Experimentation on titanic data set

Divyank Singh

21^{st} November 2020

Contents

1 Introduction

This report contains a statistical and probabilistic analysis of the 'Titanic' data set in R language.
'Titanic' data set - a well known Kaggle competition about the fate of passengers aboard the Titanic at the time of its shipwreck.
The following table describes a few entries of the data set:

Survive	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	3	Braund	male	22.00	1	0	A/5	7.2500		S
1	1	Cumings	female	38.00	1	0	PC	71.2833	C85	C
1	3	Heikkinen	female	26.00	0	0	STON	7.9250		S
0	1	McCarthy	male	54.00	0	0	17463	51.8625	E46	S
1	2	Nasser	female	14.00	1	0	237736	30.0708		C

Table 1: Few entries of data set

The first section contains the interpretation of columns. The second section is about the statistical interpretations of all the columns - graphical analysis, probabilistic and statistical observations. The third section is survival analysis where probabilities of surviving for passengers is explored.

2 Data set analysis

The data set has a dimension 891 X 12, which means 891 passenger records are available.
Finding dimensions of the data set in R

```r
data<-read.csv("E:/Jupyterfiles/ML_practice/Kaggle/Titanic/train.
    csv")
head(data) # column names and first five entries
dim(data) # returns dimensions
```

2.1 Interpreting the columns

This portion contains descriptions about the column, its data type, null values and classification of numerical data into discrete and continuous.

Column	About	D. Type	Missing vals.	Classification
Survived	Binary data depicting whether the passenger survived or not. 0: No, 1: Yes	int	-	Discrete
Pclass	Numeric column about ticket class 1: first class, 2: second class, 3: third class	int	-	Discrete
Name	Name of passenger	char	-	Discrete
Sex	Gender of passenger	char	-	Discrete
Age	Age of passenger	double	177	Continuous
SibSp	Numeric column about the number of siblings or spouses	int	-	Discrete
Parch	Number of parents or children aboard	int	-	Discrete
Ticket	Ticket number. Different for every passenger	char	-	Discrete
Fare	Passenger fare for the voyage	double	-	Continuous
Cabin	Cabin number allotted to some passengers	char	687	Discrete
Embarked	Port of embarkation C: Cherbourg, Q: Queenstown , S: Southampton	char	2	Discrete

Table 2: Column information

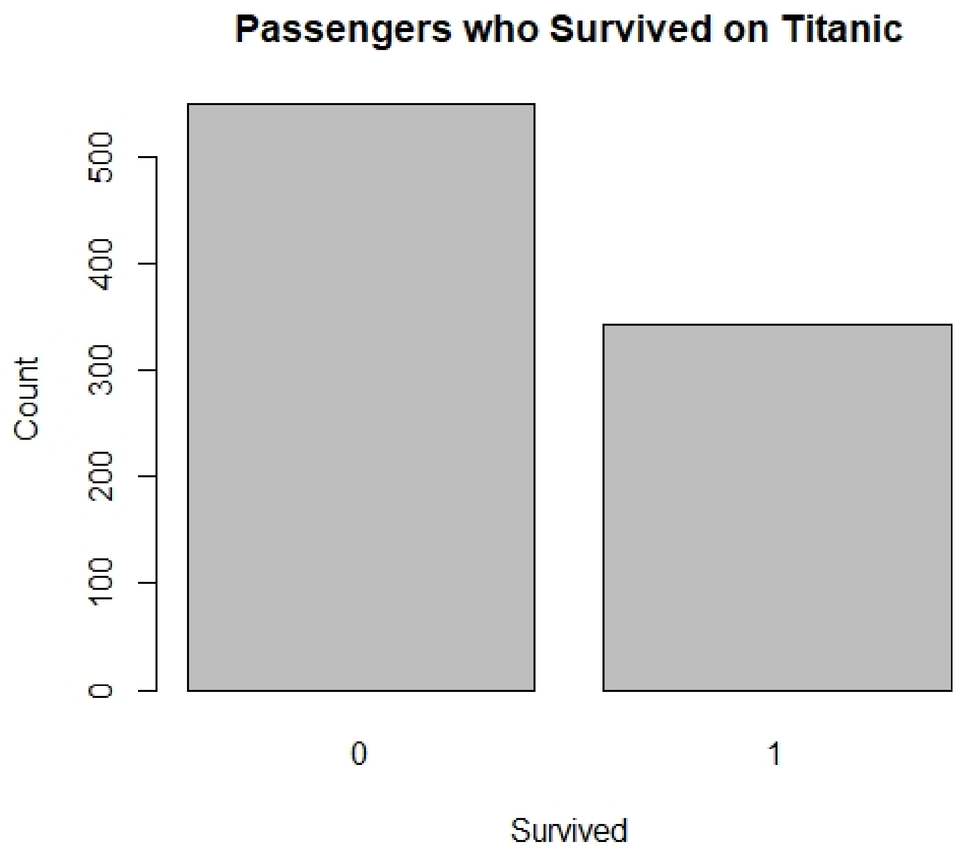
2.2 Statistical interpretation

2.2.1 Survived

The number of survivors is a discrete data type, so visualising by a bar plot[1]. The code for plotting bar plot in R language is as follows -

```
class.table = table(data$Survived)
barplot(class.table, xlab = "Survived", ylab = "Count", main = "
    Passengers who Survived on Titanic")
```

[1]A chart or graph representing comparisons in discrete data with rectangular bars horizontally or vertically having heights or lengths proportional to their values

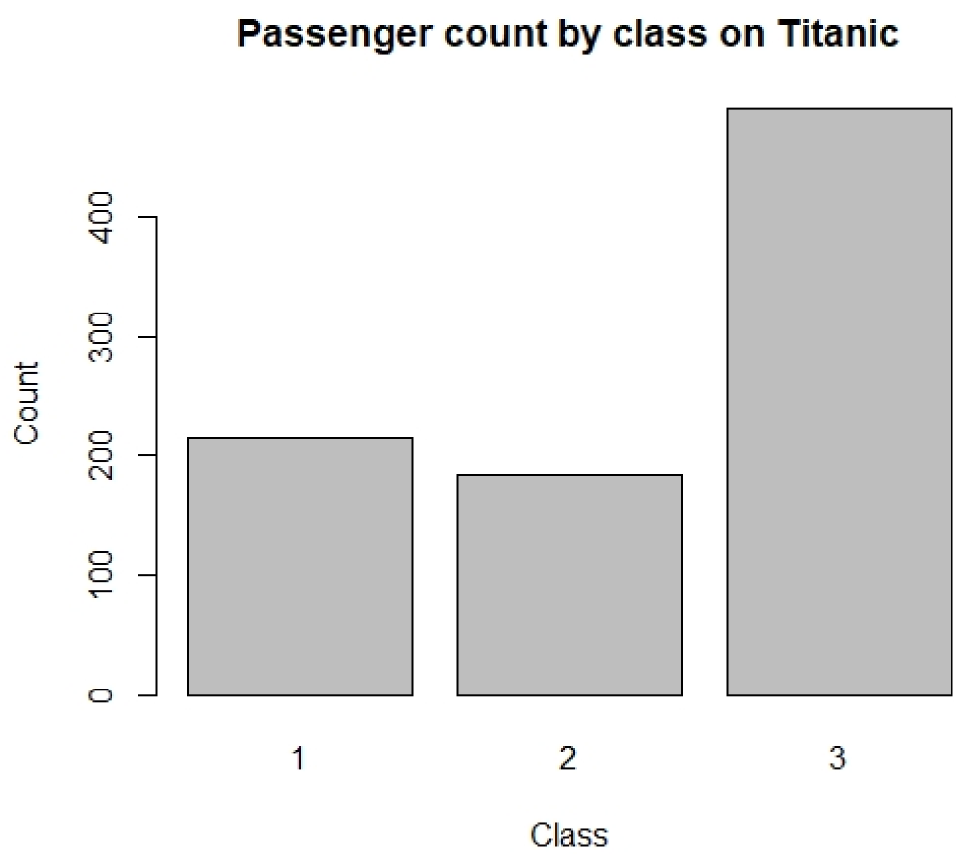


Conclusions from the bar plot -

1. Total survivors: 342
2. Probability of:
   - surviving = 0.38
   - not surviving = 0.62
3. Median for survived class: 0

2.2.2 Pclass

Pclass is a discrete column with 891 entries. Representing the number of tickets of each class purchased by passengers as a bar plot.



The R code:

```r
class.table = table(data$Pclass)
barplot(class.table, xlab = "Class", ylab = "Count", main = "
    Passenger count by class on Titanic")
```

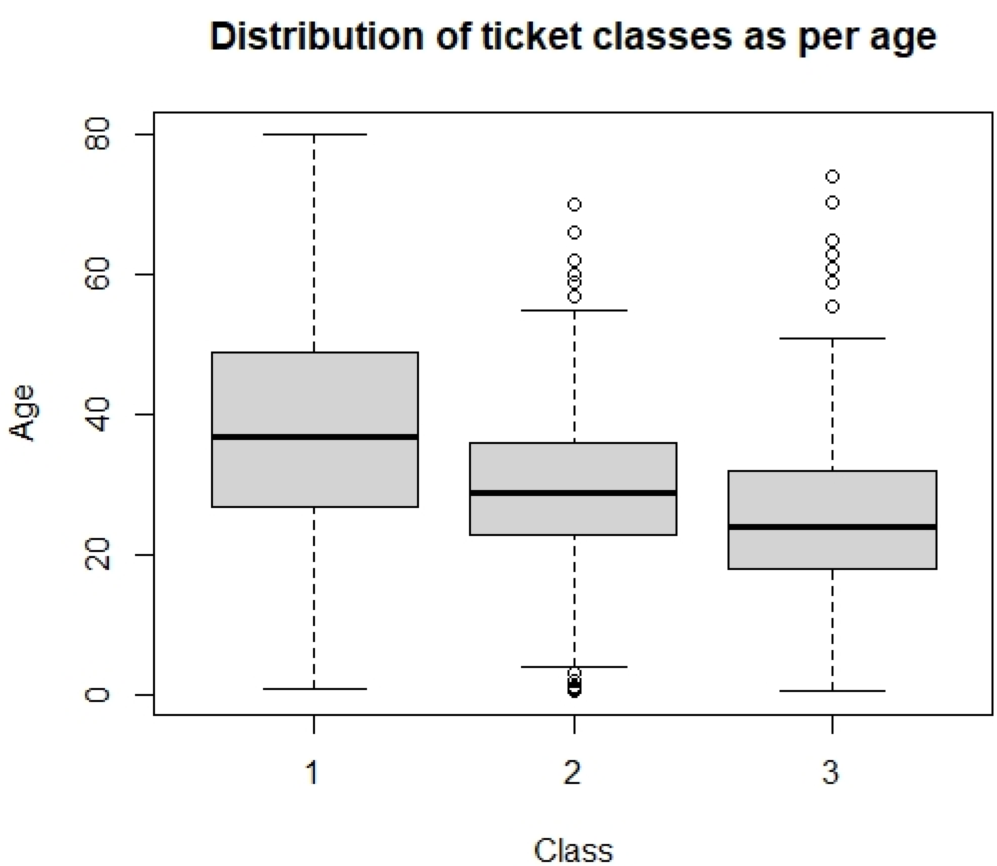
Important observations and conclusions -

1. Class wise bookings:
   - 1^{st} class traveller's $= 216$
   - 2^{nd} class traveller's $= 184$
   - 3^{rd} class traveller's $= 491$ (median)
2. Probability of passenger's travelling in various classes:
   - 1^{st} class $= 0.2424$

- 2^{nd} class = 0.2065
- 3^{rd} class = 0.5511

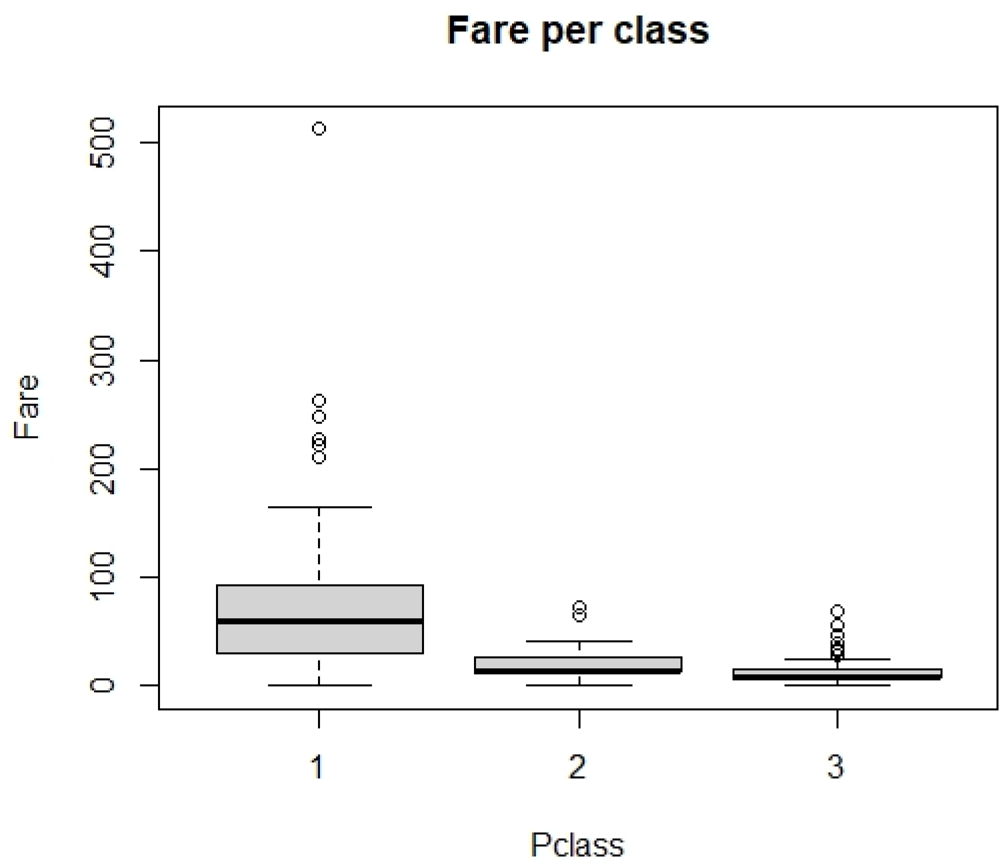
Having found out the number of passengers in each ticket class, lets find the distribution of ticket classes as per age in R using a box plot [2].

```
boxplot(data$Age~data$Pclass, main = "Distribution of ticket
    classes as per age", xlab = "Class", ylab = "Age")
```



This shows class1 had the oldest passengers and class3 had the youngest
Finding the distribution of ticket fare as per class

[2] They show five number summary of a set of data - minimum score, median, lower quartile, upper quartile and maximum score

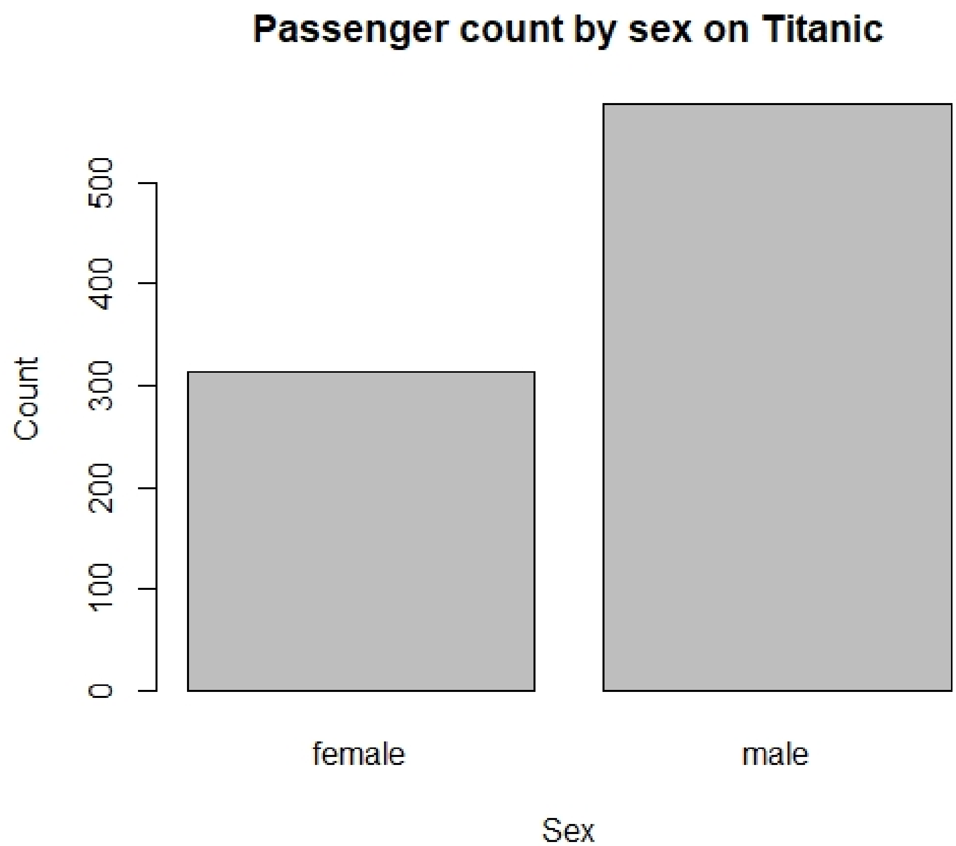


Inference: As expected class1 passengers paid the most and class3 passengers paid the least. Reaching this inference using R-

```
boxplot(data$Fare~data$Pclass, xlab = "Pclass", ylab = "Fare", main
     = "Fare per class")
```

2.2.3 Sex

Representing the number of male and female passengers out of 891 passengers aboard as a bar plot

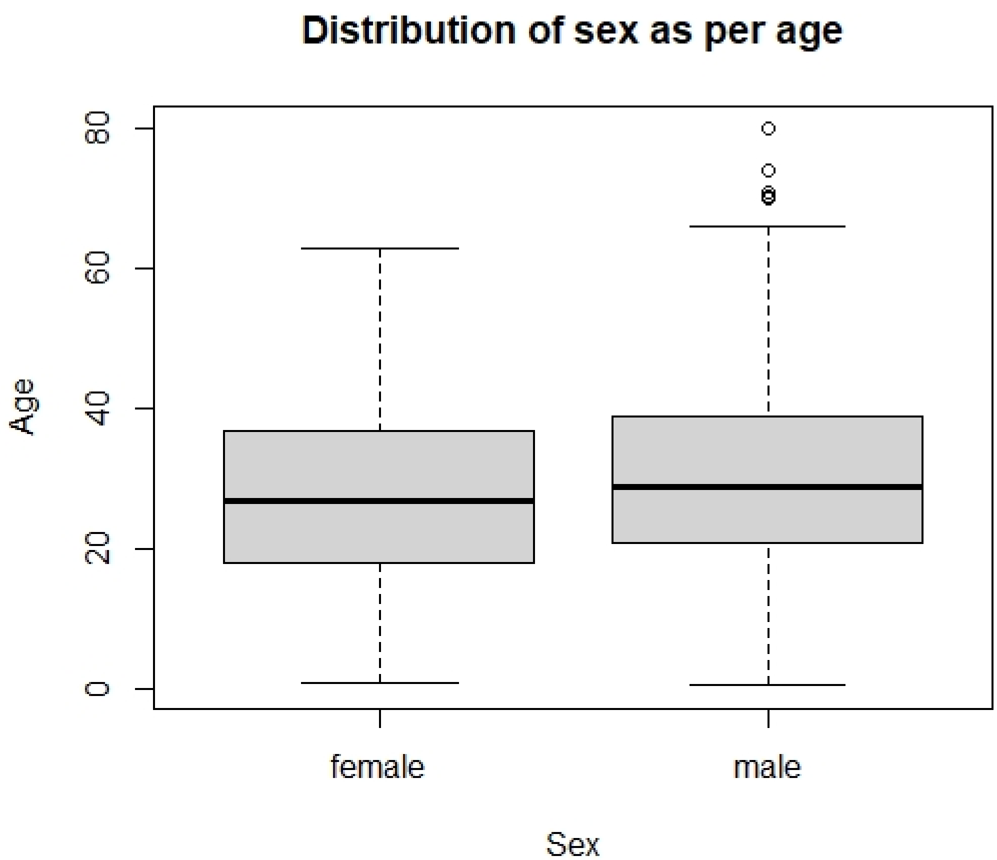


Use the following code to plot the above bar plot,

```
class.table = table(data$Sex)
barplot(class.table, xlab = "Sex", ylab = "Count", main = "
    Passenger count by sex on Titanic")
```

Having found the number of men and women it is important to find their ages. This will be done using a box plot.

```
boxplot(data$Age~data$Sex, main = "Distribution of sex as per age",
     xlab = "Sex", ylab = "Age")
```



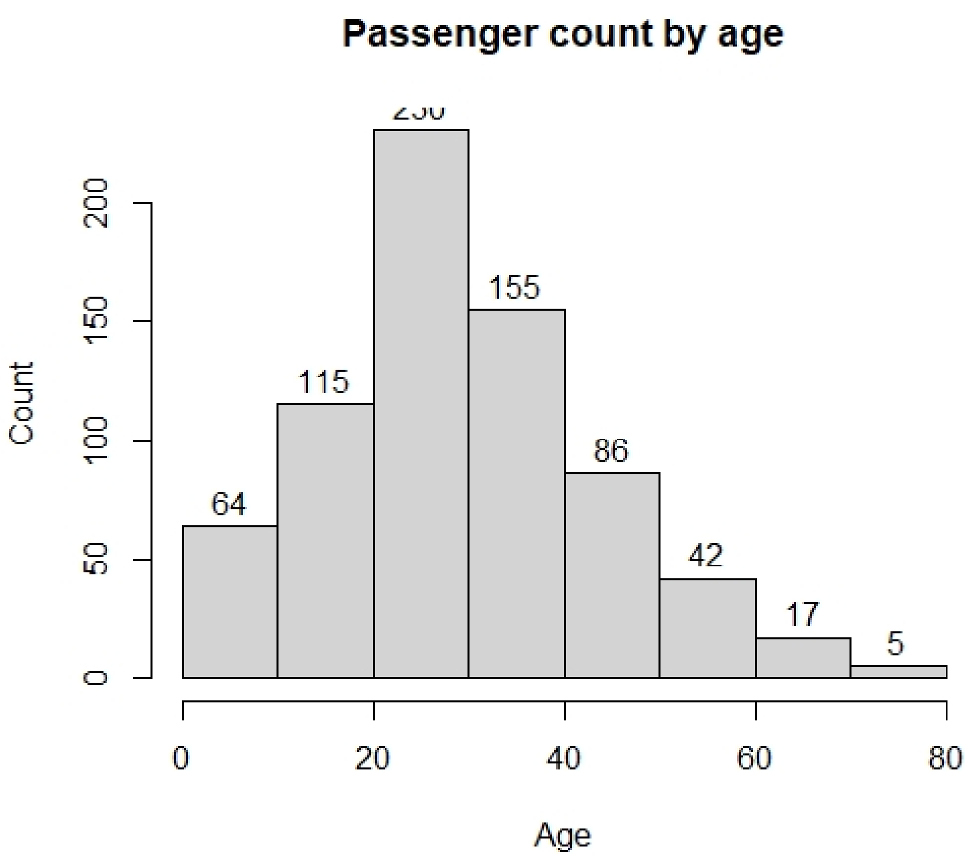
Important observations and conclusions -

1. Number of men and women
   - Female - 314
   - Male - 577 (Median)
2. Women boarding the ship were fewer and younger than men
3. Probability of a particular sex
   - Female - 0.3524
   - Male - 0.6476

2.2.4 Age

Age being a continuous data type will be visualised as a histogram[3], the histogram for 714 passengers along with the code is as follows-

```
x<-data$Age
x<- x[complete.cases(x)]
hist(x, breaks = 10, label = TRUE, xlab = "Age", ylab = "Count",
    main = "Passenger count by age on Titanic")
```



A few important points worth noting,

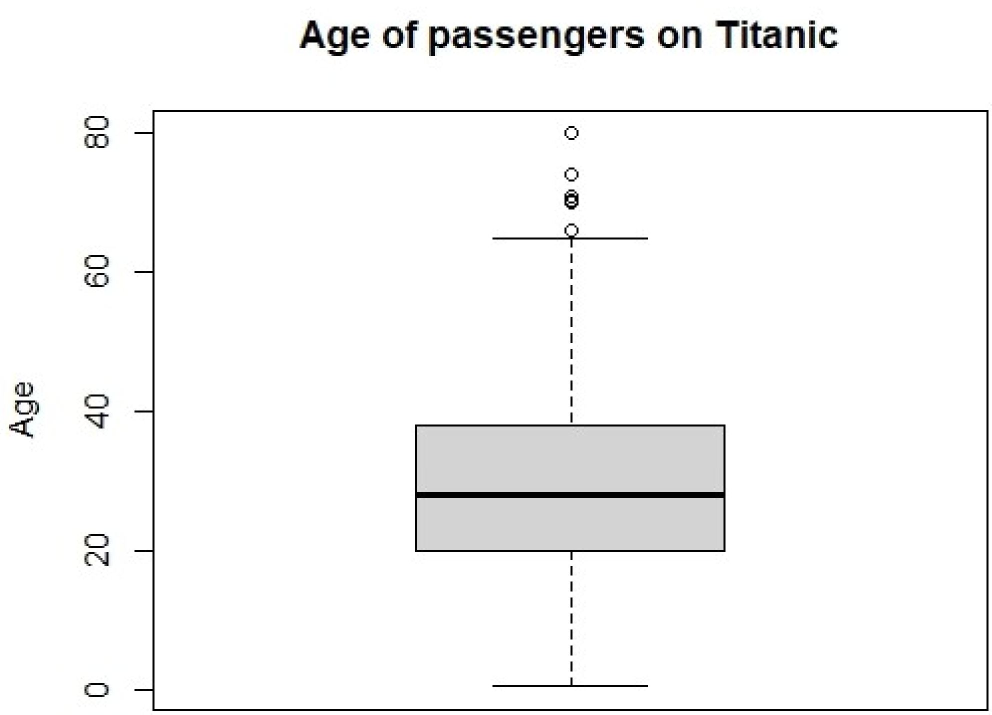
1. Total passenger's in intervals of 20:

[3]An accurate representation for estimating the probability distribution of a continuous variable. Type of a bar graph with continuous bars

- 0 - 20: 179
- 20 - 40: 385
- 40 - 60: 128
- 60 - 80: 22

2. Statistical interpretation of histogram

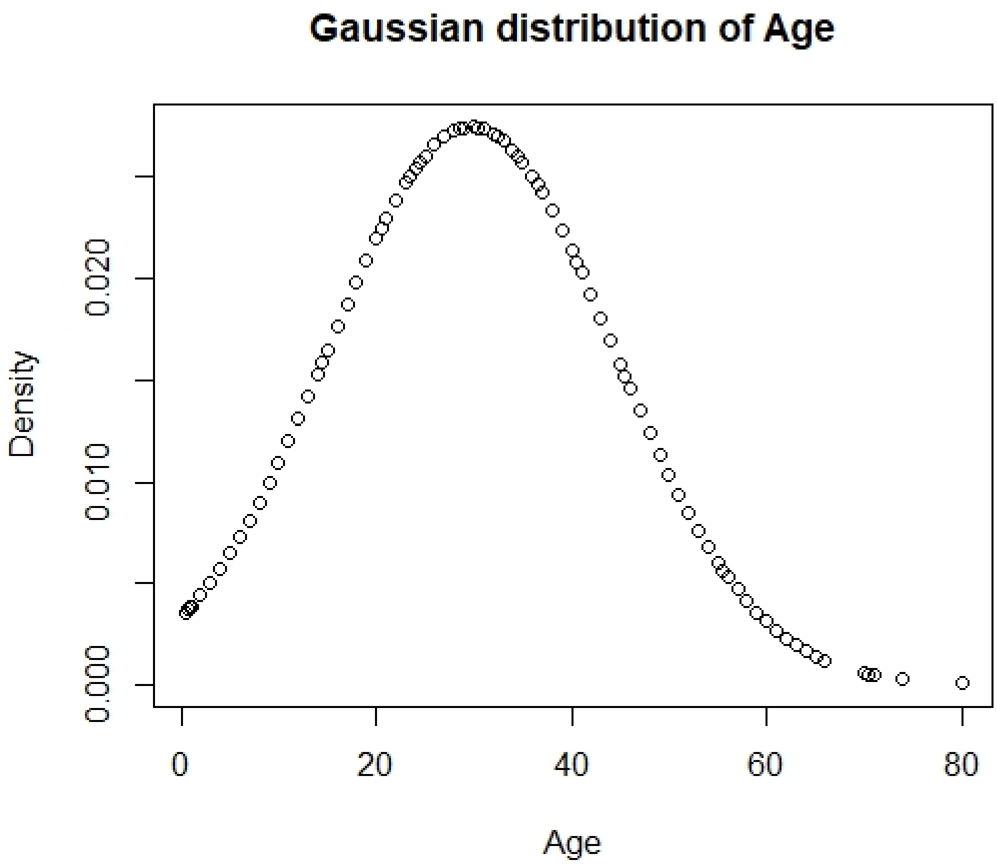
- Median class: 20 - 30, 230 passengers
- Mean age: 29.6991
- Median age: 28
- Max age: 80
- Min age: 0.42
- Variance: 211.0191
  Plotting a box plot for confirming observation 2.




```
boxplot(x, ylab = "Age", main = "Age of passengers
on Titanic")

```

3. Probability of passenger's age in intervals:
   - P(Age $\leq$ 20) = 0.2507
   - P(20 < Age $\leq$ 40) = 0.5392
   - P(40 < Age $\leq$ 60) = 0.1792
   - P(Age > 60) = 0.0302
4. Age can be represented as a Gaussian variable[4] with mean 29.6991 and standard deviation 14.5262




```
y<-dnorm(x, mean = mean(x, na.rm = TRUE), sd = sqrt(var(x,
 na.rm = TRUE))
```

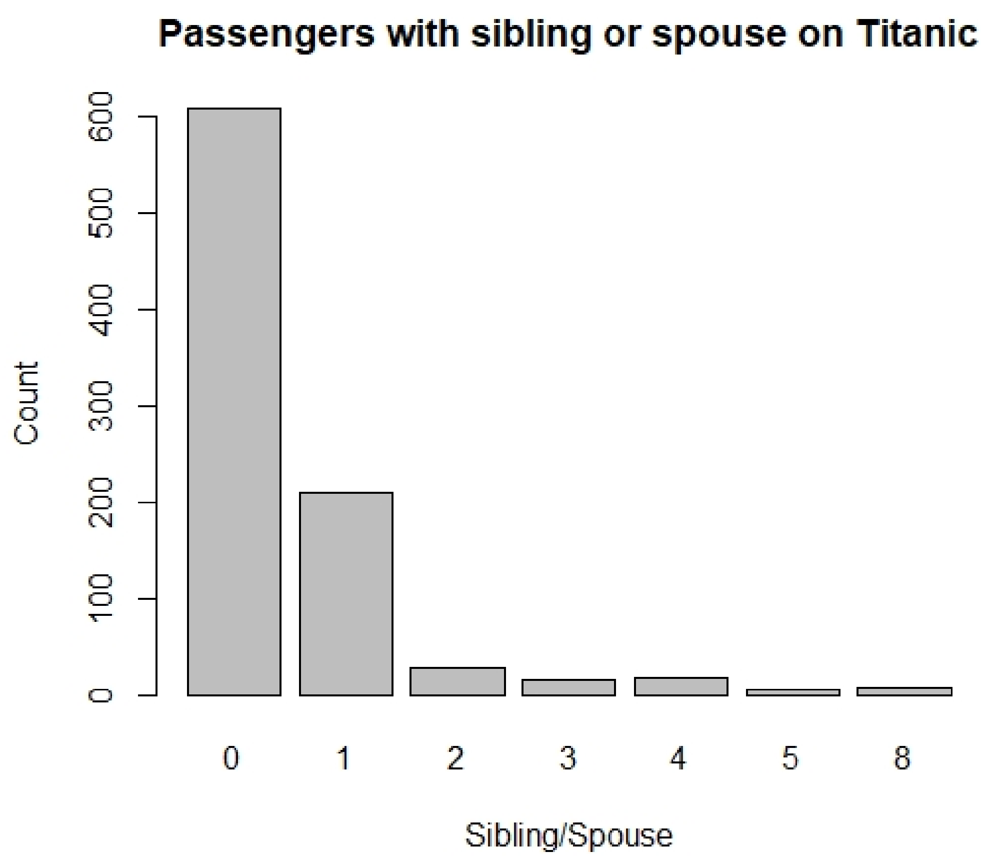
[4]A continuous variable whose pdf is of the form - f(x) = $1/\sqrt{2\pi} * e^{-(x-\mu)^2/2\sigma^2}$

```
plot(x, y, main = "Gaussian distribution of Age", xlab = "Age"
    , ylab = "Density")

```

2.2.5 SibSp

The number of passengers being accompanied by siblings or spouses will be determined by bar plot.




```
class.table = table(data$SibSp)
barplot(class.table, xlab = "Sibling/Spouse", ylab = "Count", main
    = "Passengers with sibling or spouse on Titanic")
```

Observations:

Companion	Count
0	608 (Median)
1	209
2	28
3	16
4	18
5	5
8	7

Table 3: Count of people with siblings/spouses

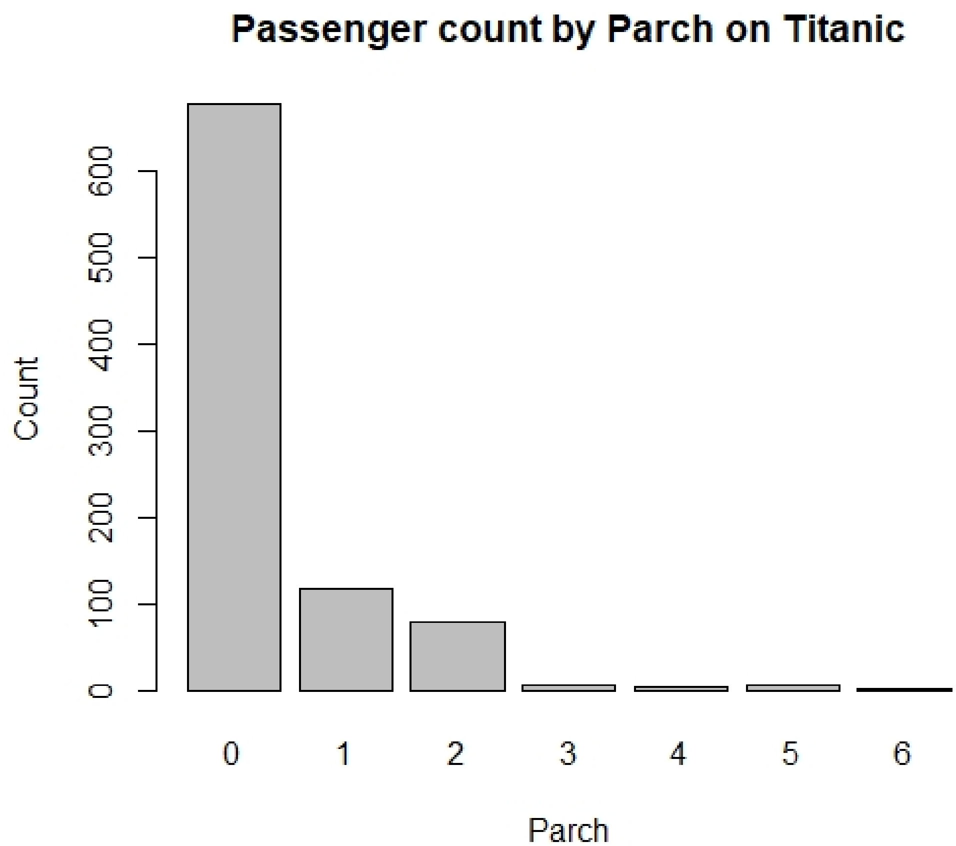
Probability of having companions

- P(SibSp = 0) = 0.6824
- P(SibSp > 0) = 0.3176

2.2.6 Parch

Like the above section where passengers with siblings or spouses was determined this section determines the number of parent-children.

```
class.table = table(data$Parch)
barplot(class.table, xlab = "Parch", ylab = "Count", main = "
    Passenger count by parch on Titanic")
```



Observations and conclusions:

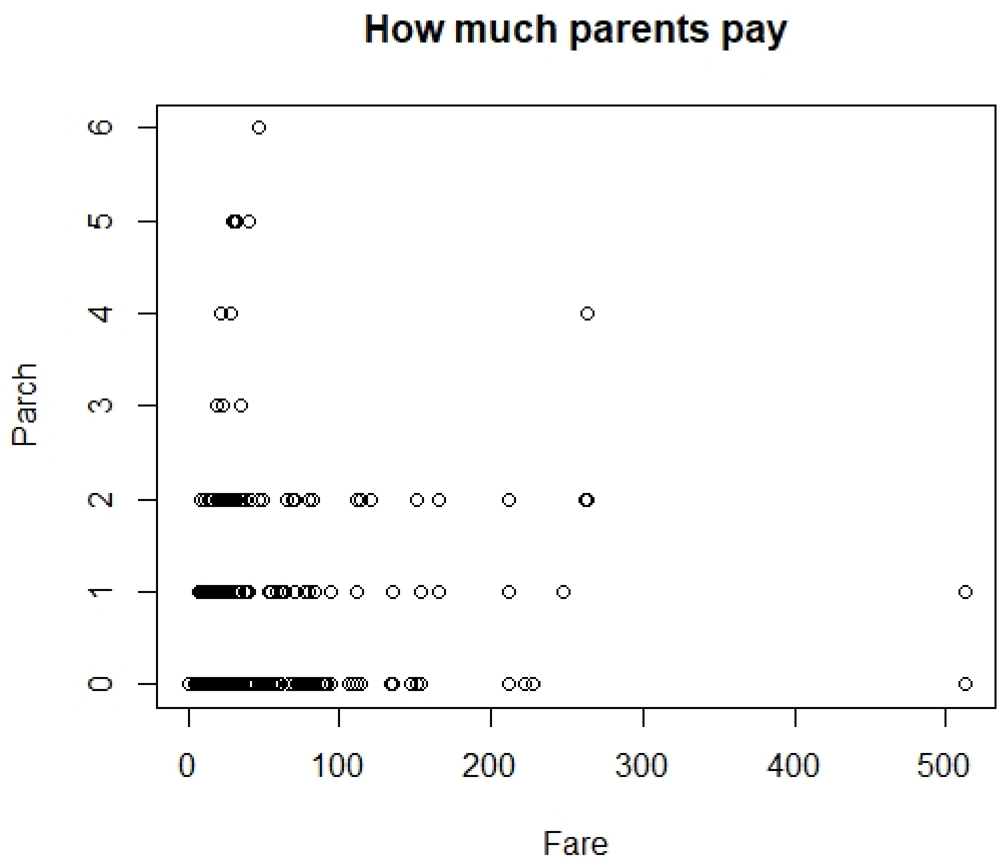
Parent and Child	Count
0	678 (Median)
1	118
2	80
3	5
4	4
5	5
6	1

Table 4: Count of parent and children

Probability of children accompanying passengers:

- P(Parch = 0) = 0.7609
- P(Parch > 0) = 0.2391

Going on a voyage with children means paying a lot for tickets. Proving/disproving this hypothesis by a scatter plot[5] and Pearson's correlation test[6]



Through the scatter plot it can be inferred parents pay less when they have 2 or more children with them
Code for scatter plot-

[5]A type of plot using Cartesian coordinates to display values for typically two variables
[6]correlation coefficients are used in statistics to determine how strong a relationship is between two variables

```
plot(data$Fare, data$Parch, xlab = "Fare", ylab = "Parch", main = "
    How much parents pay")
```

Proceeding with correlation test

```
cor.test(data$Parch, data$Fare)
```

Results:

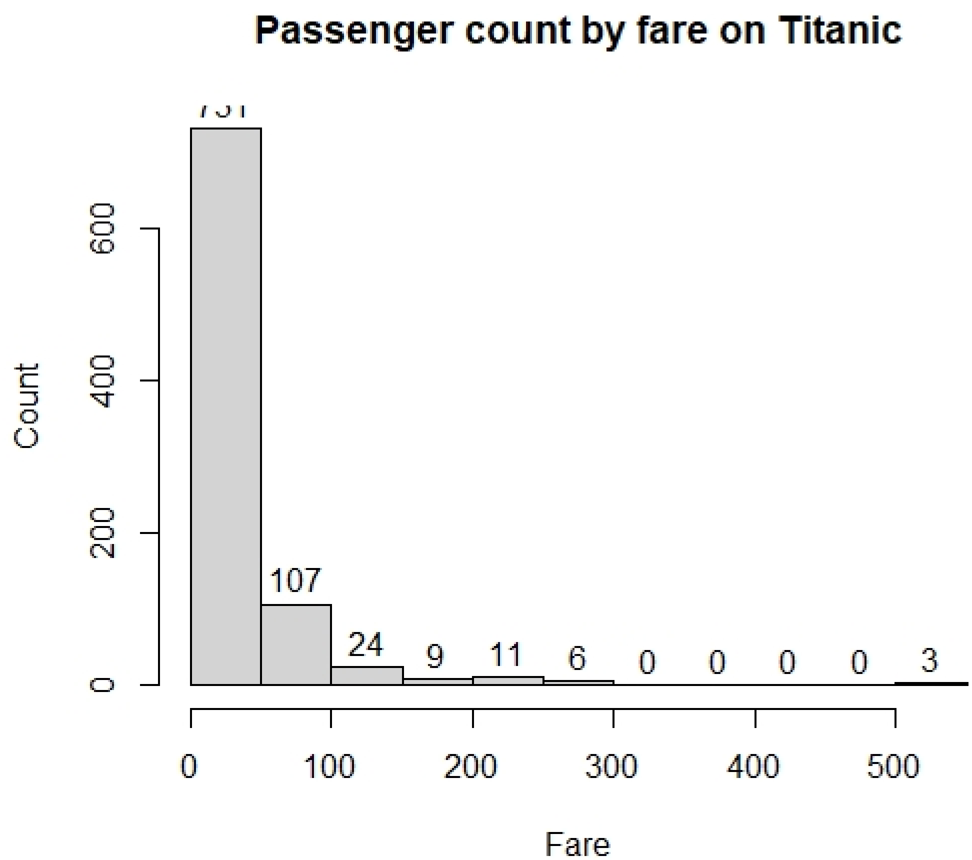
1. t = 6.6032
2. df = 889
3. p-value = 6.915e-11
4. Alternate hypothesis: true correlation is not zero
5. 95 percent confidence interval: 0.1527163 0.2779551
6. Sample estimates
    - cor 0.2162249

Conclusion: There is some correlation between fare and no. of children

2.2.7 Fare

The amount each passenger paid for their voyage is a continuous distribution and will be visualised with the help of a histogram.
The code for the histogram of 891 passengers is fairly simple, just two lines of basic R commands

```
f<-data$Fare
hist(f, xlab = "Fare", ylab = "Count", main = "Passenger count by
    fare on Titanic")
```



Observations and conclusions drawn are listed below as pointers

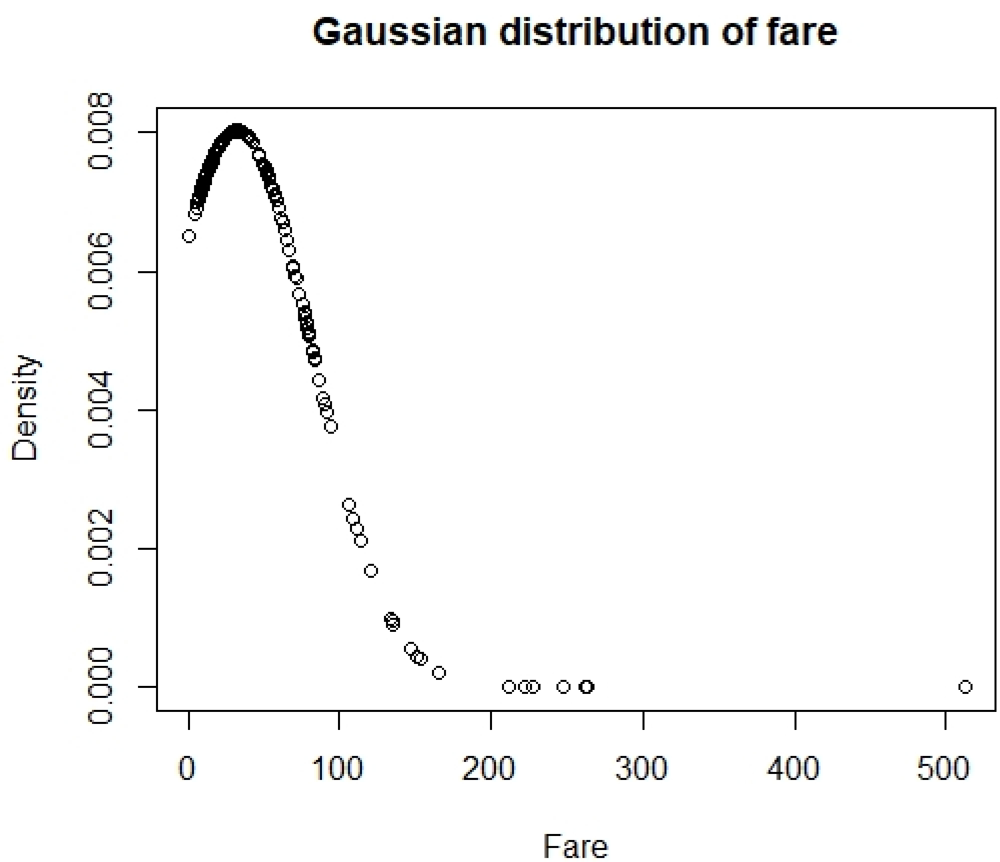
1. Statistical interpretation of histogram
   - Median class = 0 - 50, 731 entries
   - Mean fare = 32.2042
   - Median fare = 14.4542
   - Max fare = 512.3292
   - Variance = 2469.437
2. Total fare collected = 28693.95
3. Segregating the wealthy passengers from the affording passengers
   - P(fare $\leq$ 100) = 0.8194

- P(fare > 100) = 0.1706
  81% of passenger's came in affordable range

4. Fare can be represented as a Gaussian variable with mean 32.2042 and standard deviation 49.69343 using the following code

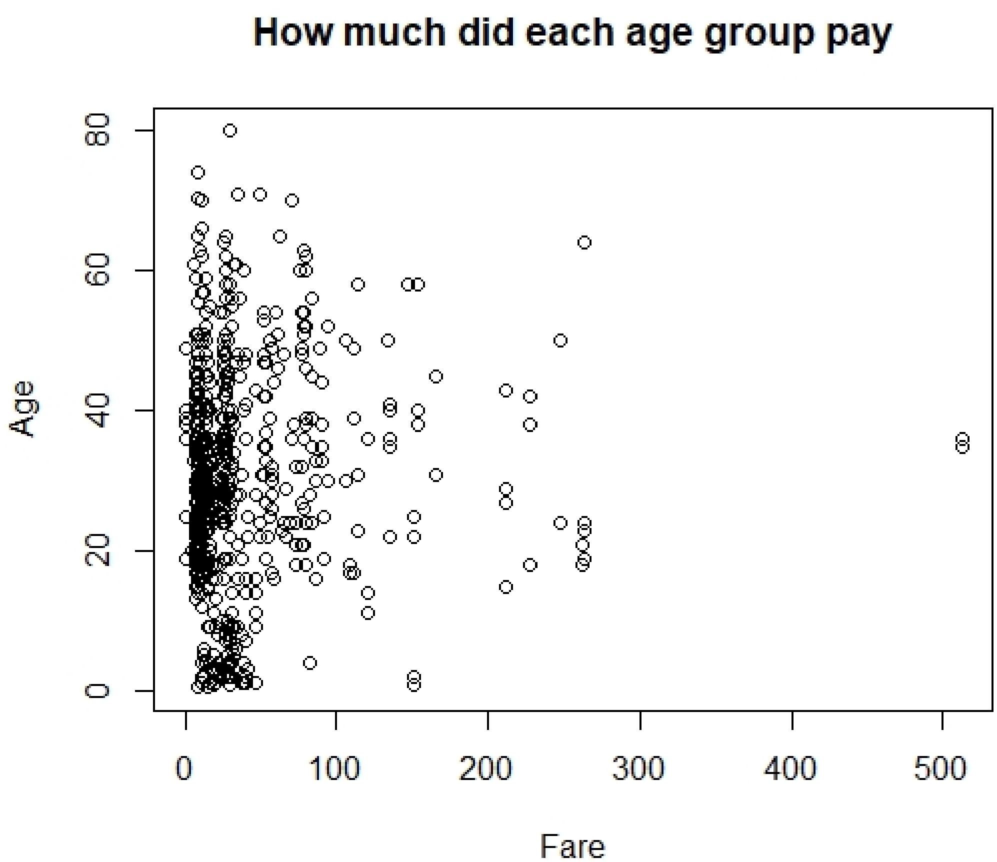
```
    y<-dnorm(f, mean = mean(f, na.rm = TRUE), sd = sqrt(var(f, na.rm = TRUE)))
plot(f, y, main = "Gaussian distribution of fare", xlab = "Fare", ylab = "Density")

```



A general rule of voyages is payment as per age. Pearson correlation test will be used to validate this rule, before validating plot a scatter plot to find conclusions visually, R code-

```r
plot(data$Fare, data$Age, xlab = "Fare", ylab = "Age", main = "How
    much did each age group pay")
```



Inference: It is difficult to find any correlation

Using Pearson correlation test to conclude the inference

```r
cor.test(train$Age, train$Fare, method = 'pearson')
```

Results of correlation test

1. t = 2.5753
2. df = 712
3. p-value = 0.01022

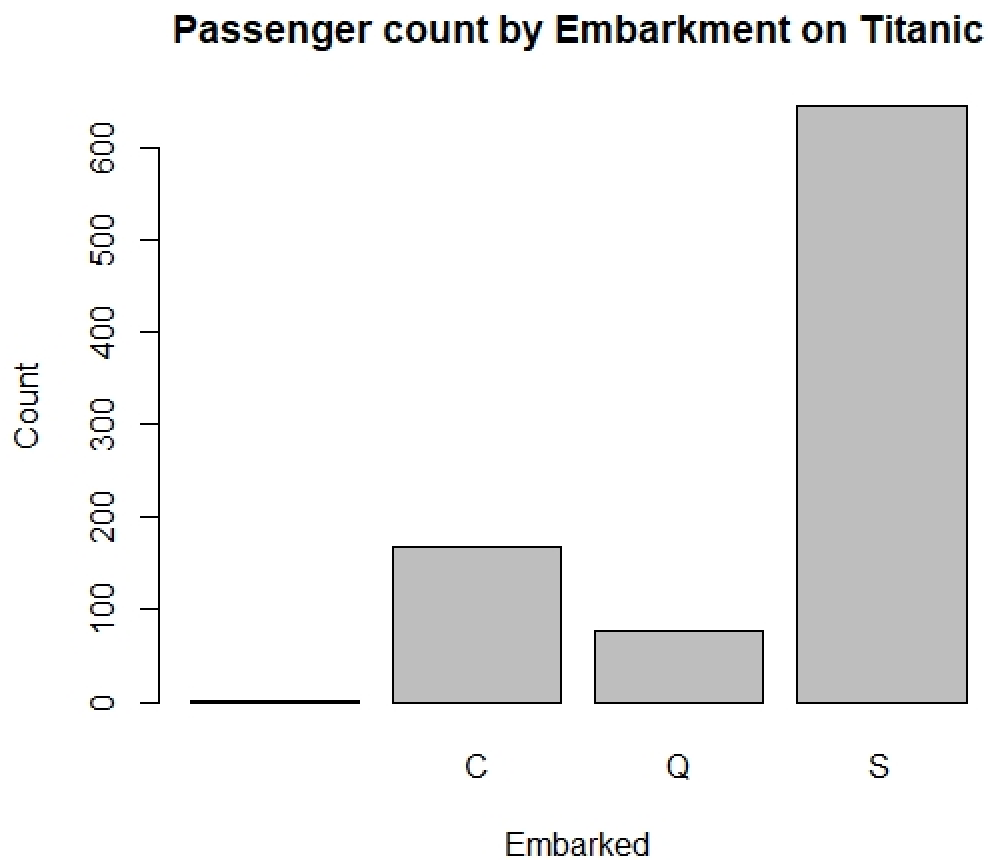
4. Alternate hypothesis: true correlation is not zero
5. 95 percent confidence interval: 0.02285549 0.16825304
6. Sample estimates
    - cor 0.09606669

Conclusion: There is very little correlation between fare and age, in fact just 9.6%
In section 2.2.2 conclusions were made about class1 having older passengers and higher fare. It was inferred that older passenger's could be paying more than younger passengers. However the correlation test disproves the inference.

2.2.8 Embarked

Finding the numeric values for 889 passengers boarding from an embarked port. The discrete data is represented as a bar plot, plot and code shown below.

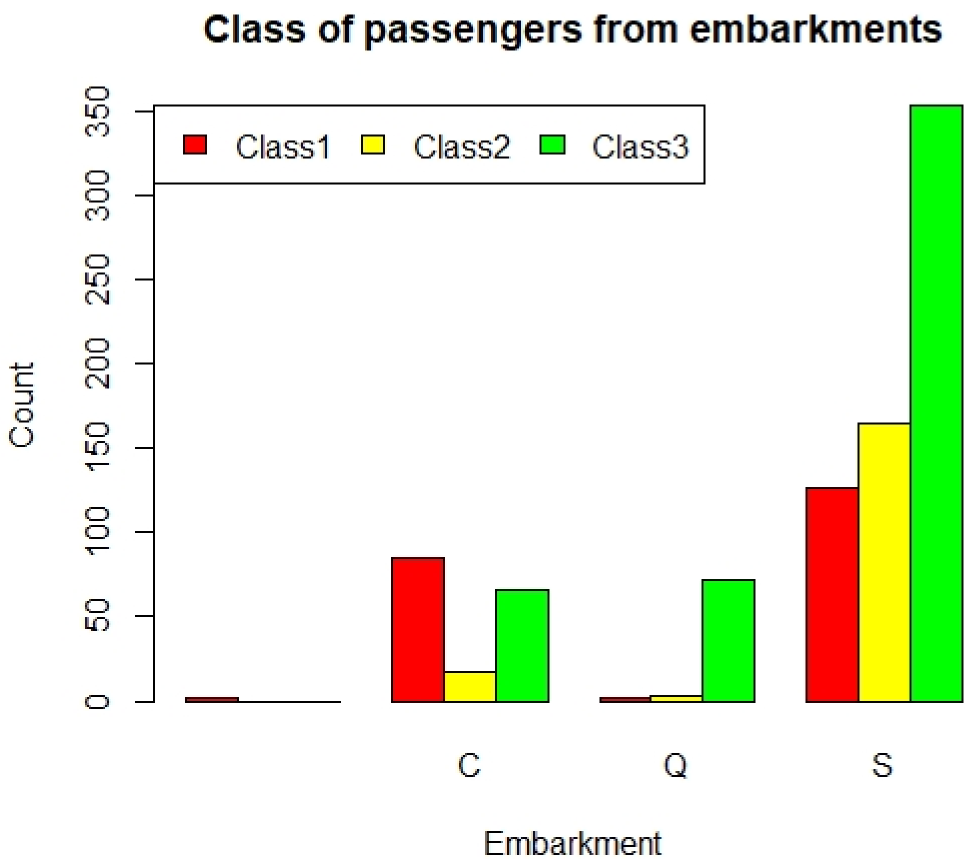
```
class.table = table(data$Embarked)
barplot(class.table, xlab = "Embarked", ylab = "Count", main = "
    Passenger count by embarkment on Titanic")
```



Observations and results -

1. Number of passenger's from embarked ports:
   - Cherbourg = 168
   - Queenstown = 77
   - Southampton = 644 (Median S)
2. Probabilities of boarded from a certain port
   - P(embarked = C) = 0.1890
   - P(embarked = Q) = 0.0866
   - P(embarked = S) = 0.7244

Finding a distribution of the passenger classes from port embarkations



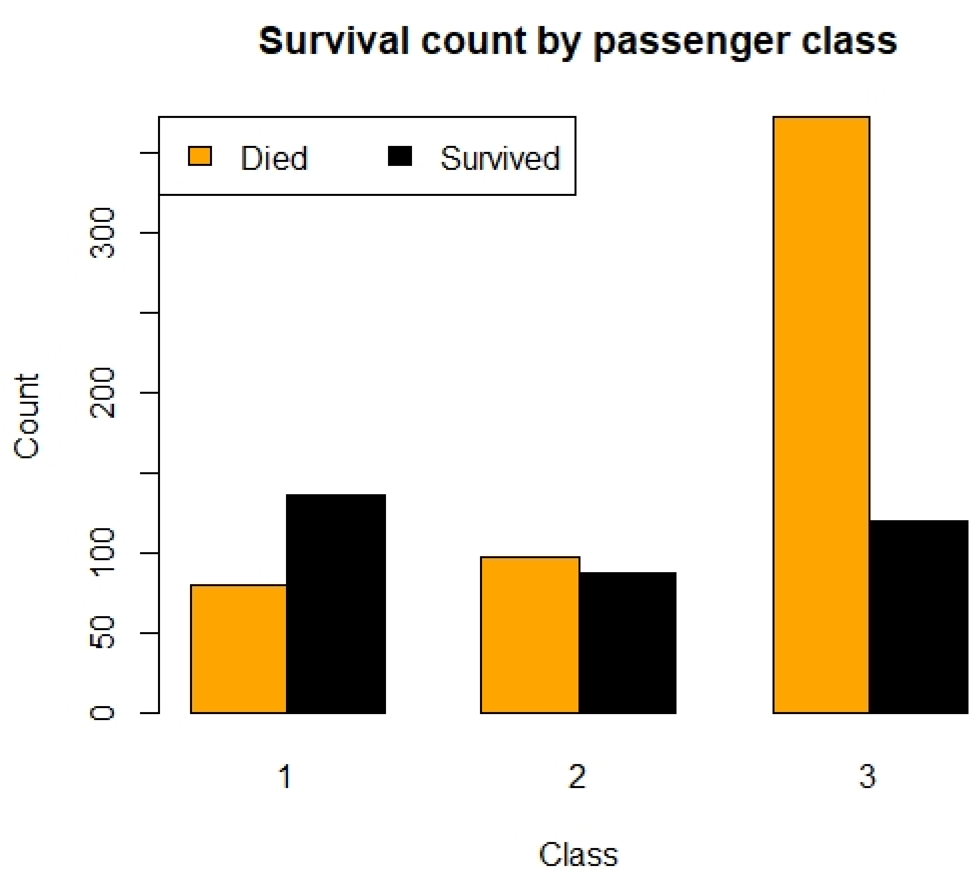

```
counts = table(data$Pclass, data$Embarked)
barplot(counts, xlab = "Embarkment", ylab = "Count", main = "Class
    of passengers from embarkments", col = c("Red", "Yellow", "
    Green"), beside = TRUE)
legend("topleft", legend = c("Class1", "Class2", "Class3"), fill=c(
    "Red", "Yellow", "Green"), horiz = TRUE)
```

3 Survival Analysis

Having found out the actual number of survivors and the details of passengers lets analyse the survivors from the columns
Pictorial representations will be used to determine the number of survivors and form conclusions. Certain hypothesis will be framed and statistical tests will be carried out to prove or disprove the hypothesis
'Probability of survival' section is about calculating the probability of survival using the concept of conditional probability

3.1 Pclass

Finding the class of tickets maximum survivors belonged to by means of a bar plot.



Use of R language in the analysis-

```
counts = table(data$Survived, data$Pclass)
barplot(counts, xlab = "Class", ylab = "Count", main= "Survival count by passenger class", col = c("Orange", "Black"), beside = TRUE)
legend("topleft", legend = c("Died", "Survived"), fill = c("Orange", "Black"), horiz = TRUE)
```

Observations:

Class	Died	Survived	D/S
1	80	136	0.5882
2	97	87	1.1149
3	372	119	3.1260

Table 5: Mathematical conclusion

Upper class passengers had a higher chance of survival. Proving by Z test[7].

- H^0: There is no significant difference in the chances of survival of upper and lower class
- H1: There is a better chance of survival for upper class passengers

```
data<-read.csv("E:/Jupyterfiles/ML_practice/Kaggle/Titanic/train.csv")
new_data<-subset(data, data$Pclass == 1)
z.test2 = function(a, b, n){
    sample_mean = mean(a)
    pop_mean = mean(b)
    c = nrow(n)
    var_b = var(b)
    zeta = (sample_mean - pop_mean) / (sqrt(var_b/c))
    return(zeta)
}
z.test2(new_data$Survived, data$Survived, new_data)
```

Result: z = 7.423828, a high z implies a low p value which affirms the observation of upper class having better chances of survival

3.1.1 Probability of survival

- 1^{st} class
  P(class = 1) = 0.2424 [from 2.2.2]
  P(survive | class1) = P(survive ∩ class1) / P(class1) —— (1)
  P(survive ∩ class1) = 136 / 891 = 0.1526 —— (2)
  Put (2) in (1)
  P(survive | class1) = 0.1526 / 0.2424 = 0.6295

[7] A statistical test to determine whether two population means are different when the variances are known and sample size is large. A z-score is a number representing the result from z-test

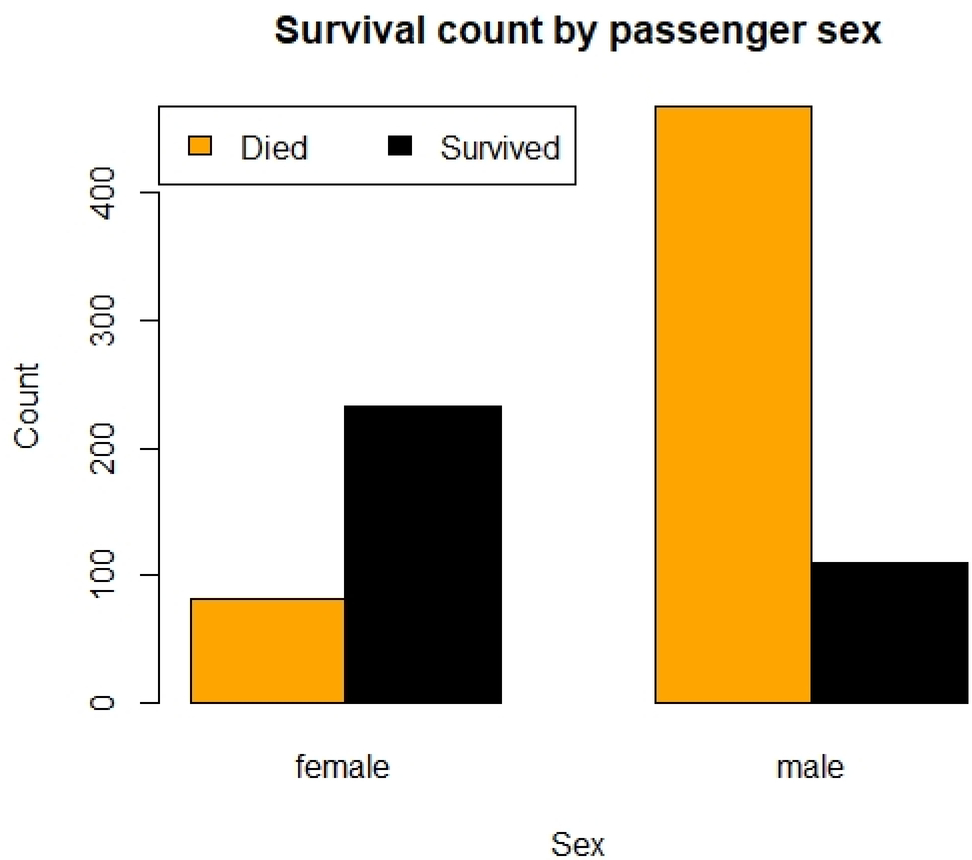
- 2^{nd} class
  P(class = 2) = 0.2065 [from 2.2.2]
  P(survive | class2) = P(survive ∩ class2) / P(class2) —– (1)
  P(survive ∩ class2) = 87 / 891 = 0.0976 —– (2)
  Put (2) in (1)
  P(survive | class2) = 0.0976 / 0.2065 = 0.4726

- 3^{rd} class
  P(class = 3) = 0.5511 [from 2.2.2]
  P(survive | class3) = P(survive ∩ class3) / P(class3) —– (1)
  P(survive ∩ class3) = 119 / 891 = 0.1336 —– (2)
  Put (2) in (1)
  P(survive | class3) = 0.1336 / 0.5511 = 0.2424

3.2 Sex

It is a well known fact at the time of evacuation women and children are given prime importance. To prove the shipwreck of the Titanic is no different lets visualise the results using a bar plot

```
counts = table(data$Survived, data$Sex)
barplot(counts, xlab = "Sex", ylab = "Count", main= "Survival count
    by passenger sex", col = c("Orange", "Black"), beside = TRUE)
legend("topleft", legend = c("Died", "Survived"), fill = c("Orange"
    , "Black"), horiz = TRUE)
```



Observations:

Sex	Died	Survived	D/S ratio
Female	81	233	0.3476
Male	468	109	4.2936

Table 6: Number of men and women saved

3.2.1 Probability of survival

- Female
  P(Female) = 0.3524 [from 2.2.3]
  P(Survive|Female) = P(Survive ∩ Female) / P(Female) —– (1)
  P(Survive ∩ Female) = 233 / 891 = 0.2615 —– (2)

Put (2) in (1)
P(Survive|Female) = 0.2615 / 0.3524 = 0.7420

- Male
  P(Male) = 0.6476 [from 2.2.3]
  P(Survive|Male) = P(Survive ∩ Male) / P(Male) —– (1)
  P(Survive ∩ Male) = 109 / 891 = 0.1223 —– (2)
  Put (2) in (1)
  P(Survive|Male) = 0.1223 / 0.6476 = 0.1888

3.3 Age

At the time of shipwreck the captain's objective was to rescue as many children as possible. Consider a hypothesis H: The Titanic survivors were younger than the passengers that died.
Using the two-group t-test[8] on the data set of survivors to evaluate the hypothesis. Assumption: the two groups - survived and age are independent of each other and data is sampled from normal populations.

```
t.test(data$Age~data$Survived)
```

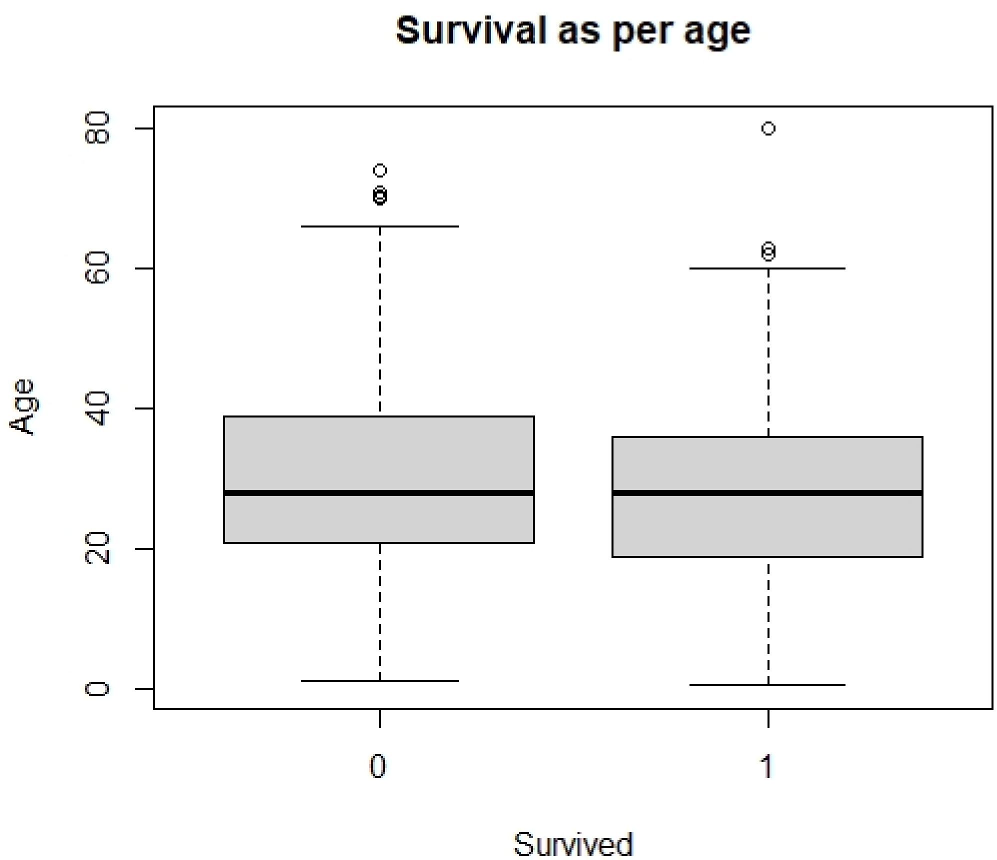
Result of t-test:

1. t = 2.046
2. df = 598.84
3. p-value = 0.04119
4. Alternate hypothesis: true difference in means is not equal to 0
5. 95 percent confidence interval: 0.09158472 4.4733946
6. Sample estimates:
   - Mean in group 0: 30.62618
   - Mean in group 1: 28.34369

Conclusion: $p > 0.001$ hence null hypothesis is rejected. There is a significant difference in the average age of survivors and non-survivors. This proves H to be true.
Visualising survival by age by means of a box plot

```
boxplot(data$Age~data$Survived, xlab = "Survived", ylab = "Age",
    main = "Survival as per age")
```

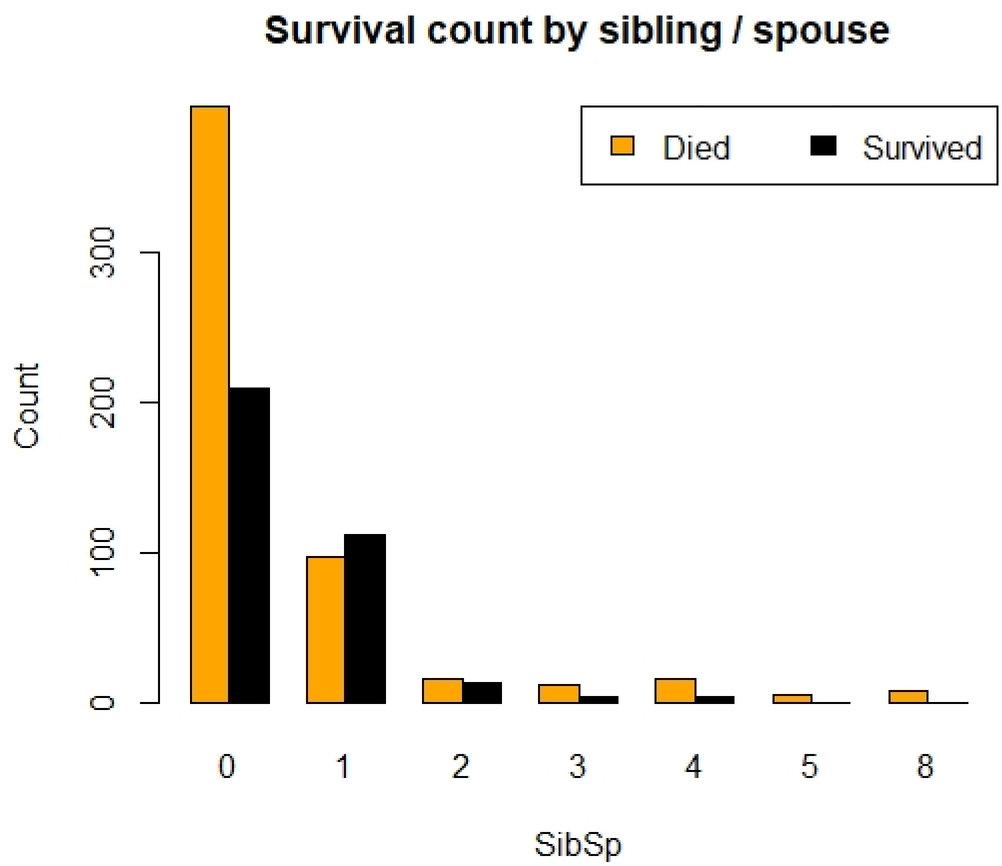
[8] Used as a hypothesis testing tool which allows testing of an assumption applicable to a population



3.4 SibSp

It had been concluded earlier that maximum passengers were without companions. Now lets check how many companions were saved in the evacuation process by means of a bar plot. The code used for plotting -

```
counts = table(data$Survived, data$SibSp)
counts
barplot(counts, xlab = "SibSp", ylab = "Count", main = "Survival
    count by sibling / spouse", col = c("Orange", "Black"), beside
    = TRUE)
legend("topright", legend = c("Died", "Survived"), fill = c("Orange
    ", "Black"), horiz = TRUE)
```



Observations

Companions	Died	Survived	D/S ratio
0	398	210	1.895
1	97	112	0.8661
2	15	13	1.1538
3	12	4	3
4	15	3	5
5	5	0	-
8	7	0	-

Table 7: Number of companions saved

1. 210 passengers without companions out of 608 survived (ratio = 0.3454)

2. 132 passengers with companions out of 283 survived (ratio = 0.4664)

3. During evacuation emphasis was given to couples
   Performing Z test

```
data<-read.csv("E:/Jupyterfiles/ML_practice/Kaggle/Titanic/train.csv")
new_data<-subset(data, data$SibSp == 0)
z.test2 = function(a, b, n){
    sample_mean = mean(a)
    pop_mean = mean(b)
    c = nrow(n)
    var_b = var(b)
    zeta = (sample_mean - pop_mean) / (sqrt(var_b/c))
    return(zeta)
}
z.test2(new_data$SibSp, data$Survived, new_data)

```

   Result: z = -19.45068, a low z implies a high p value which affirms the observation of 'Couples swim together'

3.4.1 Probability of survival

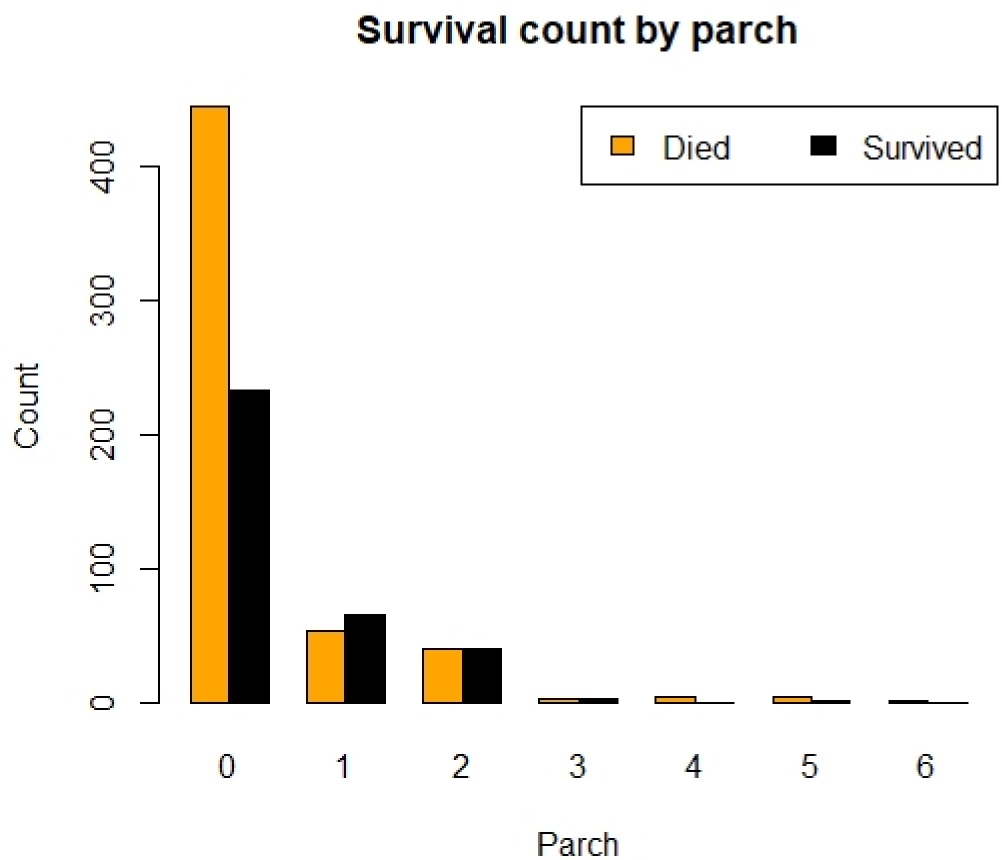
- Number of companions = 0
  P(SibSp = 0) = 0.6824 [from 2.2.5]
  P(Survive|SibSp = 0) = P(Survive ∩ SibSp) / P(SibSp = 0) —– (1)
  P(Survive ∩ SibSp) = 210 / 891 = 0.2357 —–(2)
  Put (2) in (1)
  P(Survive|SibSp = 0) = 0.2357 / 0.6824 = 0.3454

- Number of companions > 0
  P(SibSp > 0) = 0.3176 [from 2.2.5]
  P(Survive|SibSp > 0) = P(Survive ∩ SibSp) / P(SibSp > 0) —–(1)
  P(Survive ∩ SibSp) = 132 / 891 = 0.1481 —– (2)
  Put (2) in (1)
  P(Survive|SibSp > 0) = 0.1481 / 0.3176 = 0.4663

3.5 Parch

After finding the number of companions saved it is time to find the number of parent-children pair saved. To do this bar plot will be used. Code for the bar plot is as follows:

```
counts = table(data$Survived, data$Parch)
barplots(counts, xlab = "Parch", ylab = "Count", main = "Survival by parch", col = c("Orange", "Black"), beside = TRUE)
legend("topright", legend = c("Died", "Survived"), fill = c("Orange", "Black"), horiz = TRUE)
```



Observations:

Parent and child	Died	Survived	D/S ratio
0	445	233	1.9099
1	53	65	0.8154
2	40	40	1
3	2	3	0.6667
4	4	0	-
5	4	1	4
6	1	0	-

Table 8: Parent-child survival

1. For parent-child pair val: 0, 233 survived out of 678 (ratio = 0.3437)

2. For parent-child pair val > 0, 107 survived out of 213 (ratio = 0.5023)

3. During evacuation emphasis was given to parents and children

3.5.1 Probability of survival

- P(Parch = 0) = 0.7609 [from 2.2.6]
  P(Survive|Parch = 0) = P(Survive ∩ Parch) / P(Parch = 0) —– (1)
  P(Survive ∩ Parch) = 233 / 891 = 0.2615 —– (2)
  P(Survive|Parch = 0) = 0.2615 / 0.7609 = 0.3437

- P(Parch > 0) = 0.2391 [from 2.2.6] P(Survive|Parch > 0) = P(Survive ∩ Parch) / P(Parch > 0) —– (1)
  P(Survive ∩ Parch) = 109 / 891 = 0.1223 —– (2)
  P(Survive|Parch > 0) = 0.1223 / 0.2391 = 0.5115

3.6 Fare

Consider a hypothesis, H: Passengers paying less for their voyage had more chances of survival.
T test will be used to evaluate the above hypothesis.

```
t.test(data$fare~data$Survived)
```

Result of t-test:

1. t = -6.8391

2. df = 436.7

3. p-value = 2.669e-11

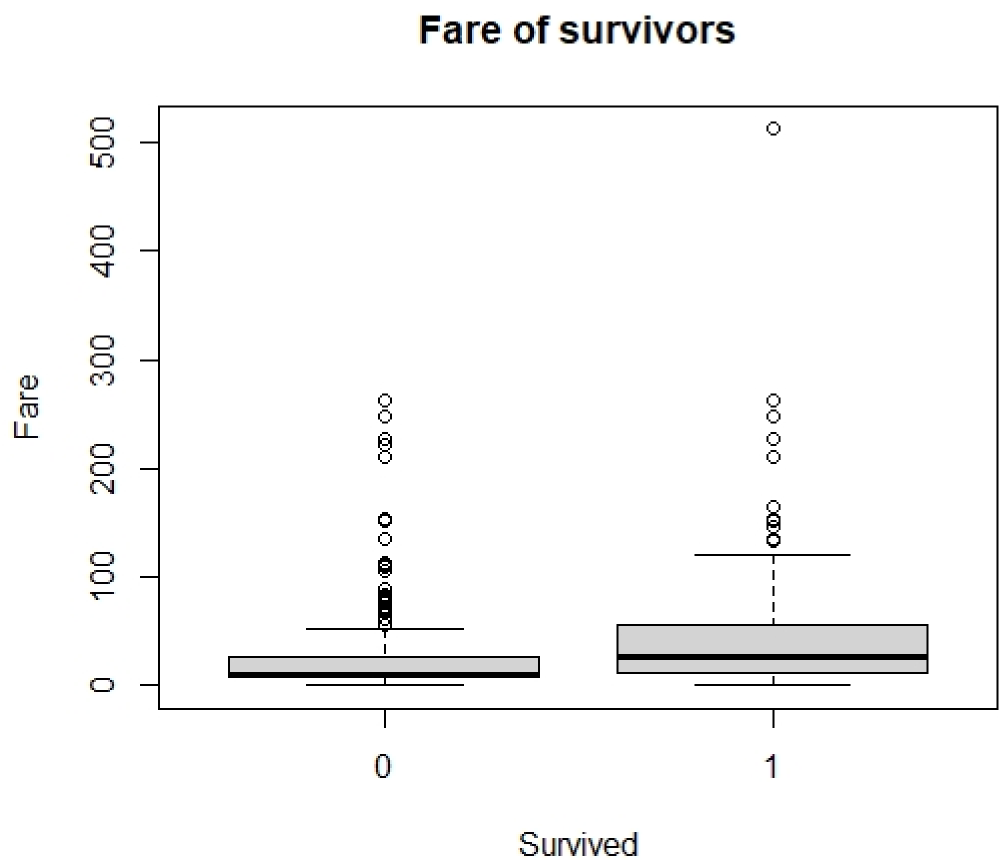
4. Alternate hypothesis: true difference in means is not equal to 0

5. 95 percent confidence interval: -33.82912 -18.72592

6. Sample estimates:

   - mean in group 0 = 22.11789
   - mean in group 1 = 48.39541

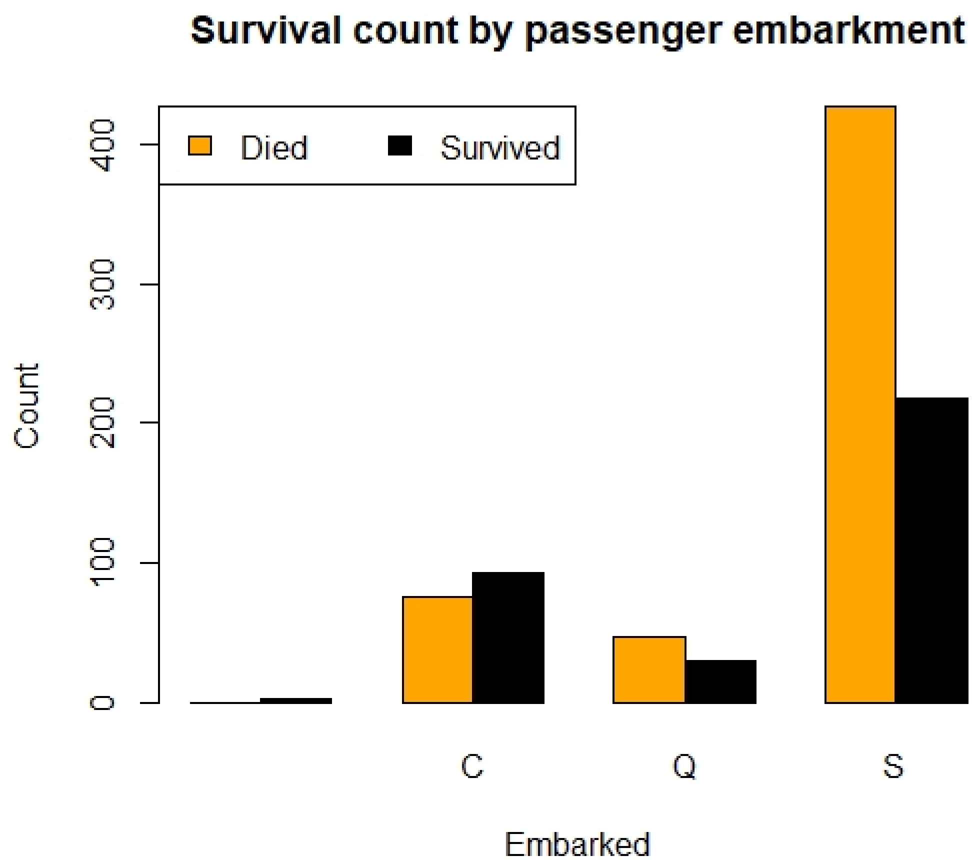
Conclusion: $p < 0.001$ hence null hypothesis is accepted. This proves H to be false.
Visualising survival by fare by means of a box plot



3.7 Embarked

In this portion the number of survivors as per the port of embarkation will be obtained. A bar plot will be used to find the lucky ports. The code to obtain the lucky port-

```
counts = table(data$Survived, data$Embarked)
barplot(counts, xlab = "Embarked", ylab = "Count", main= "Survival
    count by passenger embarkment", col = c("Orange", "Black"),
    beside = TRUE)
legend("topleft", legend = c("Died", "Survived"), fill = c("Orange"
    , "Black"), horiz = TRUE)
```



Observations:

Port	Died	Survived	D/S ratio
C	75	93	0.8065
Q	47	30	1.5667
S	427	217	1.9677

Table 9: Survivors per port

3.7.1 Probability of survival

- P(Embark = C) = 0.1890 [from 2.2.8]
  P(Survive|Embark = C) = P(Survive ∩ Embark) / P(Embark = C)——(1)
  P(Survive ∩ Embark) = 93 / 889 = 0.1046 —— (2)

Put (2) in (1)
P(Survive|Embark = C) = 0.1046 / 0.1890 = 0.5534

- P(Embark = Q) = 0.0866 [from 2.2.8]
  P(Survive|Embark = Q) = P(Survive ∩ Embark) / P(Embark = Q)——(1)
  P(Survive ∩ Embark) = 30 / 889 = 0.0337 —— (2)
  Put (2) in (1)
  P(Survive|Embark = Q) = 0.0337 / 0.0866 = 0.3891

- P(Embark = S) = 0.7244 [from 2.2.8]
  P(Survive|Embark = S) = P(Survive ∩ Embark) / P(Embark = S)——(1)
  P(Survive ∩ Embark) = 217 / 889 = 0.2441 —— (2)
  Put (2) in (1)
  P(Survive|Embark = S) = 0.2441 / 0.7244 = 0.3370

Passengers having Cherbourg as port of embarkation were lucky!

4 Conclusion

1. The graphical representations used in this report - bar plots, box plots, histograms and Gaussian distributions
   - For discrete data - bar plot
   - For continuous data - histogram
   - Box plots are used when median, max and min values are available for continuous data types
   - Gaussian variables were plotted as Gaussian distributions
2. Tests used to prove hypothesis
   - Pearson correlation test
   - T test
   - Z test
3. Data set analysis
   - 891 entries
   - The summary of column names, their data types, missing values and data classification -

Column	Data Type	Missing values	Classification	Median
Survived	int	No	Discrete	0
Pclass	int	No	Discrete	3
Name	char	No	Discrete	-
Sex	char	No	Discrete	male
Age	double	Yes	Continuous	28
SibSp	int	No	Discrete	0
Parch	int	No	Discrete	0
Ticket	char	No	Discrete	-
Fare	double	No	Continuous	14.4542
Cabin	char	Yes	Discrete	-
Embarked	char	Yes	Discrete	S

Table 10: Summary of all column entries

4. Probabilistic summarising of passengers aboard -

Description	Probability
Class 1, 2, 3	0.2424, 0.2065, 0.5511
Female	0.3524
Age $\leq$ 20	0.2507
20 $<$ Age $\leq$ 40	0.5392
40 $<$ Age $\leq$ 60	0.1792
SibSp = 0	0.6824
Parch = 0	0.7609
Fare $\leq$ 100	0.8194
Emabarked = C, Q, S	0.1890, 0.0866, 0.7244

Table 11: Passenger probabilities

5. Survival analysis
   - Probability of surviving = 0.38
     - Total survivors = 338
   - Survivors from each column -

Column	Max survivors belong to
Pclass	class1
Sex	Female
SibSp	1
Parch	3
Embarked	C

Table 12: Maximum survivors from columns

   - Probabilistic summarising of survivors -

Attribute	Probability
Class1	0.6295
Class2	0.4726
Class3	0.2424
Female	0.7420
Male	0.1888
SibSp = 0	0.3454
SibSp $>$ 0	0.4663
Parch = 0	0.3437
Parch $>$ 0	0.5115
Embarked = C	0.5534
Embarked = Q	0.3891
Embarked = S	0.3370

Table 13: Probabilities of survival