TASK 5: EXPLORATORY DATA ANALYSIS (EDA)

Objective: Extract insights using visual and statistical exploration.

Tools: Python (Pandas, Matplotlib, Seaborn)

Deliverables: Jupyter Notebook + PDF report of findings

A: Use .describe(), .info(), .value_counts()

```
import pandas as pd
train = pd.read_csv('C:\\Users\\HP\\Downloads\\train.csv')
test = pd.read_csv('C:\\Users\\HP\\Downloads\\test.csv')
gender_submission = pd.read_csv('C:\\Users\\HP\\Downloads\\gender_submission.csv')
```

```
# Import libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the data
train = pd.read_csv('C:\\Users\\HP\\Downloads\\train.csv')

# Display first few rows
train.head()
```

	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

TASK 5: EXPLORATORY DATA ANALYSIS (EDA)

```
train.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
```

```
train.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

```
]: train['Survived'].value_counts()
```

```
]: Survived
0    549
1    342
Name: count, dtype: int64
```

```
]: train['Sex'].value_counts()
```

```
]: Sex
male      577
female    314
Name: count, dtype: int64
```

```
]: train['Pclass'].value_counts()
```

```
]: Pclass
3    491
1    216
2    184
Name: count, dtype: int64
```

Observations:

- 891 records with 12 columns.
- Age, Cabin, Embarked have missing values.
- Sex is slightly unbalanced (more males).
- Pclass 3 (economy) has the most passengers.

- Fare prices have a large range (some very expensive tickets).
- Average passenger age is around 30 years.
- About 38% survived the Titanic disaster.

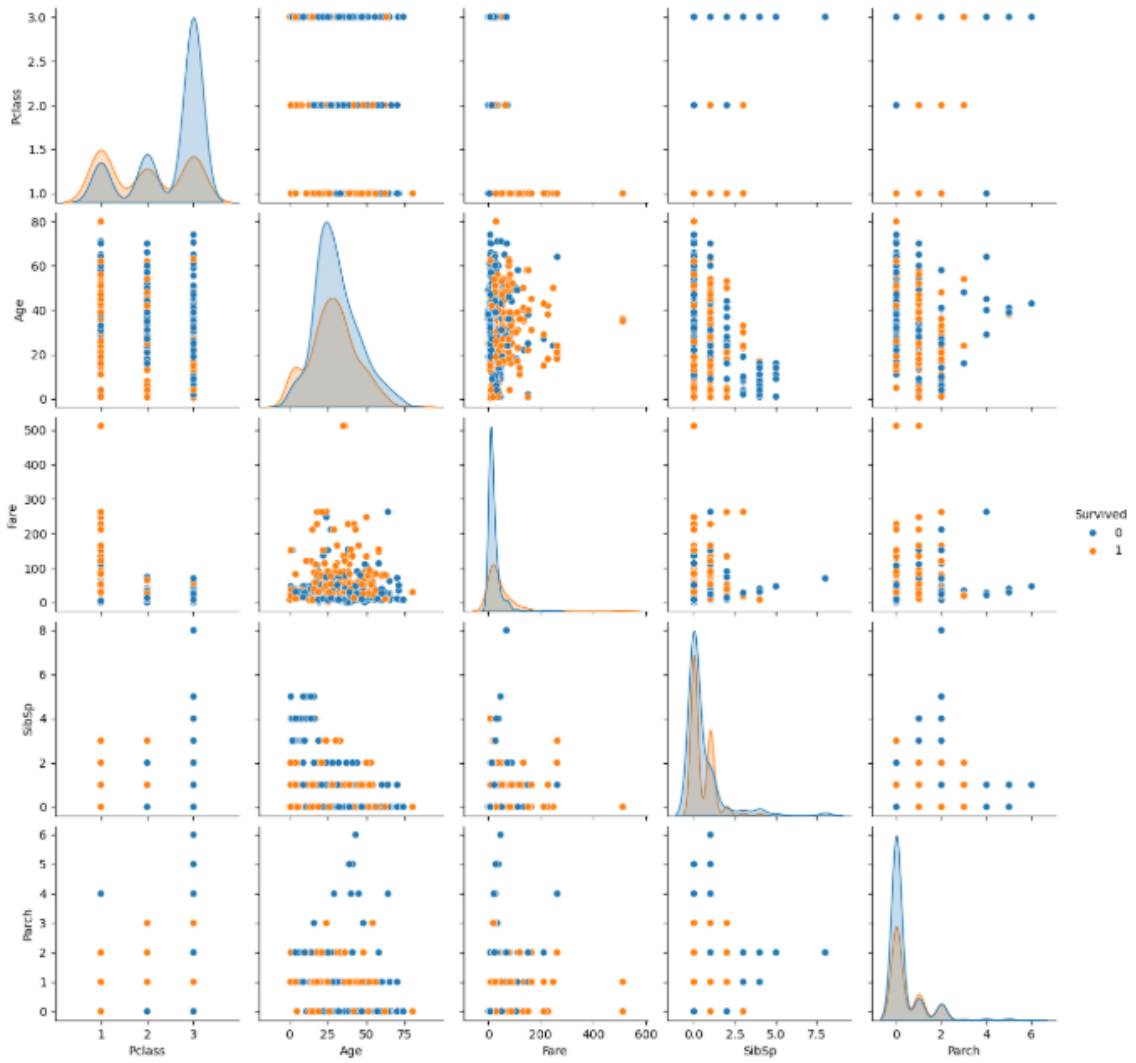
Summary

- Dataset is relatively clean but needs treatment for missing values (Age, Cabin, Embarked).
- Skew towards males and lower class passengers (Pclass 3).
- Survival rate is low (~38%) overall.

b. Use sns.pairplot(), sns.heatmap() for visualization

TASK 5: EXPLORATORY DATA ANALYSIS (EDA)

[26]:
```
sns.pairplot(train[['Survived', 'Pclass', 'Sex', 'Age', 'Fare', 'SibSp', 'Parch']], hue='Survived')
plt.show()
```



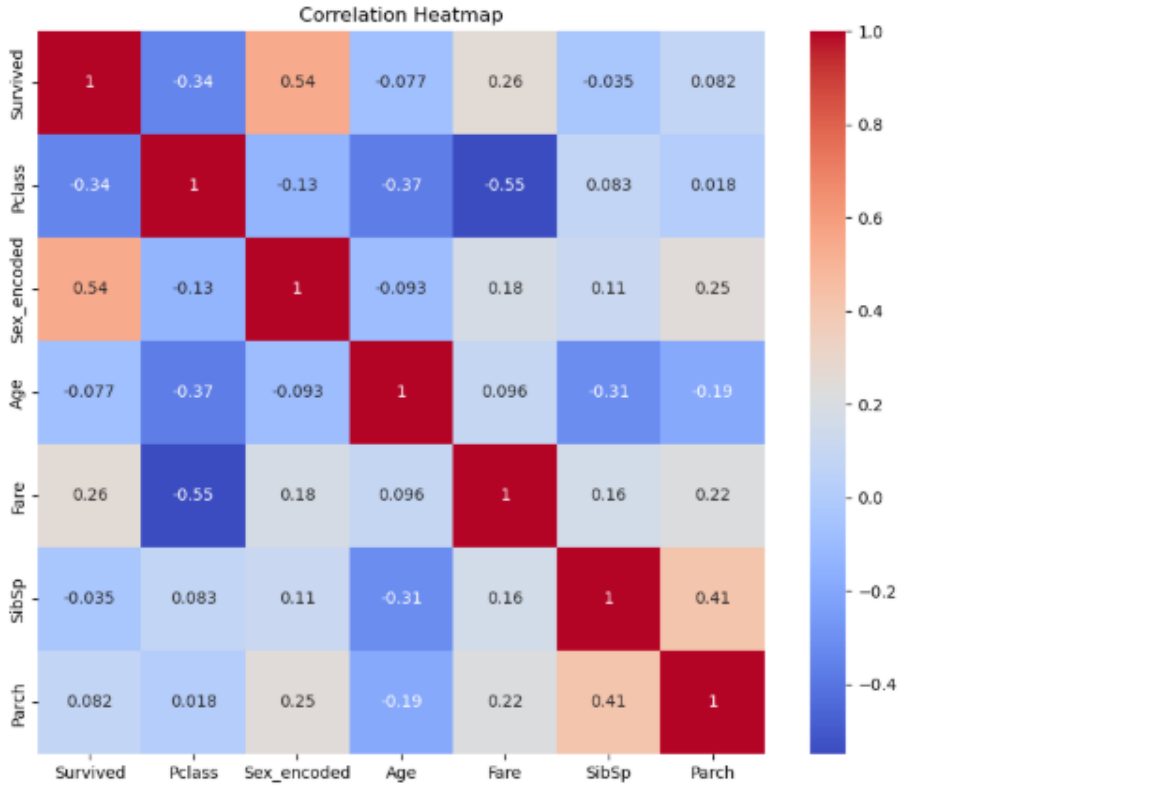
TASK 5: EXPLORATORY DATA ANALYSIS (EDA)

```
28]: train['Sex_encoded'] = train['Sex'].map({'male': 0, 'female': 1})

sns.pairplot(train[['Survived', 'Pclass', 'Sex_encoded', 'Age', 'Fare', 'SibSp', 'Parch']], hue='Survived')
plt.show()
```

TASK 5: EXPLORATORY DATA ANALYSIS (EDA)

```
[38]: plt.figure(figsize=(10,8))
sns.heatmap(train[['Survived', 'Pclass', 'Sex_encoded', 'Age', 'Fare', 'SibSp', 'Parch']].corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```



Observations:

- Sex has a strong correlation with survival (being female increased survival chances).
- Pclass (passenger class) negatively correlated with survival (lower classes had lower survival).
- Fare positively correlates with survival — richer people survived more.
- SibSp and Parch have mild impacts on survival.
- Age is weakly correlated with survival.

Summary:

- Gender and ticket class are strong predictors of survival.
- Wealthier and female passengers had a survival advantage.

c. Identify relationships and trends

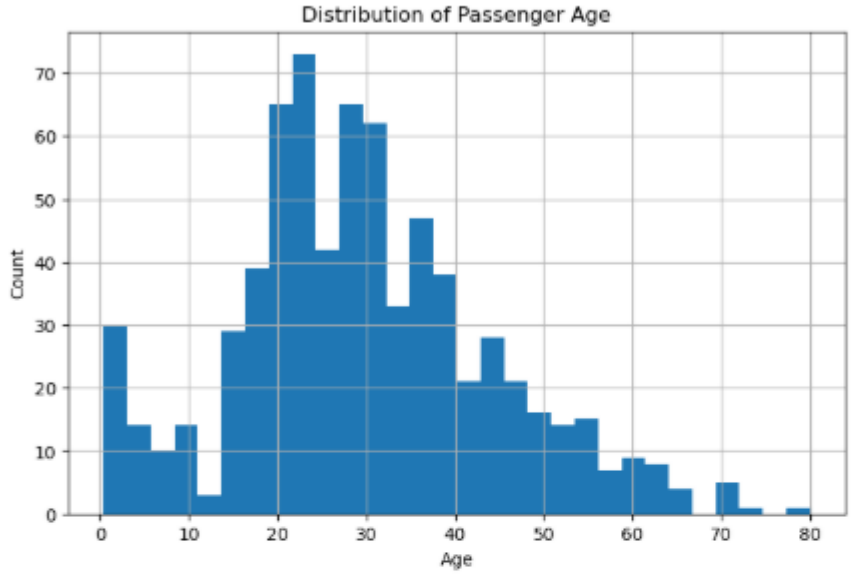
based on plots and correlations

- Females were much more likely to survive.
- Passengers in 1st class were much more likely to survive than 3rd class.
- Higher Fare = Higher survival chances.
- Young kids (very young ages) had better survival compared to middle-aged men.

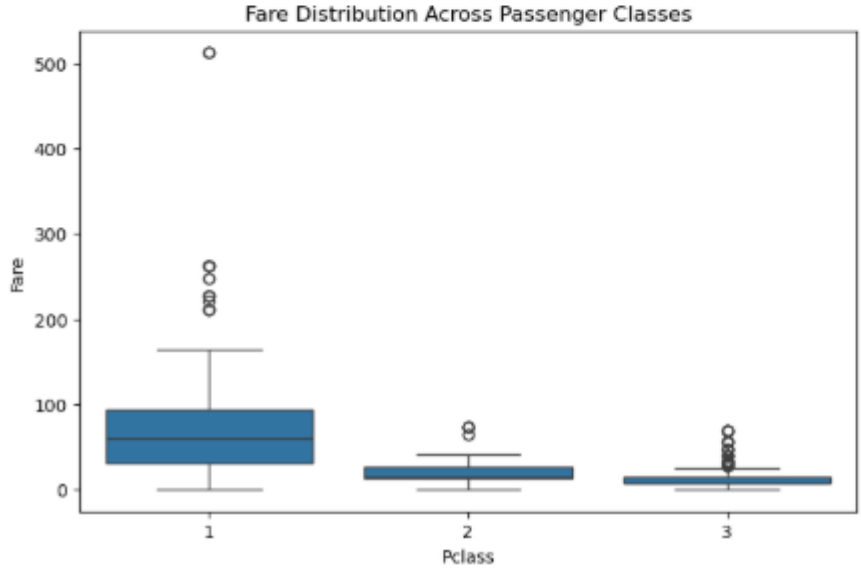
d. Plot histograms, boxplots, scatterplot

TASK 5: EXPLORATORY DATA ANALYSIS (EDA)

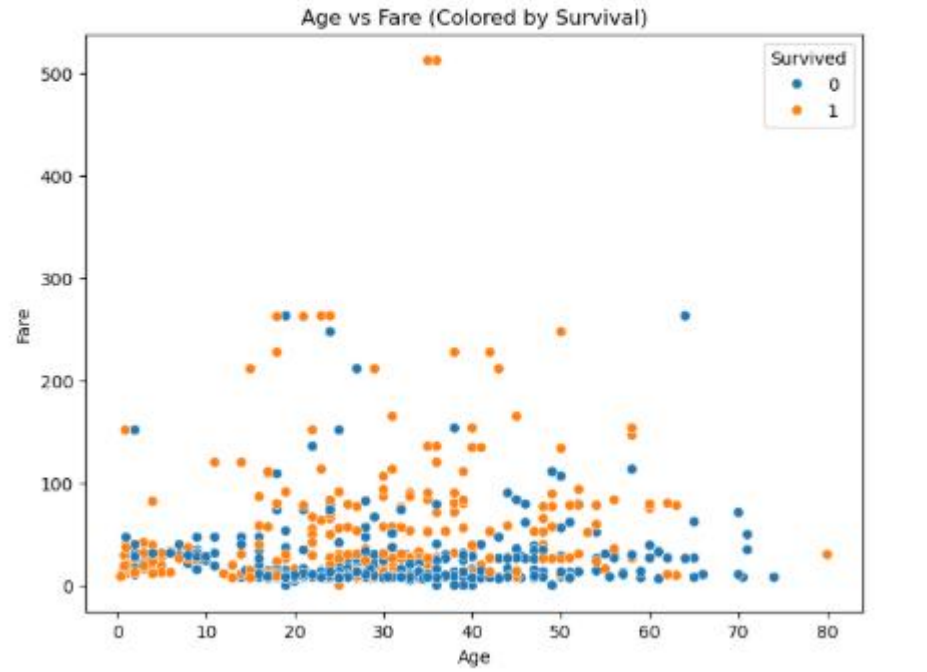
```
plt.figure(figsize=(8,5))
train['Age'].hist(bins=30)
plt.title('Distribution of Passenger Age')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()
```




```
plt.figure(figsize=(8,5))
sns.boxplot(x='Pclass', y='Fare', data=train)
plt.title('Fare Distribution Across Passenger Classes')
plt.show()
```



```
plt.figure(figsize=(8,6))
sns.scatterplot(x='Age', y='Fare', hue='Survived', data=train)
plt.title('Age vs Fare (Colored by Survival)')
plt.show()
```



Observation:

- High Fare passengers (regardless of age) had a higher chance of survival.
- Younger passengers who paid less had lower survival chance

Summary :

- Young adults made up the majority of passengers.
- Rich passengers had higher survival chances.
- Scatterplot shows survival is related to both Age and Fare, but not linearly.

FINAL OVERALL SUMMARY:

Missing Values:	Age, Cabin, and Embarked need data cleaning
Gender:	Females survived at a much higher rate
Passenger Class:	1st class passengers survived the most
Age:	Younger children survived slightly better
Fare:	Higher ticket fare correlates with survival
Trends:	Being wealthy and female increased survival chances