Data Description

The objective of exercises in this class is not to learn how to code, but to understand the theory underlying data mining tools, together with their interest and limits. Therefore, you can use any library you want. I'm guiding you to use some convenient libraries.
You can use LLM (AI), but at the condition that you understand each and every line of code written in your file.

1 Fundamentals

1. Toy dataset: Correlation
   When analyzing data, a first important aspect to understand it is to analyze the relation between variables. For instance, if I want to predict if chocolate can cure cancer, I can start by checking if people eating more chocolate tend to have less cancer. We can use correlation coefficients to do so naively. However, we will see that these coefficients should be analyzed with caution
   (a) Load the dataset `coffee_effects.csv` found on the class website
   (b) Plot the first few lines to check the content (`.head(2)`), or simply write the name of the dataframe in a notebook)
   (c) Check the Pearson correlation between variables. You can directely use `df.corr()` for instance. You can plot the correlation matrix for instance with `seaborn` library with a command such as `sns.heatmap(correlation, annot=True, cmap='coolwarm', fmt=".2f")`
   (d) Do the same with Spearman (option `method='spearman'`)
   (e) Observe the correlation values and try to interprete the relation between the variables
   (f) Draw a scatterplot with the caffeine consumtion on the x-axis, and another variable in the y-axis. Repeat for all 3 variables. Use `plotly` library to have an interactive plot, for instance with the command the command
   `px.scatter(df, x='caffeine_consumption_mg', y='productivity')`
   (g) For each relation, you should observe something unusual. Ask yourself questions such as: is the relation linear? Is it monotonous? How strong is the relation between changes in the two variables?
   (h) Using this information, can you propose a strategy to improve the productivity of a company? Should it offers free coffee to employees ? If yes, unlimited ?

2. p-values

   (a) Using `stats.pearsonr` function, obtain the p-value for the caffeine consumption VS heart rate. Conclude.
   (b) Imagine now that we have less data: take a sample of the dataset with 5, 10, 15, 20, 50, 100, 1000 random samples. For each of them, compute the correlation and p-value. Conclude
   (c) Run the experiment a few times, and compare the results.

3. Car Dataset: Exploration and Cleaning
   (a) Load the dataset `cars_synthtic.csv` .

   (b) Compute the classic descriptors of the `price` column using pandas’ `describe` function. Check the mean, std, percentiles, and extreme values...

   (c) Plot the distribution of the `price` variable using a histogram. You can directly use pandas plotting tools (`df[col].plot.hist()`). Vary the number of bins using the `bins` parameter to see if the distribution seems to follow a bell curve. Use a `kde` plot instead of a `hist` .

4. Data Cleaning

   (a) Describe the variable `length` in the same way as above. Observe that you encounter difficulties, and try to find the cause. Search for abnormal values...

   (b) Fix the problem temporarily by replacing erroneous values by `np.nan` . You can for instance use dataframe indexing, like: `df.loc[df[’B’] < threshold, ’B’] = np.nan`

   (c) Now, check the weight column. Identify the problem and solve it.

5. Quick exploration

   (a) To quickly visualize various information about your dataframe, you can use a dedicated tool. For instance, install the `dtale` package, and apply it to your dataframe using the `dtale.show(df)` function. You can directlty explore many aspects of your data directly from a browser. You could also have used packages `DataPrep` , `SweetViz` , `AutoViz` , etc.

2 Going Further

6. Mastering the scores

   To be sure to understand correctly what is going on, we will write code to compute manually simple scores. For each question, use the native function in python corresponding to those scores to check that your function is correct

   (a) Write a function to compute the variance. Compute the variance of a column.

   (b) Write a function co compute the MAD (mean absolute distance to the mean or median).

   (c) Write a function to compute the covariance between two variables

   (d) Using the covariance, write a function to compute the Pearson (linear) correlation coefficient

   (e) Using the Pearson CC function, write a function that compute the Spearman CC (you can use for instance `scipy.stats.rankdata`).