

Exercise 1:

Let us consider the two tables below:

	Company	N customers	Rating mean	Year
1	Djeezy	10.000.000	7.5	2018
2	Mobilis	9.000.000	4	2019

Table 1

	Company	N customers	Rating mean	Year
1	Ooredoo	11.000.000	8	2020
2	Mobilis	9.000.000	4	2021

Table 2

	Company	Expected revenues	Year
1	Ooredoo	12.000.000.000 DA	2023
2	Mobilis	10.800.000.000 DA	2023
1	Ooredoo	17.000.000.000 DA	2024

Table 3

- Create python script to manage these tables as data frames.
- Concatenate both tables (1 et 2) by:
 - Keeping all the information.
 - Keeping recent information.
 - Keeping the company name and the N of customers.
- Merge table 3 with the resulted table, with respect to:
 - The year and company.
 - The company
- For the available information, draw a graph representing the company, the year, and the expected revenues of each.
- If we consider that Ooredoo revenues grow up linearly, draw a line that represents the increase of revenues and calculate the expected revenue by 2025.

Exercises 2:

Let us consider the IRIS dataset (the CSV format). Implement using python the next statistical measures:

1. Min, max, mode, median, standard deviation.
2. Normalize the data using Z-score.
3. Calculate the correlation table (correlation of each two variables.).
4. Visualize the correlation of each two variables.
 - a. What do you conduct?
5. Apply PCA on the Data and keep only the two most significant dimensions.
 - a. Visualize the correlation of each two new variables.
 - b. What do you conduct?

6. Apply a random sampling to retrieve 10 data folds of data (by taking 10% of the data without repetition to compose each fold).
 - a. Create a training set by combining 9 folds among the 10 ones.
 - b. Calculate for each sample: the density, the variance.

Exercise 3:

Take the same dataset of the first PW series then:

- Check that not all numerical values include a string.
- Check that dates are correct.
- Check that the request id field respects the specific forma:
 - “US or CA-YYYY-NUMBER”.

Exercises 4:

- Retake the exercise about data smoothing in DW series and use *scipy* to smooth the data.

Exercise 5:

Let us consider the next weather dataset; using pandas (i.e., check `get_dummies` function) encode the feature **temperature** using:

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

1. Ordinal
2. One-Hot
3. Binary
4. Frequency

Exercise 6:

- Retake the same dataset of the first PW series and generalize the request date.

Exercise 7:

Let us consider the next table containing small set of mice information:

	Age (days)	Weight (grams)	Tall (cm)	Class
1	25	120	8	Obese
2	25	120	8	Obese
3	30	80	7.5	Not Obese

- Which problem it contains this data?
- Create a python function to handle it.