SI 618 Project 1 Report: FIFA Ranking and Club Transfer

Yirui Gao
October 21, 2019

1 Motivation

This project faces to the football world and leagues from different countries. In professional football leagues, transfers happen very commonly, a leaving or coming of a superstar means a loss or hope to both teams or even leagues. Among leagues all over the world, famous leagues, like EPL from England, La Liga from Spain, Serie A from Italy and Bundesliga from Germany, usually conduct the soccer market. At the same time, famous leagues might bring prosperity to its national football. The most standard ranking for national teams, the FIFA ranking, tells how each country competes with others, with the order varying year by year.

As a football fan, I have been always wondering if there is any relationship between the FIFA ranking and the transfers between leagues. Will the coming of a superstar from a league in another country promote the competitiveness of the local national team? Or does the leaving of a superstar indicate this country is having a saturation of football stars and thus a rather competitive national team? There might be interesting findings behind the ranking and the transfer statistics.

In my project, I surrounded and explored the following three instructional questions:

- Is there any relationship between the net transfer expense (overall transfer expense - overall transfer income) and the FIFA ranking for each country (involving all leagues in this country) per season?
- Does any evidence show that a country with more transfer records tends to have a stronger national team?
- Is the introduction of more Attackers or more Defenders more helpful to promote the national team?

2 Datasets

In this project, I used two datasets from Kaggle. The first one is about the top 250 transfer in each year between 2000 to 2018, while the other one shows the FIFA ranking records from the year 1993 to 2018. The two datasets are both given in csv format.

2.1 Yearly Top 250 Football transfers from 2000 to 2018

The URL link of this dataset is https://www.kaggle.com/vardan95ghazaryan/top-250-football-transfers-from-2000-to-2018.

This dataset contains the yearly top 250 most expensive football transfers from the season 2000-2001 until 2018-2019. There are 4700 transfers documented inside and 10 columns in this dataset. The columns include the following information:

- Name and position of the player (*DATA type*: String).
- Selling and buying team and league (*DATA type*: Double).
- Estimated and actual market value for that transfer (*DATA type*: Double).

2.2 FIFA Soccer Rankings from 1993 to 2018

The URL link of this dataset is https://www.kaggle.com/tadhgfitzgerald/fifa-international-soccer-mens-ranking-1993now.

This dataset gives the rankings of men's football national team over 200 countries from the year 1993 to 2018. Up to 16 columns are displayed, with the main following information that might be useful:

- Ranking and country name for that particular year (*DATA type*: String).
- Total point for ranking in that year and previous point (*DATA type*: String).
- Ranking change compared to the previous year (*DATA type*: String).
- Rank date (*DATA type*: String).

3 Data Manipulation

In this section, I will display how I managed to work on this project by breaking it into pieces. A general overview of the process is given as Fig.1. There are five sub-stages for this project, with each, I assigned specific tasks and relied on some tools. Multiple tools are used in this project and they contribute mostly.

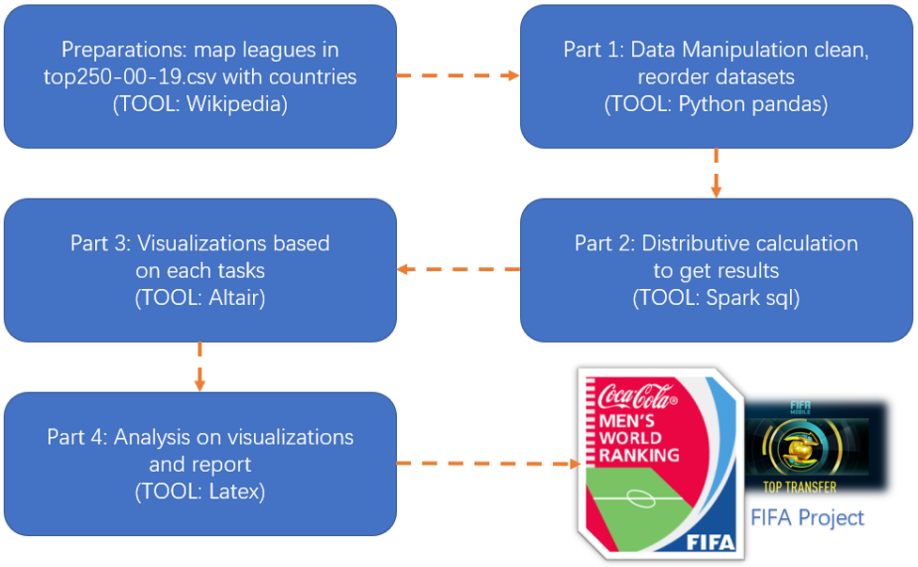

Figure 1: Project workflow.

3.1 Preparation: map leagues to countries and positions to categories

Since the transfer dataset gives the league without specifying the country, to relate this dataset to the ranking, having columns indicating the countries is necessary. This work was manually done by myself, where I used Wikipedia as my tool and found out the corresponding country for about 100 leagues. I put the league-country pairs into a new csv file named as **"league-country.csv"**.

For the third question, I would like to specify the attacker or defender category for each player. Since the transfer dataset gives only the position of the player, I manually again mapped the positions to either attacker or defender by categorizing backs, defend midfield and wings to be defenders, and the rest to the attackers.

3.2 Part 1: Use pandas to manipulate the dataset

Before feeding the overall datasets into the calculation part, I need to re-organize the datasets and make them more friendly and easier to be applied in big data calculation. There were 4 tasks that I finished in this process, with each I threw my attention on details and how I could just keep the number of columns as less as possible.

I first formed the league-country dictionary by using pandas' **to_dict()** method for later use. Then, having noticed that the string format in some columns of the transfer dataset is not so pretty with an extra blank space at the beginning, I deleted the extra blank space. For the coming task, I added the country columns into the transfer dataset and drop some unused columns like Age and Market_value. A new data frame named as **"transfer.csv"** replaced the original transfer dataset for later use.

Similarly, for the ranking dataset, I filtered out those records before the year 2001 to better match the transfer dataset and also transformed the data types of rank and rank_change from strings to integers. One difference was that the exact date for ranking and the season shown in the transfer dataset. To make them match-able and feasible to merge later, I transformed the rank date to the season by putting the rank for the whole year to the season that starts from the previous year and ends at that year. For example, for the rankings in 2005, I grouped them into the season "2004-2005". This is somewhat reasonable since the influence of transfers cannot affect in a very short time. Then a new dataset was formed called **"ranking.csv"**.

3.3 Part 2: Spark SQL to merge the datasets

Since due to the dataset reason, an operation on the sub-analysis first before a merge of two datasets would be easier, and the overall analysis was based on the sub-calculation steps. In this subsection, I will mainly discuss how my source code worked, and how I managed to use spark sql to join the datasets. The major part of the analysis will be left to the next section.

Using the Spark SQL tool for distributive calculation in **"calculations.py"**, my goal was to calculate the net transfer expense for each country involving all the leagues inside this country per season, and also count the number of transfer. As for the ranking, the average ranking for each year was to be calculated for each country. Then information would be merged as a new data frame. The final step is to calculate the attacker-defender ratio among the players coming for each country at each season, combined with the ranking, resulting in another new data frame.

A join operation based on the country and the season was performed to join the transfer table and the ranking table together, including the analysis on calculating the net transfer fee beforehand, which details will be shown in the next section. Since countries were not perfectly matched for these two tables, I used **COALESCE(buy.Count, 0)** to return a value 0 if there was a null value. I output the targeted table and named it as **"transfer_ranking.csv"**. A sample of this table can be seen in Fig.2.

```
>>> all_transfer_ranking.show()
+-----------+---------+----------+---------------+---------+----------------+------------+------------+-----------+------------------+
|    country|   season|sell_count|total_Tr_income|buy_count|total_Tr_expense|net_transfer|transfer_num|avg_ranking|avg_ranking_change|
+-----------+---------+----------+---------------+---------+----------------+------------+------------+-----------+------------------+
|      Italy|2000-2001|        10|       5.175E7|       27|       2.9799E8|    2.4624E8|          37|        5.0|               0.0|
|    England|2000-2001|        10|      1.3353E8|       34|       2.4313E8|     1.096E8|          44|       13.0|               1.0|
|   Scotland|2000-2001|         3|       1.608E7|       10|        6.913E7|     5.305E7|          13|       36.0|              -2.0|
|    Germany|2000-2001|         4|         3.0E7|       13|        6.878E7|     3.878E7|          17|       10.0|               0.0|
|     France|2000-2001|        13|      1.1834E8|       17|       1.435E8|     2.516E7|          30|        1.0|               0.0|
|     Turkey|2000-2001|         5|        2.55E7|        7|        4.305E7|     1.755E7|          12|       30.0|               1.0|
|     Greece|2000-2001|         1|      8000000.0|        4|        1.415E7|   6150000.0|           5|       52.0|              -1.0|
|Switzerland|2000-2001|         5|       1.953E7|        1|       4500000.0|    -1.503E7|           6|       60.0|               0.0|
|   Portugal|2000-2001|         6|        5.41E7|        4|         2.43E7|     -2.98E7|          10|        5.0|               0.0|
|Netherlands|2000-2001|         8|        5.23E7|        4|        2.13E7|      -3.1E7|          12|        9.0|               0.0|
|     Brazil|2000-2001|         7|       6.505E7|        1|       6250000.0|     -5.88E7|           8|        2.0|               0.0|
|      Spain|2000-2001|        17|      2.1625E8|       14|       1.3875E8|     -7.75E7|          31|        7.0|               0.0|
|  Argentina|2000-2001|        14|       1.185E8|        0|             0.0|    -1.185E8|          14|        3.0|               0.0|
|    England|2001-2002|        10|       7.318E7|       45|       3.5651E8|    2.8333E8|          55|        9.0|               0.0|
|      Spain|2001-2002|        11|      1.3732E8|       17|       2.2458E8|     8.726E7|          28|        5.0|               0.0|
|    Germany|2001-2002|         4|         3.1E7|       15|        9.652E7|     6.552E7|          19|        8.0|               1.0|
|     Greece|2001-2002|         1|      2500000.0|        4|        1.17E7|   9200000.0|           5|       52.0|               1.0|
|   Portugal|2001-2002|         5|       2.228E7|        4|        2.96E7|   7320000.0|           9|        8.0|              -1.0|
|   Scotland|2001-2002|         6|        3.43E7|        4|       2.625E7|  -8050000.0|          10|       56.0|              -1.0|
|Netherlands|2001-2002|         5|       4.833E7|        6|        2.57E7|    -2.263E7|          11|       10.0|               0.0|
+-----------+---------+----------+---------------+---------+----------------+------------+------------+-----------+------------------+
only showing top 20 rows
```

Figure 2: Transfer-ranking table.

I also adopted join operation to join the transfer table with the ranking data frame based on the country and the season as well. A calculated ratio for attacker and defender was included in the merged table, seen in Fig.3. Similarly, I output the targeted table with the name **"type_transfer_ranking.csv"**.

```
>>> type_transfer_ranking.show()
+-----------+---------+-----------+------------------+-------------------+
|    country|   season|avg_ranking|avg_ranking_change|Attack_Defend_ratio|
+-----------+---------+-----------+------------------+-------------------+
|Netherlands|2000-2001|        9.0|               0.0|                3.0|
|     Turkey|2000-2001|       30.0|               1.0|                2.5|
|     France|2000-2001|        1.0|               0.0|  1.8333333333333333|
|      Italy|2000-2001|        5.0|               0.0|  1.4545454545454546|
|      Spain|2000-2001|        7.0|               0.0|  1.3333333333333333|
|   Scotland|2000-2001|       36.0|              -2.0|                1.0|
|    England|2000-2001|       13.0|               1.0| 0.8888888888888888|
|    Germany|2000-2001|       10.0|               0.0|              0.625|
|Netherlands|2001-2002|       10.0|               0.0|                5.0|
|    Germany|2001-2002|        8.0|               1.0|                4.0|
|      Spain|2001-2002|        5.0|               0.0|  1.8333333333333333|
|      Italy|2001-2002|        8.0|              -1.0|               1.25|
|   Portugal|2001-2002|        8.0|              -1.0|                1.0|
|   Scotland|2001-2002|       56.0|              -1.0|                1.0|
|    Austria|2001-2002|       57.0|               0.0|                1.0|
|    England|2001-2002|        9.0|               0.0|  0.7307692307692307|
|     France|2001-2002|        2.0|               0.0|  0.5555555555555556|
|     Greece|2001-2002|       52.0|               1.0|  0.3333333333333333|
|      Italy|2002-2003|       12.0|               0.0|  2.3333333333333335|
|  Argentina|2002-2003|        5.0|               0.0|                2.0|
+-----------+---------+-----------+------------------+-------------------+
only showing top 20 rows
```

Figure 3: Transfer-ranking table with attacker-defender ratio.

3.4 Part 3: Visualization by altair

After using **write.csv()** method to write the output tables into csv files, I transmitted them back to my login node. I used these two new targeted data frames and loaded them into Python notebook again. Three visualizations were made to demonstrate the relationship between national team ranking and net transfer fee, transfer number, attacker-defender ratio, correspondingly. The complete three visualizations can be seen in the next section. Since there are several countries with very few records to visualize, I selected 12 of them which had relatively more data to show. The 12 countries include Argentina, Belgium, Brazil, England, France, Germany, Italy, Netherlands, Portugal, Russia, Spain and Turkey.

4 Analysis and Visualization

Analysis of the instructional questions raised at the beginning was performed step by step using Spark SQL in my **"calculations.py"** file. For each task, multiple operations were carried on to calculate and group by all the information together.

4.1 Relationship between the net transfer fee and national ranking

4.1.1 Calculation and analysis

For the first question, I intended to calculate the net transfer fee for each country at each season. The very beginning job I did was to distinguish those records between selling and buying. Thus, I first created two SQL tables for selling and buying of each country in seasons using SQL language **"group by country, season"**. I counted the number of selling or buying, and what's more, the most important part as the sum up in groups of transfer fees. Specially, in order to avoid the cases that one player transfer from a team to another team in the same country (which cannot show any influence in this topic to the national layer), I used where clause **"where Country_from != Country_to"** to filter up this case. I also sorted the two tables based on the season from the earliest and the total transfer fee (selling or buying) in the descending order.

Next comes to the calculation on the net transfer fee. Having selling and buying tables, I used a **left join** on the same country in the same season to combine these two tables. Then the net transfer fee for each country at each season was derived using the total expense minus the total income. Since I noticed that the selling record was more than the buying ones, to avoid a null value as the calculation result, I use **COALESCE** method mentioned above again, **"(COALESCE(buy.total_Tr_expense, 0) - total_Tr_income) as net_transfer"**. The results indicated that a more positive net transfer fee showed this country spent very large amount of money on introducing top players from other leagues, thus being "buying-oriented". To the contrary, a more negative value showed this country earned a lot of money by selling top players out to other leagues, thus being "selling-oriented".

Finally I merged the net transfer table with the ranking table, which has been talked about in the merging section.

4.1.2 Visualization

The direct way to visualize the relationship is to plot the two parameters in the same plot. Using altair with the 12 selected countries, I made x axis to be the season and dual y axes to coexist the two sub-plots. The bar chart with the axis on the left is the net transfer fee, where the downward direction means "selling-oriented" and the upward direction shows "buying oriented". The line chart with the axis on the right indicates the ranking. The visualization is in Fig.4.

4.1.3 Findings and conclusion

There seems no unified pattern for the relationship between the ranking and either "buying-oriented" or "selling-oriented". But to observe the trend for the ranking, combined with the current mainstream football market in Europe, there are several interesting findings:

- For the two giants in South America, Brazil and Argentina, since they are far way from the mainstream market in European, it shows that when they are less selling-oriented, the ranking will go up. This is because talented youths from these two countries are usually attracted by the big clubs in Europe. So when these two countries can have more talents staying in their leagues, the benefit to the national teams would be large.
- For large European markets, rich clubs from England enable the country to be nearly always "buying-oriented". But the evidence shows that a large number of buying transfer into England leagues is not beneficial to the national team. A reduction on absorbing superstars and give more opportunities to the native players seems to be more helpful for such rich leagues.

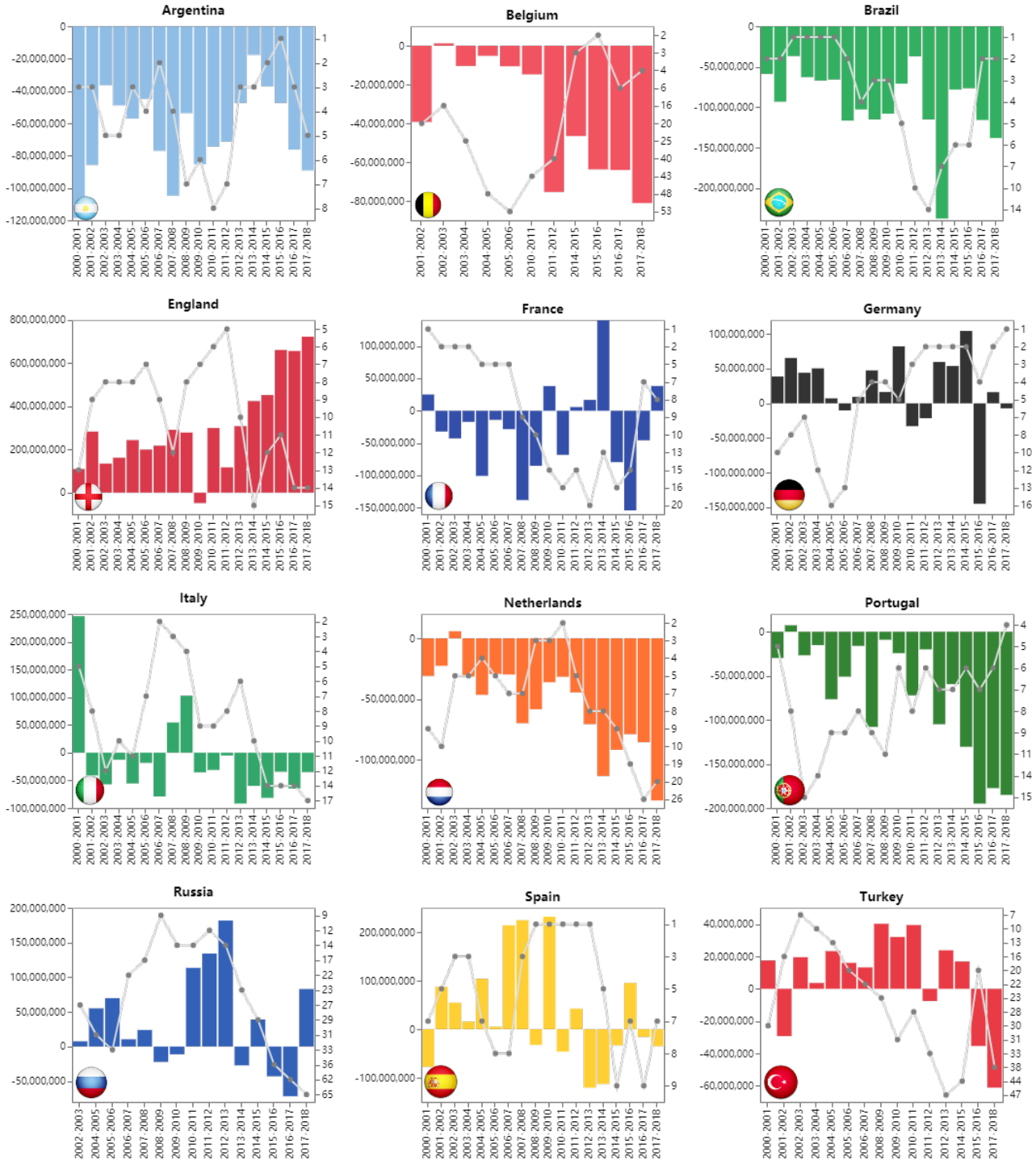

Figure 4: Relationship between the net transfer fee and ranking.

- For European giants, like Germany, Spain and Italy, they are not constantly tending to be ”selling-oriented” or ”buying oriented”, and the competitiveness of their national teams are quite stable. For these countries, the nation team's performance is not strongly affected by the oriented to buying or selling.
- For some European countries that have many talented players but their leagues are not so famous, like Belgium and Portugal, a way to enhance their national team is to send out these talented players to other high-level country leagues like England or Spain, to help improve their skills.
- But for those countries with not many talented players and not famous leagues, like Russia and Turkey, they need to spend more money to introduce more top players into their leagues, and thus enhance the competitiveness of their leagues and also the native players.

To conclude, the net transfer fee and the ranking have relationships but different in groups. For those football giants away from European, they may want more top players to stay; for rich European leagues in some country, they might want to give the native players more chances to play; for giants in European, there's not large for transfer on their ranking; for other countries, talented players should be placed at high-level leagues to help the national teams.

4.2 Relationship between th transfer number and ranking

4.2.1 Calculation and analysis

For the second question, it also digs out the relationship between transfer and ranking, but the main focus has been changed to the number of transfers, including selling records and buying records. To calculate this column, I simply added an addition operation when I merged the transfer table with the ranking table, in the code **”(transfer.sell_count + transfer.buy_count)”**. A larger number for the transfer number means a more active status to the market of top players in this country, in other words, top players come and go in this country; a lower number of transfer number means the country is unable to bring up or attract top players.

4.2.2 Visualization

Using altair again with the 12 selected countries, I kept the line chart for the ranking and change the bar chart to area chart to show the number of transfer. X axis is still the season and the area chart with the axis on the left is the transfer number. The visualization is in Fig.5.

4.2.3 Findings and conclusion

From the visualization, it is not hard to observe that a larger transfer number often leads to a more competitive national team, especially for European countries. Giants like Germany, England, Italy, France and Spain, have more top players transfer records and thus a high FIFA ranking. Other European countries with low-active market for top players like Russia and Turkey, their ranking is relatively low.

Usually the more active for the top player markets in that country, the ranking will go up. And a lower active market will lead to the decrease of the ranking. Several countries follow this rule like Belgium, France, Germany, Netherlands, Portugal, Russia and Turkey. However, there are still exceptions like Italy and Argentina, a more active market doesn't improve the national team. For Argentina, like what I analyzed in the last visualization, the transfer number is mainly the selling number, so with more top players stay, the ranking goes up, corresponding to the relationship between the net transfer fee and ranking.

4.3 Relationship between the attacker-defender ratio and ranking

4.3.1 Calculation and analysis

For the third question, the main focus is on if there is any pattern to the introduced players on their positions, that is related to the ranking. In other words, is introducing an attacker or a defender, more helpful? I calculated the ratio between the number of attackers introduced and the number of defenders introduced. To get the results, I used Spark SQL again to first extract the position type of players the in the buying table and count the number for attackers and defenders for each country per season. Then I divided the number attackers with the number of defenders using a self join as **”(t1.Pos_num / t2.Pos_num) as Attack_Defend_ratio”**, where t1 and t2 represent two entities that are actually from the same table. One thing happened that if no top defender were introduced, the number of defender would be 0, and the division would become meaningless, so I only selected the records with actual defenders being introduced. Finally I joined this table with the ranking and sorted the table according to season from the earliest and Attack_Defend_ratio in descending order.

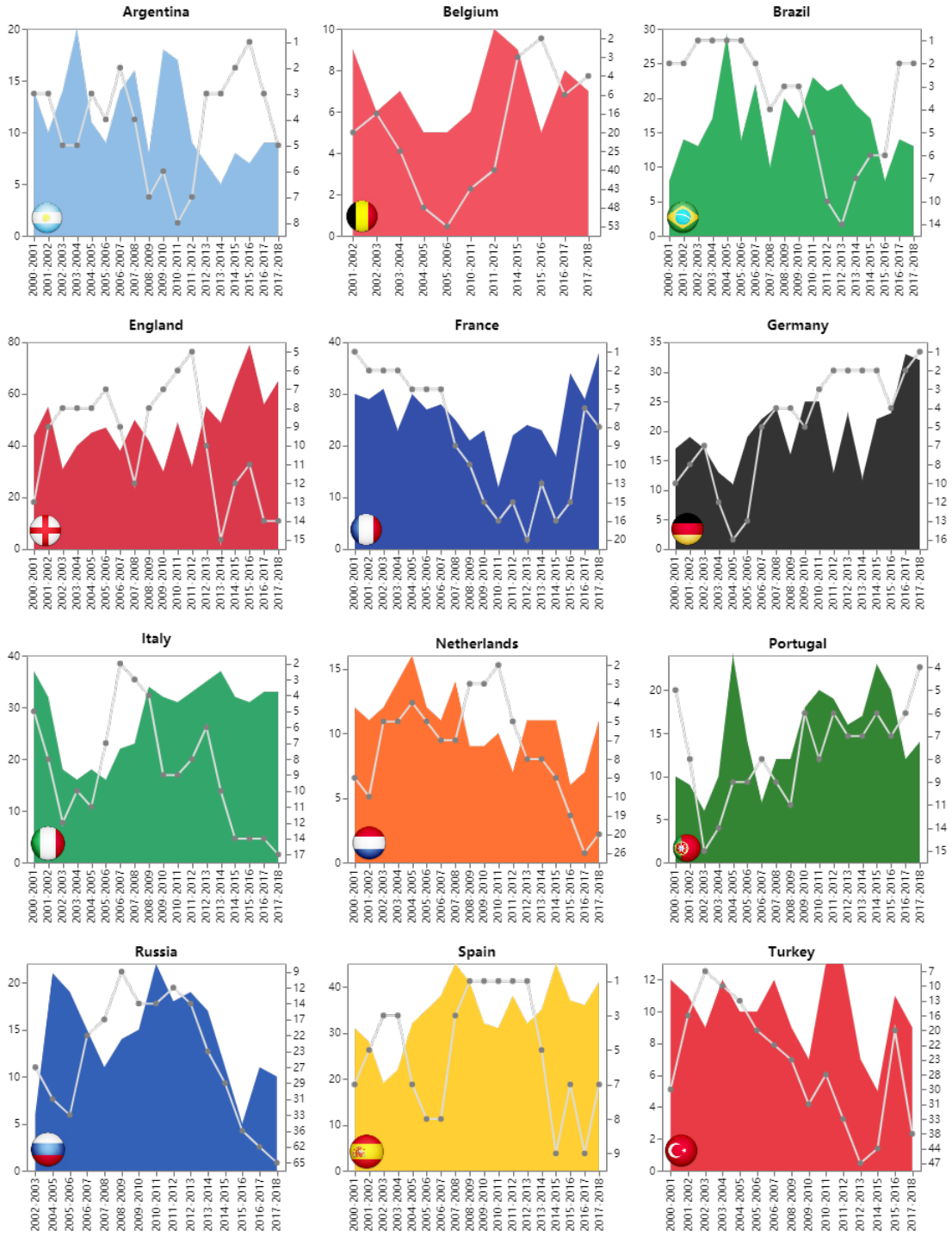

Figure 5: Relationship between the transfer number and ranking.

4.3.2 Visualization

At this time, I turned back to the bar chart to represent the attacker-defender ratio. I also kept the line chart for the ranking. X axis is still the season and the bar chart with the axis on the left is the attacker-defender ratio. The visualization is in Fig.6.

4.3.3 Findings and conclusion

This is an interesting thought, and I didn't expect any good results to visualize. But it seems that the attacker-defender ratio is somehow related to the ranking. Basically a higher ratio indicates more top attackers were introduced and a lower ratio under 1 shows more top defenders coming. Two interesting findings are listed:

- For countries with leagues that have more physical confrontation, a top attacker might be good for improving the overall quality of the leagues, and thus even the national team. Examples are England and Russia.
- For countries with leagues that one club stands out mostly with good strikers, other clubs might tend to introduce more defenders to compete with this giant club. Those the introduction of defenders might at the same time build the confidence for other teams, and also help the attackers in the giant improve themselves. Examples are Paris Saint-Germain F.C. in France and FC Bayern Munich in Germany.

5 Challenges

There are several obstacles during the process of this project. One of the challenges is the same league name from two different countries in the transfer dataset. For instance, China, Switzerland and Greece have their top leagues named as "Super League", Russia and Ukraine both have their "Premier Liga" leagues, which introduces a lot of trouble for me. To distinguish which record is for which country, I did type the team name into Wikipedia and found out the country name based on the team. It took me quite a long time to finish this job. Another challenge is about the position to divide attacker and defender. I used the difference in positions to divide, but as we know, in the modern football system, there's no strict division for attacker and defender based on the position of a player. A player at defend midfield might also be called as an attacker since he mostly undertakes the major organizing task, like Cesc Fàbregas. In my project, I neglected this influence though, it might lead to some uncertainties or errors on my analysis.

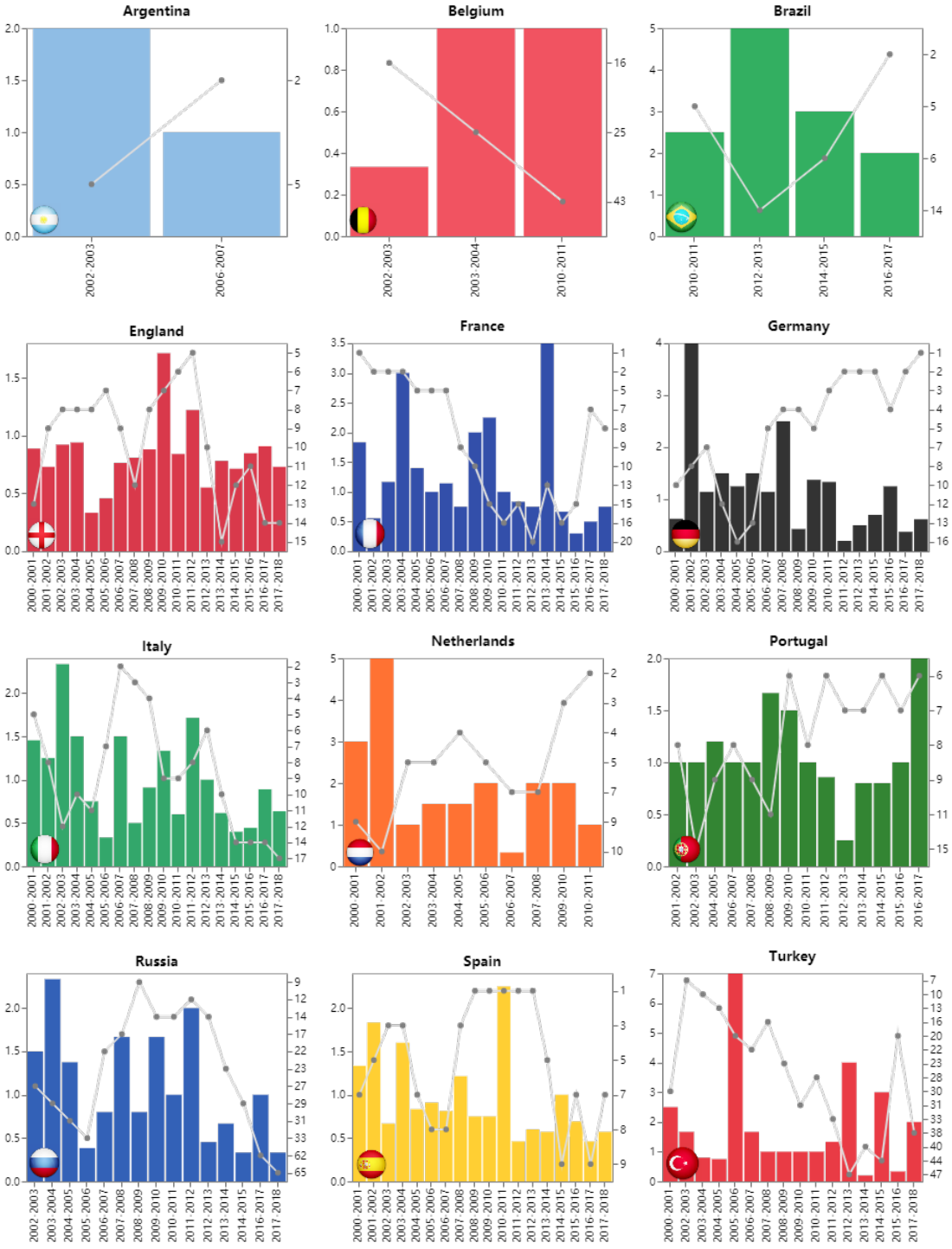

Figure 6: Relationship between the attacker-defender ratio and ranking.