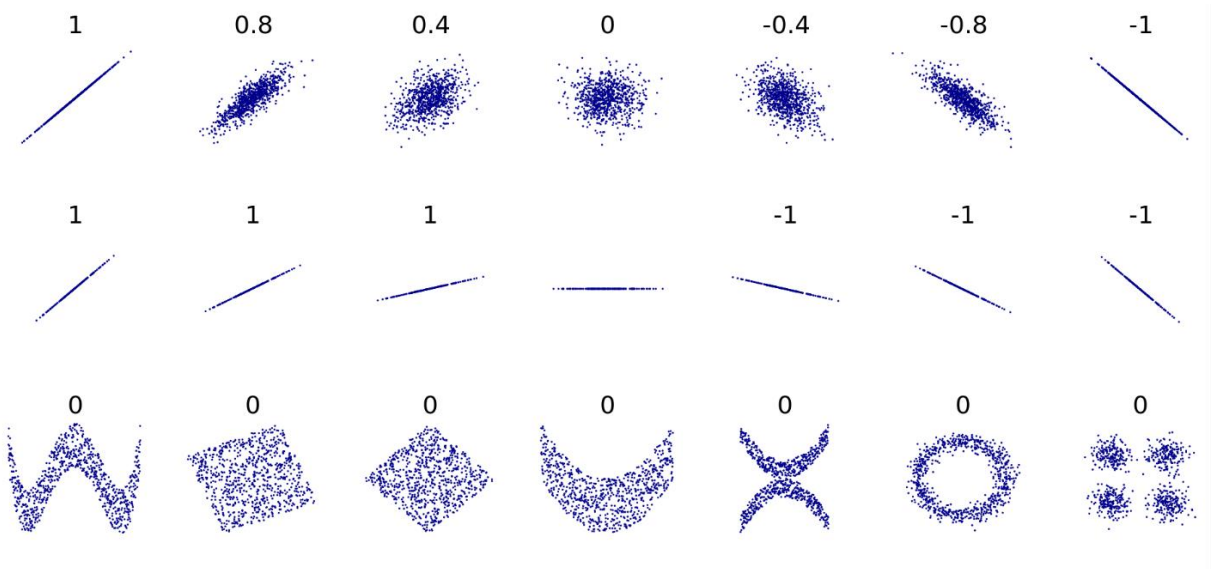
Bivariate Analysis

Correlation and Scatter Plot



Correlation

The correlation analysis examines or measures the relationship between two variables/ bivariable data. The next important question is **How strong is the Correlation?**

Correlation coefficients

The strength of the **correlation** is determined by **correlation coefficient**, which varies between -1 to +1. Correlation coefficient is used to measure the strength and direction of a linear relationship between two variables. Here is

a table that shows the strength of the correlation based on its magnitude (absolute value):

Magnitude of Correlation	Strength of Correlation
0.00 - 0.19	Very weak
0.20 - 0.39	Weak
0.40 - 0.59	Moderate
0.60 - 0.79	Strong
0.80 - 1.00	Very strong

Note: The sign of the correlation (positive or negative) indicates the direction of the relationship, while the magnitude (absolute value) indicates the strength of the relationship. A positive correlation indicates that both variables tend to increase or decrease together, while a negative correlation indicates that as one variable increases, the other tends to decrease.

Direction of the correlation

Positive Correlation: When the value of one variable increases, the value of the other variable also tends to increase. For example, there is a positive correlation between the amount of exercise a person gets and their overall health. As a person exercises more, their health tends to improve.

Negative Correlation: When the value of one variable increases, the value of the other variable tends to decrease. For example, there is a negative correlation between the number of hours a person spends watching TV and their academic performance. As a person spends more time watching TV, their academic performance tends to decrease.

No Correlation: When there is no relationship between two variables. For example, there is no correlation between a person's shoe size and their IQ. The size of a person's feet has no impact on their intelligence.

How is correlation coefficient calculated?

If you have two variables, x and y, the correlation coefficient can be determined mathematically by dividing the covariance of the two variables by the product of their standard deviations.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

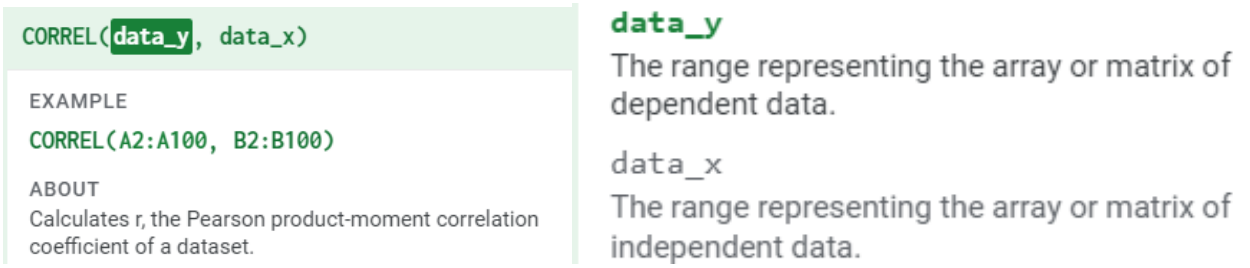
$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Correlation coefficient in Spreadsheets

To calculate the correlation coefficient in Google spreadsheets, you can use the **CORREL** function.



Formula: CORREL(array1, array2)

where array1 and array2 are the two arrays of data for which you want to calculate the correlation coefficient.

Steps to calculate the correlation coefficient in Google Spreadsheets:

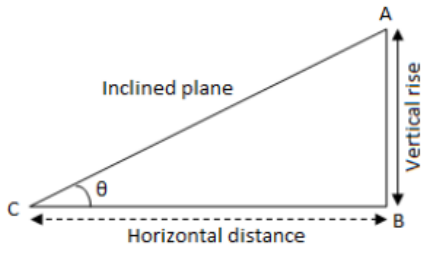
1. Open a new or existing Google Sheet.
2. Enter your data in two columns.
3. Click on an empty cell where you want to display the correlation coefficient.

```
=CORREL(A2:A2241,B2:B2241)
```

4. Type the following formula: **=CORREL(A2:A2241, B2:B2241)** where *A2:A2241* and *B2:B2241* represent the ranges of the two columns containing your data.
5. Press Enter to calculate the correlation coefficient. The result should appear in the cell.

Scatter Plot

A scatter plot visually represents the correlation between two numerical variables. In a scatter plot, each data point is plotted as a dot. It's important to note that the slope of a line is an essential concept when interpreting a scatter plot. The slope of a line refers to its steepness and is calculated by dividing the vertical change between two points by the horizontal change between the same points. Additionally, the inclination of a straight line can be represented by the angle θ, and its slope is equivalent to the tangent (tan) of that angle.



i.e., slope = $\frac{vertical\,rise}{horizontal\,distance} = \frac{AB}{BC}$ = tan θ

Positive Slope: If the line makes an acute angle in the anticlockwise direction with x-axis. Inclination θ = 45°, Therefore, slope = tan 45° = 1

Negative Slope: If the line makes an obtuse angle in the anticlockwise direction with x-axis. Inclination θ = 135° or - 45° Therefore, slope = tan (-45°) = - tan 45° = -1

Since tan θ is not defined when θ = 90°, therefore, the slope of a vertical line is not defined. i.e., slope of y-axis is m = tan 90° = ∞ i.e., not defined. Slope of x-axis is m = tan 0° = 0.

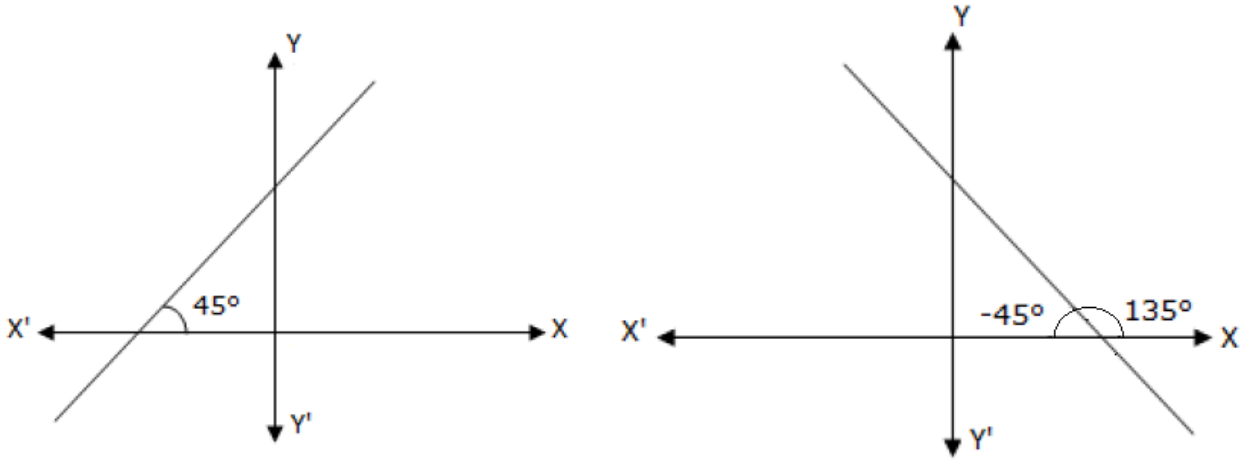


Figure (Left to Right): Positive Slope and Negative Slope

Putting it all together

Here the [Marketing dataset] is can be used to perform scatter plot and understand the type of correlation.

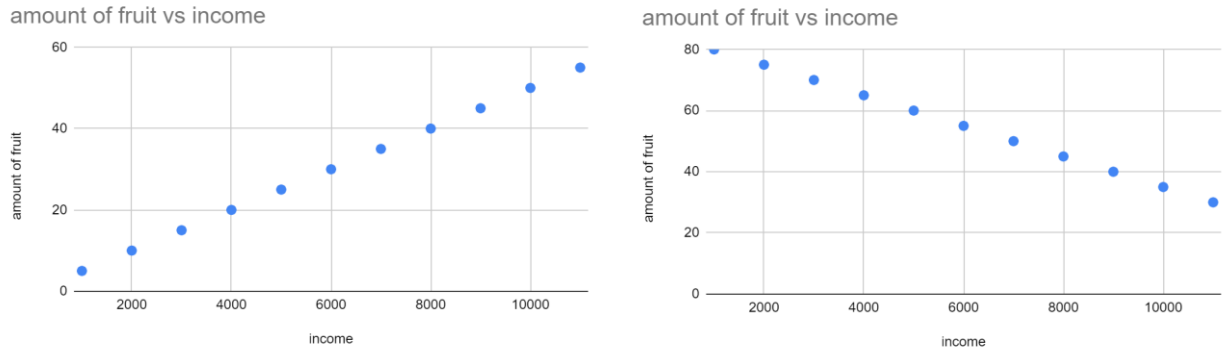
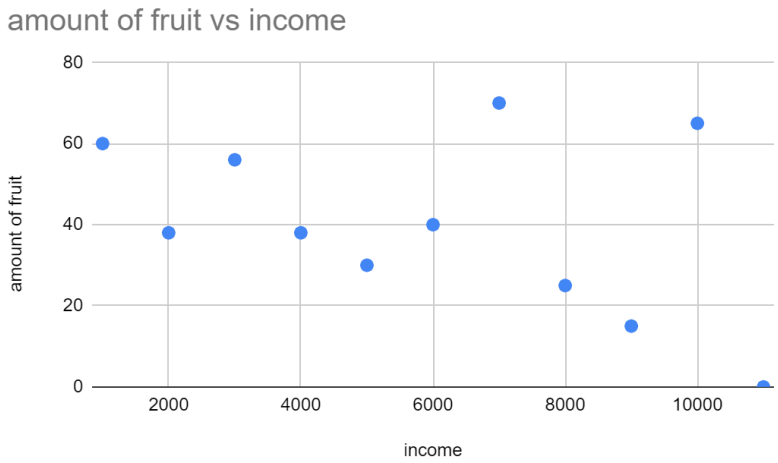


Figure: Scatter Plot & Correlation Analysis with the Marketing Dataset

The first figure on the left shows a **positive correlation** between **Income** (independent variable) and the **amount of fruit** (dependent variable). This is evident from the upward slope of the graph, indicating that as income increases, the amount of fruit also increases. In the second figure on the right, there is a **negative correlation** between **Income** (independent variable) and the **amount of fruit** (dependent variable). This can be observed from the

downward slope of the graph, indicating that as income increases, the amount of fruit decreases.



Here all the points are scattered and it is difficult to conclude whether the slope of the graph is positive or negative, hence it can be concluded that there is **no correlation**.

Now it is your turn to explore correlation analysis and scatter plots.

Reference

- **https://en.wikipedia.org/wiki/Correlation**
- **https://en.wikipedia.org/wiki/Scatter_plot**