

| | 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

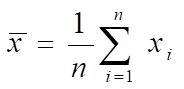
**Type of attributes :**

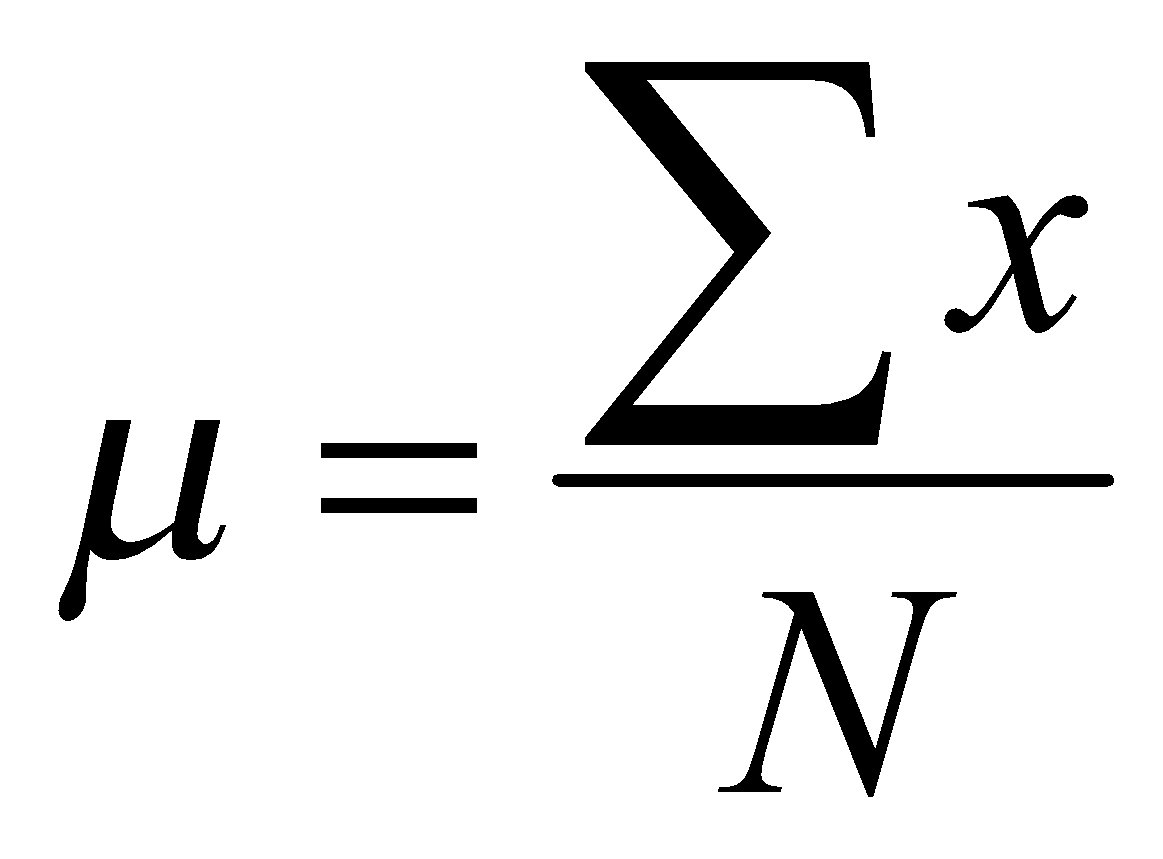
* Nominal: categories, states, or “names of things”
  + *Hair\_color =* {*auburn, black, blond, brown, grey, red, white*}
  + marital status, occupation, ID numbers, zip codes
* Binary
  + Nominal attribute with only 2 states (0 and 1)
  + Symmetric binary: both outcomes equally important
    - e.g.,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





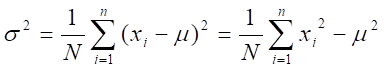
* Median:
  + Middle value if odd number of values, or average of the middle two values otherwise
  + Estimated by interpolation (for *grouped data*):
* Mode
  + Value that occurs most frequently in the data
  + Unimodal, bimodal, trimodal

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



**Measures of data dispersion**

* Quartiles, outliers and boxplots
  + Quartiles: Q1 (25th percentile), Q3 (75th percentile)
  + Inter-quartile range: IQR = Q3 –Q1
  + Five number summary: min, Q1, median,Q3, 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**)**

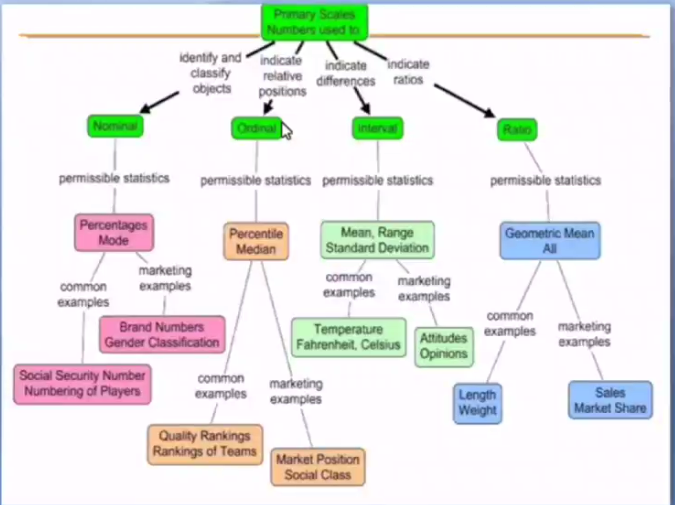




The **midrange** can also be used to assess the central tendency of a numeric data set.

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 *xi*is paired with *fi*  indicating that approximately 100 *fi* % of data are ≤ *xi*
* 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

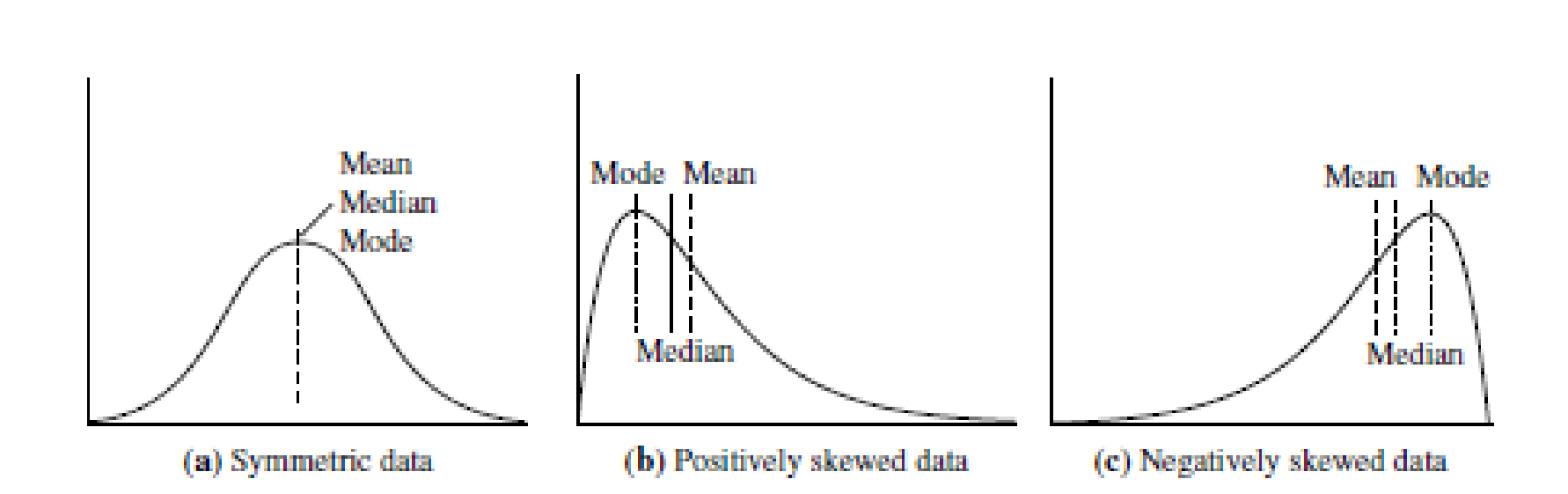
**Skewness and Kurtosis**

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)

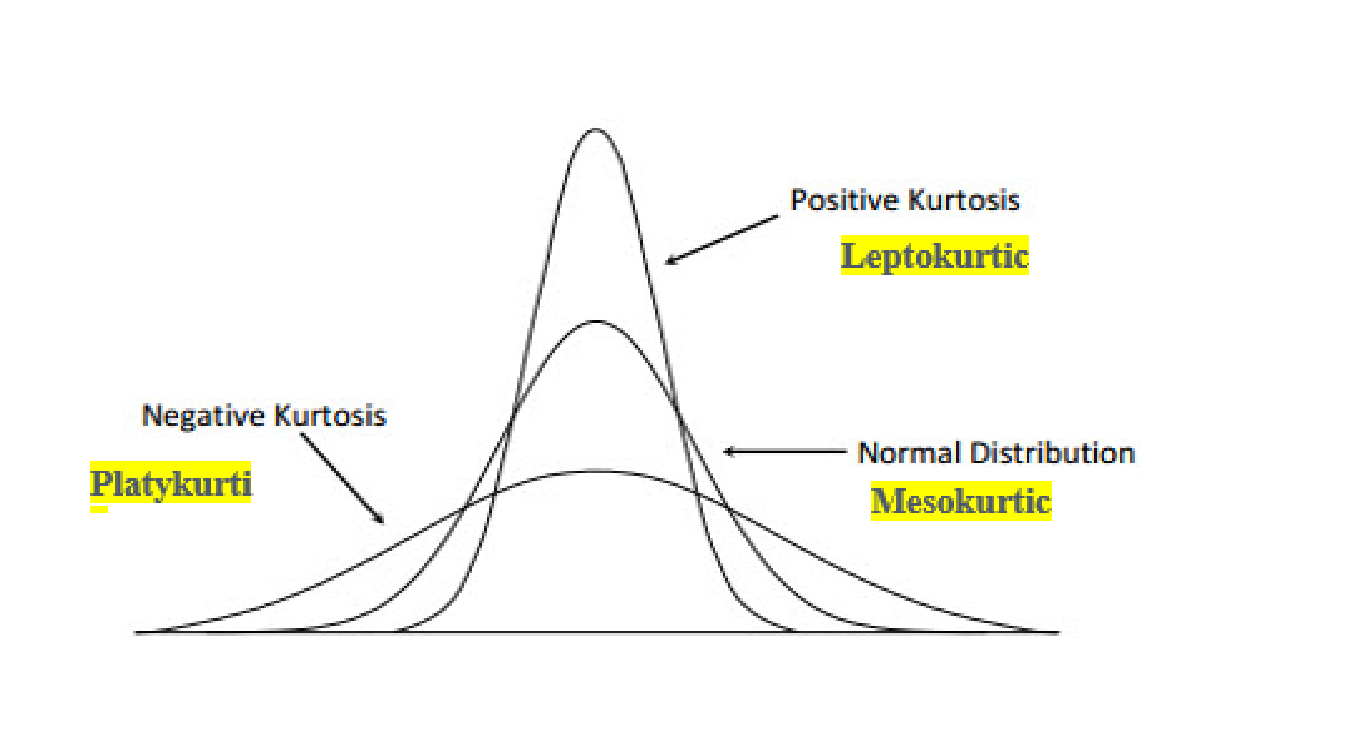


Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative

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.

1. **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)

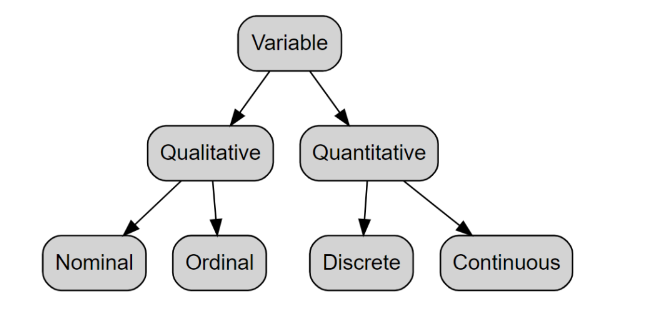
7. qsec: 1/4 mile time

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

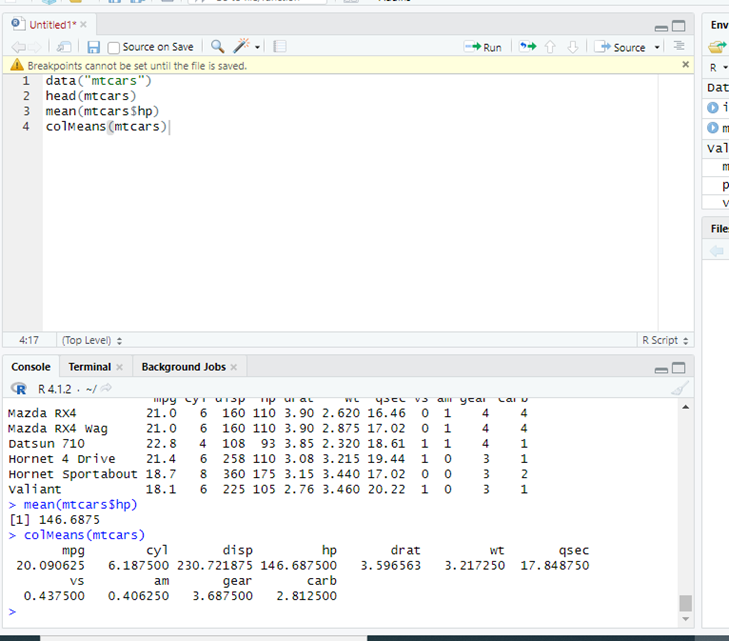
● vs: qualitative nominal

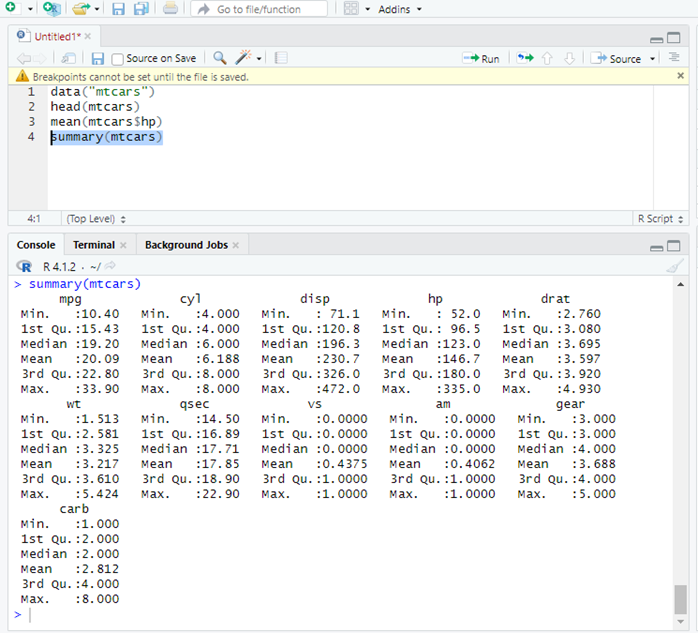
● am: Transmission (0 = automatic, 1 = manual) hence qualitative nominal

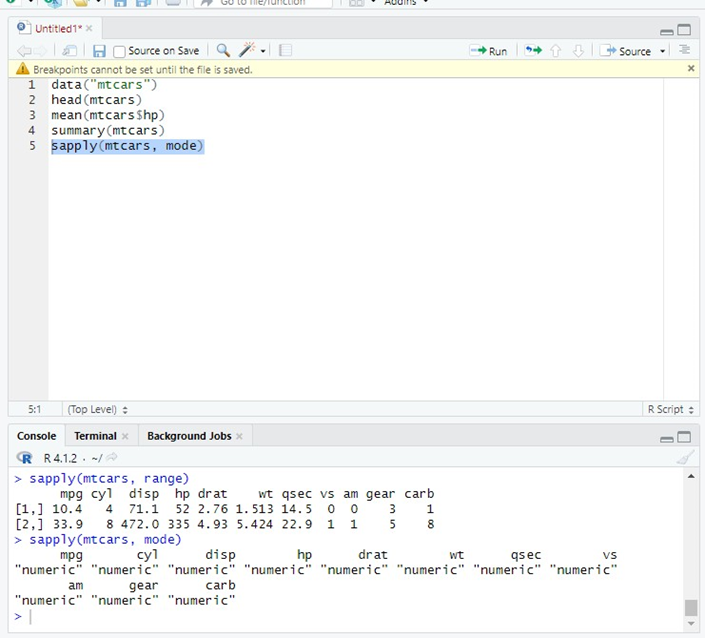
● gear: quantitative discrete

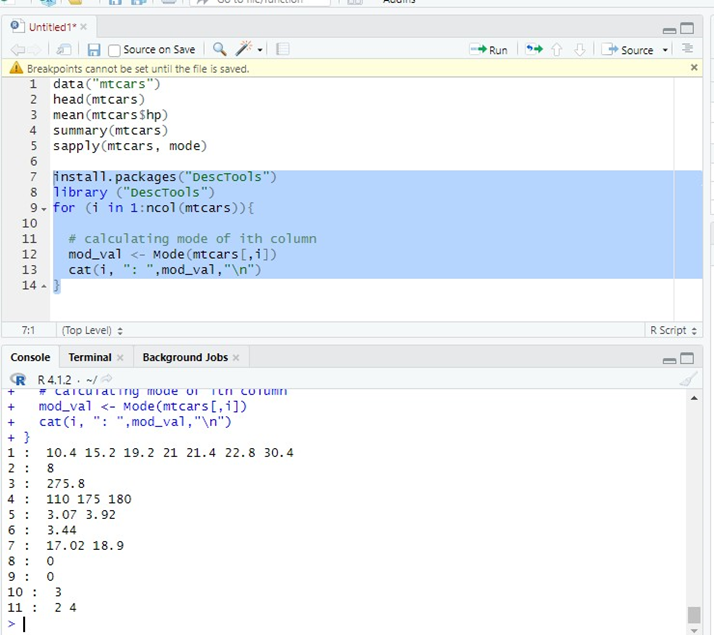
● carb: quantitative discrete

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

****

****

****

****

data("mtcars") head(mtcars) mean(mtcars$hp) summary(mtcars) sapply(mtcars, mode)

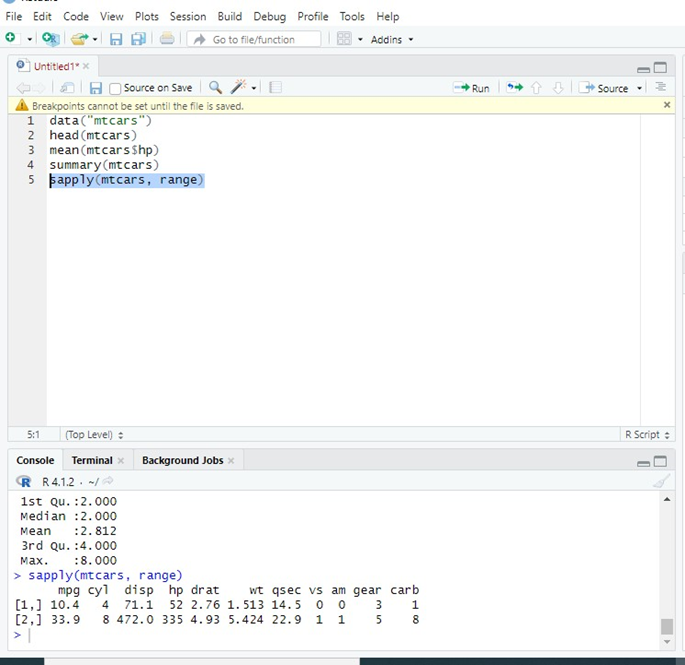
install.packages("DescTools") library ("DescTools")

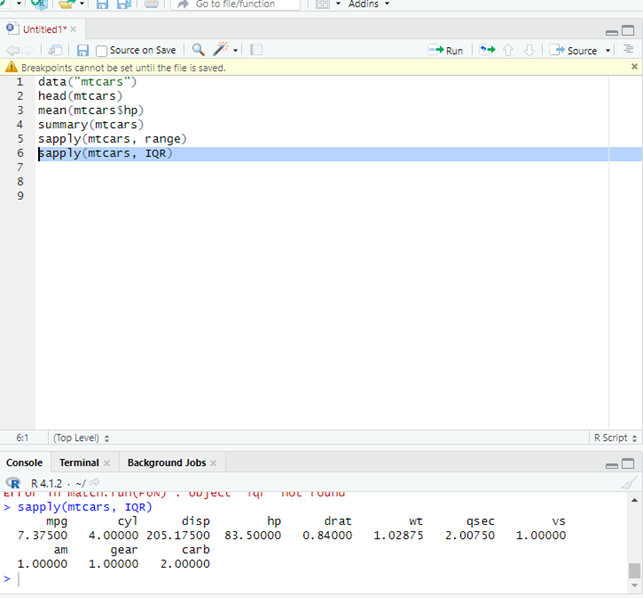
for (i in 1:ncol(mtcars)){

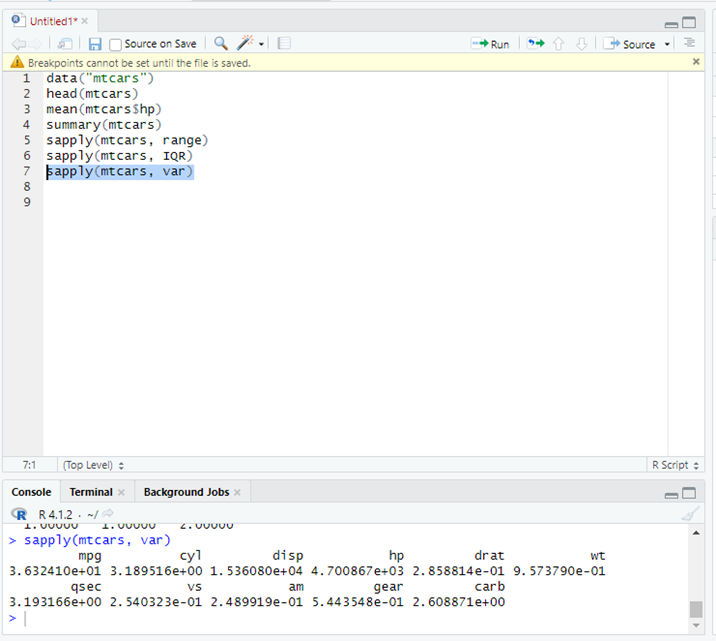
# calculating mode of ith column mod\_val <- Mode(mtcars[,i]) cat(i, ": ",mod\_val,"\n")

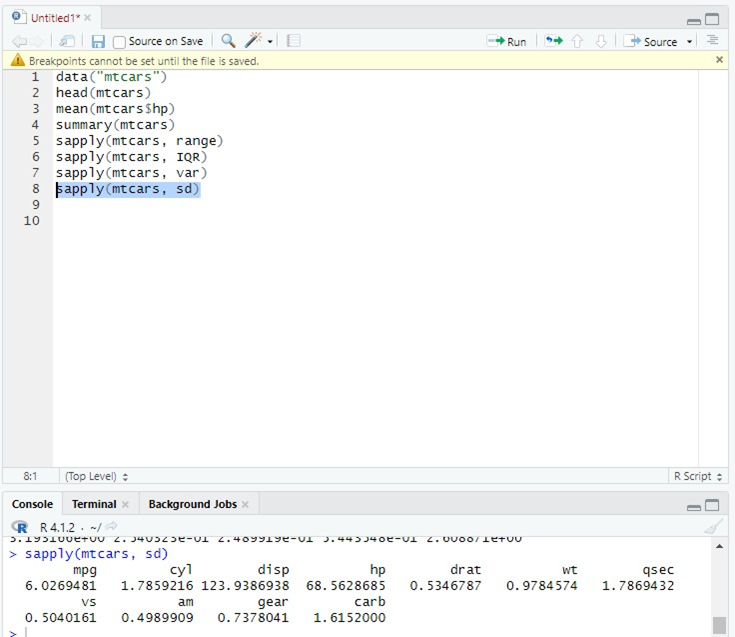
}

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

****

****

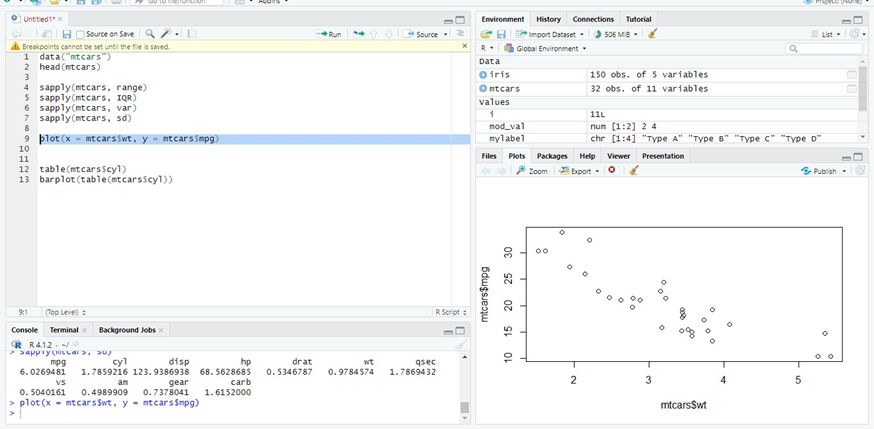
****

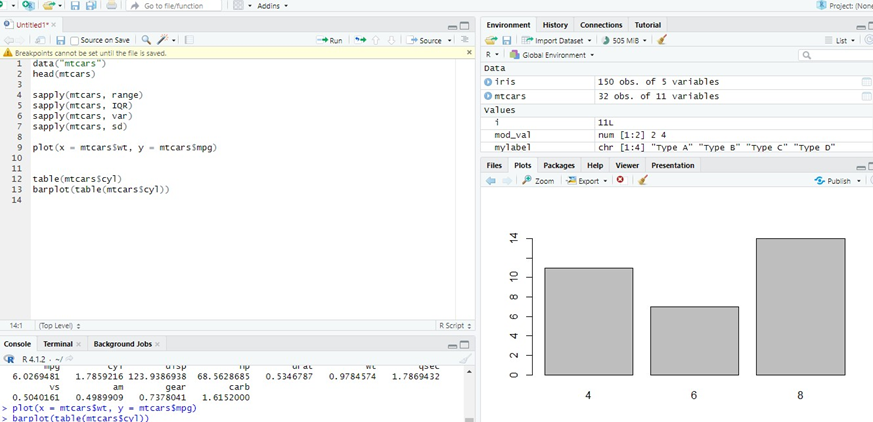
****

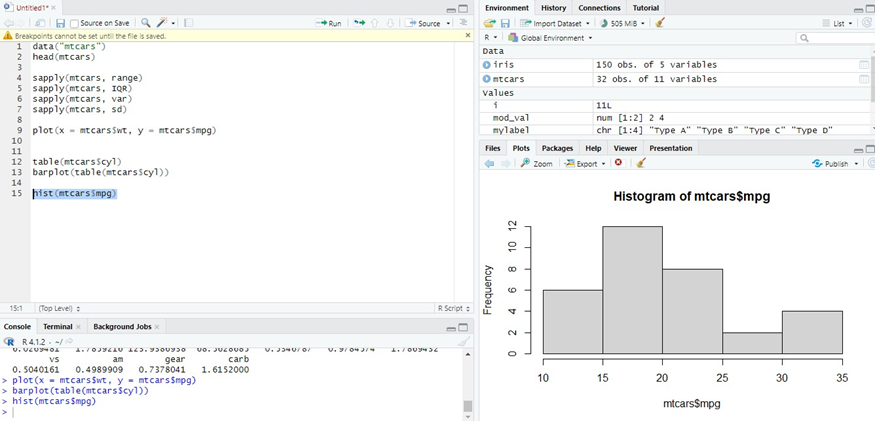
**data("mtcars") head(mtcars)**

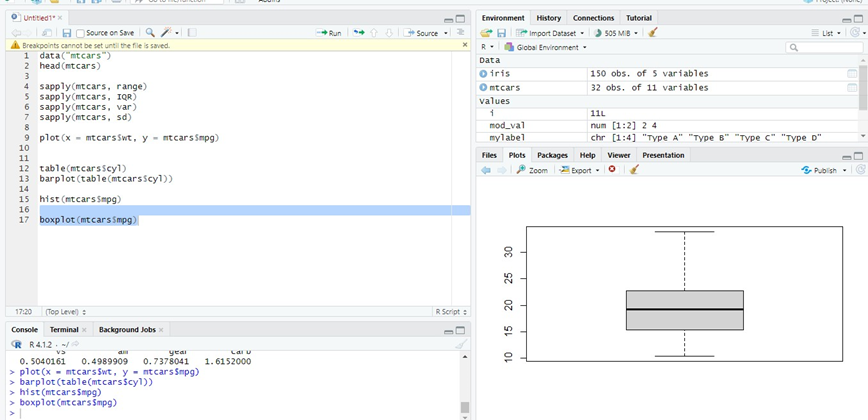
**sapply(mtcars, range) sapply(mtcars, IQR) sapply(mtcars, var) sapply(mtcars, sd)**

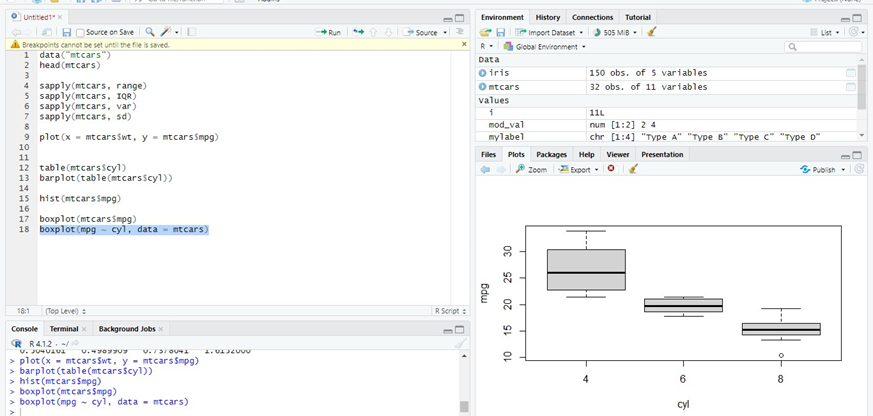
1. **Diagrammatic representation of data (Bar graphs, Pie diagram, Histogram, Box plot )**

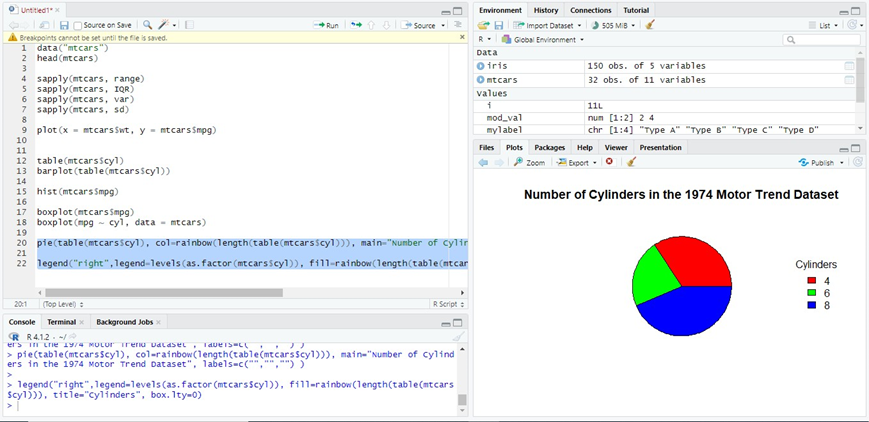
****

****

****

****

****

****

**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))**

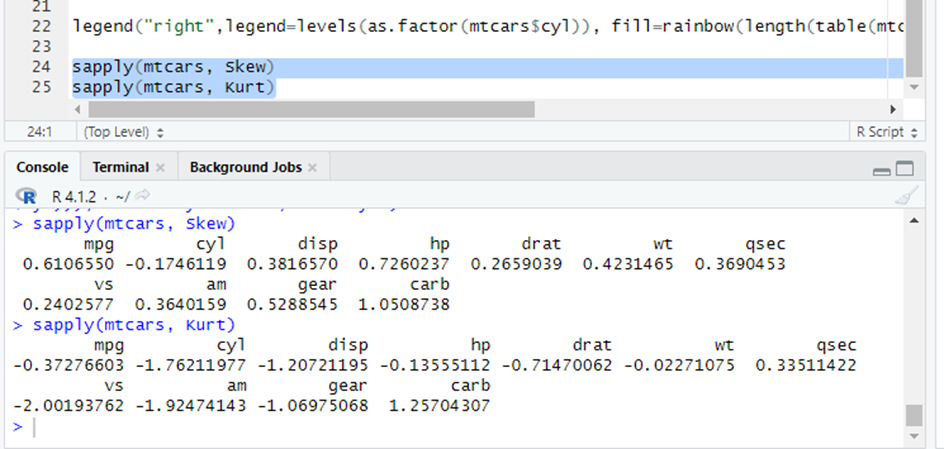
**hist(mtcars$mpg)**

**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)**

1. **Measure of Skewness and Kurtosis**

****

**sapply(mtcars, Skew) sapply(mtcars, 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.**

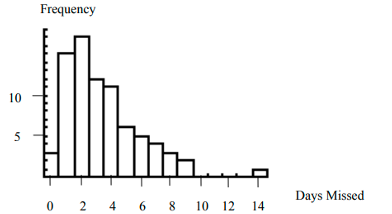
**Post lab Questions:**

1. In descriptive statistics the aim is to (multiple options may be correct) :
   1. 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.
   2. Use of probability theory to learn about population from sample data
   3. Quantitatively describe or summarize the data
   4. 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.

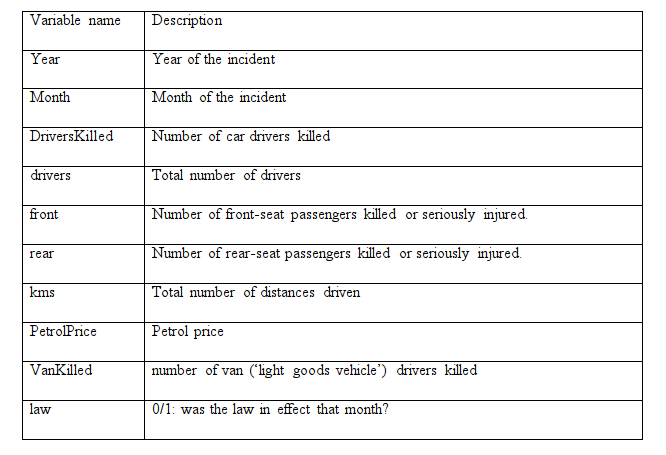
1. A boeing 747 aircaraft gets cancelled while severe snowstorms. The following histogram shows the number of days missed(per year) in last 75 years. Which of the following you use as measure to describe the center of the distribution.

****

1. Mean, because it covers information from all 75 years
2. IQR because it is unaffected by the outliers.
3. Median because the distribution is skewed to the right
4. Standard deviation, because it is unaffected by outliers and the distribution is skewed.
5. 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)
6. The variance increases
7. The mean increases
8. The median increases
9. The IQR increases.

They will all increase.

1. Refer the data set “[seatbelts.csv](https://drive.google.com/file/d/1tw4M03V9m3V_ZPwmtC8v6NxhSHJ9_3l7/view?usp=sharing)”. 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.



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

1. 90

II) 85

1. 100
2. None of the above
3. Referring the data set “[seatbelts.csv](https://drive.google.com/file/d/1tw4M03V9m3V_ZPwmtC8v6NxhSHJ9_3l7/view?usp=sharing)”. How many front seat passengers were injured or killed in the year 1984
4. 7041
5. 7047
6. 7865
7. None of the above
8. Referring the data set “[seatbelts.csv](https://drive.google.com/file/d/1tw4M03V9m3V_ZPwmtC8v6NxhSHJ9_3l7/view?usp=sharing)”. Calculate the variance for the variables “front” and “rear” and choose the correct option.
9. Variance of front seat passengers is equal to variance of rear seat passengers.
10. Variance of front seat passengers is greater than variance of rear seat passengers.
11. Variance of front seat passengers is less than the variance of rear seat passengers.
12. None of the above
13. Referring the data set “[seatbelts.csv](https://drive.google.com/file/d/1tw4M03V9m3V_ZPwmtC8v6NxhSHJ9_3l7/view?usp=sharing)”. Maximum kms driven by the driver is \_\_\_?
14. 21626
15. 17203
16. 25245
17. None of the above
18. Which of the following statement is not true about histograms?
19. Represent the frequency distribution of categorical variables
20. It is a graphical representation of data using bars of different heights
21. Groups numbers into ranges and the height of each bar depicts the frequency of each range or bin
22. Represent the frequency distribution of numerical variables