NBA project

Sanat Dusad (21352231)
Muhammad Zubair Khan (21338981)

A report submitted in part fulfilment of COMP 30780

Data Science in Practice

Demonstrator: Rian Dolphin

UCD School of Computer Science
University College Dublin

May 15, 2024

Table of Contents

Abstract

The NBA project aims to explore various aspects of basketball through data analysis, with the goal of providing insights into player performance, team strategies, and game outcomes. Through meticulous examination of player statistics, team dynamics, and game trends, the project seeks to answer four key research questions:

1) Who are the greatest scorers, playmakers, and defenders of all time?

2) What factors impact draft/trade strategies the most?

3) Has the level of the NBA gotten worse or better?

4) Which performance stats contribute to determining a team's success the most?

By addressing these research questions, the NBA project aims to provide valuable insights for players, coaches, analysts, and fans alike, enhancing understanding of the game and informing strategic decision-making processes within the basketball community. Through comprehensive data analysis and interpretation, the project endeavors to contribute to the ongoing evolution and innovation of basketball on and off the court.

By examining data from several viewpoints, we aim to give a holistic image of the world of basketball. We hope that this project can serve as a helpful resource for professionals and basketball enthusiasts. Through a planned approach we aim to portray some unique insights and to foster a greater understanding of the complex dynamics of the basketball world.

Abbreviations:

1) Points Per Game: **PPG**

2) Assists Per Game: **APG**

3) Rebounds Per Game: **RPG**

4)Minutes Per Game: **MPG**

5)Defensive Rebounds Per Game: **DRBPG**

6)Steals Per Game: **SPG**

7)Blocks Per Game: **BPG**

8)Turnovers Per Game: **ToPG**

9)Assist-Turnover Ratio: **AST/TO ratio**

10)Three Pointers Attempted: **3PA**

11)Three Pointers Made: **3PM**

12)Three Point Percentage: **3P%**

Chapter 1: **Introduction**

The analysis of NBA player statistics has captivated basketball enthusiasts and analysts for decades, offering a window into the dynamic and ever-evolving nature of the game. With the advent of advanced analytics and the proliferation of data collection methods, the study of player performance has become increasingly nuanced and insightful. This project seeks to leverage the vast repository of NBA player statistics to unravel the intricacies of player performance across different eras and illuminate the factors driving success on the basketball court.

Our investigation of NBA player performance spans multiple dimensions, including scoring, playmaking, and defensive capabilities. By utilising historical player statistics dating back to the inception of the league, we aim to discern trends, patterns, and standout performances that have left an indelible mark on the NBA landscape. From the dominant scoring prowess of prolific scorers to the visionary playmaking of masterful facilitators, and the tenacious defensive efforts of stalwart defenders, each facet of player performance offers valuable insights into the evolution of the game.

In addition to examining individual player achievements, we will delve into broader trends and shifts in player performance over time. By contextualizing statistical analyses within the historical backdrop of the NBA, we aim to unravel the impact of various factors such as rule changes, advances in training techniques, and shifts in playing styles. Through this comprehensive exploration, we seek to not only celebrate the achievements of basketball's greatest players but also gain a deeper understanding of the underlying dynamics shaping success in the NBA.

The structure of our report is as follows:

Chapter 2: Project Objectives In this chapter, we delve into the rationale driving our project selection and its potential appeal to various audiences. Additionally, we outline the project's objectives while offering a concise overview of the research questions at hand.

Chapter 3: Related Work

Chapter 4: Data Considerations

Chapter 5: Results

Chapter 6: Conclusions

Chapter 2: Project Objectives

2.1 Objectives & Motivations

Our journey into this project stemmed from a deep-seated love for basketball that runs through our veins and a desire to delve deep into the intricate world of the NBA. From heated debates over our favourite players to late-night discussions about team strategies, basketball has been more than just a sport to us. So when the opportunity arose to marry our passion for the game with our expertise in data science, we couldn't resist diving headfirst into the world of the NBA. Beyond just fandom, this project serves as a testament to our commitment to leveraging data-driven approaches to explore and understand the complex phenomena and statistics involved in this game

But our motivations run deeper than mere fandom. Sure, there's a thrill in flexing our data science muscles to analyze the ins and outs of the NBA, but there's also a genuine desire to uncover the stories lurking behind the numbers. We're not just crunching stats; we're unraveling narratives, uncovering patterns, and peeling back the layers of complexity that make the NBA such a fascinating spectacle. By undertaking this project, we aim to contribute meaningful insights that resonate not only with NBA fans but also with a broader audience curious about the involvement of data science in sport.

Our objectives extend beyond mere analysis; we aspire to captivate and engage a diverse audience, from die-hard basketball enthusiasts to casual sports fans. Through compelling visualizations and clear explanations, we aim to make complex data accessible and intriguing to a wide range of individuals. By sharing our findings in an approachable manner, we hope to spark curiosity and foster a deeper appreciation for the rich tapestry of stories woven within the NBA.

Who's our audience? Well, anyone who's ever cheered for a buzzer-beater or debated the GOAT (Greatest of All Time) over dinner with friends. From die-hard basketball junkies to casual sports enthusiasts, we're aiming to hook them all with our data-driven insights. And hey, if we can excite the general public along the way, all the better! Through compelling visualizations and clear explanations, we aim to make complex data accessible and intriguing to a wide range of individuals. We hope to spark curiosity and foster a deeper appreciation of the game, rather than just thinking that it's just 10 players running on a court.

At the heart of it all are the burning questions that keep us up at night. Who's the greatest scorer of all time? What makes a player truly clutch? These aren't just idle discussions; they're the fuel that propels us forward in our quest for basketball enlightenment. By tackling these questions head-on, armed with nothing but our laptops and a healthy dose of curiosity, we hope to shed light on some of the NBA's most enduring mysteries.

2.2 Research Questions

Here are the research questions that will be guiding our project:

1. **RQ1: Who are the greatest scorers, playmakers, and defenders of all time?** We aim to identify the top performers in the three key aspects of basketball, uncovering the legends who have left an indelible mark on the game.

2. **RQ2: What factors impact draft/trade strategies the most?** By dissecting the data behind draft picks and player trades, we seek to uncover the key determinants that shape team decision-making and recruitment in the NBA.

3. **RQ3: Has the level of the NBA gotten worse or better?** We'll delve into historical performance metrics to assess the trajectory of the NBA's overall quality over time, providing insights into the evolution of the game.

4. **RQ4: Which performance stats contribute to determining a team's success the most?** By analyzing a myriad of performance indicators, we aim to pinpoint the statistical drivers behind a team's triumphs on the court, offering valuable insights for coaches, analysts, and fans alike.

Chapter 3: Related Work

In exploring the landscape of NBA analytics, we encountered several notable projects that have contributed to understanding the intricacies of basketball dynamics. One such project, a YouTube analysis [1], offered insights into player stat correlations, minutes distribution, and the evolving trends within the NBA over the past decade. While informative, this analysis primarily focused on general trends rather than diving into specific performance metrics or predictive modeling.

Another noteworthy endeavor was a machine learning-based sports betting project hosted on GitHub [2]. This ambitious undertaking utilized neural networks to predict game outcomes and betting opportunities with a commendable accuracy of approximately 69% for money lines and 55% for under/overs. However, its primary aim was to provide insights for sports betting enthusiasts rather than offering comprehensive insights into player performance or team strategies.

Similarly, another GitHub repository [3] explored various facets of NBA analytics, including player consistency, shot clock analysis, and defensive metrics. While these analyses provided valuable insights into specific aspects of the game, they lacked a cohesive narrative that tied together different dimensions of player performance and team dynamics.

Our project aims to bridge these gaps by offering a comprehensive analysis that integrates player performance, team strategies, and historical trends to provide a holistic understanding of the NBA landscape. By leveraging advanced statistical techniques and machine learning algorithms, we seek to uncover hidden patterns and insights that go beyond conventional wisdom. Moreover, our emphasis on addressing specific research questions, such as identifying the greatest players of all time and analyzing the impact of performance metrics on team success, sets our project apart as a valuable resource for NBA enthusiasts, analysts, and coaches alike.

This narrative highlights the uniqueness and value proposition of our project compared to existing works, emphasizing our comprehensive approach and focus on addressing specific research questions.

Chapter 4: Data Considerations

4.1 Data Collection

The foundation of any data-driven analysis in the realm of sports analytics relies heavily on the quality and depth of the data collected. For our NBA data science project, we used the nba_api, an inbuilt library in python designed to scrape data directly from the official NBA website. In this report, we delve into the intricacies of our data collection methodology, detailing the categories of data collected, the tools employed, and the processes undertaken to ensure comprehensive and accurate data retrieval.

It was essential to have a thorough grasp of the NBA data landscape prior to beginning the data collection procedure. This required figuring out which major data categories would be useful for our investigation. Our data requirements, in general, fell into the below five categories:

- **Drafts**: Information related to NBA drafts, including draft picks, player selections, and draft outcomes.
- **Players**: Individual player statistics, biographical information, and performance metrics.
- **Teams**: Team-specific data including rosters, standings, and historical performance.
- **Trades** : Data on past trades across the seasons.
- **Game Data**: Detailed statistics and metrics from NBA games, including scores, player performance, and game outcomes.

The `nba_api` library emerged as the cornerstone of our data collection efforts.This Python library provides a seamless interface for interacting with the NBA's official site, allowing us to programmatically access a wealth of data directly from the source. Leveraging this library had many advantages:

- **Official Data Source**: By scraping data from the official NBA website, we ensured the authenticity and reliability of the data collected.
- **Comprehensive Coverage**: The `nba_api` library offers access to a wide array of endpoints, enabling us to retrieve data across various categories with granularity and depth.
- **Ease of Use**: The intuitive design and functionality of the library streamlined the data collection process, minimizing manual intervention and enhancing efficiency.

Lets look at how we scraped data for each of the categories mentioned above :

4.1.1 Drafts

Drafts in the NBA are pivotal events where players fresh out of college are selected by teams to join the league. In this section, we focus on scraping data related to past drafts and information about players, such as height, weight, etc. We utilize two different datasets: past drafts and player information, obtained through the ***drafthistory*** and ***draftcombinestats*** endpoints, respectively, included in the `nba_api` library. Here is what the draft dataset looks like :

	PERSON_ID	PLAYER_NAME	SEASON	ROUND_NUMBER	ROUND_PICK	OVERALL_PICK	DRAFT_TYPE	TEAM_ID	TEAM_CITY	TEAM_NAME
0	1641705	Victor Wembanyama	2023	1	1	1	Draft	1610612759	San Antonio	Spurs
1	1641706	Brandon Miller	2023	1	2	2	Draft	1610612766	Charlotte	Hornets
2	1630703	Scoot Henderson	2023	1	3	3	Draft	1610612757	Portland	Trail Blazers
3	1641708	Amen Thompson	2023	1	4	4	Draft	1610612745	Houston	Rockets
4	1641709	Ausar Thompson	2023	1	5	5	Draft	1610612765	Detroit	Pistons
...	...	...	...	...	...	...	...	...	...	...
8252	79312	Herb Wilkinson	1947	0	0	0	Draft	1610610034	St. Louis	Bombers
8253	79284	Jack Stone	1947	0	0	0	Draft	1610610025	Chicago	Stags
8254	79313	Frank Broyles	1947	0	0	0	Draft	1610610035	Toronto	Huskies
8255	79283	Hank Decker	1947	0	0	0	Draft	1610610025	Chicago	Stags
8256	76773	Harry Gallatin	1947	0	0	0	Draft	1610610024	Baltimore	Bullets

8257 rows × 14 columns

Figure 4.1: raw draft dataset

From the above Fig. 4.1, we can see the different fields in the draft dataset. However, this doesn't include all of them since there isn't enough space to show them together. Below is a list of all columns in the dataset : Fig. 4.2

```
['PERSON_ID',
 'PLAYER_NAME',
 'SEASON',
 'ROUND_NUMBER',
 'ROUND_PICK',
 'OVERALL_PICK',
 'DRAFT_TYPE',
 'TEAM_ID',
 'TEAM_CITY',
 'TEAM_NAME',
 'TEAM_ABBREVIATION',
 'ORGANIZATION',
 'ORGANIZATION_TYPE',
 'PLAYER_PROFILE_FLAG']
```

Figure 4.2: draft dataset columns

We will explain how we cleaned this dataset later on in the 'Data Preparation' section.

4.1.2 Players

To answer our research questions, we needed data on player stats before (during college) and after they were drafted into the NBA. We used the ***playercareerstats*** and ***playercareerbycollege*** endpoints to achieve the above.

Here is what the datasets look like (raw): Fig. 4.3

	PLAYER_ID	SEASON_ID	LEAGUE_ID	TEAM_ID	TEAM_ABBREVIATION	PLAYER_AGE	GP	GS	MIN	FGM	...	OREB	DREB	REB	AST	STL	BL
0	920	1985-86	0	1610612747	LAL	22.0	82	1.0	1542.0	209	...	160.0	221.0	381.0	54	49.0	49.
1	920	1986-87	0	1610612747	LAL	23.0	79	72.0	2240.0	316	...	210.0	405.0	615.0	84	70.0	80.
2	920	1987-88	0	1610612747	LAL	24.0	82	64.0	2636.0	322	...	245.0	465.0	710.0	93	87.0	45.
3	920	1988-89	0	1610612747	LAL	25.0	82	82.0	2510.0	401	...	258.0	481.0	739.0	103	94.0	55.
4	920	1989-90	0	1610612747	LAL	26.0	82	82.0	2709.0	385	...	262.0	450.0	712.0	90	66.0	50.
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
29870	980	2008-09	0	1610612739	CLE	34.0	65	65.0	1765.0	342	...	157.0	333.0	490.0	64	28.0	84.
29871	980	2009-10	0	1610612739	CLE	35.0	64	6.0	1339.0	194	...	114.0	231.0	345.0	48	14.0	50.
29872	980	2010-11	0	1610612748	MIA	36.0	72	51.0	1145.0	162	...	108.0	179.0	287.0	26	23.0	58.
29873	1629597	2019-20	0	1610612740	NOP	24.0	4	0.0	51.0	6	...	3.0	6.0	9.0	3	1.0	1.
29874	1629597	2021-22	0	1610612762	UTA	26.0	1	0.0	5.0	0	...	0.0	0.0	0.0	0	0.0	0.

29875 rows × 28 columns

Figure 4.3: Player stats

Fig. 4.4

	PLAYER_ID	PLAYER_NAME	COLLEGE	GP	MIN	FGM	FGA	FG_PCT	FG3M	FG3A	...	FT_PCT	OREB	DREB	REB	AST	STL	BLK	TOV
0	76984	Brian Heaney	Acadia (CAN)	14	70.0	13	24	0.541666	NaN	NaN	...	0.500000	NaN	NaN	4.0	6	NaN	NaN	NaN
1	78385	Bill Turner	Akron	294	4060.0	603	1481	0.407157	NaN	NaN	...	0.728119	NaN	NaN	1039.0	167	NaN	NaN	NaN
2	77684	Fred Nagy	Akron	50	NaN	94	271	0.346863	NaN	NaN	...	0.670103	NaN	NaN	NaN	68	NaN	NaN	NaN
3	76669	Ned Endress	Akron	16	NaN	3	25	0.120000	NaN	NaN	...	0.533333	NaN	NaN	NaN	4	NaN	NaN	NaN
4	202148	Mickell Gladness	Alabama A&M	26	252.0	25	59	0.423728	0.0	0.0	...	0.500000	14.0	44.0	58.0	5	4.0	20.0	7.0
5	77401	Kevin Loder	Alabama State	148	2094.0	365	791	0.461441	6.0	23.0	...	0.695000	113.0	225.0	338.0	174	67.0	43.0	143.0
6	202407	Elijah Millsap	Alabama-Birmingham	69	1120.0	93	284	0.327464	29.0	104.0	...	0.679245	37.0	154.0	191.0	78	63.0	19.0	78.0
7	101182	Donell Taylor	Alabama-Birmingham	98	834.0	109	276	0.394927	7.0	34.0	...	0.640625	30.0	76.0	106.0	90	47.0	7.0	59.0
8	1628	Alan Ogg	Alabama-Birmingham	80	657.0	75	152	0.493421	0.0	2.0	...	0.568181	48.0	85.0	133.0	13	12.0	58.0	30.0
9	1641998	Trey Jemison	Alabama-Birmingham	18	351.0	49	83	0.590361	0.0	0.0	...	0.823529	38.0	38.0	76.0	16	7.0	16.0	22.0

10 rows × 23 columns

Figure 4.4: Player College stats

As we can see both datasets have a wide variety of stats (not all of them are included in the snippets). This proved to be an invaluable asset in our analysis, offering a diverse array of statistics that enriched our understanding of player performance and contributed to the depth of our research.

4.1.3 Teams

Scraping team data was fairly straightforward as we only had to use one endpoint ***(teamyear-byyearstats)*** which allowed us to collect stats on every team across the years. These stats are similar to the ones we saw in the player dataset.

4.1.4 Trades

We identified player trades by examining instances in the player dataset where the team name associated with a player differed between consecutive rows in our dataset. This discrepancy in team names signaled a player's transfer to a new team, indicating a trade occurrence. By detecting and extracting all such instances of team name changes, we effectively identified and compiled a comprehensive list of player trades, providing valuable insights into player movement and team transactions within the NBA.

4.1.5 Game Data

To collect individual game data we used the ***leaguegamefinder*** endpoint which retrieves data on all the games played by a particular team over the years (from 1982 onwards).We passed the Ids for each team into the function and collected data on all games played by each team.

4.2 Data Preparation

In this section, we will see how we cleaned and refined the collected data across each category to ensure its accuracy, consistency, and suitability for analysis.This crucial phase involved addressing various data quality issues, such as missing values, duplicates, inconsistencies etc, to produce high-quality datasets enabling us to make meaningful insights and robust analysis. Lets see the steps we took and what the datasets ended up looking like after the cleaning process :

4.2.1 Drafts

- We first identified rows with null values and removed them from the dataset as they wouldn't be useful for our analysis.
- Since we were only analysing data after 1996, we condensed the dataset to only include data in the same time frame.
- In older seasons, some teams had different team names so we changed all of them to their current team name to maintain consistency
- Finally we merged the 'player information' dataset with drafts based on ***player_id*** to give us a broader range of stats to work with.

Here is what the dataset looked like after cleaning : Fig. 4.5

	player_id	player_name	season	round_number	round_pick	overall_pick	draft_type	team_id	team_city	team_nickname	...	standing_reach
3	1641708	Amen Thompson	2023	1	4	4	Draft	1610612745	Houston	Rockets	...	103.5
4	1641709	Ausar Thompson	2023	1	5	5	Draft	1610612765	Detroit	Pistons	...	104.0
5	1641710	Anthony Black	2023	1	6	6	Draft	1610612753	Orlando	Magic	...	102.5
7	1641716	Jarace Walker	2023	1	8	8	Draft	1610612764	Washington	Wizards	...	104.5
8	1641707	Taylor Hendricks	2023	1	9	9	Draft	1610612762	Utah	Jazz	...	107.0
...	...	...	...	...	...	...	...	...	...	...	...	...
965	2081	Ernest Brown	2000	2	23	52	Draft	1610612748	Miami	Heat	...	109.5
966	2082	Dan McClintock	2000	2	24	53	Draft	1610612743	Denver	Nuggets	...	107.0
967	2083	Cory Hightower	2000	2	25	54	Draft	1610612759	San Antonio	Spurs	...	101.5
968	2084	Chris Porter	2000	2	26	55	Draft	1610612744	Golden State	Warriors	...	103.5
969	2087	Pete Mickeal	2000	2	29	58	Draft	1610612742	Dallas	Mavericks	...	103.5

926 rows × 33 columns

Figure 4.5: Cleaned draft dataset

Here are the different columns in the dataset 4.6

```
Index(['player_id', 'player_name', 'season', 'round_number', 'round_pick',
       'overall_pick', 'draft_type', 'team_id', 'team_city', 'team_nickname',
       'team_abbreviation', 'organization', 'organization_type',
       'player_profile_flag', 'team_name', 'position', 'height',
       'height_ft_in', 'height_shoes', 'height_shoes_ft_in', 'weight',
       'wingspan', 'wingspan_ft_in', 'standing_reach', 'standing_reach_ft_in',
       'body_fat_pct', 'hand_length', 'hand_width', 'standing_vertical_leap',
       'max_vertical_leap', 'modified_lane_agility_time',
       'three_quarter_sprint', 'bench_press'],
      dtype='object')
```

Figure 4.6: Columns

4.2.2 Players

- The dataset only had a column for the ids of every team so we added a separate column with their respective names
- Many of the columns in the dataset were vaguely named which didn't clearly tell us about what the column contains data on. To resolve this we renamed the vague columns to make them easier to understand
- We removed rows in the dataset where the 'team_name' column was null as they represented players who played for teams outside of the NBA and hence, wouldn't contribute much to

our analysis.

- Similarly for the college dataset we removed rows where most of the important stats (for e.g. three_point%, field_goal% etc) were missing. We then renamed the columns to keep it consistent with the original player dataset

Here is what the two datasets looked like after cleaning : Fig 4.7 and Fig 4.8

	player_id	player_name	team_id	team_name	team_abbr	season	league_id	age	games_played	games_started	...	free_throw_pct
0	920	A.C. Green	1610612747	Los Angeles Lakers	LAL	1985-86	0	22.0	82	1.0	...	0.611
1	920	A.C. Green	1610612747	Los Angeles Lakers	LAL	1986-87	0	23.0	79	72.0	...	0.780
2	920	A.C. Green	1610612747	Los Angeles Lakers	LAL	1987-88	0	24.0	82	64.0	...	0.773
3	920	A.C. Green	1610612747	Los Angeles Lakers	LAL	1988-89	0	25.0	82	82.0	...	0.786
4	920	A.C. Green	1610612747	Los Angeles Lakers	LAL	1989-90	0	26.0	82	82.0	...	0.751
...	...	...	...	...	...	...	...	...	...	...	...	...
29870	980	Zydrunas Ilgauskas	1610612739	Cleveland Cavaliers	CLE	2008-09	0	34.0	65	65.0	...	0.799
29871	980	Zydrunas Ilgauskas	1610612739	Cleveland Cavaliers	CLE	2009-10	0	35.0	64	6.0	...	0.743
29872	980	Zydrunas Ilgauskas	1610612748	Miami Heat	MIA	2010-11	0	36.0	72	51.0	...	0.783
29873	1629597	Zylan Cheatham	1610612740	New Orleans Pelicans	NOP	2019-20	0	24.0	4	0.0	...	0.000
29874	1629597	Zylan Cheatham	1610612762	Utah Jazz	UTA	2021-22	0	26.0	1	0.0	...	0.000

26744 rows × 29 columns

Figure 4.7: Cleaned player dataset

	player_id	player_name	college	games_played	minutes_played	field_goals_made	field_goals_attempted	field_goal_pct	three_pointers_made
0	76984	Brian Heaney	Acadia (CAN)	14	70.0	13	24	0.541666	NaN
1	78385	Bill Turner	Akron	294	4060.0	603	1481	0.407157	NaN
2	77684	Fred Nagy	Akron	50	NaN	94	271	0.346863	NaN
3	76669	Ned Endress	Akron	16	NaN	3	25	0.120000	NaN
4	202148	Mickell Gladness	Alabama A&M	26	252.0	25	59	0.423728	0.0
5	77401	Kevin Loder	Alabama State	148	2094.0	365	791	0.461441	6.0
6	202407	Elijah Millsap	Alabama-Birmingham	69	1120.0	93	284	0.327464	29.0
7	101182	Donell Taylor	Alabama-Birmingham	98	834.0	109	276	0.394927	7.0
8	1628	Alan Ogg	Alabama-Birmingham	80	657.0	75	152	0.493421	0.0
9	1641998	Trey Jemison	Alabama-Birmingham	18	351.0	49	83	0.590361	0.0

10 rows × 23 columns

Figure 4.8: Cleaned college dataset

4.2.3 Teams

- As discussed before, some of the teams had different names in older seasons so we changed them to their current name.
- We also renamed columns to keep them consistent with our other datasets.
- There were only a couple columns which had null values. Since they didn't really have an impact on our analysis, we decided to keep them.

Here is what the cleaned team dataset looks like : Fig 4.9

	team_id	team_name	team_city	team_nickname	season	games_played	wins	losses	win_pct	conference_rank	...	offensive_rebounds
0	1610612737	Atlanta Hawks	Tri-Cities	Blackhawks	1949-50	64	29	35	0.453	0	...	0
1	1610612737	Atlanta Hawks	Tri-Cities	Blackhawks	1950-51	68	25	43	0.368	0	...	0
2	1610612737	Atlanta Hawks	Milwaukee	Hawks	1951-52	66	17	49	0.258	0	...	0
3	1610612737	Atlanta Hawks	Milwaukee	Hawks	1952-53	71	27	44	0.380	0	...	0
4	1610612737	Atlanta Hawks	Milwaukee	Hawks	1953-54	72	21	51	0.292	0	...	0
5	1610612737	Atlanta Hawks	Milwaukee	Hawks	1954-55	72	26	46	0.361	0	...	0
6	1610612737	Atlanta Hawks	St. Louis	Hawks	1955-56	72	33	39	0.458	0	...	0
7	1610612737	Atlanta Hawks	St. Louis	Hawks	1956-57	72	34	38	0.472	0	...	0
8	1610612737	Atlanta Hawks	St. Louis	Hawks	1957-58	72	41	31	0.569	0	...	0
9	1610612737	Atlanta Hawks	St. Louis	Hawks	1958-59	72	49	23	0.681	0	...	0

10 rows × 35 columns

Figure 4.9: Cleaned team dataset

4.2.4 Trades

- There were rows that had 'NaN' values which indicated that the player was not traded that particular season. We simply removed all such rows from the dataset.
- Apart from that, not much cleaning was required.

Here is what the dataset looked like after cleaning : Fig 4.10

	player_id	player_name	trade_year	team_traded_from	team_traded_to
0	920	A.C. Green	1993-94	Los Angeles Lakers	Phoenix Suns
1	920	A.C. Green	1996-97	Phoenix Suns	Dallas Mavericks
4	920	A.C. Green	1999-00	Dallas Mavericks	Los Angeles Lakers
5	920	A.C. Green	2000-01	Los Angeles Lakers	Miami Heat
6	2062	A.J. Guyton	2002-03	Chicago Bulls	Golden State Warriors
...	...	...	...	...	...
12520	1985	Zendon Hamilton	2005-06	Milwaukee Bucks	Cleveland Cavaliers
12521	1985	Zendon Hamilton	2005-06	Cleveland Cavaliers	Philadelphia 76ers
12523	204054	Zoran Dragic	2014-15	Phoenix Suns	Miami Heat
12525	980	Zydrunas Ilgauskas	2010-11	Cleveland Cavaliers	Miami Heat
12526	1629597	Zylan Cheatham	2021-22	New Orleans Pelicans	Utah Jazz

7702 rows × 5 columns

Figure 4.10: Cleaned trade dataset

4.2.5 Game Data

- As mentioned before, we analysed data from 1996 onwards so we condensed the dataset to only include data in the same time frame
- We updated old team names to their most recent name
- Initially there was no column for the name of the opponent so we extracted this information from the 'Matchup' column and made a separate column with the opponent's name
- The 'Matchup' column used 2 different ways of separating the 2 teams in a game : 'vs.' and '@'. We replaced the 'vs.' with '@' to ensure consistency in the column

Finally, here is what the dataset looks like : 4.11

	season_id	team_id	team_abbr	team_name	matchup	opponent_abbr	opponent_name	game_id	game_date	win/loss	...	offensive_rebounds
42	22023	1610612737	ATL	Atlanta Hawks	ATL @ WAS	WAS	Washington Wizards	22300445	2023-12-31	W	...	12.0
43	22023	1610612737	ATL	Atlanta Hawks	ATL @ SAC	SAC	Sacramento Kings	22300431	2023-12-29	L	...	18.0
44	22023	1610612737	ATL	Atlanta Hawks	ATL @ CHI	CHI	Chicago Bulls	22300408	2023-12-26	L	...	12.0
45	22023	1610612737	ATL	Atlanta Hawks	ATL @ MEM	MEM	Memphis Grizzlies	22300393	2023-12-23	L	...	13.0
46	22023	1610612737	ATL	Atlanta Hawks	ATL @ MIA	MIA	Miami Heat	22300384	2023-12-22	L	...	12.0
...	...	...	...	...	...	...	...	...	...	...	...	...
104310	21996	1610612764	WAS	Washington Wizards	WAS @ IND	IND	Indiana Pacers	29600065	1996-11-09	L	...	12.0
104311	21996	1610612764	WAS	Washington Wizards	WAS @ CHH	CHH	Charlotte Hornets	29600054	1996-11-08	L	...	25.0
104312	21996	1610612764	WAS	Washington Wizards	WAS @ SAS	SAS	San Antonio Spurs	29600042	1996-11-06	W	...	14.0
104313	21996	1610612764	WAS	Washington Wizards	WAS @ CLE	CLE	Cleveland Cavaliers	29600015	1996-11-02	L	...	13.0
104314	21996	1610612764	WAS	Washington Wizards	WAS @ ORL	ORL	Orlando Magic	29600004	1996-11-01	W	...	11.0

27350 rows × 31 columns

Figure 4.11: Cleaned Game dataset

In conclusion, the "Data Preparation" process served as a pivotal step in our NBA data science project, laying the foundation for robust analysis and insightful findings. Through cleaning and refinement of the collected data across all categories, we successfully addressed data quality issues and standardized the dataset for comprehensive exploration.

Chapter 5: Results

5.1 RQ1: Who is the greatest scorer, playmaker, and defender in NBA history?

The objective of this question is to find who the greatest scorer, playmaker, and defender of all time is. One might ask, what do these 3 terms mean?

A **scorer** is simply one player who scores a point in a game. A **playmaker** is a player who can create plays, initiate something to get a basket, and create offensive advantages for their teammates. A **defender** is a player who tries to prevent the opposition scorers from scoring a basket, either by contesting shots or by applying pressure to the opponents to make mistakes or wrong judgements.

We've always wondered if there was a way to gauge who the best scorer of all time is. Let's say, in a very unlikely scenario in which aliens were to invade planet Earth, and asked for a 1V1 basketball game with a human, putting the fate of humanity on the line, who would we entrust our lives with? Is there a way to find out?

We decided that a point allocation system based on metrics under each category would be the best way to go about finding who the best player in each category is. The way it works is each category has a number of metrics that players will be given points on based on their ranking in that metric:

Scoring Metrics:

1. Career Points
2. Highest PPG
3. FG%
4. 3PT%
5. FT%

Playmaking Metrics:

1. Most Assists
2. Most APG
3. Least ToPG
4. Highest AST/TO Ratio
5. DRBPG

Defensive Metrics:

1. Steals
2. Blocks
3. STOCKs
4. RPG

5.1.1 Data & Method

Explain in detail the data you used to answer this question. How did you use this data to answer the question?

The main dataset we used to answer this question was the **player dataset**. The dataset contains a wide variety of player statistics such as the seasons they played, their ages in each season, the amount of games they played, their scoring stats, their playmaking stats, and their defensive stats.

To start off with the scoring stats, we firstly calculated **PPG**. This was done by dividing the players' total 'points' by their number of 'games played', giving us the amount of points they scored per game. The **FG%** was obtained by dividing the 'field_goals_attempted' by the 'field_goals_made'. Similarly, the **3PT%** was calculated by dividing the 'three_pointers_attempted' by 'three_pointers_made'. Finally, the **FT%** was calculated by dividing the 'free_throws_attempted' by the 'free_throws_made'.

To explain how we used the data to answer the question it is imperative to explain what the metrics mean.A **field goal** is any shot attempt near the basket or the three point line, not including a free throw, taken by an offensive player in possession. A field goal can be worth two points or three points(if it is a three point shot). A field goal is w A **three point shot** is taken from outside the three point line, which looks like an arc.As the name suggests, it is worth three points Finally, a **free throw** is a shot taken from the free throw line, and is worth one point. It occurs if the offensive player has been fouled in the motion of shooting by an opponent. The graph 5.1 below illustrates an NBA court so readers can have a good idea of what a basketball court looks like :

In terms of the playmaking stats, **APG** was calculated by dividing the total 'assists' by the 'games played' by a player. The **ToPG** per game are calculated by dividing the total amount of 'turnovers' a player has by their 'games played'. Finally, the **AST/TO ratio** is calculated by dividing a players 'assists' by their 'turnovers'.

An **assist** is when a player passes the ball to a teammate and the teammates shot leads directly to a field goal. A **turnover** is when a player handling the ball loses possession of the basketball. The **AST/TO** ratio shows how much assists a player gets before turning the ball over. These 3 stats perfectly portray how good of a playmaker a player is.

Finally, for the defensive stats, the only calculation that needed to be made was the **STOCKs** stat, which was obtained by adding the number of 'steals' a player has to their number of 'blocks'.

A **steal** is when a player legally takes the ball away from an opponent, intercepts a pass, or otherwise obtains possession on the ball following an opponent's turnover. A **block** is when a defensive player legally deflects a field goal attempt from an offensive player to prevent a basket. Finally, a **rebound** is when a player gains possession of the ball after the opponent misses a field goal attempt.

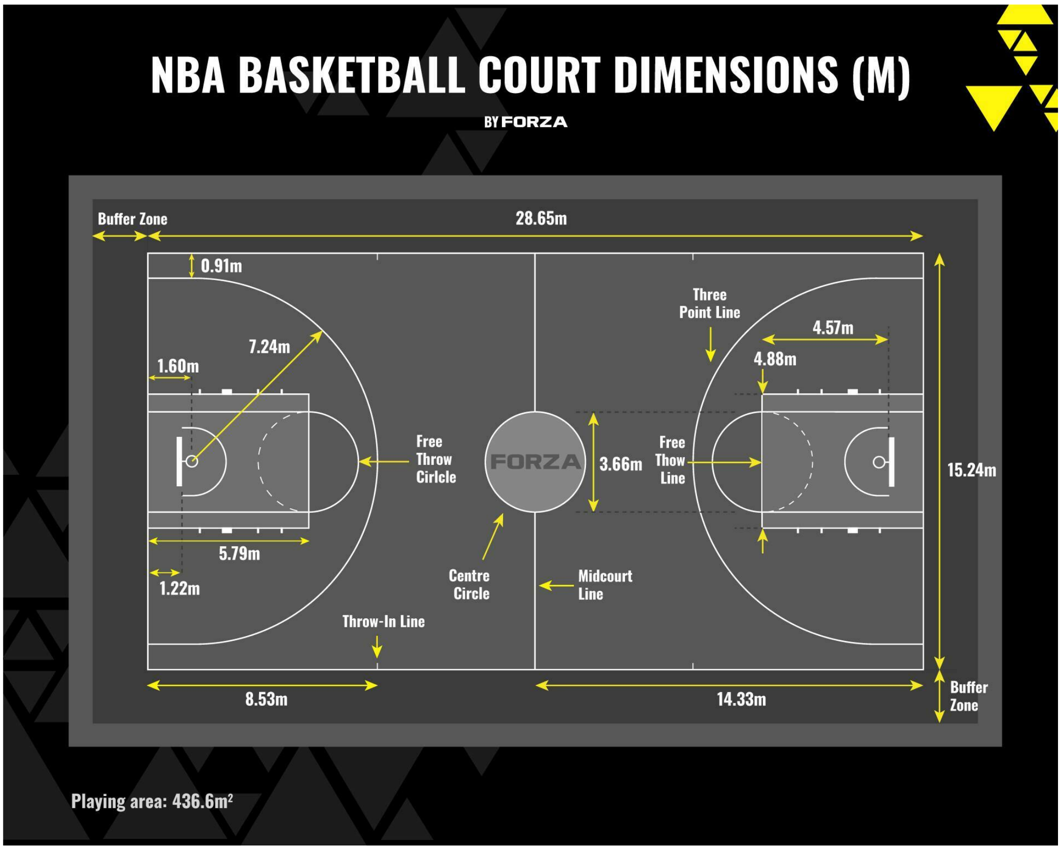

Figure 5.1: Pictorial Representation of an NBA Court

5.1.2 Results

Who is the greatest scorer of all time?

The approach to this aspect of the question was to firstly find out who the 20 players with the most points in their careers are. The idea behind this is to firstly factor in the longevity of a player and their general ability to score points. Longevity is a very important factor when discussing who the greatest player in a particular aspect of the game is. The longer a player stays in the game, the more opportunities they have to leave their mark on the game. Anyone can have a great season, but the real challenge is doing it year after year. Longevity shows consistency, durability, and adaptability – all key ingredients in the recipe for greatness.

The graph 5.2 illustrates who we got to be the 20 players with the most career points:

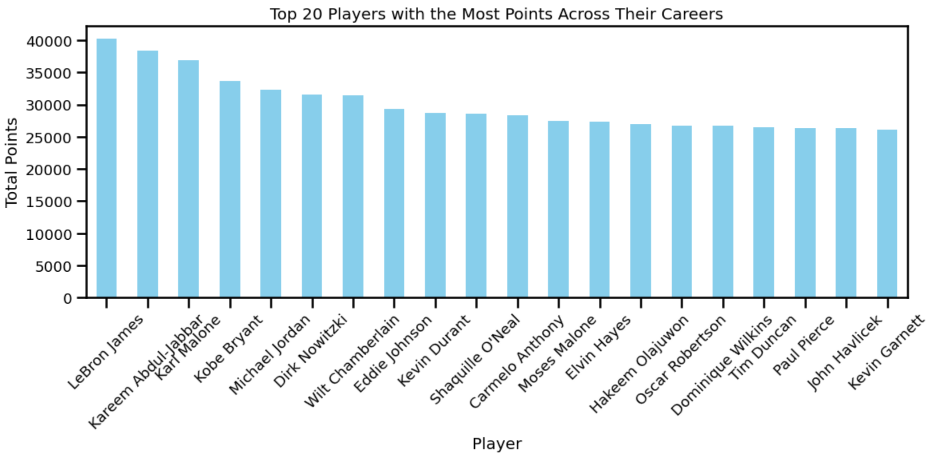

Figure 5.2: Top 20 players with most points

We allocated points to players based on their rankings for the metrics. In the case of the above metric of career points, we can see that LeBron James has the most points, so his rank would be first. In total there is 20 players in the list so LeBron would get 20 points, Kareem Abdul-Jabbar is second, meaning he would get 19 points. Karl Malone is third, meaning he would get 18 points, and so on.

Current Leaderboard:
1)LeBron James: 20 points
2)Kareem Abdul-Jabbar: 19 points
3)Karl Malone: 18 points

The next thing we did was to obtain the 20 players with the most PPG of all time.In basketball, points are the currency of success, so a high PPG usually suggests that a player is a potent offensive threat. It reflects their ability to consistently put the ball in the basket, whether through jump shots, layups, or free throws. Players with high PPG often have strong offensive skills, such as accurate shooting, good ball handling, and the ability to create scoring opportunities for themselves and their teammates. Overall, a high PPG is a sign of offensive prowess and can be a key metric for evaluating a player's scoring ability. A high PPG stat can show how consistently good a scorer is, but it may not factor in the longevity of a player. The graph 5.3 illustrates who we got to be the 20 players with the highest PPG:

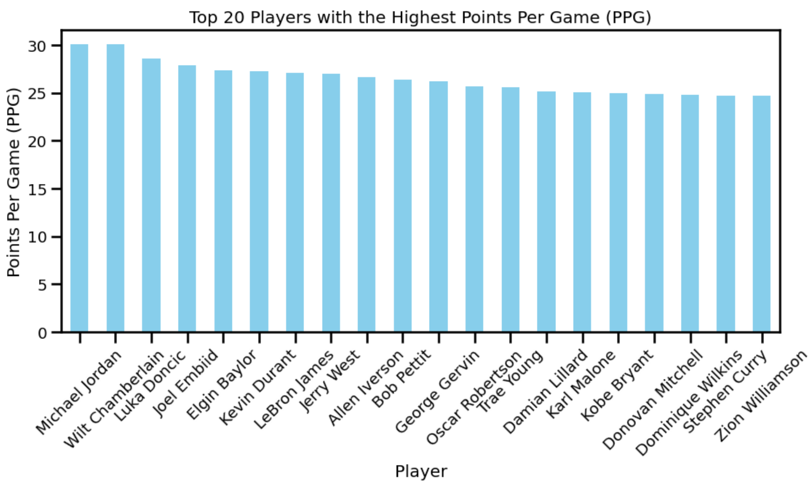

Figure 5.3: Top 20 players with highest PPG

An immediate observation we can make is that LeBron James, Karl Malone, Kobe Bryant, Michael

Jordan, Wilt Chamberlain, Kevin Durant, Oscar Robertson, and Dominique Wilkins are in both lists. Is it a very early assumption to make that the greatest scorer of all time will be one of these players? Let's find out.

Current Leaderboard:
1)Michael Jordan: 36 points
2)LeBron James: 34 points
3)Wilt Chamberlain: 33 points

Next, we combined both players into one list, with players appearing in both lists to appear once. Then we decided to compare the FG% of these players. FG% measures the efficiency of a player's shooting. A high FG% means that a player is making a large proportion of their attempted field goals. This indicates accuracy and efficiency in shooting, as the player is successfully converting their shot attempts into points for their team. Players with a high FG% are often considered reliable and effective scorers because they consistently make shots while minimizing wasted possessions. In essence, FG% reflects a player's ability to score efficiently and contribute to their team's offensive success. It is an important metric for evaluating a player's shooting proficiency and overall offensive impact on the game. The graph 5.4 illustrates which players in the list have the highest FG%:

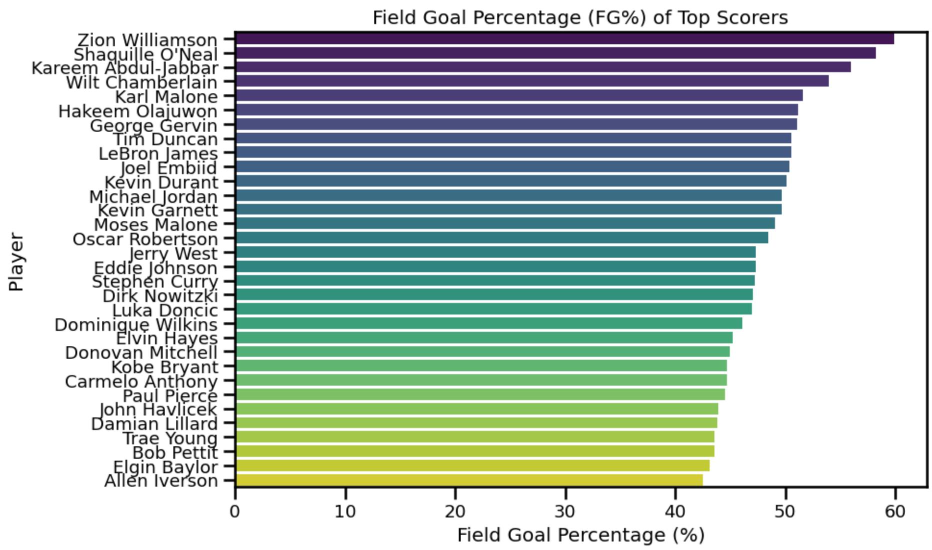

Figure 5.4: Highest FG%s

From the graph 5.4 we can see that Zion Williamson is the most efficient player from amongst the best scorers of all time. However, we realised that as he is injury prone and he's just 23 years of age may indicate that it wouldn't be ideal to call him the most efficient scorer of all time. We can also notice that players who are bigger in size are more likely to have a higher FG%, as they play closer to the rim and get easier shots. Prime examples of these are Zion Williamson, Shaquille O'Neal, and Kareem Abdul-Jabbar as well as Wilt Chamberlain. The players with the lower FG% include Damian Lillard, Allen Iverson, Paul Pierce, etc. We can infer from this that the smaller players tend to take the harder shots and therefore have the smaller FG%.

Current Leaderboard:
1)Wilt Chamberlain: 62 points
2)LeBron James: 58 points
3)Michael Jordan: 57 points

After getting some very surprising results (the Zion one), we decided to dive deeper into the data and gain more insights into why we were getting the results we were getting. We decided to take the 4 players with the most minutes played, and then decided to see how their PPG compares to their MPG. The graph 5.5 shows how as players play more minutes, of course their points scored increases.

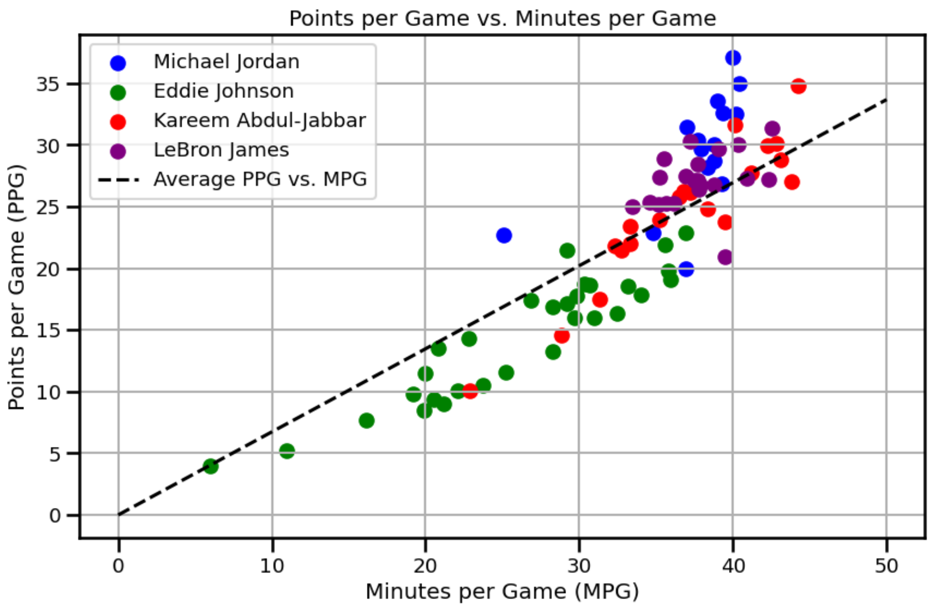

Figure 5.5: PPG vs MPG

Here is another graph 5.6 showing these players' points accumulation as their career goes on:

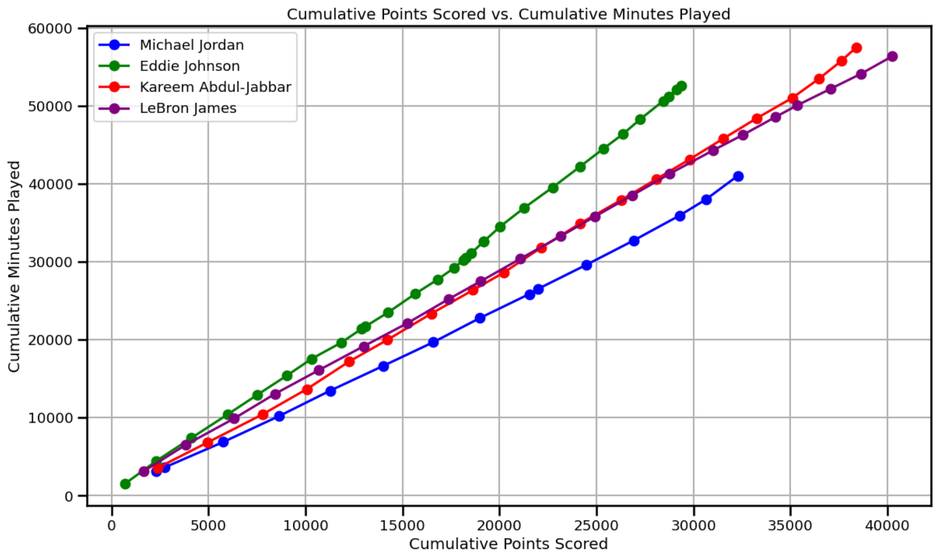

Figure 5.6: Career points progression

We can clearly expect players who play more MPG and minutes throughout their careers to have more points in their careers. Another insight we gained was how FG% varies with PPG. The figure shows a graph 5.7 plotting PPG against FG%:

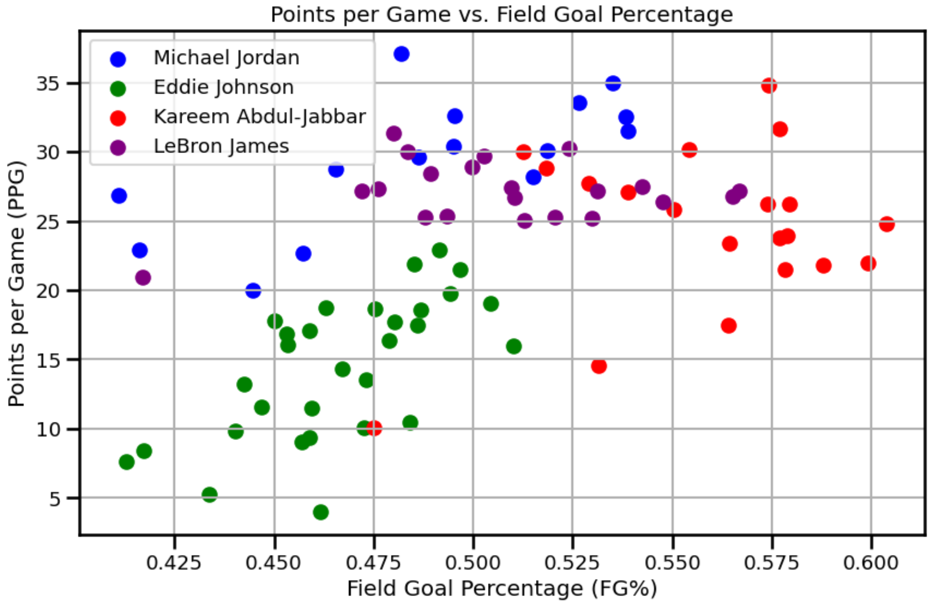

Figure 5.7: PPG VS FG%

From the above graph, we can see a common theme. As a player starts to score more PPG, their FG% decreases. This wasn't a surprise for us as we inferred this could be because of 1)tiredness and exhaustion during the game, also because 2)the risk of missing shots increases as you take more shots. We can prove this second point using a graph 5.8 here:

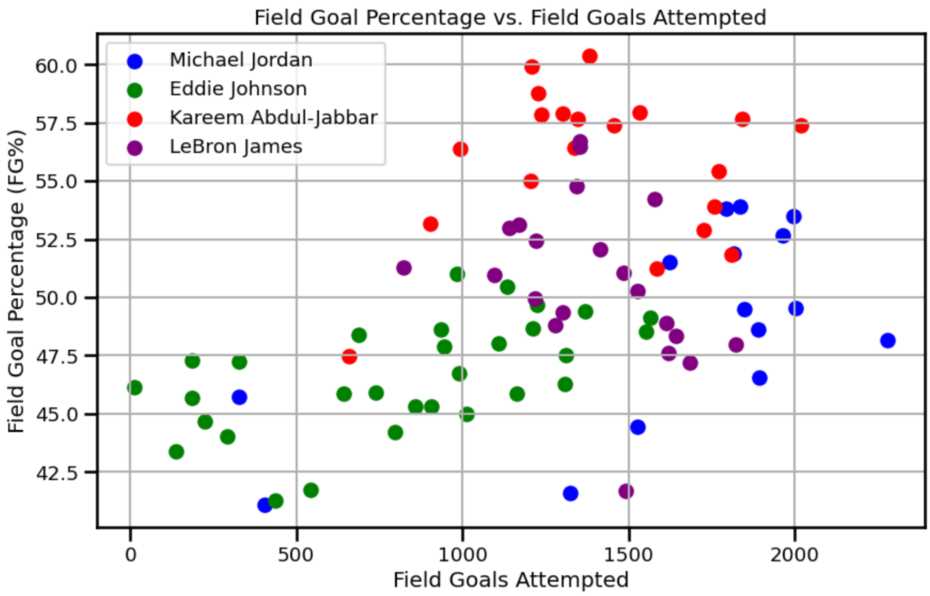

Figure 5.8: FG% VS FGA

We then moved on to 3 point shooting. We started off by analysing the consensus greatest **shooter** of all time, Stephen Curry, and compared his stats to the average player. We found out that he attempts 665.34% more 3-point shots than the average player. We also found that he is 87.34% more accurate at 3-point shots than the average player. Yes, these numbers aren't fake. And here is a graph 5.9 to prove it:

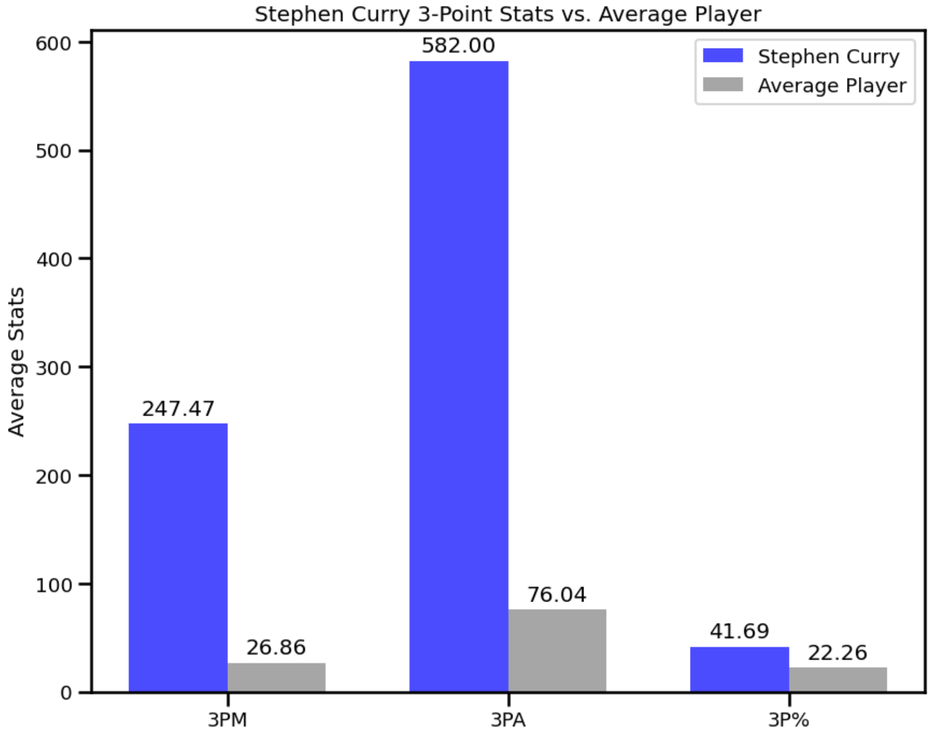

Figure 5.9: Steph Curry VS The Average Player

We then got the top 20 players with the most 3 point shots made. We then obtained their 3PT% and their 3PA and saw how the players differed in their 3PT% in the following graph 5.10:

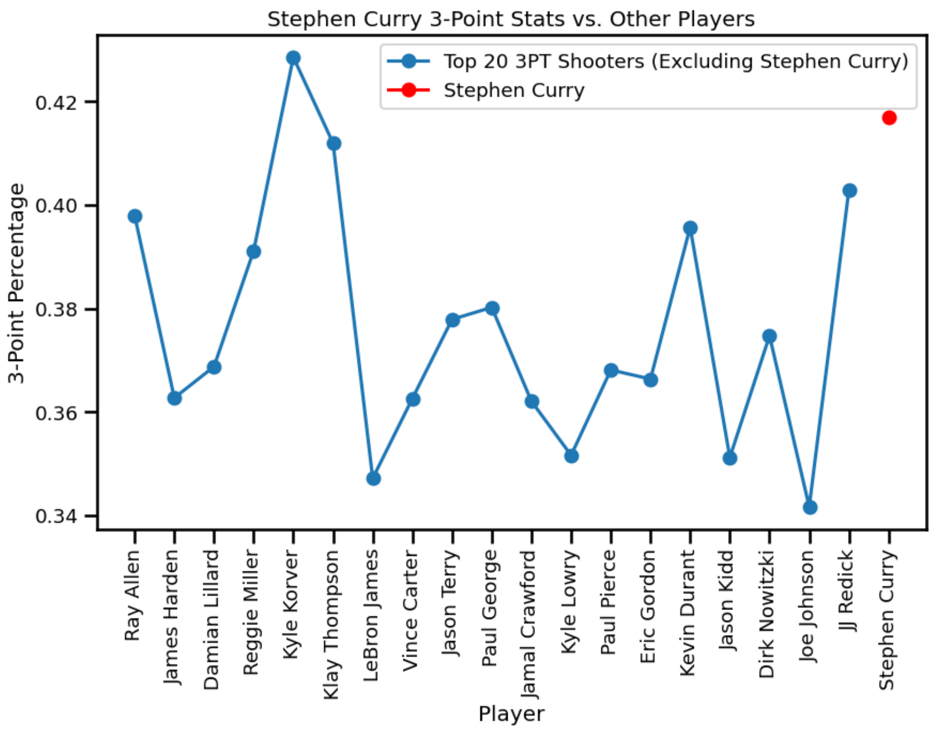

Figure 5.10: Is Steph really the G.O.A.T Shooter?

The graph shows an interesting result; Kyle Korver actually has a higher 3P% than Steph, so we dug deeper into these results and saw that Steph attempts way more 3 point shots than Kyle, which suggests that he shoots more volume in 3 point shots. We can give Steph a little bit of a leeway here as he shoots much more than Kyle yet the difference between their 3PT% is very minimal. It was quite obvious to us at this point that all our assumptions about Steph being an anomaly were true, so it was time to move on to actually addressing the question and seeing how our best scorers fared in their 3pt shooting. We got the following table 5.1 that shows our results:

Table 5.1: Three-Point Shooting Percentage by Player

Player Name	Three-Point Percentage
Stephen Curry	0.416933
Kevin Durant	0.395706
Dirk Nowitzki	0.374714
Damian Lillard	0.368750
Paul Pierce	0.368105
Donovan Mitchell	0.366286
Zion Williamson	0.362000
Trae Young	0.352667
LeBron James	0.347143
Carmelo Anthony	0.345300
Allen Iverson	0.344882
Luka Doncic	0.344000
Joel Embiid	0.343750
Kobe Bryant	0.317700
Eddie Johnson	0.297857
Michael Jordan	0.283933
Bob Pettit	0.273773
Oscar Robertson	0.273773
John Havlicek	0.273773
Jerry West	0.273773
Elgin Baylor	0.273773
Wilt Chamberlain	0.273773
Dominique Wilkins	0.273188
George Gervin	0.263000
Kevin Garnett	0.207136
Karl Malone	0.201158
Elvin Hayes	0.146200
Tim Duncan	0.138474
Hakeem Olajuwon	0.100611
Moses Malone	0.100312
Kareem Abdul-Jabbar	0.033300
Shaquille O'Neal	0.025000

It's not a suprise that Steph Curry is the best 3 point shooter amongst the best scorers of all time. It again goes to show how he's an anomaly as Kyle Korver isn't even in the list of the greatest scorers of all time, yet Steph is.

Current Leaderboard:
1)LeBron James: 82 points
2)Kevin Durant: 80 points
3)Wilt CHamberlain: 78 points

Moving on to the final metric of the scoring aspect, we will be looking through **FT%** using the graph 5.11 :

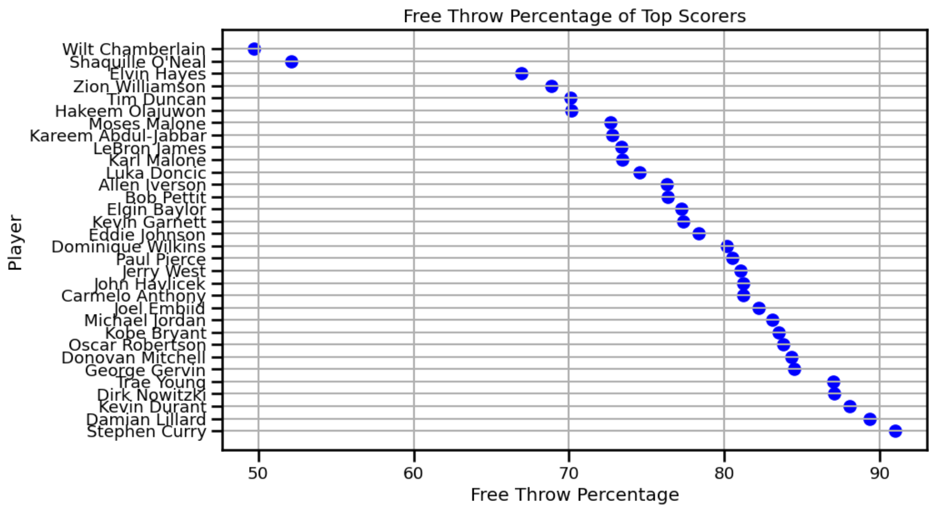

Figure 5.11: FT% of the Best Scorers

Following the results, the player we got to be the greatest scorer of all time was **Kevin Durant**, who ended with 110 points. Michael Jordan was in second place with 97 points closely followed by LeBron James with 91 points. Moving on to the next aspect of this question, **playmaking**. We decided to first get the players with the most assists of all time, as shown in the graph 5.12 :

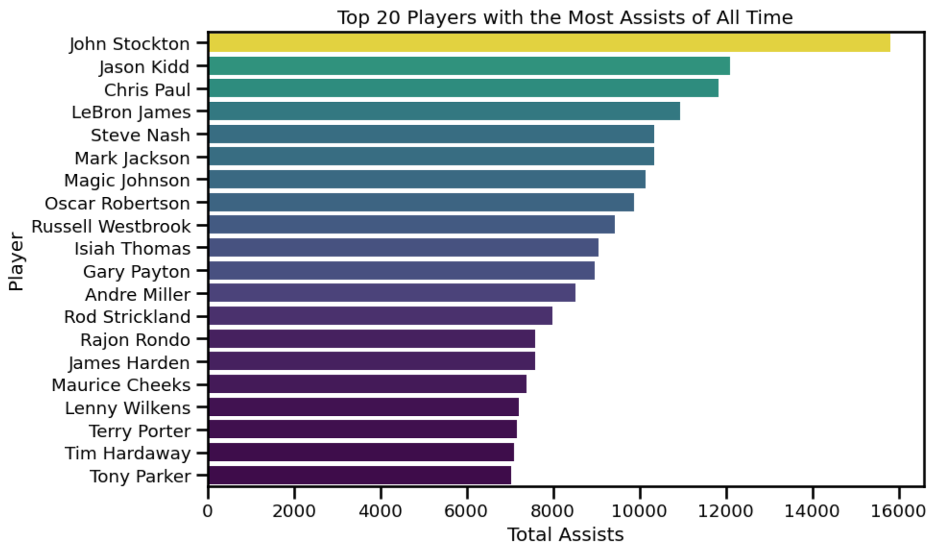

Figure 5.12: Most Assists of all Time

As we can see from our list, John Stockton has the most assists of all time. One interesting observation we can make from this graph is how LeBron James is at 4th position. This is a testament to his longevity and his skill, being a top player in two different categories of the game is no easy feat. However, he just misses out on the leaderboard:

Current Leaderboard:
1)John Stockton: 20 points
2)Jason Kidd: 19 points
3)Chris Paul: 18 points

We then got the players with the highest APG of all time. A player with high APG means that they're adept at setting up their teammates for scoring opportunities by making accurate passes. It indicates that the player has good court vision, passing ability, and an understanding of the game.

High APG values suggest that the player is actively involved in creating scoring opportunities for their team, contributing to their team's offensive success. Overall, a high APG is often associated with effective playmaking and teamwork on the basketball court. We can see the results in the graph 5.13 :

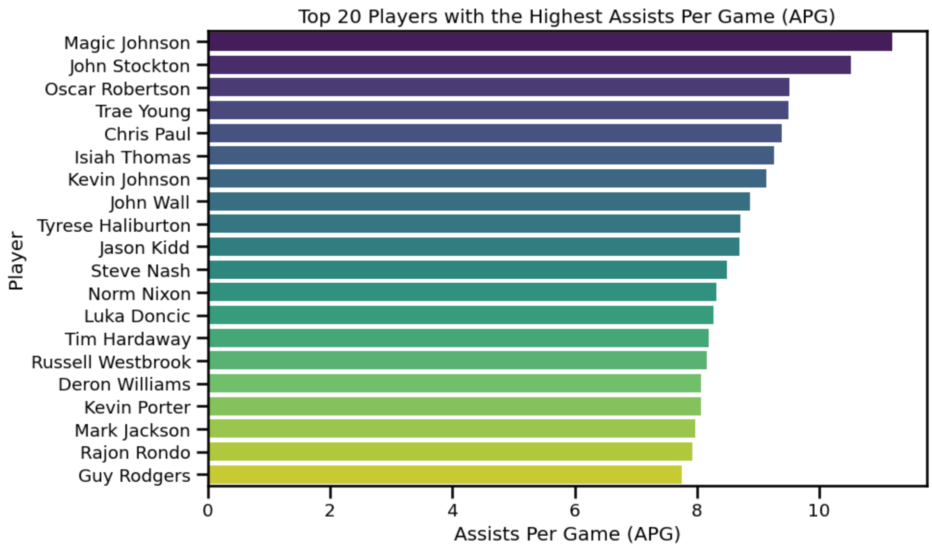

Figure 5.13: Highest APG of all Time

From this graph, seeing that John Stockton is second in APG shows us very encouraging signs that he may be the greatest playmaker of all time. We then combined the lists of the 20 players with the most assists of all time and APG of all time into one list, and players who appear twice would only appear once. The players who appeared in both lists were John Stockton, Jason Kidd, Chris Paul, Steve Nash, Mark Jackson, Magic Johnson, Oscar Robertson, Russell Westbrook, Isiah Thomas, Rajon Rondo, and Tim Hardaway. We can say for sure that these are the best **assisters** of all time.

Current Leaderboard:
1)John Stockton: 39 points
2)Magic Johnson/Chris Paul (Tied): 34 points
3)Oscar Robertson: 31 points

It's important when factoring in the best playmaker, how good you are at handling the ball. You cannot be a good playmaker but turn the ball over too much, as this leads to your opponents having more possessions of the ball. More possessions mean more scoring opportunities. As the primary ball handler(playmaker), you need to turn the ball over as less as possible. Let us analyse which players from our list are the safest ball handlers. We do this by looking firstly at the players with the least ToPG, in the table below:

Table 5.2: ToPG by Player

Player Name	ToPG
Kevin Porter	1.719272
Maurice Cheeks	2.059946
Terry Porter	2.092622
Gary Payton	2.269663
Tyrese Haliburton	2.280632
Tony Parker	2.284689
Chris Paul	2.334125
Andre Miller	2.393405
Mark Jackson	2.434414
Rod Strickland	2.616088
Rajon Rondo	2.702194
John Stockton	2.821809
Deron Williams	2.842604
Steve Nash	2.857847
Tim Hardaway	2.861592
Jason Kidd	2.877786
Kevin Johnson	3.072109
Norm Nixon	3.083333
LeBron James	3.488889
James Harden	3.663227
John Wall	3.727975
Isiah Thomas	3.760981
Magic Johnson	3.869757
Russell Westbrook	3.952381
Luka Doncic	3.984772
Trae Young	4.160891

Current Leaderboard:
1)Chris Paul/ John Stockton: 54 points
2)Jason Kidd: 41 points
3)Steve Nash: 39 points

After turnovers, we looked at the **AST/TO** ratio. A high AST/TO ratio indicates that a player is effective at distributing assists (AST) while minimizing turnovers (TO). In other words, it suggests that the player is adept at creating scoring opportunities for their teammates (assists) without committing many turnovers, which can result in lost possessions for their team. A high AST/TO ratio is often seen as a positive indicator of a player's decision-making, court vision, and overall efficiency as a playmaker. It reflects their ability to facilitate offensive plays while minimizing mistakes, contributing to their team's success on the court. We can see in the graph 5.14 below which players have the best ratios:

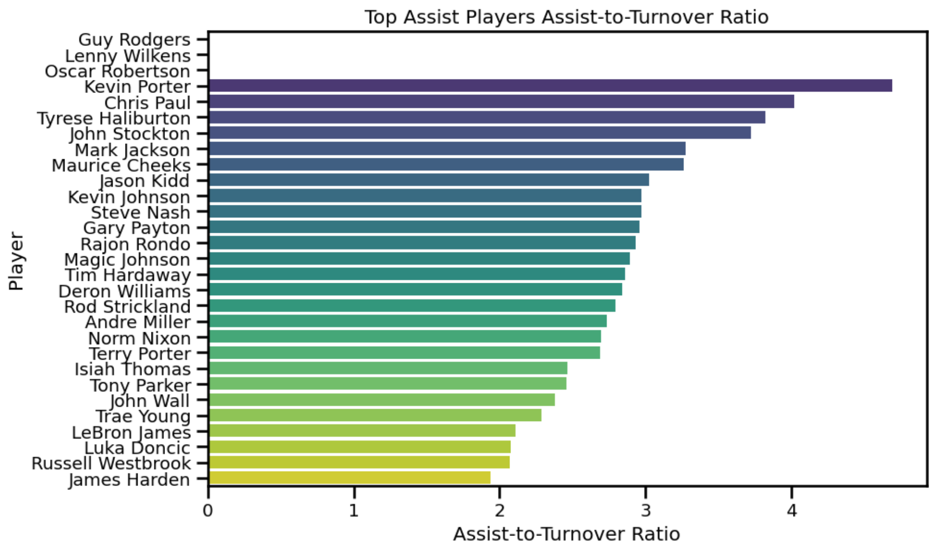

Figure 5.14: Highest AST/TO ratios

Current Leaderboard:
1)Chris Paul: 79 points
2)John Stockton: 77 points
3)Jason Kidd: 61 points

At this point we could see that the race is between Chris Paul and John Stockton. We made an interesting insight along the way with regards to turnovers. As seen in the graph **??** below, players who are their teams' main scorers tend to have more turnovers as well. They don't really share the offensive burden of the team with many people. It would've been great to also see their usage rate of the basketball, which is the percentage of the team's possessions a player ends while they're on the court.

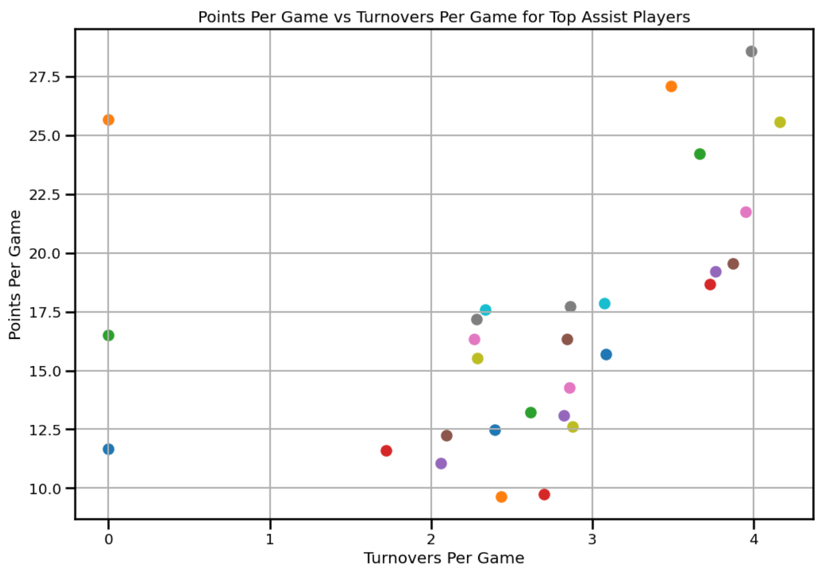

Figure 5.15: PPG VS TO

The final metric in playmaking that we analysed was **defensive rebounds**. Now, there's a case to be made that this can also be a playmaking stat.Players who have a high amount of DRBPG can be great playmakers primarily by initiating offensive opportunities. Players who secure defensive rebounds swiftly transition their team from defense to offense, often initiating fast breaks and catching opponents off-guard. Additionally, they facilitate ball movement with accurate outlet passes, advancing the ball up the court and creating scoring chances. By maintaining possession and limiting opponents' second-chance opportunities, defensive rebounders play a crucial role in extending offensive possessions. Their contributions on the defensive end indirectly translate to offensive success, making defensive rebounds a vital aspect of playmaking in basketball. We can see in the table 5.3 the players with the most DRBPG:

Table 5.3: DRBPG By Player

Player Name	**DRBPG**
Luka Doncic	7.68
LeBron James	6.34
Russell Westbrook	5.54
Magic Johnson	5.47
Jason Kidd	5.00
James Harden	4.85
Chris Paul	3.87
John Wall	3.62
Rajon Rondo	3.59
Tyrese Haliburton	3.03
Trae Young	2.92
Gary Payton	2.90
Mark Jackson	2.84
Deron Williams	2.68
Tim Hardaway	2.67
Rod Strickland	2.66
Andre Miller	2.62
Kevin Johnson	2.61
Isiah Thomas	2.58
Terry Porter	2.48
Steve Nash	2.46
Tony Parker	2.36
Maurice Cheeks	2.16
John Stockton	2.05
Norm Nixon	1.98
Kevin Porter	1.14
Lenny Wilkens	0.26
Oscar Robertson	0.20

From our results, the player we got to be the greatest playmaker of all time is **Chris Paul**, who ended up with 102 points. Jason Kidd came second with 86 points, and John Stockton was surprisingly third with 83 points. Finally, for the **defensive** aspect of the question, we started off by obtaining the players with the most steals and SPG, and combined them into a list. We then got the players with the most blocks and BPG and combined them into a list. We then combined both lists into one, meaning we got a list of players who are the best in the business at defence. We then made a new metric called STOCKs, which is Steals + Blocks. We made a graph 5.16 that plotted the players' STOCKs for their careers:

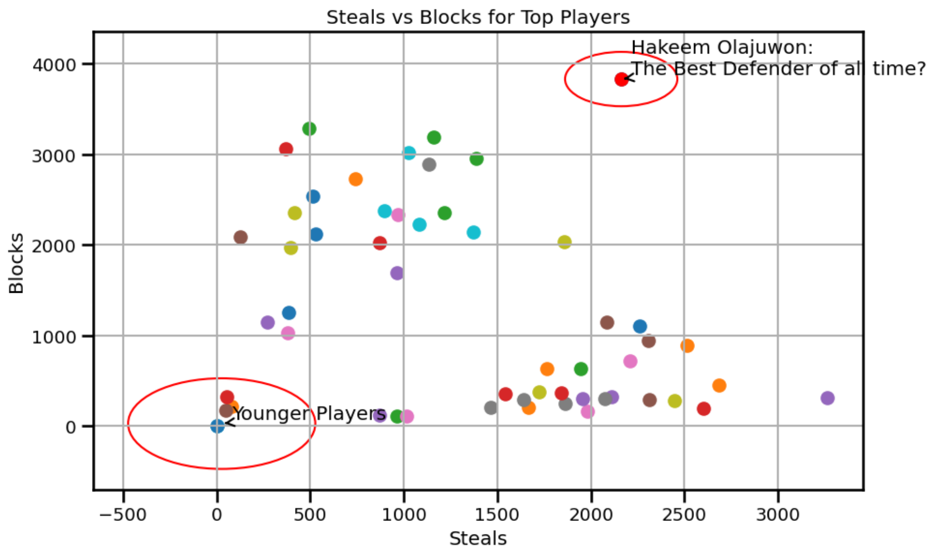

Figure 5.16: STOCKs for the best defensive players

As we can see from the graph, the player with the most STOCKs of all time was Hakeem Olajuwon, who stands alone and by himself on the graph towards the top right hand corner. A surprising thing was that we got a few of the younger players, namely players who have just played a maximum of 2 or 3 seasons in their careers to be in the list. They can be seen in the bottom left corner of the graph. If we were to guess, we would say it's because these players have a high SPG or BPG statistic, because they most definitely cannot compete with players with regards to all-time block and steal stats. We then went on to obtain the players with the most RPG, and the table below shows our results:

Player Name	RPG
Dwight Howard	11.78
Kareem Abdul-Jabbar	11.18
Hakeem Olajuwon	11.11
Shaquille O'Neal	10.85
Tim Duncan	10.84
Hassan Whiteside	10.81
David Robinson	10.64
Anthony Davis	10.62
Bill Walton	10.52
Victor Wembanyama	10.49
Dikembe Mutombo	10.33
Karl Malone	10.14
Kevin Garnett	10.03
Marcus Camby	9.78
Patrick Ewing	9.76

Table 5.4: RPG by Player

The player we got to be the greatest defender of all time, unsurprisingly was **Hakeem Olajuwon**

5.1.3 Discussion

Overall, from the question we can conclude that the greatest scorer of all time is Kevin Durant, the greatest playmaker of all time is Chris Paul, and the greatest defender of all time is Hakeem Olajuwon. Our results have shed light onto the remarkable skillsets of these 3 players. Durant's ability to consistently put up points at an elite and efficient level underscores his status as the most prolific scorers in NBA history. Chris Paul's exceptional playmaking prowess, evident in his high assist-to-turnover ratio and on-court leadership, cements his legacy as one of the premier facilitators in the game. Additionally, Hakeem Olajuwon's dominance on the defensive end, as evidenced by his impressive rebounding and shot-blocking statistics, solidifies his reputation as one of the greatest defenders to ever play the game. These findings not only offer valuable insights into the individual performances of these legendary players but also contribute to the ongoing debate surrounding the greatest players in basketball history.

5.2 RQ2 : What factors impact draft/trade strategies the most?

This research question aims to investigate the key factors that have an impact on draft/trade strategies using a comprehensive data-driven approach. Specifically, we aim to explore the multitude of factors that teams consider when formulating their draft and trade strategies, ultimately aiming to identify the primary determinants that shape these pivotal decisions. By using player statistics, team performance metrics, trade data and past draft outcomes scraped from the nba site (using the **nba_api**), we seek to provide actionable insights and recommendations for improving draft/trade decisions.

5.2.1 Data & Method

Here are the datasets we used to answer this research question :

- **Draft Dataset** :
  - This dataset contains a wide variety of attributes such as player details (for e.g. position), draft details (for e.g. round number, overall pick), team information and player bio metrics (such as height, weight, wingspan etc).
  - Initially, we leveraged this dataset to analyze the popularity of colleges in producing NBA draft picks. By aggregating draft selections based on the what college players went to, we tried to identify what colleges players are drafted from the most.
  - Subsequently, we delved into the positional trends observed in NBA drafts and trades over time.Utilizing the positional data available in the dataset, we examined the distribution of draft picks across different player positions over the years.This analysis provided insights into the evolving preferences of NBA teams when selecting players, offering valuable context for understanding positional dynamics in the draft process.
  - Lastly, we explored the correlation between player attributes and draft round picks to understand the influence of physical and performance metrics on draft selection outcomes. This analysis aimed to uncover the factors that teams prioritize when evaluating and selecting players in the draft, thereby highlighting the determinants that shape draft strategies.

- Player Dataset
  - We used the player dataset to observe the performance trends of players from each position and whether it influenced or explained why teams tend to draft/trade players from certain positions over others.
  - Through our analysis of the player dataset, we aimed to gain deeper insights into the performance dynamics within the NBA and their implications for draft/trade strategies.
- Team Dataset
  - We used the team dataset to see how teams tend to perform before and after a player is traded.
  - From this we extracted the number of successful trades (where a team has more wins after the trade than before) and saw how they varied over time.
- Trade Dataset
  - This dataset contains all the information about trades (for e.g. player name, team traded from, team traded to etc) that have happened over the years.
  - To leverage the insights provided by the trade dataset, we merged it with the previously discussed datasets containing information on team and player performance. This integration facilitated a holistic analysis, allowing us to examine the impact of various aspects of NBA on player trades.

5.2.2 Results

Results for Drafts

We first looked at what colleges had the most draft picks. The graph 5.17 below illustrates this :

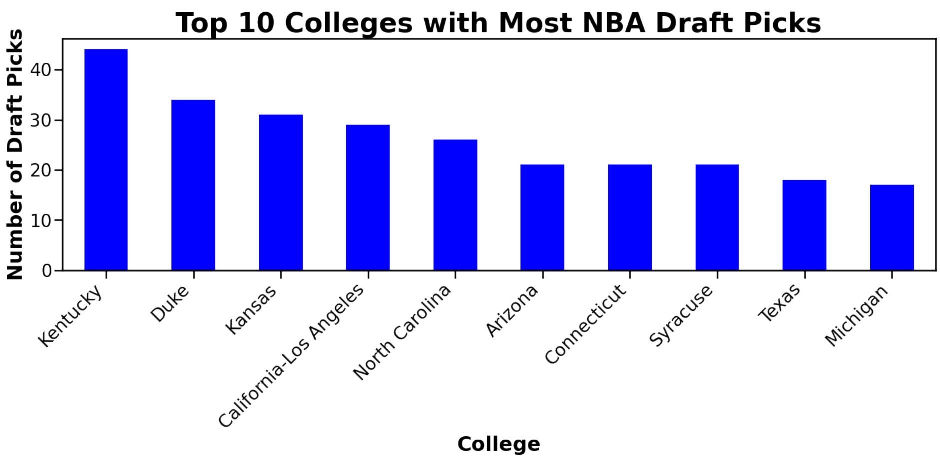

Figure 5.17: Top 10 draft picks

As we can see Kentucky has had the most drafts in total (2000-2023) followed by Duke and Kansas. However, this doesn't show us the year on year trend in draft picks from colleges where we might see a pattern.

So we visualised this further to see the year on year trend of colleges with the most draft picks :

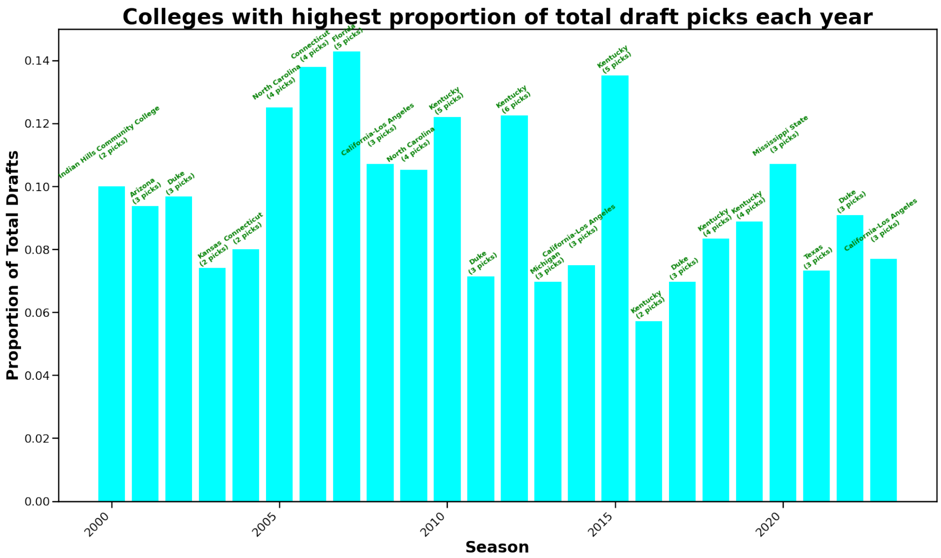

Figure 5.18: Yearly trend in drafts

From Fig 5.18 we can see Kentucky had the highest proportion of total drafts the most (6 years) followed by Duke (4 years). However, it is interesting to see how in the past 4 years Kentucky wasn't the highest even once and there is no particular college that was dominant in this time frame as each year a different college had the most drafts.

Therefore, there has been a recent change in colleges with the most draft picks and no particular college seems to be more popular for drafts than the rest

We then wanted to see if there were any patterns in positions of players drafted, so we plotted the proportion of drafts that came from each position across the years : Fig 5.19

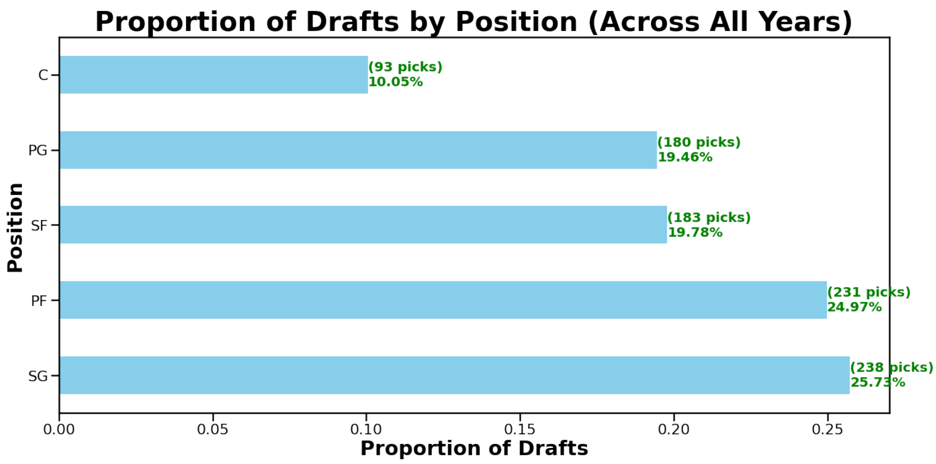

Figure 5.19: Drafts by position

It is clear from graph 5.19 that **Shooting Guards(SG)** and **Power Forwards(PF)** are the most drafted positions across the years. The least drafted position are **Centres(C)**.

To see why this is the case, we used **Points Per Game(PPG)** as a metric to compare the performance of players from each position and visualized the distribution of this metric among players, highlighting any differences or patterns : Fig 5.20

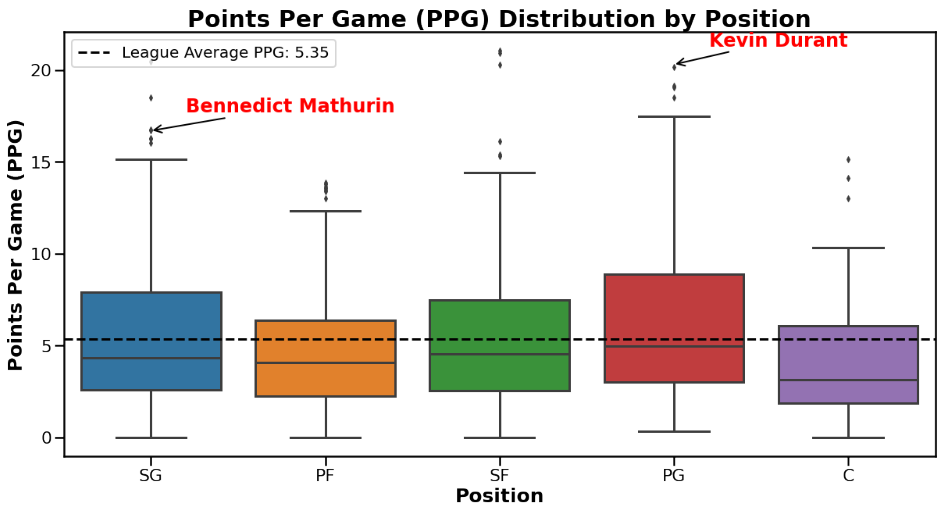

Figure 5.20: PPG Distribution

We can see that **PGs** and **SGs** have the highest PPG whereas **Cs** have the lowest. However it is interesting to see the League Average Points Per Game (PPG) for players in their first year (draft season) is only 5 points. This graph doesn't take into account the amount of time spent on court by these players which may give us a better understanding as to why the average PPG is so low.

So we then normalised points by the number of minutes spent on court and came up with a new metric **Points Per Minute (PPM)** and viewed the distribution of this among players from each position : Fig 5.21

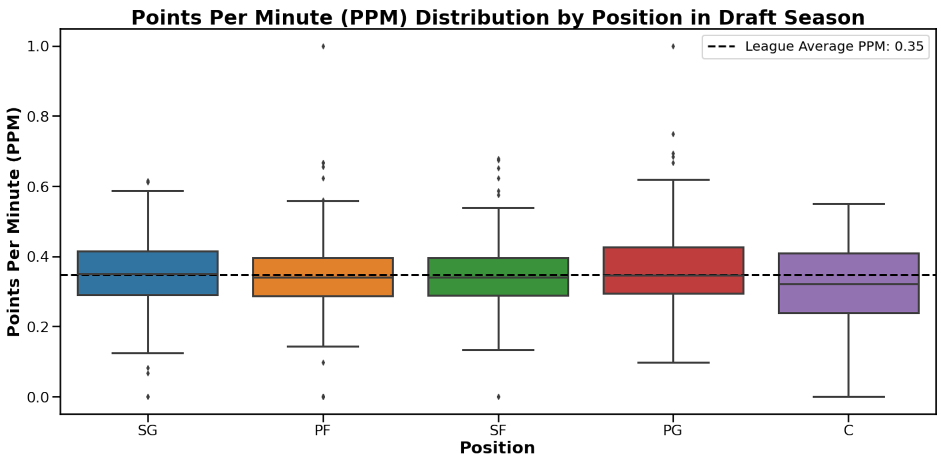

Figure 5.21: PPM Distribution

From Fig 5.21, we can see **Centres(C)** again have the lowest **PPM** which possibly explains why they are the least drafted position.

Finally, we looked at the correlation between draft picks and some of the most important player attributes for a basketball player (such as height, weight and wingspan) : Fig 5.22

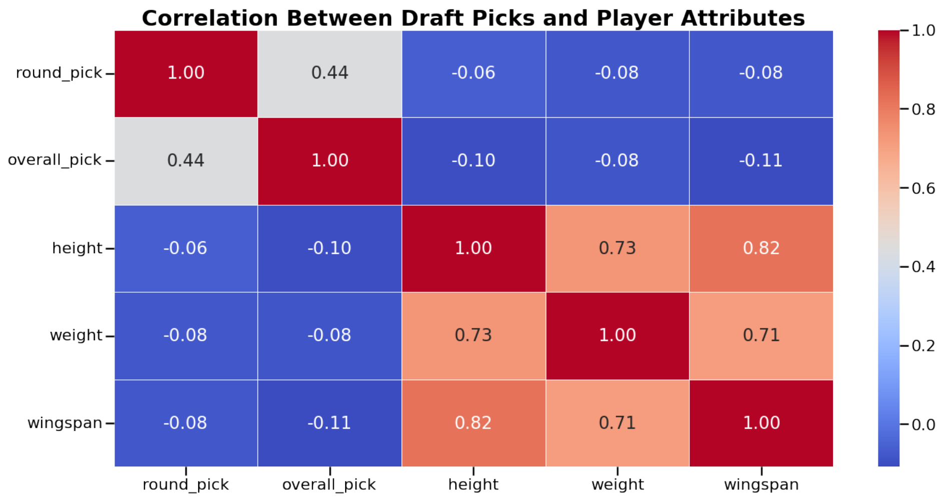

Figure 5.22: Correlation Chart

As expected height, weight and wingspan are strongly correlated. However, surprisingly there actually isn't a significant correlation between the three main physical attributes and draft picks. One possible reason for this could be that NBA teams place greater emphasis on a player's basketball skills, athleticism, shooting ability and overall basketball IQ rather than purely focusing on physical measurements like height, weight, and wingspan.

Results for Trades

We first wanted to see how **Points Per Minute(PPM)** of a traded player compares to the league average which might have an impact on trade strategies. To do this we calculated the Z score for every player and plotted these values to see how they vary from the average league PPM: Fig 5.23

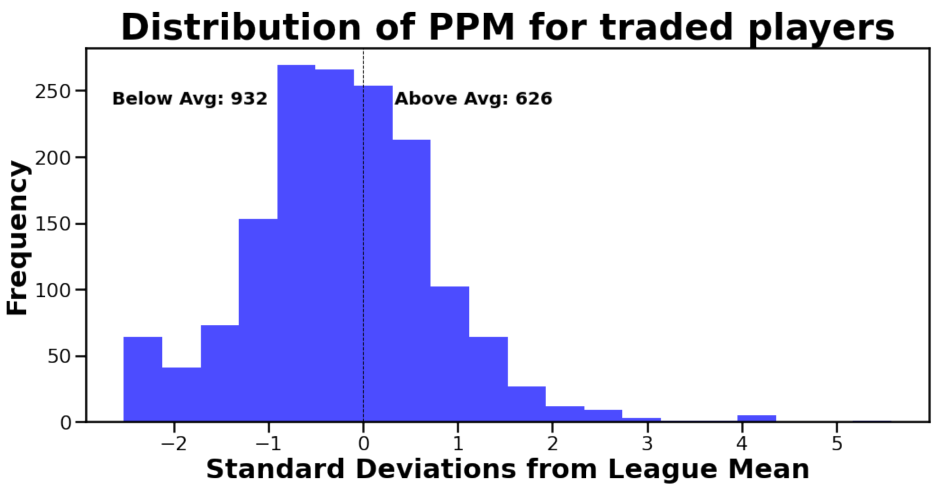

Figure 5.23: PPM Distribution

We would expect most of the trades being players scoring less than the average league PPM but surprisingly this isn't the case as there are a lot of players doing better than the average league performance who end up being traded to a different team. We can see from graph 5.23 that out of all trades, 626 of them were players who actually performed better than the average player in the league.

At the same time it makes sense for these trades to go through for the receiving team as they acquire a player who's performing better than the league average and can potentially add value to the roster.

We then looked at how player's PPM compares to their respective team's average PPM (team being traded to) to see the trend of trades. So we took the top 5 teams in terms of trades and plotted the distribution of trades : Fig 5.24

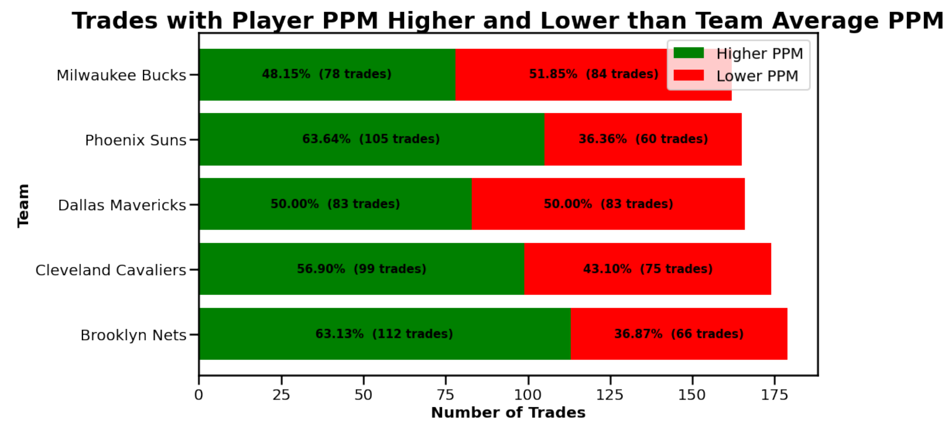

Figure 5.24: Distribution of Trades with regards to Team PPM

From Fig 5.24 we can clearly see that teams tend to accept a lot more trades where the player is performing better than the current team average (PPM) as the proportion of players traded with a higher PPM is in most cases more than that of players with a lower PPM.

Lastly, we looked at the distribution of trades by position in the last 10 years to see if they a significant impact on them : Fig 5.25

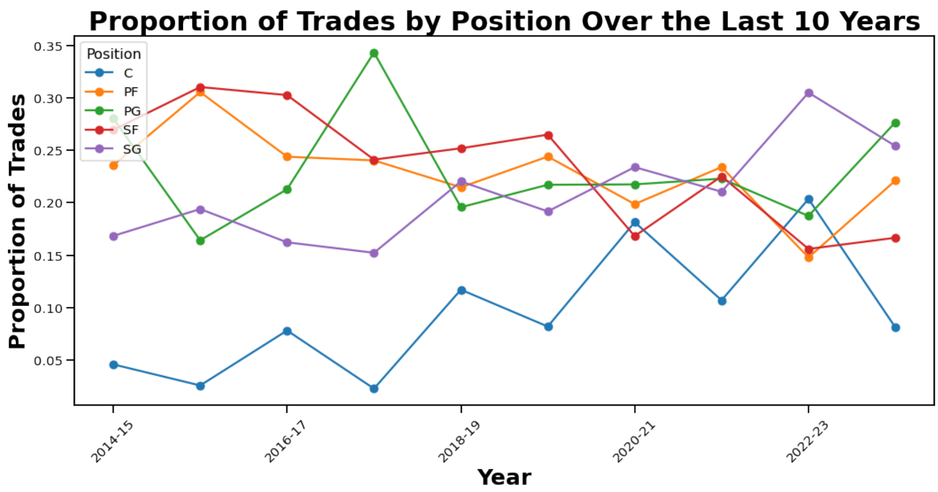

Figure 5.25: Distribution of Trades by position

We then plotted the proportion of players from each position doing better/worse than the league average who end up getting traded : Fig 5.26

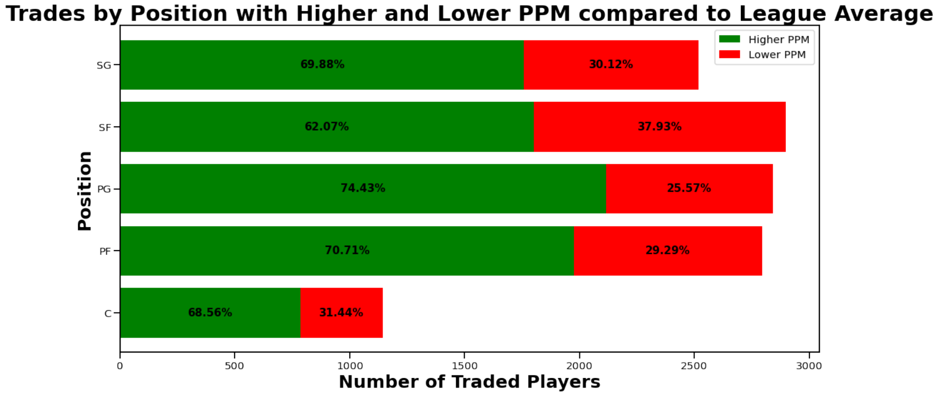

Figure 5.26: PPM Distribution by position

As we can see from both graphs, **Cs** tend to be the least traded position overall (and almost every year as well) whereas **SFs** followed by **PGs** are the most traded position.

Here are a few plausible explanations as to why this is the case :

- **Specialized Role**: Centers often have a specialized role within a team, focusing primarily on rebounding, defending the paint, and scoring close to the basket. Their specific skill set may make them less interchangeable compared to players in other positions.

- **Scarcity of Talent**: Skilled Centers are relatively rare compared to other positions. Teams may be less willing to trade away a valuable Center due to the difficulty of finding a suitable replacement.

- **Team Stability**: Teams tend to build around their Center, considering them a cornerstone of their roster. Trading a Center may disrupt team chemistry and defensive strategies, causing teams to be more cautious about trading players in this position.

However, there's one thing in common between all positions : Majority of the trades coming from each position are players who have a higher 'PPM' than the league average.

5.2.3 Discussion

Conclusion for Drafts

- **Colleges** :
  - We initially thought, after looking at the top 10 colleges with the most drafts (overall, across the seasons) that there would be a few colleges with the highest proportion of draft picks. However, when we looked at the year on year trend of colleges with the highest draft proportion there wasn't any particular college which seemed to be more popular than the others for drafting players as for most of the years a different college had the highest number of draft picks.
  - So is it fair to conclude that colleges dont have an impact on drafts at all? No, it's not accurate to say that colleges absolutely don't affect drafts. Instead, we can say that colleges alone may not be sufficient indicators for optimizing draft strategies. For e.g. : if teams focus solely on what colleges to draft players from, they might overlook other crucial factors in player evaluation, such as individual player skills, athleticism, character, and potential fit within a team's system. Therefore, teams' approach should incorporate multiple sources of talent identification and evaluation to make informed

draft selections and optimize long-term team success.

- **Positions** :
  - When we looked at the proportion of drafts by position, it was clearly evident that 'Centres(C)' were least drafted whereas 50% of the total drafts came from 'Shooting Guards(SG)' and 'Power Forwards(PF)' (25% each) meaning they were the most popular positions that teams looked to draft players from. This was also testified by our analysis on 'Points Per Game (PPG)' for players from each position where we saw that 'Centres' tend to score the least. However, 'Point Guards' had the highest 'PPG' while only being the 3rd most drafted position (19% of total drafts). To see why this was the case we normalised the points based on minutes played and came up with a new metric 'Points Per Minute(PPM)' which would give us a better undertsanding of how players from each position tend to perform. 'Point Guards' again had the highest 'PPM' so the question arises, why are they drafted so less? One possible reason for this could be that their primary focus is facilitating offense rather than being the main scorers. Other positions, such as shooting guards and power forwards, may have more scoring-oriented roles, leading to higher draft percentages.
  - Therefore, we can conclude that positions certainly have an impact on draft strategies. The analysis shows that different positions have varying levels of popularity, performance metrics, and on-court impact. By understanding these factors, teams can tailor their draft strategies to prioritize positions that align with their roster needs, playing style, and long-term goals. Positional analysis helps teams identify undervalued talent, maximize on-court efficiency, and build a balanced roster capable of competing at a high level in the NBA.
- **Physical Player Attributes** :
  - As we saw above, there was barely any correlation between the main physical player attributes and round/overall pick.
  - Therefore, while physical attributes may be considered during the evaluation process, they do not necessarily dictate a player's draft/round pick.

While our study sheds light on the relationship between certain player attributes/stats and draft picks, it's important to acknowledge that our exploration was limited to a fixed set of features. Several other factors, beyond the scope of this project, could also significantly impact draft decisions. For e.g. leadership, work ethic, and coachability can heavily influence team evaluations. Additionally, contextual factors such as team needs and even external influences like media hype and public perception may play pivotal roles in shaping draft outcomes. Therefore, while our analysis provides valuable insights, it's essential to recognize the complexity of draft decision-making and the amount of factors that teams consider when drafting players.

Conclusion for Trades

- **Performance vs. League Average PPM** : Contrary to expectations, a significant proportion of players traded to different teams have a PPM higher than the league average. However, on the flip side it makes sense for teams on the receiving end to accept the trade since the player is performing better than the average person in the league.
- **Performance vs. Team Average PPM** : Our analysis indicates that teams often accept trades where the traded player has a higher PPM than the team's current average which is a quite obvious decision. This suggests that teams are actively seeking to improve their performance by acquiring players who can contribute more efficiently than their existing roster, thereby raising the team's overall performance level. Therefore, a player's performance

with regards to the average performance of the team is most certainly a crucial factor when it comes to trades.

- **Player Position** : Our analysis reveals distinct trading patterns based on player positions in the NBA. **Centers (C)** consistently emerge as the least traded position, while **Small Forwards (SF)** and **Point Guards (PG)** are the most frequently traded positions. This finding reflects the differing functions and relative strategic importance given to various positions on teams' rosters. However, despite the differences in number of trades among positions, a common trend across all positions is the prevalence of trades involving players with higher Points Per Minute (PPM) than the league average. This consistent preference suggests that teams prioritize acquiring players who demonstrate above-average efficiency in scoring, regardless of their specific position.

It is important to acknowledge the fact that our analysis (similar to drafts) does not paint the full picture as there are various factors such as injury history, coachability etc that have not been explored which might potentially impact trade strategies for teams.

Other Considerations :

- **Team Chemistry** : The concept of team chemistry, defined by the cohesion and synergy among players on the court, is crucial in team sports like basketball. Trades that facilitate the integration of players with history of playing in the same team before or familiarity with their new teammates could potentially result in positive outcomes for the acquiring team.
- **Model to predict trade outcomes** : While we initially planned on developing a predictive model for trades, due to the complexity of this idea and time constraints we weren't able to work on it.

5.3 RQ3 : Has the level of the NBA gotten better or worse?

The aim of this question is to delve into the ever-evolving landscape of the NBA to explore whether the NBA's overall quality has seen an improvement or decline over time. With this investigation, we aim to uncover insights into the trajectory of the league's competitiveness, skill, and athletic standards. By analyzing historical performance metrics and trends, we want to provide an understanding of how the NBA has evolved over the years. Ultimately, our goal is to shed light on the dynamic nature of basketball at the highest level and offer valuable perspectives on its past, present, and future.

5.3.1 Data & Method

Explain in detail the data you used to answer this question. How did you use this data to answer the question?

We used four datasets for this question. The first one was the **playerStats** dataset, which contains a wide variety of player statistics such as the seasons they played, their ages in each season, the amount of games they played, their scoring stats, their playmaking stats, and their defensive stats. We then used the **teamStats** dataset, which contains information about team's performances, such as their games played, their win percentage, and then their scoring, playmaking, and defensive stats as a team. We also used the **gameData** dataset, which contains game-by-game stats for teams, which is also known as a box score. Finally, we used the **draftData** dataset, which contains a wide variety of player attributes such as details on their position, their draft pick number, and their biometrics such as height,weight, wingspan, etc.

5.3.2 Results

The way we approached this question was we dived into the scoring aspect of the game to see how it has evolved. Then we analysed the defensive aspect of the game, and finally the playmaking aspect of the game. We capped it off with a case study into a player called Steve Nash, where we compared his 2005/2006 MVP season to other players around that season and into the future.

Starting off with the **scoring** aspect, we obtained the top scorer from the 2023/24 season, and the top scorer from the 1999/2000 season. The top scorer in these seasons were **Luka Doncic** and **Shaquille O'Neal**, respectively. Shaquille O'Neal, who we will be calling Shaq from now on, averaged 29.67 PPG in that season, while Luka Doncic averaged 33.8 PPG in the 2023/24 season. Of course no project is complete without a graph, so we visualised their scoring statistics in the graph 5.27 below:

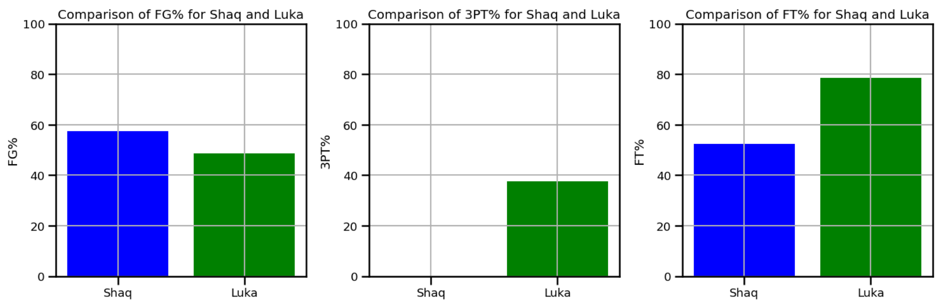

Figure 5.27: Luka VS Shaq

In the first graph on the left we can see that Shaq has a slightly higher FG% than Luka. However, that is due to him being a center and playing closer to the rim. Luka's 3PT% is much higher than Shaq's by a considerable margin. It doesnt seem likely that Shaq has a 0% 3PT%, it just cant be true that the top scorer from a year was so terrible at 3PT shooting. However, when we examined this further we found that Shaq attempted only **one** 3 point shot that **whole** season, and he made **none**. As we can see Luka is a much better FT shooter than Shaq as well. Based on this initial analysis we can see that the top scorer in 2024 is much more skilled and well-rounded than the top scorer in 2000. We got to thinking though, why did Shaq attempt so less 3 pointers? The conclusion we found ourselves at was that when you're 7'1 and weigh 160kg all you need to do is play close to the basket and score the easy points.

However, this lack of 3PT shooting by the top scorer of the year 2000 got us wondering; how has 3PT shooting evolved over the years? We decided firstly to find out how much 3 point shots people took over the years and analyse the trend, as shown in the graph 5.28 below:

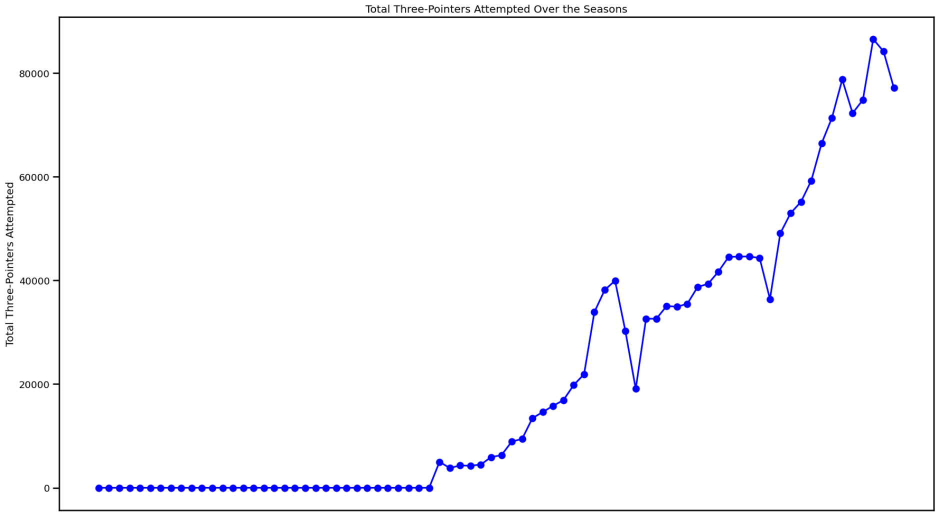

Figure 5.28: Progression of 3PA

We can also see the total 3PM in the graph 5.29 and 3PT% in the graph 5.30:

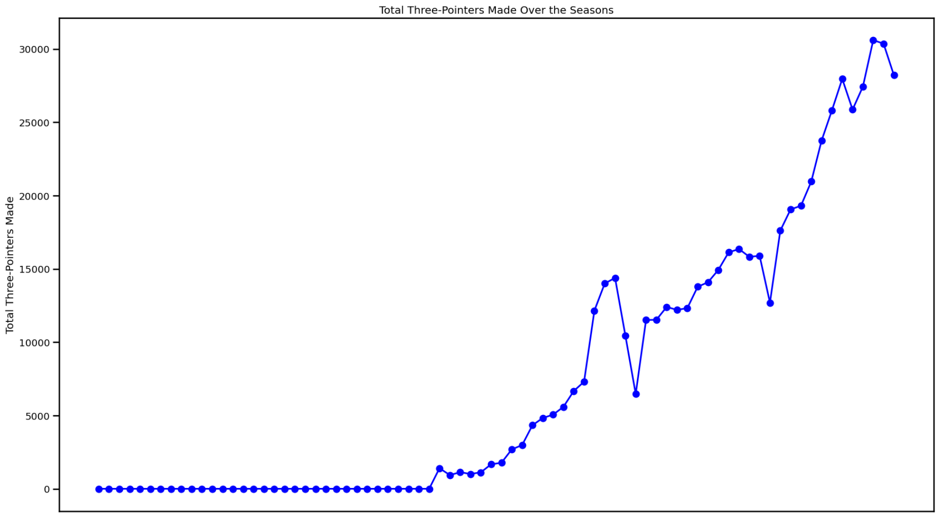

Figure 5.29: Progression of 3PM

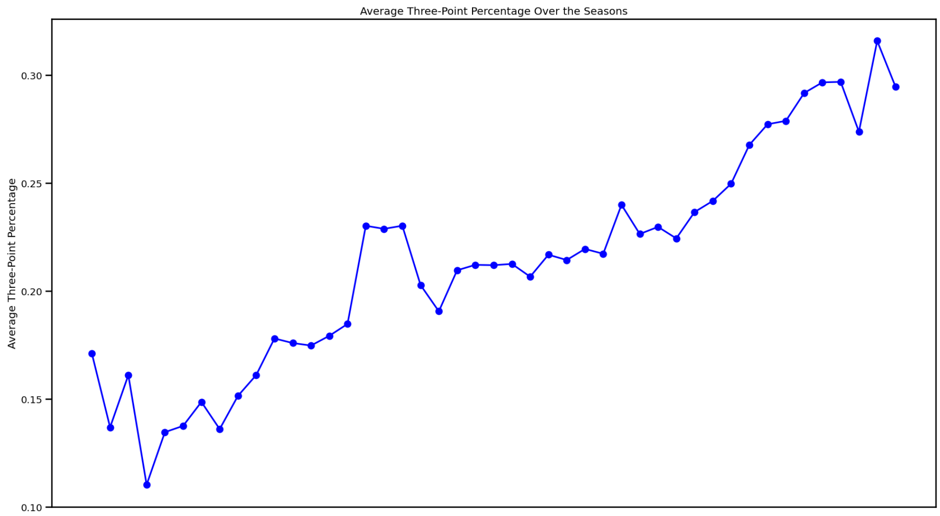

Figure 5.30: Progression of 3P%

As we can clearly see there is a trend of the 3 point shot becoming more utilised as the years have gone by. Players have also started to make more 3 pointers as the volume of shooting has increased, but we can also see that the percentage of 3 point shots being made is also increased from 26% in 2000 to over 30% in 2024. We concluded that the game is evolving to players focusing on their 3 point shot. One piece of analysis we would've loved to look at was if players get paid more if they are good 3 point shooters, but it was beyond the scope of the project.

We then decided to compare the best 3 point shooters across 3 eras: the 2000s, 2010s, and 2020s and compare their 3PM in the graph 5.31 :

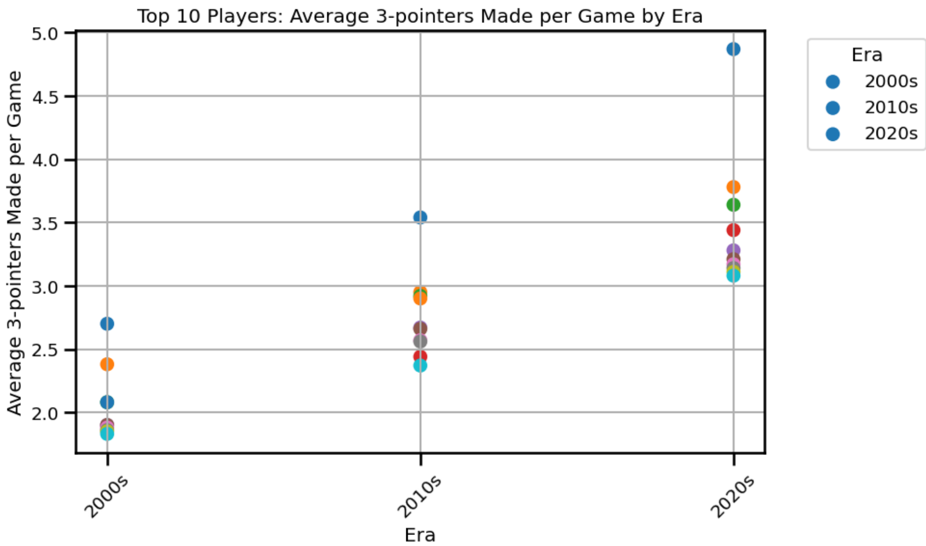

Figure 5.31: 3PM in different eras of the game

Again the graph above backs our conclusion of players depending more on 3PT shots. We then decided to look at what the average PPG of all players were in different seasons in the graph 5.32:

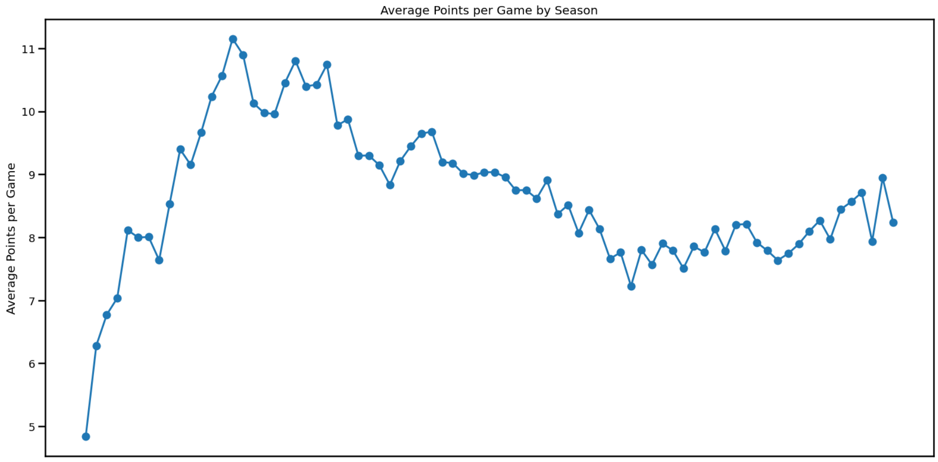

Figure 5.32: Average PPG by Player Over Time

This is quite an unexpected result, actually. We expected the average PPG to increase as time goes on. But it has actually decreased. This could be due to a number of reasons:

1. **Defensive Strategies**: Over time, defensive strategies may have become more sophisticated, leading to tighter defense and lower scoring opportunities for players.
2. **Rule Changes**: Changes in league rules, such as modifications to the shot clock, defensive rules (e.g., hand-checking), or changes in foul regulations, can impact gameplay and scoring.
3. **Evolution of Playing Style**: The style of play in basketball may have shifted over time, with teams emphasizing different aspects of the game, such as outside shooting, post play, or transition offense, which could affect scoring trends.
4. **Athleticism and Conditioning**: Improvements in athleticism and conditioning among players may have led to more competitive defense and increased physicality, making it harder for players to score.
5. **Three-Point Emphasis**: The increasing emphasis on three-point shooting in modern basketball may lead to more attempts from beyond the arc, which generally have a lower shooting percentage compared to two-point field goals, thus potentially lowering overall shooting efficiency and average PPG.
6. **Strategic Fouling**: Teams may employ intentional fouling strategies, such as the "Hack-a-Shaq" tactic, to disrupt opponents' offensive rhythm and lower their scoring output.
7. **Quality of Opposition**: As the league grows and talent becomes more evenly distributed across teams, players may face tougher defensive matchups, resulting in lower individual scoring averages.
8. **Coaching Strategies**: Coaching philosophies and game plans evolve over time, with coaches emphasizing different aspects of the game, such as team defense, ball movement, or balanced scoring, which could impact individual scoring averages.

We then decided to look at the average points being scored per game throughout the years, as shown in the graph 5.33 below:

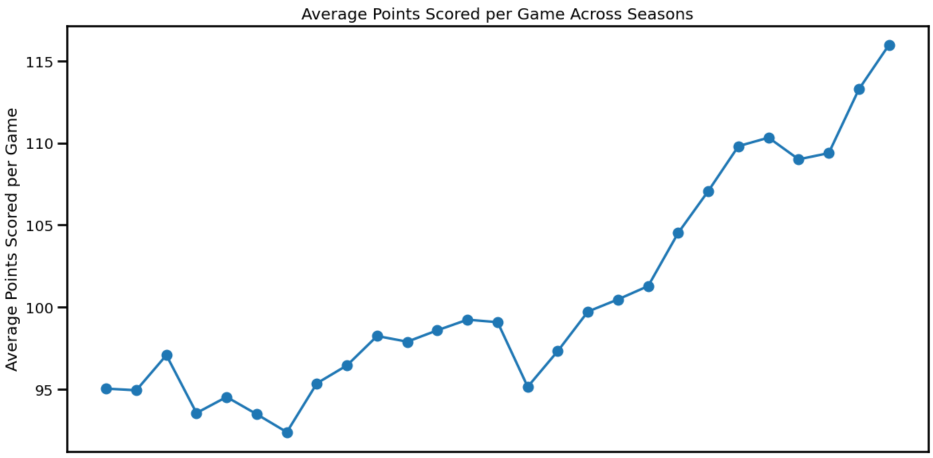

Figure 5.33: Average Team PPG

The meteoric rise in points scored by teams starts in the 2015 season. This coincidentally (note the sarcasm) happens to be the time where the Golden State Warriors introduced something called **'Positionless Basketball'**. What this means basically is that a player isn't bound to the norms of his position. For example, Centers, who are the biggest players on the team, usually around 7 feet tall, are not terrible 3 point shooters anymore. Before, they could not shoot 3 points, for example we saw with Shaq(a 147kg 7'1 Center) when comparing him to Luka that he attempted only one 3 point shot in a season where he was the top scorer. But now, if we compare him to someone like Brook Lopez(128 kg 7'1 Center), Brook's 3pt stats will surely be much better. Back then, the big men in the team had certain roles, such as Rim Protector, Pick-and-Roll Big, etc. But nowadays they're not bound by their position and can also act as players who stretch the floor due to their shooting abilities. This results in their teammates getting more open shots. Let us investigate if my claim on Brook Lopez is true, by getting a table 5.34 of his 3PA, 3PM, 3PT% compared to Shaq's.

	Player Name	3PM	3PA	3P%
0	Brook Lopez	0.0	2.0	0.000
1	Brook Lopez	0.0	2.0	0.000
2	Brook Lopez	0.0	1.0	0.000
3	Brook Lopez	0.0	0.0	0.000
4	Brook Lopez	0.0	1.0	0.000
5	Brook Lopez	0.0	1.0	0.000
6	Brook Lopez	1.0	10.0	0.100
7	Brook Lopez	2.0	14.0	0.143
8	Brook Lopez	134.0	387.0	0.346
9	Brook Lopez	112.0	325.0	0.345
10	Brook Lopez	187.0	512.0	0.365
11	Brook Lopez	102.0	325.0	0.314
12	Brook Lopez	95.0	281.0	0.338
13	Brook Lopez	19.0	53.0	0.358
14	Brook Lopez	136.0	364.0	0.374
15	Brook Lopez	129.0	358.0	0.360
16	Shaquille O'Neal	0.0	2.0	0.000
17	Shaquille O'Neal	0.0	2.0	0.000
18	Shaquille O'Neal	0.0	5.0	0.000
19	Shaquille O'Neal	1.0	2.0	0.500
20	Shaquille O'Neal	0.0	4.0	0.000
21	Shaquille O'Neal	0.0	0.0	0.000
22	Shaquille O'Neal	0.0	1.0	0.000

Figure 5.34: Brook Lopez VS Shaq 3PT Stats

The table 5.34 is the perfect representation of how even the bigger players have started to shoot the ball from 3 point range. There are a few points to note, on how the game has changed because of this fact:

1. **Floor spacing**: Traditionally, centers operated closer to the basket, but with their improved 3-point shooting, they can now stretch the defense by drawing opposing big men out to the perimeter. This creates more space for dribble penetration, cuts, and post-ups for other players on the team.
2. **Defensive adjustments**: Opposing defenses must adapt to the threat of centers shooting 3-pointers. They may need to assign a quicker defender to contest the shot or switch defensive assignments, which can open up opportunities for mismatches elsewhere on the court.
3. **Pick-and-pop effectiveness**: Centers who can shoot from beyond the 3 point line effectively become more dangerous in pick-and-pop situations. If defenders sag off to protect the paint, the center can capitalize by knocking down open shots.
4. **Offensive versatility**: Centers who add 3-point shooting to their repertoire become more versatile offensive threats. They can contribute to scoring in various ways, not just through post-ups or offensive rebounds, making them harder to defend.
5. **Defensive challenges**: Defending centers who can shoot from long range presents challenges for opposing big men. They must decide whether to stay close to contest the shot or protect the rim, potentially leaving shooters open on the perimeter.
6. **Transition offense**: Centers who can shoot 3-pointers effectively can help their team in transition by trailing the play and providing a trailer option for kick-out passes. This adds an extra dimension to the team's fast-break offense.
7. **Strategic adjustments**: Coaches may adjust their offensive systems to take advantage of their center's shooting ability. This might involve incorporating more pick-and-pop plays, spacing the floor differently, or running specific sets to create open looks for the center.
8. **Roster construction**: The ability of centers to shoot from beyond the arc may influence roster construction decisions. Teams may prioritize acquiring or developing big men who can stretch the floor that fit their play style and personnel.
9. **NOTE**: The paint is the silver box you can see around the free throw line in 5.1

Often as an NBA fan, you hear alot of talk about how back in the day, to be considered an All-Star level player, you needed to average at least 20 PPG. However, in today's game, you need to average 20 PPG to be considered an above average player. We decided to test this statement. We retrieved the number of players from the years 1980, 2000, 2020, and 2024 who averaged 20 PPG and over, and the graph **??** below illustrates the results:

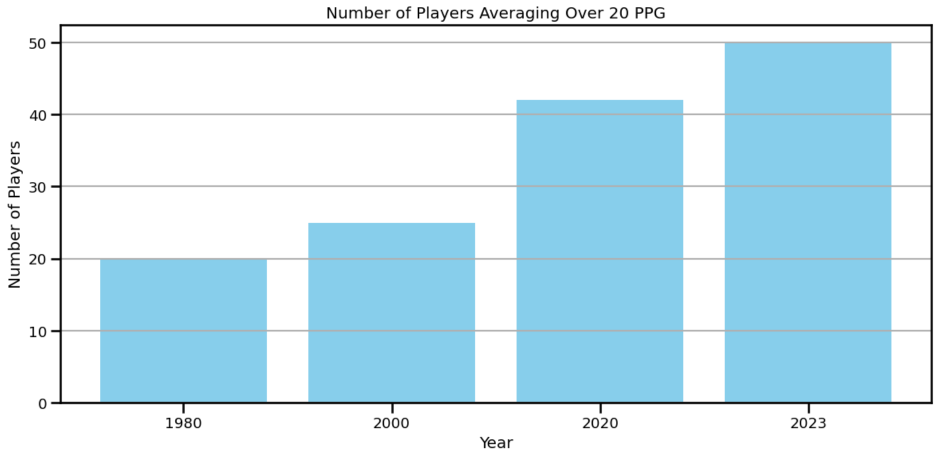

Figure 5.35: Players with >= 20 PPG

As we can see, the scoring standards of the NBA have indeed gotten higher. Players are definitely getting more skilled, **but**, could this be due to the fact that players are getting worse at defence? Let's find out by analysing the trend of average BPG in the graph 5.36 and average SPG in the graph 5.37:

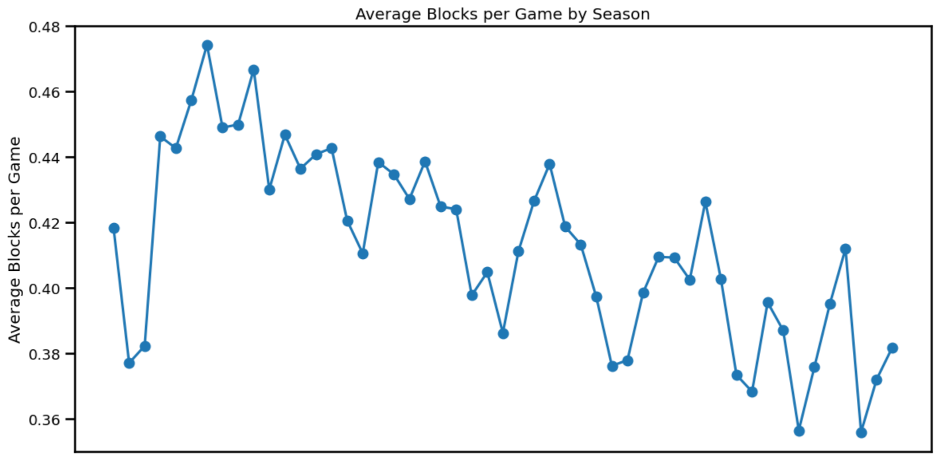

Figure 5.36: BPG Trends over the Seasons

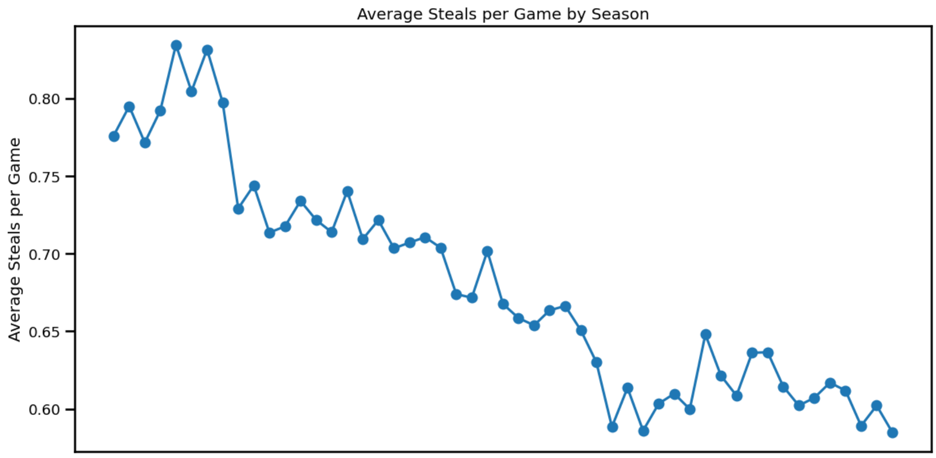

Figure 5.37: SPG Trends over the Seasons

As we can see, the average blocks and steals per game has gone down through the years. But does this really mean that the offense has gotten better? Is it a case of offensive strategies having become more sophisticated, leading to better ball movement, spacing, and scoring opportunities. As a result, defensive players could find it more challenging to make successful blocks and steals. Or is it a case of players becoming more careful due to stricter rules? With advancements in sports science and conditioning, players may prioritize maintaining their physical health and avoiding injuries. As a result, they may focus more on playing disciplined defense to prevent fouls or risky defensive maneuvers that could lead to injuries. Lets check what the STOCKs trend is and see what the correlation is with fouls per game in the following graph 5.38:

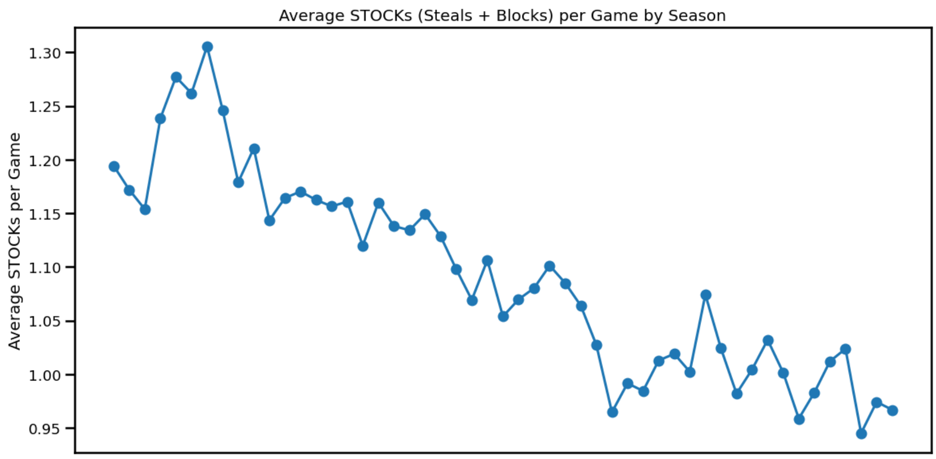

Figure 5.38: STOCKs Trend

We can observe that the STOCKs have actually dramatically decreased over time, indicating that maybe defence has gotten worse. However, we attributed it due to the game becoming "softer", meaning the referees have become more strict with foul calls. We get a correlation coefficient of 0.65 between fouls and STOCKs. This value indicates a moderately positive linear relationship between Personal Fouls and STOCKs. For context, fouls are infractions committed by players that result in an opponent being awarded free throws or the ball being given to the opposing team. It often indicates aggressive defensive play or mistakes in defensive positioning. With a correlation coefficient of 0.65:

1. There is a notable positive relationship between personal fouls and STOCKs per game. This suggests that players who are more active defensively (accumulating more steals and blocks) tend to commit more personal fouls.

2. However, the strength of the relationship is not extremely high, indicating that other factors may also influence personal fouls, and it's not solely dependent on defensive activity.
3. Overall, it implies that there is a tendency for players who are more involved in defensive actions to also incur more personal fouls, but the relationship is not absolute, and other factors likely contribute to the number of personal fouls committed by players.

We also calculated the correlation coefficient between PPG and STOCKs, and got that value to be 0.66. This value suggests a moderate to strong positive relationship between a player's offensive scoring output and their defensive activity. With a correlation coefficient of 0.66:

1. There is a notable positive relationship between a player's offensive scoring output (PPG) and their defensive activity (STOCKs per game). This suggests that players who score more points tend to also be more active on the defensive end, accumulating more steals and blocks.
2. However, while the correlation is moderate to strong, it's not perfect, indicating that other factors may also influence a player's scoring and defensive performance independently.
3. Overall, it implies that there is a tendency for players who excel offensively to also contribute significantly on the defensive end, but the relationship is not absolute, and other factors may also play a role in determining a player's offensive and defensive contributions.

Moving onto the **playmaking** aspect of the game, lets start off by looking at their APG and AST/TO ratios. We started this off with the expectation that playmaking would have stayed on the same level or maybe worse, due to the fact that more and more players were getting good at scoring, meaning their isolation basketball was improving. However, one can also argue that scoring got better because playmaking got better. Players would get better quality shots because their primary ball handlers(playmakers) were creating better looks for them. The graph 5.39 shows the trend of APG over time. We can see that it stays mainly the same after a steep increase. We then analysed how the APG of the top 10 assisters varied in different seasons in another graph 5.40:

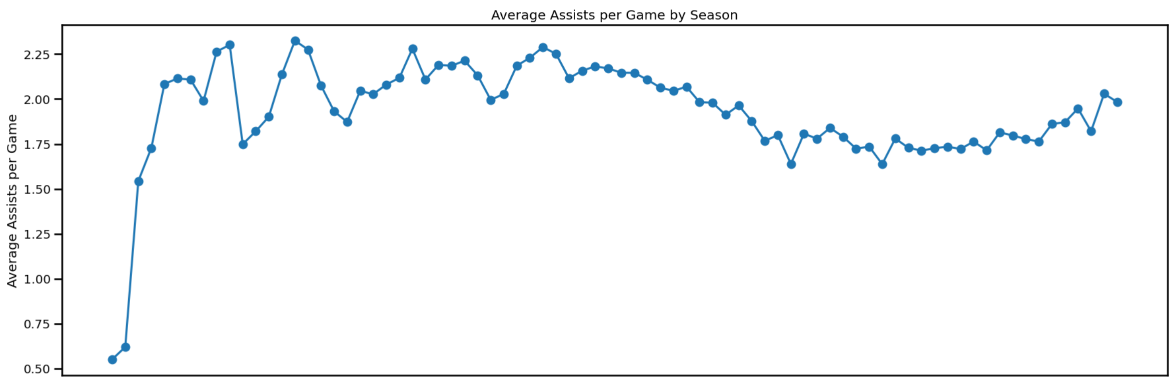

Figure 5.39: APG Trend

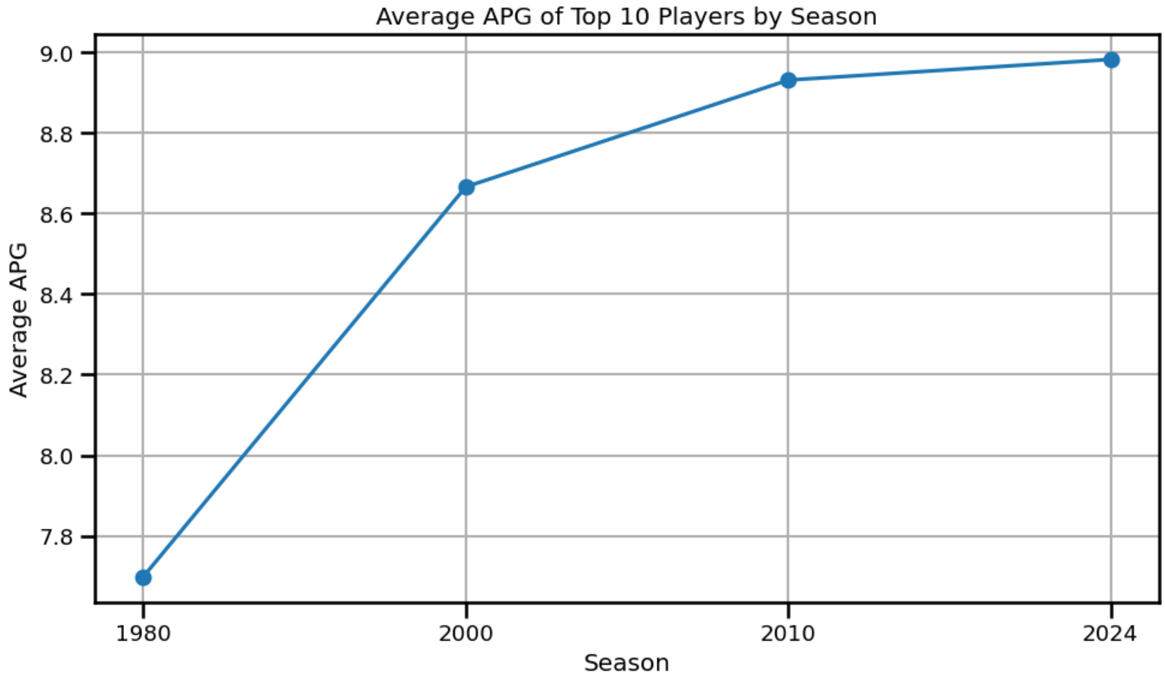

Figure 5.40: Top 10 Assisters' APG trend

From our initial assists stats analysis, we can definitely see that playmakers have also gotten better over time. However its now time to go into some other playmaking stats. We will be looking at the Assist-Turnover ratio in the graph 5.41:

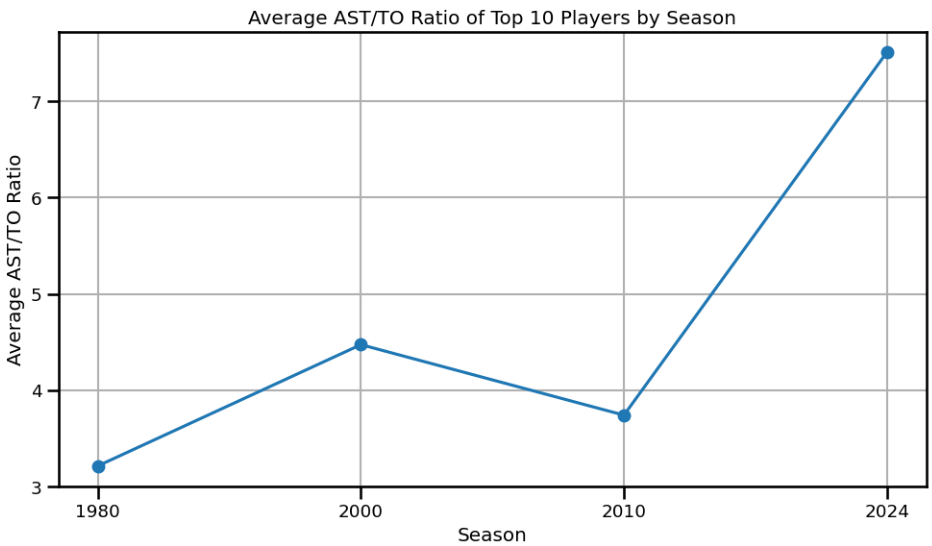

Figure 5.41: AST/TO Trend

We can surely see that the best playmakers over seasons have gotten better over time and have also become safer ball handlers over time. Another interesting thing that we did was to use our **draftData** dataset to analyse players physical attributes before they got drafted in this graph 5.42:

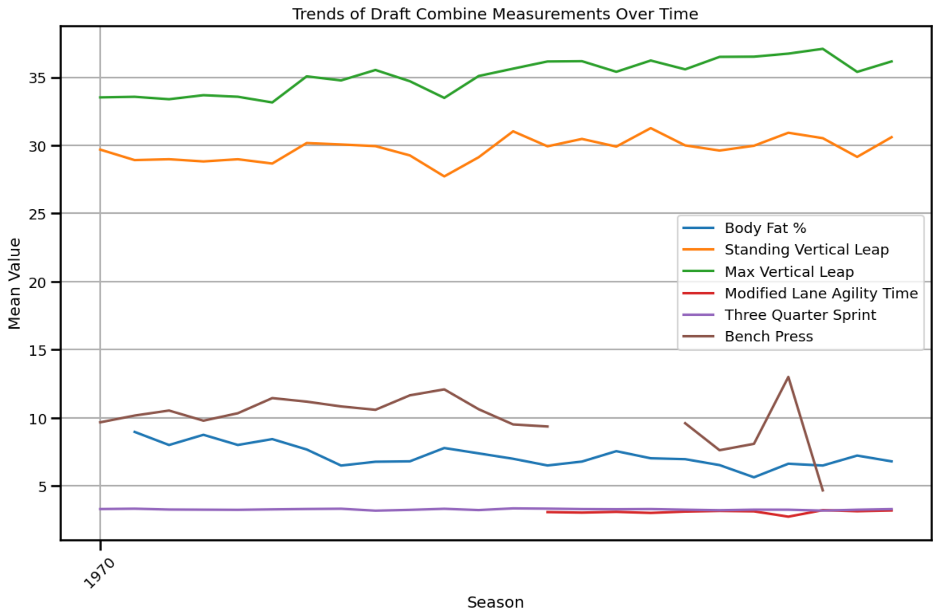

Figure 5.42: Players' Physical Attributes in College

Max Vertical Leap:

1. **Trend**: The trend of the max vertical leap shows fluctuations over time, starting around 34 in 1970 and ending around 36. There are periods of increase and decrease throughout the years, but overall, there is a slight upward trend.
2. **Implications**: The fluctuations in the max vertical leap could be influenced by various factors such as changes in training techniques, advancements in sports science, and improvements in athletic gear. It may also reflect shifts in the demographic of players entering the draft combine, with some years seeing a higher proportion of players with exceptional leaping abilities.
3. **Reasons**: Improvements in strength and conditioning programs, as well as advancements in sports technology, could contribute to the overall increase in max vertical leap over time. Additionally, changes in player recruitment strategies by teams and increased emphasis on athleticism in basketball could also impact the trend.

Standing Vertical Leap:

1. **Trend**: The standing vertical leap shows a fluctuating trend, starting just below 30 in 1970, dropping to 28 at one point, and ending around 31. There are fluctuations throughout the years, but the trend remains relatively stable.
2. **Implications**: The fluctuations in the standing vertical leap could be influenced by similar factors as the max vertical leap, including changes in training methods, advancements in sports science, and variations in the player pool. However, the relatively stable trend suggests that standing vertical leap may not have undergone significant changes over time compared to other measurements.
3. **Reasons**: The stability in the standing vertical leap trend could be due to the nature of the test, which primarily measures lower body explosive power. Changes in player physique and athleticism may have less impact on this measurement compared to others.

Bench Press:

1. **Trend**: The trend of the bench press starts at a mean value of 10 and ends at 5, showing a clear downward trend over time.
2. **Implications**: The decrease in bench press performance over time could be indicative of shifts in training focus and evaluation criteria in basketball. Teams may be placing less emphasis on upper body strength relative to other physical attributes such as speed, agility, and vertical leap.
3. **Reasons**: Changes in basketball strategy and playing style, as well as advancements in sports science, could contribute to the decline in bench press performance. Additionally, changes in player demographics and recruitment strategies may also influence the trend.

Body Fat%:

1. **Trend**: The trend of body fat percentage starts at a mean value of 9 and ends around 6.5-7, showing a consistent decrease over time.
2. **Implications**: The decreasing trend in body fat percentage reflects a potential improvement in overall fitness and conditioning levels among draft prospects. Lower body fat percentage is generally associated with better athletic performance and health.
3. **Reasons**: Increased awareness of the importance of fitness and nutrition, advancements in training methods, and stricter monitoring of player health and fitness by teams and organizations could contribute to the decline in body fat percentage over time.

Three Quarter Sprint and Modified Agility Lane Time:

1. **Trend**: Both the three-quarter sprint and modified agility lane time show relatively stable trends over time, with the measurements staying mostly constant around 3 for the three-quarter sprint and showing minimal variation for the modified agility lane time.
2. **Implications**: The stability in these measurements suggests that the athleticism required for these tests has remained consistent over time. Changes in training methods and player development may have had less impact on these aspects compared to other physical attributes.
3. **Reasons**: The consistent trends in these measurements could be attributed to the standardized nature of the tests and the fundamental skills they assess, such as speed, agility, and quickness. Minimal variation over time may indicate that these attributes have remained relatively constant in basketball players entering the draft combine.

We can conclude from this that physically, players have gotten a little bit better but the advancements in technology and facilities for training and rehabilitation have also been a big factor in contributing to this. The final aspect of the question was the Steve Nash Case Study. Now, one may be wondering why we chose Steve Nash. It's because we came across an interview [4] of his where he says he should've "shot more threes". This gave us the impression that he felt like he was too conservative in his approach to the game and he could've matched the players of today. What we did was we took his 2004/2005 season where he was the Most Valuable Player (MVP) and compared him to players who played in the same season.

We found that Steve Nash's PPG in his MVP season was 15.53, while the top 5 scorers that season had an average of 27.8 PPG. Nash averaged 12 less PPG than the best scorers that year, which can indicate that scoring was not given alot of importance and perhaps efficiency was rated higher. However, Nash shot 50.2% from the field while the top 5 scorers shot 45% from the field that year. Nash was much more efficient. One stat Nash really has going for him is his 3PT%. He shot

Stat	Steve Nash	Top 5 Scorers/Assisters
PPG	15.53	27.8
FG%	50.2	45
3PT%	43.1	36.44
APG	11.48	8.4

Table 5.5: Comparison of the 2005 MVP and top 5 scorers and assisters that season

43.1% on 3 point shots that season while the top 5 scorers shot 36.44% that season.The table below outlines the differences between Steve Nash and the top scorers and assisters that season:

What we did next was we compared Steve Nash with some of the MVPs that came after him. We decided to pick LeBron James' 2009/10 season, Kevin Durant's 2013/14 season, Russell Westbrook's 2016/17 season, and finally Nikola Jokic's 2021/22 season. We compared their statistics in their MVP seasons in the following graphs 5.43,5.44, 5.45, 5.46, 5.47, 5.48:

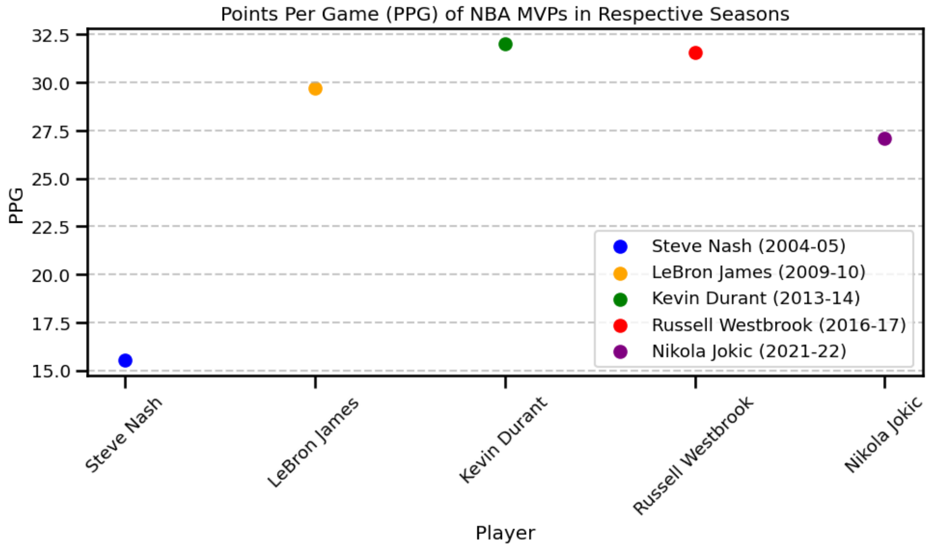

Figure 5.43: MVP's PPG

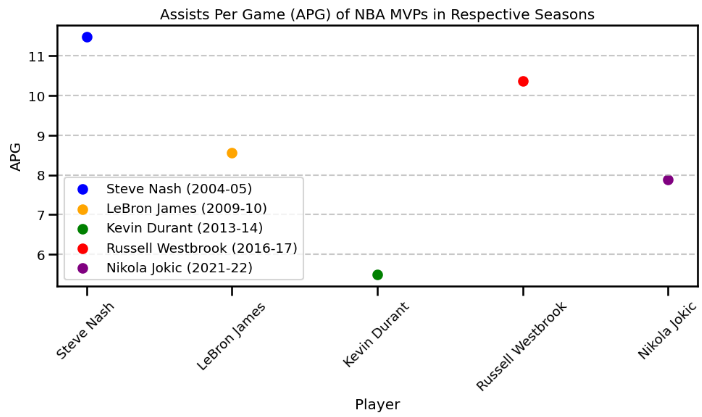

Figure 5.44: MVP's APG

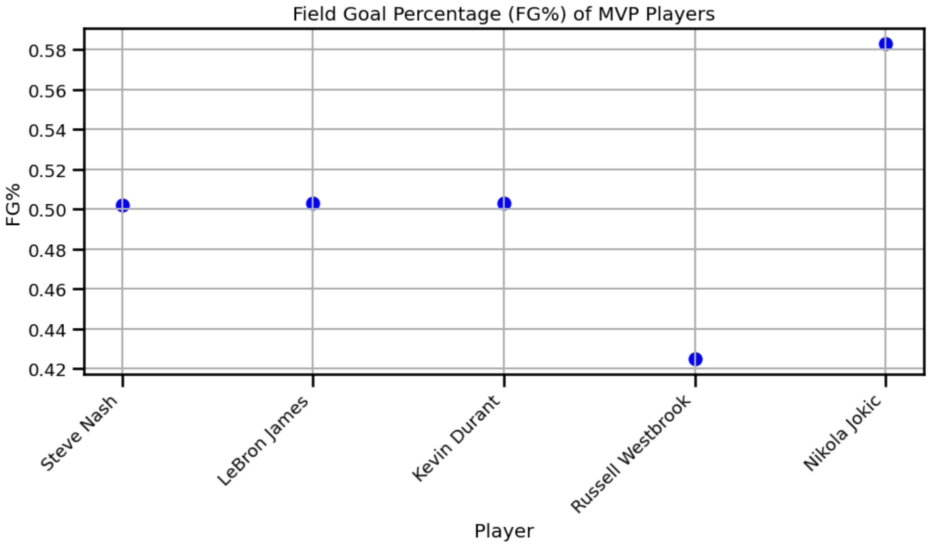

Figure 5.45: MVP's FG

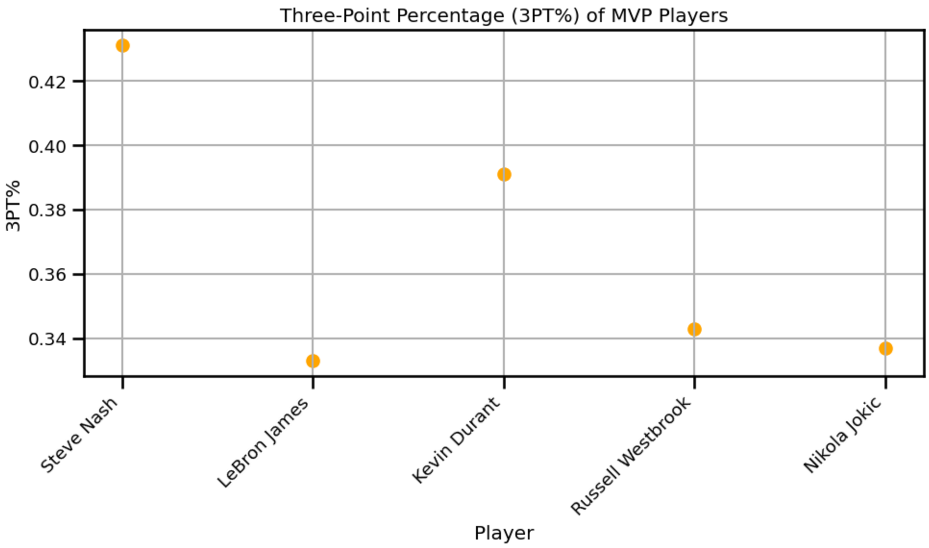

Figure 5.46: MVP's 3PT%

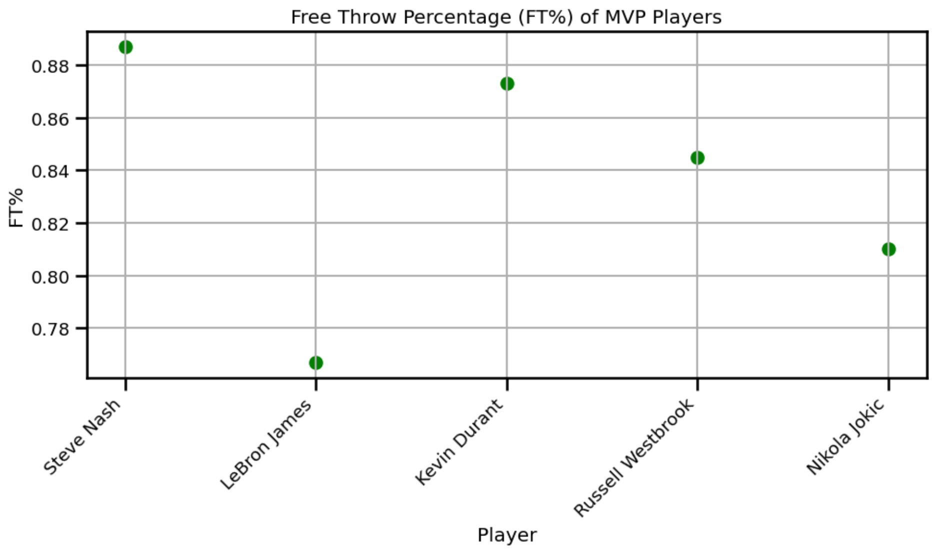

Figure 5.47: MVP's FT%

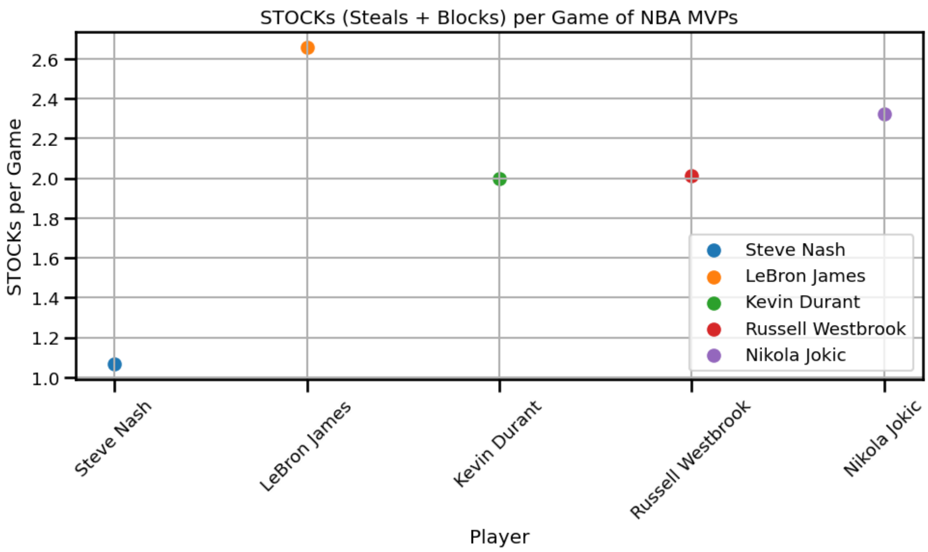

Figure 5.48: MVP's STOCKs

Comparison between MVPs' Scoring: We can see clearly that the other MVPs had much higher PPG than Steve Nash. It can highlight the fact that players needed to score more points to win games and as a result win MVP in recent years. Scoring is definitely more looked at when taking into account who the MVP should be.

1. **FG%**: their FG%s vary, with LeBron James and Kevin Durant having slightly better shooting from the field but averaging much more PPG. Russell Westbrook is considerably less efficient than Steve Nash, but we must take into account that he scored much more points. His lack in efficiency could be due to the fact that he was the team's main scorer and playmaker. Him taking more shots may mean he scores more points, but could also mean that he takes more shots to reach those points. Nikola Jokic almost shoots 60% from the field which is really crazy compared to Nash as he also has more points.

2. **3PT%**: Steve Nash was the most efficient 3PT shooter out of the 5 MVPs. No one really comes close except Kevin Durant, who comes within 4% of Nash. This could be due to the fact that Nash attempted less 3PT shots per game than the other MVPs, due to the fact

that 3PT shots were attempted way more in the modern game, and Nash himself said that he should've shot more.

3. **FT%**: Nash was also the best FT shooter out of the players. This could be due to the same reasons as the above, for 3PT shooting.

Comparison between MVPs' Playmaking: Again we can see that Nash had the highest assists per game. From our current analysis of Steve Nash against the other MVPs, we can say that they were much better scorers than Nash due to their extreme increase in PPG, but Nash is a better playmaker than all of them.

Comparison between MVPs' Defence: We can observe that Nash was a considerably worse defender than all the other MVPs, this can indicate that overall the defensive level of a player was not considered when voting for an MVP.

Finally, for this question, we compared Nash the MVP to players in general who came after that season in the following graphs 5.49,5.50, 5.51, 5.52, 5.53, 5.54 :

```
Number of players with higher PPG than Steve Nash's 2005 season:
2010: 67
2015: 62
2020: 82
2024: 92
```

Figure 5.49: Comparison of PPG

```
Number of players with higher APG than Steve Nash's 2005 season:
2010: 0
2015: 0
2020: 0
2024: 0
```

Figure 5.50: Comparison of APG

```
Number of players with higher AST/TO ratio than Steve Nash's 2005 season (minimum 5 APG):
2010: 3
2015: 3
2020: 1
2024: 12
```

Figure 5.51: Comparison of AST/TO Ratio

```
Number of players with higher FG% and higher PPG than Steve Nash's 2005 season:
2010: 16
2015: 7
2020: 15
2024: 26
```

Figure 5.52: Comparison of FG% and PPG

```
Number of players with higher 3PT% and higher PPG than Steve Nash's 2005 season:
2010: 1
2015: 6
2020: 1
2024: 1
```

Figure 5.53: Comparison of 3P% and PPG

```
Number of players with higher STOCKs than Steve Nash's 2005 season:
2010: 204
2015: 215
2020: 239
2024: 249
```

Figure 5.54: Comparison of STOCKs

From the analysis, it is clear that on the scoring side, there were 67 players in 2010 who has more PPG than Steve Nash, and that number increased to 92 in 2024. This clearly indicates that players are getting better at scoring the basketball. If we go deeper into the scoring aspect, of those players with higher PPG than Steve Nash, 16 players had a higher Field Goal Percentage than Nash, while that number increased to 26 in 2024. However, only 1 player in 2010 had a higher 3PT% than Nash in 2010, 6 in 2015, and 1 each in 2020 and 2024.

On the playmaking aspect of the game, no players had more APG than Steve Nash in either of the years. However, we can perhaps say that players became more safer handlers of the ball. The number of players who had a higher AST/TO ratio than Nash(min. 5 APG) was 3 in 2010 and increased to 12 in 2024.

On the defensive side of the game, there were 204 players in 2010 who had more STOCKs than Steve Nash, while in 2024 there were 249 players who had more STOCKs than Nash. This shows us that many players in the current day are better defenders than an MVP in 2005.

5.3.3 Discussion

From our analysis, it's evident that the NBA has experienced an overall improvement in its level of play over the years, resulting in a more captivating and entertaining spectacle for fans. The rise in scoring prowess among players indicates a significant evolution in offensive strategies, with teams increasingly prioritizing high-scoring performances over defensive prowess. This shift has led to a game where players are honing their offensive skills to unprecedented levels, making it increasingly challenging for defenses to contain them.

However, this offensive surge has also coincided with a decline in defensive efficiency, as indicated by the decrease in steals, blocks, and overall defensive impact over time. While some argue that this decline may be attributed to the increased offensive proficiency of players, others suggest that rule changes and a shift towards a more lenient approach to defense may also be contributing factors.

Despite these defensive challenges, our analysis highlights a positive trend in playmaking abilities, with players demonstrating improved passing skills and decision-making on the court. This is evident in the increasing assist-to-turnover ratios over the years, indicating a greater emphasis on ball security and strategic playmaking.

In conclusion, while the NBA has undoubtedly seen a rise in offensive firepower and playmaking prowess, the decline in defensive efficiency raises questions about the balance between offense and defense in modern basketball. Moving forward, it will be crucial for teams and players to strike a balance between offensive aggression and defensive discipline to maintain the integrity and excitement of the game.

5.4 RQ4 : Which performance stats contribute to determining a team's success the most?

This question aims to analyse various features (based on players, teams and individual games) that potentially impact a team's performance and identify the top ones which teams should focus on the most to maximise their efficiency during games.

5.4.1 Data & Method

Here are the datasets we used to answer this research question :

- **Player Dataset** : We extracted individual player statistics such as points scored, assists, rebounds, steals, and blocks from the player dataset. These metrics helped in understanding the individual contributions of players to the team's overall performance.
- **Team Dataset** : The team dataset provided aggregated statistics for each team across multiple seasons. Metrics such as win percentage, field goal percentage, three-point percentage, points scored etc were extracted from the team dataset. These metrics served as indicators of team efficiency and effectiveness in various aspects of the game. We also then used some of these metrics to build a linear regression model to predict a team's success.
- **Game Dataset** : The game dataset contained detailed information about each game, including team matchups, game outcomes (wins/losses), and individual game statistics. Analyzing game-level data allowed for a deeper understanding of how teams perform against different opponents in various game situations and under different conditions.

5.4.2 Results

It is quite obvious that for a team to perform well, individual player stats are absolutely crucial such as '3pt%', 'field goal%', 'blocks' etc. However, we thought it would be interesting to see if teams are relying on 1 or 2 players to do well or are all players in the team performing equally well, leading to the team's success.

To do this looked at the top 3 teams with the highest points per game (PPG) in the last 5 years and saw how players contributed to their respective teams :

	player_name	season	team_name	team_ppg	player_ppg	player_contribution
0	Stephen Curry	2018-19	Golden State Warriors	117.682927	27.260870	0.231647
1	Kevin Durant	2018-19	Golden State Warriors	117.682927	25.987179	0.220824
2	Klay Thompson	2018-19	Golden State Warriors	117.682927	21.538462	0.183021
3	DeMarcus Cousins	2018-19	Golden State Warriors	117.682927	16.266667	0.138225
4	Draymond Green	2018-19	Golden State Warriors	117.682927	7.363636	0.062572
5	Quinn Cook	2018-19	Golden State Warriors	117.682927	6.878378	0.058448
6	Jonas Jerebko	2018-19	Golden State Warriors	117.682927	6.287671	0.053429
7	Kevon Looney	2018-19	Golden State Warriors	117.682927	6.250000	0.053109
8	Andre Iguodala	2018-19	Golden State Warriors	117.682927	5.720588	0.048610
9	Damian Jones	2018-19	Golden State Warriors	117.682927	5.416667	0.046028
10	Damion Lee	2018-19	Golden State Warriors	117.682927	4.937500	0.041956
11	Alfonzo McKinnie	2018-19	Golden State Warriors	117.682927	4.680556	0.039773
12	Marcus Derrickson	2018-19	Golden State Warriors	117.682927	4.181818	0.035535
13	Shaun Livingston	2018-19	Golden State Warriors	117.682927	4.031250	0.034255
14	Andrew Bogut	2018-19	Golden State Warriors	117.682927	3.545455	0.030127

Above is a snippet from the table (all rows not included) on how each player in each team performed (in terms of contributing to team points) in the past 5 seasons.

We then observed that only 19 players had a 'player_contribution' of over 20% so the question arised : is it really the case that teams only require a few players to perform well for the team?

We visualised this by looking at how many players on average make up for 50% of the total points scored by their respective team : Fig 5.55

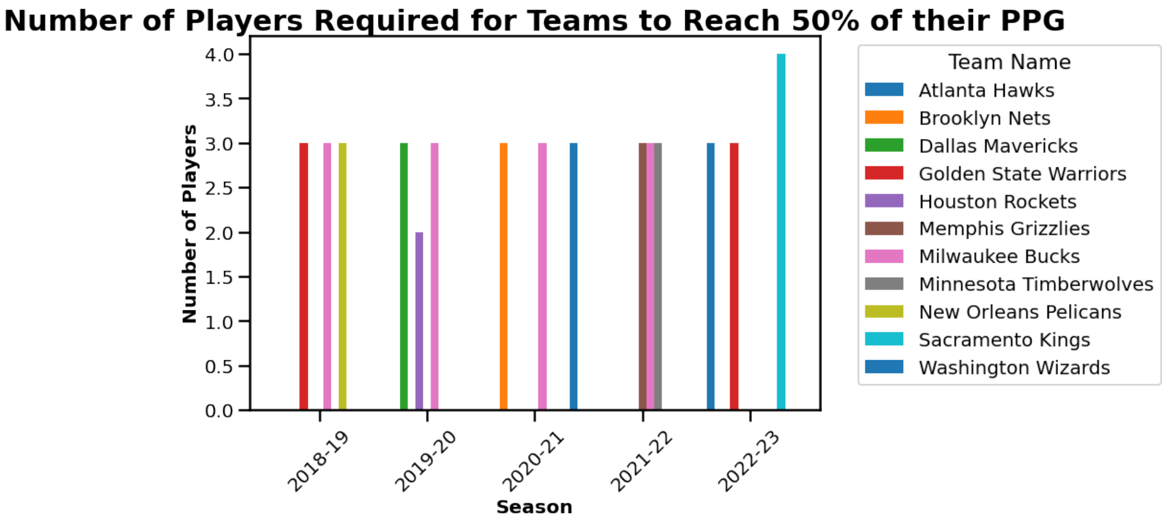

Figure 5.55: Number of Players needed to make up 50% of total points

As we can see from Fig 5.55, in most cases only 3 players (from the entire roster of around 15-18 players) make up for 50% of a team's points (PPG) throughout the season. This suggests that teams do heavily rely on a few players to perform well and carry them to victory.

We then wanted to see how individual team stats impact a team's performance. To do this we first observed the correlation between different factors such as 'field_goal_pct','three_point_pct' etc and the 'conference_rank' (chosen metric for "team's success") of teams : Fig 5.56

Note : NBA teams are divided into 2 conferences Eastern and Western with 15 teams in each conference. Therefore, Conference rank is the rank of a team in their respective conference.

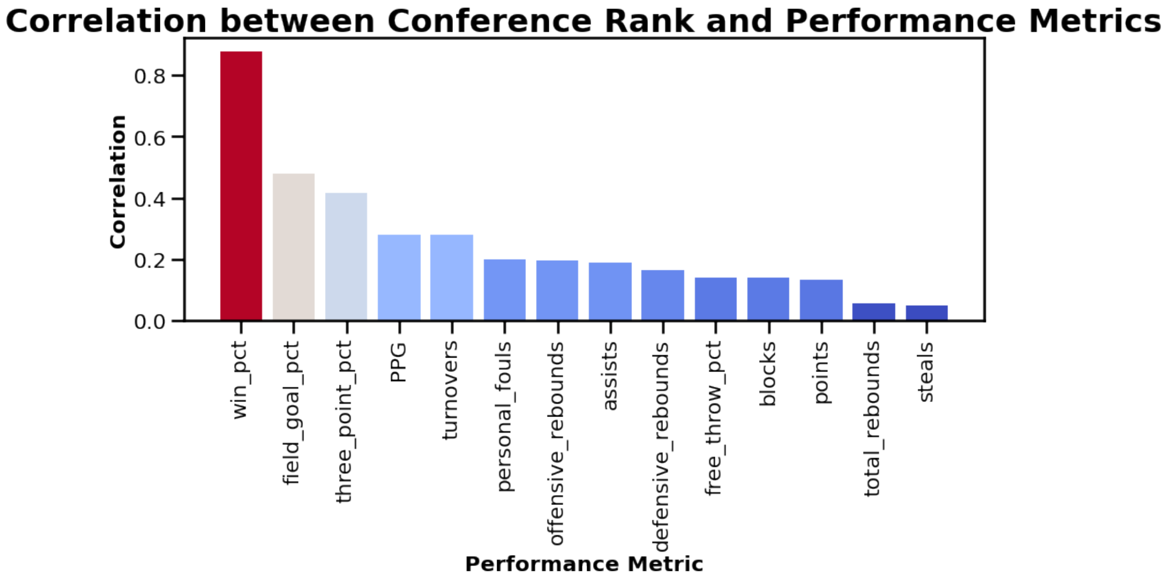

Figure 5.56: Correlation Chart

As expected ‘win_pct‘, ‘field_goal_pct‘ and ‘three_point_pct‘ seem to have the strongest correlation with ‘conference_rank‘. It is surprising to see features like ‘PPG‘ and ‘points‘ have such a low correlation as one would expect them to have the most impact on a team’s success (‘conference_rank‘ in this case).

We took the top 5 correlated features and built a Linear Regression model to see if we can somewhat accurately predict ‘conference_rank‘ based on them : Fig 5.57

```
X = team_stats[['win_pct', 'field_goal_pct', 'three_point_pct', 'PPG', 'turnovers']]
y = team_stats['conference_rank']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
r_squared = r2_score(y_test, y_pred)

print("Mean Absolute Error:", mae)
print("Mean Squared Error:", mse)
print("R-squared:", r_squared)

Mean Absolute Error: 1.344382439772509
Mean Squared Error: 2.937034725051422
R-squared: 0.8418432518256858
```

Figure 5.57: Linear Regression Model

Lets discuss the results of our model :

- **Mean Absolute Error (MAE)** : MAE measures the average absolute difference between the predicted conference ranks and the actual conference ranks. In this case, an MAE of approximately 1.34 indicates that, on average, the model’s predictions are off by around 1.34 positions in conference rank.

- **Mean Squared Error (MSE)** : MSE measures the average squared difference between the predicted conference ranks and the actual conference ranks. In this case, the MSE of approximately 2.94 indicates that, on average, the squared difference between predicted and actual conference ranks is around 2.94
- **R-squared (R^2)**: R-squared is a measure of how well the regression model explains the variance in the target variable. In this case, the R-squared value of approximately 0.84 indicates that around 84% of the variance in conference ranks is explained by the independent variables (selected metrics) included in the model.

Overall, the evaluation metrics suggest that the linear regression model performs reasonably well in predicting conference ranks based on the selected features, indicating these performance metrics have a huge impact on a team's success. However, there are still a few limitations to consider :

- **Assumptions of Linear Regression** : Linear regression assumes a linear relationship between the features and the target variable. While individual player stats like points per game (PPG) or assists may have a linear relationship with team performance to some extent, there are hidden factors to be considered. For e.g : adding a high-scoring player to a team with an already strong scoring might not have a proportional impact on conference rank
- **Correlation vs. Causation** : While the model identifies correlations between different metrics and conference ranks, it does not establish causation. Other unobserved variables or confounding factors may influence both the features and the target variable.
- **Limited Features** : Sports outcomes can be quite unpredictable and influenced by many factors, some of which may not be easily quantifiable (like player morale, injuries, etc.). So while such a model can provide useful insights, its predictions should not be taken as absolute.

Finally we looked at how individual game factors such as 'performance against specific opponents' have an impact on a team's success.

We split the data into two halves (first half containing data from years '2000-2011' and second from years '2011-2023') to see how teams have progressed over the two decades against the same opponents : Fig 5.58 , Fig 5.59

```
team_performance_2000_to_2011
✓ 0.0s
```

	team_name	opponent_name	games_played	wins	losses	avg_points_scored	avg_field_goal_pct	win_pct	loss_pct
0	Atlanta Hawks	Boston Celtics	13	5	8	95.846154	0.451538	0.384615	0.615385
1	Atlanta Hawks	Brooklyn Nets	15	3	12	94.000000	0.441067	0.200000	0.800000
2	Atlanta Hawks	Charlotte Hornets	19	14	5	95.473684	0.466789	0.736842	0.263158
3	Atlanta Hawks	Chicago Bulls	11	6	5	98.090909	0.433364	0.545455	0.454545
4	Atlanta Hawks	Cleveland Cavaliers	19	8	11	92.578947	0.452368	0.421053	0.578947
...	...	...	...	...	...	...	...	...	...
868	Washington Wizards	Portland Trail Blazers	9	4	5	88.444444	0.434000	0.444444	0.555556
869	Washington Wizards	Sacaramento Kings	6	2	4	97.166667	0.431333	0.333333	0.666667
870	Washington Wizards	San Antonio Spurs	11	4	7	91.181818	0.434818	0.363636	0.636364
871	Washington Wizards	Toronto Raptors	18	10	8	96.333333	0.444667	0.555556	0.444444
872	Washington Wizards	Utah Jazz	4	3	1	95.750000	0.407250	0.750000	0.250000

873 rows × 9 columns

Figure 5.58: 2000-2011

```
team_performance_2011_to_2023
```
✓ 0.0s

	team_name	opponent_name	games_played	wins	losses	avg_points_scored	avg_field_goal_pct	win_pct	loss_pct
0	Atlanta Hawks	Boston Celtics	12	5	7	102.666667	0.458750	0.416667	0.583333
1	Atlanta Hawks	Brooklyn Nets	16	5	11	108.750000	0.446688	0.312500	0.687500
2	Atlanta Hawks	Charlotte Hornets	25	15	10	105.240000	0.458600	0.600000	0.400000
3	Atlanta Hawks	Chicago Bulls	16	6	10	100.687500	0.437500	0.375000	0.625000
4	Atlanta Hawks	Cleveland Cavaliers	25	14	11	109.320000	0.467040	0.560000	0.440000
...	...	...	...	...	...	...	...	...	...
865	Washington Wizards	Portland Trail Blazers	5	3	2	105.800000	0.447200	0.600000	0.400000
866	Washington Wizards	Sacramento Kings	9	5	4	111.222222	0.481889	0.555556	0.444444
867	Washington Wizards	San Antonio Spurs	10	2	8	101.700000	0.455400	0.200000	0.800000
868	Washington Wizards	Toronto Raptors	18	3	15	98.888889	0.430944	0.166667	0.833333
869	Washington Wizards	Utah Jazz	7	3	4	96.857143	0.444286	0.428571	0.571429

870 rows × 9 columns

Figure 5.59: 2011-2023

We then plotted the **win%** of a few teams against each opponent in both time frames : Fig 5.60 and Fig 5.61

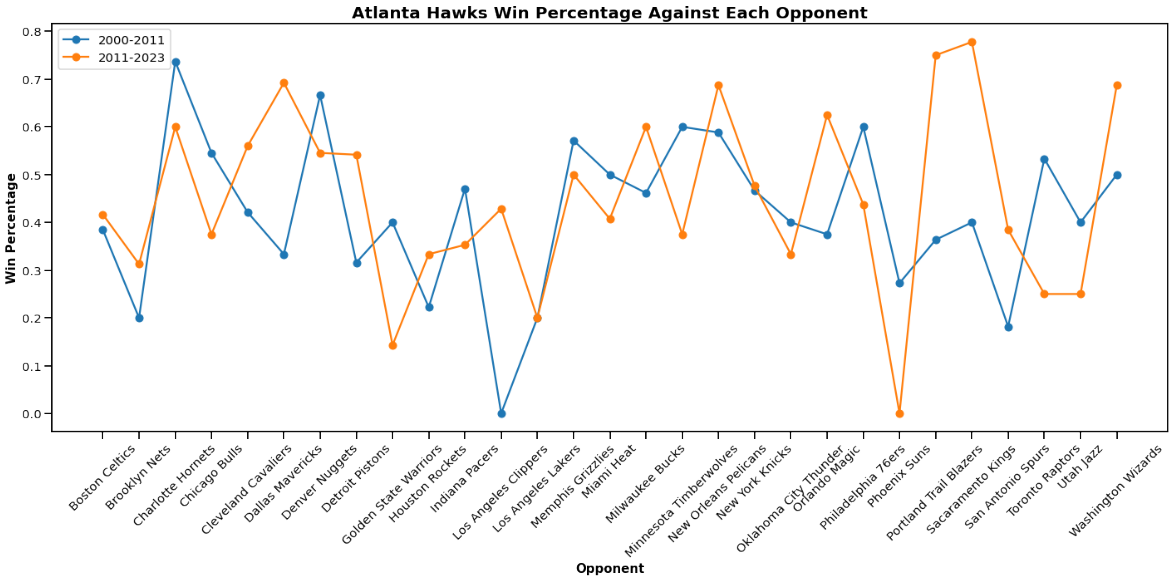

Figure 5.60: Atlanta Hawks

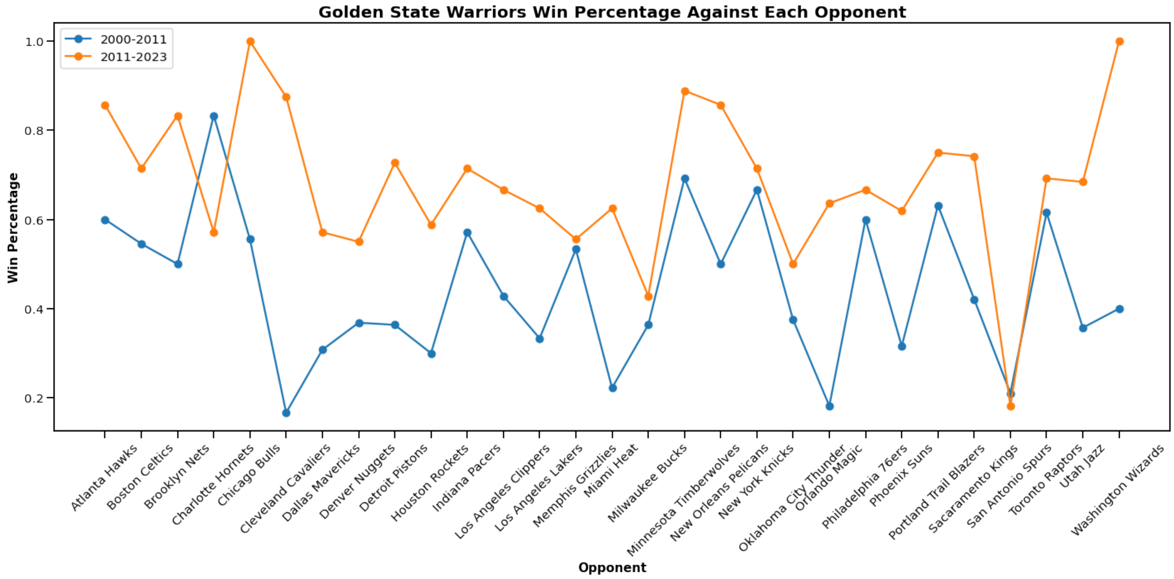

Figure 5.61: Golden State Warriors

In the graphs 5.60 and 5.61, the blue line represents the win% of the team against each opponent from ‘2000-2011‘ and the orange line represents the same information from ‘2011‘ onwards. From this we can see if the team has improved against their opponents over the years (if ‘orange point‘ lies above ‘blue point‘).

The ‘Atlanta Hawks‘ seem to have improved against some of their opponents, especially ‘Indiana Pacers‘ (win% went from 0 to 40%) and ‘Portland Trail Blazers‘ (40% to 80%) which suggests they certainly have upped their game.

On the other side, by the looks of it ‘Golden State warriors‘ have improved against almost all of their opponents which goes to show how dominant they have been in the past few years (they have won the championship 3 times in the last 5 years)

Therefore, it appears that the way teams perform against specific opponents can indeed be an important factor affecting a team’s overall performance. The examples provided above with the Atlanta Hawks and the Golden State Warriors illustrate this point well. Analyzing performance against specific opponents allows teams to identify strengths and weaknesses more accurately, leading to targeted improvements in gameplay, strategies, and player development. It emphasizes the importance of understanding matchups, exploiting opponent weaknesses, and adapting strategies based on the strengths and weaknesses of specific teams.

Lastly, before concluding our analysis, we saw if teams have improved overall by calculating the average of the difference in win_pct against each opponent (win_pct in ‘"2011-2023"‘ minus win_pct in ‘"2000-2010"‘) which will give us the average improvement in win_pct for each team : Fig 5.62

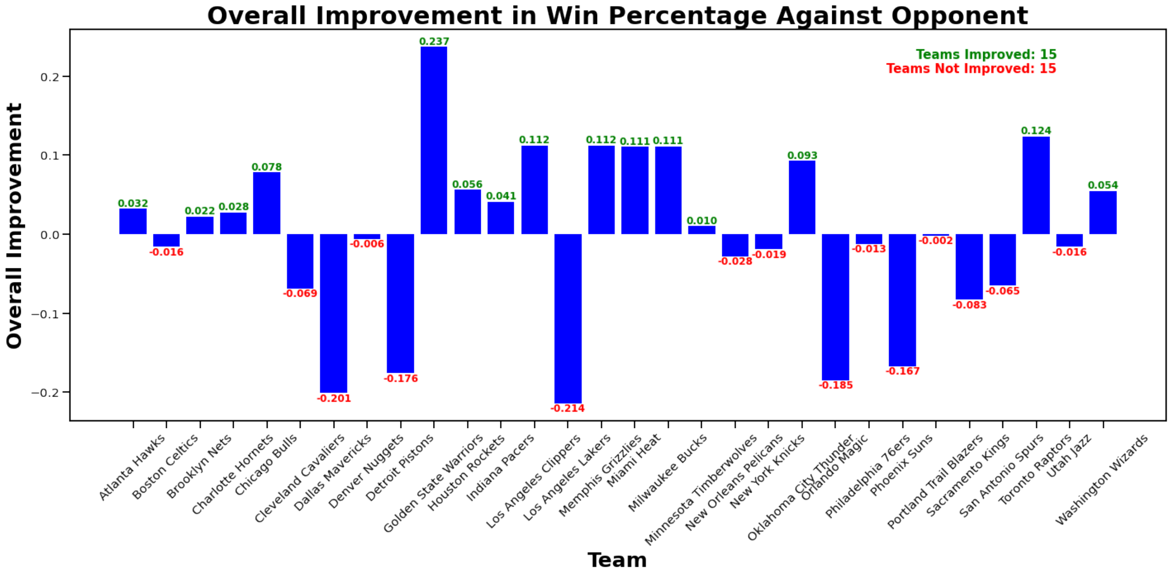

Figure 5.62: Overall Improvement

5.4.3 Discussion

From our analysis, we can conclude Team success in sports is intricately tied to star player performance, team metrics, and performance against specific opponents. Star players often elevate team performance with their exceptional skills and leadership. Monitoring key metrics like win percentage and points scored provides insights into overall team effectiveness and areas for improvement. Additionally, analyzing performance against specific opponents reveals a team's adaptability and strategic expertise, offering opportunities to capitalize on strengths and exploit weaknesses. By excelling in these areas, teams can optimize their chances for success and achieve their competitive goals.

However, while these factors certainly do influence a team's success, there are additional elements that may impact outcomes but couldn't be explored as part of our analysis as they are beyond the scope of this project. For e.g. : Factors such as team chemistry, coaching strategies, player injuries, and off-court dynamics can significantly influence a team's performance and success but may not be fully captured by traditional metrics or analysis.

Chapter 6: **Conclusions**

In summarizing our findings, it's evident that our exploration of NBA basketball has provided multifaceted insights into the game's evolution, player evaluation strategies, and team dynamics. From identifying the greatest players in NBA history to dissecting draft and trade strategies, our analysis has shed light on various aspects of the sport. Additionally, we've examined the overall trajectory of the NBA's level of play, delving into offensive and defensive trends over the years, and scrutinizing the factors contributing to team success.

Building on these conclusions, it's clear that the NBA's landscape is continually evolving, presenting both challenges and opportunities for players, teams, and analysts alike. The rise of advanced analytics and technological advancements offers unprecedented access to data and insights, revolutionizing how we understand and analyze the game. As such, future research should explore emerging trends, delve deeper into niche areas, and incorporate innovative methodologies to stay at the forefront of basketball analysis.

Furthermore, it's crucial to recognize the ethical implications of our work, particularly concerning player evaluation, team management, and fan engagement. While data-driven approaches enhance decision-making and performance optimization, they also raise concerns about privacy, fairness, and the human element of the game. As analysts, we must navigate these ethical considerations responsibly, ensuring that our methodologies and conclusions uphold integrity and respect for all stakeholders involved.

Looking ahead, there are several avenues for future research and analysis in NBA basketball. These include:

1. Investigating the impact of coaching strategies on player development and team performance.
2. Exploring the intersection of sports psychology and performance analytics to understand the mental aspects of the game.
3. Analyzing the influence of external factors such as fan engagement, media coverage, and societal trends on player and team dynamics.
4. Examining the effectiveness of different playstyles and game tactics in achieving strategic objectives on the court.
5. Delving into the global expansion of basketball and its implications for talent development, scouting, and league competitiveness.
6. By embracing these opportunities and addressing ethical considerations, we can advance our understanding of NBA basketball while promoting the integrity, inclusivity, and enjoyment of the sport for fans worldwide.

By embracing these opportunities and addressing ethical considerations, we can advance our understanding of NBA basketball while promoting the integrity, inclusivity, and enjoyment of the sport for fans worldwide.

Acknowledgements

We would like to express our sincere gratitude to our lecturer, Barry for his invaluable guidance, support, and mentorship throughout this project. His expertise in the field of Data Science provided the foundation for our research and enabled us to explore complex concepts with clarity and depth. We are truly grateful for his insightful feedback, constructive criticism, and unwavering encouragement, which have significantly enriched our learning experience and contributed to the success of this project.

We would also like to thank our demonstrator, Rian for his tireless assistance, patience, and support during the course of this project. We are grateful for his willingness to share his expertise, provide insightful explanations, and offer valuable advice whenever needed.

Bibliography

1. Sington, A. *Analyzing NBA Data in Python | NBA Data Analytics Project* (Youtube, 2023).
2. Skompinski, K. *NBA Sports Betting Using Machine Learning* (Github, 2023).
3. Jenness, C. *NBA Analytics* (Github, 2018).
4. Simmons, B. *Steve Nash Regrets Not Shooting More 3s | Bill Simmons's Book of Basketball 2.0 | The Ringer* (Youtube, 2018).

Appendices

Appendix A – Individual Contributions

Muhammad Khan : Research Questions 1 and 3
Sanat Dusad : Research Questions 2 and 4

Appendix B – Datasets & Notebooks

Please list the filenames of the main datasets used in your project along with a brief (one-sentence) description of each. Also list the notebooks with a similarly brief summary of their purpose. You can indicate too who was responsible for which dataset/notebook by indicating the initials of the student beside the dataset or notebook file.

Datasets Used

1. **draft_data.csv** – the main dataset for RQ2 which includes data on past drafts in the NBA.
2. **player_stats.csv** - dataset containing a wide array of statistics on players in the NBA used for all research questions
3. **team_game_stats.csv** - The game dataset contains detailed information about each game, including team matchups, game outcomes (wins/losses), and individual game statistics. Used for RQ2 and RQ4.
4. **team_regSeason_stats.csv** - contains detailed information about teams such as wins/losses, team performance metrics (for e.g. three_point%, field_goal% etc). Used for every RQ
5. **trade_data.csv** - This dataset contains all the information about trades (for e.g. player name, team traded from, team traded to etc) that have happened over the years. Used for RQ2.
6. **player_college_stats.csv** - contains stats of players during college (before they were drafted into the NBA). Used in RQ1 and RQ2

Notebooks Used

1. **100_downloading_draft_data.ipynb** – main notebook for scraping draft data.
2. **200_downloading_player_data.ipynb** - main notebook for scraping player data before (during college) and after being drafted.

3. **300_downloading_teamData.ipynb** - main notebook for scraping team data.
4. **400_downloading_boxScore_data.ipynb** - main notebook for scraping individual game data.
5. **500_extracting_trade_data.ipynb** - used for extracting trade data from player dataset.
6. **100_clean_draft_data.ipynb**– notebook used for cleaning draft data.
7. **200_clean_player_data.ipynb** - notebook used for cleaning player data.
8. **300_clean_team_data.ipynb** - notebook used for cleaning team data.
9. **400_clean_team_games_data.ipynb** - notebook used for cleaning game data.
10. **500_clean_trade_data.ipynb** -notebook used for cleaning trade data.
11. **ResearchQ1.ipynb**
12. **ResearchQ2.ipynb**
13. **ResearchQ3.ipynb**
14. **ResearchQ4.ipynb**