Insights into sports – why you're not an elite athlete

Introduction

Sports are ubiquitous. They take a sizable part in peoples' leisure time around the world. On one hand, sports are regarded as entertainment. People attend competitions as fans of individual athletes and teams. They follow the news related to sports, and most importantly people love to talk about sports. Hereby, the behaviour can simply be described as social belonging. You connect with people that follow the same sport and connect even closer with someone who admires the very same athlete. On the other hand, people practice sports in their leisure time. The desire to do so arises from the need to exercise physically and to maintain social contacts. Moreover, it is about competing with your friends and showing off your progress. For professional athletes this notion is intensified. For them it is about breaking records, receiving awards, earning money and becoming famous. One can argue that professional athletes have diverged from the average man in a sense that each sport became so specialized that an average person might fall behind. Nonetheless, the connection between these parts is electrified, meaning people are highly motivated to close the gap[1].

In this project, the emphasis lies on the following questions: What takes an athlete to be successful in his sport? Undeniably, experience, talent, and luck all play into success. What about height and weight? How does their impact on success differ among sports? In this project, ten different sports are looked at, ranging from team to individual sport, differing in their setting and physical requirements.

Definition of Knowledge Graph

In today's complex digital world, there has never been a greater need to achieve structure out of your content in order to organize information and establish links between diverse data types to solve real business challenges or answer research questions. When looking at data management of the past century, one sees that data was organized in silos. Therefore, it was a highly time-consuming task for every knowledge worker to identify the right dots and information pieces, to connect them, to make sense of them, and finally to communicate and interpret them the right way. On the other hand, knowledge graphs are intuitive. They express all information as relationships. All metadata and items are expressed as identifiers, so they represent real world "things" rather than just a string label. Edges build the connection between those objects. Further, sophisticated data mining methods have been developed which are employed on knowledge graphs for a variety of tasks: *Link prediction* tries to find missing edges in knowledge graphs, whereas *link correction* is about finding incorrect edges. *Entity resolution* is about mapping entities in text to knowledge graphs and *clustering* groups entities based on their similarity[2]. Many knowledge graphs are very large, and their creation is crowdsourced and/or they are generated from various sources.

One of the most well-known knowledge graphs is Wikidata. It spans across multiple domains, and touches millions of data points. It is openly curated and allows for imports of massive amount of data. Its interface for SPARQL-queries is a supportive tool in extracting the data. Wikidata introduces the concept of statements. A statement enables to store the information about an item where each item is assigned its own page. Statements can also be further described with optional qualifiers, references, and ranks. Within a SPARQL-query, statements can be used to merge items. To illustrate the scale a query can take, we look at an award won by an athlete. The statement 'awards received' entails a list a of awards. Each reward has a qualifier pointing to the 'point of time' that award was received. Then one can filter for an award, let's say 'Ballon d'or' to receive all athletes than received this award with is according date. This was one example of the multiple queries this project entails which in turn exposes the possibilities of a knowledge graph like Wikidata.

---

[1] http://freakonomics.com/podcast/sports-ep-3/

[2] Managing and Mining Graph Data, Charu C. Aggarwal & Haixun Wang (2010)

Table of content

SPARQL-Queries

The following queries not only entail a structure whose complexity was reduced to the minimum in order to receive a faster run time. For instance, 'Bind' exposes a costly computation on a SPARQL. Consequently, 'Binds' were avoided. More importantly, they entail subjective judgements about the underlying data to lower the biases and eventually extract useful knowledge. The discussed examples will motivate those judgements.

A general exclusion of athletes is made by including only athletes that were born after 1970. Before that time, sports weren't as competitive as they are now. This shall lead to more contemporary and significant results. The following aggregate functions 'sample', 'avg', 'sum', and 'group by' were extensively used as well as the predicates 'p', 'pq', and 'ps'. All queries were performed in March 2019.

Skill-level

To measure the impact of height and weight in each sport, a score that represents the skill of an athlete within the sport must be assessed. To define a skill, a criterion must be chosen. The choice of a criterion is influenced by the profession of the sport. In what setting do athletes compete? Hereby, the rules and settings of competitions are critical. Is the sport organized in teams and are there important tournaments? Is it meaningful to have participated at the Olympic games many times? Is a high ranking or a high prize money meaningful?

Number of goals

In football, the number of goals displays an indicator for success. A single football player who scores many goals supports his team to win matches. Teams are organized in leagues which functions as a classifier. A footballer who scores many goals in the third division is not comparable with one footballer that scores the same amount of goals in the first division. A quantitative difference between leagues cannot be easily interpreted. Therefore, footballers from 'LaLiga' and the 'Premier League' are considered which are the best two leagues in the world according to UEFA[3]. Query 1 shows the total goals scored divided by the number of matches played. This ratio is simply an indicator for goals per matches scored which represents the effectiveness of a player. The pool of players is decreased by taking only players who scored more than 200 goals to exclude those players that had a short career and therefore less impact on their team.

Average-ranking

For both tennis and table-tennis, candidates for representing the skill-level of a player were number of victories, number of won matches, and ranking points. Both sports are rather individual sports and organized in tournaments which take place across the globe. The ranking points turned out to be the most represented in Wikidata. They have the property of being a concise number. Over the years, players receive numerous rankings, representing their ranking at a particular point in time. In Tennis, a ranking is being composed by the 'Association of Tennis Professionals - ATP' since 1973. The ranking points are assigned according to the stage of tournament reached, and the prestige of the tournament. The four so-called 'Grand Slams' are the most prestigious tournaments. The ranking both in tennis and table-tennis are dynamic, meaning players switch their ranking quite frequently. This is amplified by the ATP in tennis that ranking points are dropped 52 weeks after being assigned. Consequently, for the skill level the average is taken by means of the aggregate function 'AVG'. Query 8 includes a filter to include only players who received at least two rankings. Otherwise, the results

[3] https://www.uefa.com/memberassociations/uefarankings/country/#/yr/2019

would be highly skewed because of players who have only one ranking in Wikidata. A filter in line 22 restricts the ranking on single and double matches. Otherwise there would be rankings numbers included from youth tournaments which are less competitive and therefore not representative.

Number of awards and number of participations

A rather straightforward measurement of skill is made by taking the number of awards. The award can be further specified to be a gold medal as in the case for swimmers, judokas, and cyclist. For these sports, the Olympic games are the most prestigious tournaments. Here, they compete against the very best from all over the world. Having received multiple gold medals over the years indicates a constant delivery of top performances. This eventually factors in the notion of having luck in sports as winning six gold medals in 6 in six consecutive years is less likely to be random but more likely to indicate a high skill.

"Peak-of-performance" age

The age when an athlete reaches its peak performance is of particular interest. It has the potential to uncover the effect of experience on the success of an athlete in a sport. As the effect of age on success might be bull-shaped for some sports, meaning that your "peak-of-performance" age lies in the middle of an athlete's career displaying the trade-off between fitness and experience. In other sports, experience plays a major (minor) and fitness a minor (major) role, meaning the relationship between age and success is a straight line with a positive (negative) slope. A second reason for a high "peak-of-performance" age other than investment in experience is a high monetary investment. This might be the case for jockeys where equipment and its maintenance are costly.

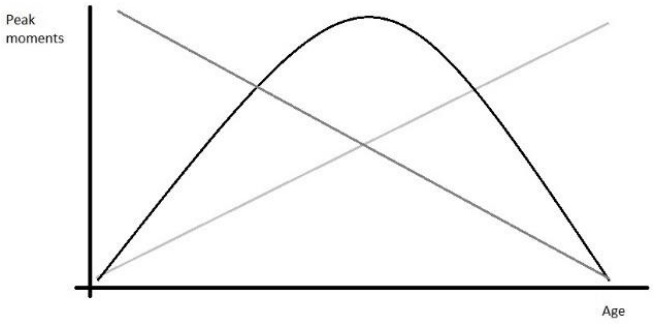


Figure 1, age vs peak moment

In order to estimate the "peak-of-performance" age of each targeted sport, the year needs to be queried in which an athlete had a peak-moment in his / her career. This can be an award, a medal or a ranking. Since an athlete might receive many gold medals, he/she reaches several peaks in his/her career. This supports the idea of separating the queries for the "peak-of-performance" age from the previously discussed queries.

Awards

Important awards are the 'Ballon d'or' for footballers, 'Player of the year' for baseballers and gold medals for cyclist, jockeys, judokas and swimmers. In sports where athletes do not aim for one award but many, a different approach must be followed. This is the case for basketball and golfers. Here, rather the year is considered in which an athlete received multiple awards. This approach requires a complex query structure. Query 23 has a sub-query in which golfers are first grouped by individual golfers and then by the year in which they won an award. This allows the sub-query to sum up the number of awards a golf received per year. In line 21, the year is restricted by considering only those years in which a golfer received at least 2 awards. In the super-query, the age at which a golfer received at least two awards is average across all golfers and grouped by gender. For golfers the average age for both female and male was around 27.

Ranking

The time when tennis players were ranked as number one should come closest in reflecting the peak moments in their careers. In Query 20 on line 17, the crucial restriction is made. The number is accessed by the predicate-prefix 'ps' in line 12 whereas the predicate-prefix 'p' accesses the QID of a statement in line 10. In line 13, a further restriction is made by accessing the 'sport' of the ranking and restricting in to 'tennis singles' only. Other 'sports' are 'tennis doubles' or 'junior tennis' which might not reflect the year of a peak moment in an athlete's career.

Height and Weight

Height and weight are both to some degree inherited trough genetics[4]. Being too small or too large places a disadvantage in many sports. For an individual athlete, this might hinder him / her from reaching the top. If you think of weight in terms of muscle mass, the advantage lies for those athletes who can more easily build muscles and less fat. As muscles weigh 12% more than fat[5], it is more likely that in professional sport a higher weight indicates higher muscle mass. If someone is not taller but shows a much higher weight, those additional muscles might place an advantage. To account for this, a common measurement is the ratio of 'Body mass index'.

$$BMI = Weight / Height$$

Gender

For all athletes, a distinction was made between male and female athletes. Not only their body differs on average, but sports also have a different setting for each gender. This is mainly because that men and women do not compete directly against each other. This is partly reflected in the different physical athletes bring into competitions.

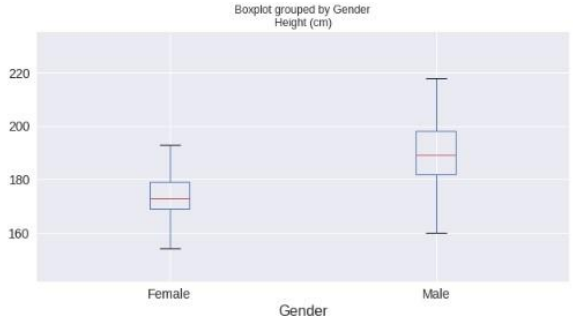


Figure 2, Height across genders

Nationality

As some sports are dominated by certain nationalities due to cultural reasons, the height and weight effect might flatten out as people in general in the countries of those nationalities have a similar average height and weight.

---

[4] Genetic and Environmental Effects on Weight, Height, and BMI Under 18 Years in a Chinese Population-Based Twin Sample, Qingqing Liu,mCanqing Yu, Wenjing Gao, Weihua Cao, Jun Lyu, Shengfeng Wang, Zengchang Pang, Liming Cong, Zhong Dong, Fan Wu, Hua Wang, Xianping Wu , Guohong Jiang, Binyou Wang and Liming Li, 2015

[5] https://www.cerascreen.de/blogs/news/muskeln-wiegen-mehr-als-fett

Countries

Another interesting aspect of sports lies on a macro level. As sports have a positive influence on well-being, I wonder whether that this is reflected in data on athletes per country and their respective human development index and life expectancy. Hereby, the causality goes in both directions meaning that a high human development index of a given country can also lead to a high number of athletes in that country. Query 27 gather all the relevant data in one handy query.

Data cleaning

Concatenation

Before all data is corrected and normalized, the tables had to simply be put together.

Exploration

After exploring the queried data, some corrections needed to be taken. For instance, both height and weight were respectively represented by different metrics. Inch had to be transformed in to cm and pound into kilogram.
Score-values that were derived from ranking points had to be reversed since a high skill-score means having a low ranking. The values were simply reversed:

$$\frac{1}{ranking\ point}$$

By means of the python library 'matplotlib', the data could easily be explored.

(https://colab.research.google.com/drive/1KLhuLi2CYxM_QdhRQAhdNii3jV1Q6Bj7)

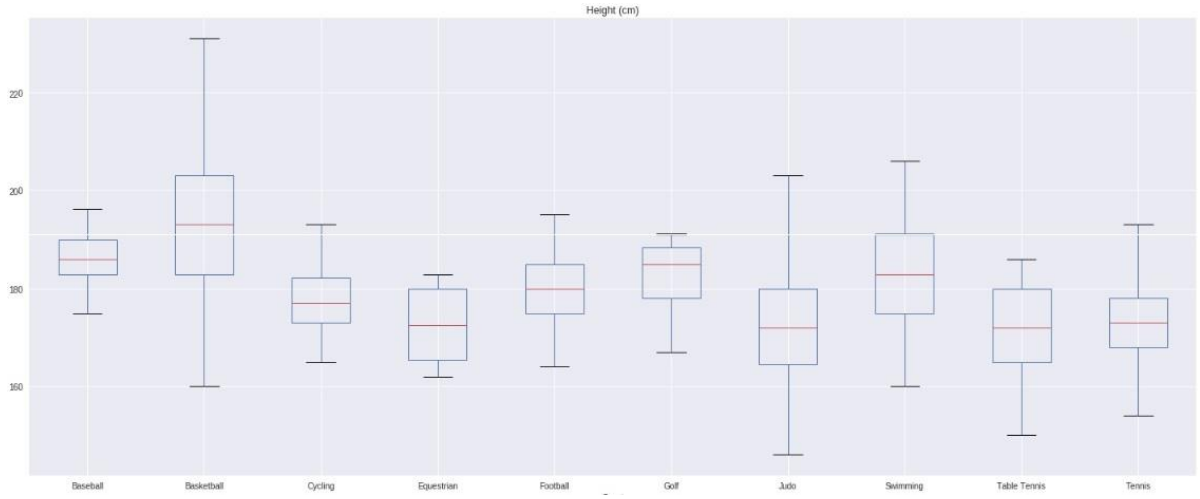


Figure 3, Height among sports (boxplots)

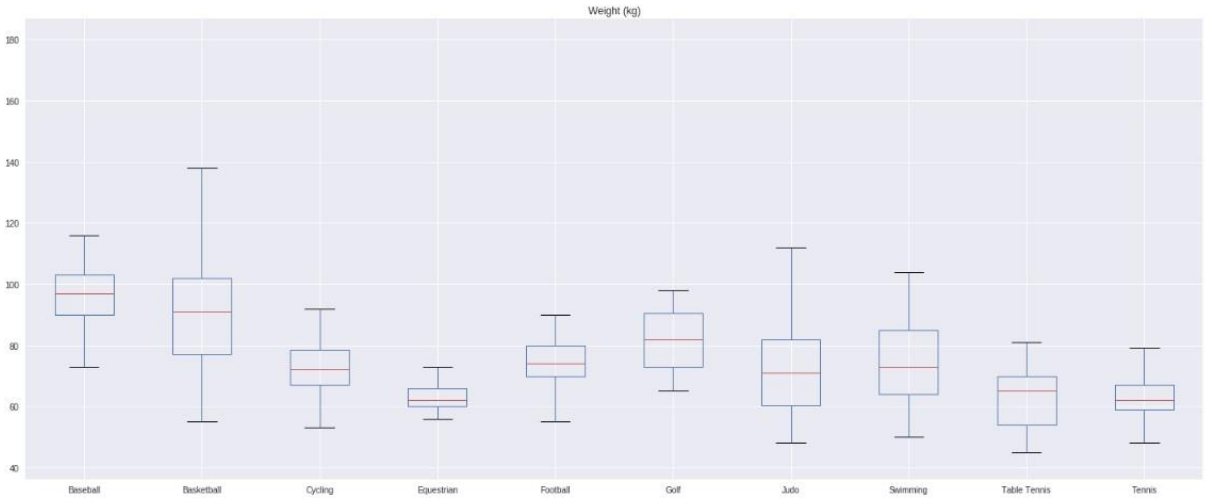


Figure 4, , Weight among sports (boxplots)

Normalization

This step concerns the values for the skill-scores. The intent is to make these values comparable among sports. The average value, μ, and the standard deviation, σ, of a subset containing all skill-values per sport is determined. Both are used to normalize each value, x, simply by using the z-score formula:

$$z = \frac{x - \mu}{\sigma}$$

The subset account for both male and female athletes expect for football and baseball where the subsets were further subdivided by their genders because different measurements were applied to obtain the skill-scores. Thus, the measure for the skill-score was not harmonized and the normalization had to be separated.

The following table shows the variation of the samples of skills scores within sports via boxplots.

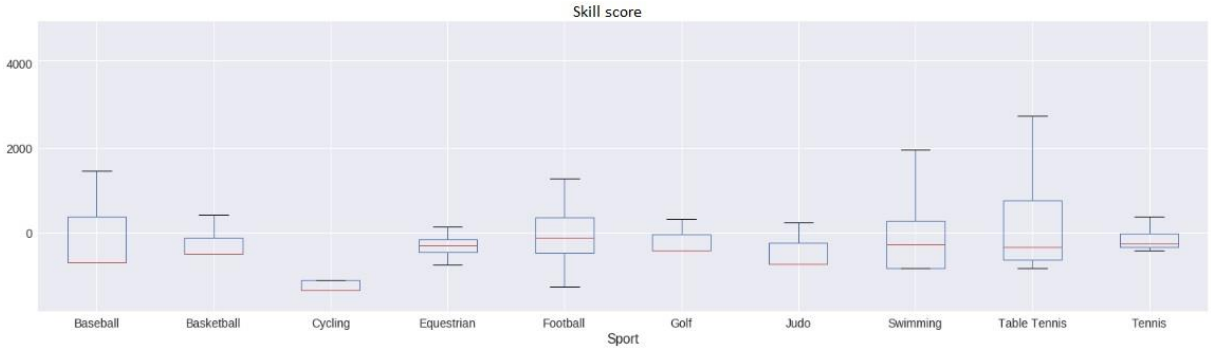


Figure 5, Skill-score among sports (boxplots)

In table-tennis a higher variation of skill-scores can be observed. This can be interpreted as having a small number of 'elite' players over the years in table-tennis. In comparison, tennis seems more competitive, as the skill-values show less variation among players indicating ups and downs of player along their career.

The table below shows the 'peak-of-performance' age which might account for the experience needed to become good in a sport. Hereby, swimmers succeed the youngest and jockeys the oldest. In

general, running-intensive sports require a certain fitness whereas other like golf and baseball require experience.

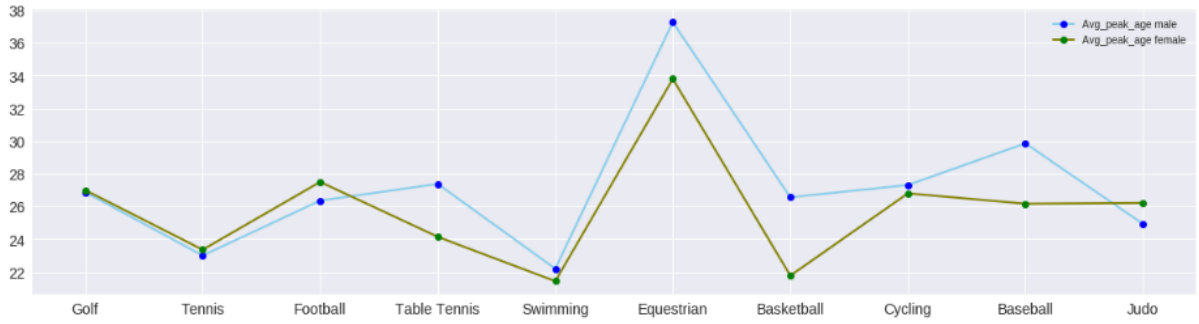


Figure 6, Peak of performance age among sports per gender

Countries - side questions

By counting the number of players per sport one obtains an approximate of the most popular sports in the world (query 24). By querying athletes of all sports available on Wikidata and grouping them by sport, and nationality (query 25), one receives the number of athletes per sport and per country. By asking who has the most athletes per inhabitants, it might point to the sportiest nations. For a population higher than 5 million, Norway, Finland, and Sweden rank the highest. These countries indeed show a high score on HDI and a high age of life expectancy of around 83.
To get insights into the most popular sports by country, query 25 is extended by 'group by ?country' . The number of players were than ranked. The top 10 sports by country were then collected and imported into Wikidata.

Import data into Wikidata

Mass import

From Wikidata I gathered data on the most popular sports of countries which were assessed by means of query 26. For instance, the category 'Sport in Spain' (Q3181499) then receives a new statement 'Sport' (P641). Inside the statement the ten most popular sports were added. The values were for instances 'Association Football' (Q2736) with the qualifier 'rankings' (P1352) and its value '1'. If you compare those rankings with the Wikipedia article 'Sport in Spain' it serves as a good representation. To import those values, I used the web application "quickstatements". To prepare the statements I used an add-on for "Google Spreadsheets" to automatically generate the URI for the countries. The procedure for the most popular sports was the following. I added a column for the predicate 'Country', a column for the predicate 'Sport', another column for a qualifier to indicate that this sport has an ordinal ranking. In the last column I added the column for the ranking.

The ambition for this was to build a link between the different of the average height and weight with the average values for each sport and the most popular sports. Unfortunately, my mass import for average height and weight was not accepted by the crowd of Wikidata.

Manual edits

For specific cases as the following I had to add triples manually. For instance, the ranking points of an important player in table-tennis history, 'Jan-Ove Waldner' (Q345360), were missing. The triples entail the statement 'ranking' (P1352) with the values for the ranking-numbers and their qualifiers 'start time' (P580).

In the data exploration phase, I found many outliers that simply entailed wrong data points. For instance, in querying the number of goals per matches in assessing the skill-scores for male footballers I came across players that had a suspiciously high number of goals. Hereby, knowledge about the domain one is analysing is very indispensable. Fortunately, Wikidata is crowdsourced and therefore simple edits are possible.

Results

Skill

By looking at the 'best of the best', an impact of height and weight within a sport is visible. A general tendency is that height and weight are beneficial for the most sports except for table tennis and equestrian sport.

This table does not support an effect of height or weight on skill across the sports. Thus, one need to look the effect within a sport and how it differs among them. For instance, Judoka Teddy Riner is much taller and heavier than the average judoka in the dataset. Higher height and weight (more muscle mass) turn out to be an advantage. The same account for Serena Williams in tennis. On the contrary, Blyth Tait is much smaller and lighter than the average jockey in the list.

Sport	Top male athlete	Score	Height above average (cm)	Weight above average (kg)	Top female athlete	Score	Height above average (cm)	Weight above average (kg)
Football	C. Ronaldo	4,29	+7	9	Anja Mittag	2,99	0	+2
Basketball	Lebron James	5,78	5	-2	Diana Taurasi	4,99	+1,5	-7,5
Baseball	Barry Bonds	2,89	1,2	11	/	/	/	/
Tennis	Rafael Nadal	2,09	-2	2	Serena Williams	8,41	+6	+10
Table tennis	Xu Xing	2,75	-4	0,5	Kasumi Ishikawa	1,33	-3	-6
Swimming	Raiens Lohte	3,65	-2	4	Jenny Thompson	4,77	+3,25	+5
Golf	Tiger Woods	3,96	1	2	/	/		/
Cycling	Eric Heiden	2,55	5	12	Antonella Bellutti	-0,39	+10	+3
Equestrian sport	Blyth Tait	0,72	-12	-7	Anky v. Grunsven	3,38	+2	+4
Judo	Teddy Riner	6,08	24	54	M. Kelmendi	0,23	-4	-14

Figure 7, Athletes with the highest score

To fully investigate the impact of height and weight on the skill, a wider analysis is need by performing a regression on all data points. By means of the python library scikit-learn data analysis is done to dig deeper into the effect of variables like height and weight.

Effect of height on skill-level

(The complete regression analysis can be seen under:
https://colab.research.google.com/drive/1KLhuLi2CYxM_QdhRQAhdNii3jV1Q6Bj7)

A regression analysis on female tennis players (code block 119) shows a clear notion: higher players seem to have more success.

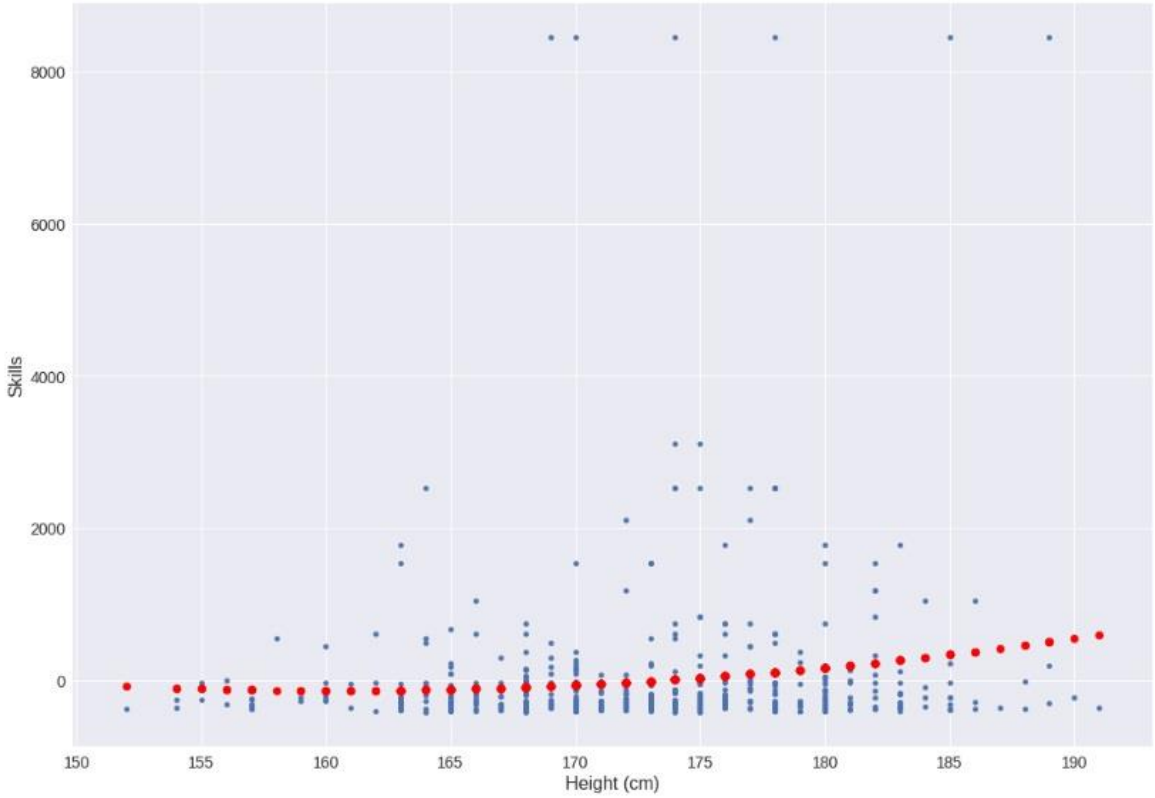


Figure 8, Female tennis players: Skill-score vs height

The next figure displays, Teddy Riner on the upper right corner. Here it seems his enormous height is simply an outlier and does not represent the sport as a whole.

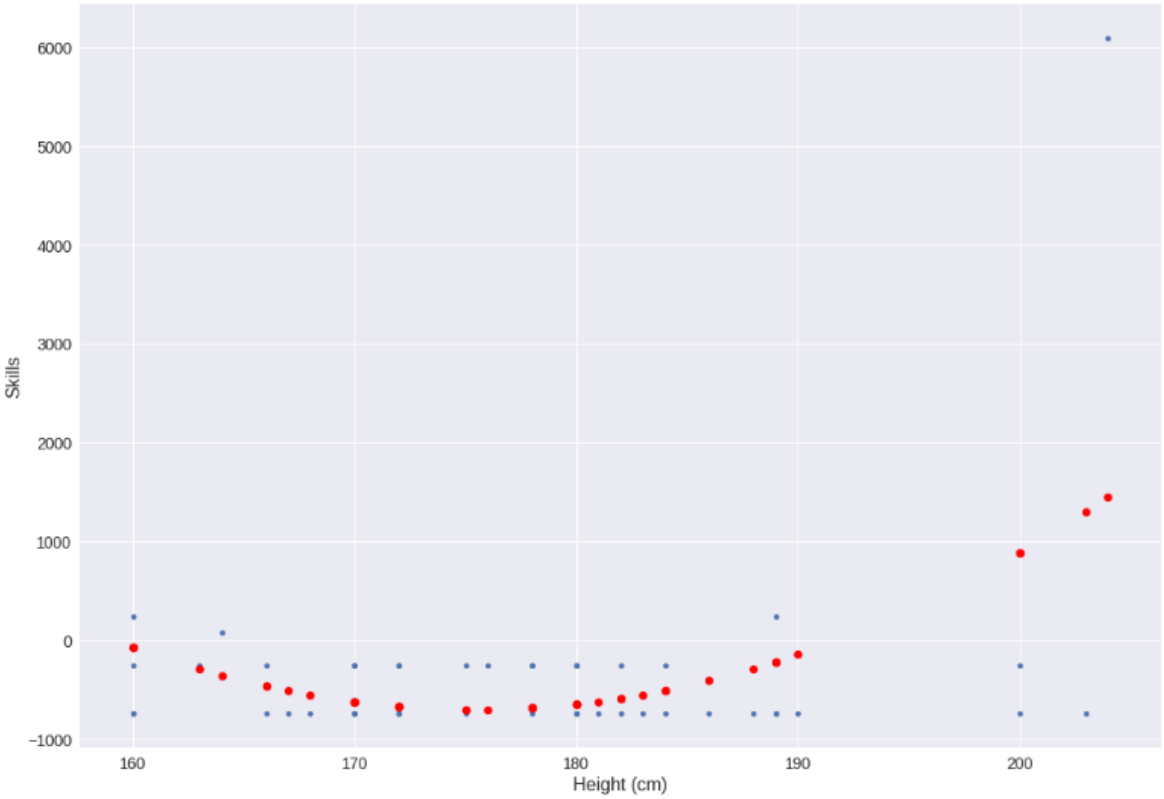


Figure 9, Male judoka: Skill-score vs height

For the remaining regression chart please refer to the Google Collaboration file.

Countries

This section investigates sports on the macro-level. Countries differ in their average height and weight. They have different life expectancy and a varying human development index. It is interesting to investigate whether there are links between these properties and sports.

The tables below indicate the effect of the number of athletes per sport and country on life expectancy and human development index of a country. As both metrics correlate high with each other (0.844), the numbers on the left and right side are similar.

```
[215] corr_matrix = countries.corr()
      corr_matrix["life_expectancy"].sort_values(ascending=False)

life_expectancy        1.000000
HDI                    0.844005
swimmers               0.228345
cyclists               0.217911
ice_hockeys            0.210216
tabletennisplayers     0.199219
volleyballs            0.184220
handballs              0.161369
Total_Athletes         0.160814
badmintons             0.157009
tennisplayers          0.153998
judokas                0.152929
basketballs            0.149950
boxers                 0.143692
footballplayers        0.135779
jockeys                0.123034
golf                   0.120577
rugbys                 0.119921
baseballs              0.116251
athletes_per_capita    0.110823
lacrosses              0.108273
fieldhockey            0.082429
avg_height             0.076620
cricketers             0.066138
aFootball              0.052186
avg_weight             0.001909
population            -0.017208
Name: life_expectancy, dtype: float64
```

```
[205] corr_matrix = countries.corr()
      corr_matrix["HDI"].sort_values(ascending=False)

HDI                    1.000000
life_expectancy        0.844005
ice_hockeys            0.269486
swimmers               0.183421
cyclists               0.175323
handballs              0.160507
basketballs            0.160157
lacrosses              0.150222
tabletennisplayers     0.147111
athletes_per_capita    0.132056
volleyballs            0.125265
rugbys                 0.124264
avg_height             0.122865
tennisplayers          0.122864
golf                   0.119811
aFootball              0.112538
Total_Athletes         0.106439
baseballs              0.090236
boxers                 0.088016
jockeys                0.086234
avg_weight             0.080458
footballplayers        0.080032
judokas                0.073411
badmintons             0.053287
fieldhockey            0.036864
cricketers             0.018327
population            -0.099946
Name: HDI, dtype: float64
```

Figure 10, Life expectancy and Human Development Index

Countries with a high number of swimmers, ice-hockey players, cyclist seem to have a higher standard of living measured by their HDI and life expectancy. In comparison, ‘athletes_per_capita’ has a far less effect on a higher standard of living.

Limitations

Technical limitation

The year-function used to assess the ‘peak-of-performance’ age does not assess the exact age. But as the average is taken, this evens out.

Athletes with no height or weight (or other variables) were not included. For some prominent missing athletes, the data was manually added, but there is clearly a bias. Some athletes had another metric for height and weight, namely inch and pound.

Wrong values for scored goals. For instance, some footballer had goals of 5 digits.

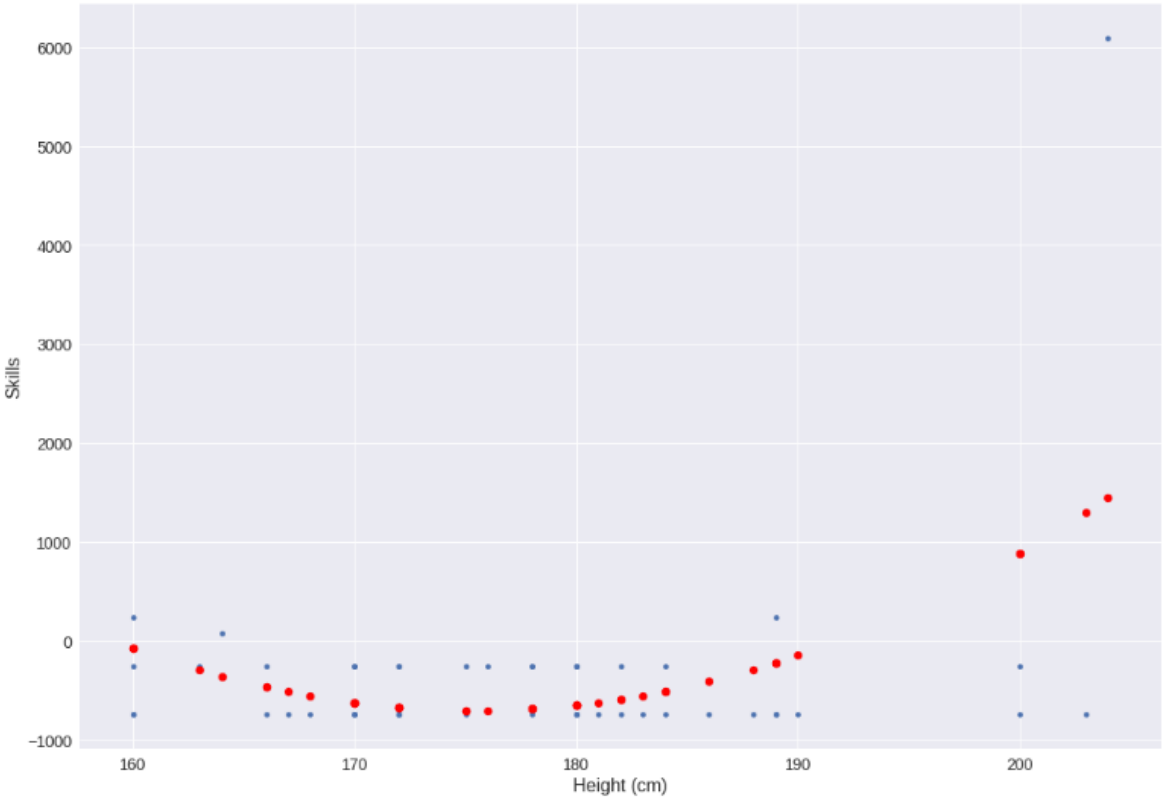
To criticize Wikidata, for some events the year was included into the label like ‘Winter Olympics 2015’. This restricts one to filter for the Winter Olympics of any year.

Legal limitation

Wikidata is full of data on prominent people including athletes. Here, the data gathered is publicly accessible and therefore align with copyright. Concerning the imports into Wikidata, the question of license is hereby self-explanatory as the source of the data is also its target.

Ethical issues

A considerable bias arises from the fact that more contributors to Wikipedia and Wikidata come from the 'western world' and are mainly male[6]. For the va



riable on nationality, the results are especially heavily biased by the persons who contributed to Wikidata.

Innovation

Developing a skill-score that is capable of comparing athletes among sports is very insightful. Further, exploring the effects of height and weight on the skill of athletes within and among sports uncovered characteristics of each investigated sport.

By transforming SPARQL-results into new datapoints and returning them back into a knowledge graph might be a new form of pre-processed data analytics. In Data Warehouses you find pre-constructed answers to frequently asked questions. The construction takes a long time. The same applies for the SPARQL-queries, especially query 26.

Conclusions

Becoming an expert in SPARQL took a long a time as the syntax requires a paradigm shift if one have worked with SQL before. Moreover, Wikidata has yet a special syntax for SPARQL. The concept of

[6] https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia

statements was crucial for this project. The prefixes for predicates ‘pq’ and ‘ps’ were needed to access and filter objects inside statements.

As soon as the queries were optimised and the desired properties were queried, Wikidata provided a very useful tool to exploit the advantages of being a knowledge graph. The pure size and reach of Wikidata enables a data gathering that not only allows to scale but also to include different domains. One can filter on specific objects, such as a gold medal awarded to all athletes of a specific sport in a specific event and gather the year at the same time. A depth of data exploration can be reached that without the property of a knowledge graph would not be possible.

The query results made a data analysis on the variables height and weight possible. The clearest effect could be seen for female tennis players. It is interesting to see how only two variables can differ so much among sports.

APPENDIX: Sparql Queries

Part 1: Skill-scores, Height, Weight, Gender, Nationality

Footballer

Query 1

```
# skill-level: number of goals scored per match played

# forwards, nationalteam, LaLiga, Premier League

# male only

(1) Select ?goals_per_matches ?football ?nationality ?height ?weight
(2) where{
(3) Bind((?total_goals / ?total_matches) as ?goals_per_matches)
(4) Filter (?total_matches > 200)
(5) Filter (?goals_per_matches > 0)
(6) {
(7) SELECT (SUM(?goals) as ?total_goals) (SUM(?matches) as ?total_matches) ?football
(8) (SAMPLE(?nationality) as ?nationality)(SAMPLE(?height) as ?height) (SAMPLE(?weight) as
    ?weight)
(9) WHERE {
(10)?football wdt:P106 wd:Q937857 ;
        i. wdt:P569 ?date_of_birth ;
        ii. wdt:P27 ?nationality;
        iii. wdt:P413 wd:Q280658;
        iv. wdt:P2048 ?height ;
        v. wdt:P2067 ?weight ;
        vi. p:P54 ?team_statement.
(11)?team_statement ps:P54 ?team.
(12)?team wdt:P118 ?league;
    a. wdt:P31 ?nationalteam.
(13)?team_statement pq:P1351 ?goals;
        i. pq:P1350 ?matches;
(14)Bind(year(?date_of_birth) as ?year_of_birth)
(15)filter(?year_of_birth > 1970)
(16)filter((?league = wd:Q324867 || ?league = wd:Q9448) && ?nationalteam = wd:Q6979593)
(17)}
(18)GROUP BY ?football
(19)}
(20)}
(21)Order By desc (?goals_per_matches)
```

Query 2

```
# skill-level: number of awards received

# female only

(1)  SELECT (COUNT (?award) as ?total_awards) ?football (SAMPLE(?nationality) as
     nationality)(SAMPLE(?height) as ?height) (SAMPLE(?weight) as ?weight)
(2)  WHERE {
(3)  ?football wdt:P106 wd:Q937857 ;
(4)  wdt:P569 ?date_of_birth ;
(5)  wdt:P27 ?nationality;
(6)  wdt:P21 wd:Q6581072;      #female
(7)  wdt:P2048 ?height ;
(8)  wdt:P2067 ?weight ;
(9)  wdt:P166 ?award.
(10)Bind(year(?date_of_birth) as ?year_of_birth)
(11)filter(?year_of_birth > 1970)
(12)}
(13)GROUP BY ?football
(14)Order by desc (?total_awards)
```

Baseballer

Query 3

```
# skill-level: number of awards

#delete Michael Jordan

SELECT (COUNT (distinct ?award) as ?total_awards)?baseball (SAMPLE(?nationality) as ?nationality)
(SAMPLE(?gender) as ?gender) (SAMPLE(?height) as ?height) (SAMPLE(?weight) as ?weight)

WHERE {

    ?baseball wdt:P106 wd:Q10871364;

            wdt:P569 ?date_of_birth;

            wdt:P2067 ?weight;

            wdt:P2048 ?height;

            wdt:P166 ?award;

            wdt:P21 ?gender;

            wdt:P27 ?nationality;

    Bind (year(?date_of_birth) as ?birth_year)

    Filter(?birth_year > 1970)

    }

    Group by ?baseball
```

Order by desc (?total_awards)

Swimmer

Query 4

Skill: number of gold medals

```
(1) SELECT (COUNT(?awarded) as ?goldmedals) ?swimmer (SAMPLE(?nationality) as
    ?nationality)(SAMPLE(?gender) as ?gender)(sample(?height) as ?height)(sample(?weight) as
    ?weight)
(2) WHERE {
(3) ?swimmer wdt:P106 wd:Q10843402;
(4)             wdt:P27 ?nationality;
(5)             wdt:P21 ?gender ;
(6)             wdt:P2048 ?height;
(7)             wdt:P2067 ?weight;
(8)             wdt:P21 ?gender;
(9)             p:P1344 ?competition_statement.
(10)?competition_statement pq:P166 ?awarded.
(11)Filter (?awarded = wd:Q15243387)
(12)}
(13)group by ?swimmer
(14)order by Desc (?goldmedals)
```

Jockey

Query 5

#skill-level: number of participations

```
(1) SELECT (COUNT (?competition) as ?competitions) ?jockey (SAMPLE(?gender) as ?gender)
    (SAMPLE(?nationality) as ?nationality)(SAMPLE(?height) as ?height) (SAMPLE(?weight) as
    ?weight)
(2) WHERE {
(3) ?jockey wdt:P106 wd:Q846750;
(4)         wdt:P27 ?nationality;
(5)         wdt:P2067 ?weight;
(6)         wdt:P2048 ?height;
(7)         wdt:P21 ?gender;
(8)         wdt:P1344 ?competition.
(9) }
(10)Group by ?jockey
(11)Order by desc(?competitions)
```

Judoka

Query 6

Skill: number of gold medals

```
(1) SELECT (COUNT(?awarded) as ?awards) ?judoka (SAMPLE(?nationality) as
    nationality)(SAMPLE(?gender) as ?gender)(sample(?height) as ?height)(sample(?weight) as
    ?weight)
(2) WHERE {
(3) ?judoka wdt:P106 wd:Q6665249;
(4)             wdt:P27 ?nationality;
(5)             wdt:P21 ?gender ;
(6)             wdt:P2048 ?height;
(7)             wdt:P2067 ?weight;
(8)             p:P1344 ?event1.
(9) ?event1 pq:P166 ?awarded.
(10)Filter (?awarded = wd:Q15243387 || ?awarded = wd:Q406039)
(11)            }
(12)group by ?judoka
(13)order by Desc (?goldmedals)
```

Cyclist

Query 7

```
# Skill: number of gold medals

# all male

(1) SELECT (COUNT(?awarded) as ?goldmedals) ?cyclist (sample(?nationality) as
    ?nationality)(SAMPLE(?gender) as ?gender) (sample(?height) as ?height)(sample(?weight) as
    ?weight)
(2) WHERE {
(3) ?cyclist wdt:P106 wd:Q2309784;
(4)             wdt:P27 ?nationality;
(5)             wdt:P21 ?gender ;
(6)             wdt:P2048 ?height;
(7)             wdt:P2067 ?weight;
(8)             wdt:P21 ?gender;
(9)             p:P1344 ?event1.
(10)?event1 pq:P166 ?awarded.
(11)Filter (?awarded = wd:Q15243387)
(12)        }
(13)group by ?cyclist
(14)order by desc(?goldmedals)
```

Tennis player

Query 8

```
# skill-level: average ranking

# 500 female only 71 male

(1) SELECT ?ranking_number ?tennis_player ?nationality ?gender ?height ?weight
(2) WHERE{
(3) ?tennis_player wdt:P21 ?gender;
(4)                     wdt:P27 ?nationality;
```

```
(5)                    wdt:P2048 ?height;
(6)                     wdt:P2067 ?weight.
(7)  Filter(?number_of_ranking_numbers > 1)
(8)          {
(9)            SELECT (AVG(?ranking_number) as ?ranking_number)(COUNT(?ranking_number) as
(10)                    ?number_of_ranking_numbers) ?tennis_player
(11)           WHERE {
(12)                    ?tennis_player wdt:P106 wd:Q10833314;
(13)                                   wdt:P569 ?date_of_birth;
(14)                                   p:P1352 ?ranking_statement.
(15)                   ?ranking_statement pq:P641 ?sport;
(16)                                      ps:P1352 ?ranking_number.
(17)                    BIND (year(?date_of_birth) as ?birthyear)
(18)                    Filter(?sport = wd:Q18123880 || ?sport = wd:Q18123885)
(19)                    Filter(?birthyear > 1970)
(20)                    }
(21)                    GROUP BY ?tennis_player
(22)         }}
(23)order by ?ranking_number
```

Table-tennis

Query 9

skill-level: average ranking

```
(1)  SELECT (AVG(?ranking_number) as ?average_ranking_number) ?table_tennis
     (SAMPLE(?nationality) as ?nationality)( SAMPLE (?gender) as ?gender) (SAMPLE (?height) as
     ?height) (SAMPLE (?weight) as ?weight)
(2)  WHERE {
(3)  ?table_tennis wdt:P106 wd:Q13382519;
(4)  wdt:P27 ?nationality;
(5)  wdt:P21 ?gender;
(6)  wdt:P569 ?birth_date;
        a.  wdt:P2048 ?height;
        b.  wdt:P2067 ?weight;
(7)  p:P1352 ?ranking_statement.
(8)  ?ranking_statement ps:P1352 ?ranking_number.
(9)  Bind (year(?birth_date) as ?birth_year)
(10)Filter (?birth_year > 1970)
(11)}
(12)Group by ?table_tennis
(13)order by asc (?average_ranking_number)
```

Basketball

Query 10

skill-level: number of awards

```
(1) SELECT (Count(distinct ?award) as ?awards)?basketball(SAMPLE(?nationality) as ?nationality)
    (SAMPLE(?gender) as ?gender)(SAMPLE(?height) as ?height) (SAMPLE(?weight) as ?weight)
(2) WHERE {
(3) ?basketball wdt:P106 wd:Q3665646;
(4)             wdt:P27 ?nationality;
(5)             wdt:P21 ?gender;
(6)             wdt:P569 ?birth_date;
(7)             wdt:P2067 ?weight;
(8)             wdt:P2048 ?height;
(9)             wdt:P166 ?award.
(10)Bind (year(?birth_date) as ?birth_year)
(11)Filter(?birth_year > 1970)
(12)        }
(13)Group by ?basketball
(14)Order by desc(?awards)
```

Golf-Players

Query 11

#Skill: number of awards

```
(1) SELECT (Count(?award) as ?awards) ?golf_player (SAMPLE(?gender) as
    ?gender)(SAMPLE(?nationality) as ?nationality)
(2) (SAMPLE(?height) as ?height) (SAMPLE(?weight) as ?weight)
(3) WHERE {
(4) ?golf_player wdt:P106 wd:Q11303721;
(5)              wdt:P569 ?date_of_birth;
(6)              wdt:P27 ?nationality;
(7)              wdt:P2067 ?weight;
(8)              wdt:P2048 ?height;
(9)              wdt:P166 ?award;
(10)             wdt:P21 ?gender.
(11)Bind (year(?date_of_birth)as?birthyear)
(12)Filter (?birthyear > 1960)
(13)        }
(14)Group by ?golf_player
(15)Order by Desc (?awards)
```

Part 2: Peak-Performance age

Football

Query 12

#age of players when they won the Ballon d'or / Fifa ballon d'or

```
(1) SELECT (AVG(?peak_age) as ?average_peak_age)
(2) WHERE {
(3) ?football wdt:P106 wd:Q937857;
(4)           wdt:P569 ?date_of_birth;
(5)           p:P166 ?award_statement.
```

```
(6)  ?award_statement pq:P585 ?award_date;
(7)                           ps:P166 ?award.
(8)  BIND (Year (?date_of_birth) as ?birthyear)
(9)  BIND (Year (?award_date) as ?eventyear)
(10)BIND ((?eventyear - ?birthyear) as ?peak_age)
(11)Filter(?award = wd:Q166177 ||?award = wd:Q2291862)
(12)Filter(?birthyear > 1970)
(13)}
```

Query 13

#age of players when they won the Fifa worldplayer / Female football player of the year

#female

```
(1)  SELECT (AVG(?peak_age) as ?average_peak_age)
(2)  WHERE {
(3)  ?football wdt:P106 wd:Q937857;
(4)            wdt:P569 ?date_of_birth;
(5)            wdt:P21 wd:Q6581072 ;
(6)            p:P166 ?award_statement.
(7)  ?award_statement pq:P585 ?award_date;
(8)                           ps:P166 ?award.
(9)  BIND (Year(?date_of_birth) as ?birthyear)
(10)BIND (Year (?award_date) as ?eventyear)
(11)BIND ((?eventyear - ?birthyear) as ?peak_age)
(12)Filter(?award = wd:Q182529 ||?award = wd:Q1478325)
(13)Filter(?birthyear > 1970)
(14)}
```

Baseball

Query 14

peak_age: age when player became 'Player of the year'

only male

```
(1)  SELECT distinct (AVG(?peak_age)as ?peak_age)
(2)  WHERE {
(3)  ?baseball wdt:P106 wd:Q10871364;
(4)            wdt:P569 ?date_of_birth;
(5)            p:P166 ?award_statement.
(6)  ?award_statement pq:P585 ?award_date;
(7)                           ps:P166 wd:Q15140008.
(8)  BIND (year(?date_of_birth) as ?birth_year)
(9)  BIND (year(?award_date) as ?award_year)
(10)BIND ((?award_year - ?birth_year) as ?peak_age)
(11)}
```

Query 15

peak_age: age when player became 'Player of the year'

only female

```
(1)  SELECT  (AVG(?peak_age) as ?peak_age)
(2)  WHERE {
(3)  ?baseball wdt:P106 wd:Q13388586;
(4)  wdt:P569 ?date_of_birth;
(5)  wdt:P1344 ?competition.
(6)  ?competition wdt:P585 ?competition_date.
(7)  BIND (year(?date_of_birth) as ?birth_year)
(8)  BIND (year(?competition_date) as ?competition_year)
(9)  BIND ((?competition_year - ?birth_year) as ?peak_age)
(10)}
```

Swimmer

Query 16

#average age of a swimmer when he/she won a gold medal

```
(1)  SELECT distinct (AVG(?peak_age)as ?peak_age) ?gender
(2)  WHERE {
(3)  ?swimmer wdt:P106 wd:Q10843402;
(4)                wdt:P21 ?gender;
(5)                wdt:P569 ?birth_date;
(6)                 p:P1344 ?competition_statement.
(7)  ?competition_statement  pq:P166 wd:Q15243387;
(8)                                  ps:P1344 ?competition.
(9)  ?competition wdt:P585 ?competition_date.
(10)Bind(year(?competition_date) as ?competition_year)
(11)Bind(year(?birth_date) as ?birth_year)
(12)Bind((?competition_year - ?birth_year) as ?peak_age)
(13)Filter(?birth_year > 1960)
(14)}
(15)Group by ?gender
```

Jockey

Query 17

#peak_age: age when jockeys won a medal at olympics

male: 37 female 34

```
(1)  SELECT (AVG(?peak_age) as ?peak_age) ?gender
(2)  WHERE {
(3)  ?jockey wdt:P106 wd:Q846750;
(4)            wdt:P569 ?birthday;
(5)            wdt:P21 ?gender;
(6)            p:P1344 ?competition_statement.
(7)  ?competition_statement pq:P166 ?award;
```

```
(8)                                ps:P1344 ?competition.
(9)  ?competition wdt:P585 ?competition_date.
(10)BIND (year(?birthday) as ?birth_year)
(11)BIND (year(?competition_date) as ?competition_year)
(12)BIND ((?competition_year - ?birth_year) as ?peak_age)
(13)            }
(14)group by ?gender
```

Judoka

Query 18

Skill: number of gold medals

```
(1)  SELECT (avg(?peak_age) as ?avg_peak_age) ?gender
(2)  WHERE {
(3)  ?judoka wdt:P106 wd:Q6665249;
(4)             wdt:P21 ?gender;
(5)             wdt:P569 ?date_of_birth;
(6)             p:P1344 ?competition_statement.
(7)  ?competition_statement pq:P166 ?awarded;
(8)                                ps:P1344 ?competition.
(9)  ?competition wdt:P585 ? competition_date.
(10)BIND (year(?date_of_birth) as ?birth_year)
(11)BIND (year(?competition_date) as ?event_year)
(12)BIND ((?event_year - ?birth_year) as ?peak_age)
(13)Filter (?awarded = wd:Q15243387 || ?awarded = wd:Q406039)
(14)            }
(15)group by ?gender
```

Cycling

Query 19

#average age of a cyclist when he/she won a gold medal

female = male

```
(1)  SELECT distinct (Avg(?peak_age)as ?peak_age) ?gender
(2)  WHERE {
(3)  ?cyclist     wdt:P106 wd:Q2309784;
(4)  wdt:P21 ?gender;
(5)  wdt:P569 ?birth_date;
(6)  p:P1344 ?competition_statement.
(7)  ?competition_statement pq:P166 wd:Q15243387;
(8)                                ps:P1344 ?competition.
(9)  ?competition wdt:P585 ?competition_date.
(10)Bind(year(?competition_date) as ?competition_year)
(11)Bind(year(?birth_date) as ?birth_year)
(12)Bind((?competition_year - ?birth_year) as ?peak_age)
(13)Filter(?birth_year > 1960)
(14)                    }
```

(15)group by ?gender

Tennis

Query 20

#Peaktime: the age when player was ranked as no.1 in tennis singles

the same for both genders: 23

```
(1)  SELECT (AVG(?peaktime) as ?peaktime) ?gender
(2)  WHERE {
(3)  ?tennis_player wdt:P106 wd:Q10833314;
(4)                 wdt:P569 ?date_of_birth;
(5)                 wdt:P21 ?gender ;
(6)                 p:P1352 ?ranking_statement.
(7)  ?ranking_statement pq:P585 ?ranking_date;
(8)                         ps:P1352 ?ranking_number;
(9)                         pq:P641 wd:Q18123880.
(10)BIND (year(?date_of_birth) as ?birth_year)
(11)BIND (year(?ranking_date) as ?ranking_year)
(12)BIND ((?ranking_year - ?birth_year) as ?peaktime)
(13)Filter(?ranking_number = 1)
(14)Filter(?birth_year > 1970)
(15)}
```

group by ?gender

Table tennis player

Query 21

peak-age: year of ranking between 1-5

male: 27 female: 24

```
(1)  SELECT distinct (AVG(?peak_age)as ?peak_age_per_gender) ?gender
(2)  WHERE {
(3)  ?tennis wdt:P106 wd:Q13382519;
(4)          wdt:P21 ?gender;
(5)          wdt:P569 ?birth_date;
(6)          p:P1352 ?ranking_statement.
(7)  ?ranking_statement pq:P580 ?ranking_date;
(8)                         ps:P1352 ?ranking.
(9)  Bind(year(?ranking_date) as ?ranking_year)
(10)Bind(year(?birth_date) as ?birth_year)
(11)Bind((?ranking_year - ?birth_year) as ?peak_age)
(12)Filter(?ranking < 6)
(13)}
(14)Group by ?gender
```

Basketball

Query 22

#Peak_age: age when basketballer received more than 3 awards

```
(1)  SELECT (AVG(?peak_age) as ?avg_peak_age) ?gender
(2)  WHERE{
(3)  BIND ((?award_year - ?birth_year) as ?peak_age)
(4)            {
(5)             SELECT distinct ?basketballer (sample(?gender) as ?gender) ?award_year
(6)                        (sample(?birth_year) as ?birth_year) (count(?award_year) as ?awardCount)
(7)             WHERE {
(8)                          ?basketballer wdt:P106 wd:Q3665646;
(9)                                        wdt:P569 ?birth_date;
(10)                                       wdt:P21 ?gender;
(11)                                       p:P166 ?award_statement .
(12)                         ?award_statement pq:P585 ?award_date.
(13)                         BIND (year(?birth_date) as ?birth_year)
(14)                         BIND ((?award_year - ?birth_year) as ?peak_age)
(15)                         Bind (year(?award_date) as ?award_year)
(16)                         Filter (?birth_year > 1960)
(17)                         }
(18)                         GROUP BY ?basketballer ?award_year
(19)                         HAVING (?awardCount > 3)
(20)            }
(21)            }
(22)GROUP BY ?gender
```

Golf

Query 23

#Peak_age: age where golfer received more than two awards

female = male

```
(1)  SELECT (AVG(?peak_age) as ?avg_peak_age) ?gender
(2)  WHERE{
(3)  BIND ((?award_year - ?birth_year) as ?peak_age)
(4)            {
(5)             SELECT distinct ?golfer (sample(?gender) as ?gender) ?award_year
(6)             (sample(?birth_year) as ?birth_year) (count(?award_year) as ?awardCount)
(7)             WHERE {
(8)             ?golfer wdt:P106 wd:Q11303721;
(9)                          wdt:P569 ?birth_date;
(10)                         wdt:P21 ?gender;
(11)                         p:P166 ?award_statement .
(12)            ?award_statement pq:P585 ?award_date.
(13)            BIND (Year(?birth_date) as ?birth_year)
(14)            BIND ((?award_year - ?birth_year) as ?peak_age)
(15)            BIND (Year (?award_date) as ?award_year)
(16)            Filter (?birth_year > 1960)
```

```
(17)                    }
(18)              GROUP BY ?golfer ?award_year
(19)              HAVING (?awardCount > 1)
(20)            }
(21)          }
(22)GROUP BY ?gender
```

Country comparison

Most played sports in the world

Query 24

```
(1)  SELECT ?sportLabel (COUNT(?player) as ?num_players)
(2)  WHERE {
(3)  ?player wdt:P106 ?sportplayer.
(4)             {
(5)               SELECT distinct ?sportplayer ?sportLabel
(6)               WHERE {
(7)               ?sportplayer wdt:P425 ?sport;
(8)                                wdt:P279 wd:Q2066131.
(9)               ?sport wdt:P31   wd:Q31629.
(10)                    }
(11)              }}
(12)group by ?sportLabel
(13)Order by ?num_players
```

Sportiest nations, number of players per country / population

Query 25

```
(1)  SELECT ?player_per_capita (sample(?countryLabel) as ?countryLabel)
(2)  WHERE {
(3)  Bind ((?num_players / population) as ?player_per_capita)
(4)             {
(5)               SELECT distinct ?countryLabel (COUNT(?player) as ?num_players)
(6)               WHERE {
(7)               ?player wdt:P106 ?sportplayer.
(8)               ?player wdt:P27 ?country.
(9)               ?country wdt:P1082 ?population.
(10)              Filter(?population > 5000000)
(11)                        {
(12)                        SELECT distinct ?sportplayer
(13)                        WHERE {
(14)                        ?sportplayer wdt:P425 ?sport;
```

```
(15)                                    wdt:P279 wd:Q2066131.
(16)                    ?sport wdt:P31   wd:Q31629.
(17)                                    }
(18)                     }}
(19)            group by ?countryLabel
(20)            }}
(21)            Order by ?player_per_capita
```

Number of players per country & sport

Query 26

```
(1)             SELECT ?countryLabel ?sportLabel (COUNT(?player) as ?num_players)
(2)              WHERE {
(3)             ?player wdt:P106 ?sportplayer.
(4)             ?player wdt:P27 ?country.
(5)                             {
(6)                              SELECT ?sportplayer ?sportLabel
(7)                              WHERE {
(8)                              ?sportplayer wdt:P425 ?sport;
(9)                                             wdt:P279 wd:Q2066131.
(10)                             ?sport wdt:P31   wd:Q31629.
(11)                                            }
(12)                             }}
(13)            group by ?countryLabel ?sportLabel
(14)            }}
(15)            Order by ?player_per_capita
```

Most played sports per countries, HDI, life expectancy population

Query 27

```
(1)  SELECT distinct ?sportLabel (COUNT(?player) as ?players) (sample(?country) as
     ?country)(sample(?population) as ?population) (sample(?life_expectancy) as
     ?life_expectancy)(sample(?HDI) as ?HDI)
(2)  WHERE {
(3)  ?player wdt:P106 ?sportplayer.
(4)  ?player wdt:P27 ?country.
(5)  Optional{ ?country wdt:P1082 ?population. }
(6)  Optional{ ?country wdt:P2250 ?life_expectancy. }
(7)  Optional{ ?country wdt:P1081 ?HDI. }
(8)             {
(9)              SELECT distinct ?sportplayer ?sportLabel
(10)             WHERE {
(11)             ?sportplayer wdt:P425 ?sport.
(12)             ?sport wdt:P31   wd:Q31629;
(13)                     wdt:P279 ?uppersport.
(14)             ?uppersport wdt:P279 wd:Q877517.
(15)             ?sportplayer wdt:P279 wd:Q2066131.
```

(16) }
(17) }}
(18)group by ?sportLabel ?country
(19)Order by ?players