Assignment Activity 2: Importing and exploring data

This assignment activity will help you to successfully complete the final assignment. The provided instructions will guide you to the minimum expected activities to complete during your exploratory data analysis and the presentation of your insights. You are encouraged to think critically and apply logic to identify patterns and trends that the organisation can use to answer specific questions. You will also use these insights to inform both your technical report and business presentation.

**Scenario**

Recall the work you’re doing for the NHS for your final assignment. Review the requirements introduced in [**Assignment: Diagnostic Analysis using Python**](https://fourthrev.instructure.com/courses/895/assignments/2971). To prepare for the data analysis, you’ll need to create a new Jupyter Notebook, import the data files, explore the data, and assess the quality.

**Objectives**

Familiarise yourself with the individual data sets. The following questions can serve as a starting point, but you should also answer the questions you documented in the first module to ensure that you can answer basic descriptive questions for each of the data sets.:

* How many locations are there in the data set?
* What are the five locations with the highest number of appointments?
* How many service settings, context types, national categories, and appointment statuses are there?
* Are there any comments regarding data quality or descriptive statistics worth noting for each of the data sets?
* Can you comment on other features (columns) that could potentially be useful in your analysis?

**Approach**

1. **Ensure** you’re prepared for your project by following the steps outlined in [**Assignment Activity 1: Exploring the data and planning your analytic approach**](https://fourthrev.instructure.com/courses/895/pages/assignment-activity-1-exploring-the-data-and-planning-your-analytic-approach). Before commencing, you should complete the following tasks:
   * Ensure you’ve downloaded and extracted all the data sets and metadata from the LSE\_DA201\_Assignment\_files.zip file into the same directory/folder you’ll use for your assignment. To ensure the file paths are maintained, keep the data and Jupyter Notebook files in the same directory/folder.
   * Remember that you will have to execute all the code in your Notebook (from the start of the Notebook to where you are currently working) every time you restart your Jupyter server. Make sure that your Notebook is in the correct state before continuing with the work for the current module.
2. **Prepare** your workstation:
   1. Launch Jupyter Notebook, and open the provided assignment template.
   2. Import the necessary libraries (e.g. Pandas and Numpy).
   3. Load and sense-check the three data sets:
      * Use the appropriate Pandas read\_() function.
      * Create three new DataFrames to store the data sets. For example, store actual\_duration.csv as ad, appointments\_regional.csv as ar, and national\_categories.xlsx as nc.
      * Sense-check the new DataFrames to determine the column names, number of rows and columns, data types, and number of missing values. (**Hint**: You can use the print() function (e.g. print(ad.columns)).
      * Determine the descriptive statistics (e.g. df.describe()) and metadata (e.g. df.info()) of each DataFrame.
3. **Explore** the data to respond to the questions posed in the objectives:
   1. Use the value.counts(), count(), and print() methods. (**Hint:** The count() method determines the number of times values meet a certain condition.)
   2. Inset an appropriate docstring to present your output in a sensible way. For example, print("Count of locations: ", …]).
   3. Use the nc DataFrame and appropriate column names (e.g. sub\_icb\_location\_name).
4. **Summarise and highlight** worthwhile trends and insights for the various data sources. You should start to think about the different stakeholders in the assignment scenario and what information would be relevant to them at this stage. The comments and observations noted at this stage will serve as input to your technical report and business presentation.
5. **Back up** your work to a safe location. This would allow you to revert to a previous state in the case of making a mistake in the code, or deleting a section by mistake. (A simple way of doing this is to save or mail a compressed version at frequent intervals).

**Information**

The raw data uses the term **count** to, for example, calculate the number of appointments. The word count refers to the number of values meeting a certain condition, while**sum** indicates the arithmetic sum of values in a column. For example, if there are five cells with different values (e.g. 20, 10, 5, 50, and 25), the count of the cells will be five, while the sum of the values will be 100.