K. J. Somaiya College of Engineering, Mumbai-77
(A Constituent College of Somaiya Vidyavihar University)
Department of Computer Engineering

Batch: A3 **Roll No.:16010121051**

Experiment No. 1

Descriptive Statistics
Title :

Aim: Use R tool to obtain measure of central tendency, dispersion, skewness, Kurtosis, and graphical representation of the given dataset.

Expected Outcome of Experiment:
CO2: Use R Programming to carry out basic statistical modelling and analysis

Books/ Journals/ Websites referred:

1. Data Mining Concepts and Techniques Jiawei Han, Michelin Kamber, Jian Pie, 3rd edition.

Aim of statistical study :

Aim of Statistical study is to be able to conduct research, to be able to infer information critical for applications and to summarize data.

Need of Descriptive statistics:

The term "Descriptive Statistics" refers to the analysis, summary, and presentation of findings related to a data set derived from a sample or entire population.
Descriptive statistics allow for the ease of data visualization. It allows for data to be presented in a meaningful and understandable way, which, in turn, allows or a simplified interpretation of the data set in question. Raw data would be difficult to analyze, and trend and pattern determination may be challenging to perform. In addition, raw data makes it challenging to visualize what the data is showing

Type of attributes :

- Nominal: categories, states, or "names of things"
    - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
    - marital status, occupation, ID numbers, zip codes
- Binary
    - Nominal attribute with only 2 states (0 and 1)
    - Symmetric binary: both outcomes equally important
        - e.g., gender

- Asymmetric binary: outcomes not equally important.
  - e.g., medical test (positive vs. negative)
  - Convention: assign 1 to most important outcome (e.g., HIV positive)
- Ordinal
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small*, *medium*, *large*}, grades, army rankings

- Quantity (integer or real-valued)
  - **Interval**
    - Measured on a scale of **equal-sized units**
    - Values have order
      - E.g., *temperature in C˚or F˚, calendar dates*
    - No true zero-point
  - **Ratio**
    - Inherent **zero-point**
    - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
      - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Measures of Central tendency:

- Mean (algebraic measure):

Note: *n* is sample size and *N* is population size.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):
- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal

For unimodal numeric data that are moderately skewed (asymmetrical), we have the following empirical relation:

$$mean - mode = 3 \times (mean - median)$$

Measures of data dispersion

- Quartiles, outliers and boxplots
  - Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$

- Five number summary: min, Q_1, median, Q_3, max
- Boxplot: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
- Outlier: usually, a value higher/lower than 1.5 x IQR

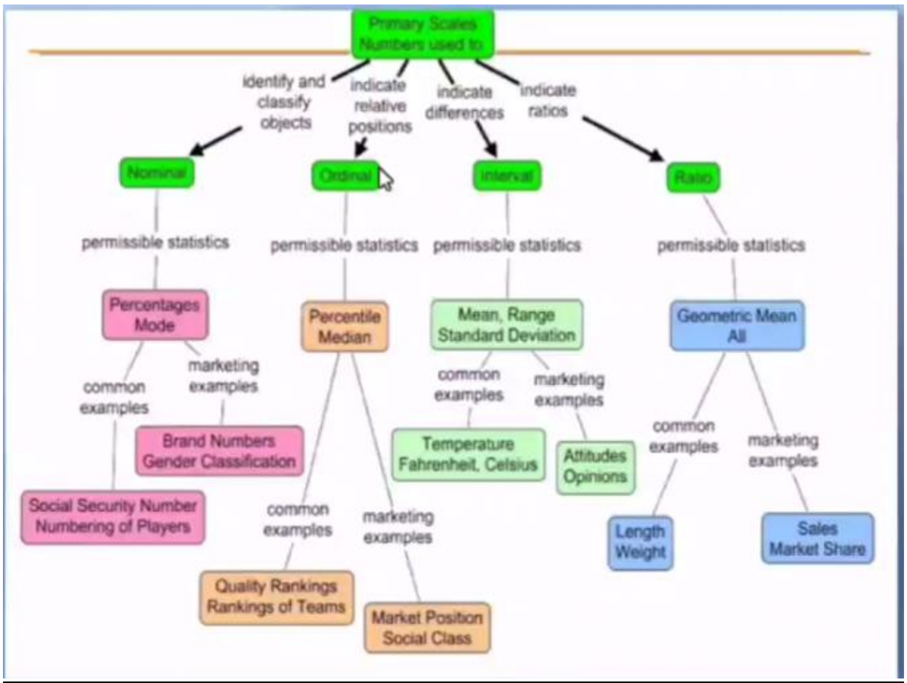
- Variance and standard deviation (*sample: s, population: σ)*
  - Variance: (algebraic, scalable computation)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

The **midrange** can also be used to assess the central tendency of a numeric data set.
It is the average of the largest and smallest values in the set.

Diagrammatic representation of measure of central tendency and operations on different of attributes



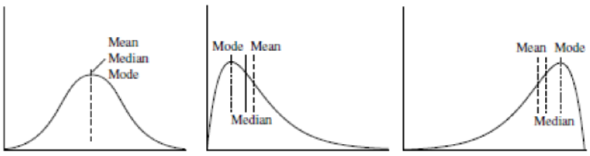
Graphical representation of data

- Boxplot: graphic display of five-number summary
- Histogram: x-axis are values, y-axis repres. frequencies

- Quantile plot: each value x_i is paired with f_i indicating that approximately 100 f_i % of data are δ x_i
- Quantile-quantile (q-q) plot: graphs the quantiles of one univariant distribution against the corresponding quantiles of another
- Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane

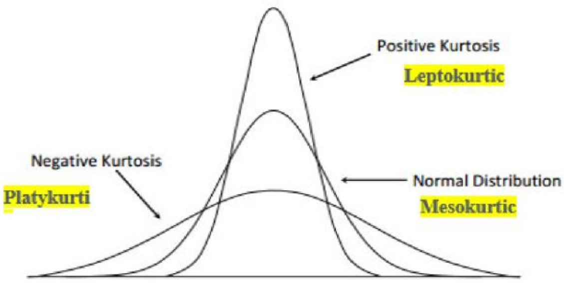
Skewness and Kurtosis

In a unimodal frequency curve with perfect symmetric data distribution, the mean,median, and mode are all at the same center value, as shown in Figure
Data in most real applications are not symmetric. They may instead be either positively skewed, where the mode occurs at a value that is smaller than the median
(or negatively skewed, where the mode occurs at a value greater than the median)



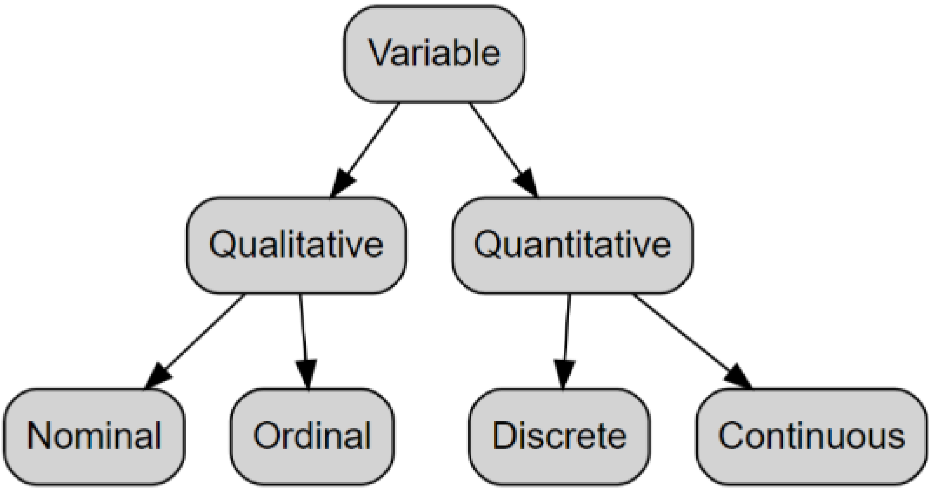
(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.



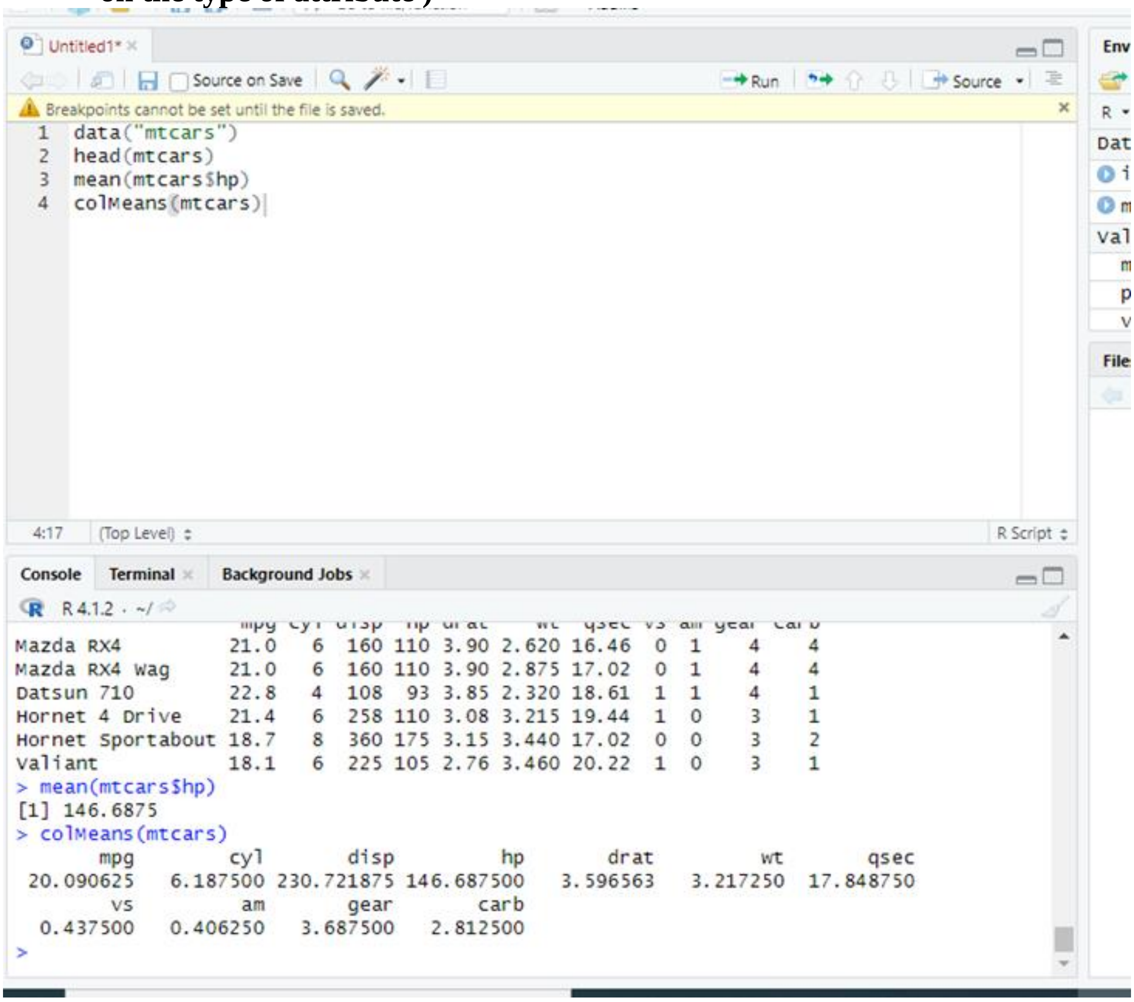
Tasks to be done in Lab

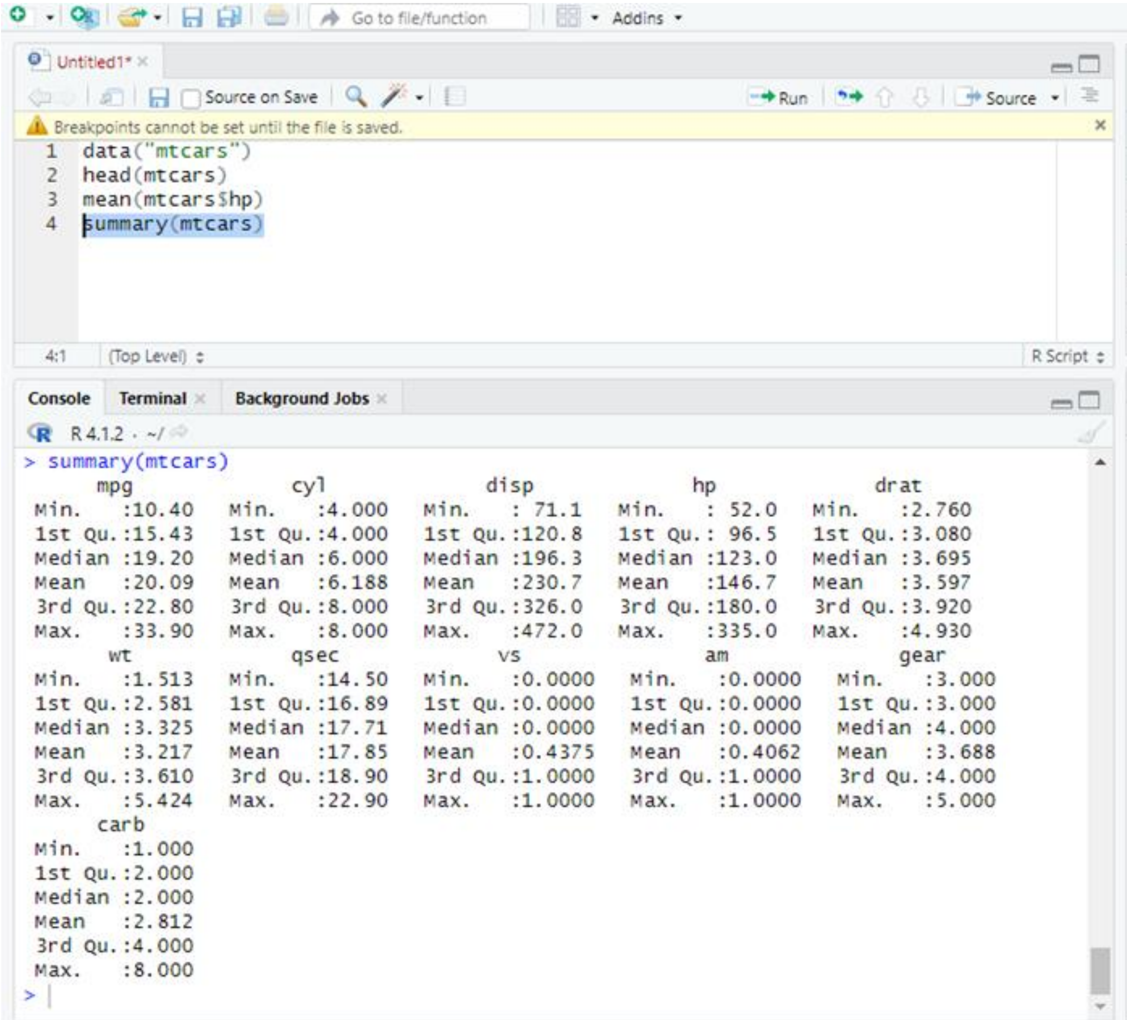
1. **Select any dataset from dataset repository.**
   mtcars selected dataset.
2. **Understand the Types of attributes in dataset (Discrete or continuous, Quantitative or qualitative, Data types-Nominal, Ordinal, interval, ratio-scale)**
   ● Description of variables:
   1. mpg: Miles/(US) gallon
   2. cyl: Number of cylinders
   3. disp: Displacement (cu.in.)
   4. hp: Gross horsepower
   5. drat: Rear axle ratio
   6. wt: Weight (1000 lbs)
   7. qsec: 1/4 mile time
   8. vs: V/S
   9. am: Transmission (0 = automatic, 1 = manual)
   10. gear: Number of forward gears
   11. carb: Number of carburetors

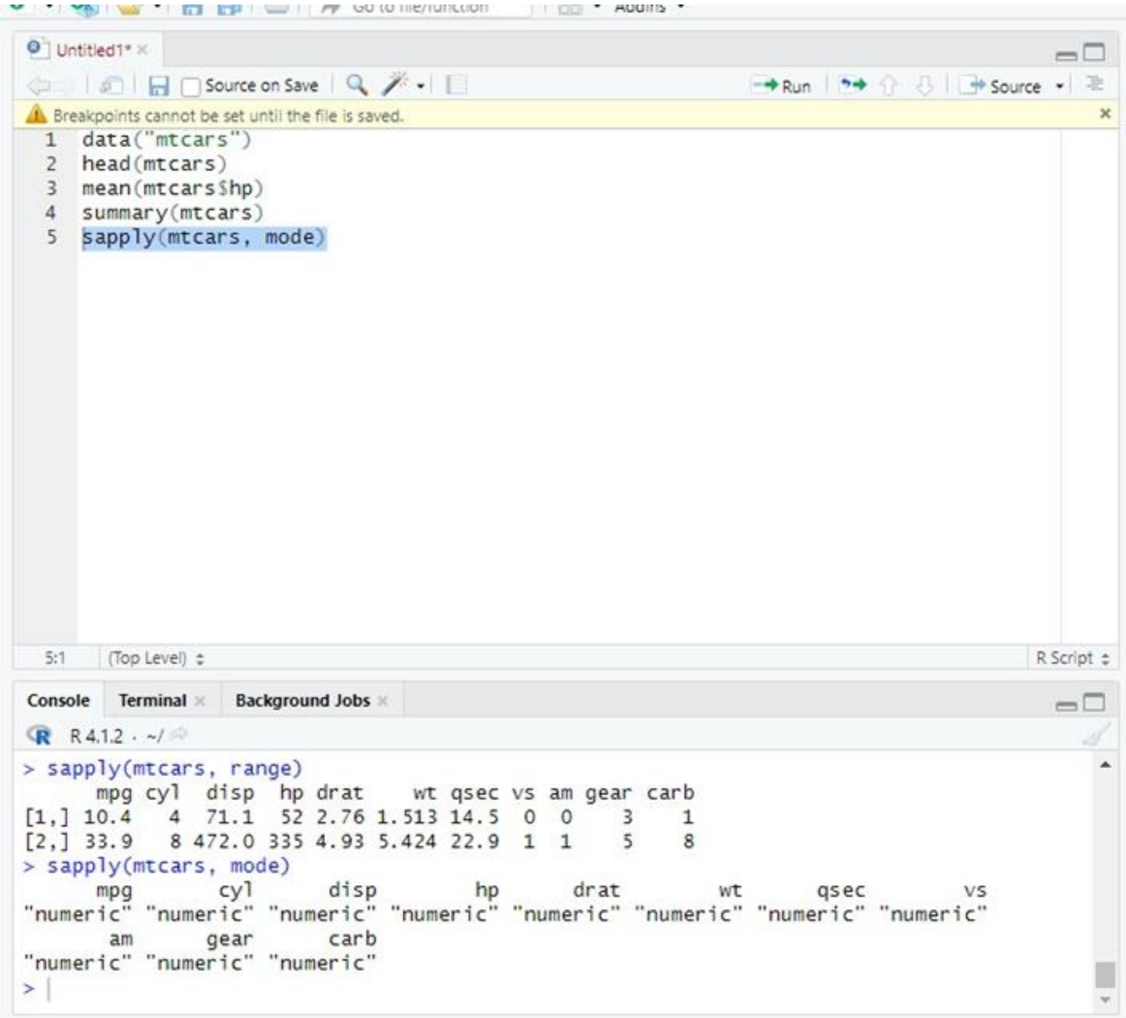


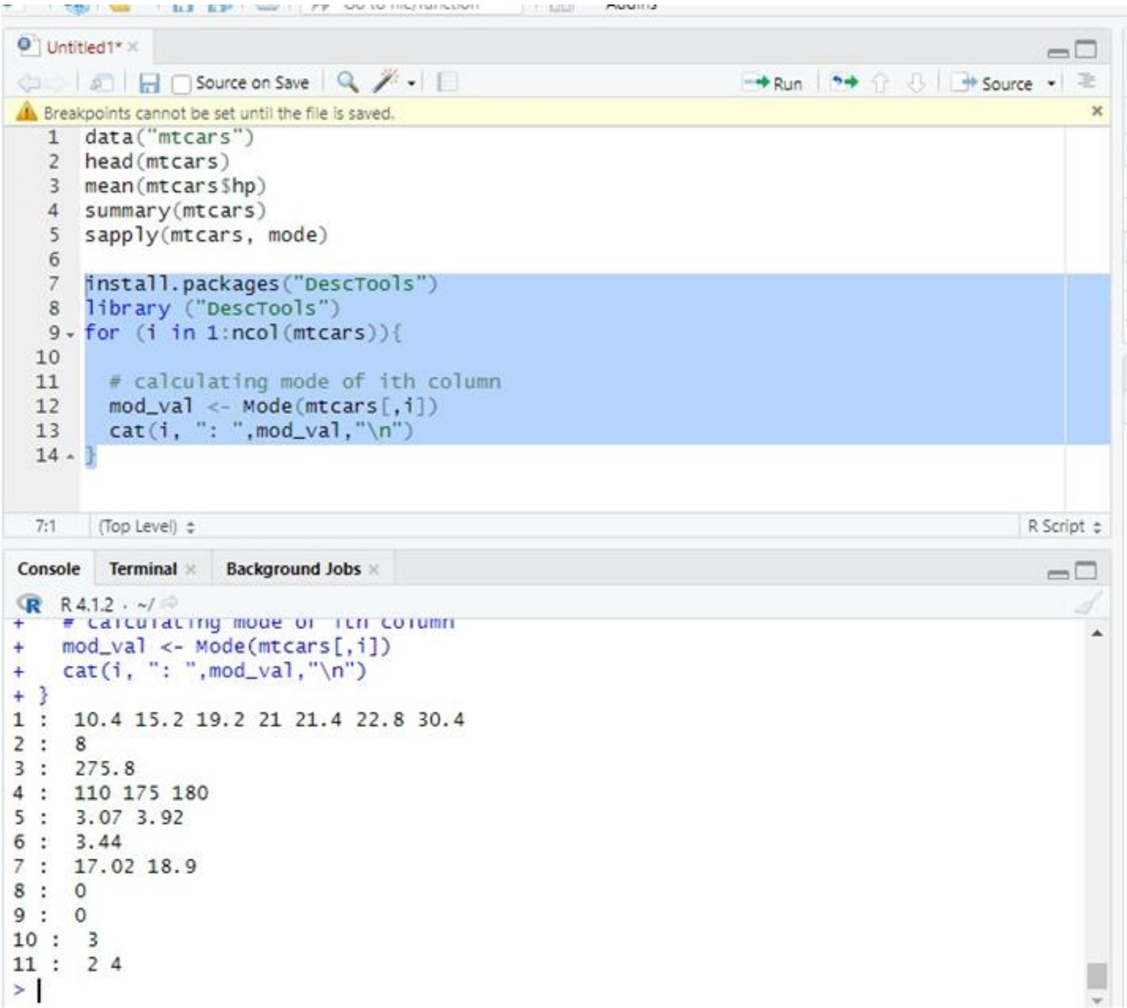
Names of the cars will be qualitative nominal data
● mpg: quantitative continous
● cyl: quantitative discrete
● disp: quantitative continous
● hp: quantitative discrete
● drat: quantitative continous
● wt: quantitative continous
● qsec: quantitative continous
● vs: qualitative nominal
● am: Transmission (0 = automatic, 1 = manual) hence qualitative nominal
● gear: quantitative discrete
● carb: quantitative discrete

3. **Measure of Central tendency (Arithmetic mean, median, Mode depending on the type of attribute)**




Go to file/function
Addins
Untitled1*
Source on Save
Run
Source
Breakpoints cannot be set until the file is saved.
1 data("mtcars")
2 head(mtcars)
3 mean(mtcars$hp)
4 summary(mtcars)
4:1
(Top Level)
R Script
Console
Terminal
Background Jobs
R 4.1.2 · ~/
> summary(mtcars)
mpg cyl disp hp drat
Min. :10.40 Min. :4.000 Min. : 71.1 Min. : 52.0 Min. :2.760
1st Qu.:15.43 1st Qu.:4.000 1st Qu.:120.8 1st Qu.: 96.5 1st Qu.:3.080
Median :19.20 Median :6.000 Median :196.3 Median :123.0 Median :3.695
Mean :20.09 Mean :6.188 Mean :230.7 Mean :146.7 Mean :3.597
3rd Qu.:22.80 3rd Qu.:8.000 3rd Qu.:326.0 3rd Qu.:180.0 3rd Qu.:3.920
Max. :33.90 Max. :8.000 Max. :472.0 Max. :335.0 Max. :4.930
wt qsec vs am gear
Min. :1.513 Min. :14.50 Min. :0.0000 Min. :0.0000 Min. :3.000
1st Qu.:2.581 1st Qu.:16.89 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:3.000
Median :3.325 Median :17.71 Median :0.0000 Median :0.0000 Median :4.000
Mean :3.217 Mean :17.85 Mean :0.4375 Mean :0.4062 Mean :3.688
3rd Qu.:3.610 3rd Qu.:18.90 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:4.000
Max. :5.424 Max. :22.90 Max. :1.0000 Max. :1.0000 Max. :5.000
carb
Min. :1.000
1st Qu.:2.000
Median :2.000
Mean :2.812
3rd Qu.:4.000
Max. :8.000
>


Untitled1*
Source on Save
Run
Source
Breakpoints cannot be set until the file is saved.
1 data("mtcars")
2 head(mtcars)
3 mean(mtcars$hp)
4 summary(mtcars)
5 sapply(mtcars, mode)
5:1
(Top Level)
R Script
Console
Terminal
Background Jobs
R 4.1.2 · ~/
> sapply(mtcars, range)
mpg cyl disp hp drat wt qsec vs am gear carb
[1,] 10.4 4 71.1 52 2.76 1.513 14.5 0 0 3 1
[2,] 33.9 8 472.0 335 4.93 5.424 22.9 1 1 5 8
> sapply(mtcars, mode)
mpg cyl disp hp drat wt qsec vs
"numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
am gear carb
"numeric" "numeric" "numeric"
>



data("mtca

rs")

head(mtcar

s)

mean(mtcar

s$hp)

summary(m

tcars)

sapply(mtc

ars, mode)

install.packages("Des

cTools") library

("DescTools")

for (i in 1:ncol(mtcars)){

calculating mode of

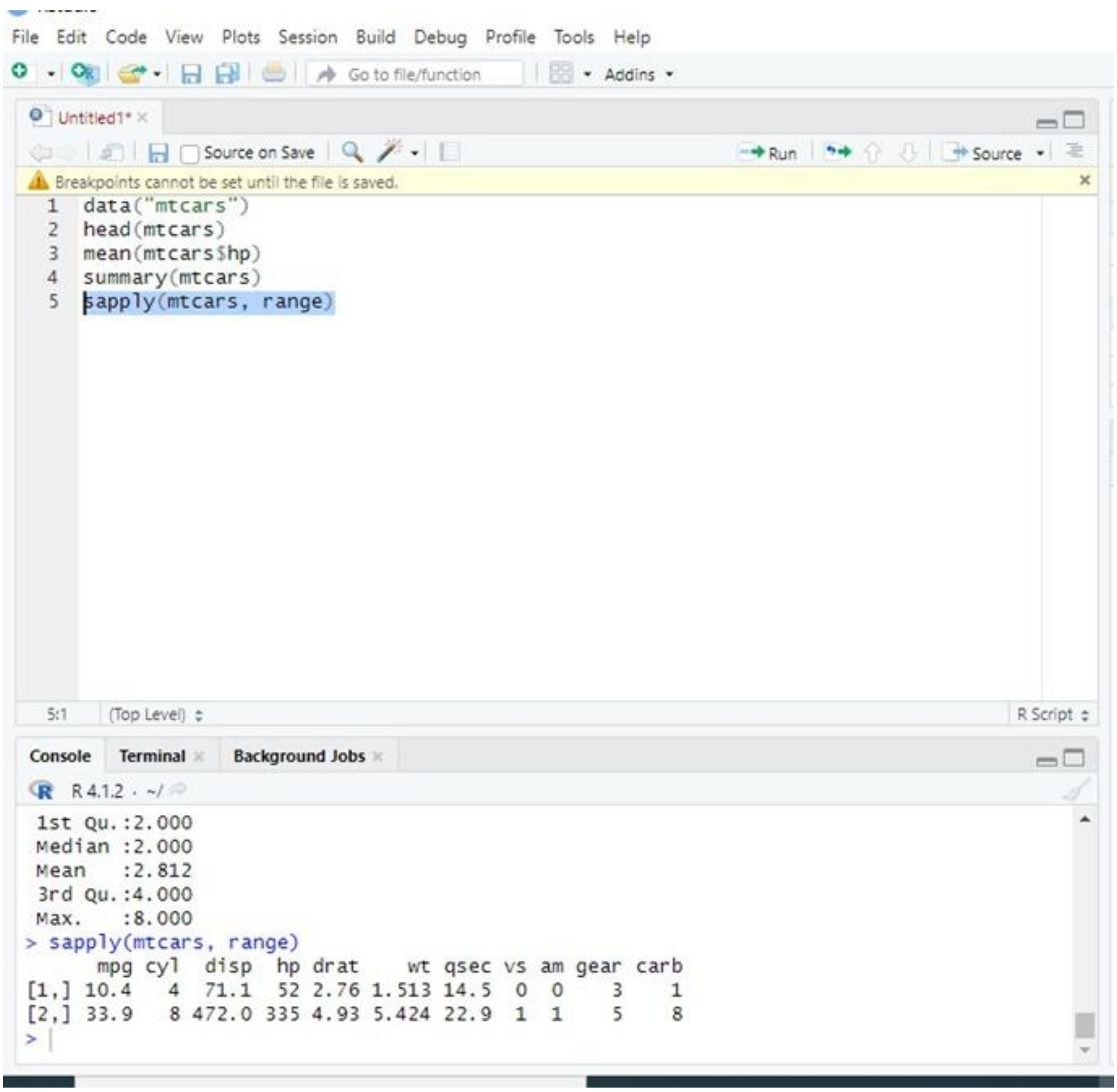
ith column mod_val <-

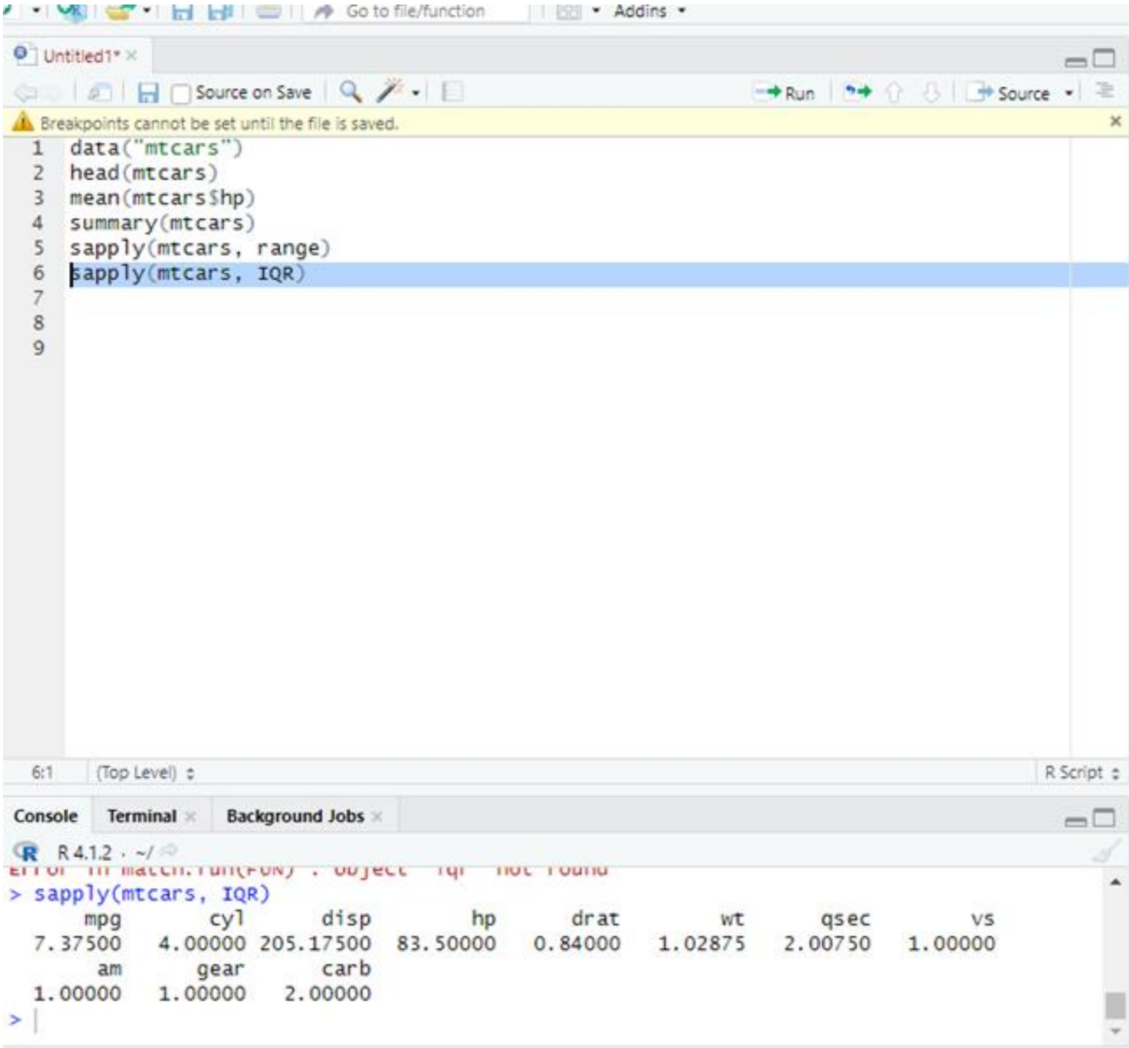
Mode(mtcars[,i]) cat(i,

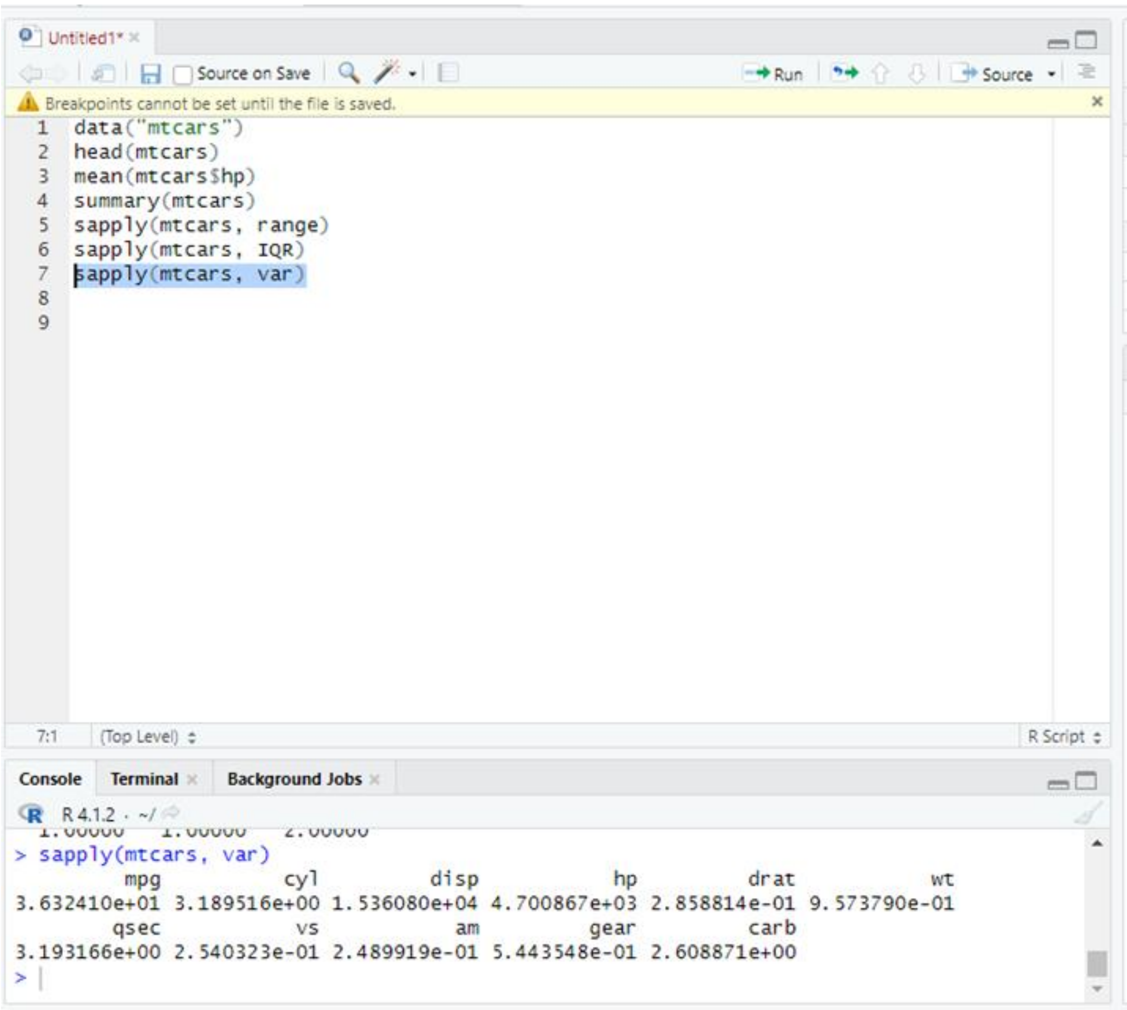
": ",mod_val,"\n")

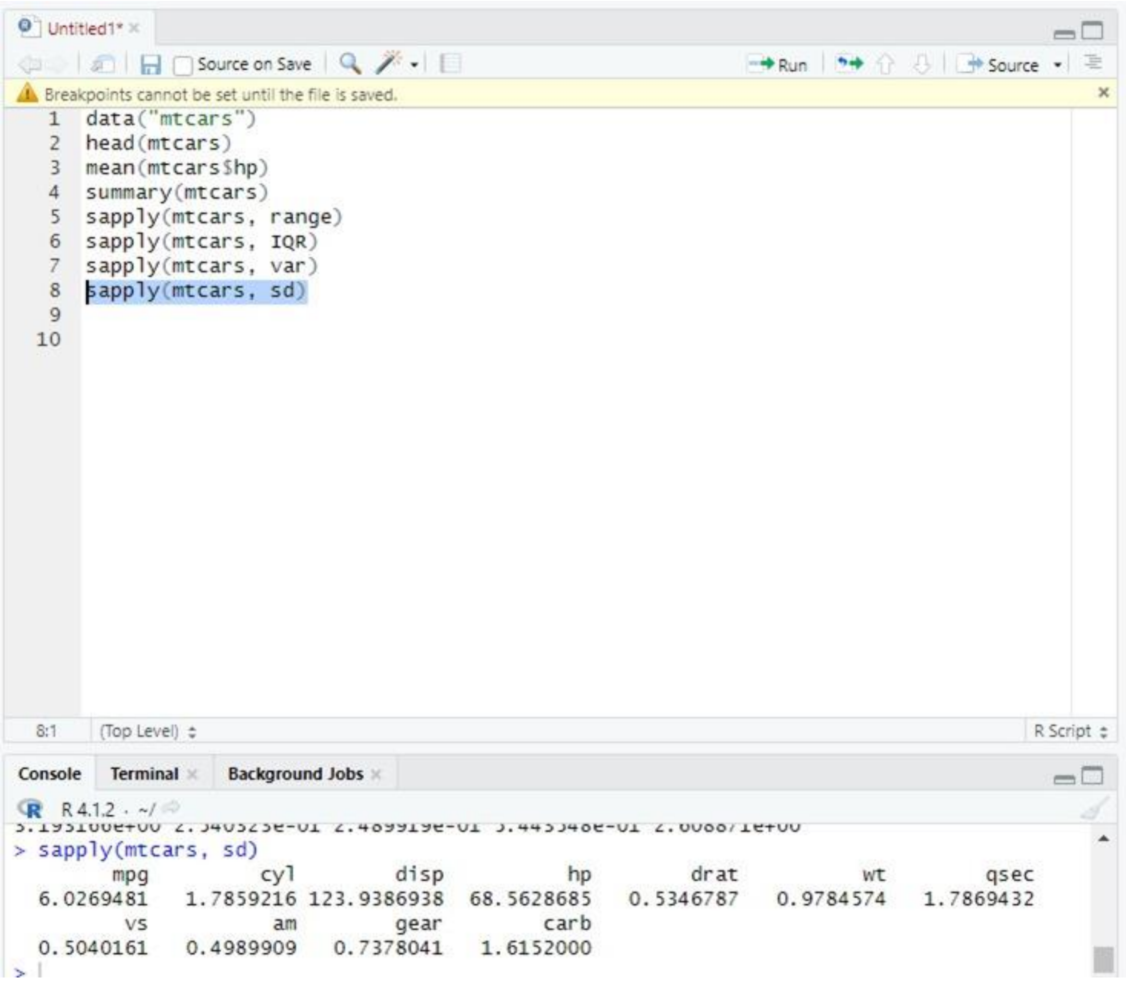
}

4. **Measures of data dispersion (Range, Interquartile range, variance, Standard deviation)**


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Untitled1*
Source on Save
Run
Source
Breakpoints cannot be set until the file is saved.
1 data("mtcars")
2 head(mtcars)
3 mean(mtcars$hp)
4 summary(mtcars)
5 sapply(mtcars, range)
5:1
(Top Level)
R Script
Console
Terminal
Background Jobs
R 4.1.2 · ~/
1st Qu.:2.000
Median :2.000
Mean :2.812
3rd Qu.:4.000
Max. :8.000
> sapply(mtcars, range)
mpg cyl disp hp drat wt qsec vs am gear carb
[1,] 10.4 4 71.1 52 2.76 1.513 14.5 0 0 3 1
[2,] 33.9 8 472.0 335 4.93 5.424 22.9 1 1 5 8
>


Go to file/function
Addins
Untitled1*
Source on Save
Run
Source
Breakpoints cannot be set until the file is saved.
data("mtcars")
head(mtcars)
mean(mtcars$hp)
summary(mtcars)
sapply(mtcars, range)
sapply(mtcars, IQR)
6:1
(Top Level)
R Script
Console
Terminal
Background Jobs
R 4.1.2 · ~/
> sapply(mtcars, IQR)
mpg cyl disp hp drat wt qsec vs
7.37500 4.00000 205.17500 83.50000 0.84000 1.02875 2.00750 1.00000
am gear carb
1.00000 1.00000 2.00000
>


Untitled1*
Source on Save
Run
Source
Breakpoints cannot be set until the file is saved.
data("mtcars")
head(mtcars)
mean(mtcars$hp)
summary(mtcars)
sapply(mtcars, range)
sapply(mtcars, IQR)
sapply(mtcars, var)
7:1
(Top Level)
R Script
Console
Terminal
Background Jobs
R 4.1.2 · ~/
> sapply(mtcars, var)
mpg cyl disp hp drat wt
3.632410e+01 3.189516e+00 1.536080e+04 4.700867e+03 2.858814e-01 9.573790e-01
qsec vs am gear carb
3.193166e+00 2.540323e-01 2.489919e-01 5.443548e-01 2.608871e+00
>



data

("m

tcar

s")

hea

d(m

tcar

s)

sapply(mtca

rs, range)

sapply(mtca

rs, IQR)

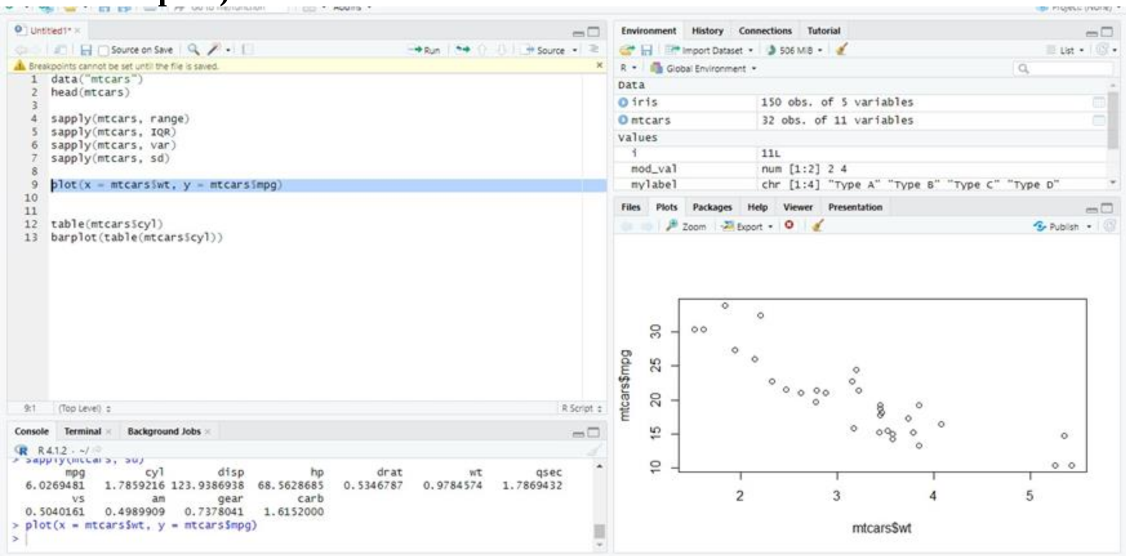
sapply(mtca

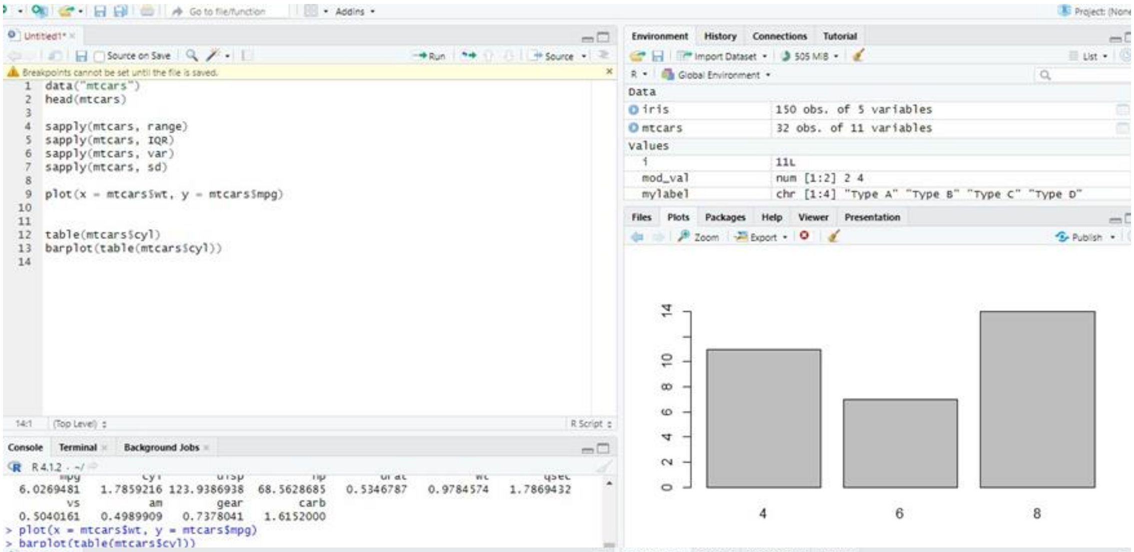
rs, var)

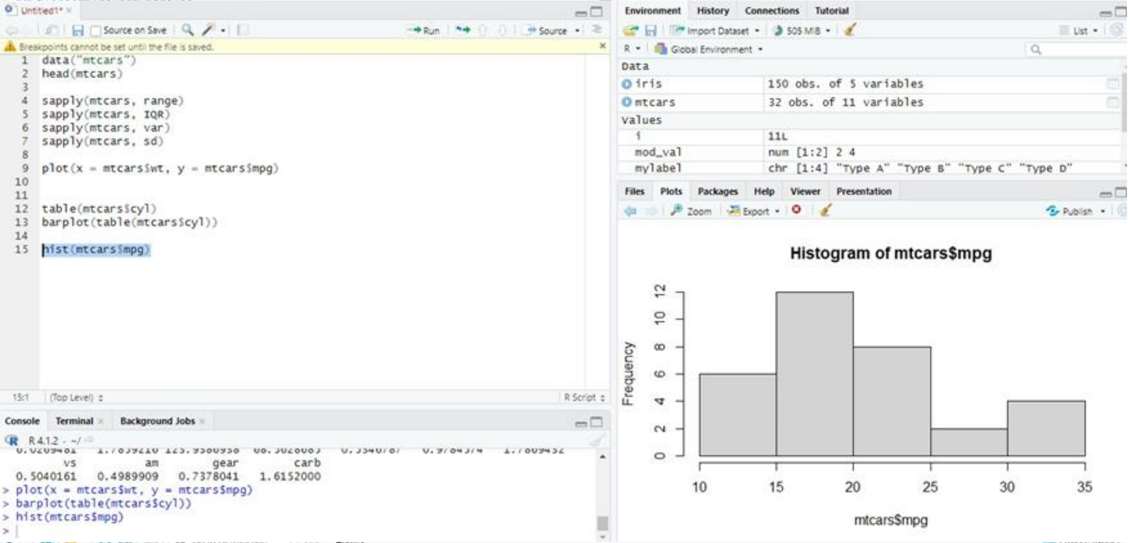
sapply(mtca

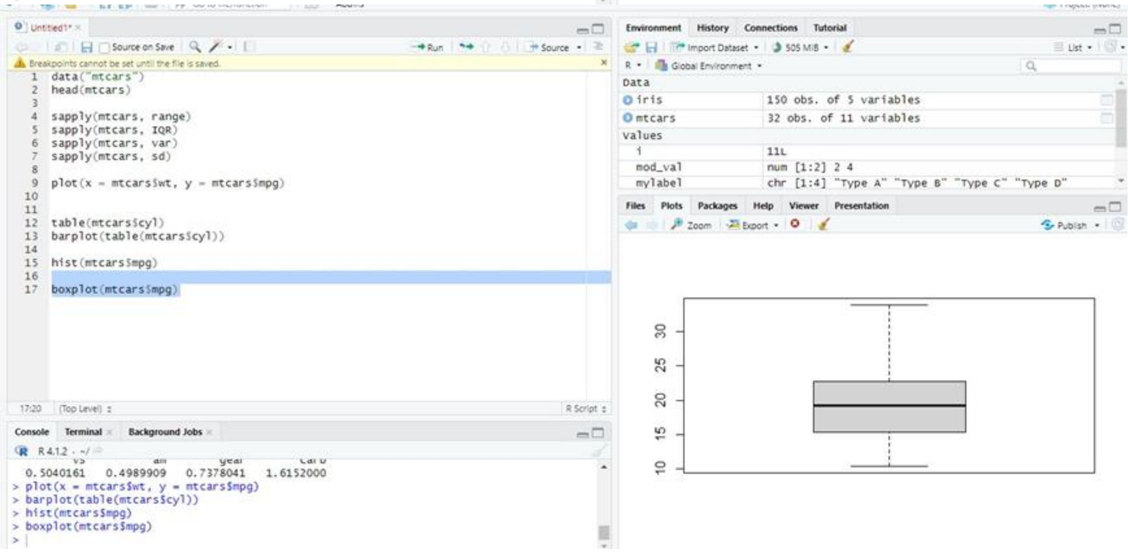
rs, sd)

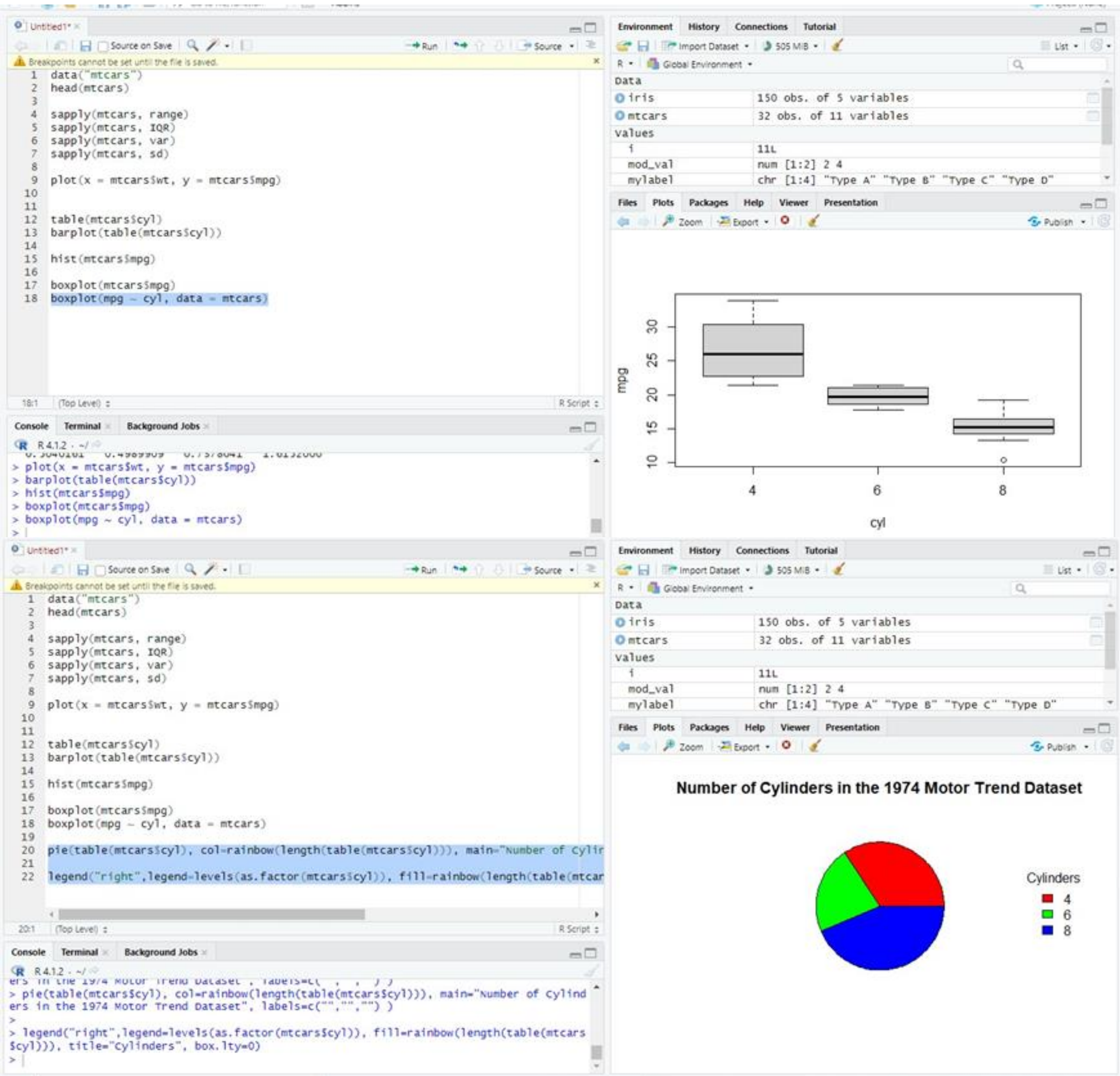
5. **Diagrammatic representation of data (Bar graphs, Pie diagram, Histogram, Box plot)**




data("mtcars")
head(mtcars)
sapply(mtcars, range)
sapply(mtcars, IQR)
sapply(mtcars, var)
sapply(mtcars, sd)
plot(x = mtcars$wt, y = mtcars$mpg)
table(mtcars$cyl)
barplot(table(mtcars$cyl))
iris 150 obs. of 5 variables
mtcars 32 obs. of 11 variables
mod_val num [1:2] 2 4
mylabel chr [1:4] "Type A" "Type B" "Type C" "Type D"
4
6
8


hist(mtcars$mpg)
Histogram of mtcars$mpg
Frequency
mtcars$mpg


boxplot(mtcars$mpg)



data

("m

tcar

s")

hea

d(m

tcar

s)

sapply(mtca

rs, range)

sapply(mtca

rs, IQR)

sapply(mtca

rs, var)

sapply(mtca

rs, sd)

plot(x = mtcars$wt, y = mtcars$mpg)

table(mtcars$c

yl)

barplot(table(

mtcars$cyl))

hist(mtcars$mpg)

boxplot(mtcars$mpg)

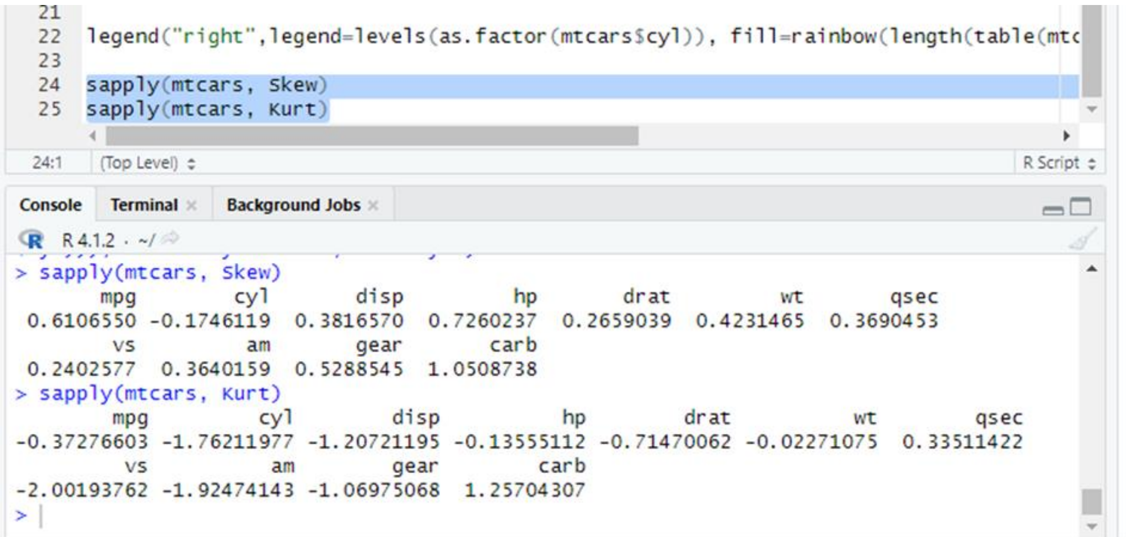
boxplot(mpg ~ cyl,

data = mtcars)

pie(table(mtcars$cyl), col=rainbow(length(table(mtcars$cyl))), main="Number of Cylinders in the 1974 Motor Trend Dataset", labels=c("","",""))

legend("right",legend=levels(as.factor(mtcars$cyl)), fill=rainbow(length(table(mtcars$cyl))), title="Cylinders", box.lty=0)

6. Measure of Skewness and Kurtosis



```
legend("right",legend=levels(as.factor(mtcars$cyl)), fill=rainbow(length(table(mtc

sapply(mtcars, Skew)
sapply(mtcars, Kurt)
```

```
> sapply(mtcars, Skew)
      mpg        cyl       disp         hp       drat         wt       qsec
0.6106550 -0.1746119  0.3816570  0.7260237  0.2659039  0.4231465  0.3690453
       vs         am       gear       carb
0.2402577  0.3640159  0.5288545  1.0508738
> sapply(mtcars, Kurt)
        mpg         cyl        disp          hp        drat          wt        qsec
-0.37276603 -1.76211977 -1.20721195 -0.13555112 -0.71470062 -0.02271075  0.33511422
         vs          am        gear        carb
-2.00193762 -1.92474143 -1.06975068  1.25704307
>
```

sapply(m

tcars,

Skew)

sapply(m

tcars,

Kurt)

Implementation using R tool:
For a given dataset/s show the code along with snapshot of output for the above tasks done in the lab. Also mention the interpretation of measures or the graph

Description of Data set used:
Title: mtcars

Source: Originally sourced from the 1974 Motor Trend magazine, Built-in dataset of Rcode.

Number of instances: 32
Number of attributes: 11

Attribute information :

1. **mpg: Miles/(US) gallon**
2. **cyl: Number of cylinders**
3. **disp: Displacement (cu.in.)**
4. **hp: Gross horsepower**
5. **drat: Rear axle ratio**
6. **wt: Weight (1000 lbs)**
7. **qsec: 1/4 mile time**

8. vs: V/S

9. am: Transmission (0 = automatic, 1 = manual)

10. gear: Number of forward gears

11. carb: Number of carburetors

Conclusion:

We have successfully implemented the experiment and learnt how to perform the above mentioned tasks using R code and importing a dataset to perform various analysis on it.

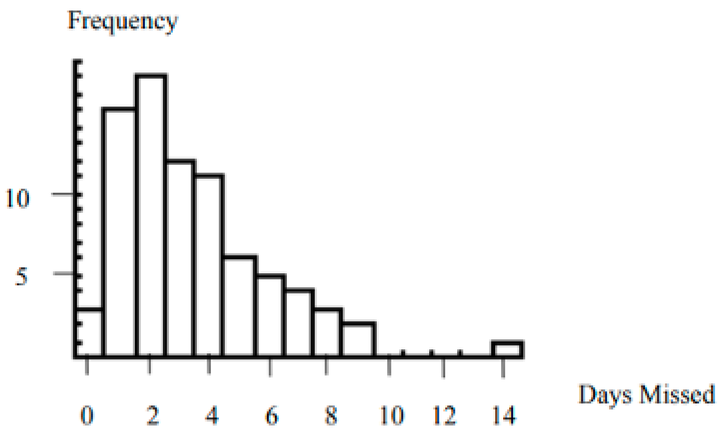
Post lab Questions:

1. In descriptive statistics the aim is to (multiple options may be correct) :
   I) Do the analysis of data that helps describe or show data in a meaningful way such that, for example, patterns might emerge from the data.
   II) Use of probability theory to learn about population from sample data
   III) Quantitatively describe or summarize the data
   IV) Describe the data by measures such as central tendency and measure of variability

   All of the options are correct.

   In descriptive statistics, the aim is to do the analysis of data that helps describe or show data in a meaningful way such that, for example, patterns might emerge from the data. Additionally, it involves the use of probability theory to learn about population from sample data, as well as to describe the data by measures such as central tendency and measure of variability.

2. A boeing 747 aircaraft gets cancelled while severe snowstorms. The following histogram shows the number of days missed(per year) in last 75 years. Which of the following you use as measure to describe the center of the distribution.



   I) Mean, because it covers information from all 75 years
   II) IQR because it is unaffected by the outliers.
   III) Median because the distribution is skewed to the right
   IV) Standard deviation, because it is unaffected by outliers and the distribution is skewed.

3. In a given data set of 100 observation, if the largest value is doubled, which of the following option is/are false (assume the largest value is non-Zero) (multiple option may be correct)
   I) The variance increases
   II) The mean increases
   III) The median increases
   IV) The IQR increases.

They will all increase.

4. Refer the data set "seatbelts.csv". Load the data set into your R workspace and answer the questions.

   The data set contains data about the road casualties in Great Britain between 1969 and 1984.

   The description of the dataset is given below: The 'Seatbelts' data set in R is a multiple time-series data set that was commissioned by the Department of Transport in 1984 to measure differences in deaths before and after front seat belt legislation was introduced on 31st January 1983. It provides monthly total numerical data on a number of incidents including those related to death and injury in Road Traffic Accidents (RTA's). The data set starts in January 1969 and observations run until December 1984.

Variable name	Description
Year	Year of the incident
Month	Month of the incident
DriversKilled	Number of car drivers killed
drivers	Total number of drivers
front	Number of front-seat passengers killed or seriously injured.
rear	Number of rear-seat passengers killed or seriously injured.
kms	Total number of distances driven
PetrolPrice	Petrol price
VanKilled	number of van ('light goods vehicle') drivers killed
law	0/1: was the law in effect that month?

The average number of car drivers killed after the law was in effect is ____?

I) 90
II) 85
II) 100
III) None of the above

5. Referring the data set “seatbelts.csv”. How many front seat passengers were injured or killed in the year 1984
   I) 7041
   II) 7047
   III) 7865
   IV) None of the above

6. Referring the data set “seatbelts.csv”. Calculate the variance for the variables “front” and “rear” and choose the correct option.
   I) Variance of front seat passengers is equal to variance of rear seat passengers.
   II) Variance of front seat passengers is greater than variance of rear seat passengers.
   III) Variance of front seat passengers is less than the variance of rear seat passengers.
   IV)
   V) None of the above

7. Referring the data set “seatbelts.csv”. Maximum kms driven by the driver is ___?
   I) 21626
   II) 17203
   III) 25245
   IV) None of the above

8. Which of the following statement is not true about histograms?
   I) Represent the frequency distribution of categorical variables
   II) It is a graphical representation of data using bars of different heights
   III) Groups numbers into ranges and the height of each bar depicts the frequency of each range or bin
   IV) Represent the frequency distribution of numerical variables