Title: Day 4- Data Analysis and Operations

Introduction:

In Day 4 of our data analysis journey, we continued our exploration of data analysis techniques, focusing on univariate and bivariate analysis. These tasks and operations are fundamental for understanding data, ensuring data quality, and setting the stage for deeper analysis.

Tasks and Operations:

1. Fetch the basic information:

In both univariate and bivariate analyses, we used dataset.info() to retrieve fundamental information about the dataset, including data types and non-null values for each attribute.

2. Get the count of quantitative, qualitative, nominal, categorical data if any:

We categorized the data into quantitative and qualitative, and then further classified qualitative data into nominal and categorical types based on the number of unique values. Our code looped through the dataset's columns to perform this classification.

3. Display the descriptive statistics:

Descriptive statistics, such as mean, standard deviation, and quartiles, were displayed using dataset.describe() to provide a summary of numerical attributes.

4. Check for duplicate values:

Duplicate rows were identified using dataset.duplicated() to ensure data integrity. We also removed duplicate rows with dataset.drop_duplicates().

5. Know the unique values in a particular column:

We examined the unique values in the 'name' column to understand the variety of data in that attribute using data['name'].unique().

6. Check for null or missing values:

Missing values were detected using dataset.isnull() to create a binary matrix of missing values. We calculated the count of missing values for each attribute using mv.sum(). Then, we replaced the missing values with zeros using dataset.fillna(0, inplace=True). We also dropped rows with missing values using dataset.dropna(axis=0).

7. **Know data types of all attributes**:

The data types of each attribute were determined using data.dtypes.

8. Data Visualization (Specific Plots):

In both univariate and bivariate analyses, we created specific plots to visualize the data:

- Histogram: Visualized the distribution of the 'amount' attribute.
- Boxplot: Investigated data dispersion and identified outliers in the 'amount' attribute.
- Scatter Plot: Explored the relationship between 'isFraud' and 'isFlaggedFraud'.
- Line Plot: Visualized the sequential relationship between 'oldbalanceOrg' and 'newbalanceOrig'.
- Barplot: Displayed the marks of different individuals in the 'name' attribute.
- These operations are integral for understanding and preparing the dataset for further analysis.

Benefits:

- These operations provide a holistic overview of the dataset, ensuring data quality and reliability.
- They enable classification and summarization of attributes, data type understanding, and handling missing or duplicate data, which are crucial steps in the data preprocessing phase.

Conclusion:

In Day 4 of our data analysis journey, we performed a range of tasks and operations to assess and prepare our dataset. By fetching basic information, classifying data, exploring descriptive statistics, identifying duplicates, handling missing values, and understanding data types, we laid a strong foundation for robust data analysis. Additionally, we created specific plots to visualize the data and gain deeper insights. These initial steps are fundamental for data analysis and decision-making.