Final Project Code

Ayesha Mulla, Harshwardhan Patil, Radhika Agarwal

2024-02-22

Read the data

```
data <- read.table("Company-data.csv", sep ='\t',header = T)
head(data)
```

```
##     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
```

Produce summary of the data

```
summary(data)
```

```
##        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
##
```

We can observe that "Income" variable has 24 NA values.

```
data <- na.omit(data)
```

Feature Engineering

```
# Tenure
data$Dt_Customer <- as.Date(data$Dt_Customer, format= "%d-%m-%Y")
days <- as.numeric(max(data$Dt_Customer) - data$Dt_Customer)
data$Tenure <- days
data$Tenure <- as.numeric(data$Tenure, errors="coerce")

# Age
data$Age <- 2014 - data$Year_Birth

# Spending
data$Spendings <- data$MntWines + data$MntFruits + data$MntMeatProducts + data$MntFishProducts +data$Mn

# Wines
data$Wines <- data$MntWines

# Fruits
data$Fruits <- data$MntFruits

# Meat
data$Meat <- data$MntMeatProducts

# Fish
data$Fish <- data$MntFishProducts

# Sweets
data$Sweets <- data$MntSweetProducts

# Gold
data$Gold <- data$MntGoldProds
```

```r
# Relationship Status
data$RelationshipStatus <- ifelse(data$Marital_Status == "Married" | data$Marital_Status == "Together",
ifelse(data$Marital_Status %in% c("Absurd", "YOLO", "Single","Alone"), "Single",
ifelse(data$Marital_Status %in% c("Widow"), "Widow",
ifelse(data$Marital_Status %in% c("Divorced"), "Divorced", "")))

data$RelStatus <- as.numeric(ifelse(data$RelationshipStatus == "Single", 1,
                         ifelse(data$RelationshipStatus == "Couple", 2,
                                ifelse(data$RelationshipStatus == "Widow", 3,
                                       ifelse(data$RelationshipStatus == "Divorced", 4, 0)))))


# Children
data$Children <- data$Kidhome + data$Teenhome

# Parent
data$Parent <- ifelse(data$Children > 0, 1, 0)

# Education
data$Education <- ifelse(data$Education %in% c("Basic", "2n Cycle"), "Undergraduate",
                         ifelse(data$Education == "Graduation", "Graduate",
                                ifelse(data$Education %in% c("Master", "PhD"), "Postgraduate", "")))

data$LevEd <- as.numeric(ifelse(data$Education == "Undergraduate",1,
                            ifelse(data$Education == "Graduate", 2,
                               ifelse(data$Education == "Postgraduate", 3, 0))))

# Campaign
data$Campaign <- data$AcceptedCmp1 + data$AcceptedCmp2 + data$AcceptedCmp3 + data$AcceptedCmp4 + data$A

# Purchases
data$Purchases <- data$NumDealsPurchases + data$NumWebPurchases + data$NumCatalogPurchases + data$NumSt

# Change names of different variables for simplicity
data$WebVisits <- data$NumWebVisitsMonth
data$Web <- data$NumWebPurchases
data$Deal<- data$NumDealsPurchases
data$Catalog <- data$NumCatalogPurchases
data$Store <- data$NumStorePurchases

# Widow or Not
data$widow = ifelse(data$RelStatus==3,1,0)
```

```r
# Remove data of customers having age greater than 80
data$Age <- ifelse(data$Age > 80, NA, data$Age)
data <- na.omit(data)
```

```r
# Remove data of customers having income greater than 170000
data$Income <- ifelse(data$Income > 170000, NA, data$Income)
data <- na.omit(data)
```

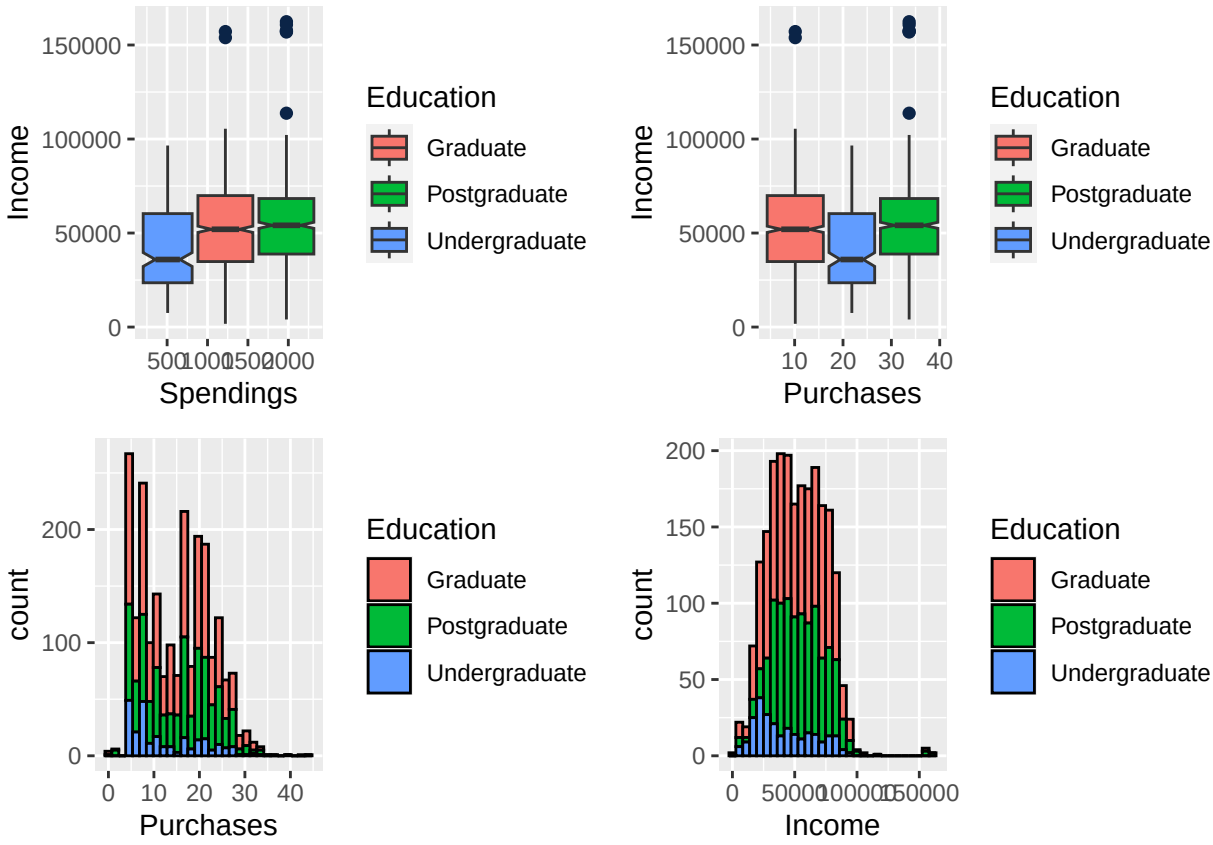
Drop unnecessary features from the data

```r
to_drop <- c("Marital_Status", "NumDealsPurchases", "NumWebPurchases", "NumCatalogPurchases", "NumStorel
data <- data[, !(names(data) %in% to_drop)]
```

```r
# Drop categorial variables to create a numerical data frame
drop <- c("Education", "RelationshipStatus")
data_numerical <- data[,!(names(data) %in% drop)]
```

Exploratory Data Analysis

```r
options(repr.plot.width=30, repr.plot.height=8)
require(gridExtra)

income_spendings_education_plot <- ggplot(data, aes(x=Spendings,y=Income,fill=Education)) +
geom_boxplot(outlier.colour="#0B2447", outlier.shape=16,outlier.size=2, notch=T)

income_purchases_education_plot <- ggplot(data, aes(x=Purchases,y=Income,fill=Education)) +
geom_boxplot(outlier.colour="#0B2447", outlier.shape=16,outlier.size=2, notch=T)

purchases_hist <- ggplot(data, aes(x=Purchases, fill=Education)) +
  geom_histogram(color="black", bins = 30)

income_hist <- ggplot(data, aes(x=Income, fill=Education)) +
  geom_histogram(color="black", bins = 30)

grid.arrange(income_spendings_education_plot, income_purchases_education_plot, purchases_hist, income_hi
```




```
df_tidy <- data %>%
  gather(key = "product", value = "amount", Wines:Gold)
```

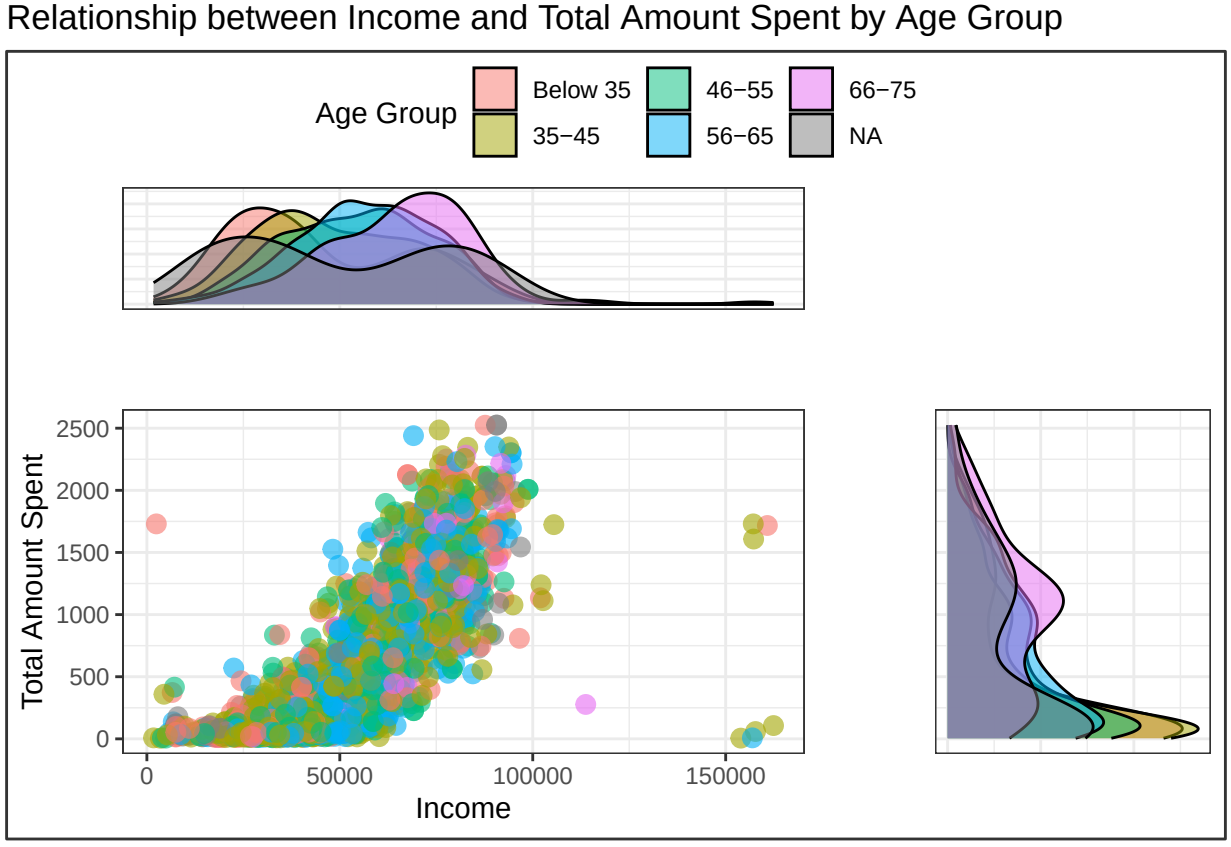
```
df_tidy <- df_tidy %>%
  mutate(age_group = cut(as.numeric(Age), breaks = c(25, 35, 45, 55, 65, 80),
                         labels = c("Below 35", "35-45", "46-55", "56-65", "66-75")))
data <- data %>%
  mutate(age_group = cut(as.numeric(Age), breaks = c(25, 35, 45, 55, 65, 80),
                         labels = c("Below 35", "35-45", "46-55", "56-65", "66-75")))
T1 <- data %>%
  mutate(total_amount_spent = Wines+Fruits+Meat+Fish+Sweets+Gold) %>%
  filter (Income < 3e+05)
```

```
colnames(T1)[colnames(T1) == "age_group"] ="Age Group"
colnames(T1)[colnames(T1) == "total_amount_spent"] ="Total Amount Spent"
```

```
data<-na.omit(data)
```

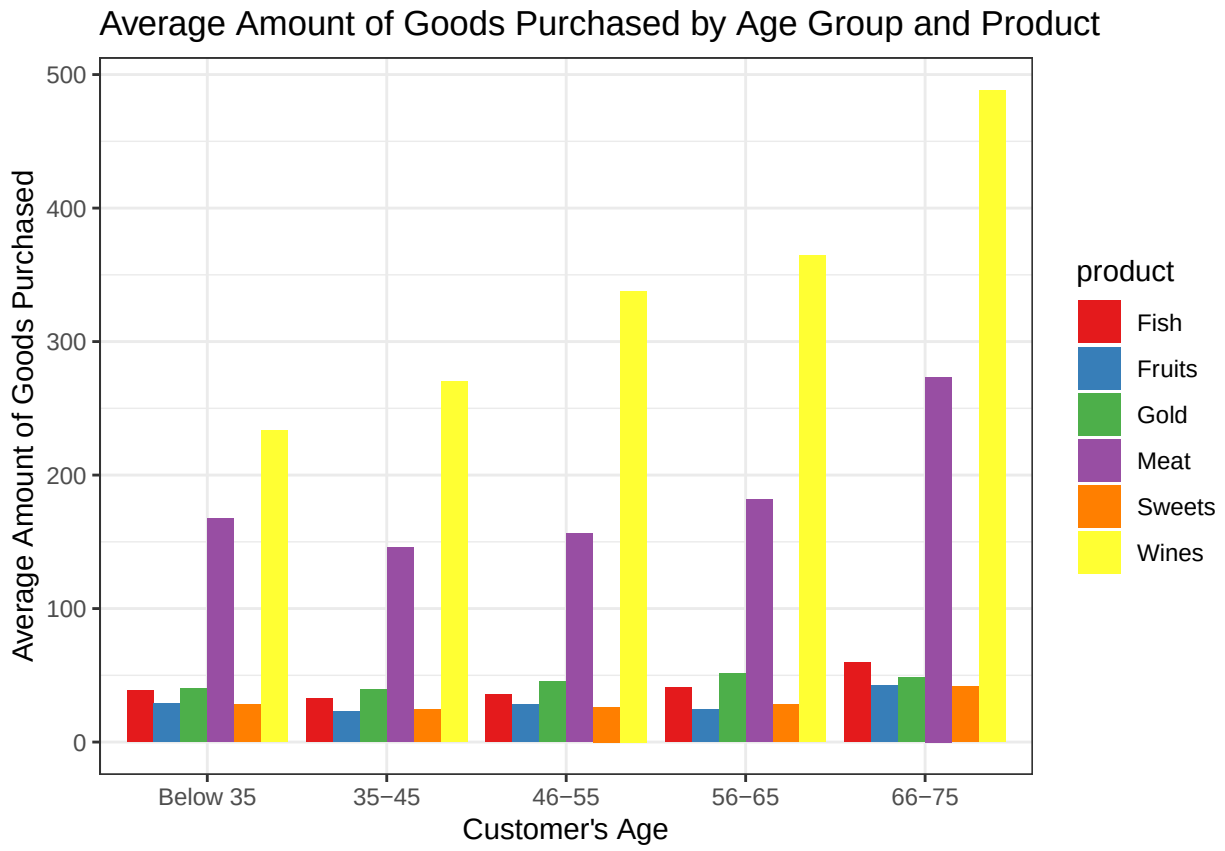
```
# Scatter plot colored by groups ("Species")
sp <- ggscatter(T1, x = "Income", y = "Total Amount Spent",
                color = "Age Group",
                size = 3, alpha = 0.6)+
  border()+ theme_bw()
```

```
# Marginal density plot of x (top panel) and y (right panel)
xplot <- ggdensity(T1, "Income", fill = "Age Group")+ theme_bw()

yplot <- ggdensity(T1, "Total Amount Spent", fill = "Age Group")+
  rotate()+ theme_bw()

# Cleaning the plots
yplot <- yplot + clean_theme()
xplot <- xplot + clean_theme()

# Give a suitable title to the plot
title <- ggtitle("Relationship between Income and Total Amount Spent by Age Group")

# Arranging the plot
ggarrange(xplot, NULL, sp, yplot,
          ncol = 2, nrow = 2,  align = "hv",
          widths = c(2, 1), heights = c(1, 2),
          common.legend = TRUE)+ theme_bw() +theme_bw()+title
```

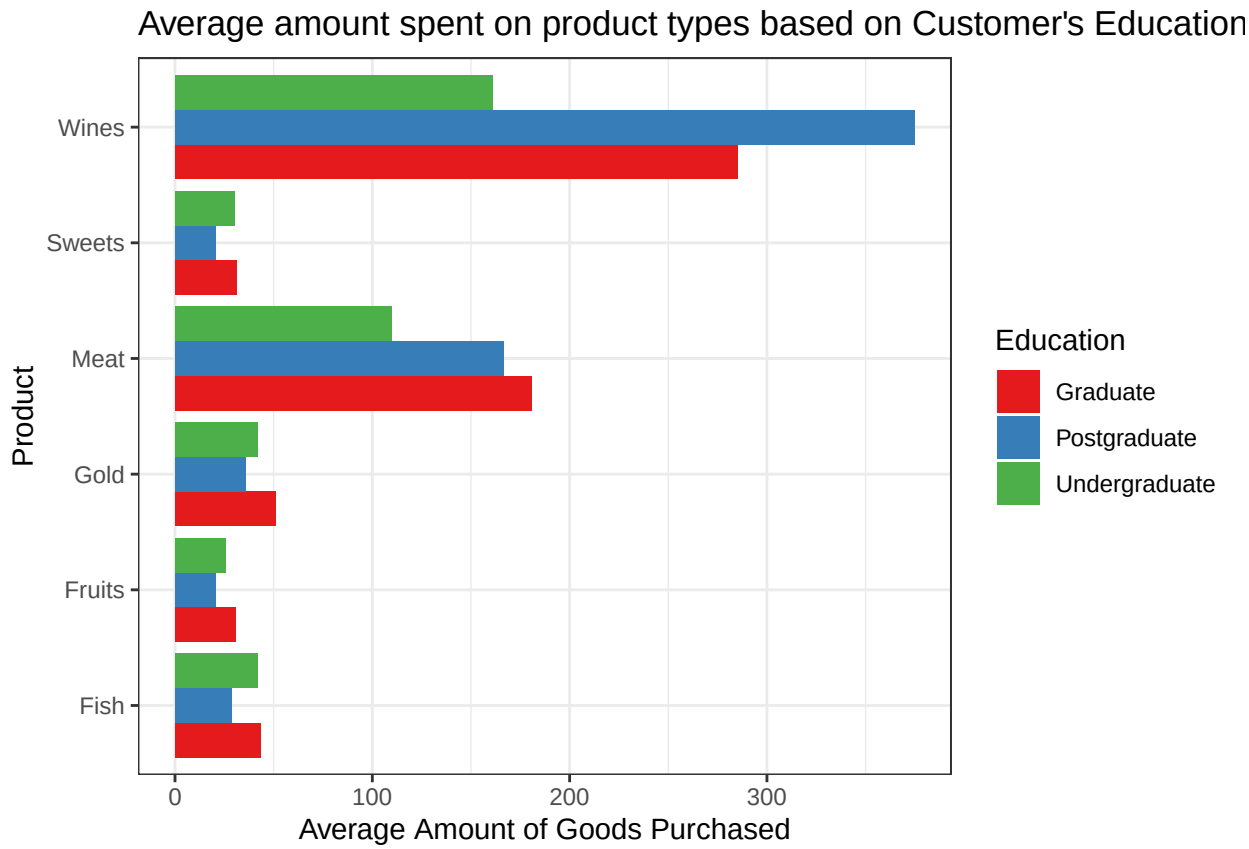



```
total_age = aggregate(amount ~ product+age_group, data = df_tidy, mean)

total_age%>%
ggplot( aes(x = age_group, y = amount, fill = product)) +
  geom_bar(stat = "identity", position = "dodge")  +
    scale_fill_brewer(palette = "Set1") +
```

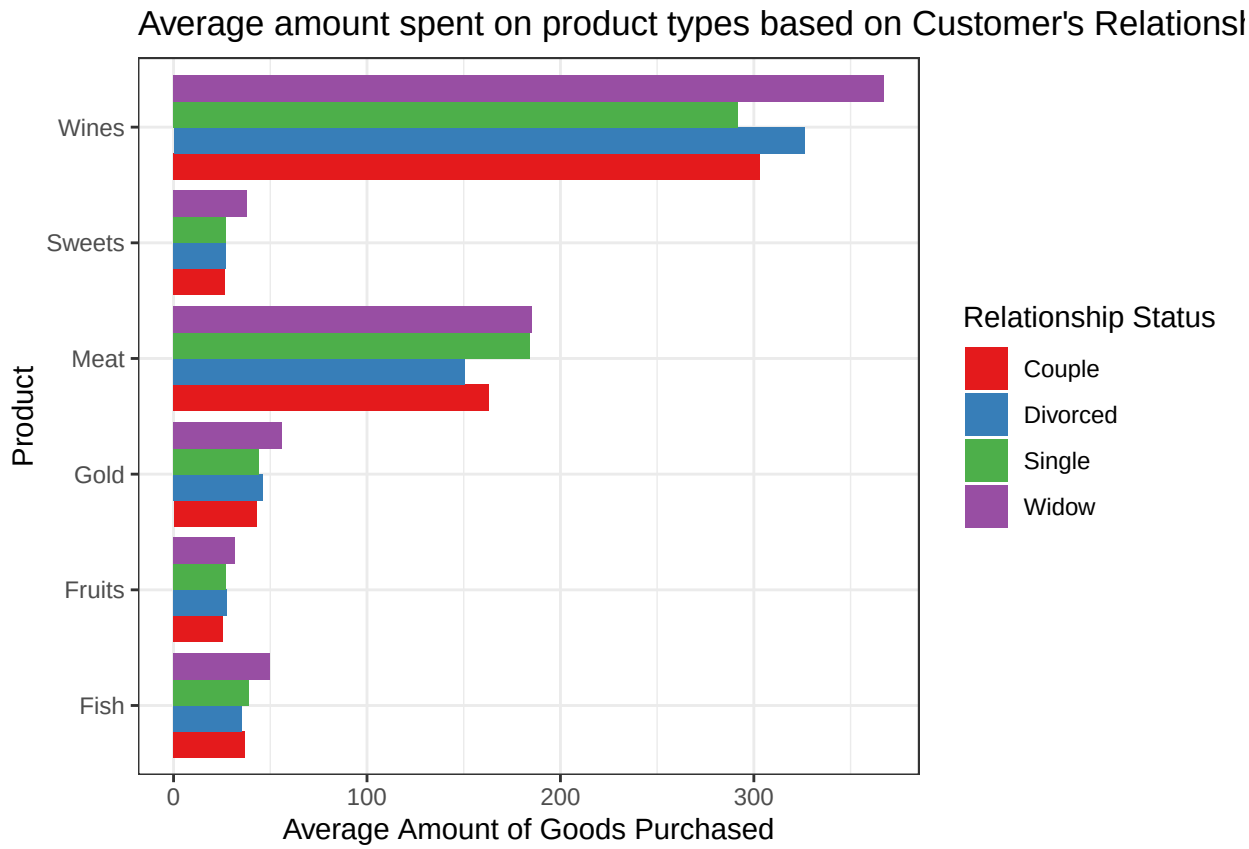
```
  labs( title="Average Amount of Goods Purchased by Age Group and Product", hjust = 0, x = "Customer's A
```




```
total_amount = aggregate(amount ~ product+Education, data = df_tidy, mean)

total_amount%>%
ggplot( aes(x = amount, y = product, fill = Education)) +
  geom_bar(stat = "identity", position = "dodge")  +
    scale_fill_brewer(palette = "Set1") +
  labs( title="Average amount spent on product types based on Customer's Education Level", hjust = 0, x
```

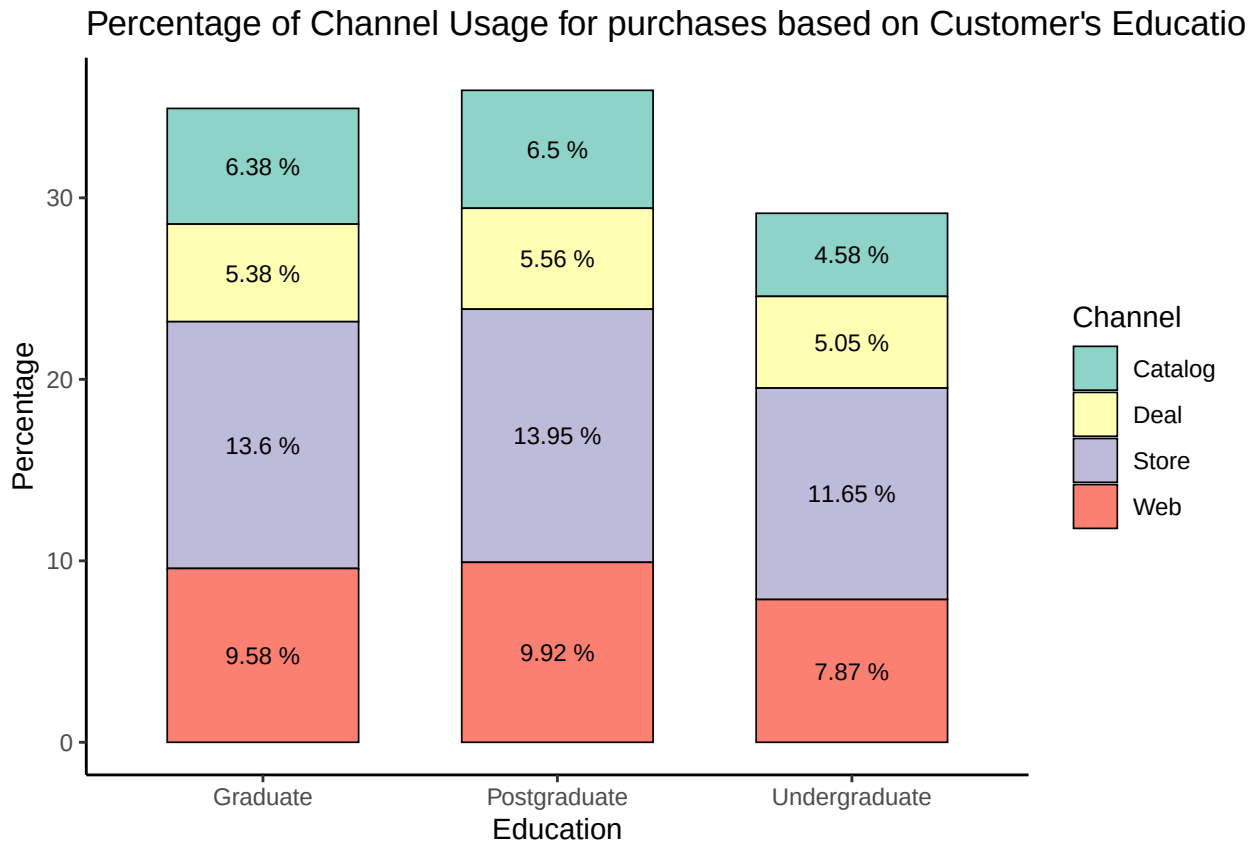



```
total_relation = aggregate(amount ~ product+RelationshipStatus, data = df_tidy, mean)

total_relation%>%
ggplot( aes(x = amount, y = product, fill = RelationshipStatus)) +
  geom_bar(stat = "identity", position = "dodge")  +
  scale_fill_brewer(palette = "Set1") +
  labs( title="Average amount spent on product types based on Customer's Relationship Status", hjust = 
```

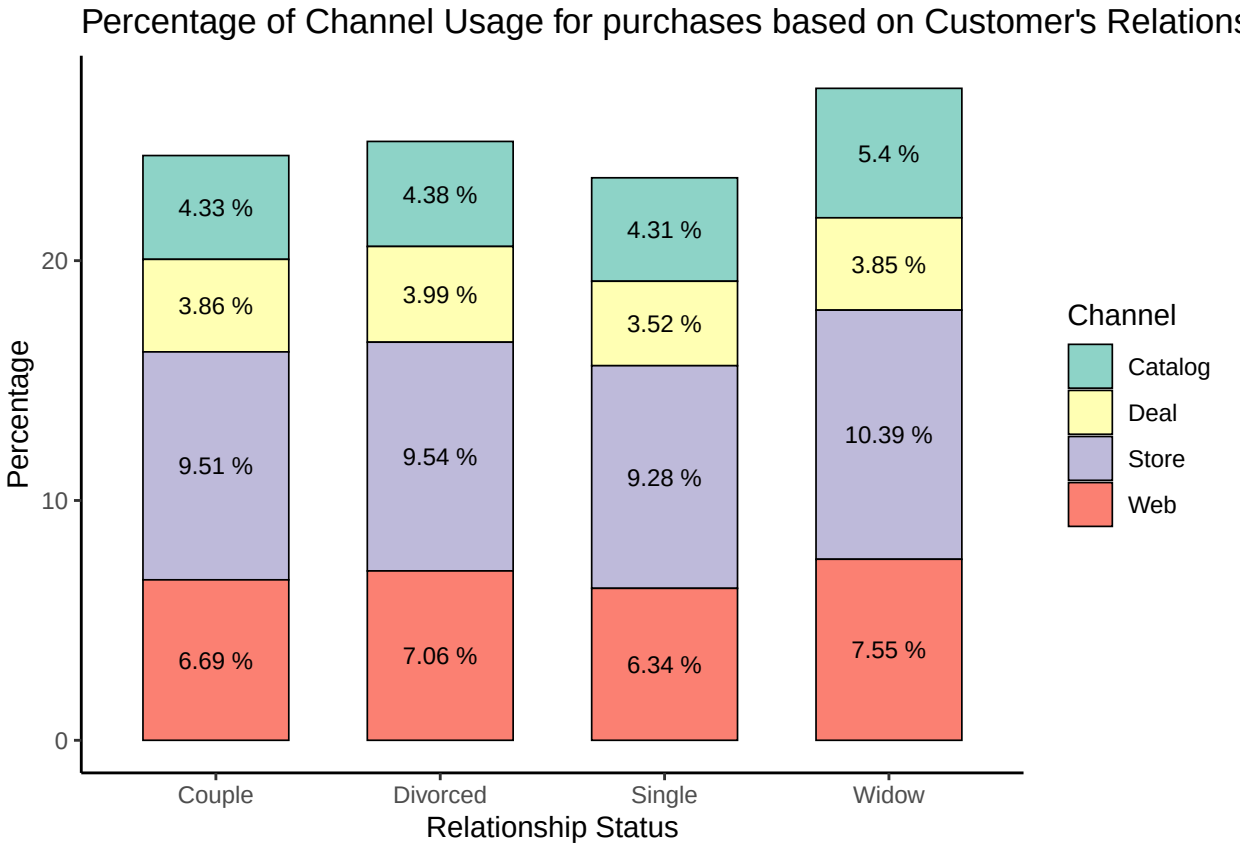



```
df_Education <- df_tidy %>%
    gather(key = "Channel", value = "Purchases", Deal, Web, Catalog, Store)%>% group_by(Education, Chan

df_Education = aggregate(Purchases ~ Channel+Education, data = df_Education, mean)

df_Education <- df_Education%>%
    mutate(Percentage = Purchases/sum(Purchases)*100)

df_Education %>%
  ggplot(aes(x=Education, y=Percentage, fill=Channel)) +
  geom_col(position="stack", color="black", width=0.65, size=0.3) +
  scale_fill_brewer(palette="Set3") +
  labs(x="Education", y="Percentage", title="Percentage of Channel Usage for purchases based on Custome
  theme_classic()+geom_text(aes(label=paste(round(Percentage,2),"%")), position=position_stack(vjust=0.5
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```
df_RelationshipStatus <- df_tidy %>%
    gather(key = "Channel", value = "Purchases", Deal, Web, Catalog, Store)%>% group_by(RelationshipSta

df_RelationshipStatus = aggregate(Purchases ~ Channel+RelationshipStatus, data = df_RelationshipStatus,
df_RelationshipStatus <- df_RelationshipStatus%>%
    mutate(Percentage = Purchases/sum(Purchases)*100)

df_RelationshipStatus %>%
  ggplot(aes(x=RelationshipStatus, y=Percentage, fill=Channel)) +
  geom_col(position="stack", color="black", width=0.65, size=0.3) +
  scale_fill_brewer(palette="Set3") +
  labs(x="Relationship Status", y="Percentage", title="Percentage of Channel Usage for purchases based
  theme_classic()+geom_text(aes(label=paste(round(Percentage,2),"%")), position=position_stack(vjust=0.5
```



Customer Segmentation Methods - PCA and K Means Clustering

```
drop <- c("age_group")
data = data[,!(names(data) %in% drop)]
```

```
subset_data = subset(data,select = !names(data) %in% c("Education", "Dt_Customer","RelationshipStatus"))
#Running a PCA.
customers_copy_pca <- PCA(subset_data, graph = FALSE)

#Exploring PCA()

# Getting the summary of the pca
summary(customers_copy_pca)
```

```
##
## Call:
## PCA(X = subset_data, graph = FALSE)
##
##
## Eigenvalues
##                        Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
## Variance               8.250   2.428   1.535   1.504   1.260   1.033   1.009
```

```
## % of var.              33.000   9.711   6.139   6.017   5.038   4.133   4.037
## Cumulative % of var.  33.000  42.710  48.850  54.866  59.904  64.037  68.074
##                        Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14
## Variance               0.922   0.802   0.779   0.747   0.658   0.637   0.576
## % of var.              3.688   3.208   3.117   2.988   2.633   2.547   2.305
## Cumulative % of var.  71.761  74.969  78.086  81.073  83.706  86.254  88.558
##                       Dim.15  Dim.16  Dim.17  Dim.18  Dim.19  Dim.20  Dim.21
## Variance               0.471   0.432   0.424   0.382   0.304   0.286   0.226
## % of var.              1.882   1.728   1.694   1.527   1.217   1.143   0.905
## Cumulative % of var.  90.440  92.169  93.863  95.390  96.607  97.750  98.654
##                       Dim.22  Dim.23  Dim.24  Dim.25
## Variance               0.179   0.158   0.000   0.000
## % of var.              0.716   0.630   0.000   0.000
## Cumulative % of var.  99.370 100.000 100.000 100.000
##
## Individuals (the 10 first)
##                Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
## 1          |  6.703 |  4.853  0.135  0.524 |  0.305  0.002  0.002 |  1.248
## 2          |  4.011 | -2.941  0.049  0.538 | -0.790  0.012  0.039 | -0.553
## 3          |  3.762 |  2.257  0.029  0.360 | -0.757  0.011  0.040 | -0.688
## 4          |  3.529 | -2.874  0.047  0.663 | -1.013  0.020  0.082 | -0.025
## 5          |  3.149 | -0.239  0.000  0.006 |  0.731  0.010  0.054 | -1.191
## 6          |  2.928 |  0.769  0.003  0.069 |  1.096  0.023  0.140 | -0.321
## 7          |  3.545 |  0.618  0.002  0.030 |  1.571  0.048  0.196 | -0.814
## 8          |  3.285 | -2.430  0.034  0.547 |  0.143  0.000  0.002 |  0.583
## 9          |  4.512 | -2.904  0.048  0.414 | -0.263  0.001  0.003 |  2.609
## 10         |  7.987 | -4.863  0.135  0.371 |  0.733  0.010  0.008 |  1.714
##              ctr   cos2
## 1          0.048  0.035 |
## 2          0.009  0.019 |
## 3          0.015  0.033 |
## 4          0.000  0.000 |
## 5          0.044  0.143 |
## 6          0.003  0.012 |
## 7          0.020  0.053 |
## 8          0.010  0.032 |
## 9          0.209  0.334 |
## 10         0.090  0.046 |
##
## Variables (the 10 first)
##              Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2
## Income     |  0.835  8.455  0.698 |  0.010  0.004  0.000 | -0.085  0.470  0.007
## Recency    |  0.009  0.001  0.000 | -0.003  0.000  0.000 | -0.310  6.252  0.096
## Complain   | -0.044  0.024  0.002 |  0.009  0.003  0.000 | -0.044  0.125  0.002
## Response   |  0.257  0.800  0.066 |  0.063  0.164  0.004 |  0.770 38.623  0.593
## Tenure     |  0.127  0.196  0.016 |  0.395  6.422  0.156 |  0.244  3.890  0.060
## Age        |  0.176  0.376  0.031 |  0.167  1.146  0.028 | -0.145  1.373  0.021
## Spendings |  0.955 11.050  0.912 |  0.068  0.190  0.005 |  0.079  0.406  0.006
## Wines      |  0.792  7.604  0.627 |  0.257  2.726  0.066 |  0.207  2.798  0.043
## Fruits     |  0.677  5.562  0.459 | -0.149  0.914  0.022 | -0.199  2.575  0.040
## Meat       |  0.819  8.136  0.671 | -0.155  0.986  0.024 |  0.031  0.062  0.001
##
## Income     |
## Recency    |
```

```
## Complain  |
## Response  |
## Tenure    |
## Age       |
## Spendings |
## Wines     |
## Fruits    |
## Meat      |
```

```
#Getting the variance of the first 7 new dimensions
customers_copy_pca$eig[,2][1:7]
```

```
##    comp 1    comp 2    comp 3    comp 4    comp 5    comp 6    comp 7
## 32.999703  9.710722  6.139137  6.016500  5.038052  4.132653  4.036781
```

```
#Getting the cummulative variance
customers_copy_pca$eig[,3][1:7]
```

```
##   comp 1   comp 2   comp 3   comp 4   comp 5   comp 6   comp 7
## 32.99970 42.71043 48.84956 54.86606 59.90411 64.03677 68.07355
```

```
#Getting the most correlated variables
dimdesc(customers_copy_pca, axes = 1:2)
```

```
## $Dim.1
##
## Link between the variable and the continuous variables (R-square)
## ================================================================================
##            correlation       p.value
## Spendings   0.95478584  0.000000e+00
## Income      0.83519207  0.000000e+00
## Catalog     0.82765387  0.000000e+00
## Meat        0.81928914  0.000000e+00
## Purchases   0.80711731  0.000000e+00
## Wines       0.79205159  0.000000e+00
## Store       0.73827366  0.000000e+00
## Fish        0.70543139 2.964987e-319
## Sweets      0.68139378 1.159298e-289
## Fruits      0.67740831 4.978610e-285
## Web         0.56934876 1.314605e-182
## Gold        0.56450347 7.071670e-179
## Campaign    0.42985424  3.620979e-96
## Response    0.25696363  2.391945e-33
## Age         0.17612089  3.027745e-16
## Tenure      0.12727947  4.013419e-09
## widow       0.06222502  4.137337e-03
## LevEd       0.04920067  2.342236e-02
## Complain   -0.04449088  4.043404e-02
## Deal       -0.09290322  1.816296e-05
## Children   -0.59269510 1.741520e-201
## WebVisits  -0.60094168 1.570792e-208
## Parent     -0.60724228 4.714693e-214
```

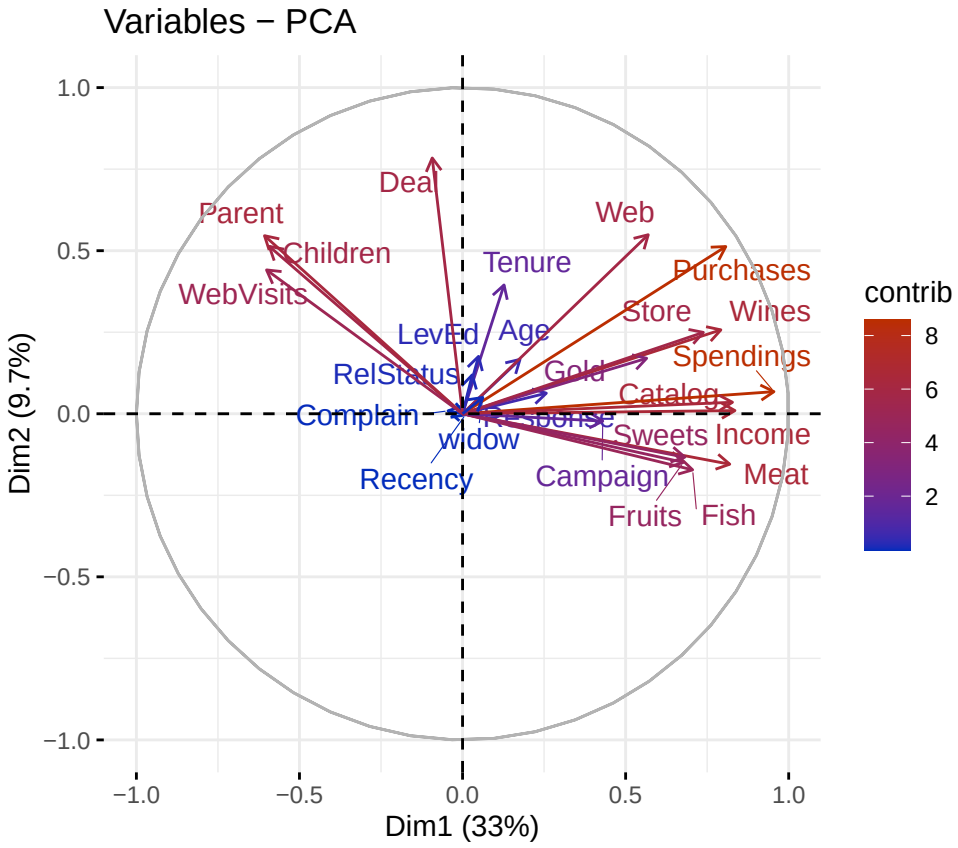
```
##
## $Dim.2
##
## Link between the variable and the continuous variables (R-square)
## ================================================================================
##             correlation       p.value
## Deal         0.78419238  0.000000e+00
## Web          0.54886044 3.029537e-167
## Parent       0.54590289 4.091940e-165
## Children     0.51328682 5.672970e-143
## Purchases    0.51263527 1.485804e-142
## WebVisits    0.44107146 9.946749e-102
## Tenure       0.39486198  4.073735e-80
## Wines        0.25725506  2.015560e-33
## Store        0.24946515  1.817152e-31
## LevEd        0.17597755  3.202131e-16
## Gold         0.17009965  3.055809e-15
## Age          0.16676121  1.062413e-14
## RelStatus    0.11908016  3.759291e-08
## Spendings    0.06782678  1.770741e-03
## Response     0.06310381  3.636629e-03
## widow        0.05001114  2.123074e-02
## Sweets      -0.13315597  7.381983e-10
## Fruits      -0.14894602  5.360846e-12
## Meat        -0.15468289  7.809640e-13
## Fish        -0.17199746  1.487876e-15
```

```
#Tracing variable contributions in customers_pca
customers_copy_pca$var$contrib
```

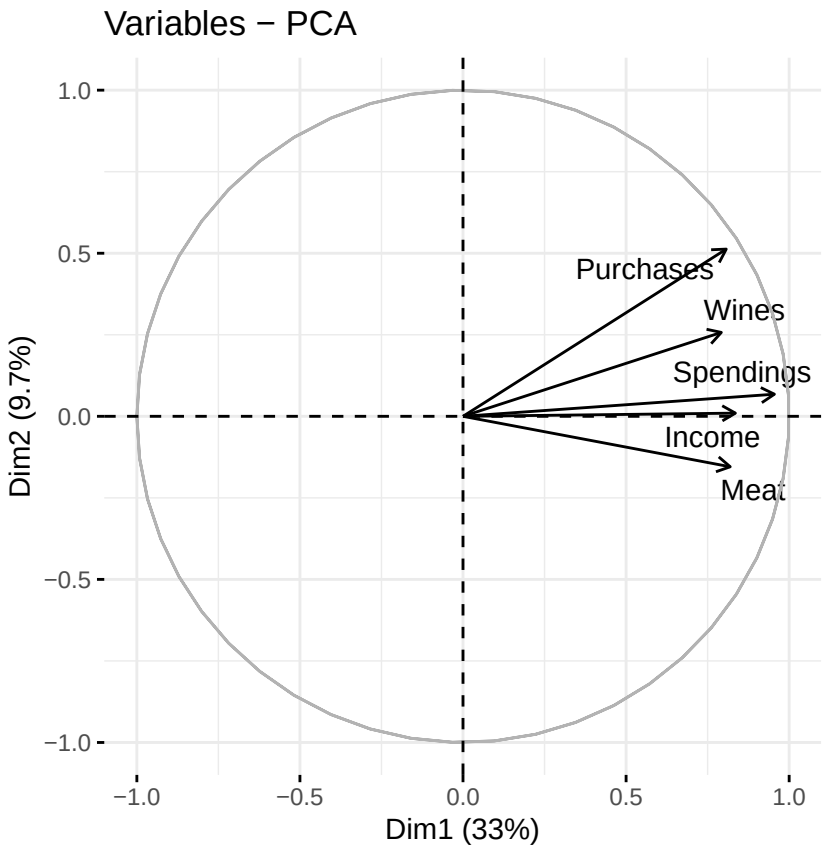
```
##                   Dim.1        Dim.2        Dim.3        Dim.4       Dim.5
## Income     8.455177e+00 3.887844e-03  0.469900516 4.291178e+00  1.11831375
## Recency    9.493675e-04 4.463176e-04  6.252201343 9.871868e-04  0.24244508
## Complain   2.399341e-02 3.276266e-03  0.125244113 5.201794e-01  0.04837416
## Response   8.003746e-01 1.640286e-01 38.623305042 2.936469e-01  1.23598182
## Tenure     1.963662e-01 6.422426e+00  3.890142468 1.495988e+01  3.67606407
## Age        3.759860e-01 1.145509e+00  1.372677441 1.938078e+01  3.70081999
## Spendings  1.104999e+01 1.895007e-01  0.405588202 9.375551e-03  0.17237076
## Wines      7.604259e+00 2.726066e+00  2.798244224 1.830455e+00  0.99705781
## Fruits     5.562256e+00 9.138318e-01  2.575228385 3.745204e+00  0.82113554
## Meat       8.136251e+00 9.855825e-01  0.062186500 3.488225e-02  0.24460962
## Fish       6.031975e+00 1.218576e+00  2.373390452 3.704340e+00  1.08706178
## Sweets     5.627899e+00 7.303479e-01  2.118758224 3.384860e+00  1.01126043
## Gold       3.862631e+00 1.191833e+00  0.362013388 3.738276e+00  1.52803376
## RelStatus  1.072216e-02 5.841001e-01  0.266283991 4.010777e+00 34.95223737
## Children   4.258068e+00 1.085247e+01  1.648360656 9.564062e-01  0.45303805
## Parent     4.469655e+00 1.227550e+01  1.728911278 5.136981e-02  0.51893315
## LevEd      2.934215e-02 1.275625e+00  2.253311685 2.577118e+01  5.24831332
## Campaign   2.239713e+00 2.082997e-02 23.257204120 1.184850e+00  0.16743619
## Purchases  7.896294e+00 1.082494e+01  0.971040529 3.001113e-03  0.49936174
## WebVisits  4.377384e+00 8.013576e+00  4.637882001 5.566767e+00  0.95570427
## Web        3.929223e+00 1.240887e+01  0.044781299 1.772480e-01  0.04385523
## Deal       1.046192e-01 2.533108e+01  0.590914723 9.434058e-01  0.14155081
## Catalog    8.303237e+00 5.120136e-02  0.007307817 2.287045e-01  0.17387255
```

```
## Store     6.606702e+00 2.563470e+00  3.130587816 1.406645e-01  0.80236566
## widow     4.693319e-02 1.030248e-01  0.034533785 5.071590e+00 40.15980308
```

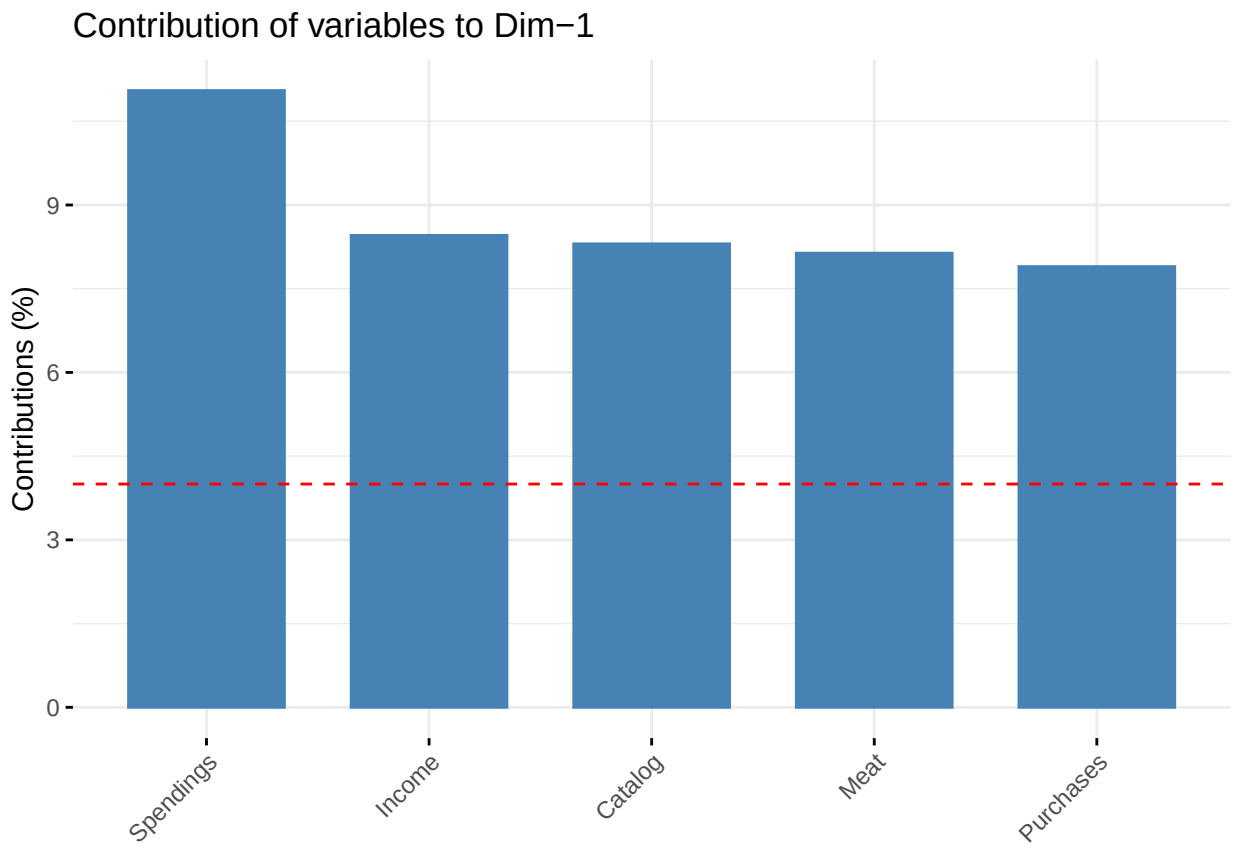
```
#Creating a factor map for the variable contributions
fviz_pca_var(customers_copy_pca, col.var = "contrib", gradient.cols = c("#002bbb", "#bb2e00"), repel = T
```



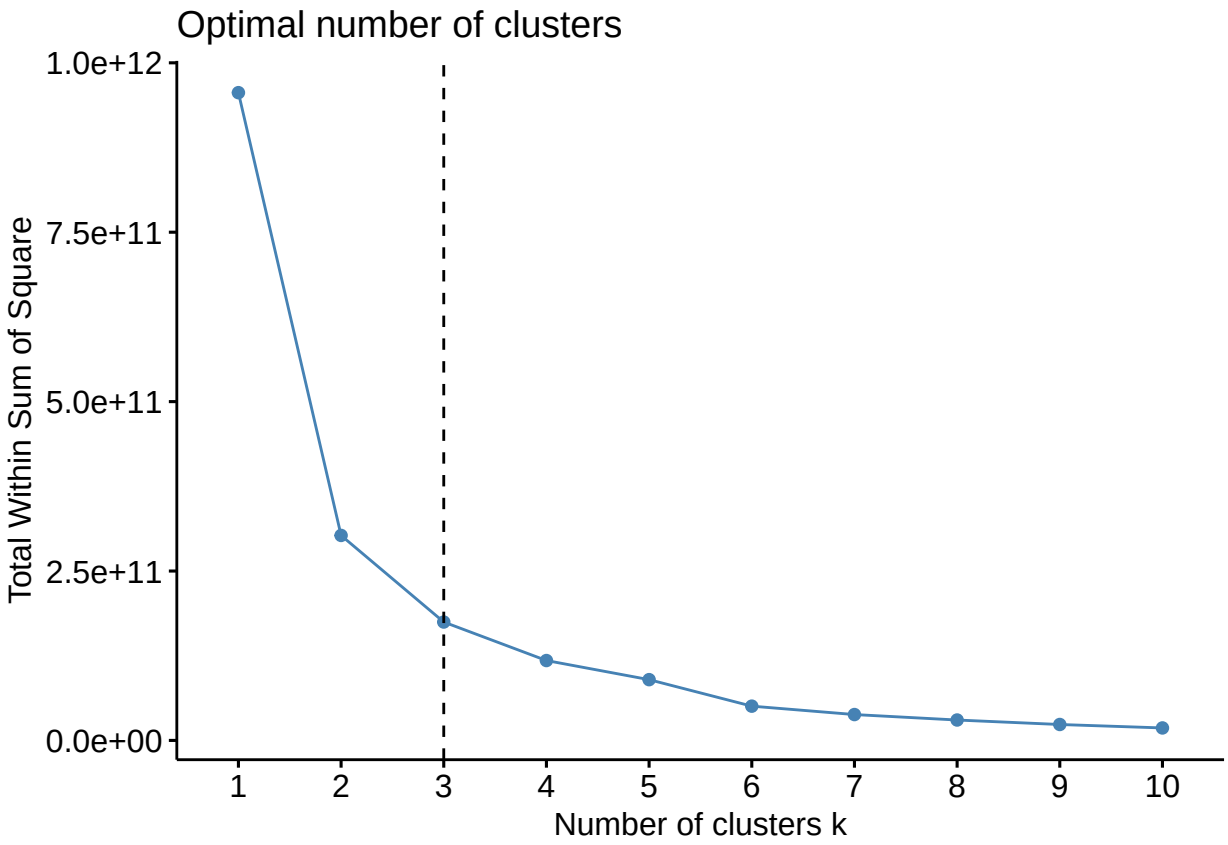

```
#Creating a factor map for the top 5 variables with the highest contributions.
fviz_pca_var(customers_copy_pca, select.var = list(contrib = 5), repel = TRUE)
```




```
fviz_contrib(customers_copy_pca, choice = "var", axes = 1, top = 5)
```




```
fviz_nbclust(subset_data,kmeans,method="wss")+geom_vline(xintercept=3,linetype=2)
```




```
# Compute correlation matrix
corr_data = subset(data,select = !names(data) %in% c("Education", "Dt_Customer","RelationshipStatus", "
```

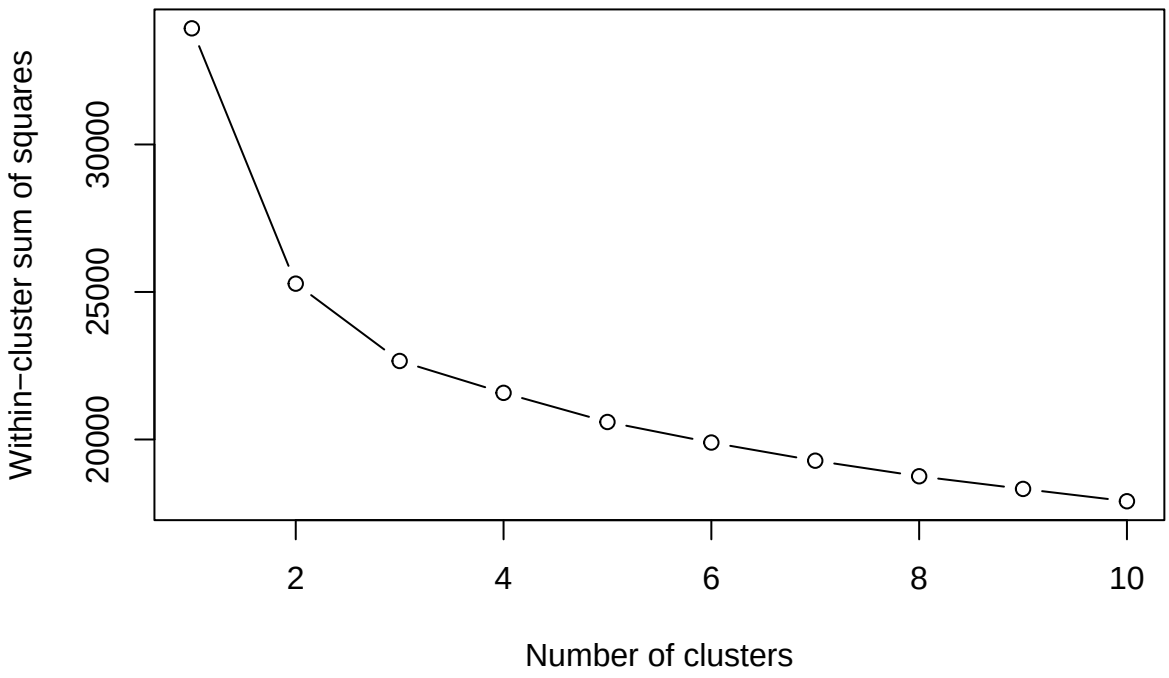
```
cor_matrix <- cor(corr_data)

highly_correlated_features <- findCorrelation(cor_matrix, cutoff = 0.70)

corr_data <- corr_data[,-highly_correlated_features]
```

```
# Normalize the data
scaled_corr_data <- scale(corr_data)
```

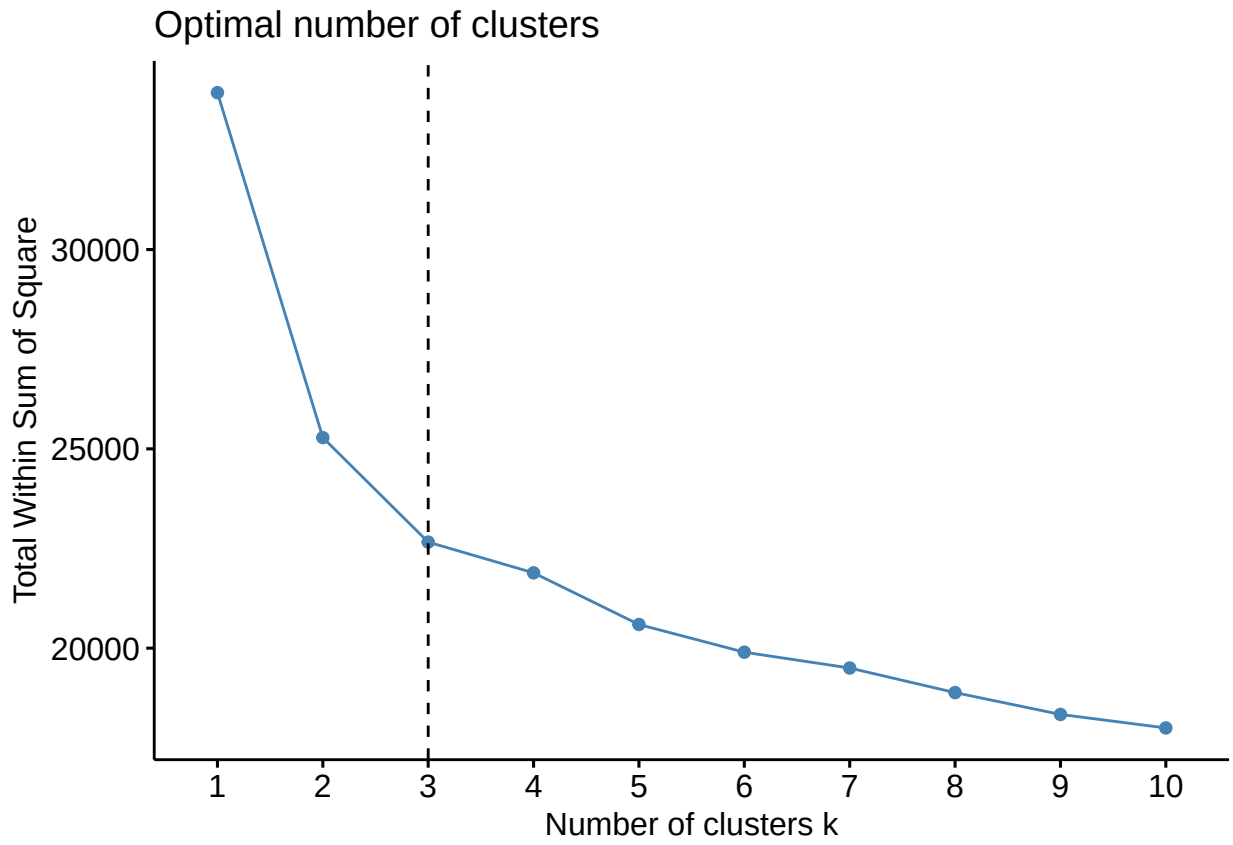
```
# Determine the optimal number of clusters using the elbow method
wss <- c()
for (i in 1:10) {
  kmeans_model <- kmeans(scaled_corr_data, centers = i, nstart = 10)
  wss[i] <- kmeans_model$tot.withinss
}
```

```
## Warning: did not converge in 10 iterations
```

```
plot(1:10, wss, type = "b", xlab = "Number of clusters", ylab = "Within-cluster sum of squares")
```

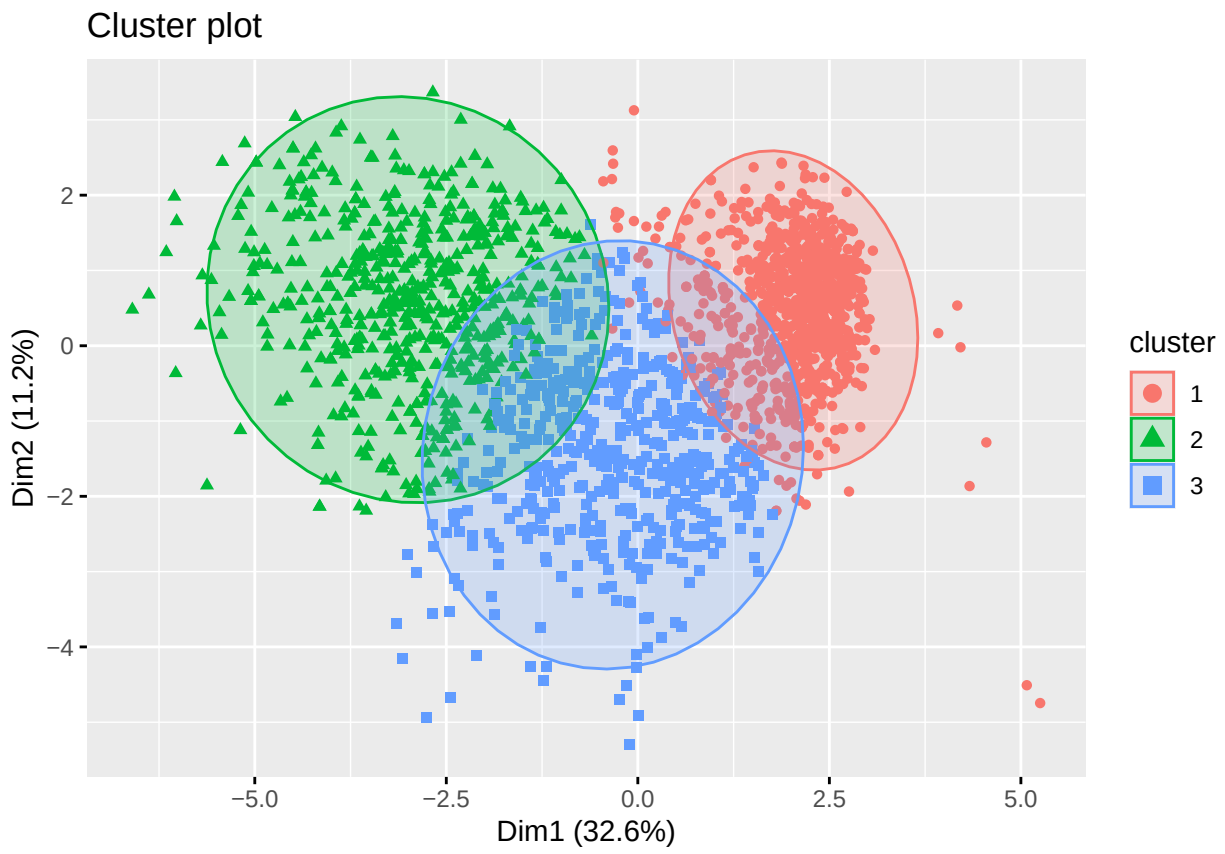



```
elbow_point <- fviz_nbclust(scaled_corr_data, kmeans, method = "wss") + geom_vline(xintercept = 3, linet
print(elbow_point)
```

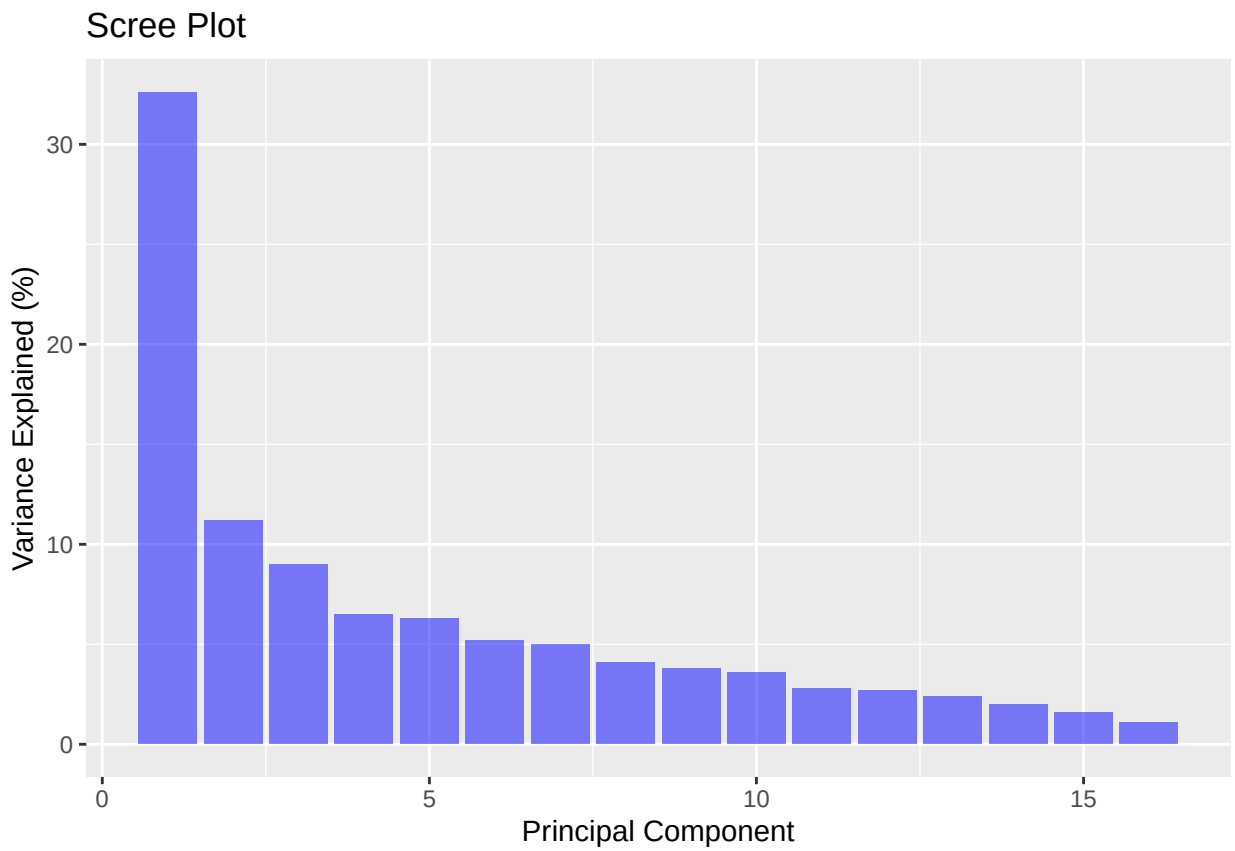



```r
set.seed(123)

# Perform k-means clustering on the dataset
kmeans_model <- kmeans(scaled_corr_data, centers = 3, nstart = 10)
cluster_assignments <- kmeans_model$cluster
```

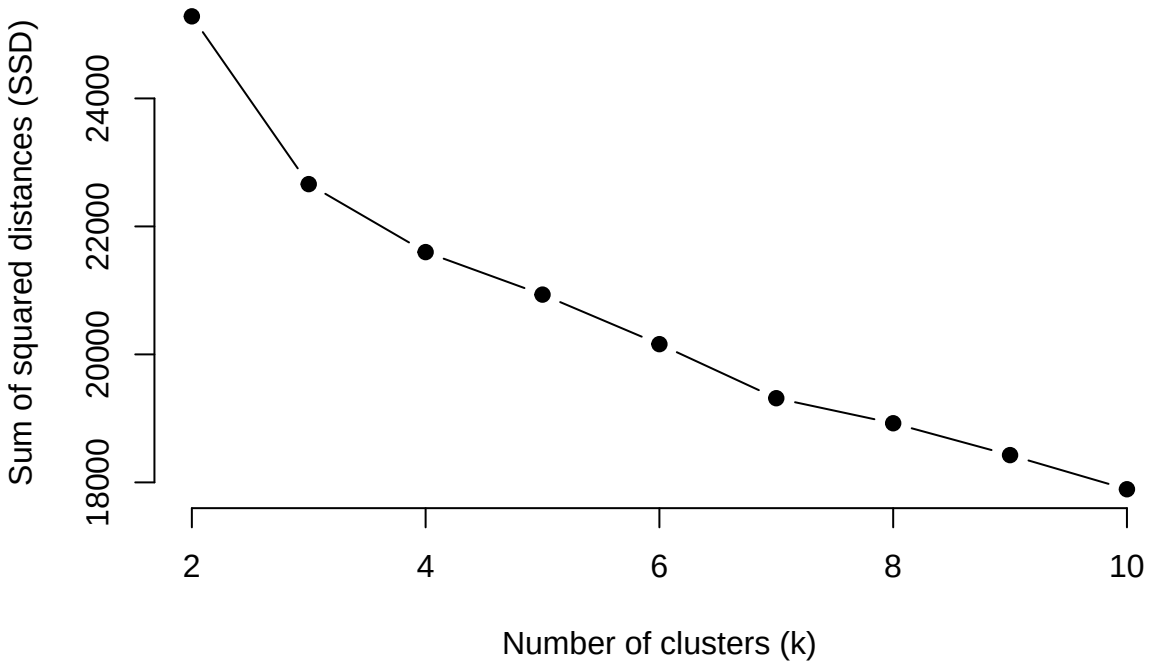
```r
fviz_cluster(kmeans_model, scaled_corr_data, geom = "point",ellipse.type = "norm", repel = TRUE)
```



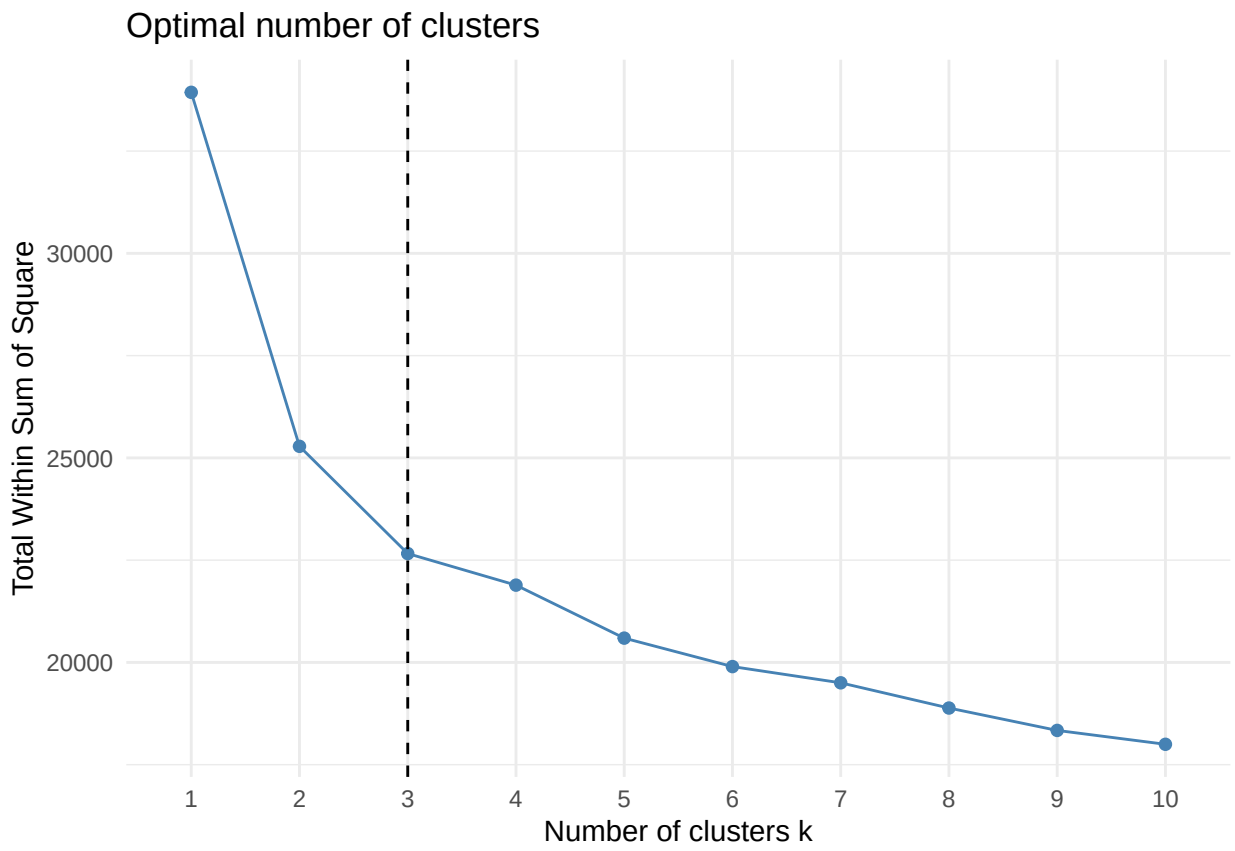

```r
pca <- prcomp(scaled_corr_data, scale = TRUE)

# Calculate variance explained by each principal component
prop_var <- round(pca$sdev^2/sum(pca$sdev^2)*100, 1)

# Plot variance explained by each principal component
var_plot <- ggplot(data.frame(PC = 1:length(prop_var), prop_var), aes(x = PC, y = prop_var)) +
  geom_bar(stat = "identity", fill = "blue", alpha = 0.5) +
  labs(x = "Principal Component", y = "Variance Explained (%)") +
  ggtitle("Scree Plot")
print(var_plot)
```




```r
# perform k-means clustering for a range of k values
k_values <- 2:10
ssd <- sapply(k_values, function(k) {
  kmeans(pca$x, centers = k)$tot.withinss
})

# plot SSD values against k values
plot(k_values, ssd, type = "b", pch = 19, frame = FALSE, xlab = "Number of clusters (k)", ylab = "Sum o
```

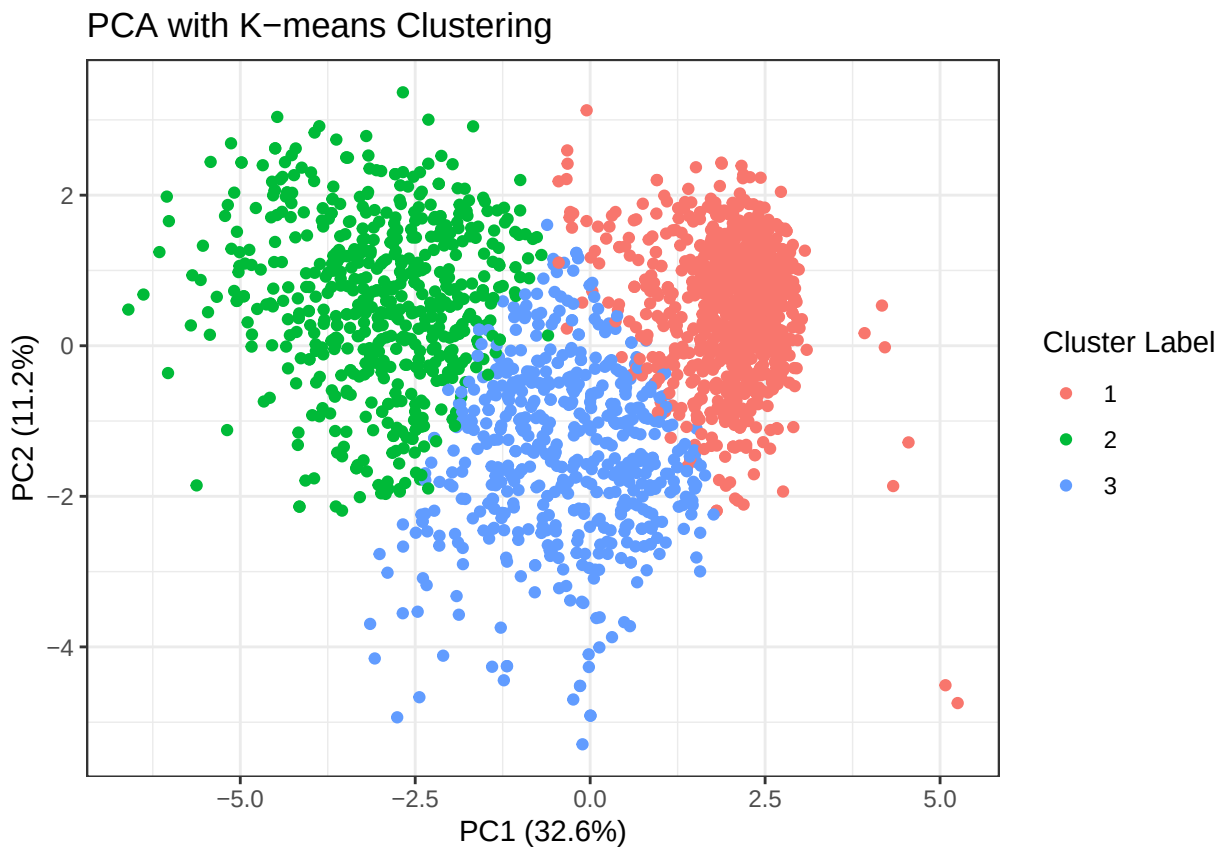



```
# identify elbow point
fviz_nbclust(pca$x, kmeans, method = "wss", k.max = 10)+ geom_vline(xintercept = 3, linetype = "dashed")
```




```r
# Choose the number of principal components
num_pc <- 10

# Extract the selected principal components
pca_sel <- data.frame(pca$x)

# Add the selected principal components to the original data
data_pca <- cbind(scaled_corr_data, pca_sel)
```

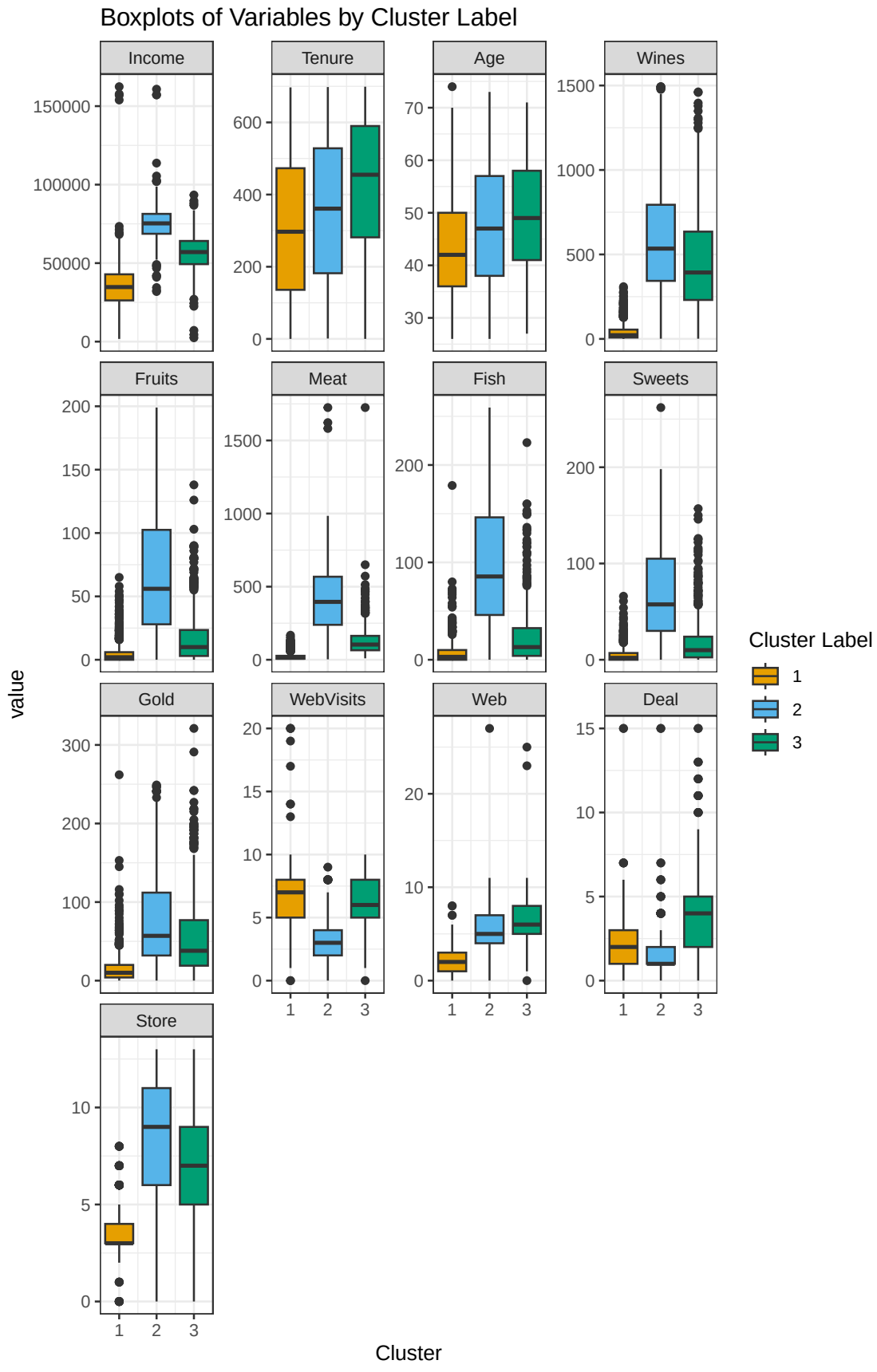
```r
set.seed(123)
kmeans_pca <- kmeans(pca_sel, centers = 3, nstart = 25)

# Add cluster labels to the original data
data_pca_clustered <- cbind(data_pca, Cluster = kmeans_pca$cluster)

# Visualize the clusters
ggplot(data_pca_clustered, aes(x = PC1, y = PC2, color = factor(Cluster))) +
  geom_point() +
  labs(x = paste0("PC1 (", prop_var[1], "%)"), y = paste0("PC2 (", prop_var[2], "%)"), title = "PCA wit
  theme_bw()
```




```
vars <- c("Income", "Tenure", "Age", "Wines", "Fruits", "Meat", "Fish", "Sweets", "Gold", "WebVisits",

cluster_data = corr_data[ ,vars]
```

```
# Add cluster labels to the original data
data_pca_clustered <- cbind(cluster_data, Cluster = kmeans_pca$cluster)
```

```
# plot boxplots for each variable, colored by cluster label
ggplot(
  melt(data_pca_clustered, id.vars = "Cluster"),
  aes(
    x = Cluster,
    y = value,
    group = Cluster,
    fill = factor(Cluster)
  )
) +
  geom_boxplot() + scale_fill_manual(values =  c("#E69F00", "#56B4E9", "#009E73")) +
  facet_wrap( ~ variable, scales = "free_y") + theme_bw() +
  labs(fill = "Cluster Label", title = "Boxplots of Variables by Cluster Label")
```


Boxplots of Variables by Cluster Label
Income
Tenure
Age
Wines
Fruits
Meat
Fish
Sweets
Gold
WebVisits
Web
Deal
Store
value
Cluster
Cluster Label
1
2
3