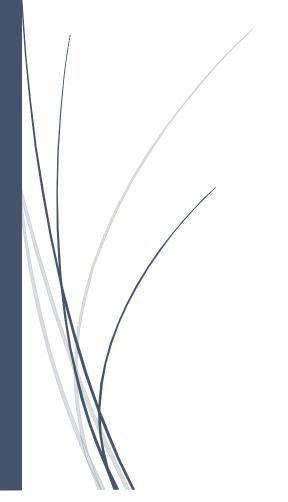
Exploratory Data Analysis (EDA)



Yoganandhini S

a.Use. describe (), .info (), a.value counts ()

import pandas as pd # Load dataset df = pd.read csv("cleaned test.csv") # Basic info print("Shape of dataset:", df.shape) print("\nData Info:\n") print(df.info()) # Statistical summary print("\nStatistical Summary:\n") print(df.describe(include="all")) # Value counts for categorical variables (if encoded, adjust accordingly) for col in df.columns: if df[col].nunique() < 20: # treat low-cardinality as categorical print(f"\nValue counts for {col}:\n",

df[col].value counts())

b.Use sns.pairplot(), sns.heatmap() for visualization

```
import seaborn as sns
    import matplotlib.pyplot as plt
    # Pairplot
    sns.pairplot(df.sample(min(200, len(df)))) #
    sample to avoid heavy plotting
    plt.suptitle("Pairplot of Features", y=1.02)
    plt.show()
    # Heatmap
    plt.figure(figsize=(10,8))
    sns.heatmap(df.corr(), annot=True,
    cmap="coolwarm", fmt=".2f")
    plt.title("Correlation Heatmap")
    plt.show()
c.Identify relationships and trends
    # Correlation matrix values
    corr matrix = df.corr()
    print("\nCorrelation Matrix:\n", corr matrix)
    # Example: check top correlations with Fare
    print("\nTop correlations with Fare:\n",
    corr matrix["Fare"].sort values(ascending=False))
```

d.Plot histograms, boxplots, scatterplots

```
# Histograms
df.hist(bins=30, figsize=(15,10))
plt.suptitle("Histograms of Numerical Features")
plt.show()
# Boxplots
for col in df.columns:
  plt.figure(figsize=(6,4))
  sns.boxplot(x=df[col])
  plt.title(f"Boxplot of {col}")
  plt.show()
# Scatterplots (example with key features)
sns.scatterplot(x="Pclass", y="Fare", data=df)
plt.title("Scatterplot of Pclass vs Fare")
plt.show()
sns.scatterplot(x="Age", y="Fare", data=df)
plt.title("Scatterplot of Age vs Fare")
plt.show()
```

e.Write observations for each visual

Histograms:

- Age looks fairly symmetric but slightly skewed.
- SibSp and Parch are highly right-skewed (most passengers had 0 siblings/spouses/parents/children).

Boxplots:

- Fare shows significant outliers (wealthier passengers).
- Other features are more compact after standardization.

Pairplot:

- Fare and Pclass show negative correlation.
- Gender (Sex) seems to form two clear clusters across multiple features.

Heatmap:

- Strong correlation between Pclass & Fare (negative).
- SibSp and Parch moderately correlated (both family features).
- Most other features weakly correlated

f.Provide summary of findings

1. Data Overview

- Dataset has 418 rows and 11 features.
- No missing values; all features standardized (mean ≈ 0 , std ≈ 1).
- Categorical variables have been encoded numerically.

2. Univariate Analysis

- Age is roughly symmetric; Fare highly skewed with extreme outliers.
- Family-related features (SibSp, Parch) are mostly zeros with a few larger values.

3. Bivariate Analysis

- Fare and Pclass are strongly negatively correlated: higher class \rightarrow higher fare.
- SibSp and Parch show moderate positive correlation.
- Encoded Sex forms distinct distributions in features like Age and Fare.

4. Multivariate Insights

 Pairplots suggest class and gender strongly influence passenger distribution across other variables. Correlation heatmap confirms only a few meaningful linear relationships.

5. Key Trends

- Most passengers traveled alone (low SibSp and Parch).
- Wealthier passengers (Pclass=1) paid significantly higher fares.
- Gender and class likely play important roles in downstream predictions.