Assignment One, Using Dataframes

Load, Clean and Standardize Data

```r
# First load data.
library(sqldf)
library(dplyr)
library(stringr)
library(tidyr)
library(ggplot2)
g_repo = "https://raw.githubusercontent.com/tonythor/cuny-datascience/"
lf =  paste(g_repo,"develop/data/lusitania_manifest.csv", sep = "")
tf =  paste(g_repo,"develop/data/titanic_lifeboats.csv", sep="")
titanic <- read.csv(tf)
lusitania <- read.csv(lf)

## Clean and standardize the titanic data set
titanic[c('last_name', 'first_name')]<- str_split_fixed(titanic$name, ',', 2)
replacements  = c('Miss.'='','Master.'='', 'Mrs.'='', 'Mister.'='', 'Mr.'='')
titanic$first_name<- str_replace_all(titanic$first_name, replacements )
titanic <- dplyr::rename(titanic, c("gender" = "sex"), c("lifeboat" = "boat"))
titanic_boat= c("titanic")
titanic$boat = titanic_boat
titanic <- titanic %>% mutate_at(c('age'), ~replace_na(.,0))
titanic <- titanic %>%   mutate(age = as.numeric(age))
# Could be any column name to filter not. Last row is empty and needs to be dropped.
titanic <- titanic %>% filter(!last_name=='')
titanic_clean <- titanic %>% select(boat,last_name, first_name,
                                    survived, gender, age, lifeboat)
## Clean and standardize the Lusitania data set
lusitania_boat= c("lusitania")
lusitania$boat = lusitania_boat
```

```r
lusitania <- dplyr::rename(lusitania, c("first_name" = "Personal.name"),
                           c("last_name" = "Family.name"),
                           c("lifeboat" = "Lifeboat"),
                           c("age" = "Age"),
                           c("gender" = "Sex"))
lusitania <- lusitania %>%  mutate(age = as.numeric(age))
lusitania <- lusitania %>% mutate_at(c('age'), ~replace_na(.,0))
lusitania = lusitania %>% mutate(survived = case_when(grepl("Saved", Fate) ~ 1,
                                                      .default = 0))
lusitania$gender <- tolower(lusitania$gender)
lusitania$last_name <- str_to_title(lusitania$last_name)
lusitania_clean <- lusitania %>% select(boat,last_name, first_name, survived,
                                        gender, age, lifeboat)
## uninion the data sets into one, add a numerical categorical flag for gender
boats = union(lusitania_clean,titanic_clean)
```