DAV Assignment 5 (Project write-up)

Group members:

Ashay Amul Shah (as3402@scarletmail.rutgers.edu)

Mansi Rajesh Khanna (mk1816@scarletmail.rutgers.edu)

Nirav Chandulal Gori (ng591@scarletmail.rutgers.edu)

Topic: FIFA players & team analysis

Technical software: We are going to use MS Excel for preprocessing, Python for data cleaning & transformation, & Tableau for data analysis, visualizations, dashboards & stories.

Dataset & our technical approach: We are going to use 2 datasets from the years 2015 & 2022 respectively, which are scraped from the official FIFA eSport game.

Here, we plan to compare player and team statistics, the overall changes that have occurred in the last 7 years, & also discover a few insights based on players' attributes, strengths, weaknesses, dominance, ratings, etc.

We will first start off with basic conditional formatting, looking up values & understanding our dataset in MS Excel. Then to preprocess, clean & transform the data further, we will be using Python. After cleaning the data and transforming the data, we will be loading our final datasets on Tableau. We will join them and then work towards insightful visualizations, dashboards & stories.

Data source, variables & limitations:

Source: https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset?select=players_22.csv

There are 16155 rows and 110 columns in the FIfa 2015 dataset, & 19,239 rows and 110 columns in the Fifa 2022 dataset.

The data allows multiple comparisons for the same players across the last 2 versions of the videogame, 2015 and 2022. Quality of the data is relatively fair when it comes to usability. We need to clean a few significant variables, & also ensure both the datasets have homogenous variables, for comparison purposes.

For preliminary analysis purposes, the variables of interest in our dataset would be:

- overall: This is a number from 0 to 100, which depicts the overall rating of the player
- international reputation: This is a rating out of 5, based on player's popularity.
- Skill_moves: This shows how skillful the player is ranging from 0 to 5.
- Work_rate: Player's work rate on the field (high, medium, low)
- Release_clause: Player's valuation in the market
- Position: player's position (left back, central midfielder, striker, goalkeeper, etc)
- Player_traits: (free kicks, long passes, leadership, flair, etc)
- Other footballing attributes like range, stamina, power, sprint, movement, crossing, shooting, etc

Graphs and charts we plan to create in our visualization:

- Histograms
- Bar charts
- Map charts
- Scatterplots
- Line Charts
- Slope Charts
- Boxplots
- Heatmaps
- Race bar charts
- Circular graphs
- And other animated, dynamic, interactive charts

Possible limitations:

- One limitation of this dataset would be that if in the future, we perform Data mining or apply machine learning techniques to this dataset, we would likely be facing the curse of dimensionality, as the number of columns are very much on the higher side.

- Another limitation is that the numerical data isn't clean enough, which we figured after doing the preliminary analysis. We will need to transform, and possibly do one-hot encoding in Python to make it analysis worthy.

- Thirdly, when we think of the objective positions of the players, the goalkeeper has the least row counts whereas midfielders have the highest counts in the dataset, so there is a likely class imbalance, and when we want to do some sort of classification analysis, data imbalance can be an issue.

Possible challenges:

- Predicting the starting lineup for any football club/country based on the player statistics, strengths, and positions of individual players of that team.
- To evaluate the ideal budget to create a competitive team.
- Comparing the top n% of players based on strengths.
- Understanding which feature has the highest impact on the player ratings.
- Understanding the key attributes required for every position on the footballing pitch.