**Data Overview:**

The use of big data has become the basketball industry’s best asset in recent years when analyzing player performance. Historically, box scores were kept to track player and team performance with limited statistical analysis. Today, player performance metrics dominate the basketball industry. Everything from on-court decision making to post-season performance evaluation are affected by these player metrics.

We chose to use the Historical NBA Raptor dataset created by FiveThirtyEight. This dataset contains individual player stats from 1977 – 2022 and contains several player performance metrics that can’t be found in a traditional box score. These player performance metrics include WAR, RAPTOR, PREDATOR, and Pace Impact, with the first three being measured offensively, defensively, and total – a combination of the two. WAR (wins above replacement) measures how many additional wins a player is worth over the course of a season compared to an average performing player at their position. This metric is often used to measure a player’s overall value to their team. RAPTOR (Robust Algorithm Player Tracking On/Off Ratings) measures the number of points a player contributes to his team per 100 possessions, as compared to a league average player. The RAPTOR calculation accounts for on-off data which measures the performance of a team when a specific player is on the court vs not playing. The PREDATOR metric measures the predicted points above average per 100 possessions. Essentially it is a prediction based on the RAPTOR metric data. Lastly Pace Impact measures the impact a player has on team possessions throughout the course of a game (48 minutes).

Other variables included in the dataset are player name, player ID, year of season, season type, NBA team, minutes played, and number of possessions. These other variables are data that can be found in traditional box scores and other widely accessible sources. We will use them here to supplement the player performance metrics and derive further insights about specific players, teams, and seasons.

After we familiarized ourselves with the dataset, we made the decisions to both supplement the data with additional variables and filter the data to focus on specific rows. We added the variables season\_count, career\_ranking, made\_playoffs, warrank, and top\_player to aid in the exploratory and explanatory visualizations that we created.

We also decided to filter out any rows where the value for minutes played (mp) was less than 200. The dataset includes a row for all players who played at least one minute in an NBA season from 1977 to 2022. There is a large number of players who played very few minutes in a given season. With a smaller sample size of playing time, the resulting values for the other metrics become less reliable. This is evident when looking at the distribution of values for raptor\_total before and after the filter is applied. After the filter is applied, the range of values shrinks quite drastically, and some outliers are removed.

Our goal is to use these player performance metrics to create visuals that discover both player and team trends over time. The two main focuses of our analysis will be the evaluation and comparison of players and teams, with the goal of tying them together visually.

**1)Correlation Matrix for Numerical Variables: (Visualization #1):**

A grid with blue dots and red text

Description automatically generated

**2)Box Plots (Minutes Played): (Visualization #2):**

As we mentioned earlier in this report, we made the decision to filter out rows where minutes played (mp) was less than 200. We felt that a larger sample of playing time would result in more accurate metrics. That hypothesis is somewhat confirmed by the box plots shown below.

The box plot on the left shows the distribution of raptor\_total for rows where mp <= 200. Notice that these rows have values for raptor \_total that range from about -100 to 80. The box plot on the right shows the distribution of raptor\_total for rows where mp >= 200. In this box plot the range of values shrinks drastically. Now they only range from 15 to -14. We believe this filter eliminates some outliers from our dataset and provides more accurate metric values for our analysis.

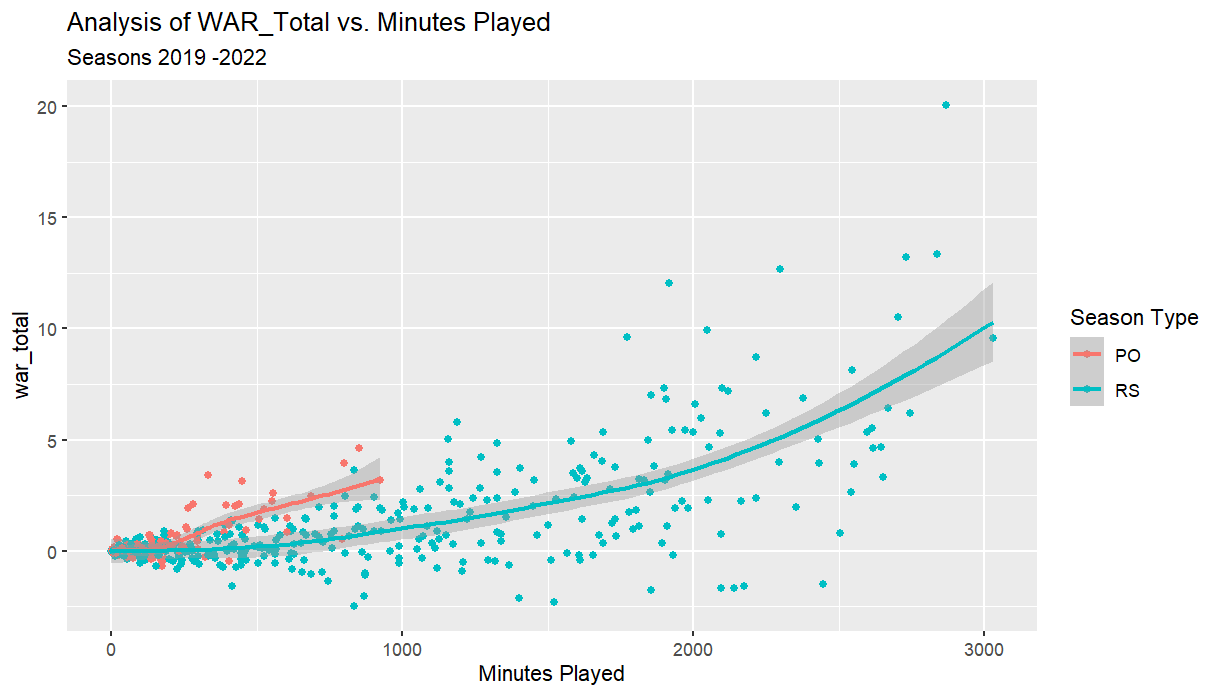
A graph with a line and a bar

Description automatically generated with medium confidence A graph with lines and dots

Description automatically generated

**Visualization XX: Player Analysis of War\_total Vs. Minutes Played Statistics**

After arriving at the conclusion to filter the data to only include players with greater than or equal to 200 minutes played, we wanted to study the relationship between the number of minutes played and the ‘war\_total’ statistic for players.We focused on data collected during the 2019 – 2022 seasons and plotted the relationship between these two variables on a scatterplot along with trendlines showing 99% confidence intervals created using the LOESS method in R. Color was used to draw attention to the data collected in the playoffs versus the regular season, as there is a bit of a skew in the data created by this separation of season type, and by identifying this difference in the data we can explain away the outliers created due to the significant difference in the number of playoff games versus regular season games. The results here indicate that there is a strong correlation between the number of minutes played and the players ‘war\_total’ statistic across both the playoffs and regular seasons. There appears to be an either exponential or polynomial relationship between these variables.



1. **Time Series: Average RAPTOR Offense & Defense (Visualization #4)**

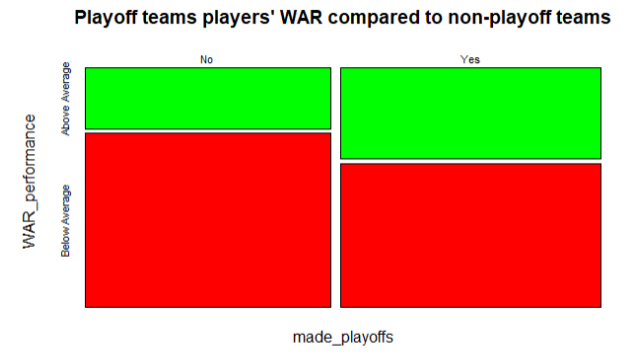
A graph showing the difference between two different lines

Description automatically generated with medium confidence

* Both avg RAPTOR Offense & Defense plotted on a single graph offering a more straightforward comparison.
* Dual-axis feature ensures that both measures share the same y-axis scale, eliminating potential confusion caused by different axis ranges in the original panel plot we created for this data.
* Each measure is distinguished by a distinct color, filtered for mp>=200.

**Visualization XX: Mosaic Plot for WAR performance [2021]**

Something that our dataset was lacking was categorical variables. In order to transform some variables into categorical variables, we decided to use R to make mosaic plots as we felt this would give me the most flexibility when working with the data. We wanted to see how certain stats affect teams’ ability to make the playoffs so for this we created a new data frame of stats from only the 2021 season. From there we created a new column for if the team that player played on made the playoffs that year or not. We also filtered the data to only show stats from the regular season to fairly compare playoff and non-playoff teams. We then created our categorical columns to compare if players were above or below the league average in various statistical categories. This was done by creating a temporary variable that gets the average of a certain statistic and creating a new column that compares each player to that average, categorizing them in either the above or below average group. This same process was done for various statistical categories including raptor offense and defense, and predator offense and defense. Although, we felt the mosaic plot for WAR performance conveyed the clearest message. The plot shows that amongst all the teams that made the playoffs versus the teams that didn’t make the playoffs, playoff teams had more players above the league average in WAR while non-playoff teams had more players below the league average in WAR. This proves that overall team WAR is an important stat to look at when assessing overall team performance in a given season.



-----------------------------------------------------------------------------------------------------------------------------------------

**Explanatory Visualizations**

**Visualization XX: Analysis of Playoff vs. Non-Playoff Teams from 2019 to 2022**

The purpose of this next visualization is to evaluate how effective the metrics war\_total and raptor\_total are at predicting which teams qualified for the playoffs in a given season. As we mentioned earlier, we only included rows where mp >= 200 when calculating the average values for raptor\_total and war\_total.

The bar charts below show the average value of raptor\_total for each NBA team for the 2019, 2020, 2021, and 2022 seasons. The average value of war\_total for each team is encoded in the color of the bar. The “Y” labels indicate the teams that qualified for the playoffs that season.

One of the main takeaways we got from this visualization is that having an average raptor\_total > 0 is a good benchmark for NBA teams to strive for. In these four seasons, 34 of the 37 teams that had an average raptor\_total > 0 made the playoffs. Additionally, 7 of the 8 conference champions and all of the 4 NBA champions in these seasons had an average raptor\_total > 0.

Looking at each of the individual bar graphs, we think the actual results of each of these NBA seasons add some context for the audience. In 2019 and 2020, there are two teams Detroit and Portland who have a noticeably low average raptor\_total. Of the 16 teams that made the playoffs in 2019, Detroit had the fewest wins. In 2020, Portland was the last team to qualify for the playoffs and needed a tie-breaker to do so.

A bar of orange and pink bars

Description automatically generated with medium confidence

A graph of a bar chart

Description automatically generated with medium confidenceA graph of a number of colored bars

Description automatically generated with medium confidence

In 2021, the NBA expanded the playoff field. In the new format, teams that finished 7th through 10th in the standings would participate in a play-in tournament to decide the final two seeds in each conference. In the bar graph for the 2021 season, there is one team with average raptor\_total > 0 that did not qualify for the playoffs, Indiana. That year, Indiana finished 9th in their conference and lost in the play-in tournament. In that same graph, the two teams that had a slightly higher average raptor\_total than the last two playoff teams (Golden State and Charlotte) finished 8th and 10th in their respective conferences and lost in the play-in tournament.

The 2022 bar graph has two non-playoffs teams with average raptor\_total > 0. The Los Angeles Clippers finished 8th in the conference and lost in the play-in tournament. The other team was New York, who finished 11th in the conference (one spot away from the play-in). The only other non-playoff team that had an average raptor\_total greater than the last playoff teams in 2022 was Cleveland, who finished 8th in their conference and lost in the play-in.

A bar of orange and pink bars

Description automatically generated with medium confidence

A graph of different colored squares

Description automatically generatedA graph of a number of colored bars

Description automatically generated with medium confidence

**Visualization XX:Analysis of Top Players by Team**

After analyzing the average player statistics, we wanted to create a visual that would connect the player statistics to their effect on overall team statistics. To do this, we created a stacked bar chart to indicate the number of “Top Players” by season and across teams. To define “Top Players” we calculated each player’s percentile rank in the ‘war\_total’ statistic across each season (2019-2022) and selected players greater than or equal to the 85th percentile in each season as “Top Players”. We filtered out players with less than 200 minutes played as explained in visualization XX to eliminate outliers. This visualization was created in Tableau, and the calculated fields “warrank” and “top\_player” were calculated in the raw excel file for the 2019 – 2022 seasons only. From this data I calculated some summary statistics to learn that the average number of “Top Players” per team is between 3 and 4 across each season. This chart shows us how many “Top Players” a team has on their in a given season and allows for comparison across teams and between seasons. We were hoping to learn a bit more about the teams that were outlier’s in visualization XX, meaning that they had a positive average RAPTOR total score and didn’t make the playoffs, but there didn’t seem to be any significant difference in the number of “Top Players” for those teams in those particular seasons compared to other teams. Specifically, in 2021 MIL had 3 Top Players and in 2022 NYK had 6 Top Players and MIN had 5 Top Players which are well within the season averages for “Top Players” by team.

A graph of different colored bars

Description automatically generated

**Visualizations XX&XX: NBA Player Career Performance Analysis**

Now that we have discussed the predictability of team performance and the overall impact players have on team’s success, we are shifting our focus towards the individual player’s career progression. When is the standard “peak” for an NBA player and how does that ultimately relate to their successes, failures, and length of career? The following visualizations detail player’s regular season performance using the WAR Total variable plotted against the number of seasons they have played. For example, if someone played for six seasons, took a break during one due to retirement/injury/etc., and then returned for two more, their performance would be reflected in the 1st to 8th seasons.

In order to create these visualizations, we had to introduce some data preprocessing steps and transformations. Neither “Seasons Played” nor “Player Rankings” were present in the original dataset. Instead of building the relationships in R or Tableau, we created these calculations in the baseline csv dataset using Excel formulas and standard NBA philosophies. “Seasons Played” was just a straight count incrementing each unique year a player was present in the RAPTOR dataset. “Player Rankings” required a bit more consideration and we settled on the following assumptions: 1) minimum of 10 seasons played and 2) average performance throughout the regular season. We wanted to align on a very strict criteria for this ranking so it could be tested and reproduced in the future.

Looking at the first visualization (below), we are using box plots by season to track the overall performance for the top 30 historic NBA players. The higher the overall WAR Total observation, the better the player performed during that regular season. This dataset reflects the careers of household names like Michael Jordan, LeBron James, and Kobe Bryant.

A graph of red and white columns

Description automatically generated with medium confidence

When creating this visualization, the drafting process primarily centered around our choice of “n” players. With only a max of 30 observations per season, our final box plots may struggle with detailing the overall performance variance (based on the IQR) of the season. However as we added more observations throughout our drafting process, we noticed that the distinct trend we observed within the top players normalized to zero. Ultimately, we decided on the top 30 historic players as a strong middle-ground.

When reviewing the overall story of this visualization there is a clear trend as the best players progress throughout their career. To no one’s surprise, even the best players have a slower start. As they continue to get experience they continue improving and normally peak between their 4th and 9th seasons with moderate regression afterwards. Regardless of the number of observations and the high variance throughout each box plot, this is a distinct trend on its own. When reviewing the minimal number of outliers, which we felt were insightful to label, it primarily reflected injuries in the worst cases and championships in the best. We believe this only further solidified our model when the biggest deviations from the trends were logical and clearly explainable.

Now that we have set a precedent for the career progression of the best NBA players, we reviewed the overall trends across the entire NBA population. Using a similar box plot graphic while only excluding the players who played less than 200 minutes in a season completely changed the visual. The variance for each box plot remained high but we lost the majority of the predictable trends we saw from the best players. Furthermore, we added the individual markers captured in the “Top 30” visualization in red to contrast the best of the best with the wholistic population.

A graph of different colored columns

Description automatically generated with medium confidence

While we were initially puzzled by the drastic differences between views, we noticed a few trends. First, the majority of the box plots are centered at or near zero. Logistically, this makes sense because the Total WAR variable is normalized to represent the average NBA player. Since most of this graphic represents the average, with the exception of those with less than 200mins, fluctuating around zero is perfectly normal. Next upon further review of both the box and whisker trends, we see some of the highest peaks between the 5th and 8th seasons. While not as definitive as the top players graphic, it is consistent and explainable if the below-average players retire quickly after their peak and fall out of this dataset.

Using these two visualizations could be imperative for predicting a prospective players success. If, across their first few years, they are trending along the path of the best or worst, NBA teams and the players themselves could make informed decisions on contract and career strategies.