ISEP
IF.1205 – Data Science Fundamentals
May 2025

DATA SCIENCE PROJECT

Principal Component Analysis (PCA) and Linear Regression

I. General instructions

The goal of this project is to apply principal component analysis and linear regression to real-world datasets. The project will be carried out by teams of **three students**. Each team will present their work during an oral presentation on **Tuesday, June 3rd.** You can work in R or Python. Instructions for the oral presentation and evaluation are detailed below.

a) Instructions for the oral defense

The oral presentation will last approximately 15 minutes per group, consisting of a 10-minute presentation and a 5-minute question period. The slides must contain the following elements:

- A cover page containing the first name, last name, and student ID of all members of the team.
- A summary.
- A short introduction.
- The body of the document (results, figures, tables, interpretations, comments, or any other element that allows questions to be answered). Answers to the questions asked in the presentation must be clearly indicated in this section.
- The conclusion.
- References.

Your R or Python code should not be included in your presentation. However, you must have this code on hand at the time of the presentation to answer any question you may have.
A single member of your team has to put the presentation file in PDF format on Moodle by **noon on Monday, June 2nd**. A Moodle repository will be created for each tutorial group (G7, G8, G9, and G10). The name of the file should follow the following format:

NameStudent_NameStudent2_NameStudent.pdf

Additionally, each team will put a csv file containing the quantitative answers to the questions asked (indicated by [qij] in the statement) on the same Moodle repository. The name of this file should be in the form:

NameStudent_NameStudent2_NameStudent.csv

A template for this file is available on Moodle (**template.csv**). A single member of your team will need to edit this csv file and upload it to Moodle **by noon on Monday, June 2nd.**

b) Instructions on the evaluation

The oral presentation will be divided in two parts: a 10-minute oral presentation and a 5-minute question period. Evaluation will be both collective, particularly regarding the overall quality and content of the presentation, and individual.
Each student will be evaluated on his contribution during both parts. Particular attention will be paid to the analytical quality of his responses.

II. Data analysis

a) The dataset

Meteorology is the science that studies atmospheric phenomena. One of its challenges is forecasting, locally and in the short term, meteorological variables such as daily precipitation, sunshine and temperature, as well as extreme events such as floods, heatwaves and cyclones. Reliable forecasts are therefore essential to ensure the safety of the population, or in sectors such as transport, particularly air transport. This reliability also has economic repercussions in areas such as agriculture.
In this context, the aim of this project is to first analyse meteorological data obtained in 2024, and then to propose a model to predict monthly temperatures in 2025.
The meteorological data is available in the files **data1.csv** and **data2.csv.**

b) Preliminary analysis: descriptive statistics

The **data1.csv** file contains meteorological data measured in France in 2024, and more specifically:

- the average minimum temperature measured in °C,
- the average maximum temperature measured in °C,
- the total rainfall measured in mm,
- the total sunshine duration measured in hours.

After importing the **data1.csv** file, answer the following questions:

1. From how many cities do the weather data is extracted [q1a]? How many cities are affected by missing measurements [q1b]? **If existing, delete these cities. They will not be considered in the following of our analysis.**

2. What is the city associated with the:
   - lowest minimum temperature [q2a],
   - highest minimum temperature [q2b],
   - lowest maximum temperature [q2c],
   - highest maximum temperature [q2d],
   - lowest rainfall [q2e],
   - highest rainfall [q2f],
   - lowest sunshine duration [q2g],
   - highest sunshine duration [q2h].

   Also indicate each associated value [q2a-h]. Comment.

3. Compute the variance of the:
   - minimum temperature [q3a],

- maximum temperature [q3b],
- total rainfall [q3c],
- sunshine duration [q3d].
Comment.

4. Compute the mean [q4a], median [q4b] and standard deviation [q4c] of the weather variable with the lowest variance computed in the previous question. Display the associated histogram. Comment.

5. Compute the mean [q5a], median [q5b] and standard deviation [q5c] of the weather variable with the highest variance computed in the previous question. Display the associated histogram. Comment.

6. We are interested in the linear correlations between the different weather variables. What are the two most positively correlated variables [q6a]? What are the two most negatively correlated variables [q6b]? What are the two least correlated variables [q6c]? Also display the associated correlation values [q6a-q6c]. Illustrate your results graphically, adding the names of the cities to each of the three figures you have created. Comment.

7. We are now interested in the linear correlations between the cities. Display the correlation matrix. Comment.

c) Principale Component Analysis (PCA)

We now apply a Principal Component Analysis (PCA) to the four weather variables in the data1.csv dataset. First, we center and reduce our data using the following formula:

$$X_i' = \frac{X_i - \mu}{\sigma}$$

where X_i represents a meteorological variable, μ its mean and σ its standard deviation.

8. Apply a PCA to the centered and reduced weather data. Display the first two principal components in the form of a scatter plot to visualise the results. Add the names of the cities to the figure. What percentage of the variance is explained by the first [q8a] and second [q8b] principal components? Also display these values on the axes of your figure. Comment.

9. Display and comment on the correlation circle.

10. Superimpose the results of questions 8 and 9 on the same figure. Can you find the results of question 2 from this figure?

d) Simple linear regression

We will now look at the maximum temperature measured in 2023 and 2024 in Paris, available in the file **data2.csv**. We will first focus on 2024.

11. Display the evolution of the temperature in 2024 in Paris in function of the month. Comment.

First, we seek to predict the maximum temperature in Paris in January 2025 using data obtained in Paris in 2024. The linear regression model we are testing is written as :

$$Maximum\ temperature = \beta 0 + \beta 1 * month\_ID + \varepsilon \qquad (1)$$

Where *month_ID* represents the position of the month in the year, ranging from 0 for January to 11 for December. We have chosen to apply simple linear regression taking into account the last n months, with n ranging from 1 (modelling from December 2024 only) to 12 (considering the months from January to December 2024). The selected model is the one with the highest value for the adjusted coefficient of determination R2.

12. Apply the n linear regressions. What is the optimal value of n [q12a]? What is the value of the associated coefficient of determination R2 - adjusted [q12b] or not [q12c]? Also give the values of β0 [q12d] and β1 [q12e] predicted by the optimal model. Analyse your results both quantitatively and visually.

13. The maximum temperature in Paris in January 2025 was 7.5 °C. What is the temperature predicted for January 2025 by the model from the previous question [q13a]? What is the difference between the predicted and actual temperature for January 2025 [q13b]?

14. Evaluate the null slope hypothesis for the coefficient β1 obtained in question 12. What is the p-value [q14a] obtained for the associated test? Is there a linear relationship between the two variables taking α=5% [q14b]?

e) Multivariate linear regression

Secondly, we plan to predict the maximum temperature in Paris in January 2025 using data obtained in Paris in 2024 and 2023.

15. Superimpose the evolution of the temperature in 2023 and 2024 in Paris in function of the month on the same figure, using a different color for each year. Comment.

We assume that there is a linear relationship between the maximum temperature measured in a given month and the maximum temperature recorded during the previous months. We train our multivariate linear regression model on the 12 months of 2024. For each month i of 2024, there are at most 12 available variables, corresponding to the maximum temperature of month i - 1 up to the maximum temperature of month i - 12. The variable to be predicted corresponds to the maximum temperature in month i.

16. How many combinations of variables are possible for predicting the temperature at a given month [q16a]? Compute the combination that maximises the adjusted R2 coefficient. What is the number of selected variables [q16b]? Display the optimal adjusted R2, the selected variables and the associated parameters. Comment on the results. Is there a linear relationship between the selected variables taking α=5%?

17. Compute the difference between the maximum temperature predicted by the model in the previous question for January 2025 and the measured maximum temperature (7.5°C) [q17]. Comment. Can this model be used to predict the maximum temperatures for February, March and April 2025 (the measured temperatures being

respectively 8.6°C, 14.6°C and 20°C)? Comment.

III. References

- https://meteofrance.com/