Project 3 - due Friday, 12/9 at 12pm

Work in groups of four people per project.

Select a dataset of at least 250,000 records, preferably at the level of individual transactions. Acquire the data, survey it, wrangle it into a suitable format, and analyze it. Present your work in a brief (8-10 minute) talk in class on Tuesday, 12/6, and write up your process and results in a reproducible notebook.

You may use any of the methods we studied in class: Unix command line tools, relational databases and dimensional models with SQLite or PostgreSQL, or Spark. The datanotebook.org server will remain available until the end of the semester and may be used for the project.

Attestation: All project team members should contribute meaningfully to the final results. Please attest to your individual contributions and that each member contributed substantially to the project in your final writeup.

Deadline: Friday, 12/9, 12pm. Submit your presentation (PDF export), your notebook, and any additionally necessary files like scripts or images together in one zip file. Only one team member should submit on behalf of all members. All team members should post their projects to GitHub for reviews after the deadline has passed.

Part 1 - Selection (30 points)

Identify and describe your dataset, its source, and what appeals to you about it. Acquire the data and perform an initial exploration to determine which themes you wish to explore. Describe the questions you want to be able to answer with the data, any concerns you have about the data, and any challenges you expect to have to overcome.

Part 2 - Wrangling (35 points)

Based on what you found above, wrangle the data into a format suitable for analysis. This may involve cleaning, filtering, merging, and modeling steps, any and all of which are valid for this project. Describe your process as you proceed, and document any scripts, databases, or other models you develop. Be specific about any key decisions to modify or remove data, how you overcame any challenges, and all assumptions you make about the meaning of variables and their values.

Verify that your wrangling steps have succeeded (for example, if you loaded the data into a dimensional model, ensure that the fact table contains the right number of records).

Part 3 - Analysis (35 points)

Explore and analyze your data in its wrangled form. Follow through on the themes you identified in Part 1 with queries or scripts that answer the questions you had in mind. Be clear about the answers you discover, discussing them and whether the results match your expectations. Include charts or other visuals that support your analysis. You may use Tableau, ggplot, or other tools we have not covered in class for visualization, but be sure to export images and to include them properly in your writeup.

Bonus - Augment (10 points)

Sometimes the most value can be gained from one dataset when it is studied alongside data drawn from other sources. Identify at least one additional data source that can complement your analysis. Pull this additional data into your chosen environment and explore at least one more theme you are able to further analyze that depends upon a combination of data from both sources.