CISC 5950 Big Data Programming Project 1

Weifeng Yan, Alma Marko, Jiarui Zhang

Part 2: NBA Shot Logs

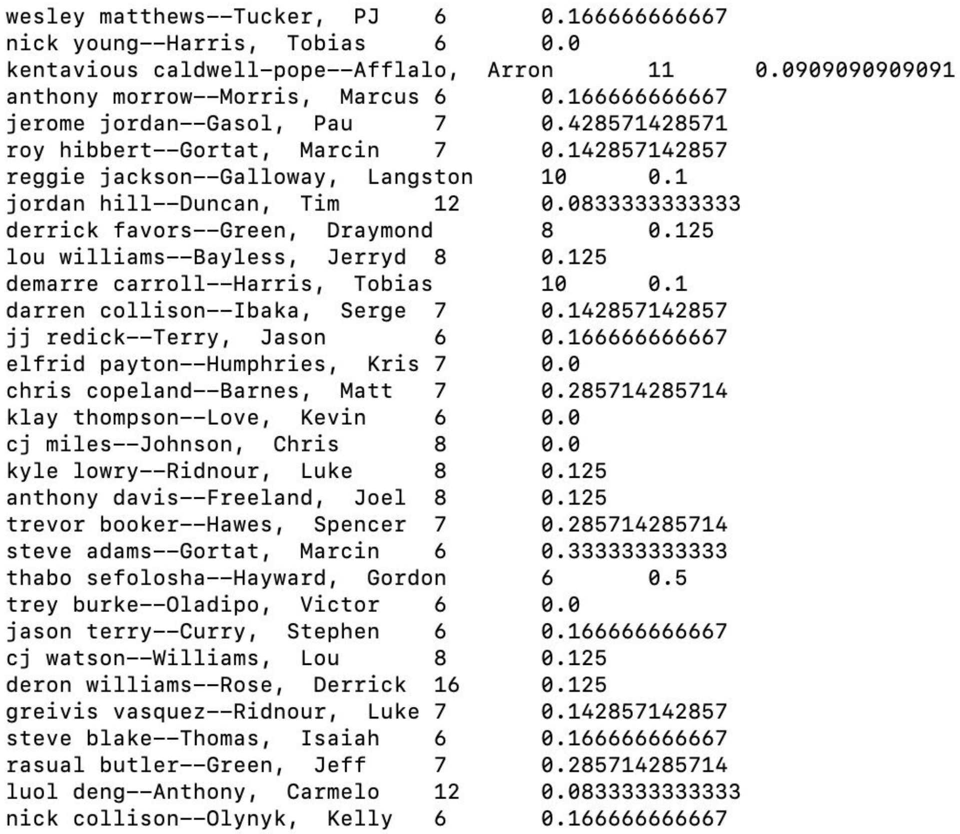
DATA:

In this part, we are working with data on shots taken during the 2014-2015 NBA season. The data set contains 21 columns and 128,071 rows. In this project, we will be mainly exploring the information from the following columns: player_id, player_name, CLOSEST_DEFENDER, CLOSEST_DEFENDER_PLAYER_ID, CLOSE_DEF_DIST, SHOT_CLOCK, SHOT_DIST, and SHOT_RESULT.

QUESTIONS

Question 1: Fear Score

For the first question, we are trying to find each player's most unwanted defender. To do this, we need to calculate the fear score for each player and defender pair. The formula we use to calculate the fear score is the number of shots with the result "made" divided by the total number of shots attempted with a specific defender. The lower the fear score, the more unwanted the defender is to the player who is shooting because the shooter is more likely to miss the shot. To solve this problem, we used two rounds of MapReduce. The first Mapper takes the csv file as input and outputs a key value pair where the key is the player-defender pair and the value is a tuple consisting of the count and running average. The first Reducer takes the mapper's output and updates the count and average hit rate for player-defender pair. It then returns a key value pair where the key is the player-defender pair and the value is a tuple of the total count and the hit rate. The second round of MapReduce is responsible for finding a player's most unwanted defender. The second Mapper takes the output from the first round of MapReduce as the input. The Mapper filters through and only prints a line of data if its total count of shots taken for a defender is greater than 5. We noticed that there are defenders where the player missed even though they only encountered the defender once, so to avoid any outliers, we chose to find the most unwanted defender by placing a condition that the player must have attempted more than 5 shots against a defender. The second Reducer then takes the mapper's output and finds each player's most unwanted defender and prints it out. The following is a portion of what the output would consist of. Essentially, the output is a player and the player's most unwanted defender along with its fear score.



```
wesley matthews--Tucker,  PJ    6       0.166666666667
nick young--Harris,  Tobias     6       0.0
kentavious caldwell-pope--Afflalo,  Arron       11      0.0909090909091
anthony morrow--Morris,  Marcus 6       0.166666666667
jerome jordan--Gasol,  Pau      7       0.428571428571
roy hibbert--Gortat,  Marcin    7       0.142857142857
reggie jackson--Galloway,  Langston     10      0.1
jordan hill--Duncan,  Tim       12      0.0833333333333
derrick favors--Green,  Draymond        8       0.125
lou williams--Bayless,  Jerryd  8       0.125
demarre carroll--Harris,  Tobias        10      0.1
darren collison--Ibaka,  Serge  7       0.142857142857
jj redick--Terry,  Jason        6       0.166666666667
elfrid payton--Humphries,  Kris 7       0.0
chris copeland--Barnes,  Matt   7       0.285714285714
klay thompson--Love,  Kevin     6       0.0
cj miles--Johnson,  Chris       8       0.0
kyle lowry--Ridnour,  Luke      8       0.125
anthony davis--Freeland,  Joel  8       0.125
trevor booker--Hawes,  Spencer  7       0.285714285714
steve adams--Gortat,  Marcin    6       0.333333333333
thabo sefolosha--Hayward,  Gordon       6       0.5
trey burke--Oladipo,  Victor    6       0.0
jason terry--Curry,  Stephen    6       0.166666666667
cj watson--Williams,  Lou       8       0.125
deron williams--Rose,  Derrick  16      0.125
greivis vasquez--Ridnour,  Luke 7       0.142857142857
steve blake--Thomas,  Isaiah    6       0.166666666667
rasual butler--Green,  Jeff     7       0.285714285714
luol deng--Anthony,  Carmelo    12      0.0833333333333
nick collison--Olynyk,  Kelly   6       0.166666666667
```

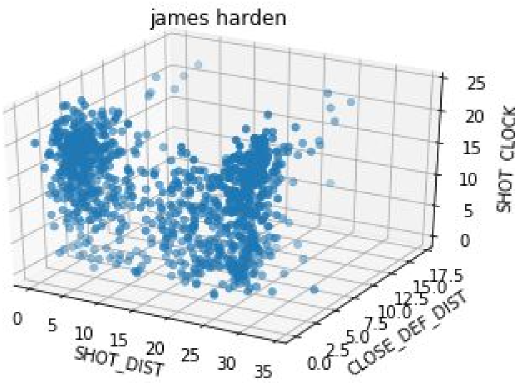
Question 2: Comfortable Zones

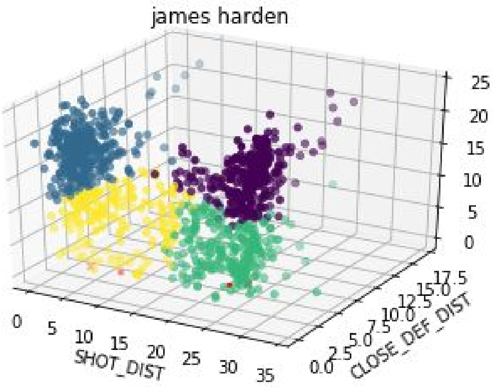
For the second question, we are trying to find a player's four comfortable zones for shooting. A comfortable zone is defined by the following matrix:

{SHOT DIST, CLOSE DEF DIST, SHOT CLOCK}

To find a player's four comfortable zones, we use K-Means. This problem requires us to build multiple MapReduce tasks. First, we would need to perform MapReduce to obtain a subset of the entire dataset that consists only of data for a given player, such as Lebron James. Next, we would randomly select four rows of data from the player-specific dataset. The three variables that define the matrix would be extracted from the four rows to create our initial four zone clusters centers. Using the four initial centers, we would run a MapReduce algorithm, where the Mapper assigns each data point from the dataset to one of the four zone clusters. The assignment is based on which of the four centers it is least farthest from. The Reducer would finally recompute the centers for each zone cluster by averaging SHOT DIST, CLOSE DEF DIST, and SHOT CLOCK with the data points that belong to the zone cluster. We repeat this last process of MapReduce either 10 times or until we see that the previous centroids are the same as the new centroids. Once we complete the MapReduce for K-Means, we would have to apply another round of MapReduce to find the best zone for a player considering his hit rate. We were able to find the best zones for James Harden, Chris Paul, Stephen Curry, and Lebron James.

1. James Harden





Summary:

Below, you will find the four cluster centers and the hit rate for each cluster. Please note that a row in the hit rate table corresponds to the same row number in the cluster center table. Essentially, the first hit rate in the table is the hit rate for the cluster associated with the first cluster center in the cluster center table. The row with the highest hit rate in the hit rate table for James Harden is the second row (0.563333333). This means that the best zone for James Harden to shoot from is in the zone with the center {4.01733333, 2.768, 17.891}. James Harden's best zone is the blue cluster shown above in the right 3D graph to the right. It can be inferred from the values of the matrix that Harden is more likely to make a successful shot if he is standing closer to the hoop. The zone we specified that was best for him, is the only one with an average hit rate greater than 50%.

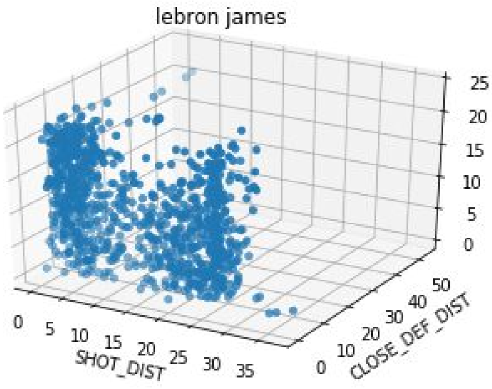
<u>Cluster Center:</u>

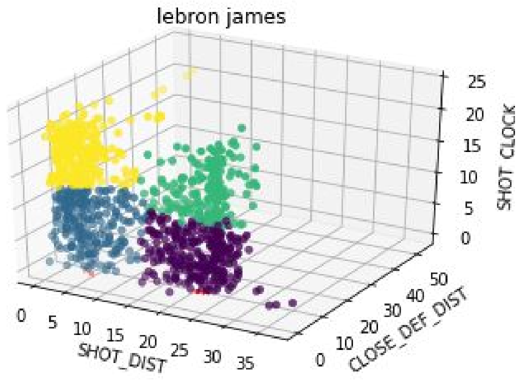
```
'SHOT_DIST', 'CLOSE_DEF_DIST', 'SHOT_CLOCK'
    [[23.81888112,  5.04020979, 16.26643357],
     [ 4.01733333,  2.768     , 17.891     ],
     [22.12163265,  4.24489796,  5.47795918],
     [ 8.30971429,  2.836     ,  8.56971429]])
```

<u>Result</u>:

```
        Hit_Rate
0       0.335664336
1       0.563333333
2       0.428571429
3       0.462857143
```

2. Lebron James





Below, you will find the four cluster centers and the hit rate for each cluster. The row with the highest hit rate in the hit rate table for Lebron James is the fourth row (0.659259259). This means that the best zone for Lebron James to shoot from is in the zone with the center {5.11148148, 4.0237037, 18.63740741}. Lebron James's best zone is the yellow cluster shown above in the right 3D graph to the right. It can be inferred from the values of the matrix that Lebron is more likely to make a successful shot if he is standing closer to the hoop, similar to James Harden.

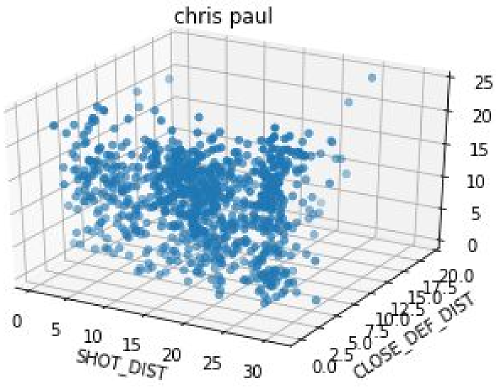
Cluster Center:

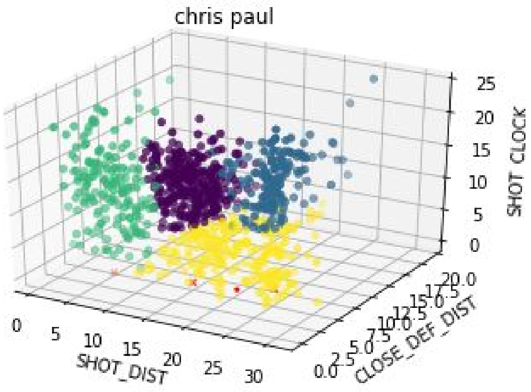
```
'SHOT_DIST', 'CLOSE_DEF_DIST', 'SHOT_CLOCK'
    [[22.42037037,  4.6662963 ,  5.73777778],
     [ 6.44912281,  2.71140351,  8.09035088],
     [23.29106145,  5.59664804, 15.7150838 ],
     [ 5.11148148,  4.0237037 , 18.63740741]])
```

Result:

```
        Hit_Rate
0       0.359259259
1       0.535087719
2       0.396648045
3       0.659259259
```

3. Chris Paul





Below, you will find the four cluster centers and the hit rate for each cluster. The row with the highest hit rate in the hit rate table for Chris Paul is the first row (0.547619048). This means that the best zone for Chris Paul to shoot from is in the zone with the center {15.52040816, 4.41292517, 14.99013605}. Chris Paul's best zone is the purple cluster shown above in the right 3D graph to the right. It is a bit difficult to determine what makes Paul more successful from the values of the matrix because there aren't any trends at first sight. The averages are scattered to maintain a relatively steady hit rate. This might be due to the fact that Paul is a traditional player.

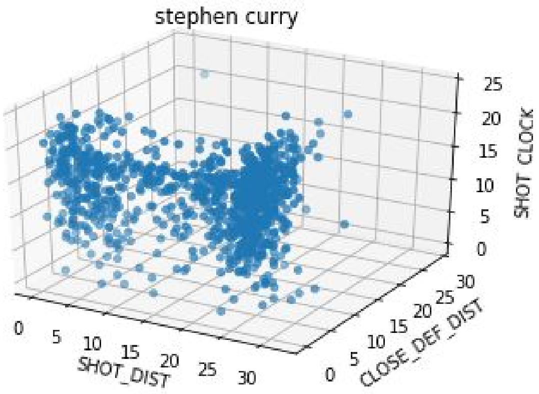
Cluster Center:

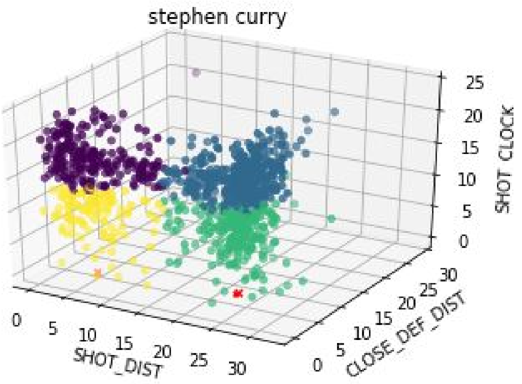
```
'SHOT_DIST', 'CLOSE_DEF_DIST', 'SHOT_CLOCK'
   [[15.52040816,  4.41292517, 14.99013605]
    [24.1,         6.2006135,  16.4196319 ]
    [ 6.39075145,  3.36300578, 13.17572254]
    [20.79230769,  4.73393665, 5.5479638 ]]
```

Result:

```
        Hit_Rate
0       0.547619048
1       0.441717791
2       0.514450867
3       0.407239819
```

4. Stephen Curry





Below, you will find the four cluster centers and the hit rate for each cluster. The row with the highest hit rate in the hit rate table for Stephen Curry is the first row (0.61965812). This means that the best zone for Stephen Curry to shoot from is in the zone with the center {5.85384615, 3.33504274, 17.80811966}. Stephen Curry's best zone is the purple cluster shown above in the right 3D graph to the right. It can be inferred from the values of the matrix that Curry is more likely to make a successful shot if he is standing closer to the hoop, similar to James Harden and Lebron James.

Cluster Center:

```
'SHOT_DIST', 'CLOSE_DEF_DIST', 'SHOT_CLOCK'
    [[ 5.85384615,  3.33504274, 17.80811966]
     [23.75529101,  5.41798942, 18.05582011]
     [23.64886878,  5.01040724,  9.45294118]
     [ 6.27407407,  2.75185185,  9.22777778]]
```

Result:

```
        Hit_Rate
0       0.61965812
1       0.444444444
2       0.389140271
3       0.601851852
```