Final Report

NBA 2K

Lucas Tetrault *(Author)*
IUPUI
Indianapolis, IN

Jay Dawson *(Author)*
IUPUI
Indianapolis, IN

***Abstract*—This electronic document is a "live" template. The various components of your paper [title, text, heads, etc.] are already defined on the style sheet, as illustrated by the portions given in this document. DO NOT USE SPECIAL CHARACTERS, SYMBOLS, OR MATH IN YOUR TITLE OR ABSTRACT.** *(Abstract)*

***Index Terms*—Component, formatting, style, styling, insert.** *(key words)*

I. Objective

The objective of this data mining project is to clean, process, and analyze the NBA 2K roster and player statistic datasets obtained from Kaggle. Our aim is to derive insights into players' attributes and connections to enhance the realism and immersion of the gaming experience provided by NBA 2K. Specifically, we intend to identify patterns, trends, and discrepancies within the dataset to gain a deeper understanding of player characteristics and the distribution of ratings. This analysis will encompass various player attributes, including but not limited to their physical attributes, skill ratings, and performance statistics within the game. Additionally, we will explore game statistics such as points per game by age and points per game by team. We aim to delve into the realm of game statistics to further enrich our analysis. Specifically, we will examine the relationship between players' ages and their respective scoring performance, shedding light on how age may influence player effectiveness on the court. Furthermore, we will investigate the distribution of scoring points per game across different teams to discern patterns and variations in team performance. This exploration will allow us to understand the strengths and weaknesses of individual teams, as well as identify any disparities or trends that may exist across the league.

II. Data

The dataset appears pretty well-organized and usable for the most part. We did encounter some duplicates towards the end, originating from other games. Initially, we considered removing them, but they ended up being useful for comparing players across different years of the game. Additionally, we found some blank cells that needed to be filled in. During the cleaning process, we identified redundant attributes, which we removed to streamline our dataset. Other attributes required refinement and simplification. After these adjustments, we believe the dataset is up to standard and meets our needs for this assignment.

We first replaced the blank cells in the "college" column with either "International" if they are from overseas or "High School" if they came straight to the league without attending any university. Then we decided to rename the attribute to "education" because it fit the description better. After that we decided to delete the "draft_round" column as a whole because it seemed like unnecessary noise and redundant since we already have a "draft_pick" column (originally draft_peak before we fixed it). Lastly, we filled in all the blank cells in the team column with "unsigned" because they are not signed to any NBA teams.

III. Methodology

With our dataset now cleaned and prepared, we're ready to dive into analysis using a combination of Excel and R. We'll employ various methods to uncover patterns and correlations among players. Visualization will play a key role, utilizing graphs, charts, tables, and trees to present our findings. We'll use scatter plots with regression lines, a technique we've recently learned in class, which we believe will be particularly useful for this analysis. These visualizations will help us understand relationships between different player attributes and identify any trends within the data. Additionally, we'll leverage a range of functions to compute and execute different tasks efficiently.

IV. Hypothesis

Our null hypothesis suggests that there is not sufficient evidence to support a correlation between a player's salary and their overall rating. Conversely, our alternative hypothesis proposes that there is indeed a correlation between a player's salary and their overall rating. We can apply this formula to test many theories, but our primary focus at the moment is on this specific relationship.

A. Additional Hypotheses

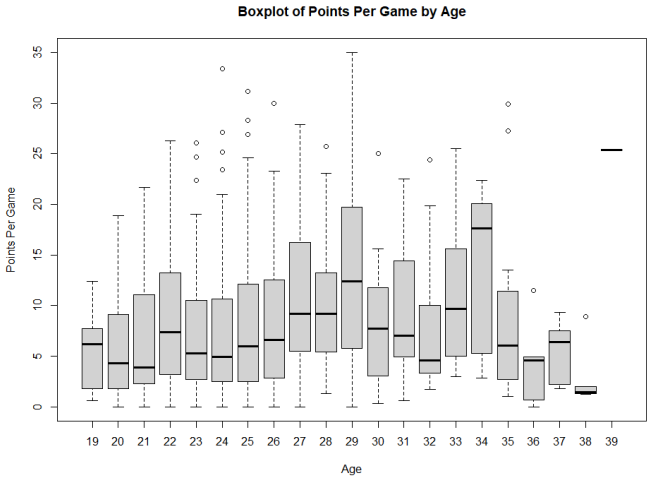
Null hypothesis: There is no significant difference in the average points per game across different age groups in the NBA. Alternative hypothesis: There is a significant difference in the average points per game across different age groups in the NBA. Could also use this same format to compare scoring between the top 8 teams in the league.

DESCRIPTIVE STATISTICS OF DATA OBJECTS

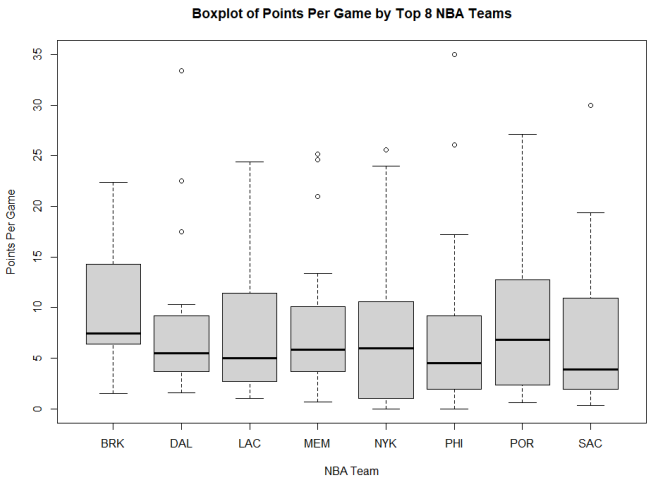
```
 full_name            rating          jersey             team
Length:464         Min.   :67.00   Length:464         Length:464
Class :character   1st Qu.:73.00   Class :character   Class :character
Mode  :character   Median :76.00   Mode  :character   Mode  :character
                   Mean   :77.57
                   3rd Qu.:80.00
                   Max.   :97.00
  position           birthday            height           weight
Length:464         Length:464         Length:464         Length:464
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character


   salary            country           draft_year    draft_pick
Length:464         Length:464         Min.   :2001   Length:464
Class :character   Class :character   1st Qu.:2011   Class :character
Mode  :character   Mode  :character   Median :2015   Mode  :character
                                      Mean   :2014
                                      3rd Qu.:2017
                                      Max.   :2019
 education          version
Length:464         Length:464
Class :character   Class :character
Mode  :character   Mode  :character
```

This scatter plot illustrates a pretty clear relationship between player rating and salary. The positive slope of the regression line indicates that as player ratings increase, so do their salaries. This makes intuitive sense, as higher-rated players are typically more valuable to teams and therefore command higher salaries. The trendline provides an average representation of this relationship, showing that, on average, as a player's rating goes up, their salary tends to increase accordingly. However, there is still variability in salaries for players with the same rating, which could be influenced by factors such as contract negotiations, market demand, or individual player characteristics beyond their in-game ratings. There are some outliers present, where players have unexpectedly high or low salaries for their rating, possibly due to endorsement deals, team dynamics, or other unique circumstances. Overall, the scatterplot suggests that player ratings are generally correlated with salaries in NBA 2K, with higher-rated players earning higher salaries on average, though it is not a super close fit with a lot of variability as well.

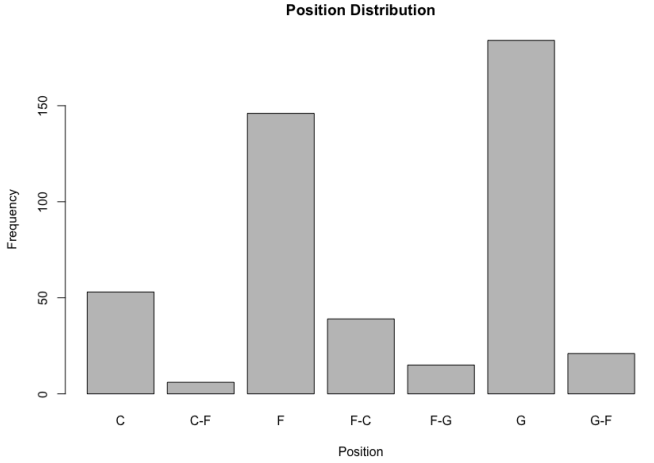


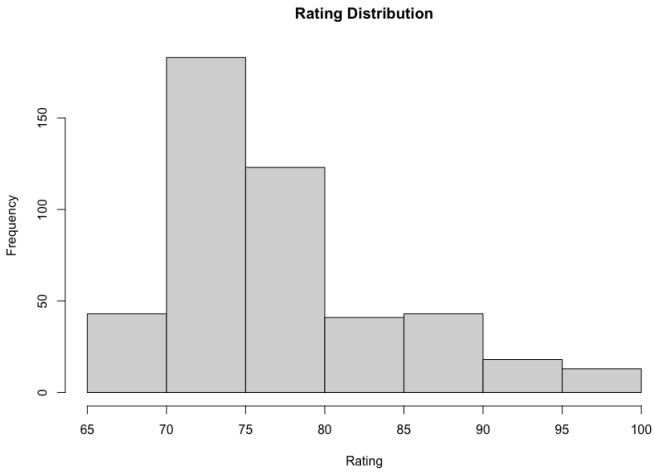
This box plot reveals several key insights about each NBA player's points per game average based on their age. Longer boxes such as ages 29 and 34 suggest greater variability in PPG within those age groups. The whiskers extending from the box indicate the range of the data, excluding outliers, which are plotted individually beyond the whiskers. Players like Luka Doncic who are still very young and are around the top of the league in scoring stand out in this graph. Overall, the plot allows us to observe the trend of PPG across different age groups, typically peaking during a player's prime years and then declining as they get older. One outcast in this case is LeBron James, who is 39 and still averaging 25 PPG. Comparing these medians and the spread of PPG between age groups provides insights into how player performance changes with age. Outliers beyond the whiskers represent players who perform significantly better or worse compared to the majority of their age group, highlighting exceptional performers or players experiencing significant decline. This distribution gives a sense of the typical career trajectory of NBA players in terms of scoring performance as they age.

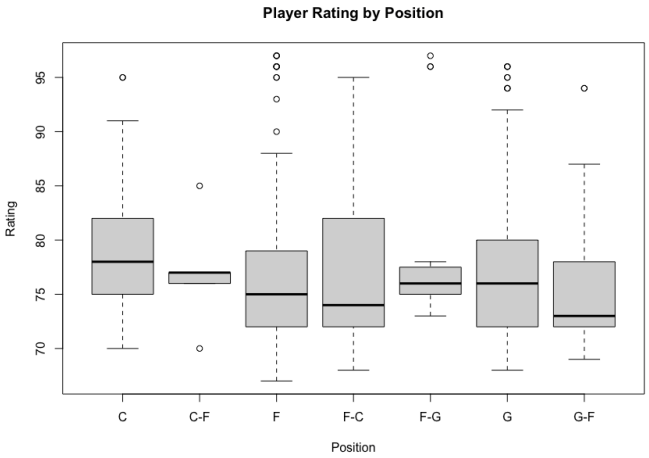


This box plot is similar to the previous one, but it is focusing on each player's points per game average that play

for the top 8 teams in the NBA. There is a lot less variability in these boxes compared to the other box plot which suggests that these teams are more evenly distributed compared to the entire league for each age group. Comparing the box plots of different teams allows us to observe differences in scoring distribution among the top teams. Teams with higher median PPG and wider IQRs may have more balanced scoring attacks, with several players contributing consistently. On the other hand, teams with lower median PPG and narrower IQRs may rely heavily on a few high-scoring players. Teams like the Dallas Mavericks and Philadelphia 76ers for example have the least overall variability and look like the lowest performers before you see some of the massive outliers on their teams like Joel Embiid and Tyrese Maxey on the 76ers and Luka Doncic for the Mavericks who tend to carry the scoring load for their respective teams. Overall, the box plot provides a snapshot of how scoring is distributed among players on the top teams in the NBA, offering insights into each team's offensive dynamics and the roles played by different players in contributing to their team's success.



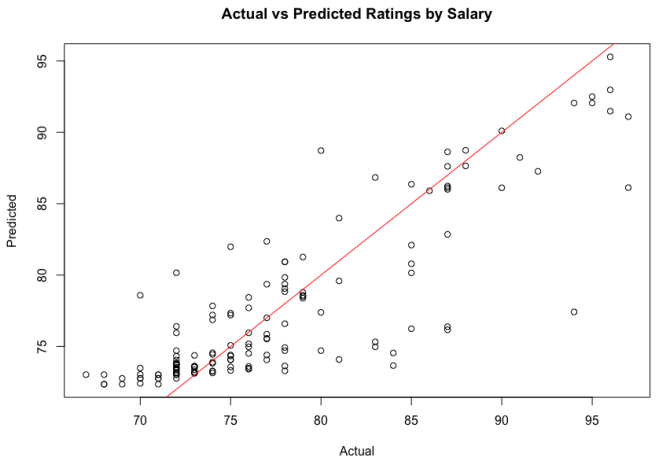




We used three graphs in this case to look at the position distribution as well as the spread of overall ratings for NBA players and then compare them using another box plot. The first graph tells us that most players are guards or forwards, while centers and other hybrid positions are a lot less common. The second graph shows that the most common NBA player rating is between 70 and 75, with a decent amount between 75 and 80 as well. The rest of the player ratings are pretty evenly distributed throughout the league. We then used the box plot above to combine the two graphs and see how they relate to each other. It looks like the center position is usually rated the highest, with most of them being rated at least a 75 overall and the highest median of them all. The forward-center hybrid on the other hand has the most variability and one of the lowest medians. Guards and forwards look pretty similar with the most outliers that represent the superstars of the league.

Top Ten Highest Overall Ratings in NBA 2K

	full_name	rating
103	LeBron James	97
113	LeBron James	97
183	Giannis Antetokounmpo	97
338	Kawhi Leonard	97
12	James Harden	96
13	James Harden	96
46	Stephen Curry	96
114	LeBron James	96
127	Giannis Antetokounmpo	96
188	Giannis Antetokounmpo	96



This scatter plot compares NBA player's actual ratings and their predicted ratings based on their salary. There is a positive relationship between the two and we found the Mean Squared Error (MSE) to be around 15 which is not too high, suggesting that the predictions were generally pretty good and not very far off the actual ratings. There are still a decent amount of outliers, especially near the bottom right corner of the graph. These points represent highly rated players that are technically underpaid based on the salary rate of the league overall. There are also a decent amount of players that are well over the regression line, suggesting that they are overpaid and predicted to be rated higher than they actually are based on how much they are being paid to play. There are definitely some players that significantly outperform or underperform based on their contracts while in the game. Overall, the trend suggests that some higher-paid players tend to have lower actual ratings in the game than predicted and vice versa, indicating that salary alone does not entirely determine a player's in-game performance rating.

Conclusion

Despite facing challenges during the coding and visualization process in R, we are pleased with our analysis and overall findings. Our main focus was on exploring the relationship between players' overall ratings, salary levels, and positions on the court in NBA 2K20. The final graph revealed that lower-rated players are often overpaid but show significant variability, while most top players are technically underpaid. Although only a few players perfectly fit this trend, our main alternative hypothesis was largely confirmed. Additionally, we tested several supplemental theories that furthered our understanding of trends and patterns throughout the NBA. Our project aimed to enhance the realism and engagement of the NBA 2K gaming experience by analyzing the game's dataset alongside external player performance data to gain deeper insights into the league. Through rigorous data cleaning and analysis, we explored various factors including players' skills, age, points scored per game, and identified intriguing trends and disparities between teams. Our analysis provided insights into how players' ages impact their performance and what factors contribute to team success within the game. In summary, our work contributes to enriching the NBA 2K gaming experience by providing players with a more authentic and enjoyable virtual basketball experience. By uncovering and understanding the underlying patterns and dynamics of player performance and team dynamics, we aim to make the gaming experience more immersive and reflective of real-world basketball.

References

[1] Isaienkov, K. (2022, December 25). NBA 2K20 Player dataset. Kaggle.

[2] Chung, B. W. (2024, January 26). NBA player stats dataset for the 2023-2024. Kaggle.