Task 2: Exploratory Data Analysis (EDA)

Problem Statement

Perform exploratory data analysis on a chosen dataset.

Steps Completed

1. **Data Loading**: The `cleaned_titanic.csv` dataset (output from Task 1) was loaded using Pandas.
2. **Data Overview**: Basic information about the dataset, including data types, non-null values, and memory usage, was displayed using `df.info()`. Descriptive statistics were generated using `df.describe()` to understand the central tendency, dispersion, and shape of the dataset's distribution.
3. **Missing Values Check**: The number of missing values for each column was verified using `df.isnull().sum()`.
4. **Visualizations**: Several visualizations were created using Matplotlib and Seaborn to gain insights into the data:
    - **Distribution of Age and Fare**: Histograms with KDE (Kernel Density Estimate) plots were generated to show the distribution of these numerical features.
    - **Survival Rate by Sex and Pclass**: Bar plots were used to visualize the survival rates based on gender and passenger class, highlighting potential relationships between these features and survival.
    - **Correlation Heatmap**: A heatmap was generated to display the correlation matrix of numerical features, helping to identify strong relationships between variables.

Code Implementation

The EDA steps are implemented in the `eda_titanic.py` script.

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

def perform_eda(input_filepath):
    df = pd.read_csv(input_filepath)
```

```
    print("\n--- Dataset Info ---")
    df.info()

    print("\n--- Dataset Description ---")
    print(df.describe())

    print("\n--- Missing Values ---")
    print(df.isnull().sum())

    # Visualizations
    plt.figure(figsize=(12, 6))

    # Distribution of Age
    plt.subplot(1, 2, 1)
    sns.histplot(df["Age"], kde=True)
    plt.title("Distribution of Age")

    # Distribution of Fare
    plt.subplot(1, 2, 2)
    sns.histplot(df["Fare"], kde=True)
    plt.title("Distribution of Fare")
    plt.tight_layout()
    plt.savefig("age_fare_distribution.png")
    plt.show()

    plt.figure(figsize=(12, 6))
    # Survival by Sex
    plt.subplot(1, 2, 1)
    sns.barplot(x="Sex_male", y="Survived", data=df)
    plt.title("Survival Rate by Sex (0=Female, 1=Male)")

    # Survival by Pclass
    plt.subplot(1, 2, 2)
    sns.barplot(x="Pclass", y="Survived", data=df)
    plt.title("Survival Rate by Pclass")
    plt.tight_layout()
    plt.savefig("survival_by_sex_pclass.png")
    plt.show()

    # Correlation Heatmap
    plt.figure(figsize=(10, 8))
    sns.heatmap(df.corr(numeric_only=True), annot=True,
cmap="coolwarm", fmt=".2f")
    plt.title("Correlation Heatmap")
    plt.savefig("correlation_heatmap.png")
    plt.show()

    print("\n--- Insights ---")
    print("1. Age and Fare distributions are shown.")
    print("2. Survival rates by Sex and Pclass are visualized.")
    print("3. A correlation heatmap provides insights into
feature relationships.")
```

```
if __name__ == "__main__":
    input_file = "../task1_data_preprocessing/
cleaned_titanic.csv"
    perform_eda(input_file)
```

Output

Upon execution, the script prints dataset information, descriptive statistics, and missing values. It also generates three image files: * `age_fare_distribution.png`: Histograms showing the distribution of Age and Fare. * `survival_by_sex_pclass.png`: Bar plots showing survival rates by Sex and Pclass. * `correlation_heatmap.png`: A heatmap illustrating the correlation between numerical features.

Console output includes:

```
 --- Dataset Info ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 887 entries, 0 to 886
Data columns (total 8 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Survived                 887 non-null    int64
 1   Pclass                   887 non-null    int64
 2   Name                     887 non-null    object
 3   Age                      887 non-null    float64
 4   Siblings/Spouses Aboard  887 non-null    int64
 5   Parents/Children Aboard  887 non-null    int64
 6   Fare                     887 non-null    float64
 7   Sex_male                 887 non-null    bool
dtypes: bool(1), float64(2), int64(4), object(1)
memory usage: 49.5+ KB
--- Dataset Description ---
         Survived      Pclass  ...  Parents/Children
Aboard        Fare
count  887.000000  887.000000  ...               887.000000
887.000000
mean     0.385569    2.305524  ...                 0.383315
0.063056
std      0.487004    0.836662  ...                 0.807466
0.097168
min      0.000000    1.000000  ...                 0.000000
0.000000
25%      0.000000    2.000000  ...                 0.000000
0.015469
50%      0.000000    3.000000  ...                 0.000000
```

```
0.028213
75%      1.000000    3.000000  ...                0.000000
0.060776
max      1.000000    3.000000  ...                6.000000
1.000000
[8 rows x 6 columns]
--- Missing Values ---
Survived                    0
Pclass                      0
Name                        0
Age                         0
Siblings/Spouses Aboard     0
Parents/Children Aboard     0
Fare                        0
Sex_male                    0
dtype: int64
--- Insights ---
1. Age and Fare distributions are shown.
2. Survival rates by Sex and Pclass are visualized.
3. A correlation heatmap provides insights into feature
relationships.
```

These outputs provide a comprehensive overview of the dataset's characteristics and relationships between variables, fulfilling the requirements of an EDA task.