When I watched the movie "Titanic" twenty years ago, I had wondered whether most survivors were women and children. Today, thanks to this project, I can figure it out. Also, I'm interested in finding out whether the richer with a first class ticket had a higher chance of survival compared to the other 2 classes. Finally, I wanted to analyze how ticket price index differs in 3 different cities. I will focus on the data I retrieved from https://d17h27t6h515a5.cloudfront.net/topher/2016/September/57e9a84c_titanic-data/titanic-data.csv (data description reference: https://www.kaggle.com/c/titanic/data), which contains demographics and passenger information from 891 of the 2224 passengers and crew on board the Titanic.

So, here are my questions:

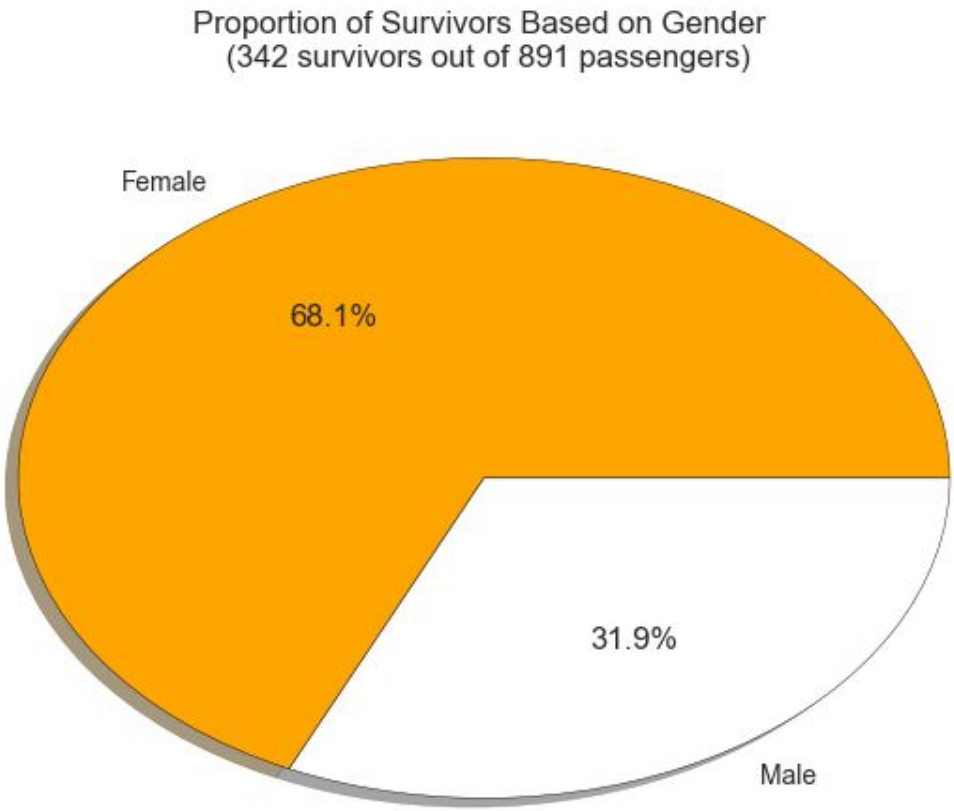
1. What proportion of total survivors are women?
2. Is the survival rate of children (under 12) higher than the others?
3. Did the passengers of first class have a higher chance of survival compared to the other 2 classes and the crew?
4. How did the price index differ in 3 different cities? In other words, how many tickets of class 3 could be bought if we paid the price of class 1 in 3 different cities?

To answer the questions listed above, I first looked into the data frame of the csv file with pandas module in Python:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.6+ KB
```

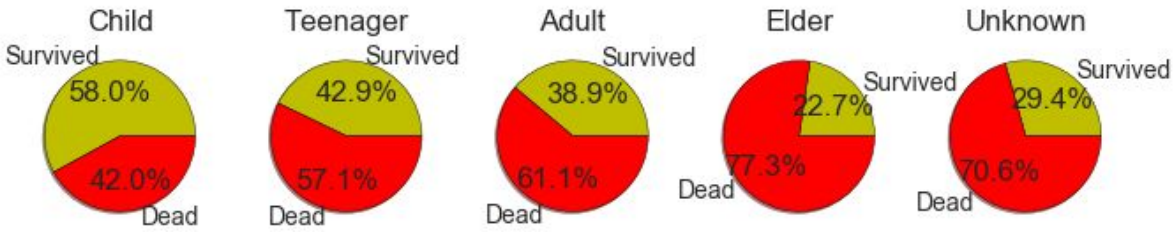
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

For Q1, I separated the samples into two gender groups and summed up the survivors, then output the pie chart illustrating proportion of survivors based on gender as below:



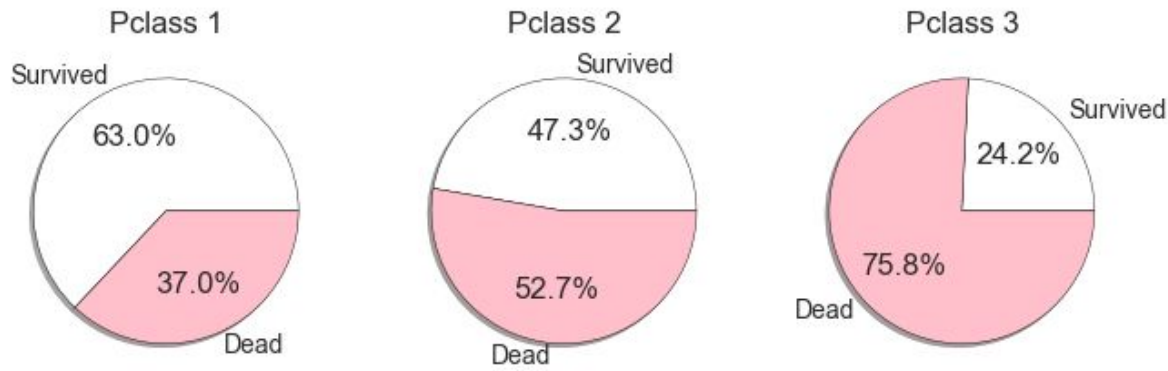
As you can see from the pie chart, female survivors made up of 68.1% from the total. This result is below my expectation, because when we say "most", we usually think of it as being more than 80%. In this case, the percentage was less than 70%. However, we must also keep in mind that we only have information from 891 of the 2224 passengers and crew on board the Titanic, the result might not represent the trend of the whole population properly.

After knowing the proportion of the survivors based on gender, I wanted to explore whether children aboard the Titanic had a higher chance of survival. To group the samples by age, I made a list and checked every sample to see which group each survivor should be included into before adding them into the data frame. Afterwards, I counted the total passengers of each group and calculated their survival rates. Finally, I created the following five pie charts to compare the survival rates of each group:



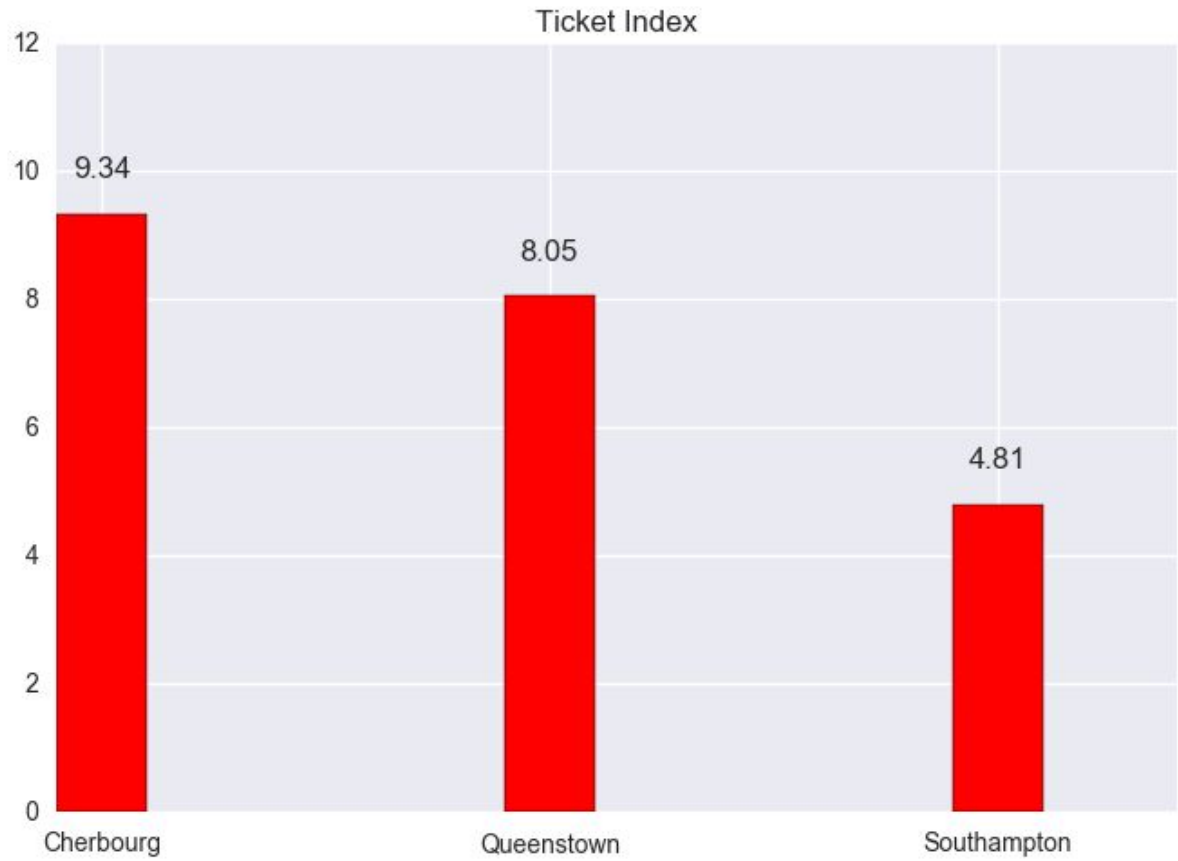
From the pie charts above calculated from the survival rates out of the 891 passengers, not surprisingly children aboard the Titanic had the highest chance of survival and the eldery had the lowest. This information and the first chart on proportion based on gender give support to the quote in the movie "Women and children first" and make the movie more consistent with the fact. Although the statistical results seem reasonable, this may not be the complete picture given there are 177 passengers without age information.

Now I know children had a higher chance of survival than the other age groups and female survivors were more likely to survive than male, but what about the ticket class? Do the passengers in first class have a higher chance of survival than those in second class and third class? To figure this out, I categorized the samples by class, first by separating the passengers into each category and then calculating their survival rates. The pie charts are shown below:



Compared to the survival rate of passengers in third class, those in first class had a much higher chance of surviving the tragedy, which was a sad but realistic fact: If you pay a higher price, you would receive better service, or in other words, a higher chance of survival.

The final part which also happened to be the most enjoyable was to compute the ticket index; that is, how many third class tickets can be bought for the price of one first class ticket in each of the three different cities. I divided the mean of the first class ticket price by the mean of the third class. The outcome is presented with a bar chart:



From the bar chart, we can see that Cherbourg had the highest index value (each first class ticket could be exchanged for 9.34 third class tickets), which is much higher than the 4.81 in Southampton.

This statistic implies that the wealth disparity in Cherbourg was much greater than Southampton, but can we say it for sure? When making this chart, I also extracted the standard deviations of the fare of tickets in first and third class, and noticed the standard deviations were quite large. To explore more and make sure that I derived a firm conclusion, I wrote some more code lines to plot the distribution by port and passenger ticket class. Firstly, I created two lists to store extracted arrays of ticket price and port information. Then I applied a loop to put every necessary data into dictionaries with a plotting module named Plotly. Finally, I represented the distribution of data points with a boxplot shown below (please refer to the ipython file to see more details):

(Red: first class, Green: second class, Orange: third class)

As you can see, there are many outliers in all cities, the maximum in Cherbourg is even up to $512 for one first class ticket while the median is about $78. Besides the outliers, the samples of first and second class in Queenstown are too few to build a strong and reliable correlation. Finally, the sample size should be 891 passengers, but there were 2 passengers without port data and not grouped in any class, which may affect our conclusion as well.

After going through all the investigating procedures based on 342 survivors out of the 891 passengers, here are my findings:

1. Female survivors made up 68.1% of the total;
2. Compared to the other age ranges, children aboard the Titanic had a higher chance (58.0%) of survival;
3. Compared to the other two ticket classes, the passengers with first class ticket had a higher chance (63.0%) of survival; and
4. Cherbourg had the highest index value (each first class ticket could be exchanged for 9.34 third class tickets), which is much higher than the 4.81 in Southampton.

However, we should keep in mind that:

1. Not all 2224 people aboard the Titanic are included; the sample consists of only 891 passengers and zero crew, which may not properly draw a whole picture about the people on board;
2. There are many missing values; for example, the age and port information are unrecorded for 177 and 2 passengers, respectively, which may make our findings unreliable; and
3. There are many outliers in some categories and undersized samples in some others when exploring the data of ticket fare, which may lead us to incorrect conclusions as well.