Data Wrangling Exercise 2

Swati Jani Joshi

June 17, 2016

Titanic

R markdown containing the complete code for the Titanic data set in exercise 2.

Load Data

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
df <- read.csv("~/Documents/Data Wrangling/DW_2/titanic_original.csv")
```

1. Port of Embarkation

```
is.na(df$embarked) <- df$embarked == ""
df$embarked[is.na(df$embarked)] <- "S"
```

2. Age

a.

```
age_mean <- mean(df$age, na.rm = TRUE)
df$age[is.na(df$age)] <- age_mean
```

b. Other ways the missing values could have been populated by using the median value or the mode value instead of the mean value.

3. Lifeboat

```
is.na(df$boat) <- df$boat == ""
df$boat <- as.character(df$boat)
df$boat[is.na(df$boat)] <- "None"
```

4. Cabin

It only makes sense to fill in the missing cabin numbers with NA.

A missing value for cabin numbers most likely means that the person did not survive.

```
df$has_cabin_number <- ifelse(df$cabin == '', 0, 1)
```

Clean data file

```
write.csv(df, file = "titanic_clean.csv")
```