Titanic Dataset Analysis

This notebook presents an exploratory data analysis (EDA) of the Titanic dataset. The primary goal is to uncover meaningful insights about the passengers, identify patterns that influenced survival rates, and visualize key relationships within the data. This project is part of an internship task under Prodigy InfoTech.

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

from google.colab import files
uploaded = files.upload()
```

```
<IPython.core.display.HTML object>
```

```
Saving Titanic(1).xlsx to Titanic(1).xlsx
```

```
{'titanic.csv': b'...'}
```

```
{'titanic.csv': b'...'}
```

```
uploaded.keys()
```

```
dict_keys([])
```

```
df = pd.read_excel('Titanic.xlsx')
```

```
!pip install openpyxl
```

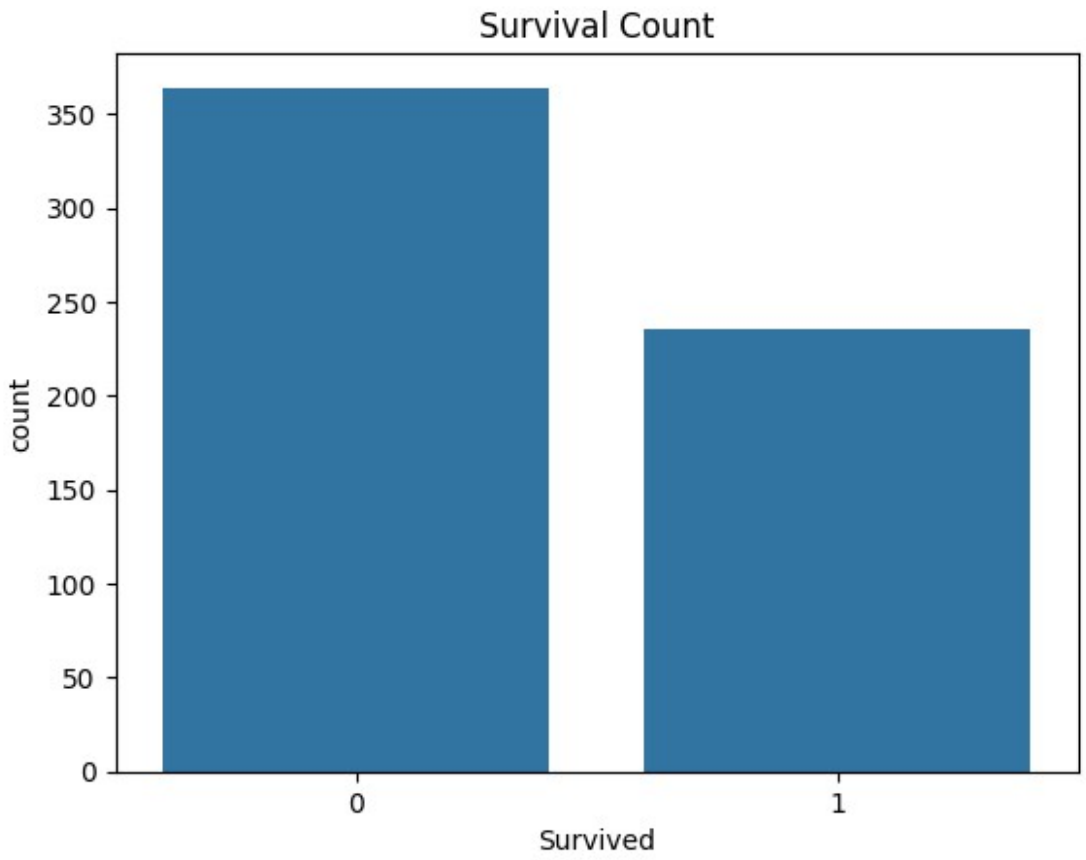
```
Requirement already satisfied: openpyxl in
/usr/local/lib/python3.11/dist-packages (3.1.5)
Requirement already satisfied: et-xmlfile in
/usr/local/lib/python3.11/dist-packages (from openpyxl) (2.0.0)
```

```
from google.colab import files
uploaded = files.upload()
```

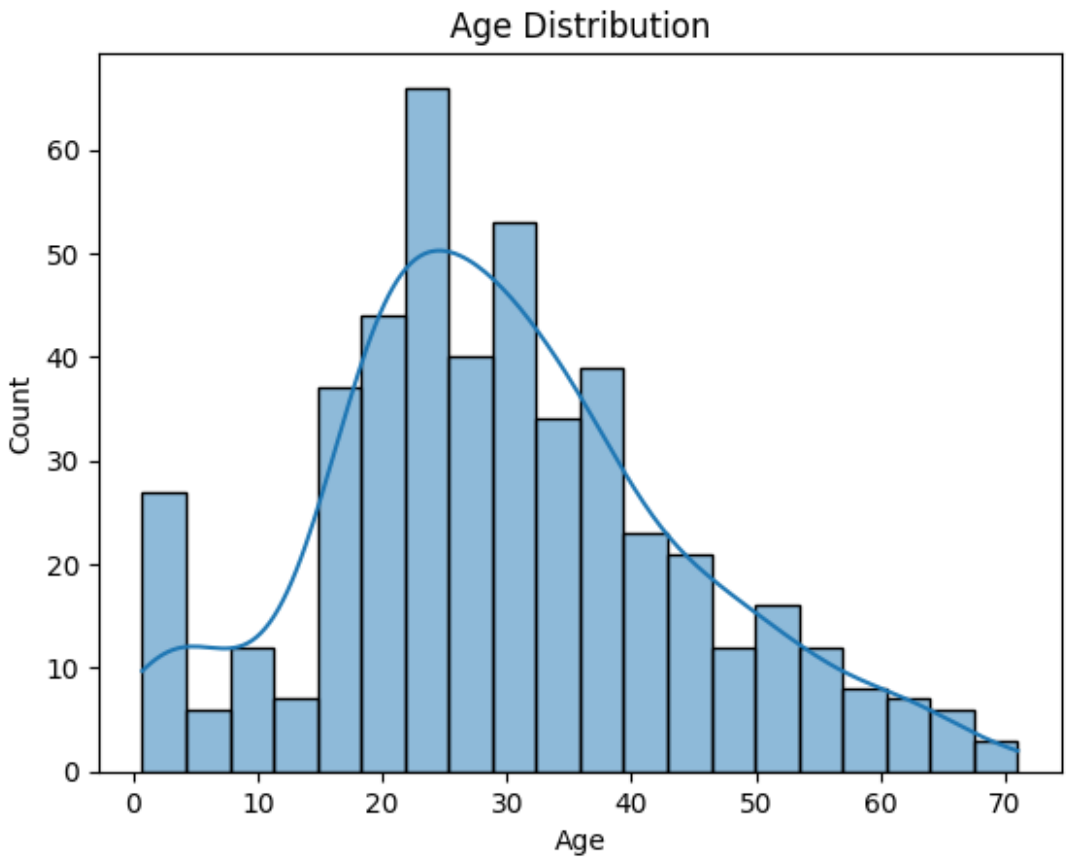
```
import pandas as pd
df = pd.read_excel('Titanic (1).xlsx')  # Use the exact file name from
upload

df.info()
df.describe()
df.columns
```

```
<IPython.core.display.HTML object>
```

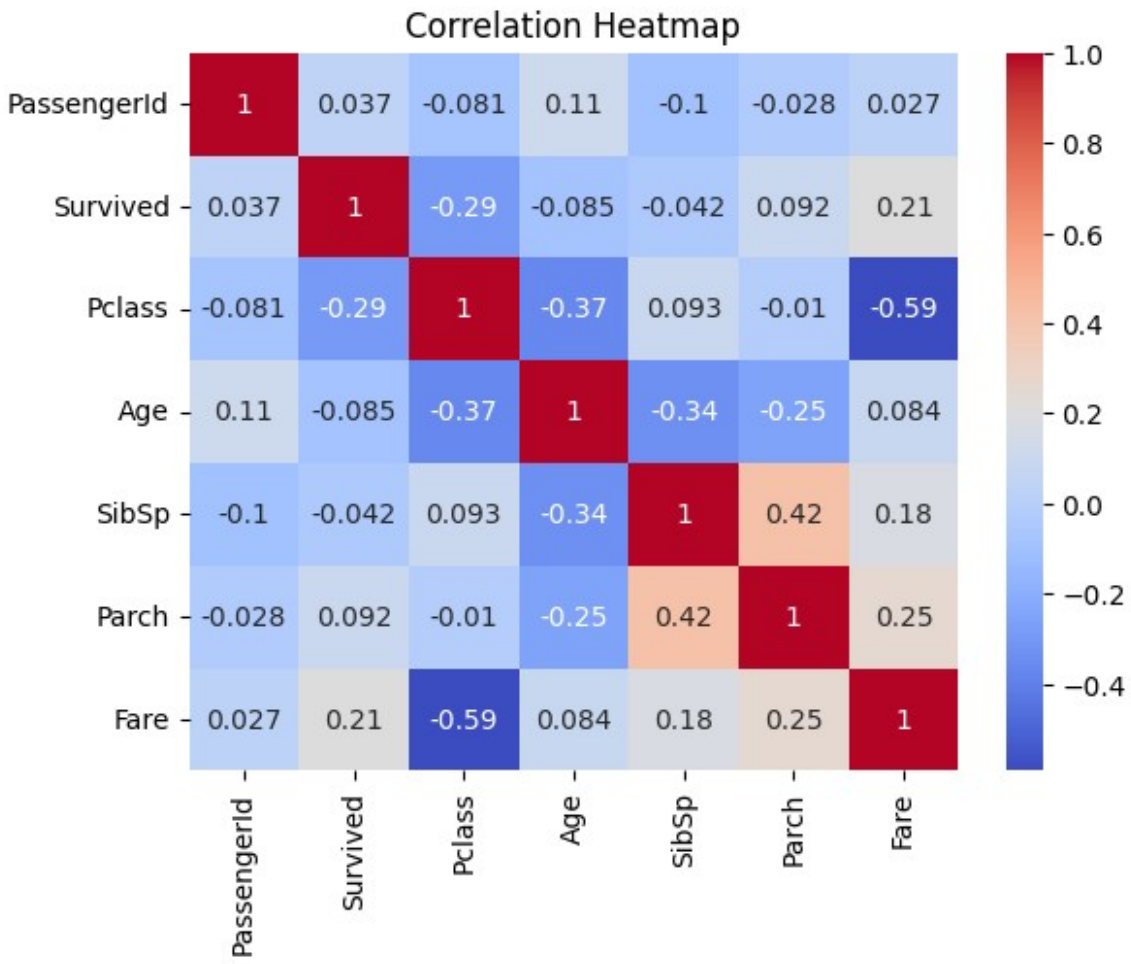
```
Saving Titanic.xlsx to Titanic (2).xlsx
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 599 entries, 0 to 598
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  599 non-null    int64
 1   Survived     599 non-null    int64
 2   Pclass       599 non-null    int64
 3   Name         599 non-null    object
 4   Sex          599 non-null    object
 5   Age          473 non-null    float64
 6   SibSp        599 non-null    int64
 7   Parch        599 non-null    int64
 8   Ticket       599 non-null    object
 9   Fare         599 non-null    float64
 10  Cabin        136 non-null    object
 11  Embarked     598 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 56.3+ KB

Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age',
'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')

import pandas as pd

xls = pd.ExcelFile('Titanic (1).xlsx')
print(xls.sheet_names)

['Sheet1']

['Sheet1']

['Sheet1']

df = pd.read_excel('Titanic (1).xlsx', sheet_name='Sheet1')

print(xls.sheet_names)

['Sheet1']

import seaborn as sns
import matplotlib.pyplot as plt

# Assuming 'df' is your DataFrame
sns.countplot(x='Survived', data=df)
plt.title('Survival Count')
plt.show()
```




```
sns.histplot(df['Age'].dropna(), bins=20, kde=True)
plt.title('Age Distribution')
plt.show()
```




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

# Select only numeric columns for correlation
numeric_df = df.select_dtypes(include='number')

# Plot the heatmap
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```



Conclusion

Through this analysis of the Titanic dataset, we have examined various factors that impacted survival, including gender, class, age, fare, and family size. Key takeaways show that women and children had higher survival rates, and passengers in higher classes were more likely to survive. These insights demonstrate how data analysis can be leveraged to understand historical events and inform predictive modeling tasks.