Could not connect to the reCAPTCHA service. Please check your internet connection and reload to get a reCAPTCHA challenge.

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

Choose Files train.csv

- **train.csv**(text/csv) - 61194 bytes, last modified: 11/8/2025 - 100% done

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

```
{'train.csv': 'train.csv'}
```

```
{'train.csv': 'train.csv'}
```

```
import pandas as pd

df = pd.read_csv("train.csv")
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	female	35.0	1	0	113803	53.1000	C123	S

Next steps: Generate code with df View recommended plots New interactive sheet

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load dataset
df = pd.read_csv("train.csv")

# Show first rows
df.head()

# Info and basic stats
df.info()
df.describe()
df.isnull().sum()  # Check missing values
```

```
<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
```

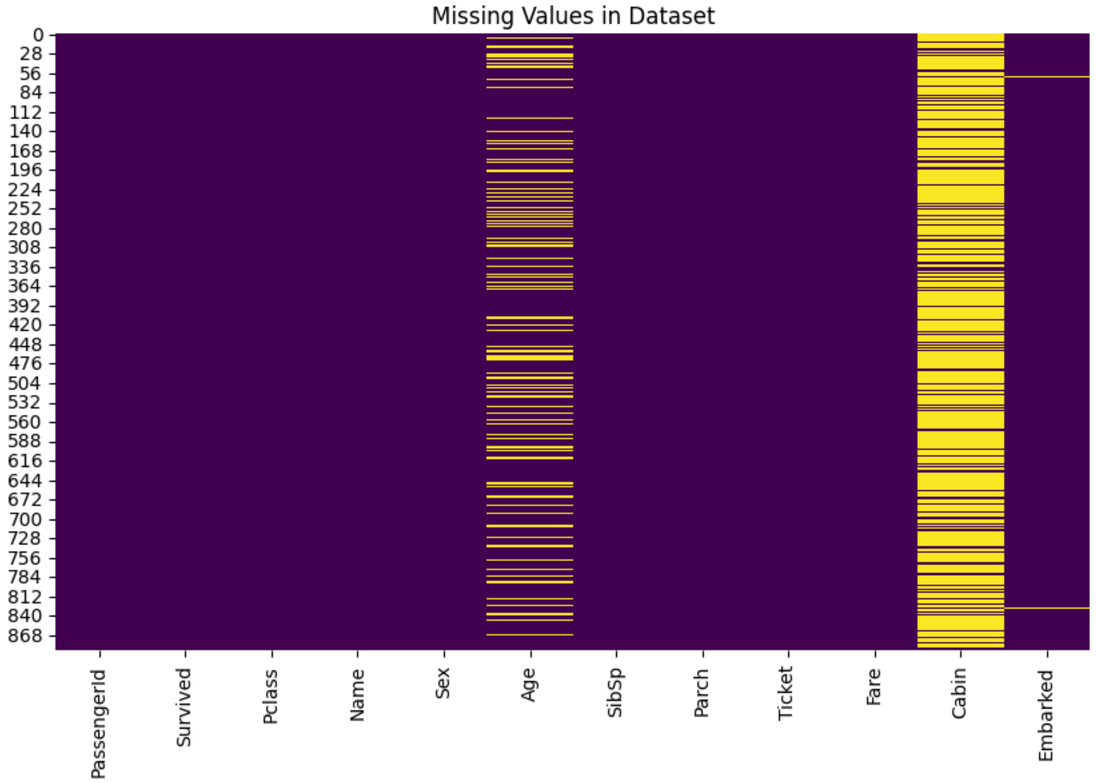
	0
PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	2

dtype: int64

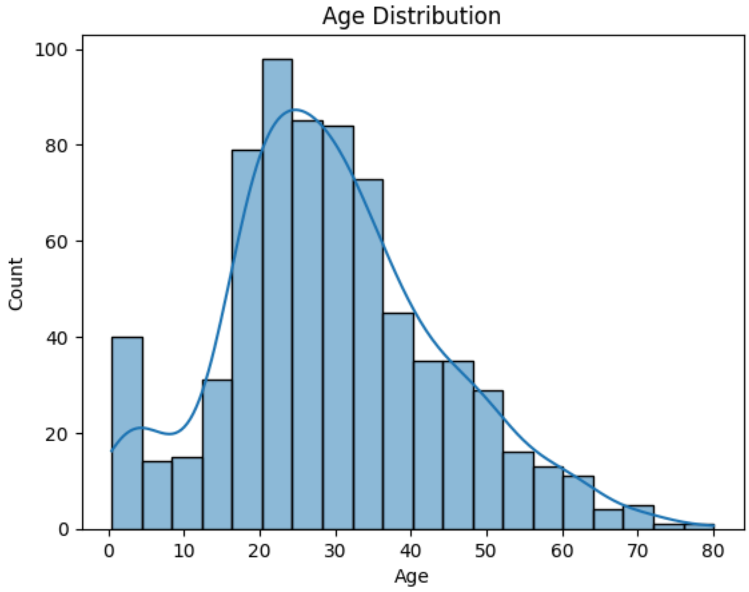
```
print("Sex:\n", df['Sex'].value_counts(), "\n")
print("Pclass:\n", df['Pclass'].value_counts(), "\n")
print("Embarked:\n", df['Embarked'].value_counts(), "\n")
```

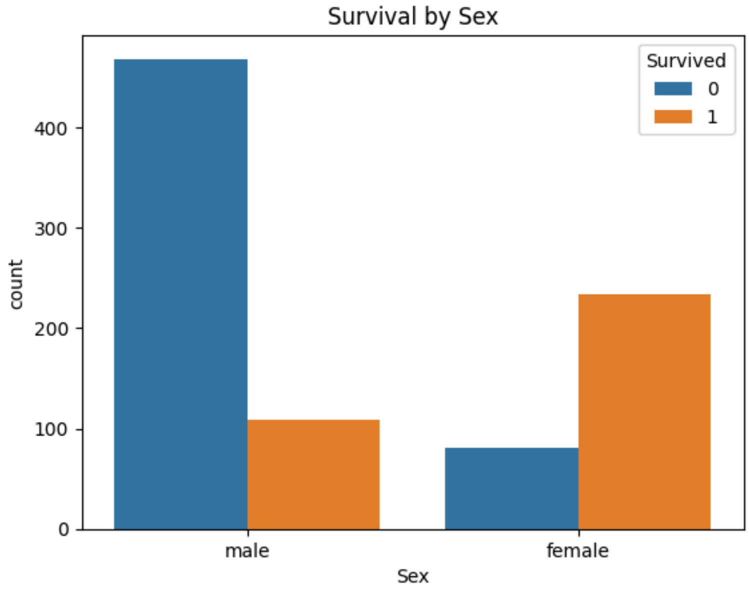
```
Sex:
 Sex
male      577
female    314
Name: count, dtype: int64 

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

Embarked:
 Embarked
S    644
C    168
Q     77
Name: count, dtype: int64 
```

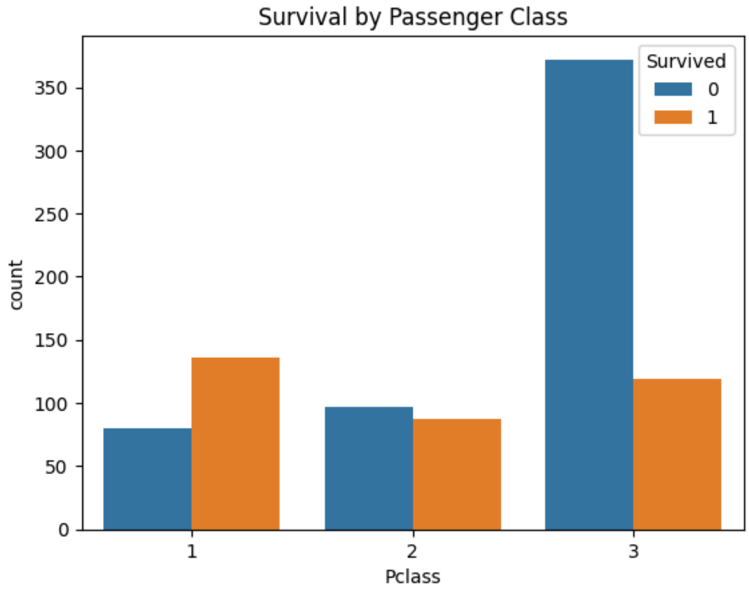
```
plt.figure(figsize=(10,6))
sns.heatmap(df.isnull(), cbar=False, cmap='viridis')
plt.title("Missing Values in Dataset")
plt.show()
```

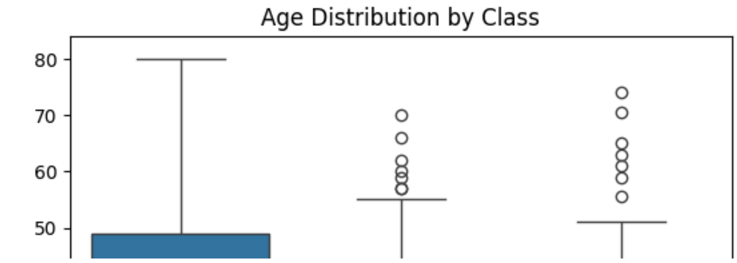


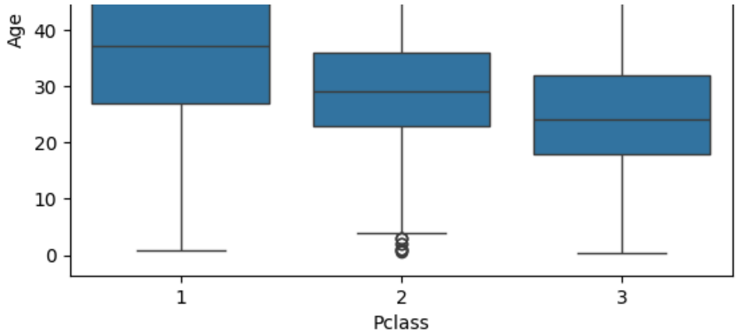
```
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder

# Age Distribution
sns.histplot(df['Age'], kde=True)
plt.title("Age Distribution")
plt.show()

# Survival by Sex
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title("Survival by Sex")
plt.show()

# Survival by Passenger Class
sns.countplot(x='Pclass', hue='Survived', data=df)
plt.title("Survival by Passenger Class")
plt.show()

# Age Distribution by Class
sns.boxplot(x='Pclass', y='Age', data=df)
plt.title("Age Distribution by Class")
plt.show()

# Correlation Heatmap (numeric only)
plt.figure(figsize=(8,6))
sns.heatmap(df.select_dtypes(include=['number']).corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap (Numeric Only)")
plt.show()

# Correlation Heatmap (with encoded categorical)
df_encoded = df.copy()
for col in df_encoded.select_dtypes(include=['object']).columns:
    df_encoded[col] = LabelEncoder().fit_transform(df_encoded[col].astype(str))

plt.figure(figsize=(8,6))
sns.heatmap(df_encoded.corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap (With Encoded Categories)")
plt.show()
```

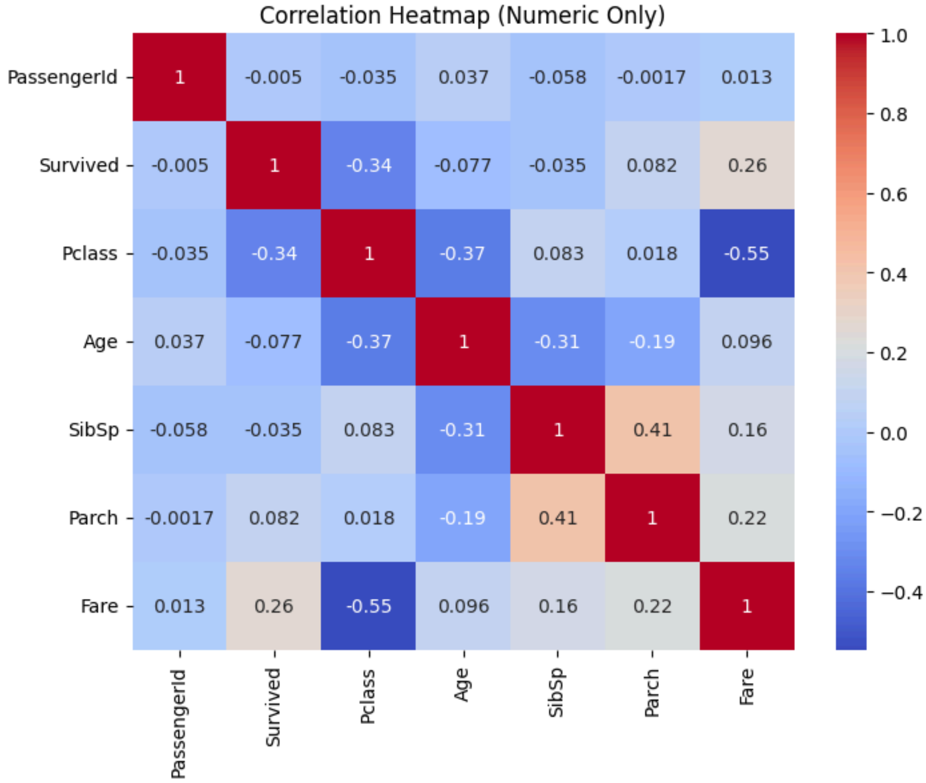

Age Distribution

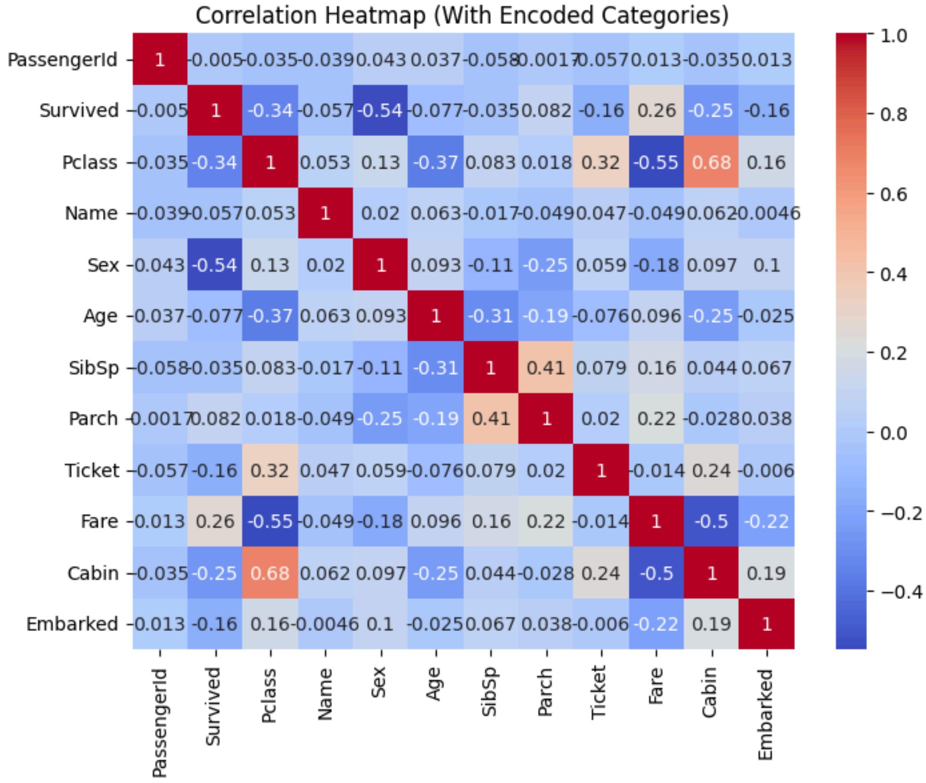

Survival by Sex


Survival by Passenger Class


Age Distribution by Class


Age
40
30
20
10
0
1
2
3
Pclass


Correlation Heatmap (Numeric Only)
PassengerId 1 -0.005 -0.035 0.037 -0.058 -0.0017 0.013
Survived -0.005 1 -0.34 -0.077 -0.035 0.082 0.26
Pclass -0.035 -0.34 1 -0.37 0.083 0.018 -0.55
Age 0.037 -0.077 -0.37 1 -0.31 -0.19 0.096
SibSp -0.058 -0.035 0.083 -0.31 1 0.41 0.16
Parch -0.0017 0.082 0.018 -0.19 0.41 1 0.22
Fare 0.013 0.26 -0.55 0.096 0.16 0.22 1
PassengerId
Survived
Pclass
Age
SibSp
Parch
Fare
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Correlation Heatmap (With Encoded Categories)
PassengerId 1 -0.005 -0.035 -0.039 0.043 0.037 -0.058 -0.0017 -0.057 0.013 -0.035 0.013
Survived -0.005 1 -0.34 -0.057 -0.54 -0.077 -0.035 0.082 -0.16 0.26 -0.25 -0.16
Pclass -0.035 -0.34 1 0.053 0.13 -0.37 0.083 0.018 0.32 -0.55 0.68 0.16
Name -0.039 -0.057 0.053 1 0.02 0.063 -0.017 -0.049 0.047 -0.049 0.062 -0.0046
Sex -0.043 -0.54 0.13 0.02 1 0.093 -0.11 -0.25 0.059 -0.18 0.097 0.1
Age -0.037 -0.077 -0.37 0.063 0.093 1 -0.31 -0.19 -0.076 0.096 -0.25 -0.025
SibSp -0.058 -0.035 0.083 -0.017 -0.11 -0.31 1 0.41 0.079 0.16 0.044 0.067
Parch -0.0017 0.082 0.018 -0.049 -0.25 -0.19 0.41 1 0.02 0.22 -0.028 0.038
Ticket -0.057 -0.16 0.32 0.047 0.059 -0.076 0.079 0.02 1 -0.014 0.24 -0.006
Fare -0.013 0.26 -0.55 -0.049 -0.18 0.096 0.16 0.22 -0.014 1 -0.5 -0.22
Cabin -0.035 -0.25 0.68 0.062 0.097 -0.25 0.044 -0.028 0.24 -0.5 1 0.19
Embarked -0.013 -0.16 0.16 -0.0046 0.1 -0.025 0.067 0.038 -0.006 -0.22 0.19 1
PassengerId
Survived
Pclass
Name
Sex
Age
SibSp
Parch
Ticket
Fare
Cabin
Embarked
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

```
observations = """
1. Most passengers are between 20–40 years old.
2. Females have a higher survival rate than males.
3. First-class passengers survived more often than third-class passengers.
4. Age distribution is higher in 1st class compared to 3rd class.
5. Fare shows a positive correlation with survival.
"""
print(observations)
```

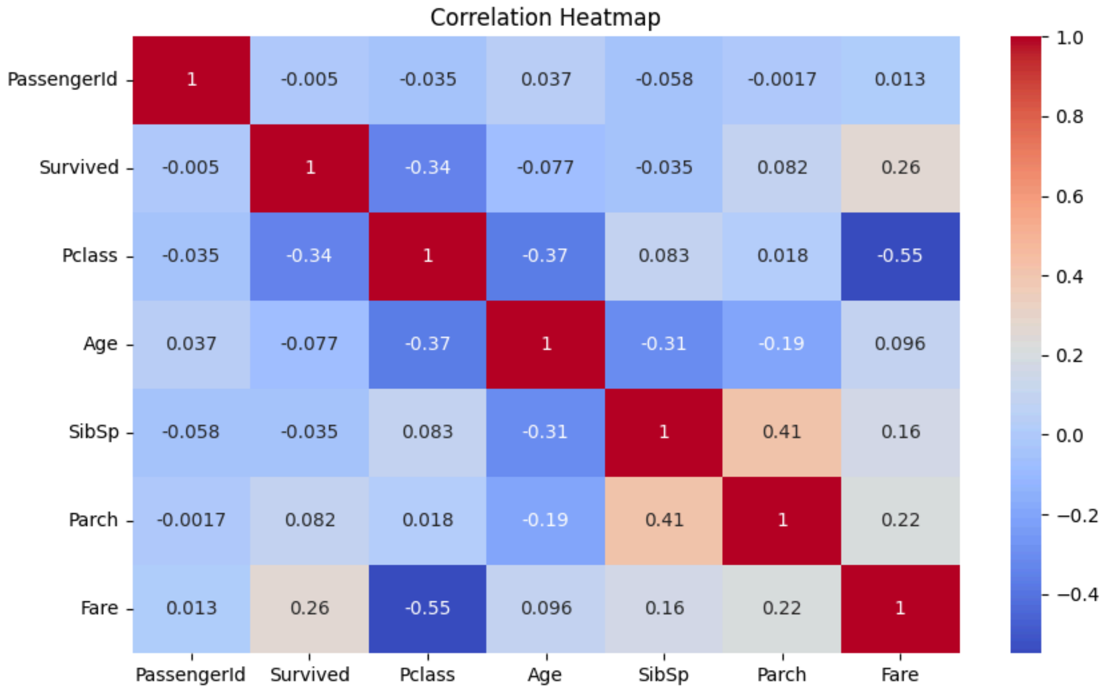
```
1. Most passengers are between 20–40 years old.
2. Females have a higher survival rate than males.
3. First-class passengers survived more often than third-class passengers.
4. Age distribution is higher in 1st class compared to 3rd class.
5. Fare shows a positive correlation with survival.
```

```
for col in df.select_dtypes(include='object').columns:
    print(f"\nValue counts for {col}:\n")
    print(df[col].value_counts())
```

```
Value counts for Name:

Name
Dooley, Mr. Patrick                                    1
Braund, Mr. Owen Harris                                1
Cumings, Mrs. John Bradley (Florence Briggs Thayer)    1
Heikkinen, Miss. Laina                                 1
Futrelle, Mrs. Jacques Heath (Lily May Peel)           1
                                                      ..
Hewlett, Mrs. (Mary D Kingcome)                        1
Vestrom, Miss. Hulda Amanda Adolfina                   1
Andersson, Mr. Anders Johan                            1
Saundercock, Mr. William Henry                         1
Bonnell, Miss. Elizabeth                               1
Name: count, Length: 891, dtype: int64

Value counts for Sex:

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

Value counts for Ticket:

Ticket
347082              7
1601                7
CA. 2343            7
3101295             6
CA 2144             6
                   ..
PC 17590            1
17463               1
330877              1
373450              1
STON/O2. 3101282    1
Name: count, Length: 681, dtype: int64

Value counts for Cabin:

Cabin
G6             4
C23 C25 C27    4
B96 B98        4
F2             3
D              3
              ..
E17            1
A24            1
C50            1
B42            1
C148           1
Name: count, Length: 147, dtype: int64

Value counts for Embarked:
```

```
# Keep only numeric columns for correlation
numeric_df = df.select_dtypes(include=['number'])
```

```
plt.figure(figsize=(10,6))
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```



	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1	-0.005	-0.035	0.037	-0.058	-0.0017	0.013
Survived	-0.005	1	-0.34	-0.077	-0.035	0.082	0.26
Pclass	-0.035	-0.34	1	-0.37	0.083	0.018	-0.55
Age	0.037	-0.077	-0.37	1	-0.31	-0.19	0.096
SibSp	-0.058	-0.035	0.083	-0.31	1	0.41	0.16
Parch	-0.0017	0.082	0.018	-0.19	0.41	1	0.22
Fare	0.013	0.26	-0.55	0.096	0.16	0.22	1

```
# Pairplot
sns.pairplot(df)
plt.show()
```