Data Processing

Shannon Rush

July 10, 2014

This document details the steps taken to process the Titanic Survival dataset obtained from the Kaggle Titanic competition into the data needed for this Shiny application.

Load the raw data obtained from Kaggle

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

Create new dataframe using just columns needed for the Shiny application, with tidier column names

```
new.data <- data.frame(survived=data$Survived,
                       class=data$Pclass,
                       gender=data$Sex,
                       age=data$Age)
```

Keep only observations that include data for each selected feature

```
titanic.data <- new.data[complete.cases(new.data),]
```

Fractional ages are not necessary for this analysis, round ages to nearest integer

```
titanic.data$age <- round(titanic.data$age)
```

Revalue survival factor for easier labeling

```
library(plyr)
titanic.data$survived <- as.character(titanic.data$survived)
titanic.data$survived <- revalue(titanic.data$survived, c("0"="Did Not Survive","1"="Survived"))
```

Save as .RData file for easy loading

```
save(titanic.data, file="titanic_data.RData")
```