Fifa Ratings

Abstract + Background:

Fifa player ratings are used to rank different players in a game called Fifa. Every year, Fifa creates a new game that includes updated rankings. These rankings are mainly developed by attributes such as dribbling, shooting, passing, speed, etc. EA Sports uses various ways such as data, scouting results, and analysis to rank a player on a scale from 0-99. Although player ratings are mainly used within the game, it is also used for recruiting and "player of the week". To delve deeper into the flaws of Fifa player ratings, we must first determine how the ratings and stats are decided. Firstly, the Fifa data analysis board consists of over 6,000 Fifa Data Reviewers from around the world that tracks each player. They collect the rating of each attribute listed above. Depending on their position, their CB coefficient changes for each attribute, and the total rating is shown. However, some of the ratings may be different from the calculated value because it may shift a little from their reputation. Michael Muller-Mohring, the head of data collection, said that he is yet to meet a player that is satisfied with their rating, however, some ratings seem completely off from their attributes. For example, Thomas Muller is a big exception to the Fifa Ratings. His overall rating is 87, but none of his attribute ratings are close to that overall rating and all of them are below. Muller-Mohring said that this player was one of the hardest to rate because he is not good in a specific attribute, but he excels in his ability to assess the play and has exceptional positional awareness. Fifa ratings have a lot to improve, and this is just one aspect where Fifa ratings may not be accurate.

Specific methods can be utilized to determine why this might be and know which players are outliers in the Fifa Ratings The complete set of player ratings for Fifa 22 is used with several machine learning techniques to determine the outliers in the data by comparing the overall rating with the other attributes that affect it. The main method that should be used by Fifa for player ratings is the Elo method which is used widely in games like Chess and any video game with a ranking system. In order to find the outliers in the data we used DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and LOF (Local Outlier Factor) to compare these machine learning techniques and analyze the best technique for finding outliers. Also, methods that aren't specifically used for finding outliers are used to check the accuracy of the methods. K-means is basically a vector quantization that partitions different attributes and observations into specific clusters.

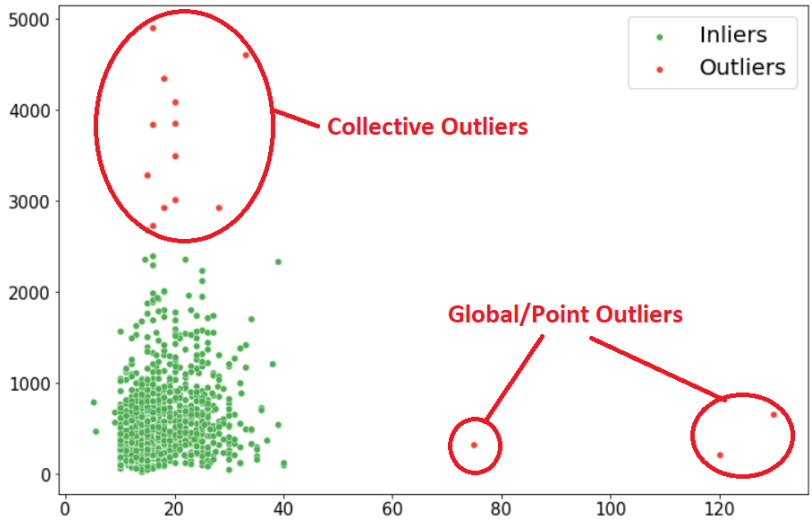
How do inflation and bias affect the accuracy of Fifa Ratings and what methods can be used to spot these outliers?

Main Methods:

PCA (Principle component analysis) is a method that is used in Python to convert a multidimensional array and convert it lower to a compatible array. In this case, the data needs to

be converted to a 2d array in order to use the methods that analyze outliers in the Fifa player data. In the data, only specific attributes in the Fifa 22 dataset are used because some of the data are unimportant information.

DBSCAN and LOF are the main methods that are used to analyze player data. They are both clustering methods that use machine-learning techniques to make a simulation that checks if there are nearby data points within a certain range. However, each method differs in many ways. DBSCAN is a clustering algorithm that groups data points based on the radius around them and separates regions that are of high density. For example, if there is a data point within a specific radius in the graph, those 2 are grouped. The user needs to use trial and error to determine the right radius and amount of data points in one group possible so that it confirms the outliers. LOF is slightly different from DBSCAN in that it checks to see which groups have the lowest density and compares all of the groupings. The data points with significantly lower density compared to neighbors are considered outliers. The downside of this method is that it may require a lot of experimentation to discern the perfect parameters for the method. Below is an example of what the graph would look like if there were two attributes to the player.



https://www.analyticsvidhya.com/blog/2022/10/outliers-detection-using-iqr-z-score-lof-and-dbscan/

DBSCAN utilizes two parameters in order for the method to work. The first parameter is Epsilon which is the maximum distance that each point is to the other in order for them to be part of the same group. Any points that aren't close enough to any other points counts as an outlier. The next parameter is the minimum amount of points necessary for a specific group to be called a "group". For example, if there are 2 neighbors and the minimum amount of points necessary is 3, then those 2 points are counted as outliers. There are also three different types of data points to keep in mind when clustering data. The first data point is the one that is included inside a group. The next 2 data points are counted as outliers but aren't actually part of the group. One of them

has a neighbor that is included in the group but isn't actually part of the group because the minimum amount of points nearby isn't big enough. Finally, the last one isn't near any other point and is a complete outlier from the rest.

LOF is not actually a clustering method but a method that is specifically for finding outliers. It also takes in similar parameters like DBSCAN but instead of checking for neighbors, it also takes in a parameter that checks for the density of a group. For example, the parameter will determine how many dots can be in a group before it is considered an outlier.

Another method that utilizes clustering differently but can be used to analyze the player data is K-Means. The main process of this method is to take data points as inputs and groups and put them into "k-clusters" which are essentially clusters that group different attributes into sections. To follow the same idea as before, these clusters can be used to find the outliers in the player data but also find potential in a specific group. For example, if goalkeepers are grouped and the age difference is huge, there may be a slight chance of bigger potential compared to other players.

Elo Method assigns ratings to players based on performance in matches and changes based on matches played. The Elo rating system is a popular method for ranking players and teams in various sports, including soccer. It assigns ratings to players based on their performance in matches and adjusts these ratings based on the results of subsequent matches. This method could be used to assess the accuracy of FIFA ratings by comparing them to Elo ratings and examining any discrepancies. Most likely Fifa used the Elo method to create the rankings

Other Machine Learning Techniques:

Neural Networks are an artificial intelligence way of teaching computers how to analyze data that is similar to the human brain. It is basically a type of deep learning that creates interconnected nodes in the structure of the human brain. It continues to improve itself and is made as an adaptive system. For example, ChatGPT also uses this method and adapts to the situation as it continues to interact with the user.

R's Graphical Techniques:

Regression analysis is a method that was in Baseball methods using R. It is mainly used to graph data and figure out the process of independent and dependent variables. The techniques can predict the potential of players which may be used in this research paper and depending on what parameters you put, it can also determine outliers as well. There are 2 main forms of regression analysis which are linear and logistic analysis. The linear analysis basically predicts the continuous dependent variable while the logistic analysis has many different dependent variables that can change over time. The best way of using regression analysis is probably to create simulation models and graph it.

Data:

The most important data is the Fifa 22 complete data set which includes the entire dataset of the players in the game. It included all the attributes of the player and all the information necessary for completing the analysis. The best way to import the data so that the player ratings can be grouped and compared is to use the Python library, pandas. In the library, some methods convert data frames of CSV files to NumPy array. With the data sorted in a multidimensional array that contains all the attributes of the player including the overall score, the main methods can be utilized to check for outliers in the player ratings. Below are the different attributes of each player when determining the overall score.

pace	shooting	passing	dribbling	defending	physic	attacking_crossing	attacking_finishing	attacking_heading_accuracy	attacking_short_passing	attacking_volleys
85	92	91	95	34	65	85	95	70	91	88
78	92	79	86	44	82	71	95	90	85	89
87	94	80	88	34	75	87	95	90	80	86
91	83	86	94	37	63	85	83	63	86	86
76	86	93	88	64	78	94	82	55	94	82
						13	11	15	43	13
97	88	80	92	36	77	78	93	72	85	83
						15	13	25	60	11
						18	14	11	61	14

skill_dribbling	skill_curve	skill_fk_accuracy	skill_long_passing	skill_ball_control	movement_acceleration	movement_sprint_speed	movement_agility	movement_reactions	movement_balance	power_shot_power	power_jumping	power_stamina
96	93	94	91	96	91	80	91	94	95	86	68	72
85	79	85	70	88	77	79	77	93	82	90	85	76
88	81	84	77	88	85	88	86	94	74	94	95	77
95	88	87	81	95	93	89	96	89	84	80	64	81
88	85	83	93	91	76	76	79	91	78	91	63	89
12	13	14	40	30	43	60	67	88	49	59	78	41
93	80	69	71	91	97	97	92	93	83	86	78	88
30	14	11	68	46	54	60	51	87	35	68	77	43
21	18	12	63	30	38	50	39	86	43	66	79	35

power_strength	power_long_shots	mentality_aggression	mentality_interceptions	mentality_positioning	mentality_vision	mentality_penalties	mentality_composure	defending_marking_awareness
69	94	44	40	93	95	75	96	20
86	87	81	49	95	81	90	88	35
77	93	63	29	95	76	88	95	24
53	81	63	37	86	90	93	93	35
74	91	76	66	88	94	83	89	68
78	12	34	19	11	65	11	68	27
77	82	62	38	92	82	79	88	26
80	16	29	30	12	70	47	70	17
78	10	43	22	11	70	25	70	25
85	86	80	44	94	87	91	91	50
72	65	93	91	72	78	54	84	90
82	79	63	39	90	87	84	90	43

?

Because of the empty blocks of data, oftentimes, there is a difficulty when trying to manipulate the data into the form of a 2-dimensional array. This is why the players are split into non-goalkeepers and goalkeepers. Below are the attributes that are used when comparing the

players in each section.

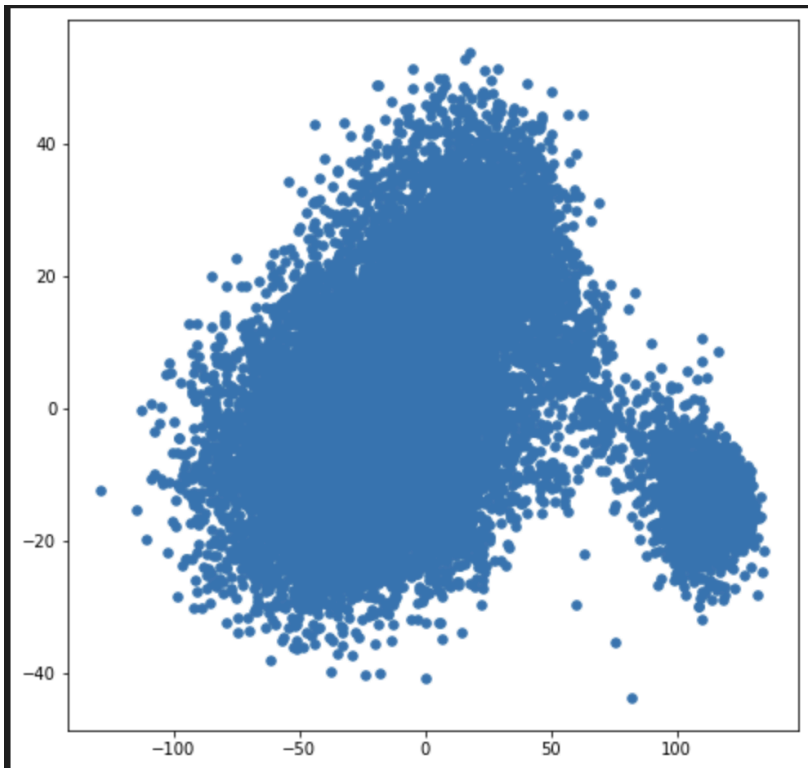
Goal Keepers	Non-Goal Keepers
`fifaData_GK = df[['overall', 'goalkeeping_diving', 'goalkeeping_handling', 'goalkeeping_kicking', 'goalkeeping_positioning', 'goalkeeping_reflexes']].values`	`fifaData_P = df_P[['overall', 'pace', 'shooting', 'passing', 'dribbling', 'defending', 'physic', 'attacking_crossing', 'attacking_finishing', 'attacking_heading_accuracy', 'attacking_short_passing', 'attacking_volleys', 'skill_dribbling', 'skill_curve', 'skill_fk_accuracy', 'skill_long_passing', 'skill_ball_control', 'movement_acceleration', 'movement_sprint_speed', 'movement_agility', 'movement_reactions', 'movement_balance', 'power_shot_power', 'power_jumping', 'power_stamina', 'power_strength', 'power_long_shots', 'mentality_aggression', 'mentality_interceptions', 'mentality_positioning', 'mentality_vision', 'mentality_penalties', 'mentality_composure', 'defending_marking_awareness', 'defending_standing_tackle', 'defending_sliding_tackle']].values`

Results:

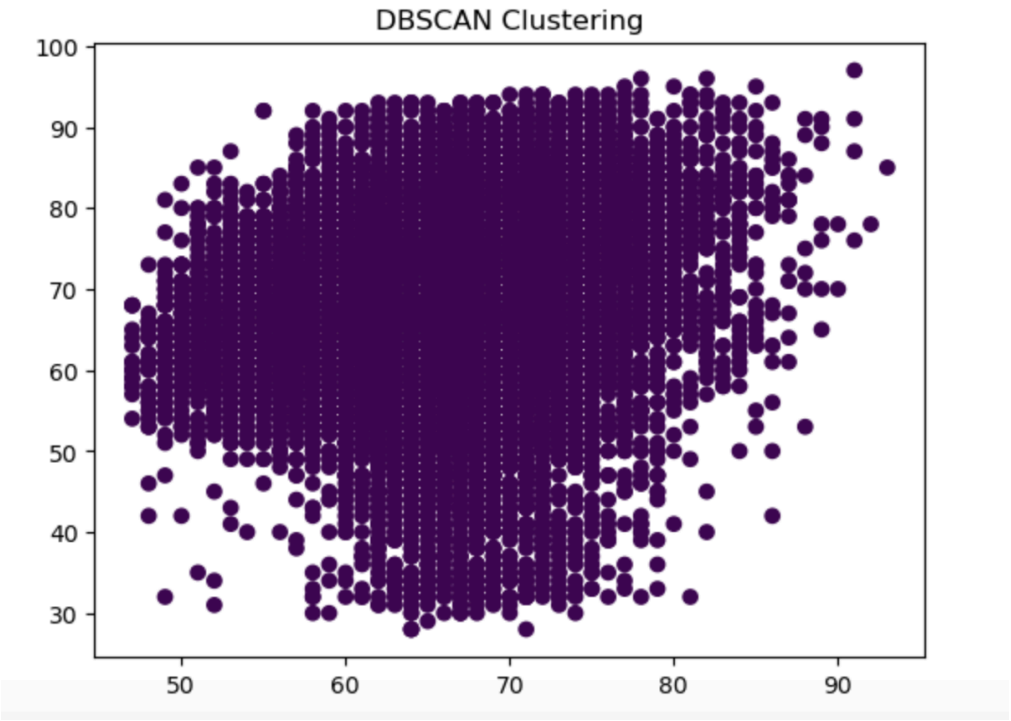
The data above is first converted into a 2-dimensional array. Each array is a player and

they all consist of the overall rating and the several attributes that make up the array. Once it is converted, the data can then be graphed using a form of K-means which plots each dot and the outliers can be easily observed. However, this is not enough to tell which players are specifically outliers and the problems that occur because of these outliers.

While trying to plot the data into a graph, there were some problems that occurred. When there was a null in the data, there was an issue with PCA which made it so that the data wouldn't be able to change into 2-dimensional arrays due to there being missing data. Therefore, the data was split into 2 different sections to make it much easier to track and not make any errors. It was split into the goalkeeper section and the nongoalkeepers section. Each one had different data and players that had missing data were either removed or modified so it fits in the data and doesn't affect the outlier detector.



Next up is the DBSCAN clustering method. This method, as explained before, is a widely used method for determining outliers. In this clustering graph, the minimum number of samples to be in a group is 3, and the maximum distance for there to be a neighbor is 2.

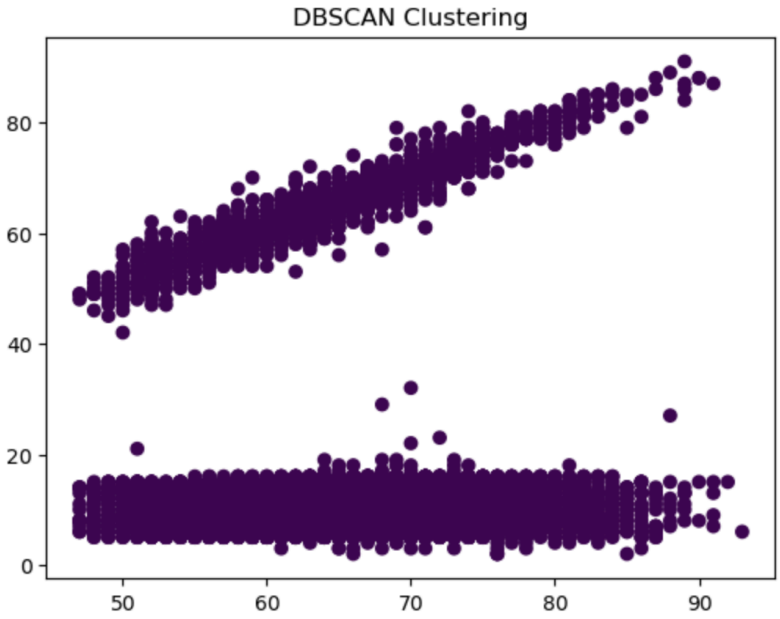


The most difficult part of the DBSCAN is definitely choosing the parameters. Depending on what parameters you choose, the outliers are completely different and it will completely change which players are going to be targeted. Just from the number of outliers alone, using these parameters, there are around 17000 outliers. And just some of the several outlier points are shown below.

```
Number of outliers: 17107
Outlier points:
[93. 85. 92. 91. 95. 34. 65. 85. 95. 70. 91. 88. 96. 93. 94. 91. 96. 91.
 80. 91. 94. 95. 86. 68. 72. 69. 94. 44. 40. 93. 95. 75. 96. 20. 35. 24.]
[92. 78. 92. 79. 86. 44. 82. 71. 95. 90. 85. 89. 85. 79. 85. 70. 88. 77.
 79. 77. 93. 82. 90. 85. 76. 86. 87. 81. 49. 95. 81. 90. 88. 35. 42. 19.]
[91. 87. 94. 80. 88. 34. 75. 87. 95. 90. 80. 86. 88. 81. 84. 77. 88. 85.
 88. 86. 94. 74. 94. 95. 77. 77. 93. 63. 29. 95. 76. 88. 95. 24. 32. 24.]
[91. 91. 83. 86. 94. 37. 63. 85. 83. 63. 86. 86. 95. 88. 87. 81. 95. 93.
 89. 96. 89. 84. 80. 64. 81. 53. 81. 63. 37. 86. 90. 93. 93. 35. 32. 29.]
[91. 76. 86. 93. 88. 64. 78. 94. 82. 55. 94. 82. 88. 85. 83. 93. 91. 76.
 76. 79. 91. 78. 91. 63. 89. 74. 91. 76. 66. 88. 94. 83. 89. 68. 65. 53.]
[91. 97. 88. 80. 92. 36. 77. 78. 93. 72. 85. 83. 93. 80. 69. 71. 91. 97.
 97. 92. 93. 83. 86. 78. 88. 77. 82. 62. 38. 92. 82. 79. 88. 26. 34. 32.]
[90. 70. 91. 83. 83. 47. 83. 80. 94. 86. 85. 88. 83. 83. 65. 86. 85. 65.
 74. 71. 92. 70. 91. 79. 83. 85. 86. 80. 44. 94. 87. 91. 91. 50. 36. 38.]
[90. 78. 66. 75. 82. 87. 83. 68. 65. 54. 82. 56. 79. 49. 49. 79. 81. 82.
 75. 82. 93. 92. 71. 77. 97. 72. 65. 93. 91. 72. 78. 54. 84. 90. 93. 86.]
```

When changing the parameters of the maximum distance to 5, the number of outliers decreases drastically and decreases to around 1000 outliers. Each outlier point has a specific reason for the overall rating.

The graph for just the goalkeepers is shown below and there are only a few outliers due to there being fewer player attributes that make up a goalie as shown in the data section.



Related Work

Murphy, R. (2021, October 2). *FIFA player ratings explained: How are the card number & stats decided?*. Goal.com US. https://www.goal.com/en-us/news/fifa-player-ratings-explained-how-are-the-card-number--stats-decided/1hszd2fgr7wgf1n2b2yjdpgynu

The main topic of this article is to describe how the OVR or the Fifa player ratings are created. It explains the number of scouts needed for Fifa to create these ratings and the CB coefficient that is used for each specific stat whether it be dribbling, passing, or strength. It mainly uses ratings with real-life player abilities and using player statistics. They collect the statistics and also use the World Cup, and different leagues from different levels. They use different attributes and establish rankings with benchmarks and comparisons. It also addresses the different debates that people are having when they see the FIFA player ratings. This article is important to my research paper because it includes the basis for how FIFA ratings are created and how they are used.

Brophy, J. (2022, September 30). *Meet EA Guru who decides FIFA ratings - and why Thomas Muller remains one exception*. talkSPORT. https://talksport.com/football/1206820/fifa-23-player-ratings-decide-system-pace-michael-muller-mohring-ea-thomas-muller/#:~:text=%22Each%20footballer%20is%20made%20up,%2C%20free%20kicks%2C%20etc.%22

This article is mainly about the low scores and the outliers of a player's performance. There is an interview with a FIFA guru who decides the FIFA ratings. The main statement of the article is talking about how several teams have faced criticism from players that were unhappy with ratings. They mainly talk about the most difficult attribute which is jumping. Since a player can get lucky sometimes and jump with an incredible goal, it may be difficult because that goal may have been a fluke. For example, Muller's FIFA card has no core stat over 87 but that is his overall rating. This article is crucial because of how it talks about the main part of my research paper which is the outliers and how they come up with the ratings even though the algorithm doesn't quite add up.

Marchi, M., Albert, J., & Baumer, B. (2019). *Analyzing baseball data with R*. CRC Press, Taylor & Francis Group.

The main idea of this book is the application of R, a programming language, and analyzing baseball data. It provides a guide on how R is used in different types of baseball analysis when it comes to player analysis and statistics. The book mainly covers data manipulation, visualization, and statistics which are used to determine how the player plays and what they do. It talks about many topics regarding the history of baseball and how it was used to track the methods every year. Although this book is mainly about baseball, I can use these algorithms in my paper and use statistical methods to explore FIFA ratings. The most important part, however, is the baseline and how I can use it as a source to learn about the topic of statistical analysis.

Wolf, Stephan, Schmitt, Maximilian, and Schuller, Björn. 'A Football Player Rating System'. 1 Jan. 2020 : 243 - 257

This article mainly talks about the Football Player Rating System. They created a method that would analyze the different football teams and players within them. They changed their results depending on the league that the article uses. They used several leagues like the UEFA Champions and Europa Leagues to compare the players. However, their main objective was to create a rating system for the players just like the Fifa player ratings. Elo algorithms are a major part of the rating system and are the easiest ways to graph onto a valuation model. They also use these models to talk about the skills that may not be listed in the attributes, talking about how each skill is evaluated clearly and is distinguished from the other. They also say that there is never an attribute that is chosen that doesn't reflect a player and each one is important in their own way. This is important to my research paper because it talks about the football player rating system and the good and bad parts when trying to find the correct algorithm.

Predicting the Potential of Professional Soccer Players Ruben Vroonen1 , Tom Decroos1 , Jan Van Haaren2 , and Jesse Davis1 1 KU Leuven, Department of Computer Science, Celestijnenlaan 200A, 3001 Leuven, Belgium 2 SciSports, Hengelosestraat 500, 7251 AN Enschede, The Netherlands

This paper talks about machine learning techniques and these techniques are used to predict the potential of professional soccer players. This study was conducted in Belgium and Netherlands

and they worked towards making a new and innovative model to predict potential. The first part of the paper talked about the dataset that they used to contain information about player statistics and groups. They mainly included attributes that were also on the FIFA website. They also used scouting reports to get assessments of player potential. In the next section of their paper, they trained a model that would evaluate player performance. This study changed the way people viewed players' potential using the prediction model. The research also talks about the potential machine learning methods that FIFA used for their player rankings since it isn't fully disclosed yet. The main conclusion that they made was the evaluation of the APROPOS projection system which made predictions with the k-nearest neighbors approach. Overall, they used multiple methods to predict player potential and created the most accurate possible methods. These scouting methods are very important to my research paper because it talks about the player's potential. Since I will be talking about the outliers and the potential using machine learning methods from the baseball book, I will need to use this research article to have a good understanding of what I will be facing and the problems that will occur when I research the potential of players.

Using FIFA Soccer video game data for soccer analytics Leonardo Cotta Computer Science Universidade Federal De Minas Gerais, Pedro O.S. Vaz de Melos, Fabrício Benevenuto, Antonio A.F. Loureiro

This paper talks about the use of data from the FIFA Soccer game for analytical purposes. This study was mainly shown from the popularity of the soccer players and how many people wanted to leverage the data. It makes insights of real-world soccer strategies and player performance. They collected data from FIFA Soccer matches, including attributes of players and teams. This data mostly talks about the player ratings, positions, and attributes. The study revealed several intriguing patterns and relationships according to the article. The researchers discovered that certain player characteristics, including quickness and shooting accuracy and had a huge impact on a team's chance of winning a match. The article emphasizes the possibilities of using games for statistics and although it doesn't really talk about real-world soccer dynamics, it is a quick and easy resource for doing research and making predictions. However, this study has consequences for soccer clubs, coaches, and analysts. It makes use of better knowledge of player performance, team strategies, and other factors of video game data discoveries. It also includes the importance of data-driven approaches. Researchers talked about hidden patterns and links in large-scale datasets by exploring analytical approaches. This leads to a better understanding of the study and the field. The study expands on and improves the field of sports analytics by providing new insights into player performance, team dynamics, and strategic decision-making in soccer. Overall, this article is important to my research paper because it talks about the use of FIFA soccer game ratings for analytical purposes and it directly relates to my topic.

Clustering Football Players by Using FIFA 19 Data Oğuz Can Yurteri Oğuz Can Yurteri Data Analytics Team Lead at Pensa Systems.
https://www.linkedin.com/pulse/clustering-football-players-using-fifa-19-data-o%C4%9Fuz-can-yurteri/

The main idea of this article was to create a machine-learning project that talks about football skills and includes players based on their attributes This data uses info from FIFA 19 with almost 18,000 players and millions of data points. The original dataset is grouped into several different attributes similar to the other research papers above. The article uses data analysis to compare the relationship between different players. It talks about the things that they observed and how they are used in the modeling stage while also giving a graph of the number of players in each section. Each graph that is listed in the article has a pair plot of positional skills and demonstrates the ability of a player as he plays in these different positions. The results are shown in each of the playing positions of the player but it doesn't really matter that much. Then there are modeling outliers and models that show the quartiles of each specific topic. Finally, it shows the results and graphs with attributes and players. This article is very useful because it uses machine learning techniques to cluster players which directly aligns with the techniques I am trying to use for my research paper.