Dara Hashemi

NBA Player Performance Before and After COVID-19 Season

Motivation

The 2019-2020 NBA season, like many other things during the year 2020, was put on hold due to the COVID-19 pandemic. Fortunately, the season was able to resume later in the year for teams which were close to or had already qualified for the playoffs. A majority of teams' seasons were cut short, giving them one of the longest off-seasons in NBA History to rest, recover, vacation, spend time with family, etc. while the rest of the teams were given the opportunity to play beyond the regular season for a chance at a championship. While getting a chance to play in the NBA playoffs is an impressive achievement, the 2019-2020 playoffs came with a large burden on players. Teams who were in the playoffs, more specifically the Los Angeles Lakers and Miami Heat (who made it to the championship), had the shortest off-season in NBA History- around 72 days, whereas in the previous offseason, players were given double that time for recovery. Teams who didn't make the playoffs in the 2019-2020 season were given a 287-day off-season. For this project, I aim to compare the best NBA players' statistics and performance based on the 2019-2020 season, which consisted of a normal offseason beforehand, to the statistics of the 2020-2021 season, which consisted of the shortened or extended off-season depending on whether the team made the playoffs or not. Coming off a non-typical year which saw the NBA season cut short due to COVID, I am looking to see if there were positive or negative differences in player performance from the extended or shortened off-season.

Description of Data Sources

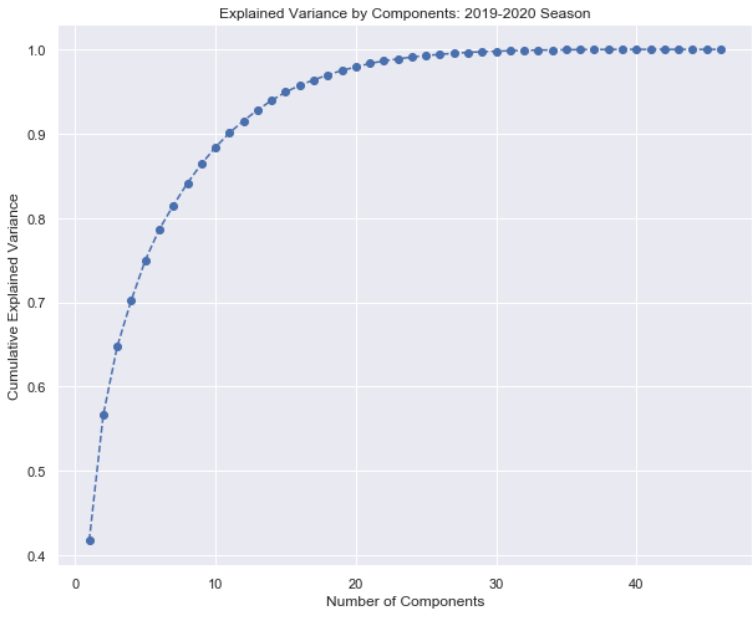
The data sources used consist of taking the statistics of 2019-2020 NBA players' and merging this data with players' advanced statistics coming off of a normal off-season (about 140 days of rest) from basketball-reference.com and comparing this data to the statistics of 2020-2021 NBA players and merging their data with advanced statistics from the same website as well. Both datasets have the exact same features associated with them, the only difference is that rows may be different due to new players entering the league or retiring between seasons, as well as the values for each player will be somewhat different between the two seasons. Basketball-reference.com is a great resource for anything basketball related (NBA, college, etc.) because it is always up to date with the most accurate basketball statistics and it includes many advanced metric calculations in addition to all the surface level statistics. Many analysts create

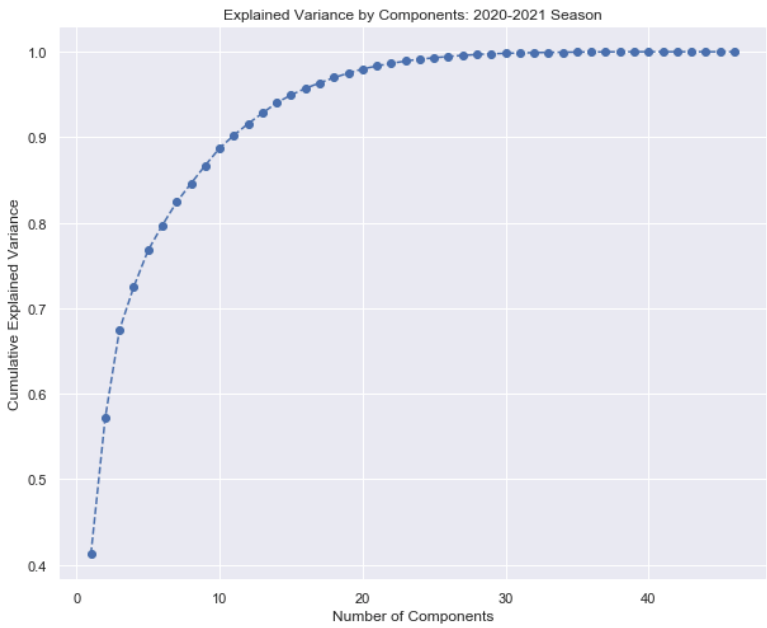
visualizations or player models from this website's data. The 2019-2020 dataset is 529 rows by 50 columns while the 2020-2021 dataset is 540 rows by 50 columns.

Analysis Performed

The project starts with scraping all the data and merging the proper datasets with each other, giving us two compete datasets to use. After cleaning some NaN's, NA's, and blank values throughout the data, I decide to cluster the players using K-Means Clustering and Principal Component Analysis. The reason I decided to cluster all the players in both seasons is because I want to group the most valuable players in the NBA together for each season. From there, I can analyze which individual players were in the top tier cluster for one of the season's and whether they were in the top tier cluster for the other season as well. After finding the top tier clusters of players, we can also execute more analysis on these groups specifically to see the differences in performance between the premier players of the league. Clustering also allows us to compare top players relative to each specific season, so if it was a down year for player performance in general, the clusters will be grouping players based on all players' performance for that season.

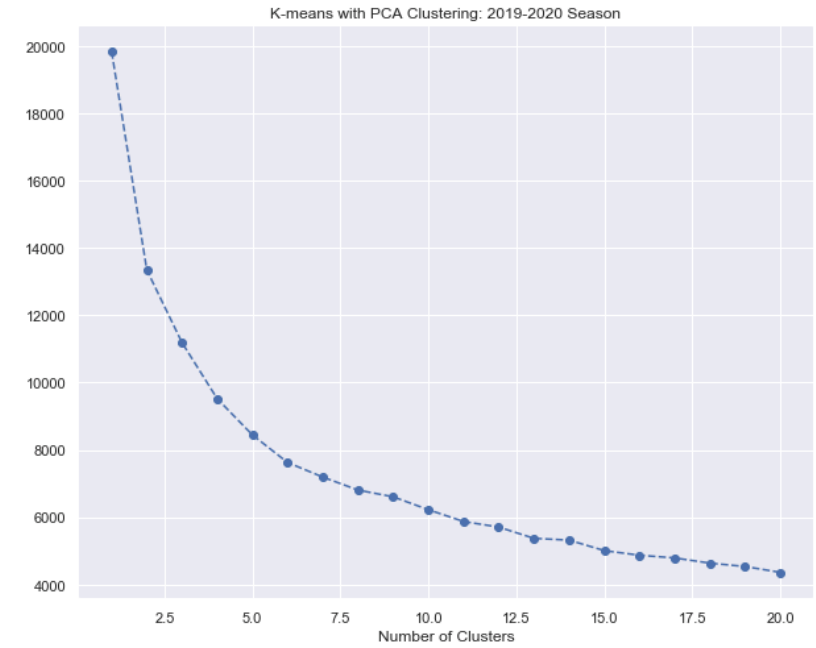
Once we have our data, we begin the clustering method in Python by standardizing our data to keep all our values in the same numerical range. After standardizing our data, we must reduce the dimensions of our data using Principal Component Analysis. We reduce the dimensionality of data because we will certainly have too many variables which have some correlation to each other, so by combining some of these features together, we can reduce the complexity of the clustering problem by reducing the amount of features. After running this, we need to decide how many features we'd like to keep based on a cumulative variance plot.

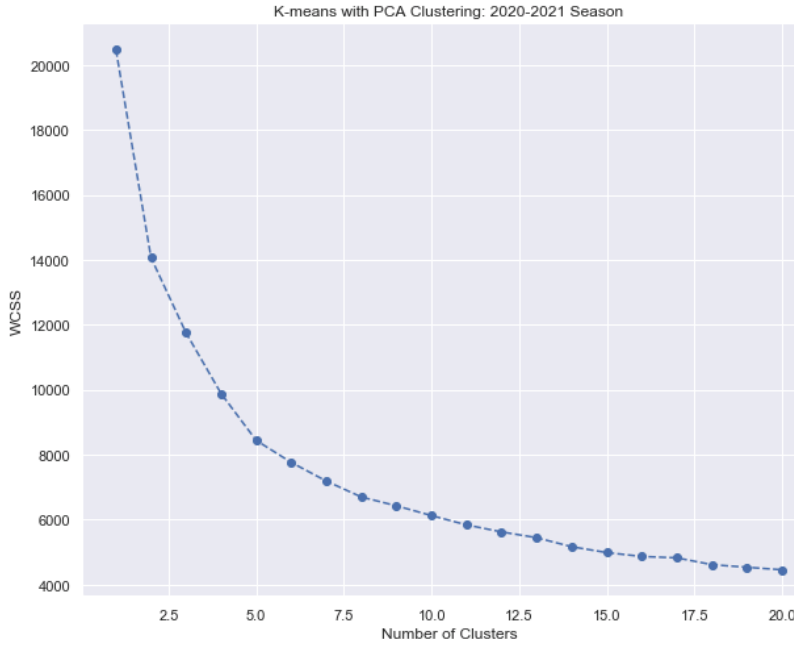




The plot above shows the amount of variance captured depending on the number of components we include. Being that generally we would like to keep around 80% of variance, we need to select the amount of components which gives is as little as possible but also hits the variance threshold. To also keep things constant between the two seasons, we chose 7 components for both datasets. Next, we calculate PCA with our number of components and obtain PCA scores we will use in the K-Means Clustering algorithm.

After finding the PCA scores, we test how many clusters would be ideal for our two models using the elbow-method. We are looking for the "elbow" of the graph which is where our line begins to level out. We would like to choose the least amount of clusters where the line is no longer declining at as steep of a rate.





We want the WCSS or Within-Cluster Sum of Square to be as low as possible, but we also would like a reasonable amount of clusters. Therefore, we find the area where the steep decline ends to choose our amount of clusters. This is somewhat arbitrary as well, so to keep analysis consistent we will use 6 clusters for both seasons. Once this is done, we calculate the K-Means algorithm with the PCA scores. This gives us our clusters.

Dara Hashemi

After we create our clusters of NBA Players for each season, we see that clusters are created based on player type, the multiple clusters include one for top tier talent, one for tall players who play mostly inside the three-point line, one for starter-type players, one for people who come off the bench, and last for players who don't get much playing time. Once analyzing this, we are only curious about the cluster with the top talent. After extracting the clusters with premier players from both seasons, we look to see who was in the cluster for the 2019-2020 season and who ended up not being in the top cluster for the 2020-2021 season. After finding individual players who were top tier for only one season, we dive into some individual cases to determine what statistics were bringing their overall performance down. We also calculate averages for both seasons' top tier players and see if there is any difference in performance between the two.

Conclusions

There were some effects of the shortened off-season just when looking at the surface level data. When analyzing these two clusters as a whole, there are some minor differences in the averages but nothing too drastic. The averages for all the regular statistics/metrics are the same for the most part, but the advanced statistics/metrics tend to lean a small amount in favor of the year with the shortened off-season. Metrics involving True Shooting Percentage, which calculates player efficiency in shooting the ball, is about 2% higher on average in addition to Player Efficiency Rating being about 2 points higher on average. Even usage for these top tier players went up close to 2% in the year of the shortened off-season. These are minimal differences and there is likely no significant difference between the two. When looking at averages of top tier players for each season, there was not much of a difference in performance.

Players who the NBA love to market in Jimmy Butler, LeBron James, and Anthony Davis all sustained some form of longer term injury during the 2020-2021 season and were forced to sit out a majority of the season. These three players also happened to play the most minutes in the playoffs last season, as they led both their teams to the championship. There are also more players in the top tier cluster who suffered long-term injury following the shortened off-season, effectively leading them out of the top tier cluster for the 2020-2021 season including Jamal Murray and T.J. Warren. For those who have very little injury history prior to the 2021 season, it

is worth noting the fact that these players had half the time to rest and recover than they normally would. There were multiple injuries in the season following a shortened offseason. An NBA season is usually 82 games long which spans over 7-8 months of these athletes putting heavy burden on their body. There needs to be proper time for players, especially those the NBA loves to market their brand with (and who ultimately end up finishing their season the latest every year), to rest and recover before the new season begins. While there does not seem to be any dip in general player performance from the top tier talent, my biggest finding was that there was a higher risk of injury after the condensed off-season. The 2020-2021 season's top tier players, based on our model's clustering, missed 551 of a possible 2,880 games (19%), whereas the 2019-2020 season's top tier players, based on our clustering model, missed 678 of a possible 4,060 games (16%). And according to the Elias Sports Bureau, the average number of players sidelined per game due to injury, non-COVID-19 illness or rest this season was 5.1, the highest since it started being tracked in 2009-10. That does not include games missed by players in the health and safety protocols. The next highest season was 4.8, so 2020-21 was 5% higher. In addition, this season's All-Stars missed 370 of a possible 1,944 games (19%), the highest percentage in a season in NBA history (Elias Sports Bureau). A long-term injury to a star player can cost the NBA millions of dollars in revenue, so it is in the best interest of the league, AND their players, to allow athletes the proper amount of time to recover from the season.