**BUSINESS INTELLIGENCE LAB**

**EXPERIMENT NO:4**

**AIM:**To perform exploratory data analysis and data visualization using python.

**THEORY:**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations.

Measures of central tendency

A measure of central tendency (also referred to as measures of center or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or center of its distribution.

There are three main measures of central tendency:

* mode: The mode is the most commonly occurring value in a distribution.
* median: The median is the middle value in distribution when the values are arranged in ascending or descending order.
* mean: The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.

Each measure describes a different indication of the typical or central value in the distribution.

Dispersion:

Dispersion of data used to understand the distribution of data. It helps to understand the variation of data and provides a piece of information about the distribution data. Range, IOR, Variance, and Standard Deviation are the methods used to understand the distribution data.

* The range is the easiest dispersion of data or measure of variability. To calculate a range, prepare all the values in ascending order, then subtract the lowest from the highest value.
* Range = Highest\_value – Lowest\_value
* IQR is a range (the boundary between the first and second quartile) and Q3 (the boundary between the third and fourth quartile).IQR is preferred over a range as, like a range, IQR is not influenced by outliers. IQR is used to measure variability by splitting a data set into four equal quartile
* Variance is a simple measure of dispersion. Variance measures how far each number in the dataset is from the mean. To compute variance first, calculate the mean and squared deviations from a mean.
* Standard deviation is a square root of the variance to get original values. A low standard deviation indicates data points close to the mean.

Correlation:

Correlation explains how one or more variables are related to each other. These variables can be input data features that have been used to forecast our target variable.

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

Data visualization

Data visualization is a graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. This blog on data visualization techniques will help you understand detailed techniques and benefits.

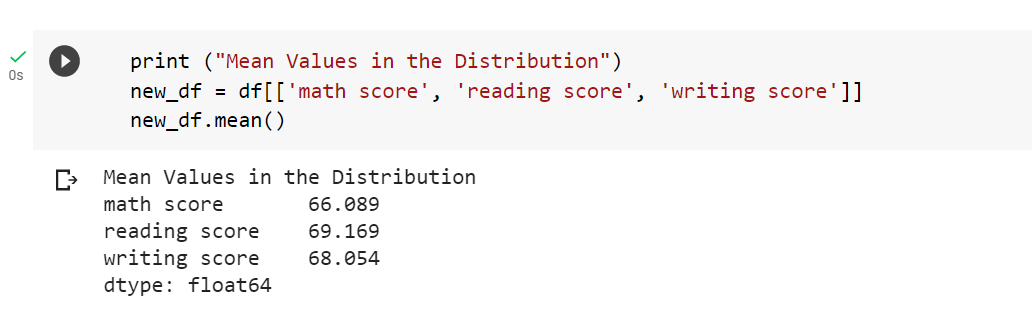
Some Data Visualization Techniques:

* Box plots
* Histograms
* Heat maps
* Charts
* Treemaps
* Word Cloud/Network diagram

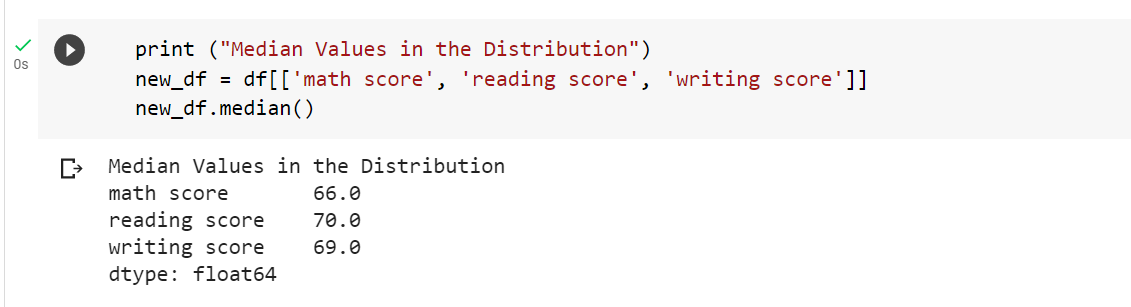
**IMPLEMENTATION:**

1. **Descriptive analysis - statistical measures of data (Central tendency)**

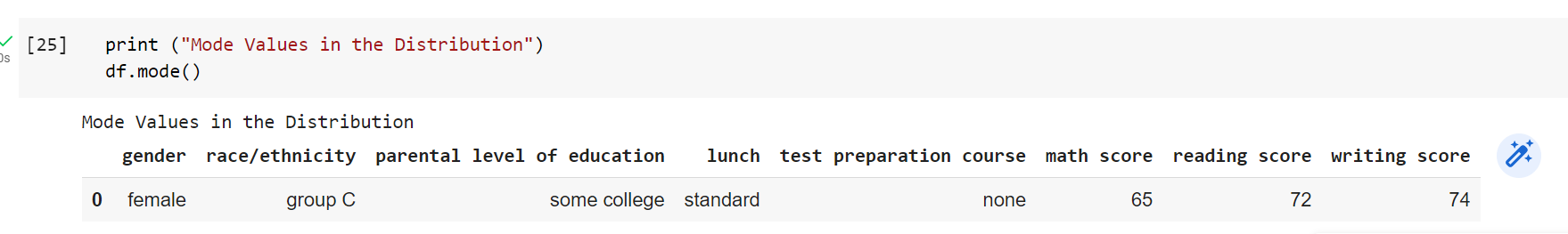
1. Mean



1. Median

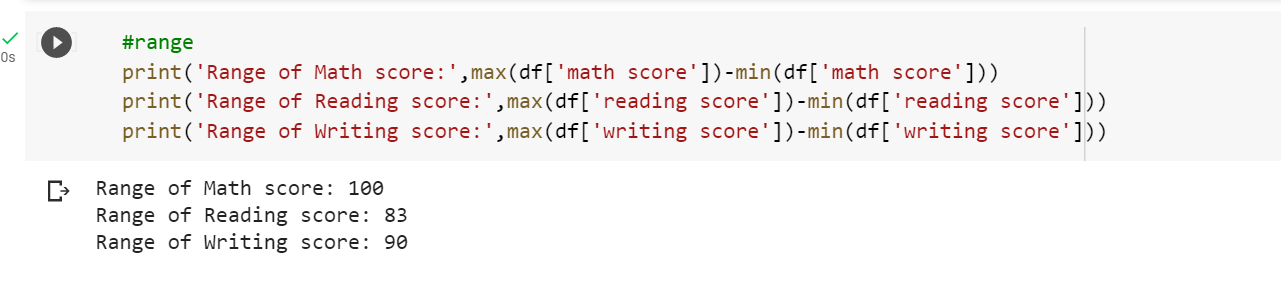


1. Mode

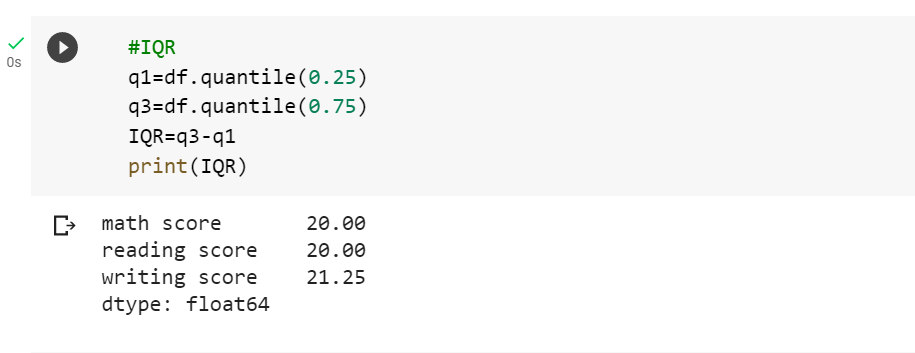


2. **Descriptive analysis - statistical measures of data (Dispersion)**

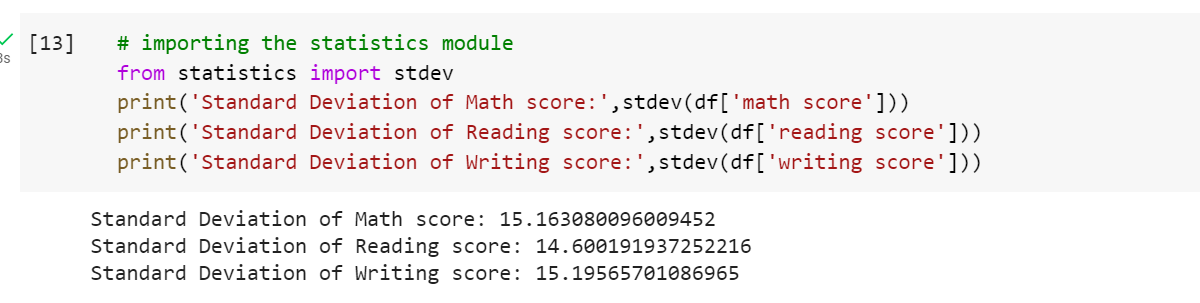
1. Range



1. Interquartile range

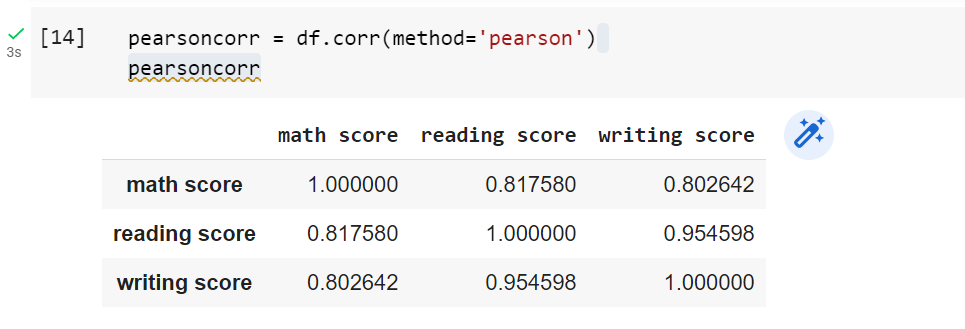


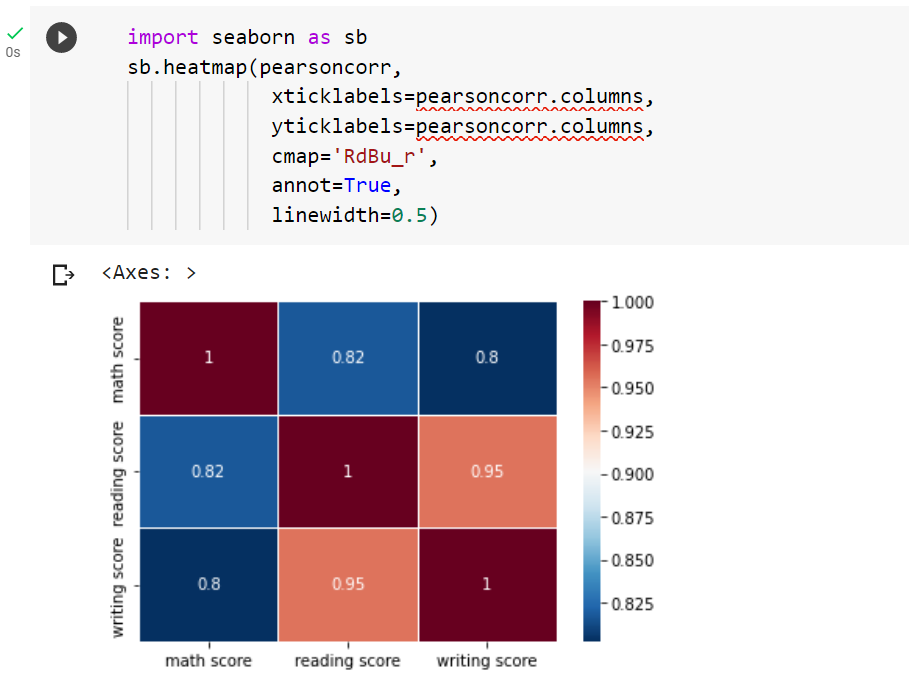
1. Standard deviation



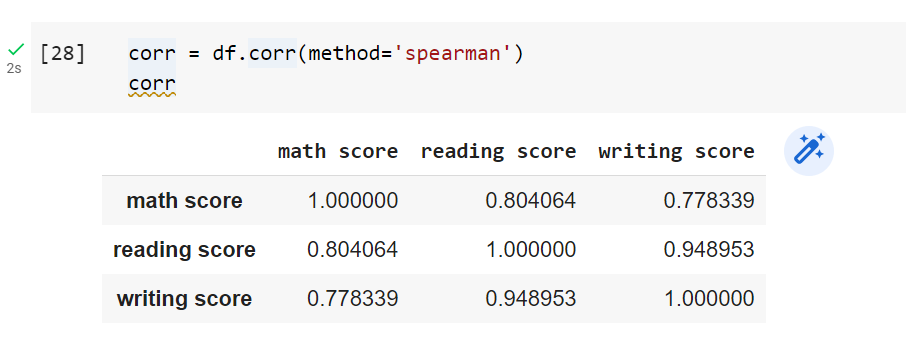
**3. Correlation between attributes**

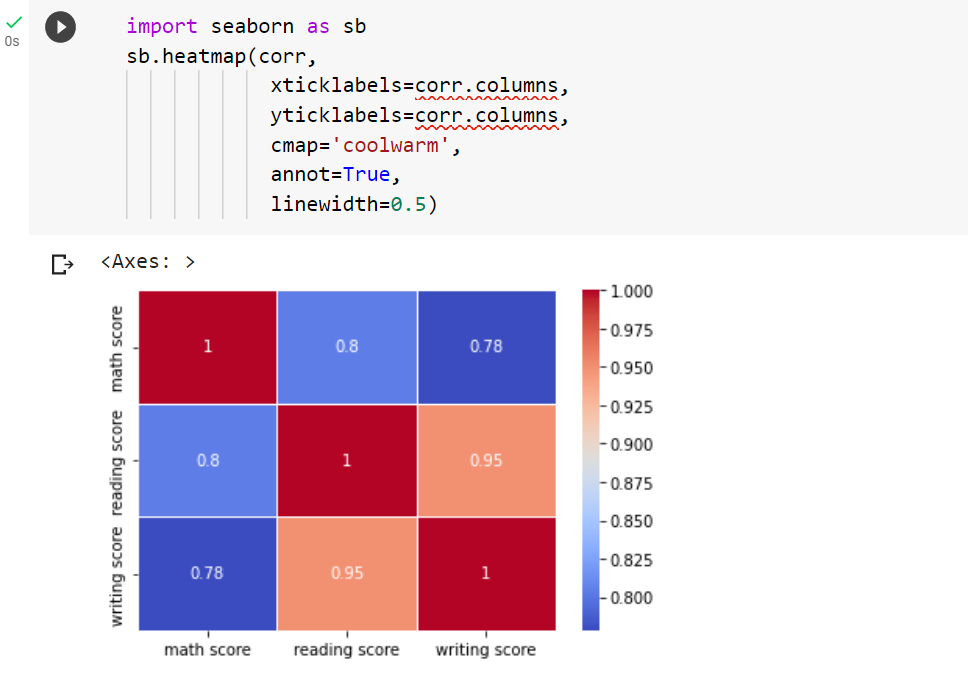
1. Pearson's correlation coefficient



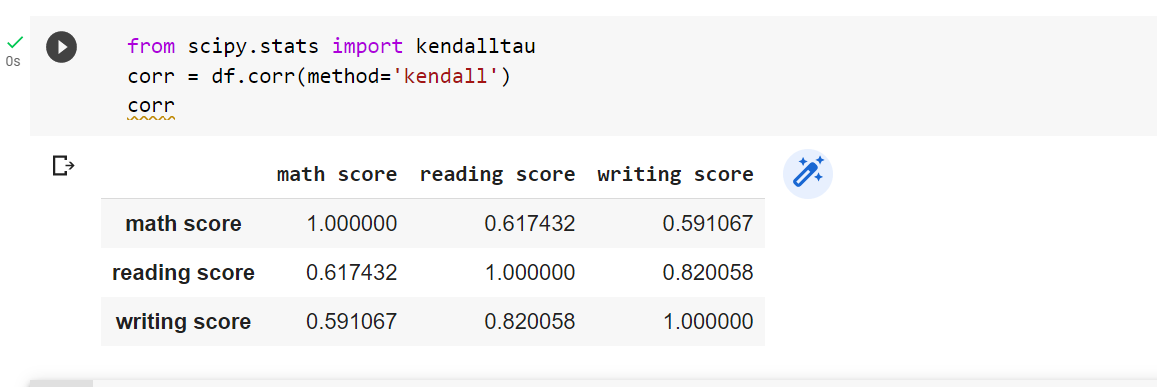


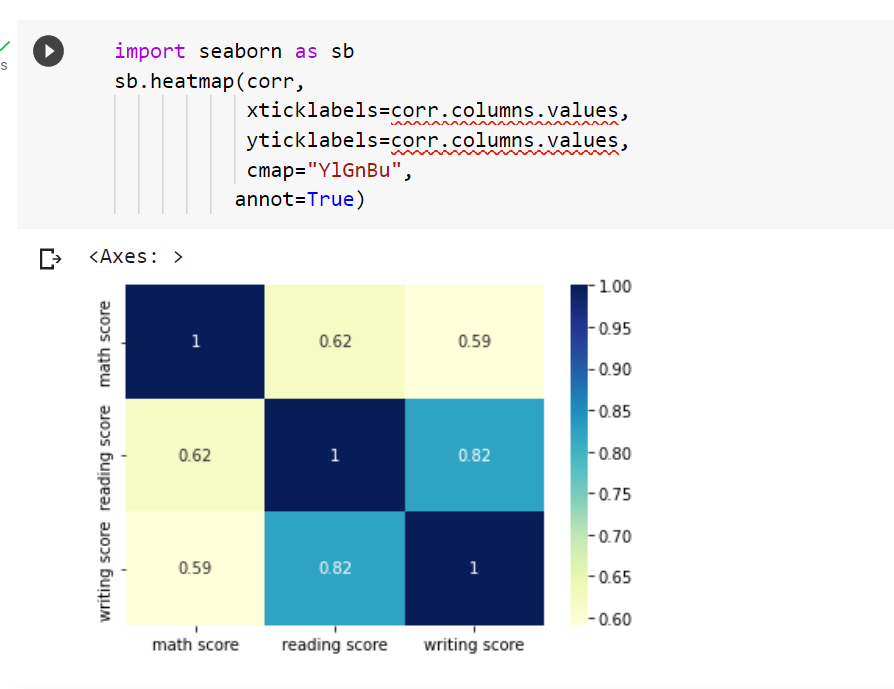
1. Spearman's rank correlation coefficient





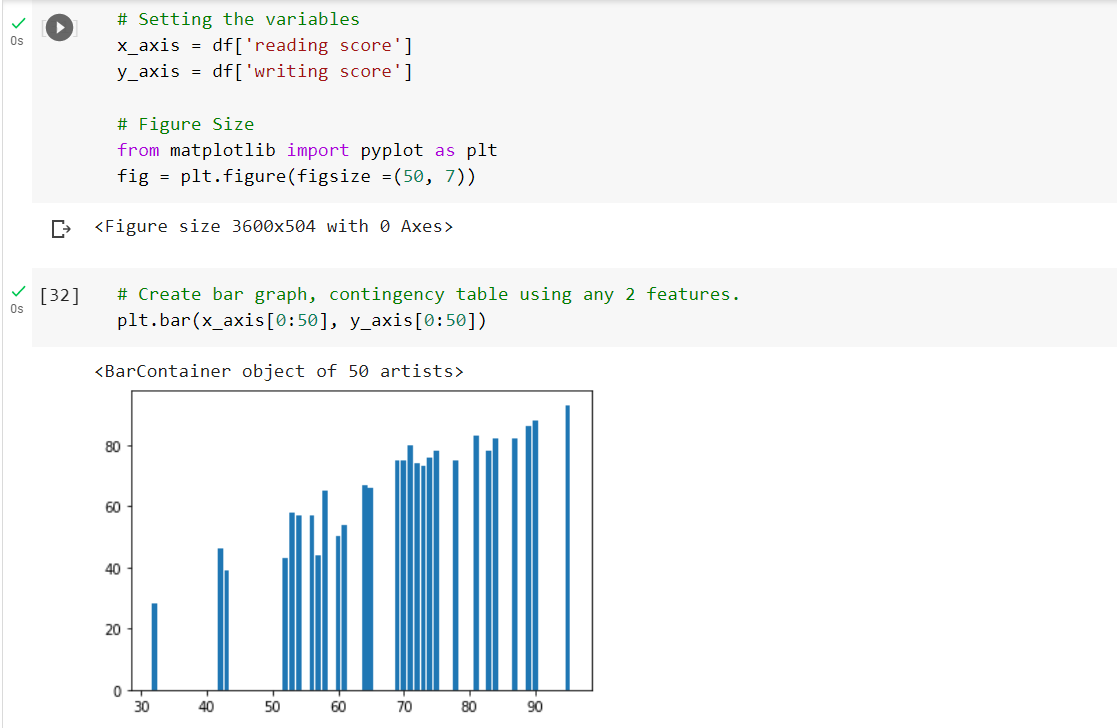
1. Kendall's tau correlation coefficient





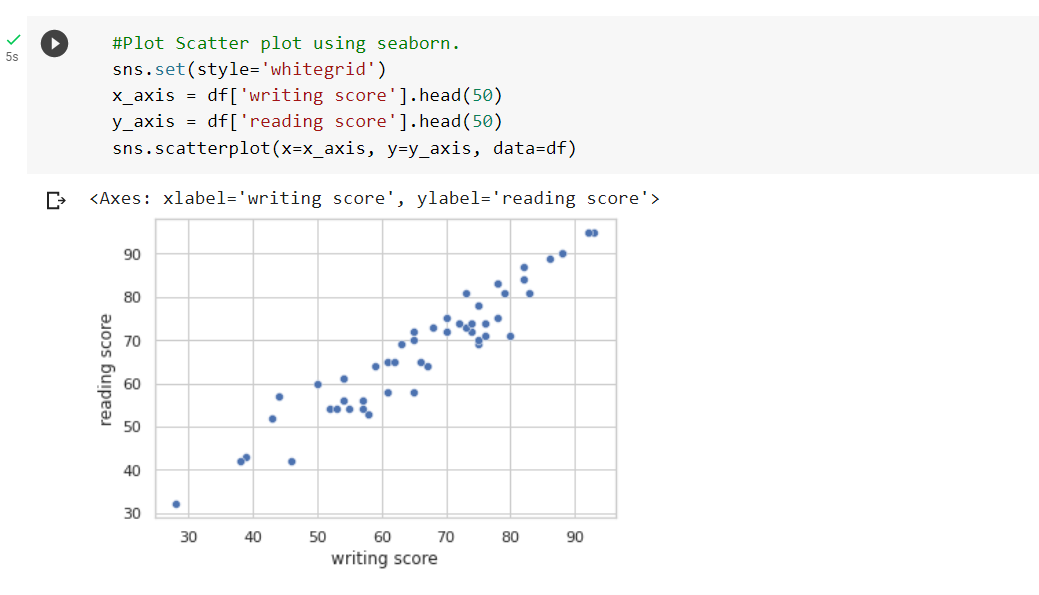
4. Visualization techniques

1. Bar graph



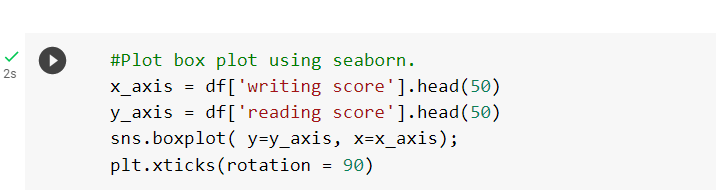
From the above bar graph, we understand that with the increasing value of the reading score, the value of the writing score is also increasing.

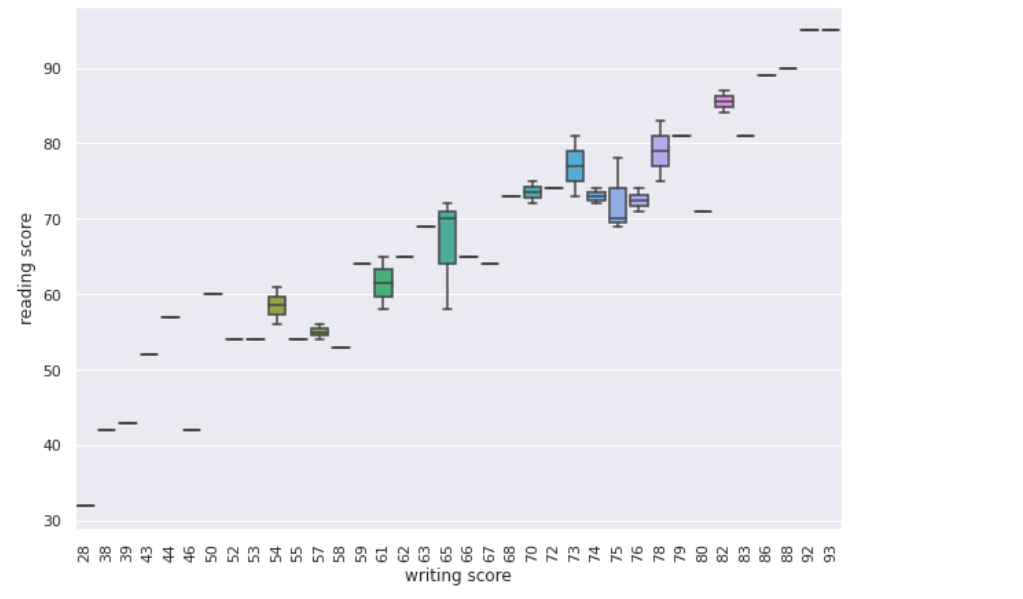
1. Scatter plot



In the above scatter plot, the points are arranged in a linear pattern, which may indicate a strong positive or negative correlation between the writing score and reading score.

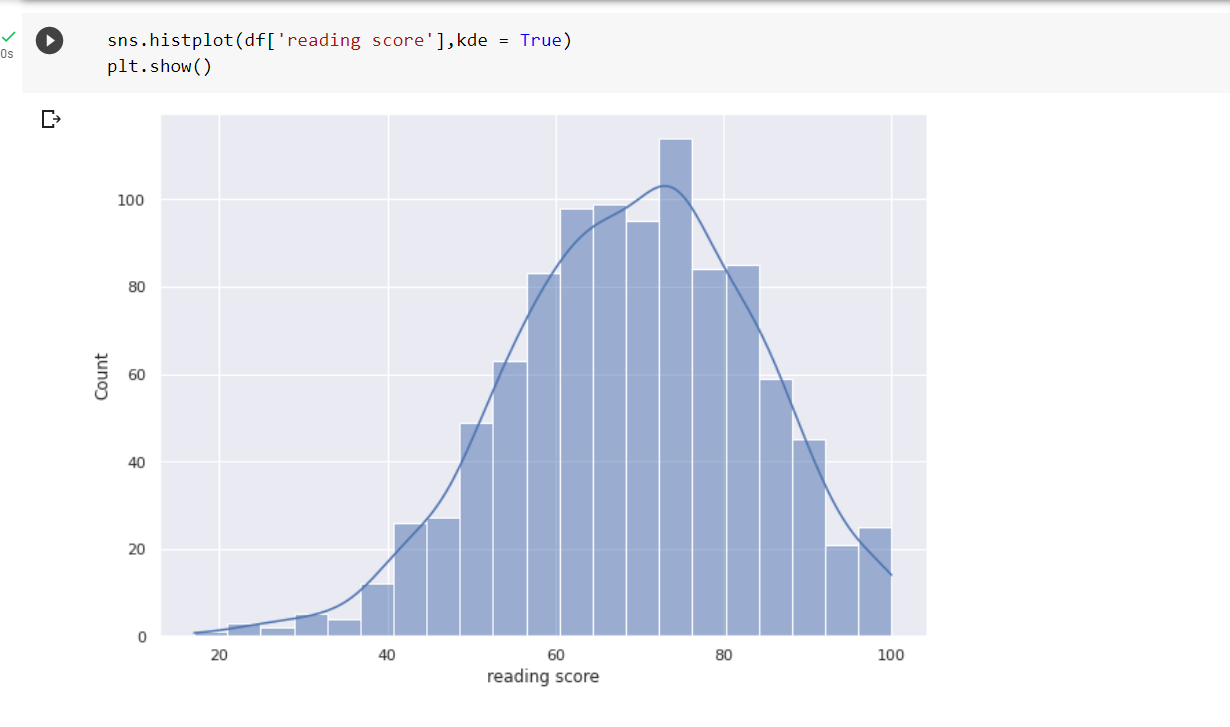
1. Box plot

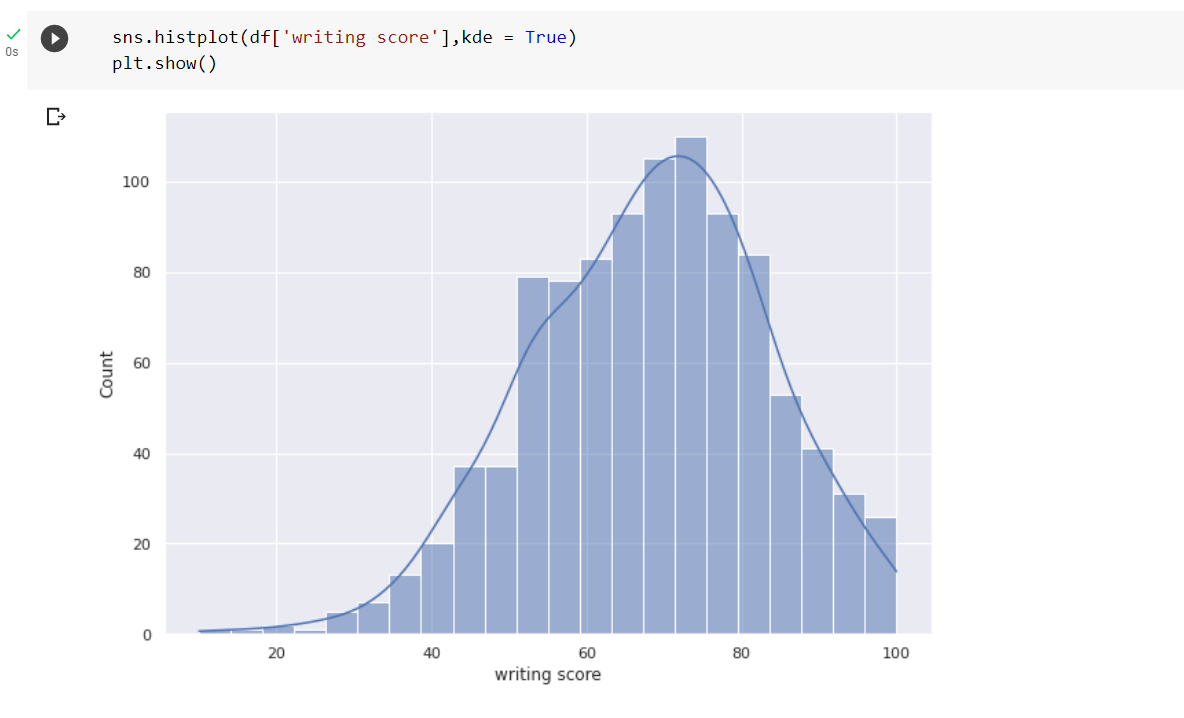




The data points outside of the whiskers of box plot are considered outliers .

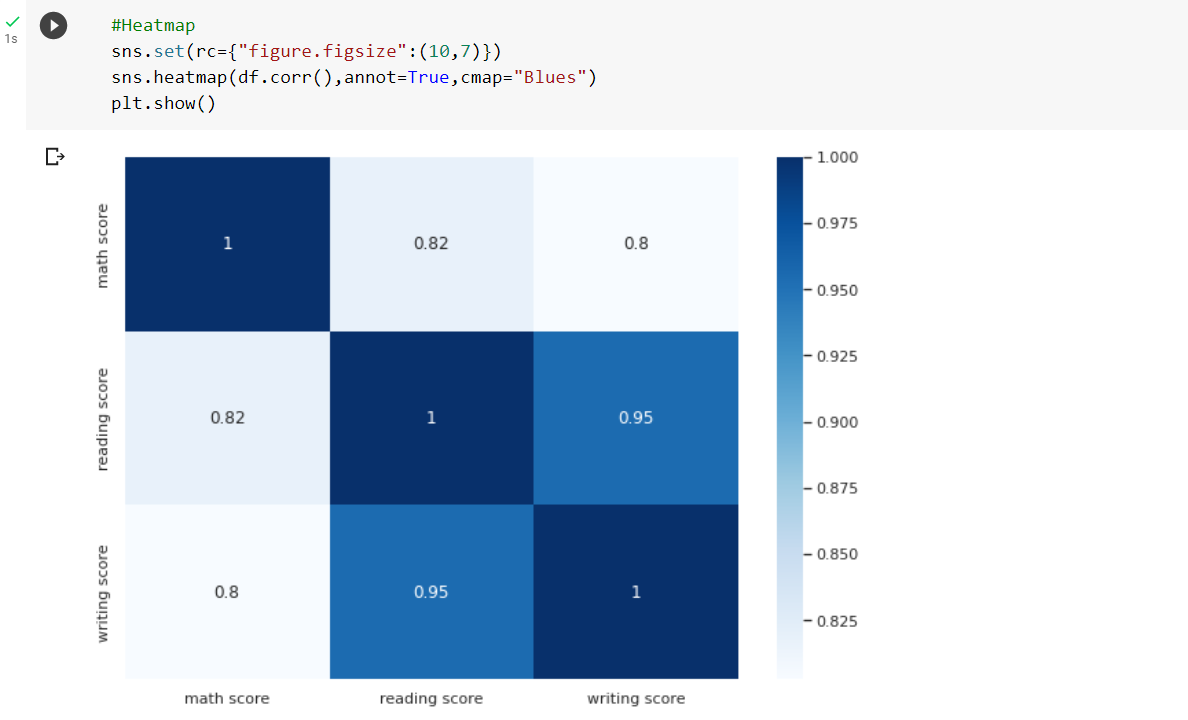
1. Histogram





The histogram of the writing score and the reading score is a skewed histogram . It is negatively skewed as the tail increases towards left.

1. Heatmap



From the above heatmap, we understand that reading scores and writing scores are highly correlated with each other.

**CONCLUSION**: In this experiment, we have performed exploratory data analysis and data visualization using python.