College Basketball Stats Translating to the NBA

By Andrew Wolfe

I. Motivation

The overall theme for this project was to analyze college basketball statistics and see how they translated to the NBA from 2009-2021. The first goal of the project was to see which statistics translated from college to the NBA. The second goal of mine was to see which college basketball statistics contributed most to NBA Win-Share. NBA Win-Share is based on offense, defense, and playing time statistics. It is worth one-third of a time win (a 40-win team has 120-win shares to distribute to its players). My initial hypothesis was that shooting statistics, such as EFG%, FT%, and 3P%, would translate positively to the NBA and would have great correlations to a higher win-share percentage. This was my reasoning because it would seem to make sense that if someone can shoot the ball well in college then they would be able to translate it to the NBA. Lastly, I am a manager for the Men's Basketball Team at Michigan, explaining my motivation for the project!

II. Data Sources

Draft-data-20-years: This CSV is a dataset that reveals the draft pick, college, and NBA statistics of every NBA player (that was drafted) since 1990. The CSV sheet can be read easily through pandas in python on Jupyter Notebook. There are 1868 rows and 26 columns in the initial dataset for manipulating it. The statistics in it are the player's NBA statistics. The important variables that I used were the following:

- **"Player":** the name of the player drafted
- **"FG%":** Field goal percentage in the NBA
- **"FT":** free throw percentage in the NBA
- **"WS:** Win shares in the NBA
- **"WS/48":** Win shares in the NBA per 48 minutes (1 game)

College Basketball Players 2009-2021 CSV: The second dataset I used had information containing college basketball statistics. The dataset contained college basketball players' statistics from 2009-2021. The CSV can also be read through pandas in python on Jupyter Notebook. There are 61,061 rows and 66 columns in the initial dataset before manipulating it and the statistics were the college

statistics for each player. The variables that I used can be described as the following:

- **"player_name":** the name of the player
- **"conf":** The conference the college player is in
- **"TS_per":** True college shooting percentage of the player
- **"eFG":** College effective-field goal percentage
- **"FT_per":** College free throw percentage
- **"Rec Rank":** High School Recruiting Rank of College Player
- **"TP_per":** College 3-point percentage
- **"ORB_per":** College offensive rebounding percentage
- **"DRB_per":** College defensive rebounding percentage

III. Data Manipulation

To manipulate the data, the first thing I needed to do was load the packages, such as pandas, NumPy, seaborn, and matplotlib. The next step was loading the two datasets, 'draft-data-20-years.csv' and 'CollegeBasketballPlayers2009-2021.csv'. After this was done, I had to merge the two datasets through left-join so that I could have a dataset with the player's name, their college basketball statistics, and their NBA statistics.

The first problem that I ran into was that I had too many columns. Because of this, I had to start dropping columns on my dataset using pandas. This included the columns ''ht', 'MPG', 'College', 'usg','GP', 'TOTMP', 'TOTPTS', 'dunksmade', and 'dunksmiss+dunksmade' and more. In my mind, this was extremely important because it gave me a clearer picture of what variables I was working with. Another problem that I encountered was that some of the column values were filled with NA sections. Because of this, I had to delete all of the "NA" rows from 1990-2009 (the gap in the two datasets). Now that I had a dataset with no missing data, I could go on ahead and begin to dive into it.

The first thing I wanted to take a deeper look into was how conferences affected a player's chances of entering the NBA. By filtering and only referring to Power 5 conferences, the B10, ACC, SEC, B12, and Pac-12, I discovered which conference had the highest number of NBA drafted players since 2009. I used a pie chart for this part of the project. Next, I wanted to see the average high school recruit rank of each player in each conference. After that, I began to see if there was a change in three point and free throw percentage from college to the NBA. To do this, I had to create another variable that was equal to the NBA Average

number and subtract it from the College Average number. Bar graphs were made to display these relationships and I changed the matplotlib style to 'classic'.

Another aspect that I investigated was how NBA Win-Share per 48 minutes varied within each of the power five conferences. To find these values, I used the group by function and then calculated the mean NBA Win-Share per 48 Minutes from 2009-2021. A bar chart was used to show this relationship. Now that I was finished looking at how conference play affected NBA players, I wanted to see the independent correlations with all the variables, more particularly 'NBA W/S 48' and 'NBA_FGPer'. I used the .corr() function and once identifying the highest correlations, I created a scatterplot to more clearly show these relationships. Seaborn was used to depict this. The purpose of this was to discover which college statistics translated to the NBA and to discover which statistics contributed most to winning games in the NBA. However, I needed to keep in mind that "correlation is not causation".

The last piece of information that I wanted to investigate was if I could create a linear model for any of the NBA and college statistics. As time went on, I discovered a potential relationship between NBA Field Goal Percentage and college offensive rebounding and two-point percentage. For this, I had to use the 'sn.ols' function to find my p values and r-squared valued and import statsmodels. I did not consider the linearity, normality, and independence for the linear model and this does need to be investigated in the future to see if a true relationship exists.

IV. Analysis and Visualization

As mentioned in Section III, the first thing that I investigated was what college conferences NBA players were coming from. As seen below, the highest percentage of drafted NBA players comes from the ACC, followed by the SEC and B10. This is shown in Figure 1.

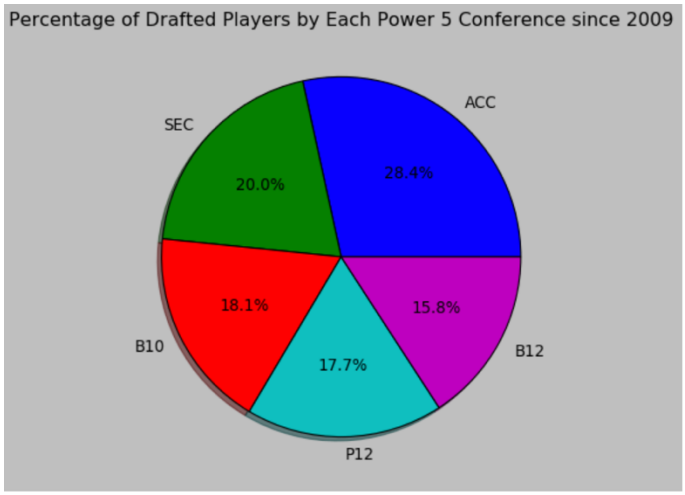

Figure 1

As shown by the pie chart above, about 28.4% of the players in the NBA that played college basketball in the Power 5 (since 2009) came from the ACC. The SEC had about 20% and the B10 had 18.1%. These numbers are significant because they reveal the power and stability of the ACC conference over the last 12 years in College Basketball. Moreover, I wanted to see the reasoning behind this number and thought it would be interesting to see the average high school recruiting rank for each player in each conference when coming into college. Looking at Figure 2 below, we can see that the Power 5 Conferences have been relatively receiving equal talent from high school. The B12 and ACC lead the way here for the average rank of their high school basketball recruits. One interesting note here is that the B12 has received the best recruits on average for the past 12 years, yet has the

least to show for it translating to the NBA level.

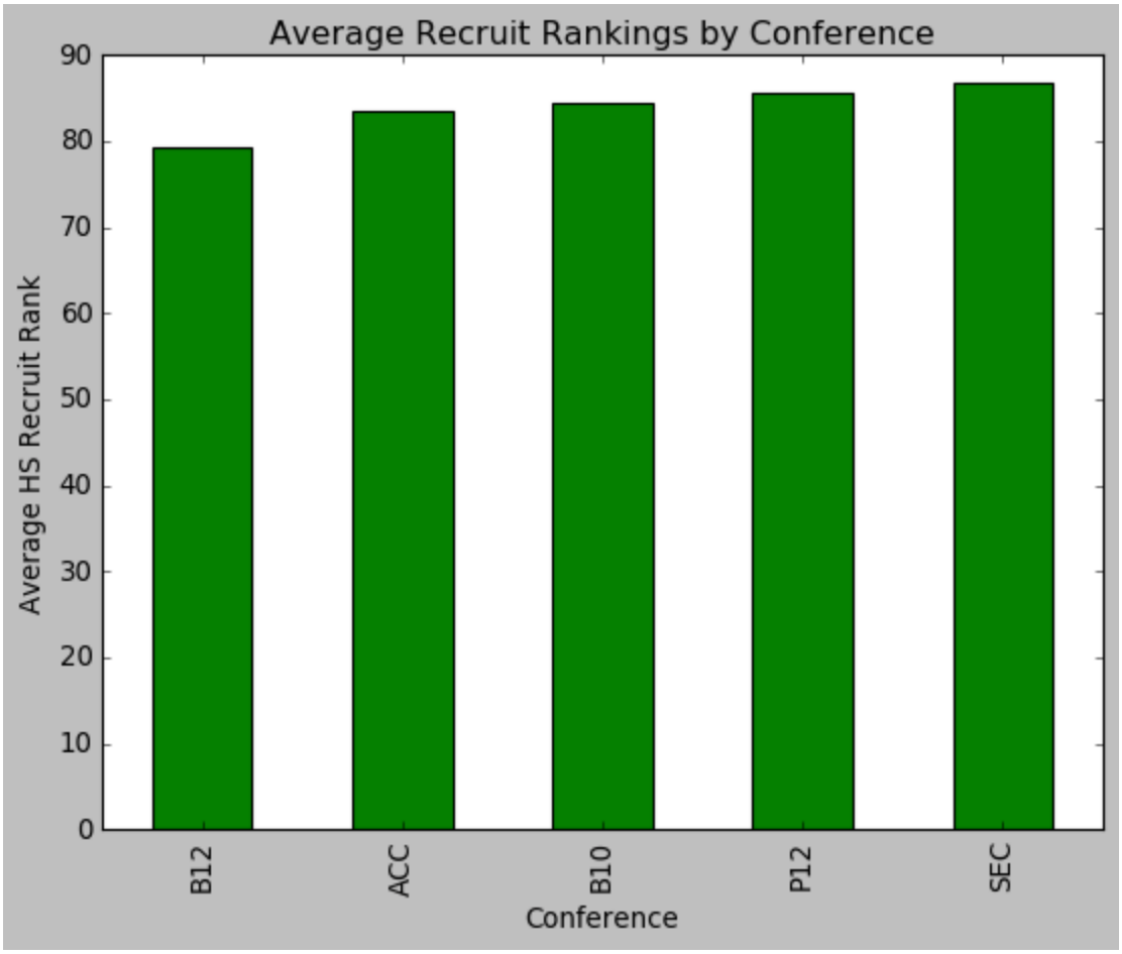

Figure 2

Moreover, I examined the change in free throw percentage and three-point percentage in college and the NBA. By looking at Figure 4 and 5 below, we can see that on average free throws and 3-point percentage for players relatively stayed the same when going from college to the NBA as the numbers changed by less than 1 percent.

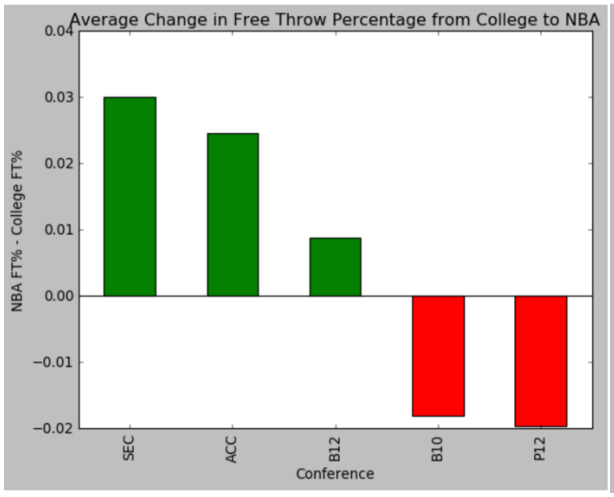

Figure 3

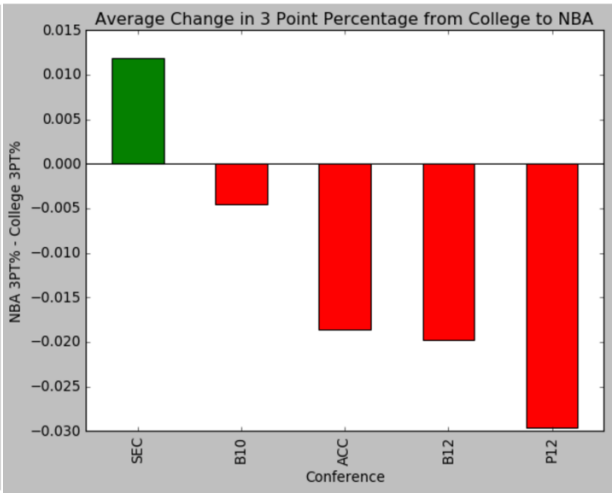

Figure 4

Additionally, when looking at Win-Share per 48 Minutes for each Power 5 Conference and their players in the NBA, we obtain a similar pattern to Figure 2 and Figure 1. When looking at Figure 5, we can see that the lowest average number of Win-Shares per 48 minutes are players that played in the B12 Conference.

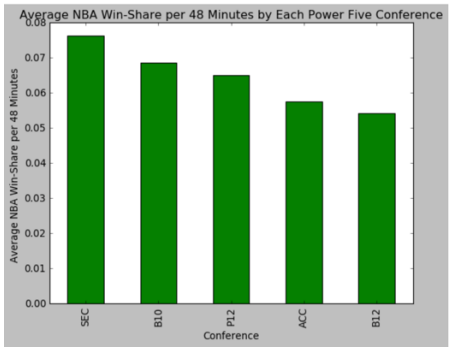

Figure 5

While looking at the average NBA W/S per 48 Minutes, I also calculated the correlations for all the variables in my dataset:

	Ortg	eFG	TS_per	ORB_per	DRB_per	AST_per	TO_per	FT_per	twoP_per	TP_per	...	VORP	NBA_DraftYr
Ortg	1.000000	0.732599	0.816095	0.152425	0.072803	-0.013924	-0.609490	0.359144	0.512113	0.203184	...	0.105637	-0.014026
eFG	0.732599	1.000000	0.934496	0.334443	0.273252	-0.233268	-0.184029	-0.017497	0.815880	0.011825	...	0.081062	0.042627
TS_per	0.816095	0.934496	1.000000	0.241712	0.225757	-0.135871	-0.218822	0.259470	0.730894	0.115520	...	0.084461	0.049365
ORB_per	0.152425	0.334443	0.241712	1.000000	0.684711	-0.515348	0.017329	-0.437351	0.499871	-0.433108	...	0.159525	-0.175182
DRB_per	0.072803	0.273252	0.225757	0.684711	1.000000	-0.331623	0.028906	-0.247639	0.358395	-0.222946	...	0.186235	-0.052242
AST_per	-0.013924	-0.233268	-0.135871	-0.515348	-0.331623	1.000000	0.175323	0.232212	-0.283797	0.258558	...	0.075774	0.025459
TO_per	-0.609490	-0.184029	-0.218822	0.017329	0.028906	0.175323	1.000000	-0.259334	-0.058421	-0.197179	...	0.028674	-0.023571
FT_per	0.359144	-0.017497	0.259470	-0.437351	-0.247639	0.232212	-0.259334	1.000000	-0.193693	0.442380	...	-0.058217	0.137556
twoP_per	0.512113	0.815880	0.730894	0.499871	0.358395	-0.283797	-0.058421	-0.193693	1.000000	-0.235736	...	0.138534	0.012057
TP_per	0.203184	0.011825	0.115520	-0.433108	-0.222946	0.258558	-0.197179	0.442380	-0.235736	1.000000	...	-0.116573	0.099103
blk_per	0.008148	0.274971	0.182358	0.639269	0.554454	-0.473771	0.124247	-0.387069	0.410361	-0.427120	...	0.117940	-0.085429
stl_per	-0.100069	-0.150354	-0.137181	-0.197786	-0.204646	0.383109	0.181841	-0.009393	-0.137482	0.088394	...	0.126655	-0.005529
ftr	0.093760	0.162552	0.275930	0.459069	0.322638	-0.087381	0.193869	-0.212738	0.257063	-0.280487	...	0.175391	-0.182546
porpag	0.706992	0.423735	0.571208	-0.056952	0.037088	0.350708	-0.465644	0.448097	0.243051	0.299652	...	0.124829	-0.092851
adjoe	0.845471	0.587833	0.718539	0.153601	0.155078	0.184555	-0.549709	0.394352	0.407803	0.212486	...	0.126165	-0.030348
year	0.082149	0.116654	0.148102	-0.150252	0.003257	0.071206	-0.064492	0.192690	0.063967	0.136187	...	-0.269545	0.952378
Rec Rank	0.019454	0.015575	0.014944	0.072253	0.074168	-0.030349	0.043576	-0.069066	0.017553	-0.090521	...	0.086927	-0.082249
ast/tov	0.266589	-0.137933	-0.085241	-0.549872	-0.431317	0.723388	-0.237516	0.262100	-0.247555	0.314033	...	0.037853	0.068029
pick	-0.161725	-0.195094	-0.191076	-0.171697	-0.094118	-0.019327	-0.021830	0.060302	-0.188695	0.053644	...	-0.288437	0.124113
drtg	0.035037	-0.085403	-0.027833	-0.447287	-0.523749	0.231486	-0.112420	0.279200	-0.156274	0.217955	...	-0.231244	0.262400
adrtg	0.015416	-0.093182	-0.033202	-0.448712	-0.511028	0.233284	-0.103320	0.283936	-0.163234	0.245959	...	-0.240017	0.249299
dporpag	0.284218	0.129365	0.219653	-0.048496	0.182107	0.286597	-0.168050	0.222394	0.038544	0.196776	...	0.176045	-0.084699
Pk	-0.154751	-0.189359	-0.185793	-0.175460	-0.097582	-0.016099	-0.027853	0.058036	-0.182251	0.052513	...	-0.284926	0.122422
NBA_FGPer	0.131894	0.322410	0.239594	0.493759	0.377627	-0.294018	-0.012547	-0.289815	0.439508	-0.374753	...	0.248960	0.009960
NBA_3P%	0.078137	-0.096059	-0.022219	-0.338057	-0.328146	0.142016	-0.161705	0.330066	-0.140857	0.355136	...	0.079827	0.076138
NBA_FT%	0.093564	-0.096259	0.012204	-0.340875	-0.344585	0.229050	-0.078101	0.380228	-0.168437	0.193002	...	0.123773	-0.009857
NBA_WS	0.105099	0.085586	0.083418	0.222046	0.186234	0.011458	-0.003726	-0.103125	0.154036	-0.159999	...	0.891421	-0.460053

Figure 6

This piece of information was very telling to me because it explained how there were high correlations between NBA Win Share per 48 Minutes and college offensive rebound percentage, defensive rebound percentage, and two-point percentage. Another interesting relationship shown in the correlations in Figure 6 was two-point percentage in college and NBA Field Goal Percentage. When creating a scatterplot for each of these relationships, we can see this more clearly.

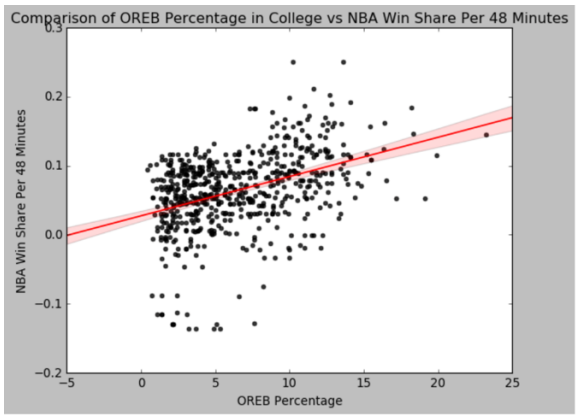

Figure 7

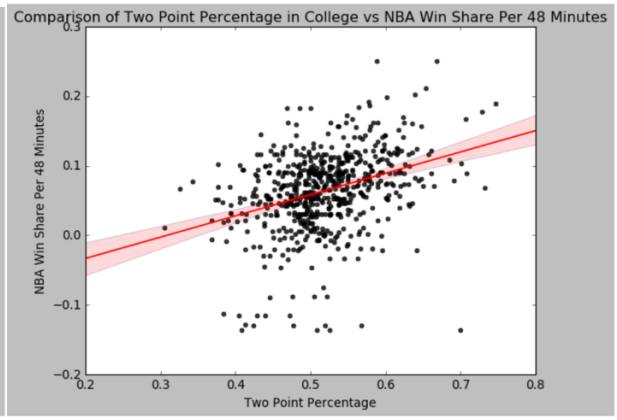

Figure 8

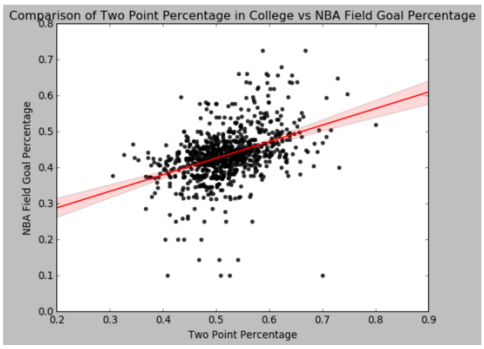

Figure 9

In each of the situations, there is a medium to strong positive relationship, meaning that as the x variable increases, so does the y variable. However, we need to keep in mind that correlation does not mean causation. Because of this, I wanted to see if I could create a linear model for predicting NBA Field Goal Percentage through college offensive rebounding and two-point percentage. Looking at Figure 10, we can see that our R-squared value is 0.27, explaining that there is a weak effect between college offensive rebounding and two-point percentage for predicting NBA Field Goal Percentage. We do need to keep in mind that to talk about this model with high confidence, we would need to check for linearity, normality, independence, and randomness.

OLS Regression Results

Dep. Variable:	NBA_FGPer	R-squared:	0.272
Model:	OLS	Adj. R-squared:	0.270
Method:	Least Squares	F-statistic:	157.9
Date:	Thu, 14 Apr 2022	Prob (F-statistic):	5.30e-59
Time:	22:20:31	Log-Likelihood:	1147.4
No. Observations:	850	AIC:	-2289.
Df Residuals:	847	BIC:	-2275.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.2550	0.018	14.125	0.000	0.220	0.290
twoP_per	0.2612	0.038	6.961	0.000	0.188	0.335
ORB_per	0.0069	0.001	10.973	0.000	0.006	0.008

Omnibus:	170.067	Durbin-Watson:	2.045
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1009.461
Skew:	-0.763	Prob(JB):	6.29e-220
Kurtosis:	8.116	Cond. No.	144.

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figure 10

V. Conclusion

In conclusion, I examined the relationship between college basketball and NBA statistics. I first examined the effect that college basketball conferences have on the NBA and discovered that through the past decade, the ACC has sent the most players to the NBA. Also, I examined the change in certain statistics from college to the NBA, such as free throw percentage and three-point percentage and found that there is evidence for shooting carrying over to the NBA. Another aspect I investigated was if it was possible to create a linear model to predict NBA

field goal percentage through college statistics. I found it was possible through using college two-point percentage and offensive rebounding percentage. However, in the future when I look more into this model I will need to check for linearity, normality, independence, and randomness. Lastly, I assessed NBA Win-Shares and found that it correlated positively with college offensive rebound percentage, defensive rebound percentage, and two-point percentage. In the future, I am looking forward to continuing to learn more about these statistics and updating my models and predictions.

VI. References

Draft-data-20-years: https://www.kaggle.com/datasets/benwieland/nba-draft-data

College Basketball Players 2009-2021 CSV:
https://www.kaggle.com/datasets/adityak2003/college-basketball-players-20092021