Practical No: 1

Part A:

Aim: Perform descriptive analysis on data.

Theory:

Descriptive Analysis:

Descriptive analysis is the elementary transformation of data in a way that describes the basic characteristics such as central tendency, distribution, and variability.

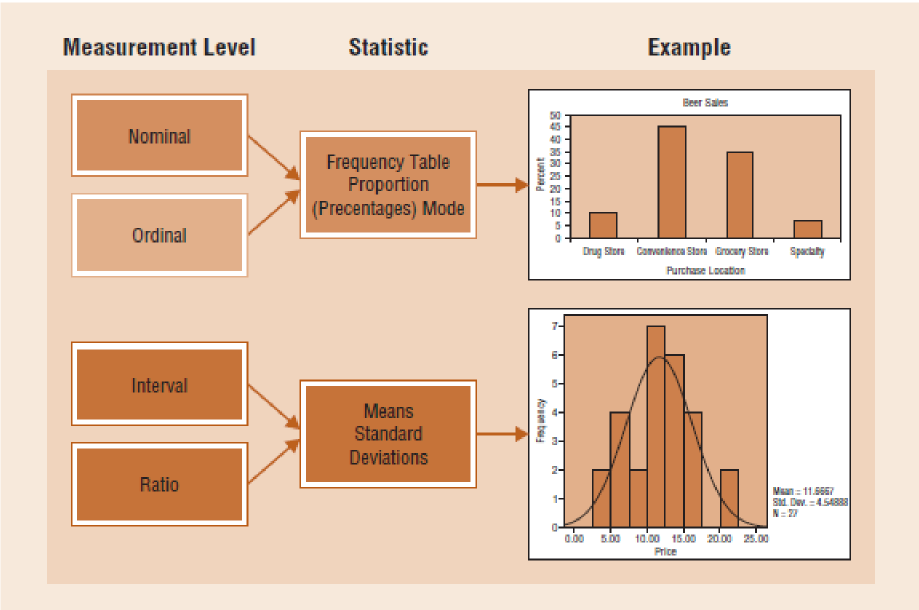


EXHIBIT 20.1
Levels of Scale Measurement and Suggested Descriptive Statistics

Mean:

The mean is simply the arithmetic average, and it is perhaps the most common measure of central tendency.

Researchers generally wish to know the population mean μ, (lowercase Greek letter mu), which is calculated as follows:

$$\mu = \frac{\sum_{i=1}^{n} X_i}{N}$$

Where

N= number of all observations in the population

Often we will not have the data to calculate the population mean μ, we will calculate the sample mean, $\bar{x}$(read "X bar"), with the following formula:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Where

n= number of observations made in the sample.

Median:

The next measure of central tendency, the **median**, is the midpoint of the distribution, or the 50^{th} percentile. In other words, the median is the value below which half the values in the sample fall, and above which half of the values fall.

Mode:

In apparel, mode refers to the most popular fashion. In statistics the **mode** is the measure of central tendency that identifies the value that occurs most often.

Range:

The simplest measure of dispersion is the **range**. It is the distance between the smallest and the largest values of a frequency distribution.

Standard error:

The **standard error** of the mean, using the following formula

$$S_{\overline{X}} = \frac{S}{\sqrt{n}}$$

Variance:

A measure of variability or dispersion. Its square root is the standard deviation.

$$\text{Variance} = S^2 = \frac{\Sigma(X_i - \overline{X})^2}{n - 1}$$

Variance is a very good index of dispersion. The variance, $\boldsymbol{S^2}$,will equal zero if and only if each and every observation in the distribution is the same as the mean. The variance will grow larger as the observations tend to differ increasingly from one another and from the mean.

Standard deviation:

A quantitative index of a distribution's spread, or variability; the square root of the variance for a distribution.

For the standard deviation is

$$S = \sqrt{S^2} = \sqrt{\frac{\Sigma(X_i - \overline{X})^2}{n - 1}}$$

Skewness:

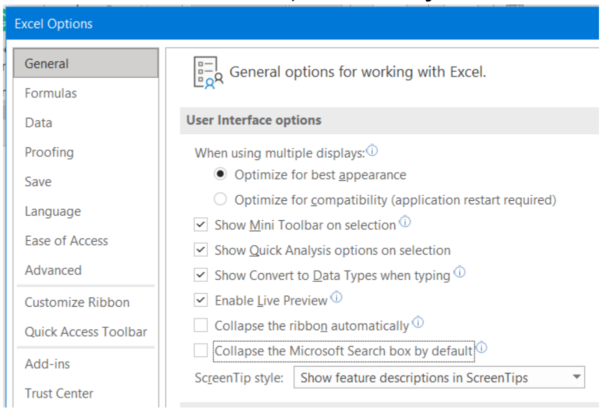
Skewness is a measure of the symmetry in a distribution. A symmetrical dataset will have a skewness equal to 0. So, a normal distribution will have a skewness of 0. Skewness essentially measures the relative size of the two tails.

Kurtosis:

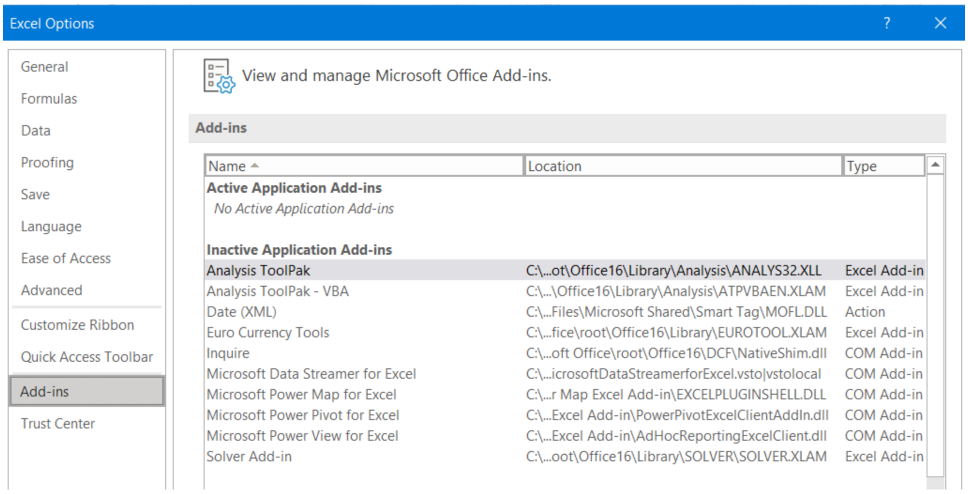
Kurtosis is a measure of the combined sizes of the two tails. It measures the amount of probability in the tails. The value is often compared to the kurtosis of the normal distribution, which is equal to 3. If the kurtosis is greater than 3, then the dataset has heavier tails than a normal distribution (more in the tails). If the kurtosis is less than 3, then the dataset has lighter tails than a normal distribution (less in the tails).

Steps:

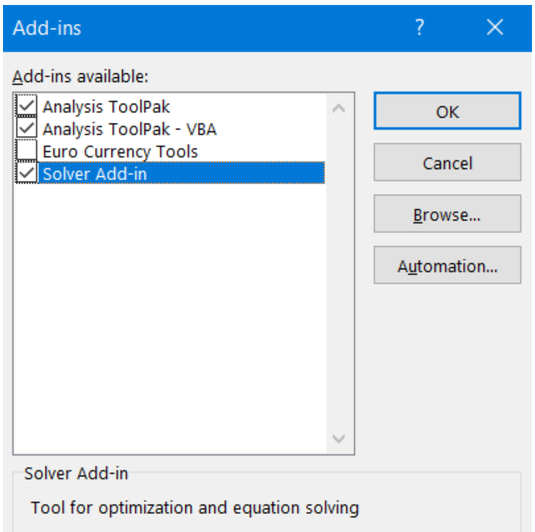
1. Open MS-Excel - (Click on [office button]) or directly - File menu - Select 'option' tab



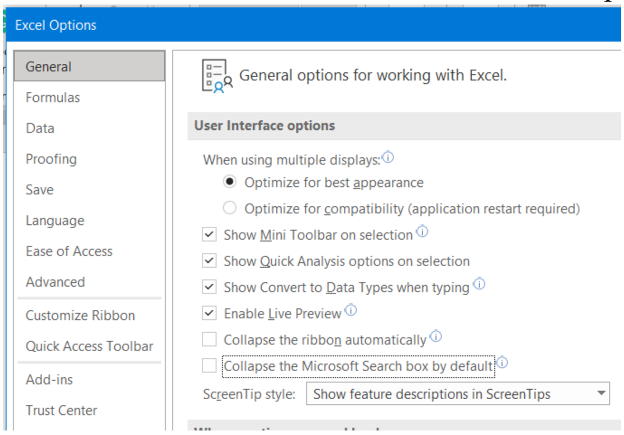
2. In excel option window click on 'Add-Ins' tab



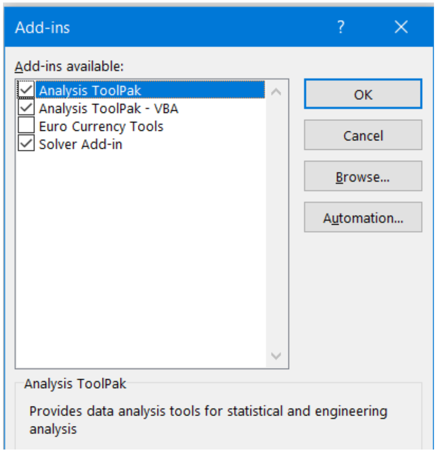
3. In the add-ins tab select analysis tool pack and click on go. A pop-up window opens select all the tools except euro currency tool, click on ok.



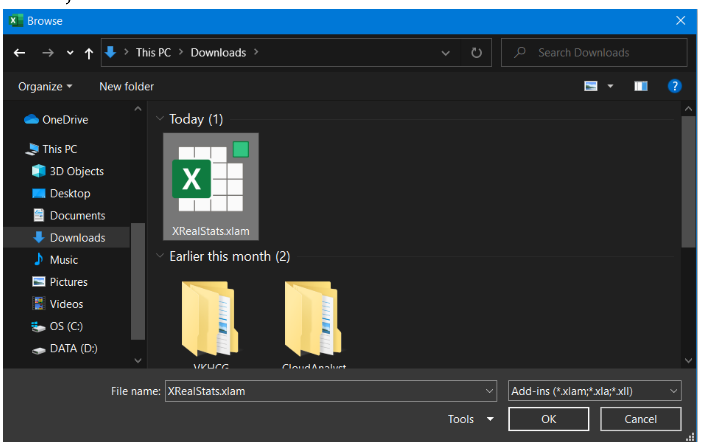
4. Open MS-Excel □ Click on [office button] - File menu - Select ‘option’ tab



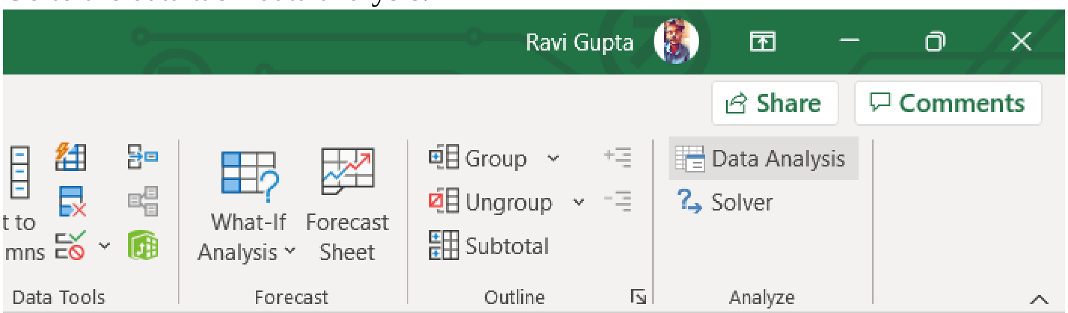
5. In excel option window click on ‘ Add-Ins’ tab – Go



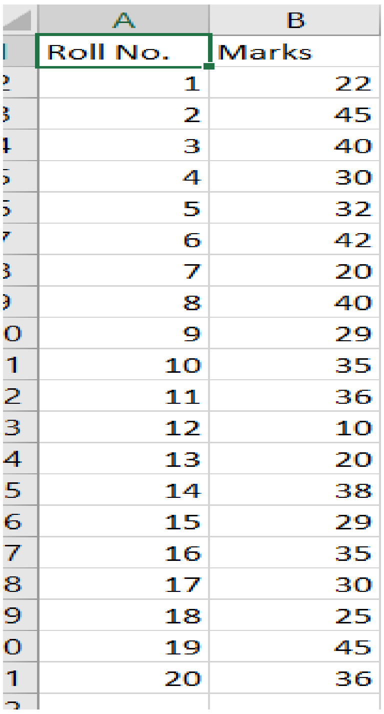
6. Browse the xlam file, Click Ok.



7. Go to the data tab - data analysis.

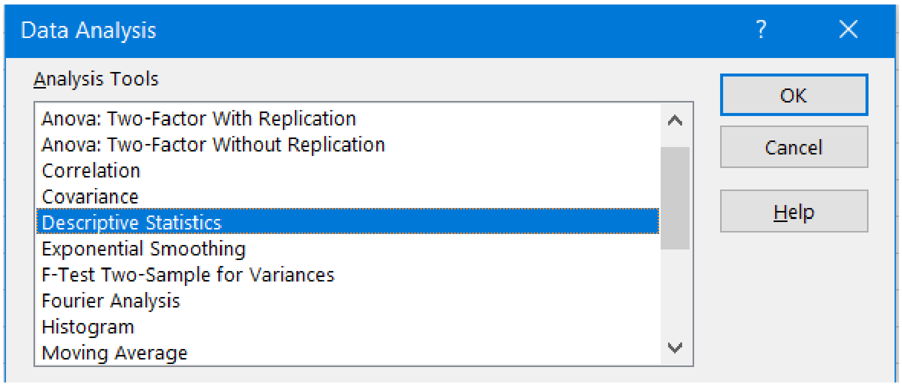


8. In the excel sheet add the data required for analysis.

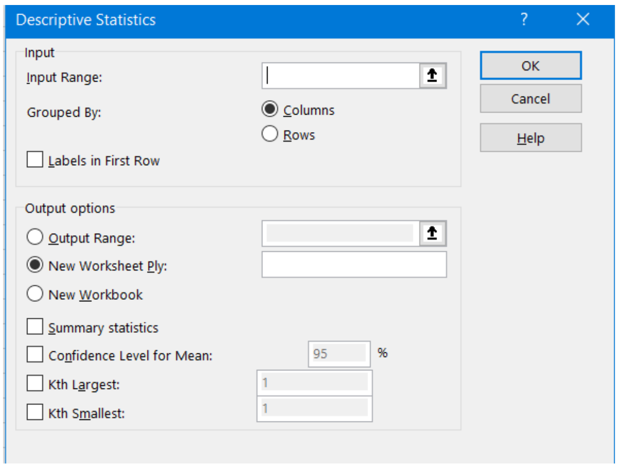


Roll No.	Marks
1	22
2	45
3	40
4	30
5	32
6	42
7	20
8	40
9	29
10	35
11	36
12	10
13	20
14	38
15	29
16	35
17	30
18	25
19	45
20	36

9. Once the data is added go to the 'Data' tab, select 'Data Analysis' from the pop-up that appears select the desired technique (for this practical- 'Descriptive Statistics'



10. Another popup window will open.



11. In the open popup window do the following:
1. Select the input range
2. In output options select output range radio button. Select a cell for output.
3. Select the summary statistic checkbox, click on ok. [Make sure ‘Labels in first row’ checkbox is selected.]

16_Ravi Gupta	
Marks	
Mean	31.95
Standard Error	2.051283
Median	33.5
Mode	45
Standard Deviation	9.173618
Sample Variance	84.15526
Kurtosis	0.172569
Skewness	-0.64093
Range	35
Minimum	10
Maximum	45
Sum	639
Count	20
Largest(2)	45
Smallest(2)	20

Part B:

Aim: Import data from different data sources (from Excel, csv).

Theory:

Matplotlib:

Matplotlib is one of the most popular Python packages used for data visualization. It is a cross platform library for making 2D plots from data in arrays. Matplotlib is written in Python and makes use of NumPy, the numerical mathematics extension of Python. It provides an object-oriented API that helps in embedding plots in applications using Python GUI toolkits such as PyQt, WxPythonotTkinter. It can be used in Python and IPython shells, Jupyter notebook and web application servers also.

Matplotlib has a procedural interface named the Pylab, which is designed to resemble MATLAB, a proprietary programming language developed by MathWorks. Matplotlib along with NumPy can be considered as the open source equivalent of MATLAB.

Pandas:

Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays. As one of the most popular data wrangling packages.

Pandas works well with many other data science modules inside the Python ecosystem.

show( ):

It will display the current figure that you are working on. plt.draw() will re-draw the figure. This allows you to work in interactive mode and, should you have changed your data or formatting, allow the graph itself to change.

plot( ):

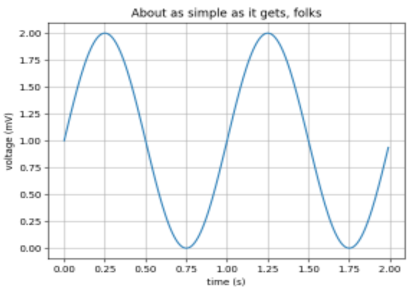
A plot is a graphical technique for representing a data set, usually as a graph showing the relationship between two or more variables.

Plots play an important role in statistics and data analysis. The procedures here can broadly be split into two parts: quantitative and graphical plot.

Different types of Plots:

- **Line plot:**

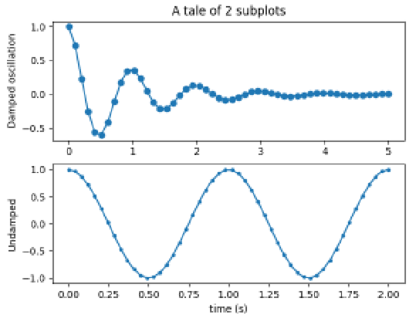
  Create a line plot with text tables using **plot().**



Simple Plot

- **Multiple subplots plot:**

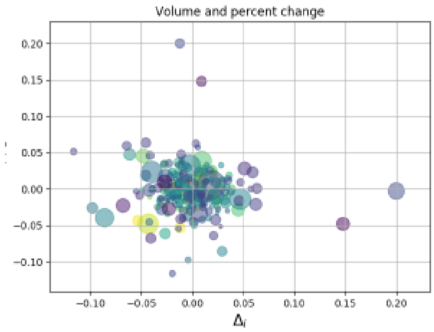
  Multiple axes are created with the **subplot()** function.



Subplot

- **Scatter plots:**

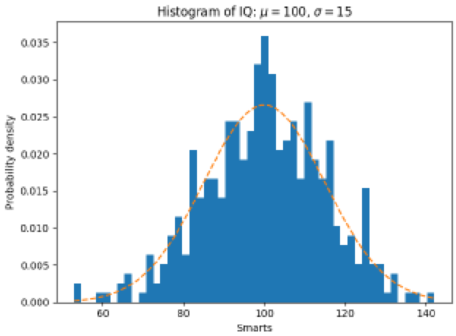
  The **scatter()** function makes a scatter plot with optional (size) and colour arguments. A scatter plot of y vs. x with varying marker size and colour.Here the alpha attributes is used to make semitransparent circle markers.



Scatter plot

- **Histograms:**

  The **hist()** function automatically generates histograms and returns the bin counts or probabilities.



Histogram Features

read_excel(): It read Excel file into a pandas DataFrame. Supports xls, xlsx, odf etc file extension read from local filesystem or URL. An option to read a single sheet or list of sheets.

to_excel(): Write object to an Excel sheet. To a single object to an Excel .xlsx file it is only necessary to specify a target file name. To write to multiple sheets it is necessary to create an ExcelWriter object with a target file name, and specify a sheet in the file to write to.

read_csv(): Read a comma-seprated value (csv) file into DataFrame. Also supports optionally iterating or breaking of file into chunks.

Steps:

1. **Install files:**

pip install xlrd

```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.19043.1320]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Ravi Gupta\AppData\Local\Programs\Python\Python37\Scripts>pip install xlrd
Collecting xlrd
  Downloading xlrd-2.0.1-py2.py3-none-any.whl (96 kB)
     |████████████████████████████████| 96 kB 1.7 MB/s
Installing collected packages: xlrd
Successfully installed xlrd-2.0.1
WARNING: You are using pip version 21.2.4; however, version 21.3.1 is available.
You should consider upgrading via the 'c:\users\ravi gupta\appdata\local\programs\python\python37\python.exe -m pip inst
all --upgrade pip' command.

C:\Users\Ravi Gupta\AppData\Local\Programs\Python\Python37\Scripts>
```

pip install openpyxl

```
C:\Windows\System32\cmd.exe
all --upgrade pip' command.

C:\Users\Ravi Gupta\AppData\Local\Programs\Python\Python37\Scripts>pip install Openpyxl
Requirement already satisfied: Openpyxl in c:\users\ravi gupta\appdata\local\programs\python\python37\lib\site-packages
(3.0.9)
Requirement already satisfied: et-xmlfile in c:\users\ravi gupta\appdata\local\programs\python\python37\lib\site-package
s (from Openpyxl) (1.1.0)
WARNING: You are using pip version 21.2.4; however, version 21.3.1 is available.
You should consider upgrading via the 'c:\users\ravi gupta\appdata\local\programs\python\python37\python.exe -m pip inst
all --upgrade pip' command.
```

2. **Create and save an *.xls file and *.csv file.**

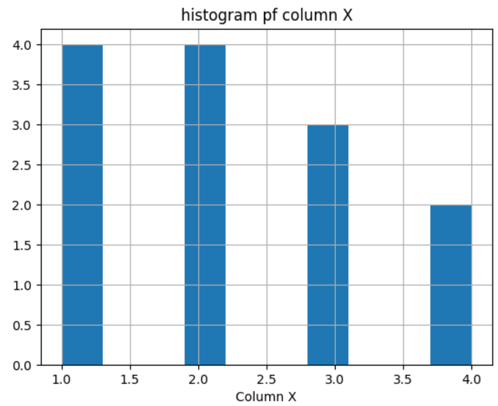
	A	B	C
1	X	Y	Z
2	1	1	9
3	1	2	5
4	1	5	7
5	1	4	8
6	2	8	10
7	2	6	10
8	2	5	10
9	2	4	10
10	3	6	10
11	3	5	5
12	3	5	9
13	4	8	7
14	4	8	7
15			

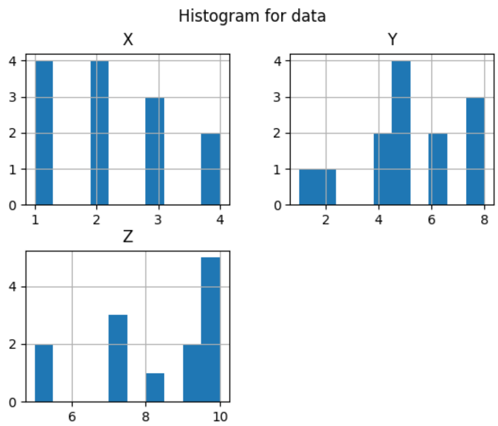
Code:

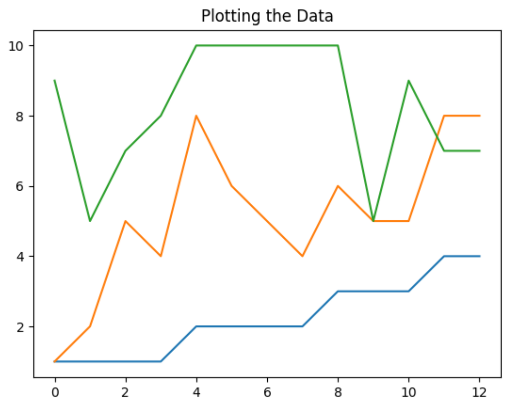
```
import pandas as pd
import scipy
import matplotlib.pyplot as plt
print("Ravi Gupta , Roll No:16")
df = pd.read_excel("Data.xlsx", engine='openpyxl')
print("Reading and printing Data in the excel file:")
print(df)
print("Reading and printing Data in the excel file using head method:")
print(df.head(5))
df.hist(column ="X")
plt.title("histogram pf column X")
plt.xlabel("Frequency")
plt.xlabel("Column X")
plt.show()
df.hist()
plt.suptitle("Histogram for data")
plt.show()
plt.plot()
plt.title("Plotting the Data")
plt.plot(df)
plt.show()
plt.scatter(df["Y"], df["Z"])
plt.title("Scatter Plot from Y and Z")
plt.show()
print("storing data from python to excel")
df.to_excel("ProcessedData.xls", sheet_name="ProcData")
print("Reading and prinitg Data from csv file:")
df1 = pd.read_csv('samplecsv.csv')
print(df1)
print("Reading and Printing Data from csv file using head method:")
print(df1.head(2))
plt.plot(df1["X"], label=["X"])
plt.plot(df1["Y"], label=["Y"])
plt.legend()
plt.show()
```

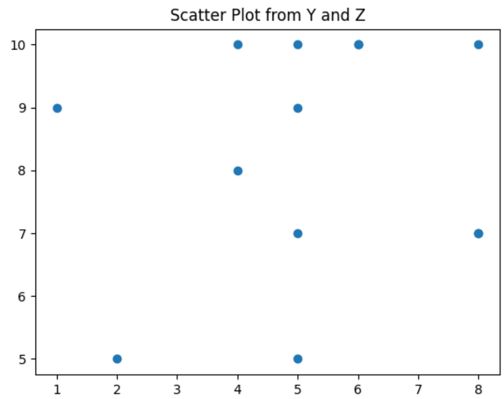
Output:

```
========= RESTART: C:/Users/Ravi Gupta/Desktop/RIC Practical/pr1.py =========
Ravi Gupta , Roll No:16
Reading and printing Data in the excel file:
    X  Y   Z
0   1  1   9
1   1  2   5
2   1  5   7
3   1  4   8
4   2  8  10
5   2  6  10
6   2  5  10
7   2  4  10
8   3  6  10
9   3  5   5
10  3  5   9
11  4  8   7
12  4  8   7
Reading and printing Data in the excel file using head method:
   X  Y   Z
0  1  1   9
1  1  2   5
2  1  5   7
3  1  4   8
4  2  8  10
```









Practical No: 2

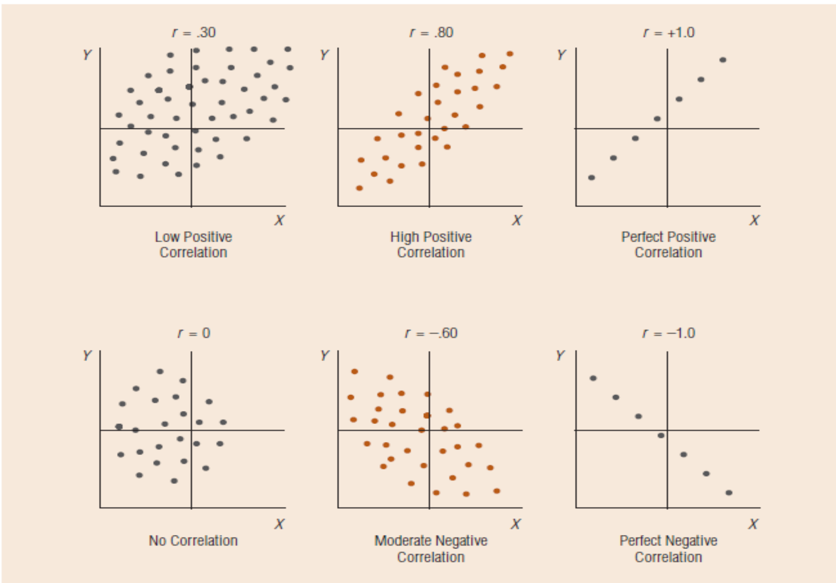
Aim: Compute different types of correlation.

Theory:

Correlation:

- Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. For example, height and weight are related; taller people tend to be heavier than shorter people. The relationship isn't perfect. People of the same height vary in weight, and you can easily think of two people you know where the shorter one is heavier than the taller one. Nonetheless, the average weight of people 5'5'' is less than the average weight of people 5'6'', and their average weight is less than that of people 5'7'', etc. Correlation can tell you just how much of the variation in peoples' weights is related to their heights.
- Correlation works for quantifiable data in which numbers are meaningful, usually quantities of some sort. It cannot be used for purely categorical data, such as gender, brands purchased, or favourite colour.
- The main result of a correlation is called the correlation coefficient (or "r"). It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related.
- If r is close to 0, it means there is no relationship between the variables. If r is positive, it means that as one variable gets larger the other gets larger. If r is negative it means that as one gets larger, the other gets smaller (often called an "inverse" correlation).
- Numpy implements a corrcoef() function that returns a matrix of correlations of x with x, x with y, y with x and y with y. We're interested in the values of correlation of x with y (so position (1, 0) or (0, 1)).

EXHIBIT 23.2
Scatter Diagram to Illustrate Correlation Patterns



Methods Used:

np.random.seed ( ):

This function is for pseudo-random numbers in Python. It sets the random seed of the NumPy pseudo number generator. This is used in the generation of a pseudo-random encryption key. Encryption keys are an important part of computer security. These are the kind of secret keys which used to protect data from unauthorized access over the internet.

np.random.randint ( ):

It return random integers from low to high. Return random integers from the “discrete uniform” distribution of the specified dtype in the “half-open” interval(low, high).

numpy.random.randint(low, high=None, size=None, dtype=’l’)

np.random.normal ( ):

Draw random samples from a normal Gaussian distribution. The probability density function of the normal distribution, first derived by De Moivre and 200 years later by both Gauss and Laplace independently , is often called the bell curve because of its characteristic shape (see the example below).The normal distributions occurs often in nature. For example, it describes the commonly occurring distribution of samples influenced by a large number of tiny, random disturbances, each with its own unique distribution.

numpy.random.normal(loc=, scale=, size=)

plt.scatter( ):

Scatter plots are used to observe relationship between variables and uses dots to represent the relationship between them. The **scatter( )** method in the matplotlib library is used to draw a scatter plot. Scatter plots are widely used to represent relation among variables and how change in one affects the other.

np.corrcoef( ):

The **corrcoef( )** returns the correlation matrix, which is a two-dimensional array with the correlation coefficients.The main diagonal of the matrix is equal to 1. The upper left value is the correlation coefficient for x and x. Similarly, the lower right value is the correlation coefficient for y and y.

numpy.corrcoef(x, y=None, rowvar=True, bias=<no value>, ddof=<no value>, *, dtype=None)

plt.show( ):

It will display the current figure that you are working on. plt.draw() will re-draw the figure. This allows you to work in interactive mode and, should you have changed your data or formatting, allow the graph itself to change.

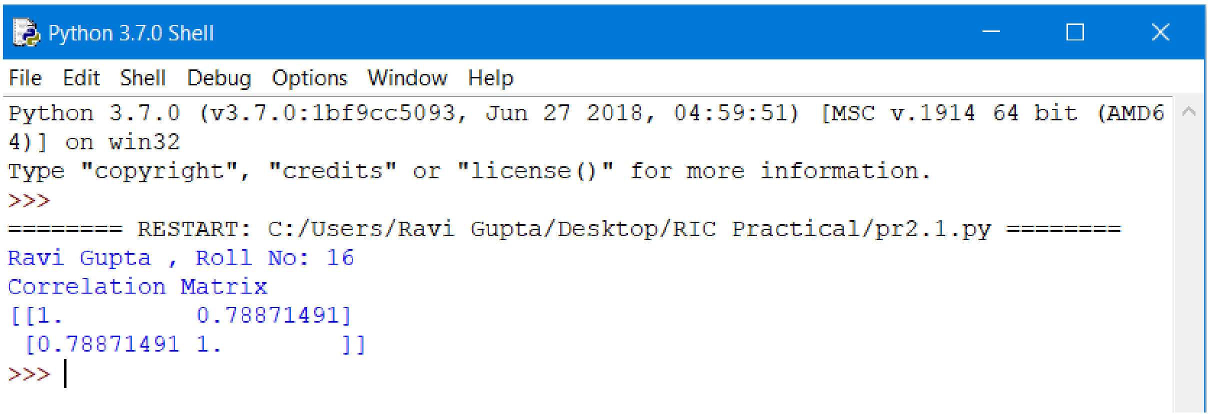
Part A:

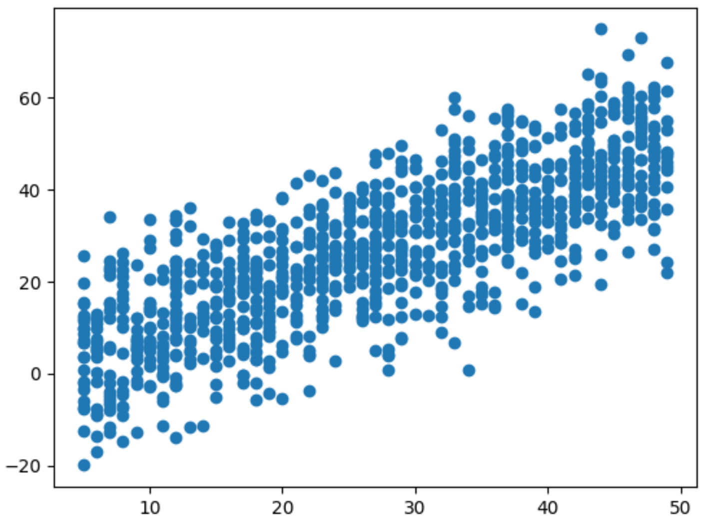
Aim: Demonstrate Positive Correlation.

Code:

```
import numpy as np
import matplotlib.pyplot as plt
print("Ravi Gupta , Roll No: 16")
np.random.seed(1)
# 1000 random integers between 0 and 50
X= np.random.randint(5,50,1000)
# positive COrrelation with some noise
Y= X + np.random.normal(0,10,1000)
print("Correlation Matrix")
print(np.corrcoef(X,Y))
plt.scatter(X,Y)
plt.show()
```

Output:





Part B:

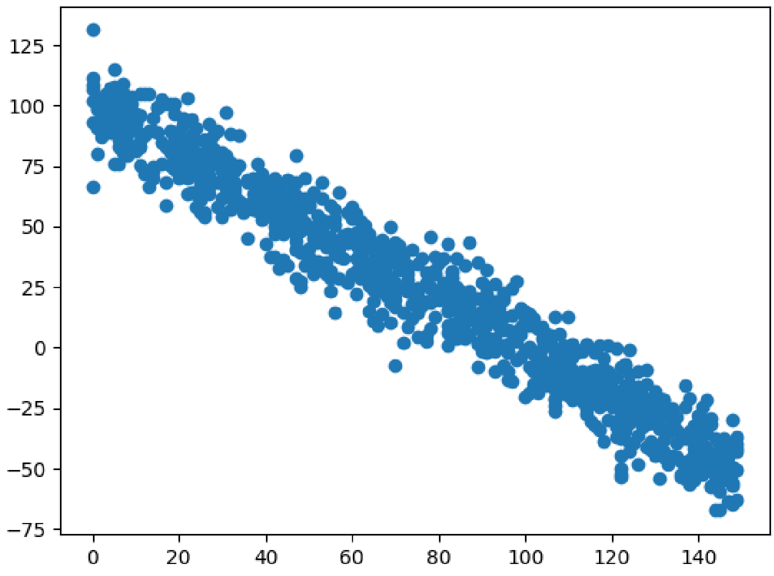
Aim: Demonstrate Negative Correlation.

Code:

```
import numpy as np
import matplotlib.pyplot as plt
print("Ravi Gupta , Roll No: 16")
#generate 1000 random numbers in the ranges of 0 to 150
X= np.random.randint(0,150,1000)
#negative correlation with some noise
#np.rndom,normal(0,1,1000)-- generate 1000 random numbers in the range of 0 to 1000
#Y= 100-X +random no ... is formula for negative correlation
Y=100-X +np.random.normal(0,10,1000)
print(np.corrcoef(X,Y))
plt.scatter(X,Y)
plt.show()
```

Output:

```
>>>
======== RESTART: C:/Users/Ravi Gupta/Desktop/RIC Practical/pr2.2.py ========
Ravi Gupta , Roll No: 16
[[ 1.         -0.97594641]
 [-0.97594641  1.        ]]
>>>
```



Part C:

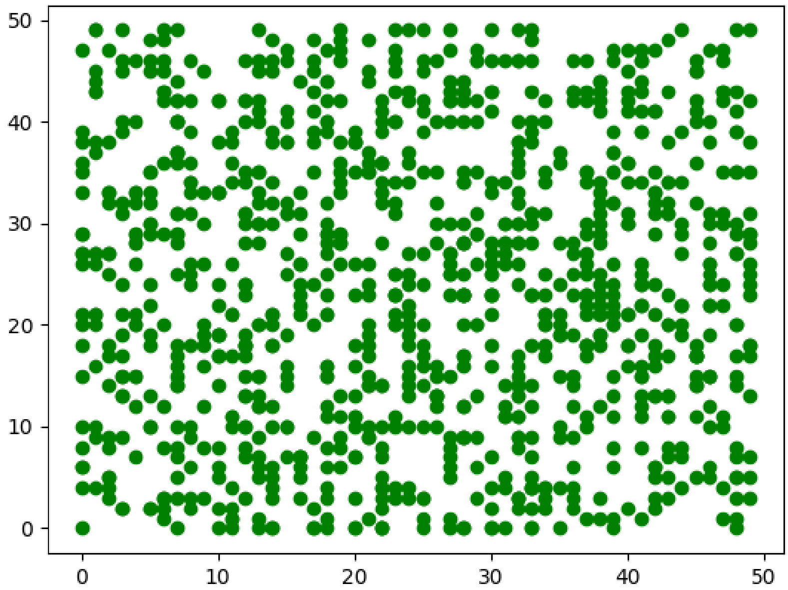
Aim: Demonstrate No/weak Correlation.

Code:

```
import numpy as np
import matplotlib.pyplot as plt
print("Ravi Gupta Roll NO:16")
np.random.seed(1)
X=np.random.randint(0,50,1000)
Y=np.random.randint(0,50,1000)
print("Correlation Matrix")
print(np.corrcoef(X,Y))
plt.scatter(X,Y,color="green")
plt.show()
```

Output:

```
======== RESTART: C:/Users/Ravi Gupta/Desktop/RIC Practical/pr2.3.py ========
Ravi Gupta Roll NO:16
Correlation Matrix
[[1.         0.00404702]
 [0.00404702 1.        ]]
>>> |
```



Practical No: 3

Theory:

Regression:

- In statistical modelling, regression analysis is used to estimate the relationships between two or more variables:
- Dependent variable (i.e., criterion variable) is the main factor you are trying to understand and predict. Independent variables (i.e. explanatory variables, or predictors) are the factors that might influence the dependent variable.
- Regression analysis helps you understand how the dependent variable changes when one of the independent variables varies and allows to mathematically determine which of those variables really has an impact.
- Simple linear regression models the relationship between a dependent variable and one independent variables using a linear function. If you use two or more explanatory variables to predict the dependent variable, you deal with multiple linear regression.
- Linear regression, coefficients a and b such that: **y = bx + a**
- Polynomial Regression is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial.
- In general, we can model it for *nth* value/degree. **y = a + b1x + b2x^2 +....+ bnx^n**

Regarding Output:

a) Regression analysis output: **Summary Output**

This part tells you how well the calculated linear regression equation fits your source data.

Multiple R:

It is the **Correlation Coefficient** that measures the strength of a linear relationship between two variables. The correlation coefficient can be any **value between -1 and 1**, and its absolute value indicates the relationship strength. The larger the absolute value, the stronger the relationship:

1 means a strong positive relationship

-1 means a strong negative relationship

0 means no relationship at all

R Square:

It is the **Coefficient of Determination**, which is used as an indicator of the goodness of fit. It shows how many points fall on the regression line. The R2 value is calculated from the total sum of squares, more precisely, it is the sum of the squared deviations of the original data from the mean.

In our example, R2 is 0.91 (rounded to 2 digits), which is fairy good. It means that 91% of our values fit the regression analysis model. In other words, 91% of the dependent variables (y-values) are explained by the independent variables (x-values). Generally, R Squared of 95% or more is considered a good fit.

Adjusted R Square:

It is the R square adjusted for the number of **independent variable** in the model. You will want to use this value instead of R square for multiple regression analysis.

Standard Error:

It is another **goodness-of-fit measure** that shows the precision of your regression analysis - the smaller the number, the more certain you can be about your regression equation. While R2 represents the percentage of the dependent variables variance that is explained by the model, Standard Error is an absolute measure that shows the average distance that the data points fall from the regression line.

Observations: It is simply the number of observations in your model.

b) Regression analysis output: **ANOVA**

1. The second part of the output is **Analysis of Variance** (ANOVA):
2. Basically, it splits the sum of squares into individual components that give information about the levels of variability within your regression model:\
    - df is the number of the degrees of freedom associated with the sources of variance.
    - SS is the sum of squares. The smaller the Residual SS compared with the Total SS, the better your model fits the data.
    - MS is the mean square.
    - F is the F statistic, or F-test for the null hypothesis. It is used to test the overall significance of the model.
    - Significance F is the P-value of F.
3. The ANOVA part is rarely used for a simple linear regression analysis in Excel, but you should definitely have a close look at the last component. The Significance F value gives an idea of how reliable (statistically significant) your results are. If Significance F is less than 0.05 (5%), your model is OK. If it is greater than 0.05, you'd probably better choose another independent variable.

c) Regression analysis output: **coefficients**

1. This section provides specific information about the components of your analysis:
2. The most useful component in this section is Coefficients. It enables you to build a linear regression equation in Excel: **y = bx + a**
3. For our data set, where y is the number of umbrellas sold and x is an average monthly rainfall, our linear regression formula goes as follows: **Y = Rainfall Coefficient * x + Intercept**
4. Equipped with a and b values rounded to three decimal places, it turns into: **Y=0.45*x-19.074**
5. For example, with the average monthly rainfall equal to 82 mm, the umbrella sales would be approximately 17.8: **0.45*82-19.074=17.8**
6. In a similar manner, you can find out how many umbrellas are going to be sold with any other monthly rainfall (x variable) you specify.

d) Regression analysis output: **residuals**

1. If you compare the estimated and actual number of sold umbrellas corresponding to the monthly rainfall of 82 mm, you will see that these numbers are slightly different:

- Estimated: 17.8 (calculated above)
- Actual: 15 (row 2 of the source data)

2. Why's the difference? Because independent variables are never perfect predictors of the dependent variables. And the residuals can help you understand how far away the actual values are from the predicted values:

3. For the first data point (rainfall of 82 mm), the residual is approximately -2.8. So, we add this number to the predicted value, and get the actual value: 17.8 - 2.8 = 15.

Part A:

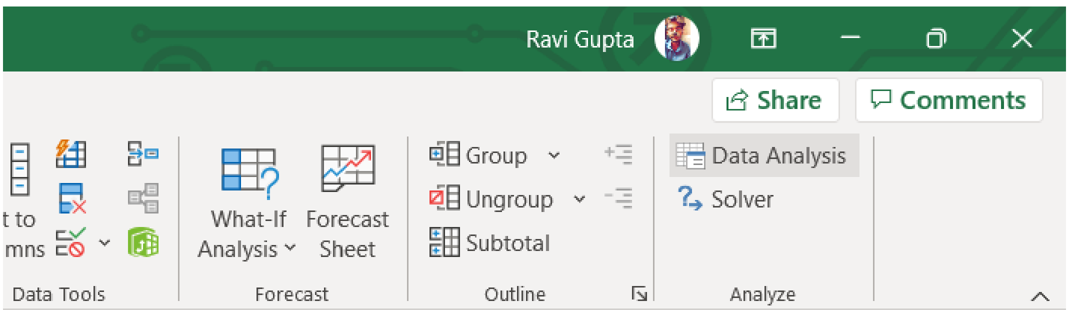
Aim: Perform Linear Regression for Prediction.

Steps:

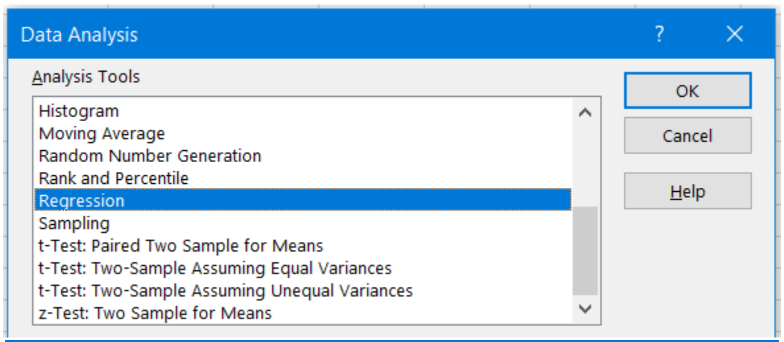
- Enter the required data:
  Umbrella Sold is dependent variable, rainfall is the independent variable
  X=Rainfall
  Y=Umbrella sold

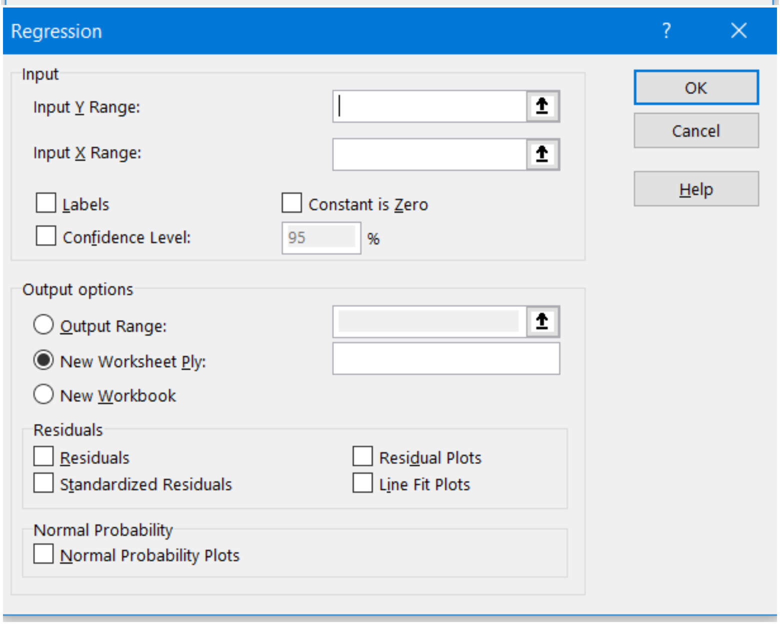
	A	B	C
1	Month	Railfall (mm)	Umbrellas sold
2	Jan	82	15
3	Feb	92.5	25
4	Mar	83.2	17
5	Apr	97.7	28
6	May	131.9	41
7	Jun	141.3	47
8	Jul	165.4	50
9	Aug	140	46
10	Sep	126.7	37
11	Oct	97.8	22
12	Nov	86.2	20
13	Dec	99.6	30
14	Jan	87	14
15	Feb	97.2	27
16	Mar	88.2	14
17	Apr	102.7	30
18	May	123	43
19	Jun	146.3	49
20	Jul	160	49

- Go to the data tap - Data Analysis

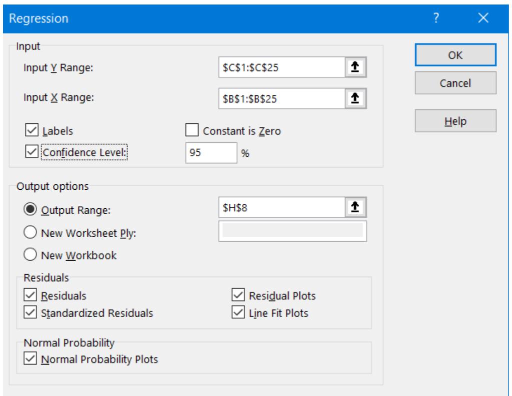


- Select regression - click on OK





- Select Y range - select X Range - check labels - select output range - select output residuals and normal probability - click on OK



Output:

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.957573162
R Square	0.916946361
Adjusted R Square	0.913171196
Standard Error	3.585286718
Observations	24

ANOVA

	df	SS	MS	F	Significance F
Regression	1	3122.164155	3122.164155	242.889057	2.26957E-13
Residual	22	282.7941788	12.85428085		
Total	23	3404.958333			

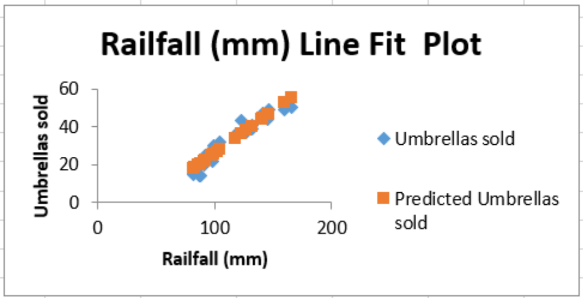
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	ower 95.0%	pper 95.0%
Intercept	-19.04669663	3.374426334	-5.644425079	1.1241E-05	-26.04482852	-12.0486	-26.0448	-12.0486
Railfall (mm)	0.449810426	0.028861942	15.58489836	2.26957E-13	0.389954423	0.509666	0.389954	0.509666

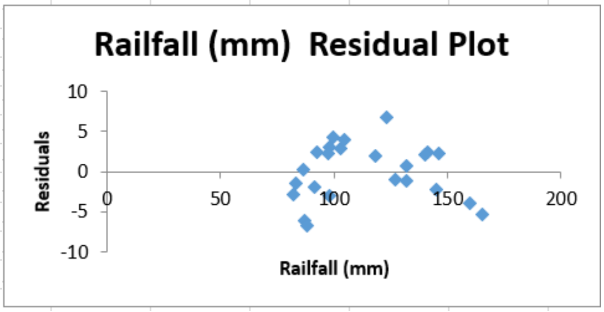
RESIDUAL OUTPUT

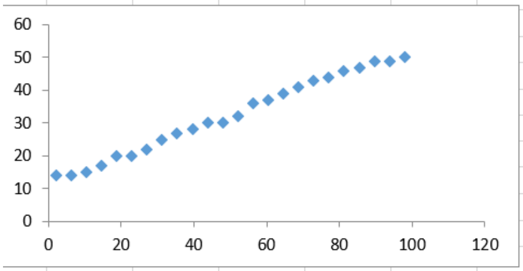
Observation	dicted Umbrellas s	Residuals	Standard Residuals
1	17.83775831	-2.83775831	-0.809289842
2	22.56076778	2.439232217	0.695635653
3	18.37753082	-1.377530821	-0.392852942
4	24.899782	3.100218001	0.884139755
5	40.28329857	0.716701431	0.204393442
6	44.51151657	2.488483426	0.709681424
7	55.35194784	-5.351947842	-1.526302296
8	43.92676302	2.07323698	0.591258819
9	37.94428435	-0.944284354	-0.269296978
10	24.94476304	-2.944763041	-0.839806128
11	19.7269621	0.273037901	0.077866674
12	25.75442181	4.245578192	1.21078081
13	20.0 Horizontal (Value) Axis	4	-1.735875054
14	24.67487679	2.325123214	0.663093327
15	20.62658295	-6.626582951	-1.889810789
16	27.14883413	2.851165871	0.813113495
17	36.27998578	6.720014222	1.916456109
18	46.7605687	2.239431295	0.638655164
19	52.92297154	-3.922971541	-1.118777807
20	46.17581515	-2.175815151	-0.62051271
21	40.19333648	-1.193336484	-0.340323237
22	34.03093365	1.969066353	0.561550782
23	21.97601423	-1.976014229	-0.563532221
24	28.00347394	3.996526062	1.13975455

PROBABILITY OUTPUT

Percentile	mbrellas sold
2.083333333	14
6.25	14
10.41666667	15
14.58333333	17
18.75	20
22.91666667	20
27.08333333	22
31.25	25
35.41666667	27
39.58333333	28
43.75	30
47.91666667	30
52.08333333	32
56.25	36
60.41666667	37
64.58333333	39
68.75	41
72.91666667	43
77.08333333	44
81.25	46
85.41666667	47
89.58333333	49
93.75	49
97.91666667	50







Part B:

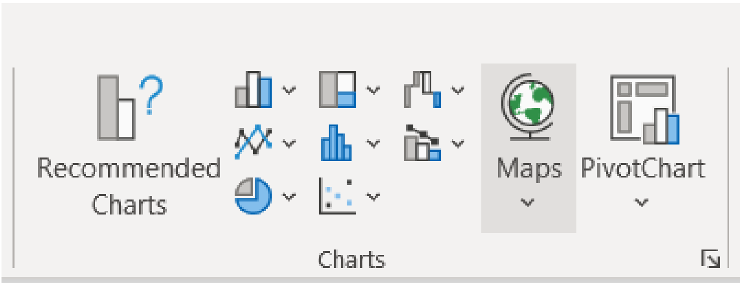
Aim: Perform the Polynomial Regression for Prediction.

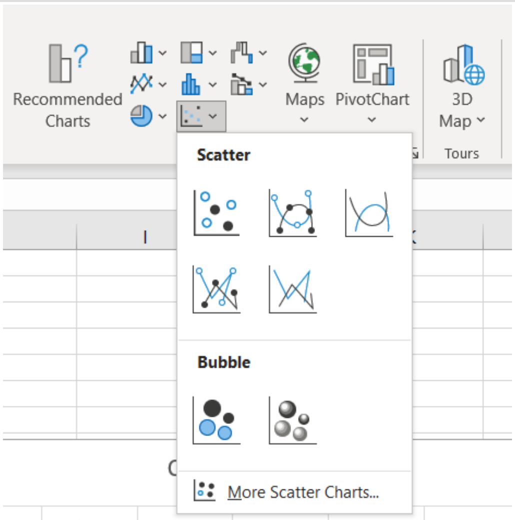
Steps:

1. In Excel: Insert the data

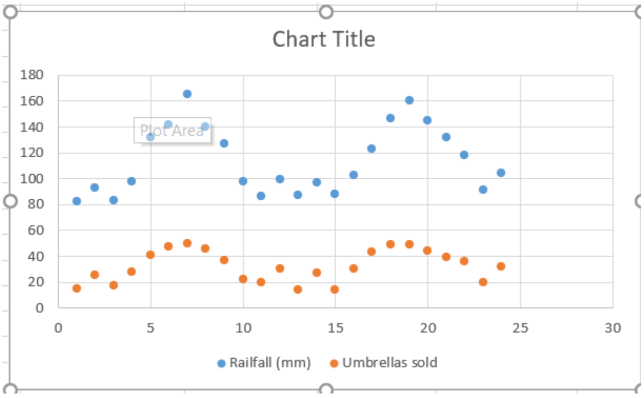
	A	B	C
1	Month	Railfall (mm)	Umbrellas sold
2	Jan	82	15
3	Feb	92.5	25
4	Mar	83.2	17
5	Apr	97.7	28
6	May	131.9	41
7	Jun	141.3	47
8	Jul	165.4	50
9	Aug	140	46
10	Sep	126.7	37
11	Oct	97.8	22
12	Nov	86.2	20
13	Dec	99.6	30
14	Jan	87	14
15	Feb	97.2	27
16	Mar	88.2	14
17	Apr	102.7	30
18	May	123	43
19	Jun	146.3	49
20	Jul	160	49

2. Select the data - insert tab - in chart - select Scatter plot

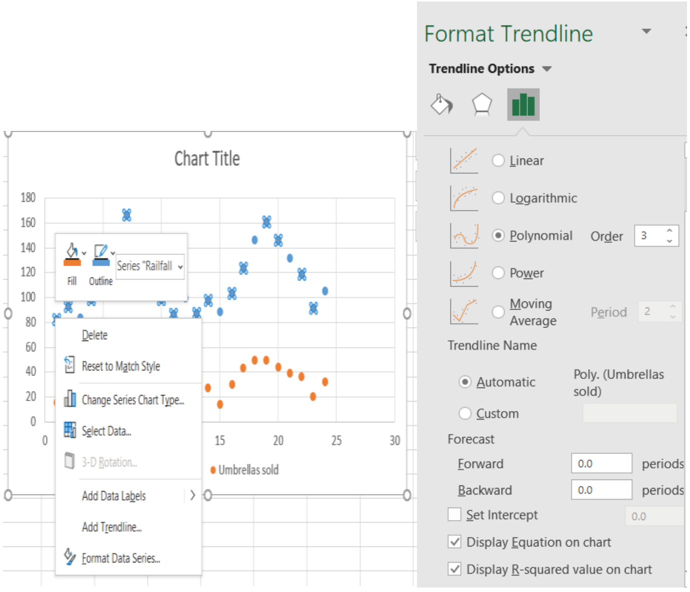


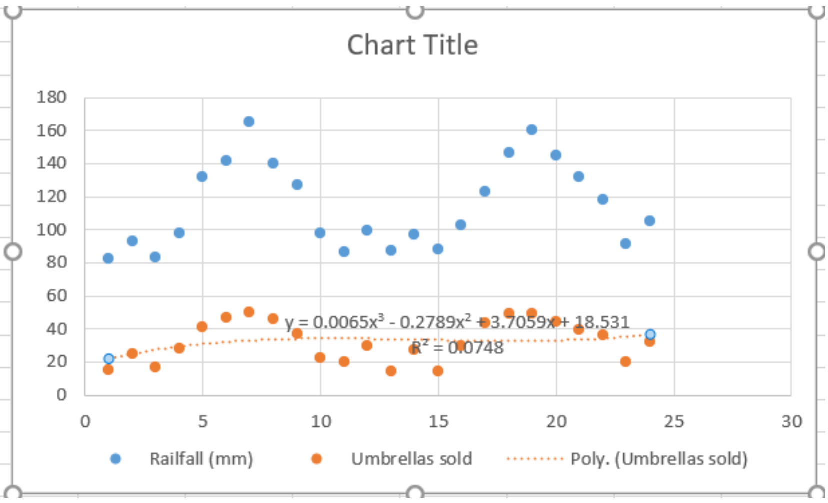


3. Scatter Plot will appear (Note: if no data is selected, scatter plot will be empty)



4. On the scatter plot > select any one point > (all points will get selected) > right Click > Add Trendline > (A dialog box will open)> Select Polynomial > Select order (2 or 3 or 4) > Check the display equation on chart and Display R- squared value on chart check boxes.





Practical No: 4

Aim: Perform Multiple Linear Regression.

Theory:

Multiple Linear Regression:

Multiple regression analysis is an extension of simple regression analysis allowing a metric dependent variable to be predicted by **multiple independent variables**. An analysis of association in which the effects of two or more independent variables on a single, interval-scaled dependent variable are investigated simultaneously.

Linear Regression: 1 Dependent Variable and 1 Independent Variable.
Multiple Regression: 1 Dependent Variable and Many [2 or more] Independent Variables.

Variables
Test Score →Dependent → y
IQ → Independent →x1
Study Hours → Independent → x2

The first task in our analysis is to define a linear, least-squares regression equation to predict test score, based on IQ and study hours. Since we have two independent variables, the equation takes the following form:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

In this equation, **$\hat{y}$ is the *predicted* test score. The independent variables are IQ and study hours, which are denoted by x_1 and x_2, respectively. The regression coefficients are b_0, b_1, and b_2.** On the right side of the equation, the only unknowns are the regression coefficients; so to specify the equation, we need to assign values to the coefficients.

Here, we see that the regression intercept (b_0) is 23.156, the regression coefficient for IQ (b_1) is 0.509, and the regression coefficient for study hours (b_2) is 0.467. So the least-squares regression equation can be re-written as: Y is Test Score

Y= 23.15614055 + 0.509433962 *IQ + 0.467133657 * Study Hours
Y = 23.1561 + 0.5094 * IQ + 0.4671 * Study Hours
If IQ= 104 and Study Hours=40 what is Predicted Test Score?
Predicted value is 94.8177
Actual value is 95
Residual Value = Actual – Predicted = 95 – 94.8177 = 0.1823

This is the only linear equation that satisfies a least-squares criterion. That means this equation fits the data from which it was created better than any other linear equation.

How well does our equation fit the data? To answer this question, researchers look at the coefficient of multiple determination (R^2). The coefficient of multiple determination measures the proportion of variation in the dependent variable that can be predicted from the set of independent variables in the regression equation. When the regression equation fits the data well, R^2 will be large (i.e., close to 1); and vice versa.

The coefficient of muliple determination is 0.905. For our sample problem, this means 90.5% of test score variation can be explained by IQ and by hours spent in study.

The F statistic (33.4) is big, and the p value (0.00026) is small. This indicates that one or both independent variables has explanatory power beyond what would be expected by chance.

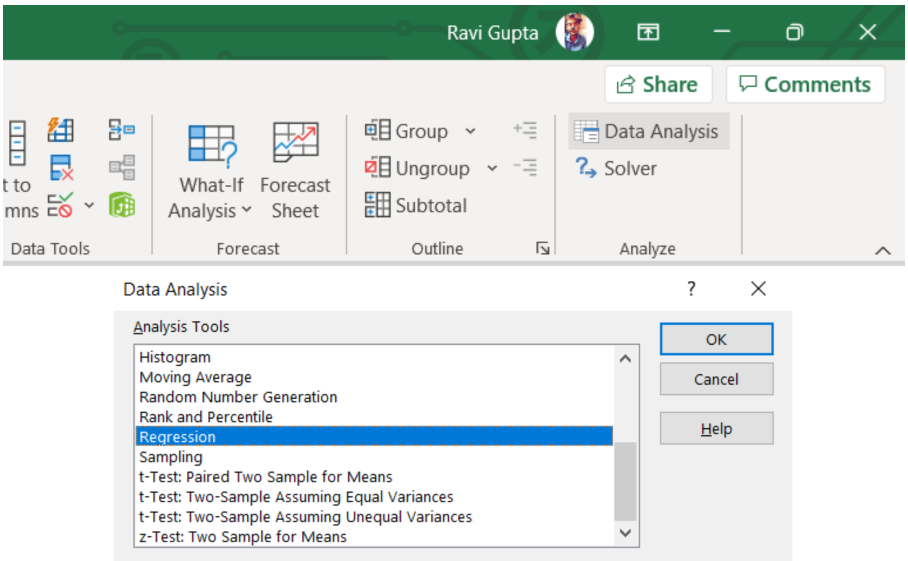
The regression coefficients table shows the following information for each coefficient: its value, its standard error, a t-statistic, and the significance of the t-statistic. In this example, the t-statistics for IQ and study hours are both statistically significant at the 0.05 level. This means that IQ contributes significantly to the regression after effects of study hours are taken into account. And study hours contribute significantly to the regression after effects of IQ are taken into account.

Steps:

1. Open Excel with the required Data.

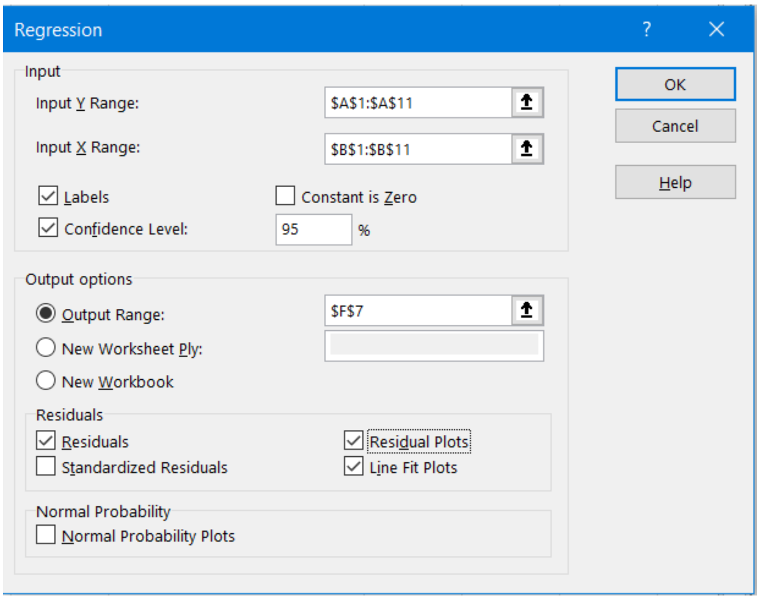
	A	B	C
1	Test Score	IQ	Study Hours
2	100	125	30
3	95	104	40
4	92	110	25
5	90	105	20
6	85	100	20
7	80	100	20
8	78	95	15
9	75	95	10
10	72	85	0
11	65	90	5

2. Go to ‘Data Tap’> Go to the Data analysis > Select Regression > Click OK



3. Regression Box will appear. Enter following things:

- Y Range: Test Score.
    - X Range: IQ and Study Hours.
    - Check the ‘Labels’ Check box.
    - Check the ‘Confidence Interval’ Check box [95% is the value]
    - Select ‘Output Range’ radio button in ‘Output Options’.
    - Provide a empty cell in ‘Output Range’ Text box.
    - Select Following Checkboxes: Residuals, Residual Plots and Line Fit Plots.
    - Click on “Ok”.



Output:

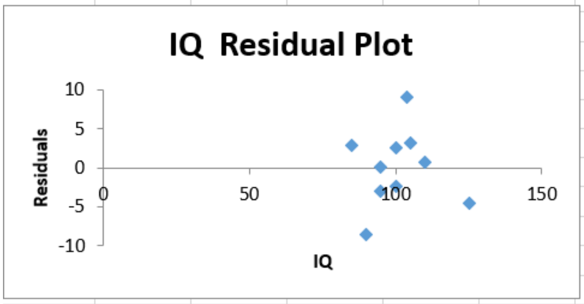
SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.89744
R Square	0.805398
Adjusted R Square	0.781073
Standard Error	5.195314
Observations	10

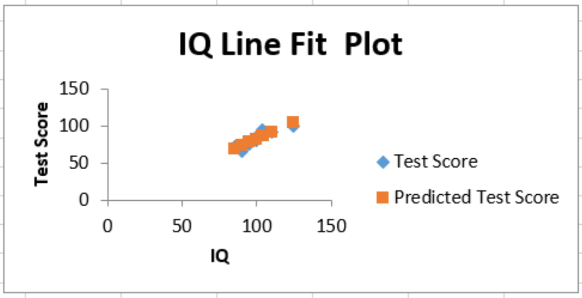
ANOVA					
	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	893.6697	893.6697	33.10957	0.000427
Residual	8	215.9303	26.99128		
Total	9	1109.6			

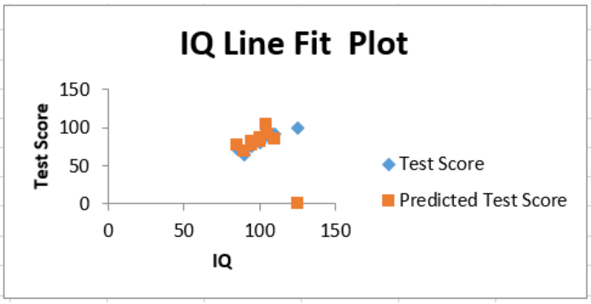
	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	-6.41566	15.66066	-0.40967	0.6928	-42.5292	29.6979	-42.5292	29.6979
IQ	0.888163	0.154353	5.754091	0.000427	0.532224	1.244103	0.532224	1.244103

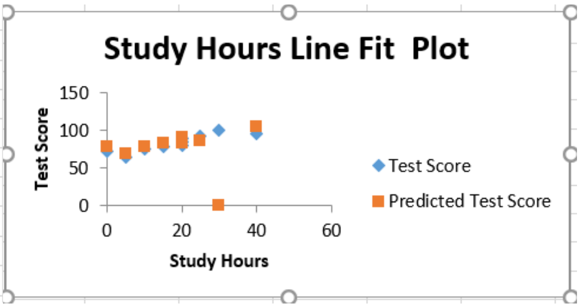
RESIDUAL OUTPUT

Observation	*icted Test S*	*Residuals*
1	104.6047	-4.60473
2	85.95331	9.046694
3	91.28228	0.717716
4	86.84147	3.158531
5	82.40065	2.599347
6	82.40065	-2.40065
7	77.95984	0.040162
8	77.95984	-2.95984
9	69.07821	2.921794
10	73.51902	-8.51902









Practical No: 5

Aim: Perform the Random Sampling for the given data and analyze it.

Theory:

All probability sampling techniques are based on chance selection procedures. Because the probability sampling process is random, the bias inherent in nonprobability sampling procedures is eliminated.

Simple Random Sampling:

The sampling procedure that ensures each element in the population will have an equal chance of being included in the sample is called simple random sampling. Examples include drawing names from a hat and selecting the winning raffle ticket from a large drum. If the names or raffle tickets are thoroughly stirred, each person or ticket should have an equal chance of being selected. In contrast to other, more complex types of probability sampling, this process is simple because it requires only one stage of sample selection.

Although drawing names or numbers out of a fishbowl, using a spinner, rolling dice, or turning a roulette wheel may be an appropriate way to draw a sample from a small population, when populations consist of large numbers of elements, sample selection is based on tables of random numbers or computer-generated random numbers.

Stratified Sampling:

- The usefulness of dividing the population into subgroups, or strata, whose members are more or less equal with respect to some characteristic was illustrated in our discussion of quota sampling.
- The first step is the same for both stratified and quota sampling: choosing strata on the basis of existing information—for example, classifying retail outlets based on annual sales volume. However, the process of selecting sampling units within the strata differs substantially. In stratified sampling, a subsample is drawn using simple random sampling within each stratum. This is not true of quota sampling.
- The reason for taking a stratified sample is to obtain a more efficient sample than would be possible with simple random sampling. Suppose, for example, that urban and rural groups have widely different attitudes toward energy conservation, but members within each group hold very similar attitudes. Random sampling error will be reduced with the use of stratified sampling, because each group is internally homogeneous but there are comparative differences between groups. More technically, a smaller standard error may result from this stratified sampling because the groups will be adequately represented when strata are combined.
- Another reason for selecting a stratified sample is to ensure that the sample will accurately reflect the population on the basis of the criterion or criteria used for stratification. This is a concern because occasionally simple random sampling yields a disproportionate number of one group or another and the sample ends up being less representative than it could be.
- A researcher can select a stratified sample as follows. First, a variable (sometimes several variables) is identified as an efficient basis for stratification. A stratification

variable must be a characteristic of the population elements known to be related to the dependent variable or other variables of interest. The variable chosen should increase homogeneity within each stratum and increase heterogeneity between strata. The stratification variable usually is a categorical variable or one easily converted into categories (that is, subgroups). For example, a pharmaceutical company interested in measuring how often physicians prescribe a certain drug might choose physicians' training as a basis for stratification. In this example the mutually exclusive strata are MDs (medical doctors) and ODs (osteopathic doctors).

- Next, for each separate subgroup or stratum, a list of population elements must be obtained. (If such lists are not available, they can be costly to prepare, and if a complete listing is not available, a true stratified probability sample cannot be selected.) Using a table of random numbers or some other device, a separate simple random sample is then taken within each stratum.

Methods Used:

- **plt.rcParams():** An instance of RcParams for handling default Matplotlib values.
- **sns.color_palette():** This function provides an interface to most of the possible ways that one can generate color palettes in seaborn. And it's used internally by any function that has a palette argument.
- **sns.set_style('darkgrid'):** Darkgrid appear on the sides of the plot on setting it as set_style('darkgrid'). Palette attribute is used to set the color of the bars. It helps to distinguish between chunks of data.
- **sns.color_palette():** This function provides an interface to most of the possible ways that one can generate color palettes in seaborn. And it's used internally by any function that has a palette argument.
- **corr():** It is used to find the pairwise correlation of all columns in the dataframe. Any na values are automatically excluded.
- **plt.subplots():** pyplot, subplots creates a figure and a grid of subplots with a single call, while providing reasonable control over how the individual plots are created.
- **sns.heatmap():** Heatmaps are used in various forms of analytics but are most commonly used to show user behaviour on specific webpages or webpage templates. Heatmaps can be used to show where users have clicked on a page, how far they have scrolled down a page, or used to display the results of eye-tracking tests.
- **sns.distplot():** It is used to plot the distplot. The distplot represents the univariate distribution of data i.e. data distribution of a variable against the density distribution.

Steps:

- Open the Data in an Excel File.

	A	B	C	D	E	F	G	H	I	J
1	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_in	median_house_value	ocean_proximity
2	-118.49	34.02	27	4725	1185	1945	1177	4.1365	470800	<1H OCEAN
3	-119.95	37.47	32	1312	315	600	265	1.5	91500	INLAND
4	-121.76	37.29	15	2267	348	1150	327	7.1267	277900	<1H OCEAN
5	-121.49	38.56	52	1777	368	624	350	3.6729	137800	INLAND
6	-119.73	34.35	20	1648	319	905	307	4.375	335200	NEAR OCEAN
7	-118.15	34.15	49	806	199	698	172	2.3654	137500	<1H OCEAN
8	-122.22	37.81	52	1971	335	765	308	6.5217	273700	NEAR BAY
9	-118.35	34.17	47	858	170	365	171	2.0385	225000	<1H OCEAN
10	-121.44	38.63	38	1402	370	970	382	1.6343	71000	INLAND
11	-117.06	34.87	14	3348	619	1756	557	3.5987	91400	INLAND
12	-118.14	34.09	20	3447	1007	2622	934	2.918	208700	<1H OCEAN
13	-120.93	39.96	15	1666	351	816	316	2.9559	118800	INLAND
14	-121.95	37.78	4	14652	2826	5613	2579	6.3942	356700	<1H OCEAN
15	-121.52	37.75	18	1544	272	825	286	4.3229	327300	INLAND
16	-118.5	34.05	36	4152	542	1461	550	15.0001	500001	<1H OCEAN
17	-120.88	38.58	8	3417	604	1703	623	4.0827	170700	INLAND
18	-122.32	41.31	45	1393	294	521	249	1.1915	71900	INLAND
19	-121.15	38.89	20	2024	313	879	309	5.2903	239400	INLAND
20	-117.95	33.9	15	3057	479	1679	498	6.8429	372600	<1H OCEAN
21	-121.62	41.78	40	3272	663	1467	553	1.7885	43500	INLAND
22	-121.13	37.74	28	409	104	244	98	3.4643	90900	INLAND
23	-117.1	32.68	49	1412	350	1200	332	2.0398	93600	NEAR OCEAN
24	-121.48	38.49	26	3165	806	2447	752	1.5908	78600	INLAND
25	-122.15	37.45	52	568	91	219	75	6.1575	500001	NEAR BAY
26	-118.04	33.83	19	4526	830	2318	748	4.6681	320700	<1H OCEAN
27	-118.34	34.19	43	1029	252	613	255	2.6827	219900	<1H OCEAN
28	-122.31	37.56	45	1685	321	815	314	4.2955	309700	NEAR OCEAN
29	-118.15	34.21	34	2765	515	1422	438	5.4727	238900	INLAND

housing Sheet1

- Insert a column after total_bedrooms → label it as "Random" → Apply rand() Function to "Random" Column.

	A	B	C	D	E	F	G	H	I	J	K
1	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	Random	population	households	median_in	median_house_value	ocean_proximity
2	-118.49	34.02	27	4725	1185		1945	1177	4.1365	470800	<1H OCEAN
3	-119.95	37.47	32	1312	315		600	265	1.5	91500	INLAND
4	-121.76	37.29	15	2267	348		1150	327	7.1267	277900	<1H OCEAN
5	-121.49	38.56	52	1777	368		624	350	3.6729	137800	INLAND
6	-119.73	34.35	20	1648	319		905	307	4.375	335200	NEAR OCEAN
7	-118.15	34.15	49	806	199		698	172	2.3654	137500	<1H OCEAN
8	-122.22	37.81	52	1971	335		765	308	6.5217	273700	NEAR BAY
9	-118.35	34.17	47	858	170		365	171	2.0385	225000	<1H OCEAN
10	-121.44	38.63	38	1402	370		970	382	1.6343	71000	INLAND
11	-117.06	34.87	14	3348	619		1756	557	3.5987	91400	INLAND
12	-118.14	34.09	20	3447	1007		2622	934	2.918	208700	<1H OCEAN
13	-120.93	39.96	15	1666	351		816	316	2.9559	118800	INLAND
14	-121.95	37.78	4	14652	2826		5613	2579	6.3942	356700	<1H OCEAN
15	-121.52	37.75	18	1544	272		825	286	4.3229	327300	INLAND
16	-118.5	34.05	36	4152	542		1461	550	15.0001	500001	<1H OCEAN
17	-120.88	38.58	8	3417	604		1703	623	4.0827	170700	INLAND
18	-122.32	41.31	45	1393	294		521	249	1.1915	71900	INLAND
19	-121.15	38.89	20	2024	313		879	309	5.2903	239400	INLAND
20	-117.95	33.9	15	3057	479		1679	498	6.8429	372600	<1H OCEAN
21	-121.62	41.78	40	3272	663		1467	553	1.7885	43500	INLAND
22	-121.13	37.74	28	409	104		244	98	3.4643	90900	INLAND
23	-117.1	32.68	49	1412	350		1200	332	2.0398	93600	NEAR OCEAN
24	-121.48	38.49	26	3165	806		2447	752	1.5908	78600	INLAND
25	-122.15	37.45	52	568	91		219	75	6.1575	500001	NEAR BAY
26	-118.04	33.83	19	4526	830		2318	748	4.6681	320700	<1H OCEAN
27	-118.34	34.19	43	1029	252		613	255	2.6827	219900	<1H OCEAN
28	-122.31	37.56	45	1685	321		815	314	4.2955	309700	NEAR OCEAN
29	-118.15	34.21	34	2765	515		1422	438	5.4727	238900	INLAND

housing Sheet1

- Use rand() Function to generate random numbers.

	A	B	C	D	E	F	G	H	I	J	K
1	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	Random	population	households	median_in	median_house_value	ocean_proximity
2	-118.49	34.02	27	4725	1185	0.926961914	1945	1177	4.1365	470800	<1H OCEAN
3	-119.95	37.47	32	1312	315	0.270141266	600	265	1.5	91500	INLAND
4	-121.76	37.29	15	2267	348	0.981698662	1150	327	7.1267	277900	<1H OCEAN
5	-121.49	38.56	52	1777	368	0.915638017	624	350	3.6729	137800	INLAND
6	-119.73	34.35	20	1648	319	0.455158857	905	307	4.375	335200	NEAR OCEAN
7	-118.15	34.15	49	806	199	0.929315412	698	172	2.3654	137500	<1H OCEAN
8	-122.22	37.81	52	1971	335	0.856667146	765	308	6.5217	273700	NEAR BAY
9	-118.35	34.17	47	858	170	0.862154838	365	171	2.0385	225000	<1H OCEAN
10	-121.44	38.63	38	1402	370	0.185929876	970	382	1.6343	71000	INLAND
11	-117.06	34.87	14	3348	619	0.536309746	1756	557	3.5987	91400	INLAND
12	-118.14	34.09	20	3447	1007	0.137314182	2622	934	2.918	208700	<1H OCEAN
13	-120.93	39.96	15	1666	351	0.755716815	816	316	2.9559	118800	INLAND
14	-121.95	37.78	4	14652	2826	0.002237073	5613	2579	6.3942	356700	<1H OCEAN
15	-121.52	37.75	18	1544	272	0.344654716	825	286	4.3229	327300	INLAND
16	-118.5	34.05	36	4152	542	0.589714512	1461	550	15.0001	500001	<1H OCEAN
17	-120.88	38.58	8	3417	604	0.681910231	1703	623	4.0827	170700	INLAND
18	-122.32	41.31	45	1393	294	0.847328083	521	249	1.1915	71900	INLAND
19	-121.15	38.89	20	2024	313	0.360885972	879	309	5.2903	239400	INLAND
20	-117.95	33.9	15	3057	479	0.35258545	1679	498	6.8429	372600	<1H OCEAN
21	-121.62	41.78	40	3272	663	0.799390933	1467	553	1.7885	43500	INLAND
22	-121.13	37.74	28	409	104	0.501913083	244	98	3.4643	90900	INLAND
23	-117.1	32.68	49	1412	350	0.3865471	1200	332	2.0398	93600	NEAR OCEAN
24	-121.48	38.49	26	3165	806	0.459021199	2447	752	1.5908	78600	INLAND
25	-122.15	37.45	52	568	91	0.047981063	219	75	6.1575	500001	NEAR BAY
26	-118.04	33.83	19	4526	830	0.887348873	2318	748	4.6681	320700	<1H OCEAN
27	-118.34	34.19	43	1029	252	0.331881371	613	255	2.6827	219900	<1H OCEAN
28	-122.31	37.56	45	1685	321	0.63064047	815	314	4.2955	309700	NEAR OCEAN
29	-118.15	34.21	34	2765	515	0.966709926	1422	438	5.4727	238900	INLAND

housing Sheet1

- Sort the "Random" Column.

F
Random
0.163554876
0.920919712
0.806609665
0.189370409
0.472226519
0.681769855
0.064506841
0.22385648
0.063461706
0.831534805
0.338574006
0.884926596
0.985649517
0.202429326
0.680944572
0.536907316
0.117174622
0.812115687
0.251631875
0.702301906
0.154152506
0.249175009
0.804467644
0.595162055
0.160725968
0.908466674
0.965296918
0.680069648

- Select First 100 Rows.

	A	B	C	D	E	F	G	H	I	J	K
1	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	Random	population	households	median_in	median_house_value	ocean_proximity
2	-118.45	34.06	20	3367	1264	0.163554876	2667	1131	2.2444	500000	<1H OCEAN
3	-118.42	34.05	52	2533	402	0.920919712	981	386	7.8164	500001	<1H OCEAN
4	-118.04	34.07	39	2451	649	0.806609665	2536	648	2.3098	173100	<1H OCEAN
5	-119.64	36.35	30	1765	310	0.189370409	746	298	2.8125	70200	INLAND
6	-118.15	34.59	33	2111	429	0.472226519	1067	397	3.7344	111400	INLAND
7	-121.2	37.97	39	440	83	0.681769855	270	97	6.0582	157700	INLAND
8	-124.16	40.6	39	1322	283	0.064506841	642	292	2.4519	85100	NEAR OCEAN
9	-121.65	39.32	40	812	154	0.22385648	374	142	2.7891	73500	INLAND
10	-119.59	36.57	19	1733	303	0.063461706	911	281	3.5987	131700	INLAND
11	-117.12	32.75	17	2060	633	0.831534805	1251	602	1.9886	119200	NEAR OCEAN
12	-117.72	34.03	17	2902	476	0.338574006	1652	479	5.6029	161800	INLAND
13	-122.64	38.87	16	1177	240	0.884926596	519	199	1.5739	73500	INLAND
14	-122.23	38.1	46	4143	895	0.985649517	2240	847	2.4201	92800	NEAR BAY
15	-117.06	32.72	31	2669	514	0.202429326	1626	499	3.1923	116900	NEAR OCEAN
16	-121.3	38.64	20	5001	830	0.680944572	2330	830	4.0833	160000	INLAND
17	-118.28	33.96	37	1812	500	0.536907316	1640	447	1.9348	99100	<1H OCEAN
18	-119.18	34.18	31	2636	638	0.117174622	2695	614	3.2196	175800	NEAR OCEAN
19	-118.45	34.07	13	4284	1452	0.812115687	3806	1252	1.3125	350000	<1H OCEAN
20	-122.27	37.9	52	2079	273	0.251631875	684	275	7.9556	374400	NEAR BAY
21	-117.93	33.86	36	1672	318	0.702301906	1173	337	4.5774	182100	<1H OCEAN
22	-121.87	36.55	20	10053	1768	0.154152506	3083	1621	5.1506	387500	NEAR OCEAN
23	-117.84	33.75	16	4367	1161	0.249175009	2164	1005	4.0214	139500	<1H OCEAN
24	-117.89	33.92	17	2936	555	0.804467644	1381	535	5.4617	190300	<1H OCEAN
25	-117.02	32.66	19	771		0.595162055	376	108	6.6272	273600	NEAR OCEAN
26	-117.16	32.71	5	2508	827	0.160725968	2066	761	1.3092	325000	NEAR OCEAN
27	-117.9	33.65	24	4496	877	0.908466674	1928	855	4.6808	245500	<1H OCEAN
28	-118.23	33.9	45	1285	238	0.965296918	840	211	3.4107	112500	<1H OCEAN
29	-116.92	32.76	9	1859	307	0.680069648	947	304	5.9202	181300	<1H OCEAN

housing Sheet1

Practical No: 6

Part A:

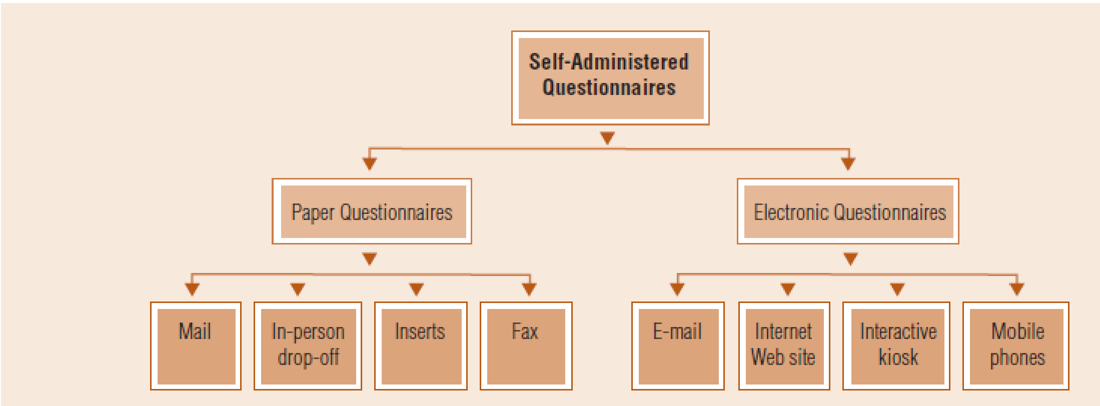
Aim: Design a survey form for a case study, collect the primary data and analyse it.

Theory:

Primary Data Analysis:

Many surveys do not require an interviewer's presence. Researchers distribute questionnaires to consumers through the mail and in many other ways (see Exhibit 10.1). They insert questionnaires in packages and magazines. They may place questionnaires at points of purchase or in high-traffic locations in stores or malls. They may even fax questionnaires to individuals. Questionnaires can be printed on paper, but they may be posted on the Internet or sent via e-mail. No matter how the self-administered questionnaires are distributed, they are different from interviews because the respondent takes responsibility for reading and answering the questions.

EXHIBIT 10.1 Self-Administered Questionnaires Can Be Either Printed or Electronic



e-mail surveys: Surveys distributed through electronic mail.
Internet survey: A self-administered questionnaire posted on a Web site.

Questionnaire:

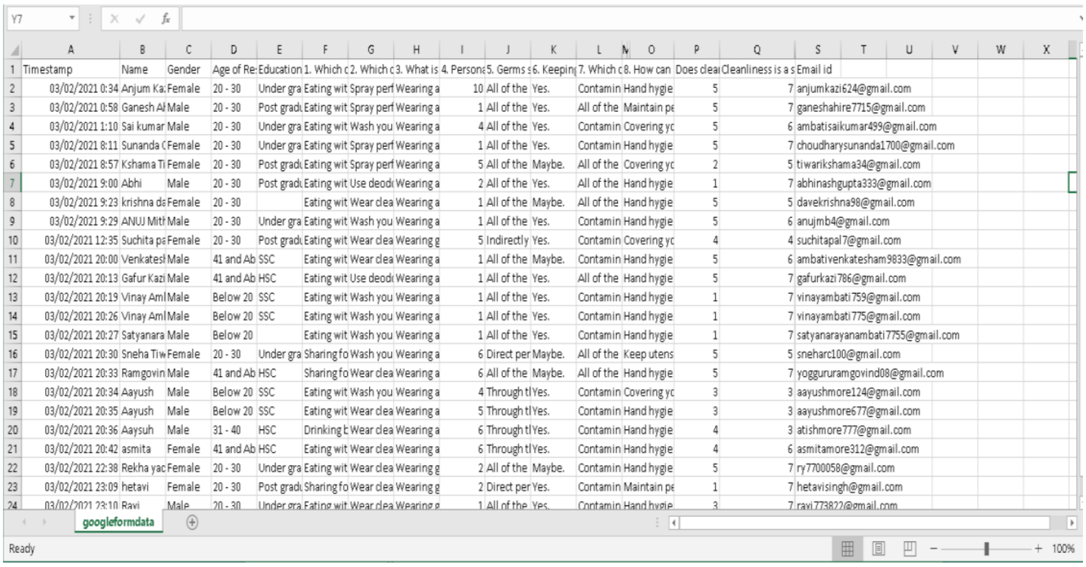
The research questionnaire, development stage is critically important as the information provided is only as good as, the questions asked. However, the importance of question wording is easily, and far too often, overlooked. Businesspeople who are inexperienced at research frequently believe that constructing a questionnaire is a simple task. Amateur researchers think a short questionnaire can be written in minutes. Unfortunately, newcomers who naively believe that good grammar is all a person needs to construct a questionnaire generally end up with useless results. Ask a bad question, get bad results. Good questionnaire design requires far more than correct grammar. People don't understand questions just because they are grammatically correct. Respondents simply may not know what is being asked. They may be unaware of the business issue or topic of interest. They may confuse the subject with something else. The question may not mean the same thing to everyone interviewed. Finally, people may refuse to answer personal questions. Most of these problems can be minimized, however, if a skilled researcher composes the questionnaire.

For a questionnaire to fulfill a researcher's purposes, the questions must meet the basic criteria of **relevance and accuracy.** To achieve these ends, a researcher who is systematically planning a questionnaire's design will be required to make several decisions:

- What should be asked?
- How should questions be phrased?
- In what sequence should the questions be arranged?
- What questionnaire layout will best serve the research objectives?
- How should the questionnaire be pretested? Does the questionnaire need to be revised?

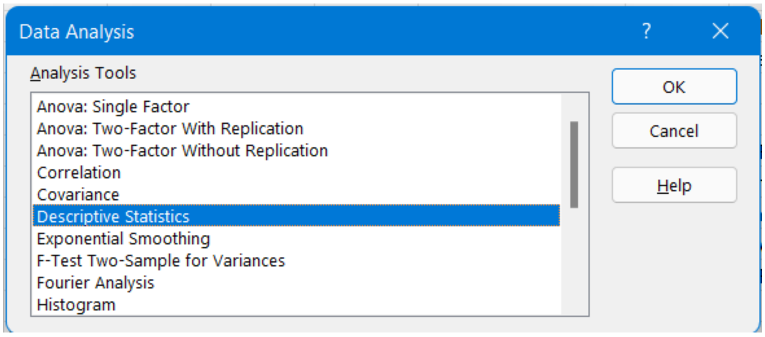
Steps:

1. Open browser → search for google form → click on google form survey for personal use →make google form for questionnaire →download the data into the Excel and name file with extension (.csv)
2. Display the collected data in excel:

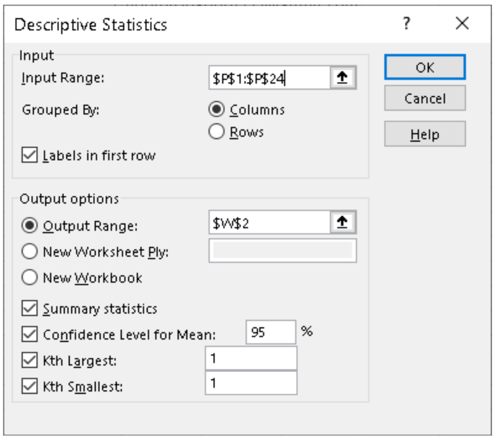


Timestamp	Name	Gender	Age of Re	Education	1. Which c	2. Which c	3. What is	4. Persona	5. Germs s	6. Keepin	7. Which c	8. How can	Does clea	Cleanliness is a s	Email id
03/02/2021 0:34	Anjum Ka	Female	20 - 30	Under gra	Eating wit	Spray perf	Wearing a	10	All of the	Yes.	Contamin	Hand hygie	5	7	anjumkazi624@gmail.com
03/02/2021 0:58	Ganesh Al	Male	20 - 30	Post gradu	Eating wit	Spray perf	Wearing a	1	All of the	Yes.	All of the	Maintain pe	5	7	ganeshahire7715@gmail.com
03/02/2021 1:10	Sai kumar	Male	20 - 30	Under gra	Eating wit	Wash you	Wearing a	4	All of the	Yes.	Contamin	Covering yc	5	6	ambatisaikumar499@gmail.com
03/02/2021 8:11	Sunanda (	Female	20 - 30	Under gra	Eating wit	Spray perf	Wearing a	1	All of the	Yes.	Contamin	Hand hygie	5	7	choudharysunanda1700@gmail.com
03/02/2021 8:57	Kshama Ti	Female	20 - 30	Post gradu	Eating wit	Spray perf	Wearing a	5	All of the	Maybe.	All of the	Covering yc	2	5	tiwarikshama34@gmail.com
03/02/2021 9:00	Abhi	Male	20 - 30	Post gradu	Eating wit	Use deodo	Wearing a	2	All of the	Yes.	All of the	Hand hygie	1	7	abhinashgupta333@gmail.com
03/02/2021 9:23	krishna da	Female	20 - 30		Eating wit	Wear clea	Wearing a	1	All of the	Maybe.	All of the	Hand hygie	5	5	davekrishna98@gmail.com
03/02/2021 9:29	ANUJ Mith	Male	20 - 30	Under gra	Eating wit	Wash you	Wearing a	1	All of the	Yes.	Contamin	Hand hygie	5	6	anujmb4@gmail.com
03/02/2021 12:35	Suchita pa	Female	20 - 30	Post gradu	Eating wit	Wear clea	Wearing g	5	Indirectly	Yes.	Contamin	Covering yc	4	4	suchitapal7@gmail.com
03/02/2021 20:00	Venkatesh	Male	41 and Ab	SSC	Eating wit	Wear clea	Wearing a	1	All of the	Maybe.	Contamin	Hand hygie	5	6	ambativenkatesham9833@gmail.com
03/02/2021 20:13	Gafur Kazi	Male	41 and Ab	HSC	Eating wit	Use deodo	Wearing a	1	All of the	Yes.	All of the	Hand hygie	5	7	gafurkazi786@gmail.com
03/02/2021 20:19	Vinay Aml	Male	Below 20	SSC	Eating wit	Wash you	Wearing a	1	All of the	Yes.	Contamin	Hand hygie	1	7	vinayambati759@gmail.com
03/02/2021 20:26	Vinay Aml	Male	Below 20	SSC	Eating wit	Wash you	Wearing a	1	All of the	Yes.	Contamin	Hand hygie	1	7	vinayambati775@gmail.com
03/02/2021 20:27	Satyanara	Male	Below 20		Eating wit	Wash you	Wearing a	1	All of the	Yes.	Contamin	Hand hygie	1	7	satyanarayanambati7755@gmail.com
03/02/2021 20:30	Sneha Tiw	Female	20 - 30	Under gra	Sharing fo	Wash you	Wearing a	6	Direct per	Maybe.	All of the	Keep utens	5	5	sneharc100@gmail.com
03/02/2021 20:33	Ramgovin	Male	41 and Ab	HSC	Sharing fo	Wear clea	Wearing a	6	All of the	Maybe.	All of the	Hand hygie	5	7	yoggururamgovind08@gmail.com
03/02/2021 20:34	Aayush	Male	Below 20	SSC	Eating wit	Wash you	Wearing a	4	Through tl	Yes.	Contamin	Covering yc	3	3	aayushmore124@gmail.com
03/02/2021 20:35	Aayush	Male	Below 20	SSC	Eating wit	Wear clea	Wearing a	5	Through tl	Yes.	Contamin	Hand hygie	3	3	aayushmore677@gmail.com
03/02/2021 20:36	Aaysuh	Male	31 - 40	HSC	Drinking b	Wear clea	Wearing a	6	Through tl	Yes.	Contamin	Hand hygie	4	3	atishmore777@gmail.com
03/02/2021 20:42	asmita	Female	41 and Ab	HSC	Eating wit	Wear clea	Wearing a	6	Through tl	Yes.	Contamin	Hand hygie	4	6	asmitamore312@gmail.com
03/02/2021 22:38	Rekha yac	Female	20 - 30	Under gra	Eating wit	Wear clea	Wearing g	2	All of the	Maybe.	Contamin	Hand hygie	5	7	ry7700058@gmail.com
03/02/2021 23:09	hetavi	Female	20 - 30	Post gradu	Sharing fo	Wear clea	Wearing g	2	Direct per	Yes.	Contamin	Maintain pe	1	7	hetavisingh@gmail.com
03/02/2021 23:10	Ravi	Male	20 - 30	Under gra	Eating wit	Wear clea	Wearing g	1	All of the	Yes.	Contamin	Hand hygie	3	7	ravi773822@gmail.com

3. Go to data tab → click on data analysis →select Descriptive analysis → ok



4. In the open popup window do the following:
   a) select the input range
   b) select output range radio button and click on a cell for output
   c) select the summary statistic, confidence level for mean, kth largest, kth smallest checkboxes and click on ok.



Output:

Does cleanliness affects the environment	
Mean	3.608695652
Standard Error	0.342849778
Median	4
Mode	5
Standard Deviation	1.644249772
Sample Variance	2.703557312
Kurtosis	-1.168978733
Skewness	-0.714497305
Range	4
Minimum	1
Maximum	5
Sum	83
Count	23
Largest(1)	5
Smallest(1)	1
Confidence Level(95.0%)	0.71102692

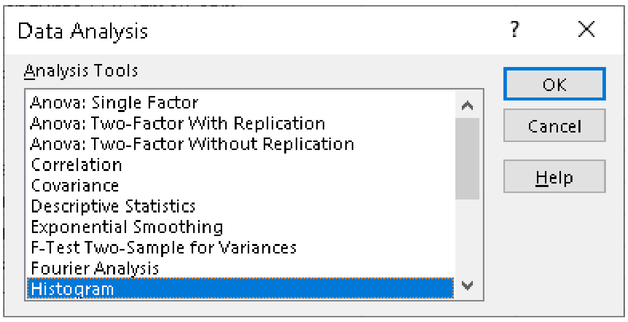
Plot Charts:

Steps:

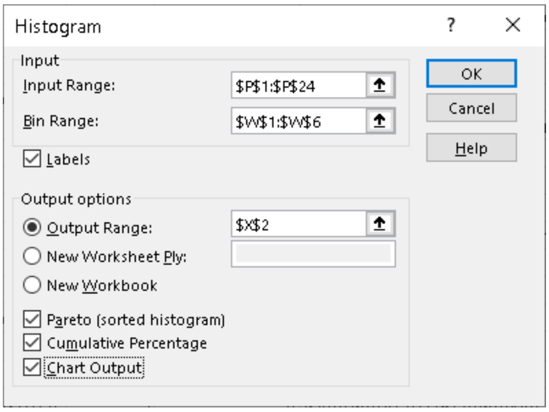
1. Create new column (name- option)- which column we are selected check out the number which are in the selected column → then write in “Option column”.

option
1
2
3
4
5

2. Go to the data tab → click on data analysis → Select histogram → OK

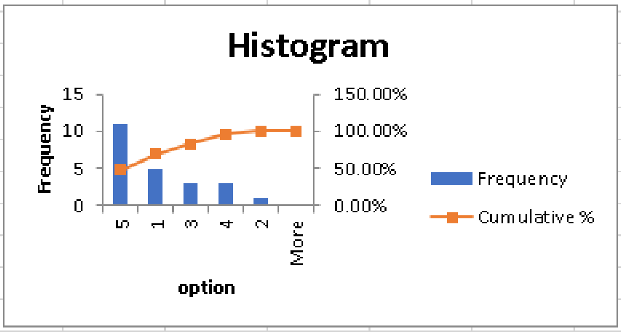


3. In input range → select the “Does cleanliness affects the environment ” and in bin range →“option column”→ tick all the checkboxes below.



Output:

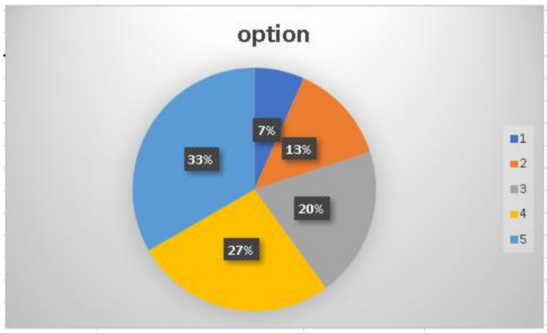
option	Frequency	Cumulative %	option	Frequency	Cumulative %
1	5	21.74%	5	11	47.83%
2	1	26.09%	1	5	69.57%
3	3	39.13%	3	3	82.61%
4	3	52.17%	4	3	95.65%
5	11	100.00%	2	1	100.00%
More	0	100.00%	More	0	100.00%



4. For pie chart→ select option column.

option
1
2
3
4
5

5. Select Option column on given output → go to the insert tab → click on the pie chart.



Part B:

Aim: **Perform suitable analysis of given secondary data.**

Theory:

Secondary Data Analysis:

Data that have been previously collected for some purpose other than the one at hand.

EXHIBIT 8.2

Common Research Objectives for Secondary-Data Studies

Broad Objective	Specific Research Example
Fact-finding	Identifying consumption patterns Tracking trends
Model building	Estimating market potential Forecasting sales Selecting trade areas and sites
Database marketing	Enhancing customer databases Developing prospect lists

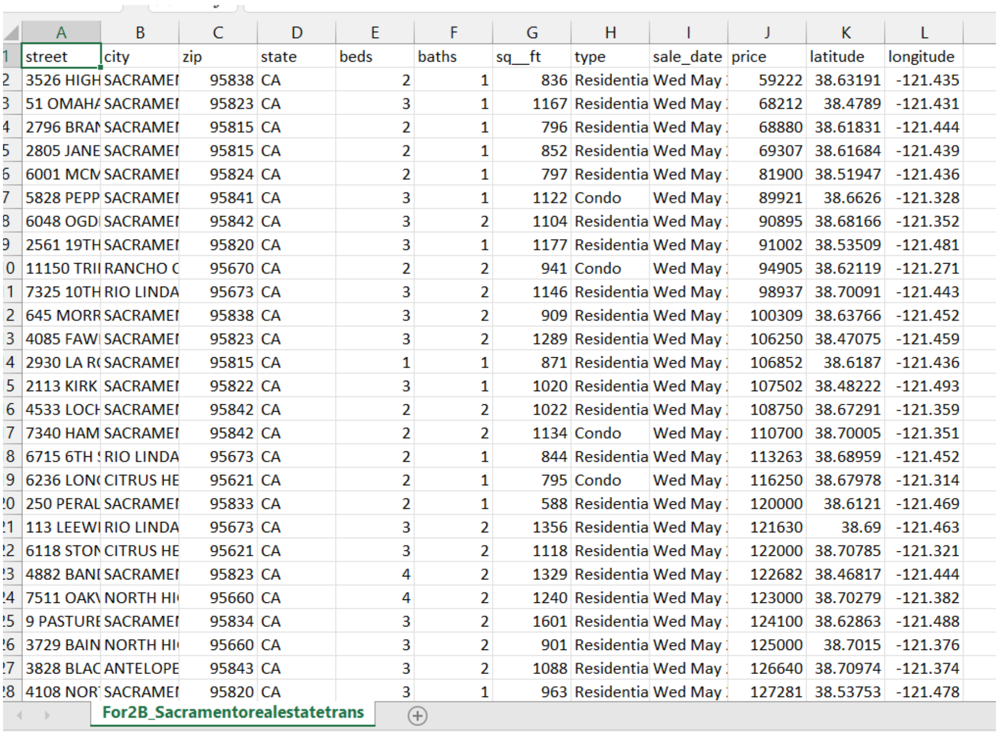
Steps in Secondary Data Analysis:

1. **Determine your research question:** Knowing exactly what you are looking for.
2. **Locating data:** Knowing what is out there and whether you can gain access to it. A quick Internet search, possibly with the help of a librarian, will reveal a wealth of options.
3. **Evaluating relevance of the data:** Considering things like the data's original purpose, when it was collected, population, sampling strategy/sample, data collection protocols, operationalization of concepts, questions asked, and form/shape of the data.
4. **Assessing credibility of the data:** Establishing the credentials of the original researchers, searching for full explication of methods including any problems encountered, determining how consistent the data is with data from other sources, and discovering whether the data has been used in any credible published research.
5. **Analysis:** This will generally involve a range of statistical processes.

Descriptive Analysis on Beds:

Steps:

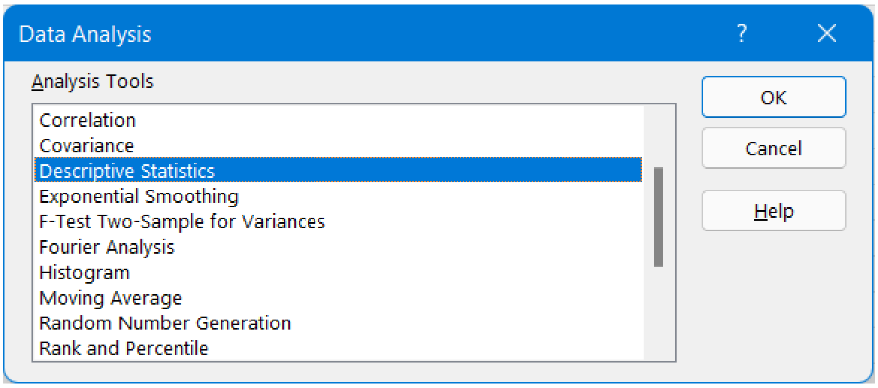
1. Open the data set in the excel.



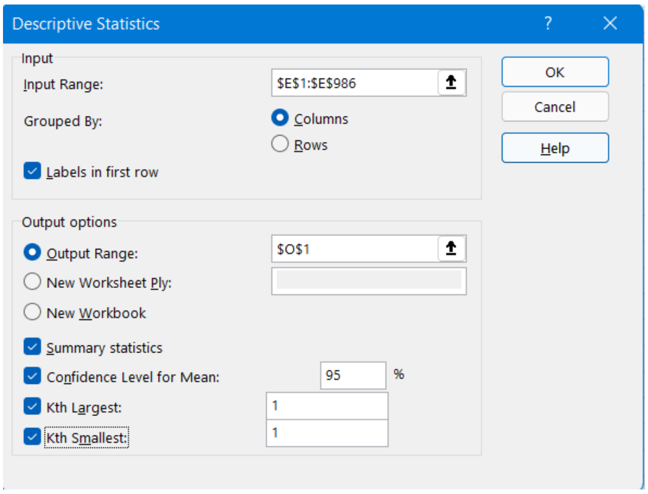
	A	B	C	D	E	F	G	H	I	J	K	L
1	street	city	zip	state	beds	baths	sq__ft	type	sale_date	price	latitude	longitude
2	3526 HIGH	SACRAMEI	95838	CA	2	1	836	Residentia	Wed May	59222	38.63191	-121.435
3	51 OMAHA	SACRAMEI	95823	CA	3	1	1167	Residentia	Wed May	68212	38.4789	-121.431
4	2796 BRAI	SACRAMEI	95815	CA	2	1	796	Residentia	Wed May	68880	38.61831	-121.444
5	2805 JANE	SACRAMEI	95815	CA	2	1	852	Residentia	Wed May	69307	38.61684	-121.439
6	6001 MCM	SACRAMEI	95824	CA	2	1	797	Residentia	Wed May	81900	38.51947	-121.436
7	5828 PEPP	SACRAMEI	95841	CA	3	1	1122	Condo	Wed May	89921	38.6626	-121.328
8	6048 OGDI	SACRAMEI	95842	CA	3	2	1104	Residentia	Wed May	90895	38.68166	-121.352
9	2561 19TH	SACRAMEI	95820	CA	3	1	1177	Residentia	Wed May	91002	38.53509	-121.481
0	11150 TRII	RANCHO C	95670	CA	2	2	941	Condo	Wed May	94905	38.62119	-121.271
1	7325 10TH	RIO LINDA	95673	CA	3	2	1146	Residentia	Wed May	98937	38.70091	-121.443
2	645 MORR	SACRAMEI	95838	CA	3	2	909	Residentia	Wed May	100309	38.63766	-121.452
3	4085 FAWI	SACRAMEI	95823	CA	3	2	1289	Residentia	Wed May	106250	38.47075	-121.459
4	2930 LA R(	SACRAMEI	95815	CA	1	1	871	Residentia	Wed May	106852	38.6187	-121.436
5	2113 KIRK	SACRAMEI	95822	CA	3	1	1020	Residentia	Wed May	107502	38.48222	-121.493
6	4533 LOCH	SACRAMEI	95842	CA	2	2	1022	Residentia	Wed May	108750	38.67291	-121.359
7	7340 HAM	SACRAMEI	95842	CA	2	2	1134	Condo	Wed May	110700	38.70005	-121.351
8	6715 6TH	RIO LINDA	95673	CA	2	1	844	Residentia	Wed May	113263	38.68959	-121.452
9	6236 LON(	CITRUS HE	95621	CA	2	1	795	Condo	Wed May	116250	38.67978	-121.314
20	250 PERAL	SACRAMEI	95833	CA	2	1	588	Residentia	Wed May	120000	38.6121	-121.469
21	113 LEEWI	RIO LINDA	95673	CA	3	2	1356	Residentia	Wed May	121630	38.69	-121.463
22	6118 STON	CITRUS HE	95621	CA	3	2	1118	Residentia	Wed May	122000	38.70785	-121.321
23	4882 BANI	SACRAMEI	95823	CA	4	2	1329	Residentia	Wed May	122682	38.46817	-121.444
24	7511 OAK\	NORTH HI	95660	CA	4	2	1240	Residentia	Wed May	123000	38.70279	-121.382
25	9 PASTURE	SACRAMEI	95834	CA	3	2	1601	Residentia	Wed May	124100	38.62863	-121.488
26	3729 BAIN	NORTH HI	95660	CA	3	2	901	Residentia	Wed May	125000	38.7015	-121.376
27	3828 BLAC	ANTELOPE	95843	CA	3	2	1088	Residentia	Wed May	126640	38.70974	-121.374
28	4108 NOR	SACRAMEI	95820	CA	3	1	963	Residentia	Wed May	127281	38.53753	-121.478

For2B_Sacramentorealestatetrans

2. Go to the data tab → click on data analysis → Select Descriptive statistics→ OK.



3. In the open popup window do the following:
   a) Select the input range.
   b) Select output range radio button and click on a cell for output.
   c) Select the summary statistic, confidence level for mean, kth largest, kth smallest checkboxes and click on ok.



Output:

beds	
Mean	2.911675127
Standard Error	0.041674186
Median	3
Mode	3
Standard Deviation	1.307932232
Sample Variance	1.710686724
Kurtosis	0.62980724
Skewness	-0.794780303
Range	8
Minimum	0
Maximum	8
Sum	2868
Count	985
Largest(1)	8
Smallest(1)	0
Confidence Level(95.0%)	0.081780496

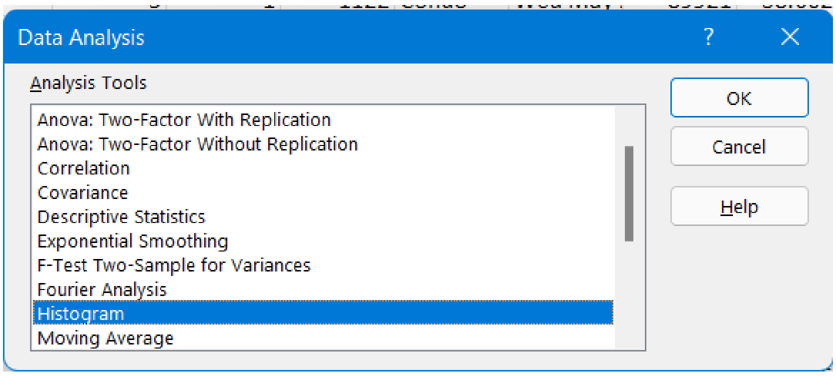
Histogram on Beds:

Steps:

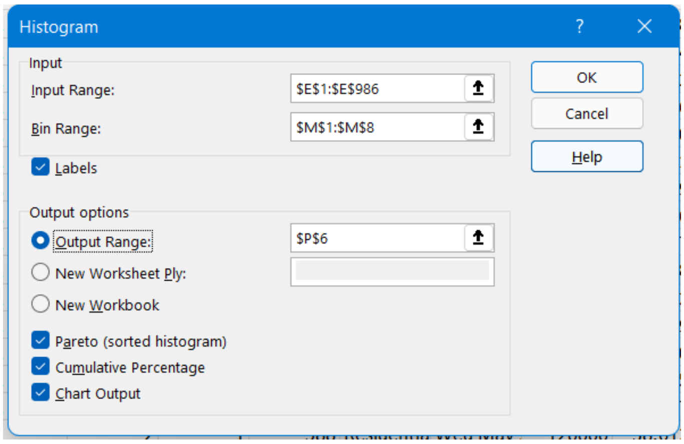
1. Add One Column name=number of beds → enter the values.

number of beds
0
1
2
3
4
5
6

2. Go to the data tab → click on data analysis → Select Histogram → Ok.

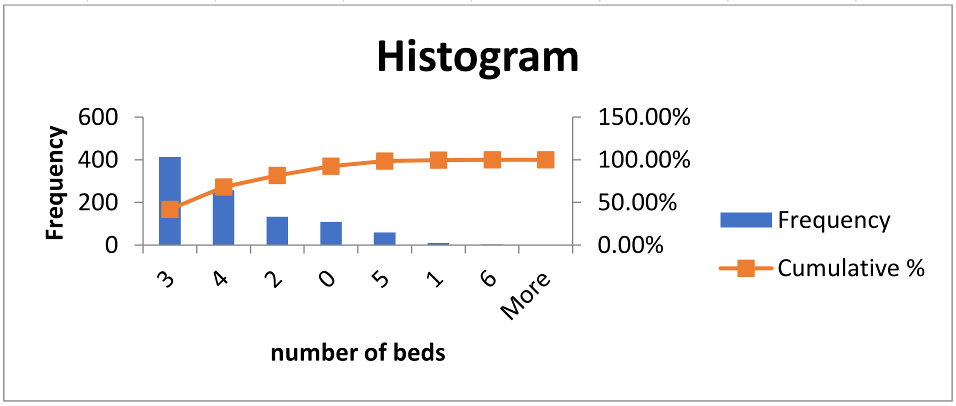


3. In input range → select the "bed " and in bin range → "number of beds"→ tick all the checkboxes below.



Output:

number of be	Frequency	umulative	mber of be	Frequency	umulative %
0	108	10.96%	3	413	41.93%
1	10	11.98%	4	258	68.12%
2	133	25.48%	2	133	81.62%
3	413	67.41%	0	108	92.59%
4	258	93.60%	5	59	98.58%
5	59	99.59%	1	10	99.59%
6	3	99.90%	6	3	99.90%
More	1	100.00%	More	1	100.00%



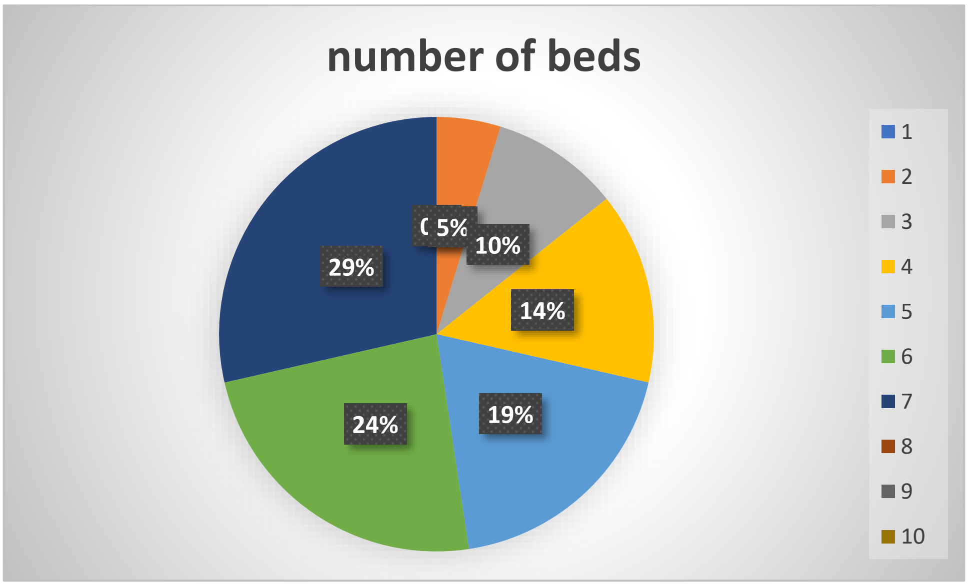
Pie Chart on Beds:

Steps:

1. Add One Column name=number of beds → enter the values.

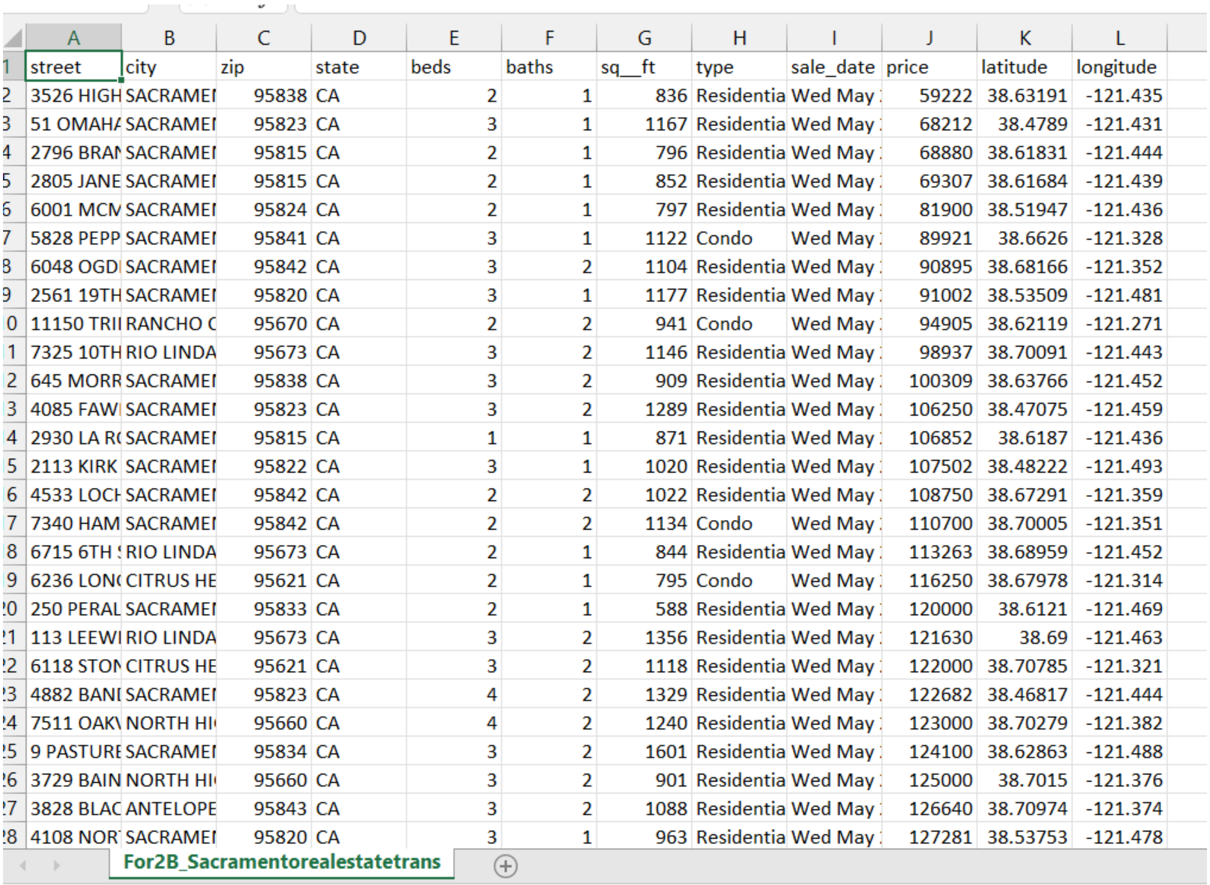
number of beds
0
1
2
3
4
5
6

2. Select the number of bed column → insert tab →select pie chart.



Descriptive Analysis on Baths:

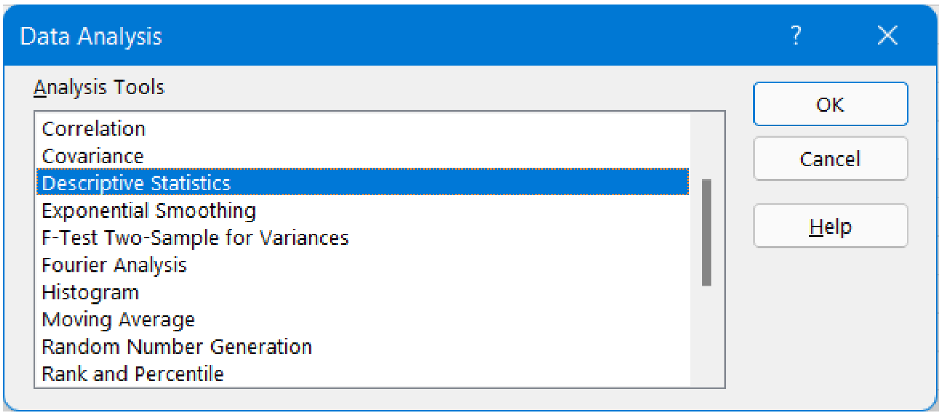
1. Open the data set in the excel.



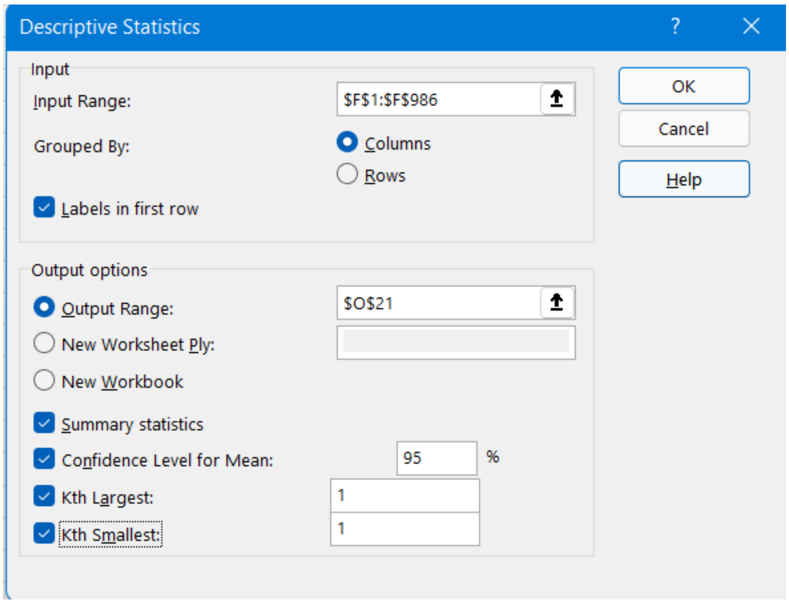
	A	B	C	D	E	F	G	H	I	J	K	L
1	street	city	zip	state	beds	baths	sq__ft	type	sale_date	price	latitude	longitude
2	3526 HIGH	SACRAMEI	95838	CA	2	1	836	Residentia	Wed May	59222	38.63191	-121.435
3	51 OMAHA	SACRAMEI	95823	CA	3	1	1167	Residentia	Wed May	68212	38.4789	-121.431
4	2796 BRAN	SACRAMEI	95815	CA	2	1	796	Residentia	Wed May	68880	38.61831	-121.444
5	2805 JANE	SACRAMEI	95815	CA	2	1	852	Residentia	Wed May	69307	38.61684	-121.439
6	6001 MCM	SACRAMEI	95824	CA	2	1	797	Residentia	Wed May	81900	38.51947	-121.436
7	5828 PEPP	SACRAMEI	95841	CA	3	1	1122	Condo	Wed May	89921	38.6626	-121.328
8	6048 OGD	SACRAMEI	95842	CA	3	2	1104	Residentia	Wed May	90895	38.68166	-121.352
9	2561 19TH	SACRAMEI	95820	CA	3	1	1177	Residentia	Wed May	91002	38.53509	-121.481
0	11150 TRII	RANCHO C	95670	CA	2	2	941	Condo	Wed May	94905	38.62119	-121.271
1	7325 10TH	RIO LINDA	95673	CA	3	2	1146	Residentia	Wed May	98937	38.70091	-121.443
2	645 MORR	SACRAMEI	95838	CA	3	2	909	Residentia	Wed May	100309	38.63766	-121.452
3	4085 FAW	SACRAMEI	95823	CA	3	2	1289	Residentia	Wed May	106250	38.47075	-121.459
4	2930 LA R	SACRAMEI	95815	CA	1	1	871	Residentia	Wed May	106852	38.6187	-121.436
5	2113 KIRK	SACRAMEI	95822	CA	3	1	1020	Residentia	Wed May	107502	38.48222	-121.493
6	4533 LOCH	SACRAMEI	95842	CA	2	2	1022	Residentia	Wed May	108750	38.67291	-121.359
7	7340 HAM	SACRAMEI	95842	CA	2	2	1134	Condo	Wed May	110700	38.70005	-121.351
8	6715 6TH	RIO LINDA	95673	CA	2	1	844	Residentia	Wed May	113263	38.68959	-121.452
9	6236 LON	CITRUS HE	95621	CA	2	1	795	Condo	Wed May	116250	38.67978	-121.314
0	250 PERAL	SACRAMEI	95833	CA	2	1	588	Residentia	Wed May	120000	38.6121	-121.469
1	113 LEEWI	RIO LINDA	95673	CA	3	2	1356	Residentia	Wed May	121630	38.69	-121.463
2	6118 STON	CITRUS HE	95621	CA	3	2	1118	Residentia	Wed May	122000	38.70785	-121.321
3	4882 BANI	SACRAMEI	95823	CA	4	2	1329	Residentia	Wed May	122682	38.46817	-121.444
4	7511 OAK	NORTH HI	95660	CA	4	2	1240	Residentia	Wed May	123000	38.70279	-121.382
5	9 PASTURE	SACRAMEI	95834	CA	3	2	1601	Residentia	Wed May	124100	38.62863	-121.488
6	3729 BAIN	NORTH HI	95660	CA	3	2	901	Residentia	Wed May	125000	38.7015	-121.376
7	3828 BLAC	ANTELOPE	95843	CA	3	2	1088	Residentia	Wed May	126640	38.70974	-121.374
8	4108 NOR	SACRAMEI	95820	CA	3	1	963	Residentia	Wed May	127281	38.53753	-121.478

For2B_Sacramentorealestatetrans

2. Go to the data tab → click on data analysis → Select Descriptive statistics→ OK.



3. In the open popup window do the following:
   - **a)** Select the input range.
   - **b)** Select output range radio button and click on a cell for output.
   - **c)** Select the summary statistic, confidence level for mean, kth largest, kth smallest checkboxes and click on ok.



Output:

baths	
Mean	1.776649746
Standard Error	0.028528906
Median	2
Mode	2
Standard Deviation	0.895371422
Sample Variance	0.801689984
Kurtosis	0.361496017
Skewness	-0.236131499
Range	5
Minimum	0
Maximum	5
Sum	1750
Count	985
Largest(1)	5
Smallest(1)	0
Confidence Level(95.0%)	0.05598449

Practical No: 7

Hypothesis

Theory:

Hypothesis:

A Null Hypothesis, proposes that no significant difference exists in a set of given observations. For the purpose of these tests in general.

Null: Given two sample means are equal
Alternate: Given two sample means are not equal

For rejecting a null hypothesis, a test statistic is calculated. This test-statistic is then compared with a critical value and if it is found to be greater than the critical value the hypothesis is rejected. “In the theoretical underpinnings, hypothesis tests are based on the notion of critical regions: the null hypothesis is rejected if the test statistic falls in the critical region. The critical values are the boundaries of the critical region. If the test is one-sided (like a $\chi 2$ test or a one-sided t-test) then there will be just one critical value, but in other cases (like a two-sided t-test) there will be two”. The *t* distribution provides a good way to perform one sample tests on the mean when the population variance is not known provided the population is normal or the sample is sufficiently large so that the Central Limit Theorem applies.

Critical Value:

A critical value is a point (or points) on the scale of the test statistic beyond which we reject the null hypothesis, and, is derived from the level of significance α of the test. Critical value can tell us, what is the probability of two sample means belonging to the same distribution. Higher, the critical value means lower the probability of two samples belonging to same distribution. The general critical value for a two-tailed test is 1.96, which is based on the fact that 95% of the area of a normal distribution is within 1.96 standard deviations of the mean.

Critical values can be used to do hypothesis testing in following way:

1. Calculate test statistic
2. Calculate critical values based on significance level alpha
3. Compare test statistic with critical values.

If the test statistic is lower than the critical value, accept the hypothesis or else reject the hypothesis.

A univariate ***t*-test** is appropriate for testing hypotheses involving some observed mean against some specified value. The ***t*-distribution**, like the standardized normal curve, is a symmetrical, bell-shaped distribution with a mean of 0 and a standard deviation of 1.0. When sample size (*n*) is larger than 30, the *t*-distribution and *Z*-distribution are almost identical.

One Sample t-Test: The One Sample *t* Test determines whether the sample mean is statistically different from a known or hypothesised population mean. The One Sample *t* Test is a parametric test. The degrees of freedom are determined by the number of distinct calculations that are possible given a set of information. In the case of a univariate *t*-test, the degrees of freedom are equal to the sample size (*n*) minus one.

The calculation of t closely resembles the calculation of the Z-value. To calculate t, use the formula

$$t = \frac{\overline{X} - \mu}{S_{\overline{X}}}$$

with $n - 1$ degrees of freedom.

[So, if the current result of 2.01E-08 is actually .0000000201. That is very, very small for a p-value – much less than our cutoff of 0.05. Because our p-value is less than .05, then we would consider our results statistically significant. If a one-sample t-test result is statistically significant, we would say that our mean is significantly different than the chosen value.]
if p value is less than alpha reject the null hypothesis.

One Sample Test:

We make an inference to a population in comparison to some set value. For example, we might be interest in knowing whether the dissolved oxygen levels in a lake meet a state standard of 5 mg/L.

Two Independent Sample Test:

In this test, we collect two independent samples to test whether there is a difference in means between two populations (or if one population mean is greater or less than the other). Comparing GRE scores between men and women is an example of a two independent sample test.

For the unequal variance t test, the null hypothesis is that the two population means are the same but the two population variances may differ. ... The unequal variance t test reports a confidence interval for the difference between two means that is usable even if the standard deviations differ.

Assuming unequal variances, the test statistic is calculated as:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

$$s_1^2 = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2}{n_1 - 1}$$

$$s_2^2 = \frac{\sum_{j=1}^{n_2} (x_j - \bar{x}_2)^2}{n_2 - 1}$$

Paired or Repeated Measure Test:

This test compares paired data, such as data collected before and after a treatment.

Example: A comparison of NOx emissions from randomly selected automobiles before and after an additive is added to the fuel.

One-Sided vs. Two-Sided Comparison of Means Tests:

For a comparison of means test, you may use either a one-sided or two-sided test. A one-sided test (leading to a one-sided p-value) examines whether one mean is greater (or less than) the other mean. If you want to test whether the mean of population A is greater (or less) than the mean of population B, this is a one-sided test. If you want to test whether there is a difference between two means (without any directionality), then you use a two-sided test (and subsequently a two-sided p-value (see below). The null and alternative hypotheses should reflect whether or not you are using a one- or two-sided comparison of means test.

Packages Used:

a. **scipy.stats:** This module contains a large number of probability distributions, summary and frequency statistics, correlation functions and statistical tests, masked statistics, kernel density estimation, quasi-Monte Carlo functionality, and more.

Methods Used:

1. **np.mean( ):** This function is used to compute the arithmetic mean along the specified axis.This function returns the average of the array elements. By default, the average is taken on the flattened array. Else on the specified axis, float 64 is intermediate as well as return values are used for integer inputs.
2. **numpy.genfromtxt():** This method is the source of the data. The data can be string, text file, list of strings etc. If we provide the URL for the data then it is downloaded and use the current working directory.
3. **ttest_1samp():** This is a two-sided test for the null hypothesis that the expected value (mean) of a sample of independent observations a is equal to the given population mean, popmean.

Part A:

Aim: **Perform Testing of Hypothesis using One Sample t-Test.**

Code:

```
from scipy.stats import ttest_1samp
import numpy as np
ages = np.genfromtxt('ages.csv')
print("Executed By 16_Ravi Gupta")
print(ages)
ages_mean = np.mean(ages)
print("Actual Average of our data: ")
print(ages_mean)
muavg=30
print("In null hypothesis we assume the average to be:")
```

```
print(muavg)
tset, pval = ttest_1samp(ages, muavg)
print('p-values == ',pval)
if pval< 0.05: # alpha value is 0.05
   print("reject null hypothesis")
else:
   print("accept null hypothesis")
```

Output:

```
= RESTART: C:/Users/Ravi Gupta/Desktop/Research in Computing Practical/Practical No 7/7A.py
Executed By 16_Ravi Gupta
[20. 30. 25. 13. 16. 17. 34. 35. 38. 42. 43. 45. 48. 49. 50. 51. 54. 55.
 56. 59. 61. 62. 18. 22. 29. 30. 31. 39. 52. 53. 67. 36. 47. 54. 40. 40.
 35. 22. 59. 58. 30. 43. 22. 45. 21. 59. 51. 47. 25. 58. 50. 23. 24. 45.
 37. 59. 28. 28. 48. 42. 54. 36. 36. 24. 26. 24. 50. 48. 34. 44. 56. 55.
 35. 33. 39. 53. 34. 28. 56. 24. 21. 29. 28. 58. 35. 57. 26. 25. 59. 56.
 22. 57. 48. 33. 23. 26. 57. 32. 53. 31. 35. 44. 54. 25. 31. 58. 26. 32.
 26. 50. 41. 49. 26. 33. 34. 24. 43. 42. 51. 36. 38. 38. 40. 38. 56. 39.
 23. 33. 53. 30. 38.]
Actual Average of our data:
39.47328244274809
In null hypothesis we assume the average to be:
30
p-values ==  5.362905195437013e-14
reject null hypothesis
>>> |
```

Null hypothesis: the mean of ages is equal to 30

Alternative hypothesis: Mean of ages is not equal to 30

p-value: 5.362905195437013e-14

Condition: If p-value is less than alpha(0.05) then Reject Null Hypothesis and if P-value is greater than alpha(0.05) the Accept Null Hypothesis.

Result: Reject null hypothesis.

Part B:

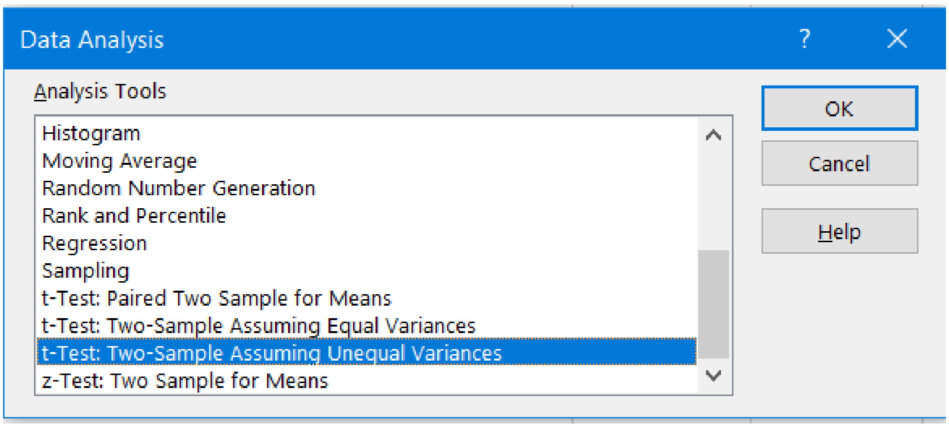
Aim: Perform Testing of Hypothesis using Two Sample t-Test.

Steps:

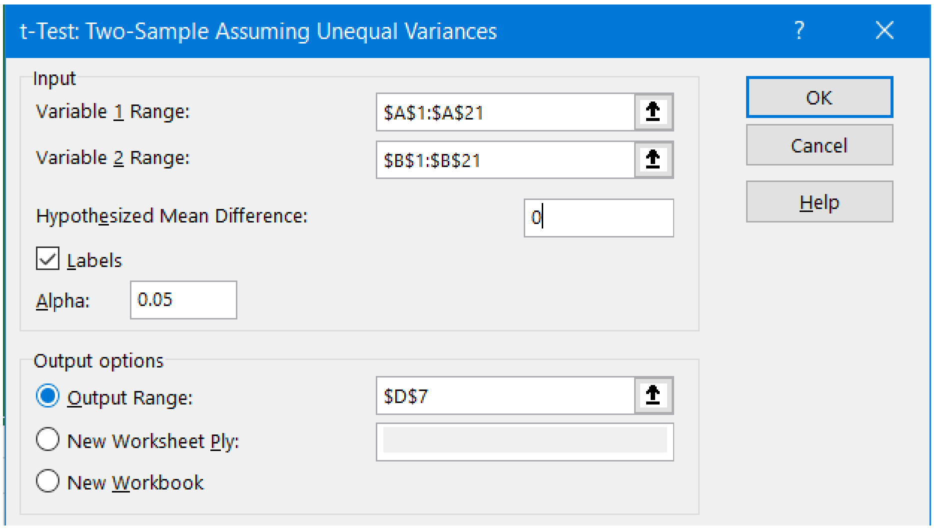
1. In excel sheet enter the required data for analysis.

Experimental	Comparision
35	2
40	27
12	38
15	31
21	1
14	19
46	1
10	34
28	3
48	1
16	2
30	3
32	2
48	1
31	2
22	1
12	3
39	29
19	37
25	2

2. Go to the data tab and click on Data Analysis → select t-Test: Two-Sample Assuming Unequal Variances and click on ok.



3. In the open popup window do the following:
   - **a)** select the input range
   - **b)** Hypothesized Mean Difference = 0 and Alpha= 0.05
   - **c)** select output range



Output:

t-Test: Two-Sample Assuming Unequal Variances		
	Experimental	*Comparision*
Mean	27.15	11.95
Variance	156.45	213.5236842
Observations	20	20
Hypothesized Mean Difference	0	
df	37	
t Stat	3.534053898	
P(T<=t) one-tail	0.000559265	
t Critical one-tail	1.68709362	
P(T<=t) two-tail	0.00111853	
t Critical two-tail	2.026192463	

Null hypothesis: Given two sample's average means are equal.

Alternative hypothesis: Given two sample's average means are not equal. [can be less than or greater than -→ 2 tail]

p-value: 0.000754802

Condition: If test-statistic is found to be greater than the critical value the null hypothesis is rejected, else accepted.

t–Stat : 3.640758601 [Calculated using formula]

t critical two tail: 2.01954097 [Taken from Table (A3)]

Result: Reject null hypothesis.

Part C:

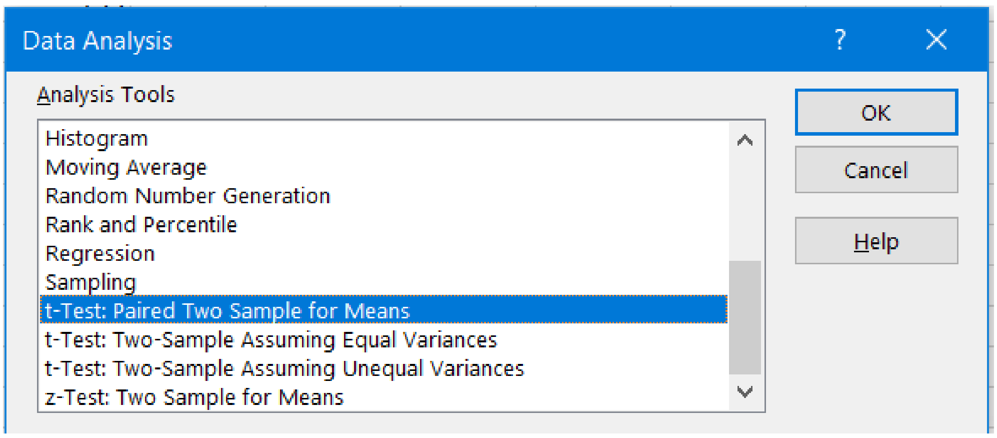
Aim: Perform Testing of Hypothesis using Paired t-Test.

Steps:

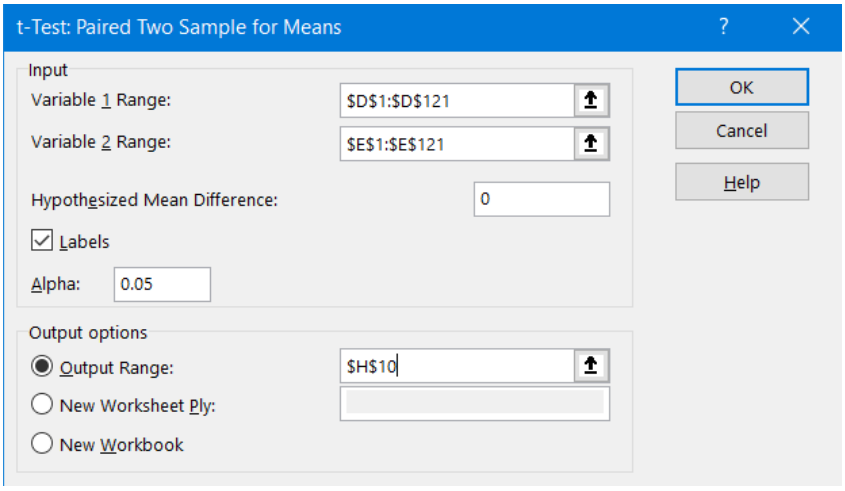
1. Open the data into Excel File.

A	B	C	D	E
patient	gender	agegrp	bp_before	bp_after
1	Male	30-45	143	153
2	Male	30-45	163	170
3	Male	30-45	153	168
4	Male	30-45	153	142
5	Male	30-45	146	141
6	Male	30-45	150	147
7	Male	30-45	148	133
8	Male	30-45	153	141
9	Male	30-45	153	131
10	Male	30-45	158	125
11	Male	30-45	149	164
12	Male	30-45	173	159
13	Male	30-45	165	135
14	Male	30-45	145	159
15	Male	30-45	143	153

2. Go to the data tab and click on Data Analysis → select t-Test: Paired Two Sample for Means → click on ok.



3. In the open popup window do the following:
   - **a)** Select the Input Range
   - **b)** Write Hypothesized Mean Difference = 0 and Alpha = 0.05
   - **c)** Select Output Range



Output:

t-Test: Paired Two Sample for Means		
	bp_before	bp_after
Mean	156.45	151.3583333
Variance	129.7285714	201.004972
Observations	120	120
Pearson Correlation	0.159118103	
Hypothesized Mean Difference	0	
df	119	
t Stat	3.337187051	
P(T<=t) one-tail	0.000564896	
t Critical one-tail	1.657759285	
P(T<=t) two-tail	0.001129791	
t Critical two-tail	1.980099876	

Null hypothesis: No difference.

Alternative hypothesis: There is a significant difference and bp_after is less than bp_before.

p-value: 0.000564896

Condition: if test-statistic is found to be greater than the critical value the null hypothesis is rejected else Accepted.

t Stat: 3.337187051

t Critical one-tail: 1.657759285

Result: Reject Null Hypothesis. Accepting Alternative Hypothesis.

Practical No: 8

Theory:

Distribution:

The *Z*-distribution and the *t*-distribution are very similar, and thus the *Z*-test and *t*-test will provide much the same result in most situations. However, when the population standard deviation (σ) is known, the *Z*-test is most appropriate. When σ is unknown (the situation in most marketing research studies), and the sample size greater than 30, the *Z*-test also can be used. When σ is unknown and the sample size is small, the *t*-test is most appropriate. Since the two distributions are similar with larger sample sizes, the two tests often yield the same conclusion.

Use a Z test if:

- **Your sample size is greater than 30. Otherwise, use a t-test.**
- **Data points should be independent from each other. In other words, one data point isn't related or doesn't affect another data point.**
- **Your data should be normally distributed. However, for large sample sizes (over 30) this doesn't always matter.**
- **Your data should be randomly selected from a population, where each item has an equal chance of being selected.**
- **Sample sizes should be equal if at all possible.**

Methods Used:

pd.read.csv(): It is used to import data from a csv file. This function can take many arguments, but the most important is file which is the name of file to be read. This function reads the data as a dataframe. If the values are separated by a comma use read.csv().

describe(): This method is used for calculating some statistical data like **percentile, mean** and **std** of the numerical values of the Series or DataFrame. It analyzes both numeric and object series and also the DataFrame column sets of mixed data types.

Part A:

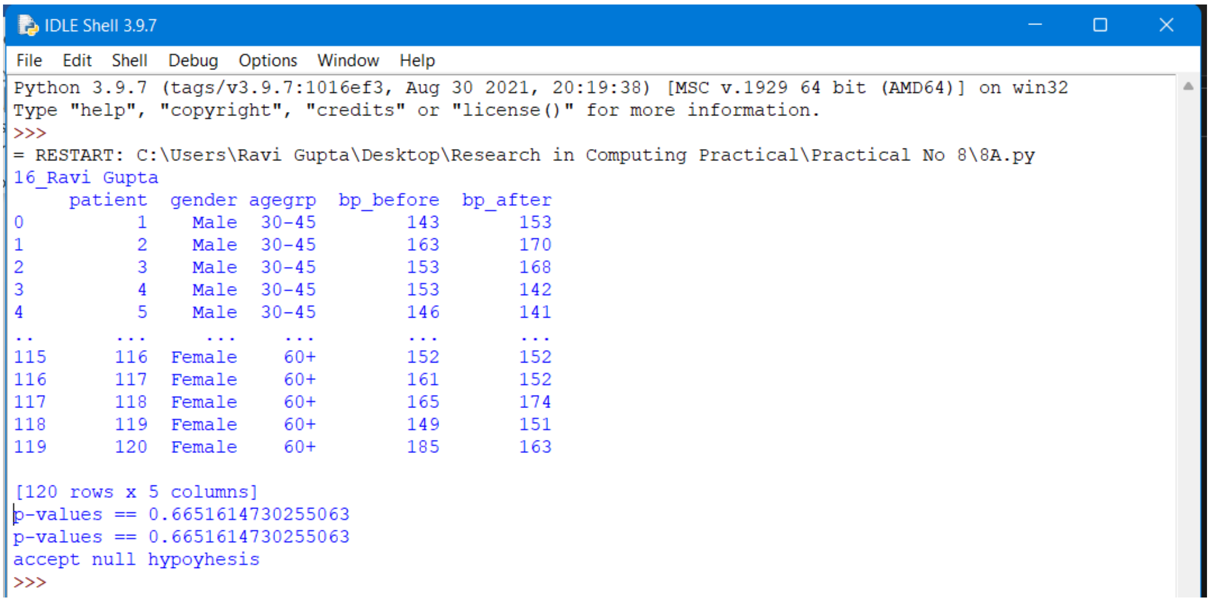
Aim: **Perform testing of hypothesis using one sample Z-test.**

Code:

```
from statsmodels.stats import weightstats as stests
import pandas as pd
from scipy import stats
print("16_Ravi Gupta")
df = pd.read_csv("blood_pressure.csv")
df[['bp_before', 'bp_after']].describe()
print(df)
#sample value with some standard value or standard mean it is one test
#null hypothesis is that sample mean is equal to 156
ztest, pval = stests.ztest(df['bp_before'], x2=None, value=156)
```

```
print('p-values ==', float(pval))
print('p-values ==', pval)
#0.05 is your alpha value confidence level -- 95% --> 100-95= 5%= 5/100= 0.05
if pval<0.05:
   print("reject null hypothesis")
else:
   print("accept null hypoyhesis")
```

Output:



```
IDLE Shell 3.9.7
File  Edit  Shell  Debug  Options  Window  Help
Python 3.9.7 (tags/v3.9.7:1016ef3, Aug 30 2021, 20:19:38) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
= RESTART: C:\Users\Ravi Gupta\Desktop\Research in Computing Practical\Practical No 8\8A.py
16_Ravi Gupta
     patient  gender agegrp  bp_before  bp_after
0          1    Male  30-45        143       153
1          2    Male  30-45        163       170
2          3    Male  30-45        153       168
3          4    Male  30-45        153       142
4          5    Male  30-45        146       141
..       ...     ...    ...        ...       ...
115      116  Female    60+        152       152
116      117  Female    60+        161       152
117      118  Female    60+        165       174
118      119  Female    60+        149       151
119      120  Female    60+        185       163

[120 rows x 5 columns]
p-values == 0.6651614730255063
p-values == 0.6651614730255063
accept null hypoyhesis
>>>
```

Part B:

Aim: Perform testing of hypothesis using two sample Z-test.

Code:

```
from statsmodels.stats import weightstats as stests
from scipy import stats
import pandas as pd
print("16_Ravi Gupta")
df = pd.read_csv("blood_pressure.csv")
df[['bp_before', 'bp_after']].describe()
print(df)
#Null hypothesis is that there is no difference in two observations.
ztest, pval = stests.ztest(df['bp_before'], x2=df['bp_after'], value=0, alternative = 'two-sided')
pval = float(pval)
print('p-values == ', pval)
if pval < 0.05:
    print("Rreject null hypothesis")
    print("Accept Alternate hypothesis")
else:
    print("Accept null hypothesis")
    print("Reject Alternative hypothesis")
```

Output:

```
...
====== RESTART: C:/Users/Ravi Gupta/Desktop/Research in Computing Practical/Practical No 8/8B.py ======
16_Ravi Gupta
     patient  gender agegrp  bp_before  bp_after
0          1    Male  30-45        143       153
1          2    Male  30-45        163       170
2          3    Male  30-45        153       168
3          4    Male  30-45        153       142
4          5    Male  30-45        146       141
..       ...     ...    ...        ...       ...
115      116  Female    60+        152       152
116      117  Female    60+        161       152
117      118  Female    60+        165       174
118      119  Female    60+        149       151
119      120  Female    60+        185       163

[120 rows x 5 columns]
p-values ==  0.002162306611369422
Rreject null hypothesis
Accept Alternate hypothesis
>>>
```

Practical No: 9

Theory:

Chi-Squared:

The chi-squared goodness-of-fit test is an analog of the one-way t-test for categorical variables: it tests whether the distribution of sample categorical data matches an expected distribution. For example, you could use a chi-squared goodness-of-fit test to check whether the race demographics of members at your church or school match that of the entire U.S. population or whether the computer browser preferences of your friends match those of Internet uses as a whole.

When working with categorical data, the values themselves aren't of much use for statistical testing because categories like "male", "female," and "other" have no mathematical meaning. Tests dealing with categorical variables are based on variable counts instead of the actual value of the variables themselves.

If: chi-squared statistic exceeds the critical value[table value], we'd reject the null hypothesis that the two distributions are the same.

Chi-squared goodness-of-fit test:

- The actual χ^2 value is computed using the following formula:

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i}$$

where

χ^2 = chi-square statistic

O_i = observed frequency in the ith cell

E_i = expected frequency in the ith cell

- Like many other probability distributions, the χ^2 distribution is not a single probability curve, but a family of curves. These curves vary slightly with the degrees of freedom. In this case, the degrees of freedom can be computed as

$$df = k - 1$$

where

k = number of cells associated with column or row data.

Chi-squared Test of Independence:

- The expected values are what we would find if there is no relationship between the two variables.
- The expected values for each cell can be computed easily using this formula:

$$E_{ij} = \frac{R_i C_j}{n}$$

where

R_i = total observed frequency count in the ith row

C_j = total observed frequency count in the jth column

n = sample size

- The actual χ^2 value is computed using the following formula:

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E_i}$$

where

χ^2 = chi-square statistic

O_i = observed frequency in the ith cell

E_i = expected frequency in the ith cell

The Number of degrees of freedom: (R-1) (C-1).

Part A:

Aim: Perform testing of hypothesis using chi-squared goodness-of-fit test.

Steps:

1. **Insert data in Excel. (Observed Values)**

	A	B
1	System type	Oi
2	Windows	20
3	Mac	60
4	Linux	20
5		

2. **Total the Oi.**
   Formula: =SUM(B2:B4)

B5 | =SUM(B2:B4)

	A	B	C	D	E
1	System type	Oi			
2	Windows	20			
3	Mac	60			
4	Linux	20			
5		100			

3. **Calculate Ei.**
   Formula: =100/3

C2 =100/3

	A	B	C	D
1	System type	Oi	Ei	
2	Windows	20	33.33333	
3	Mac	60	33.33333	
4	Linux	20	33.33333	
5		100		

4. **Perform steps to calculate Chi-Square value.**
    a) **Oi-Ei**
    b) **(Oi-Ei)^2**
    c) **(Oi-Ei)^2/Ei**
    d) **Take Sum of last step.**
    Formula: =SUM(F2:F4)

F5 =SUM(F2:F4)

	B	C	D	E	F	G
1	Oi	Ei	Oi-Ei	(Oi-Ei)^2	(Oi-Ei)^2/Ei	
2	20	33.33333333	-13.33333333	177.7777778	5.333333333	
3	60	33.33333333	26.66666667	711.1111111	21.33333333	
4	20	33.33333333	-13.33333333	177.7777778	5.333333333	
5	100			Summation:	32	CHI-Squared Calculated Value
6						

5. **Calculate df=k-1.**
    df=3-1=2

DF=K-1	df=3-1	df=2		
k= number of cells associated with either row or column data				

6. **Use function “CHIINV” or “CHISQ.INV.RT” to calculate the table value of chi-square.**
    df=2 and alpha=0.05

Chi-square table value	5.991464547

7. **Conditions and Conclusion:**
    Condition: If chi square statistics is greater than table value reject null hypothesis.
    Conclusion: Calculated value is greater than tabular value so reject null hypothesis.
8. **Check the conclusion using p-value.**
    Formula: CHISQ.TEST(Obs. Value, Exp. Value)

Formula: =CHISQ.TEST(B2:B4,C2:C4)

P-value	1.12535E-07

p-value= 1.12535E-07
Conclusion: P-value is less than alpha so we Reject the Null Hypothesis.
Null Hypothesis: Users equally prefer all three types of systems.
Alternative: Users prefer some systems over others.

Part B:

Aim: Perform testing of hypothesis using chi-squared Test of Independence.

1. **Insert data in Excel sheet.**

1	Observed Values		
2	Location	Profitable	Non-Profitable
3	Stand alone	50	10
4	Shopping Center	15	25

2. **Calculate all the row and column totals.**

D5 =SUM(D3:D4)

	A	B	C	D
1	Observed Values			
2	Location	Profitable	Non-Profitable	Total
3	Stand alone	50	10	60
4	Shopping Center	15	25	40
5	Total	65	35	100

3. **Calculate the expected values.**

Formula: =(Row(Total)*Column(Total))/100

C10 =(D4*C5)/100

6					
7	Expected Values				Ri*Cj/Total
8	Location	Profitable	Non-Profitable	Total	
9	Stand alone	39	21	60	
10	Shopping Center	26	14	40	
11	Total	65	35	100	
12					
13					

4. **Perform steps to calculate Chi-Squared Statistics.**
   a) **Oi-Ei**
   b) **(Oi-Ei)^2**
   c) **(Oi-Ei)^2/Ei**
   d) **Take Sum of last step.**

12					
13	Oi-Ei	(Oi-Ei)^2	(Oi-Ei)^2/Ei		df=(R-1(C-1)
14	11	121	3.102564103		
15	-11	121	5.761904762		
16	-11	121	4.653846154		
17	11	121	8.642857143		
18		Chi-square stat/calculated value	22.16117216		
19					

5. **Calculate df=(R-1)(C-1).**
   R=2
   C=2
   df=(2-1)*(2-1)=1*1=1.

df=(R-1(C-1)	(2-1)*(2-1)	1*1	1

6. **Use function "CHIINV" or "CHISQ.INV.RT" to calculate the table value of chi-square.**
   df=1 and alpha=0.05

Chi-square table value	3.841458821

7. **Conditions and Conclusion:**
   Condition: If chi square statistics is greater than table values reject null hypothesis.
   Conclusion: Calculated value is greater than tabular value so reject null hypothesis.

8. **Check the conclusion using p-value.**
   Formula: CHISQ.TEST(Obs. Value, Exp. Value)
   Formula: =CHISQ.TEST(B3:C4,B9:C10)

P-value	2.50693E-06

p-value= 2.50693E-06
Conclusion: P-value is less than alpha so we Reject the Null Hypothesis.
Null Hypothesis: Location has no effect on the profitability of shop.
Alternative: Location affects the profitability of shop.