JN file

May 20, 2025

[16]:
```
import seaborn as sns
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

[8]:
```
import pandas as pd

# Load the Titanic datasets
train_df = pd.read_csv("train.csv")
test_df = pd.read_csv("test.csv")
gender_submission_df = pd.read_csv("gender_submission.csv")

# Display basic info for each dataset
train_info = train_df.info()
test_info = test_df.info()
gender_submission_info = gender_submission_df.info()

# Display summary statistics for training data
train_description = train_df.describe(include='all')

# Display first few rows of the training data
train_head = train_df.head()

(train_info, test_info, gender_submission_info, train_description, train_head)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
```

```
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  418 non-null    int64
 1   Pclass       418 non-null    int64
 2   Name         418 non-null    object
 3   Sex          418 non-null    object
 4   Age          332 non-null    float64
 5   SibSp        418 non-null    int64
 6   Parch        418 non-null    int64
 7   Ticket       418 non-null    object
 8   Fare         417 non-null    float64
 9   Cabin        91 non-null     object
 10  Embarked     418 non-null    object
dtypes: float64(2), int64(4), object(5)
memory usage: 36.1+ KB
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  418 non-null    int64
 1   Survived     418 non-null    int64
dtypes: int64(2)
memory usage: 6.7 KB
```

[8]:
```
(None,
 None,
 None,
         PassengerId    Survived      Pclass                 Name   Sex  \
 count    891.000000  891.000000  891.000000                  891   891
 unique          NaN         NaN         NaN                  891     2
 top             NaN         NaN         NaN  Dooley, Mr. Patrick  male
 freq            NaN         NaN         NaN                    1   577
 mean     446.000000    0.383838    2.308642                  NaN   NaN
 std      257.353842    0.486592    0.836071                  NaN   NaN
 min        1.000000    0.000000    1.000000                  NaN   NaN
 25%      223.500000    0.000000    2.000000                  NaN   NaN
 50%      446.000000    0.000000    3.000000                  NaN   NaN
```

```
75%      668.500000    1.000000    3.000000                NaN   NaN
max      891.000000    1.000000    3.000000                NaN   NaN

               Age       SibSp       Parch  Ticket        Fare Cabin Embarked
count   714.000000  891.000000  891.000000     891  891.000000   204      889
unique         NaN         NaN         NaN     681         NaN   147        3
top            NaN         NaN         NaN  347082         NaN    G6        S
freq           NaN         NaN         NaN       7         NaN     4      644
mean     29.699118    0.523008    0.381594     NaN   32.204208   NaN      NaN
std      14.526497    1.102743    0.806057     NaN   49.693429   NaN      NaN
min       0.420000    0.000000    0.000000     NaN    0.000000   NaN      NaN
25%      20.125000    0.000000    0.000000     NaN    7.910400   NaN      NaN
50%      28.000000    0.000000    0.000000     NaN   14.454200   NaN      NaN
75%      38.000000    1.000000    0.000000     NaN   31.000000   NaN      NaN
max      80.000000    8.000000    6.000000     NaN  512.329200   NaN      NaN
,
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S  )
```

[12]:
```
train_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
```

```
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

[21]:
```
test_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 11 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  418 non-null    int64
 1   Pclass       418 non-null    int64
 2   Name         418 non-null    object
 3   Sex          418 non-null    object
 4   Age          332 non-null    float64
 5   SibSp        418 non-null    int64
 6   Parch        418 non-null    int64
 7   Ticket       418 non-null    object
 8   Fare         417 non-null    float64
 9   Cabin        91 non-null     object
 10  Embarked     418 non-null    object
dtypes: float64(2), int64(4), object(5)
memory usage: 36.1+ KB
```

[22]:
```
gender_submission_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  418 non-null    int64
 1   Survived     418 non-null    int64
dtypes: int64(2)
memory usage: 6.7 KB
```

[14]:
```
train_df.describe()
```

[14]:
```
       PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000
mean    446.000000    0.383838    2.308642   29.699118    0.523008
```

```
std     257.353842    0.486592    0.836071   14.526497    1.102743
min       1.000000    0.000000    1.000000    0.420000    0.000000
25%     223.500000    0.000000    2.000000   20.125000    0.000000
50%     446.000000    0.000000    3.000000   28.000000    0.000000
75%     668.500000    1.000000    3.000000   38.000000    1.000000
max     891.000000    1.000000    3.000000   80.000000    8.000000

            Parch        Fare
count  891.000000  891.000000
mean     0.381594   32.204208
std      0.806057   49.693429
min      0.000000    0.000000
25%      0.000000    7.910400
50%      0.000000   14.454200
75%      0.000000   31.000000
max      6.000000  512.329200
```

[23]:
```
test_df.describe()
```

[23]:
```
       PassengerId      Pclass         Age       SibSp       Parch        Fare
count   418.000000  418.000000  332.000000  418.000000  418.000000  417.000000
mean   1100.500000    2.265550   30.272590    0.447368    0.392344   35.627188
std     120.810458    0.841838   14.181209    0.896760    0.981429   55.907576
min     892.000000    1.000000    0.170000    0.000000    0.000000    0.000000
25%     996.250000    1.000000   21.000000    0.000000    0.000000    7.895800
50%    1100.500000    3.000000   27.000000    0.000000    0.000000   14.454200
75%    1204.750000    3.000000   39.000000    1.000000    0.000000   31.500000
max    1309.000000    3.000000   76.000000    8.000000    9.000000  512.329200
```

[24]:
```
gender_submission_df.describe()
```

[24]:
```
       PassengerId    Survived
count   418.000000  418.000000
mean   1100.500000    0.363636
std     120.810458    0.481622
min     892.000000    0.000000
25%     996.250000    0.000000
50%    1100.500000    0.000000
75%    1204.750000    1.000000
max    1309.000000    1.000000
```

[10]:
```
train_df.value_counts()
```

[10]:
```
PassengerId  Survived  Pclass  Name
Sex     Age   SibSp  Parch  Ticket    Fare      Cabin  Embarked
2            1         1       Cumings, Mrs. John Bradley (Florence Briggs
Thayer)   female  38.0  1      0      PC 17599  71.2833   C85    C           1
```

```
572          1         1       Appleton, Mrs. Edward Dale (Charlotte Lamson)
female  53.0  2      0      11769     51.4792   C101   S           1
578          1         1       Silvey, Mrs. William Baird (Alice Munger)
female  39.0  1      0      13507     55.9000   E44    S           1
582          1         1       Thayer, Mrs. John Borland (Marian Longstreth
Morris)  female  39.0  1      1      17421     110.8833  C68    C           1
584          0         1       Ross, Mr. John Hugo
male    36.0  0      0      13049     40.1250   A10    C           1
                                                                        ..
328          1         2       Ball, Mrs. (Ada E Hall)
female  36.0  0      0      28551     13.0000   D      S           1
330          1         1       Hippach, Miss. Jean Gertrude
female  16.0  0      1      111361    57.9792   B18    C           1
332          0         1       Partner, Mr. Austen
male    45.5  0      0      113043    28.5000   C124   S           1
333          0         1       Graham, Mr. George Edward
male    38.0  0      1      PC 17582  153.4625  C91    S           1
890          1         1       Behr, Mr. Karl Howell
male    26.0  0      0      111369    30.0000   C148   C           1
Name: count, Length: 183, dtype: int64
```

[25]:
```
test_df.value_counts()
```

[25]:
```
PassengerId  Pclass  Name
Sex     Age   SibSp  Parch  Ticket         Fare      Cabin            Embarked
904          1       Snyder, Mrs. John Pillsbury (Nelle Stevenson)
female  23.0  1      0      21228          82.2667   B45              S
1
906          1       Chaffee, Mrs. Herbert Fuller (Carrie Constance Toogood)
female  47.0  1      0      W.E.P. 5734    61.1750   E31              S
1
916          1       Ryerson, Mrs. Arthur Larned (Emily Maria Borie)
female  48.0  1      3      PC 17608       262.3750  B57 B59 B63 B66  C
1
918          1       Ostby, Miss. Helene Ragnhild
female  22.0  0      1      113509         61.9792   B36              C
1
920          1       Brady, Mr. John Bertram
male    41.0  0      0      113054         30.5000   A21              S
1
..
1296         1       Frauenthal, Mr. Isaac Gerald
male    43.0  1      0      17765          27.7208   D40              C
1
1297         2       Nourney, Mr. Alfred (Baron von Drachstedt")"
male    20.0  0      0      SC/PARIS 2166  13.8625   D38              C
1
```

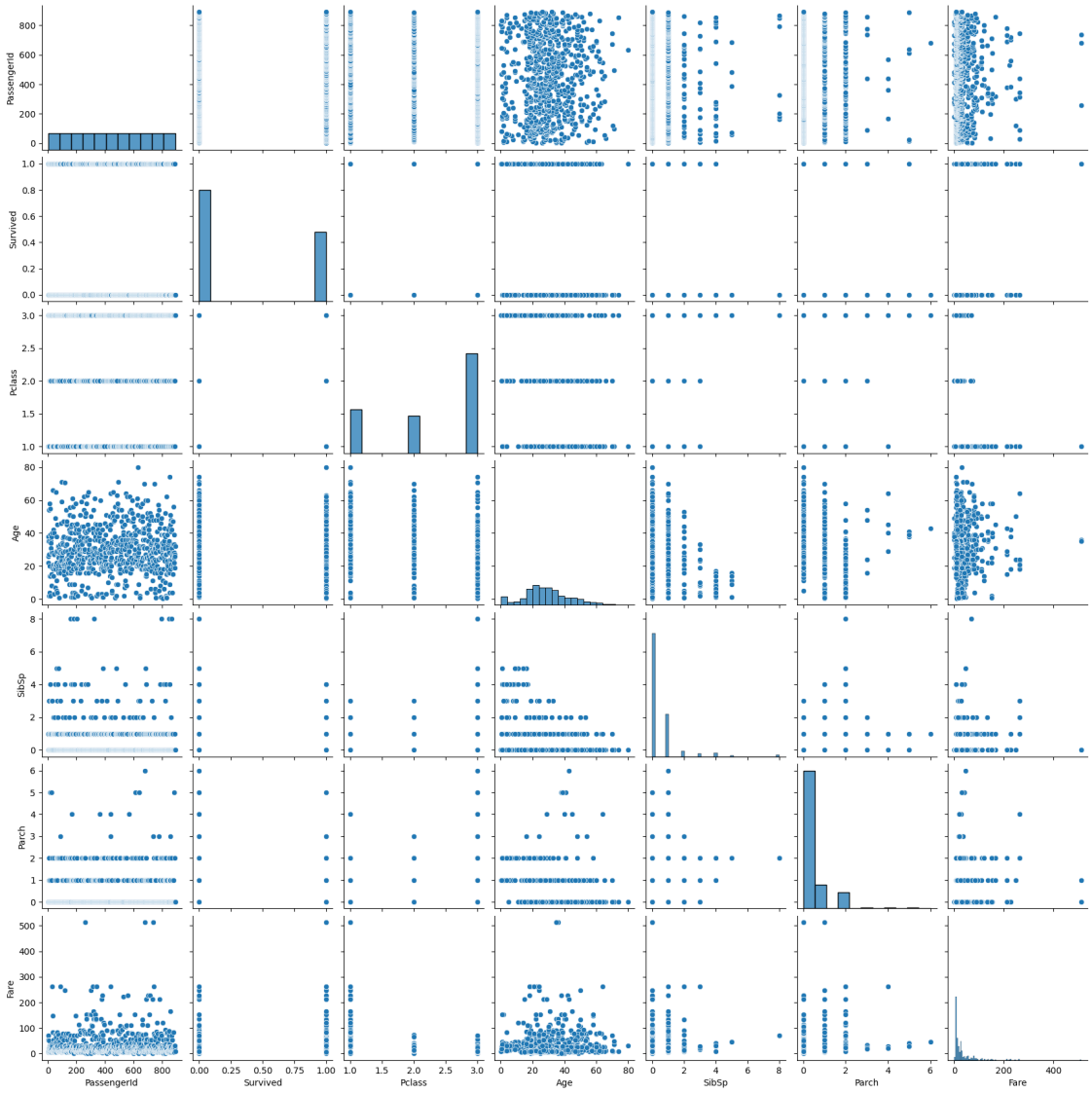
```
1299         1       Widener, Mr. George Dunton
male    50.0  1      1      113503         211.5000  C80              C
1
1303         1       Minahan, Mrs. William Edward (Lillian E Thorpe)
female  37.0  1      0      19928          90.0000   C78              Q
1
1306         1       Oliva y Ocana, Dona. Fermina
female  39.0  0      0      PC 17758       108.9000  C105             C
1
Name: count, Length: 87, dtype: int64
```

[26]:
```
gender_submission_df.value_counts()
```

[26]:
```
PassengerId  Survived
1309         0           1
892          0           1
1293         0           1
1292         1           1
1291         0           1
                        ..
898          1           1
897          0           1
896          1           1
895          0           1
894          0           1
Name: count, Length: 418, dtype: int64
```

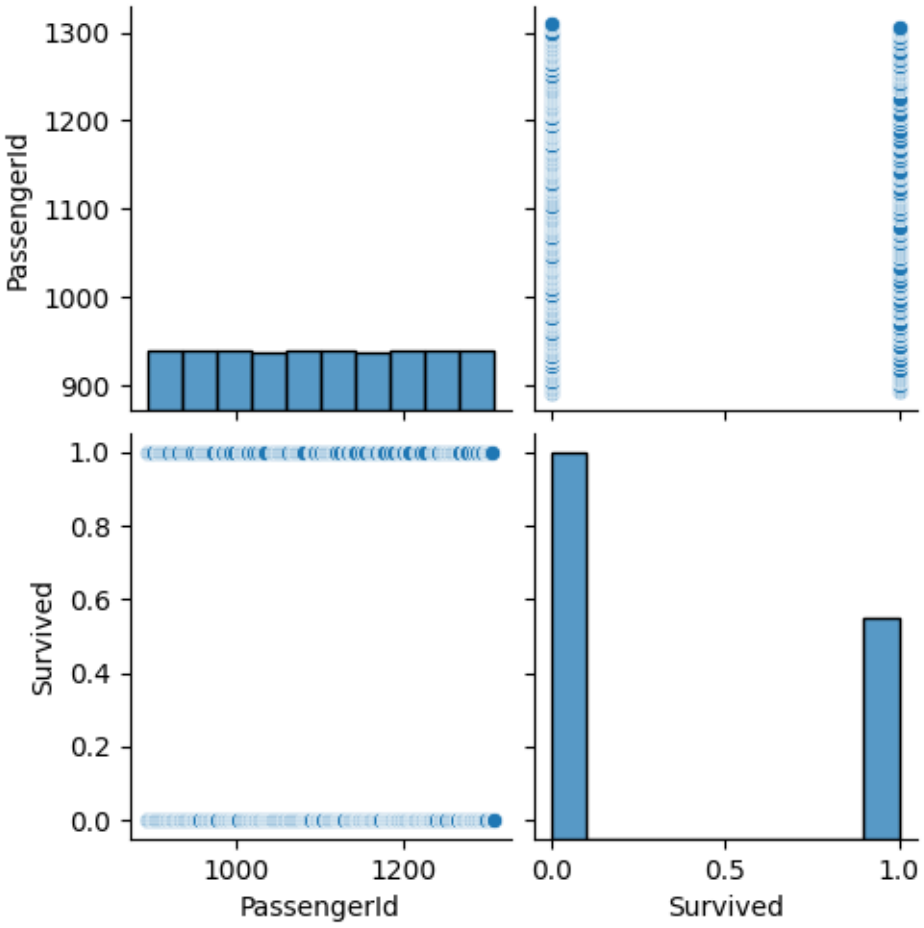
[9]:
```
sns.pairplot(train_df)
```

[9]:
```
<seaborn.axisgrid.PairGrid at 0x20da1762240>
```



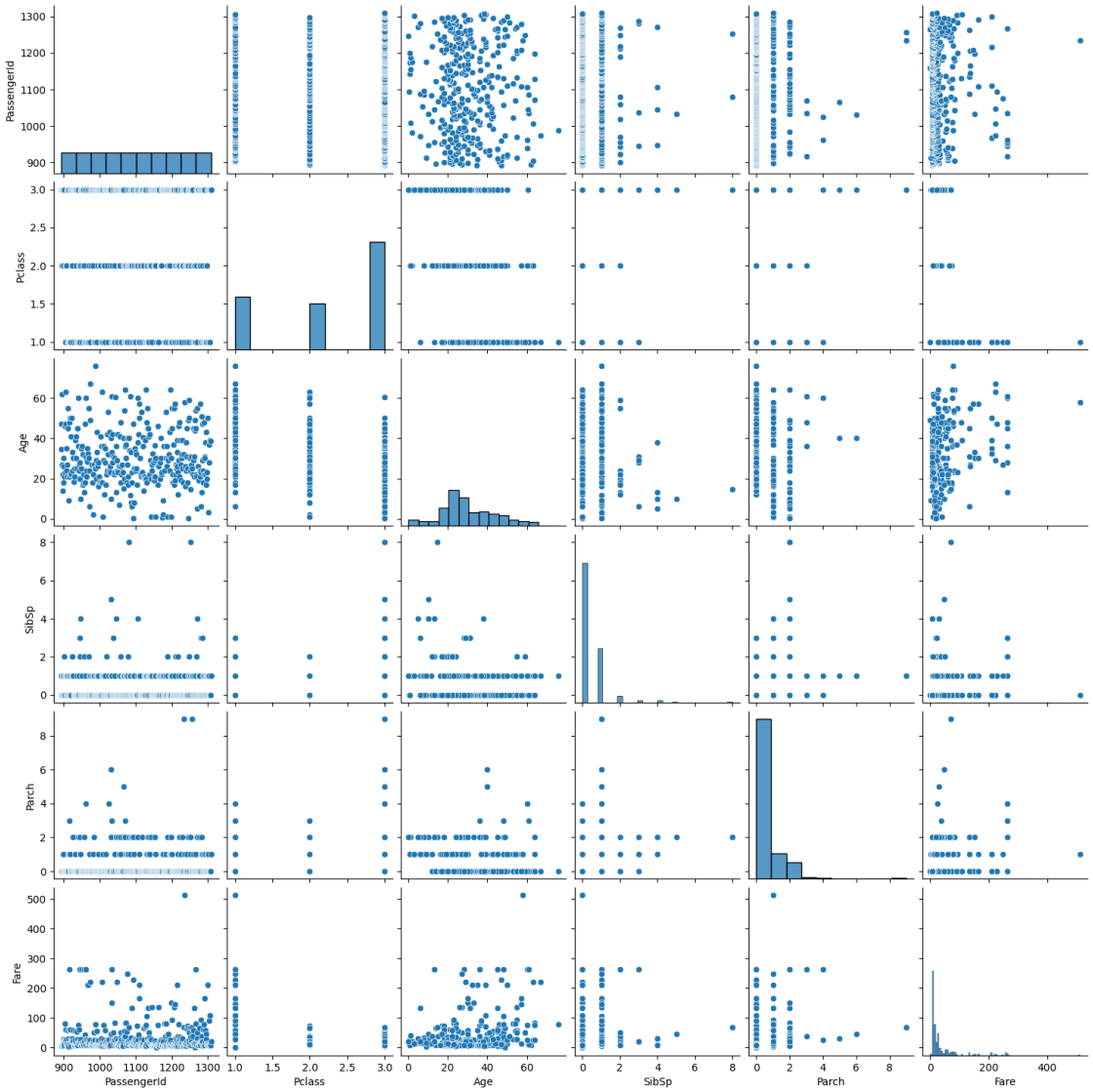
[28]:
```
sns.pairplot(gender_submission_df)
```

[28]: <seaborn.axisgrid.PairGrid at 0x20dabb00c80>



[27]:
```
sns.pairplot(test_df)
```

[27]: <seaborn.axisgrid.PairGrid at 0x20da8eb7b90>

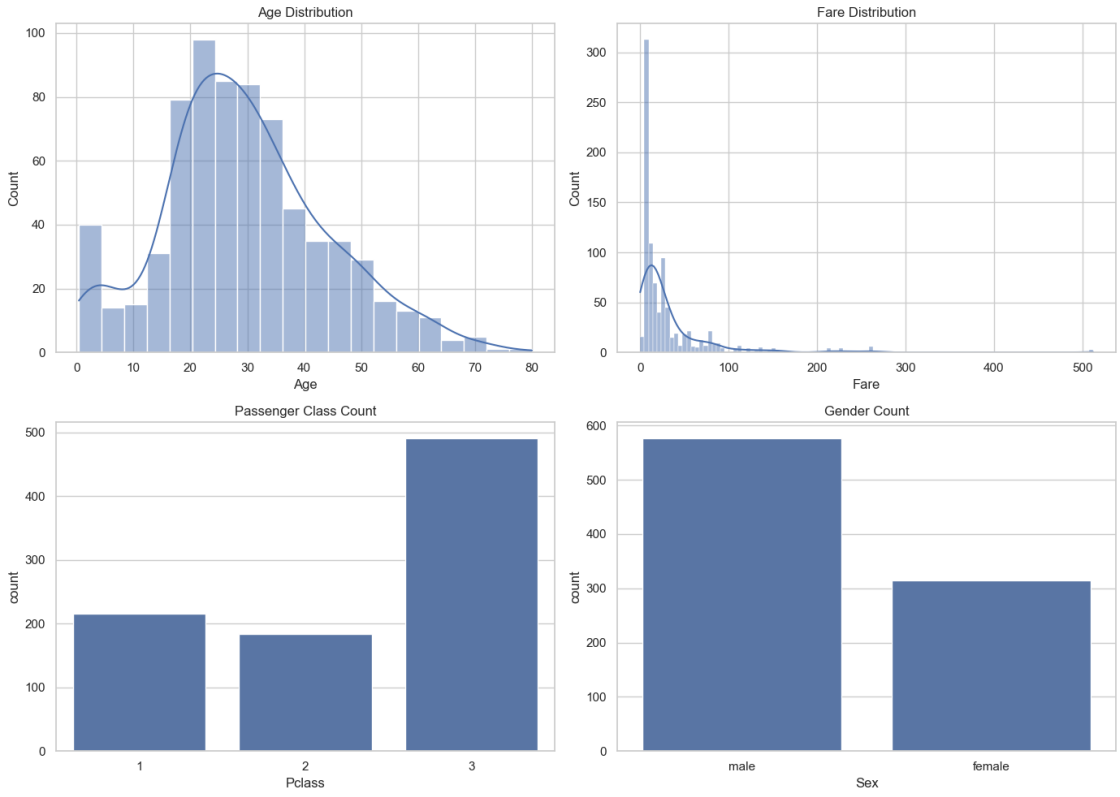


[ ]:

[ ]:

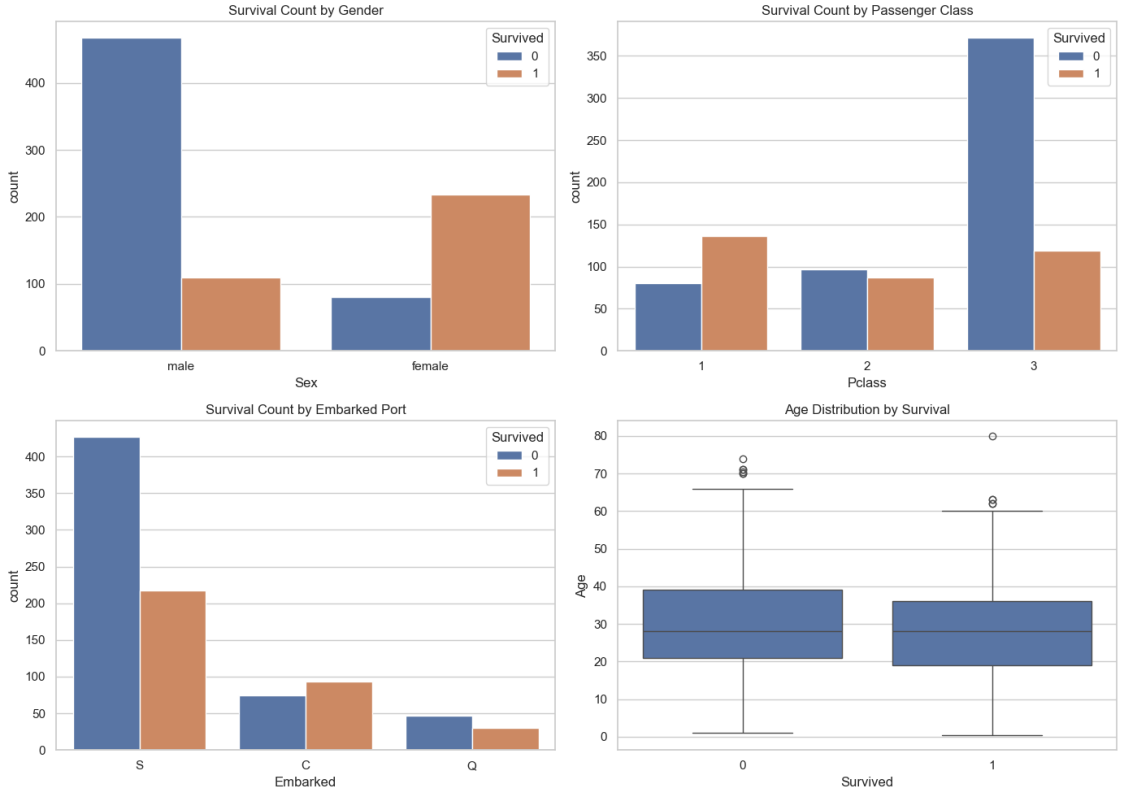
[2]:
```
import seaborn as sns
import matplotlib.pyplot as plt

# Set visual style
sns.set(style="whitegrid")

# Plot distribution of numerical features
fig, axes = plt.subplots(2, 2, figsize=(14, 10))
```

```
sns.histplot(train_df['Age'].dropna(), kde=True, ax=axes[0, 0]).set_title('Age␣
 ↪Distribution')
sns.histplot(train_df['Fare'], kde=True, ax=axes[0, 1]).set_title('Fare␣
 ↪Distribution')
sns.countplot(x='Pclass', data=train_df, ax=axes[1, 0]).set_title('Passenger␣
 ↪Class Count')
sns.countplot(x='Sex', data=train_df, ax=axes[1, 1]).set_title('Gender Count')

plt.tight_layout()
plt.show()
```



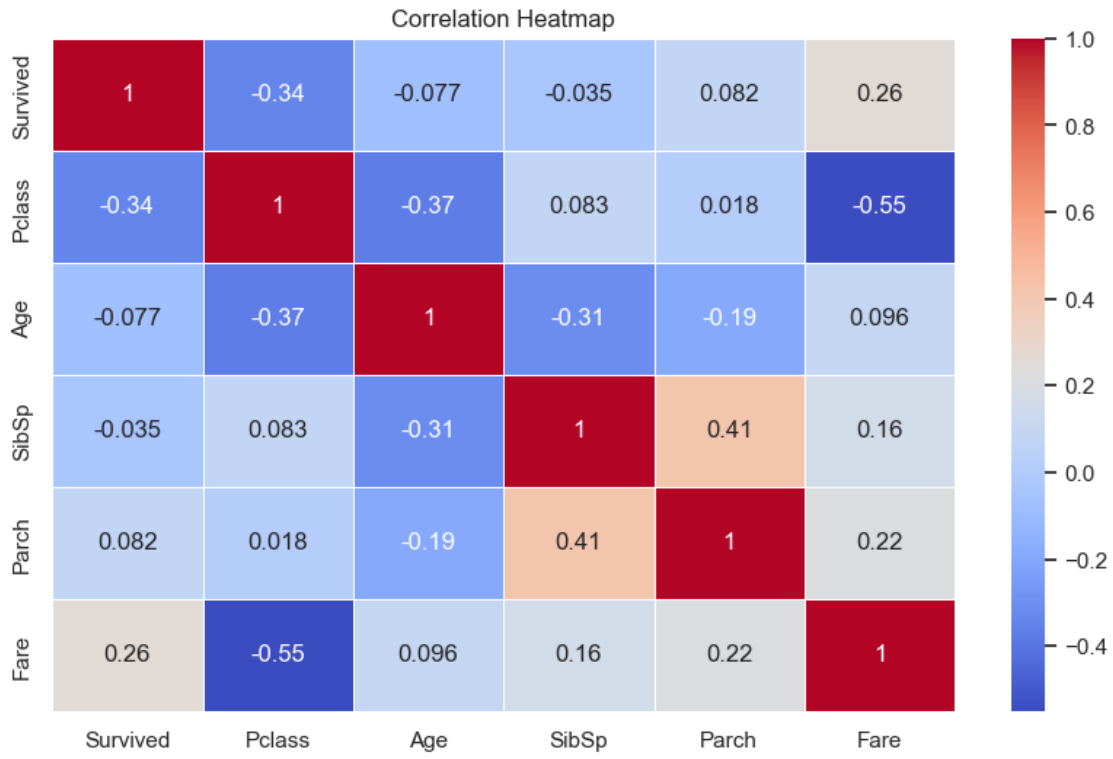
[3]:
```
# Bivariate analysis: survival rate by category

fig, axes = plt.subplots(2, 2, figsize=(14, 10))

# Survival by Sex
sns.countplot(x='Sex', hue='Survived', data=train_df, ax=axes[0, 0])
axes[0, 0].set_title('Survival Count by Gender')

# Survival by Pclass
sns.countplot(x='Pclass', hue='Survived', data=train_df, ax=axes[0, 1])
```

```
axes[0, 1].set_title('Survival Count by Passenger Class')

# Survival by Embarked
sns.countplot(x='Embarked', hue='Survived', data=train_df, ax=axes[1, 0])
axes[1, 0].set_title('Survival Count by Embarked Port')

# Boxplot of Age by Survival
sns.boxplot(x='Survived', y='Age', data=train_df, ax=axes[1, 1])
axes[1, 1].set_title('Age Distribution by Survival')

plt.tight_layout()
plt.show()
```

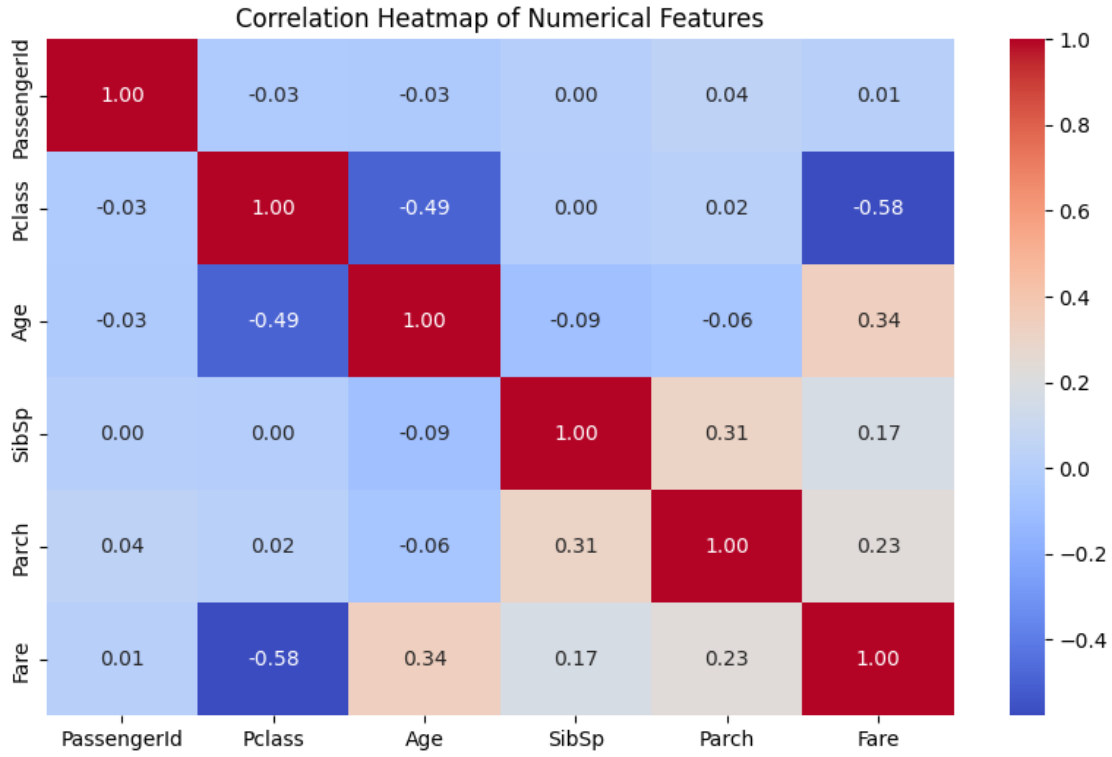


[4]:
```
# Compute correlation matrix for numerical features
corr_matrix = train_df[['Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare']].
  ↪corr()

# Heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Heatmap')
```

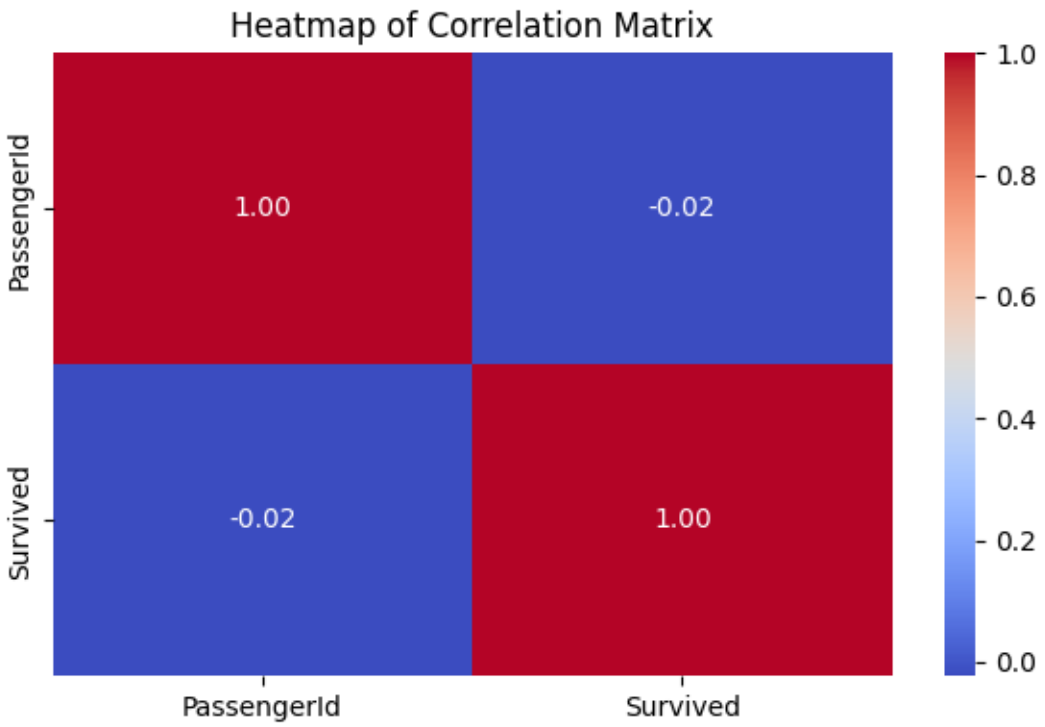
```
plt.show()
```



[19]:
```
# Load the data
df = pd.read_csv("test.csv")

# Select only numerical columns for the correlation matrix
numeric_df = df.select_dtypes(include=['float64', 'int64'])

# Compute the correlation matrix
correlation_matrix = numeric_df.corr()

# Set up the matplotlib figure
plt.figure(figsize=(10, 6))

# Draw the heatmap
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")

# Add a title
plt.title("Correlation Heatmap of Numerical Features")

# Show the plot
plt.show()
```



[20]:
```python
# Load the CSV file
df = pd.read_csv('gender_submission.csv')

# Generate and plot the heatmap
plt.figure(figsize=(6, 4))
correlation_matrix = df.corr(numeric_only=True)
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Heatmap of Correlation Matrix')
plt.tight_layout()
plt.show()
```



[ ]:
```
--THANKS FOR YOUR TIME AND PATIENCE.
```