Project 1: Medicare Data (BigQuery Dataset)

In the United States, Medicare is a single-payer, national social insurance program administered by the U.S. federal government since 1966. Healthcare is an increasingly important issue for many Americans; the Centers for Medicare and Medicaid Services estimate over 41 million Americans were enrolled in Medicare prescription drug coverage programs as of October 2016. It provides health insurance for Americans aged 65 and older who have worked and paid into the system through the payroll tax. Unfortunately, with modern US Healthcare programs' complexity and sophistication, fraud losses in healthcare cost US taxpayers a staggering amount, to quote from the Justice Department,

> "Health care fraud costs the United States tens of billions of dollars each year. Some estimates put the figure close to $100 billion a year. It is a rising threat, with national health care expenditures estimated to exceed $3 trillion in 2014."
> - U.S. Department of Justice

This project aims to tackle this data using a data-driven approach, particularly we hope to:

- Detect patterns of fraud medicare providers.
- Build classification models to detect these providers.

Solutions to all problems start with gathering data and seeing the big picture through big data analytics lens, here I employed a combination of data including CMS Medicare 2014 Part D data from Google BigQuery, Medicare Exclusion list from the Office of Inspector General, and a geographical dataset.

Part D prescriptions contain a detail of drug prescription per beneficiary per drug type. The description includes npi of the provider, location of the provider, drug name, total drug amount supplied, total duration supplied, and total cost of the drug. The exclusion list include npi of providers who has been classified as fraud providers by the DOJ.

Topics that will be covered using these datasets include

1. Exploration of medicare drug prescription and cost across states
2. Exploration of excluded drug providers
3. Feature selection and engineering
4. Selection and optimization of classifications models

Data source:
1. https://cloud.google.com/bigquery/public-data/medicare
2. https://oig.hhs.gov/exclusions/exclusions_list.asp#instruct
3. https://simplemaps.com/data/us-cities