Mall_data

Vijaya Suresh

2023-02-20

```
#Import all the required libraries
install.packages("readr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

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

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(readr)
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
```

```
#import the data
data <- read_csv("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.
```

```
print(data)
```

```
## # A tibble: 200 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
##   7 0007       Female    35                   18                        6
##   8 0008       Female    23                   18                       94
##   9 0009       Male      64                   19                        3
##  10 0010       Female    30                   19                       72
## # ... with 190 more rows
```

```
#BASIC_INSIGHTS
glimpse(data)                      #DATA_TYPE
```

```
## Rows: 200
## Columns: 5
## $ CustomerID               <chr> "0001", "0002", "0003", "0004", "0005", "0006~
## $ Genre                    <chr> "Male", "Male", "Female", "Female", "Female",~
## $ Age                      <dbl> 19, 21, 20, 23, 31, 22, 35, 23, 64, 30, 67, 3~
## $ `Annual Income (k$)`     <dbl> 15, 15, 16, 16, 17, 17, 18, 18, 19, 19, 19, 1~
## $ `Spending Score (1-100)` <dbl> 39, 81, 6, 77, 40, 76, 6, 94, 3, 72, 14, 99, ~
```

```
summary(data)                      #STATISTICAL_SUMMARY
```

```
##   CustomerID           Genre                Age        Annual Income (k$)
##  Length:200         Length:200         Min.   :18.00   Min.   : 15.00  
##  Class  :character  Class  :character  1st Qu.:28.75   1st Qu.: 41.50  
##  Mode   :character  Mode   :character  Median :36.00   Median : 61.50  
##                                        Mean   :38.85   Mean   : 60.56  
##                                        3rd Qu.:49.00   3rd Qu.: 78.00  
##                                        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         
```

```
#MISSING_VALUE
sum(is.na(data))                  #SUM_OF_MISSING_VALUE
```

```
## [1] 0
```

```
#DISTINCT_DATA
distinct(data)                     #DISTINCT_DATA
```

```
## # A tibble: 200 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
##  7 0007       Female    35                   18                        6
##  8 0008       Female    23                   18                       94
##  9 0009       Male      64                   19                        3
## 10 0010       Female    30                   19                       72
## # ... with 190 more rows
```

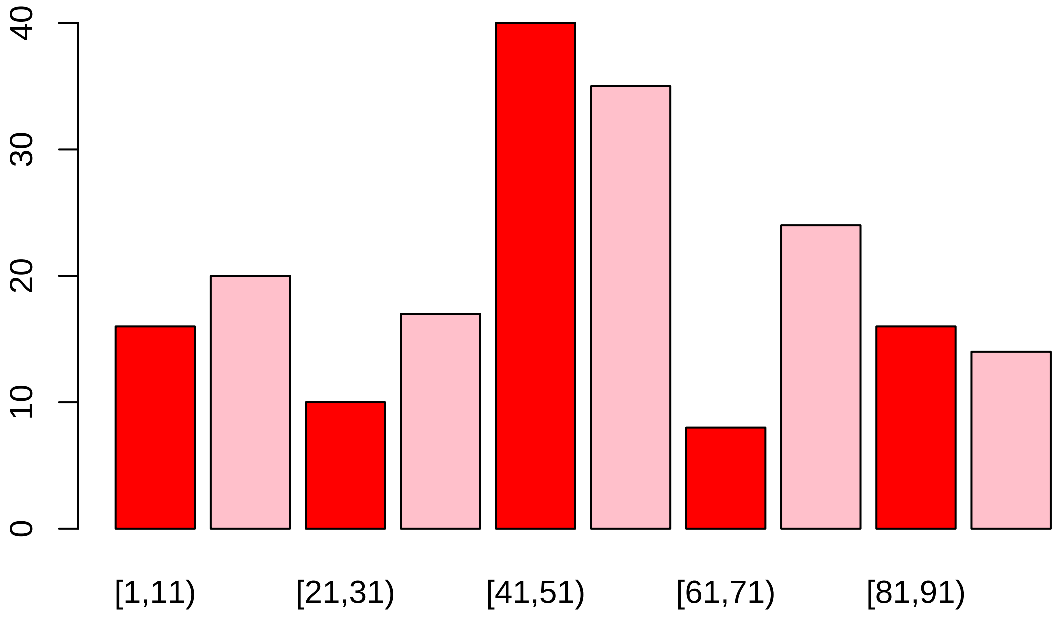
```
#RENAMING_COLUMN
colnames(data)[4]= "Income"
colnames(data)[5]="Score"
head(data)
```

```
## # A tibble: 6 x 5
##   CustomerID Genre    Age Income Score
##   <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
```

```
#EXPLORATORY DATA ANALYSIS
#BAR_PLOT
score = pull(data,Score)
score_1=cut(score,breaks=seq(1,101,by=10),right=FALSE)
table(score_1)
```

```
## score_1
##    [1,11)  [11,21)  [21,31)  [31,41)  [41,51)  [51,61)  [61,71)  [71,81)
##        16       20       10       17       40       35        8       24
##   [81,91) [91,101)
##        16       14
```

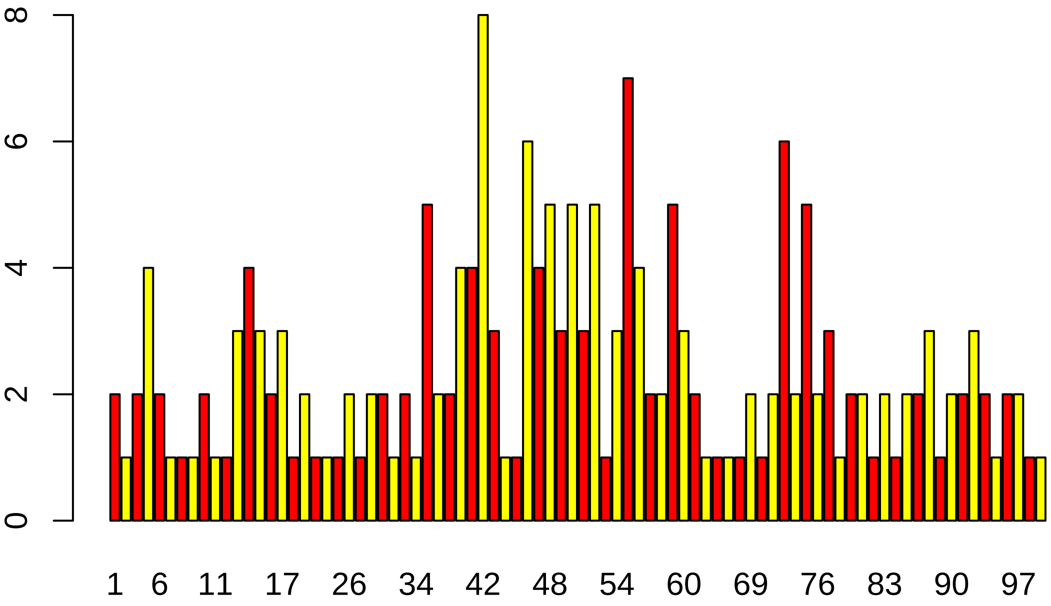
```
barplot(table(score_1),col=c('red','pink'))
```



```
table(score)
```

```
## score
##  1  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 20 22 23 24 26 27 28 29 31
##  2  1  2  4  2  1  1  1  2  1  1  3  4  3  2  3  1  2  1  1  1  2  1  2  2  1
## 32 34 35 36 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
##  2  1  5  2  2  4  4  8  3  1  1  6  4  5  3  5  3  5  1  3  7  4  2  2  5  3
## 61 63 65 66 68 69 71 72 73 74 75 76 77 78 79 81 82 83 85 86 87 88 89 90 91 92
##  2  1  1  1  1  2  1  2  6  2  5  2  3  1  2  2  1  2  1  2  2  3  1  2  2  3
## 93 94 95 97 98 99
##  2  1  2  2  1  1
```

```
barplot(table(score),col=c('red','yellow'))
```



```
#BOXPLOT
boxplot(score)

#SUBSETTING INTERQUARTILE DATA OF SCORE
df =filter(data,Score>=35 & Score<=73)
glimpse(df)
```

```
## Rows: 105
## Columns: 5
## $ CustomerID <chr> "0001", "0005", "0010", "0017", "0018", "0021", "0022", "00~
## $ Genre      <chr> "Male", "Female", "Female", "Female", "Male", "Male", "Male~
## $ Age        <dbl> 19, 31, 30, 35, 20, 35, 25, 31, 35, 21, 30, 65, 48, 31, 24,~
## $ Income     <dbl> 15, 17, 19, 21, 21, 24, 24, 25, 28, 30, 34, 38, 39, 39, 39,~
## $ Score      <dbl> 39, 40, 72, 35, 66, 35, 73, 73, 61, 73, 73, 35, 36, 61, 65,~
```

```
summary(df)
```

```
##   CustomerID           Genre                 Age            Income     
##  Length:105         Length:105         Min.   :18.00   Min.   : 15.0  
##  Class :character   Class :character   1st Qu.:27.00   1st Qu.: 44.0  
##  Mode  :character   Mode  :character   Median :38.00   Median : 54.0  
##                                        Mean   :40.72   Mean   : 54.4  
##                                        3rd Qu.:51.00   3rd Qu.: 63.0  
##                                        Max.   :70.00   Max.   :103.0  
##      Score      
##  Min.   :35.00  
##  1st Qu.:43.00  
##  Median :50.00  
##  Mean   :51.63  
##  3rd Qu.:58.00
##  Max.   :73.00
```

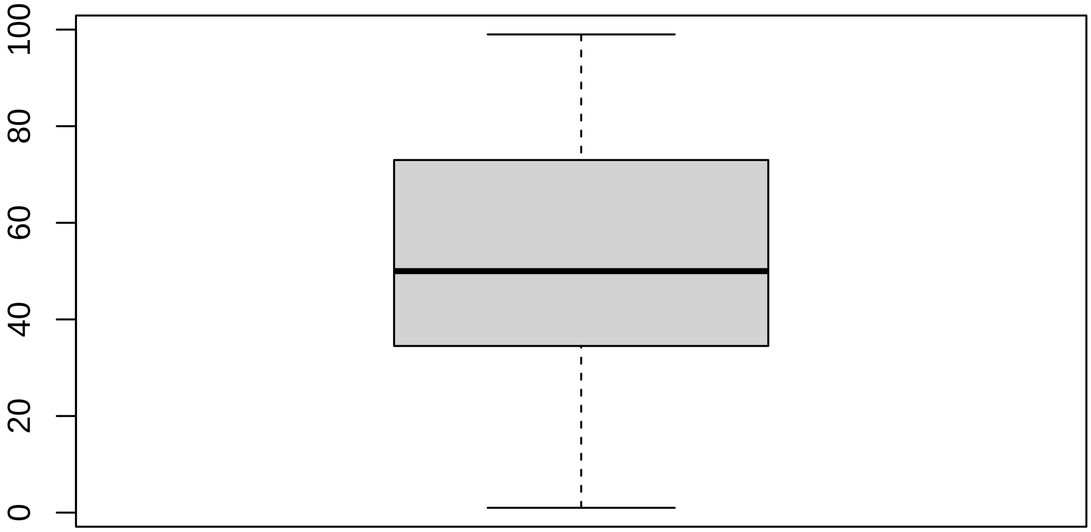
```
print(df)
```

```
## # A tibble: 105 x 5
##    CustomerID Genre    Age Income Score
##    <chr>      <chr>  <dbl>  <dbl> <dbl>
##  1 0001       Male      19     15    39
##  2 0005       Female    31     17    40
##  3 0010       Female    30     19    72
##  4 0017       Female    35     21    35
##  5 0018       Male      20     21    66
##  6 0021       Male      35     24    35
##  7 0022       Male      25     24    73
##  8 0024       Male      31     25    73
##  9 0028       Male      35     28    61
## 10 0032       Female    21     30    73
## # ... with 95 more rows
```

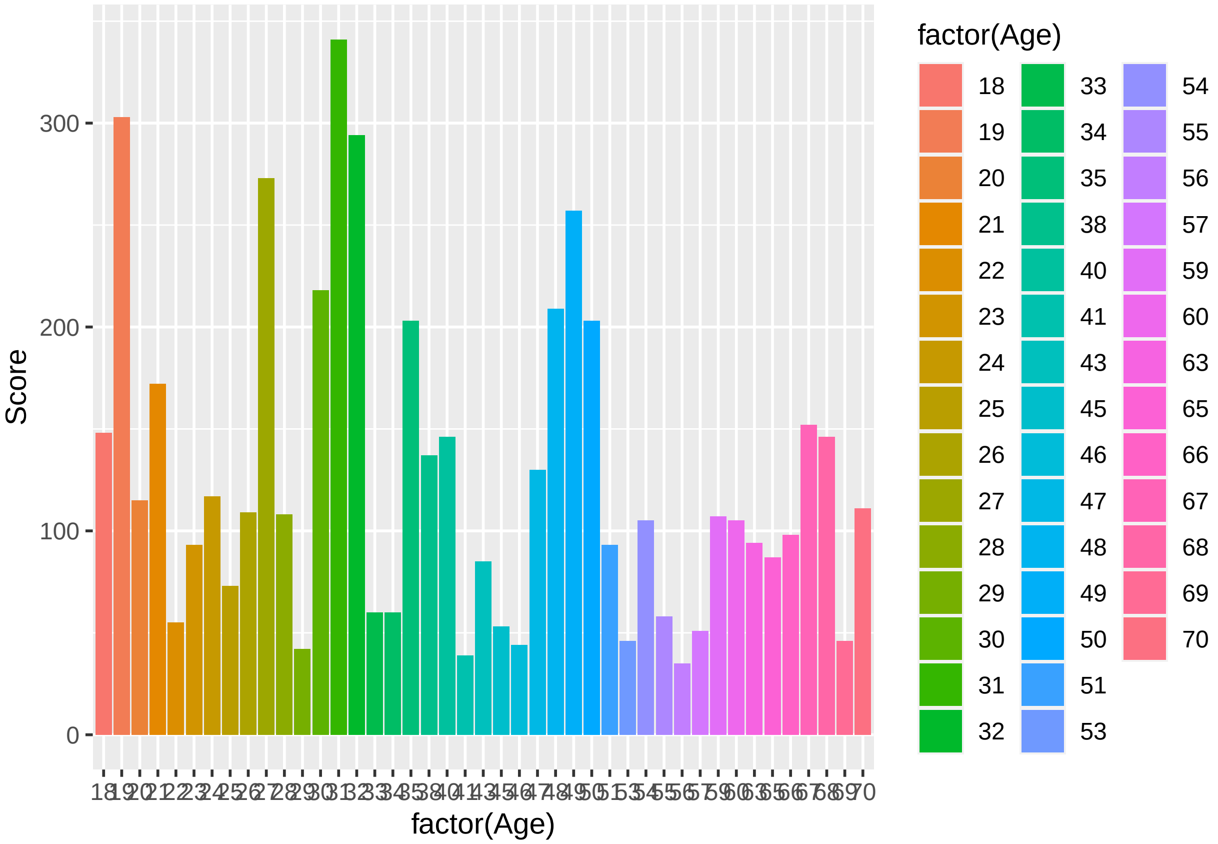
```
#BAR CHART OF TWO ATTRIBUTES
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

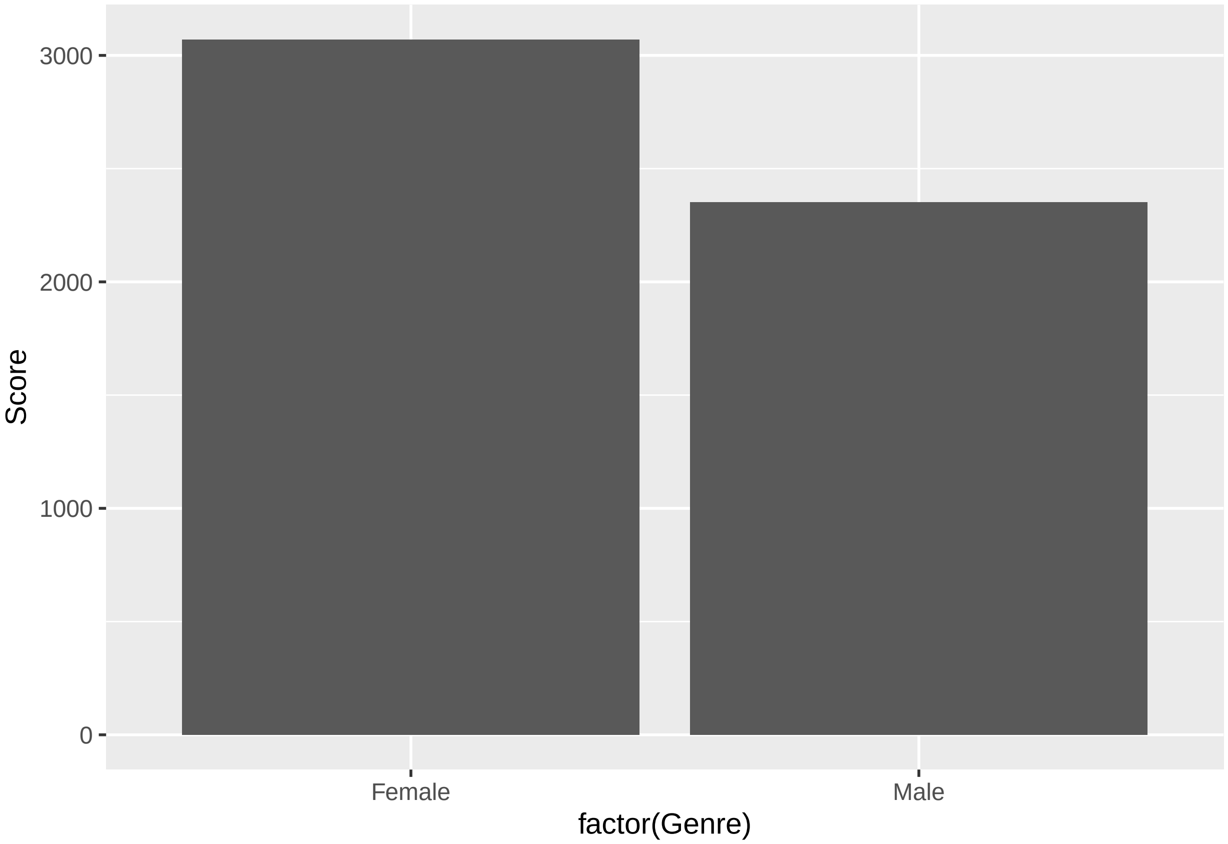
```
library(ggplot2)
```



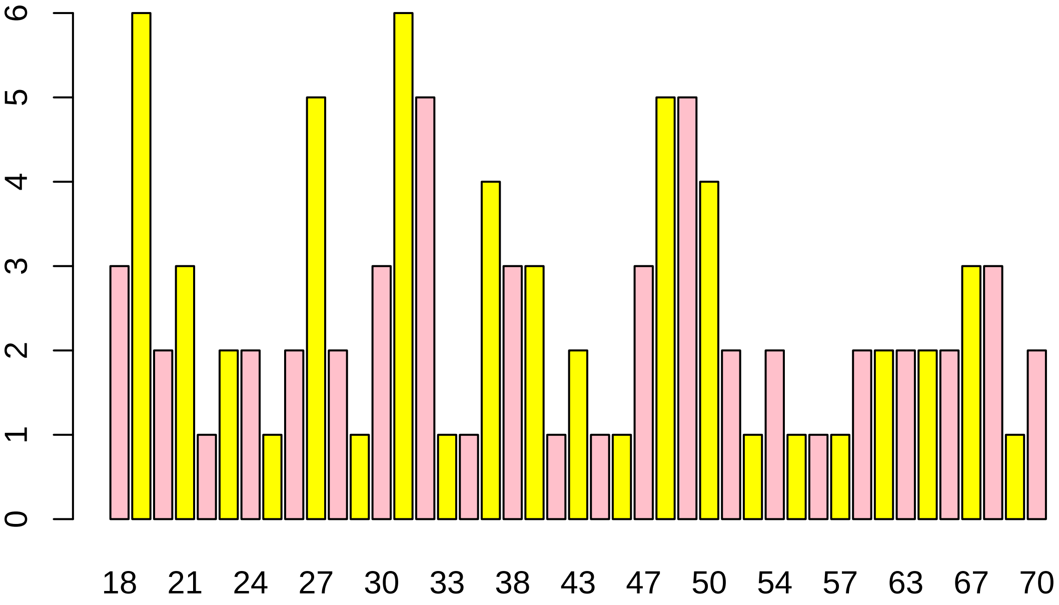
```
ggplot(df, aes(x =factor(Age), y = Score,fill=factor(Age))) +
  geom_bar(stat = "identity")
```



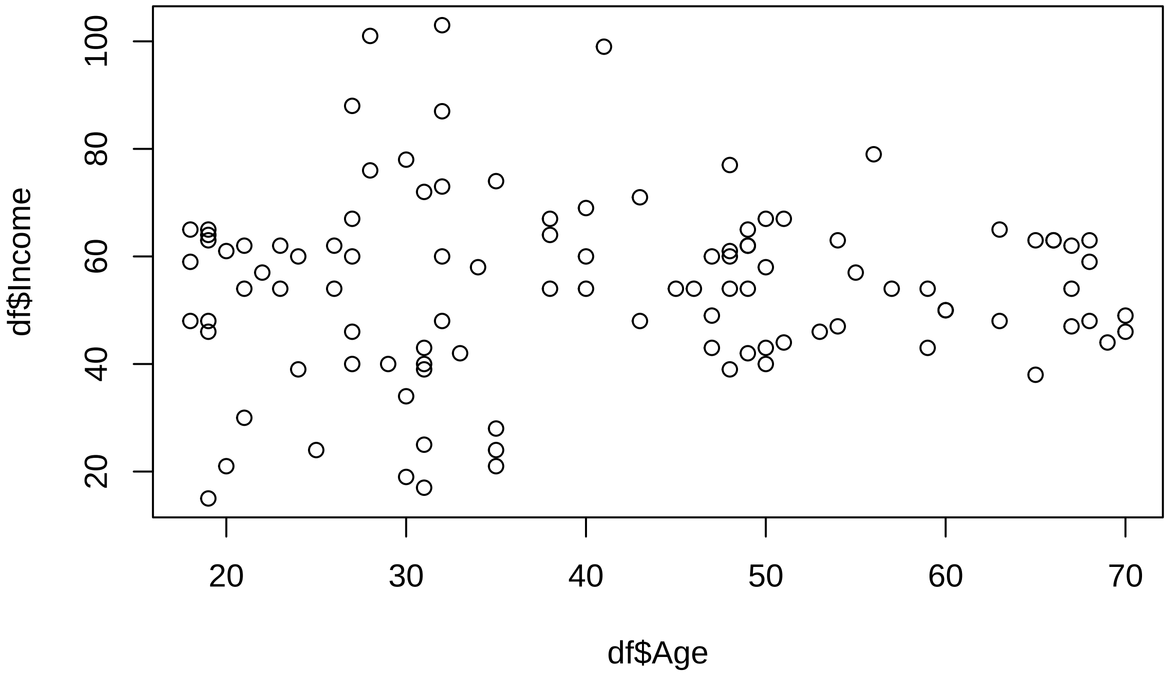
```
ggplot(df, aes(x =factor(Genre), y = Score,factor(Genre))) +
  geom_bar(stat = "identity")
```



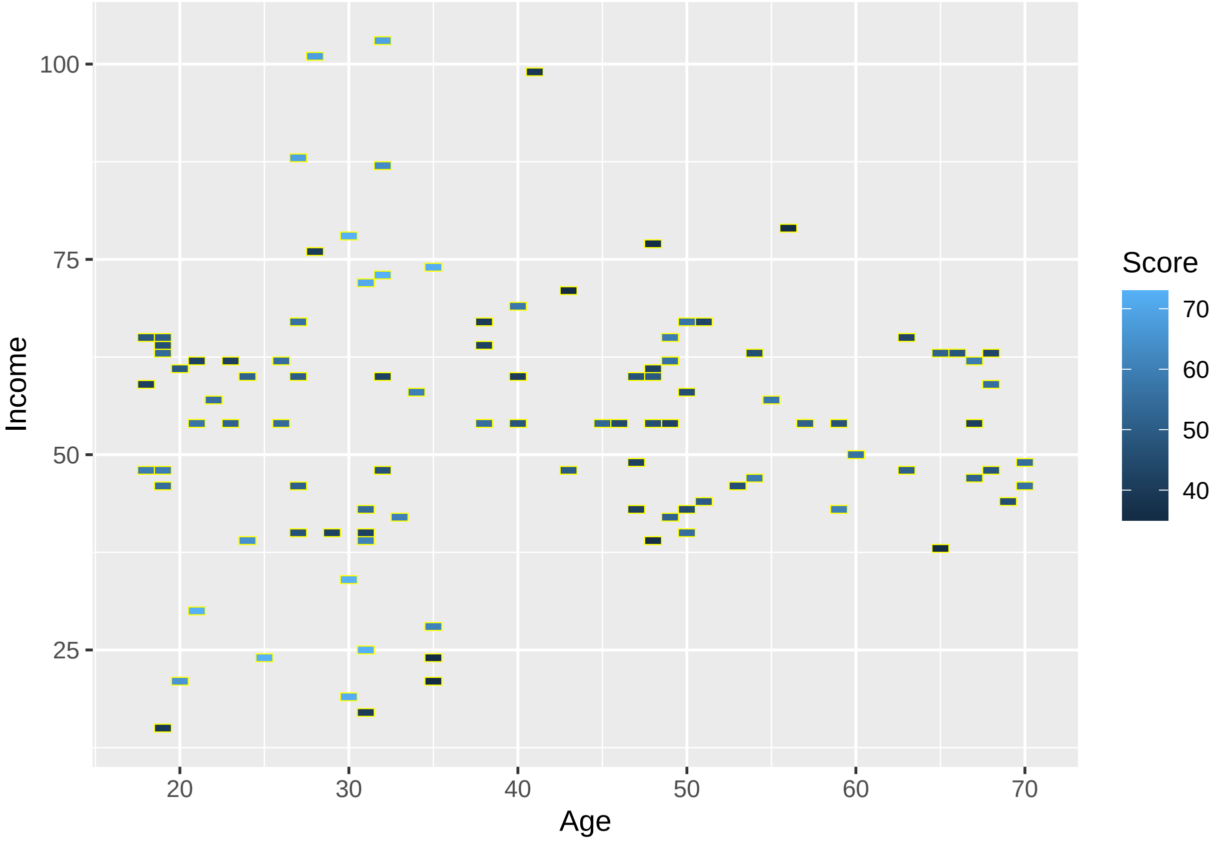
```
#BARCHART
age = pull(df,Age)
barplot(table(age),col=c('pink','yellow'))
```



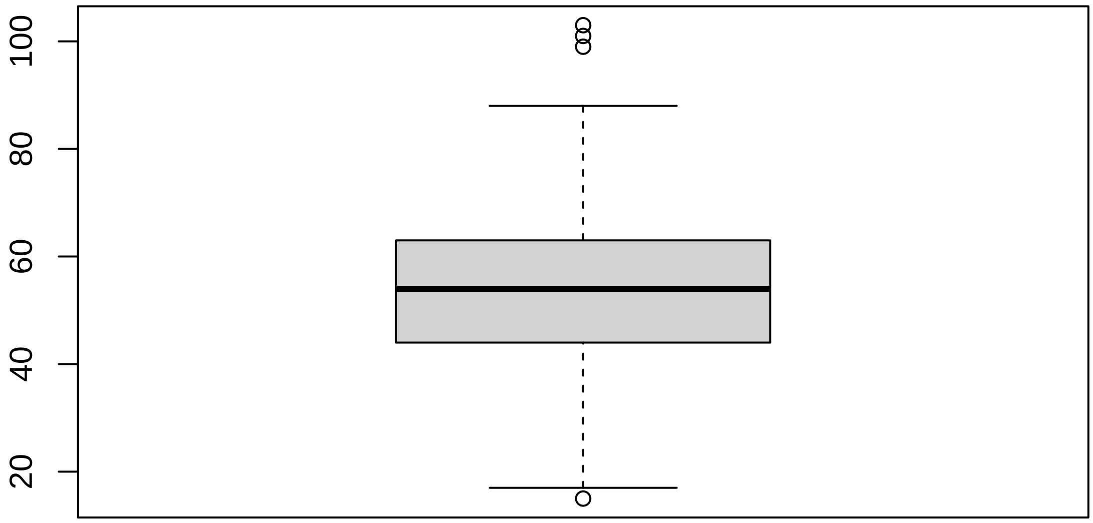
```
#SCATTER_PLOT
plot(x=df$Age,y=df$Income)
```



```
#HEAT_MAP
ggplot(df,aes(x=Age,y=Income,fill =Score))+
  geom_tile(color='yellow',linewidth=0.2)
```



```
#BOXPLOT
boxplot(df$Income)
```



```
#subsetting  female customers
female = filter(df,Genre=='Female')
head(female)
```

```
## # A tibble: 6 x 5
##   CustomerID Genre    Age Income Score
##   <chr>      <chr>  <dbl>  <dbl> <dbl>
## 1 0005       Female    31     17    40
## 2 0010       Female    30     19    72
## 3 0017       Female    35     21    35
## 4 0032       Female    21     30    73
## 5 0038       Female    30     34    73
## 6 0041       Female    65     38    35
```

```
#basic insight
summary(female)
```

```
##   CustomerID           Genre                Age            Income      
##  Length:60          Length:60          Min.   :18.00   Min.   : 17.00  
##  Class :character   Class :character   1st Qu.:29.75   1st Qu.: 43.75  
##  Mode  :character   Mode  :character   Median :38.00   Median : 55.50  
##                                        Mean   :39.98   Mean   : 54.80  
##                                        3rd Qu.:50.00   3rd Qu.: 64.25  
##                                        Max.   :68.00   Max.   :103.00  
##      Score      
##  Min.   :35.00  
##  1st Qu.:42.00  
##  Median :50.00  
##  Mean   :51.17  
##  3rd Qu.:57.00  
##  Max.   :73.00
```

```
#Correlation
install.packages("ggpubr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(ggpubr)

cor(female$Income,female$Score)
```

```
## [1] -0.0008112764
```

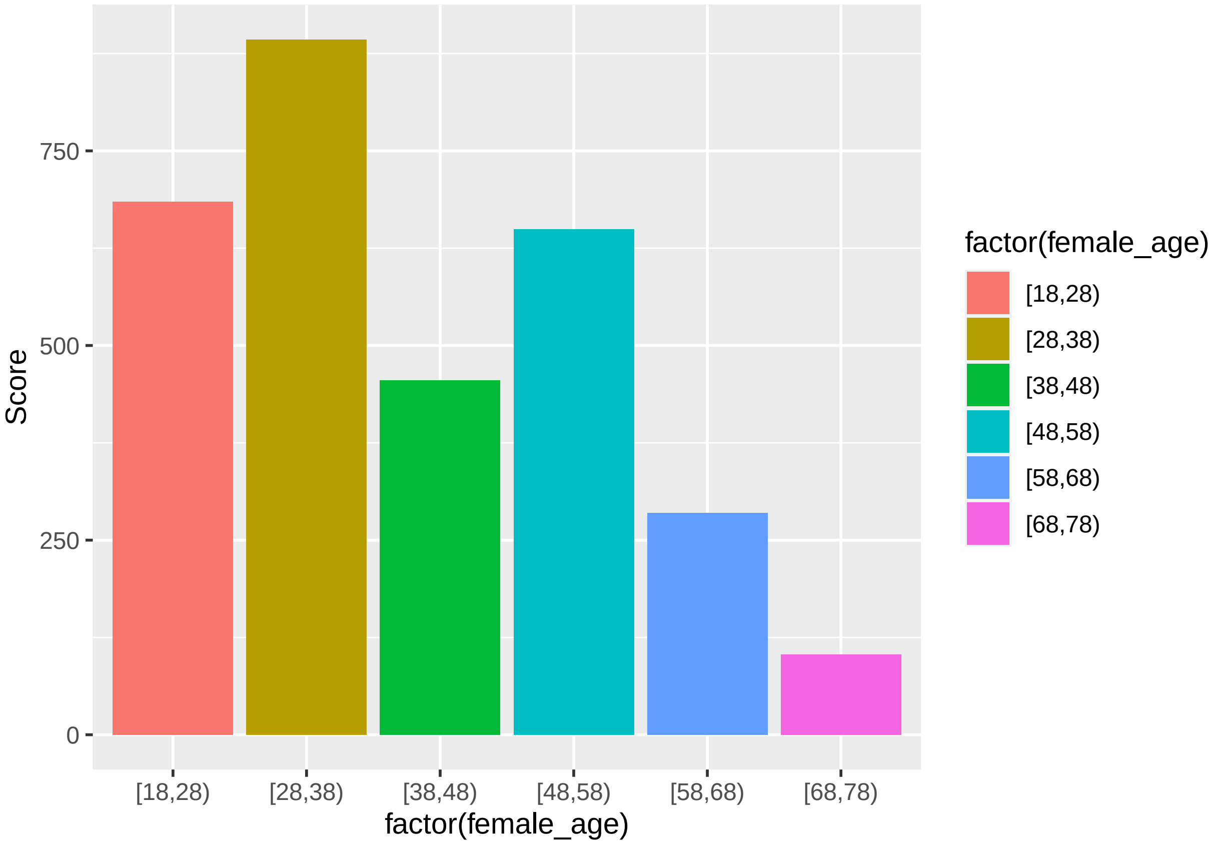
```
cor.test(female$Income,female$Score)
```

```
##
##  Pearson's product-moment correlation
##
## data:  female$Income and female$Score
## t = -0.0061785, df = 58, p-value = 0.9951
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2546835  0.2531656
## sample estimates:
##           cor
## -0.0008112764
```

```
#BAR CHART OF TWO ATTRIBUTES
female_age = pull(female,Age)
female_age=cut(female_age,breaks=seq(18,80,by=10),right=FALSE)
table(female_age)
```

```
## female_age
## [18,28) [28,38) [38,48) [48,58) [58,68) [68,78)
##      13      16      10      13       6       2
```

```
ggplot(female, aes(x =factor(female_age), y = Score,fill=factor(female_age))) +
  geom_bar(stat = "identity")
```



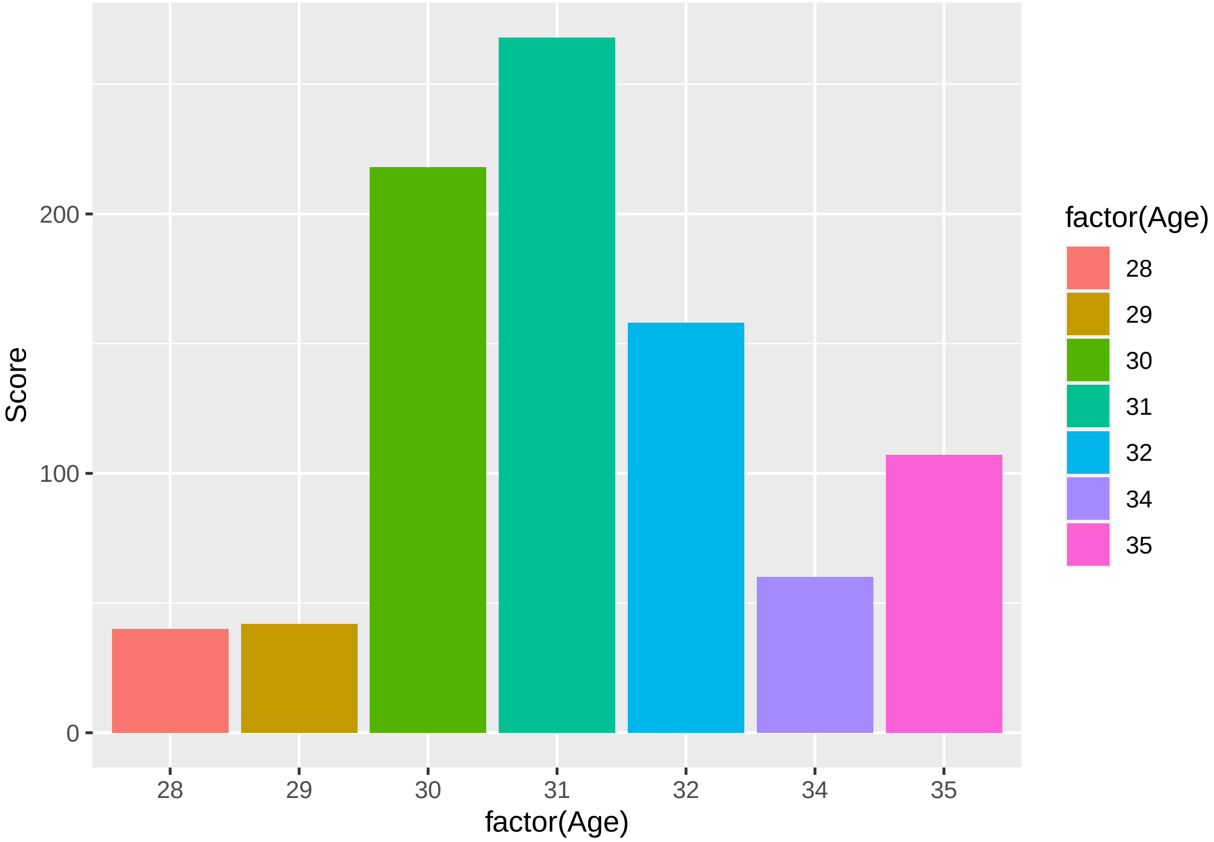
```
#subsetting female customers based on age group of highest score
female_filtered=filter(female,Age>=28 & Age<38)
head(female_filtered)
```

```
## # A tibble: 6 x 5
##   CustomerID Genre    Age Income Score
##   <chr>      <chr>  <dbl>  <dbl> <dbl>
## 1 0005       Female    31     17    40
## 2 0010       Female    30     19    72
## 3 0017       Female    35     21    35
## 4 0038       Female    30     34    73
## 5 0044       Female    31     39    61
## 6 0049       Female    29     40    42
```

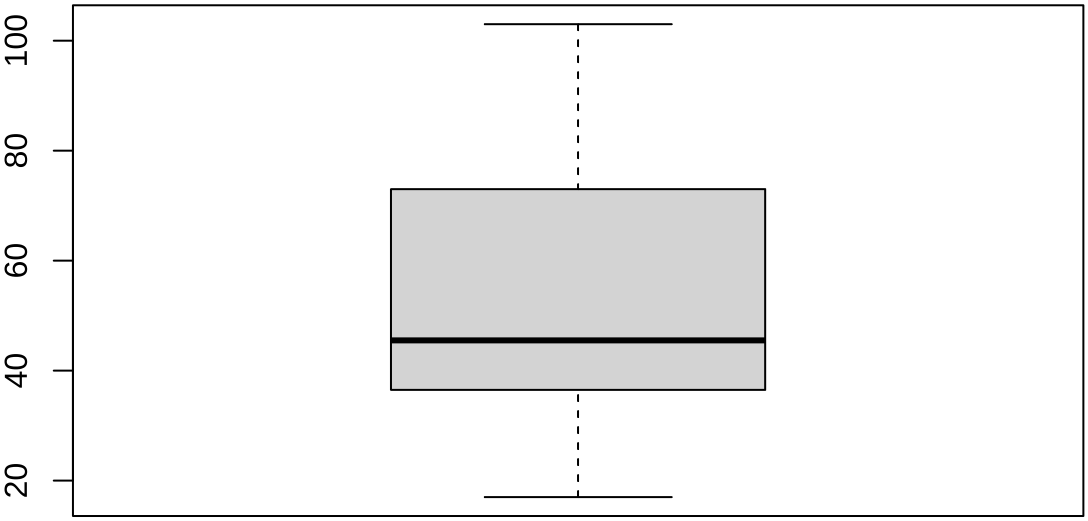
```
summary(female_filtered)
```

```
##   CustomerID           Genre                Age             Income      
##  Length:16          Length:16          Min.   :28.00   Min.   : 17.00  
##  Class :character   Class :character   1st Qu.:30.00   1st Qu.: 37.75  
##  Mode  :character   Mode  :character   Median :31.00   Median : 45.50  
##                                        Mean   :31.38   Mean   : 51.38  
##                                        3rd Qu.:32.00   3rd Qu.: 72.50  
##                                        Max.   :35.00   Max.   :103.00  
##      Score      
##  Min.   :35.00  
##  1st Qu.:42.00  
##  Median :57.00  
##  Mean   :55.81  
##  3rd Qu.:71.25
##  Max.   :73.00
```

```
#BAR CHART OF TWO ATTRIBUTES
ggplot(female_filtered, aes(x =factor(Age), y = Score,fill=factor(Age))) +
  geom_bar(stat = "identity")
```



```
#boxplot
boxplot(female_filtered$Income)
```



```r
#Correlation
cor(female_filtered$Income,female_filtered$Score)
```

```
## [1] 0.3731415
```

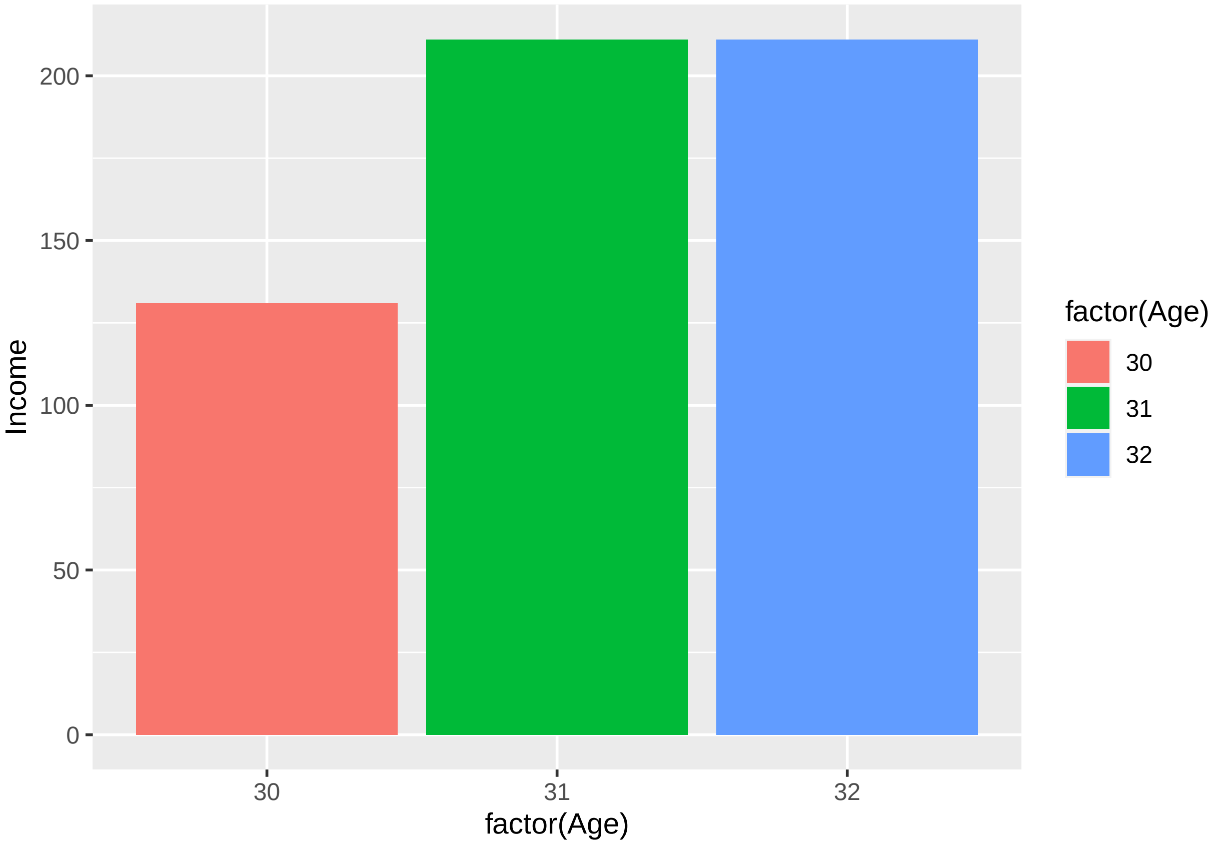
```r
cor.test(female_filtered$Income,female_filtered$Score)
```

```
##
##  Pearson's product-moment correlation
##
## data:  female_filtered$Income and female_filtered$Score
## t = 1.5049, df = 14, p-value = 0.1546
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1503792  0.7332238
## sample estimates:
##       cor
## 0.3731415
```

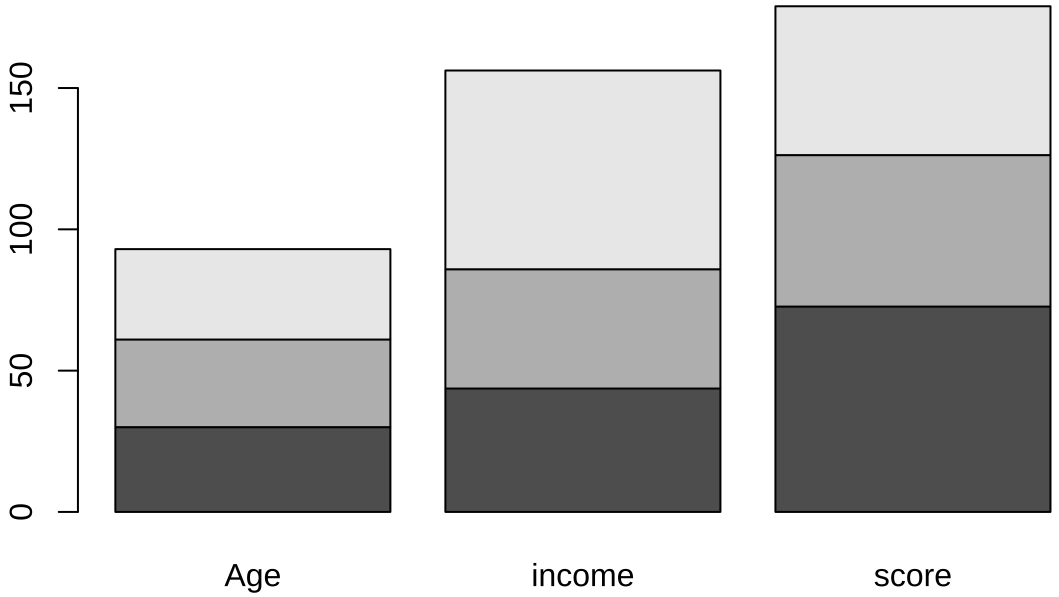
```r
#subsetting female customers based on age of highest score
female_filtered_score = filter(female_filtered,Age>=30 & Age<=32)
head(female_filtered_score)
```

```
## # A tibble: 6 x 5
##   CustomerID Genre    Age Income Score
##   <chr>      <chr>  <dbl>  <dbl> <dbl>
## 1 0005       Female    31     17    40
## 2 0010       Female    30     19    72
## 3 0038       Female    30     34    73
## 4 0044       Female    31     39    61
## 5 0050       Female    31     40    42
## 6 0053       Female    31     43    54
```

```r
#BAR CHART OF TWO ATTRIBUTES
ggplot(female_filtered_score, aes(x =factor(Age), y = Income,fill=factor(Age))) +
  geom_bar(stat = "identity")
```



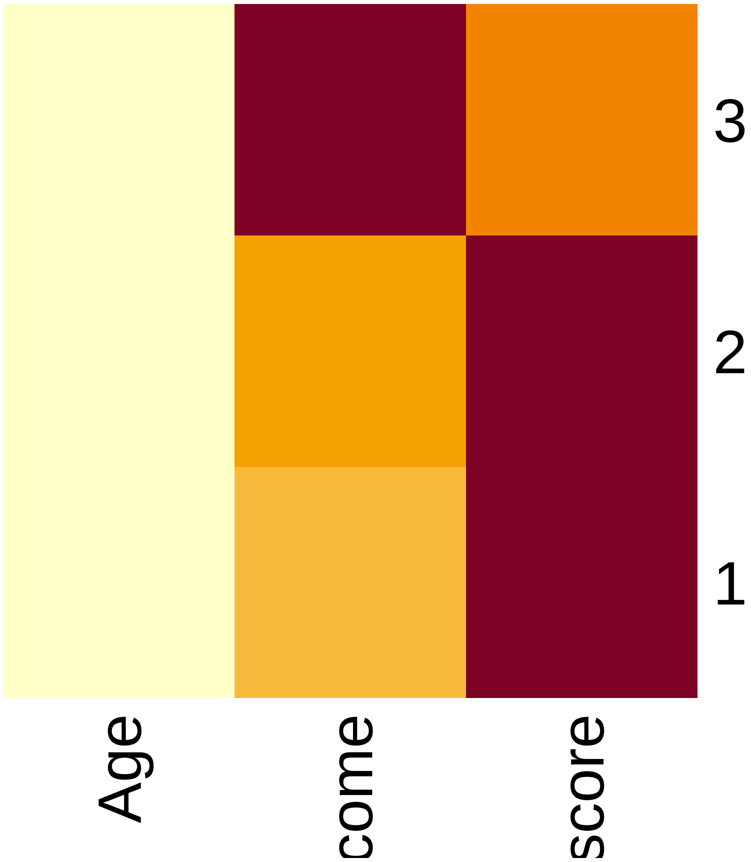
```
#group_by age and summarise
y=female_filtered_score%>%group_by(Age)%>%
  summarise(income=mean(Income))
z=female_filtered_score%>%group_by(Age)%>%
  summarise(score=mean(Score))
average_1=merge(y,z)

#barplot
barplot(as.matrix(average_1,col=c("orange","white","green")))
```



```
#Correlation btw average income and average score
cor(y,z)
```

```
##              Age      score
## Age    1.000000 -0.8859572
## income 0.841943 -0.4956923
```

```
#heatmap
heatmap(as.matrix(average_1),Rowv = NA, Colv = NA)
```



```
#subsetting male customers
male =filter(df,Genre=='Male')
#BASIC INSIGHTS
print(male)
```

```
## # A tibble: 45 x 5
##    CustomerID Genre   Age Income Score
##    <chr>      <chr> <dbl>  <dbl> <dbl>
##  1 0001       Male     19     15    39
##  2 0018       Male     20     21    66
##  3 0021       Male     35     24    35
##  4 0022       Male     25     24    73
##  5 0024       Male     31     25    73
##  6 0028       Male     35     28    61
##  7 0043       Male     48     39    36
##  8 0052       Male     33     42    60
##  9 0054       Male     59     43    60
## 10 0056       Male     47     43    41
## # ... with 35 more rows
```

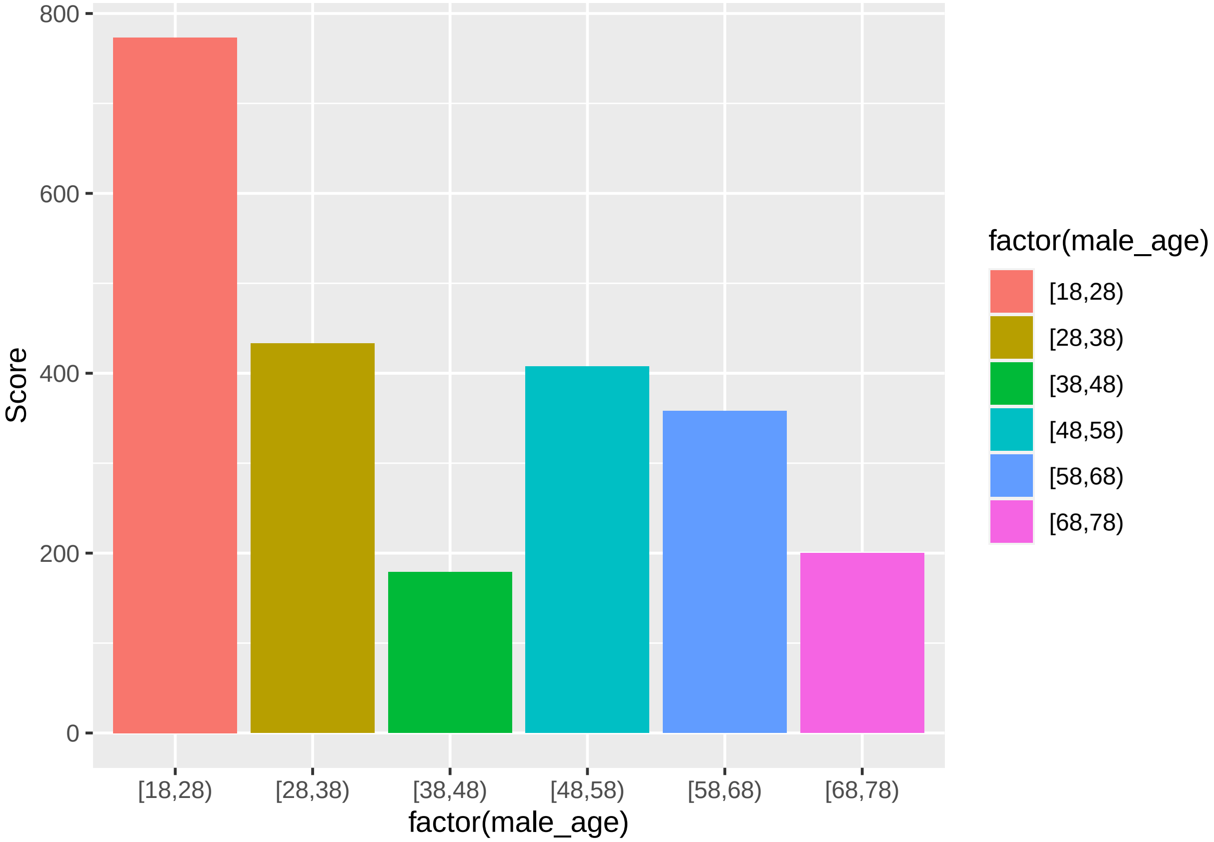
```
summary(male)
```

```
##   CustomerID           Genre                Age            Income     
##  Length:45          Length:45          Min.   :18.00   Min.   : 15.00  
##  Class :character   Class :character   1st Qu.:26.00   1st Qu.: 46.00  
##  Mode  :character   Mode  :character   Median :40.00   Median : 54.00  
##                                        Mean   :41.71   Mean   : 53.87  
##                                          3rd Qu.:57.00   3rd Qu.: 63.00
##                                          Max.   :70.00   Max.   :101.00
##      Score
##  Min.   :35.00
##  1st Qu.:46.00
##  Median :52.00
##  Mean   :52.24
##  3rd Qu.:59.00
##  Max.   :73.00
```

```
#BAR CHART OF TWO ATTRIBUTES
male_age = pull(male,Age)
male_age=cut(male_age,breaks=seq(18,80,by=10),right=FALSE)
table(male_age)
```

```
## male_age
## [18,28) [28,38) [38,48) [48,58) [58,68) [68,78)
##      14       7       4       9       7       4
```

```
ggplot(male, aes(x =factor(male_age), y = Score,fill=factor(male_age))) +
  geom_bar(stat = "identity")
```



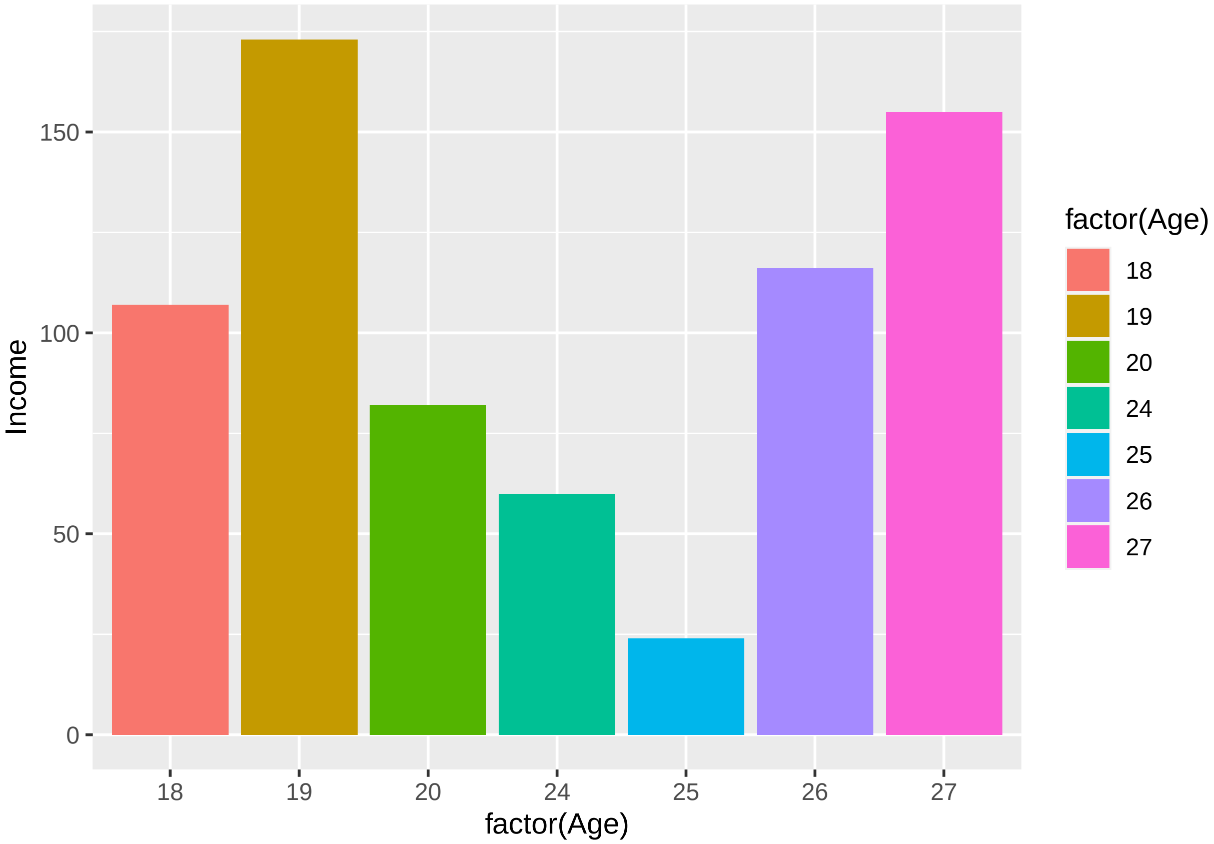
```
#subsetting male customers based on age group highest score
male_filtered=filter(male,Age>=18 & Age<28)
print(male_filtered)
```

```
## # A tibble: 14 x 5
##    CustomerID Genre   Age Income Score
##    <chr>      <chr> <dbl>  <dbl> <dbl>
##   1 0001       Male     19     15    39
##   2 0018       Male     20     21    66
##   3 0022       Male     25     24    73
##   4 0062       Male     19     46    55
##   5 0066       Male     18     48    59
##   6 0069       Male     19     48    59
##   7 0076       Male     26     54    54
##   8 0092       Male     18     59    41
##   9 0096       Male     24     60    52
##  10 0100       Male     20     61    49
##  11 0104       Male     26     62    55
##  12 0114       Male     19     64    46
##  13 0121       Male     27     67    56
##  14 0178       Male     27     88    69
```

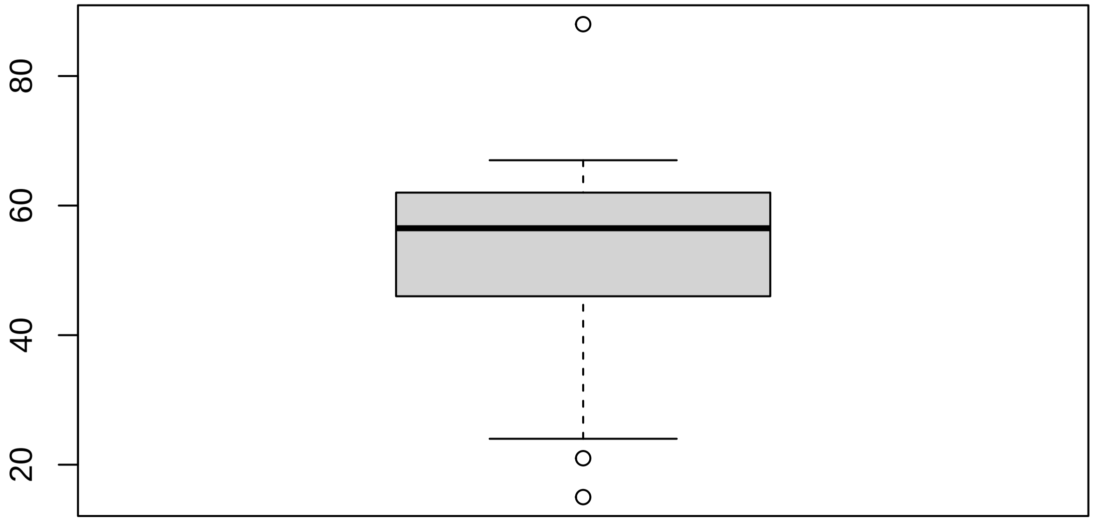
```
summary(male_filtered)
```

```
##   CustomerID           Genre                Age            Income     
##  Length:14          Length:14          Min.   :18.00   Min.   :15.00  
##  Class :character   Class :character   1st Qu.:19.00   1st Qu.:46.50  
##  Mode  :character   Mode  :character   Median :20.00   Median :56.50  
##                                        Mean   :21.93   Mean   :51.21  
##                                        3rd Qu.:25.75   3rd Qu.:61.75  
##                                        Max.   :27.00   Max.   :88.00  
##      Score      
##  Min.   :39.00  
##  1st Qu.:49.75  
##  Median :55.00  
##  Mean   :55.21  
##  3rd Qu.:59.00  
##  Max.   :73.00
```

```
#BAR CHART OF TWO ATTRIBUTES
ggplot(male_filtered, aes(x =factor(Age), y = Income,fill =factor(Age) )) +
  geom_bar(stat = "identity")
```



```
#boxplot - Income
boxplot(male_filtered$Income)
```



```
#correlation
cor(male_filtered$Income,male_filtered$Score)
```

```
## [1] -0.01957335
```

```
cor.test(male_filtered$Income,male_filtered$Score)
```

```
##
##  Pearson's product-moment correlation
```

```
##
## data:  male_filtered$Income and male_filtered$Score
## t = -0.067817, df = 12, p-value = 0.947
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5444981  0.5163688
## sample estimates:
##         cor
## -0.01957335
```

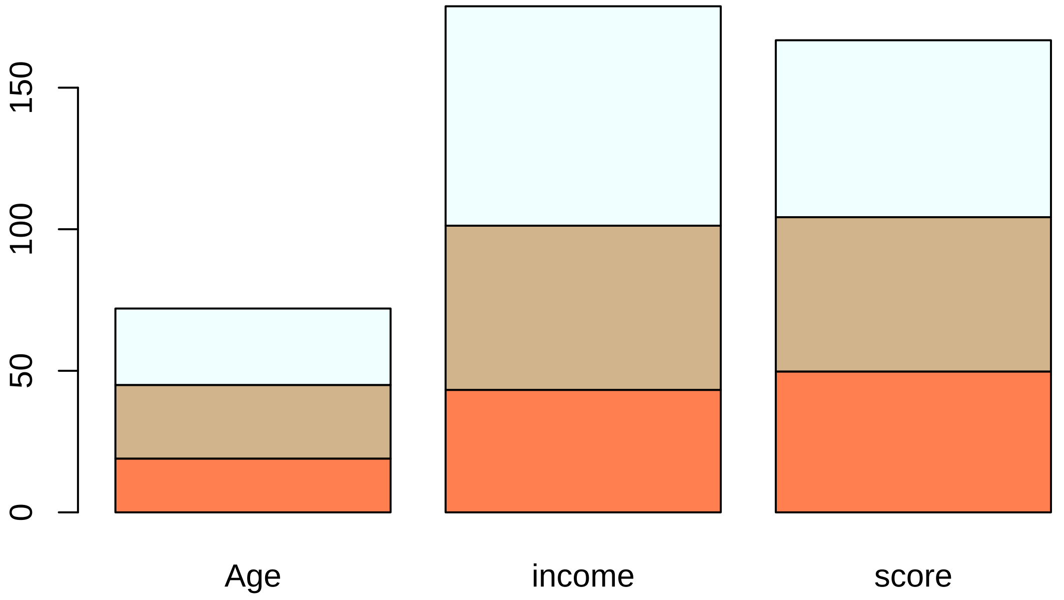
```
#subsetting male customers based on age of highest score
male_filtered_score = filter(male_filtered,Age==19 | Age==26 |Age ==27)
head(male_filtered_score)
```

```
## # A tibble: 6 x 5
##   CustomerID Genre   Age Income Score
##   <chr>      <chr> <dbl>  <dbl> <dbl>
## 1 0001       Male     19     15    39
## 2 0062       Male     19     46    55
## 3 0069       Male     19     48    59
## 4 0076       Male     26     54    54
## 5 0104       Male     26     62    55
## 6 0114       Male     19     64    46
```

```
#group_by age and summarise
a=male_filtered_score%>%group_by(Age)%>%
  summarise(income=mean(Income))

b=male_filtered_score%>%group_by(Age)%>%
  summarise(score=mean(Score))
average =merge(a,b)
average
```

```
##   Age income score
## 1  19  43.25 49.75
## 2  26  58.00 54.50
## 3  27  77.50 62.50
```

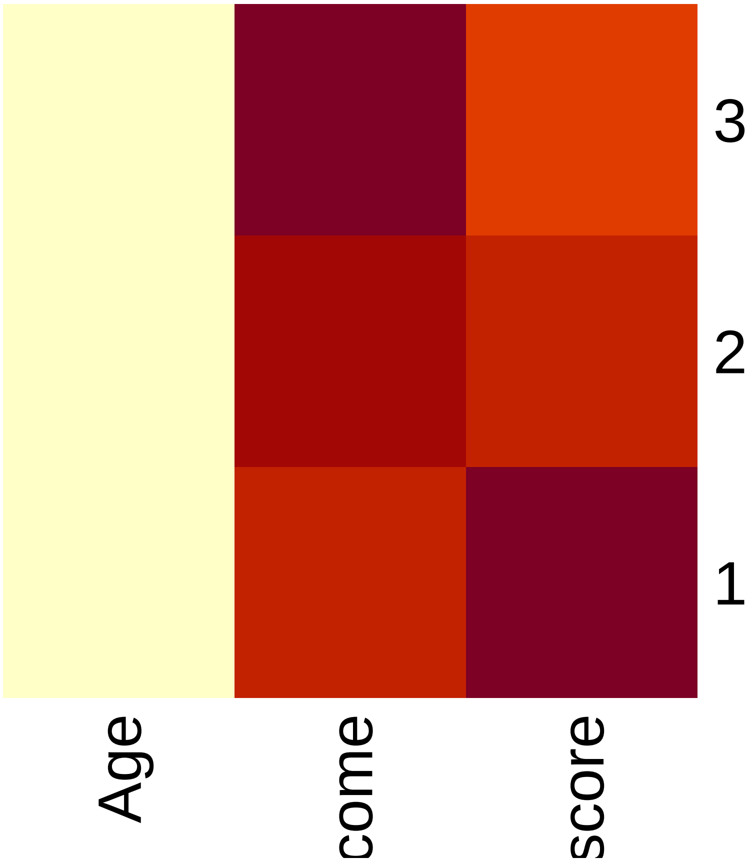
```
#barplot
barplot(as.matrix(average),col=c("coral","tan","azure"))
```



```
#Correlation btw average income and average score
cor(a,b)
```

```
##             Age     score
## Age     1.00000 0.8500286
## income  0.88302 0.9978083
```

```
#Heatmap
heatmap(as.matrix(average),Rowv = NA, Colv = NA)
```


3
2
1
Age
come
score