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.

*# Load packages*

**library**('ggplot2') *# visualization*

**library**('ggthemes') *# visualization*

**library**('scales') *# visualization*

**library**('dplyr') *# data manipulation*

**library**('mice') *# imputation*

**library**('randomForest') *# classification algorithm*

train <- **read.csv**('../input/train.csv', stringsAsFactors = F)

test <- **read.csv**('../input/test.csv', stringsAsFactors = F)

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

*# check data*

**str**(full)

*# Grab title from passenger names*

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

*# Show title counts by sex again*

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

*# Finally, grab surname from passenger name*

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

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

**cat**(**paste**('We have <b>', **nlevels**(**factor**(full**$**Surname)), '</b> unique surnames. I would be interested to infer ethnicity based on surname --- another time.'))

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

setwd("C:/Users/sriram.sampath/Desktop/Kaggle/Titanic")

train <- read.csv("train.csv")

colnames(train)

table(train$Survived)

table(train$Pclass)

train$Pclass <- as.factor(train$Pclass) table(train$Pclass,train$Survived)

train$Survived <- as.factor(train$Survived) plot(train$Pclass,train$Survived)

train$Sex <- as.factor(train$Sex) table(train$Sex)

table(train$Sex, train$Survived)

plot(train$Sex, train$Survived)

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

graphs showing Age distribution before and after imputation.

**library**(Amelia)

**library**(mice)

**library**(ggplot2)

**library**(lattice)

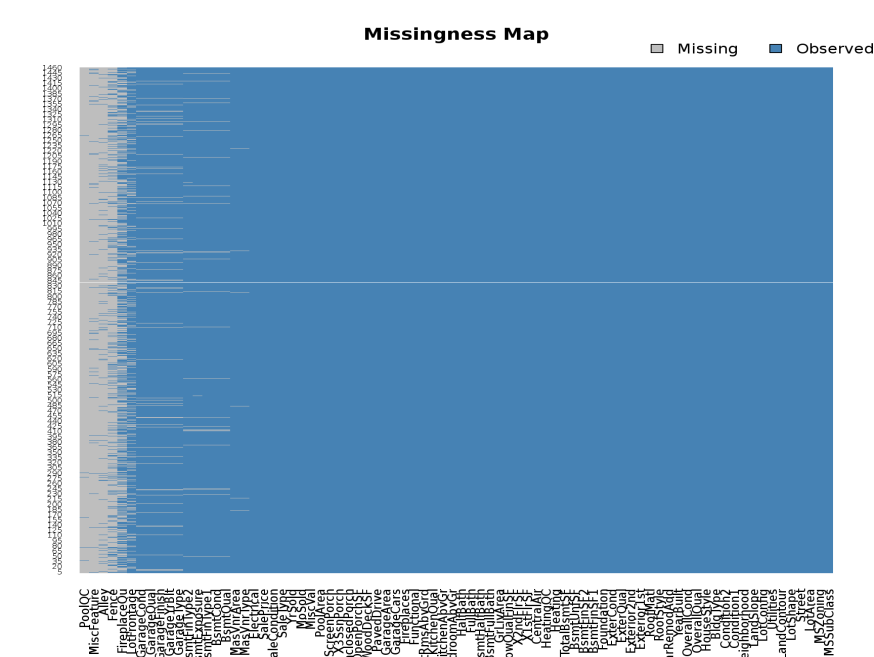
train\_raw <- read.csv('../input/train.csv')

train\_raw$MSSubClass <- as.factor(train\_raw$MSSubClass)

train\_raw$MoSold <- as.factor(train\_raw$MoSold)

train\_raw$YrSold <- as.factor(train\_raw$YrSold)

missmap(train\_raw[-1], col=c('grey', 'steelblue'), y.cex=0.5, x.cex=0.8)

**

*# Let's also get some hard numbers*

sort(sapply(train\_raw, **function**(x) { sum(is.na(x)) }), decreasing=TRUE)