Mini Project

Dataframe:

A Dataframe is a data structure that organizes data into a 2-dimensional table of rows and columns, much like a spreadsheet. Dataframes are one of the most common data structures used in modern data analytics because they are a flexible and intuitive way of storing and working with data. A data frame is a special case of the list in which each component has equal length. A data frame is used to store data table and the vectors which are present in the form of a list in a data frame, are of equal length. In a simple way, it is a list of equal length vectors. A matrix can contain one type of data, but a data frame can contain different data types such as numeric, character, factor, etc.

Program-

```
library(dplyr)
sanika<-data.frame(
  id =c(1,2,3,4,5,6,7,8,9,10),
  product_line = c("Sports","Travel","Electronic Accessories","Home and
lifestyle","Fashion accessories","Health","Beauty","Food and
beverages","Furniture","Books and toys"),
  customer = c("Member", "Normal", "Normal","Member", "Normal",
"Member","Normal", "Normal","Member", "Normal"),
  unit_price = c(74.69,15.28,46.33,58.22,86.31,85.39,68.84,73.56,36.26, 54.84),
  quantity = c( 7, 5, 7, 8, 7, 7, 6, 10, 2, 3),
  gender = c("female", "female", "male", "male", "male", "female", "female",
"male","female", "male"),
  branch = c("A"," C"," A", "A"," C"," B", "B"," C"," A"," B"),
  city = c("Mumbai", "Vashi", "Panvel","Mumbai", "Vashi", "Panvel","Mumbai",
"Vashi", "Panvel", "Mumbai" ),
  tax = c(26.12, 3.82, 16.21, 23.28, 30.20, 29.88, 20.65, 36.78, 11.73, 24.12),
  payment = c("Cash", "Ewallet", "Card","Cash", "Ewallet", "Card","Cash", "Ewallet",
"Card","Cash"),
  gross_margin = c(4.761,4.761,4.761,4.761,4.761,4.761,4.761,4.761,4.761,4.761),
  gross_income = c(26.12, 3.82, 16.21, 23.28, 30.20, 29.88, 20.65, 36.78, 11.73,
24.12),
  rating = c(9.1, 9.6, 7.4, 8.4, 5.3, 4.1, 5.8, 8, 7.2, 5.9),
  date = c(01-05-2019,01-05-2019,03-08-2019, 1-27-2019, 02-08-2019, 3-25-2019, 2-
25-2019, 03-09-2019, 02-12-2019, 02-07-2019),
  total = c(548.9715, 80.22, 340.5255, 489.048, 634.3785, 627.6165, 433.692, 772.38,
76.146, 172.746)
  )
#print data of first four sales
a = print(head(sanika,4))

#print data of last three sales
```

```
a = print(tail(sanika,3))

#print minimum unit price of product line
a = print(min(sanika $unit_price))

#print maximum unit price of product line
a = print(max(sanika $unit_price))

#print mean and median of unit price
a = print(mean(sanika $unit_price))

a = print(median(sanika $unit_price))

#print range of unit price, gross income and product line
print(range(sanika $unit_price))

print(range(sanika $gross_income))

print(range(sanika $product_line))

#variance
print(var(sanika $unit_price))

#Table
print(table(sanika $product_line))

#standard deviation
print(sd(sanika $unit_price))

print("Sanika Mhatre")
timestamp()
```

Output:

```
> library(dplyr)
> sanika<-data.frame(
+   id =c(1,2,3,4,5,6,7,8,9,10),
+   product_line = c("Sports","Travel","Electronic Accessories","Home and
lifestyle","Fashion accessories","Health","Beauty","Food and beverages","F
urniture","Books and toys"),
+   customer = c("Member", "Normal", "Normal","Member", "Normal", "Member"
,"Normal", "Normal","Member", "Normal"),
+   unit_price = c(74.69,15.28,46.33,58.22,86.31,85.39,68.84,73.56,36.26,
54.84),
+   quantity = c( 7, 5, 7, 8, 7, 7, 6, 10, 2, 3),
+   gender = c("female", "female", "male", "male", "male", "female", "fema
le", "male","female", "male"),
+   branch = c("A"," C"," A", "A"," C"," B", "B"," C"," A"," B"),
+   city = c("Mumbai", "Vashi", "Panvel","Mumbai", "Vashi", "Panvel","Mumb
ai", "Vashi", "Panvel", "Mumbai" ),
+   tax = c(26.12, 3.82, 16.21, 23.28, 30.20, 29.88, 20.65, 36.78, 11.73,
24.12),
```

```
+   payment = c("Cash", "Ewallet", "Card","Cash", "Ewallet", "Card","Cash"
, "Ewallet", "Card","Cash"),
+   gross_margin = c(4.761,4.761,4.761,4.761,4.761,4.761,4.761,4.761,4.761
,4.761),
+   gross_income = c(26.12, 3.82, 16.21, 23.28, 30.20, 29.88, 20.65, 36.78
, 11.73, 24.12),
+   rating = c(9.1, 9.6, 7.4, 8.4, 5.3, 4.1, 5.8, 8, 7.2, 5.9),
+   date = c(01-05-2019,01-05-2019,03-08-2019, 1-27-2019, 02-08-2019, 3-25
-2019, 2-25-2019, 03-09-2019, 02-12-2019, 02-07-2019),
+   total = c(548.9715, 80.22, 340.5255, 489.048, 634.3785, 627.6165, 433.
692, 772.38, 76.146, 172.746)
+   )
> #print data of first four sales
> a = print(head(sanika,4))
  id           product_line customer unit_price
1  1                 Sports   Member      74.69
2  2                 Travel   Normal      15.28
3  3 Electronic Accessories   Normal      46.33
4  4     Home and lifestyle   Member      58.22
  quantity gender branch   city   tax payment
1        7 female      A Mumbai 26.12    Cash
2        5 female      C  Vashi  3.82 Ewallet
3        7   male      A Panvel 16.21    Card
4        8   male      A Mumbai 23.28    Cash
  gross_margin gross_income rating  date
1        4.761        26.12    9.1 -2023
2        4.761         3.82    9.6 -2023
3        4.761        16.21    7.4 -2024
4        4.761        23.28    8.4 -2045
     total
1 548.9715
2  80.2200
3 340.5255
4 489.0480
>
> #print data of last three sales
> a = print(tail(sanika,3))
   id       product_line customer unit_price
8   8 Food and beverages   Normal      73.56
9   9          Furniture   Member      36.26
10 10     Books and toys   Normal      54.84
   quantity gender branch   city   tax payment
8        10   male      C  Vashi 36.78 Ewallet
9         2 female      A Panvel 11.73    Card
10        3   male      B Mumbai 24.12    Cash
   gross_margin gross_income rating  date
8         4.761        36.78    8.0 -2025
9         4.761        11.73    7.2 -2029
10        4.761        24.12    5.9 -2024
     total
8  772.380
9   76.146
10 172.746
>
> #print minimum unit price of product line
> a = print(min(sanika $unit_price))
[1] 15.28
>
> #print maximum unit price of product line
> a = print(max(sanika $unit_price))
[1] 86.31
>
> #print mean and median of unit price
> a = print(mean(sanika $unit_price))
[1] 59.972
>
> a = print(median(sanika $unit_price))
[1] 63.53
>
```

```
> #print range of unit price, gross income and product line
> print(range(sanika $unit_price))
[1] 15.28 86.31
>
> print(range(sanika $gross_income))
[1]  3.82 36.78
>
> print(range(sanika $product_line))
[1] "Beauty" "Travel"
>
> #variance
> print(var(sanika $unit_price))
[1] 510.5338
>
> #Table
> print(table(sanika $product_line))

                Beauty         Books and toys
                     1                      1
Electronic Accessories    Fashion accessories
                     1                      1
    Food and beverages              Furniture
                     1                      1
                Health     Home and lifestyle
                     1                      1
                Sports                 Travel
                     1                      1
>
> #standard deviation
> print(sd(sanika $unit_price))
[1] 22.595
>
>
> print("Sanika Mhatre")
[1] "Sanika Mhatre"
> timestamp()
##------ Thu Apr 13 03:17:30 2023 ------##
```

EDA:

Exploratory data analysis (EDA) is a crucial step in data analysis that can be performed using R programming language. R provides a wide range of powerful tools and packages for EDA, including data visualization libraries like ggplot2, data manipulation libraries like dplyr, and statistical computing libraries like tidyr.

To perform EDA in R, the first step is to import the data into R using functions like read.csv(), read_excel(), or read.table(). Then, data cleaning and preprocessing techniques can be applied to handle missing values, remove outliers, and transform variables.

After data preprocessing, data visualization can be performed using ggplot2, which provides an extensive set of tools for creating a wide range of charts, histograms, and other plots. Summary statistics like mean, median, standard deviation, and correlation coefficients can be calculated using built-in R functions.

EDA can also involve identifying patterns and relationships in the data using techniques like clustering, principal component analysis, or factor analysis. These techniques can be implemented using R packages like cluster, factoextra, or psych.

Data Science:

Data science in R programming refers to the use of R, a popular programming language for statistical computing and data analysis, for tasks related to data science. R provides a comprehensive set of tools and libraries for data science, including data wrangling, exploratory data analysis, statistical modeling, machine learning, and data visualization.

R has a rich ecosystem of packages, which are collections of functions and data sets, that make it easy to perform complex data science tasks. Some popular packages for data science in R include:

dplyr: for data manipulation and filtering

ggplot2: for data visualization

tidyr: for data cleaning and reshaping

caret: for machine learning and predictive modeling

shiny: for creating interactive web applications with data

tidyr: for data cleaning and reshaping

caret: for machine learning and predictive modelling

shiny: for creating interactive web applications with data

R is also highly extensible, which means that users can create their own packages and functions to extend its functionality for specific tasks. This makes it a powerful tool for data scientists who need to work with large and complex data sets.

Overall, R is a popular choice for data science due to its flexibility, extensibility, and comprehensive set of tools and libraries.

Data manipulation:

It refers to the process of changing or transforming data to prepare it for analysis. It involves a wide range of techniques and methods for modifying and organizing data, such as cleaning, filtering, merging, and reshaping data sets.

Data manipulation is an essential step in the data analysis process because data is often messy, incomplete, or unstructured. Before data can be analyzed, it needs to be cleaned, formatted, and transformed to ensure that it is consistent and relevant to the research question or problem being addressed.

Some common techniques used in data manipulation include:

Cleaning: removing or correcting errors, inconsistencies, or missing data from the data set.

Filtering: selecting a subset of the data based on specific criteria, such as date range or category.

Merging: combining data from different sources or tables based on common variables or keys.

Reshaping: changing the structure or format of the data to better suit the analysis, such as pivoting or aggregating data.

Transforming: creating new variables or variables based on calculations or transformations of existing variables.

Pipelining:

Pipelining in R is a technique used for chaining together multiple operations to process data more efficiently and to make code more readable. It involves passing the output of one operation as the input of the next operation in a sequence, allowing for a streamlined workflow. In R, pipelining can be achieved using the %>% operator

Dplyr-

Select() is a function from dplyr R package that is used to select data frame variables by name, by index, and also is used to rename variables while selecting, and dropping variables by name.

filter() method in R is used to subset a data frame based on a provided condition. If a row satisfies the condition, it must produce TRUE . Otherwise, non-satisfying rows will return NA values. Hence, the row will be dropped.

mutate() function in R programming to add new variables in the specified data frame. These new variables are added by performing the operations on present variables.

arrange() function in R programming that arranges datasets in ascending order.

ggplot2 -

ggplot2 is a R package dedicated to data visualization. It can greatly improve the quality and aesthetics of your graphics, and will make you much more efficient in creating them. ggplot2 allows to build almost any type of chart.

Data visualization-

R, we can create visually appealing data visualizations by writing few lines of code. For this purpose, we use the diverse functionalities of R. Data visualization is an efficient technique for gaining insight about data through a visual medium.

Program:

```
library(dplyr)

sanika<-data.frame(

  id =c(1,2,3,4,5,6,7,8,9,10),
```

```
  product_line = c("Sports","Travel","Electronic Accessories","Home and lifestyle","Fashion accessories","Health","Beauty","Food and beverages","Furniture","Books and toys"),

  customer = c("Member", "Normal", "Normal","Member", "Normal", "Member","Normal", "Normal","Member", "Normal"),

  unit_price = c(74.69,15.28,46.33,58.22,86.31,85.39,68.84,73.56,36.26, 54.84),

  quantity = c( 7, 5, 7, 8, 7, 7, 6, 10, 2, 3),

  gender = c("female", "female", "male", "male", "male", "female", "female", "male","female", "male"),

  branch = c("A"," C"," A", "A"," C"," B", "B"," C"," A"," B"),

  city = c("Mumbai", "Vashi", "Panvel","Mumbai", "Vashi", "Panvel","Mumbai", "Vashi", "Panvel", "Mumbai" ),

  tax = c(26.12, 3.82, 16.21, 23.28, 30.20, 29.88, 20.65, 36.78, 11.73, 24.12),

  payment = c("Cash", "Ewallet", "Card","Cash", "Ewallet", "Card","Cash", "Ewallet", "Card","Cash"),

  gross_margin = c(4.761,4.761,4.761,4.761,4.761,4.761,4.761,4.761,4.761,4.761),

  gross_income = c(26.12, 3.82, 16.21, 23.28, 30.20, 29.88, 20.65, 36.78, 11.73, 24.12),

  rating = c(9.1, 9.6, 7.4, 8.4, 5.3, 4.1, 5.8, 8, 7.2, 5.9),

  date = c(01-05-2019,01-05-2019,03-08-2019, 1-27-2019, 02-08-2019, 3-25-2019, 2-25-2019, 03-09-2019, 02-12-2019, 02-07-2019),

  total = c(548.9715, 80.22, 340.5255, 489.048, 634.3785, 627.6165, 433.692, 772.38, 76.146, 172.746)

  )

#print table

print(sanika)


dataset=sanika

a<-dataset


#Give the dimensions of the dataset

cat ("Dimension:",dim(a))


#Give number of rows in the dataset

cat ("\nrow:", nrow(a))
```

```
#Give number of columns in the dataset
cat("\ncolumn:",ncol (a))

#Questions on functions of dplyr
#Display sales details arranged in ascending order as per their unit price
a%>%arrange (gross_income)->a1
print (a1)

#Display sales details arranged in ascending order as per their unit price
a%>%arrange (unit_price)->a2
print (a2)

#Display the total gross of the sales
a %>% mutate(a, Total_gross=gross_income+gross_margin)->a3
print(a3)

#To display the id,product line,total gross of female customer with quantity > 6
a3%>%select (id, product_line, gender, quantity, Total_gross)%>%filter(gender=="female" &
quantity>6 )->a4
print(a4)

#Display id and product line of students whose total gross is greater than
a3%>%select(id, product_line, Total_gross)%>%filter (Total_gross>25)->a5
print (a5)

#Display the total sales of the sales
a %>% mutate(a, Total_sales=unit_price+tax)->a6
print(a6)
print("Sanika Mhatre")
timestamp()
```

Output:

```
> library(dplyr)
> sanika<-data.frame(
+   id =c(1,2,3,4,5,6,7,8,9,10),
+   product_line = c("Sports","Travel","Electronic Accessories","Home and
lifestyle","Fashion accessories","Health","Beauty","Food and beverages","F
urniture","Books and toys"),
+   customer = c("Member", "Normal", "Normal","Member", "Normal", "Member"
,"Normal", "Normal","Member", "Normal"),
+   unit_price = c(74.69,15.28,46.33,58.22,86.31,85.39,68.84,73.56,36.26,
54.84),
+   quantity = c( 7, 5, 7, 8, 7, 7, 6, 10, 2, 3),
+   gender = c("female", "female", "male", "male", "male", "female", "fema
le", "male","female", "male"),
+   branch = c("A"," C"," A", "A"," C"," B", "B"," C"," A"," B"),
+   city = c("Mumbai", "Vashi", "Panvel","Mumbai", "Vashi", "Panvel","Mumb
ai", "Vashi", "Panvel", "Mumbai" ),
+   tax = c(26.12, 3.82, 16.21, 23.28, 30.20, 29.88, 20.65, 36.78, 11.73,
24.12),
+   payment = c("Cash", "Ewallet", "Card","Cash", "Ewallet", "Card","Cash"
, "Ewallet", "Card","Cash"),
+   gross_margin = c(4.761,4.761,4.761,4.761,4.761,4.761,4.761,4.761,4.761
,4.761),
+   gross_income = c(26.12, 3.82, 16.21, 23.28, 30.20, 29.88, 20.65, 36.78
, 11.73, 24.12),
+   rating = c(9.1, 9.6, 7.4, 8.4, 5.3, 4.1, 5.8, 8, 7.2, 5.9),
+   date = c(01-05-2019,01-05-2019,03-08-2019, 1-27-2019, 02-08-2019, 3-25
-2019, 2-25-2019, 03-09-2019, 02-12-2019, 02-07-2019),
+   total = c(548.9715, 80.22, 340.5255, 489.048, 634.3785, 627.6165, 433.
692, 772.38, 76.146, 172.746)
+   )
> #print table
> print(sanika)
   id           product_line customer unit_price quantity gender branch
city   tax payment gross_margin gross_income
1   1                 Sports   Member      74.69        7 female      A Mu
mbai 26.12    Cash        4.761        26.12
2   2                 Travel   Normal      15.28        5 female      C  V
ashi  3.82 Ewallet        4.761         3.82
3   3 Electronic Accessories   Normal      46.33        7   male      A Pa
nvel 16.21    Card        4.761        16.21
4   4     Home and lifestyle   Member      58.22        8   male      A Mu
mbai 23.28    Cash        4.761        23.28
5   5    Fashion accessories   Normal      86.31        7   male      C  V
ashi 30.20 Ewallet        4.761        30.20
6   6                 Health   Member      85.39        7 female      B Pa
nvel 29.88    Card        4.761        29.88
7   7                 Beauty   Normal      68.84        6 female      B Mu
mbai 20.65    Cash        4.761        20.65
8   8     Food and beverages   Normal      73.56       10   male      C  V
ashi 36.78 Ewallet        4.761        36.78
9   9              Furniture   Member      36.26        2 female      A Pa
nvel 11.73    Card        4.761        11.73
10 10          Books and toys   Normal      54.84        3   male      B Mu
mbai 24.12    Cash        4.761        24.12
   rating  date    total
1     9.1 -2023 548.9715
2     9.6 -2023  80.2200
3     7.4 -2024 340.5255
4     8.4 -2045 489.0480
5     5.3 -2025 634.3785
6     4.1 -2041 627.6165
7     5.8 -2042 433.6920
8     8.0 -2025 772.3800
9     7.2 -2029  76.1460
10    5.9 -2024 172.7460
>
> dataset=sanika
```

```
> a<-dataset
>
> #Give the dimensions of the dataset
> cat ("Dimension:",dim(a))
Dimension: 10 15>
> #Give number of rows in the dataset
> cat ("\nrow:", nrow(a))

row: 10>
> #Give number of columns in the dataset
> cat("\ncolumn:",ncol (a))

column: 15>
> #Questions on functions of dplyr
> #Display sales details arranged in ascending order as per their unit pri
ce
> a%>%arrange (gross_income)->a1
> print (a1)
   id           product_line customer unit_price quantity gender branch
city   tax payment gross_margin gross_income
1   2                 Travel   Normal      15.28        5 female      C  V
ashi  3.82 Ewallet        4.761         3.82
2   9              Furniture   Member      36.26        2 female      A Pa
nvel 11.73    Card        4.761        11.73
3   3 Electronic Accessories   Normal      46.33        7   male      A Pa
nvel 16.21    Card        4.761        16.21
4   7                 Beauty   Normal      68.84        6 female      B Mu
mbai 20.65    Cash        4.761        20.65
5   4     Home and lifestyle   Member      58.22        8   male      A Mu
mbai 23.28    Cash        4.761        23.28
6  10          Books and toys   Normal      54.84        3   male      B Mu
mbai 24.12    Cash        4.761        24.12
7   1                 Sports   Member      74.69        7 female      A Mu
mbai 26.12    Cash        4.761        26.12
8   6                 Health   Member      85.39        7 female      B Pa
nvel 29.88    Card        4.761        29.88
9   5    Fashion accessories   Normal      86.31        7   male      C  V
ashi 30.20 Ewallet        4.761        30.20
10  8     Food and beverages   Normal      73.56       10   male      C  V
ashi 36.78 Ewallet        4.761        36.78
   rating  date    total
1     9.6 -2023  80.2200
2     7.2 -2029  76.1460
3     7.4 -2024 340.5255
4     5.8 -2042 433.6920
5     8.4 -2045 489.0480
6     5.9 -2024 172.7460
7     9.1 -2023 548.9715
8     4.1 -2041 627.6165
9     5.3 -2025 634.3785
10    8.0 -2025 772.3800
>
> #Display sales details arranged in ascending order as per their unit pri
ce
> a%>%arrange (unit_price)->a2
> print (a2)
   id           product_line customer unit_price quantity gender branch
city   tax payment gross_margin gross_income
1   2                 Travel   Normal      15.28        5 female      C  V
ashi  3.82 Ewallet        4.761         3.82
2   9              Furniture   Member      36.26        2 female      A Pa
nvel 11.73    Card        4.761        11.73
3   3 Electronic Accessories   Normal      46.33        7   male      A Pa
nvel 16.21    Card        4.761        16.21
4  10          Books and toys   Normal      54.84        3   male      B Mu
mbai 24.12    Cash        4.761        24.12
5   4     Home and lifestyle   Member      58.22        8   male      A Mu
mbai 23.28    Cash        4.761        23.28
```

```
6   7                 Beauty   Normal      68.84        6 female      B Mu
mbai 20.65    Cash        4.761        20.65
7   8     Food and beverages   Normal      73.56       10   male      C  V
ashi 36.78 Ewallet        4.761        36.78
8   1                 Sports   Member      74.69        7 female      A Mu
mbai 26.12    Cash        4.761        26.12
9   6                 Health   Member      85.39        7 female      B Pa
nvel 29.88    Card        4.761        29.88
10  5    Fashion accessories   Normal      86.31        7   male      C  V
ashi 30.20 Ewallet        4.761        30.20
   rating  date    total
1     9.6 -2023  80.2200
2     7.2 -2029  76.1460
3     7.4 -2024 340.5255
4     5.9 -2024 172.7460
5     8.4 -2045 489.0480
6     5.8 -2042 433.6920
7     8.0 -2025 772.3800
8     9.1 -2023 548.9715
9     4.1 -2041 627.6165
10    5.3 -2025 634.3785
>
> #Display the total gross of the sales
> a %>% mutate(a, Total_gross=gross_income+gross_margin)->a3
> print(a3)
   id           product_line customer unit_price quantity gender branch
city   tax payment gross_margin gross_income
1   1                 Sports   Member      74.69        7 female      A Mu
mbai 26.12    Cash        4.761        26.12
2   2                 Travel   Normal      15.28        5 female      C  V
ashi  3.82 Ewallet        4.761         3.82
3   3 Electronic Accessories   Normal      46.33        7   male      A Pa
nvel 16.21    Card        4.761        16.21
4   4     Home and lifestyle   Member      58.22        8   male      A Mu
mbai 23.28    Cash        4.761        23.28
5   5    Fashion accessories   Normal      86.31        7   male      C  V
ashi 30.20 Ewallet        4.761        30.20
6   6                 Health   Member      85.39        7 female      B Pa
nvel 29.88    Card        4.761        29.88
7   7                 Beauty   Normal      68.84        6 female      B Mu
mbai 20.65    Cash        4.761        20.65
8   8     Food and beverages   Normal      73.56       10   male      C  V
ashi 36.78 Ewallet        4.761        36.78
9   9              Furniture   Member      36.26        2 female      A Pa
nvel 11.73    Card        4.761        11.73
10 10          Books and toys   Normal      54.84        3   male      B Mu
mbai 24.12    Cash        4.761        24.12
   rating  date    total Total_gross
1     9.1 -2023 548.9715      30.881
2     9.6 -2023  80.2200       8.581
3     7.4 -2024 340.5255      20.971
4     8.4 -2045 489.0480      28.041
5     5.3 -2025 634.3785      34.961
6     4.1 -2041 627.6165      34.641
7     5.8 -2042 433.6920      25.411
8     8.0 -2025 772.3800      41.541
9     7.2 -2029  76.1460      16.491
10    5.9 -2024 172.7460      28.881
>
> #To display the id,product line,total gross of female customer with quan
tity > 6
> a3%>%select (id, product_line, gender, quantity, Total_gross)%>%filter(g
ender=="female" & quantity>6 )->a4
> print(a4)
  id product_line gender quantity Total_gross
1  1       Sports female        7      30.881
2  6       Health female        7      34.641
>
```

```
> #Display id and product line of students whose total gross is greater th
an
> a3%>%select(id, product_line, Total_gross)%>%filter (Total_gross>25)->a5
> print (a5)
  id        product_line Total_gross
1  1              Sports      30.881
2  4  Home and lifestyle      28.041
3  5 Fashion accessories      34.961
4  6              Health      34.641
5  7              Beauty      25.411
6  8  Food and beverages      41.541
7 10      Books and toys      28.881
>
> #Display the total sales of the sales
> a %>% mutate(a, Total_sales=unit_price+tax)->a6
> print(a6)
   id           product_line customer unit_price quantity gender branch
city   tax payment gross_margin gross_income
1   1                 Sports   Member      74.69        7 female      A Mu
mbai 26.12    Cash        4.761        26.12
2   2                 Travel   Normal      15.28        5 female      C  V
ashi  3.82 Ewallet        4.761         3.82
3   3 Electronic Accessories   Normal      46.33        7   male      A Pa
nvel 16.21    Card        4.761        16.21
4   4     Home and lifestyle   Member      58.22        8   male      A Mu
mbai 23.28    Cash        4.761        23.28
5   5    Fashion accessories   Normal      86.31        7   male      C  V
ashi 30.20 Ewallet        4.761        30.20
6   6                 Health   Member      85.39        7 female      B Pa
nvel 29.88    Card        4.761        29.88
7   7                 Beauty   Normal      68.84        6 female      B Mu
mbai 20.65    Cash        4.761        20.65
8   8     Food and beverages   Normal      73.56       10   male      C  V
ashi 36.78 Ewallet        4.761        36.78
9   9              Furniture   Member      36.26        2 female      A Pa
nvel 11.73    Card        4.761        11.73
10 10         Books and toys   Normal      54.84        3   male      B Mu
mbai 24.12    Cash        4.761        24.12
   rating  date    total Total_sales
1     9.1 -2023 548.9715      100.81
2     9.6 -2023  80.2200       19.10
3     7.4 -2024 340.5255       62.54
4     8.4 -2045 489.0480       81.50
5     5.3 -2025 634.3785      116.51
6     4.1 -2041 627.6165      115.27
7     5.8 -2042 433.6920       89.49
8     8.0 -2025 772.3800      110.34
9     7.2 -2029  76.1460       47.99
10    5.9 -2024 172.7460       78.96
>
>
> print("Sanika Mhatre")
[1] "Sanika Mhatre"
> timestamp()
##------ Thu Apr 13 03:34:17 2023 ------##
```

Histogram-

A histogram is a graphical representation of the distribution of a numerical variable. It displays the frequencies or relative frequencies of a set of continuous or discrete data. A histogram consists of a series of bars, where each bar represents a range of values and the height of the bar corresponds to the frequency or proportion of data points in that range.

Histograms are commonly used to examine the shape of a distribution, identify any outliers or gaps, and determine the central tendency and variability of a dataset. They are also useful for comparing the distributions of different datasets or subgroups within a dataset.

Syntax: hist(v, main, xlab, xlim, ylim, breaks, col, border)

Parameters:

v: This parameter contains numerical values used in histogram.

main: This parameter main is the title of the chart.

col: This parameter is used to set color of the bars.

xlab: This parameter is the label for horizontal axis.

border: This parameter is used to set border color of each bar.

xlim: This parameter is used for plotting values of x-axis.

ylim: This parameter is used for plotting values of y-axis.

breaks: This parameter is used as width of each bar.

Program:

```
gross_income = c(26.12, 3.82, 16.21, 23.28, 30.20, 29.88, 20.65, 36.78, 11.73, 24.12)

result <-hist(gross_income,

        main = "Histogram of sales",

        xlab = "Gross income",

        col="purple",

        xlim = c(0,30),

        ylim = c(0,2.5),

        border="black",

        breaks=11
```

)

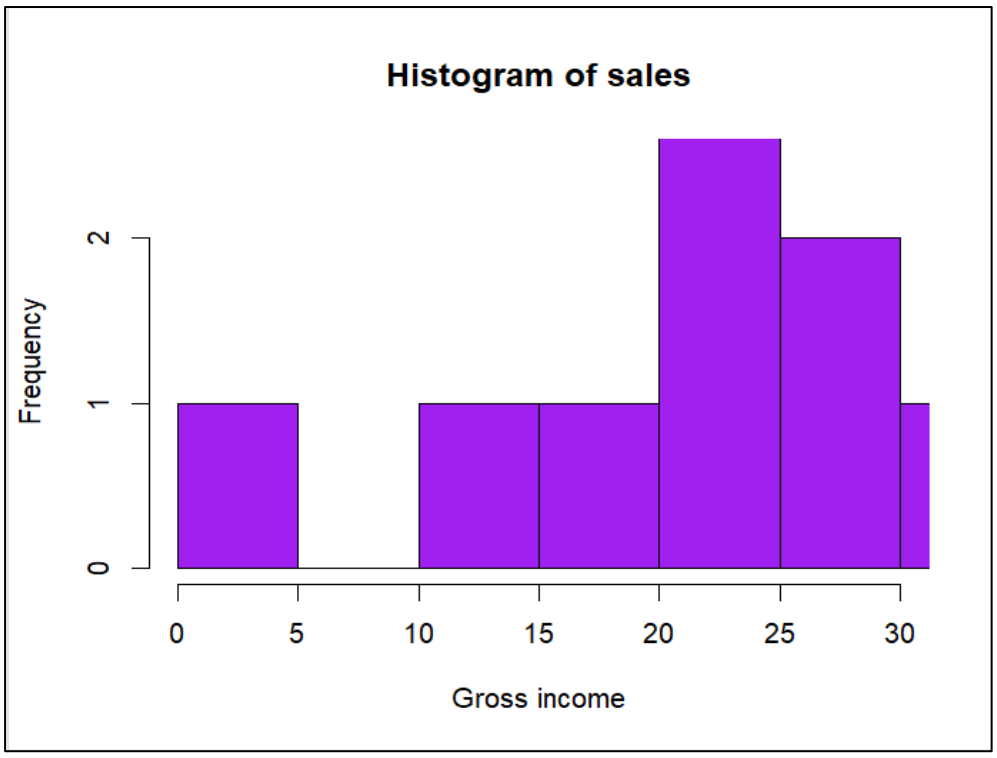
print(result)

print("Sanika Mhatre")

timestamp()

Output:

```
> gross_income = c(26.12, 3.82, 16.21, 23.28, 30.20, 29.88, 20.65, 36.78,
11.73, 24.12)
> result <-hist(gross_income,
+               main = "Histogram of sales",
+               xlab = "Gross income",
+               col="purple",
+               xlim = c(0,30),
+               ylim = c(0,2.5),
+               border="black",
+               breaks=11
+ )
> print(result)
$breaks
[1]  0  5 10 15 20 25 30 35 40

$counts
[1] 1 0 1 1 3 2 1 1

$density
[1] 0.02 0.00 0.02 0.02 0.06 0.04 0.02 0.02

$mids
[1]  2.5  7.5 12.5 17.5 22.5 27.5 32.5 37.5

$xname
[1] "gross_income"

$equidist
[1] TRUE

attr(,"class")
[1] "histogram"
>
>
> print("Sanika Mhatre")
[1] "Sanika Mhatre"
> timestamp()
##------ Thu Apr 13 03:52:06 2023 ------##
```



Scatterplot:

Scatter plots are the graphs that present the relationship between two variables in a data-set. It represents data points on a two-dimensional plane or on a Cartesian system. The independent variable or attribute is plotted on the X-axis, while the dependent variable is plotted on the Y-axis.

Syntax: plot(x, y, main, xlab, ylab, xlim, ylim, axes)
Parameters:

- x: This parameter sets the horizontal coordinates.
- y: This parameter sets the vertical coordinates.
- xlab: This parameter is the label for horizontal axis.
- ylab: This parameter is the label for vertical axis.
- main: This parameter main is the title of the chart.
- xlim: This parameter is used for plotting values of x.
- ylim: This parameter is used for plotting values of y.
- axes: This parameter indicates whether both axes should be drawn on the plot.

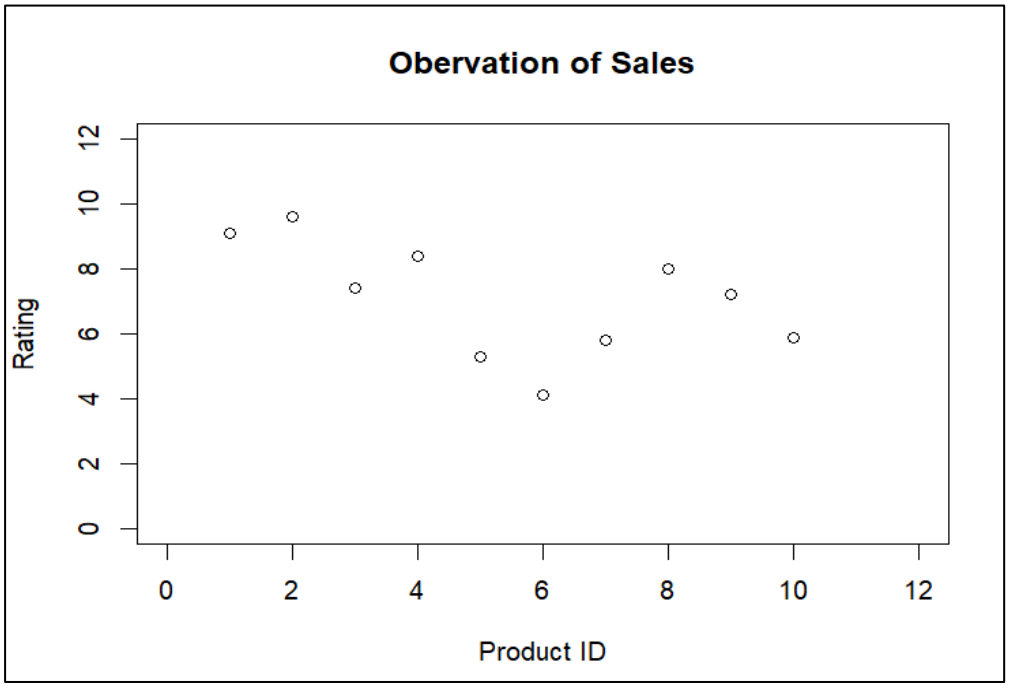
Program:

```
id = c(1,2,3,4,5,6,7,8,9,10)
rating = c(9.1, 9.6, 7.4, 8.4, 5.3, 4.1, 5.8, 8, 7.2, 5.9)
plot(id , rating,
    xlab= "Product ID",
    ylab = "Rating",
    xlim = c(0, 12),
    ylim = c(0, 12),
    main = "Obervation of Sales"
)
```

```
print("Sanika Mhatre")
timestamp()
```

Output:

```
> id = c(1,2,3,4,5,6,7,8,9,10)
> rating = c(9.1, 9.6, 7.4, 8.4, 5.3, 4.1, 5.8, 8, 7.2, 5.9)
> plot(id , rating,
+      xlab= "Product ID",
+      ylab = "Rating",
+      xlim = c(0, 12),
+      ylim = c(0, 12),
+      main = "Obervation of Sales"
+ )
>
> print("Sanika Mhatre")
[1] "Sanika Mhatre"
> timestamp()
##------ Thu Apr 13 04:39:01 2023 ------##
```



Boxplot-

Boxplots are a measure of how well data is distributed across a data set. This divides the data set into three quartiles. This graph represents the minimum, maximum, average, first quartile, and the third quartile in the data set.

A boxplot is a graph that gives us a good indication of how the values in the data are spread out.
Box plots provide some indication of the data's symmetry and skew-ness.

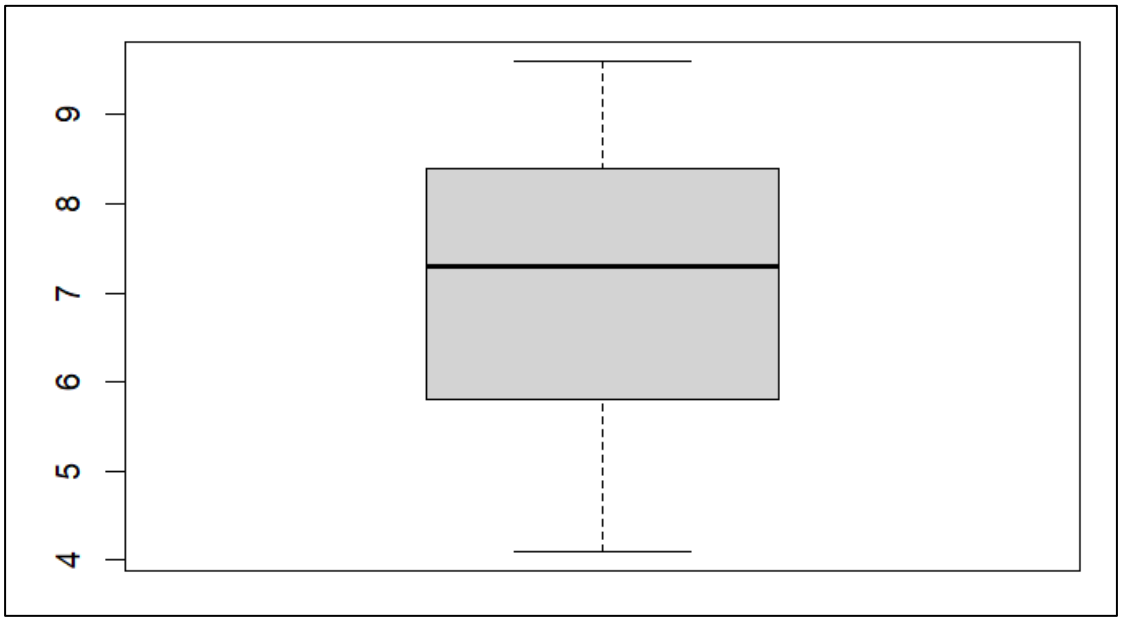
Program-

```
id = c(1,2,3,4,5,6,7,8,9,10)

rating = c(9.1, 9.6, 7.4, 8.4, 5.3, 4.1, 5.8, 8, 7.2, 5.9)


plot (id , rating,
     xlab = "Product ID",
     ylab = "Rating",
     ylim = c(0, 12),
     xlim = c(0, 12),
     main = "Obervation of Sales"
)
print(boxplot(rating))
print("Sanika Mhatre")
timestamp()
```

Output:

```
> id = c(1,2,3,4,5,6,7,8,9,10)
> rating = c(9.1, 9.6, 7.4, 8.4, 5.3, 4.1, 5.8, 8, 7.2, 5.9)
>
> plot (id , rating,
+       xlab = "Product ID",
+       ylab = "Rating",
+       ylim = c(0, 12),
+       xlim = c(0, 12),
+       main = "Obervation of Sales"
+ )
> print(boxplot(rating))
$stats
     [,1]
[1,]  4.1
[2,]  5.8
[3,]  7.3
[4,]  8.4
[5,]  9.6

$n
[1] 10

$conf
         [,1]
[1,] 6.000936
[2,] 8.599064

$out
numeric(0)

$group
numeric(0)
```

```
$names
[1] "1"

> print("Sanika Mhatre")
[1] "Sanika Mhatre"
> timestamp()
##------ Thu Apr 13 04:44:51 2023 ------##
```



Linear Regression-

Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation.

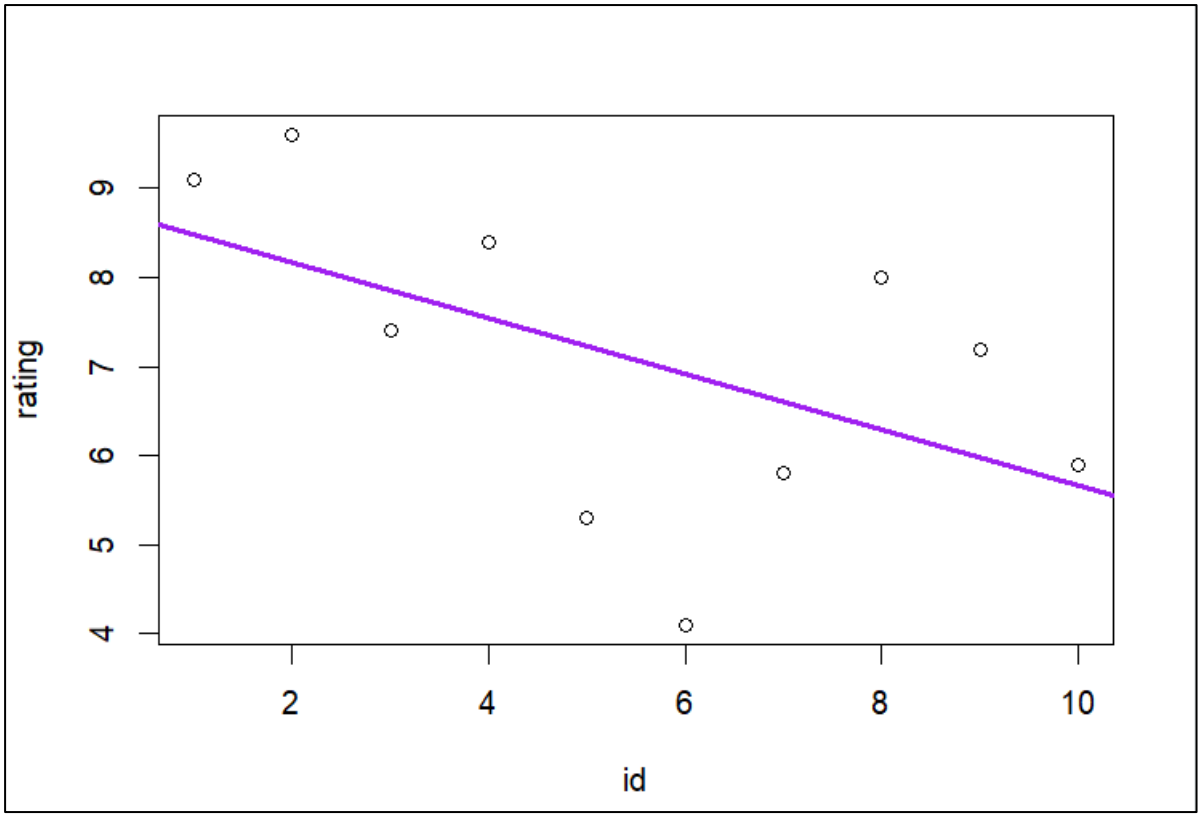
Program-

```
id = c(1,2,3,4,5,6,7,8,9,10)

rating = c(9.1, 9.6, 7.4, 8.4, 5.3, 4.1, 5.8, 8, 7.2, 5.9)


plot (id, rating)

cor (id, rating)


result=lm(rating~id)

print (result)


abline(result, lwd=3,col="purple")
```

```
a=data.frame(id=8)


predict(result,a)


print("Sanika Mhatre")

timestamp()
```

Output:

```
> id = c(1,2,3,4,5,6,7,8,9,10)
> rating = c(9.1, 9.6, 7.4, 8.4, 5.3, 4.1, 5.8, 8, 7.2, 5.9)
>
> plot (id, rating)
> cor (id, rating)
[1] -0.5347429
>
> result=lm(rating~id)
> print (result)

Call:
lm(formula = rating ~ id)

Coefficients:
(Intercept)          id
    8.8000      -0.3127

>
> abline(result, lwd=3,col="purple")
>
> a=data.frame(id=8)
>
> predict(result,a)
       1
6.298182
>
> print("Sanika Mhatre")
[1] "Sanika Mhatre"
> timestamp()
##------ Thu Apr 13 04:59:38 2023 ------##
```


rating
9
8
7
6
5
4
2
4
6
8
10
id