IST 652 Final Project

Factors That Contribute to NBA Scoring

Jack Dolitsky, Bourama Sidibe, and Lagarthucin Legrand

Scoring is one of the most important skills to have in the NBA. It is the most fundamental component of the game. Afterall, the way to win the game is by scoring more than your opponent. Elite scorers in the NBA have come in a variety of different backgrounds, body types, and archetypes. Wilt Chamberlain, a 7-foot giant, was one of the leading scorers in NBA history, but rarely ever scored from outside a few feet of the basket. Stephen Curry has led the league in scoring, while breaking the record for most three-pointers made in a season. Our project aims to look at scoring trends, and what factors might contribute to being a good scorer. For the purpose of this project, scoring will be evaluated on points averaged per game over the course of the player's career. The dataset that we gathered was the "NBA NCAA Comparisons" project from dataworld.com. The original dataset includes almost every NBA player that was drafted in the late 1940's until 2017. It includes the player's career stats in the NBA as well as in the NCAA, if applicable. The dataset contains 4576 rows and 34 columns. The dataset contains rows for the player's name and URL (to link to their statistics pages). It contains rows for the height, weight, birthdate, college, and position of the player. The rest of the rows are the player statistics. This includes the players' three-pointers, field goals, and free throws, attempted and made per game, as well the percentage for each. There is also a row for effective field goal percentage, which weighs three-pointers as 1.5 times more than a two-pointers, which is a better efficiency metric than field goal percentage. These stats are used for both the player's NBA and NCAA stats.

The original dataset that we were going to analyze was a healthy aging dataset from data.gov. We read in the data and started to try to answer our data questions, but we ran into many issues. The main issue was there was no sample size data. This

meant we could not accurately compare any data, or do any proper grouping. Another problem was there was inconsistent data. Some questions such as “do you experience cognitive decline?” may have only included data for certain locations, genders, or ages, but not others. This led to us not being able to deduct any reasonable conclusions, so we then found a different dataset with more consistent data.

Since we are only looking at scoring in the NBA, the first transformation that we did was dropping all of the columns that had to do with the NCAA. We also dropped the column for URL, birthdate, and college, as it was not relevant for the analysis we were looking to conduct since we are focusing on their scoring in the nba.

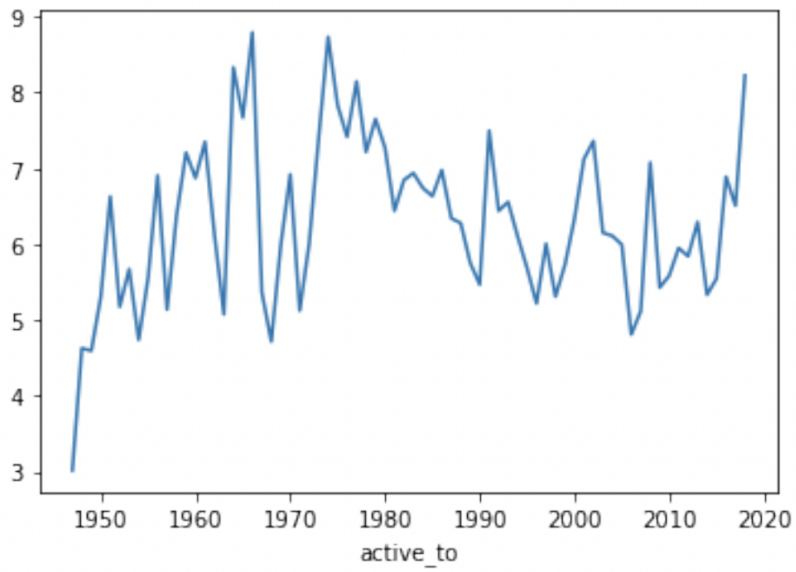
Analyzing position and height required some transformation as well. One investigation we were looking to conduct was whether having multiple positions had an impact on scoring. In order to do that we created a new column called “position class.” This column would tell us whether the given player had a singular position, or multiple positions. Following the creation of the “position class” column, we split it up into two more columns for regression analysis. One column was called “multi” and the other “single.” If the player had multiple positions, there would be a 1 in the “multi” column, and if not, then there is a “0.” This was the same rule applied to the “single.”
We did a similar transformation for height. The height groups we created were “under 6 feet,” “6-0 to 6-5,” “6-6 to 6-11,” and “7 feet and up.” This would give more data points for each group, as 20 different height points would not make for sufficient conclusions. We then did the same thing we did for positions, by creating 4 new columns that contained 0s and 1s for regression analysis.

In order to look at players who scored 20 points a game by position, we created a sub dataframe. This data frame contained all players who averaged over 20 points per game in their career. This was so we could easily compare this specific sub-group and

One of the first things we wanted to analyze when it came to NBA scoring is the scoring trends of different eras. As recently stated, our data set covers information about the NBA from around the creation of the league in 1946 until 2017. So we decided to look at the average points per game for every single player to ever qualify in the NBA during that time period. We analyzed this data by creating a plot with every year on the x-axis and points per game on the y-axis.

```
[133]: nba_by_year = nba.groupby("active_to")["NBA_ppg"].mean()
       nba_by_year.plot()
```

```
[133]: <AxesSubplot:xlabel='active_to'>
```



As you can see from the graph above, the highest average point per game that was had by the players in the NBA was around nine points in the mid to late 1960s, mid-to-late 1970s, and 2017. Scoring was at its lowest around the beginning of the creation of the

league so players simply just were not that good. Most players treated the league as their side job, and did not take it as seriously as today's players do. This is also because the NBA wasn't nearly as successful as it is today and players were not making nearly the money they are making today. They just didn't have as much incentive. As time went by and players became better and better, scoring trends started to increase which is why you can see it peaked around 1965. RIght after peaking, scoring dropped drastically in the late 60s early 1970s. It then peaked again at around 9 points near 1976. This can be explained by the fact that something major happened for the NBA this year. This was the year that the NBA merged with another major basketball company called the ABA. This merger combined the players from both leagues, bringing in a lot of new great players including all time great scorers like Julius Erving.

As years passed by, more and more talent began to join the NBA. In the 1980s, scoring decreased but averages still stayed higher than they were before the NBA/ABA merger. This can also be attributed to the fact that the three point line was added to the NBA in 1979. Before this, only two point field goals and free throws were recorded. Then came the 1990s. This time period is widely known as probably the most physical era of basketball there was. A lot less fouls were called and defenders got away with a lot more contact. This made scoring during this era incredibly challenging and averages began to drop. Which is also why this era is widely known as the era of the big man. Since the game was so physical, teams targeted the most physical players they can possibly get and usually played through them. Centers and forwards were the focal points of offenses simply because they were bigger. This in turn slowed the pace of the game tremendously which gave players less opportunity to score. During the mid

2000s, scoring began to increase again and has not looked back. This can be attributed to many different factors. The first being rule changes. The NBA realized the game was too physical and thought low scoring games were not as entertaining. They began to give defenders a lot less leeway by calling more fouls and eliminating things defenders were allowed to do. This made scoring much easier because it was harder to play defense. Along with the fact that players also were becoming better, the league also became a lot faster. Guards started to become the focal point of offenses along with the fact that teams incentives players to shoot more three pointers. Overall, this graph was very important. Although this graph did not lead us to any solidified conclusion, it was able to tell the story of how different basketball eras played a role in how well players were able to score.

The next data question was which positions in the NBA average the most points.

```
#Average ppg by position
position_points = nba.groupby("position")["NBA_ppg"].mean()
position_points.head(10)
```

```
position
C      5.157937
C-F    7.859361
F      5.364688
F-C    8.073590
F-G    9.426267
G      6.303084
G-F    8.053203
```

Here we see F-G average the most points, and Centers average the least. The most points are averaged among players with multiple positions. A possible reason for this is players that are more versatile are likely to have more ways to score and take advantage of position mismatches. In order to see if versatility played a factor in

scoring, we grouped the players into categories of whether they played multiple positions or a single position.

```
#Average ppg by group
group_position_points = nba.groupby("position_class")["NBA_ppg"].mean()
group_position_points
```

```
position_class
multi     8.275527
single    5.774019
Name: NBA_ppg, dtype: float64
```

From this we can see that players who played multiple positions averaged about 8.3 points per game, significantly higher than the 5.8 average of the players who played a singular position. In today's NBA this is what would be expected, because team's value multi-faceted players more than ever. LeBron James, for example, has won numerous awards, and is still being paid among the highest in the entire league, at the age of 37. James is known for being able to play every single position on the floor. Another example is Kevin Durant, who is almost 7 feet tall, but has the skills of a smaller guard, and has led the league in scoring multiple times. This grouping is especially interesting, because in the older NBA, teams were a lot more traditional and tended to keep players in their primary position, but as seen above, being able to play multiple positions can definitely have a positive effect on being able to score.

We also wanted to see the effects that height of the player had on scoring.

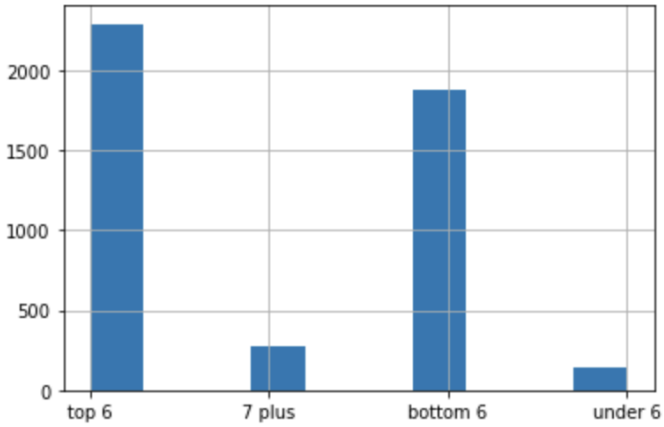
```
height_points = nba.groupby("height")["NBA_ppg"].mean()
height_points.head(30)
```

```
height
5-10     6.252083
5-11     5.418033
5-3      3.850000
5-5      8.900000
5-6      5.700000
5-7      2.983333
5-8     13.166667
5-9      7.780000
6-0      6.265333
6-1      6.349565
6-10     6.034462
6-11     6.255217
6-2      6.022936
6-3      6.542040
6-4      6.549275
6-5      6.578014
6-6      6.781728
6-7      6.778992
6-8      6.468531
6-9      6.541920
7-0      6.006707
7-1      5.279245
7-2      5.403846
7-3      5.861538
7-4      9.900000
7-5      0.750000
7-6     13.550000
7-7      6.200000
```

The most points averaged come from players who were 5-8 and 7-6, but this is not reliable because there were not many players at that height in NBA history. Most players were in the 6 foot range. Within the 6 foot range, the most points were averaged among players who were 6-6 and 6-7. A reason likely for this is smaller players tend to be more skilled but don't have the height advantage. These heights are a good blend of players who are likely skilled and have a relatively tall height. To see a more clear difference in scoring, we grouped all the players into 4 height categories. The categories were “under 6 feet tall” (under_6), “between 6-0 and 6-5” (bottom_6), “between 6-6 and 6-11” (top_6), and “7 feet and over” (7_plus). The following bar chart will show the distributions of these heights among the dataset.

```
#distribution of heights
nba["height_class"].hist()
```

```
<AxesSubplot:>
```



As stated previously, the vast majority of players fall into the 6-foot range. There are still about 250 players who were 7-foot and taller, and 150-200 players who were under 6-foot tall. The following code will show the scoring distributions among the height groups.

```
group_height_points = nba.groupby("height_class")["NBA_ppg"].mean()
group_height_points
```

```
height_class
7 plus      5.848519
bottom 6    6.415344
top 6       6.519066
under 6     6.013139
Name: NBA_ppg, dtype: float64
```

As seen here the players in the 6 foot range average more points that the more extreme heights (whether tall or short) even though individually, the highest averages were among players in those height ranges. This goes to show even further, that having a happy medium between height and skill is a major factor that goes into scoring.

Our main regression model looks at both position categories and height categories as the independent variables, with points per game as the dependent variable.

OLS Regression Results

Dep. Variable:	NBA_ppg	**R-squared:**	0.055
Model:	OLS	**Adj. R-squared:**	0.054
Method:	Least Squares	**F-statistic:**	66.05
Date:	Sun, 01 May 2022	**Prob (F-statistic):**	2.12e-54
Time:	21:45:37	**Log-Likelihood:**	-13483.
No. Observations:	4576	**AIC:**	2.698e+04
Df Residuals:	4571	**BIC:**	2.701e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.4989	0.281	19.543	0.000	4.947	6.050
top_6	0.1374	0.298	0.461	0.645	-0.448	0.723
bottom_6	0.4502	0.300	1.500	0.134	-0.138	1.039
multi	2.5514	0.159	16.060	0.000	2.240	2.863
under_6	0.2908	0.484	0.601	0.548	-0.657	1.239

We used 7_plus as our dummy variable for the height group and single as our dummy variable for the position group. We see here that the only significant variable here is multi, which also has the highest coefficient at about 2.5. The second highest coefficient is bottom_6 with 0.45, but it just fails out of the 0.1 p-value (10% significance level) and has 0 in its confidence interval, which means it is not something to rely on, but it is still worth noting. The regression as a whole only has a 0.054 adjusted R-squared, which is very low. Overall, there is no set conclusion we can derive from this regression, but we do see some interesting notes again on the effect of being able to play multiple positions, and having a relatively normal height have on scoring in the NBA.

One of the ways that we wanted to define whether a player is a good scorer is to see whether or not they averaged 20 points. So the next thing that we wanted to determine is which position has the most 20 point scorers.

position	bottom_6	top_6	7_plus
C	10	10	10
C-F	6	6	6
F	11	11	11
F-C	7	7	7
F-G	12	12	12
G	19	19	19
G-F	6	6	6

Here as you can see, we decided to create a sub data frame that contains only players who averaged 20 points or more during their career. We did so while keeping the players grouped by position. The table above shows that the G position has the most total 20 point scorers at 19. They are followed up by the F-G position at 12 and the F position at 11. But this table does not tell the entire story and we knew that there was still more digging to be done. Next we decided to use the “count()” function in order to count the total number of NBA players. Again, the players were grouped by their listed position.

position	bottom_6	top_6	7_plus
C	504	504	504
C-F	219	219	219
F	1297	1297	1297
F-C	390	390	390
F-G	217	217	217
G	1589	1589	1589
G-F	359	359	359

This table gives our previous table a lot more meaning. As you can see here, the G position has had the most total players in NBA history. This can be explained by the fact that most people on the planet or on the bottom_6 in height and players this height mostly play the G. So this can explain why that specific position has the most 20 point scorers, partly because of the fact that they simply had the most players. This doesn't necessarily mean that this position has the best scorers. Although, their work can't be all the way discredited. Being a smaller player in the NBA is usually a disadvantage so players this size usually have to be a lot more skilled than bigger players.

Since neither graph told the entire story, we decided to analyze one last thing when it comes to 20 point scorers in the NBA. Next we decided to analyze which position has the highest percentage of 20 point scorers. This can prove more useful because it may show which positions are scoring at a higher rate. We assumed it would make the total number of players less of a factor.

```
[4]: #finding percent of each position that averaged 20 points
position_percent = pd.concat([twenty_point_scorers['name'],␣
 ↪position_count['height']], axis = 1)

position_percent['percent'] = position_percent['name'] /␣
 ↪position_percent['height'] * 100
position_percent
```

[4]:

position	name	height	percent
C	10	504	1.984127
C-F	6	219	2.739726
F	11	1297	0.848111
F-C	7	390	1.794872
F-G	12	217	5.529954
G	19	1589	1.195721
G-F	6	359	1.671309

As you can see from the code above, we used "height" to represent the total number of players, "name" to represent the total number of 20 point scorers, and "percent" to represent the percentage of 20 point scorers in each position. Although the G position

had the highest total number of 20 point scorers, the table above shows that the F-G position has the highest percentage of 20 point scorers at about 5.5 percent and no other position comes close. The next closest position being C-F at 2.7 percent. The total number of players evidently plays a large role again because it can explain why the F-G position has the highest percentage. This position has had the least total number of players. This still can not take away from the fact that the F-G may be the best overall scoring position. The G position has 19 total players averaging more than 20 points per game and about 1500 total players The F-G position has had almost 1200 less total players, but still has a total of 12 players with 20 points. Although this position is greatly outnumbered in players by other positions, it still has the second most 20 point scorers. This shows how unique these types of players usually are. They usually have the size and athleticism of players who play the F position combined with the speed/skill of players who play the G position, creating a great scoring threat.

Another thing we wanted to look at were how different variables in the dataframe correlated with scoring so we could take a deeper dive into those variables.

	NBA_ppg
weight	-0.018133
active_from	-0.046379
active_to	0.087976
NBA__3ptapg	0.427647
NBA__3ptpg	0.429600
NBA__3ptpct	0.176916
NBA_efgpct	0.321084
NBA_fg%	0.342878
NBA_fg_per_game	0.991255
NBA_fga_per_game	0.968965
NBA_ft%	0.351308
NBA_ft_per_g	0.902089
NBA_fta_p_g	0.877710
NBA_g_played	0.720605
NBA_ppg	1.000000
single	-0.231272
multi	0.231272
under_6	-0.015152
bottom_6	-0.001138
top_6	0.020562
7_plus	-0.030297

Every number displayed above below shows how its corresponding column correlates to points per game. The highest correlation here is NBA_fg_per_game, which just means how many field goals (shots) were made per game, so this is self explanatory for why the correlation is so high at 99%. Two interesting variables to note are NBA_3ptpg and NBA_ft_per_g, which are how many three-pointers and free throws the player made per game, respectively. three-pointers having a correlation of 43% including the years before the 3-point-line was introduced means that the number is probably higher. A large amount of points that were scored in the NBA in this time period, are scored with 2-point field goals, so seeing free throws made have 90% correlation and three-pointers having at least a 43% correlation is very notable.

```
#three pointers made per game among positions
three_point_scorers = three_point_era.groupby("position")["NBA__3ptpg"].mean()
three_point_scorers.head(10)
```

```
position
C      0.024642
C-F    0.116129
F      0.257423
F-C    0.106341
F-G    0.573404
G      0.500202
G-F    0.540323
```

Here we see centers average the least amount of three-pointers with 0.02 per game, while F-G average the most with 0.57 per game. This is consistent with what we saw for points averaged per game among positions. Since three-pointers are worth 50% more than a two-point shot, it is an extremely valuable tool for a player to have. In today's NBA, players are attempting more three-pointers than ever before in history, so it is even more valuable to have in today's game, even among centers, which is not included in this dataset.

```
#free throws made among each position
free_throw_scorers = nba.groupby("position")["NBA_ft_per_g"].mean()
free_throw_scorers.head(10)
```

```
position
C      1.104762
C-F    1.673973
F      1.059985
F-C    1.747179
F-G    1.967742
G      1.225110
G-F    1.630362
```

F-G again averaged the most free throws made per game with 1.97, while pure forwards averaged the least with 1.06. The top 4 positions with the most free throws are all multi positions, which is very consistent with what we have seen up to this point. A possible reason for this is that these players are more unpredictable due to having a varied skill set, and can draw more fouls, which seems to be an extremely important part of scoring as seen from our correlation matrix.

Overall, there is no definitive conclusion we can draw about what exactly is the important factor in scoring. There are a multitude of reasons for this. Firstly, scoring is such a vast component of the game, and can happen in so many different ways, that you can not mark it down to one specific factor. Some players thrive from deep range, others with 2-point midrange shots or layups, and some excel at drawing foul calls and getting to the free throw line. Further, this dataset is missing many other components to scoring. Usage percentage is the amount of possessions a given player ends by either scoring, missing a shot, getting a turnover, or drawing a foul. This is a major component to scoring because, in simpler terms, it is how much a player has the ball in their hand, and it is not included in the dataset. Passing statistics are also not included in this dataset. Although intuitively, passing has nothing to do with scoring, the threat of being able to make a pass to anywhere makes the defense adjust to block the pass, leading to easier scoring opportunities.

Although there was no set conclusion, we got many meaningful insights. We saw Forward-Guards consistently thrived in all of the scoring categories, and more so players that play multiple positions. In the linear regression, we saw that multiple positions had by far the highest coefficient. Being able to have the height of forward, but also having some of passing and ball handling skills that are more typical among smaller positions (guards), can lead to endless scoring opportunities. If you can shoot the three, it brings defense out, which makes it easier to get inside to get an easy layup or draw a foul to get free throws. The players that have multiple positions usually mean they have the height of one position, but have the skills to also play another position. We see F-G being better scorers than F-C in general, because guards are usually more

focused on scoring while centers are usually more focused on defense. Overall, If an NBA team were in need of more scoring on the team, our analysis has shown that a good place to start would be to look at versatile multi-position players in the 6-foot range, who have a knack for getting to the free throw line.

If we had more time to do more extensive research, we would look more extensively at different eras in the NBA. The game is always changing and evolving, so the way player's score in today's NBA could be completely different than 30 years in the future or in the past. We also would be interested in using the dataset to look at the NCAA data, to see how certain variables translate from college to the NBA. This would help teams in need of a scorer see if free-throws average in the NCAA correlated highly with free-throws averaged in NBA, for example. In other words, teams in need of scorers could use that analysis to see if the factors we deemed important in the NBA, were able to be predicted from their college career.

References

"NBA NCAA Comparisons - Project by BGP12." *Data.world*, 6 Feb. 2020, https://data.world/bgp12/nbancaacomparisons.