Titanic Dataset – EDA Summary Report

Pandas, Matplotlib, Seaborn is the python library is used for **EDA** in this titanic dataset.

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns
```

```
df.info()
```

```
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Passenger_Id  418 non-null    int64
 1   Survived      418 non-null    int64
 2   P_class       418 non-null    int64
 3   Name          418 non-null    object
 4   Sex           418 non-null    object
 5   Age           332 non-null    float64
 6   SibSp         418 non-null    int64
 7   Parch         418 non-null    int64
 8   Ticket        418 non-null    object
 9   Fare          417 non-null    float64
 10  Cabin         91 non-null     object
 11  Embarked      418 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 39.3+ KB
```

```
df.describe()
```

	Passenger_Id	Survived	P_class	Age	SibSp	Parch	Fare
count	418.000000	418.000000	418.000000	332.000000	418.000000	418.000000	417.000000
mean	1100.500000	0.363636	2.265550	30.272590	0.447368	0.392344	35.627188
std	120.810458	0.481622	0.841838	14.181209	0.896760	0.981429	55.907576
min	892.000000	0.000000	1.000000	0.170000	0.000000	0.000000	0.000000
25%	996.250000	0.000000	1.000000	21.000000	0.000000	0.000000	7.895800
50%	1100.500000	0.000000	3.000000	27.000000	0.000000	0.000000	14.454200
75%	1204.750000	1.000000	3.000000	39.000000	1.000000	0.000000	31.500000
max	1309.000000	1.000000	3.000000	76.000000	8.000000	9.000000	512.329200

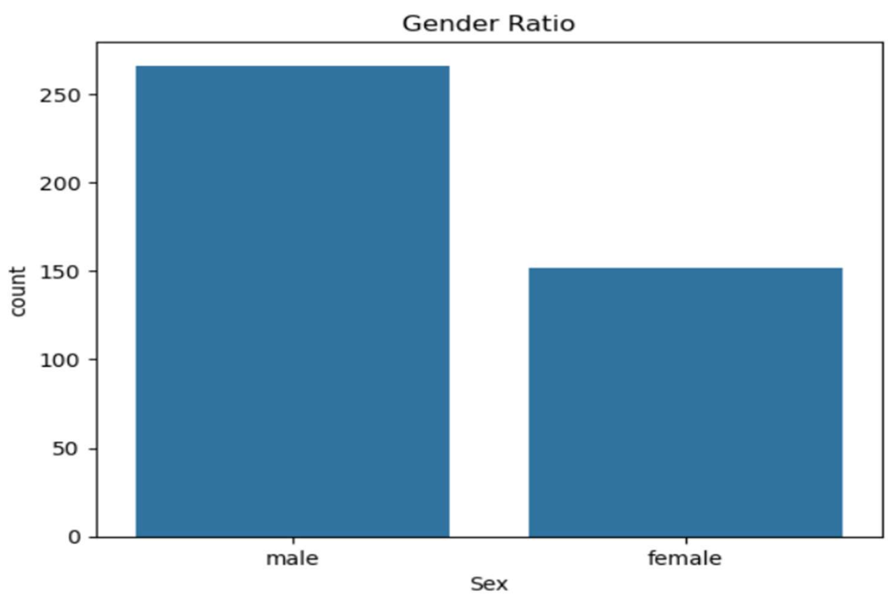
```
sns.countplot(data=df,x="Sex")
plt.title("Gender Ratio")
plt.show()
```

What the plot does:

It counts how many males and females were on board the Titanic.

Insights:

- There are **more male passengers** than female passengers.
- This helps in understanding why survival rates differ later.



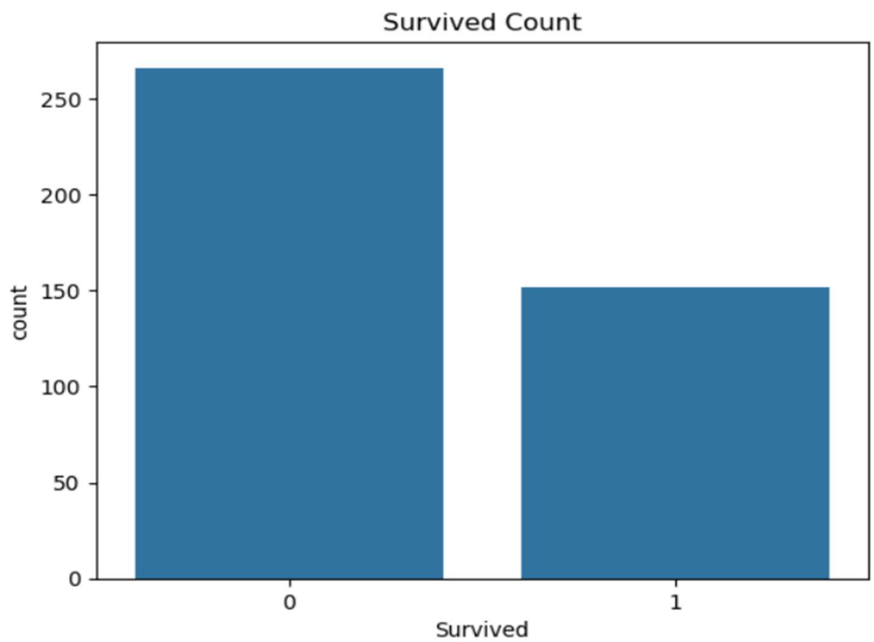

```
sns.countplot(data=df,x="Survived")
plt.title("Survived Count")
plt.show()
```

What the plot does:

Shows how many passengers survived (1) vs did not survive (0).

Insights:

- **More people died** than survived.
- The survival rate is clearly low.



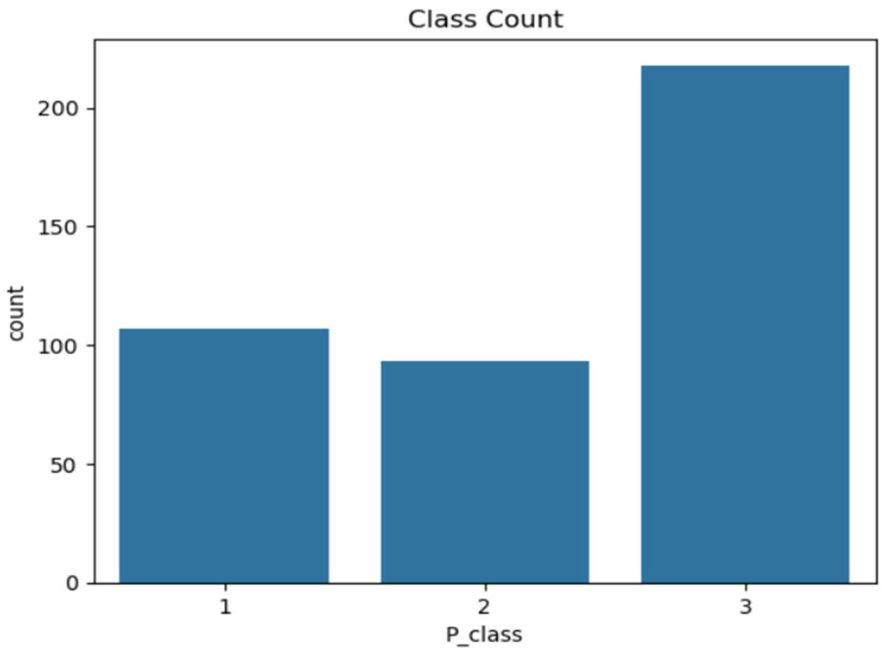

```
sns.countplot(data=df,x="P_class")
plt.title("Class Count")
plt.show()
```

What the plot does:

Shows number of passengers in 1st, 2nd, and 3rd class.

Insights:

- Most passengers were in **3rd class**.
- Least passengers were in **1st class**.



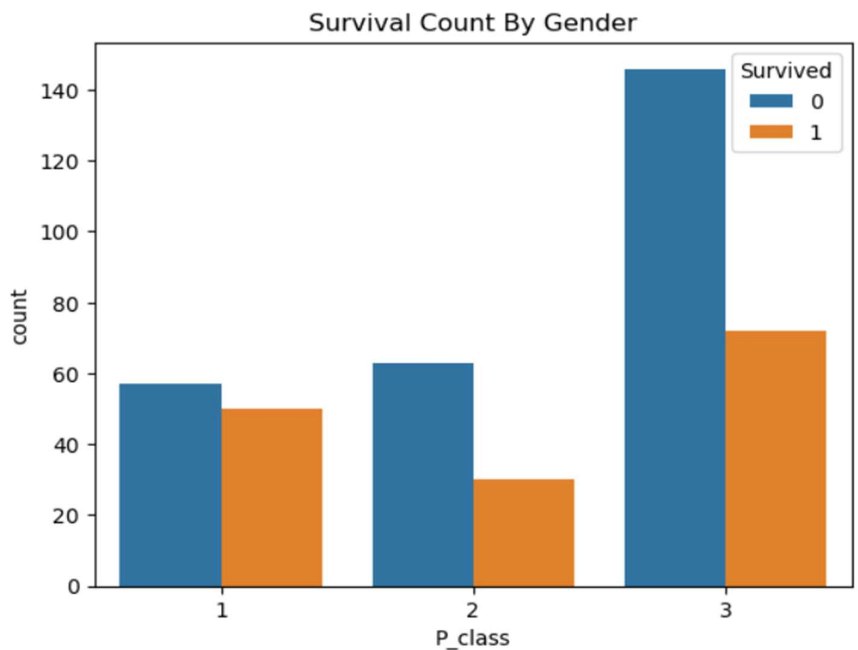

```
sns.countplot(data=df, x="P_class", hue="Survived")
plt.title("Survival Count By Gender")
plt.show()
```

What the plot does:

Compares survival numbers across different passenger classes.

Insights:

- **1st class passengers survived the most.**
- **3rd class passengers died the most.**
- Shows strong relationship between wealth (class) and survival.



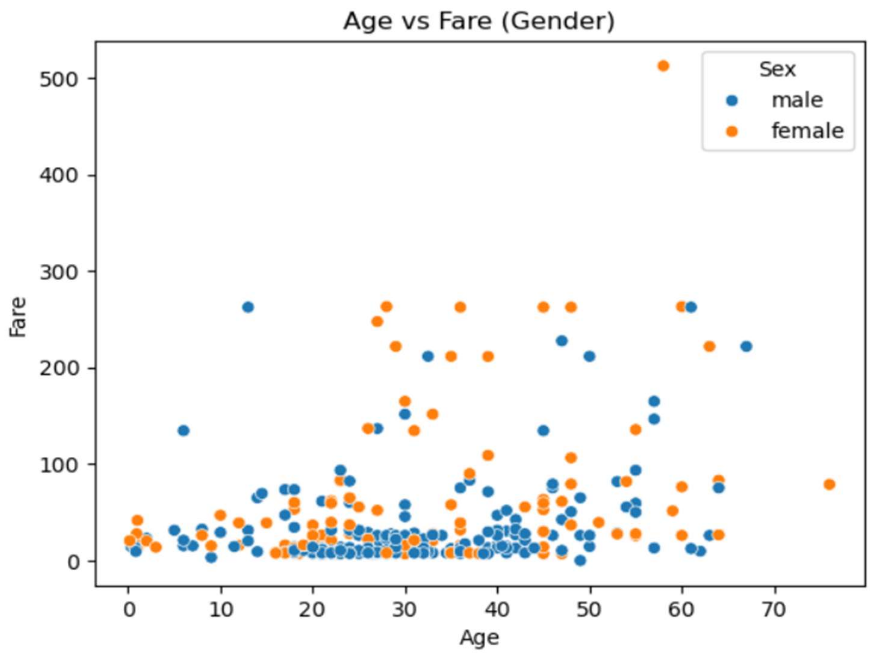

```
sns.scatterplot(data=df,x="Age",y="Fare", hue="Sex")

plt.title("Age vs Fare (Gender)")

plt.show()
```

What the plot does:

Plots relationship between Age and Fare, colored by gender.

Insights:

- Females generally paid **higher fares** (more seen in upper class).
- Younger children appear across all fare ranges.



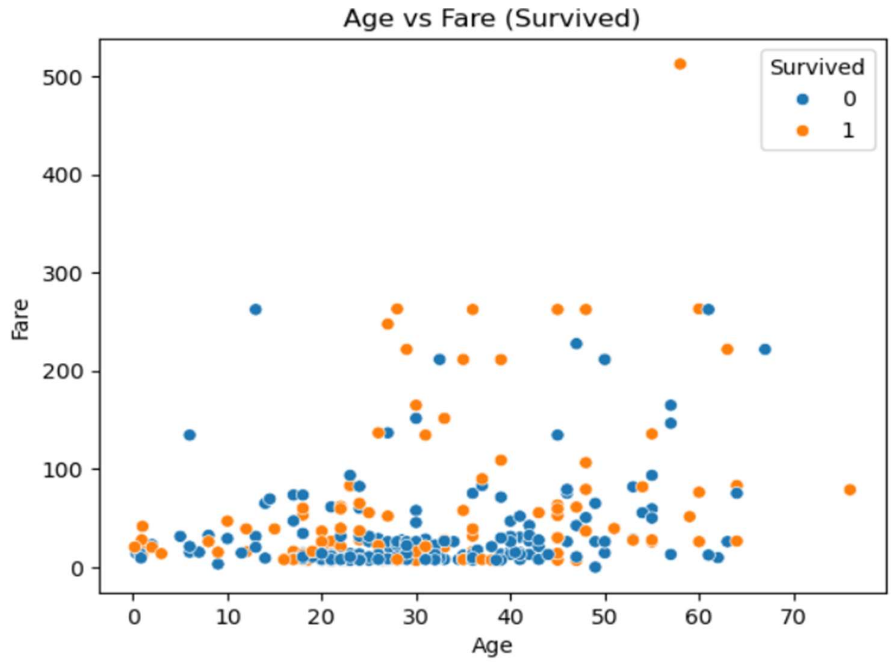

```
sns.scatterplot(data=df, x="Age", y="Fare", hue="Survived")
plt.title("Age vs Fare (Survived)")
plt.show()
```

What the plot does:

Shows how Age and Fare relate to survival.

Insights:

- Many survivors paid **higher fares** → more likely in **1st class**.
- People who paid very low fares mostly **did not survive**.



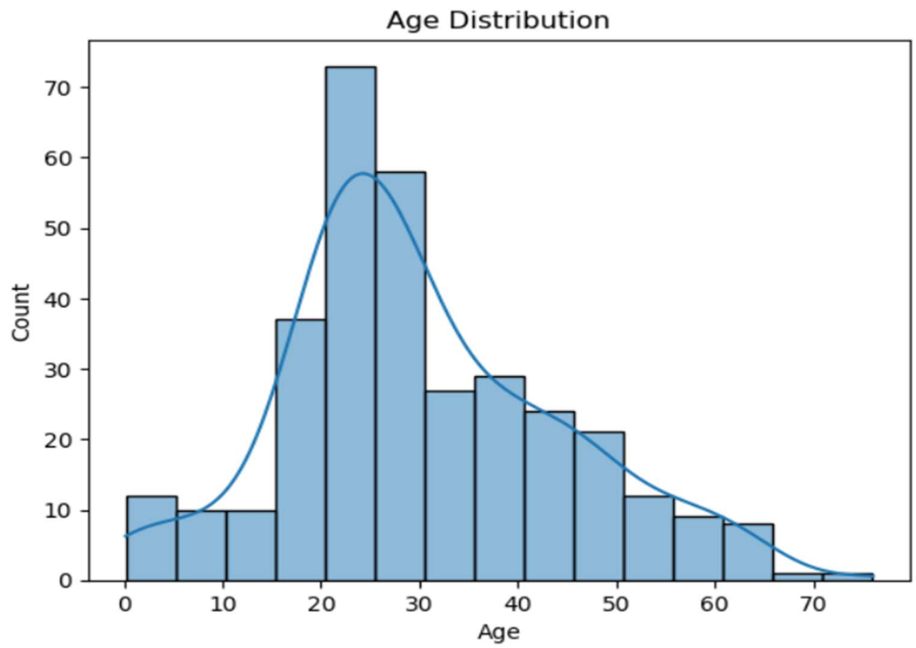

```
sns.histplot(df["Age"],kde=True)

plt.title("Age Distribution")

plt.show()
```

What the plot does:

Shows how ages of passengers are distributed.

Insights:

- Most passengers are between **20 to 40 years old**.
- There are fewer children and elderly passengers.



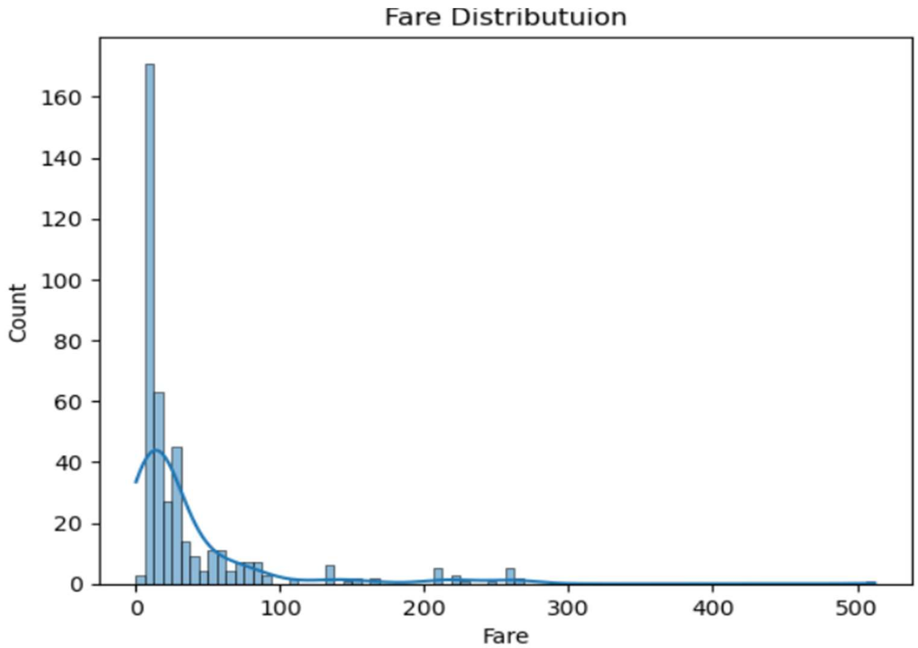

```
sns.histplot(df["Fare"],kde=True)

plt.title("Fare Distributuion")

plt.show()
```

What the plot does:

Shows the distribution of ticket prices.

Insights:

- Most passengers paid **low fares**.
- A few paid extremely high fares → these are generally **1st class**.



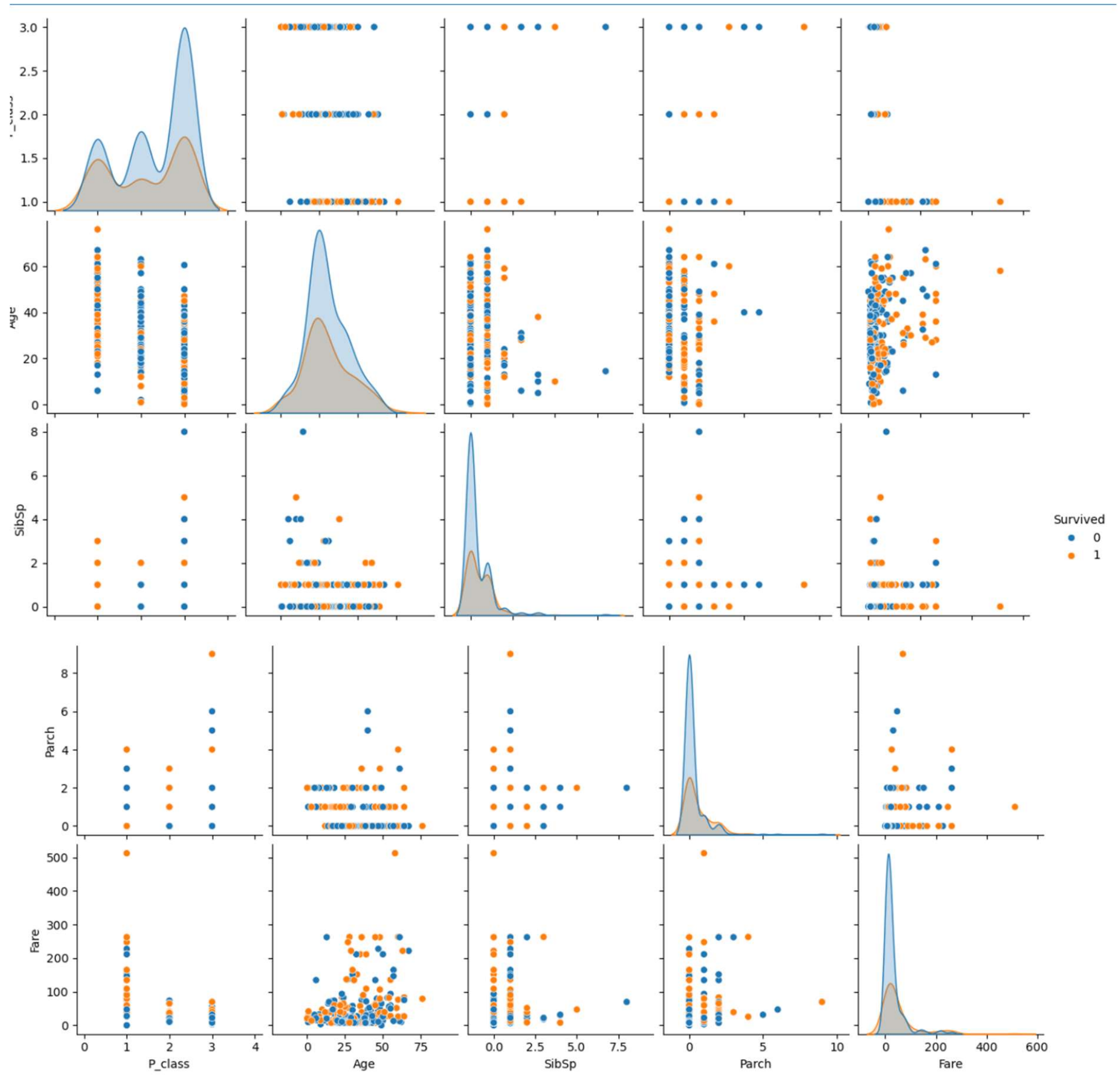

```
sns.pairplot(df[["Survived", "P_class", "Age", "SibSp", "Parch", "Fare"]], hue="Survived")
plt.show()
```

What the plot does:

Creates multiple scatterplots to see relationships among all numeric variables.

Insights:

- Strong visible separation between Fare and Survived.
- P_class strongly influences Fare and Survival.
- Age has weak correlation with survival.




```
plt.figure(figsize=(8,5))

sns.heatmap(df[["Survived","P_class","Age","SibSp","Parch","Fare"]].corr(), annot=True,
cmap="coolwarm")

plt.title("Correlation Heatmap")

plt.show()
```

What the plot does:

Shows correlation values between numeric features.

Key Correlations:

- **P_class and Fare** → strong negative correlation
  (Higher class number = lower fare)

- **Fare and Survived** → positive correlation
  (Higher fare = higher survival)
- **P_class and Survived** → negative correlation
  (Lower class = higher survival)
- **SibSp & Parch** → positive correlation
  (family members often travel together)

Insights:

- Wealth/class strongly affected survival chances.
- Fare is a good predictor for survival.
- Age does not strongly correlate with survival.