Data Science | 30 Days of Machine Learning | Day - 6

Educator Name: Nishant Dhote
Support Team: **+91-7880-113-112**

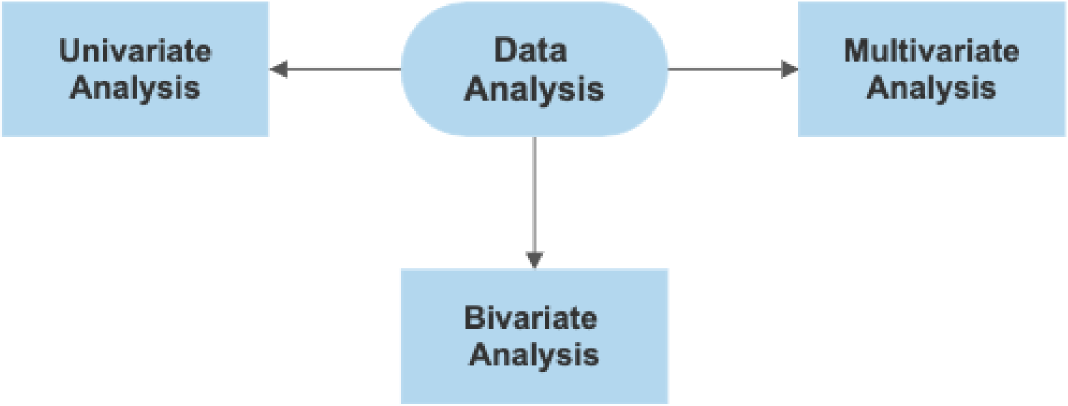
----Today Topics | Day 06----

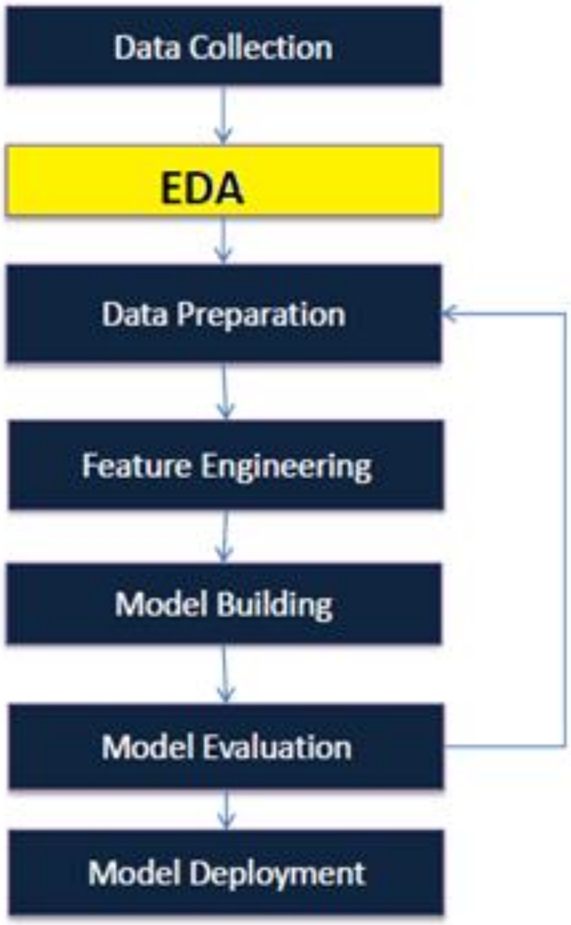
- **EDA: Exploratory Data Analysis**

- **EDA Univariate Analysis**
- **EDA Bivariate Analysis**
- **EDA Multivariate Analysis**
- **# How We Understand the Data?**
- **Feature Engineering**

Exploratory data analysis (EDA) is an approach to analysing data sets to summarize their main characteristics, often with visual methods. Before applying any ML algorithms in data, we need to understand the data which we are going to follow. Without data understanding there will be a possibility of ML model failure. The understanding of data is nothing but this Exploratory Data Analysis (EDA).

It is always better to explore each data set using multiple exploratory techniques and compare the results. Once the data set is fully understood, it is quite possible that data scientist will have to go back to data collection and cleansing phases in order to transform the data set according to the desired business outcomes. The goal of this step is to become confident that the data set is ready to be used in a machine learning algorithm.





Data Collection

EDA

Data Preparation

Feature Engineering

Model Building

Model Evaluation

Model Deployment

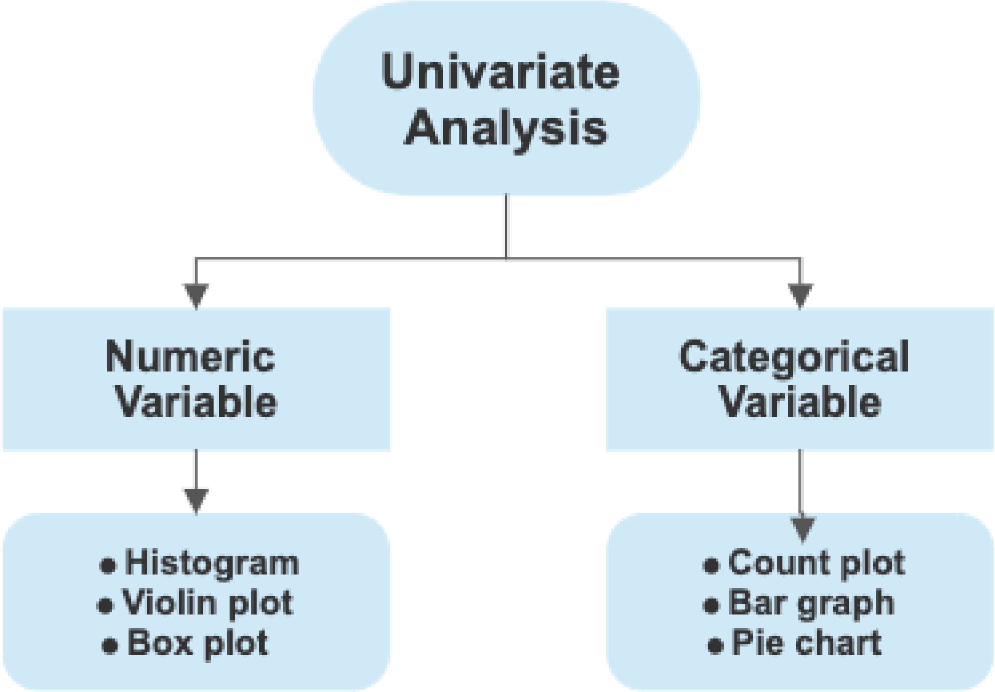
TYPES OF EXPLORATORY DATA ANALYSIS:

1. Univariate Analysis
2. Bivariate Analysis
3. Multivariate Analysis

Univariate EDA involves looking at a single variable at a time. Univariate EDA can help you understand the data distribution and identify any outliers.

Bivariate EDA involves looking at two variables at a time. Bivariate EDA can help you understand the relationship between two variables and identify any patterns that might exist.

Multivariate EDA involves looking at three or more variables at a time. Multivariate EDA can help you understand the relationships between several variables and identify any complex patterns or outliers that might exist.



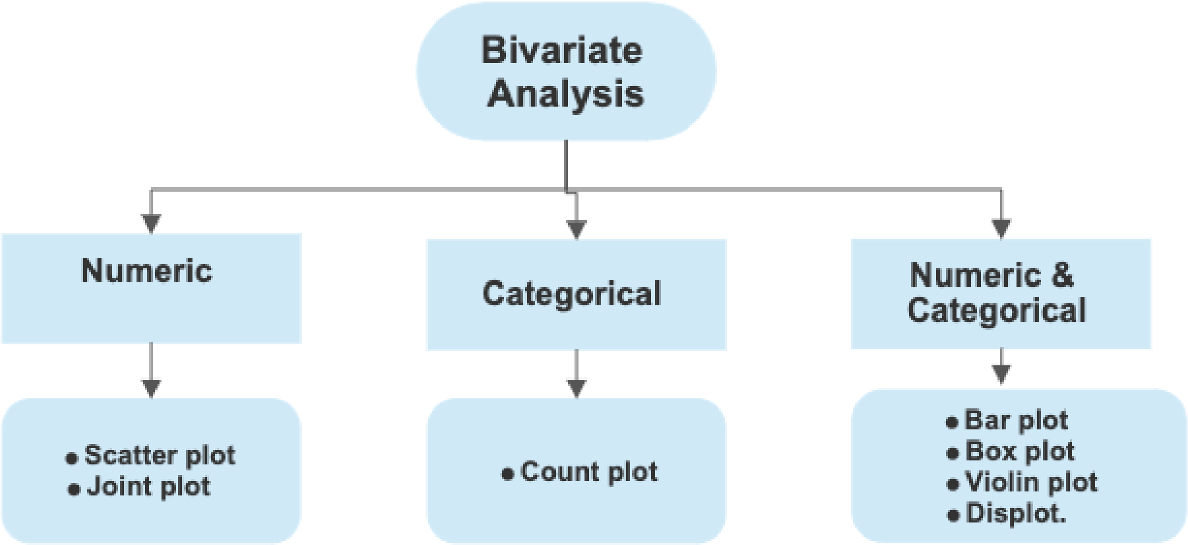
Univariate Analysis

Numeric Variable

- Histogram
- Violin plot
- Box plot

Categorical Variable

- Count plot
- Bar graph
- Pie chart



Bivariate Analysis

Numeric

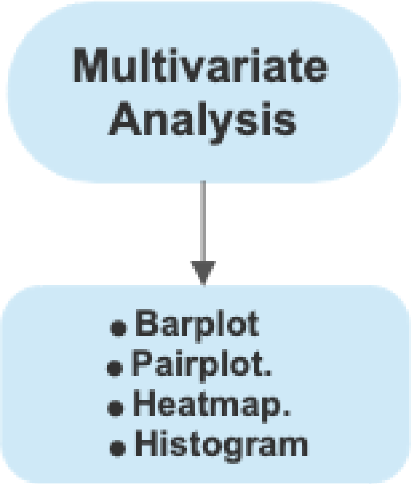
- Scatter plot
- Joint plot

Categorical

- Count plot

Numeric & Categorical

- Bar plot
- Box plot
- Violin plot
- Displot.



Multivariate Analysis

- Barplot
- Pairplot.
- Heatmap.
- Histogram

Dataset Link Kaggle: https://www.kaggle.com/competitions/titanic

GitHub Link: https://github.com/TheiScale/30_Days_Machine_Learning/tree/main/Day%206%20ML

#Import Library and Dataset

```
import pandas as pd
df = pd.read_csv('train.csv')
```

1. **Data Size:**
   - What is the magnitude of the dataset in terms of its rows and columns?
   - How big is the data?

```
df.shape
```

2. **Data Visualization:**
   - Can you provide a visual representation or description of the dataset's structure?
   - How does the data look like?

```
df.sample(5)
```

3. **Column Data Varieties:**
   - What variations in data types exist among the columns in the dataset?
   - What is the data types of cols?

```
df.info()
```

4. **Existence of Missing Values:**
   - Are there any absent values within the dataset? If so, which columns have them, and how many are missing?
   - Are there any missing values?

```
df.isnull().sum()
```

5. **Mathematical Representation:**
   - How is the data characterized mathematically, including measures such as mean, median, and standard deviation?

- How does the data look mathematically?

```
df.describe()
```

6. **Identification of Redundancies:**
   - Is there any repetition or duplication of values within the dataset?
   - Are there duplicate values?

```
df.duplicated().sum()
```

7. **Correlation Examination:**
   - What level of correlation exists between different columns in the dataset?
   - How is the correlation between columns?

```
df.corr(numeric_only = True)
```

1. EDA in Univariate Analysis:

```
<Start Coding>
```

#Import Library

```
import pandas as pd
import seaborn as sns
```

#Define Data Frame as “df”

```
df = pd.read_csv('train.csv')

---
df.head()
```

#1. Categorical Data

#a. Countplot

```
sns.countplot(x='Survived', data=df)
df['Survived'].value_counts()

OR--

sns.countplot(x='Pclass', data=df)
df['Pclass'].value_counts()
```

```
OR--

sns.countplot(x='Embarked', data=df)
df['Embarked'].value_counts()
```

#b. Piechart

```
df['Sex'].value_counts().plot(kind='pie',autopct='%.2f')
```

2. NUMERICAL DATA

a. Histogram

```
 import matplotlib.pyplot as plt
 plt.hist(df['Age'],bins=5)
```

b. Distplot / Histplot

```
sns.distplot(df['Age'])
```

c. Boxplot

Article 1: click here to read more boxplot or Article 2 click here

```
sns.boxplot(df['Age'])
```

df['Age'].min()

df['Age'].max()

df['Age'].mean()

df['Age'].skew()

Recommended Datasets for practice:

Air Passengers- https://www.kaggle.com/code/chandrimad31/flight-passenger-satisfaction-eda-and-prediction

2.EDA Bivariate Analysis <Day 7>

3.EDA Multivariate Analysis <Day 7>

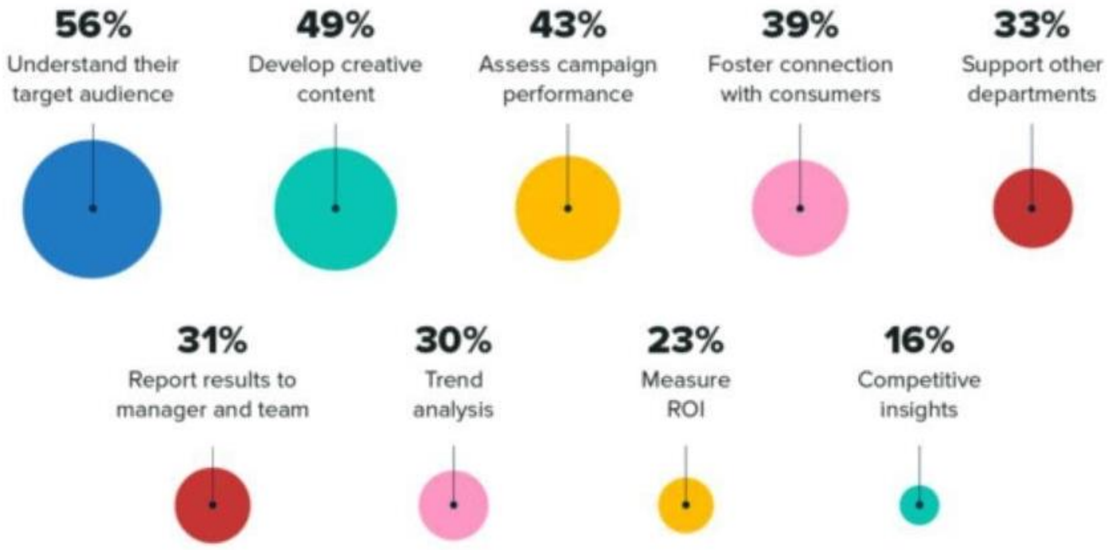
Data Story Telling (Day 6): Curious Data Minds

How is data science used in media and entertainment industry?

Read Blogs: https://hevodata.com/learn/data-analytics-in-media/#:~:text=Here%20are%20a%20couple%20of,in%20listening%20or%20viewing%20habits.

https://www.polestarllp.com/blog/big-data-analytics-media-and-the-entertainment-industry

How We use Social Data


Understand their target audience
Develop creative content
Assess campaign performance
Foster connection with consumers
Support other departments
Report results to manager and team
Trend analysis
Measure ROI
Competitive insights