Matthew Li

Liam Robertson

Ethan Sartory

CSCI 185: Web and Data Mining

June 7th, 2024

Analysis of NBA Teams Player's Stats

Introduction:

For our project, we decided to scrape NBA player statistic data to see if there were any correlations between various player attributes. Our main goal behind this project was to answer two questions regarding the individual players and overall team performance respectfully. We first wanted to determine whether factors such as player age, height, or weight correlate with **player performance**. We can use the observations from this analysis to determine whether player performance can be a measure of physical traits or if other factors play a role in player performance. Secondly, we wanted to determine if the total number of **offensive** or **defensive rebounds** has an impactful effect on the team's total points scored.

To achieve our goals, we created a Python algorithm that scrapes statistical data from basketballreference.com. For each interaction, our search was narrowed to all of the players for a certain Team during a certain season. After we performed web scraping, we cleaned and processed the data which allowed us to easily manipulate and use it to perform our analysis. To achieve both goals, our main focus was to find correlations between a variety of attributes, such as age, height, weight, rebounds (offensively and defensively), and avg points score

Hypothesis (Goal 1):

We hypothesized that the player's height will always have a positive correlation with the amount of points scored (in a given range). Our theory was based on the stereotype where tall people are considered very well-suited to playing basketball due to being able to shoot the ball into the hoop more easily compared to someone who was not as tall. As with all datasets we are aware of outliers, and recognize that at a certain point height may prove to be a detriment, and this positive correlation may not be seen in reality.

We predicted that the player's weight would have a positive impact on the amount of points they scored, up to a certain point. A player at a lower weight would probably struggle to match the physical intensity of stronger players. At the same time, once going beyond a certain weight, a player would probably struggle to keep up athletically with their peers.

Between player age and correlation, we theorized that age will have both a positive and negative correlation with the amount of points scored. Peak physical condition among athletes normally occurs in the late 20's. So up to this point we would likely see a positive correlation. However, after this we may in fact see a negative correlation. As older players would be more out of their physical prime, they may not perform as well.

Hypothesis (Goal 2):

Our hypothesis is that a team on average scores more points when they have more offensive rebounds than defensive rebounds. We are making this assumption because offensive rebounds give the team another opportunity to score when they are at their own basket. However, we do acknowledge that defensive rebounds are important as well because it denies the

enemy team the chance to score for that possession. We still believe the former is more important because scoring points is done on “offense” not “defense”.

Implementation:

One of the preliminary steps to web scraping is identifying a common pattern that would allow for a versatile web scraper. We initially analyzed the structure of our target website and determined the web pages we needed to scrape the information from. The home page of https://www.basketball-reference.com/ provides the option to look at the data of players, teams, leaders, scores, drafts, etc. When looking at the data of specific basketball teams, we noticed a pattern with the **path name** in the url. Typically urls are structured in the following format: <host name>/<path name>. In our case the host name is https://www.basketball-reference.com and the **path name** was everything that followed it. For this website, the data we were attempting to scrape had the **path name** in the format: /teams/<team name here>/<season year>.html. As the name suggests, it is in an HTML format so we would extract using BeautifulSoup down the line. The HTML itself contained many different tables including per games, totals, per 36 games, and play-by-play. We specifically focused on the per games table, which contained offensive/defensive rebounds, as well as pts. However, we noticed that there was no information on each players’ weight and height. All information of the player’s height and weight was instead relegated to a separate bio page.

Following this finding, our next step was to perform the actual request itself. We started off with a list **wiz_info** (this would store all of the players’ information for each attribute), then importing the request library, we added the following line: wiz_url = (f'https://www.basketball-reference.com/teams/WAS/2021.html'). After that, we performed

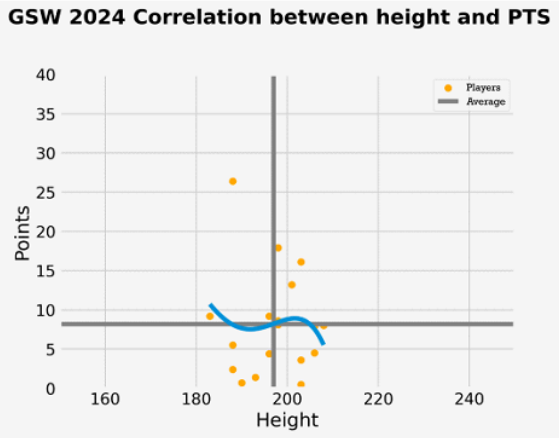
wiz_res = requests.get(wiz_url). An obstacle that we had encountered during testing was that there were some limitations on how many requests we could make at a time. A request number of 200 meant it was successful, but 429 meant that we had hit the request limit and subsequently got blocked by basketballreference.com. So in order to circumvent this request limit we had to continuously change our IP addresses utilizing a VPN. As for trying to test any team multiple times, we had settled on writing all of the information scraped to csv files. For subsequent scrapes of various teams and seasons, we had to implement a consistent and automatic file naming scheme. Using the **path name** from the url, we extracted the team name and seasons and stored this data into a variable (tid). We were then able to append this variable to each file name to organize the data extracted from subsequent tests.

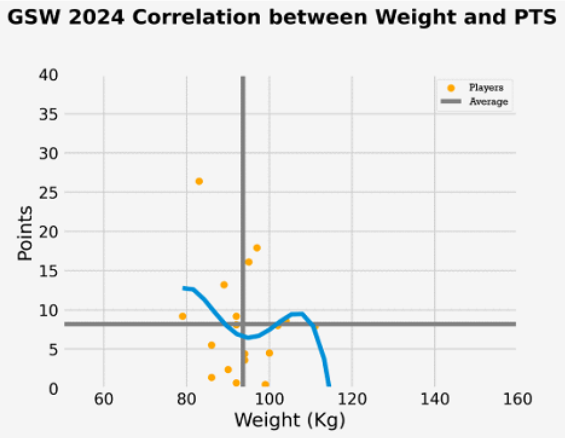
As for finding which table we needed, we ran the following line: wiz_per_game = wiz_soup.find(name = 'table', attrs = {'id' : 'per_game'}). Looping through, we'd add the needed attributes to a player dictionary and then add the dictionary to the wiz_info list once we finished looking at each row. Following that, we decided to take it upon ourselves to try to add in our own attributes of height and weight. Fortunately basketballreference's player bios are also fairly uniform, the player names on the per game table were links to their bios. As such what we did to extract from the HTML was look for patterns with the inspect element tool, gather the elements corresponding to height and weight and then add as additional dictionary entries to the player. Finally, with the ts variable, we would write to a new csv file.

When it came to finding correlations of our target attributes, it was not as grueling of a task as the web scraping. We correctly performed the correlation analysis on our target attributes and then utilized matplotlib to neatly graph the correlations. To perform the correlation analysis, we used the pandas library to read csv files which we then convert into dataframes to work with.

Results and Analysis:

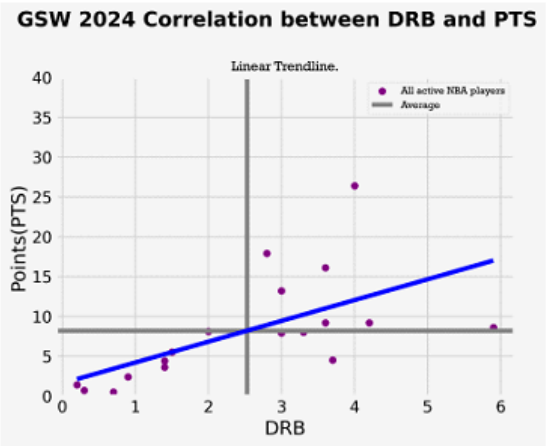
During our analysis, we were hoping to discover whether individual attributes such as age, height and weight exhibited correlation with overall player performance. We also looked into whether offensive or defensive rebounds had a stronger impact on the overall team's performance. Contrary to our original hypothesis that height and points scored would have a positive correlation and weight/age with respect to points scored would have a correlation resemblement to a normalized curve, we noticed that the correlations weren't always uniformly positive or negative. Based on our findings, we can conclude that correlation does not tell the full story, that it does not imply causation.

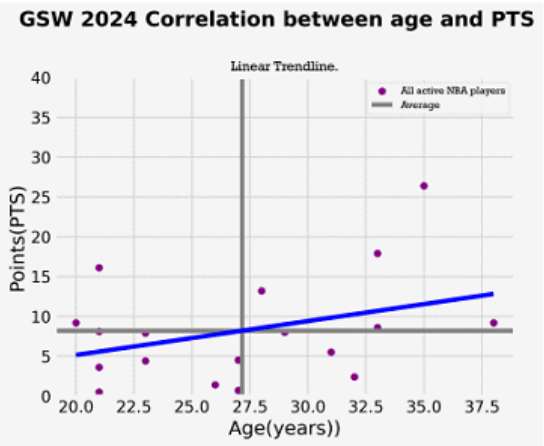


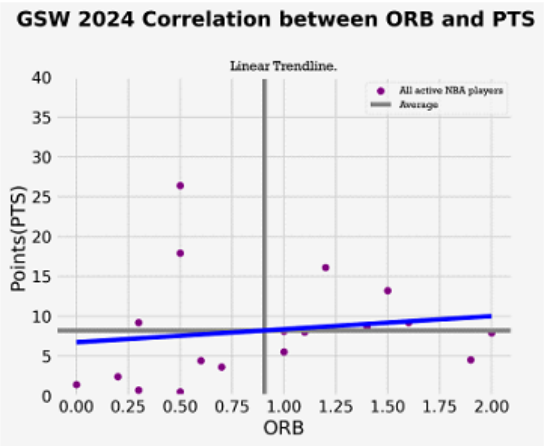


For the correlation between Weight and PTS, when we tested the Golden State Warriors, our findings were as we expected: The correlation between weight and points were negative, with a correlation coefficient of r = -0.17196385717733648. When correlation is negative, it implies that as one attribute increases, the other will decrease. However, when we tested the correlation between Height and PTS, we found that the correlation between the two attributes was r =-0.06370661268412987. Factors that may have impacted the coefficient value could include player experience, that just because someone is taller does not necessarily mean that they are more skilled in basketball and able to score more points.

With age, just measuring the two numerical values does not tell the full picture. Like one of the oldest players on the Golden State Warriors, Steph Curry at 35, scored the most points of the rest of his team. Then, looking at the overall GSW 2024 season, r=0.3537014858112542.

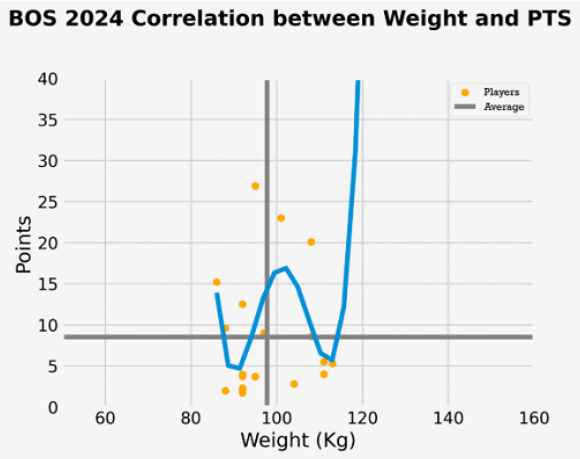


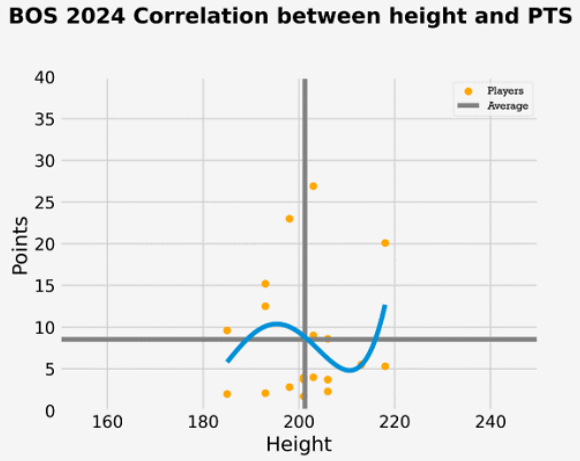




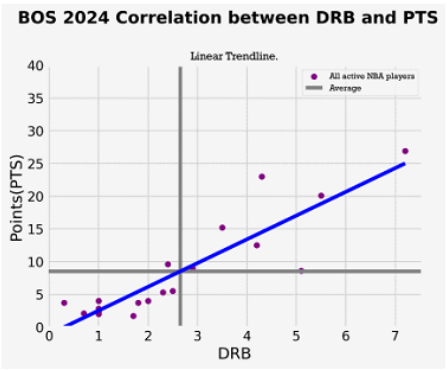
Defensive rebounds (DRB) and offensive rebounds (ORB) tended to have a positive correlation. What stood out to us though was that going into the specific coefficient values, it showed, as seen with the GSW 2024 season, that between DRB and PTS r=0.6066825382542943 while for ORB and PTS r=0.14602684388251608, these values going in contrast to our previous hypothesis that playing offensively would have a greater impact on points scored. The reason why this may have happened is because of factors such as providing

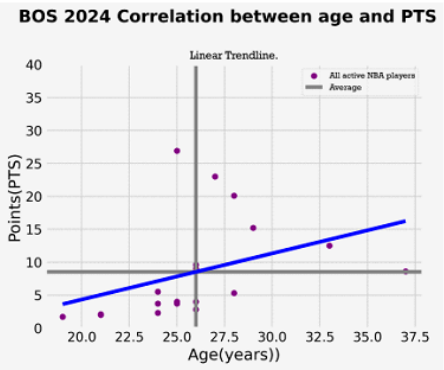
opportunity creation, control of pace, or quick game transitions, none of these could be obtained through correlation coefficients.

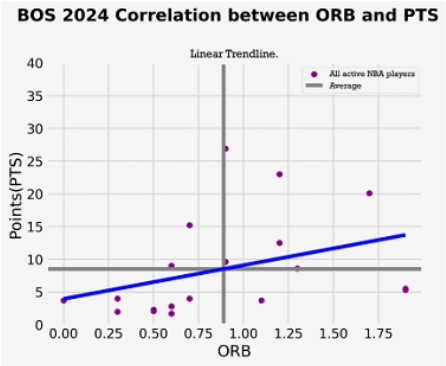




When we ran our analysis on the Boston Celtics for the 2024 season, we noticed a small, yet positive relationship between weight and points with a value of r = 0.05088254665309752. This positive correlation means that if a player weighs more, then they would be more likely to score more points than if they weighed less. The explanations for this value may not be fully obtained through numeric calculations, like maybe the players who weighed more had more opportunities to make a shot or they have more experience. We also noticed a small, yet positive correlation between height and points scored with an r value of r = 0.0840111510230281. This also means that a taller player will likely score more points on average than a shorter player. It must be stressed however that these r values are small, so our attributes in question cannot be considered strongly correlated.







In all three comparisons, we noticed a positive correlation between the attributes. Defensive rebounds and points scored can be considered strongly correlated with the r value of r = 0.8891594546364995. This is a high value for correlation and it shows that when players score more defensive rebounds, the team will on averager score more points. We noticed an interesting relationship when we compared offensive rebounds with points and age with points. These attributes had an r value of r = 0.362215889173508 and r = 0.37524610751256393 respectively. These values are still both positively correlated and their values are quite similar. This shows that both age and offensive rebounds have an effect on total points scored, although not as large of an effect as defensive rebounds do.

Conclusion:

Our original hypotheses were the following: Player height and points scored would always have a positive correlation while player weight/age with respect to points scored would have a varying correlation (light or too heavy weight has negative with pts, medium weight has positive). Then, we also hypothesized that teams would score more points overall if they had more offensive rebounds than defensive rebounds.

Our results are inconclusive when proving our first hypothesis. We noticed a negative correlation with the Warriors between weight with points scored and height with points scored. Surprisingly, the Celtics had a positive correlation with respect to both of these attributes. This shows that our measure of player performance or our assumption of player performance is not accurate. As it turns out, there would be more (non-numeric) factors to have considered, like player experience or opportunities to shoot the ball.

Our results clearly show that our second hypothesis was incorrect. It is clear, after analyzing both the Golden State Warriors and Boston Celtics in the 2024 season, that defensive rebounds led to more points scored by the team. With the Warriors, defensive rebounds were approximately 4.15x more correlated to points scored than offensive rebounds. This factor stays relatively high with the Celtics which had a value of 2.45x. It is important to note that in each case, the correlations of ORB and DRB were still positive which shows that both types of rebounds are important for scoring more points, however, defensive rebounds will lead to more points scored than offensive. The reasoning for this discrepancy in the findings versus our original results, much like height/weight and points, could be attributed to non-numeric results. It could be a number of factors such as player build, experience, and what type of shot scored.

Putting the findings of our two hypotheses together, we can conclude that correlation does not tell the full picture, that there are other non-numeric factors that can affect the outcome.

Appendix

Player performance: In our project we will consider player performance to be a measure of a variety of individual attributes. Eg. A player with more points scored and more rebounds would be more performant than a player who had less points and less rebounds.

Path name: A web page is composed of many different objects. These objects can be HTML files, images, audio files, etc.. To access any of these objects, the universal resource locator (url) must be known. A url has two main components to it, a host name and a path name. The host name provides the address to the website one is attempting to visit. The path name provides the address to a specific object.