Importing the required libraries

```
library(ggplot2)
library(readxl)
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
```

Read the Data set

```
dt <- read.csv(file.choose())
```

```
head(dt)
```

```
##     Date Day      Time    Region CardType Gender BuyCategory ItemsOrdered
## 1 06-Mar Mon   Morning      West  Loyalty Female        High            4
## 2 06-Mar Mon   Morning      West  Loyalty Female      Medium            1
## 3 06-Mar Mon Afternoon      West  Loyalty Female      Medium            5
## 4 06-Mar Mon Afternoon NorthEast  Loyalty Female         Low            1
## 5 06-Mar Mon Afternoon      West  Loyalty   Male      Medium            4
## 6 06-Mar Mon Afternoon NorthEast    Other Female      Medium            5
##   TotalCost HighItem
## 1    136.97    79.97
## 2     25.55    25.55
## 3    113.95    90.47
## 4      6.82     6.82
## 5    147.32    83.21
## 6    142.15    50.90
```

Structure of the data

```
str(dt)
```

```
## 'data.frame':    403 obs. of  10 variables:
##  $ Date        : chr  "06-Mar" "06-Mar" "06-Mar" "06-Mar" ...
##  $ Day         : chr  "Mon" "Mon" "Mon" "Mon" ...
##  $ Time        : chr  "Morning" "Morning" "Afternoon" "Afternoon" ...
##  $ Region      : chr  "West" "West" "West" "NorthEast" ...
##  $ CardType    : chr  "Loyalty" "Loyalty" "Loyalty" "Loyalty" ...
##  $ Gender      : chr  "Female" "Female" "Female" "Female" ...
##  $ BuyCategory : chr  "High" "Medium" "Medium" "Low" ...
##  $ ItemsOrdered: int  4 1 5 1 4 5 1 4 4 2 ...
##  $ TotalCost   : num  136.97 25.55 113.95 6.82 147.32 ...
##  $ HighItem    : num  79.97 25.55 90.47 6.82 83.21 ...
```

summary of the data

```
summary(dt)
```

```
##      Date               Day                Time              Region         
##  Length:403         Length:403         Length:403         Length:403        
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##    CardType            Gender          BuyCategory         ItemsOrdered   
##  Length:403         Length:403         Length:403         Min.   : 1.000  
##  Class :character   Class :character   Class :character   1st Qu.: 2.000  
##  Mode  :character   Mode  :character   Mode  :character   Median : 3.000  
##                                                           Mean   : 3.476  
##                                                           3rd Qu.: 4.000  
##                                                           Max.   :11.000  
##    TotalCost         HighItem     
##  Min.   :-90.00   Min.   :  6.82  
##  1st Qu.: 82.86   1st Qu.: 56.28  
##  Median :126.16   Median : 83.62  
##  Mean   :152.08   Mean   :100.61  
##  3rd Qu.:204.02   3rd Qu.:119.46  
##  Max.   :485.01   Max.   :381.33  
```

Assign factors to character values

```
dt$Date <- as.factor(dt$Date)
dt$Day <- as.factor(dt$Day)
dt$Time <- as.factor(dt$Time)
dt$CardType <- as.factor(dt$CardType)
dt$Gender <- as.factor(dt$Gender)
dt$BuyCategory <- as.factor(dt$BuyCategory)
dt$Region <- as.factor(dt$Region)
```

```
#structure after factor the character columns
```

```
str(dt)
```

```
## 'data.frame':    403 obs. of  10 variables:
##  $ Date        : Factor w/ 112 levels "01-Apr","01-Jun",..: 18 18 18 18 18 18 22 22 26 22 ...
##  $ Day         : Factor w/ 7 levels "Fri","Mon","Sat",..: 2 2 2 2 2 2 6 6 6 6 ...
##  $ Time        : Factor w/ 3 levels "Afternoon","Evening",..: 3 3 1 1 1 1 2 2 2 2 ...
##  $ Region      : Factor w/ 4 levels "MidWest","NorthEast",..: 4 4 4 2 4 2 4 3 3 4 ...
##  $ CardType    : Factor w/ 2 levels "Loyalty","Other": 1 1 1 1 1 2 2 2 2 1 ...
##  $ Gender      : Factor w/ 2 levels "Female","Male": 1 1 1 1 2 1 2 2 2 2 ...
##  $ BuyCategory : Factor w/ 3 levels "High","Low","Medium": 1 3 3 2 3 3 2 1 1 2 ...
##  $ ItemsOrdered: int  4 1 5 1 4 5 1 4 4 2 ...
##  $ TotalCost   : num  136.97 25.55 113.95 6.82 147.32 ...
##  $ HighItem    : num  79.97 25.55 90.47 6.82 83.21 ...
```

checking for null values in the data set

```
colSums(is.na(dt))
```

```
##         Date          Day         Time       Region     CardType       Gender 
##            0            0            0            0            0            0 
##  BuyCategory ItemsOrdered    TotalCost     HighItem 
##            0            0            0            0
```

```
unique(dt$Time)
```

```
## [1] Morning   Afternoon Evening  
## Levels: Afternoon Evening Morning
```

Using Dplyr package

```
data <- dt
```

filter the data whose Buy category is High & analysis with the gender & Items Ordered column

```
data %>% filter(BuyCategory == 'High') %>%
  select(Gender,ItemsOrdered,BuyCategory)%>%
  arrange(desc(ItemsOrdered))
```

```
##     Gender ItemsOrdered BuyCategory
## 1     Male           11        High
## 2     Male           10        High
## 3     Male           10        High
## 4   Female            9        High
## 5   Female            9        High
## 6   Female            9        High
## 7   Female            9        High
## 8   Female            9        High
## 9   Female            9        High
## 10    Male            9        High
## 11  Female            8        High
## 12  Female            8        High
## 13    Male            8        High
## 14  Female            8        High
## 15    Male            8        High
## 16  Female            7        High
## 17  Female            7        High
## 18  Female            7        High
## 19    Male            7        High
## 20    Male            7        High
## 21  Female            7        High
## 22  Female            7        High
## 23  Female            7        High
## 24  Female            7        High
## 25    Male            7        High
## 26  Female            7        High
## 27    Male            7        High
## 28  Female            7        High
## 29  Female            7        High
## 30  Female            7        High
## 31  Female            7        High
## 32    Male            7        High
## 33    Male            6        High
## 34    Male            6        High
## 35  Female            6        High
## 36  Female            6        High
## 37  Female            6        High
## 38  Female            6        High
## 39  Female            6        High
## 40    Male            6        High
## 41  Female            6        High
## 42    Male            6        High
## 43    Male            6        High
## 44  Female            6        High
## 45  Female            6        High
## 46  Female            6        High
## 47  Female            6        High
## 48  Female            6        High
## 49  Female            6        High
## 50  Female            6        High
## 51  Female            5        High
## 52  Female            5        High
## 53  Female            5        High
## 54    Male            5        High
## 55  Female            5        High
## 56    Male            5        High
## 57  Female            5        High
## 58    Male            5        High
## 59  Female            5        High
## 60  Female            5        High
## 61    Male            5        High
## 62    Male            5        High
## 63  Female            5        High
## 64  Female            5        High
## 65  Female            5        High
## 66  Female            5        High
## 67  Female            5        High
## 68  Female            4        High
## 69    Male            4        High
## 70    Male            4        High
## 71    Male            4        High
## 72    Male            4        High
## 73  Female            4        High
## 74  Female            4        High
## 75    Male            4        High
## 76  Female            4        High
## 77  Female            4        High
## 78  Female            4        High
## 79  Female            4        High
## 80    Male            4        High
## 81  Female            4        High
## 82  Female            4        High
## 83    Male            4        High
## 84  Female            4        High
## 85    Male            4        High
## 86    Male            4        High
## 87  Female            4        High
## 88    Male            4        High
## 89  Female            4        High
## 90  Female            4        High
## 91    Male            4        High
## 92    Male            4        High
## 93    Male            4        High
## 94  Female            4        High
## 95  Female            4        High
## 96    Male            3        High
## 97    Male            3        High
## 98  Female            3        High
## 99  Female            3        High
## 100   Male            3        High
## 101   Male            3        High
## 102 Female            3        High
## 103   Male            3        High
## 104 Female            3        High
## 105   Male            3        High
## 106   Male            3        High
## 107 Female            3        High
## 108 Female            3        High
## 109 Female            3        High
## 110   Male            3        High
## 111 Female            3        High
## 112 Female            2        High
## 113 Female            2        High
## 114   Male            2        High
## 115 Female            2        High
## 116   Male            2        High
## 117 Female            2        High
## 118 Female            2        High
## 119 Female            2        High
## 120 Female            1        High
## 121 Female            1        High
```

filter the Gender whose Buy category is High & Items Ordered is greater than 7

```
data %>% filter(BuyCategory == 'High') %>%
  filter(Gender == 'Male' & ItemsOrdered >= 7)%>%
  select(Gender,ItemsOrdered,BuyCategory)%>%
  arrange(desc(ItemsOrdered))
```

```
##    Gender ItemsOrdered BuyCategory
## 1    Male           11        High
## 2    Male           10        High
## 3    Male           10        High
## 4    Male            9        High
## 5    Male            8        High
## 6    Male            8        High
## 7    Male            7        High
## 8    Male            7        High
## 9    Male            7        High
## 10   Male            7        High
## 11   Male            7        High
```

slice() gets row by index position

```
data %>% slice(1,3,5)
```

```
##     Date Day      Time Region CardType Gender BuyCategory ItemsOrdered
## 1 06-Mar Mon   Morning   West  Loyalty Female        High            4
## 2 06-Mar Mon Afternoon   West  Loyalty Female      Medium            5
## 3 06-Mar Mon Afternoon   West  Loyalty   Male      Medium            4
##   TotalCost HighItem
## 1    136.97    79.97
## 2    113.95    90.47
## 3    147.32    83.21
```

Get columns that end with given string:

```
data %>% select(ends_with("Type")) %>% head()
```

```
##   CardType
## 1  Loyalty
## 2  Loyalty
## 3  Loyalty
## 4  Loyalty
## 5  Loyalty
## 6    Other
```

Get columns that match a string or regular expression:

```
data %>% select(matches("Buy")) %>% head()
```

```
##   BuyCategory
## 1        High
## 2      Medium
## 3      Medium
## 4         Low
## 5      Medium
## 6      Medium
```

Mutate() to add new variables to an existing data frame.

```
data %>% mutate(TotalCost_all_item = ItemsOrdered * TotalCost) %>% head()
```

```
##     Date Day      Time    Region CardType Gender BuyCategory ItemsOrdered
## 1 06-Mar Mon   Morning      West  Loyalty Female        High            4
## 2 06-Mar Mon   Morning      West  Loyalty Female      Medium            1
## 3 06-Mar Mon Afternoon      West  Loyalty Female      Medium            5
## 4 06-Mar Mon Afternoon NorthEast  Loyalty Female         Low            1
## 5 06-Mar Mon Afternoon      West  Loyalty   Male      Medium            4
## 6 06-Mar Mon Afternoon NorthEast    Other Female      Medium            5
##   TotalCost HighItem TotalCost_all_item
## 1    136.97    79.97             547.88
## 2     25.55    25.55              25.55
## 3    113.95    90.47             569.75
## 4      6.82     6.82               6.82
## 5    147.32    83.21             589.28
## 6    142.15    50.90             710.75
```

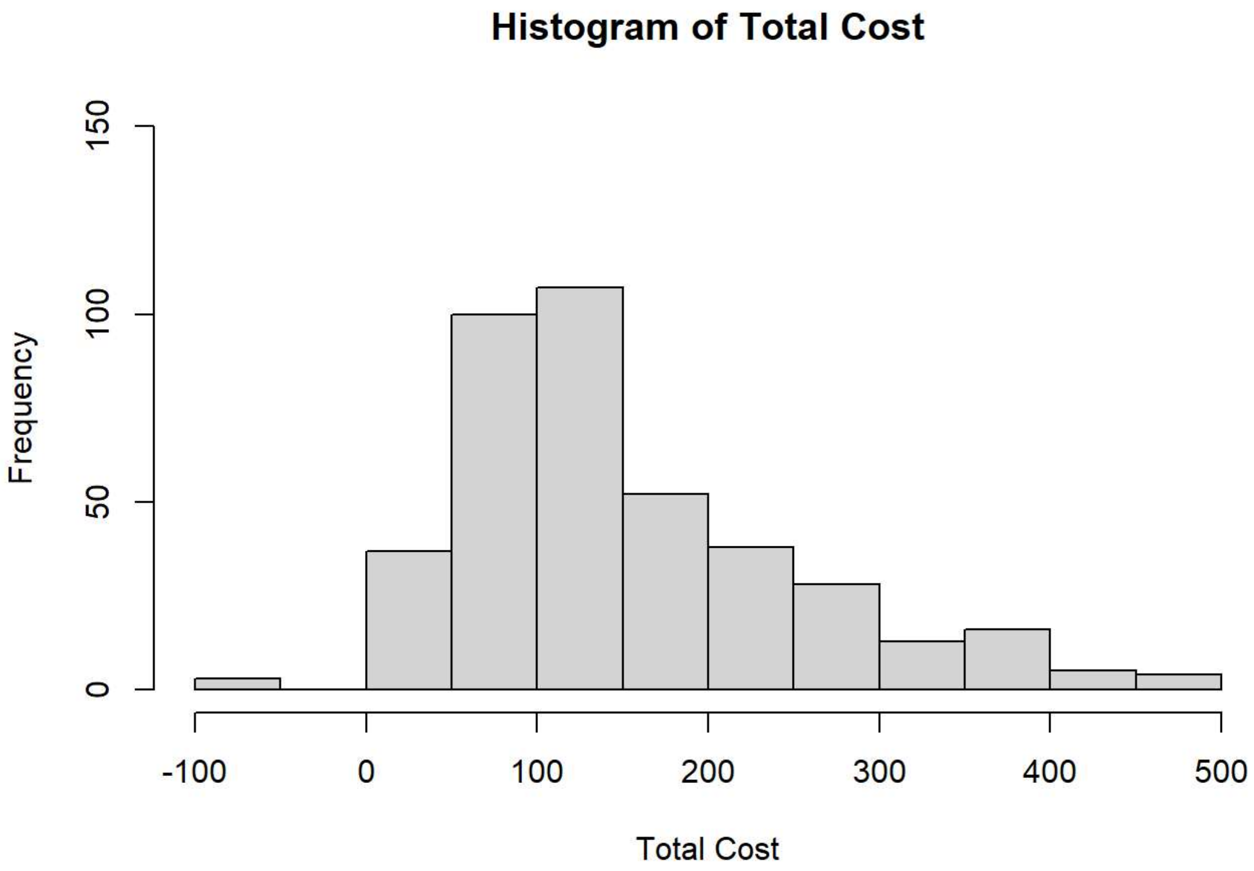
summarize()

```
data %>% summarize( Avg_TotalCost= mean(TotalCost))
```

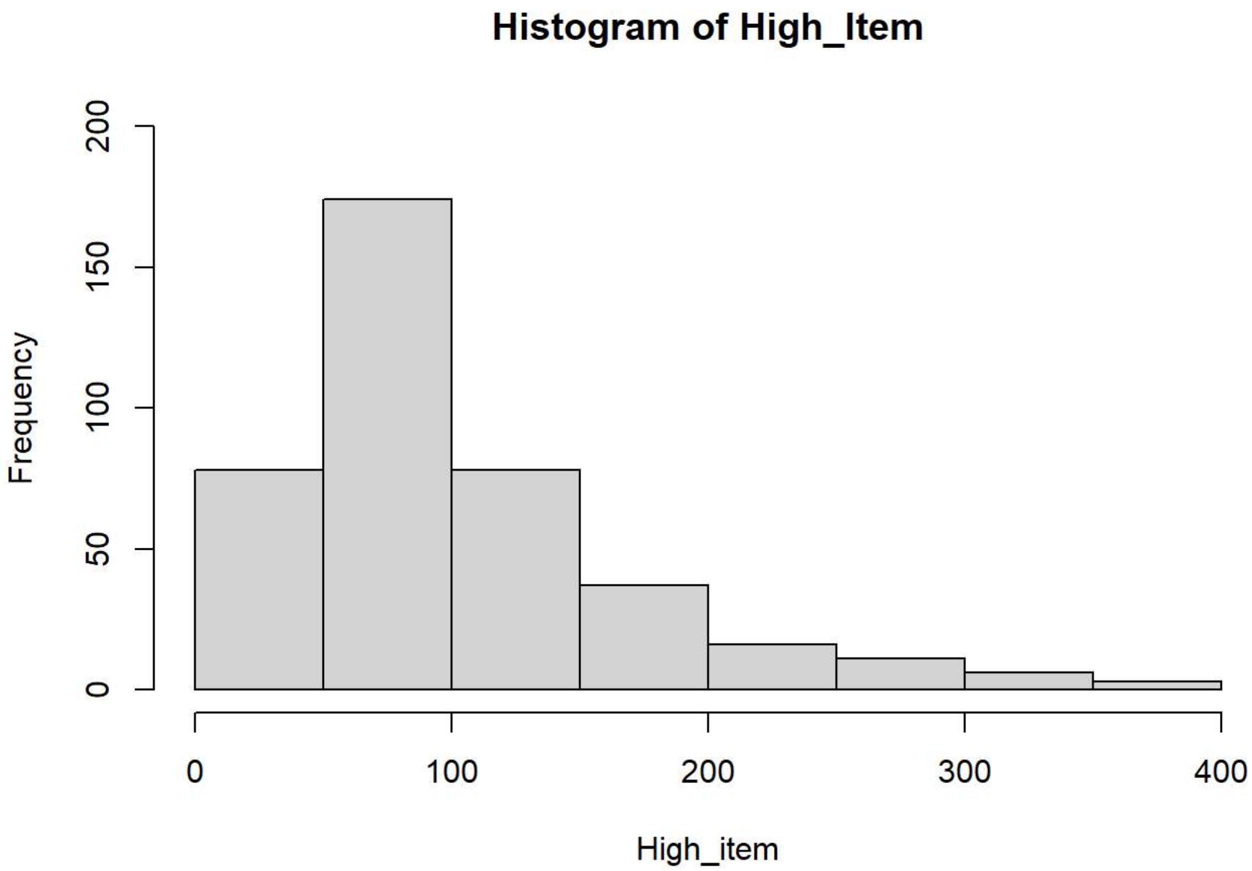
```
##   Avg_TotalCost
## 1      152.0849
```

Data visualization

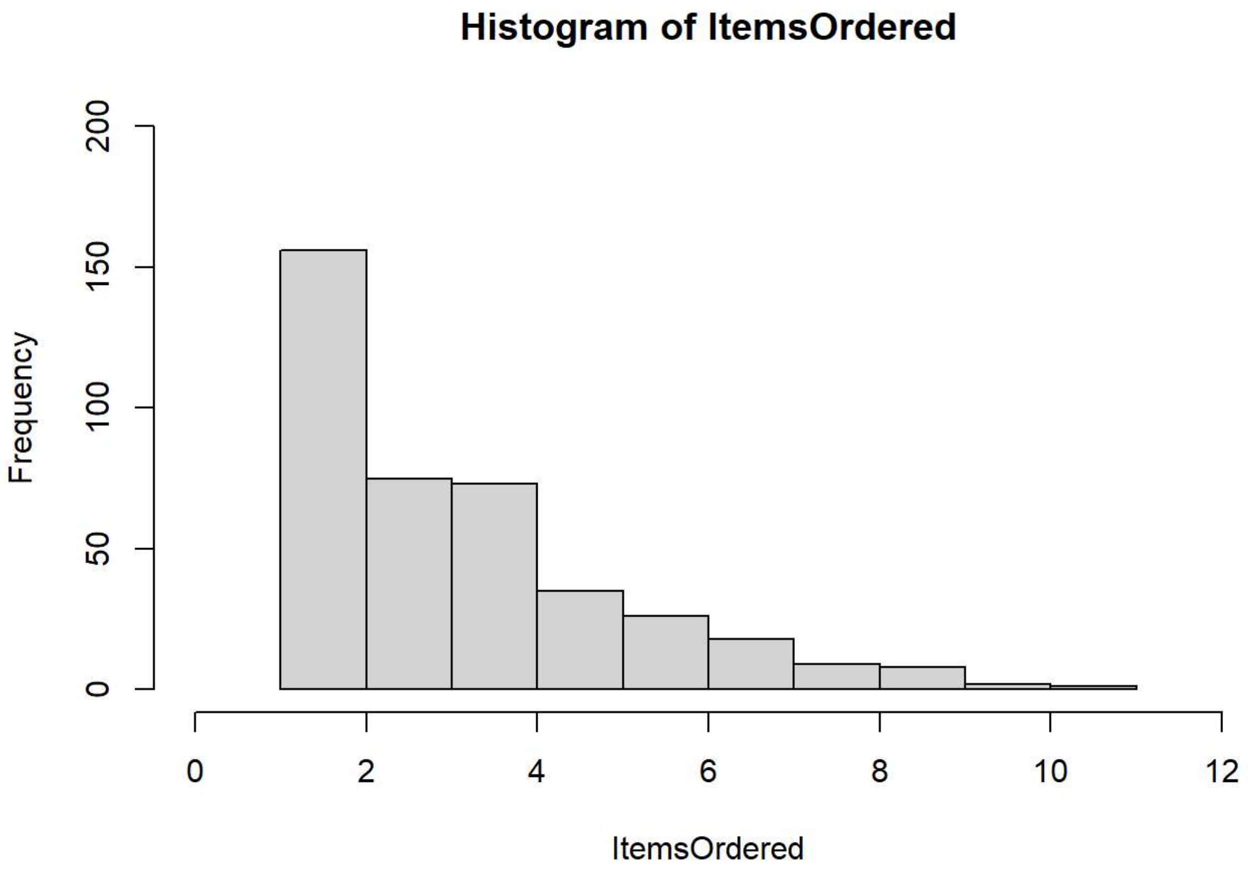
```
hist(dt$TotalCost,xlab = "Total Cost", main ="Histogram of Total Cost",ylim = c(0,150))
```



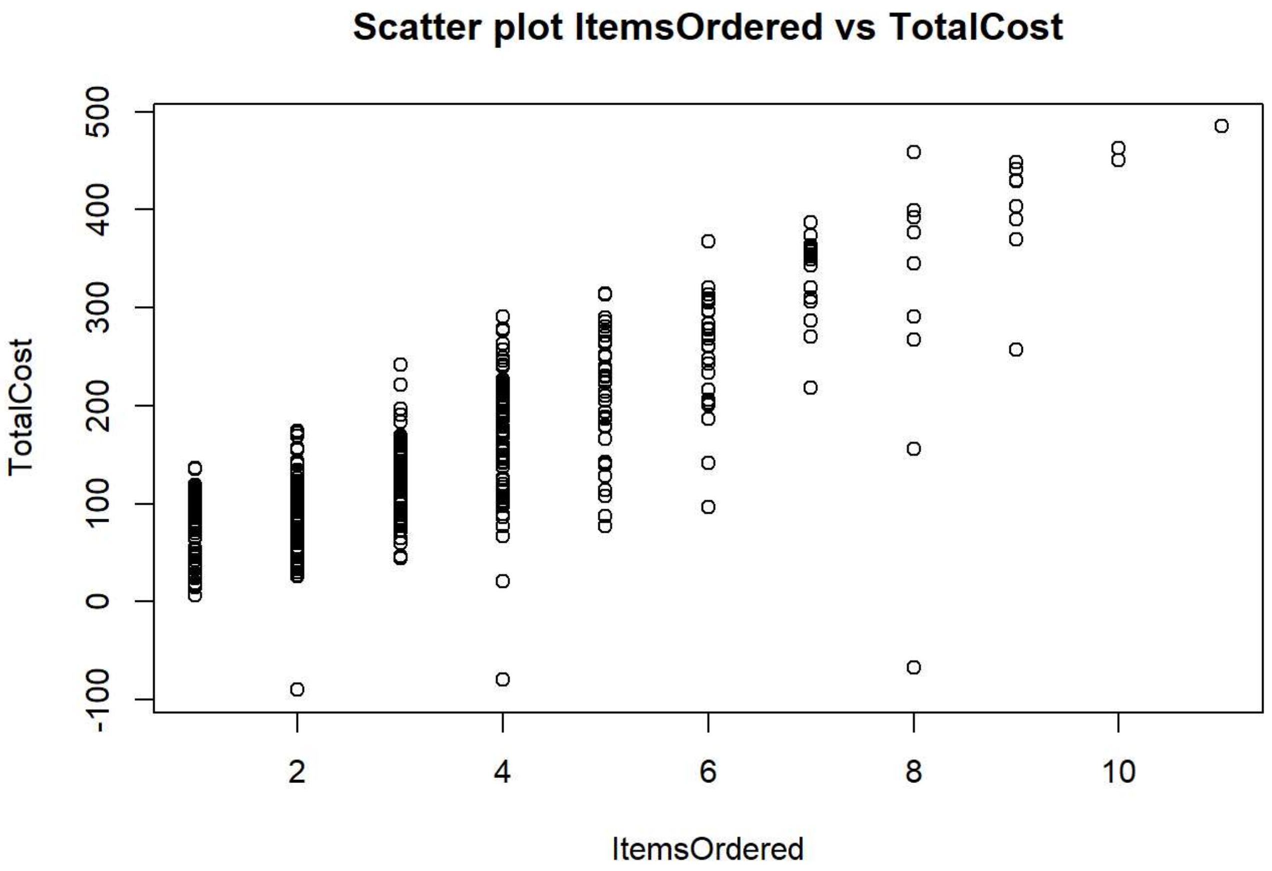
```
hist(dt$HighItem,main="Histogram of High_Item",xlab="High_item", ylim = c(0,200))
```



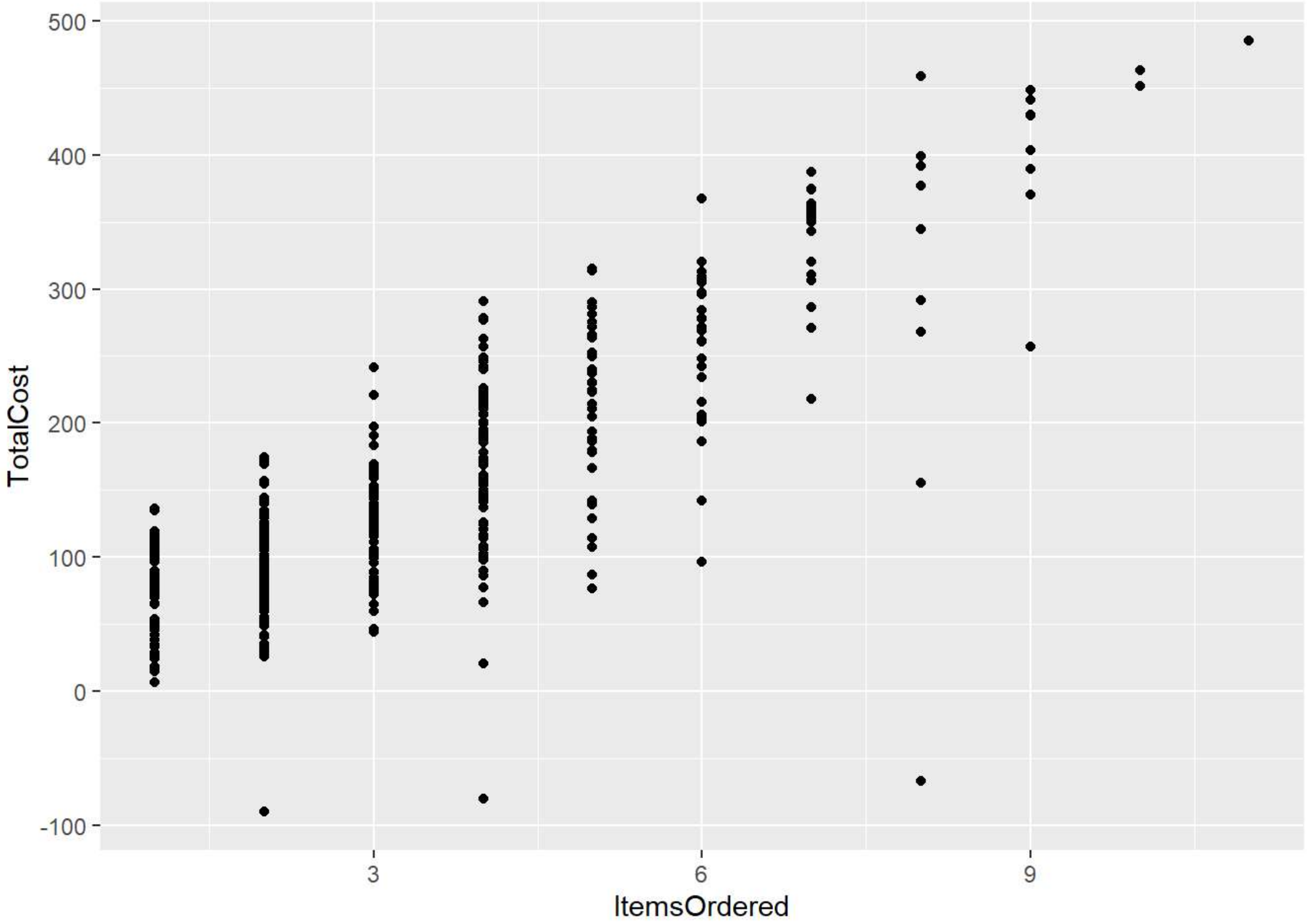
```
hist(dt$ItemsOrdered,main="Histogram of ItemsOrdered",xlab="ItemsOrdered",ylim = c(0,200),xlim =
c(0,12))
```



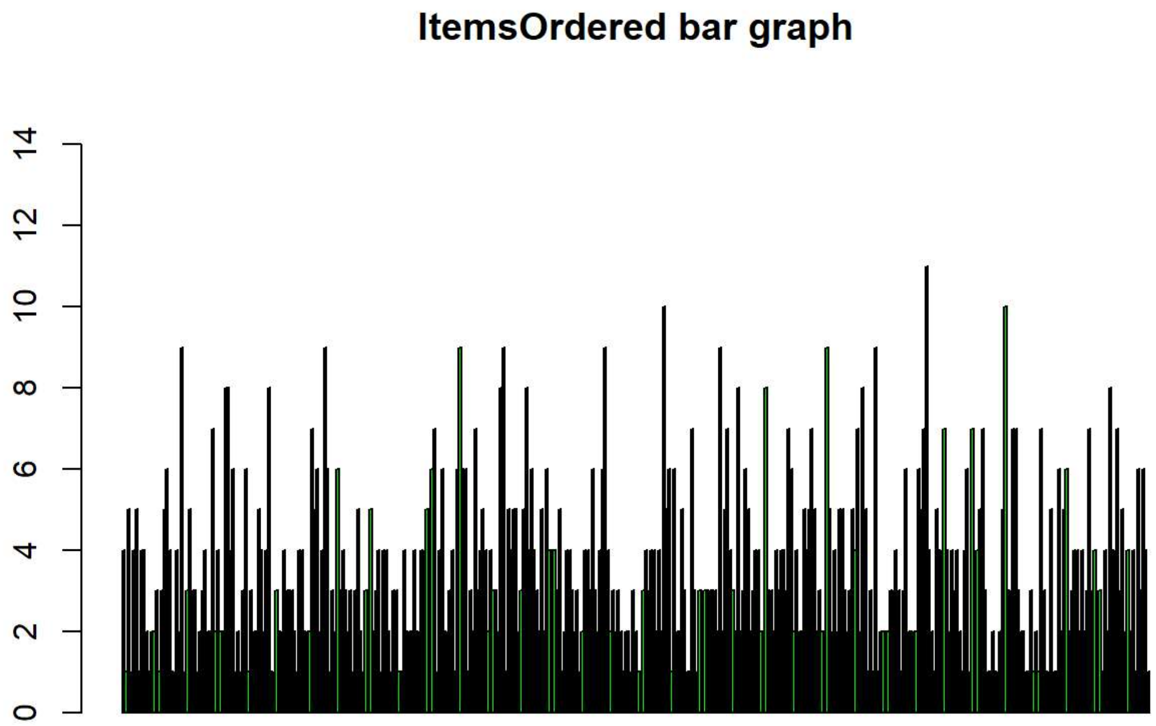
```
plot(dt$ItemsOrdered,dt$TotalCost,,main = "Scatter plot ItemsOrdered vs TotalCost",xlab = "Items
Ordered",ylab = "TotalCost", type= "p")
```



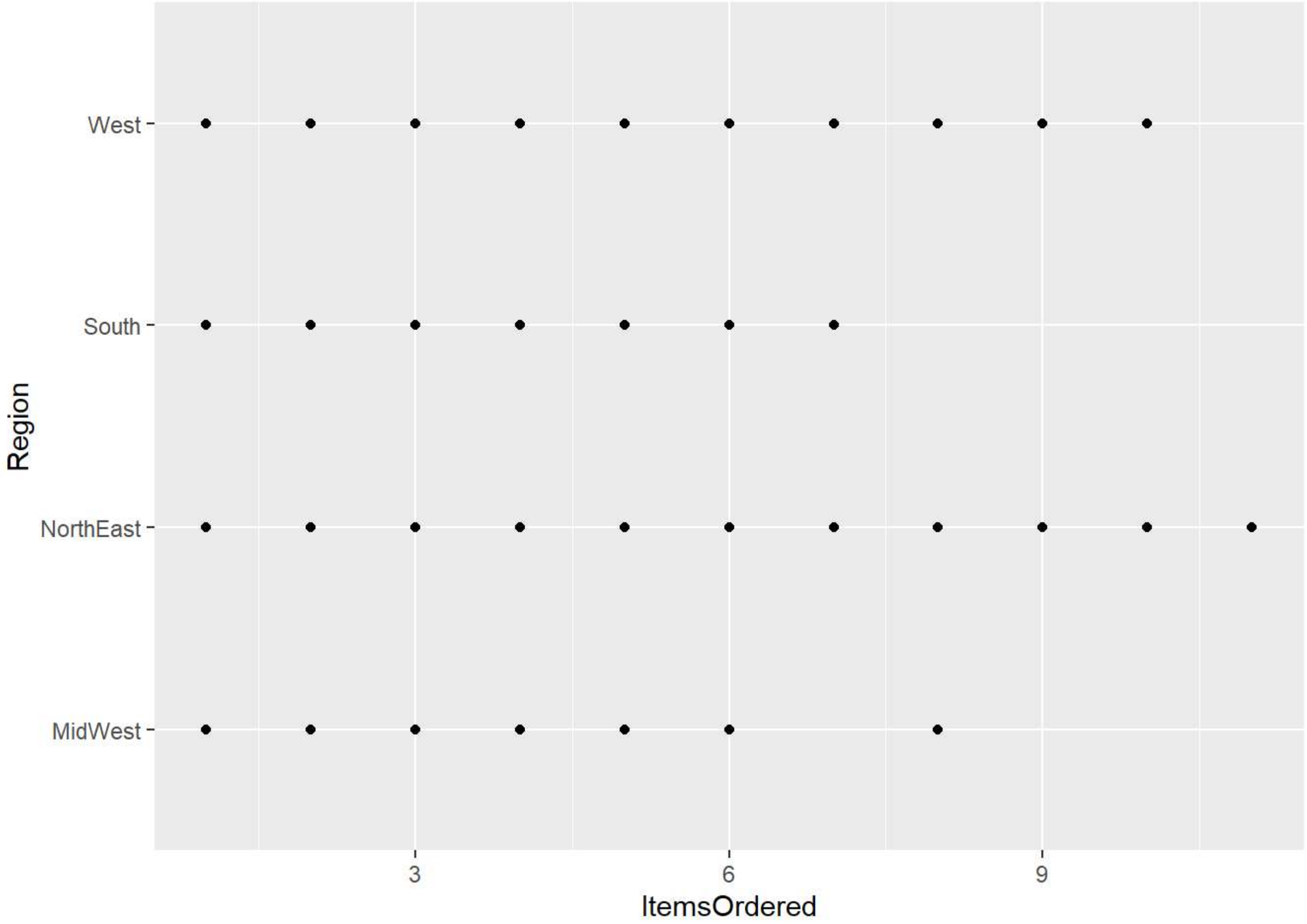
```
ggplot(dt,aes(y = TotalCost, x = ItemsOrdered)) +geom_point()
```



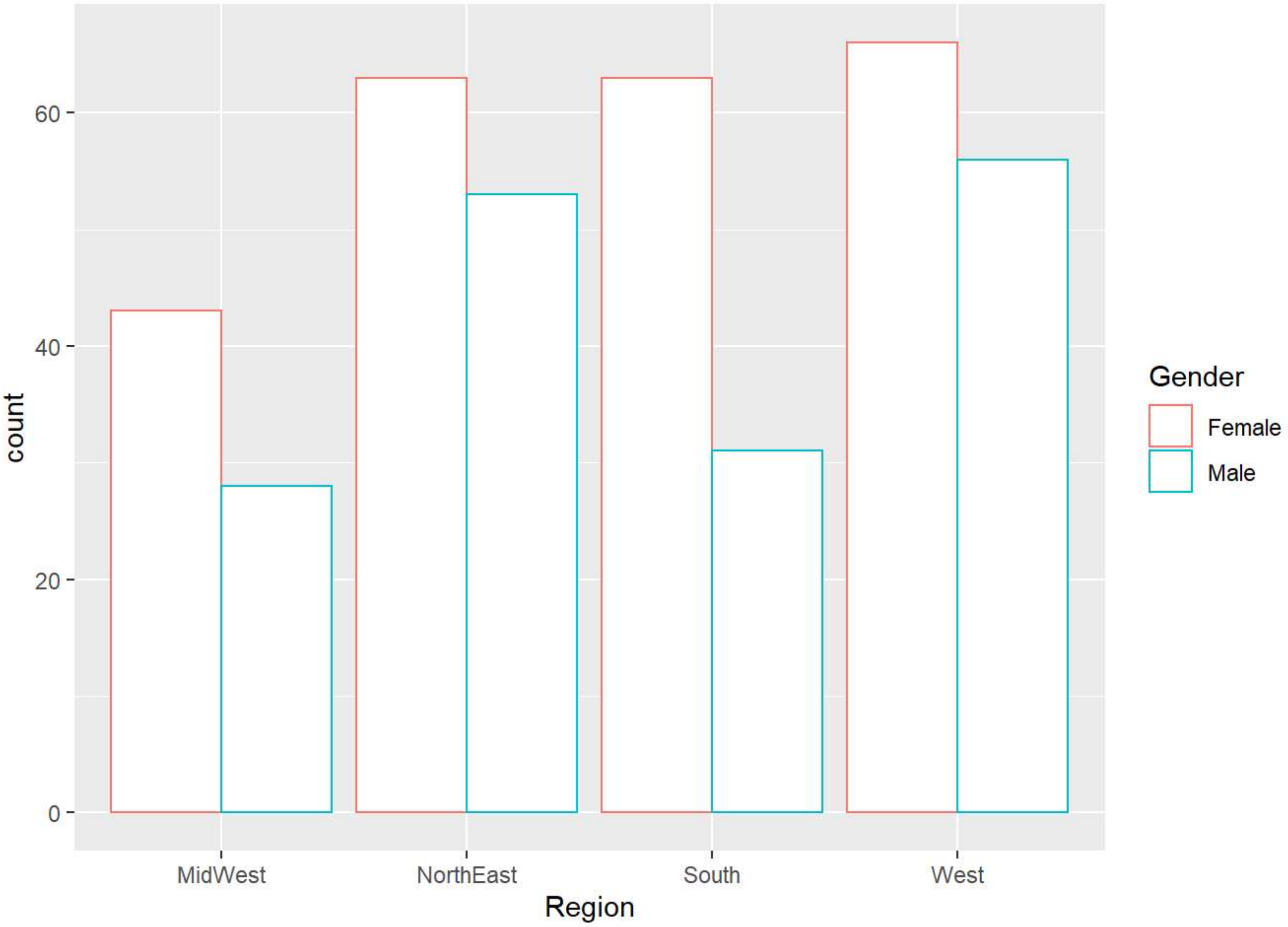
```
barplot(dt$ItemsOrdered,main = "ItemsOrdered bar graph",col = 'green',ylim = c(0,15))
```



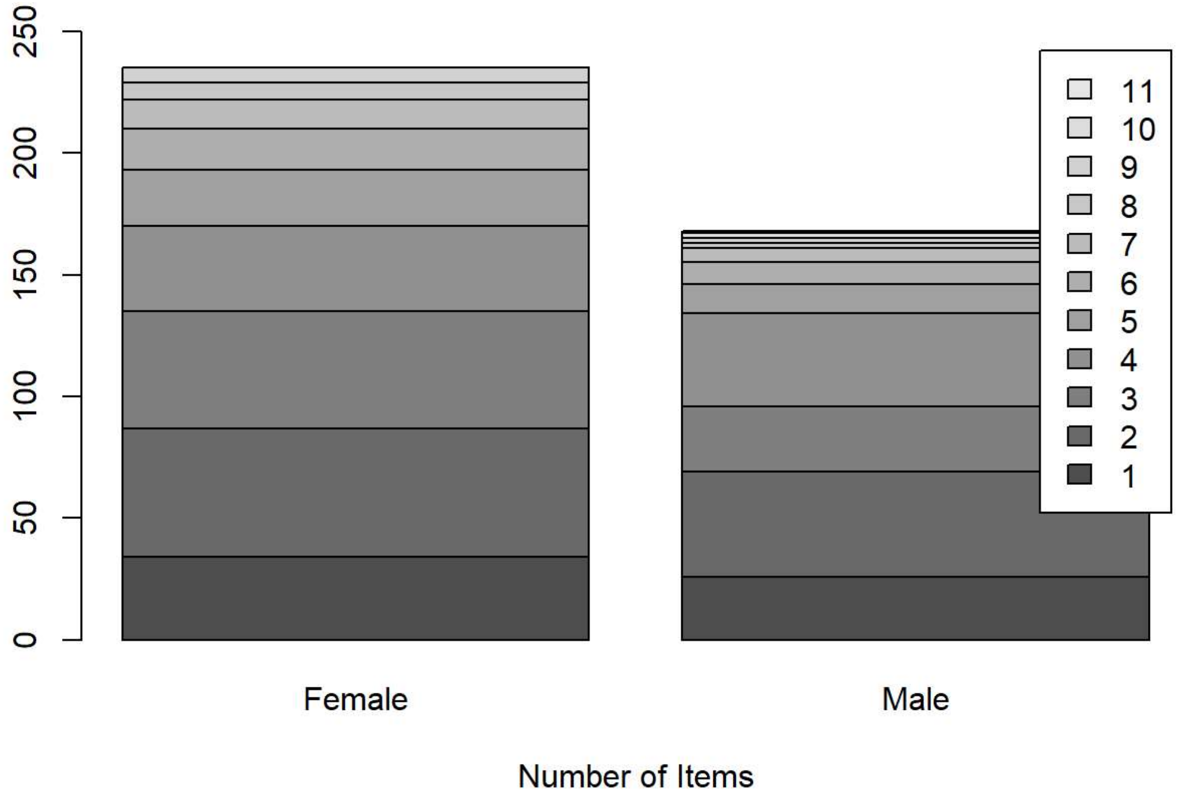
```
ggplot(dt,aes(y = Region, x = ItemsOrdered)) +geom_point()
```



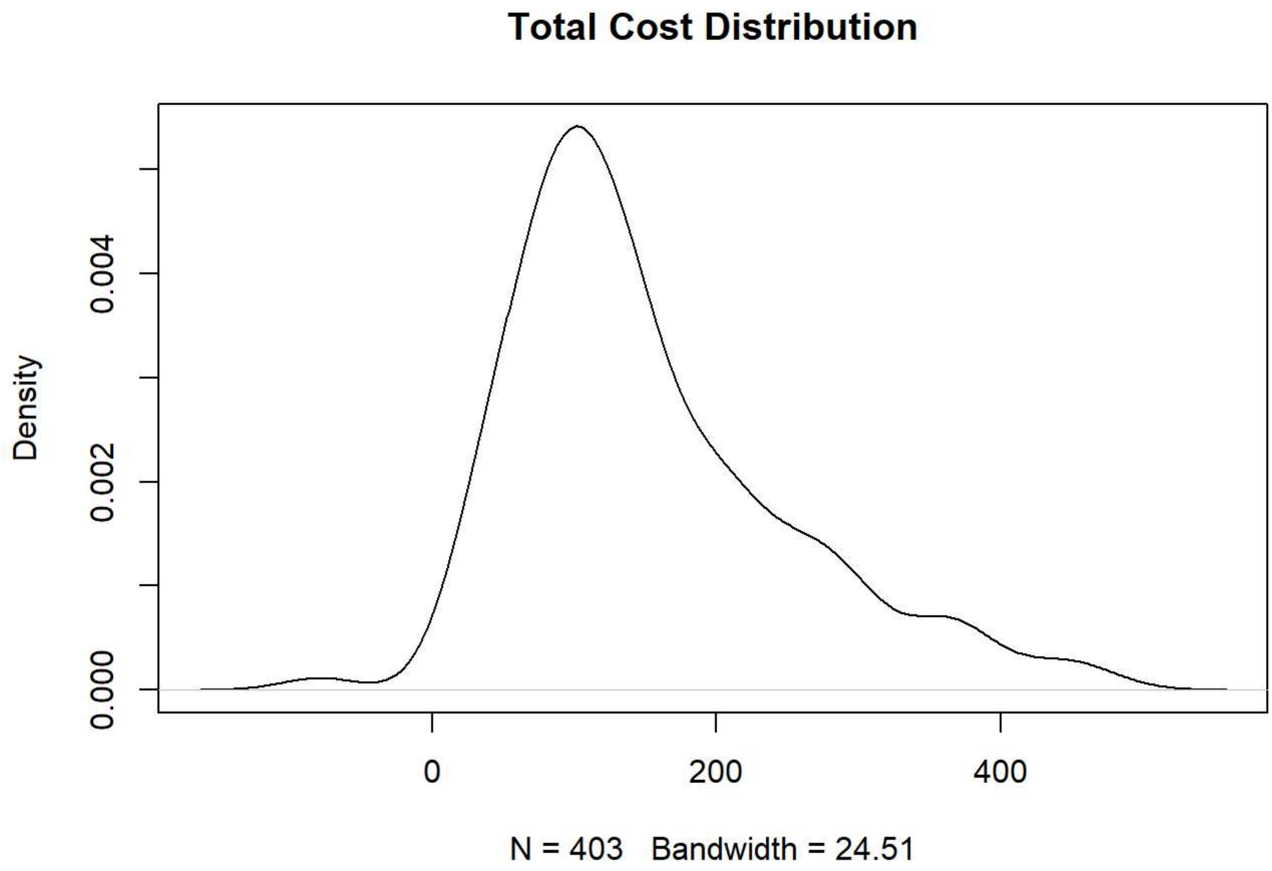
```
ggplot(dt) + geom_bar(mapping = aes(x=Region, color=Gender), fill='white', position='dodge')
```



```
#table(), performs categorical tabulation of data with the variable and its frequency
counts <- table(dt$ItemsOrdered, dt$Gender)
barplot(counts, main = '',xlab="Number of Items",legend = rownames(counts),ylim = c(0,250))
```



```
plot(density(dt$TotalCost), main='Total Cost Distribution')
```


Total Cost Distribution
Density
0.000
0.002
0.004
0
200
400
N = 403   Bandwidth = 24.51