Graded Project

Working with data using Python Libraries, Visualization, EDA & Data Preprocessing.

Context:

A new football club named ‘GL United FC’ has just been inaugurated. This club does not have a team yet. The team is looking to hire players for their roster. Management wants to make such decisions using a data based approach.

During a recent hiring drive, you were selected for the Data Science team as a data scientist. Your team has been tasked with creating a report which recommends players for the main team. To start with, a total of 15 players are required. Player data for all teams has been acquired from FIFA. This data contains information about the players, the clubs they are currently playing for and various performance measures. The team needs 20 possible players to choose from. You have been requested to do the analysis and formulate a report in order to help the management make a decision regarding potential players.

Dataset Description:

The data contains details for over 18,000 players playing in various football clubs in Europe. It contains information on age, skill rating, wages and player value, etc. The files provided are as follows:

fifa.csv – data file.

fifa_ variable_information.csv - information on individual variables.

Questions: (Total points: 50)

- Load and explore data (4 points)
  1. Import the required libraries and read the dataset. (1 point)
  2. Check the first few samples, shape, info of the data and try to familiarize yourself with different features. (2 points)
  3. Drop the columns which you think redundant for the analysis. (1 point) [Hint: columns like 'Photo','Flag','Club Logo']

- Data Cleaning and Preprocessing (18 points)

  4. Convert the columns "Value", "Wage", "Release Clause" to float datatype after getting rid of currency symbol and suffix. (6 points)
     - Note: When the record/entry has "M"(indicates millions) as suffix you need to multiply that value with 1000000
     - When the record/entry has "K"(indicates thousands) as suffix you need to multiply that value with 1000
  5. Convert the column "Joined" into integer data type with keeping only the year. (2 points)
  6. Convert the column "Contract Valid Until" to pandas datetime type. (2 points)
  7. The column 'Height' is in inches with a quotation mark, Convert to float with decimal points. (2 points)
  8. The column "Weight" has the suffix as lbs, remove the suffix and convert to float. (2 points)
  9. Check for the percentage of missing values and impute them with appropriate imputation techniques. (4 points)

- Exploratory Data Analysis (28 points)

10. Plot the distribution of Overall rating for all the players and write your findings. (2 points)
11. Retrieve the names of top20 players based on the Overall rating. (2 points)
12. Generate a dataframe which should include all the information of the Top 20 players based on the Overall rating. (4 points)
13. What is the average "Age" and "Wage" of these top 20 players? (use the data frame created in the question 11) (2 points)
14. Among the top 20 players based on the Overall rating, which player has the highest wage? Display the name of the player with his wage. (2 points)
15. Generate a dataframe which should include the "Player name", "Club Name", "Wage", and 'Overall rating'. (4 Points)
    - i) find the average Overall rating for each club.
    - ii) Display the average overall rating of Top10 Clubs using a plot
16. What is the relationship between age and individual potential of the player? Visualize the relationship with appropriate plot and Comment on the same. (2 points)
17. Which features directly contribute to the wages of the players? Support your answer with a plot and a metric. (2 points) (hint: use potential, Overall, value, international reputation, and Release Clause)
18. Find the position in the pitch where the maximum number of players play and the position where the minimum number of players play? Display it using a plot. (2 points)
19. How many players are from the club 'Juventus' and the wage is greater than 200K? Display all the information of such players. (2 points)

20. Generate a data frame containing top 5 players by Overall rating for each unique position. (2 Points)
21. What is the average wage one can expect to pay for the top 5 players in every position? (use the data frame created in Q19) (2 points)

Submission:

- Please submit the solution file in .html and .ipynb format on Olympus
- Add necessary comments wherever required