**Data Exploratory Phase: Understanding the data to predict Opioid Drug prescribers**

The document contains the steps that was implemented for the critical assessment of the datasets. The primary focus of the project is to study the rate of prescription drug related fatalities and attempt to predict future opioid prescribers and isolate the physicians who are high risk opioid prescribers.

The document highlights the initial investigations conducted on datasets to identify definite patterns can that can emerge. It also outlines the process adopted for spotting redundant columns and data anomalies. It provides an overview of the process used for tuning the data for building predictive models on the hypothesis.

**Initial process**

Several open source datasets related to opioid prescriptions were eliminated due to the following factors contributing towards the rejection:

* Very small datasets
* Highly imbalanced datasets with very limited negative or positive outcomes
* Normalized datasets which contained only tabulated features

**Selecting the dataset for the predictive model (data preparation)**

**Primary source:**

<https://www.kaggle.com/apryor6/us-opiate-prescriptions/data>

**Additional dataset used for validating stead increase in death caused due to overdose of prescription opioids**

<https://data.world/datasets/opioids>

The source from where the datasets were obtained states that Opioid prescriptions and fatalities related to prescription drug overdose are becoming an overwhelming problem for the United States. The crisis is escalating at a monumental rate every year. The model attempts to use an open source dataset consisting of prescription data along with overdose deaths and details of opioid prescriptions to detect generous “opioid drug prescribers”. The intention is to alert the behavior rather than find fault with the prescription behavior which will eventually save lives and decrease the alarming rate of addictive prescription drug related fatalities.

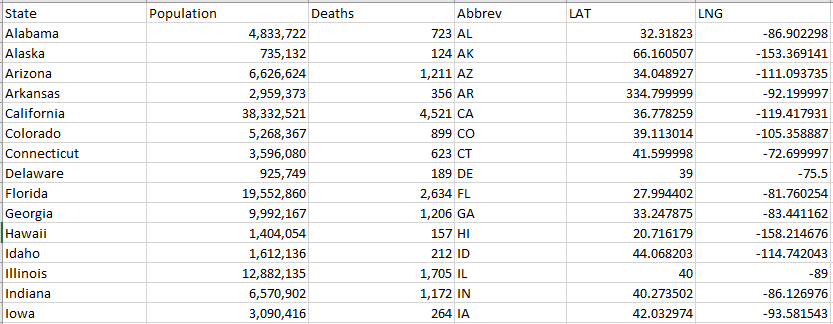
The model uses 4 different files to run the analytics namely:

* Overdose death rate information for the year 2014
* Overdose death rates from 1999- 2015 to confirm if this is an ongoing issue worthy of study
* Prescription details for the year 2014 comprising of both Opioids and non-opioid prescriptions
* Opioid drug list that would provide useful information to segregate opioid drugs versus the non-opioid drugs in the prescription details file

The following data sets were used for the exploratory phase to detect opioid prescribers:

1. **Overdose death dataset 1**

The data lists the total number of deaths related to prescription opioid overdosing for all the major states in the USA. The dataset will not be used in the final prediction to forecast opioid prescribers, but the dataset will be used to prove that the deaths related to opioid overdosing is a real issue that needs to be solved. Two overdose death files were used to scope the project and to test the hypothesis on whether there is a real time crisis related to death caused by excessive opioid prescriptions in the united states. I have included two additional columns labeled LAT and LNG in the dataset so the data can be processed for visualization which will be used to in an interactive Choropleth map



**Observed features of the dataset**

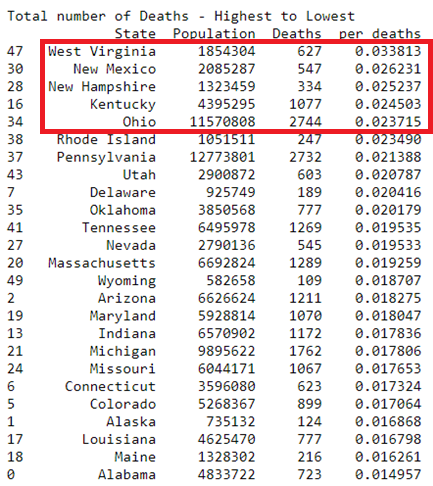
* In order to calculate the percentage death based on population, the columns Population and Deaths must be converted to a numeric value
* There are no missing values and the data follows a uniform consistency
* To use Folium, a latitude and longitude field along with a Geo Json file needs to be included
* The data is already aggregated and does not require any group by calls
* There are no complex datasets in the file that needs to be decomposed to multiple parts
* Data normalization is not required to tabulate any new columns

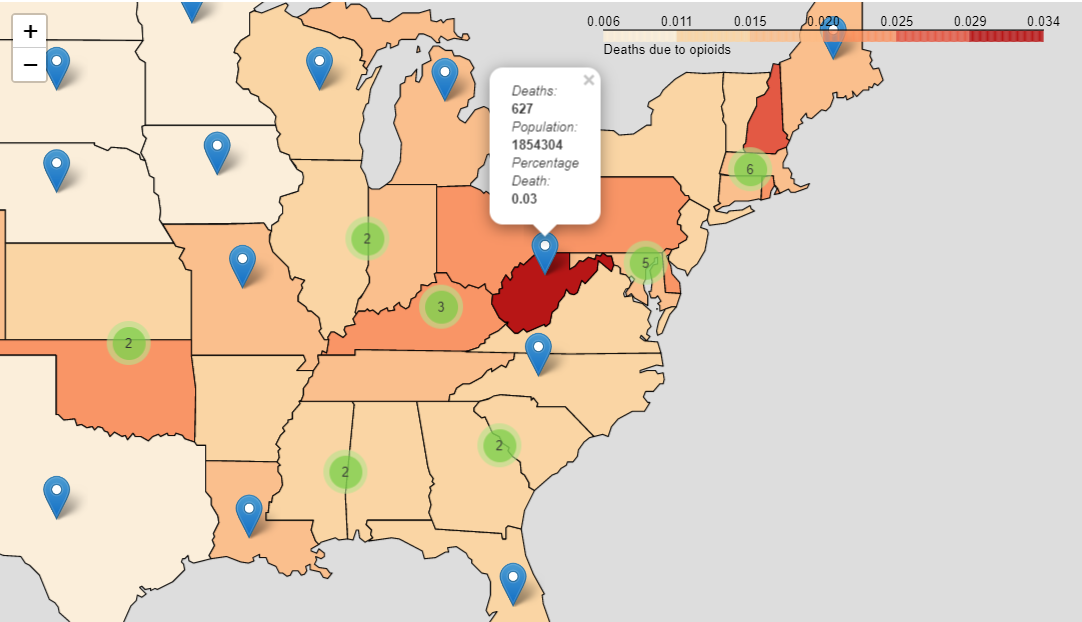
**Interesting aspect of the data**:

The total number of deaths related to prescription drugs overdose is a vital column to be examined. This is the primary column that would indicate the importance of the study

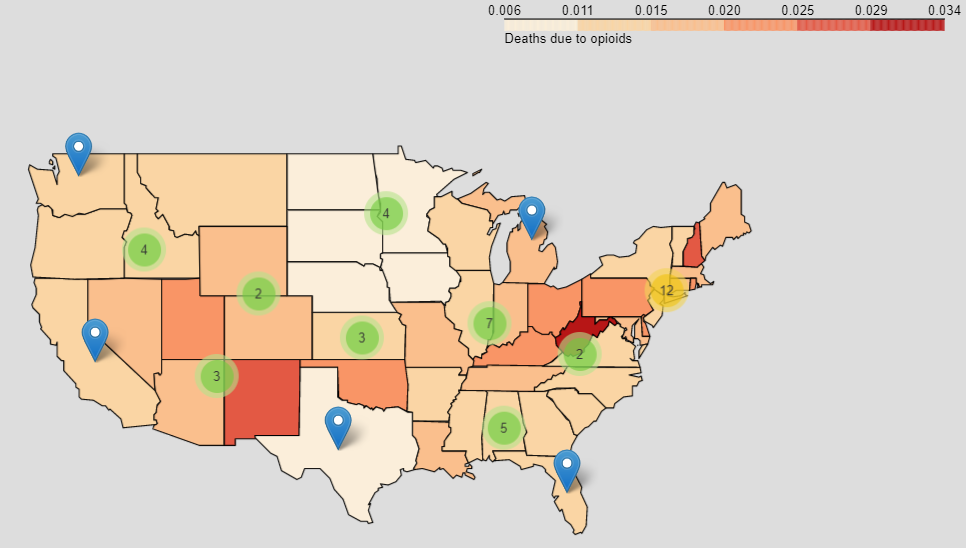
**Analysis of the data**:

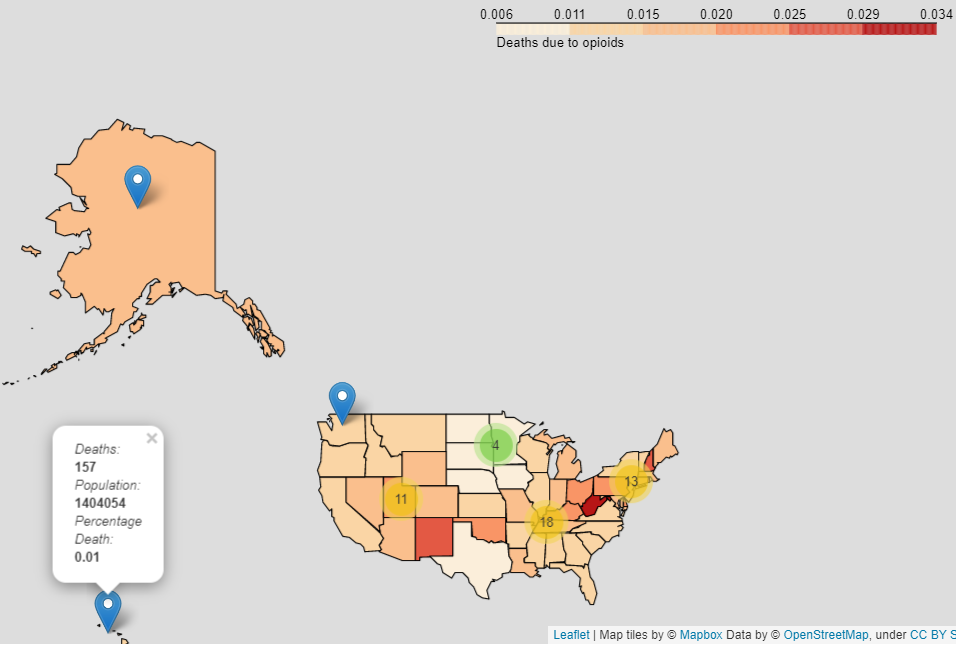
At a first glance we notice that California has the highest death rate due to Opioid overdosing. However, the population of California is larger than the population of other States. Th death rate is obviously based on the percentage of the total population of the state. Another column was thus introduced into the dataset which produced the percentage score of deaths based on the total population figure. We now notice that ***West Virginia, New Mexico, New Hampshire, Kentucky, and Ohio*** topped the list of total number of deaths caused by prescription opioid overdosing as shown in the following figures:





The final map of overdosing issue in the united states appears as indicated below:





1. **Overdose death dataset 2**

I wanted to dig deeper into the overdose death rates caused by prescription drugs as the dataset that I an investigating comprised of a single year figure (2014).

The primary question at this point was who many deaths occurred due to opioid prescriptions prior to or after 2014? Was there an exponential increase or a steady growth or was there a dip in the figures? For instance, if the values were exponentially larger prior to 2014 which were then slowly tapering off, then it is a clear indication that the problem is slowly resolving itself and it may not be a life-threatening crisis.

To address this predicament, a second dataset was also examined to zoom into the top five states which produced the largest number of opioid overdose deaths based on the population figures.

**Observed features of the dataset**

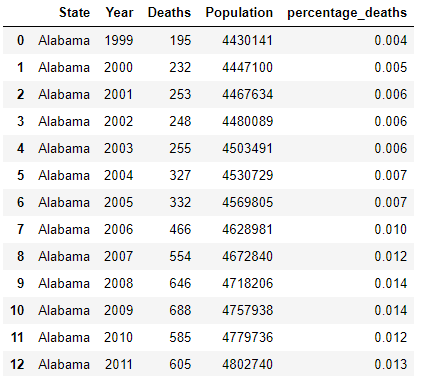
* There are no missing values and the data follows a uniform consistency
* The data is already aggregated and does not require any group by year
* There are no complex datasets in the file that needs to be decomposed to multiple parts
* Data normalization is not required to tabulate any new columns

**Interesting aspect of the data**:

As this is also a dataset that addressed the total number of deaths related to prescription drugs, the total number of deaths is a vital column to be examined. This is the primary column that would indicate the importance of the study

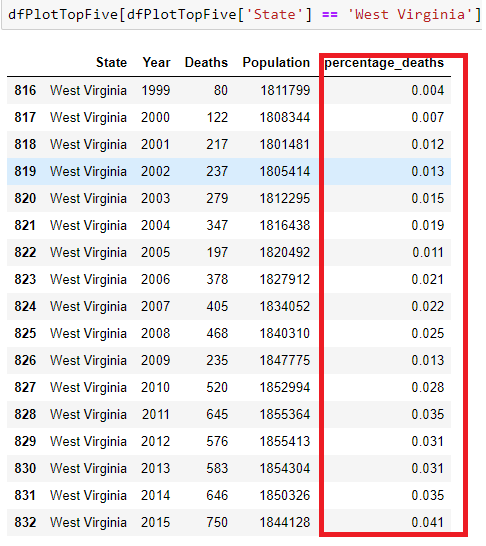
**Analysis of the data**:

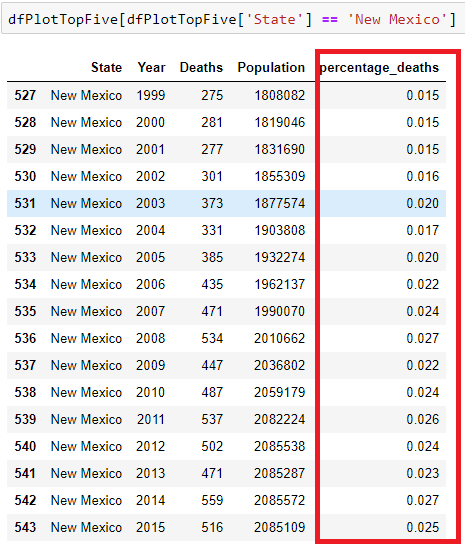
This dataset was primarily used to confirm if there is a steady increase in opioid overdose deaths. Like the previous sample, an additional column – percentage\_deaths, was engineered to ensure that the death rates accurately reflected the population figures.

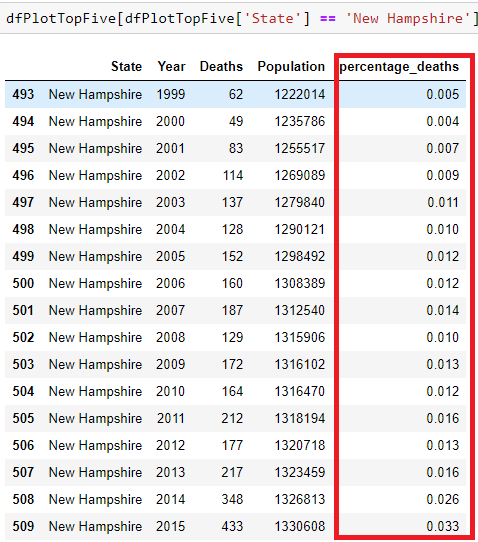


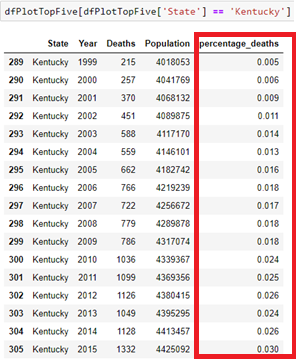
Inference from the dataset:

The top 5 states were microscopically investigated to examine the death rates over the course of year from 1999 – 2015. The dataset revealed that in all cases there was indeed a steady increase of prescription drug overdose death every year – while there were a few instances where the increase was stable, most of the values clearly indicates the fact that there is still an ongoing crisis on generous prescription of opioid drugs that needs to be resolved.

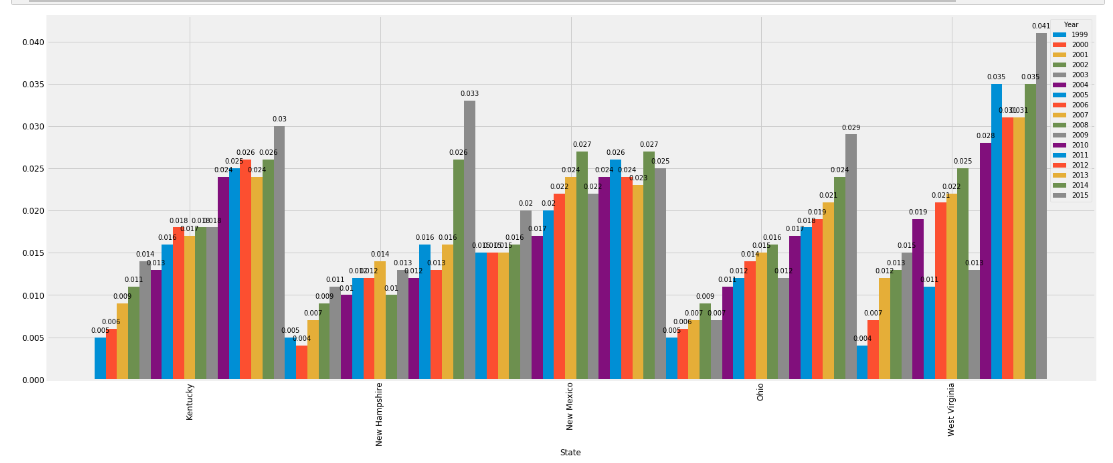








The outcome of this investigation is therefore to continue the study further as there is a definite issue related to opioid prescriptions as indicated in the graph below for the top 5 states with the greatest number of prescription drug related death rates in the year 2014:

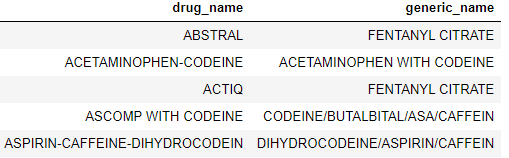


1. **Opioid dataset**

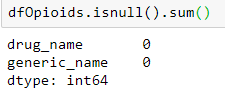
The opioid dataset lists a set of opioid prescription drugs that were directly linked to the prescription datasets. A sample of the data is shown in the figure below.

**Interesting aspect of the data**:

This is a straightforward dataset comprising of the prescription drug name and the generic name for the drug. On further investigation it was observed that these were the drugs that appeared in the primary prescription dataset that will be used for analysis for predicting overdose prescriptions



The dataset provided was completed with no null values that had to be removed from the dataset



**Analysis of the data**:

On further investigation it was also observed that these were the data leak columns in the prescription dataset that allowed the model to make unrealistically good predications with very high scores. While testing the datasets with various classification algorithms it was observed that the prescription models produced 99% to 100% accuracy and precision scores. This dataset was therefore provided for eliminating the prescription drugs from the prescription dataset

1. **Prescribers information**

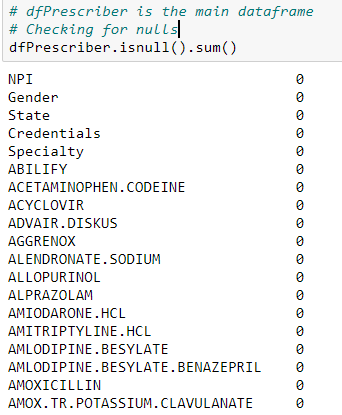
**About the dataset**

The dataset contains summaries of prescription records for 250 common opioid and non-opioid drugs prescribed by 25,000 unique licensed medical professionals for the year 2014 in the United States primarily for citizens covered under Part D Medicare formulary list. It also contains some metadata about the doctors.

This is the primary dataset that will be used for classifying opioid prescribers versus non prescribers. The dataset comprises of 25,000 rows and 256 columns. The dataset is large enough to be meaningful.



There were no null values present in the dataset- so there was no need to remove columns with more than 50% missing values

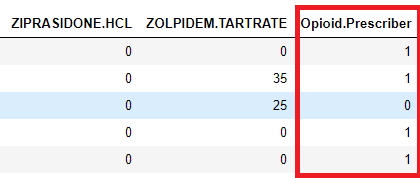


The data comprises of both categorical and continuous variable. After carefully examining the dataset.

**Deciding on a target column**

Our primary purpose of this model is to predict who is more likely to generously prescribe opioids and who is not prone to prescribe opioids and we need to find a suitable column that meets this requirement.

The binary column labeled Opioid.Prescriber will be the dependent predictive column while the other columns will be used as features to predict Opioid.Prescriber.



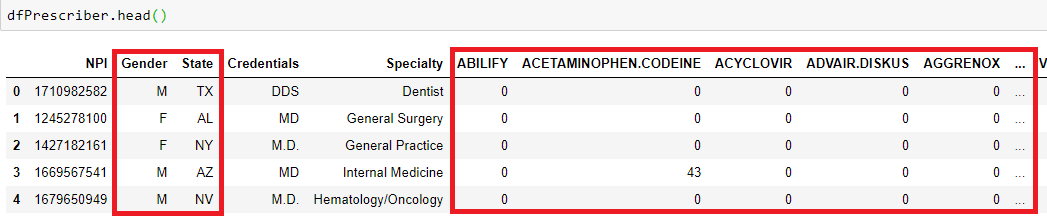
**Observed features of the dataset**

* There are no missing values and the data follows a uniform consistency
* Data aggregation is required for preliminary analysis to compare prescribers and non-prescribers by state and gender
* There are no complex datasets in the file that needs to be decomposed to multiple parts
* Opioid prescription columns do no match the values presented in the opioid datasets so some transformation of the columns via python code is required to eliminate these columns from the dataset
* Data normalization is not required to tabulate any new columns and most of the values are of a binary format
* Discretization of the data is required to transform categorical values to continuous numeric values

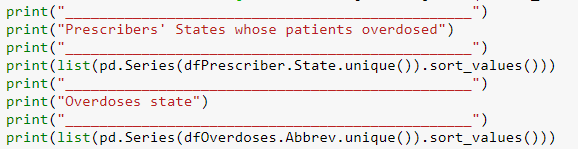
**Analysis of the dataset**:

Examining the datasets and narrowing down columns for cleaning and elimination

The dataset comprises of both categorical and continuous variables- a One Hot Encoding or factorization technique is required to convert string values to Integers



The prescription information available for the respective states were compared with the overdose death files:



The dataset comprises of the following columns:

* Physician details such as state, NPI number, Specialty
* Opioid prescriptions – drugs were derived from the opioid file with a binary 0 and 1 indicating if the drug was prescribed or not
* Non opioid prescription details- drugs that were not included in the opioid file were marked with a 0 and 1 indicating if the drug was prescribed or not
* Opioid.Prescription column that determines if the prescriber is an opioid prescriber or not (the supervised learning algorithm will use this column to learn about the dataset to predict unknown scenarios)

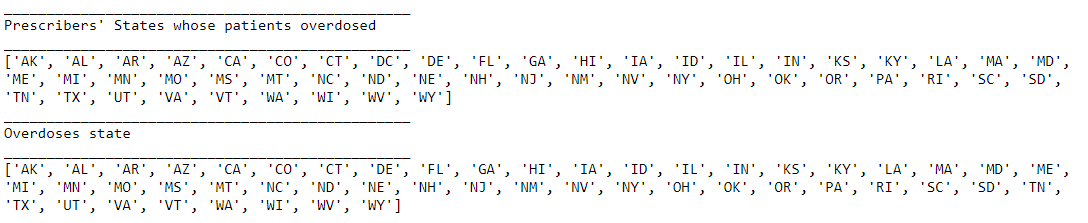
**Cleaning the dataset for analytics**:

* **Removing redundant feature states without enough of overdose death information**:

There were a few States identified that were not present in any of the overdose death, and these states were therefore removed from the study. They were as follows:

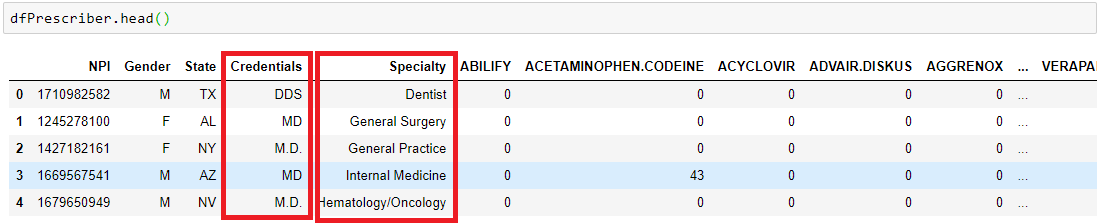
* PR: Puerto Rico
* AE: Armed Forces Europe
* GU: Guam
* VI: Virgin Islands
* ZZ: ???





* **Removing the column labeled Credentials**

From the dataset it is obvious that both Specialty and Credentials are redundant as they both contain the same values. Credentials was therefore eliminated from the dataset prior to analysis

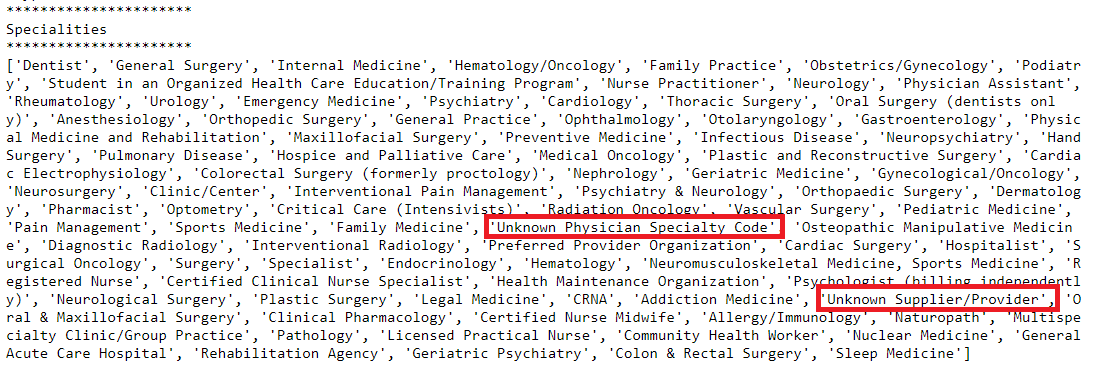


* **Removing the column labeled NPI**

The NPI number does not seem to add any value to the dataset this was also eliminated prior to predictions

* **Specialty feature Column**

The data comprises of two unknown specialties- on further investigation and running crude analysis on the dataset I decided to keep these two columns. I feel that there may be situations in the future which could present itself with similar values and my models should be able to predict if the prescriber is an opioid prescriber or not based on all values, so therefore after careful consideration I decided to keep both these values for analytics



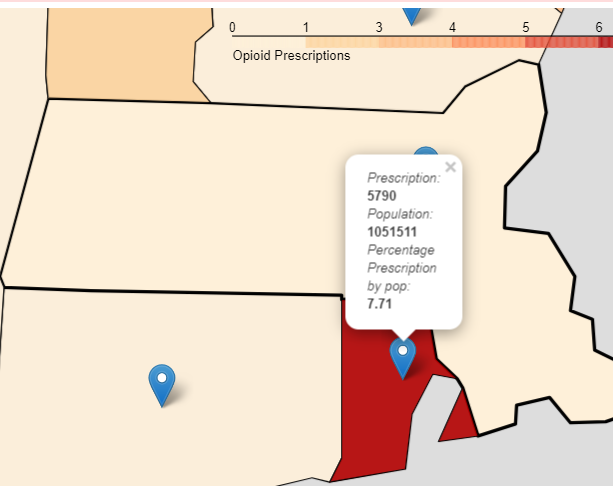
* **Analyzing opioid prescribers by state and state and gender**

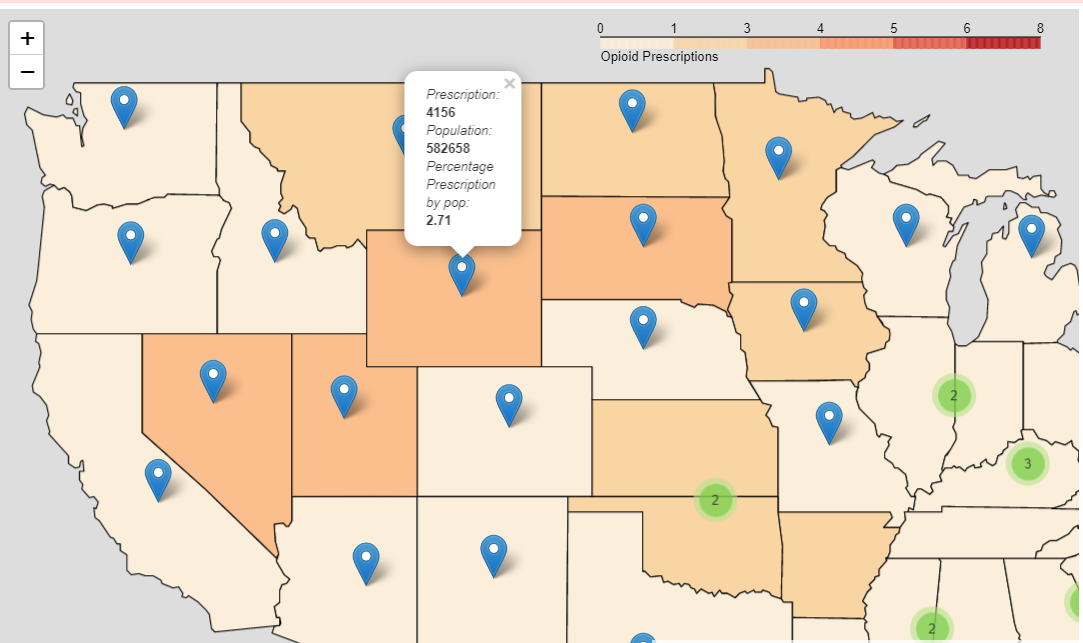
The prescriber dataset was cleansed for opioid drugs obtained from the opioid file and the values were investigated further. The primary intention of the study was to match the overdose information obtained from the prescription data with that of the overdose death rates for each state.

**Conclusion**: The top 5 prescription drug overdose details for the year 2014 did not match with the prior results obtained by evaluating the top 5 death rates. After further investigation, I was able to find out that Kaggle had not posted the entire dataset online, but a subset of the original dataset was provided. A link was provided directing me to the original dataset, but I was not able to find the parent files.

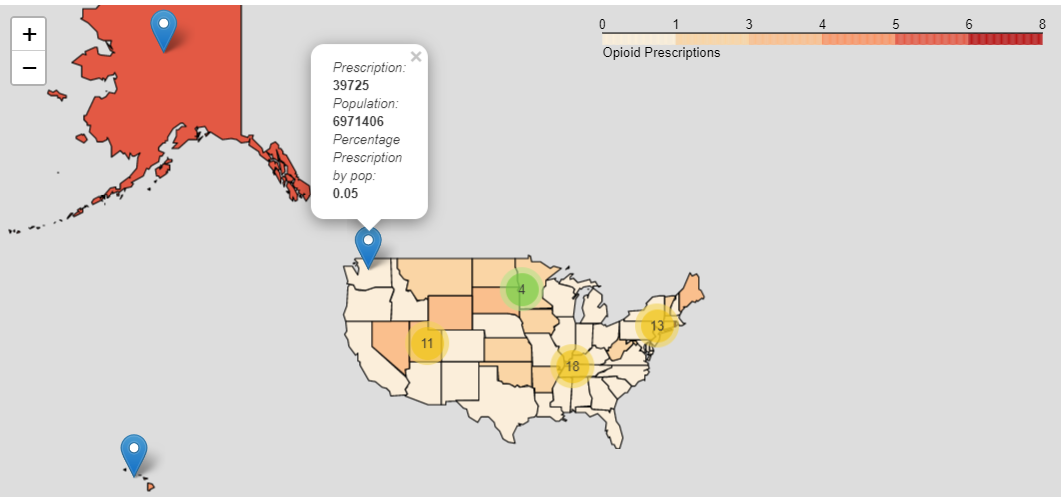
The following are the interesting observation obtained from analyzing the prescription information

* The largest number of opioid prescribers were from Rhode Island as indicated in the choropleth map





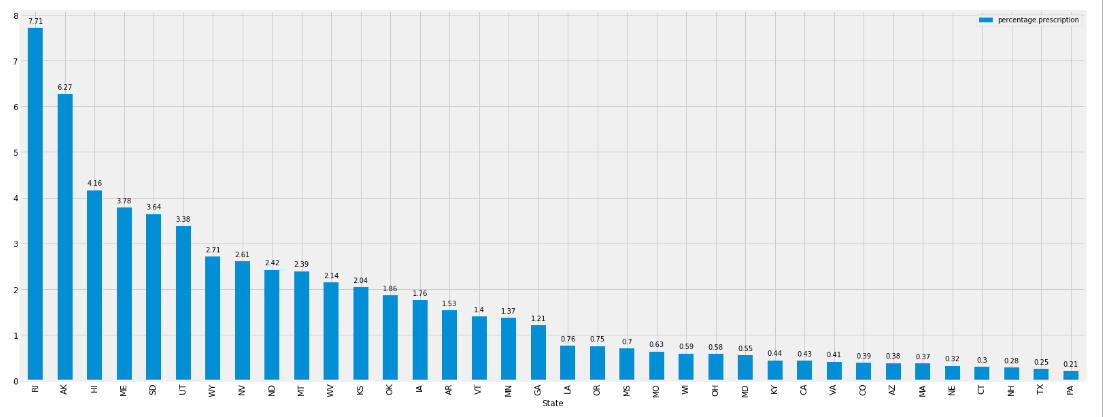
The map for opioid prescribers by State:



Data for opioid prescription for a state based on population was derived and plotted. Prescriptions lower than 0.02% were removed from the graph for visualization purpose

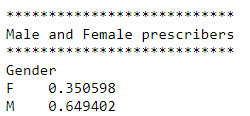
The top 5 states with the greatest number of opioid prescribers where from:

* Rhode Island
* Arkansas
* Hawaii
* Maine
* South Dakota

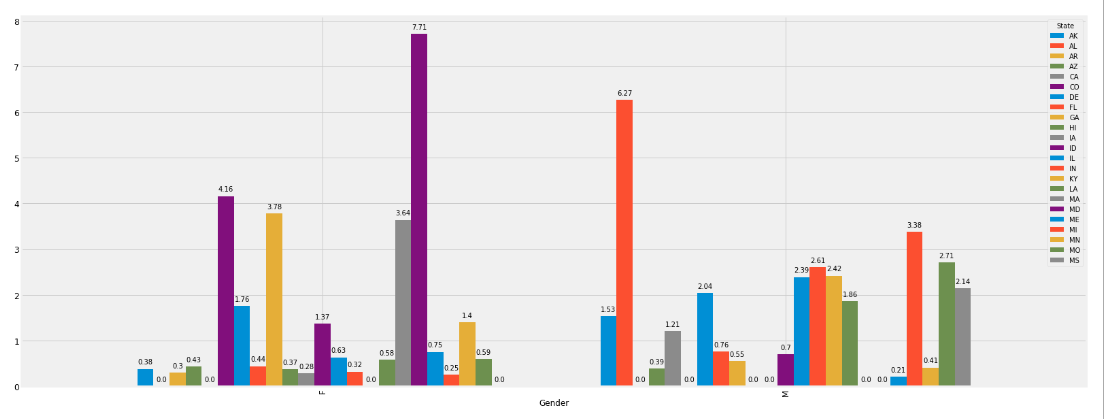


In order to find out if opioid prescriptions were related to gender, further analysis was conducted by grouping the dataset by gender and state. The following were observed:

Only 35% of the opioid prescriptions were from female prescribers while 64% of the prescriptions for opioids originated from male physicians



A graph showing the differences between male and female prescription drug prescribers is shown below:



The largest number of male opioid prescribers hailed from Alabama and the largest number of female opioid prescribers hailed from Alabama. Interestingly enough, it can be observed that the largest number of opioid prescriptions was from the female population (7.71 female versus 6.23 male)

The data for the graphs were derived by:

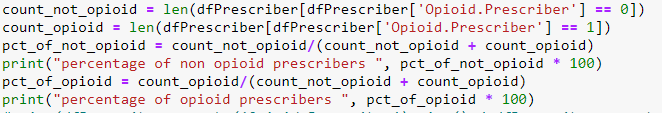
* Eliminating non opioid drug prescriptions
* Selecting only the opioid drug prescription listed in the opioid file
* Computing the total of all the opioid prescriptions for every State and State + Gender to obtain the respective data frames
* Joining the overdose dataset with the prescription dataset to obtain the population information for the respective states
* Calculating the percentage prescription for every state based on the total obtained for the total number of opioid prescriptions for every State and State + Gender
* The final results obtained for percentage prescription based on population information for State and State + Gender were then plotted in the graph shown above

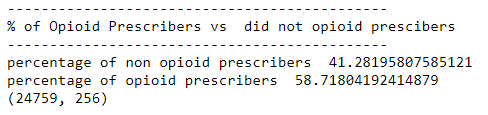
**Conclusion**: Gender is also a significant feature and a strong determinant of the outcome and hence it cannot be eliminated from the final prediction dataset.

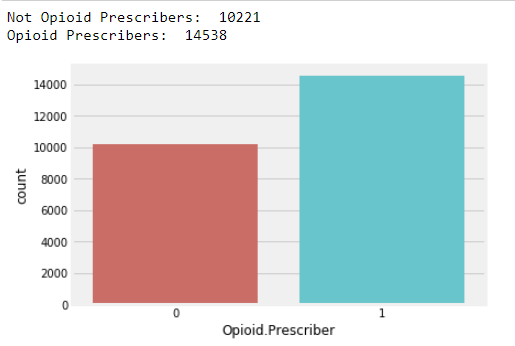
* **Analyzing opioid prescribers versus non opioid prescribers**

Since the primary focus of the study is to discover opioid prescribers versus non opioid prescribers for unknown or new situations, additional investigation was conducting to examine the data

41.29% of the prescribers were non opioid prescribers while almost 59% of the dataset comprised of opioid prescribers as shown below:







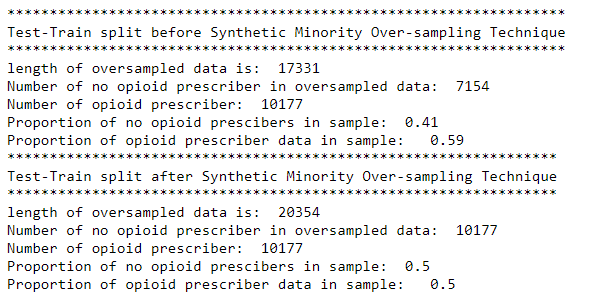
Although the dataset is somewhat balanced (this is not a highly imbalanced dataset), I feel that I need to create a balanced test – train split before passing the data to classification algorithms, especially because this is a sensitive model and we want to ensure that the test and train split is valid and highly balanced.

Before passing the prescriber data to classification algorithms the following critical steps are applied to tune the dataset for making it more prediction worthy:

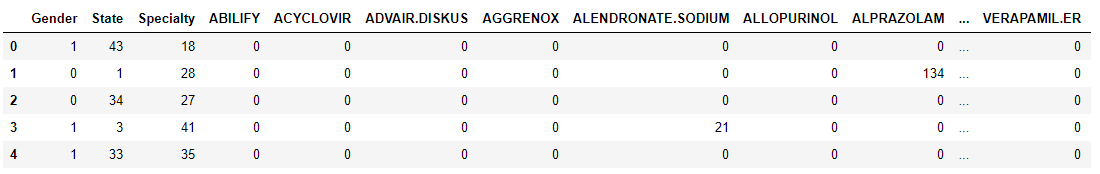
* Substituting missing values for the opioid and non-opioid columns with 0. A point to me noted that although this was applied while reading the data into the dataframe, there were not many missing data in the provided dataset
* Remove Opioid prescription column values to avoid data leak
* Remove Credential and NPI columns from the feature variables
  + Credential and Specialty data overlap
  + Gender, State, and Specialty are better predicators than NPI numbers which does not add any valuable meaning to the features
* Remove states that are not included in our overdose files
* Factorize the categorical variables to convert it to numerical values
* Identify the dependent target column for prediction of unknown cases (Opioid.Prescriber)
* Use SMOTE (Synthetic minority over sampling technique) to create a balanced 50:50 test-train split

Presenting the dataset details before applying it to the classifier:

**A perfect 50:50 ratio of opioid prescribers and opioid non prescribers**



**Removing Opioid prescription details, NPI, and Credentials and converting categorical to continuous variables**



**The following prediction models will be used to determine the binary classifier outcome**

* Logistic Regression
* Random Forest Classifier
* Decision Tree Classifier
* Gaussian Naïve Bayes
* Gradient Boosting
* KNN
* Linear Discriminant
* Bagging Classifier

**The metrics to be gathered which will determine the top algorithms to be used for unknown situations:**

* Average Precision score
* Recall Score
* Accuracy Score
* Confusion Matrix

Additionally, file 2 (overdose death rates from 1999 to 2015) will be used to forecast future deaths based on current trends. The time series models such as AR, ARMA, ARIMA will be used to project the data into the future to forecast future trends. I have decided to eliminate SARIMA as the data is not granular enough to observe trends related to seasonal influence