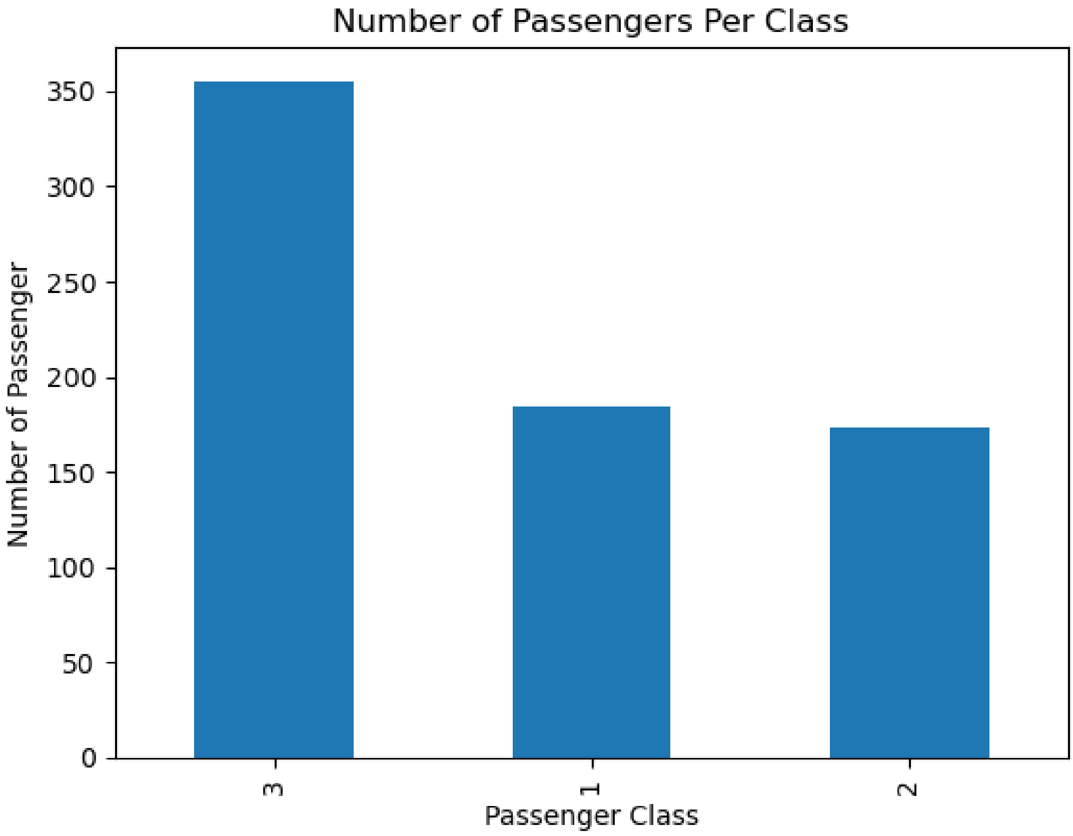
DJS Compute Task 3 – Data Visualization

Conducted on the Titanic Dataset

a. What is the distribution of passengers by class?

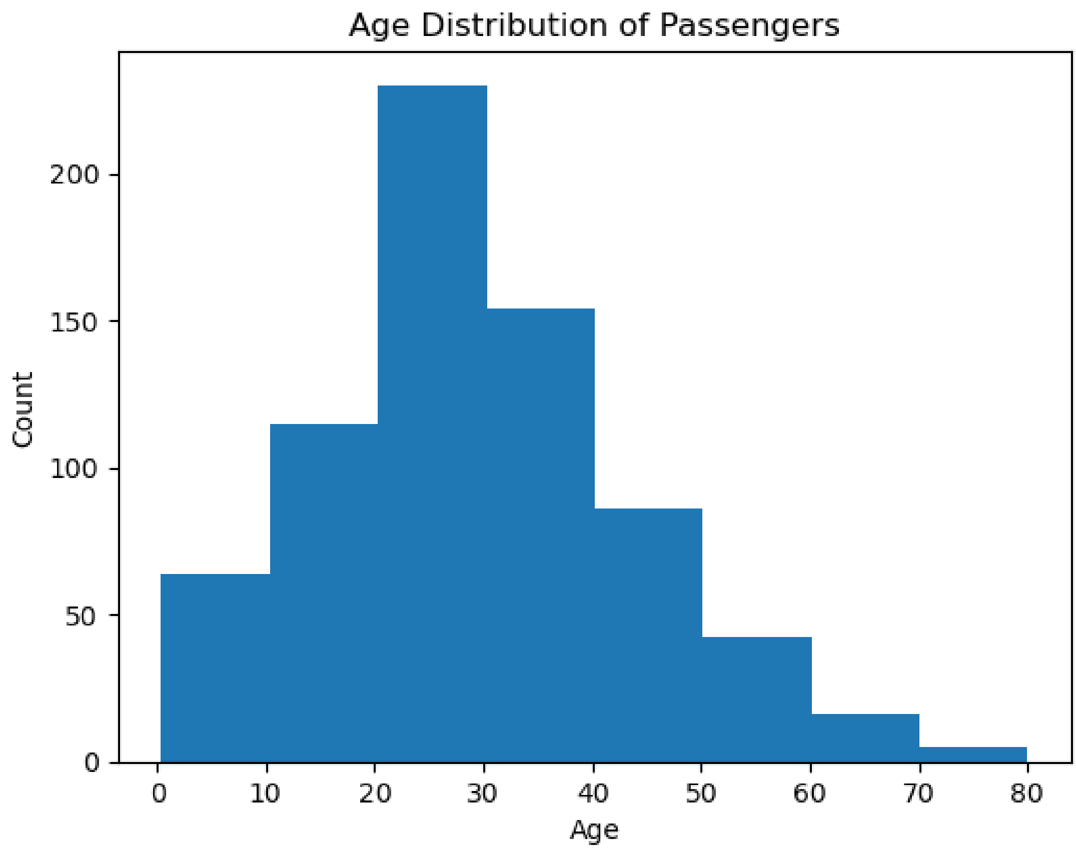


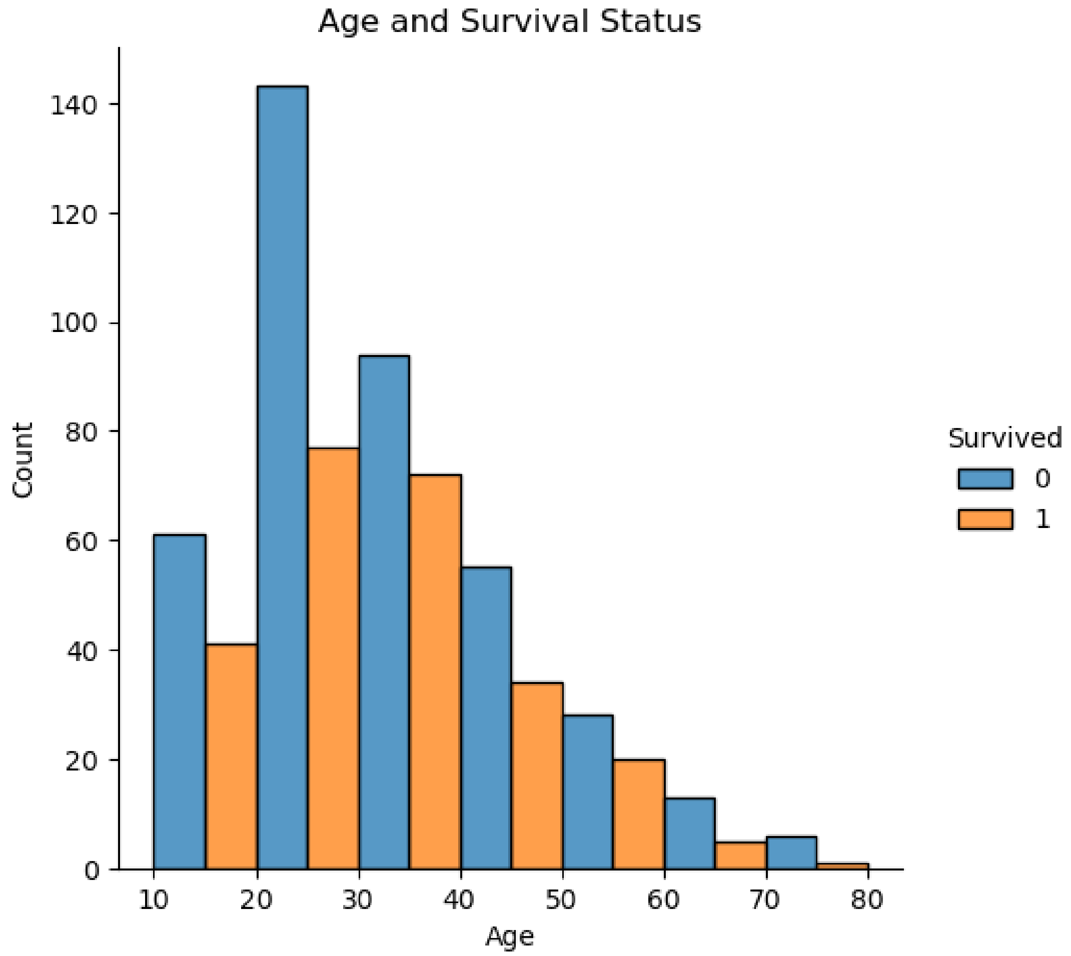
Most of the passengers on board were from class 3 (355). Class 1 had the second most number of passengers (184). 173 passengers were from Class 2. There isn't much difference between the number of passengers in Class 2 and Class 1.

b. What is the age distribution of passengers on the Titanic?

After cleaning the data, the mean age of passengers onboard comes out to be 29.64 years. 50% of the passengers were below the age of 28 and 75% were below the age of 38. The oldest passenger on board was 80 years old and the smallest age value in the dataset is 0.42 – this can be either a mistake or there was a 5-month-old baby on board.

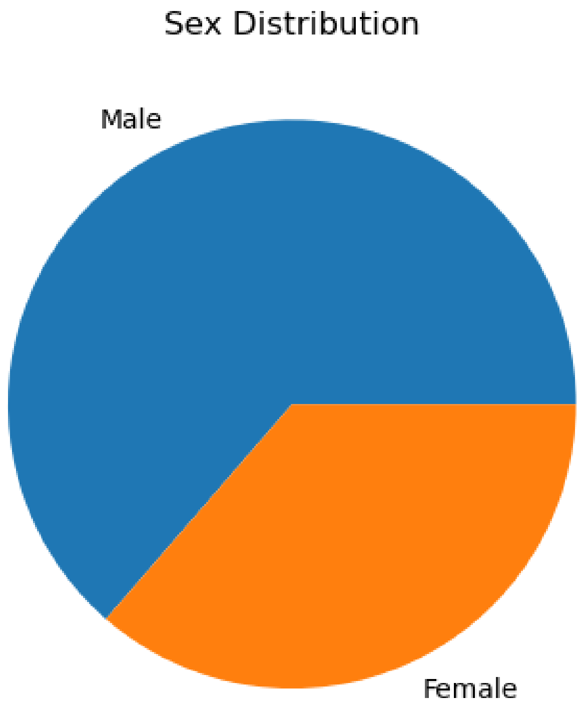
Most of the passengers on board were between the ages of 20 – 30. After this, the most prominent age group is 30 – 40 years, followed by 10-20 years. The age group that had the least amount of passenger on board was between 70 – 80 years.



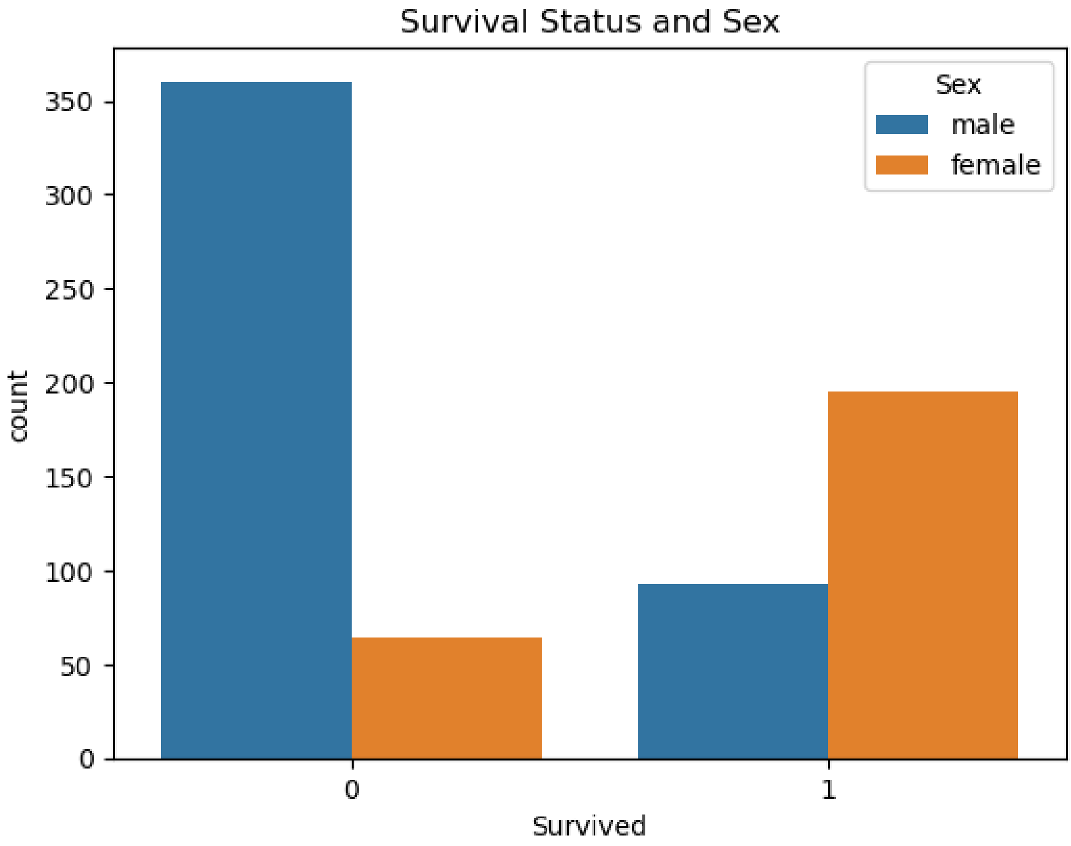


The highest number of deaths were from the age group 20 – 30 years, followed by 30 – 40, which makes sense since most of the passengers on board were in this age range. Note that the highest number of survivors also came from this age range.

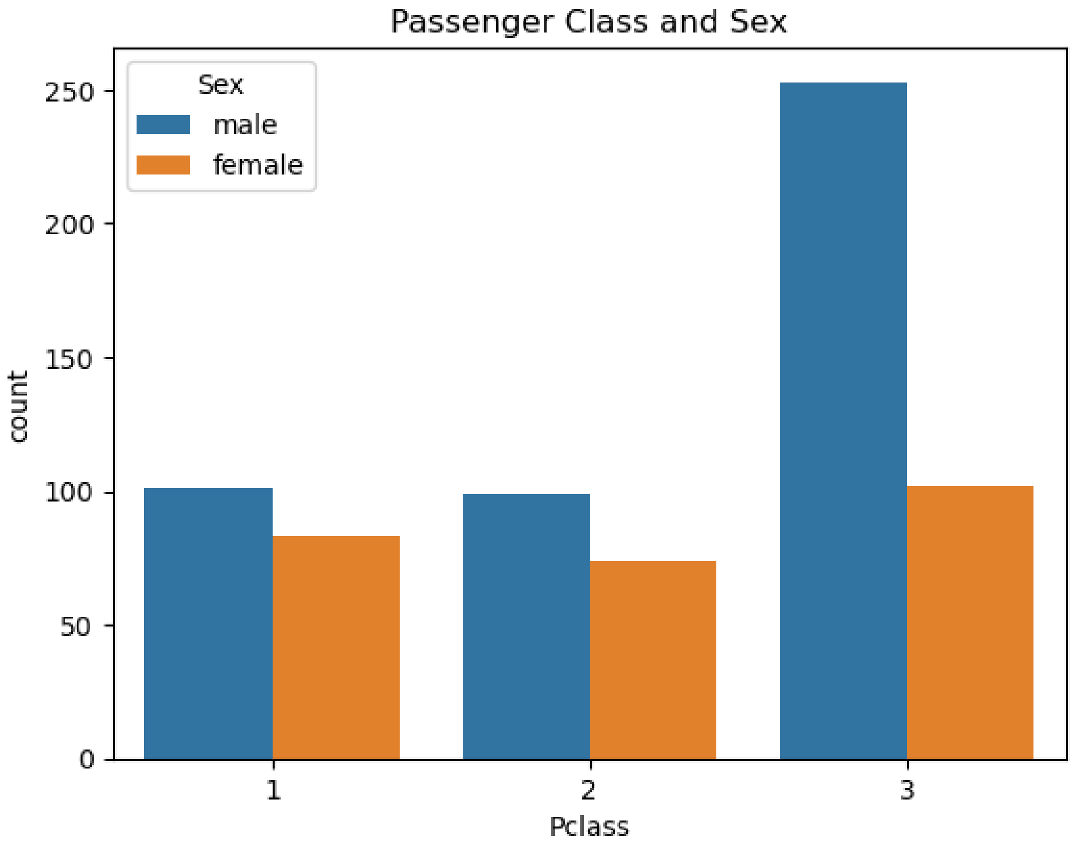
c. What is the gender distribution among passengers?



Almost 64% of passengers onboard were male (453 out of 712).

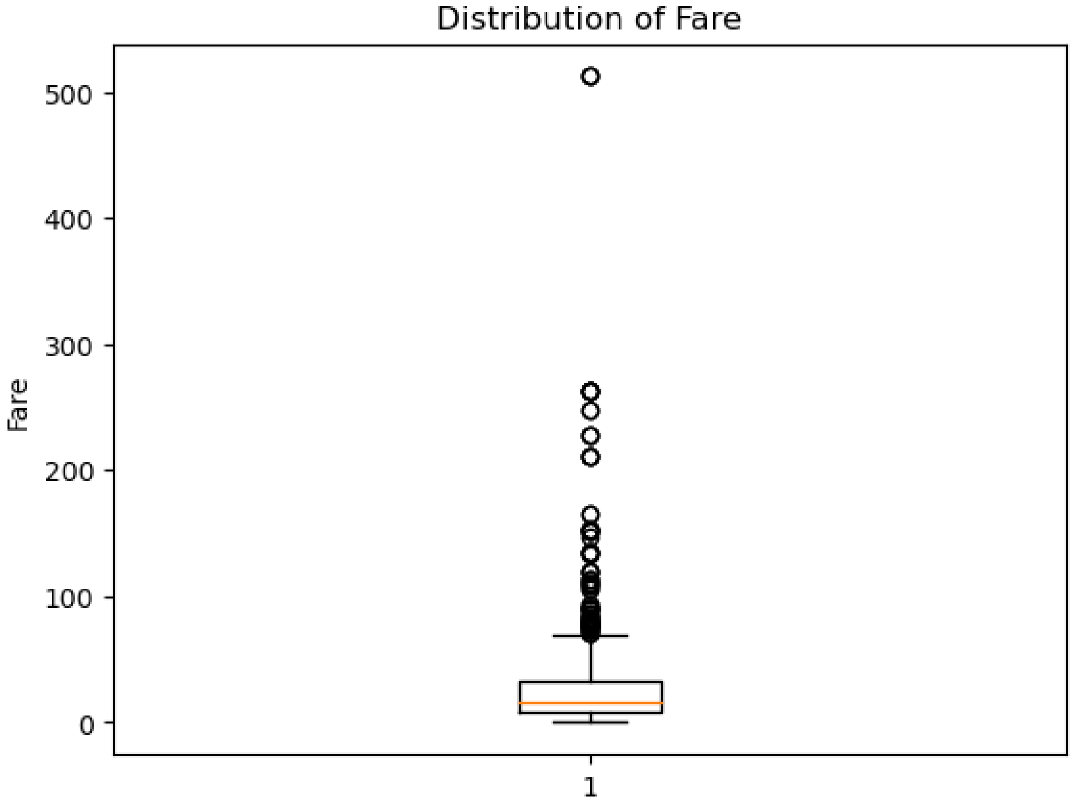


Most of the passengers who died were male, which makes sense since most of the passengers on board were male. An interesting thing to note is that the number of female survivors was almost double that of the male survivors and more than 2/3 of the female passengers onboard survived.



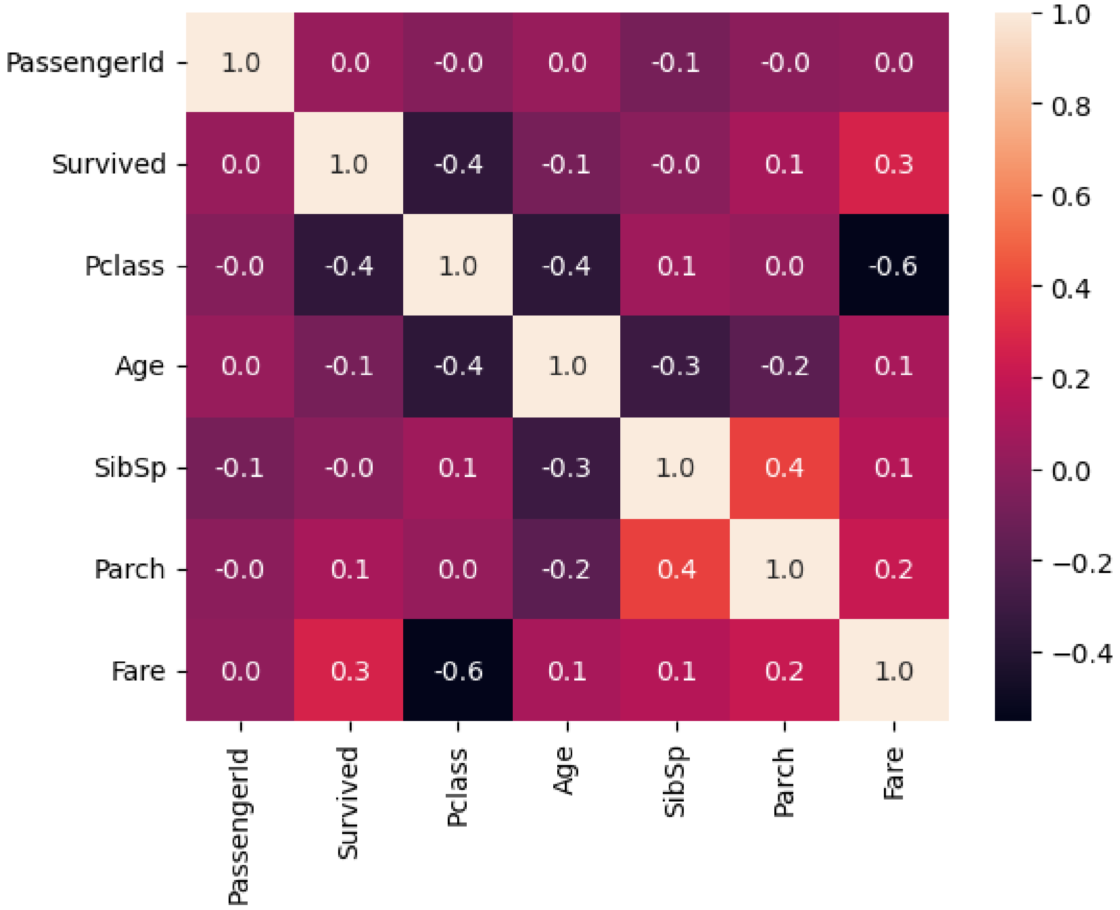
The most number of male passengers were in Class 3 which again makes sense, since the most number of passengers were in class 3 and most of the passengers are male. Interestingly, there isn't much difference in the number of female passengers in each class.

d. How does fare vary on the Titanic?



The distribution of fare is shown using a box plot (aka a box-and-whisker plot) here the box represents the interquartile range (Q3 – Q1) and the whiskers extend to the max and min values respectively. The orange line represents the median and all the points extending past the max are considered outliers. We know from the describe() function that the max value of fare is £512 and the minimum was £0, the mean was £34.
From the box plot, we can determine that all values above 100 are considered outliers and that the median is less than 50.

e. Is there a correlation between different numerical features in the dataset?



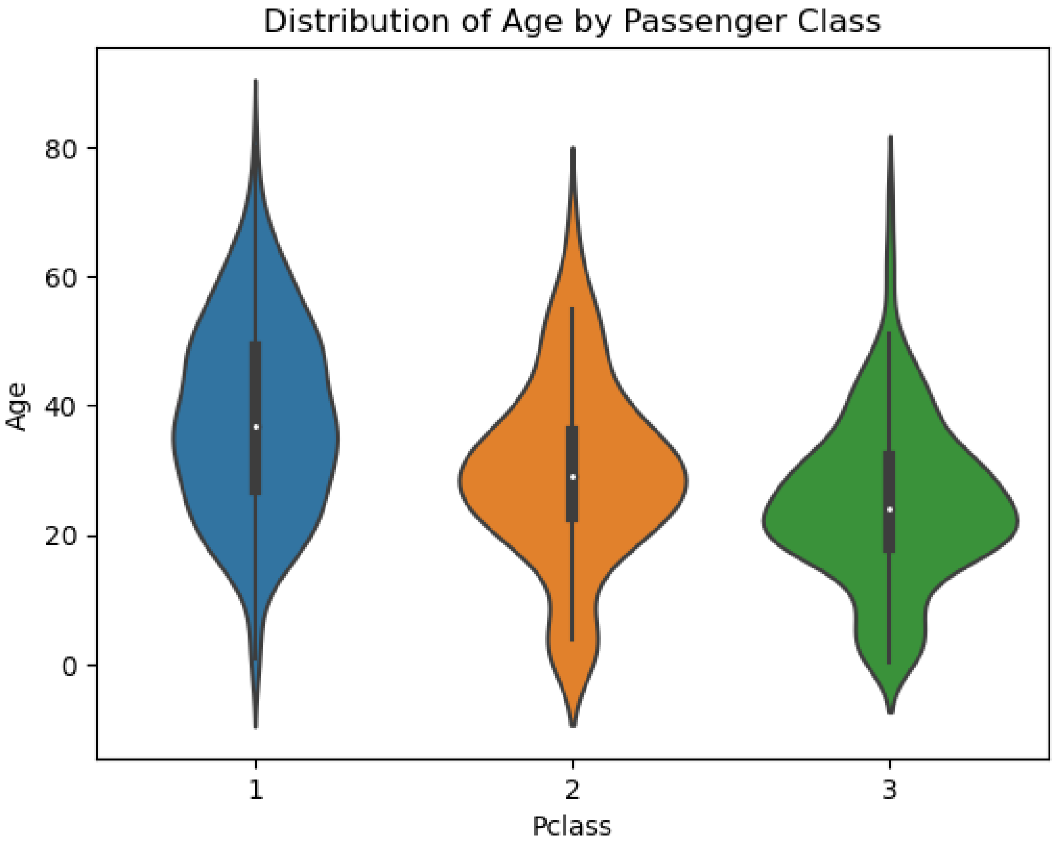
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.0	0.0	-0.0	0.0	-0.1	-0.0	0.0
Survived	0.0	1.0	-0.4	-0.1	-0.0	0.1	0.3
Pclass	-0.0	-0.4	1.0	-0.4	0.1	0.0	-0.6
Age	0.0	-0.1	-0.4	1.0	-0.3	-0.2	0.1
SibSp	-0.1	-0.0	0.1	-0.3	1.0	0.4	0.1
Parch	-0.0	0.1	0.0	-0.2	0.4	1.0	0.2
Fare	0.0	0.3	-0.6	0.1	0.1	0.2	1.0

1. The strongest positive correlation (0.4) exists between the SibSp (No. of siblings/spouses on board) and Parch (No. of parents/children on board) which makes sense.
2. There also seems to be a positive correlation between the survival status and fare, implying that the people who paid higher fares were more likely to survive
3. The strongest negative correlation (-0.6) is between Pclass (Passenger Class) and Fare – this also makes sense as I assume that the fare would be more for 1st class and less for 2nd and 3rd.
4. There is also a strong negative correlation between Pclass and Age (-0.4), implying that as we move from 1st to 3rd class, the average age of the passengers decreases. This again makes sense since older people are more likely to have more money.
5. There is a strong negative correlation between Survived and Pclass (-0.4) implying that that more people from 1st class survived compared to 2nd and 3rd class. A point to note here is that we have previously established that the number of passengers

onboard that had 3rd class tickets was significantly greater than the number that had 1st or 2nd class tickets. So, it stands to reason that the death rate in class 3 was also higher. It is also plausible that people from 1st class were given preference during rescue operations.

6. Continuing from point 2, where I noted that higher fare prices correlate to a higher survival rate and from point 3 where we established that fare prices decreased from 1st class to 3rd. We also established in point 5 that people from 1st class were more likely to survive. So, all these points support each other.

f. How does age distribution vary by passenger class?



The median age of passengers in each class decreases as we move from 1st to 3rd class. The maximum age of a passenger in 1st class is also more than the max age of passengers in class 2 and 3. The highest number of passengers below the age of 40 is also in class 3, followed by class 2. So, class 2 and 3 had a greater number of young passengers compared to class 1, an observation that was made before when we looked at the heatmap.

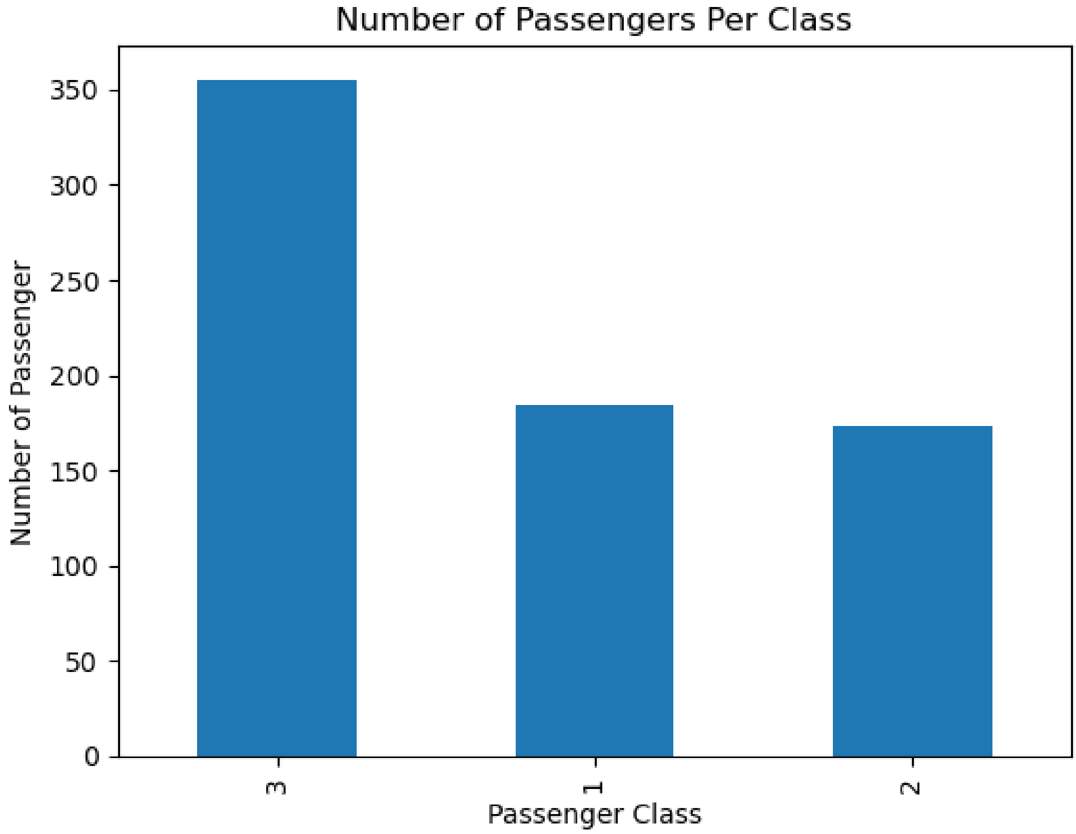
g. What is the survival rate for each passenger class?
We know from the below graph and from when we ran the value_counts() function on the Pclass column of the dataset, that class 3 had the greatest number of passengers (355), followed by class 1 (184) and class 2 (173). The 2nd graph below implies that most of the passengers who died where from class 3, which again makes sense as it had the highest number of passengers onboard anyway. The survival rate of passengers of class 1 was higher than the other 2 classes, a point that has been noted before in the heatmap.

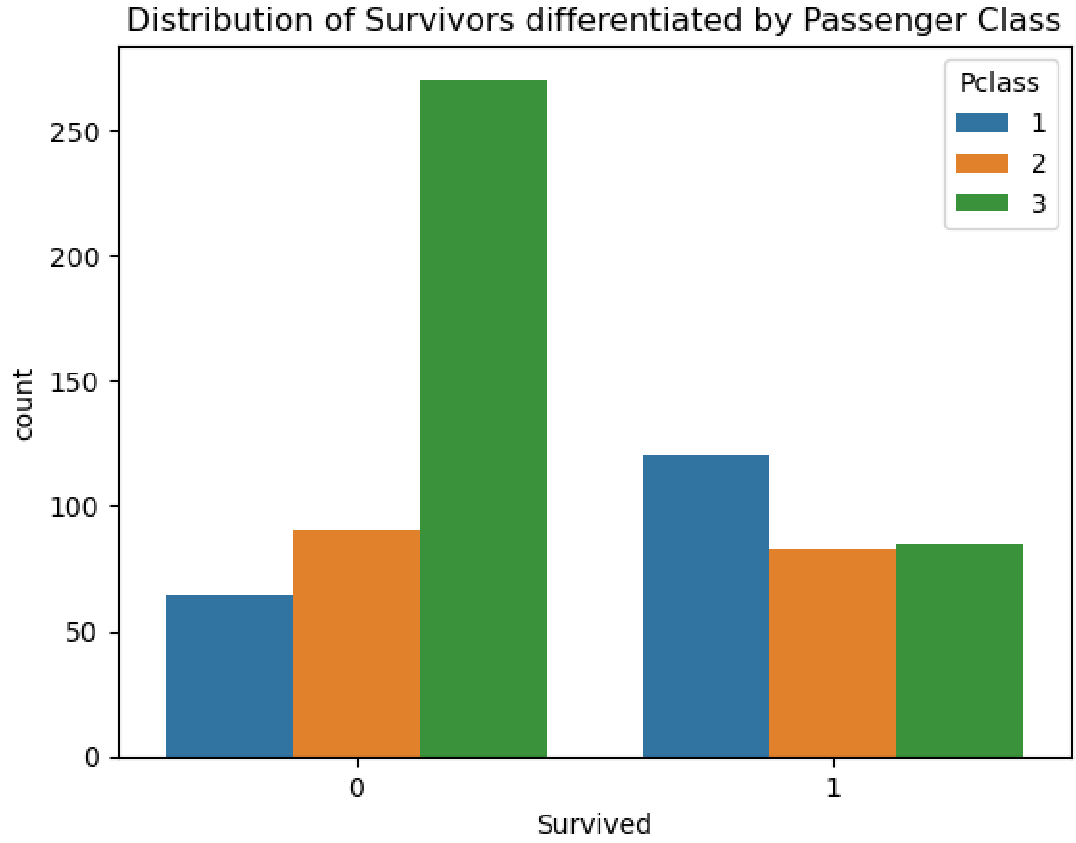
Approx. survival rate:

1st class = 68%

2nd class = 46%

3rd class = 23%





h. Are there any interesting relationships between numerical features in the dataset?
None, that I haven't already mentioned in the heatmap answer.