Practical No: 2-2

AIM: *Perform data analysis using R programming*

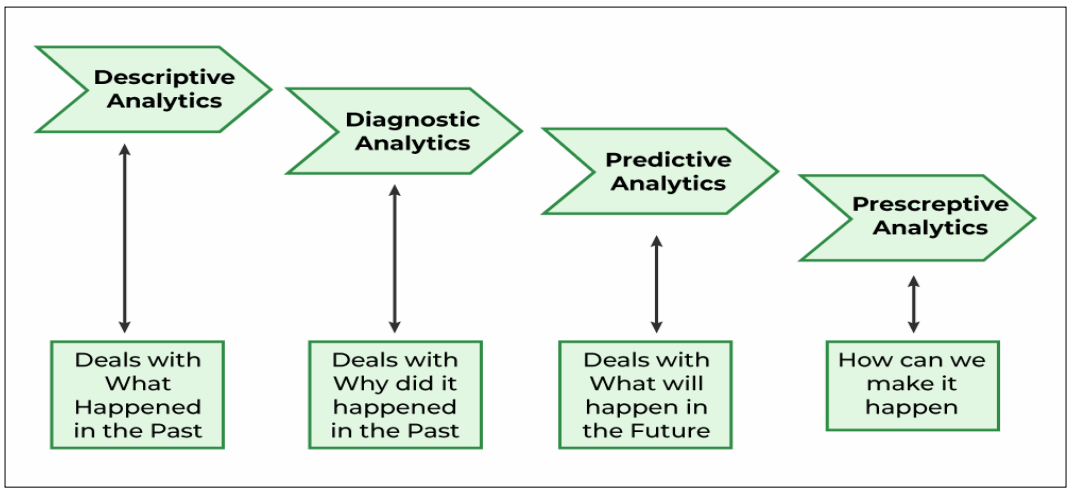
THEORY:

Data analysis using R: Data Analysis is a subset of data analytics, it is a process where the objective has to be made clear, collect the relevant data, preprocess the data, perform analysis(understand the data, explore insights), and then visualize it. The last step visualization is important to make people understand what's happening in the firm.

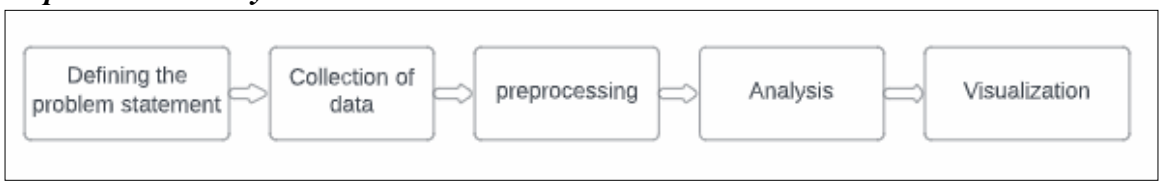
Types of Data Analytics

There are four major types of data analytics:

1. Predictive (forecasting)
2. Descriptive (business intelligence and data mining)
3. Prescriptive (optimization and simulation)
4. Diagnostic analytics



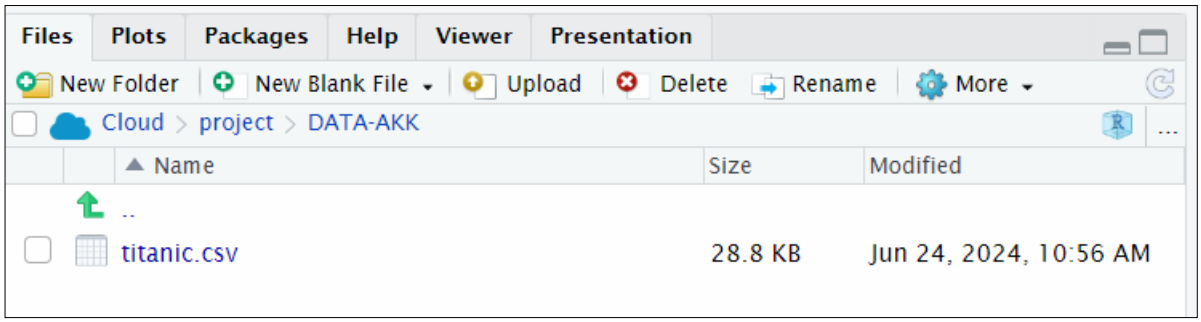
Steps in Data Analysis



Data Analysis using the Titanic dataset:

Save the dataset in the current working directory, now we will start analysis (getting to know our data).

Students can prefer the Free Posit Cloud. Posit Cloud (formerly RStudio Cloud) lets you access Posit's powerful set of data science tools right in your browser – no installation or complex configuration required. And can choose to sign in for free. (read site instructions carefully)




```
titanic=read.csv("train.csv")
```

```
head(titanic)
```

```
> titanic=read.csv("titanic.csv")
> head(titanic)
  PassengerId Survived Pclass                                            Name    Sex
1         892        0      3                                 Kelly, Mr. James   male
2         893        1      3                 Wilkes, Mrs. James (Ellen Needs) female
3         894        0      2                        Myles, Mr. Thomas Francis   male
4         895        0      3                                 Wirz, Mr. Albert   male
5         896        1      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female
6         897        0      3                       Svensson, Mr. Johan Cervin   male
   Age SibSp Parch  Ticket    Fare Cabin Embarked
1 34.5     0     0  330911  7.8292              Q
2 47.0     1     0  363272  7.0000              S
3 62.0     0     0  240276  9.6875              Q
4 27.0     0     0  315154  8.6625              S
5 22.0     1     1 3101298 12.2875              S
6 14.0     0     0    7538  9.2250              S
```

To understand the class(data type) of each column **sapply()** method can be used.

```
sapply(titanic,class)
```

```
> sapply(titanic, class)
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp
  "integer"   "integer"   "integer" "character" "character"   "numeric"   "integer"
      Parch      Ticket        Fare       Cabin    Embarked
  "integer" "character"   "numeric" "character" "character"
> |
```

To analyze data using a summary of all the columns, their values, and data types. summary() can be used for this purpose.

summary(titanic)

```
> summary(titanic)
  PassengerId        Survived          Pclass          Name
 Min.   : 892.0   Min.   :0.0000   Min.   :1.000   Length:418
 1st Qu.: 996.2   1st Qu.:0.0000   1st Qu.:1.000   Class :character
 Median :1100.5   Median :0.0000   Median :3.000   Mode  :character
 Mean   :1100.5   Mean   :0.3636   Mean   :2.266
 3rd Qu.:1204.8   3rd Qu.:1.0000   3rd Qu.:3.000
 Max.   :1309.0   Max.   :1.0000   Max.   :3.000

     Sex                 Age            SibSp            Parch
 Length:418         Min.   : 0.17   Min.   :0.0000   Min.   :0.0000
 Class :character   1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000
 Mode  :character   Median :27.00   Median :0.0000   Median :0.0000
                    Mean   :30.27   Mean   :0.4474   Mean   :0.3923
                    3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.0000
                    Max.   :76.00   Max.   :8.0000   Max.   :9.0000
                    NA's   :86
    Ticket               Fare            Cabin             Embarked
 Length:418         Min.   :  0.000   Length:418         Length:418
 Class :character   1st Qu.:  7.896   Class :character   Class :character
 Mode  :character   Median : 14.454   Mode  :character   Mode  :character
                    Mean   : 35.627
                    3rd Qu.: 31.500
                    Max.   :512.329
                    NA's   :1
```

From the above summary Students to extract below observations:

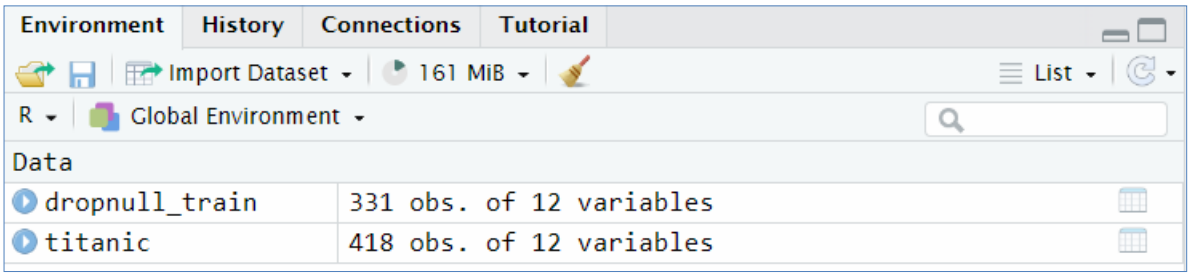
- Total passengers: 891
- The number of total people who survived: 342
- Number of total people dead: 549
- Number of males in the titanic: 577
- Number of females in the titanic: 314
- Maximum age among all people in titanic: 80
- Median age: 28

Preprocessing of the data is important before analysis, so null values have to be checked and removed.

sum(is.na(train))

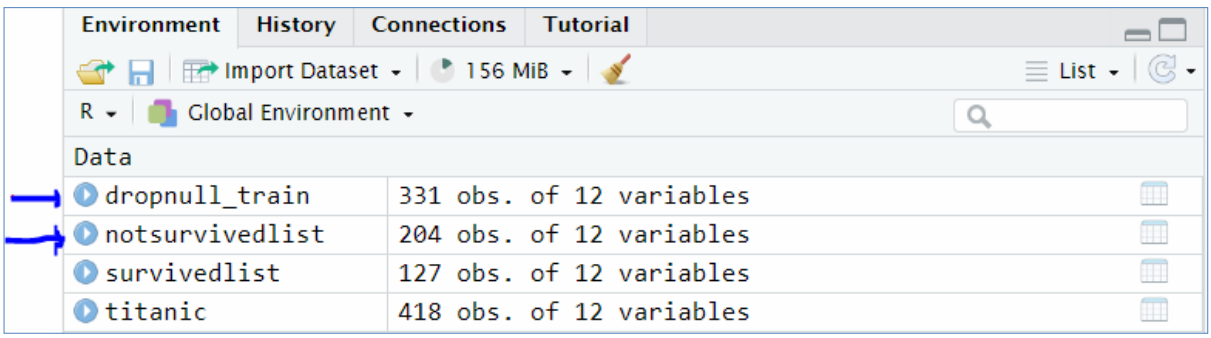
dropnull_train=titanic[rowSums(is.na(titanic))<=0,]

- dropnull_train contains only 331 rows because (total rows in dataset (418) – null value rows (87) = remaining rows (331))
- Now lets will divide survived and dead people into a separate list from 331 rows.



survivedlist=dropnull_train[dropnull_train$Survived == 1,]
notsurvivedlist=dropnull_train[dropnull_train$Survived == 0,]

```
> survivedlist=dropnull_train[dropnull_train$Survived == 1,]
> notsurvivedlist=dropnull_train[dropnull_train$Survived == 0,]
>
```



Visualization:

Now to visualize the number of males and females dead and survived using *bar plots, histograms, and piecharts.*

Bar charts are a popular and effective way to visually represent categorical data in a structured manner. R stands out as a powerful programming language for data analysis and visualization.

A bar chart also known as bar graph is a pictorial representation of data that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. In other words, it is the

pictorial representation of the dataset. These data sets contain the numerical values of variables that represent the length or height.

R uses the barplot() function to create bar charts. Here, both vertical and Horizontal bars can be drawn.

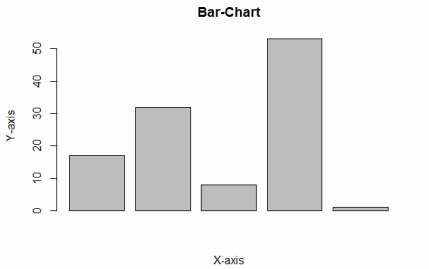
Syntax: barplot(H, xlab, ylab, main, names.arg, col, horiz = TRUE)

Parameters:
H: This parameter is a vector or matrix containing numeric values which are used in bar chart.
xlab: This parameter is the label for x axis in bar chart.
ylab: This parameter is the label for y axis in bar chart.
main: This parameter is the title of the bar chart.
names.arg: This parameter is a vector of names appearing under each bar in bar chart.
col: This parameter is used to give colors to the bars in the graph.
horizontal = `TRUE`

Ex:

```
# Create the data for the chart
A <- c(17, 32, 8, 53, 1)

# Plot the bar chart
barplot(A, xlab = "X-axis", ylab = "Y-axis", main ="Bar-Chart")
```



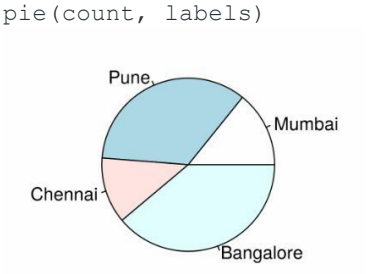
A **pie chart** is a circular statistical graphic, which is divided into slices to illustrate numerical proportions. It depicts a special chart that uses "pie slices", where each sector shows the relative sizes of data. A circular chart cuts in the form of radii into segments describing relative frequencies or magnitude also known as a circle graph. R Programming Language uses the function pie() to create pie charts. It takes positive numbers as a vector input.

Syntax: pie(x, labels, radius, main, col, clockwise)

Parameters:
x: This parameter is a vector that contains the numeric values which are used in the pie chart.
labels: This parameter gives the description to the slices in pie chart.
radius: This parameter is used to indicate the radius of the circle of the pie chart.(value between -1 and +1).
main: This parameter is represents title of the pie chart.
clockwise: This parameter contains the logical value which indicates whether the slices are drawn clockwise or in anti clockwise direction.
col: This parameter give colors to the pie in the graph.

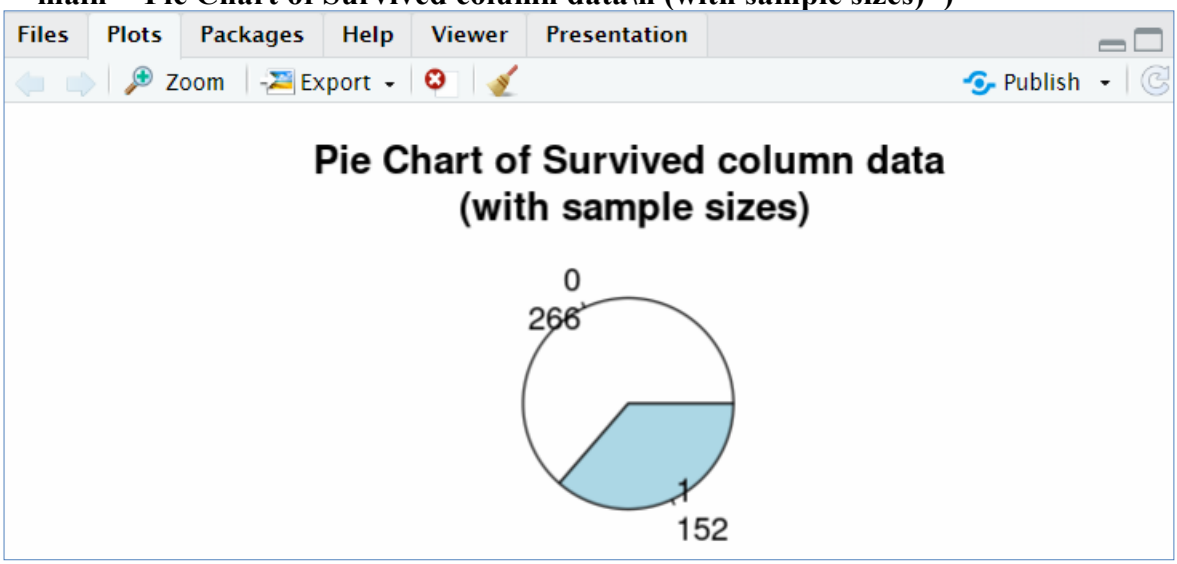
```
Ex:
# Create data for the graph.
Count<- c(23, 56, 20, 63)
labels <- c("Mumbai", "Pune", "Chennai", "Bangalore")

# Plot the chart.
pie(count, labels)
```



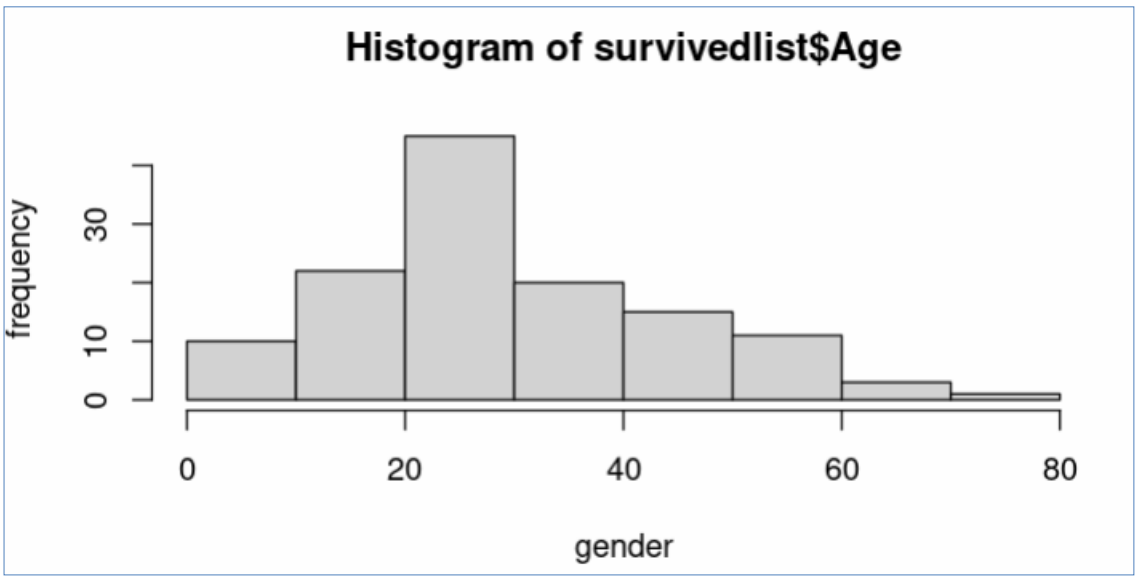
For the Titanic data set, creating a pie chart to visualize the number of males and females dead and survived.

```
mytable <- table(titanic$Survived)
lbls <- paste(names(mytable), "\n", mytable, sep="")
pie(mytable,
   labels = lbls,
   main="Pie Chart of Survived column data\n (with sample sizes)")
```




```
hist(survivedlist$Age,
    xlab="gender",
    ylab="frequency")
```

```
> hist(survivedlist$Age,
+      xlab="gender",
+      ylab="frequency")
> |
```



```
barplot(table(notsurvivedlist$Sex),
      xlab="gender",
      ylab="frequency")
```

Inclass Assessment:

1. Draw barplot to Analyse males and females those who not survived in titanic.
2. There are some passengers who are charged extremely high. So, these values can affect the analysis as they are outliers. Confirm their presence using a boxplot.