**K-Means Clustering for Sports Teams Budgets**

**Project Overview**

In sports there has always existed the saying that money buys championships. The idea here is that teams with higher budgets get more wins, and thus eventually championships. I wanted to test that saying by seeing if teams with similar budgets perform similarly in terms of wins. This analysis will be done in Microsoft Excel to demonstrate the technical understanding of what is occurring in a K-Means clustering algorithm.

**Table of Contents**

1. [Introduction](#introduction)
2. Plotting
3. Initialize and Iterate
4. Conclusion
5. Installation
6. Project Structure
7. Contact

**Introduction**

To get started with this analysis we utilize the dataset below. This dataset shows player payroll and season wins for all teams in the Liga MX Soccer League for the 22-23 season.

|  |  |  |
| --- | --- | --- |
| Team ID | Player Payroll | Season Wins |
| 1 | $20,430,000 | 16 |
| 2 | $18,220,000 | 22 |
| 3 | $17,330,000 | 15 |
| 4 | $12,610,000 | 12 |
| 5 | $10,940,000 | 8 |
| 6 | $9,970,000 | 13 |
| 7 | $9,690,000 | 14 |
| 8 | $9,380,000 | 10 |
| 9 | $9,135,000 | 15 |
| 10 | $8,990,000 | 11 |
| 11 | $8,810,000 | 13 |
| 12 | $8,780,000 | 7 |
| 13 | $8,350,000 | 7 |
| 14 | $8,250,000 | 10 |
| 15 | $7,430,000 | 8 |
| 16 | $7,330,000 | 11 |
| 17 | $7,280,000 | 4 |
| 18 | $6,250,000 | 8 |

I will perform K-means clustering on the given dataset to identify if any teams fall into clusters based on similarities between the two variables in the dataset. I will use the variables "Player Payroll" and "Season Wins" as the features for clustering. First, we start by plotting our current data in a scatter plot.

**Plotting**

Initially we can see a general trend that indicates that more money will equal more wins. However, we are trying to see more precisely how closely these teams are related to each other. To do this we will assume that we want to divide the data by three clusters (K=3).

**Initialize and Iterate**

Randomly select three different data points as the initial centroids. We choose IDs 3, 9, and 17 as our centroids.

|  |  |  |
| --- | --- | --- |
| Center 1 | $17,330,000 | 15 |
| Center 2 | $8,990,000 | 11 |
| Center 3 | $6,250,000 | 8 |

Now we calculate the distance between each data point and the centroids. We assign each data point to the cluster with the nearest centroid. After calculating the distance for between each data point the data set will look like this

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Team ID | Distance 1 | Distance 2 | Distance 3 | Cluster |
| 1 | $3,100,001 | $11,440,005 | $14,180,008 | 1 |
| 2 | $890,007 | $9,230,011 | $11,970,014 | 1 |
| 3 | $0 | $8,340,004 | $11,080,007 | 1 |
| 4 | $4,720,003 | $3,620,001 | $6,360,004 | 2 |
| 5 | $6,390,007 | $1,950,003 | $4,690,000 | 2 |
| 6 | $7,360,002 | $980,002 | $3,720,005 | 2 |
| 7 | $7,640,001 | $700,003 | $3,440,006 | 2 |
| 8 | $7,950,005 | $390,001 | $3,130,002 | 2 |
| 9 | $8,195,000 | $145,004 | $2,885,007 | 2 |
| 10 | $8,340,004 | $0 | $2,740,003 | 2 |
| 11 | $8,520,002 | $180,002 | $2,560,005 | 2 |
| 12 | $8,550,008 | $210,004 | $2,530,001 | 2 |
| 13 | $8,980,008 | $640,004 | $2,100,001 | 2 |
| 14 | $9,080,005 | $740,001 | $2,000,002 | 2 |
| 15 | $9,900,007 | $1,560,003 | $1,180,000 | 3 |
| 16 | $10,000,004 | $1,660,000 | $1,080,003 | 3 |
| 17 | $10,050,011 | $1,710,007 | $1,030,004 | 3 |
| 18 | $11,080,007 | $2,740,003 | $0 | 3 |

We assign the cluster based on the closest distance to the centroid.

Here we begin the iteration process where we recalculate the centroids by taking the mean of all the data points in each cluster and recalculating the distance from the centroid. This process repeats until the clustering assignment no longer varies between iterations. In this case we ran 4 iterations until the clusters no longer changed.

**Conclusion**

The original plotted data showed that there is a positive correlation between team budget and season wins. However, our clustered plotted data shows that budgets between $6m - $8.5m are considered low and have the lowest performance. Incrementing the budget from the highest point in cluster three by $4m will land you inside the second cluster. In the second cluster a team can expect to earn about 3 more wins on average over teams in the third cluster. The biggest jump is seen in the first cluster. The first cluster has an average budget of $18.6m and has on average 7 more wins than the second cluster.

In conclusion we can see that only a substantial amount of money will provide a substantial increase in performance. Marginal increases in player payroll will generally have mild to no impact on performance.

**Installation**

To get started with this project, you'll need to have the following software installed:

* Microsoft Excel

**Project Structure**

* data/: Directory containing the dataset and exercise details.

**Contact**

For any questions or feedback, please contact ivan231r@mail.com.

**Sources**

Capology (2023, January 31). 2022-2023 Liga MX Payrolls. Retrieved July 12, 2023, from https://www.capology.com/mx/liga-mx/payrolls/

FlashScore (2023, May 31). Liga MX Standings. Retrieved July 12, 2023, from https://www.flashscore.com/football/mexico/liga-mx/standings/