# segment 2

assignment for Hight Dimensional Data Anlysis

2025-02-05

## Meta data

Dataset Description: E-commerce Customer Behavior

Overview: This dataset provides a comprehensive view of customer behavior within an e-commerce platform. Each entry in the dataset corresponds to a unique customer, offering a detailed breakdown of their interactions and transactions. The information is crafted to facilitate a nuanced analysis of customer preferences, engagement patterns, and satisfaction levels, aiding businesses in making data-driven decisions to enhance the customer experience.

Columns: - Customer ID (Type: Numeric): Description: A unique identifier assigned to each customer, ensuring distinction across the dataset.

- Gender: Type: Categorical (Male, Female) Description: Specifies the gender of the customer, allowing for gender-based analytics.
- Age: Type: Numeric Description: Represents the age of the customer, enabling age-group-specific insights.
- City: Type: Categorical (City names) Description: Indicates the city of residence for each customer, providing geographic insights.
- Membership Type: Type: Categorical (Gold, Silver, Bronze) Description: Identifies the type of membership held by the customer, influencing perks and benefits.
- Total Spend: Type: Numeric Description: Records the total monetary expenditure by the customer on the e-commerce platform.
- Items Purchased: Type: Numeric Description: Quantifies the total number of items purchased by the customer.
- Average Rating: Type: Numeric (0 to 5, with decimals) Description: Represents the average rating given by the customer for purchased items, gauging satisfaction.
- Discount Applied: Type: Boolean (True, False) Description: Indicates whether a discount was applied to the customer's purchase, influencing buying behavior.
- Days Since Last Purchase: Type: Numeric Description: Reflects the number of days elapsed since the customer's most recent purchase, aiding in retention analysis.
- Satisfaction Level: Type: Categorical (Satisfied, Neutral, Unsatisfied) Description: Captures the overall satisfaction level of the customer, providing a subjective measure of their experience.

## LOad necessaries libraries

```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(ggplot2)
library(scales)
```

# LOad the data set

```r
df = read.csv("E-commerce_Customer_Behavior.csv")
str(df)
```

```
## 'data.frame':    350 obs. of  11 variables:
##  $ Customer.ID             : int  101 102 103 104 105 106 107 108 109 110 ...
##  $ Gender                  : chr  "Female" "Male" "Female" "Male" ...
##  $ Age                     : int  29 34 43 30 27 37 31 35 41 28 ...
##  $ City                    : chr  "New York" "Los Angeles" "Chicago" "San Francisco" ...
##  $ Membership.Type         : chr  "Gold" "Silver" "Bronze" "Gold" ...
##  $ Total.Spend             : num  1120 780 511 1480 720 ...
##  $ Items.Purchased         : int  14 11 9 19 13 8 15 12 10 21 ...
##  $ Average.Rating          : num  4.6 4.1 3.4 4.7 4 3.1 4.5 4.2 3.6 4.8 ...
##  $ Discount.Applied        : logi  TRUE FALSE TRUE FALSE TRUE FALSE ...
##  $ Days.Since.Last.Purchase: int  25 18 42 12 55 22 28 14 40 9 ...
##  $ Satisfaction.Level      : chr  "Satisfied" "Neutral" "Unsatisfied" "Satisfied" ...
```

```r
summary(df)
```

```
##   Customer.ID       Gender               Age             City          
##  Min.   :101.0   Length:350         Min.   :26.0   Length:350        
##  1st Qu.:188.2   Class :character   1st Qu.:30.0   Class :character  
##  Median :275.5   Mode  :character   Median :32.5   Mode  :character  
##  Mean   :275.5                      Mean   :33.6                     
##  3rd Qu.:362.8                      3rd Qu.:37.0                     
##  Max.   :450.0                      Max.   :43.0                     
##  Membership.Type     Total.Spend     Items.Purchased Average.Rating 
##  Length:350         Min.   : 410.8   Min.   : 7.0    Min.   :3.000  
##  Class :character   1st Qu.: 502.0   1st Qu.: 9.0    1st Qu.:3.500  
##  Mode  :character   Median : 775.2   Median :12.0    Median :4.100  
##                     Mean   : 845.4   Mean   :12.6    Mean   :4.019  
##                     3rd Qu.:1160.6   3rd Qu.:15.0    3rd Qu.:4.500  
##                     Max.   :1520.1   Max.   :21.0    Max.   :4.900  
##  Discount.Applied Days.Since.Last.Purchase Satisfaction.Level
##  Mode :logical    Min.   : 9.00            Length:350        
##  FALSE:175        1st Qu.:15.00            Class :character  
##  TRUE :175        Median :23.00            Mode  :character  
##                   Mean   :26.59                              
##                   3rd Qu.:38.00                              
##                   Max.   :63.00                              
```

This data set counts 11 characteristics with 350 custmers. These characteristics are categoricals (Gender,City ,Membership.Type, Satisfaction.Level ), logique variable ( Discount.Applied) and numericals (Age, Total.spend, Items.purchased, Average.Rating and Days.Since.Last.Purchase).

Let's transform all catagorical variable as factor.

```
df$Gender = as.factor(df$Gender)
df$City = as.factor(df$City)
df$Membership.Type = as.factor(df$Membership.Type)
df$Satisfaction.Level = as.factor(df$Satisfaction.Level)
```

## EXploratory of data

### Univariate

```
summary(df)
```

```
##   Customer.ID       Gender          Age                     City    Membership.Type
##  Min.   :101.0   Female:175   Min.   :26.0   Chicago      :58   Bronze:116     
##  1st Qu.:188.2   Male  :175   1st Qu.:30.0   Houston      :58   Gold  :117     
##  Median :275.5                Median :32.5   Los Angeles  :59   Silver:117     
##  Mean   :275.5                Mean   :33.6   Miami        :58                  
##  3rd Qu.:362.8                3rd Qu.:37.0   New York     :59                  
##  Max.   :450.0                Max.   :43.0   San Francisco:58                  
##   Total.Spend     Items.Purchased Average.Rating  Discount.Applied
##  Min.   : 410.8   Min.   : 7.0    Min.   :3.000   Mode :logical   
##  1st Qu.: 502.0   1st Qu.: 9.0    1st Qu.:3.500   FALSE:175       
##  Median : 775.2   Median :12.0    Median :4.100   TRUE :175       
##  Mean   : 845.4   Mean   :12.6    Mean   :4.019                   
##  3rd Qu.:1160.6   3rd Qu.:15.0    3rd Qu.:4.500                   
##  Max.   :1520.1   Max.   :21.0    Max.   :4.900                   
##  Days.Since.Last.Purchase   Satisfaction.Level
##  Min.   : 9.00                     :  2       
##  1st Qu.:15.00            Neutral    :107       
##  Median :23.00            Satisfied  :125       
##  Mean   :26.59            Unsatisfied:116       
##  3rd Qu.:38.00                                
##  Max.   :63.00
```

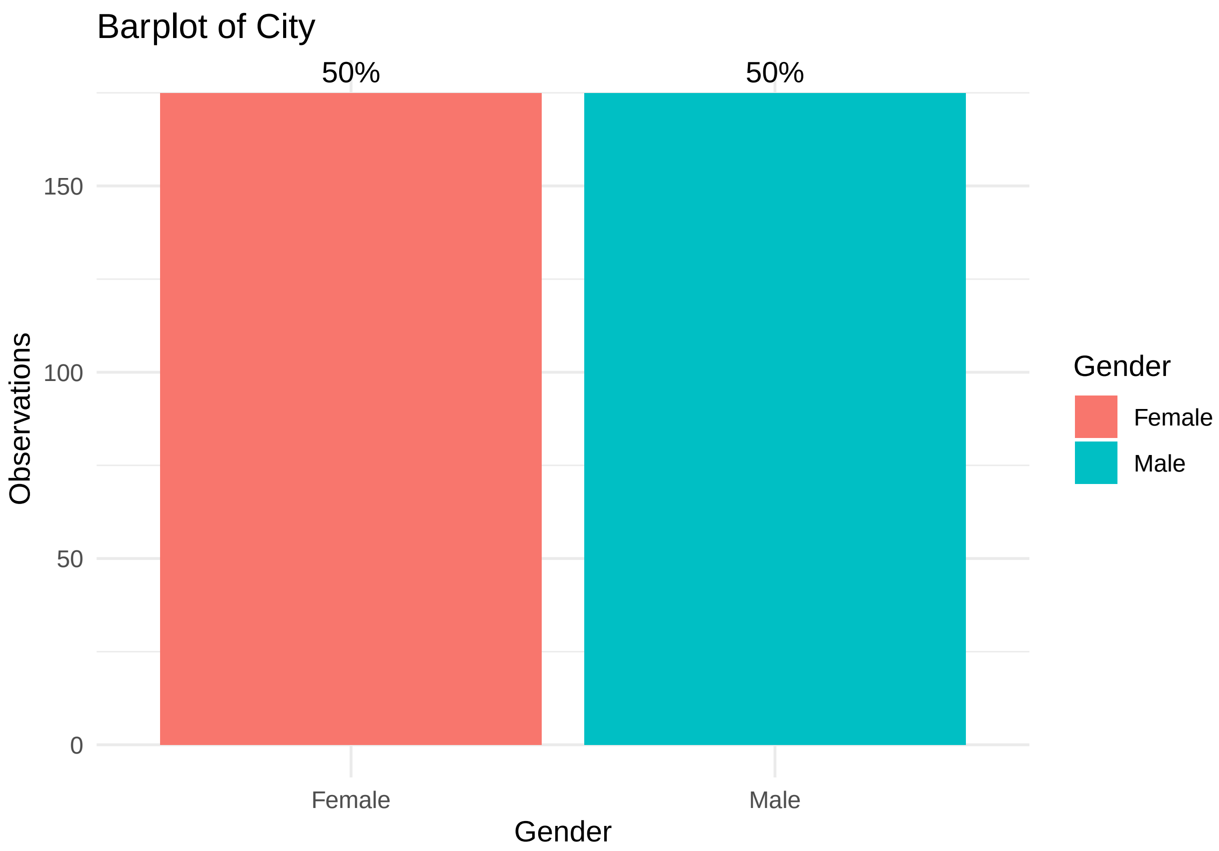
```
duplicated(df)
```

```
##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [193] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [205] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [217] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [229] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [241] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [253] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [265] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [277] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [289] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [301] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [313] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [325] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [337] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [349] FALSE FALSE
```

*Gender*

```r
df_counts <- df %>%
  group_by(Gender) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

#barplot
ggplot(df_counts, aes(x = Gender, y = count, fill = Gender)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), vjust = -0.5) +
  labs(title = "Barplot of City", x = "Gender", y = "Observations") +
  theme_minimal()
```

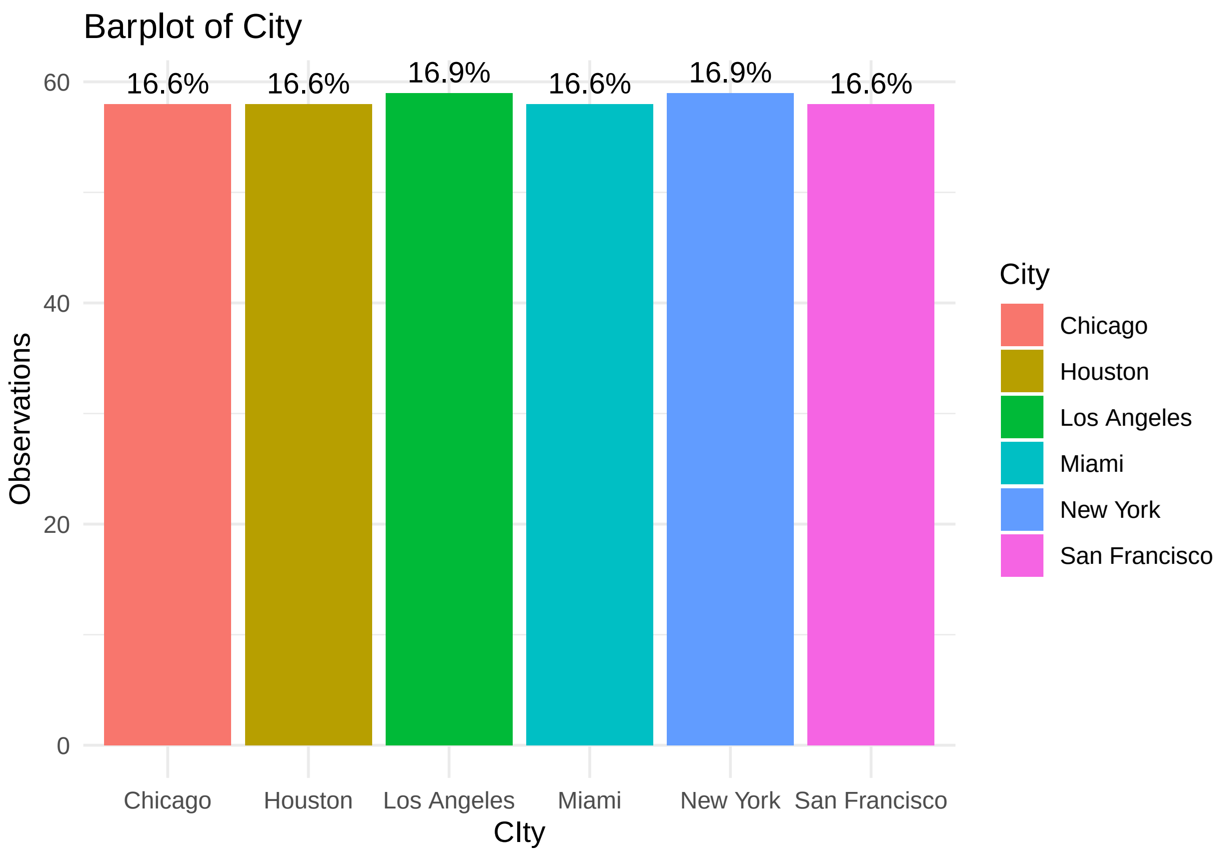


We have as many women as men in our dataset.

*City*

```
df_counts <- df %>%
  group_by(City) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

#barplot
ggplot(df_counts, aes(x = City, y = count, fill = City)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), vjust = -0.5) +
  labs(title = "Barplot of City", x = "CIty", y = "Observations") +
  theme_minimal()
```

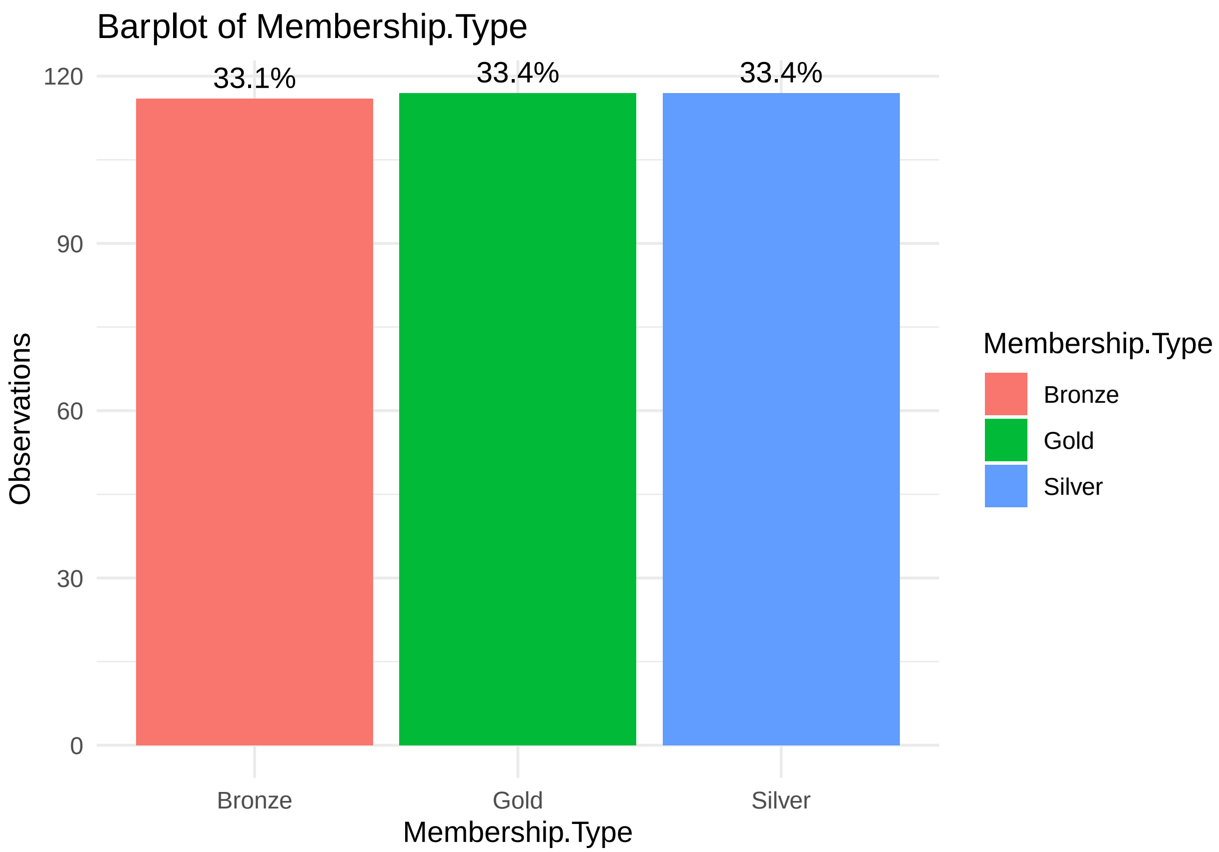


These customers are from 06 city that are Chicago, Houston, LOs angles, Miami, New york and San Francisco. We have more people who are from New york (16.9%) and LOs ANgeles (16.9%) than anothers city.

*Membership.Type*

```
df_counts <- df %>%
  group_by(Membership.Type) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

#barplot
ggplot(df_counts, aes(x = Membership.Type, y = count, fill = Membership.Type)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), vjust = -0.5) +
  labs(title = "Barplot of Membership.Type", x = "Membership.Type", y = "Observations") +
  theme_minimal()
```

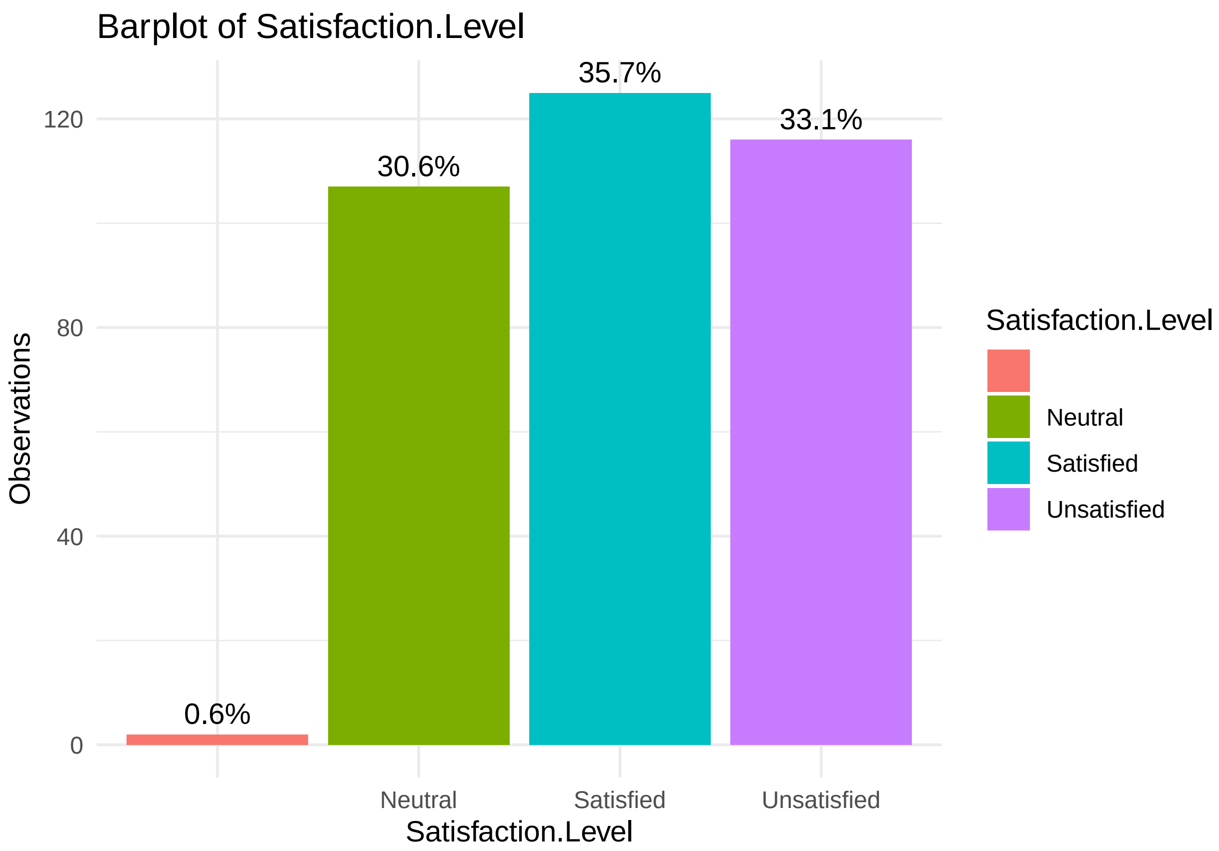


We have 03 types's membership of the customer, and their numbers are almost the same (33.1% for Bronze, 33.4% for Gold and 33.4% for Silver)

*Satisfaction.Level*

```
#summary(df$Satisfaction.Level)
df_counts <- df %>%
  group_by(Satisfaction.Level) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

#barplot
ggplot(df_counts, aes(x = Satisfaction.Level, y = count, fill = Satisfaction.Level)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")), vjust = -0.5) +
  labs(title = "Barplot of Satisfaction.Level", x = "Satisfaction.Level", y = "Observations") +
  theme_minimal()
```



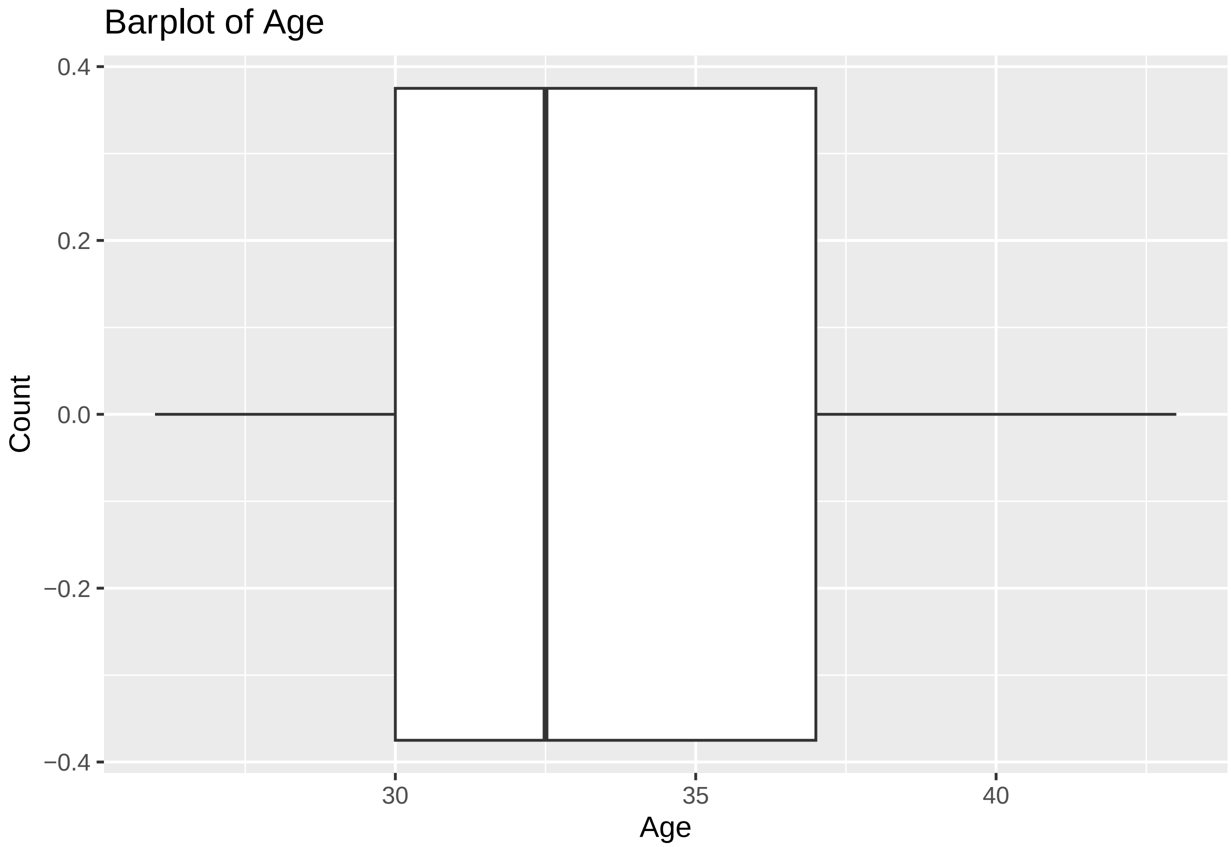
We have 03 levels of satisfactions: Neutral, Satisfied and Unsatified. Also, there are some customers who didn't give their satisfaction's level. There are more people who are satified (35.7%) than the other level (30.6% for Neutral, 33.1% for Unsatified, and 0.6% for those who didn't give their levels.).

*Age*

```
summary(df$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    26.0    30.0    32.5    33.6    37.0    43.0
```

```
ggplot(df, aes(x=Age)) + geom_boxplot() + labs(title = "Barplot of Age", x="Age", y="Count")
```



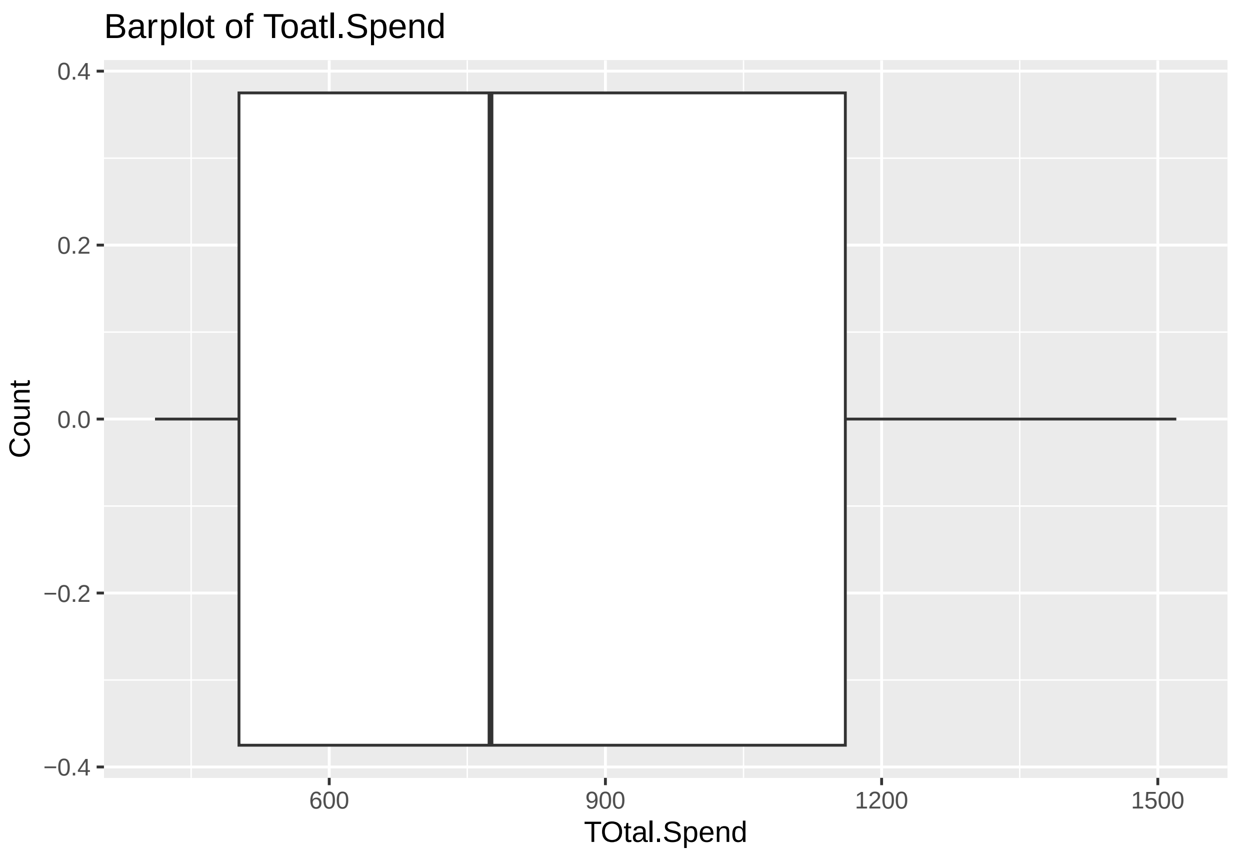
THe customer's age vary between 26 and 43 years old. And the clientele is mostly made up of young to middle-aged adults, mainly in the second age bracket (30-40 years).

## TOtal spend

```
summary(df$Total.Spend)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   410.8   502.0   775.2   845.4  1160.6  1520.1
```

```
ggplot(df, aes(x=Total.Spend)) + geom_boxplot() + labs(title="Barplot of Toatl.Spend", x="TOtal.Spend",
```

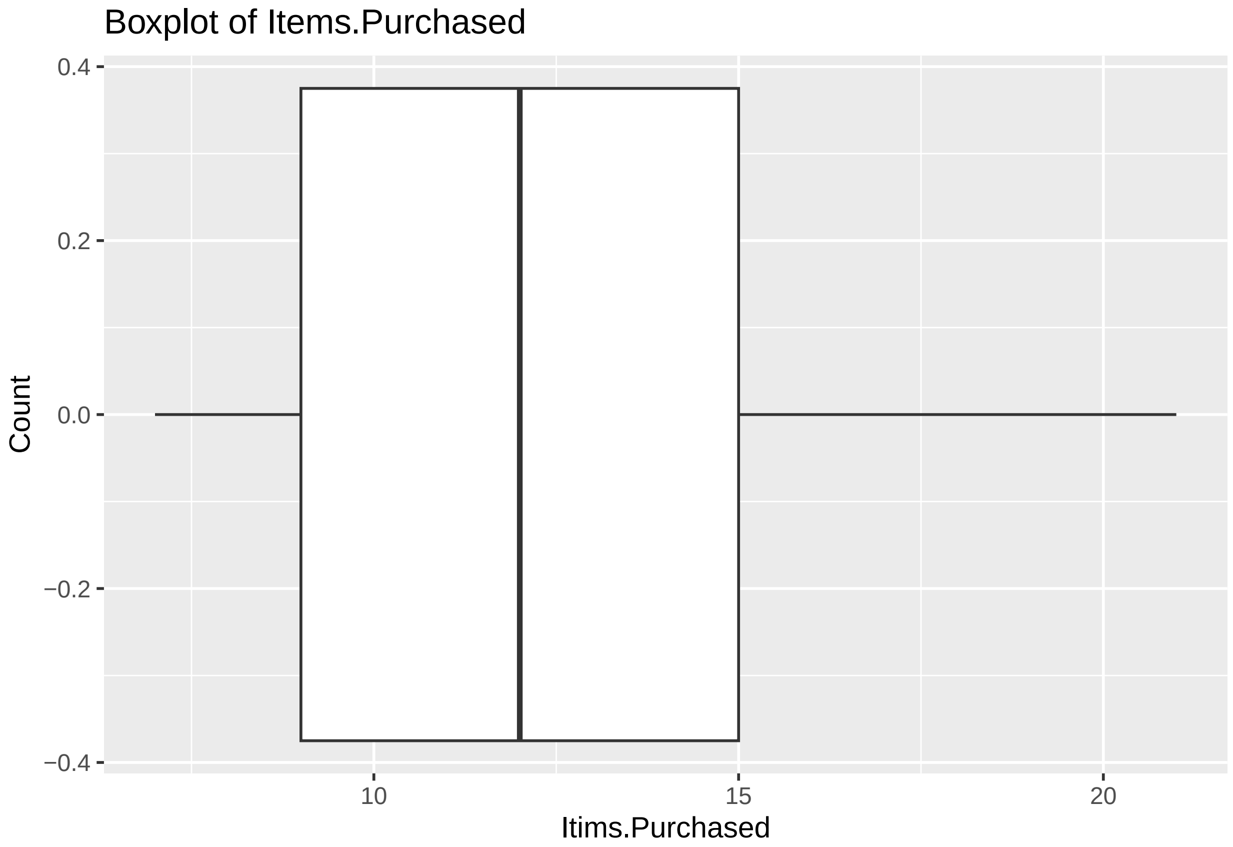


There is a wide dispersion of expenditure between customers. The spend vary between $410,8 and $1 520, AS there is a diffferemce between the mean and the median it means that there are some customers who have the big spend. Also 25% des clients dépensent plus de 1 160,6.

```
summary(df$Items.Purchased)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     7.0     9.0    12.0    12.6    15.0    21.0
```

```
ggplot(df, aes(x=Items.Purchased)) + geom_boxplot()+ labs(title = "Boxplot of Items.Purchased", x="Itim
```



```
summary(df)
```

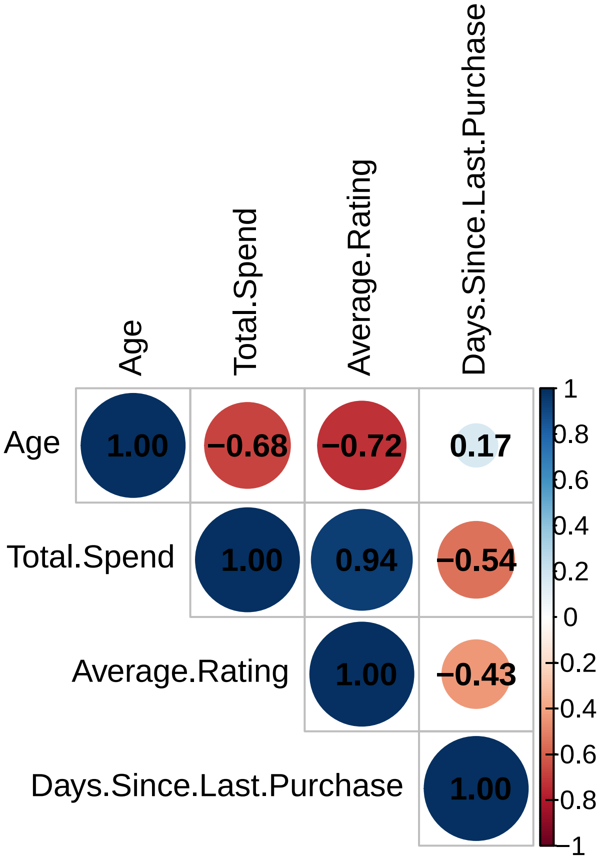
```
##   Customer.ID       Gender         Age                    City    Membership.Type
##  Min.   :101.0   Female:175   Min.   :26.0   Chicago      :58   Bronze:116     
##  1st Qu.:188.2   Male  :175   1st Qu.:30.0   Houston      :58   Gold  :117     
##  Median :275.5                Median :32.5   Los Angeles  :59   Silver:117     
##  Mean   :275.5                Mean   :33.6   Miami        :58                  
##  3rd Qu.:362.8                3rd Qu.:37.0   New York     :59                  
##  Max.   :450.0                Max.   :43.0   San Francisco:58                  
##   Total.Spend     Items.Purchased Average.Rating  Discount.Applied
##  Min.   : 410.8   Min.   : 7.0    Min.   :3.000   Mode :logical   
##  1st Qu.: 502.0   1st Qu.: 9.0    1st Qu.:3.500   FALSE:175       
##  Median : 775.2   Median :12.0    Median :4.100   TRUE :175       
##  Mean   : 845.4   Mean   :12.6    Mean   :4.019                   
##  3rd Qu.:1160.6   3rd Qu.:15.0    3rd Qu.:4.500                   
##  Max.   :1520.1   Max.   :21.0    Max.   :4.900                   
##  Days.Since.Last.Purchase   Satisfaction.Level
##  Min.   : 9.00                       :  2     
##  1st Qu.:15.00             Neutral    :107     
##  Median :23.00             Satisfied  :125     
##  Mean   :26.59             Unsatisfied:116     
##  3rd Qu.:38.00                                 
##  Max.   :63.00
```

### Bivariate

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

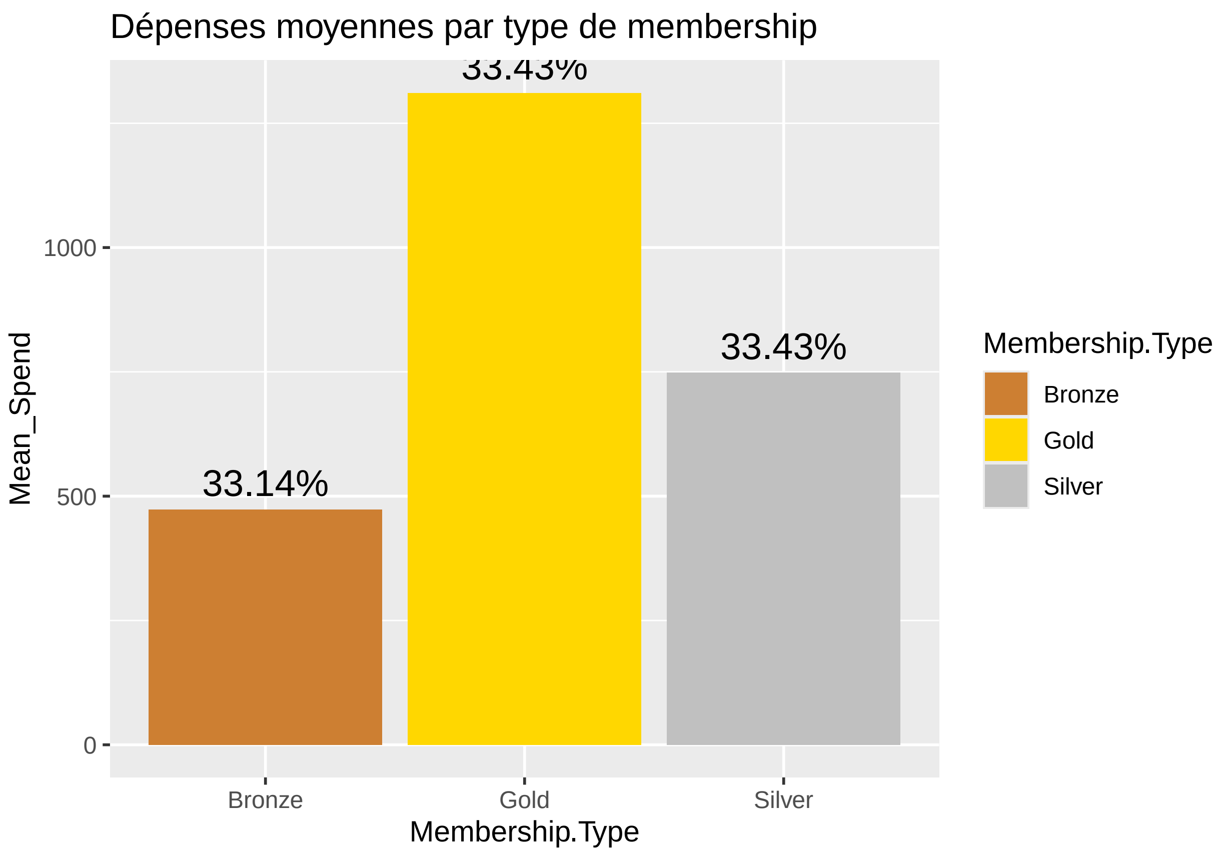
```
    corr_matrix<-cor(df[, c(3,6,8,10)])
    corrplot(corr_matrix,method = "circle",type = "upper", tl.col = "black",  addCoef.col = "black")
```



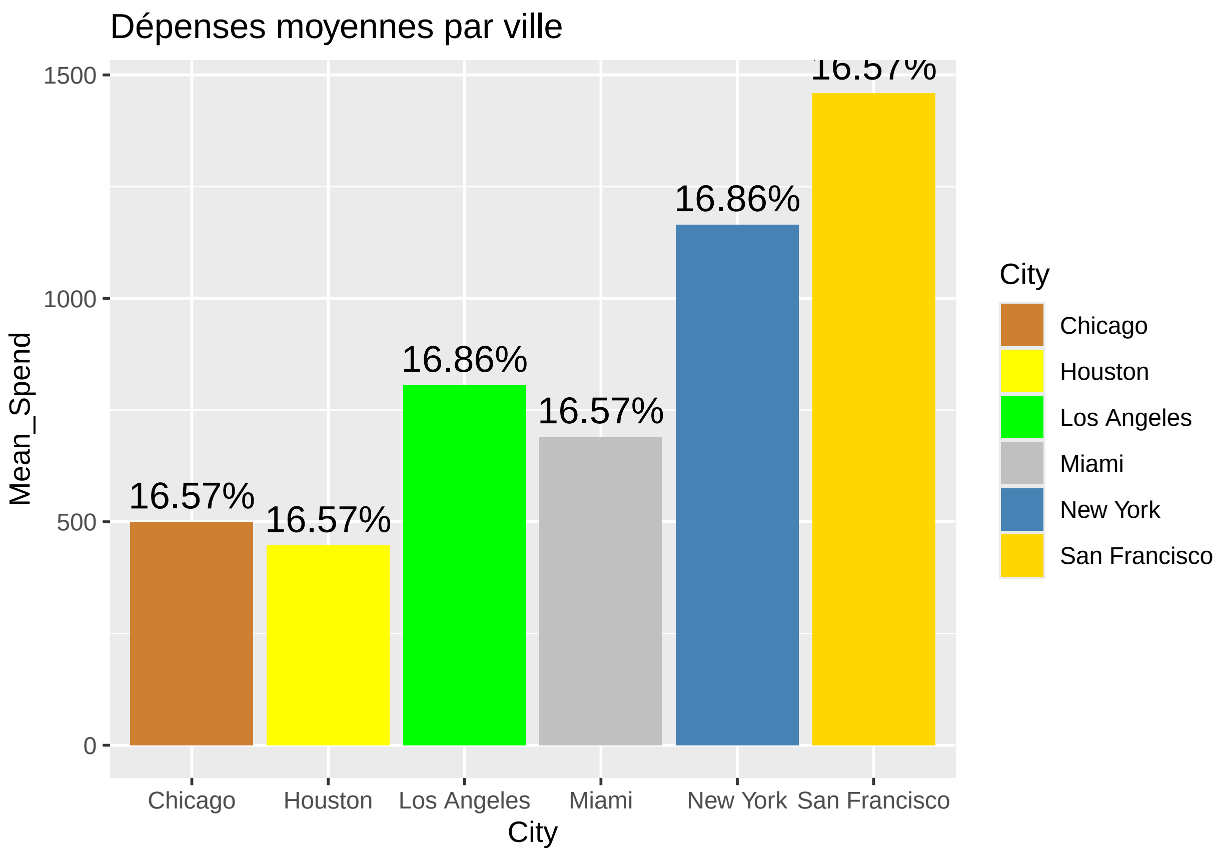
We note that there is a strong positive dependency (0.94) between Average.Rating and Total.Spend. This means that one increases with the growth of the other. Average.Rating also decreases with increasing age (cor = -0.72). The same applies to depnese, which decreases with increasing age (-0.68). The greater the Days.since.Last.Purchase., the lower the expenses ( cor=-0.54).

### Analsis of expenditure

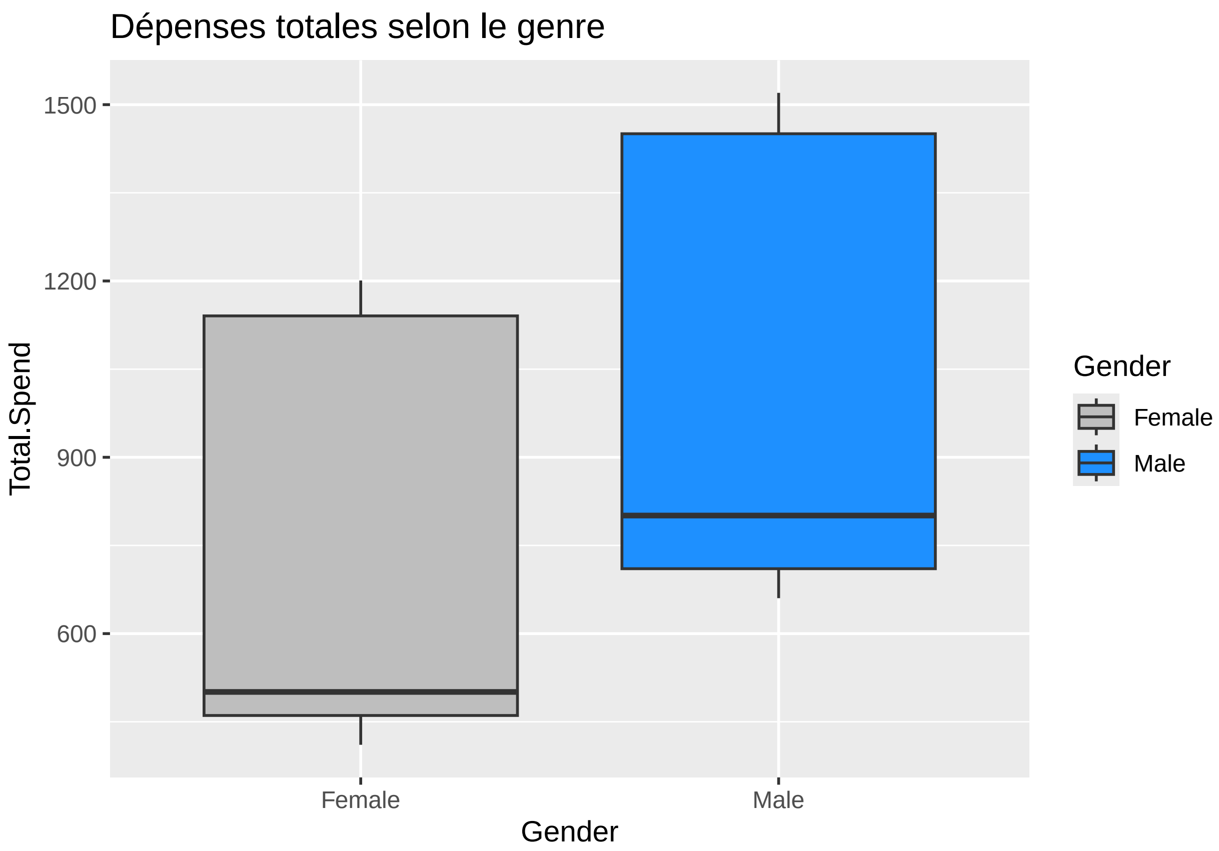
```
# Moyenne des dépenses par type de membership
df %>%
  group_by(Membership.Type) %>%
  summarise(Mean_Spend = mean(Total.Spend), Count = n()) %>%
  mutate(Percentage = percent(Count / sum(Count))) %>%
  ggplot(aes(x=Membership.Type, y=Mean_Spend, fill=Membership.Type)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=Percentage), vjust=-0.5, color="black", size=5) +
  scale_fill_manual(values=c("Bronze"="#CD7F32", "Gold"="#FFD700", "Silver"="#C0C0C0")) +
  ggtitle("Dépenses moyennes par type de membership")
```



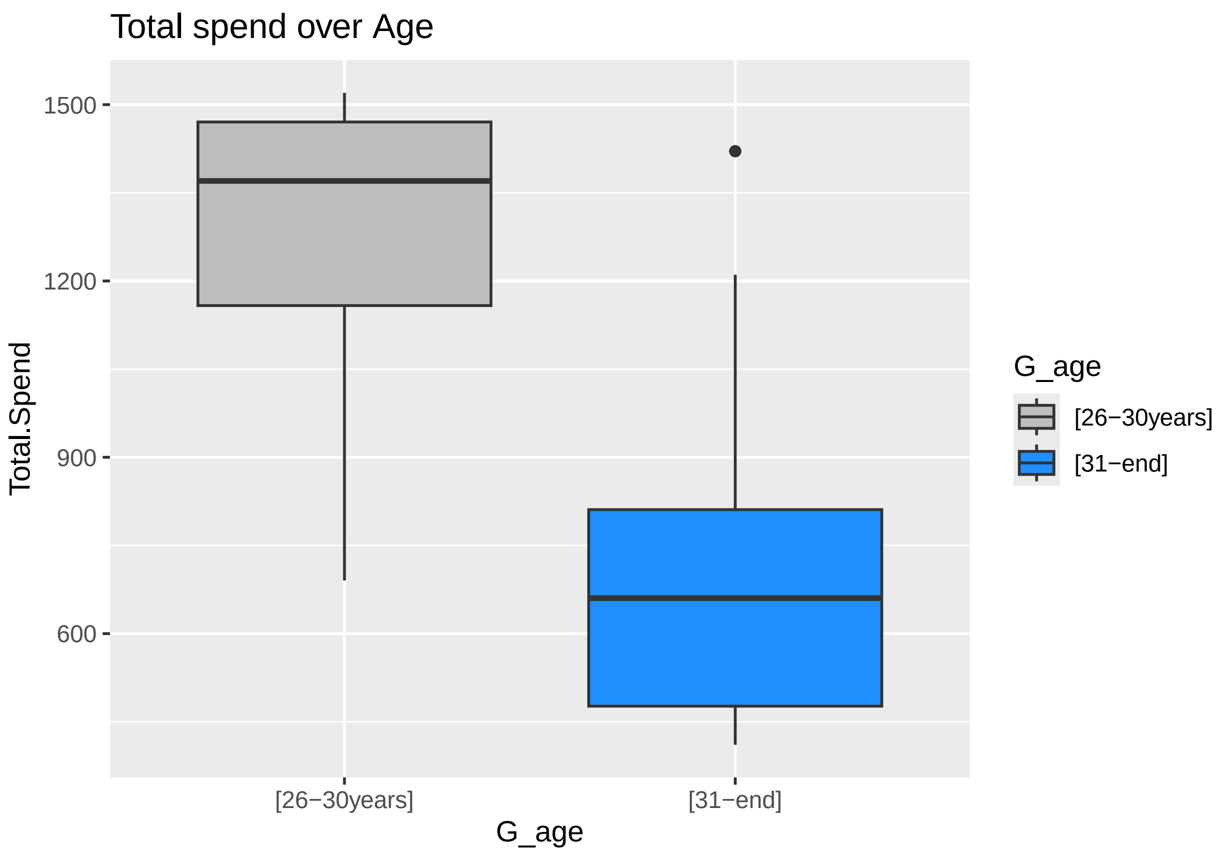
```r
# Depense moyenne par ville
df %>%
  group_by(City) %>%
  summarise(Mean_Spend = mean(Total.Spend), Count = n()) %>%
  mutate(Percentage = percent(Count / sum(Count))) %>%
  ggplot(aes(x=City, y=Mean_Spend, fill=City)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=Percentage), vjust=-0.5, color="black", size=5) +
  scale_fill_manual(values=c("Chicago"="#CD7F32", "Houston"= "yellow","Los Angeles"= "green", "Miami"="#
  ggtitle("Dépenses moyennes par ville")
```



```r
# Boxplot des dépenses en fonction du genre
ggplot(df, aes(x=Gender, y=Total.Spend, fill=Gender)) +
  geom_boxplot() +
  scale_fill_manual(values=c("Female"="grey", "Male"="#1E90FF")) +
  ggtitle("Dépenses totales selon le genre")
```



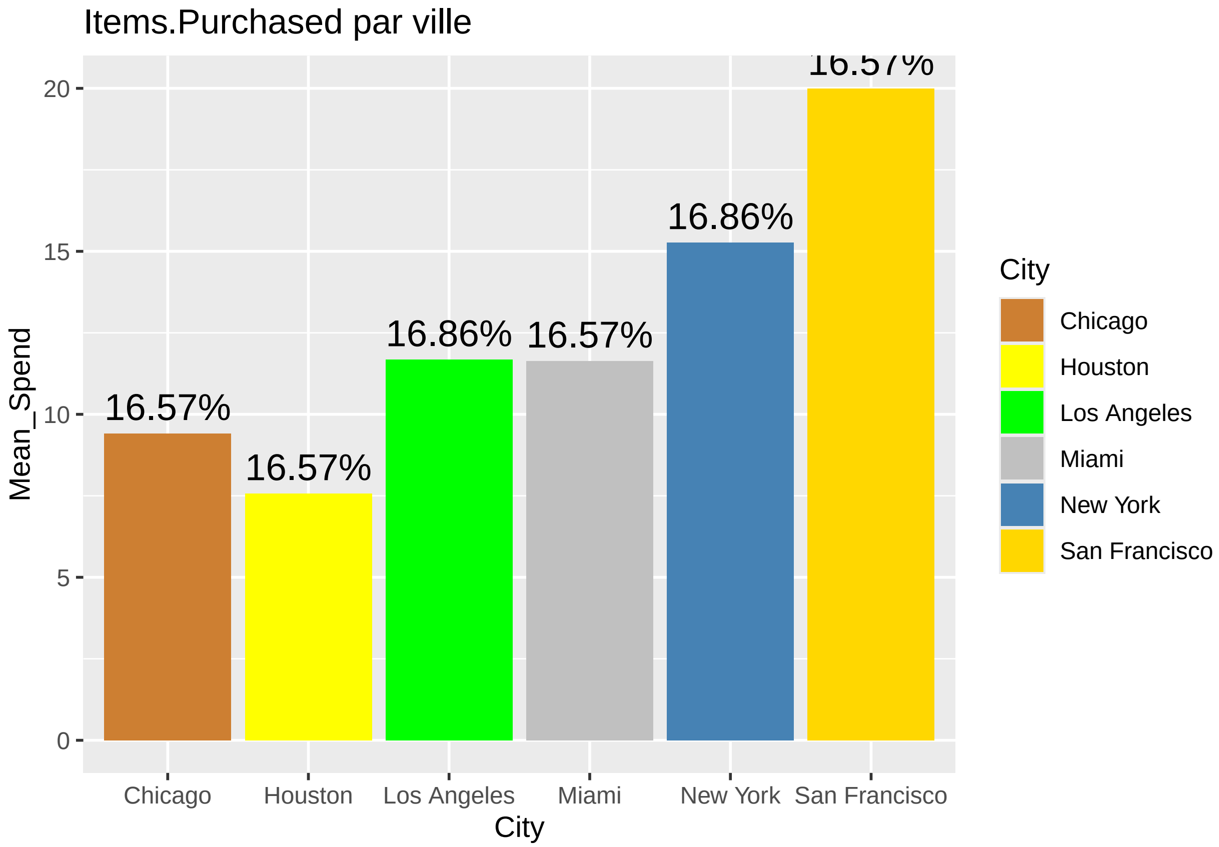
```r
#spending by age
df$G_age = ifelse(df$Age<= 30 & df$Age> 26,"[26-30years]", "[31-end]")
ggplot(df, aes(x=G_age, y=Total.Spend, fill=G_age)) +
  geom_boxplot() +
  scale_fill_manual(values=c("[26-30years]"="grey", "[31-end]"="#1E90FF")) +
  ggtitle("Total spend over Age")
```



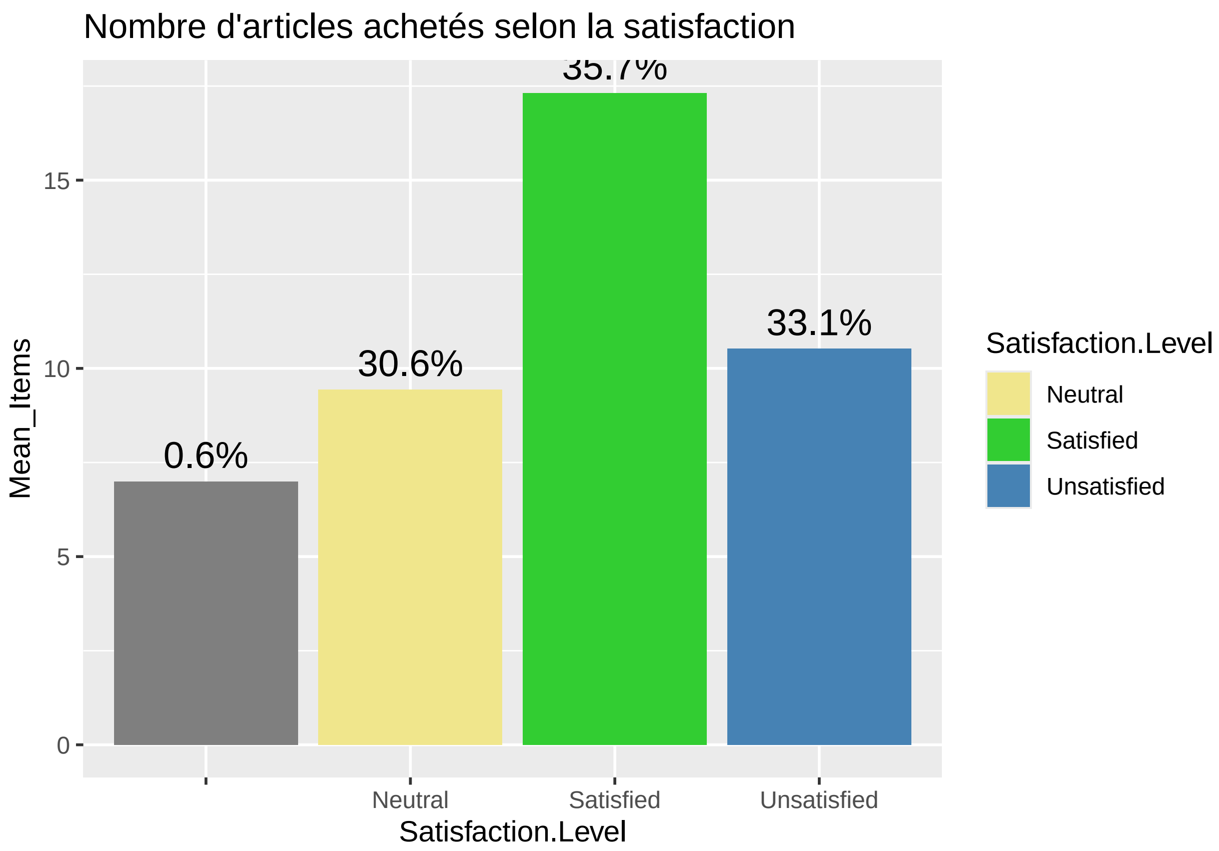
**Interretation** - In average Gold (more than $1250 ) menbership spend more than silver (almost $750) and Bronze (less than $500). The Gold is our potential clients that spend a lot in our shop.

- Men tend to spend more than women.
- San Francisco ($1459,772) is the biggest spender, followed by New York (average $1165,036) and Houston (less than $500).
- The average spending of those aged between 26 and 30 is very different from that of those aged between 31 and 43.

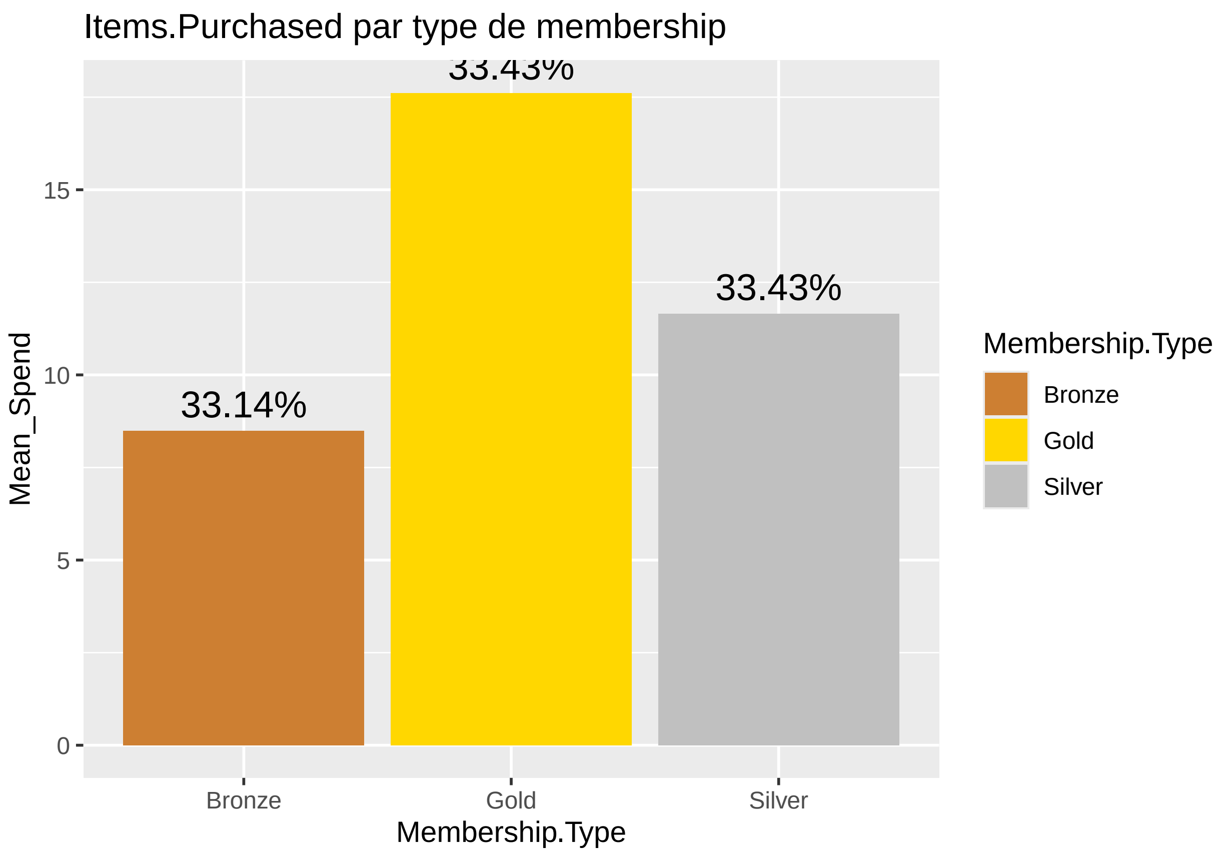
```
# Relation entre la ville et le nombre d'articles achetés
df %>%
  group_by(City) %>%
  summarise(Mean_Spend = mean(Items.Purchased), Count = n()) %>%
  mutate(Percentage = percent(Count / sum(Count))) %>%
  ggplot(aes(x=City, y=Mean_Spend, fill=City)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=Percentage), vjust=-0.5, color="black", size=5) +
  scale_fill_manual(values=c("Chicago"="#CD7F32", "Houston"= "yellow","Los Angeles"= "green", "Miami"="
  ggtitle("Items.Purchased par ville")
```



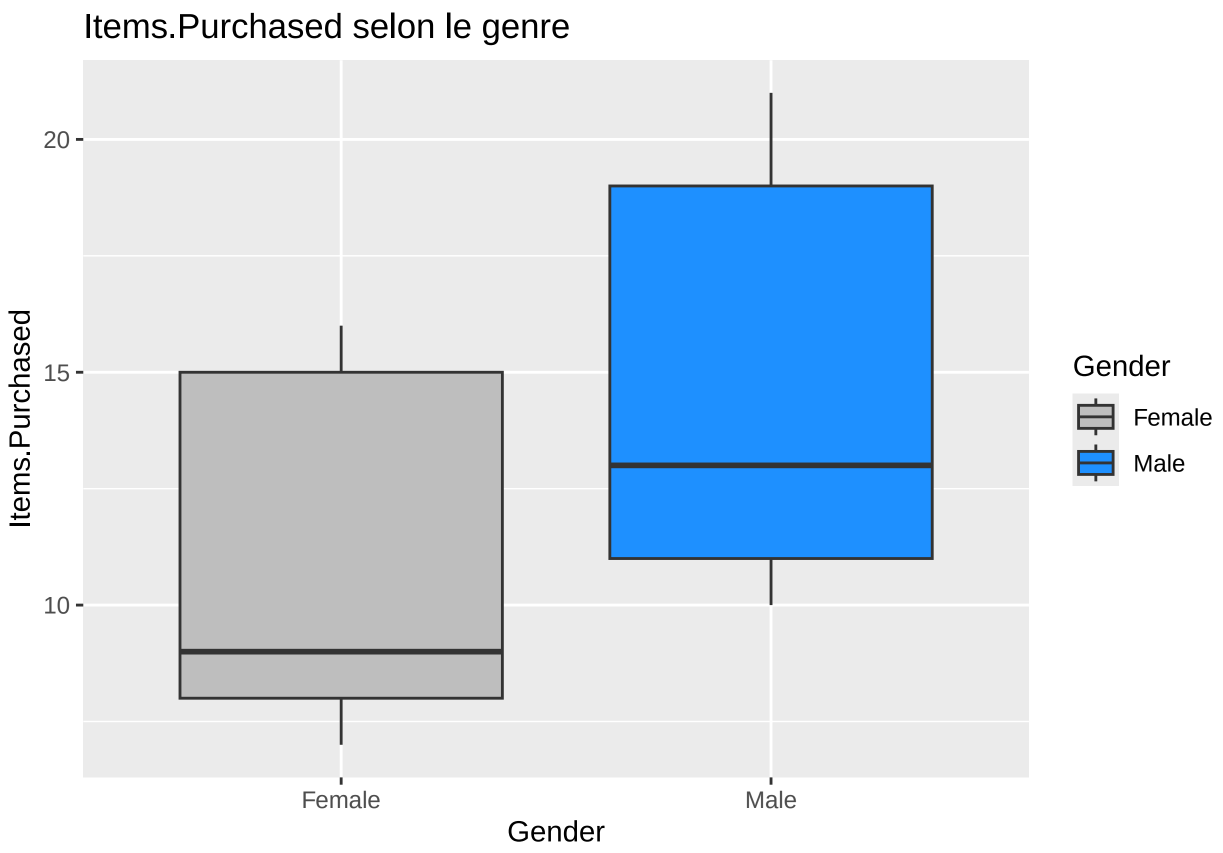
```
# Relation entre la satisfaction et le nombre d'articles achetés
df %>%
  group_by(Satisfaction.Level) %>%
  summarise(Mean_Items = mean(Items.Purchased), Count = n()) %>%
  mutate(Percentage = percent(Count / sum(Count))) %>%
  ggplot(aes(x=Satisfaction.Level, y=Mean_Items, fill=Satisfaction.Level)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=Percentage), vjust=-0.5, color="black", size=5) +
  scale_fill_manual(values=c("Neutral"="#F0E68C", "Satisfied"="#32CD32", "Unsatisfied"="steelblue")) +
  ggtitle("Nombre d'articles achetés selon la satisfaction")
```



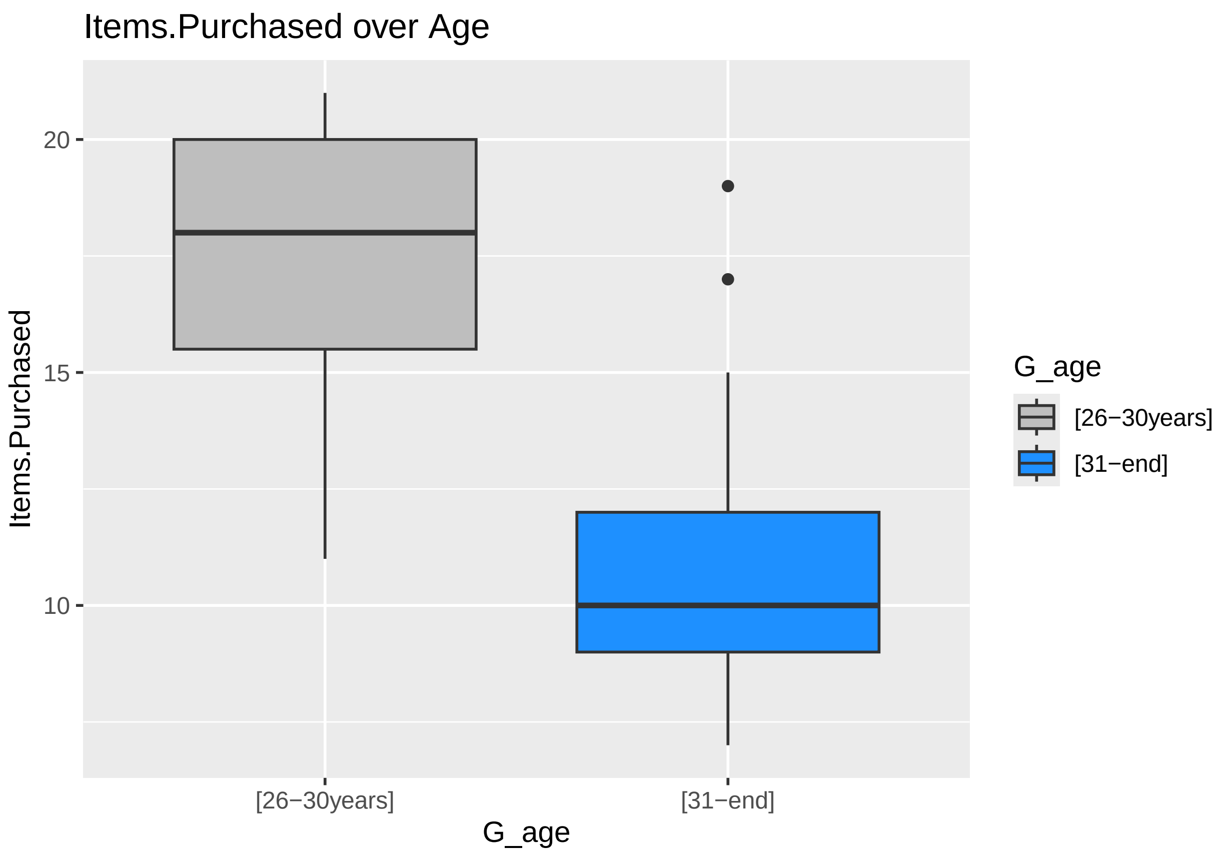
```r
# Items.Purchased par type de membership
df %>%
  group_by(Membership.Type) %>%
  summarise(Mean_Spend = mean(Items.Purchased), Count = n()) %>%
  mutate(Percentage = percent(Count / sum(Count))) %>%
  ggplot(aes(x=Membership.Type, y=Mean_Spend, fill=Membership.Type)) +
  geom_bar(stat="identity") +
  geom_text(aes(label=Percentage), vjust=-0.5, color="black", size=5) +
  scale_fill_manual(values=c("Bronze"="#CD7F32", "Gold"="#FFD700", "Silver"="#C0C0C0")) +
  ggtitle("Items.Purchased par type de membership")
```



```r
# Boxplot Items.Purchased en fonction du genre
ggplot(df, aes(x=Gender, y=Items.Purchased, fill=Gender)) +
  geom_boxplot() +
  scale_fill_manual(values=c("Female"="grey", "Male"="#1E90FF")) +
  ggtitle("Items.Purchased selon le genre")
```



```
#Items.Purchased by age
#df$G_age = ifelse(df$Age<= 30 & df$Age> 26,"[26-30years]", "[31-end]")
ggplot(df, aes(x=G_age, y=Items.Purchased, fill=G_age)) +
  geom_boxplot() +
  scale_fill_manual(values=c("[26-30years]"="grey", "[31-end]"="#1E90FF")) +
  ggtitle("Items.Purchased over Age")
```



- Satisfied customers are those who buy the most products on average. There are also some people who buy on average more than 10 item but they aren't satified.

## Forcasting

```
df2 = df[, -c(1,8,9,11,5,12)]
model = lm(Total.Spend ~ ., data=df2)
summary(model)
```

```
##
## Call:
## lm(formula = Total.Spend ~ ., data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.916  -8.529  -0.316   8.800  45.382
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                239.7018    37.3449   6.419 4.62e-10 ***
## GenderMale                   3.5790    10.4668   0.342   0.7326
## Age                          1.7448     0.7004   2.491   0.0132 *
## CityHouston                -19.5708     6.8438  -2.860   0.0045 **
## CityLos Angeles            226.3216    13.0664  17.321  < 2e-16 ***
## CityMiami                  159.8701    13.1617  12.147  < 2e-16 ***
## CityNew York               516.1913     9.8119  52.608  < 2e-16 ***
## CitySan Francisco        673.2759    17.5519  38.359  < 2e-16 ***
## Items.Purchased           25.3327     1.0814  23.427  < 2e-16 ***
## Days.Since.Last.Purchase  -1.2754     0.1811  -7.042 1.05e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.61 on 340 degrees of freedom
## Multiple R-squared:  0.9984, Adjusted R-squared:  0.9984
## F-statistic: 2.377e+04 on 9 and 340 DF,  p-value: < 2.2e-16
```

```r
df2 = df2[, -1]
model2 = lm(Total.Spend ~ ., data=df2)
summary(model2)
```

```
##
## Call:
## lm(formula = Total.Spend ~ ., data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.925  -8.624  -0.221   8.809  45.390
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              240.3319    37.2511   6.452 3.79e-10 ***
## Age                        1.7359     0.6990   2.483  0.01350 *
## CityHouston              -19.7215     6.8207  -2.891  0.00408 **
## CityLos Angeles          229.7003     8.5385  26.902  < 2e-16 ***
## CityMiami                163.3058     8.4902  19.235  < 2e-16 ***
## CityNew York             516.0848     9.7943  52.692  < 2e-16 ***
## CitySan Francisco        676.6168    14.5620  46.465  < 2e-16 ***
## Items.Purchased           25.3292     1.0799  23.455  < 2e-16 ***
## Days.Since.Last.Purchase  -1.2808     0.1802  -7.109 6.87e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.59 on 341 degrees of freedom
## Multiple R-squared:  0.9984, Adjusted R-squared:  0.9984
## F-statistic: 2.681e+04 on 8 and 341 DF,  p-value: < 2.2e-16
```

**Interpretation**: - For each additional year of age, total expenditure increases by an average of 1.74 units (e.g. dollars), all else being equal. - Customers in San Francisco and New York spend significantly more than those in other cities. - For each additional item purchased, total expenditure increases by an average of 25.33 units. - For each additional day since the last purchase, total expenditure decreases by an average of 1.28 units.

### Implications for the company

*Priority targets:* Customers in San Francisco and New York spend significantly more. The company could focus its marketing efforts on these regions. Older customers and those who buy more items are also priority targets.

*Loyalty:* Customers who return more frequently (fewer days since last purchase) spend more. Loyalty programs could encourage more frequent purchases.

_Gender:___ Gender has no significant impact on spending. Marketing strategies should therefore not be differentiated according to gender.

```r
anova(model, model2)
```

```
## Analysis of Variance Table
##
## Model 1: Total.Spend ~ Gender + Age + City + Items.Purchased + Days.Since.Last.Purchase
## Model 2: Total.Spend ~ Age + City + Items.Purchased + Days.Since.Last.Purchase
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    340 72591
## 2    341 72616 -1   -24.963 0.1169 0.7326
```