Lab 6

Enter Your Name and UNI Here

February 22nd, 2018

Instructions

Make sure that you upload an RMarkdown file to the canvas page (this should have a .Rmd extension) as well as the PDF of HTML output after you have knitted the file (this will have a .pdf extension or .html). Note that since you have already knitted this file, you should see both a **Lab6_UNI.pdf** and a **Lab6_UNI.Rmd** file in your UN2102 folder. Click on the **Files** tab to the right to see this. The files you upload to the Canvas page should be updated with commands you provide to answer each of the questions below. You can edit this file directly to produce your final solutions. The lab is due 11:59pm on Tuesday, March 27th.

Titanic

In this lab we will be studying a data set which provides information on the survival rates of passengers on the fatal voyage of the ocean liner *Titanic*. The dataset provides information on each passenger including, for example, economic status, sex, age, cabin, name, and survival status. This is a training dataset taken from the Kaggle competition website; for more information on Kaggle competitions, please refer to https://www.kaggle.com. Students should download the data set on Canvas.

Tasks

1) Load the **Titanic** data set.

```
# Read in data
setwd("/Users/salimmjahad/Desktop/STAT_COMP/lab6")
titanic <- read.table("Titanic.txt", header = TRUE, as.is = TRUE)
dim(titanic)
```

```
## [1] 891  12
```

2) Look at the first 10 entries of the variable **Name**. Notice that each person has a *title*, i.e., Mr., Mrs. Miss., etc. . .

```
## R Code ------------------------
head(titanic$Name, 10)
```

```
##  [1] "Braund, Mr. Owen Harris"
##  [2] "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
##  [3] "Heikkinen, Miss. Laina"
##  [4] "Futrelle, Mrs. Jacques Heath (Lily May Peel)"
##  [5] "Allen, Mr. William Henry"
##  [6] "Moran, Mr. James"
##  [7] "McCarthy, Mr. Timothy J"
##  [8] "Palsson, Master. Gosta Leonard"
##  [9] "Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)"
## [10] "Nasser, Mrs. Nicholas (Adele Achem)"
```

3) Create a new variable of the **titanic** dataframe called **Title** that gives the appropriate title of each passenger. The variable **Title** should have 5 levels: **Miss**, **Mrs**, **Mr**, **Master**,and **Other**. Display the first 10 entries of the new variable **Title**.

```
## R Code ------------------------
get1 <- function(vec) {
  return(vec[1])
}
get2 <- function(vec) {
  return(vec[2])
}
repOther <- function(char) {
  if (char %in% c("Mr", "Mrs", "Miss", "Master")) {
    return(char)
  }
  else {
    return("Other")
  }
}

titanic$Title = sapply(strsplit(sapply(strsplit(titanic$Name, split = ", "), get2), split = ". "), get1)

titanic$Title = as.factor(sapply(titanic$Title, repOther))

head(titanic$Title, 10)
```

```
##  [1] Mr     Mrs    Miss   Mrs    Mr     Mr     Mr     Master Mrs    Mrs
## Levels: Master Miss Mr Mrs Other
```

4) Create a table showing the counts for each level of the variable **Title**. Also create a table showing the number of passengers that survived split by their *title*.

```
## R Code ------------------------
lvl <- c("Mr", "Mrs", "Miss", "Master", "Other")
lvl_count <- c(length(which(titanic$Title=="Mr")),
         length(which(titanic$Title=="Mrs")),
         length(which(titanic$Title=="Miss")),
         length(which(titanic$Title=="Master")),
         length(which(titanic$Title=="Other")))
cnt <- data.frame(lvl, lvl_count)
cnt
```

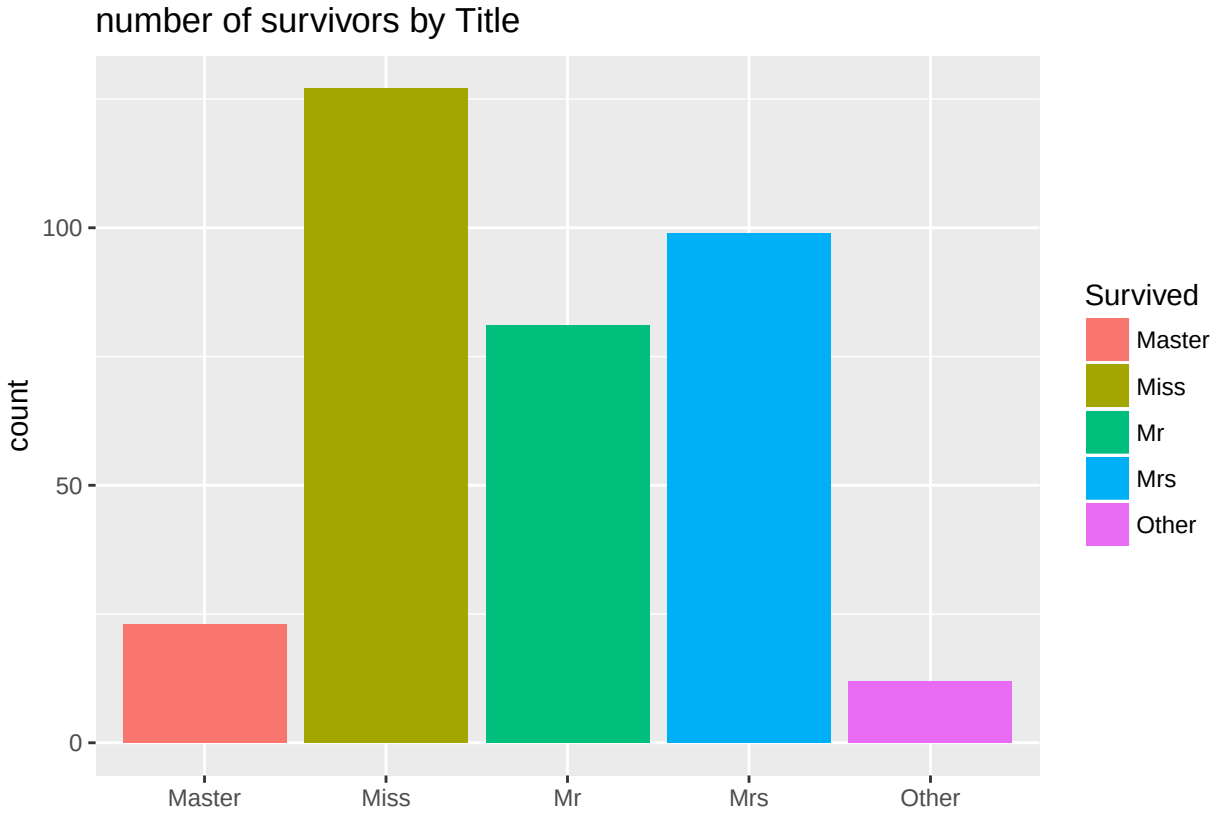
```
##      lvl lvl_count
## 1     Mr       517
## 2    Mrs       125
## 3   Miss       182
## 4 Master        40
## 5  Other        27
```

```
survivors <- subset(titanic, Survived==1)
surv_count <- c(length(which(survivors$Title=="Mr")),
                length(which(survivors$Title=="Mrs")),
                length(which(survivors$Title=="Miss")),
                length(which(survivors$Title=="Master")),
                length(which(survivors$Title=="Other")))
surv_cnt <- data.frame(lvl, surv_count)
surv_cnt
```

```
##      lvl surv_count
## 1     Mr         81
## 2    Mrs         99
## 3   Miss        127
## 4 Master         23
## 5  Other         12
```

5) Plot the number of passengers that survived split by the variable **Title**. Use **ggplot** with the geometric object **geom_bar**.

```
## R Code ------------------------
library("ggplot2")
ggplot(data=survivors) +
  geom_bar(aes(x=factor(Title), fill=Title))+
  labs(title = "number of survivors by Title",fill="Survived", x="")
```



6) Display all of the names that correspond to **Other**. How many cases fall in the this category and what are the name titles corresponding to the level **Other**? Note: you can just identify the names by inspection.

```
## R Code ------------------------
repNOther <- function(char) {
  if (char %in% c("Mr", "Mrs", "Miss", "Master")) {
    return("Other")
  }
  else {
    return(char)
  }
}
```

```
tmp <- sapply(strsplit(sapply(strsplit(titanic$Name, split = ", "), get2), split = ". "), get1)
tmp <- sapply(tmp, repNOther)
names(tmp) <- NULL
levels(as.factor(tmp))
```

```
##  [1] "Capt"     "Col"      "Don"      "Dr"       "Jonkheer" "Lady"
##  [7] "Major"    "Mlle"     "Mme"      "Ms"       "Other"    "Rev"
## [13] "Sir"      "th"
```

"Capt" "Col" "Don" "Dr" "Jonkheer" "Lady" "Major"
"Mlle" "Mme" "Ms" "Rev" "Sir" "th"

There are 13 levels in Other, some due to some cases that don't have the name in the same format.

7) Create a new variable of the titanic data frame called **Last_name** that gives the last name of passenger. Display the first 10 entries of the new variable **Last_name**.

```
## R Code ------------------------
head(titanic$Name)
```

```
## [1] "Braund, Mr. Owen Harris"
## [2] "Cumings, Mrs. John Bradley (Florence Briggs Thayer)"
## [3] "Heikkinen, Miss. Laina"
## [4] "Futrelle, Mrs. Jacques Heath (Lily May Peel)"
## [5] "Allen, Mr. William Henry"
## [6] "Moran, Mr. James"
```

```
titanic$Last_name <- sapply(strsplit(titanic$Name, split = ", "), get1)
head(titanic$Last_name, 10)
```

```
##  [1] "Braund"    "Cumings"   "Heikkinen" "Futrelle"  "Allen"
##  [6] "Moran"     "McCarthy"  "Palsson"   "Johnson"   "Nasser"
```

8) Display the first 8 most common last names.

```
## R Code ------------------------
head(summary(as.factor(titanic$Last_name)),8)
```

```
## Andersson      Sage    Carter   Goodwin   Johnson    Panula     Skoog
##         9         7         6         6         6         6         6
##      Rice
##         5
```