# Analysis

January 09, 2020

The purpose of this document is to visually analyze all of the FIDE data files collected in the previous step's folder.

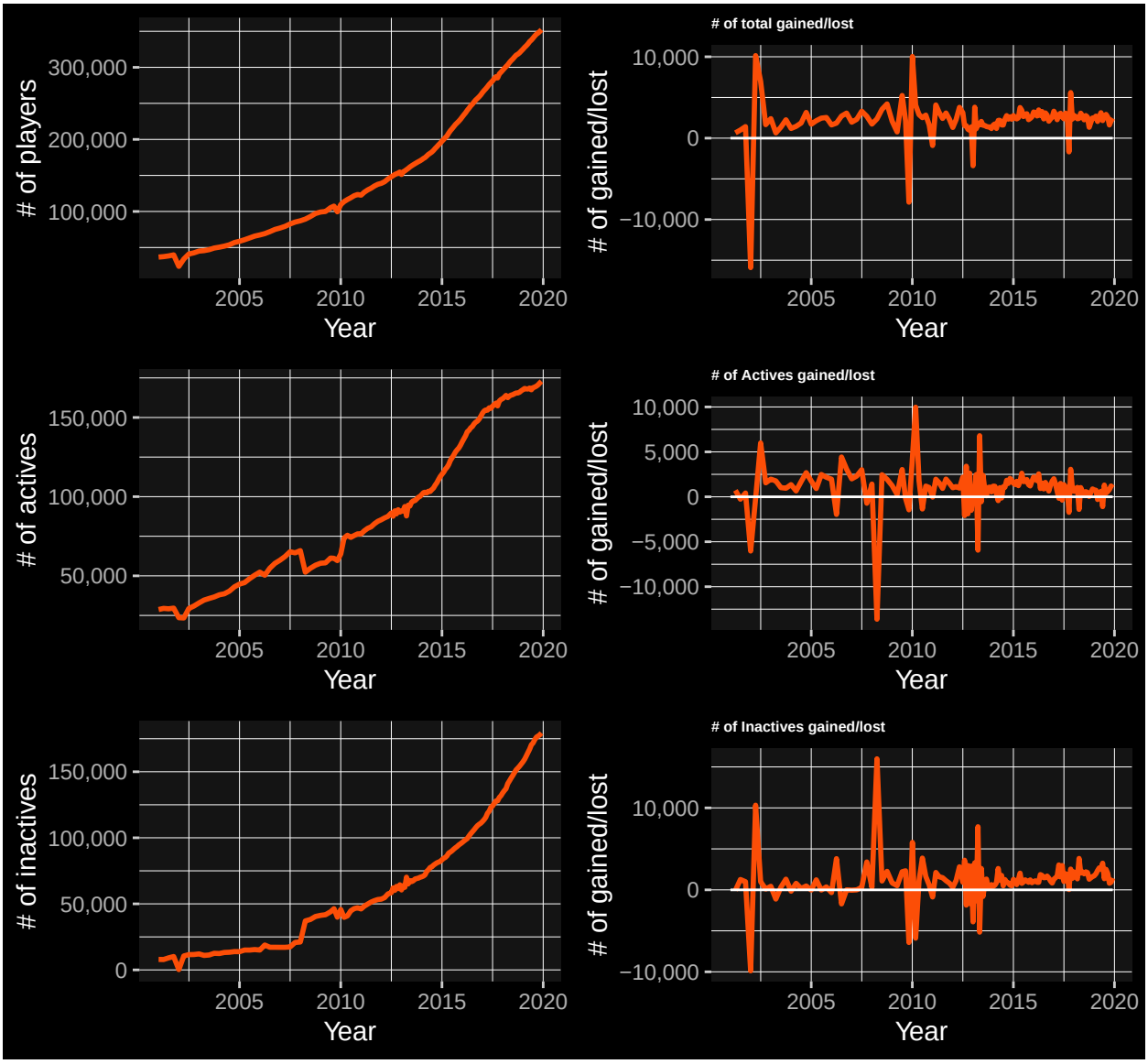
## Irregular values by year

| Date | # of irregular values |
|---|---|
| 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 |

As we can see from the table above, most of irregular values in the files come from early on (2001 - 2005) rather than the latest files.

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.

## Total player count over time

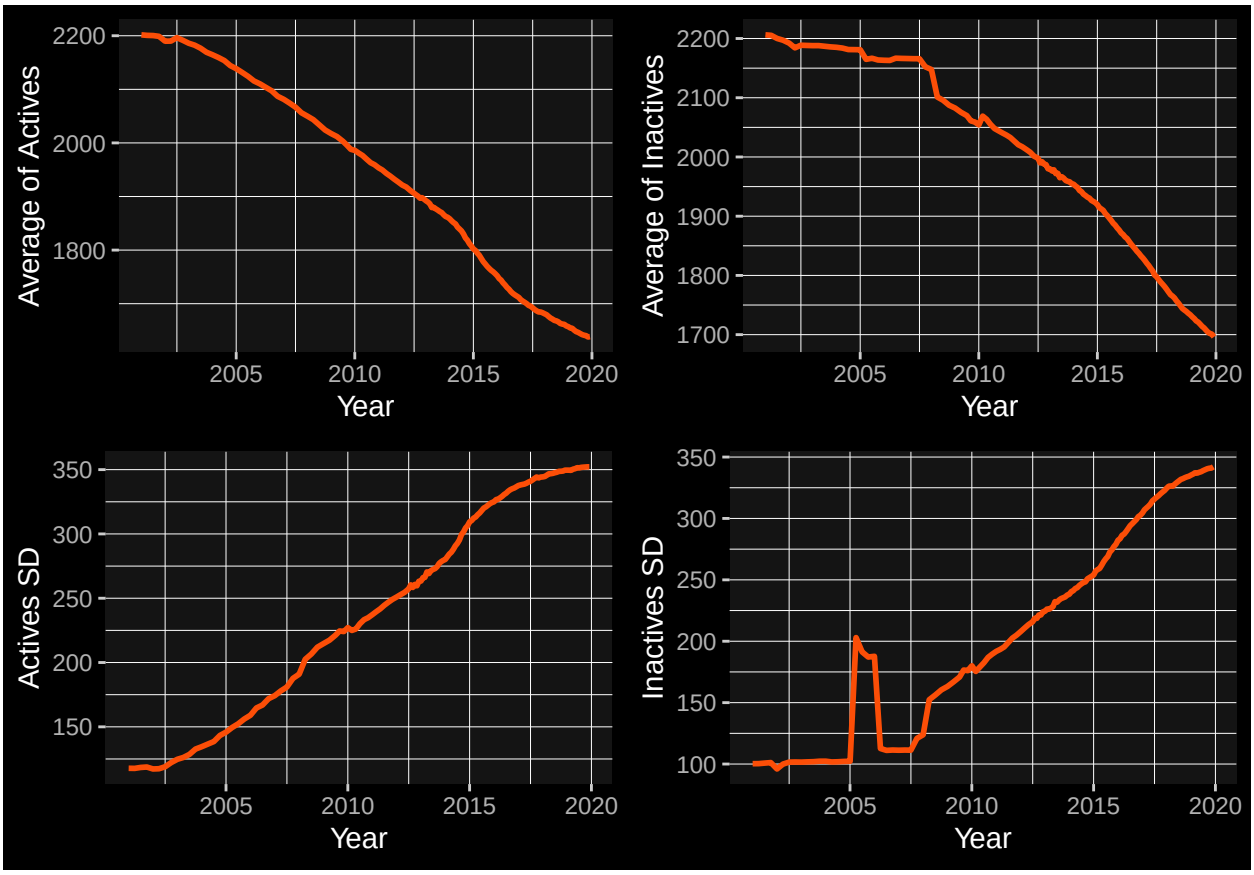



The four charts above reveal the total count of FIDE's members over time. Chart `A` shows a smooth gradual increasing growth curve among total players. This may lead you to believe that more chess players are playing tournaments, but a player's **activity** is a better metric to go by.

**Activity** is defined by if a given FIDE player had played a rated FIDE game within the past 12 months. If we take this into account, charts `C` and `E` show how the total active & inactive player count increase over time. Both charts show a fairly linear trend over time, but from 2007 to 2010 in each graph, there was a noticeable drop off in the active player base and increase in the inactive player base. I have my doubts on if there is faulty data here because chart `A` shows no irregularity during that time. This dip may be due to the economic crash during that time, but I need to do more exploration on this topic before making any definitive statements. Exploring variation by country may also be worth doing.

Charts `B`, `D` and `F` show the number of players gained and lost over time. The most relevant of the 3 graphs is `B` which shows several instances where total player counts dropped off. I genuinely don't if the data is faulty because of my doing or if FIDE is providing incomplete data sets based off of the charts.

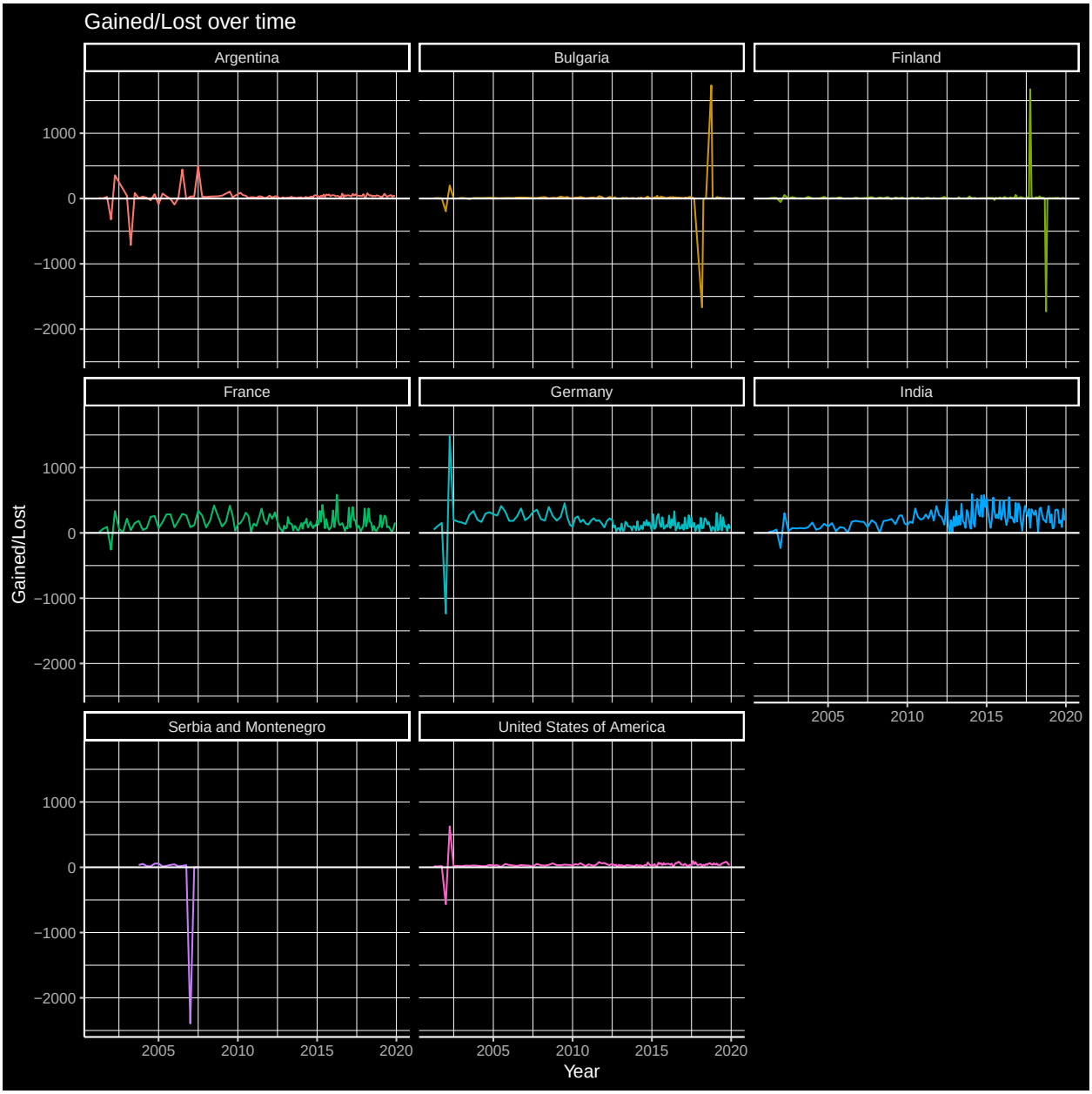
## Rating stability over time (misleading)




Each chart above shows how the Active and Inactive player's average rating and rating standard deviation have progressed over time. For the most part, it is a meaningless metric because FIDE has brought in more **lower** rated chess players into the player pool over time.

This explains the steady decrease in average rating and increase in standard deviation over time.

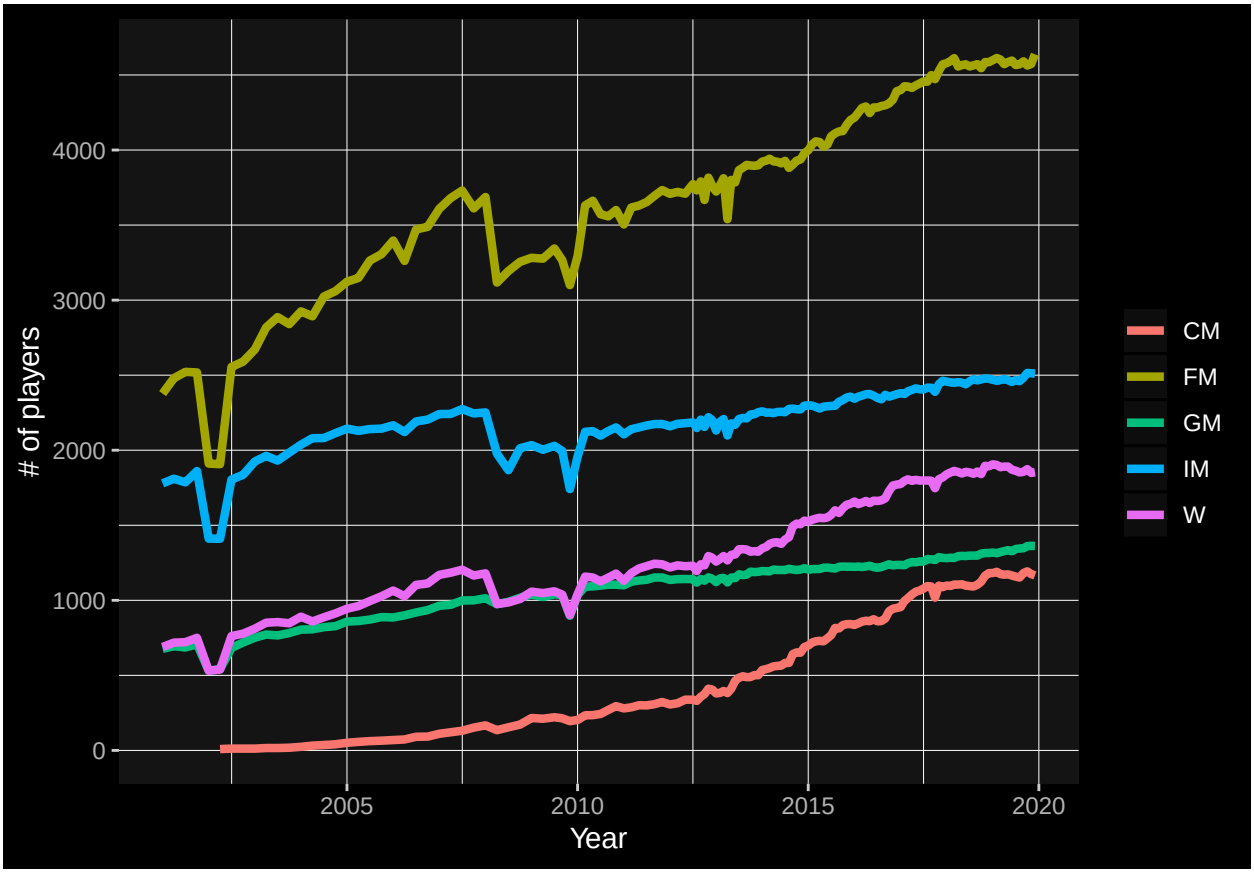
## Which countries have seen the greatest changes in player counts?




A consistent theme we will see in every graph is a potential corruption of data. The graph above displays the of players gained and lost over time by country.

- Bulgaria, Finland, Germany, Serbia and Montenegro, and the United State of America saw unusual spikes & dips at various junctures (2002, 2007, 2017)

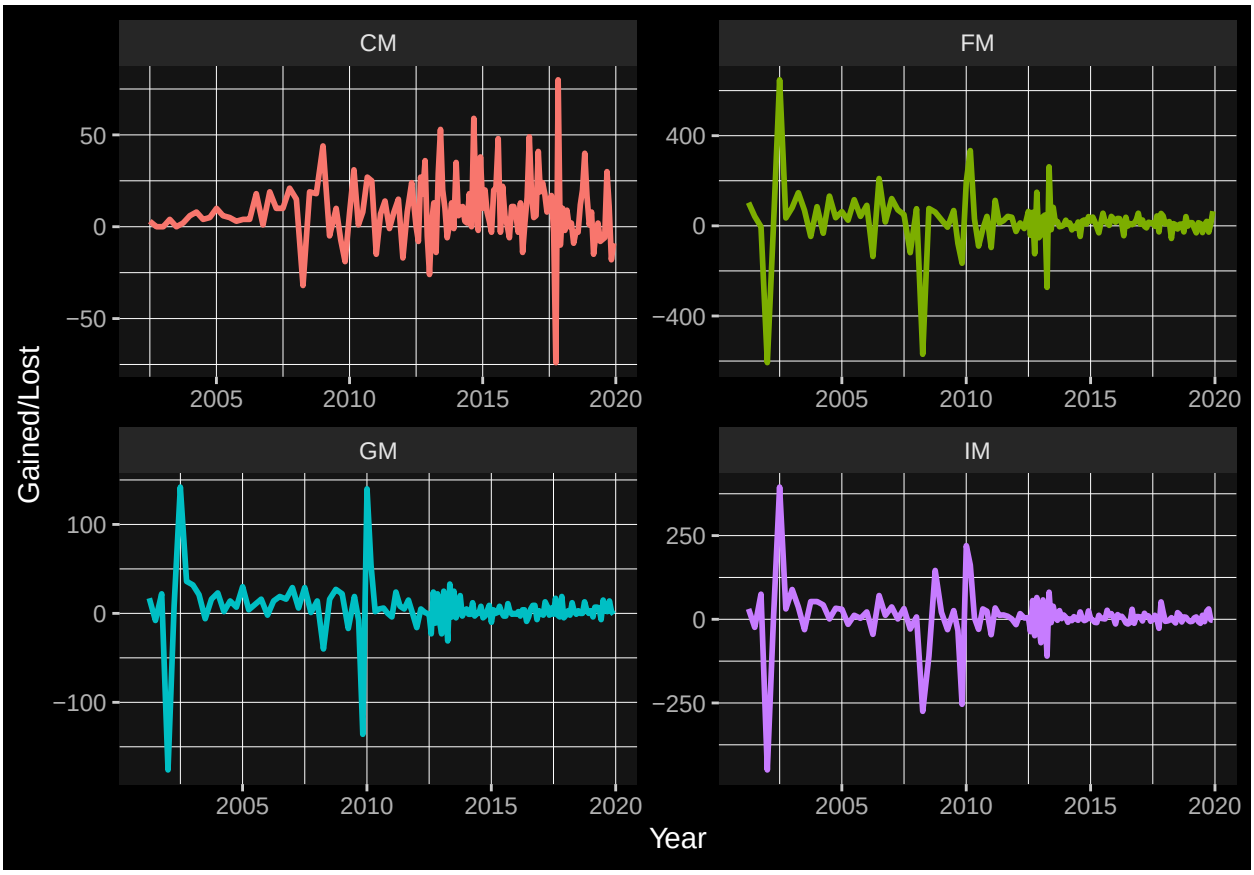
## Titled player count over time




The plot above shows a few interesting trends (W = All women's titles):

- All titled player groups have significantly increased in size over the past 20 years
- Around both 2002 and 2007, there were significant dips in all categories, except in `GM`s and `CM`s.
- There are about as many `GM`s as `CM`s. It goes to show how stigmatized the CM title is.

## Titled players gained/lost over time



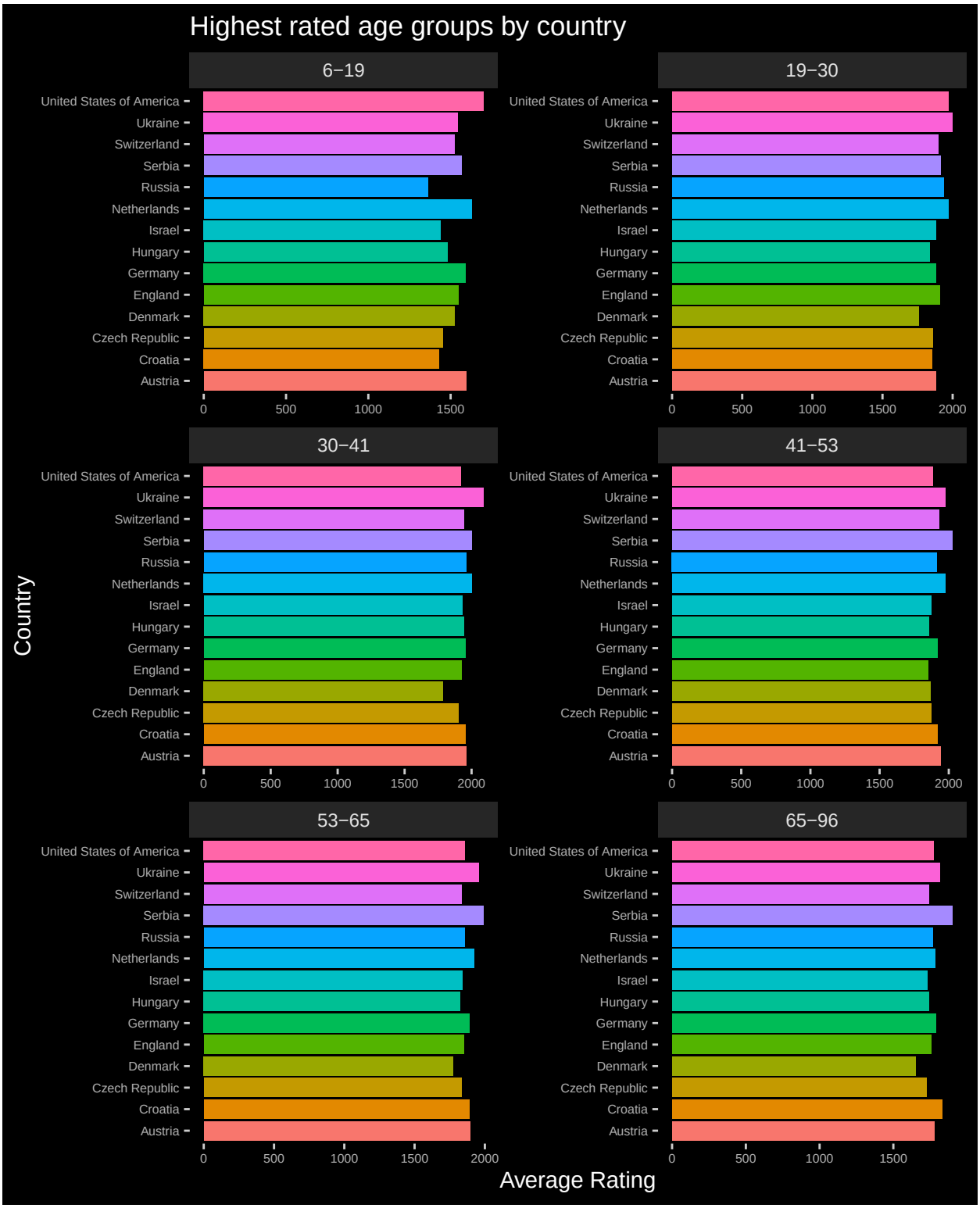

The graph above is graphs the rolling difference of titled player counts over time. As we saw before, there are a few noticeable dips in the graph. Most notably they occur around 2002, 2008 and 2010 in varying degrees.

## Titled player counts by country


Notable observations:

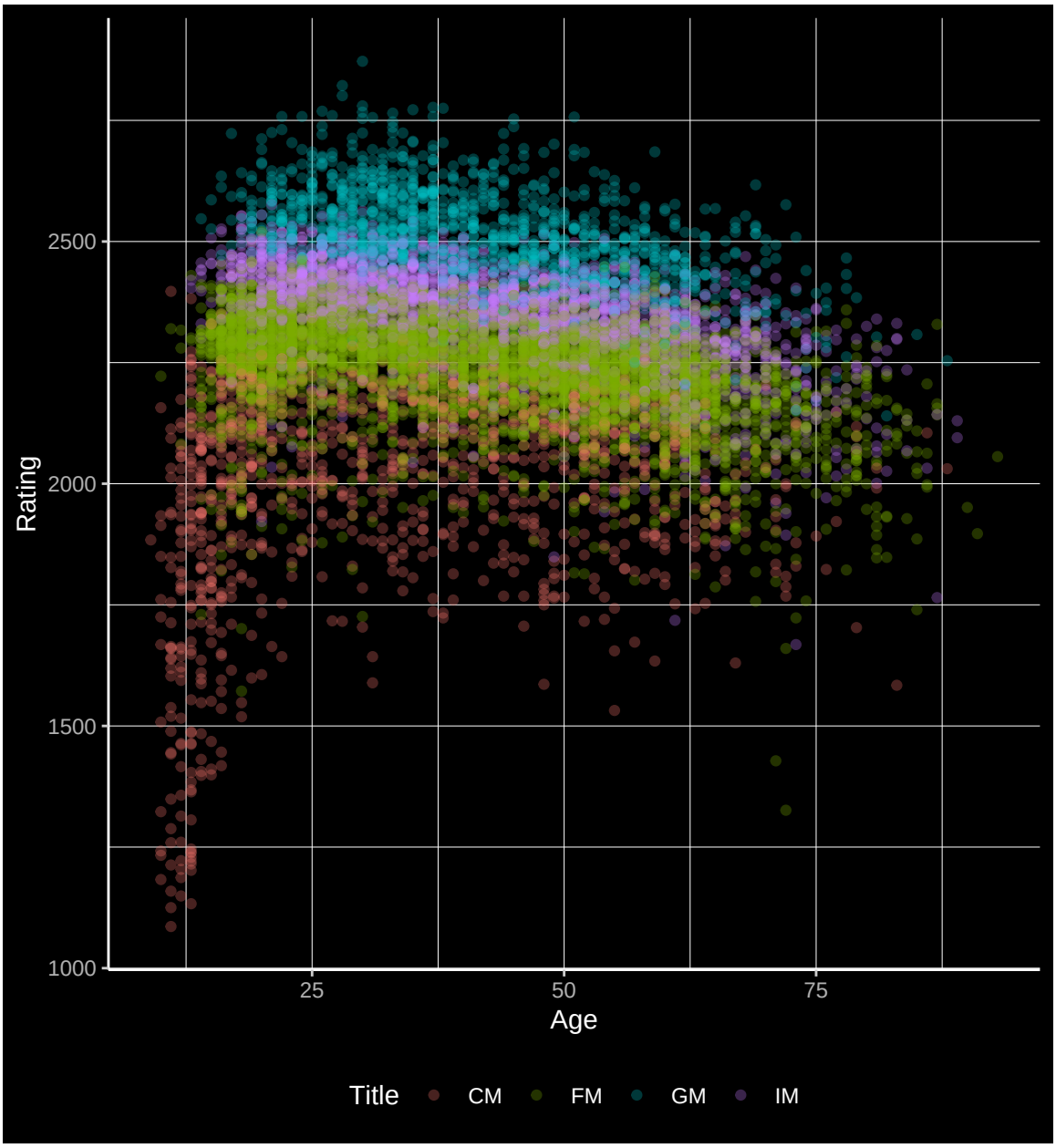
- Clearly, among `FM`, `IM` and `GM` players, Russia has the most titled players by a large margin. This is not surprising because of Russia's longstanding history (USSR and Soviet Union) of chess.
- Germany is a clear 2nd amongst `FM`, `IM` and `GM` players. They also lead the pack in `CM` players. This is surprising to me since I've never thought of Germany as being a bastion of strong players.
- Every other country lags behind these top-tier powerhouses.

## Strongest countries by age group (Work in progress)




This plot would be substantially better at showing relative strengths among countries if I could scale it differently. This will adjusted in future iterations, probably through scaling.

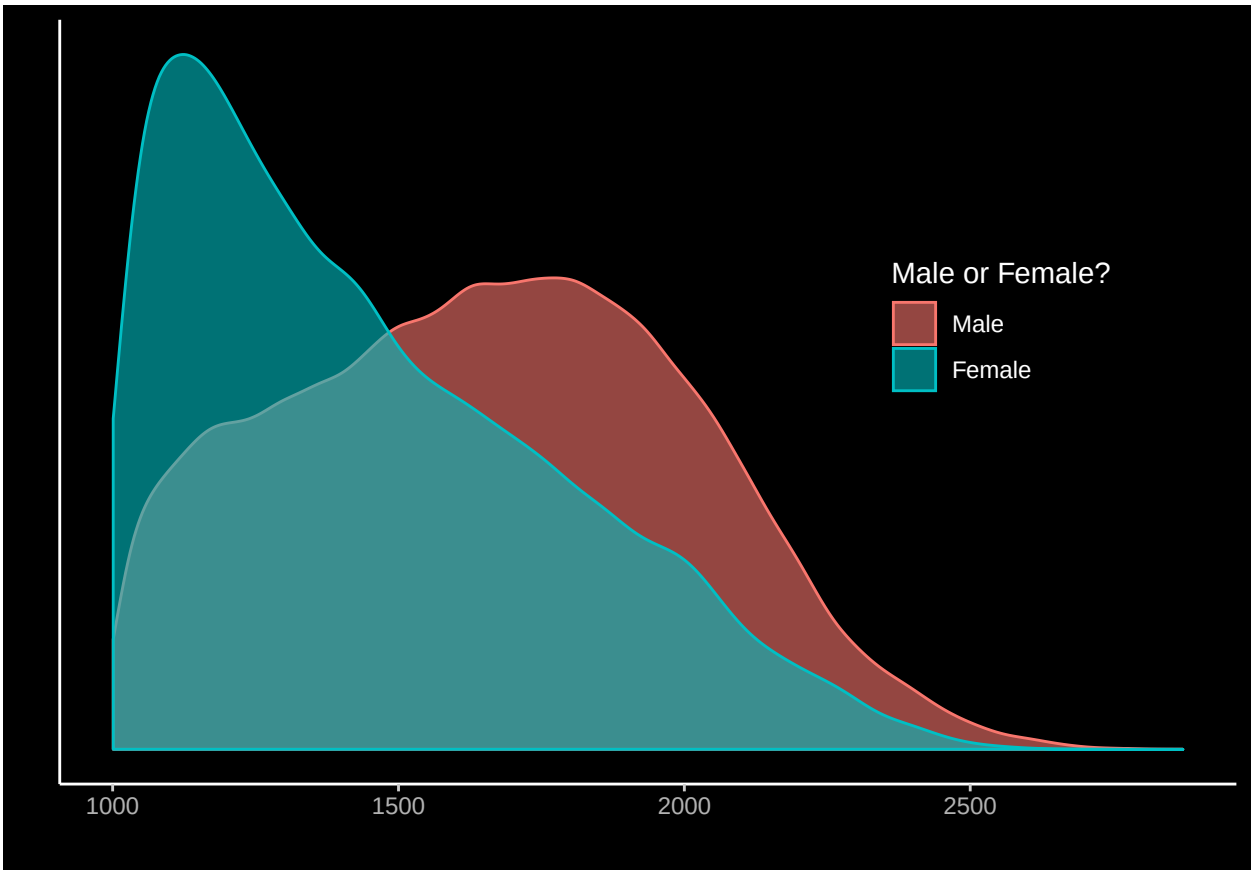
## Age vs Rating of titled players (December 2019)



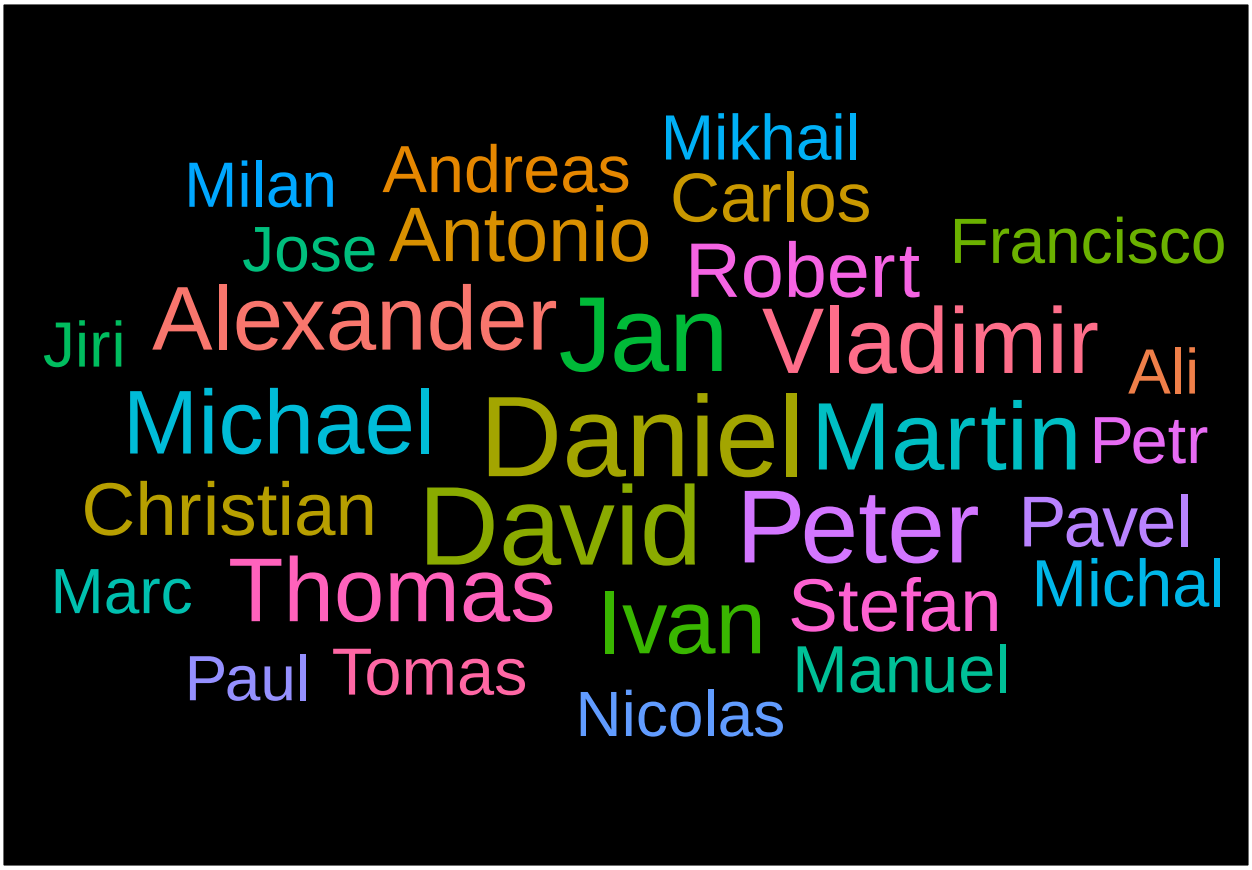

The most aesthetically pleasing graph to look at is the one above. It reveals a few aesthetically pleasing observations:

- Bands of players can be separated by titles categories: an obvious points is that the top blue band is all GMs, the highest rated group. Below the blue band are IMs (purple band), FMs (green band) and CMs (red band).
- CMs and FMs vary greatly across rating categories because it has become much easier for lower rated players to acquire titles in youth tournaments and via inter zonal tournaments.
- There is a slight negative correlation, among all titled player groups, between Age and Rating.

## Males/Females rating distributions (December 2019)




Above is a plot showing the stark difference between male and female rating distributions for the previous month. I'll look to explore why the female rating distribution is sharply skewed right compared to the male counterpart.


Mikhail
Milan
Andreas
Carlos
Jose
Antonio
Francisco
Robert
Alexander
Jan
Vladimir
Jiri
Ali
Michael
Daniel
Martin
Petr
Christian
David
Peter
Pavel
Marc
Thomas
Michal
Ivan
Stefan
Paul
Tomas
Manuel
Nicolas


Andersson
Martin
Johansson
Horvath
Garcia
Toth
Schmidt
Mueller
Szabo
Silva
Nielsen
Wang
Larsen
Kim
Chen
Hansen
Lee
Tan
Liu
Nguyen
Li
Andersen
Pedersen
Jensen
Ivanov
Christensen
Zhang
Nagy
Sorensen
Da Silva