TITANIC DATASET ANALYSIS REPORT

INTRODUCTION

This report explores the Titanic passenger dataset using statistical and visual techniques in Python (Pandas, Matplotlib, and Seaborn). The aim is to identify trends, relationships, and key factors correlating with survival on the Titanic, with all code steps and graphics thoroughly explained.

1. DATA LOADING AND BASIC EXPLORATION

Code:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv("train.csv")
```

Explanation:

This loads the CSV data file into a Pandas DataFrame called df. Pandas is used for data manipulation and analysis.

Data Information and Summary:

Code:

```
print(df.info())
print(df.describe(include="all"))
```

Explanation:

- .info() displays row/column counts, datatypes, and missing values.
- .describe(include="all") gives summary statistics for both numeric and categorical columns: mean, min, max, count, unique values, etc.

Findings:

- 891 total records, 12 columns (e.g., Survived, Pclass, Sex, Age, Fare).
- Missing values in columns like Age and Cabin.
- Survived is the target variable (1 = survived, 0 = did not).

2. VALUE COUNTS FOR CATEGORICAL COLUMNS

Code:

```
for col in ["Sex", "Embarked", "Pclass"]:
print(df[col].value_counts())
```

Explanation:

Shows how many passengers fall into each category (e.g., number of males/females, embarkation point, passenger class).

Findings:

- More males than females.
- Most passengers embarked from Southampton ('S').
- Majority were third-class passengers.

3. MISSING VALUE TREATMENT

Code:

```
df["Age"] = df.groupby("Pclass")["Age"].transform(lambda
x:x.fillna(x.median()))
df["Embarked"] = df["Embarked"].fillna(df["Embarked"].mode()[0])
df["Cabin"] = df["Cabin"].fillna("Unknown")
print(df.isnull().sum())
```

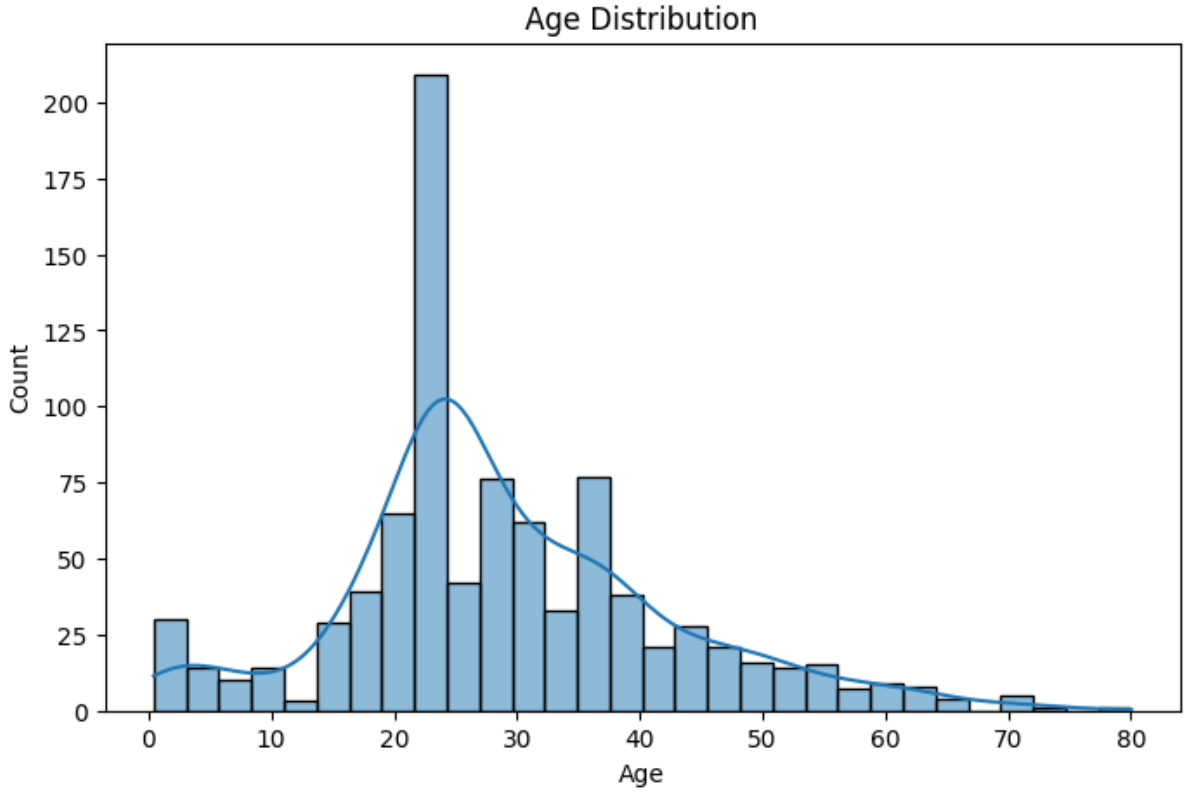
Explanation:

- Fills missing Age with median for each Pclass (to reflect different class demographics).
- Fills missing Embarked value with the most common port.
- Fills missing Cabin entries with 'Unknown'.
- Prints remaining missing values to verify completeness.

4. VISUAL EXPLORATION AND EXPLANATION

a. age distribution (histogram)

```
import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(8,5))
sns.histplot(df["Age"], bins=30, kde=True)
plt.title("Age Distribution")
plt.show()
```

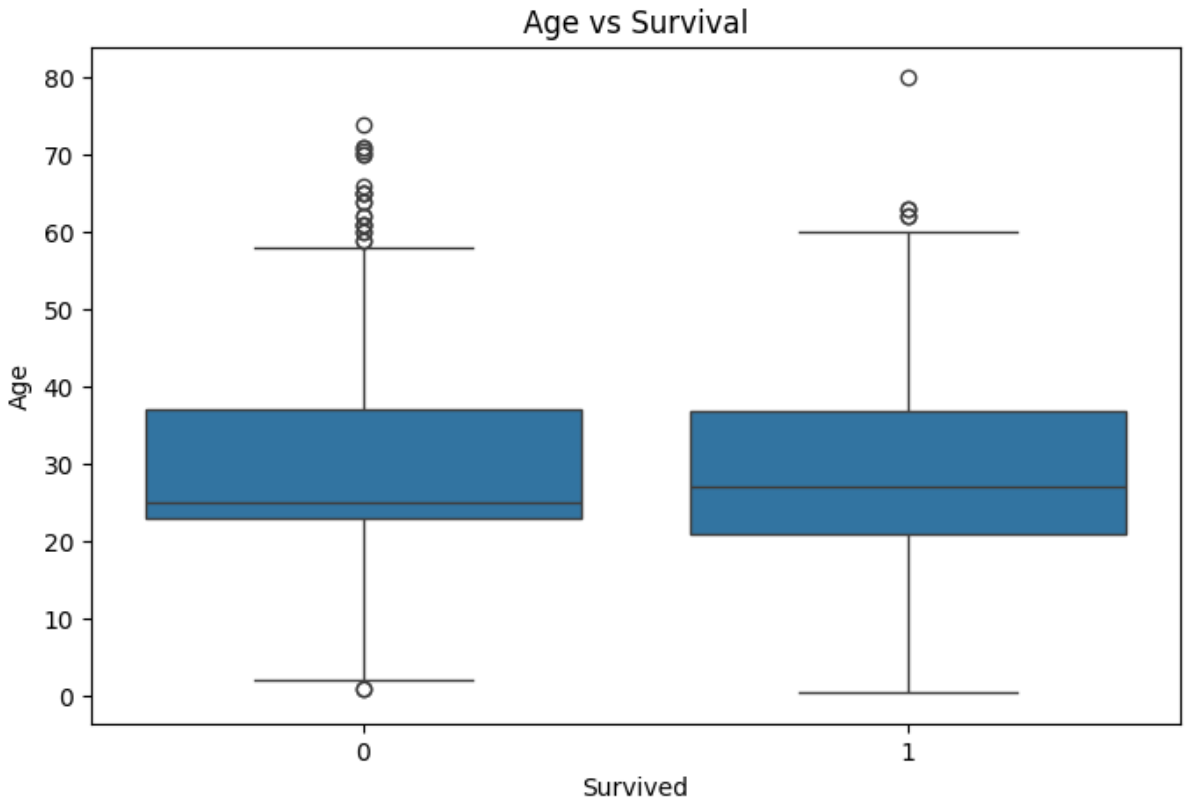


Explanation & Insights:

Shows the spread of passenger ages. The distribution is right-skewed, with most passengers between 20–40 years old.

b. Age vs. Survival (Boxplot)

```
plt.figure(figsize=(8,5))
sns.boxplot(x="Survived", y="Age", data=df)
plt.title("Age vs Survival")
plt.show()
```



Explanation & Insights:

Displays median and spread of ages for survivors and non-survivors. Median ages are fairly close, but slightly more young survivors, with more outliers among the older who did not survive.

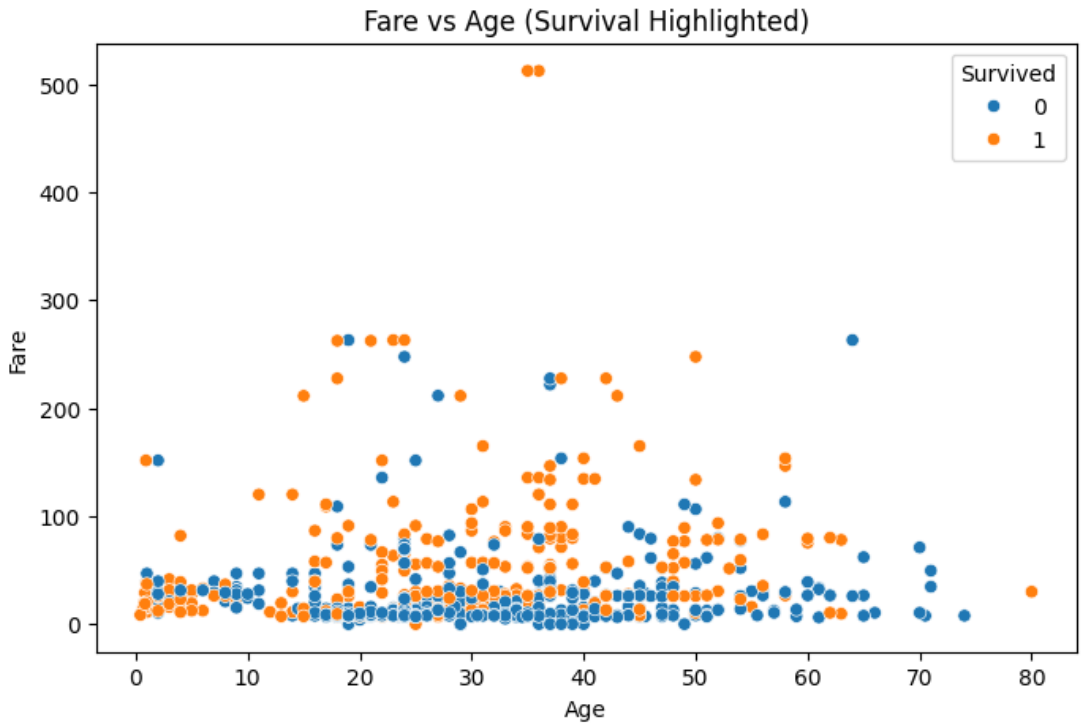
c. Survival Count by Sex (Countplot)

```
plt.figure(figsize=(6,4))
sns.countplot(x="Sex", hue="Survived", data=df)
plt.title("Survival Count by Sex")
plt.show()
```

Explanation & Insights:

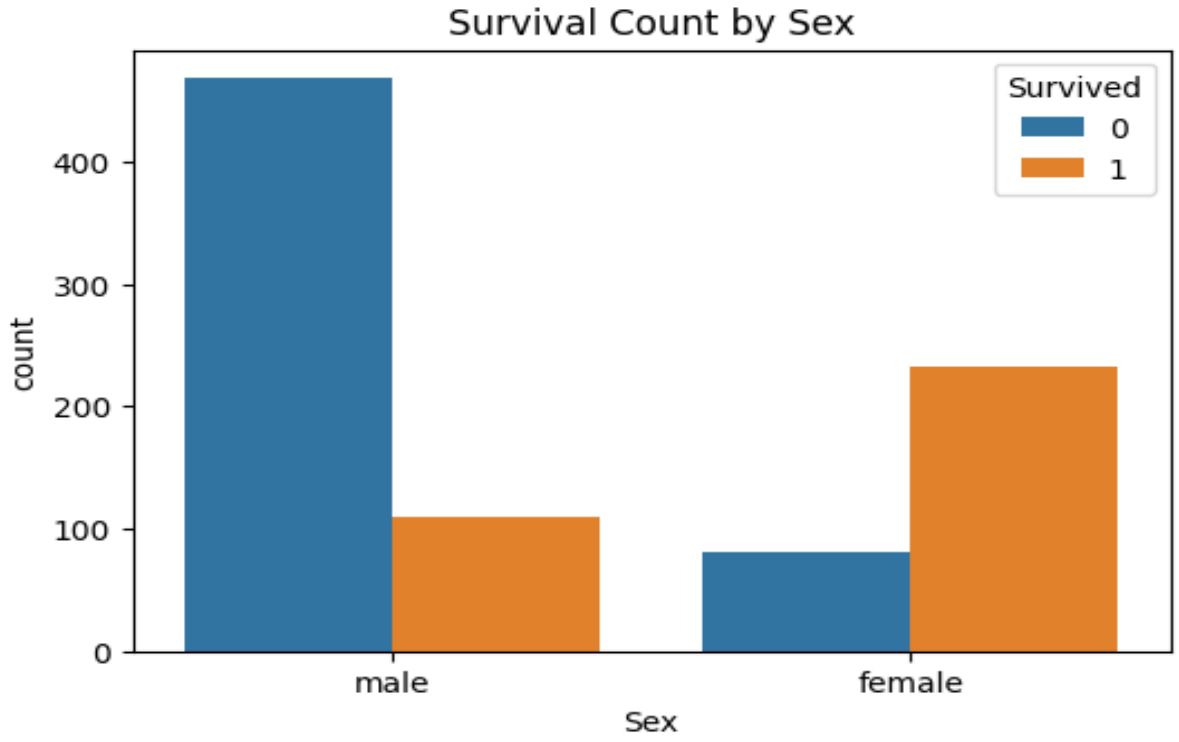
Clearly shows women survived at a much higher rate than men, reflecting the "women and children first" policy during evacuation.

d.



Survival Count by Passenger Class (Countplot)

plt.figure(figsize=(6,4))
sns.countplot(x="Pclass", hue="Survived", data=df)
plt.title("Survival Count by Passenger Class")
plt.show()

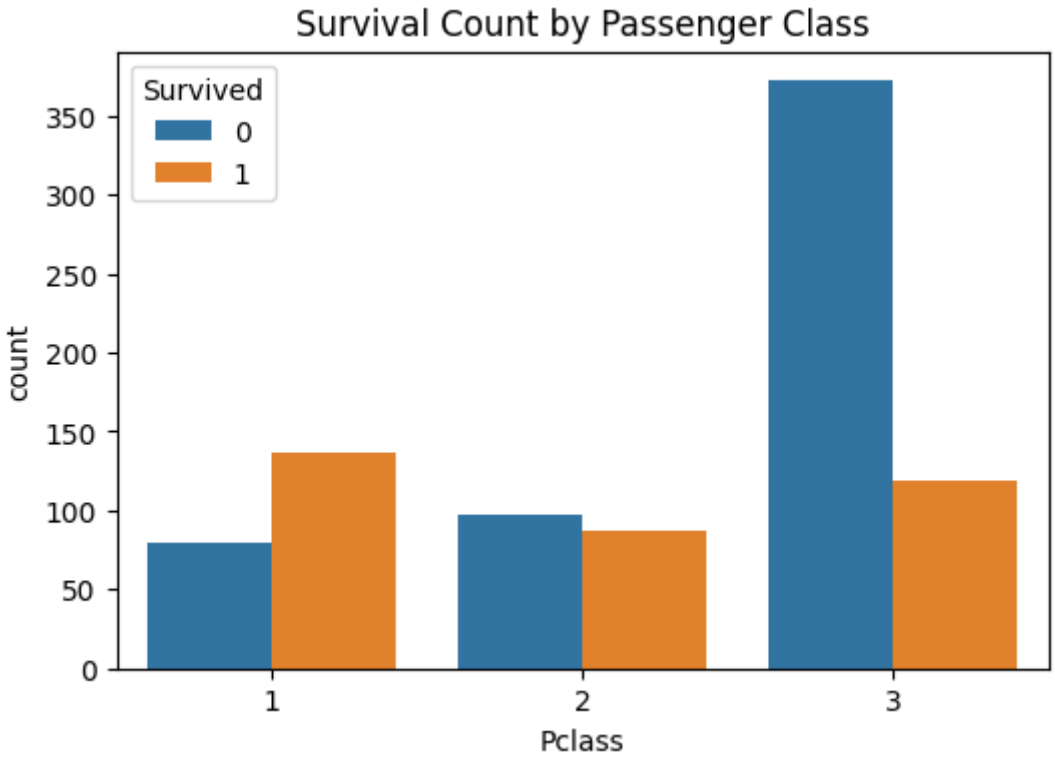


Explanation & Insights:

First-class passengers had the highest survival rates, followed by second, then third class. Highlights the role of socio-economic status in survival chances.

e. Fare vs Age by Survival (Scatterplot)

```
plt.figure(figsize=(8,5))
sns.scatterplot(x="Age", y="Fare", hue="Survived", data=df)
plt.title("Fare vs Age (Survival Highlighted)")
plt.show()
```

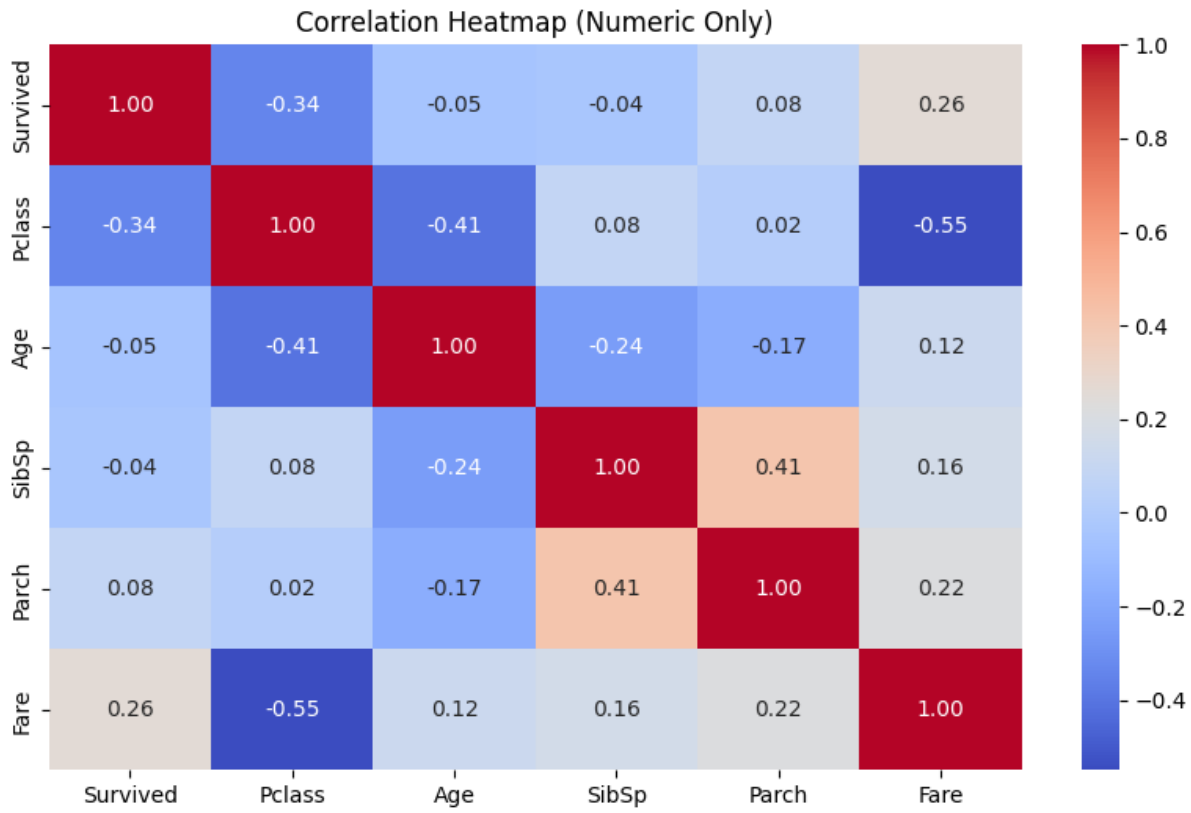


Explanation & Insights:

Most survivors paid higher fares (likely first class or wealthy), and the very young also have higher survival, irrespective of fare.

f. Correlation Heatmap

```
numeric_cols = ['Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare']
corr = df[numeric_cols].corr()
plt.figure(figsize=(10,6))
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap (Numeric Only)')
plt.show()
```



Explanation & Insights:

- Survived is positively associated with higher fare and negatively with Pclass (lower class number = higher status).
- Strong negative correlation between Fare and Pclass.
- Moderate positive correlation between number of siblings/spouses (SibSp) and number of parents/children (Parch) aboard.

FINDINGS AND INSIGHTS

- Sex: Females had a significantly higher chance of survival.
- Pclass: First class survival rates far higher than third.
- Age: Slight edge for children and young adults.
- Fare: Wealthier passengers (higher fare, higher class) had greater survival odds.
- Family: Some correlation, but less direct impact on survival.
- Social Status: Titanic tragedy mirrored early 20th-century social hierarchies.

CONCLUSION

By sequentially analyzing, cleaning, and visualizing the data, this report has shown that survival on the Titanic was most affected by gender, class, and socio-economic status. Visual explorations directly supported and illustrated the statistical findings.

This workflow, and its combination of code, analysis, and visualization, is fully reproducible, so further features and scenarios may be investigated in future work.

REFERENCES:

- Kaggle Titanic Dataset
- Python libraries: Pandas, Matplotlib, Seaborn