Data Science Exercise 2

Kathryn Gadberry

April 6, 2016

Project Set-Up

```
titanic_clean <- read.csv("titanic_clean.csv")
titanic_original <- read.csv("titanic_original.csv")
glimpse(titanic_clean)
```

```
## Observations: 1,310
## Variables: 16
## $ X         (int) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1...
## $ pclass    (int) 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ survived  (int) 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1...
## $ name      (fctr) Allen, Miss. Elisabeth Walton, Allison, Master. Hud...
## $ sex       (fctr) female, male, female, male, female, male, female, m...
## $ age       (dbl) 29.0000, 0.9167, 2.0000, 30.0000, 25.0000, 48.0000, ...
## $ sibsp     (int) 0, 1, 1, 1, 1, 0, 1, 0, 2, 0, 1, 1, 0, 0, 0, 0, 0, 0...
## $ parch     (int) 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1...
## $ ticket    (fctr) 24160, 113781, 113781, 113781, 113781, 19952, 13502...
## $ fare      (dbl) 211.3375, 151.5500, 151.5500, 151.5500, 151.5500, 26...
## $ cabin     (fctr) B5, C22 C26, C22 C26, C22 C26, C22 C26, E12, D7, A3...
## $ embarked  (fctr) S, S, S, S, S, S, S, S, S, C, C, C, C, S, S, S, C, ...
## $ boat      (fctr) 2, 11, NA, NA, NA, 3, 10, NA, D, NA, NA, 4, 9, 6, B...
## $ body      (int) NA, NA, NA, 135, NA, NA, NA, NA, NA, 22, 124, NA, NA...
## $ home.dest (fctr) St Louis, MO, Montreal, PQ / Chesterville, ON, Mont...
## $ has_cabin (int) 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 1...
```

Importing Data

```
titanic_original <- read.csv("titanic_original.csv")
View(titanic_original)
titanic_clean <- data.frame(titanic_original)
suppressMessages(library(dplyr))
suppressMessages(library(tidyr))
```

Port of Embarkation: Missing Values

```
titanic_original[285, 12]
```

```
## [1]
## Levels:  C Q S
```

```
S <- NULL
S <- vector(mode = "character", length = length(S))
S <- "S"
titanic_clean[285,12] <- S
titanic_clean[285, 12]
```

```
## [1] S
## Levels:  C Q S
```

Finding Mean with Missing Age Values

```
titanic_original$age[1:25]
```

```
##  [1] 29.0000  0.9167  2.0000 30.0000 25.0000 48.0000 63.0000 39.0000
##  [9] 53.0000 71.0000 47.0000 18.0000 24.0000 26.0000 80.0000      NA
## [17] 24.0000 50.0000 32.0000 36.0000 37.0000 47.0000 26.0000 42.0000
## [25] 29.0000
```

```
mean(titanic_clean$age, na.rm = TRUE)
```

```
## [1] 29.88113
```

The median would have also been an appropriate method for finding the average age of passengers. Median is a better method when extreme outliers in the distribution could skew the average. For the titanic data set, mean would not significantly skew the results, and was therefore an appropriate method.

```
summary(titanic_clean$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.1667 21.0000 28.0000 29.8800 39.0000 80.0000     264
```

```
titanic_clean <- titanic_clean %>% mutate(age = ifelse(is.na(age), 30, age))
titanic_clean$age[1:20]
```

```
##  [1] 29.0000  0.9167  2.0000 30.0000 25.0000 48.0000 63.0000 39.0000
##  [9] 53.0000 71.0000 47.0000 18.0000 24.0000 26.0000 80.0000 30.0000
## [17] 24.0000 50.0000 32.0000 36.0000
```

Populating Values with Lifeboats

```
titanic_clean$boat[titanic_clean$boat == ""] <- NA
titanic_clean$boat[1:25]
```

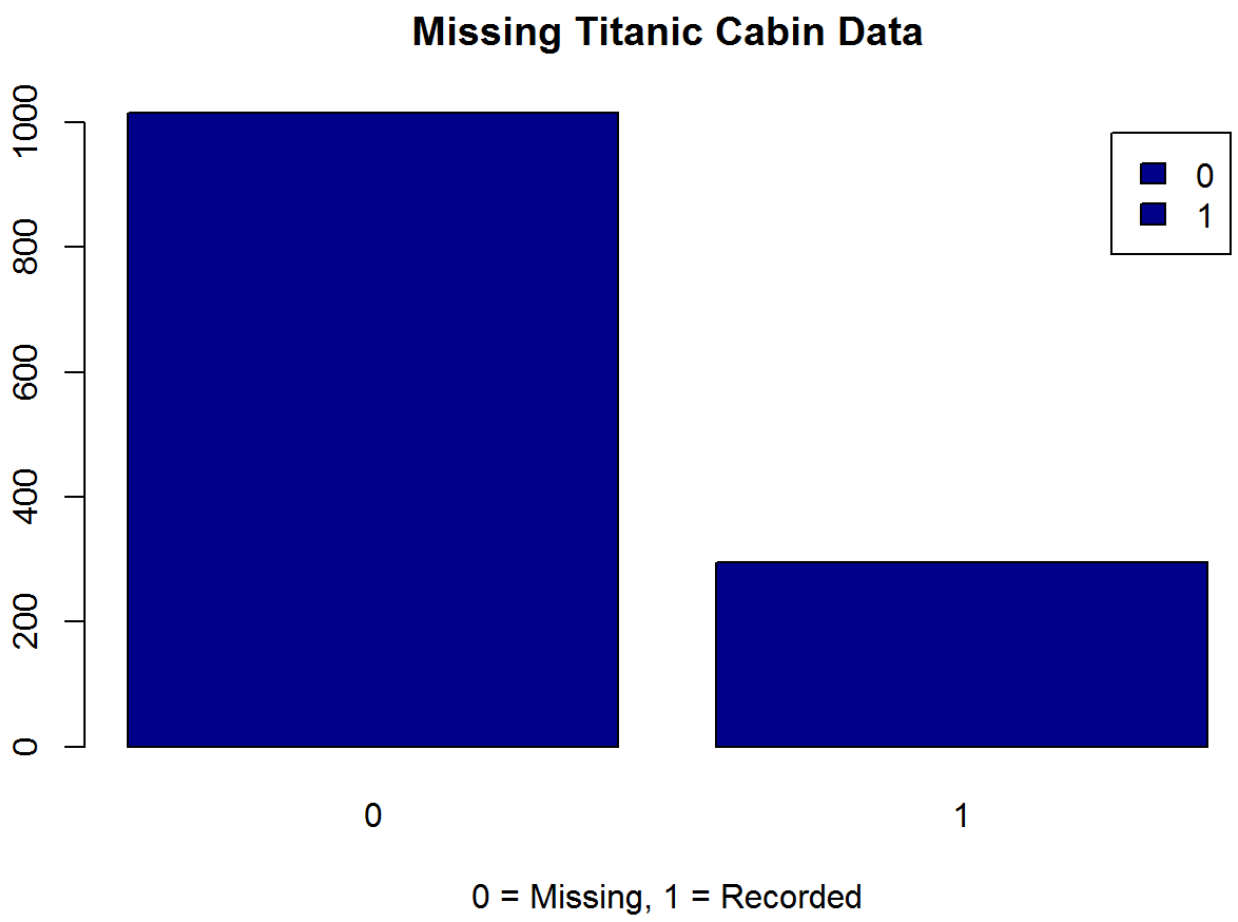
```
##  [1] 2    11   <NA> <NA> <NA> 3    10   <NA> D    <NA> <NA> 4    9    6
## [15] B    <NA> <NA> 6    8    A    5    5    5    4    8
## 28 Levels:  1 10 11 12 13 13 15 13 15 B 14 15 15 16 16 2 3 4 5 5 7 ... D
```

Binomial Distribution with Cabins

```
titanic_clean$boat[titanic_clean$boat == ""] <- NA
cabin_search <- NULL
cabin_search <- vector(mode = "integer", length = length(cabin_search))
titanic_clean$cabin[titanic_clean$cabin == ""] <- NA
has_cabin <- NULL
has_cabin <- titanic_clean$cabin
cabin_search <- as.numeric(has_cabin, rm.na = FALSE)
has_cabin <- ifelse(is.na(cabin_search), 0, 1)
titanic_clean$has_cabin <- has_cabin
titanic_clean$has_cabin[1:25]
```

```
##  [1] 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 0 1 1 1 1 1 1 1 0 1
```

```
counts <- table(titanic_clean$has_cabin)
barplot(counts, main="Missing Titanic Cabin Data",
    xlab="0 = Missing, 1 = Recorded", col=c("darkblue","darkblue"),
      legend = rownames(counts), beside=TRUE)
```


Missing Titanic Cabin Data
0
200
400
600
800
1000
0
1
0
1
0 = Missing, 1 = Recorded