A Needle in a Data Haystack project

1.1 Writeup

- Project Title: Sports – Stats & Facts
- Team member info:

  - Lior Cohen, lior.cohen8@mail.huji.ac.il, lior_13
  - Yaacov Yonatan Goldberger, yaacov.goldberger@mail.huji.ac.il, jjgold
  - Ori Broda, ori.broda@mail.huji.ac.il, orib

- Link to the project repository on github: https://github.com/jjgold012/data-science
  Important: Limited by size, the code we provided in the code_group11 file contains only the .py files we used in our project without the data. Thus, in order to enjoy the full functionality, one should download the repository's contents from the above link.
- We also included a video sample of the project's GUI found as 'project_clip.mp4' in the repository.
- Problem description: We set to evaluate the precision in predicting a sport event result (who'll be the winner). Our initial assumption is that sport events are hard to predict because there are a lot of variables to consider, and sport involve a large portion of luck.
- Data: We found many data on sport types in .csv and .xls formats.
  Those files were found in different betting websites and forums that share them, such as bet365.
  Each excel contains data on many matches occurred over years/seasons (we got from about 4 years of data in hockey to about 40 years in football) and betting odds from betting websites. The main attributes we addressed are: match date, the league/tournament, names of competing teams, score of each team, the winner and the betting odds.
  The different sport types we collected and the amount of records are:
  soccer – 113157
  football – 35795 for american and 1508 for australian
  basketball – 11021
  hockey – 6848
  rugby – 2563
  tennis – 41187 for men and 24705 for women
  cricket – 194
  Summing up for a total of 236978 records.
  Total amount of space used for repository: 466MB

- We took all the excel files we downloaded and inserted into SQL tables, with scripts we wrote, the columns we thought are interesting the most which are mentioned above, and created a table for each sport type. All can be found in the git folder. After doing so we were able to perform any SQL queries we wanted and display it in different ways. We decided to display the queries as facts and as graphs.
  Every detail can be observed using an easy-to-use software we developed, which can hopefully be installed and executed in your computer (in case problems occur, we also added a nice sample video demo ☺)

- Experiments: Our project was involved mostly handling and creating the database, it required an extensive search online for data from betting companies. Given the data we managed to put our hand on, we did came to a conclusion that sport is very unpredictable (about 25 precent on avg of the betting companies prediction were wrong as shown by the query result on our database below).

- Future work: First, there can be more features to add to the GUI such as more charts involving betting odds statistics (we only included about 15 charts below) and find more cool things to do with the data. It is possible to extend the tables we built and add a column of the referee and reveal dark secrets about them, like their favorite team that barely lose while the referee in charge. Also, we can decide, by the months matched where played, what is the season with the highest chance to predict scores and make a lot of money. Can look for the temperature in that day and check if the weather affect game result. Another possibility for future work is to expend the data to other sport types, like cycling, horse racing, F1 and the Olympics.

- Conclusion: We managed to collect various data and establish a decent sports DB. Using the data we were able to obtain valuable and solid information about teams/players over the course of more than a decade. We believe that it is possible to predict future game results at some high probability based on this data, and maybe even help the betting sites to be more accurate (for example, we saw in the 'underdog' chart (see below) that the amount of wrong predictions is roughly around 25% for the majority of sports).

README

####### Project technical information #######

- The database is stored on localhost MySQL Server 5.7

- We used python 2.7.13 for establishing the database, while the GUI runs with python 3.

- The database runs on 'root' user with no password using charset 'utf-8'.

- Some of the queries we used can be found in 'queries.txt' which is placed in the main folder 'data-science'.

####### Installation instructions #######

1. Download and install MySQL Server 5.7: https://dev.mysql.com/downloads/installer/

2. Download and install the prefered python version: https://www.python.org/downloads/

3. Install the required packages (see below).

4. Run the file 'run.py'

   This should create and fill the database which then can be queried using SQL statements.

5. Run the file 'projectGUI.py'

   This should display the GUI as shown in the video.

####### Pip modules required for the project to run #######

DB modules:

- pymysql

- xlrd

- dateutil

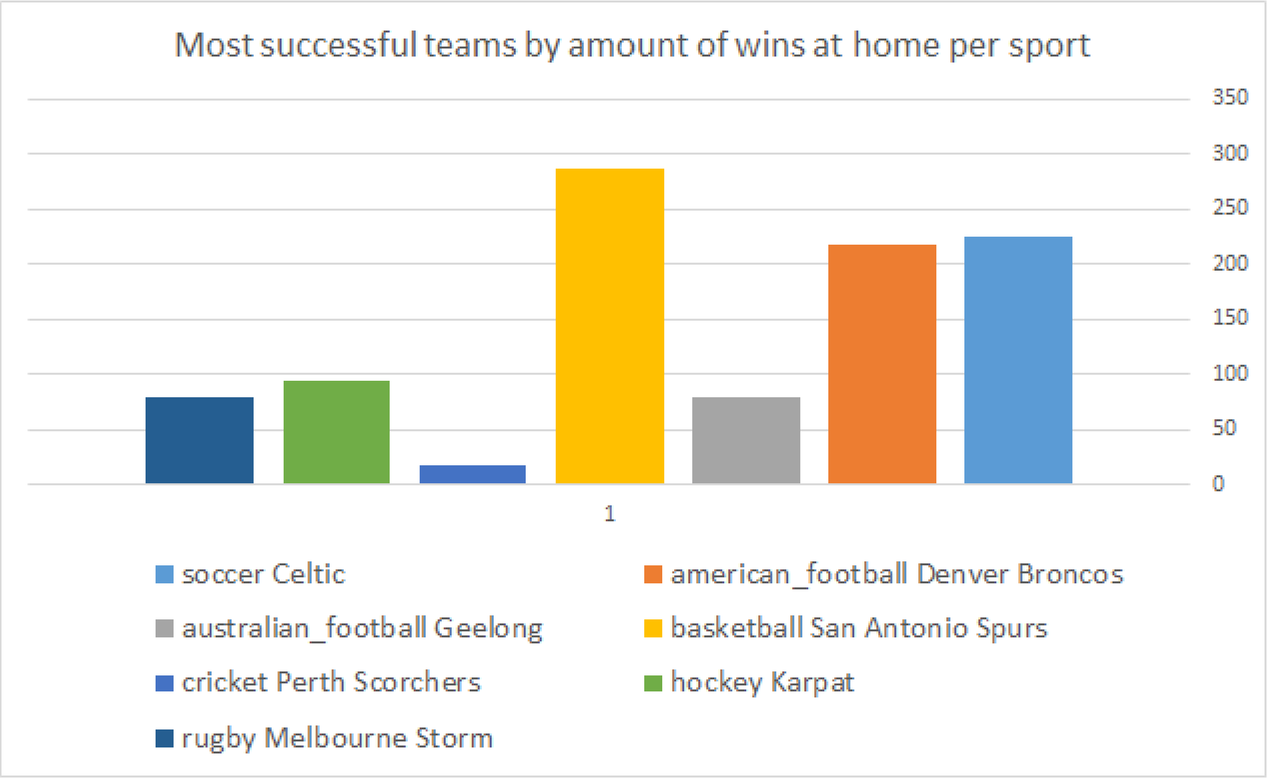
- calendar

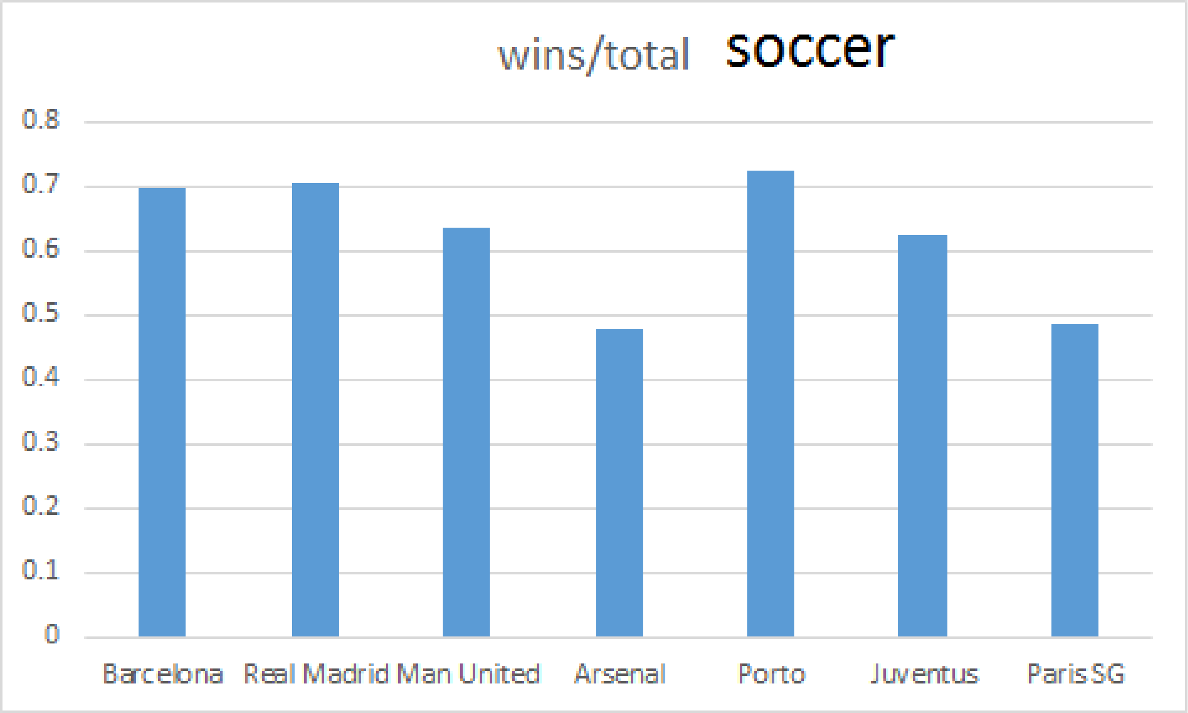
GUI modules:

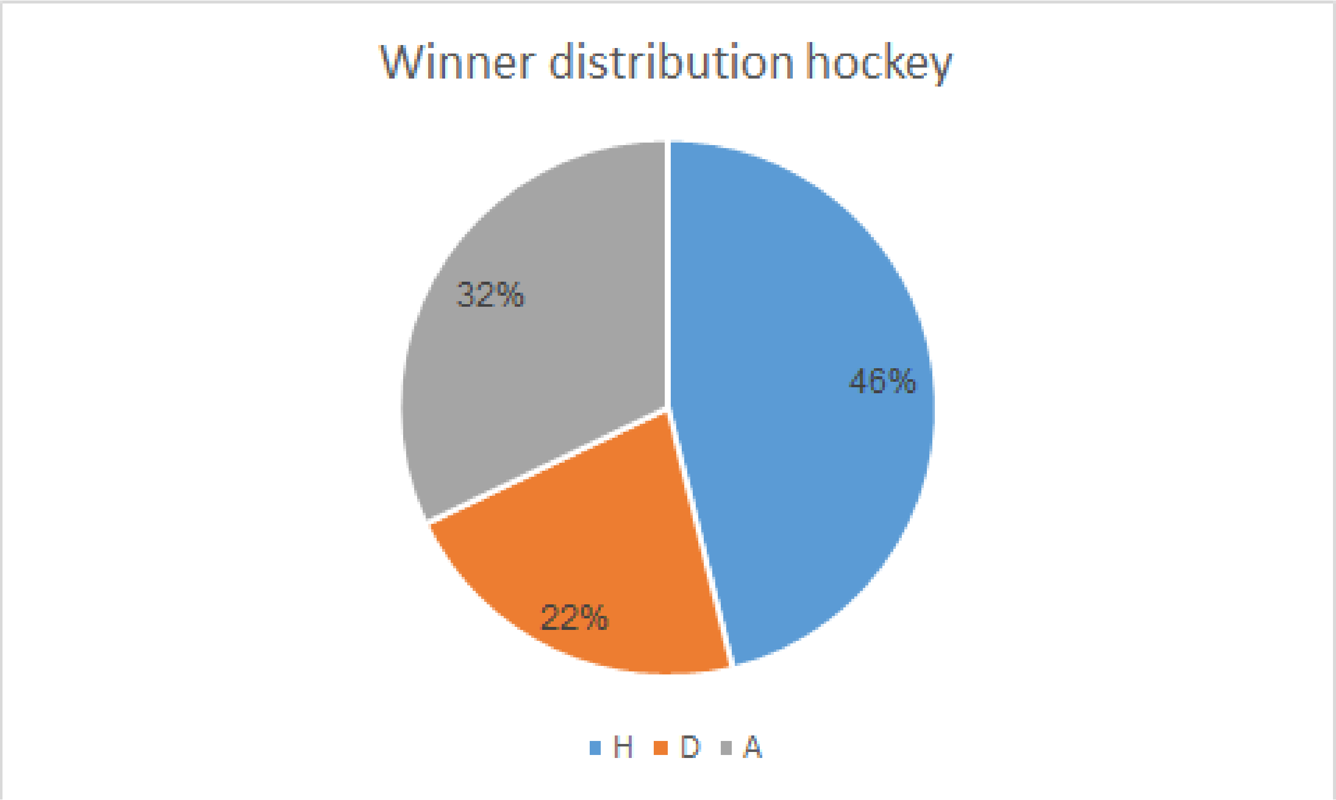
- tkinter

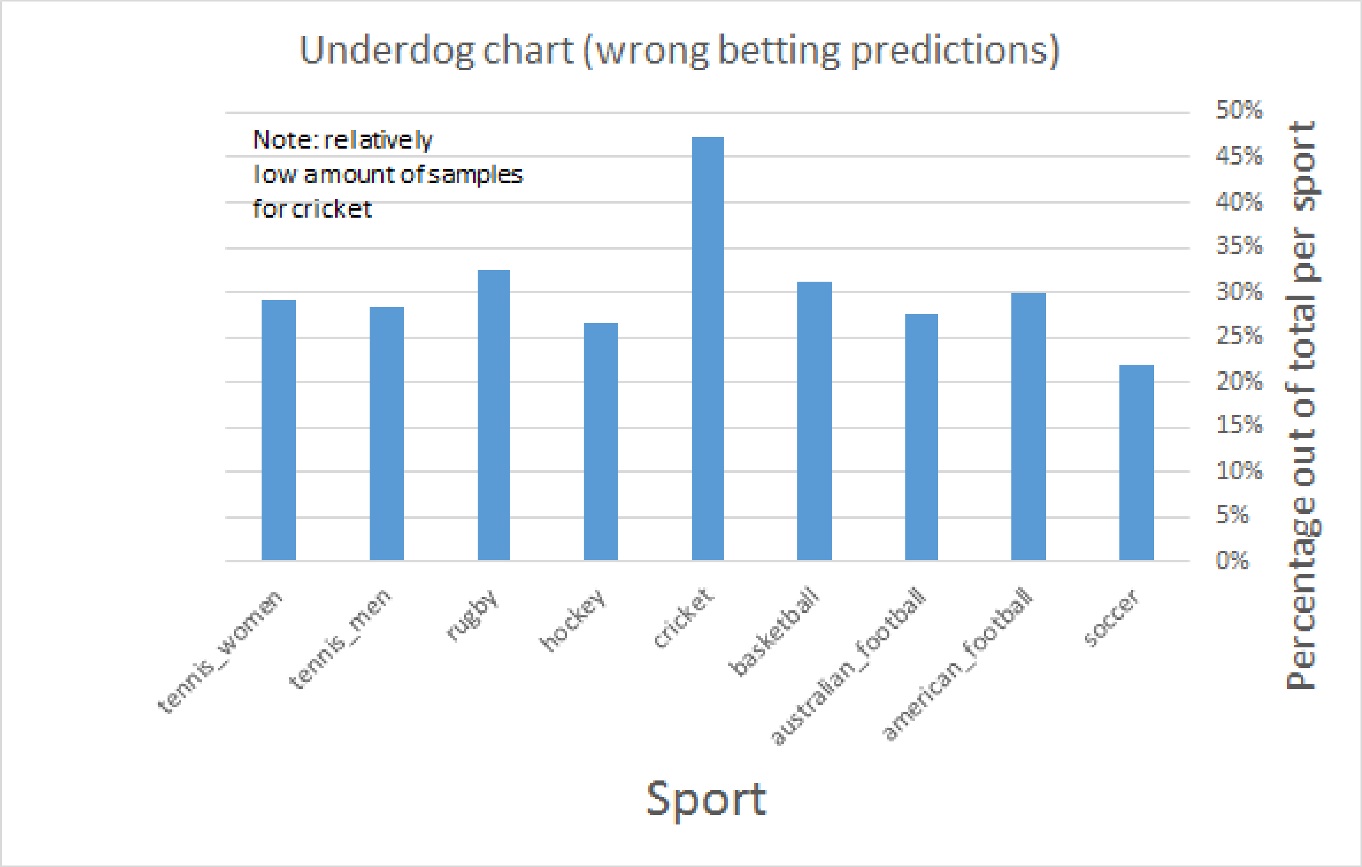
- matplotlib

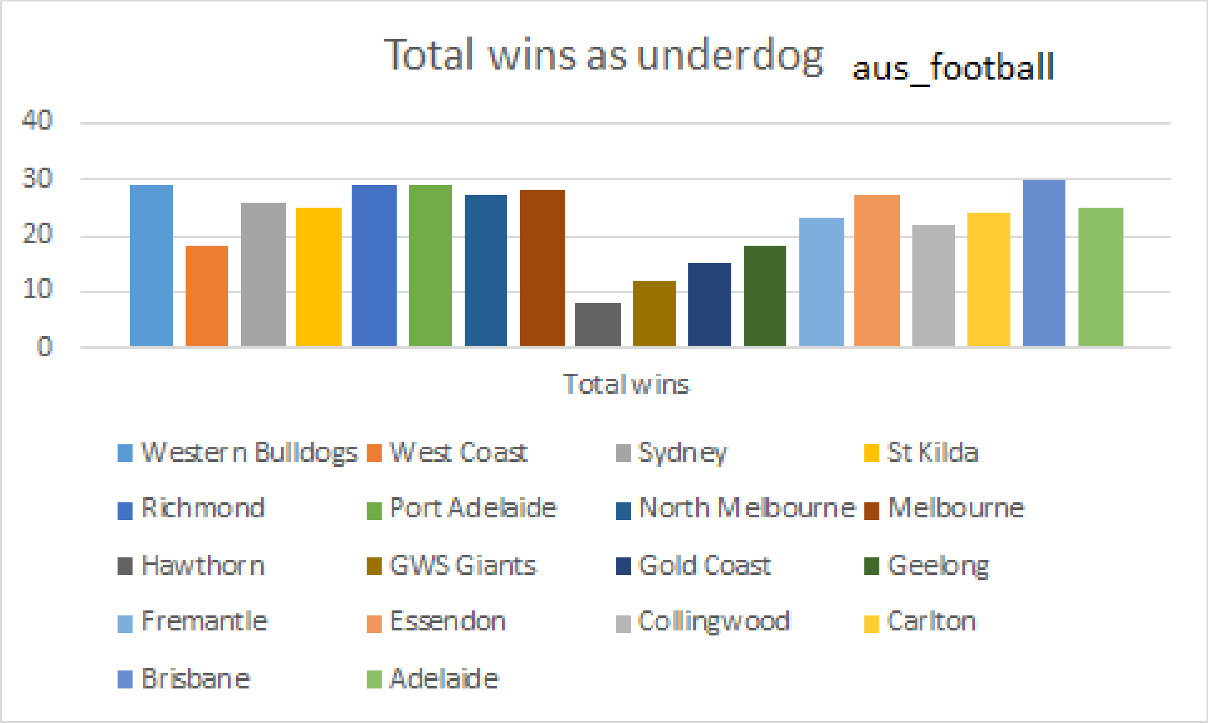
- pymysql

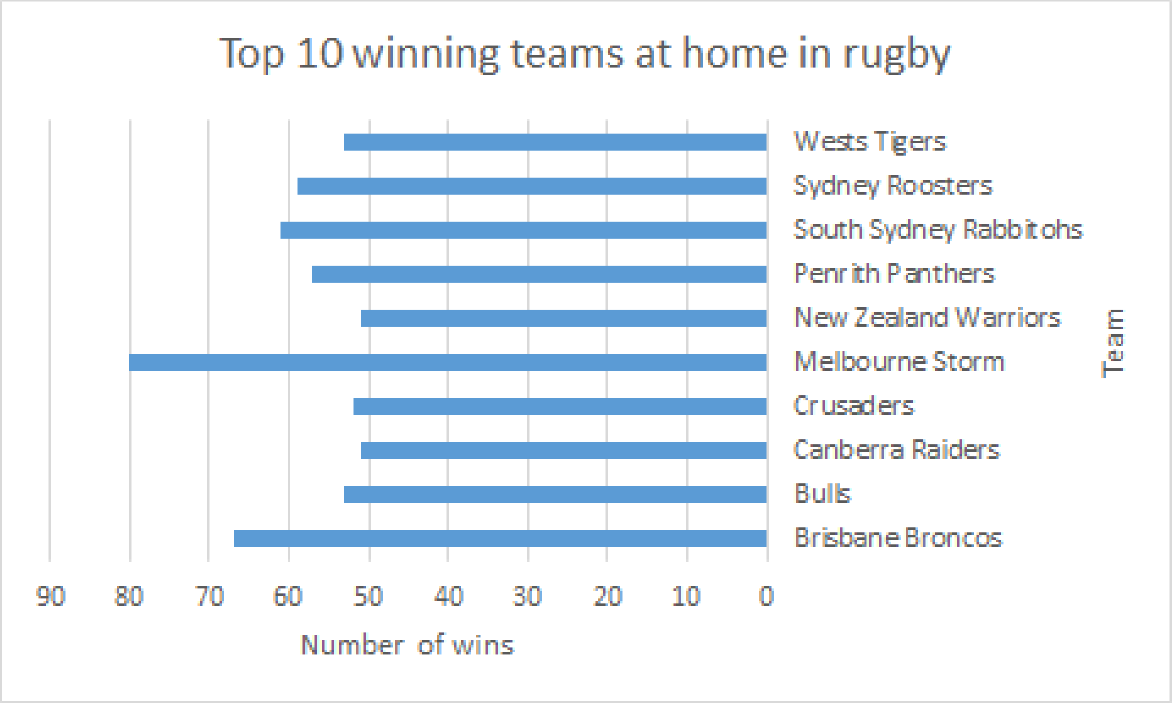

Most successful teams by amount of wins at home per sport
350
300
250
200
150
100
50
0
1
soccer Celtic
american_football Denver Broncos
australian_football Geelong
basketball San Antonio Spurs
cricket Perth Scorchers
hockey Karpat
rugby Melbourne Storm

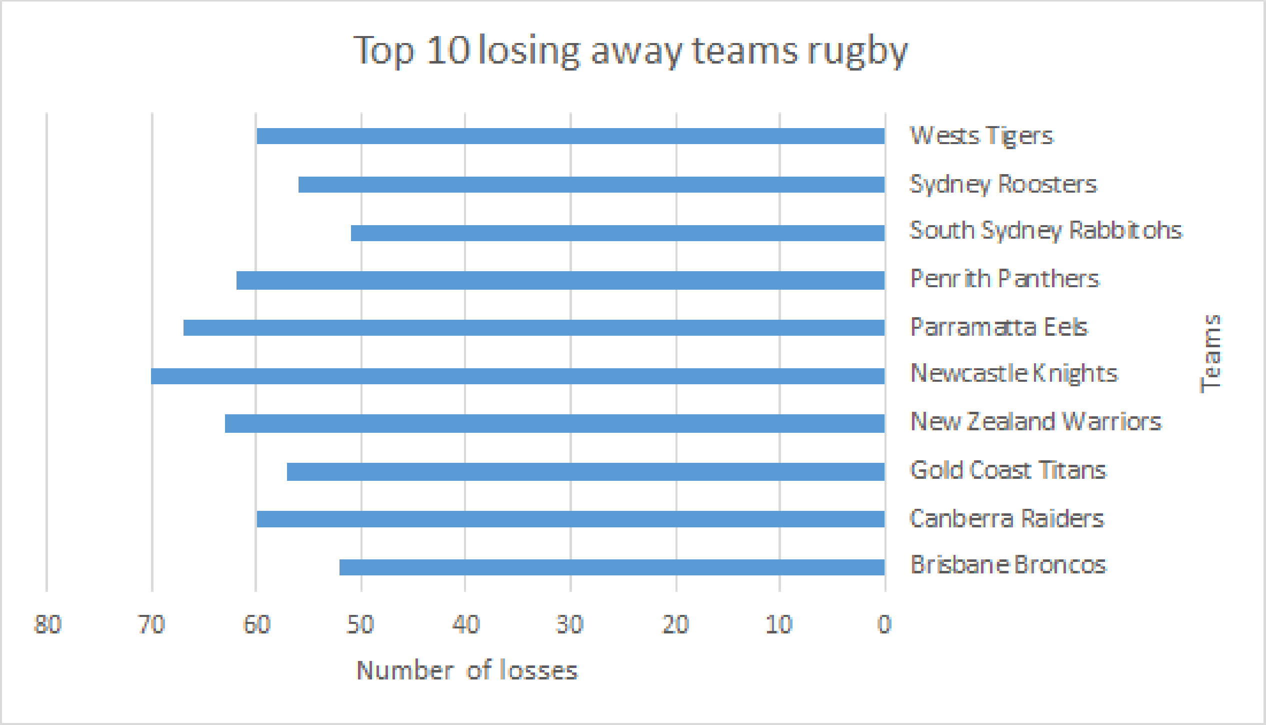

wins/total
soccer
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Barcelona
Real Madrid
Man United
Arsenal
Porto
Juventus
Paris SG

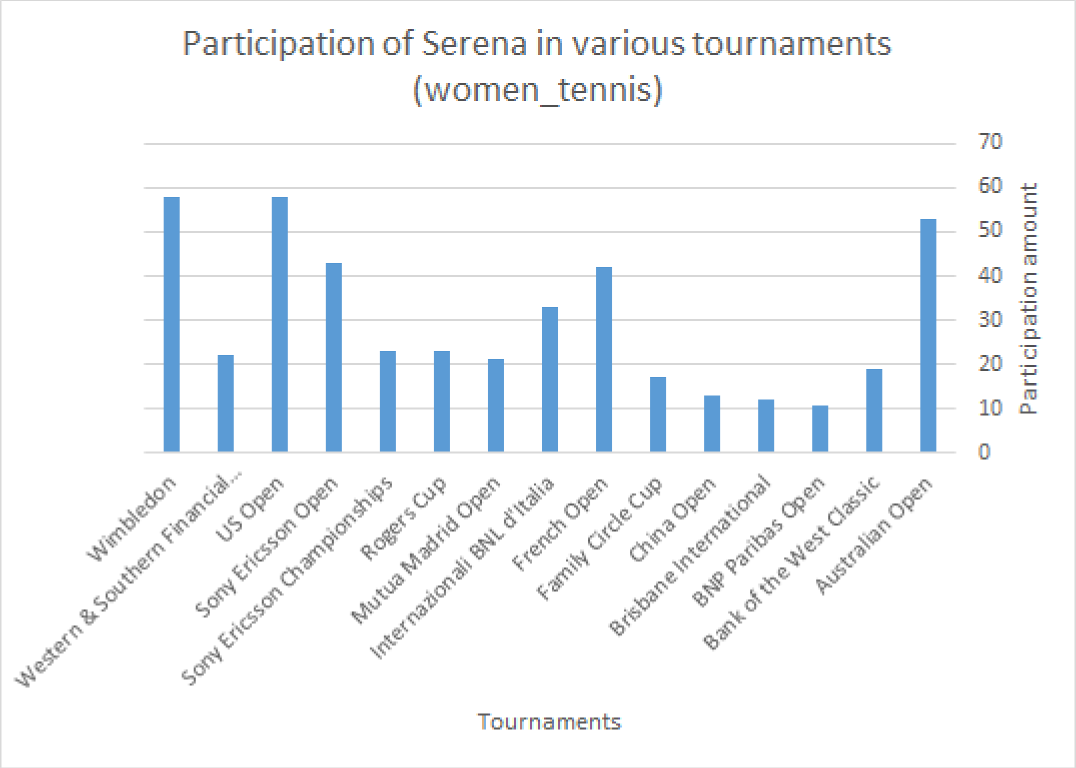

Winner distribution hockey
32%
46%
22%
■ H ■ D ■ A

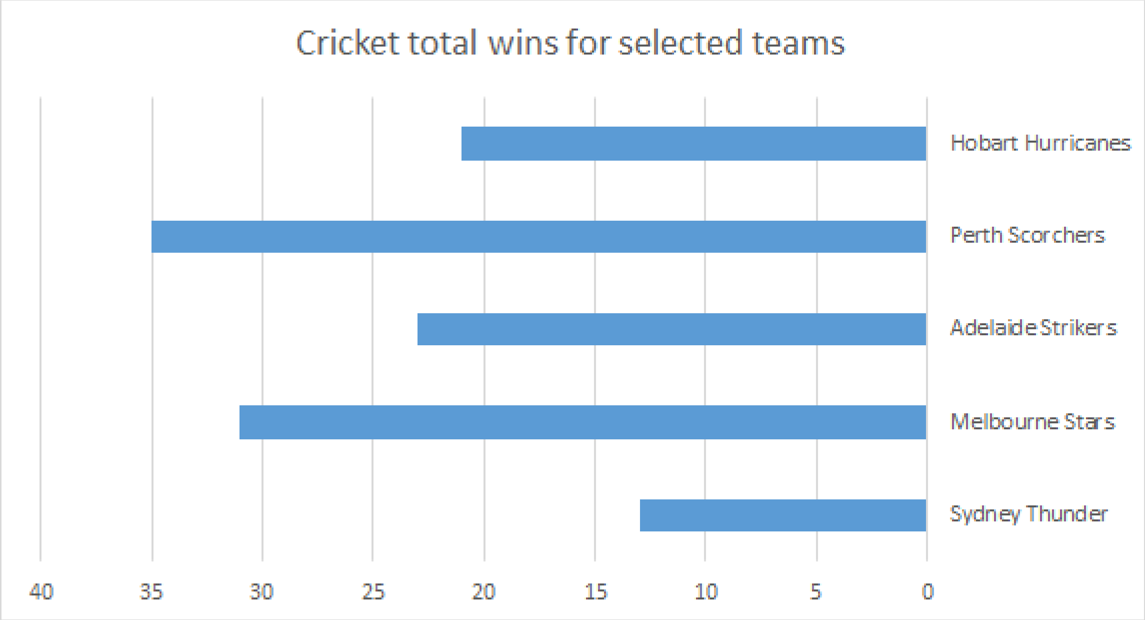

Underdog chart (wrong betting predictions)
Note: relatively low amount of samples for cricket
50%
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Percentage out of total per sport
tennis_women
tennis_men
rugby
hockey
cricket
basketball
australian_football
american_football
soccer
Sport

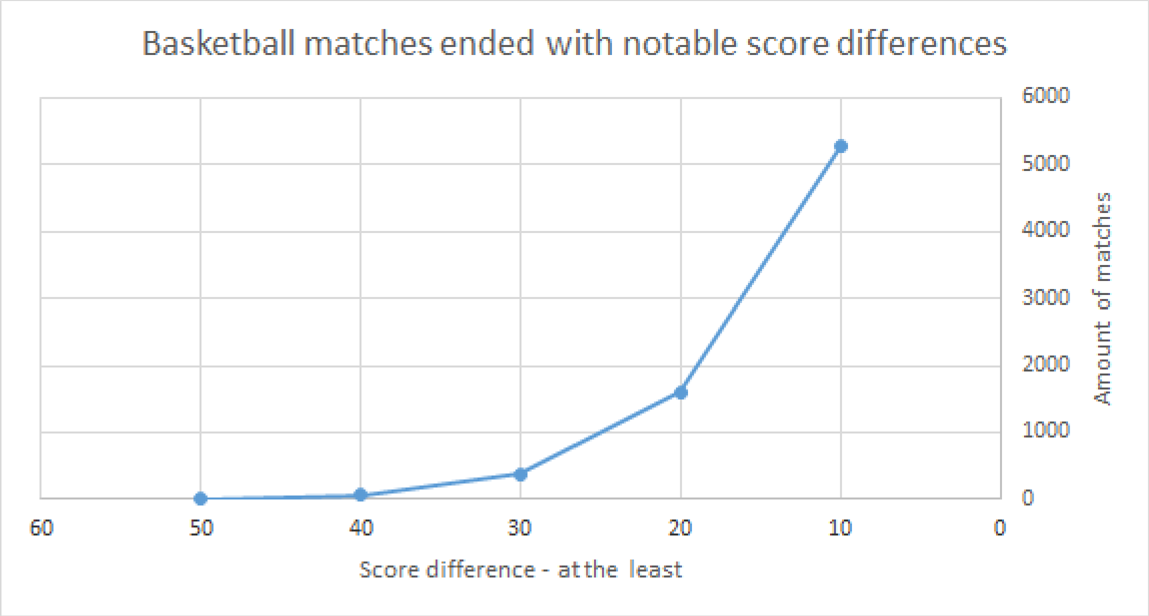

Total wins as underdog
aus_football
40
30
20
10
0
Total wins
Western Bulldogs
West Coast
Sydney
St Kilda
Richmond
Port Adelaide
North Melbourne
Melbourne
Hawthorn
GWS Giants
Gold Coast
Geelong
Fremantle
Essendon
Collingwood
Carlton
Brisbane
Adelaide

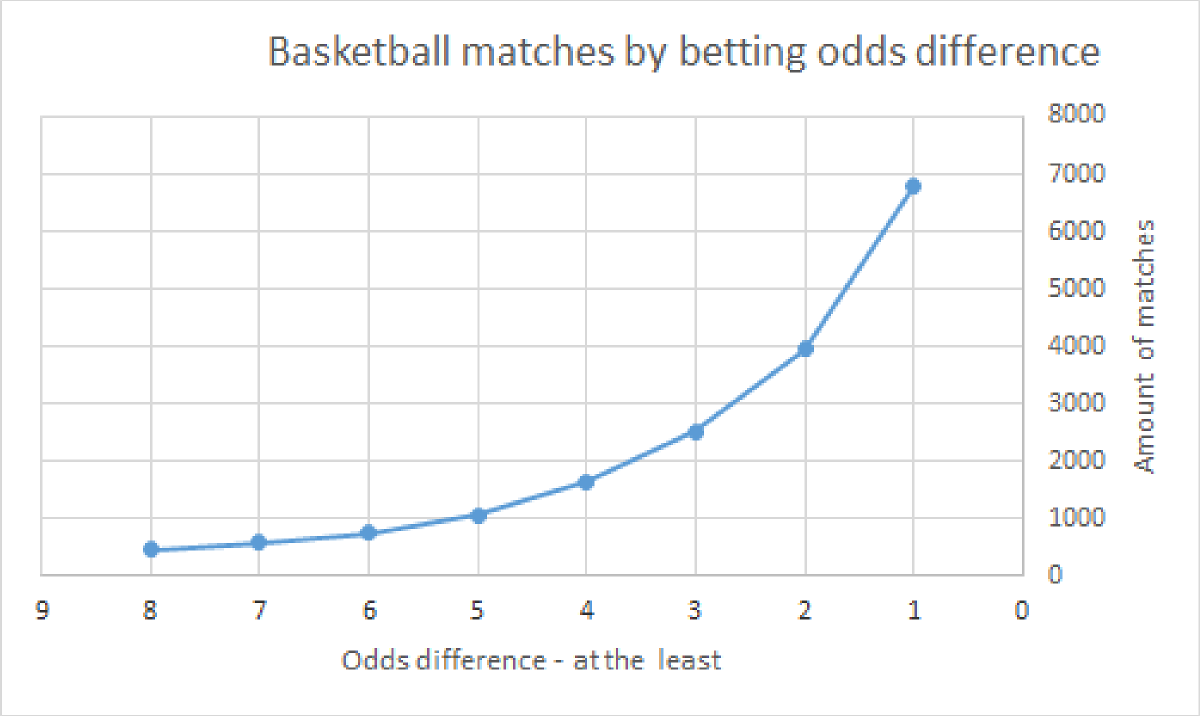

Top 10 winning teams at home in rugby
Wests Tigers
Sydney Roosters
South Sydney Rabbitohs
Penrith Panthers
New Zealand Warriors
Melbourne Storm
Crusaders
Canberra Raiders
Bulls
Brisbane Broncos
Team
90
80
70
60
50
40
30
20
10
0
Number of wins

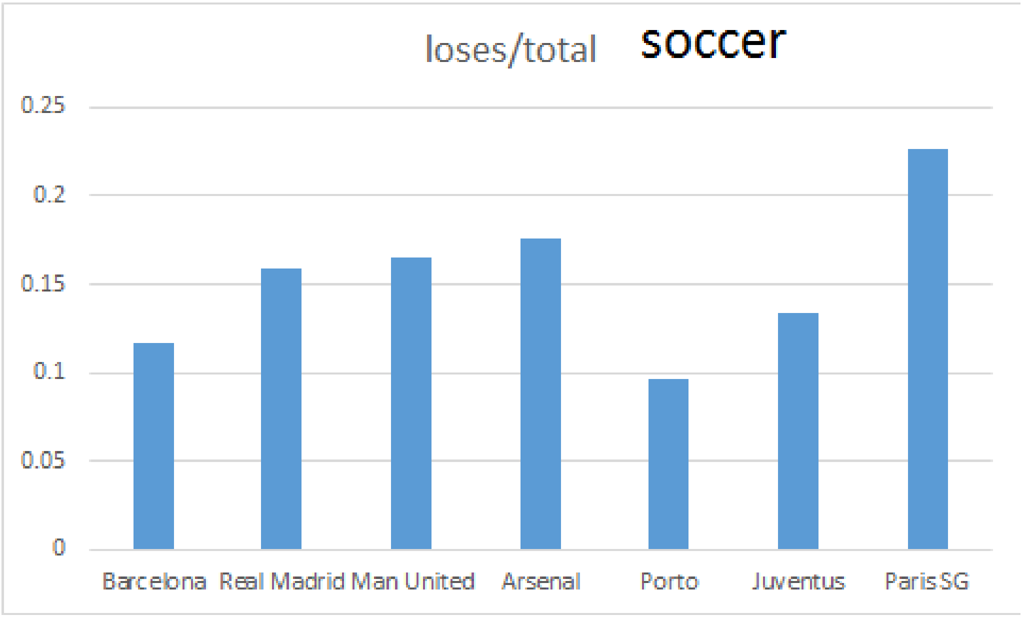

Top 10 losing away teams rugby
Wests Tigers
Sydney Roosters
South Sydney Rabbitohs
Penrith Panthers
Parramatta Eels
Newcastle Knights
New Zealand Warriors
Gold Coast Titans
Canberra Raiders
Brisbane Broncos
Teams
80
70
60
50
40
30
20
10
0
Number of losses

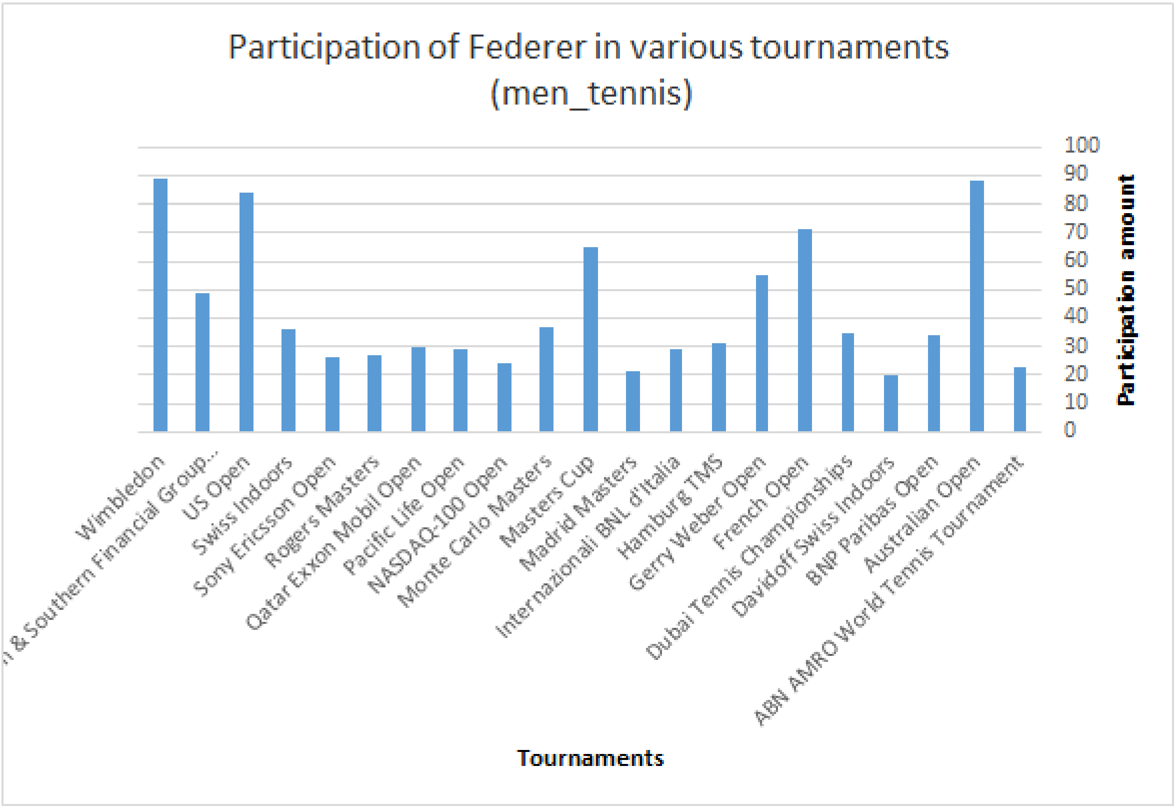

Participation of Serena in various tournaments (women_tennis)
70
60
50
40
30
20
10
0
Participation amount
Wimbledon
Western & Southern Financial...
US Open
Sony Ericsson Open
Sony Ericsson Championships
Rogers Cup
Mutua Madrid Open
Internazionali BNL d'Italia
French Open
Family Circle Cup
China Open
Brisbane International
BNP Paribas Open
Bank of the West Classic
Australian Open
Tournaments

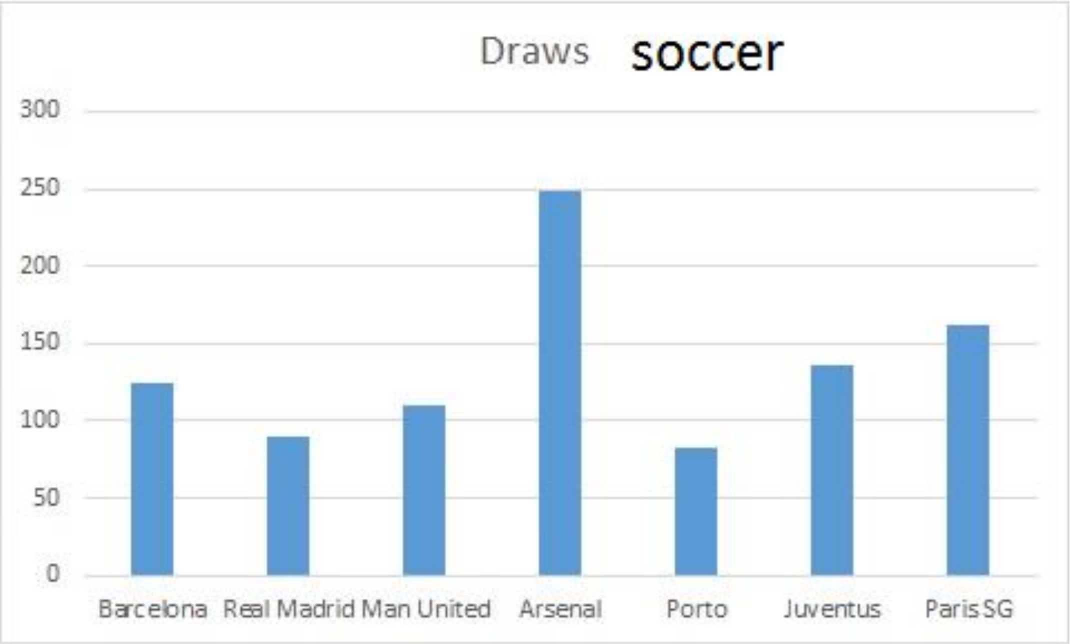

Cricket total wins for selected teams
Hobart Hurricanes
Perth Scorchers
Adelaide Strikers
Melbourne Stars
Sydney Thunder
40
35
30
25
20
15
10
5
0

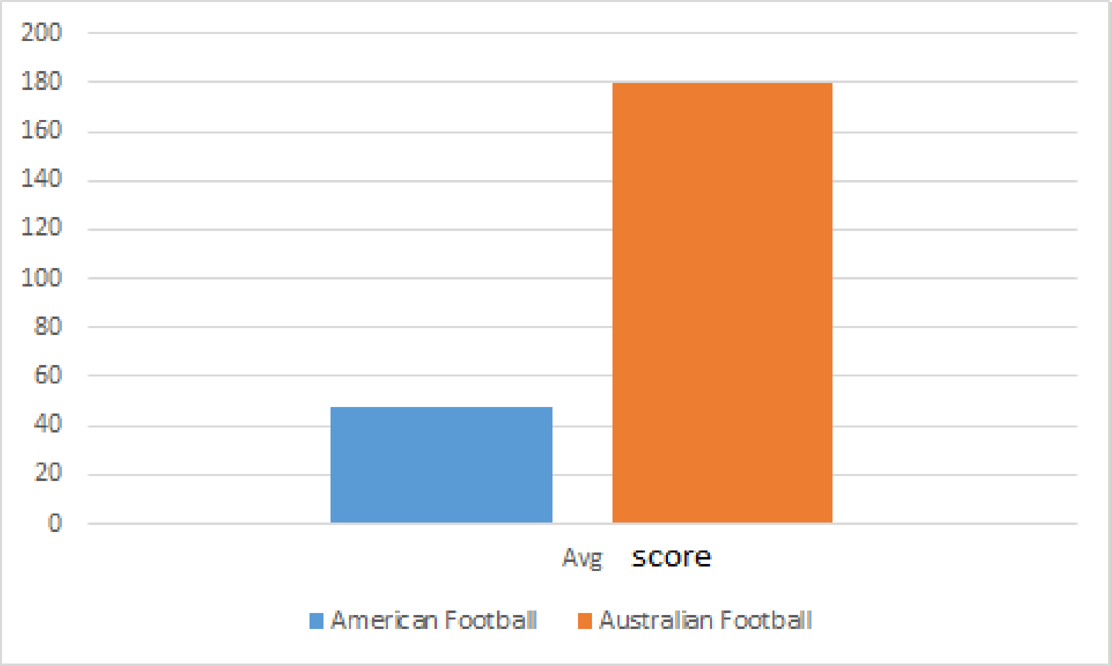

Basketball matches ended with notable score differences
6000
5000
4000
3000
2000
1000
0
Amount of matches
60
50
40
30
20
10
0
Score difference - at the least


Basketball matches by betting odds difference
8000
7000
6000
5000
4000
3000
2000
1000
0
Amount of matches
9
8
7
6
5
4
3
2
1
0
Odds difference - at the least


loses/total
soccer
0.25
0.2
0.15
0.1
0.05
0
Barcelona
Real Madrid
Man United
Arsenal
Porto
Juventus
Paris SG


Participation of Federer in various tournaments (men_tennis)
100
90
80
70
60
50
40
30
20
10
0
Participation amount
Wimbledon
n & Southern Financial Group...
US Open
Swiss Indoors
Sony Ericsson Open
Rogers Masters
Qatar Exxon Mobil Open
Pacific Life Open
NASDAQ-100 Open
Monte Carlo Masters
Masters Cup
Madrid Masters
Internazionali BNL d'Italia
Hamburg TMS
Gerry Weber Open
French Open
Dubai Tennis Championships
Davidoff Swiss Indoors
BNP Paribas Open
Australian Open
ABN AMRO World Tennis Tournament
Tournaments


Draws soccer
300
250
200
150
100
50
0
Barcelona
Real Madrid
Man United
Arsenal
Porto
Juventus
Paris SG


200
180
160
140
120
100
80
60
40
20
0
Avg score
🟦 American Football
🟧 Australian Football