Prac1

what does numpy and pandas does explain in very short

- NumPy: Python library for numerical computing, providing support for arrays, matrices, and mathematical functions to operate on these data structures efficiently.
- Pandas: Python library for data manipulation and analysis, offering data structures like DataFrame for handling tabular data and tools for cleaning, transforming, and analyzing datasets easily.

explain what does describe function do explaiin in short

The describe() function in pandas generates descriptive statistics of a DataFrame or Series, including count, mean, standard deviation, minimum, maximum, and percentiles, giving a quick summary of the data's distribution and central tendencies.

what are quantiles?Q1,Q2,Q3 explain in short

Quantiles are points in a dataset that divide the data into equally sized subsets. The three most common quantiles are:

Q1 (First Quartile): It represents the value below which 25% of the data falls. It's also known as the 25th percentile.

Q2 (Second Quartile or Median): It's the middle value in the dataset when it's sorted in ascending order. Fifty percent of the data falls below and above this point.

Q3 (Third Quartile): It indicates the value below which 75% of the data falls. It's also called the 75th percentile.

what does fillna and dropna do explain in short

fillna(): A method in pandas used to fill missing (NaN) values in a DataFrame or Series with specified values, such as a single scalar value or a dictionary mapping column names to values.

dropna(): A method in pandas used to remove rows or columns containing missing (NaN) values from a DataFrame. It provides options to drop rows or columns based on different criteria like any NaN present or only if all values are NaN in a row or column.

what is qualitative and quantitative data explain with example in short

Qualitative Data: Qualitative data describes qualities or characteristics and cannot be measured numerically. It includes attributes like colors, feelings, or opinions. For example, the color of a car, the taste of food, or the sentiment in a text review.

Quantitative Data: Quantitative data consists of numerical measurements or counts that can be subjected to mathematical operations. It represents quantities or amounts. Examples include the height of a person, the temperature of a room, or the number of cars in a parking lot.

explain Data Formatting and Data Normalization in short

Data Formatting: Data formatting involves arranging or modifying data to adhere to a specific structure, style, or format. It ensures consistency and readability. For example, formatting dates as "YYYY-MM-DD" or currency values with a specific number of decimal places.

Data Normalization: Data normalization is the process of rescaling the values of features (variables) in a dataset to a standard range. It helps in bringing all features to a similar scale, preventing any one feature from dominating others. Common normalization techniques include Min-Max scaling and Z-score normalization.

Prac2

what is outliers explain inshort

Outliers are data points that significantly differ from other observations in a dataset. They can skew statistical analyses and distort the interpretation of results. Outliers can occur due to measurement errors, natural variability, or rare events. Identifying and handling outliers is important to ensure accurate analysis and modeling of data.

What is IQR method explain in short

The Interquartile Range (IQR) method is a statistical technique used to identify outliers in a dataset. It involves calculating the difference between the third quartile (Q3) and the first quartile (Q1) of the data. Then, a data point is considered an outlier if it falls below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR. This method is based on the spread of the middle 50% of the data and is less sensitive to extreme values compared to other methods.

what are the methods to handle outlierrs like IQR

Z-score Method: Calculate the Z-score for each data point and remove or mark data points with Z-scores above or below a certain threshold (e.g., 3 standard deviations from the mean).

Percentile Method: Remove data points that fall above or below a certain percentile threshold (e.g., 99th or 1st percentile).

explain the steps in IQR in very short

- Calculate Q1 and Q3: Compute the first quartile (Q1) and third quartile (Q3) of the dataset.
- Compute IQR: Calculate the Interquartile Range (IQR) by subtracting Q1 from Q3.
- Identify Outliers: Any data point below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR is considered an outlier.
- Handle Outliers: Remove, mark, or replace outliers based on the specific requirements of the analysis.

what is data transformation explain in short

Data transformation involves modifying or converting the original data into a different format or representation to improve its quality, compatibility, or usefulness for analysis. It can include tasks like scaling, encoding, standardizing, or normalizing data to meet specific requirements or to make it suitable for certain algorithms or analyses.

what is scaling explain it in short

Scaling is the process of adjusting the range of values in a dataset to a standard scale, typically to facilitate better comparison or analysis. It ensures that all features contribute equally to the analysis by bringing them to a similar numerical range. Common scaling techniques include Min-Max scaling, Standardization (Z-score scaling), and Robust scaling.

what is use of sklearn,seaborn,matplotlib in short

- Scikit-learn (sklearn): A machine learning library in Python used for building and applying various machine learning models, including classification, regression, clustering, and dimensionality reduction.
- Seaborn: A statistical data visualization library in Python built on top of matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
- Matplotlib: A comprehensive plotting library in Python for creating static, interactive, and publication-quality visualizations. It offers a wide range of plotting functions for various types of data and customization options.

Prac3

what does groupby( ) fo explain inshort

The groupby() function in pandas is used to group data in a DataFrame based on specified columns or criteria. It creates groups of rows that share the same values in the specified columns and allows for applying aggregate functions, such as sum, mean, or count, to each group separately, facilitating data analysis and summary.

what does .std gives explain in short

The .std() method in pandas calculates the standard deviation of numeric data in a DataFrame or Series. Standard deviation measures the dispersion or spread of values around the mean, providing insight into the variability of the data**.**

Prac4,5

what is regression explain in short

Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It aims to predict the value of the dependent variable based on the values of the independent variables. Regression analysis helps understand how changes in independent variables affect the dependent variable and is commonly used for prediction, forecasting, and understanding the underlying relationships in data.

what is dependent and independant variable explain in short

Dependent Variable: Also known as the response variable, it's the outcome or target variable in a statistical model. Its value depends on the values of other variables in the model. For example, in a study examining the effect of study time on exam scores, the exam score is the dependent variable because it depends on the amount of study time.

Independent Variable: Also known as the predictor variable, it's a variable that stands alone and isn't affected by other variables in the model. It's manipulated or controlled in an experiment or study to observe its effect on the dependent variable. In the example above, study time is the independent variable because it's manipulated to observe its effect on exam scores.

what are the types of regression explain inshort

Linear Regression: It models the relationship between a dependent variable and one or more independent variables by fitting a straight line or hyperplane to the data.

Logistic Regression: It's used when the dependent variable is binary (e.g., yes/no, true/false). It models the probability of the occurrence of a certain event.

Polynomial Regression: It models the relationship between the dependent variable and independent variable as an nth-degree polynomial.

what does train_test_split() do explain in short

The train_test_split() function in machine learning is used to split a dataset into two separate sets: one for training a model and the other for testing its performance. It randomly partitions the data into training and testing subsets based on the specified ratio, allowing for the evaluation of the model's generalization ability on unseen data.

what is mean square and r square explain inshort

- Mean Square Error (MSE): It measures the average squared difference between the actual values and the predicted values in a regression model. A lower MSE indicates better model performance, with zero being the perfect score.
- R-squared (R^2): It represents the proportion of variance in the dependent variable that is explained by the independent variables in a regression model. It ranges from 0 to 1, where 1 indicates that the model explains all the variability in the data, and 0 indicates that it explains none.

explain Confusion matrix TP, FP, TN, FN, Accuracy, Error rate, Precision, Recall in short

- True Positive (TP): The number of correct positive predictions by the model.
- False Positive (FP): The number of incorrect positive predictions by the model.
- True Negative (TN): The number of correct negative predictions by the model.
- False Negative (FN): The number of incorrect negative predictions by the model.
- Accuracy: The proportion of correct predictions out of the total predictions made by the model.
- Error Rate: The proportion of incorrect predictions out of the total predictions made by the model.
- Precision: The proportion of true positive predictions out of all positive predictions made by the model.
- Recall: The proportion of true positive predictions out of all actual positive instances in the data.
- These metrics help assess the performance of classification models by evaluating their ability to correctly classify instances into different classes.

Prac6

Explain gaussianNB in short

Gaussian Naive Bayes (GaussianNB) is a variant of the Naive Bayes algorithm used for classification tasks. It assumes that the features follow a Gaussian (normal) distribution and calculates the likelihood of a given feature value belonging to a particular class using the Gaussian probability density function. Despite its simplicity and the "naive" assumption of feature independence, GaussianNB can perform well on a wide range of classification problems, especially when dealing with continuous data.

Prac7

explain Tokenization, POS Tagging, stop words removal, Stemming and Lemmatization in short

Tokenization: It's the process of breaking text into smaller units, such as words or sentences, called tokens.

POS Tagging (Part-of-Speech Tagging): It's the process of assigning grammatical tags (e.g., noun, verb, adjective) to each word in a sentence to indicate its syntactic role.

Stop Words Removal: It involves filtering out common words (e.g., "the", "and", "is") from text data as they don't contribute much to the meaning of the text.

Stemming: It's the process of reducing words to their root or base form by removing suffixes and prefixes. For example, "running" becomes "run".

Lemmatization: It's similar to stemming but produces valid words (lemmas) by considering the word's meaning and context. For example, "running" becomes "run".

explain Term Frequency and Inverse Document Frequency.

Term Frequency (TF): It measures the frequency of a term (word) within a document. It indicates how often a term appears in a document relative to the total number of terms in the document.

Inverse Document Frequency (IDF): It measures the importance of a term in a collection of documents. It is calculated by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of the result. This helps in reducing the weight of terms that appear in many documents and increasing the weight of terms that appear in few documents.

TF-IDF (Term Frequency-Inverse Document Frequency) is a commonly used weighting scheme that combines TF and IDF to evaluate the importance of terms in a document collection.

what is corpous explain in very short

A corpus is a collection of text documents or linguistic data, often used for analysis, research, or training machine learning models. It serves as a representative sample of a language or domain, containing various texts such as articles, books, conversations, and more.

Prac8 9 10

explain countplot histplot scatterplot boxplot in short

Countplot: A countplot is a type of categorical plot that shows the counts of observations in each category using bars.

Histplot: A histplot is a type of plot that shows the distribution of a continuous variable by binning its values into intervals and plotting the frequency of observations in each interval as bars.

Scatterplot: A scatterplot is a type of plot that shows the relationship between two continuous variables by representing each data point as a dot on the plot.

Boxplot: A boxplot is a type of plot that shows the distribution of a continuous variable using quartiles. It displays the median, quartiles, and any outliers in the data.