BIG DATA TOOLS FOR MANAGERS

Unit-3 : Introduction to R

by
Ankit Velani
Adjunct Faculty, Dept. of MBA,
Siddaganga Institute of Technology,Tumkur

Session-3 : Basic Graphs

- Basic Graphs in R
  - Table
  - Histogram
  - Boxplot
  - Pie
  - Bar graph
  - Scatter plot
  - Line chart

Table

View() function to display the data in Table format.

Example:

```
v_data = read.csv("C:/dataset/VEHICLE_PARK.csv")

View(v_data)
```

Table

table() function used to calculate the frequency count for the categorical variable.

Example:

```
v_data = read.csv("C:/dataset/VEHICLE_PARK.csv")

# Count the frequency for the Vehicle Type column
table(v_data$VEHICLE_TYPE)
```

Output:

```
table(data$VEHICLE_TYPE)

      BUSE  FOUR WHEELER        OTHERS         TRUCK  TWO WHEELER
      5412          4510          4510          5412          2706
```

Table - prop.table

Example: prop.table

```
# Count the frequency for the Vehicle Type column
v_freq = table(v_data$VEHICLE_TYPE)

# Display data in fractions/percentage
prop.table(v_freq)
```

Output:

```
prop.table(v_freq)

      BUSE FOUR WHEELER       OTHERS        TRUCK  TWO WHEELER
0.24              0.20         0.20         0.24         0.12
```

Marketing Data

- This dataset is about monthly marketing spend for generating sales for each month. So here Sales is a dependent variable and Spends is an independent variable.
- **Columns :**

Month	Spend	Sales
1	1000	9914
2	4000	40487
3	5000	54324
4	4500	50044
5	3000	34719
6	4000	42551
7	9000	94871
8	11000	118914
9	15000	158484
10	12000	131348
11	7000	78504
12	3000	36284

Read Marketing data

```
# Importing CSV file in R
data = read.csv("C:/dataset/marketing-spend.csv")
```

Histogram

hist() function to display the histogram for any dataset variable.

Example:

```
# Show Histogram for Sales
hist(data$Sales)
```

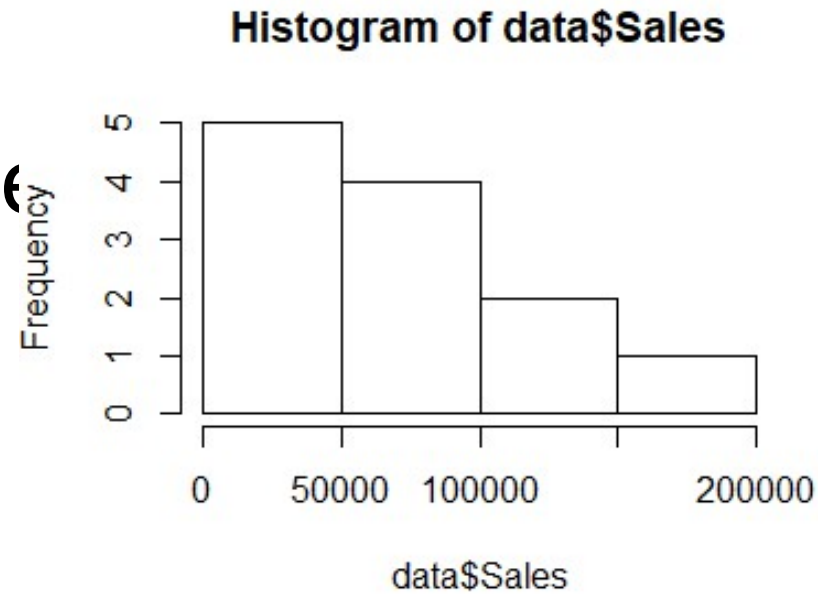
Histogram

hist() function to display the histogram for any dataset variable.

Example:

Show Histogram for Sale

hist(data$Sales)



Histogram

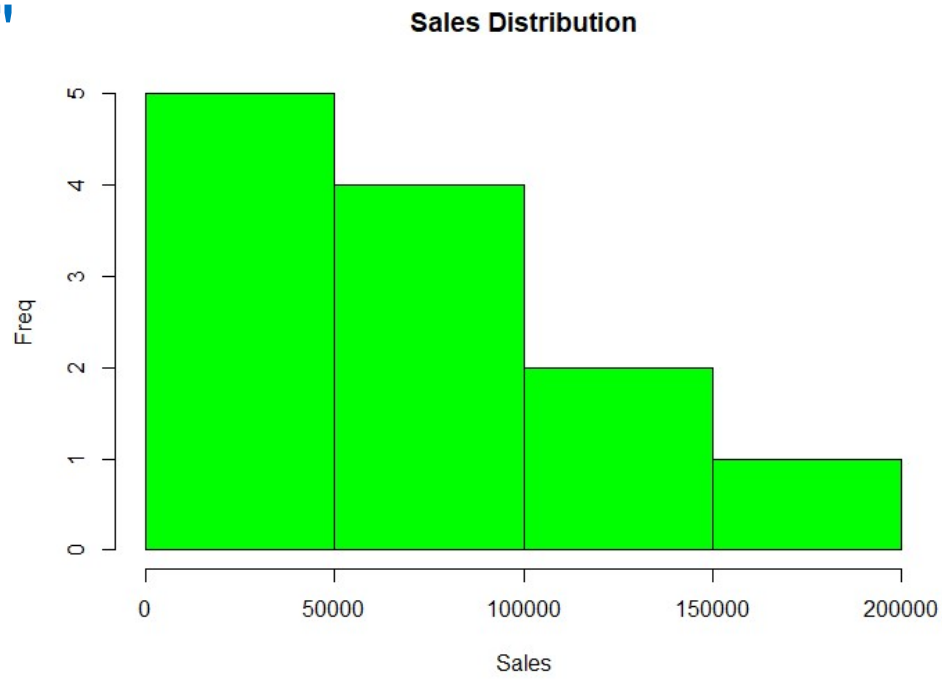
hist() function to display the histogram for any dataset variable.

Customize histogram by setting up the title, color, x or y axis label

Histogram

Show Histogram for Sales with colors and title

```
hist(data$Sales,
        col    = "green",
        main = "Sales Distribution"
        xlab  = "Sales",
        ylab  = "Freq" )
```



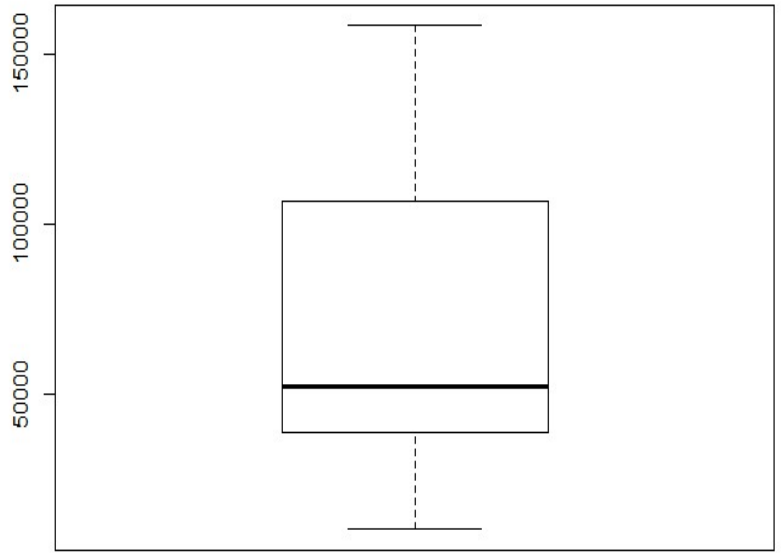
Boxplot

boxplot() function to display the boxplot for numeric variable.

Example:

Show Boxplot for Sales

boxplot(data$Sales)



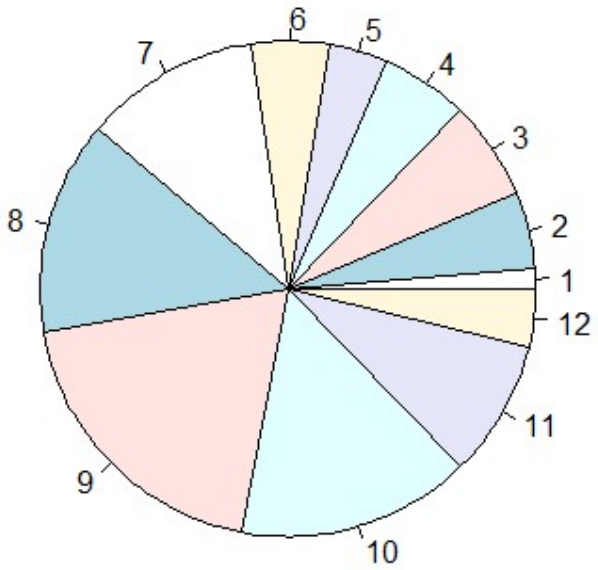
Pie Chart

pie() function to display the pie chart for numeric & categorical variable

Example:

Display Pie chart for Monthly Spend

```
pie(data$Spend, data$Month)
```



Bar Graph

barplot() function to display the bar graph for numeric & categorical variable

Example:

```
# Display Bar graph for Monthly Spend
barplot(data$Spend,
        names.arg = data$Month)
```

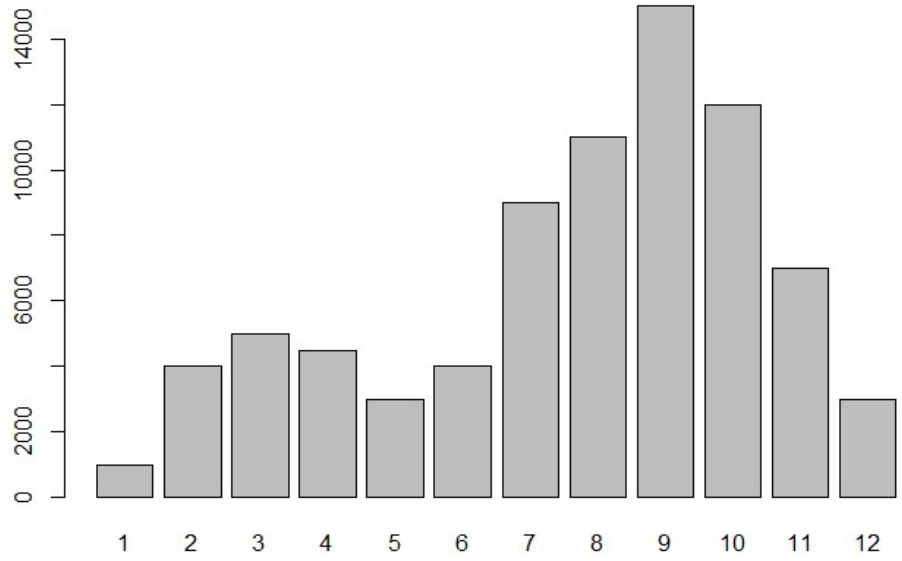
Bar Graph

barplot() function to display the bar graph for numeric & categorical variable

Example:

Display Bar graph for Monthly Spend

```
barplot(data$Spend,
        names.arg = data$Month)
```



Bar Graph

barplot() function to display the bar graph for numeric & categorical variable

Example:

```
# Display Bar graph for Monthly Spend
barplot(data$Spend,
        names.arg = data$Month,
        horiz= TRUE)
```

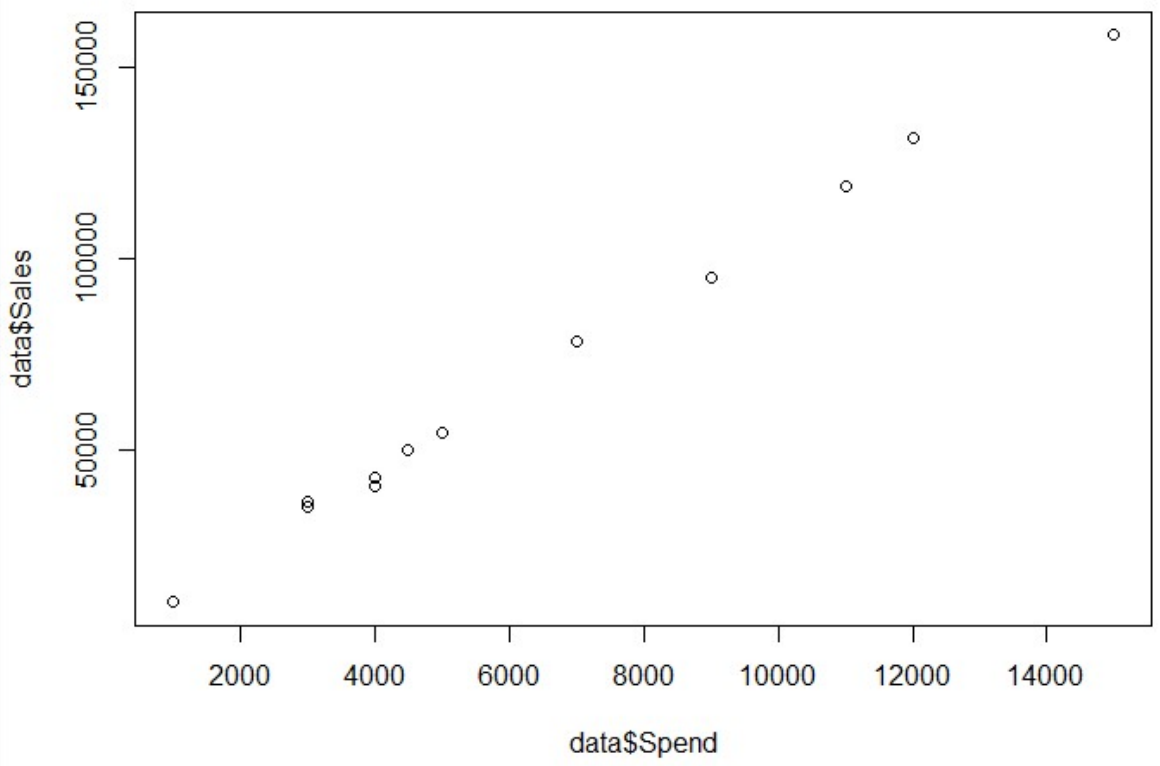
Convert Bar graph from vertical to horizontal

Scatter plot

plot() function will display scatter plot if both the variables are numeric.

Example:

```
# Display scatter plot for Spend vs Sales
plot(data$Spend,
     data$Sales)
```



Line Chart

plot() function with additional parameters will display line chart.

Example:

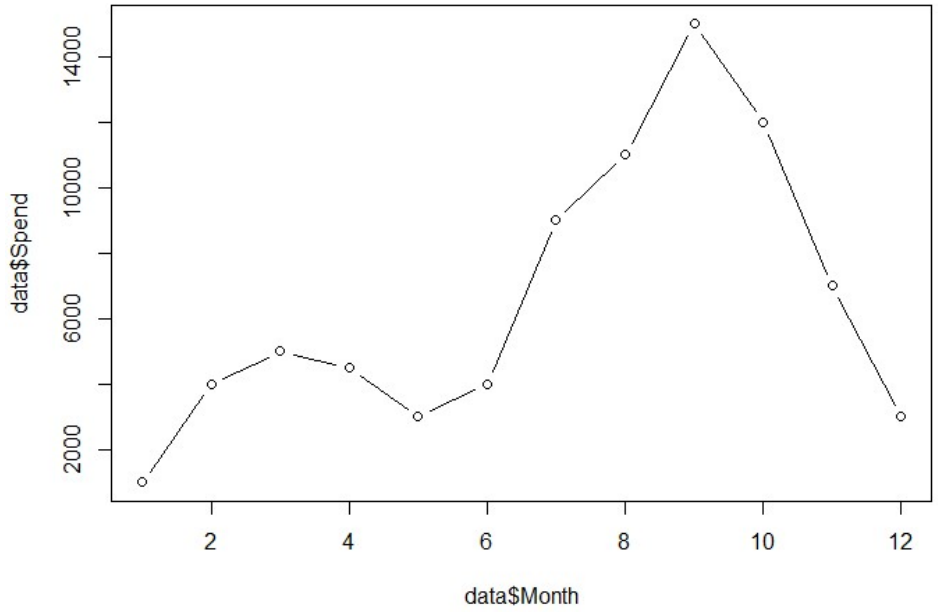
```
# Display Line Chart Month vs Spend
plot(data$Month,
     data$Spend,
     type='b') # Type : p l b c o h s S
```

Line Chart

plot() function with additional parameters will display line chart.

Example:

```
# Display Line Chart Month vs Spend
plot(data$Month,
     data$Spend,
     type='b') # Type : p l b c o h s S
```



Regression

lm() function helps us to create regression model in R with given formula in the form of **Y ~ X+X2+X3+X4…etc**

summary() functions to look the model and it's parameters such as formula, coefficients,standard error, residual, multiple/adjusted R-Square..etc to analyze regression model

predict() function used to make a prediction on new data, and we can dervided formula forprediction y = b0 + b1*x1* + *b2*x2 + b3*x3..etc

Correlation

cor() function helps us to get the correlation for the variables.

Example:

```
cor(data$Spend,data$Sales)
```

#Default method is pearson correlation

Correlation

cor() function helps us to get the correlation for the variables.

Example:

```
cor(data$Spend,data$Sales)

# Methods: "pearson", "kendall", "spearman"
cor(data$Spend,data$Sales, method = "spearman")
```

Regression

Parameters for lm() function

- Dependent variable
- Indepedent Variable
- Data Source

Example:

```
model_1 <- lm(Sales~Spend, data)  #Simple Linear Regression

model_2 <- lm(Sales~Spend+Month, data) #Multiple Linear Regression
```

Regression

Parameters for lm() function

- Dependent variable
- Indepedent Variable
- Data Source

Example:

```
model_1 <- lm(Sales~Spend, data)  #Simple Linear Regression

model_2 <- lm(Sales~Spend+Month, data) #Multiple Linear Regression
```

Regression - Summary

Example:

model_1 <- lm(Sales~Spend, data) #Simple Linear Regression

summary(model_1)

```
> summary(model_1)

Call:
lm(formula = Sales ~ Spend, data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -3385  -2097    258   1726   3034

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1383.4714  1255.2404   1.102    0.296
Spend         10.6222     0.1625  65.378 1.71e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2313 on 10 degrees of freedom
Multiple R-squared:  0.9977,    Adjusted R-squared:  0.9974
F-statistic:  4274 on 1 and 10 DF,  p-value: 1.707e-14
```