**ETL Project Write Up**

**Cristalle Pronier**

**Overview**

Increased scrutiny on the potential over-prescription of habit-forming drugs such as opioids has led to more examination of reported data. Our team choose to pull multiple healthcare datasets to create a database to examine this issue. We choose to narrow our data to records for California only for this project due to the sheer volume of records. This write up describes the dataset extracted and transformed by Cristalle.

This national dataset is published by the Centers for Medicare & Medicaid Services (CMS) and contains the names of all medical providers and their corresponding pharmaceutical prescriptions with related prescriber data who bill medical services under a CMS certified facility. This dataset also contains the National Provider Identifier (NPI) which is a unique number assigned to each provider. The second dataset in this project extracted and transformed by Hung also includes the NPI which will be a key link in joining the datasets prior to analysis.

**Extract**

The original dataset contains over 25 million records and is publicly available at the following link: <https://data.cms.gov/Medicare-Part-D/Medicare-Provider-Utilization-and-Payment-Data-201/77gb-8z53>. To pull data the CMS API (run by Socrata) was used with a filter query to return only California records using the two-digit postal abbreviation CA. The API documentation for developers was consulted to figure out the API url to pull only California records (<https://data.cms.gov/resource/bwz9-yuau.json?nppes_provider_state=CA>).

Initially, only 1,000 California records were being returned from the API call which is much lower than the expected amount of records. To troubleshoot this issue, I contacted the data technical assistance team and searched the documentation. I found in the documentation (team also emailed me the following day to confirm what I figured out) that there is a default limit of 1,000 records. This could be changed to the maximum of 50,000 with an additional limit defined on the API url (documentation: <https://dev.socrata.com/docs/paging.html>). The API url was configured to pull the maximum allowed records from California cases (<https://data.cms.gov/resource/bwz9-yuau.json?nppes_provider_state=CA&$limit=50000>). Data were returned in JSON format.

**Transform**

The JSON data was transformed into a Pandas Dataframe prior to cleaning the data. The dataset documentation was reviewed to better understand the data and to make informed decisions on data cleaning.

Three columns of unnecessary data were dropped from the Dataframe: bene\_count\_ge65\_suppress\_flag, description\_flag and ge65\_suppress\_flag. These data referred to the “suppression” flags and reasons if they were flagged. In this dataset, some of the data are suppressed meaning that it is not reported for one of several reasons. The unreported data shows up as “NaN.” The most common reason for suppressed data is a low number reported, in the case of this dataset, any number reported between 1-10 will be suppressed and shown as NaN. Null values were left as they were originally reported as some columns have a substantial amount. It was not appropriate to delete all records with a null in one or more column since there are so many and it was not appropriate to assign a value of zero because that could skew numeric analysis done with the data. Instead, future analysis using these data in the constructed database should consider how to handle the null values on a case-by-case basis.

**Load**

The cleaned dataset is loaded into a MySQL database. MySQL was chosen because the existing tabular data structure would work best. Loading Pandas Dataframes into MySQL is a straightforward process. With both datasets loaded into MySQL they can be queried, joined and reviewed as needed.