laSalle
UNIVERSITAT RAMON LLULL

Assessment 3.
Descriptive analytics
GIA04 – Knowledge-based systems

The dataset *educacio.csv* contains the results obtained by sixth grade students in the evaluation of basic skills and knowledge at the end of primary education since 2018.

We want to analyze this dataset to provide information about the profile of 6th grade students and their annual evolution.

The description of the dataset columns is as follows:

1. ANY: year in which the evaluation took place.
2. PCAT: weighted overall score of linguistic competence in Catalan.
3. PCAST: weighted overall score of linguistic competence in Spanish.
4. PMAT: weighted global score of mathematical competence.
5. PANG: weighted overall score for English language proficiency.
6. PMED: overall score of the competence associated with the natural environment's knowledge.
7. GENERE: gender of the student who submits to the evaluation.
8. ANY_NAIXEMENT: year of birth of the student who submits to the evaluation.
9. AREA_TERRITORIAL: region where the student's center taking the assessment is located.
10. NATURALESA: determines if the student's center is public or private.
11. HABITAT: municipalities by population groups.

You are asked to perform the following analysis:

1. Student analysis

1.1. Display in different graphs the number of students per year, by year and nature (public or private), by year and gender, and by year and habitat.
1.2. Discuss your conclusions.

2. Analysis of the average of the grades

2.1. Calculate the average of the 5 grades (PCAST, PCAT, PANG, PMAT, and PMED) and them to your data.
2.2. Display a histogram of the average of the 5 grades you computed.
2.3. Discuss your conclusions.

3. Analysis of the quartiles of the grades

3.1. Display the boxplots of the 5 tests and the average per year.
3.2. Display the boxplots of the 5 tests and the average by gender.
3.3. Display the boxplots of the 5 tests and the average by NATURALESA (public or private).
3.4. Display the boxplots of the 5 tests and the average per municipality.
3.5. Display the boxplots of the 5 tests and the average by region.
3.6. Discuss your conclusions.

4. Correlation analysis

4.1. Display the correlation matrix of the variables. Convert to numeric the ones you think are necessary (justify it).

4.2. Discuss your conclusions.

5. Additional analyses

5.1. Add any additional analyses that you deem appropriate.

5.2. Discuss your conclusions.

6. Identification of student types (OPTIONAL)

6.1. Apply the clustering algorithm k-means on the grades of the 5 subjects. Analyse the results obtained with different numbers of clusters and choose the grouping that you consider most appropriate. Justify the answer.

6.2. About the chosen solution, explain what type of student belongs to each cluster group. You can use boxplots or histograms discriminating by the cluster identifier (you will have to add the cluster assigned to each row to the student table). To save the cluster assigned to each student in the data table, you can do (*data* is the dataframe that contains the students' data and *kmeans* the KMeans object)*:* data['cluster'] = kmeans.labels_

Deliverable

A single .ipynb file **must be submitted**, containing:

1. **Implementation and conclusions obtained from each of questions 1 to 5 (8 points)**
   - A Jupyter Notebook using the Python programming language.
2. **Implementation and conclusions obtained from question 6 (2 points, optional)**
   - To be done in the same Jupyter Notebook above, using the Python programming language.

Considerations

- In the **eStudy** you will find **the Jupyter Notebook** and the **dataset**.
- **Implement with the** Python **programming language** in the **provided Jupyter Notebook**.
- **It is done individually.**
- **The conclusions** of the results obtained in each question will be placed in the same **Jupyter Notebook.**
- The **conclusions** must be **well written** and without spelling mistakes.