Analysis

December 26, 2019

The purpose of this document is to make use of the data stored within the `FIDE` data folder.

Before beginning, we need to import several packages that help run the code below.

```
library(tidyverse)
library(data.table)
library(xts)
library(dygraphs)
library(foreach)
library(doParallel)
library(zoo)
library(stringi)
library(cowplot)
library(gganimate)
library(lubridate)
library(knitr)
require(scales)
```

`tidyverse` imports dplyr, tidyr and several other useful packages for data wrangling and manipulation.

`data.table` is used purely for speeding up imports.

Access the data

The data is stored in a path we have to manipulate to get to.

```
path = "~/GitHub/FIDE/Chess Scripts/Step 4 - Cleaning/Cleaned csvs/"
opts_knit$set(root.dir = path)
temp = list.files(path = path, pattern = "*.csv")
proper_temp <- paste(path, temp, sep = "")
```

Create functions to rename datasets

Below, I define a few functions that help us rename the datasets eventually

```
num <- function(x) match(tolower(x), tolower(month.abb))
month <- function(x){return(substr(x, 1, 3))}
add_zero <- function(x){if (x <= 9){x = paste("0", x, sep = "")}; return(x)}
```

Get the month number of all of the files in the dataset

Below, I rename create the variables names when they are imported in memory.

```
temp%>%
sapply(month)%>%
sapply(num)%>%
sapply(add_zero)%>%
paste("20", substr(temp, 4, 5), "-", ., "-", "01", sep = "")-> month_num

head(month_num)
```

```
## [1] "2001-04-01" "2002-04-01" "2003-04-01" "2004-04-01" "2005-04-01"
## [6] "2006-04-01"
```

We can see that the datasets correspond to dates on which the is recorded.

Import a sample of datasets

Due to the data being quite large, your RAM may be used heavily.

I'll look to investigate if parallelizing `rbindlist()` is possible at some point. Please bear with the computation time.

```
for(i in 1:length(proper_temp)) {
assign(month_num[i], fread(proper_temp[i], sep = "*", data.table = FALSE,
                           strip.white = TRUE, blank.lines.skip = TRUE))
}

#statement below takes a very long time to run because of rbindlist()
FIDE <- mget(ls(pattern = "[0-9][0-9]-[0-9][0-9]"))%>%
        rbindlist(fill = TRUE)

rm(list=setdiff(ls(), c("FIDE")))
```

#write dataframe to folder

```
# fwrite(FIDE, "FIDE.csv")
# FIDE <- fread("FIDE.csv", data.table = FALSE)
```

Now that we have the data in a more clean, usable format, it's time we analyzed it.

Set up basic filters

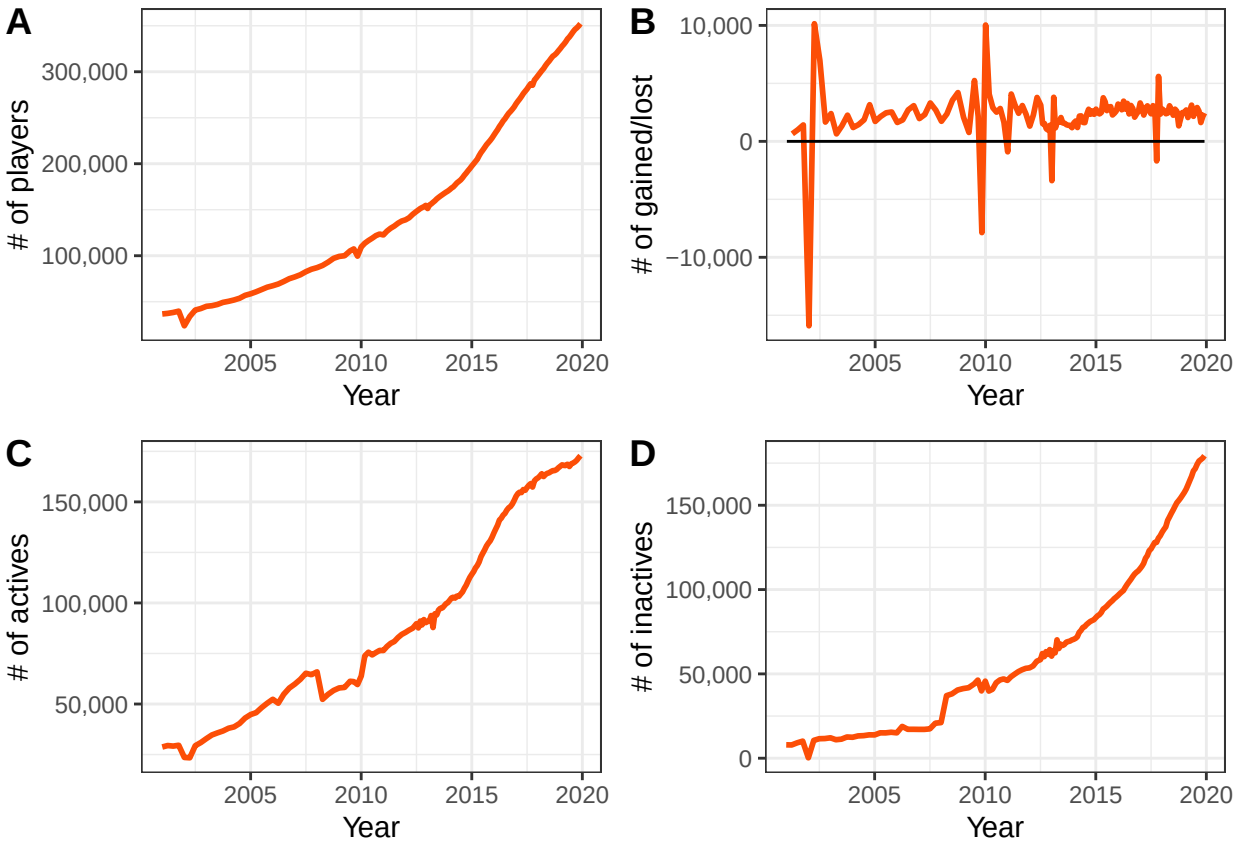
```
Titles = c("CM", "WCM",  "WGM", "WFM", "IM", "FM", "WIM", "GM", "")
Titles_only = c("CM", "WCM",  "WGM", "WFM", "IM", "FM", "WIM", "GM")
Active = c("", "w")
Inactive = c("i", "wi")
```

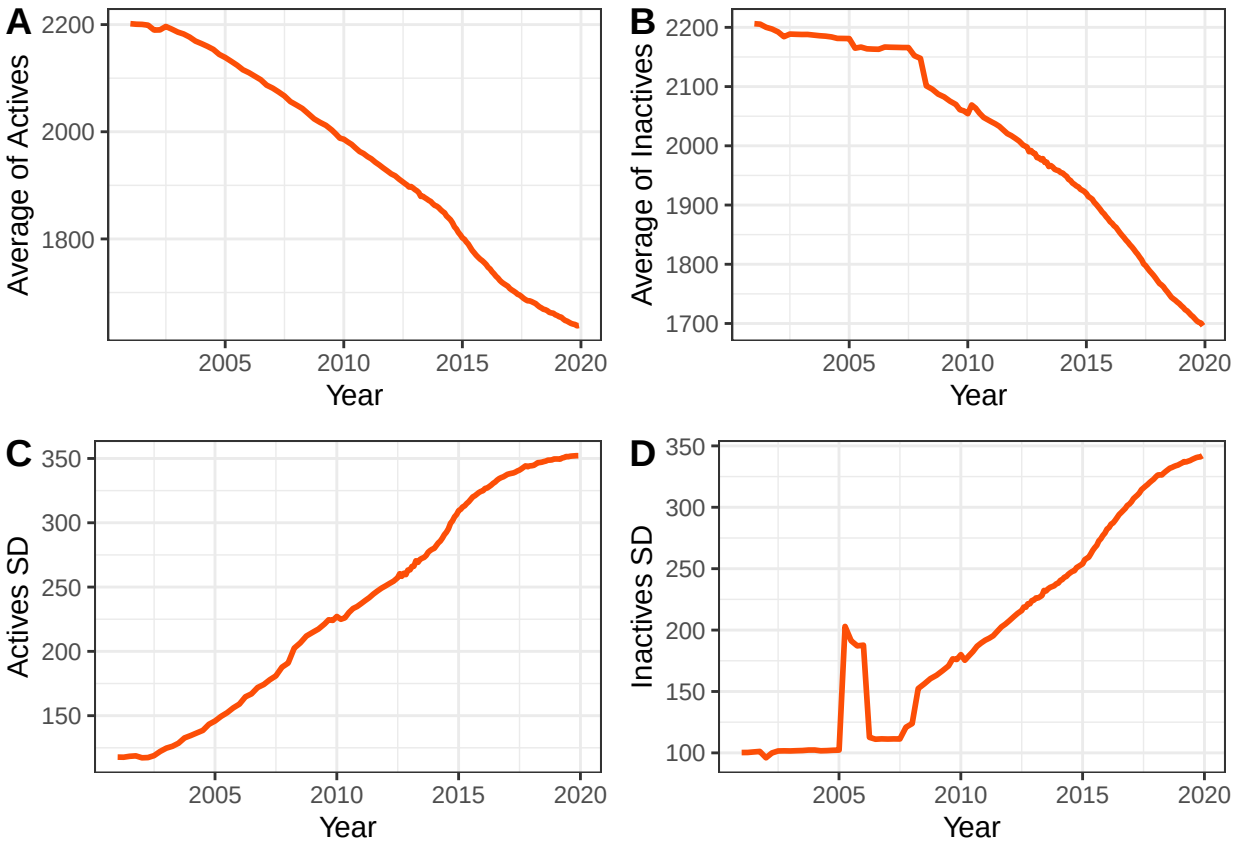
Irregular values by year

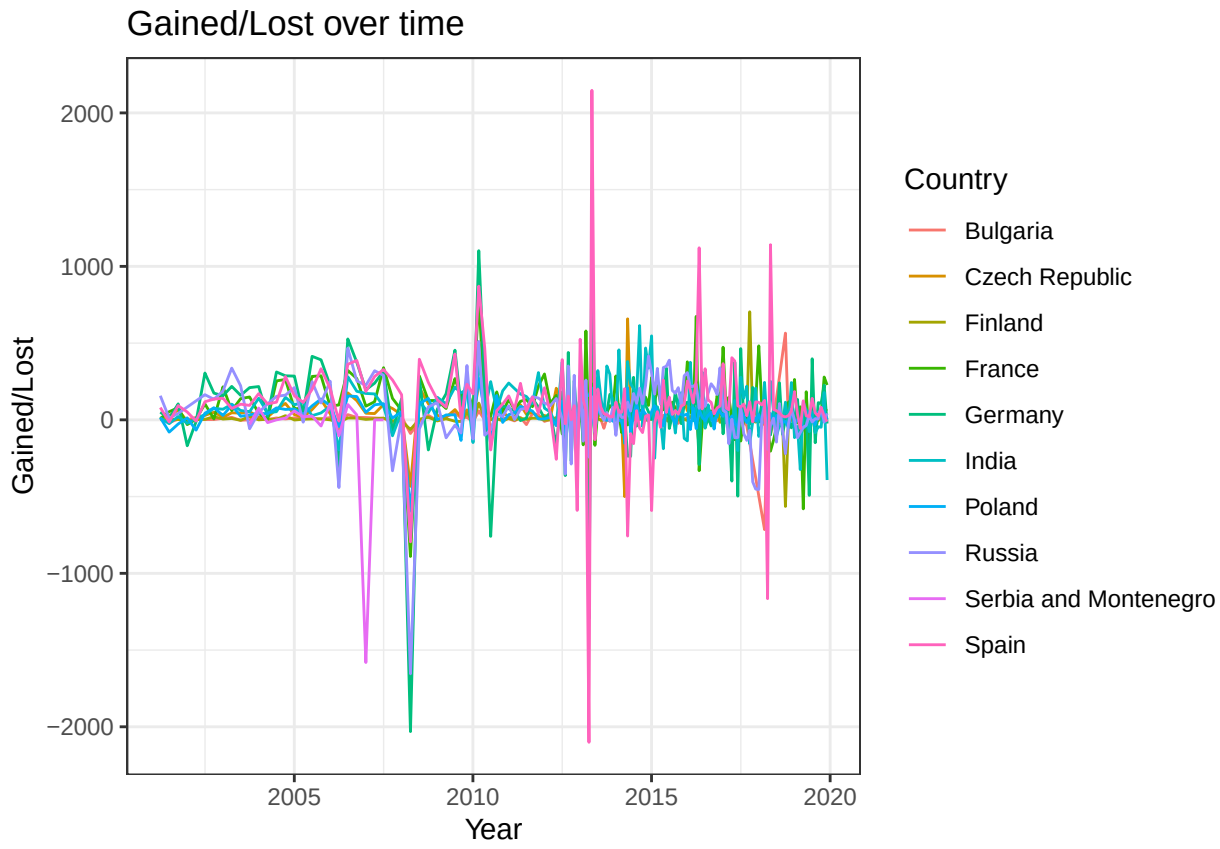
Date_numeric	n
2002.249	7435
2002.003	5455
2001.003	306
2001.249	305
2001.497	304
2005.497	244
2005.003	194
2004.751	170
2004.500	136
2004.251	114

I'll look to address many of the values in the early datasets eventually. For now though, over 99.9% of the data is interpretable.

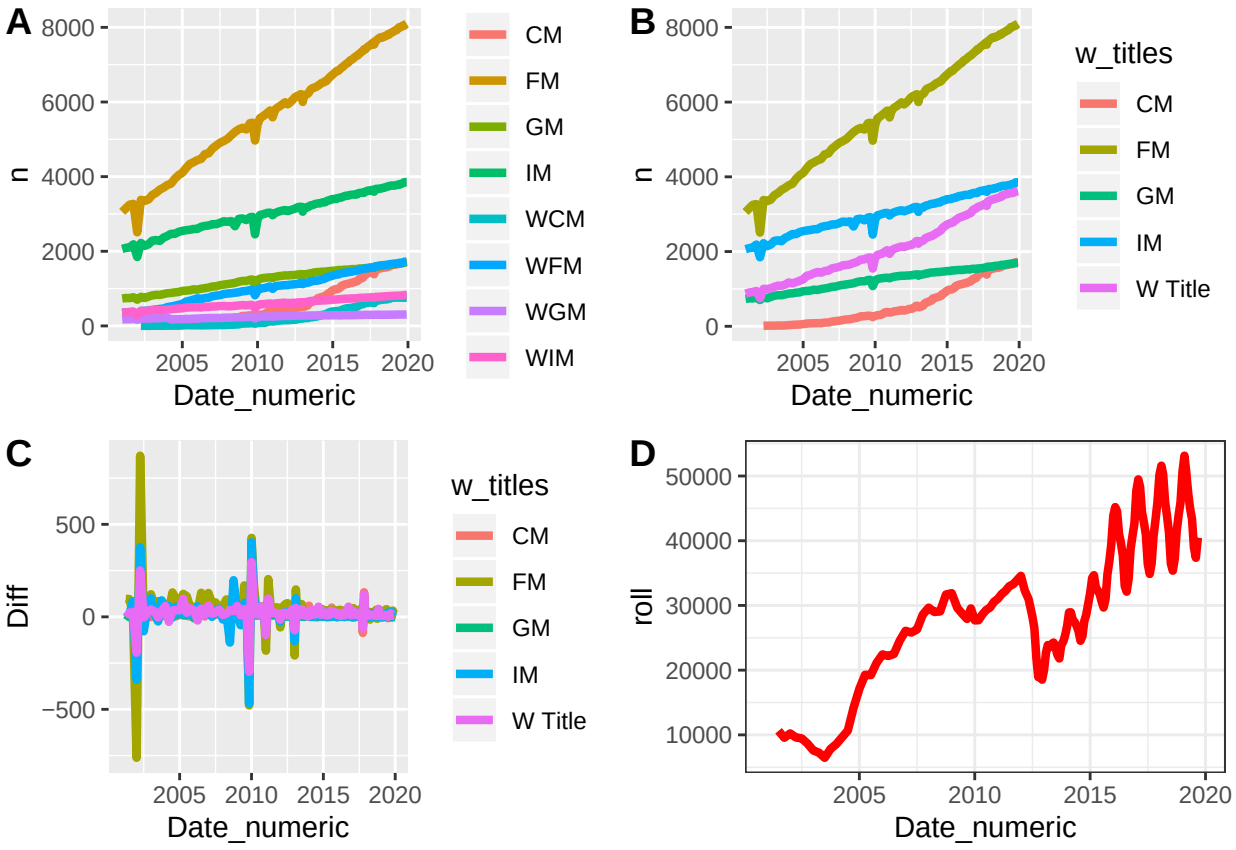
Player counts by Activity

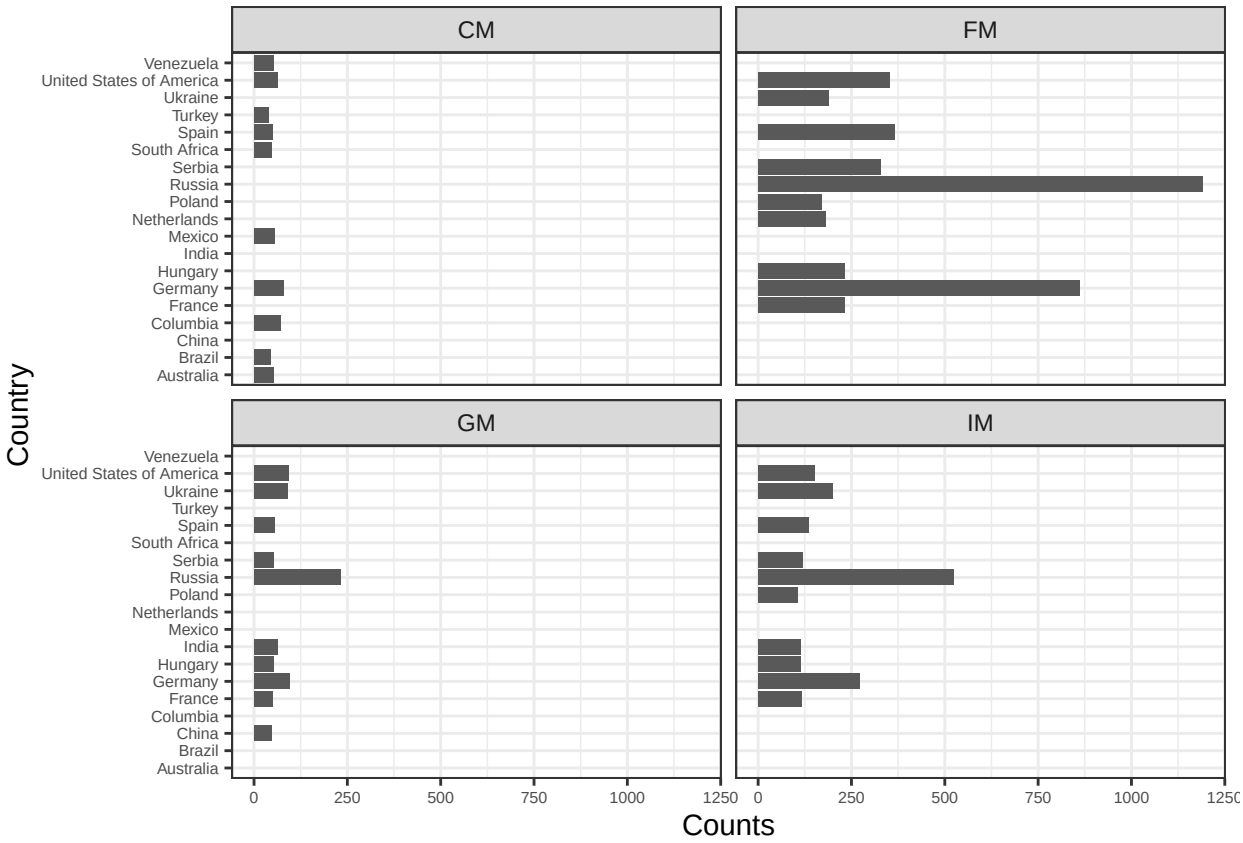


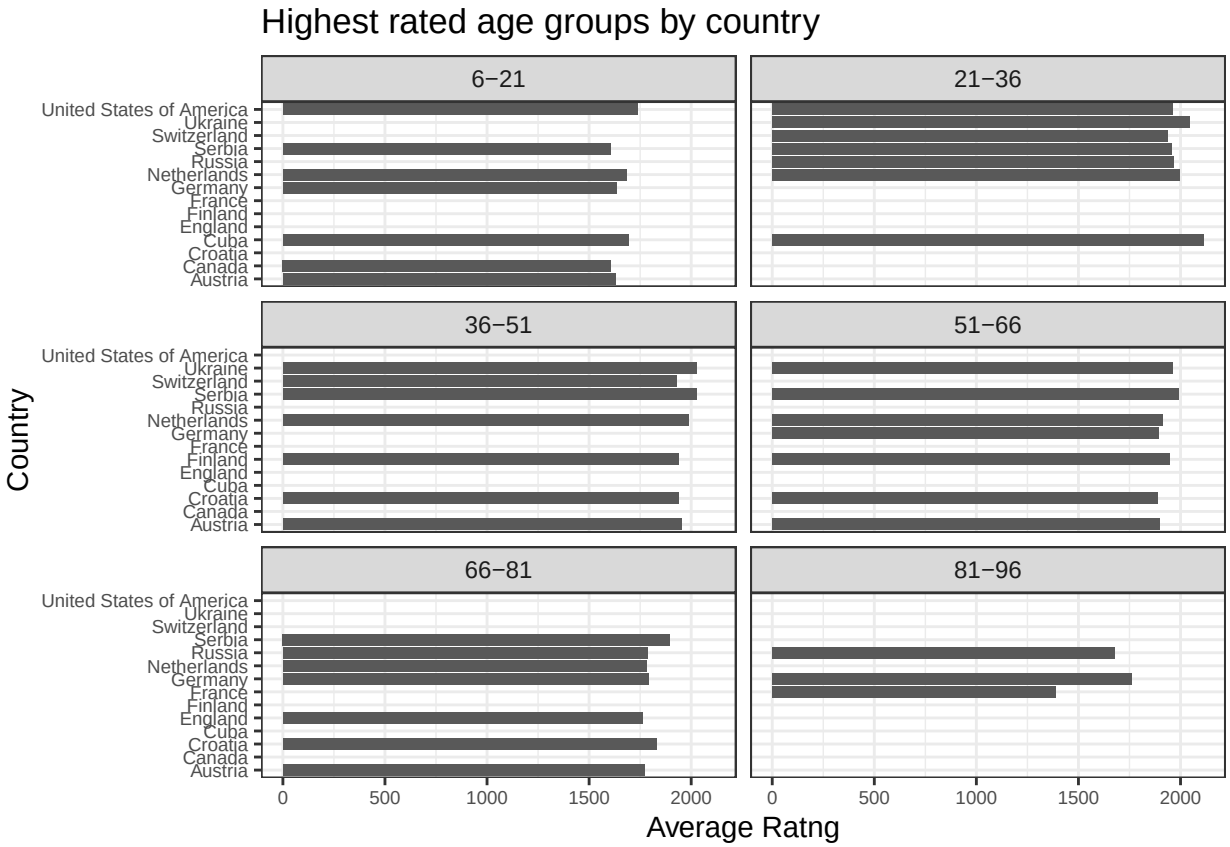

A
Average of Actives
2200
2000
1800
2005
2010
2015
2020
Year
B
Average of Inactives
2200
2100
2000
1900
1800
1700
2005
2010
2015
2020
Year
C
Actives SD
350
300
250
200
150
2005
2010
2015
2020
Year
D
Inactives SD
350
300
250
200
150
100
2005
2010
2015
2020
Year

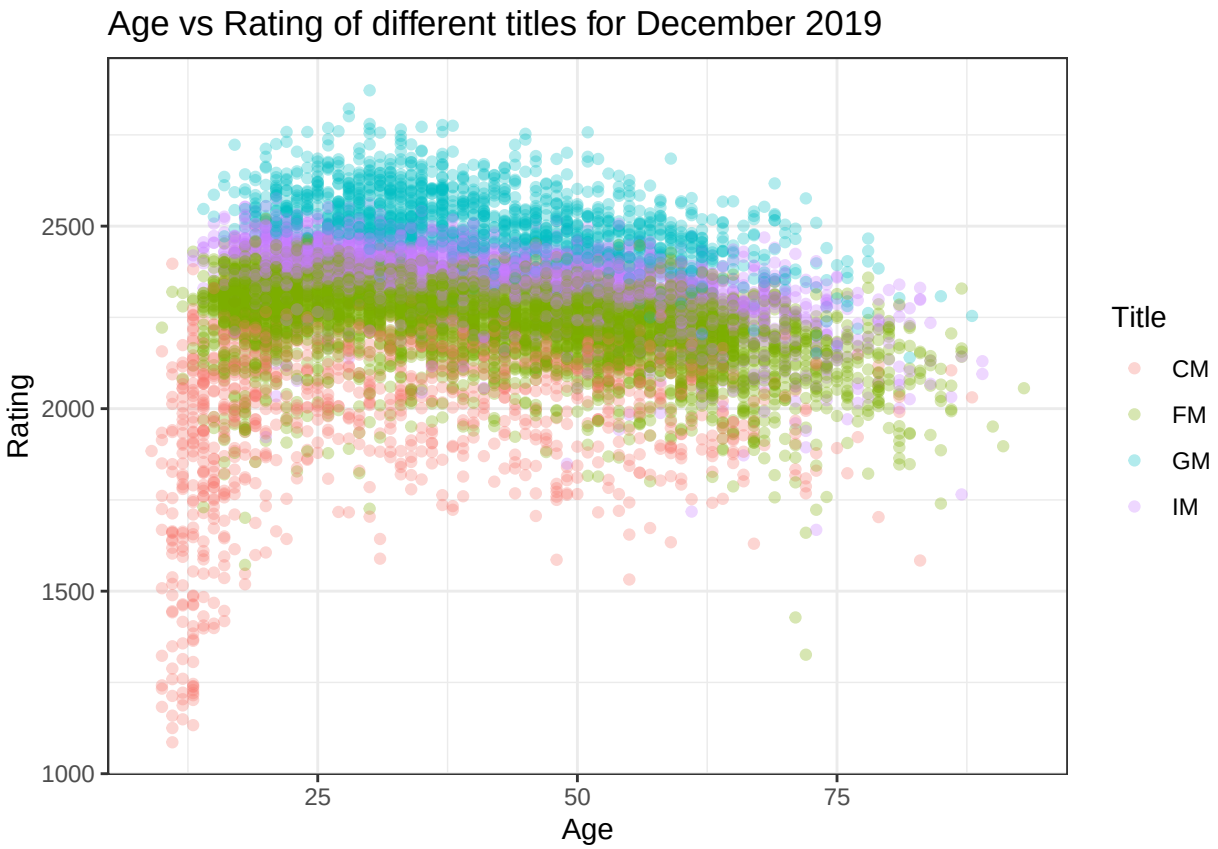


#Titled player count over time


A
B
C
D
n
Diff
roll
Date_numeric
w_titles
CM
FM
GM
IM
WCM
WFM
WGM
WIM
W Title
8000
6000
4000
2000
0
500
−500
50000
40000
30000
20000
10000
2005
2010
2015
2020


CM
FM
GM
IM
Country
Counts
Venezuela
United States of America
Ukraine
Turkey
Spain
South Africa
Serbia
Russia
Poland
Netherlands
Mexico
India
Hungary
Germany
France
Columbia
China
Brazil
Australia
0
250
500
750
1000
1250


Highest rated age groups by country
6−21
21−36
36−51
51−66
66−81
81−96
United States of America
Ukraine
Switzerland
Serbia
Russia
Netherlands
Germany
France
Finland
England
Cuba
Croatia
Canada
Austria
Country
0
500
1000
1500
2000
Average Ratng


Age vs Rating of different titles for December 2019
Rating
2500
2000
1500
1000
25
50
75
Age
Title
CM
FM
GM
IM