**HYPOTHESIS TESTING ON TITANIC DATA**

**MG 220**

Submitted By:

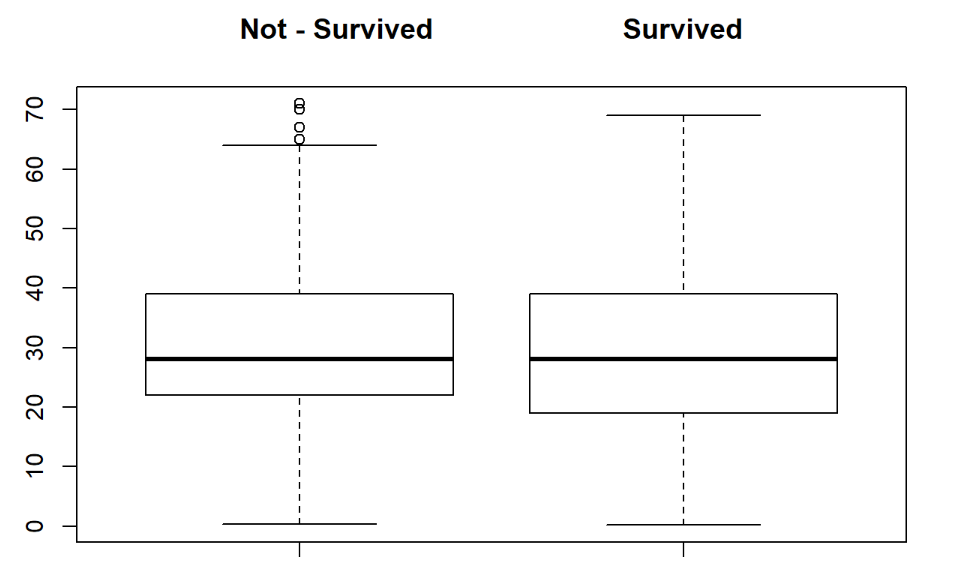
1. Kumar Saurav

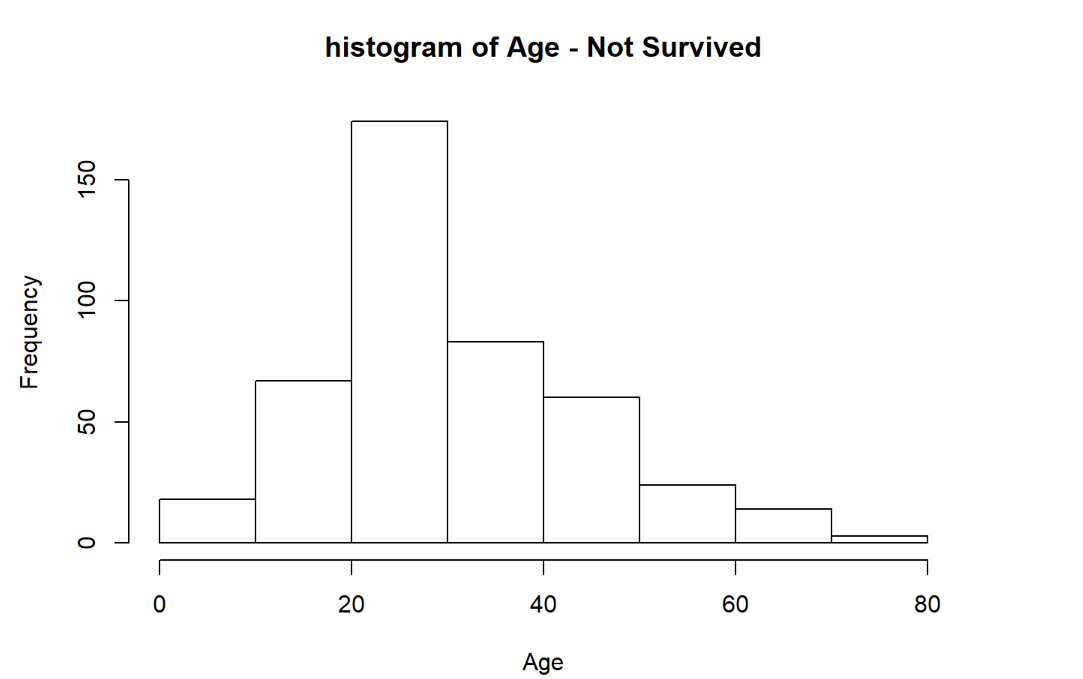
HYPOTHESIS TESTING ON TITANIC DATA

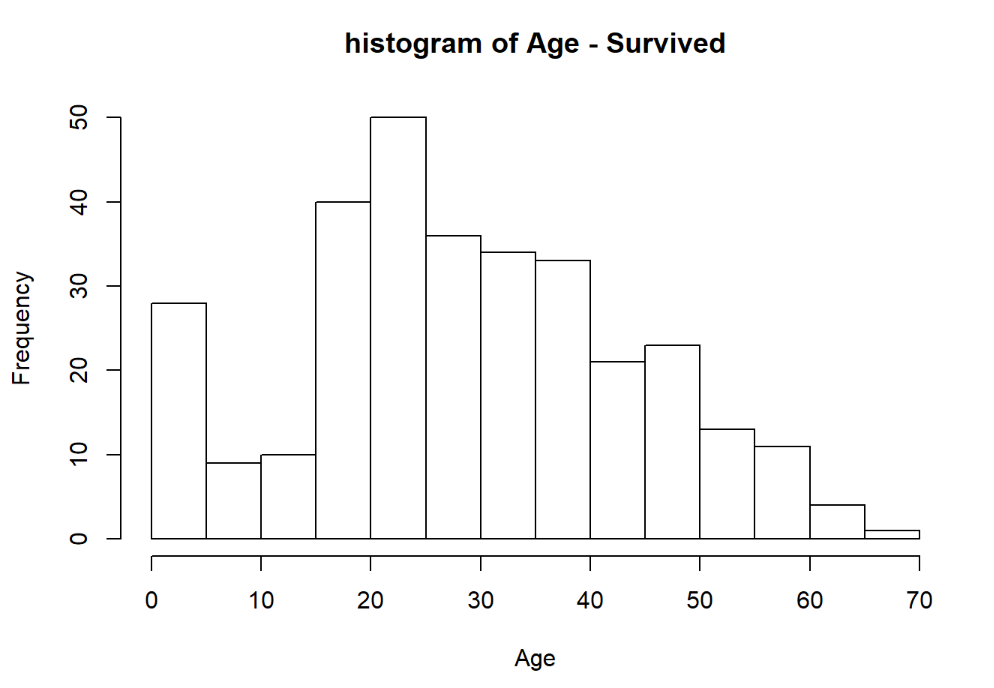
MG 220

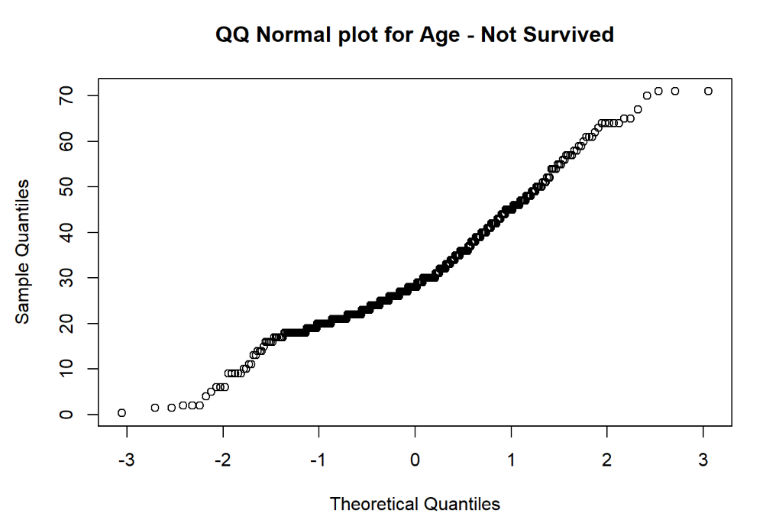
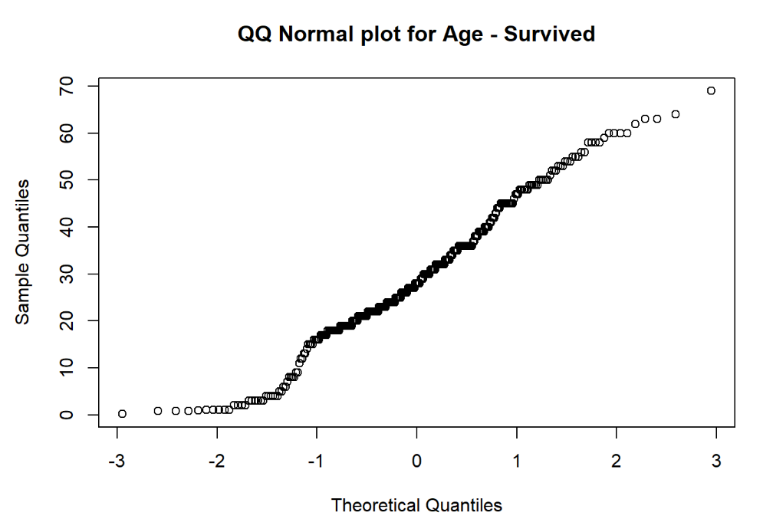
1. Is there a significant difference in Age distribution between those who survived and those who did not?

* Firstly, we observe the distribution of data







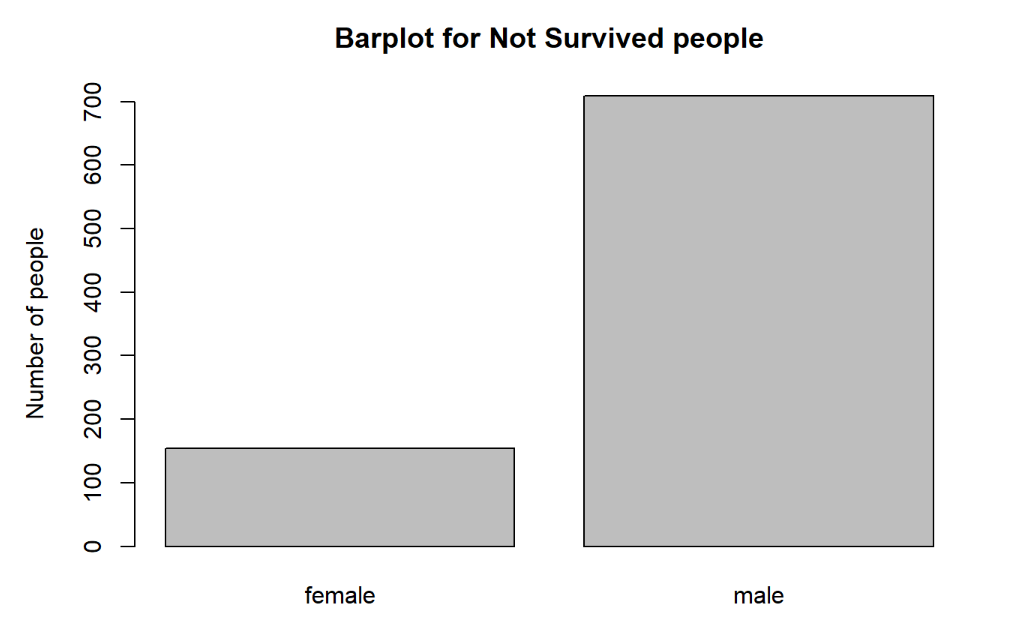


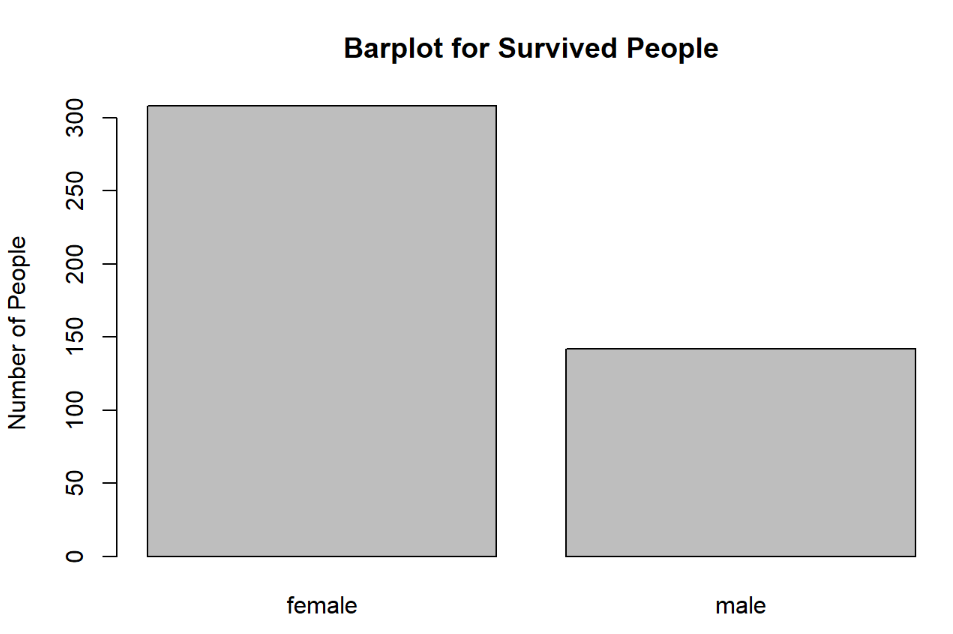
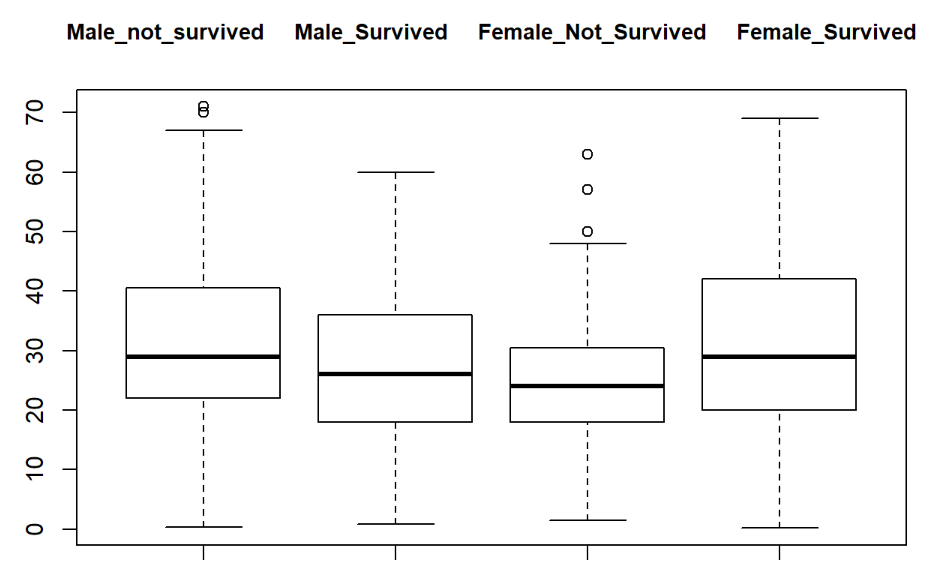
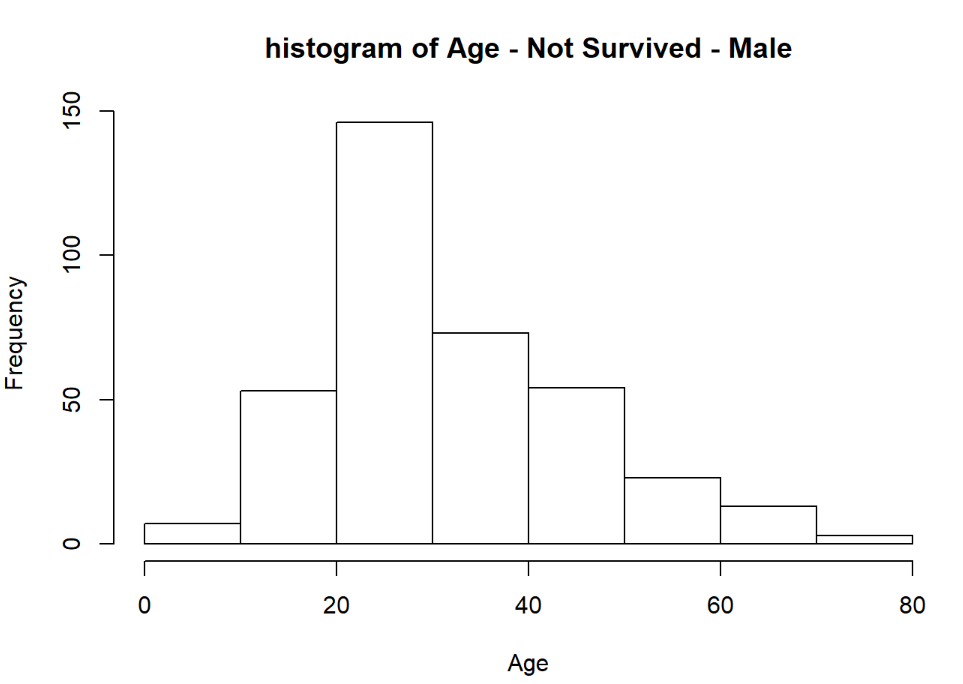
* For people who survived:
* Method P.value
* 1 Shapiro-Wilk normality test 0.0004996137
* 2 Anderson-Darling normality test 0.0017831268
* 3 Cramer-von Mises normality test 0.0068092810
* 4 Lilliefors (Kolmogorov-Smirnov) normality test 0.0322461921
* 5 Shapiro-Francia normality test 0.0020591830
* For people who did not survive:
* Method P.value
* 1 Shapiro-Wilk normality test 1.461444e-09
* 2 Anderson-Darling normality test 5.014279e-16
* 3 Cramer-von Mises normality test 7.370000e-10
* 4 Lilliefors (Kolmogorov-Smirnov) normality test 2.646866e-16
* 5 Shapiro-Francia normality test 1.719147e-08
* The P-Value of the Normality tests are less than 0.05, strong evidence against null hypothesis that the data follows normal distribution
* This implies both the data for survived and not survived people did not follow normal distribution
* HYPOTHESIS TESTING:
  + **Null Hypothesis**: There is a no difference in Age distribution between those who survived and those who did not.
  + **Alternate Hypothesis**: There is a difference in Age Distribution between those who survived and those who did not
  + Firstly, we applied Flinger’s Test to compare the variances of the two distributions
  + Fligner-Killeen: med chi-squared = 513.45, df = 1, p-value < 2.2e-16
  + Therefore, strong evidence against the null hypothesis that the variances of these two distributions are equal.

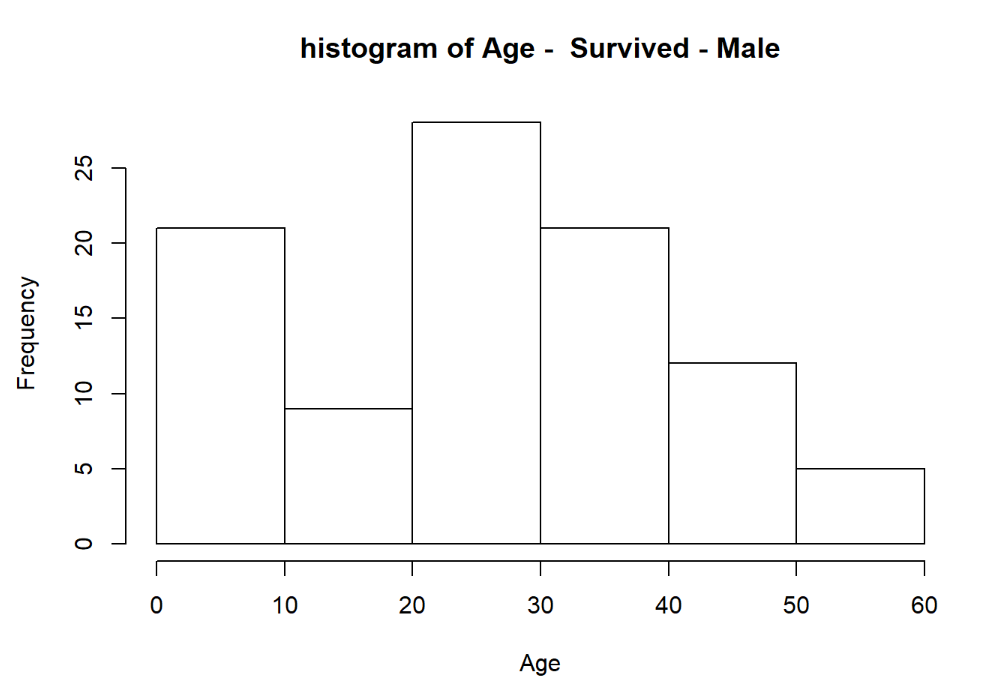
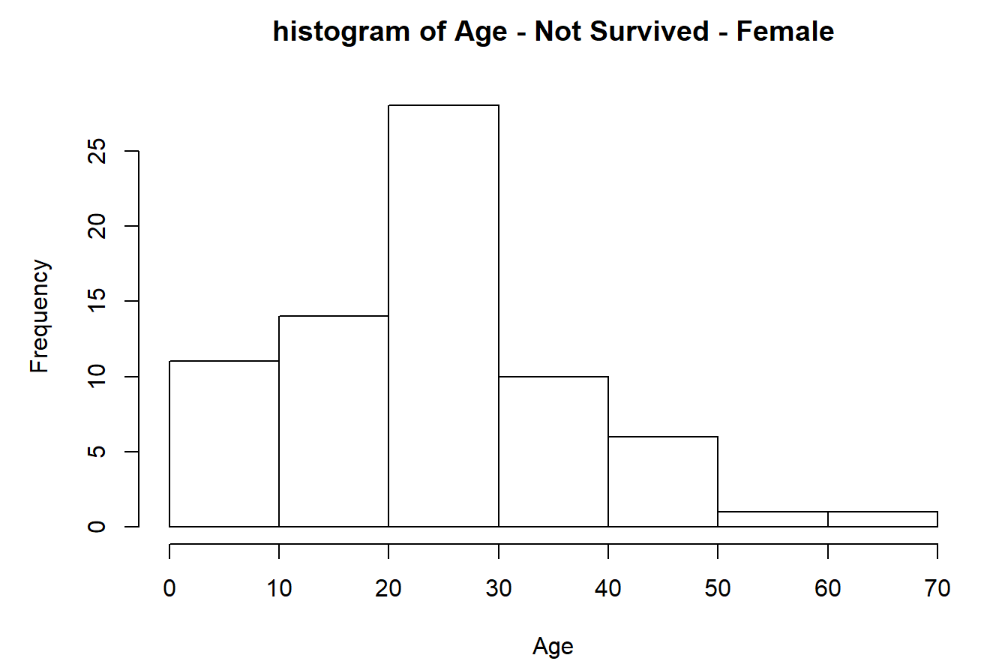
* Now, we apply Wilcoxon Rank sum Test.
* Wilcoxon rank sum test with continuity correction
* W = 73190, p-value = 0.1917
* alternative hypothesis: true location shift is not equal to 0
* The P-Value of the Wilcoxon rank sum test is 0.192. There is no strong evidence against Null hypothesis.
* This implies there is a **No significant difference**between the age of the people who have survived and who have not.

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

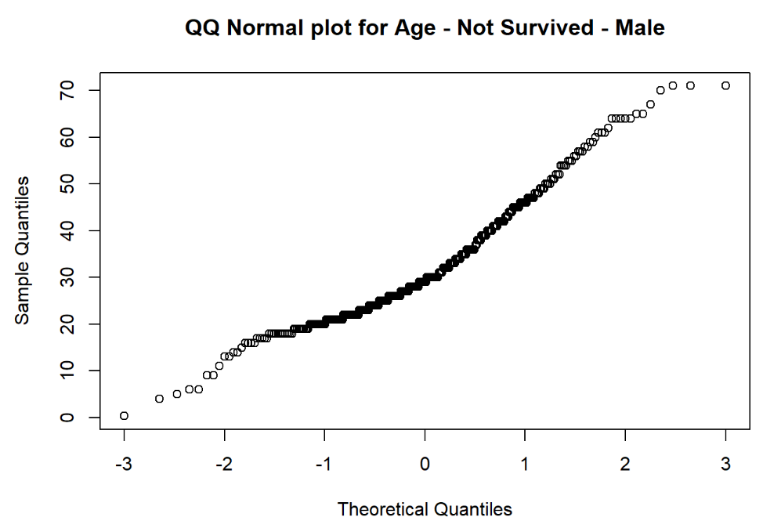
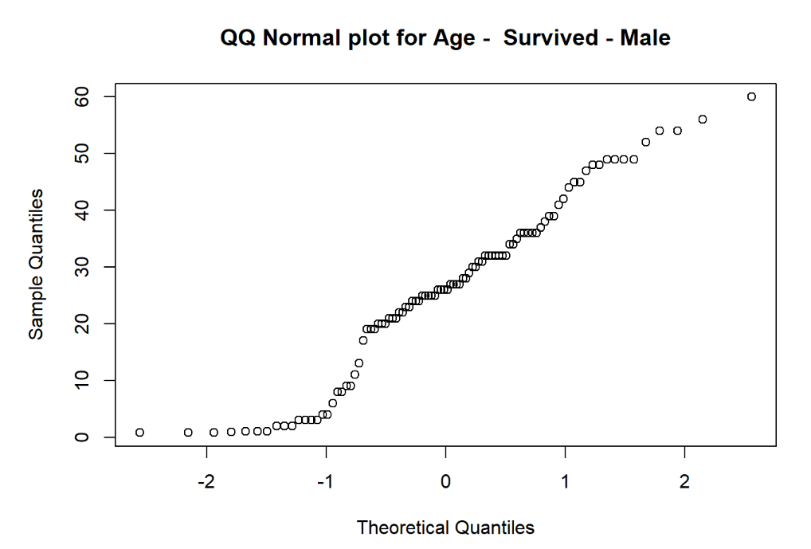
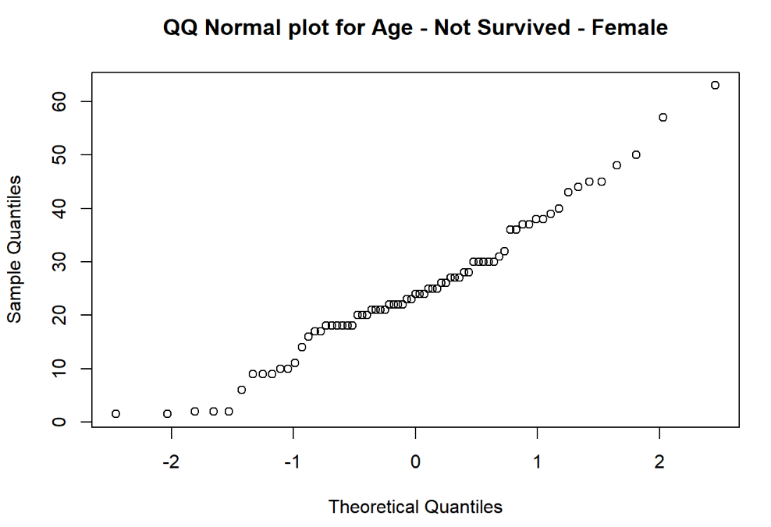
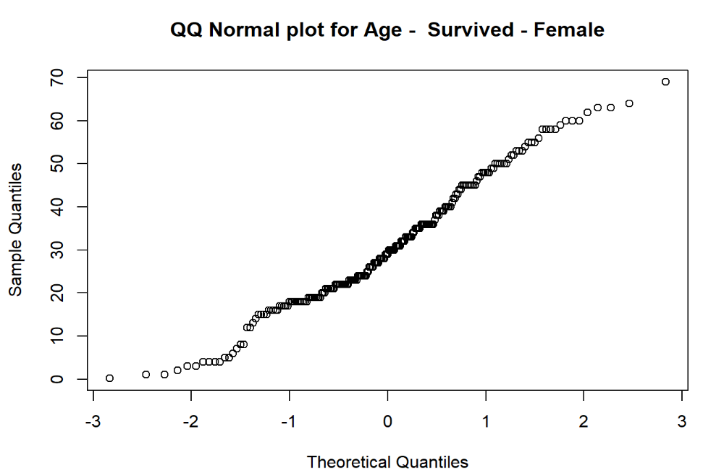
* Again, we observe the distribution of data across the two genders





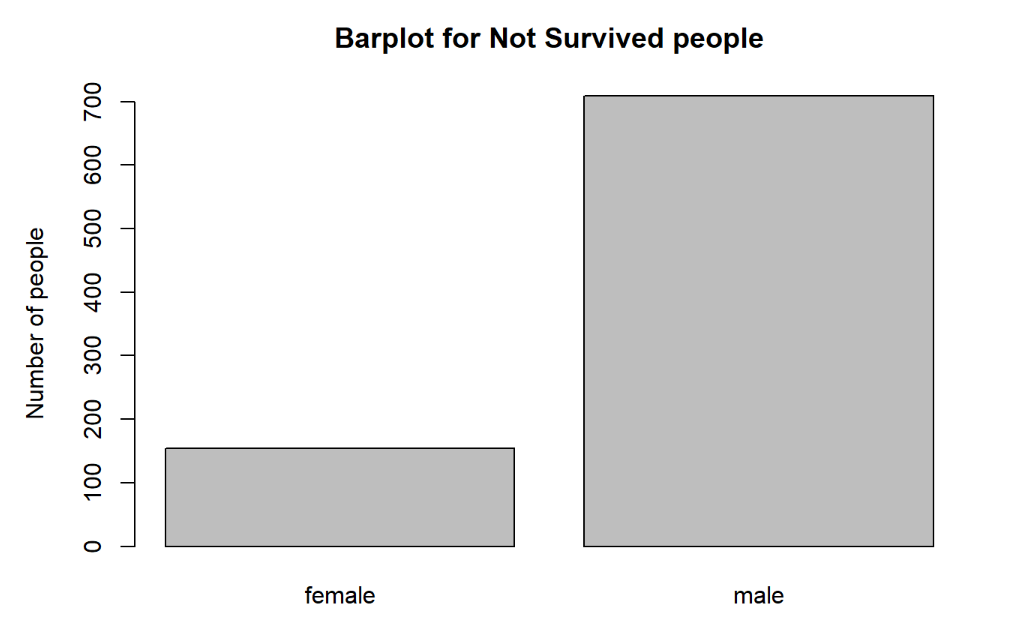
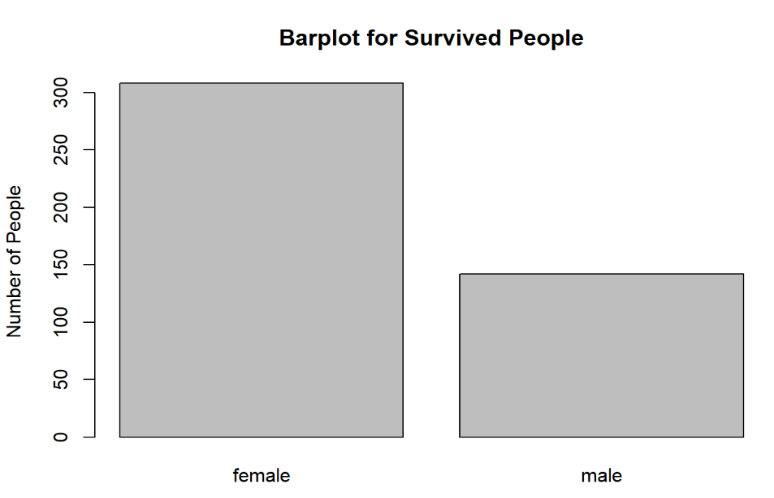






* For male passengers who did not survive:
  + Method P.value
  + Shapiro-Wilk normality test 6.368376e-10
  + Anderson-Darling normality test 2.227363e-16
  + Cramer-von Mises normality test 7.370000e-10
  + Lilliefors (Kolmogorov-Smirnov) normality test 6.088379e-15
  + Shapiro-Francia normality test 8.325134e-09
* For male passengers who survived:
  + - Method P.value
    - Shapiro-Wilk normality test 0.004201276
    - Anderson-Darling normality test 0.008390771
    - Cramer-von Mises normality test 0.045051620
    - Lilliefors (Kolmogorov-Smirnov) normality test 0.059854415
    - Shapiro-Francia normality test 0.013508441
* For female passengers who survived:
  + Method P.value
  + Shapiro-Wilk normality test 0.0026901879
  + Anderson-Darling normality test 0.0006357403
  + Cramer-von Mises normality test 0.0008234442
  + Lilliefors (Kolmogorov-Smirnov) normality test 0.0001661670
  + Shapiro-Francia normality test 0.0077707718
* For female passengers who did not survive:
  + Method P.value
  + Shapiro-Wilk normality test 0.11296551
  + Anderson-Darling normality test 0.11530637
  + Cramer-von Mises normality test 0.08483381
  + Lilliefors (Kolmogorov-Smirnov) normality test 0.12109238
  + Shapiro-Francia normality test 0.11744076
* Female Passengers who did not survive have a normally distributed data.
* All the other categories like Male survived, Male not survived, Female survived data are not normal.
* HYPOTHESIS TESTING – FOR MALE
  + **Null Hypothesis:**There is no significant difference in Age distribution between those who survived and those who did not for male.
  + **Alternative Hypothesis:**There is significant difference in Age distribution between those who survived and those who did not for male.
  + Method P.Value
  + Asymptotic two-sample Kolmogorov-Smirnov test 0.002006277
  + Wilcoxon rank sum test with continuity correction 0.003961685
  + The P-Value of both the two-sample KS test and Wilcoxon rank sum test were found to be less than 0.05.
  + There is strong evidence against the Null Hypothesis.
  + There is significant difference in Age distribution between those who survived and those who did not for male.
* HYPOTHESIS TESTING – FOR FEMALE
  + **Null Hypothesis:**There is no significant difference in Age distribution between those who survived and those who did not for Female
  + **Alternative Hypothesis:**There is significant difference in Age distribution between those who survived and those who did not for Female
  + Method P.Value
  + Asymptotic two-sample Kolmogorov-Smirnov test 0.013259850
  + Wilcoxon rank sum test with continuity correction 0.005118923
  + The P-Value of both the two-sample KS test and Wilcoxon rank sum test were found to be less than 0.05.
  + There is strong evidence against the Null Hypothesis.
  + There is significant difference in Age distribution between those who survived and those who did not for Female.

1. Is there a significant difference in Survival Probabilities for the two Genders?



* HYPOTHESIS TESTING
  + **Null Hypothesis:** There is no difference in the survival probability between the two genders
  + **Alternative Hypothesis:** There is a difference in the survival probability between the two genders
  + Method P.Value
  + Pearson's Chi-squared test with Yates' continuity correction 1.040403e-73
  + Fisher's Exact Test for Count Data 4.826448e-74
  + Both the Chi-Sq. test and Fisher test are giving P-Value less than 0.05.
  + There is strong evidence against Null Hypothesis
  + The Survival Rate is dependent on the Gender based on chi square test for independence and Fisher’s Exact

1. Is there a significant difference in Survival Probabilities for the two Genders even after taking the effect of Passenger Class into account?
   * Firstly, we check whether there is significant difference in Survival Probabilities across the Passenger Class
   * Survived Un-Survived
   * Class-1 193 129
   * Class-2 119 161
   * Class-3 138 573
   * HYPOTHESIS TESTING
     + **Null Hypothesis:** The Survival probability is same across all the passenger class
     + **Alternate Hypothesis:** The Survival probability is not same across all the passenger class
     + Method P.Value
     + Pearson's Chi-squared test 3.852316e-38
     + Fisher's Exact Test for Count Data 2.791493e-38
     + Both the Chi-Sq. test and Fisher test are giving P-Value less than 0.05.
     + There is strong evidence against Null Hypothesis
     + Therefore, the survival probabilities are not same across the passenger classes
   * Now, we check whether there is significant difference in Survival Probabilities for the two Genders after taking the effect of Passenger Class into account.
     + CLASS 1 Passengers

Female Male

* + - * Survived 9 120
      * Not-Survived 134 59
      * HYPOTHESIS TESTING:
        + **Null Hypothesis:** There is no difference between the survival proportion between the male and female passenger of class 1
        + **Alternate Hypothesis:** There is a significant difference between the survival proportion between the male and female passenger of class 1
        + Method P.Value
        + Pearson's Chi-squared test 7.576739e-28
        + Fisher's Exact Test 1.086515e-31
        + The P-Value of both the Chi-Sq test and Fisher Exact Test is less than 0.05
        + There is strong evidence against Null Hypothesis
        + There is a significant difference between the Survival Probability of Male and Female of the class-1 Passengers.
    - CLASS 2 Passengers:

Survived Not-Survived

* + - * Female 13 94
      * Male 148 25
      * HYPOTHESIS TESTING:
        + **Null Hypothesis:** There is no difference between the survival proportion between the male and female passenger of class 2
        + **Alternate Hypothesis:** There is a significant difference between the survival proportion between the male and female passenger of class 2
        + Method P.Value
        + Pearson's Chi-squared test 6.633397e-33
        + Fisher's Exact Test 5.079791e-36
        + The P-Value of both the Chi-Sq test and Fisher Exact Test is less than 0.05
        + There is strong evidence against Null Hypothesis
        + There is a significant difference between the Survival Probability of Male and Female of the class-2 Passengers.
    - CLASS 3 Passengers:

Survived Not-Survived

* + - * Female 132 80
      * Male 441 58
      * HYPOTHESIS TESTING:
        + **Null Hypothesis:** There is no difference between the survival proportion between the male and female passenger of class 3
        + **Alternate Hypothesis:** There is a significant difference between the survival proportion between the male and female passenger of class 3
        + Method P.Value
        + Pearson's Chi-squared test 1.866807e-15
        + Fisher's Exact Test 1.184368e-14
        + The P-Value of both the Chi-Sq test and Fisher Exact Test is less than 0.05
        + There is strong evidence against Null Hypothesis
        + There is a significant difference between the Survival Probability of Male and Female of the class-3 Passengers.

1. How the Survival Probability got affected by the joint (combined) effect of Gender, Age and Passenger Class?

* Firstly, we discretize the Age into bins: Young(Age < 18), Adult(18<Age<45) and Old(Age>45)

Young Adult Old

* + Survived 58 181 52
  + Not-Survived 38 309 68
* HYPOTHESIS TESTING: (For Age Group)
  + - * **Null Hypothesis:** Survival probabilities and age groups are independent
      * **Alternate Hypothesis:** Survival Probability is dependent on age group

Method P.Value

* + - * Pearson's Chi-squared test 9.471394e-05
      * Fisher's Exact Test 1.093512e-04
      * Since, p value is less than 0.05, Survival probability is dependent on age groups
* HYPOTHESIS TESTING: (For Gender)
  + **Null Hypothesis:** Survival probabilities and gender are independent
  + **Alternative Hypothesis:** Survival Probability is dependent on gender

Method P.Value

* + Pearson's Chi-squared test 1.040403e-73
  + Fisher's Exact Test for Count Data 4.826448e-74
  + Both the Chi-Sq. test and Fisher test are giving P-Value less than 0.05.
  + There is strong evidence against Null Hypothesis
  + The Survival Rate is dependent on the Gender based on chi square test for independence and Fisher’s Exact
* HYPOTHESIS TESTING (For P-Classes)
  + - **Null Hypothesis:** Survival probabilities and Passenger classes are independent
    - **Alternate Hypothesis:** Survival probability is dependent on P-class.

Method P.Value

* + - Pearson's Chi-squared test 3.852316e-38
    - Fisher's Exact Test for Count Data 2.791493e-38
    - Both the Chi-Sq. test and Fisher test are giving P-Value less than 0.05.
    - There is strong evidence against Null Hypothesis
    - Therefore, the survival probabilities are dependent on P-Class
* Since, Survival probability is dependent on Gender, Age and P-class, we can use these features in a **logistic regression model** in order to predict the survival probabilities.
* glm(formula = V5 ~ V4 + V3 + V2, family = binomial(link = "logit"), data = df)
  + where V2 is the passenger class, V3 is the age and V4 is the gender

Coefficients Estimate Std. Error z-value Pr(>|z|)

* + (Intercept) 3.759662 0.397567 9.457 < 2e-16
  + Gender -2.631357 0.201505 -13.058 < 2e-16
  + Age -0.039177 0.007616 -5.144 2.69e-07
  + P-2nd class -1.291962 0.260076 -4.968 6.78e-07
  + P-3rd class -2.521419 0.276657 -9.114 < 2e-16