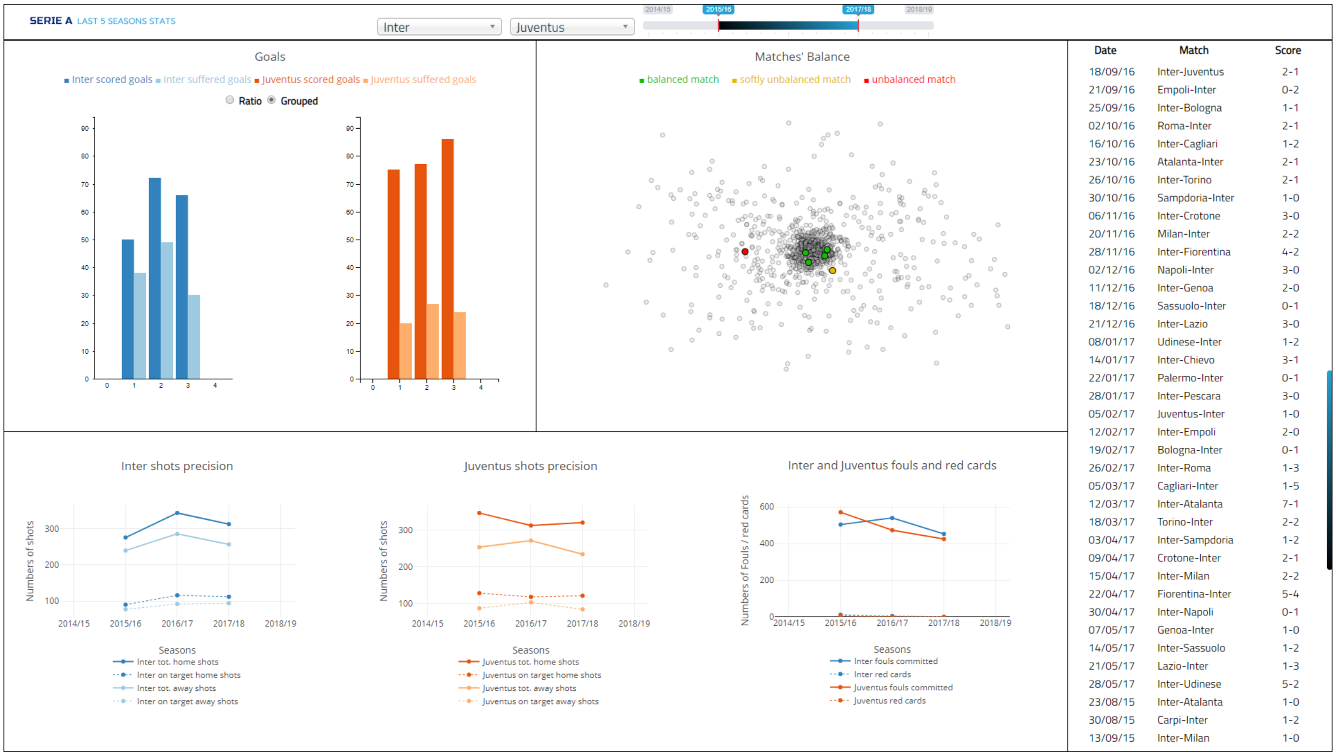
Visual Analytics project: last five Serie A seasons statistics

Engineering in Computer Science

Umberto di Canito
Daniele Buonadonna

a.y. 2019/20



Date	Match	Score
18/09/16	Inter-Juventus	2-1
21/09/16	Empoli-Inter	0-2
25/09/16	Inter-Bologna	1-1
02/10/16	Roma-Inter	2-1
16/10/16	Inter-Cagliari	1-2
23/10/16	Atalanta-Inter	2-1
26/10/16	Inter-Torino	2-1
30/10/16	Sampdoria-Inter	1-0
06/11/16	Inter-Crotone	3-0
20/11/16	Milan-Inter	2-2
28/11/16	Inter-Fiorentina	4-2
02/12/16	Napoli-Inter	3-0
11/12/16	Inter-Genoa	2-0
18/12/16	Sassuolo-Inter	0-1
21/12/16	Inter-Lazio	3-0
08/01/17	Udinese-Inter	1-2
14/01/17	Inter-Chievo	3-1
22/01/17	Palermo-Inter	0-1
28/01/17	Inter-Pescara	3-0
05/02/17	Juventus-Inter	1-0
12/02/17	Inter-Empoli	2-0
19/02/17	Bologna-Inter	0-1
26/02/17	Inter-Roma	1-3
05/03/17	Cagliari-Inter	1-5
12/03/17	Inter-Atalanta	7-1
18/03/17	Torino-Inter	2-2
03/04/17	Inter-Sampdoria	1-2
09/04/17	Crotone-Inter	2-1
15/04/17	Inter-Milan	2-2
22/04/17	Fiorentina-Inter	5-4
30/04/17	Inter-Napoli	0-1
07/05/17	Genoa-Inter	1-0
14/05/17	Inter-Sassuolo	1-2
21/05/17	Lazio-Inter	1-3
28/05/17	Inter-Udinese	5-2
23/08/15	Inter-Atalanta	1-0
30/08/15	Carpi-Inter	1-2
13/09/15	Inter-Milan	1-0

Figure 1: The full web application interface. Every view is updated when a user choose some options (Team A, Team B, Seasons)

Abstract

The main goal of this application is related to visualization of the latest five years statistic of first italian soccer league Serie A, in order to offer visual informations about datas of this competition. Moreover, a filter menu is setted in the beginning of the web page in which the user can choose one or two teams and one or more seasons (from 2014/15 to 2018/19 seasons).

1 Introduction

The world of sport today is increasingly made up of numbers and information useful in different contexts and applications. One of the sports that has always used this information, but especially in recent years is without a doubt football.

In more detail, statistics are now used in different ways and sectors in the world football scene, from the choice of players in the youth sectors, through the customization of training and the preparation of matches, up to the forecast of future events relating to direct clashes between teams in the various competitions that teams week by week are facing. Not only that, because today the data are also used by journalists, commentators, bloggers (but also simple football fans) to write articles on the matches, reports on the players and comments on the performance of the team, the players or the tactics.

So it is essential to provide the right information (or statistics) to professionals and to fans of the sport in general, in order to better address these features.

However, often the data are available on the net in a tabular, textual or in another format which makes often difficult to pull out of the analysis. For this reason, today Visual Analytics technologies meet us to be able to dynamically display data, through the use of graphic libraries and intelligent algorithms that can facilitate the use of information.

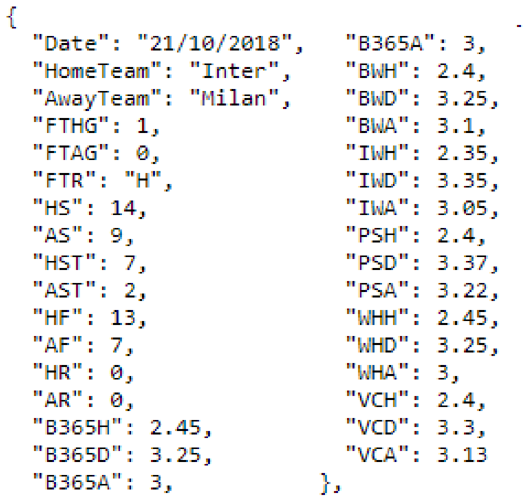
Our dataset represents the data related to goals, shots' accuracy, the relation between fouls and red cards, match results and how much they're balanced, of one or two teams.

2 Dataset

The dataset used in this project is taken originally from Datahub.io [1] and subsequently edited in order to satisfy our purpose [2]. It contains 1905 tuples[1], each with 32 attributes. What we obtained is a consistent and more manageable dataset with an AS index of 60960.

About the attributes we can distinguish between:

- categorical attributes
  - `Date`, that indicates the date of played match
  - `HomeTeam`, `AwayTeam` that indicate respectively the home and the away teams
  - `FTR` that indicates the full time result: H=Home Win, D=Draw, A=Away Win
- quantitative attributes
  - `FTHG`, `FTAG` that indicate respectively the scored goal by the teams, home and away cases.
  - `HS`, `AS`, that indicate rispectively the home and away team shots and `HST`, `AST` that indicate rispectively the home and away shots on target.
  - a list of odd quotes for home win, draw and away win from 6 different odd agencies: `B365H`, `B365D`, `B365A`, `BWH`, `BWD`, `BWA`, `IWH`, `IWD`, `IWA`, `PSH`, `PSD`, `PSA`, `WHH`, `WHD`, `WHA`, `VCH`, `VCD`, `VCA` [2]



```
{
  "Date": "21/10/2018",    "B365A": 3,
  "HomeTeam": "Inter",     "BWH": 2.4,
  "AwayTeam": "Milan",     "BWD": 3.25,
  "FTHG": 1,               "BWA": 3.1,
  "FTAG": 0,               "IWH": 2.35,
  "FTR": "H",              "IWD": 3.35,
  "HS": 14,                "IWA": 3.05,
  "AS": 9,                 "PSH": 2.4,
  "HST": 7,                "PSD": 3.37,
  "AST": 2,                "PSA": 3.22,
  "HF": 13,                "WHH": 2.45,
  "AF": 7,                 "WHD": 3.25,
  "HR": 0,                 "WHA": 3,
  "AR": 0,                 "VCH": 2.4,
  "B365H": 2.45,           "VCD": 3.3,
  "B365D": 3.25,           "VCA": 3.13
  "B365A": 3,            },
```

Figure 2: A row example of the dataset (JSON)

3 Visualization

The visualization that we have developed is composed by 5 views, each showing different informa-

[1] **381 x 5=1905**: *381* because Serie A is composed by 10 league day matches, for a total of 38 match days in total, (with 20 teams playing 2 times for season) ; *5* as we are considering the latest 5 seasons of competition (from 2014/15 to 2018/19 seasons)

[2] The 6 odds agencies are: *Bet365, Bet&Win, Interwetten, Pinnacle, William Hill, VC Bet*

tion of the dataset. All of them are coordinated with each other, so choosing some options relative to teams and seasons on the menu bar in the top of page will update all views.

At the beginning, on the launch of the web app, the views concerning the goals, accuracy of shots, relation between fouls/goals and match results do not show any information, as initially no team has been selected. With regard to the view concerning the balance of the matches, all the matches of the seasons 2015/16, 2016/17 and 2017/18 are visible, as the slider in the menu concerning the choice of the seasons is set by default in that range.

The user has the ability to choose some options, available in the menu present on the top of the page. In detail, he can choose one or two teams, using the two dropdown menus and also one or more seasons from 2014/15 to 2018/19, using the slider. After these choices all the views are updated accordingly.

About the design pattern, we choose to make a white-based layout to make visible as better as possible the details of the differents charts. Not only, because we choose to use the same colors for representing all the information of the relative selected team. Better explaining with an example, after choosing the `Team A`, all the lines/traces of multiline charts (precision of shots and relationship between fouls/red cards charts) and bars of the bar chart (goals chart) have the same color o chromatic scale (this last as the case of the shots on the target, present in the precision of shots view). The choose of colors was made with the ColorBrewer tool [9]

3.1 Grouped/Stacked Barchart: scored and suffered goals

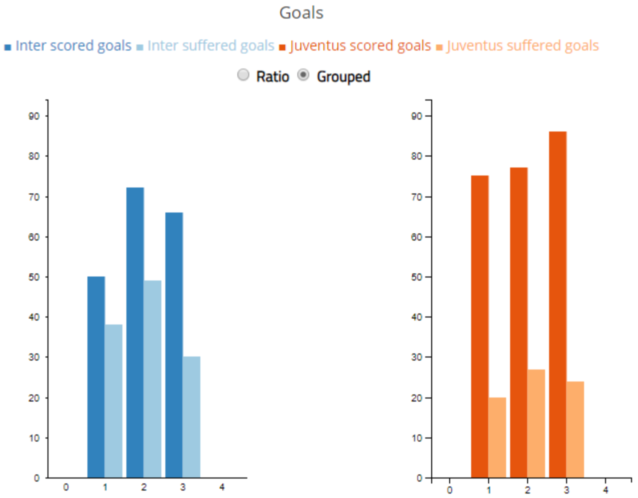


Figure 3: Grouped/Stacked Barchart in which are represented the scored and suffered goals of selected team/s

In this view we wanted to analyze the number of goals, scored and suffered, of the selected teams. In particular, the chart consists of two subcharts, one for `Team A` and the other for `Team B`. In addition, for each chart are shown the number of goals scored and suffered per season (season that we remind to be chosen in the bar menu at the top by the user).

As for the technical details of the visualization, both subcharts have on the x-axis the seasons, from 2014/15 to 2018/19, and on the y-axis the numerical value of the goals, and were made with javascript library `D3.js` [3]. Not only that, as the user can choose to visualize both charts in two different ways, using a form containing three radio buttons:

- `Grouped`, in which two different bars are shown side by side for each season: one for scored goals and the other for suffered goals.
- `Stacked`, in which for each season is shown a single bar but divided logically and visually into two by two different colors. This is useful in order to show the ratio between goals scored and suffered per team.

In reality, the stacked version shown a ratio view, in which the values along the y-axis are represented in relation of the two different type of goals category.

3.2 Scatterplot: matches balance

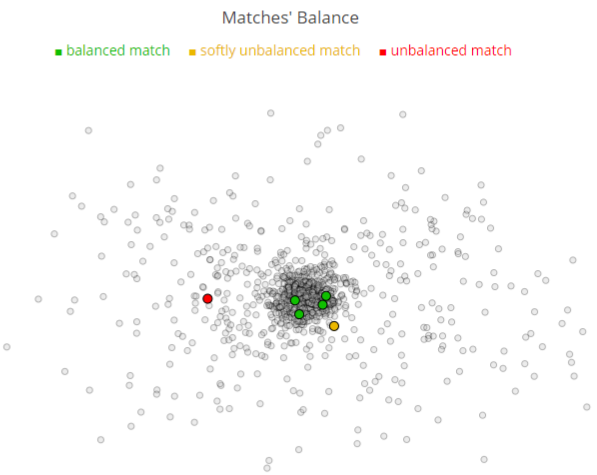


Figure 4: Scatterplot in which are represented the matches balance of selected team/s

In this view we wanted to analyze the matches balance between teams, in the selected seasons. To obtain this type of result we used a chart of the scatter type, which shows a density of points along two dimensions, represented by the axes y and x. It was made with javascript library `D3.js` [3]

Each point represents a match between two teams. Initially, at the start of the web app, the chart shows all matches from the season 2015/16 to 2017/18, set by default in the slider at the top. Then the user will be free to choose one or both teams in the menu, so can restrict the range of the visualization, focusing only on that/those particular teams.

The chart visualizes three different categories in which, as shown by the legend, are differentiated between balanced, softly unbalanced and unbalanced matches, through the use of three different colors of the points represented.

How the data are computed? As previously said, for each match we know the odds of six different betting agency. We can look for the greatest value and the lowest value, then compute the difference. The current value is, indeed, an approximate estimation of how much a match is balanced. In fact, if two teams with the same chances to win play, the agency will give them similar odds, making the difference computed very low. On the contrary, if two teams play and one of them is clearly the favorite, the odd difference will grow.
So, computed these estimation we need a dissimilarity matrix to give to the Multidimensional scaling (MDS) algorithm in order to work and classify the matches. The dissimilarity between two matches is again computed as the difference between the estimations of the matches. The output of the MDS algorithm do not show two or more well define clusters, although, as clearly seen in previous figure. But, what it is useful to notice is that more a match is unbalanced more is far from the origin of the axes.

3.3 Multiline charts: precision of shots and relation between fouls/red cards

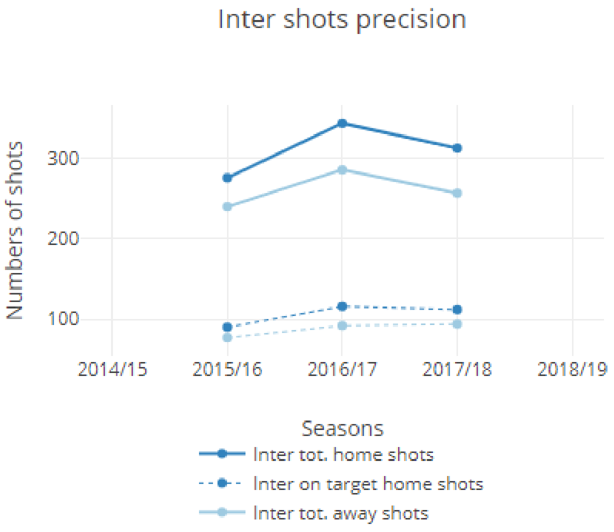


Figure 5: Multiline chart in which is represented the shots precision of selected team/s

The first view of the multiline chart type we are introducing is intended to represent the accuracy of the shots of one or both selected teams.

In reality, as before, it also consists of two sub-charts, one for `Team A` and one for `Team B`. In each of them you can see the accuracy of the shots, understood as the ratio between the total shots and those on target, both at home and away from home, in the selected range of seasons.

This type of view allows to see the general trend during the seasons in the shots of both teams, thanks to the use of lines in fact that allow the immediate fruition of data. In addition the user, through a clickable legend, can decide to filter the visualization of only some of the available traces, making the data even more analyzable. When he choose to click on these traces, the chart will fit in a way to better see the remaining lines.

Finally, each user can see the exact number of shots (both total and on target and both at home and abroad) by positioning himself on the points corresponding to the seasons, thanks to the use of a tooltip that will show this information.

About the second multiline chart:

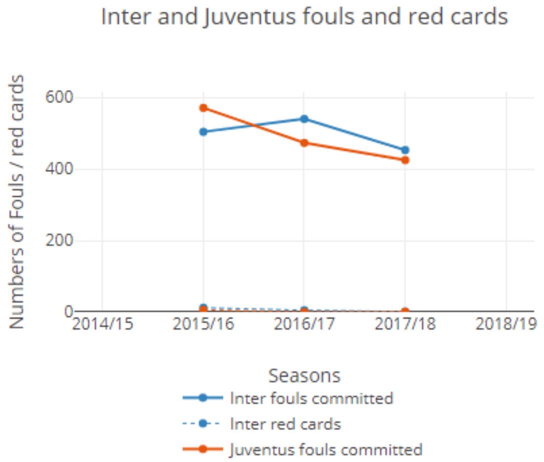
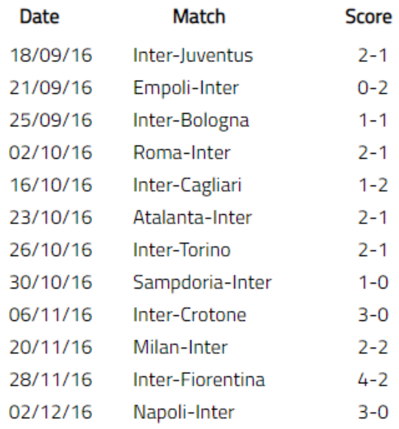


Figure 6: Multiline chart in which is represented the relation between fouls and red cards of selected team/s

It aims to represent the relationship between the fouls committed and the red cards received of one or both of the selected teams.

However, in this case there is no subdivision into subcharts, as it is a single chart but, as before, it maintains the same graphical features. In fact, also in this case the user can see the tracks, can filter only some of the four traces and see the exact number of fouls committed and/or red cards received.

3.4 Table: matches results



Date	Match	Score
18/09/16	Inter-Juventus	2-1
21/09/16	Empoli-Inter	0-2
25/09/16	Inter-Bologna	1-1
02/10/16	Roma-Inter	2-1
16/10/16	Inter-Cagliari	1-2
23/10/16	Atalanta-Inter	2-1
26/10/16	Inter-Torino	2-1
30/10/16	Sampdoria-Inter	1-0
06/11/16	Inter-Crotone	3-0
20/11/16	Milan-Inter	2-2
28/11/16	Inter-Fiorentina	4-2
02/12/16	Napoli-Inter	3-0

Figure 7: Table in which are represented the results of all matches of a team or the matches results between two selected teams

In this view, the aim was to visualize, for one or both selected teams, the results of the matches during the selected seasons. In more detail, if user select only one team will be visualized all the matches of that team. If the user select both teams will be visualized all the details about the direct matches between that two teams.

In order to do this, we have created a table with three columns: `Date` (i.e. the date of the matches), `Match` (i.e. the teams that have collided, respecting the order of home and away) and `Score` (i.e. the result of matches).

In addition, the table is strictly related to the view inherent to the balance of matches, because when the user is positioned on the single match in that graph, the row inherent to that match is automatically selected in the table. In addition, the automatically selected line is underlined respecting the color inherent in the classification of the match (balanced, softly unbalanced and unbalanced).

4 Analytics

The dataset[1] taken from DataHub is been considerably reduced of not interesting data with respect to our goal. The output is been stored in a JSON file on an online repository[2], accessible through the internet. The data from this repository is loaded and locally stored, in order to be used to retrieve and compute on the fly everything is needed to be visualize. The only exception is been the MDS related visualization, where all the points are been calculated using a python script: the coordinates are been stored in another online repository[6]. This is due to the fact that MDS computation will take time and so it is not feasible to make it on the fly. Also, since input data do not change, it has no meaning to compute each time MDS algorithm.

All other computations, like how many goals the team have scored and suffered, how many shots were done in the target area etc. are done dynamically, respecting the changing of the input of the user, both about teams and seasons selections.

5 Conclusions

What we have done in this project can be a useful tool for anyone who wants to read the data of the last five years of the Serie A in an easy way from the point of view of the acquisition of information, through the use of graphs that facilitate its use.

In particular, it can be useful to make predictions about future matches of the new season of the current championship, trying to predict likely results or outcomes of the matches based on past results.

References

[1] Dataset (original version): https://datahub.io/sports-data/italian-serie-a

[2] Dataset (edit version, repositary on JSONSTORAGE.NET): https://jsonstorage.net/api/items/881a4adb-14b5-47b2-907f-f5ca2f0a1366

[3] D3.js (API): https://github.com/d3/d3/blob/master/API.md

[4] Plot.ly (API): https://plot.ly/javascript/

[5] MDS Algorithm: https://en.wikipedia.org/wiki/Multidimensional_scaling

[6] Matches Coordinates for MDS (JSONSTORAGE.NET): https://jsonstorage.net/api/items/e3c00004-c5da-449a-b7de-4ca7a63e845f

[7] Ion.RangeSlider (API): http://ionden.com/a/plugins/ion.rangeSlider/api.html

[8] Chosen (API): https://harvesthq.github.io/chosen/

[9] ColorBrewer: http://colorbrewer2.org/