Stats Project: Data Handling

Q1. How would you handle missing values in a dataset? Describe at least two methods.
Ans: Handling missing values is a crucial step in data preprocessing. Handling missing values in a dataset depends on the type of variable, i.e., numerical or character.

Method 1: Removal
This is the easiest way to handle missing values. You can either remove rows or columns containing missing values. However, this can result in the loss of important data points and can lead to less accuracy. This approach works well when there are only a very few rows that contain missing values or columns that contain missing data that are not important

Method 2: Mean/Median/Mode Imputation
This method involves replacing missing values with a suitable substitute value, such as the mean, median, or mode of the respective feature.

- Mean Imputation: Replace missing values with the mean of the feature. This method is suitable for continuous features.
- Median Imputation: Replace missing values with the median of the feature. This method is suitable for continuous features and is more robust to outliers.
- Mode Imputation: Replace missing values with the mode of the feature. This method is suitable for categorical features.

Q2. Explain why it might be necessary to convert data types before performing analysis.
Ans: Converting data types before performing analysis is often necessary to ensure accurate and meaningful results.

Common data type conversions include:
- Converting string data to numeric or date data
- Converting date data to timestamp data

By converting data types before performing analysis, you can ensure that your data is in the correct format, which can lead to more accurate and reliable results.

Stats Project: Statistical Analysis

Q1. What is a T-test, and in what scenarios would you use it? Provide an example based on sales data.

Ans: A T-test is a statistical hypothesis test used to determine if there's a significant difference between the means of two groups. It's commonly used in various fields, including medicine, social sciences, and business.

"Test done in python file shared"

Q2. Describe the Chi-square test for independence and explain when it should be used. How would you apply it to test the relationship between shipping mode and customer segment?

Ans: The Chi-square test for independence is a statistical test used to determine whether there is a significant association between two categorical variables. It compares the observed frequencies in a contingency table to the expected frequencies under the assumption that the variables are independent.

"Test done in python file shared"

StatsProject: Univariate and Bivariate Analysis

Q1. What is univariate analysis, and what are its key purposes?

Ans: Univariate analysis is the simplest form of data analysis, focusing on examining and summarizing a single variable. The term "univariate" refers to "one variable," meaning that this analysis does not deal with relationships or interactions between variables—only the distribution, characteristics, and structure of one variable at a time.

Key Purposes of Univariate Analysis

1. **Understanding Distribution**:
    - To assess how the data values of a single variable are spread or distributed (e.g., normal, skewed).
    - Helps identify patterns or trends in the data.
2. **Summarizing Data**:
    - Use measures of central tendency (mean, median, mode) to summarize the data.
    - Use measures of dispersion (range, variance, standard deviation) to understand variability.
3. **Detecting Outliers**:
    - Helps identify extreme values that deviate significantly from the rest of the data.
4. **Determining Data Type**:
    - Helps determine whether the variable is categorical, ordinal, or numerical, which guides further analysis or modeling techniques.
5. **Visualization**:
    - Enables visual representation to understand the variable's behavior and distribution using tools such as:
        - Histograms
        - Box plots
        - Bar charts
        - Pie charts
6. **Initial Data Exploration**:
    - Provides insights into the quality of data, such as missing values, data ranges, and errors, for cleaning and preprocessing.

Q2, Explain the difference between univariate and bivariate analysis. Provide an example of each.

The main difference between univariate and bivariate analysis is the number of variables being analysed:

Univariate analysis

Analyses a single variable. The purpose of univariate analysis is to describe data by summarizing it and finding patterns.

Bivariate analysis

Analyses two variables. The purpose of bivariate analysis is to explain the relationship between the two variables, and to determine if one variable can be predicted from the other.

Stats Project: Data Visualization

Q1. What are the benefits of using a correlation matrix in data analysis? How would you interpret the results?

Ans: A correlation matrix is a powerful tool in data analysis that offers several benefits,

- Benefits of using a correlation matrix:

1. Identifies relationships: A correlation matrix helps identify the strength and direction of linear relationships between variables.

2. Multivariate analysis: It enables the analysis of multiple variables simultaneously, providing a comprehensive understanding of the data.

3. Variable selection: Correlation matrices aid in selecting relevant variables for further analysis, such as regression or clustering.

4. Data quality check: It can help detect errors, outliers, or inconsistencies in the data.

5. Feature engineering: Correlation matrices can inform the creation of new features or the transformation of existing ones.

- Interpreting the results:

1. Correlation coefficients: Values range from -1 (perfect negative correlation) to 1 (perfect positive correlation). A value of 0 indicates no correlation.

2. Color-coding: Many correlation matrices use color-coding to represent the strength of correlations. Typically, darker colors indicate stronger correlations.

3. Significance testing: Some correlation matrices include p-values or significance levels to indicate whether the observed correlations are statistically significant.

4. Clusters and patterns: Look for clusters of highly correlated variables, which can indicate underlying patterns or relationships in the data.