IMPORT LIBRARIES

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(style="whitegrid")
```

LOAD THE DATASET

```
df=pd.read_csv('train.csv')
df.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500

Next steps: Generate code with df New interactive sheet

**BASIC DATA UNDERSTANDING **

```
df.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
```

SUMMARY STATISTICS

```
df.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

MISSING VALUES

```
df.isnull().sum()
```

	0
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

Value counts for categorical columns

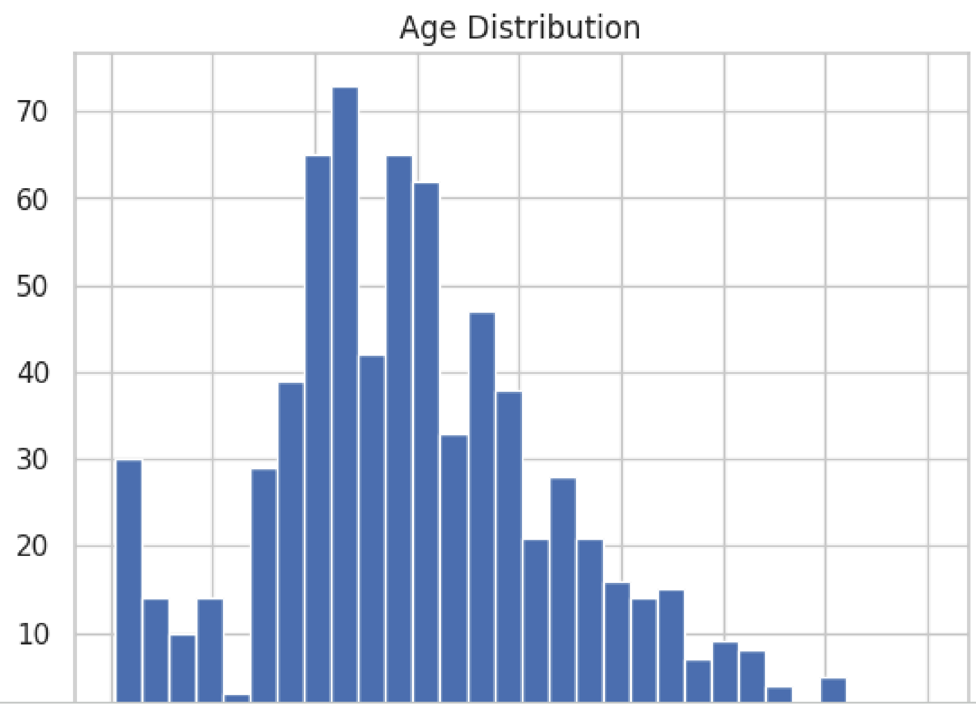
```
df['Survived'].value_counts()
df['Sex'].value_counts()
df['Pclass'].value_counts()
```

	count
Pclass	
3	491
1	216
2	184

dtype: int64

ANALYSIS HISTOGRAM

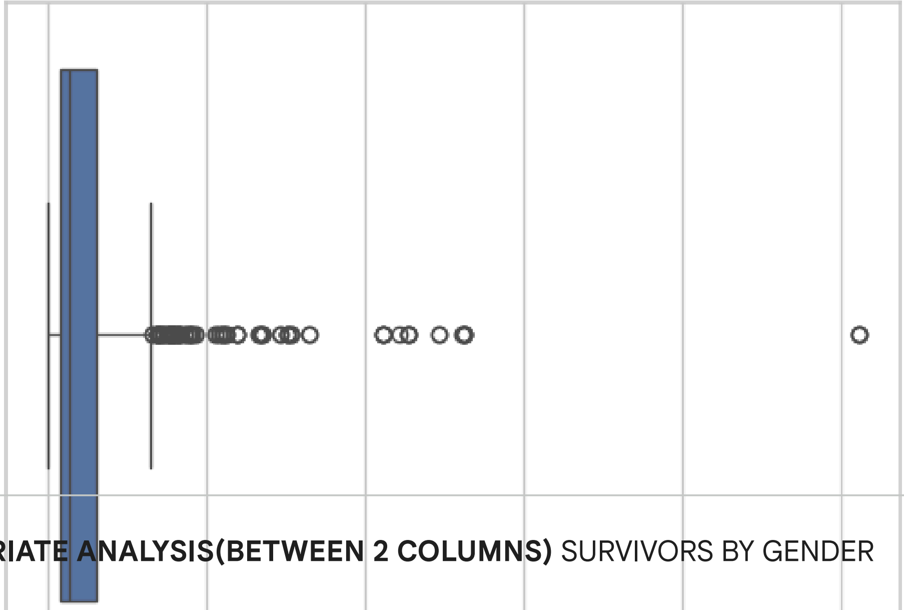
```
df["Age"].hist(bins=30)
plt.title("Age Distribution")
plt.show()
```



BOXPLOT

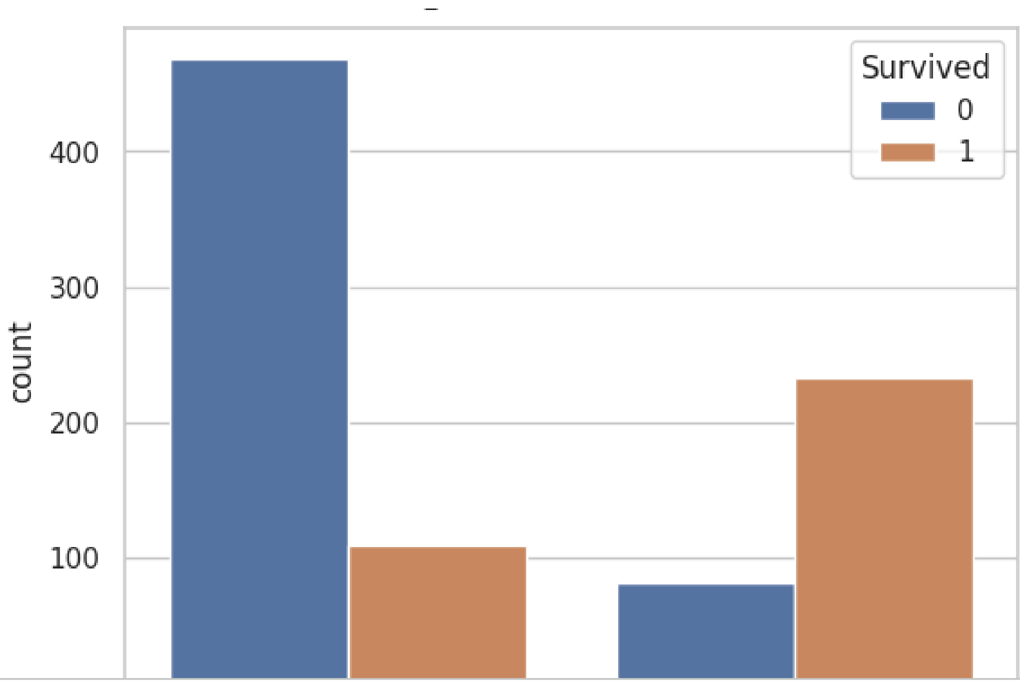
```
sns.boxplot(x=df["Fare"])
```

```
<Axes: xlabel='Fare'>
```



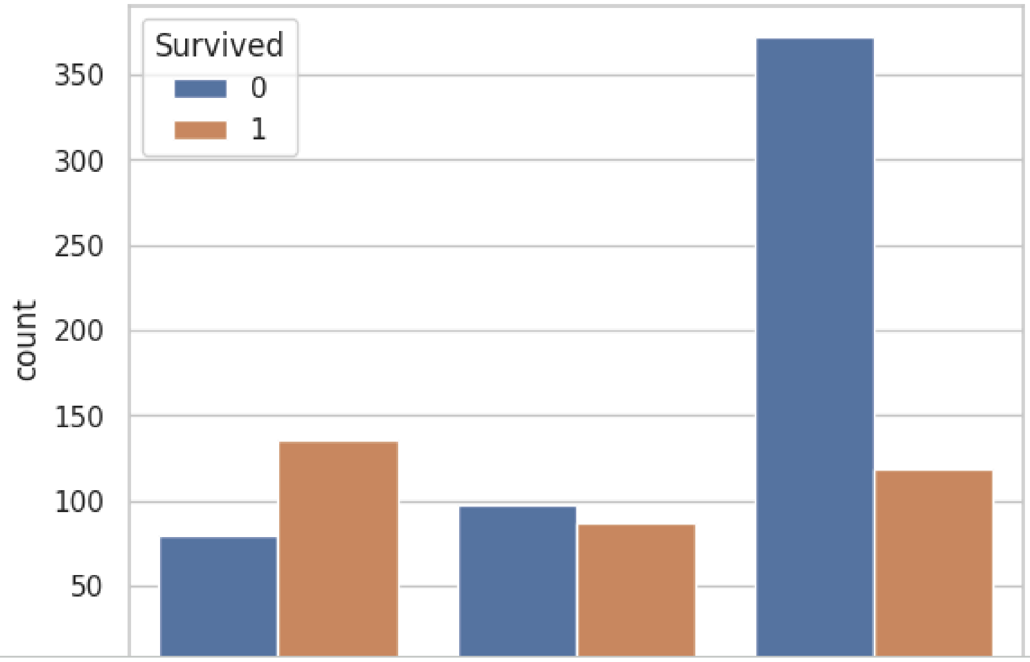
BIVARIATE ANALYSIS(BETWEEN 2 COLUMNS) SURVIVORS BY GENDER

```
sns.countplot(x="Sex",hue="Survived",data=df)
plt.show()
```



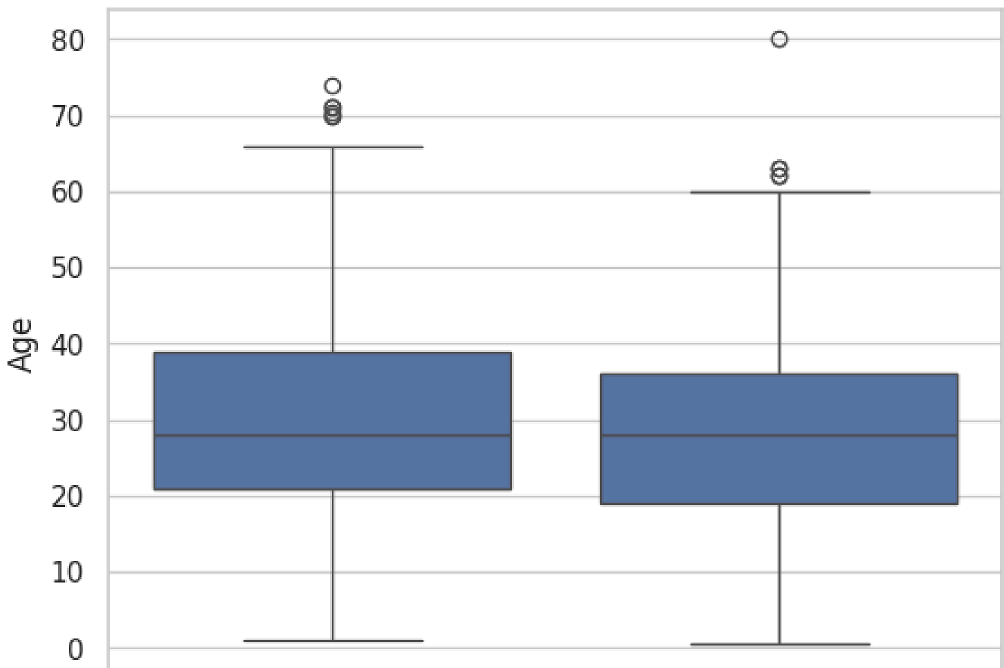
SURVIVORS BY CLASS

```
sns.countplot(x="Pclass",hue="Survived",data=df)
plt.show()
```



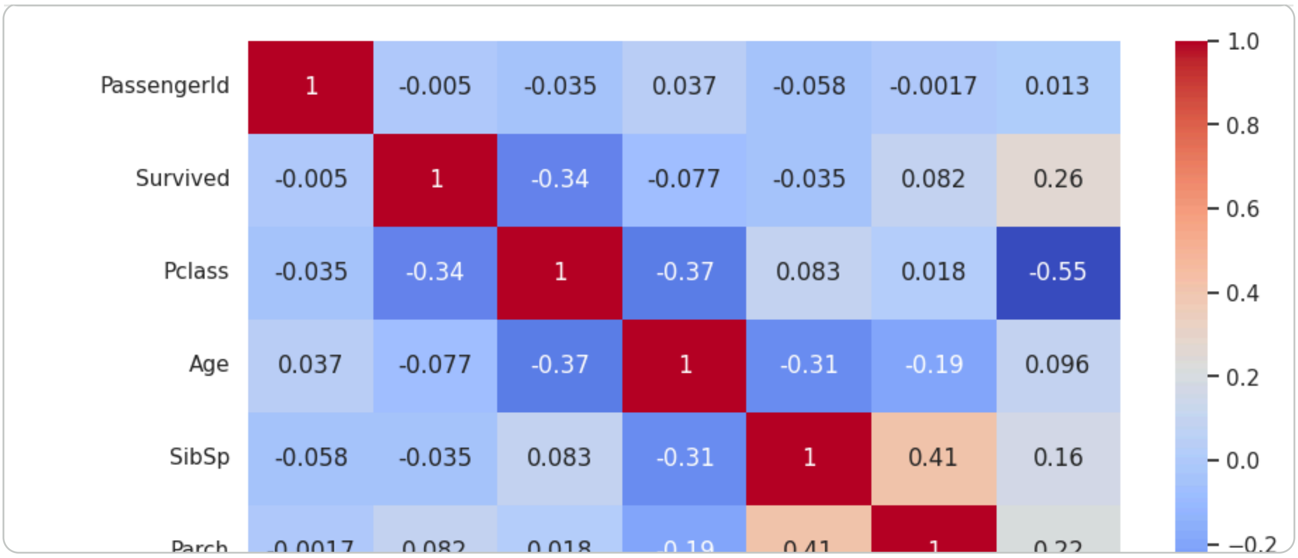
AGE VS SURVIVED

```
sns.boxplot(x="Survived",y="Age",data=df)
plt.show()
```



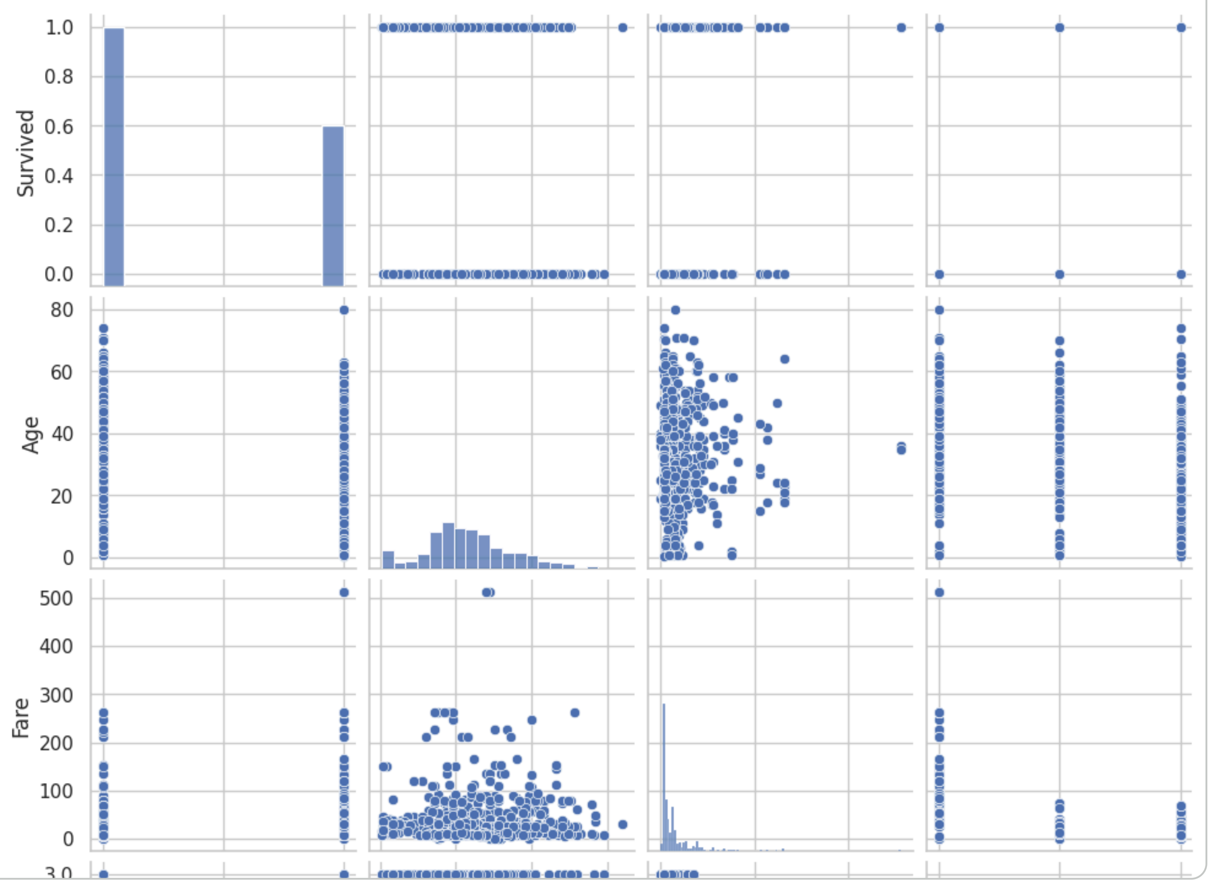
CORRELATION HEATMAP

```
plt.figure(figsize=(10,6))
sns.heatmap(df.select_dtypes(include=np.number).corr(),annot=True,cmap="coolwarm")
plt.show()
```



PAIRPLOT

```
sns.pairplot(df[['Survived','Age','Fare','Pclass']])
plt.show()
```



SUMMARY Exploratory Data Analysis (EDA) — Summary

OBJECTIVE:

To explore the Titanic dataset using statistical and visual analysis to identify patterns, trends, and factors affecting passenger survival.

Tools Used:

Python

Pandas

Matplotlib

Seaborn

COLAB Notebook

DATA UNDERSTANDING

Dataset contains 891 rows and 12 columns.

Important columns: Survived, Pclass, Sex, Age, Fare, Embarked.

Missing values found in Age, Cabin, and Embarked.

KEY ANALYSIS PERFORMED

1. Descriptive Statistics

Used .info(), .describe() to understand data types and distributions.

Identified missing data and basic summary metrics.

2. Univariate Analysis

Plotted histograms for Age, Fare.

Plotted boxplots to detect outliers.

Observed:

Age distribution is mostly between 20–40.

Fare values have high variance.

3. Bivariate Analysis

Survival by gender:

Females had much higher survival rate than males.

Survival by passenger class:

1st-class passengers survived more than 2nd and 3rd-class.

Age vs Survival:

Younger passengers had slightly better survival chance.

Fare vs Survival:

Higher fare → more survival (likely due to class difference).

4. Correlation Analysis

Heatmap showed:

Survival is negatively correlated with Pclass.

Fare and Survival are positively correlated.

Strong correlation between Fare and Pclass.

5. Pairplot

Visualized relationships between Age, Fare, Pclass, and Survival.

INSIGHTS & OBSERVATIONS

Gender is a major survival factor: Females survived significantly more.

Class also matters: 1st class passengers had best survival outcomes.

Economic status & fare: Higher fare passengers were safer.

Children had slightly better survival probability.

Many cabins missing → cannot use effectively for analysis.

CONCLUSION

The EDA shows that gender, passenger class, and ticket fare are the most influential factors in determining survival on the Titanic. Higher-class and female passengers had better chances of survival, showing possible priority in rescue operations.