CST8333 21S Dataset Attribution

Attribution and License

The dataset for use in CST8333 22W comes from the Open Government of Canada, published by Public Health Agency of Canada.

You can obtain the dataset here:
Canada Energy Regulator **Pipeline Incident Data**. [webpage]
Retrieved on JAN 8, 2022 from https://open.canada.ca/data/en/dataset/7dffedc4-23fa-440c-a36d-adf5a6cc09f1

- The specific resource downloaded from the web page is the Dataset as CSV in English.

You need to review the Open Government License, which is found here:
http://open.canada.ca/en/open-government-licence-canada

It is your responsibility to attribute the data set source as per the Open Government License within your program source code and project documentation.

Additional Information

The dataset file is named “pineline-incidents-comprehensive-data.csv”

While the file will open by default in MS Excel, it is actually only a text file, use a text editor like Notepad++ or Visual Studio Code to review it.

Notes

- The demonstration program(s) provided use a data set from a previous run of this course i.e. Canadian Cheeses, and are provided to show case a non-layered program similar to assignment 1, which is then refactored most of the way into a layered program (assignment 2). **You are required to use the data set assigned above for your programming work in this course, do not use Cheeses or any other dataset.**

List of Columns to use:
You are required to the following columns for the purposes of Create, Read, Update, Delete functionality in your program:

A, B,C,D,E,F,K,M,R

Each date entry also seems to report the incidents on the specific date only, so to see total number of incidents over a period of time you may need to aggregate the data (perform a running sum).

It may be interesting to use the data here to create graphs or charts of the aggregate data over time on a province by province basis.

What is a Dataset?

Within the context of this course, a data set is a collection of information stored as records with a known sequence of columns.

A CSV file is a plain text file that has records stored as lines of text, corresponding to the records of a data set.

Each line in the file represents one record, each record's fields are separated by comma characters.

While Microsoft Excel is associated with *.csv file extension by the operating system, and will be the program opened if you double-click on the *.csv file, it is important to open the *.csv file in a text editor to see the actual structure.

The dataset provided has the column titles (field names) provided as the first line of text in the file. The second line of text might be data, or could be the column names in French. The remaining lines of text in the file are the records.

Typically a program will read in the lines of text, parse the text using the comma as a delimiter character to split the fields before storing the records into some sort of data structure and then manipulating the data further.

It's recommended to use a CSV API library instead of writing your own customized code, as text enclosed in double quotes may have commas intended to be included as part of the data can be mistaken for delimiters, and there can be other issues as well that a CSV API will know how to handle. Writing your own code can work, but may need to be extensively tested as well as maintained frequently.

Lastly, and importantly, your program code needs to be based on the dataset. You are obligated to use the data set column names in your code to indicate to a reviewer that your program was custom made for this dataset. Highly abstract programs that can process any data set, and have none of the column names from this dataset within the code, are subject to a loss of marks.