Report on Titanic Dataset

Variables

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

Pclass: A proxy for socio-economic status (SES)

1st = Upper, 2nd = Middle, 3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of 0.5

Sibsp: The dataset defines family relations in this way...

Sibling = brother, sister, stepbrother, stepsister

Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...

Parent = mother, father

Child = daughter, son, stepdaughter, stepson

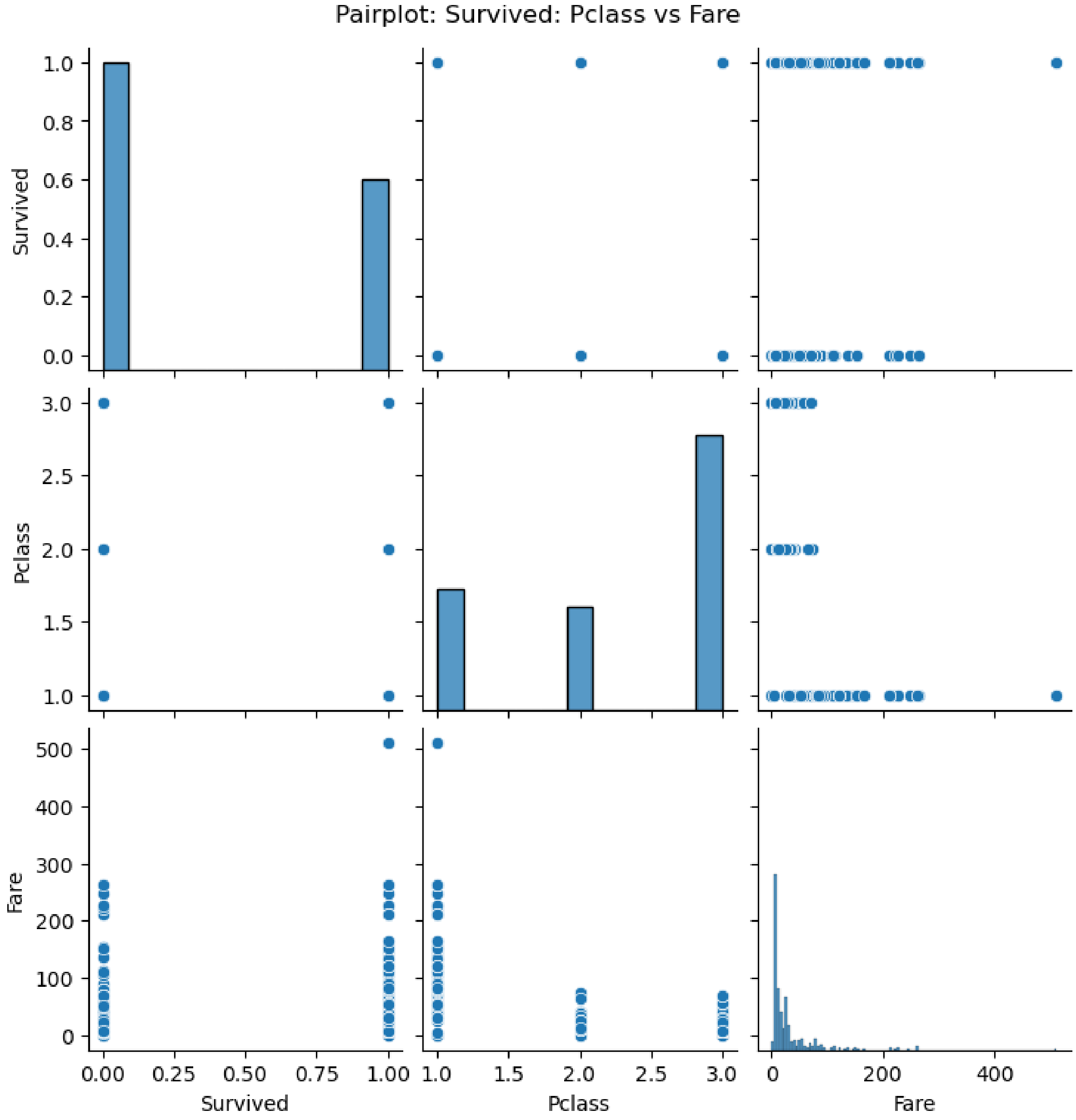
Some children travelled only with a nanny, therefore parch=0 for them.

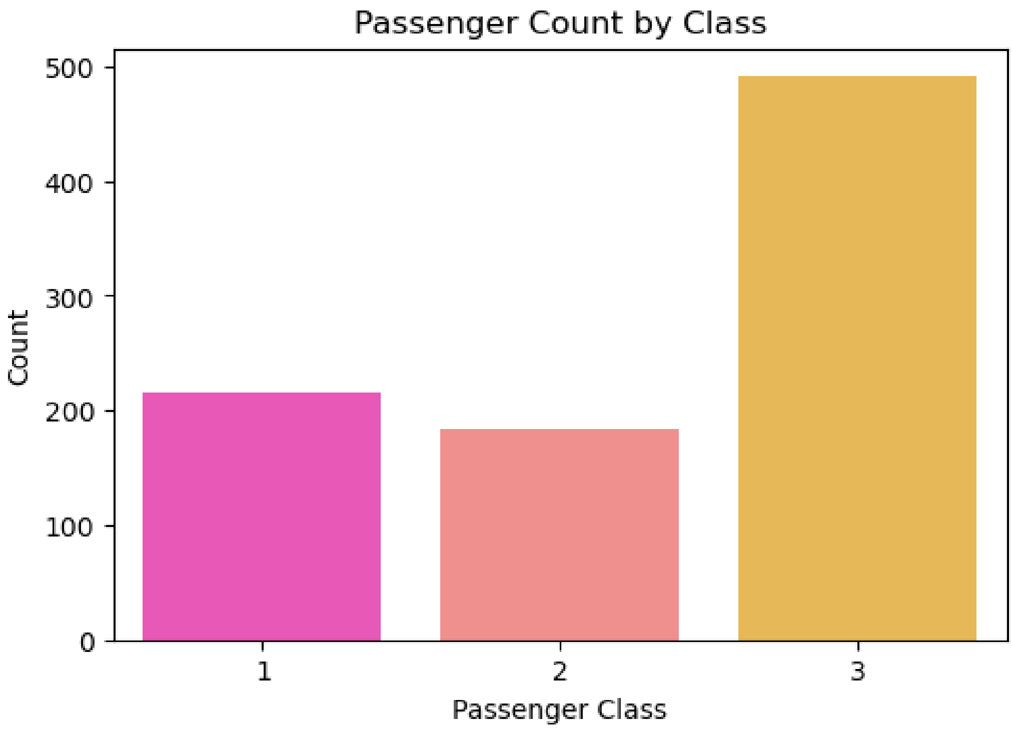
Analysis

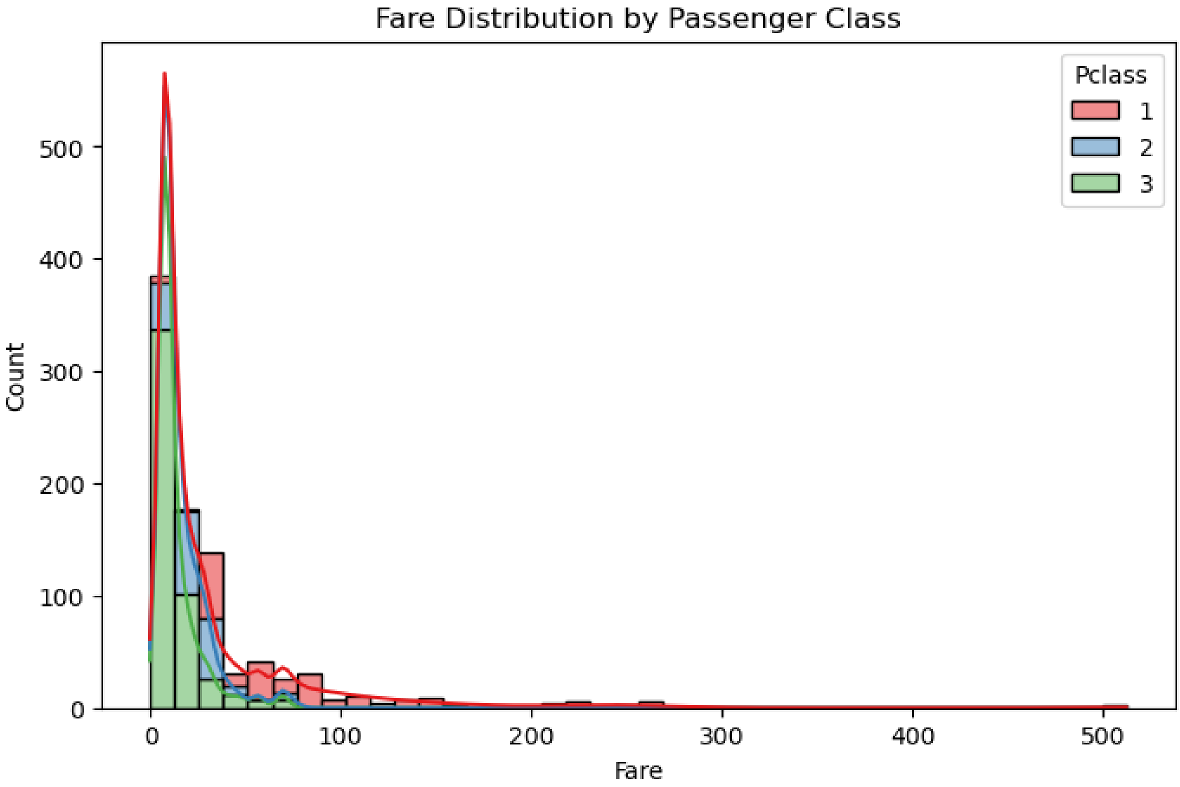
- Libraries used : Pandas, Matplotlib, Seaborn
- Headings, data structure(info) and Statistic analysis using describe()

- There are 891 entries for all headings except Age which had some null values and Cabin had only 204 entries, so Cabin column as whole was dropped, while Age null values were filled with NaN values which can then be dropped when analysing

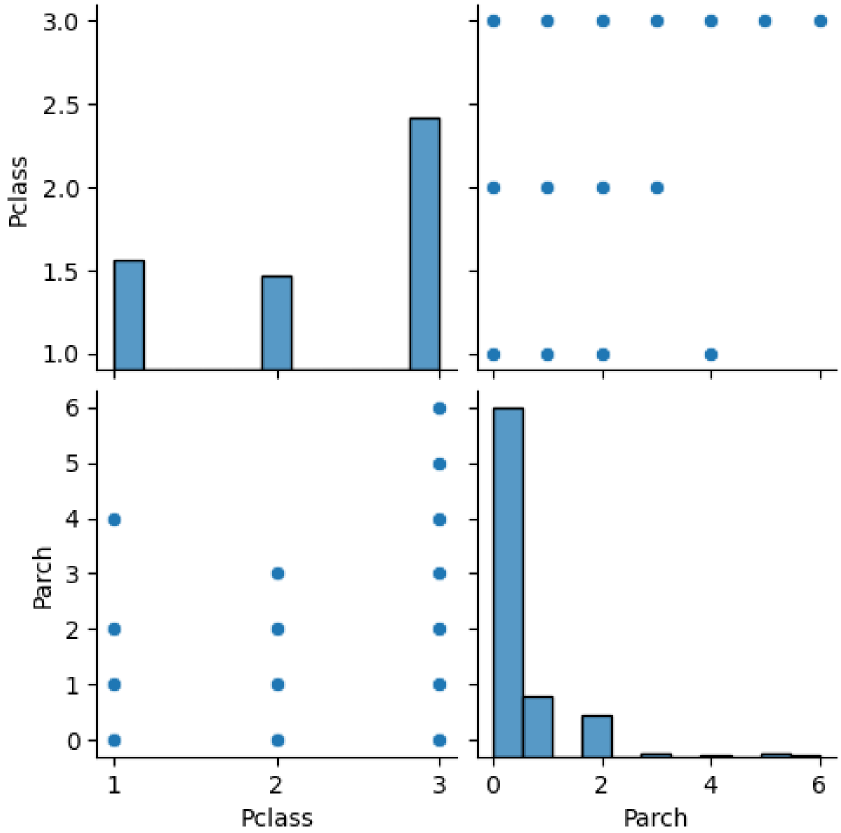
- Pair Plots, count plots, scatter-plots, box plots, histograms and correlation matrices with hues(heat maps) were made for analysing:-



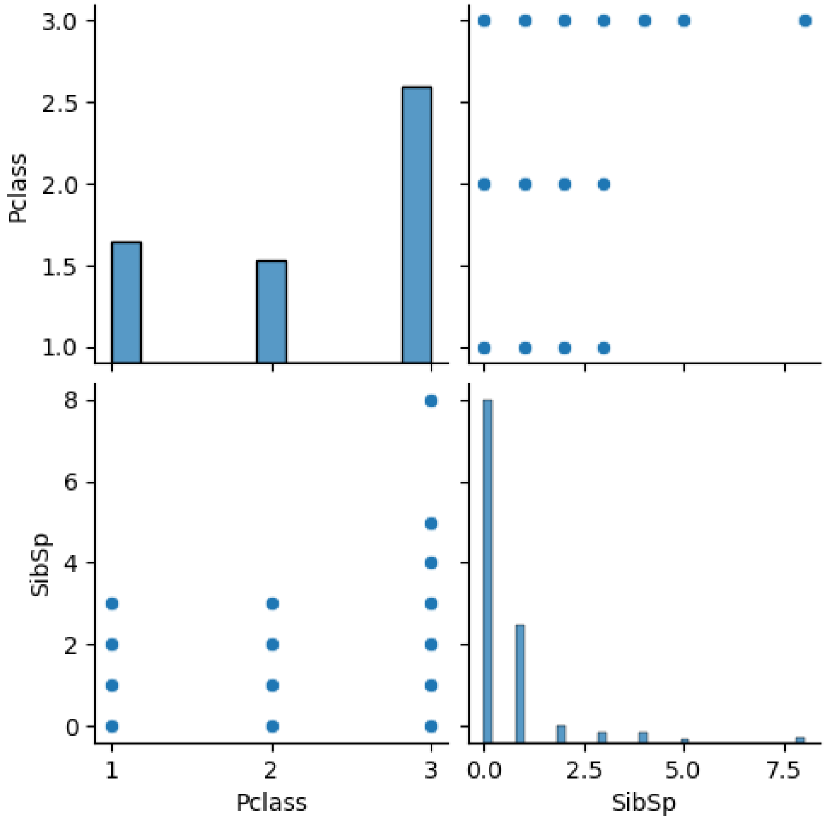




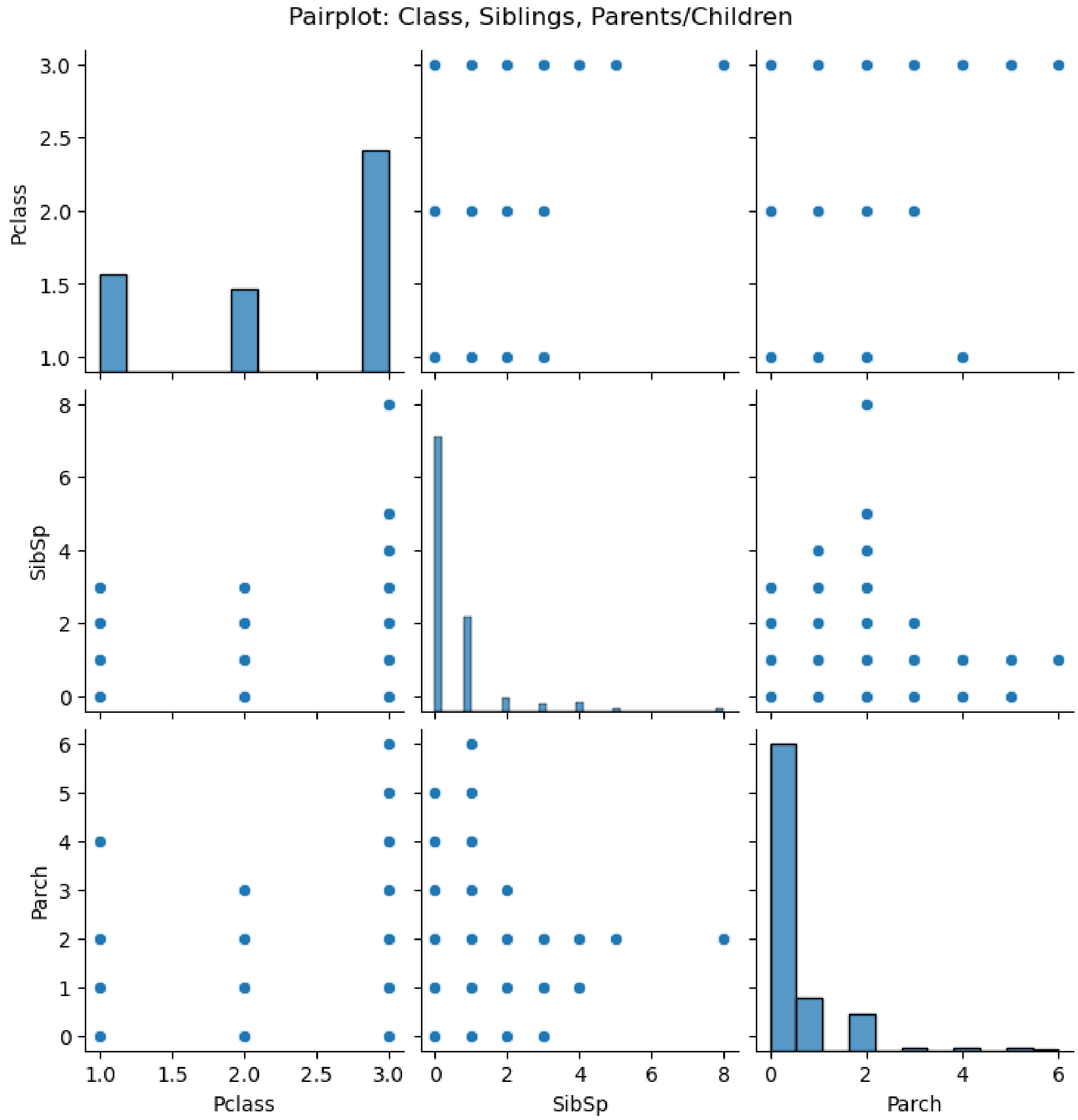
We can see that most of the passengers were from 3rd class. Also from the pair plot we can see that there is a 1st class outlier who survived.



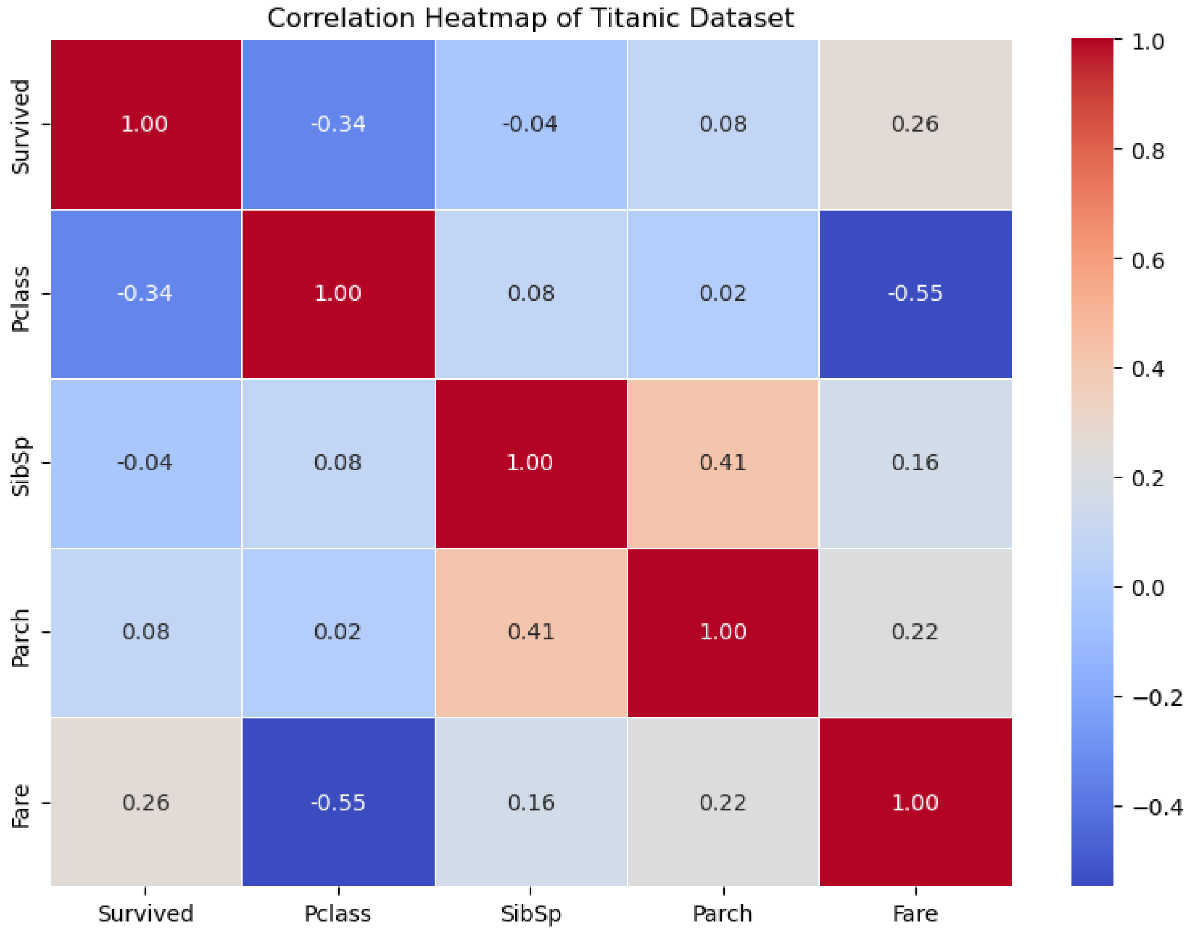
Pairplot: Class, Parents/Children



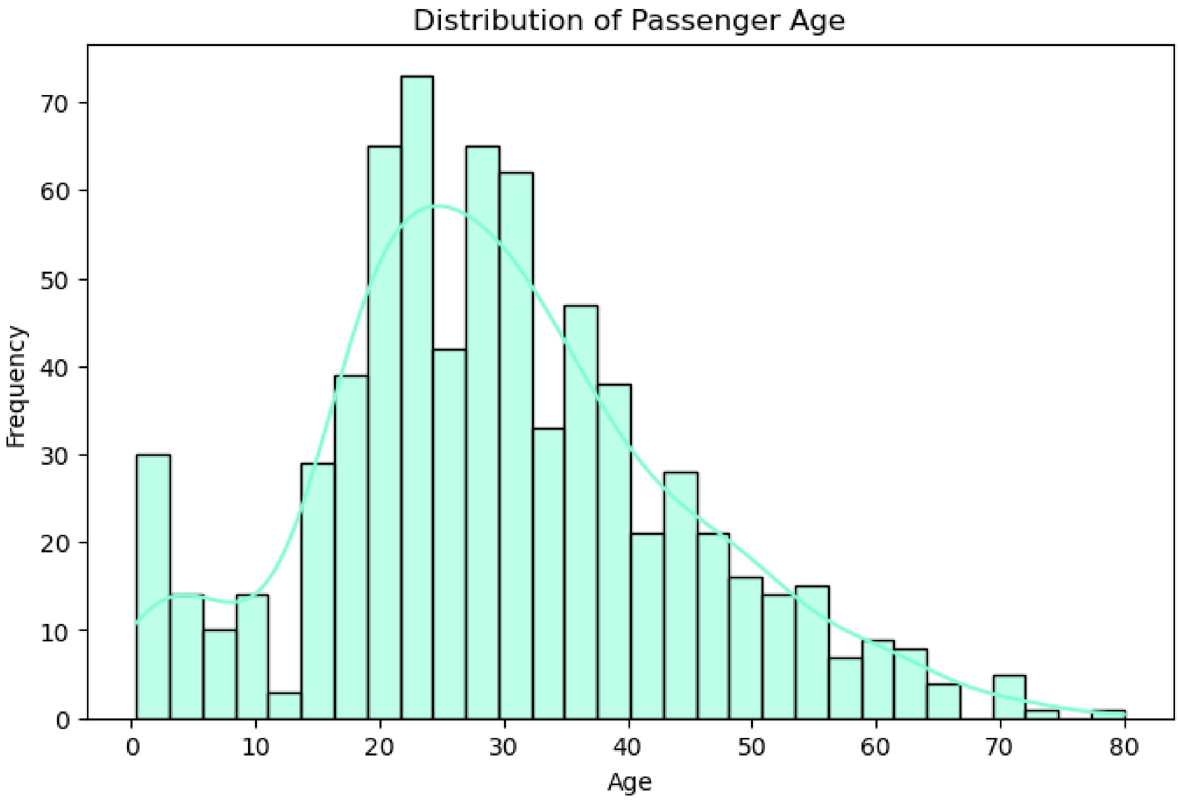
Pairplot: Class, Siblings



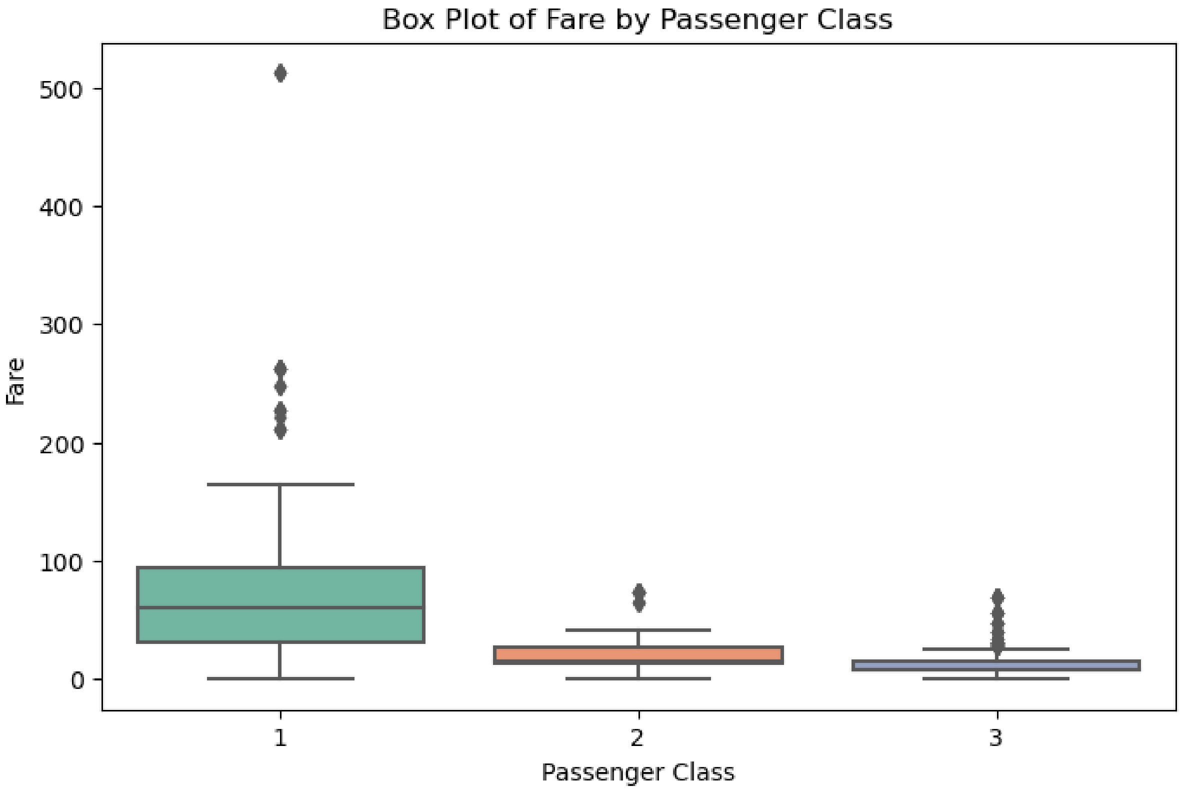
Here we can see that the outlier of the 3rd class who survived was travelling with siblings/spouses



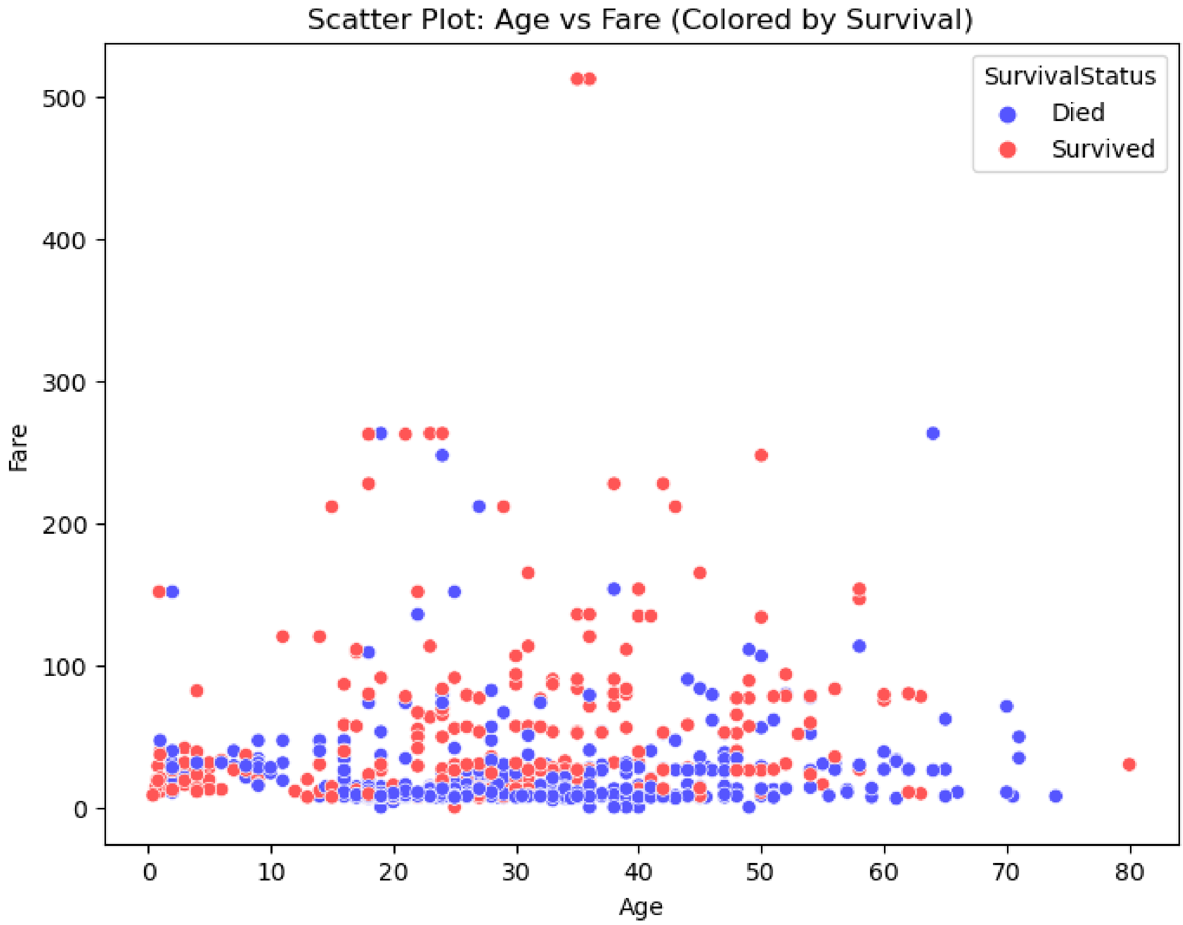
There is mostly negative correlation between these values means that two variables move in opposite directions; i.e., when one increases the other decreases, and vice-versa. Strong negative correlation between Pclass and Fare means higher classes paid more; lower classes paid much less. Also Moderate Negative correlation between Pclass and survived means that as class increases(3rd class) chance of survival decreases. We can see moderate correlation between Parch and SibSp means that passengers with siblings/spouses aboard also had parents/children aboard.



Distribution of Age shows that it is very much varied from 0 to 80. But most of the passengers were from 20-40 years of age. There were more children than elderly on board Titanic.



1. **Fare decreases as class increases**
2. **Presence of Outliers (especially in 1st class) There were some very expensive 1st class tickets, possibly for luxury suites or wealthy passengers (above 200 and even 500)**
3. **Distribution is skewed All three box plots show right skewness — long upper tails, indicating a few very high fares. Most passengers in each class paid fares near the lower end. The Titanic fare structure clearly reflected social class.**



Here the majority of people who died had paid less fare irrespective of age. But most children survived, while mostly 2nd class passengers (who paid fare between 50-200) and 3rd class passengers who were young also survived.

CONCLUSIONS:

1. There were more 3rd & 2nd class passengers than 1st class, probably because providing too much luxury on board a ship is not easy.
2. There were more children than elderly aboard, as children accompanied the families and elderly will have problem moving on the ship
3. The chance of surviving is negatively correlated with class i.e., rich passengers had better chance of survival/survived
4. There were many families aboard rather than only spouses, parents, children with nannies or flying solo.
5. The survived were either children, rich or young who could survive in the cold water when Titanic sunk.