Data Analysis with Python (1): Pandas

Practice questions

Question 1: Columns as variables

Try saving a column as a variable. Print the new variable you've saved.

Question 2: Unique years

We can sometimes consider the year to be a categorical variable. How many unique years are there in the data set?

If you're forgetting what the column names are, remember to use `.columns`!

Question 3: Summarizing data

Select the columns `age5_surviving`, `gdp_per_day`, and `gdp_per_capita`, and print out summary statistics for these columns.

Question 4: Subsetting

Create a subset of data set including only samples from the region `'Asia'`. What is the mean of the `'life_expectancy'` column of this subset?

Question 5: Multiple subsets

Take a subset of df_hungary where the `life_expectancy` column is below 70.

What is the mean of `population` for this subset? Is it different than the mean population of `df_hungary`?

How many data points are remaining? (hint: use `.shape`)

Question 6: Shape after merge

Using `.shape` to compare the size of `merged_left` and `merged_right`. Are the numbers of rows the same? What about the number of columns? If there is a difference, why is there a difference?

Question 7: Comparing merges

Compare the two types of joins. Which type of join results in more rows with missing data? Can you think of situations where one type of join might be more useful than the other?

Question 8: Putting it together

Some data was collected by 5 different researchers on deer population sizes. Below, this data was recorded with accompanying temperature data and dates in three lists. In each list, one researcher is associated with the same index.

0. Haley McCann
1. Siena Welch
2. Jaylin Mercado
3. Ismael Hayden
4. Nina Bright

Transfer these data into a Pandas dataframe. Display the data frame, and export it as a .csv file.

As a reminder, each list is in the same order as the researchers name -> all of Haley McCann's data is at index `0`.

```
researchers = ['Haley McCann', 'Siena Welch',
    'Jaylin Mercado', 'Ismael Hayden', 'Nina Bright']

temperatures = [29.75, 12.63, 31.58, 7.16, 32.51]

populations = [442, 336, 505, 913, 933]

dates = ['5/25/2022','3/18/2022','6/28/2022','11/11/2022','7/6/2023']
```

Question 10: Your own data

Upload your own .csv file onto Google Colab. Make sure you have a single row for column names. Try manipulating it and summarizing your data.