TD1/5: Project

Exercise 1: Set up

1. Load data:
   masse-salariale-et-assiette-chomage-partiel-mensuelles-du-secteur-prive_modif.csv[1]

2. See number of samples (rows) and features (columns)

3. See data type

4. Set *dernier_jour_du_mois* as index

5. Cast index as datetime

6. Sort index in ascending order

Exercise 2: Data Analysis

1. Discover data:

   - Visualize (plot) data (can be done in one simple line of code)

Exercise 3: Data Cleaning

1. Check for missing values (one might be more subtle than a yelling NaN)

2. Impute these missing values with at least 2 methods seen in the lectures, don't delete them in this project (imputing is more difficult than deleting)

3. Check and treat outlier(s)

Exercise 4: Feature Engineering

1. Add a feature *is_year_end*

   - 1 when month is november or december
   - 0 otherwise

Exercise 5: Prediction

1. Split your data into a train set (70% of data) and a test set (30%)

2. Use a linear regression to predcit *part_assiette_chomage_partiel* 1 month ahead

   - you should shift your features (in time) compared to your target
   - find tutorials, there are a lot of them, its the only way toward autonomous learning!

[1] Data is a modified version from this source

3. How good is your prediction?
   - Use metric(s) to evaluate your model on both the train and test sets
   - Interpret the results
   - Give advices to your (hypothetical) colleague to continue your work

Exercise 5:.1 Bonus

1. Make a prediction without the added variable *is_year_end*
2. Use a Ridge regression in place of the Linear regression (you might become happy about the results!)
3. Use a **polynomial** regression to predcit 1 month ahead (find tutorials, there are a lot of them, and its the only way to learn autonomously!)
4. Predict 2 months ahead, then 3 and 4 months ahead. If your code is written correctly, it should only require to manually change the value of a constant.