Data Analysis Project on Laptop Dataset Using R

Laptop dataset is CSV(Comma-separated Values) format typically includes information about different laptop models.

It allows users to analyze and compare different laptops based on various attributes, which can be useful for **decision-making, market analysis and research.**

It contains 23 variables and 896 Observations

#Structure of dataset

```
'data.frame': 896 obs. of  23 variables:
 $ brand_name     : chr  "Lenovo" "Lenovo" "Avita" "Avita" ...
 $ model          : chr  "A6-9225" "Ideapad" "PURA" "PURA" ...
 $ processor_brand: chr  "AMD" "AMD" "AMD" "AMD" ...
 $ processor_name : chr  "A6-9225 Processor" "APU Dual" "APU Dual" "APU Dual" ...
 $ processor_gnrtn: chr  "10th" "10th" "10th" "10th" ...
 $ ram_gb         : chr  "4 GB GB" "4 GB GB" "4 GB GB" "4 GB GB" ...
 $ ram_type       : chr  "DDR4" "DDR4" "DDR4" "DDR4" ...
 $ ssd            : chr  "0 GB" "0 GB" "128 GB" "128 GB" ...
 $ hdd            : chr  "1024 GB" "512 GB" "0 GB" "0 GB" ...
 $ os             : chr  "Windows" "Windows" "Windows" "Windows" ...
 $ os_bit         : chr  "64-bit" "64-bit" "64-bit" "64-bit" ...
 $ graphic_card_gb: int  0 0 0 0 0 0 0 0 4 0 ...
 $ weight         : chr  "ThinNlight" "Casual" "ThinNlight" "ThinNlight" ...
 $ display_size   : num  15.1 15.1 15.1 15.1 15.1 ...
 $ warranty       : int  0 0 0 0 0 0 0 0 0 1 ...
 $ Touchscreen    : chr  "No" "No" "No" "No" ...
 $ msoffice       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ latest_price   : int  24990 19590 19990 21490 24990 24990 20900 21896 26899
31990 ...
```

```
 $ old_price     : int  32790 21325 27990 27990 33490 33490 22825 0 27668 36990 ...
 $ discount      : int  23 8 28 23 25 25 8 0 2 13 ...
 $ star_rating   : num  3.7 3.6 3.7 3.7 3.7 3.7 3.9 3.9 0 4.2 ...
 $ ratings       : int  63 1894 1153 1153 1657 1657 1185 219 0 76 ...
 $ reviews       : int  12 256 159 159 234 234 141 18 0 13 ...
```

Basic Operations

##Loading the dataset

```
Lap<-read.csv('C:/Users/G Sirisha/OneDrive/Desktop/DSST/Laptop.csv')
```

#No.of rows

```
> nrow(Lap)
[1] 896
```

#No.of columns

```
> ncol(Lap)
[1] 23
```

#Colnames

```
> colnames(Lap)
 [1] "brand"          "model"          "processor_brand"
 [4] "processor_name"  "processor_gnrtn" "ram_gb"
 [7] "ram_type"       "ssd"            "hdd"
[10] "os"             "os_bit"         "graphic_card_gb"
[13] "weight"         "display_size"    "warranty"
[16] "Touchscreen"     "msoffice"        "latest_price"
[19] "old_price"       "discount"        "star_rating"
```

#Summary

```
> summary(Lap)
    brand           model          processor_brand
 Length:896        Length:896        Length:896
```

```
Class :character   Class :character   Class :character  
Mode  :character   Mode  :character   Mode  :character  



processor_name     processor_gnrtn      ram_gb         
Length:896         Length:896         Length:896        
Class :character   Class :character   Class :character  
Mode  :character   Mode  :character   Mode  :character  



  ram_type             ssd                hdd           
Length:896         Length:896         Length:896        
Class :character   Class :character   Class :character  
Mode  :character   Mode  :character   Mode  :character  



    os                os_bit          graphic_card_gb
Length:896         Length:896         Min.   :0.000  
Class :character   Class :character   1st Qu.:0.000  
Mode  :character   Mode  :character   Median :0.000  
                                      Mean   :1.199  
                                      3rd Qu.:2.000  
                                      Max.   :8.000  
   weight          display_size         warranty    
Length:896         Length:896         Min.   :0.000  
Class :character   Class :character   1st Qu.:0.000  
Mode  :character   Mode  :character   Median :1.000  
```

```
                          Mean   :0.692
                          3rd Qu.:1.000
                          Max.   :3.000
 Touchscreen        msoffice        latest_price
 Length:896        Length:896        Min.   : 13990
 Class :character   Class :character   1st Qu.: 45490
 Mode  :character   Mode  :character   Median : 63494
                          Mean   : 76310
                          3rd Qu.: 89090
                          Max.   :441990
   old_price       discount      star_rating     ratings
 Min.   :     0   Min.   : 0.00   Min.   :0.00   Min.   :     0.0
 1st Qu.: 54941   1st Qu.:11.00   1st Qu.:0.00   1st Qu.:     0.0
 Median : 78053   Median :19.00   Median :4.10   Median :    19.0
 Mean   : 88134   Mean   :18.53   Mean   :2.98   Mean   :  367.4
 3rd Qu.:111020   3rd Qu.:26.00   3rd Qu.:4.40   3rd Qu.:  179.5
 Max.   :377798   Max.   :57.00   Max.   :5.00   Max.   :15279.0
    reviews
 Min.   :   0.00
 1st Qu.:   0.00
 Median :   3.00
 Mean   :  46.15
 3rd Qu.:  23.25
 Max.   :1947.00
```

#head(): it returns the first few rows to understand it's structure

```
> head(Lap)
   brand   model processor_brand    processor_name processor_gnrtn
1 Lenovo A6-9225           AMD A6-9225 Processor           10th
2 Lenovo Ideapad           AMD        APU Dual           10th
3  Avita    PURA           AMD        APU Dual           10th
```

```
4  Avita    PURA          AMD        APU Dual          10th
5  Avita    PURA          AMD        APU Dual          10th
6  Avita    PURA          AMD        APU Dual          10th
  ram_gb ram_type    ssd     hdd      os os_bit graphic_card_gb
1 4 GB GB     DDR4   0 GB 1024 GB Windows 64-bit               0
2 4 GB GB     DDR4   0 GB  512 GB Windows 64-bit               0
3 4 GB GB     DDR4 128 GB    0 GB Windows 64-bit               0
4 4 GB GB     DDR4 128 GB    0 GB Windows 64-bit               0
5 4 GB GB     DDR4 256 GB    0 GB Windows 64-bit               0
6 8 GB GB     DDR4 256 GB    0 GB Windows 64-bit               0
     weight display_size warranty Touchscreen msoffice latest_price
1 ThinNlight      Missing        0          No       No        24990
2     Casual      Missing        0          No       No        19590
3 ThinNlight      Missing        0          No       No        19990
4 ThinNlight      Missing        0          No       No        21490
5 ThinNlight      Missing        0          No       No        24990
6 ThinNlight           14        0          No       No        24990
  old_price discount star_rating ratings reviews
1     32790       23         3.7      63      12
2     21325        8         3.6    1894     256
3     27990       28         3.7    1153     159
4     27990       23         3.7    1153     159
5     33490       25         3.7    1657     234
6     33490       25         3.7    1657     234
```

Data Cleaning

#Checking NA values in a dataset

```
> is.null(Lap)
[1] FALSE
```

```
> colSums(is.na(Lap))
        brand          model processor_brand
processor_name
           0             0             0             0
processor_gnrtn         ram_gb       ram_type           ssd
           0             0             0             0
          hdd            os         os_bit graphic_card_gb
           0             0             0             0
        weight    display_size      warranty    Touchscreen
           0             0             0             0
      msoffice    latest_price      old_price       discount
           0             0             0             0
    star_rating        ratings        reviews
           0             0             0
```

Here, there is no NA values in the dataset

But the null values are displayed as “Missing”.

#max: it is used to find the max in the columns. If there is any Missing values, it returns as “Missing”.

```
> max(Lap$discount)
[1] 57
> max(Lap$old_price)
[1] 377798
> max(Lap$latest_price)
[1] 441990
> max(Lap$display_size)
[1] "Missing"
```

Now, we replacing the string “Misssing” with NA

```
Lap[Lap=='Missing']<-NA
View(Lap)
```

#Checking NA values are inserted or not

```
> colSums(is.na(Lap))
        brand          model processor_brand  processor_name
            0             95               0               0
processor_gnrtn         ram_gb        ram_type             ssd
          239              0               0               0
          hdd             os          os_bit graphic_card_gb
            0              0               0               0
       weight   display_size        warranty     Touchscreen
            0            332               0               0
     msoffice   latest_price       old_price        discount
            0              0               0               0
  star_rating        ratings         reviews
            0              0               0
```

There are NA values in Three Columns – **model, display_size, processor_gnrtn**

Firstly, consider the column model, in this we don't about the brand which is missed so we play the **string "None"** in the Missing place

#Model Column

```
> class(Lap$model)
[1] "character"
> Lap$model
  [1] "A6-9225"     "Ideapad"     "PURA"
  [4] "PURA"        "PURA"        "PURA"
  [7] "APU"         "APU"         "Athlon"
 [10] "Aspire"      "ExpertBook"  NA
 [13] "v15"         "ExpertBook"  "VivoBook"
 [16] "EeeBook"     "EeeBook"     "ExpertBook"
```

#Replacing NA values by using the string “None”

```
> Lap$model[is.na(Lap$model)]<-'None'
> Lap$model
  [1] "A6-9225"    "Ideapad"    "PURA"
  [4] "PURA"       "PURA"       "PURA"
  [7] "APU"        "APU"        "Athlon"
 [10] "Aspire"     "ExpertBook"  "None"
 [13] "v15"        "ExpertBook"  "VivoBook"
 [16] "EeeBook"    "EeeBook"    "ExpertBook"
 [19] "None"       "Aspire"     "Nitro"
> table(is.na(Lap$model))
FALSE
  896
```

Now, we successfully inserted the None in the place of NA values

#Display_Size

```
> class(Lap$display_size)
[1] "character"
> Lap$display_size
  [1] NA     NA     NA     NA     NA     "14"
  [7] "14"   NA     "14"   NA     "15.6"  NA
 [13] "15.6"  NA     NA     NA     NA     "14"
 [19] "14"   NA     NA     NA     NA     NA
```

#checking null values

```
 > table(is.na(Lap$display_size))
 FALSE  TRUE
   564   332
```

#converting the class of display_size into “numeric”

```
> Lap$display_size<-as.numeric(Lap$display_size)
> class(Lap$display_size)
[1] "numeric"
```

#mean

```
> display_mean=mean(Lap$display_size, na.rm=TRUE)
> display_mean
[1] 15.12202
```

#Replacing the null values with mean values

```
> Lap$display_size[is.na(Lap$display_size)]<-display_mean
> Lap$display_size
  [1] 15.12202 15.12202 15.12202 15.12202 15.12202
  [6] 14.00000 14.00000 15.12202 14.00000 15.12202
 [11] 15.60000 15.12202 15.60000 15.12202 15.12202
```

#checking na values are replaced or not

```
> table(is.na(Lap$display_size))
FALSE
  896
```

#Processor_generatipon

##Checking null values

```
> class(Lap$processor_gnrtn)
[1] "character"
> Lap$processor_gnrtn
  [1] "10th" "10th" "10th" "10th" "10th" "10th" "10th"
  [8] "10th" "10th" "10th" "10th" "10th" "10th" "10th"
 [15] NA     NA     NA     "10th" "10th" "10th" "10th"
 [22] "10th" "10th" NA     NA
```

#Here tha generation contains 7th, 8th, 11th, 12th. So I’m replacing the missing the values with 12th generation

```
> Lap$processor_gnrtn[is.na(Lap$processor_gnrtn)]<-'12th'
> Lap$processor_gnrtn
  [1] "12th" "12th" "12th" "12th" "12th" "12th" "12th"
  [8] "12th" "12th" "12th" "12th" "12th" "12th" "12th"
 [15] "12th" "12th" "12th" "12th" "12th" "12th" "12th"
 [22] "12th" "12th" "12th"
```

#checking null values are replaced or not

```
> table(is.na(Lap$processor_gnrtn))
FALSE
  896
```

We removed the NA values successfully. So, the Dataset is very Clean. There is no inconsistent data anymore.

#Changing the datatypes of some columns

```
> class(Lap$Touchscreen)
[1] "character"
> #table
> table(Lap$Touchscreen)
 No Yes
793 103
> Lap$Touchscreen<-as.factor(Lap$Touchscreen)
> class(Lap$Touchscreen)
[1] "factor"
> class(Lap$msoffice)
[1] "character"
> #table
> table(Lap$msoffice)
```

```
 No Yes
606 290
> Lap$msoffice<-as.factor(Lap$msoffice)
> #class
> class(Lap$msoffice)
[1] "factor"
```

Data Analysis

Data Analysis is the process of examining, cleaning, and interpreting data to extract usefyl information and make informed decisions.

#Considering any one of the column and performing the mathematical operations like mean, median, var, range, IQR …

```
#latest_price column
```

#min – it gives the min value

```
> min(Lap$latest_price)
[1] 13990
```

#mean

```
> mean(Lap$latest_price)
[1] 76309.86
```

#median – middle most value

```
> median(Lap$latest_price)
[1] 63494
```

#range – it gives the min and max values

```
> range(Lap$latest_price)
[1]  13990 441990
```

#Inter quartile range

```
> IQR(Lap$latest_price)
[1] 43600
```

```
> var(Lap$latest_price)
[1] 2172804805
```

#summary

```
> summary(Lap)
    brand              model           processor_brand
 Length:896         Length:896         Length:896
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character



 processor_name     processor_gnrtn       ram_gb
 Length:896         Length:896         Length:896
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character



   ram_type             ssd                hdd
 Length:896         Length:896         Length:896
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character



      os               os_bit          graphic_card_gb
 Length:896         Length:896         Min.   :0.000
 Class :character   Class :character   1st Qu.:0.000
 Mode  :character   Mode  :character   Median :0.000
                                       Mean   :1.199
```

```
                              3rd Qu.:2.000
                              Max.   :8.000
   weight          display_size     warranty
 Length:896        Min.   :12.20   Min.   :0.000
 Class :character   1st Qu.:15.12   1st Qu.:0.000
 Mode  :character   Median :15.12   Median :1.000
                    Mean   :15.12   Mean   :0.692
                    3rd Qu.:15.60   3rd Qu.:1.000
                    Max.   :17.30   Max.   :3.000
```

#Considering the brand , model, processor_gnrtn, it returns the 12th generation, Ryzen model for every brand

```
> Lap%>% select(brand,processor_gnrtn,model)%>%
    filter(model=="Ryzen"&processor_gnrtn=="12th")->Ryzen
> View(Ryzen)
> Ryzen
   brand processor_gnrtn model
1     HP          12th Ryzen
2   ASUS          12th Ryzen
3   DELL          12th Ryzen
4   ASUS          12th Ryzen
5     HP          12th Ryzen
6     HP          12th Ryzen
```

#returns the Dell brand having Windows Operatinf system

```
> Lap%>%filter(brand=="DELL"&os=="Windows")->Dell_os
> Dell_os
  brand              os
   1   DELL       Windows
   2   DELL       Windows
   3   DELL       Windows
   4   DELL       Windows
  5   DELL        Windows
```

```
> ###all brands
> Lap%>%select(brand,ratings)%>%
+   group_by(brand)%>%
+   summarise(total=sum(ratings))->tol_ratings
> tol_ratings
# A tibble: 21 × 2
   brand     total
   <chr>     <int>
 1 ALIENWARE    40
 2 APPLE     33811
 3 ASUS      93386
 4 Avita      6998
 5 DELL      18904
 6 HP        75896
 7 Infinix    2882
 8 LG           83
 9 Lenovo    25437
10 MICROSOFT   114
# ℹ 11 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Data Manipulation

Data Manipulation involves adjusting or organising data to make it more useful for analysis like select, filter, renames, arrange …….

#select

```
> select(Lap,2,4)
         model    processor_name
```

```
1     A6-9225 A6-9225 Processor
2     Ideapad         APU Dual
3        PURA        APU Dual
4        PURA        APU Dual
5        PURA        APU Dual
6        PURA        APU Dual
```

#starts_with : returns the columns which starts with “r”

```
> Lap%>%select(starts_with("r"))
     ram_gb ram_type ratings reviews
1    4 GB GB     DDR4      63      12
2    4 GB GB     DDR4    1894     256
3    4 GB GB     DDR4    1153     159
4    4 GB GB     DDR4    1153     159
5    4 GB GB     DDR4    1657     234
```

#ends_with : returns the columns which ends with “e”

```
> Lap%>%select(ends_with("e"))
      processor_name ram_type display_size msoffice
1   A6-9225 Processor     DDR4     15.12202       No
2          APU Dual     DDR4     15.12202       No
3          APU Dual     DDR4     15.12202       No
4          APU Dual     DDR4     15.12202       No
```

#filter

```
 > Lap%>%filter(Touchscreen=="Yes" &
 +                latest_price<=50000)
     brand    model processor_brand
 processor_name
 1     DELL Inspiron         Intel       Core i3
```

```
2 Smartron   t.book          Intel       Core m3

  processor_gnrtn  ram_gb ram_type    ssd
hdd      os

1          12th 8 GB GB     DDR4 256 GB 0 GB
Windows

2          12th 4 GB GB     DDR3 128 GB 0 GB
Windows

  os_bit graphic_card_gb weight display_size
warranty

1 64-bit             0 Casual        14.0       1

2 64-bit             0 Casual        12.2       0

  Touchscreen msoffice latest_price old_price
discount

1        Yes      No       49990     68658      27

2        Yes      No       42990     45990       6

  star_rating ratings reviews

1        4.2      24       5

2        3.5      25       6

>


>Lap%>%filter(Touchscreen=="Yes"&star_rating>=4.0)

>Lap%>%filter(brand=="HP"&ratings>1000)->b

>b


> Lap%>%filter(brand=="DELL"&os=="Windows")->Dell_os

> Dell_os

  brand                os

   1   DELL        Windows
   2   DELL        Windows
   3   DELL        Windows
   4   DELL        Windows
   5   DELL        Windows
```

#rename

```
> Lap<-rename(Lap, "brand_name"="brand")
> Lap
  brand_name     model processor_brand
1     Lenovo   A6-9225           AMD
2     Lenovo   Ideapad           AMD
3      Avita      PURA           AMD
4      Avita      PURA           AMD
5      Avita      PURA           AMD
6      Avita      PURA           AMD
```

#arrange

```
> Lap%>%select(brand_name,star_rating)%>%
+   arrange(star_rating)
   brand_name star_rating
1          HP         0.0
2        ASUS         0.0
3        ASUS         0.0
4        ASUS         0.0
5        ASUS         0.0
6      Lenovo         0.0
> Lap%>%select(brand_name, latest_price, old_price)%>%
+   arrange(latest_price, old_price)
   brand_name latest_price old_price
1       iball        13990     19999
2      Lenovo        16990     24840
3       Avita        17490     23490
4        ASUS        17990     21990
5        ASUS        18990     22990
6      Lenovo        19590     21325
```

Data Visualization

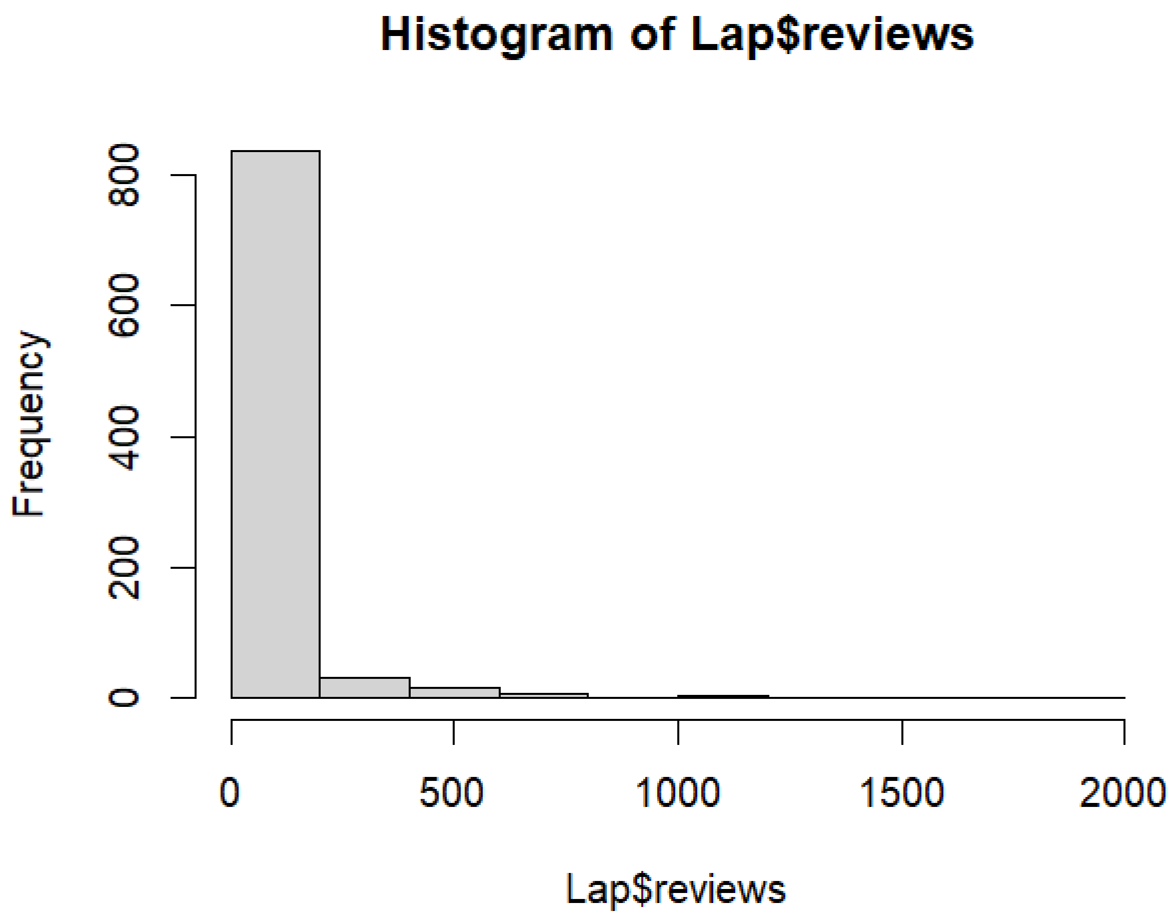
Data Visualization is the process of creating visual representations of data to make it easier to understand and analyze.

library(ggplot2)

#Histogram

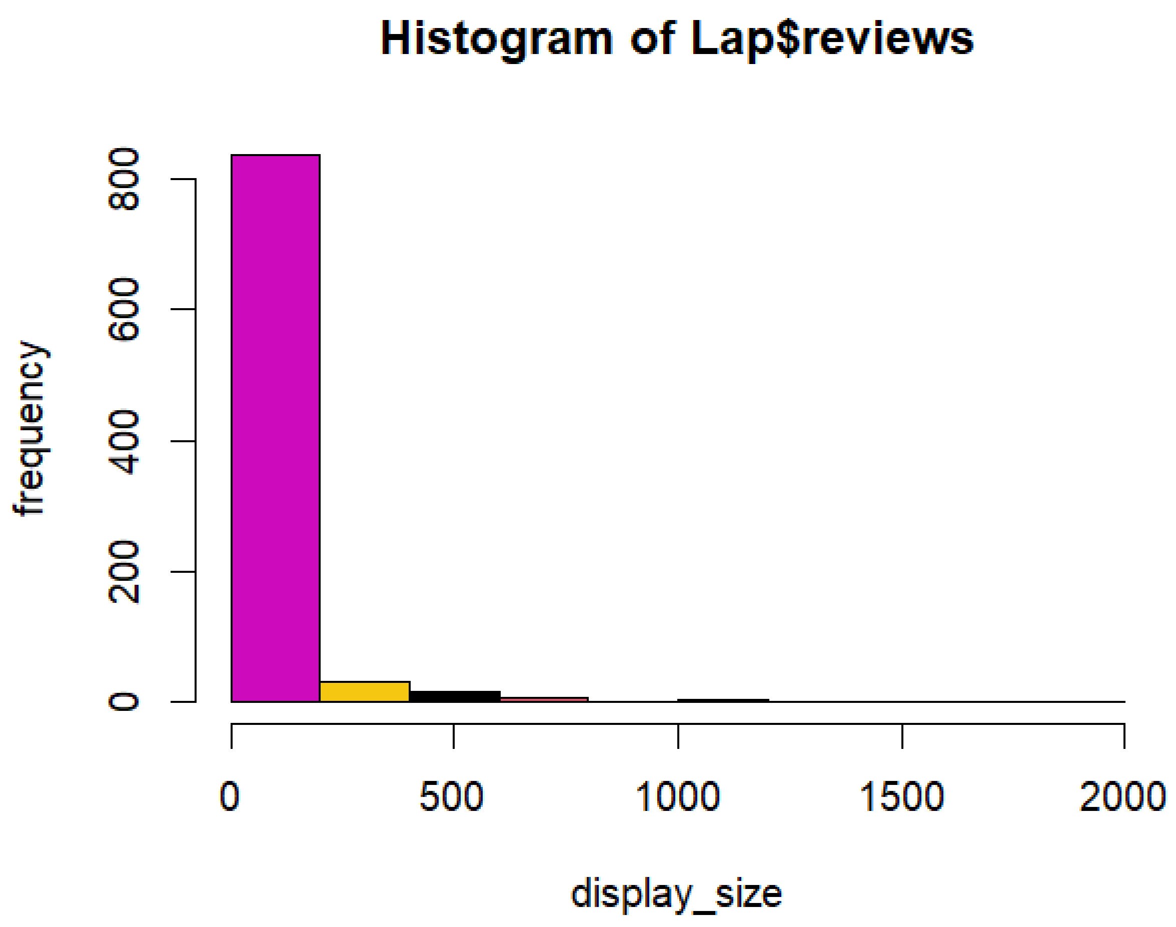
##without libray

hist(Lap$reviews)



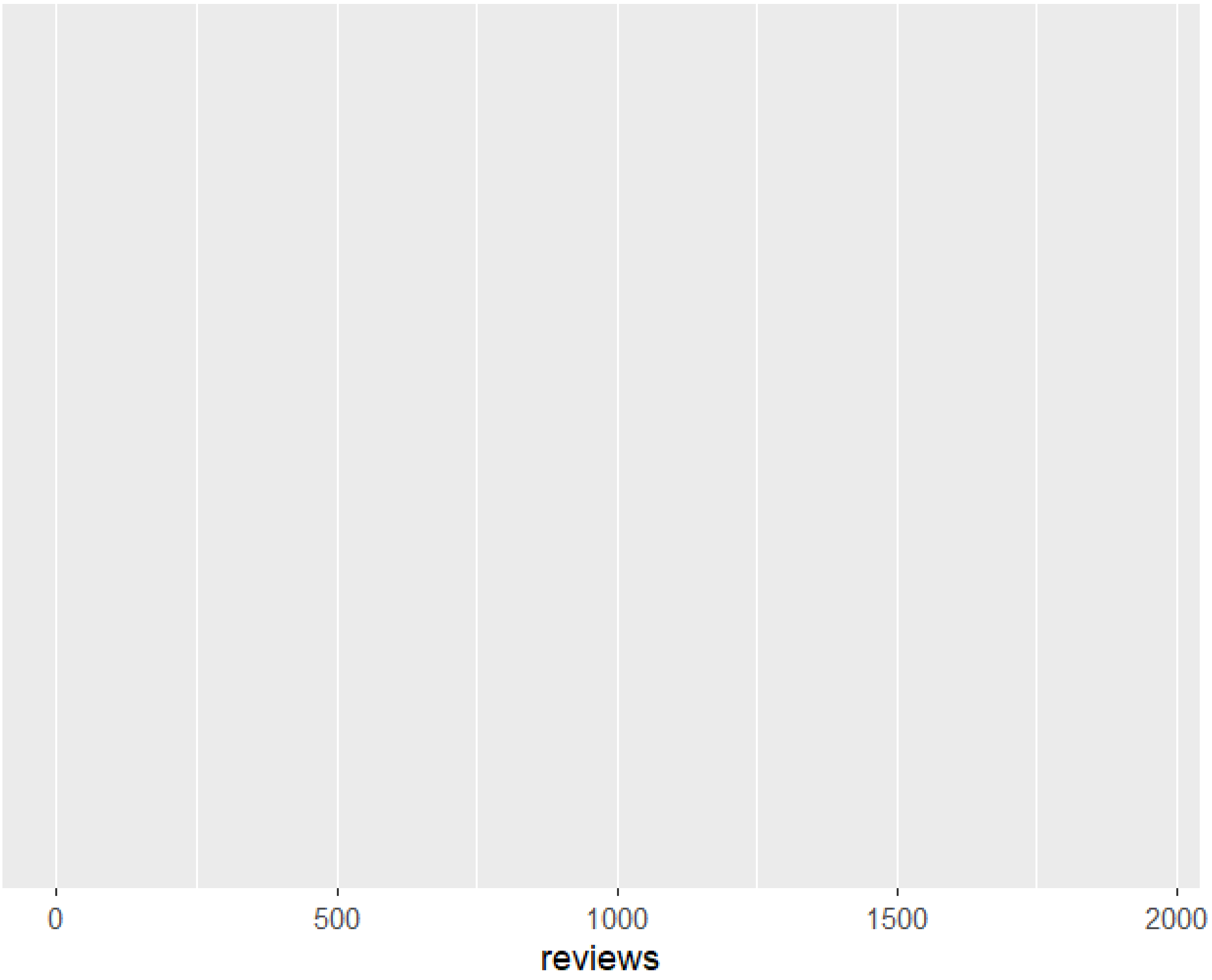
##adding x and y axis and color

```
hist(Lap$reviews,xlab = "display_size",
    ylab = "frequency",col = c(6,7,9,2,4,5))
```



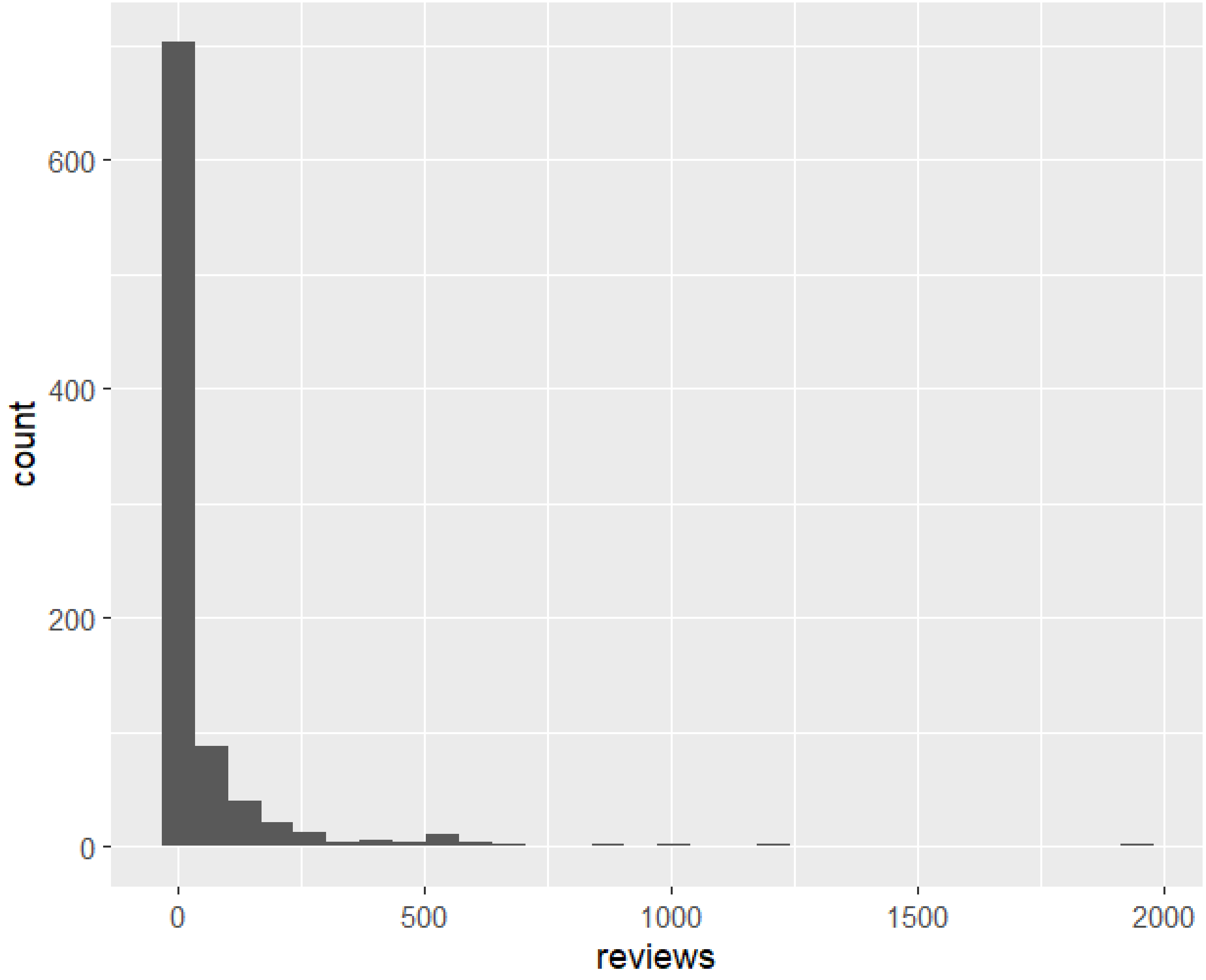
#histogram

```
ggplot(data=Lap,aes(x=reviews))
```



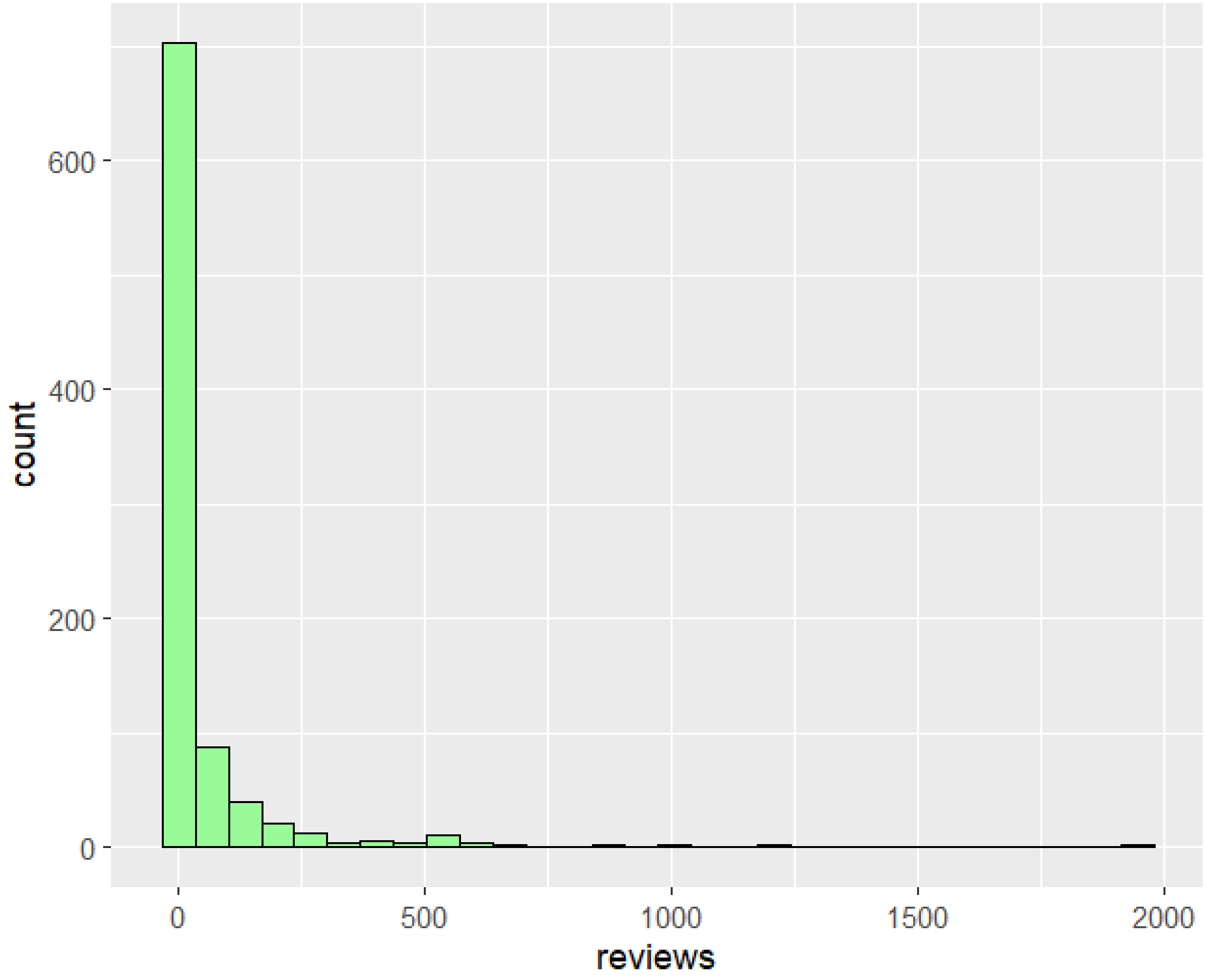
#adding geometric

```
ggplot(data=Lap,aes(x=reviews))+
  geom_histogram()
```

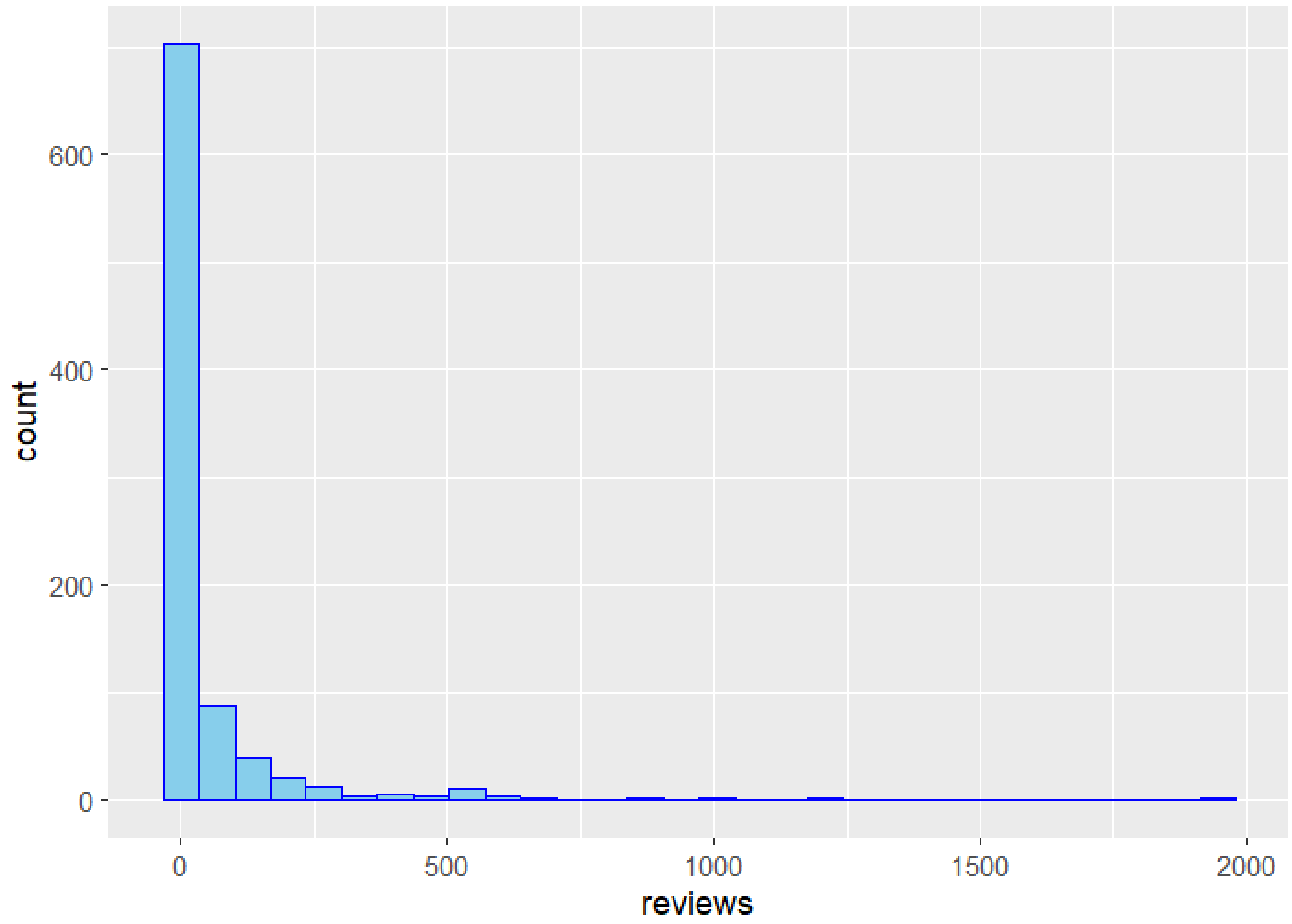


##adding color and border

```
ggplot(data=Lap,aes(x=reviews))+
  geom_histogram(fill="palegreen",col="black")
```

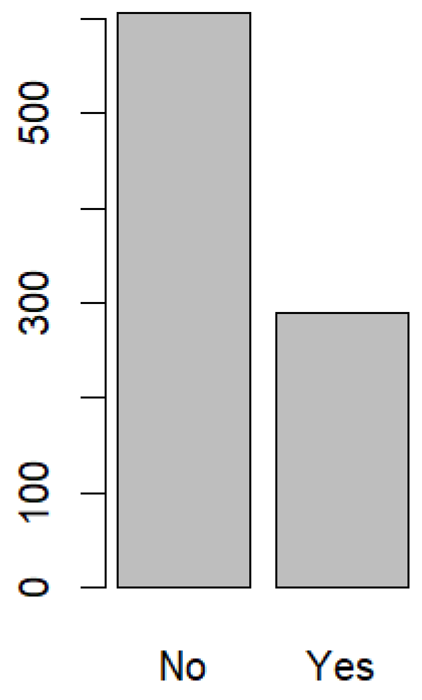


```
ggplot(data=Lap,aes(x=reviews))+
geom_histogram(fill="skyblue",col="blue")
```

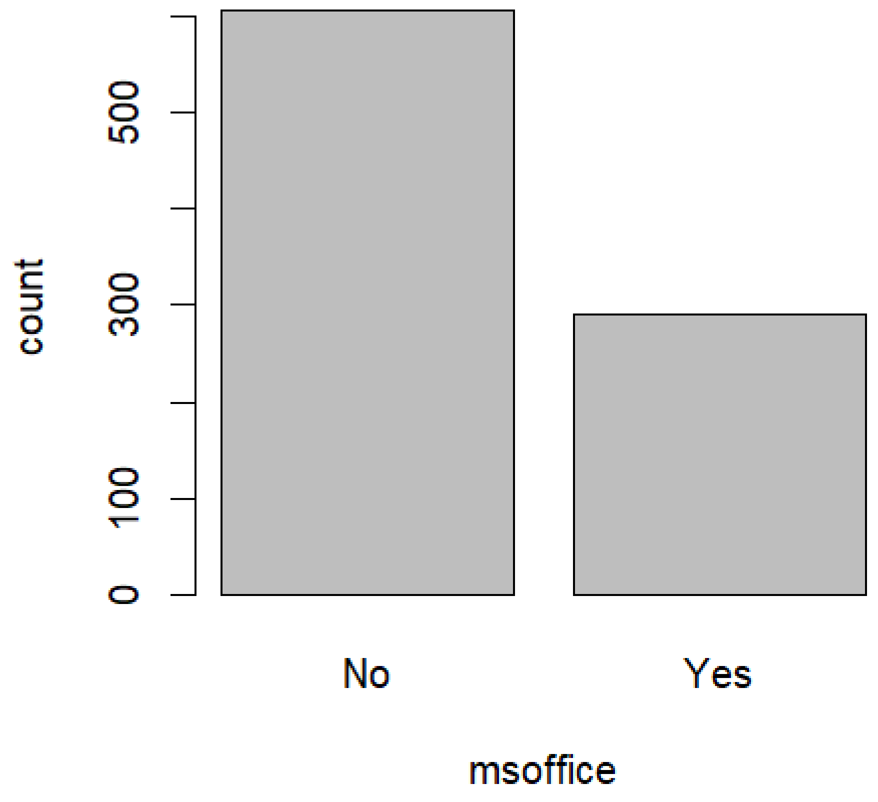


#Barplot

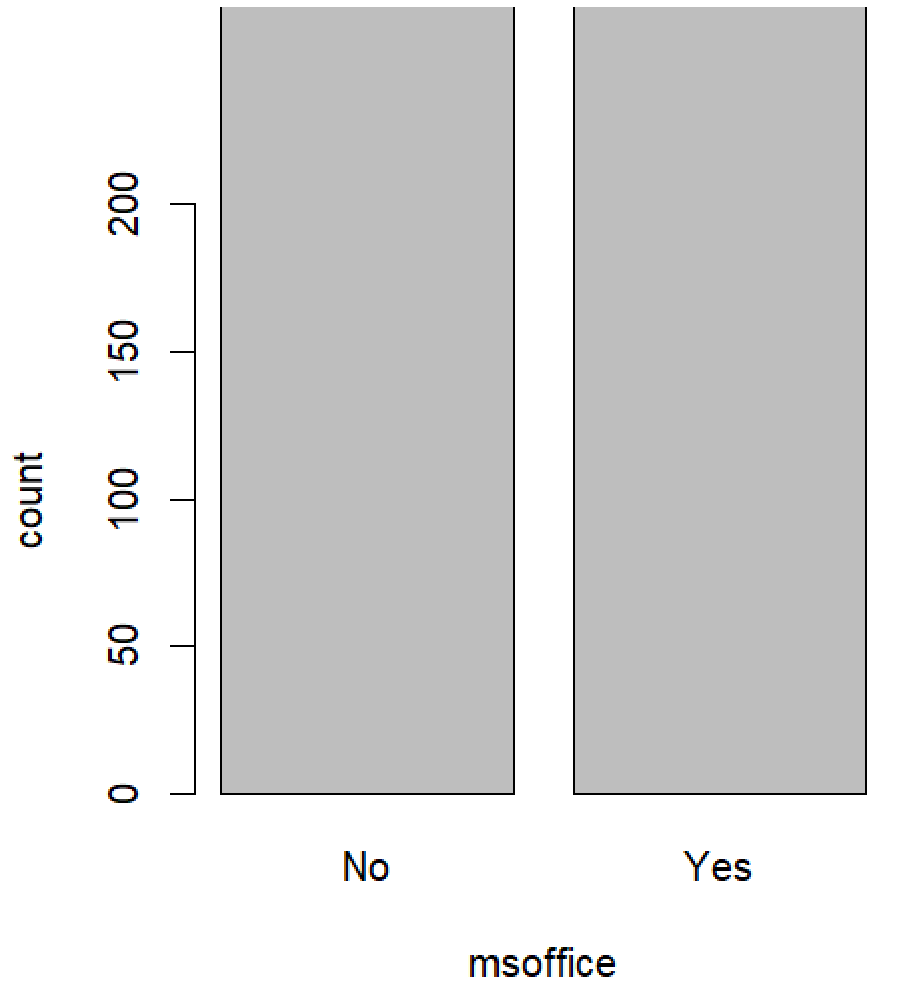
```
> a<-table(Lap$msoffice)

> a

 No Yes

606 290

> barplot(a)
```



```
> barplot(a,xlab = "msoffice",ylab = "count")
```

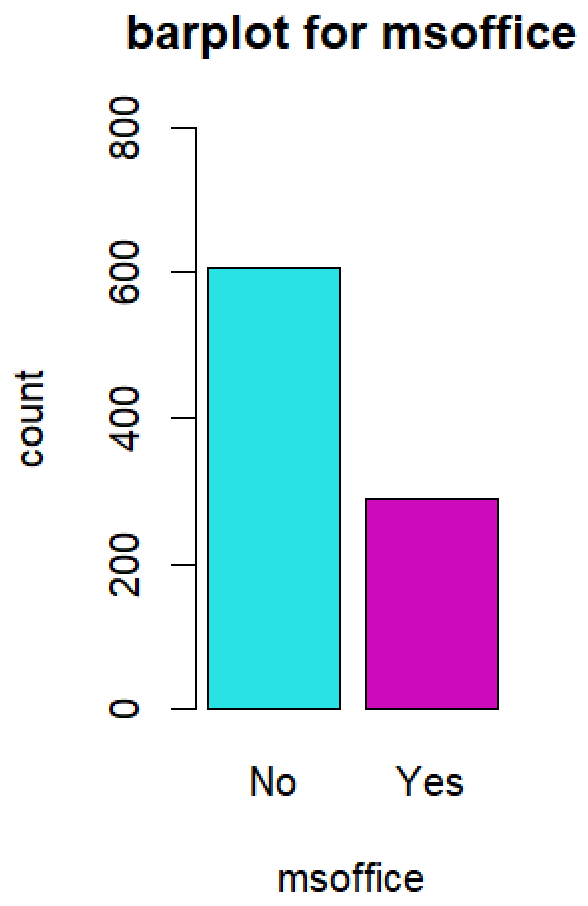


```
> barplot(a,xlab = "msoffice",ylab = "count",
+       ylim=c(0,200))
```



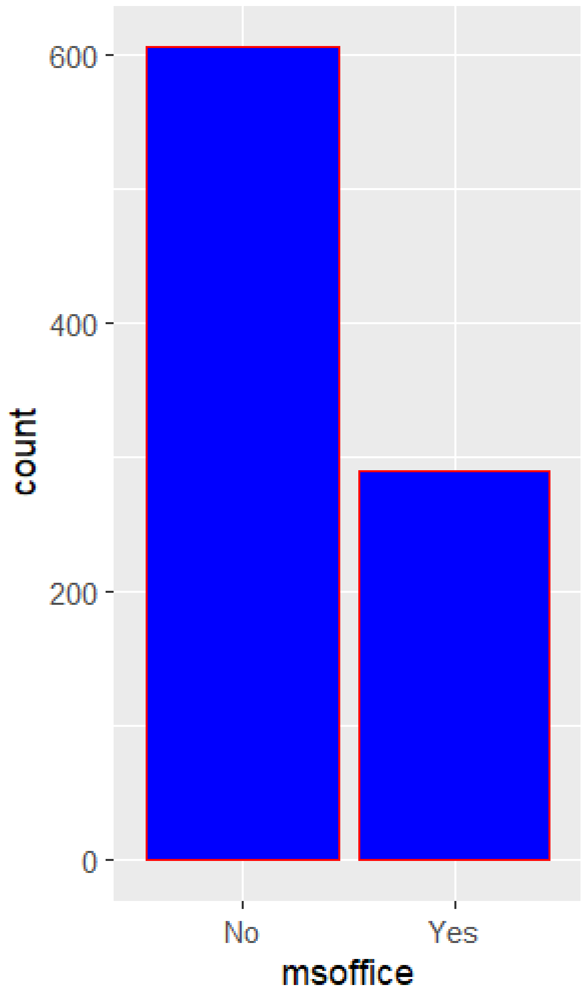
#adding color

```
> barplot(a,xlab = "msoffice",ylab = "count",
+       ylim=c(0,800),col=c(5,6),
+       main="barplot for msoffice")
```



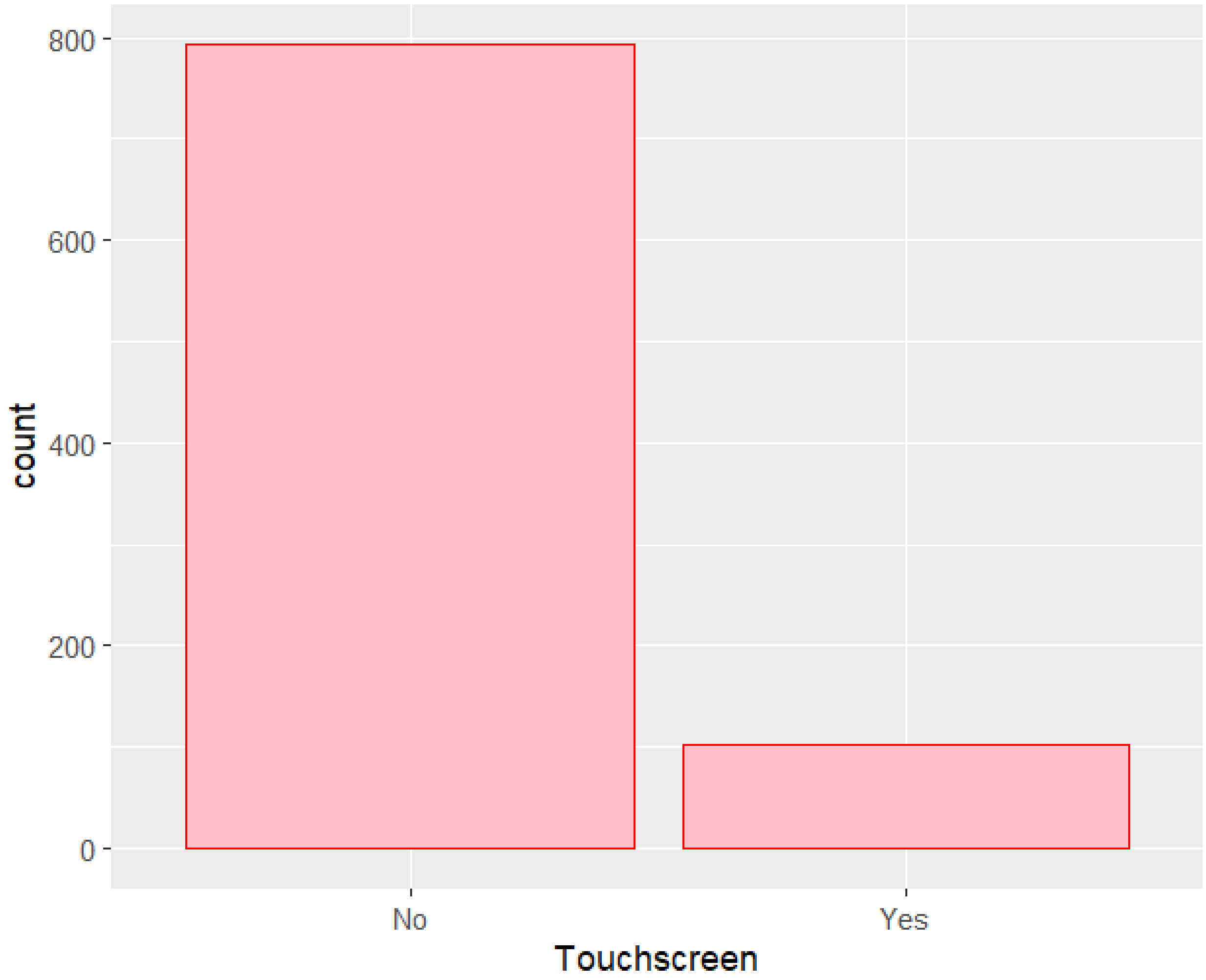
#geom_bar()

```
ggplot(data=Lap,aes(x=msoffice))+
+   geom_bar(fill=”blue”, color=”red”)
```



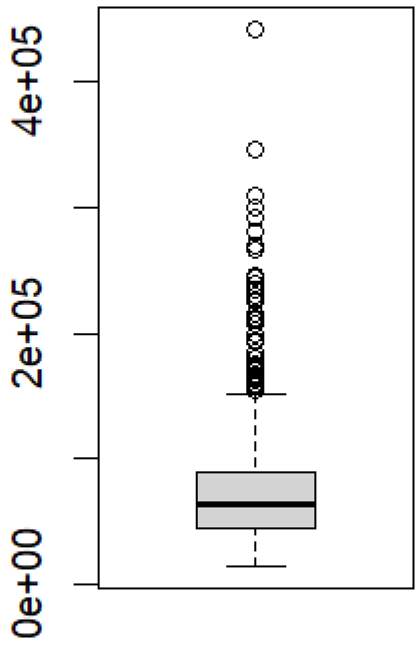
#adding color

```
ggplot(data=Lap,aes(x=Touchscreen))+
  geom_bar(fill="pink", color="red")
```

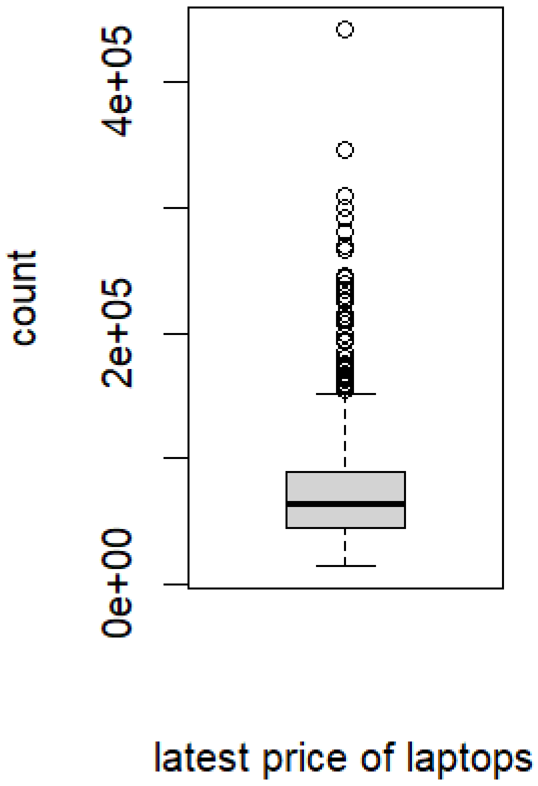


#Boxplot

```
> boxplot(Lap$latest_price)
```

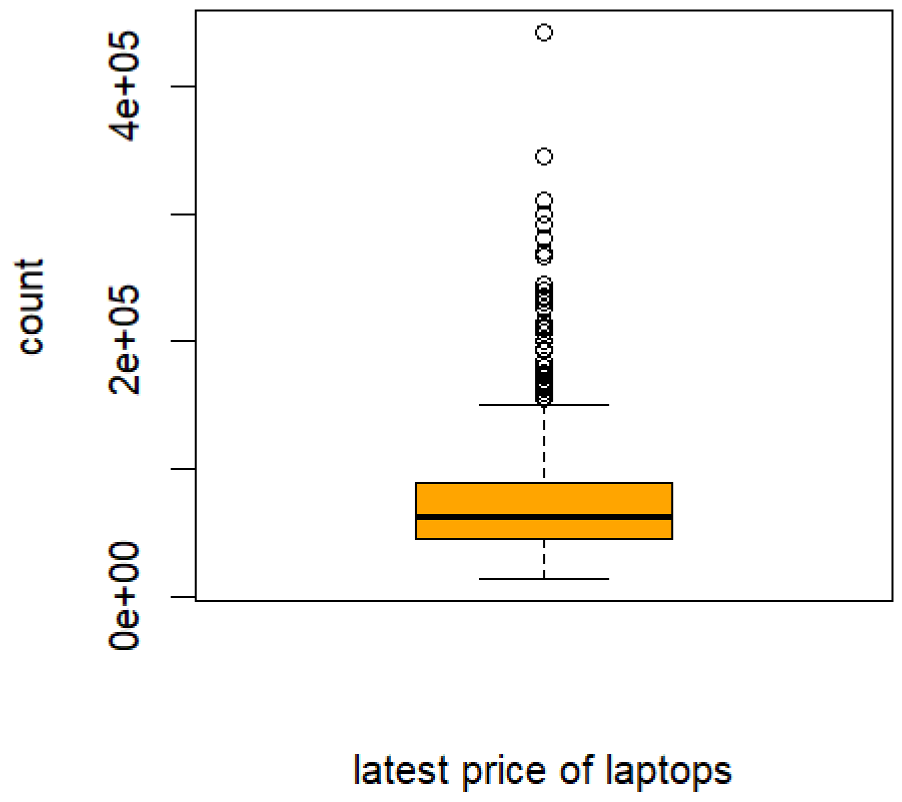


```
> boxplot(Lap$latest_price,
+       xlab="latest price of laptops",
+       ylab="count")
```



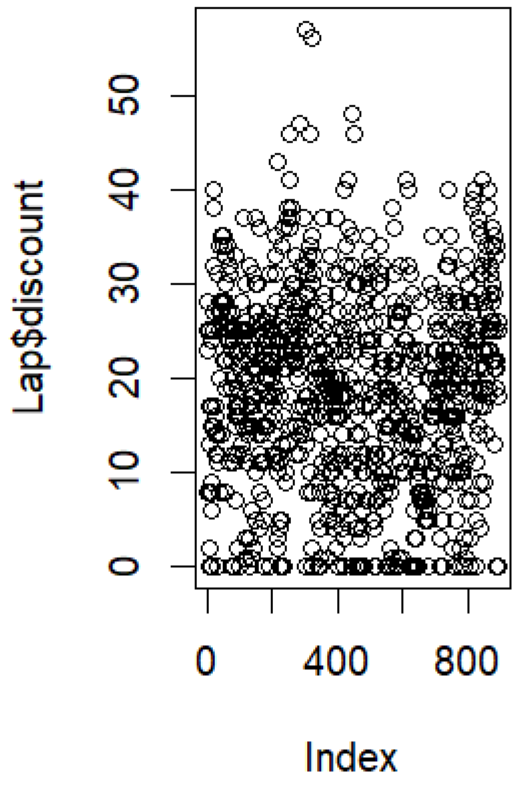
#adding color

```
boxplot(Lap$latest_price,
      xlab="latest price of laptops",
ylab="count", col="orange")
```

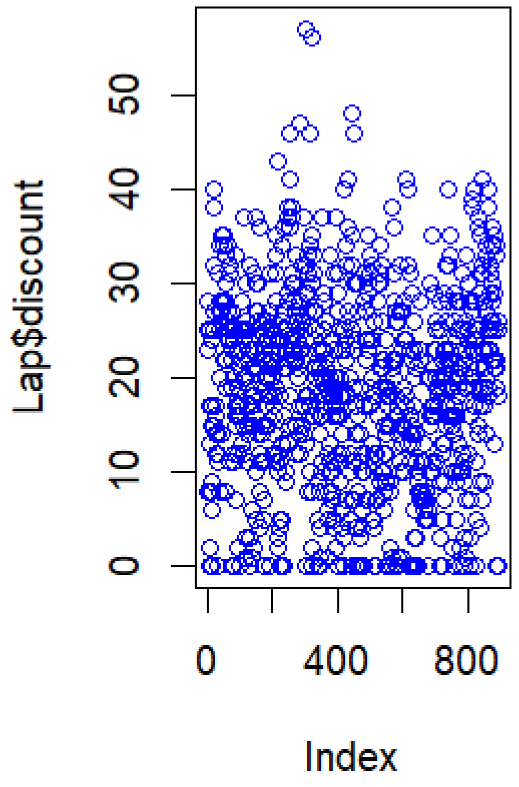


#Scatterplot

```
> plot(Lap$discount)
```

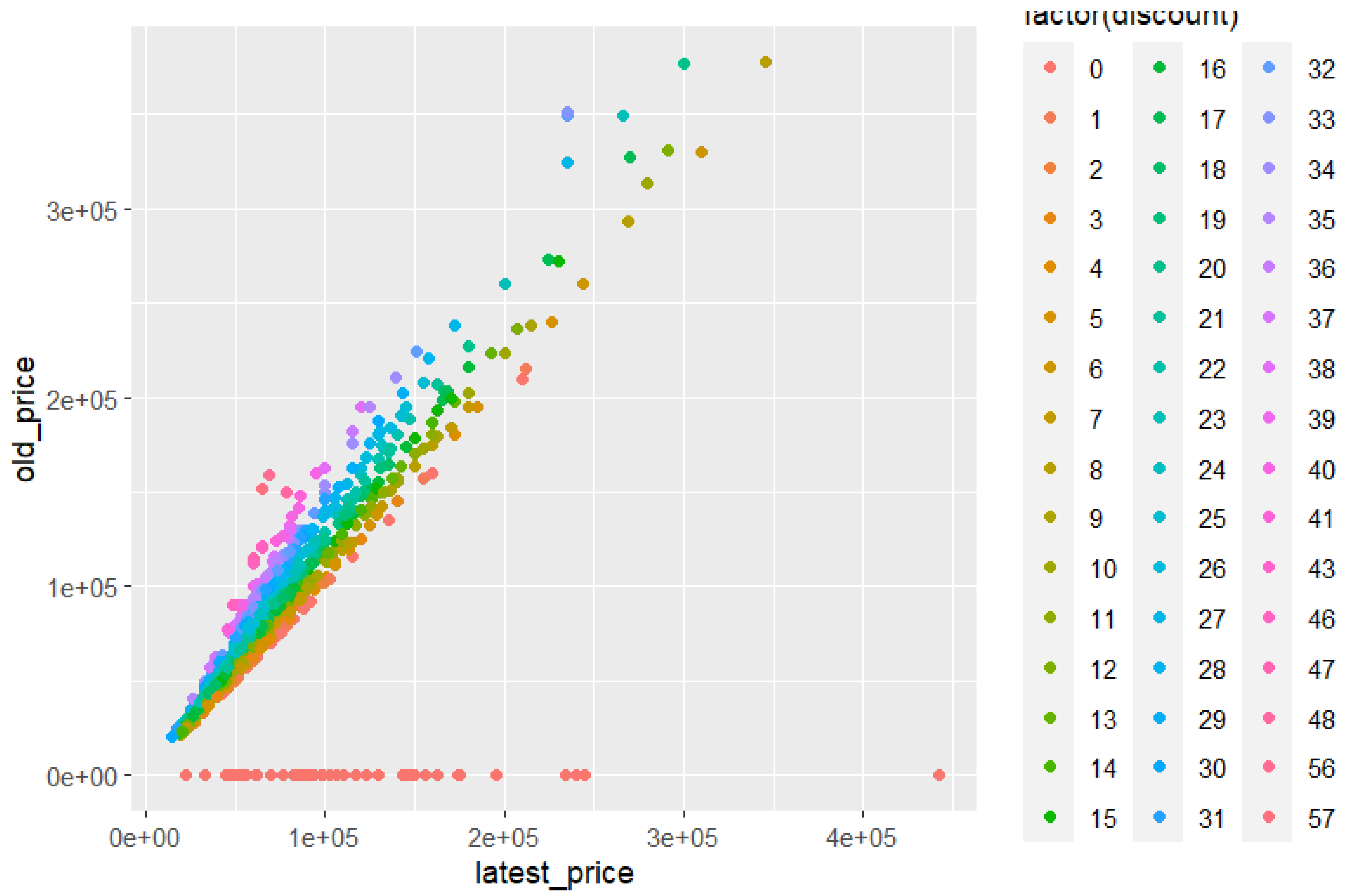


```
#adding color
plot(Lap$discount, col="blue")
```

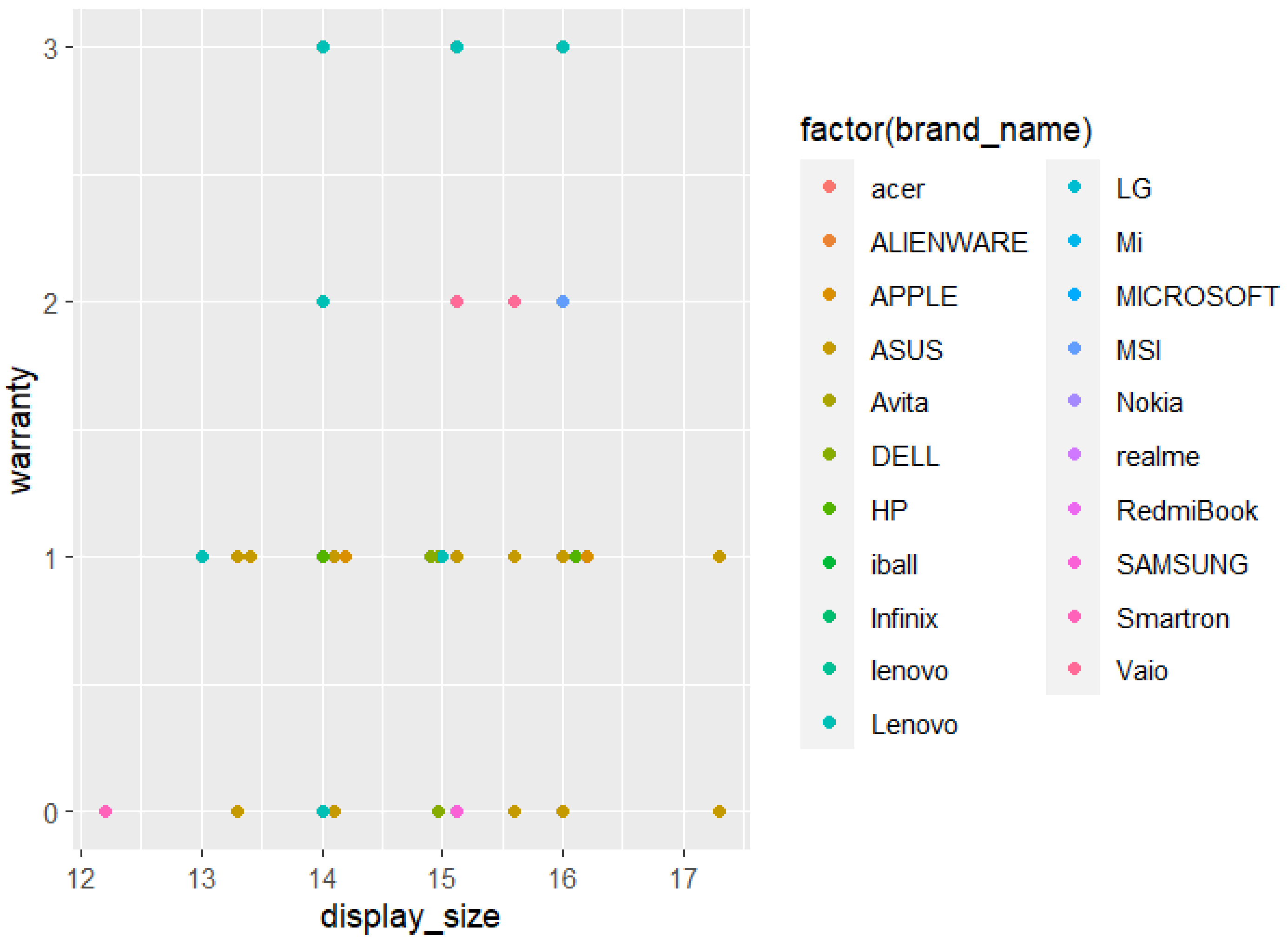


#using library

```
> ggplot(data=Lap,aes(x=latest_price,y=old_price,
+                  col=factor(discount)),pch=3)+
+   geom_point()
```



```
> ggplot(data=Lap,aes(x=display_size,y=warranty,
+                  col=factor(brand_name)),pch=3)+
+   geom_point()
```



Thank you Sir,

Name : G.Sirisha

Group : III Data Science

Branch : ADCTUNI.