ProtATonce data analysis assignment

Introduction

The goal of this assignment is to write code in order to analyse two different datasets. Keep in mind that there is no correct way to perform the required tasks. Your assessment will be based on your creativity, code organization, the ability to describe your methodology, and your way of thinking. This assignment has two parts, in the first part you are given a dataset of NBA players' statistics, while in the second part you are given a dataset of single-cell measurements. Perform as many of the tasks for each part of the assignment as possible, while describing your reasoning in detail.

Part A. NBA dataset.

For this part, you are given three datasets, which contain information for individual basketball players from 67 NBA seasons. A glossary of the column names and what they represent can be found here https://www.basketball-reference.com/about/glossary.html. Your assignment is to write code in order to perform the following tasks:

- Preprocess and merge the datasets in order to transform them into a tidy dataset. More specifically, the final data table should contain player statistics per year and each column should represent a unique feature.

- Perform data exploration by analyzing specific features of the dataset. For example, aggregate statistics, visualizations, etc. (Limit your analysis to 3 features)

- Form a statistical hypothesis and design a test for its significance. Examples of statistical hypotheses could be:
    - The efficiency of all players that started in more than 10 games a season in the period of 1990-2000 is higher than the period of 1970-1980.
    - The height of players is associated with the position that the play.

Part B. Single-cell dataset.

For this part, you are given a dataset containing "single-cell" measurements: 20 features (proteins) that describe each cell (features x_1, x_2, etc.). Your assignment is to write code in order to perform the following tasks:

- Identify potentially different types of cells (groups) based on the cell features, assign each cell to a group and visualize the results. You are free to use any method(s) you want, but make sure that you explain the reasoning and thought process behind your choice.

- Identify the 5 most important features of each group (use the given column names) and provide a representative set (vector) of features for each group.

If you do not manage to perform the first task, in order to continue to the second, you can randomly assign cells to a fixed number of groups.

Deliverables and notes.

Deliverables: For each part of the assignment, submit a report that shows all the code, plots and written summary of your reasoning and methodology. You can either submit a pdf, an R markdown file or a Jupyter notebook. Additionally, GitHub project submissions are also acceptable.

Notes: Keep your code clean and write detailed comments in order to make it comprehensive.

Due date: Please submit your work before Sunday, January 31stth at 12 pm (GMT+2).

Good luck,

Sincerely,

The ProtAtOnce team.