CMSE 402: Final Writeup

Jacob Morrison

3 May 2018

1 Dataset

For the final project, I chose a dataset that included statistics on all international soccer matches that have occurred between 1872 and January of 2018. I modified the results slightly to split the date up into a separate column for the year, month, and day, rather than one column listing the date. Otherwise, I made no modifications to the dataset. A link to the original dataset can be found here:

https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017.

2 Questions

I ended up asking essentially the same questions as I posed in the written checkup. The final form of the questions, as shown in the slides and on the poster, are:

1. How has goal scoring changed over time?
2. What country has hosted the most international matches?
3. Is homefield advantage a real phenomenon?
4. What is the record for each team that has played in the World Cup?

3 Findings

It wasn't until very late in the project that I found a way to tie the questions together. Over the past few years, I've noticed that people get really into soccer around the time the World Cup happens, but tend to be less interested in the time between Cups. For that reason, I thought the questions I asked and the conclusions I drew from them would serve as interesting factoids to show people who are starting to get excited about soccer again as the World Cup returns this summer. The conclusions that I drew from each visualization are:

1. **How has goal scoring changed over time?** While the average number of goals scored has gone down over time, I noticed that the average has flattened out over the past 30 years or so. One thought I had on this trend was that, in recent years, hundreds of international matches are played each year, compared to only a handful in the early years. In general, soccer tends to be a low scoring affair, and with many games being played, the sheer number of games with "normal" scores will wash out the pull that lopsided matches have on the average.

2. **What country has hosted the most international matches?** Interestingly, the United States has hosted the most international soccer matches over the years, hosting almost twice as many matches as the next country, France. I wonder if this stems from the quality of the infrastructure in the U.S. for hosting large scale tournaments, compared with many of the other countries in the Western Hemisphere. On the other hand, in the Eastern Hemisphere, many European countries are on par with the U.S. in terms of the infrastructure needed to host internationals soccer matches, causing them to be spread among the countries, rather than one nation getting the lion's share of matches.

3. **Is homefield advantage a real phenomenon?** Before getting to the conclusion for this question, a little background is needed. First, the equation for calculating the winning percentage counts a tie as one-half of a win, so the equation looks like this:

$$\frac{wins \ + \ ties/2}{total \ games \ played}. \tag{1}$$

   For the actual comparison, I looked at the winning percentage of the team who is playing in their home country and then compared that to the winning percentage of the listed home team if both teams are playing in a country that is not their own. I also compared the winning percentages when teams are playing in a tournament or a friendly match, in addition to doing a combined analysis. To first order (comparing whether the error bars overlap or not), I found that for both tournament matches and the combined analysis, a team that plays in their home country has a higher chance of winning, or at least tying, than playing in a neutral country. The error bars were large enough when broken down by friendly matches that I couldn't, with any level of certainty, come to this same conclusion.

4. **What is the record for each team that has played in the World Cup?** Unsurprisingly, I found that Germany and Brazil have dominated the FIFA World Cup since its inception. The thing that I found most surprising when answering this question was that there have been teams who have played in the World Cup once or twice and have never tied, let alone won, a match.

4 What I Learned Through This Project

One of the biggest things I learned through this project was the difficulty of taking a dataset, then determining a set of questions that were focused on a central theme for that dataset. I believe much of my difficulty stems from a lack of practice in that department, but I also didn't have a cohesive topic that tied my questions together until later on in the project. With regards to having a predefined topic, I believe that having such a topic to guide your questioning is useful, but, as it was in my project, it is possible to start out by asking a series of questions and then finding a topic based on the results that you find. In the future, I will need to be sure to plan ahead when either determining a theme for my questions or using my questions to guide my choice of theme.

Another thing I learned through this project was the art of using visualizations to tell a narrative. Most of the time, when I make plots and graphs for my research group, I'm more trying to quickly glean information either for myself or my group and less trying to tell a cohesive story to the wider scientific community. Using these plots to tell a story to those outside my collaboration does happen, but it occurs at irregular intervals. Therefore, at times, it can be hard to find the narrative among the flood of graphs. This project showed me that, by finding a theme amongst the data and using it to ask key questions, a narrative for the data can be created and presented to others.

5 Expectation vs Reality

In general, most things went as expected with the project. Two things which went differently than originally anticipated are the amount of time it took me to find a theme for my international soccer match data and the time I had to work on my project. In the previous section, I talked about the difficulty I had in finding a theme and what I learned from that. The reason behind what I learned is that it took much longer for me than I had originally thought it would to find a theme that linked the four questions I had determined early on in the project. That being said, I felt that I had a lot more time to devote to the project than I had anticipated going into it. Part of that was due to my schedule for my outside-of-class research projects and part was due to the due dates which forced me to work on the project consistently over the past two months or so.