Task 5 – Exploratory Data Analysis (EDA) on Titanic Dataset

1 Objective

Perform Exploratory Data Analysis (EDA) on the Titanic Dataset to discover patterns and relationships that affected passenger survival.

The aim is to demonstrate practical skills in data exploration, cleaning, and visual interpretation using Python.

2 Tools & Libraries Used

Python 3	-	**Programming language**
Pandas	-	**Data loading & manipulation**
NumPy	-	**Numeric operations**
Matplotlib	-	**Basic visualizations**
Seaborn	-	**Statistical plots & heatmaps**
Google Colab / Jupyter Notebook	-	**Interactive coding environment**

3 Dataset Overview

Source: Kaggle – Titanic Dataset

File used: train.csv

Rows: 891 **Columns:** 12

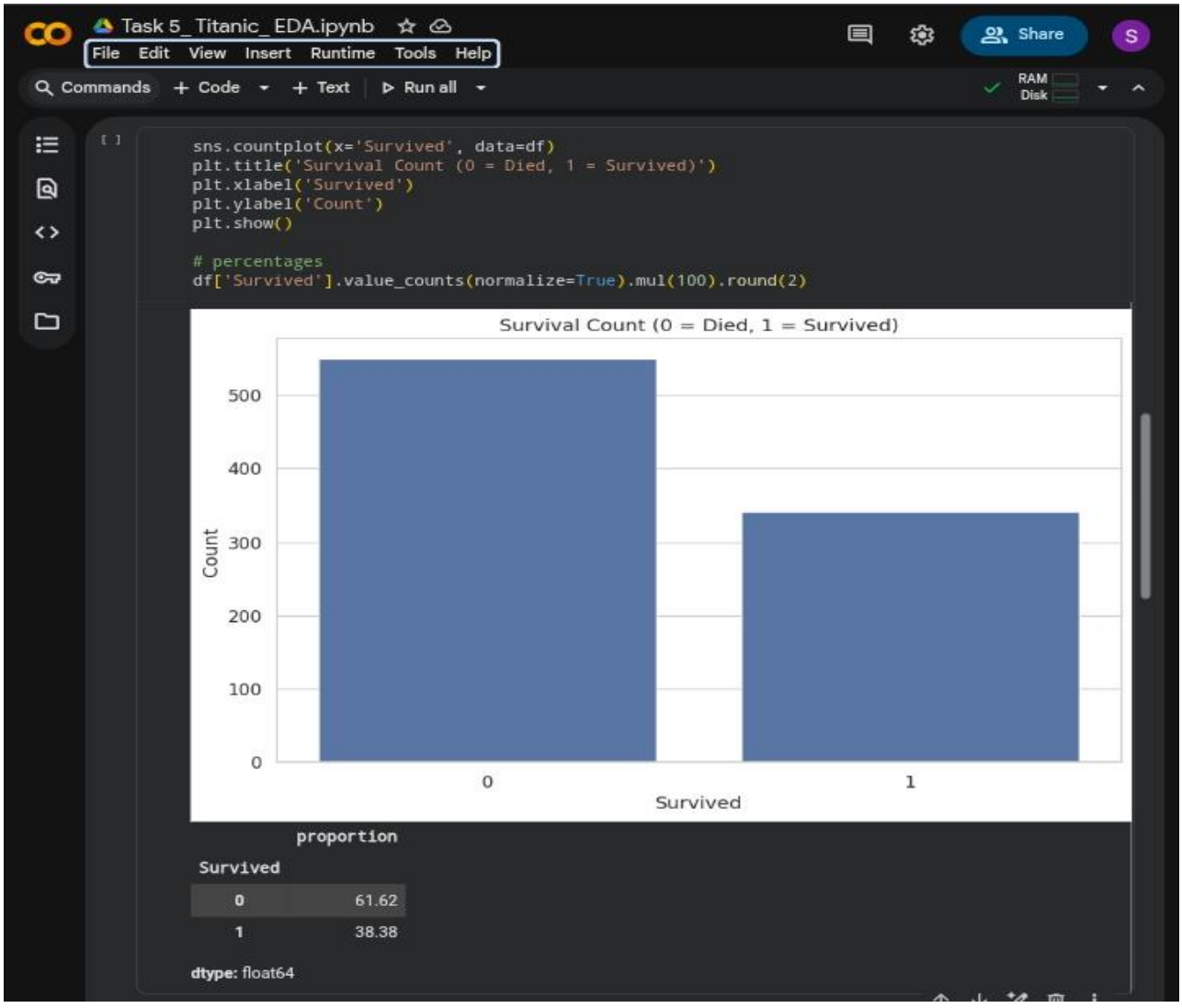
Main Columns:

Survived → Target (1 = Survived, 0 = Did not)

Pclass → Ticket class (1 = Upper, 2 = Middle, 3 = Lower)

Sex, Age, SibSp, Parch, Fare, Embarked, Cabin

4 Visual Highlights



Task 5_ Titanic_ EDA.ipynb

File Edit View Insert Runtime Tools Help

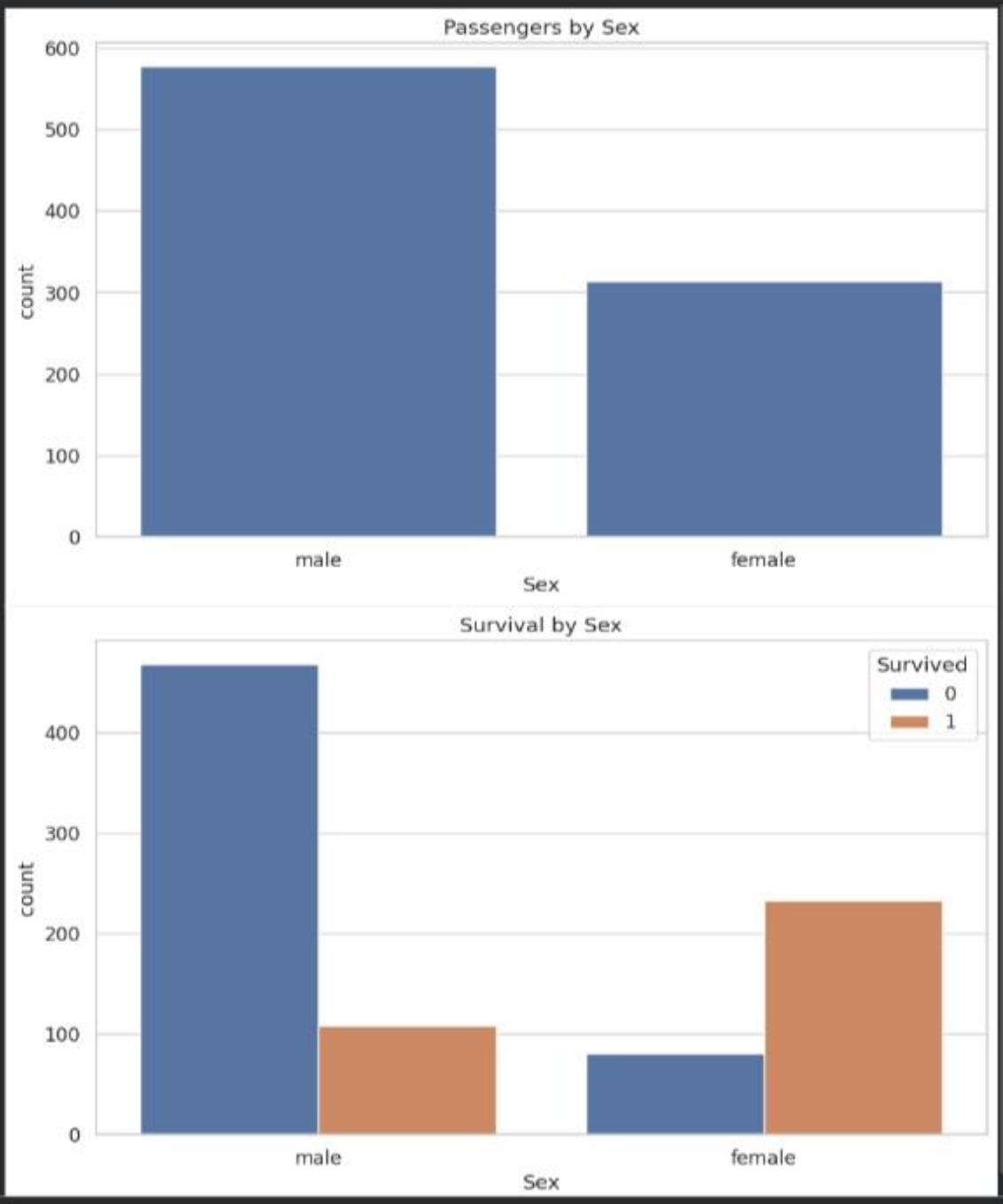
Share

Commands + Code + Text Run all

RAM Disk

```
sns.countplot(x='Sex', data=df)
plt.title('Passengers by Sex')
plt.show()

sns.countplot(x='Sex', hue='Survived', data=df)
plt.title('Survival by Sex')
plt.legend(title='Survived')
plt.show()
```



Observations

- Male had ~74 % survival rate; females ~19 %.

Task 5_Titanic_EDA.ipynb

File Edit View Insert Runtime Tools Help

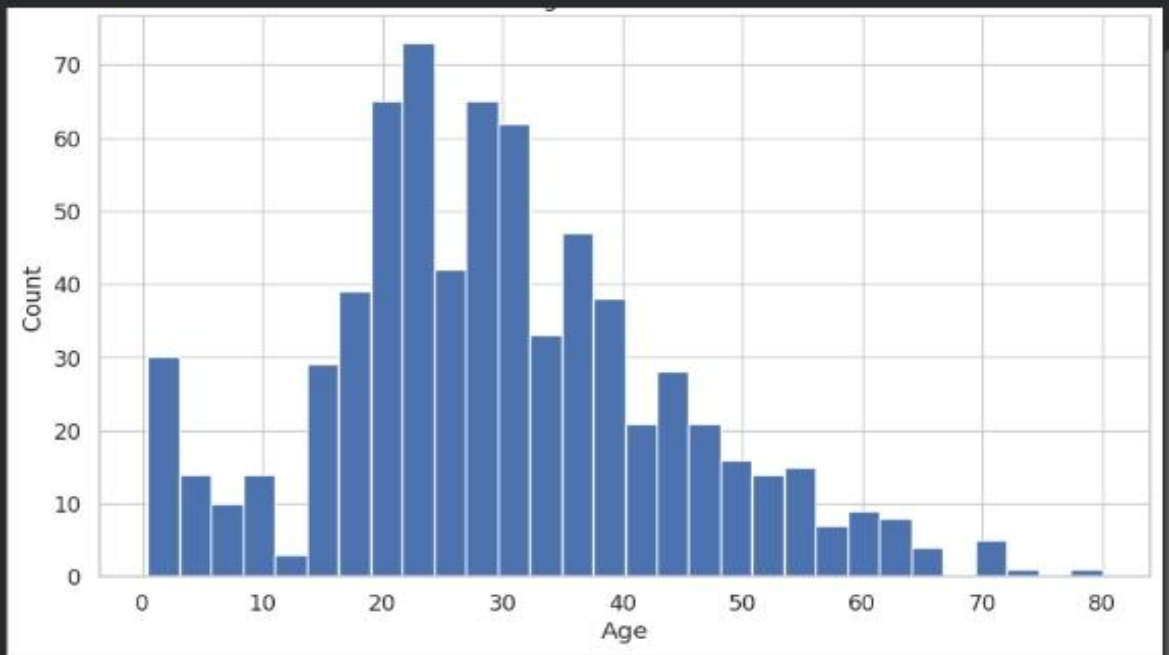
Share

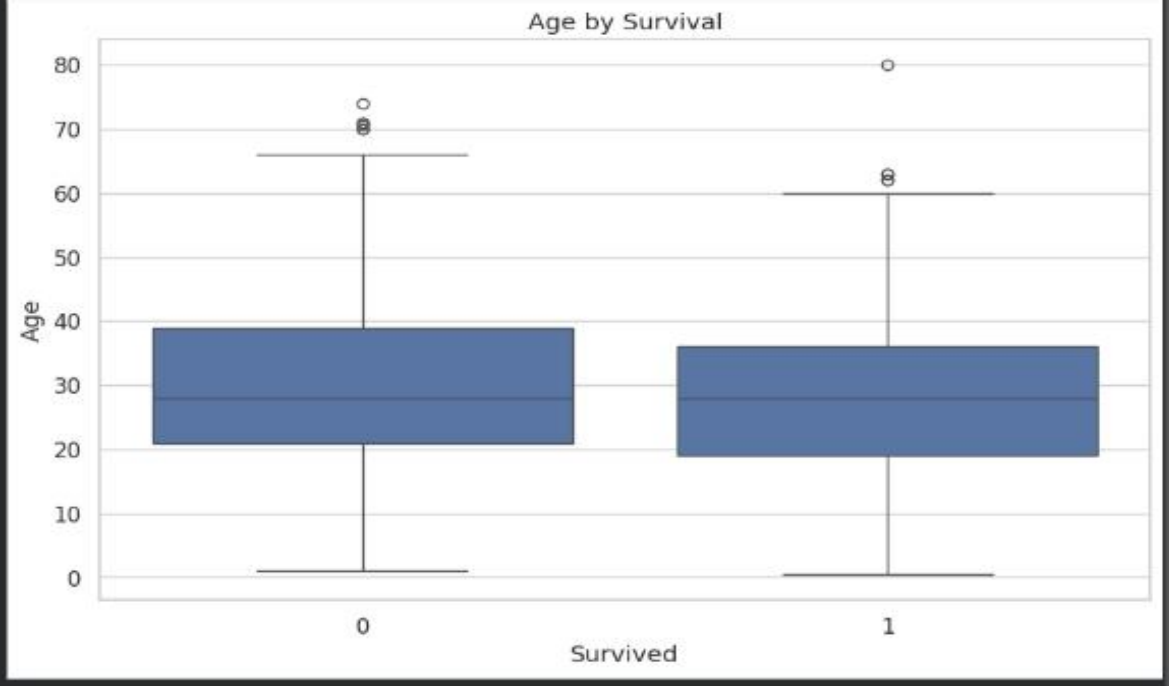
Commands + Code + Text Run all

RAM Disk

```
# histogram
df['Age'].hist(bins=30)
plt.title('Age distribution')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()

# boxplot by survival
sns.boxplot(x='Survived', y='Age', data=df)
plt.title('Age by Survival')
plt.show()
```





Observations -Middle Age peoples are mostly Present.

Task 5_ Titanic_ EDA.ipynb

File Edit View Insert Runtime Tools Help

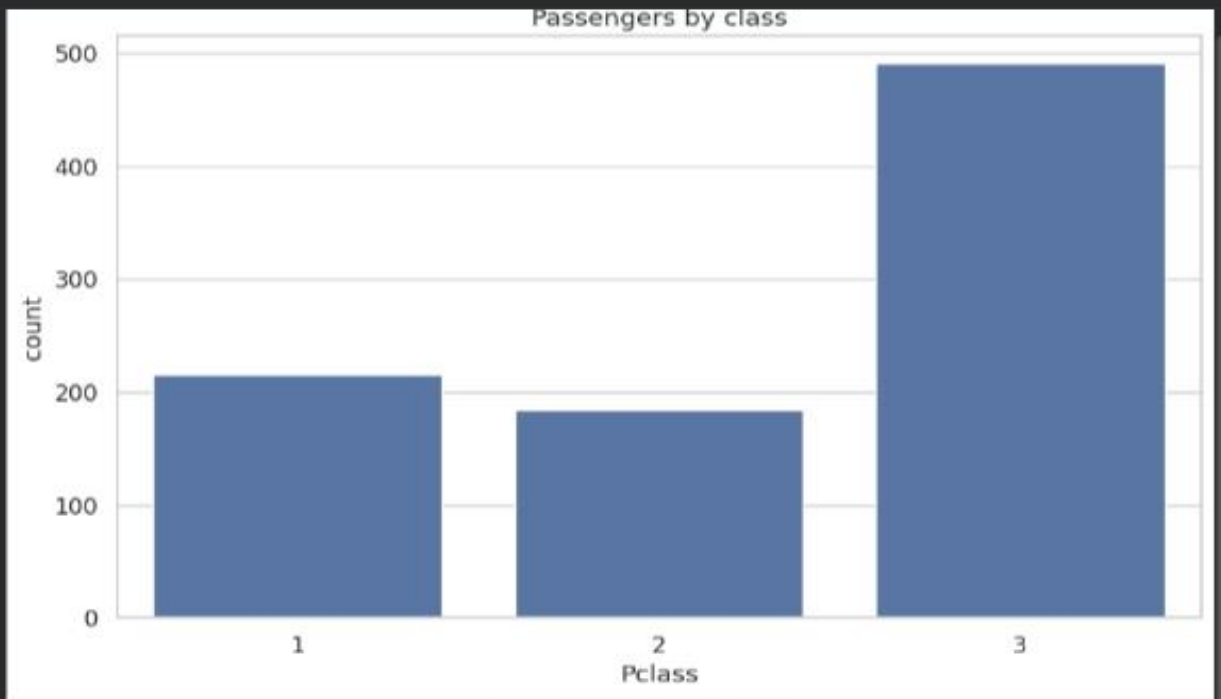
Share

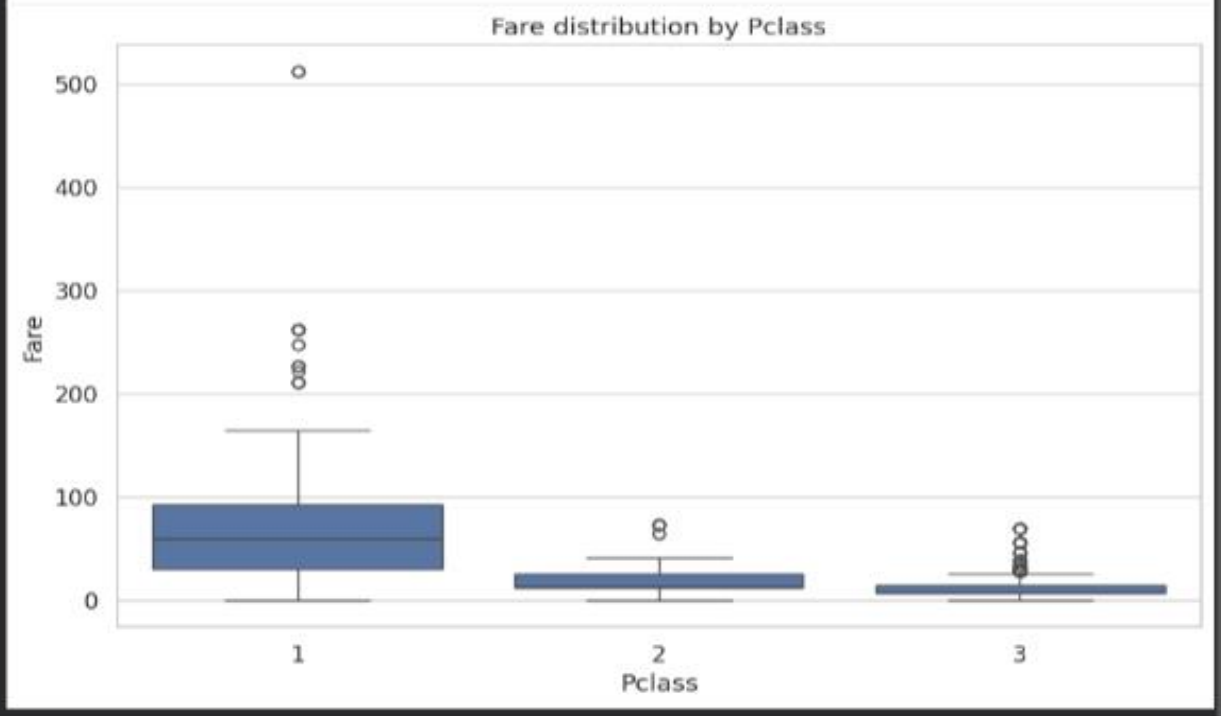
Commands + Code + Text Run all

RAM Disk

```
sns.countplot(x='Pclass', data=df)
plt.title('Passengers by class')
plt.show()

sns.boxplot(x='Pclass', y='Fare', data=df)
plt.title('Fare distribution by Pclass')
plt.show()
```





Observations -No of Passengers count in first class is More than the third class.

Task 5_ Titanic_ EDA.ipynb

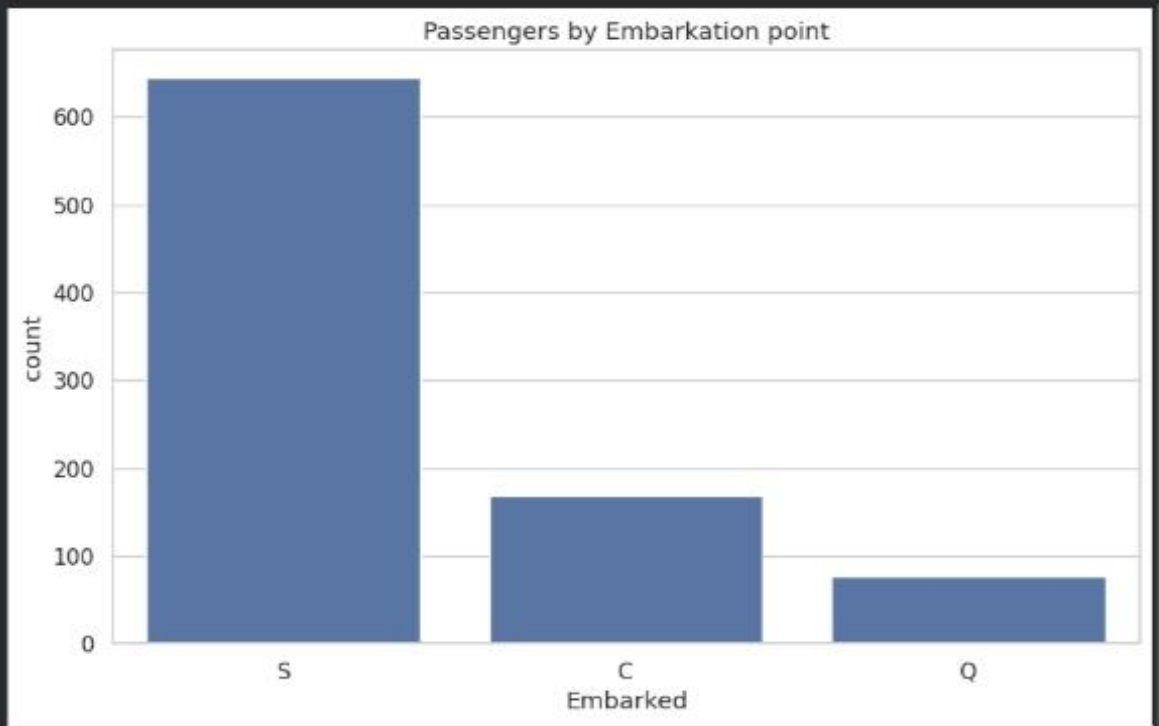
File Edit View Insert Runtime Tools Help

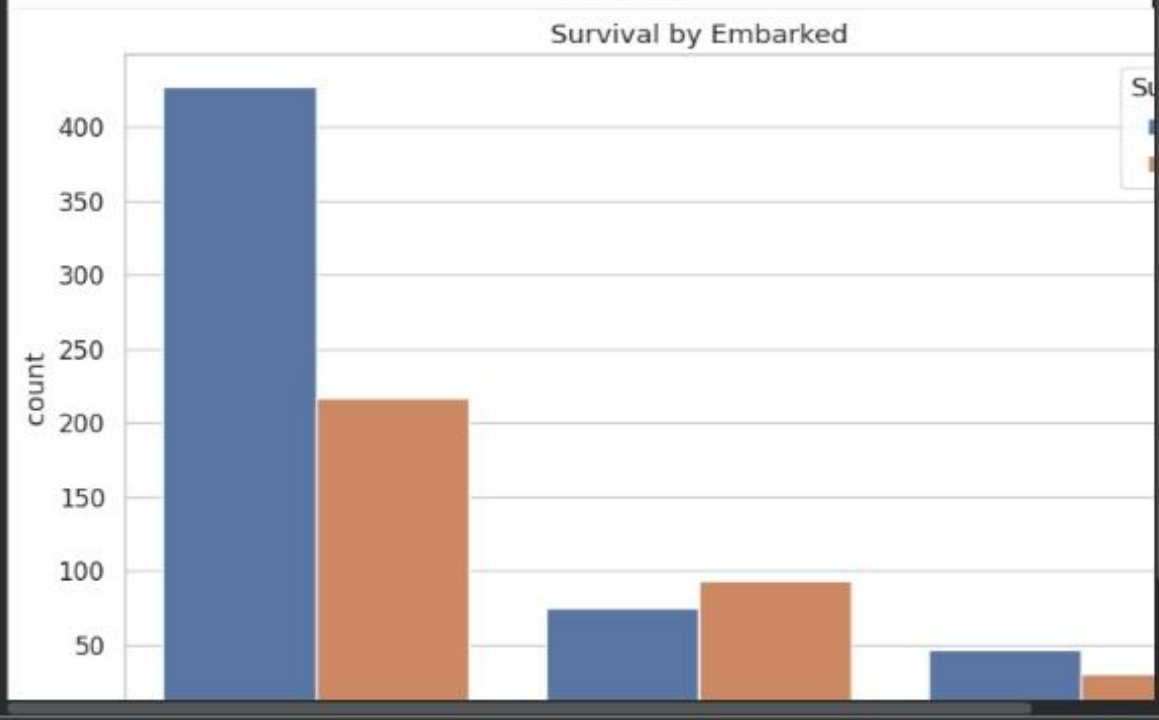
Share

Commands + Code + Text Run all

RAM Disk

```
sns.countplot(x='Embarked', data=df)
plt.title('Passengers by Embarkation point')
plt.show()
sns.countplot(x='Embarked', hue='Survived', data=df)
plt.title('Survival by Embarked')
plt.show()
```





Task 5_ Titanic_ EDA.ipynb

File Edit View Insert Runtime Tools Help

Commands + Code + Text Run all

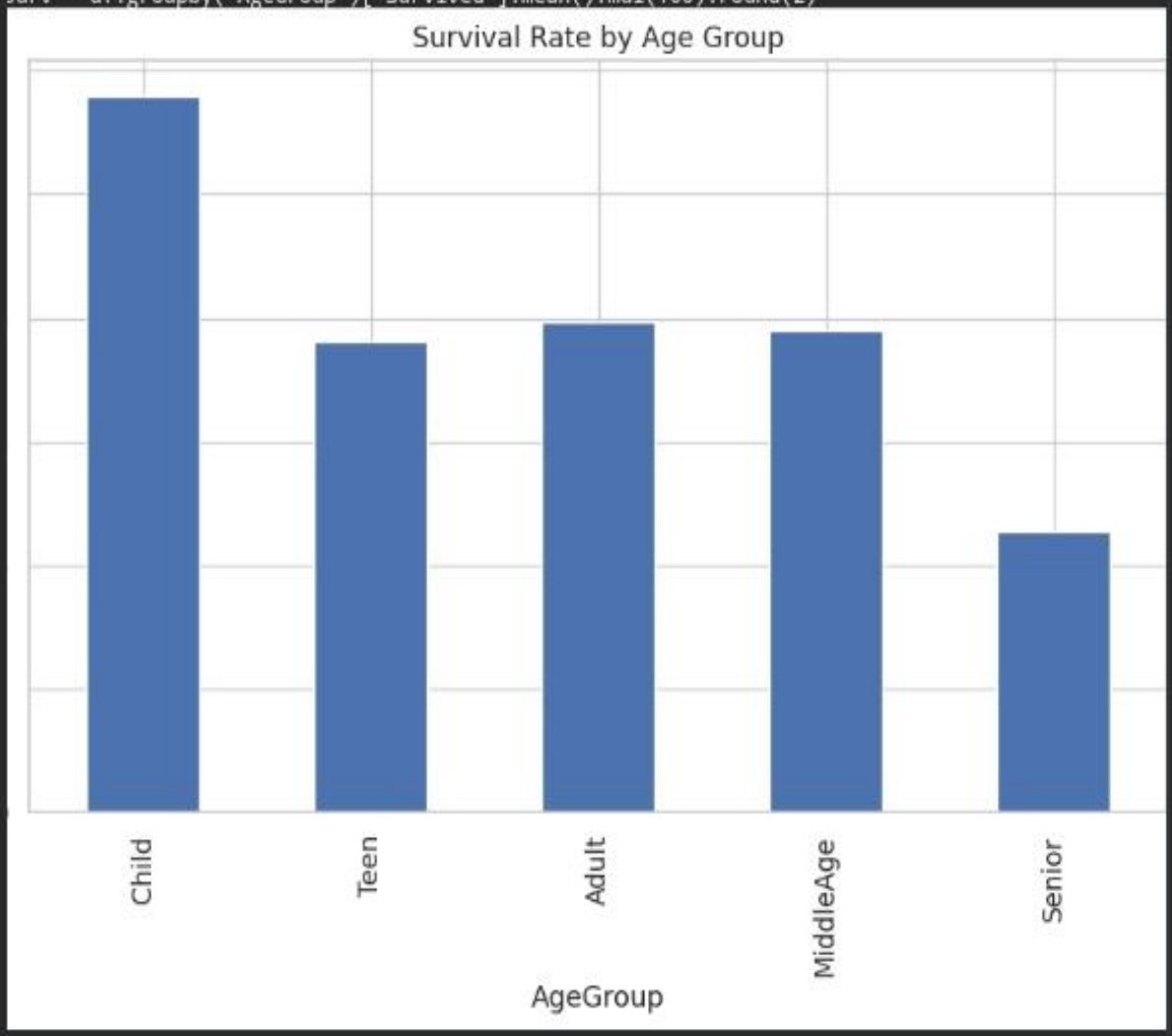
RAM Disk

Share

```
# create age bins
bins = [0,12,20,40,60,200]
labels = ['Child','Teen','Adult','MiddleAge','Senior']
df['AgeGroup'] = pd.cut(df['Age'], bins=bins, labels=labels)

# survival rate by age group
age_surv = df.groupby('AgeGroup')['Survived'].mean().mul(100).round(2)
age_surv
age_surv.plot(kind='bar')
plt.ylabel('Survival Rate (%)')
plt.title('Survival Rate by Age Group')
plt.show()
```

```
ython-input-2794545954.py:7: FutureWarning: The default of observed=False is deprecated and w
surv = df.groupby('AgeGroup')['Survived'].mean().mul(100).round(2)
```



Observations

- Younger passengers had better survival chances.

Task 5_ Titanic_ EDA.ipynb

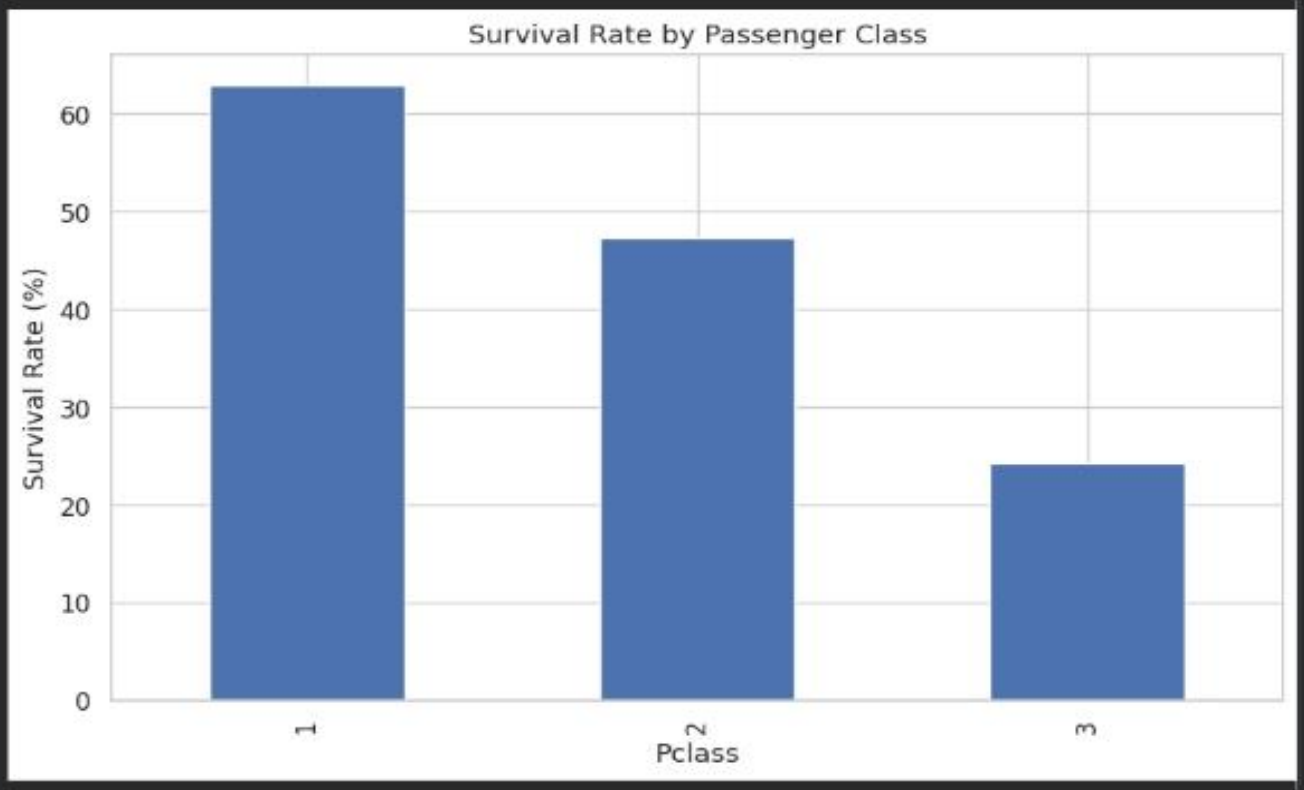
File Edit View Insert Runtime Tools Help

Share

Commands + Code + Text Run all

RAM Disk

```
pclass_surv = df.groupby('Pclass')['Survived'].mean().mul(100).round(2)
pclass_surv.plot(kind='bar')
plt.ylabel('Survival Rate (%)')
plt.title('Survival Rate by Passenger Class')
plt.show()
```



Observations

- 1st class passengers survived more often than 3rd class.

Task 5_ Titanic_ EDA.ipynb

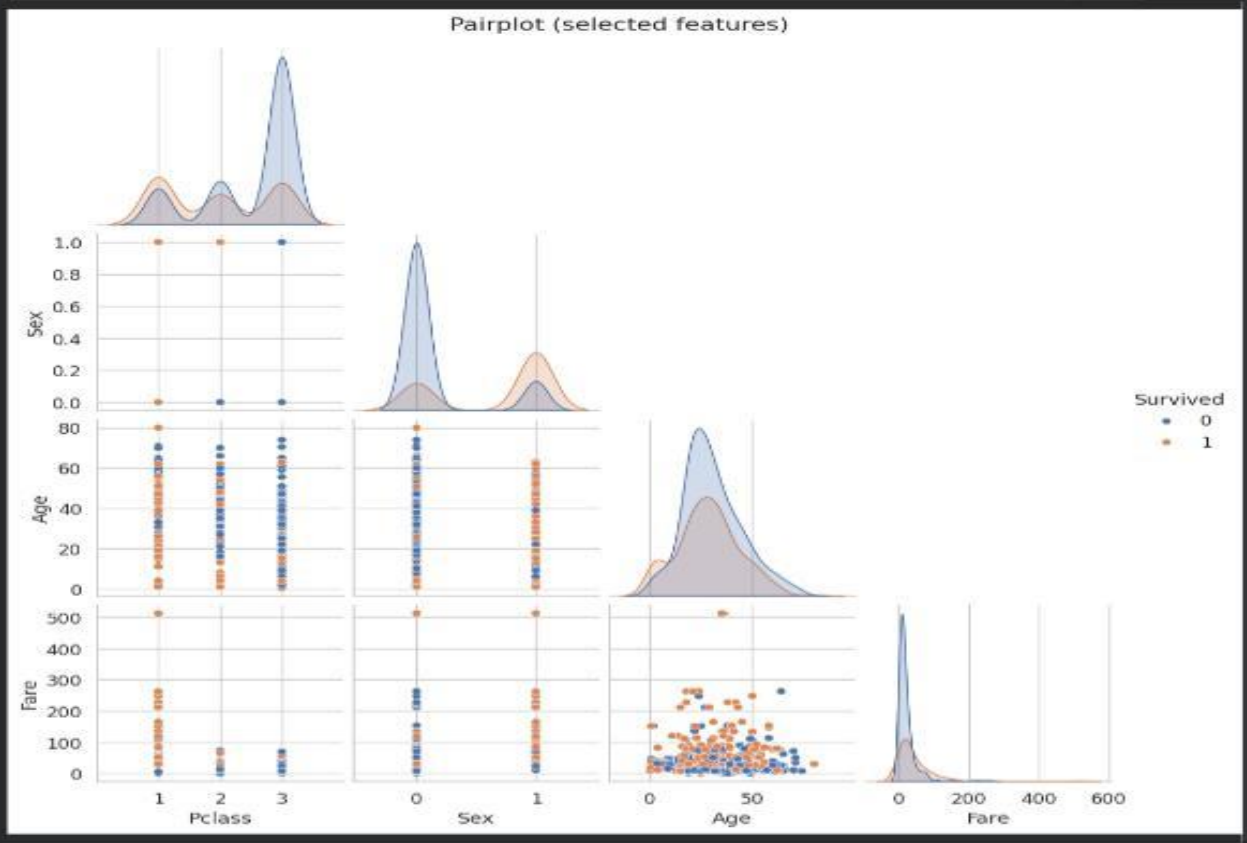
File Edit View Insert Runtime Tools Help

Share

Commands + Code + Text Run all

RAM Disk

```
sns.pairplot(corr_df[['Survived','Pclass','Sex','Age','Fare']], hue='Survived', corner=True)
plt.suptitle('Pairplot (selected features)', y=1.02)
plt.show()
```




```
# family size
df['FamilySize'] = df['SibSp'] + df['Parch'] + 1

# title extraction
df['Title'] = df['Name'].str.extract(r',\s*([^\.]+)\.', expand=False)
df['Title'] = df['Title'].replace(['Lady','Countess','Capt','Col','Don','Dr','Major','Rev','
df['Title'] = df['Title'].replace({'Mlle':'Miss','Ms':'Miss','Mme':'Mrs'})

df[['Title','Survived']].groupby('Title').mean().sort_values(by='Survived', ascending=False)
```

Title	Survived
the Countess	1.000000
Mrs	0.793651
Miss	0.702703
Master	0.575000
Rare	0.318182
Mr	0.156673

Summary of Findings

- Sex and class strongly affect survival.
- Fare correlates positively with survival.
- Missing Cabin data could be replaced by "HasCabin" flag.

5 Key Findings & Insights

Women and children had significantly higher survival rates.

1st-class passengers survived far more than lower classes.

Higher fare = better survival chance.

Age and family size show moderate influence.

Missing Cabin information reduces data completeness.

6 Conclusion

EDA on the Titanic dataset revealed that gender, passenger class, and fare amount were the most important factors influencing survival.

This analysis demonstrates the importance of data visualization and cleaning for discovering patterns before machine-learning modelling.