Assignment 5 (Group Project, part 1) - Transaction Database

Objectives: Gain experience building a relational database using transactional data. Select and identify a publicly available dataset, model it, and load it into a relational database. Explore the data. Record the whole process in a notebook so it may be reproducible and easily communicated to others. Gain experience working collaboratively to achieve these goals.

Grading criteria: The tasks should all be completed. As everyone will work with data of their choosing, not everyone will have the same results. The narrative surrounding code should inform the reader about the process, from sources of data through any decisions made about techniques and tools. The notebook itself should be reproducible; from start to finish, another person should be able to use the same code to obtain the same data and same results that you obtain, or at least similar data and similar results, should the data sources chosen vary over time.

Give your notebook a clear name like "`assignment-05-our-last-names`". Document the contributions of each group member. Acknowledge any assistance you received. Zip your .ipynb along with any scripts or other files together. Upload your zipfile to blackboard by the deadline. Only one group member should submit on behalf of your group.

Deadline: Friday, October 28, 8am

Part 1 - Select, procure, and describe data

Find and select a publicly available source of transactional data. The records must be of individual transactions, not summaries. For example, the Connecticut boating record data we used in Assignment 4 was summary data; as such it is an example of data not to use. Instead, look for something like the Capital Bikeshare trip data, wherein each individual row identifies a specific transaction. There are at least a few sets of transactional data on data.gov (e.g. http://catalog.data.gov/dataset/building-violations-f0f5e); Shmuel Ben-Gad at Gelman Library provided a good list of sources (e.g. https://wrds-web.wharton.upenn.edu/wrds/index.cfm) and may be available to help; you may find other ideas through the links below.

Potential sources for data (via Prof. Kanungo)

- https://www.opensciencedatacloud.org/publicdata/
- http://www.kdnuggets.com/datasets/index.html
- http://datascience.berkeley.edu/open-data-sets/
- http://www.datasciencecentral.com/profiles/blogs/big-data-sets-available-for-free
- http://www.datasciencecentral.com/profiles/blogs/great-github-list-of-public-data-sets
- http://www.datascienceweekly.org/data-science-resources/data-science-datasets
- http://www.statsci.org/datasets.html
- http://blog.bigml.com/2013/02/28/data-data-data-thousands-of-public-data-sources/

Separately, if you might be interested in working with social media data, you can find samples from Twitter here:

- https://archive.org/details/twitterstream
- See Dan C. for more

Once you have chosen a dataset to work with, describe its purpose and origin, provide proper links or citations, get the data, and begin to describe its components. Tools like `wget` and `csvkit` may be ideal for this purpose, but you may use others, provided you show your work, preferably in a notebook.

Part 2 - Model and load database

Design and implement a relational database schema for this dataset. This schema should include at least one table. A tool like `csvsql` may be ideal for this purpose as well, with the same caveat - show your work, and show the schema you create.

Load the data into the database. If you pre-process the data to remove, rename, or re-arrange columns, or otherwise merge, split, or combine the data with any other source, document these steps in reproducible detail within your notebook, and by including any external scripts you develop.

You may choose to create indexes on your data; document these as well.

Part 3 - Explore data from database

With your data loaded, perform some exploratory queries using SQL. Verify that the data loaded successfully by comparing counts with the source data files you described in Part 1. Identify key variables of interest and note their ranges along with other useful descriptive statistics. For this part, you might find it helpful to produce one or more plots.

At the end of this step, you and each of your group members should feel confident about your dataset choice; you should have a documented, reproducible process that allows each group memeber to work with the data on their own and with a database; you should have a sense of the kinds of stories you might want to tell with your data.