CS 564 Spring 2017: Data Analysis with and Beyond SQL

Project Grade Weight: 6% of the total grade

Due: May 4, 2017 2:00 PM

Introduction

Now that you know about databases and querying, it time to put your skills to work for something bigger. The setup for this project is likely familiar to you: You have just got your hands on an interesting dataset, and you now want to find interesting insights hidden in that dataset.

In such endeavors, you will first need to *explore* the dataset by posing queries against the dataset. Then, you may want to visualize the query results. Later you may build machine learning models. You may iterate through the previous steps, and finally summarize your findings. In this project, you get to undertake one such effort, using a popular tool called Jupyter notebook (formerly called iPython notebook). This notebook is part of the Python ecosystem.

If you don't know Python, don't worry. It is easy to pick up, and we will use a very small surface of that language. There are a number of quick start guides for Python, including: https://www.stavros.io/tutorials/python/ and https://www.learnpython.org. If you have not worked with Python before, go ahead and read these guides now.

If you are not familiar with Jupyter notebook, don't worry, we will get you started with a skeleton notebook, and go over it in the discussion session. You can also find numerous tutorials on the web. Look for tutorials on "Jupyter notebook."

Note this project is somewhat open-ended (by design) and meant to give you a way to creatively explore a popular way to build a data science pipeline. We hope what you learn in this project becomes a skill that you can take with you well beyond this semester – so please, try to go beyond what is in this project and see if you can explore more advanced analysis tools (e.g. scikit-learn) as part of this project.

The dataset

For this assignment, we have prepared a dataset in the popular comma-separated-file (CSV) format. This data is at: http://pages.cs.wisc.edu/~jignesh/cs564/projects/notebook/AQI.csv.gz. Uncompress this file to produce a .csv file that you can load into a database via code written into a "*code cell*" in your notebook.

This dataset was derived from raw data published by the Environmental Protection Agency at https://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html. The CSV file has data about the Air Quality Index (AQI) for each week since 1997, and for each EPA monitoring station. See https://airnow.gov/index.cfm?action=aqibasics.aqi to understand the AQI scores.

The data file is quite large, so you may have trouble opening it in your favorite spreadsheet program. There are even larger datasets that EPA produces, including hourly data. That dataset would have been too large to use for this assignment as disk space on lab machines is limited. The dataset that we provide, however is large enough to do some interesting analysis.

Your Assignment

As mentioned above, by design, this assignment is open-ended. We provide you with a starting notebook. This notebook is at:
http://www.cs.wisc.edu/~jignesh/cs564/projects/notebook/Starting_Template.ipynb.
The PDF version of this notebook is at:
http://www.cs.wisc.edu/~jignesh/cs564/projects/notebook/Starting_Template.pdf

If you are doing this project on a CSL machine, then you should use the notebook at:
http://www.cs.wisc.edu/~jignesh/cs564/projects/notebook/Starting_template_v3.ipynb

Before you can run that notebook, you will need to run: `pip install --user Flask-SQLAlchemy` to install SQL Alchemy. Then, you can start the notebook by typing in:

```
ipython notebook Starting_template_v3.ipynb
```

You task is to turn in a completed notebook that must have at least the following five elements:

1. Well-defined commenting of the Python code that you write, and the use of "markdown" cells to explain each step in your notebook. In the very first markdown cell, you must note your name, and which version of the notebook package/software you used. There are three acceptable software packages for this assignment. These two are:
    a. Pre-installed version of IPython on CSL machines: You can start a notebook, on a CSL machines by run: `ipython notebook <name_of_ipythonbook>.ipynb`
    b. Anaconda v. 3.6: There is a free distribution of Python related tools called Anaconda, which is available for download from https://www.continuum.io/downloads . This package bundles in a number of popular Python tools/libraries, including Jupyter (and Python itself). If you choose to do your assignment using the Anaconda approach, please use the Python 3.6 version.
    c. Anaconda v. 2.7: Same as above but with the older version of Python.
2. You must load the data from the CSV file above into a SQLite3 database. See code in our starter template to get you started on this aspect.
3. Once you get setup, dive into analytics mode:
    a. Issue SQL queries against the database to understand what is in the database.
    b. Find **at least three** interesting insights. These insights need not be too complicated. Explain to the reader of the notebook what you find (hopefully you can see why the notebook is such a powerful tool for data analysis). Be creative.
    c. Visualize the output of at least one of the queries using a graph.
4. You must have a final "*markdown*" cell that contains a summary of the interesting insights that you find from your analysis (see the template that we provide for an example). You can use *<html>* formatting in markdown cells.

If all you care about is getting a high score on this assignment, you can probably complete this assignment in a few days using our starting template. But, we really hope to sign off this semester with the message that the best think a good college can teach you is: how to learn by yourself and to constantly keep learning. In that spirit, we hope you go beyond the bare minimum that you need to do for this assignment and produce a insightful report that digs out the hidden secrets that this dataset may have. Feel free to explore building some machine learning models using `scikit-learn` (http://scikit-learn.org/stable/) Simple models like decision trees (see https://tinyurl.com/mggvfcv) can tell you a lot – feel free to try it!

***Helpful Tip**: When you press the run button associated with a notebook cell, sometimes it may seem like nothing happens, or you think the code in the cell has completed running (so you move on to the next cell). When the code in the cell is running, the notebook will put a * sign on the top-left corner, just outside the cell. This sign indicates that the code is still running. Wait till that * changes to a cell number. Now the code in the cell has finished execution.*

You may run into this problem with the cell in which the data is loaded into the database.

Submission Instructions:

1) Place your notebook in a directory. Name this directory using the format: `<lastname>_<firstname>_P5 (e.g. Cook_Tim_P5).` The notebook should be self-contained (that is the point of the notebook). Also "print" your notebook after running all the code cells (so the pictures are printed) as a pdf file, and place the pdf file in this directory. Call the pdf file `AQI.pdf`
2) Run:
   `tar –czvf <lastname>_<firstname>_P5.tar.gz /path-to-project/lastname_firstname_P5`
3) Submit the tar file.
4) To check you can un compress the tar file.
   (run: `tar –xzvf <lastname>_<firstname>_P5.tar.gz`).

To hand in your work, please go to the Canvas: Assignment 5 (Data Analysis and Beyond SQL) to upload your files. Your files must be uploaded by the deadline stated on the first page.