FIFA Players Rating and Wage Analysis

Yuchen Wu, Xinmeng Zhang

- **Introduction and Motivation**

  FIFA has been one of the most famous and popular soccer video games in the world. It is well known not only for its excellent gameplay, but also for its large player dataset, which includes over 20,000 players around the world. In this dataset, the player ratings are well based on the actual performance of soccer players, which can reflect the ability of these soccer players in the real world.

  Therefore, besides the basic information that we want to collect from the players, we also would like to dive into this dataset and try to get an insight into the real world such as the trend of most valued attributes of players or the reasonability of the players' wage. For example, is it true that speed is more valued than other skills of players compared with before? Also, does higher wage mean that the player is more skilled and valuable to the team? Answering these questions is not only fun, but also offers a better understanding of the situation in the real soccer world.

- **Summary of research questions**
  1. Compare the attributes such as Agility, BallControl, or Strength of the top 5% players over the years. Has the indicator of good players changed? I.e. Can we observe which attributes have become more/less important?
  2. What is the most preferred foot of players in 2022?
  3. What are the Top 10 countries that have the most players each year?
  4. What is the trend of average height and weight of players by their nationality?
  5. What is the age distribution of current players in 2022 and the wage distribution by age and nationality?
  6. Has Messi and C. Ronaldo's ability on each attribute improved over the years? Summarize their personal characteristics based on the attribute value distribution.
  7. Can the wage amount paid to the players reflect their true abilities?

- **Method**

  Numbers are corresponding to that of research questions

1. Use pandas to select the top 5% players by their overall ratings from 2015 to 2022, then use seaborn and matplotlib to visualize the change of some key attributes over years in a line chart. Attributes include “pace”, “physic”, “shooting”, “passing”, “dribbling”, and “defending”
2. Use pandas to filter 2022 data by the column “preferred_foot” and then plot a Pie Chart using seaborn library to see the distribution of the most preferred foot used by players.
3. Use pandas to get the number of players for each country, then use seaborn and matplotlib to first make a cloudmap, and then a histogram to see the top 10 countries that have the most players.
4. Use pandas to group each year’s data by the column “nationality_name” and then calculate the mean of height and weight. Then combine the calculated outcome of each year together, then plot a line chart using seaborn with Year as x-axis and weight(kg) & height(cm) as two y-axis, color hue by height and weight.
5. We first use pandas to filter 2022 data by the column “age”, “nationality_name” and “wage”, then use seaborn and matplotlib to make a histogram of player distribution by age, a scatter plot of wage and age, and a scatter plot of wage and nationality
6. Use pandas to select the row of Messi and C. Ronaldo on each year and combine them together into one dataframe. Then group by their names to create a line plot using seaborn library to track their six attributes over the years. We can also make a polar graph using the Matplotlib library to see their current attribute distribution. We can then summarize their characteristics based on these graphs.
7. We will build a LASSO regression model using scikit-learn library in pandas, with wage as labels, overall rating and potential rating as features, using the 2015 - 2021 data to fit the model. Then we use the model to predict the wage of the 2022 data and see how well it predicts by looking at the mean squared error. If MSE is large, then we might be able to conclude that the wage is inflated and does not imply a corresponding improvement of the players’ abilities.

- **Dataset Used**

  We will use the sofifa dataset on Kagel to complete our analysis. There are multiple datasets that range from 2015 to 2022. The major variables in the

dataset include player's name, player id, overall rating, value, wage, league, and nationality and their attribute data such as Attack, Power and Shooting etc. The website to access the dataset is https://www.kaggle.com/stefanoleone992/fifa-22-complete-player-dataset

- **Challenge goals**
    1) Messy Data and Multiple datasets

        In our project, we'll be using multiple datasets from 2015 to 2022. We will try to combine these datasets first since we want to know the player ratings over the past few years instead of a single year. Besides, the dataset contains too much information and some of the values are missing. There are more than a hundred columns, and values include names, numbers, data, and URLs. To utilize the dataset, we need to make a lot of modifications first.

    2) Machine learning

        We want to use Machine Learning methods with the help of the scikit-learn library in pandas to create a LASSO model and fit the 2015 - 2021 data to see the predictability of the model on wage. Then we want to utilize the model to predict the wage in 2022 to see if players' ability and overall performance rating score can explain their wage amount, thus to tell if the wage is inflated. The difficulty is to determine what we want to predict and decide which feature and labels to use.

    3) Data Visualization

        We need to choose the appropriate type of graph when plotting based on the need of research questions. Which one most efficiently and accurately conveys the information that we want to get is a key and hard task to accomplish.

- **Work Plan**
    1) Data wrangling (2 - 3h)

Deal with all the garbleds and missing values in the data. Rearrange, filter or combine necessary datasets for time series analysis research questions.

2) Data Visualization coding (5 - 7h)

   Split the parts where need visualization into half and each one of us takes responsibility for our own parts. We can work individually at first and discuss it on Discord if we come up with any difficulties.

3) Machine Learning coding (3 - 4h)

   We want to do this together since we want to discuss in detail which labels to be included in the LASSO model and the interpretation of the outcome.

4) Script and interpret data analysis outcome (4 - 5h)

   Do this together using collaborative documents such as Google Doc.

- **Primary development environment**

  We plan to work on Ed workspace for the coding part and write papers on google doc which can be then uploaded to the Ed workspace. We also create a Github repository as a backup collaboration method if we find that necessary libraries won't work on Ed.