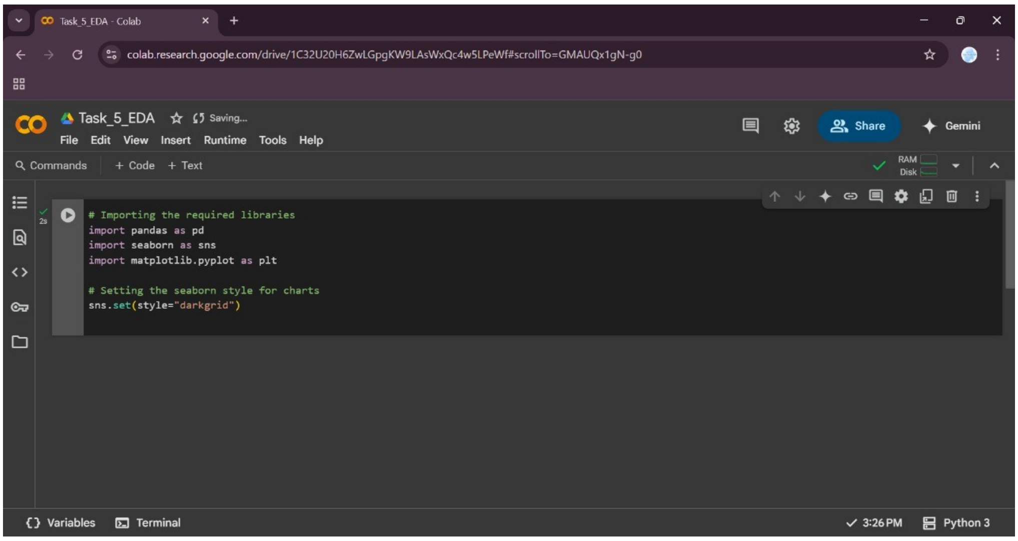
Task 5 – Exploratory Data Analysis (EDA)

Tool Used: Google Colab (Python, Pandas, Seaborn, Matplotlib)

1. **Import Libraries**

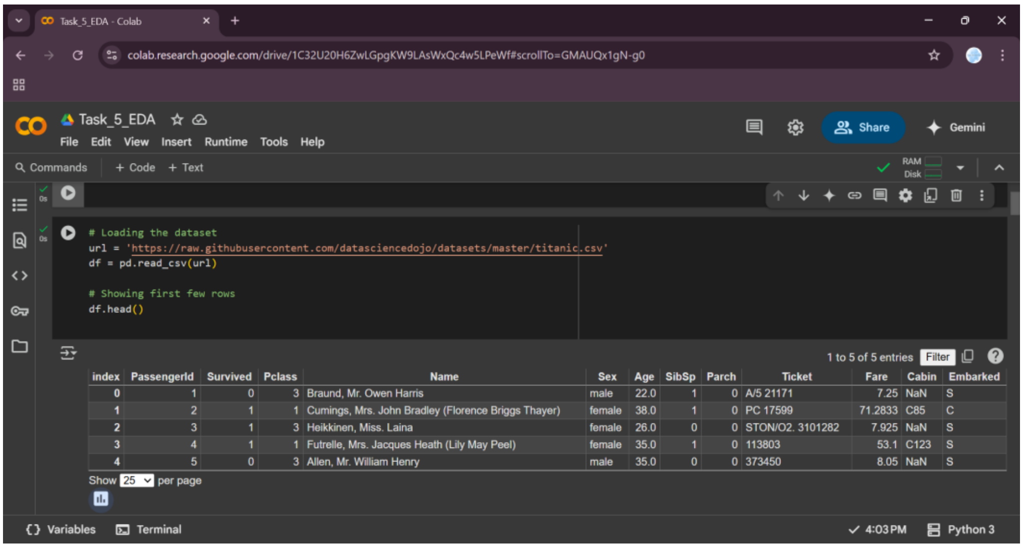


```
# Importing the required libraries
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Setting the seaborn style for charts
sns.set(style="darkgrid")
```

Task_5_EDA - Colab
colab.research.google.com/drive/1C32U20H6ZwLGpgKW9LAsWxQc4w5LPeWf#scrollTo=GMAUQx1gN-g0
Task_5_EDA
Saving...
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
RAM
Disk
Variables
Terminal
3:26 PM
Python 3

2. **df.head()**



```
# Loading the dataset
url = 'https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv'
df = pd.read_csv(url)

# Showing first few rows
df.head()
```

1 to 5 of 5 entries

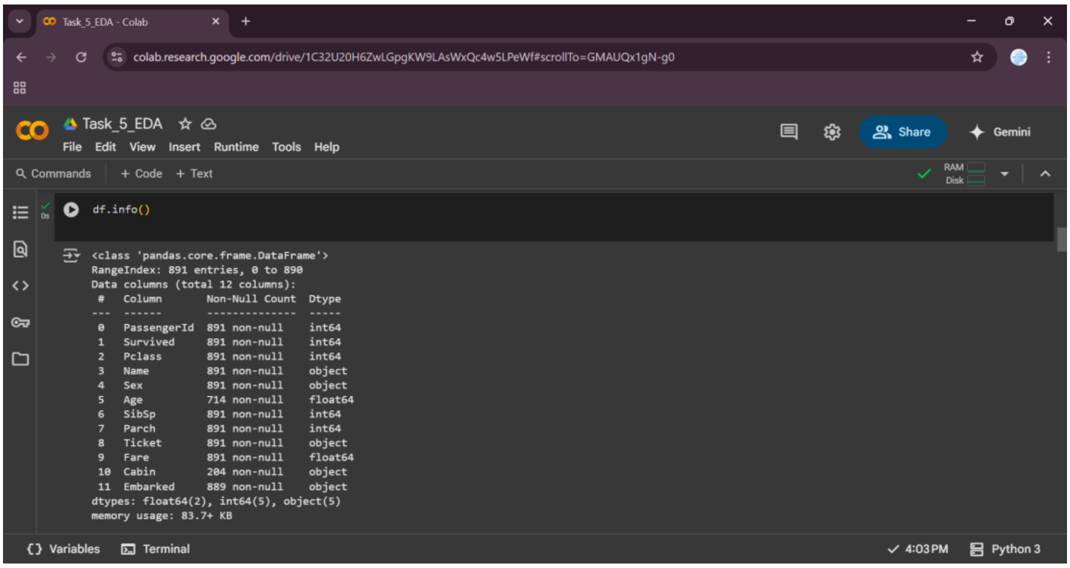
index	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05	NaN	S

Show 25 per page

Task_5_EDA - Colab
colab.research.google.com/drive/1C32U20H6ZwLGpgKW9LAsWxQc4w5LPeWf#scrollTo=GMAUQx1gN-g0
Task_5_EDA
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
RAM
Disk
Filter
Variables
Terminal
4:03 PM
Python 3

Observation:
Shows the first 5 rows of the dataset. Columns like Survived, Pclass, Name, Sex, Age, Fare are available.

3. **df.info()**

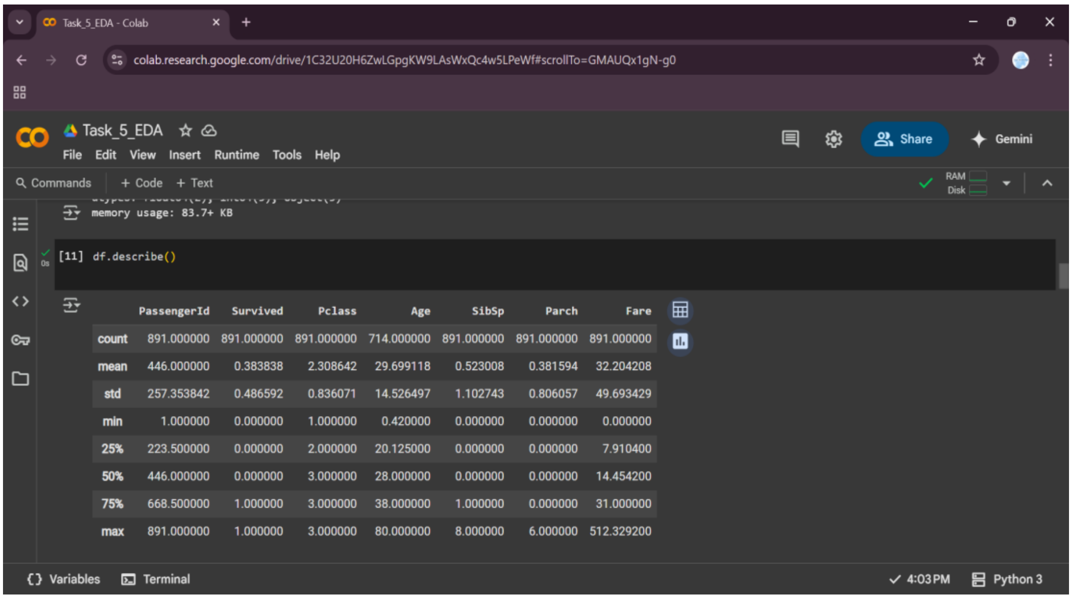

Task_5_EDA - Colab
colab.research.google.com/drive/1C32U20H6ZwLGpgKW9LAsWxQc4w5LPeWf#scrollTo=GMAUQx1gN-g0
Task_5_EDA
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
RAM
Disk
Variables
Terminal
4:03 PM
Python 3

```
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
```

Observation:
Shows that Age and Cabin columns have missing values. Data types are mostly correct.

4. **df.describe()**


Task_5_EDA - Colab
colab.research.google.com/drive/1C32U20H6ZwLGpgKW9LAsWxQc4w5LPeWf#scrollTo=GMAUQx1gN-g0
Task_5_EDA
File Edit View Insert Runtime Tools Help
Share
Gemini
Commands
+ Code
+ Text
RAM
Disk
Variables
Terminal
4:03 PM
Python 3

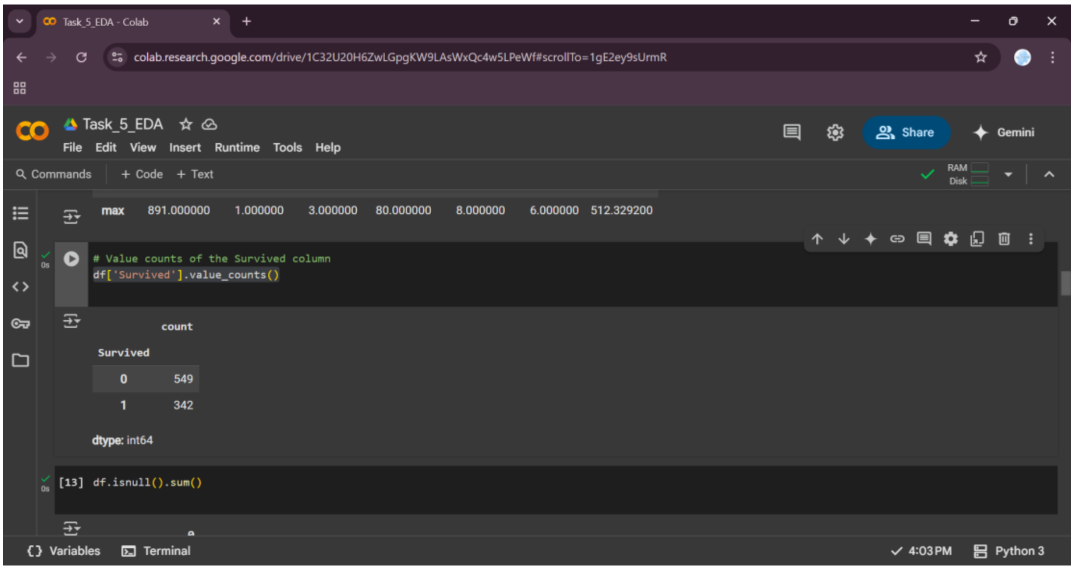
```
memory usage: 83.7+ KB
```

```
[11] 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

Observation:
Summary statistics show how Fare has a high range, and Age has outliers and a wide distribution.

5. **df['Survived'].value_counts()**



```
max    891.000000    1.000000    3.000000    80.000000    8.000000    6.000000  512.329200
```

```
# Value counts of the Survived column
df['Survived'].value_counts()
```

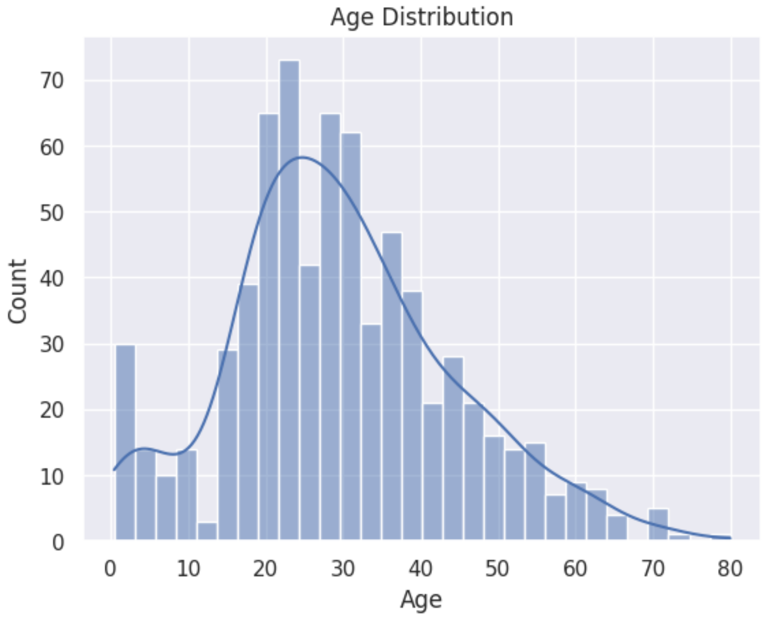
```
          count
Survived
   0        549
   1        342

dtype: int64
```

```
[13] df.isnull().sum()
```

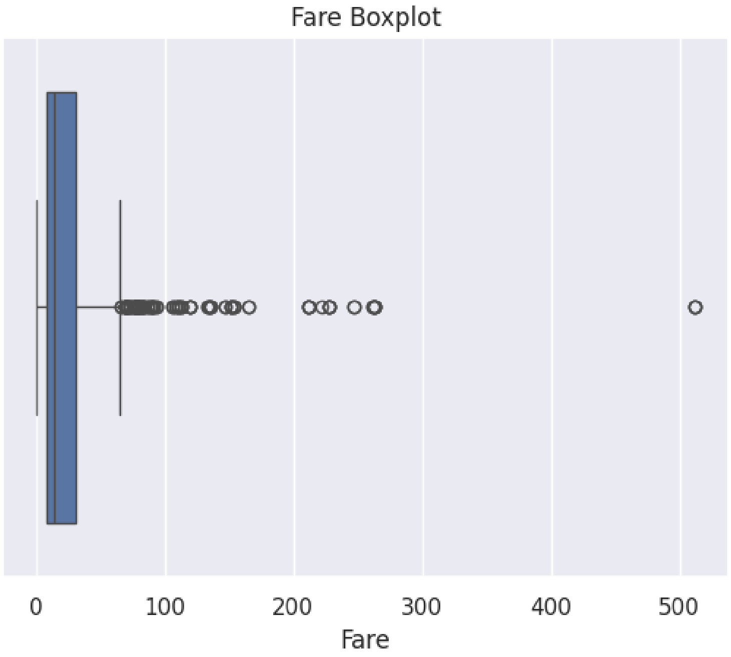
Observation:
More people didn’t survive (0) than survived (1). Helps understand the class imbalance.

6. **Age Histogram**



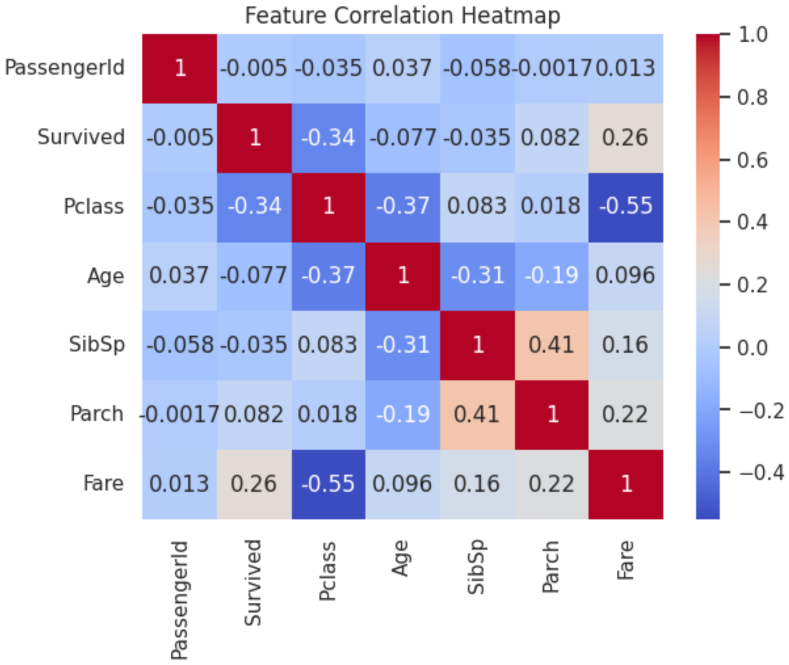
Observation:
Most passengers are between 20 and 40 years old. Very few children or seniors.

7. **Fare Boxplot**



Observation:
There are a few very high fare values (outliers), indicating some rich passengers.

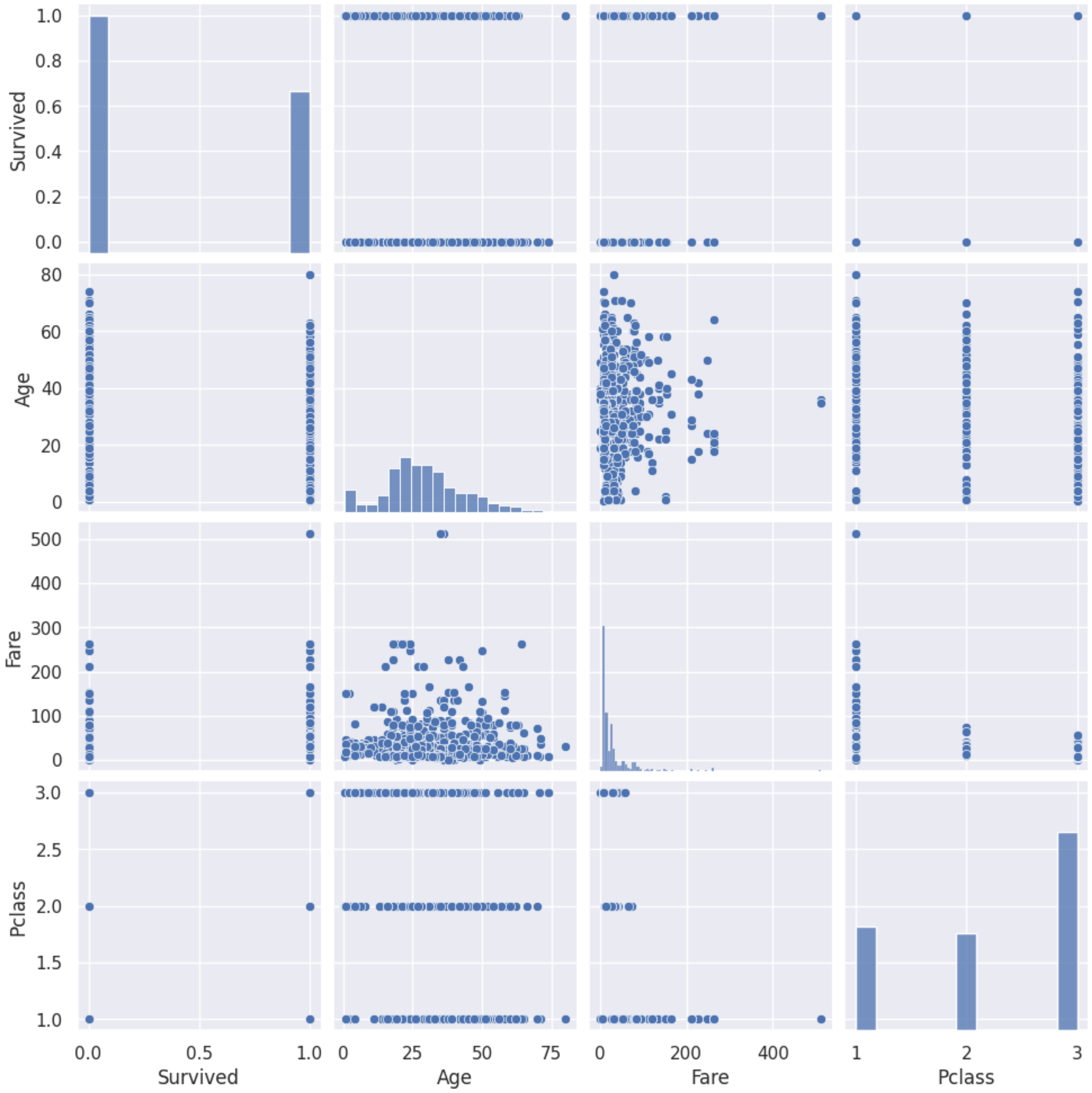
8. **Correlation Heatmap**



	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

Observation:
Survived has slight positive correlation with Fare and negative correlation with Pclass.

9. **Pairplot**



Observation:
Shows relationship between variables. Survival might relate to Fare and Pclass.

Final Summary:

-The dataset contains 891 passengers with 12 columns.

- Missing values found in 'Age', 'Cabin', and 'Embarked'.
- Most passengers are in the age group 20–40.
- Survival rate is low; more people did not survive.
- Fare has large outliers; some paid extremely high amounts.
- Correlation shows survival is affected by Fare and Pclass.
- Females had higher survival rates than males.