DATA ANALYSIS AND PREPROCESSING

Date : 20-02-2025 Full Name: Vighnesh Reddy

1. **What is descriptive statistics?**

   Descriptive statistics is a branch of statistics that focuses on summarizing and describing the basic features of a dataset in a clear and understandable way. It provides a clear picture of the data by organizing, displaying, and quantifying its main characteristics, without making inferences or predictions about a larger population.

   For example, if we have a list of test scores of a class, it tells us the highest and the lowest score, the average score, and also the spread of the score(for example where do most of the students lie)

   It also tells us about the distribution of the data, the frequency of the data and where the center of the data lies

2. **Explain mean, median, mode, standard deviation, and variance**

   Mean : The sum of all data points divided by the number of data points. It’s the "central value" of the dataset.
   $\mu = (x_1 + x_2 + ... + x_n) / n$

   Mode : The value that appears most often in the dataset. There can be one mode, multiple modes, or no mode if all values appear equally often.

   Median : The middle value when the data is sorted in ascending order. If there’s an even number of data points, it’s the average of the two middle values.

   Variance : Measures how much the data points differ from the mean, on average. It’s the average of the squared differences from the mean.
   $\sigma^2 = \Sigma(x_i - \mu)^2 / n$

   Standard Deviation : The square root of the variance. It’s a measure of spread in the same units as the data.

3. **What is normalization? What are the different methods of Normalization?**

   Normalization is a way to adjust data so it fits a standard scale, like 0 to 1 or a normal distribution. It makes numbers easier to compare by removing differences in size or units. Common methods use minimums, maximums, averages, or ranges to rescale the data simply and consistently.

Normalization is essentially done so that we can compare the data in a much more easier and efficient way.

Different methods of normalization :

Min-Max: X_normalized = (X - X_min) / (X_max - X_min)

Z-Score: X_normalized = (X - μ) / σ

Decimal Scaling: X_normalized = X / 10^j

Robust Scaling: X_normalized = (X - X_median) / (Q3 - Q1)
Q1 = first quartile (25th percentile)
Q3 = third quartile (75th percentile)

Max-Absolute Scaling : X_normalized = X / Max(|X|)

4. What are outliers, and different methods of outlier removal?

Outliers are data points that differ significantly from the rest of the dataset. They're the odd ones out—either much larger, much smaller, or just not fitting the pattern. These outliers can usually mess up mean, variance and other trends.

Methods of outlier removal :

IQR Method (Interquartile Range):
This method uses the middle 50% of the data between the 25th and 75th percentiles to establish a normal range. It calculates the spread of this central portion and extends it by a factor, typically 1.5, to create boundaries. Anything beyond these limits is removed. It's highly efficient because it doesn't rely on the average, which outliers can distort, making it reliable for skewed or messy data.

Z-Score Method (Standard Deviation-Based):

This approach checks how far each point is from the average, measured in standard deviations. Points too far away often more than 3 standard deviations are flagged as outliers and dropped. It's efficient for data that's roughly bell-shaped, as it leverages standard statistical properties to quickly identify extremes. It's straightforward to apply and performs well when outliers aren't too numerous, though it's less robust if the data deviates from normality.

Winsorizing:

Rather than removing outliers, this method caps them by setting upper and lower limits, often based on percentiles or a range from the data's center. Extreme values are replaced with these boundary values, keeping the dataset intact but reducing the influence of outliers.

5. What are common errors found in raw data and what are the methods used to fix them?

The most common errors that are found in raw data are dulplicates, missing values, outliers, inconsistent formats, incorrect data types, typing errors.

Common Methods used to fix them :
Duplicates in the dataset can be replaced with the average, mean, mode or median.
Delete or merge data which are duplicated.
You can use the above mentioned methods to remove outliers
We can convert all the data into a uniform style
We can use spell checking, cross validation and fuzzy matching to eliminate typographical errors
We change change all formats of data types to match exactly its purpose.

6. What is probability distribution and its different types?

A probability distribution describes how the values of a random variable are spread out and how likely each value (or range of values) is to occur. Think of it as a map showing the chances of different outcomes in an uncertain situation like rolling a die or measuring heights in a group. It tells you what to expect and how much variation there might be.

Different Types :

- Uniform Distribution : Every outcome has the same chance. Picture rolling a fair die—each number (1 to 6) is equally likely.
- Binomial Distribution : Counts successes in a fixed number of yes/no trials, where each trial has the same probability of success. Think of flipping a coin 10 times and tracking heads—it shows how likely 0, 1, 2, etc., heads are
- Poisson Distribution : Counts successes in a fixed number of yes/no trials, where each trial has the same probability of success. Think of flipping a coin 10 times and tracking heads—it shows how likely 0, 1, 2, etc., heads are.
- Geometric Distribution : Tracks how many tries it takes to get the first success in repeated yes/no trials. For example, how many coin flips until you get heads.

7. What is correlation analysis, its significance, and the applications?

Correlation analysis is basically finding out how two variables move together or don't. It checks if one variable changes when another does, and if so, how strongly and in what direction. Its more about finding patterns but not proving why it happens

Significance : Correlation analysis is significant because it helps us spot connections in data without jumping to conclusions about why they exist.

It reveals relationships, simplifies complexity, avoids assumptions and helps us to be more efficient in working with the particular data set

Applications :

Business and Marketing: Links ad spending to sales to optimize budgets.
Finance: Tracks stock movements with market trends for smarter investing.
Healthcare: Connects lifestyle factors to disease rates for research leads.
Science and Research: Ties environmental changes to outcomes like growth or health.
Economics: Relates unemployment to inflation for economic forecasting.
Machine Learning: Identifies key features affecting predictions to refine models.

8. What is regression in detail?

Regression is a statistical method used to model the relationship between one or more independent variables (predictors) and a dependent variable (outcome). It's about finding a mathematical equation that best describes how the predictors influence the outcome, so you can predict or understand that outcome based on new data.

Types of Regression :

1. Simple Linear Regression
2. Multiple Linear Regression : more factors which influences the model
3. Polynomial Regression : relationship isn't exactly straight.
4. Logistic Regression : gives us a yes/no or 0/1
5. Non-linear Regression : for relationships with weird and custom curves

9. What is overfitting and underfitting? How to avoid it?

Overfitting: Overfitting happens when a model learns the training data too well, capturing every tiny detail including noise instead of the general trend. It performs great on the data it's trained on but fails miserably with anything new. The model becomes overly complex.It's like overanalyzing something simple until it's unrecognizable. This leads to predictions that are spot-on for the known stuff but way off for anything else, making it unreliable for new data.

Underfitting : Underfitting is when the model learns the data too simply, in this process it missed out on a lot. It misses the pattern and can make unrealistic assumptions. Even if the data has got curves and twists, it will just give us a straight line.

Ways to avoid overfitting and underfitting :
Reduce Model Complexity (to Prevent Overfitting)
Increase Model Complexity (to Address Underfitting)
Acquire More Data
Apply Regularization : essentially techniques like lasso and ridge
Adjust model parameters iteratively to balance bias and variance.
Preprocess Data : Remove outliers and irrelevant features to minimize noise, aiding in avoiding overfitting and supporting underfitting correction.
Enhance feature engineering and merge one or more parameters to make a better model
Monitor metrics like R-square score, RMSE scores

Metrics :
Mean Squared Error (MSE) : the more closer to zero the more better
Mean Absolute Error (MAE) : the more closer to zero the more better
R^2 (Coefficient of Determination): between 0 and 1. The closer to 1, the better is the model
Root Mean Squared Error (RMSE): the more closer to zero the better