Zishen Zhang zz363
MSDS
Data Wrangling and Husbandry 16:954:597
April. 26, 2019

Final Project: The Relationship between ATP Matches and World Ranking

Part I: Intuition and Brief Introduction to the Project:

Tennis, originated in England from 19th century, has become one of the most famous sports that people enjoy themselves while playing or watching the games. Among a great number of professional tennis players in the world, Roger Federer must be considered as one of the best players in the last 15 years. 38 years old sounds unreasonable for an athlete to be at the top of his/her field, but Federer simply does so. While watching Federer's games, audiences easily become addicted to the skills he applies during the game and strong performance of his basic training.

The most exciting moments for tennis fans each year must be Grand Slams: Australian Open, Roland Garros, Wimbledon, and US Open. Many professional players consider Grand Slams as platforms to demonstrate their professional abilities as well as the effort they have made so far. Even more intriguing, each title of Grand Slams counts as 2,000 points, which almost takes ¼ of the total annual points for many top players. What's more, there's a series of games, ATP1000 Masters Series, that professional players have to attend each year since such points are required to be counted to world ranking. For some of those Master Series tournaments, many professional tennis players consider them as warm-up opportunities before Grand Slams since the court condition is quite similar.

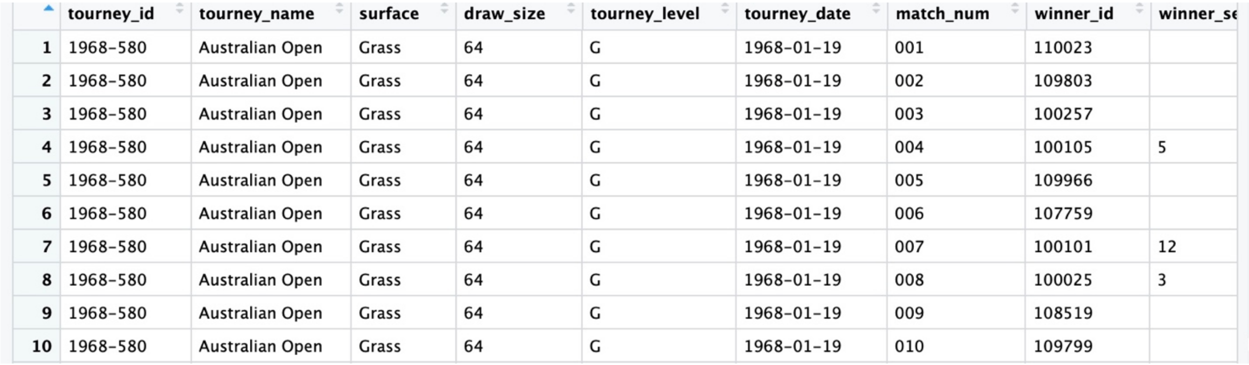
Since Federer got his first title of Grand Slams by the end of 2002, historical data from 2003 to 2016 is chosen from ATP.csv, which contains all matches' data from 1968 to 2016, to figure out the winners of Grand Slams and Masters Series. ATP official website provides year-end world rankings. Surprisingly, the result of world ranking data from ATP website shows some interesting patterns, which will be discussed later in this paper.

Part II: Data Cleaning for ATP.csv from the Kaggle Website and Rankings from ATP

For every project, it's always nice to do some preprocessing over the dataset we get since not all variables in the dataset are useful for upcoming analysis. Redundant data not only occupies extra spaces of computer memory but also raises the probability of choosing incorrect variables that creates some misleading results then.

After taking a look at ATP.csv file inside RStudio, it can be easily found that many columns have no values recoded (NA), which requires to be deleted from the original file for convenience. The original csv file has 49 columns, which has been greatly modified to only 24 columns then. Moreover, while extracting Grand Slams records, some misnomer problems have to be resolved as well. For example, "Australian Chps." appears in the column "tourney_name" because Australian Open used to be Australasian Championships, with the name change occurred after 1969. Similarly, some data in the table is recorded as "Us Open" instead of "US Open," which will cause troubles matching the name of the title winner with the title names. Last but not least, the "tourney_date" variable in the table was once considered as integers data type, which ought to be clarified by switching it into year-month-data format. After a thorough cleaning on the original dataset, the modified version seems to be more concise, which will benefit others who might be interested in the same topic as well.

While looking for year-end world ranking data from ATP official website, only top 10 players are chosen as candidates since the total points other players get by the end of the year become too low to be considerable on finding the trend of rankings and titles. Low points imply that such players either didn't do well in Grand Slams and Masters Series or they were absent from those games. After scrapping the data from the website and opened in RStudio, there are 2 columns that consist of missing values. Deleting those 2 columns make the dataset look tidy. The timestamp selected is the first week of the next year in order to make sure that all matches' results from the last year count.



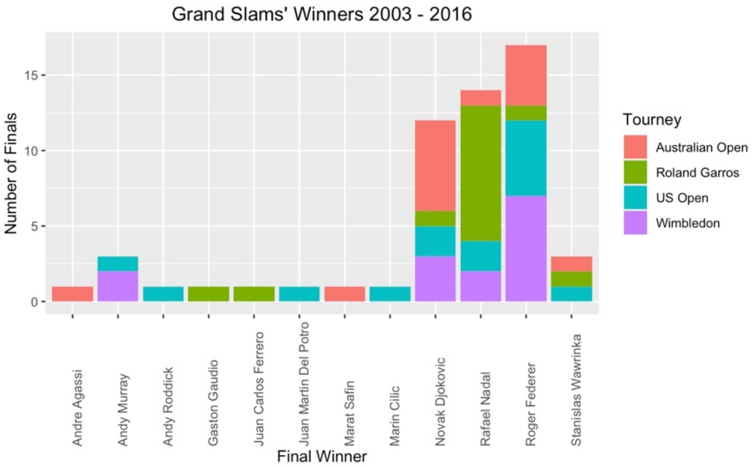
	tourney_id	tourney_name	surface	draw_size	tourney_level	tourney_date	match_num	winner_id	winner_se
1	1968-580	Australian Open	Grass	64	G	1968-01-19	001	110023	
2	1968-580	Australian Open	Grass	64	G	1968-01-19	002	109803	
3	1968-580	Australian Open	Grass	64	G	1968-01-19	003	100257	
4	1968-580	Australian Open	Grass	64	G	1968-01-19	004	100105	5
5	1968-580	Australian Open	Grass	64	G	1968-01-19	005	109966	
6	1968-580	Australian Open	Grass	64	G	1968-01-19	006	107759	
7	1968-580	Australian Open	Grass	64	G	1968-01-19	007	100101	12
8	1968-580	Australian Open	Grass	64	G	1968-01-19	008	100025	3
9	1968-580	Australian Open	Grass	64	G	1968-01-19	009	108519	
10	1968-580	Australian Open	Grass	64	G	1968-01-19	010	109799	

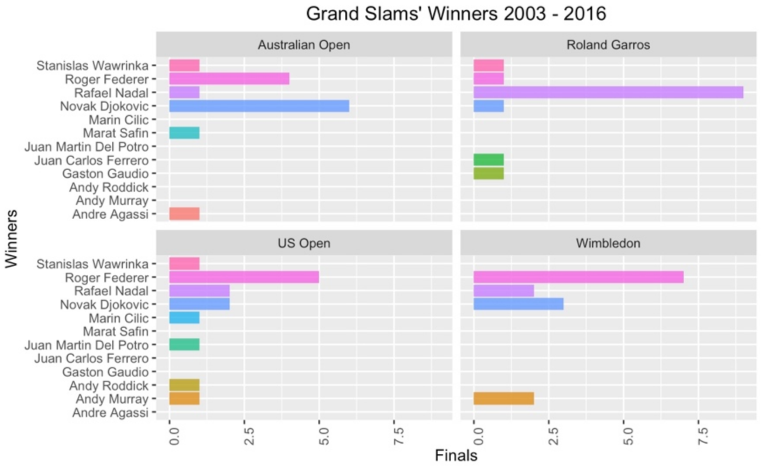
Part III: Analysis Based on Grand Slams Results between 2003-2016:

Grand Slams consist of 4 tournaments: Australian Open, Roland Garros, Wimbledon, and US Open, with Australian Open and US Open playing on hard courts, while Roland Garros playing on clay courts and Wimbledon on grass courts. Hard courts are more common, as most of the courts in US are. Players who are good at finding opportunities near baseline get used to this kind of courts the best. Clay courts benefit more for players who are trying to make strong spins on their shots such as Rafael Nadal. Grass courts must be the most challenging one because the patterns of balls' movement are less predictable, which requires players to have comprehensive abilities on dealing with any issues that happen during the games.

To grasp matches data from all historical records, filter setting was made at first by choosing the data with tourney_level='G', which stands for Grand Slams. Some variables like tourney_id and date range are also set to make the searching result more specific. Furthermore, by selecting 9 correlated variables from the filtered data frame and tournaments after 2003 only, the desired data frame is achieved. Tidyverse, a useful package which allows users to get structured data frame and to operate on visualization more easily, is applied in reaching the desired result.

By creating a general bar plot and a set of bar plots with each tournament respectively, it can be easily seen that Roger Federer, Rafael Nadal, and Novak Djokovic are doing so well that the total number of titles each of the three players obtained equals to or even greater than the sum of the numbers of titles of others players in the same graph. Roger Federer, the first man who got 17 Grand Slams titles in ATP history, has dominating performance on hard courts and grass courts, but clay courts seem to be a little tough for him. Rafael Nadal, the most interesting player in the graph, got 14 titles in total, while 8 were earned at Roland Garros. Novak Djokovic's situation is quite similar to Roger Federer. He has already got 12 grand Slam titles by the end of 2016, while he is actually 6 years younger than Federer. Perhaps he just needs more time to catch the pace of Roger Federer in the near future.

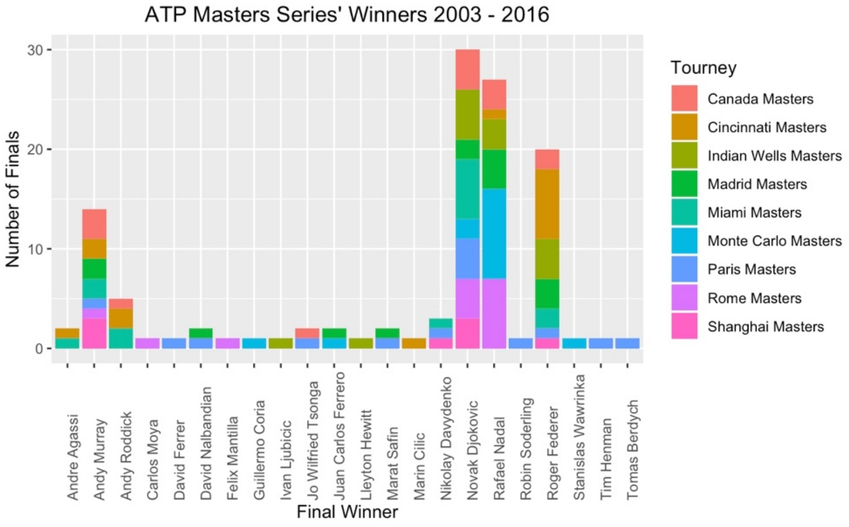


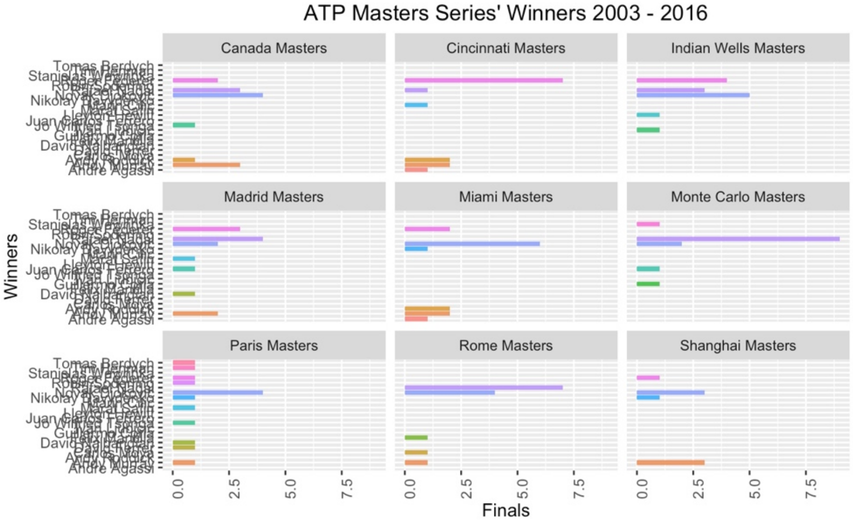


Part IV: Analysis Based on ATP1000 Masters Series:

Checking basic information about ATP1000 Masters Series, there used to be 10 tournaments included in Masters Series. However, starting from 2009, Hamburg Tournament was no longer considered as Masters Series. To ignore the effect of ATP regulation, those data points corresponding to Hamburg Masters are deleted from the modified ATP dataset.

By sifting valuable data points with "tourney_level" = "M," which represents Master Series and setting other filters as exactly the same as Grand Slams' analysis does, the bar plots that correspond to sifted data on ATP1000 Masters Series demonstrate that great performance of Roger Federer, Rafael Nadal, and Novak Djokovic still lasts, with Djokovic earning 30 titles, Rafael Nadal 27, and Federer 20. Besides the 3 MVPs above, Andy Murray also appears to have excellent performance over other players on the graph, with 14 titles obtained as the 4th place on the graph. Though Andy Murray didn't have as many Grand Slams titles as 3 MVPs, his great performance at Masters Series, especially between 2015-2016 when he got 5/14 titles in two years, definitely makes people curious about his world rankings of 2015 and 2016.





Part V: ATP World Ranking Analysis:

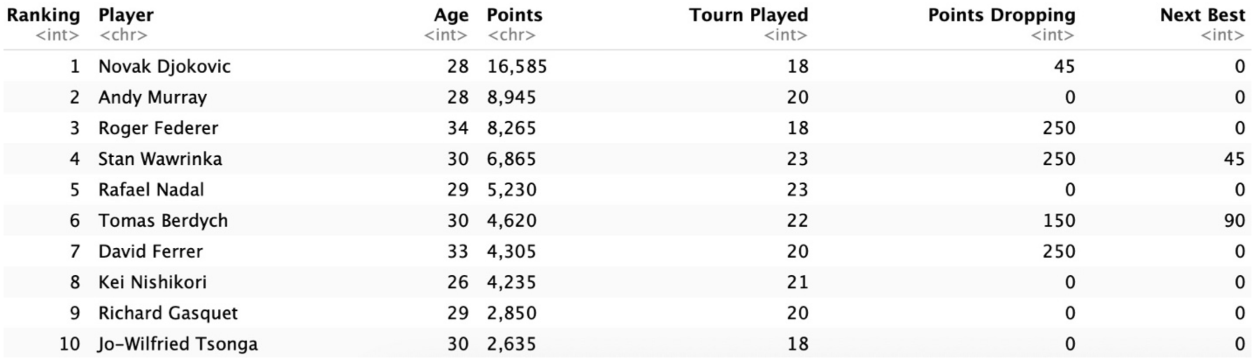
With the intuition as described in part V, the top-10 ranking data of 2015 and 2016 are scrapped from ATP website. An interesting pattern is observed in ATP 2016 ranking table: though Andy Murray just played 16 tournaments in 2016, he actually got the highest year-end point, which allowed him to be world no.1. What's more, Federer's name didn't appear in the top-10 list, which seems to be a little bit strange as well.

The unusual behavior of Federer's slip on world ranking is more straightforward. Federer got a knee surgery in February, which not only prevented him from attending Rio Olympic Games but also forced him to forfeit all tournaments for recovery. Less tournaments attendance definitely means less points given.

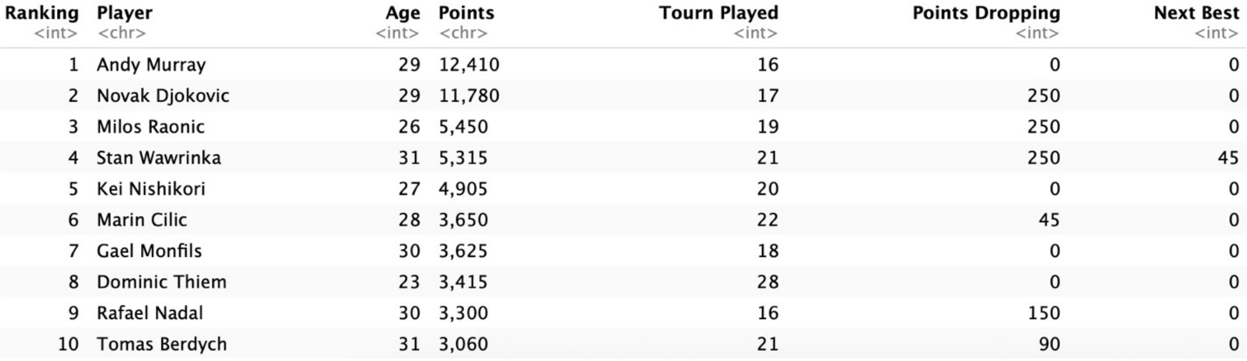
To understand the unusual behavior Andy Murray did in 2016, ATP tournaments points calculation method must be introduced. To avoid players dealing with tournaments passively after they've already got a lot of points, the point-calculation method is set to be dynamics, which means that ATP will choose best 18 results of tournaments, 13 compulsory tournaments (Grand Slams and Masters Series) and 5 best tournaments besides 13 compulsory ones, within the last 52 weeks (13 months).

52 weeks guarantee to have the results of a certain tournament twice, this year's result and that of the same tournament a year ago. If this year's result is better than that of the last year, the targeted player should be good. However, if there's a slip on this year's result, then certain points will be deducted based on which round of the tournament the player eventually entered. Therefore, either keep the same pace of certain tournament as a player did last year or do a better job on such tournament is the choice to save the place on ranking.

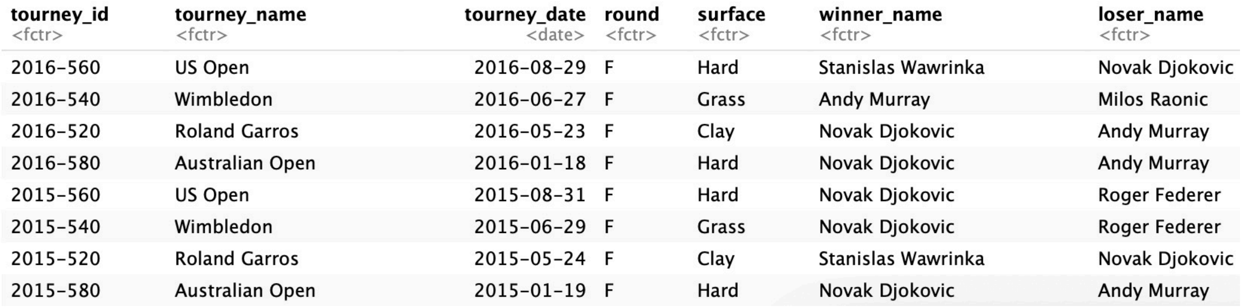
Now going back to Andy Murray's problem. Comparing Murray's performance during 2015 and 2016 on Grand Slams and ATP1000 Masters Series, it's surprised to find out that luck also behaves as a significant factor of his world ranking. From the statistical results, Andy Murray not only did pretty good jobs on Grand Slams and Maters Series in 2016 but also improved his performance on most of the tournaments he attended in 2015. His great competitor, Novak Djokovic, unfortunately, got a more challenging year in 2016, since he did extraordinary job on almost all tournaments he attended in 2015, which indicates that he had to do the same good job or even a better one to reduce the risk of losing points. The tradeoff between Murray and Djokovic is quite interesting, which requires calculation and effort on holding the world ranking.



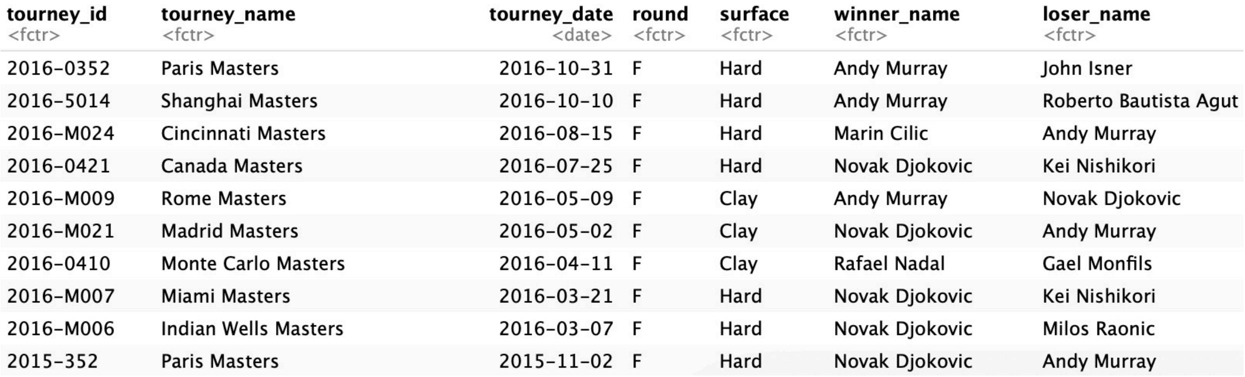
Ranking	Player	Age	Points	Tourn Played	Points Dropping	Next Best
<int>	<chr>	<int>	<chr>	<int>	<int>	<int>
1	Novak Djokovic	28	16,585	18	45	0
2	Andy Murray	28	8,945	20	0	0
3	Roger Federer	34	8,265	18	250	0
4	Stan Wawrinka	30	6,865	23	250	45
5	Rafael Nadal	29	5,230	23	0	0
6	Tomas Berdych	30	4,620	22	150	90
7	David Ferrer	33	4,305	20	250	0
8	Kei Nishikori	26	4,235	21	0	0
9	Richard Gasquet	29	2,850	20	0	0
10	Jo-Wilfried Tsonga	30	2,635	18	0	0



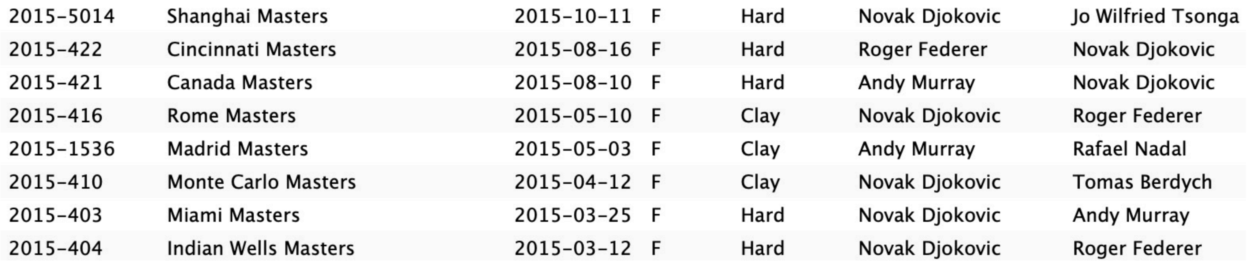
Ranking	Player	Age	Points	Tourn Played	Points Dropping	Next Best
<int>	<chr>	<int>	<chr>	<int>	<int>	<int>
1	Andy Murray	29	12,410	16	0	0
2	Novak Djokovic	29	11,780	17	250	0
3	Milos Raonic	26	5,450	19	250	0
4	Stan Wawrinka	31	5,315	21	250	45
5	Kei Nishikori	27	4,905	20	0	0
6	Marin Cilic	28	3,650	22	45	0
7	Gael Monfils	30	3,625	18	0	0
8	Dominic Thiem	23	3,415	28	0	0
9	Rafael Nadal	30	3,300	16	150	0
10	Tomas Berdych	31	3,060	21	90	0



tourney_id	tourney_name	tourney_date	round	surface	winner_name	loser_name
<fctr>	<fctr>	<date>	<fctr>	<fctr>	<fctr>	<fctr>
2016-560	US Open	2016-08-29	F	Hard	Stanislas Wawrinka	Novak Djokovic
2016-540	Wimbledon	2016-06-27	F	Grass	Andy Murray	Milos Raonic
2016-520	Roland Garros	2016-05-23	F	Clay	Novak Djokovic	Andy Murray
2016-580	Australian Open	2016-01-18	F	Hard	Novak Djokovic	Andy Murray
2015-560	US Open	2015-08-31	F	Hard	Novak Djokovic	Roger Federer
2015-540	Wimbledon	2015-06-29	F	Grass	Novak Djokovic	Roger Federer
2015-520	Roland Garros	2015-05-24	F	Clay	Stanislas Wawrinka	Novak Djokovic
2015-580	Australian Open	2015-01-19	F	Hard	Novak Djokovic	Andy Murray



tourney_id	tourney_name	tourney_date	round	surface	winner_name	loser_name
<fctr>	<fctr>	<date>	<fctr>	<fctr>	<fctr>	<fctr>
2016-0352	Paris Masters	2016-10-31	F	Hard	Andy Murray	John Isner
2016-5014	Shanghai Masters	2016-10-10	F	Hard	Andy Murray	Roberto Bautista Agut
2016-M024	Cincinnati Masters	2016-08-15	F	Hard	Marin Cilic	Andy Murray
2016-0421	Canada Masters	2016-07-25	F	Hard	Novak Djokovic	Kei Nishikori
2016-M009	Rome Masters	2016-05-09	F	Clay	Andy Murray	Novak Djokovic
2016-M021	Madrid Masters	2016-05-02	F	Clay	Novak Djokovic	Andy Murray
2016-0410	Monte Carlo Masters	2016-04-11	F	Clay	Rafael Nadal	Gael Monfils
2016-M007	Miami Masters	2016-03-21	F	Hard	Novak Djokovic	Kei Nishikori
2016-M006	Indian Wells Masters	2016-03-07	F	Hard	Novak Djokovic	Milos Raonic
2015-352	Paris Masters	2015-11-02	F	Hard	Novak Djokovic	Andy Murray



2015-5014	Shanghai Masters	2015-10-11	F	Hard	Novak Djokovic	Jo Wilfried Tsonga
2015-422	Cincinnati Masters	2015-08-16	F	Hard	Roger Federer	Novak Djokovic
2015-421	Canada Masters	2015-08-10	F	Hard	Andy Murray	Novak Djokovic
2015-416	Rome Masters	2015-05-10	F	Clay	Novak Djokovic	Roger Federer
2015-1536	Madrid Masters	2015-05-03	F	Clay	Andy Murray	Rafael Nadal
2015-410	Monte Carlo Masters	2015-04-12	F	Clay	Novak Djokovic	Tomas Berdych
2015-403	Miami Masters	2015-03-25	F	Hard	Novak Djokovic	Andy Murray
2015-404	Indian Wells Masters	2015-03-12	F	Hard	Novak Djokovic	Roger Federer

Part VI: Summaries and Conclusions:

Summarizing the work in this project, data cleaning process allows researchers who aim to find the relationship between games results and ranking to start the analysis more easily; the selection of significant tournaments ensures the validity of analysis since those tournaments not only represent the best level of ATP world but also make greater contribution to the point-counting process compared to smaller tournaments; the introduction to point-counting criterion of ATP tournaments illustrates the spirit of sports: hardworking and never give up.

For professional tennis players, it’s sometimes difficult for them to find the balance between the number of games played and potential point-loss risk. Wise players have to thoroughly think about the tournaments they ought to attend within a year based on estimated performance of such tournaments in the last few years as well as cautious point-counting calculations. Injury is evitable for sports players as well, so making a good choice might extend a player’s career life a lot, like Roger Federer did in 2016.

Hopefully from this project, people who have keen interesting in tennis can get a clearer view of how game results are closely correlated to the world ranking in ATP world, while people who are not so familiar with tennis can discover the wisdom in tennis.

Part VII: References:

VM, Sijo. "ATP Matches, 1968 to 2017." Kaggle. March 30, 2017. https://www.kaggle.com/sijovm/atpdata.

"Rankings | Singles | ATP Tour | Tennis." ATP Tour. https://www.atptour.com/en/rankings/singles?rankDate=2017-01-02.

"Rankings | Singles | ATP Tour | Tennis." ATP Tour. https://www.atptour.com/en/rankings/singles?rankDate=2016-01-04.

.