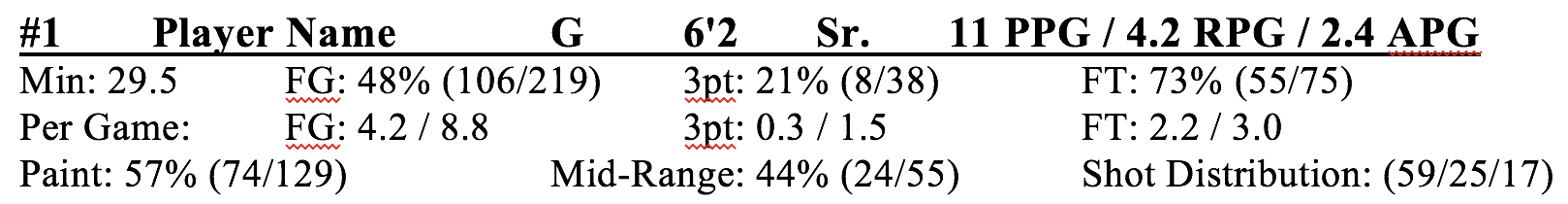
**Introduction**

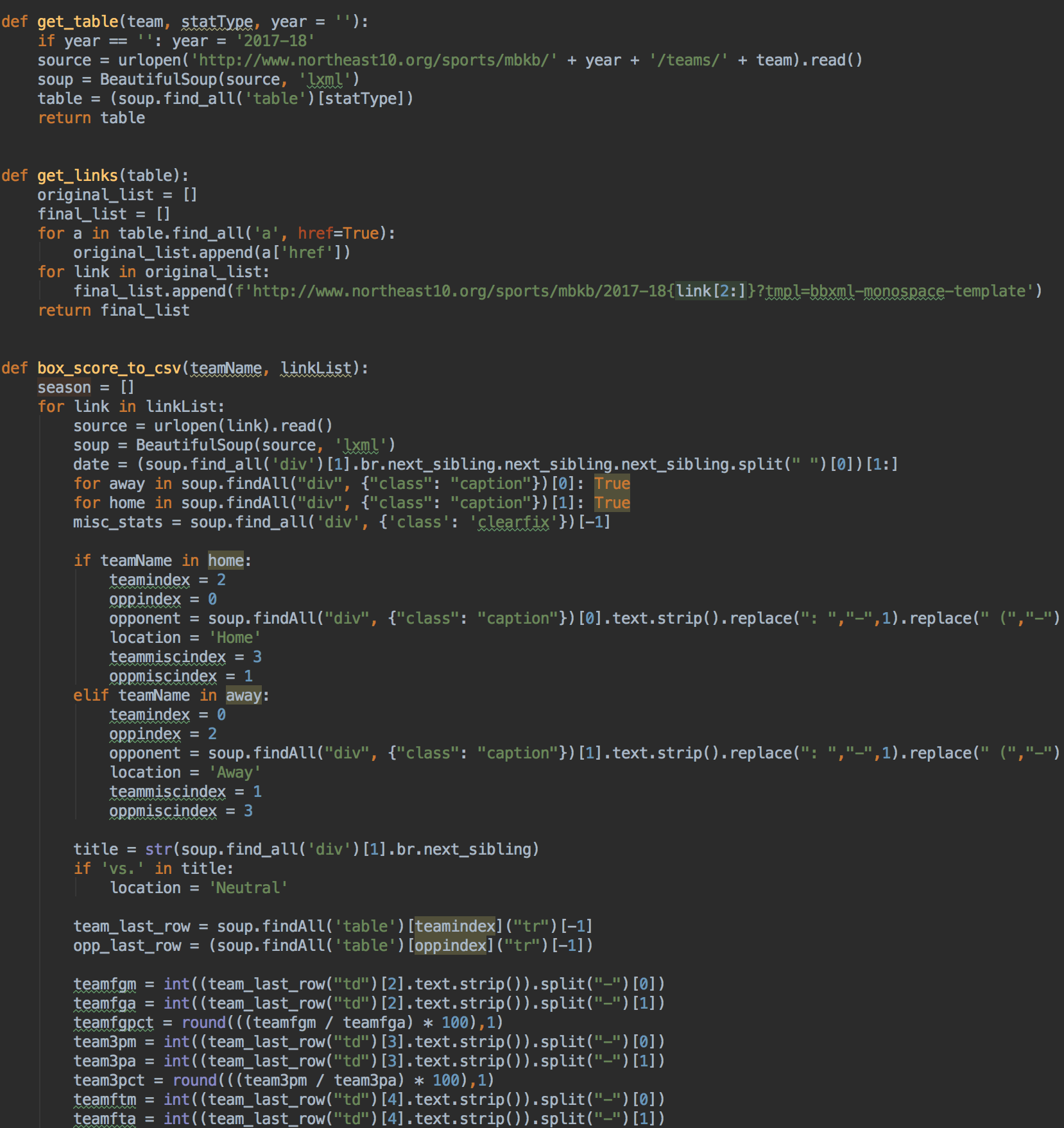
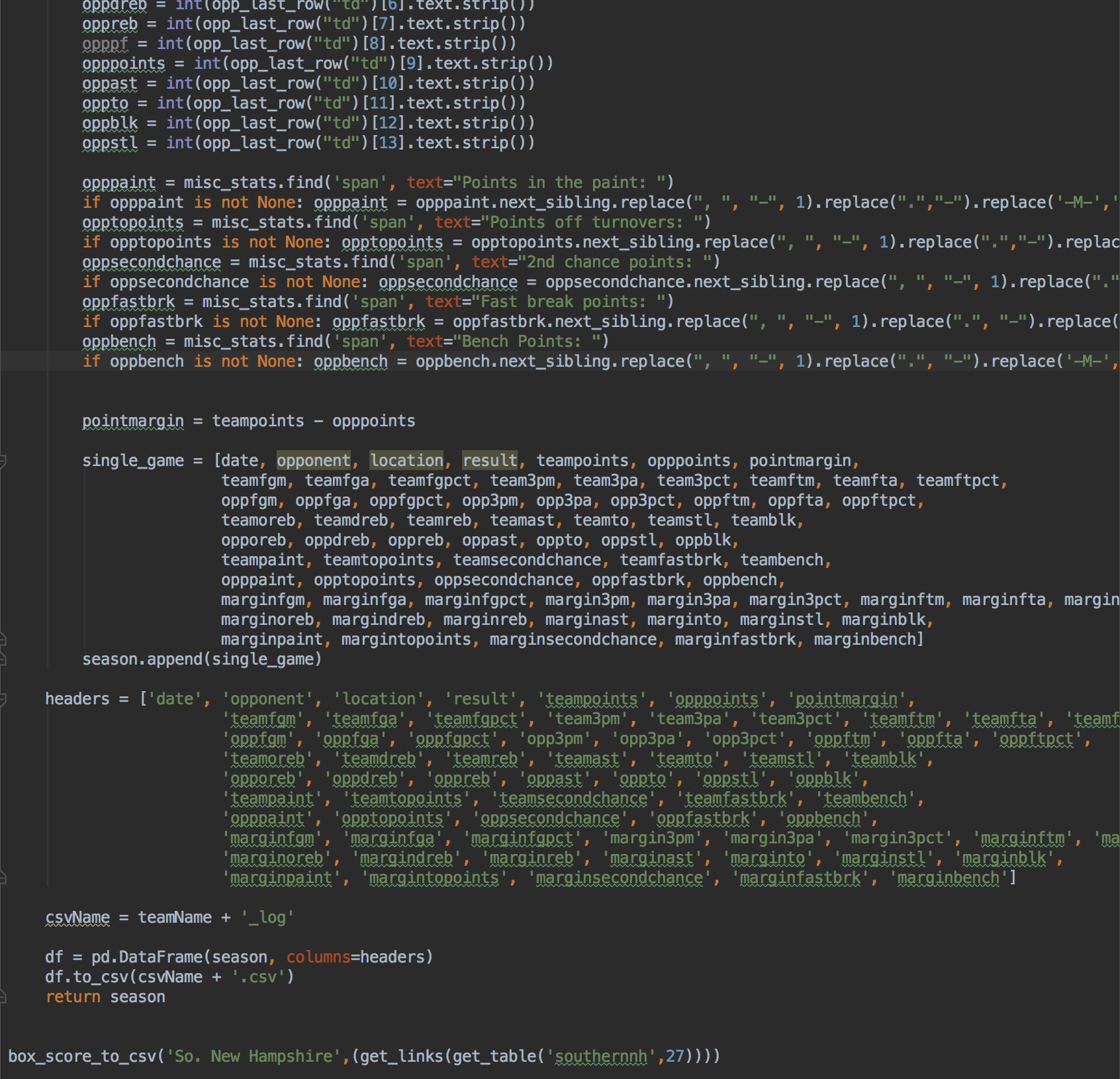
In my brief time as a graduate assistant for the Men’s Basketball Team at The College of Saint Rose, numbers and basic statistics have been utilized to save many hours in the film room. My typical role is to create scouting reports using basic statistics such as points scored, assists, and rebounds in each game, as well as shooting percentages to understand tendancies without having to ever watch a second of game film. The same concepts can be applied to teams on the whole.



The idea for this project was simple. How can I utilize the advanced concepts learned in class in order to gain a better understanding for not only our opponents, but our own team as well?

This question seemed pretty simple and easy enough to answer, but it turned out to prove incredibly challenging at the start. The three questions that had to be answered were who would data be collected on (team or player?), what source would it be collected from (advanced metrics on synergy sports or basic box scores?), and how would it be collected (copy and paste, or python scraping)?

Eventually, I decided to limit the scope to team data, taken from box scores, scraped using the BeautifulSoup package from Python. The reason I decided to choose this is because players are relatively easy to understand without using advanced statistics. For example, if a player shoots 80% of his shots from the 3 point line, we already know to run him off of the 3 point line and make him dribble inside. Or if a big man has many turnovers, we may decide to double team him when he gets the ball. Teams however are much more complex. It may take many hours of film study to understand what makes a team tick, why they win, and the reasons why they lose. Box score data what selected because it is very easy to understand and exists for every game, but the drawback is that it is hard to get it in a usable format for data science, thus the need for Python.

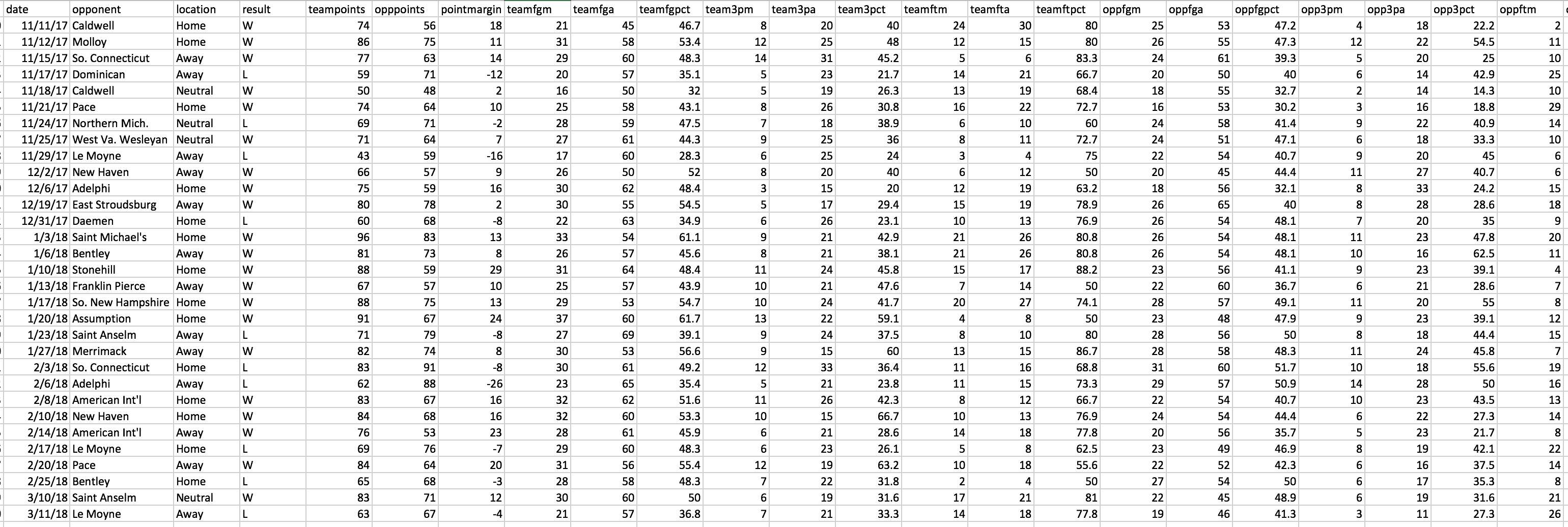
**Beautiful Soup**

The web scraping tool I decided to use was Python’s Beautiful Soup package. I had some very basic experience with Python, but never any web scraping. Overall I found the code to be fairly straight forward minus some minor struggles. Most of the data I was interested in was located in different tables, so it was just a matter of getting the links for each game, identifying the tables, and assigning variables to the data in a specific row of interest. Unfortunately, the home team was always first and the away team was always second, so I had to include additional logic to determine where the tables were located for the team I was interested in based on the location of the game.

Overall, it took me about 10 hours of googling, searching Stack Overflow, and coding. Much of the code was straightforward, but it did take some time to iron out certain bugs. 200 semi-poorly written lines of code later, I had game logs saved to CSV for 10 of the 15 teams in our league. Five teams had a bug that I was unable to figure out and still need to fix, but I was still left with enough data for this project.

**Data**

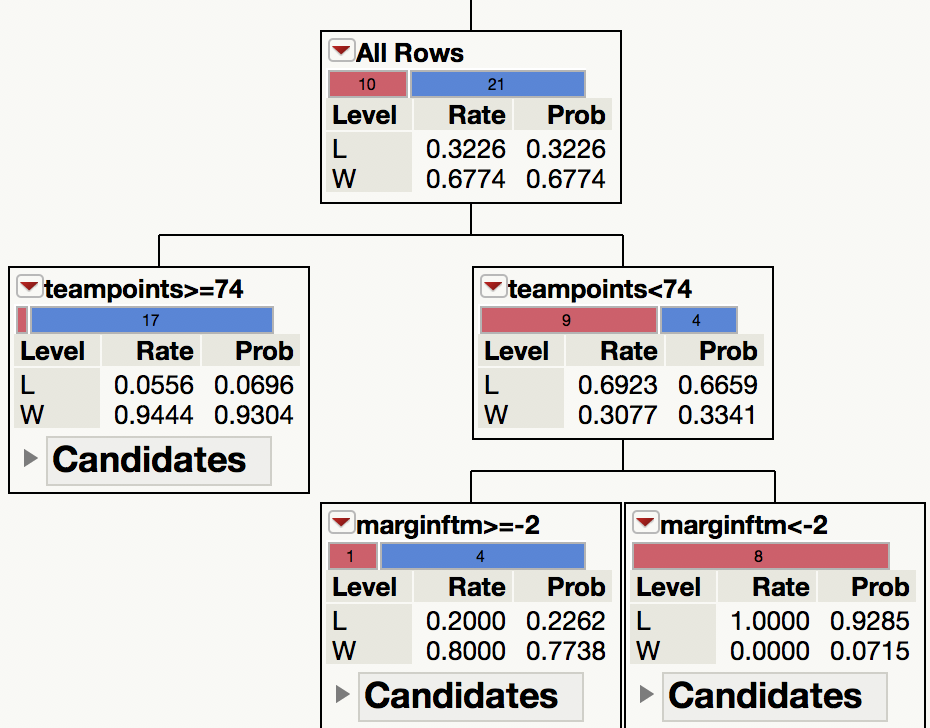
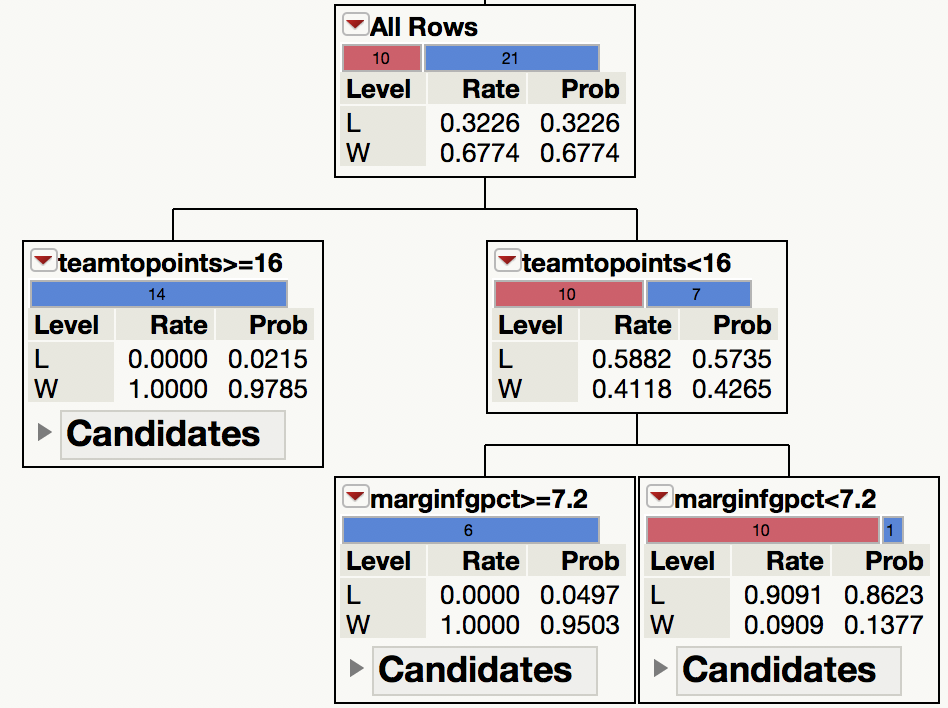
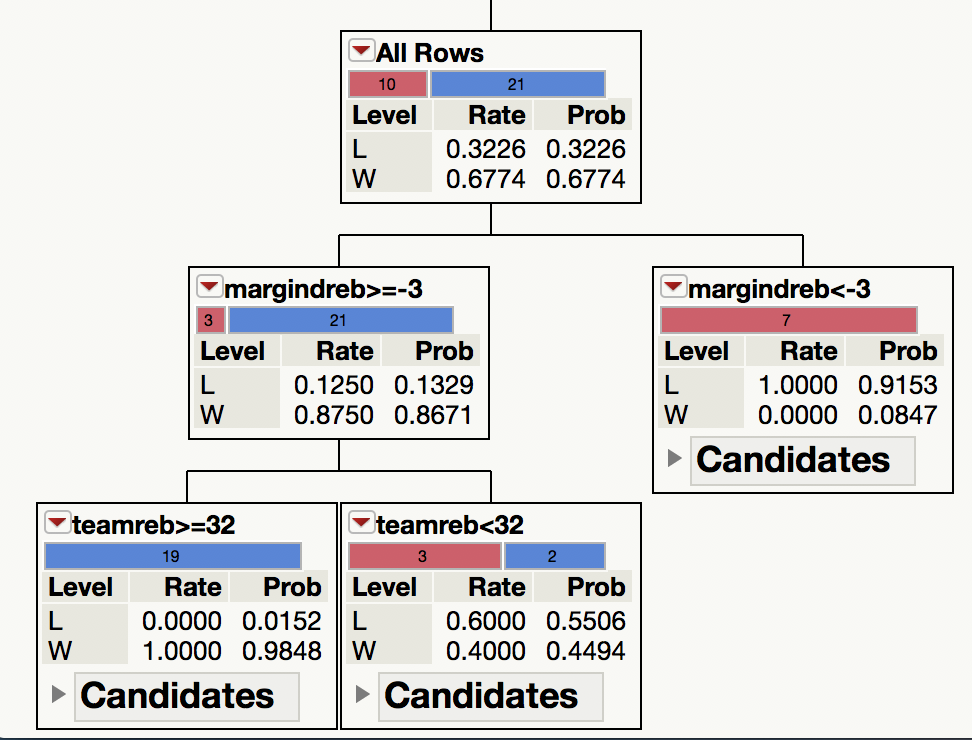
After scraping, I was left with 25-30 observations for each team, with 70 variables. These variables included date, location, opponent, and result for each game, as well as box score stats for both teams (points, rebounds, assists, etc), and the margins between them. Margins seemed to be useful because 30 rebounds might seem like a lot, but if the other team has 35, they were outperformed in that category. I thought that getting more rebounds than the other team in a given game might be a much more important metric than getting a certain number of them.



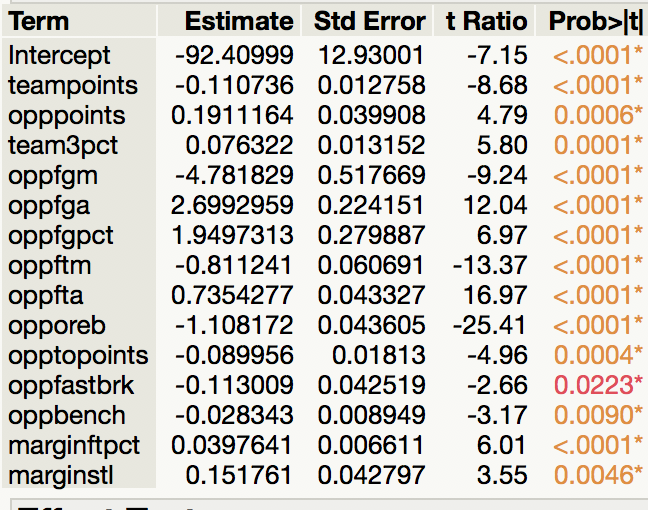
**Methods**

I found decision trees to be most useful because not only were they very easy to set up and provided great insights into the data, but they would likely be very easy to explain to someone who might not understand the math and numbers behind them (such as the coaching staff). I first decided to analyze the Saint Rose data to get a better understanding of our own strenghts and weaknesses. The numbers show that we went 0-7 when the defensive rebound margin was -3 or lower, and we were 21-3 when the margin was between +∞ and -3. Additionally, in those 24 games, we were 19-0 if we gathered 32 or more rebounds, or 2-3 if not. The variables for defensive rebound margin and team rebounds were removed, and the decision trees were rerun. The tree then split on 16+ points off of turnovers, where we were 14-0, or 7-10 with less than 16 points. In those 17 games, if the margin of field goal percent favored us by 7.2% or more, we were 6-0, or 1-10 otherwise. The two variables were removed, and the model was created again. This time, it showed that we were 17-1 when scoring 74+ points, or 4-9 when scoring less than 74, and of those 13 games, we were 4-1 when the free throw margin was greater than 1, or 0-8 otherwise.

This insight is incredibly helpful and right away tells us a lot about our team. While it’s hard from a coaching standpoint to just say “go out and get 30 rebounds,” there are definitely things a coach may decide to do from a strategic standpoint that can increase the likelihood of rebounds, turnovers, free throws, shot selection, etc. Pace of play may be one of them, for example. Faster shots may lead to more shot attempts during the course of the game, which could also provide the team with more defensive rebounding opportunities. Changing these strategies would certainly be a bit above my pay grade, but arming a head coach with this knowledge would be nothing but positive.

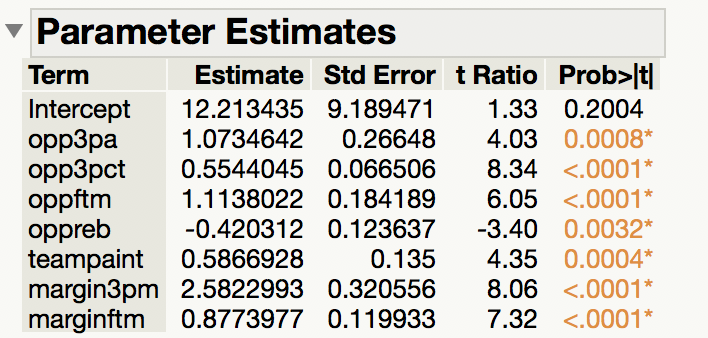


Next, I decided to take a look at the defensive rebounding variable from the first decision tree, using stepwise regression analysis. I wanted to gain a better understanding of the other variables that may cause an increase in the number of defensive rebounds, which these decision trees deemed to be so important to our team’s success.



Right away, I notice some pretty obvious metrics that go into the stat. The most important seem to be opponent offensive rebounds, which makes sense because the more offensive rebounds an opponent gets, the less defensive rebounds you can get, as well as opponent field goal percentages, which have a similar relationship. The more the opponent’s shots are going in, the less opportunities you’ll have to gather a rebound. The variables I found to be interesting however, revolve around free throws, and opponent steals / points in transition. It seems obvious that free throws would act in a similar manner to field goals, however I never stopped to consider the fact that free throws are much less likely to go in at this Division 2 level, as opposed to the NBA which I frequently spend time watching on TV, due to lower percentages. Additionally, more opponent steals may lead to transition baskets, which lead to more efficient shots by the opponent, which would limit rebounding opportunities.

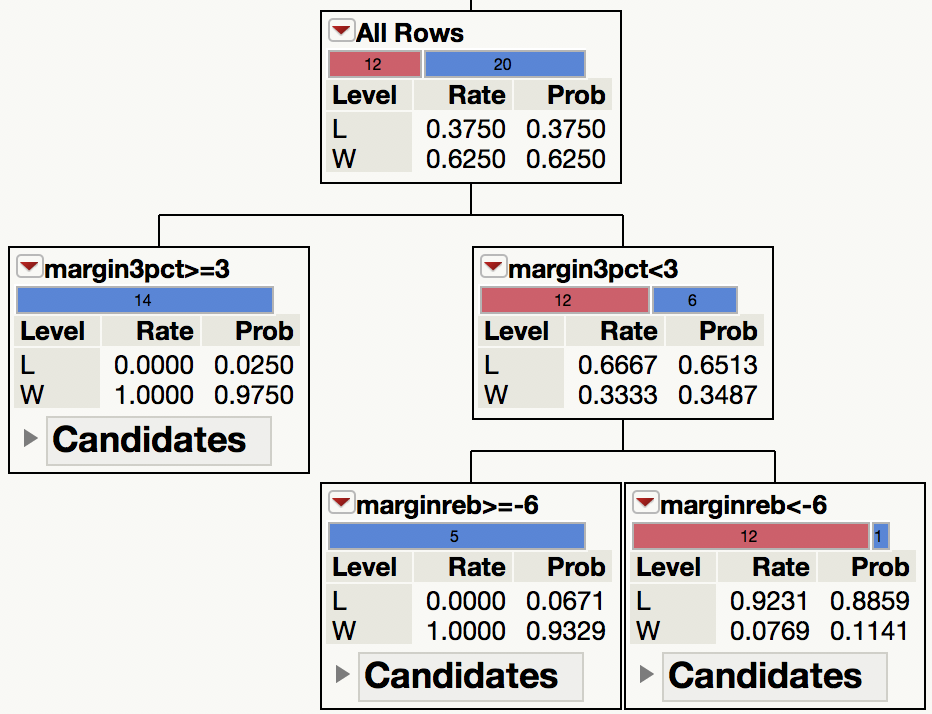
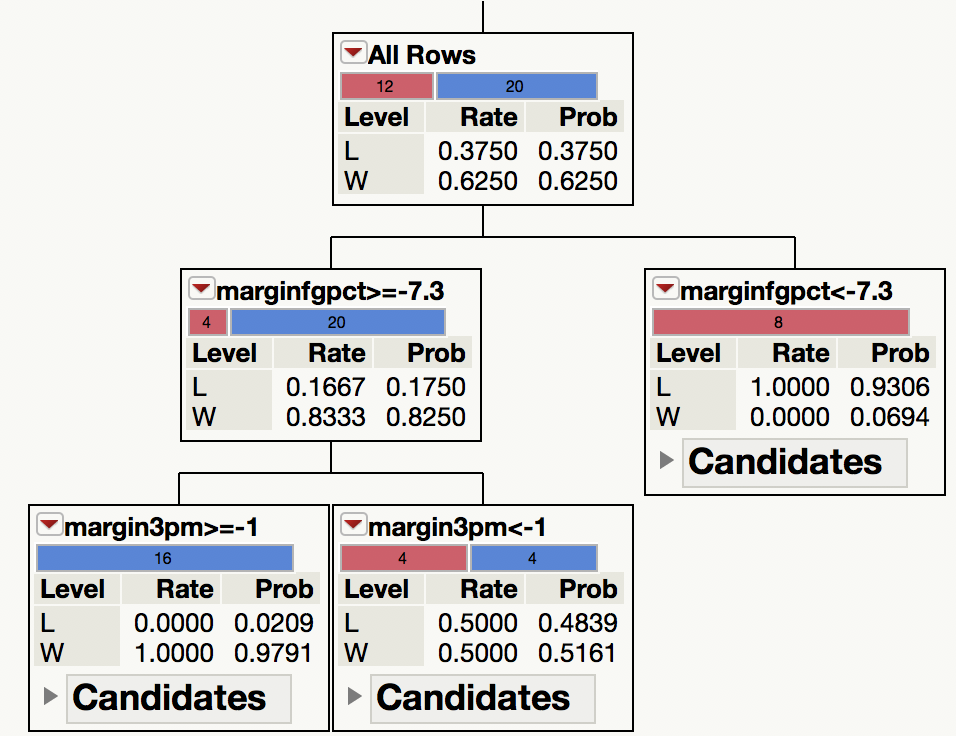
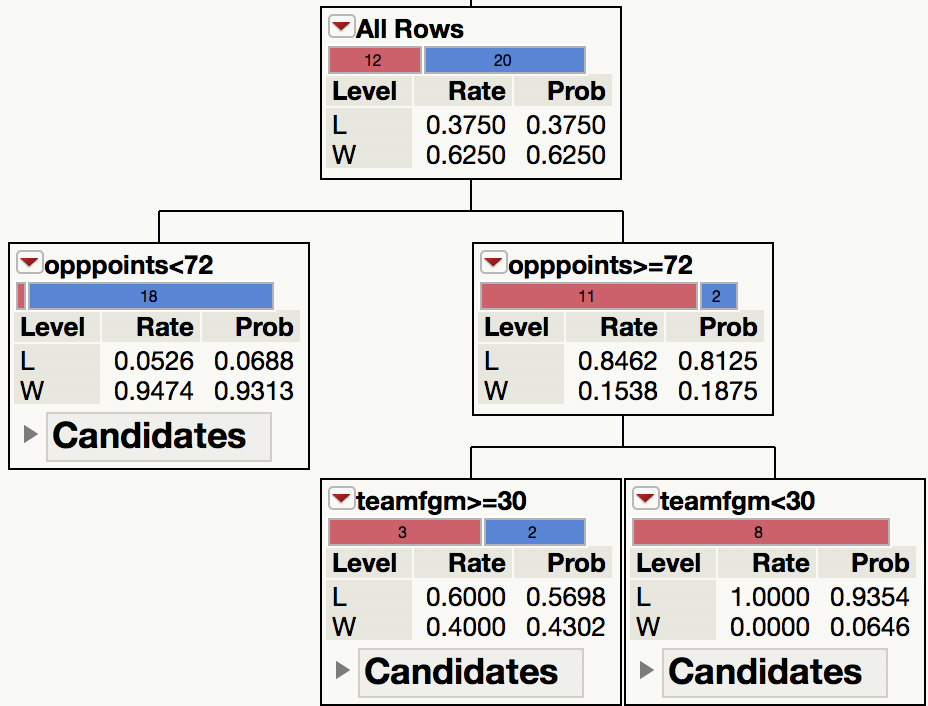
Lastly, I took a look at points scored, the same way, removing obvious variables such as shooting percentages and team points.



These results also seem to have some interesting insights. Many of the opponent’s shooting stats seemed to impact our points scored, which is interesting, but may be because if the score is higher we are looking to play faster and score more points to catch up to the opponent. I also thought that points in the paint was a significant factor, which might mean that plays should be run for big men inside to increase the likelihood of high percentage shots.

**“Opposition Research”**

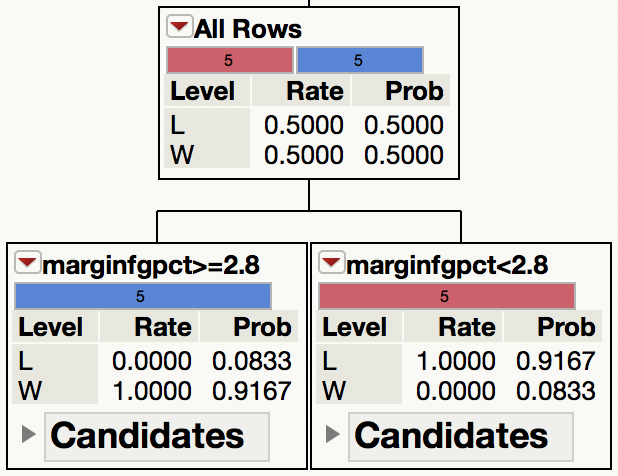
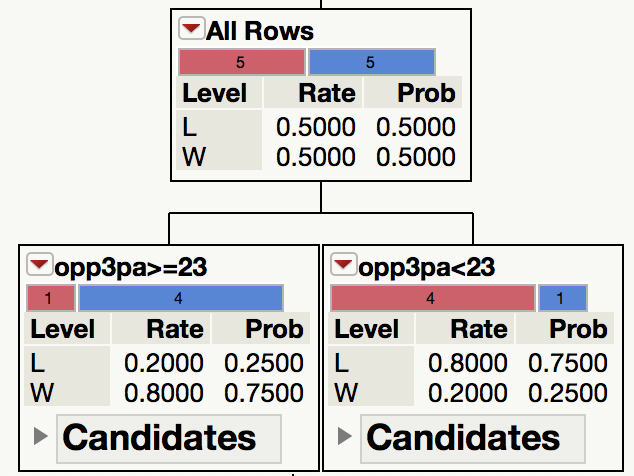
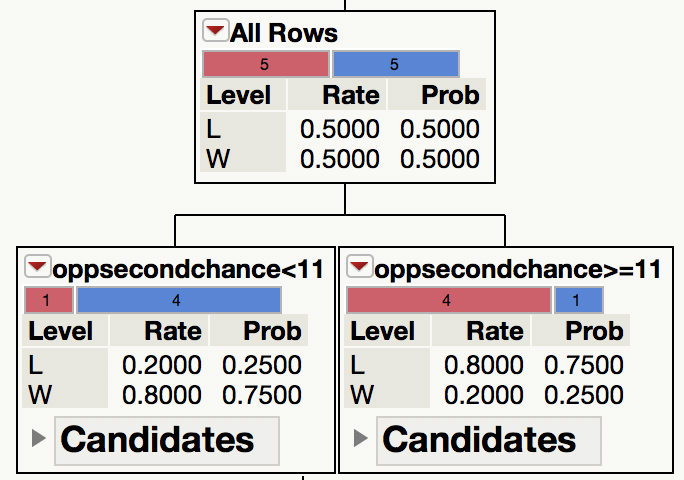
Decision trees were also run on an opposing team, the Merrimack College Warriors. The analysis revealed, as expected, variables different from Saint Rose that were keys to the success of the team.



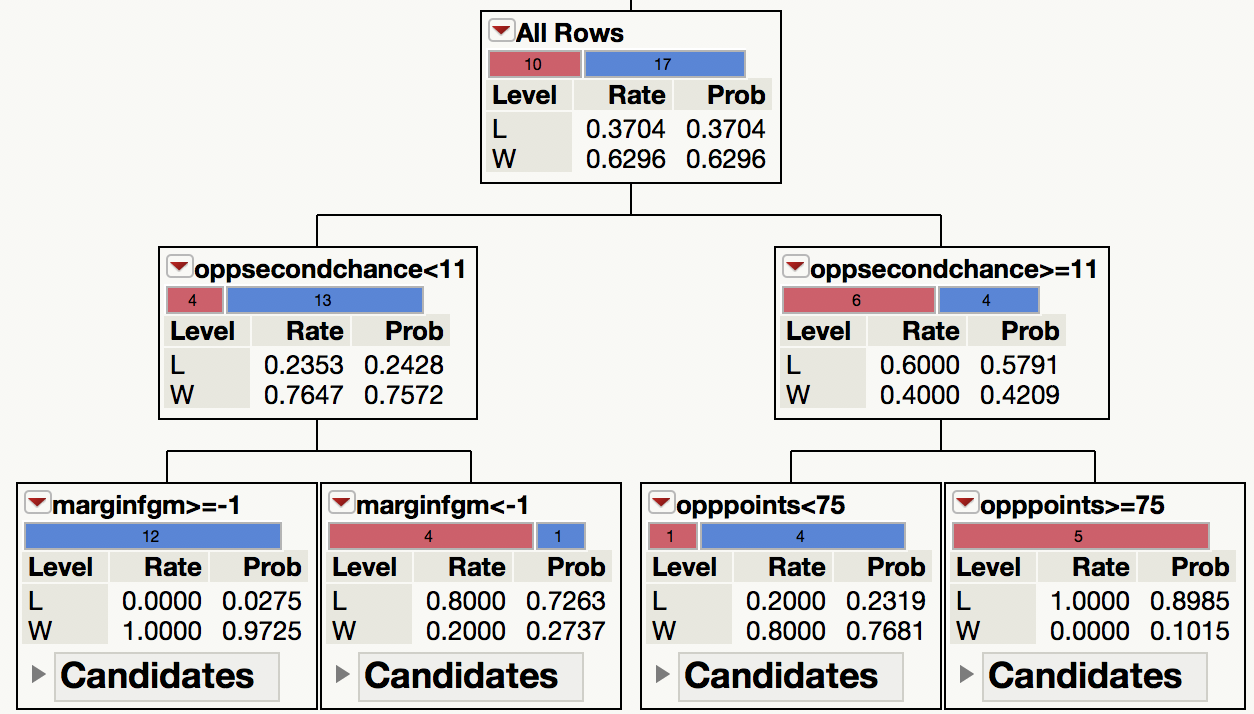
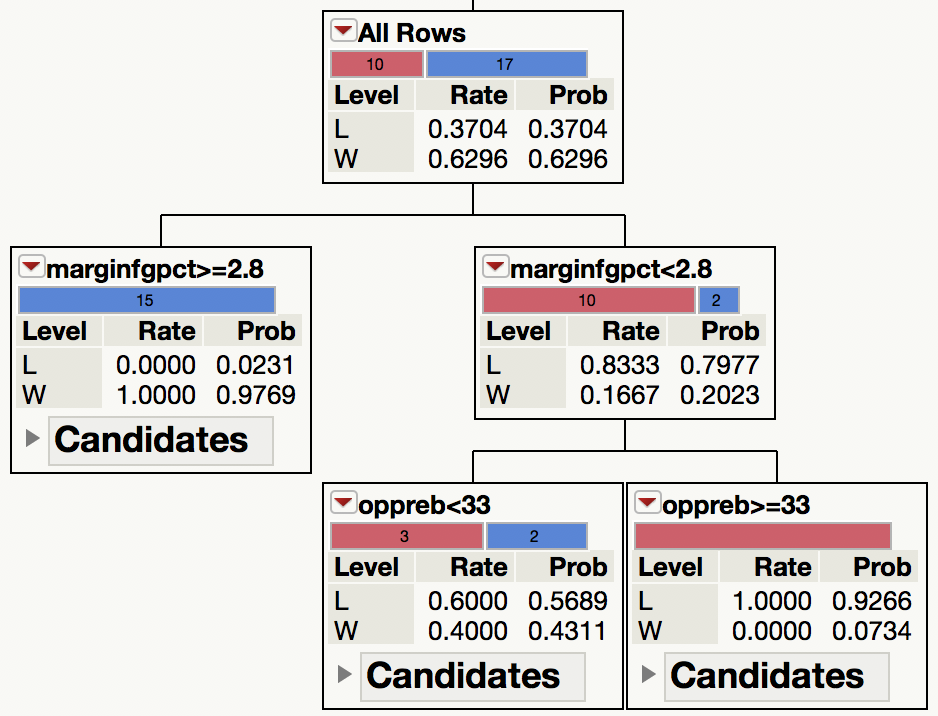
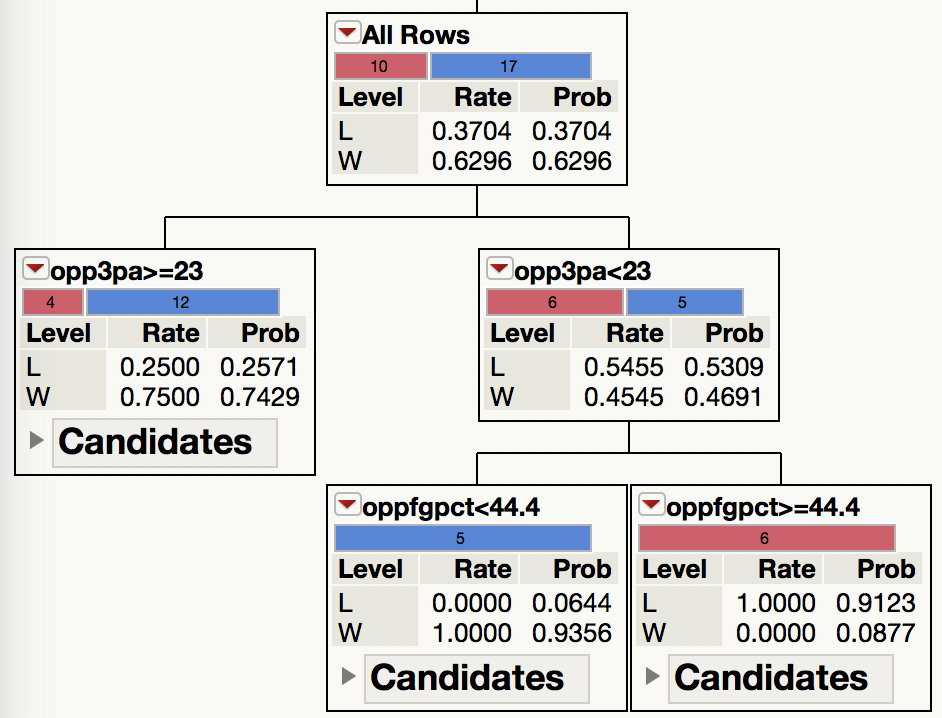
Right away, it appears as though a game between Saint Rose and Merrimack would favor us in a high tempo, high scoring duel with many shots taken. As before, a coach with many years of experience would need to take a look at this data to better craft a strategy for a potential game.

**Possible Issues**

While this data makes sense after the conclusion of a season, it may be hard to analyze early on in the year. Decision trees were unable to be split with less than 10 observations. I ran analysis for two different teams, once through 10 games, and once during the full season, to see how well the decision tree could predict the team’s strenghts and weaknesses early in the season. The first team, New Haven, was 5-5 through the first 10 games, so the system had an evenly split number of results to filter data on.

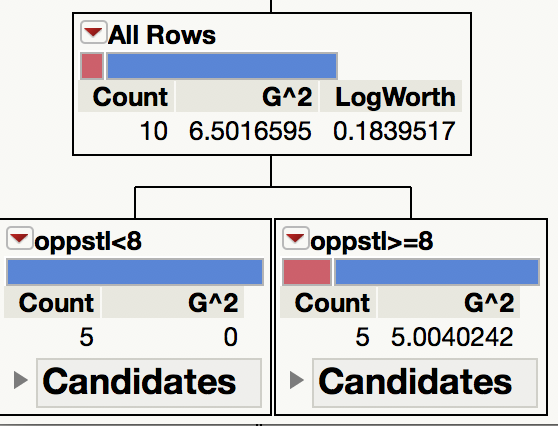
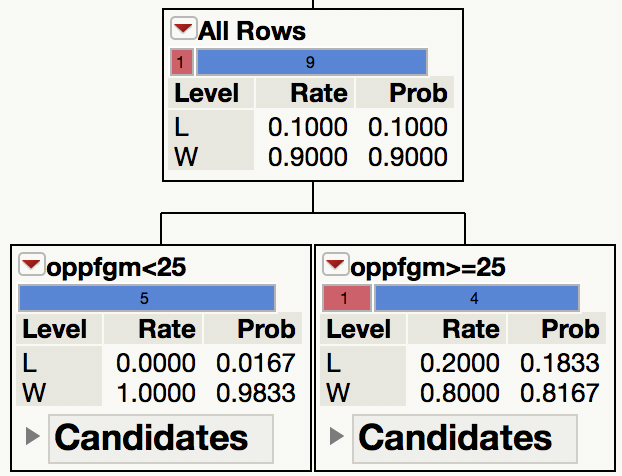
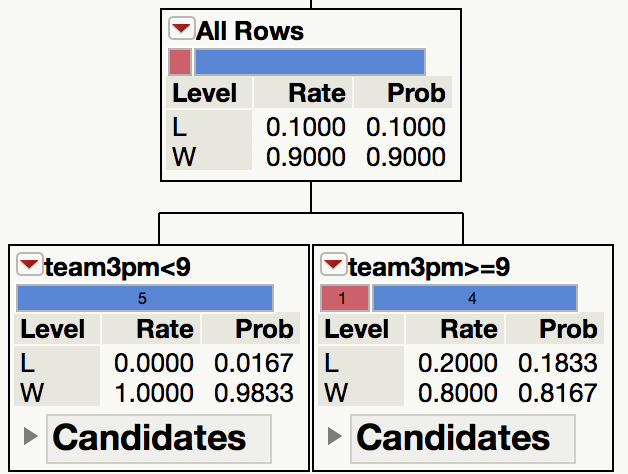
 

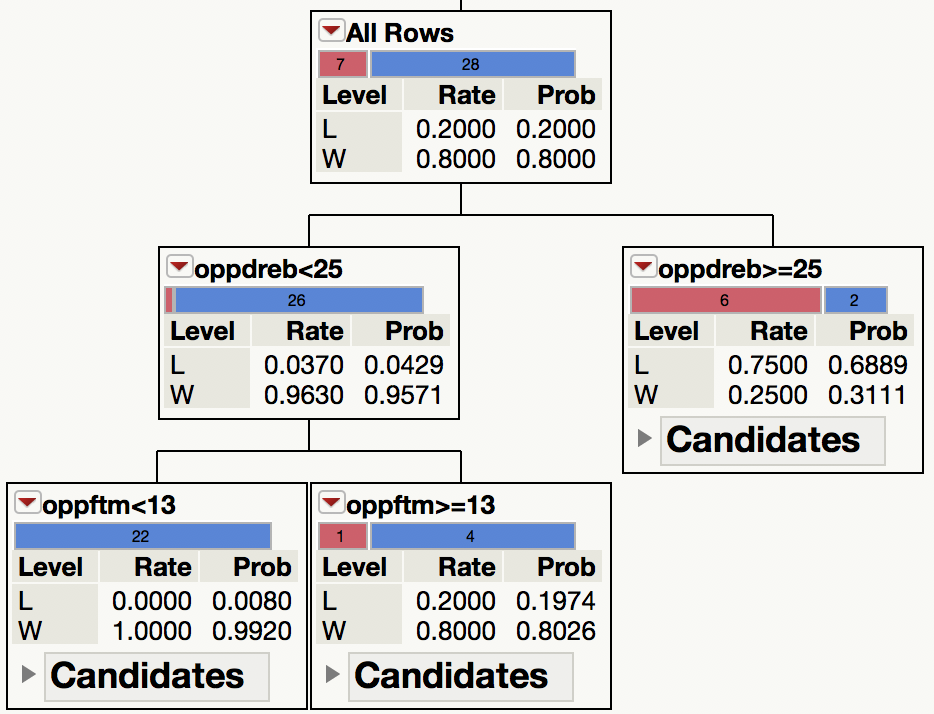
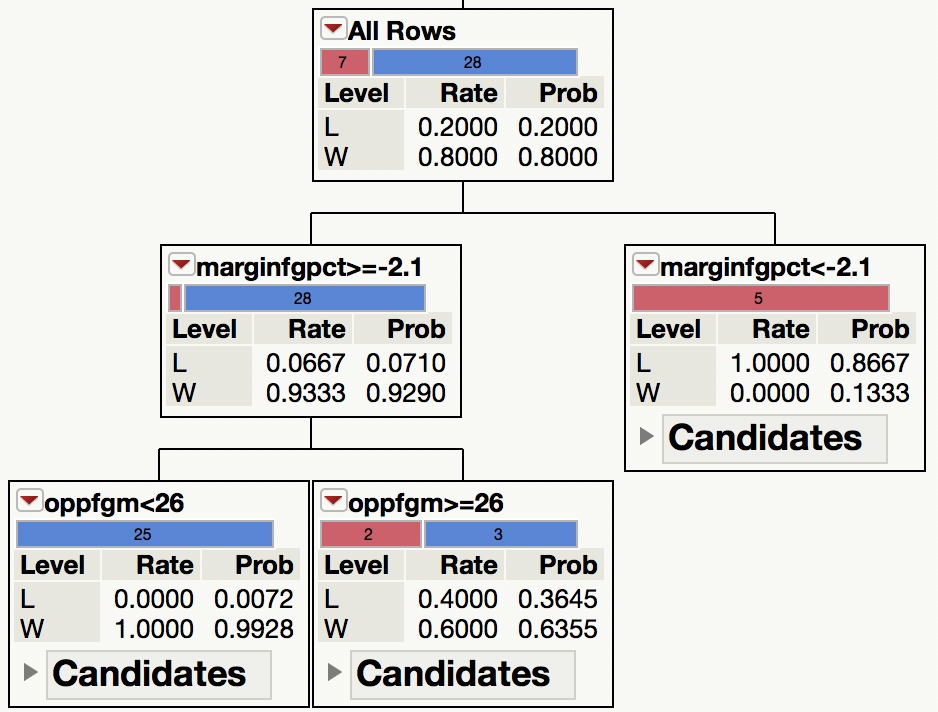
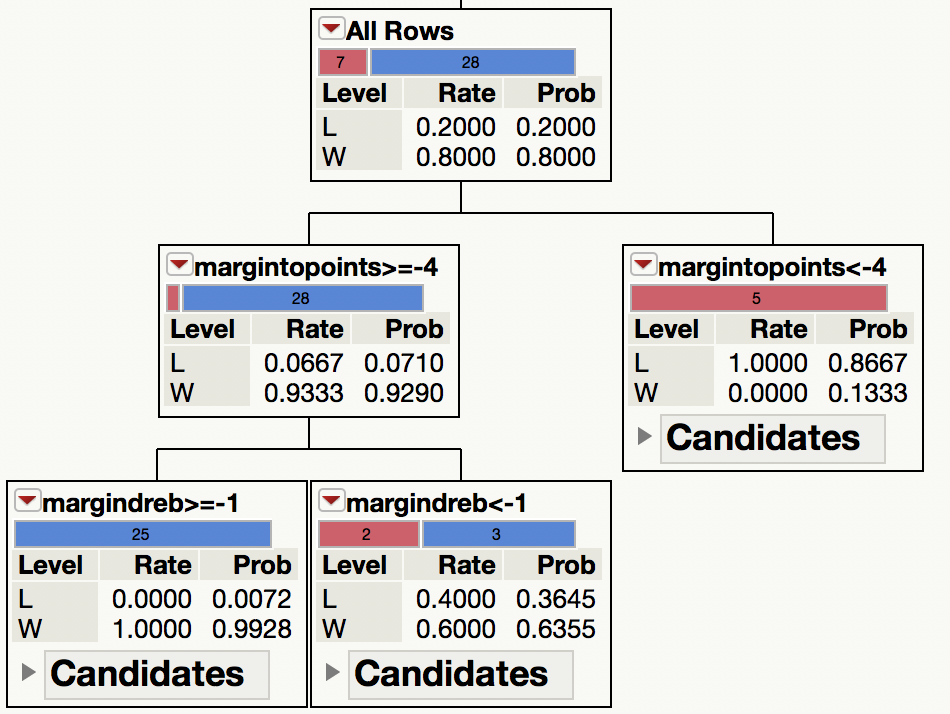
There is a decent amount of data to work with here, as in the first games of the season, the team needs to limit opponent 3 point field goals, keep their field goal margin relatively even, and limit second chance points.



When split again on the same variables at the end of the season, the data stays surprisingly consistent for the most part. Field goal percent margin is still a very important variable, and while the other two vary, it seems like the data taken from a 10 game sample size can still be considered usable so early in the season. That is, as long as the team has a somewhat even number of wins and losses to compare, which leads me to my next test.

The LeMoyne Dolphins were 9-1 after the 10th game, and an incredibly good team. They would eventually go on to win the conference championship and represent our region in the Elite Eight in North Dakota. I decided to run the analysis after game 10, and again after the full season, but this time instead of splitting at the same variables I wanted to see if the system would select any of the same without my input.





Interestingly enough, opponent field goals made was a variable for the full season data set, but only on the second split. None of the other original variables seemed to make an impact on the season in its entirety. Perhaps a stepwise regression on point margin would be a better test early on in the season, especially for teams that have lost few games. Maybe the question should be “how do we keep this game close” rather than “what are some weaknesses we can exploit.”

**Logistic Regression**

Unfortunately, and much to my surprise, logisitic regression was not applicable to this data set. It didn’t seem like the model was able to decipher the important data from the massive amount of variables it had to sort through. This could possibly be because there are just too few observations in a given season, and because there are so many variables fluctuating wildly between games, it’s just not possible to understand how the variables impact the result. I may find however that this model could be used in a more broad sense with hundreds if not thousands of observations to find how a team typically wins games in the league.

**Conclusion and Future Possibilities**

Overall, I think the decision trees are an incredibly powerful model that I will most definitely be using next season, and the regression models provide solid insights as well. Given the fact that I will be in my same role next season, I would love to continue working on this project and make it better by the start of next year. One way I could do that is to add more Python. As I’m the only one on the coaching staff that knows much about computers, the less explainations and technical support I have to do with this analysis, the better. I feel as though a Python front end and Python data analysis packages would be a great way to keep everything contained in one program without the need for JMP or complicated Python code ever being seen by end users.

Additionally, I could potentially find ways to utilize more data and possibly more models. I would love to add player data to my datasets, as well as advanced metrics like PACE, ORTG, and DRTG, which are calculations using existing data from box scores. I could look to fix the issue regarding the first couple of games in the season, which could be solved by possibly implementing regressions on the score margin, or analyzing data from past seasons to see if similar play styles can be expected on a year to year basis.

Lastly, existing code would likely need to be improved. Since many of the table locations are hard coded into my Python script, one slight change to the website can ruin the entire thing. A better understanding of Python and BeautifulSoup may lead to better data scraping, and better longevity for the system as a whole.