Ibrahim Hassanain,Tszchung Cheung, Amrit Singh,
Hamzah Saleh, Lifu Tao, Shazid Rahman
City College of New York
Professor Retemadpour
23 April 2021
Group Assignment 2

Group Assignment PCA

Link to Graph:
https://colab.research.google.com/drive/1RhUP5PjFQXeQzQz_xCqM_z-z4Oh0j29g?usp=sharing
Link to Github: https://github.com/asingh2066/Visualization---PCA-Group-Assignment-2-.git
Link to Code: https://colab.research.google.com/drive/1RhUP5PjFQXeQzQz_xCqM_z-z4Oh0j29g?usp=sharing

Introduction:

The overall goal of our data was to create a 2 component PCA on NBA player data based on the shots taken, and each player's mean performance. Taking this data, appropriate cluster groups, and PCA values were calculated. In our assignment, we decided to use the R example you provided as well as other sources to create a 2 component PCA visual. The reason for creating a 2 component PCA was because our variance for the **2 PCA's was totaled to be above 70%**. We approached this conclusion since we know that for the PCA, we want to use the least amount of components possible to explain the most amount of variance. After calculating and graphing the principal and cumulative components, we figured out that our proportion of Variance to PCA was 7. However, after calculating the variance for each individual PCA value, we realized that **the variance for PC1, PC2, and PC3 totaled up to around 93%.** This value is well above the 70% mark, meaning the data from the other PCA components is not significant to us when it comes to keeping mean data. The 7 clusters were groups that were assigned to each NBA-player based on their mean statistical performance per season. Each player used in the data for the chart was assigned a PCA value based on their performance, and that PCA value was assigned a group based on if it fell within the range for the cluster group.

After acquiring the PCA and variance, we moved on to clustering our data to see how distinct groups we can create for NBA players from the years 2016- 2019. We decided to use k-mean inorder to find the clusters, after graphing we learned that using 7 clusters worked best when it came to separating the data. After creating a 2 component PCA visual, we decided that this worked fine with our data since the clusters were distinct. We did not need a 3D visual since our **PC1 and PC2 variance was above 75%**, meaning this was enough to represent our data. The rest of the documentation is just some images from the code, and small explanations for each step. The assignment focuses on using principal component analysis and clustering on NBA player data from 2013-2019 to figure out the kinds of players that exist (when it comes to offense), in the NBA right now. The dataset attempts to fully describe the kinds of shots players take and the way they end up making their points. The dataset includes every player stepped on the court for gameplay for every season from 2013-2019. The seven clusters are just the different groups of performance, each cluster is based on mean performance, so out of all the NBA players in the seasons from 2016-2019, you can categorize their offensive performance into 7 groups. With each NBA-player averaging in their own group. Each point in the cluster represents a different player based on which PCA value the cluster averages fall within.

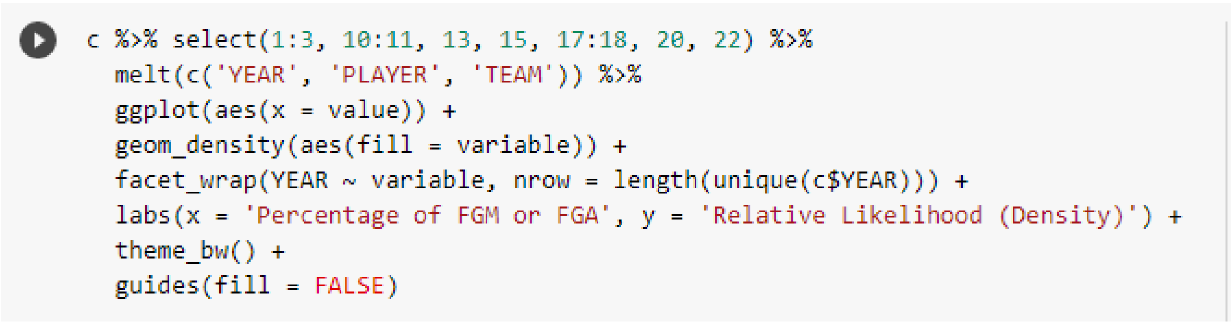
Determining Usable Data: (Figure 1 below)

Initially, we wanted to use all the data we have, but realized that we need to filter and clean the data if we wanted to have distinctive clusters. Inorder to achieve this, we had to see if the data for every season was usable. This depended on how the sport has evolved over time when it comes to the level of offense players bring to the court. We wanted to ensure that we use the maximum amount of data, but also ensuring that the data is averaged throughout the years. After doing some reading, we decided to create a histogram to compare the different statistics we have for each season. From left to right, the **histograms depict**: % of Field Goal Attempts That are 3PT Shots, % of Points That Come From 2PT Shots, % of Points That Come From 3PT Shots, % of Points That Come From Free Throws, % of Points That Occur In The Paint, % of 2PT Field Goals Made That Were Assisted, % of 3PT Field Goals Made That Were Assisted, and % of Total Field Goals Made That were Assisted.

Looking at the histogram, the majority of the distributions remain similar for each season, such as % of 2PT FG's assisted or % of PTS in the paint. However, a few distributions, such as the **3PT-related distributions start to change.** This showcases the shifting emphasis of NBA offenses. Looking back at the 2013-2014 season of the dataset, the % of FGA's that are 3PT shots (the first distribution, shown in red) and % of PTS coming from 3PT shots (the third distribution, shown in green) both are at 0. However, looking at 2016-2017, players were more likely to have taken 30-40% of their shots from the 3PT range, as the density in that range is greater than the density at 0. For the first time in the dataset, players were equally likely to have

25-30% of their points come from 3PT shots, compared to before where 0% was from the 3 PT range. During the more recent seasons (2017-2019), players were more likely to have 50% of their FGA come from 3PT range than they were to have 0% of their FGA come from that range.

Something else to note is that the distribution of % of made 3PT shots that were assisted (the second distribution from the right, shown in purple). In the 2013-2014 data, there's a gradual increasing slope from 50% of made 3PT field goals assisted to 100% of made 3PT field goals assisted, whereas in 2018-2019, this increase is much steeper, demonstrating the increased number of players who are assisted on almost all of their 3PT field goals. This will be further looked at in one of the clusters we identify later on in the analysis. After looking at this data, we concluded that the years **2016-2019 had similar data, so we filtered out the previous seasons.**



```
c %>% select(1:3, 10:11, 13, 15, 17:18, 20, 22) %>%
  melt(c('YEAR', 'PLAYER', 'TEAM')) %>%
  ggplot(aes(x = value)) +
  geom_density(aes(fill = variable)) +
  facet_wrap(YEAR ~ variable, nrow = length(unique(c$YEAR))) +
  labs(x = 'Percentage of FGM or FGA', y = 'Relative Likelihood (Density)') +
  theme_bw() +
  guides(fill = FALSE)
```

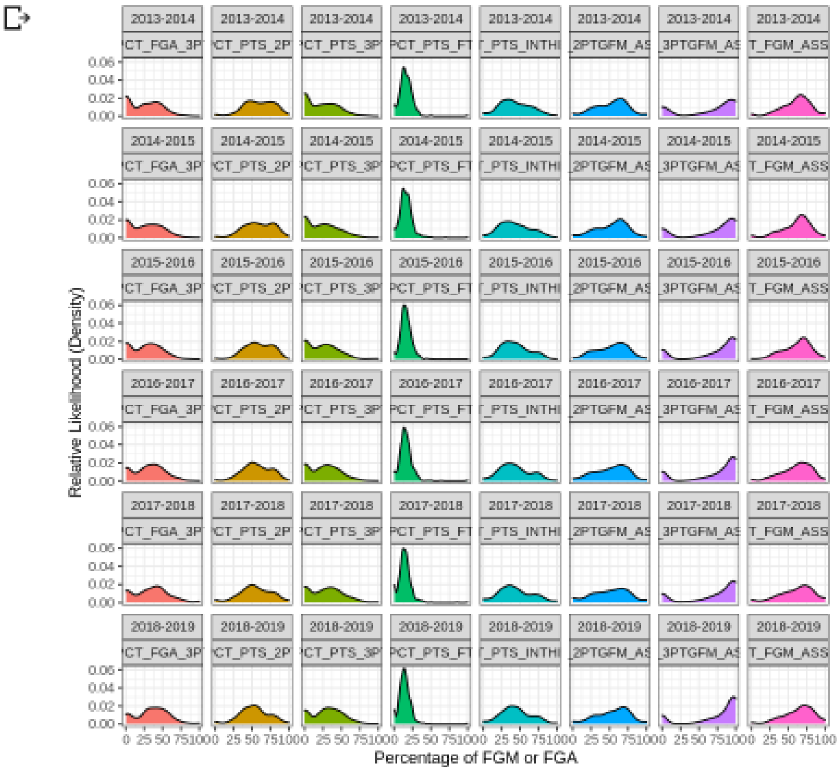
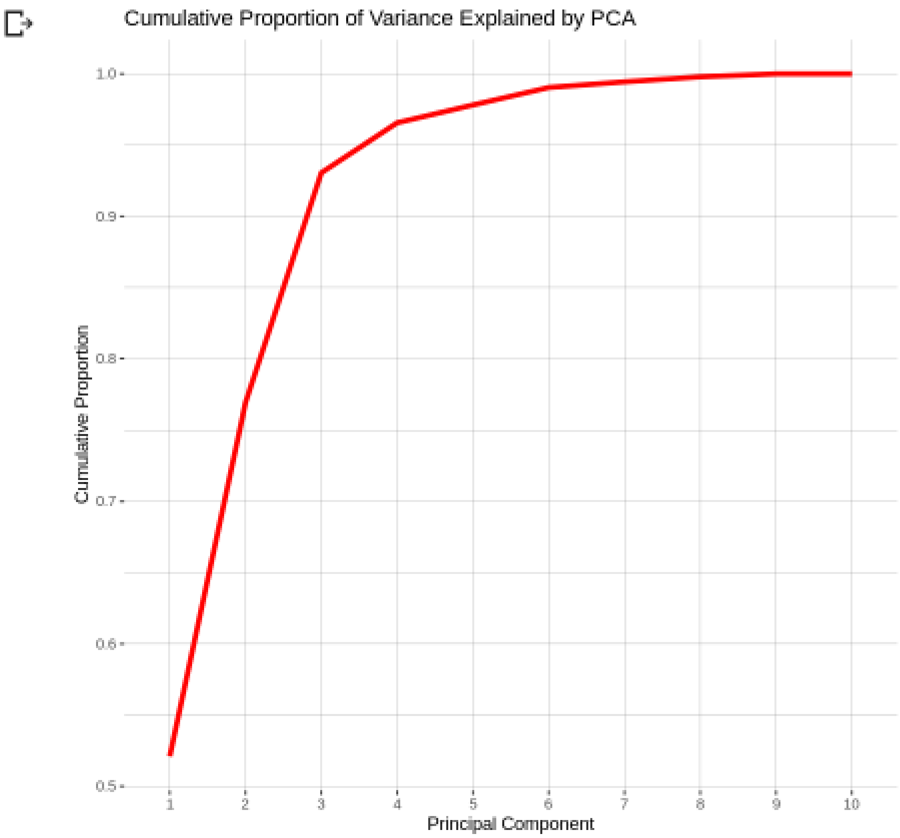


Figure 1: ***Histogram of various data from the years 2013-2019***

Principal Component Analysis:

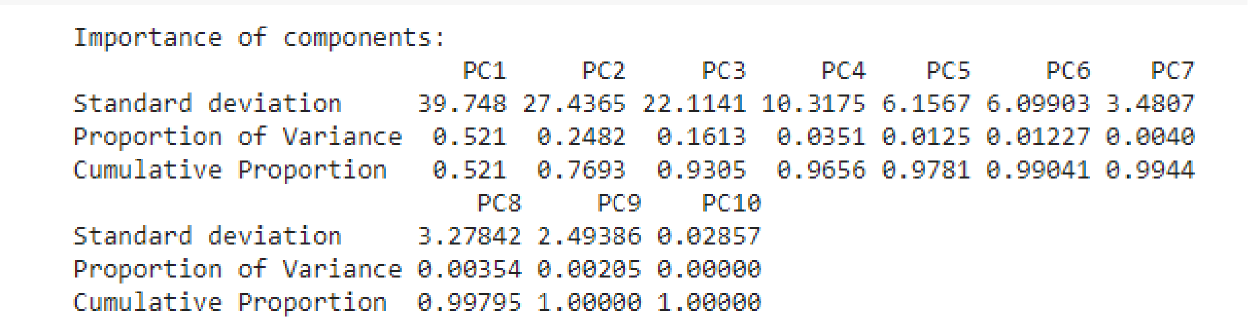
The purpose of Principal component analysis is to simplify a dataset by compressing the information found in a large number of variables into a smaller number of variables. Each variable in a dataset represents a dimension in space (if your dataset has 3 variables, it would correspond with 3D space). In our case we decided to use 2 variables, so the direction would take the form a * var1 + b * var2). If one component can accurately summarize the dataset by sufficiently minimizing the average squared distance, we can reduce all of the data points to a single value, meaning we can substitute the 3 variables into the linear combination, making a three-dimensional point into a one-dimensional point.



Graph 1: ***Proportion of Variance to PCA***

PCA Analysis of NBA Shot Type Dataset

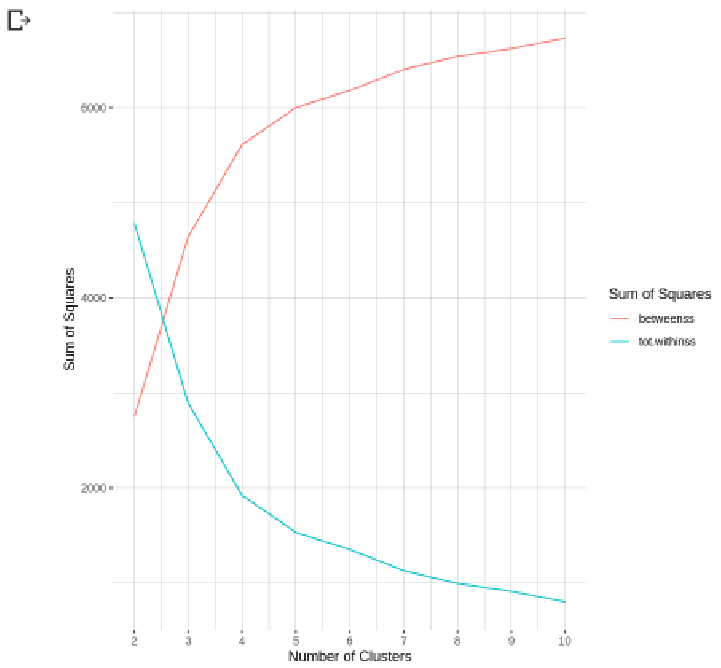
In PCA, we want to use the least amount of components possible to explain the most amount of variance. This is the table that represents different values for each PCA, specifically the variance for each amount of information captured by each added component, we only need the first two components since their variance added is above 70%.



```
Importance of components:
                           PC1     PC2     PC3     PC4    PC5     PC6    PC7
Standard deviation     39.748 27.4365 22.1141 10.3175 6.1567 6.09903 3.4807
Proportion of Variance  0.521  0.2482  0.1613  0.0351 0.0125 0.01227 0.0040
Cumulative Proportion   0.521  0.7693  0.9305  0.9656 0.9781 0.99041 0.9944
                           PC8     PC9    PC10
Standard deviation     3.27842 2.49386 0.02857
Proportion of Variance 0.00354 0.00205 0.00000
Cumulative Proportion  0.99795 1.00000 1.00000
```

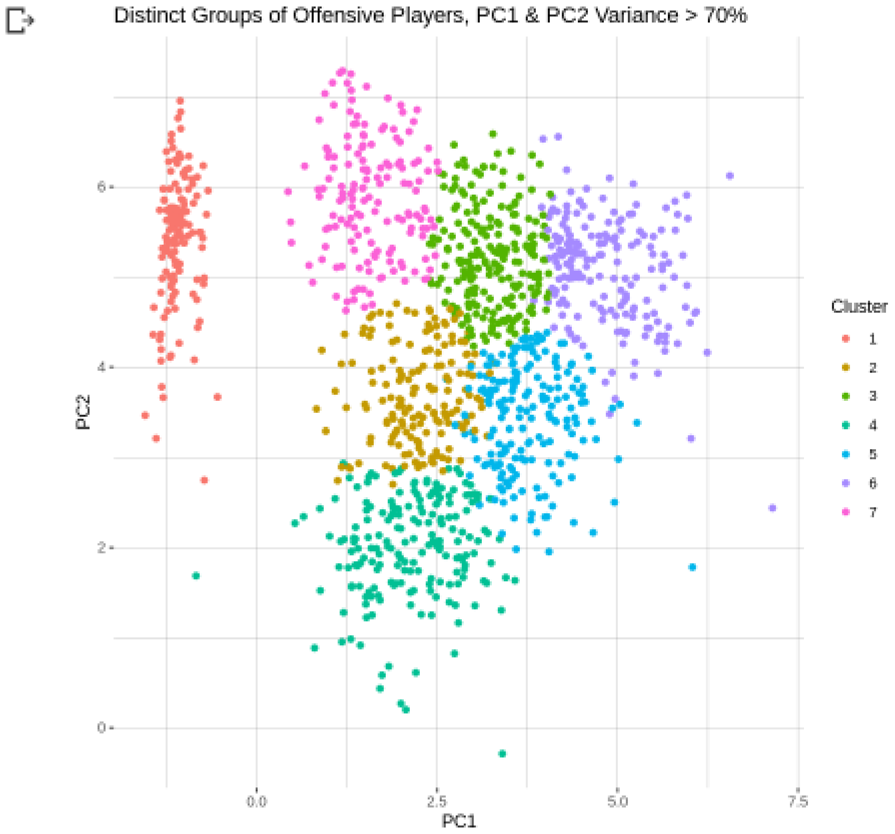
K-Means Clustering Analysis:

Now that we have our simpler and nearly informative dataset, we can begin our clustering work. We want to figure out how many distinct groups of offensive players truly exist in the NBA, using the data we have. This is done by plotting the sum of all the cluster variances we want to minimize the value as much as possible, this would allow our groups to be consistent. We also want the cluster variance in between to maximize so that our clusters are not overlapping and are distinct against the number of clusters we specify. We want to balance the two kinds of variances so we use a graph to approximate the optimal number of clusters.



Graph 2: ***The blue line shown here shows the sum of the groups within the cluster variances for each number of clusters given, while the red line gives us the between-cluster variance.***

Two-Component Visualization:

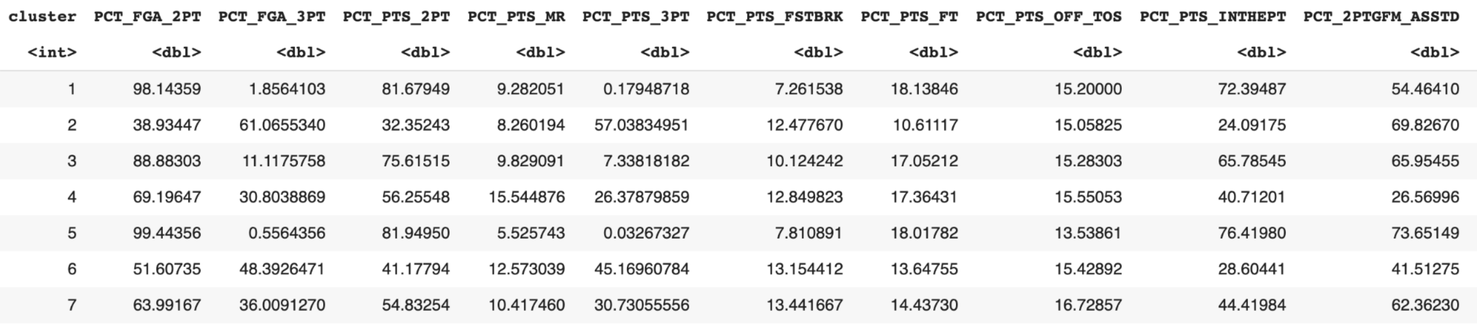


Graph 3: ***The graph shows the calculated 7 clusters for the 2 PCA components.***

Conclusions:

You can conclude from this graph that there are 7 different range groups for offensive performance in the NBA. Looking at the two component PCA produced above, after doing the calculations, this was the most optimal way to approach a solution. We decided to have 7 clusters, each group representing a cluster of players. Each cluster represents a different range of mean statistics for NBA players. Based on these mean statistics, each NBA is given a distinct cluster, with very few outliers. The seven clusters are just the different groups of performance, each cluster is based on mean performance, so out of all the NBA players in the seasons from 2016-2019, you can categorize their offensive performance into 7 groups. With each NBA-player averaging in their own group. Each point in the cluster represents a different player based on which PCA value the cluster averages fall within. This means, after averaging the stats for every NBA player they are assigned a PCA value, and based on that PCA range they assigned a cluster group that is closest to that mean value (correlates to that PCA range). Each cluster group is just based on mean statistical performance of the players, the table for this can be seen in the assignment file. Looking at the graph above, we did not need to create a three dimensional graph since the variance of PC1 and PC2 was greater than 70%.

You can view the table below to see the different mean values for each cluster group. To see the full table look at the assignment file. And to see the table which tells you which player received which group based on performance, you can view the file.



cluster	PCT_FGA_2PT	PCT_FGA_3PT	PCT_PTS_2PT	PCT_PTS_MR	PCT_PTS_3PT	PCT_PTS_FSTBRK	PCT_PTS_FT	PCT_PTS_OFF_TOS	PCT_PTS_INTHEPT	PCT_2PTGFM_ASSTD
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	98.14359	1.8564103	81.67949	9.282051	0.17948718	7.261538	18.13846	15.20000	72.39487	54.46410
2	38.93447	61.0655340	32.35243	8.260194	57.03834951	12.477670	10.61117	15.05825	24.09175	69.82670
3	88.88303	11.1175758	75.61515	9.829091	7.33818182	10.124242	17.05212	15.28303	65.78545	65.95455
4	69.19647	30.8038869	56.25548	15.544876	26.37879859	12.849823	17.36431	15.55053	40.71201	26.56996
5	99.44356	0.5564356	81.94950	5.525743	0.03267327	7.810891	18.01782	13.53861	76.41980	73.65149
6	51.60735	48.3926471	41.17794	12.573039	45.16960784	13.154412	13.64755	15.42892	28.60441	41.51275
7	63.99167	36.0091270	54.83254	10.417460	30.73055556	13.441667	14.43730	16.72857	44.41984	62.36230