Processing UK Prescribing Data Using SparkSQL : CS784 Project Proposal

Derek Paulsen

March 30, 2019

1 Problem

There is a large amount of freely available data from various governments which provide detailed information about important aspects of life (e.g. healthcare, crime, etc.). The problem is then to process the data and integrate it with other sources to find meaningful interactions, which could inform policy and other decisions.

Attempting solve this problem poses a few particular challenges,

1.1 Data Size

Many of the datasets of interest are large (e.g. greater than a GB in size). While dealing with a couple GB of data is possible on a single machine, depending on the relationship between data sets, 2GB of data can easily turn into 20GB after being joined, making scale out the only viable option. This problem is typically solved by using a distributed execution framework (e.g. MapReduce, TensorFlow, Spark), for our purposes interative data querying is vital for effeciently exploring the data hence we opted to use SparkSQL.

1.2 Finding Linkable Datasets

While there is a massive amount of publicly available data from many sources, it is very difficult to find datasets that make sense to link. That is, finding two datasets which are from the same time period, concerned with the same population and may have some interesting interaction between them is far from trivial. This problem was address with a lot of googling and paitience.

1.3 Joining Datasets

Even when datasets make sense to join it is frequently the case the there isn't a simple way to do so. In particular it was found that many of the datasets which have geographic location, have different granularity. For example, census data is typically the most fine grained geographic data, but statistics like mortality from diseases are much coarser grain and less frequently updated. To address this problem, extensive data transformation and aggregation. In particular, the granularity of geographic data was modified to such that all the datasets use keyed by year and the outer post code.

1.4 Finding Interactions

After finding datasets and figuring out how to link them, there still is the problem of finding meaningful/interesting interactions and insight. The number of possible directions for where to look for interactions in datasets is an absolutely massive search space when there are 3 or 4 datasets. This problem is probably the most difficult to tackle since there really isn't a standard way of addressing this problem. For our case we found it effecitve to first interacively query and get summary statistics of the data that we have to get a sense for possible directions and trends in the data. After doing this we ran many queries with cheap to compute statistical measures, in particular we looked at the pearson's correlation between two varibles to as kind of a filter for places to do finer analysis. Additionally we also had to use a lot of simply guessing as to where to look based on common sense. For example, higher crime might lead to greater stress which might lead to more heart medication being prescribed on average.

2 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)?

3 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.

4 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/.