```
# Exploratory Data Analysis (EDA) on Titanic.

# Step 1: Import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline

# Step 2: Load data
df = pd.read_csv('Titanic.csv')
print("Dataset loaded successfully!")
```

```
Dataset loaded successfully!
```

```
# Step 3: First look at the data
print("Shape:", df.shape)
df.head()
```

```
Shape: (891, 12)
```

	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

```
# Step 4: Basic info
df.info()
```

```
<class'.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
dtype: float64(2), int64(5), object(5)
memory usage: 83.7 KB
```

```
# Step 5: Check missing values
df.isnull.sum()
```

	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

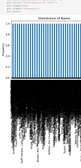
```
# Step 6: Explore the Survival Column
df['Survived'].value_counts
```

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

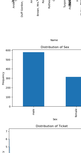
```
# Step 7: Percentage of survival
df['Survived'].value_counts(normalize=True)*100
```

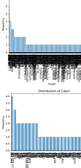
```
Survived
0    61.616162
1    38.383838
Name: count, dtype: float64
```

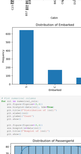
```
# Step 8: Visualize the target
plt.figure(figsize=(6,4))
sns.countplot(x='Survived', data=df)
plt.title('Survival Count')
plt.show()
```

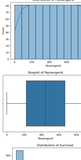


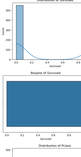

```
# Step 9: Numerical features
num = ['Survived', 'Pclass', 'Age', 'SibSp', 'ParCh', 'Fare']
for col in num:
    plt.figure(figsize=(6,4))
    sns.histplot(data=df, x=col, kde=True)
    plt.title('Distribution of {col}')
    plt.show()
```

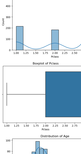




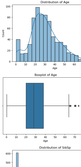


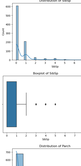




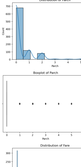


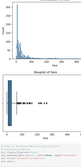

```
# Step 10: Categorical features
cat = ['Sex', 'Embarked']
for col in cat:
    plt.figure(figsize=(6,4))
    sns.countplot(x=col, hue='Survived', data=df)
    plt.title('Survival by {col}')
    plt.show()
```



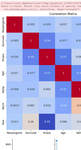



```
# Step 11: Correlation
plt.figure(figsize=(8,6))
sns.heatmap(df[num].corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

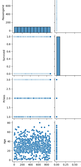





```
# Step 12: Survival by Sex
plt.figure(figsize=(6,4))
sns.barplot(x='Sex', y='Survived', data=df)
plt.title('Survival Rate by Sex')
plt.show
```




```
# Step 13: Survival by Pclass
plt.figure(figsize=(6,4))
sns.barplot(x='Pclass', y='Survived', data=df)
plt.title('Survival Rate by Pclass')
plt.show
```



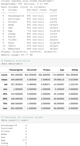

```
# Step 14: Survival by Age
plt.figure(figsize=(6,4))
sns.boxplot(x='Survived', y='Age', data=df)
plt.title('Survival by Age')
plt.show
```

```
# Step 15: Survival by Embarked
plt.figure(figsize=(6,4))
sns.barplot(x='Embarked', y='Survived', data=df)
plt.title('Survival Rate by Embarked')
plt.show
```

```
# Step 16: Survival by Sex and Pclass
plt.figure(figsize=(6,4))
sns.barplot(x='Sex', y='Survived', hue='Pclass', data=df)
plt.title('Survival by Sex and Pclass')
plt.show
```



```
# Step 17: Scatterplot: Passenger vs Survived
plt.figure(figsize=(6,4))
sns.scatterplot(x='PassengerId', y='Survived', data=df)
plt.title('Scatterplot: Passenger vs Survived')
plt.show
```



Observations for Each Visual

1. Categorical Columns (like Pclass) Sex

There are more males than females onboard

Embarked

Most passengers embarked from port S (Southampton)

Cabin

Most entries are missing for Cabin information

2. Numerical Columns (Histograms + Boxplots) Age

Age can mostly between 20 and 40 years.

Outliers exist (e.g., passengers aged 70–80)

Fare

Most fares are low (under 100), but there are some very high fares (over 500), indicating outliers.

SibSp (Siblings/Spouse aboard)

Most people travelled alone (0 siblings/spouses).

Parch (Parents/Children aboard)

Most passengers also had no parents or children with them.

Passenger Class (Pclass)

The largest group travelled in third class.

3. Correlation Matrix (Heatmap) Fare is negatively correlated with Pclass (lower class number → higher fare).

SibSp and Parch show some positive correlation (families tend to travel together).

Survived is somewhat correlated with Pclass and Fare (higher class and fare increased survival chances).

4. Pairplot Survived seems to have higher fares and belong more to first class.

Age is spread across both survivors and non-survivors.

5. Scatterplot (Passengers vs Survived) No visible relationship between PassengerId and Survival (as expected). ID is just an index.

Summary of Findings

Category: Finding Sentiangulation: Gives passengers were young males between 20-40 years old. Fare Class: Mainly travelled in 3rd class. Pclass: Most fares were low, but some high outliers exist. Family: Most passengers were travelling alone. Survival: Higher survival rates were linked to higher passenger classes and fares. Data Quality: Significant missing values in Age, Cabin, and Embarked columns. Correlation: Fare and Pclass are strongly negatively correlated.