Assignment-1

Data Analysis Principles

Data Analysis Principles involve systematically applying statistical and logical techniques to describe, condense, and evaluate data. Key principles include understanding the data's source, context, and quality, cleaning the data to remove errors, exploring the data using descriptive statistics and visualization techniques, modeling the data with statistical models for predictions or inferences, and interpreting results to draw meaningful conclusions and make informed decisions.

Statistical Analysis

Statistical Analytics uses statistical methods to collect, review, analyze, and draw conclusions from data. This includes descriptive statistics (mean, median, mode, range, variance, standard deviation) to summarize data features, inferential statistics (hypothesis testing, confidence intervals, regression analysis) to extend conclusions beyond immediate data, predictive analytics to forecast future outcomes, and prescriptive analytics to recommend actions based on data analysis.

Hypothesis Testing

Hypothesis Testing is a method for making decisions using data from experiments or studies. It involves a null hypothesis (H0) of no effect or difference and an alternative hypothesis (H1) of an effect or difference. The p-value indicates the probability of observing the data if H0 is true, with small p-values suggesting strong evidence against H0. Type I errors (false positives) occur when H0 is wrongly rejected, while Type II errors (false negatives) occur when H0 is wrongly not rejected. The significance level (α), commonly set at 0.05, is the threshold for rejecting H0.

Regression

Regression analysis helps understand relationships between dependent and independent variables. Linear regression fits a linear equation to data, multiple regression uses multiple independent variables, logistic regression predicts probabilities for categorical outcomes, and polynomial regression models relationships as nth degree polynomials.

Correlation

Correlation measures the strength and direction of relationships between two variables using the correlation coefficient (r), ranging from -1 to 1. A positive correlation means both variables move in the same direction, while a negative correlation means one increases as the other decreases. No correlation indicates no relationship. Importantly, correlation does not imply causation; it simply shows a relationship between variables.

Anova

ANOVA (Analysis of Variance) is a method for comparing means across multiple groups to determine if at least one group mean differs significantly. One-way ANOVA compares means across one factor with multiple levels, while two-way ANOVA examines the influence of two categorical variables. ANOVA relies on assumptions of normality, homogeneity of variances, and independence of observations. The F-statistic, the ratio of variance between group means to variance within groups, determines the p-value for the test.

5V’s of Big Data

* **Volume**: This refers to the vast amounts of data generated every second from various sources such as social media, sensors, transactions, and more. The sheer scale of data that organizations have to handle and analyze is massive.
* **Velocity**: This describes the speed at which data is generated, collected, and processed. In many cases, data needs to be processed in real-time or near-real-time to be useful, such as in financial transactions, social media feeds, and IoT applications
* **Variety**: This refers to the different types of data that are available. Data can be structured (like databases), semi-structured (like XML files), and unstructured (like text, video, and audio files). The diversity of data types presents challenges in terms of storage, processing, and analysis.
* **Veracity**: This pertains to the quality and accuracy of the data. High veracity means the data is trustworthy and accurate, while low veracity indicates a higher level of uncertainty and the potential presence of noise or errors in the data.
* **Value**: This is about the usefulness of the data. Data alone doesn't hold value; it needs to be processed and analyzed to extract actionable insights that can drive business decisions and strategies. The ultimate goal of big data initiatives is to derive significant value from the data.