BHARAT INTERN TASK 1

DATA SCIENCE INTERN

TITANIC SURVIVAL PREDICTION

Dataset link: https://www.kaggle.com/datasets/yasserh/titanic-dataset

```
R 4.3.1 · ~/
> titanic=read.csv("C:/Users/sibir/Downloads/Titanic-Dataset.csv")
> head(titanic)
  PassengerId Survived Pclass                                                Name    Sex Age SibSp Parch           Ticket    Fare Cabin Embarked
1           1        0      3                             Braund, Mr. Owen Harris   male  22     1     0        A/5 21171  7.2500              S
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1     0         PC 17599 71.2833   C85        C
3           3        1      3                              Heikkinen, Miss. Laina female  26     0     0 STON/O2. 3101282  7.9250              S
4           4        1      1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1     0           113803 53.1000  C123        S
5           5        0      3                            Allen, Mr. William Henry   male  35     0     0           373450  8.0500              S
6           6        0      3                                    Moran, Mr. James   male  NA     0     0           330877  8.4583              Q
> tail(titanic)
    PassengerId Survived Pclass                                     Name    Sex Age SibSp Parch     Ticket   Fare Cabin Embarked
886         886        0      3      Rice, Mrs. William (Margaret Norton) female  39     0     5     382652 29.125              Q
887         887        0      2                    Montvila, Rev. Juozas   male  27     0     0     211536 13.000              S
888         888        1      1             Graham, Miss. Margaret Edith female  19     0     0     112053 30.000   B42        S
889         889        0      3 Johnston, Miss. Catherine Helen "Carrie" female  NA     1     2 W./C. 6607 23.450              S
890         890        1      1                     Behr, Mr. Karl Howell   male  26     0     0     111369 30.000  C148        C
891         891        0      3                       Dooley, Mr. Patrick   male  32     0     0     370376  7.750              Q
> View(titanic)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500		S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250		S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.0500		S
6	6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583		Q
7	7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.8625	E46	S
8	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750		S
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.1333		S
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.0708		C

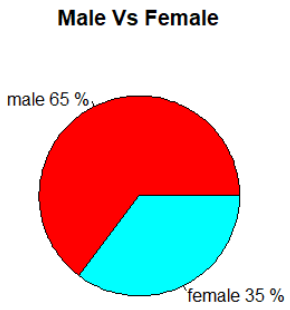
```
> summary(titanic)
  PassengerId       Survived         Pclass          Name               Sex                 Age              SibSp           Parch           Ticket              Fare       
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891         Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000   Length:891         Min.   :  0.00  
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000   Class :character   1st Qu.:  7.91  
 Median :446.0   Median :0.0000   Median :3.000   Mode  :character   Mode  :character   Median :28.00   Median :0.000   Median :0.0000   Mode  :character   Median : 14.45  
 Mean   :446.0   Mean   :0.3838   Mean   :2.309                                       Mean   :29.70   Mean   :0.523   Mean   :0.3816                      Mean   : 32.20  
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000                                       3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000                      3rd Qu.: 31.00  
 Max.   :891.0   Max.   :1.0000   Max.   :3.000                                       Max.   :80.00   Max.   :8.000   Max.   :6.0000                      Max.   :512.33  
                                                                                      NA's   :177                                                                 
    Cabin             Embarked        
 Length:891         Length:891        
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

```
> str(titanic)
'data.frame':   891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...
 $ 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" ...
```

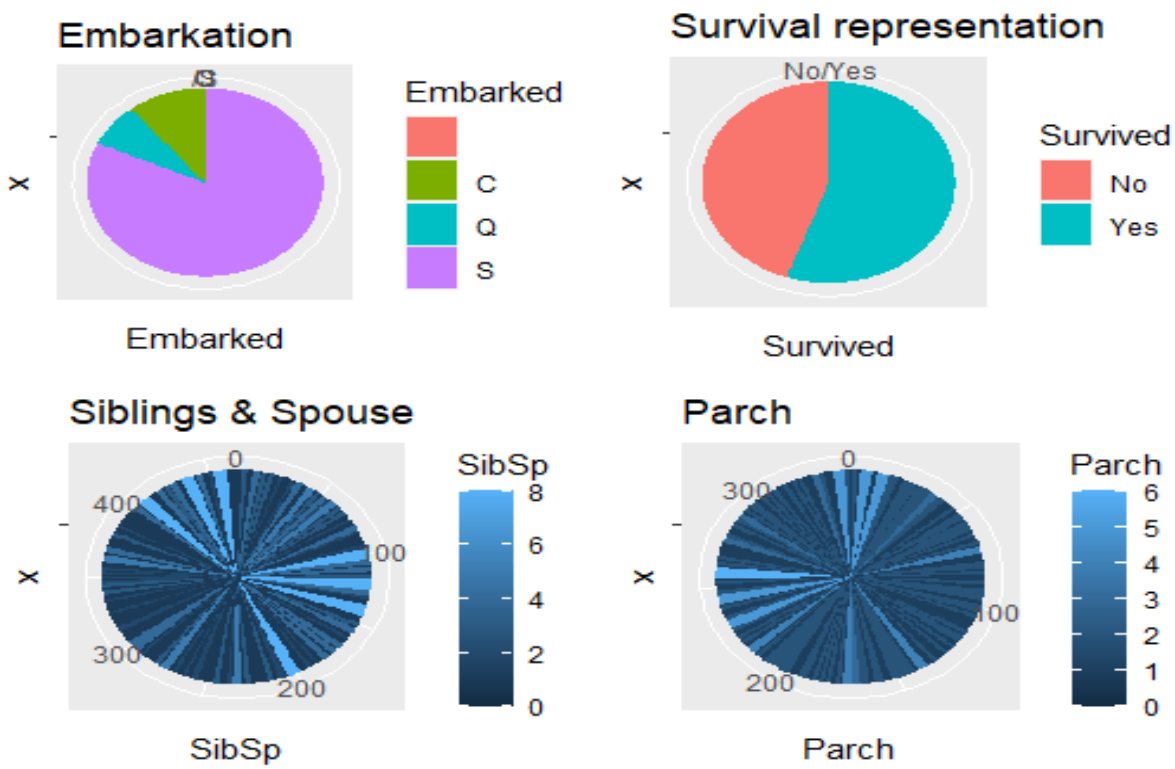
```
> colnames(titanic)
 [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"         "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"        "Cabin"       "Embarked"
> #To check for missing values in the dataset
> sum(is.na(titanic))
[1] 177
> sum(is.na(titanic$Age))
[1] 177
> #Filling the missing values of age with the mean value of rest of the ages
> titanic$Age[is.na(titanic$Age)]=mean(titanic$Age,na.rm = TRUE)
> head(titanic)
  PassengerId Survived Pclass                                                Name    Sex      Age SibSp Parch           Ticket    Fare Cabin Embarked
1           1        0      3                             Braund, Mr. Owen Harris   male 22.00000     1     0        A/5 21171  7.2500              S
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38.00000     1     0         PC 17599 71.2833   C85        C
3           3        1      3                              Heikkinen, Miss. Laina female 26.00000     0     0 STON/O2. 3101282  7.9250              S
4           4        1      1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.00000     1     0           113803 53.1000  C123        S
5           5        0      3                            Allen, Mr. William Henry   male 35.00000     0     0           373450  8.0500              S
6           6        0      3                                    Moran, Mr. James   male 29.69912     0     0           330877  8.4583              Q
> sum(is.na(titanic$Age))
[1] 0
> #remove the column "Cabin" as it has missing characters
> titanic=subset(titanic,select = -Cabin)
> #changing to factor variables
> titanic$Survived=factor(titanic$Survived,levels = c(0,1),labels = c("No","Yes"))
> titanic$Name=as.factor(titanic$Name)
> titanic$Sex=as.factor(titanic$Sex)
> titanic$Ticket=as.factor(titanic$Ticket)
> titanic$Embarked=as.factor(titanic$Embarked)
> str(titanic)
'data.frame':   891 obs. of  11 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 1 1 2 2 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age        : num  22 38 26 35 35 ...
 $ 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     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
> head(titanic)
  PassengerId Survived Pclass                                                Name    Sex      Age SibSp Parch           Ticket    Fare Embarked
1           1       No      3                             Braund, Mr. Owen Harris   male 22.00000     1     0        A/5 21171  7.2500        S
2           2      Yes      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female 38.00000     1     0         PC 17599 71.2833        C
3           3      Yes      3                              Heikkinen, Miss. Laina female 26.00000     0     0 STON/O2. 3101282  7.9250        S
4           4      Yes      1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.00000     1     0           113803 53.1000        S
5           5       No      3                            Allen, Mr. William Henry   male 35.00000     0     0           373450  8.0500        S
6           6       No      3                                    Moran, Mr. James   male 29.69912     0     0           330877  8.4583        Q
```

EXPLORATORY DATA ANALYSIS

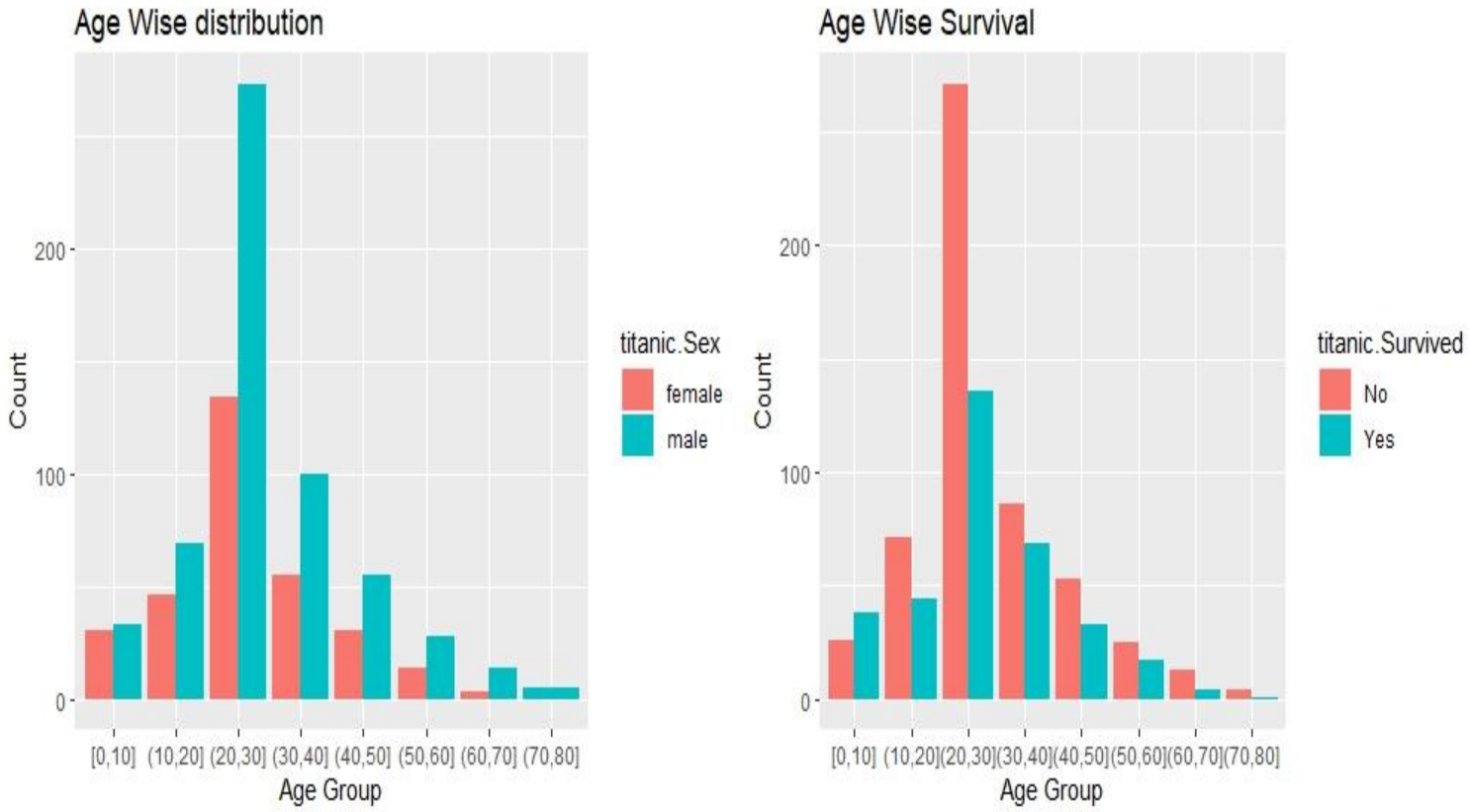
```
> #Exploratory Data Analysis
> #Gender representation
> d1=data.frame(sex=c("male","female"),count=c(sum(titanic$Sex=="male"),sum(titanic$Sex=="female")))
> pct=round((d1$count/sum(d1$count))*100)
> pie(d1$count,labels=paste(d1$sex,sep=" ",pct,"%"),main="Male Vs Female",col=rainbow(length(d1$sex)))
```



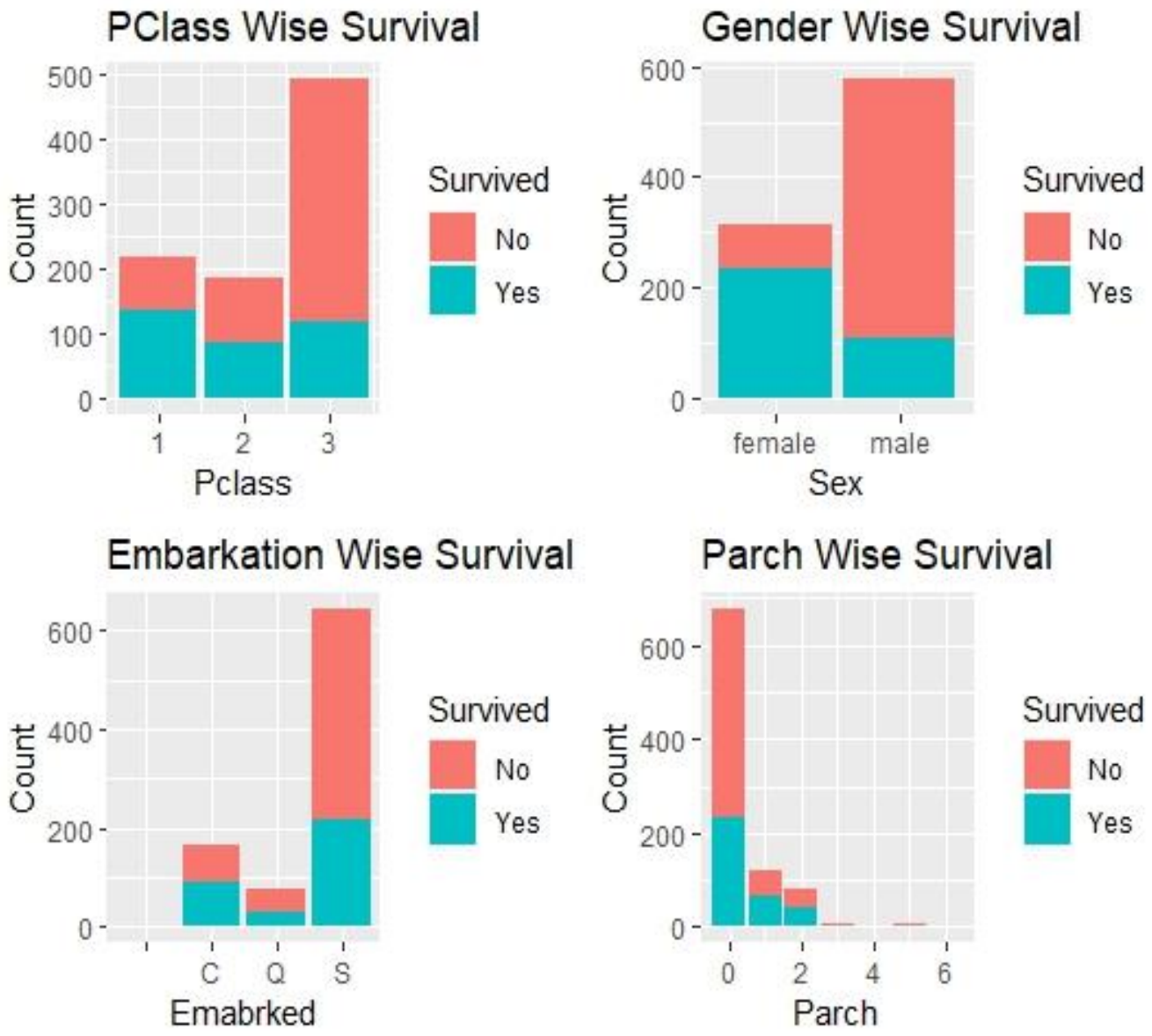
```
> library(tidyverse)
> g1=ggplot(titanic, aes(x="", y=Embarked, fill=Embarked)) + geom_bar(stat="identity", width=1)  + coord_polar("y", start=0) + ggtitle("Embarkation")
> g2=ggplot(titanic, aes(x="", y=Survived, fill=Survived)) + geom_bar(stat="identity", width=1)  + coord_polar("y", start=0) + ggtitle("Survival representation")
> g3=ggplot(titanic, aes(x="", y=SibSp, fill=SibSp)) + geom_bar(stat="identity", width=1)  + coord_polar("y", start=0) + ggtitle("Siblings & Spouse")
> g4=ggplot(titanic, aes(x="", y=Parch, fill=Parch)) + geom_bar(stat="identity", width=1)  + coord_polar("y", start=0) + ggtitle("Parch")
> library(gridExtra)
> grid.arrange(g1,g2,g3,g4,ncol=2)
```



```
> Age_grp=cut(titanic$Age,seq(0,80,10),include.lowest = TRUE)
> df2=data.frame(titanic$Age,titanic$Sex,Age_grp)
> g5=ggplot(df2, aes(x=Age_grp, fill=titanic.Sex)) + geom_bar(position="dodge") + ggtitle("Age Wise distribution") + xlab("Age Group") + ylab("Count")
> df3=data.frame(titanic$Age,titanic$Survived,Age_grp)
> g6=ggplot(df3, aes(x=Age_grp, fill=titanic.Survived)) + geom_bar(position="dodge") + ggtitle("Age Wise Survival") + xlab("Age Group") + ylab("Count")
> grid.arrange(g5,g6,ncol=2)
```



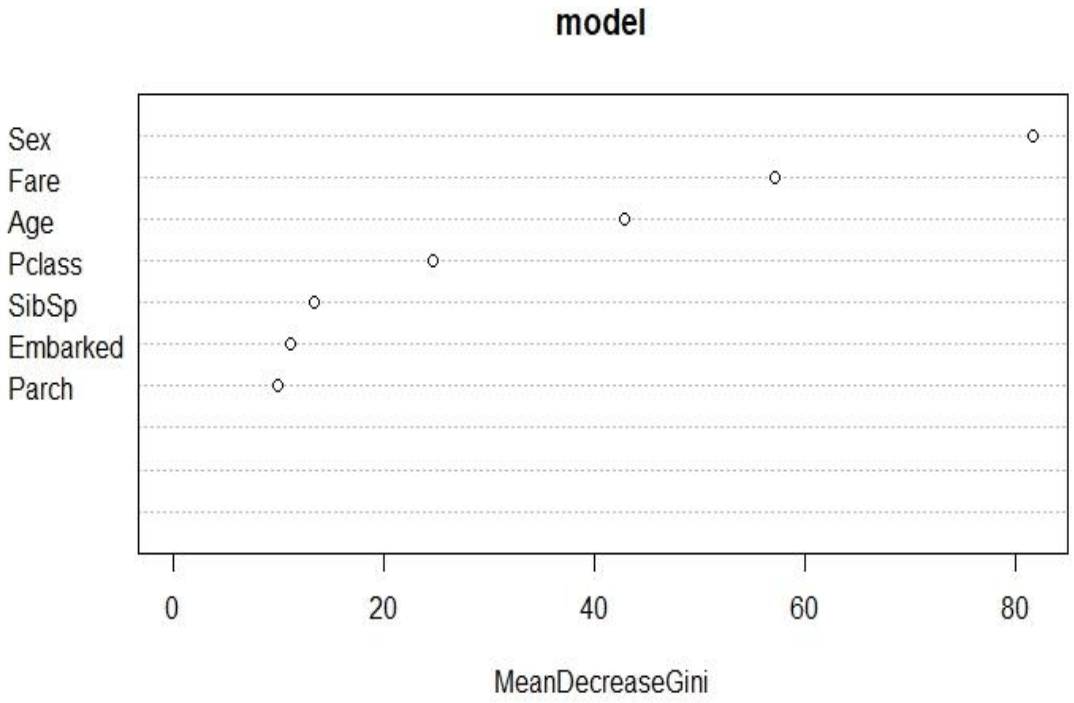
```
› g7=ggplot(data = titanic) + geom_bar(mapping = aes(x = Pclass,fill=Survived)) + ggtitle("PClass Wise Survival") + xlab("Pclass") + ylab("Count")
› g8=ggplot(data = titanic) + geom_bar(mapping = aes(x = Sex,fill=Survived)) + ggtitle("Gender Wise Survival") + xlab("Sex") + ylab("Count")
› g9=ggplot(data = titanic) + geom_bar(mapping = aes(x = Embarked,fill=Survived)) + ggtitle("Embarkation Wise Survival") + xlab("Emabrked") + ylab("Count")
› g10=ggplot(data = titanic) + geom_bar(mapping = aes(x = Parch,fill=Survived)) + ggtitle("Parch Wise Survival") + xlab("Parch") + ylab("Count")
› grid.arrange(g7,g8,g9,g10,ncol=2)
```




```
> #Splitting training and testing data
> obs=nrow(titanic)
> split=round(obs*0.8)
> train=titanic[1:split,]
> test=titanic[(split+1):nrow(titanic),]
> dim(train)
[1] 713  11
> dim(test)
[1] 178  11
```

RANDOM FOREST ALGORITHM

```
> #Using Random Forest algorithm to predict the model
> library(randomForest)
> set.seed(123)
> model=randomForest(formula=Survived~Sex+Fare+Age+Pclass+SibSp+Embarked+Parch,data=train,ntree=500)
> model

Call:
 randomForest(formula = Survived ~ Sex + Fare + Age + Pclass +      SibSp + Embarked + Parch, data = train, ntree = 500)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 17.53%
Confusion matrix:
     No Yes class.error
No  399  35  0.08064516
Yes  90 189  0.32258065
> model$importance
         MeanDecreaseGini
Sex             81.682279
Fare            57.176561
Age             42.924628
Pclass          24.635322
SibSp           13.495989
Embarked        11.149023
Parch            9.955924
```



```
> library(caret)
> confusionMatrix(predict(model,test),test$Survived)
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  106  15
       Yes   9  48

               Accuracy : 0.8652
                 95% CI : (0.8061, 0.9117)
    No Information Rate : 0.6461
    P-Value [Acc > NIR] : 4.077e-11

                  Kappa : 0.6987

 Mcnemar's Test P-Value : 0.3074

            Sensitivity : 0.9217
            Specificity : 0.7619
         Pos Pred Value : 0.8760
         Neg Pred Value : 0.8421
             Prevalence : 0.6461
         Detection Rate : 0.5955
   Detection Prevalence : 0.6798
      Balanced Accuracy : 0.8418

       'Positive' Class : No
```

From the confusion matrix, it is evident that the accuracy of Random Forest algorithm is **86.52%**

LOGISTIC REGRESSION MODEL

```
> #Using Logistic Regression Model
> logisfit=glm(formula = titanic$Survived~Sex+Age+Pclass+Parch+Fare+Embarked+SibSp,family = 'binomial',data = titanic)
> logisfit

Call:  glm(formula = titanic$Survived ~ Sex + Age + Pclass + Parch +
    Fare + Embarked + SibSp, family = "binomial", data = titanic)

Coefficients:
(Intercept)      Sexmale          Age       Pclass        Parch         Fare    EmbarkedC    EmbarkedQ    EmbarkedS
  17.572941    -2.718695    -0.039901    -1.100058    -0.092602     0.001918   -12.287753   -12.321829   -12.706570
      SibSp
  -0.325777

Degrees of Freedom: 890 Total (i.e. Null);  881 Residual
Null Deviance:      1187
Residual Deviance: 784.2        AIC: 804.2
```

```
> summary(logisfit)

Call:
glm(formula = titanic$Survived ~ Sex + Age + Pclass + Parch +
    Fare + Embarked + SibSp, family = "binomial", data = titanic)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  17.572941 610.227586   0.029   0.9770
Sexmale      -2.718695   0.200783 -13.540  < 2e-16 ***
Age          -0.039901   0.007854  -5.080 3.77e-07 ***
Pclass       -1.100058   0.143529  -7.664 1.80e-14 ***
Parch        -0.092602   0.118708  -0.780   0.4353
Fare          0.001918   0.002376   0.807   0.4194
EmbarkedC   -12.287753 610.227400  -0.020   0.9839
EmbarkedQ   -12.321829 610.227452  -0.020   0.9839
EmbarkedS   -12.706570 610.227384  -0.021   0.9834
SibSp        -0.325777   0.109384  -2.978   0.0029 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.66  on 890  degrees of freedom
Residual deviance:  784.19  on 881  degrees of freedom
AIC: 804.19

Number of Fisher Scoring iterations: 13

> logistrain=predict(logisfit,type='response')
> tapply(logistrain, titanic$Survived, mean)
       No       Yes
0.2268479 0.6358494
> logispred=predict(logisfit,newdata=test,type='response')
> summary(logispred)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.006723 0.105609 0.229980 0.372123 0.647668 0.999999
> test[logispred<=0.5,"Survivepred"]="No"
> test[logispred>0.5,"Survivepred"]="Yes"
> View(test)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	Survivepred
714	714	No	3	Larsson, Mr. August Viktor	male	29.00000	0	0	7545	9.4833	S	No
715	715	No	2	Greenberg, Mr. Samuel	male	52.00000	0	0	250647	13.0000	S	No
716	716	No	3	Soholt, Mr. Peter Andreas Lauritz Andersen	male	19.00000	0	0	348124	7.6500	S	No
717	717	Yes	1	Endres, Miss. Caroline Louise	female	38.00000	0	0	PC 17757	227.5250	C	Yes
718	718	Yes	2	Troutt, Miss. Edwina Celia "Winnie"	female	27.00000	0	0	34218	10.5000	S	Yes
719	719	No	3	McEvoy, Mr. Michael	male	29.69912	0	0	36568	15.5000	Q	No
720	720	No	3	Johnson, Mr. Malkolm Joackim	male	33.00000	0	0	347062	7.7750	S	No
721	721	Yes	2	Harper, Miss. Annie Jessie "Nina"	female	6.00000	0	1	248727	33.0000	S	Yes
722	722	No	3	Jensen, Mr. Svend Lauritz	male	17.00000	1	0	350048	7.0542	S	No
723	723	No	2	Gillespie, Mr. William Henry	male	34.00000	0	0	12233	13.0000	S	No
724	724	No	2	Hodges, Mr. Henry Price	male	50.00000	0	0	250643	13.0000	S	No

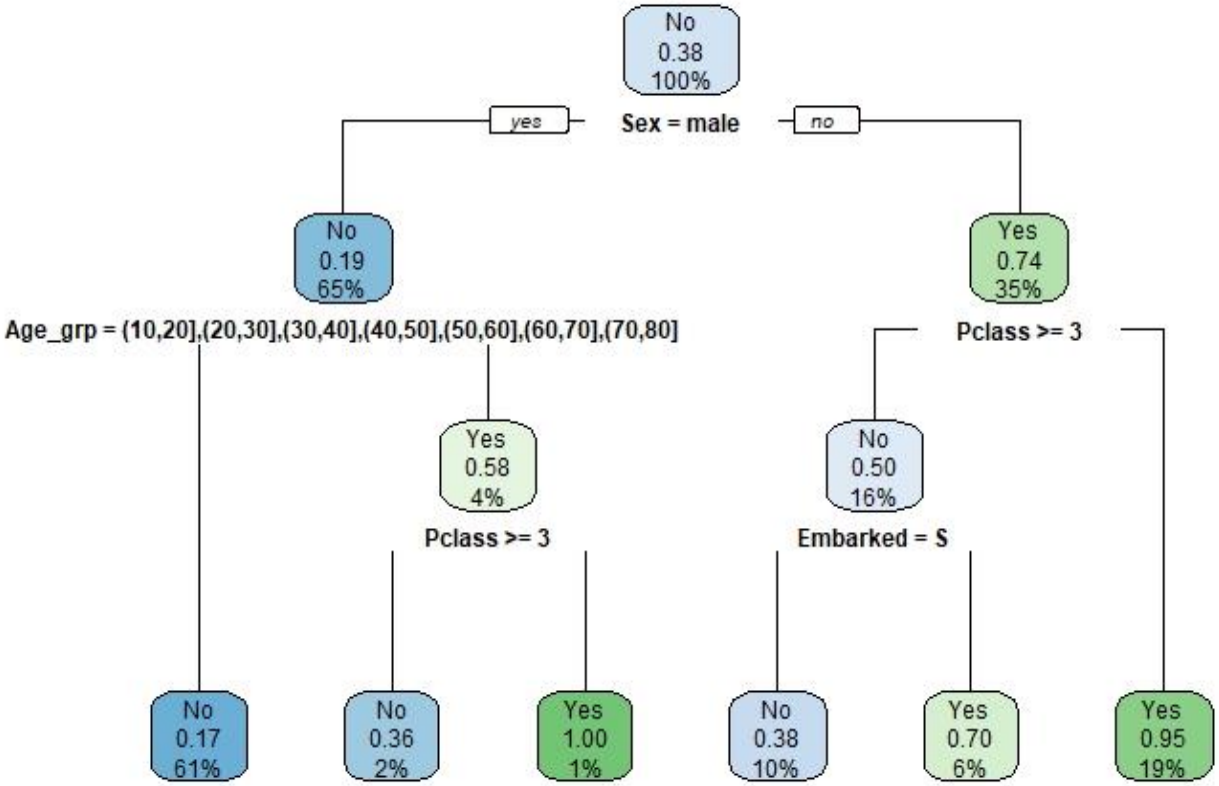
```
> confusionMatrix(table(test$Survived,test$Survivepred),positive = "No")
Confusion Matrix and Statistics


      No Yes
No   104  11
Yes   16  47

               Accuracy : 0.8483
                 95% CI : (0.787, 0.8976)
    No Information Rate : 0.6742
    P-Value [Acc > NIR] : 1.022e-07

                  Kappa : 0.6623

 Mcnemar's Test P-Value : 0.4414

            Sensitivity : 0.8667
            Specificity : 0.8103
         Pos Pred Value : 0.9043
         Neg Pred Value : 0.7460
             Prevalence : 0.6742
         Detection Rate : 0.5843
   Detection Prevalence : 0.6461
      Balanced Accuracy : 0.8385

       'Positive' Class : No
```

From the confusion matrix, it is evident that the accuracy of Logistic Regression model is **84.83%**

DECISION TREE ALGORITHM

```
> #Using Decision tree algorithm for prediction
> library(rpart)
> library(rpart.plot)
> titanic2=data.frame(titanic,Age_grp)
> View(titanic2)
> tree=rpart(Survived~Age_grp+Sex+Embarked+Pclass+Parch,titanic)
> pred=data.frame(Age_grp=c("(20,30]"),Sex=c("female"),Embarked=c("C"),Pclass=c(1),Parch=c(1))
> result=predict(tree,pred)
> result
         No       Yes
1 0.05294118 0.9470588
> rpart.plot(tree)
```

The probability that a female belonging to the age group of 20-30 embarked at ‘C’, belonging to passenger class 1 and Parch 1 will survive titanic is **94.7%**. This means she will **survive**.


No
0.38
100%
yes
Sex = male
no
No
0.19
65%
Yes
0.74
35%
Age_grp = (10,20],(20,30],(30,40],(40,50],(50,60],(60,70],(70,80]
Pclass >= 3
Yes
0.58
4%
No
0.50
16%
Pclass >= 3
Embarked = S
No
0.17
61%
No
0.36
2%
Yes
1.00
1%
No
0.38
10%
Yes
0.70
6%
Yes
0.95
19%