**Mathematics For Computer Science**

**UE23MA242A**

**Teaching Assistants: Shubhika Verma, Abdullah Khan, Shruti Singh**

Problem Set: FIFA 20 Player Analysis Case Study

**Introduction**

The sections below, Background and Case Study, provide context for this data science assignment. This exercise will allow you to explore a real-world dataset from FIFA 20 using Python and analyze various features through descriptive statistics, tables, and visualizations. Happy coding!

**Background**

The dataset provided is derived from the popular video game FIFA 20 and includes detailed information on professional football players, capturing aspects such as demographics, physical attributes, skill ratings, and market values. By analyzing this data, sports analysts and gaming enthusiasts can gain insights into player characteristics, performance potential, and trends across different leagues, clubs, and nationalities. Additionally, this dataset enables football clubs and game developers to make data-driven decisions about player representation, game balance, and strategy development.

**Case Study**

A researcher, using the FIFA 20 player dataset, seeks to perform statistical analysis to explore various player characteristics. The goal is to uncover relationships among factors like age, skill ratings, physical attributes, and overall player value, as well as to identify patterns that could impact team composition or player development both in the game and in real-life sports contexts.

**Dataset Description**

The dataset contains the following primary variables:

1. **sofifa\_id**: A unique identifier for each player.
2. **Short Name**: The commonly known name of the player.
3. **Long Name**: The full name of the player.
4. **Age**: The player’s age at the time of data collection.
5. **DOB**: The date of birth of the player in DD-MM-YYYY format.
6. **Height (cm)**: The player’s height in centimeters.
7. **Weight (kg)**: The player’s weight in kilograms.
8. **Nationality**: The country represented by the player.
9. **Club**: The football club for which the player currently plays.
10. **Overall**: A numerical rating representing the player’s overall skill level.
11. **Potential**: A numerical rating representing the player’s potential skill level.
12. **Value (EUR)**: The estimated market value of the player, expressed in euros.
13. **Wage (EUR)**: The player’s weekly wage, expressed in euros.
14. **Preferred Foot**: Indicates whether the player is left-footed or right-footed.
15. **International Reputation**: A rating indicating the player’s global profile and popularity.
16. **Weak Foot**: A rating of the player’s ability to use their non-dominant foot.
17. **Skill Moves**: A rating indicating the player's skill move abilities.

**Problem Set**

Solve the below given questions. Bonus questions, if any can be solved if you have additional time left but are not mandatory. Questions that aren’t applicable to the dataset provided, or questions that yield irregular output maybe omitted/the irregular output will be accepted.

**Unit 1: Data Understanding and Preprocessing**

1. **Feature Classification**:
   * Group the variables in the FIFA dataset into suitable data categories (ordinal, nominal, interval, or ratio). Explain why each category applies to the variables you assign to it.
2. **Data Quality Issues**:
   * Identify and explain any issues or inconsistencies with the dataset. Outline the steps you would take to clean and prepare the data, making it ready for accurate and reliable analysis.
3. **Summary Statistics**:
   * Provide a summary of each feature’s central tendency and dispersion. Describe the best measures of central tendency for each attribute, and calculate standard deviation and range for all numerical columns.
4. **Visualizations**:
   * Create histograms and box plots for **'Age'** and **'Overall'**.
     + **i)** Describe the type of distribution you observe for each variable.
       - *Hint: Adjust the visual scale for better clarity.*
     + **ii)** Calculate the count of outliers in each variable.
     + **iii)** *Bonus*: After determining the type of distribution, adjust and re-plot the histograms and box plots for **'Overall'** based on any notable characteristics in the distributions.
5. **Outlier Handling**:
   * Explain the actions you would take to address outliers in the dataset. Show the before and after results with updated visualizations, like histograms and box plots.
6. **Normal Probability Plot (Q-Q Plot)**:
   * Create and interpret a Q-Q plot for **'Overall'**. Based on the plot's shape, describe the conclusions you can draw about its distribution.
7. **Correlation Analysis**:
   * Find the correlation between **'Age'** and other numerical variables, identifying the attribute with the strongest correlation to **'Overall'**.
8. **Pair Plot Analysis**:
   * Select a random sample of 10,000 entries and create a pair plot using **'Age'** and **'Overall'** with **'Potential'** as the hue. Describe any relationships or patterns you notice.

**Unit 2: Hypothesis Testing**

1. **Hypothesis Testing**:
   * Formulate a null and alternative hypothesis to check for any significant difference in median **'Overall'** scores across different age groups. Use the Mann-Whitney U test to evaluate this relationship.
2. **Margin of Error**:
   * Determine the margin of error to evaluate the precision of your previous analysis. Explain what your findings suggest about the reliability of the results.

**Unit 3: Prediction and Feature Engineering**

1. **Linear Regression Analysis**:
   * Use **'Age'** and **'Potential'** to predict **'Overall'** ratings with linear regression. Plot the predicted values against actual ones and compute both MSE and RMSE.
2. **Feature Engineering**:
   * With **'Age'**, **'Overall'**, and **'Potential'** representing a range of player attributes, suggest two new features that could enhance predictions of **'Overall'**.