Univariate Exploratory Data Analysis

Vinaykumar Pandey

08/04/2021

#Include the library file

```
library(dplyr)
```

```
## Warning: package ’dplyr’ was built under R version 4.0.3
```

```
##
## 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
```

```
library(ggplot2)
```

```
## Warning: package ’ggplot2’ was built under R version 4.0.4
```

```
library(tidyr)
```

```
## Warning: package ’tidyr’ was built under R version 4.0.3
```

```
library(plotrix)
```

```
## Warning: package ’plotrix’ was built under R version 4.0.3
```

#Read the dataset

```
export_data <- read.csv("india_export.csv")
```

#Storing file in data frame

```
export_data <- data.frame(export_data)
```

Display all the columns of the dataset

```
colnames(export_data)
```

```
## [1] "Year"
## [2] "Consumer.Electronics"
## [3] "Industrial.Electronics"
## [4] "Computer.Hardware"
## [5] "Communication...Broadcast.Equipments"
## [6] "Electronics.Components"
## [7] "Sub.Total"
## [8] "Computer.Software"
## [9] "Total"
```

#Display structure of the dataset

```
str(export_data)
```

```
## 'data.frame':    17 obs. of  9 variables:
##  $ Year                                : chr  "2000-01" "2001-02" "2002-03" "2003-04" ...
##  $ Consumer.Electronics                : int  648 700 750 825 1150 2000 1500 1600 2600 3000 ...
##  $ Industrial.Electronics              : int  500 950 1400 1515 1500 2300 3000 3885 4200 3500 ...
##  $ Computer.Hardware                   : int  1250 1800 550 1440 1200 1025 1500 990 1650 1900 ...
##  $ Communication...Broadcast.Equipments: int  550 150 500 165 350 500 650 625 12280 7800 ...
##  $ Electronics.Components              : int  1840 2200 2400 3755 3800 3800 5850 6100 10500 9700 ...
##  $ Sub.Total                           : int  4788 5800 5600 7700 8000 9625 12500 13200 31230 25900
##  $ Computer.Software                   : int  28350 36500 46100 58240 80180 104100 141000 164400 216
##  $ Total                               : int  33138 42300 51700 65940 88180 113725 153500 177600 247
```

#Grouping the NA values

```
export_data<-data.frame(lapply(export_data,function(x) {gsub("#N/A",NA,x)}))
```

Total NA values

```
total_NA<-sapply(export_data,function(x) sum(is.na(x)==TRUE))

total_NA
```

```
##                                 Year                 Consumer.Electronics
##                                    0                                    4
##               Industrial.Electronics                    Computer.Hardware
##                                    4                                    4
## Communication...Broadcast.Equipments               Electronics.Components 
##                                    4                                    4 
##                            Sub.Total                    Computer.Software 
##                                    2                                    0 
##                                Total 
##                                    0
```

#finding the column with Null values percentge

```
for(i in 1:ncol(export_data)) {
  colum_name <- colnames(export_data[i])
  Null_perc <- sum(is.na(export_data[,i]))/length(export_data[,i])
  if (Null_perc > 0.05) {
    print(paste("Column ", colum_name, " has ", round(Null_perc*100, 3), "% Null values"))
  }
}
```

```
## [1] "Column  Consumer.Electronics  has  23.529 % Null values"
## [1] "Column  Industrial.Electronics  has  23.529 % Null values"
## [1] "Column  Computer.Hardware  has  23.529 % Null values"
## [1] "Column  Communication...Broadcast.Equipments  has  23.529 % Null values"
## [1] "Column  Electronics.Components  has  23.529 % Null values"
## [1] "Column  Sub.Total  has  11.765 % Null values"
```

#separate the date column

```
export_data<- separate(export_data, col=Year, into = c("Year", "Year_To"), sep = "-")
```

```
export_data
```

```
##     Year Year_To Consumer.Electronics Industrial.Electronics Computer.Hardware
## 1   2000      01                  648                    500              1250
## 2   2001      02                  700                    950              1800
## 3   2002      03                  750                   1400               550
## 4   2003      04                  825                   1515              1440
## 5   2004      05                 1150                   1500              1200
## 6   2005      06                 2000                   2300              1025
## 7   2006      07                 1500                   3000              1500
## 8   2007      08                 1600                   3885               990
## 9   2008      09                 2600                   4200              1650
## 10  2009      10                 3000                   3500              1900
## 11  2010      11                 1400                   4500              1300
## 12  2011      12                 1227                   5600              2100
## 13  2012      13                 1600                   5900              2400
## 14  2013      14                 <NA>                   <NA>              <NA>
## 15  2014      15                 <NA>                   <NA>              <NA>
## 16  2015      16                 <NA>                   <NA>              <NA>
## 17  2016      17                 <NA>                   <NA>              <NA>
##    Communication...Broadcast.Equipments Electronics.Components Sub.Total
## 1                                   550                   1840      4788
## 2                                   150                   2200      5800
## 3                                   500                   2400      5600
## 4                                   165                   3755      7700
## 5                                   350                   3800      8000
## 6                                   500                   3800      9625
## 7                                   650                   5850     12500
## 8                                   625                   6100     13200
## 9                                 12280                  10500     31230
## 10                                 7800                   9700     25900
## 11                                14800                  18400     40400
## 12                                18200                  15500     42627
## 13                                20900                  13200     44000
## 14                                 <NA>                   <NA>     46704
## 15                                 <NA>                   <NA>     36692
## 16                                 <NA>                   <NA>      <NA>
## 17                                 <NA>                   <NA>      <NA>
##    Computer.Software  Total
## 1              28350  33138
## 2              36500  42300
## 3              46100  51700
## 4              58240  65940
## 5              80180  88180
## 6             104100 113725
## 7             141000 153500
## 8             164400 177600
## 9             216190 247420
## 10            237000 262900
## 11            268610 309010
## 12            332769 375396
## 13            412191 456191
## 14            527292 573996
## 15            600000 636692
## 16            700000 700000
## 17            779200 779200
```

```
export_data <- select(export_data, -Year_To)

export_data
```

```
##     Year Consumer.Electronics Industrial.Electronics Computer.Hardware
## 1   2000                  648                    500              1250
## 2   2001                  700                    950              1800
## 3   2002                  750                   1400               550
## 4   2003                  825                   1515              1440
## 5   2004                 1150                   1500              1200
## 6   2005                 2000                   2300              1025
## 7   2006                 1500                   3000              1500
## 8   2007                 1600                   3885               990
## 9   2008                 2600                   4200              1650
## 10  2009                 3000                   3500              1900
## 11  2010                 1400                   4500              1300
## 12  2011                 1227                   5600              2100
## 13  2012                 1600                   5900              2400
## 14  2013                 <NA>                   <NA>              <NA>
## 15  2014                 <NA>                   <NA>              <NA>
## 16  2015                 <NA>                   <NA>              <NA>
## 17  2016                 <NA>                   <NA>              <NA>
##    Communication...Broadcast.Equipments Electronics.Components Sub.Total
## 1                                   550                   1840      4788
## 2                                   150                   2200      5800
## 3                                   500                   2400      5600
## 4                                   165                   3755      7700
## 5                                   350                   3800      8000
## 6                                   500                   3800      9625
## 7                                   650                   5850     12500
## 8                                   625                   6100     13200
## 9                                 12280                  10500     31230
## 10                                 7800                   9700     25900
## 11                                14800                  18400     40400
## 12                                18200                  15500     42627
## 13                                20900                  13200     44000
## 14                                 <NA>                   <NA>     46704
## 15                                 <NA>                   <NA>     36692
## 16                                 <NA>                   <NA>      <NA>
## 17                                 <NA>                   <NA>      <NA>
##    Computer.Software  Total
## 1              28350  33138
## 2              36500  42300
## 3              46100  51700
## 4              58240  65940
## 5              80180  88180
## 6             104100 113725
## 7             141000 153500
## 8             164400 177600
## 9             216190 247420
## 10            237000 262900
## 11            268610 309010
## 12            332769 375396
## 13            412191 456191
## 14            527292 573996
## 15            600000 636692
## 16            700000 700000
## 17            779200 779200
```

#Excluding the NA values

```
export_dataset <- na.exclude(export_data)
```

```
View(export_dataset)
```

```
year_export=as.Date(export_dataset$Year,'%Y')

year_export=as.numeric(format(year_export,'%Y'))

export_dataset["year_export"]=NA
export_dataset$year_export=year_export
export_dataset$year_export=as.integer(export_dataset$year_export)
```

```
export_dataset$Consumer.Electronics<-as.numeric(export_dataset$Consumer.Electronics)
export_dataset$Industrial.Electronics<-as.numeric((export_dataset$Industrial.Electronics))
export_dataset$Computer.Hardware<-as.numeric(export_dataset$Computer.Hardware)
export_dataset$Communication...Broadcast.Equipments<-as.numeric(export_dataset$Communication...Broadcas
export_dataset$Electronics.Components<-as.numeric(export_dataset$Electronics.Components)
export_dataset$Sub.Total<-as.numeric(export_dataset$Sub.Total)
export_dataset$Computer.Software<-as.numeric(export_dataset$Computer.Software)
export_dataset$Total<-as.numeric((export_dataset$Total))
```

```
head(export_dataset)
```

```
##   Year Consumer.Electronics Industrial.Electronics Computer.Hardware
## 1 2000                  648                    500              1250
## 2 2001                  700                    950              1800
## 3 2002                  750                   1400               550
## 4 2003                  825                   1515              1440
## 5 2004                 1150                   1500              1200
## 6 2005                 2000                   2300              1025
##   Communication...Broadcast.Equipments Electronics.Components Sub.Total
## 1                                  550                   1840      4788
## 2                                  150                   2200      5800
## 3                                  500                   2400      5600
## 4                                  165                   3755      7700
## 5                                  350                   3800      8000
## 6                                  500                   3800      9625
##   Computer.Software  Total year_export
## 1             28350  33138        2000
## 2             36500  42300        2001
## 3             46100  51700        2002
## 4             58240  65940        2003
## 5             80180  88180        2004
## 6            104100 113725        2005
```

```
export_dataset<-select(export_dataset, -year_export)
```

```
export_dataset
```

```
##    Year Consumer.Electronics Industrial.Electronics Computer.Hardware
## 1  2000                  648                    500              1250
## 2  2001                  700                    950              1800
## 3  2002                  750                   1400               550
## 4  2003                  825                   1515              1440
## 5  2004                 1150                   1500              1200
## 6  2005                 2000                   2300              1025
## 7  2006                 1500                   3000              1500
## 8  2007                 1600                   3885               990
## 9  2008                 2600                   4200              1650
## 10 2009                 3000                   3500              1900
## 11 2010                 1400                   4500              1300
## 12 2011                 1227                   5600              2100
## 13 2012                 1600                   5900              2400
##    Communication...Broadcast.Equipments Electronics.Components Sub.Total
## 1                                   550                   1840      4788
## 2                                   150                   2200      5800
## 3                                   500                   2400      5600
## 4                                   165                   3755      7700
## 5                                   350                   3800      8000
## 6                                   500                   3800      9625
## 7                                   650                   5850     12500
## 8                                   625                   6100     13200
## 9                                 12280                  10500     31230
## 10                                 7800                   9700     25900
## 11                                14800                  18400     40400
## 12                                18200                  15500     42627
## 13                                20900                  13200     44000
##    Computer.Software  Total
## 1              28350  33138
## 2              36500  42300
## 3              46100  51700
## 4              58240  65940
## 5              80180  88180
## 6             104100 113725
## 7             141000 153500
## 8             164400 177600
## 9             216190 247420
## 10            237000 262900
## 11            268610 309010
## 12            332769 375396
## 13            412191 456191
```

#Finading the Correlation

```
cor(export_dataset$Consumer.Electronics, export_dataset$Industrial.Electronics, method="pearson")
```

```
## [1] 0.5003069
```

```
cor(export_dataset$Consumer.Electronics, export_dataset$Computer.Hardware, method="pearson")
```

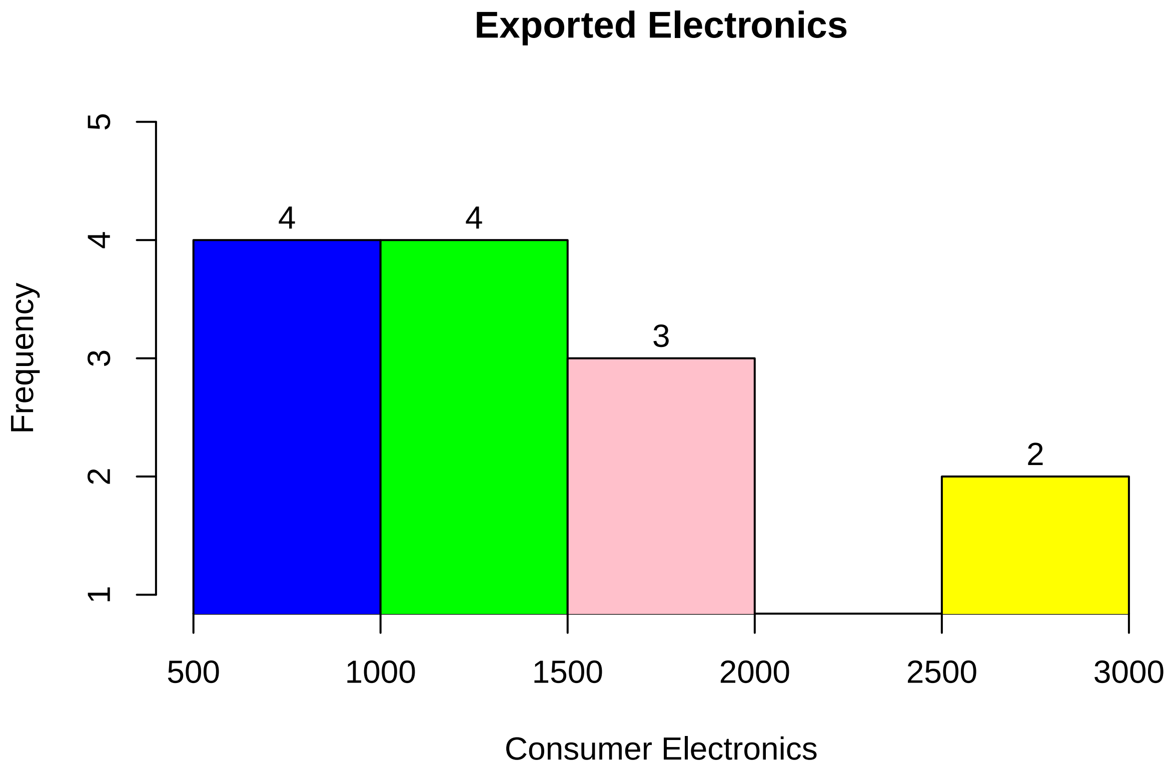
```
## [1] 0.2841414
```

```
cor(export_dataset$Consumer.Electronics, export_dataset$Communication...Broadcast.Equipments, method =
```

```
## [1] 0.336246
```

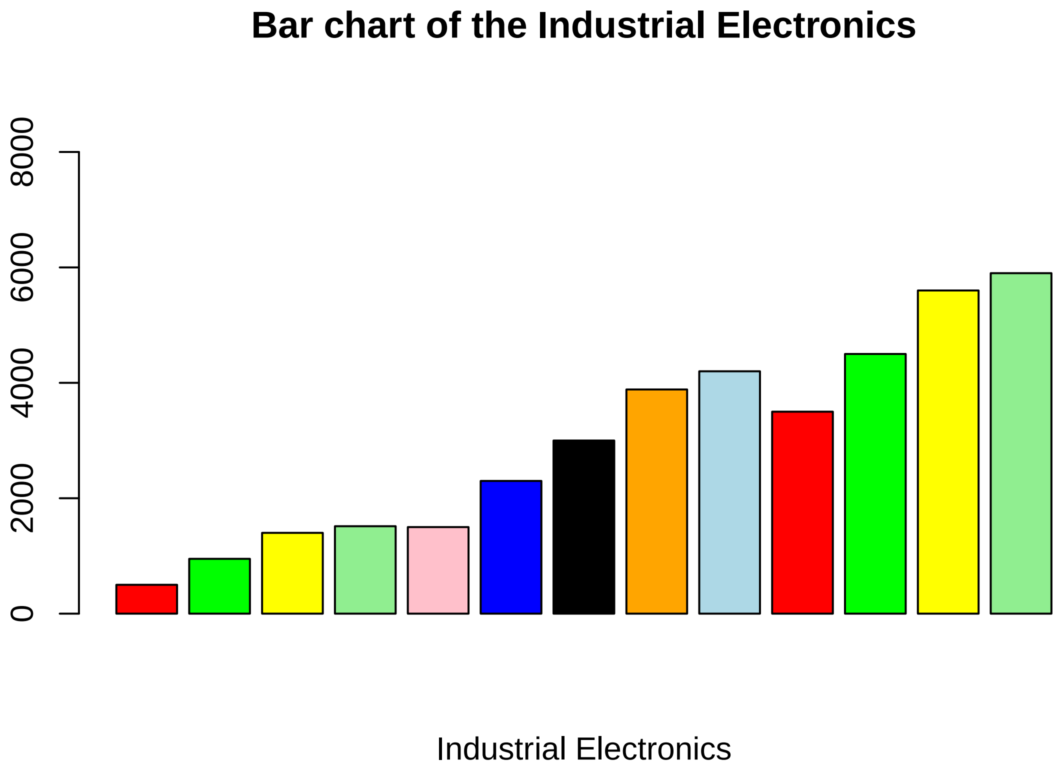
#The histogram

```
hist(export_dataset$Consumer.Electronics, xlab = "Consumer Electronics", col = c("blue","green","pink",
```



```
colour <- c("red","green","yellow","light green", "pink", "blue", "black","orange","light blue")
```

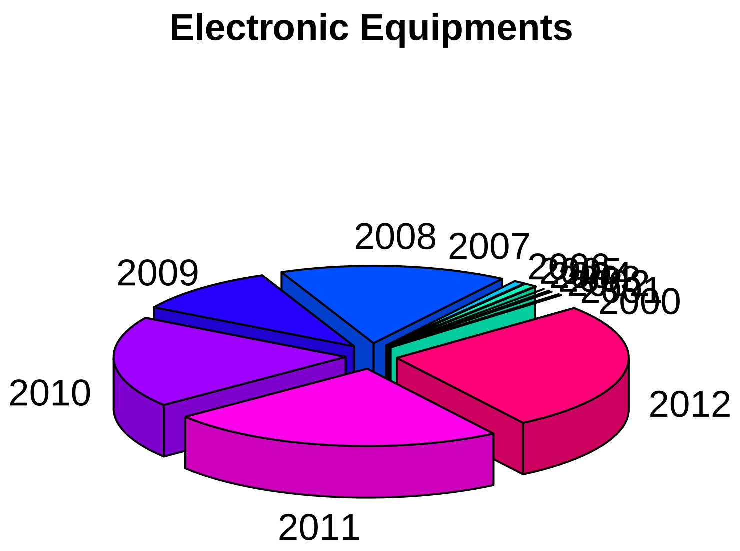
```
barplot(export_dataset$Industrial.Electronics, col=colour, main = "Bar chart of the Industrial Electron
```



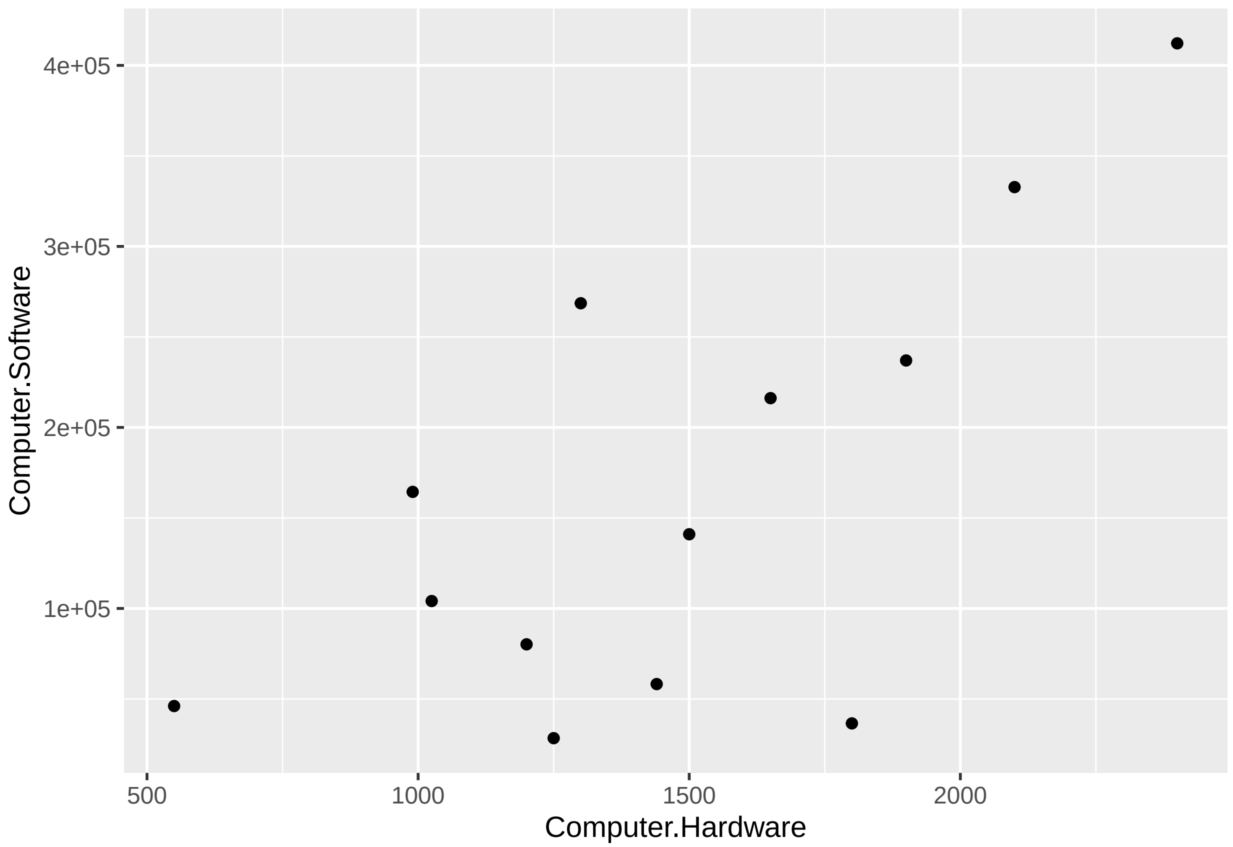
```
T_Elect_eq <- export_dataset$Communication...Broadcast.Equipments
T_El <- table(T_Elect_eq)
```

```
#pie(export_dataset$Communication...Broadcast.Equipments, col = colour)


pie3D(export_dataset$Communication...Broadcast.Equipments, labels = export_dataset$Year, main = "Electr
```



```r
ggplot(export_dataset, aes(x=Computer.Hardware, y=Computer.Software)) + geom_point()
```


4e+05
3e+05
2e+05
1e+05
Computer.Software
500
1000
1500
2000
Computer.Hardware