Report of Project 2

Question:

What kinds of passenger have most probability to survive from Titanic?

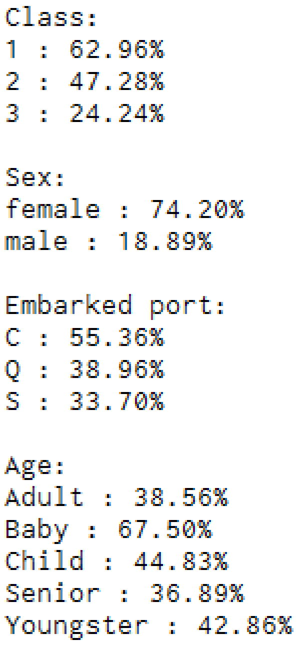
Data cleaning:

In the column of “Age”, some data is missing. When I try to analyze the relation between “Age ” and “survival probability”, I have to drop all the NAN.

After that, I classify each passenger into categories.

Data exploration:

1 Dimension:



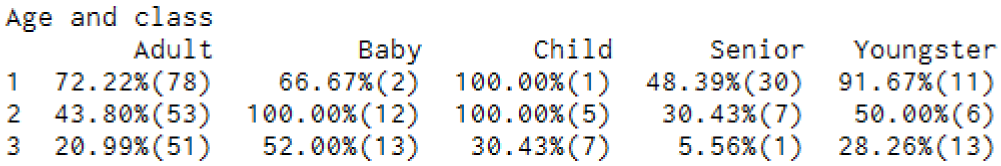
```
Class:
1 : 62.96%
2 : 47.28%
3 : 24.24%

Sex:
female : 74.20%
male : 18.89%

Embarked port:
C : 55.36%
Q : 38.96%
S : 33.70%

Age:
Adult : 38.56%
Baby : 67.50%
Child : 44.83%
Senior : 36.89%
Youngster : 42.86%
```

2 Dimension:

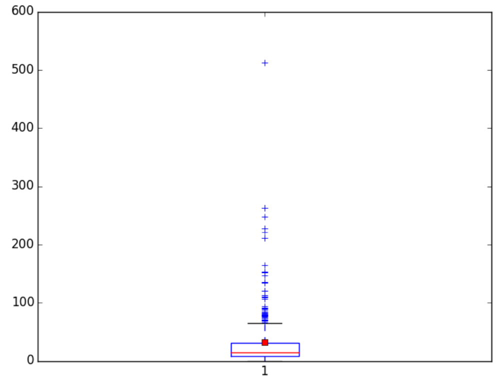


```
Age and class
        Adult         Baby       Child      Senior   Youngster
1  72.22%(78)    66.67%(2)  100.00%(1)  48.39%(30)  91.67%(11)
2  43.80%(53)  100.00%(12)  100.00%(5)   30.43%(7)   50.00%(6)
3  20.99%(51)   52.00%(13)   30.43%(7)    5.56%(1)  28.26%(13)
```

The number in the bracket means the number of survived people in this cell.

Exploration Phase

Single Variable Explorations:



This is the box plot of Titanic „Fare“ column.

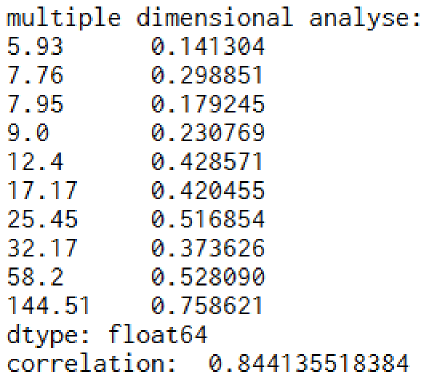
The mean fare: 32.2

The median fare : 14.45

25 percentile is 7.91

75 percentile is 31

multiple-variable exploration phase:



```
multiple dimensional analyse:
5.93      0.141304
7.76      0.298851
7.95      0.179245
9.0       0.230769
12.4      0.428571
17.17     0.420455
25.45     0.516854
32.17     0.373626
58.2      0.528090
144.51    0.758621
dtype: float64
correlation:  0.844135518384
```

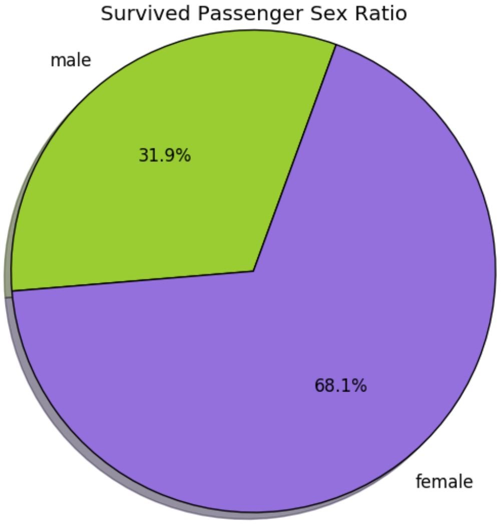
I separate all the passengers into 10 groups according to their ticket price.

Data at left is the average price of each group, and right column shows the survive rate of each group.

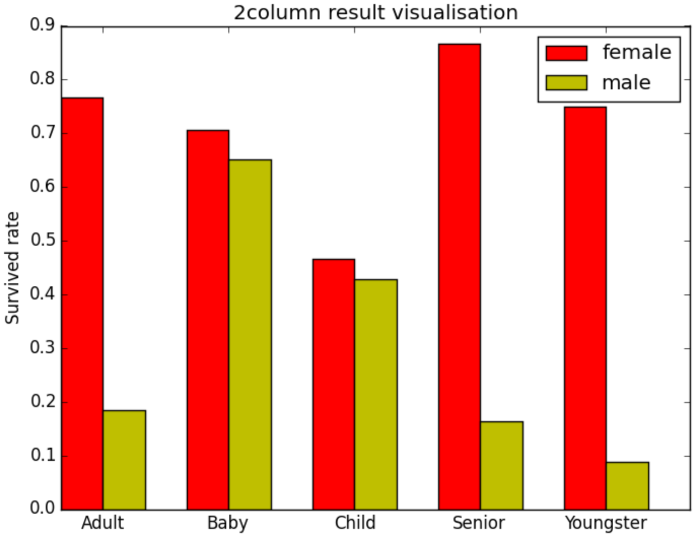
The correlation between them is 0.844, which means a strong correlation.

We can say that, high price ticket is strongly related to a high survive rate.

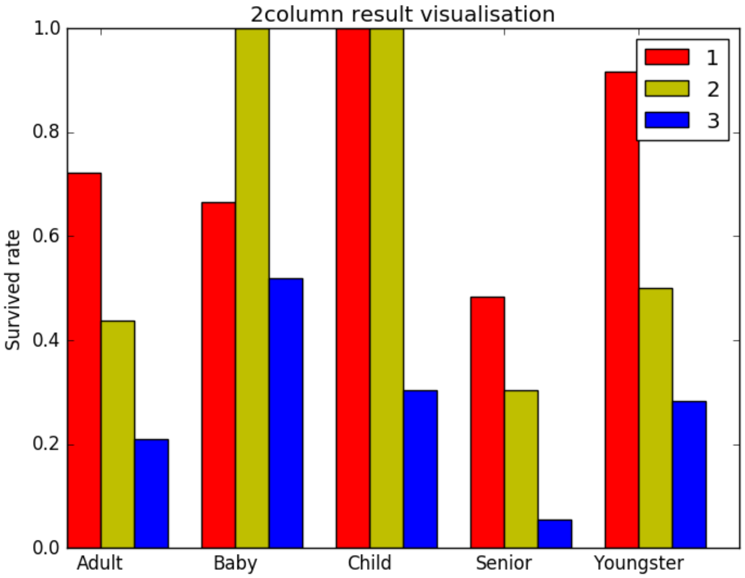
Visualization:



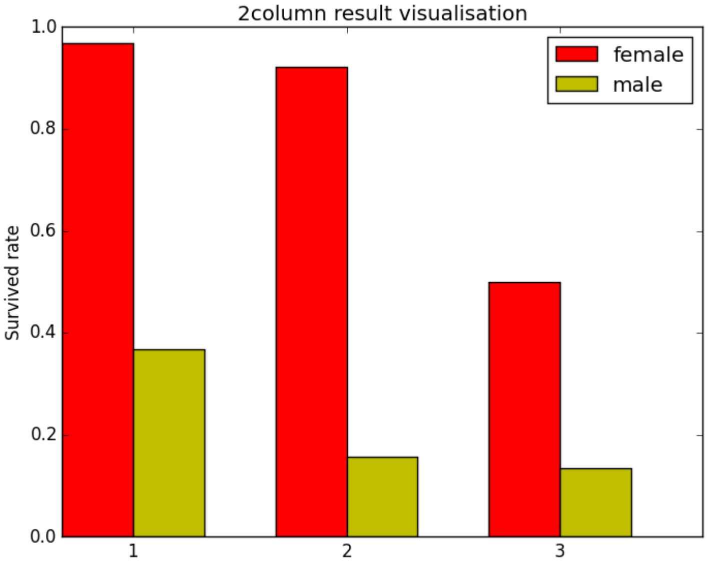
In the survived population, the number of female is more than 2 times of the number of male.



I classify all the passengers, who recorded their age, into different groups: Baby, child, youngster, adult and senior.



This shows the survival rate of passengers in different class and age.



The number of x-ticks-labels shows the class of the passengers.

Conclusion:

Limitation:

1. When the population is divided into smaller groups, the sample size is not large enough to show the pattern. In other words, the data is hugely influenced by fortuity.

2. When i clean the data within the column „Age“, omitting the value may lead to some error at last.

3. If we have the information about height, weight or the descendant number of each passenger, we can find more interesting result.

Summary

Generally, female has much higher survival rate than male, extremely for senior passengers, which is almost 5 times.

Almost in any age, people in a higher class also have higher survival rate than lower class.

(Except for baby, it may be caused by fortuity.)

For babies, the survival rate between male and female are not obvious.

Statistical Test

We can use a Chi-Squared Test to find whether there are associations between 2 variables of the passengers.