Dr. Semmelweis and the Discovery of Handwashing

To complete this project you need to know some Python and be familiar with `pandas` DataFrames and bootstrap analysis. Here are relevant DataCamp exercises if you need to brush up your skills:

- From **Data Manipulation with pandas**
    - **Reading in a CSV**
    - **Subsetting rows**
    - **Inspecting a DataFrame**
- From **Statistical Thinking in Python (Part 2)**
    - **Bootstrap analysis**

Even if you've taken these courses you will still find this project challenging unless you use some external *documentation*. Here is a **pandas cheat sheet** summarizing the basics of pandas DataFrames. (You could also look at the **official pandas documentation** but be aware that it is *very technical*.)

Finally, know that *Google is your friend* and a good search pattern is **example of ??? in pandas** where **???** is whatever you need to do. For example, if you need to read in a csv file you could search for ***example of reading a csv file in pandas***.

The Solution of the below mentioned Tasks of this project have been uploaded in notebook.

Tasks for this project:

Task 1: Instructions

Load in the dataset with the yearly number of deaths.

- Import the `pandas`, aliasing it as `pd`.
- Read in `datasets/yearly_deaths_by_clinic.csv` and assign it to the variable `yearly`.
- Print out `yearly`.

Task 2: Instructions

Calculate the yearly proportion of deaths.

- Calculate the proportion of `deaths` per number of `births` and store the result in a new column named `proportion_deaths`.
- Extract the rows from Clinic 1 into `clinic_1` and the rows from Clinic 2 into `clinic_2`.
- Print out `clinic_1`.

Task 3: Instructions

Plot the yearly proportion of deaths for both clinics.

- Import `matplotlib.pyplot` as `plt`.
- Plot `proportion_deaths` by `year` for the two clinics in a single plot. Use the DataFrame `.plot()` method.
  - Label the plotted lines using the `label` argument to `.plot()`.
  - Change the y-axis label to `"Proportion deaths"` using the `ylabel` parameter in your second call of `.plot()`.
- Save the Axes object returned by the `plot` method into the variable `ax`.

Task 4: Instructions

Load in the dataset with the monthly number of deaths for Clinic 1.

- Read in `datasets/monthly_deaths.csv` and assign it to the variable `monthly`. Make sure to tell `read_csv` to parse the `date` column as a date.
- Calculate the proportion of `deaths` per number of `births` and store the result in the new column `monthly["proportion_deaths"]`.
- Print out the first rows in `monthly` using the `.head()` method.

Task 5: Instructions

Plot the monthly proportion of deaths for Clinic 1.

- Plot `proportion_deaths` by `date` for the `monthly` date using the DataFrame `.plot()` method.
  - Change the y-axis label to `"Proportion deaths"`
- Save the Axes object returned by the `.plot()` method into the variable `ax`.

Task 6: Instructions

Make a plot that highlights the effect of handwashing. *The code to define `handwashing_start` is already provided to you using `pandas`' `to_datetime()` function.*

- Split `monthly` into `before_washing` (the rows in `monthly` before `handwashing_start`) and `after_washing` (the rows in `monthly` at and after `handwashing_start`).
- Using the same approach you used in Task 3, plot `proportion_deaths` in `before_washing` and `after_washing` into the same plot. Again, use the DataFrame `.plot()` method twice, saving the Axes object returned by the first call of `.plot()` into the variable `ax`.
    - Label the plotted lines using the `label` argument to `.plot()`.
    - Change the y-axis label to `"Proportion deaths"` in your second call of `.plot()`.

Task 7: Instructions

Calculate the average reduction in proportion of deaths due to handwashing.

- Select the column `proportion_deaths` in `before_washing` and assign it to `before_proportion`.
- Do the same for `proportion_deaths` in `after_washing` and assign it to `after_proportion`.
- Calculate the difference in mean monthly proportion of deaths as mean `after_proportion` minus mean `before_proportion`.

Task 8: Instructions

Make a bootstrap analysis of the difference in mean monthly proportion of deaths.

- Within your `for` loop:
    - `boot_before` and `boot_after` should be sampled with replacement from `before_proportion` and `after_proportion`.
    - The difference in means should be appended to `boot_mean_diff`.
- Calculate a 95% `confidence_interval` as the 2.5% and 97.5% quantiles of `boot_mean_diff`.

Task 9: Instructions

- Given the data Semmelweis collected, is it `True` or `False` that doctors should wash their hands?