**Information Systems – Spring 2019**

**HW2: Jupyter + Pandas (25 points)**

**Due: Jan 24, 2019**

The data for this assignment is taken from data.gov which includes (1) baby names in the US over the past 100 years and (2) accidental drug deaths in CT 2012-June 2017. The datasets are posted on D2L.

You will work with this data to answer questions about each of these datasets using the Jupyter notebook.

Submission:

* Submit your **Jupyter notebook to the D2L dropbox by Jan 24 at 1600**. The notebook must have all the cells executed (with output shown).
* Turn in a **hardcopy of this worksheet with your answers written in**, in class on Jan 24.

Collaboration:

You may collaborate on this assignment with students currently enrolled in this course. You are encouraged to post and answer questions on Piazza related to this assignment. You may use code presented in class or posted on the class website as examples. You may use the Internet as a resource for code documentation and ideas. Direct copying from the Internet (or anyone) is never allowed, and you must be able to cite and explain any code you have incorporated within your own program. If in doubt, ask your instructor.

Grading:

In addition to correctness of your answers, you will be graded on your ability to communicate your work in your Jupyter notebook. Your code should be well commented and include markdown cells to indicate what questions you are answering and the conclusion you are drawing from your Python code. You should consider your notebook a report and that should be readable by anyone, even if they don’t understand Python.

**Part I – Baby Names**  
Use the babynames.csv file (posted on D2L) to answer the following questions.

1. What years does this dataset cover? (Use python to answer this question)
2. How many unique *boy* names are there in this dataset?
3. How many babies were named Taylor in this dataset?
4. How many more *boys* were named Logan than *girls* in 1998?
5. What were the 3 top fe*male names* in 1998? Give counts for each name.
6. How many unisex names are in this dataset? Make a list of the unisex names.

Hint: <https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.intersect1d.html>

To answer this question, tell me how many unisex names there are and the first three and last three that appear in the list.

1. Tell me something interesting about *your* name based on this dataset. (It must be something more than just reading a single row off the dataset). If your name does not appear in the dataset, then use something close to your name.

**Part II – Drug Deaths**

Use the drugdeaths.csv file (posted on D2L) to answer the following questions.

Start by preparing the dataset:

Import the data – be sure to:

- Make the dates of data type pandas.tslib.Timestamp

- Make date column the index column

- Print the head of the resulting data frame in your Jupyter notebook

Make a month column – be sure to:

- Create a "Month" column whose values are Jan, Feb,...,Nov, Dec

- Print the head of the data frame in your Jupyter notebook

1. How many accidental drug deaths are there for each month in this dataset?

|  |  |  |  |
| --- | --- | --- | --- |
| **Month** | **# of deaths** | **Month** | **# of deaths** |
| January |  | July |  |
| February |  | August |  |
| March |  | September |  |
| April |  | October |  |
| May |  | November |  |
| June |  | December |  |

1. What is the city with the most accidental deaths in this dataset? How many deaths did this city have?
2. Find the average age of accidental deaths for all cities in the dataset.

Hint: use the aggregate method.

* 1. What city has the oldest average age of accidental deaths? What is the average age of accidental deaths for this city?
  2. What city has the youngest average age of accidental deaths? What is the average age of accidental deaths for this city?

1. Make a bar chart to display the frequency of the ages that occur in this dataset. Comment on the chart in 1-2 sentences.