TD1/5: Project

Exercise 1: Set up

1. Load data:
   masse-salariale-et-assiette-chomage-partiel-mensuelles-du-secteur-prive.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)
   - Use standard functions to get descriptive statistics of each variable

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 4 months ahead
   - you should shift your features (in time) compared to your target

[1] Data is a modified version from this source

- find tutorials, there are a lot of them, its the only way toward autonomous learning!

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
4. Make a prediction without the added variable *is_year_end*

Exercise 5:.1 Bonus

1. Use a **polynomial** regression to predcit 4 months ahead (find tutorials, there are a lot of them, and its the only way to learn autonomously!)
2. Predict 1 month ahead, then 3 months ahead. If your code is written correctly, it should only require to set a constant to 4, 1 or 3.