Python for Data Analysis and Visualization

Instructor: Claudia Carroll
Spring 2024

Session 3 (April 1)

Transdisciplinary Institute *in* Applied Data Sciences (TRIADS)

Arts & Sciences at Washington University in St. Louis
Signature Initiative

Today's Lesson Plan

1. Review class 2 exercises

2. Brief introduction to Python dataframes and Pandas

3. Demo: Pandas Basics

4. Exercises: Working with Pandas

Class 2 Exercise 2: Lists in Dictionaries

1. Add a dictionary ‘Addresses’ to the personDict dictionary that in turn contains two dictionaries, ‘Home’ and ‘Work’ that use key:value pairs to track addresses under the following values: ‘Street’, ‘City’, ‘Postcode’. (You may make up the addresses!)

2. Print out the postcode for the work address.

3. Print out the names of the children **on separate lines** (i.e. not as a list)

Class 2 Exercise 2 Solution:

```
personDict['Addresses'] = {'Home' : {‘Street' : '23
acacia ave.', ‘City' : 'Romford', 'PostCode' : 'RO6
5WR'}, 'Work' : {‘Street' : '19 Orford Road.', ‘City' :
'London', 'PostCode' : 'EC4J 3XY'} }

print(personDict['Addresses']['Work']['PostCode’])

for child in personDict['Children’]:
    print(child)
```

Class 2 Exercise 3: Extracting from JSON files

Write the code to extract each crop (‘D_curr_crop’) grown by each farm (‘id’) in the JSON file and the plot number in which they grow it. The output should be in the following string format”

```
Farm no. 1 grows maize in plot 1.
```

Hint 1: You will need to create a counter to track plot number.
Hint 2: There will be several nested loops and conditional statements!

Exercise 3 (Extracting from JSON files): Solution

```
unique_outputs = set()

for farms in d:
    plot_no = 0
    id = farms['A03_quest_no']
    if 'D_plots' in farms:
        plot = farms['D_plots']
        for crops in plot:
            plot_no += 1
            if 'D_crops' in crops:
                crop = crops['D_crops']
                for curr_crops in crop:
                    if 'D_curr_crop' in curr_crops:
                        combination = (id, curr_crops['D_curr_crop'], plot_no)
                        if combination not in unique_outputs:
                            print("Farm no ", id," grows ", curr_crops['D_curr_crop']," in plot", plot_no , ".")
                            unique_outputs.add(combination)
```

Introduction to Pandas

What is Pandas?

- The most common Python *library* for basic data manipulation
- Built on the NumPy library
- Often used in conjunction with additional libraries for advanced data analysis and visualization, such as matplotlib or SciPy
- Particularly suited to working with *tabular* data (csv, tsv, excel etc.)
- Reads tabular data as a *dataframe*
- A dataframe is a 2-dimensional array in python

Uses of Pandas?

- Reading data stored in CSV files (other file formats can be read as well)
- Slicing and subsetting data
- Dealing with missing data
- Inserting and deleting columns from data structures
- Aggregating data
- Joining of datasets (after they have been loaded into Dataframes)

Pandas Classes

Monday:

- Reading tabular data
- Slicing
- Basic data calculations (means etc.)

Wednesday:

- Dealing with null values
- Aggregating data in new dataframes
- Joining or merging dataframes

Setup

- Go to my GitHub repository for this class:

  https://github.com/ClaudiaECarroll/python_data_class

- From the Class 3 folder, download the files 'gdp_africa.csv' and 'gdp_europe.csv'.

- Put these files in your Desktop folder for this workshop.

Demo 1: Pandas Basics

Exercise 1:

1. Write the code to print each country in the data file gdp_africa.csv and that country's mean gdp between 1952 and 1982.

2. Now write the code to print each year (as represented by the column headings) in gdp_africa.csv and the mean GDP in Africa that year. How does your code differ from part one?

Exercise 1 Part 1: Solution

Write the code to print each country in the data file gdp_africa.csv and that country’s mean gdp between 1952 and 1982.

```
for x in countries:
        y = df.loc[x, "gdpPercap_1952": "gdpPercap_1982"].mean()
        print(x, y)
```

Exercise 1 Part 2: Solution

Now write the code to print each year (as represented by the column headings) in gdp_africa.csv and the mean GDP in Africa that year. How does your code differ from part one?

```
for x in years:
        y = df[x].mean()
        print(x, y)
```

Homework Exercises

1. Write the code the get the minimum value of a row “year” in a dataframe df_years
2. If you are about to write some code using the pandas library, what is the first line of code you have to enter in your program?
3. Below is a table containing the population of major cities in millions by year. Write the code to print which cities in the following table had a population greater than 20 million in the year 2010.

	London	**Paris**	**New York**	**Tokyo**
2020	9.0	10.9	18.9	37.5
2010	8.9	10.4	19.2	37.4
2000	8.9	10.5	19.1	37.1
1990	8.8	10.4	19.0	36.5