

Assignment #2 - Viewing and Manipulating Data

March 8, 2022

[]:

1 Assignment #2: Data Exploration

In this assignment, you will demonstrate your knowledge of the Python and Pandas skills we've learned so far. These include:

- Getting an overview of your data
- Extracting columns
- Getting fundamental statistical information of a series, such as the mean and median
- Getting a random sample of your data
- Creating a subset of your data using indexing with boolean lists
- Sorting your data
- Indexing by matching a string

You will not need to formulate or answer a research question for this assignment. You simply need to demonstrate your ability to perform specific operations on the data set you have selected. In the next assignment, your midterm, you will create a narrative exploration of the data using many of the same methods, but in such a way as to use narrative along with data analysis to answer a research question.

You may use the data set you imported in assignment #1, or you may choose another data set. That means that you can reuse code you wrote in Assignment #1 to import the data.

In the below notebook, I will ask for you to perform a task. Each task will be followed by one or more blank code cells. Use as many cells as you need to complete the task. If you know some other way to complete the task with Python and/or Pandas that is not the way we learned in class, that is also acceptable. Feel free to use Google searches to refresh your memory as to how to perform each task. You will also likely wish to consult the class notebooks, which have been emailed to you after class and which can be found in [our class repository on GitHub](#).

If your data set does not have a needed form of data for a task, write code that would perform the task if a column with that data existed, and provide an explanation as a comment or markdown cell.

You may receive partial credit for some incomplete or incorrect answers. Please feel free to add comments about your thought process, which will make it more likely that you will receive partial credit.

The last cell is a bonus. You will not be penalized for not attempting or completing it.

Import Pandas and read your data in as a data frame. Assign the data frame to a variable, such as `df`.

```
[ ]:
```

```
[ ]:
```

Make the data frame visible by placing the data frame variable in a cell by itself.

```
[ ]:
```

What columns are in your data frame? Use a function or method that shows all the columns.

```
[ ]:
```

Write Python code to output the number of rows in your data frame.

```
[ ]:
```

Write code to output a column from your data frame as a series. (Extract a column.)

```
[ ]:
```

Pick a column in your data set with numeric data (integers or floats). Output the mean (average) and median of that column.

```
[ ]:
```

```
[ ]:
```

Use a conditional with a column to create a boolean series. Save the boolean series to a variable.

```
[ ]:
```

Use your boolean series and index your data frame to create a subset of the data. Save the resulting data frame to a variable.

```
[ ]:
```

Output the length of the resulting data frame (the subset of your data for which the conditional was true).

```
[ ]:
```

Sort your data frame by a particular column and output the result.

```
[ ]:
```

Output the first ten rows of your data set, the last ten rows of your data set, and a random ten rows of your data set (a sample of your data).

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```

1.1 Bonus

In the remaining cells of this notebook, use your original data frame and create a subset of the data that matches a specific string in a column. That is, extract all rows of the original data frame that contain a specific string in one of the columns.

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```