Exploratory Data Analysis (EDA) on Titanic Dataset

By Mayank Shukla

Reading csv file through file path.

[1]:
```
import numpy as np
import pandas as pd
df=pd.read_csv("C:\\Users\\mayan\\Downloads\\titanic\\train.csv")
df.head()
```

[1]:

	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

Question 1. Use .describe(), .info(), .value_counts().

```
[4]: df.describe()
```

[4]:

	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

```
[5]: 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
```

```
[27]: df.value_counts()
```

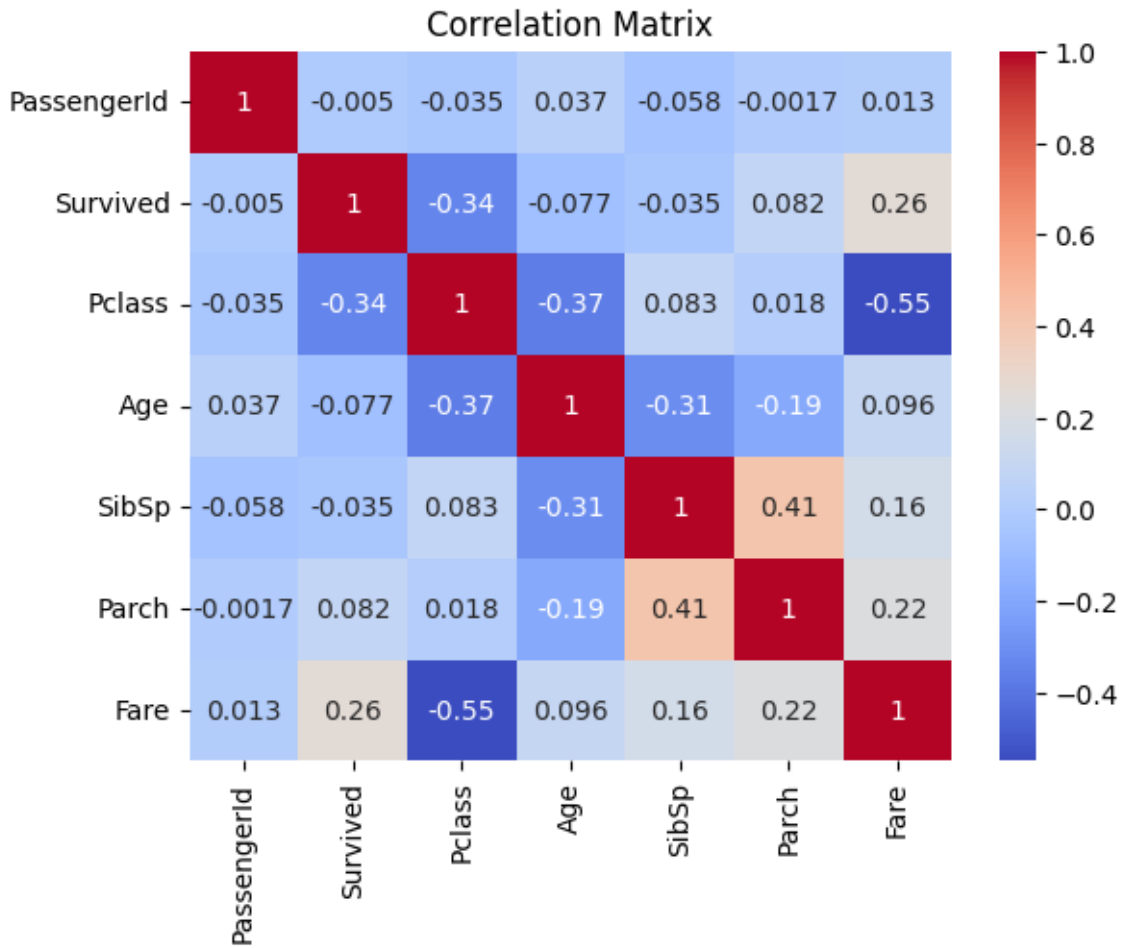
```
[27]: PassengerId  Survived  Pclass  Name                                                Sex     Age   SibSp  Parch  Ticket    Fare      Cabin  Embarked
      2            1         1       Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38.0  1      0      PC 17599  71.2833   C85    C         1
      572          1         1       Appleton, Mrs. Edward Dale (Charlotte Lamson)       female  53.0  2      0      11769     51.4792   C101   S         1
      578          1         1       Silvey, Mrs. William Baird (Alice Munger)           female  39.0  1      0      13507     55.9000   E44    S         1
      582          1         1       Thayer, Mrs. John Borland (Marian Longstreth Morris) female  39.0  1      1      17421     110.8833  C68    C         1
      584          0         1       Ross, Mr. John Hugo                                 male    36.0  0      0      13049     40.1250   A10    C         1
                                                                                                                                                    ..
      328          1         2       Ball, Mrs. (Ada E Hall)                             female  36.0  0      0      28551     13.0000   D      S         1
      330          1         1       Hippach, Miss. Jean Gertrude                        female  16.0  0      1      111361    57.9792   B18    C         1
      332          0         1       Partner, Mr. Austen                                 male    45.5  0      0      113043    28.5000   C124   S         1
      333          0         1       Graham, Mr. George Edward                           male    38.0  0      1      PC 17582  153.4625  C91    S         1
                   1         1       Behr, Mr. Karl Howell                               male    26.0  0      0      111369    30.0000   C148   C         1
      Name: count, Length: 183, dtype: int64
```

Expand Output

Question 2. Use sns.pairplot(), sns.heatmap() for visualization.

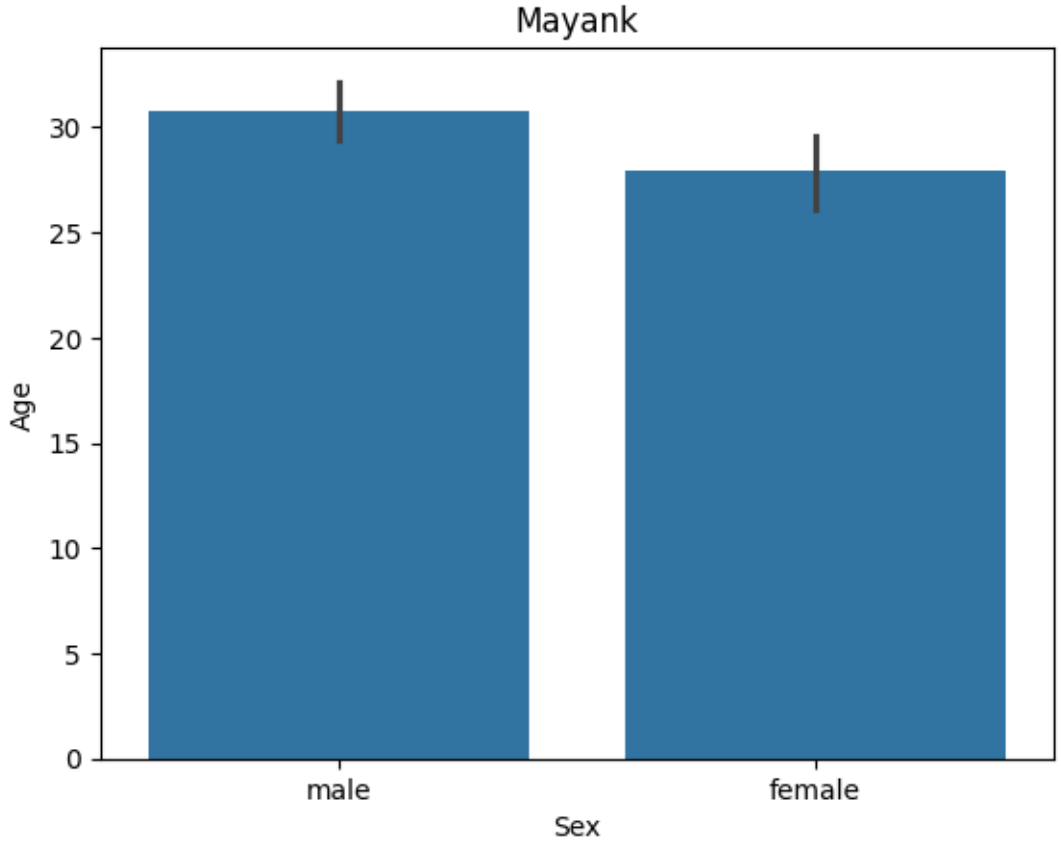
Pairplot :-

Heatmap :-


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

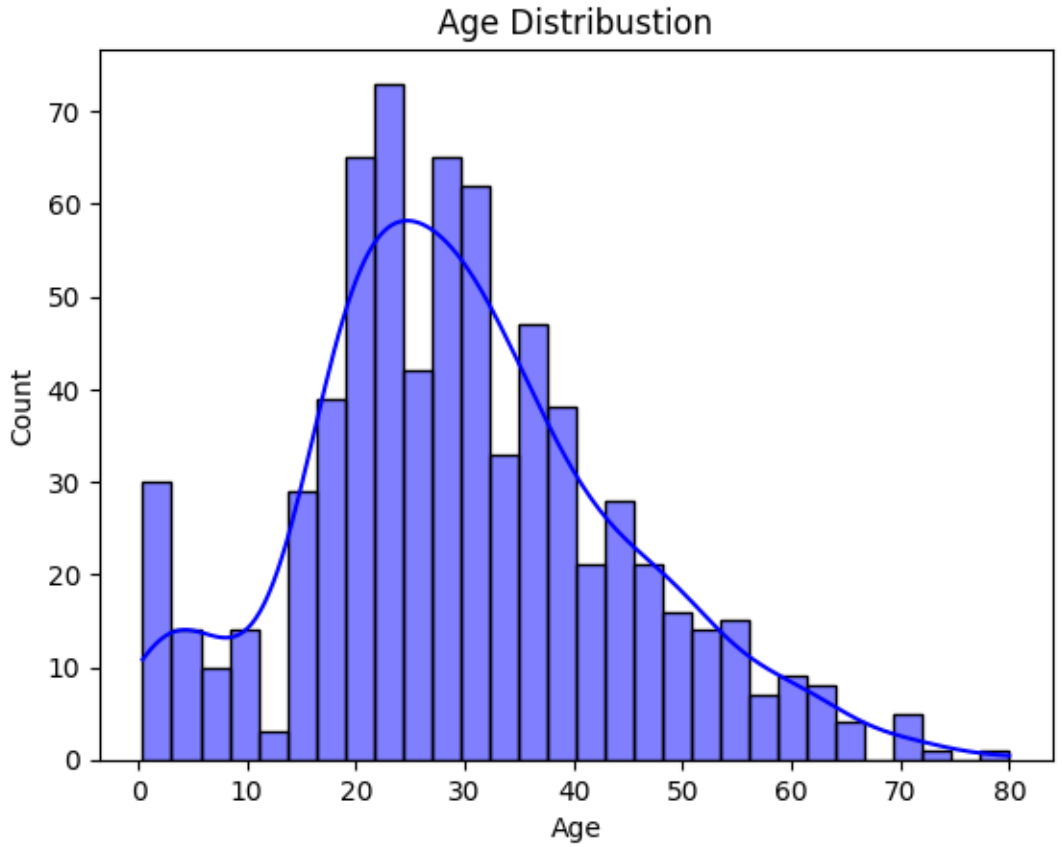
Question 3. Identify relationships and trends.

Barplot :-

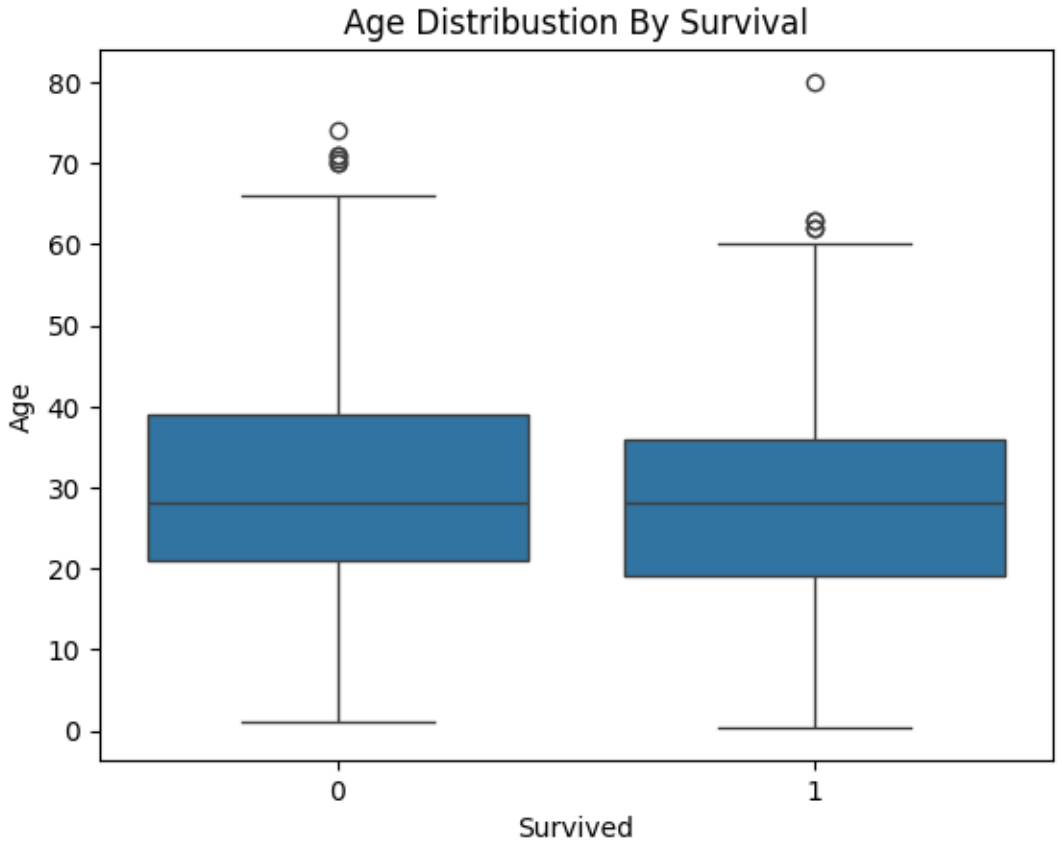


Question 4. Plot histograms, boxplots, scatterplots

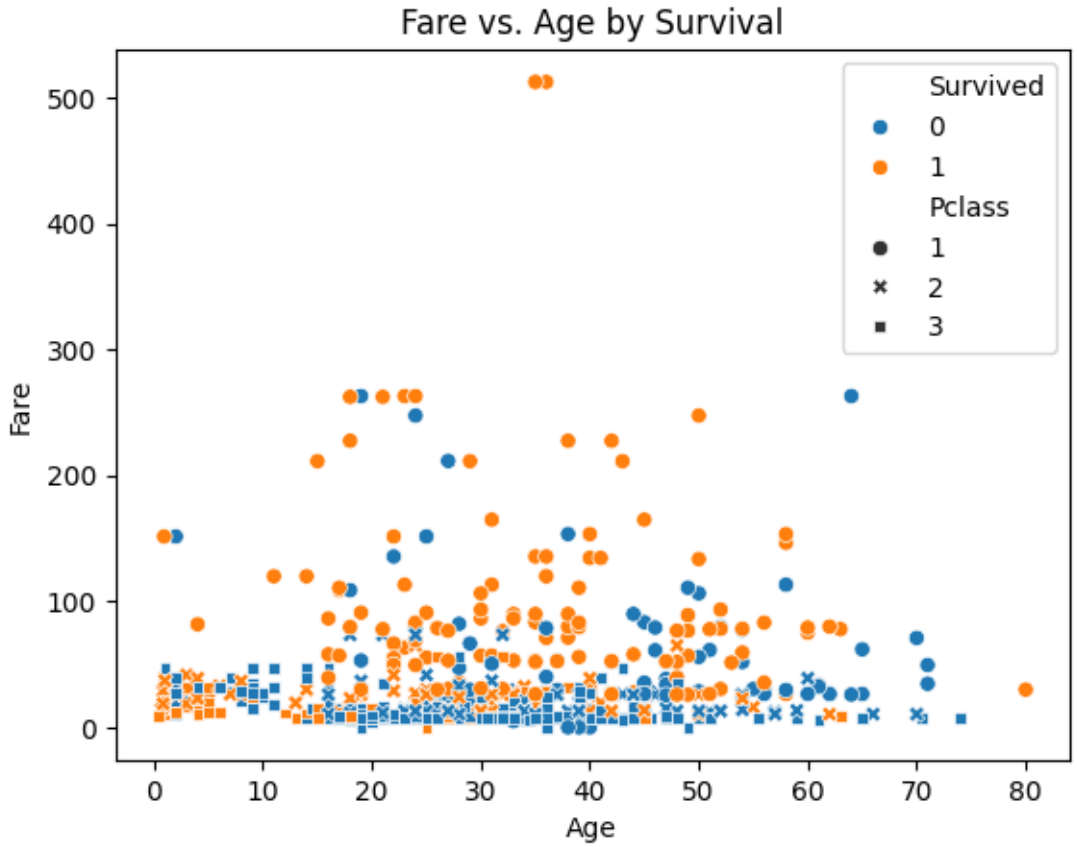
Histogram :-



Boxplot : -


Age Distribustion By Survival
80
70
60
50
40
30
20
10
0
Age
0
1
Survived

Scatter Plot : -



Summary : -

Passengers in higher classes, who paid higher fares, were more likely to survive. Younger individuals also showed higher survival rates. Most passengers were concentrated in the 20-30 age range and paid lower fares, but survival was skewed towards those with higher fares and better class. Gender and age played roles in survival outcomes, with variability seen in the data. 🚢✨

Thank You