Exploratory Data Analysis

	Survived	Pclass	Sex	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare
0	0	3	male	22.0	1	0	7.2500
1	1	1	female	38.0	1	0	71.2833

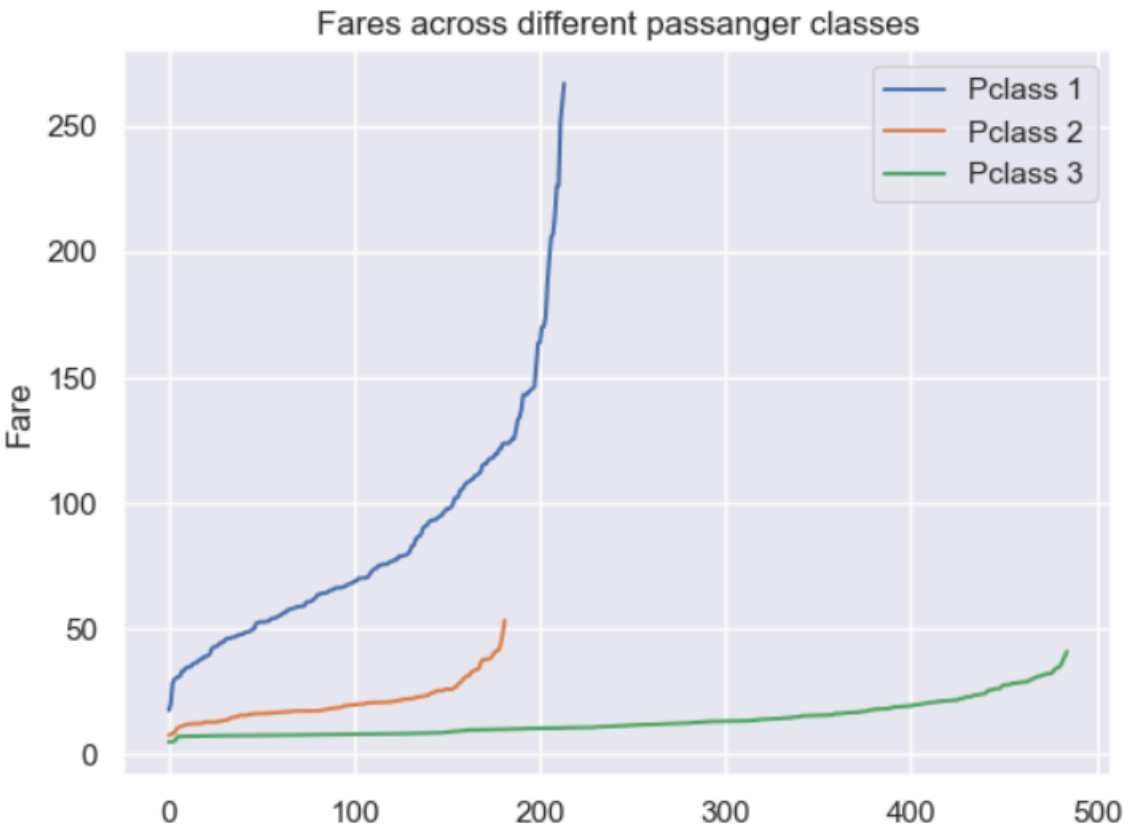
1). Features in the original dataset contain 7 columns.

2). "Survived" being the dependent feature with label 1 and 0 for yes and no respectively.

3). "Pclass"(passenger class), "Sex","Siblings/Spouses Aboard","Parents/Children Aboard" being categorical features.

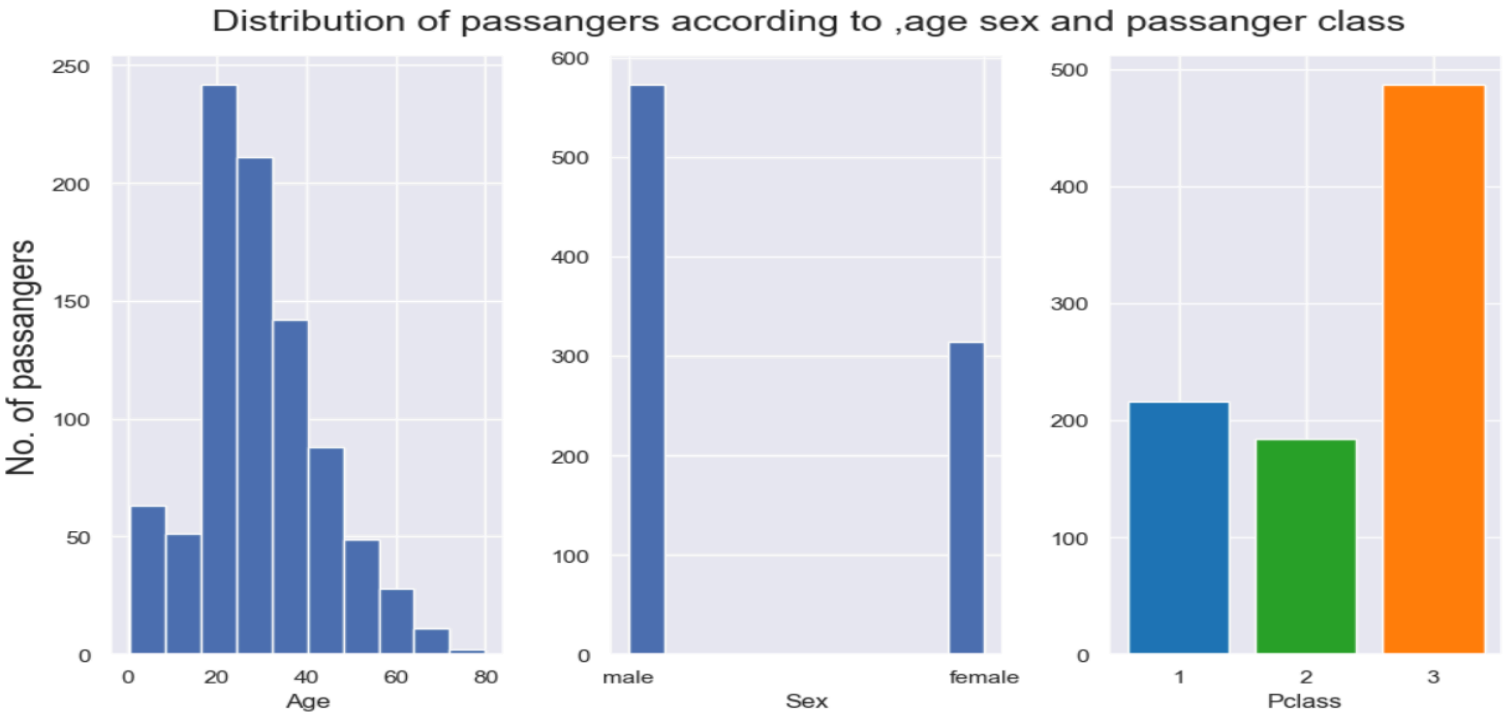
4). "Age" and "Fare" are numerical features.

5). There are 3 passenger classes 1,2,3. 1 being the most expensive and 3 being the least.

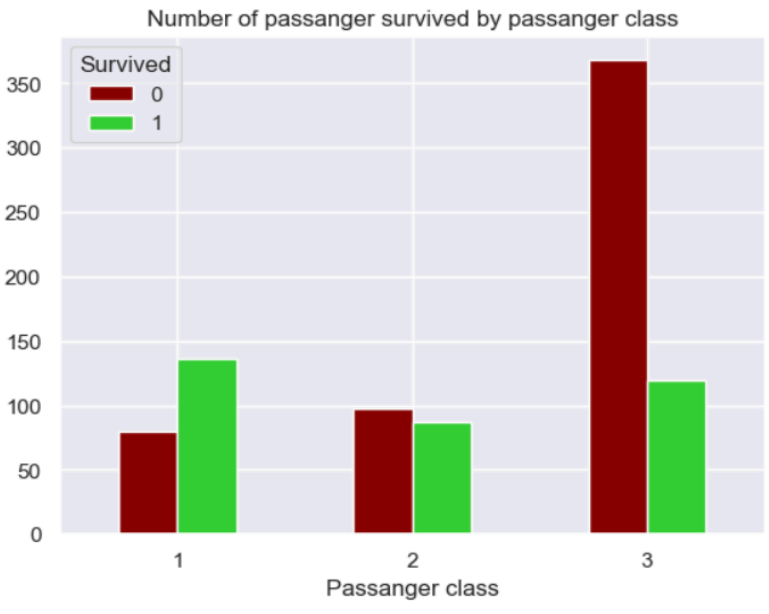


6). The high deviation in the fares of passenger class 1 class suggests existence of sub classes in passenger class 1.

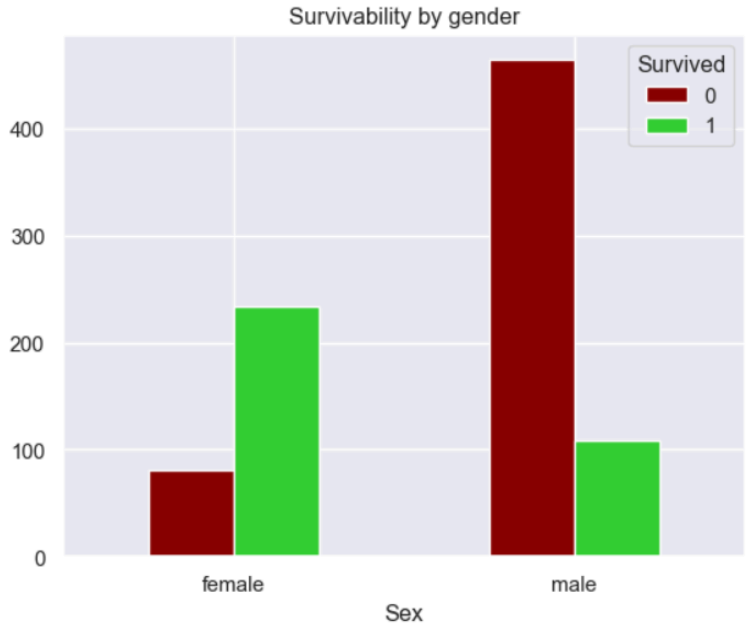
Demographics on the Titanic.



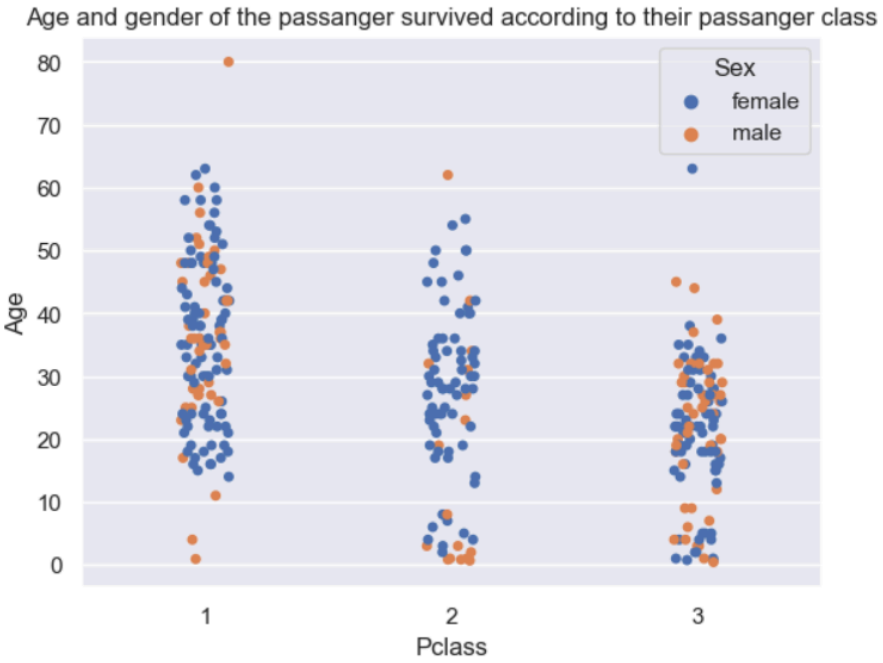
Titanic contained mostly males in passenger class 3 in the age range of 20-30.



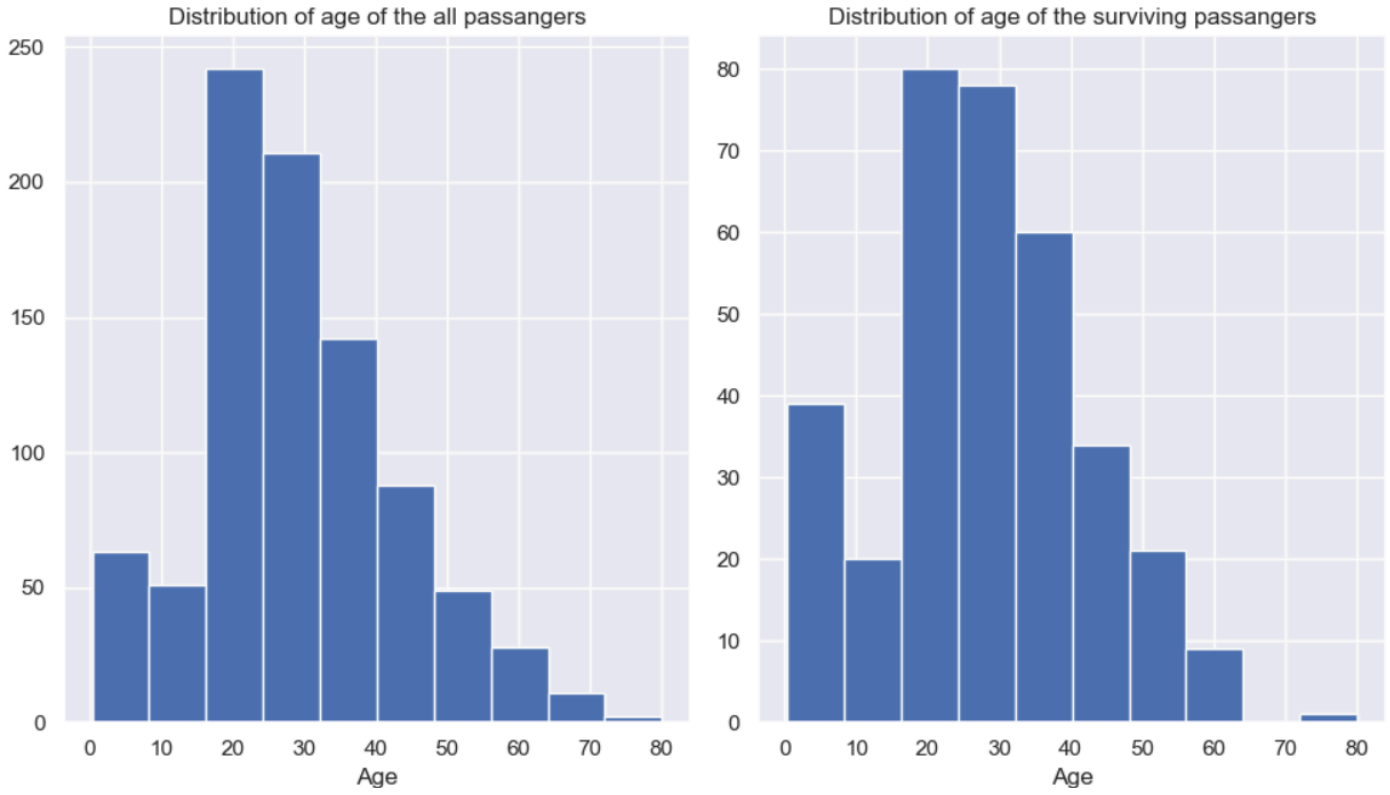
Survival rate of passenger class 1 is highest at around 63% i.e. being in pclass 1 improved the odds of surviving.



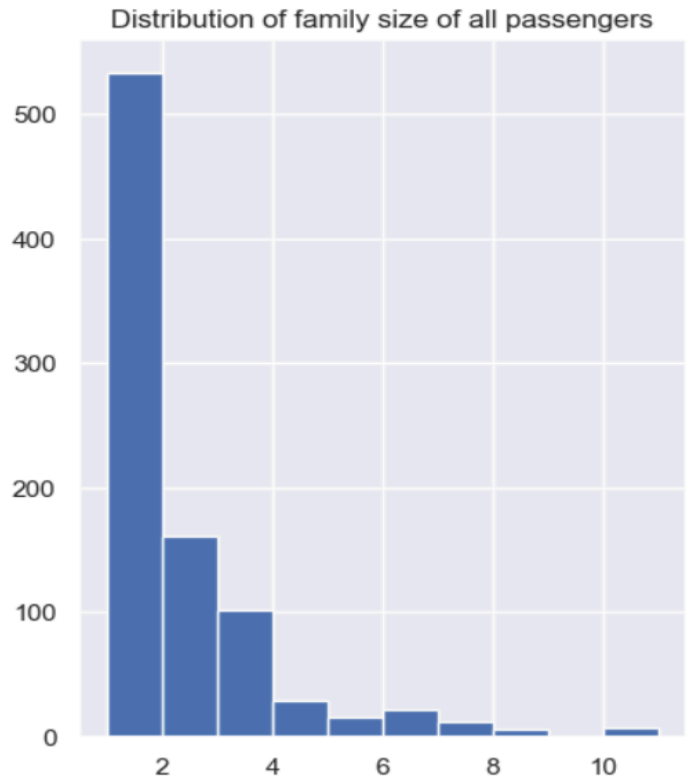
Survival rate for female passengers is much higher than of male passengers.

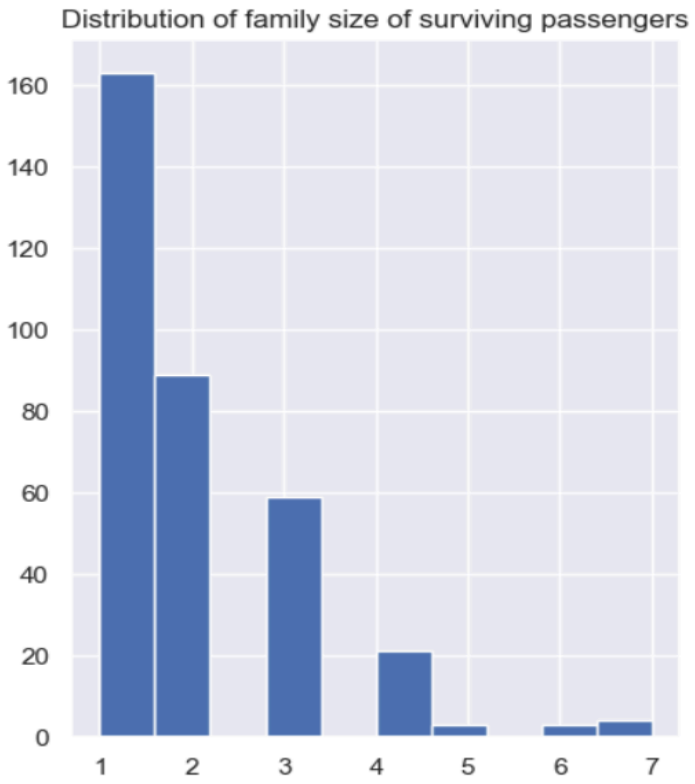


This swarmplot shows individual data of surviving passengers.
A few outliers can be spotted from the visualisation.

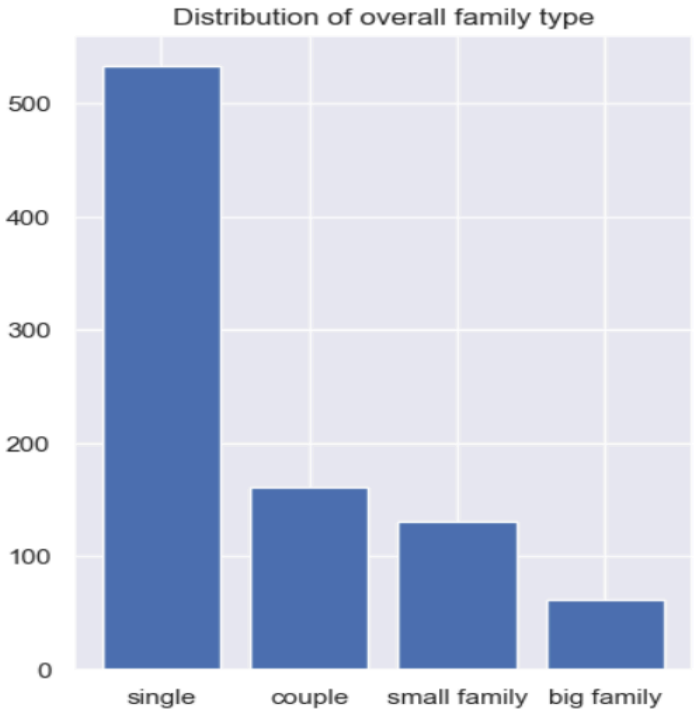


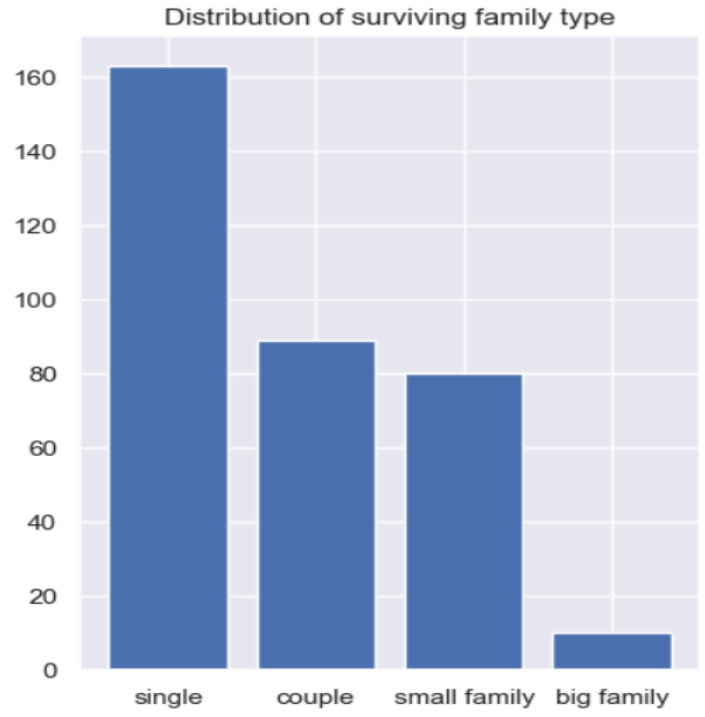
The frequency of babies and infants(age 0-10) has increased in the surviving passengers with the rest of the distribution remaining the same with the exception of age 65-70.





The family size is calculated as 1+siblings/spouses+parents/children.



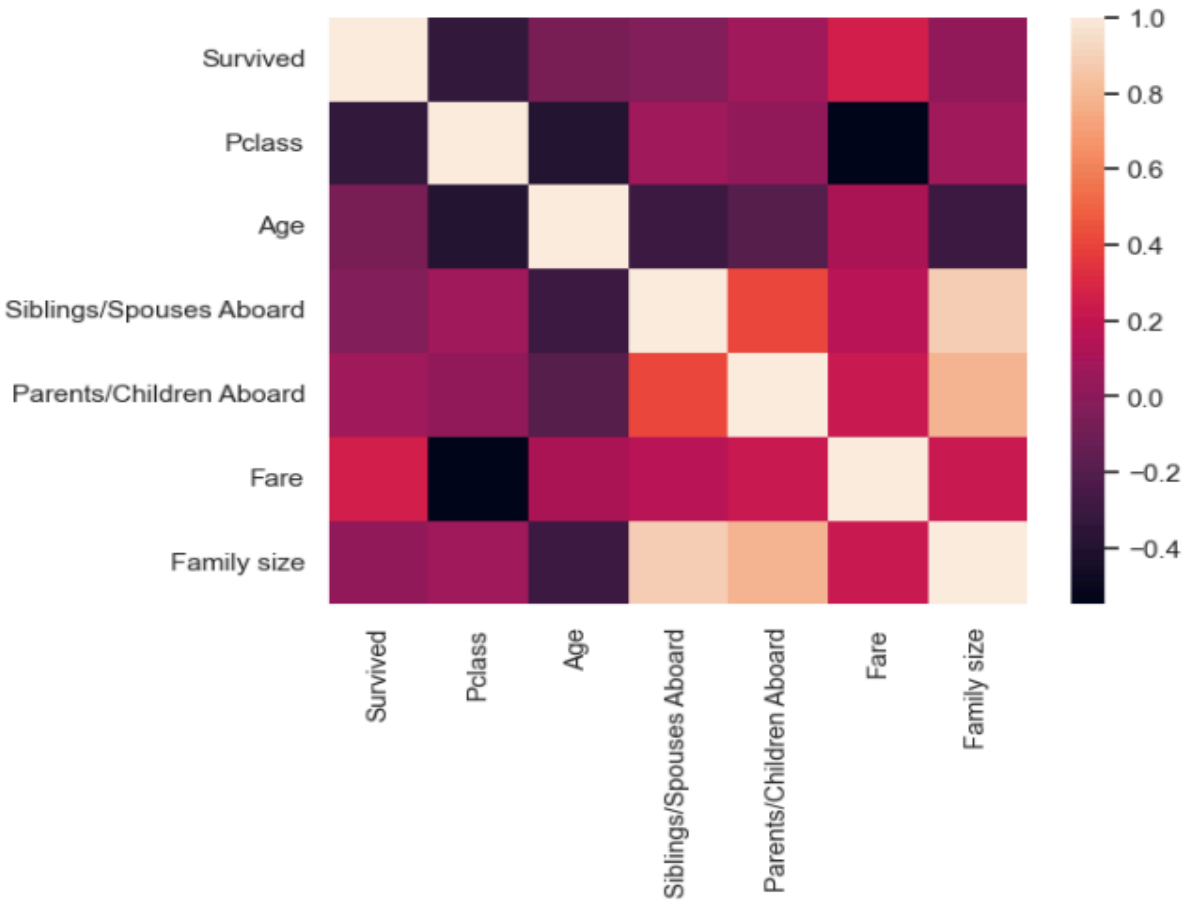


Family size is grouped into 4 family types, single(1),couple(2),small family(4),big family(4+).

The frequency of couple and small families has increased in survivors list whereas that of big family has decreased.

Correlation Matrix.

	Survived	Pclass	Age	Siblings/Spouses Aboard	Parents/Children Aboard	Fare	Family size
Survived	1.000000	-0.336528	-0.059665	-0.037082	0.080097	0.256179	0.014673
Pclass	-0.336528	1.000000	-0.391492	0.085026	0.020252	-0.548919	0.068244
Age	-0.059665	-0.391492	1.000000	-0.297669	-0.193741	0.112329	-0.300297
Siblings/Spouses Aboard	-0.037082	0.085026	-0.297669	1.000000	0.414244	0.158839	0.890595
Parents/Children Aboard	0.080097	0.020252	-0.193741	0.414244	1.000000	0.215470	0.782864
Fare	0.256179	-0.548919	0.112329	0.158839	0.215470	1.000000	0.216250
Family size	0.014673	0.068244	-0.300297	0.890595	0.782864	0.216250	1.000000



Several features are correlated to each other and can be ignored to prevent multicollinearity.

The features “Siblings/Spouses Aboard”,”Parents/Children Aboard” and “Family size” are combined to make a new feature “Family type”.