IB Extended Essay:

Analyzing the impact of individual and team statistics on winning in the NBA

Research Question: Can NBA conference finalists be best determined by individual or team success?

Subject: Mathematics

Word Count: 3714

Table of Contents

1. Introduction

Basketball is a sport where the epitome of collaboration and athleticism shine through, offering moments that showcase the sheer creativity of humans through activity. There are many aspects to this complicated and intricate game, with the offensive and defensive parts of it spearheading it all. Although at the end of the day, the ultimate goal is to see who is able to put the ball through the basket more, statistics also show a large part of the game where the eye-test doesn't provide.

In the world's premier league of basketball – the NBA (National Basketball Association), each team looks to build rosters that sustain success and emit talent as they compete for a championship. The ultimate goal in the NBA is to raise the Larry O'Brien Trophy – to win the NBA playoffs, beating 4 teams in 4 rounds of 7-game series.

To build a roster, teams will have to find talented players to lead their team to success. A really capable player is often needed to be able to be a leader in games, showing up in clutch moments in a game. However, on the other side of the spectrum, team building should also have in mind the chemistry of the team itself, as it is a team after all, requiring multiple different people to be successful. In teambuilding, it is hard to balance the two, a good player is rare to stumble upon, and to keep him will need a hefty price often, which leaves less to be given out to other players to fill the team, sacrificing some of the quality of the teammates and depth. Another approach could be to establish a well-rounded team, which is to have many elite level players around the same caliber to create a homogenous squad that plays well together. This sacrifices the ability to obtain the top of the top talent, which might help you to make the toughest shots or stops in grueling moments.

Then it begs the question, what is more valuable, team stats that show how proficient a team is as a whole, or the stats of the top few individuals on a team who are key contributors, to the success of an NBA team for them to make the finals?

This extended essay will tackle the question: Can NBA conference finalists be best determined by individual or team success?

This study will be in analyzing the past 24 years of basketball, starting from the 2000-2001 season to the 2023-2024 season. obtaining a score that shows the difference between individual stats, and team stats with their impact on winning.

1.1 Methodology

I exported the data of all 716 teams over the last 25 seasons onto an excel document, then proceeded to convert the file's format to CSV (comma separated value), which is easier for my python application to recognize.

I will use python as a medium to create machine learning models that can process and analyze the data effectively. Correlation coefficient, decision trees, k-mean clusters, logistic regression, and normalization will be used to determine the value in which a stat relates to a team making the finals. In addition, I will be using the model to fabricate a score for teams and individuals that signifies how valuable they are in predicting and influencing conference finalists. The DRtg (defensive rating), ORtg (offensive rating), pace, and SRS (simple ratings system), are stats that would be utilized upon my research. On the other hand, VORP (value over replacement), BPM (box plus minus), PER (player efficiency rating), and WS (win share), will be used for the individual side of the research (Glossary, n.d.). Reference the appendix for the definition of the stats. More analyzation will occur by seeing how proficient a player is and the team's defense, moreover, considering their team performance and opponents. All the statistics will be obtained from Stathead.com, the data provided by Sportradar, the official data provider of the NBA. Research papers will also be referenced and examined upon.

Correlation Coefficient is a concept utilized frequently while tackling the research question (Glen, 2024).

Another mathematical concept that I used was the value of different stats. For example, the three-point shot is more valuable than the normal two-point shot, therefore when you want to calculate the efficiency of a player, it is more accurate to portray it in eFG%, where the value of the three-point shot is taken into account when calculating the player's overall efficiency, therefore their shot diet contributes to the final eFG% (Glossary, n.d.). In this case different weights will be used to act as a balance for score calculation and aggregation will be used. A comprehensive score would be given to the teams in leu of their stats. Then the process would be reversed to try to verify if the top scorers’ teams would be the same ones that made the finals.

2 Mathematical Exploration

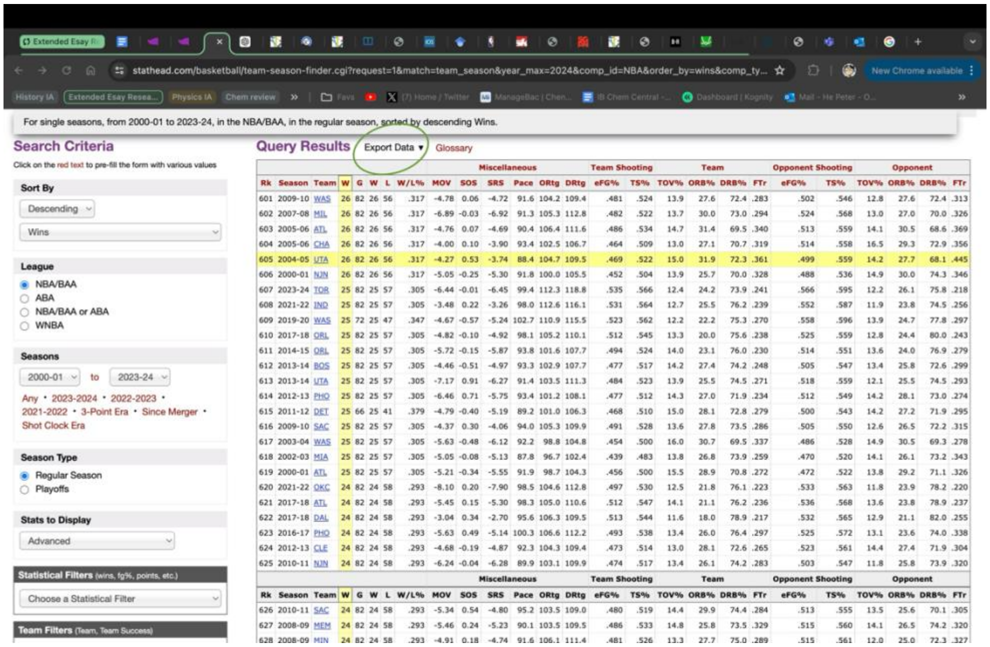
2.1 Literature Review

One research article - “Game statistics that discriminate winning and losing at the NBA level of basketball competition”, showed that the flow of the game decreases in the playoffs significantly (Cabarkapa et al., 2022). One conclusion that they came to was that the players play more tenacious defense, as they will not allow for an open shot from the opponent that could alter the trajectory of their season. The stakes of playoff games are significantly higher than the 82 games each team play in the regular season, therefore their statistics may take a rise or drop in different categories. From this one assumption can be made is that since every team in the playoffs would play defense at an elevated level, regardless of personnel, offensive tough shot making is what is extremely valued in the playoffs. The research also found that there are less FGA (field goal attempts) and FGM (field goal made). One could assume that although team stats are important, offensive factors such as a team effort in creating an open shot, and defensive factors like team defensive rotations will be less valuable in front of tough shot making, which is an individual aspect of the game.

Another research article – “Exploring Game Performance in the National Basketball Association Using Player Tracking Data”, showed that All-star players were more valuable in terms of producing a 12-foot shot make (Sampaio et al., 2015). Through other means of data analysis, such as k-cluster means using a statistical model, showed different groups of players that contribute differently to different aspects. This allows teams to be able to gameplan better against the opponents.

2.2 Data Collection

Here is a page on Stathead, the website where all the data is obtained from. Here a query search had been entered and I narrow down the criteria of the results on the left. Then when I have obtained the right data, I export the data into an excel sheet.



stathead.com/basketball/team-season-finder.cgi?request=1&match=team_season&year_max=2024&comp_id=NBA&order_by=wins&comp_ty...

For single seasons, from 2000-01 to 2023-24, in the NBA/BAA, in the regular season, sorted by descending Wins.

Search Criteria

Click on the red text to pre-fill the form with various values

Sort By

Descending

Wins

League

● NBA/BAA
○ ABA
○ NBA/BAA or ABA
○ WNBA

Seasons

2000-01 to 2023-24

Any • 2023-2024 • 2022-2023 • 2021-2022 • 3-Point Era • Since Merger • Shot Clock Era

Season Type

● Regular Season
○ Playoffs

Stats to Display

Advanced

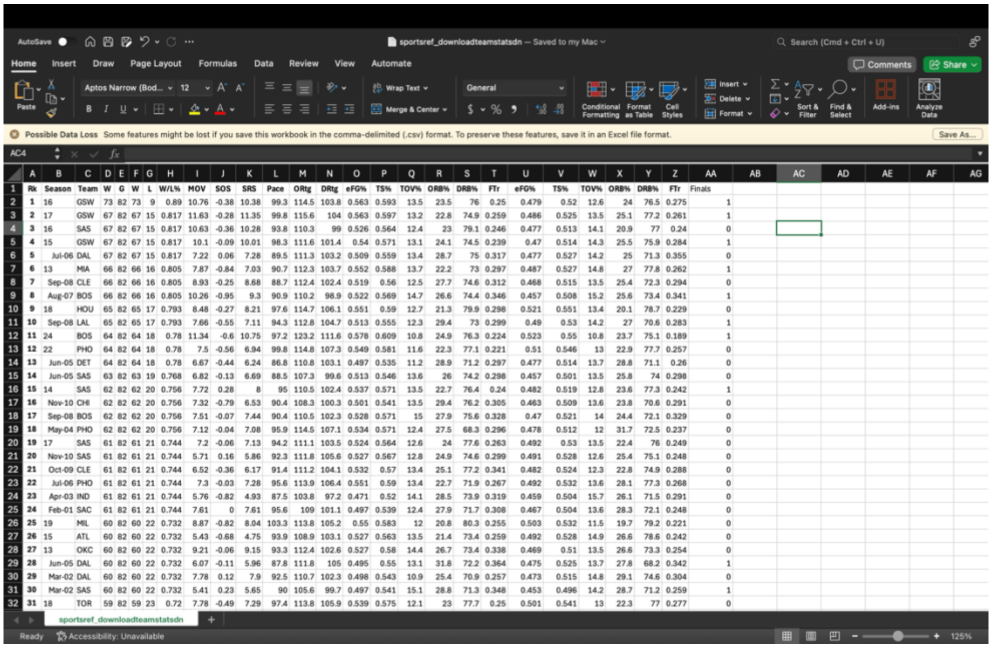
Statistical Filters (wins, fg%, points, etc.)

Choose a Statistical Filter

Team Filters (Team, Team Success)

Query Results

								Miscellaneous						Team Shooting		Team				Opponent Shooting		Opponent			
Rk	Season	Team	W	G	W	L	W/L%	MOV	SOS	SRS	Pace	ORtg	DRtg	eFG%	TS%	TOV%	ORB%	DRB%	FTr	eFG%	TS%	TOV%	ORB%	DRB%	FTr
601	2009-10	WAS	26	82	26	56	.317	-4.78	0.06	-4.72	91.6	104.2	109.4	.481	.524	13.9	27.6	72.4	.283	.502	.546	12.8	27.6	72.4	.313
602	2007-08	MIL	26	82	26	56	.317	-6.89	-0.03	-6.92	91.3	105.3	112.8	.482	.522	13.7	30.0	73.0	.294	.524	.568	13.0	27.0	70.0	.326
603	2005-06	ATL	26	82	26	56	.317	-4.76	0.07	-4.69	90.4	106.4	111.6	.486	.534	14.7	31.4	69.5	.340	.513	.559	14.1	30.5	68.6	.369
604	2005-06	CHA	26	82	26	56	.317	-4.00	0.10	-3.90	93.4	102.5	106.7	.464	.509	13.0	27.1	70.7	.319	.514	.558	16.5	29.3	72.9	.356
605	2004-05	UTA	26	82	26	56	.317	-4.27	0.53	-3.74	88.4	104.7	109.5	.469	.522	15.0	31.9	72.3	.361	.499	.559	14.2	27.7	68.1	.445
606	2000-01	NJN	26	82	26	56	.317	-5.05	-0.25	-5.30	91.8	100.0	105.5	.452	.504	13.9	25.7	70.0	.328	.488	.536	14.9	30.0	74.3	.346
607	2023-24	TOR	25	82	25	57	.305	-6.44	-0.01	-6.45	99.4	112.3	118.8	.535	.566	12.4	24.2	73.9	.241	.566	.595	12.2	26.1	75.8	.218
608	2021-22	IND	25	82	25	57	.305	-3.48	0.22	-3.26	98.0	112.6	116.1	.531	.564	12.7	25.5	76.2	.239	.552	.587	11.9	23.8	74.5	.256
609	2019-20	WAS	25	72	25	47	.347	-4.67	-0.57	-5.24	102.7	110.9	115.5	.523	.562	12.2	22.2	75.3	.270	.558	.596	13.9	24.7	77.8	.297
610	2017-18	ORL	25	82	25	57	.305	-4.82	-0.10	-4.92	98.1	105.2	110.1	.512	.545	13.3	20.0	75.6	.238	.525	.559	12.8	24.4	80.0	.243
611	2014-15	ORL	25	82	25	57	.305	-5.72	-0.15	-5.87	93.8	101.6	107.7	.494	.524	14.0	23.1	76.0	.230	.514	.551	13.6	24.0	76.9	.279
612	2013-14	BOS	25	82	25	57	.305	-4.46	-0.51	-4.97	93.3	102.9	107.7	.477	.517	14.2	27.4	74.2	.248	.505	.547	13.4	25.8	72.6	.299
613	2013-14	UTA	25	82	25	57	.305	-7.17	0.91	-6.27	91.4	103.5	111.3	.484	.523	13.9	25.5	74.5	.271	.518	.559	12.1	25.5	74.5	.293
614	2012-13	PHO	25	82	25	57	.305	-6.46	0.71	-5.75	93.4	101.2	108.1	.477	.512	14.3	27.0	71.9	.234	.512	.549	14.2	28.1	73.0	.274
615	2011-12	DET	25	66	25	41	.379	-4.79	-0.40	-5.19	89.2	101.0	106.3	.468	.510	15.0	28.1	72.8	.279	.500	.543	14.2	27.2	71.9	.295
616	2009-10	SAC	25	82	25	57	.305	-4.37	0.30	-4.06	94.0	105.3	109.9	.491	.528	13.6	27.8	73.5	.286	.505	.550	12.6	26.5	72.2	.315
617	2003-04	WAS	25	82	25	57	.305	-5.63	-0.48	-6.12	92.2	98.8	104.8	.454	.500	16.0	30.7	69.5	.337	.486	.528	14.9	30.5	69.3	.278
618	2002-03	MIA	25	82	25	57	.305	-5.05	-0.08	-5.13	87.8	96.7	102.4	.439	.483	13.8	26.8	73.9	.259	.470	.520	14.1	26.1	73.2	.343
619	2000-01	ATL	25	82	25	57	.305	-5.21	-0.34	-5.55	91.9	98.7	104.3	.456	.500	15.5	28.9	70.8	.272	.472	.522	13.8	29.2	71.1	.326
620	2021-22	OKC	24	82	24	58	.293	-8.10	0.20	-7.90	98.5	104.6	112.8	.497	.530	12.5	21.8	76.1	.223	.533	.563	11.8	23.9	78.2	.220
621	2017-18	ATL	24	82	24	58	.293	-5.45	0.15	-5.30	98.3	105.0	110.6	.512	.547	14.1	21.1	76.2	.236	.536	.568	13.6	23.8	78.9	.237
622	2017-18	DAL	24	82	24	58	.293	-3.04	0.34	-2.70	95.6	106.3	109.5	.513	.544	11.6	18.0	78.9	.217	.532	.565	12.9	21.1	82.0	.255
623	2016-17	PHO	24	82	24	58	.293	-5.63	0.49	-5.14	100.3	106.6	112.2	.493	.538	13.4	26.0	76.4	.297	.525	.572	13.1	23.6	74.0	.338
624	2012-13	CLE	24	82	24	58	.293	-4.68	-0.19	-4.87	92.3	104.3	109.4	.473	.514	13.0	28.1	72.6	.265	.523	.561	14.4	27.4	71.9	.304
625	2010-11	NJN	24	82	24	58	.293	-6.24	-0.04	-6.28	89.9	103.1	109.9	.474	.517	13.4	26.1	74.2	.283	.503	.547	11.8	25.8	73.9	.320
626	2010-11	SAC	24	82	24	58	.293	-5.34	0.54	-4.80	95.2	103.5	109.0	.480	.519	14.4	29.9	74.4	.284	.513	.555	13.5	25.6	70.1	.305
627	2008-09	MEM	24	82	24	58	.293	-5.46	0.24	-5.23	90.1	103.5	109.5	.486	.533	14.8	25.8	73.5	.329	.515	.560	14.1	26.5	74.2	.320
628	2008-09	MIN	24	82	24	58	.293	-4.91	0.18	-4.74	91.6	106.1	111.4	.481	.526	13.3	27.7	75.0	.289	.515	.561	12.0	25.0	72.3	.327



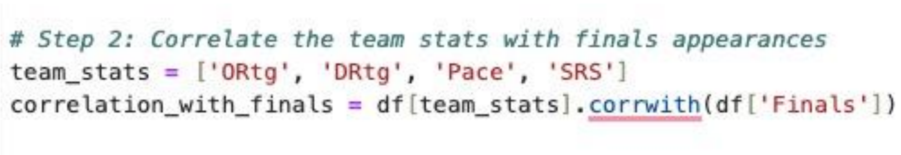
Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

Rk	Season	Team	W	G	W	L	W/L%	MOV	SOS	SRS	Pace	ORtg	DRtg	eFG%	TS%	TOV%	ORB%	DRB%	FTr	eFG%	TS%	TOV%	ORB%	DRB%	FTr	Finals
1	16	GSW	73	82	73	9	0.89	10.76	-0.38	10.38	99.3	114.5	103.8	0.563	0.593	13.5	23.5	76	0.25	0.479	0.52	12.6	24	76.5	0.275	1
2	17	GSW	67	82	67	15	0.817	11.63	-0.28	11.35	99.8	115.6	104	0.563	0.597	13.2	22.8	74.9	0.259	0.486	0.525	13.5	25.1	77.2	0.261	1
3	16	SAS	67	82	67	15	0.817	10.63	-0.36	10.28	93.8	110.3	99	0.526	0.564	12.4	23	79.1	0.246	0.477	0.513	14.1	20.9	77	0.24	0
4	15	GSW	67	82	67	15	0.817	10.1	-0.09	10.01	98.3	111.6	101.4	0.54	0.571	13.1	24.1	74.5	0.239	0.47	0.514	14.3	25.5	75.9	0.284	1
5	Jul-06	DAL	67	82	67	15	0.817	7.22	0.06	7.28	89.5	111.3	103.2	0.509	0.559	13.4	28.7	75	0.317	0.477	0.527	14.2	25	71.3	0.355	0
6	13	MIA	66	82	66	16	0.805	7.87	-0.84	7.03	90.7	112.3	103.7	0.552	0.588	13.7	22.2	73	0.297	0.487	0.527	14.8	27	77.8	0.262	1
7	Sep-08	CLE	66	82	66	16	0.805	8.93	-0.25	8.68	88.7	112.4	102.4	0.519	0.56	12.5	27.7	74.6	0.312	0.468	0.515	13.5	25.4	72.3	0.294	0
8	Aug-07	BOS	66	82	66	16	0.805	10.26	-0.95	9.3	90.9	110.2	98.9	0.522	0.569	14.7	26.6	74.4	0.346	0.457	0.508	15.2	25.6	73.4	0.341	1
9	18	HOU	65	82	65	17	0.793	8.48	-0.27	8.21	97.6	114.7	106.1	0.551	0.59	12.7	21.3	79.9	0.298	0.521	0.551	13.4	20.1	78.7	0.229	0
10	Sep-08	LAL	65	82	65	17	0.793	7.66	-0.55	7.11	94.3	112.8	104.7	0.513	0.555	12.3	29.4	73	0.299	0.49	0.53	14.2	27	70.6	0.283	1
11	24	BOS	64	82	64	18	0.78	11.34	-0.6	10.75	97.2	123.2	111.6	0.578	0.609	10.8	24.9	76.3	0.224	0.523	0.55	10.8	23.7	75.1	0.189	1
12	22	PHO	64	82	64	18	0.78	7.5	-0.56	6.94	99.8	114.8	107.3	0.549	0.581	11.6	22.3	77.1	0.221	0.51	0.546	13	22.9	77.7	0.257	0
13	Jun-05	DET	64	82	64	18	0.78	6.67	-0.44	6.24	86.8	110.8	103.1	0.497	0.535	11.2	28.9	71.2	0.297	0.477	0.514	13.7	28.8	71.1	0.26	0
14	Jun-05	SAS	63	82	63	19	0.768	6.82	-0.13	6.69	88.5	107.3	99.6	0.513	0.546	13.6	26	74.2	0.298	0.457	0.501	13.5	25.8	74	0.298	0
15	14	SAS	62	82	62	20	0.756	7.72	0.28	8	95	110.5	102.4	0.537	0.571	13.5	22.7	76.4	0.24	0.482	0.519	12.8	23.6	77.3	0.242	1
16	Nov-10	CHI	62	82	62	20	0.756	7.32	-0.79	6.53	90.4	108.3	100.3	0.501	0.541	13.5	29.4	76.2	0.305	0.463	0.509	13.6	23.8	70.6	0.291	0
17	Sep-08	BOS	62	82	62	20	0.756	7.51	-0.07	7.44	90.4	110.5	102.3	0.528	0.571	15	27.9	75.6	0.328	0.47	0.521	14	24.4	72.1	0.329	0
18	May-04	PHO	62	82	62	20	0.756	7.12	-0.04	7.08	95.9	114.5	107.1	0.534	0.571	12.4	27.5	68.3	0.296	0.478	0.512	12	31.7	72.5	0.237	0
19	17	SAS	61	82	61	21	0.744	7.2	-0.06	7.13	94.2	111.1	103.5	0.524	0.564	12.6	24	77.6	0.263	0.492	0.53	13.5	22.4	76	0.249	0
20	Nov-10	SAS	61	82	61	21	0.744	5.71	0.16	5.86	92.3	111.8	105.6	0.527	0.567	12.8	24.9	74.6	0.299	0.491	0.528	12.6	25.4	75.1	0.248	0
21	Oct-09	CLE	61	82	61	21	0.744	6.52	-0.36	6.17	91.4	111.2	104.1	0.532	0.57	13.4	25.1	77.2	0.341	0.482	0.524	12.3	22.8	74.9	0.288	0
22	Jul-06	PHO	61	82	61	21	0.744	7.3	-0.03	7.28	95.6	113.9	106.4	0.551	0.59	13.4	22.7	71.9	0.267	0.492	0.532	13.6	28.1	77.3	0.268	0
23	Apr-03	IND	61	82	61	21	0.744	5.76	-0.82	4.93	87.5	103.8	97.2	0.471	0.52	14.1	28.5	73.9	0.319	0.459	0.504	15.7	26.1	71.5	0.291	0
24	Feb-01	SAC	61	82	61	21	0.744	7.61	0	7.61	95.6	109	101.1	0.497	0.539	12.4	27.9	71.7	0.308	0.467	0.504	13.6	28.3	72.1	0.248	0
25	19	MIL	60	82	60	22	0.732	8.87	-0.82	8.04	103.3	113.8	105.2	0.55	0.583	12	20.8	80.3	0.255	0.503	0.532	11.5	19.7	79.2	0.221	0
26	15	ATL	60	82	60	22	0.732	5.43	-0.68	4.75	93.9	108.9	103.1	0.527	0.563	13.5	21.4	73.4	0.259	0.492	0.528	14.9	26.6	78.6	0.242	0
27	13	OKC	60	82	60	22	0.732	9.21	-0.06	9.15	93.3	112.4	102.6	0.527	0.58	14.4	26.7	73.4	0.338	0.469	0.51	13.5	26.6	73.3	0.254	0
28	Jun-05	DAL	60	82	60	22	0.732	6.07	-0.11	5.96	87.8	111.8	105	0.495	0.55	13.1	31.8	72.2	0.364	0.475	0.525	13.7	27.8	68.2	0.342	1
29	Mar-02	DAL	60	82	60	22	0.732	7.78	0.12	7.9	92.5	110.7	102.3	0.498	0.543	10.9	25.4	70.9	0.257	0.473	0.515	14.8	29.1	74.6	0.304	0
30	Mar-02	SAS	60	82	60	22	0.732	5.41	0.23	5.65	90	105.6	99.7	0.497	0.541	15.1	28.8	71.3	0.348	0.453	0.496	14.2	28.7	71.2	0.259	1
31	18	TOR	59	82	59	23	0.72	7.78	-0.49	7.29	97.4	113.8	105.9	0.539	0.575	12.1	23	77.7	0.25	0.501	0.541	13	22.3	77	0.277	0

Here is the data I have input into the Excel sheet, in the screenshot above, it's all the teams from the past 24 seasons sorted by regular season wins, with the Finals column entered manually. After all of the data from the 2000-01 season to the 2023-24 season has been compiled into two separate documents, with 5749 player seasons for the individual data and 716 team seasons for the team data.

2.3 Score Creation

2.31 Correlation



```
# Step 2: Correlate the team stats with finals appearances
team_stats = ['ORtg', 'DRtg', 'Pace', 'SRS']
correlation_with_finals = df[team_stats].corrwith(df['Finals'])
```

Firstly, the chosen 4 stats for team and individual numbers will be correlated to the Finals column which indicates whether or not the team or player made the finals that season, 1 indicating they did, 0 meaning otherwise.

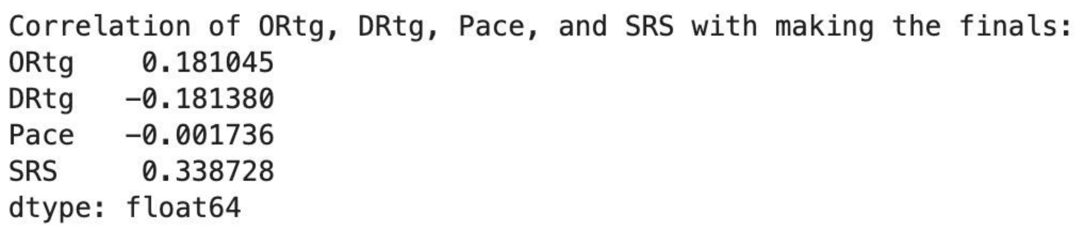
Offensive Rating (ORtg), Defensive Rating (DRtg), Pace, and Simple Rating Score (SRS) were chosen for team stats.

Value Over Replacement (VORP), Box Plus Minus (BPM), Player Efficiency Rating (PER), and Win Share (WS) were chosen for the individual stats.

The aim was to obtain a relationship between the separate stat categories and making the finals or not. The Correlation Formula is as follows:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]} \; (Equation\ \#1)$$

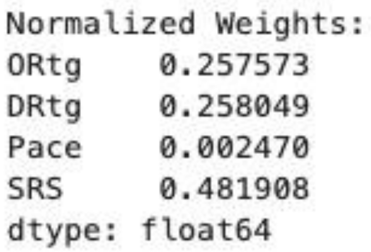
See appendix for explanation of the equation.



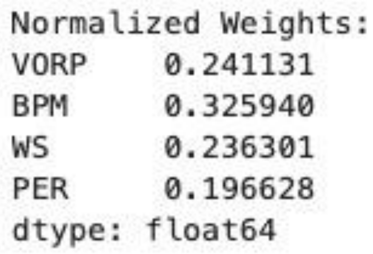
```
Correlation of ORtg, DRtg, Pace, and SRS with making the finals:
ORtg    0.181045
DRtg   -0.181380
Pace   -0.001736
SRS     0.338728
dtype: float64
```

All the calculations done are carried out by Python through the code.

2.32 Normalized Weights



```
Normalized Weights:
ORtg     0.257573
DRtg     0.258049
Pace     0.002470
SRS      0.481908
dtype: float64
```



```
Normalized Weights:
VORP     0.241131
BPM      0.325940
WS       0.236301
PER      0.196628
dtype: float64
```

The correlation of the team stats communicates that offensive and defensive ratings are virtually the same in terms of their relations on finals teams both scoring similar scores of 0.181 and -0.181 respectively. The data shows that SRS (Simple Ratings System) has the strongest correlation to whether a team makes the finals or not. The Simple Ratings System is to show the average margin of victory over opponents and then adjust to opponent strength and other factors (Kubatko, 2008). The league's leader in SRS has won the championship 54% of the time (33 times). A higher correlation could be that SRS is directly attributed to winning as it is capped higher if you win games and lower if you win less games.

The normalized correlation is to help with the overall cohesion and the integrity of the data, making it more legible and easier to interpret. With the data closer to 1 being more impactful. Now the normalized correlation can be used as weights to calculate an overall score for the teams and players.

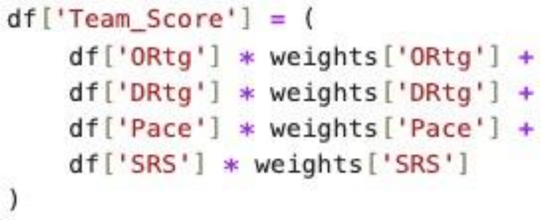
2.33 Individual and Finals Score

After the weights have been obtained, the weights are used to calculate the Team Score and the Individual score. The equation for the weighted sum is:

$$W = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad \text{(Equation \#2)}$$

See appendix for the equation’s explanation.

The code is as shown below:



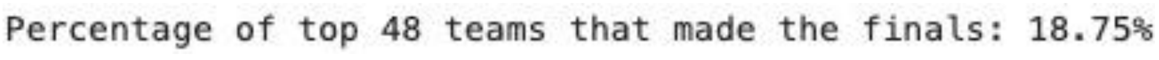
```
df['Team_Score'] = (
    df['ORtg'] * weights['ORtg'] +
    df['DRtg'] * weights['DRtg'] +
    df['Pace'] * weights['Pace'] +
    df['SRS'] * weights['SRS']
)
```

With the scores calculated for each team and player, they are then added to the respective excel sheet under a new column.

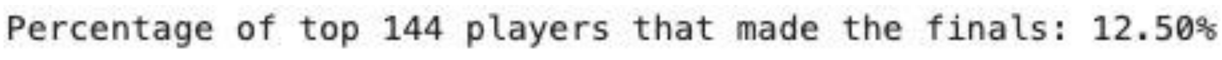
Then both sheets are then rearranged into descending order of Team Score and Individual Score.

The top 48 teams are selected since there are 2 teams that make the finals every year and the data is over 24 seasons (2*24 = 48).

The top 144 players are selected as it was decided that there should be presumed that 3 players from each team are chosen, therefore (2*3*24 = 144). The results are as shown below when they are printed:



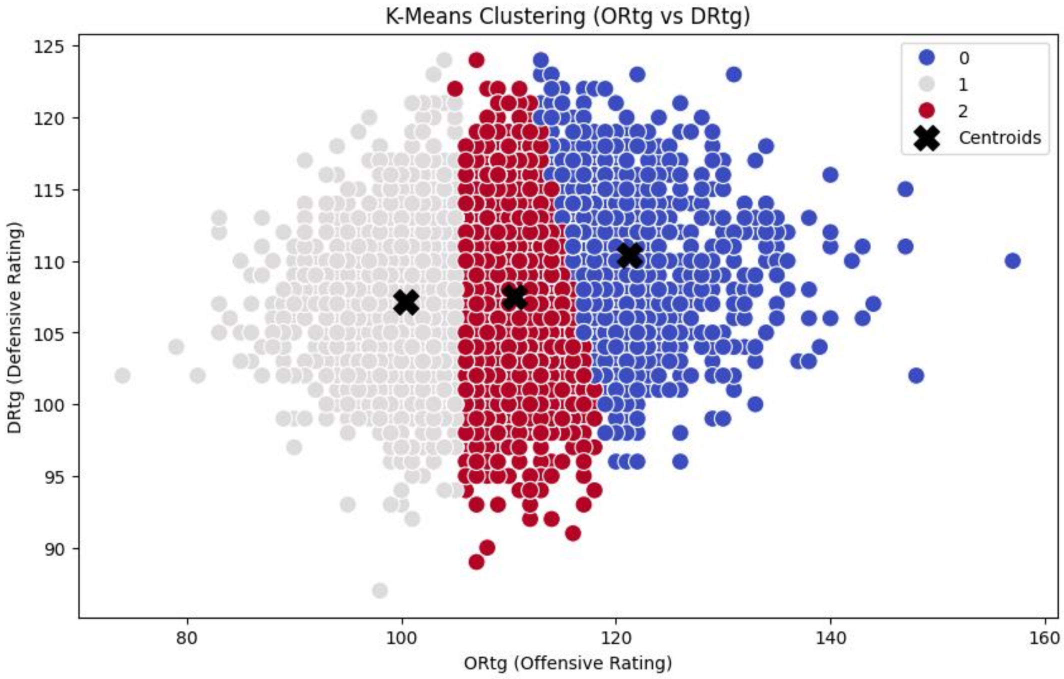
```
Percentage of top 48 teams that made the finals: 18.75%
```



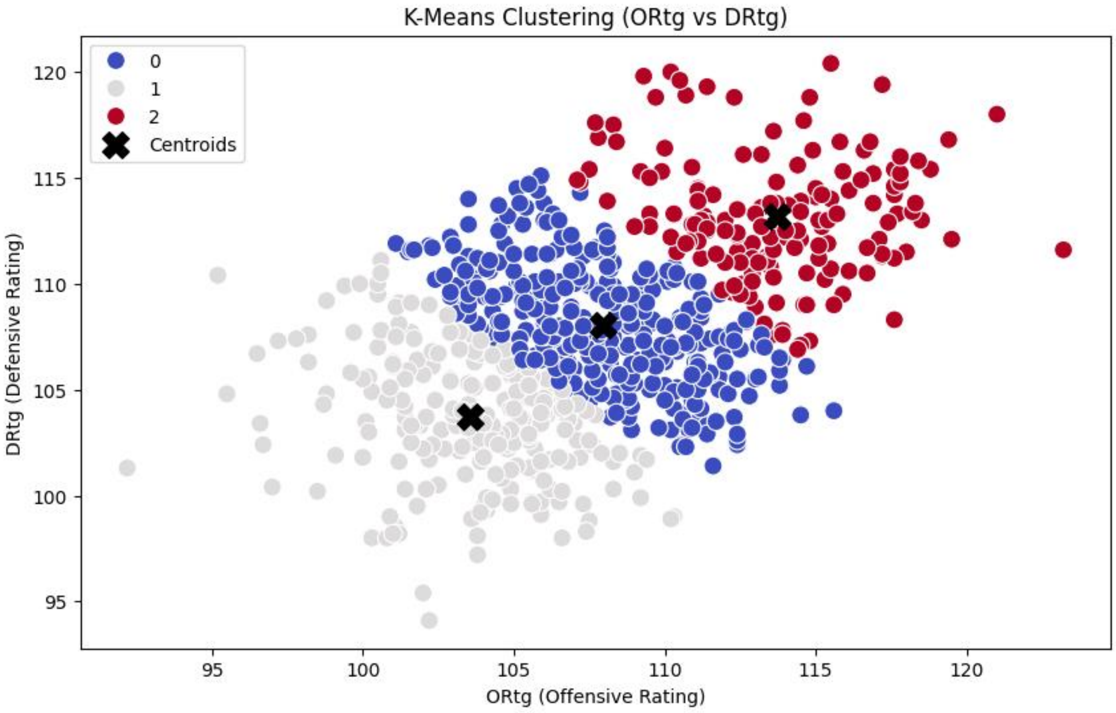
```
Percentage of top 144 players that made the finals: 12.50%
```

2.4 K-Mean Clusters

Then using visualization through k-mean clustering, the data are split into three groups each by machine learning. Three groups were chosen as it is the easiest way of categorizing teams into great, average, and bad teams and players. The ORtg and the DRtg is taken for both the players and the teams as the x and y values.

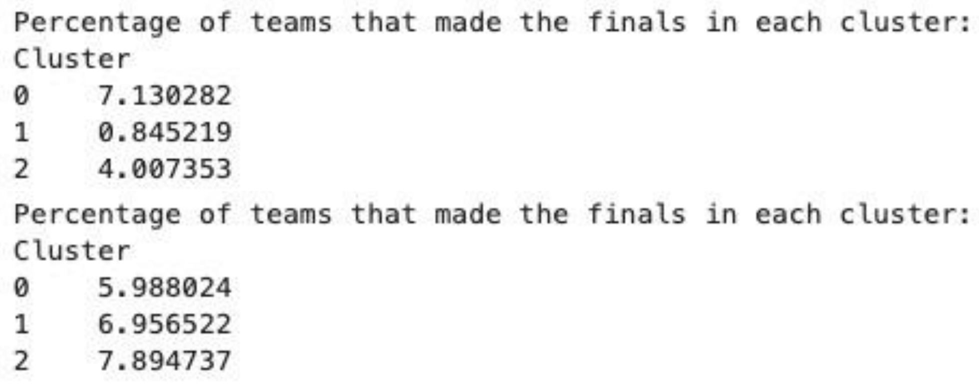


The individuals' clusters display three distinct clusters. The 0 cluster shows a group where offense is more of emphasis as they are the right-most cluster for ORtg, however, a large amount of the data congregates at the top, signifying the players aren't proficient in defense as much, may be due to their focus on offense. For the 2 cluster, it shows a more balanced group where they don't excel in offense or defense. Lastly, the 1 cluster shows players that are more defensive-minded and are significantly less in offensive production.



For the team cluster, there is a more linear relationship that can be examined. Cluster 2 is the teams that have the worst defense, as they are significantly more offensive focused. This might be due to the change in style of play with the 3-point era ushering in over the last decade, where defense is less of a focal point for teams. Cluster 0 shows the more balanced teams, with a cluster of teams that is adequate in defense towards the right and them also being average in defense. Cluster three show teams that are very defensive focused and lacks in offense efficiency.

The different clusters show groups of players/teams that made the finals. Despite the percentage being relatively low due to the immense amount of data, the percentages are still compared to one another which can show which cluster showed more impact under team or individual efficiency. The results for the k-mean clusters are shown below (Individual on top and Team on bottom):



```
Percentage of teams that made the finals in each cluster:
Cluster
0    7.130282
1    0.845219
2    4.007353
Percentage of teams that made the finals in each cluster:
Cluster
0    5.988024
1    6.956522
2    7.894737
```

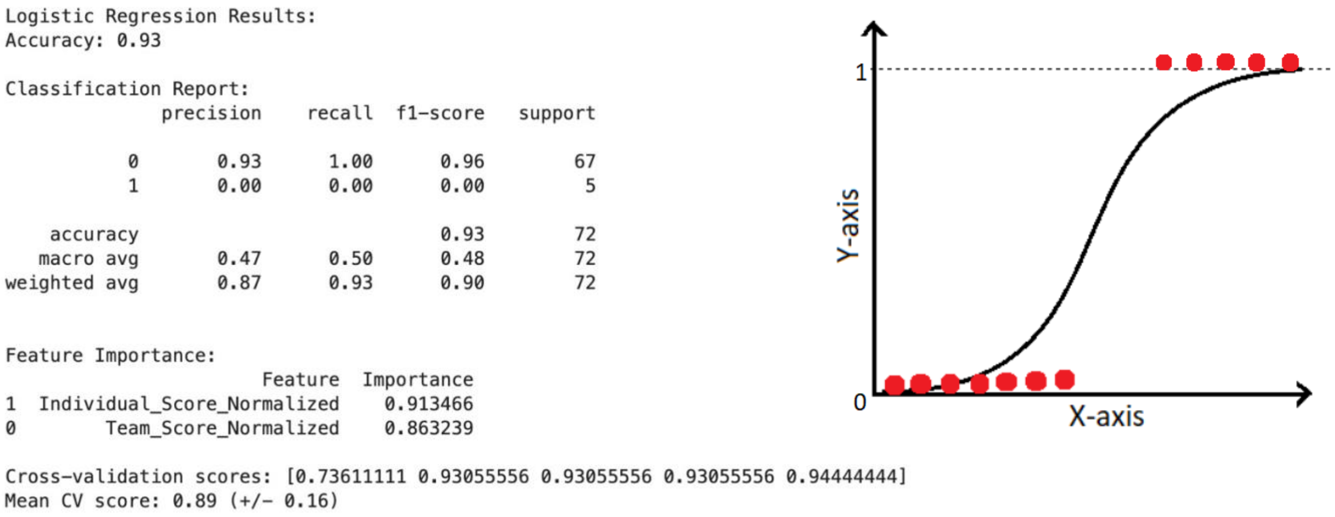
2.5 Logistic Regression

Logistic regression was also used to predict the probability a team makes the finals. The equation for the log odds is as shown:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \ \beta_1 \mathrm{X}\ (\mathit{Equation\#3})$$

Check appendix for the explanation of the equation.

A logistic regression model and the results are shown below:

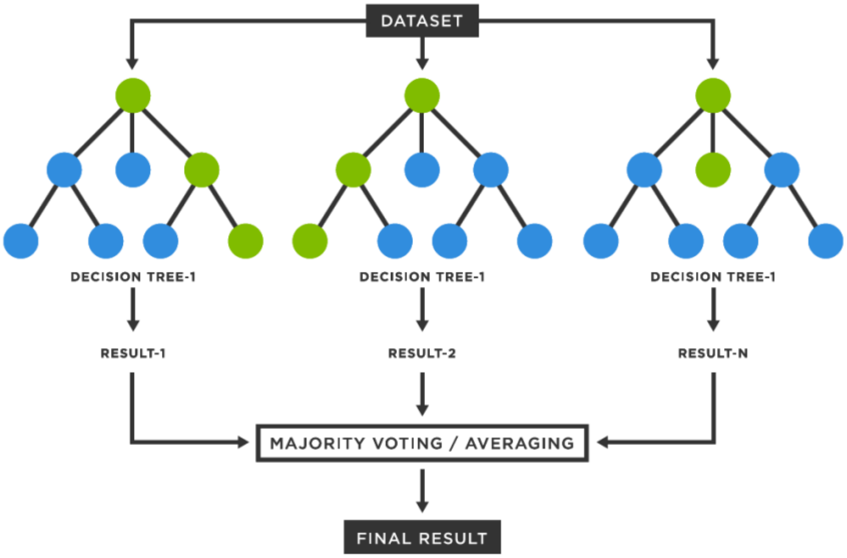


```
Logistic Regression Results:
Accuracy: 0.93

Classification Report:
              precision    recall  f1-score   support

           0       0.93      1.00      0.96        67
           1       0.00      0.00      0.00         5

    accuracy                           0.93        72
   macro avg       0.47      0.50      0.48        72
weighted avg       0.87      0.93      0.90        72


Feature Importance:
                     Feature  Importance
1  Individual_Score_Normalized    0.913466
0        Team_Score_Normalized    0.863239

Cross-validation scores: [0.73611111 0.93055556 0.93055556 0.93055556 0.94444444]
Mean CV score: 0.89 (+/- 0.16)
```

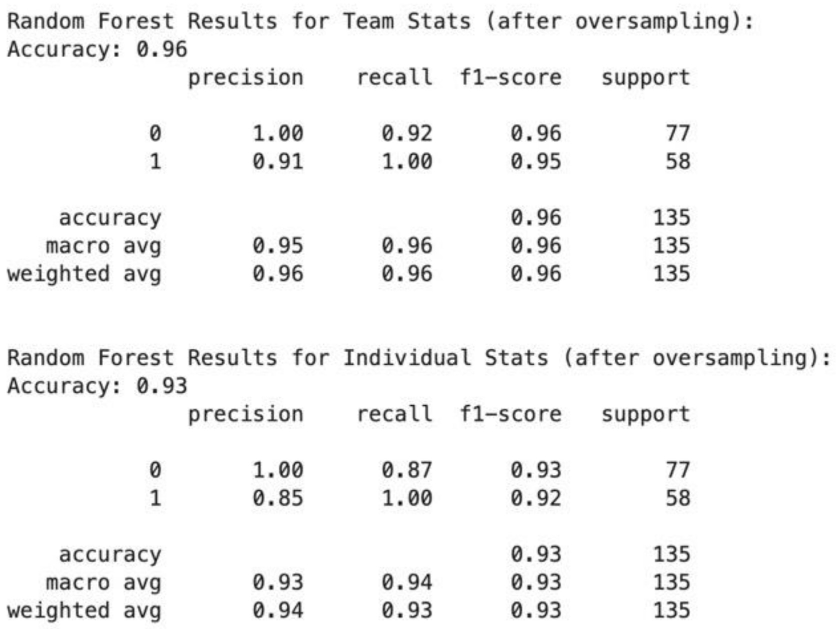
ß is the coefficient that is obtained from the model, in this case 0.913 for Individual Score and 0.863 for Team Score. By applying this to the log odds function shown above a probability will be returned. In this case Individual Score seems to have a higher impact on making the finals when compared to the Team Score. The model itself also had a 93% in terms of its accuracy in predicting if a team can or cannot make the finals. However the classification report yielded a result that it was unable to predict any of the teams that made the finals, this may be due to the immense amount of data in favor of teams that don't make the finals to teams that did.

2.6 Random Forest

Random Forest was also used to predict the outcomes. A Random Forest model is where multiple decision trees are used to produce an average prediction. Each decision tree is trained on random subsets of data to produce an outcome.



In this case the Random Forest is used to predict finals appearances for teams and individuals' stats. The datasets used for the Random Forest is overwhelmingly in favor of teams that didn't make the finals; therefore, the oversampling is used to balance out the data. RandomOverSampler is used so that instances from the minority class, in this case teams or individuals that made the finals, will be duplicated to address the imbalance.



```
Random Forest Results for Team Stats (after oversampling):
Accuracy: 0.96
              precision    recall  f1-score   support

           0       1.00      0.92      0.96        77
           1       0.91      1.00      0.95        58

    accuracy                           0.96       135
   macro avg       0.95      0.96      0.96       135
weighted avg       0.96      0.96      0.96       135


Random Forest Results for Individual Stats (after oversampling):
Accuracy: 0.93
              precision    recall  f1-score   support

           0       1.00      0.87      0.93        77
           1       0.85      1.00      0.92        58

    accuracy                           0.93       135
   macro avg       0.93      0.94      0.93       135
weighted avg       0.94      0.93      0.93       135
```

Both Random Forest for individual and team showed high accuracy with 93% and 97% respectively.

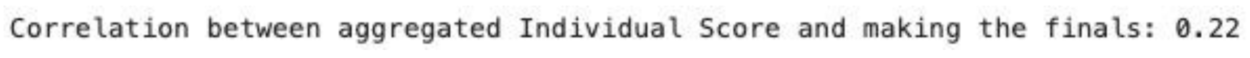
Team stats showed 91% when predicting true positives for finals teams, as well as successfully predicting 100% of finals teams. For the f1-score, team stats were at 95%.

Individually, the stats successfully predicted 0.85% of the true positives, while predicting 100% of finals teams. For the f1-score, individual stats were at 92%.

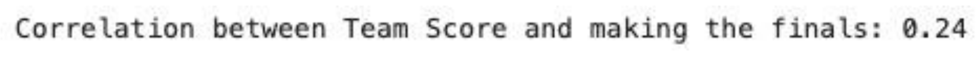
Overall, the team stats yielded a stronger precision in terms of predicting the finals teams with 95% as its f1-score compared to 92% for the individual stats.

2.7 Individual and Team Score Correlation

Lastly, the Individual and Team Scores are taken as a whole to correlate to Finals appearances. For Individual Scores it is aggregated as there are far more players over the 24 seasons than teams. The Individual Score for each player on the same team in one season is added up together and then correlated to the Finals Column. This is so that there is a level of plain field when the two correlations are analyzed. The results are shown below:



```
Correlation between aggregated Individual Score and making the finals: 0.22
```



```
Correlation between Team Score and making the finals: 0.24
```

3. Discussion

Overall, the data suggests that team stats and contributions correlate to making the finals more than individual success does.

Firstly, the top 48 teams and 144 players from 2000-01 to 2023-24 are evaluated, as they are the top ranked when calculated for their Individual Score and Team Score. The results were that 18.75% of the top teams made the finals, whereas 12.50% of the top players made it. Regarding to Team Score and Individual Score, team wise there seems to be more of a winning impact.

Secondly, the k-mean clusters were analyzed to provide a detailed result for different levels of impact. 3 clusters were used individual and team stats, and the ORtg and the DRtg were analyzed. Both were calculated with the team in mind since ORtg and DRtg for individual players are based off the team's ratings when they are on the floor. The results showed that the best clusters (#0 cluster for individual and #2 cluster for team) had the team stats with the 7.89% teams making finals appearances compared to the 7.13% of the individual teams making the finals. The second-best clusters (#2 cluster for individual and #0 cluster for team) had the team stats with the 5.99% teams making finals appearances compared to the 4.01 %. The third-best clusters (#1 cluster both individual and team) had the team stats with the 6.96% teams making finals appearances compared to the 0.85 %. This shows that in each cluster, team stats had higher percentage of teams making the finals, implying that team offensive and defensive efficiency had a higher contribution. Since individual and team ORtg and DRtg are strongly related to one another, it isn't surprising that the results showed similar percentages.

Thirdly, the logistic regression model showed that Team Score had an 86% probability of making the finals whereas Individual Score had a 91% chance. This result contradicts all the other models as this suggests that individual stats are more impactful than team stats. This might be due to the linear nature of the logistic regression model as it doesn't take into account the relationship of predictor variables such as ORtg and DRtg which have interconnectedness. This is why models such as the random forest producing a different outcome with the Team Stats being more favoured when predicting finals teams. Synergy is not considered for logistic regression; therefore, individual stat was displayed to be higher value.

Fourthly, the random forest model predicted that team stats were more accurate and precise when it comes to predicting if a team makes a finals appearance, as it showed a higher precision and recall with the team precision being 6% higher. The calculated f1-score also returned a higher score for team stats as it is 3% higher than the individual stats at 95%.

Lastly, Team Score and Individual Score were correlated to finals appearances, and Team Score had a higher correlation of 0.24 compared to the 0.22 of Individual Score. Although the two correlations coefficients are very close in value, Team Score still display a higher impact to finals appearances.

3.1 Case Study

The results that were obtained indicates that team stats overall were more impactful in terms of making finals appearances. The results that were collected seems to suggest that a different perspective to looking at team or individual success' impact might be more holistic, that the to work in tandem, with team stats being influenced heavily by individual stats but also vice versa. Team success could be argued to be superior with the backing of the data, with many examples of the playoffs to support the claim. Looking at the most recent finals between the Boston Celtics and the Dallas Mavericks, it can be easily argued that Luka Doncic was the best players in the matchup for both teams as the leading scorer of the league during the regular season, with Kyrie Irving next to him contributing significantly during the playoffs, newly acquired well-rounded wing PJ Washington, high-flying wing Derrick Jones Jr., and the impactful rookie Derrick Lively II, the Mavs are a fierce force that rose out of the talented west, with the two All Stars in Irving and Doncic spearheading the team (Basketball Reference, 2024). On the other side of the equation the Celtics have All-NBA talent Jayson Tatum and Jaylen Brown leading the way on dominant force. They also have Jrue Holiday and Derrick White, both being perimeter defensive specialists who also can provide spacing with their elite shooting. Lastly Kristaps Porzingis who is a 7'3 mismatch against anyone in the finals, punishing players with his post up and touch around the rim. Despite the Mavericks arguably possessing the top talent, the starting lineup for the Celtics were all all-star caliber, and each player played into a lesser role of what they could potentially be as the main focal point. The collective sacrifice that the Celtics made outpowered the perennial offensive firepower that the duo of Doncic and Irving brought and won the series easily in 5 games. This showed that although the Mavericks were an excellent team, that had a top ten offense in the league in the regular season, it was almost entirely contributed by Doncic and Irving being the engine. On the other hand, the Celtics had a number 1 offense in the league in the regular season not by piloted by a top offensive superstar, but multiple contributing elite players, who all played into a less significant role to achieve a more cohesive and unified offense and defense. Having a great team will mean a higher chance to win a championship, you can win it without a superstar player with an excellent team. However, you can’t win a championship with a superstar with no support next to them.

3.2 Limitations and Improvements

There are many limitations to my mathematical approach and the models I built. One main limitation to statistics is that there is such a substantial amount of it that it will be hard to include all of it. The selected period that was researched was only over the past 24 years, which excludes the 50 plus years of the NBA that came before, which contained countless valuable stats. Another aspect that limited the ability for statistics to determine something is that it undermines any outside factors. The way players perform and the way games are played have many unseen factors surrounding it that shaped the outcome of games. Injuries, coaching, and other influences are not captured by statistics. The interpretation of the data is also a contested feature of statistics, as the correlation might only suggest a connection with an outcome but doesn't necessarily mean that it is the causation of it. In this particular case, correlation on itself may be faint, as the correlation between the calculated Team Score and the Individual Score are only 0.24 and 0.22 respectively. On its own they are rather slight when they are correlated in making a finals appearance, but in this extended essay correlation is used more of a tool of comparison between individual and collective stats, as well as signifying one of the many factors of winning in basketball.

4. Conclusion

For this extended essay, the goal was to answer the question: Can NBA conference finalists be best determined by individual or team success? I used a variety of numerical models and machine learning models to determine it by analyzing 24 years of data from the 2000-01 to 2023-24 season.

The results showed that team stats have a higher impact on making the finals or not, with SRS being the most impactful stat. On the other hand, individual stats showed a higher value in logistic regression, however for all the other models was less valuable in comparison to team stats, with BPM being the most significant factor as a stat.

To surmise, the outcome this extended essay displayed that team success is a more significant predictor for NBA success, which is making the finals and ultimately winning it. The findings emphasize the notion that basketball is a collective sport, where team productivity excels when compared to individual brilliance. Individual stats such as PER or BPM may highlight a player's impact, the team stats seem to have a more significant effect on the path to victory with stats such as SRS. The results produced are a solid foundation for more discussion and research on the factors of winning in basketball.

5. Appendix

5.1 Code:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.cluster import KMeans
from sklearn.metrics import accuracy_score, classification_report
from sklearn.preprocessing import StandardScaler
from scipy import stats

# Load team and individual data
team_file_path = 'updated_sportsref_with_team_scores.csv'
individual_file_path = 'updated_player_data_with_individual_scores.csv'
df_team = pd.read_csv(team_file_path)
df_individual = pd.read_csv(individual_file_path)

# Inspect the 'Season' column and clean it as needed
def clean_season(season):
    if isinstance(season, str):
        if '-' in season:
            return season.split('-')[0]
        elif len(season) == 4 and season.isdigit():
            return season
    return np.nan

# Apply the cleaning function and remove NaN values
df_team['Clean_Season'] = df_team['Season'].apply(clean_season)

# Convert the cleaned season to numeric (we can use int or str based on preference)
df_team['Clean_Season'] = pd.to_numeric(df_team['Clean_Season'], errors='coerce')

# Drop rows where 'Clean_Season' is still NaN
df_team = df_team.dropna(subset=['Clean_Season'])

# Convert 'Clean_Season' to int for further processing
df_team['Clean_Season'] = df_team['Clean_Season'].astype(int)

# Handle invalid date formats
try:
    df_team['Season_Date'] = pd.to_datetime(df_team['Clean_Season'], format='%Y',
errors='coerce')
except ValueError as e:
    print(f"Error occurred: {e}")

# Drop rows where 'Season_Date' could not be parsed
df_team = df_team.dropna(subset=['Season_Date'])

# Now continue with your analysis as normal
# Function to calculate and print correlation with finals
def analyze_correlation(df, stats, target='Finals'):
    correlation = df[stats].corrwith(df[target])
    weights = correlation.abs() / correlation.abs().sum()
    print(f"Correlation with {target}:")
    print(correlation)
    print("\nNormalized Weights:")
    print(weights)
    return weights

# Analyze team stats
team_stats = ['ORtg', 'DRtg', 'Pace', 'SRS']
team_weights = analyze_correlation(df_team, team_stats)

# Calculate Team Score
df_team['Team_Score'] = np.dot(df_team[team_stats], team_weights)

# Analyze individual stats
individual_stats = ['VORP', 'BPM', 'WS', 'PER']
individual_weights = analyze_correlation(df_individual, individual_stats)

# Calculate Individual Score (if not already done)
if 'Individual_Score' not in df_individual.columns:
    df_individual['Individual_Score'] = np.dot(df_individual[individual_stats],
individual_weights)

# Aggregate individual scores to team level
team_individual_scores = df_individual.groupby(['Team',
'Season'])['Individual_Score'].sum().reset_index()
df_team = pd.merge(df_team, team_individual_scores, on=['Team', 'Season'], how='left')

# Normalize scores
scaler = StandardScaler()
df_team['Team_Score_Normalized'] = scaler.fit_transform(df_team[['Team_Score']])
df_team['Individual_Score_Normalized'] = scaler.fit_transform(df_team[['Individual_Score']])

# Logistic Regression
```

```
X = df_team[['Team_Score_Normalized', 'Individual_Score_Normalized']]
y = df_team['Finals']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

log_reg = LogisticRegression(random_state=42)
log_reg.fit(X_train, y_train)
y_pred = log_reg.predict(X_test)

print("\nLogistic Regression Results:")
print(f"Accuracy: {accuracy_score(y_test, y_pred):.2f}")
print("\nClassification Report:")
print(classification_report(y_test, y_pred))

# This returns probabilities, not class labels
probabilities = log_reg.predict_proba(X_test)

# Example: print the probability of making the finals for the first 5 instances
print(probabilities[:5])

# Feature importance
feature_importance = pd.DataFrame({
    'Feature': X.columns,
    'Importance': abs(log_reg.coef_[0])
})
feature_importance = feature_importance.sort_values('Importance', ascending=False)
print("\nFeature Importance:")
print(feature_importance)

# Cross-validation
cv_scores = cross_val_score(log_reg, X, y, cv=5)
print(f"\nCross-validation scores: {cv_scores}")
print(f"Mean CV score: {cv_scores.mean():.2f} (+/- {cv_scores.std() * 2:.2f})")

# Random Forest for comparison
rf = RandomForestClassifier(random_state=42)
rf.fit(X_train, y_train)
rf_pred = rf.predict(X_test)

print("\nRandom Forest Results:")
print(f"Accuracy: {accuracy_score(y_test, rf_pred):.2f}")
print("\nClassification Report:")
print(classification_report(y_test, rf_pred))

# Time series analysis
yearly_finals = df_team.groupby('Season_Date')['Finals'].sum().reset_index()
plt.figure(figsize=(12, 6))
```

```
sns.lineplot(x='Season_Date', y='Finals', data=yearly_finals)
plt.title('Number of Finals Teams Over Time')
plt.ylabel('Number of Finals Teams')
plt.show()

# Clustering
kmeans = KMeans(n_clusters=3, random_state=42)
df_team['Cluster'] = kmeans.fit_predict(df_team[['ORtg', 'DRtg']])

plt.figure(figsize=(10, 6))
sns.scatterplot(x='ORtg', y='DRtg', hue='Cluster', data=df_team, palette='coolwarm',
size='Finals', sizes=(50, 200))
plt.title('K-Means Clustering (ORtg vs DRtg)')
plt.xlabel('ORtg (Offensive Rating)')
plt.ylabel('DRtg (Defensive Rating)')
plt.show()

# Statistical tests
t_stat, p_value = stats.ttest_ind(
    df_team[df_team['Finals'] == 1]['Team_Score'],
    df_team[df_team['Finals'] == 0]['Team_Score']
)
print(f"\nt-test for Team Score: t-statistic = {t_stat:.4f}, p-value = {p_value:.4f}")

t_stat, p_value = stats.ttest_ind(
    df_team[df_team['Finals'] == 1]['Individual_Score'],
    df_team[df_team['Finals'] == 0]['Individual_Score']
)
print(f"t-test for Individual Score: t-statistic = {t_stat:.4f}, p-value = {p_value:.4f}")

# Correlation heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(df_team[team_stats + ['Team_Score', 'Individual_Score', 'Finals']].corr(),
annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()

# Print summary statistics
print("\nSummary Statistics:")
print(df_team.describe())

# Box plots for comparing finals vs non-finals teams
plt.figure(figsize=(12, 6))
sns.boxplot(x='Finals', y='Team_Score', data=df_team)
plt.title('Team Score Distribution: Finals vs Non-Finals Teams')
plt.show()
plt.figure(figsize=(12, 6))
sns.boxplot(x='Finals', y='Individual_Score', data=df_team)
plt.title('Individual Score Distribution: Finals vs Non-Finals Teams')
plt.show()

# Time series of Team Score and Individual Score
df_team['Year'] = df_team['Season_Date'].dt.year
yearly_scores = df_team.groupby('Year')[['Team_Score',
'Individual_Score']].mean().reset_index()

plt.figure(figsize=(12, 6))
sns.lineplot(x='Year', y='Team_Score', data=yearly_scores, label='Team Score')
sns.lineplot(x='Year', y='Individual_Score', data=yearly_scores, label='Individual Score')
plt.title('Average Team and Individual Scores Over Time')
plt.legend()
plt.show()

# Feature engineering: Offensive-Defensive Rating Difference
df_team['ORtg_DRtg_Diff'] = df_team['ORtg'] - df_team['DRtg']

# Analyze the new feature
print("\nCorrelation of ORtg-DRtg Difference with Finals appearances:")
print(df_team['ORtg_DRtg_Diff'].corr(df_team['Finals']))

plt.figure(figsize=(10, 6))
sns.boxplot(x='Finals', y='ORtg_DRtg_Diff', data=df_team)
plt.title('ORtg-DRtg Difference: Finals vs Non-Finals Teams')
plt.show()
```

```
import pandas as pd
import numpy as np
import re
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import RandomOverSampler
from sklearn.metrics import classification_report, accuracy_score

# Load the team and individual data
team_file_path = 'updated_sportsref_downloadteamdfinal.csv'
individual_file_path = 'updated_player_data_with_individual_scores.csv'
df_team = pd.read_csv(team_file_path)
df_individual = pd.read_csv(individual_file_path)
```

```
# More robust cleaning function using regex to capture only valid years
def clean_season(season):
    if isinstance(season, str):
        match = re.search(r'\d{4}', season)
        if match:
            return match.group(0)
    return np.nan

# Apply the cleaning function
df_team['Clean_Season'] = df_team['Season'].apply(clean_season)
df_team = df_team.dropna(subset=['Clean_Season'])
df_team['Clean_Season'] = df_team['Clean_Season'].astype(int)

# Merge team and individual data
team_individual_scores = df_individual.groupby(['Team',
'Season'])['Individual_Score'].sum().reset_index()
df_team = pd.merge(df_team, team_individual_scores, on=['Team', 'Season'], how='left')

# Normalize scores
scaler = StandardScaler()
df_team['Team_Score_Normalized'] = scaler.fit_transform(df_team[['Finals_Score']])
df_team['Individual_Score_Normalized'] = scaler.fit_transform(df_team[['Individual_Score']])

# Define features and target for the Team Stats model
X_team = df_team[['Team_Score_Normalized']]
y_team = df_team['Finals']

# Define features and target for the Individual Stats model
X_individual = df_team[['Individual_Score_Normalized']]
y_individual = df_team['Finals']

# Step 3: Implement oversampling using RandomOverSampler
ros = RandomOverSampler(random_state=42)

# For Team Stats
X_team_resampled, y_team_resampled = ros.fit_resample(X_team, y_team)

# For Individual Stats
X_individual_resampled, y_individual_resampled = ros.fit_resample(X_individual,
y_individual)

# Train-test split for Team Stats
X_train_team, X_test_team, y_train_team, y_test_team = train_test_split(X_team_resampled,
y_team_resampled, test_size=0.2, random_state=42)
```

```
# Train-test split for Individual Stats
X_train_individual, X_test_individual, y_train_individual, y_test_individual =
train_test_split(X_individual_resampled, y_individual_resampled, test_size=0.2,
random_state=42)

# Random Forest for Team Stats
rf_team = RandomForestClassifier(random_state=42)
rf_team.fit(X_train_team, y_train_team)
y_pred_team = rf_team.predict(X_test_team)

print("\nRandom Forest Results for Team Stats (after oversampling):")
print(f"Accuracy: {accuracy_score(y_test_team, y_pred_team):.2f}")
print(classification_report(y_test_team, y_pred_team))

# Random Forest for Individual Stats
rf_individual = RandomForestClassifier(random_state=42)
rf_individual.fit(X_train_individual, y_train_individual)
y_pred_individual = rf_individual.predict(X_test_individual)

print("\nRandom Forest Results for Individual Stats (after oversampling):")
print(f"Accuracy: {accuracy_score(y_test_individual, y_pred_individual):.2f}")
print(classification_report(y_test_individual, y_pred_individual))
```

5.2 **Definitions**

VORP (value over replacement): According to Basketball reference: “A box score estimates of the points per 100 team possessions that a player contributed above a replacement-level (-2.0) player, translated to an average team and prorated to an 82-game season” (Glossary, n.d.).

BPM (box plus minus): According to an article by Daniel Myers, the creator of BPM, it is “a box score estimates of the points per 100 possessions that a player contributed above a league-average player, translated to an average team.” (Myers, 2020).

PER (player efficiency rating): John Hollinger, the creator of PER, describes it as: "The PER sums up all a player's positive accomplishments, subtracts the negative accomplishments, and returns a per-minute rating of a player's performance." (Calculating PER, n.d.; Myers, 2020).

WS (win share): Is an estimation of the number of wins that is contributed by a player (Glossary, n.d.).

ORtg (offensive rating): According to Dean Oliver, the creator of ORtg and DRtg: "for players it is points produced per 100 possessions; teams it is points scored per 100 possessions (Calculating Individual Offensive and Defensive Ratings, n.d.; Glossary, n.d.).

DRtg (defensive rating): Similarly to ORtg, it is points allowed per 100 possessions

Pace: Estimate of the number of possessions per 48 minutes by a team (Calculating Individual Offensive and Defensive Ratings, n.d.; Glossary, n.d.).

SRS (simple ratings system): Is a rating that considers the average point differential and strength of schedule (The Simple Rating System» Basketball-Reference.com Blog» Blog Archive, 2008).

eFG% (effective field goal percentage): This statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal (Glossary, n.d.; The Simple Rating System» Basketball-Reference.com Blog» Blog Archive, 2008).

5.3 Equations:

$$r = \frac{n(\sum xy)-(\sum x)(\sum y)}{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}\ (\textit{Equation \#1})$$

n	Quantity of Information
Σx	Total of the First Variable Value
Σy	Total of the Second Variable Value
Σxy	Sum of the Product of & Second Value
Σx^2	Sum of the Squares of the First Value
Σy^2	Sum of the Squares of the Second Value

(Correlation Coefficient Formula for Pearson's, Linear, Sample and Population Correlation Coefficients, n.d.)

$$W = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}\ \text{(Equation \#2)}$$

w_i = the weight for each data point.
X_i = the value of each data point.
n = the number of terms to be averaged.
(Frost, 2022)

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \mathrm{X}\ (\textit{Equation\#3})$$

p is the probability of having the outcome.

$\frac{p}{1-p}$ is the odds of the outcome.

β_0 is the intercept of the regression
(Choueiry, n.d.)

6. Bibliography

ADAM FROMAL, JANUARY 27, 2012, *Understanding the NBA: Explaining Advanced Defensive Stats and Metrics.* Bleacher report https://bleacherreport.com/articles/1040309-understanding-the-nba-explaining-advanced-defensive-stats-and-metrics

Stathead Basketball. "Player Season & Career Stats Finder - Basketball." Sports Reference, 2024, https://stathead.com/basketball/player-season-finder.cgi?request=1&order_by=def_rtg&year_min=2024&year_max=2024&qual=mp_per_g_req&display_type=per_poss_100&ccomp%5B1%5D=gt&cval%5B1%5D=1000&cstat%5B1%5D=mp

Sampaio, J., McGarry, T., Calleja-González, J., Sáiz, S.J., Schelling i del Alcázar, X., & Balciunas, M. (2015). Exploring game performance in the National Basketball Association using player tracking data. PLOS ONE, 10(7), e0132894. https://doi.org/10.1371/journal.pone.0132894

Scanlan, Aaron T., et al. "The Underpinning Factors of NBA Game-Play Performance: A Systematic Review (2001-2020)." ResearchGate, Mar. 2021, https://www.researchgate.net/publication/350100390_The_Underpinning_Factors_of_NBA_Game-Play_Performance_A_Systematic_Review_2001-2020

Cabarkapa, D., Deane, M. A., Fry, A. C., Jones, G. T., Cabarkapa, D. V., Philipp, N. M., & Yu, D. (2022). Game statistics that discriminate winning and losing at the NBA level of basketball competition. *PLOS ONE*, *17*(8), e0273427. https://doi.org/10.1371/journal.pone.0273427

Author(s). "Counterpoints: Advanced Defensive Metrics for NBA Basketball." DocsLib, [publication date], https://docslib.org/doc/3691413/counterpoints-advanced-defensive-metrics-for-nba-basketball.

Koster, J., & Aven, B. (2018). The effects of individual status and group performance on network ties among teammates in the National Basketball Association. *PLOS ONE*, *13*(4), e0196013. https://doi.org/10.1371/journal.pone.0196013

Martin, J. (2012, August 1). *What's More Important to NBA Team Success, Depth or Star Power?* Bleacher Report. https://bleacherreport.com/articles/1281737-whats-more-important-to-nba-team-success-depth-or-star-power

The Simple Rating System» Basketball-Reference.com Blog» Blog Archive. (2008, March 10). Www.basketball-Reference.com. https://www.basketball-reference.com/blog/indexba52.html?p=39

Basketball Reference. (2024). *2024 NBA Finals - Mavericks vs. Celtics | Basketball-Reference.com*. Basketball-Reference.com. https://www.basketball-reference.com/playoffs/2024-nba-finals-mavericks-vs-celtics.html

Glen, S. (2024). *Correlation Coefficient: Simple Definition, Formula, Easy Steps*. Statistics How To. https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

Glossary. (n.d.). Basketball-Reference.com. https://www.basketball-reference.com/about/glossary.html#:~:text=VORP%20%2D%20Value%20Over%20Replacement%20Player

Calculating PER. (n.d.). Basketball-Reference.com. https://www.basketball-reference.com/about/per.html

Myers, D. (2020, February). *About Box Plus/Minus (BPM)*. Basketball-Reference.com. https://www.basketball-reference.com/about/bpm2.html

Calculating Individual Offensive and Defensive Ratings. (n.d.). Basketball-Reference.com. https://www.basketball-reference.com/about/ratings.html

Correlation Coefficient Formula For Pearson's, Linear, Sample and Population Correlation Coefficients. (n.d.). BYJUS. https://byjus.com/correlation-coefficient-formula/

Frost, J. (2022, November 26). *Weighted Average: Formula & Calculation Examples*. Statistics by Jim. https://statisticsbyjim.com/basics/weighted-average/

Choueiry, G. (n.d.). *Interpret the Logistic Regression Intercept – Quantifying Health*. Quantifying Health. https://quantifyinghealth.com/interpret-logistic-regression-intercept/