Capstone 1: Predicting Prescriptions, Data Wrangling

Rika Yatchak

January 25, 2019

1 Introduction

In this document we'll describe the different data-wrangling steps taken to prepare the Kaggle Opioid Prescriber Dataset for EDA and further analysis for my Springboard Capstone 1 project. The main goal of this project is to use historical provider drug prescription numbers to predict future demand.

2 Dataset Description

The dataset was uploaded to Kaggle in 2016. The dataset was compiled by AJ Pryor and is sourced from cms.gov. There are three different files:

- `opioids.csv` – Drug and generic names for 113 different opioids.
- `overdoses.csv` – Number of opioid-related overdose deaths in different states (no territories!) in the US in 2014.
- `prescribers.csv` – Prescriber information for 20,000 different prescribers in the US. Includes credentials, gender, state, specialty, and then the number of prescriptions given in 2014 for the 250 most commonly prescribed drugs. Finally, the boolean column `Opioid.Prescriber` indicates whether each provider prescribes opioids or not.

3 Data Wrangling Steps

Since the dataset originates from a Kaggle challenge, it is in very good condition: the csv files are easily read into pandas dataframes using `pd.read_csv()`. There are no missing values. The main concerns in this dataset are twofold: identifying outliers, and string processing of opioid names.

The `Overdoses` table requires no data wrangling. We will have to pay attention to outliers in the `Prescribers` table, and some string processing will be required in the `opioids.csv` table.

3.1 Outliers

One type of outlier is the provider state. Providers from certain US Territories and some small US states do not occur frequently in the dataset.

If provider state is used as a feature, it is important to remove providers from all territories and possibly also Alaska and Wyoming from the dataset when doing any regression tasks.

Another concern is the prescriber specialty. There are many specialties that are not well-represented in the dataset: out of **109** different specialties recorded in the `prescribers.csv` dataset, only **51** of them are represented by 30 or more providers in the dataset. Thus, for any logistic regression tasks, the dataset is cut to contain only the prescribers with these well-represented specialties.

Abbreviation	State/Territory	Count
AK	Alaska	39.0
WY	Wyoming	38.0
VI	U.S. Virgin Islands	3.0
GU	Guam	2.0
ZZ	Unknown	2.0
AE	U.S. Armed Forces (Overseas)	2.0
AA	U.S. Armed Forces (Americas)	1.0

Table 1: Territories and States with less than 50 providers in the dataset.

3.2 `Opioid.Prescriber` and string processing

The dataset documentation describes `Opioid.Prescriber` as a boolean value. The value 1 is assigned when the prescriber in question prescribed any opioid in 2014, and 0 otherwise. However, it's not clear which of the 250 drugs for which we have prescription records are opioids. For this, we need to compare the relevant columns of the `Prescribers` table and the drug names contained in `Opioids`.

Here, we run into a formatting problem: the drug names contained in `Prescribers` are formatted in a different way than the drug names in `Opioids`. For example, the opioid acetaminophen codeine is represented by `ACETAMINOPHEN.CODEINE` in `Prescribers` and `ACETAMINOPHEN-CODEINE` in `Opioids`. Hyphens, spaces, and withs are a problem. So we remove these and then split the drug names into tuples, so `ACETAMINOPHEN.CODEINE` becomes the tuple `(ACETAMINOPHEN, CODEINE)`. Then we compare the drug names in `Prescribers` and `Opioids`. It turns out that many providers have no recorded opioid prescriptions, but are still marked as opioid prescribers.