Predicting NBA results based on Common Basketball Statistics

Antonio Balseiro Vilela

June 2021

Introduction

The National Basketball Association is one of the major sports leagues in The United States of America, and thus, it generates large amounts of money every year. The teams that compose the NBA are constantly looking for new sources of information, specifically statistical data, that can help them evaluate and improve performance. New technologies and methodologies are constantly being developed and used by statisticians to track each aspect of the game and assess which factors are key when it comes to winning games. The purpose of this research is to identify which variables mostly affect the number of wins each team obtains throughout the NBA Regular Season. Data has been obtained from the NBA official website. The information collected represents data from the 2020-2021 season. In order to perform the analysis, I have run a Regression Model that illustrates how close the chosen variables are from depicting real life results. In addition, flaws of the model have also been considered and discussed throughout the paper.

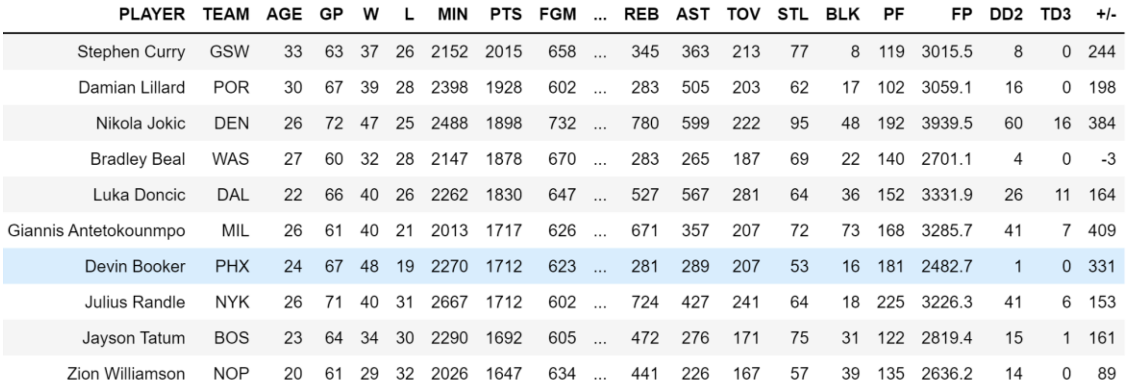
Data Section

For the project Predicting NBA results based on Common Basketball Statistics, I have extracted data from the NBA official website. Common Statistics from every single player in the League has been retrieved, such as points scored, turnovers, rebounds,etc. These statistics (indepedent variables) will be exploited in order to find out what is their relation towards the wins obtained (independent variable). An explanation of each of these statistics will be presented in the project. Throughout the paper, the original Datafreame has been modified to adapt the data to the paper requisites. Also, new information has been added to add value to the already existing used Dataframe.

Methodology and Results

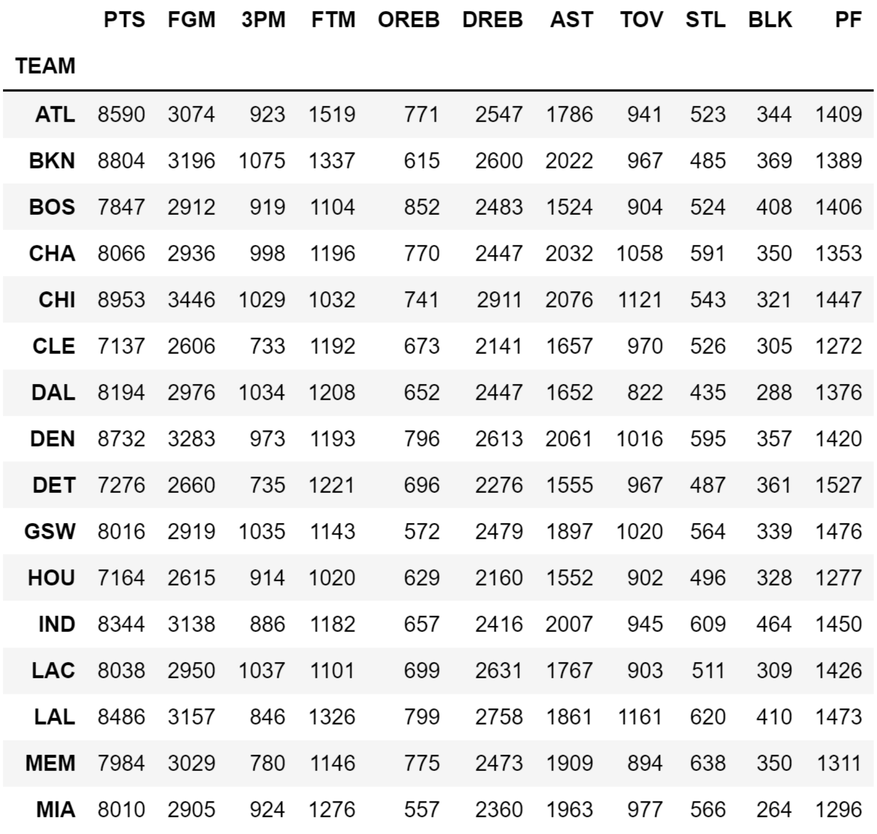
In order to understand the relationship between the independent variables and the dependent variable (wins), a regression analysis has been carried out. The original Dataframe imported into the project consisted in a raw data table containing 30 possible independent variables that can affect the outcome of the dependent variable.

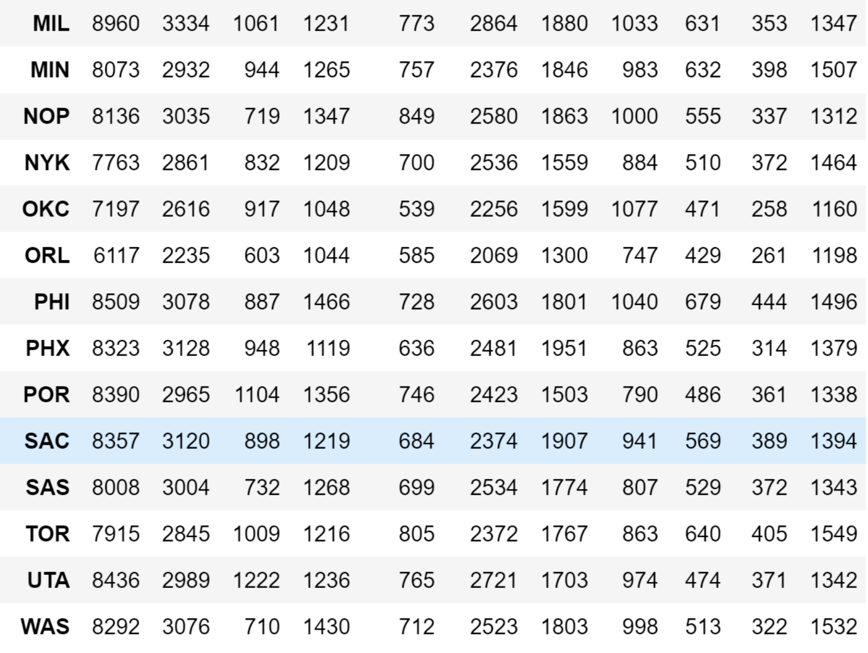
First of all, I imported all the necessary libraries to work with Python and I retrieved the dataset of the basic statistis of all the NBA players tat featured in the 2020-2021 NBA Regular Season.



PLAYER	TEAM	AGE	GP	W	L	MIN	PTS	FGM	...	REB	AST	TOV	STL	BLK	PF	FP	DD2	TD3	+/-
Stephen Curry	GSW	33	63	37	26	2152	2015	658	...	345	363	213	77	8	119	3015.5	8	0	244
Damian Lillard	POR	30	67	39	28	2398	1928	602	...	283	505	203	62	17	102	3059.1	16	0	198
Nikola Jokic	DEN	26	72	47	25	2488	1898	732	...	780	599	222	95	48	192	3939.5	60	16	384
Bradley Beal	WAS	27	60	32	28	2147	1878	670	...	283	265	187	69	22	140	2701.1	4	0	-3
Luka Doncic	DAL	22	66	40	26	2262	1830	647	...	527	567	281	64	36	152	3331.9	26	11	164
Giannis Antetokounmpo	MIL	26	61	40	21	2013	1717	626	...	671	357	207	72	73	168	3285.7	41	7	409
Devin Booker	PHX	24	67	48	19	2270	1712	623	...	281	289	207	53	16	181	2482.7	1	0	331
Julius Randle	NYK	26	71	40	31	2667	1712	602	...	724	427	241	64	18	225	3226.3	41	6	153
Jayson Tatum	BOS	23	64	34	30	2290	1692	605	...	472	276	171	75	31	122	2819.4	15	1	161
Zion Williamson	NOP	20	61	29	32	2026	1647	634	...	441	226	167	57	39	135	2636.2	14	0	89

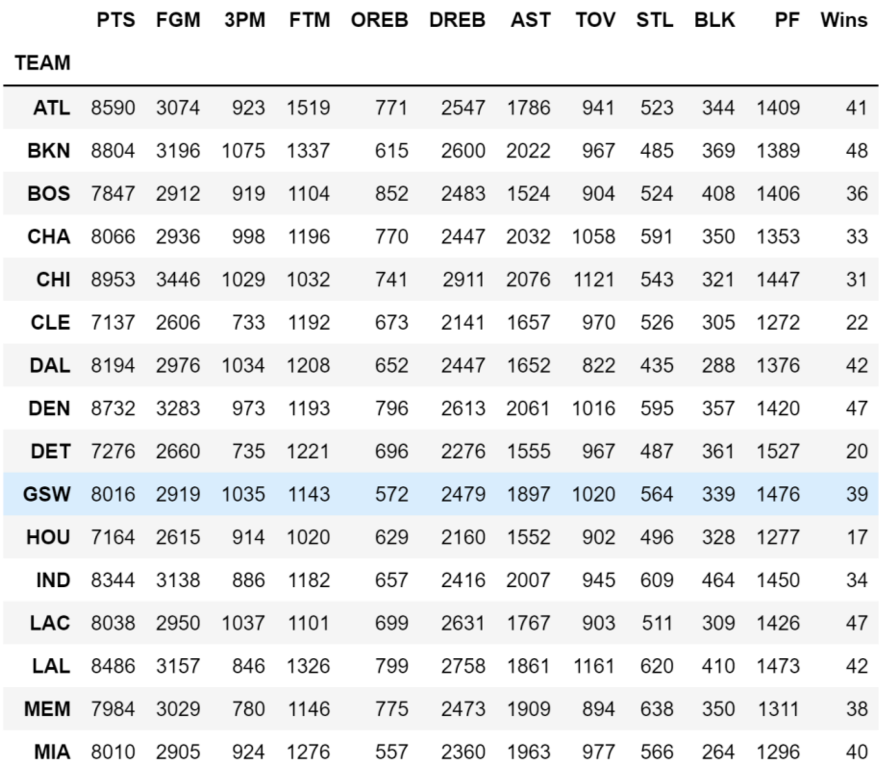
After training and testing the dataset, I selected only a number of statistics that would be useful for the analysis, such as Points scored, rebounds obtained, assists to other teammates or steals. Some of the other statitics were not chosen as they would bias the result: instead of choosing "rebounds", I decided to include "Offensive rebounds" and "Defensive rebouds", as both these stats comprise the first one. As the original dataset was comprised of players and I needed the data grouped by team, I grouped the dataframe by each team and I summed all the other variables that were chosen for the analysis.

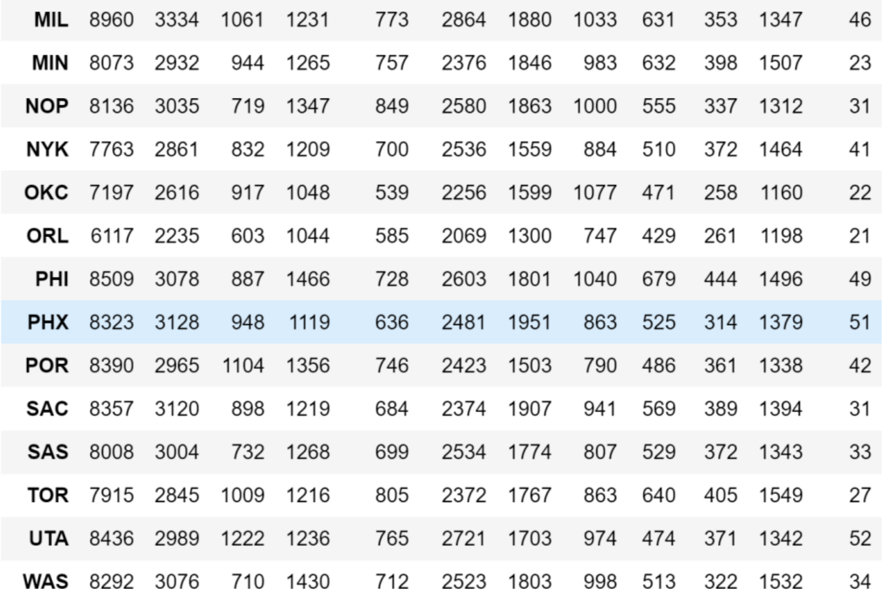




	PTS	FGM	3PM	FTM	OREB	DREB	AST	TOV	STL	BLK	PF
TEAM											
ATL	8590	3074	923	1519	771	2547	1786	941	523	344	1409
BKN	8804	3196	1075	1337	615	2600	2022	967	485	369	1389
BOS	7847	2912	919	1104	852	2483	1524	904	524	408	1406
CHA	8066	2936	998	1196	770	2447	2032	1058	591	350	1353
CHI	8953	3446	1029	1032	741	2911	2076	1121	543	321	1447
CLE	7137	2606	733	1192	673	2141	1657	970	526	305	1272
DAL	8194	2976	1034	1208	652	2447	1652	822	435	288	1376
DEN	8732	3283	973	1193	796	2613	2061	1016	595	357	1420
DET	7276	2660	735	1221	696	2276	1555	967	487	361	1527
GSW	8016	2919	1035	1143	572	2479	1897	1020	564	339	1476
HOU	7164	2615	914	1020	629	2160	1552	902	496	328	1277
IND	8344	3138	886	1182	657	2416	2007	945	609	464	1450
LAC	8038	2950	1037	1101	699	2631	1767	903	511	309	1426
LAL	8486	3157	846	1326	799	2758	1861	1161	620	410	1473
MEM	7984	3029	780	1146	775	2473	1909	894	638	350	1311
MIA	8010	2905	924	1276	557	2360	1963	977	566	264	1296
MIL	8960	3334	1061	1231	773	2864	1880	1033	631	353	1347
MIN	8073	2932	944	1265	757	2376	1846	983	632	398	1507
NOP	8136	3035	719	1347	849	2580	1863	1000	555	337	1312
NYK	7763	2861	832	1209	700	2536	1559	884	510	372	1464
OKC	7197	2616	917	1048	539	2256	1599	1077	471	258	1160
ORL	6117	2235	603	1044	585	2069	1300	747	429	261	1198
PHI	8509	3078	887	1466	728	2603	1801	1040	679	444	1496
PHX	8323	3128	948	1119	636	2481	1951	863	525	314	1379
POR	8390	2965	1104	1356	746	2423	1503	790	486	361	1338
SAC	8357	3120	898	1219	684	2374	1907	941	569	389	1394
SAS	8008	3004	732	1268	699	2534	1774	807	529	372	1343
TOR	7915	2845	1009	1216	805	2372	1767	863	640	405	1549
UTA	8436	2989	1222	1236	765	2721	1703	974	474	371	1342
WAS	8292	3076	710	1430	712	2523	1803	998	513	322	1532

Once the dataframe was updated, I needed to retrieve the number of wins that each team obtained at the end of the Regular Season. As the original database was comprised of player data and not team data, this information had to be obtained from other source in the NBA official website. A new dataframe was created by inserting a new column with the number of wins obtained by each team.

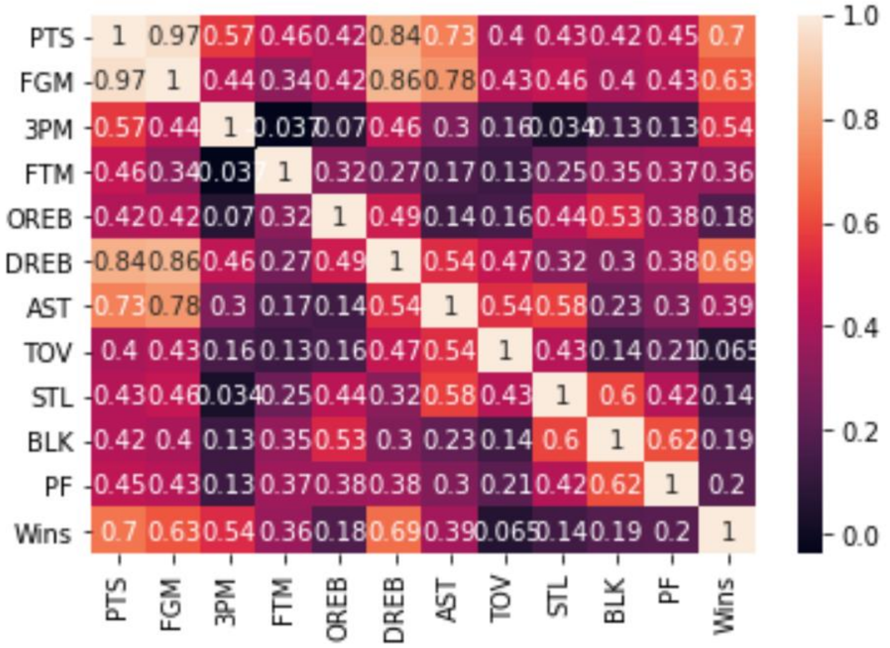




TEAM	PTS	FGM	3PM	FTM	OREB	DREB	AST	TOV	STL	BLK	PF	Wins
ATL	8590	3074	923	1519	771	2547	1786	941	523	344	1409	41
BKN	8804	3196	1075	1337	615	2600	2022	967	485	369	1389	48
BOS	7847	2912	919	1104	852	2483	1524	904	524	408	1406	36
CHA	8066	2936	998	1196	770	2447	2032	1058	591	350	1353	33
CHI	8953	3446	1029	1032	741	2911	2076	1121	543	321	1447	31
CLE	7137	2606	733	1192	673	2141	1657	970	526	305	1272	22
DAL	8194	2976	1034	1208	652	2447	1652	822	435	288	1376	42
DEN	8732	3283	973	1193	796	2613	2061	1016	595	357	1420	47
DET	7276	2660	735	1221	696	2276	1555	967	487	361	1527	20
GSW	8016	2919	1035	1143	572	2479	1897	1020	564	339	1476	39
HOU	7164	2615	914	1020	629	2160	1552	902	496	328	1277	17
IND	8344	3138	886	1182	657	2416	2007	945	609	464	1450	34
LAC	8038	2950	1037	1101	699	2631	1767	903	511	309	1426	47
LAL	8486	3157	846	1326	799	2758	1861	1161	620	410	1473	42
MEM	7984	3029	780	1146	775	2473	1909	894	638	350	1311	38
MIA	8010	2905	924	1276	557	2360	1963	977	566	264	1296	40
MIL	8960	3334	1061	1231	773	2864	1880	1033	631	353	1347	46
MIN	8073	2932	944	1265	757	2376	1846	983	632	398	1507	23
NOP	8136	3035	719	1347	849	2580	1863	1000	555	337	1312	31
NYK	7763	2861	832	1209	700	2536	1559	884	510	372	1464	41
OKC	7197	2616	917	1048	539	2256	1599	1077	471	258	1160	22
ORL	6117	2235	603	1044	585	2069	1300	747	429	261	1198	21
PHI	8509	3078	887	1466	728	2603	1801	1040	679	444	1496	49
PHX	8323	3128	948	1119	636	2481	1951	863	525	314	1379	51
POR	8390	2965	1104	1356	746	2423	1503	790	486	361	1338	42
SAC	8357	3120	898	1219	684	2374	1907	941	569	389	1394	31
SAS	8008	3004	732	1268	699	2534	1774	807	529	372	1343	33
TOR	7915	2845	1009	1216	805	2372	1767	863	640	405	1549	27
UTA	8436	2989	1222	1236	765	2721	1703	974	474	371	1342	52
WAS	8292	3076	710	1430	712	2523	1803	998	513	322	1532	34

3.1 Correlation analysis

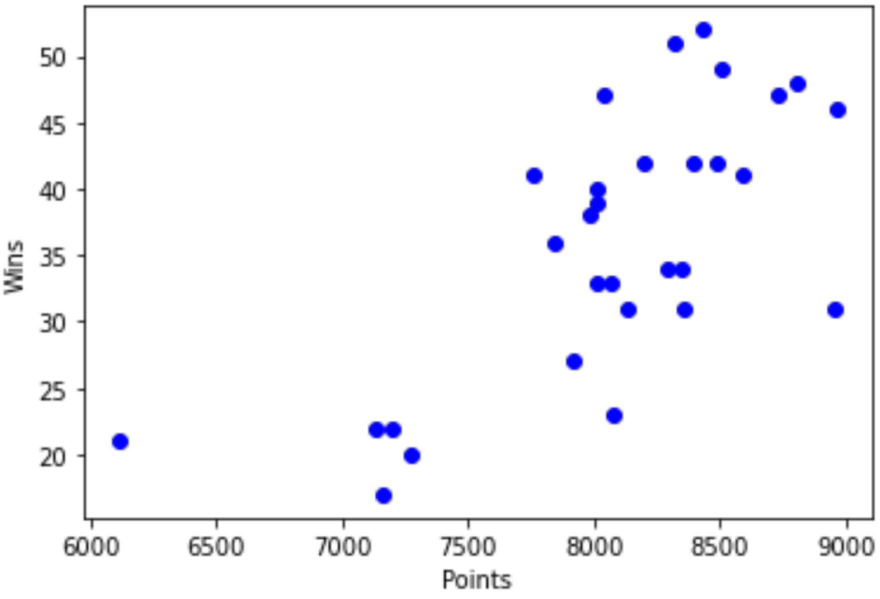
Next, I ran a correlation matrix to quantify the degree to which two variables are related. Through the correlation analysis, the goal was to evaluate the correlation coefficient that tells us how much one variable changes when the other one does. Correlation analysis provides with a linear relationship between two variables. The results showed that some variables had a higher degree of correlation to wins than others.



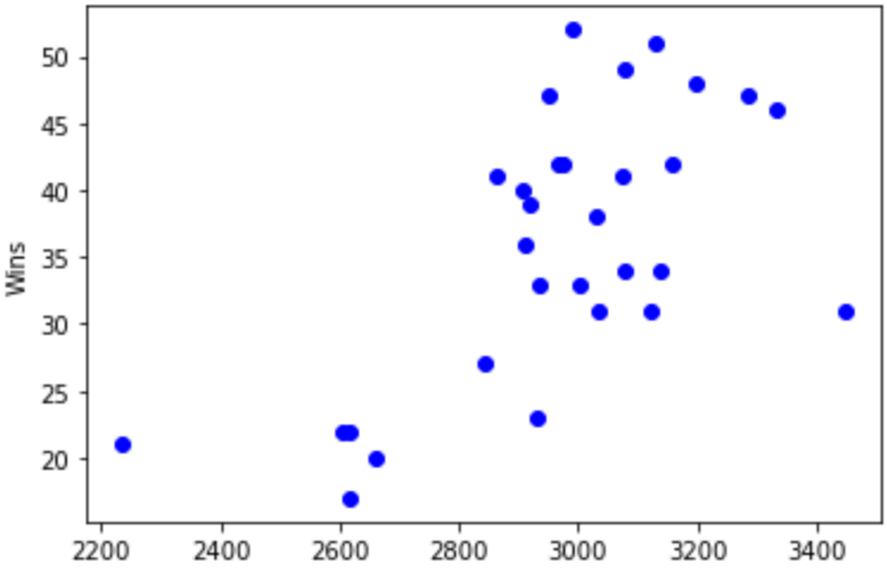
	PTS	FGM	3PM	FTM	OREB	DREB	AST	TOV	STL	BLK	PF	Wins
PTS	1	0.97	0.57	0.46	0.42	0.84	0.73	0.4	0.43	0.42	0.45	0.7
FGM	0.97	1	0.44	0.34	0.42	0.86	0.78	0.43	0.46	0.4	0.43	0.63
3PM	0.57	0.44	1	0.037	0.07	0.46	0.3	0.16	0.034	0.13	0.13	0.54
FTM	0.46	0.34	0.037	1	0.32	0.27	0.17	0.13	0.25	0.35	0.37	0.36
OREB	0.42	0.42	0.07	0.32	1	0.49	0.14	0.16	0.44	0.53	0.38	0.18
DREB	0.84	0.86	0.46	0.27	0.49	1	0.54	0.47	0.32	0.3	0.38	0.69
AST	0.73	0.78	0.3	0.17	0.14	0.54	1	0.54	0.58	0.23	0.3	0.39
TOV	0.4	0.43	0.16	0.13	0.16	0.47	0.54	1	0.43	0.14	0.21	0.065
STL	0.43	0.46	0.034	0.25	0.44	0.32	0.58	0.43	1	0.6	0.42	0.14
BLK	0.42	0.4	0.13	0.35	0.53	0.3	0.23	0.14	0.6	1	0.62	0.19
PF	0.45	0.43	0.13	0.37	0.38	0.38	0.3	0.21	0.42	0.62	1	0.2
Wins	0.7	0.63	0.54	0.36	0.18	0.69	0.39	0.065	0.14	0.19	0.2	1

3.2 Scatterplot Testing

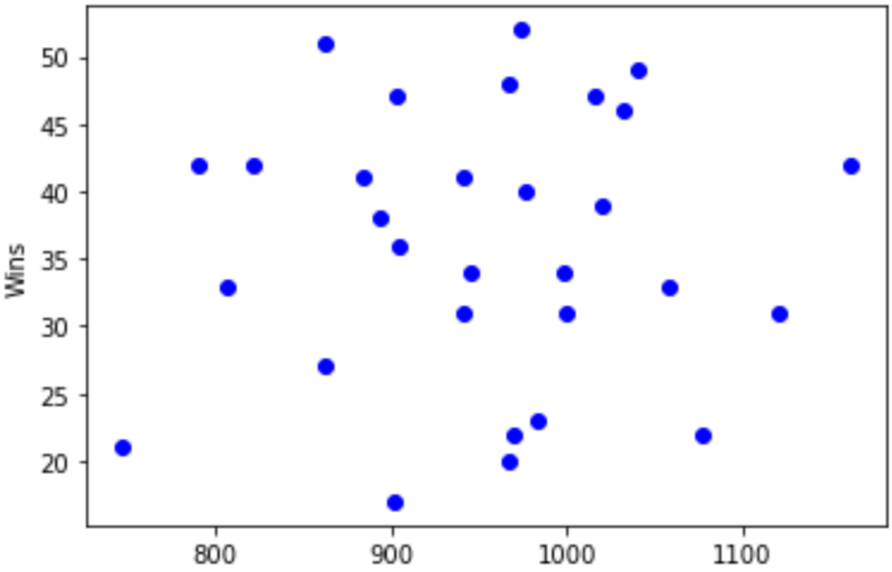
I again ran a test and train analysis on the new dataframe. Based on these results, a scatterplot was performed with the variables that obtained a higher correlation to wins. In the first case, I compared wins to points obtained. Although the sample size is relatively small, we can observe a linear realation between both variables.



The same results showed by performing a correlation analysis between both Field Goals Made / Defensive Rebounds and wins.

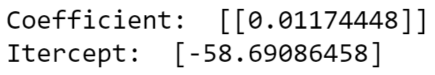


Another scatterplot was also performed to show the results of a variable that had barely no correlation to wins. As we can see, Turnovers, which happen every time the team in possesion of the ball loses it, do not really relate to the number of wins obtained, as there is a lot of dispersion and the plotting does not seem to fit a linear line.

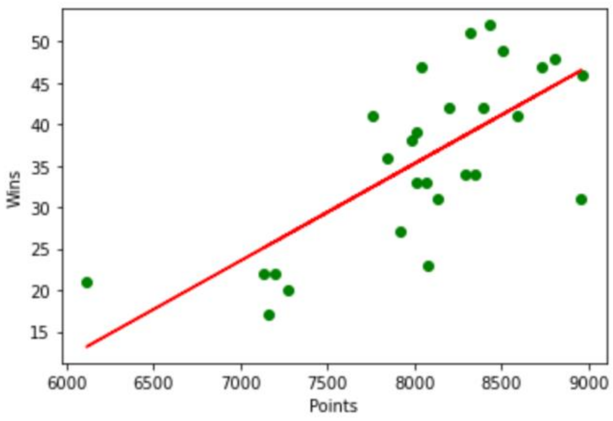


3.3. Simple and Multiple Regression Models

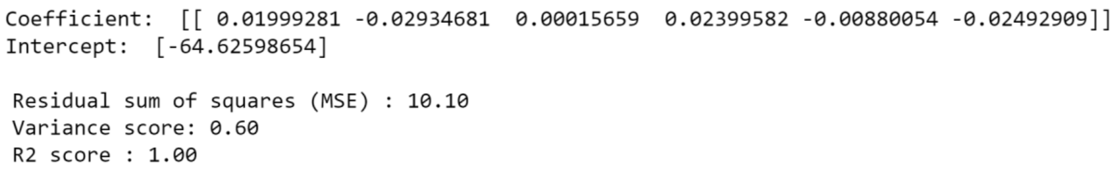
After performing this analysis, I ran a simple linear regression model to see how one single variable can explain the behaviour of the dependent variable. But as the sample size is quite small, as there are only 30 teams in the league, the results were not very promising.



```
Coefficient:  [[0.01174448]]
Itercept:  [-58.69086458]
```



Then, I ran a multiple linear regression analysis by choosing the variables that had a higher correlation to wins. This model fitted much nicely, as all these different variables do explain why teams obtain more wins than others thanks to the results retrieved by performing a Variance and an R2 analysis. The latter is a goodness-of-fit variation that indicates the percentage of the variance in the dependent variable that the independent variables explain collectively, while the latter explains the variability between the different variables.



```
Coefficient:  [[ 0.01999281 -0.02934681  0.00015659  0.02399582 -0.00880054 -0.02492909]]
Intercept:  [-64.62598654]

 Residual sum of squares (MSE) : 10.10
 Variance score: 0.60
 R2 score : 1.00
```

3.4. Results

After performing all the mentioned analysis following the regression model methodology, the results showed that the variables chosen do, in fact, explain the number of wins that each team obtained by the end of the Regular Season. The variables chosen to perform this multiple linear regression analysis were:

Points scored, Field Goals Made, 3 Point Field Goals Made, Defensive Rebounds ,Assists, Offensive Rebounds

By performing the multiple linear regression analysis, I came to the conclusion that these variables do explain the behaviour of the dependent variable, that is, the wins obtained by each team, by performing a Variance and an R2 analysis. Even though the sampe size is relatively small, we obtained a Variance of 0.6 and an R2 of 1.00. This R2 result can bias the outcome of the analysis due to the small sample size used to perform the analysis. However, we can see that this model, with the flaws that it may have, does in fact fit the hypothesis that these variables explain why some teams obtain more wins than others.

Discussion and Conclusion

During this report I analyzed how some statistics obtained by NBA teams throught the 2020-2021 regular season can explain the number of wins they obtained. We identified as key independent variables the Points scored, Field Goals Made, 3 Point Field Goals Made, Defensive Rebounds, Assists & Offensive Rebounds. The results showed that these statistics really do explain a teams´ performance. Still, the model has some flaws that should be addressed in future reports. Firstly, the sameple size is too small. On this ocassion, I chose to ran the analysis on team performance, but we could obtain better results if the dataset was comprised of players and not teams. Also, there are advanced statistics that were not chosen in the analysis, and some others that are intangible and cannot (or are very hard) to obtain, such as players mentalityt, sickness, personal situation or, one of the most important ones, injuries. All these factors also contribute to the number of wins ay team obtains, and should also be taken into account when performing an anlysis such as this one.