DATA ANALYTICS WITH R, EXCEL AND TABLAEU

ASSIGNMENT 6.1 ANSWERS

By ASHISH S SHANBHAG

ashishshanbhag108@gmail.com

Question no:

5)

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

Perform the following:

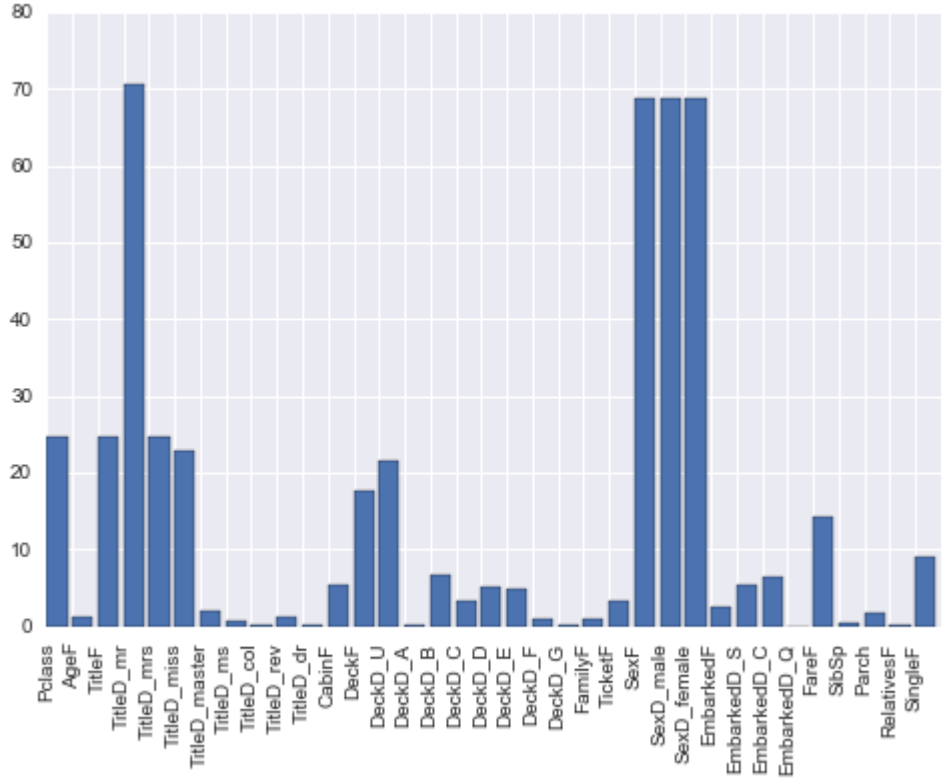
a.Preprocess the passenger names to come up with a list of titles that represent families and represent using appropriate visualization graph.

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.

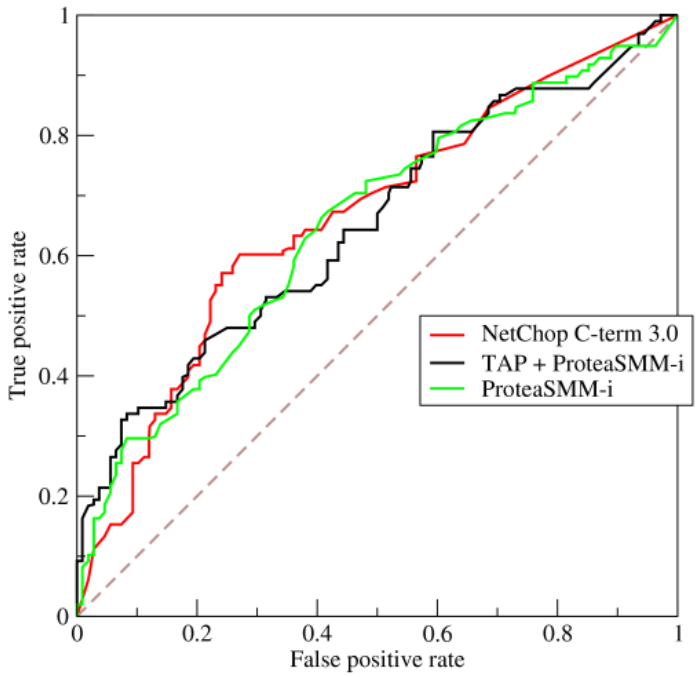
Ans:

```
from sklearn.feature_selection import SelectKBest
selector = SelectKBest(k=5)
selector.fit(train_data_munged[predictors], train_data_munged["Survived"])

scores = -np.log10(selector.pvalues_)

plt.bar(range(len(predictors)), scores)
plt.xticks(range(len(predictors)), predictors, rotation='vertical')
plt.show()
```



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

Ans



```
tempData <- mice(data,m=5,maxit=50,meth='pmm',seed=500)
summary(tempData)

Multiply imputed data set
Call:
mice(data = data, m = 5, method = "pmm", maxit = 50, seed = 500)
Number of multiple imputations:  5
Missing cells per column:
  Ozone Solar.R    Wind    Temp
     37       7       7       5
Imputation methods:
  Ozone Solar.R    Wind    Temp
  "pmm"   "pmm"   "pmm"   "pmm"
VisitSequence:
  Ozone Solar.R    Wind    Temp
      1       2       3       4
PredictorMatrix:
        Ozone Solar.R Wind Temp
Ozone       0       1    1    1
Solar.R     1       0    1    1
Wind        1       1    0    1
Temp        1       1    1    0
Random generator seed value:  500
```