Project report on Titanic survivor data

Titanic survivor data

On 15th April 1912, Titanic a British passenger liner was on its maiden voyage from Southampton to New York city. After a collision with an iceberg, it sank in the North Atlantic Ocean. It is considered the deadliest peacetime disaster in modern history as out of 2224 passengers and crew members more than 1500 survived[1]. Analysis in this project is based on titanic_survivor.csv data which contains information about 891 on-board passengers. The questions asked in this study are what factors effected the survival of passengers titanic_survivor data. In current study we address following questions:

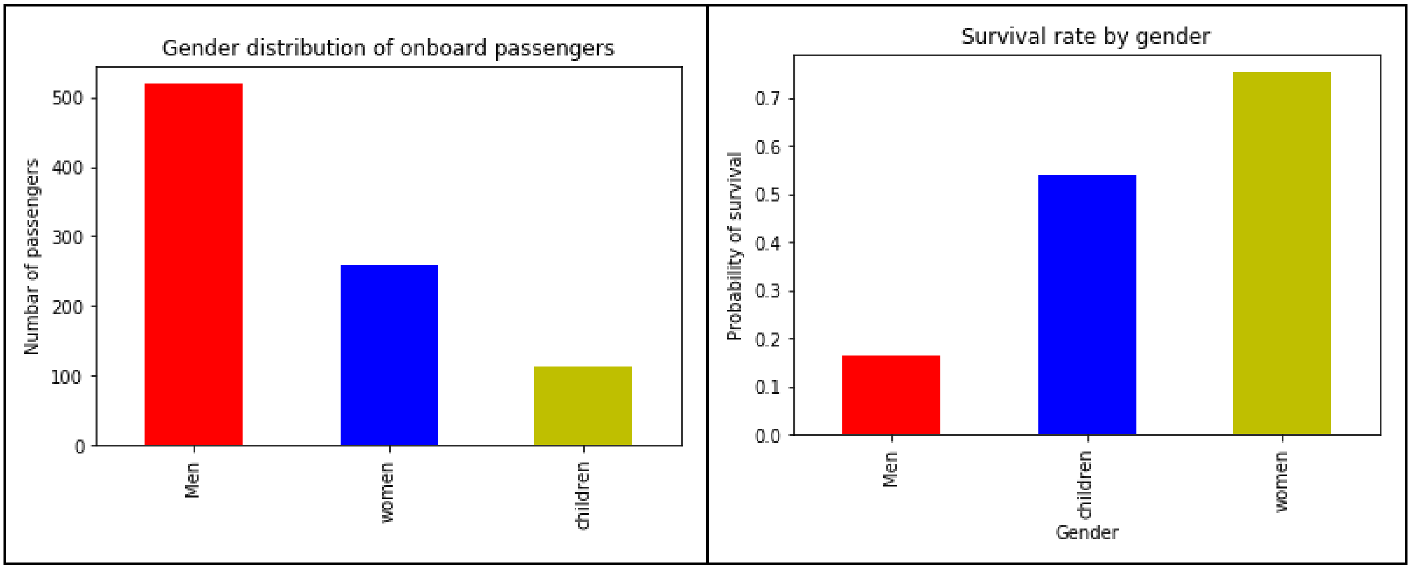
- Which gender and passenger class had better chances of survival?
- Which age group had best chances of survival?
- Passengers from which port of embarkation survived most and least?
- What is the Effect of on-board relatives on survival of passengers?

For the comparison of survival 'Survival probability' has been used which is mean of 'Survived'column.

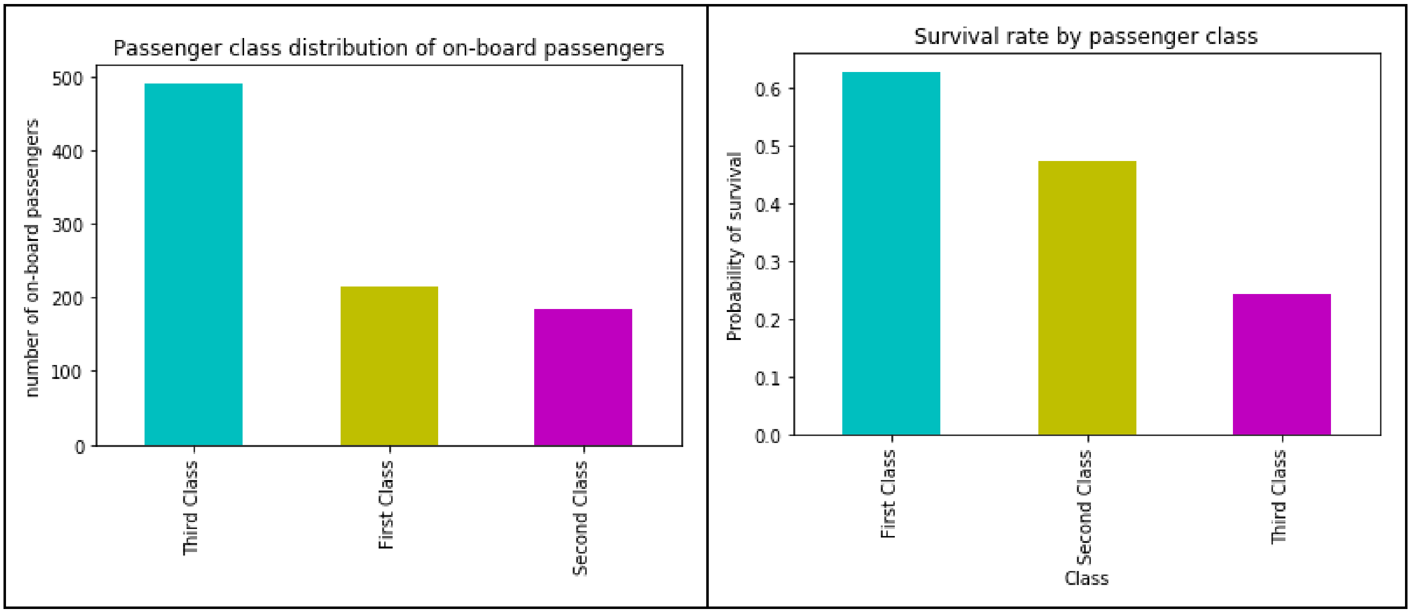
Data wrangling:

- The column 'Sex' of titanic_survivor data contains only male and female. A new column 'Gender' has been added showing three values kids, male and female representing minor, adult male and adult female.
- Filling the missing value in column 'Age' required to be done very carefully due to large numbers(177) of blanks. If it is replaced by mean it will give an anomaly at mean and will affect any exploration done significantly. In order to fill the missing values without disturbing the statistics of 'Age' column positive random numbers has been generated taking in to account mean and std.
- Column 'Cabin has been removed as it have not been utilized in the current study and there were too many missing values.
- Column 'Embarked' had two blank rows with same ticket that means these two passengers were traveling together and embarked from same port. Therefore, blank rows has been filled with mode of 'Embarked' which is 'S'.
- A new column 'Survival has been added showing 'Died' and 'Survived' corresponding to 0 and 1 in 'Survived' column.
- A new column 'Class' has been added with values 'First class', 'Second class' and 'Third class' corresponding to 1,2 and 3 values of 'Pclass' column.
- Column Has_parch has been added showing boolean values True if a passenger has a non-zero value in 'Parch' column and False otherwise.
- Column Has_sibsp has been added showing boolean values True if a passenger has a non-zero value in 'SibSp' column and False otherwise.
- Column Has_relative has been added showing boolean values True if a passenger has non-zero value in columns 'Parch' or 'SibSp' and False otherwise.

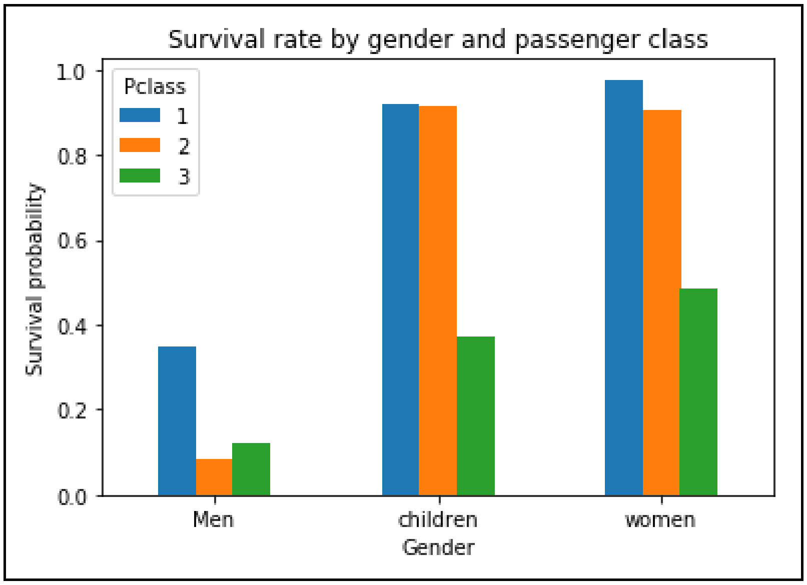
What is the effect of gender and passenger class on survival of passengers?



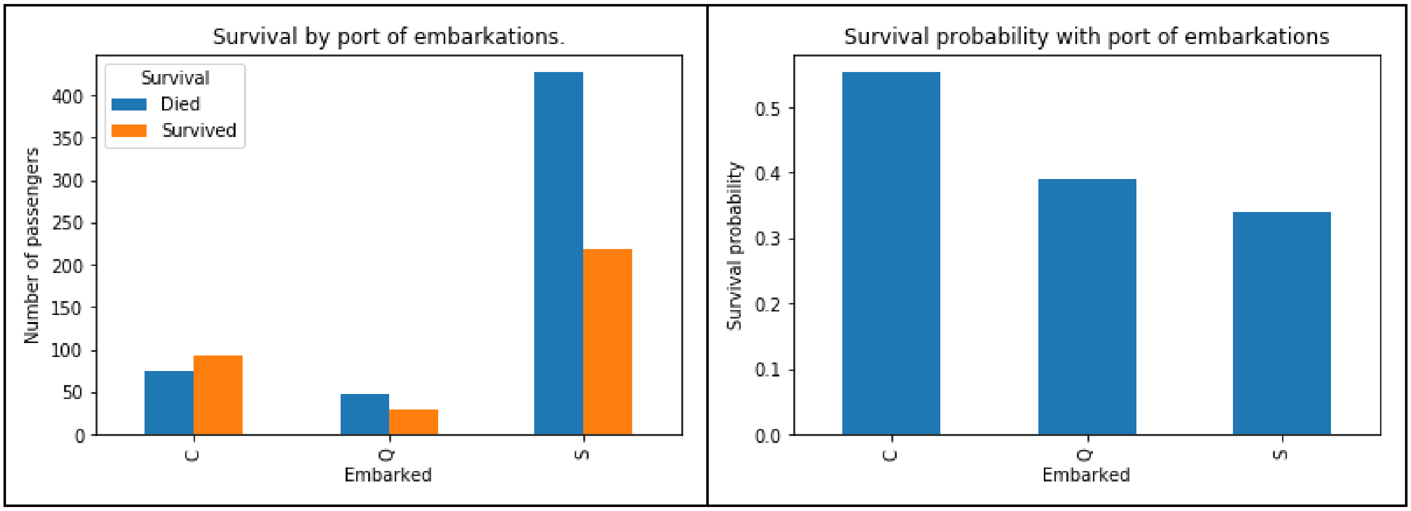
Total on-board passengers were 891 wherein 519 were men, 259 women and 113 children. Out of 259 women, 195 survived, in 113 children, 61 survived while out of 519 men only 86 survived. In the above figure, the right plot shows the survival probability which has been calculated as mean of 'Survived' column. Women have the highest survival probability of 0.75 while men have the lowest survival probability. Survival probability of children is 0.54. To find out if there were any preference according to passenger's class, class wise passenger distribution and survival probability have been given in following bar charts.



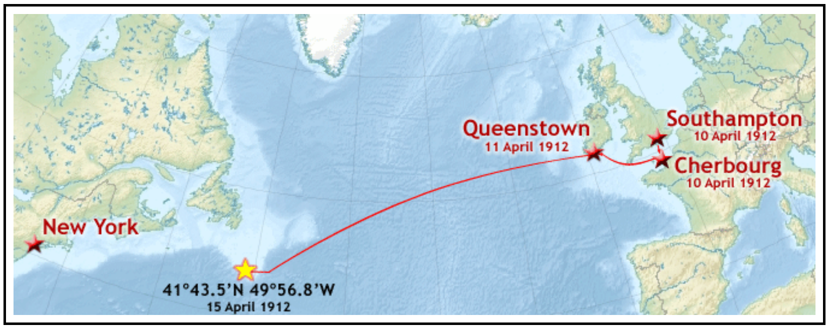
In class wise distribution of total on-board passengers, there were 216 in first class, 184 in second class and 491 in third class. Out of total survived passengers 136 were from first class, 87 were from second class and 119 were from third class thus survival probability of first class, being 0.63, is the highest and that of third class is the lowest being 0.24. Looking at this data we may want to conclude that the passengers from higher socioeconomic class were given preference. But we also know from the gender data that women and kids were given preference so the first class having the highest survival probability may be because there might have been more women and kids in first class. To clarify the actual effect of class and gender on survival probabilities we must see the gender wise passenger distribution in different passenger classes as given in following figure.



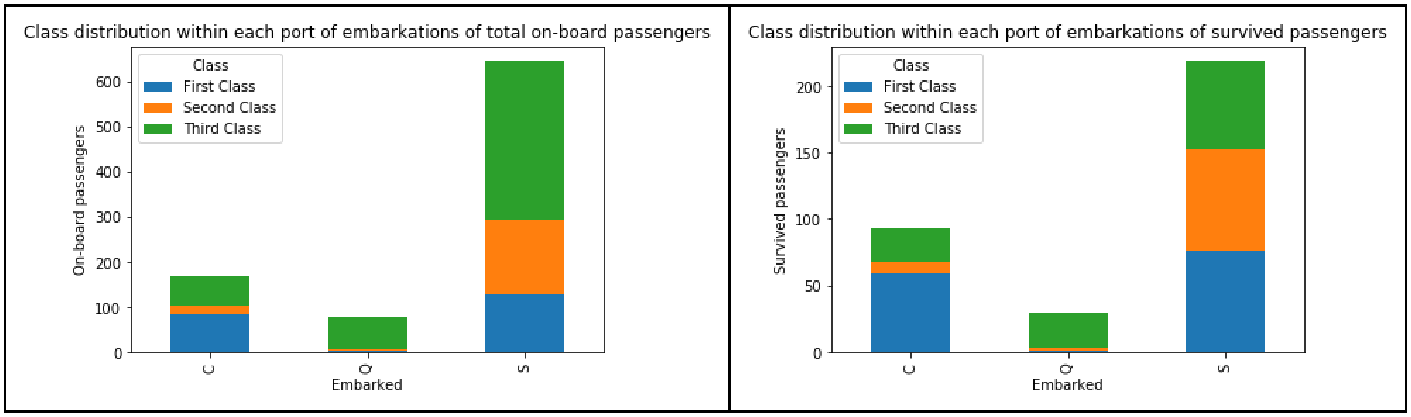
Some interesting observations can be made from this chart- 1. Survival probability of women and children in 1st and second class is almost same and between 0.9 to 1 but for the third class percentage comes down to 0.4 to 0.5. 2. Survival probability of men in first class is higher than the men in second and third class. The most interesting data is that the survival percentage of men in 2nd class is the lowest. The data of gender and class concludes that the preference were given to women and children and higher passenger class. Further, The correlation of age and relatives with this data can be of interest which is covered in latter sections. ## Which port of embarkation had most number of survivors? In this data there are three port of embarkations- C = Cherbourg, Q = Queenstown, S = Southampton.



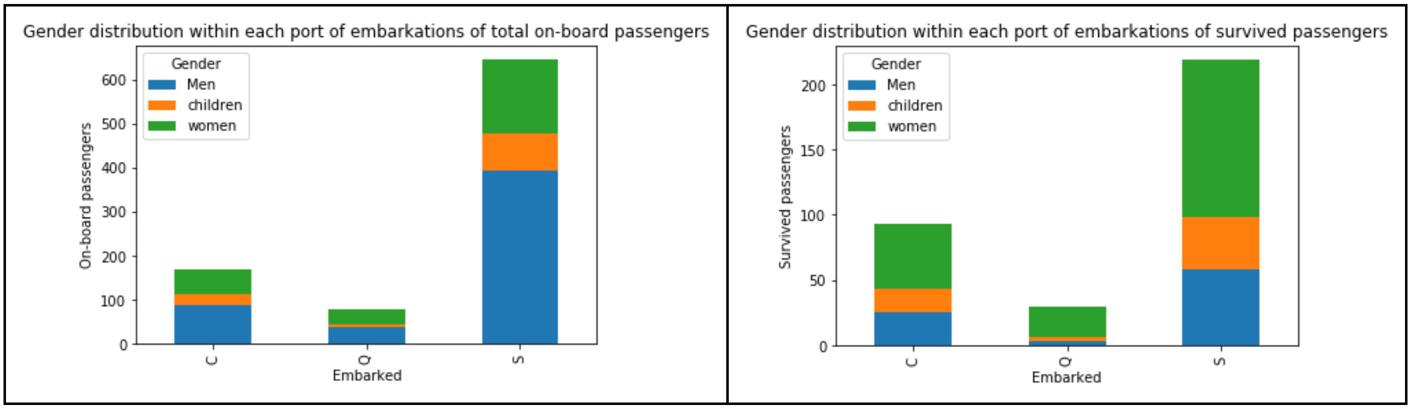
Looking at this chart we see that though largest number of passengers embarked and survived from Southampton, the survival probability (mean of column 'Survived') for Southampton is the least. Whereas, survival probability for passengers boarding from Cherbourg is the highest. If we relate port of embarkation to the passengers' nationality i.e. passengers boarding on Cherbourgh are French nationals, Queenstown are Irish and Southampton are of british nationality respectively. We can say that nearly 43% of French nationals survived, 33 % of English survived and 26% of Irish survived. Following figure shows the voyage map of titanic[1].



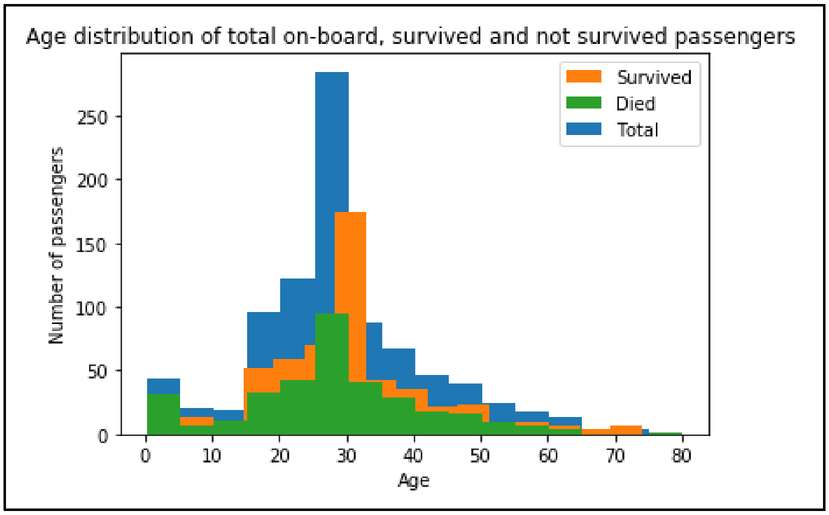
There is no way to relate why any nationals will be given preference of life boats therefore it's relation with other variables with preferences should be comprehended. Thus we look at class and gender distribution of passengers from different port of embarkations. In following charts we can see the passenger class and gender distributions within port of embarkations for total and survived passengers:



It can be seen that nearly half of the French passengers were in 1st class, one third of all the English passengers were in 1st class and while majority of the Irish passengers were in third class.



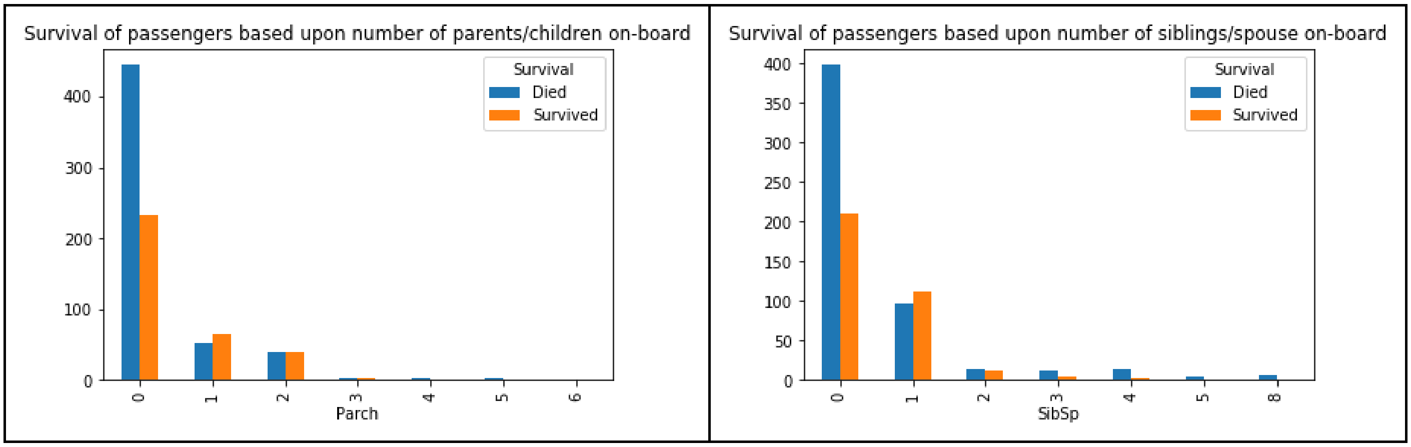
In gender distributions, the difference between port of embarkations is not as drastically different as in class distributions. Still, women+children to men ratio in Cherbourg and Queenstown is nearly 1:1 while in Southampton this ratio is nearly 1:2. Thus, both the gender and class distribution accounts for the significant difference between French nationals and English while poor survival probability of Irish nationals is mainly due to the fact that majority were in third class. This data quite explains why survival probability for Irish passengers were the lowest and that for French passengers were the highest. ## Which age group had best chances of survival?: The following histograms show age distribution of total on-board passengers and survived passengers.



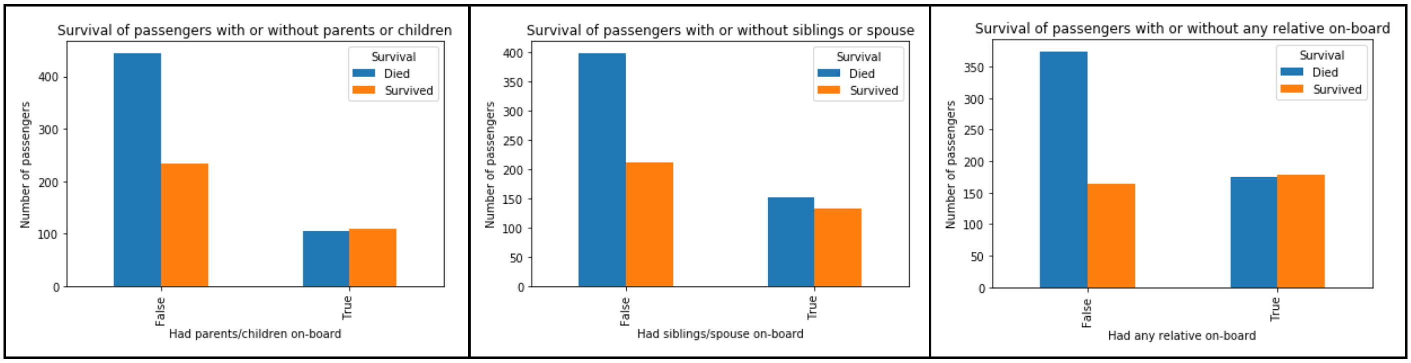
Above histogram shows age distribution of 714 on-board passengers whose ages are entered in 'Age' column in the given csv file. The mean of on-board passengers is 29.69, while the same for survived and not-survived passengers are 28.55 and 30.42 respectively. The lower mean of survived passengers may be due to survival of most of the children. Both the youngest and oldest person aged 0.42 years and 80 years survived. From the modes of the three histograms, it can be interpreted that most of all the on-board passengers were in 25-30 years age group while the most of survived passengers were in 30-35 years age group and of the not-survived passengers list most were in 25 - 30 years age group. The other noticeable data from not-survived passenger is that most of the kids who couldn't survive were under 5 years.

What is the Effect of on-board relatives on survival of passengers?:

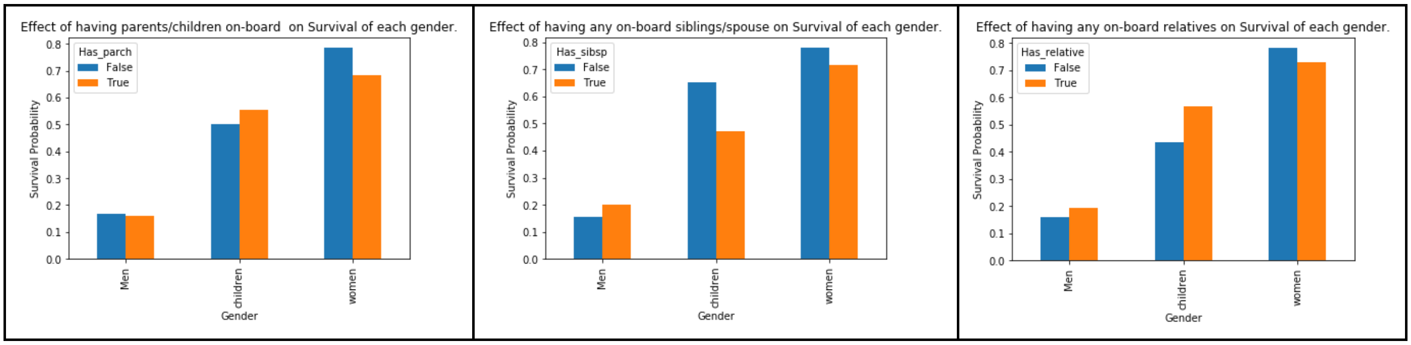
Following figure shows how many passengers had their parent/child or sibling/ spouse on-board.



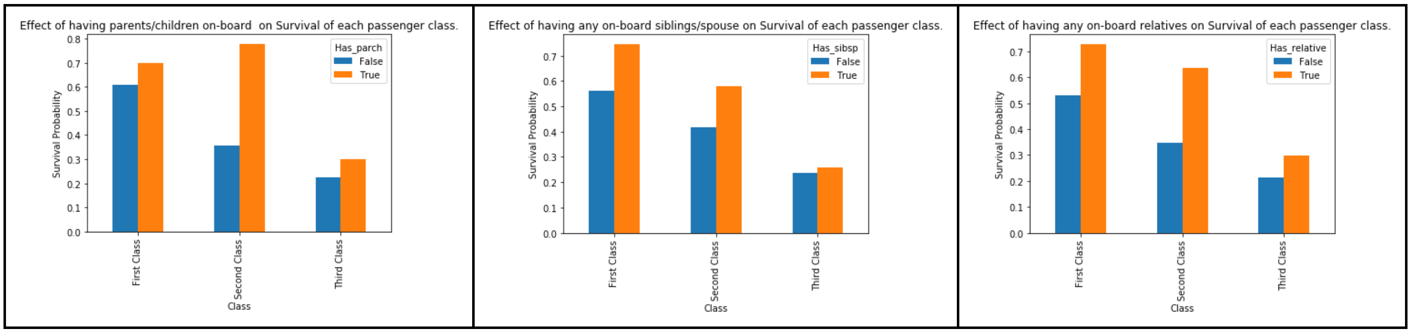
In the first plot passengers without any parent/child on-board have poor probability of survival than passengers with parent/child. However, The chances get poorer as number of parent/child increase. The following chart shows effect of on-board relative on survival of a passenger. the first plot shows effect of parent/child, second plot shows effect of sibling/spouse and third pot shows effect of having any relative('Parch' or 'SibSp') on-board.



These plots show that effect of having a parent/child, sibling/spouse or any of these relative improves the chances of survival of passengers. The effect is most prominent in the passengers with parent/child present. It can be interpreted as the families with just one child had better chances of survival than the families with multiple children as the latter ones had more people to take to safety. Sibling/spouse plot shows similar results, except that the survival probability reduces even faster as numbers in 'SibSP' increase. The reason being 2 siblings are equivalent to 3 in children. Now, effect of having a parent on-board for a child may be positive for his/her survival but that might not be the case for a mother or a father. Therefore, effect of having relative on-board on different gender has been shown in the bar chart below.



From first plot, we can see that children who had their parents traveling with them had better chances of survival while the survival percentage in adult females is less than if they have their parent/child traveling with them. One of the reasons for this data might be that the mothers were more worried about the survival of their children than about themselves. This data does not show any significant difference in men survival data. On the other hand the survival data of sibling/spouse shows that men who had their siblings or spouses with them have better survival probability but kids and women with siblings and spouses on-board has poorer probability of survival. It might be because if a parent has multiple children, as his attention gets divided, the probability of survival of children decreases. Worry about the safety of their husbands might be the reason of decreased survival probability in women. Now combining these two plots, plot three shows effect of having any relative on-board on survival of each gender. Men and kids had better chances of survival if they had any relatives on-board while women have poorer chances of survival than if they had any relatives present. To separate the effect of class from the effect of having relative on-board, the survival probability of passengers with or without relatives has been shown in following plots.



All of the three plots give the same result i.e. in all three classes having any kind of relatives on-board improves the chances of survival. Having parent/children in second class drastically improves the survival probaility.

Conclusion:

The current study and the conclusions below are based on this sample data and can be very different from the full titanic_survivor data.

- The women and kids percentage of survival are way more than men.
- The survival percentage of 1st class passengers is the highest and that of 2nd class is the lowest which is due to very low survival percentage of men in 2nd class.
- Survival percentage of men in second class is the lowest. One of the reasons may be that they offered their seats to women and kids of third class.
- Most embarkations from Queenstown was in 3rd class hence the survival percentage for Queenstown is the lowest among the three port of embarkations.
- Most embarkations from Cherbourg were in 1st and 2nd class that's why the survival percentage from Cherbourg is the highest.
- Most of the children who died were under 5 years of age.
- Survival percentage for children traveling with family is higher than the ones who were traveling without the family, however, impact of parent on-board is positive while impact of siblings is negative. Survival percentage of women traveling with relative is lower than the women who were traveling without the family while that of men is better with relatives on-board. In other words, having relative on-board lowers the chances of survival in adult with higher preference and increases the same in adult with lower preference.

Source:

In [13]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
titanic_survivor=pd.read_csv('/Users/admin/Desktop/DAND/Projects/titanic
_survivor/titanic-data.csv')

#titanic_survivor nan or missing values have been relaced with mode of e
mbarked column.
titanic_survivor['Embarked'].mode()

titanic_survivor['Embarked']=titanic_survivor['Embarked'].fillna('S')

# Create Survival Label Column
titanic_survivor['Survival'] = titanic_survivor.Survived.map({0 : 'Died'
, 1 : 'Survived'})

# Create Pclass Label Column
titanic_survivor['Class'] = titanic_survivor.Pclass.map({1 : 'First Clas
s', 2 : 'Second Class', 3 : 'Third Class'})

# Create Embarked Labels Column
titanic_survivor['Ports'] = titanic_survivor.Embarked.map({'C' : 'Cherbo
urg', 'Q' : 'Queenstown', 'S' : 'Southampton'})
titanic_survivor.head()

# data wrangling for 'Age'. Each blank has been replaced by positive ran
dom number which does not
#change the mean or std of 'Age' column

import random
def rand_age(mean, std):
     age = -1
     while age < 0: # makes sure the number is not negative, which is po
ssible in a normal distribution
          age = random.normalvariate(mean, std)
     return age
titanic_survivor.Age.apply(lambda x:rand_age(titanic_survivor.Age.mean
(), titanic_survivor.Age.std()) if x!=x else x).head(10)


#define a function to devide the Sex column in to male female and kids
def calculate_kid(row):
    if row['Age'] < 18:
        return 'children'
    elif row['Sex']=='male':
        return 'Men'
    elif row['Sex']=='female':
        return 'women'

# make new column ['Gender'] based on above function.
titanic_survivor['Gender']=titanic_survivor.apply(lambda row: calculate_
kid (row),axis=1)

# get the gender counts and visualization of the same
gender_count=titanic_survivor['Gender'].value_counts()
print(gender_count)
```

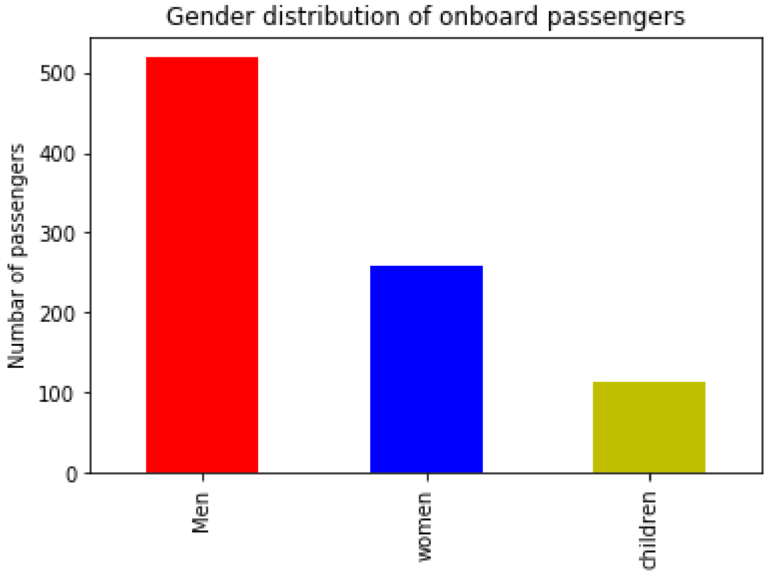
```
gender_count.plot(kind='bar',color=('r','b','y'))

plt.title('Gender distribution of onboard passengers')

plt.ylabel('Numbar of passengers')
# survival rate by gender

plt.show()

# survival rate by gender

titanic_survivor.groupby(['Gender'])['Survived'].mean().plot(kind='bar',color=('r','b','y'))
plt.ylabel('Probability of survival')
plt.title('Survival rate by gender')
plt.show()

# get the passenger class count and visualization
titanic_survivor['Class'].value_counts().plot(kind='bar',color=('c','y','m'))
plt.title('Passenger class distribution of on-board passengers')
plt.ylabel('number of on-board passengers')
plt.show()

# survival rate by passenger class
titanic_survivor.groupby(['Class'])['Survived'].mean().plot(kind='bar',color=('c','y','m'))
plt.title('Survival rate by passenger class')
plt.ylabel('Probability of survival')
plt.show()

# calculate survival probability with gender in different class
titanic_survivor.pivot_table(index='Gender', columns='Pclass', values='Survived', aggfunc='mean').plot.bar(rot=0)
plt.title('Survival rate by gender and passenger class')
plt.ylabel('Survival probability')
plt.show()

##########   portof embarkations   ###############################

# calculate total onboard passengers with port of embarkations
#following lines calculate survival percentage for different port of embarkation.
#survival percent=survived for port/total embarkations on port
emb_total=(titanic_survivor.groupby(['Embarked'])['Name'].count())
emb_survived=titanic_survivor.groupby(['Embarked'])['Survived'].sum()

#visualize survival by port of embarkation
titanic_survivor.pivot_table(index='Embarked', columns='Survival', values='Survived', aggfunc='count').plot(kind='bar')
plt.title( 'Survival by port of embarkations.')
plt.ylabel('Number of passengers')
plt.show()

# Survival probability by port of embarkation
titanic_survivor.groupby(['Embarked'])['Survived'].mean().plot(kind='bar')
```

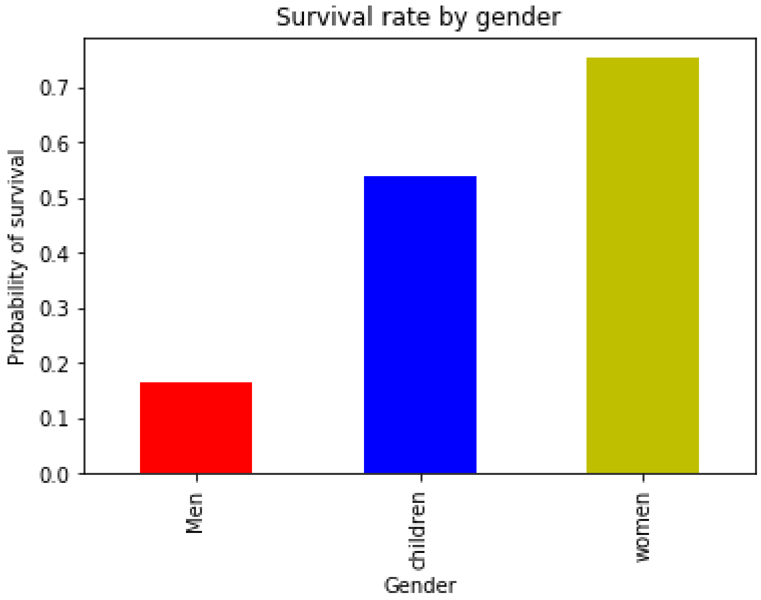
```
plt.ylabel('Survival probability')
plt.title('Survival probability with port of embarkations')
plt.show()

#Class distribution within each port of embarkations of total on-board p
assengers'
titanic_survivor.pivot_table(index='Embarked', columns='Class', values=
'Survived',aggfunc='sum').plot(kind='bar',stacked=True)
plt.ylabel('Survived passengers')
plt.title('Class distribution within each port of embarkations of surviv
ed passengers')
plt.show()
titanic_survivor.pivot_table(index='Embarked', columns='Class', values=
'Survived',aggfunc='count').plot(kind='bar',stacked=True)
plt.ylabel('On-board passengers')
plt.title('Class distribution within each port of embarkations of total
 on-board passengers')
plt.show()

#Gender distribution within each port of embarkations of total on-board
 passengers
titanic_survivor.pivot_table(index='Embarked', columns='Gender', values=
'Survived',aggfunc='sum').plot(kind='bar',stacked=True)
plt.ylabel('Survived passengers')
plt.title('Gender distribution within each port of embarkations of survi
ved passengers')
plt.show()
titanic_survivor.pivot_table(index='Embarked', columns='Gender', values=
'Survived',aggfunc='count').plot(kind='bar',stacked=True)
plt.ylabel('On-board passengers')
plt.title('Gender distribution within each port of embarkations of total
 on-board passengers')

plt.show()

######  having relatives  #####################

# to see effect of having parent/child calculate total passengers with n
umber of parents children on-board

titanic_survivor.groupby(['Parch']).apply(len).plot(kind='bar')
plt.title('On-board passengers with number of parents/children ')
plt.ylabel('Number of on-board passengers')
plt.show()

#Calulate Survival of passengers based upon number of parents/children o
n-board
titanic_survivor.pivot_table(index='Parch', columns='Survival', values=
'Survived', aggfunc='count').plot(kind='bar')
plt.title('Survival of passengers based upon number of parents/children
 on-board')
plt.show()

# comparision of survival probability with or without any parents/childr
en
# we make a new column Has_parch with values true or false depending upo
n value in 'Parch' is zero
```

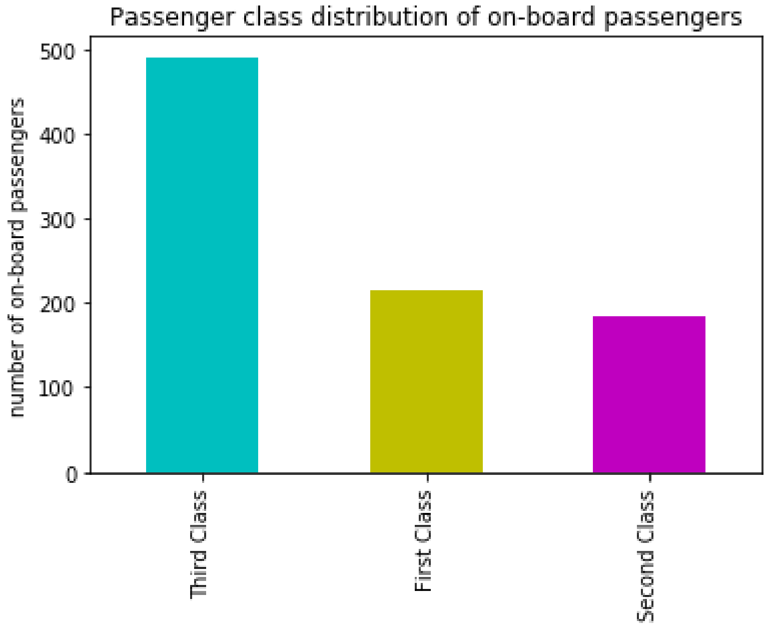
```
#or non-zero. Define a function  has_parch for checking a passenger has any par/ch or not.
def has_parch(row):
     return row['Parch']!=0
titanic_survivor['Has_parch']=titanic_survivor.apply(lambda row: has_parch(row), axis=1)
titanic_survivor.pivot_table(index='Has_parch',columns='Survival',values='Survived',aggfunc='count').plot(kind='bar')
plt.ylabel('Number of passengers')
plt.xlabel('Had parents/children on-board')
plt.title('Survival of passengers with or without parents or children')
plt.show()

#comparision of survival with or without siblings/spouse
# comparision of survival probability with or without any parents/children

def has_sibsp(row):
     return row['SibSp']!=0
titanic_survivor['Has_sibsp']=titanic_survivor.apply(lambda row: has_sibsp(row), axis=1)
titanic_survivor.pivot_table(index='Has_sibsp',columns='Survival',values='Survived',aggfunc='count').plot(kind='bar')
plt.ylabel('Number of passengers')
plt.xlabel('Had siblings/spouse on-board')
plt.title('Survival of passengers with or without siblings or spouse')
plt.show()

# survival of passengers who had any relatives onboard
# make a column Has_relative with values true or false whether or not they had any relatives on-board.
#define a function which gives boolean value of (Parch or SibSp)
def has_relative(row):
    return row['Parch']!=0 or row['SibSp']!=0
titanic_survivor['Has_relative']=titanic_survivor.apply(lambda row: has_relative(row), axis=1)
titanic_survivor.pivot_table(index='Has_relative',columns='Survival',values='Survived',aggfunc='count').plot(kind='bar')
plt.ylabel('Number of passengers')
plt.xlabel('Had any relative on-board')
plt.title('Survival of passengers with or without any relative on-board')
plt.show()

#Effect of having any on-board relatives on Survival of each gender
titanic_survivor.pivot_table(index='Gender', columns= 'Has_parch', values='Survived',aggfunc='mean').plot(kind='bar')
plt.title('Effect of having parents/children on-board  on Survival of each gender.')
plt.ylabel('Survival Probability')
titanic_survivor.pivot_table(index='Gender', columns= 'Has_sibsp', values='Survived',aggfunc='mean').plot(kind='bar')
plt.title('Effect of having any on-board siblings/spouse on Survival of each gender.')
plt.ylabel('Survival Probability')
titanic_survivor.pivot_table(index='Gender', columns= 'Has_relative', values='Survived',aggfunc='mean').plot(kind='bar')
```

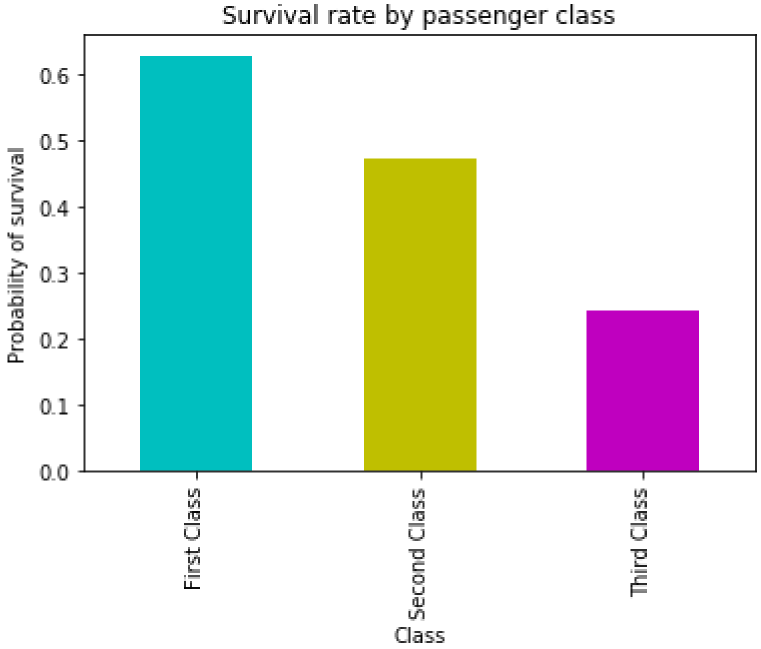
```
plt.title('Effect of having any on-board relatives on Survival of each g
ender.')
plt.ylabel('Survival Probability')
plt.show()
```

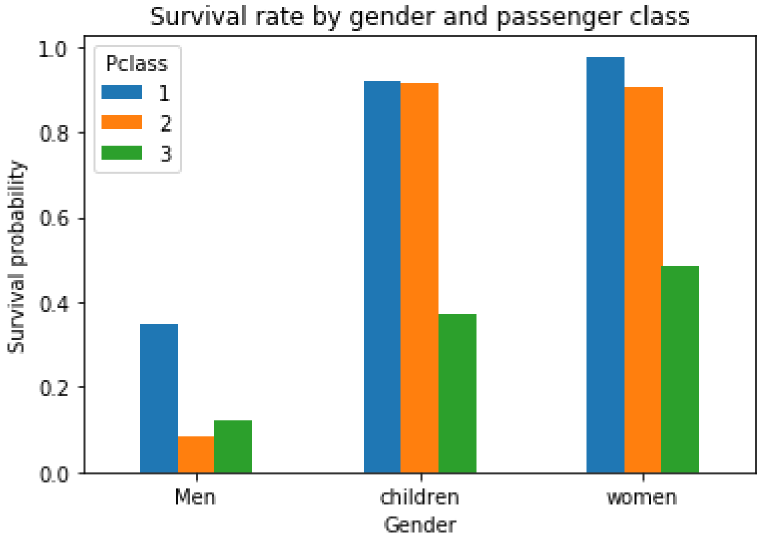
```
Men         519
women       259
children    113
Name: Gender, dtype: int64
```

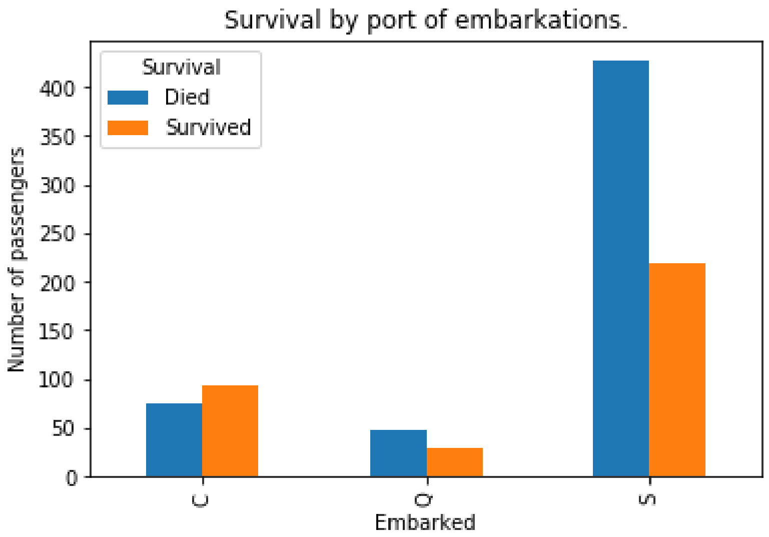


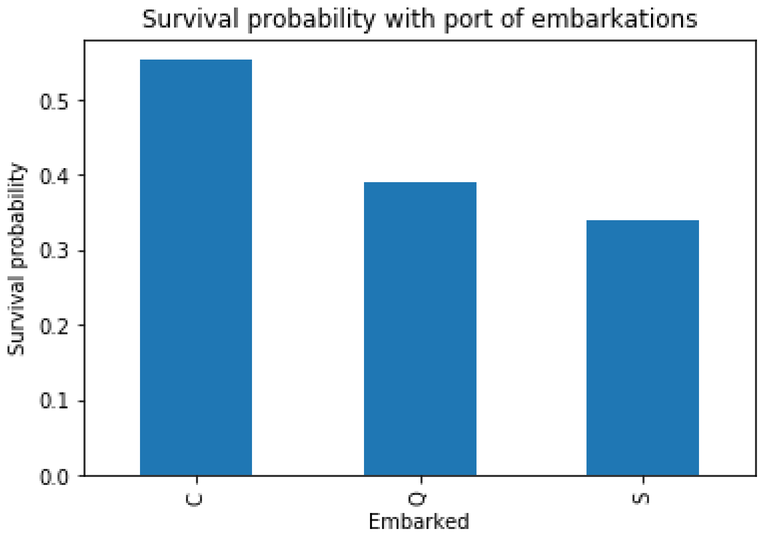


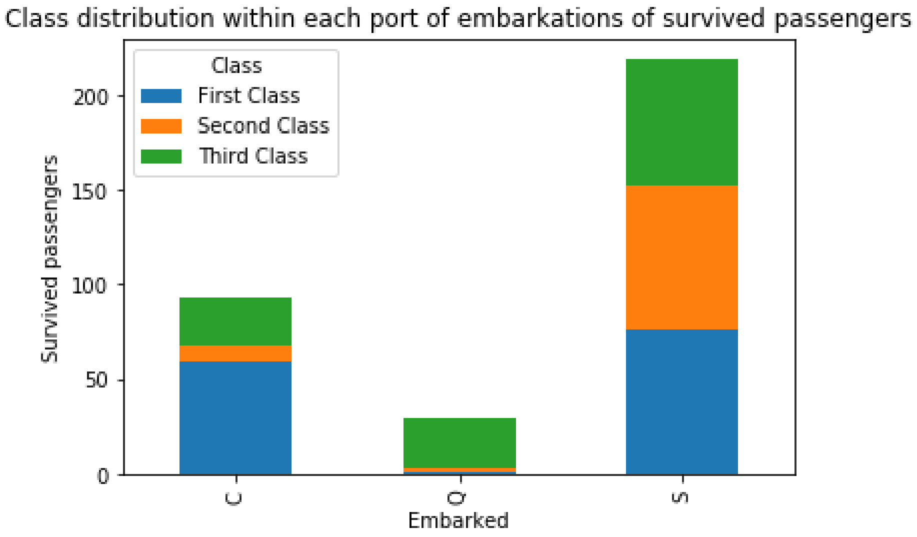

Passenger class distribution of on-board passengers
number of on-board passengers
0
100
200
300
400
500
Third Class
First Class
Second Class

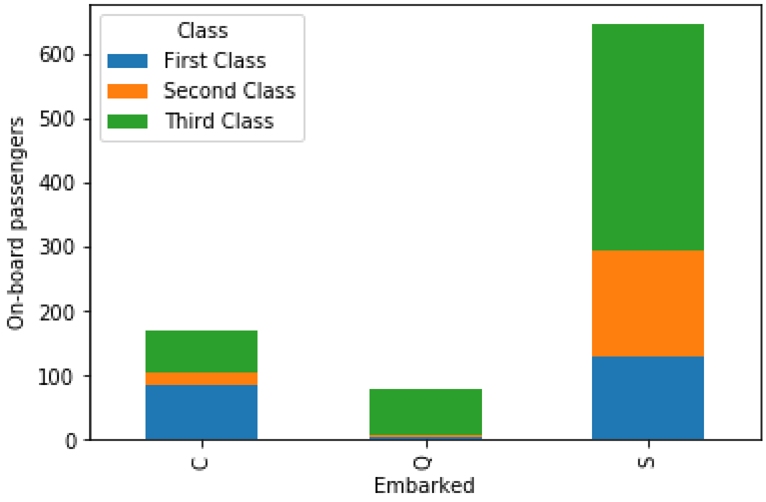

Survival rate by passenger class
Probability of survival
0.0
0.1
0.2
0.3
0.4
0.5
0.6
First Class
Second Class
Third Class
Class


Survival rate by gender and passenger class
Survival probability
0.0
0.2
0.4
0.6
0.8
1.0
Pclass
1
2
3
Men
children
women
Gender

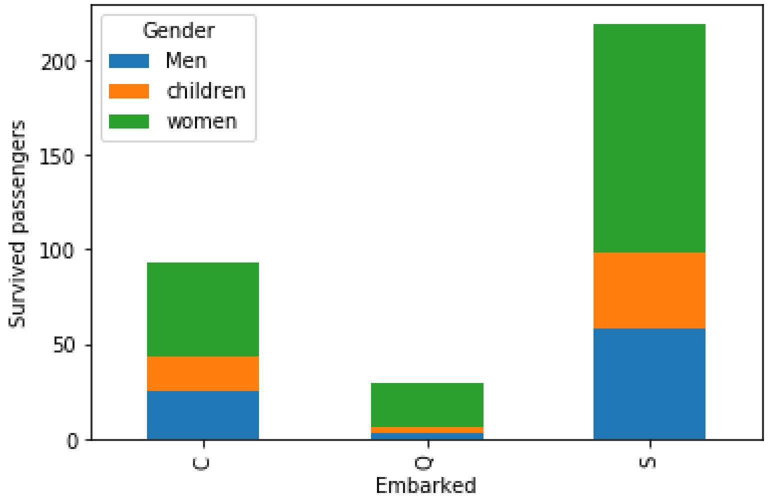

Survival by port of embarkations.


Survival probability with port of embarkations

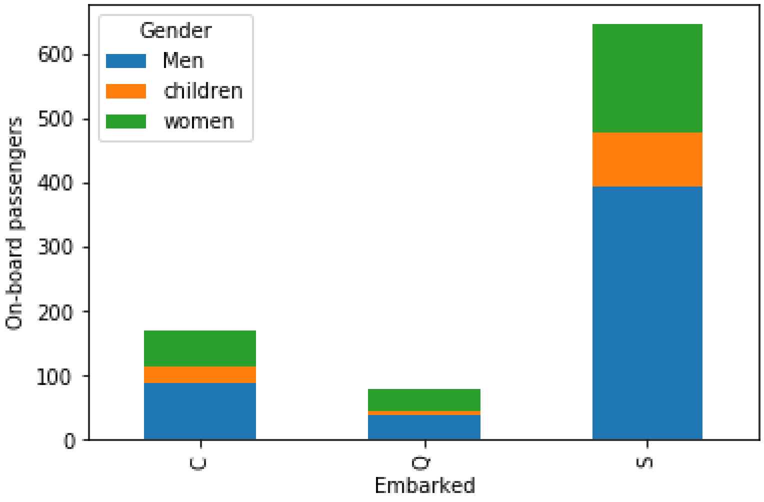

Class distribution within each port of embarkations of survived passengers



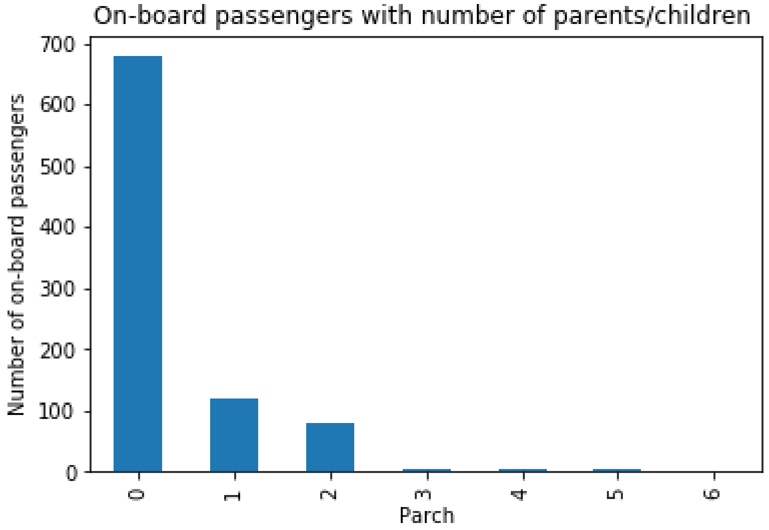
Class distribution within each port of embarkations of total on-board passengers

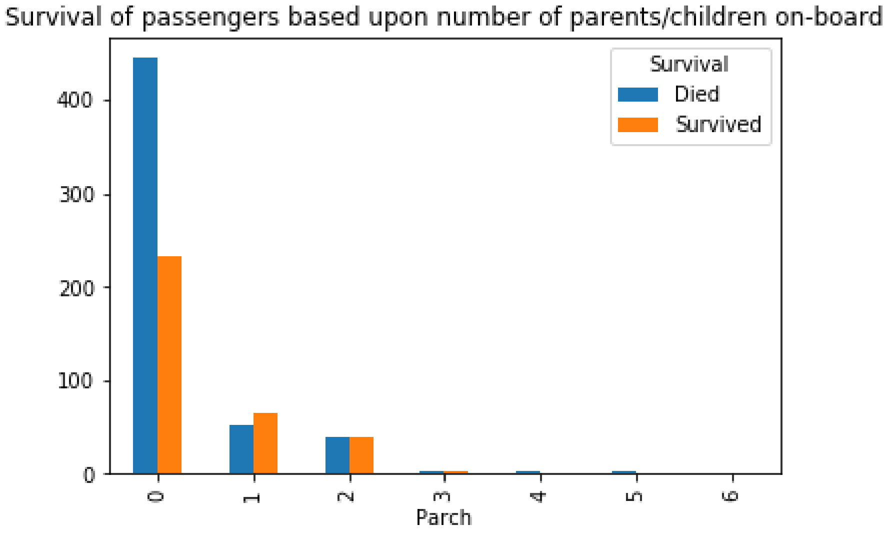


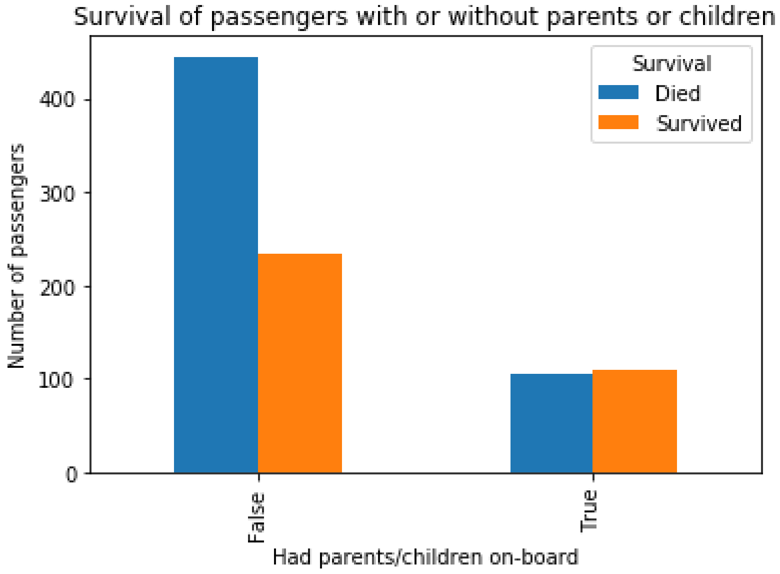
Gender distribution within each port of embarkations of survived passengers

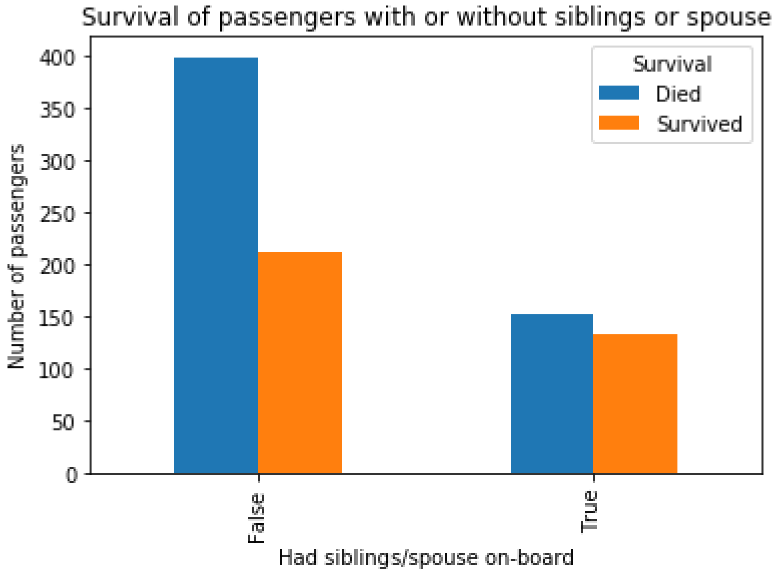


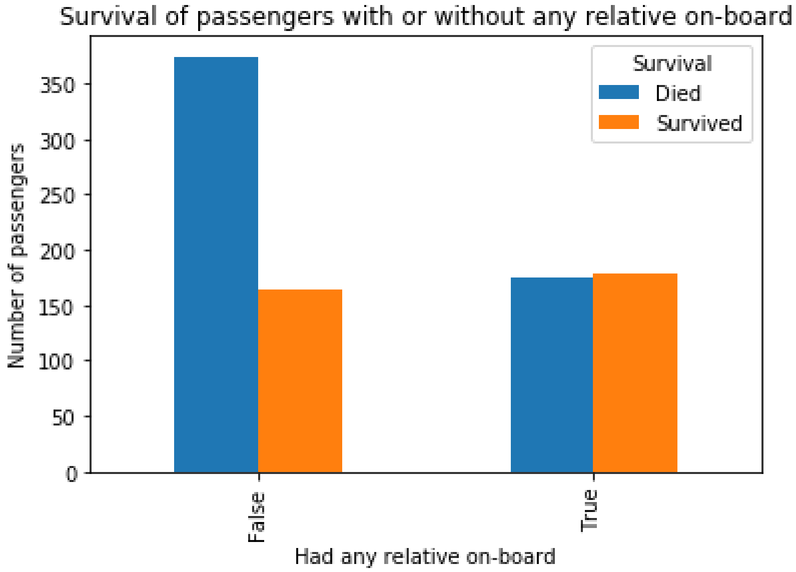
Gender distribution within each port of embarkations of total on-board passengers

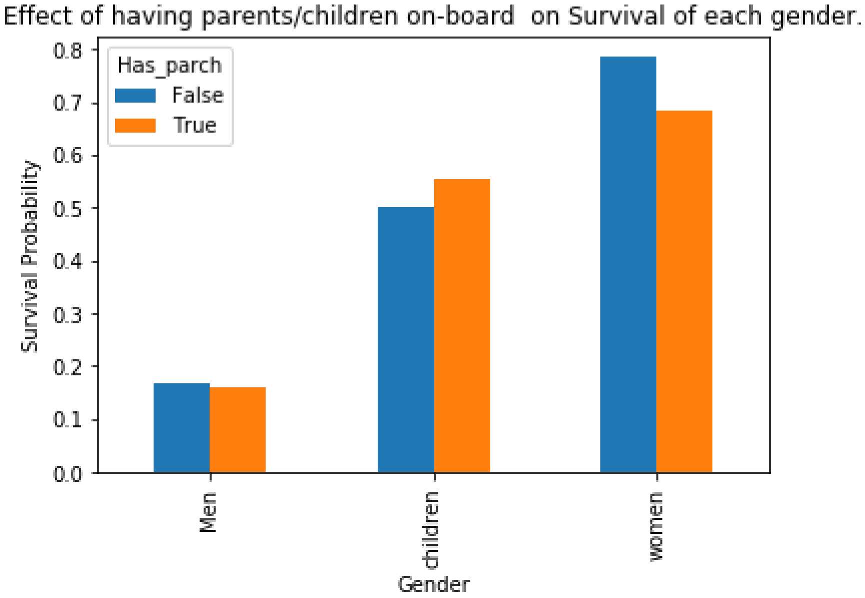

On-board passengers with number of parents/children
Number of on-board passengers
700
600
500
400
300
200
100
0
0
1
2
3
4
5
6
Parch

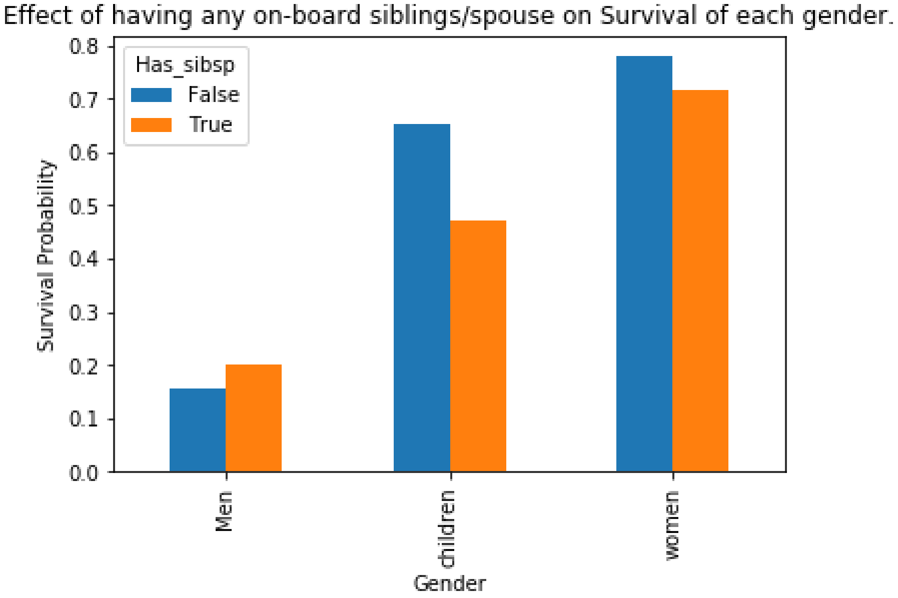

Survival of passengers based upon number of parents/children on-board
Survival
Died
Survived
400
300
200
100
0
0
1
2
3
4
5
6
Parch

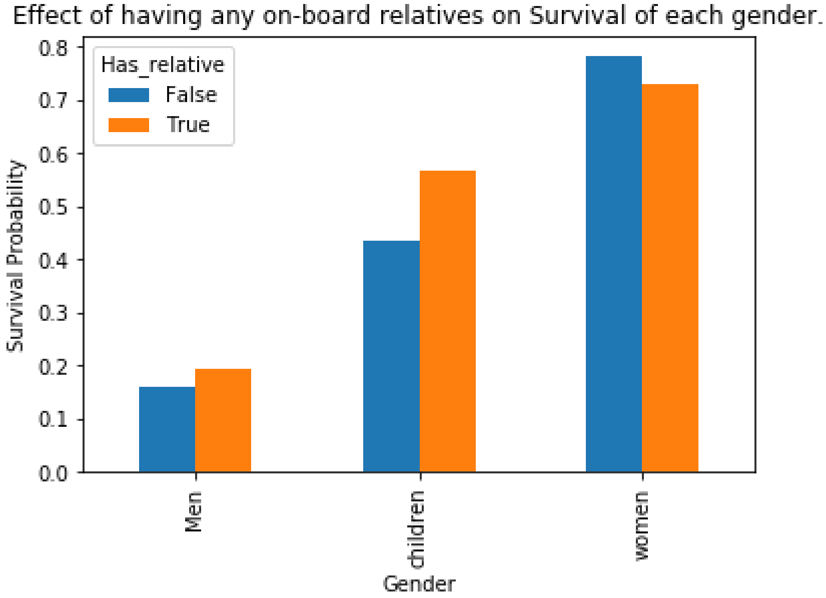

Survival of passengers with or without parents or children
Survival
Died
Survived
Number of passengers
400
300
200
100
0
False
True
Had parents/children on-board


Survival of passengers with or without siblings or spouse
Number of passengers
0
50
100
150
200
250
300
350
400
Survival
Died
Survived
False
True
Had siblings/spouse on-board


Survival of passengers with or without any relative on-board
Number of passengers
0
50
100
150
200
250
300
350
Survival
Died
Survived
False
True
Had any relative on-board


Effect of having parents/children on-board on Survival of each gender.
Survival Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Has_parch
False
True
Men
children
women
Gender


Effect of having any on-board siblings/spouse on Survival of each gender.
Has_sibsp
False
True
Survival Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Men
children
women
Gender


Effect of having any on-board relatives on Survival of each gender.
Has_relative
False
True
Survival Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Men
children
women
Gender

```
In [ ]:
```