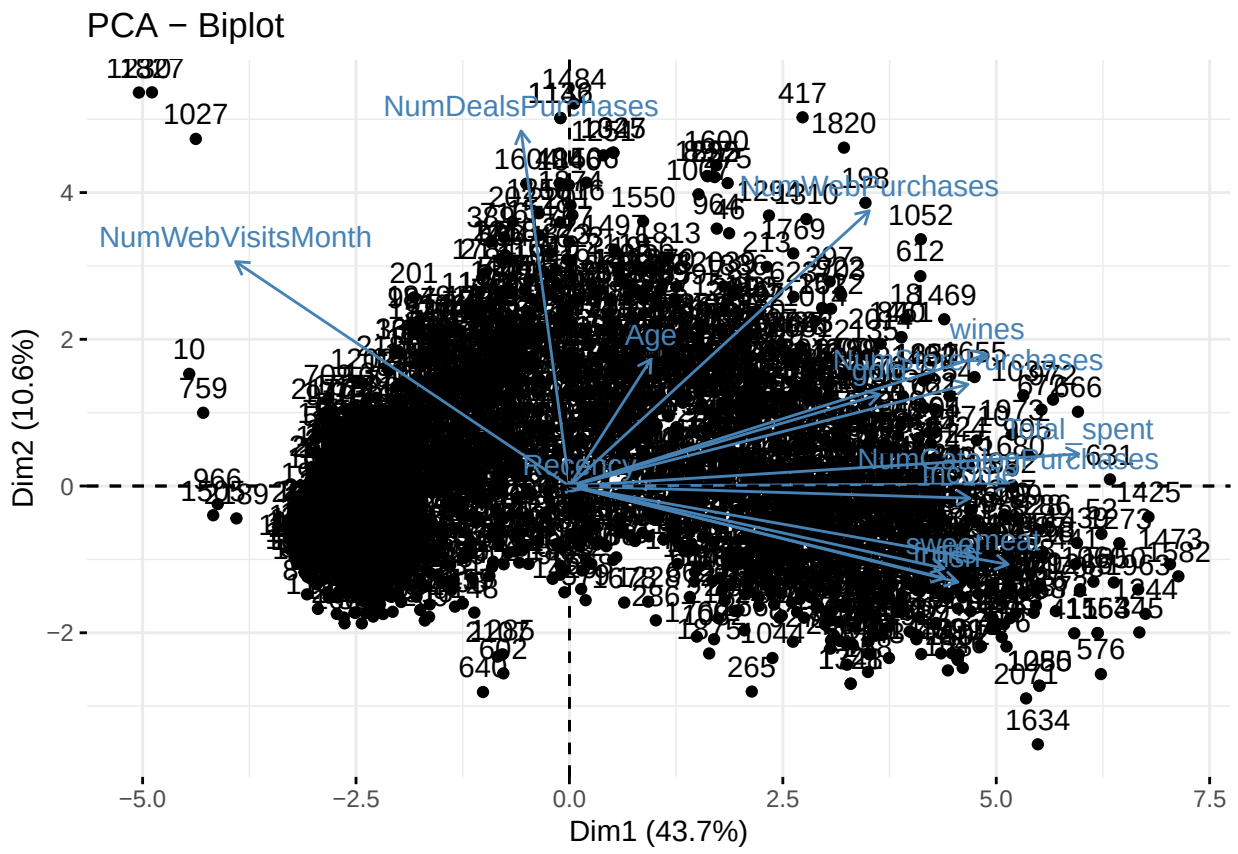



```
# Barplotting the contributions of variables
fviz_contrib(customers_pca, choice = "var", axes = 1, top = 5)
```



Biplots

```
fviz_pca_biplot(customers_pca)
```

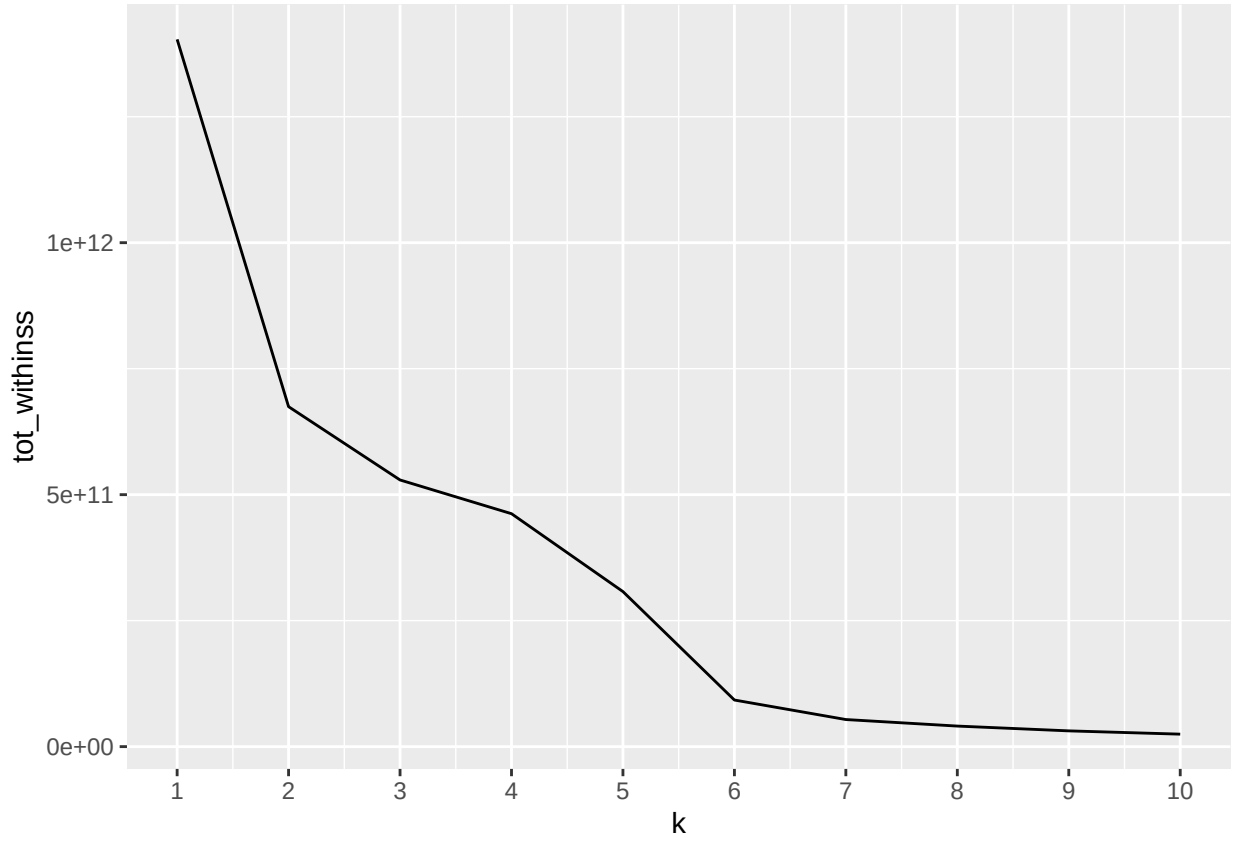


K-Means Clustering

```
# Elbow method to determine optimal k
tot_withinss <- map_dbl(1:10, function(k){
  model <- kmeans(x = customers[, c(3, 6:17, 25:26)], centers = k)
  model$tot.withinss
})

elbow_df <- data.frame(k = 1:10, tot_withinss = tot_withinss)
head(elbow_df)
```

```
##   k tot_withinss
## 1 1 1.403512e+12
## 2 2 6.746498e+11
## 3 3 5.291487e+11
## 4 4 4.622088e+11
## 5 5 3.076275e+11
## 6 6 9.247805e+10
```

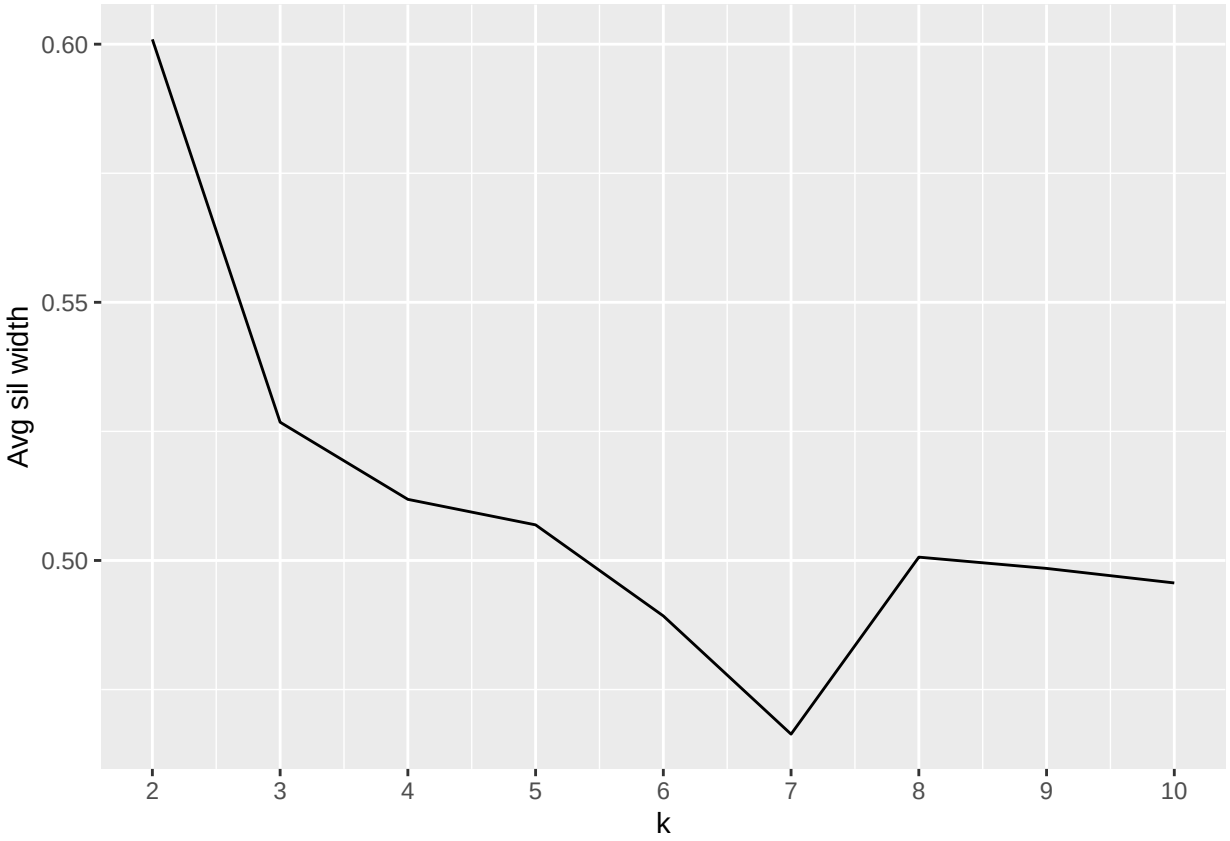
```
# Plotting the elbow plot
ggplot(elbow_df, aes(k, tot_withinss)) + geom_line() + scale_x_continuous(breaks = 1:10)
```




```
# Silhouette analysis
sil_width <- map_dbl(2:10, function(k){
    model <- pam(customers[, c(3, 6:17, 25:26)], k = k)
    model$silinfo$avg.width
})

sil_df <- data.frame(k = 2:10, sil_width = sil_width)
head(sil_df)
```

```
##   k sil_width
## 1 2 0.6009380
## 2 3 0.5267777
## 3 4 0.5118380
## 4 5 0.5068921
## 5 6 0.4892537
## 6 7 0.4663557
```

```
ggplot(sil_df, aes(k, sil_width)) + geom_line() + scale_x_continuous(breaks = 2:10) + labs(y = "Avg sil
```




```r
# K-means clustering with k=2
set.seed(77)
customers_cluster <- kmeans(customers[, c(3, 6:17, 25:26)], centers = 2)
```