Titanic data analysis

Step 1:

Loading the dataset into the jupyter notebook using pandas and saving the data into a data frame

Step2:

Finding the summary statistics using .describe(),.info() and .value\_counts() and isnull().sum()

Step 3:

We have found that there are null values in 3 columns they are age , cabin and embarked , now we have to fill these null values

Step 4:

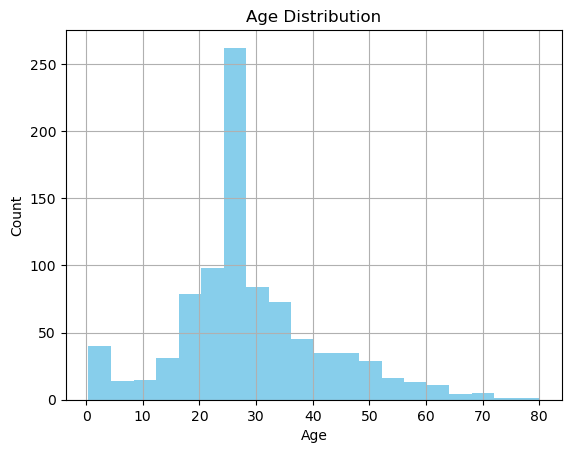
Since the age column has outliers we have to use median of age column to fill it

Next the cabin column has more than 600 null values which is more than 80% so we have to drop the column

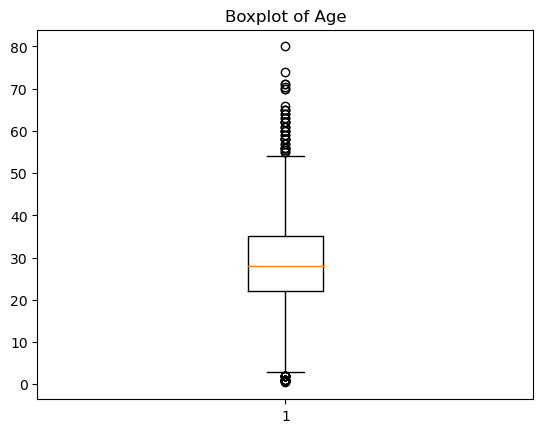
Now to fill the null values in the embarked column we will use the mode that column

Step 5 :

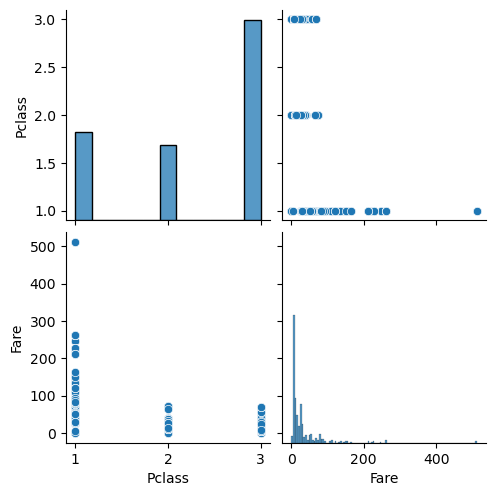
Since the data is clean now , we can continue with the visualization



So in this graph we can see the age VS count distribution. The age of passengers lie between 0-80 and most of the passengers are of the age between 20 – 30 (youngsters).



This boxplot represents the age column . The median age is around 28 years. There are several outliers, especially passengers above 65 years. The age range is mostly between 20 and 40, showing a tight central spread. Indicates the presence of elderly passengers who are rare.



Histogram of fare:

Right-skewed: Most fares are low (under 100). A few extreme values (outliers) exist: fares > 300 (probably first-class VIPs).

Histogram of pclass:

It shows that Most passengers prefer 3rd class, followed by 1st, then 2nd. It’s a discrete distribution (only 1, 2, 3).

Scatterplot (fare vs pclass):

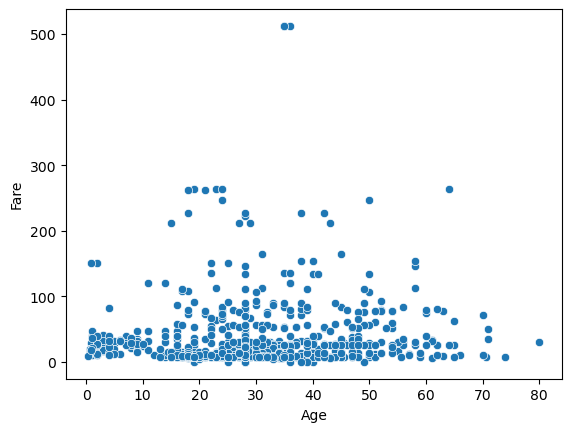
This scatter plot shows clear inverse relationship:

1st class passengers (pclass = 1) tend to have high fares.

3rd class passengers (pclass = 3) paid lowest fares.

This confirms that passenger class strongly influenced ticket price.

Few overlapping fare ranges between classes, but overall trend is clear.

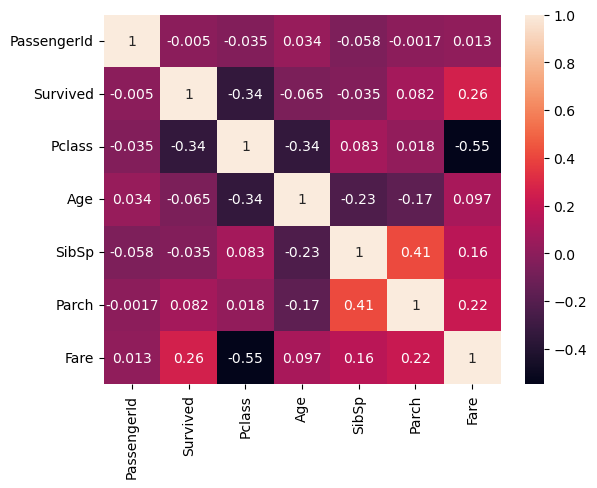


The plot shows no clear trend: passengers of all ages paid a wide range of fares. Both young and old passengers could have paid low or high fares.

Most passengers are Aged 20–40 Paid a fare of under ₹100 (or equivalent currency units)

A few high fare outliers exist (₹200–₹500+). Many of them are older passengers, likely 1st class VIPs. Several younger passengers (children) appear to have paid low fares — possibly traveling with families.

Age does not determine fare directly. Instead, passenger class (not shown in this plot in the last graph) is a stronger factor affecting fare.



Darker red colour represents Strong negative correlation (near -1)

Darker blue represents Strong positive correlation (near +1)

White/light colours represent Weak or no correlation (near 0)

|  |  |
| --- | --- |
| fare vs pclass | Higher the class lower the fare (1st class = 1) |
| fare vs survived | Higher fare passengers more likely to survive |
| pclass vs survived | 3rd class less likely to survive |
| age vs fare | Very weak positive correlation |
| age vs survived | Very weak (older passengers less likely to survive) |

Bottom of Form