PROJECT

Dr. Matthieu Cisel

Bachelor 2

1 Pandas exercises : mastery of the basics

In this learning unit, you will demonstrate your ability to analyze a dataset in a semi-autonomous manner. Given the fact that mastery of pandas for data wrangling is one of the main objectives of this class, we will start with exercises that you can do with ChatGPT, notably, and move afterwards to the semi-autonomous analysis of a real life dataset. You must submit an ipynb version of your work for the first part of the learning unit, and both an ipynb and a small report that follows the IMRAD structure, as far as the second part is concerned.

1.1 Simplest operations

1. Create two simple datasets df1, df2 (a few lines, a few columns), similar in terms of dimensions
2. Each column should have a name (like a, b, c, d, etc.), values should be numbers generated randomly using numpy
3. For df1, change the value of the first line, first column, to NaN (missing value)
4. Using fillna, change this missing value to 0
5. Export df1 as a csv database
6. Merge these datasets, once vertically, once horizontally, and display the results in your notebook
7. Gather all the columns of df into only one column using the melt function
8. Create df3, a dataset modified from df1, where you have shifted the values of one column by one, downwards

1.2 A geographical approach to completion rates in France

You must first download the dataset called "Formations engagées", from the CDC open data repository. We want to see what proportion of learners who have registered actually complete the training, depending on the "department" (geographical location). For statut_dossier : Réalisation totale : fully completed the course. Réalisation partielle : did not fully complete the training. Annulation titulaire : the registrant cancelled the training session. The training center can also cancel the training session for a given person before it even begins.

1. Show the head of intermediate datasets whenever possible, for the sake of clarity
2. Do not forget to reset the index when relevant

1.2.1 Focusing on subsetting

1. Summarize the price of an average training session for different departments (a line = a training session) for a subset of the dataset (Departments = Paris U Ain). Use the isin function
2. Repeat the previous operation, but compute the cost of the average training session for a sample of the original dataset (25%) - all departments together. Do it one with the query function, and once without it
3. Compute the cost of the average training session for training sessions where cost is above 1000 euros (entire France)
4. Compute the cost of the average training session for the top 20 most expensive training sessions
5. Compute the cost of the average training session for the dataset where intitule_formation includes the word "Yoga"
6. Replace all instances where the "Yoga" pattern appear by "Cours de yoga"
7. Compute the cost of the average training session for the subset of the dataset where intitule_formation beginning with the word "Certificat"
8. Compute the cost of the average training session for n the subset of the dataset where intitule_formation ending with the word "compétences"
9. For the latest subset, in addition to the average cost, plot a distribution of the cost through a histogram

1.2.2 Using pandas on a real life dataset

1. Screen the dataset for duplicates (there are none, but still)
2. Create a new variable by multiplying nb_dossiers (number of files) with prix_moyen (average cost), and assess to what extent this new variable (whose name you will choose wisely) matches montant_engage. Use the == at some point
3. Create a table where you compute the cumulative sum of montant_engage for all departments, using groupby, and then subset this table so that you have only the departement of Paris
4. Perfom a value count of the number of training sessions per region in 2022 (one line is one training session), using the command size. It creates the df called train_reg. Then download the population size of the Region (at 1st January, 2023) from this address. Merge both datasets (train_reg and regions) to create an intermediate where the number of row corresponds to the number of distinct regions. Specify which type of merge you did and explain why you used this one. Based on this new dataset, compute for each region the % of training sessions per habitant
5. Go back to the original dataset "Fondations engagées". Remove data where status_dossier is not "Clos" (by the way, why is it useful if we want to compute a completion rate ?). Use a str.contains
6. Develop a strategy to have the % of "Réalisation totale" (with regards to the 4 types of "dossiers" that are closed), according to the department. Provide a justification

2 Exploratory analysis

Stick to these dataset. Use the skills that you have acquired in the previous exercises to provide a few analyses of your choices, along with graphs made from matplotlib. Make a short introduction, a paragraph or two on methods (data, libraries, etc.). Describe your results, and in a discussion part, provide an interpretation. This is the IMRAD structure.