Tennis Group Project

Samuel Stockman, Angie Gray, Jack Simons

10/12/2020

Let's first load in the relevant data (see `tennis.R`)

```
file_list = list.files()
```

Now we want to compile the data into a better format. We need the input `variables` which should be equal to something of the form: `variables = c('surface', 'winner_id', 'loser_id', 'loser_age', 'winner_age')`

```
load_data <- function(file_list, variables) {
  no_files <- length(file_list)
  test_file <- file_list[[no_files]]
  file_list <- file_list[1:(no_files-1)]
  for (file_index in 1:(no_files-1)) {
    temp_file <- read.csv(file_list[[file_index]])
    reduced_file <- subset(temp_file, select = variables)
    if (file_index==1) {
      melted_files <- reduced_file
    } else {
      melted_files <- rbind(melted_files, reduced_file)
    }
  }
  return(melted_files)
}
```

Let's now use this function

```
variables <- c('surface', 'winner_id', 'loser_id', 'winner_age', 'loser_age')
my_data <- load_data(file_list, variables)
```

Let's see what the data looks like

```
head(my_data, 30)
```

```
##    surface winner_id loser_id winner_age loser_age
## 1     Clay    200781   200795   31.91786  16.18617
## 2     Clay    202512   200862   21.97673  19.29363
## 3     Clay    200869   200863   23.09925  17.72211
## 4     Clay    200274   200834   22.85010  21.00205
## 5     Clay    200791   202513   29.56331  17.44011
## 6     Clay    200870   200781   19.49624  31.91786
## 7     Clay    202512   200869   21.97673  23.09925
## 8     Clay    200274   200791   22.85010  29.56331
## 9     Clay    200317   202514   26.92402  16.03012
## 10    Clay    200870   202512   19.49624  21.97673
## 11    Clay    200317   200274   26.92402  22.85010
```

```
## 12    Clay    200317   200870   26.92402  19.49624
## 13    Clay    200791   200862   29.59069  19.32101
## 14    Clay    200834   200795   21.02943  16.21355
## 15    Clay    200274   202514   22.87748  16.05749
## 16    Clay    200870   200863   19.52361  17.74949
## 17    Clay    200834   200791   21.02943  29.59069
## 18    Clay    200274   202512   22.87748  22.00411
## 19    Clay    200869   200317   23.12663  26.95140
## 20    Clay    200870   200834   19.52361  21.02943
## 21    Clay    200274   200869   22.87748  23.12663
## 22    Clay    200870   200274   19.52361  22.87748
## 23   Grass    200130   200131         NA  22.69405
## 24   Grass    200133   200132         NA        NA
## 25   Grass    200134   200135   18.09172  30.32170
## 26   Grass    200137   200136         NA        NA
## 27   Grass    200139   200138   21.89185        NA
## 28   Grass    200140   200141         NA  17.67556
## 29   Grass    200143   200142   21.22930        NA
## 30   Grass    200144   200145   25.51129  18.33265
```

We notice that the winner and loser ages contain `NA`, so let's set these missing values equal to their mean

```
missing_winner_ages <- which(is.na(my_data$winner_age))
missing_loser_ages <- which(is.na(my_data$loser_age))
my_data$winner_age[missing_winner_ages] <- mean(my_data$winner_age[!is.na(my_data$winner_age)])
my_data$loser_age[missing_loser_ages] <- mean(my_data$loser_age[!is.na(my_data$loser_age)])
head(my_data, 30)
```

```
##    surface winner_id loser_id winner_age loser_age
## 1     Clay    200781   200795   31.91786  16.18617
## 2     Clay    202512   200862   21.97673  19.29363
## 3     Clay    200869   200863   23.09925  17.72211
## 4     Clay    200274   200834   22.85010  21.00205
## 5     Clay    200791   202513   29.56331  17.44011
## 6     Clay    200870   200781   19.49624  31.91786
## 7     Clay    202512   200869   21.97673  23.09925
## 8     Clay    200274   200791   22.85010  29.56331
## 9     Clay    200317   202514   26.92402  16.03012
## 10    Clay    200870   202512   19.49624  21.97673
## 11    Clay    200317   200274   26.92402  22.85010
## 12    Clay    200317   200870   26.92402  19.49624
## 13    Clay    200791   200862   29.59069  19.32101
## 14    Clay    200834   200795   21.02943  16.21355
## 15    Clay    200274   202514   22.87748  16.05749
## 16    Clay    200870   200863   19.52361  17.74949
## 17    Clay    200834   200791   21.02943  29.59069
## 18    Clay    200274   202512   22.87748  22.00411
## 19    Clay    200869   200317   23.12663  26.95140
## 20    Clay    200870   200834   19.52361  21.02943
## 21    Clay    200274   200869   22.87748  23.12663
## 22    Clay    200870   200274   19.52361  22.87748
## 23   Grass    200130   200131   23.50854  22.69405
## 24   Grass    200133   200132   23.50854  23.45292
## 25   Grass    200134   200135   18.09172  30.32170
## 26   Grass    200137   200136   23.50854  23.45292
```

```
## 27   Grass    200139   200138   21.89185  23.45292
## 28   Grass    200140   200141   23.50854  17.67556
## 29   Grass    200143   200142   21.22930  23.45292
## 30   Grass    200144   200145   25.51129  18.33265
```