Report Assignment 2

NBA Data Analysis

Clustering, Linear Regression, Panel Data

Emery Ong A0136591B
Emile Brès A0132365L
Simon Helmlinger A0134470M
Juan Manuel Muñoz Perez A0134739X

Monday 30th March, 2015

Clustering. Which players are similar? (15 points)

Introduction Using the `stats` library in R, the purpose of this part is to determine the 'closest' players thanks to the kmeans algorithm. We should not remember the main idea of the NBA analysis started in the assignment 1: *explain the factors that influence the player's salary.* After defining 'closest', we will explain the approach used in order to conduct the analysis.

In the context of NBA players, two players are close if their statistics (`weight`, `height`, `age`, `experience` and about the games such `wins`, `forward`, `steal`, and so on) are similar. The following aims to cluster the current NBA active players in order to understand the characteristics of the differents groups and how it influence their salary.

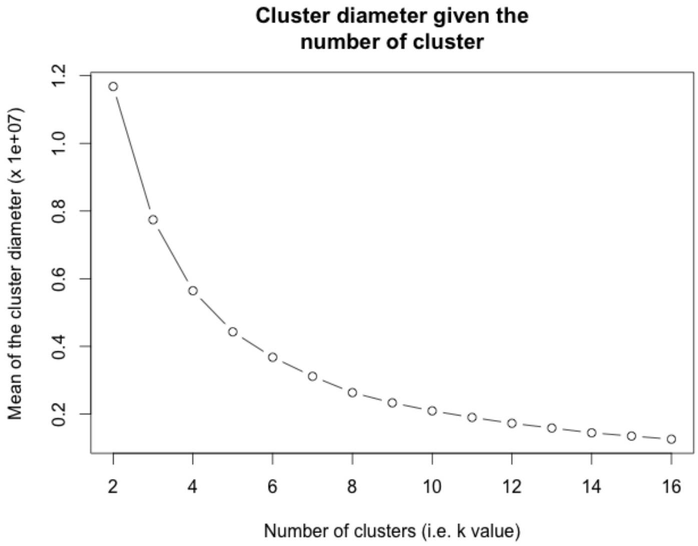
Data The first step to conduct the analysis is to build the dataset. To do that, we used the data scraped during the assignment 1 as follows:

- Filtering
  1. extract the profile of the active players into the `active_player_profile` dataframe(attributes: `PlayerID`, `name`, `shoots`, `weight`, `height`, `dob`, `birth_city`, `birth_state`, `experience` and `age`)
  2. extract the most recent salary recorded for the active players into the `active_salaries` dataframe (attributes: `PlayerID`, `Season`, `Team`, `FranchiseID` and `Salary`)
  3. extract the totals statistics for the current active players into the `active_totals_final` dataframe (attributes: `PlayerID`, `Season`, `Age`, `FranchiseID`, `Lg`, `Pos`, `G`, `GS`, `MP`, `FG`, `FGA`, `FG%`, `X3P`, `X3PA`, `X3P%`, `X2P`, `X2PA`, `X2P%`, `eFG%`, `FT`, `FTA`, `FT%`, `ORB`, `DRB`, `TRB`, `AST`, `STL`, `BLK`, `TOV`, `PF` and `PTS`)
- Merging
  1. merge `active_player_profile` with `active_salaries` into the `player_information_inter` dataframe
  2. merge `player_information_inter` with `active_totals_final` into the `player_information` dataframe

The `player_information` dataframe contains 600 active players but at the end the kmeans algorithm is applied to only 539 active players. Indeed, the dataset is build so that for each active player we keep its last salary recorded - note that the salary of the current season can be missing - with its corresponding totals statistics for the same team and season - some players change team during the season and so have two records in salaries and teams for the same season.

Attributes chosen to explain the player's salary The salary is influenced directly by the `experience` of the player and by all the statisticals attributes that gather data about the player's performance (see `active_totals_final` for the list).

kmeans algorithm The main issue in dealing with the kmeans algorithm is the difficulty in finding the optimal number of centrois (`k`). In order to find the better parameter we use the cluster-diameter mean analysis. The figure **??** below show the result for our dataset.



From that figure we cannot read obviously the optimal `k`. A deep analysis is done. The idea is to compute the divergence of the slope from the `k`$^{\text{th}}$ to the `k+1`$^{\text{th}}$. clusters. The table below shows the results computed.

`k`	`diameter`	`slope`	`variation` (%)	`divergence` (%)	`diff_div` (%)
2	11680914	0.00	0.000000	0.00000	0.00000000
3	7742893	-3938021.36	-Inf	0.00000	0.00000000
4	5644795	-2098098.18	46.722021	46.72202	46.72202131
5	4429310	-1215484.24	42.067333	69.13465	22.41262459
6	3677566	-751744.45	38.152678	80.91061	11.77595914
7	3113996	-563570.26	25.031670	85.68900	4.77839436
8	2633222	-480773.84	14.691410	87.79149	2.10248776
9	2330790	-302431.24	37.094905	92.32022	4.52873621
10	2093675	-237115.06	21.597036	93.97883	1.65860413
11	1899240	-194435.25	17.999620	95.06262	1.08378818

The `variation` is the % of variation between the `k`$^{\text{th}}$ and the `k-1`$^{\text{th}}$ slopes. The `divergence` is the variance of each slope from the first one. Finally, the `diff_div` is the variation between the `k`$^{\text{th}}$ and the `k-1`$^{\text{th}}$ of the `divergence`.

Theses numbers conduct us to choose $k = 6$. Indeed, for that value, the slope is up to 81% different from the first slope. For $k = 7$, the difference is of 85.7%. Thus, the variance of the `divergence` start becoming insignificant (about 4.7% compare to previous which are about > 12%).

Cluster centroids With $k = 6$ and the previous attributes chosen, the kmeans algorithm output the same centers (it is 'stable'). The data below shows the difference from the mean.

```
> km_final$centers
       Salary     weight      height experience        age         G
1 -1796384.28 -0.3256596 -2.92232958  0.1824454 -0.1335164  1.199887
2 -3013671.09  0.1906397 -0.06971589 -1.9161103 -0.9032700 -7.110760
3  2944608.94  0.1305320 -1.79908192  2.2301966  1.5640596  8.234934
4  7231951.43 -1.7428654  6.30508475  2.4661741  1.2319281  4.138780
5 13826170.47  0.4130320  8.31841808  3.1420716  1.3221846  7.864934
```

```
6    67988.91  0.5091240 -0.13169686  0.7296578  0.2513800  5.433210
          GS          MP          FG         FGA        X3P       X3PA
1 -1.6590457   -17.75074   -8.767042   -17.67844  -2.952233  -5.719221
2 -7.9715261 -311.13325  -57.616928  -123.64942 -11.010633 -30.181503
3  8.1495115  367.00992   64.626942   142.05535  21.107639  54.038518
4 16.5513545  370.29958   75.619731   150.50847   0.991453   4.316964
5 24.8426365  643.31804  172.798192   359.58847  18.648889  51.924143
6  0.6371193  148.42425   18.100721    43.48549   8.750958  22.930580
         X2P       X2PA          FT         FTA         ORB          PF
1  -5.814809 -11.95922   -7.560660   -8.029586   0.0530475   0.9898798
2 -46.606295 -93.46792  -28.379378  -35.958352 -13.8642266 -22.2423341
3  43.519303  88.01683   30.527395   35.792844  10.9232286  23.7870468
4  74.628278 146.19151   36.672027   51.408228  25.2361292  23.0502680
5 154.149303 307.66433  119.181770  148.632844  32.4751036  32.0564218
6   9.349763  20.55491    4.330736    4.172154   4.5606208  15.1722839
         DRB        TRB       AST         STL      BLK       TOV        PTS
1  -4.401007  -4.347959  -3.01709  -0.8676804 -1.08407  -1.97136  -28.04698
2 -45.046135 -58.910361 -33.95444 -10.1036953 -5.92591 -19.98455 -154.62387
3  42.724723  53.647952  30.13609   9.9164901  4.22345  17.47066  180.88892
4  72.749964  97.986093  60.06678  14.7362017 15.70983  29.32362  188.90294
5 121.165348 153.640452 108.44422  21.9577401 14.18470  62.22003  483.42704
6  14.767647  19.328268   5.02169   5.1287746  1.34378   6.55704   49.28314
```

Interpretation By analyzing the centers coordinates, we can derive the caracteristics of the different player clusters. We can assume there are three super groups in the data: **rookies** (cluster 2), **experienced** (clusters 1 and 6) and **senior** players (clusters 3,4 and 5). The experience is the attribute that influence the most the player salary. The more experienced a player is, the higher salary he earns - senior players are above 3 millions. At the opposite, the rookies are under 3 millions the mean salary. Note also that the age is strongly correlated with experience. The more experienced a player is, the higher the probability to be older and finally the higher salary he earns. Let's understand deeper the variances into super groups.

Rookies

Since rookies have played lesser games than other players, their statistics are lower - the centroid's coordinates are all below the mean. But be careful, that doesn't mean the players are bad. It just translates a lack of experience compared to experienced and seniors.

Experienced

Seniors

Question 2

toto

Question 3

toto

Question 4

toto