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.

Ans:

# Load packages

**library**('ggplot2')

**library**('ggthemes')

**library**('scales')

**library**('dplyr')

**library**('mice')

**library**('randomForest')

train <- **read.csv**('C:\Users\jai sriram\Desktop\Acadgild/train.csv', stringsAsFactors = F)

test <- **read.csv**('C:\Users\jai sriram\Desktop\Acadgild/test.csv', stringsAsFactors = F)

full <- **bind\_rows**(train, test) *# bind training & test data*

**str**(full)

full**$**Title <- **gsub**('(.\*, )|(\\..\*)', '', full**$**Name)

**table**(full**$**Sex, full**$**Title)

rare\_title <- **c**('Dona', 'Lady', 'the Countess','Capt', 'Col', 'Don', 'Dr', 'Major', 'Rev', 'Sir', 'Jonkheer')

**table**(full**$**Sex, full**$**Title)

full**$**Surname <- **sapply**(full**$**Name,

**function**(x) **strsplit**(x, split = '[,.]')[[1]][1])

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

Ans:

full**$**Fsize <- full**$**SibSp **+** full**$**Parch **+** 1

full**$**Family <- **paste**(full**$**Surname, full**$**Fsize, sep='\_')

**ggplot**(full[1**:**891,], **aes**(x = Fsize, fill = **factor**(Survived))) **+**

**geom\_bar**(stat='count', position='dodge') **+**

**scale\_x\_continuous**(breaks=**c**(1**:**11)) **+**

**labs**(x = 'Family Size') **+**

**theme\_few**()

c. Impute the missing values in Age variable using Mice Library, create two different

graphs showing Age distribution before and after imputation.

Ans:

**sum**(**is.na**(full**$**Age))

factor\_vars <- **c**('PassengerId','Pclass','Sex','Embarked',

'Title','Surname','Family','FsizeD')

full[factor\_vars] <- **lapply**(full[factor\_vars], **function**(x) **as.factor**(x))

mice\_mod <- **mice**(full[, **!names**(full) **%in%** **c**('PassengerId','Name','Ticket','Cabin','Family','Surname','Survived')], method='rf')

mice\_output <- **complete**(mice\_mod)

**# Plot age distributions**

**par**(mfrow=**c**(1,2))

**hist**(full**$**Age, freq=F, main='Age: Original Data',

col='darkgreen', ylim=**c**(0,0.04))

**hist**(mice\_output**$**Age, freq=F, main='Age: MICE Output',

col='lightgreen', ylim=**c**(0,0.04))