DATA COLLECTION REPORT

FIFA PLAYER RATING AND WAGE ANALYSIS

Data Source and Dimensions:

The dataset chosen for this project is the FIFA 19 complete player dataset obtained from Kaggle(https://www.kaggle.com/karangadiya/fifa19). This dataset includes game statistics of all players and roughly has 18,200 rows and 89 columns.

Data Format/Representation:

The data obtained from Kaggle is in CSV format. We process this dataset and convert it into a matrix representation where each row represents a player. Each column corresponds to a player's various attributes representing overall rating, wages, age and football skills like shooting, passing etc. We omit certain features that are guaranteed not to influence the analysis like Club logo, Name, Photo etc.

The process of converting the original dataset into a matrix representation involves conversion and normalization of various textual/unnormalized columns into normalized ones. In particular, textual features like Preferred Foot, Work Rate, Position etc. are converted into numeric fields based on some enumerated field mappings. Similarly, date features like Joining Date, Contract Validity etc. are also converted into numerical values such as time period in terms of days/years. After converting these columns, we normalize them since the values might be in different ranges. Since most of the skill features under consideration are in the scale of 100 by default, we normalize these textual/date converted columns to that same scale so that there is no bias between the features.

This normalized matrix representation with numerical features makes it convenient to perform dimensionality reduction techniques like PCA on the data as they are in the same units. Moreover, since the features are normalized, the rows can be considered as vectors and Euclidean distance relations can be applied while performing clustering on the data.

Data modelling:

The real dataset contains game statistics of players which are majorly numeric values. Each feature has some correlation with other features like a player whose position feature is 'CF' is more likely to have high values for Shooting and Dribbling features than for Defense and Tackling feature. There are many such feature correlations that can be drawn based on domain knowledge which gives us a way to model the data. This structure can be used to analyze the results as well as to generate more data for testing the scalability of the technique.