Titanic_data_analysis.R

STEVE

2025-01-02

```
### set the working directory:

setwd("C:/Users/Steve/Desktop/Business_analyst/R  project")


### load the required packages for the analysis
library(tidyverse)
library(readxl)


### import the dataset int the R environment
titanic_data  <-  read.csv("titanic.csv")
titanic_data
```

```
##     PassengerId Survived Pclass
## 1             1        0      3
## 2             2        1      1
## 3             3        1      3
## 4             4        1      1
## 5             5        0      3
## 6             6        0      3
## 7             7        0      1
## 8             8        0      3
## 9             9        1      3
## 10           10        1      2
## 11           11        1      3
## 12           12        1      1
## 13           13        0      3
## 14           14        0      3
## 15           15        0      3
## 16           16        1      2
## 17           17        0      3
## 18           18        1      2
## 19           19        0      3
## 20           20        1      3
## 21           21        0      2
## 22           22        1      2
## 23           23        1      3
## 24           24        1      1
## 25           25        0      3
## 26           26        1      3
## 27           27        0      3
## 28           28        0      1
```

## 29	29	1	3
## 30	30	0	3
## 31	31	0	1
## 32	32	1	1
## 33	33	1	3
## 34	34	0	2
## 35	35	0	1
## 36	36	0	1
## 37	37	1	3
## 38	38	0	3
## 39	39	0	3
## 40	40	1	3
## 41	41	0	3
## 42	42	0	2
## 43	43	0	3
## 44	44	1	2
## 45	45	1	3
## 46	46	0	3
## 47	47	0	3
## 48	48	1	3
## 49	49	0	3
## 50	50	0	3
## 51	51	0	3
## 52	52	0	3
## 53	53	1	1
## 54	54	1	2
## 55	55	0	1
## 56	56	1	1
## 57	57	1	2
## 58	58	0	3
## 59	59	1	2
## 60	60	0	3
## 61	61	0	3
## 62	62	1	1
## 63	63	0	1
## 64	64	0	3
## 65	65	0	1
## 66	66	1	3
## 67	67	1	2
## 68	68	0	3
## 69	69	1	3
## 70	70	0	3
## 71	71	0	2
## 72	72	0	3
## 73	73	0	2
## 74	74	0	3
## 75	75	1	3
## 76	76	0	3
## 77	77	0	3
## 78	78	0	3
## 79	79	1	2
## 80	80	1	3
## 81	81	0	3
## 82	82	1	3

```
## 83                83             1          3
##                                                                 Name
## 1                                          Braund, Mr. Owen Harris
## 2          Cumings, Mrs. John Bradley (Florence Briggs Thayer)
## 3                                            Heikkinen, Miss. Laina
## 4                    Futrelle, Mrs. Jacques Heath (Lily May Peel)
## 5                                          Allen, Mr. William Henry
## 6                                                  Moran, Mr. James
## 7                                           McCarthy, Mr. Timothy J
## 8                                    Palsson, Master. Gosta Leonard
## 9             Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)
## 10                              Nasser, Mrs. Nicholas (Adele Achem)
## 11                                  Sandstrom, Miss. Marguerite Rut
## 12                                         Bonnell, Miss. Elizabeth
## 13                                   Saundercock, Mr. William Henry
## 14                                      Andersson, Mr. Anders Johan
## 15                             Vestrom, Miss. Hulda Amanda Adolfina
## 16                                 Hewlett, Mrs. (Mary D Kingcome)
## 17                                             Rice, Master. Eugene
## 18                                     Williams, Mr. Charles Eugene
## 19   Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)
## 20                                         Masselmani, Mrs. Fatima
## 21                                             Fynney, Mr. Joseph J
## 22                                            Beesley, Mr. Lawrence
## 23                                      McGowan, Miss. Anna "Annie"
## 24                                     Sloper, Mr. William Thompson
## 25                                    Palsson, Miss. Torborg Danira
## 26 Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)
## 27                                          Emir, Mr. Farred Chehab
## 28                                     Fortune, Mr. Charles Alexander
## 29                                     O'Dwyer, Miss. Ellen "Nellie"
## 30                                              Todoroff, Mr. Lalio
## 31                                        Uruchurtu, Don. Manuel E
## 32              Spencer, Mrs. William Augustus (Marie Eugenie)
## 33                                          Glynn, Miss. Mary Agatha
## 34                                           Wheadon, Mr. Edward H
## 35                                           Meyer, Mr. Edgar Joseph
## 36                                    Holverson, Mr. Alexander Oskar
## 37                                                Mamee, Mr. Hanna
## 38                                          Cann, Mr. Ernest Charles
## 39                                  Vander Planke, Miss. Augusta Maria
## 40                                       Nicola-Yarred, Miss. Jamila
## 41               Ahlin, Mrs. Johan (Johanna Persdotter Larsson)
## 42  Turpin, Mrs. William John Robert (Dorothy Ann Wonnacott)
## 43                                              Kraeff, Mr. Theodor
## 44                          Laroche, Miss. Simonne Marie Anne Andree
## 45                                    Devaney, Miss. Margaret Delia
## 46                                          Rogers, Mr. William John
## 47                                                Lennon, Mr. Denis
## 48                                         O'Driscoll, Miss. Bridget
## 49                                              Samaan, Mr. Youssef
## 50                  Arnold-Franchi, Mrs. Josef (Josefine Franchi)
## 51                                      Panula, Master. Juha Niilo
## 52                                     Nosworthy, Mr. Richard Cater
```

```
##                                                  Name    Sex   Age SibSp Parch
## 1                             Braund, Mr. Owen Harris   male 22.00     1     0
## 2   Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38.00     1     0
## 3                              Heikkinen, Miss. Laina female 26.00     0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.00     1     0
## 5                            Allen, Mr. William Henry   male 35.00     0     0
## 6                                    Moran, Mr. James   male    NA     0     0
## 7                             McCarthy, Mr. Timothy J   male 54.00     0     0
## 8                      Palsson, Master. Gosta Leonard   male  2.00     3     1
## 9   Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) female 27.00     0     2
## 10                 Nasser, Mrs. Nicholas (Adele Achem) female 14.00     1     0
## 11                Sandstrom, Miss. Marguerite Rut female  4.00     1     1
## 12                       Bonnell, Miss. Elizabeth female 58.00     0     0
## 13                 Saundercock, Mr. William Henry   male 20.00     0     0
## 14                    Andersson, Mr. Anders Johan   male 39.00     1     5
## 15           Vestrom, Miss. Hulda Amanda Adolfina female 14.00     0     0
## 16               Hewlett, Mrs. (Mary D Kingcome) female 55.00     0     0
## 17                   Rice, Master. Eugene   male  2.00     4     1
## 18           Williams, Mr. Charles Eugene   male    NA     0     0
## 19  Vander Planke, Mrs. Julius (Emelia Maria Vande) female 31.00     1     0
## 20                        Masselmani, Mrs. Fatima female    NA     0     0
## 21                       Fynney, Mr. Joseph J   male 35.00     0     0
## 22                      Beesley, Mr. Lawrence   male 34.00     0     0
##                                                  Name    Sex   Age SibSp Parch
## 53                     Harper, Mrs. Henry Sleeper (Myna Haxtun) female 49.00     1     0
## 54   Faunthorpe, Mrs. Lizzie (Elizabeth Anne Wilkinson) female 29.00     1     0
## 55                         Ostby, Mr. Engelhart Cornelius   male 65.00     0     1
## 56                                Woolner, Mr. Hugh   male    NA     0     0
## 57                             Rugg, Miss. Emily female 21.00     0     0
## 58                           Novel, Mr. Mansouer   male 28.50     0     0
## 59                   West, Miss. Constance Mirium female  5.00     1     2
## 60                Goodwin, Master. William Frederick   male 11.00     5     2
## 61                             Sirayanian, Mr. Orsen   male 22.00     0     0
## 62                               Icard, Miss. Amelie female 38.00     0     0
## 63                     Harris, Mr. Henry Birkhardt   male 45.00     1     0
## 64                              Skoog, Master. Harald   male  4.00     3     2
## 65                             Stewart, Mr. Albert A   male    NA     0     0
## 66                          Moubarek, Master. Gerios   male    NA     1     1
## 67                  Nye, Mrs. (Elizabeth Ramell) female 29.00     0     0
## 68                        Crease, Mr. Ernest James   male 19.00     0     0
## 69           Andersson, Miss. Erna Alexandra female 17.00     4     2
## 70                               Kink, Mr. Vincenz   male 26.00     2     0
## 71                   Jenkin, Mr. Stephen Curnow   male 32.00     0     0
## 72                Goodwin, Miss. Lillian Amy female 16.00     5     2
## 73                         Hood, Mr. Ambrose Jr   male 21.00     0     0
## 74                 Chronopoulos, Mr. Apostolos   male 26.00     1     0
## 75                                 Bing, Mr. Lee   male 32.00     0     0
## 76                     Moen, Mr. Sigurd Hansen   male 25.00     0     0
## 77                               Staneff, Mr. Ivan   male    NA     0     0
## 78                     Moutal, Mr. Rahamin Haim   male    NA     0     0
## 79           Caldwell, Master. Alden Gates   male  0.83     0     2
## 80                  Dowdell, Miss. Elizabeth female 30.00     0     0
## 81                          Waelens, Mr. Achille   male 22.00     0     0
## 82                 Sheerlinck, Mr. Jan Baptist   male 29.00     0     0
## 83                 McDermott, Miss. Brigdet Delia female    NA     0     0
```

```
##  23 female 15.00     0     0
##  24   male 28.00     0     0
##  25 female  8.00     3     1
##  26 female 38.00     1     5
##  27   male    NA     0     0
##  28   male 19.00     3     2
##  29 female    NA     0     0
##  30   male    NA     0     0
##  31   male 40.00     0     0
##  32 female    NA     1     0
##  33 female    NA     0     0
##  34   male 66.00     0     0
##  35   male 28.00     1     0
##  36   male 42.00     1     0
##  37   male    NA     0     0
##  38   male 21.00     0     0
##  39 female 18.00     2     0
##  40 female 14.00     1     0
##  41 female 40.00     1     0
##  42 female 27.00     1     0
##  43   male    NA     0     0
##  44 female  3.00     1     2
##  45 female 19.00     0     0
##  46   male    NA     0     0
##  47   male    NA     1     0
##  48 female    NA     0     0
##  49   male    NA     2     0
##  50 female 18.00     1     0
##  51   male  7.00     4     1
##  52   male 21.00     0     0
##  53 female 49.00     1     0
##  54 female 29.00     1     0
##  55   male 65.00     0     1
##  56   male    NA     0     0
##  57 female 21.00     0     0
##  58   male 28.50     0     0
##  59 female  5.00     1     2
##  60   male 11.00     5     2
##  61   male 22.00     0     0
##  62 female 38.00     0     0
##  63   male 45.00     1     0
##  64   male  4.00     3     2
##  65   male    NA     0     0
##  66   male    NA     1     1
##  67 female 29.00     0     0
##  68   male 19.00     0     0
##  69 female 17.00     4     2
##  70   male 26.00     2     0
##  71   male 32.00     0     0
##  72 female 16.00     5     2
##  73   male 21.00     0     0
##  74   male 26.00     1     0
##  75   male 32.00     0     0
##  76   male 25.00     0     0
```

```
## 77     male    NA     0     0
## 78     male    NA     0     0
## 79     male  0.83     0     2
## 80   female 30.00     0     0
## 81     male 22.00     0     0
## 82     male 29.00     0     0
## 83   female    NA     0     0
##                  Ticket     Fare
## 1             A/5 21171   7.2500
## 2              PC 17599  71.2833
## 3  STON/O2. 3101282   7.9250
## 4              113803  53.1000
## 5              373450   8.0500
## 6              330877   8.4583
## 7               17463  51.8625
## 8              349909  21.0750
## 9              347742  11.1333
## 10             237736  30.0708
## 11           PP 9549  16.7000
## 12             113783  26.5500
## 13          A/5. 2151   8.0500
## 14             347082  31.2750
## 15             350406   7.8542
## 16             248706  16.0000
## 17             382652  29.1250
## 18             244373  13.0000
## 19             345763  18.0000
## 20               2649   7.2250
## 21             239865  26.0000
## 22             248698  13.0000
## 23             330923   8.0292
## 24             113788  35.5000
## 25             349909  21.0750
## 26             347077  31.3875
## 27               2631   7.2250
## 28              19950 263.0000
## 29             330959   7.8792
## 30             349216   7.8958
## 31          PC 17601  27.7208
## 32          PC 17569 146.5208
## 33             335677   7.7500
## 34       C.A. 24579  10.5000
## 35          PC 17604  82.1708
## 36             113789  52.0000
## 37               2677   7.2292
## 38       A./5. 2152   8.0500
## 39             345764  18.0000
## 40               2651  11.2417
## 41               7546   9.4750
## 42              11668  21.0000
## 43             349253   7.8958
## 44     SC/Paris 2123  41.5792
## 45             330958   7.8792
## 46  S.C./A.4. 23567   8.0500
```

```
##  47            370371  15.5000
##  48             14311   7.7500
##  49              2662  21.6792
##  50            349237  17.8000
##  51           3101295  39.6875
##  52        A/4. 39886   7.8000
##  53         PC 17572  76.7292
##  54              2926  26.0000
##  55            113509  61.9792
##  56             19947  35.5000
##  57        C.A. 31026  10.5000
##  58              2697   7.2292
##  59        C.A. 34651  27.7500
##  60            CA 2144  46.9000
##  61              2669   7.2292
##  62            113572  80.0000
##  63             36973  83.4750
##  64            347088  27.9000
##  65         PC 17605  27.7208
##  66              2661  15.2458
##  67        C.A. 29395  10.5000
##  68         S.P. 3464   8.1583
##  69           3101281   7.9250
##  70            315151   8.6625
##  71        C.A. 33111  10.5000
##  72            CA 2144  46.9000
##  73      S.O.C. 14879  73.5000
##  74              2680  14.4542
##  75              1601  56.4958
##  76            348123   7.6500
##  77            349208   7.8958
##  78            374746   8.0500
##  79            248738  29.0000
##  80            364516  12.4750
##  81            345767   9.0000
##  82            345779   9.5000
##  83            330932   7.7875
##              Cabin Embarked
##  1                        S
##  2               C85      C
##  3                        S
##  4              C123      S
##  5                        S
##  6                        Q
##  7               E46      S
##  8                        S
##  9                        S
##  10                       C
##  11               G6      S
##  12             C103      S
##  13                       S
##  14                       S
##  15                       S
##  16                       S
```

```
## 17                        Q
## 18                        S
## 19                        S
## 20                        C
## 21                        S
## 22            D56         S
## 23                        Q
## 24             A6         S
## 25                        S
## 26                        S
## 27                        C
## 28 C23 C25 C27            S
## 29                        Q
## 30                        S
## 31                        C
## 32            B78         C
## 33                        Q
## 34                        S
## 35                        C
## 36                        S
## 37                        C
## 38                        S
## 39                        S
## 40                        C
## 41                        S
## 42                        S
## 43                        C
## 44                        C
## 45                        Q
## 46                        S
## 47                        Q
## 48                        Q
## 49                        C
## 50                        S
## 51                        S
## 52                        S
## 53            D33         C
## 54                        S
## 55            B30         C
## 56            C52         S
## 57                        S
## 58                        C
## 59                        S
## 60                        S
## 61                        C
## 62            B28
## 63            C83         S
## 64                        S
## 65                        C
## 66                        C
## 67            F33         S
## 68                        S
## 69                        S
## 70                        S
```

```
## 71                      S
## 72                      S
## 73                      S
## 74                      C
## 75                      S
## 76        F G73         S
## 77                      S
## 78                      S
## 79                      S
## 80                      S
## 81                      S
## 82                      S
## 83                      Q
##  [ reached 'max' / getOption("max.print") -- omitted 808 rows ]
```

```
View(titanic_data)

### inspect the titanic  dataset
str(titanic_data)
```

```
## 'data.frame':    891 obs. of  12
##  $ PassengerId: int  1 2 3 4 5 6
##  $ Survived   : int  0 1 1 1 0 0
##  $ Pclass     : int  3 1 3 1 3 3
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

```
### the head function diplay the first few rows of the dataset
head(titanic_data)
```

```
##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                  Name
## 1                             Braund, Mr. Owen Harris
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer)
## 3                              Heikkinen, Miss. Laina
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel)
## 5                           Allen, Mr. William Henry
## 6                                   Moran, Mr. James
##      Sex Age SibSp Parch
## 1   male  22     1     0
## 2 female  38     1     0
## 3 female  26     0     0
## 4 female  35     1     0
```

```
## 5     male  35     0     0
## 6     male  NA     0     0
##                 Ticket     Fare Cabin
## 1           A/5 21171   7.2500
## 2            PC 17599  71.2833   C85
## 3 STON/O2. 3101282   7.9250
## 4              113803  53.1000  C123
## 5              373450   8.0500
## 6              330877   8.4583
##   Embarked
## 1        S
## 2        C
## 3        S
## 4        S
## 5        S
## 6        Q
```

```
summary(titanic_data)
```

```
##     PassengerId       Survived
##   Min.   :  1.0   Min.   :0.0000
##   1st Qu.:223.5   1st Qu.:0.0000
##   Median :446.0   Median :0.0000
##   Mean   :446.0   Mean   :0.3838
##   3rd Qu.:668.5   3rd Qu.:1.0000
##   Max.   :891.0   Max.   :1.0000
##
##       Pclass
##   Min.   :1.000
##   1st Qu.:2.000
##   Median :3.000
##   Mean   :2.309
##   3rd Qu.:3.000
##   Max.   :3.000
##
##       Name
##   Length:891
##   Class :character
##   Mode  :character
##
##
##
##
##       Sex
##   Length:891
##   Class :character
##   Mode  :character
##
##
##
##
##        Age              SibSp
##   Min.   : 0.42   Min.   :0.000
##   1st Qu.:20.12   1st Qu.:0.000
##   Median :28.00   Median :0.000
```

```
##  Mean   :29.70      Mean   :0.523
##  3rd Qu.:38.00      3rd Qu.:1.000
##  Max.   :80.00      Max.   :8.000
##  NA's   :177
##       Parch
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.3816
##  3rd Qu.:0.0000
##  Max.   :6.0000
##
##      Ticket
##  Length:891
##  Class :character
##  Mode  :character
##
##
##
##
##         Fare
##  Min.    :  0.00
##  1st Qu.:  7.91
##  Median : 14.45
##  Mean   : 32.20
##  3rd Qu.: 31.00
##  Max.   :512.33
##
##      Cabin
##  Length:891
##  Class :character
##  Mode  :character
##
##
##
##
##     Embarked
##  Length:891
##  Class :character
##  Mode  :character
##
##
##
##
```

```
### displays the number of rows in the datasets
nrow(titanic_data)
```

```
## [1] 891
```

```
### ncol display the number of columns in the dataset
ncol(titanic_data)
```

```
## [1] 12
```

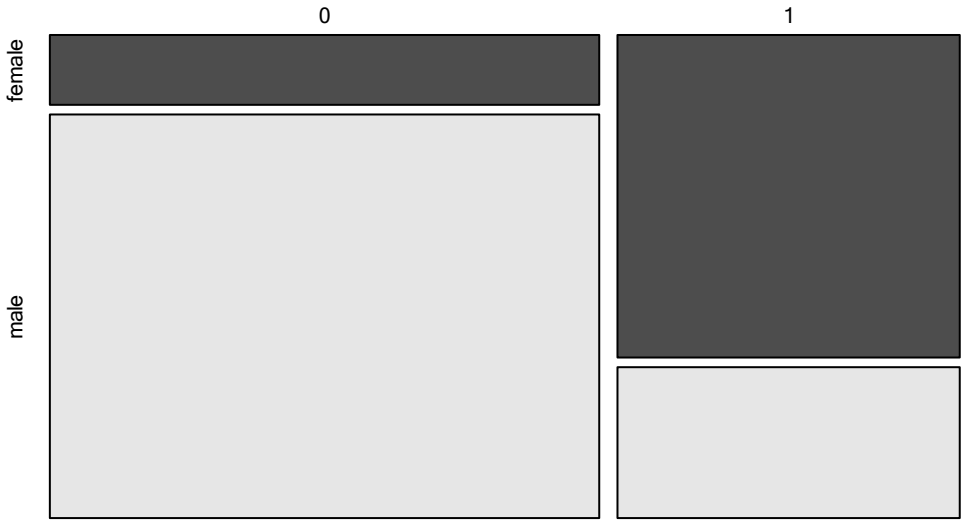
```
### checking the total number of missing values in each column

colSums(is.na(titanic_data))
```

```
##  PassengerId       Survived
##                0            0
##       Pclass           Name
##                0            0
##          Sex            Age
##                0          177
##        SibSp          Parch
##                0            0
##       Ticket           Fare
##                0            0
##        Cabin       Embarked
##                0            0
```

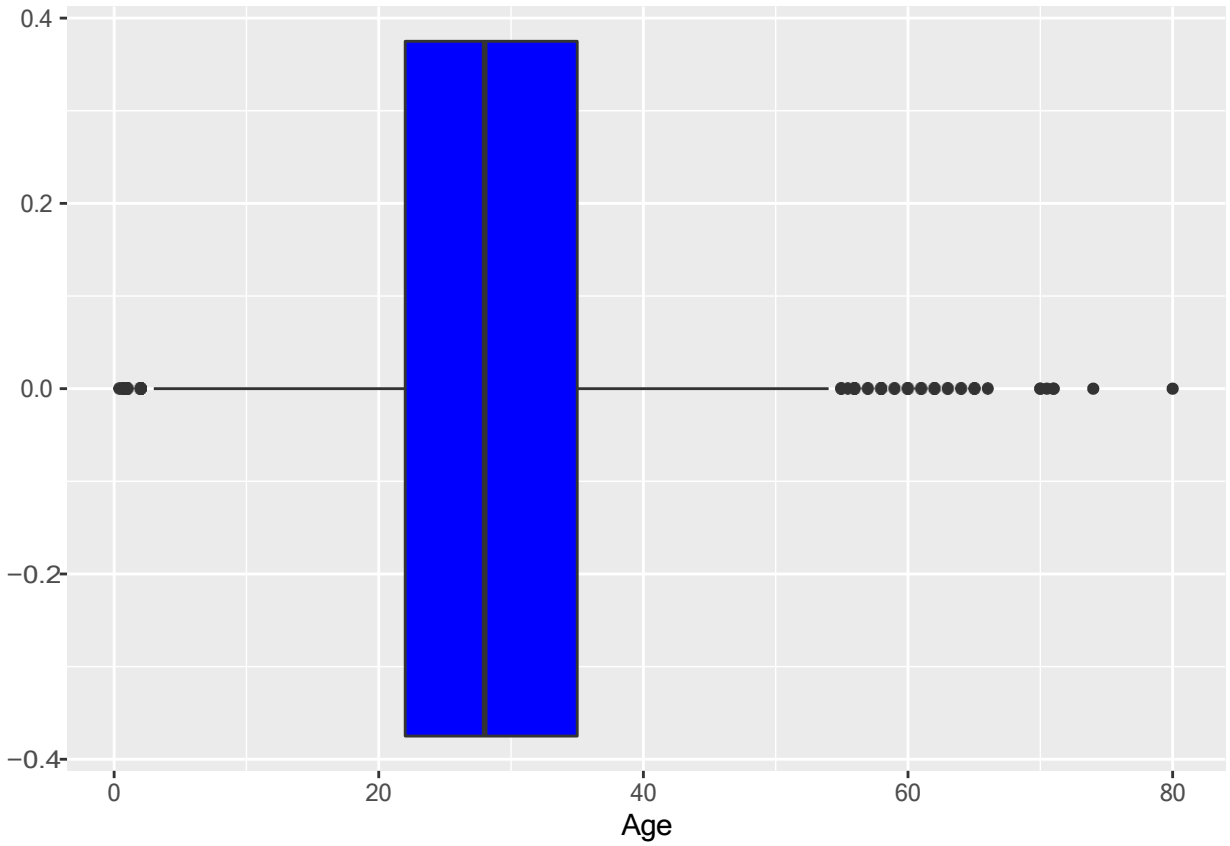
```
###  Data cleaning and Transformation

### Transform the variable into the correct datatype

titanic_data[, c("Survived", "Pclass")] <- lapply(titanic_data[,c("Survived", "Pclass")],
                                                  as.factor)


### Replacing the missing values in the age columns with the median
### of the colunm
titanic_data$Age[is.na(titanic_data$Age)] <- median(titanic_data$Age,
                                                    na.rm = TRUE)

Contingency_table_of_gender_vs_survival <- table(titanic_data$Survived,
                                                 titanic_data$Sex)

prop.table(Contingency_table_of_gender_vs_survival, margin = 2)
```

```
##
##          female       male
##    0  0.2579618  0.8110919
##    1  0.7420382  0.1889081
```

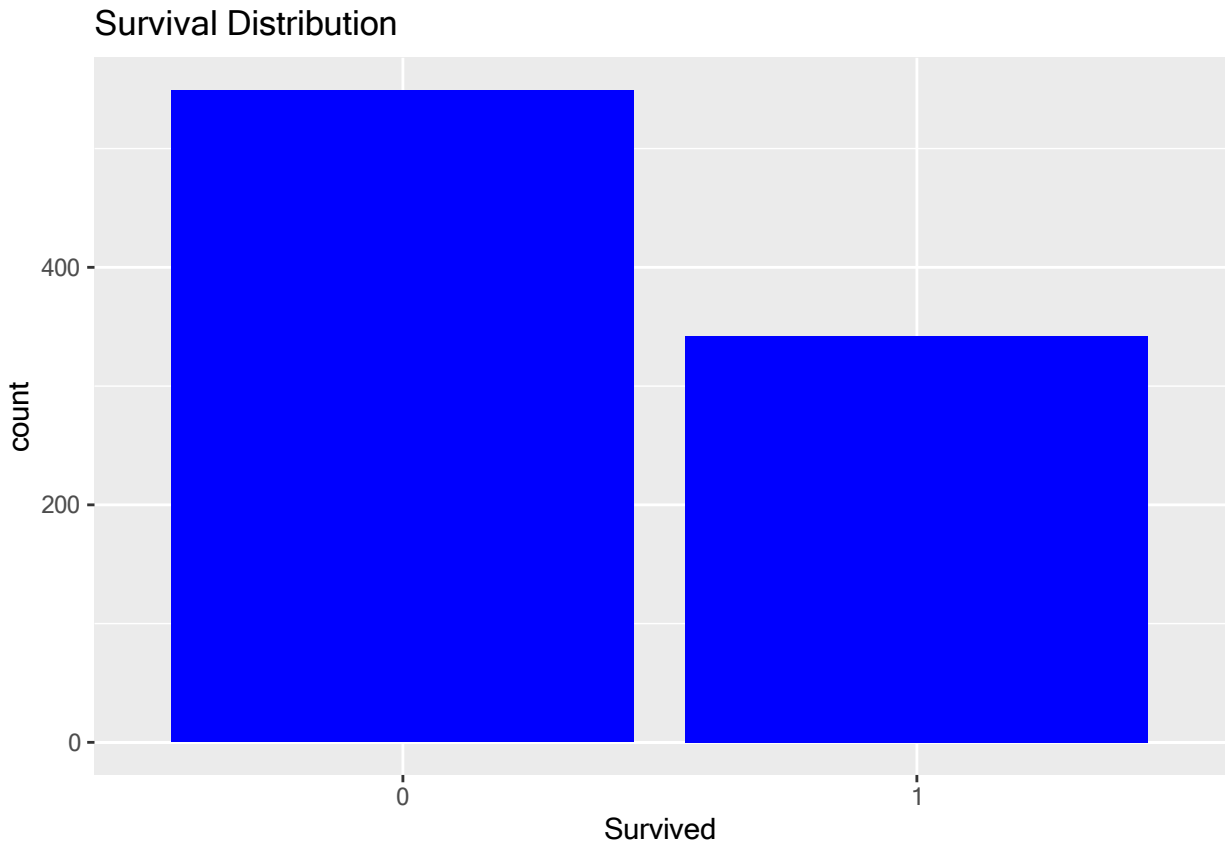
```
### Visualize the Number of people who

mosaicplot(Contingency_table_of_gender_vs_survival, main = "Survival by Sex",
           color=TRUE)
```



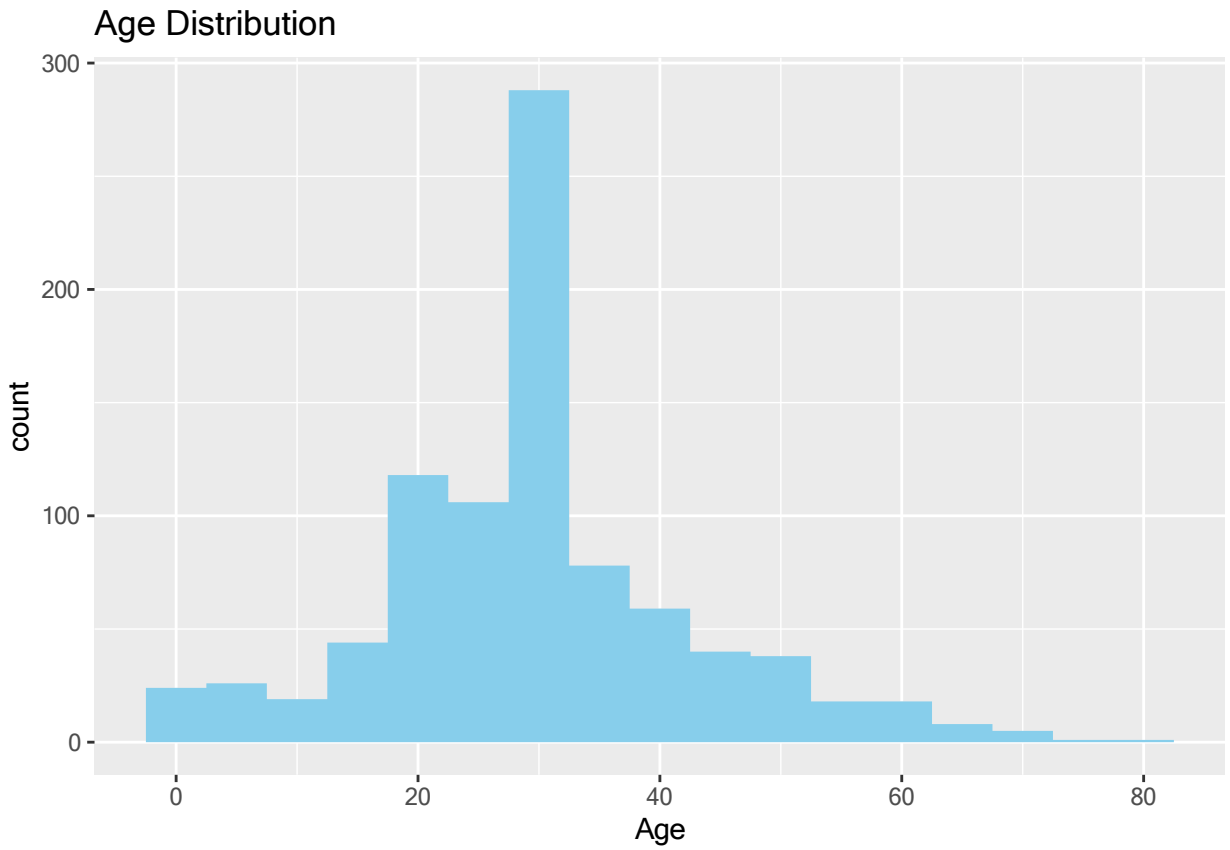

```
### checking out for outliers in the age column

ggplot(titanic_data,   aes(x=Age))+
                        geom_boxplot( fill="blue")
```



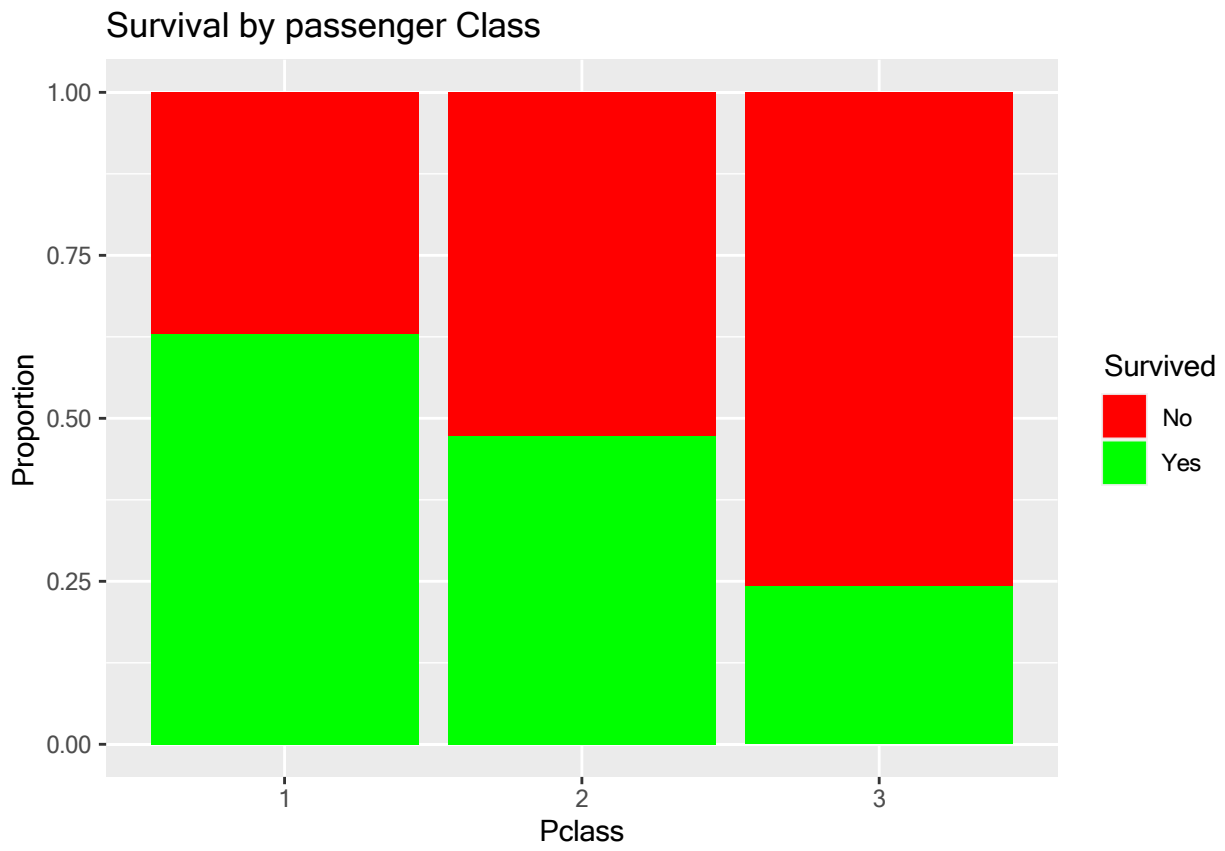

```
### Visulization of the distribution of survival status and age
ggplot(titanic_data, aes(x= Survived))+
  geom_bar(fill="blue")+
  ggtitle("Survival Distribution")+
  labs(X="Survival status (0=No, 1=Yes", Y="Counts")
```



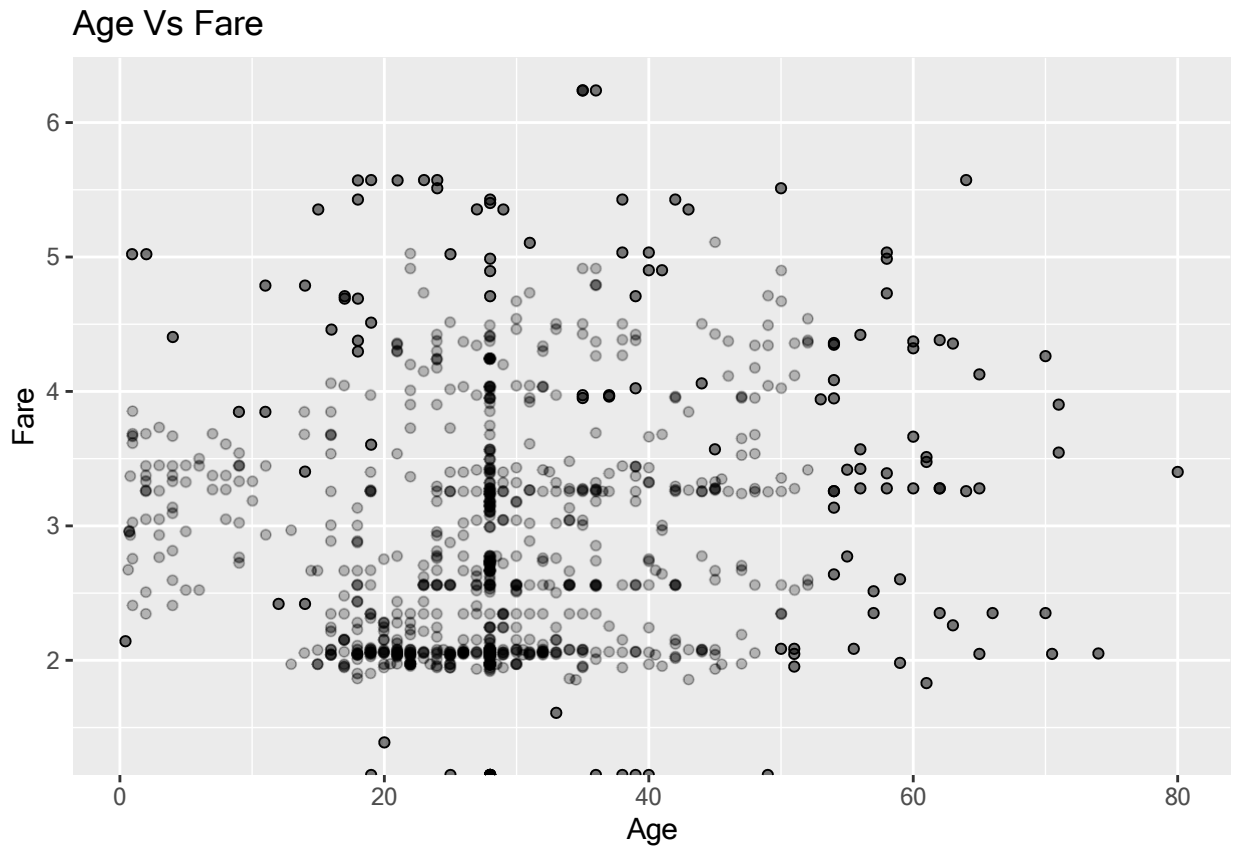

```
### Distribution by age
ggplot(titanic_data, aes(x= Age))+
  geom_histogram(binwidth = 5, fill="skyblue")+
  ggtitle("Age Distribution")
```



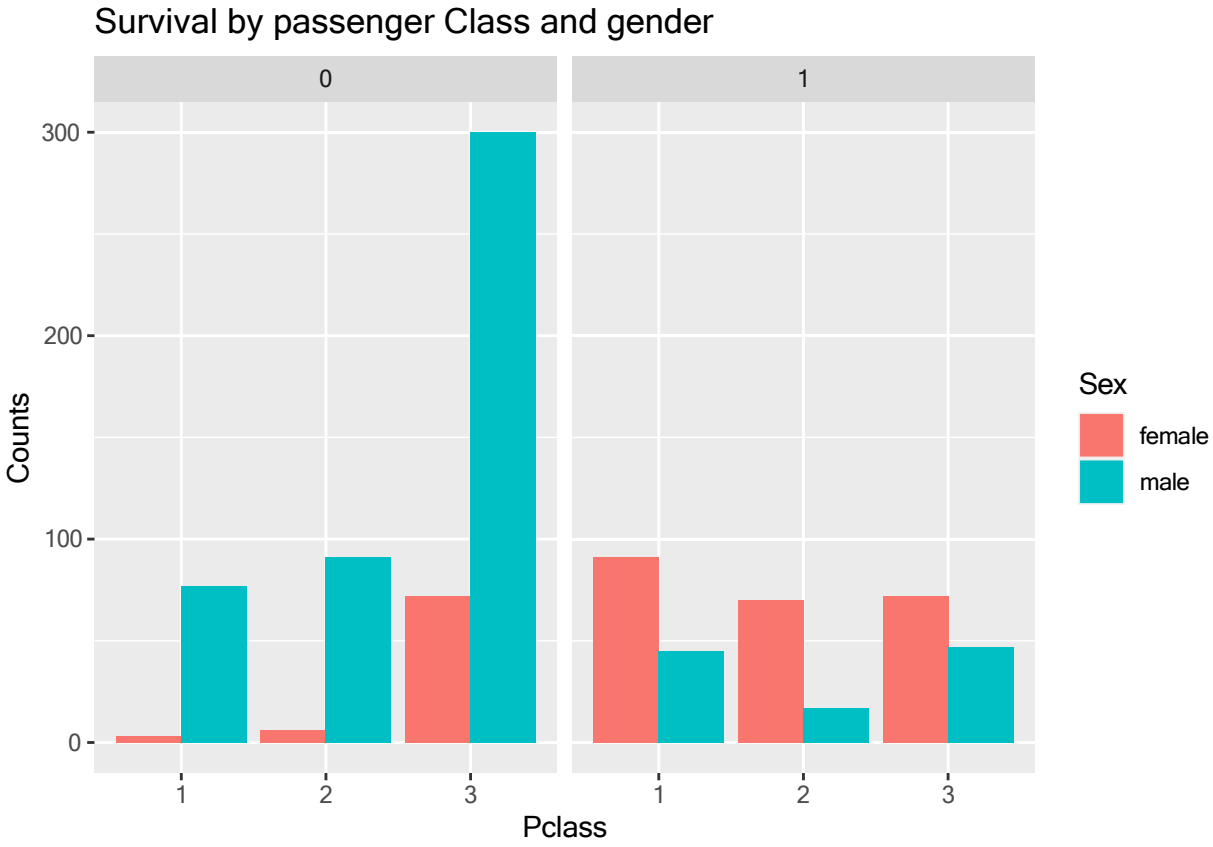

```
### Visualization of bivariate varibles
ggplot(titanic_data, aes(x= Pclass, fill= Survived))+
  geom_bar(position = "fill")+
  labs(title = "Survival by passenger Class", X= "Class",y="Proportion")+
  scale_fill_manual(values = c("red","green"), labels=c("No","Yes"))
```




```
### Age vs Fare

ggplot(titanic_data, aes(x= Age, y= log(Fare)))+
  geom_point(alpha=0.5)+
  labs(title = "Age Vs Fare", x="Age", y="Fare")
```




```
ggplot(titanic_data, aes(x= Pclass, fill= Sex))+
  geom_bar(position = "dodge")+
  facet_wrap(~Survived)+
  labs(title = "Survival by passenger Class and gender", X= "Class",y="Counts")
```


Survival by passenger Class and gender
0
1
300
200
100
0
Counts
1
2
3
1
2
3
Pclass
Sex
female
male