MG 221

Applied Probability and Statistics

Assignment 2

Department of Management Studies

Indian Institute of Science, Bangalore

Submitted By:

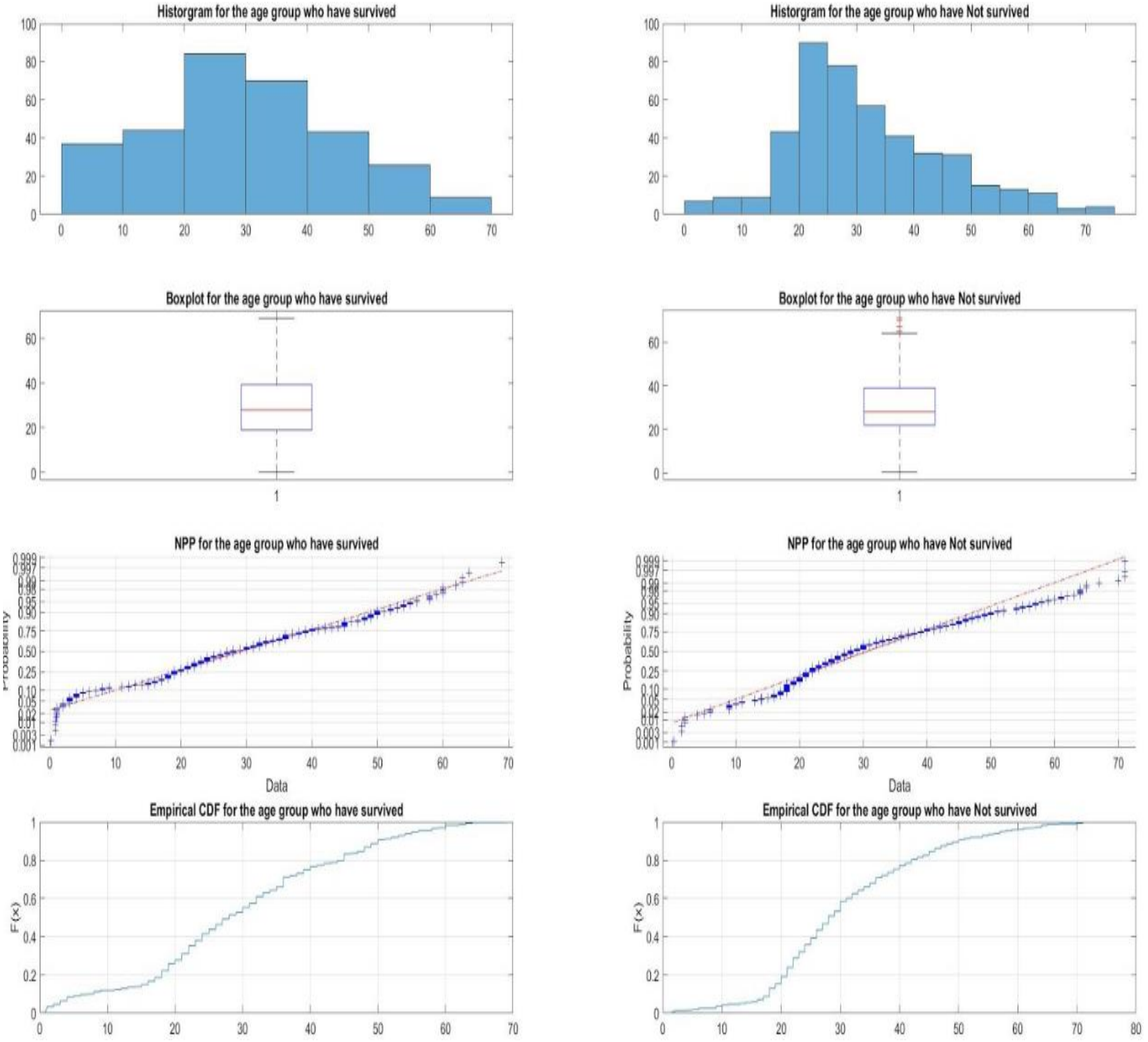
Kumar Prerak

Under the guidance of:

Prof. Chiranjit Mukhopadhyay

Q1) Is there a significant difference in Age distribution between those who survived and those who did not?

Graphical method to check if Age distribution of those who survived and that of those who did not, are normal or not:



Inference from graphical representation: It looks like distribution of people who survived appears to be close to normal but not exactly while distribution of people who did not survive appears to be “not normal”.

Tests of Normality:

Method	p Value (normtest(survage))	p Value (normtest(nonsurvage))
Shapiro-Wilk normality test	0.000499614	1.72E-08
Anderson-Darling normality test	0.00176702	0.0005
Kolmogorov-Smirnov normality test	1.07E-250	0

Conclusion: Neither 'Age of Survived' nor 'Age of Not Survived' have normal distributions.

We therefore use the following:

```
>wilcox.test(survage,nsurvage,alt="two.sided")
```

Wilcoxon rank sum test with continuity correction:

> Ho: There is no difference in distribution of age in both cases-survived and not survived.

> Ha: There is a difference in distribution of age in both cases-survived and not survived.

data: survived$Age and Not_Survived$Age

W= 65469 , **p-value = 0.1917**

alternative hypothesis: true location shift is not equal to 0

Conclusion: Therefore, **no enough evidence to prove that there is any significant difference in the age distribution between those who survived and those who did not.**

Z-test:

Since this is a **large sample**, we have performed a **z test**

> Ho: There is no difference in distribution of age in both cases-survived and not survived.

> Ha: There is a difference in distribution of age in both cases-survived and not survived.

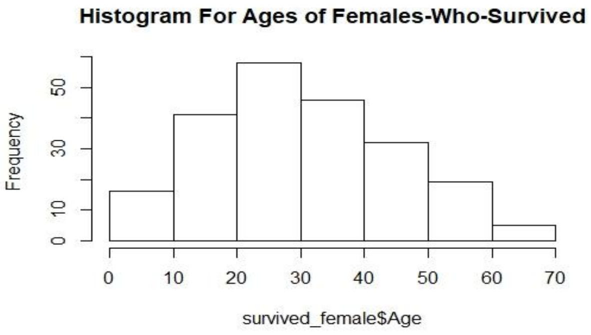
p value = 0.09935709

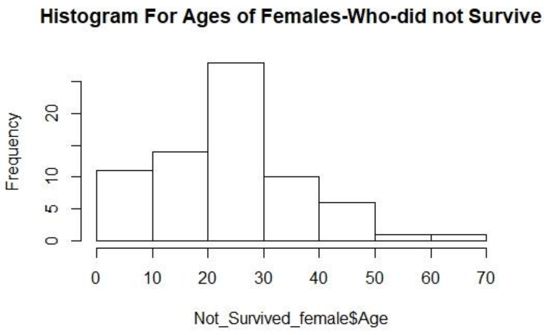
Conclusion: Keeping the significance level at α=0.05, **there is not enough evidence to prove that there is any significant difference in the age distribution between those who survived and those who did not** (p-value is >0.05).

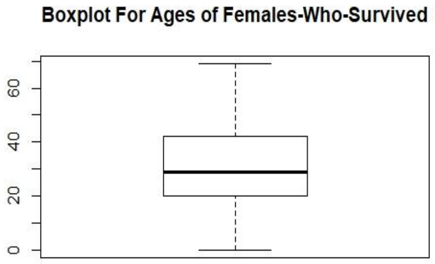
Q2. Answer the same as above after controlling for Gender.

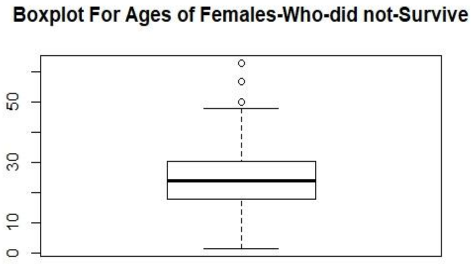
i) Comparison of age distributions of survived Female passengers and those female passengers who did not survive.

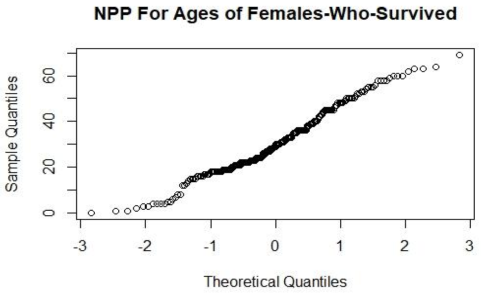
<u>Graphical techniques</u>

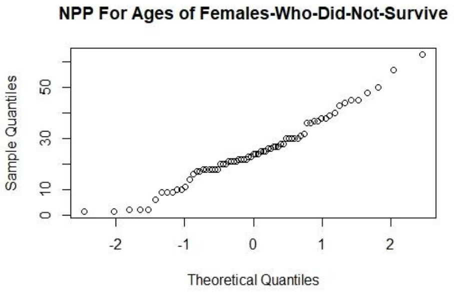












From the above three plots, for survived and not survived case based on gender (Female), normality is not observed.

<u>Tests for normality-</u>

```
> normtest(survived_female$Age)#not normal
> normtest(Not_Survived_female$Age)
```

Method	p Value (survived)	p Value (non-survived)
Shapiro-Wilk normality test	0.0026901879	0.11296551
Anderson-Darling normality test	0.0006357403	0.11530637
Cramer-von Mises normality test	0.0008234442	0.08483381
Kolmogorov-Smirnov normality test	0.0001661670	0.12109238
Shapiro-Francia normality test	0.0077707718	0.11744076

Conclusion: In case “Age of Survived” for female passengers have non-normal distribution and in case of “Age of Not Survived” for female passengers have normal distributions.

WILCOXON TEST-

Hypothesis testing for female-

> **Ho=**There is no difference in distribution of age in both cases-survived and not survived- for females

> **Ha=**There is a difference in distribution of age in both cases-survived and not survived- for female

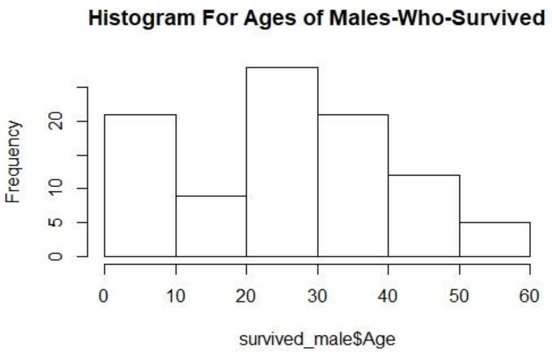
```
> wilcox.test(survived_female$Age,Not_Survived_female$Age)

 Wilcoxon rank sum test with continuity correction

 data:  survived_female$Age and Not_Survived_female$Age

W = 9408.5, p-value = 0.005119
```

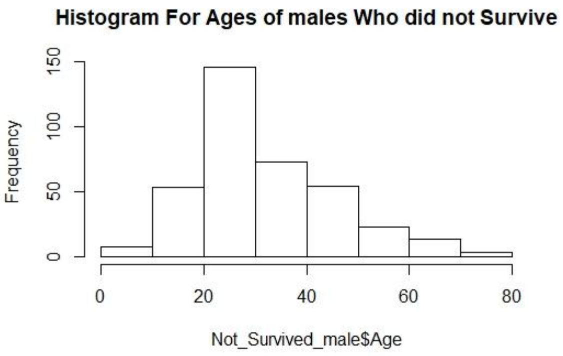
Alternative Hypothesis: true location shift is not equal to 0

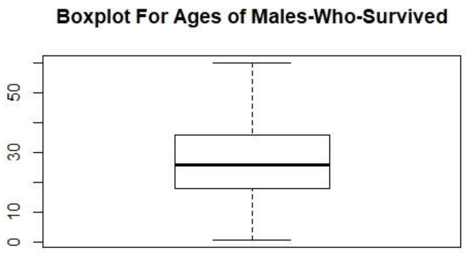
Conclusion: P value is 0.005119**, there is enough evidence to assume that there is a significant difference between age of those females who survived and those who did not**.

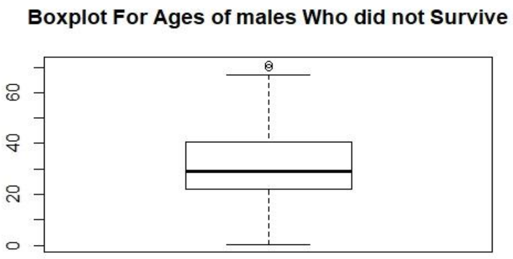
ii) Comparison of age distributions of survived Male passengers and those male passengers who did not survive.

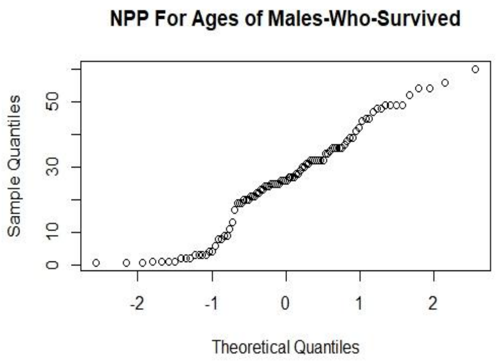
- **<u>Graphical techniques</u>**

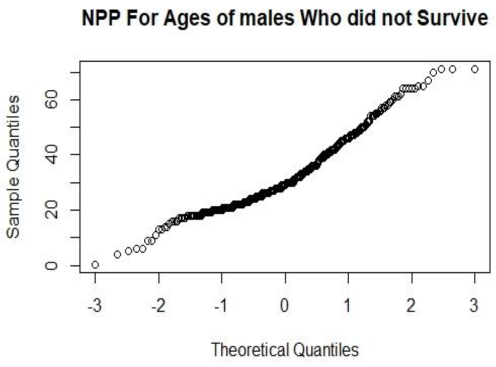












From the above three plots, for survived and not survived case based on gender (Male), normality is not observed.

Tests of Normality:

> normtest(Not_Survived_male$Age)#Not normal

Method	p Value (survived)	p Value (non-survived)
Shapiro-Wilk normality test	0.004201276	6.368376e-10
Anderson-Darling normality test	0.008390771	2.227363e-16
Cramer-von Mises normality test	0.045051620	7.370000e-10
Kolmogorov-Smirnov normality test	0.059854415	6.088379e-15
Shapiro-Francia normality test	0.013508441	8.325134e-09

Conclusion: Neither ‘Age of Survived - Males’ nor ‘Age of Not Survived - Males’ have normal distributions.

Hypothesis testing-Male-

> **Ho=**There is no difference in distribution of age in both cases-survived and not survived- for males

> **Ha=**There is a difference in distribution of age in both cases-survived and not survived- for males

> **Wilcoxon-Rank Sum Test** used here. Distributions aren’t normally distributed

> wilcox.test(survived_male$Age,Not_Survived_male$Age)

Wilcoxon rank sum test with continuity correction

data: survived_male$Age and Not_Survived_male$Age

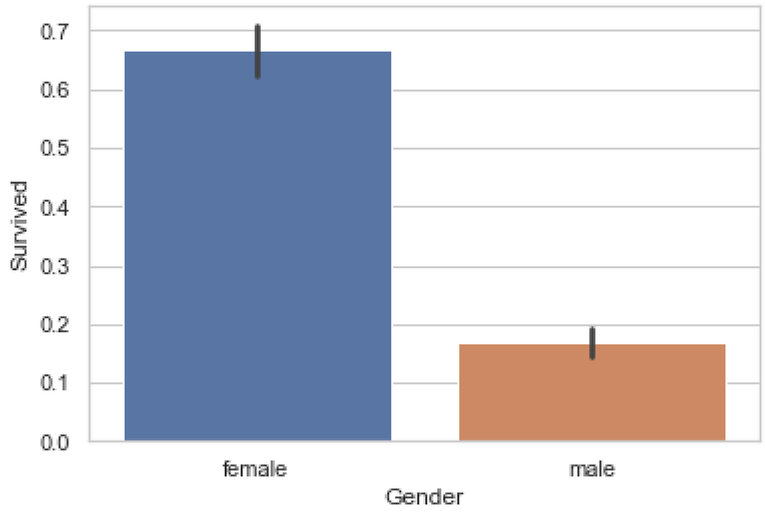
W = 14453, p-value = 0.003962

alternative hypothesis: true location shift is not equal to 0

Conclusion: P value is p-value = 0.003962. Indicating a strong evidence against Ho. **There is enough evidence to assume that there is a significant difference between age of those males who survived and those who did not**.

Q.3) Is there a significant difference in Survival Probabilities for the two Genders?

Graphical representation:



Inference from graphical representation: It appears that survival probability of female is more than that of male.

```
> cont_table

        Gender   Survived   Not_Survived

        Female    308         154

          Male    142         709
```

> Hypothesis Testing

> Ho: There is no significant difference in survival probability between the two genders.

> Ha: There is a significant difference in survival probability between the two genders.

```
> fisher.test(cont_table[,-1])

Fisher's Exact Test for Count Data

data:  cont_table[, -1]

p-value < 2.2e-16

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval: 7.601263 - 13.122462

sample estimates:

odds ratio

 9.965185
```

> Hypothesis Testing

> Ho: There is no significant difference in survival probability between the two genders.

> Ha: There is a significant difference in survival probability between the two genders.

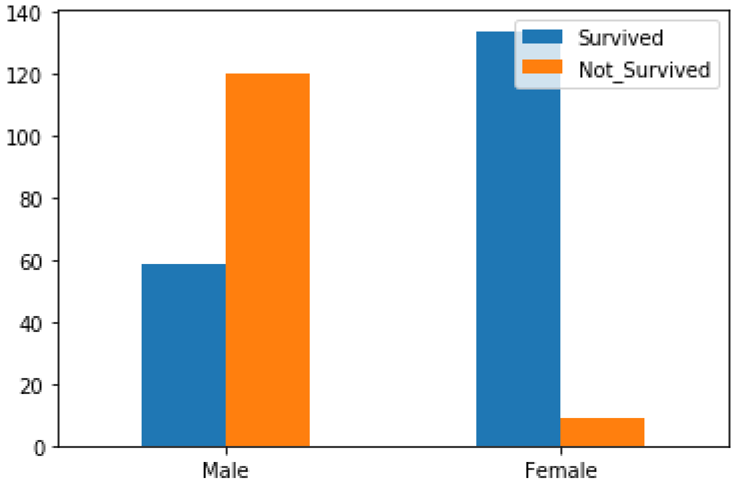
```
> chisq.test(cont_table[,-1])

 Pearson's Chi-squared test with Yates' continuity correction

 data:  cont_table[, -1]

Chi-squared = 329.84, df = 1, p-value < 2.2e-16
```

Conclusion: An extremely low p-value < 2.2e-16 is observed. There is a **significant difference in survival probability between the two genders**. Therefore, the survival probability of an individual is dependent on the gender. Survival probability of female is more than that of male.

Q.4) Is there a significant difference in Survival Probabilities for the two Genders even after taking the effect of Passenger Class into account?

Class 1-

Graphical representation:



Inference: It appears that Survival probability of female is greater than that of male in PClass 1st.

```
> cont_table #class1
```

```
  Gender     Survived     Not_Survived
1 female         134            9
2   male         59           120
```

> Hypothesis Testing

> Ho: There is no significant difference in survival probability in class 1 between the two genders.

> Ha: There is a significant difference in survival probability in class 1 between the two genders.

```
> chisq.test(cont_table[,-1])
```

```
Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table[, -1]

X-squared = 119.64, df = 1, p-value < 2.2e-16
```

>Hypothesis Testing

>Ho: There is no significant difference in survival probability in class 1 between the two genders.

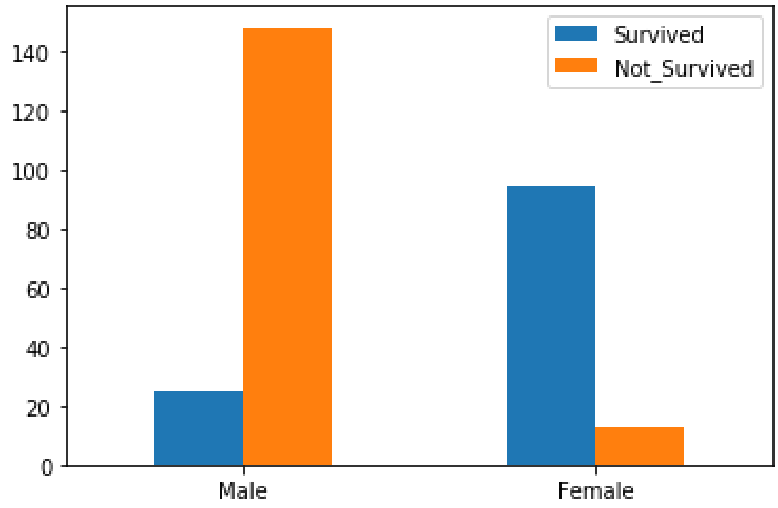
>Ha: There is a significant difference in survival probability in class 1 between the two genders.

```
> fisher.test(cont_table[,-1])

Fisher's Exact Test for Count Data

data:  cont_table[, -1]

p-value < 2.2e-16

Alternative hypothesis: true odds ratio is not equal to 1

95 % CI: [14.00568, 71.53596]

sample estimates:

odds ratio - 29.901
```

Conclusion: A very low p-value < 2.2e-16 is observed. **Strong evidence against the null hypothesis. Significant difference in survival probability between the two genders for class 1.**

Class 2-

Graphical representation:



Inference: It appears that Survival probability of female is greater than that of male in PClass 2nd.

```
> cont_table #class2

  Gender    Survived     Not_Survived

3 female       94                13

4   male       25                148
```

> Hypothesis Testing

> Ho: There is no significant difference in survival probability in class 2 between the two genders.

> Ha: There is a significant difference in survival probability in class 2 between the two genders.

```
> fisher.test(cont_table[,-1])

Fisher's Exact Test for Count Data

data:  cont_table[, -1]

p-value < 2.2e-16

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

19.87698 94.70673

sample estimates:

odds ratio - 41.80798
```

> Hypothesis Testing

> Ho: There is no significant difference in survival probability in class 2 between the two genders.

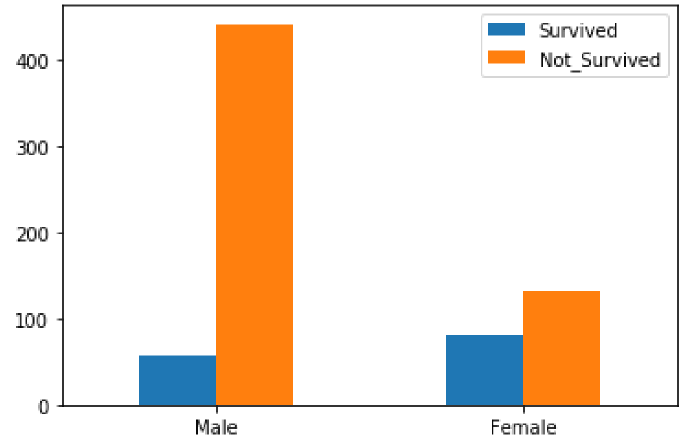
> Ha: There is a significant difference in survival probability in class 2 between the two genders.

```
> chisq.test(cont_table[,-1])

Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table[, -1]

X-squared = 142.76, df = 1, p-value < 2.2e-16
```

Conclusion: A very low p-value < 2.2e-16 is observed. **Significant difference in survival probability between the two genders for class 2.**

Class 3 –

Graphical representation:



Inference: It appears as if although there is difference between survival probabilities of two genders, more passengers in each group died than those who survived. We can say that most passengers in Class 3 could not survive.

```
> cont_table class3

  Gender          Survived          Not_Survived

5 female             80                        132

6   male             58                        441
```

> Hypothesis Testing

> Ho: There is no significant difference in survival probability in class 3 between the two genders.

> Ha: There is a significant difference in survival probability in class 3 between the two genders.

> fisher.test(cont_table[,-1])

Fisher's Exact Test for Count Data

data: cont_table[, -1]

p-value = 1.184e-14

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

3.061374 6.939940

sample estimates:

odds ratio - 4.596261

> **Hypothesis Testing**

> **Ho:** There is no significant difference in survival probability in class 3 between the two genders.

> **Ha:** There is a significant difference in survival probability in class 3 between the two genders.

```
> chisq.test(cont_table[,-1])

Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table[, -1]

X-squared = 63.201, df = 1, p-value = 1.867e-15
```

Conclusion: A very low p-value is observed. **Significant difference in survival probability between the two genders for class 3.**

Overall Conclusion on PClass from Graphical interpretation: Survival probability of

PClass 1> PClass 2> PClass 3

Q.5) Analyse how the Survival Probability got affected by the joint (combined) effect of Gender, Age and Passenger Class, and based on this analysis, write a brief and consolidated report on how the Survival Probabilities of Titanic passengers got affected by these three (independent) variables.

Step-1

Missing value treatment for Age

In the given titanic datasheet, some of the age values are missing which are shown as NA in the original datasheet. To avoid the loss the data we did data maniputation in the following way:

- Categorized the data into 12 groups based on PClass, gender, Survival (3*2*2=12).
- NA in each group was replaced by the mean of respective groups excluding NA values.

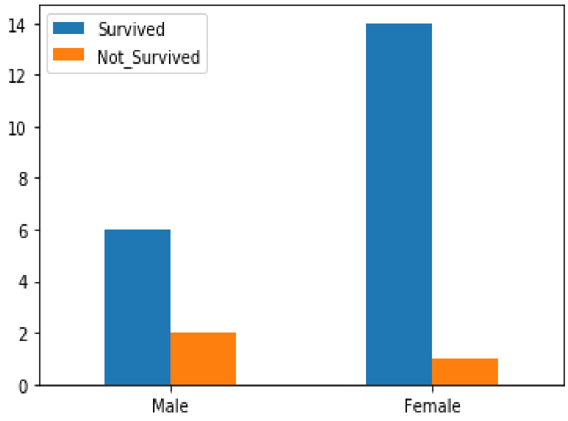
Step-2

- Categorized the Ages into three different groups such as teenagers (age<= 20 years), adults (age between 20 to 50 years) and old-ages (age >50).

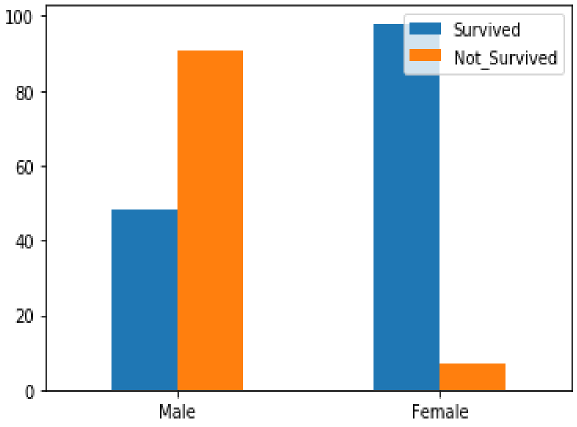
Step-3

Graphical representation

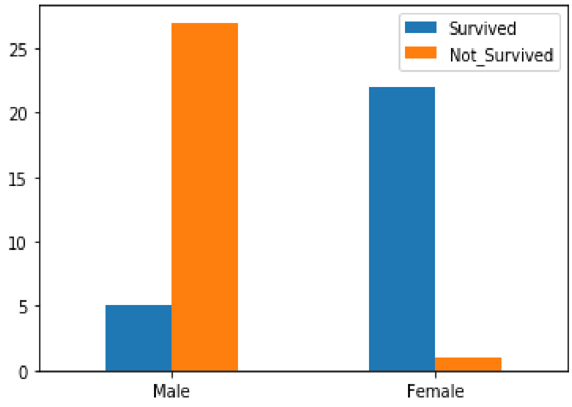
PClass-1 And teenagers



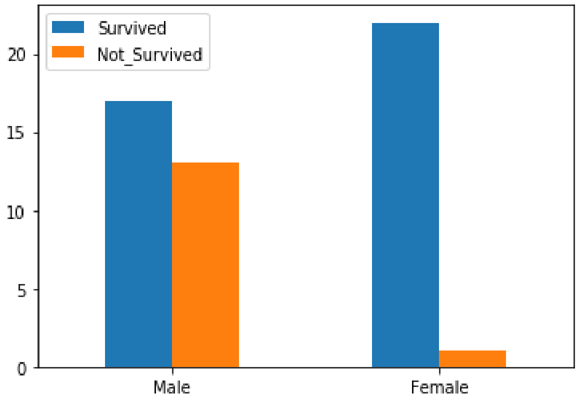
PClass-1 And Adults



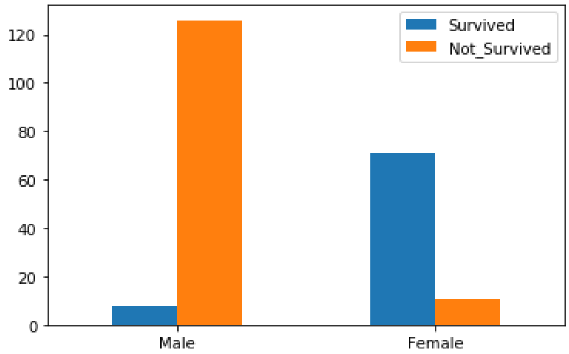
PClass-1 And Old-age



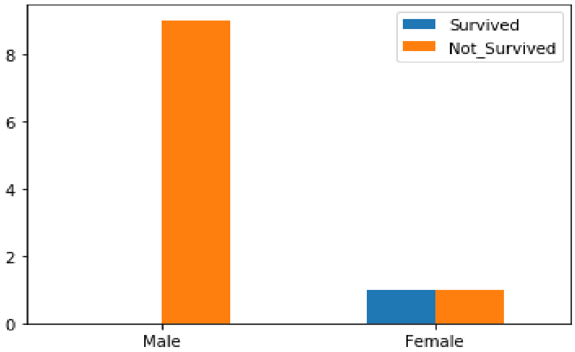
PClass-2 And teenagers



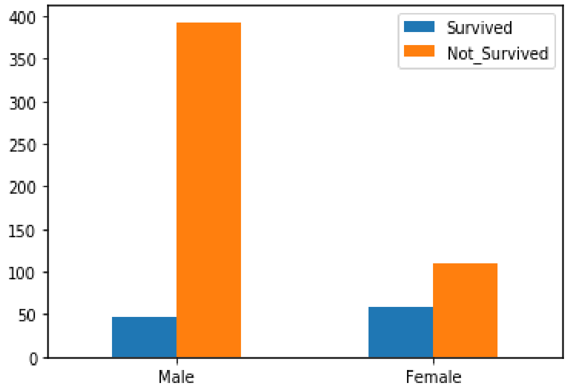
PClass-2 And Adults



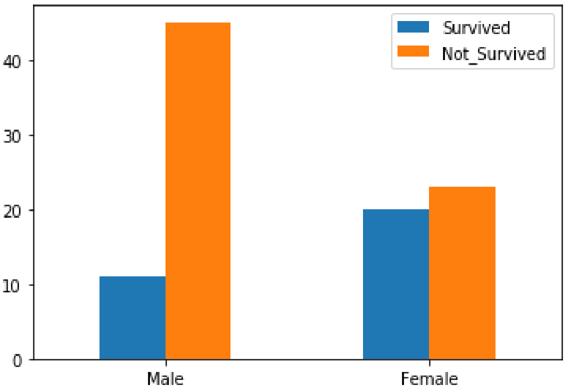
PClass-2 And old-age



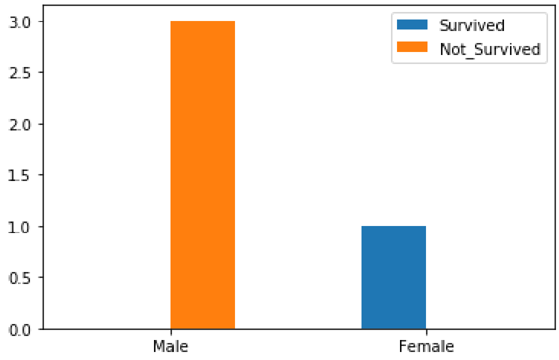
PClass-3And Teenagers



PClass-3And Adults



PClass-3 And old-age



Step-4

Hypothesis testing

Ho: There is no significant difference between the gender's survival rate in each of the 9 classes (3 PClasses X 3 Age- groups).

Ha: Females' survivability is more than that of men

- For every PClass corresponding to an Age-group, the survival and non-survival count of male and female group is obtained in the form of a 2 by 2 contingency table.

- Fisher’s one-sided exact test was performed for every 2 by 2 contingency table and p-values are reported as follows:

P-values:

	1st class	2nd class	3rd class
teenagers	0.268775	0.001466	0.008124
Adults	1.36E-22	1.32E-35	2.04E-11
Old age	1.27E-09	0.181818	0.25

Inferences

Based on the Fisher’s test performed above the following inferences are drawn.

1) **Among the teenagers:**
   - In PClass-1, since p-value is 0.26, we can’t say that there is significant difference between the male and female survivability.
   - In PClass 2 and 3, since p-value is very small, there is strong evidence that female survivability is significantly more than that of men.

2) **Among the adults:**
   - In all the PClasses, since p-value is negligible, there is very strong evidence that there is a significant difference in Survival Probability of males and females, with female survivability more than that of men.

3) **Among the old-age people:**
   - In PClass-1, since p-value is very less, there is very strong evidence that there is a significant difference in Survival Probability of males and females, with female survivability more than that of men.
   - In class 2 and 3, There is not sufficient evidence to say that there is significant difference in the male and female survival probability.