Stats 133 Final Project Report

William Cheung & Matthew Weglicki

Due December 14, 2015

Introduction

Tennis is one of the most popular sports in the world, drawing in billions of viewers every year, and so the amount of statistics available on the sport is appropriately plentiful. Thus, our undertaking in this project is going to be analyze the performance of individual countries since 2010 in the sport's most popular event - men's singles. In particular, the question we want to answer is, are there any countries which have dominated the sport in men's singles since the turn of this decade, and if so who are they? We have taken a two prong approach to analyzing "dominance" - analyzing both year-end ranking statistics, and comprehensive match statistics since 2010.

In terms of rankings, we set a standard of only considering the top 100 players in the world in a given year and proceeded to investigated three aspects of the rankings which we thought would effectively reflect dominance. First of all we investigated how much representation each nation had in the top 100, top 50, top 20, and top 10, in order to analyze the depth of quality that each nation had. In other words to provide a much more probing analysis by seeing whether dominant nations had strong representations in all positions within the rankings and to expose countries who on an absolute scale might seem to be "dominant" within the top 100, but in fact have 9 out of 10 of their players ranked below position 90, for example. Secondly, we looked at the evolution of top 100 rankings over the years, to see which countries have consistently had strong representation, and which might have appeared to be dominant only for brief amounts of time. Lastly, we used the ranking points of the top 100 players to measure the average number of points each country had per player in the top 100, and used that as a proxy for how good the average player in each country is, which allowed for further interesting comparisons.

The second criteria we used to determine which countries were most dominant were match statistics of every single match played since 2010. Since individual match wins are accounted for in the ranking points, we used the total number of titles each country has won since 2010 as the measure of success in this category. Tennis is also a discipline which is played on three different court surfaces - Clay, Grass, and Hard - as well as having three different "levels" of tournaments on the professional tour - ATP, Masters, and Grandslam. The level of each tournament is indicated by the amount of points it offers in each round of the tournament, with Grandslams offering the most points to participants (and hence being the highest level of play), followed by Masters events, and then ATP tournaments offering the least amount of points on the tennis pro tour. With this in mind it seemed appropriate to offer sub-questions of whether the dominant nations in tennis had a particular affinity to certain surfaces and what how their performance changes depending on the level of the tournament. This was achieved by examining how many of the total titles that a country won were from a particular surface and level.

Data Extraction and Cleaning

We were very fortunate in that we chose a topic which did not have a lack of data. The source of our data was Jack Sackmann's (a professional sports analyst) public GitHub repository, found at https://github.com/JeffSackmann/tennis_atp. There we found raw data containing comprehensive men's singles rankings and match statistics for the years 2010 - 2015 (his larger tennis repository dated back to rankings and statistics for both men and women's singles back to 1968!). All of the data collected, although bountiful, was very messy, at times cryptic and required significant manipulation to be transformed into the "clean" data we could use. For example, the match statistics for each year contained well over 2000 entries, over 49 different columns, being so detailed as to include the handedness and serving percentages of each player involved

in the match. The extraction and cleaning of this particular data was simply achieved by downloading the raw data into the appropriate raw data folder via the download.file() function in R and then creating a new csv file in our clean data folder via the write.csv() function, selecting only the columns applicable to our analysis (tourney name, surface, tourney level, match num, winner id, winner ioc). At this stage we excluded tournaments of inappropriate level that shouldn't have made it into the raw data (Challenger tournaments), and exclusive tournaments where gauging a nation's strength relative to all other countries is difficult to analyze (Davis Cup and Tour Finals). Other things to note are that during the downloading of data, warning messages popped up indicating that there were some problems. Upon inspection this was due to a large number of "NAs" showing up in the latter, very specific categories of the data, categories which were superfluous to our needs and hence did not affect our analysis. The rankings raw data was much more awkward to manipulate. The rankings were given on a weekly basis of over 1000 players involved in the professional tennis tour, and to add to the complication the rankings for 2010-2014 were in the same file and did not have the names or nation representations included. They only included the week of the ranking, ranking number, player id and ranking points. Thus cleaning involved selecting only the top 100 players for the very last week of rankings for each year, and creating separate dataframes for the years 2010-2014. Once that was achieved we matched the player_id of everyone in the top 100 rankings with their ID found in the players.csv file from the same GitHub repository and so imported the names and countries of each player into our rankings data.

Data Manipulation and Processing

Match Statistics Processing

For the match statistics analysis, the main task was converting data on every match played in a given year, to matches which represented the finals of tournaments, and hence could be representative of how many titles a certain country won. The process for doing so began with first creating the functions which when given the match statistics from certain years, are able to extract those matches which were "title matches", where the winner won the entire tournament, and output the winning nationalities, surface, and level of each tournament in that year. The we move onto applying those functions in order to create dataframes for each year containing information about the winning nationality, surface and level of each tournament in that year. We then consolidate all these dataframes into one comprehensive dataframe, from which we will be able to conduct our appropriate analysis.

Rankings Processing

For the rankings analysis, we start with the clean data which includes the top 100 players with their ID, rank, points, date of birth, and country. From this we will most closely use rank, points, and country. For this rankings analysis, we wanted to analyze a country's depth, meaning if they have produced multiple top players, or if the country only has really one or two top players only. We break it down into three parts.
For the first part, we looked at the rankings from 2015 and created a mapping using a list, where each item of the list is a country, then its elements are the ranking of their players in the top 100. For the second part, we wanted to track the evolution of the rankings between 2010 to 2015. Based off the top 100 rankings for those years, we could count how many of players those countries had in the top 100. Then, with the different years, we can see how consistent those country's players performed. For the third part, we use the number of points each player received for playing and winning tournaments. We can use this to determine the average number points a player in the top 100 accumulated, and then compare this average to the players from each country. This is to determine which country produces players in comparison to the average. Therefore, we can determine the top countries by the points accumulated.

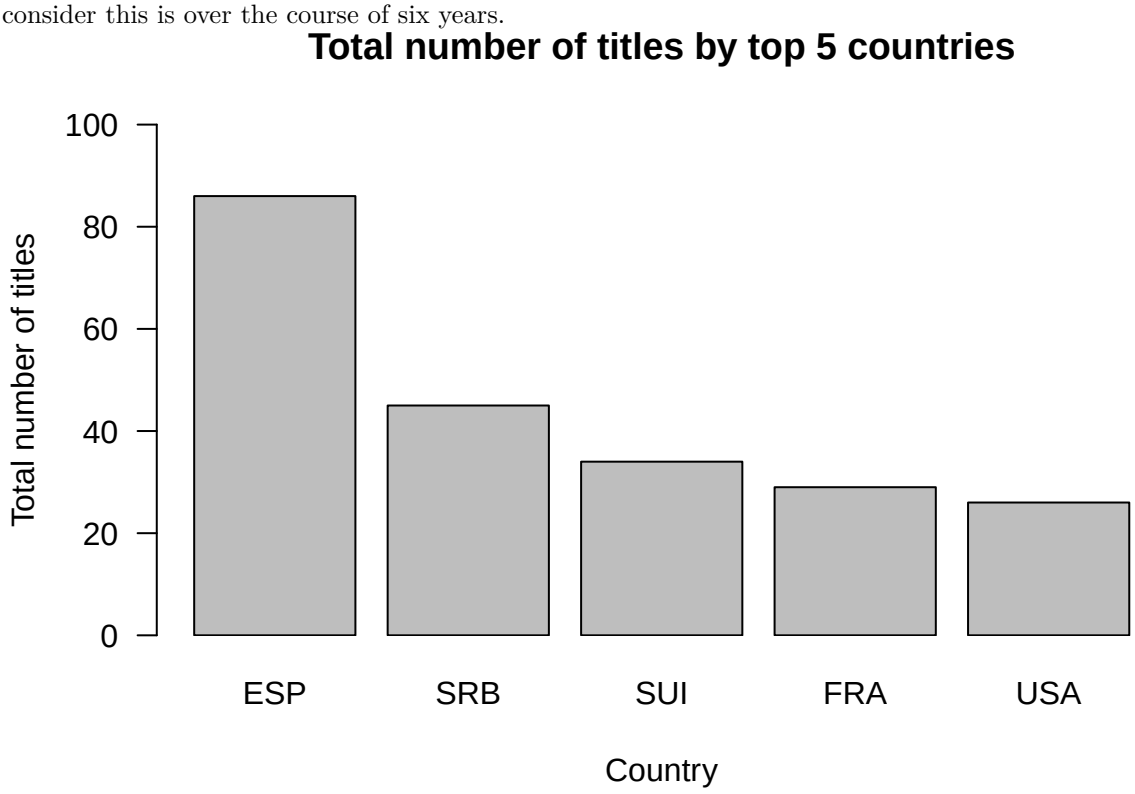
Data Analysis

Match Analysis

From the match analysis first of all we see that in the 2010-2015 period a total of 32 different countries have won titles. However, an overwhelming large proportions of titles have gone to the top 10 nations as is demonstrated by the organised list of countries with their appropriate number of titles below.

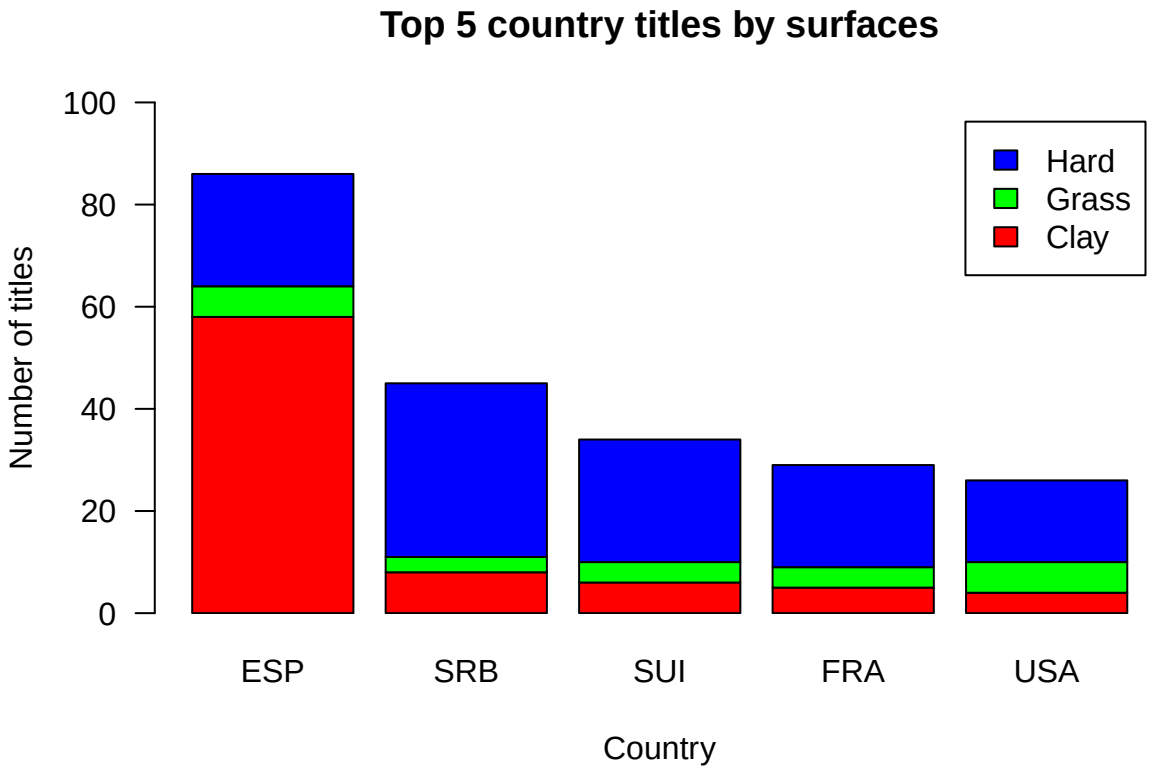
```
## ESP SRB SUI FRA USA ARG GBR CRO CZE JPN RUS GER CAN SWE ITA AUS LAT AUT
##  86  45  34  29  26  24  22  15  11   9   9   8   7   7   6   6   6   6
## BUL UKR URU SVK RSA BRA KAZ POR BEL NED CYP FIN UZB DOM
##   4   4   3   3   3   3   2   2   2   2   1   1   1   1
```

Furthermore, we see Spain in particular has been very dominant in the sport during this time, obtaining almost twice as many titles as Serbia in second place, as demonstrated in the barplot below. After Spain however, there is a group of countries bunched together, with not much separating them, especially when you consider this is over the course of six years.



Next we can evaluate that once within this exclusive group of top 5 countries whether any of them have been performing particularly well on certain court surfaces by first of all looking at the decomposition of the top 5 countries' titles by surface in table form, and then analyzing another barplot of the data.

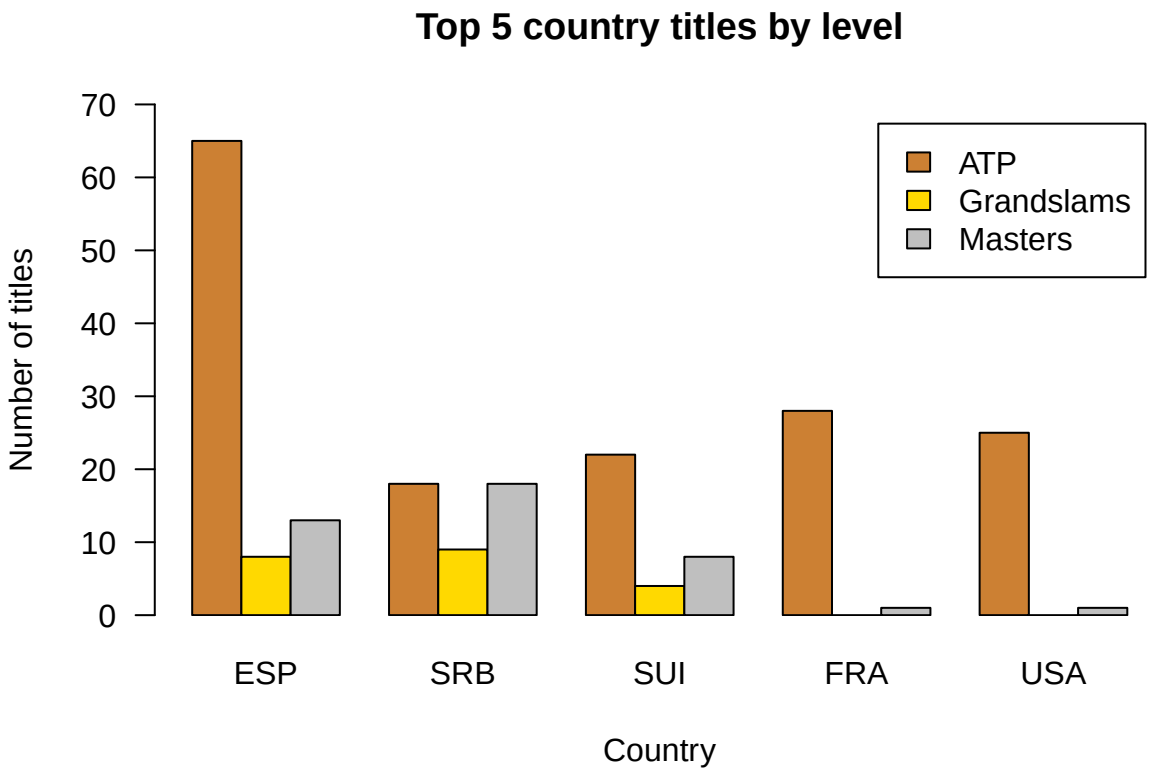
```
##   country Clay Grass Hard
## 1     ESP   58     6   22
## 5     SRB    8     3   34
## 6     SUI    6     4   24
## 7     FRA    5     4   20
## 3     USA    4     6   16
```



Overall, the general trend within this data is that most of the top five countries have won most of their titles on hard courts, followed by clay courts, and least of all on grass courts. This is unsurprising given that the distribution of matches palyed on surfaces follows that exact trend - most matches are played on hard, followed by clay, and least on grass throughout the tennis season. The major buck in this trend however is in the case of Spain, who have won a staggering 58 out of 86 titles on clay. Coupling this with the fact that clay court is not the most played on surface during the tennis season seems to highlight that Spain's affinity to clay courts has played a very significant role in its dominance, in terms of total titles won, of men's singles since 2010. Other trends to note are that Serbia appears to have a particularly high success rate on hard courts, winning 76% of its titles on the surface, and that relative to how little time is allocated to grass courts during the season (about 6 weeks of the year), the USA appears to favor the surface significantly over clay courts.

Lastly for match analysis it is time to see how the top 5 countries perform at varying levels of the game. Whether certain countries shine when it really matters, and whether others perform better when there is less pressure. Below are once again the total titles won by each of the top 5 countries, this time separated by the level. Once again this is presented in table for first and then with a graphical representation of the data.

```
##   country ATP Grandslams Masters
## 1     ESP  65          8      13
## 5     SRB  18          9      18
## 6     SUI  22          4       8
## 7     FRA  28          0       1
## 3     USA  25          0       1
```



The overall trend that can be deduced from this data is such that for most countries, the majority of their titles come from lower level tournaments (ATPs), which is unsurprising, as once again the make up the vast majority (60%) of the tennis annual tour. There are several other things to point out about this data. First of all there appears to be a division of quality between the top 3 and France and the USA. This analysis comes from the fact that although France and the US are still close to total number of titles of Switzerland and Serbia, neither of the bottom two appear to perform particularly well on large occasions, with neither winning any Grandslams, and aminiscule amount of Masters titles. Secondly, although Spain appears to have an overwhelmingly large number of ATP titles compared tot the other countries, it is not so dominant when it comes to higher level Masters and Grandslam events. Lastly, it is apparent that Serbia performs well under pressure, having won the majority of its titles not in ATP events (despite that making up most of the tour), and having the greatest number of Grandslams of all the nations in the top 5. What can be drawn from this analysis is that when it comes to the big occasions, Serbia is most certainly the dominant nation, Spain still puts up very respectable performances at the big events, but dominates the lower ATP events (winning over 40% of all those tournaments since 2010), whereas France and the USA are nations which do significantly better when the stakes are lower. In terms of segregation by level, Switzerland follows a very typical trend - seeing a steady decrease in the number of titles as the level of the tournament increases.

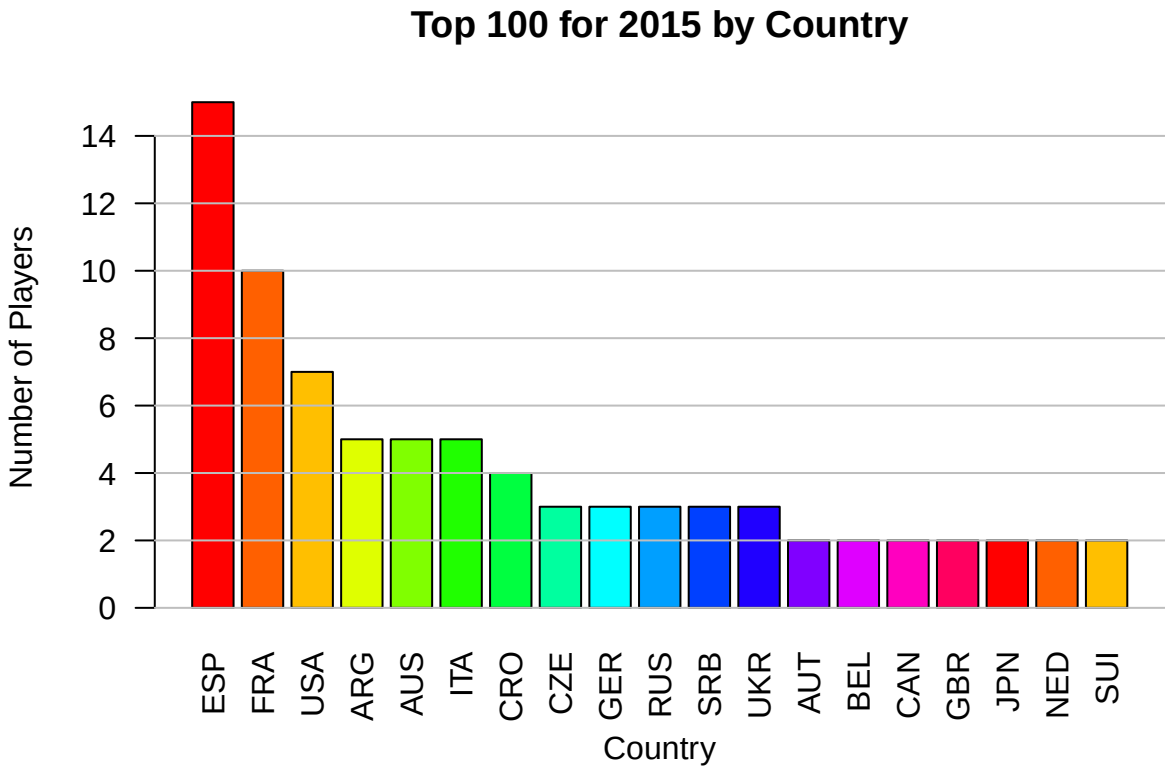
Rankings Analysis

We will use the Tennis ATP Rankings from 2015 to determine the competitive depth of a country's tennis representation. First we will read in the data from clean_data directory. We want the top 100 players from each year, and their corresponding country they represent. We have three parts to this analysis.

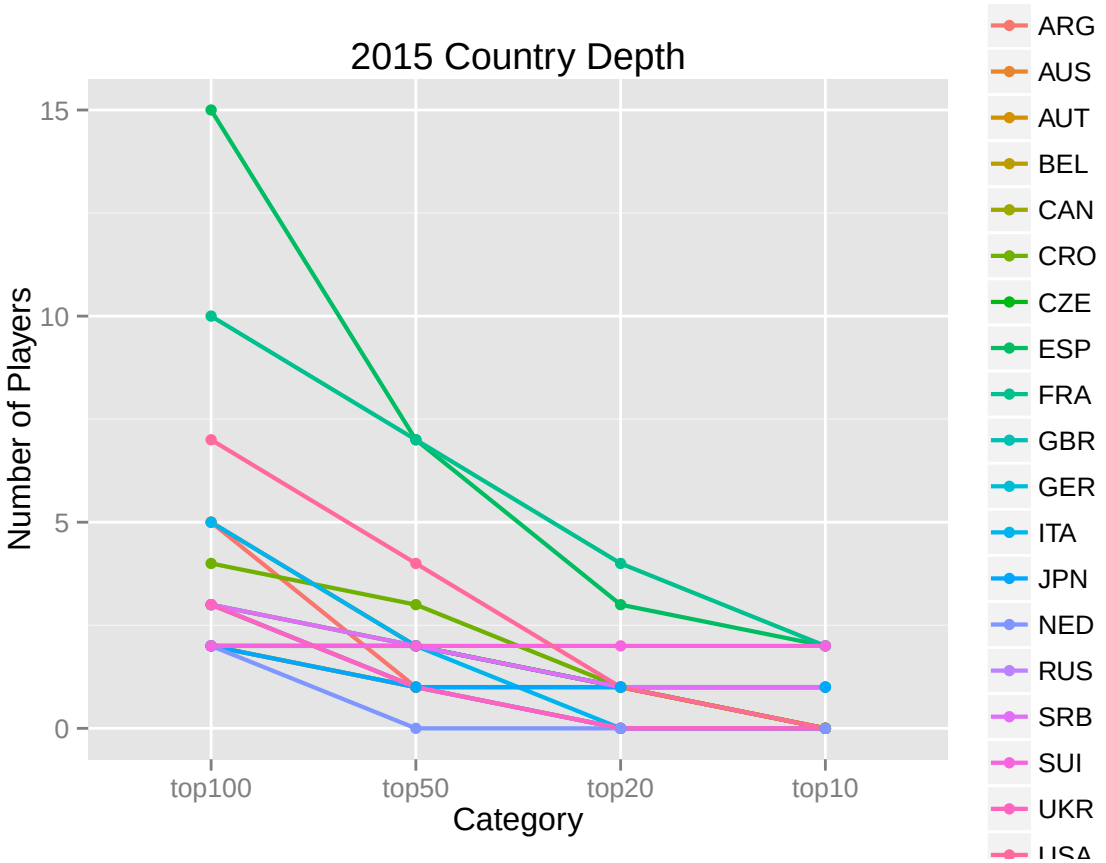
Part I was accomplished by manipulating our data first into a table and then into a data frame. First we defined depth as a country with more than one player in the top 100. This means the country has more than one representative in the top 100. Then we went even more specific to look at the top 50, top20 and top10. The following table (only showing first 6 rows) represents each country and the amount of players in each category for 2015. The amount of countries shown is filtered by having at least 2 players represnted in the top 100.

```
##     top100 top50 top20 top10
## ESP     15     7     3     2
## FRA     10     7     4     2
## USA      7     4     1     0
## ARG      5     1     0     0
## AUS      5     2     1     0
## ITA      5     2     0     0
```

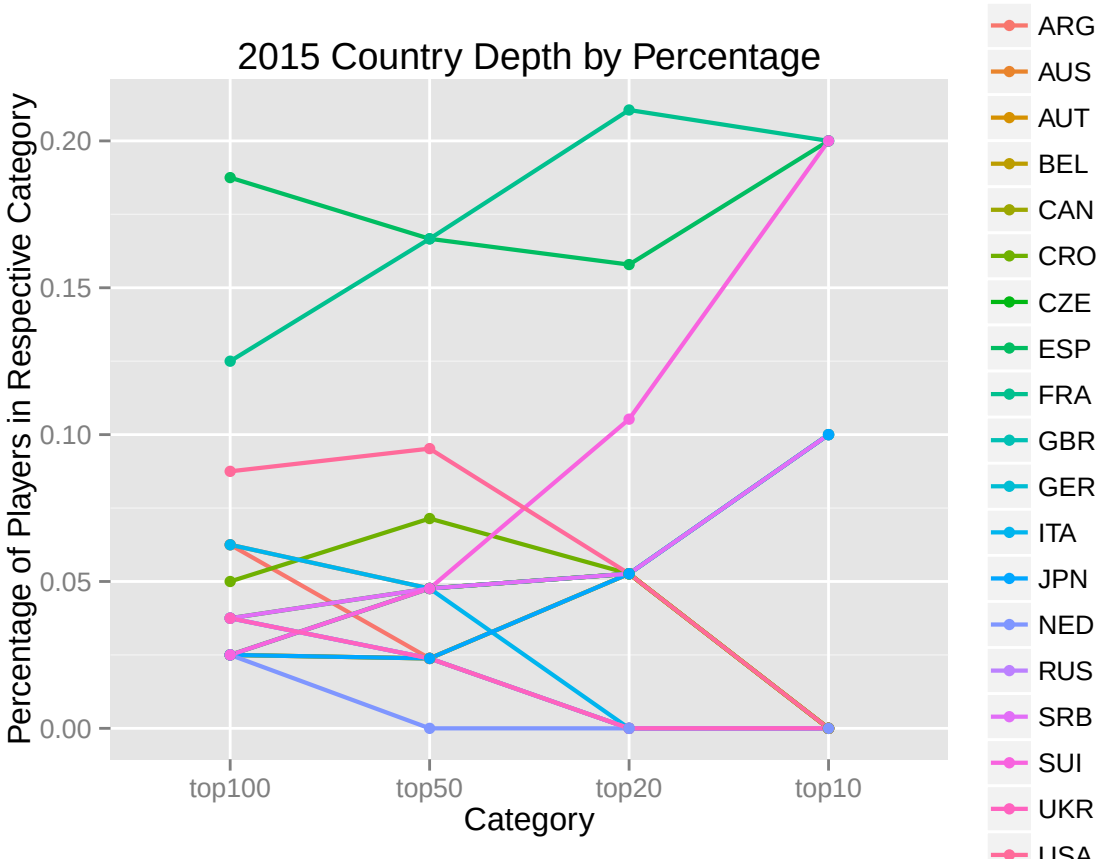
We created basic bar plots to visualize these categories individually. We will only show top 100 bar graph for this writeup, since these graphs were intermediate for further analysis. Please see /code/analysis_rankings.Rmd for further details of these graphs.



After viewing the top 100, 50, 20, and 10 in 2010 by country on their own, we will now see if there is a difference in depth by tracking how many players were in the top 100, 50, 20, and 10 per each country together. We will do this by raw number of players, then percentage of players in that respective category. We use ggplot to combine all these findings and figure out the dominant counties.

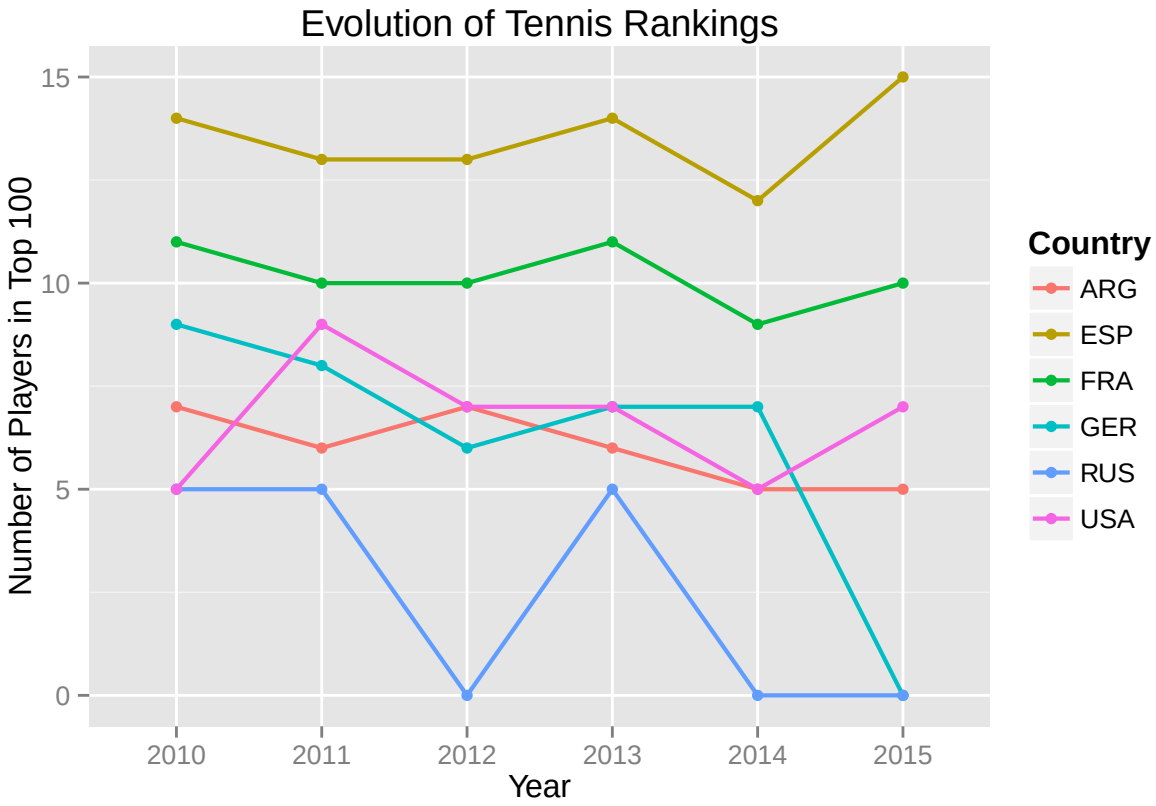


From Part I, we can look at our graph "2015 Country Depth" and see that Spain (ESP) was the most dominant, akin the most players in the top 100, and held consistently even in with 2 players in the top 10. Also, France (FRA) had a strong showing, with 10 players in the top 100 and most players (4) in the top 20, and a tied for most players in the top 10 as well. Another thing to note is that Switzerland (SUI) has only 2 players in the top 100, but those 2 players also happen to be in the top 10! Not only did we find the raw number of players in each category, but we can try to see the weight of each country in the category as well, so we used percentage of players in the groups.



We can look at our other graph, "2015 Country Depth by Percentage" to see that indeed, Spain (ESP), France (FRA), and Switzerland (SUI) dominated the top 10. One interesting note is the USA dropped out of the top 10 after a solid showing in the top 100 and top 50.

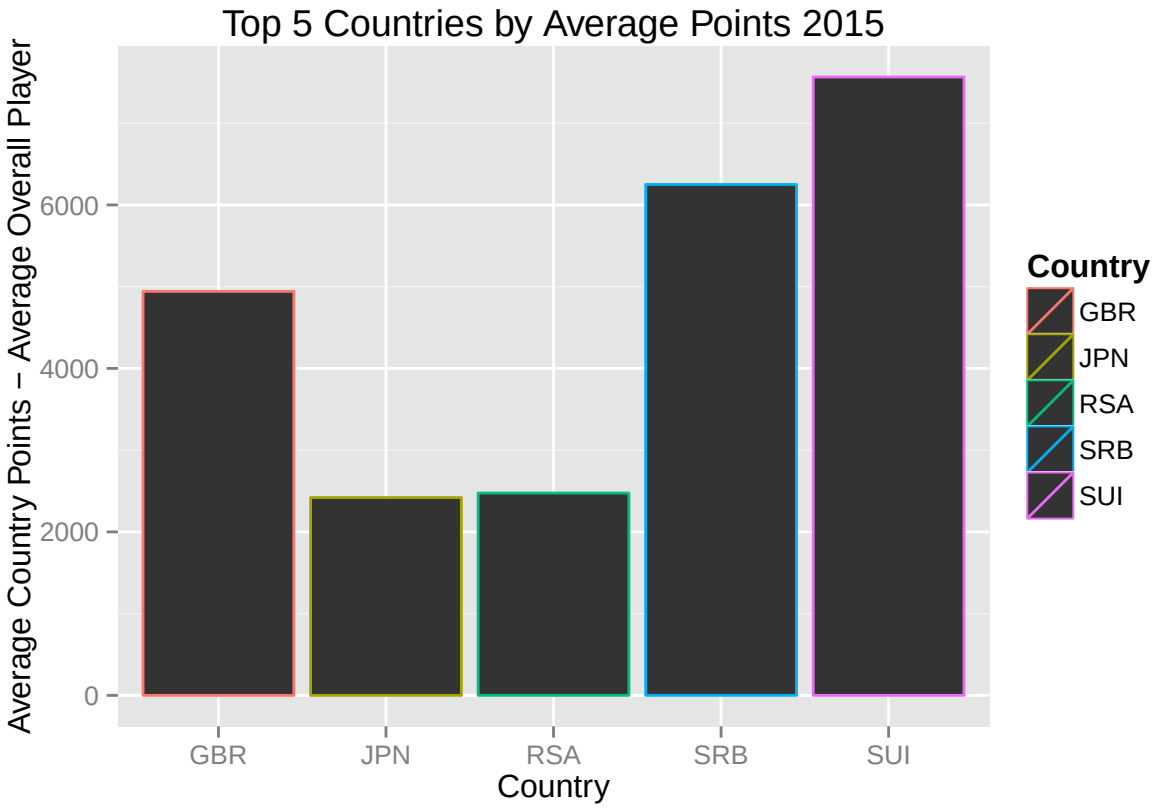
For Part II we had to merge the different data frames into one so we could visualize this in one graph. Therefore, we took advantage of the table and merge functions in R. We also had to reshape our new table to be ready to use ggplot correctly. We want to show the evolution of rankings - a year on year change of how many players in each category to show most consistent countries, based off the starting point of year 2010.



From Part II, we analyzed the movement and evolution of rankings of these countries from 2010 - 2015. We use "Evolution of Tennis Rankings" to determine the consistency of these countries. Just from looking at the graph, we can note that Spain was consistently on the top, with France a consistent second rank. After those top two, it gets a bit fuzzier. USA jumps from having 5 to 9 players in the top 100 from 2010-2015. But what is discouraging for Germany is that they are trending downwards. From 9 in the top100 in 2010, they dipped to 6 in 2012 and struggled to having no more players in the top 100 by 2015.

For Part III, we want to find out how good the average player is; in other words, how much a player contributes to his country's ranking on average. We will use data from 2015. Average number of points per any country: this is the sum of all points / number of unique countries = averae points for a country. Our metric for determining how good a player is will be comparing the number of points that player has versus the numbr of points the average player has. Therefore we have: number of points of the country / # of players in that country = how good that player is.

```
##     difference
## SUI       7565
## SRB       6250
## GBR       4942
## RSA       2475
## JPN       2419
```



From Part III, we look at our graph "Top 5 Countries by Average Points 2015" to see which are the strongest overall countries by the amount of points they accumulated in comparison to the average. Immediately we see Switzerland, Serbia, and Germany in the top three. Something interesting is that Spain is no longer a part of the top 5 countries, though previously they showed in the depth rankings. If we looked at our table, "country vs avg 2015 sorted" we would notice that Spain was ranked 9th in overall points.

Conclusion

From the analysis conducted from match statistics it can be deduced that even within the group of countries that appear to dominate the sport in men's singles, there is a division between how much they divide. Analysis of the total number of titles showed that Spain, as number one, appears to completely overwhelm the others in terms of absolute number of titles won, which after further analysis can be attributed to particular dominance on clay courts and lower level tournaments. From the analysis of titles by level of tournament, we can also conclude that there is a significant divide between the top 3 nations -Spain, Serbia, and Switzerland- and the rest of the world in terms of quality of players, with those nations having dominated most of the higher level tournaments in the sport since 2010. From these rankings we can conclude that the dominant countries of tennis, from 2010 to 2015 are, in no particular order: Spain, France, Switzerland, and Serbia. These findings do make sense if we look at the individual players. For Spain, top players like Rafael Nadal, Feliciano Lopez, and Roberto Bautista Agut boost the rankings for Spain. For France, Fichard Hasquet, Jo Wilfried Tsonga, Gilles Simon, Benoit Paire, among others help their strong showing at the top. Switzerland boasts two of the top 5 players in Roger Federer and Stanislas Wawrinka, but apart from those two, Switzerland doesn't have much depth. Finally, Novak Djokovic, the current World Number 1, has bolstered Serbia's rankings mostly by himself, with fellow countryman Viktor Troicki also in the top 50.

Final Remarks

By doing this project to look through a mine of data, we could really analyze and draw strong conclusions for tennis. Sports and data is such a thriving place for exploration, and we are glad that we chose this topic for our project. By using R, we have gained the tools to do even more data mining in the future. If we had more time, we could go in more depth with the individual's statistics over their career, as that could be an interesting path to investigate. Overall, we are impressed by how much data is available, as we barely scratched the surface of the abundant amount of analysis we could do. But we are proud of our work, as we both learned a lot about tennis and about data mining by doing this proejct!