GE2324
The Art and Science of Data
MoneyBall Project

Project by:
Aarnav Gupta (aarngupta2) - 55990960 - K-Means Clustering [25%]
Aaron Shanly George (ashanlyge2) - 55773349 - Correlation [25%]
Aryan Girish Kasliwal (akasliwal2) - 55972222 - Introduction and Initial Calculations [25%]
MASHKIN Ivan (imashkin2) - 55844838 - Network Theory [25%]

1. MoneyBall: the Problem

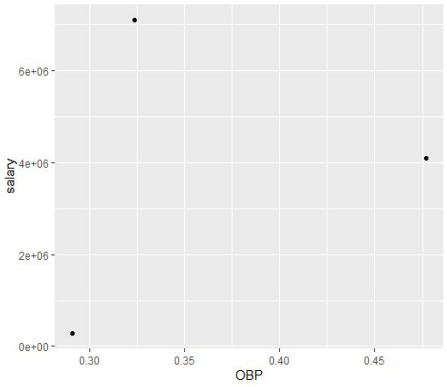
In 2001, Oakland Athletics had a remarkable season when they reached the playoffs of Major League Baseball (MLB). This season was soon followed by the departure of their three key players, Johnny Damon, Jason Giambi, and Jason Isringhousen. Being the underdogs in 2001, the Oakland A's budget was only USD 44 Million as against the big teams who had a budget of over USD 100 Million. It was expected that the Oakland A's would have a tough time replacing their star players with the limited budget. This formed the basis of the MoneyBall problem. The manager of Oakland A's, Billy Beane, took up the responsibility of finding players that not only excelled the three players that left but also had low salaries in order to fit the budget.

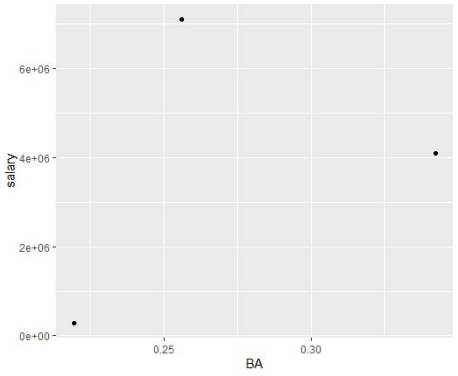
Until this point in Baseball history, the performance of every player was judged majorly based on their At Bats' throughout the year. Billy Beane's ideology was that many more statistics define a player's performance than just their At Bats'. He assembled a team that did rigorous statistical analysis on all the available players and included statistics such as On Base Percentage (OBP), Slugging Percentage (SLG), and Batting Average (BA).

Throughout this project, our group has analysed all players from 1985-2001, evaluated their statistics and found interesting relations between baseball characteristics and salaries of the players. Our results may/may not match the actual signings by Oakland A's that year since we do not account for the availability of players at that particular time.

2. Evaluation of the lost players

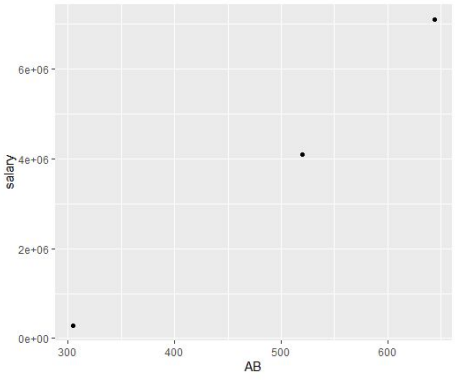
In the first part of our project, we evaluated the players that left the team. For each player, we calculated their On Base Percentage (OBP), Slugging Percentage (SLG) and Batting Average (BA). We then found out the average of these statistics in order to compare it to the possible replacement players. Further, we calculated the total of the salaries for these three players in order to meet the budget with the replacement players.





The graphs above compare the salary of each lost player to their On Base Percentage and Batting Average. It can be seen that there is a player who has a comparatively low OBP and BA and yet he receives a high salary. From this, we can conclude that the relation between OBP or BA and salary is not strong.

The graph on the right compares salary with At Bats. It can be inferred that the trend that exists between OBP, BA and salary does not exist between AB and salary. Rather, salary constantly increases with the increase in AB. This shows that baseball players were largely evaluated based on their AB and not as much based on other characteristics. In the later part of the project, a similar observation has been shown with correlation and with k-mean clustering between OBP and salary and AB and salary with a larger data set.



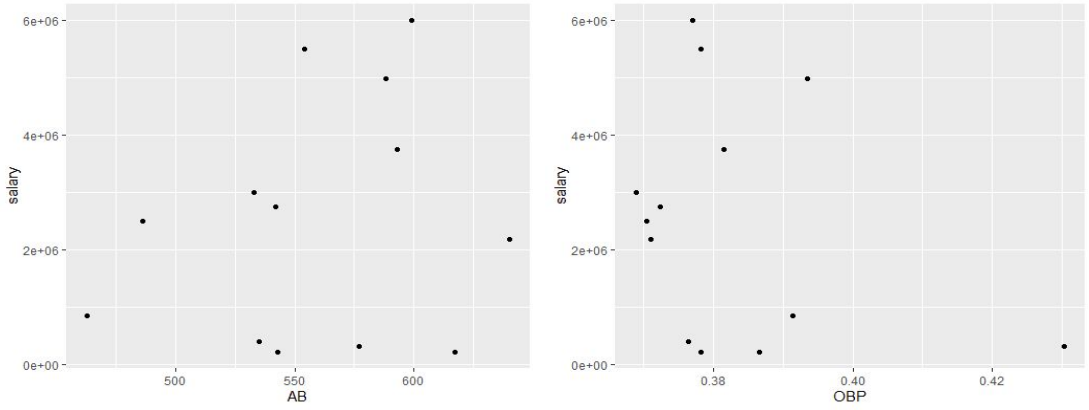
3. Narrowing down to top 40 replacements

Once we had the baseball statistics of the three lost players, we collected data of about 98,000 different performances by all players in their complete careers from 1871 to 2013. We then collected data on the salaries of these players from 1985 to 2013 and eliminated all players before 1985 as they would have already retired from Baseball by 2001. For each of the left over players, we then evaluated their On Base Percentage, Slugging Percentage and Batting Average. Further, we also evaluated every player's salary and added it to our data set of baseball statistics. Once all these statistics were calculated, we started analysing players that had a higher OBP along with a lower Slugging Percentage and salary than the average of the players that left. These constraints narrowed our list of replacements down to top 40 players. This data set of 40 players was then divided into different Age groups; age below 22, age between 23-27, age between 28-32, age between 33-37, and age above 38.

a. Age group below 22

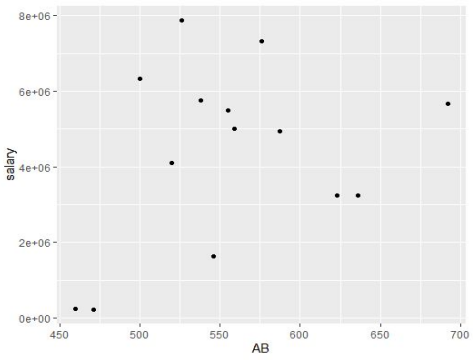
No real conclusion can be drawn from the graph for players below the age of 22 because there are not enough data points (only 1 player) in this age group.

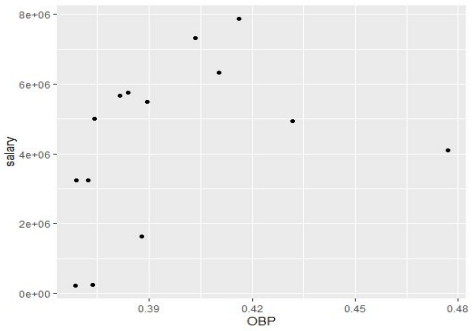
b. Age group 23 to 27



The graphs above show two sets of scatter plots, one is between salary and At Bats, the other is between salary and On Base Percentage. It can be inferred from the graphs that there are no players with high salary and low At Bats, but there are players with high salary and low OBP as well as low salary and high OBP. Hence, to replace the three lost players, it can be possible to get players from this age group that have a high OBP and less salary (less than 1 Million dollars) such that the new players fit the teams budget.

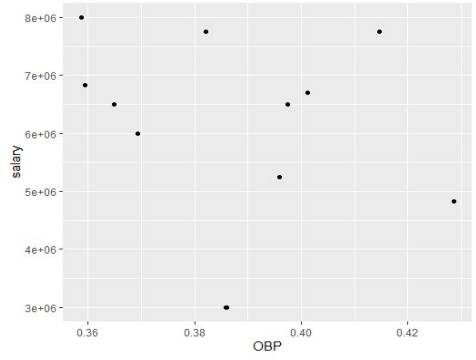
c. Age group 28 to 32

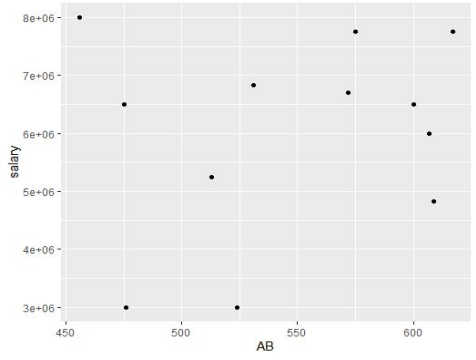




The graphs above show the same two sets of scatter plots as before, but for a different age group. It can be inferred from these graphs that there are no players with high AB and low salary as well as there are no players with high OBP and relatively low salary. Hence, we can conclude that in order to get a player with high OBP from this age group, they must have a moderate At Bats and salary in the range of 3 to 5 Million dollars.

c. Age between 33 to 37





The graphs above show the same two sets of scatter plots as before, but for a different age group. From the graphs, it can be deduced that there are no players with high At Bats and low salary as well as no players with high OBP and low salary. In comparison to the age groups earlier, it can be inferred that the salary gradually increases with age even if the performance does not change. Hence, any replacement player from this age group would have a high salary and a moderate performance rate in terms of AB and OBP.

d. Age group above 38

No real conclusion can be drawn from the graph for players below the age of 22 because there are not enough data points (1 player) in this age group.

4. Correlation

Note: All correlations stated below are Pearson's Correlations and are **not** restricted to the players performance in 2001 only. This has been done in order to be value sensitive for baseball statistics.

a. Age group below 22

Since there is only one player in this age group, that too with data entries of only one year, we can not perform correlation analysis.

b. Age group 23 to 27

i. Correlation between AB and salary: 0.411.
ii. Correlation between OBP and salary: 0.234

c. Age group 28 to 32

i. Correlation between AB and salary: 0.551
ii. Correlation between OBP and salary: 0.357

d. Age Group 33 to 37

i. Correlation between AB and salary: 0.326
ii. Correlation between OBP and salary: 0.498

e. Age group above 38

Though this age group too has only one player, data entries from multiple years can be used for finding correlation.

i. Correlation between AB and salary: 0.453
ii. Correlation between OBP and salary: 0.535

Observation: It can be inferred from the correlation data above that the correlation between At Bats and salary, although not strong, is higher for younger players, especially below the age 32. This strengthens the fact that baseball players were largely evaluated based on their At Bats and not as much based on their other characteristics.

However, another interesting observation that can be made is that as the age increases, the correlation between OBP and salary tends to increase. A similar observation was stated earlier in this report. This is because, irrespective of the change in a player's performance, many other factors such as experience, game knowledge and connections, longer training durations, personal influence etc have an impact on the high salary rates for players in the higher age groups. This indicates that any player who falls in the age group above 33 tends to have a higher salary than before, irrespective of his performance.

5. K-Means Clustering

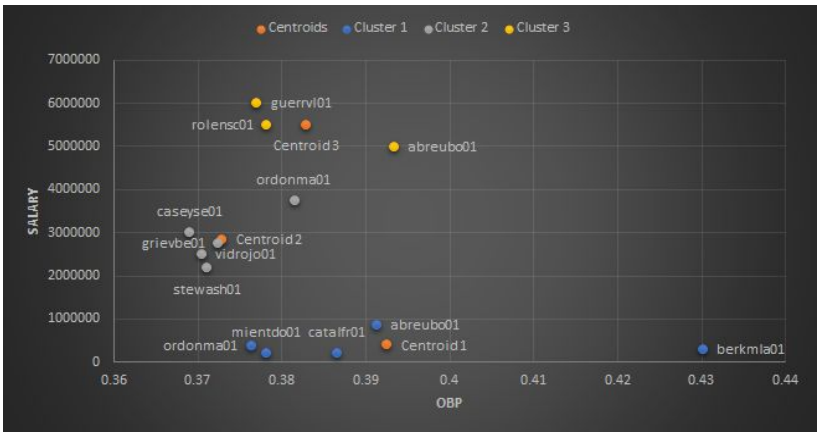
We took the data for the latest available year for each player in different age groups, and performed K-Means clustering to obtain 3 centroids & clusters. For each age group, clustering was performed twice, once with Salary and OBP; and again with Salary and AB. This has been done to verify the correlation data obtained earlier, and to see if the players can even be grouped together with these data (if the clusters follow a similar pattern).

a. Age Group below 22

Since there is only one player in the given age group, we can not perform K-Means clustering here.

b. Age Group 23 to 27

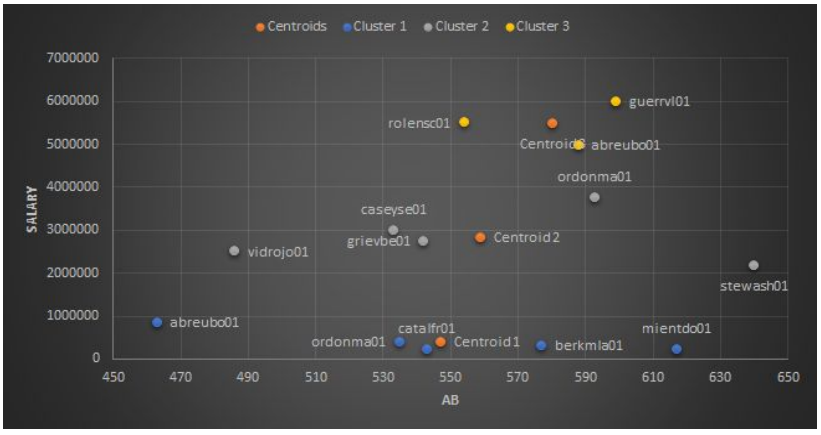
i. OBP and Salary



This shows us that Salary and OBP are not that well linked to each other. If they were, our cluster centroids would have shown some trend. Such as, a cluster would have lower salary and/or high OBP; or vice versa. But the centroids we got are not following such a trend. The middle salary corresponds to the lowest OBP cluster (Centroid 2), while the middle OBP corresponds to the highest salary cluster (Centroid 3).

This shows that there is very little correlation between OBP and Salary, which we already confirmed by calculating the Pearson correlation between the two, which had come out to be 0.234.

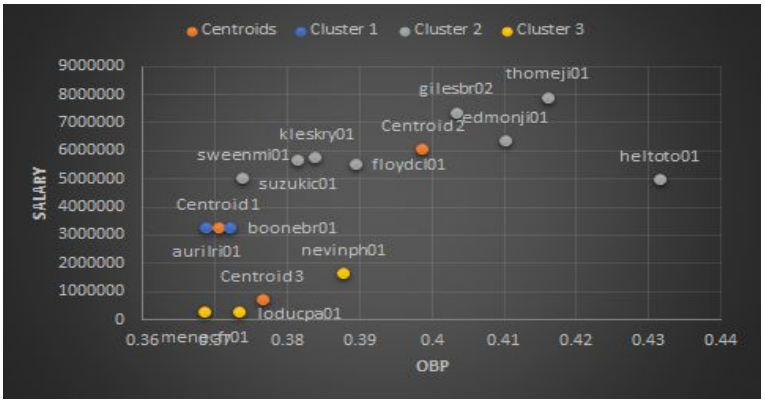
ii. AB and Salary



This time, we find very similar results, with Y coordinate of the centroids not changing at all, and the same players belonging to the same clusters. Hence, it indicates that AB or OBP does not matter much in cluster formation, as much as Salary does.

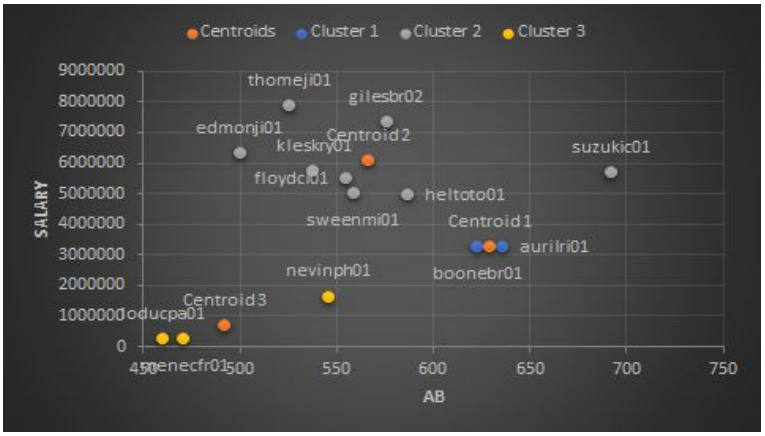
c. Age Group 28 to 32

i. OBP and Salary



As expected by the correlation of the same data (= 0.357), there is no specific trend observed in the cluster centroids. In other words, a cluster is not able to be representative of any specific group. Centroid 2 seems to represent OBPs of all kinds of values, both low and high, but only high salary; hence there's very little correlation.

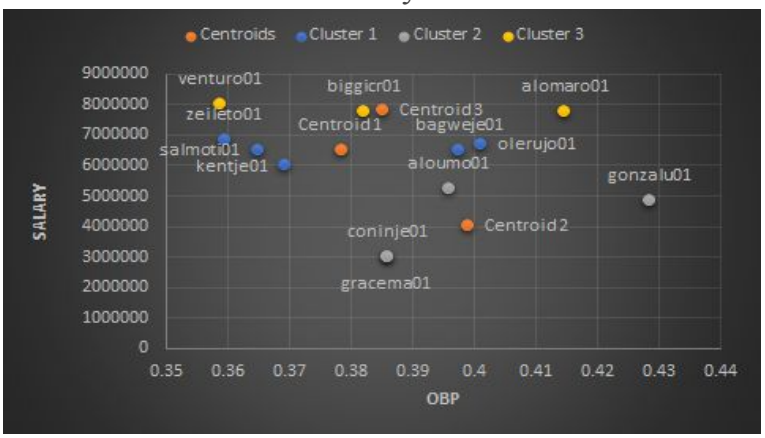
ii. AB and Salary



The same trend as earlier is again observed.

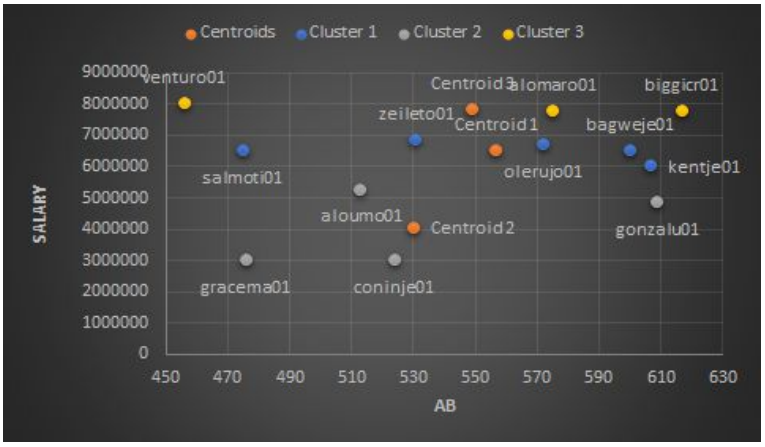
d. Age Group 33 to 37

i. OBP and Salary



This time too, the centroids are more representative of the salary, and barely representative of OBP. For example, players in centroid 3 (cluster 3) vary from lowest OBP to second highest OBP.

ii. AB and Salary



The same trend as the previous groups is observed in this too.

e. Age Group above 38

Since there is only one player in this age group, K-Means Clustering is not possible.

6. Network Analysis

For part four, we have taken the 'latest statistics' (year 2001) of the players from all age ranges and calculated their OPS, which as mentioned previously is "On base percentage plus Slugging percentage". OPS is a great method to estimate the offensive skills of an individual player, a value of 0.9 or above is valued as the offensive skills of a great professional MLB player.

$$OPS = OBP + SLG$$

$$OBP = \frac{H + BB + HBP}{AB + BB + SF + HBP}$$

Where:

- "$H + BB + HBP$" basically means the number of times the batter gets to at least the first base after hitting the ball.
- "$AB + BB + SF + HBP$" basically means total plate opportunities to take the bases from the hit.

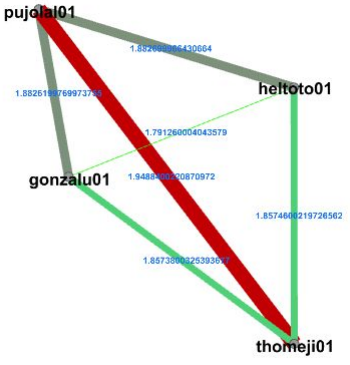
We have later calculated each individual OPS through the data we have collected. Our plan was to find synergies between the players by seeing how aggressive they play, the hypothesis is that players that play more aggressively will play well along other offensive players as they would have common experience of playing the game, common skill and easier to communicate their strategies with somebody with the same mindset. We have taken the 4 best players with the OBP and evaluated their OPS.

To form a visual representation, we have taken advantage of the matrix table and gephi.To analyse the synergy, we have taken the mean between 2 individuals, to find their mean score - the greater the value, the better synergy.

However, for the matrix, the OPS was inverted to make the greatest synergy to be represented by the shorter distance between the players; The distance between the players was exaggerated slightly to show which players had best and worst synergies.

Furthermore, color scheme and width were used to further visualise the synergy between players: small width and green path exemplified best synergy, while largest width and red color represented worst synergy.

	pujolal01	heltoto01	thomeji01	gonzalu01
pujolal01		0.94135	0.97442	0.94131
heltoto01	0.94135		0.92873	0.89563
thomeji01	0.97442	0.92873		0.92869
gonzalu01	0.94131	0.89563	0.92869	



From this observation, we can conclude that overall, the best player which has the best synergy with the rest of the player is "Gonzalu01", who is Luis Gonzalez. The second best player is "Heltoto01", who is Todd Helton.

The same concept could also be applied on a larger scale, but of course would require greater CPU power, which of course was a great limitation to this part of the project.

7. Conclusion

Before 2002, all baseball players were judged based on very few characteristics, mainly At Bats. When we performed correlation between At Bats and salary and OBP and salary, the results were in the range of low correlation to medium. Further, while we narrowed our list of players down to 40, we noticed a general trend of increase in salary with age irrespective of performance. Similarly, when we did K means clustering, we found that salary increases with age irrespective of performance. Firstly, from the low correlation, we can conclude that characteristics such as OBP and AB are nor enough to judge a player. As shown with network analysis, many other unseen factors such as synergy are also important. The trend of increasing salary with age tells us that many off field characteristics such as experience, personal influence and training period also affects the salary.

Hence we can conclude that new factors such as OPS and synergy can be used to judge the players and decide their salary in a better way. As for the three replacement players that we had to find, the graphs of narrowing our players down to 40 tell us the categories of players we can pick from each age group. Any team will have to balance their player selection in their own way and that is why we have left that part open ended.

8. References

a. Gadoci, B. (2016–1871, June 20). Lahman's Baseball Database (Version 2016) [Baseball statistics from 1871 to 2015.]. SeanLahman.com. https://data.world/bgadoci/lahmans-baseball-database

b. Teknomo, K. (n.d.). Online K Means Clustering. Revoledu. Retrieved July 24, 2020, from https://people.revoledu.com/kardi/tutorial/kMean/Online-K-Means-Clustering.html