EDA-Univariate Analysis

Univariate analysis is a fundamental step in exploratory data analysis (EDA) that involves examining the distribution and properties of a single variable. This type of analysis helps to identify trends, patterns, and relationships within the data.

Example:

Suppose we have a dataset containing information about customer orders, including the order amount (in dollars). We want to perform univariate analysis on this variable to understand its characteristics.

Descriptive Statistics:

1. **Mean**: Calculate the average order amount.
2. **Median**: Find the middle value of the order amounts (i.e., the 50th percentile).
3. **Mode**: Determine the most frequently occurring order amount.

```
 import pandas as pd

# Sample dataset
data = {
    'Order Amount': [100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]
}

df = pd.DataFrame(data)

# Calculate descriptive statistics
mean_order_amount = df['Order Amount'].mean()
median_order_amount = df['Order Amount'].median()
mode_order_amount = df['Order Amount'].mode().values[0]

print("Mean Order Amount:", mean_order_amount)
print("Median Order Amount:", median_order_amount)
print("Mode Order Amount:", mode_order_amount)
```

Output:

```
 Mean Order Amount: 500.0
Median Order Amount: 500.0
Mode Order Amount: 100
```

Visualization:

To further understand the distribution of the order amounts, we can create a histogram:

```
 import matplotlib.pyplot as plt

# Create a histogram
plt.hist(df['Order Amount'], bins=10, edgecolor='black')
plt.xlabel('Order Amount (in dollars)')
plt.ylabel('Frequency')
plt.title('Distribution of Order Amounts')
plt.show()
```

This histogram reveals that the order amounts are clustered around $500, with a few outliers. The univariate analysis helps us identify patterns in the data and provides a foundation for further analysis.

Additional Measures:

Depending on the research question or problem statement, additional measures may be used in univariate analysis, such as:

1. **Standard Deviation**: A measure of the spread of the data.
2. **Interquartile Range (IQR)**: The difference between the 75th and 25th percentiles.
3. **Skewness**: A measure of the asymmetry of the distribution.

By applying these measures, you can gain a deeper understanding of your data and make informed decisions about further analysis or visualization.

Curated by Brajesh Kumar