Data Wrangling of the project "Prescribing of Opioid among Medical Professionals"

By: Yonaton Heit

1) Data Acquiring:

The data set comes from a 2013 Medicare Part D report on the opioid prescription of medical professionals. The data reports the health care provider's name, state, the zip code, the specialty, the number of opioid prescribed (new prescriptions and refills), the total of prescription, and the percentage of drugs prescribed that are opioid. This data set contains over 1 million different health care providers. It was downloaded using the Socrata module and the query limit set to 1.5 million so all data points included.

2) Data Wrangling and Cleaning

The first thing I did was change the name of the columns to something more descriptive. Each health care provider was given a unique identification number. This number was used as the index for the pandas dataframe. The name catalogue was divided into two columns, a first and last name. In the cases where health care provider name listed were the health center rather than the name of the individual, the first name was empty. The first and last names were combined into a single column in order to remove empty columns.

At several entries, the number and percentages of opioids prescribed were empty. These were assumed to be and filled with zeros. Two versions of this dataframe were created, one that contained health care providers that prescribed opioids and one that included all health care providers. The complete dataframe contains 1049326 health care providers, the dataframe that only includes health care providers that prescribed opioids has 496744 providers.

For each unique specialty description, an integer identification number was given. These numbers will be used in order to categorize the health care providers. Note that several providers with similar specialty were given different number. For example, there are several different identification numbers for nurse. Several of these categories will most likely be combined in the future based on keywords (such as nurses and surgeons). There are currently 246 specialties found, only 169 of which prescribe opioids.

The population of each zip code was reported in the 2010 census. The population of zip codes in the same city were summed over and then assigned to the health care provider using their zip code. After this, of the 1049326 health providers, 1836 (0.17%) had unknown area population. Some operated outside the US and therefore not included in the census. Some health care providers were in the zip codes missing from the census.

3) Outlier Detection:

Since detecting outliers is the purpose of this analysis, for the purpose of cleaning the data, outlier detection was limited to filtering the data points outside of the feasible range. The only concern was if the percentages of opioids prescriptions were more than 100% or below 0%. The maximum percentage of opioid prescription was 100% and the minimum was 0%.