Scott Jue
MSDS 456
3/12/2023

Assignment 2 Option B: Team Construction and Player Types

Introduction

The objective of this analysis will be to determine what factors it takes to win in the current NBA climate and how we can use a modern approach to approach player classification that can be used to determine player types and improve team construction. Determining which factors it takes to win will not only help the team prioritize what areas of the game are most important for winning, but it will also provide guidance in how to evaluate player performance. Previous research by Dean Oliver has found that a Four Factor Model can be used to predict the number of wins for a team. As such, I will replicate this model using the 2021-2022 regular season data to determine where the Golden State Warriors were ranked in the NBA last season and where the team currently stands this season. Additionally, there is belief that the standard basketball positions do not accurately reflect the game today. Therefore, I will also build a classification model using machine learning methods to determine a new classification framework for identifying player types. Previous research by Alex Cheng utilized a KMeans Clustering method to create 8 different position types based on a variety of player stats. I will use a similar approach to develop a player classification model using 2021-2021 and 2021-2022 regular season data. With this analysis, the team may determine the optimal mix of positions to insert into a 5-man lineup to increase the chances of winning using position types that are a more accurate representation of how the game is played today.

Replication of the Four Factor Model

The data I used to replicate the Four-Factor Model is 2021-2022 regular season data acquired from NBA.com. NBA.com provides a specific table that has the variables needed for the Four-Factor Model. The table contains Effective Field Goal (EFG), Opponent Effective Field Goal

(OEFG), Offensive Turnover Percentage (TPP), Defensive Turnover Percentage (DTPP), Offensive Rebound Percentage (ORP), Defensive Rebound Percentage (DRP), Free Throw Rate (FTR), and Opponent Free Throw Rate (OFTR).As such, I have chosen to use this table as it provides the exact data needed for the analysis and requires minimal cleaning or filtering of the data besides once selecting the season and season type (regular vs playoffs) from the drop-down selection on the website. I also changed the variables names to match the names used in Dean Oliver's Four-Factor model. Additionally, the raw data had some percentages in percent format and some in decimal format. So, I converted everything to decimal format which is how the original Four-Factor model was formatted. I handled these data cleaning processes in Excel before loading the data into R. Additionally, prior to modeling I also create four new variables for the differentials between the respective offensive and defensive values for team stat. The variables are Shooting Rate Diff, Turnover Rate Diff, Rebounding Rate Diff, and Free Throw Rate Diff. These will be the variables used in the Four-Factor Model.

We can see how the overall performance metrics are related to a team's win in the below correlation matrices in Figure 1 and Figure 2.

	W	EFG	FTR	TPP	ORP	OEFG	OFTR	DTPP	DRP
W	1.00	0.65	-0.06	-0.40	0.17	-0.65	-0.15	0.25	-0.28
EFG	0.65	1.00	0.13	-0.01	-0.13	-0.31	-0.21	-0.12	-0.24
FTR	-0.06	0.13	1.00	0.26	0.03	0.23	0.10	-0.18	-0.16
TPP	-0.40	-0.01	0.26	1.00	-0.03	0.15	0.31	0.06	0.01
ORP	0.17	-0.13	0.03	-0.03	1.00	0.01	0.10	0.42	0.01
OEFG	-0.65	-0.31	0.23	0.15	0.01	1.00	0.24	-0.03	0.14
OFTR	-0.15	-0.21	0.10	0.31	0.10	0.24	1.00	0.53	0.20
DTPP	0.25	-0.12	-0.18	0.06	0.42	-0.03	0.53	1.00	0.41
DRP	-0.28	-0.24	-0.16	0.01	0.01	0.14	0.20	0.41	1.00

Figure 1. Correlation Matrix of overall team performance metrics.

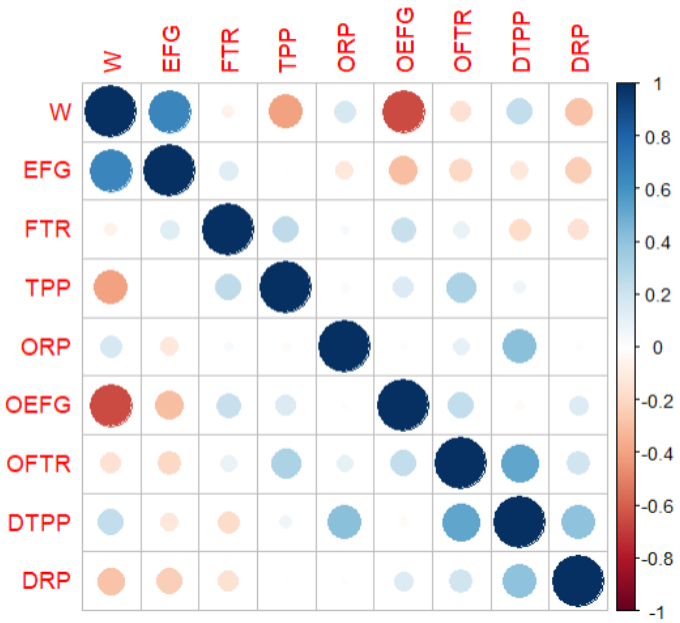


Figure 2. Correlation Heatmap of overall team performance metrics.

Effective Field Goal percent (EFG) has the strongest positive correlation to wins at 0.65. And Opponent's Effective Field Goal percent (OEFG) has the strongest negative correlation to wins at -0.65. Turnover Percent (TPP) has a moderate negative correlation to wins at -0.4 and Defensive Turnover Percent (DTPP) has a low positive correlation to wins at 0.25. Offensive Rebound Percent (ORP) has a low positive correlation to wins at 0.17 and Defensive Rebound Percent (DRP) has a low negative correlation to wins at -0.28. Lastly, Free Throw Rate (FTR) has almost no correlation to wins at -0.06 and Opponent Free Throw Rate (OFTR) has a low negative correlation to wins at -0.15. Based on these correlations, the biggest contributing factor is the team's scoring ability (EFG) and the team's ability to minimize the opposing team's scoring (OEFG). The next important factor, is the team's ability to limit turnovers (TPP) while also causing turnovers defensively (DTPP). Next is the team's rebounding performance followed by the team's ability to minimize the opposing team's free throw rate. As mentioned previously, the team's free throw rate does show to be correlated to wins.

	W	shootdiff	TOdiff	rebdiff	FTdiff
W	1.00	0.81	-0.47	0.29	0.08
shootdiff	0.81	1.00	-0.01	0.04	0.19
TOdiff	-0.47	-0.01	1.00	-0.14	0.37
rebdiff	0.29	0.04	-0.14	1.00	0.08
FTdiff	0.08	0.19	0.37	0.08	1.00

Figure 3. Correlation Matrix of the Four Factors.

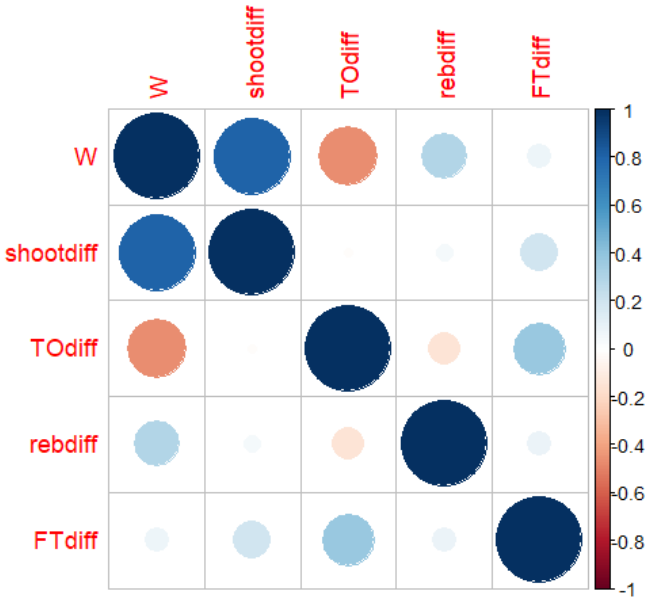


Figure 4. Correlation Heatmap of the Four Factors.

The correlation matrices from Figure 3 and Figure 4 show the relationships of the four factors which are the differences between the offensive and defensive stat for each type of variable (i.e. shooting, turnover, etc.). We can see how this is provides a more simplified version of the relationships between the variables by using the differences instead. The difference in shooting rates (EFG- OEFG) is the most correlated variable to wins at 0.81. The difference in turnover rates (TPP-DTPP) is negatively correlated at -0.47. The difference in rebound rates (ORP-DRP) has a positive correlation to wins at 0.29. Lastly, the difference in free throw rates does not appear to be correlated to wins, since the coefficient is 0.08. It is also important to note

that none of the explanatory variables show a strong correlation with each other. As such, we do not have issues of collinearity by using these four variables for the model.

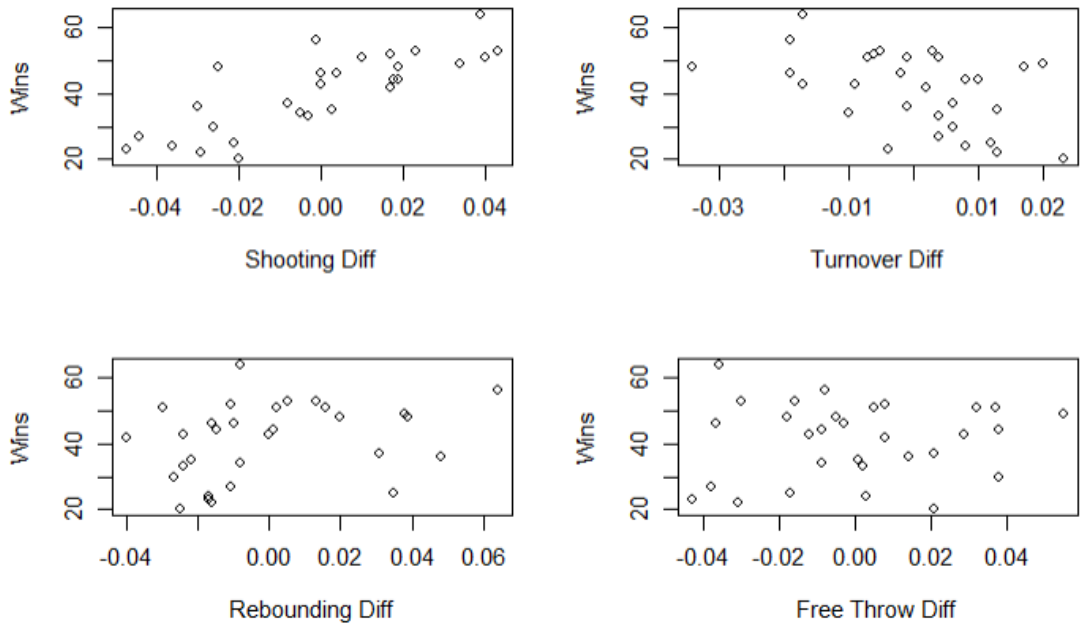


Figure 5. Scatterplots of Four Factors vs Wins

From the above scatterplots (Fig. 5) of the four factors we can see the relationship that these variables have to the number of wins. The linear patters or lack thereof are consistent with the correlation matrixes seen previously. Shooting Diff shows the strongest linear pattern to wins.

Using the below table (Fig. 6), we can see where each team ranks for each of the overall performance metrics of the Four-Factor Model from the previous season. We can see that the Warriors are rank 3rd in EFG and 2nd in OEFG. For TPP, the Warriors are ranked 29th so this is an area I would recommend to improve to further increase the team's winning percentage. The teams DTPP is tide for 6th in the league. The team is in the middle of the league (16th) in ORP, but 6th in the league for DRP. Lastly, the Warriors are 23rd in FTR and 26th in OFRT. The free

throw rates could be improved, but shouldn't take priority over TPP as the free throw rates have shown to not be as correlated to wins as TPP.

TEAM <chr>	W <dbl>	EFG_R <dbl>	FTR_R <dbl>	TPP_R <dbl>	ORP_R <dbl>	OFTR_R <dbl>	OEFG_R <dbl>	DTPP_R <dbl>	DRP_R <dbl>
Phoenix Suns	64	4.0	29.0	4.5	21.0	22.5	3.0	8.0	13.0
Memphis Grizzlies	56	23.0	19.0	7.0	1.0	18.0	10.0	4.0	15.0
Golden State Warriors	53	3.0	23.0	29.0	16.0	26.0	2.0	6.5	6.0
Miami Heat	53	5.0	12.5	28.0	10.0	27.0	11.5	3.0	8.0
Dallas Mavericks	52	13.5	14.5	7.0	24.0	10.0	7.5	17.0	10.0
Boston Celtics	51	9.0	21.5	13.5	11.5	9.0	1.0	11.0	16.5
Milwaukee Bucks	51	6.0	9.0	10.5	16.0	3.5	19.0	27.0	2.0
Philadelphia 76ers	51	16.5	2.0	4.5	30.0	13.5	11.5	17.0	18.5
Utah Jazz	49	2.0	4.0	21.0	5.0	1.5	7.5	29.0	5.0
Denver Nuggets	48	1.0	20.0	25.5	18.5	12.0	20.0	27.0	1.0
Toronto Raptors	48	27.0	24.5	3.0	2.0	17.0	17.5	1.0	22.5
Chicago Bulls	46	10.0	16.5	7.0	27.5	15.5	22.5	22.5	8.0
Minnesota Timberwolves	46	12.0	10.0	17.5	7.0	29.0	17.5	2.0	28.0
Brooklyn Nets	44	11.0	18.0	17.5	8.5	20.0	7.5	20.5	30.0
Cleveland Cavaliers	44	13.5	7.0	27.0	11.5	5.0	5.0	13.5	18.5
Atlanta Hawks	43	8.0	11.0	1.0	16.0	6.0	24.0	27.0	11.5
Charlotte Hornets	43	7.0	24.5	9.0	14.0	11.0	25.0	5.0	28.0
LA Clippers	42	19.5	26.0	13.5	26.0	1.5	4.0	15.0	28.0
New York Knicks	37	26.0	3.0	12.0	6.0	24.0	7.5	24.5	4.0
New Orleans Pelicans	36	24.0	5.5	21.0	4.0	13.5	26.0	9.0	3.0
Washington Wizards	35	18.0	12.5	10.5	29.0	15.5	13.0	30.0	11.5
San Antonio Spurs	34	21.0	30.0	2.0	8.5	8.0	14.5	17.0	24.0
Los Angeles Lakers	33	15.0	8.0	19.0	24.0	22.5	21.0	13.5	21.0
Sacramento Kings	30	22.0	5.5	15.5	22.0	7.0	27.0	19.0	22.5
Portland Trail Blazers	27	25.0	16.5	25.5	18.5	28.0	30.0	10.0	20.0
Indiana Pacers	25	19.5	21.5	24.0	3.0	21.0	28.0	20.5	16.5
Oklahoma City Thunder	24	29.0	27.5	15.5	24.0	3.5	16.0	22.5	14.0
Detroit Pistons	23	30.0	14.5	21.0	13.0	30.0	22.5	6.5	26.0
Orlando Magic	22	28.0	27.5	23.0	27.5	19.0	14.5	24.5	8.0
Houston Rockets	20	16.5	1.0	30.0	20.0	25.0	29.0	12.0	25.0

Figure 6. Team Rankings for overall performance metrics.

The rankings for the teams using the four factor variables can be seen below (Fig. 7). The Warriors were ranked first in shooting. Since this is the highest correlated variable to wins, being best in the league in this stat should lead to more wins. For turnovers, the Warriors are right in the middle of the pack being ranked 16th. As mentioned previously, this is an area where the team can improve as this variable has the second strongest correlation to wins. For rebounding, the Warriors are in the top third of the league being ranked 10th. Lastly, the Warriors are in the bottom third of the league for free throws being ranked 25th. However, this isn't the biggest of concerns as free throws are not as strongly correlated to wins as the other variables.

TEAM <chr>	W <dbl>	shootdiff_r <dbl>	TOdiff_r <dbl>	rebdiff_r <dbl>	FTdiff_r <dbl>
Phoenix Suns	64	3.0	5.0	15.0	27.0
Memphis Grizzlies	56	18.0	3.0	1.0	18.0
Golden State Warriors	53	1.0	16.0	10.0	25.0
Miami Heat	53	5.0	10.0	9.0	22.0
Dallas Mavericks	52	10.0	9.0	18.0	10.5
Boston Celtics	51	2.0	13.5	11.0	12.0
Milwaukee Bucks	51	11.5	18.5	8.0	4.0
Philadelphia 76ers	51	11.5	8.0	29.0	5.0
Utah Jazz	49	4.0	29.0	4.0	1.0
Denver Nuggets	48	6.5	28.0	7.0	17.0
Toronto Raptors	48	24.0	1.0	3.0	24.0
Chicago Bulls	46	16.0	12.0	20.5	16.0
Minnesota Timberwolves	46	13.0	2.0	16.0	28.0
Brooklyn Nets	44	6.5	22.0	19.0	19.5
Cleveland Cavaliers	44	8.0	24.0	12.0	2.5
Atlanta Hawks	43	16.0	7.0	13.0	6.0
Charlotte Hornets	43	16.0	4.0	25.0	21.0
LA Clippers	42	9.0	15.0	30.0	10.5
New York Knicks	37	21.0	20.0	6.0	7.0
New Orleans Pelicans	36	27.0	13.5	2.0	9.0
Washington Wizards	35	14.0	26.5	24.0	15.0
San Antonio Spurs	34	20.0	6.0	14.0	19.5
Los Angeles Lakers	33	19.0	17.0	26.0	14.0
Sacramento Kings	30	25.0	21.0	28.0	2.5
Portland Trail Blazers	27	29.0	18.5	17.0	29.0
Indiana Pacers	25	23.0	25.0	5.0	23.0
Oklahoma City Thunder	24	28.0	23.0	23.0	13.0
Detroit Pistons	23	30.0	11.0	22.0	30.0
Orlando Magic	22	26.0	26.5	20.5	26.0
Houston Rockets	20	22.0	30.0	27.0	8.0

Figure 7. Team Rankings for the Four Factors.

Variable Importance of Four Factor Model

To create the Four Factor Model, I used the 4 variables from above as explanatory variables to create a multiple linear regression model for predicting the number of wins using the regular season data from the 2021-2022 season. The model coefficients can be seen in the below table (Fig. 8). Using these coefficients, we can determine how these variables impact the number of predicted wins and the relative importance.

Intercept	41.041
Shooting Diff	361.174
Turnover Diff	-417.319
Rebound Diff	82.24
Free Throw Diff	38.11

Figure 8. Four Factor Model Coefficients

This model results in an R-squared value of 0.906 which means that these four variables are able to explain approximately 91% of the variance observed in regular season wins during the 2021-2022 season. The standard error for this model is 3.83.

The shooting differential was shown to have a 0.81 correlation to wins and by itself explains 65% of the variation in wins. The turnover differential was shown to have a -0.47 correlation with wins and by itself explains 21% of the variation in wins. The rebound differential was shown to have a 0.29 correlation with wins and by itself explains 4% of the variation in wins. Lastly, the free throw differential was shown to have a 0.08 correlation with wins and by itself explains 1% of the variation in wins. From this analysis we can conclude that a team's shooting percentage differential is the most important factor for winning NBA games.

We can see that the intercept value is 41.041. This suggests that if a team performed equally on the offensively and defensively resulting in a 0 for each variable, the baseline number of predicted wins is approximately 41. The relative importance for the four variables is summarized below.

- A .01 increase in Shooting Diff is worth 3.6 wins. This can be achieved by:
    1. Improve our EFG by 1%
    2. Reduce our opponent's EFG by 1%
    3. Improve our EFG by 0.5% and reduce our opponent's EFG by 0.5%

  The above improvements could on average translate to 3.6 more wins.
- A .01 increase in Turnover Diff is worth 4.1 wins. This can be achieved by:
    1. Committing one less turnover per 100 possessions

2. Committing one less turnover per 200 possessions and forcing on more turn over per 200 defensive possessions.

The above improvements could on average translate to 4.1 more wins.

- A .01 increase in Rebounding Diff is worth 0.8 wins. This can be achieved by:
    1. One more offensive rebound per 100 missed shots
    2. One more defensive rebound per 100 missed shots by the opponent
    3. One more offensive rebound per 200 missed shots and one more defensive rebound per 200 missed shots by the opponent

The above improvements could on average translate to 0.8 more wins.

- A .01 increase in Free Throw Diff is worth 0.4 wins. This can be achieved by:
    1. One more Free Throw made per 100 FG attempts
    2. One less Free Throw given up per 100 FG attempts by an opponent
    3. One more Free Throw made per 200 FG attempts and one less FT given up per 200 FG attempts

The above improvements could on average translate to 0.4 more wins.

Using this model, we can the team statistics in each of these four factors to predict the amount of games the team will win a season. A graph of the predicted wins vs the actuals wins for the 2021-2022 season is provided below (Fig. 9). We can see that the plot of points shows a linear pattern which is a good sign that the model's predictions are reasonable and that a linear regression model provides a good overall fit for data.

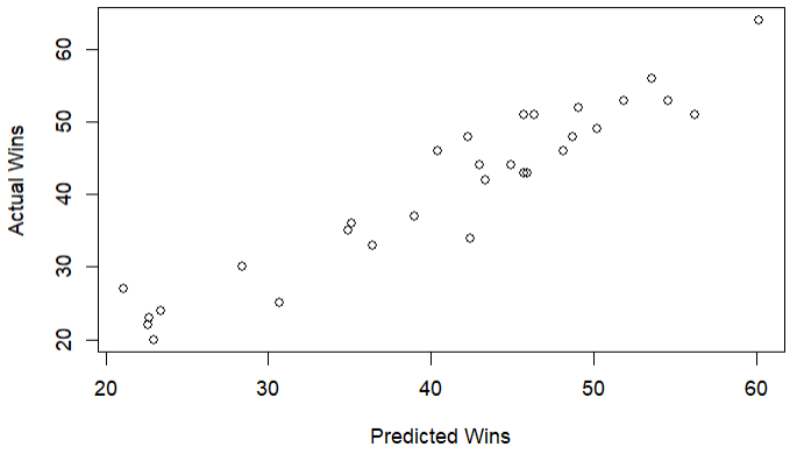


Figure 9 Predicted Wins vs Actual Wins (2021-2022 regular season)

In the case of the Warriors, the team had 53 wins during the 2021-2022 season and the model predicted the Warriors to have approximately 55 wins using the four factors. Thus, the model was only off by 2 wins which is within 1 standard deviation. Furthermore, as previously mentioned this model has a R-squared of 0.91 with a standard error of 3.83. Therefore, the model is fairly accurate in predicting the number of teams. In addition to determining how certain factors can impact winning, this model can also be used during the season by using the current values of the four factors to predict the final total wins of the team for the entire season. For the current season, as of the time of writing this report, the Warriors are predicted to win approximately 39 games. The team currently has 35 wins through 68 games played. This prediction seems reasonable given that the standard error is 3.83 wins.

Determination of Classification Framework

After determining the important factors that relate to the overall team statistics, I created a classification model that aims to create new position types that reflect how the game is played today and group current NBA players into there new categories. Similar analysis and research have been done by Cheng (2017) in which a k-means clustering model was used to determine 8

new positions types and grouped NBA players into these new positions. As such, I will use a similar modeling approach for determining a new classification framework for position types.

For this model I acquired 2020-2021 and 2021-2022 regular season data from basketball-reference.com. I used two different datasets from the website. First, I acquired the data for Per 100 Possessions for each season which is similar to the game stats but is summarized by per 100 possessions. Then, I also included the Shooting data for each season which has more shot specific data such as distance and location of shots for players during the course of the season. I merged these two data sets together and removed variables that were not necessary or useable in the model such as player rank, age, games started, and team. This resulted in a data set of 601 player observations and 36 variables.

Next, I filtered the data to only include players that played more than 40 games for the 2021-2022 season and 36 games for the 2020-2021 season since it was a shortened season due to COVID-19. This was done to handle and potential outliers and is also consistent with the data cleaning approach by Cheng (2017). I then aggregated the data and took the average from the two seasons by player, so that there wouldn't be any duplicate players (one for each season). Lastly, I created a subset of the data which only included the player name and 32 explanatory variables removing variables such as games played and minutes played.

Due to the high number of explanatory variables that are correlated with one another, I have decided to perform dimensionality reduction on these variables to reduce the amount of input variables used in the classification model. For this process, I used Linear Discriminant Analysis (LDA) to reduce the number of dimensions in the data set while also retaining as much

information as possible. This methodology is also used by Cheng (2017) in his analysis to reduce the amount of variables used since he started with 56 variables.

Prior reducing the dimensions, I scaled the explanatory variables using the StandardScaler function from sklearn. Using the current position designations as the target variable, I then used LDA to reduce the 32 variables into 2 dimensions while still capturing approximately 62% of the data. Cheng (2017) also reduced the variables into 2 dimensions, so I felt this was an appropriate number of dimensions as it allows for better interpretation. I also tested using Principal Component Analysis (PCA) to reduce the dimensions; however, this resulted in a cumulative explained variance of approximately 53% for 2 dimensions. These results were similar to the analysis by Cheng (2017). As such, the LDA method is able to explain more of the variance observed in the data set and will be used to reduce the input variables for the classification model. Below is a plot of players (Fig. 10) using the two dimensions from the LDA model. We can see that this plot produces clusters of players with similar position types. However, the clusters seem to overlap, so using a KMeans Clustering model we can determine new clusters (position types) that do not overlap based on these 2 dimensions. This is because KMeans Clustering is an iterative algorithm that assigns each data point to a cluster center and calculates the mean difference between the data points and cluster centers. Then it is able to find the cluster centers that optimizes the position of the centroids, so that they best represent certain regions of the data (Cheng 2017).

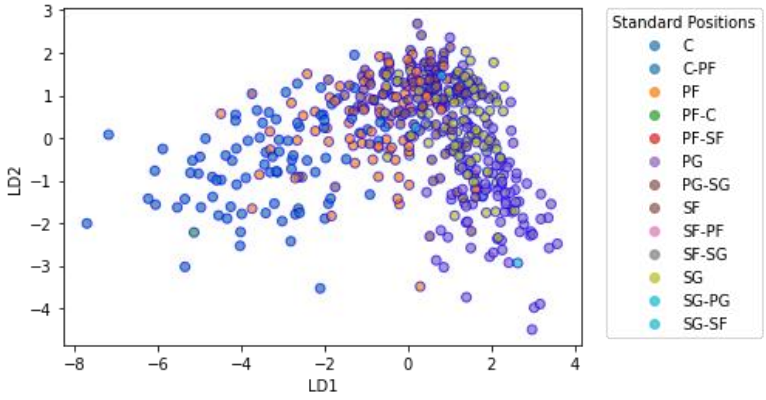


Figure 10. Scatterplot of Players using 2 dimensions from LDA by original position type

Using a silhouette score I determined that having 7 clusters would be appropriate for the KMeans Clustering model. Silhouette scores measures the similarity of an object to its own cluster compared to other clusters. Cheng (2017) also used silhouette score to determine the number of clusters to be used in the KMeans models; however, he decided to have 8 clusters based on the score observed in his analysis. In the below plot (Fig. 11) we can see the same plot of players using the 2 dimensions, but now the points are grouped into the 7 different clusters. Compared to the original plot, there players are grouped so that there is no overlap between clusters.

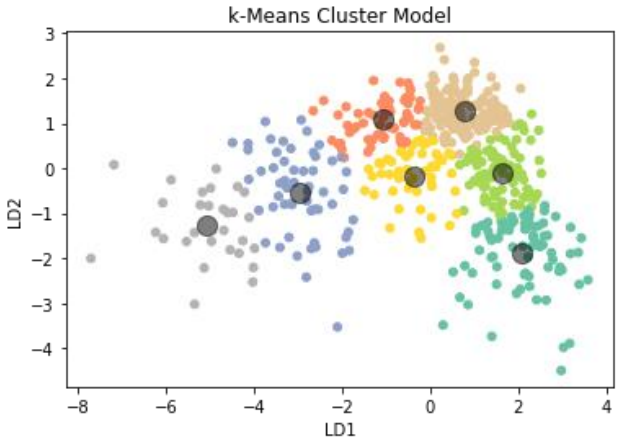


Figure 11. Scatterplot of Players using 2 dimensions from LDA by cluster classification

Since LDA reduced the variables into 2 dimensions, I used PCA to extract the most important features from the dimensions to aide with interpretability for defining each cluster. Using these important features, I created new position names that describes the position type that the cluster represents which are Scoring-Playmaker Wing, Versatile Guard, Inside Scoring Forward, Versatile Wing, Offensive Big, Defensive Centers, and Ball-Handling Scorers. In the scatterplot below (Fig. 12), we can see where each of these clusters appear on for these position types.

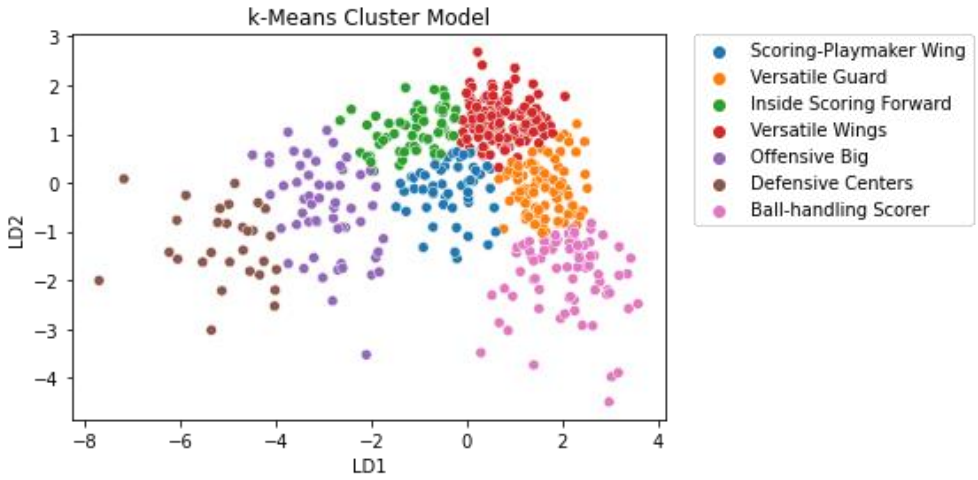
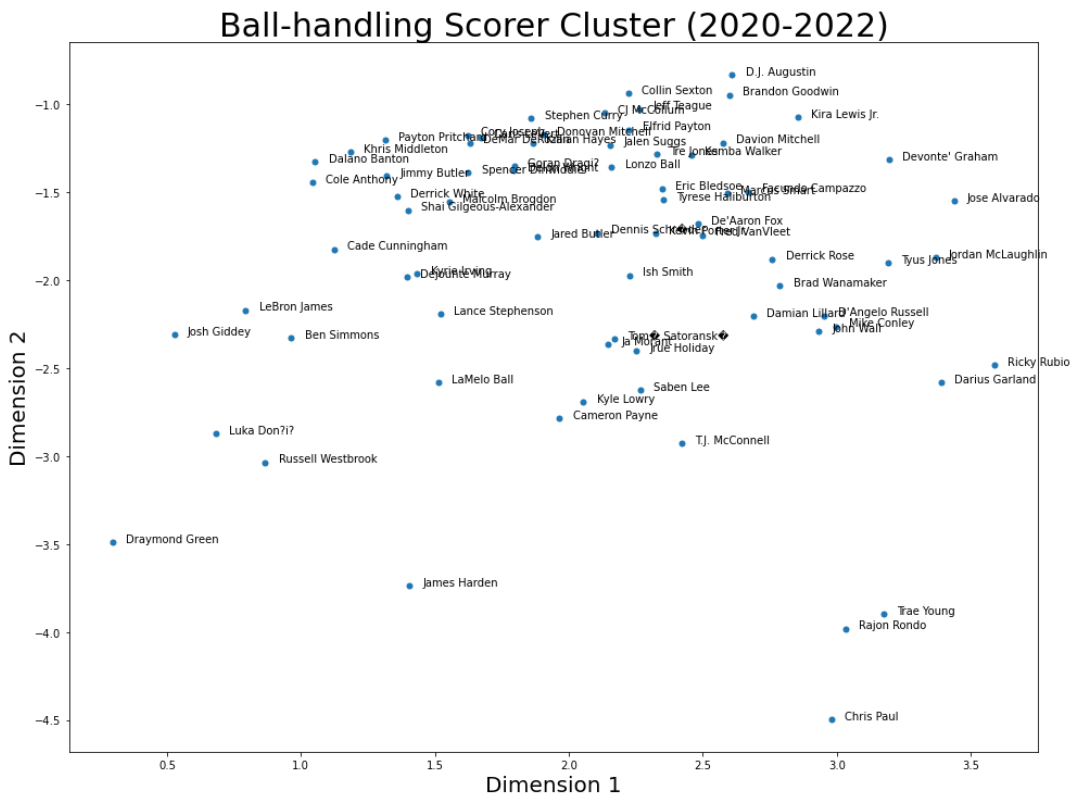
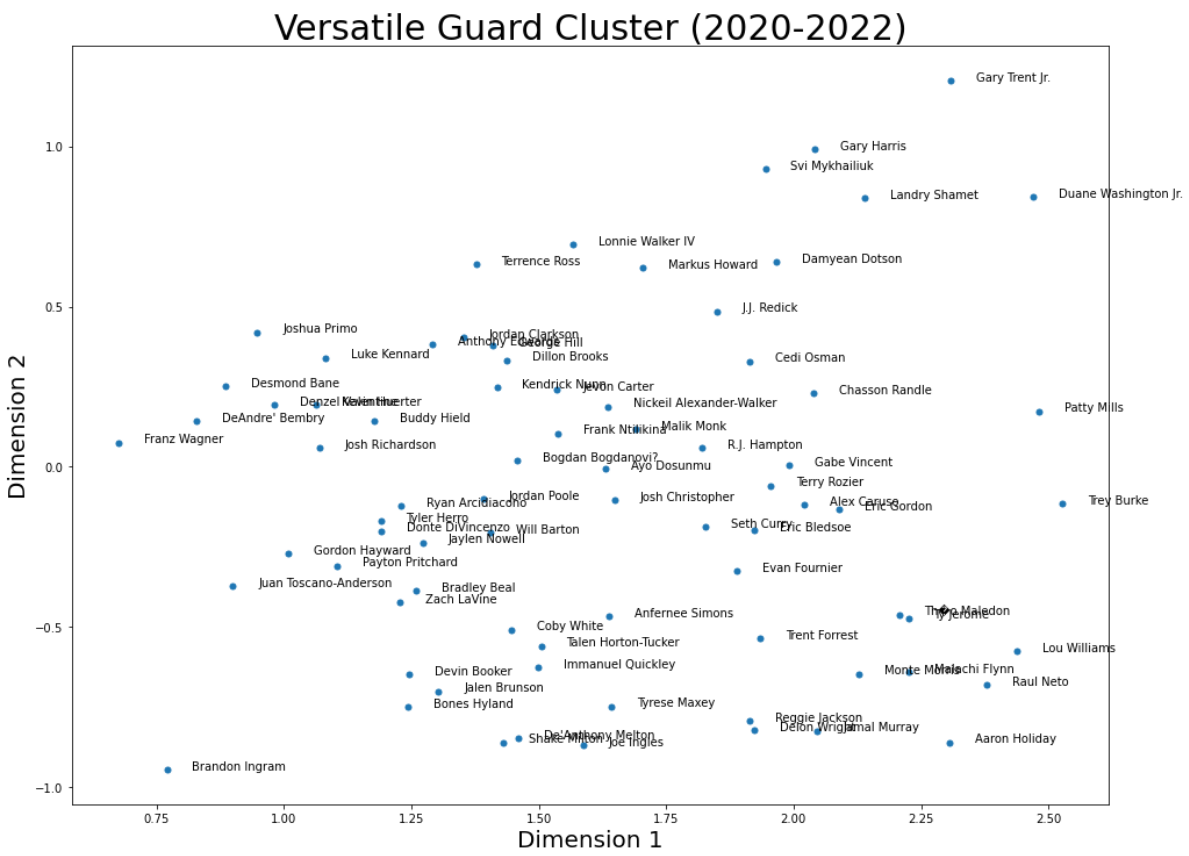


Figure 12. Scatterplot of Players using 2 dimensions from LDA by cluster label

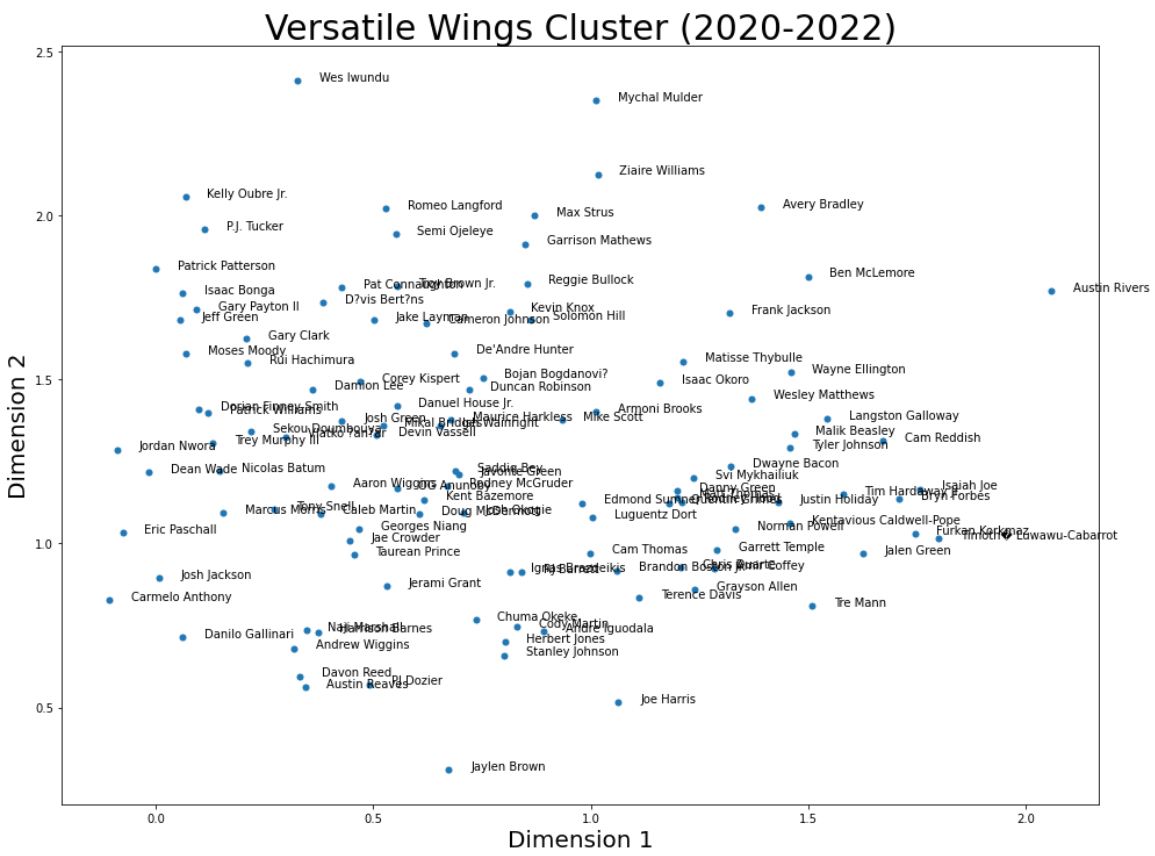
Once I determined these new positions, I then used the model to classify current NBA players into these 7 positions. Below are plots for each of these positions and with each point representing a player from that position type or cluster.



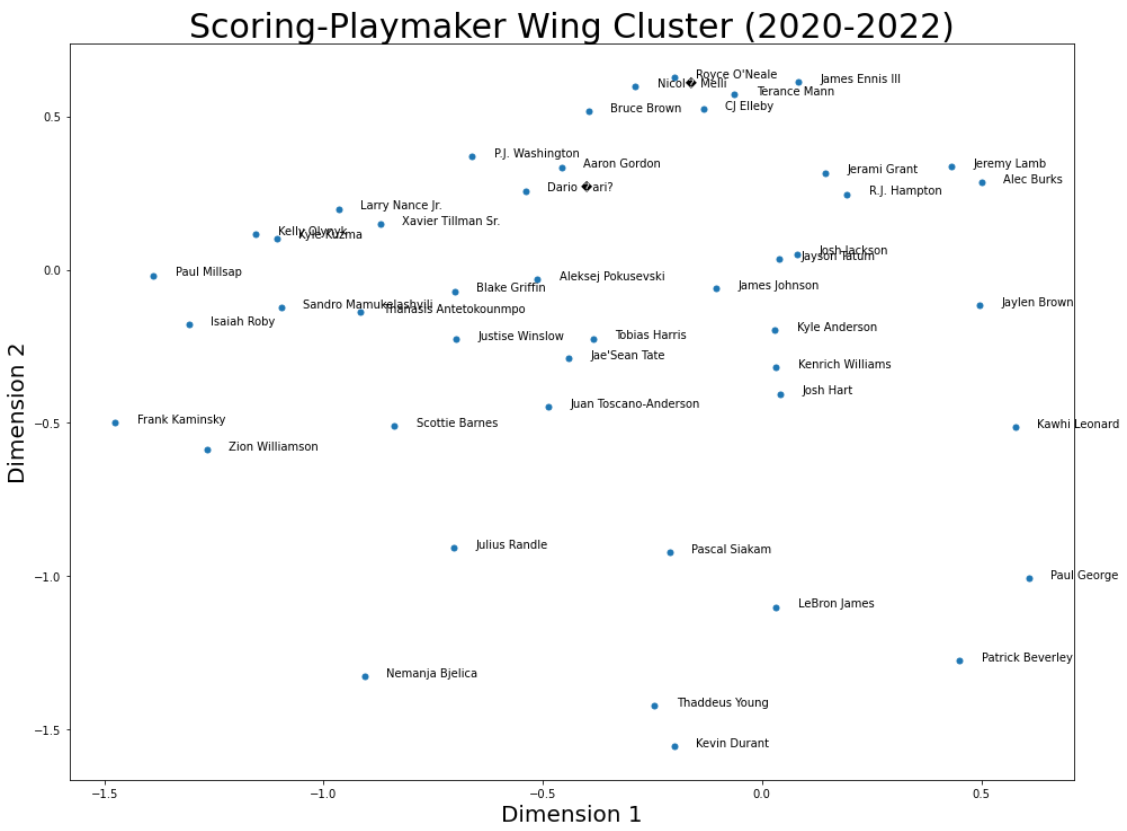
For the Ball-handling Scorer position the notable players are LeBron James, Stephen Curry, Luka Doncic, James Harden, and Kyrie Irving. These players prefer to possess the ball to create offense and are one of the primary scorers on their respective team. Some of the notable features from this cluster are points, FG made, FG attempts, and FT attempts.



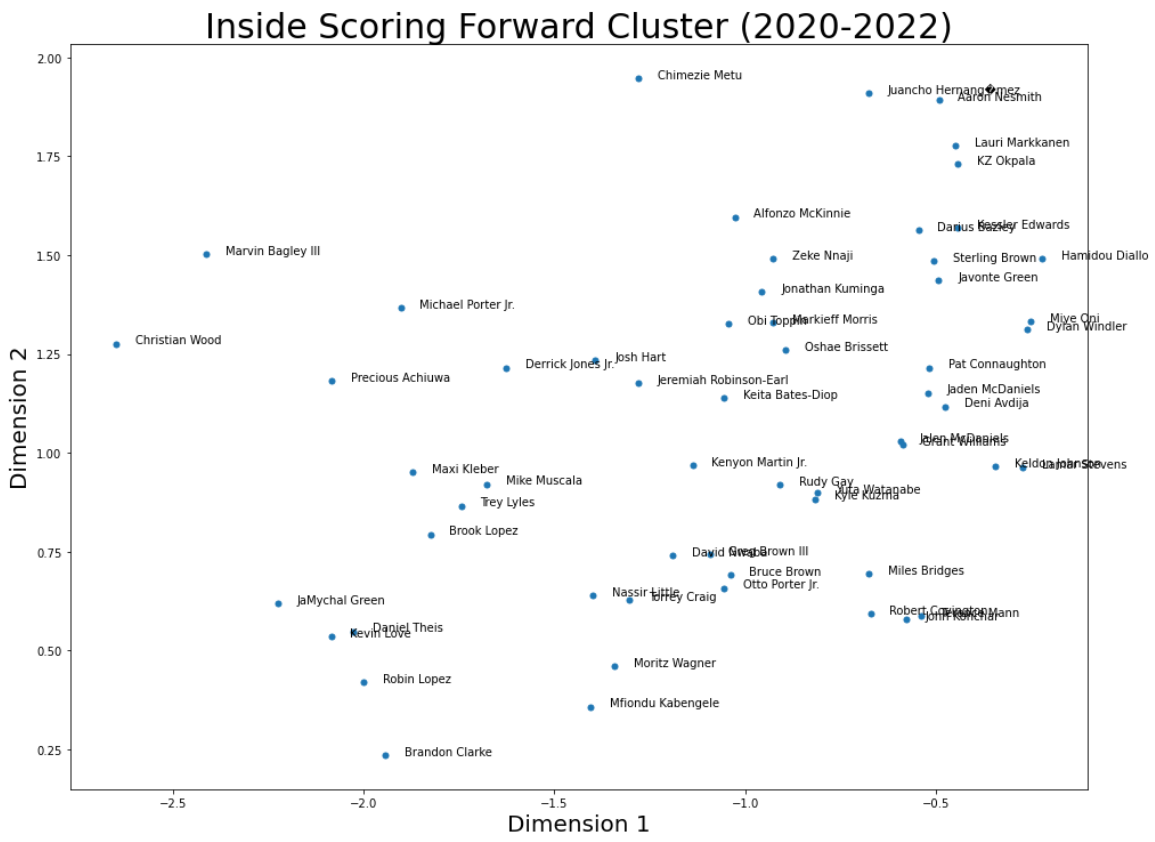
For Versatile Guards, the notable players are Bradley Beal, Devin Booker, and Jordan Poole. These are players that can provide versatility from the traditional Guard position. They can provide scoring from mid-range or driving and attacking the rim. Some of the notable features from this cluster are FG made, PTS, and 2-point attempts.



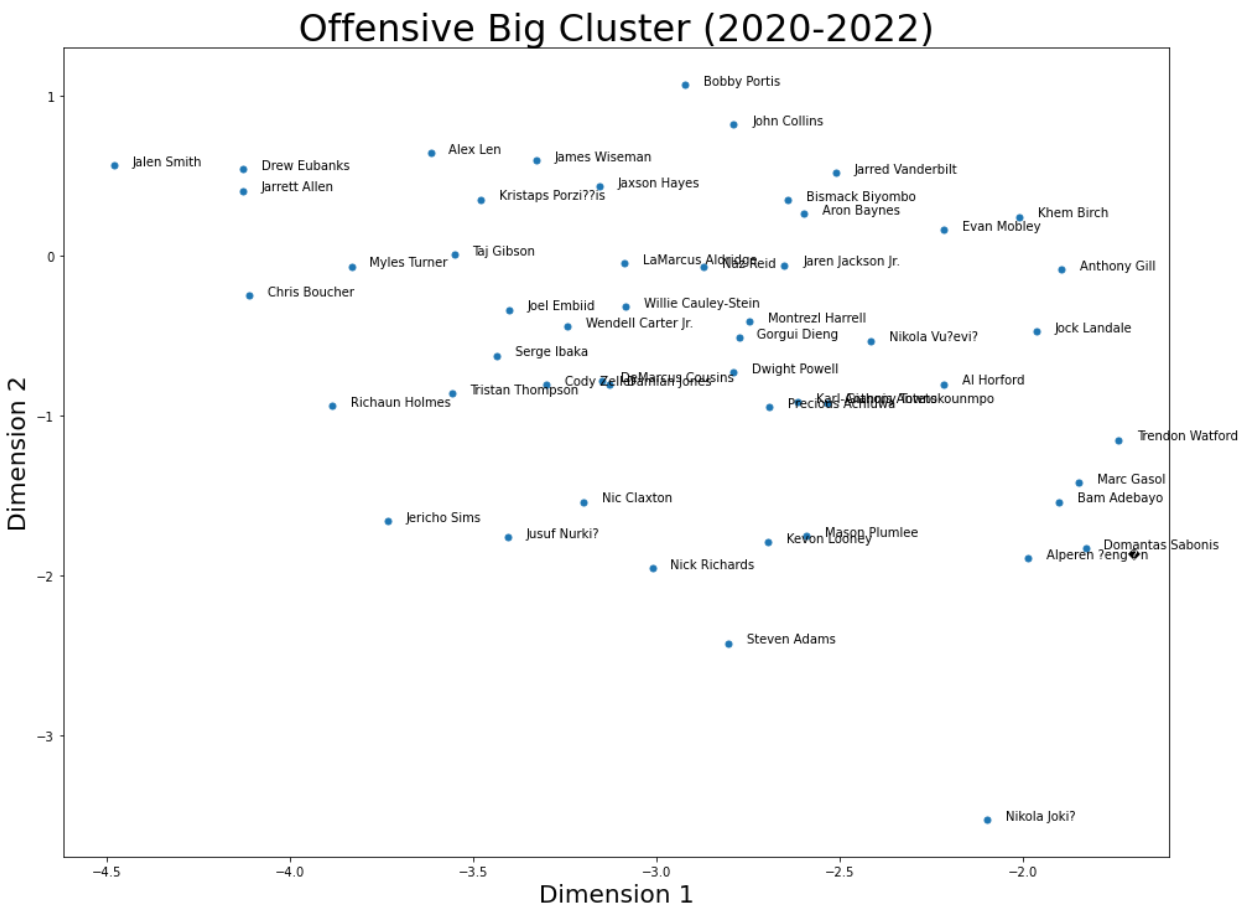
For Versatile Wings, the notable players are Jaylen Brown, Andrew Wiggins, and Carmelo Anthony. These are players that can provide versatility from a wing-style position. Similar to Versatile Guards, they can provide scoring from shooting or driving and attacking the rim. This position is less likely to be handling the ball during a team's offensive position. Some of the notable features from this cluster are 2-point attempts, FG made, FT attempts.



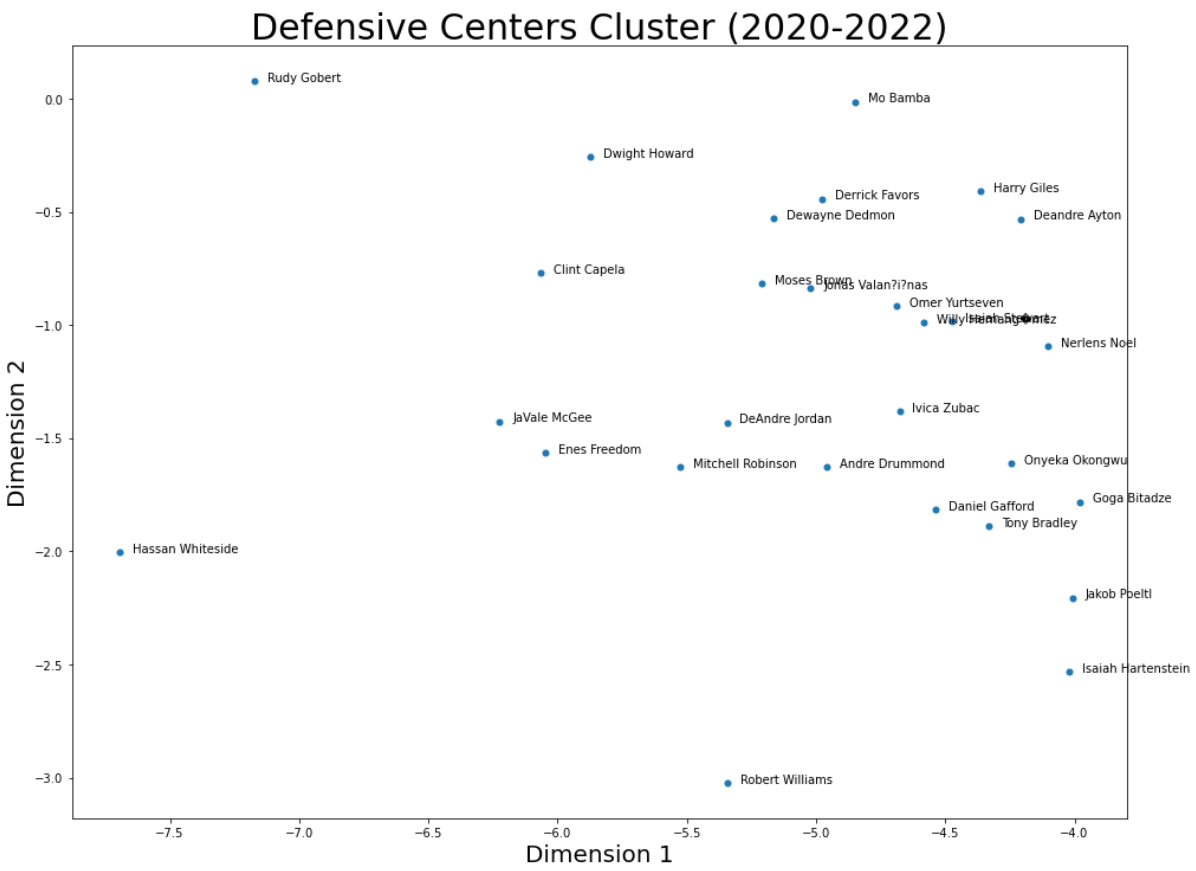
For Scoring-Playmaker Wings, the notable players are Zion Williamson, Kevin Durant, Jayson Tatum, and Kawhi Leonard. These players play a similar style to the Ball-handling Scorers; however, they typically are bigger in size and will set up more from the wing position. In addition to providing offense through scoring, these players also possess strong passing and playmaking abilities to also add to the team's overall offense. Some notable features from this cluster are PTS, FG attempts, FG mates, and FT made.



For Inside Scoring Forwards, the notable players are Kevin Love, Michael Porter Jr., and Jonathan Kuminga. These players play a style to similar to the tradition forward position and primarily score in the paint around the rim. However, they are typically not relied on for scoring as these players are not the highest scorers on the team. However, they also have higher than average amount of dunks and can also provide a higher than average number of offensive rebounds. Some notable features from this cluster are, 2-point attempts, FG%, Dunks, Offensive rebounds, and percent of shots take between 0 and 3 feet.



For Offensive Big, the notable players are Joel Embiid, Gianni Antetokounmpo, and Nikola Jokic. These players play a similar style to the traditional center or power forward position; however, these players have above average offensive capabilities. They are usually relied on for their high scoring and can score from outside the 3-point line as well as from inside. Some notable features from this cluster are FG attempts, points, FG made, 2-point attempts, and 3-point attempts.



For Defensive Centers, the notable players are Jonas Valanciunas, JaVale McGee, and Deandre Ayton. These players are utilized for their size advantage and ability to defend the rim. This position is least likely to be playing on the outside during an offensive possession and will often set up in the paint around the rim. Furthermore, these players also have a higher than average turnover to assist ratio which means that they are less likely to be handling the ball and making plays. Some notable features from this cluster are average distance of shots, 3-point shots per 100 possessions, and 3-point shots per game.

These player positions differ from the positions determined by Cheng (2017). However, classification models are somewhat subjective as determining inputs such as number of clusters are up to the modeler's discretion. Moreover, it is also a subjective process when creating descriptions and labels for each of the clusters/newly created. Lastly, Cheng (2017) used data

from multiple and different years as well as included the Advanced Stats for each player. Therefore, it is expected that our clusters would be different. However, there are a few similarities in the positions such as Defensive Centers, Offensive Bigs would be equivalent to Offensive Centers, Scoring-playmaker Wing would be equivalent to Scoring Wings, Versatile Guard would be equivalent to Combo Guard, and Ball-handling Scorer would be equivalent to Floor Generals.

Recommendations, Limitations and Future Directions of Roster Construction

Using these new positions, we can then see which combination of positions leads to the most optimal 5-man lineup. Since the Warriors won the championship in the 2021-2022 season, I looked to see which 5-man line up produced the best +/- rating during the post season. Out of the 5 most commonly used 5-man lineups as determined by games played, the lineup consisting of Stephen Curry, Klay Thompson, Draymond Green, Andrew Wiggins, and Kevon Looney had the best +/- during the post season. We can see the position types for each of these players using the new classification model in Figure 13. From these initial results, this lineup has 2 Ball-handling Scorers. This could be redundant to have 2 of these position types in the same 5-man lineup. However, I suspect Draymond Green might have been misclassified which I will touch on shortly.

Stephen Curry	Ball-handling Scorer
Klay Thompson	Scoring-Playmaker Wing
Draymond Green	Ball-handling Scorer
Andrew Wiggins	Versatile Wing
Kevon Looney	Offensive Big

Figure 13. Warrior's best 5-man lineup during 2021-2022 post season

This analysis can be extended further to look at the top 5-man lineups for other teams to see what other combination of player types can also lead to success in today's game. The Warriors

are known to use a slightly smaller lineup as this works better for their style of play of quick ball movement and perimeter shooting. So, this could be the reason why there is no defensive center or inside scoring forward in this 5-man lineup. However, each coach and respective teams have different playing styles and game strategies, so it is important to note that this is not a one size fits all approach as a specific lineup combination may not work for all teams. With that said, we could potentially look at teams with similar playing styles and see which lineup combination from those teams are most successful using the classification model. Then, recommendations could be made to either mirror similar lineup constructions by these new player position types or seek to acquire players to help construct a similar lineup if the team currently does not have a particular player type on the roster. Additionally, the team can also use this information to scout potential replacements for a current player on the roster that may be leaving the team in the future due to retirement, trade, free agency etc.

There are several limitations to this current analysis. First, it was difficult to acquire the data as I had to resort to copy and pasting data from basketball-reference.com in CSV format into an Excel file. Then, I had to use the Text-to-Column feature to format the data into workable table. As this was time-consuming, I only included data from 2 seasons. However, ideally, I would have included data from additional seasons. Another limitation to this model was that due to needing to keep the position variable for the LDA process, some players were listed as different positions for different seasons which ended up creating duplicate records for players (one for each season). For example, LeBron James is listed as a PG in one year and a PF in another year. I did not want to delete duplicates, because I wanted to retain as many observations as possible after already removing players with less than half the season played. As such, I noticed that when I classified the players, LeBron James showed up in 2 different clusters. So, it is possible that a

player can change his playing style from year to year. This is another reason as to why I would like to include additional seasons so that we can look averages over a course of several seasons. As previously mentioned, the cluster labels were subjectively named based on the important features of each cluster. Players that are grouped into a particular cluster may not be an appropriate classification of that player all things considered. For example, Draymond Green was classified as a Ball-Handling Scorer. However, I think most experts would agree that this would be a bit of stretch to classify Draymond Green as that position type despite the result of the model. Another limitation is that determining the most successful 5-man lineup combination can be dependent on matchups which this analysis ignores. For example, the Warrior's 5-man lineup from Figure 13 does not have Defensive Center. However, if the opponent has an Offensive Big in the lineup, then the Warriors might consider inserting a Defensive Center into the lineup to defend against this particular player. As such, when applying this model to determine the optimal lineup, the team should also consider how the lineup would matchup against the opponent's lineup. Finally, I found it difficult to assess the importance each of the new positions as I did not include any advanced stats data in my analysis. In contrast, Cheng (2017) was able to use Average Win Shares by Position, Value Over Replacement Player by Position, and Player Efficient Rating (PER) by Position, to assess the value of and importance of each position relative to the other positions. In hindsight, I would have also included this data primarily for this purpose. I had initially chosen to omit this data from the model as I did not want to overcomplicate the model and I also felt that most of the advanced stats didn't describe how the player played the game.

As previously mentioned, I would include data from additional seasons as well as advanced stats for position evaluation purposes in future research. I would also spend some more time

cleaning the data to handle duplicate players that have multiple position designations. I would likely choose the most recent position and make that the player's position for all seasons. Lastly, I would perform a more in-depth analysis on each cluster to see if I can refine the cluster/position label. This part was also a bit challenging for me as a few of the clusters shared some similarities, so I think with some more exploration into each cluster along with some input from someone with more knowledge of the game would lead to more accurate cluster labels.

Overall, this analysis shows by using the Four-Factor Model there are certain factors of the game that are more important than other in terms of impact on games won. This allows the team to focus and prioritize the factors that contribute most to winning. We can also see how each team ranks for a particular factor and address areas that the team is weak in to increase the amount of games won. Additionally, the KMeans Clustering model suggests that there are more than 5 types of player positions in today's game. This modeling approach is able to classify players into position types that are a better representation of how each player plays the game. With this, teams can have a better understanding of types of players they have and can use this knowledge to optimize lineup constructions and also put players in a position to maximize their strengths and minimize their weaknesses. This analysis should ultimately improve the team's performance and hopefully lead to an increase in the number of wins for the team.

References

Cheng, Alex. 2017. “Using Machine Learning to Find the 8 Types of Players in the NBA.” *Medium*. March 9. https://medium.com/fastbreak-data/classifying-the-modern-nba-player-with-machine-learning-539da03bb824.