Any surprises from your domain from these data?

- I guess the biggest surprise from these data are that we could never find a single source for all the data we needed. The basketball-refernce.com website had any player statistic imaginable except for player salary. We had to get salary values from ESPN. We tried other sources to get the data we needed but because we are looking at some very specific statistics, most sites did not track these values. For example, blocks and field goals are very common statistics and appear on most websites. Defensive box plus minus, however, was only available on the basketball-reference site. Therefore, we were relegated to use multiple sites to complete the necessary dataset for this analysis. Maybe we should start publishing this complete dataset for others to use in the future.

The dataset is what you thought it was?

- Once we got the ESPN and basketball-reference data combined, we noticed that there were a lot of unaccounted for records. Either ESPN was reporting a player salary that did not correlate with any player statistics from the basketball-reference data or vice versa. As a result, we were forced to delete about 10 records because they were either missing salary or statistical data. Furthermore, there are some missing values in some of the data columns. These missing values will likely just be set to zero because they appear as a – symbol in the data columns. What this suggests is that no value was actually recorded. In other words, the value should be zero.

Have you had to adjust your approach or research questions?

- We haven't had to adjust our approach as of yet because we haven't really started on any of the actual analysis. One thing I guess that we did end up changing is that we went from manually building the two datasets to using a web scraper to do the work. The manual method was technically quicker, but if we want to continue to use this model in the future, a web scraping approach will be much better. A potential issue with this method, however, is that the web page structure needs to remain the same from year-to-year for the scraper to work correctly. This is easy to verify once we run a scrape, so I don't see this being too big of an issue. Additionally, I couldn't get our scraping script to run on my PC, but it worked just fine on my Mac. The other guys in our group were able to get their PCs to run the code just fine. This isn't a big deal, but something that might be an issue if anyone besides us wants to use the code.

Is your method working?

- As I stated earlier, I don't really have much to report on regarding our methods working or not. The one thing we have implement so far, the web scraping tool, is working correctly and is a pretty slick piece of code. One thing that isn't working, however, is the simple task of manually joining the data. For example, we noticed some mistakes because some players were listed multiple times in the ESPN data. Each time a player was traded throughout the year, they would show up as a separate salary entry in the data. This caused more than a few issues when it came time to match this player up with his statistics.

What challenges are you having?

- The major challenge so far is getting the two disparate datasets to match up. When we first did it by hand (prior to using the web scraper), doing a basic join on the data, based on player name, seemed straightforward enough, but we still ended up having to do a lot of manual inspection of the data to ensure that things matched up correctly. For example, some players had special characters in their names on the ESPN salary data, like Nikola Vučević. These symbols were not included in the basketball-reference dataset. Because of this, the joining of the two datasets did not always work correctly. What is really confusing is that sometimes the matching would work out just fine, even with the special symbols. I guess it's something to do with the joining algorithm. In the future, we should probably find a way to simply remove any special symbols from the data before we do a manual join.