Data Science for AAE

AAE 718 – Summer 2024

Worksheet 3

Pandas

1 Reading

1.1 Python Data Science Handbook

The following sections are all from Chapter 3. Between this worksheet and Worksheet 4, you'll be reading basically the entire chapter.

This cheatsheet could also be helpful

- Installing and using Panads
- Pandas Objects
- Data Indexing and Selection
- Operating on data
- Handling Missing Data
- Hierarchical Indexing
- Aggregation and Grouping

2 Daily Goals

- Loading a CSV
- Indicies and columns
- Extracting columns
- Extracting rows (filter, loc, iloc, query)
- Adding new columns
- Grouping and aggregation
- Plotting from a dataframe
- Understand operating on a copy vs the dataframe

3 The Climate dataset

Load the file "Methane_final.csv" from the Climate directory. This dataset has information about the methane gas emissions from various sources. Here is a link to the dataset. The dataset should display itself so you can view it.

There are a few things to notice in this,

1. There is a column called "Unnamed: 0"
2. There is one quantitative type, emissions
3. Some segments are labeled "Total", are these aggregates?
4. "World" is a region, is that an aggregate?

When first loading a dataset I want to identify important columns and drop those that aren't important. It's also a good rule of thumb to identify if some data is an aggregate (summation or average) of other data. You usually are better off dropping this data and remaking it, but you should try to verify it's correct.

Problem 1 (4 pt) Write a function called `methane` that takes one argument, a file path, and returns a dataframe of the "Methane_final.csv" data set.

The dataframe you return should not have the unnamed column.[1]

Here is a pro-tip. You'll be putting the previous function in a `.py` file. You can import this file in a Jupyter notebook with the code `from file_name.py import *`. This will make it so you can use all the functions defined in `file_name.py` in your Jupyter notebook.

Problem 2 (5 pt) Write a function called `methane_aggregation` with one input, a file path, and returns a number.

Load the Methane dataset (use the function you just created). Extract the rows where `region` is not "World" and `type` is "Agriculture". Find the sum of the emissions in this subset.

Also find the World total Agriculture emissions. Return the difference of these two numbers.

In the PDF submission, state the number you found and discuss it (it won't be zero). I'm not looking for a paper, just a paragraph. Discuss the implications of the non-zero number and theorize why this may have occurred.

Problem 3 (3 pt) In the previous problem you should have found a non-zero number. Perhaps the subset you created had extra segments we didn't consider.

Write a function called `problem_03` that returns an array of the unique entries in the `segment` column of your subset.

Problem 4 (5 pt) Let's explore average emissions by region. Write a function called `region_mean` that returns a dataframe with region averages.

The goal of this problem is to use the `groupby` command. In the PDF for this problem create a table with the numbers you found.

Problem 5 (5 pt) The numbers in the previous problem could be wrong due to the aggregated rows. Write a function called `region_total_mean` that first filters non-total `segments` and returns a dataframe with region averages.

Again, create a table in the PDF with the numbers. Discuss the differences between the tables for the two problems.

Problem 6 (5 pt) In this problem we're going to make some boxplots. Call this function `methane_graphs`. Create the following boxplots, include each in your PDF (be sure to label them). Also briefly (1-2 sentences) discuss each graph, what do you notice?

1. Aggregated by region
2. Aggregated by region, excluding World
3. Aggregated by region, excluding World and only including Total
4. Aggregated by segment, excluding World and Total.
5. Make one you find interesting.

As some suggestions, I recommend you make your bar graphs horizontal and adjust the figsize so they are larger/more detailed.

[1] It's incredibly common to write a function to load and clean your data. It can greatly simplify your analysis code.

4 Animal Crossing

Link to dataset description

Problem 7 (5 pt) Write a function called `animal_crossing` that takes a file path and returns the animal crossing accessories dataset.

In the PDF discuss what you see in the columns.

Problem 8 (3 pt) What item has the largest sell price?

Write a function, `sell_price` to answer this question. Return the item name. Include the item name in your PDF.

Problem 9 (5 pt) What item has the smallest difference between buy and sell price?[2]

Write a function, `smallest_diff` to answer this question, return the product `Name`. Include the item name in your PDF.

You should do this without any warnings from Python.

[2] you may wonder why this problem is 5 points. It is deceptively difficult