Final Project Report: CS210
Opioid Abuse Visual Analysis

By: Sana Rahman

Contents

I. Reassessment of Project Proposal
II. Data Collection
III. Data Storage
IV. Data Cleaning
V. Data Integration
VI. Exploratory Data Analysis (EDA)

Github Repo Link: https://github.com/sanar812/cs210-opioid-analysis.git

I. Reassessment of Project Proposal

While reviewing the proposal I made for this project, I noticed several factors that might be too complex or unrealistic, particularly given the constraints of time, experience, and data. My initial project definition read as follows:

> *"Assess existing demographic, social, and economic data to predict risk levels for substance use behaviors. The primary strategy to be used is predictive modeling, which will have to assign different combinations of data to risk levels. This project will use structured data in order to do this, though a limitation is the lack of flexibility of this data. Data collection methods will include surveys, and potentially APIs. Furthermore, this project will use public data, so the privacy of all sensitive data should be ensured."*

My initial plan was to code this project in Java, and because it is not a data language, predictive modeling would have been extremely difficult to achieve. For this reason, I chose not to do predictive analysis, instead focusing on data integration and visualization. However, I decided to use Python and PostgreSQL anyway, as Python is a data science language, and I found it to be much more suitable for data analysis.

I chose to use public data for analysis— while my hope was to perform advanced data integration on JSONs, CSVs, and other data types, the data I had access to was limited— I will elaborate on this in Section II. Consequently, I chose to perform integration on several different datasets with overlapping data.

My project proposal listed "Existing Issues in Current Data Management Practices" as follows:

1. Inconsistent data formats (e.g. CSV, Excel, JSON, etc.) across different sources
2. Inconsistent quality (e.g. variations in capitalization, number formats, etc.) across different sources
3. Lack of integration among datasets from various agencies (healthcare, law enforcement, etc.)
4. Ethical concerns regarding privacy and use of personal data

While I planned to tackle (2) during data cleaning and (4) was a non-issue because the data I chose to use contained no personal information, I decided (1) and (3) were not feasible to handle at this time. My reason for dismissing (1) is explained above. (3) had similar issues: I was only able to find datasets with *relevant* overlapping data from the healthcare department; I will also elaborate on this in Section II.

Finally, I chose to narrow down "Substance Abuse Predictive Analysis" to "Opioid Use Visual Analysis" in accordance with the data I was able to find.

II. Data Collection

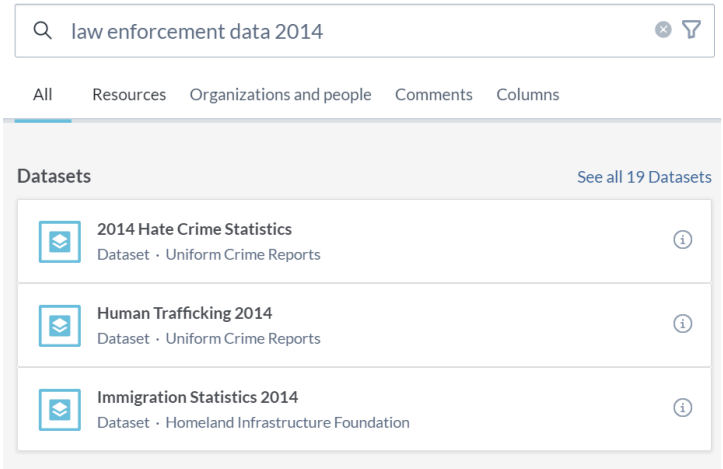
My initial search for data I could use in this project was simple: "drug abuse." My logic was that once I found a particularly useful source, I could find similar datasets from different sources to integrate with it. Resultantly, I found two separate datasets from the same source, one dated from 2012 to 2014, and the other from 2014 to 2016.

I believed this would be perfect for data cleaning and integration, as I found that there were differences in the data collected in both time frames. The former dataset presented statistics from two different drug abuse issues: (1) nonmedical use of pain relievers, and (2) tobacco product use. However, the latter dataset presented only tobacco product use data. I decided it would make sense to find more data correlating with tobacco use in this time frame, but found no additional data in this domain. Instead, I found a dataset containing a summary of opioid overdose deaths by state, from 1999 to 2014.

There were two overlaps between this summary and the dataset from 2012 to 2014. Firstly, the years matched up. Because I decided to focus on a time frame rather than a location, age group, etc., it no longer made sense to use the 2014 to 2016 dataset. Secondly, the summary specified *opioid* overdose deaths, while the 2012 to 2014 dataset had data for nonmedical use of pain relievers, which is otherwise known as illicit opioid use. Although both were in a CSV format, I decided this project would not need to integrate different formats of data given that these datasets would work well together and that searches for varying data *types* were fruitless.

As stated in Section I, data was very limited. While there are certainly problems with the lack of integration among datasets from various agencies as well as inconsistent data formats, the data I chose to focus on was presented in similar formats from healthcare agencies.

For example, the searches I performed after narrowing down to opioid use from 2012 to 2014 included keywords such as "2014" and "opioid." When narrowing the search to a *different* agency, I searched "law enforcement data 2014," yielding the following results:



Similar data was found for "2012" and "2013." It is evident that this data is irrelevant to opioid abuse, so I concluded that I would be unable to integrate data from different agencies.

I continued my searches along these parameters. I realized that the 2012 to 2014 dataset was only a sample size of 10 rows, but was able to find an elaborate dataset of 465 rows titled “Nonmed Use of Painrelievers” — the exact data from the 2012 to 2014 dataset but reduced, and presenting ratios rather than percentages (an easy fix: multiply each value by 100, but this data converted to percentages itself when I downloaded the CSV). I concluded my search when I had at least three related datasets that could be used for my project.

III. Data Storage

To store the data from several CSV files, I decided to load the CSVs into PostgreSQL, which I determined to be the most suitable application for my project given that I will be working with external data sources:

```
postgres=# create database opioid_data;
CREATE DATABASE
```

My datasets are listed as follows:

1. Opioid Deaths 1999-2014
2. 2012-2014 Nonmedical Use of Pain Relievers
3. Admissions aged 12 and older, by primary substance of abuse: Number, 2004-2014
4. Admissions aged 12 and older, by primary substance of abuse: Percent distribution, 2004-2014

```
opioid_data=# create table opioid_deaths(
opioid_data(# state varchar(200), year int, deaths int, populati
on int,
opioid_data(#  crude_rate decimal(4,1), crude_rate_lower_CI deci
mal(4,1),
opioid_data(# crude_rate_upper_CI decimal(4,1), prescription_yea
rly int);
CREATE TABLE
```

One immediate conflict was the values in the CSVs— the tables included words such as “unreliable” and “suppressed,” which did not match the `int` or `decimal` variables specified in their respective columns. For this reason, I decided to load the CSVs into a staging table with type `TEXT` for all columns:

```
opioid_data=# create table opioid_deaths_staging(
opioid_data(# state text, year text, deaths text, population tex
t,
opioid_data(#  crude_rate text, crude_rate_lower_CI text,
opioid_data(# crude_rate_upper_CI text, prescription_yearly text
);
CREATE TABLE
opioid_data=# \copy opioid_deaths_staging(state, year, deaths, p
opulation, crude_rate, crude_rate_lower_CI, crude_rate_upper_CI,
 prescription_yearly) from 'C:\Users\user\Downloads\Multiple Cau
se of Death, 1999-2014 v1.1.csv' delimiter ',' csv header;
COPY 816
```

I encountered several problems trying to upload the second dataset: (1) the dataset contained percentages, with the "%" symbol, so the values were not compatible with the `decimal` variable type of each column. (2) The dataset was actually downloaded as an Excel file. I simply chose to convert the file to a CSV, which made no changes to the data and was a simple enough solution so that I could continue using PostgreSQL. I had to create a staging table for this dataset, too, because of the percentages.

```
opioid_data=# create table nonmed_opioid_use(
opioid_data(# data_order int, state varchar(200),
opioid_data(# substate varchar(200), small_area_est decimal(5,2)
,
opioid_data(# lower_CI decimal(5,2), upper_CI decimal(5,2),
opioid_data(# map_group int);
CREATE TABLE
```

```
opioid_data=# create table nonmed_opioid_use_staging(
opioid_data(# data_order int, state varchar(200),
opioid_data(# substate varchar(200), small_area_est text,
opioid_data(# lower_CI text, upper_CI text,
opioid_data(# map_group int);
CREATE TABLE
```

```
opioid_data=# \copy nonmed_opioid_use_staging(data_order, state, substate, small_area_est, lower_CI, upper_CI, map_group) fr
om 'C:\Users\user\Downloads\2012-2014_Substate_SAE_Table_8.csv' delimiter ',' csv header;
COPY 465
```

The third dataset had numerical values that included commas, which did not match the `int` type of the columns. I had begun by creating a staging table for this dataset because of the peculiar way the data was formatted in the CSV:

	Table 1.1a. Admissions aged 12 and older, by primary substance of abuse: Number, 2004-2014		
Primary substance	2004 2005 2006 2007 2008 2009 2010 2011 2012	2013	2014
	Total 1,808,469 1,895,917 1,962,674 1,974,739 2,076,291 2,052,174 1,928,829 1,928,123 1,825,970	1,736,547	1,614,358
Alcohol	729,366 746,057 780,815 806,323 860,427 855,294 781,692 756,533 707,863	647,989	585,024
Alcohol only	402,999 412,323 433,612 448,707 483,494 479,134 431,490 416,950 391,899	367,484	327,694
Alcohol w/secondary drug	326,367 333,734 347,203 357,616 376,933 376,160 350,202 339,583 315,964	280,505	257,330
Opiates	323,409 332,179 353,671 364,629 408,578 434,536 436,473 478,621 479,709	494,247	489,680
Heroin	262,518 260,759 268,443 263,118 282,724 287,783 267,572 282,841 299,674	333,250	357,293
Other opiates/synthetics	60,891 71,420 85,228 101,511 125,854 146,753 168,901 195,780 180,035	160,997	132,387
Non-RX methadone	3,157 4,133 5,051 5,876 6,488 6,385 6,490 6,860 5,978	5,095	4,212
Other opiates/synthetics	57,734 67,287 80,177 95,635 119,366 140,368 162,411 188,920 174,057	155,902	128,175
Cocaine	248,492 268,573 278,258 260,849 239,581 193,269 158,960 151,910 125,995	105,392	87,510
Smoked cocaine	179,091 193,159 198,642 186,842 170,885 138,693 112,223 105,031 86,287	71,546	57,493
Non-smoked cocaine	69,401 75,414 79,616 74,007 68,696 54,576 46,737 46,879 39,708	33,846	30,017
Marijuana/hashish	285,193 303,649 313,521 317,307 359,157 373,257 357,952 351,896 317,739	288,917	247,461
Stimulants	143,525 172,778 164,148 151,886 131,739 120,084 119,190 117,432 127,268	141,155	144,427
Methamphetamine	124,500 154,057 155,987 143,338 122,797 111,769 108,894 107,242 117,594	131,270	135,264
1			
Other amphetamines	18,010 17,723 6,941 6,620 6,910 7,285 9,054 8,632 8,637	9,006	8,395
Other stimulants	1,015 998 1,220 1,928 2,032 1,030 1,242 1,558 1,037	879	768
Other drugs	28,270 28,703 28,764 29,892 37,034 43,614 47,593 46,601 42,191	37,818	33,511
Tranquilizers	8,169 8,712 10,302 11,672 13,497 15,615 17,242 19,217 18,066	15,916	15,106
Benzodiazepines	7,500 8,163 9,767 11,128 12,967 15,048 16,716 18,781 17,664	15,600	14,851

It can be seen that the values for 2004 to 2012 are crowded into one column, the second row contains the CSV headers instead of the first, the third row is formatted differently than the rest due to the presentation of totals, and one row has completely random inputs of null data and the number "1" — staging and cleaning would certainly be needed. However, I did not anticipate the commas posing an issue, as the formatting of the cells displays commas while the raw inputs do not contain commas. Because the table was already created at this point, I simply had to alter the column type of the last two columns. The final `admissions_by_number` table would have a column for `year_2012` rather than `combined_years`, as this project focuses on the time frame of 2012 to 2014.

```
opioid_data=# create table admissions_by_number_staging(
opioid_data(# primary_substance varchar(250), combined_years text, year_2013 int, year_2014 int);
CREATE TABLE
```

```
opioid_data=# \copy admissions_by_number_staging(primary_substance, combined_years, year_2013, year_2014) from 'C:\Users\use
r\Downloads\Table 1.1a.csv' delimiter ',' csv header;
ERROR:  invalid input syntax for type integer: "1,736,547"
CONTEXT:  COPY admissions_by_number_staging, line 3, column year_2013: "1,736,547"
opioid_data=# alter table admissions_by_number_staging alter column year_2013 type text, alter column year_2014 type text;
ALTER TABLE
opioid_data=# \copy admissions_by_number_staging(primary_substance, combined_years, year_2013, year_2014) from 'C:\Users\use
r\Downloads\Table 1.1a.csv' delimiter ',' csv header;
COPY 32
```

```
opioid_data=# create table admissions_by_number(
opioid_data(# primary_substance varchar(250), year_2012 int, year_2013 int, year_2014 int);
CREATE TABLE
```

The final dataset is the same as the third but with percentages replacing the numbers.

```
opioid_data=# create table admissions_by_percentage_staging(
opioid_data(# primary_substance varchar(250), combined_years text, year_2013 text, year_2014 text);
CREATE TABLE
opioid_data=# \copy admissions_by_percentage_staging(primary_substance, combined_years, year_2013, year_2014) from 'C:\Users
\user\Downloads\Table 1.1b.csv' delimiter ',' csv header;
COPY 32
opioid_data=# create table admissions_by_percentage(
opioid_data(# primary_substance varchar(250), year_2012 decimal(3,1), year_2013 decimal(3,1), year_2014 decimal(3,1));
CREATE TABLE
```

By the end of the data storage stage, the `opioid_data` database had the following schema:

```
opioid_data=# \dt
                    List of relations
 Schema |                Name                | Type  |  Owner
--------+------------------------------------+-------+----------
 public | admissions_by_number               | table | postgres
 public | admissions_by_number_staging       | table | postgres
 public | admissions_by_percentage           | table | postgres
 public | admissions_by_percentage_staging   | table | postgres
 public | nonmed_opioid_use                  | table | postgres
 public | nonmed_opioid_use_staging          | table | postgres
 public | opioid_deaths                      | table | postgres
 public | opioid_deaths_staging              | table | postgres
(8 rows)
```

IV. Data Cleaning

`1. opioid_deaths`

The primary concerns with this table were the string values ‘Unreliable’ and ‘Suppressed’ which could not be inserted into the `int` or `decimal` columns. To narrow this, first, I inserted any rows *without* these strings into the final `opioid_deaths` table from the `opioid_deaths_staging` table, casting all values from `text` to their respective types.

```
opioid_data=# insert into opioid_deaths(state, year, deaths, population, crude_rate, crude_rate_lower_ci, crude_rate_upper_ci, p
rescription_yearly)
opioid_data-# select
opioid_data-# cast(state as varchar(200)),
opioid_data-# cast(year as int),
opioid_data-# cast(deaths as int),
opioid_data-# cast(population as int),
opioid_data-# cast(crude_rate as decimal(4,1)),
opioid_data-# cast(crude_rate_lower_ci as decimal(4,1)),
opioid_data-# cast(crude_rate_upper_ci as decimal(4,1)),
opioid_data-# cast(prescription_yearly as int)
opioid_data-# from opioid_deaths_staging
opioid_data-# where deaths not like
opioid_data-# 'Unreliable' and deaths not like 'Suppressed'
opioid_data-# and population not like 'Unreliable' and populatio
n not like 'Suppressed'
opioid_data-# and crude_rate not like 'Unreliable' and crude_rat
e not like 'Suppressed'
opioid_data-# and crude_rate_lower_ci not like 'Unreliable' and
crude_rate_lower_ci not like 'Suppressed'
opioid_data-# and crude_rate_upper_ci not like 'Unreliable' and
crude_rate_upper_ci not like 'Suppressed';
INSERT 0 772
```

However, I still had to include the remaining data, so I chose to replace all string values with `null` values, sending them into the `opioid_deaths` table. I created a unique constraint on the table to avoid repeats of any state-year combinations, as `state` and `year` each repeat numerous times, but a combination of both will always be unique in the table.

```
opioid_data=# alter table opioid_deaths add constraint opioid_de
aths_unique_state_year unique (state, year);
ALTER TABLE
opioid_data=# insert into opioid_deaths(state, year, deaths, population, crude_rate, crude_rate_lower_ci, crude_rate_upper_ci, prescr
iption_yearly)
opioid_data-# select state, year::int,
opioid_data-# case when deaths = 'Unreliable' or deaths = 'Suppressed' then null else deaths::int end,
opioid_data-# case when population = 'Unreliable' or population = 'Suppressed' then null else population::int end,
opioid_data-# case when crude_rate = 'Unreliable' or crude_rate = 'Suppressed' then null else crude_rate::decimal(4,1) end,
opioid_data-# case when crude_rate_lower_ci = 'Unreliable' or crude_rate_lower_ci = 'Suppressed' then null else crude_rate_lower_ci::
decimal(4,1) end,
opioid_data-# case when crude_rate_upper_ci = 'Unreliable' or crude_rate_upper_ci = 'Suppressed' then null else crude_rate_upper_ci::
decimal(4,1) end,
opioid_data-# prescription_yearly::int
opioid_data-# from opioid_deaths_staging
opioid_data-# on conflict(state, year) do nothing;
INSERT 0 44
```

The final 44 rows in `opioid_deaths` showed:

Alaska	2001	17	633714	1.6	4.3	138
Alaska	2002		642337			142
Alaska	2003	14	648414	1.2	3.6	149
Alaska	2004	11	659286	0.8	3.0	155
Alaska	2005	19	666946	1.7	4.4	163
Alaska	2007	16	680300	1.3	3.8	184
District of Columbia	2009	13	592228	1.2	3.8	202
Iowa	1999	15	2917634	0.3	0.8	116
Iowa	2000	19	2926324	0.4	1.0	126
Mississippi	1999	16	2828408	0.3	0.9	116
Mississippi	2000	14	2844658	0.3	0.8	126
Montana	1999	16	897507	1.0	2.9	116
Montana	2000	13	902195	0.8	2.5	126
Nebraska	1999		1704764			116
Nebraska	2000	16	1711263	0.5	1.5	126
Nebraska	2003	14	1738643	0.4	1.4	149
North Dakota	1999		644259			116
North Dakota	2000		642200			126
North Dakota	2001		639062			138
North Dakota	2002	14	638168	1.2	3.7	142
North Dakota	2003	10	638817	0.8	2.9	149
North Dakota	2004	13	644705	1.1	3.4	155
North Dakota	2005		646089			163
North Dakota	2006		649422			174
North Dakota	2007	17	652822	1.5	4.2	184
North Dakota	2009	18	664968	1.6	4.3	202
North Dakota	2010	18	672591	1.6	4.2	210
North Dakota	2011	10	683932	0.7	2.7	219
North Dakota	2012		699628			217
North Dakota	2013	14	723393	1.1	3.2	207
South Dakota	1999		750412			116
South Dakota	2000		754844			126
South Dakota	2001	10	757972	0.6	2.4	138
South Dakota	2002	13	760020	0.9	2.9	142
South Dakota	2003	10	763729	0.6	2.4	149
Vermont	1999	17	604683	1.6	4.5	116
Wyoming	1999		491780			116
Wyoming	2000	10	493782	1.0	3.7	126
Wyoming	2001		494657			138
Wyoming	2002	17	500017	2.0	5.4	142
Wyoming	2003		503453			149
Wyoming	2004	14	509106	1.5	4.6	155
Wyoming	2005	10	514157	0.9	3.6	163
Wyoming	2006	14	522667	1.5	4.5	174

At this point, both the CSV and `opioid_deaths` table contained 816 rows of data.

2. nonmed_opioid_use

This dataset contained percentage symbols, so I simply removed them as I loaded the `nonmed_opioid_use` table.

```
opioid_data=# insert into nonmed_opioid_use(data_order, state, substa
te, small_area_est, lower_ci, upper_ci, map_group)
opioid_data-# select data_order, state, substate,
opioid_data-# replace(small_area_est, '%', '')::decimal(5,2),
opioid_data-# replace(lower_ci, '%', '')::decimal(5,2),
opioid_data-# replace(upper_ci, '%', '')::decimal(5,2),
opioid_data-# map_group
opioid_data-# from nonmed_opioid_use_staging;
INSERT 0 465
```

The `nonmed_opioid_use_staging` table contains 465 rows. Below is a preview:

```
opioid_data=# select * from nonmed_opioid_use_staging;
 data_order |        state         |                    substate                    | small_area_est | lower_ci | upper_ci | map_group
------------+----------------------+------------------------------------------------+----------------+----------+----------+-----------
          1 | Total United States  | Total United States                            | 4.31%          | 4.17%    | 4.45%    |         1
          2 | Northeast            | Northeast                                      | 3.82%          | 3.57%    | 4.10%    |         2
          3 | Midwest              | Midwest                                        | 4.21%          | 4.00%    | 4.43%    |         2
          4 | South                | South                                          | 4.30%          | 4.09%    | 4.51%    |         2
          5 | West                 | West                                           | 4.78%          | 4.48%    | 5.10%    |         2
          6 | Alabama              | Alabama                                        | 5.24%          | 4.40%    | 6.24%    |         3
          7 | Alabama              | Region 1                                       | 4.88%          | 3.77%    | 6.29%    |         4
          8 | Alabama              | Region 2                                       | 5.62%          | 4.48%    | 7.03%    |         4
          9 | Alabama              | Region 3                                       | 5.12%          | 3.99%    | 6.56%    |         4
         10 | Alabama              | Region 4                                       | 5.27%          | 4.13%    | 6.72%    |         4
         11 | Alaska               | Alaska                                         | 4.72%          | 3.95%    | 5.64%    |         3
         12 | Alaska               | Anchorage                                      | 5.02%          | 4.02%    | 6.24%    |         4
         13 | Alaska               | Northern                                       | 4.61%          | 3.59%    | 5.90%    |         4
         14 | Alaska               | South Central                                  | 4.58%          | 3.55%    | 5.89%    |         4
         15 | Alaska               | Southeast                                      | 4.15%          | 3.10%    | 5.53%    |         4
         16 | Arizona              | Arizona                                        | 5.18%          | 4.29%    | 6.24%    | -- More  --
```

The `nonmed_opioid_use` table now also contains 465 rows, with `small_area_est`, `lower_ci`, and `upper_ci` of type `numeric(5,2)`:

```
opioid_data=# select * from nonmed_opioid_use;
 data_order |        state         |                    substate                    | small_area_est | lower_ci | upper_ci | map_group
------------+----------------------+------------------------------------------------+----------------+----------+----------+-----------
          1 | Total United States  | Total United States                            |           4.31 |     4.17 |     4.45 |         1
          2 | Northeast            | Northeast                                      |           3.82 |     3.57 |     4.10 |         2
          3 | Midwest              | Midwest                                        |           4.21 |     4.00 |     4.43 |         2
          4 | South                | South                                          |           4.30 |     4.09 |     4.51 |         2
          5 | West                 | West                                           |           4.78 |     4.48 |     5.10 |         2
          6 | Alabama              | Alabama                                        |           5.24 |     4.40 |     6.24 |         3
          7 | Alabama              | Region 1                                       |           4.88 |     3.77 |     6.29 |         4
          8 | Alabama              | Region 2                                       |           5.62 |     4.48 |     7.03 |         4
          9 | Alabama              | Region 3                                       |           5.12 |     3.99 |     6.56 |         4
         10 | Alabama              | Region 4                                       |           5.27 |     4.13 |     6.72 |         4
         11 | Alaska               | Alaska                                         |           4.72 |     3.95 |     5.64 |         3
         12 | Alaska               | Anchorage                                      |           5.02 |     4.02 |     6.24 |         4
         13 | Alaska               | Northern                                       |           4.61 |     3.59 |     5.90 |         4
         14 | Alaska               | South Central                                  |           4.58 |     3.55 |     5.89 |         4
         15 | Alaska               | Southeast                                      |           4.15 |     3.10 |     5.53 |         4
         16 | Arizona              | Arizona                                        |           5.18 |     4.29 |     6.24 |         3
```

3. `admissions_by_number`

This dataset contains several problems. To review:

- Values for 2004 to 2012 are crowded into one column
- Row 1 contains the CSV headers
- Row 2 is formatted differently than the rest due to the presentation of totals
- The row between `primary_substance = 'Methamphetamine'` and `primary_substance = 'Other amphetamines'` has a random input of '1'

See below:

```
opioid_data=# select * from admissions_by_number_staging;
     primary_substance     |                                         combined_years                                          | year_2013 | year_2014
---------------------------+-------------------------------------------------------------------------------------------------+-----------+-----------
 Primary substance         | 2004 2005 2006 2007 2008 2009 2010 2011 2012                                                    | 2013      | 2014
                           | Total 1,808,469 1,895,917 1,962,674 1,974,739 2,076,291 2,052,174 1,928,829 1,928,123 1,825,970 | 1,736,547 | 1,614,358
 Alcohol                   | 729,366 746,057 780,815 806,323 860,427 855,294 781,692 756,533 707,863                         | 647,989   | 585,024
 Alcohol only              | 402,999 412,323 433,612 448,707 483,494 479,134 431,490 416,950 391,899                         | 367,484   | 327,694
 Alcohol w/secondary drug  | 326,367 333,734 347,203 357,616 376,933 376,160 350,202 339,583 315,964                         | 280,505   | 257,330
 Opiates                   | 323,409 332,179 353,671 364,629 408,578 434,536 436,473 478,621 479,709                         | 494,247   | 489,680
 Heroin                    | 262,518 260,759 268,443 263,118 282,724 287,783 267,572 282,841 299,674                         | 333,250   | 357,293
 Other opiates/synthetics  | 60,891 71,420 85,228 101,511 125,854 146,753 168,901 195,780 180,035                            | 160,997   | 132,387
 Non-RX methadone          | 3,157 4,133 5,051 5,876 6,488 6,385 6,490 6,860 5,978                                           | 5,095     | 4,212
 Other opiates/synthetics  | 57,734 67,287 80,177 95,635 119,366 140,368 162,411 188,920 174,057                             | 155,902   | 128,175
 Cocaine                   | 248,492 268,573 278,258 260,849 239,581 193,269 158,960 151,910 125,995                         | 105,392   | 87,510
 Smoked cocaine            | 179,091 193,159 198,642 186,842 170,885 138,693 112,223 105,031 86,287                          | 71,546    | 57,493
 Non-smoked cocaine        | 69,401 75,414 79,616 74,007 68,696 54,576 46,737 46,879 39,708                                  | 33,846    | 30,017
 Marijuana/hashish         | 285,193 303,649 313,521 317,307 359,157 373,257 357,952 351,896 317,739                         | 288,917   | 247,461
 Stimulants                | 143,525 172,778 164,148 151,886 131,739 120,084 119,190 117,432 127,268                         | 141,155   | 144,427
 Methamphetamine           | 124,500 154,057 155,987 143,338 122,797 111,769 108,894 107,242 117,594                         | 131,270   | 135,264
 1                         |                                                                                                 |           |
 Other amphetamines        | 18,010 17,723 6,941 6,620 6,910 7,285 9,054 8,632 8,637                                         | 9,006     | 8,395
 Other stimulants          | 1,015 998 1,220 1,928 2,032 1,030 1,242 1,558 1,037                                             | 879       | 768
 Other drugs               | 28,270 28,703 28,764 29,892 37,034 43,614 47,593 46,601 42,191                                  | 37,818    | 33,511
 Tranquilizers             | 8,169 8,712 10,302 11,672 13,497 15,615 17,242 19,217 18,066                                    | 15,916    | 15,106
 Benzodiazepines           | 7,500 8,163 9,767 11,128 12,967 15,048 16,716 18,781 17,664                                     | 15,600    | 14,851
 Other tranquilizers       | 669 549 535 544 530 567 526 436 402                                                             | 316       | 255
 Sedatives/hypnotics       | 4,179 4,513 4,149 4,601 4,973 5,335 4,439 3,971 3,456                                           | 3,354     | 2,821
 Barbiturates              | 1,294 1,402 1,053 1,095 1,204 1,343 1,402 947 773                                               | 1,004     | 1,119
 Other sedatives/hypnotics | 2,885 3,111 3,096 3,506 3,769 3,992 3,037 3,024 2,683                                           | 2,350     | 1,702
 Hallucinogens             | 2,290 2,046 1,645 1,655 1,905 1,874 1,782 1,991 2,141                                           | 2,157     | 1,864
 PCP                       | 3,236 2,888 2,869 3,207 4,059 4,436 4,739 5,749 5,847                                           | 5,324     | 4,910
 Inhalants                 | 1,191 1,372 1,126 1,140 1,383 1,611 1,540 1,256 1,147                                           | 940       | 791
 Over-the-counter          | 828 811 991 923 1,157 1,722 2,286 1,315 1,076                                                   | 1,060     | 911
 Other                     | 8,377 8,361 7,682 6,694 10,060 13,021 15,565 13,102 10,458                                      | 9,067     | 7,108
 None reported             | 50,214 43,978 43,497 43,853 39,775 32,120 26,969 25,130 25,205                                  | 21,029    | 26,745
(32 rows)
```

First, I removed the rows with no data:

```
opioid_data=# delete from admissions_by_number_staging
opioid_data-# where primary_substance = 'Primary substance' or primary_substance = '1';
DELETE 2
opioid_data=# select * from admissions_by_number_staging;
     primary_substance     |                                         combined_years                                          | year_2013 | year_2014
---------------------------+-------------------------------------------------------------------------------------------------+-----------+-----------
                           | Total 1,808,469 1,895,917 1,962,674 1,974,739 2,076,291 2,052,174 1,928,829 1,928,123 1,825,970 | 1,736,547 | 1,614,358
 Alcohol                   | 729,366 746,057 780,815 806,323 860,427 855,294 781,692 756,533 707,863                         | 647,989   | 585,024
 Alcohol only              | 402,999 412,323 433,612 448,707 483,494 479,134 431,490 416,950 391,899                         | 367,484   | 327,694
 Alcohol w/secondary drug  | 326,367 333,734 347,203 357,616 376,933 376,160 350,202 339,583 315,964                         | 280,505   | 257,330
 Opiates                   | 323,409 332,179 353,671 364,629 408,578 434,536 436,473 478,621 479,709                         | 494,247   | 489,680
 Heroin                    | 262,518 260,759 268,443 263,118 282,724 287,783 267,572 282,841 299,674                         | 333,250   | 357,293
 Other opiates/synthetics  | 60,891 71,420 85,228 101,511 125,854 146,753 168,901 195,780 180,035                            | 160,997   | 132,387
 Non-RX methadone          | 3,157 4,133 5,051 5,876 6,488 6,385 6,490 6,860 5,978                                           | 5,095     | 4,212
 Other opiates/synthetics  | 57,734 67,287 80,177 95,635 119,366 140,368 162,411 188,920 174,057                             | 155,902   | 128,175
 Cocaine                   | 248,492 268,573 278,258 260,849 239,581 193,269 158,960 151,910 125,995                         | 105,392   | 87,510
 Smoked cocaine            | 179,091 193,159 198,642 186,842 170,885 138,693 112,223 105,031 86,287                          | 71,546    | 57,493
 Non-smoked cocaine        | 69,401 75,414 79,616 74,007 68,696 54,576 46,737 46,879 39,708                                  | 33,846    | 30,017
 Marijuana/hashish         | 285,193 303,649 313,521 317,307 359,157 373,257 357,952 351,896 317,739                         | 288,917   | 247,461
 Stimulants                | 143,525 172,778 164,148 151,886 131,739 120,084 119,190 117,432 127,268                         | 141,155   | 144,427
 Methamphetamine           | 124,500 154,057 155,987 143,338 122,797 111,769 108,894 107,242 117,594                         | 131,270   | 135,264
 Other amphetamines        | 18,010 17,723 6,941 6,620 6,910 7,285 9,054 8,632 8,637                                         | 9,006     | 8,395
 Other stimulants          | 1,015 998 1,220 1,928 2,032 1,030 1,242 1,558 1,037                                             | 879       | 768
 Other drugs               | 28,270 28,703 28,764 29,892 37,034 43,614 47,593 46,601 42,191                                  | 37,818    | 33,511
 Tranquilizers             | 8,169 8,712 10,302 11,672 13,497 15,615 17,242 19,217 18,066                                    | 15,916    | 15,106
-- More  --
```

Then, I altered the data in row 1 (previously row 2), so that "Total" would be in the `primary_substance` column, instead. This can be seen in the final row of `admissions_by_number_staging`:

```
opioid_data=# update admissions_by_number_staging
opioid_data-# set primary_substance = 'Total',
opioid_data-# combined_years = substring(combined_years from 'Total (.*)')
opioid_data-# where combined_years like 'Total%';
UPDATE 1
opioid_data=# select * from admissions_by_number_staging;
     primary_substance      |                                      combined_years                                       | year_2013 | year_2014
----------------------------+-------------------------------------------------------------------------------------------+-----------+-----------
 Alcohol                    | 729,366 746,057 780,815 806,323 860,427 855,294 781,692 756,533 707,863                   | 647,989   | 585,024
 Alcohol only               | 402,999 412,323 433,612 448,707 483,494 479,134 431,490 416,950 391,899                   | 367,484   | 327,694
 Alcohol w/secondary drug   | 326,367 333,734 347,203 357,616 376,933 376,160 350,202 339,583 315,964                   | 280,505   | 257,330
 Opiates                    | 323,409 332,179 353,671 364,629 408,578 434,536 436,473 478,621 479,709                   | 494,247   | 489,680
 Heroin                     | 262,518 260,759 268,443 263,118 282,724 287,783 267,572 282,841 299,674                   | 333,250   | 357,293
 Other opiates/synthetics   | 60,891 71,420 85,228 101,511 125,854 146,753 168,901 195,780 180,035                      | 160,997   | 132,387
 Non-RX methadone           | 3,157 4,133 5,051 5,876 6,488 6,385 6,490 6,860 5,978                                     | 5,095     | 4,212
 Other opiates/synthetics   | 57,734 67,287 80,177 95,635 119,366 140,368 162,411 188,920 174,057                       | 155,902   | 128,175
 Cocaine                    | 248,492 268,573 278,258 260,849 239,581 193,269 158,960 151,910 125,995                   | 105,392   | 87,510
 Smoked cocaine             | 179,091 193,159 198,642 186,842 170,885 138,693 112,223 105,031 86,287                    | 71,546    | 57,493
 Non-smoked cocaine         | 69,401 75,414 79,616 74,007 68,696 54,576 46,737 46,879 39,708                            | 33,846    | 30,017
 Marijuana/hashish          | 285,193 303,649 313,521 317,307 359,157 373,257 357,952 351,896 317,739                   | 288,917   | 247,461
 Stimulants                 | 143,525 172,778 164,148 151,886 131,739 120,084 119,190 117,432 127,268                   | 141,155   | 144,427
 Methamphetamine            | 124,500 154,057 155,987 143,338 122,797 111,769 108,894 107,242 117,594                   | 131,270   | 135,264
 Other amphetamines         | 18,010 17,723 6,941 6,620 6,910 7,285 9,054 8,632 8,637                                   | 9,006     | 8,395
 Other stimulants           | 1,015 998 1,220 1,928 2,032 1,030 1,242 1,558 1,037                                       | 879       | 768
 Other drugs                | 28,270 28,703 28,764 29,892 37,034 43,614 47,593 46,601 42,191                            | 37,818    | 33,511
 Tranquilizers              | 8,169 8,712 10,302 11,672 13,497 15,615 17,242 19,217 18,066                              | 15,916    | 15,106
 Benzodiazepines            | 7,500 8,163 9,767 11,128 12,967 15,048 16,716 18,781 17,664                               | 15,600    | 14,851
 Other tranquilizers        | 669 549 535 544 530 567 526 436 402                                                       | 316       | 255
 Sedatives/hypnotics        | 4,179 4,513 4,149 4,601 4,973 5,335 4,439 3,971 3,456                                     | 3,354     | 2,821
 Barbiturates               | 1,294 1,402 1,053 1,095 1,204 1,343 1,402 947 773                                         | 1,004     | 1,119
 Other sedatives/hypnotics  | 2,885 3,111 3,096 3,506 3,769 3,992 3,037 3,024 2,683                                     | 2,350     | 1,702
 Hallucinogens              | 2,290 2,046 1,645 1,655 1,905 1,874 1,782 1,991 2,141                                     | 2,157     | 1,864
 PCP                        | 3,236 2,888 2,869 3,207 4,059 4,436 4,739 5,749 5,847                                     | 5,324     | 4,910
 Inhalants                  | 1,191 1,372 1,126 1,140 1,383 1,611 1,540 1,256 1,147                                     | 940       | 791
 Over-the-counter           | 828 811 991 923 1,157 1,722 2,286 1,315 1,076                                             | 1,060     | 911
 Other                      | 8,377 8,361 7,682 6,694 10,060 13,021 15,565 13,102 10,458                                | 9,067     | 7,108
 None reported              | 50,214 43,978 43,497 43,853 39,775 32,120 26,969 25,130 25,205                            | 21,029    | 26,745
 Total                      | 1,808,469 1,895,917 1,962,674 1,974,739 2,076,291 2,052,174 1,928,829 1,928,123 1,825,970 | 1,736,547 | 1,614,358
(30 rows)
```

Finally, I inserted the data into `admissions_by_number` while removing commas, casting, and retrieving only the substring after the last ‘ ‘ char in the `combined_years` column.

```
opioid_data=# insert into admissions_by_number(primary_substance, year_2012, year_2013, year_2014)
opioid_data-# select primary_substance,
opioid_data-# cast(reverse(substring(reverse(replace(combined_years, ',', '')) from '^[^ ]*')) as int)
opioid_data-# ,cast(replace(year_2013, ',', '') as int),
opioid_data-# cast(replace(year_2014, ',', '') as int)
opioid_data-# from admissions_by_number_staging;
INSERT 0 30
```

I only chose to acquire the substring for the year 2012 because of the time frame I selected.

The final `admissions_by_number` table can be seen below:

```
opioid_data=# select * from admissions_by_number;
     primary_substance      | year_2012 | year_2013 | year_2014
----------------------------+-----------+-----------+-----------
 Alcohol                    |    707863 |    647989 |    585024
 Alcohol only               |    391899 |    367484 |    327694
 Alcohol w/secondary drug   |    315964 |    280505 |    257330
 Opiates                    |    479709 |    494247 |    489680
 Heroin                     |    299674 |    333250 |    357293
 Other opiates/synthetics   |    180035 |    160997 |    132387
 Non-RX methadone           |      5978 |      5095 |      4212
 Other opiates/synthetics   |    174057 |    155902 |    128175
 Cocaine                    |    125995 |    105392 |     87510
 Smoked cocaine             |     86287 |     71546 |     57493
 Non-smoked cocaine         |     39708 |     33846 |     30017
 Marijuana/hashish          |    317739 |    288917 |    247461
 Stimulants                 |    127268 |    141155 |    144427
 Methamphetamine            |    117594 |    131270 |    135264
 Other amphetamines         |      8637 |      9006 |      8395
 Other stimulants           |      1037 |       879 |       768
 Other drugs                |     42191 |     37818 |     33511
 Tranquilizers              |     18066 |     15916 |     15106
 Benzodiazepines            |     17664 |     15600 |     14851
 Other tranquilizers        |       402 |       316 |       255
 Sedatives/hypnotics        |      3456 |      3354 |      2821
 Barbiturates               |       773 |      1004 |      1119
 Other sedatives/hypnotics  |      2683 |      2350 |      1702
 Hallucinogens              |      2141 |      2157 |      1864
 PCP                        |      5847 |      5324 |      4910
 Inhalants                  |      1147 |       940 |       791
 Over-the-counter           |      1076 |      1060 |       911
 Other                      |     10458 |      9067 |      7108
 None reported              |     25205 |     21029 |     26745
 Total                      |   1825970 |   1736547 |   1614358
(30 rows)
```

4. admissions_by_percentage

As mentioned above, the `admissions_by_percentage` table is in the same format as the `admissions_by_number` table, with percentages replacing raw numbers. The CSV can be seen below:

	Table 1.1b. Admissions aged 12 and older, by primary substance of abuse: Percent distribution, 2004-2014		
Primary substance	2004 2005 2006 2007 2008 2009 2010 2011 2012	2013	2014
	Total 100.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0	100	100
Alcohol	40.3 39.4 39.8 40.8 41.4 41.7 40.5 39.2 38.8	37.3	36.2
Alcohol only	22.3 21.7 22.1 22.7 23.3 23.3 22.4 21.6 21.5	21.2	20.3
Alcohol w/secondary drug	18.0 17.6 17.7 18.1 18.2 18.3 18.2 17.6 17.3	16.2	15.9
Opiates	17.9 17.5 18.0 18.5 19.7 21.2 22.6 24.8 26.3	28.5	30.3
Heroin	14.5 13.8 13.7 13.3 13.6 14.0 13.9 14.7 16.4	19.2	22.1
Other opiates/synthetics	3.4 3.8 4.3 5.1 6.1 7.2 8.8 10.2 9.9	9.3	8.2
Non-RX methadone	0.2 0.2 0.3 0.3 0.3 0.3 0.3 0.4 0.3	0.3	0.3
Other opiates/synthetics	3.2 3.5 4.1 4.8 5.7 6.8 8.4 9.8 9.5	9	7.9
Cocaine	13.7 14.2 14.2 13.2 11.5 9.4 8.2 7.9 6.9	6.1	5.4
Smoked cocaine	9.9 10.2 10.1 9.5 8.2 6.8 5.8 5.4 4.7	4.1	3.6
Non-smoked cocaine	3.8 4.0 4.1 3.7 3.3 2.7 2.4 2.4 2.2	1.9	1.9
Marijuana/hashish	15.8 16.0 16.0 16.1 17.3 18.2 18.6 18.3 17.4	16.6	15.3
Stimulants	7.9 9.1 8.4 7.7 6.3 5.9 6.2 6.1 7.0	8.1	8.9
Methamphetamine	6.9 8.1 7.9 7.3 5.9 5.4 5.6 5.6 6.4	7.6	8.4
	1		
Other amphetamines	1.0 0.9 0.4 0.3 0.3 0.4 0.5 0.4 0.5	0.5	0.5
Other stimulants	0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1	0.1	*
Other drugs	1.6 1.5 1.5 1.5 1.8 2.1 2.5 2.4 2.3	2.2	2.1
Tranquilizers	0.5 0.5 0.5 0.6 0.7 0.8 0.9 1.0 1.0	0.9	0.9
Benzodiazepines	0.4 0.4 0.5 0.6 0.6 0.7 0.9 1.0 1.0	0.9	0.9

In addition to the existing concerns of the previous dataset, this dataset also contains `'*'` values. This is because, according to the data summary, "Admissions for which values were not collected, unknown, or missing are excluded from the percentage base (denominator)."[1] The `admissions_by_percentage_staging` table is as follows:

```
opioid_data=# select * from admissions_by_percentage_staging;
     primary_substance     |                      combined_years                       | year_2013 | year_2014
---------------------------+-----------------------------------------------------------+-----------+-----------
 Primary substance         | 2004 2005 2006 2007 2008 2009 2010 2011 2012              | 2013      | 2014
                           | Total 100.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0 | 100.0     | 100.0
 Alcohol                   | 40.3 39.4 39.8 40.8 41.4 41.7 40.5 39.2 38.8              | 37.3      | 36.2
 Alcohol only              | 22.3 21.7 22.1 22.7 23.3 23.3 22.4 21.6 21.5              | 21.2      | 20.3
 Alcohol w/secondary drug  | 18.0 17.6 17.7 18.1 18.2 18.3 18.2 17.6 17.3              | 16.2      | 15.9
 Opiates                   | 17.9 17.5 18.0 18.5 19.7 21.2 22.6 24.8 26.3              | 28.5      | 30.3
 Heroin                    | 14.5 13.8 13.7 13.3 13.6 14.0 13.9 14.7 16.4              | 19.2      | 22.1
 Other opiates/synthetics  | 3.4 3.8 4.3 5.1 6.1 7.2 8.8 10.2 9.9                      | 9.3       | 8.2
 Non-RX methadone          | 0.2 0.2 0.3 0.3 0.3 0.3 0.3 0.4 0.3                       | 0.3       | 0.3
 Other opiates/synthetics  | 3.2 3.5 4.1 4.8 5.7 6.8 8.4 9.8 9.5                       | 9.0       | 7.9
 Cocaine                   | 13.7 14.2 14.2 13.2 11.5 9.4 8.2 7.9 6.9                  | 6.1       | 5.4
 Smoked cocaine            | 9.9 10.2 10.1 9.5 8.2 6.8 5.8 5.4 4.7                     | 4.1       | 3.6
 Non-smoked cocaine        | 3.8 4.0 4.1 3.7 3.3 2.7 2.4 2.4 2.2                       | 1.9       | 1.9
 Marijuana/hashish         | 15.8 16.0 16.0 16.1 17.3 18.2 18.6 18.3 17.4              | 16.6      | 15.3
 Stimulants                | 7.9 9.1 8.4 7.7 6.3 5.9 6.2 6.1 7.0                       | 8.1       | 8.9
 Methamphetamine           | 6.9 8.1 7.9 7.3 5.9 5.4 5.6 5.6 6.4                       | 7.6       | 8.4
                           | 1                                                         |           |
 Other amphetamines        | 1.0 0.9 0.4 0.3 0.3 0.4 0.5 0.4 0.5                       | 0.5       | 0.5
 Other stimulants          | 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1                       | 0.1       | *
 Other drugs               | 1.6 1.5 1.5 1.5 1.8 2.1 2.5 2.4 2.3                       | 2.2       | 2.1
 Tranquilizers             | 0.5 0.5 0.5 0.6 0.7 0.8 0.9 1.0 1.0                       | 0.9       | 0.9
 Benzodiazepines           | 0.4 0.4 0.5 0.6 0.6 0.7 0.9 1.0 1.0                       | 0.9       | 0.9
 Other tranquilizers       | *  *  *  *  *  *  *  *  *                                 | *         | *
 Sedatives/hypnotics       | 0.2 0.2 0.2 0.2 0.2 0.3 0.2 0.2 0.2                       | 0.2       | 0.2
 Barbiturates              | 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.0  *                        | 0.1       | 0.1
 Other sedatives/hypnotics | 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.1                       | 0.1       | 0.1
 Hallucinogens             | 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1                       | 0.1       | 0.1
 PCP                       | 0.2 0.2 0.1 0.2 0.2 0.2 0.2 0.3 0.3                       | 0.3       | 0.3
 Inhalants                 | 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1                       | 0.1       | *
 Over-the-counter          | *  *   0.1  *   0.1   0.1   0.1   0.1   0.1               | 0.1       | 0.1
 Other                     | 0.5 0.4 0.4 0.3 0.5 0.6 0.8 0.7 0.6                       | 0.5       | 0.4
 None reported             | 2.8 2.3 2.2 2.2 1.9 1.6 1.4 1.3 1.4                       | 1.2       | 1.7
(32 rows)
```

[1] SOURCE: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, Treatment Episode Data Set (TEDS). Data received through 02.01.16. https://data.world/samhsa/admissions-substance-of-abuse/workspace/project-summary?agentid=samhsa&datasetid=admissions-substance-of-abuse.

Similar data cleaning measures were used on this dataset.

```
opioid_data=# delete from admissions_by_percentage_staging
opioid_data-# where primary_substance = 'Primary substance' or combined_years = '1';
DELETE 2
opioid_data=# update admissions_by_percentage_staging
opioid_data-# set primary_substance = 'Total',
opioid_data-# combined_years = substring(combined_years from 'Total (.*)')
opioid_data-# where combined_years like 'Total%';
UPDATE 1
opioid_data=# select * from admissions_by_percentage_staging;
     primary_substance     |                    combined_years                    | year_2013 | year_2014
---------------------------+------------------------------------------------------+-----------+-----------
 Alcohol                   | 40.3 39.4 39.8 40.8 41.4 41.7 40.5 39.2 38.8         | 37.3      | 36.2
 Alcohol only              | 22.3 21.7 22.1 22.7 23.3 23.3 22.4 21.6 21.5         | 21.2      | 20.3
 Alcohol w/secondary drug  | 18.0 17.6 17.7 18.1 18.2 18.3 18.2 17.6 17.3         | 16.2      | 15.9
 Opiates                   | 17.9 17.5 18.0 18.5 19.7 21.2 22.6 24.8 26.3         | 28.5      | 30.3
 Heroin                    | 14.5 13.8 13.7 13.3 13.6 14.0 13.9 14.7 16.4         | 19.2      | 22.1
 Other opiates/synthetics  | 3.4 3.8 4.3 5.1 6.1 7.2 8.8 10.2 9.9                 | 9.3       | 8.2
 Non-RX methadone          | 0.2 0.2 0.3 0.3 0.3 0.3 0.3 0.4 0.3                  | 0.3       | 0.3
 Other opiates/synthetics  | 3.2 3.5 4.1 4.8 5.7 6.8 8.4 9.8 9.5                  | 9.0       | 7.9
 Cocaine                   | 13.7 14.2 14.2 13.2 11.5 9.4 8.2 7.9 6.9             | 6.1       | 5.4
 Smoked cocaine            | 9.9 10.2 10.1 9.5 8.2 6.8 5.8 5.4 4.7                | 4.1       | 3.6
 Non-smoked cocaine        | 3.8 4.0 4.1 3.7 3.3 2.7 2.4 2.4 2.2                  | 1.9       | 1.9
 Marijuana/hashish         | 15.8 16.0 16.0 16.1 17.3 18.2 18.6 18.3 17.4         | 16.6      | 15.3
 Stimulants                | 7.9 9.1 8.4 7.7 6.3 5.9 6.2 6.1 7.0                  | 8.1       | 8.9
 Methamphetamine           | 6.9 8.1 7.9 7.3 5.9 5.4 5.6 5.6 6.4                  | 7.6       | 8.4
 Other amphetamines        | 1.0 0.9 0.4 0.3 0.3 0.4 0.5 0.4 0.5                  | 0.5       | 0.5
 Other stimulants          | 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1                  | 0.1       | *
 Other drugs               | 1.6 1.5 1.5 1.5 1.8 2.1 2.5 2.4 2.3                  | 2.2       | 2.1
 Tranquilizers             | 0.5 0.5 0.5 0.6 0.7 0.8 0.9 1.0 1.0                  | 0.9       | 0.9
 Benzodiazepines           | 0.4 0.4 0.5 0.6 0.6 0.7 0.9 1.0 1.0                  | 0.9       | 0.9
 Other tranquilizers       | *  *  *  *  *  *  *  *  *                            | *         | *
 Sedatives/hypnotics       | 0.2 0.2 0.2 0.2 0.2 0.3 0.2 0.2 0.2                  | 0.2       | 0.2
 Barbiturates              | 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.0  *                   | 0.1       | 0.1
 Other sedatives/hypnotics | 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.2 0.1                  | 0.1       | 0.1
 Hallucinogens             | 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1                  | 0.1       | 0.1
 PCP                       | 0.2 0.2 0.1 0.2 0.2 0.2 0.2 0.3 0.3                  | 0.3       | 0.3
 Inhalants                 | 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1                  | 0.1       | *
 Over-the-counter          | *  *   0.1  *   0.1   0.1   0.1   0.1   0.1         | 0.1       | 0.1
 Other                     | 0.5 0.4 0.4 0.3 0.5 0.6 0.8 0.7 0.6                  | 0.5       | 0.4
 None reported             | 2.8 2.3 2.2 2.2 1.9 1.6 1.4 1.3 1.4                  | 1.2       | 1.7
 Total                     | 100.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0 100.0 | 100.0     | 100.0
(30 rows)
```

When creating the `admissions_by_percentage` table, I did not account for the value `'100.0'` when setting the type to `decimal(3, 1)`. Consequently, I had to alter the table:

```
opioid_data=# alter table admissions_by_percentage alter column year_2012 type decimal(5,1), alter column year_2013 type decimal(
5,1), alter column year_2014 type decimal(5,1);
ALTER TABLE
```

Finally, I loaded the data into `admissions_by_percentage`:

```
opioid_data=# insert into admissions_by_percentage(primary_substance, year_2012, year_2013, year_2014)
opioid_data-# select primary_substance,
opioid_data-# cast(nullif(reverse(substring(reverse(replace(combined_years, '*', '')) from '^[^ ]*')), '') as decimal(5,1)),
opioid_data-# cast(nullif(replace(year_2013, '*', ''), '') as decimal(5,1)),
opioid_data-# cast(nullif(replace(year_2014, '*', ''), '') as decimal(5,1))
opioid_data-# from admissions_by_percentage_staging;
INSERT 0 30
opioid_data=# select * from admissions_by_percentage;
     primary_substance     | year_2012 | year_2013 | year_2014
---------------------------+-----------+-----------+-----------
 Alcohol                   |      38.8 |      37.3 |      36.2
 Alcohol only              |      21.5 |      21.2 |      20.3
 Alcohol w/secondary drug  |      17.3 |      16.2 |      15.9
 Opiates                   |      26.3 |      28.5 |      30.3
 Heroin                    |      16.4 |      19.2 |      22.1
 Other opiates/synthetics  |       9.9 |       9.3 |       8.2
 Non-RX methadone          |       0.3 |       0.3 |       0.3
 Other opiates/synthetics  |       9.5 |       9.0 |       7.9
 Cocaine                   |       6.9 |       6.1 |       5.4
 Smoked cocaine            |       4.7 |       4.1 |       3.6
 Non-smoked cocaine        |       2.2 |       1.9 |       1.9
 Marijuana/hashish         |      17.4 |      16.6 |      15.3
 Stimulants                |       7.0 |       8.1 |       8.9
 Methamphetamine           |       6.4 |       7.6 |       8.4
 Other amphetamines        |       0.5 |       0.5 |       0.5
 Other stimulants          |       0.1 |       0.1 |
 Other drugs               |       2.3 |       2.2 |       2.1
 Tranquilizers             |       1.0 |       0.9 |       0.9
 Benzodiazepines           |       1.0 |       0.9 |       0.9
 Other tranquilizers       |           |           |
 Sedatives/hypnotics       |       0.2 |       0.2 |       0.2
 Barbiturates              |           |       0.1 |       0.1
 Other sedatives/hypnotics |       0.1 |       0.1 |       0.1
 Hallucinogens             |       0.1 |       0.1 |       0.1
 PCP                       |       0.3 |       0.3 |       0.3
 Inhalants                 |       0.1 |       0.1 |
 Over-the-counter          |       0.1 |       0.1 |       0.1
 Other                     |       0.6 |       0.5 |       0.4
 None reported             |       1.4 |       1.2 |       1.7
 Total                     |     100.0 |     100.0 |     100.0
(30 rows)
```

V. Data Integration

Initially, I wanted to combine several different datasets all into one; however, I realized that because different data goes by different "keys," this would not be possible. So, I combined some datasets but not others using `JOIN` statements.

```
opioid_data=# create table admissions_data as
opioid_data-# select
opioid_data-# an.primary_substance as primary_substance,
opioid_data-# an.year_2012 as number_2012,
opioid_data-# an.year_2013 as number_2013,
opioid_data-# an.year_2014 as number_2014,
opioid_data-# ap.year_2012 as percentage_2012,
opioid_data-# ap.year_2013 as percentage_2013,
opioid_data-# ap.year_2014 as percentage_2014
opioid_data-# from admissions_by_number an
opioid_data-# join admissions_by_percentage ap
opioid_data-# on an.primary_substance = ap.primary_substance;
SELECT 32
```

```
opioid_data=# select * from admissions_data;
     primary_substance     | number_2012 | number_2013 | number_2014 | percentage_2012 | percentage_2013 | percentage_2014
---------------------------+-------------+-------------+-------------+-----------------+-----------------+-----------------
 Alcohol                   |      707863 |      647989 |      585024 |            38.8 |            37.3 |            36.2
 Alcohol only              |      391899 |      367484 |      327694 |            21.5 |            21.2 |            20.3
 Alcohol w/secondary drug  |      315964 |      280505 |      257330 |            17.3 |            16.2 |            15.9
 Opiates                   |      479709 |      494247 |      489680 |            26.3 |            28.5 |            30.3
 Heroin                    |      299674 |      333250 |      357293 |            16.4 |            19.2 |            22.1
 Other opiates/synthetics  |      180035 |      160997 |      132387 |             9.5 |             9.0 |             7.9
 Other opiates/synthetics  |      180035 |      160997 |      132387 |             9.9 |             9.3 |             8.2
 Non-RX methadone          |        5978 |        5095 |        4212 |             0.3 |             0.3 |             0.3
 Other opiates/synthetics  |      174057 |      155902 |      128175 |             9.5 |             9.0 |             7.9
 Other opiates/synthetics  |      174057 |      155902 |      128175 |             9.9 |             9.3 |             8.2
 Cocaine                   |      125995 |      105392 |       87510 |             6.9 |             6.1 |             5.4
 Smoked cocaine            |       86287 |       71546 |       57493 |             4.7 |             4.1 |             3.6
 Non-smoked cocaine        |       39708 |       33846 |       30017 |             2.2 |             1.9 |             1.9
 Marijuana/hashish         |      317739 |      288917 |      247461 |            17.4 |            16.6 |            15.3
 Stimulants                |      127268 |      141155 |      144427 |             7.0 |             8.1 |             8.9
 Methamphetamine           |      117594 |      131270 |      135264 |             6.4 |             7.6 |             8.4
 Other amphetamines        |        8637 |        9006 |        8395 |             0.5 |             0.5 |             0.5
 Other stimulants          |        1037 |         879 |         768 |             0.1 |             0.1 |
 Other drugs               |       42191 |       37818 |       33511 |             2.3 |             2.2 |             2.1
 Tranquilizers             |       18066 |       15916 |       15106 |             1.0 |             0.9 |             0.9
 Benzodiazepines           |       17664 |       15600 |       14851 |             1.0 |             0.9 |             0.9
 Other tranquilizers       |         402 |         316 |         255 |                 |                 |
 Sedatives/hypnotics       |        3456 |        3354 |        2821 |             0.2 |             0.2 |             0.2
 Barbiturates              |         773 |        1004 |        1119 |                 |             0.1 |             0.1
 Other sedatives/hypnotics |        2683 |        2350 |        1702 |             0.1 |             0.1 |             0.1
 Hallucinogens             |        2141 |        2157 |        1864 |             0.1 |             0.1 |             0.1
 PCP                       |        5847 |        5324 |        4910 |             0.3 |             0.3 |             0.3
 Inhalants                 |        1147 |         940 |         791 |             0.1 |             0.1 |
 Over-the-counter          |        1076 |        1060 |         911 |             0.1 |             0.1 |             0.1
 Other                     |       10458 |        9067 |        7108 |             0.6 |             0.5 |             0.4
 None reported             |       25205 |       21029 |       26745 |             1.4 |             1.2 |             1.7
 Total                     |     1825970 |     1736547 |     1614358 |           100.0 |           100.0 |           100.0
(32 rows)
```

From the number of rows, I realized that there were two extra rows (expected no. of rows = 30). There were two instances of `'Other opiates/synthetics'` in each table joined in `admissions_data`, resulting in duplicates. The following code removed these duplicates:

```
opioid_data=# with duplicates as(
opioid_data(# select
opioid_data(# ctid, primary_substance, number_2012, number_2013, number_2014, percentage_2012, percentage_2013, percentage_2014,
opioid_data(# row_number() over (partition by number_2012 order by ctid) as row_num
opioid_data(# from admissions_data)
opioid_data-# delete from admissions_data where ctid in (
opioid_data(# select ctid from duplicates where row_num > 1);
DELETE 2
```

Note that I partitioned by `number_2012` rather than `primary_substance` due to the double instances of `'Other opiates/synthetics'`. In this case, at 1 decimal places, percentages were too imprecise to use `percentage_2012` or any other percentage data.

`opioid_deaths` and `nonmed_opioid_use` could not be joined into a table, because the former measured by year from 2012 to 2014, while the latter gave general data for the period of 2012 to 2014.

VI. Exploratory Data Analysis (EDA)

In order to use `Pandas` to do EDA, I had to export by database to CSVs.

```
opioid_data=# \copy admissions_data to 'C:\Users\user\Documents\CS210\admissions.c
sv' delimiter ',' csv header;
COPY 30
```

I removed some data from the `opioid_deaths` table and `nonmed_opioid_use` data before exporting, as I did not want to visualize unnecessary data.

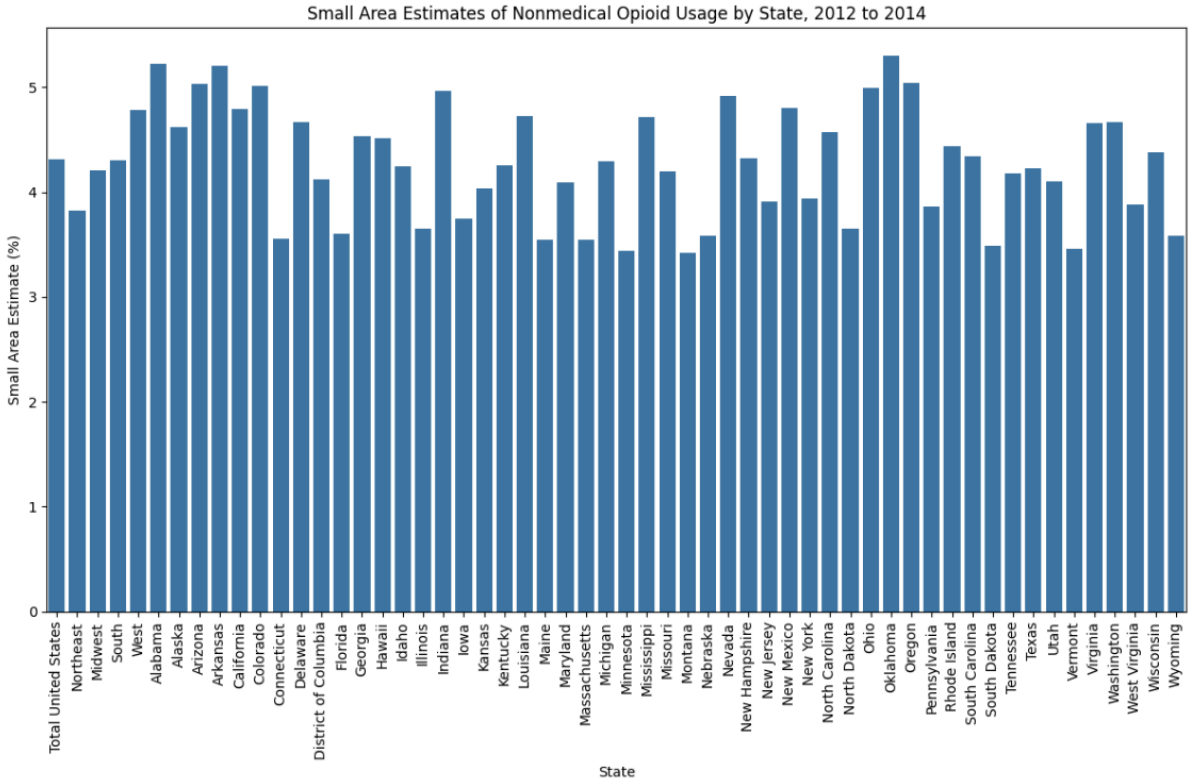
```
opioid_data=# \d opioid_deaths
                      Table "public.opioid_deaths"
       Column        |          Type          | Collation | Nullable | Default
---------------------+------------------------+-----------+----------+---------
 state               | character varying(200) |           |          |
 year                | integer                |           |          |
 deaths              | integer                |           |          |
 population          | integer                |           |          |
 crude_rate          | numeric(4,1)           |           |          |
 crude_rate_lower_ci | numeric(4,1)           |           |          |
 crude_rate_upper_ci | numeric(4,1)           |           |          |
 prescription_yearly | integer                |           |          |
Indexes:
    "opioid_deaths_unique_state_year" UNIQUE CONSTRAINT, btree (state, year)


opioid_data=# alter table opioid_deaths
opioid_data-# drop column crude_rate_lower_ci, drop column crude_rate_upper_ci;
ALTER TABLE
opioid_data=# \copy opioid_deaths to 'C:\Users\user\Documents\CS210\opioid_deaths.csv' delimiter ',' csv header;
COPY 816
opioid_data=# \d nonmed_opioid_use
                   Table "public.nonmed_opioid_use"
     Column     |          Type          | Collation | Nullable | Default
----------------+------------------------+-----------+----------+---------
 data_order     | integer                |           |          |
 state          | character varying(200) |           |          |
 substate       | character varying(200) |           |          |
 small_area_est | numeric(5,2)           |           |          |
 lower_ci       | numeric(5,2)           |           |          |
 upper_ci       | numeric(5,2)           |           |          |
 map_group      | integer                |           |          |


opioid_data=# alter table nonmed_opioid_use
opioid_data-# drop column data_order, drop column map_group, drop column lower_ci,
 drop column upper_ci;
ALTER TABLE
opioid_data=# \copy nonmed_opioid_use to 'C:\Users\user\Documents\CS210\nonmed_pain_relievers.csv' delimiter ',' csv header;
COPY 465
```

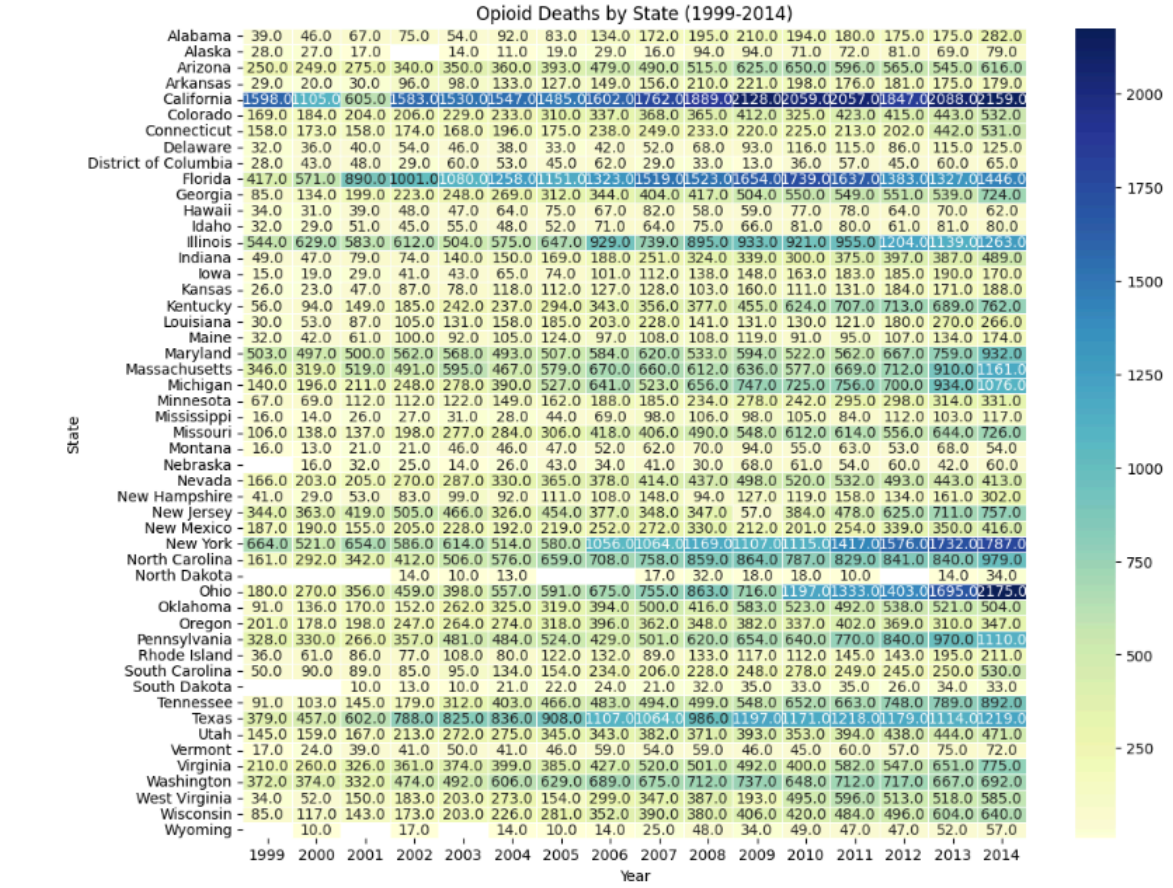
Using Pandas, Matplotlib, and Seaborn, I was able to perform data visualization on the exported CSV files.

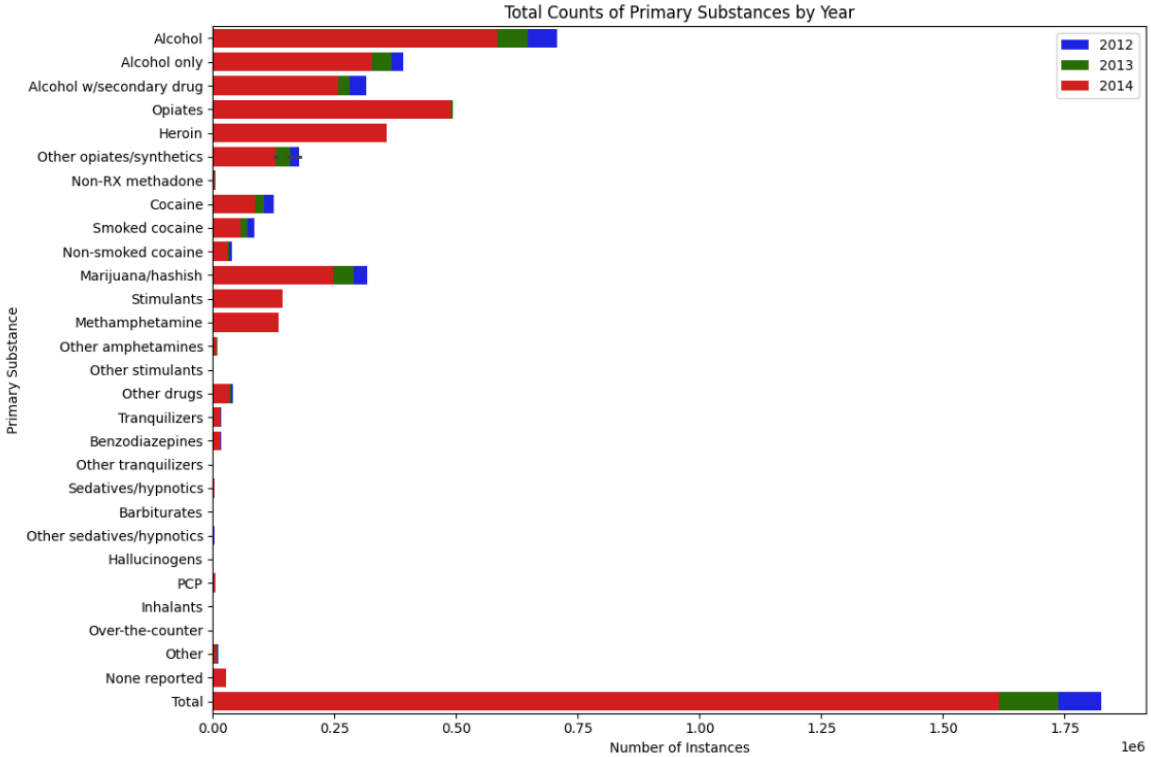
Trends in 2012 to 2014 data show that while opioid-related deaths are steadily increasing, admission into care centers are decreasing. Certain regions are evidently more at-risk of opioid use, such as Western states like California and Oregon, and states in the South and Northeast such as Texas and New York, respectively.

Source code attached.

Below are the visuals:




Opioid Deaths by State (1999-2014)
State
Year
1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Alabama 39.0 46.0 67.0 75.0 54.0 92.0 83.0 134.0 172.0 195.0 210.0 194.0 180.0 175.0 175.0 282.0
Alaska 28.0 27.0 17.0 14.0 11.0 19.0 29.0 16.0 94.0 94.0 71.0 72.0 81.0 69.0 79.0
Arizona 250.0 249.0 275.0 340.0 350.0 360.0 393.0 479.0 490.0 515.0 625.0 650.0 596.0 565.0 545.0 616.0
Arkansas 29.0 20.0 30.0 96.0 98.0 133.0 127.0 149.0 156.0 210.0 221.0 198.0 176.0 181.0 175.0 179.0
California 1598.0 1105.0 605.0 1583.0 1530.0 1547.0 1485.0 1602.0 1762.0 1889.0 2128.0 2059.0 2057.0 1847.0 2088.0 2159.0
Colorado 169.0 184.0 204.0 206.0 229.0 233.0 310.0 337.0 368.0 365.0 412.0 325.0 423.0 415.0 443.0 532.0
Connecticut 158.0 173.0 158.0 174.0 168.0 196.0 175.0 238.0 249.0 233.0 220.0 225.0 213.0 202.0 442.0 531.0
Delaware 32.0 36.0 40.0 54.0 46.0 38.0 33.0 42.0 52.0 68.0 93.0 116.0 115.0 86.0 115.0 125.0
District of Columbia 28.0 43.0 48.0 29.0 60.0 53.0 45.0 62.0 29.0 33.0 13.0 36.0 57.0 45.0 60.0 65.0
Florida 417.0 571.0 890.0 1001.0 1080.0 1258.0 1151.0 1323.0 1519.0 1523.0 1654.0 1739.0 1637.0 1383.0 1327.0 1446.0
Georgia 85.0 134.0 199.0 223.0 248.0 269.0 312.0 344.0 404.0 417.0 504.0 550.0 549.0 551.0 539.0 724.0
Hawaii 34.0 31.0 39.0 48.0 47.0 64.0 75.0 67.0 82.0 58.0 59.0 77.0 78.0 64.0 70.0 62.0
Idaho 32.0 29.0 51.0 45.0 55.0 48.0 52.0 71.0 64.0 75.0 66.0 81.0 80.0 61.0 81.0 80.0
Illinois 544.0 629.0 583.0 612.0 504.0 575.0 647.0 929.0 739.0 895.0 933.0 921.0 955.0 1204.0 1139.0 1263.0
Indiana 49.0 47.0 79.0 74.0 140.0 150.0 169.0 188.0 251.0 324.0 339.0 300.0 375.0 397.0 387.0 489.0
Iowa 15.0 19.0 29.0 41.0 43.0 65.0 74.0 101.0 112.0 138.0 148.0 163.0 183.0 185.0 190.0 170.0
Kansas 26.0 23.0 47.0 87.0 78.0 118.0 112.0 127.0 128.0 103.0 160.0 111.0 131.0 184.0 171.0 188.0
Kentucky 56.0 94.0 149.0 185.0 242.0 237.0 294.0 343.0 356.0 377.0 455.0 624.0 707.0 713.0 689.0 762.0
Louisiana 30.0 53.0 87.0 105.0 131.0 158.0 185.0 203.0 228.0 141.0 131.0 130.0 121.0 180.0 270.0 266.0
Maine 32.0 42.0 61.0 100.0 92.0 105.0 124.0 97.0 108.0 108.0 119.0 91.0 95.0 107.0 134.0 174.0
Maryland 503.0 497.0 500.0 562.0 568.0 493.0 507.0 584.0 620.0 533.0 594.0 522.0 562.0 667.0 759.0 932.0
Massachusetts 346.0 319.0 519.0 491.0 595.0 467.0 579.0 670.0 660.0 612.0 636.0 577.0 669.0 712.0 910.0 1161.0
Michigan 140.0 196.0 211.0 248.0 278.0 390.0 527.0 641.0 523.0 656.0 747.0 725.0 756.0 700.0 934.0 1076.0
Minnesota 67.0 69.0 112.0 112.0 122.0 149.0 162.0 188.0 185.0 234.0 278.0 242.0 295.0 298.0 314.0 331.0
Mississippi 16.0 14.0 26.0 27.0 31.0 28.0 44.0 69.0 98.0 106.0 98.0 105.0 84.0 112.0 103.0 117.0
Missouri 106.0 138.0 137.0 198.0 277.0 284.0 306.0 418.0 406.0 490.0 548.0 612.0 614.0 556.0 644.0 726.0
Montana 16.0 13.0 21.0 21.0 46.0 46.0 47.0 52.0 62.0 70.0 94.0 55.0 63.0 53.0 68.0 54.0
Nebraska 16.0 32.0 25.0 14.0 26.0 43.0 34.0 41.0 30.0 68.0 61.0 54.0 60.0 42.0 60.0
Nevada 166.0 203.0 205.0 270.0 287.0 330.0 365.0 378.0 414.0 437.0 498.0 520.0 532.0 493.0 443.0 413.0
New Hampshire 41.0 29.0 53.0 83.0 99.0 92.0 111.0 108.0 148.0 94.0 127.0 119.0 158.0 134.0 161.0 302.0
New Jersey 344.0 363.0 419.0 505.0 466.0 326.0 454.0 377.0 348.0 347.0 57.0 384.0 478.0 625.0 711.0 757.0
New Mexico 187.0 190.0 155.0 205.0 228.0 192.0 219.0 252.0 272.0 330.0 212.0 201.0 254.0 339.0 350.0 416.0
New York 664.0 521.0 654.0 586.0 614.0 514.0 580.0 1056.0 1064.0 1169.0 1107.0 1115.0 1417.0 1576.0 1732.0 1787.0
North Carolina 161.0 292.0 342.0 412.0 506.0 576.0 659.0 708.0 758.0 859.0 864.0 787.0 829.0 841.0 840.0 979.0
North Dakota 14.0 10.0 13.0 17.0 32.0 18.0 18.0 10.0 14.0 34.0
Ohio 180.0 270.0 356.0 459.0 398.0 557.0 591.0 675.0 755.0 863.0 716.0 1197.0 1333.0 1403.0 1695.0 2175.0
Oklahoma 91.0 136.0 170.0 152.0 262.0 325.0 319.0 394.0 500.0 416.0 583.0 523.0 492.0 538.0 521.0 504.0
Oregon 201.0 178.0 198.0 247.0 264.0 274.0 318.0 396.0 362.0 348.0 382.0 337.0 402.0 369.0 310.0 347.0
Pennsylvania 328.0 330.0 266.0 357.0 481.0 484.0 524.0 429.0 501.0 620.0 654.0 640.0 770.0 840.0 970.0 1110.0
Rhode Island 36.0 61.0 86.0 77.0 108.0 80.0 122.0 132.0 89.0 133.0 117.0 112.0 145.0 143.0 195.0 211.0
South Carolina 50.0 90.0 89.0 85.0 95.0 134.0 154.0 234.0 206.0 228.0 248.0 278.0 249.0 245.0 250.0 530.0
South Dakota 10.0 13.0 10.0 21.0 22.0 24.0 21.0 32.0 35.0 33.0 35.0 26.0 34.0 33.0
Tennessee 91.0 103.0 145.0 179.0 312.0 403.0 466.0 483.0 494.0 499.0 548.0 652.0 663.0 748.0 789.0 892.0
Texas 379.0 457.0 602.0 788.0 825.0 836.0 908.0 1107.0 1064.0 986.0 1197.0 1171.0 1218.0 1179.0 1114.0 1219.0
Utah 145.0 159.0 167.0 213.0 272.0 275.0 345.0 343.0 382.0 371.0 393.0 353.0 394.0 438.0 444.0 471.0
Vermont 17.0 24.0 39.0 41.0 50.0 41.0 46.0 59.0 54.0 59.0 46.0 45.0 60.0 57.0 75.0 72.0
Virginia 210.0 260.0 326.0 361.0 374.0 399.0 385.0 427.0 520.0 501.0 492.0 400.0 582.0 547.0 651.0 775.0
Washington 372.0 374.0 332.0 474.0 492.0 606.0 629.0 689.0 675.0 712.0 737.0 648.0 712.0 717.0 667.0 692.0
West Virginia 34.0 52.0 150.0 183.0 203.0 273.0 154.0 299.0 347.0 387.0 193.0 495.0 596.0 513.0 518.0 585.0
Wisconsin 85.0 117.0 143.0 173.0 203.0 226.0 281.0 352.0 390.0 380.0 406.0 420.0 484.0 496.0 604.0 640.0
Wyoming 10.0 17.0 14.0 10.0 14.0 25.0 48.0 34.0 49.0 47.0 47.0 52.0 57.0
2000
1750
1500
1250
1000
750
500
250


Total Counts of Primary Substances by Year
Primary Substance
Number of Instances
1e6
2012
2013
2014
Alcohol
Alcohol only
Alcohol w/secondary drug
Opiates
Heroin
Other opiates/synthetics
Non-RX methadone
Cocaine
Smoked cocaine
Non-smoked cocaine
Marijuana/hashish
Stimulants
Methamphetamine
Other amphetamines
Other stimulants
Other drugs
Tranquilizers
Benzodiazepines
Other tranquilizers
Sedatives/hypnotics
Barbiturates
Other sedatives/hypnotics
Hallucinogens
PCP
Inhalants
Over-the-counter
Other
None reported
Total
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75