Matplotlib in Python:

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.

Matplotlib was originally written by John D. Hunter in 2003. The current stable version is 2.2.0 released in January 2018.

Installation of Matplotlib

```
pip install matplotlib
```

Importing and checking matplotlib version:

Program:

```
import matplotlib
print(matplotlib.__version__)
```

Output:

```
2.0.0
```

PyPlot

Plots (graphics), also known as charts, are a visual representation of data in the form of colored (mostly) graphics. It tells its audience the story about the data relationship through data points, lines, symbols, labels, and numbers so that professionals and anyone with limited knowledge of reading data can get a fair idea of what the data is trying to show. We can use the Matplotlib visualization library in Python to portray the graphs.

Most of the *Matplotlib* utilities lies under the *pyplot* submodule, and are usually imported under the *plt* alias:

```
import matplotlib.pyplot as plt
```

Plot Types:

The six most commonly used Plots come under Matplotlib. These are:

- Bar Plot
- Line Plot
- Scatter Plot
- Pie Plot
- Histogram Plot

➢ Bar plot:

A bar plot or bar chart is a graph that represents the category of data with rectangular bars.A bar chart describes the comparisons between the discrete categories. One of the axis of the plot represents the specific categories being compared, while the other axis represents the measured values corresponding to those categories.

The **matplotlib** API in Python provides the bar() function which can be used in MATLAB style use or as an object-oriented API. The syntax of the bar() function to be used with the axes is as follows:-

Syntax:

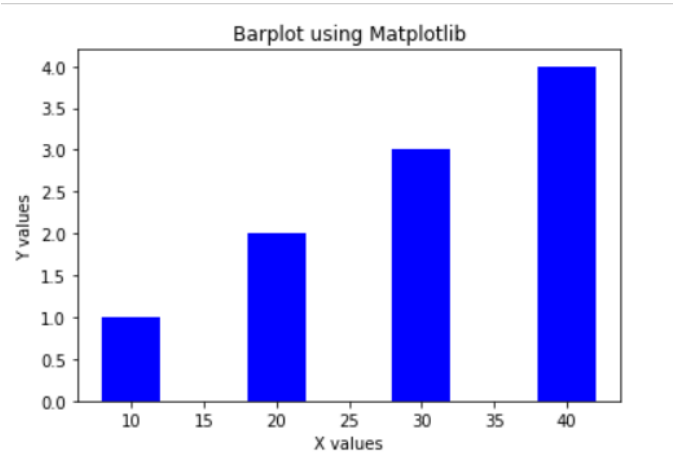
```
plt.bar(x, height, width, bottom, align)
```

The function creates a bar plot bounded with a rectangle depending on the given parameters.

Program:

```
import numpy as np
import matplotlib.pyplot as plt
x=[10,20,30,40]
y=[1,2,3,4]
plt.bar(x,y,color="blue",width= 4)
plt.xlabel("X values")
plt.ylabel("Y values")
plt.title("Barplot using Matplotlib")
plt.show()
```

Output:



- **Line Plot:**

The plot() function is used to draw points (markers) in a diagram. By default, the plot() function draws a line from point to point. The function takes parameters for specifying points in the diagram. Parameter 1 is an array containing the points on the x-axis. Parameter 2 is an array containing the points on the y-axis.

Syntax:

```
plt.plot(x,y, scalex=True, scaley=True)
```

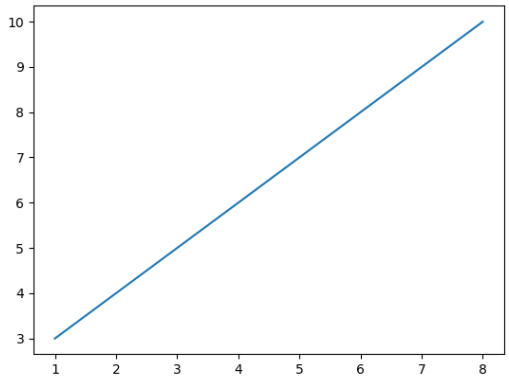
Program:

```
import matplotlib.pyplot as plt
import numpy as np

xpoints = np.array([1, 8])
ypoints = np.array([3, 10])

plt.plot(xpoints, ypoints)
plt.show()
```

Output:



- **Scatter Plot:**

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.

Syntax:

```
plt.scatter(x_axis_data, y_axis_data,)
```

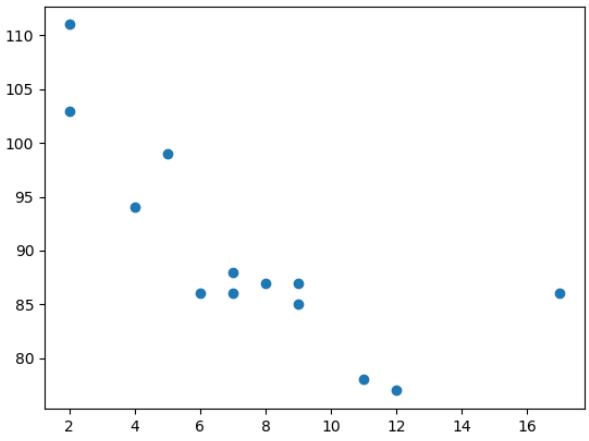
Program:

```
import matplotlib.pyplot as plt
import numpy as np

x = np.array([5,7,8,7,2,17,2,9,4,11,12,9,6])
y = np.array([99,86,87,88,111,86,103,87,94,78,77,85,86])

plt.scatter(x, y)
plt.show()
```

Output:



➢ **Pie Plot:**

A Pie Chart is a circular statistical plot that can display only one series of data. The area of the chart is the total percentage of the given data. The area of slices of the pie represents the percentage of the parts of the data. The slices of pie are called wedges. The area of the wedge is determined by the length of the arc of the wedge. The area of a wedge represents the relative percentage of that part with respect to whole data. Pie charts are commonly used in business presentations like sales, operations, survey results, resources, etc as they provide a quick summary.

Syntax:

```
plt.pie(data)
```

Program:

```
import matplotlib.pyplot as plt
import numpy as np

y = np.array([35, 25, 25, 15])

plt.pie(y)
plt.show()
```

Output:

➢ **Histogram Plot:**

A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable. It is a kind of bar graph.

To construct a histogram, follow these steps −

- **Bin** the range of values.
- Divide the entire range of values into a series of intervals.
- Count how many values fall into each interval.

The bins are usually specified as consecutive, non-overlapping intervals of a variable.

The **matplotlib.pyplot.hist()** function plots a histogram. It computes and draws the histogram of x.

Program:

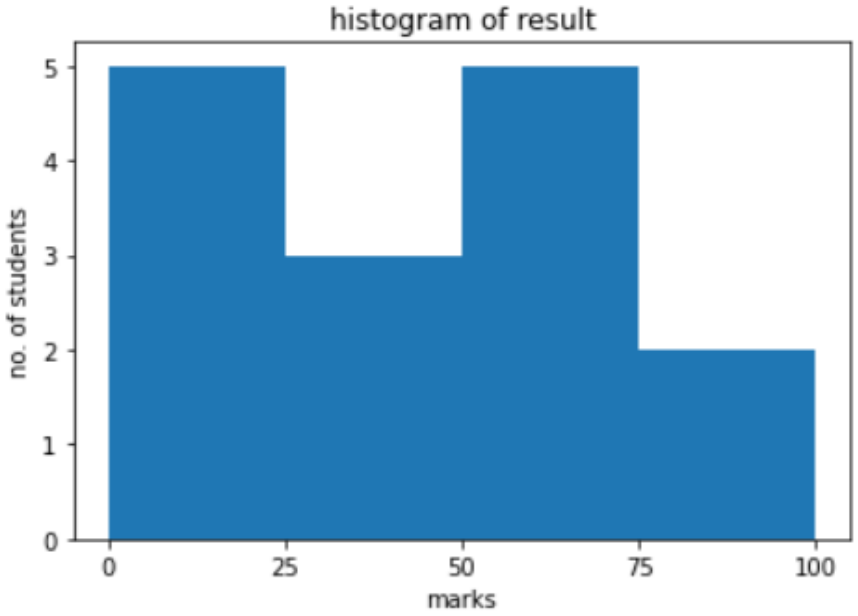
```
from matplotlib import pyplot as plt
import numpy as np
fig,ax = plt.subplots(1,1)
a = np.array([22,87,5,43,56,73,55,54,11,20,51,5,79,31,27])
```

```
ax.hist(a, bins = [0,25,50,75,100])

ax.set_title("histogram of result")
ax.set_xticks([0,25,50,75,100])
ax.set_xlabel('marks')
ax.set_ylabel('no. of students')

plt.show()
```

Output:



Polygons:

To plot shapely polygons and objects using matplotlib, the steps are as follows −

- Create a polygon object using (x, y) data points.
- Get x and y, the exterior data, and the array using polygon.exterior.xy.
- Plot x and y data points using plot() method with red color.

Syntax:

```
patches.Polygon(xy, *, closed=True, **kwargs)
```

Program:

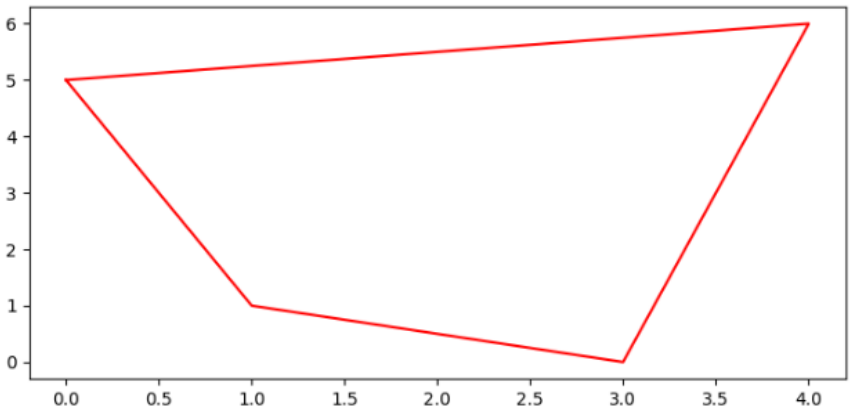
```
from shapely.geometry import Polygon
import matplotlib.pyplot as plt

plt.rcParams["figure.figsize"] = [7.00, 3.50]
plt.rcParams["figure.autolayout"] = True

polygon1 = Polygon([ (0, 5), (1, 1), (3, 0), (4, 6) ])
x, y = polygon1.exterior.xy

plt.plot(x, y, c="red")
plt.show()
```

Output:



➢ Box plots / Quartiles:

A Box Plot is also known as Whisker plot is created to display the summary of the set of data values having properties like minimum, first quartile, median, third ***quartile*** and maximum.

Creating Box Plot:

The *matplotlib.pyplot* module of *matplotlib* library provides *boxplot( )* function with the help of which we can create box plots.

Syntax:

```
plt.boxplot(data, patch_artist=None, widths=None)
```

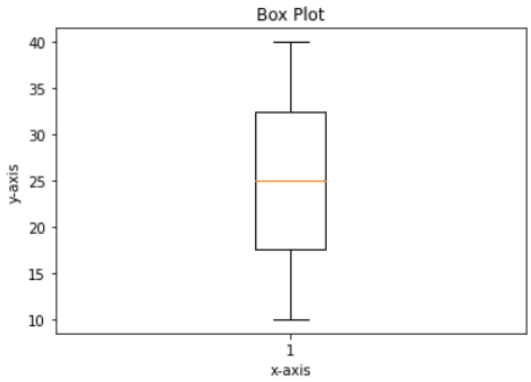
Program:

```
import matplotlib.pyplot as plt
import numpy as np

data=[10,20,30,40]
plt.boxplot(data)

plt.show()
```

Output:



Heat Maps:

A 2-D Heatmap is a data visualization tool that helps to represent the magnitude of the phenomenon in form of colors. In python, we can plot 2-D Heatmaps using Matplotlib package. There are different methods to plot 2-D Heatmaps.

Method 1: Using *matplotlib.pyplot.imshow()* Function

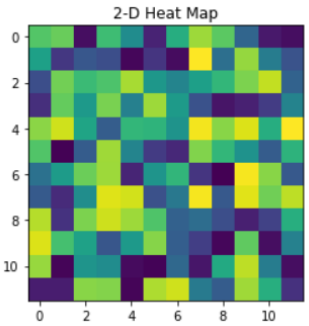
Program:

```
import numpy as np
import matplotlib.pyplot as plt

data = np.random.random(( 12 , 12 ))
plt.imshow( data)

plt.title( "2-D Heat Map" )
plt.show()
```

Output:



Method 2: Using Seaborn Library

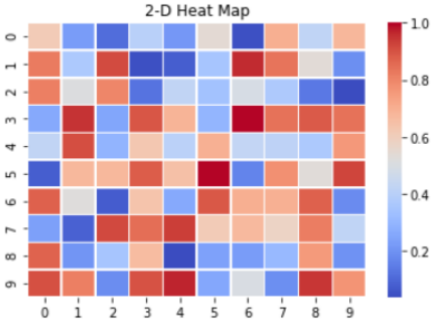
Program:

```
import numpy as np
import seaborn as sns
import matplotlib.pylab as plt

data_set = np.random.rand( 10 , 10 )
ax = sns.heatmap( data_set , linewidth = 0.5 , cmap = 'coolwarm' )
plt.title( "2-D Heat Map" )

plt.show()
```

Output:



Simple Hypothesis Testing:

Statistics is the science of analyzing huge amounts of data. In the real world, it is nearly impossible to deduce statistics about the entire population. And this huge amount of data needs interpretation to draw meaningful conclusions. Hence, we take some random samples from the population, derive some statistical measures (e.g. mean, standard deviation, variance), and draw conclusions about relationships from the data collected.

Data can be interpreted by assuming a specific outcome and use statistical methods to confirm or reject the assumption. This assumption is called a hypothesis and the statistical test used for this purpose is called hypothesis testing.

In statistics, a hypothesis is a statement about a population that we want to verify based on information contained in the sample data.

Hypothesis testing quantifies an observation or outcome of an experiment under a given assumption. The result of the test enables us to interpret whether the assumption holds true or false. In other words, it signifies if the hypothesis can be confirmed or rejected for the observation made.

An observation or outcome of an experiment is known as a test statistic, which is a statistic measure or a standardized value that is calculated from sample data of the underlying population.

Terminology

To understand hypothesis testing, there's some terminology that you have to understand:

- **Null Hypothesis:** the hypothesis that sample observations result purely from chance. The null hypothesis tends to state that there's no change.
- **Alternative Hypothesis:** the hypothesis that sample observations are influenced by some non-random cause.
- **P-value:** the probability of obtaining the observed results of a test, assuming that the null hypothesis is correct; a smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.
- **Alpha:** the significance level; the probability of rejecting the null hypothesis when it is true — also known as Type 1 error.

I'll use the coin example again so that you can understand these terms better:

- The **null hypothesis** in our example is that the coin is a fair coin and that the observations are purely from chance.
- The **alternative hypothesis** would then be that the coin is **not** fair, and thus, the observations did not happen by chance.
- The **p-value** in the scenario of flipping tails 2 times in a row is 25% and 6 times in a row is 1.56%.
- The **alpha** or **level** of significance would be 5%.

Reject or Do not Reject?

The main rule in determining whether you reject the null is simple, PGADRN

If the **P**-value is Greater than the Alpha, **D**o not **R**eject the **N**ull.

In the case of flipping tails 2 times in a row, we would not reject the null since 25% > 5%. However, in the case of flipping tails 6 times in a row, we would reject the null since 1.56% < 5%.

What is the point of Significance Testing?

So now that you understand the use of hypothesis testing through the coin toss example, know the relevant terminology, and know the main rule to determine whether to reject the null or not, let's dive into significance testing.

What is the point of significance testing? It's used to determine how likely or unlikely a hypothesis is for a given sample of data. The last part of the statement, 'for a given sample of data' is key because more often than not, you won't be able to get an infinite amount of data or data that represents the entire population.

Steps for Hypothesis Testing

Here are the steps to performing a hypothesis test:

1. State your null and alternative hypotheses. To reiterate, the null hypothesis typically states that everything is as normally was — that nothing has changed.
2. Set your significance level, the alpha. This is typically set at 5% but can be set at other levels depending on the situation and how severe it is to committing a type 1 and/or 2 error.
3. Collect sample data and calculate sample statistics.
4. Calculate the p-value given sample statistics. Once you get the sample statistics, you can determine the p-value through different methods. The most common methods are the T-score and Z-score for normal distributions.
5. Reject or do not reject the null hypothesis.

T Test (Students T test)

A t-test is a type of inferential statistic which is used to determine if there is a significant difference between the means of two groups which may be related in certain features. It is mostly used when the data sets, like the set of data recorded as outcome from flipping a coin a 100 times, would follow a normal distribution and may have unknown variances. T test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population.

T-test has 2 types:

1. One sampled t-test
2. two-sampled t-test.

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.

Example:- you have 10 ages and you are checking whether avg age is 30 or not

```
from scipy.stats import ttest_1samp
import numpy as np
ages = np.genfromtxt(“ages.csv”)
print(ages)
ages_mean = np.mean(ages)
print(ages_mean)
tset, pval = ttest_1samp(ages, 30)
print(“p-values”,pval)
if pval < 0.05: # alpha value is 0.05 or 5%
      print(" we are rejecting null hypothesis")
else:
      print("we are accepting null hypothesis")
```

Two sampled T-test:- The Independent Samples t Test or 2-sample t-test compares the means of two independent groups in order to determine whether there is statistical evidence that the associated population means are significantly different. The Independent Samples t Test is a parametric test. This test is also known as: Independent t Test.

Example: is there any association between week 1 and week 2

```
from scipy.stats import ttest_ind
import numpy as np
week1 = np.genfromtxt("week1.csv", delimiter=",")
week2 = np.genfromtxt("week2.csv", delimiter=",")
print(week1)
print("week2 data :-\n")
print(week2)
week1_mean = np.mean(week1)
week2_mean = np.mean(week2)
print("week1 mean value:",week1_mean)
print("week2 mean value:",week2_mean)
week1_std = np.std(week1)
week2_std = np.std(week2)
print("week1 std value:",week1_std)
print("week2 std value:",week2_std)
ttest,pval = ttest_ind(week1,week2)
print("p-value",pval)
if pval <0.05:
      print("we reject null hypothesis")
else:
      print("we accept null hypothesis")
```

Paired sampled t-test:- The paired sample t-test is also called dependent sample t- test. It's a uni variate test that tests for a significant difference between 2 related variables. An example of this is if you where to collect the blood pressure for an individual before and after some treatment, condition, or time point.

H0 :- means difference between two sample is 0

H1:- mean difference between two sample is not 0

Example:

```
import pandas as pd
from scipy import stats
df = pd.read_csv("blood_pressure.csv")
df[['bp_before','bp_after']].describe()
ttest,pval = stats.ttest_rel(df['bp_before'], df['bp_after'])
print(pval)
if pval<0.05:
      print("reject null hypothesis")
else:
      print("accept null hypothesis")
```

U Test:

The Mann-Whitney U Test, also known as the Wilcoxon Rank Sum Test, is a non-parametric statistical test used to compare two samples or groups.

The Mann-Whitney U Test assesses whether two sampled groups are likely to derive from the same population, and essentially asks; do these two populations have the same shape with regards to their data? In other words, we want evidence as to whether the groups are drawn from populations with different levels of a variable of interest. It follows that the hypotheses in a Mann-Whitney U Test are:

The null hypothesis (H0) is that the two populations are equal.

The alternative hypothesis (H1) is that the two populations are not equal.

Some researchers interpret this as comparing the medians between the two populations (in contrast, parametric tests compare the means between two independent groups). In certain situations, where the data are similarly shaped (see assumptions), this is valid – but it should be noted that the medians are not actually involved in calculation of the Mann-Whitney U test statistic. Two groups could have the same median and be significantly different according to the Mann-Whitney U test.

Mann-Whitney U Test Assumptions

Some key assumptions for Mann-Whitney U Test are detailed below:

- The variable being compared between the two groups must be **continuous** (able to take any number in a range – for example age, weight, height or heart rate). This is because the test is based on ranking the observations in each group.
- The data are assumed to take a **non-Normal**, or skewed, distribution. If your data are normally distributed, the unpaired Student's t-test should be used to compare the two groups instead.
- While the data in both groups are not assumed to be Normal, the data are assumed to be **similar in shape** across the two groups.

- The data should be two randomly selected **independent** samples, meaning the groups have no relationship to each other. If samples are paired (for example, two measurements from the same group of participants), then a paired samples t-test should be used instead.
- Sufficient **sample size** is needed for a valid test, usually more than 5 observations in each group.

Example:

```
test_team=[6.2, 7.1, 1.5, 2,3 , 2, 1.5, 6.1, 2.4, 2.3, 12.4, 1.8, 5.3, 3.1,
9.4, 2.3, 4.1]
developer_team=[2.3, 2.1, 1.4, 2.0, 8.7, 2.2, 3.1, 4.2, 3.6, 2.5, 3.1, 6.2,
12.1, 3.9, 2.2, 1.2 ,3.4]
ttest,pvalue = stats.mannwhitneyu(test_team,developer_team, alternative="two-
sided")
print("p-value:%.4f" % pvalue)
if pvalue <0.05:
    print("Reject null hypothesis")
else:
    print("Fail to recejt null hypothesis")
```

Correlation and Covariance:

Correlation:

Correlation, statistical technique which determines how one variables moves/changes in relation with the other variable. It gives us the idea about the degree of the relationship of the two variables. It's a bi-variate analysis measure which describes the association between different variables. In most of the business it's useful to express one subject in terms of its relationship with others.

For example: Sales might increase if lot of money is spent on product marketing.

Why it is useful?

1. If two variables are closely correlated, then we can predict one variable from the other.
2. Correlation plays a vital role in locating the important variables on which other variables depend.
3. It's used as the foundation for various modeling techniques.
4. Proper correlation analysis leads to better understanding of data.
5. Correlation contribute towards the understanding of causal relationship (if any).

Relationship of Correlation and Covariance

Before diving more into correlation, let's get the understanding of covariance.

Covariance: The prefix 'Co' defines some kind of joint action and variance refers to the change or variation. So it says, two variables are related based on how these variables change in relation with each other.

But wait, is covariance same as correlation?

As covariance says something on same lines as correlation, correlation takes a step further than covariance and also tells us about the strength of the relationship.

Both can be positive or negative. Covariance is positive if one increases other also increases and negative if one increases other decreases.

Covariance is calculated as

$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

X_i= Observation point of variable X
$\bar{x}$= Mean of all observations(X)
Y_i= Observation point of variable Y
$\bar{y}$ = Mean of all observations(Y)
n= Number of observations

Decoding the covariance formula: Covariance between two variables x and y is the sum of the products of the differences of each item and their respective means divided by the number of items in the dataset minus one..

Getting better understanding with a simple example of sample data:

Following data shows the number of customers with their corresponding temperature.

Temperature(X)	No. Of Customers(Y)
97	14
86	11
89	9
84	9
94	15
74	7

First find means of both the variables subtract each of the item with its respective mean and multiply it together as follows

Mean of X, $\bar{x}$ = (97+86+89+84+94+74)/6 = 524/6= 87.333
Mean of Y, $\bar{Y}$ = (14+11+9+9+15+7)/6 = 65/6= 10.833

Temperature (x- $\bar{X}$)	No of Customers (y- $\bar{Y}$)	Product (x- $\bar{X}$)(y-$\bar{Y}$)
97-87.33 = 9.67	14-10.83 = 3.17	30.65
86-87.33 = -1.33	11-10.83 = 0.17	-0.22
89-87.33 = 1.67	9-10.83 = -1.83	-3.05
84-87.33 = -3.33	9-10.83 = -1.83	6.09
94-87.33 = 6.67	15-10.83 = 4.17	27.81
74-87.33 = -13.33	7-10.83 = -3.83	51.05

COV(x, y) = 112.33/(6–1) = 112.33/5 = 22.46

The covariance between the temperature and customers is 22.46. Since the covariance is positive, temperature and number of customers have a positive relationship. As temperature rises, so does the number of customers.

But here there is no information about how strong the relationship is, and that's where correlation comes into the picture.

Correlation coefficient is the term used to refer the result of any correlation measurement methods.

So here, the sample Correlation coefficient is calculated as

$$Correlation = \frac{Cov\ (x,y)}{\sigma x * \sigma y}$$

COV(x, y) = covariance of the variables x and y
σx = sample standard deviation of variable x
σy = sample standard deviation of variable y

Temperature (x)	Customer (Y)	Temperature (x- $\bar{x}$)²	Customer(y- $\bar{Y}$)²
97	14	93.50	10.04
86	11	1.76	0.02
89	9	2.78	3.34
84	9	11.08	3.34
94	15	44.48	17.38
74	7	177.68	14.66

COV(x, y) = 22.46
σx = 331.28/5=66.25= 8.13
σy = 48.78/5=9.75=3.1
correlation = 22.46/(8.13x 3.1)= 22.46/25.20 =0.8

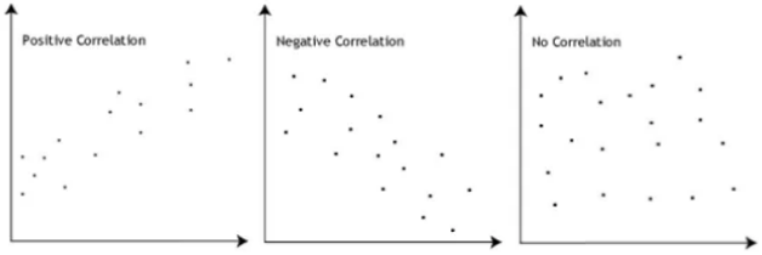
0.8 shows that strength of the correlation between temperature and number of customers is very strong.

Sample correlation coefficient can be used to estimate the population correlation coefficient.

Different methods exist to calculate correlation coefficient between two subjects. Some of the methods are:

1. Pearson Correlation Coefficient

It captures the strength and direction of the linear association between two continuous variables. It tries to draw the line of best fit through the data points of two variables. Pearson correlation coefficient indicates how far these data points are away from the line of best fit. The relationship is linear only when the change in one variable is proportional to the change in another variable.



Pearson Correlation Coefficient calculated as

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

r = Pearson Correlation Coefficient
n = number of observations
∑xy = sum of the products of x and y values
∑x = sum of x values
∑y = sum of y values
∑x2= sum of squared x values
∑y2= sum of squared y values

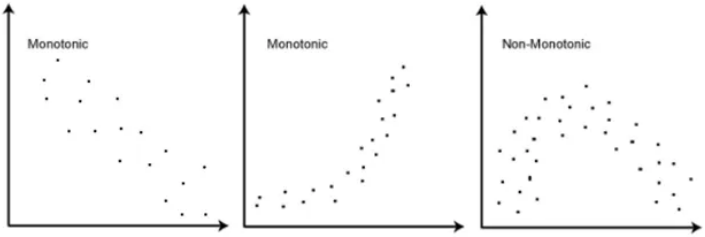
Example:

```
import os
import sys
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
X = np.random.rand(50)       #random number
Y = 2 * X + np.random.normal(0, 0.1, 50)
#covariance
cov_matrix = np.cov(X, Y)      #calculate covariance between x & y
print('The Covariance of X and Y: %.2f'%cov_matrix[0, 1])
# output:  The covariance of X and Y: 0.21
#Correlation
cor_matrix = np.corrcoef(X, Y) #calculate correlation between x & y
print(Correlation of X and Y: %.2f'%cor_matrix[0, 1])
# output: Correlation of X and Y: 0.99
```

Spearman's Correlation Coefficient

It tries to determine the strength and the direction of the monotonic relationship which exists between two ordinal or continuous variables. In a monotonic relationship two variables tend to change together but not with the constant rate. It's calculated on the ranked values of the variables rather than on the raw data.

Monotonic and non- monotonic relationships are shown below:



$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

ρ= Spearman rank correlation coefficient

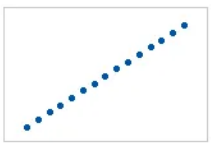
di= the difference between the ranks of corresponding variables

n= number of observations

Comparison: Pearson and Spearman correlation coefficient

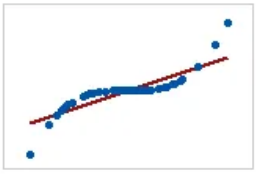
Pearson and Spearman correlation coefficient can take values from -1 to 1.

(i) If one variable increases with the other variable at the consistent rate then Pearson coefficient would be 1, which results in a perfect line. In this case Spearman coefficient would also be 1.



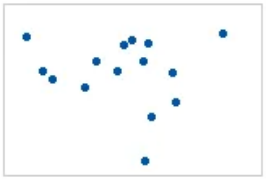
Pearson = +1, Spearman = +1

(ii) If one variable increases with the other variable but not with the consistent rate then Pearson coefficient would be positive but less than 1. In this case Spearman coefficient would be still 1.



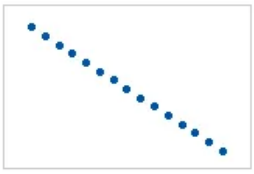
Pearson = +0.851, Spearman = +1

(iii) If the relationship is random then both the coefficients would be near 0.



Pearson = −0.093, Spearman = −0.093

(iv) If the relationship between the variables is a perfect line but with a decreasing relationship then both the coefficients would be -1.



Pearson = −1, Spearman = −1

(v) If the relationship between two variables is such that one variable decreases when the other increases but not with the consistent rate, then Pearson coefficient would be negative but greater than -1. Spearman coefficient would be -1 in this case.

When to use what?

Pearson correlation describes linear relationships and spearman correlation describes monotonic relationships. A scatter plot would be helpful to visualize the data and understand which correlation coefficient should be used. Other way of doing is to apply both the methods and check which is performing well. For instance if results show spearman correlation coefficient is greater than Pearson coefficient, it means our data has monotonic relationships and not linear.

Test for Association:

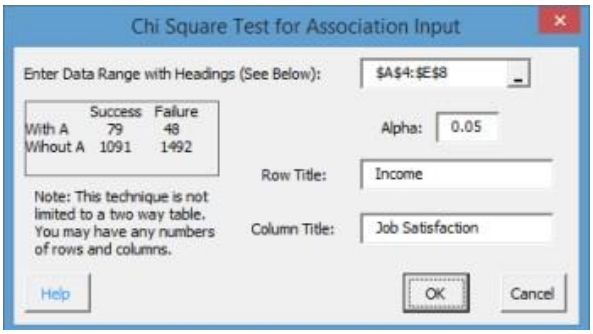
The Chi-Square Test for Association is used to determine if there is any association between two variables. It is really a hypothesis test of independence. The null hypothesis is that the two variables are not associated, i.e., independent. The alternate hypothesis is that the two variables are associated. The example below shows how to do this test using the SPC for Excel software.

A survey was done to determine if job satisfaction was related to income. A total of 901 people participated in the survey. The data are shown below. We will use the Chi-Square Test for Association to determine if the two variables are associated.

1. Enter the data into an Excel worksheet as shown below.

Income	Very Dissatisfied	Little Dissatisfied	Moderately Satisfied	Very Satisfied
<6000	20	24	80	82
6000 - $15000	22	38	104	125
15000 - 25000	13	28	81	113
>25000	7	18	54	92

2. Select all the data in the table above including the headings.
3. Select "Misc. Tools" from the "Statistical Tools" panel on the SPC for Excel ribbon.
4. Select the "Chi Square Test for Association" option and then OK.



- Enter Data Range with Labels: enter the range containing the data and the labels; default is the range selected on the worksheet.
- Alpha: this is the confidence level; 1-alpha is the confidence interval. Default is 0.05 for 95% confidence.
- Row Title: Enter the title of the rows; default is value in first row of selected data.
- Column Title: enter the title of the columns; default is value in row above second column.
- Select OK to generate the results.
- Select Cancel to end the program.

Chi Square Test for Association Output

The output from the Chi-Square Test for Association is shown below. An explanation of the output follows.

Chi Square Results for Association

Income		Job Satisfaction				Row Total
		Very Dissatisfied	Little Dissatisfied	Moderately Satisfied	Very Satisfied	
<6000	Observed	20	24	80	82	206
	Expected	14.18	24.69	72.93	94.20	
	Contribution to χ^2	2.393	0.019	0.684	1.579	
6000 - $15000	Observed	22	38	104	125	289
	Expected	19.89	34.64	102.32	132.15	
	Contribution to χ^2	0.225	0.326	0.028	0.387	
15000 - 25000	Observed	13	28	81	113	235
	Expected	16.17	28.17	83.20	107.46	
	Contribution to χ^2	0.622	0.001	0.058	0.286	
>25000	Observed	7	18	54	92	171
	Expected	11.77	20.50	60.54	78.19	
	Contribution to χ^2	1.931	0.304	0.707	2.438	
Column Total		62	108	319	412	901

Alpha	0.05
χ^2	11.989
Degrees of Freedom	9
$\chi^2_{(0.05,\ 9)}$	16.919
p Value	0.2140

Residuals

Income	Job Satisfaction			
	Very Dissatisfied	Little Dissatisfied	Moderately Satisfied	Very Satisfied
<6000	5.82	-0.69	7.07	-12.20
6000 - $15000	2.11	3.36	1.68	-7.15
15000 - 25000	-3.17	-0.17	-2.20	5.54
>25000	-4.77	-2.50	-6.54	13.81

The null hypothesis is not rejected.
There is no evidence that Income and Job Satisfaction are associated

The top part of the output contains the data with the observed and expected values as well as the contribution of each to $\chi 2$. The row and column totals are also given.

The middle portion of the output contains the following:

- Alpha (entered)
- The calculated $\chi 2$
- The degrees of freedom
- The critical $\chi 2$ value based on alpha and the degrees of freedom
- The calculated p value (will be in red if $\leq$ alpha)

The bottom portion of the output contains the residuals. The residuals are the difference between the observed and the expected values. The conclusion is then given based on the values of alpha and the p value. The null hypothesis (that the variables are not associated) is rejected if the p value < alpha.