

# Python Programming

## Assignment 1 – 20%

Deadline: 28<sup>th</sup> April 2025

### I. (3 points)

Write a Python program to collect footballer player statistical data with the following requirements:

- Collect statistical data [\*] for all players who have played more than 90 minutes in the 2024-2025 English Premier League season.
- Data source: <https://fbref.com/en/>
- Save the result to a file named 'results.csv', where the result table has the following structure:
  - Each column corresponds to a statistic.
  - Players are sorted alphabetically by their first name.
  - Any statistic that is unavailable or inapplicable should be marked as "N/a".
- [\*] The required statistics are:
  - **Nation**
  - **Team**
  - **Position**
  - **Age**
  - **Playing Time:** matches played, starts, minutes
  - **Performance:** goals, assists, yellow cards, red cards
  - **Expected:** expected goals (xG), expected Assist Goals (xAG)
  - **Progression:** PrgC, PrgP, PrgR
  - **Per 90 minutes:** Gls, Ast, xG, xGA
  - **Goalkeeping:**
    - Performance: goals against per 90mins (GA90), Save%, CS%
    - Penalty Kicks: penalty kicks Save%
  - **Shooting:**
    - Standard: shoots on target percentage (SoT%), Shoot on Target per 90min (SoT/90), goals/shot (G/sh), average shoot distance (Dist)
  - **Passing:**
    - Total: passes completed (Cmp), Pass completion (Cmp%), progressive passing distance (TotDist)
    - Short: Pass completion (Cmp%),
    - Medium: Pass completion (Cmp%),
    - Long: Pass completion (Cmp%),
    - Expected: key passes (KP), pass into final third (1/3), pass into penalty area (PPA), CrsPA, PrgP
  - **Goal and Shot Creation:**
    - SCA: SCA, SCA90
    - GCA: GCA, GCA90

- **Defensive Actions:**
  - Tackles: Tkl, TklW
  - Challenges: Att, Lost
  - Blocks: Blocks, Sh, Pass, Int
- **Possession:**
  - Touches: Touches, Def Pen, Def 3rd, Mid 3rd, Att 3rd, Att Pen
  - Take-Ons: Att, Succ%, Tkld%
  - Carries: Carries, ProDist, ProgC, 1/3, CPA, Mis, Dis
  - Receiving: Rec, PrgR
- **Miscellaneous Stats:**
  - Performance: Fls, Fld, Off, Crs, Recov
  - Aerial Duels: Won, Lost, Won%
- Reference: <https://fbref.com/en/squads/822bd0ba/Liverpool-Stats>

## II. (2 points)

- Identify the top 3 players with the highest and lowest scores for each statistic. Save result to a file name 'top\_3.txt'
- Find the median for each statistic. Calculate the mean and standard deviation for each statistic across all players and for each team. Save the results to a file named 'results2.csv' with the following format:

|     |        | Median of<br>Attribute 1 | Mean of<br>Attribute 1 | Std of<br>Attribute 1 | ... | ... |
|-----|--------|--------------------------|------------------------|-----------------------|-----|-----|
| 0   | all    |                          |                        |                       |     |     |
| 1   | Team 1 |                          |                        |                       |     |     |
| ... |        |                          |                        |                       |     |     |
| n   | Team n |                          |                        |                       |     |     |

- Plot a histogram showing the distribution of each statistic for all players in the league and each team.
- Identify the team with the highest scores for each statistic. Based on your analysis, which team do you think is performing the best in the 2024-2025 Premier League season?
- Histogram Plot: [https://matplotlib.org/stable/api/as\\_gen/matplotlib.pyplot.hist.html](https://matplotlib.org/stable/api/as_gen/matplotlib.pyplot.hist.html)

## III. (3 points)

- Use the K-means algorithm to classify players into groups based on their statistics.
- How many groups should the players be classified into? Why? Provide your comments on the results.
- Use PCA to reduce the data dimensions to 2, then plot a 2D cluster of the data points.

#### IV. (2 point)

- Collect player transfer values for the 2024-2025 season from <https://www.footballtransfers.com>. Note that only collect for the players whose playing time is greater than 900 minutes
- Propose a method for estimating player values. How do you select feature and model?

#### Submission Instructions:

- The submission should include Python code.
- A report (.pdf) including your justification and results .
- Submit your work to your personal github and send me the link.
- Thank you!