

```
# task2.py

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset
df = pd.read_csv("TRAIN1.csv")

print("==== RAW DATA SAMPLE ====")
print(df.head())

print("\n==== DATASET INFO ====")
print(df.info())

print("\n==== MISSING VALUES ====")
print(df.isnull().sum())

# === Clean missing data ===

# Fill missing Age and Embarked
df['Age'] = df['Age'].fillna(df['Age'].median())
df['Embarked'] = df['Embarked'].fillna(df['Embarked'].mode()[0])

# Drop Cabin due to high nulls
df = df.drop(columns=['Cabin'])

# === Plotting ===

# 1. Survival Count
sns.countplot(x='Survived', data=df)
```

```
plt.title("Survival Count")
plt.savefig("survival_count.png")
plt.show()
```

2. Survival by Gender

```
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title("Survival by Gender")
plt.savefig("survival_by_gender.png")
plt.show()
```

3. Survival by Class

```
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title("Survival by Passenger Class")
plt.savefig("survival_by_class.png")
plt.show()
```

4. Age Distribution

```
sns.histplot(df['Age'], kde=True)
plt.title("Age Distribution")
plt.savefig("age_distribution.png")
plt.show()
```

5. Correlation Heatmap (NUMERIC COLUMNS ONLY)

```
numeric_df = df.select_dtypes(include=['int64', 'float64'])
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.savefig("correlation_heatmap.png")
plt.show()
```