Titanic - EDA from Disaster

Data Dictionary

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

importing Libraries

In [1]:

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

importing Dataset

In [2]:

```python
titanic = pd.read_csv('train.csv')
```

In [3]:

```python
df = titanic.copy()
```

In [4]:

```python
df.sample(15)
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
141	142	1	3	Nysten, Miss. Anna Sofia	female	22.0	0	0	347081	7.7500	NaN	S
788	789	1	3	Dean, Master. Bertram Vere	male	1.0	1	2	C.A. 2315	20.5750	NaN	S
853	854	1	1	Lines, Miss. Mary Conover	female	16.0	0	1	PC 17592	39.4000	D28	S
794	795	0	3	Dantcheff, Mr. Ristiu	male	25.0	0	0	349203	7.8958	NaN	S
224	225	1	1	Hoyt, Mr. Frederick Maxfield	male	38.0	1	0	19943	90.0000	C93	S
850	851	0	3	Andersson, Master. Sigvard Harald Elias	male	4.0	4	2	347082	31.2750	NaN	S
241	242	1	3	Murphy, Miss. Katherine "Kate"	female	NaN	1	0	367230	15.5000	NaN	Q
712	713	1	1	Taylor, Mr. Elmer Zebley	male	48.0	1	0	19996	52.0000	C126	S
863	864	0	3	Sage, Miss. Dorothy Edith "Dolly"	female	NaN	8	2	CA. 2343	69.5500	NaN	S
847	848	0	3	Markoff, Mr. Marin	male	35.0	0	0	349213	7.8958	NaN	C
26	27	0	3	Emir, Mr. Farred Chehab	male	NaN	0	0	2631	7.2250	NaN	C
758	759	0	3	Theobald, Mr. Thomas Leonard	male	34.0	0	0	363294	8.0500	NaN	S
735	736	0	3	Williams, Mr. Leslie	male	28.5	0	0	54636	16.1000	NaN	S
412	413	1	1	Minahan, Miss. Daisy E	female	33.0	1	0	19928	90.0000	C78	Q
239	240	0	2	Hunt, Mr. George Henry	male	33.0	0	0	SCO/W 1585	12.2750	NaN	S

Data Preprocessing

Data Preprocessing

In [5]:

```
df.shape
```

Out[5]:

```
(891, 12)
```

In [6]:

```
df.describe()
```

Out[6]:

	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

In [7]:

```
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
```

Data Refining

In [8]:

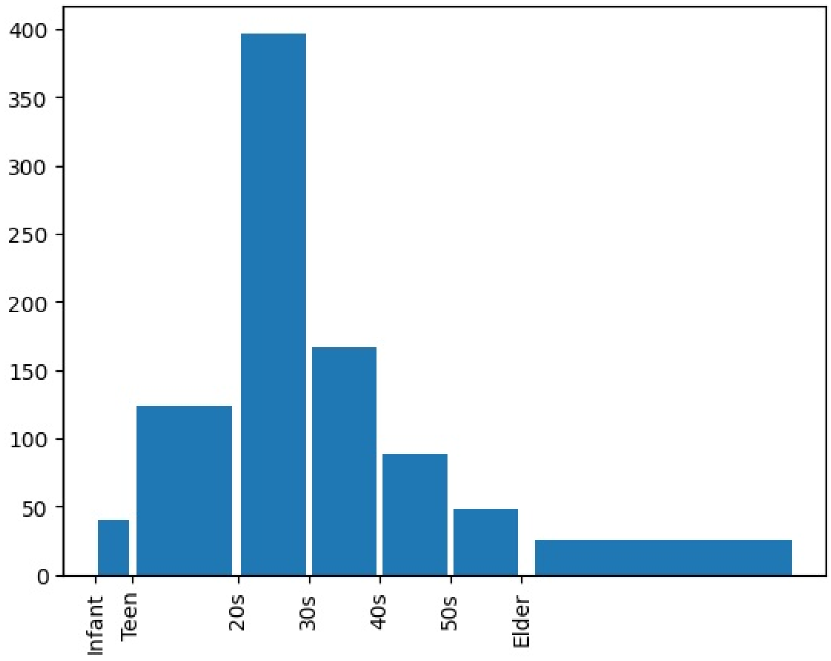
```
# replacing the missing values
df['Age'] = df['Age'].replace(np.nan,df['Age'].median(axis=0))
df['Embarked'] = df['Embarked'].replace('s','S')
```

In [9]:

```
#type casting Age to integer
df['Age'] = df['Age'].astype(int)
```

In [10]:

```
bins_level=[0, 5, 20 , 30, 40, 50, 60, 100]
plt.hist(df['Age'], bins = bins_level, rwidth = 0.9)
bins_label = ['Infant', 'Teen', '20s', '30s', '40s', '50s','Elder']
plt.xticks(bins_level[:-1],bins_label,rotation='vertical')
plt.show()
```



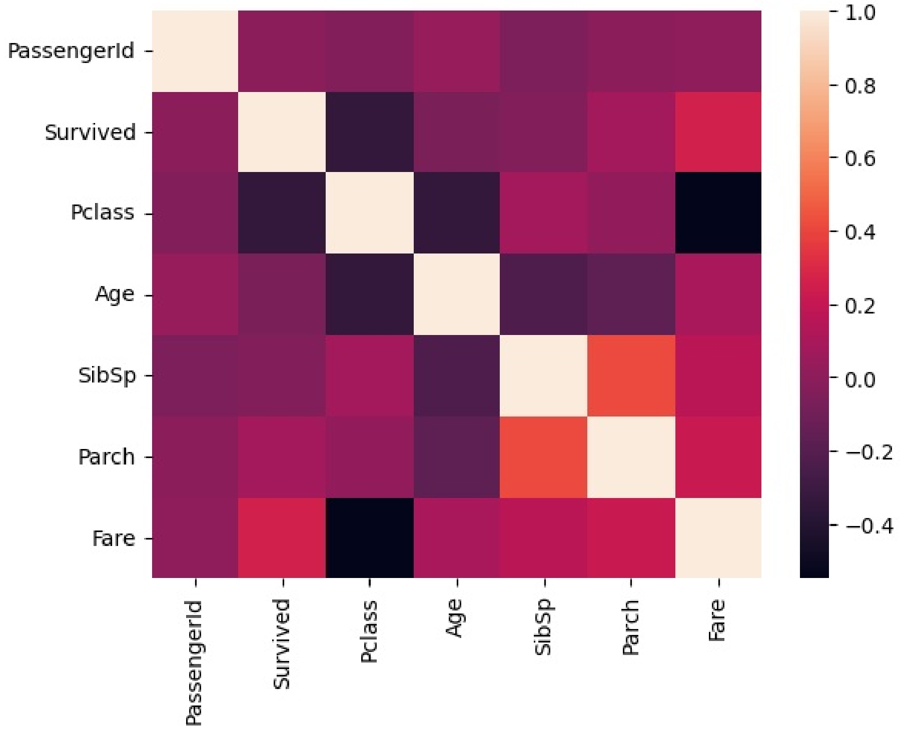
Visualisation using corelation with the help of heatmap

```
In [11]: corelation = df.corr()
```

```
C:\Users\SOUMEN MONDAL\AppData\Local\Temp\ipykernel_2860\2195490469.py:1: FutureWarning: The default value of n
umeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid c
olumns or specify the value of numeric_only to silence this warning.
  corelation = df.corr()
```

```
In [12]: sns.heatmap(corelation)
```

```
Out[12]: <AxesSubplot: >
```



Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js