```
 # installation of libraries

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Display settings
sns.set_style('whitegrid')
plt.style.use("fivethirtyeight")
```

```
# Step 2: Upload files

from google.colab import files
uploaded = files.upload()
```

Choose Files No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving train.csv to train (1).csv
```

```
# Step 3: Read the CSV file

df = pd.read_csv('/content/train.csv')

# First 5 rows
df.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

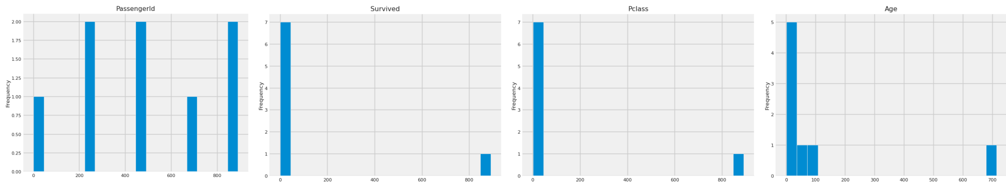
```
# Step 4a: General info about the dataset
df.info()
```

```
<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
```

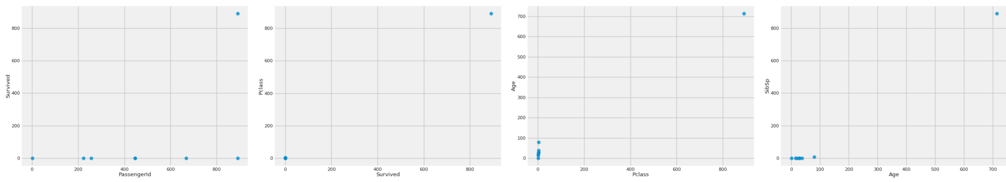
```
# Step 4b: Statistical summary
df.describe()
```

	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

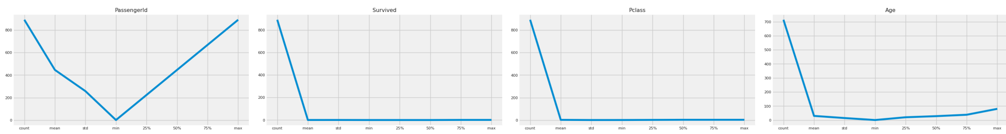
Distributions



2-d distributions



Values




```
# Step 4c: Value counts for categorical columns
print(df['Survived'].value_counts())
print(df['Pclass'].value_counts())
print(df['Sex'].value_counts())
```

```
Survived
0    549
1    342
Name: count, dtype: int64
Pclass
3    491
1    216
2    184
Name: count, dtype: int64
Sex
male      577
female    314
Name: count, dtype: int64
```

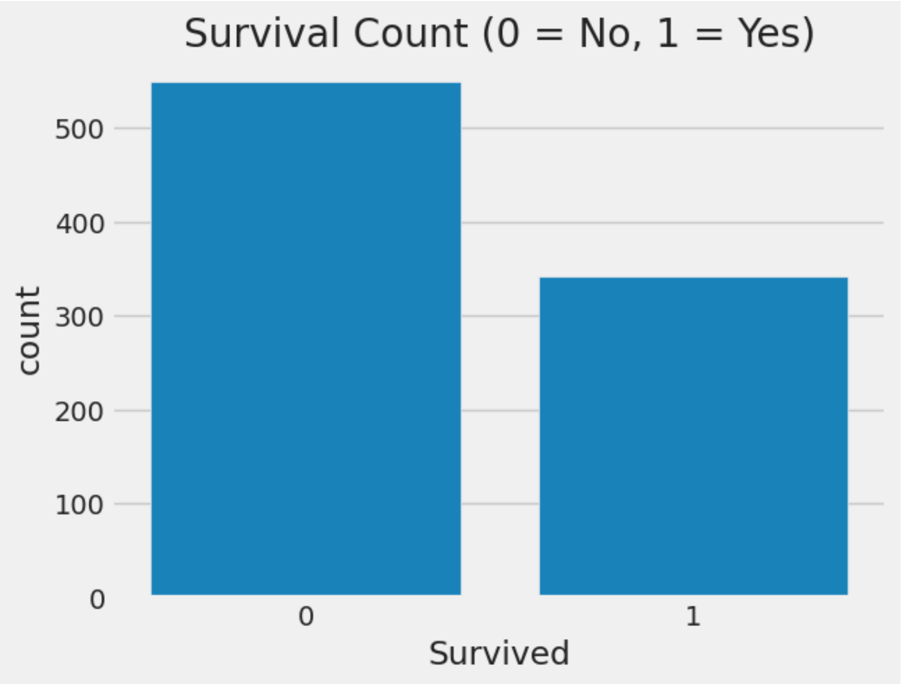
```
# Step 5: Check missing values
df.isnull().sum()
```

	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

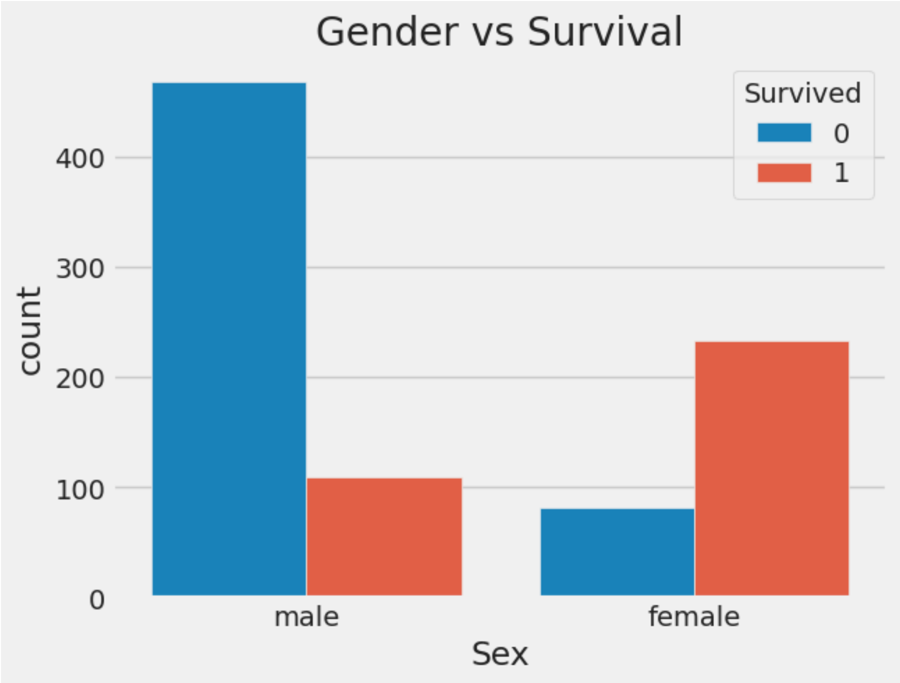
Columns like Age, Cabin, and Embarked will show missing values.

✅ Shows how passengers are spread by age.

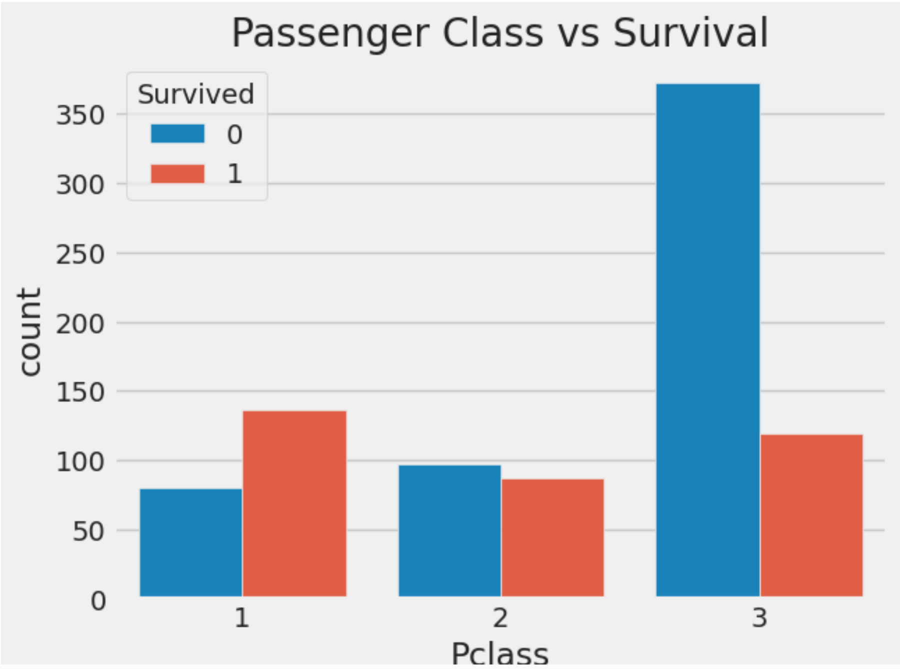
```
# Step 6b: Survival count
sns.countplot(data=df, x='Survived')
plt.title('Survival Count (0 = No, 1 = Yes)')
plt.show()
```




```
# Step 6c: Gender vs Survival
sns.countplot(data=df, x='Sex', hue='Survived')
plt.title('Gender vs Survival')
plt.show()
```




```
# Step 6d: Pclass vs Survival
sns.countplot(data=df, x='Pclass', hue='Survived')
plt.title('Passenger Class vs Survival')
plt.show()
```



Pairplot

```
# Step 7a: Pairplot
sample_df = df[['Survived', 'Pclass', 'Age', 'Fare']]
sns.pairplot(sample_df, hue='Survived')

plt.show()
```

Observation:

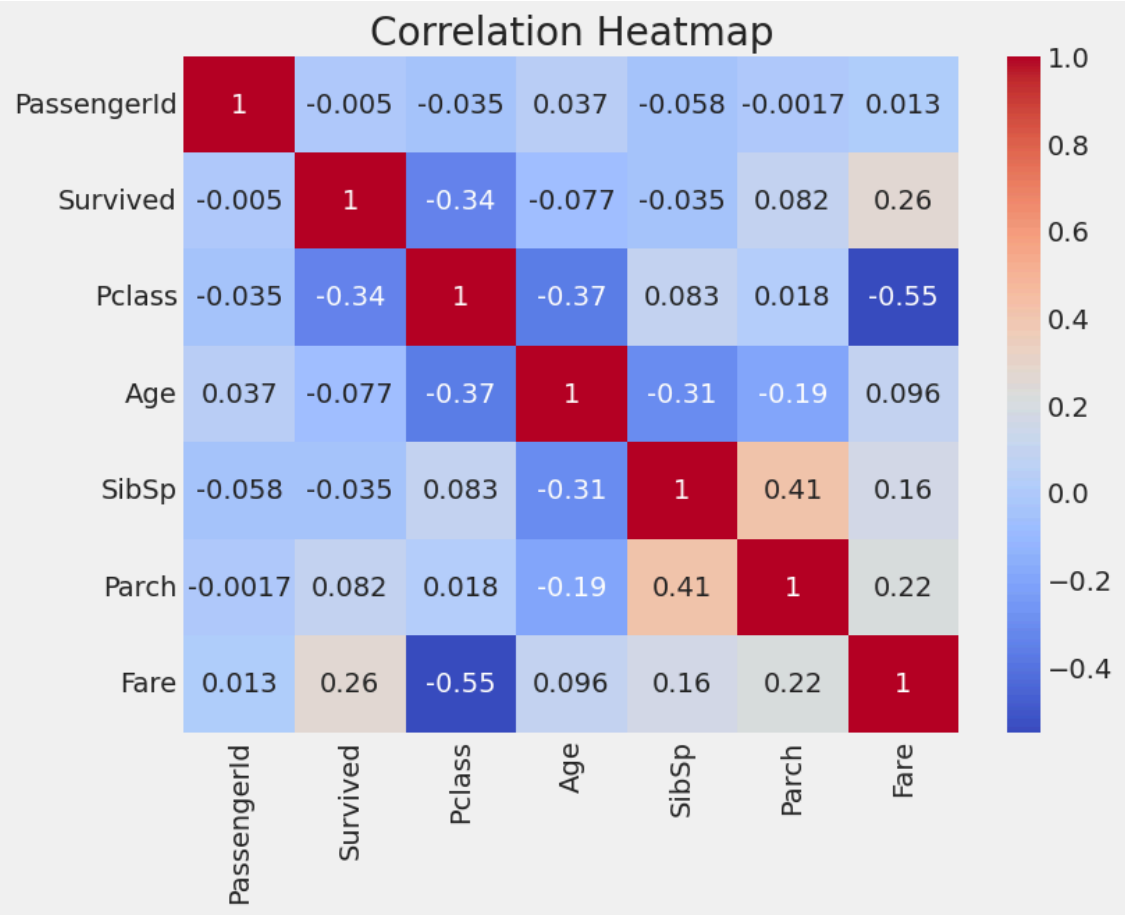
1st class passengers (lower Pclass value) had higher survival rates.

Higher Fare passengers had better survival rates.

Age distribution is more spread out and doesn't strongly separate survivors and non-survivors.

✅ Shows relationships between Age, Fare, Class and Survival.

```
# Step 7b: Correlation Heatmap
plt.figure(figsize=(8,6))
numeric_df = df.select_dtypes(include=np.number) # Select only numerical features for correlation
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```



Observation:

Fare and Pclass are strongly negatively correlated (r ≈ -0.55).

Survived has positive correlation with Fare and weak negative correlation with Pclass.

Age has almost no strong correlation with survival.

From the Pairplot:

Survived vs Fare: Passengers who paid higher Fare were more likely to survive.

Survived vs Pclass: Passengers from 1st class (Pclass = 1) survived more than 2nd or 3rd class.

Survived vs Age: Younger passengers show a slightly higher chance of survival.

From the Heatmap:

Fare and Pclass: Strong negative correlation (≈ -0.55) → Higher class (1st class) paid higher fares.

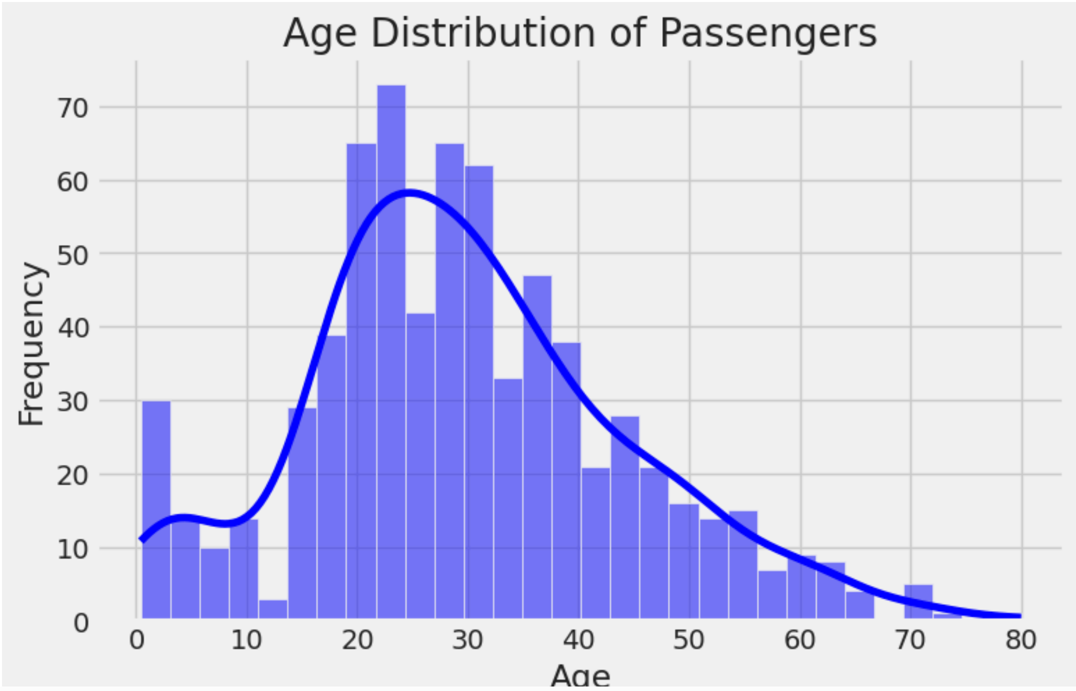
Survived and Pclass: Negative correlation with Pclass → Lower Pclass (higher number, like 3) → lower survival.

Survived and Fare: Positive correlation → Paying more = better chance of surviving.

Survived and Age: Very weak negative correlation — age doesn't strongly affect survival.

Histogram

```
# Histogram of Age
plt.figure(figsize=(8,5))
sns.histplot(df['Age'].dropna(), kde=True, bins=30, color='blue')
plt.title('Age Distribution of Passengers')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```

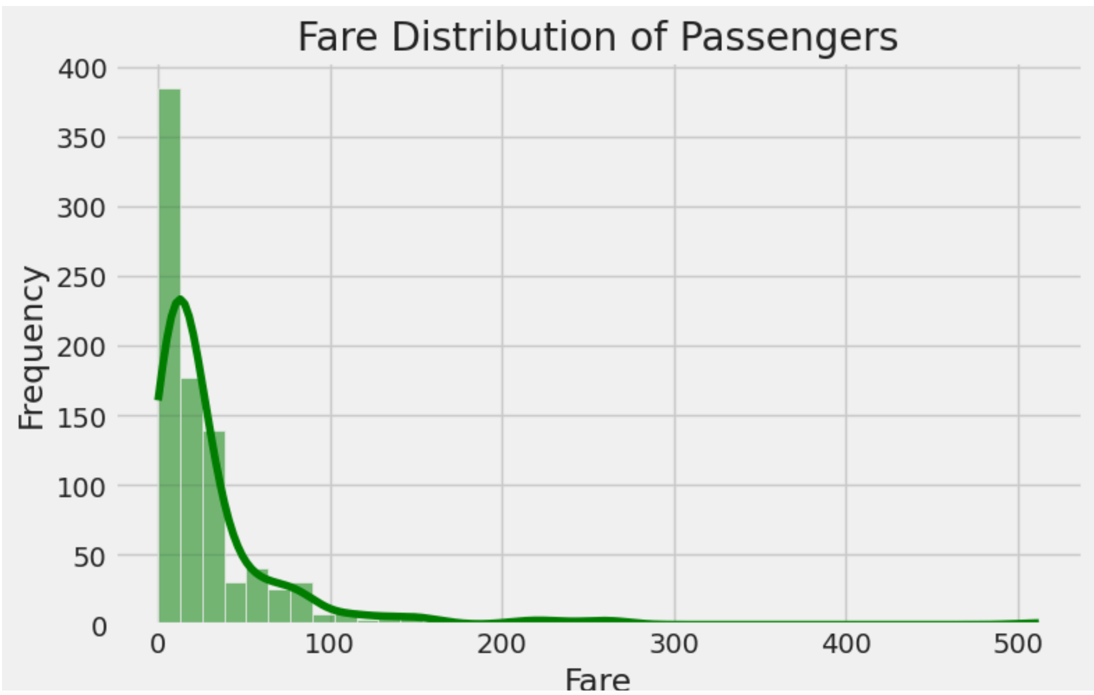


Observation:

Most passengers were aged between 20 to 40 years.

Very few passengers were either very young (below 10) or very old (above 60).

```
# Histogram of Fare
plt.figure(figsize=(8,5))
sns.histplot(df['Fare'], kde=True, bins=40, color='green')
plt.title('Fare Distribution of Passengers')
plt.xlabel('Fare')
plt.ylabel('Frequency')
plt.show()
```



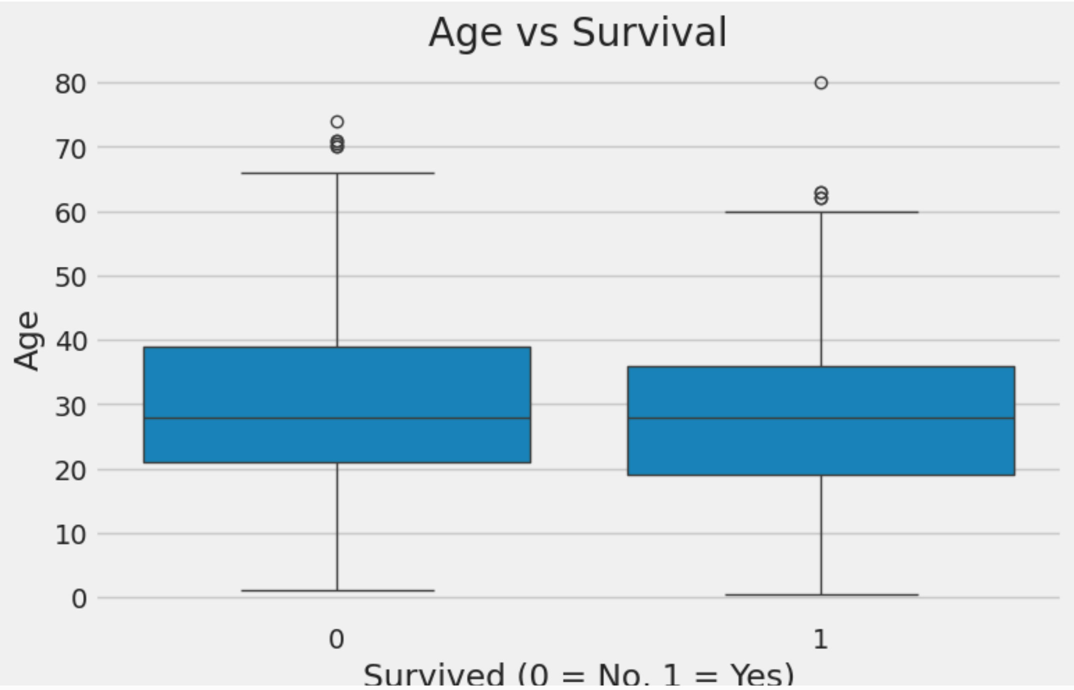
Observation:

Most passengers paid a fare between $0 and 100$.

There are a few passengers who paid extremely high fares (outliers above $200).

BoxPloat

```
# Boxplot of Age vs Survived
plt.figure(figsize=(8,5))
sns.boxplot(data=df, x='Survived', y='Age')
plt.title('Age vs Survival')
plt.xlabel('Survived (0 = No, 1 = Yes)')
plt.ylabel('Age')
plt.show()
```

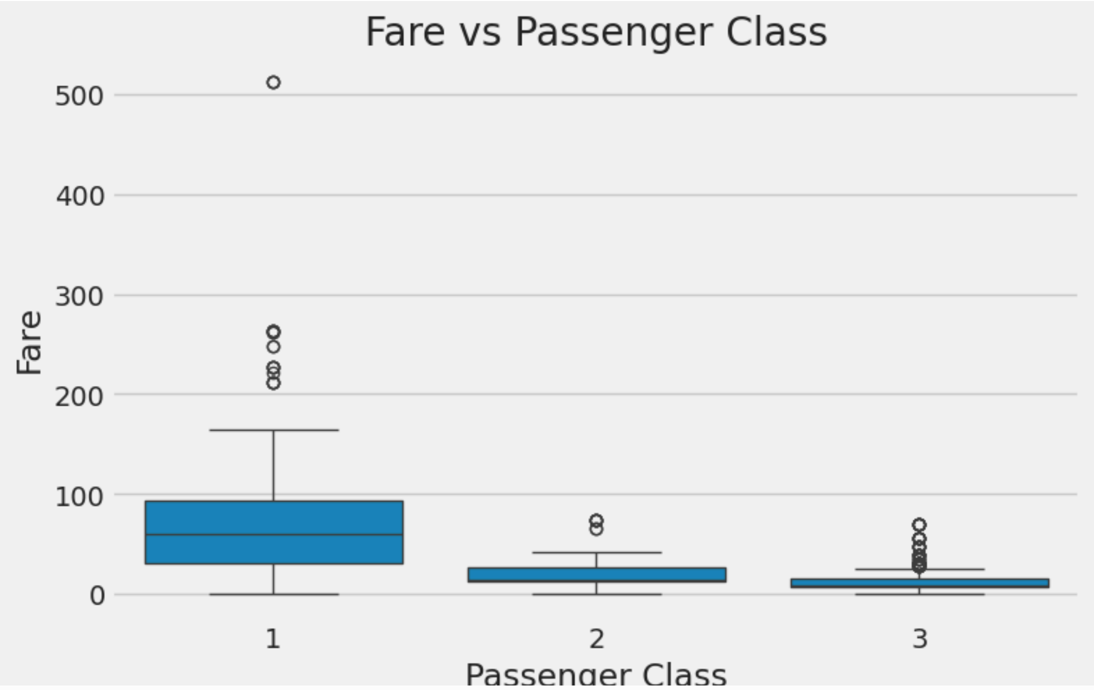


Observation:

The median age of survivors is slightly lower than that of non-survivors.

Younger passengers had a higher chance of surviving.

```
# Boxplot of Fare vs Pclass
plt.figure(figsize=(8,5))
sns.boxplot(data=df, x='Pclass', y='Fare')
plt.title('Fare vs Passenger Class')
plt.xlabel('Passenger Class')
plt.ylabel('Fare')
plt.show()
```



Observation:

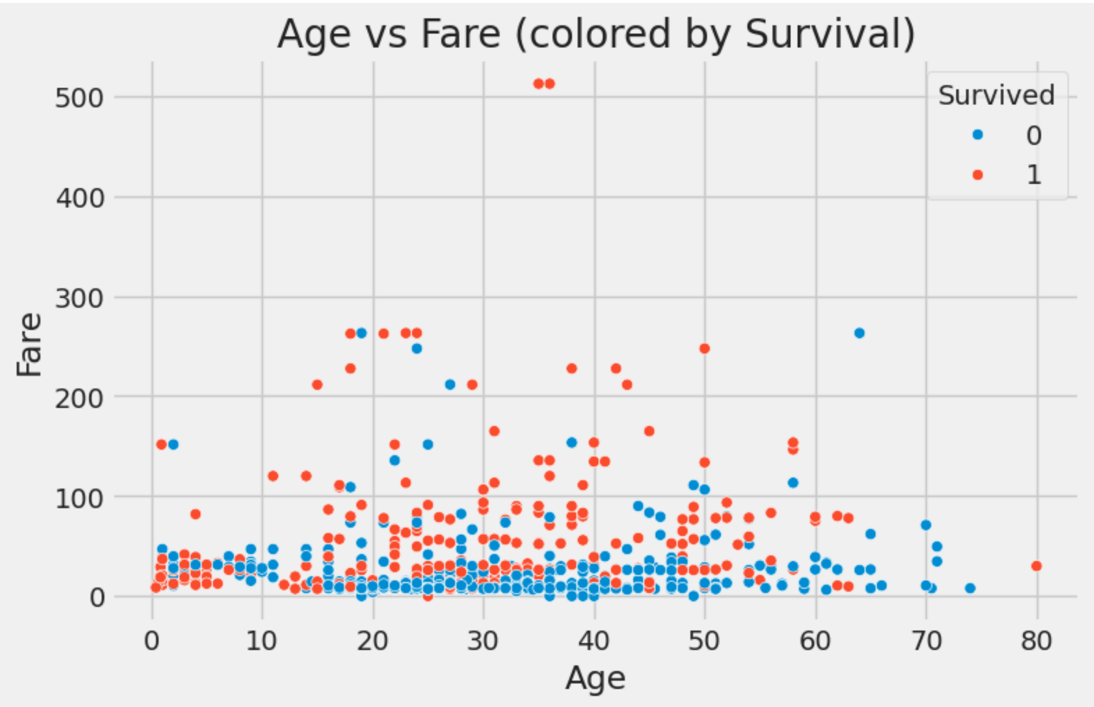
1st class passengers paid the highest fares on average.

3rd class passengers paid the lowest fares.

There are more outliers (very high fares) in 1st class.

ScatterPloat

```
# Scatterplot of Age vs Fare
plt.figure(figsize=(8,5))
sns.scatterplot(data=df, x='Age', y='Fare', hue='Survived')
plt.title('Age vs Fare (colored by Survival)')
plt.xlabel('Age')
plt.ylabel('Fare')
plt.legend(title='Survived')
plt.show()
```



Observation:

Passengers who paid higher fares were more likely to survive.

No strong pattern is visible between Age and Survival.

Some very young passengers (children) also had high survival rates

Histogram vs ScatterPloat observation