Assignment #1 - Evaluate Data Set

February 18, 2022

1 Assignment #1: Evaluate Data Set

In your first graded assignment for this course, you will find and evaluate a data set on a topic of your choosing. You will be given specific questions related to the data set, and you will also be tasked with importing the data from your data set into your Jupyter Notebook environment.

1.1 Choose a Broad Topic

First, choose a topic of interest to you, such as a specific social issue, a political or cultural trend, a community, or a hobby or personal interest. Ideally, it should be something relatively broad that society or culture is having some sort of conversation about, or which many people think about on a regular basis. Possible examples might be:

- Housing and rental prices in New York City
- A popular genre of music
- A sport
- Food prices
- Crime
- Government spending
- Popularity of movies or TV shows
- Bestselling books

1.2 Find a Data Set

Search for terms related to your topic on Google, both generally and by adding specific names of websites where data will commonly be found. The following websites are frequently places to find data sets, and it sometimes makes sense to both search directly on the site and to do regular Google searches with the site name:

- Google Data Set Search
- GitHub
- Kaggle
- Data.gov

For example, if your topic is bestselling books, you might try these searches on Google:

- bestselling books data set
- bestselling books data set kaggle
- bestselling books data set github
- bestselling books data set

You're not likely to find bestselling books data on data.gov, but you might find a lot of info related to health, the economy, and demographics, since that's what the government is most concerned about.

What is your topic? (Enter responses to questions like these in the cell provided below.)

What drew you to this topic, or why did you choose it?

1.3 Choose a Data Set

Choose a data set to use for the rest of this assignment. The data set doesn't need to be perfect, but ideally it should interest you and be in one of the discussed formats (CSV, TSV, Excel, JSON, .txt). If the data is in another format, either reach out to me to ask about it or choose another data set for this assignment. You may want to look ahead to the rest of this assignment to make sure nothing about the data set will make it difficult to answer the questions or to import the data.

What is the name of your data set? Or provide a one-sentence description.

What is the URL of the page on which you found the data set? (Paste the full URL.)

1.4 Evaluation

In this step, you will evaluate the data to the extent possible without using data science tools like Python or Pandas. You should download the data to your computer, and use both the data set itself as downloaded and the page you found it on to answer these questions. If you can't answer a question, write what you tried to do to answer it—don't give up right away, and try to think of other ways to answer the question. (In some cases, you can even contact the person who created the data set—if you do so, feel free to copy me on the email.)

What is the file format or file extension of the data set? Examples might be .csv (Comma Seperated Values), .tsv (Tab Separated Values), .xlsx (Excel workbook), .txt (plain text file), or JSON (JavaScript Object Notation).

What is the size in megabytes (MB)or gigabytes (GB) of the data set?

How many columns or fields is the data set? (Columns or fields are different types of data—for example, a book dataset might have title, author, and year as columns.)

List out the columns in the data set. (You can put each column on a line, or you can separate the columns with commas.)

How many rows is the data set? (To use the book data set example again, each row might represent one book.)

What types of data appear in the data set? (You can use Python terms, like "integer," "float," "boolean," "string," or you can use other descriptive terms, like "numeric data" or "text data." Try to be as comprehensive as possible in your answer.

On initial inspection, does anything appear to be missing or wrong in the data set? (Don't spend too long on this.)

What kinds of questions could you answer with this data set? In answering this question, write at least one paragraph of at least 150 words.

Do you see any issues or limitations with the data set? Alternatively, what do you wish was included in the data set that is not included? (Write at least one paragraph of at least 150 words.)

1.5 Reading the Data in Python

Using as many cells as you need in the rest of this notebook, load the data into Python. You will probably want to use Pandas to load the data. Some example code is provided for you below.

To use the example code in a Jupyter Notebook on your computer, you will need to make sure your data set is in the same folder as your notebook, and that you get the filename *exactly* right, including the extension. Here is example code for a Jupyter Notebook—this is just to get you started, and you are responsible for getting this working, which may involve looking up how to import data using Python and Pandas on Google.

```
import pandas

df = pandas.read_csv('name_of_data_file.csv')

df
```

If your data is in another format, you will need to use the Pandas function related to that format. For example, to import JSON:

```
import pandas

df = pandas.read_json('name_of_data_file.json')

df
```

If you're on Google Colab, you will need to upload your dataset to Colab. To do so, you can use some special code—when you run it, a "choose file" dialog will open below the cell. Click the "choose file" button and select your file from your hard drive.

```
import pandas
from google.colab import files

upload = files.upload()
```

Once you have uploaded the file, run this code to import it into a data frame. You will need to change the 'name_of_uploaded_file.csv' string to the exact name of the file you uploaded.

```
import io

filename = 'name_of_uploaded_file.csv'

df = pandas.read_csv(io.BytesIO(upload[filename]))

df
```

Use as much space as you need below to import the file into Pandas (as above). Make sure the dataframe (df) is shown as an output at the end.

[ ]:

[ ]:

[ ]: