Problem 1- Data Exploratory Data Analysis

INTRODUCTION

OVERVIEW OF DATASET

The data used for this exploratory data analysis is publicly available on Kaggle website as titanic passengers' data. This dataset contains the information about the unforgettable event of past that was the titanic immersion. The data includes the information about each passenger in terms of their unique identifier i.e., passenger ID, the class in which they were travelling (1^{st},2^{nd} & 3^{rd}), name of person, the gender, age, number of sibling or the life partner, number of kids the person have with them, the price they paid for the journey, and lastly the location from which they got on board for the journey on the boat towards the final destination. Most importantly the data about the individual whether they pull through from that event or not.

Although the chances of getting out from the event were less and based of the person's destiny. But from the available data we can figure out few patterns that might have helped the individual to keep their body and soul together. This was we are going to explore about the data by seeing the patterns of each variable in relation to the survival chances.

SCOPE OF ANALYSIS

The scope of this report is limited to the exploratory data analyses of training dataset and finding the way in which this data can be used for the prediction. We took the training data set into consideration as it contains most data of the full data set and has the target variable which can also be used to see its relationship with other variables. This can help us to see how other variables can be taken into consideration for the prediction process while exploring the dataset. All the variables will be explored in relation to each other and with the target variable i.e., survivors. The purpose of this report is to fetch knowledge about the variables of dataset by visualization and to see the commonalities in the variables while exploring it using different visualizations.

TOOLS USED FOR ANALYSIS

To do all this analysis, the Python libraries such as NumPy and pandas are used for this exploration and seaborn and matplotlib is used to visually present the relationship of the variables with one another.

DATASET SOURSE

The dataset is collected from the Kaggle which is from a competition using titanic data.

DESCRIPTION STATISTIC ABOUT THE VARIABLES

	count	mean	std	min	25%	50%	75%	max
PassengerId	891.000000	446.000000	257.353842	1.000000	223.500000	446.000000	668.500000	891.000000
Survived	891.000000	0.383838	0.486592	0.000000	0.000000	0.000000	1.000000	1.000000
Pclass	891.000000	2.308642	0.836071	1.000000	2.000000	3.000000	3.000000	3.000000
Age	714.000000	29.699118	14.526497	0.420000	20.125000	28.000000	38.000000	80.000000
SibSp	891.000000	0.523008	1.102743	0.000000	0.000000	0.000000	1.000000	8.000000
Parch	891.000000	0.381594	0.806057	0.000000	0.000000	0.000000	0.000000	6.000000
Fare	891.000000	32.204208	49.693429	0.000000	7.910400	14.454200	31.000000	512.329200

All the variables in the dataset have significant meaning. Upon interpretating the statistical analysis of the data. We can see the following information about the passengers in it.

Passenger ID- This is the unique identifier in the dataset. So it doesn't have any meaning attached to it except looking at the min and max values we can say that the dataset contains total 891 observations about the passengers that were travelling on that day.

Pclass- This variable shows the class in which the passenger was travelling. Again looking at other statistical analysis it doesn't make sense as it is a categorial variable. The data has the lowest class names as the 3rd class passenger and the highest class is the 1st class passenger with addition utilities given to them. Upon further interpretation, it is visible that the majority of the people belong to the 3rd class, that contains more than half number of passengers. Additionally, the average value is 2.3 which shows that the passengers are evenly distributed in the given categories of the classes available to them for the journey.

Survived- it is categorical data that has only two values. Such as 1 a showing the survived and 0 indicating the person who didn't survived. It can interpret that 38% people were able to pull through the disaster and remaining majority did 'not.

Age- this shows the age of the passengers travelling that day. The minimum values shows the youngest person in the journey was the infant. And the max value indicates that the oldest person in the journey was the people whose age were 80 years at most. The variation in the dataset is high indicating that the variety of ages groups of people were travelling.

SibSp- This shows the number of siblings/ spouses. The min values shows that there are people who were alone in the journey and max value show that there is at least a passenger with 8 cousins with them on board. The data shows that majority that is more than 2/3rd of the people didn't have any siblings or spouses with them on board.

Parch- this shows the number of people with parents or kids with them. The min value is 0 showing that there are people with no parents or kids with them on the other hand the max value us 6 indicating that minimum 1 passenger has max 6 kids and parents with them. similar to the SibSp the majority of the people are alone and don't carry along their parets or kids in the journey.

Fare- This has the minimum value of 0 indicating that few passengers got lucky in may be getting the ticket for free in offer or some promotional campaign from the boat company. And the most expensive ticket price was of value 512.32 showing that this could the price of the person travelling in the 1st class or has the longest journey.

CORRELATION ANALYSIS

From the above presented visual about the correlation among the variables. It can be interpreted that the survival and the class have negative correlation indicating that as the class of the passenger increases so then the chance of survival increases. And the obvious interpretation can be seen that the passengerID and survival have very weak correlation between them, reflecting that there is no relation between them. It is true that unique identifier will not have any sort of relation with the chances of survival.

Similar to the P class earlier relation with the survival, the pclass also seems to have negative correlative with the Fare, as the fair of ticket increases the class in which the person is travelling is increases as the higher class contains more facilities and perks as compared to the low class ticket. This makes it obvious to have the negative correlation between them.

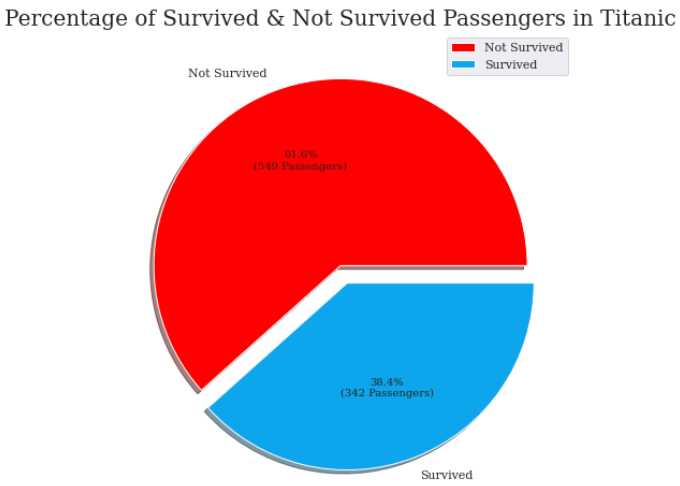
Looking at the age of the passenger and the survival, it can be seen that as the person is younger they have less chances of survival as compared to the elders. This is also true to the reality as the elders were more capable of enduring the ice water during the submerge of the titanic as compared to the kids. There is little relation between the SibSp and Parch and Survival variables showing that as the people have a greater number of people on board with them they had less changes of surviving as they wont be able to fully focus on them to save and unless would might have panicked looking at their close one die in front of them.

Moreover, the fare and survived has the positive correlation. Which shows that those who had paid more for ticket has more chances o survival which can be also seen in a way as seen earlier with the class and the survival. Usually, higher class tickets are expensive.

NUMERIC DATA EXPLORATION

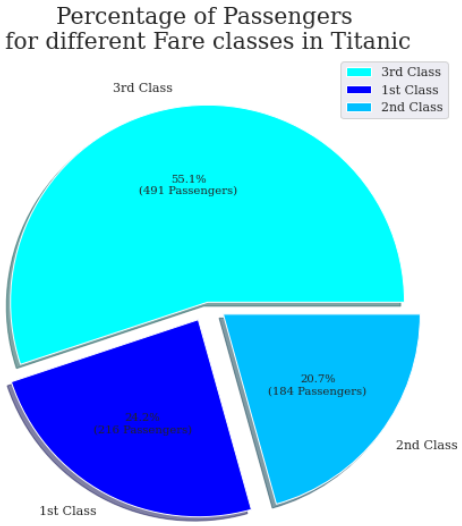
UNIVARIATE ANALYSIS

Survived



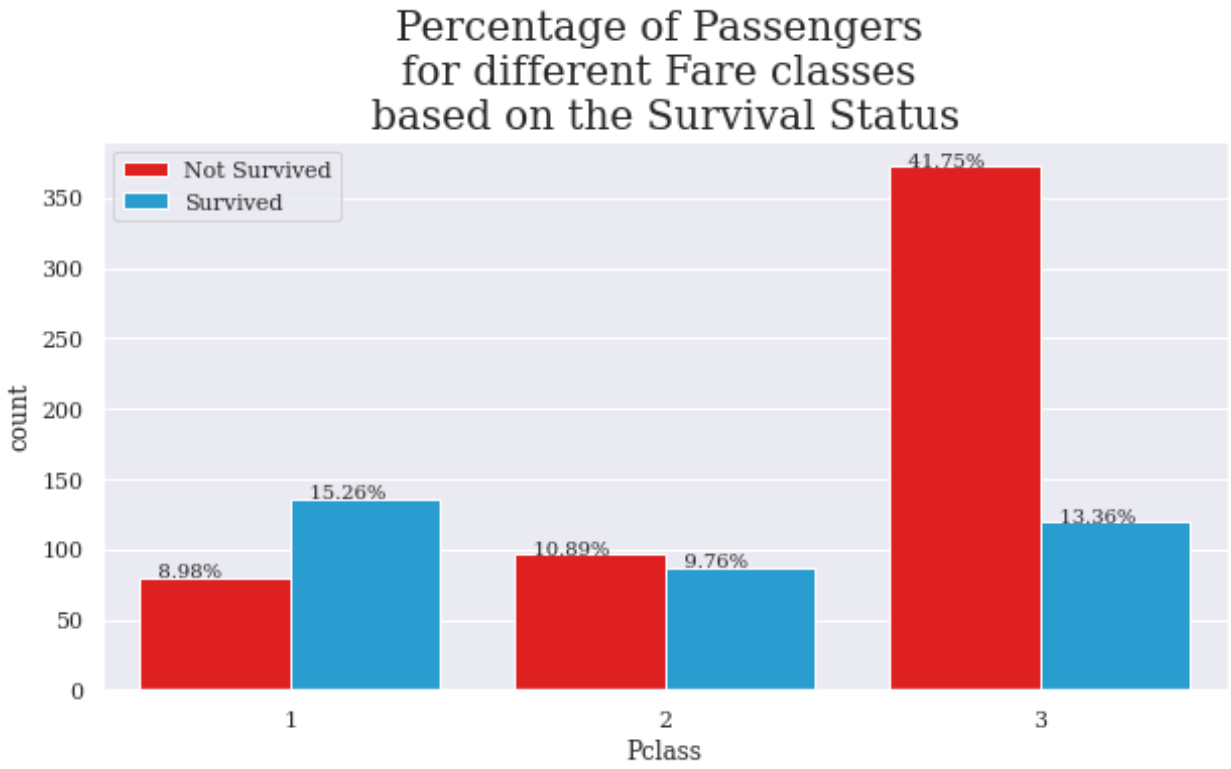
From the above pie chart, we can see the division among the people who survived that disaster and the people who didn't. as the disaster was a huge event in the history as the large amount of people died. This is also seen in the above chart that the people who died were about 61.6% of the population and only the 38.4% people were able to make out of that disaster from the dataset we are given. As the chances of survival was less so we see more people died than the people who survived.

Passenger class



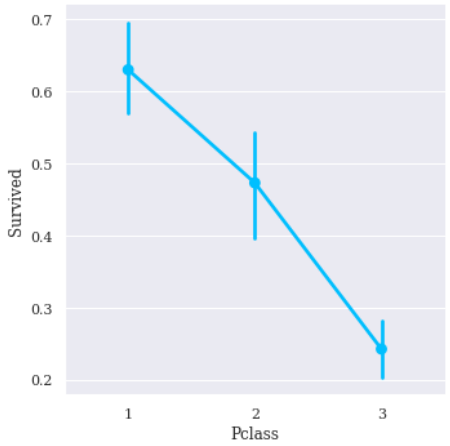
This pie chart represents the people % of people who were traveling in different type of ticket groups. And we can see that majority of the people were travelling from the 3rd class ticket as it is obvious that the lower the class the lower the fare from the correlation as well. So, we can say that more people tend to get the less priced ticket. And about same proportion of people from the 1st and 2nd class i.e., 24.2% and 20.7% were travelling at higher priced ticket. And they are almost equally distributed in same proportion in both the categories.

Passenger class & survived



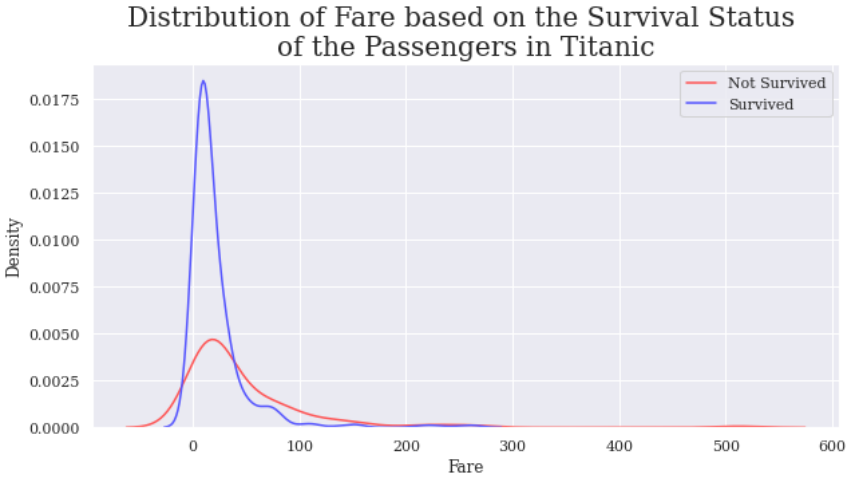
This bar chart represents the percentage of people travelling from the different category of class of ticket and their survival status in each category. We know that majority of people travelled from the 3rd class ticket so, we can see that 41.75% died in that class people and only the 13.36% people survived. Majorly the 3rd class people died. Secondly the 2nd class people died in terms of proportion. And we can see that the class impacted it. As the 1st class people survived more a even though there were less number of people travelling from that ticket of class. It can be interpretated that the higher-class people were given priority or they might have special survival kit available to them as perk on the journey. So, they could have get more access to the life jackets than the people who were travelling from the 3rd class ticket.

Passenger class & survival rate



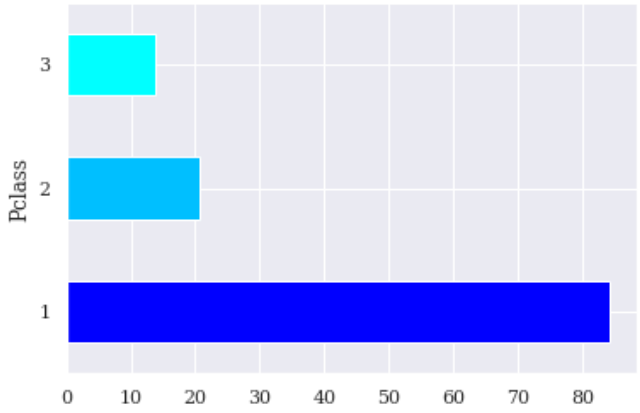
From the above line chart as well, we can see that the rate of survival decreased from as the passenger is travelling from the lower class. We see a significant drop-in rate as the 1st class people has higher rate of survival and the 3rd class has the lower rate of survival in terms of the people travelled from that category and percentage of people died out of that category.

Fare- survival



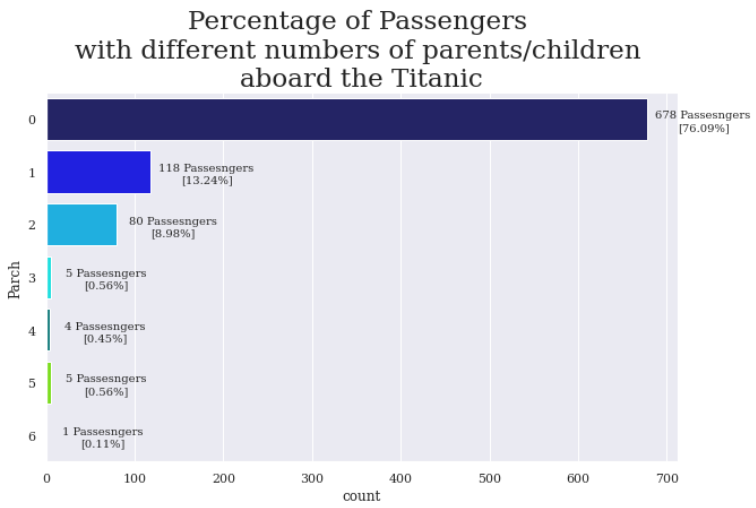
The red line indicates the not survived and the blue shows the people who survived. And this is validated from the above graph that the people with higher class of ticket that means the higher fare ticket survived more in number as compared to the people with the lower-class ticket or the cheaper fare ticket were small in number in the status of their survival. There is a huge peak in the survival line as the fare increases then the not survived people.

Passenger Class Mean Fare



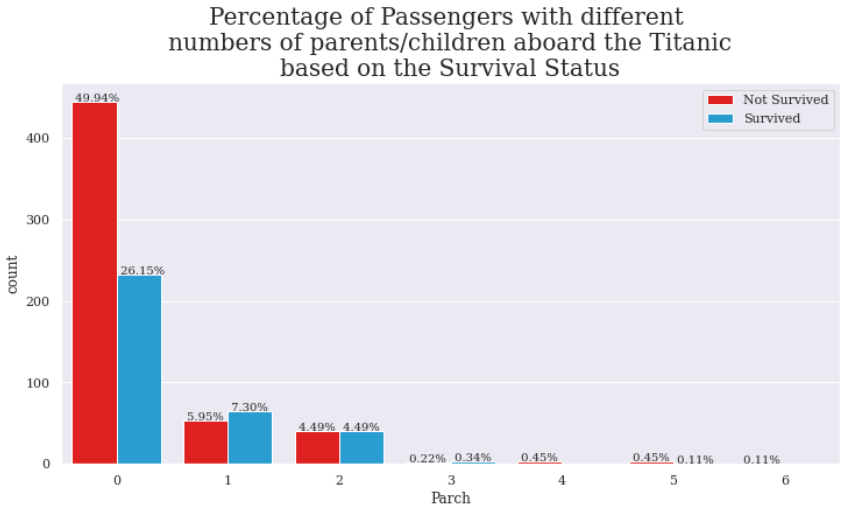
This shows the fare of the ticket on the mean level. And it validates the obvious fact that the 1st class has highest mean fare of above 80 dollars. 2nd class has second highest fare of around 21 dollar. And lastly the cheapest ticket of 3rd class is the fare with the lowest mean of around the 15 dollars. We saw the mean as the tickets are sold in offers during promotion and few get in for less. But even after the reduced price the distinction of price between the 1st ,2nd, 3rd remains the same as only few gets offer.

Passengers with different number of parents/ children



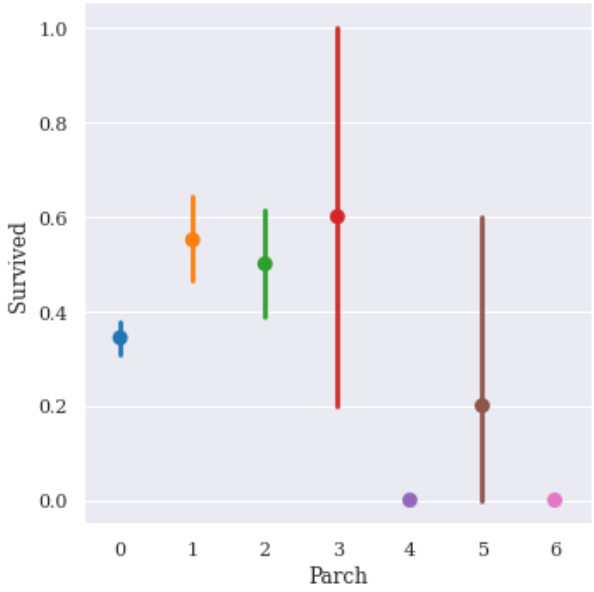
This shows the count of people in each of the category of number of companions they were travelling with. We have 0 to 6 number of companions in terms of parent or as a child with the passenger. And from the above graph it can be seen that a huge number of people that is 76.09% of passengers were travelling alone and had no kids or parents along with them. And only one passenger had 6 number of people along with them in the journey. And the number decreases after the 0 to 5 on the smaller scale.

Passengers with different number of parents/ children & survival



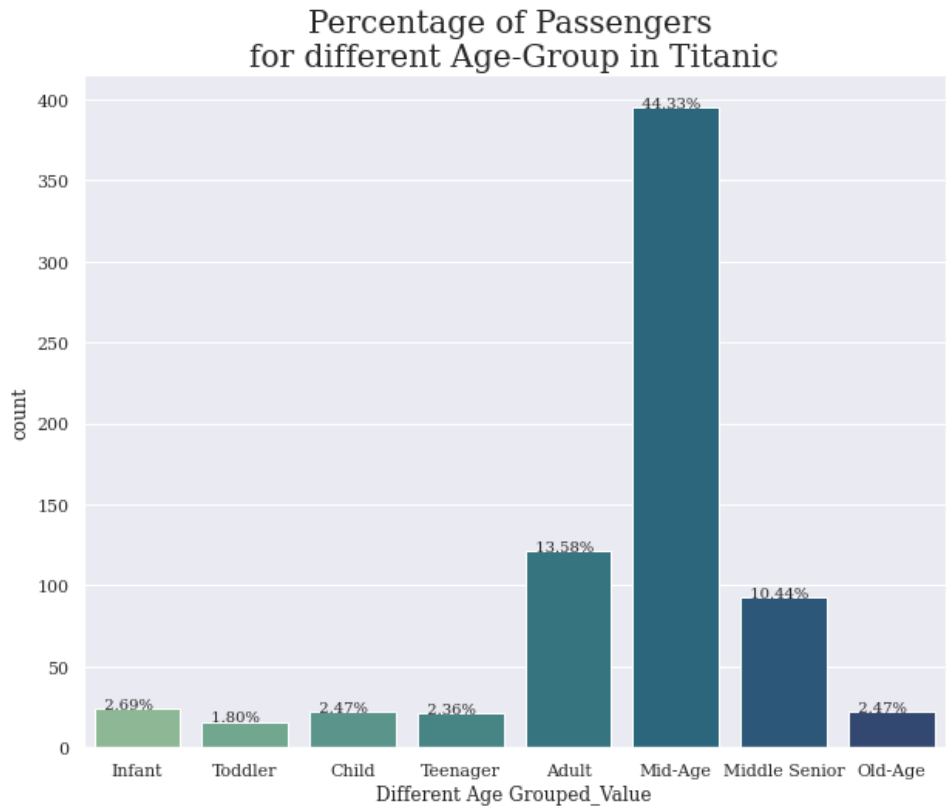
Now looking at the comparison of the people who survived or not survived with the number of companions they had with them on the journey. We can see that count of people who were alone died on high number of 49.94%. Those who had 1 companion with them also did not survive on 5.95%. we can see that the people with 0, 1,2 number of companions were the people who died more. And there were less dead percentage of people in the 3,4,5,6 number of companion.

Passengers with different number of parents/ children & survival rate



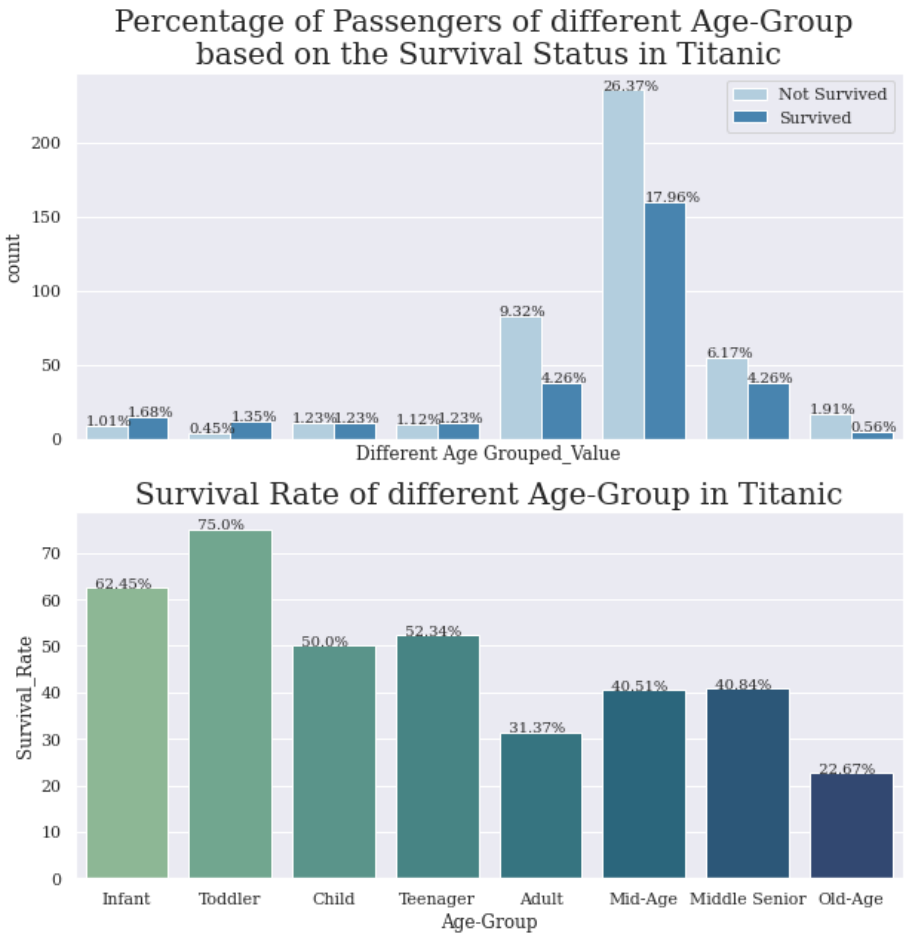
Now looking at the rate of survival based on the total number and out which who survived. This gives us better view of looking at people companion impact on the survival. This can be observed from the above graph that people with 4 and 6 parents/children had lowest survival chance. It is obvious that people gave priority of their kids and parents to survive instead of prioritizing themselves out of love. And the people with 3 parents/kids were the people who had more survival rate as they were less people and they might have handled each other with less number of available life jackets.

Age group based using ranges



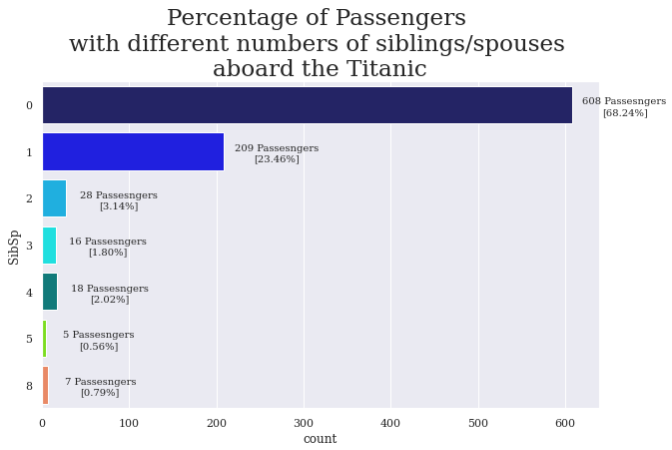
Age is divided into different categories using it numeric values given to us. This division is made based on the universal categorization of people from infant to the old age based in the age number range. We can see that there were around 44% of mid age people on board. This was the highest among the traveler in the boat on the day of journey. As there are the people who might have taken out their family member out for holiday with there kids or parents or alone. So this accounts makes them the highest among all the dataset. Toddler were the least in percentage in total of all the passengers. Same reason can be taken for the adults like the mid age people who might have been alone or with their parents were the second highest category on board in the journey.

Age group & survival percentage and rate

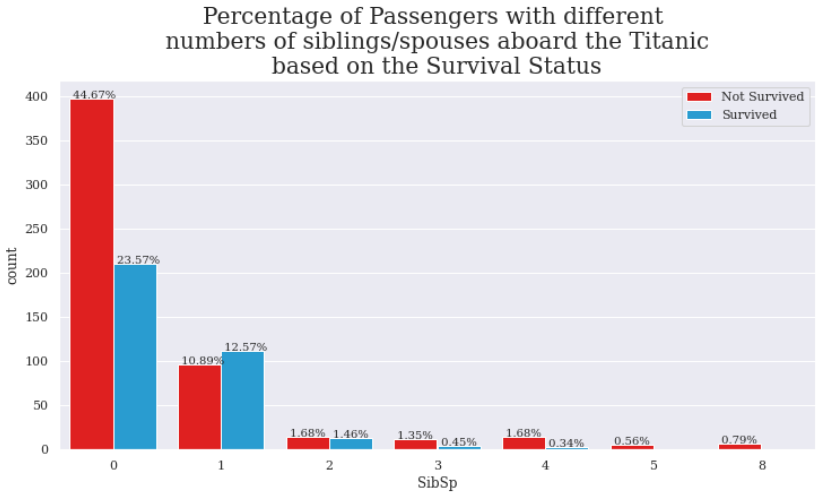


From the top chart we can see that the highest number of people were survived from the middle age category but more than that there were people who died in that category. Looking at the second graph gives us the better look on the people survival rate in each category of age groups. This can be observed that toddlers had the highest survival rate in the dataset. As we know that the kids were given priority for the rescue for survival as compared to others. Moreover, we can see the same for the infant as well. That is why its is on the 2nd highest survival based on the comparison with others. Thirdly we see the teenagers and child who were able to be survived on that day may be because of the endurance or the ability to swim with the support of other life jackets available. Lastly the old age people were the most vulnerable to death. As they were not given priority for giving life jackets, they also had less endurance capability and also were not able to swim as well.

No. of siblings / spouses



This shows the count of people in each of the category of number of companions they were travelling with. We have 0 to 5 and 8 number of companions in terms of siblings or as a spouses with the passenger. And from the above graph it can be seen that a huge number of people that is 68.24% of passengers were travelling alone and had no siblings or spouses along with them. And only seven passengers had 8 number of people along with them in the journey. And the number decreases after the 0 to 5 on the smaller scale.

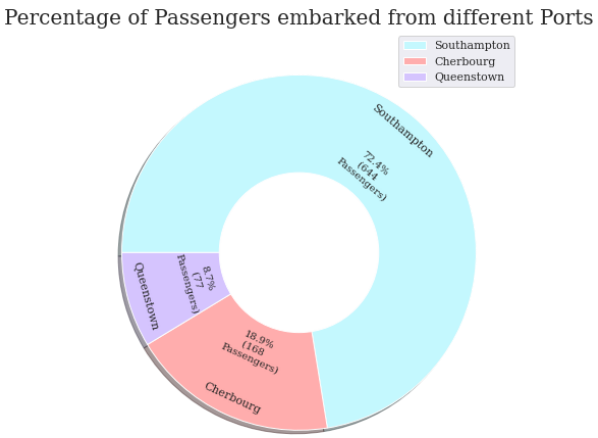


Now looking at the comparison of the people who survived or not survived with the number of companions they had with them on the journey. We can see that count of people who were alone died on high number of 44.67%. Those who had 1 companion with them also did not survive on 10.89%. we can see that the people with 0, 1,2,3 and 4 number of companions were the people who died more. And there were less dead percentage of people in the 5 and 6 number of companion.

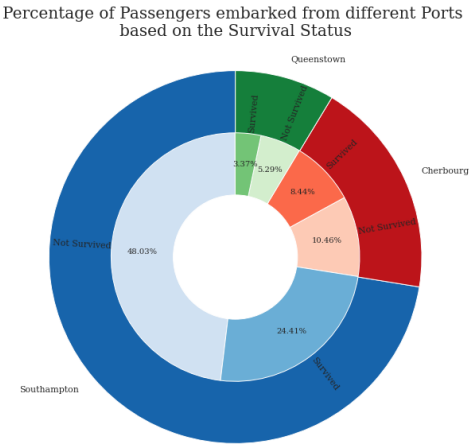
CATEGORICAL DATA EXPLORATION

Univariate analysis

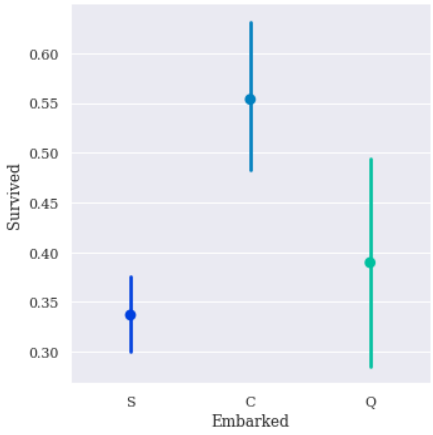
Port of embarkment



The above pie chart shows that majority of the people embarked their journey from the port named as Southampton which was 72.4% of the dataset. Secondly, the 18.9% of people started their journey from Cherbourg port. And only 8.7% of people embarked their journey from the Queenstown port.

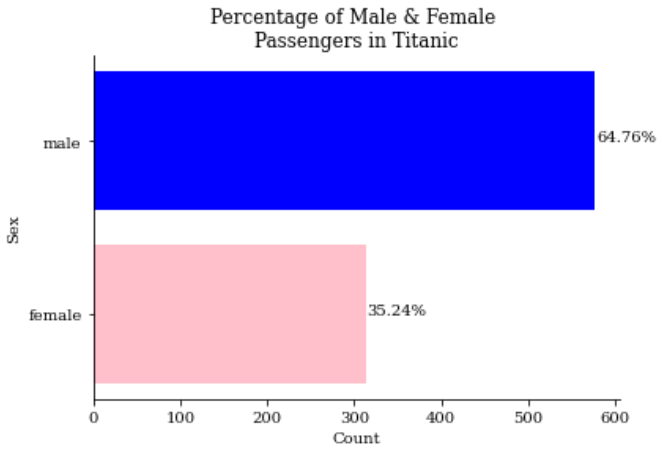


Now looking at the donut chart given above we can see that highest number of people not survived were from Southampton being the percentage of 48.03%. followed by the Cherbourg and lastly the Queenstown.

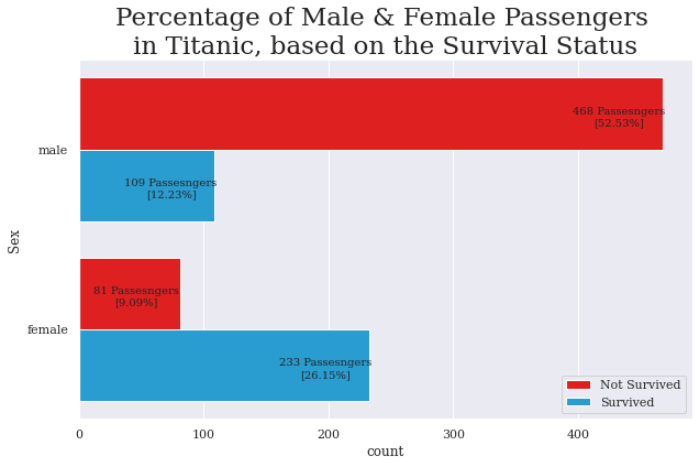


From the above survival rate of people based on their embarked port we can figure out the higher chances if the survival were for the people embarked their journey from Cherbourg port. May be the people board from this port were of adult or younger generation. As they have more chances of survival. Another reason could be the people who boarded from this port might be the last one. And the other who were travelling from the long hour of journey were already tired and couldn't physically help themselves on that day for the survival of their own lives. Lastly most of the people boarded from the Southampton port had the least chances of survival. As they could the people who were with their families. And we know that people alone had more chances of survival than the people with the family members.

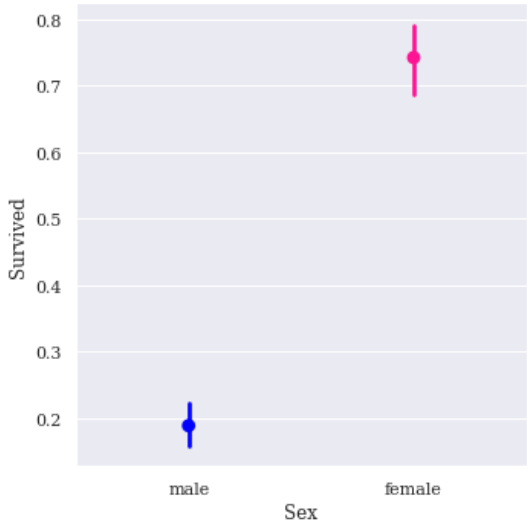
Gender



We can see the count of each category of gender in the above database. This can be seen that there were more males then the females in the journey. 64.76% of people were males and 35.24 were females.



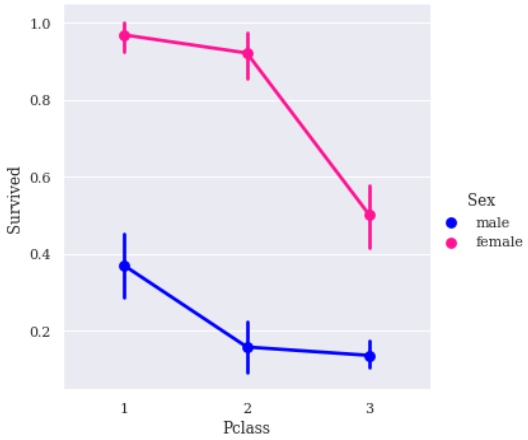
Looking at the gender of passengers who survived that day. We see that there were majority of the male people in the boat. And a smaller number of females. Moreover, we see that the died passengers were highest from the male category. However, there were large number of female who survived as compared to the males in even though they were small in number and less probability of more survival chances.



This shows the rate of survival of both the genders. The female have much higher rate of survival as compared to men's. We know for a fact that on that day the females and kids were given priority in the disaster that happened. That fact can be seen here as well in the above graph looking at the survival rate of females as compared to the males. The females survived more than the male who were able to survive as they were the priority in giving life jackets.

NUMERIC AND CATEGORICAL EXPLORATION

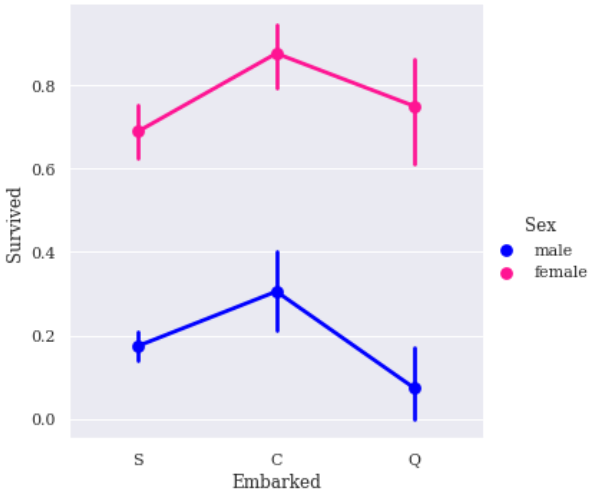
Passenger class along with survival rate and sex



Here in the plot again we can see that upon comparison of female and male. The females had higher survival rate as compared to males. In the female's category, those who were traveling in first class has the highest survival rate and the females travelling from the 3rd class had lowest in survival but still greater than the male's overall survival rate in all the classes of tickets.

In male's category the males travelling in 1st class had more survival rate and 2nd and 3rd had similar survival rate. We can see that same pattern as female had better survival rate than males. On the top of that classes had the similar pattern, higher the class so the greater chances of survival.

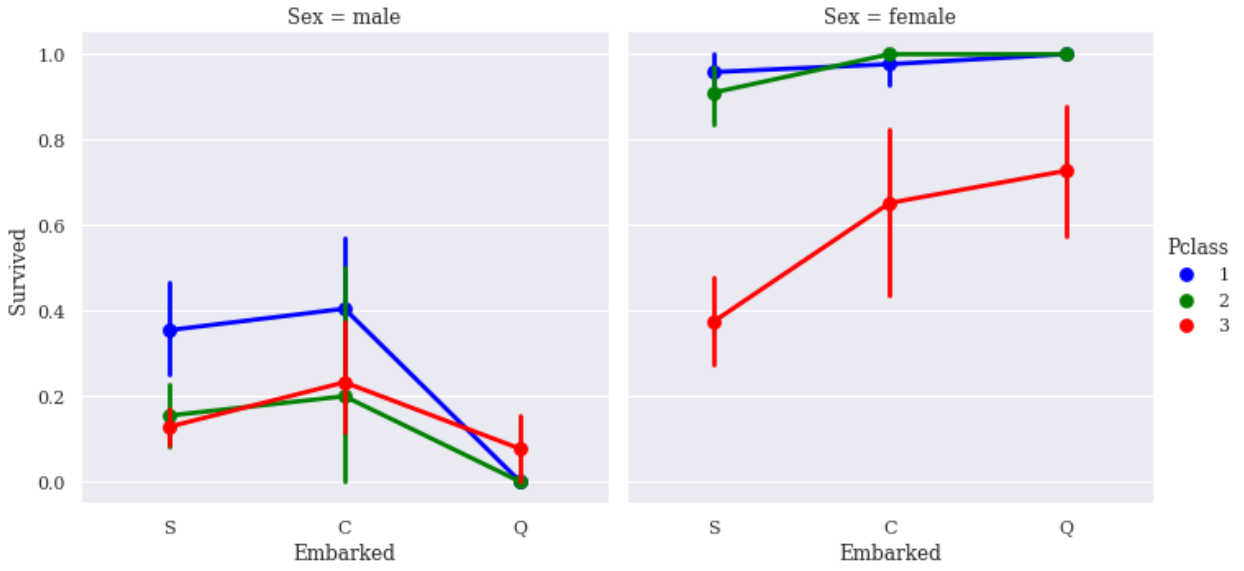
Port of embarkment along with survival rate and sex



From the above plot, if we observe the embarked port, sex and chances of survival. The females had higher survival rate as compared to males. In the female's category, those who were traveling from Cherbourg has the highest survival rate and the females travelling from Southampton had lowest in survival but still greater than the male's overall survival rate from all the respective ports.

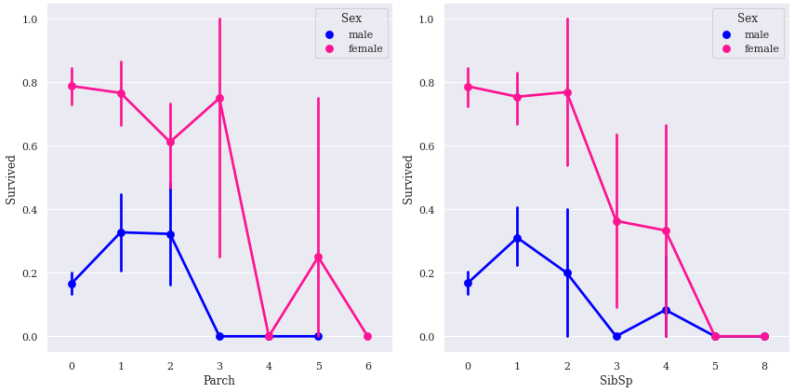
In male's category the males travelling from Cherbourg had more survival rate and Queenstown had lowest survival rate. We can see that same pattern as female had better survival rate than males. On the top of that ports had the similar pattern, Cherbourg is the top in greater chances of survival.

Passenger class, port of embarkment along with sex and survival rate



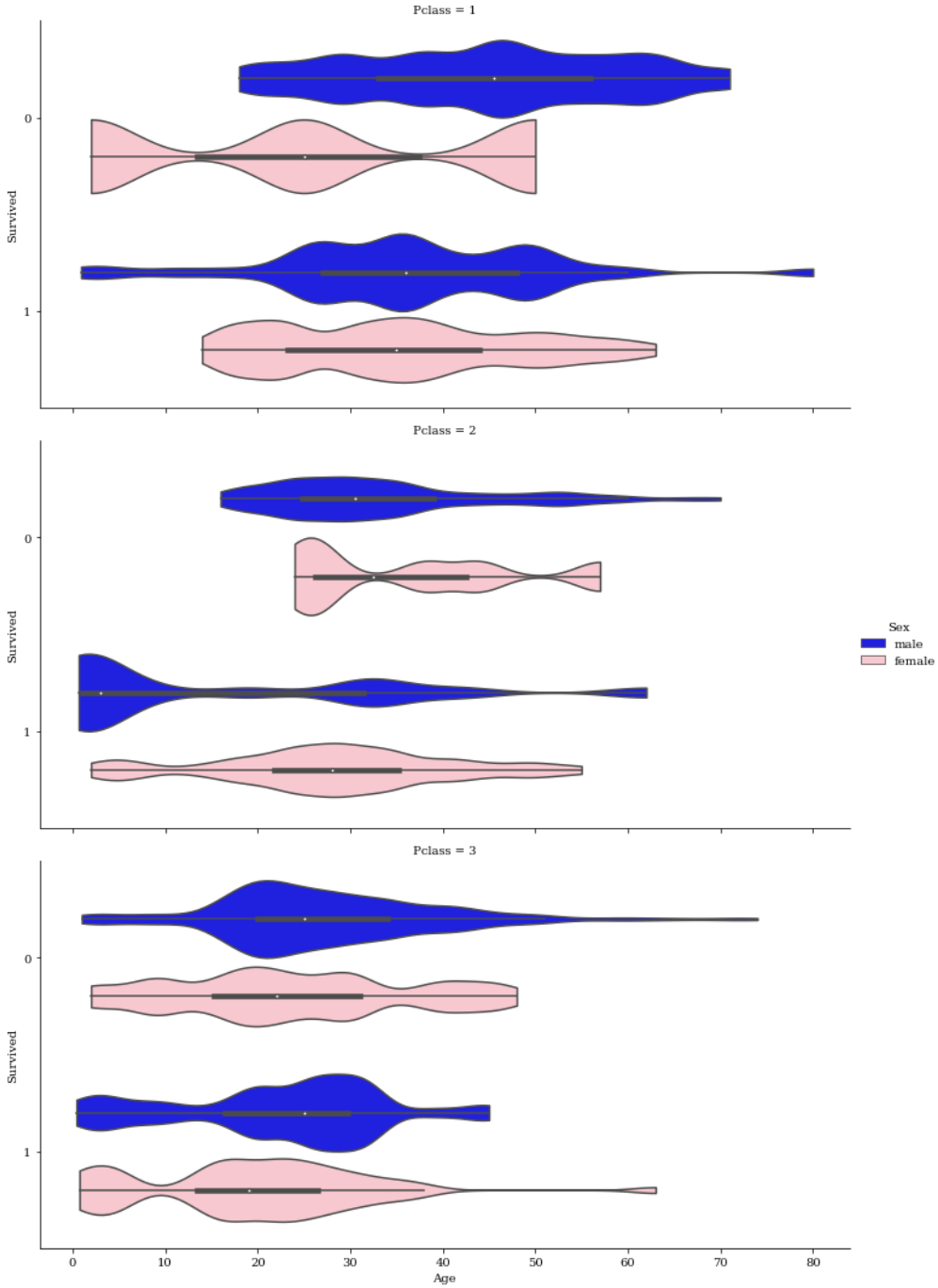
This this chart if we observe the top chances of survival in males, we see that males who were from 1st class ticket and ported from Cherbourg is at the top. And the least chances of survival were for the male traveling from Queenstown in 2nd class ticket. Moreover, if we see for the females we observe that they had top chances of survival from the 2nd class ticket travelling from Cherbourg. And the least chances of survival from the 3rd class ticket porting from Southampton. Overall, irrespective of the class and embarkment port the females have still more chances of survival.

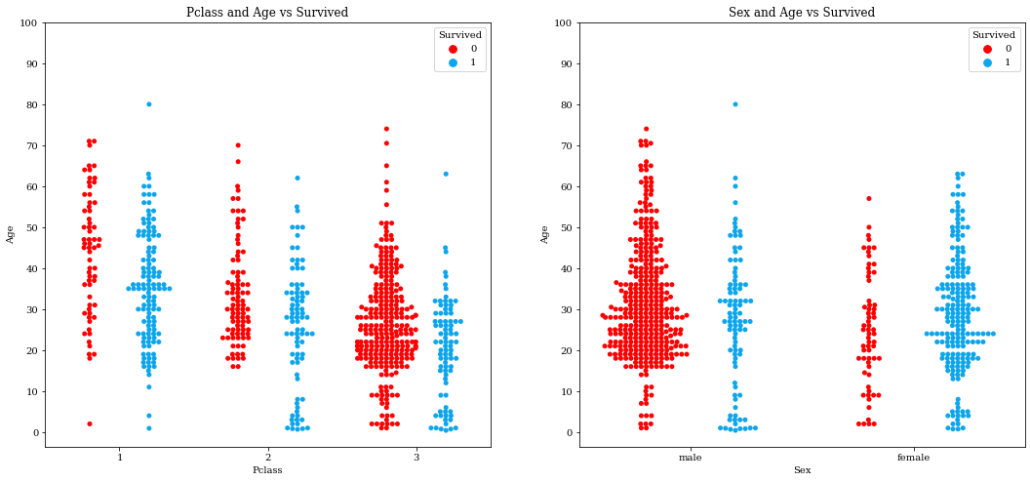
No. of sibling/children/parents/spouse along with sex and survival rate



Females with no companion along with them had highest survival rate and lowest when they had 4 or more companion along with them in terms of parents/spouse/kids. On the top of that males who had highest survival rate when they have 1 or 2 companions along with them and lowest when 3 or more companions along.

Age, passenger class, sex along with survival rate





Now taking age, sex, ticket class as the factor of survival. As we see in the graphs the line for survival rate is dropping as the age is increasing this shows that infants had higher chances of survival than the elders. in the males the passengers from the 2nd class tickets who were elders has the lowest survival rate. In opposition to this the females from the 1st class of old age had the highest survival rate. From all the diagonals we see that females and kids were advantage in this disaster as they were given more priority than others.

CONCLUSION

SUMMARY OF FINDINGS AND INSIGHTS FROM THE EDA

The EDA involves the class in which they were travelling (1st,2nd & 3rd), name of person, the gender, age, number of sibling or the life partner, number of kids the person have with them, the price they paid for the journey, and lastly the location from which they got on board for the journey on the boat towards the final destination. Using the bar graphs, pie charts, line charts and histograms to explore the data. It can be seen that the survival rate was effected by the gender, ticket class, and most importantly the number of companions with they were travelling. The females and younger people had more chances of survival. And the ticket class also had impact on the survival as the top-class ticket had more chances of survival than the lower-class ticket.

RECOMMENDATIONS FOR FURTHER ANALYSIS (PREDICTIVE ANALYSIS)

This analysis can be used while model building for the prediction of the dataset for those who survived and those who didn't survived that day. Upon the exploration of data set, we see that all the variables are present and only the age is not given for all the datasets. We can remove that and build the logistic model for prediction keeping in mind that sex, age, port of embarkment, are important feature to be considered as the significant predictor. However, the fare and p class seems to be same role in the analysis so we can remove anyone of them for the prediction. And on the top of that while we predict the data, we should not use the ids as it is of no use and use categorical data as the factor. Taking these into the consideration we can build the predictive analysis model more effective.