A Statistical Study on the Development of Modern Tennis

FIT5147 Project Report

Monash University

Name : Rongkai Shi

Student ID: 30937604

Introduction

The origin of tennis dates back thousands of years and it all started from a 12th century handball game in France called "Paume". From there, it was first adopted by the royal families and then became a prevalent sport in Europe (The Origins of Tennis - History of Tennis). Nowadays, it is one of the most valued sports in the world and its influence has even spread among non-European countries. However, traditionally, tennis is often considered as a richman's sport, which means it is only popular among the rich people (Tangirala). Based on that, the correlation between wealthness and tennis development is discussed here. GDP per capita and number of active professional players are used as the indicators here. Also, as a sport which requires intensive physical endurance and talents, it is natural to assume that players with an ideal body shape tend to have higher achievements in the professional career (Genclein97, 2015). The validity of this assumption is assessed in this report as well. Moreover, considering the fact that tennis is a relatively new sport to players from the East, this report seeks to make a comparison between eastern professional players and western professional players (Ning, 2018). Lastly, while tennis increases its worldwide popularity by having more tournaments in Asia and longer seasons, players are starting to complain on the long season and excessive tournaments which might harm or shorten their professional tennis career (Mcgrogan, 2009). Nonetheless, prize money also increased significantly with more tournaments and longer season and this provides players with better physiology and team support. So has the longer season shortened or lengthened tennis career life? Here, the report makes a comparison between the year of 2010 and 2020 to discover the changes in tennis player's career life. Therefore, in summary, this report aims to discuss the following questions:

1. Is it true that tennis is more popular in developed countries? How is GDP per capita a factor in determining the popularity of tennis in different countries?
2. What is the most ideal body shape for a modern tennis player in terms of height and weight?
3. How have eastern players and western players performed differently in their career paths?
4. Since 2010, has the increased number of tournaments lengthened or shortened the career length of tennis players?

Data Wrangling

Part 1. Wrangling with Raw data

In total, there are 10 sets of raw data used to answer the aforementioned questions. However, the raw datasets have different formats and large amount of unsorted information. Significant amount of data wrangling is needed in order to answer those questions. The 10 raw datasets are listed below:

1. Female player profile (~30000 rows × 13columns, tabular data with spatial elements and texts, https://www.rank-tennis.com/en/profile/wta)
   This dataset contains the name, nationality, current rank, career high rank, date of birth, height and weight of all female professional players. As we are only interested in the active players, the list of active players could be retrieved through ordering the data by their rankings from the website and only take those with valid rankings. The data was then processed in Excel to reformat the table heading into one single row and stored as a csv file. The wrangling process was completed via R studio. as.character(), as.Date(), as.integer() were used to convert columns to their correct format. Furthermore, rows with Nan data are removed via na.omit(subset()). The end result was a dataframe with 1733 objects and 13 variables. A screenshot of the data is available in Figure 21 of the Appendix.

2. Male player profile (~31000 rows × 13 columns, tabular data with spatial elements and texts, https://www.rank-tennis.com/en/profile/atp)
   The wrangling process is similar to female player profile. The end result was a dataframe with 2958 objects and 13 variables. A screenshot of the data is available in Figure 22 of the Appendix.
3. GDP per capita (xml data, contains spatial data and numbers, http://api.worldbank.org/v2/en/indicator/NY.GDP.PCAP.CD?downloadformat=xml)
   The xml data was converted to a dataframe in Python using the package xml.etree.cElementTree. Then store as a csv file with pandas library. The end result was a dataframe with 239 objects and 4 variables. A screenshot of the data is available in Figure 23 of the Appendix.
4. List of developed countries (Text data in PDF, https://www.un.org/en/development/desa/policy/wesp/wesp_current/2014wesp_country_classification.pdf)
   The list of developed countries was extracted from the PDF file using Python's camelot library and then reformatted and stored as a csv file using pandas. The end result was a dataframe with 36 objects and 1 variable. A screenshot of the data is available in Figure 24 of the Appendix.
5. Population (xml data contains spatial data and numbers, http://api.worldbank.org/v2/en/indicator/SP.POP.TOTL?downloadformat=xml)
   Wrangling process is similar to GDP. The end result was a dataframe with 264 objects and 4 variables as can be seen in Figure 25 of the Appendix.
6. Continents and countries (227 rows × 3 columns, a cleaned csv file with spatial information and text, https://www.kaggle.com/tomvebrcz/countriesandcontinents)
   The dataset can be readily used in R and a screenshot is in Figure 26 of the Appendix.
7. Female player ranking and age of year 2010(1147 rows × 6 columns, tabular data in html, https://www.rank-tennis.com/en/history/official)
   The dataset is copied to excel and saved as a csv file. In R, its Nan data was cleaned with na.omit(subset()). Also, the 'Age' column has a string format of '25y346d' and since we only care about the years for the age, the column was converted to an integer age via the function as.integer(substr(as.character(),1, 2), where the first two characters were extracted and converted to integers. The end result can be seen in Figure 27 of the Appendix.
8. Female player ranking and age of year 2020 (1333 rows × 6 columns)
9. Male player ranking and age of year 2010 (1795 rows × 6 columns)
10. Male player ranking and age of year 2020 (1936 rows × 6 columns)

Datasets 8, 9, 10 can be derived from the same link as dataset 7 and the processing procedure was also the same.

Part 2 Wrangling with exploration

Question 1.

For question one, we wish to check the player distribution among developed countries and developing countries as compared to population distribution. So firstly, a list of characters of developed countries are created from dataset 4 with as.list() in R. Then based on this list, a function with if loop to check whether a country name column is a developed country was defined in R. This function is then used to create a new column 'country_type' for male player profile, female player profile and population. The code is

available in Figure 28 of the Appendix. There are some mismatches for the country names which lead to Na values. The Na values' index were retrieved and the names were manually edited in Excel. The female data and male data are combined with rbind() to form a table for all players. Now that all countries are categorised as either developed or developing, the percentages of players or population in developed countries and developing countries are calculated based on the group_by and summarise function. Since, we want the player and population to be factors in a variable column, the dataframe of percentages were then melted with country type as the ID and we obtain a dataframe as shown below in Figure 1. From there, we can use this table for a stacked bar plot to compare the percentages.

	country_type	variable	value
1	developed	population	13.25758
2	developing	population	86.74242
3	developed	player	64.54914
4	developing	player	35.45086

Figure 1 Data of Melted Table

Next, using rgdal package in R, the the GDP was mapped with the world shape file for plotting spatial data with leaflet. The number of players within each country can be aggregated with group_by function and with tmaptool's geocode_OSM and a for loop, the latitude and longitude columns for the countries can be derived. Then the population dataframe was left joined on to the aggregated player table on country name such that with the new table, the tennis player ratio can be calculated by dividing the number of players with total population of each country. The final table was shown in Figure 2.

IOC	number_player	lng	lat	country_type	ratio
<chr>	<int>	<dbl>	<dbl>	<fct>	<dbl>
Algeria	5	3.00	28.0	developing	1.18
Argentina	149	-65.0	-35.0	developing	33.5
Armenia	1	44.7	40.8	developing	3.39

Figure 2 Ratio table after left join

The ratio table then leftjoined with GDP on country names such that a scatter plot can be drawn for ratio and GDP.

Question 2.

To distinguish female and male players in the player table, a sex factor column is created simply using a replicate( )function. A body ratio column was also added by dividing the weight with the height. The new table was then melted with the ID variables being CH(career high ranking) and sex such that we do separate plots for the height, weight and body ratio against the CH. The dataframe was shown in Figure 3.

	CH	sex	variable	value
1	1	Female	Height	166
2	1	Female	Height	186
3	1	Female	Height	180
4	1	Female	Height	168
5	3	Female	Height	174

Figure 3 Body shape table after melting

Question 3

Here, eastern players are defined as players from Asia and western players are defined as players from Europe, Oceania and North America. So by left joining the player profile with the continent dataframe on country names, the continent information for each player can be obtained. Then, all players from the relevant continents are subsetted from the joined table. Using a nested for loop and if loop we could add a background column based on the continents to indicate whether they are eastern or western. Next, difftime is used to get the age at which they achieve their career high ranking. After grouping the table by career high age and background, the number of players for each group was obtained via summarise() function. With a nested for-if loop, the percentages of players who achieved their career high ranking at different ages are calculated for each background respectively. Then we have a table that looks like Figure 4(player_number represents percentage here) and after making the percentages of eastern players negative, we can use that to plot a population pyramid. We can also repeat the same steps to get proportions of players of different backgrounds at different career high ranking and plot another population pyramid. The table was shown in Figure 5.

CH_age	Background	player_number
<fct>	<fct>	<dbl>
14	West	0.0607
15	East	-0.331
15	West	0.425
16	East	-1.82

Figure 4 Table for different background and CH age

CH	Background	player_number
<fct>	<fct>	<dbl>
1	East	-0.164
1	West	0.453
2	West	0.151

Figure 5 Table for different background and CH

Question 4

In this part, to discover the difference in player career length between 2010 and 2020, we use the age of all active players as an indicator of player's career length. However, it was necessary to categorize theses players based on sex. So we label dataset 7, 8, 9, and 10 with a column for 'Year and sex' and use rbind() function to combine these tables. Then they were further categorised with a new factor column called 'Age_group" with a nested for-if loop. Furthermore, using similar steps as the previous questions, the percentages of players at different age groups are put into a column via group_by(), summarise() and a nested for-if loop as shown in Figure 6. The table was then used for plotting a bar chart. In addition, to understand the average rank of players of different ages, a new table shown in Figure 7 was obtained with similar steps.

	YearSex	Age_group	pop	Percentage
	<fct>	<fct>	<int>	<dbl>
1	2010 Female	<20	468	40.8
2	2010 Female	>=35	5	0.436
3	2010 Female	20~24	446	38.9
4	2010 Female	25~29	190	16.6

Figure 6 Table for age and player distribution

Age	YearSex	average_rank	Sex
<int>	<fct>	<dbl>	<fct>
14	2010 Female	924.	Female
14	2020 Female	979	Female
15	2010 Female	872.	Female

Figure 7 Table for average rank at different ages

Data Checking

All data was checked and the main errors and outliers exist for weight and height of the player profile as well as the tennis player ratio for different countries. There are a lot of 0 values in the height and weight columns as theses data might be unavailable for some players. These rows are removed by a subset() function when discussing question 2. After that, the boxplots are shown below in Figure 8:

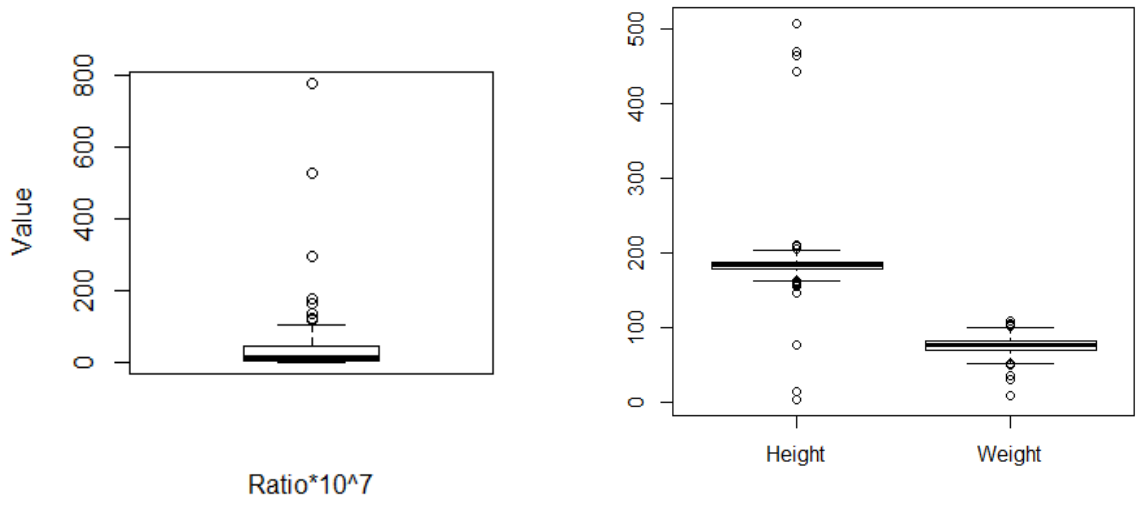

Figure 8 Boxplots for tennis player ratio, height and weight

In Figure 8, we can see there were a few outliers for tennis player ratio and these values were removed using code:

```
outliers = boxplot(player_count$ratio, plot=FALSE)$out
tempdata<-player_count
tempdata<- subset(tempdata, !tempdata$ratio %in% outliers)
```

Figure 9 Code for removing outliers

The reason why these outliers were removed was that they might disturb the real trend show on the map. Extremely high ratios could be a result of small population and it was not accurate enough in representing the trend. For height and weight, it could be observed some outliers are significant in the real world as it was reasonable to have a player around 210 cm or 100 kg. It was not a good decision to remove them. Hence, an if loop was used to only remove heights that are outside 150~210 cm and weights that are below 40kg.

Data Exploration

A comparison was first made between the composition of tennis player and world population in terms of developed countries and developing countries. A stacked bar chart was used to make the comparison and as shown in Figure 10, there was a huge disparity between their compositions. Only 13% of the world population are from the developed countries as compared to the 64% for tennis players. So in this case, it was true that tennis is a sport for the richman as people in the developed countries are generally more wealthy and it also shows to a great extent that the popularity of tennis in developed countries is higher.

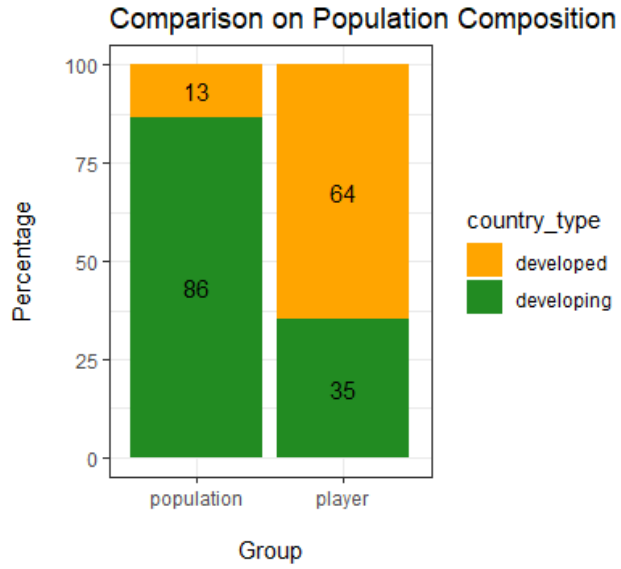

Figure 10 Comparison on population composition

To have a better understanding on the relationship between wealthiness and popularity of tennis, a chloropleth leatlet map of GDP per capita was drawn as shown in Figure 11, the larger circles represents higher ratio of active tennis players in the corresponding country. So the map seems to unveil the trend that countries with higher GDP per capita tend to have larger circles, which suggests higher popularity of tennis. However, tennis was born in the western countries and this has undoubtedly influenced its popularity in these areas, which happens to be the regions with higher GDP. After zooming in the map to get a closer insight on the situation in Asia, the trend observed previously was proven to be correct. The zoom-in version can be seen in Figure 12.

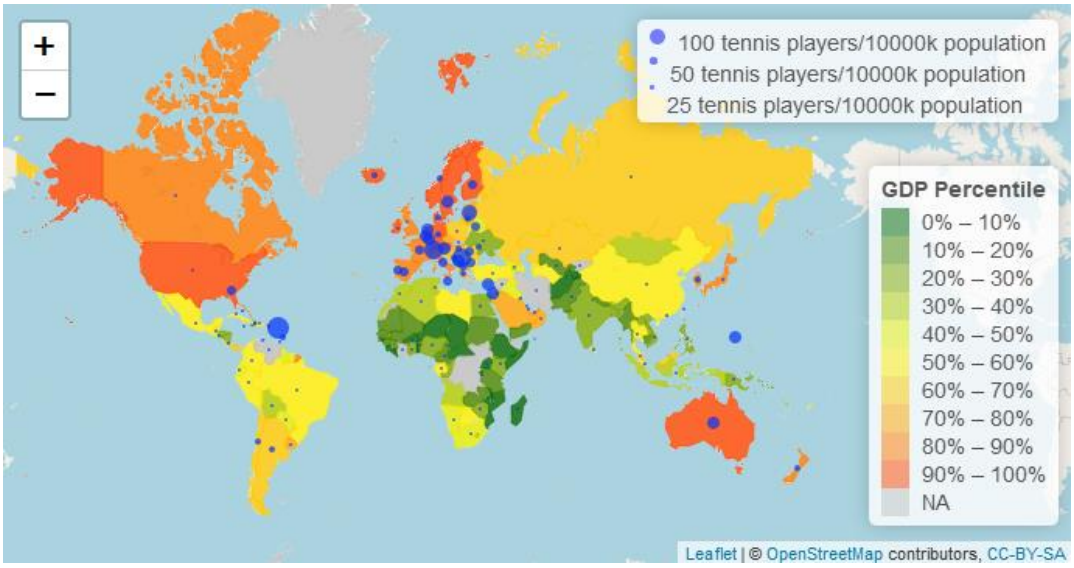

Figure 11 GDP and tennis popularity

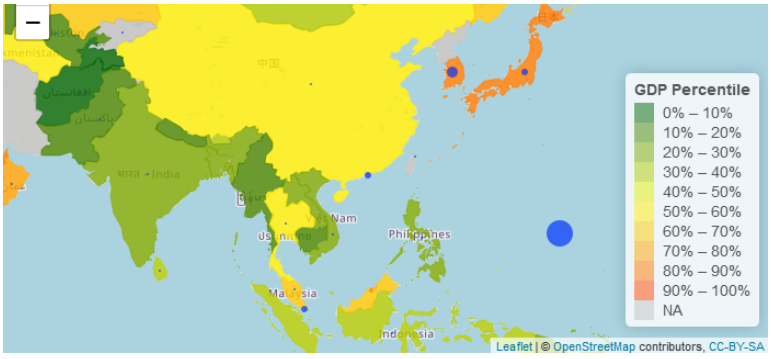

Figure 12 Asia-focus version of Figure 11

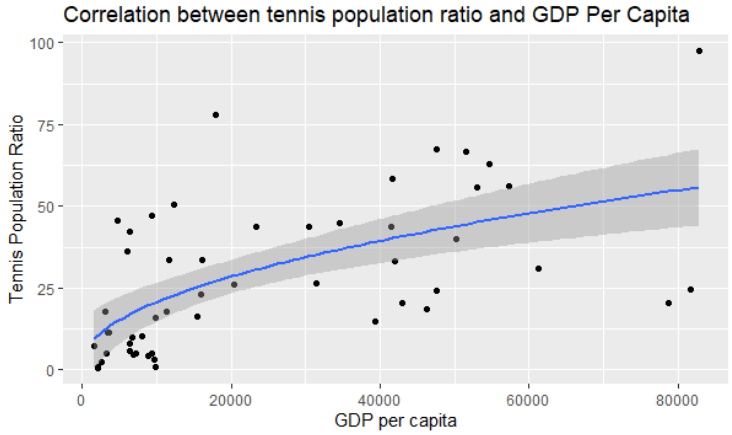

Figure 13 Correlation between tennis player ratio and GDP

To further investigate the correlation between GDP per capita and ratio of tennis players, a scattered plot was drawn with ggplot() and 'lm' smoother is used with formula = y~sqrt(x). From Figure 13, We can obviously see a positive correlation between the ratio of tennis player and GDP per capita. Hence, countries with higher GDP are more likely to have higher tennis player ratio and higher tennis popularity. This father proves the influence of richness on tennis popularity.

Regarding the ideal body shape for a tennis player, this report seeks to investigate it based on height, weight and body ratio (weight/height) . The data are categorised into female and male groups as the ideal body shape should differ for male and female. To assess the successfulness of a tennis player, the career high ranking was used as a valid indicator of their capability and successfulness. A faceted scatter plot was drawn as shown in Figure 14. As there are too many points, we used a smoother line of 'gam' and a formula of y~poly(x, 2) to observe the trend. So it can be observed that for female players that their career high rank decreases as height increases or weight increases. Whereas, the rank increases with higher body ratio. It seems that height plays the most significant role in determining a female player's capability as it has the steepest slope, followed by weight and body ratio. Thus, for female players, it is desirable to have higher height. Weight and body ratio do influence a bit, but not very significant. For male players, we can see the trend was more obvious, which means that body shape plays a greater role in determining their capability. The rank decreases with higher height, but higher weight and higher body_ratio are not always desirable. If we took the minimum point of the smooth line, the most ideal weight would be around 90kg and the most ideal body shape would be around 0.46. Below these 2 values, the heavier the better, while above these two values, the slimmer the better.

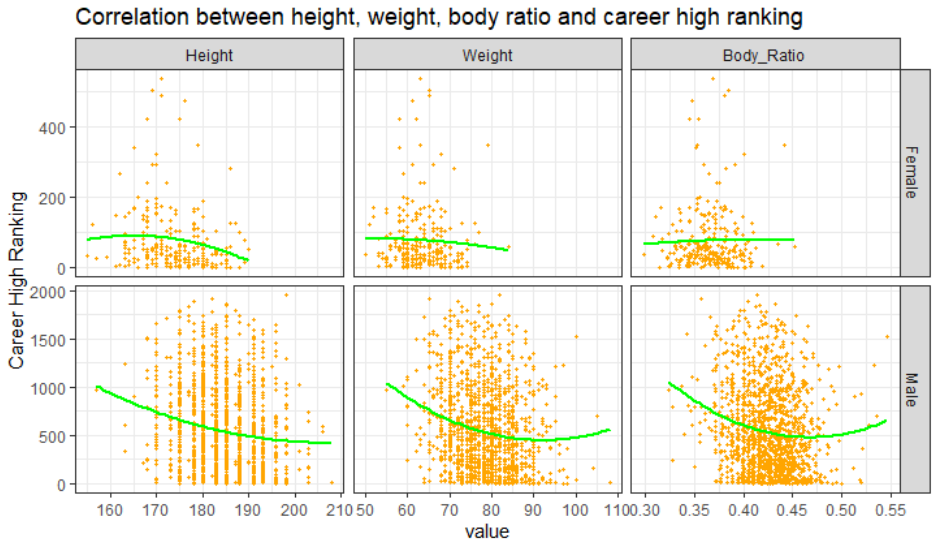

Figure 14 Scatter plot for correlation between height, weight, body ratio and career high ranking

For instigating the differences in the career development of eastern players and western players, again, the career high rank was taken as a key milestone in their professional tennis career. This report investigates on when they achieved the biggest success in their career and whether the western players have outperformed the eastern players.

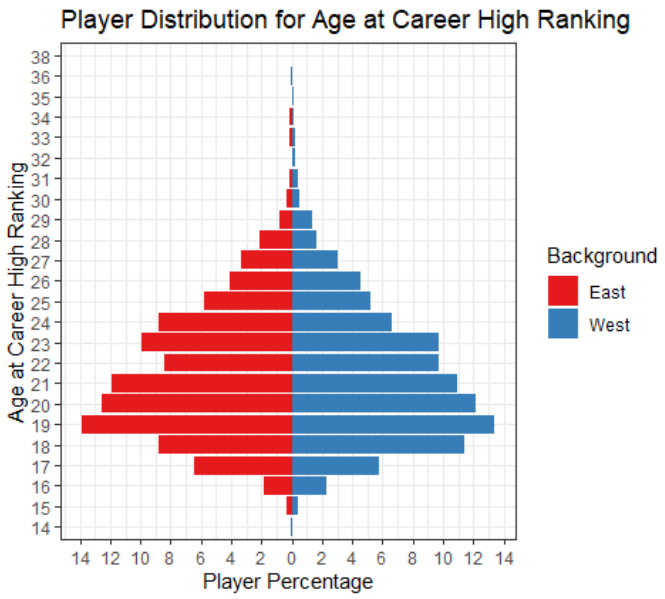

Figure 15 Population pyramid for eastern and western players based on age at career high rank

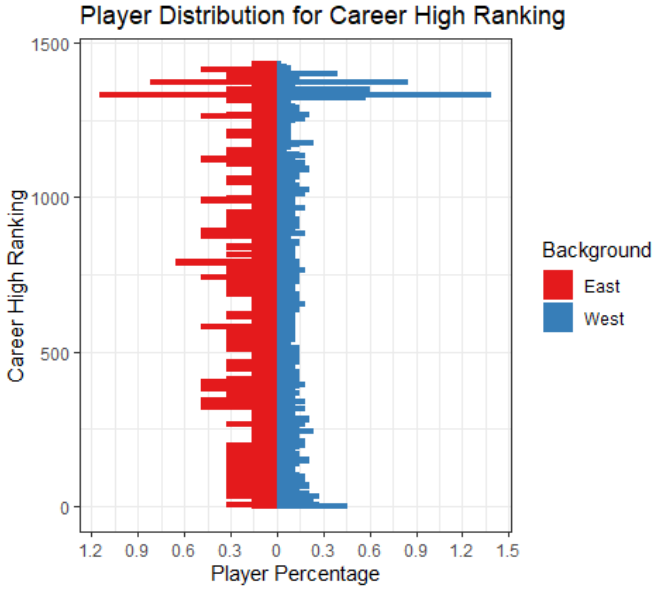

Figure 16 Population pyramid for player distribution on career high rank

Figure 15 illustrated the percentage of players achieved their career high rank at different ages. A population pyramid was used to show the differences in the pattern. It can be observed that eastern players generally mature later than western players. Western has a significantly higher percentage for the age of 18(11% as compared to 9 % for eastern players), whereas eastern players have higher percentages for ages >= 24. Moreover, Figure 16 shows the percentage of players that achieved different career high ranks. It was hard to judge whether the western players have outperformed eastern players because although there are more western players who achieved their career high rank in the top section, eastern players have a more even distribution and lesser proportion of players at the bottom section as well. Thus, overall, eastern players mature later and western players are more polarised in their career achievement.

Players nowadays play more tournaments and earn more prize money. This report also looks into the age evolution of the tennis players since 2010. A comparison on player distribution among different ages was done and the players are grouped based on sex as female players tend to retire early due to family issues. A boxplot was first drawn to instigate the age data, yet the outliers are relevant cases in this discussion and are not removed. It can be seen in Figure 17 that both male and female players have transformed into an 'aging population'. The average age of players has increased since 2010, especially for the female players. To gain a better understanding, a multi-set bar plot was used to break the distribution into different age groups. As shown in Figure 18, we could see that the percentage of players below 20 has decreased significantly for both male and female(especially female). There was not much

change for the group of 20~24, though the male group did increase quite a bit. For the age group of 25~34, it did not vary much for the male but the female proportion of age around 25 to 34 had increased much more significantly. Furthermore, both female and male players have doubled proportions of players aged above 35 now.

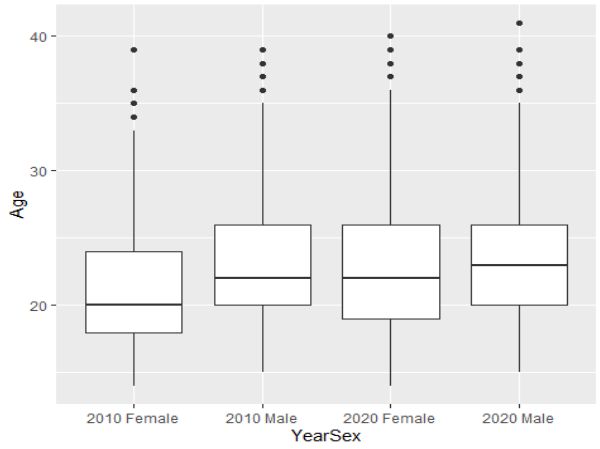

Figure 17 Boxplot for player age

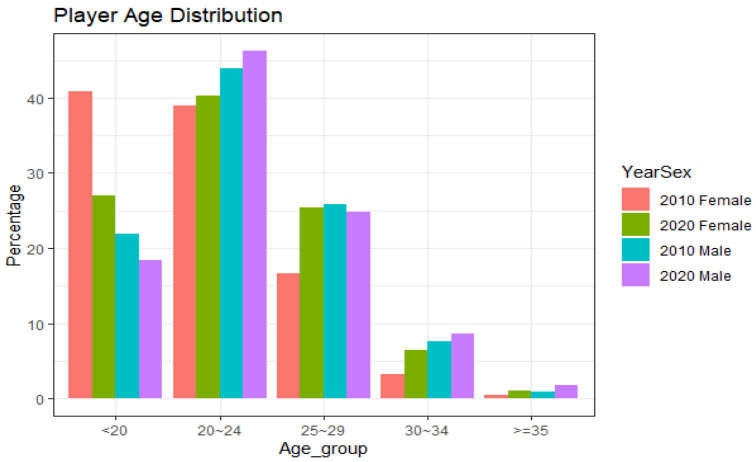

Figure 18 Player Age Distrbution

Although the proportion of aged tennis players has increased over the years, but the question of whether they are still performing at a competitive level remains. Thus, by using 'loess' smoother on the scattered plot, the average rank of players of different ages are compared as shown in Figure 19. It was noticeable that female players aged above 35 are performing much better now as compared to 2010 and the same trend can be seen for male players above 30 years old. It can be doubted whether these 2 numbers should have been higher because the number of players also increased over the years and the longer ranking list might have pulled up the average ranking. Similarly, if we plot the average age of player at different ranking as shown in Figure 20(the graph was purposely unfaceted to show the comparison among male and female players), we could observe that the players' rank generally increases with their age and female players have aged much more than male players. Moreover, the age gap between female and ,male players has shortened. This could attribute to the decreased income gap between female and male players (Staff, 2020). Therefore, if the age of players was used to infer their career length, the players' career life was indeed lengthened and at the same time, aged players still remained competitive in professional tennis. The trend was more obvious among female players.

Figure 19 'loess' smoother line for rank against age

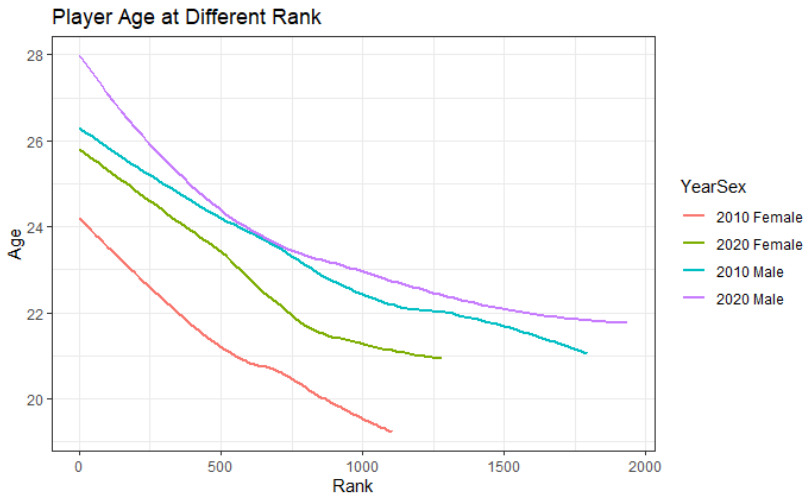

Figure 20 'loess' smoother line for age against rank

Conclusion

In this report, by using ratio of active professional tennis players, it was discovered that tennis was indeed more popular in developed countries and GDP per capita was positively correlated to the ratio of active tennis players. This answered the first question I proposed. Then the ideal body shape was assessed with career high rank as a measure of capability, the trend discovered was that for female players, the taller the better . Weight and body ratio did not affect much. For male players, height was also positively correlated to capability whereas this was not true for weight and body ratio. The ideal shape was at 90 kg and 0.46. This answered the second question. As for the third question, it was discovered that eastern players mature later and western players are more polarised in their career achievement. Lastly, the 4th question was discussed base on the distribution of players among age groups. The career life of tennis players has lengthened since the average age of active players increased and the aged players were still performing well.

Reflection

This project taught me how to use R studio and Python to do data wrangling, cleaning and plotting graphs. I became more familiar with various R commands and new functions which I had to learn to process my data. I also gained a better understading on modern tennis development. Nontheless, due to time constraint and page limit, I could only dicuss each question with limited datasets and few perspectives. Real-world questions are always complicated and I could have introduced more variables to each question to dicuss them from different points of view. For example, I could only use career high rank as a measure of their capability whereas it could be measure with more things like career titles , win/loss ratio. Some players are good at doubles, but I only discussed singles' players here. For the last question, I could further investigate the correlation between prize money and career length as it seemed that players who earn more prize money tend to have a longer career life.

Bibliography

BaneMichael. (2015.6.21). Rich rewards for those at the top in tennis, but what of the rest? Date accessed: 2020.4.29， source: The Conversation: https://theconversation.com/rich-rewards-for-those-at-the-top-in-tennis-but-what-of-the-rest-35961

Genclein97. (2015.3.27). The Body of an Elite Tennis Player. Date accessed: 2020.04.29， source: Elite Tennis Players: https://elitetennisplayers.wordpress.com/2015/05/27/9/

McgroganED. (2009.09.19). Viewpoint: Players Need to Stop Complaining. Date accessed: 2020.04.29， source: Tennis.com: https://www.tennis.com/players/2009/11/viewpoint-players-need-to-stop-complaining/18216/

NingPan. (2018.1.23). Hyeon Chung: Rising Asian players boosting the Australian Open on and off the court. Date accessed: 2020.04.29 ， source: NEWS: https://www.abc.net.au/news/2018-01-23/asian-dynamics-boosting-the-australian-open-on-and-off-the-court/9349116

StaffWTA. (2020.3.11). Venus on equal prize money - 'I’m happy I was able to do my part'. Date accessed : 2020.4.29 ， source: WTA: https://www.wtatennis.com/news/1644365/venus-on-equal-prize-money-i-m-happy-i-was-able-to-do-my-part-

TangiralaPrasant. (n.d.). Race and Color in Tennis: Whither Diversity? Date accessed: 2020.04.29， source: Bleacher Report: https://bleacherreport.com/articles/181775-race-and-color-in-tennis-whither-diversity

The Origins of Tennis - History of Tennis. (n.d.). Date accessed: 2020.04.29，source: History of Tennis: https://www.tennistheme.com/tennishistory.html

Totalsportstek2. (2020.01.20). Highest Prize Money In Tennis Grand Slams (2020). Date accessed: 2020.04.29 ， source: TOTAL SPORTTEK: https://www.totalsportek.com/money/highest-prize-money-in-tennis-grand-slams/

Appendix

```
'data.frame':   1733 obs. of  13 variables:
 $ Name    : chr  "Ashleigh BARTY" "Karolina PLISKOVA" "Naomi OS
AKA" "Simona HALEP" ...
 $ IOC     : Factor w/ 88 levels "Algeria","Argentina",..: 4 21
41 69 14 84 21 79 55 86 ...
 $ DOB     : Date, format: "1996-04-24" "1992-03-21" ...
 $ Height  : int  166 186 180 168 170 174 182 175 182 175 ...
 $ Weight  : int  62 72 69 60 0 60 70 63 74 70 ...
 $ Turn.Pro: int  0 2009 0 2006 2017 2010 2006 0 2009 1995 ...
 $ Prz     : int  16515667 19512517 14417479 35108020 6720038 19
234960 31066637 7846208 10730150 92543816 ...
 $ Rk      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ CH      : int  1 1 1 1 4 3 2 7 4 1 ...
 $ Date    : Date, format: "2019-06-24" "2017-07-17" ...
 $ Titles  : int  6 15 5 19 3 13 27 4 9 72 ...
 $ W       : int  240 527 217 493 137 368 527 266 426 826 ...
 $ L       : int  90 289 130 214 50 191 233 144 251 142 ...
```

Figure 21 Female Player Dataframe

```
'data.frame':   2958 obs. of  13 variables:
 $ Name    : chr  "Rafael NADAL" "Novak DJOKOVIC'
" "Daniil MEDVEDEV" ...
 $ IOC     : Factor w/ 114 levels "Algeria","Anti
",..: 96 90 99 88 5 42 40 54 96 37 ...
 $ DOB     : Date, format: "1986-06-03" ...
 $ Height  : int  185 188 185 198 185 193 198 196
 $ Weight  : int  85 77 85 83 79 85 90 95 75 85 .
 $ Turn.Pro: int  2001 2003 1998 2014 2011 2016 2
004 ...
 $ Prz     : int  115466561 136954944 127504891 9
 7209605 18900563 3580862 11725808 17615816 ...
 $ Rk      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ CH      : int  1 1 1 4 4 5 3 8 9 6 ...
 $ Date    : Date, format: "2008-08-18" ...
 $ Titles  : int  84 77 103 7 16 3 11 3 9 8 ...
 $ W       : int  970 889 1235 131 272 100 221 61
 $ L       : int  196 185 268 73 146 61 112 41 17
```

Figure 22 Male Player Dataframe

```
'data.frame':   239 obs. of  4 variables:
 $ Country: Factor w/ 239 levels "Afghanistan",
 9 10 4 7 ...
 $ Item   : Factor w/ 1 level "GDP per capita (
1 1 1 1 1 1 1 1 ...
 $ Year   : int  2018 2018 2018 2018 2018 2018
18 ...
 $ Value  : num  521 3432 5269 42030 6609 ...
```

Figure 23 GDP

```
'data.frame':   36 obs. of  1 variable:
 $ Developed.economies: Factor w/ 36 le
",..: 2 3 9 11 12 13 14 17 18 22 ...
```

Figure 24 Developed countries

```
'data.frame':   264 obs. of  4 variables:
 $ Country: Factor w/ 264 levels "Afghanistan
250 9 10 4 ...
 $ Item   : Factor w/ 1 level "Population, to
1 1 1 ...
 $ Year   : int  2018 2018 2018 2018 2018 201
18 ...
 $ Value  : num  105845 37172386 30809762 286
```

Figure 25 Population

```
'data.frame':   227 obs. of  3 variables:
 $ Country   : Factor w/ 227 levels "Afghar
 34 36 38 40 41 45 ...
 $ Region    : Factor w/ 7 levels "AFRICA"
1 1 1 1 1 ...
 $ Population: int  12127071 7862944 16398
340702 420979 4303356 9944201 690948 ...
```

Figure 26 Continents

```
'data.frame':   1146 obs. of  6 variables:
 $ Rank : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Point: int  8645 6375 6230 5930 5626 5100 4930
 ...
 $ Name : Factor w/ 1146 levels "Abigail Spears",
 1002 1073 304 1085 451 5 909 ...
 $ IOC  : Factor w/ 77 levels "Albania","Argentin
 58 74 58 6 59 55 4 ...
 $ Trn  : int  17 22 0 17 17 0 15 21 20 23 ...
 $ Age  : int  28 19 23 24 29 28 20 25 21 26 ...
```

Figure 27 Female ranking 2010

```
type_function = function(x) {
  if (as.character(x) %in% try) {
    return("developed")
  } else {
    return("developing")
  }
}
wta$country_type = as.factor(apply(wta,1,function(x) type_function(x['IOC'])))
head(wta)
atp$country_type = as.factor(apply(atp,1,function(x) type_function(x['IOC'])))
```

Figure 28 codes for deriving country type