DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

19AD504 – DATA VISUALIZATION

UNIT – 4

4.4 COMMUNICATING CORRELATION AND REGRESSION

Correlation is a statistical measure that indicates the extent to which two or more variables are related.

It is a powerful tool for understanding the relationships between different data points and can be used to identify patterns and trends in large data sets.

Scatter Plots

One of the most common ways to visualize correlation is through scatter plots.

Scatter plots are graphs that display the relationship between two variables, with one variable plotted on the x-axis and the other variable plotted on the y-axis.

By plotting the data points on a graph, it becomes easier to see if there is a correlation between the two variables.
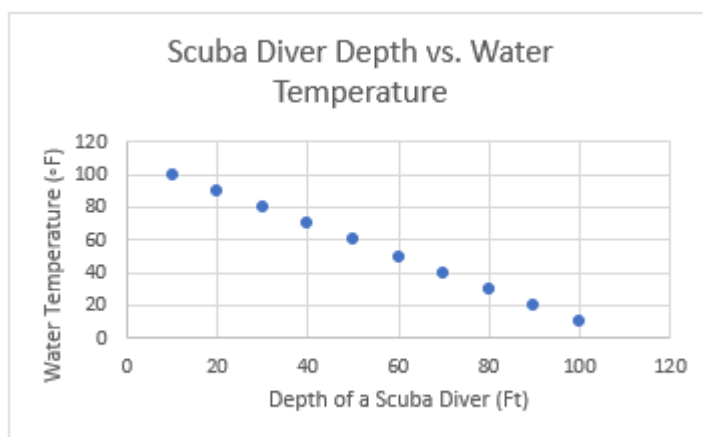
The Correlation Coefficient

The correlation coefficient is a statistical measure that ranges from -1 to 1 and indicates the strength and direction of the correlation between two variables.

A correlation coefficient of 1 indicates a perfect positive correlation, while a correlation coefficient of -1 indicates a perfect negative correlation.

A correlation coefficient of 0 indicates no correlation between the two variables.
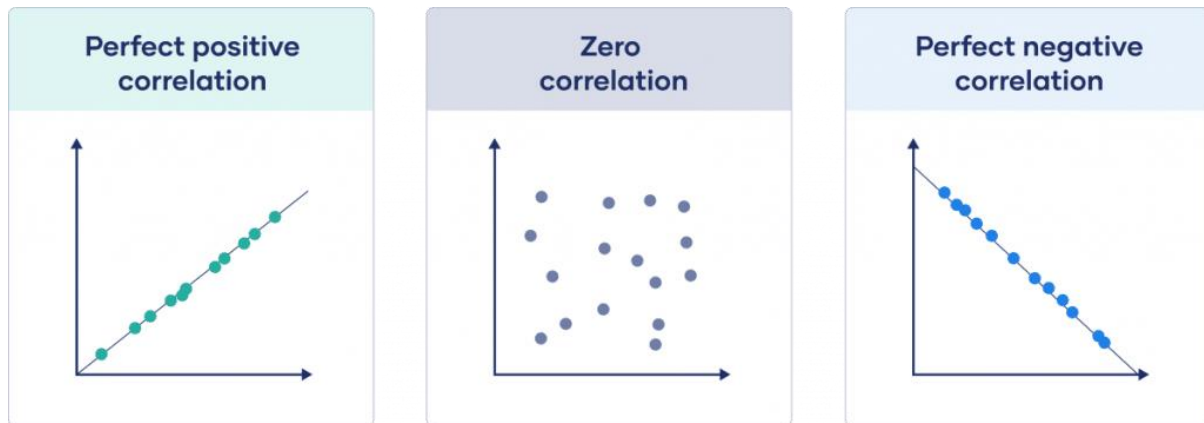
Scatter Plots

Scatter plots are a powerful tool for visualizing the relationship between two variables.

They are commonly used in correlation and regression analysis to identify patterns and trends in data.

Outliers and Influential Points

Outliers and influential points can have a significant impact on the interpretation of a scatter plot.

Outliers are data points that fall far outside the expected range of values, while influential points are data points that have a disproportionate effect on the slope of the trend line.

It is important to identify and address these points when interpreting scatter plots.

Interpreting Scatter Plots

When interpreting scatter plots, it is important to look for patterns and trends that can provide insights into the relationship between the two variables.

A scatter plot with a positive slope indicates a positive correlation, while a negative slope indicates a negative correlation.

A scatter plot with no clear pattern or trend indicates little or no correlation.

What is the Correlation Coefficient?

The correlation coefficient is a statistical measure that indicates how strongly two variables are related.

How is the Correlation Coefficient Calculated?

The correlation coefficient is calculated by dividing the covariance of the two variables by the product of their standard deviations.

What Does the Correlation Coefficient Tell Us?

The correlation coefficient ranges from -1 to 1. A value of 1 indicates a perfect positive correlation, while a value of -1 indicates a perfect negative correlation. A value of 0 indicates no correlation between the two variables.

What is Regression?

Regression is a statistical method used to determine the relationship between two or more variables. It is often used to predict the value of one variable based on the value of another variable.

Linear Regression

Linear regression is a type of regression analysis that models the relationship between two variables by fitting a linear equation to the observed data.

The goal of linear regression is to find the best-fitting straight line through the data points.

Linear regression is a statistical approach to modeling the relationship between a dependent variable and one or more independent variables.

It is used to predict the value of the dependent variable based on the values of the independent variables.

In linear regression, a line is fitted to the data points in such a way that the sum of the squared distances between the line and the data points is minimized.

The line is then used to make predictions about the dependent variable based on the values of the independent variables.