Titanic (Train)

Session 6 Assignment 1

1. Import the Titanic Dataset from the link Titanic Data Set.

Perform the following:

a. Preprocess the passenger names to come up with a list of titles that represent families and represent using appropriate visualization graph.

b. Represent the proportion of people survived from the family size using a graph.

c. Impute the missing values in Age variable using Mice Library, create two different graphs showing Age distribution before and after imputation.

a. Preprocess the passenger names to come up with a list of titles that represent families and represent using appropriate visualization graph.

Preprocessing Data

```
library(reshape)
library(caret)
d <- train
d.nrow<-seq(1, nrow(d)) # save the number of  rows in the train dataset
d.miss <- melt(apply(d[, -2], 2, function(x) sum(is.na(x) | x=="")))
cbind(row.names(d.miss)[d.miss$value>0], d.miss[d.miss$value>0,])
```

```
     [,1]       [,2]
[1,] "Age"      "177"
[2,] "Cabin"    "687"
[3,] "Embarked" "2"
```

```
#Variable "Cabin"
#"Cabin" has missed about 80% values. We will not use this variable.

#Variable "Embarked"
#Update missing Embarked value with the most common value:
```

```
#table(d$Embarked)
#Variable "Price"
#Some Fare values contains sum for tickets were purchased in groups. Introduc
e a new variable "Price" that will be Fare per person.

d$Fare[which(is.na(d$Fare))] <- 0 # Update missing Fare value with 0.
# calculate Ticket Price (Fare per person)
ticket.count <- aggregate(d$Ticket, by=list(d$Ticket), function(x) sum( !is.n
a(x) ))
d$Price<-apply(d, 1, function(x) as.numeric(x["Fare"]) / ticket.count[which(t
icket.count[, 1] == x["Ticket"]), 2])
```

```

        Capt          Col          Don           Dr     Jonkheer         Lady 
           1            2            1            7            1            1 
       Major       Master         Miss         Mlle          Mme           Mr 
           2           40          182            2            1          517 
         Mrs           Ms          Rev          Sir the Countess 
         125            1            6            1            1 
```

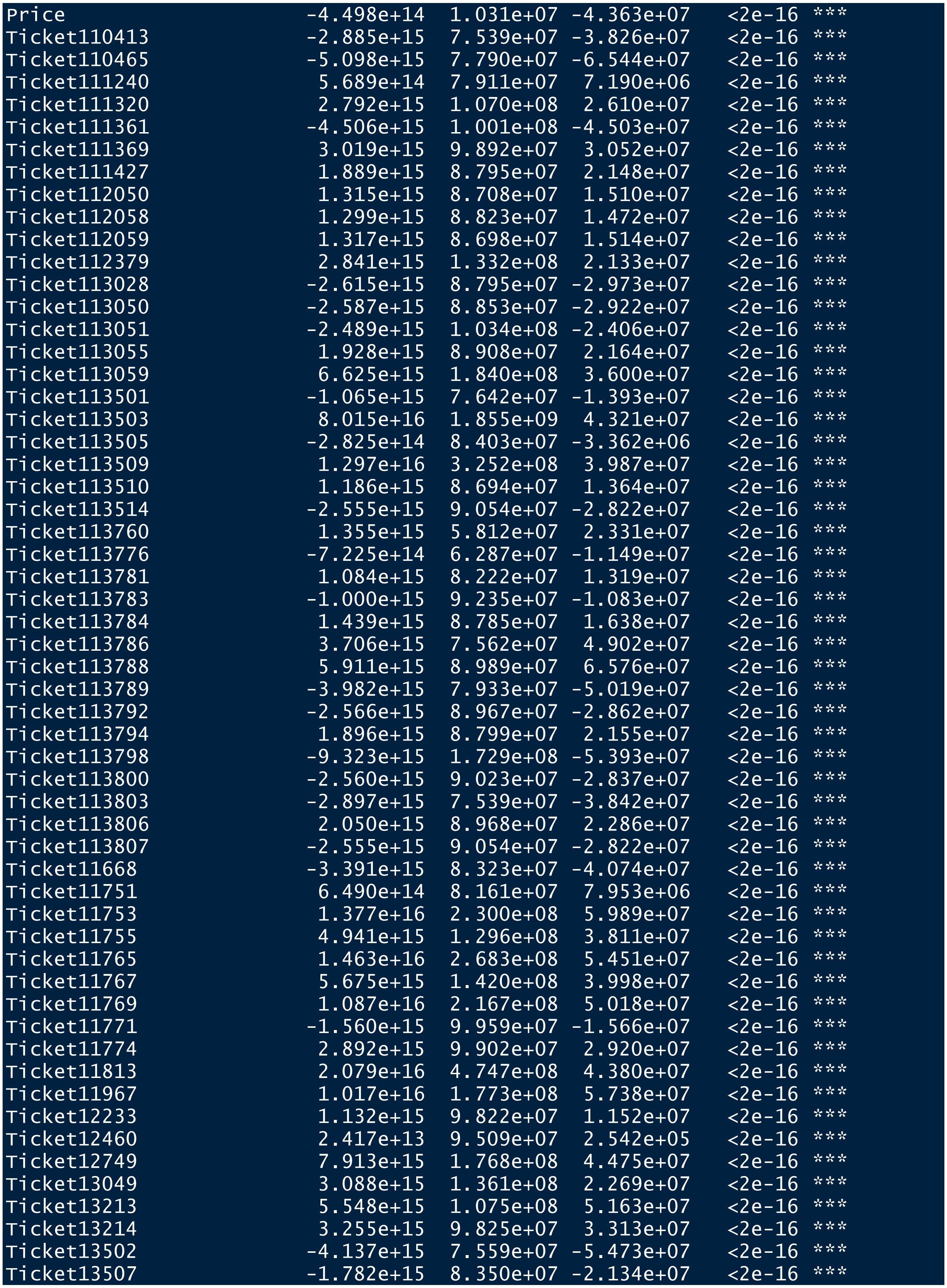
```
#Price related to passenger class. Missig price values (price=0) we can updat
e with median price per passenger class:

pclass.price<-aggregate(d$Price, by = list(d$Pclass), FUN = function(x) media
n(x, na.rm = T))
d[which(d$Price==0), "Price"] <- apply(d[which(d$Price==0), ] , 1, function(x
) pclass.price[pclass.price[, 1]==x["Pclass"], 2])
#Variable "Title"
#Extract title of each persons name to a new variable "Title"
d$Title<-regmatches(as.character(d$Name),regexpr("\\,[A-z ]{1,20}\\.", as.cha
racter(d$Name)))
d$Title<-unlist(lapply(d$Title,FUN=function(x) substr(x, 3, nchar(x)-1)))
table(d$Title)
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

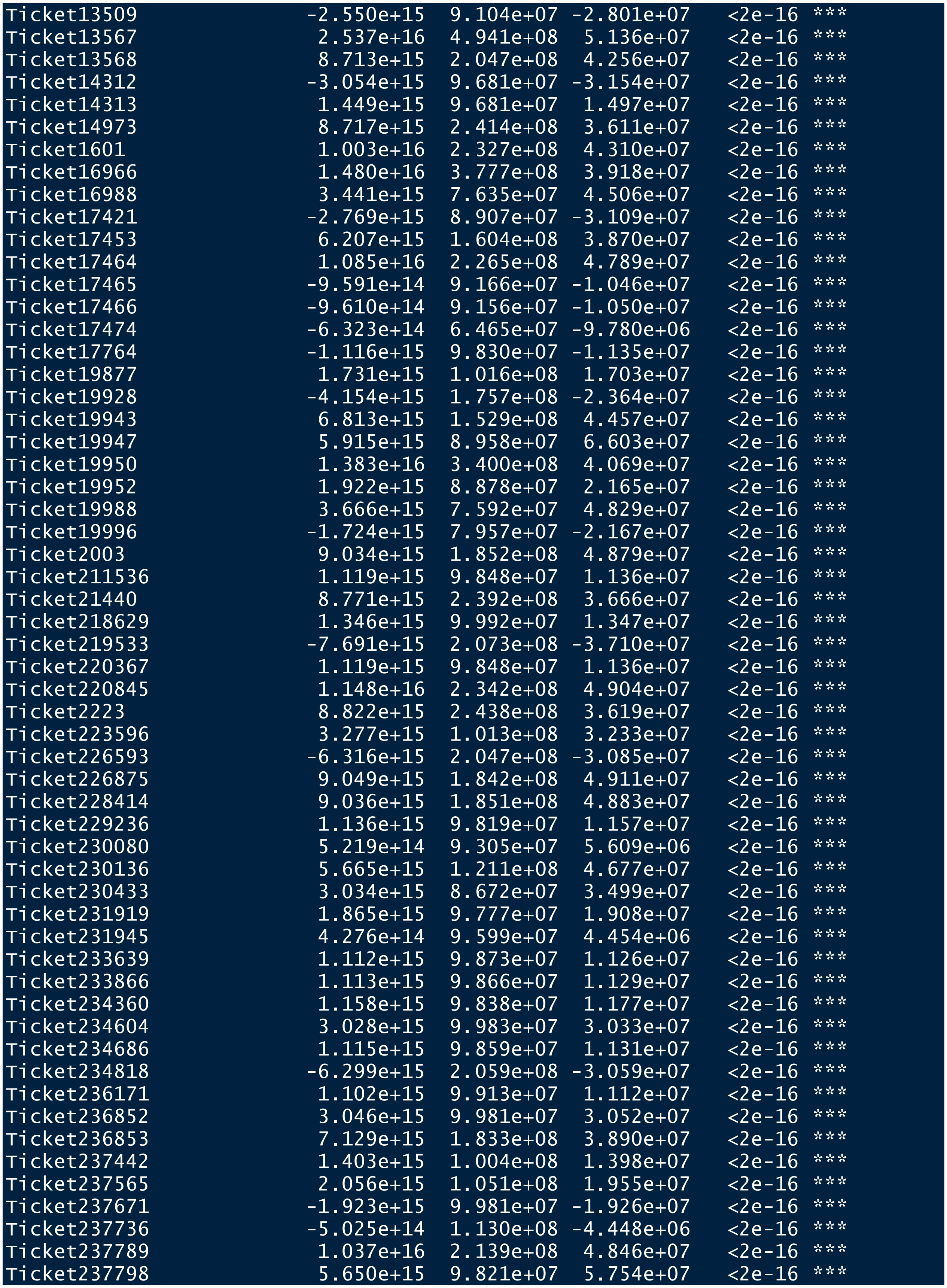
```
#Merge 17 different title groups to the most common 4 groups.

d$Title[which(d$Title %in% c("Mme", "Mlle"))] <- "Miss"
d$Title[which(d$Title %in% c("Lady", "Ms", "the Countess", "Dona"))] <- "Mrs"
d$Title[which(d$Title=="Dr" & d$Sex=="female")] <- "Mrs"
d$Title[which(d$Title=="Dr" & d$Sex=="male")] <- "Mr"
d$Title[which(d$Title %in% c("Capt", "Col", "Don", "Jonkheer", "Major", "Rev"
, "Sir"))] <- "Mr"
d$Title<-as.factor(d$Title) #convert to factor variable
```

```
#Variable "Age"
#Update unknown age with median age for each group of title:

title.age<-aggregate(d$Age,by = list(d$Title), FUN = function(x) median(x, na
.rm = T))
d[is.na(d$Age), "Age"] <- apply(d[is.na(d$Age), ] , 1, function(x) title.age[
title.age[, 1]==x["Title"], 2])
#Split train and test data
#We merged train and test data at the begining of preprocess. Now we will spl
it it back to "t" and "d" Data frame variables.
#Data frame "t" has no "Survival" values and will be used to predict "Surviva
l" and submit on Kaggle.
#Data frame "d" that contains train data we also split to test prediction mod
els.

t <- d[-d.nrow, ] # test data. It has no "Survival" values.
d <- d[d.nrow, ] #Train data
set.seed(1234)
inTrain<-createDataPartition(d$Survived, p = 0.8)[[1]]
#Fitting a linear model that includes all variables.
fit.8 <- glm(Survived ~ Pclass+Sex+Age+SibSp+Parch+Embarked+Title+Price+Ticke
t, data=d[inTrain,], family=binomial(("logit")))
summary(fit.8)
#Fitting a linear model that includes 5 statistically significant variable an
d "Ticket" converted to a factor variable.
fit.6.grp <- glm(Survived ~ Pclass+Age+SibSp+Parch+Title+I(Ticket>2), data=d[
inTrain,], family=binomial)
summary(fit.6.grp)
```

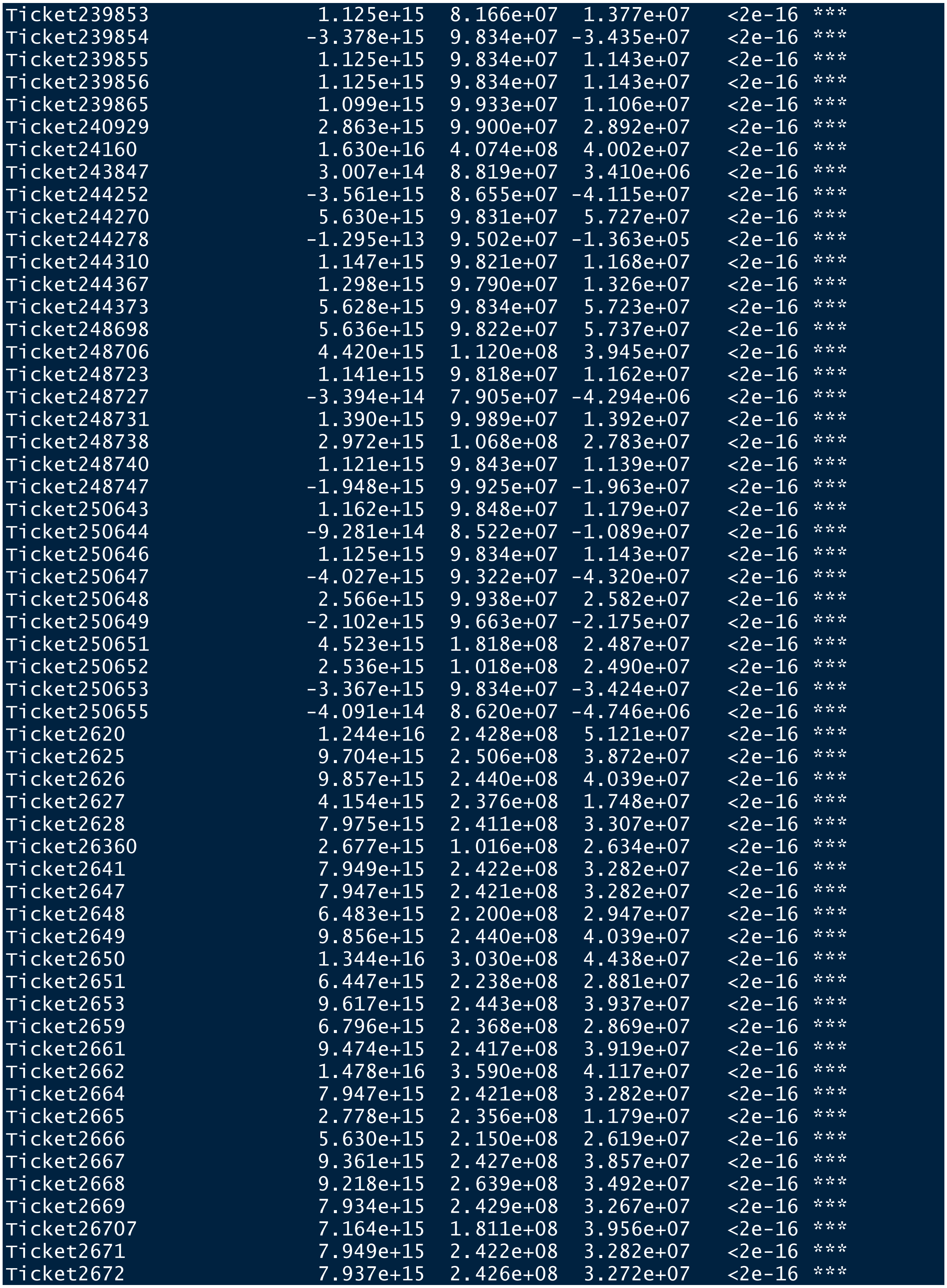
```
Call:
glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch + 
    Embarked + Title + Price + Ticket, family = binomial(("logit")), 
    data = d[inTrain, ])

Deviance Residuals: 
   Min      1Q  Median      3Q     Max  
 -8.49    0.00    0.00    0.00    8.49  

Coefficients: (4 not defined because of singularities)
                           Estimate Std. Error    z value Pr(>|z|)    
(Intercept)               2.746e+16  5.354e+08  5.130e+07   <2e-16 ***
Pclass                   -9.834e+15  2.233e+08 -4.404e+07   <2e-16 ***
Sexmale                   6.646e+14  2.627e+07  2.531e+07   <2e-16 ***
Age                      -1.852e+12  6.684e+05 -2.771e+06   <2e-16 ***
SibSp                    -1.666e+14  1.213e+07 -1.373e+07   <2e-16 ***
Parch                     4.025e+12  1.418e+07  2.839e+05   <2e-16 ***
EmbarkedQ                 8.159e+15  2.468e+08  3.306e+07   <2e-16 ***
EmbarkedS                -4.140e+14  6.218e+07 -6.659e+06   <2e-16 ***
TitleMiss                 4.578e+14  2.026e+07  2.260e+07   <2e-16 ***
TitleMr                  -3.269e+15  2.549e+07 -1.282e+08   <2e-16 ***
TitleMrs                         NA         NA         NA       NA    
```



```
Price                    -4.498e+14  1.031e+07 -4.363e+07   <2e-16 ***
Ticket110413             -2.885e+15  7.539e+07 -3.826e+07   <2e-16 ***
Ticket110465             -5.098e+15  7.790e+07 -6.544e+07   <2e-16 ***
Ticket111240              5.689e+14  7.911e+07  7.190e+06   <2e-16 ***
Ticket111320              2.792e+15  1.070e+08  2.610e+07   <2e-16 ***
Ticket111361             -4.506e+15  1.001e+08 -4.503e+07   <2e-16 ***
Ticket111369              3.019e+15  9.892e+07  3.052e+07   <2e-16 ***
Ticket111427              1.889e+15  8.795e+07  2.148e+07   <2e-16 ***
Ticket112050              1.315e+15  8.708e+07  1.510e+07   <2e-16 ***
Ticket112058              1.299e+15  8.823e+07  1.472e+07   <2e-16 ***
Ticket112059              1.317e+15  8.698e+07  1.514e+07   <2e-16 ***
Ticket112379              2.841e+15  1.332e+08  2.133e+07   <2e-16 ***
Ticket113028             -2.615e+15  8.795e+07 -2.973e+07   <2e-16 ***
Ticket113050             -2.587e+15  8.853e+07 -2.922e+07   <2e-16 ***
Ticket113051             -2.489e+15  1.034e+08 -2.406e+07   <2e-16 ***
Ticket113055              1.928e+15  8.908e+07  2.164e+07   <2e-16 ***
Ticket113059              6.625e+15  1.840e+08  3.600e+07   <2e-16 ***
Ticket113501             -1.065e+15  7.642e+07 -1.393e+07   <2e-16 ***
Ticket113503              8.015e+16  1.855e+09  4.321e+07   <2e-16 ***
Ticket113505             -2.825e+14  8.403e+07 -3.362e+06   <2e-16 ***
Ticket113509              1.297e+16  3.252e+08  3.987e+07   <2e-16 ***
Ticket113510              1.186e+15  8.694e+07  1.364e+07   <2e-16 ***
Ticket113514             -2.555e+15  9.054e+07 -2.822e+07   <2e-16 ***
Ticket113760              1.355e+15  5.812e+07  2.331e+07   <2e-16 ***
Ticket113776             -7.225e+14  6.287e+07 -1.149e+07   <2e-16 ***
Ticket113781              1.084e+15  8.222e+07  1.319e+07   <2e-16 ***
Ticket113783             -1.000e+15  9.235e+07 -1.083e+07   <2e-16 ***
Ticket113784              1.439e+15  8.785e+07  1.638e+07   <2e-16 ***
Ticket113786              3.706e+15  7.562e+07  4.902e+07   <2e-16 ***
Ticket113788              5.911e+15  8.989e+07  6.576e+07   <2e-16 ***
Ticket113789             -3.982e+15  7.933e+07 -5.019e+07   <2e-16 ***
Ticket113792             -2.566e+15  8.967e+07 -2.862e+07   <2e-16 ***
Ticket113794              1.896e+15  8.799e+07  2.155e+07   <2e-16 ***
Ticket113798             -9.323e+15  1.729e+08 -5.393e+07   <2e-16 ***
Ticket113800             -2.560e+15  9.023e+07 -2.837e+07   <2e-16 ***
Ticket113803             -2.897e+15  7.539e+07 -3.842e+07   <2e-16 ***
Ticket113806              2.050e+15  8.968e+07  2.286e+07   <2e-16 ***
Ticket113807             -2.555e+15  9.054e+07 -2.822e+07   <2e-16 ***
Ticket11668              -3.391e+15  8.323e+07 -4.074e+07   <2e-16 ***
Ticket11751               6.490e+14  8.161e+07  7.953e+06   <2e-16 ***
Ticket11753               1.377e+16  2.300e+08  5.989e+07   <2e-16 ***
Ticket11755               4.941e+15  1.296e+08  3.811e+07   <2e-16 ***
Ticket11765               1.463e+16  2.683e+08  5.451e+07   <2e-16 ***
Ticket11767               5.675e+15  1.420e+08  3.998e+07   <2e-16 ***
Ticket11769               1.087e+16  2.167e+08  5.018e+07   <2e-16 ***
Ticket11771              -1.560e+15  9.959e+07 -1.566e+07   <2e-16 ***
Ticket11774               2.892e+15  9.902e+07  2.920e+07   <2e-16 ***
Ticket11813               2.079e+16  4.747e+08  4.380e+07   <2e-16 ***
Ticket11967               1.017e+16  1.773e+08  5.738e+07   <2e-16 ***
Ticket12233               1.132e+15  9.822e+07  1.152e+07   <2e-16 ***
Ticket12460               2.417e+13  9.509e+07  2.542e+05   <2e-16 ***
Ticket12749               7.913e+15  1.768e+08  4.475e+07   <2e-16 ***
Ticket13049               3.088e+15  1.361e+08  2.269e+07   <2e-16 ***
Ticket13213               5.548e+15  1.075e+08  5.163e+07   <2e-16 ***
Ticket13214               3.255e+15  9.825e+07  3.313e+07   <2e-16 ***
Ticket13502              -4.137e+15  7.559e+07 -5.473e+07   <2e-16 ***
Ticket13507              -1.782e+15  8.350e+07 -2.134e+07   <2e-16 ***
```



```
Ticket13509              -2.550e+15  9.104e+07 -2.801e+07   <2e-16 ***
Ticket13567               2.537e+16  4.941e+08  5.136e+07   <2e-16 ***
Ticket13568               8.713e+15  2.047e+08  4.256e+07   <2e-16 ***
Ticket14312              -3.054e+15  9.681e+07 -3.154e+07   <2e-16 ***
Ticket14313               1.449e+15  9.681e+07  1.497e+07   <2e-16 ***
Ticket14973               8.717e+15  2.414e+08  3.611e+07   <2e-16 ***
Ticket1601                1.003e+16  2.327e+08  4.310e+07   <2e-16 ***
Ticket16966               1.480e+16  3.777e+08  3.918e+07   <2e-16 ***
Ticket16988               3.441e+15  7.635e+07  4.506e+07   <2e-16 ***
Ticket17421              -2.769e+15  8.907e+07 -3.109e+07   <2e-16 ***
Ticket17453               6.207e+15  1.604e+08  3.870e+07   <2e-16 ***
Ticket17464               1.085e+16  2.265e+08  4.789e+07   <2e-16 ***
Ticket17465              -9.591e+14  9.166e+07 -1.046e+07   <2e-16 ***
Ticket17466              -9.610e+14  9.156e+07 -1.050e+07   <2e-16 ***
Ticket17474              -6.323e+14  6.465e+07 -9.780e+06   <2e-16 ***
Ticket17764              -1.116e+15  9.830e+07 -1.135e+07   <2e-16 ***
Ticket19877               1.731e+15  1.016e+08  1.703e+07   <2e-16 ***
Ticket19928              -4.154e+15  1.757e+08 -2.364e+07   <2e-16 ***
Ticket19943               6.813e+15  1.529e+08  4.457e+07   <2e-16 ***
Ticket19947               5.915e+15  8.958e+07  6.603e+07   <2e-16 ***
Ticket19950               1.383e+16  3.400e+08  4.069e+07   <2e-16 ***
Ticket19952               1.922e+15  8.878e+07  2.165e+07   <2e-16 ***
Ticket19988               3.666e+15  7.592e+07  4.829e+07   <2e-16 ***
Ticket19996              -1.724e+15  7.957e+07 -2.167e+07   <2e-16 ***
Ticket2003                9.034e+15  1.852e+08  4.879e+07   <2e-16 ***
Ticket211536              1.119e+15  9.848e+07  1.136e+07   <2e-16 ***
Ticket21440               8.771e+15  2.392e+08  3.666e+07   <2e-16 ***
Ticket218629              1.346e+15  9.992e+07  1.347e+07   <2e-16 ***
Ticket219533             -7.691e+15  2.073e+08 -3.710e+07   <2e-16 ***
Ticket220367              1.119e+15  9.848e+07  1.136e+07   <2e-16 ***
Ticket220845              1.148e+16  2.342e+08  4.904e+07   <2e-16 ***
Ticket2223                8.822e+15  2.438e+08  3.619e+07   <2e-16 ***
Ticket223596              3.277e+15  1.013e+08  3.233e+07   <2e-16 ***
Ticket226593             -6.316e+15  2.047e+08 -3.085e+07   <2e-16 ***
Ticket226875              9.049e+15  1.842e+08  4.911e+07   <2e-16 ***
Ticket228414              9.036e+15  1.851e+08  4.883e+07   <2e-16 ***
Ticket229236              1.136e+15  9.819e+07  1.157e+07   <2e-16 ***
Ticket230080              5.219e+14  9.305e+07  5.609e+06   <2e-16 ***
Ticket230136              5.665e+15  1.211e+08  4.677e+07   <2e-16 ***
Ticket230433              3.034e+15  8.672e+07  3.499e+07   <2e-16 ***
Ticket231919              1.865e+15  9.777e+07  1.908e+07   <2e-16 ***
Ticket231945              4.276e+14  9.599e+07  4.454e+06   <2e-16 ***
Ticket233639              1.112e+15  9.873e+07  1.126e+07   <2e-16 ***
Ticket233866              1.113e+15  9.866e+07  1.129e+07   <2e-16 ***
Ticket234360              1.158e+15  9.838e+07  1.177e+07   <2e-16 ***
Ticket234604              3.028e+15  9.983e+07  3.033e+07   <2e-16 ***
Ticket234686              1.115e+15  9.859e+07  1.131e+07   <2e-16 ***
Ticket234818             -6.299e+15  2.059e+08 -3.059e+07   <2e-16 ***
Ticket236171              1.102e+15  9.913e+07  1.112e+07   <2e-16 ***
Ticket236852              3.046e+15  9.981e+07  3.052e+07   <2e-16 ***
Ticket236853              7.129e+15  1.833e+08  3.890e+07   <2e-16 ***
Ticket237442              1.403e+15  1.004e+08  1.398e+07   <2e-16 ***
Ticket237565              2.056e+15  1.051e+08  1.955e+07   <2e-16 ***
Ticket237671             -1.923e+15  9.981e+07 -1.926e+07   <2e-16 ***
Ticket237736             -5.025e+14  1.130e+08 -4.448e+06   <2e-16 ***
Ticket237789              1.037e+16  2.139e+08  4.846e+07   <2e-16 ***
Ticket237798              5.650e+15  9.821e+07  5.754e+07   <2e-16 ***
```



```
Ticket239853             1.125e+15  8.166e+07  1.377e+07   <2e-16 ***
Ticket239854            -3.378e+15  9.834e+07 -3.435e+07   <2e-16 ***
Ticket239855             1.125e+15  9.834e+07  1.143e+07   <2e-16 ***
Ticket239856             1.125e+15  9.834e+07  1.143e+07   <2e-16 ***
Ticket239865             1.099e+15  9.933e+07  1.106e+07   <2e-16 ***
Ticket240929             2.863e+15  9.900e+07  2.892e+07   <2e-16 ***
Ticket24160              1.630e+16  4.074e+08  4.002e+07   <2e-16 ***
Ticket243847             3.007e+14  8.819e+07  3.410e+06   <2e-16 ***
Ticket244252            -3.561e+15  8.655e+07 -4.115e+07   <2e-16 ***
Ticket244270             5.630e+15  9.831e+07  5.727e+07   <2e-16 ***
Ticket244278            -1.295e+13  9.502e+07 -1.363e+05   <2e-16 ***
Ticket244310             1.147e+15  9.821e+07  1.168e+07   <2e-16 ***
Ticket244367             1.298e+15  9.790e+07  1.326e+07   <2e-16 ***
Ticket244373             5.628e+15  9.834e+07  5.723e+07   <2e-16 ***
Ticket248698             5.636e+15  9.822e+07  5.737e+07   <2e-16 ***
Ticket248706             4.420e+15  1.120e+08  3.945e+07   <2e-16 ***
Ticket248723             1.141e+15  9.818e+07  1.162e+07   <2e-16 ***
Ticket248727            -3.394e+14  7.905e+07 -4.294e+06   <2e-16 ***
Ticket248731             1.390e+15  9.989e+07  1.392e+07   <2e-16 ***
Ticket248738             2.972e+15  1.068e+08  2.783e+07   <2e-16 ***
Ticket248740             1.121e+15  9.843e+07  1.139e+07   <2e-16 ***
Ticket248747            -1.948e+15  9.925e+07 -1.963e+07   <2e-16 ***
Ticket250643             1.162e+15  9.848e+07  1.179e+07   <2e-16 ***
Ticket250644            -9.281e+14  8.522e+07 -1.089e+07   <2e-16 ***
Ticket250646             1.125e+15  9.834e+07  1.143e+07   <2e-16 ***
Ticket250647            -4.027e+15  9.322e+07 -4.320e+07   <2e-16 ***
Ticket250648             2.566e+15  9.938e+07  2.582e+07   <2e-16 ***
Ticket250649            -2.102e+15  9.663e+07 -2.175e+07   <2e-16 ***
Ticket250651             4.523e+15  1.818e+08  2.487e+07   <2e-16 ***
Ticket250652             2.536e+15  1.018e+08  2.490e+07   <2e-16 ***
Ticket250653            -3.367e+15  9.834e+07 -3.424e+07   <2e-16 ***
Ticket250655            -4.091e+14  8.620e+07 -4.746e+06   <2e-16 ***
Ticket2620               1.244e+16  2.428e+08  5.121e+07   <2e-16 ***
Ticket2625               9.704e+15  2.506e+08  3.872e+07   <2e-16 ***
Ticket2626               9.857e+15  2.440e+08  4.039e+07   <2e-16 ***
Ticket2627               4.154e+15  2.376e+08  1.748e+07   <2e-16 ***
Ticket2628               7.975e+15  2.411e+08  3.307e+07   <2e-16 ***
Ticket26360              2.677e+15  1.016e+08  2.634e+07   <2e-16 ***
Ticket2641               7.949e+15  2.422e+08  3.282e+07   <2e-16 ***
Ticket2647               7.947e+15  2.421e+08  3.282e+07   <2e-16 ***
Ticket2648               6.483e+15  2.200e+08  2.947e+07   <2e-16 ***
Ticket2649               9.856e+15  2.440e+08  4.039e+07   <2e-16 ***
Ticket2650               1.344e+16  3.030e+08  4.438e+07   <2e-16 ***
Ticket2651               6.447e+15  2.238e+08  2.881e+07   <2e-16 ***
Ticket2653               9.617e+15  2.443e+08  3.937e+07   <2e-16 ***
Ticket2659               6.796e+15  2.368e+08  2.869e+07   <2e-16 ***
Ticket2661               9.474e+15  2.417e+08  3.919e+07   <2e-16 ***
Ticket2662               1.478e+16  3.590e+08  4.117e+07   <2e-16 ***
Ticket2664               7.947e+15  2.421e+08  3.282e+07   <2e-16 ***
Ticket2665               2.778e+15  2.356e+08  1.179e+07   <2e-16 ***
Ticket2666               5.630e+15  2.150e+08  2.619e+07   <2e-16 ***
Ticket2667               9.361e+15  2.427e+08  3.857e+07   <2e-16 ***
Ticket2668               9.218e+15  2.639e+08  3.492e+07   <2e-16 ***
Ticket2669               7.934e+15  2.429e+08  3.267e+07   <2e-16 ***
Ticket26707              7.164e+15  1.811e+08  3.956e+07   <2e-16 ***
Ticket2671               7.949e+15  2.422e+08  3.282e+07   <2e-16 ***
Ticket2672               7.937e+15  2.426e+08  3.272e+07   <2e-16 ***
```

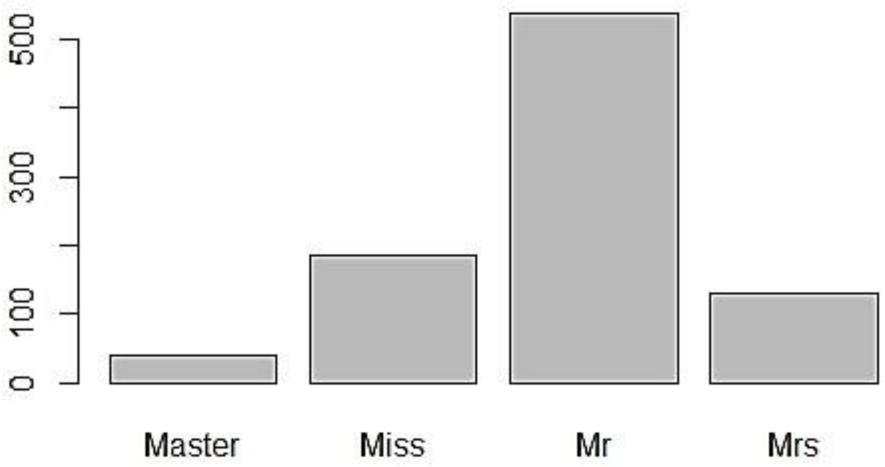
```
Ticket2674                 7.947e+15  2.421e+08  3.282e+07   <2e-16 ***
Ticket2678                 3.103e+15  2.368e+08  1.311e+07   <2e-16 ***
Ticket2680                 1.136e+16  2.982e+08  3.808e+07   <2e-16 ***
Ticket2683                 7.601e+15  2.360e+08  3.221e+07   <2e-16 ***
Ticket2685                 7.949e+15  2.422e+08  3.282e+07   <2e-16 ***
Ticket2686                 7.949e+15  2.422e+08  3.282e+07   <2e-16 ***
Ticket2687                 9.359e+15  2.429e+08  3.853e+07   <2e-16 ***
Ticket2689                 8.772e+15  2.997e+08  2.927e+07   <2e-16 ***
Ticket2690                 8.087e+15  2.386e+08  3.389e+07   <2e-16 ***
Ticket2691                 2.862e+15  2.352e+08  1.217e+07   <2e-16 ***
Ticket2693                 7.937e+15  2.427e+08  3.270e+07   <2e-16 ***
Ticket2694                 7.947e+15  2.421e+08  3.282e+07   <2e-16 ***
Ticket2695                 8.083e+15  2.388e+08  3.385e+07   <2e-16 ***
Ticket2697                 7.946e+15  2.423e+08  3.279e+07   <2e-16 ***
Ticket2699                 8.901e+15  2.556e+08  3.482e+07   <2e-16 ***
Ticket2700                 7.949e+15  2.422e+08  3.282e+07   <2e-16 ***
Ticket27042                3.533e+15  8.017e+07  4.407e+07   <2e-16 ***
Ticket27267                2.571e+15  9.948e+07  2.584e+07   <2e-16 ***
 [ reached getOption("max.print") -- omitted 375 rows ]
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 943.08  on 711  degrees of freedom
Residual deviance: 576.70  on 141  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 1718.7

Number of Fisher Scoring iterations: 16
```

```
Call:
glm(formula = Survived ~ Pclass + Age + SibSp + Parch + Title + 
    I(Ticket > 2), family = binomial, data = d[inTrain, ])

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.6268  -0.5138  -0.3596   0.5342   2.6923  

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)        5.69803    0.72768   7.830 4.86e-15 ***
Pclass            -1.09370    0.16973  -6.444 1.16e-10 ***
Age               -0.02927    0.01073  -2.726 0.006403 ** 
SibSp             -0.48954    0.13350  -3.667 0.000246 ***
Parch             -0.26606    0.14547  -1.829 0.067404 .  
TitleMiss         -0.31742    0.54436  -0.583 0.559818    
TitleMr           -3.33644    0.59484  -5.609 2.04e-08 ***
TitleMrs           0.56676    0.61818   0.917 0.359237    
I(Ticket > 2)TRUE -0.82075    0.34158  -2.403 0.016271 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 945.03  on 712  degrees of freedom
Residual deviance: 567.25  on 704  degrees of freedom
AIC: 585.25

Number of Fisher Scoring iterations: 5
```



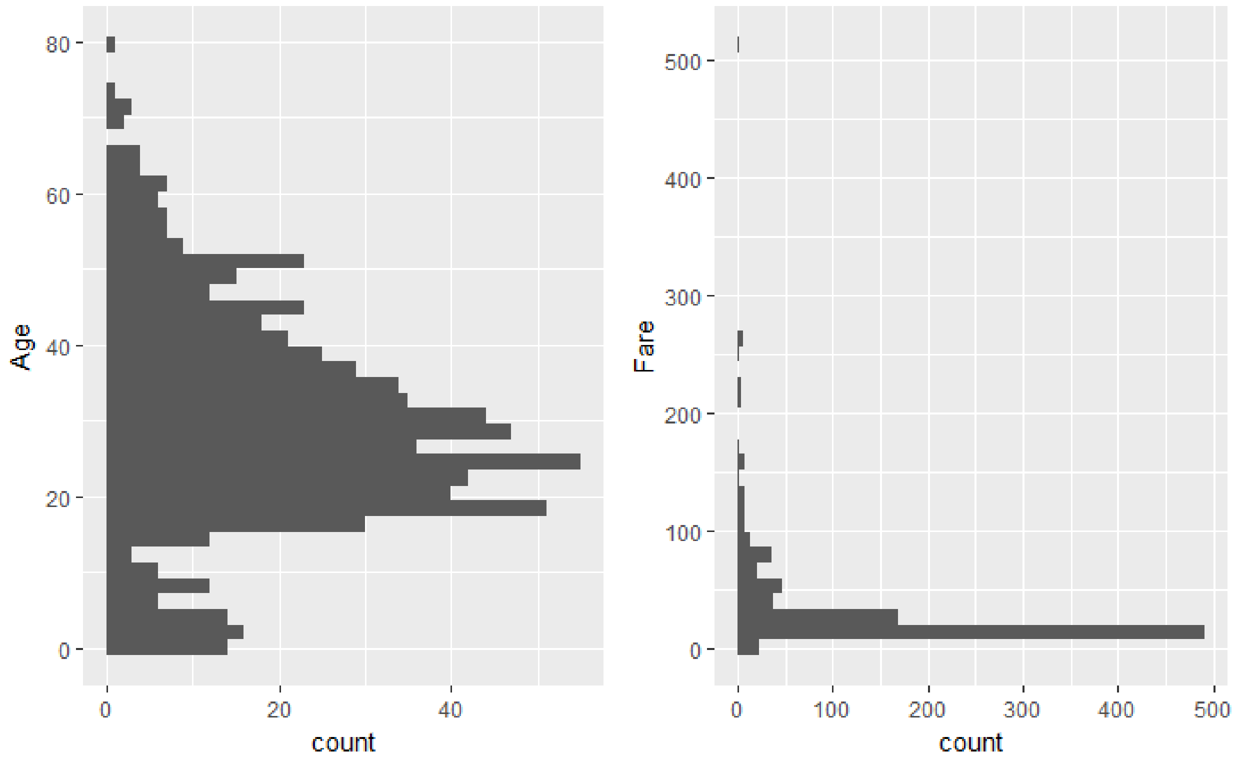
p1 <- ggplot(data=train,aes(x=Age)) + geom_histogram(aes(fill=Survived),bins = 40) + coord_flip()

p2 <- ggplot(data=train,aes(x=Fare)) + geom_histogram(aes(fill=Survived),bins = 40) + coord_flip()

grid.arrange(p1,p2,nrow=1)

summary(train$Fare)

```
get_legend<-function(myggplot){
  tmp <- ggplot_gtable(ggplot_build(myggplot))
  leg <- which(sapply(train, function(x) x$name) == "guide-box")
  legend <- tmp$grobs[[leg]]
  return(legend)
}
p <- lapply(X = c('Pclass','Sex','SibSp','Parch','Embarked'),
            FUN = function(x) ggplot(data = train)+
                aes_string(x=x,fill='Survived')+
                geom_bar(position="dodge")+
                theme(legend.position="none"))
```

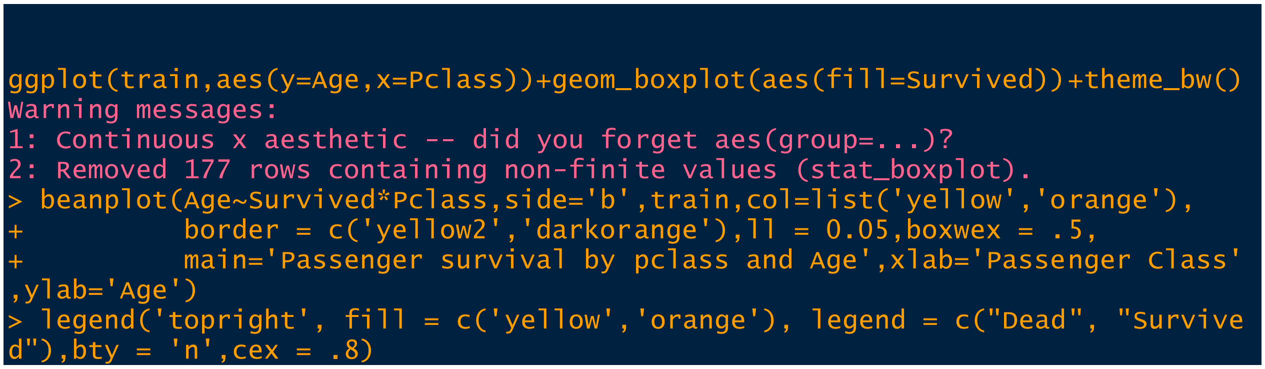


~~~
```{r}

 summary(train$Embarked)

train.imp <- train

train.imp$Embarked[is.na(train.imp$Embarked)]  <- 'S'

train.imp$title  <- str_extract(pattern = '[a-zA-Z]+(?=\\.)',string = train.imp$Name)

train.imp$title  <- as.factor(train.imp$title)

ggplot(train.imp,aes(x=title,y=Age))+

   geom_jitter(shape=21,alpha=.6,col='blue')+

   stat_summary(aes(y = Age,group=1), fun.y=median, colour="red",
geom="point",group=1)+

   theme_bw()+

   theme(axis.text.x = element_text(angle = 45, hjust = 1),legend.position="none")+

   labs(caption='red points are median values')

train.imp$title  <- as.character(train.imp$title)

train.imp$title[train.imp$title   %in% c('Capt','Col','Major')] <- 'Officer'

train.imp$title[train.imp$title   %in%
c('Don','Dr','Rev','Sir','Jonkheer','Countess','Lady','Dona')] <- 'Royalty'

train.imp$title[train.imp$title   %in% c('Mrs','Mme')] <- 'Mrs'

train.imp$title[train.imp$title   %in% c('Ms','Mlle')] <- 'Miss'

train.imp$title  <- as.factor(train.imp$title)

ggplot(train.imp,aes(x=title,y=Age))+

   geom_jitter(color='blue',shape=21,alpha=.7)+

   stat_summary(aes(y = Age,group=1), fun.y=median, colour="red",
geom="point",group=1)+

   theme_bw()+

   theme(axis.text.x = element_text(angle = 45, hjust = 1))+

   labs(caption='red points are median values')
~~~

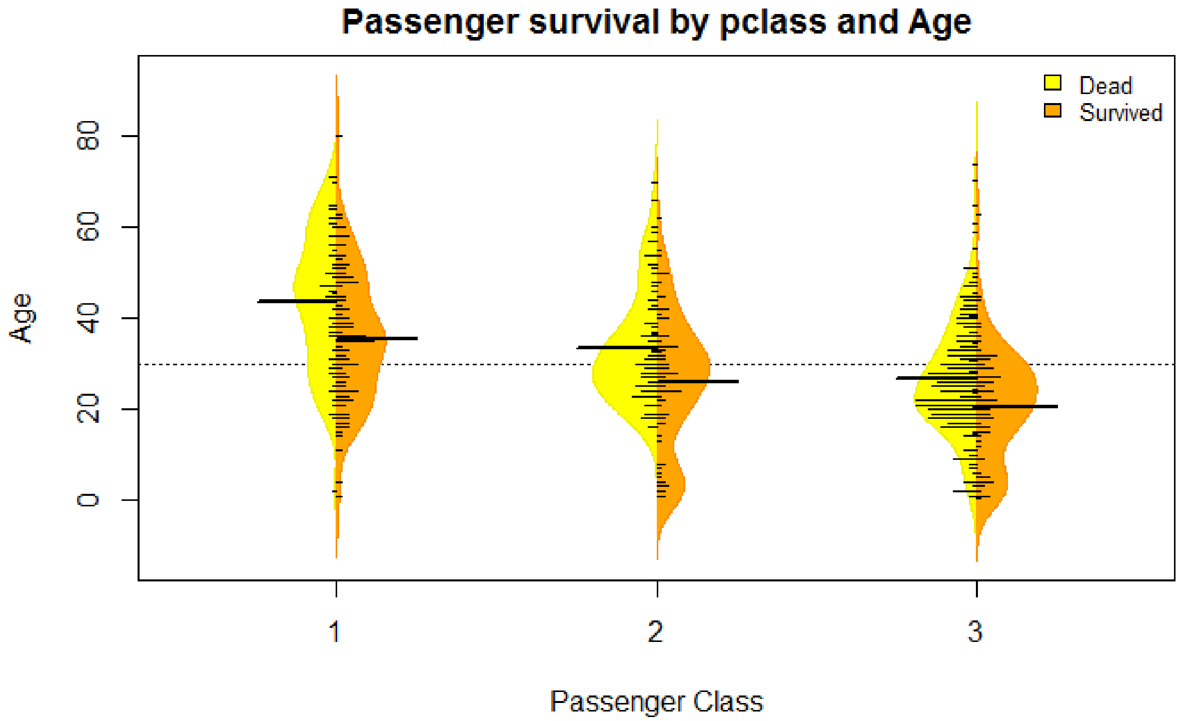
```
 age.predictors <- train.imp  %>%
   dplyr::select(-Survived,-Cabin,-Ticket,-Name)  %>%
   dplyr::filter(complete.cases(.))
ctrl <- trainControl(method = "repeatedcv",
                     repeats = 5)
rpartGrid <- data.frame(maxdepth = seq(2,10,1))
rpartFit_ageimputation <- train(x=age.predictors[,-3],
                      y=age.predictors$Age,
                      method='rpart2',
                      trControl = ctrl,
                      tuneGrid = rpartGrid
                      )
rpartFit_ageimputation
```

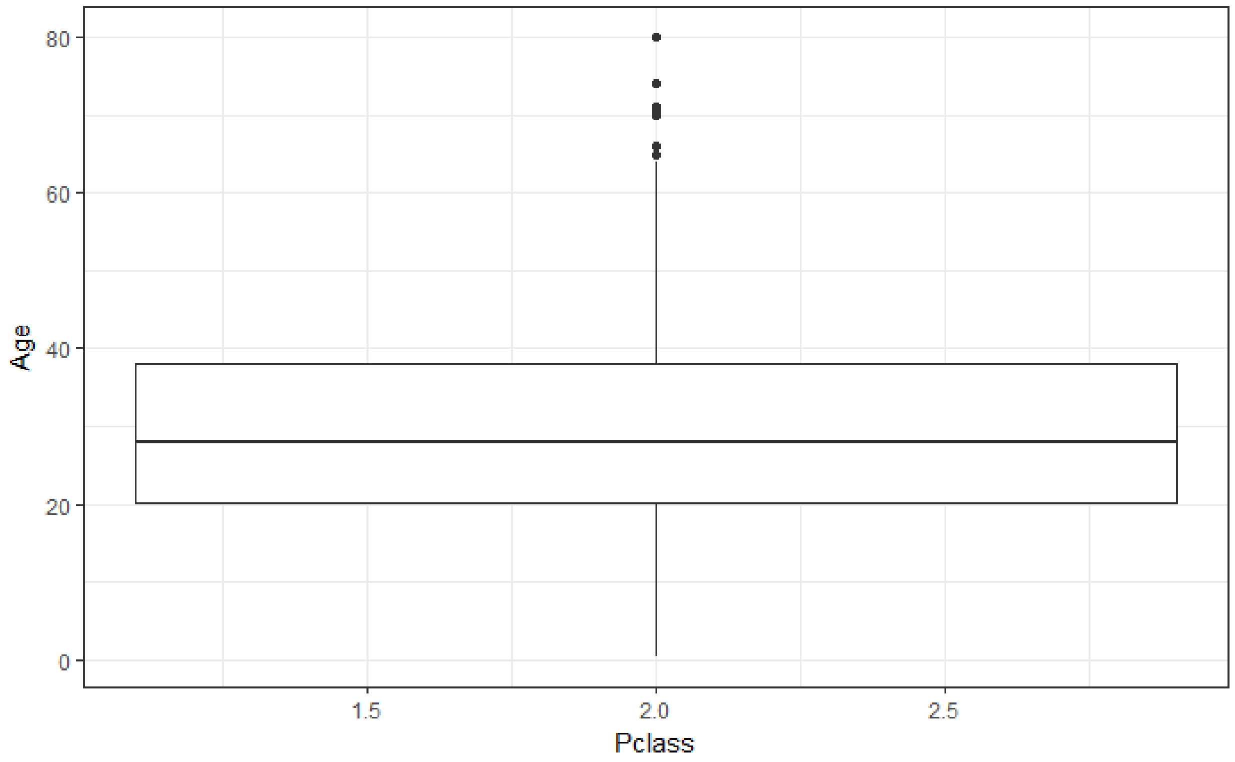
```
## CART
##
## 508 samples
##   7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 457, 457, 457, 457, 457, 457, ...
## Resampling results across tuning parameters:
##
##   maxdepth  RMSE      Rsquared   MAE
##    2        12.02414  0.3171031  9.443687
##    3        11.30498  0.3985131  8.707856
##    4        11.42463  0.3882499  8.782511
##    5        11.27085  0.4038018  8.639549
##    6       11.39825  0.3930011  8.720958
##    7       11.43177  0.3890118  8.744528
##    8       11.47797  0.3851413  8.783542
##    9       11.48005  0.3848860  8.783870
##   10        11.48005  0.3848860  8.783870
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was maxdepth = 5.
plot(rpartFit_ageimputation)
rpart.plot::rpart.plot(rpartFit_ageimputation$finalModel, extra=101, box.palette="GnBu")
save(rpartFit_ageimputation,file = 'rpartFit_ageimputation')
missing_age <- is.na(train.imp$Age)
age.predicted <- predict(rpartFit_ageimputation, newdata = train.imp[missing_age,])
train.imp[missing_age,'Age'] <- age.predicted


train.imp %>%
   mutate(Age_Imputed = missing_age) %>%
   ggplot(aes(x=title,y=Age))+
   stat_summary(aes(y = Age,group=1), fun.y=median, colour="red",
geom="point",group=1)+
   geom_jitter(aes(y=Age,col=Age_Imputed,shape=Age_Imputed))+
   theme_bw()+
   theme(axis.text.x = element_text(angle = 45, hjust = 1),legend.position="none")+
   labs(caption='green points are imputed values')
train.imp$child <- 0
train.imp$child[train.imp$Age<18] <- 1
train.imp$Seniors <- ifelse(train.imp$Age>60,1,0)
train.imp$TotalFam <- train.imp$SibSp + train.imp$Parch + 1
```

~~~
train.imp$LargeFamily <- ifelse(train.imp$TotalFam>4,1,0)

train.imp$Name <- NULL

train.imp$CabinCode <- map_chr(train$Cabin,~str_split(string = .x,pattern = '')[[1]][1])

train.imp$CabinCode[is.na(train.imp$CabinCode)] <- 'U'

train.imp$CabinCode <- as.factor(train.imp$CabinCode)


train.imp$CabinNum <- as.numeric(map_chr(train$Cabin,~str_split(string = .x,pattern =
'[a-zA-Z]')[[1]][2]))

train.imp$CabinNum <- map_int(train.imp$CabinNum, ~as.integer(str_split(.x,pattern =
'',simplify = T)[1][1]))

train.imp$CabinNum[is.na(train.imp$CabinNum)]  <- 0


train.imp$TopDeck <- ifelse(train.imp$CabinCode %in% c('A','B'),1,0)

train.imp$MidDeck <- ifelse(train.imp$CabinCode %in% c('C','D'),1,0)

train.imp$LowerDeck <- ifelse(train.imp$TopDeck==0 & train.imp$MidDeck==0 ,1,0)


train.imp$NumberofCabins <- map_int(train$Cabin,~str_split(string = .x,pattern = ' ')[[1]]
%>% length)

train.imp$Cabin <- NULL

train.imp$Ticket  %>% table() %>% as.numeric() %>% table()

## .

##   1   2   3   4   5   6   7

## 430  60  15   3   1   1   1

train.imp %>% group_by(Pclass) %>% dplyr::select(Ticket,Pclass)  %>% sample_n(5)




```
~~~

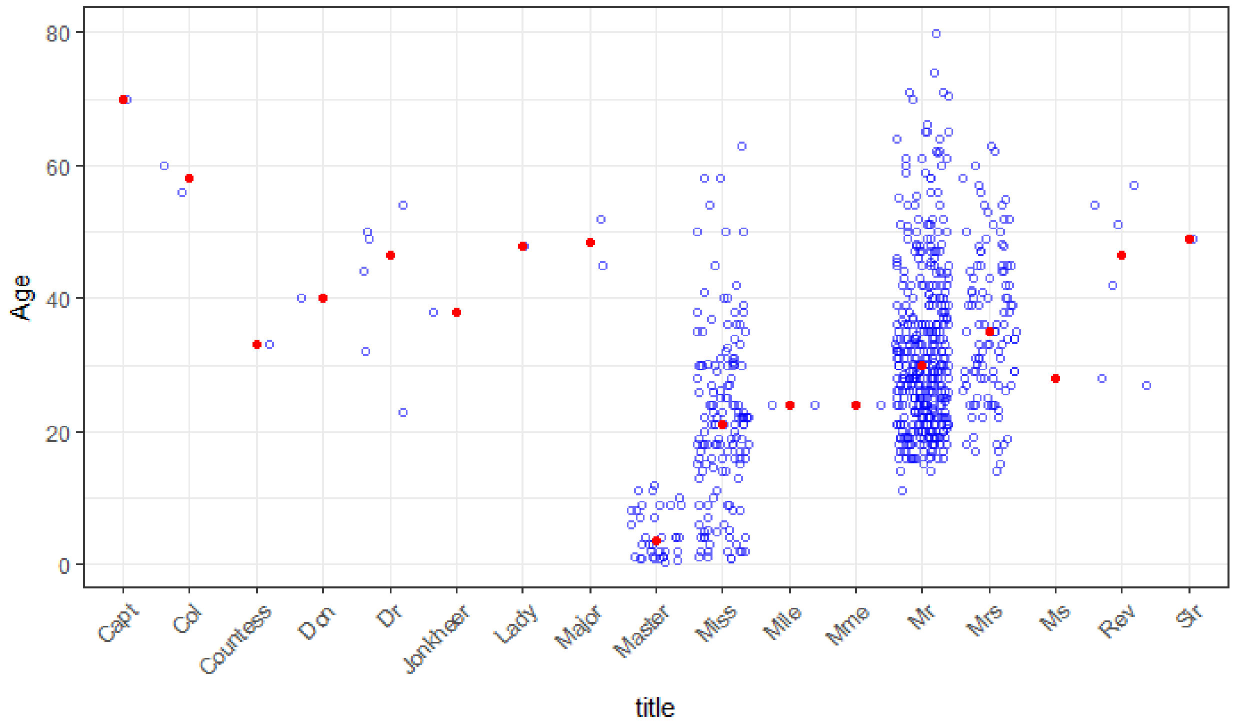


```
ggplot(train,aes(y=Age,x=Pclass))+geom_boxplot(aes(fill=Survived))+theme_bw()
Warning messages:
1: Continuous x aesthetic -- did you forget aes(group=...)?
2: Removed 177 rows containing non-finite values (stat_boxplot).
> beanplot(Age~Survived*Pclass,side='b',train,col=list('yellow','orange'),
+          border = c('yellow2','darkorange'),ll = 0.05,boxwex = .5,
+          main='Passenger survival by pclass and Age',xlab='Passenger Class'
,ylab='Age')
> legend('topright', fill = c('yellow','orange'), legend = c("Dead", "Survive
d"),bty = 'n',cex = .8)
```



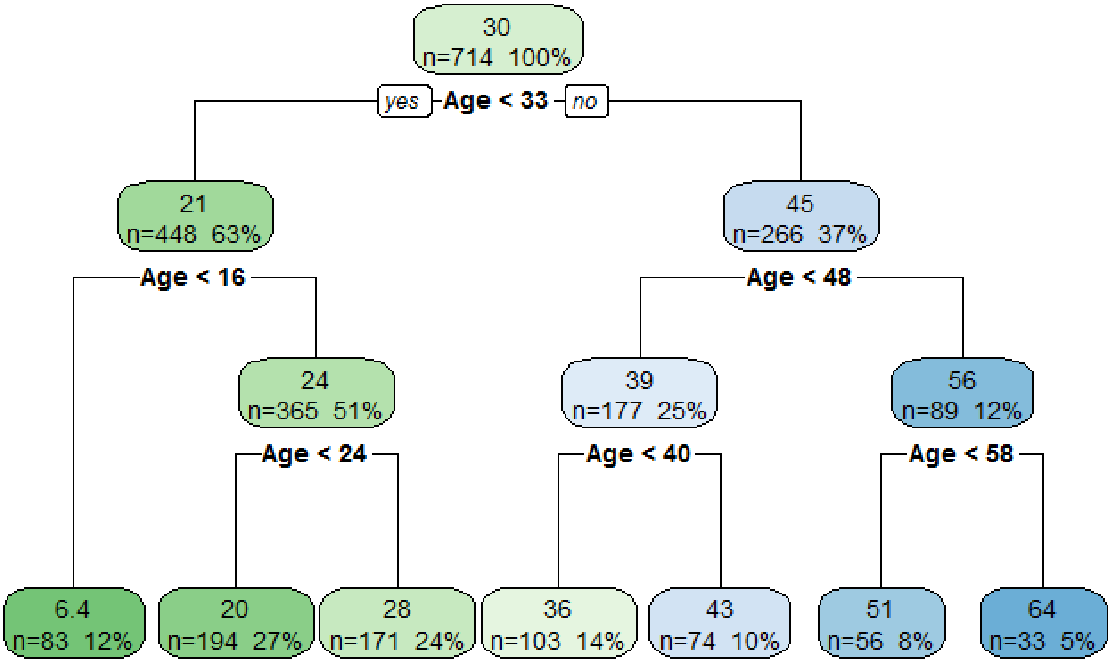


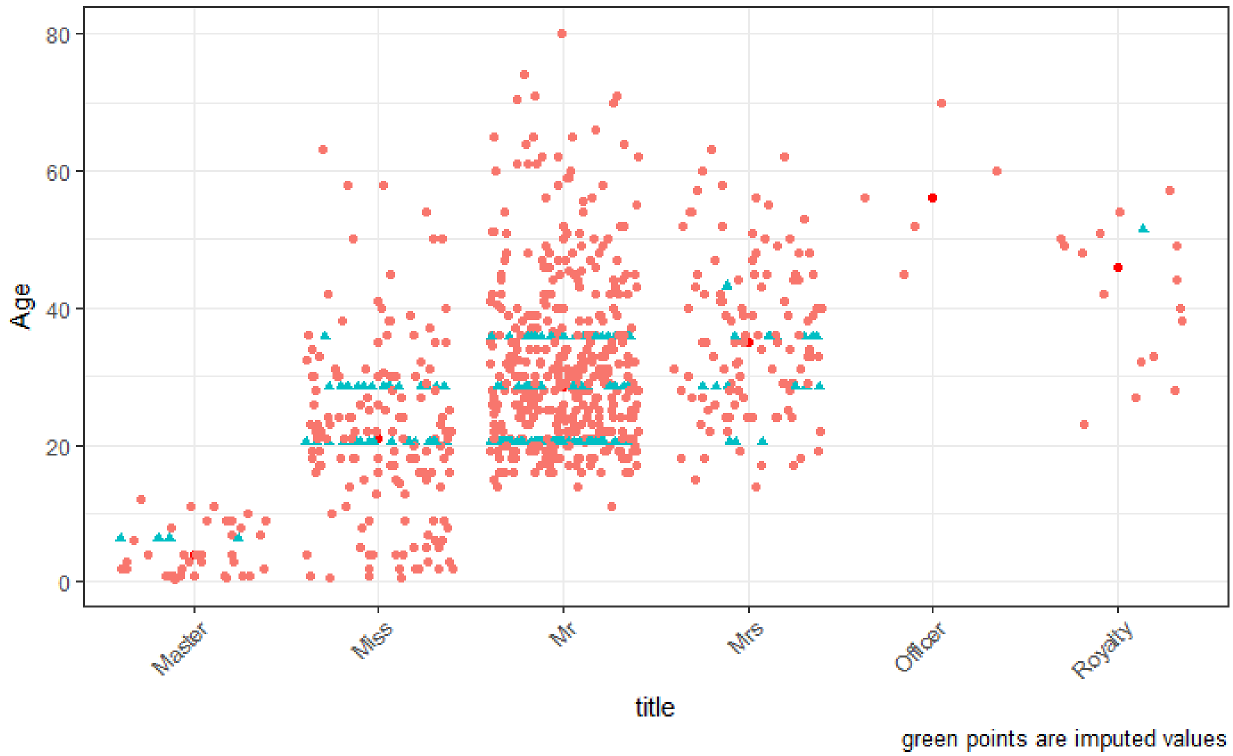
```
stat_summary(aes(y  = Age,group=1),  fun.y=median,  colour="red",  geom="point",group=1)+
theme_bw()+  theme(axis.text.x  = element_text(angle  =  45,  hjust  =
1),legend.position="none")+  labs(caption='red  points  are  median  values')
```



red points are median values

```
## CART
##
## 508 samples
##   7 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 457, 457, 457, 457, 457, 457, ...
## Resampling results across tuning parameters:
##
##   maxdepth  RMSE      Rsquared   MAE
##    2        12.02414  0.3171031  9.443687
##    3        11.30498  0.3985131  8.707856
##    4        11.42463  0.3882499  8.782511
##    5        11.27085  0.4038018  8.639549
##    6        11.39825  0.3930011  8.720958
##    7        11.43177  0.3890118  8.744528
##     8          11.47797  0.3851413  8.783542
##     9          11.48005  0.3848860  8.783870
##    10          11.48005  0.3848860  8.783870
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was maxdepth = 5.
plot(rpartFit_ageimputation)
rpart.plot::rpart.plot(rpartFit_ageimputation$finalModel, extra=101, box.pale
tte="GnBu")
save(rpartFit_ageimputation,file = 'rpartFit_ageimputation')
```





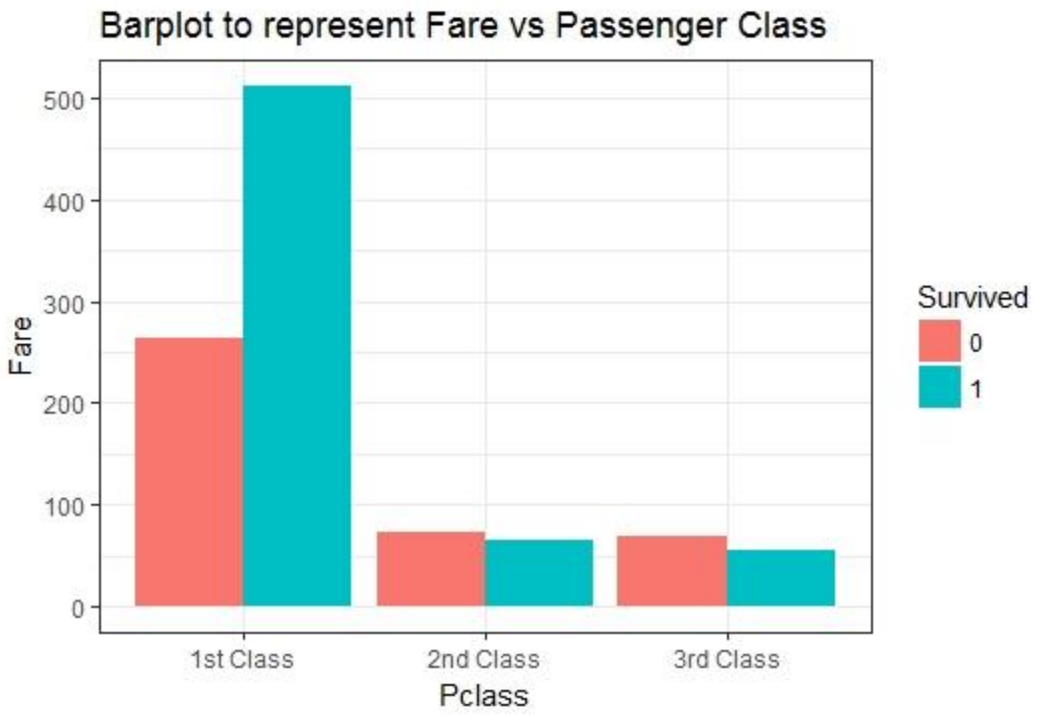
Ticket <chr>	Pclass <int>
113767	1
17421	1
PC 17582	1
113510	1
13507	1
S.O.C. 14879	2
244373	2
239853	2
C.A. 31921	2
236853	2

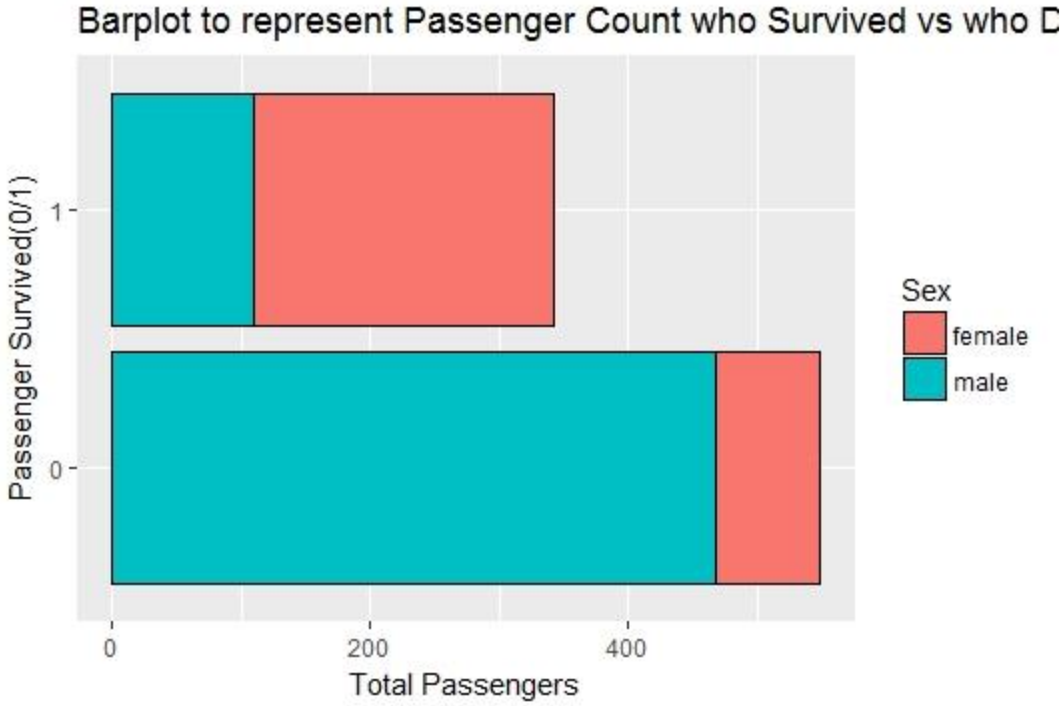
Next
12
Previous
1-10 of 15 rows

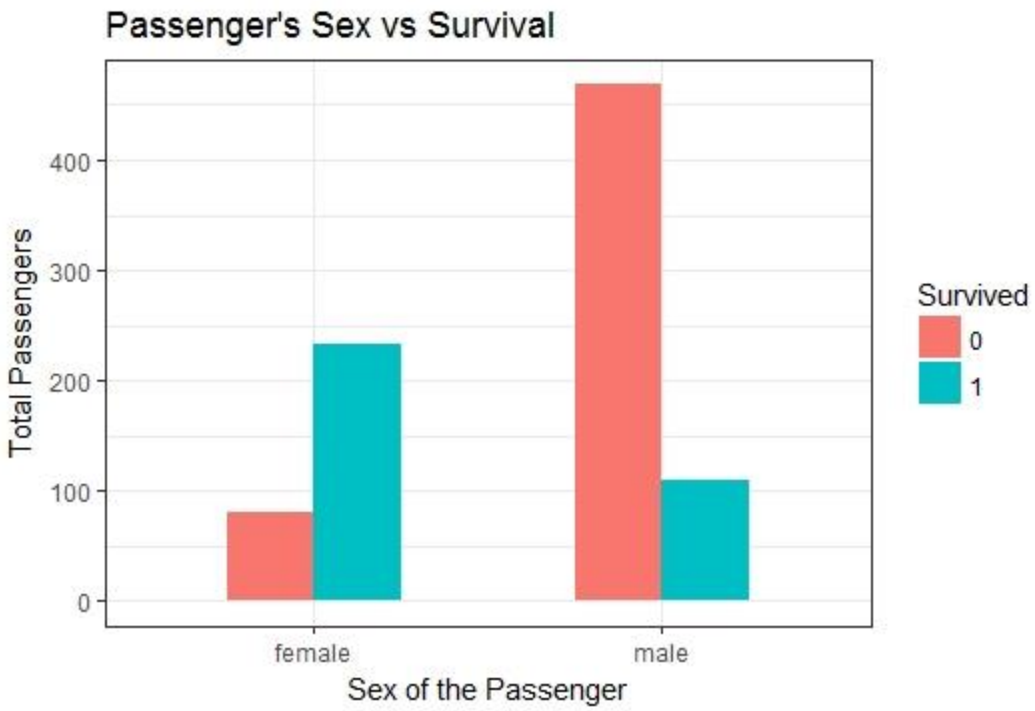
Ticket <chr>	Pclass <int>
2665	3
2691	3
A/4 48871	3
349204	3
349248	3

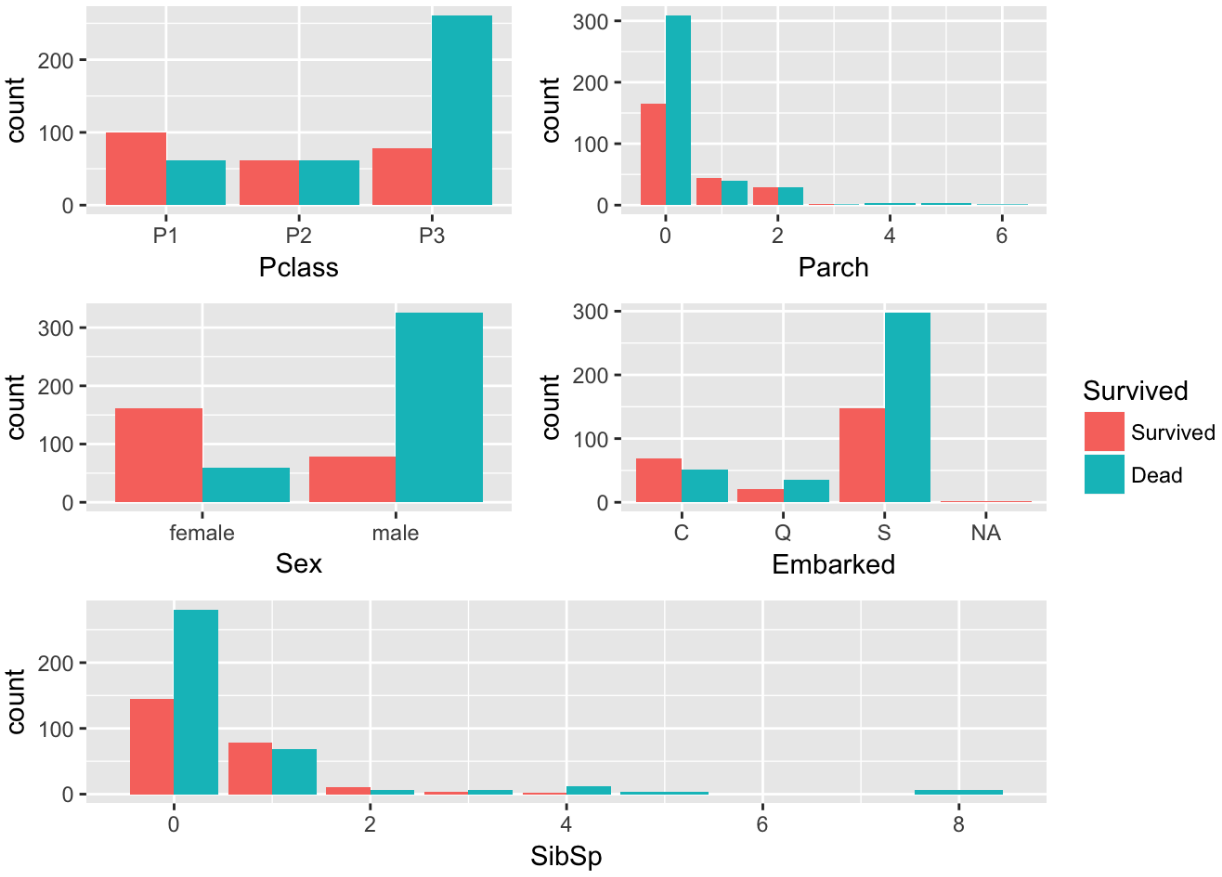
b. Represent the proportion of people survived from the family size using a graph.

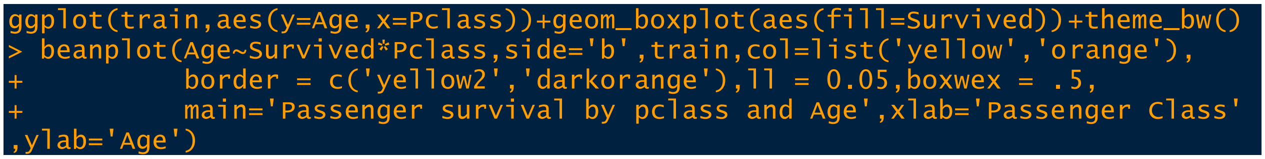
```
get_legend<-function(myggplot){
  tmp <- ggplot_gtable(ggplot_build(myggplot))
  leg <- which(sapply(tmp$grobs, function(x) x$name) == "guide-box")
  legend <- tmp$grobs[[leg]]
  return(legend)
}p <- lapply(X = c('Pclass','Sex','SibSp','Parch','Embarked'),
      FUN = function(x) ggplot(data = train)+
        aes_string(x=x,fill='Survived')+
        geom_bar(position="dodge")+
        theme(legend.position="none"))
legend <- get_legend(ggplot(data = train,aes(x=Pclass,fill=Survived))+geom_bar())
grid.arrange(p[[1]],p[[2]],p[[3]],p[[4]],p[[5]], legend,layout_matrix = cbind(c(1,2,3),
c(4,5,3),   c(6,6,6)), widths=c(3,3,1))
```


Barplot to represent Fare vs Passenger Class

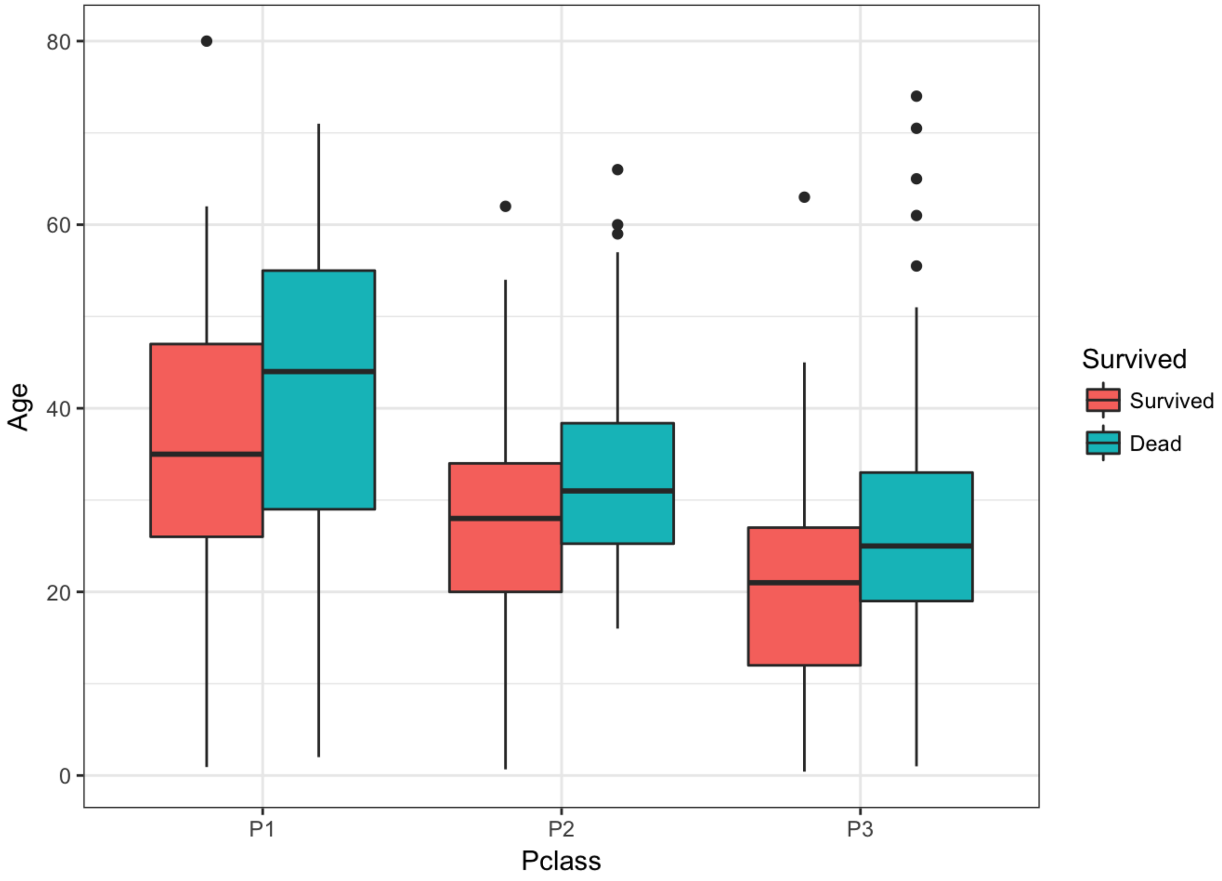

Barplot to represent Passenger Count who Survived vs who D


Passenger's Sex vs Survival
Total Passengers
400
300
200
100
0
female
male
Sex of the Passenger
Survived
0
1

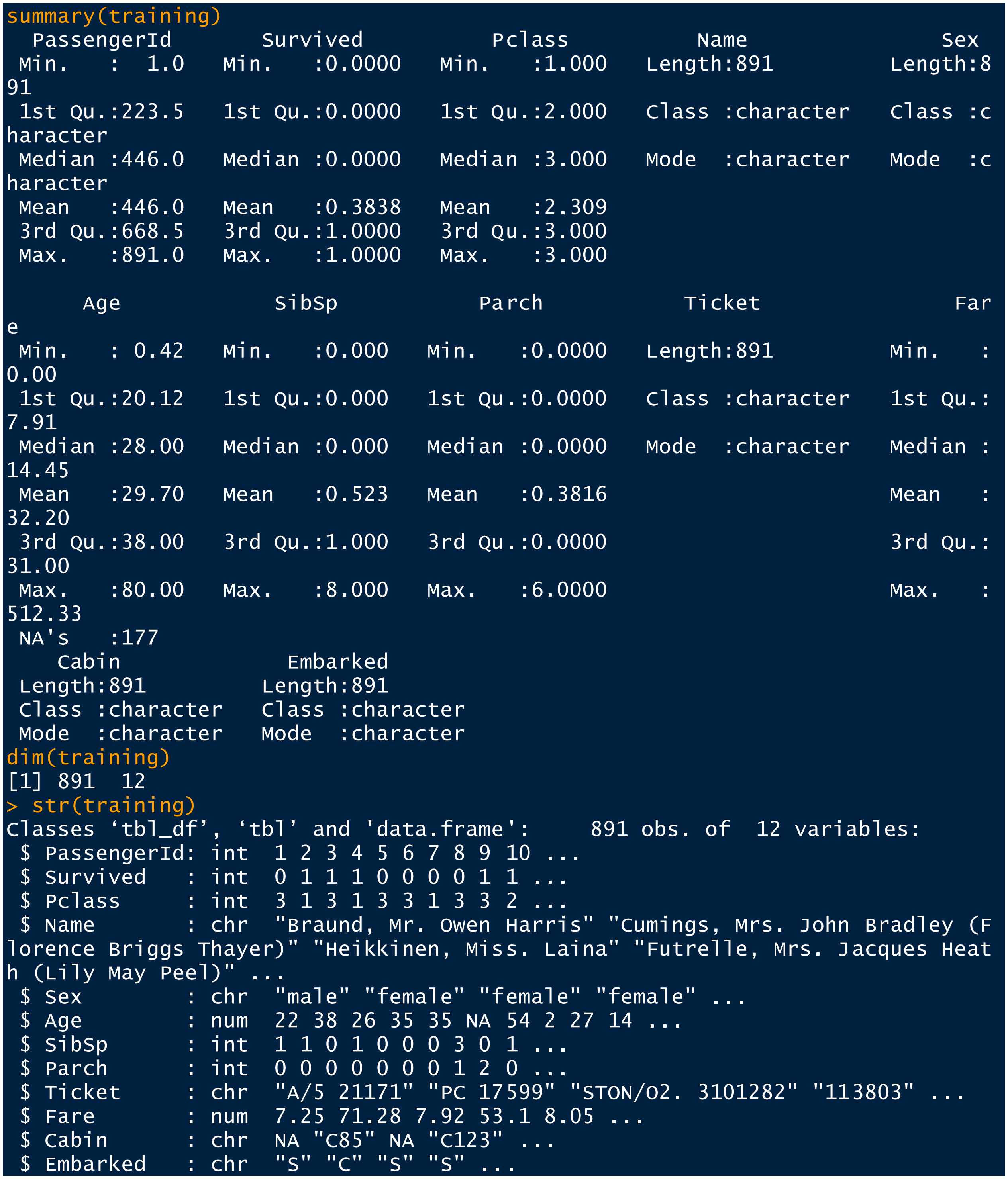

count
200
100
0
P1
P2
P3
Pclass
count
300
200
100
0
0
2
4
6
Parch
count
300
200
100
0
female
male
Sex
count
300
200
100
0
C
Q
S
NA
Embarked
Survived
Survived
Dead
count
200
100
0
0
2
4
6
8
SibSp



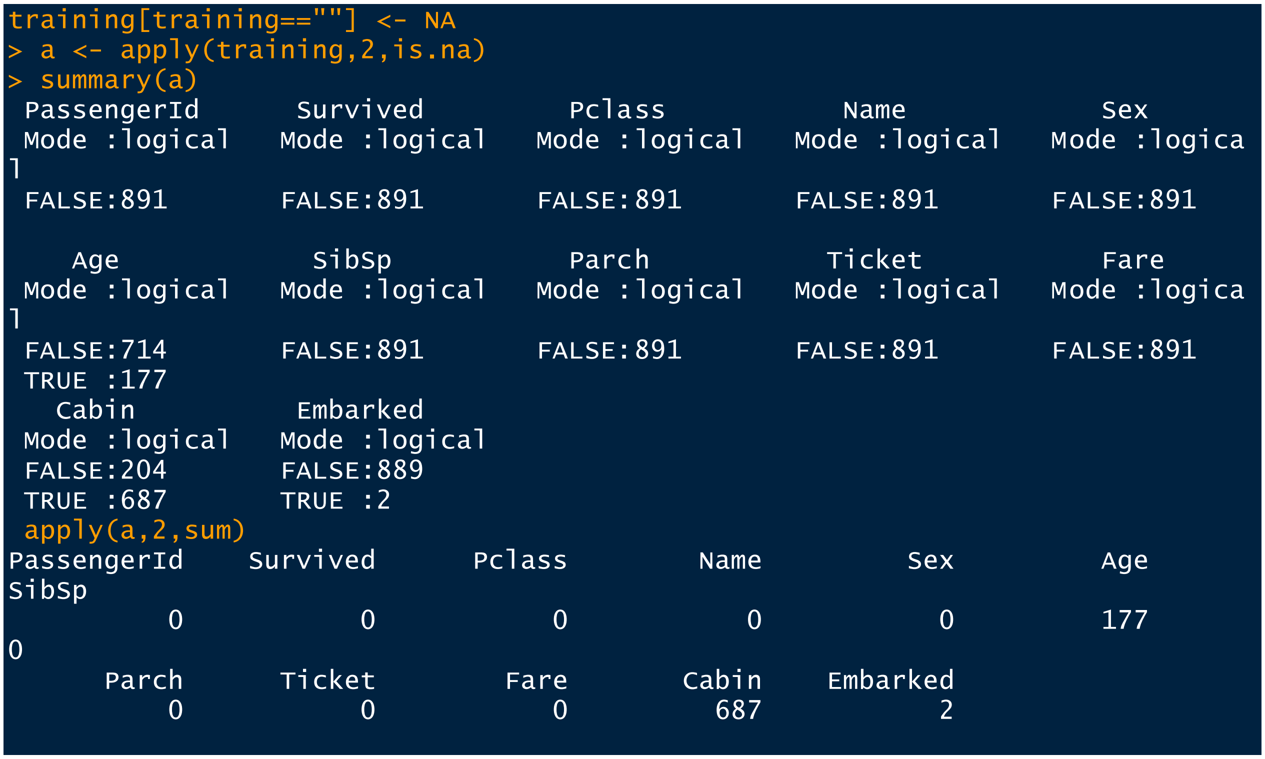
```
ggplot(train,aes(y=Age,x=Pclass))+geom_boxplot(aes(fill=Survived))+theme_bw()
> beanplot(Age~Survived*Pclass,side='b',train,col=list('yellow','orange'),
+          border = c('yellow2','darkorange'),ll = 0.05,boxwex = .5,
+          main='Passenger survival by pclass and Age',xlab='Passenger Class'
,ylab='Age')
```



c. Impute the missing values in Age variable using Mice Library, create two different graphs showing Age distribution before and after imputation

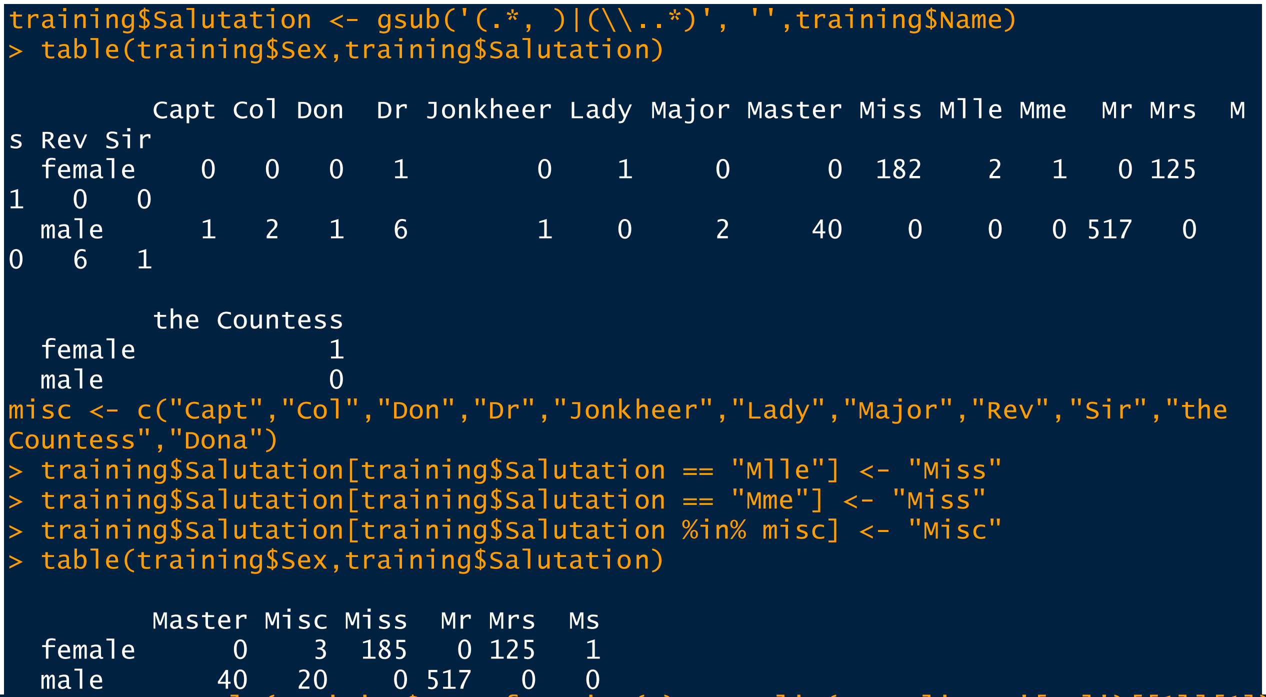


```
summary(training)
  PassengerId       Survived          Pclass           Name                Sex
 Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891         Length:8
91
 1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character   Class :c
haracter
 Median :446.0   Median :0.0000   Median :3.000   Mode  :character   Mode  :c
haracter
 Mean   :446.0   Mean   :0.3838   Mean   :2.309
 3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
 Max.   :891.0   Max.   :1.0000   Max.   :3.000

      Age            SibSp           Parch            Ticket                Far
e
 Min.   : 0.42   Min.   :0.000   Min.   :0.0000   Length:891         Min.   :
0.00
 1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000   Class :character   1st Qu.:
7.91
 Median :28.00   Median :0.000   Median :0.0000   Mode  :character   Median :
14.45
 Mean   :29.70   Mean   :0.523   Mean   :0.3816                      Mean   :
32.20
 3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000                      3rd Qu.:
31.00
 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
dim(training)
[1] 891  12
> str(training)
Classes ‘tbl_df’, ‘tbl’ and '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 (F
lorence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heat
h (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  NA "C85" NA "C123" ...
 $ Embarked   : chr  "S" "C" "S" "S" ...
```

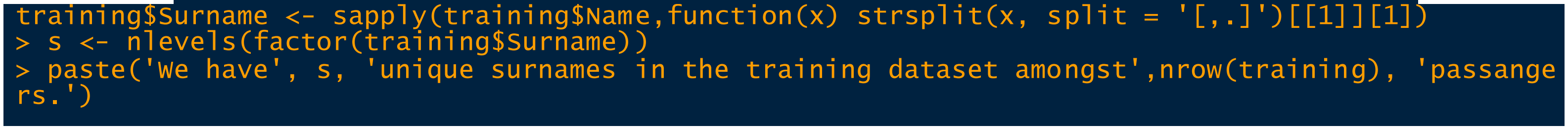


```
training[training==""] <- NA
> a <- apply(training,2,is.na)
> summary(a)
 PassengerId      Survived         Pclass           Name             Sex
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logica
l
 FALSE:891       FALSE:891       FALSE:891       FALSE:891       FALSE:891

    Age            SibSp           Parch           Ticket           Fare
 Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logica
l
 FALSE:714       FALSE:891       FALSE:891       FALSE:891       FALSE:891
 TRUE :177
   Cabin          Embarked
 Mode :logical   Mode :logical
 FALSE:204       FALSE:889
 TRUE :687       TRUE :2
 apply(a,2,sum)
PassengerId    Survived      Pclass        Name         Sex         Age
SibSp
          0           0           0           0           0         177
0
      Parch      Ticket        Fare       Cabin    Embarked
          0           0           0         687           2
```

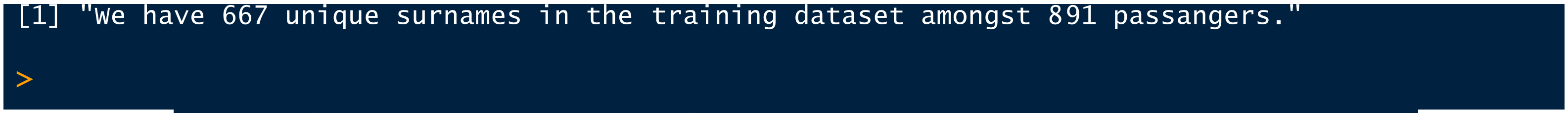
It can be seen that Age,Cabin and Embarked variables have missing values. Cabin has most number of missing values.These missing values can be found by using ‘Multivariate Imputation by Chained Equations (MICE)’ package



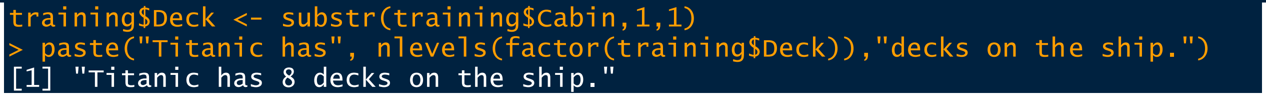
```
training$Salutation <- gsub('(.*, )|(\\..*)', '',training$Name)
> table(training$Sex,training$Salutation)

         Capt Col Don  Dr Jonkheer Lady Major Master Miss Mlle Mme  Mr Mrs   M
s Rev Sir
  female    0   0   0   1        0    1     0      0  182    2   1   0 125
1   0   0
  male      1   2   1   6        1    0     2     40    0    0   0 517   0
0   6   1

         the Countess
  female            1
  male              0
misc <- c("Capt","Col","Don","Dr","Jonkheer","Lady","Major","Rev","Sir","the
Countess","Dona")
> training$Salutation[training$Salutation == "Mlle"] <- "Miss"
> training$Salutation[training$Salutation == "Mme"] <- "Miss"
> training$Salutation[training$Salutation %in% misc] <- "Misc"
> table(training$Sex,training$Salutation)

         Master Misc Miss   Mr Mrs   Ms
  female      0    3  185    0 125    1
  male       40   20    0 517    0    0
```



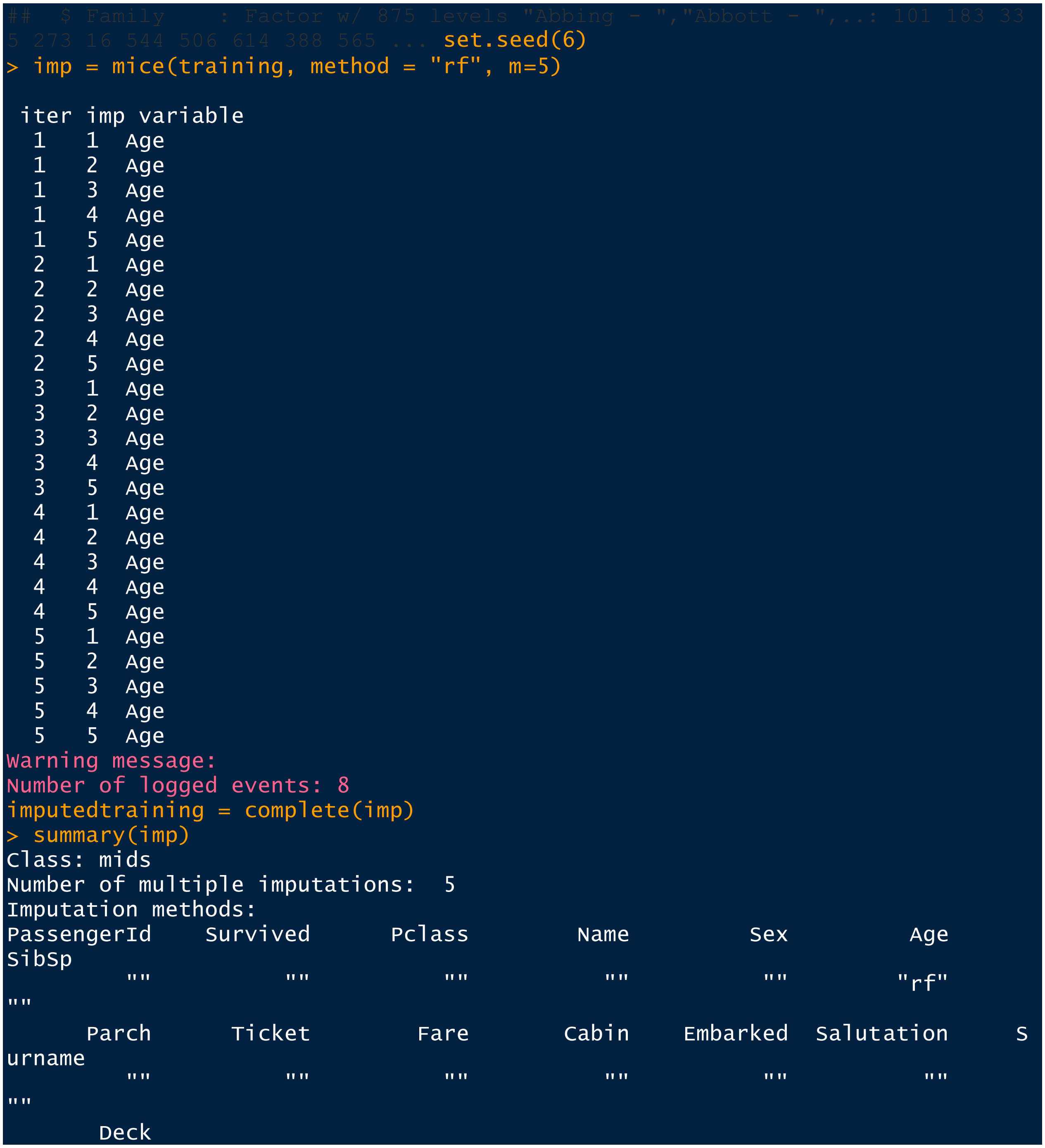
```
training$Surname <- sapply(training$Name,function(x) strsplit(x, split = '[,.]')[[1]][1])
> s <- nlevels(factor(training$Surname))
> paste('We have', s, 'unique surnames in the training dataset amongst',nrow(training), 'passange
rs.')
```



```
[1] "We have 667 unique surnames in the training dataset amongst 891 passangers."

>
```



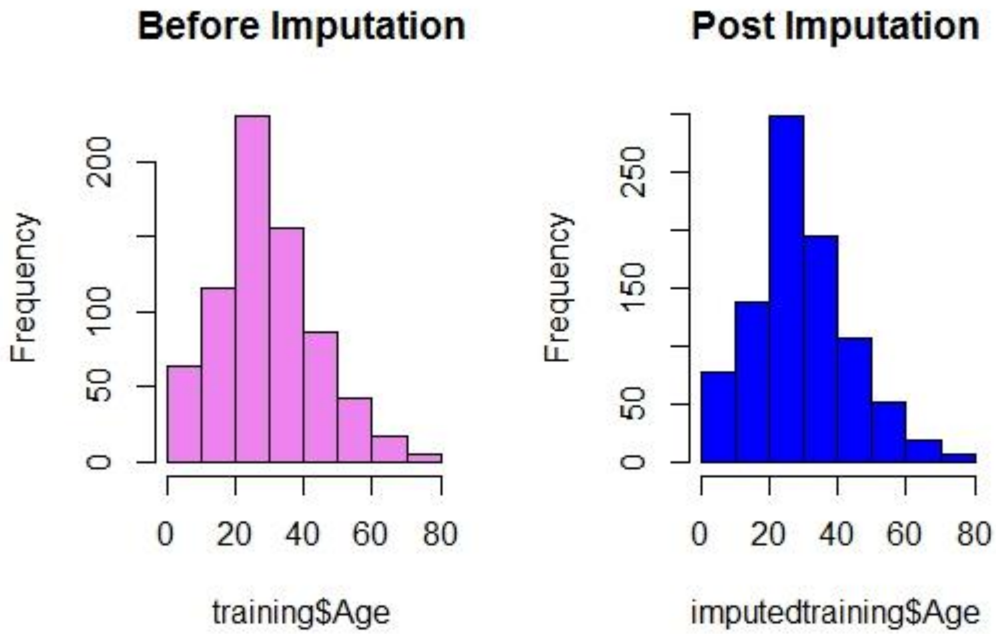
```
training$Deck <- substr(training$Cabin,1,1)
> paste("Titanic has", nlevels(factor(training$Deck)),"decks on the ship.")
[1] "Titanic has 8 decks on the ship."
```

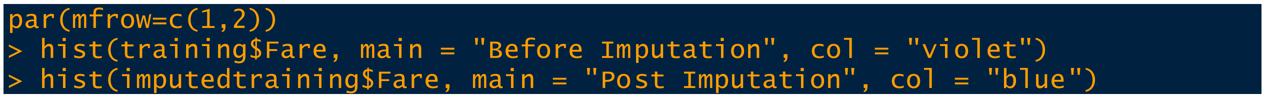


```
##  $ Family     : Factor w/ 875 levels "Abbing - ","Abbott - ",..: 101 183 33
5 273 16 544 506 614 388 565 ... set.seed(6)
> imp = mice(training, method = "rf", m=5)

 iter imp variable
  1   1  Age
  1   2  Age
  1   3  Age
  1   4  Age
  1   5  Age
  2   1  Age
  2   2  Age
  2   3  Age
  2   4  Age
  2   5  Age
  3   1  Age
  3   2  Age
  3   3  Age
  3   4  Age
  3   5  Age
  4   1  Age
  4   2  Age
  4   3  Age
  4   4  Age
  4   5  Age
  5   1  Age
  5   2  Age
  5   3  Age
  5   4  Age
  5   5  Age
Warning message:
Number of logged events: 8
imputedtraining = complete(imp)
> summary(imp)
Class: mids
Number of multiple imputations:  5
Imputation methods:
PassengerId    Survived      Pclass        Name         Sex         Age
SibSp
         ""          ""          ""          ""          ""        "rf"
""
      Parch      Ticket        Fare       Cabin    Embarked  Salutation     S
urname
         ""          ""          ""          ""          ""          ""
""
       Deck
```

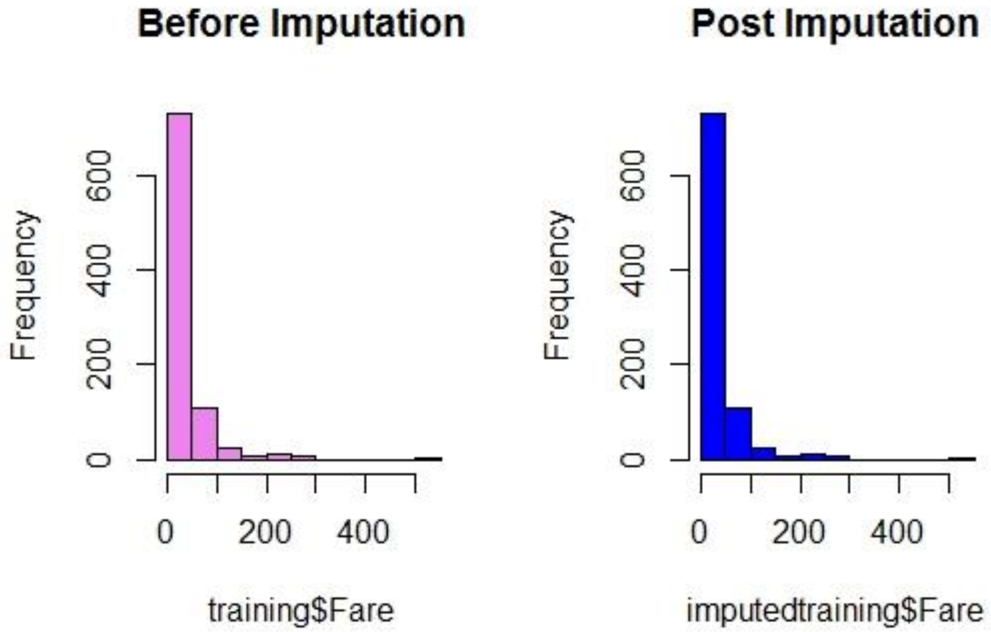
```
         ""
PredictorMatrix:
            PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket Fare
Cabin Embarked
PassengerId           0        1      1    0   0   1     1     1      0    1
0        0
Survived              1        0      1    0   0   1     1     1      0    1
0        0
Pclass                1        1      0    0   0   1     1     1      0    1
0        0
Name                  1        1      1    0   0   1     1     1      0    1
0        0
Sex                   1        1      1    0   0   1     1     1      0    1
0        0
Age                   1        1      1    0   0   0     1     1      0    1
0        0
            Salutation Surname Deck
PassengerId          0       0    0
Survived             0       0    0
Pclass               0       0    0
Name                 0       0    0
Sex                  0       0    0
Age                  0       0    0
Number of logged events:  8
  it im dep     meth        out
1  0  0     constant       Name
2  0  0     constant        Sex
3  0  0     constant     Ticket
4  0  0     constant      Cabin
5  0  0     constant   Embarked
6  0  0     constant Salutation
apply(apply(imputedtraining,2,is.na),2,sum)
PassengerId    Survived      Pclass        Name         Sex         Age
SibSp
          0           0           0           0           0           0
0
      Parch      Ticket        Fare       Cabin    Embarked  Salutation     S
urname
          0           0           0         687           2           0
0
       Deck
        687
```

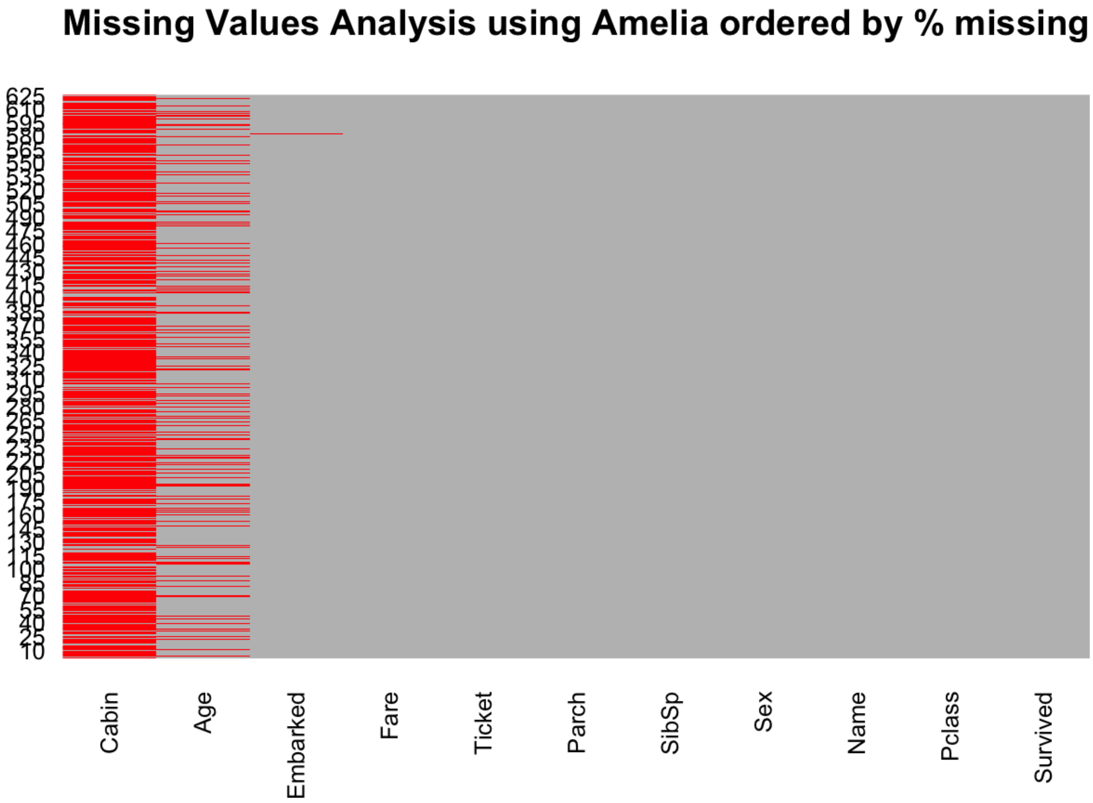
```
par(mfrow=c(1,2))
>
> hist(training$Age, main = "Before Imputation", col = "violet")
> hist(imputedtraining$Age, main = "Post Imputation", col = "blue")
```





```
par(mfrow=c(1,2))
> hist(training$Fare, main = "Before Imputation", col = "violet")
> hist(imputedtraining$Fare, main = "Post Imputation", col = "blue")
```





```
#Missing cases (numbers):

map_int(train.raw,~sum(is.na(.x)))

## Survived    Pclass      Name       Sex       Age     SibSp     Parch    Ticket

##        0         0         0         0       117         0         0         0

##     Fare     Cabin  Embarked

##        0       478         2
```

Cabin has a large number of missing values (77% missing). Imputing this variable may prove challenging or even useless. Age (19.9% missing) and Embarked (0.2%) missi

Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
{r cars} summary(cars)
```

Including Plots

You can also embed plots, for example:

```
{r pressure, echo=FALSE} plot(pressure)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.