Proposal for Multivariate Data Analysis Project

Jerry Li

jl4533@drexel.edu

Dataset Selection and Relevance

For this project, I have selected the **FIFA 22** dataset, which contains extensive data on professional soccer players. The dataset includes 65 features, such as player attributes, ratings, and physical characteristics, with 16,710 samples. As a soccer enthusiast, I find this dataset particularly interesting because it provides a rich source of data that can be analyzed to uncover patterns and insights within the world of soccer. The diverse features in this dataset allow for a comprehensive exploration of how various attributes correlate with player performance, market value, and potential growth.

Source of Dataset

My dataset is download from Kaggle's FIFA Player Stats Database.

Description of Variables

The FIFA 22 dataset contains a comprehensive set of 65 features. Due to space limitations, I will highlight a few key variables as examples:

- **Overall Rating** : A numerical score representing the player's overall performance level.
- **Potential** : The projected future rating of the player, indicating their potential for improvement.
- **Position** : The specific role or position that the player occupies on the field (e.g., Forward, Midfielder, Defender).
- **Height and Weight** : Physical measurements of the players that could impact their style of play and effectiveness.
- **Skill Moves and Weak Foot** : Technical attributes that measure the player’s dribbling ability and proficiency with their non-dominant foot.
- **International Reputation** : A measure of how well-known the player is on the international stage, potentially impacting their market value.

These variables help us understand how different qualities affect a player's success in soccer. The variety in the dataset allows us to look closely at what influences both individual and team performance.

Initial Hypotheses

Based on the variables provided in the dataset, the following hypotheses will guide the analysis:

- **Hypothesis 1** : Physical attributes such as height and weight are more influential for players in defensive positions than those in offensive positions, as physicality is often crucial for defensive roles.
- **Hypothesis 2** : Goalkeepers with higher reflex ratings tend to have better overall ratings, indicating that reflexes are a crucial attribute for goalkeeping success.
- **Hypothesis 3** : Players from more prestigious clubs (based on club reputation) tend to have higher market values, suggesting that club prestige influences player valuation.

Expected Outcomes

Through this analysis, I want to confirm or reject these hypotheses by exploring the relationships between key variables. The findings will highlight factors that contribute to a player's success and offer insights for optimizing player development and scouting strategies in professional soccer.

Code

You can access the code for this project on my GitHub repository. It includes all the scripts used for data analysis, from data cleaning to the final models. You can find it here.