**Offense vs Defense - Final Analysis**

**The code from the project was written exclusively by Jake Rosen**

**Part 1: Introduction and Research Questions**

We have decided to complete a statistical analysis on the 2020-2021 NBA regular season. Due to the recovery from the 2019-2020 Pandemic Season, this is the first time an NBA season has been shortened to 72 games (excluding lockout seasons). We have decided to choose this season as we feel it is the most interesting and relevant as it is the most recent NBA season with a full, complete and unchangeable set of statistics. The 2021-2022 NBA season is still ongoing and thus not as relevant as its predecessor. We have decided to look into, in this season specifically, what led to more overall team success, offense or defense? Some example research questions we have used so far for this analysis are below:

* How does a team’s offensive rating impact their overall win count or percentage?
* How does a team’s defensive rating impact their overall win count or percentage?
* How do specific team’s defensive statistics impact team success? Is it more important to make impact plays (I.e turnovers) or to limit shooting percentages
  + These stats will be broken up into Steals, Blocks, Defensive Rebounds, Forced Turnovers, etc.
* What teams are strongest offensively and defensively?
* How do our own offensive and defensive indexes correlate with win count?

The link to our team’s git repository for the project can be found here: <https://coursework.cs.duke.edu/rafael.adi/cs216-project>

The outlined questions specifically allow our team to focus on specific NBA players and teams that have a large set of data (for example they played a lot of games and several minutes in those games), and thus will provide us with the best sample when analyzing the best of the best during the 2020-2021 NBA season. We should be able to feasibly retrieve summaries for our questions as the diverse datasets described below outline every statistic for every player described above. Finally, the questions provide for more interesting and relevant information as it is following the biggest names and teams in the current NBA and focuses on the most recent full season set of data.

**Part 2: Data Sources**

Our data sources come from Sports Reference and more specifically, their Basketball Reference Subdivision. Links to the drawn data are on the bottom of this section. Each data set relates to the 2020-2021 NBA season and thus are relevant to the research questions about the NBA we have described above. We chose this season specifically, because it is the most relevant data that is not changing, since the NBA’s current data changes on a day-to-day basis. Each dataset is in our GitLab Repository Below and descriptions for each dataset are as follows:

* **player\_per\_game** - File that shows each NBA player that played in any regular season game for the 2020-2021 season and their standard average statistics per game (points, rebounds, assists, steals, etc.)
* **player\_advanced** - File that shows each NBA player that played in any regular season game for the 2020-2021 season and their advanced statistics (i.e. statistics that combine a wide variety of variables like effective field goal percentage, plus-minus, etc.)
* **player\_per\_100\_pos** - File that shows each NBA player that played in any regular season game for the 2020-2021 season and their average statistics scaled to if they were to play 100 possessions during every game.
* **team\_per\_game** - File that shows each NBA team and their standard average statistics per game (points, rebounds, assists, steals, etc.)
* **team\_advanced** - File that shows each NBA team during the 2020-2021 regular season and their advanced statistics (i.e. statistics that combine a wide variety of variables like effective field goal percentage, plus-minus, etc.)
* **team\_per\_100\_pos** - File that shows each team during the 2020-2021 regular season and their average statistics scaled to if they were to play 100 possessions during every game.
* **team\_rating\_and\_rankings** - File that shows each team during the 2020-2021 regular season and their team rating and record rankings at the end of the season.
* **Team\_per\_100\_pos** - File that shows each team’s opponents’ statistics scaled to if they were to play 100 possessions during every game

Player Data: <https://www.basketball-reference.com/leagues/NBA_2021_totals.html#totals_stats>

Team Data:

<https://www.basketball-reference.com/leagues/NBA_2022.html>

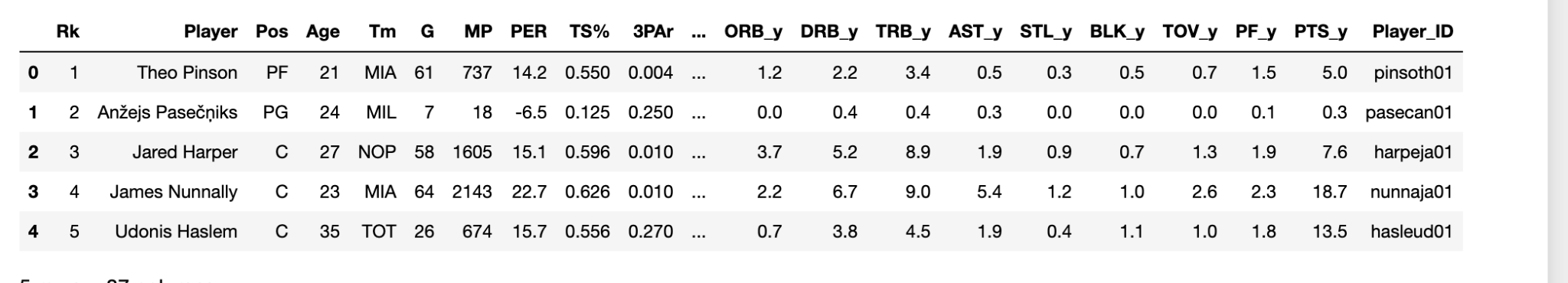
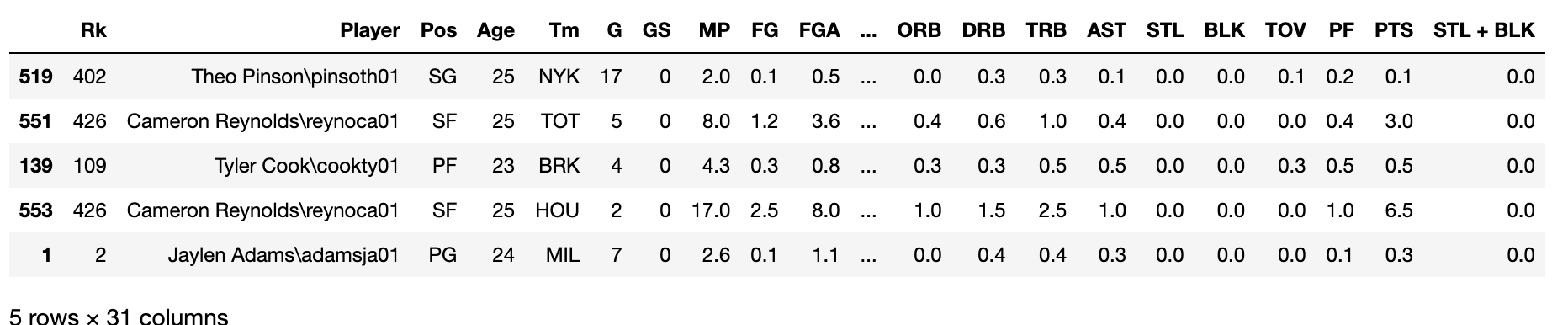
**Part 3: Results and Methods**

We began our project by importing data sets from basketball reference, a website that contains a variety of different data sets about the NBA. We used two types of datasets from this website, player specific data and team specific data.

Our player specific data is the player\_per\_game data set that contains per game stats from every NBA player and the player\_advanced dataset that contains advanced statistics on players. We then merged these datasets into a greater dataset called player\_data.

Our team specific data comes from the team\_per\_game data set that contains per game data on each team, team\_advanced dataset that contains advanced statistics for each team, and the team\_per\_100 data set that normalizes the data for each team to their statistics per 100 possessions. We merged this data into a dataset called team\_data. We also collect the dataset that keeps track of the per 100 possessions statistics by the opponents of each team that can be useful to identify how well their defense fared during the season.

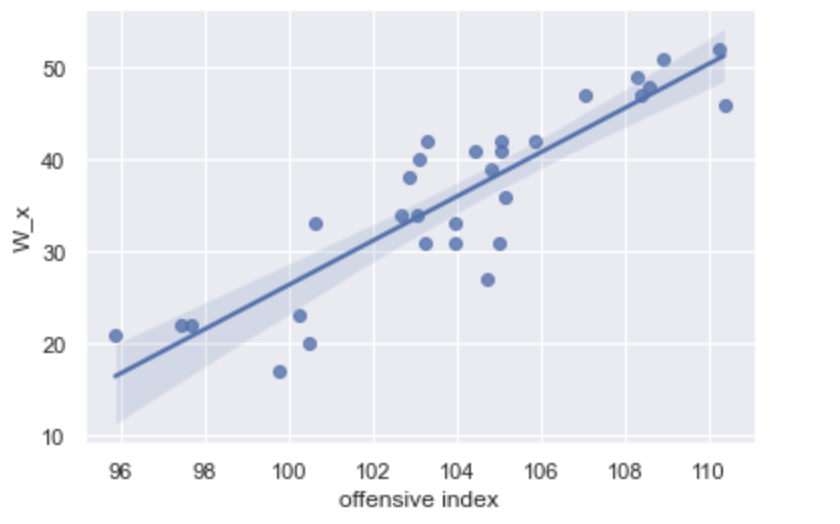
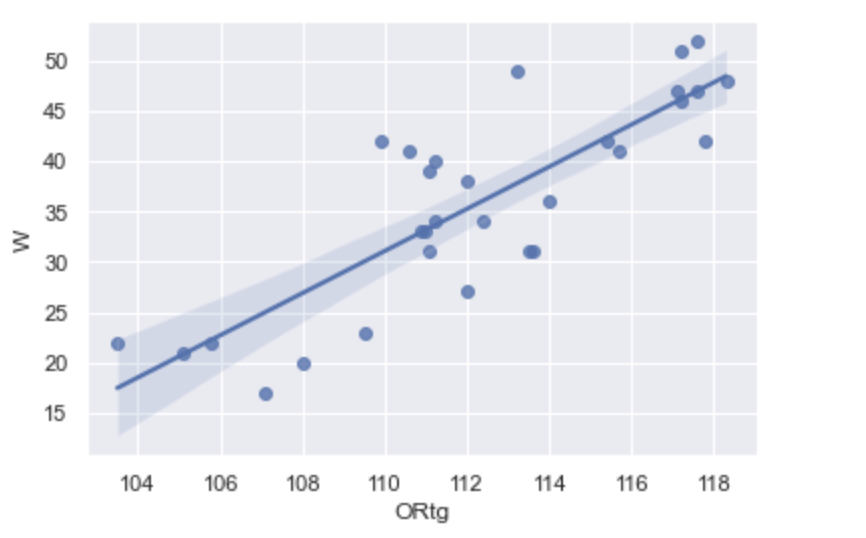
We began our exploration of the dataset by importing each individual csv file. Then, we began to aggregate the data into two distinct datasets, player\_data and team\_data. Some of the data needed to be cleaned in order to do this. For example, in the player data there was an ID code that followed the players name. We split these two pieces of data (the name and the ID) into two separate columns. We were able to do this by using the String split method to split the string at the back slash “\”.

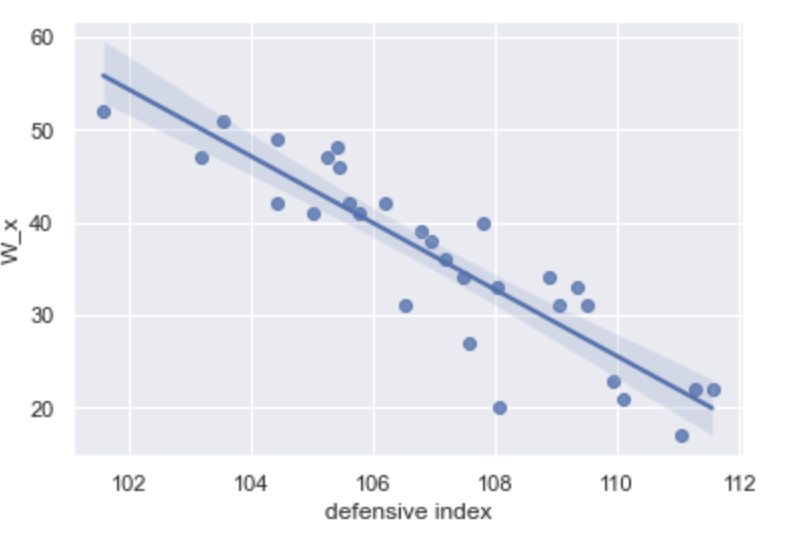
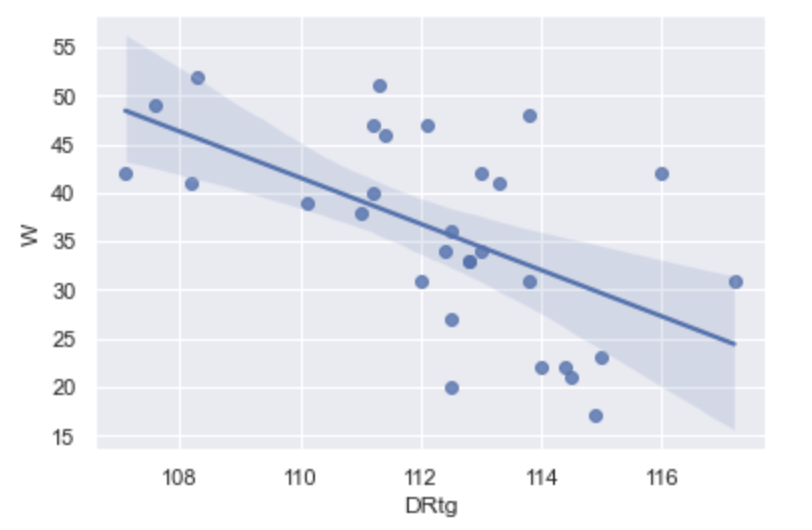


For our team data, some of the files contained an asterisk following the team name that we needed to delete to match it with the data from the other data set. To begin seeing the trends between defensive statistics and winning we created multiple graphs.

`In order to truly analyze the most important component aspect of a team’s game, we strived to create our own offensive and defensive indexes. We created graphs using the Seaborn library which allows us to graph data presented from a dataframe. We plotted our indexes against wins and compared them to other graphs where we plotted the NBA advanced statistics for offensive and defensive rating. Our index corresponded much more closely to the wins number than the traditional offensive and defensive rating that is compiled by basketball reference. Here is a comparison of the graphs

**NBA Advanced O/D rating**  **Our Offensive/Defensive Index**

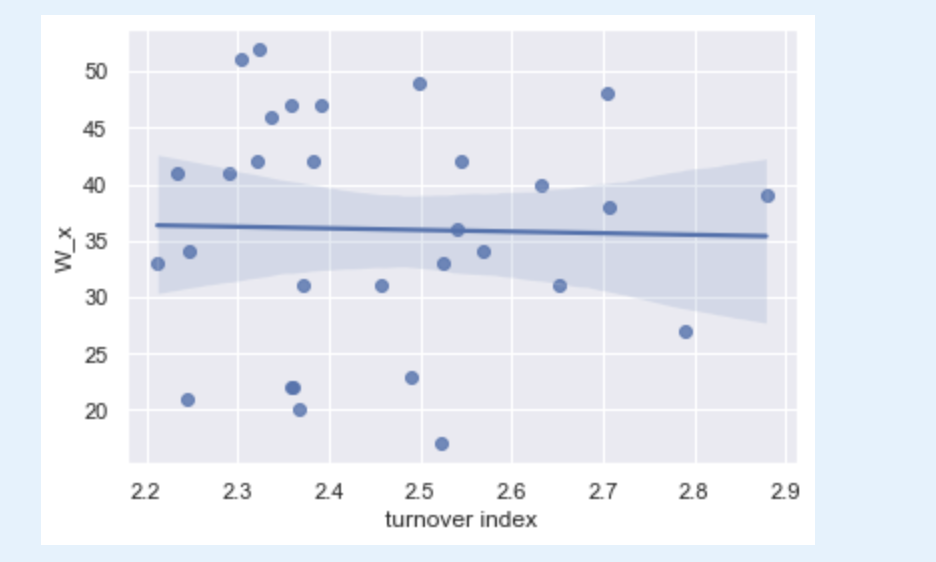




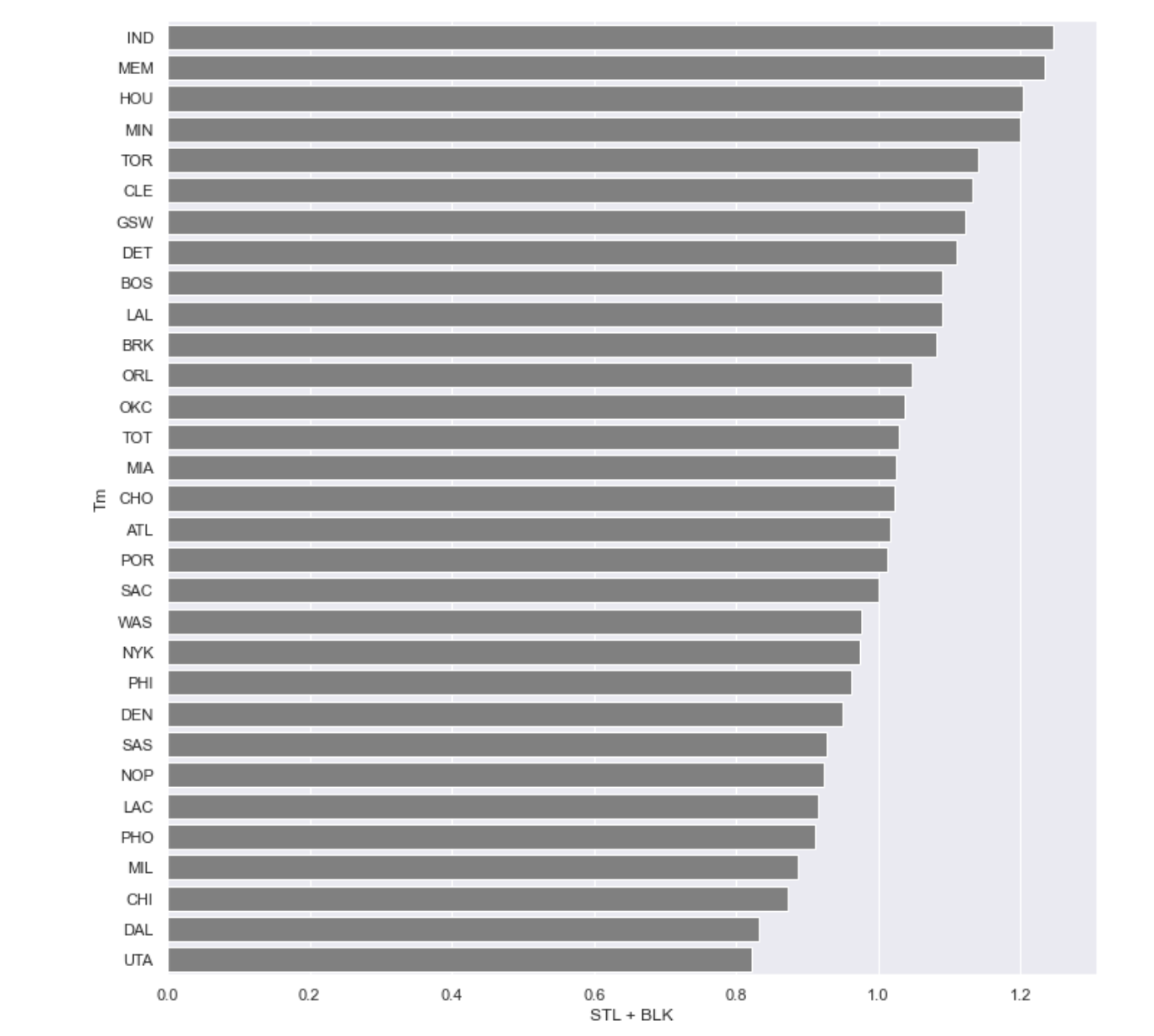
As you can see, the offensive and defensive indexes that we created show far less volatility when being used to directly compare against the total wins of a team than the traditional offensive rating and defensive rating used for advanced NBA statistics.

Our offensive and defensive indexes were created using the same components as each other, but we alternate between the team\_per\_100 data set and the opponent\_per\_100 set where it is fitting to do so. As in, for the offensive index we use the teams offensive shooting percentages while for the defensive index we use their opponents shooting percentages. The four main statistics used were points per game, 3 point percentage, 2 point percentage, and the advanced stat, effective field goal percentage that weighs three pointers more heavily (which we also do manually when we add 3 point percentage into the index). We normalize all of these values manually to the max value in the data set for each statistic. This value then gets multiplied by 25 so it can be within a similar range as the currently used “defensive rating”. When analyzing our data, we wanted to find which side of the ball had the greater impact on wins, offensive or defense. To discover this, we found the slope of the lines in our offensive/defensive rating versus wins graphs. For offensive rating, the slope was 2.41. On the opposite side, the slope of the defensive rating graph was -3.6 (negative because a higher defensive rating is a negative). Since the absolute value of slope of the defensive graph is higher, there is a steeper correlation between wins and defensive index which supports the claim that defense is the better indicator of higher win counts than offense.

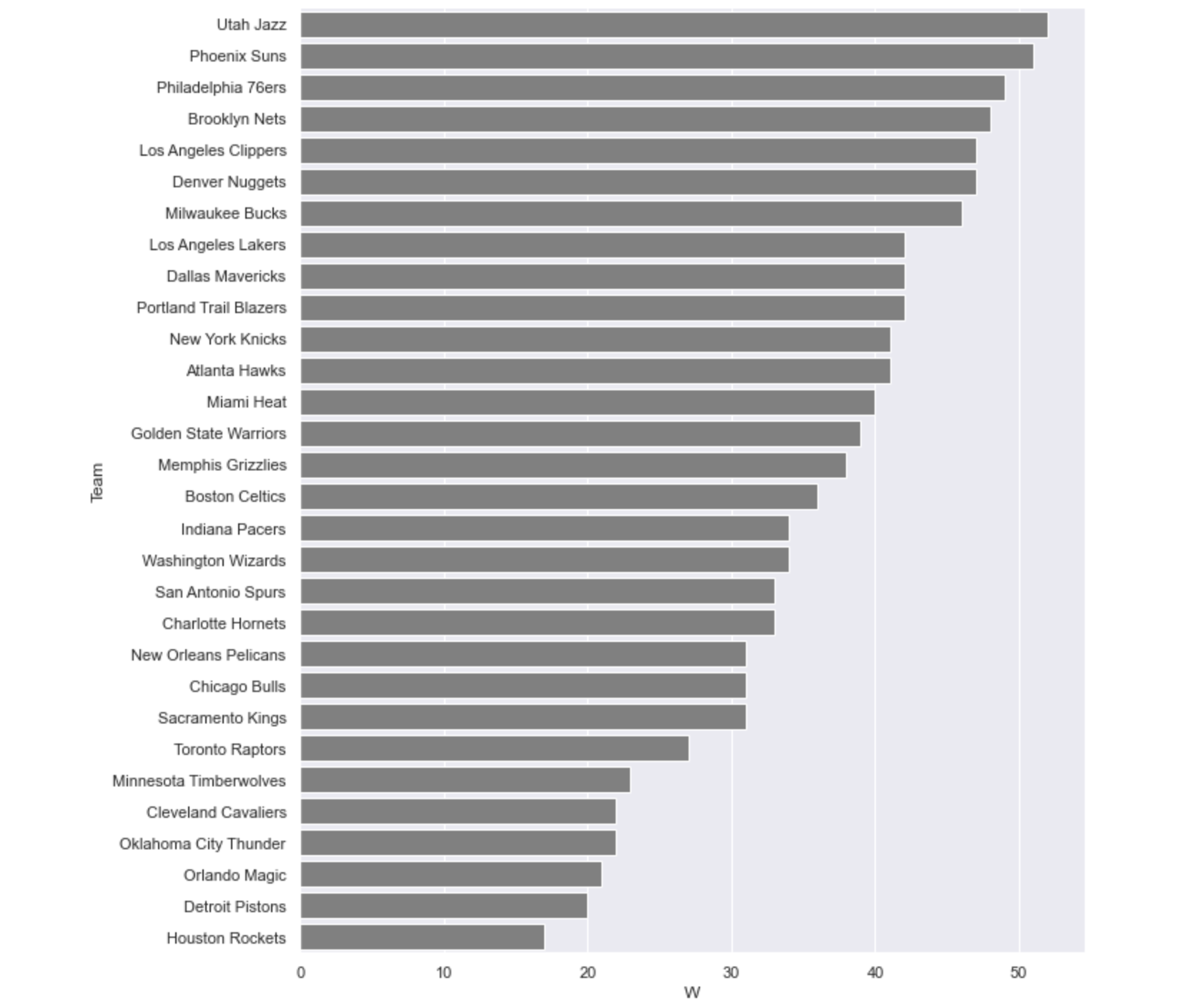
Our index does not include anything about turnover percentages or blocks because when added the data becomes diluted and has much more variability. As seen here when we map out a “turnover index” that we created which includes steals per 100 possessions, blocks per 100 possessions, and turnover percentage all normalized to the maximum of the data set we see almost no correlation between it and wins.



We then created a bar graph that visualizes the newly created column “STL + BLK” which is the number of blocks and steals per game. We created the new column by adding together the blocks per game and steals per game for each player in the data frame. We then grouped the players by team and found the average per game Steal + block number for each team. These values were then put into the bar graph.



While we were unable to place these values in the same graph, here is a bar graph visualizing the number of wins by each team. Immediately you can notice some clear opposites. For example, the team with the least STL+BLK, the Utah Jazz are also the team with the most wins. Additionally the pacers who lead the league in our STL + BLK category are bottom half in the league in wins. Thus, we find that there is a very weak correlation between turnover and block numbers and wins within this dataset.



**Part 4: Limitations and Future Work**

Our approach to analyzing the given datasets was a holistic one. We had to consider several variables, such as number of steals and blocks and shooting percentages, to create new statistics that could be used to measure an NBA team’s success as a whole. Some of these new statistics included defensive, offensive, and turnover indexes, which allowed us to draw new conclusions about these teams and their dynamics. While we tried to be holistic as possible, the datasets and methods we used pose some limitations to our research. A major limitation of our work was that the data we used is from only one season, the 2020-2021 NBA season. Although the goal of this project was partly to analyze the 2020-2021 NBA season, this means that our results were far from all-encompassing. The season we analyzed was also unique in that it occurred during the COVID year, so regulations and other factors at play may have impacted how certain teams played or which players were active throughout the season. Therefore, it is more difficult to make generalizations about team success across seasons if the data is isolated due to these circumstances. From our analysis, we can see there were clear relationships with offense and winning as was there with defense. However, we cannot be certain that the good defensive teams did not just happen to be better offensive teams or if there was another unforeseen factor that really led to their overall success. These uncertainties allow us to form strong follow-up research questions. For example, how would our indexes shift if we looked at multiple NBA seasons? What impact do individual players have on overall team success? Our future work would likely be conducting statistical analysis on similar data, but more encompassing of different factors that we previously discussed in our limitations. All things considered, while our own research was strong in that we considered multiple variables across several datasets, we could incorporate current limitations of only using one season’s data and not looking at individual player’s impact into our future work to make more holistic statistical analyses and inferences.

**Part 5: Conclusion**

In summary, we attempted to solve the age old question in basketball: what has a bigger impact, offense or defense? We chose the NBA because it is the highest level of professional basketball and has the deepest selection of data for us to analyze. When we set out to answer our research questions, the first one we wanted to tackle was offense. Our composite offensive rating consisted of: points per game, 2 point shooting percentage, 3 point shooting percentage, and effective field goal percentage in the 2020-2021 NBA season. Our defensive index consisted of the opponents metrics for all of these categories (i.e. points allowed, opponent field goal percentage etc). We have found substantial evidence that having a lower defensive index (better rating) correlates to more wins in the NBA than a high offensive index. Additionally, our metric was better at predicting wins than the offensive rating metric currently used by the NBA. Our defensive index was an excellent predictor of wins. In fact, it was such a strong predictor that it had a larger correlation with wins than our offensive rating did. That leads us to conclude that our defensive rating stat is more effective at predicting a teams number of wins. After making these findings, we decided to dig deeper into defensive metrics. First we looked at turnover data including turnovers per game, steals, blocks and more. We charted this data against wins and found no correlation between turning the ball over and winning basketball games. Our conclusion from our bar and line plots was that steals, turnovers, and blocks don’t actually affect winning. Ultimately, lowering an opposing team's offensive rating (our version) is the most important aspect of the NBA game and has the highest correlation with winning. In other words, defenses should strive to reduce opposing teams shooting percentages and do not necessarily need to focus on the turnover “big play” aspect of NBA defense in order to succeed.