Processing UK Prescribing Data Using SparkSQL : CS784 Project Proposal

Derek Paulsen

February 24, 2019

1 Goals of the Project

The primary goal of this project to examine trends of drug prescribing in the UK by postal code and attempt to find interesting correlations between characteristics of postal codes. Current directions for where to look for correlations are,

- Latitude - Does living farther north or south increase the rates a which certain medications (or classes of medications) are prescribed?
- Income - Does the median income of a postal code relate to which drugs are being prescribed there? Furthermore can any disparities be explained by other data (e.g. disease prevalence rates)?
- Disease Prevalence - Are people is certain postal codes more likely to seek treatment (i.e. be prescribed medication for) for a given disease than others?
- Education - How does average level of education of an area affect what drugs are being prescribed in an area? Additionally, how are the effects tied into other possible causes (e.g. average income)?

2 Approach

The UK government provide many large, reliable, and representative datasets about health care in the UK [2]. Some of these including statistics about healthcare, environment, and education. These datasets (along with possibly others) will be linked via postal code to try to find interesting interactions and correlationals.

The current plan is to use SparkSQL [1] (with python) to do the data processing, on a cluster of a few machines on Cloud Lab. SparkSQL is ideal for this sort of exploratory data analysis for a few reasons. First, the data being processed will be structured as a relational table, which SparkSQL is intended to handle efficiently. Second, the data analysis is very much exploratory, hence being able to interactively query the data using python is very useful from a productivity stand point. Finally, the datasets are quite large, (over a GB per month), hence being able to run the scripts on a cluster is essential for completing the analysis in a reasonable time.

3 Measuring the Outcome

A serious shortcoming of the approach in this project is that there is that there is very limited ability to make causal inferences with the analysis. This being the case, the ideal outcome of this project would be to find an interesting correlation between two datasets and then be able to gather more evidence for a causal relationship. For example, this could be finding that the ratio of people with heart disease to the number of people being prescribed medications to treat heart disease differ based on the median income of the postal code. Then begin able to find other outside sources that have found similar results via different methods (e.g. a medical journal study).

References

[1] Michael Armbrust et al. "Spark SQL: Relational Data Processing in Spark". In: SIGMOD '15 (2015), pp. 1383–1394. DOI: 10.1145/2723372.2742797. URL: http://doi.acm.org/10.1145/2723372.2742797.

[2] *UK Open Data*. https://data.gov.uk/.