Introduction to Univariate Analysis

Univariate analysis involves examining and summarizing data for a single variable. It focuses on the distribution, central tendency, and dispersion of the data. Common techniques include measures like mean, median, mode, standard deviation, and visualizations like histograms, boxplots, and KDE plots.

Common Use Cases

- Identifying the shape and distribution of the data.
- Detecting outliers and anomalies.
- Summarizing data with descriptive statistics.

Key Visualizations and Methods in Univariate Analysis

1. **Descriptive Statistics**:
   - mean, median, mode, standard deviation, variance.
2. **Plots**:
   - Histogram, KDE (Kernel Density Estimation), Boxplot.

Code Example: Univariate Analysis in Python

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Sample Data
data = {
   "Age": [23, 45, 31, 35, 27, 56, 32, 41, 29, 40, 60, 36, 55, 34, 42]
}
# Create DataFrame
df = pd.DataFrame(data)
# Descriptive Statistics
print("Descriptive Statistics:")
print(df["Age"].describe())

# Histogram
plt.figure(figsize=(8, 6))
sns.histplot(df["Age"], kde=True, color="blue", bins=10)
plt.title("Age Distribution")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.show()

# Boxplot
plt.figure(figsize=(6, 4))
```

```
sns.boxplot(data=df, x="Age", color="orange")
plt.title("Boxplot of Age")
plt.xlabel("Age")
plt.show()
```

Output

Descriptive Statistics

```
Descriptive Statistics:
count    15.000000
mean     39.666667
std      11.269428
min      23.000000
25%      31.000000
50%      36.000000
75%      42.000000
max      60.000000
```

1. Descriptive Statistics

- **df["Age"].describe()**:
  - Provides a summary of statistics, including:
    - **count**: Number of observations.
    - **mean**: Average value.
    - **std**: Standard deviation.
    - **min**, **max**, **25%**, **50% (median)**, and **75%** values.

2. Histogram with KDE

- **sns.histplot()**:
  - Creates a histogram to visualize the distribution of the Age variable.
  - **kde=True**: Adds a smooth Kernel Density Estimation curve.
  - **bins=10**: Divides the data into 10 bins for the histogram.

3. Boxplot

- **sns.boxplot()**:
  - Plots a boxplot to visualize the spread of the data and detect outliers.
  - Key Components:
    - **Box**: Represents the interquartile range (IQR).
    - **Whiskers**: Extend to show the range within 1.5 * IQR.
    - **Dots**: Represent outliers.

Histogram

- A plot showing the frequency of ages with a KDE curve overlay.

Boxplot

- A single box showing:
  - Median value of the dataset.
  - Range and potential outliers.

Key Insights from the Analysis

1. **Central Tendency**:
   - The average age is approximately 39.67.
2. **Spread**:
   - The standard deviation is 11.27, indicating moderate variability.
3. **Outliers**:
   - The boxplot helps visually identify any outliers in the dataset.

Key Syntax for Univariate Analysis

Descriptive Statistics

```
df["column_name"].describe()  # Summary statistics
df["column_name"].mean()      # Mean
df["column_name"].median()    # Median
df["column_name"].std()       # Standard deviation
```

Visualization

1. **Histogram**:

   ```
   sns.histplot(data=df, x="column_name", kde=True, bins=10)
   ```

2. **Boxplot**:

   ```
   sns.boxplot(data=df, x="column_name", color="orange")
   ```

Advantages of Univariate Analysis

1. Simple and quick to perform.
2. Provides foundational insights for further analysis.
3. Helps identify potential data cleaning tasks, like handling outliers.