Data Mining Project

POLITÉCNICA

LORENZO LEUZZI

December 2023

Contents

1 Introduction

In this final project, I undertake an individual role where I simulate being a prospective contractor, tasked with demonstrating the value of my analytical skills to a potential client. The project is divided into four integral sections. The first section, "Business Understanding", involves identifying and defining business goals, formulating relevant questions, and establishing a corresponding data mining goal. The second section, "Data Understanding: Descriptive and Diagnostic Analysis", requires a detailed examination of the dataset through various analytical techniques supported by statistical validation. The third section, "Modelling: Predictive Analysis", involves the application and comparison of multiple models, focusing on deriving insights and accurate predictions. Lastly, the "Evaluation" section connects the analytical findings to the original business objectives, proposing actionable business strategies and demonstrating how these insights address the initial business queries. This structured approach ensures a comprehensive analysis, aligning technical data mining processes with practical business applications.

The technical execution of the project is carried out using the *IBM SPSS Modeler*[1].

1.1 Dataset Overview

As a passionate enthusiast of the NBA and a basketball player myself, I've embarked on a project to uncover the correlation between the *NBA 2K*[2] rating system and real NBA performance metrics.

For this purpose I've decided to focus on the dataset *NBA 2K Ratings with Real NBA Stats*[3] from *Kaggle*[4], a data science platform. It includes NBA 2K Ratings for each player along with their real-life NBA statistics for the corresponding season. Notably, NBA 2K establishes player ratings prior to the onset of each season, prompting the need to merge each player's stats from the preceding season with their game ratings. For instance, the ratings in NBA 2K21 are based on the NBA 2019-20 season statistics. The dataset spans from the 2014-15 season (NBA 2K16) to the 2019-20 season (NBA 2K21). The author meticulously gathered the data from https://hoopshype.com/ and the official NBA statistics portal at https://stats.nba.com/players/.

The primary goal of this dataset analysis is to decode the factors in real-life NBA performance that significantly influence the NBA 2K ratings.

The dataset contains 2411 player-season instances and Table 1 is an exhaustive explanation of the 32 fields in the dataset.

2 Business Understanding

2.1 Business Goals

For the NBA 2K dataset, which includes both in-game ratings and real-life statistics of NBA players, we can focus on several actionable and promising business goals. These goals should be feasible and have the potential to lead to concrete actions such as marketing strategies, investment decisions, or the development of new products or features. Here are some suggested business goals:

- **Talent Scouting and Analytics Tool**: Utilize the dataset to create a talent scouting tool that helps identify underrated or upcoming players in the NBA. This tool can be marketed to professional sports analysts, fantasy league players, or even real-world NBA teams. By comparing in-game ratings with real-life statistics, the tool can highlight players who may be performing well in certain areas but are not yet widely recognized.

- **Enhancing Player Engagement and Game Development**: Utilize the dataset to analyze which aspects of a player's real-life performance most impact their in-game ratings, with a focus on identifying gaps or inconsistencies between these two dimensions. This analysis will inform the development of new in-game features, campaigns, and more sophisticated algorithms that closely link real-life NBA performances with the game.

[1] https://www.ibm.com/products/spss-modeler
[2] https://nba.2k.com/
[3] https://www.kaggle.com/datasets/willyiamyu/nba-2k-ratings-with-real-nba-stats
[4] https://www.kaggle.com/

Field	Description
ID	A unique identifier for each player's performance for a specific season.
PLAYER	The name of the NBA player (unique player identifier).
TEAM	The NBA team that the player is part of.
AGE	The age of the player during the season.
SEASON	The specific NBA season for the statistics.
GP (Games Played)	The total number of games in which the player appeared.
W (Wins)	Games won by the player's team when the player participated.
L (Losses)	Games lost by the player's team when the player participated.
MIN (Minutes Played)	The average minutes the player was on the court.
PTS (Points)	The average of points scored by the player per game.
FGM (Field Goals Made)	The average number of field goals made by the player per game.
FGA (Field Goals Attempted)	The average field goal attempts by the player per game.
FG% (Field Goal Percentage)	The average percentage of field goals made per game.
3PM (3-Point Field Goals Made)	The average number of 3-point field goals made.
3PA (3-Point Field Goals Attempted)	The average number of 3-point field goal attempts.
3P% (3-Point Field Goal Percentage)	The average percentage of 3-point field goals made.
FTM (Free Throws Made)	The average number of free throws made by the player.
FTA (Free Throws Attempted)	The average total free throw attempts.
FT% (Free Throw Percentage)	The average percentage of free throws made.
OREB (Offensive Rebounds)	The average number of offensive rebounds collected.
DREB (Defensive Rebounds)	The average number of defensive rebounds collected.
REB (Rebounds)	Total average number of rebounds collected.
AST (Assists)	The average number of assists made by the player.
TOV (Turnovers)	The average number of times the player lost possession.
STL (Steals)	The average number of times the player took the ball from an opponent.
BLK (Blocks)	The average number of times the player stopped an opponent's field goal attempt.
PF (Personal Fouls)	The average number of personal fouls committed.
FP (Fantasy Points)	A value representing the average overall performance for fantasy basketball per game. Formula: $(FP = Pts + 1.2Rebs + 1.5Ast + 3Stl + 3Blocks - TO)$
DD2 (Double Doubles)	Games with double digits in two key statistical categories.
TD3 (Triple Doubles)	Games with double digits in three key statistical categories.
+/- (Plus/Minus)	The point differential when the player is on the court.
Rankings	The player's overall rating in NBA 2K.

Table 1: Fields description

- **Betting Analysis**: Levarge the dataset to enhance the accuracy of betting odds for individual player performance, which in turn affects overall game outcomes.
- **Data-Driven Salary Assignemt**: Utilize the dataset to develop a systematic approach for determining player salaries based on predictive analytics.

2.2 Business Questions

2.2.1 Descriptive Analysis

- How are the 2K ratings distributed among NBA players?
- What is the average in-game performance rating for an NBA player as per the 2K dataset?
- Who are the top and bottom-tier players according to the 2K rankings, and how do these rankings compare to each other?
- Which NBA teams can be considered as the best and worst, and what factors contribute to these rankings?
- Do teams exhibit distinct distributions of player statistics?
- What is the typical progression of a player's 2K ratings with respect to their age?
- Do statistics of player differenciate among different season?

2.2.2 Diagnostic Analysis

The diagnostic analysis is a phase in the project where I seek to gain a deeper understanding of the factors that have the most significant impact on player rankings and those that do not. Essentially, it involves investigating which variables or aspects of player performance are strongly correlated with their 2K rankings and which ones have little to no influence. This phase helps us identify the key determinants of player rankings and provides insights into what makes a player's rating high or low in the context of the game.

2.2.3 Predictive Analysis

In the predictive analysis phase, our primary objective is to develop models and techniques that allow us to predict 2K rankings based on the information we have at our disposal. This often involves using statistical or machine learning models to establish a relationship between player performance metrics and their corresponding 2K ratings. By building these predictive models, we can make accurate forecasts of player rankings, enabling us to assess how well the game's ratings align with real-world player statistics. This phase is crucial for enhancing the realism and accuracy of the game's player ratings, ultimately leading to a more immersive gaming experience for users.

2.2.4 Prescriptive analysis

Here are potential business actions that can be derived from the results of the preceding analysis:

- Feature Engineering and Data Quality Enhancement: Begin by improving the quality and relevance of data used for player ratings. Identify and collect additional player statistics that have a significant impact on rankings, based on the diagnostic analysis findings. This may include incorporating more advanced metrics or gameplay data that better reflect a player's in-game performance.
- Real-time Updates: Implement a system that allows for real-time updates of player ratings based on their actual in-season performance. Continuously feed new statistics into the predictive model to adjust player ratings as the NBA season progresses. This ensures that player ratings remain dynamic and responsive to changes in real-world player performance.

- User Feedback Integration: Actively engage with the gaming community and collect user feedback regarding player ratings. Integrate user opinions and insights into the prescriptive analysis to identify discrepancies between in-game ratings and player expectations. This feedback loop can help fine-tune player ratings over time.

- Player Development Mechanics: Incorporate player development mechanics into the game based on the natural player trajectory. Allow users to train or improve their in-game players based on their in-game performance and achievements. This can add depth to the gaming experience and encourage users to strategize and invest in their player roster.

- Injury and Performance Impact: Implement an injury system that accounts for real-life injuries and their impact on player ratings. Injuries can lead to temporary decreases in player performance, adding realism to the game.

3 Data Understanding: Descriptive and Diagnostic Analysis

3.1 First steps

The dataset in question was found to be in a robust and ready-to-use state, requiring no additional preprocessing steps. All entries were complete, correctly formatted, and well-structured, ensuring seamless integration into the analysis pipeline.

In this NBA 2K dataset, a player name may appear multiple times but with different NBA seasons. Each row represents the player's performance and in-game rating for a specific season. These entries are intentionally kept separate rather than merged such as by taking the mean of their stats across seasons because each season is considered a distinct instance of the player's career performance. Merging these rows would dilute the individual impact of a single season's performance and mask the progression or regression of a player's skills and contributions over time. By retaining separate entries for each season, we can analyze and track a player's development, noting improvements, slumps, and other trends that may correlate with their NBA 2K ratings. For example, a player might have had a standout season due to exceptional performance or conversely, may have experienced a dip in form due to injury or other factors. Such nuances would be lost if the data across seasons were merged.

3.2 Outliers

In statistical analysis, outliers are data points that significantly differ from other observations. They can occur due to variability in the measurement or may indicate experimental errors. For continuous variables, a scatter plot can help visualize the relationship between two variables and identify potential outliers. Points that fall far from the main cluster of data may be considered outliers. For instance, if plotting players' Fantasy Points per game against minutes played, an outlier might be a player with an unusually high value of FP despite playing very few minutes. A player isn't considered an outlier solely due to an exceptionally high FP score; instead, such a player is recognized as a superstar. This potential last anomaly is illustrated in Figure 1, while Figure 2 and Figure 3 are the statistics plotted against the 2k rankings.

The data points predominantly cluster around the main body of the dataset, suggesting the absence of clear outliers. Should there be any data points that significantly deviate from this cluster, it would still be prudent to retain them. The data collection by NBA analysts is highly reliable since these statistics are sourced directly from the NBA. Therefore, any anomalous performances captured in the data are likely to be true representations of rare occurrences rather than errors. These outliers could provide crucial insights into exceptional player performances that are not appreciated enough or atypical game situations, which could be invaluable for comprehensive analysis and understanding of the game's nuances.

3.3 Statistics and Graphs

3.3.1 How are the 2K ratings distributed among NBA players?

In this dataset of NBA 2K player rankings, the lowest recorded ranking is 62. On the other end of the spectrum, the highest ranking achieved is an impressive 98, reflecting near-perfect player attributes and

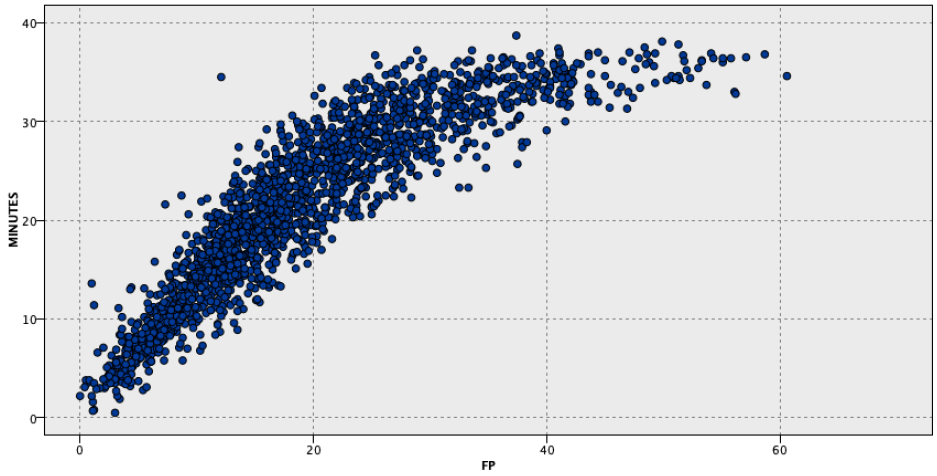


Figure 1: Scatterplot of FP against MIN

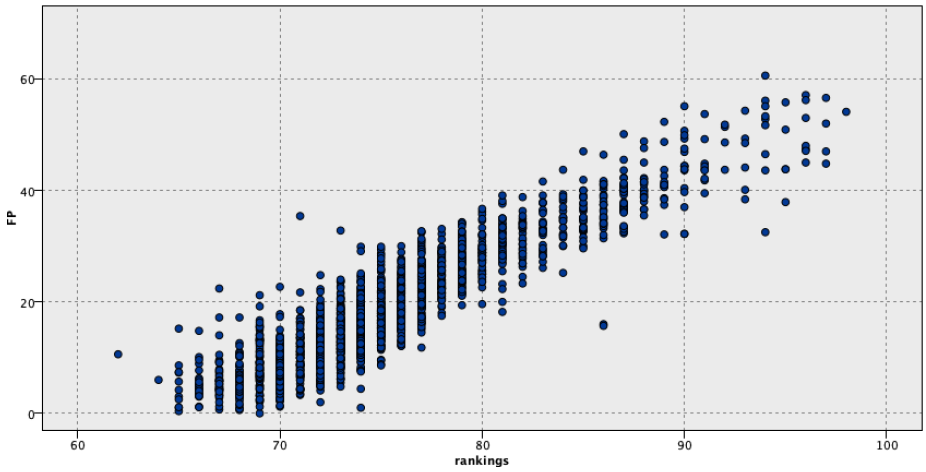


Figure 2: Scatterplot of FP against rankings

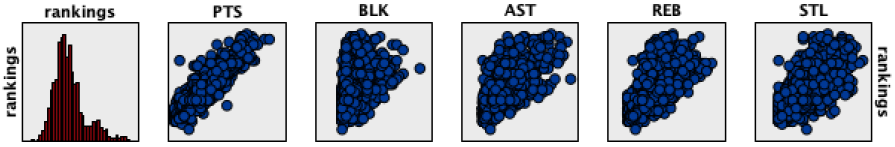


Figure 3: Scatterplots of PTS, BLK, AST, REB, STL against rankings

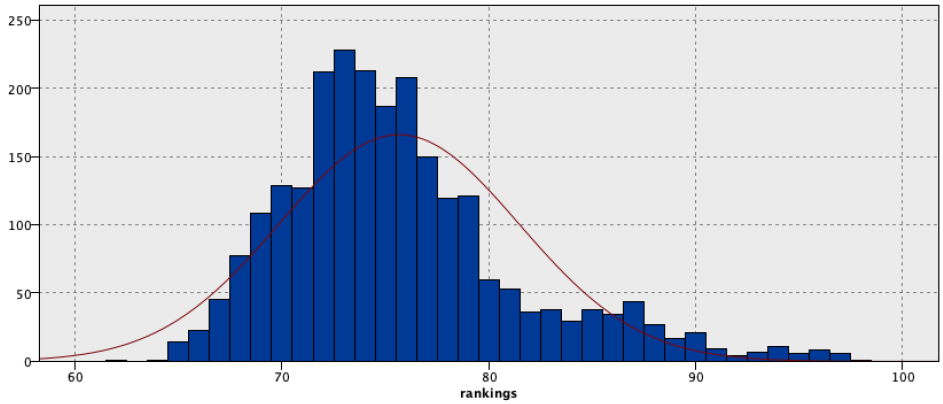


Figure 4: Rankings distribution among players

skills within the game. The mean ranking sits at 75.657, which suggests that the average competency level skews towards the higher end of the scale. The standard deviation of the rankings is 5.798, denoting that the majority of the players' rankings are spread within a relatively narrow range around the mean, pointing to a certain level of homogeneity in the in-game abilities of NBA players as rated by NBA 2K.

The histogram in Figure 4 showcases the distribution of NBA 2K player rankings with a superimposed normal distribution curve for comparison. It reveals a concentration of scores in the range $[70, 80]$, signifying a commonality in rankings within this segment of players. The shape of the distribution, with a skewness of 1.1, indicates a positive skew, where a greater number of players have rankings below the mode, and higher rankings are less common, as shown by the tail that stretches towards the higher end of the scale. The distribution's breadth spans from the lower 60s to the high 90s, and there's a noticeable decline in frequency as rankings rise, especially nearing the 90s, suggesting that top-tier rankings are rare. Such a distribution pattern suggests that analysts and teams might interpret rankings in the upper 80s as substantially above the norm, given the overall skewness of the data.

3.3.2 What is the average in-game performance rating for an NBA player as per the 2K dataset?

Taking into account the key statistics, the typical performance for a player over the course of a season averages out to the following:

- GP: 55.381, (Std. Dev. 22.817, 41.20%)
- W: 27.982, (Std. Dev. 15.299, 54.67%)
- L: 27.399, (Std. Dev. 13.728, 50.10%)
- MIN: 21.516, (Std. Dev. 8.763, 40.73%)
- PTS: 9.445, (Std. Dev. 6.029, 63.83%)
- REB: 3.966, (Std. Dev. 2.507, 63.21%)
- AST: 2.041, (Std. Dev. 1.826, 89.47%)
- TOV: 1.227, (Std. Dev. 0.812, 66.18%)
- STL: 0.689, (Std. Dev. 0.423, 61.39%)
- BLK: 0.443, (Std. Dev. 0.437, 98.65%)
- FP: 19.424, Std. Dev. 10.780, 55.50%)

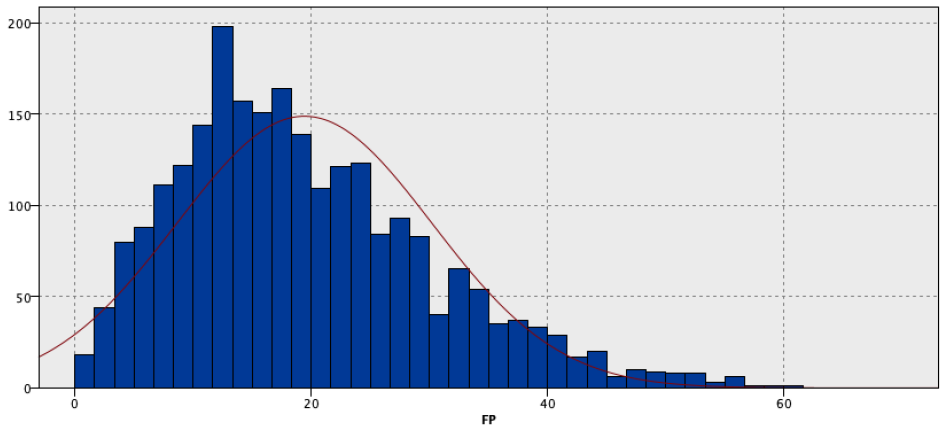


Figure 5: FP distribution among players

ID	PLAYER	TEAM	AGE	SEASON	GP	W	L	MIN	PTS
1130....	LeBron James	CLE	33....	2017-18	82....	50....	32....	36....	27....
709.0...	Kawhi Leonard	TOR	28....	2018-19	60....	41....	19....	34....	26....
173.0...	Giannis Antetokounmpo	MIL	25....	2019-20	63....	51....	12....	30....	29....
1528....	LeBron James	CLE	32....	2016-17	74....	51....	23....	37....	26....
739.0...	LeBron James	LAL	34....	2018-19	55....	28....	27....	35....	27....
1109....	Kevin Durant	GSW	29....	2017-18	68....	49....	19....	34....	26....
316.0...	LeBron James	LAL	35....	2019-20	67....	50....	17....	34....	25....
1904....	LeBron James	CLE	31....	2015-16	76....	56....	20....	35....	25....
212.0...	James Harden	HOU	30....	2019-20	68....	43....	25....	36....	34....
1507....	Kevin Durant	GSW	28....	2016-17	62....	51....	11....	33....	25....

Figure 6: Highest ranked players

Higher standard deviation percentages denote significant variation around the mean, implying that the average is not the most representative value of common performance, with the median and mean diverging. Illustrating this variance with individual graphs for each statistic would be cumbersome and potentially perplexing. Hence, in Figure 5, we present the distribution of Fantasy Points, a composite statistic that encapsulates the essence of other individual performance metrics, offering a overview of player performance variability. This histogram illustrates the distribution of fantasy points among players, revealing a skewed pattern, motivating the high Std. Dev. The bulk of the data is clustered on the left side, showing that most players score lower on the fantasy points scale, with a gradual decrease in the number of players achieving higher points. The tallest bars are concentrated in the lower scoring ranges, suggesting that a majority of players have a modest impact in terms of fantasy scoring metrics. The long tail towards the right indicates a few players achieve exceptionally high fantasy points, likely reflecting standout individual performances or star players with a dominant role in their teams. This distribution indicates that while the average player garners only a moderate number of fantasy points, there is a small proportion of players whose contributions are significantly above average, highlighting the disparity in performance levels within the league.

3.3.3 Who are the top and bottom-tier players according to the 2K rankings, and how do these rankings compare to each other?

The elite players in the NBA as indicated by the 2K rankings, are showcased in Figure 6. These individuals are not just familiar names but superstars who have been at the forefront of the league for a considerable time. On the other end of the spectrum are players with less recognition, occupying the lower ranks in the referenced figures (Figure 7).

To dissect the disparities between these two groups, we will utilize along the rankings, the Fantasy Points metric, which aggregates key statistics for an accessible overview. The bottom 100 players have an average ranking of 66.570 with a standard deviation of 1.057. In stark contrast, the top 100

ID	PLAYER	TEAM	AGE	SEASON	GP	W	L	MIN	PTS
2355....	Sean Kilpatrick	MIN	25....	2014-15	4.0...	2.0...	2.0...	17....	5....
2227....	Johnny O'Bryant III	MIL	22....	2014-15	34....	17....	17....	10....	2....
1146....	Marcus Georges-...	MIN	24....	2017-18	42....	24....	18....	5.3...	1....
1593....	RJ Hunter	CHI	23....	2016-17	3.0...	2.0...	1.0...	3.1...	0....
1964....	RJ Hunter	BOS	22....	2015-16	36....	21....	15....	8.7...	2....
1684....	Aaron Harrison	CHA	21....	2015-16	21....	15....	6.0...	4.4...	0....
1028....	Isaiah Hicks	NYK	23....	2017-18	18....	4.0...	14....	13....	4....
2091....	Bruno Caboclo	TOR	19....	2014-15	8.0...	7.0...	1.0...	2.9...	1....
2088....	Brandon Rush	GSW	29....	2014-15	33....	27....	6.0...	8.2...	0....
1169....	Naz Mitrou-Long	UTA	24....	2017-18	1.0...	0.0...	1.0...	0.5...	3....

Figure 7: Lowest ranked players

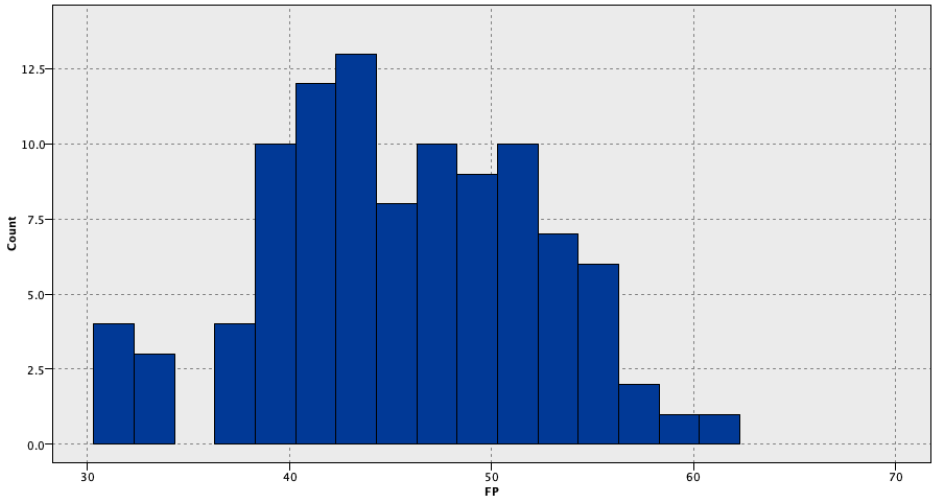


Figure 8: FP distribution for highest ranked players

players boast a much higher average ranking of 91.730, coupled with a standard deviation of 2.874. This stark disparity in rankings is mirrored in their fantasy point scores, with a meager 5.950 for the lower-ranked players against a robust 45.530 for the top performers. Interestingly, the lower-ranked players are, on average, 3.32 years younger than their top-ranked counterparts. The Fantasy Points distribution patterns for both player groups, as depicted in Figure 9 and Figure 8, show these clear differences.

Analyzing the teams these players play for, Figure 10 reveals a concentration of top players in the Golden State Warriors, reflecting the team's championship victories in 2015, 2017, and 2018, three out of the six seasons covered in this dataset. For the lower-ranked players, certain teams like IND and WAS stand have more instances (Figure 11), but without significant playing time, these players have had minimal impact on their teams' overall performance.

3.3.4 Which NBA teams are ranked as the best and worst?

Let's defined a rankings of team based on their wins a losses records. For each team we aggregate the number of wins a losses recorded by different players, then we calculate the record with the following formula.

$$Record = \frac{\#wins}{\#wins + \#losses} \tag{1}$$

As we were expected the Golden State Warriors are the best team winning around 75% of the games, while the New York Knicks won only 32% of their games (Figure 12).

Delving deeper into the composition of the Golden State Warriors and setting aside the less influential players, specifically those averaging fewer than 12 minutes per game or participating in less than 20 games in a regular season, we observe that the team's mean rankings and fantasy points exceed the

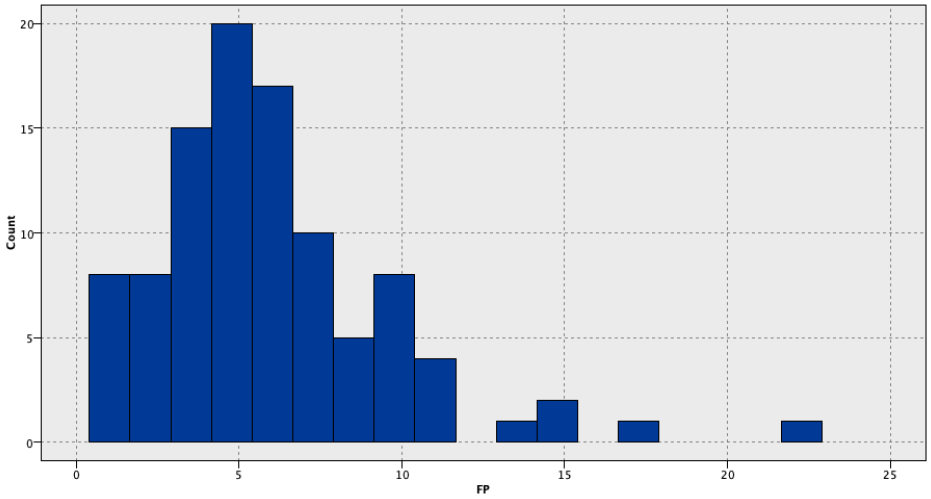


Figure 9: FP distribution for lowest ranked players

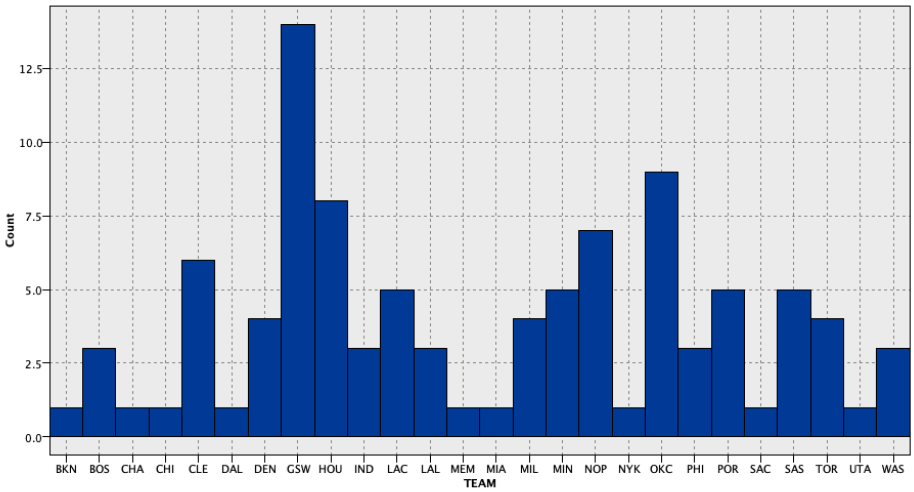


Figure 10: Team distribution for highest ranked players

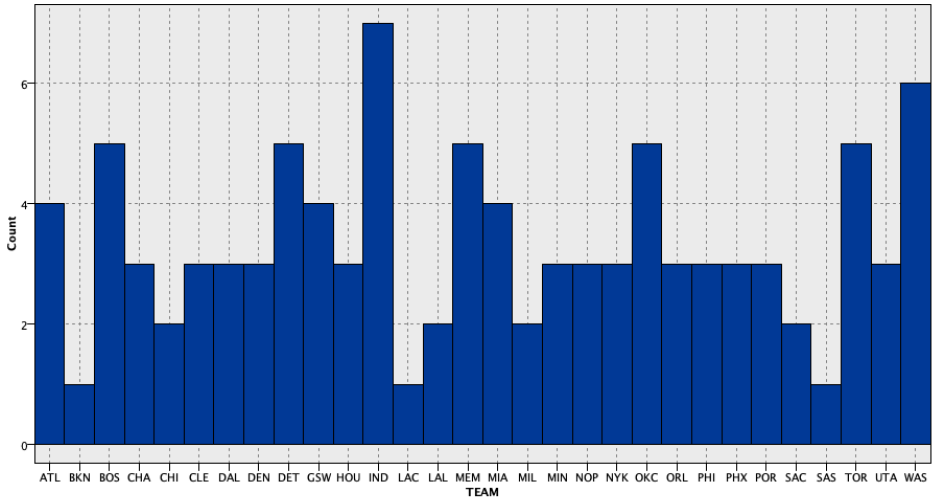


Figure 11: Team distribution for lowest ranked players

TEAM	W_S...	W_Mean	W_SDev	L_Sum	L_Mean	L_SDev	Record_Count	RECORD
GSW	3572....	42.024	22.298	1225....	14.412	9.838	85	0.745
TOR	3149....	37.488	16.389	1536....	18.286	9.625	84	0.672
SAS	2861....	39.192	16.568	1558....	21.342	10.190	73	0.647
HOU	2358....	31.440	18.034	1462....	19.493	11.973	75	0.617
LAC	2710....	36.622	14.419	1724....	23.297	9.688	74	0.611
BOS	3076....	35.356	13.935	1974....	22.690	9.791	87	0.609
OKC	2930....	34.881	14.719	2051....	24.417	10.312	84	0.588
UTA	2612....	32.247	14.274	1982....	24.469	11.912	81	0.569
MIL	2646....	33.494	14.055	2024....	25.620	12.122	79	0.567
POR	2578....	32.225	14.150	2073....	25.913	10.587	80	0.554
IND	2424....	30.300	13.523	2035....	25.438	11.746	80	0.544
MIA	2263....	27.938	13.189	2046....	25.259	11.797	81	0.525
CLE	2394....	28.500	16.774	2255....	26.845	12.830	84	0.515
DEN	2366....	28.506	13.573	2241....	27.000	12.085	83	0.514
MEM	2138....	24.575	14.378	2250....	25.862	13.148	87	0.487
ATL	2127....	27.269	15.514	2298....	29.462	12.716	78	0.481
WAS	2191....	26.398	14.414	2399....	28.904	12.739	83	0.477
NOP	1932....	25.091	11.635	2197....	28.532	11.712	77	0.468
DET	1898....	24.025	12.822	2279....	28.848	13.677	79	0.454
DAL	2029....	26.351	12.576	2438....	31.662	13.212	77	0.454
CHA	2004....	26.368	11.385	2417....	31.803	11.268	76	0.453
CHI	1961....	23.071	12.381	2539....	29.871	12.422	85	0.436
PHI	1822....	23.359	14.627	2386....	30.590	17.789	78	0.433
LAL	1857....	21.593	13.642	2609....	30.337	16.824	86	0.416
ORL	1586....	21.726	10.701	2371....	32.479	15.279	73	0.401
BKN	1727....	20.318	11.204	2645....	31.118	15.629	85	0.395
MIN	1637....	20.987	12.688	2541....	32.577	14.864	78	0.392
SAC	1853....	22.060	9.982	2893....	34.440	13.565	84	0.390
PHX	1395....	18.355	9.757	2682....	35.289	16.735	76	0.342
NYK	1397....	17.462	9.086	2956....	36.950	13.811	80	0.321

Figure 12: Teams ranked for win %

Figure 13: Rankings trajectory across age

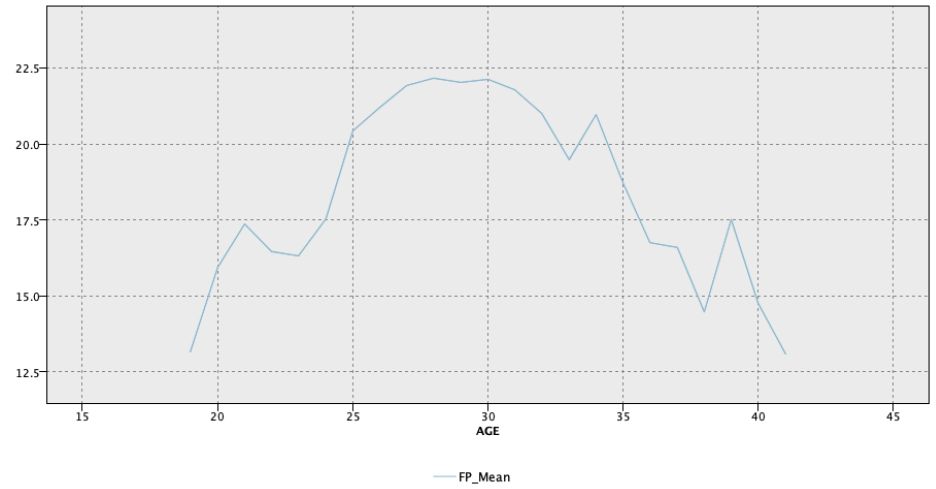


Figure 14: FP trajectory across age

league average, with values of 80.407 (Std. Dev. 7.859) for rankings and 24.341 (Std. Dev. 11.941) for fantasy points. The substantial standard deviations underscore the diversity of roles within the team and the wide dispersion of individual performances, reflecting a team with a variety of specialized and impactful player contributions.

3.3.5 What is the typical progression of a player's 2K ratings with respect to their age, and how does aging impact their in-game performance metrics?

As a player progresses through their NBA career, they typically experience phases of development and decline. To investigate this trend, I aggregated the players' data by their rankings and fantasy points, calculating the average for each age group. The resulting plot, as seen in Figure 13 and Figure 14, illustrates a common trajectory for NBA players: starting from their entry into the league, there is a phase of improvement where their performance enhances, usually peaking around the age of 30. After reaching this peak, the trend generally reverses, and a gradual decline in performance ensues. This pattern underscores the natural arc of an athlete's career, marked by initial growth, peak performance, and eventual regression.

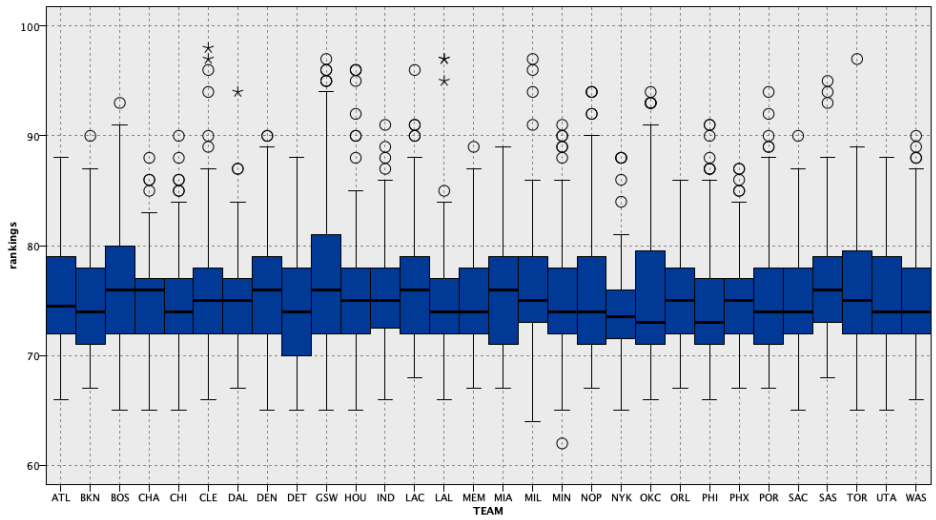


Figure 15: Boxplot of players's rankings across different teams

3.3.6 Do teams exhibit distinct distributions of player statistics?

The boxplot analysis reveals, Figure 15, the distribution of player rankings across different NBA teams, shedding light on the comparison between teams, relative player's performances, and stratification within each team. It appears that the player rankings across different NBA teams are quite consistent, the interquartiles ranges (the heights of the boxes) fall in the interval from 70 to 80 and do not vary dramatically from one team to another, suggesting that the central 50% of players on each team have comparable rankings distributions. Also, the central tendency of the rankings, marked by the median line within each box, falls within a similar range for most teams, around 75.

Exceptional cases, particularly those above the upper whisker, often represent star players whose rankings far exceed their team's average, emphasizing their pivotal role in the team's success. Conversely, data points below the lower whisker could indicate players that performs worst than their teammates. It's essential to understand that while some data points might seem to stand out as outliers, they actually represent unique instances within their teams and align consistently with the rankings of other players. These are not statistical anomalies but notable cases where individual player performance may deviate from the norm, reflecting the distinct skillset of different players.

3.3.7 Do players' statistics differentiate across different season?

In this scenario, the boxplot analysis (referenced as Figure 16) seemingly indicates that the distribution of players' rankings across various editions of the game and different seasons is fairly consistent. However, to accurately assess any potential differences in distributions among the levels of these categorical variables, additional analysis, such as ANOVA tests (in Section 3.4), is essential.

3.4 ANOVA tests

To accurately determine whether different levels of categorical variables (such as team and season) exhibit distinct distributions, we can employ ANOVA tests.

The first test was done to assess if there's a consistent distribution of player rankings or Fantasy Points across various teams. The findings presented in Figure 17 reveal that when it comes to Fantasy Points, there is no clear indication of any significant differences between the means. However, for rankings, the results are more ambiguous (marginal significance), suggesting that while the means may be similar, it's not definitive. The analysis of the boxplot in Figure 15 and the plot in Figure 10 highlighted that the Golden State Warriors team displayed a notably different interquartile range compared to other teams, along with having the most highly rated players. Consequently, I chose to exclude the GSW players from the dataset and re-conducted the ANOVA test. The revised results, as shown in Figure 18, provide clearer evidence that the distributions across teams likely share a common

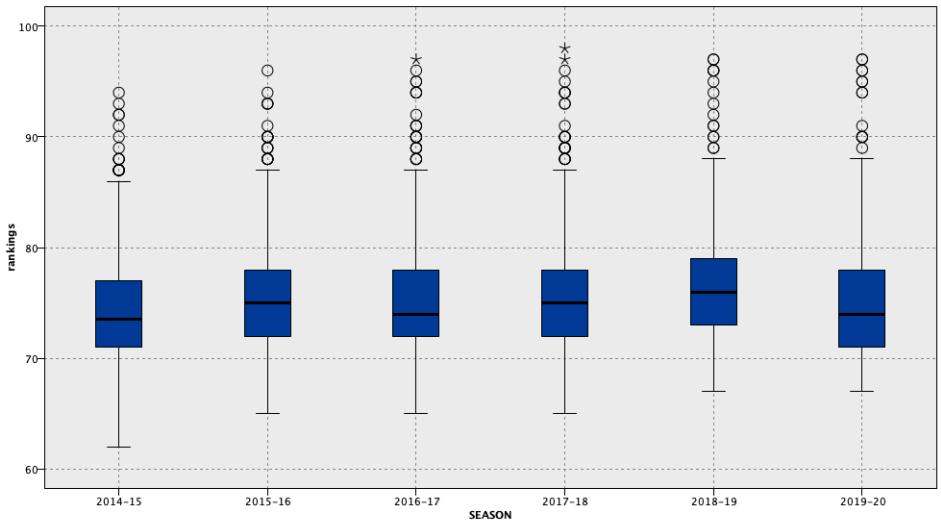


Figure 16: Boxplot of players's rankings across different game editions

F-Test	df	Importance
1.453	29, 2382	0.944 Marginal
0.495	29, 2382	0.011 Unimport...

Figure 17: ANOVA test results to test wether the mean of rankings and FP is different across teams

mean for both rankings and FP, as: $1 - Importance = p\text{-}value > 0.05$. This suggests that any previous uncertainties regarding team distributions were influenced by the inclusion of players from the GSW team.

Following the steps of the descriptive analysis, the subsequent investigation focused on determining whether a player's performance in terms of Fantasy Points and rankings was influenced by the season they played in. This was explored using an ANOVA test to analyze the impact of different seasons on these variables. The ANOVA results in Figure 19 present a compelling case, with the F-test values exceeding 2.5 for both rankings and Fantasy Points (FP), indicating that there are statistically significant differences in the means across different NBA seasons. This points to a genuine variation in player rankings and FP from season to season. Such differences may be attributed to changes in player performance, league dynamics, or other external factors affecting the game across the years.

F-Test	df	Importance
1.159	28, 2298	0.742 Unimport...
0.514	28, 2298	0.016 Unimport...

Figure 18: ANOVA test results to test wether the mean of rankings and FP is different across teams excluding GSW

Field	2014-15*	2015-16*	2016-17*	2017-18*	2018-19*	2019-20*	F-Test	df	Importance
rankings	74.609	75.725	75.723	75.904	76.912	75.207	6.662	5, 2406	1.000
	5.657	5.539	5.842	5.982	5.783	5.741			Import...
	0.296	0.290	0.295	0.293	0.303	0.255			
	366	364	393	418	364	507			
FP	19.729	19.734	19.138	19.245	21.286	18.013	4.134	5, 2406	0.999
	9.963	10.360	10.845	10.646	11.053	11.325			Import...
	0.521	0.543	0.547	0.521	0.579	0.503			
	366	364	393	418	364	507			

Figure 19: ANOVA test results to test wether the mean of rankings and FP is different across season

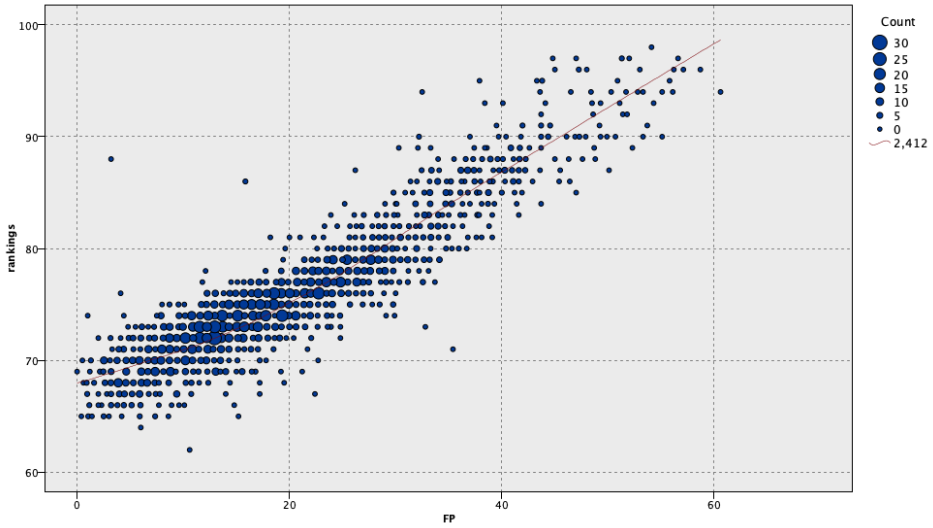


Figure 20: Scatterplot showing the strong correlation between FP and ranking

3.5 Correlations

Given that the dataset predominantly comprises continuous values, an in-depth correlation analysis is essential to discern the factors that substantially influence the 2k rankings of players.

In the realm of player performance, 2K rankings offer a multifaceted reflection of a player's actual contributions on the court, the Pearson coefficient was calculated between the rankings and all the others statistics (Table 2). The minutes played (MINUTES), points scored (PTS), field goals made (FGM), and attempts (FGA) showcase a robust correlation with player rankings, underscoring the pivotal role of active playtime and scoring efficiency in boosting a player's stature within the video-game. Similarly, free throws made (FTM) and attempted (FTA), as well as defensive rebounds (DREB), reveal a strong relationship with rankings, suggesting that a player's defensive prowess; offensive rebounds (OREB) and blocks (BLK), which are less correlated with player rankings, are commonly associated with taller players who may not excel in other skill areas. Turnovers (TO) in basketball, though typically negative, correlate positively with higher player rankings in 2K because players who handle the ball most tend to have a greater impact on the game through scoring and assists, despite a higher chance of losing possession. Essentially, the frequency of turnovers is offset by the significant positive contributions these key players make. The fantasy points (FP) metric stands out with an exceptionally strong correlation, indicating it as a comprehensive indicator of a player's overall impact and effectiveness in games. We can easily visualize this strong correlation in Figure 20. Contrastingly, metrics like age, losses (L), field goal percentage (FG%), three-point percentage (3P%), free throw percentage (FT%), and triple-doubles (TD3) show weaker correlations, implying these aspects, while significant, may not be as influential in determining a player's rank. It's these nuanced correlations that help paint a complete picture of a player's ranking, where not just the points they score but also how they play the game holistically contributes to their standing in the league.

Analyzing the interactions among the various variables within the dataset is also crucial. The Scatterplot Matrix, as illustrated in Figure 21, effectively showcases these interrelationships. Notably,

Rankings	1.000/Perfect
AGE	0.149/Weak
GP	0.462/Medium
W	0.484/Medium
L	0.229/Weak
MINUTES	0.779/**Strong**
PTS	0.895/**Strong**
FGM	0.894/**Strong**
FGA	0.850/**Strong**
FG%	0.275/Weak
3PM	0.472/Medium
3PA	0.469/Medium
3P%	0.162/Weak
FTM	0.823/**Strong**
FTA	0.826/**Strong**
FT%	0.285/Weak
OREB	0.415/Medium
DREB	0.712/**Strong**
REB	0.659/Medium
AST	0.645/Medium
TOV	0.771/**Strong**
STL	0.622/Medium
BLK	0.446/Medium
PF	0.543/Medium
FP	0.917/**Strong**
DD2	0.634/Medium
TD3	0.316/Weak
+/-	0.414/Medium

Table 2: Correlation coefficients between rankings and all the others features in the dataset

the strong correlations identified in the correlation matrix (Table 3) are the following.

- PTS and FGM: 0.987/Strong
- PTS and TOV: 0.833/Strong
- PTS and MINUTES: 0.869/Strong
- FGM and TOV: 0.822/Strong
- FGM and MINUTES: 0.872/Strong
- AST and TOV: 0.836/Strong
- AST and STL: 0.667/Strong
- REB and BLK: 0.704/Strong
- TOV and MINUTES: 0.765/Strong
- STL and MINUTES: 0.723/Strong

The strong correlations observed in the dataset reveal compelling insights into player performances and their in-game contributions. Points scored and field goals made exhibit an almost perfect correlation, trivially indicating that as players successfully make more shots, their point totals inevitably increase. Additionally, the correlation between minutes and the following statistics: points scored, field goals made, turnovers and steals indicate that active engagement in the game leads to higher statistics in these categories both for better and worse. A significant relationship is evident between field goals made (and thus points scored) and turnovers, indicating players with higher scoring often incur more turnovers. Similarly, assists show a correlation with turnovers, which highlights the dual nature of a playmaker's role in generating scoring chances and the associated risk of losing possession. The minutes a player spends on the court act as a *lurking variable* in these scenarios, suggesting that increased playing time may lead to correlate variables that counterintuitively are not. Assists also correlate well with steals (STL), suggesting that players who are good at distributing the ball are also adept at disrupting the opposing team's offense. Finally, The relationship between rebounds (REB) and blocks (BLK) is strong, pointing to a probably taller player who has defensive prowess and ability to control the paint.

The Pearson coefficients for the Fantasy Points attribute are not computed, as this statistic is derived from a linear combination of other stats in the dataset. Consequently, their correlation is inherently implied, rendering the calculation of these coefficients redundant.

For every correlation discussed, it's noteworthy that the *p-values* were consistently low, always falling below 0.05. This indicates a significant level of statistical significance for each correlation, reinforcing the reliability of the relationships identified between various variables within themselves and the 2k rankings of players.

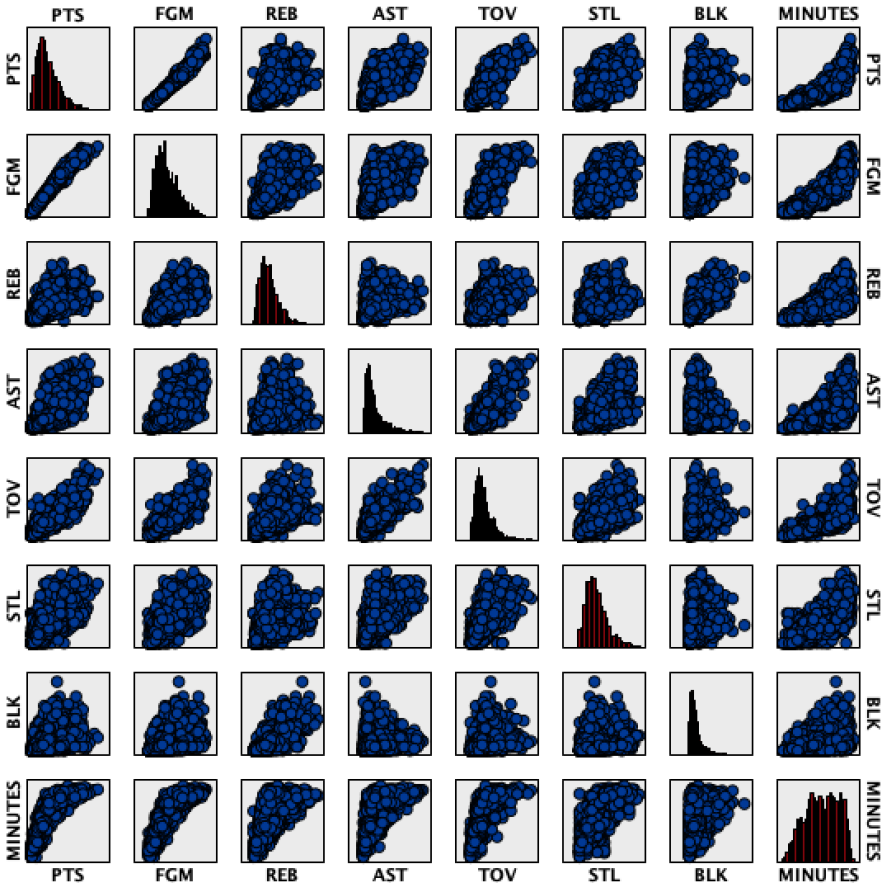


Figure 21: Scatterplot Matrix showing the correlations between different statistics

	PTS	FGM	REB	AST	TOV	STL	BLK	MINUTES
PTS	1.000/Perfect	0.987/**Strong**	0.573/Medium	0.673/**Strong**	0.833/**Strong**	0.634/Medium	0.318/Weak	0.869/**Strong**
FGM	0.987/**Strong**	1.000/Perfect	0.622/Medium	0.649/Medium	0.822/**Strong**	0.624/Medium	0.365/Medium	0.872/**Strong**
REB	0.573/Medium	0.622/Medium	1.000/Perfect	0.250/Weak	0.517/Medium	0.379/Medium	0.704/**Strong**	0.613/Medium
AST	0.673/**Strong**	0.649/Medium	0.250/Weak	1.000/Perfect	0.836/**Strong**	0.667/**Strong**	0.019/Weak	0.654/Medium
TOV	0.833/**Strong**	0.822/**Strong**	0.517/Medium	0.836/**Strong**	1.000/Perfect	0.665/Medium	0.260/Weak	0.765/**Strong**
STL	0.634/Medium	0.624/Medium	0.379/Medium	0.667/**Strong**	0.665/Medium	1.000/Perfect	0.194/Weak	0.723/**Strong**
BLK	0.318/Weak	0.365/Medium	0.704/**Strong**	0.019/Weak	0.260/Weak	0.194/Weak	1.000/Perfect	0.344/Medium
MINUTES	0.869/**Strong**	0.872/**Strong**	0.613/Medium	0.654/Medium	0.765/**Strong**	0.723/**Strong**	0.344/Medium	1.000/Perfect

Table 3: Correlation matrix for PTS, FGM, REB, AST, TOV, STL, BLK, MIN

ID	PLAYER	rankings	OD	OFF_SCORE_BIN	DEF_SCORE_BIN
1613....	Russell Westbrook	94.000	4.600	5	4
1215....	Russell Westbrook	93.000	4.600	5	4
812.0...	Russell Westbrook	90.000	4.600	5	4
173.0...	Giannis Antetokounmpo	97.000	4.600	5	4
525.0...	Anthony Davis	94.000	4.400	4	5
1311....	Anthony Davis	94.000	4.400	4	5
636.0...	Giannis Antetokounmpo	96.000	4.400	4	5
686.0...	Joel Embiid	91.000	4.400	4	5
892.0...	Anthony Davis	94.000	4.400	4	5
958.0...	DeMarcus Cousins	90.000	4.400	4	5
2128....	DeMarcus Cousins	87.000	4.400	4	5
1987....	Russell Westbrook	93.000	4.200	5	3
2003....	Stephen Curry	94.000	4.200	5	3
2195....	James Harden	92.000	4.200	5	3
316.0...	LeBron James	97.000	4.200	5	3
322.0...	Luka Doncic	94.000	4.200	5	3
1479....	John Wall	90.000	4.200	5	3
1049....	James Harden	96.000	4.200	5	3
739.0...	LeBron James	97.000	4.200	5	3
667.0...	James Harden	96.000	4.200	5	3

Figure 22: Players ranked by their OD score

3.6 Offense-Defense Analysis

As specified before the Fantasy Points score was calculated as $FP = Pts + 1.2Rebs + 1.5Ast + 3Stl + 3Blocks - TO$. This formula provided a comprehensive measure of a player's overall performance in the game. To extend this approach, we can introduce a similar quantitative ranking and grouping mechanism, inspired by the Recency-Frequency-Monetary (RFM) Analysis method, by dividing the components of the FP score into two distinct categories: Offense and Defense. By doing so, we can create two new scores, *Offense Score* and *Defense Score*, which isolate and quantify the offensive and defensive contributions of players separately. The Offense Score would encapsulate the player's performance during offensive plays, primarily focusing on points scored and assists, while the Defense Score would concentrate on the player's defensive actions, like rebounds, steals, and blocks. This division allows for a more nuanced analysis and comparison of players based on their specific strengths in either offense or defense. In particular these scores are calculated as following.

$$O = Pts + 1.5Ast \tag{2}$$

$$D = 1.2Rebs + 3Stl + 3Blocks \tag{3}$$

Subsequently, we categorized these values into quintiles (5 bins), in a manner similar to the RFM analysis. The final Offensive-Defensive (OD) scores were derived, giving slightly more weight to offensive performance.

$$OD = 0.6O + 0.4D \tag{4}$$

The players rankings using the new system based on the OD score is shown in Figure 22. Interestingly, this table lists two instances of the Most Valuable Player (MVP, award given to the best player of the season) winners: Giannis Antetokounmpo for the 2019-2020 season and Russell Westbrook for the 2016-2017 season—within the top entries. In contrast, the table in Figure 6 featured only one MVP in the top ten-plus entries. This could imply that the OD score serves as a superior predictor for the MVP award winner. However, given the limited number of instances for this award (only one per season), it is not conclusive to make such a claim, as it could be a result of chance.

To draw comparisons between the NBA2k rankings and the OD score outcomes, I normalized the OD score to match the 62-98 range of the rankings using this formula.

$$\text{Transformed } OD = 62 + \frac{(\text{Original } OD - 1)}{(5 - 1)} \times (98 - 62) \tag{5}$$

As observed in Figure 23, there is a strong correlation of 0.883 between the OD score and the rankings. This indicates that the OD analysis is well-aligned, especially given that the rankings are a solid reflection of the players' real-life capabilities.

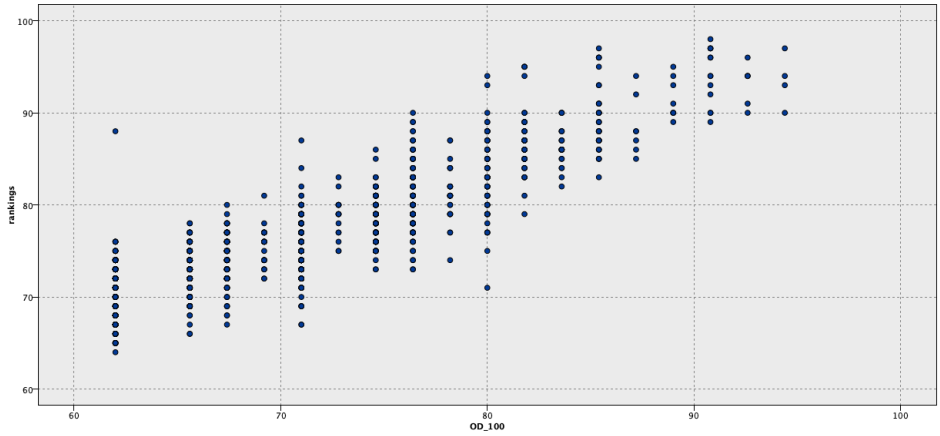


Figure 23: Scatterplot showing correlation between OD and rankings

3.6.1 Players Segmentation

The Offense-Defense (OD) analysis has yielded distinct segments of players within the league, categorized by their offensive and defensive score bins. Each segment encapsulates a specific player archetype, reflecting their impact and role in the league:

- **SUPERSTAR** (4-5 Offense; 3-5 Defense): These players are the cream of the crop, consistently demonstrating dominance in the league. They excel on both ends of the court, making significant contributions to their teams' successes.
- **SCORER** (4-5 Offense; 1-2 Defense): Players in this category are offensive juggernauts, often leading their teams in scoring. Some, like Stephen Curry, possess such a remarkable scoring ability that they border on superstar status, despite not having the highest defensive scores.
- **SUPERDEFENDER** (1-2 Offense; 4-5 Defense): These players may not light up the scoreboard, but their defensive prowess is invaluable. They specialize in shutting down opponents, altering the game's dynamic with their defensive skills.
- **POTENTIAL** STAR (3 Offense; 3-5 Defense): These players show promise on both sides of the game. They are solid contributors and have the potential to rise to superstar status with further development and consistency.
- **ROLE PLAYER** (2 Offense; 2-3 Defense): Integral to any team, these players are typically found coming off the bench or in a rotational role. They are reliable in their contributions and can be counted on to play their part effectively.
- **MARGINAL** (1 Offense; 1-2 Defense): These players have a limited influence on their team's performance. Their contributions are minimal, and they often have a negligible impact on the outcome of games.
- The remaining players, who don't fit into these categories, are classified as **STARTER**. These are the good, reliable players who may not exhibit the extremes of the segments above but are nonetheless starters due to their well-rounded game and consistent performance.

Each player was so tagged within they respective segment. The pie chart, in Figure 24, illustrates the distribution of player segments based on this tagging. The largest segment is MARGINAL, followed by ROLE suggesting that a significant portion of players are less influential yet very important for a team composition. STARTER and POTENTIAL STAR, form substantial parts of the pie, indicating a healthy number of players with potential for significant impact. Segments like SCORER, SUPERDEFENDER and SUPERSTAR are smaller, highlighting that such specialized roles or exceptional roles are less common. Now analysis how are composed each segment let's see their rankings and FP distribution. The histogram in Figure 25 showcases the distribution of players' rankings from

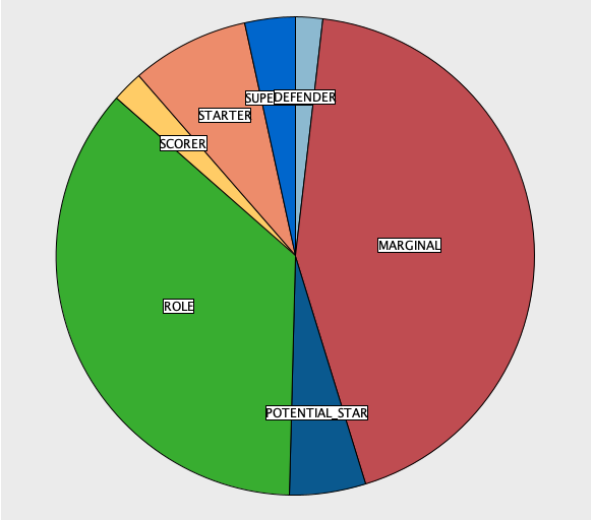


Figure 24: OD segments pie distribution

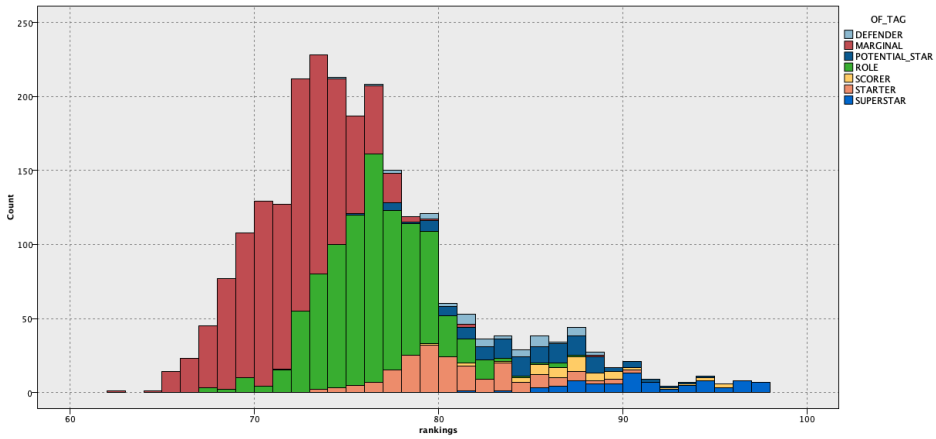


Figure 25: Ranking distribution colored by OD segment

a dataset, with each color representing different segments based on an Offense-Defense analysis. The MARGINAL segment (red) dominates the lower-ranking spectrum and as the rankings improve, we see a gradual transition through ROLE, STARTER, segments, reflecting a decrease in player count as performance criteria tighten. POTENTIAL_STAR, SCORER, and SUPERDEFENDER are present in the medium-high spectrum of the ranking. Notably, the SUPERSTAR segment (dark blue) is only present at the highest ranking levels, underscoring the rarity and exclusivity of top-performing players. The same behaviour is followed for the FP distribution in Figure 26. Overall, this histograms visually summarizes the relationship between a player's ranking and their designated Offense-Defense category, highlighting the pyramid-like structure of talent distribution in the dataset. From the table in Figure 27, SUPERSTARS are the top performers with the highest average ranking of 91 and a substantial Fantasy Points average of approximately 46.9, suggesting they are highly influential in games. POTENTIAL_STAR, SCORER and SUPERDEFENDER segments follow, indicating players with significant contributions reflected in high rating and high FP. STARTER and ROLE are mid-tier segments, representing reliable players with average stats, whereas ROLE players have lower mean rankings and Fantasy Points, implying they have specialized roles or less game impact. MARGINAL players have, trivially, the lowest averages in both rankings and Fantasy Points. The standard deviations across all categories are relatively small, which demonstrates the segmentation system's effectiveness in categorizing players with minimal dispersion. The exception to this trend is found in the FP score for marginal players, suggesting that some of them might somehow influence the game. Examining also

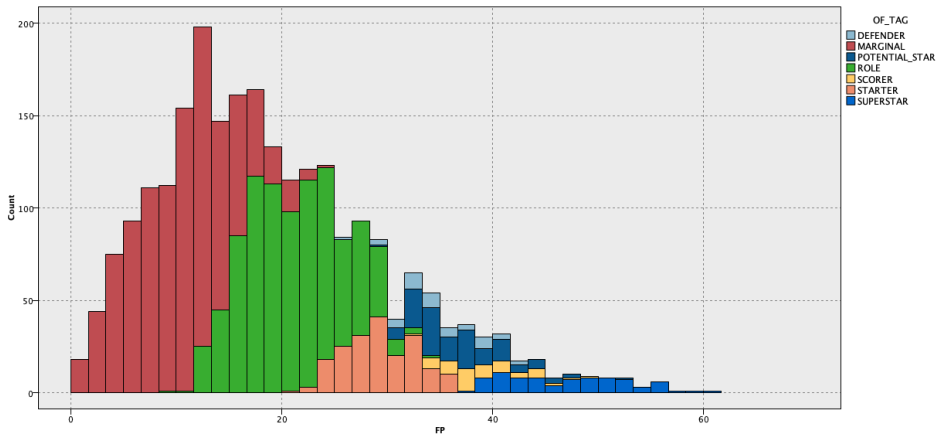


Figure 26: FP distribution colored by OD segment

FP_Mean	FP_SDev	rankings_Mean	rankings_SDev	OF_TAG	Record_Count
46.898	5.760	91.000	3.846	SUPERSTAR	82
36.682	4.245	84.185	3.864	POTENTIAL_STAR	124
38.932	3.582	87.320	3.508	SCORER	50
34.978	3.810	82.711	3.375	DEFENDER	45
29.265	3.429	80.487	3.466	STARTER	193
20.982	4.525	75.863	2.749	ROLE	871
10.518	4.518	71.522	2.803	MARGINAL	1047

Figure 27: FP and ranking aggregated by OD segment

the dataset's other statistics against the OD tagging, we can conclude that they closely align with our expectations (Figure 28, Figure 29, and Figure 30).

Specifically, SUPERSTARS category not only display the highest scores but also feature impressive statistics, like SCORER, but also feature impressive overall statistic. Conversely, the SUPERDEFENDER and POTENTIAL STAR categories balance their performance with supplementary statistics. STARTER players, on the whole, outperform ROLE players. MARGINAL players generally have lower statistics in most aspects of the game, with higher relative Std. Dev. meaning that they can produce in some cases decent performance.

3.7 Clustering

The goal of this section is to delineate unique clusters of players based on their real-life statistical data and to discern the principal characteristics of these clusters to see if they aligned with the 2K rankings or the Offense-Defense analysis. Another potential objective for the clustering output could be for the algorithm to effectively identify basketball playing positions such as Point Guard (PG), Shooting Guard (SG), Small Forward (SF), Power Forward (PF), and Center (C).

To enhance the visualization of clusters in relation to the rankings, I introduced an additional field, STATUS, to group the ranking values. This grouping was carried out manually using the *Derive*

MINUTES_Mean	MINUTES_SDev	PTS_Mean	PTS_SDev
34.767	1.610	24.528	3.962
32.340	2.575	17.730	3.090
34.116	2.013	23.284	3.289
30.329	2.705	13.307	2.181
31.212	2.851	16.491	2.968
24.890	4.556	10.280	2.482
13.621	5.381	4.462	1.905

Figure 28: MIN and PTS aggregated by OD segment

REB_Mean	REB_SDev	AST_Mean	AST_SDev	TOV_Mean	TOV_SDev
7.655	2.615	6.420	2.411	3.287	0.884
7.976	2.377	3.548	1.835	2.184	0.607
4.138	0.705	6.162	1.278	2.912	0.586
11.142	2.354	1.393	0.653	1.742	0.364
3.756	1.066	4.347	1.669	2.136	0.510
4.290	1.966	2.120	1.050	1.285	0.413
2.654	1.614	0.859	0.574	0.633	0.329

Figure 29: REB, AST, and TOV aggregated by OD segment

STL_Mean	STL_SDev	BLK_Mean	BLK_SDev
1.498	0.429	0.791	0.569
1.105	0.468	0.983	0.547
1.144	0.232	0.310	0.136
0.882	0.366	1.778	0.625
1.009	0.334	0.289	0.166
0.780	0.331	0.442	0.368
0.412	0.257	0.329	0.326

Figure 30: STL and BLK aggregated by OD segment

Node, as the *Binning* Node did not produce the desired output, likely due to the skewed distribution of rankings. Therefore, I created bins following the pattern illustrated in Figure 31.

In all the autoclustering processes, the parameter controlling the number of clusters falls within the range of $[2, 15]$. The clustering algorithms employed are *Kohonen* (fixed width of 3), *K-means*, and *TwoStep* (Figure 32) and their results are evaluated and ranked based on their silhouette scores.

3.7.1 Clustering using Players' Statistics

Initially, I filtered irrelevant variables, such as ID and PLAYER, that had no bearing on the study. Following this, I undertook data preparation, which involved cleansing and transforming the dataset (e.g. putting continuos input field on a common scale) in readiness for analysis. The possibility of developing dummy variables for the categorical attribute representing the season was considered. However, as illustrated in Figure 33, the data preparation node indicated that the SEASON attribute would likely be the least significant predictor for rankings and intuitively the season has nothing to do with the player role; hence, I chose to exclude it from our consideration. The ANOVA analysis implied differences in the players' mean performances across seasons, yet, aligning with our expectations, numerous other variables proved to be more predictive and influential. For the same reason, I also excluded age, team, and +/-, opting, for now, to focus on the remaining variables for the analysis. Following the completion of the autoclustering process, the generated models are illustrated in Figure 34. However, the cluster with the highest silhouette score, which was the k-means with 2 clusters, proved to be overly general as it merely distinguished between the top half and the bottom half of players. Consequently, this model was rejected. Subsequently, when exploring the k-means model with 3 clusters, as shown in the rankings distribution in Figure 35, it became evident that three distinct categories of players had been identified: Good players, Average players, and Non-influential players. This hierarchical structure was consistently reflected in all the statistical measures, Figure 36. Another model results analysis involved the utilization of k-means with 7 clusters which which reached a silhouette of 0.312, the visualizations of these clusters colored by status are provided in Figure 37.

By examining the table presented in Figure 38, we can provide the following observations regarding

Set field to	If this condition is true
MARGINAL	rankings <= 72
ROLE	rankings > 72 and rankings <= 78
STARTER	rankings > 78 and rankings <= 88
SUPERSTAR	rankings > 88

Figure 31: Ranking's binning criteria

Use?	Model type	Model parameters	No of models
☑	Kohonen	Specify...	14
☑	K-means	Specify...	14
☑	TwoStep	Specify...	14

Figure 32: Clustering algorithm for the *Autoclustering* node

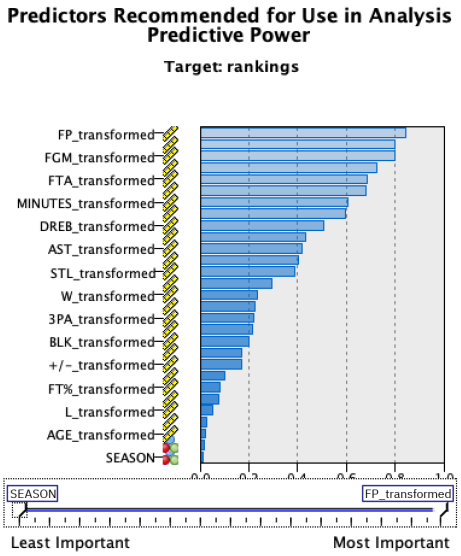


Figure 33: Predictors reccomended by the *Auto Data Preparation* node

Sort by: Use ◉ Ascending ○ Descending ✕ Delete Unused Models View: Training set

Use?	Graph	Model	Build Time (mins)	Silhouette	Number of Clusters	Smallest Cluster (N)	Smallest Cluster (%)	Largest Cluster (N)	Largest Cluster (%)	Smallest/ Largest	Importance
☑		K-...	< 1	0.420	2	1138	47	1274	52	0.893	1.0
☐		Two...	< 1	0.417	2	988	40	1424	59	0.694	1.0
☐		K-...	< 1	0.37	3	618	25	1143	47	0.541	1.0
☐		Two...	< 1	0.336	3	242	10	1312	54	0.184	1.0
☐		K-...	< 1	0.312	7	111	4	685	28	0.162	1.0
☐		Two...	< 1	0.311	5	216	8	944	39	0.229	1.0
☐		K-...	< 1	0.309	4	327	13	794	32	0.412	1.0

Figure 34: Clustering models

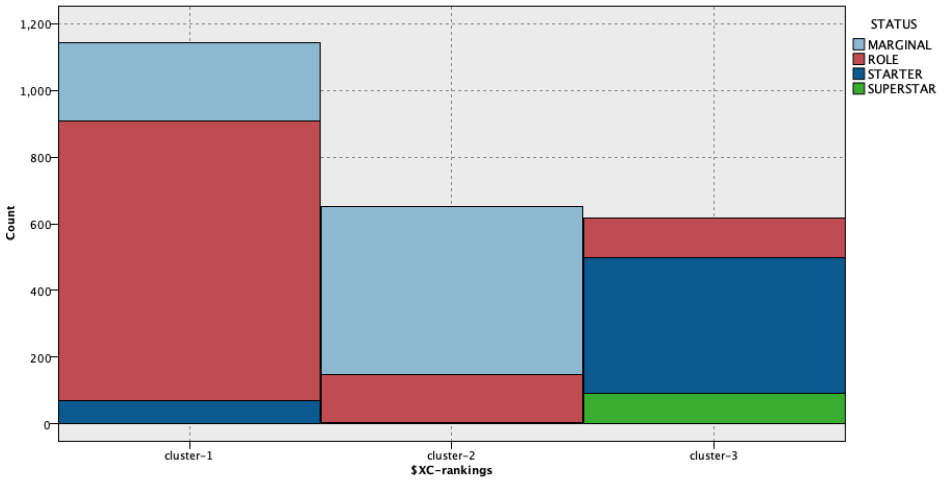


Figure 35: Clustering distributions colored by STATUS for k-means 3

FP_Mean	FP_SDev	rankings_Mean	rankings_SDev	$XC-rankings	Record_Count
34.037	7.354	83.073	5.149	cluster-3	618
8.787	4.395	70.573	2.806	cluster-2	651
17.581	4.656	74.542	2.534	cluster-1	1143

Figure 36: Statistics aggregated by clusters for k-means 3

FP_Mean	FP_SDev	rankings_Mean	rankings_SDev	$XC-rankings	Record_Count
39.928	6.762	87.275	4.512	cluster-5	240
23.586	5.893	76.484	4.214	cluster-6	155
30.187	6.639	80.345	3.907	cluster-3	220
23.310	4.172	76.965	2.816	cluster-1	536
4.476	3.069	68.910	2.843	cluster-2	111
14.286	3.371	73.372	2.244	cluster-4	685
9.018	3.516	70.634	2.504	cluster-7	465

Figure 37: Clustering distributions colored by STATUS for k-means 7

FP_Mean	FP_SDev	rankings_Mean	rankings_SDev	$XC-rankings	Record_Count
39.928	6.762	87.275	4.512	cluster-5	240
23.586	5.893	76.484	4.214	cluster-6	155
30.187	6.639	80.345	3.907	cluster-3	220
23.310	4.172	76.965	2.816	cluster-1	536
4.476	3.069	68.910	2.843	cluster-2	111
14.286	3.371	73.372	2.244	cluster-4	685
9.018	3.516	70.634	2.504	cluster-7	465

Figure 38: Statistics aggregated by clusters for k-means 7

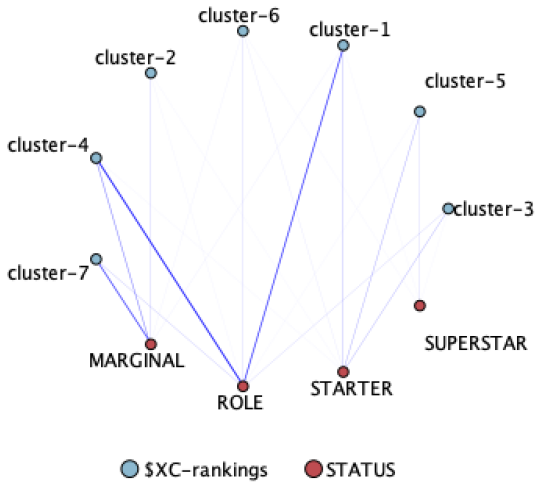


Figure 39: Web association between the clusters and players' STATUS

the clustering.

- Cluster 1: These are generally good players with high minutes, slightly above average (similar to cluster 6).
- Cluster 2: Bottom-tier players with overall lower statistics.
- Cluster 3: Exceptional players, but not quite superstars. They have high usage minutes and excel in collecting rebounds and blocking shots, indicating strong defensive abilities.
- Cluster 4: Rotational players who maintain overall decent statistics.
- Cluster 5: Superstars with outstanding overall statistics, often playing the most minutes. Slightly more centered in the offensive part of the game.
- Cluster 6: Similar to cluster 1.
- Cluster 7: Marginal players who receive some minutes but have a limited impact on the game.

Cluster 1 and 6 can be surely merged.

Figure 39 illustrates the visual representation of the relationships between the binned rankings and cluster divisions. The stronger relationship are between ROLE player and cluster 1 and 4, as well as between MARGINAL and cluster 4 and 7. STARTER are mostly concentrated in cluster 1, 3 and 5 while SUPERSTAR players appear only in cluster 3.

This model summary is available in Figure 40. The pie chart, in Figure 41 displays the distribution of sizes for seven different clusters generated by a k-means clustering analysis. The largest cluster, making up 28.4% of the data, significantly outweighs the smallest cluster, which comprises only 4.6%. The ratio of the size of the largest cluster to the smallest is 6.17, indicating substantial variation in cluster sizes. The various predictors importance is depicted in Figure 42.

3.7.2 Clustering using Offense-Defense Analysis

I attempted another clustering approach using the Offense and Defense scores discussed in Section 3.6, while maintaining the previously described settings. The most effective clustering techniques, as depicted in Figure 43, turned out to be the Kohonen map and TwoStep, both utilizing a relatively high number of clusters, ranging from 14 to 17. In these cases, the silhouette scores approached near-perfect values of 1.0. However, it's essential to consider that there are 25 potential combinations for our two parameters (since the scores fall within the range of $[1, 5]$) – in our case there's only 20 of those combinations – and typically, we aim for the number of clusters to be significantly less than the

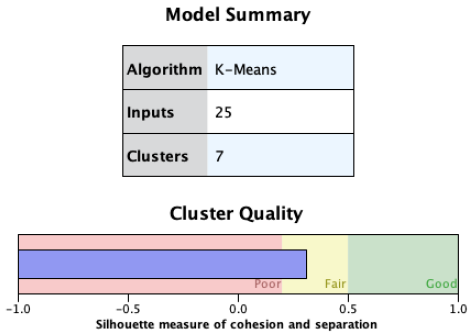


Figure 40: Model summary for k-means 7

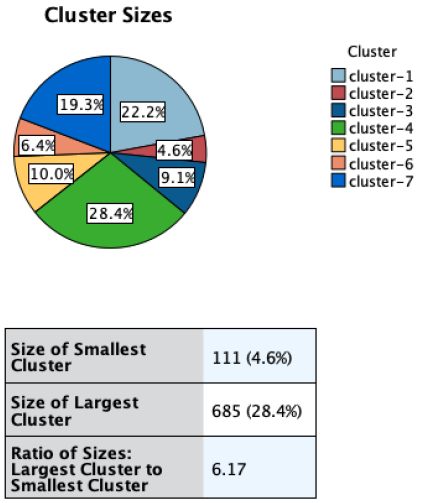


Figure 41: Cluster sizes for k-means 7

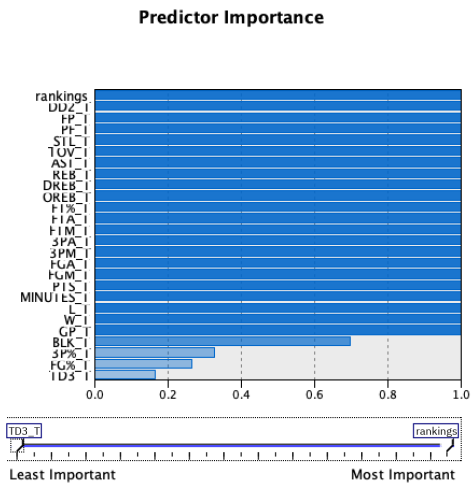


Figure 42: Predictors importance for k-means 7

Sort by: Use | Ascending | Descending | Delete Unused Models | View: Training set

Use?	Graph	Model	Build Time (mins)	Silhouette	Number of Clusters	Smallest Cluster (N)	Smallest Cluster (%)	Largest Cluster (N)	Largest Cluster (%)	Smallest/ Largest	Importance
☑		Koh...	< 1	0.986	17	3	0	635	26	0.005	1.0
☐		Two...	< 1	0.98	15	15	0	635	26	0.024	1.0
☐		Koh...	< 1	0.979	17	3	0	635	26	0.005	1.0
☐		Koh...	< 1	0.976	16	3	0	635	26	0.005	1.0
☐		Koh...	< 1	0.973	16	1	0	635	26	0.002	1.0
☐		Two...	< 1	0.971	14	19	0	635	26	0.030	1.0
☐		Koh...	< 1	0.969	15	3	0	635	26	0.005	1.0
☐		Koh...	< 1	0.967	14	1	0	635	26	0.002	1.0

Figure 43: Clustering models

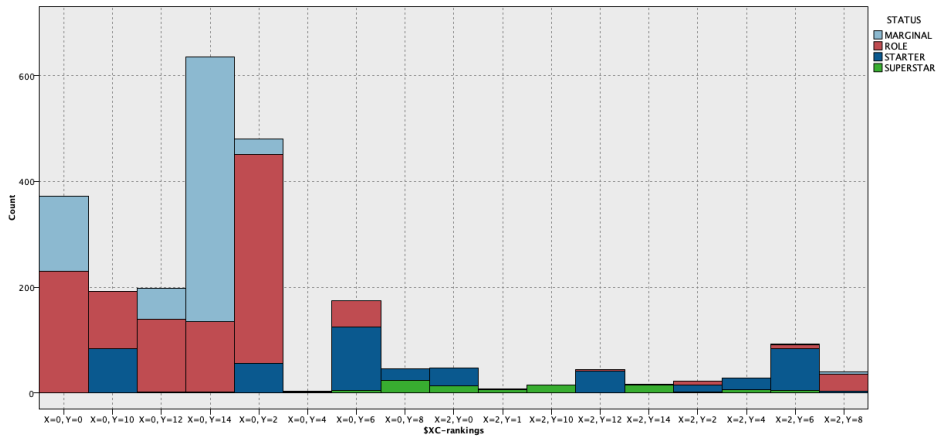


Figure 44: Clustering distributions colored by STATUS for Kohoen map 17

number of possible values. This precaution ensures that the algorithm does not overfit the data by creating a one-to-one mapping.

For instance, the Kohonen map with 17 clusters exhibited the highest silhouette score (0.986); the Figure 44 shows the clusters colored by player status and it is evident that while similar players were grouped closely on the map, this does not provide substantial insights, thus leading to discount these models for further analysis. Specifically, for this model, cluster $X = 2, Y = 10$ exclusively consists of players with an Offensive score of 5 and a Defensive score of 3.

Consequently, I delved deeper into the model with much fewer clusters e.g. 4 clusters TwoStep (silhouette 0.642). The distribution of clusters for this model is depicted in Figure 45. This time, the clustering proved to be more meaningful. Upon sorting the table based on the OD score and scrolling through it, it became apparent that different vertical sections of the table corresponded to different clusters, aligning with the segmentation results. The differentiation is evident in Figure 46, revealing a hierarchical structure. In broad terms, cluster-1 players exhibit higher performance than cluster-2 players, who, in turn, outperform cluster-3 players, and so forth. It's worth emphasizing that the actual SUPERSTARS are somewhat mixed in with the good players, implying that the cluster-1 group at the top is too generalized then.

The next model under consideration was the TwoStep method with 9 cluster, which yielded a commendable silhouette score of 0.909, distribution in Figure 47. Although initially certain clusters seemed to be potentially redundant and candidates for merging, a more thorough examination revealed that they offered valuable insights. This led me to select it as the most favorable clustering outcome.

FP_Mean	FP_SDev	rankings_Mean	rankings_SDev	$XC-rankings	Record_Count
39.928	6.762	87.275	4.512	cluster-5	240
23.586	5.893	76.484	4.214	cluster-6	155
30.187	6.639	80.345	3.907	cluster-3	220
23.310	4.172	76.965	2.816	cluster-1	536
4.476	3.069	68.910	2.843	cluster-2	111
14.286	3.371	73.372	2.244	cluster-4	685
9.018	3.516	70.634	2.504	cluster-7	465

Figure 45: Clustering distributions colored by STATUS for TwoStep 4

FP_Mean	FP_SDev	rankings_Mean	rankings_SDev	$XC-rankings	Record_Count
39.928	6.762	87.275	4.512	cluster-5	240
23.586	5.893	76.484	4.214	cluster-6	155
30.187	6.639	80.345	3.907	cluster-3	220
23.310	4.172	76.965	2.816	cluster-1	536
4.476	3.069	68.910	2.843	cluster-2	111
14.286	3.371	73.372	2.244	cluster-4	685
9.018	3.516	70.634	2.504	cluster-7	465

Figure 46: Statistics aggregated by clusters for TwoStep 4

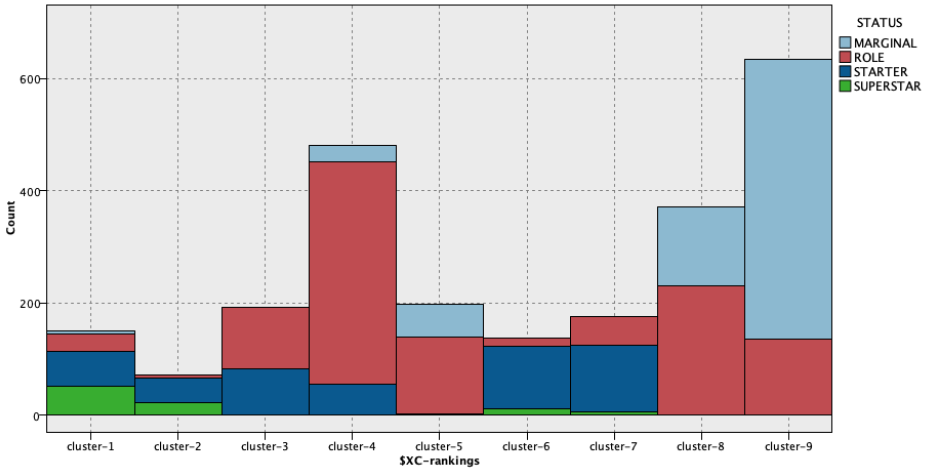


Figure 47: Clustering distributions colored by STATUS for TwoStep 9

FP_Mean	FP_SDev	rankings_Mean	rankings_SDev
41.018	8.849	85.944	5.395
36.345	10.866	85.413	7.181
34.767	5.375	83.370	4.113
29.761	3.159	80.674	3.509
26.286	2.862	78.141	2.194
20.920	3.121	75.815	2.424
15.989	2.504	73.773	2.224
14.048	2.260	72.992	2.187
7.841	3.179	70.420	2.515

OFF_SCORE_Mean	OFF_SCORE_SDev	DEF_SCORE_Mean	DEF_SCORE_SDev	OD_Mean	OD_SDev	$XC-rankings	Record_Count
21.464	8.991	21.865	3.169	3.366	0.659	cluster-2	71
26.671	11.986	12.224	2.506	3.059	0.814	cluster-1	150
22.888	2.599	14.043	4.488	2.977	0.341	cluster-6	138
23.158	2.610	8.772	1.527	2.600	0.000	cluster-7	175
13.942	2.474	13.766	1.519	2.400	0.000	cluster-3	192
13.712	2.727	8.505	1.624	2.000	0.000	cluster-4	481
12.381	2.324	4.759	0.891	1.600	0.000	cluster-5	198
6.961	1.582	7.871	1.496	1.400	0.000	cluster-8	372
4.906	2.395	3.473	1.404	1.000	0.000	cluster-9	635

Figure 48: Statistics aggregated by clusters for TwoStep 9

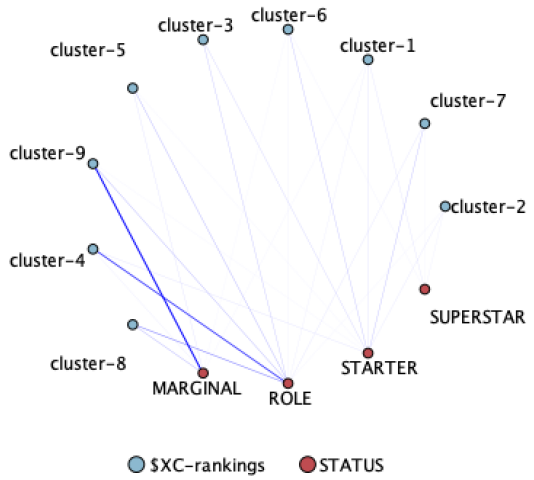


Figure 49: Web association between the clusters and players' STATUS

Interpreting the table in Figure 48.

- Cluster 1: These players are considered superstars, but they tend to focus on scoring and assisting, which may result in more turnovers. For instance, athletes like James Harden, LeBron James, and Stephen Curry fall into this category.
- Cluster 2: This group comprises overall all-stars with high ratings and fantasy points. They excel in the defensive aspect of the game, often contributing with more rebounds and blocks. Typically, they are taller players who cover the roles of center or power forward. Notable examples include Giannis Antetokounmpo, Nikola Jokic, and Joel Embiid.
- Cluster 3: These players are rotational (role players) and perform adequately, particularly excelling in defense.
- Cluster 4: Rotational players who perform at an average level.
- Cluster 5: Rotational players who perform mostly average but struggle in defensive aspects.
- Cluster 6: These individuals show the potential to become stars or starters, and they tend to focus more on defense.
- Cluster 7: Similar to Cluster 6, great players but with a slightly weaker emphasis on defense.
- Cluster 8: Marginal or bottom-tier players.
- Cluster 9: Players in this category are at the bottom tier in terms of performance.

One option to consider is combining paired clusters where one is more defense-oriented. For instance, merging cluster 1 with 2, merging 3 with 4 and 5, and then combining 6 with 7, while leaving clusters 8 and 9 as they are. However, I believe that maintaining the offense-defense distinction provides valuable insights. The cluster divisions, as explained before, and their associations with rankings, Offensive score, Defensive score, and OD tag can be observed in Figure 49, Figure 50, and Figure 51, respectively. The summary of this last model is available in Figure 52 and Figure 53 provides a visual representation of the distribution of clusters for this TwoStep 9 cluster analysis, with each segment representing a different cluster and its relative size within the dataset. The largest cluster, Cluster 1, comprises a significant 26.3% of the data, whereas the smallest cluster, Cluster 9, contains only 2.9%. The two predictors importance is depicted in Figure 54.

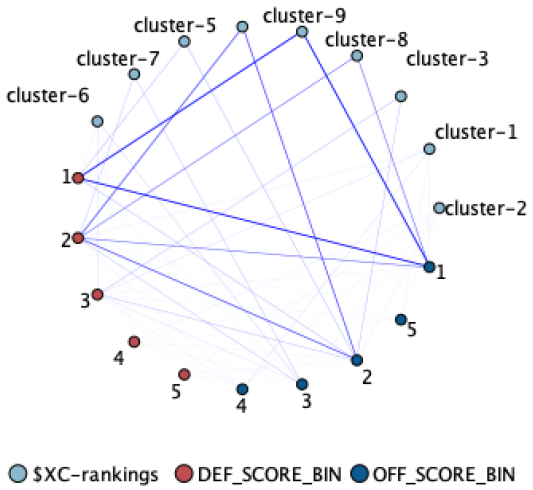


Figure 50: Web association between the clusters and players' OFF_SCORE and DEF_SCORE

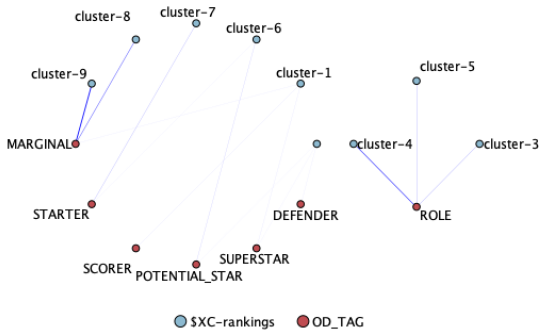


Figure 51: Web association between the clusters and players' OD_TAG

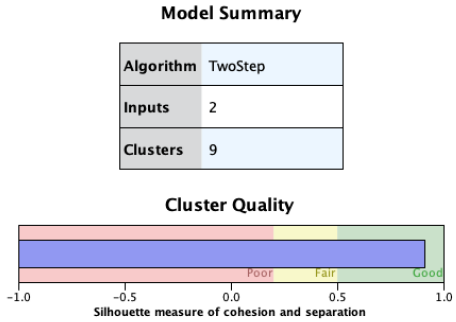
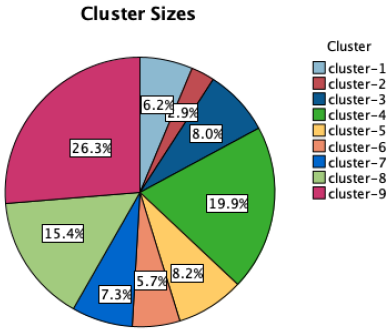


Figure 52: Model summary for TwoStep 9



Size of Smallest Cluster	71 (2.9%)
Size of Largest Cluster	635 (26.3%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	8.94

Figure 53: Cluster sizes for TwoStep 9

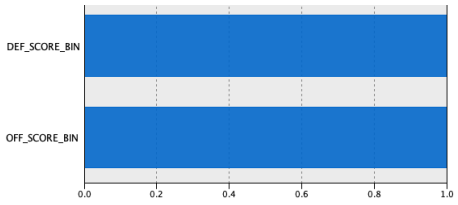


Figure 54: Predictors importance for TwoStep 9

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.944[a]	.891	.890	1.927005

a. Predictors: (Constant), +/−, OREB, FT%, TD3, AGE, STL, 3P%, FG%, GP, FTM, BLK, PF, 3PM, AST, DD2, L, DREB, FGM, TOV, MINUTES, 3PA, FTA, FGA, REB, FP, PTS

Figure 55: Summary for the linear regression model using all the variables

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	67.439	.431		156.358	<.001
	AGE	.021	.010	.015	2.036	.042
	GP	.050	.004	.198	13.117	<.001
	L	−.075	.006	−.178	−11.758	<.001
	MINUTES	−.118	.015	−.179	−7.677	<.001
	PTS	1.827	.574	1.900	3.181	.001
	FGM	−1.265	1.047	−.472	−1.208	.227
	FGA	−.579	.089	−.457	−6.526	<.001
	FG%	−.006	.007	−.010	−.888	.375
	3PM	−1.232	.614	−.172	−2.007	.045
	3PA	.158	.136	.057	1.161	.246
	3P%	−.010	.004	−.025	−2.866	.004
	FTM	−.883	.547	−.213	−1.614	.107
	FTA	−.076	.164	−.023	−.463	.643
	FT%	−.001	.003	−.003	−.293	.770
	OREB	−1.955	.811	−.272	−2.412	.016
	DREB	−1.897	.811	−.605	−2.338	.019
	REB	2.292	.839	.991	2.732	.006
	AST	.518	.419	.163	1.236	.217
	TOV	−.339	.309	−.048	−1.098	.272
	STL	1.008	.840	.073	1.200	.230
	BLK	1.082	.836	.082	1.295	.195
	PF	−.291	.095	−.036	−3.068	.002
	FP	.065	.279	.122	.235	.815
	DD2	−.013	.009	−.022	−1.482	.138
	TD3	−.022	.028	−.006	−.772	.440
	+/−	.027	.021	.014	1.285	.199

Figure 56: Coefficients for the linear regression model using all the variables

4 Modelling: Predictive Analysis

The primary aim of this section is to predict NBA 2K player rankings and comprehend the contributions of predictors to these rankings.

4.1 Linear Regression

As observed in the Descriptive Analysis section, the variables exhibited substantial correlations with the player rankings, suggesting that predicting these rankings might entail a linear problem.

Initially, I trained a model using all available predictors, which explained a significant amount of the variance, as shown by the high R^2 value (0.891) in Figure 55. However, as expected, many of these predictors turned out to be insignificant, indicating the presence of multicollinearity (interactions between variables). This multicollinearity issue is evident in Figure 56, where the coefficients of certain predictors have associated *p-value* exceeding 0.05. Notably, Figure 57 illustrate that the most influential predictors were PTS, REB, and AST.

Subsequently, I embarked on a model-variables selection process by building models for various splits of variables using the Stepwise method. This procedure led to the identification of the best-performing model, which boasted a similar R^2 value (0.890), as depicted in Figure 58. This optimal model incorporated 11 parameters, as outlined in Figure 59, all of which exhibited statistical significance. Intriguingly, the most impactful predictor (Figure 60) was Fantasy Points (FP), which itself can be viewed as a linear regression model with its given coefficients. While our primary goal is accurate ranking prediction, we also aim to comprehend variable interactions and significance.

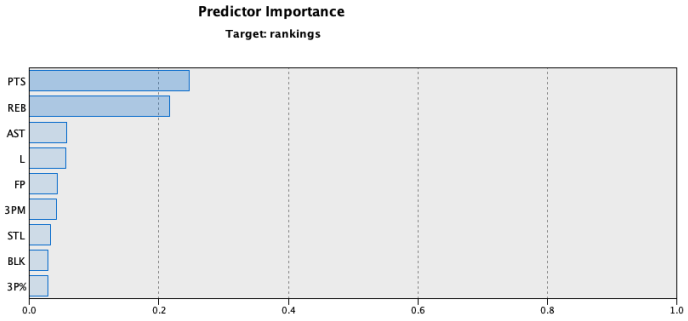


Figure 57: Most important predictors for the linear regression model using all the variables

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.917[a]	.841	.841	2.309414
2	.926[b]	.857	.857	2.190417
3	.931[c]	.868	.867	2.111185
4	.936[d]	.876	.876	2.041638
5	.940[e]	.884	.884	1.977984
6	.941[f]	.886	.886	1.958980
7	.943[g]	.888	.888	1.940378
8	.942[h]	.888	.888	1.940965
9	.943[i]	.889	.888	1.936906
10	.943[j]	.889	.889	1.933576
11	.943[k]	.890	.889	1.930080

a. Predictors: (Constant), FP
b. Predictors: (Constant), FP, +/−
c. Predictors: (Constant), FP, +/−, PTS
d. Predictors: (Constant), FP, +/−, PTS, MINUTES
e. Predictors: (Constant), FP, +/−, PTS, MINUTES, W
f. Predictors: (Constant), FP, +/−, PTS, MINUTES, W, FGA
g. Predictors: (Constant), FP, +/−, PTS, MINUTES, W, FGA, L
h. Predictors: (Constant), FP, PTS, MINUTES, W, FGA, L
i. Predictors: (Constant), FP, PTS, MINUTES, W, FGA, L, PF
j. Predictors: (Constant), FP, PTS, MINUTES, W, FGA, L, PF, 3P%
k. Predictors: (Constant), FP, PTS, MINUTES, W, FGA, L, PF, 3P%, FTA

Figure 58: Summary for the linear regression model selecting the variables with the Stepwise method

11	(Constant)	67.689	.145		466.099	<.001
	FP	.395	.014	.734	27.361	<.001
	PTS	.745	.053	.775	14.164	<.001
	MINUTES	−.111	.014	−.168	−8.208	<.001
	W	.054	.003	.142	17.517	<.001
	FGA	−.488	.056	−.385	−8.721	<.001
	L	−.029	.003	−.070	−8.651	<.001
	PF	−.304	.084	−.038	−3.632	<.001
	3P%	−.012	.003	−.029	−3.696	<.001
	FTA	−.187	.060	−.056	−3.117	.002

Figure 59: Coefficients for the linear regression model selecting the variables with the Stepwise method

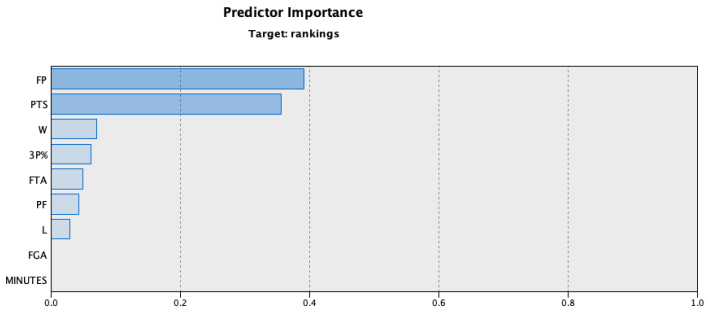


Figure 60: Most important predictors for the linear regression model selecting the variables with the Stepwise method

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.895[a]	.801	.801	2.586551
2	.913[b]	.833	.833	2.371467
3	.925[c]	.856	.856	2.200894
4	.929[d]	.864	.864	2.141197
5	.933[e]	.871	.871	2.085985
6	.936[f]	.875	.875	2.050448
7	.937[g]	.879	.878	2.022478
8	.940[h]	.883	.883	1.984918
9	.941[i]	.886	.885	1.964621
10	.942[j]	.888	.887	1.947405
11	.943[k]	.889	.888	1.939061
12	.943[l]	.889	.888	1.939107
13	.943[m]	.889	.889	1.935394
14	.943[n]	.889	.889	1.932550
15	.943[o]	.890	.889	1.929872
16	.943[p]	.890	.889	1.928640

Figure 61: Summary for the linear regression model selecting the variables with the Stepwise method and excluding out FP

In pursuit of a deeper understanding, I trained a Linear Regression model while excluding FP as a predictor, and the results remained qualitatively consistent, as shown in Figure 61. In this scenario, the best model incorporated 16 parameters (Figure 62). This model provided more insightful variable importance since, unlike the previous cases, there was no variables interactions between variables: the variance that was initially attributed to Fantasy Points is now explained by individual variables. As depicted in Figure 63, the most crucial predictors, in order, include PTS, REB, AST, W, BLK, and STL, aligning well with basketball knowledge. This model is leveraged for further prediction analysis.

The table in Figure 64 presents the results comparing the predictions with actual rankings for two data partitions: training and testing (75%-25% split). The model, as the two previous ones, shows a high degree of correlation between the predicted and actual rankings in both partitions, with a linear correlation of 0.945 for training and 0.94 for testing. These strong correlations indicate that the model can reliably predict rankings based on the input features. The mean errors are close to zero, which suggests that there is no significant bias in the predictions across the dataset; however, the presence of non-zero minimum and maximum errors indicates some variance in prediction accuracy. The mean absolute error values are low (1.411 for training and 1.442 for testing), which further suggests that the model's predictions are very close to the actual values on average.

The standard deviation values are under 2 for both partitions, indicating moderate variability in the prediction errors. The occurrence numbers show how many instances were evaluated in each partition, with 1,770 in training and 642 in testing, demonstrating that a substantial amount of data was used to validate the model's performance.

The prediction quality can also be visualized in Figure 65, illustrating how closely the data points

16	(Constant)	67.189	.284		236.209	<.001
	PTS	1.147	.054	1.193	21.117	<.001
	REB	.431	.030	.187	14.447	<.001
	AST	.609	.048	.192	12.642	<.001
	FGA	-.481	.061	-.380	-7.872	<.001
	BLK	1.304	.135	.098	9.643	<.001
	W	.081	.005	.213	17.069	<.001
	MINUTES	-.116	.014	-.176	-8.146	<.001
	STL	1.239	.147	.090	8.399	<.001
	GP	-.028	.003	-.111	-8.155	<.001
	TOV	-.477	.141	-.067	-3.379	<.001
	3P%	-.012	.003	-.029	-3.668	<.001
	PF	-.276	.092	-.034	-3.010	.003
	FTM	-.219	.081	-.053	-2.713	.007
	AGE	.020	.010	.015	2.016	.044

Figure 62: Coefficients for the linear regression model selecting the variables with the Stepwise method and excluding out FP

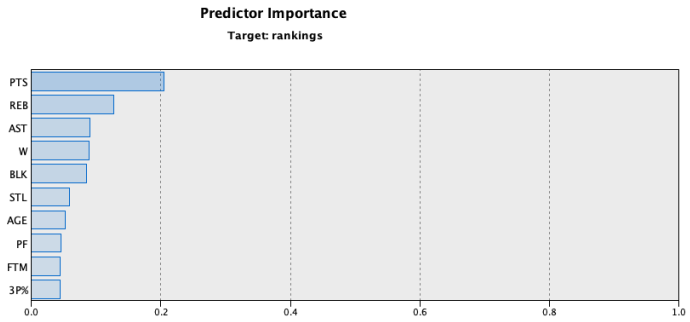


Figure 63: Most important predictors for the linear regression model selecting the variables with the Stepwise method and excluding out FP

Comparing $E-rankings with rankings

'Partition'	1_Training	2_Testing
Minimum Error	-13.374	-8.768
Maximum Error	19.401	12.004
Mean Error	0.005	-0.014
Mean Absolute Error	1.411	1.442
Standard Deviation	1.938	1.883
Linear Correlation	0.945	0.94
Occurrences	1,770	642

Figure 64: Error analysis for the chosen linear regression model

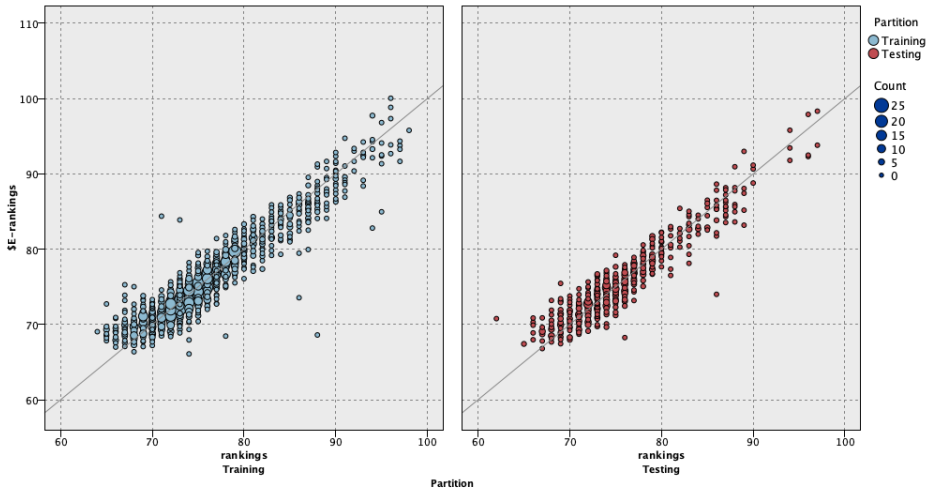


Figure 65: Scatterplot showing the difference between the predictions and the targets for the chosen linear regression model. Both for training (left) and testing (right)

align with the ideal regression line.

Since I already solved a linear regression task with excellent metrics and the target variable is continuous, I choose not to bin the target values and attempt logistic regression. This is because binning the target values can introduce information loss. When a continuous target variable is discretize into categories, it loses the granularity and precision of the original data. Also Linear regression models are often more interpretable than logistic regression models, especially for stakeholders who are not well-versed in machine learning.

While these results are satisfactory, the following subsections will explore various other algorithms and models.

4.2 Decision Tree

Subsequently, I opted to experiment with decision trees as an alternative approach. Specifically, I constructed nodes for four distinct decision tree models: QUEST, C5.0, C&R Tree, and CHAID; using a partition of 70% for the training and 30% for testing. It's worth noting that QUEST and C5.0 are specifically suitable for categorical outputs. Consequently, I employed the STATUS variable, which represents binned rankings, as the target variable for these models. The other decision tree models could accommodate continuous output. For better understanding the relations between individual variables and the rankings the Fantasy Points was not used as an input.

The decision tree generated using the QUEST algorithm proved to be concise and easily interpretable, as illustrated in Figure 66. It efficiently categorized players based on their scoring performance. Players with high scoring were labeled as either SUPERSTAR or STARTER, while those with low scoring were further classified based on additional criteria like FM, W, and Double Double (DD2). MARGINAL players were identified when all checked statistics in a node were lower, while in other cases, players were assigned roles, with the best-case scenario being a STARTER. This model exhibited decent performance, achieving around 75% accuracy and a confusion matrix with a well-populated main diagonal, as displayed in Figure 67.

I also experimented with the C5.0 model, which was more complex and challenging to interpret due to its 75 nodes and 11 levels. In the initial 4 levels, as shown in Figure 68, the primary decision was still based on PTS, with SUPERSTAR players passing additional checks related to DD2. For other cases, criteria such as FGM, +/-, AST, DREB, and OREB were examined to determine player roles. Similar to the QUEST model, this one also achieved decent results, closely resembling the previous model's performance, as seen in Figure 69.

Comparing the predictions of these two classification decision tree models, as shown in Figure 70, they exhibited agreement approximately 82% of the time on the test data. Additionally, when comparing the agreement to the STATUS variable, they predicted correctly of 86% of the time. The

- PTS <= 13.998 [Mode: ROLE] (902)
  - FGM <= 2.187 [Mode: MARGINAL] (367)
    - W <= 34.837 [Mode: MARGINAL] (308)
      - FGM <= 1.404 [Mode: MARGINAL] ⇨ **MARGINAL** (183; 0.896)
      - FGM > 1.404 [Mode: MARGINAL] (125)
        - W <= 24.922 [Mode: MARGINAL] ⇨ **MARGINAL** (88; 0.75)
        - W > 24.922 [Mode: ROLE] ⇨ **ROLE** (37; 0.541)
    - W > 34.837 [Mode: ROLE] ⇨ **ROLE** (59; 0.576)
  - FGM > 2.187 [Mode: ROLE] (535)
    - DD2 <= 13.057 [Mode: ROLE] ⇨ **ROLE** (492; 0.758)
    - DD2 > 13.057 [Mode: STARTER] ⇨ **STARTER** (43; 0.605)
- PTS > 13.998 [Mode: STARTER] (251)
  - PTS <= 23.218 [Mode: STARTER] ⇨ **STARTER** (205; 0.717)
  - PTS > 23.218 [Mode: SUPERSTAR] ⇨ **SUPERSTAR** (46; 0.804)

Figure 66: Resulting decision tree for QUEST

Comparing $C-STATUS with STATUS

'Partition'	1_Training		2_Testing	
Correct	1,370	83.28%	572	74.58%
Wrong	275	16.72%	195	25.42%
Total	1,645		767	

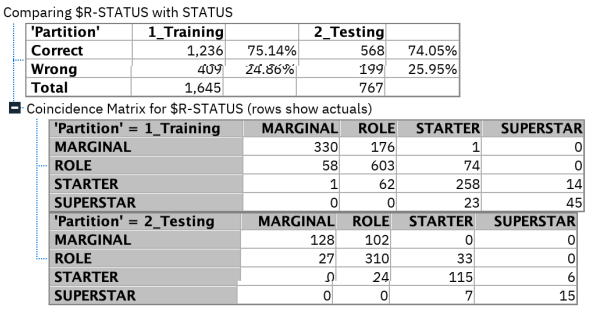
Coincidence Matrix for $C-STATUS (rows show actuals)

'Partition' = 1_Training	MARGINAL	ROLE	STARTER	SUPERSTAR
MARGINAL	456	91	0	0
ROLE	76	621	38	0
STARTER	1	49	281	4
SUPERSTAR	0	1	15	52
'Partition' = 2_Testing	**MARGINAL**	**ROLE**	**STARTER**	**SUPERSTAR**
MARGINAL	155	75	0	0
ROLE	39	302	29	0
STARTER	0	36	100	9
SUPERSTAR	0	0	7	15

Figure 67: Error analysis for QUEST

- PTS <= 10.750 [Mode: ROLE] (1,076)
  - FGM <= 1.850 [Mode: MARGINAL] ⇨ **MARGINAL** (414; 0.831)
  - FGM > 1.850 [Mode: ROLE] (662)
    - OREB <= 2.150 [Mode: ROLE] (619)
      - AST <= 3.050 [Mode: ROLE] (545)
      - AST > 3.050 [Mode: ROLE] ⇨ **ROLE** (74; 0.797)
    - OREB > 2.150 [Mode: ROLE] (43)
      - DREB <= 7.050 [Mode: ROLE] ⇨ **ROLE** (40; 0.8)
      - DREB > 7.050 [Mode: STARTER] ⇨ **STARTER** (3; 1.0)
- PTS > 10.750 [Mode: STARTER] (569)
  - PTS <= 16.150 [Mode: ROLE] (331)
    - DREB <= 5.450 [Mode: ROLE] (271)
      - +/- <= 0.550 [Mode: ROLE] ⇨ **ROLE** (155; 0.794)
      - +/- > 0.550 [Mode: STARTER] (116)
    - DREB > 5.450 [Mode: STARTER] ⇨ **STARTER** (60; 0.917)
  - PTS > 16.150 [Mode: STARTER] (238)
    - PTS <= 22.800 [Mode: STARTER] (166)
      - GP <= 7.500 [Mode: ROLE] ⇨ **ROLE** (4; 0.5)
      - GP > 7.500 [Mode: STARTER] (162)
    - PTS > 22.800 [Mode: SUPERSTAR] (72)
      - DD2 <= 12.500 [Mode: STARTER] (29)
      - DD2 > 12.500 [Mode: SUPERSTAR] ⇨ **SUPERSTAR** (43; 0.93)

Figure 68: Resulting decision tree for C5.0



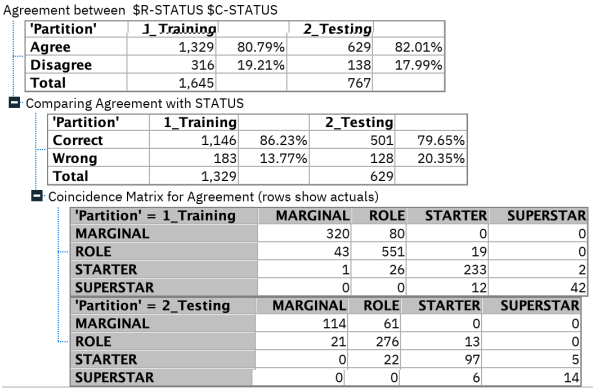
Comparing $R-STATUS with STATUS

'Partition'	1_Training		2_Testing	
Correct	1,236	75.14%	568	74.05%
Wrong	409	24.86%	199	25.95%
Total	1,645		767	

Coincidence Matrix for $R-STATUS (rows show actuals)

'Partition' = 1_Training	MARGINAL	ROLE	STARTER	SUPERSTAR
MARGINAL	330	176	1	0
ROLE	58	603	74	0
STARTER	1	62	258	14
SUPERSTAR	0	0	23	45
'Partition' = 2_Testing	MARGINAL	ROLE	STARTER	SUPERSTAR
MARGINAL	128	102	0	0
ROLE	27	310	33	0
STARTER	0	24	115	6
SUPERSTAR	0	0	7	15

Figure 69: Error analysis for C5.0



Agreement between $R-STATUS $C-STATUS

'Partition'	1_Training		2_Testing	
Agree	1,329	80.79%	629	82.01%
Disagree	316	19.21%	138	17.99%
Total	1,645		767	

Comparing Agreement with STATUS

'Partition'	1_Training		2_Testing	
Correct	1,146	86.23%	501	79.65%
Wrong	183	13.77%	128	20.35%
Total	1,329		629	

Coincidence Matrix for Agreement (rows show actuals)

'Partition' = 1_Training	MARGINAL	ROLE	STARTER	SUPERSTAR
MARGINAL	320	80	0	0
ROLE	43	551	19	0
STARTER	1	26	233	2
SUPERSTAR	0	0	12	42
'Partition' = 2_Testing	MARGINAL	ROLE	STARTER	SUPERSTAR
MARGINAL	114	61	0	0
ROLE	21	276	13	0
STARTER	0	22	97	5
SUPERSTAR	0	0	6	14

Figure 70: Agreements of the two decision tree classifiers

confusion matrices showed elevated values along the diagonal, indicating solid predictive performance.

Moving to the regression aspect, I employed the C&R tree, which generated a decision tree comprising 42 nodes distributed over 5 levels. Figure 71 illustrates the first 4 levels of this tree. Similar to the classification models, this regression model used PTS as the root node and subsequently considered factors such as FM, W, AGE, DREB, and ASSISTS to make predictions. It intuitively indicated that the more nodes a player traversed towards the high end of the tree, the better their predicted ranking would be. The model yielded highly favorable results, as depicted in Figure 72, with a Mean Absolute Error of approximately 2.

For the regression analysis, I also employed the CHAID model, which featured only 4 levels but expanded horizontally to encompass 69 nodes. This model initially categorized potentially based on points, then examined additional statistics like AST, REB, W, and others (Figure 73). Once again, the results were highly acceptable, yielding an absolute error of around 2, as illustrated in Figure 74. Furthermore, when comparing these two regression decision tree models, their agreement was substantial, resulting in improved individual scores and achieving an error rate of approximately 1.7 in the test data, as shown in Figure 75.

In the decision tree models, certain predictors that did not appear as significant in the best linear regression model gained importance. Variables such as FM, +/-, OREB, DREB, and DD2 played substantial roles in these decision trees. This contrast likely arises from the removal of these variables to address multicollinearity issues in the linear regression model. For instance, points are interrelated with Field Goal Made, offensive rebounds (OREB) and defensive rebounds (DREB) with rebounds, while DD2 exhibits correlations with many other statistics by definition.

4.3 Neural Network

The next phase of the Predictive Analysis involved experimenting with Neural Networks. This models can be perceived as black boxes due to their complex and layered structures, making it challenging to understand how they arrive at specific predictions. The intricate interactions among numerous

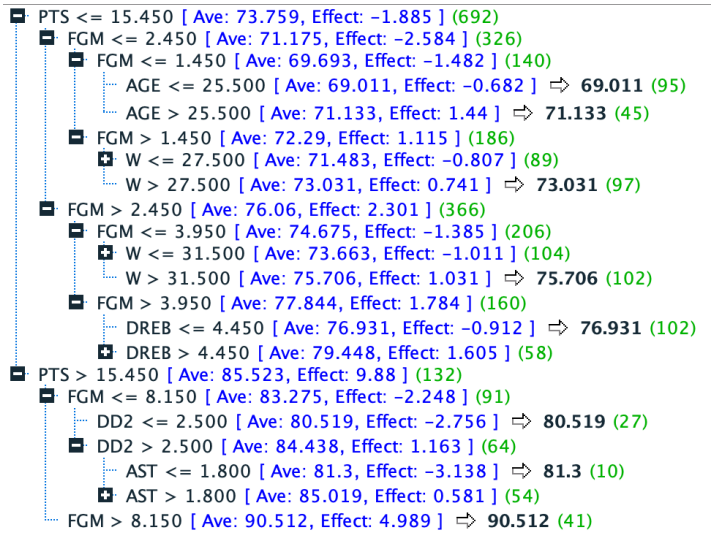


Figure 71: Resulting decision tree for C&R Tree

'Partition'	1_Training	2_Testing
Minimum Error	-9.931	-10.512
Maximum Error	16.867	14.481
Mean Error	0.082	0.063
Mean Absolute Error	1.73	2.055
Standard Deviation	2.32	2.722
Linear Correlation	0.919	0.88
Occurrences	1,145	1,267

Figure 72: Error analysis for C&R Tree

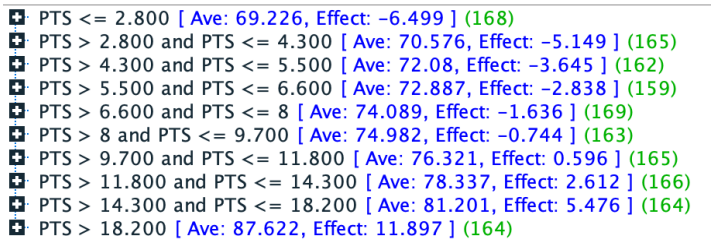


Figure 73: Resulting decision tree for CHAID

'Partition'	1_Training	2_Testing
Minimum Error	-12.091	-11.0
Maximum Error	18.817	12.724
Mean Error	0.0	-0.034
Mean Absolute Error	1.557	1.958
Standard Deviation	2.108	2.593
Linear Correlation	0.934	0.893
Occurrences	1,145	1,267

Figure 74: Error analysis for CHAID

Agreement between $R-rankings $R1-rankings

Comparing Agreement with rankings

'Partition'	1_Training	2_Testing
Minimum Error	-11.46	-8.273
Maximum Error	18.993	11.693
Mean Error	0.055	-0.007
Mean Absolute Error	1.503	1.687
Standard Deviation	2.035	2.236
Linear Correlation	0.94	0.914
Occurrences	1,645	767

Figure 75: Agreements of the two decision tree regressors

Model Summary

Target	rankings
Model	Multilayer Perceptron
Stopping Rule Used	Error cannot be further decreased
Hidden Layer 1 Neurons	13

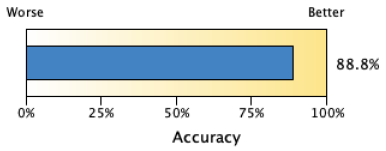


Figure 76: Summary for the neural network

neurons and weights make it difficult to trace the decision-making process and identify the specific features driving their outputs. For this reason I choose to incorporate the Fantasy Points variable into the analysis in an attempt to enhance predictive capabilities. This time the data partition was: 70% for training, 10% for validation and 20% for testing. The model created, as depicted in the summary provided in Figure 76, took the form of a Multilayer Perceptron featuring a single layer and 13 nodes. Notably, the most influential predictors, as evidenced in Figure 77, included points, alongside new variables not present in the previous model, such as Free Throws Made, Double Double, and Triple Double. The model demonstrated remarkable predictive prowess, as illustrated in Figure 78, with predictions exhibiting a very high correlation (0.929 in testing) with the target variable and a mean absolute error of approximately 1.5 during testing. This exceptional predictive performance is also visually evident in the scatterplot depicted in Figure 79.

Despite the complexity and impressive predictive capabilities of neural networks, it's important to

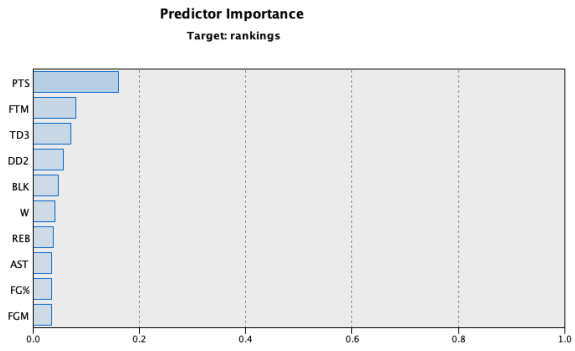


Figure 77: Most important predictors for the neural network

'Partition'	1_Training	2_Testing	3_Validation
Minimum Error	-11.545	-7.931	-7.472
Maximum Error	18.885	12.055	5.563
Mean Error	0.076	0.024	-0.056
Mean Absolute Error	1.454	1.534	1.476
Standard Deviation	1.968	2.068	1.947
Linear Correlation	0.943	0.929	0.93
Occurrences	1,645	530	237

Figure 78: Error analysis for the neural network

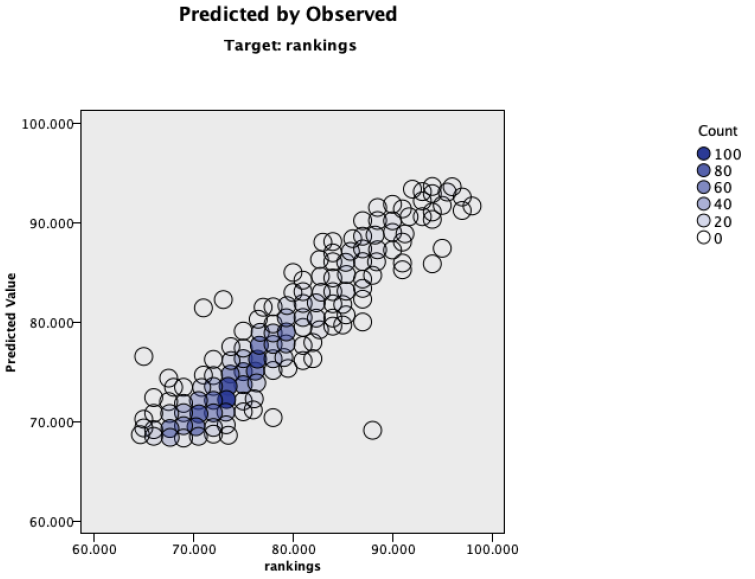


Figure 79: Scatterplot showing the difference between the predictions and the targets for the neural network

Model Information

Target Field	rankings
Model Building Method	Random Trees Regression
Number of Predictors Input	28
Relative Error	0.186
Variance Explained	0.814

Figure 80: Summary for the random tree model

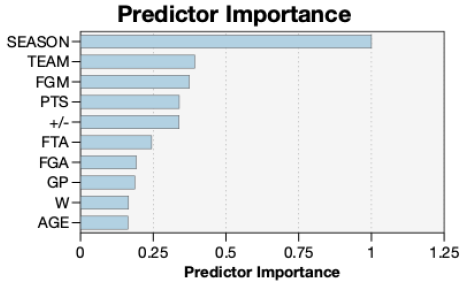


Figure 81: Most important predictors for the random tree model

highlight that the superior metrics continue to favor the Linear Regression model.

4.4 Random Forest

Another quick experiment was done using random forest, the model summary and the predictors importance are in Figure 80 and 81. Once again the results were good (Figure 82) but worst compared to the other models.

4.5 Ensemble

For ensemble techniques I tried first Boosting and Bagging techniques for neural networks. Bagging involves training multiple neural networks independently on bootstrapped subsets of the training data and averaging their predictions, while Boosting sequentially trains neural networks, giving more weight to previously misclassified samples to improve overall model performance. In particular boosting can enanche model accuracy and begging can enanche model stability. But neither of these two techniques showed improved results, as we can see in Figure 83 for boosting and Figure 84 for bagging.

After conducting individual performance assessments for various methods, the next step involved putting all of them together using the *Ensemble* Node. I connected each regression model (excluding the decision trees and the bagged model), and in Figure 85, you can see an evaluation of the final predictions, which represent the mean of predictions from the ensemble technique. When compared to the results presented in Figure 64 for linear regression, it's apparent that these predictions are marginally more accurate in terms of Mean Absolute Error but exhibit a higher Mean Error. The great results can be visualized in Figure 86, where most of the points cluster near the ideal regression line. For both in the linear regression and in the ensemble model, the results are impressive, with the models predicting with an average discrepancy of less than 1.5 ranking points.

'Partition'	1_Training	2_Testing
Minimum Error	-6.053	-9.424
Maximum Error	10.347	11.933
Mean Error	0.025	0.021
Mean Absolute Error	1.221	1.687
Standard Deviation	1.645	2.227
Linear Correlation	0.963	0.915
Occurrences	1,645	767

Figure 82: Error analysis for the random tree model

'Partition'	1_Training	2_Testing	3_Validation
Minimum Error	-2.021	-8.442	-6.606
Maximum Error	2.277	12.844	7.249
Mean Error	0.021	-0.066	-0.091
Mean Absolute Error	0.401	1.588	1.584
Standard Deviation	0.522	2.118	2.012
Linear Correlation	0.996	0.926	0.926
Occurrences	1,645	530	237

Figure 83: Error analysis for the boosted model

'Partition'	1_Training	2_Testing	3_Validation
Minimum Error	-4.085	-8.295	-6.773
Maximum Error	4.463	16.461	8.189
Mean Error	-0.001	-0.053	-0.026
Mean Absolute Error	0.773	2.005	1.821
Standard Deviation	0.996	2.673	2.303
Linear Correlation	0.986	0.882	0.908
Occurrences	1,645	530	237

Figure 84: Error analysis for the bagged model

'Partition'	1_Training	2_Testing	3_Validation
Minimum Error	-5.259	-7.811	-5.618
Maximum Error	7.697	13.549	6.672
Mean Error	0.031	-0.027	-0.047
Mean Absolute Error	0.778	1.442	1.326
Standard Deviation	1.024	1.968	1.71
Linear Correlation	0.986	0.936	0.947
Occurrences	1,645	530	237

Figure 85: Error analysis for the ensemble model

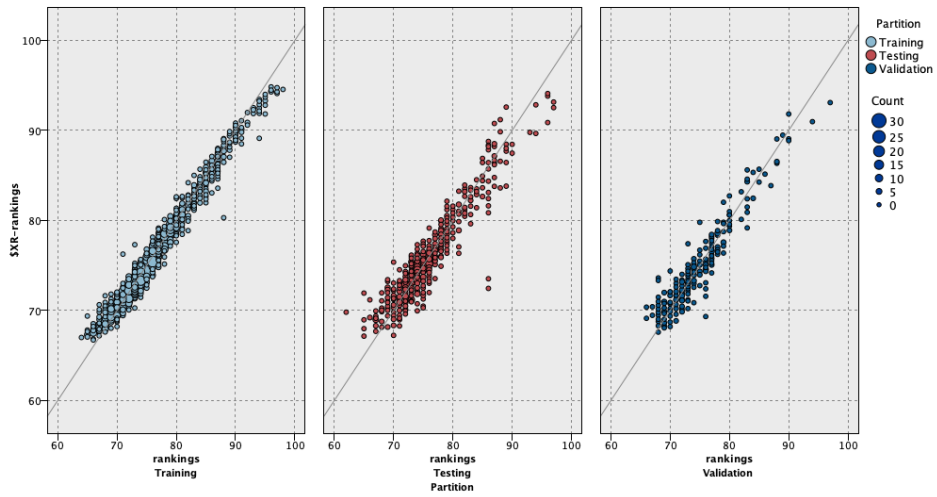


Figure 86: Scatterplot showing the difference between the predictions and the targets for the ensemble model

ID	PLAYER	preds_diff	$E-rankings	rankings
1016....	Gordon Hayward	376.386	68.599	88.000
655.0...	JaKarr Sampson	178.856	84.374	71.000
2324....	Paul George	155.126	73.545	86.000
1899....	Lance Stephenson	144.089	73.996	86.000
1101....	Kawhi Leonard	125.571	82.794	94.000
1142....	MarShon Brooks	118.213	83.873	73.000
440.0...	Stephen Curry	100.875	84.956	95.000
872.0...	Aaron Jackson	91.245	68.448	78.000
2355....	Sean Kilpatrick	76.870	70.768	62.000
938.0...	Damion Lee	68.374	75.269	67.000
2238....	Julius Randle	62.700	66.082	74.000
1650....	Treveon Graham	60.066	68.250	76.000
1135....	Luke Kornet	59.591	72.720	65.000
1060....	Jayson Tatum	49.807	79.943	87.000
1519....	Kyle O'Quinn	49.129	75.009	68.000
231.0...	Jeremiah Martin	47.213	73.871	67.000
606.0...	Draymond Green	42.782	79.459	86.000
1487....	Josh Huestis	41.547	75.446	69.000
2295....	Michael Beasley	38.256	73.185	67.000
195.0...	J.R. Smith	36.779	67.935	74.000

Figure 87: Players with least accurate predictions: largest difference between the predicted predictions and the target ranking

5 Evaluation

5.1 Final Model Selection

For the final model selection, I would opt for the linear regression approach employing stepwise methods for predictor selection, excluding the aggregated Fantasy Points statistic. Given the similar predictive quality to the ensemble model, a more efficient and less resource-intensive option is preferable. Linear regression is notably straightforward and interpretable, ideal for linear relationships and smaller datasets, which happed to be the case, with lower computational demands and reduced overfitting risk compared to more complex models.

5.2 Error Analysis

Let's delve into the most significant discrepancies and explore why they occurred. Figure 87 shows a list of players sorted by the largest squared differences between the predicted value and the target. At the top of the list is Gordon Hayward during the 2017-18 season, a player with low statistics but a rating of 88. This may seem puzzling, but it can be attributed to the fact that rankings are assigned at the start of the season. This player case: *"Hayward* [5] *played seven seasons with the Jazz, and was selected to the 2017 NBA All-Star Game. In the 2017 off-season, Hayward signed as a free agent with the Celtics, but was ruled out for the remainder of the 2017–18 season after suffering a fractured tibia and dislocated ankle only five minutes into the season opener."*
This situation highlights the potential of using rankings to detect injuries. As observed in Figure 88 and Figure 89 the predominant factor contributing to both player underrating and overrating appears to be injuries. A significant number of data points cluster around a limited number of games played, making it challenging to accurately understand their in-game performance and subsequently predict their capabilities in the video game.

Additionally, the disparities between our predictions and the actual 2K rankings serve as a valuable tool for identifying players who might be underperforming compared to their assigned ratings. This insight can be applied broadly to detect instances where a player's on-court performance falls short of the expectations set by their rating in the game. On the flip side, this approach is equally useful for pinpointing underrated players who are surpassing expectations with their performance on the court.

When talking about overrated players (Figure 90) we can see some cases of rookies. These players are in their first year in the league, and their rankings are assigned before the season starts, while I am using the real statistics at the end of that season. This situation sheds light on how often the expectations for young players frequently deviate slightly from the actual outcomes. Furthermore, there are instances of overrated players who happen to be superstars. Their impact on the game can

[5] https://en.wikipedia.org/wiki/Gordon_Hayward

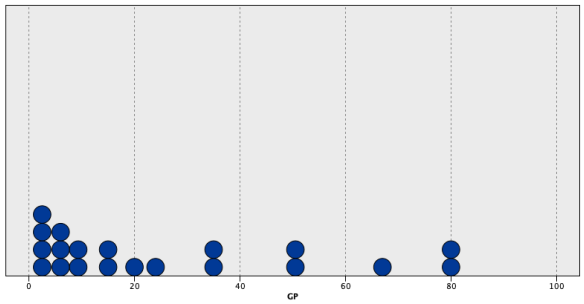


Figure 88: GP distribution for underrated players

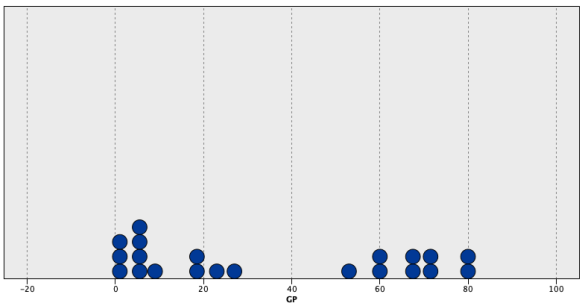


Figure 89: GP distribution for overrated players

be challenging to quantify solely through statistics, but it becomes evident when you watch them play. Take, for example, two-time NBA champion Kawhi Leonard, renowned for his exceptional all-around skills, particularly on the defensive end. Although his defensive prowess may not always manifest in traditional statistics like steals, blocks, or rebounds, he has a consistent ability to neutralize the players he guards. Similarly, players like Kyrie Irving are offensive powerhouses, characterized by extraordinary ball-handling skills and a knack for scoring in spectacular fashion on multiple occasions.

Upon examining the underrated players in Figure 91, it becomes evident that, aside from the discussed injured players, these individuals primarily fall into the categories of marginal and role players. Therefore, discerning the specific reasons behind these disparities in their rankings proves to be quite challenging. Nevertheless, it can offer valuable insights for a team aiming to complete its roster with a constrained budget.

As for players who achieve a perfect match with their rankings, you can find a comprehensive list in Figure 92. These players deliver in-game performances that align precisely with the game's expectations. Additionally, the top superstar players consistently received correct high rankings, as illustrated in Figure 93.

The distribution of squared errors is depicted in Figure 94, revealing that nearly all instances exhibit a discrepancy of within 5 squared ranking points. This outcome is highly favorable and indicative of the model's strong performance.

ID	PLAYER	GP	preds_diff
35.000	Blake Griffin	18....	4.887
65.000	Chandler Parsons	5.0...	4.671
195.0...	J.R. Smith	6.0...	6.065
440.0...	Stephen Curry	5.0...	10.044
487.0...	Victor Oladipo	19....	4.518
606.0...	Draymond Green	66....	6.541
709.0...	Kawhi Leonard	60....	5.349
752.0...	Marcus Smart	80....	4.769
872.0...	Aaron Jackson	1.0...	9.552
1014....	Glenn Robinson III	23....	4.607
1016....	Gordon Hayward	1.0...	19.401
1058....	Jaylen Brown	70....	4.690
1060....	Jayson Tatum	80....	7.057
1101....	Kawhi Leonard	9.0...	11.206
1114....	Klay Thompson	73....	4.915
1123....	Kyrie Irving	60....	4.621
1650....	Treveon Graham	27....	7.750
1897....	Kyrie Irving	53....	5.821
1899....	Lance Stephenson	69....	12.004
2238....	Julius Randle	1.0...	7.918

Figure 90: Overrated players

ID	PLAYER	GP	preds_diff
160.0...	Frank Mason	9.0...	-4.839
231.0...	Jeremiah Martin	9.0...	-6.871
370.0...	Mychal Mulder	7.0...	-5.184
655.0...	JaKarr Sampson	4.0...	-13.374
938.0...	Damion Lee	15....	-8.269
1135....	Luke Kornet	20....	-7.720
1142....	MarShon Brooks	7.0...	-10.873
1197....	RJ Hunter	5.0...	-4.785
1487....	Josh Huestis	2.0...	-6.446
1519....	Kyle O'Quinn	79....	-7.009
1964....	RJ Hunter	36....	-4.638
1966....	Rakeem Chris...	1.0...	-4.800
2198....	James Michae...	15....	-4.869
2222....	Joel Anthony	49....	-4.920
2227....	Johnny O'Brya...	34....	-5.036
2273....	Luc Mbah a M...	67....	-4.757
2295....	Michael Beasley	24....	-6.185
2354....	Ryan Kelly	52....	-4.849
2355....	Sean Kilpatrick	4.0...	-8.768
2368....	Steve Blake	81....	-4.684

Figure 91: Underrated players

1064....	Jeremy Lamb	0.000	78.021	78.000
823.0...	Solomon Hill	0.000	69.980	70.000
1840....	Jeremy Evans	0.000	70.020	70.000
1670....	Wayne Selden	0.000	69.981	70.000
1733....	Cameron Payne	0.000	72.982	73.000
760.0...	Maxi Kleber	0.000	74.015	74.000
1864....	Jose Calderon	0.000	74.015	74.000
1962....	Paul Pierce	0.000	72.987	73.000
1619....	Serge Ibaka	0.000	79.988	80.000
97.000	Darius Garland	0.000	73.989	74.000
311.0...	LaMarcus Aldridge	0.000	84.010	84.000
1694....	Allen Crabbe	0.000	74.990	75.000
964.0...	Dennis Smith Jr.	0.000	77.991	78.000
1248....	Tobias Harris	0.000	81.991	82.000
1329....	Brandon Knight	0.000	74.008	74.000
1436....	JJ Redick	0.000	78.994	79.000
119.0...	Derrick Walton Jr.	0.000	68.994	69.000
1243....	Thon Maker	0.000	72.005	72.000
1011....	Georges Niang	0.000	69.001	69.000
1602....	Richaun Holmes	0.000	75.999	76.000

Figure 92: Players with most accurate predictions: smallest difference between the predicted value and the target ranking

ID	PLAYER	preds_diff	$E-rankings	rankings
316.0...	LeBron James	10.244	93.799	97.000
1904....	LeBron James	12.372	92.483	96.000
212.0...	James Harden	7.984	98.826	96.000
1507....	Kevin Durant	4.647	93.844	96.000
717.0...	Kevin Durant	13.732	92.294	96.000
1049....	James Harden	1.768	97.330	96.000
279.0...	Kawhi Leonard	11.018	92.681	96.000
636.0...	Giannis Antetokounmpo	3.577	97.891	96.000
667.0...	James Harden	16.306	100.038	96.000
1499....	Kawhi Leonard	13.393	91.340	95.000
23.000	Anthony Davis	0.830	94.089	95.000
1231....	Stephen Curry	6.197	92.511	95.000
440.0...	Stephen Curry	100.875	84.956	95.000
1449....	James Harden	3.245	96.801	95.000
826.0...	Stephen Curry	6.997	92.355	95.000
86.000	Damian Lillard	0.321	93.433	94.000
2003....	Stephen Curry	13.621	97.691	94.000
2271....	LeBron James	4.911	91.784	94.000
1101....	Kawhi Leonard	125.571	82.794	94.000
1629....	Stephen Curry	3.500	92.129	94.000
1613....	Russell Westbrook	14.215	97.770	94.000
322.0...	Luka Doncic	0.079	94.281	94.000
892.0...	Anthony Davis	3.198	95.788	94.000
525.0...	Anthony Davis	0.288	94.537	94.000
1311....	Anthony Davis	0.310	93.443	94.000
1013....	Giannis Antetokounmpo	0.765	93.125	94.000
793.0...	Paul George	0.239	92.511	93.000
1123....	Kyrie Irving	21.351	88.379	93.000
1987....	Russell Westbrook	0.391	92.374	93.000
2367....	Stephen Curry	0.677	92.177	93.000
1876....	Kawhi Leonard	14.906	89.139	93.000
1883....	Kevin Durant	1.498	94.224	93.000
1215....	Russell Westbrook	0.024	92.845	93.000
2195....	James Harden	1.866	93.366	92.000

Figure 93: Difference between the predicted value and the target ranking for the best players

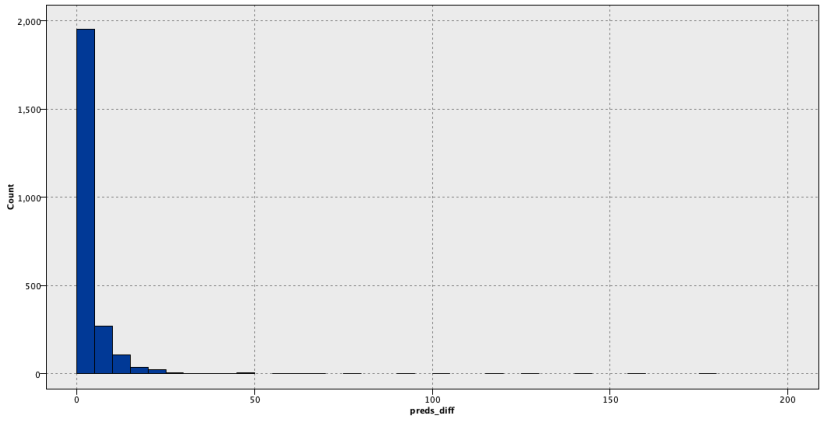


Figure 94: Difference between the predicted value and the target ranking distribution

5.3 Business Goals

- **Talent Scouting and Analytics Tool**: I was able to provide a somehow valuable insights into player performance. By identifying the reasons for a player's underperformance or overperformance through the analysis of prediction discrepancies, scouts and coaches can better understand a player's true potential and areas for improvement. This data-driven approach allows teams to make more informed decisions when evaluating prospects and assessing existing talent, ultimately enhancing their ability to build well-rounded and successful rosters.

- **Enhancing Player Engagement and Game Development**: I understood the role of each variable impacting the 2K rankings through correlations and ANOVA analysis and this can have a significant impact on enhancing player engagement and game development. By pinpointing which in-game statistics most strongly influence player rankings, game developers can fine-tune the gameplay experience to align with player expectations, making the game more immersive and enjoyable. Additionally, this insight can help create more accurate rankings system, leading to increased player engagement and satisfaction, ultimately driving the success and longevity of the game.

- **Betting Insight**: Adjust the betting odds for specific events, like player performance, based on the importance of relevant statistics. For instance, I concluded that points scored (PTS) is a highly influential statistic, consider modifying odds related to a player's scoring performance.

- **Data-driven Salary Assignment**: Player salaries can be determined through a function of the calculated player rankings. For example:

$$salary = base\_salary + (player\_ranking \times salary\_multiplier) \tag{6}$$

  where $base\_salary \sim 500k$ and $salary\_multiplier \in [10k, 30k]$.
  Teams have the flexibility to offer lower salaries to players who are underperforming. Similarly, players whose talent is not yet well-recognized can also receive lower salaries, representing a potential bargain for the team.

6 Conclusion

Throughout this project, I embarked on a comprehensive journey in the realm of Data Mining. It all began with an understanding of the business goals and then traversed through various stages of data exploration, encompassing diverse graphs and statistical analyses. The exploration of relationships within the dataset, affecting player rankings, was conducted using various methodologies like correlations and ANOVAs. A notable success was achieved through the somehow innovative Offense-Defense Analysis, which was an adaptations of the RFM analysis for this task, resulting in meaningful clusters.

However, the most promising part of this endeavor was the predictive aspect. I achieved remarkable results by accurately aligning 2K ratings with a simple linear regression model, recognizing that the 2K company likely employs far more complex machine learning models for their assessments.

In conclusion, this project was enjoyable, particularly due to my passion for the NBA. Delving into the analytical characteristics of players offered a unique perspective and a rewarding experience, ultimately broadening my understanding of the game and its statistical intricacies.