Arnav Tumbde
RCOEM Nagpur

Domain : Data Science

#Task 2

- Perform data cleaning and exploratory data analysis (EDA) on a dataset of your choice, such as the Titanic dataset from Kaggle. Explore the relationships between variables and identify patterns and trends in the data.

Screenshots of Source Code Jupyter Notebook :

Loading the Dataset

In [3]:
```
import pandas as pd
titanic = pd.read_csv("D:\\Semester - IV\\Prodigy Internship\\Task 2\\archive\\Titanic-Dataset.csv")
```

Data Cleaning

Check for Missing Values

In [4]:
```
titanic.isnull().sum()
```

Out[4]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Handling Missing Values

In [9]:
```
titanic['Age'].fillna(titanic['Age'].median(), inplace=True)
titanic.dropna(subset=['Embarked'], inplace=True)
```

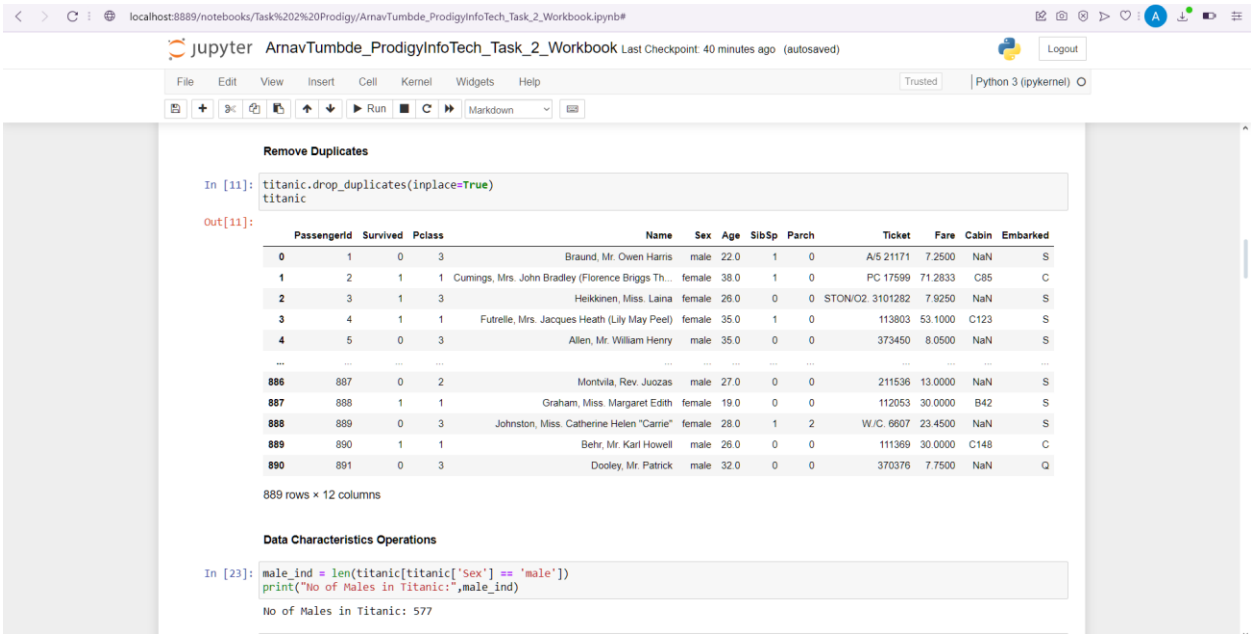
Data type conversions if necessary

In [10]:
```
titanic['Fare'] = titanic['Fare'].astype(float)
titanic
```

Out[10]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	28.0	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

889 rows × 12 columns



EDA Performing

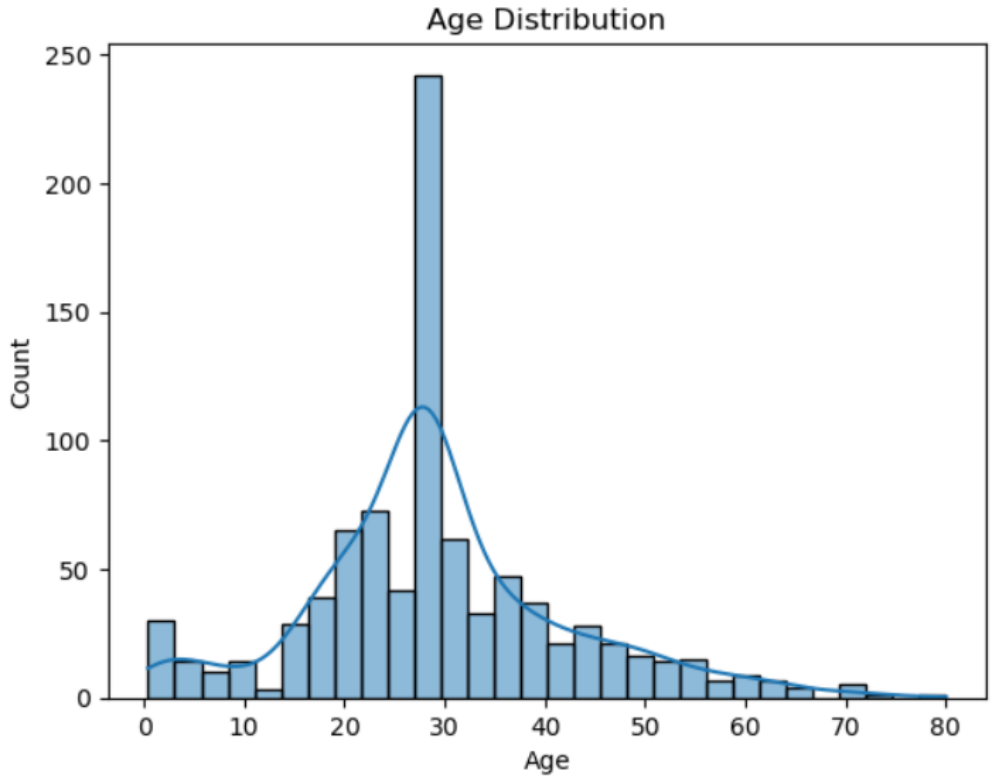
In [8]:
```
titanic.describe()
```

Out[8]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	889.000000	889.000000	889.000000	889.000000	889.000000	889.000000	889.000000
mean	446.000000	0.382452	2.311586	29.315152	0.524184	0.382452	32.096681
std	256.998173	0.486260	0.834700	12.984932	1.103705	0.806761	49.697504
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	224.000000	0.000000	2.000000	22.000000	0.000000	0.000000	7.895800
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.000000	1.000000	3.000000	35.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Screenshots of Visualizations:

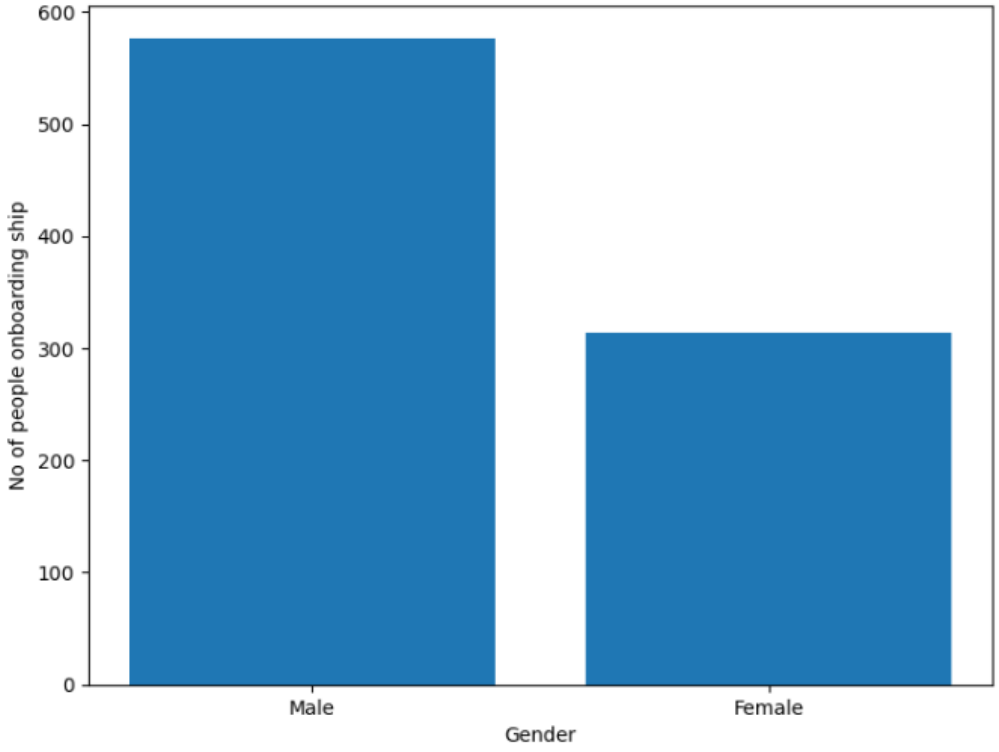
In [16]:
```
sns.histplot(titanic['Age'], bins=30, kde=True)
plt.title('Age Distribution')
plt.show()
```



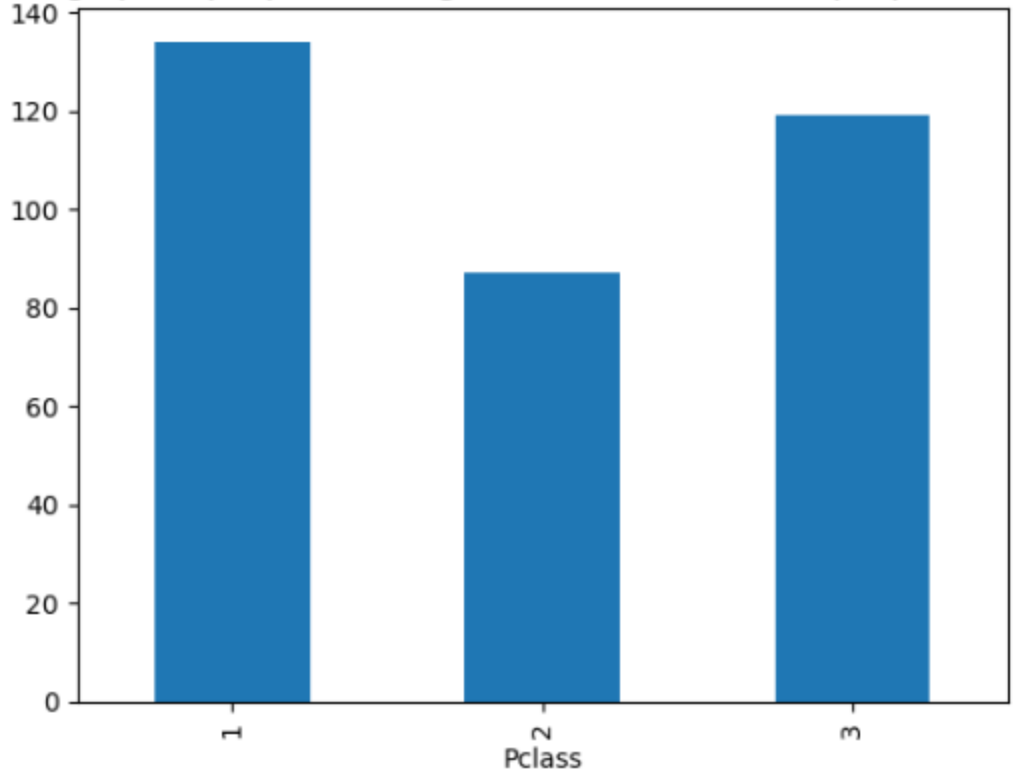
Visualization Using Seaborn and MatplotLib

In [15]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

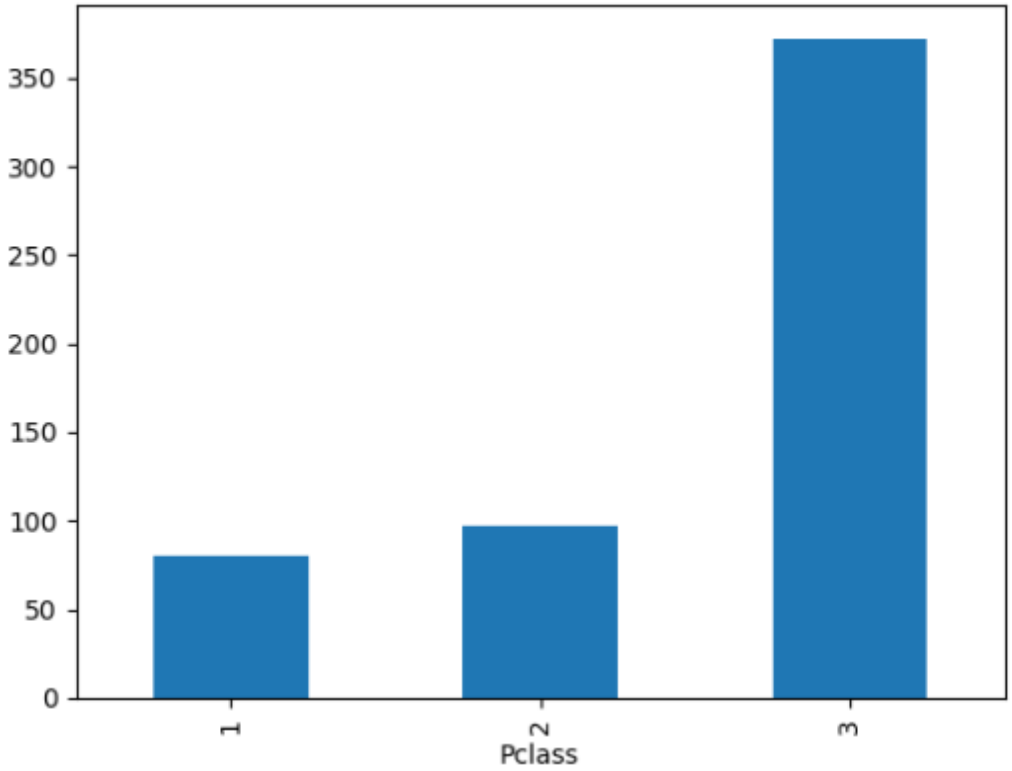
In [33]:
```
fig = plt.figure()
ax = fig.add_axes([0,0,1,1])
gender = ['Male','Female']
index = [577,314]
ax.bar(gender,index)
plt.xlabel("Gender")
plt.ylabel("No of people onboarding ship")
plt.show()
```

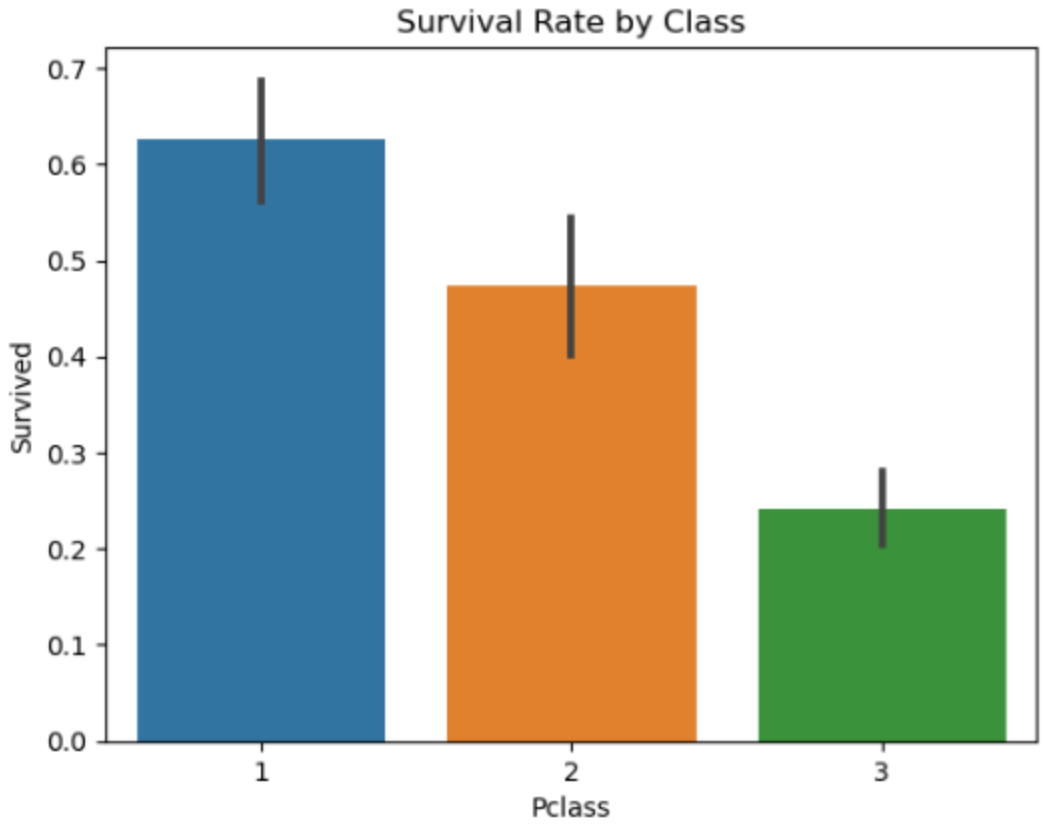


Bar graph of people accrding to ticket class in which people survived

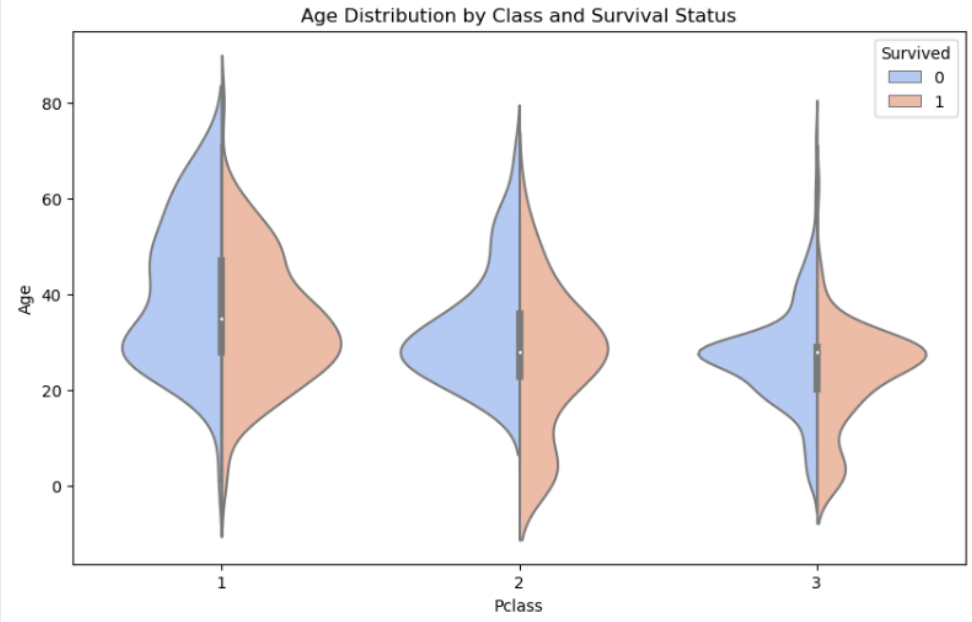


Bar graph of people accrding to ticket class in which people couldn't survive





In [22]:
```
# Violin plot for Age distribution by class and survival status
plt.figure(figsize=(10, 6))
sns.violinplot(x='Pclass', y='Age', hue='Survived', data=titanic, split=True, palette='coolwarm')
plt.title('Age Distribution by Class and Survival Status')
plt.show()
```



a. About the Dataset

For this task, I used the Titanic dataset from Kaggle. This dataset contains information about the passengers aboard the Titanic, which sank in 1912. The dataset includes the following columns:

- **PassengerId**: A unique identifier for each passenger.
- **Survived**: Survival status (0 = No, 1 = Yes).
- **Pclass**: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd).
- **Name**: The name of the passenger.
- **Sex**: The gender of the passenger.
- **Age**: The age of the passenger.
- **SibSp**: The number of siblings or spouses aboard the Titanic.
- **Parch**: The number of parents or children aboard the Titanic.
- **Ticket**: The ticket number.
- **Fare**: The fare paid for the ticket.
- **Cabin**: The cabin number.
- **Embarked**: The port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton).

b. Explanation of the Concept

The primary objective of this task was to perform data cleaning and exploratory data analysis (EDA) to understand the relationships between different variables in the dataset and identify patterns and trends. Data cleaning involves handling missing values, correcting data types, and filtering out irrelevant data. EDA includes visualizing data distributions, exploring relationships between variables, and identifying significant trends.

c. Outcome of the Analysis

1. **Handling Missing Values**: I identified and handled missing values in the 'Age', 'Cabin', and 'Embarked' columns. Missing 'Age' values were filled with the median age, missing 'Embarked' values with the mode, and the 'Cabin' column was dropped due to a high percentage of missing values.
2. **Distribution of Passengers by Class**: A count plot showed the distribution of passengers across the three ticket classes. Most passengers were in the 3rd class, followed by the 1st and 2nd classes.
3. **Survival Rate by Gender**: A bar chart highlighted the survival rate for males and females. The survival rate for females was significantly higher than for males.
4. **Age Distribution**: A histogram showed the age distribution of passengers. The distribution revealed that most passengers were between 20 and 40 years old.
5. **Fare Distribution**: A KDE plot indicated that the majority of passengers paid lower fares, with a few outliers who paid significantly higher fares.

6. **Correlation Matrix**: A heatmap of the correlation matrix revealed relationships between numerical variables. For example, 'Pclass' and 'Fare' had a negative correlation, indicating higher class passengers paid higher fares.
7. **Survival Rate by Class and Embarkation Point**: A heatmap showed the survival rate for passengers based on their class and port of embarkation. This revealed that 1st class passengers had a higher survival rate, particularly those who embarked from Cherbourg.

d. Conclusion

- **Handling Missing Data**: Addressing missing values is crucial for accurate analysis. Different techniques, such as filling with median/mode or dropping columns, were applied based on the context.
- **Passenger Class Distribution**: Most passengers traveled in the 3rd class, reflecting the socio-economic status of the majority of passengers.
- **Gender and Survival Rate**: The analysis revealed a significant gender disparity in survival rates, with females having a higher likelihood of survival.
- **Age and Fare Distributions**: Understanding the age and fare distributions helps in identifying the demographic and economic profile of the passengers.
- **Correlations and Survival Trends**: The correlation matrix and survival analysis by class and embarkation point provided insights into the factors influencing survival rates, such as class and port of embarkation.

These visualizations and analyses help uncover important patterns and trends, providing a deeper understanding of the Titanic dataset and the factors affecting passenger survival.