**A picture containing text, sign

Description automatically generated**

|  |  |
| --- | --- |
| **2023 MFRE SUMMER PROGRAM** |  |
| **Python Workshop 2: Data Manipulation with Pandas** | |
| **CASE** | |
| Now we’re getting into the fun stuff! We’ll introduce the `pandas` package and its two main data structures: DataFrames, and Series. Then, we’ll explore the workflow for importing, investigating, adding to and removing from our data. All of these steps are vital for setting data up for modelling or visualization at a later point. | |

1. **Introduction to `pandas`**
   1. DataFrames
   2. Series
2. **Data Cleaning and Exploratory Analysis**
   1. Data Import with `pd.read\_csv()`
   2. Checking rows and columns with `.shape`
   3. Summing columns with `.sum()`
   4. Checking column max/min with `.max()`, `.idxmax()`, etc
   5. **Slicing and selection with `.loc[]`, `.iloc[]`**
   6. **Filtering with Boolean conditions**
   7. **Column arithmetic operations**
      1. **E, F, G: Your absolute core operations for working with DataFrames! If you learn anything, learn these!**
   8. Summary statistics
   9. Open-ended exercise

**Case Study: Exploring Canadian Carbon Dioxide Emissions**

**Source of Dataset:** https://www.canada.ca/en/environment-climate-change/services/environmental-indicators/greenhouse-gas-emissions.html

**Sample Exercises**

1. Get the shape of `regional\_emissions` dataset. Is it bigger or smaller than `sector\_emissions`?
2. Get the column names and data types of the `regional\_emissions` dataset.
3. Calculate the sum of each column, then row. What do these values mean?
4. Two exercises with DataFrame indices:
   1. Create a DataFrame with “Region” as its index. Set an index on another dataset, then reset.
   2. Take the `.max()`, `.idxmax()`, `.min()`, and `.idxmin()` of `regional\_emissions\_index`. Do you notice any interesting patterns?
5. Create a DataFrame with only values from 2000 onwards. Sum these columns one at a time, then add columns to get total emissions from 2000 onwards.
6. Filtering by singular and compound conditions.
7. Calculating percentage change in regional emissions from 1990 to 2005.
8. Calculating general summary statistics over full DataFrame.
9. Identify all industries which have seen a decrease in emissions since 2000, with any technique.