# EDA-Categorical Variable Analysis

# Categorical Variable Analysis (CVA)

Categorical Variable Analysis is a technique used in Exploratory Data Analysis (EDA) to understand the distribution and relationships between categorical variables. CVA helps identify patterns, trends, and correlations within categorical data.

## Why CVA?

Categorical variables are common in many datasets, but they can be challenging to analyze due to their discrete nature. CVA provides a way to:

1. Understand the distribution of each categorical variable.
2. Identify relationships between categorical variables.
3. Visualize complex patterns and correlations.

## Types of Analysis

There are two primary types of analysis in CVA:

### 1. Univariate Analysis

This involves analyzing each categorical variable separately to understand its:

- Distribution (e.g., frequency, proportions)
- Central tendency (e.g., mode, median)
- Variability (e.g., standard deviation)

**Example: Univariate Analysis**

```
 import pandas as pd

# Sample dataset with a categorical variable 'Color'
data = {
    'Name': ['John', 'Mary', 'David', 'Emily', 'Michael'],
    'Color': ['Red', 'Blue', 'Green', 'Red', 'Blue']
}

df = pd.DataFrame(data)

# Univariate analysis of the 'Color' variable
print(df['Color'].value_counts())  # Frequency distribution

print(df['Color'].describe())  # Central tendency and variability
```

Output:

```
 Name: Color, dtype: int64
Red    2
Blue   2
Green  1
Name: Color, dtype: int64

count     5.000000
unique    3.000000
top       Red
freq      2.000000
dtype: object
```

## 2. Bivariate Analysis

This involves analyzing the relationship between two categorical variables to understand:

- Association (e.g., correlation, contingency table)
- Dependence (e.g., mutual exclusivity)

**Example: Bivariate Analysis**

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

# Sample dataset with two categorical variables 'Color' and 'Shape'
data = {
    'Name': ['John', 'Mary', 'David', 'Emily', 'Michael'],
    'Color': ['Red', 'Blue', 'Green', 'Red', 'Blue'],
    'Shape': ['Circle', 'Square', 'Triangle', 'Circle', 'Square']
}

df = pd.DataFrame(data)

# Bivariate analysis of the relationship between 'Color' and 'Shape'
sns.set()
plt.figure(figsize=(8, 6))
sns.countplot(x='Color', hue='Shape', data=df)
plt.title('Relationship between Color and Shape')
plt.show()
```

This code creates a count plot to visualize the association between the categorical variables.

By performing CVA, you can gain insights into the patterns and relationships within your categorical data, which can inform further analysis or modeling.

---

*Curated by Brajesh Kumar*