Lab 2: Exploratory Data Analysis (EDA)

Objective

The objective of this lab is to perform an in-depth **Exploratory Data Analysis (EDA)** on the Titanic dataset to:

1. Understand the structure and basic features of the dataset.
2. Perform **Univariate Analysis** to study individual variables.
3. Conduct **Bivariate Analysis** to examine relationships between two variables.
4. Explore **Multivariate Analysis** to identify patterns involving multiple variables.
5. Visualize data effectively using various plots and derive actionable insights.

By the end of this lab, you will have a thorough understanding of the dataset and its underlying patterns.

Libraries Installation

Before starting, ensure that the required libraries are installed in your Python environment. Run the following command:

pip install pandas numpy matplotlib seaborn

Explanation of Libraries

1. **pandas**: For data manipulation and analysis.
2. **numpy**: For numerical computations.
3. **matplotlib**: For creating static visualizations.
4. **seaborn**: For advanced and aesthetically pleasing statistical plots.

1. Boxplot

- **Purpose**:
  - Identifies **outliers** (extreme data points).
  - Summarizes the distribution of data based on the minimum, first quartile (Q1), median, third quartile (Q3), and maximum.
- **When to Use**:
  - To visualize and compare the spread of numerical data across categories.

2. Histogram

- **Purpose**:
  - Displays the **frequency distribution** of a single numerical variable.
  - Helps understand data spread, shape, skewness, and peaks (modes).
- **When to Use**:
  - To observe how frequently values occur within specified ranges (bins).

3. Distplot/Histplot

- **Purpose**:
  - Combines a histogram with a **KDE (Kernel Density Estimate)** to show the probability density of a variable.
- **When to Use**:
  - To understand both the frequency and density of a numerical variable.

4. Heatmap

- **Purpose**:
  - Visualizes the **correlation** between numerical variables using color intensity.
  - Shows which variables are positively or negatively correlated.

- **When to Use**:
  - To identify patterns or relationships between variables.

5. Pie Chart

- **Purpose**:
  - Displays proportions or percentages of categories as slices of a circle.
- **When to Use**:
  - To visualize the **composition** of a single categorical variable.

6. Countplot

- **Purpose**:
  - Displays the **frequency** of each category in a categorical variable.
- **When to Use**:
  - To count and compare occurrences of categories.

7. Scatterplot

- **Purpose**:
  - Plots individual data points to show the **relationship** between two numerical variables.
  - Identifies **trends, clusters**, or **outliers**.
- **When to Use**:
  - To examine how one variable changes with another.

8. Pairplot

- **Purpose**:
  - Displays pairwise scatterplots for all numerical variables in the dataset.
  - Useful for detecting patterns, correlations, and outliers across multiple features.
- **When to Use**:
  - To perform a comprehensive comparison of all numerical variables.

9. Bar Chart

- **Purpose**:
  - Visualizes data for **categorical variables** as bars, where the height represents the frequency or value.
- **When to Use**:
  - To compare the size or count of different categories.

10. Line Plot

- **Purpose**:
  - Shows trends over time or a sequence by connecting data points with a line.
- **When to Use**:
  - To analyze time-series data or sequential patterns.

11. Violin Plot

- **Purpose**:
  - Combines a boxplot and KDE to show the distribution of data, including its density and spread.
- **When to Use**:
  - To understand how data is distributed, especially when comparing multiple groups.

12. KDE Plot

- **Purpose**:
  - Represents the **probability density function** of a variable, showing where data points are concentrated.
- **When to Use**:
  - To smooth out the data and observe the overall distribution.

13. Jointplot

- **Purpose**:
  - Combines scatterplots and histograms (or KDEs) to display the relationship between two numerical variables, along with their individual distributions.
- **When to Use**:
  - To analyze the relationship between two variables while examining their marginal distributions.

14. Stacked Bar Chart

- **Purpose**:
    - Displays proportions of categories within each group in a stacked format.
- **When to Use**:
    - To show both the total and the composition of groups.

Dataset Overview

The dataset contains the following columns:

Column Name	Description
PassengerId	Unique ID for each passenger
Survived	Survival status (0 = No, 1 = Yes)
Pclass	Passenger class (1 = First, 2 = Second, 3 = Third)
Name	Full name of the passenger
Sex	Gender of the passenger
Age	Age of the passenger
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Ticket	Ticket number
Fare	Ticket fare paid
Cabin	Cabin number
Embarked	Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Basic Exploration

Display Basic Dataset Information

```
# Import necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load dataset
df = pd.read_csv('train.csv')

# Display the first 5 rows
print(df.head())

# Display dataset information
print(df.info())

# Display descriptive statistics
print(df.describe())
```

Explanation:

1. **head()**: Displays the first five rows of the dataset for a quick overview.
2. **info()**: Provides information about column data types and non-null counts, helping identify missing values.
3. **describe()**: Summarizes numerical columns with statistics like mean, median, standard deviation, min, and max values.

Basic Plotting

1. Boxplot

```
sns.boxplot(x=df['Fare'])
plt.title("Boxplot of Fare")
plt.xlabel("Fare")
plt.show()
```

Description:

- **Boxplot** helps detect **outliers** (data points that are significantly different from others) and shows the distribution's spread, median, and quartiles.

2. Distplot

```
sns.histplot(df['Age'], kde=True, bins=20, color='skyblue')
plt.title("Age Distribution with KDE")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.show()
```

Description:

- Combines a **histogram** and a **KDE curve** to visualize the distribution of the numerical data.
- Identifies skewness, modality, and overall spread of the data.

3. Heatmap

```
sns.heatmap(df.corr(), annot=True, cmap='coolwarm', fmt='.2f')
plt.title("Correlation Heatmap")
plt.show()
```

Description:

- **Heatmap** shows pairwise correlations between numerical variables.
- Highlights strong positive or negative relationships for further analysis.

4. Histogram

```
df['Age'].plot(kind='hist', bins=20, color='green', alpha=0.7)
plt.title("Histogram of Age")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

Description:

- Visualizes the frequency distribution of **Age**.
- Useful for identifying common age ranges and trends in passenger demographics.

5. Pie Chart

```
survival_counts = df['Survived'].value_counts()
labels = ['Did Not Survive', 'Survived']
plt.pie(survival_counts, labels=labels, autopct='%1.1f%%', startangle=140, colors=['lightcoral', 'skyblue'])
plt.title("Survival Proportion")
plt.show()
```

Description:

- Highlights proportions or percentages of passengers who survived versus those who did not.

Section 1: Univariate Analysis

Numerical Data

```
sns.histplot(df['Fare'], kde=True, color='blue')
plt.title("Distribution of Fare")
plt.show()

sns.boxplot(x=df['Age'])
plt.title("Boxplot of Age")
plt.show()
```

Description:

- **Histograms** visualize the frequency distribution, while **boxplots** detect outliers and show data spread.

Categorical Data

```
sns.countplot(x='Pclass', data=df, palette='viridis')
plt.title("Passenger Count by Class")
plt.xlabel("Passenger Class")
plt.ylabel("Count")
plt.show()

df['Embarked'].value_counts().plot(kind='bar', color='orange')
plt.title("Embarked Port Count")
plt.xlabel("Port")
plt.ylabel("Count")
plt.show()
```

Description:

- **Countplots** display the frequency of categorical variables like passenger class.
- **Bar charts** highlight the distribution of values across categories.

Section 2: Bivariate Analysis

Numerical-Numerical

```
sns.scatterplot(x='Age', y='Fare', data=df)
plt.title("Scatterplot of Age vs Fare")
plt.xlabel("Age")
plt.ylabel("Fare")
plt.show()
```

Description:

- **Scatterplots** highlight relationships between two numerical features.

Categorical-Categorical

```
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title("Survival Count by Passenger Class")
plt.xlabel("Passenger Class")
plt.ylabel("Count")
plt.show()
```

Description:

- **Grouped countplots compare survival counts across classes.**

Numerical-Categorical

```
sns.boxplot(x='Survived', y='Age', data=df)
plt.title("Boxplot of Age by Survival Status")
plt.xlabel("Survived")
plt.ylabel("Age")
plt.show()
```

Description:

- **Boxplots** compare distributions (e.g., Age) across survival categories.

Section 3: Multivariate Analysis

```
sns.pairplot(df[['Age', 'Fare', 'Pclass', 'Survived']], hue='Survived', palette='coolwarm')
plt.suptitle("Pairplot of Selected Features", y=1.02)
plt.show()
```

Description:

- **Pairplots** allow simultaneous comparison of multiple features to detect relationships and patterns.

Task 1: Dataset Exploration

1. **Question:** What is the structure of the Titanic dataset?

    - Use basic functions like head(), info(), and describe() to understand the dataset's structure, data types, and summary statistics.
    - Identify missing values and duplicates in the dataset.

Task 2: Numerical Data Analysis

2. **Question:** What are the key statistical properties and distributions of the numerical columns?

    - Analyze columns such as Age, Fare, and Parch using summary statistics and visualizations.
    - Plot histograms, boxplots, and KDE plots to understand distributions and detect outliers.

Task 3: Categorical Data Analysis

3. **Question:** What is the frequency distribution of categorical columns?

    - Explore Pclass, Sex, Embarked, and Survived using value_counts().
    - Visualize the distributions using countplots and pie charts.

Task 4: Relationship Between Variables (Bivariate Analysis)

4. **Question:** How are variables related to each other?

    Analyze the relationship between:

    - **Numerical-Numerical:** Age vs Fare (scatter plot and correlation).
    - **Categorical-Categorical:** Pclass vs Survived (grouped bar chart).
    - **Numerical-Categorical:** Age distribution across survival status (boxplot).

Task 5: Multivariate Analysis

5. **Question:** How do multiple variables interact with each other?

   - Perform a pairwise analysis on selected columns such as Age, Fare, Pclass, and Survived using pairplots.
   - Visualize overall correlations using a heatmap.