Open Dataset Research: Drug Overdose Deaths

User Guide & Analyses

INST 490-0104

Table of Contents

Introduction	**2**
Drug Overdose Deaths	**3**
Provisional Drug Overdose Death Counts	**6**
Accessing the Dataset	6
Accessing the Dataset Information	6
Dashboards & Data Tables	13
Technical Notes	16
Nature & Sources of Data	16
Cause of Death Classification	16
Selection of States/Jurisdictions to Report	16
Percentages of Records Pending Investigation	17
Percent Completeness	17
Drug Specificity	17
Improvements in Data Quality	17
Delayed Reporting Adjustments	19
Differences Between Provisional & Final Data	19
References & Resources	19
Dataset	**20**
Background	20
Format	20
Things to Know	20
What is Provisional Data?	20
Definition of Drug Deaths	21
Context	21
Metrics	21
Appendix	21
Data Cleaning/Scrubbing	**24**
Example Data Analyses	**28**
Power BI Dashboard	28
Tableau Dashboard 1	30
Tableau Dashboard 2	33
Python Analysis	34
Version history	**42**

Introduction

The Open Research Dataset project revolves around focusing on an open dataset of choice, while creating relevant documentation and a dataset information repository. The goal of this project is to make the selected open dataset more understandable for individuals interested in utilizing or researching the data. This will be accomplished by creating a user guide to the data in order to help users understand the content and navigate easier. Additionally, open datasets may require additional cleaning. The proceeding step in the project will be fixing any errors in the data. After cleaning is completed, sample data analyses will be created with the cleaned data, acting as creation guides for users interested in visually handling the dataset.

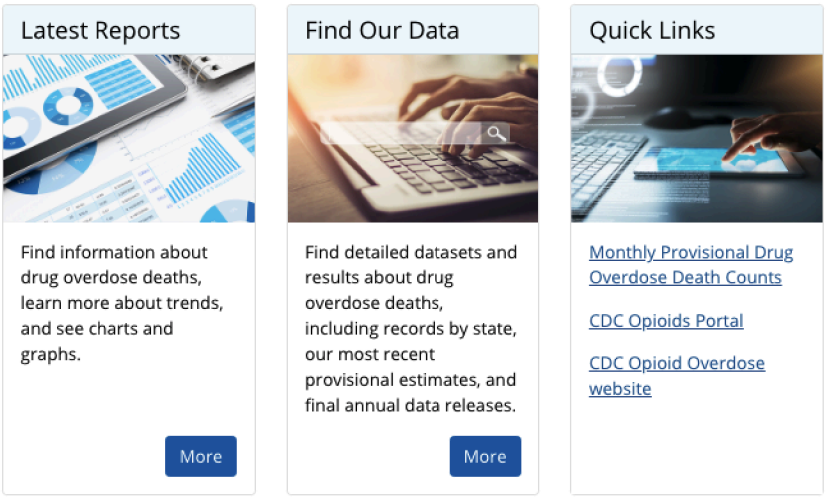
For this project, the topic of concentration will be drug overdose deaths, specifically utilizing the 'VSRR Provisional Drug Overdose Death Counts' (Vital Statistics Rapid Release) dataset. This dataset shows relevant information about the deaths from various forms of drugs in different states throughout the USA. There will be documentation on research about the data source and content, an assessment of the dataset(s) included in the source, analysis of the capabilities of the data source, documentation regarding the data sets and capabilities such as a user guide, and technical data analysis examples of the type of research that could be done with the available data. Access to relevant dataset and project files can be found in the project's GitHub repository.

Drug Overdose Deaths

"The National Vital Statistics System provides timely, reliable data to monitor and track drug overdose deaths." —NVSS (National Vital Statistics System)

The NVSS 'Drug Overdose Deaths' page provides an overview of the purpose for providing and tracking data regarding drug overdoses in the United States. NVSS states that "monitoring deaths from drug overdose helps us understand the epidemic's impact on the U.S. population and tell us how the crisis is evolving," and emphasizes the importance of knowing what substances are being associated with deaths, how they are used, and where in the nation overdose deaths are happening[1].

The drug overdose page has three sections: 'Latest Reports,' 'Find Our Data,' and 'Quick Links.'



***As of 11/16/2022, an additional link has been added to the top of the "Quick Links" section titled, "VSRR Provisional Drug Overdose Death Surveillance"*

Clicking on 'Latest Reports' takes the user to a specific part of the same page, under the section 'Publications,' (*see screenshot below*) where there are links to four data briefs, ranging from information on drug overdose deaths in the United States from 1999-2020, a study on the co-involvement of opioids in drug overdose deaths involving cocaine and psychostimulants, information on the urban-rural differences in drug overdose death rates from 1999-2019, and a report on the early provisional estimates of drug overdoses, suicide, and transportation-related deaths. Following 'Latest Reports' in a horizontal-style navigator are links where users can find

[1] For more information, see the NVSS's Drug Overdose Deaths homepage

relevant publications for the year 2021, 2020, 2019, 2018, 2017, 2016, and 2015 & prior.

***Latest Reports subject to change over time*

***As of 11/20/2024, Latest Reports includes publications from 2022, 2023, and 2024*

The 'Find Our Data' button takes the user to the 'Find Our Data' section of the page (*see screenshot below*), which has subsections of 'Data Releases,' 'Provisional Estimates (Vital Statistics Rapid Release),' 'Online Databases: Mortality,' and 'Tabulated Data.' In the 'Data Releases' subsection are four links, with three linking to cause of death data files for the years 2020, 2019, and 2018 respectively, and another linking to the new expanded 2016 mortality file including the literal text from the cause-of-death Section of the death certificate. The 'Provisional Estimates (Vital Statistics Rapid Release)' sub-section has three links. One linking to the VSRR Provisional Drug Overdose Death Surveillance, one linking to the VSRR Monthly Provisional Drug Overdose Death Counts, which will be the main emphasis of this project document, and one that links to Provisional County-Level Drug Overdose Death Counts. The 'Online Databases: Mortality' subsection has two links, one of which takes the user to CDC Wonder, which is an online query tool that searches and customizes tabulated data in regard to causes of death, and another link which navigates to WISQARS (Web-based Injury Statistics Query and Reporting System), which is an interactive database that has fatal/nonfatal relevant data and information. The 'Tabulated Data' sub-section has two links, one of which navigates to an information page relating to a dataset dealing with drug poisoning deaths by state, and another which navigates to an information page relating to a dataset dealing with drug poisoning deaths by county.

Find Our Data

Data Releases

- 2018 Multiple Cause of Death Data File (1/2020)
- Redacted Death Certificate Literal Text (7/16/2019)
  The new expanded 2016 mortality file includes the literal text from the cause-of-death Section of the death certificate.

Provisional Estimates (Vital Statistics Rapid Release)

- VSRR Provisional Drug Overdose Death Counts
  Provisional (preliminary) estimates for drug overdose deaths based on a current flow of mortality data. Counts for the most recent final annual data are provided for comparison.

Online Databases: Mortality

- CDC WONDER
  Interactive, online query tool that contains searchable and customizable tabulated data about leading causes of death
- CDC Web-based Injury Statistics Query and Reporting System (WISQARS)
  Interactive, online database that provides fatal and nonfatal injury, violent death, and cost of injury data from a variety of trusted sources

Tabulated Data

- NCHS – Drug Poisoning Mortality by State: United States
  Drug poisoning deaths at the U.S. and state level by selected demographic characteristics. Includes age-adjusted death rates for drug poisoning.
- NCHS – Drug Poisoning Mortality by County: United States
  Drug poisoning deaths at the county level by selected demographic characteristics. Includes age-adjusted death rates for drug poisoning from 1999 to 2015.

***Data Releases and Provisional Estimates (Vital Statistics Rapid Release) subject to change over time.*

The ‘Quick Links’ part of the page contains four links. The first link, 'VSRR Provisional Drug Overdose Death Surveillance,' takes the user to a page that provides access to surveillance data on drug overdose deaths based on a current flow of mortality data in the National Vital Statistics System. The second link, ‘Monthly Provisional Drug Overdose Death Counts,’ once again takes the user to a drug overdose dataset documentation page (this will be discussed in more detail following this section). The third link, ‘CDC Opioids Portal,’ navigates to the CDC opioid website, where users can find opioid related data, research, partnerships, support, resources, guidelines, and other information regarding opioids. The fourth link, ‘CDC Opioid Overdose website,’ navigates to the ‘Opioid Overdose’ section of the CDC opioid website, where the user can access fast facts, preventative tips, resources, data, and other information for victims, providers, or people seeking general information.

Additional resources, such as state and local resources, include a death certificate completion guide, and modernization guides relating to the NVSS and death reporting.

It is important to note that the 'Drug Overdose Deaths' page links to a specific dataset documentation page more than once, specifically the 'VSRR Provisional Drug Overdose Death Counts' page. This page, as well as its dataset, will be the main emphasis of this project document.

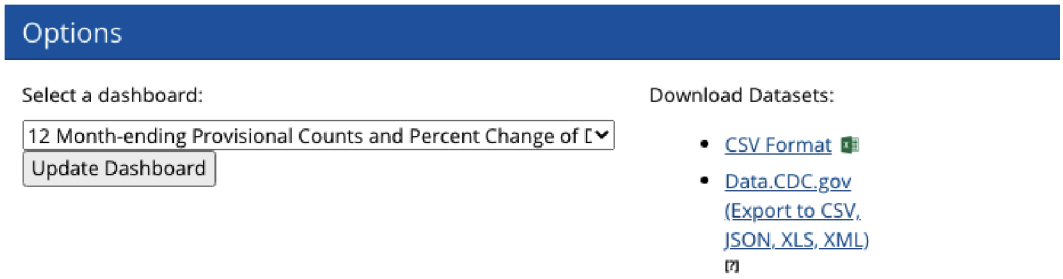
Provisional Drug Overdose Death Counts

"The National Vital Statistics System collects and shares critical information on deaths from drug overdoses, such as what substances were used and where deaths are happening in America." —NVSS (National Vital Statistics System)

NVSS has a program that releases a variety of provisional data, including drug overdose deaths, called Viral Statistics Rapid Release (VSRR). The NVSS VSRR's dataset documentation page on provisional drug overdose death counts provides a summary of the page, two fabricated dashboards using the dataset, several data tables, technical notes, informative charts, and references/resources to further support their data. The spotlight of the documentation page are the dataset itself, dataset information, and two fabricated dashboards using the dataset.

Accessing the Dataset

The dataset CSV (comma-separated values) file can be accessed by navigating to the 'Options' section below the page summary (*see screenshot below*). To the right of the dashboard drop down menu is a bulleted list of links, with one being 'CSV Format' and the other being 'Data.CDC.gov (Export to CSV, JSON, XLS, XML).' The dataset is within the 'CSV Format' link. Upon clicking that, users will automatically download the dataset.

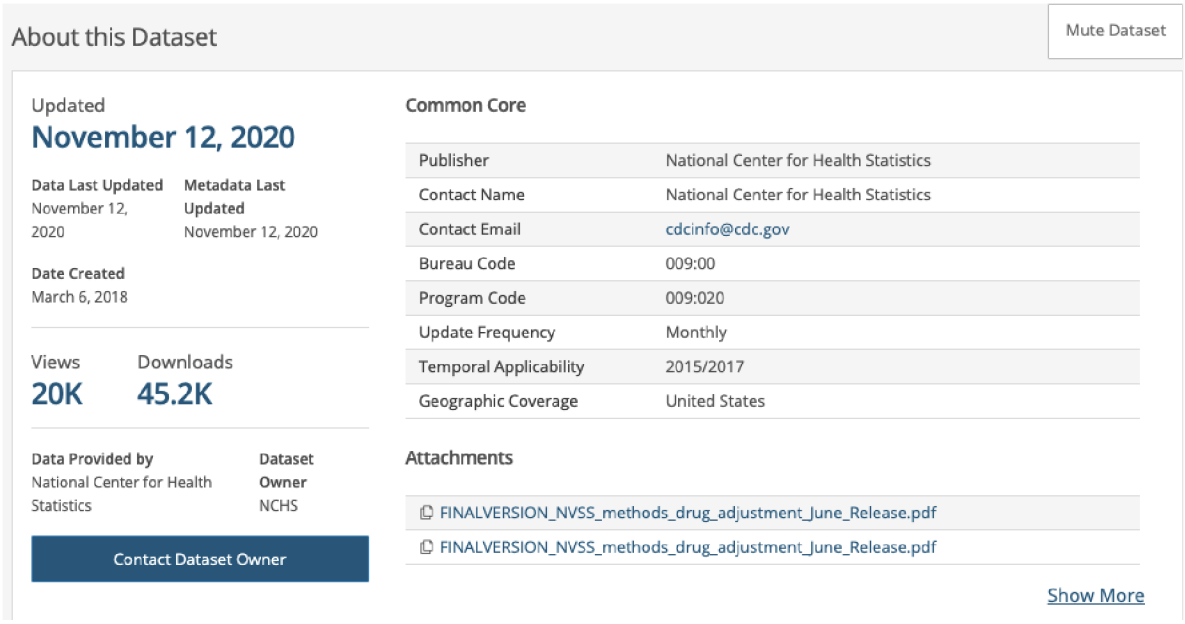


Accessing the Dataset Information

For information on the dataset specifically, clicking the 'Data.CDC.gov (Export to CSV, JSON, XLS, XML)' link takes the user to an overview page which supplies a general summary of the dashboards, a link to the external dashboard(s) used on the documentation page, and dataset specific information.

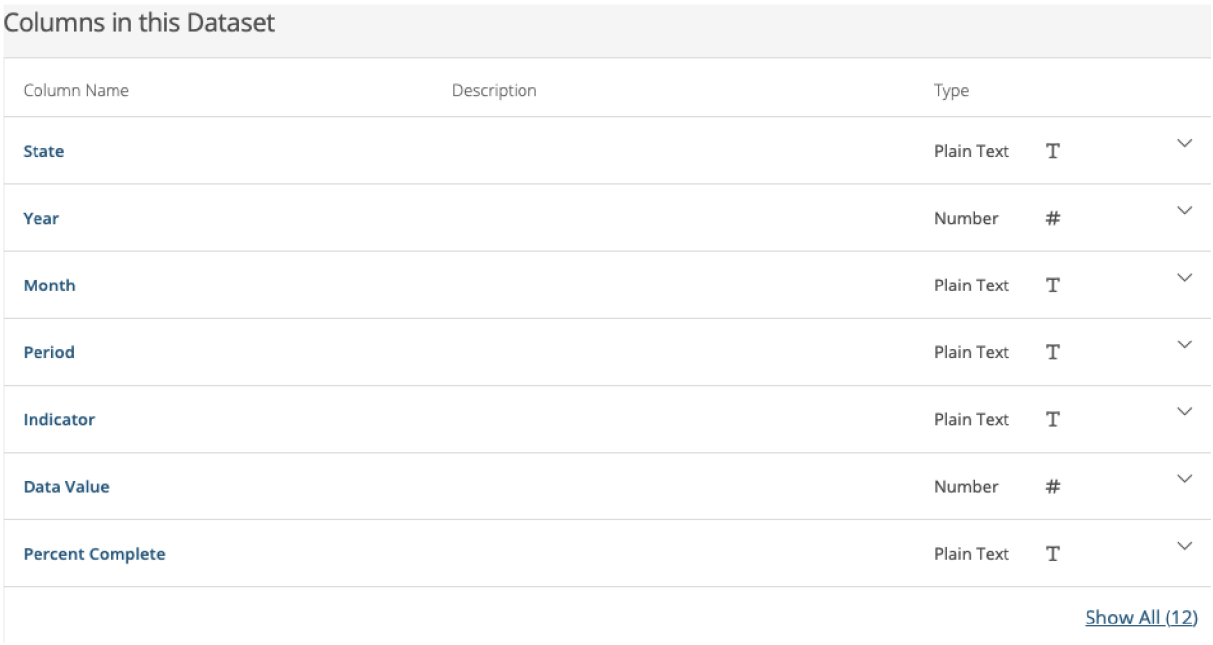
In the section 'About this Dataset' (*see screenshot below*) is information such as when the dataset was created (March 6, 2018), when it was last updated (in terms of when the user is accessing the site), how many views and downloads the dataset/documentation has, common core information (such as publisher, contact email, etc.), additional file attachments, relevant tags

(such as deaths, drug, drug overdose, heroin, cocaine, etc.), and licensing/attribution information (the license is owned by the Public Domain US Government).



***Updated dates subject to change over time*

In the section 'Columns in this Dataset' (*see screenshot below*) is data-metric specific information, such as the column names in the dataset, what that column's data type is (for instance, text or number), and the column's API field name (in instances where users would want to use the dataset's API).



The 'Table Preview' section shows snippets of the actual dataset into manageable, 'next' or 'previous' page views (*see screenshot below*). Each page preview shows the 12 columns with

14 rows of data, with the ability to sort each column by ascending or descending order (if numerical, sorted by numerical order; if alphabetic, sorted by alphabetical order).

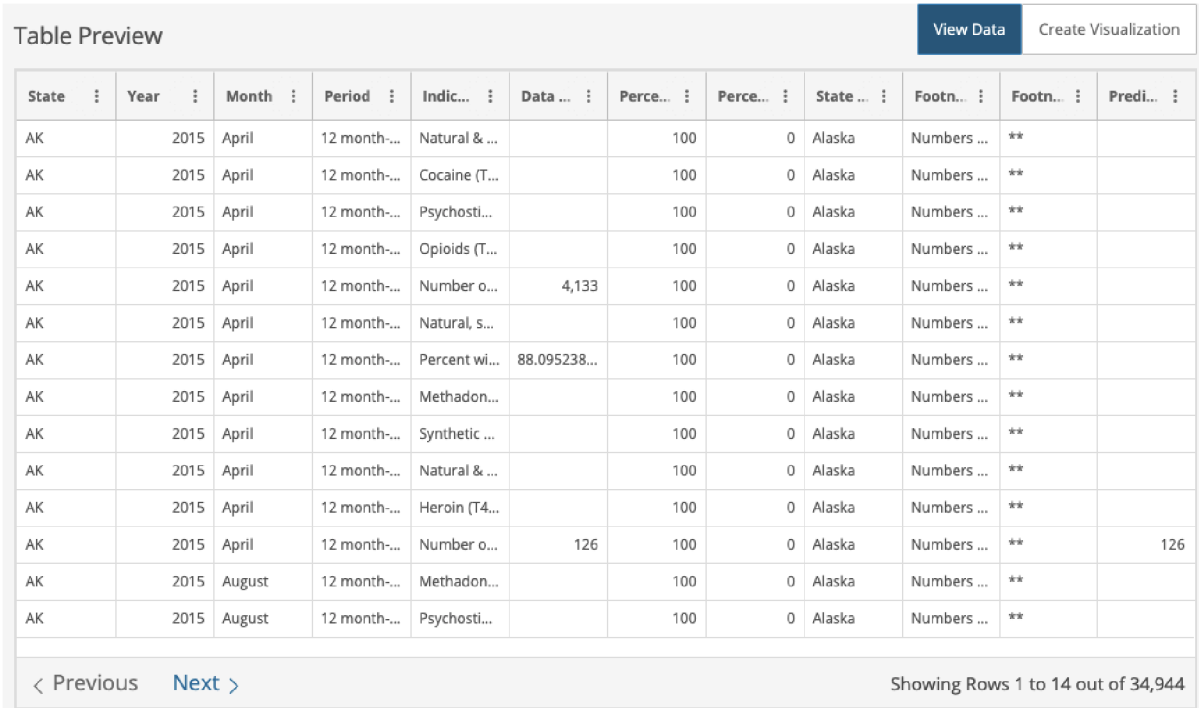


Table Preview

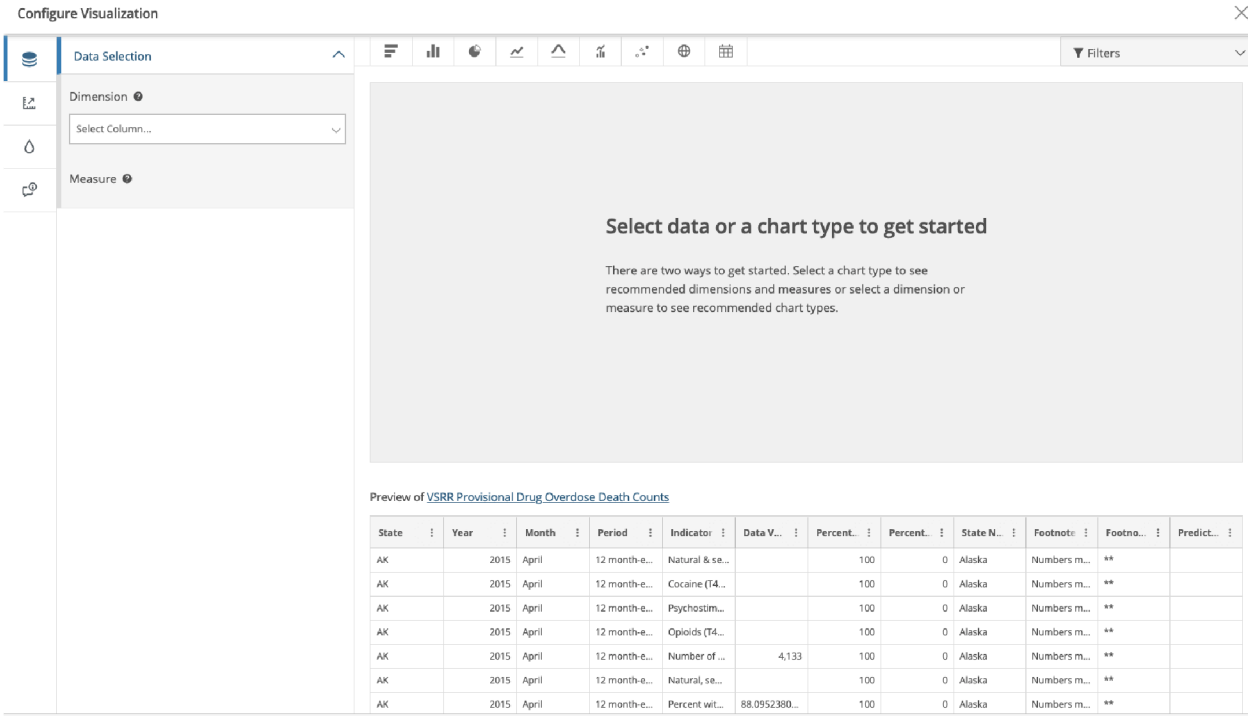
View Data | Create Visualization

State	Year	Month	Period	Indic...	Data ...	Perce...	Perce...	State ...	Footn...	Footn...	Predi...
AK	2015	April	12 month-...	Natural & ...		100	0	Alaska	Numbers ...	**	
AK	2015	April	12 month-...	Cocaine (T...		100	0	Alaska	Numbers ...	**	
AK	2015	April	12 month-...	Psychosti...		100	0	Alaska	Numbers ...	**	
AK	2015	April	12 month-...	Opioids (T...		100	0	Alaska	Numbers ...	**	
AK	2015	April	12 month-...	Number o...	4,133	100	0	Alaska	Numbers ...	**	
AK	2015	April	12 month-...	Natural, s...		100	0	Alaska	Numbers ...	**	
AK	2015	April	12 month-...	Percent wi...	88.095238...	100	0	Alaska	Numbers ...	**	
AK	2015	April	12 month-...	Methadon...		100	0	Alaska	Numbers ...	**	
AK	2015	April	12 month-...	Synthetic ...		100	0	Alaska	Numbers ...	**	
AK	2015	April	12 month-...	Natural & ...		100	0	Alaska	Numbers ...	**	
AK	2015	April	12 month-...	Heroin (T4...		100	0	Alaska	Numbers ...	**	
AK	2015	April	12 month-...	Number o...	126	100	0	Alaska	Numbers ...	**	126
AK	2015	August	12 month-...	Methadon...		100	0	Alaska	Numbers ...	**	
AK	2015	August	12 month-...	Psychosti...		100	0	Alaska	Numbers ...	**	

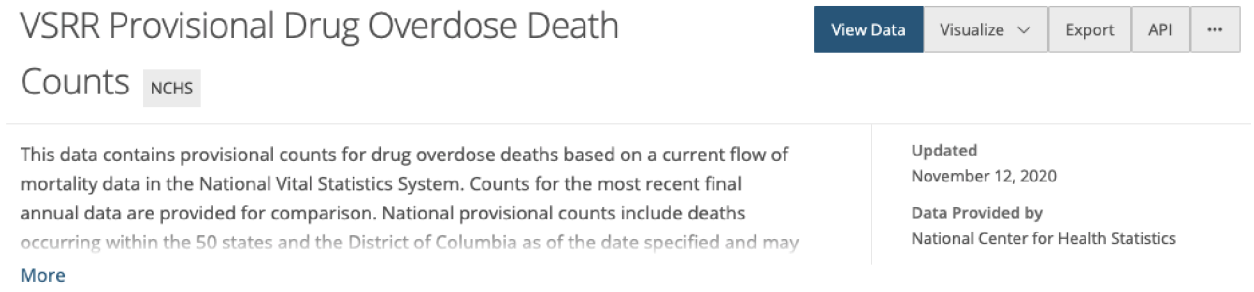
< Previous Next >

Showing Rows 1 to 14 out of 34,944

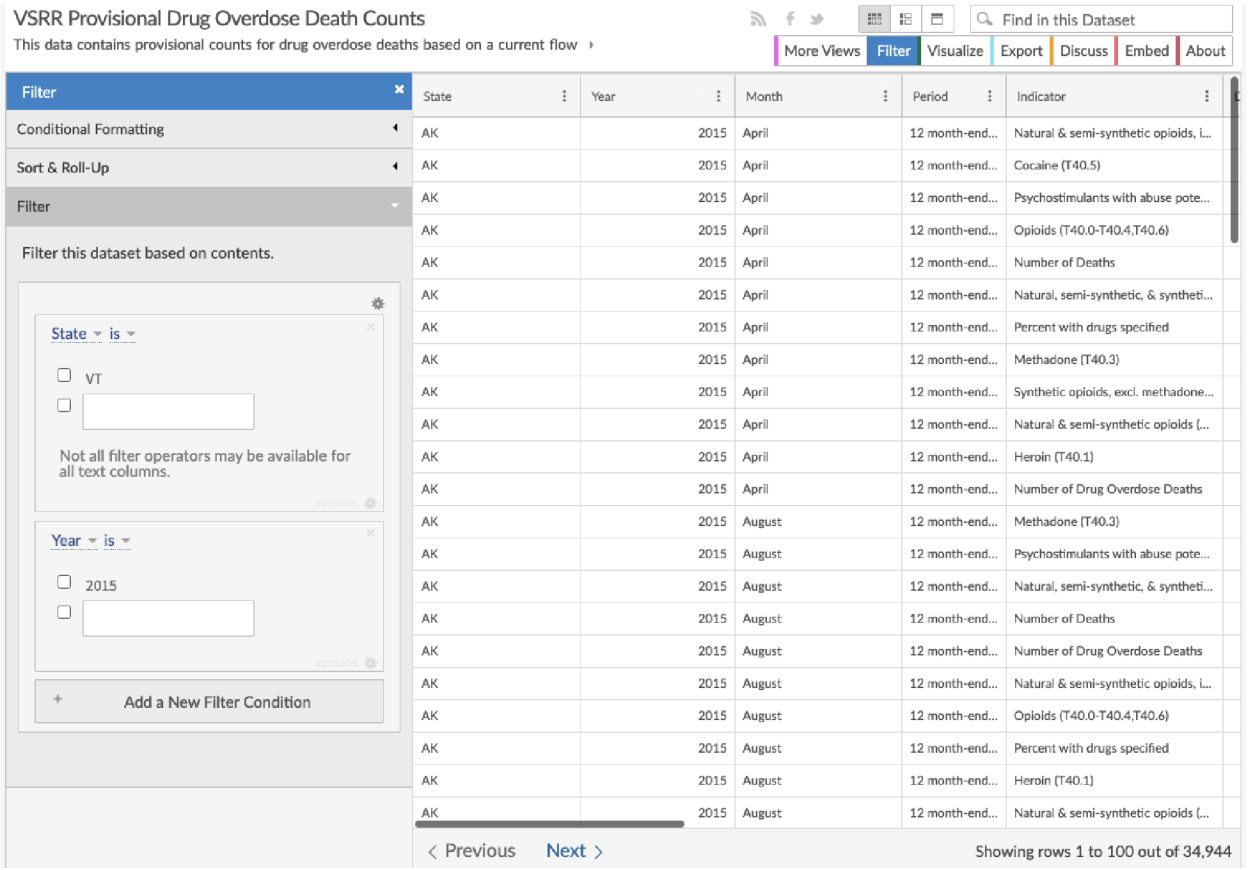
Clicking on the ‘Create Visualization’ link above the top right of the page preview takes the user to a CDC-hosted data visualization creator (*see screenshot below*), where the user can select certain data columns, axis, presentation styles, and create legends. Visualizations that the user can create include bar charts, column charts, pie charts, timeline charts, histograms, combination charts, scatter charts, a map, and a calendar. Depending on what visualization the user wishes to create, specific data input parameters vary.



At the very top of the dataset information page are five links users can interact with: 'View Data,' 'Visualize,' 'Export,' 'API,' and '...'.



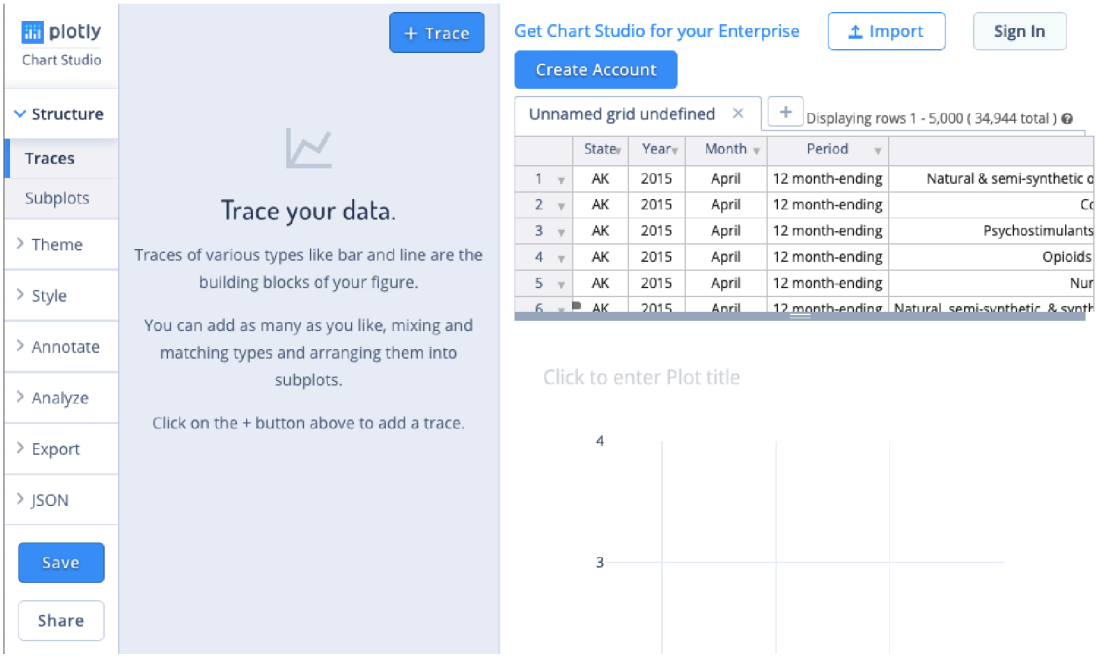
The 'View Data' button takes the user to a Socrata-powered data viewer (*see screenshot below*), which shows the drug overdose dataset in an interactive manner.



Users can filter using conditional formatting (which changes the background color of rows based on specific criteria), sorting, and filtering by state or year. Additional views of the dataset can be accessed to see how the dataset was like during previous years (dataset snapshots). If a user would like to visualize the data, clicking 'Visualize' takes the user to the same CDC-hosted data visualization creator mentioned before. 'Export' allows for the user to download the dataset in various formats, like CSV, JSON, RDF, RSS, TSV, and XML. Users interested in sharing the dataset on personal websites or the internet can do so by clicking on 'Embed,' which allows for the user to customize the dataset in different preview sizes. 'About' shows a link to the dataset information page, a description which is a word-for-word post of the dataset documentation page, visit counts, download counts, metadata, and similar information found on previously mentioned pages.

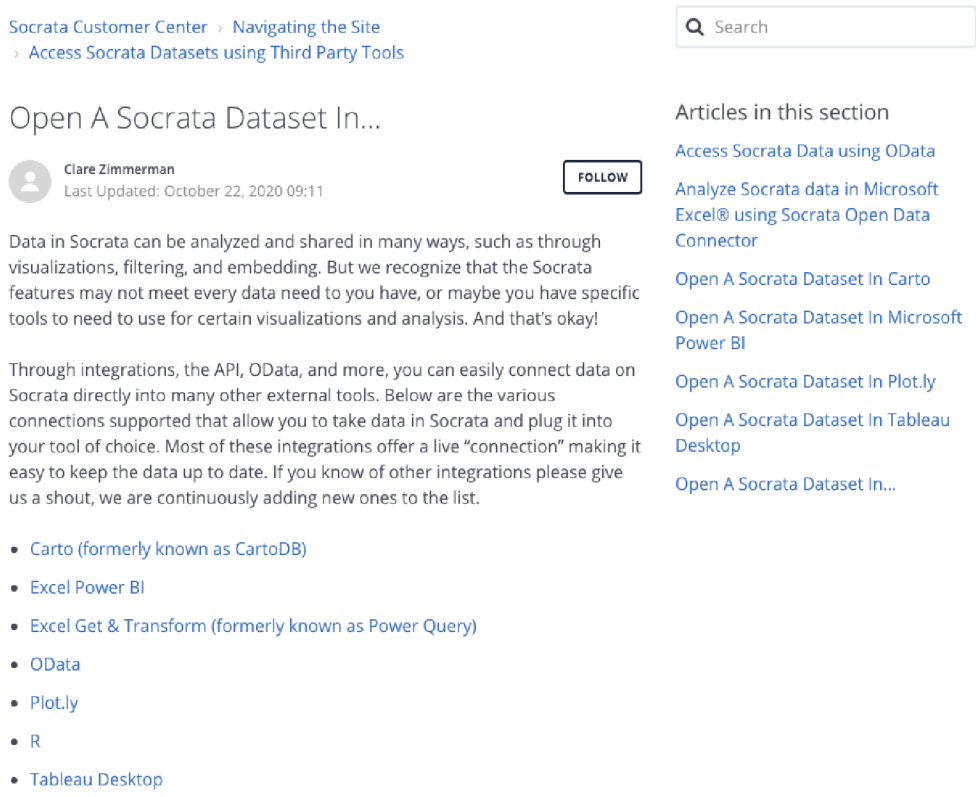
The 'Visualize' button, once clicked, shows a dropdown of more links, including 'Create Visualization,' 'Plot.ly,' and 'More…'. Clicking on 'Create Visualization' takes the user once more to the CDC-hosted data visualization creator (*see above*).

Clicking on 'Plot.ly' warns users that they will be taken to an external website. From there, clicking 'Open' takes the user to Plot.ly, which is another data visualization creating tool (*see screenshot below*).

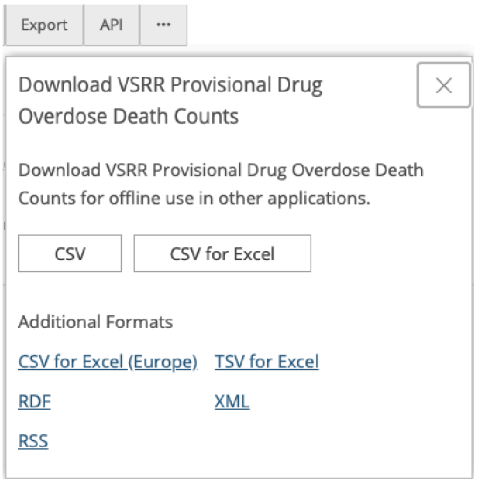


In Plot.ly, users have the ability to view the dataset, as well as Plot.ly's various data creation options. 'Structure' allows for users to 'Trace' data and 'Subplots.' 'Theme' allows users to choose what kind of visualization they want to create with the data. 'Style' enables the user to fine-tune minute visualization details, such as background/margin colors, colorscales, text, titles, axes and more. 'Annotate' allows for adding text or explanations to visualizations. 'Analyze' works as a regression calculator to find trends or averages. 'Export' gives the option of saving/downloading the visualization(s) as either an image or HTML file. 'JSON' allows users to access the Javascript tree-setup of their visualization(s).

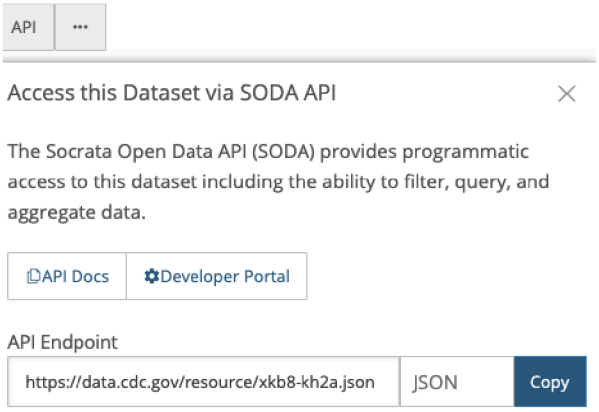
Going back to the 'Visualization' button on the dataset information page, the 'More…' button takes the user to a page (*see screenshot below*) that informs them of the various visualization tools that can be used (such as Carto, Power BI, Excel, OData, Plot.ly, R, and Tableau).



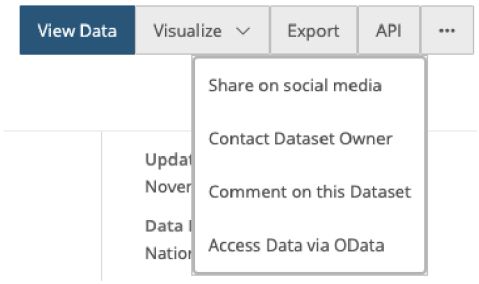
The 'Export' button (*see screenshot below*) allows for users to download the dataset in multiple files, such as CSV, TSV, RDF, XML, or RSS.



The 'API' button once clicked (*see screenshot below*) shows a mini window, including a dataset API link (with word for word documentation as the original dataset documentation page), a developer portal link to guide users on how to use API, and an option to choose either a JSON or CSV API endpoint URL.

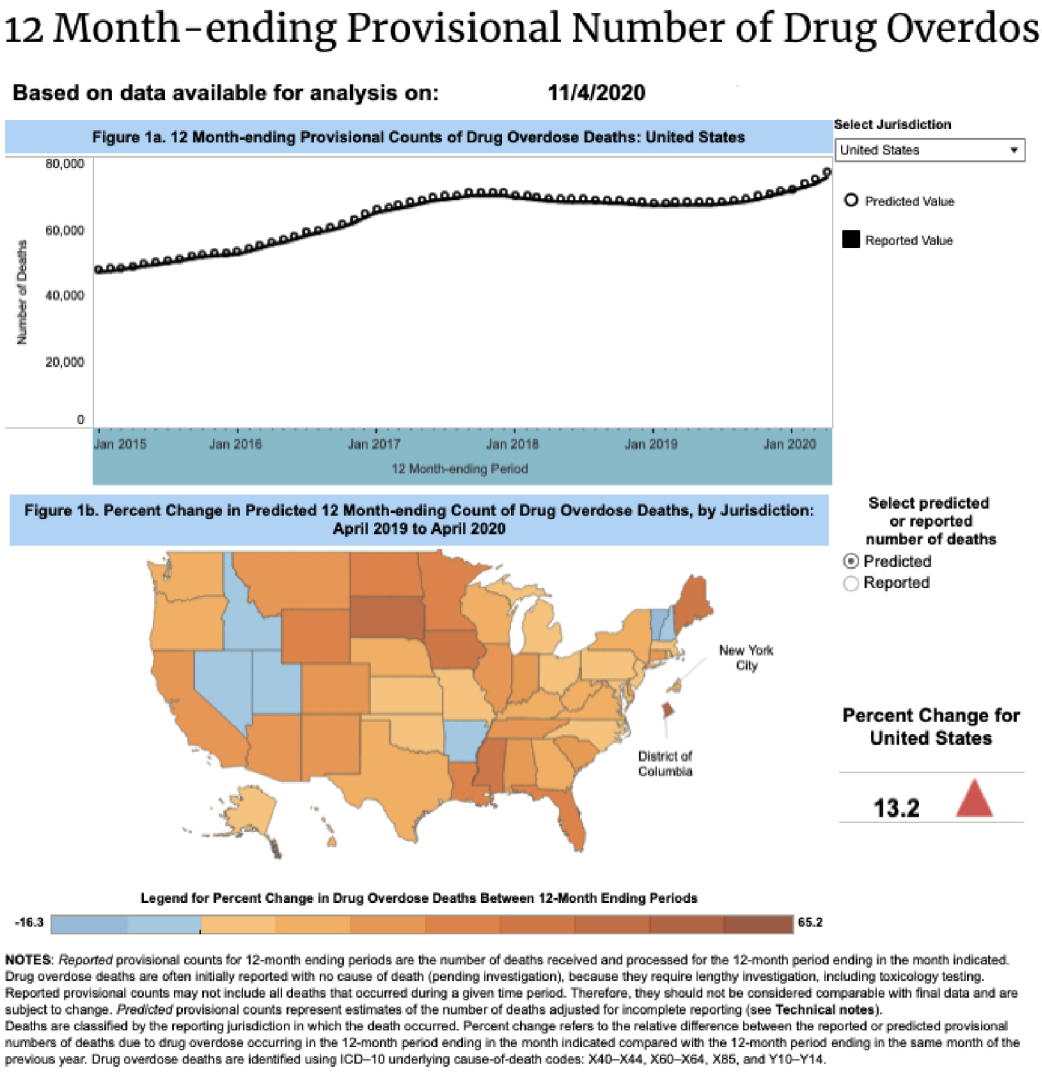


The '...' button, once clicked (*see screenshot below*), gives the user the ability to share the page on social media (shareable via Facebook, Twitter, or email), contact the dataset owner (a message window is shown where the user can input a subject, message, and email), comment on the dataset (which is unavailable as of November 2020), and access the data using other services (like OData, for using the dataset within software like Tableau or Power BI).



Dashboards & Data Tables

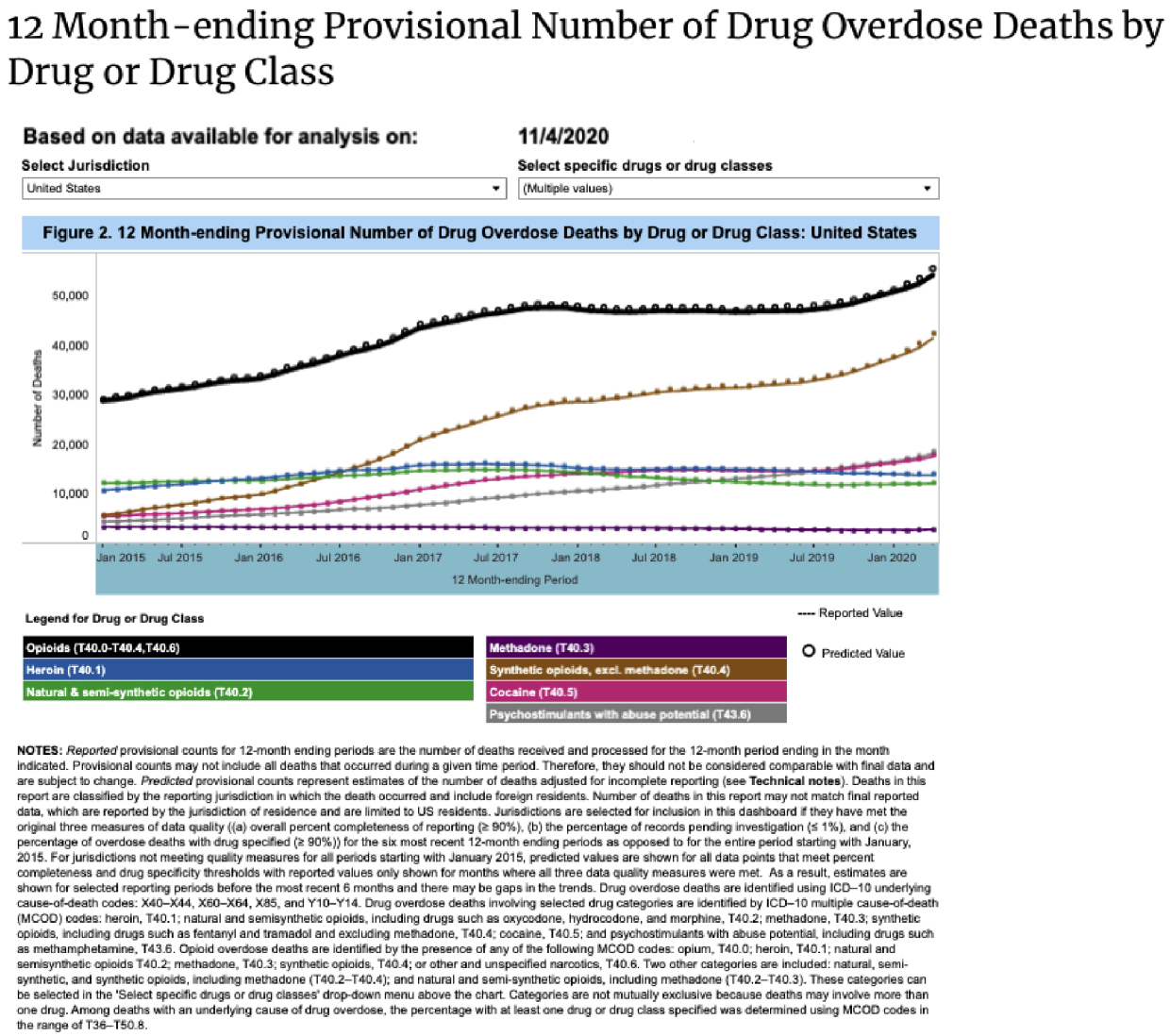
The first dashboard concentrates on '12 Month-ending Provisional Counts and Percent Change of Drug Overdose Deaths,' which shows two figures (*see screenshot below*). The first figure, Figure 1a, shows both the reported and predicted counts of drug overdose deaths as a line graph (which can be filtered by state or nationally). The y-axis is the count of deaths, ranging from 0 to 80,000, and the x-axis shows the month and year (which are navigable using a '-' and '+' button, with '-' being more date specific and '+' being less date specific). The second figure, Figure 1b, portrays a hoverable/clickable US map that shows percent changes of drug deaths from the current year/month to the previous year/month per state, as well as predicted or reported counts (user has the option to choose either) of deaths from the current year/month to the previous year/month per state.



Following the first dashboard are data tables specific to the dashboard. Data table for Figure 1a shows the used data columns from the dataset to develop the dashboard. Columns such as 'Jurisdiction' (United States or state name), 'Month and Year' (for instance, April 2020), 'Reported Value' (death count), 'Predicted Value' (predicted death count), 'Percent Pending Investigation' (for instance, 0.2), and 'Footnote' (which informs users of any underreporting or similar). Data table for Figure 1b follows the same setup, using columns such as 'Jurisdiction' (United States or state name), 'Reported or Predicted,' 'Current 12 Month-ending' (death counts for the current year), 'Previous 12 Month-ending' (death counts for the previous year), 'Percent Change' (for instance, 13.2), and 'Footnote' (which informs users of any underreporting or similar).

The second dashboard concentrates on '12 Month-ending Provisional Number of Drug Overdose Deaths by Drug or Drug Class,' which consists of one figure, Figure 2 (*see screenshot below*), showing the reported and predicted counts of deaths from drug overdoses with specific drug categories (per state and nationally). Users are able to select specific drugs or drug classes

to view, as well as nationally or within specific states. The legend associated with this dashboard identifies specific drug categories with specific colors, for instance, the color black standing for opioids or blue for heroin. Users are able to hover over the dashboard for specific data, including reported and predicted death counts, percent pending investigation, and percent with drugs specified. The y-axis of this dashboard is the count of deaths ranging from 0 to 60,000, and the x-axis shows the month and year (which are navigable using a '-' and '+' button, with '-' being more date specific and '+' being less date specific).



Similar with the first dashboard, a data table is provided specific to the second dashboard. Data table for Figure 2 shows the used data columns from the dataset to develop the dashboard. Columns such as 'Jurisdiction' (United States or state name), 'Indicator' (drug specific cause of death, such as Psychostimulants with abuse potential (T43.6)), 'Month and Year' (for instance, April 2020), 'Reported Value' (death count), 'Predicted Value' (predicted death count), 'Percent

Pending Investigation' (for instance, 0.2), 'Percent with Drugs Specified,' and 'Footnote' (which informs users of any underreporting or similar).

Technical Notes

After the dashboards and their respective data tables are technical notes. The technical notes explain the nature and sources of data, cause of death classifications, selection of states/jurisdictions to report, percentage of records pending investigation, percent completeness, drug specificity, improvements in data quality, delayed reporting adjustments, and the differences between provisional and final data.

Nature & Sources of Data

The data is obtained from the National Center for Health Statistics (NCHS) through death records and death certificates, originally taken from state offices. In most cases, initial counts of deaths from drug overdoses are reported around six months after the actual date because of long investigations[2].

Cause of Death Classification

Causes of deaths have to be recorded in a way that follows World Health Organization regulations, usually by International Statistical Classification of Diseases protocol. This provides a universal baseline in classifying deaths around the world. In terms of drug overdose deaths, ICD's (International Classification of Disease) are followed with cause of death codes, as follows[3]:

- X40–X44 (unintentional)
- X60–X64 (suicide)
- X85 (homicide)
- Y10–Y14 (undetermined)

Drug categories presented include:

- Heroin (T40.1)
- Natural & semi-synthetic opioids (T40.2)
- Natural & semi-synthetic opioids, including methadone (T40.2, T40.3)
- Synthetic opioids, excluding methadone (T40.4)
- Methadone (T40.3)
- Cocaine (T40.5)
- Opioids (T40.0-T40.4,T40.6)
- Psychostimulants with abuse potential (T43.6)
- Natural, semi-synthetic, & synthetic opioids, including methadone (T40.2-T40.4)

[2] For more information, please refer to NVSS VSRR's Technical notes, under the 'Nature and sources of data' section

[3] See NVSS VSRR's Technical notes, under the 'Cause-of-death classification and definition of drug deaths' section

Selection of States/Jurisdictions to Report

Initial provisional drug overdose death counts are given by the state where the death happened.

Percentages of Records Pending Investigation

Drug overdose deaths usually need long or extended periods of investigations. Death certificates can potentially be filed with 'pending investigation,' or an initial unknown cause of death. If a state's percent of records report high amounts of 'pending investigation,' the amount of deaths caused by drug overdose is very likely to be underestimated. States reporting less than 1/100 of their records as 'pending investigation' tend to have less final counts of drug overdose deaths. States reporting more than 1/100 of their records as 'pending investigation' tend to generally underestimate the final count of drug overdose deaths by as much as 30/100.[4]

Percent Completeness

The National Center for Health Statistics takes periodically updated counts of the deaths from each state. This stands for the most suitable estimate of how many deaths occurred in a specific state within a specific month. The percent completeness is created by first dividing the number of deaths for each state and then multiplying by 100[5].

Drug Specificity

Percent of drug overdose death records in terms of a specific drug varies depending on state. States involved in this dataset have met certain required thresholds for specific drugs reported[6]. This is tracked by percent of drug overdose death records with at least one drug recorded.

Improvements in Data Quality

In order for drug overdose death records for a state to be reported to the NCHS VSRR provisional drug overdose dashboard(s), there are requirements that states have to meet in order to ensure data timeliness and consistency. These requirements relate to the continuous relative time of data submission (for instance, the same time each month)[7]. Because of this, states may be added or removed based on whether they pass or fail the criteria. The data table following this technical note shows data quality measures for all states, including 'State Name' (like Alaska), 'Year' (like 2020), 'Month' (like January), 'Period' (12 month ending), 'Percent with drugs specified' (for instance, 95.52), 'Percent Complete' (for instance, 100), and 'Percent Pending Investigation' (for instance, 0.02).

Following this are three tables relating to completeness coefficients of provisional counts after investigative lag, by month and percent pending (involving opioids), and by month and percent pending (for specific drugs).

---

[4] See NVSS VSRR's Technical notes, under the 'Percentage of records pending investigation' section

[5] See NVSS VSRR's Technical notes, under the 'Percent completeness' section

[6] For more information, see NVSS VSRR's Technical notes, under the 'Drug specificity' section

[7] For more information on the requirements, see NVSS VSRR's Technical notes, under the 'Improvements in Data Quality' section

The table relating to completeness of provisional counts after investigating lag (Table 1) has the columns 'Reporting Jurisdiction' (for instance, United States or state), and the months ranging from January to December (there is a column for each month). The values are percentages of completeness.

The table relating to completeness of provisional data by month-ending and percent pending (involving opioids) (Table 2) has the columns 'Model Parameters' (Intercept, Month, and Percent Pending), 'Drug Overdose,' 'Any opioids (T40.0-T40.4,T40.6),' 'Natural, semi-synthetic, and synthetic opioids, including methadone (T40.2-T40.4),' and 'Natural & semi-synthetic opioids and methadone (T40.2-T40.3).' The values are calculations of standard errors (*see screenshot below*).

Table 2. Model results of the completeness of provisional data by month-ending and percent pending: Drug overdose deaths and deaths involving any opioid. Values are estimated coefficients (robust standard errors).

Model Parameters	Drug overdose	Any opioids (T40.0-T40.4,T40.6)	Natural, semi-synthetic, and synthetic opioids, including methadone (T40.2-T40.4)	Natural & semi-synthetic opioids and methadone (T40.2-T40.3)
Intercept	101 (0.1)	100.9 (0.1)	100.9 (0.1)	100.6 (0.1)
Feb	-0.3 (0)	-0.4 (0.1)	-0.5 (0.1)	-0.6 (0.1)
Mar	-0.7 (0.1)	-0.8 (0.1)	-0.9 (0.1)	-0.9 (0.1)
Apr	-0.6 (0.1)	-0.8 (0.1)	-0.9 (0.1)	-0.8 (0.1)
May	-0.7 (0.1)	-0.8 (0.1)	-0.9 (0.1)	-0.8 (0.1)
Jun	-0.6 (0.1)	-0.7 (0.1)	-0.8 (0.1)	-0.7 (0.1)
Jul	-0.7 (0.1)	-0.7 (0.1)	-0.8 (0.1)	-0.7 (0.1)
Aug	-0.8 (0.1)	-0.8 (0.1)	-0.8 (0.1)	-0.8 (0.1)
Sep	-0.6 (0.1)	-0.7 (0.1)	-0.7 (0.1)	-0.8 (0.1)
Oct	-0.7 (0.1)	-0.8 (0.1)	-0.8 (0.1)	-0.8 (0.1)
Nov	-0.5 (0.1)	-0.5 (0.1)	-0.5 (0.1)	-0.4 (0.1)
Dec	-0.3 (0.1)	-0.3 (0.1)	-0.2 (0.1)	-0.1 (0.1)
Percent Pending	-11.9 (0.1)	-11.8 (0.1)	-12.1 (0.1)	-10.3 (0.1)

SOURCE: NCHS, National Vital Statistics System, 2016-2018.

The table relating to completeness of provisional data by month-ending and percent pending (involving specific drugs) (Table 3) has the columns 'Model Parameters' (Intercept, Month, and Percent Pending), 'Heroin (T40.1),' 'Natural & semi- synthetic opioids (T40.2),' 'Methadone (T40.3),' 'Synthetic opioids, excl. methadone (T40.4),' 'Cocaine (T40.5),' and 'Psychostim. w/ abuse potential (T43.6).' The values are calculations of standard errors (*see screenshot below*).

Table 3. Model results of the completeness of provisional data by month-ending and percent pending: deaths involving specific drugs and drug classes. Values are estimated coefficients (robust standard errors).

Model Parameters	Heroin (T40.1)	Natural & semi- synthetic opioids (T40.2)	Methadone (T40.3)	Synthetic opioids, excl. methadone (T40.4)	Cocaine (T40.5)	Psychostim. w/ abuse potential (T43.6)
Intercept	100.9 (0.1)	100.6 (0.1)	100.4 (0.1)	100.6 (0.1)	100.7 (0.1)	100.3 (0.1)
Feb	-0.3 (0.1)	-0.5 (0.1)	-0.6 (0.1)	-0.4 (0.1)	-0.4 (0.1)	-0.5 (0.1)
Mar	-0.8 (0.1)	-0.9 (0.1)	-1 (0.1)	-0.8 (0.1)	-0.7 (0.1)	-0.8 (0.1)
Apr	-0.8 (0.1)	-0.8 (0.1)	-1 (0.1)	-0.7 (0.1)	-1.2 (0.2)	-0.7 (0.1)
May	-0.9 (0.1)	-0.8 (0.1)	-1.1 (0.1)	-0.8 (0.1)	-1.5 (0.2)	-0.8 (0.1)
Jun	-0.8 (0.1)	-0.7 (0.1)	-0.7 (0.1)	-0.5 (0.1)	-1.2 (0.2)	-0.6 (0.1)
Jul	-0.9 (0.1)	-0.7 (0.1)	-0.7 (0.1)	-0.6 (0.1)	-1.2 (0.2)	-0.7 (0.1)
Aug	-0.9 (0.1)	-0.8 (0.1)	-0.7 (0.1)	-0.6 (0.1)	-1 (0.2)	-0.6 (0.1)
Sep	-0.6 (0.1)	-0.9 (0.1)	-0.8 (0.1)	-0.4 (0.1)	-0.6 (0.2)	-0.4 (0.1)
Oct	-0.8 (0.1)	-0.9 (0.1)	-0.7 (0.1)	-0.6 (0.1)	-0.6 (0.2)	-0.6 (0.1)
Nov	-0.6 (0.1)	-0.5 (0.1)	-0.3 (0.1)	-0.4 (0.1)	-0.2 (0.2)	-0.1 (0.1)
Dec	-0.3 (0.1)	-0.1 (0.1)	0 (0.1)	-0.1 (0.1)	0 (0.1)	0.2 (0.1)
Percent Pending	-11.1 (0.1)	-10.3 (0.1)	-9.9 (0.1)	-12.4 (0.1)	-11.5 (0.2)	-11.4 (0.1)

SOURCE: NCHS, National Vital Statistics System, 2016-2018.

Delayed Reporting Adjustments

The way for balancing the initial and final death counts involves calculating drug overdose predicted counts by creating 'multiplication factors' constructed from the severity of underreported provisional data compared with final data[8]. The provisional drug overdose death counts are multiplied by the created factor to make predicted counts that adjust for investigative delays.

Differences Between Provisional & Final Data

Provisional drug overdose death counts are generally lower estimates compared to actual final counts. Final counts are accrued at the end of a certain time period or a given year. The severity of underestimation is mainly calculated by the percent of records that are still in the process of investigation and depends on variables such as state, year, and month of recorded death[9]. The amount of drug overdose deaths will be underestimated to a larger extent in places/states with higher amounts of records reported as 'pending investigation.'

References & Resources

Year 2019 and 2020 death count estimates are based on provisional data, whereas years 2015 to 2018 death count estimates are based on final data. The CDC provides for eight additional references for the user, ranging from timeliness of death certificate data to drug overdose death data briefs from years 1999-2018.

Dataset

[8] For more information on how these factors are calculated, see NVSS VSRR's Technical notes, under the 'Adjustments for delayed reporting' section

[9] Additional information on provisional/final data can be accessed in NVSS VSRR's Technical notes, under the 'Differences between final and provisional data' section

"Monitoring deaths from drug overdose helps in understanding the epidemic's impact on the U.S. population and shows how the crisis is changing over time." —NVSS (National Vital Statistics System)

Background

The NVSS (National Vital Statistics System) is an inter-governmental data sharing system specializing in the topic of public health[10]. Their data is primarily pulled from the NCHS (National Center for Health Statistics) and various US state systems[11]. NVSS has a program that releases a variety of provisional data, including drug overdose deaths, called Viral Statistics Rapid Release (VSRR)[12]. The NVSS VSRR's dataset documentation on provisional drug overdose death counts is the main focus of this user guide. The dataset itself holds information regarding drug overdose deaths ranging from the years 2015 to 2020 from all over the nation.

The dataset of interest for this project, 'VSRR Provisional Drug Overdose Death Counts,' provides provisional estimates for drug overdose deaths based on a pattern of mortality data.

The data within this dataset is received and analyzed from death certificates and cause of death information. Information pertaining to which drugs are involved in a death comes from the cause of death documentation. Due to the long process of drug overdose death investigations, data is continuously updated as new information is received from sources.

Format

The 'VSRR Provisional Drug Overdose Death Counts' dataset was created on March 6, 2018 and has been continuously updated on a monthly basis, published by the NCHS as a .CSV file. The dataset (as of October 2020) has 32,736 rows of data and 12 data columns, including 'State,' 'Year,' 'Month,' 'Period,' 'Indicator,' 'Data Value,' 'Percent Complete,' 'Percent Pending Investigation,' 'State Name,' 'Footnote,' 'Footnote Symbol,' and 'Predicted Value.' For more information on these data columns, such as column data type, description, and examples, refer to the 'Appendix' section.

Things to Know

What is Provisional Data?

Provisional data is data that may not be complete or requires constant updates from several sources (for instance, from government systems or US state jurisdictions)[13]. Provisional data reports are based on new/updated records received from US states or trusted sources. Data may be released monthly or quarterly and can change as information continues to be collected, analyzed, and reported. They are estimated counts that may be different from the final count.

It is important to note that provisional data or information may change as the data becomes more complete. The pro of recording provisional data is promptness and accessibility. Final, complete data takes time to build accuracy and completeness. Faster, more frequent data records could show changes in trends and provide hints about public health patterns (such as

[10] More information about NVSS can be found here
[11] More information about NCHS can be found here
[12] More information about VSRR can be found here
[13] See Footnote 9

drug overdoses). Preliminary data allows for tracking and monitoring the ongoing impact of a crisis.

Definition of Drug Deaths

It is not uncommon for drug overdose deaths to involve multiple drugs. One death might be included in multiple categories when more than one drug is found to be the cause of death. For instance, an overdose that involves heroin and cocaine will be included in the number of drug overdose deaths involving heroin *and* the number of drug overdose deaths involving cocaine. Drug overdose deaths are sorted by categories, identified by specific (ICD) codes. Drug categories presented include[13]:

- Heroin (T40.1)
- Natural & semi-synthetic opioids (T40.2)
- Natural & semi-synthetic opioids, including methadone (T40.2, T40.3)
- Synthetic opioids, excluding methadone (T40.4)
- Methadone (T40.3)
- Cocaine (T40.5)
- Opioids (T40.0-T40.4,T40.6)
- Psychostimulants with abuse potential (T43.6)
- Natural, semi-synthetic, & synthetic opioids, including methadone (T40.2-T40.4)

Context

The dataset shows counts of drug overdose deaths in regards to periodically updated data. Provisional counts include deaths reported in the fifty states (and certain large-population cities). The dataset includes reported and predicted counts of deaths due to drug overdose occurring in the United States, the percent changes in drug overdose deaths, and the predicted counts of drug overdose deaths involving specific drugs.

Metrics

There are certain metrics to ensure the quality of this data. This includes the percent complete in death reporting, percent of deaths needing further investigation, and the percent of drug overdose deaths with specific drugs. These measures help improve the accuracy of drug overdose counts. Reports of specific drugs and drug classes involved in deaths vary by state.

Appendix

The dataset has 32,736 entries for 12 different data columns, including ‘State,’ ‘Year,’ ‘Month,’ ‘Period,’ ‘Indicator,’ ‘Data Value,’ ‘Percent Complete,’ ‘Percent Pending Investigation,’ ‘State Name,’ ‘Footnote,’ ‘Footnote Symbol,’ and ‘Predicted Value.’ The reported and predicted provisional counts represent the numbers of deaths in consequence of drug overdoses. Deaths are reported by the state/jurisdiction where the death happened.

Here is what each of the data columns in the dataset stand for, with their respective data type, description, and example:

[13] See Footnote 3

Data Column	Data Type	Description/Example
State	*Text/Object*	• State abbreviation (e.g.: Maryland is abbreviated as MD) • Example: MD
Year	*Number/Integer*	• Year of data entry, ranges from 2015-2020 • Example: 2015
Month	*Text/Object*	• Month of data entry, ranges from January-December • Example: January
Period	*Text/Object*	• All data 32,736 entries are based on a 12-month calendar • Example: 12 month-ending
Indicator	*Text/Object*	• Drug overdose deaths categories, identified by specific multiple cause-of-death codes. Drug categories presented include[14]: ◦ Heroin (T40.1) ◦ Natural & semi-synthetic opioids (T40.2) ◦ Natural & semi-synthetic opioids, including methadone (T40.2, T40.3) ◦ Synthetic opioids, excluding methadone (T40.4) ◦ Methadone (T40.3) ◦ Cocaine (T40.5) ◦ Opioids (T40.0-T40.4,T40.6) ◦ Psychostimulants with abuse potential (T43.6)

[14] See Footnote 3

		○ Natural, semi-synthetic, & synthetic opioids, including methadone (T40.2-T40.4)

Data Value	*Number/Integer*	• Calculated provisional count of deaths • Provisional drug overdose death counts are based on death records received and processed by the National Center for Health Statistics • Example: 126
Percent Complete	*Text/Object*	• Obtained by dividing the number of death records in the NVSS database for each state for each 12 month period by the control counts and multiplying by 100. • Example: 100
Percent Pending Investigation	*Number/Integer*	• More so than not, drug overdose deaths need long periods of investigations. Death certificates may be filed with 'pending investigation,' or with an initial unknown cause of death. • Example: 0
State Name	*Text/Object*	• State spelled in its entirety • Example: Maryland
Footnote	*Text/Object*	• Text-based column telling users to refer to 'Technical Notes' in the database website or if the data entry is of high/low quality or incomplete. • Example: 'See Technical Notes'
Footnote Symbol	*Text/Object*	• * or ** symbol to relay to users which Footnote to refer to
Predicted Value	*Number/Integer*	• Methods to adjust provisional counts of drug overdose deaths for underreporting

		• Example: 126

Data Cleaning/Scrubbing

The data cleaning process is done within Jupyter Notebook using Python. Python libraries used for this process are Pandas and NumPy. For extra documentation on these libraries, please refer to the official Python documentation manuals for each respective library. To view the Python Notebook for the data cleaning process, please navigate to the 'Cleaning Notebook.ipynb' file in the project's GitHub repository.

In order for certain software or programming languages to work with data of any kind, the data must be software- or programming language-friendly. There are different directions in terms of 'cleaning' or 'scrubbing' data. The most common reasons for 'cleaning' data are to make sure there is no:

- Missing information
- Repetitive data
- Spelling errors
- Human errors (such as incorrectly inputting a number)

Inclusion of any of these errors or similar will render the involved data to be of poor quality and can cause misinformation or a skew in research results. In order to clean the dataset of interest for this project, the Python programming language will be used within a software called Jupyter Notebook. The Python libraries being used will be Pandas and NumPy.

The original dataset has a total of 32,736 entries/rows with 12 columns, including including 'State,' 'Year,' 'Month,' 'Period,' 'Indicator,' 'Data Value,' 'Percent Complete,' 'Percent Pending Investigation,' 'State Name,' 'Footnote,' 'Footnote Symbol,' and 'Predicted Value' (*see screenshot below*).

	State	Year	Month	Period	Indicator	Data Value	Percent Complete	Percent Pending Investigation	State Name	Footnote	Footnote Symbol	Predicted Value
0	AK	2015	April	12 month-ending	Psychostimulants with abuse potential (T43.6)	NaN	100	0.0	Alaska	Numbers may differ from published reports usin...	**	NaN
1	AK	2015	April	12 month-ending	Cocaine (T40.5)	NaN	100	0.0	Alaska	Numbers may differ from published reports usin...	**	NaN
2	AK	2015	April	12 month-ending	Methadone (T40.3)	NaN	100	0.0	Alaska	Numbers may differ from published reports usin...	**	NaN
3	AK	2015	April	12 month-ending	Synthetic opioids, excl. methadone (T40.4)	NaN	100	0.0	Alaska	Numbers may differ from published reports usin...	**	NaN
4	AK	2015	April	12 month-ending	Heroin (T40.1)	NaN	100	0.0	Alaska	Numbers may differ from published reports usin...	**	NaN

Ten of the twelve columns (*see screenshot below*) are categorized as 'object' data types. This poses a problem that will be discussed later. One column, Year, is an integer data type. One column, Percent Pending Investigation, is a float/decimal data type.

```
RangeIndex: 32736 entries, 0 to 32735
Data columns (total 12 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   State                           32736 non-null  object
 1   Year                            32736 non-null  int64
 2   Month                           32736 non-null  object
 3   Period                          32736 non-null  object
 4   Indicator                       32736 non-null  object
 5   Data Value                      27149 non-null  object
 6   Percent Complete                32736 non-null  object
 7   Percent Pending Investigation   32736 non-null  float64
 8   State Name                      32736 non-null  object
 9   Footnote                        32736 non-null  object
 10  Footnote Symbol                 32736 non-null  object
 11  Predicted Value                 20750 non-null  object
dtypes: float64(1), int64(1), object(10)
```

Looking for null, missing, or NaN values shows that the dataset has a total of 17,573 NaN values (not a number or null value), 5,587 of which come from the 'Data Value' column, and 11,986 of which come from the 'Predicted Value' column (*see screenshot below*).

```
State                              0
Year                               0
Month                              0
Period                             0
Indicator                          0
Data Value                      5587
Percent Complete                   0
Percent Pending Investigation      0
State Name                         0
Footnote                           0
Footnote Symbol                    0
Predicted Value                11986
dtype: int64
```

The 'Data Value' is the count of provisional deaths, whereas 'Predicted Value' are estimated counts. Because the analysis of this dataset relies heavily on the presence of available 'Data Value' and 'Predicted Value' entries, we will be dropping all entries with NaN. We do not want to insert false death data into the analysis, so we choose not to fill the NaN values with other alternatives and simply remove them from the dataset.

Upon inspection of the 'VSRR Provisional Drug Overdose Death Counts' dataset, we additionally chose to remove the 'Footnote,' 'Footnote Symbol,' and 'Period' columns. The 'Footnote' column tells users to refer to 'Technical Notes' which are outside of the dataset. The 'Footnote Symbol' column also tells users to refer to sources outside of the dataset. The 'Period' column has the value of '12 month-ending' for all 32,736 entries, which is self-explanatory and unnecessary for data analytics.

After removing the missing data and unnecessary columns, 20,577 data entries and 9 columns remain (*see screenshots below*).

	State	Year	Month	Indicator	Data Value	Percent Complete	Percent Pending Investigation	State Name	Predicted Value
8	AK	2015	April	Number of Drug Overdose Deaths	126	100	0.0	Alaska	126
21	AK	2015	August	Number of Drug Overdose Deaths	124	100	0.0	Alaska	124
34	AK	2015	December	Number of Drug Overdose Deaths	121	100	0.0	Alaska	121
36	AK	2015	February	Number of Drug Overdose Deaths	127	100	0.0	Alaska	127
52	AK	2015	January	Number of Drug Overdose Deaths	126	100	0.0	Alaska	126

```
Int64Index: 20577 entries, 8 to 32735
Data columns (total 9 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   State                        20577 non-null  object 
 1   Year                         20577 non-null  int64  
 2   Month                        20577 non-null  object 
 3   Indicator                    20577 non-null  object 
 4   Data Value                   20577 non-null  object 
 5   Percent Complete             20577 non-null  object 
 6   Percent Pending Investigation  20577 non-null  float64
 7   State Name                   20577 non-null  object 
 8   Predicted Value              20577 non-null  object 
dtypes: float64(1), int64(1), object(7)
```

Further analysis reveals that the actual data types of the columns are not as they outwardly seem to be, which relates to the problem of having ‘object’ data types (see first paragraph of Data Cleaning). One would assume that columns such as ‘Data Value,’ ‘Percent Complete,’ and ‘Predicted Value’ hold numeric values-- when in actuality, are ‘object’ data types within the dataset. Object data types cannot be analyzed numerically, because they are not integer/numeric/float data types. This is a problem because the general consensus of these columns are indeed integers upon first glance but need to be programmed into integer/number format in order to run them into any software/analytical tools.

The ‘Data Value’ column is listed as an ‘object’ data type because of one reason: there are commas within the data values. Generally speaking, commas within numbers make the numbers easier to read, but are not software/programming friendly. To fix this, the commas within the ‘Data Value’ entries will be replaced with an empty string, such as ‘ ‘ (*see screenshot below*). For example, the number 10,000 will then turn into 10000. Now the ‘Data Value’ column can be converted into an integer/numerical data type.

```
# fixing the data value column to convert it into int type
fff['Data Value'] = fff['Data Value'].str.replace(',', '')
fff['Data Value'] = fff['Data Value'].astype(int)
```

The ‘Percent Complete’ column is listed as an ‘object’ data type because there are certain entries incorporating ‘+,’ such as ‘+99.5.’ The use of ‘+’ makes the entries impossible to analyze as numeric/integer data types. To fix this, ‘+’ will be replaced with an empty string, such as ‘ ‘ (*see screenshot below*). For example, the value +99.5 will turn into 99.5. Now the ‘Percent Complete’ column can be converted into an integer/numerical data type.

```
# fixing the percent complete column to convert it into float type
fff['Percent Complete'] = fff['Percent Complete'].str.replace('+', '')
fff['Percent Complete'] = fff['Percent Complete'].astype(float)
```

The ‘Predicted Value’ column is listed as an ‘object’ data type for the same reason ‘Data Value’ was. The use of a comma ‘,’ turns the number into an ‘object’ within programming software/tools. To fix this, the same method will be used by replacing the comma with an empty string, ‘ ‘ (*see screenshot below*), and will turn a number like 90,000 into 90000. Now the ‘Predicted Value’ column can be converted into an integer/numerical data type.

```
# fixing the predicted value column to convert it into int type
fff['Predicted Value'] = fff['Predicted Value'].str.replace(',', '')
fff['Predicted Value'] = fff['Predicted Value'].astype(int)
```

Re-evaluating the data types for the columns after fixing these issues shows that 4 columns are objects (State, Month, Indicator and State Name are all text-based entries), 3 are integer/numeric, and 2 are float/decimal types (*see screenshot below*).

```
Int64Index: 20577 entries, 8 to 32735
Data columns (total 9 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   State                          20577 non-null  object
 1   Year                           20577 non-null  int64
 2   Month                          20577 non-null  object
 3   Indicator                      20577 non-null  object
 4   Data Value                     20577 non-null  int64
 5   Percent Complete               20577 non-null  float64
 6   Percent Pending Investigation  20577 non-null  float64
 7   State Name                     20577 non-null  object
 8   Predicted Value                20577 non-null  int64
dtypes: float64(2), int64(3), object(4)
```

Integer/numeric/float/decimal data types can all be run in numerical calculations and analyses. Checking for null/NaN values shows that there are no missing values (*see screenshot below*).

```
State                          0
Year                           0
Month                          0
Indicator                      0
Data Value                     0
Percent Complete               0
Percent Pending Investigation  0
State Name                     0
Predicted Value                0
dtype: int64
```

As a recap, the data cleaning process deals with:

- Identifying the dataset data types
- Finding null/NaN values and removing them
- Removing unnecessary data columns
- Adjusting the data columns to proper data types
- Re-checking for null/NaN values and proper data types

Now, the data is cleaned and is ready to be analyzed and worked with. The cleaned data will be saved as 'CleanedData1.csv,' and can be accessed within the project's GitHub repository.

Example Data Analyses

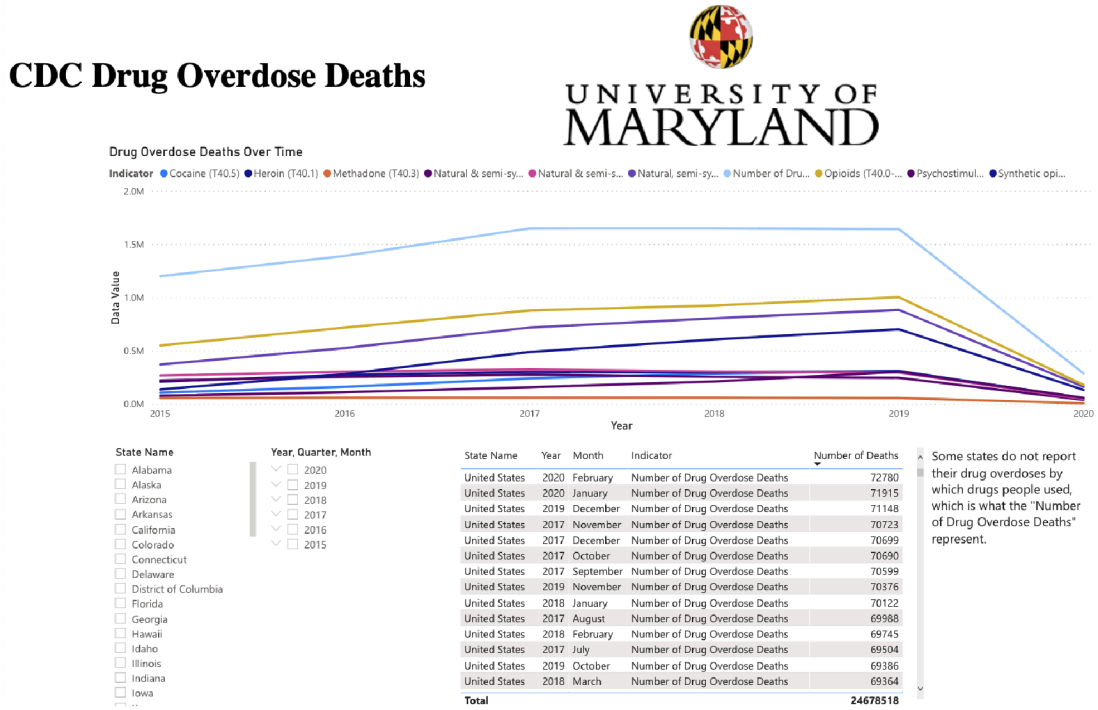
Users interested in this dataset can utilize the information in order to analyze possible relationships, trends, or predictions relating to drug overdose deaths. Analyses can incorporate the curation of dashboards, charts, graphs, or statistical analyses obtained by using the dataset in software such as Tableau, Power BI, or programming languages like Python, R, and SQL. These data analyses examples use the cleaned version (CleanedData1.csv) of the original dataset.

Power BI Dashboard

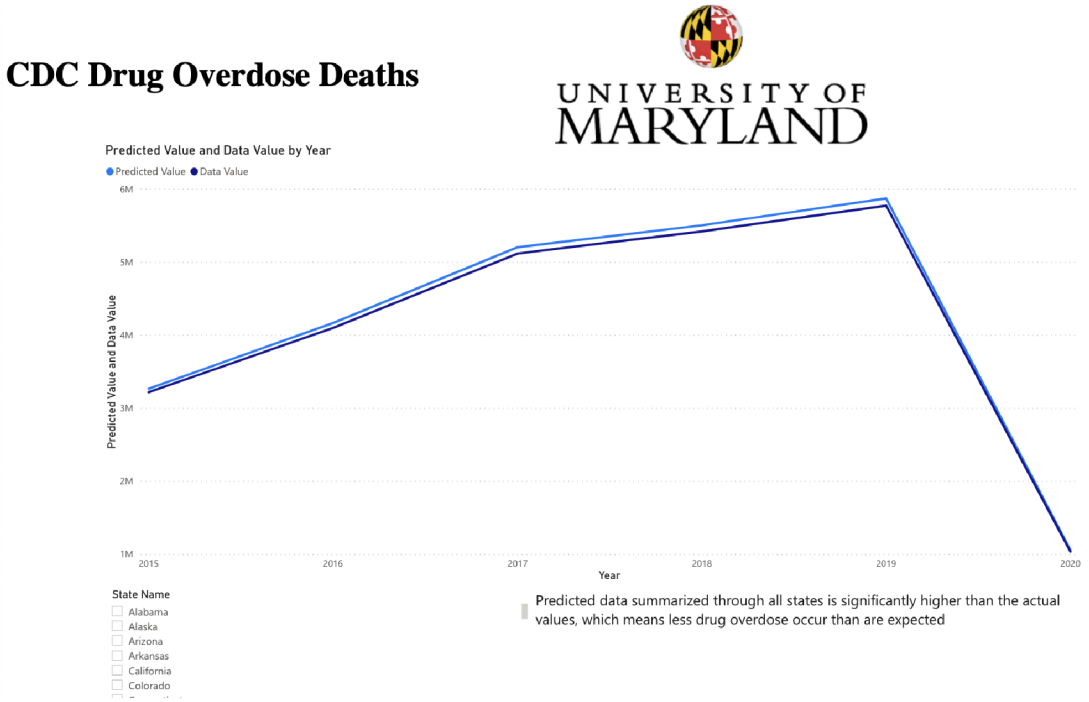
Microsoft Power BI is a software that provides users tools for aggregating, analyzing, visualizing and sharing data. Microsoft Power BI is used to find insights within a particular dataset. Power BI helps connect datasets, transform, clean the data and create charts or graphs to provide visuals of the data.

The goal of this Microsoft Power BI report is to allow users to see what can be done with the dataset. The three pages of the Power BI report created are just three examples of what can be done.

The first page (*see screenshot below*) is an overview of the entire dataset. The filters at the bottom left of the page allow the user to filter the data down to visualize trends over certain date ranges and based on the state. If the user hovers over the top graphic, they will be able to see the specific data point for each type of drug overdose for their selected time range. In addition to this, the chart on the bottom is responsive to the filters, which also allows for an in depth look at the number of deaths in a date range and what their cause is.



The second page (*see screenshot below*) shows the summary of the predicted value of drug overdoses, with the provisional value. The filter at the bottom once again allows for filtering by state, which gives the user the opportunity to zoom in on their area of study. One thing to note about this visualization is that this is a summary of all states over the course of the entire 5 years from the dataset, so it is indicative of an overall trend in the USA over those 5 years.



The third page (*see screenshot below*) shows the comparison of each type of drug overdose death by the state, which shows where the most drug overdose deaths are occurring. As noted on page 2, some states do not report their drug overdose deaths by which type of drug, so there is a large number of deaths under the category, "Number of Drug Overdose Deaths." Other than that minor issue, this visual allows users to filter by the type of drug that people overdosed on, the state you want to see, as well as the time period. This visual also shows you the total number of drug overdose deaths, which is responsive to the filters so you can have an accurate sum of each state, drug type, or time frame.

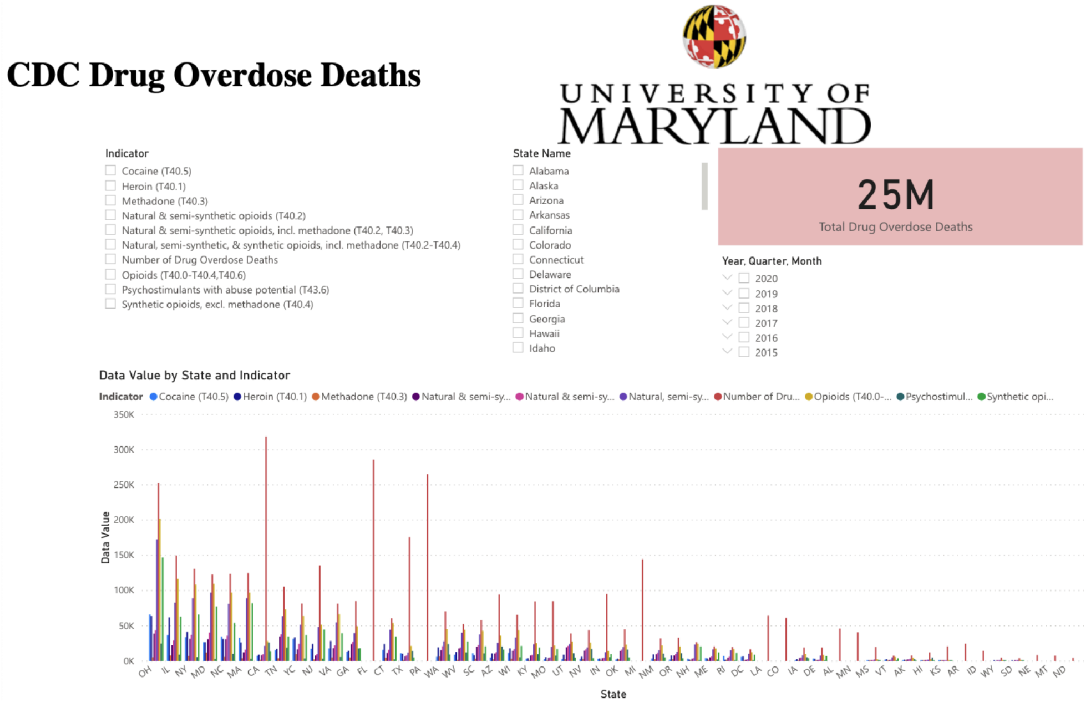
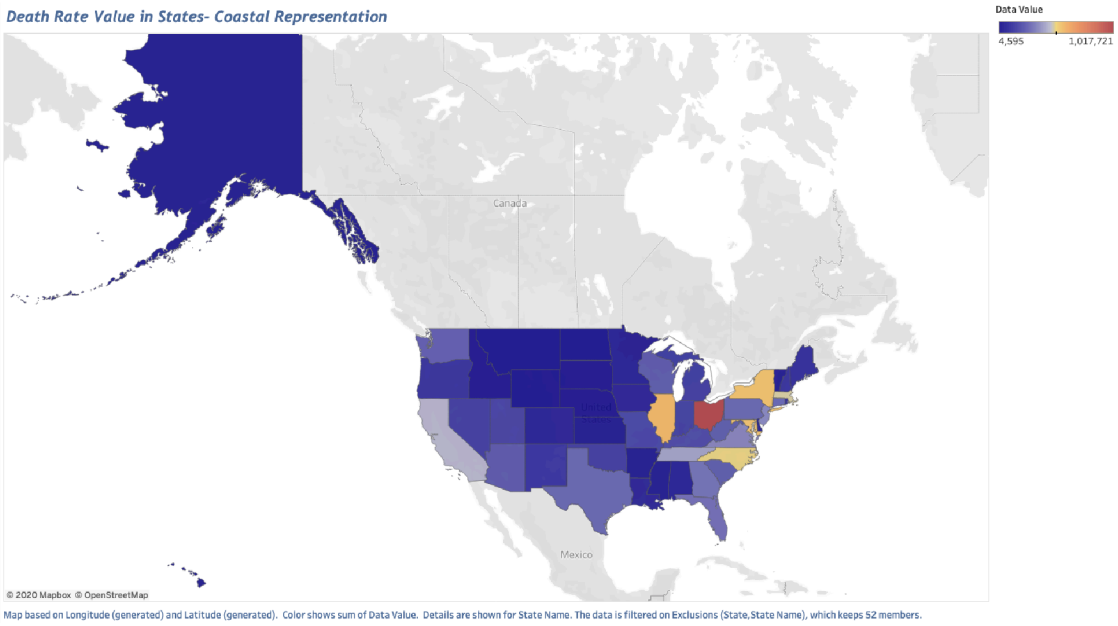


Tableau Dashboard 1

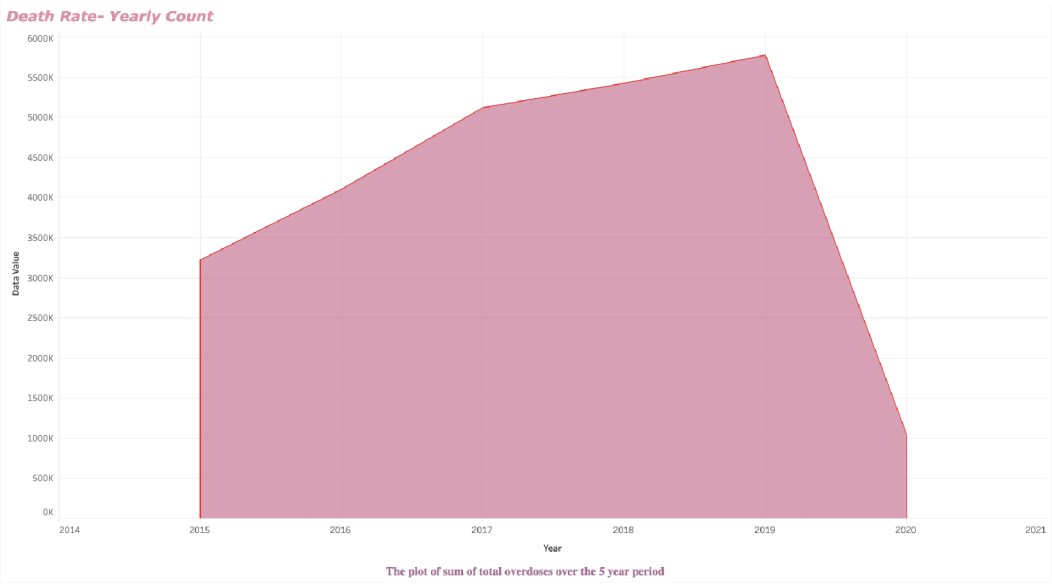
Tableau is a user-friendly interactive data visualization and analysis software. Its tendency to create interactive visual analytics in the form of dashboards has made it popular for users to utilize this application for understanding processed data. These dashboards make it easier for non-technical individuals and users to change data into interactive and understandable concepts. With the assistance of Power BI and Tableau, users can draw concise analysis on project assignments.

The first Tableau visualization (*see screenshot below*) illustrates the calculated provisional count of deaths across all 50 states within the United States. Provisional drug overdose death counts are based on death records received and processed by the National Center for Health Statistics (NCHS). Some filters and tools within the first visualization include a search bar, a home button to reset, and zoom in and out for easy clarification and analysis. On the top right part of the visualization, there is a red-black diverging palette tool with ranges between

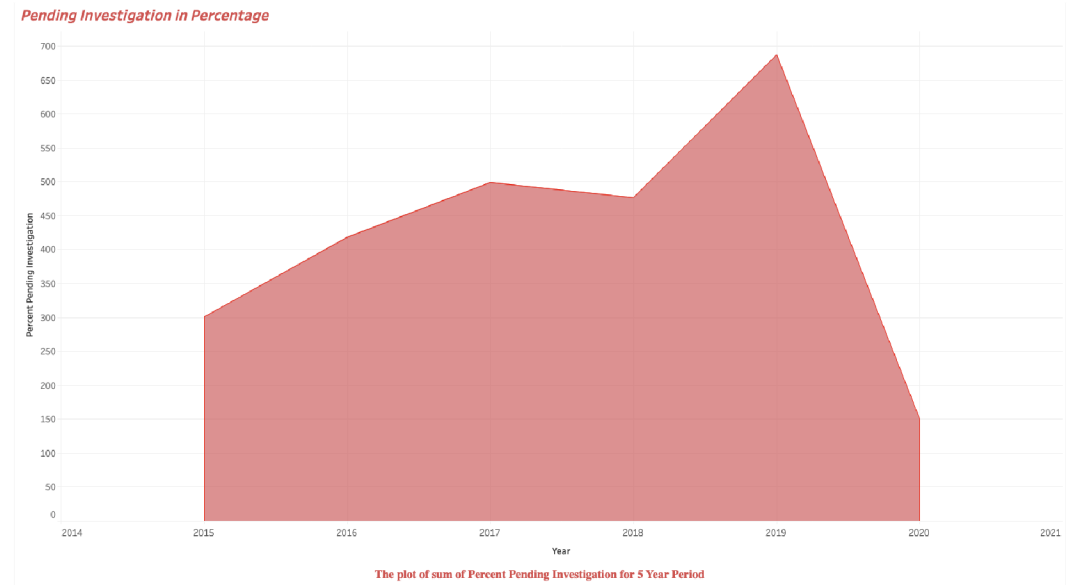
4000 and 1,018,000 counts of death with red (through white) to black representations respectively. This allows easy reading of death rates within the states and regions. When observed, there are less drug overdoses in Alaska and the mid-northern part of the United States. However, users can see an increased number of cases towards the north-eastern and south-eastern part of the United States. There is not a precise pattern in the states with most rates.



The subsequent Tableau visualization (*see screenshot below*) on the second tab illustrates the death rate over the period of 5 years. From this graphical analysis, we observed a growth of almost a million death rates from 2015 to 2017. Between 2017 to 2019, there was a slight decrease in death rates. The trend was still increasing but not at a rapid rate as the previous years. Between 2019 and 2020, there was a drastic decrease of drug overdose related deaths of about 4 million.



The third Tableau visualization (*see screenshot below*) on the third tab elaborates the pending investigation in percentage in years. Note that the percentage exceeds 100%. Mode of measure is increased to allow easy reading of the visualization. The trend shows selected reporting periods before the most recent six months. In this analysis, it is realized that there were about 100 pending cases every year from 2015 to 2017. From 2017 to 2018, there was a slight decrease in cases and then an increase of almost 200 cases from 2018 to 2019. A massive decrease of pending cases were reported within 2019 and 2020. From the second and third Tableau analysis, users can observe an increase in death rates and pending investigations between the first 2 or 3 consecutive years with a drastic decrease from 2019 to 2020 in both graphs.



The fourth visualization using Tableau (*see screenshot below*) shows the adjusted provisional counts of drug overdose deaths of various drug types.

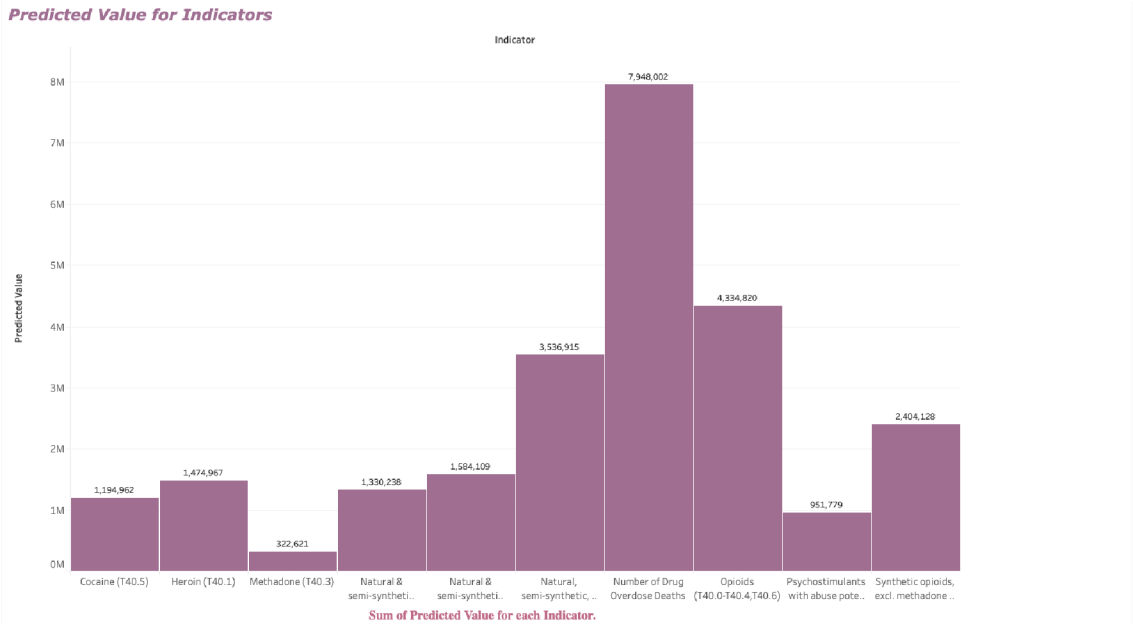
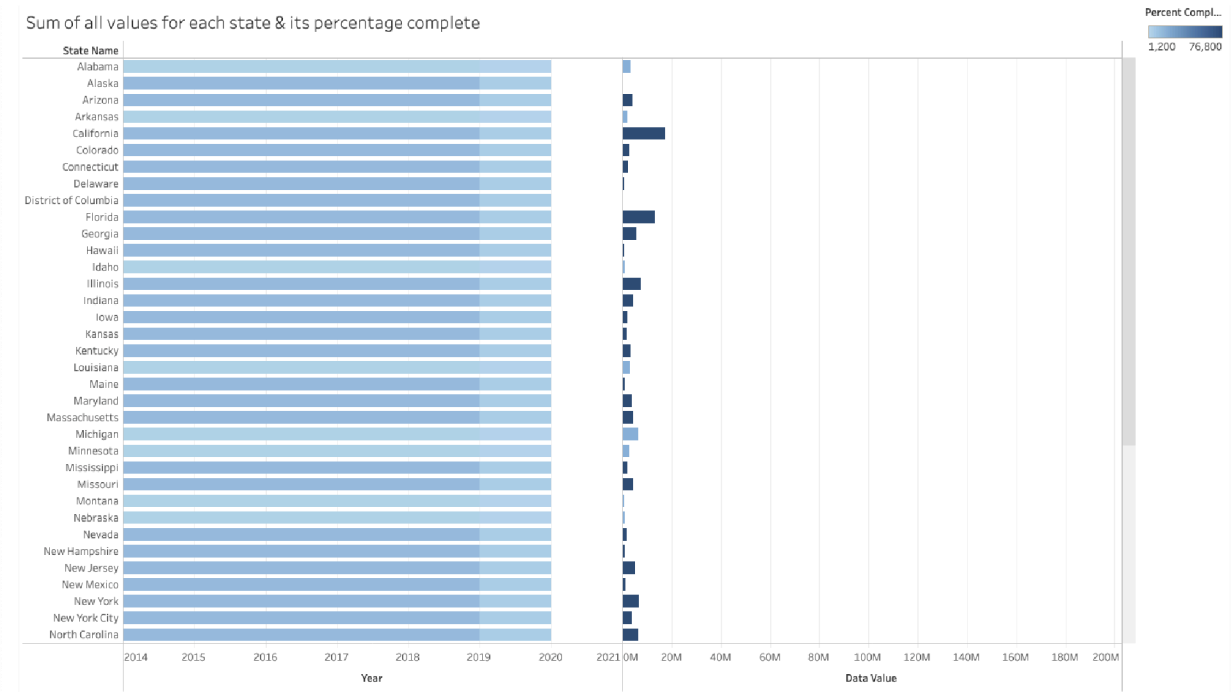


Tableau Dashboard 2

This analysis also uses Tableau as a data visualization tool. Tableau allows users to input data from a .CSV/Excel file and then create visualizations based on the variables in the data. It enables the creation of many different types of graphs from maps to bar charts.

This dataset shows and visualizes two very important indicators (*see screenshot below*): the percent completeness and the data value (the number value of overdoses). It is crucial to show the completeness for a state because it shows how much data is left to be included. For example, a state could have a very low number of overdoses and a researcher might think that they should not give funding to that state because it may not be an issue there.



However, the state might only be reporting a fraction of the data. If all of the data is actually reported then the state could potentially be having a lot of overdose deaths and it may turn into a major issue. One of the limitations is population. Although some states like California might stand out in terms of drug overdose death counts, it is important to note that they have the highest population rate compared to other states. It is important to compare the number and data value of the overdoses to how big the state actually is and how many people the state actually has.

Python Analysis

The Jupyter Notebook file for this analysis is 'Data Analysis.ipynb'. To view the Python Notebook, please navigate to the 'Data Analysis.ipynb' file in the project's GitHub repository.

In addition to utilizing software such as Tableau and Power BI, users can also work with data using programming languages such as Python, R, and SQL. For this analysis, Python is used. Python is a general-purpose coding language and can be used for other types of programming and software development besides web development. In this case, Python will be used for processing data and performing mathematical computations and developing charts, by using its specific libraries such as NumPy, Pandas, and Matplotlib (*for documentation of these libraries, as well as the Python language, please refer to their official guides*). NumPy is a library used for working with arrays, Pandas library is used for data analysis and manipulation, and Matplotlib is a library for creating static, animated, and interactive visualizations.

To integrate all of these libraries and their outputs into one environment, Jupyter Notebook will be used. Jupyter Notebook is an open-source, interactive web tool that combines live code, graphics, visualizations, and text in shareable notebooks that run in a web browser.

In ‘Data Analysis.ipynb,’ the Jupyter Notebook file for this analysis, the first step is to import all necessary Python packages/libraries (in this case it will be NumPy, Pandas, and Matplotlib). This is done in the section ‘Importing Packages.’

Importing Packages

```
# importing packages

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

Then, in the ‘Reading in Data’ section, a data frame is created and the ‘CleanedData1.csv’ file is read and previewed (*see screenshot below*).

Reading in Data

```
# reading in data

df = pd.read_csv("CleanedData1.csv")
df.head()
```

	Unnamed: 0	State	Year	Month	Indicator	Data Value	Percent Complete	Percent Pending Investigation	State Name	Predicted Value
0	8	AK	2015	April	Number of Drug Overdose Deaths	126	100.0	0.0	Alaska	126
1	21	AK	2015	August	Number of Drug Overdose Deaths	124	100.0	0.0	Alaska	124
2	34	AK	2015	December	Number of Drug Overdose Deaths	121	100.0	0.0	Alaska	121
3	36	AK	2015	February	Number of Drug Overdose Deaths	127	100.0	0.0	Alaska	127
4	52	AK	2015	January	Number of Drug Overdose Deaths	126	100.0	0.0	Alaska	126

Acquiring general information/statistics about the cleaned dataset begins. The ‘Dataset Information/Statistics’ section (*see screenshot below*) tells users there are 20,577 rows and 9 columns in the cleaned dataset. Out of these 9 columns, 4 are object data types, 3 are integer data types, and 2 are float data types.

Dataset Information/Statistics

```
# seeing how many rows and columns are in the dataset

df.shape
```

```
(20577, 9)
```

```
# see what data types there are in the dataset

df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20577 entries, 0 to 20576
Data columns (total 9 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   State                          20577 non-null  object
 1   Year                           20577 non-null  int64
 2   Month                          20577 non-null  object
 3   Indicator                      20577 non-null  object
 4   Data Value                     20577 non-null  int64
 5   Percent Complete               20577 non-null  float64
 6   Percent Pending Investigation  20577 non-null  float64
 7   State Name                     20577 non-null  object
 8   Predicted Value                20577 non-null  int64
dtypes: float64(2), int64(3), object(4)
memory usage: 1.4+ MB
```

Following this are brief statistical summaries (count, mean, standard deviation, minimum, 25%, 50%, and 75% quartiles, and maximum) of the 9 columns in the dataset (it is recommended to ignore these statistics for 'Year,' as 'Year' only ranges from years 2015-2020).

```
# see statistical summary for all columns (ignore year, only ranges from 2015–2020)

df.describe()
```

	Year	Data Value	Percent Complete	Percent Pending Investigation	Predicted Value
count	20577.000000	20577.000000	20577.000000	20577.000000	20577.000000
mean	2017.299995	1199.325363	99.998591	0.123051	1218.960052
std	1.483763	5042.060784	0.033766	0.166313	5117.793705
min	2015.000000	10.000000	98.900000	0.000000	10.000000
25%	2016.000000	89.000000	100.000000	0.017647	90.000000
50%	2017.000000	286.000000	100.000000	0.052672	289.000000
75%	2019.000000	695.000000	100.000000	0.159506	710.000000
max	2020.000000	72780.000000	100.000000	1.411568	74144.000000

Further analysis of the values and value counts for each column is made. For instance, the states and counts of states can be found for the 'State' column, the year and year counts for 'Year,' month and month counts for 'Month,' indicator and indicator counts for 'Indicator,' data values and data value counts for 'Data Value,' and so forth.

Before beginning analysis of the data, the 'Data Wrangling' section filters the cleaned dataset by 'Year' (*see screenshot below*). This will create a data frame for each year that is in the cleaned dataset ('y2015' will only consist of data from year 2015, 'y2016' from year 2016, etc.).

Data Wrangling

```
# organizing the data by year

y2015= df.loc[df['Year']==2015]
y2016= df.loc[df['Year']==2016]
y2017= df.loc[df['Year']==2017]
y2018= df.loc[df['Year']==2018]
y2019= df.loc[df['Year']==2019]
y2020= df.loc[df['Year']==2020]
```

The ‘Data Analytics’ will consist of three examples: finding the average predicted values per month in 2015, finding the average predicted value per month in 2015 for Maryland, and finding the amount of each drug overdose per month in 2016.

The analytics for finding the average predicted values per month in 2015 begins by creating a new variable that only consists of the ‘Month’ and ‘Predicted Value’ columns from the ‘y2015’ data frame. This new variable will show the mean/average of ‘Predicted Value’ per month in 2015 (*see screenshot below*).

Data Analytics

Finding the average predicted values per month in 2015

```
# now that we have dataset variables for each year, we can dive deeper into analytics!
# for example, let's see the average 'Predicted Value' for each month in 2015:

pv_y2015 = y2015[['Month','Predicted Value']].groupby(['Month']).agg('mean')
pv_y2015
```

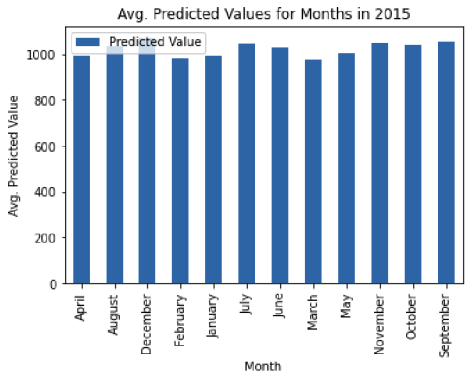
Month	Predicted Value
April	993.796992
August	1035.356618
December	1071.626335
February	982.200803
January	994.404959
July	1043.045627
June	1029.859316
March	975.844358
May	1004.423221
November	1049.517857
October	1040.771429
September	1054.602941

From the calculated mean/average of ‘Predicted Value’ for each month in 2015, users can then identify which month has the highest average and which month has the lowest. In this case, December has the highest average ‘Predicted Value’ of 1071.626335 overdose deaths, whereas March has the lowest average ‘Predicted Value’ of 975.844358. Users can see a histogram of this data to better visualize the differences in means (*see screenshot below*).

```
# let's try visualizing that:
# note that we may need to reset index if we want to use an index as a graph variable

pv1_y2015 = pv_y2015.reset_index()
graph = pv1_y2015.plot.bar(x='Month', y='Predicted Value')
graph.set_ylabel('Avg. Predicted Value')
graph.set_xlabel('Month')
graph.set_title('Avg. Predicted Values for Months in 2015')
```

```
Text(0.5, 1.0, 'Avg. Predicted Values for Months in 2015')
```



The analytics for finding the average predicted value per month in 2015 for Maryland also uses the 'y2015' data frame, however it is filtered to only include data where 'State Name' is 'Maryland.' This is defined as a new variable, 'MDy2015.' Following similar steps as the previous analysis, 'Month' and 'Predicted Value' from 'MDy2015' will be grouped together and will show the mean/average 'Predicted Value' by 'Month' (*see screenshot below*).

Finding the average predicted value per month in 2015 for Maryland

```
# let's try seeing the avg. predicted values for Maryland in each month of 2015
# first we use the y2015 data and filter for the State Name 'Maryland'

MDy2015= y2015.loc[y2015['State Name']=='Maryland']

# then let's see the average 'Predicted Value' for each month in 2015 for Maryland:

MD_y2015 = MDy2015[['Month','Predicted Value']].groupby(['Month']).agg('mean')
MD_y2015
```

Month	Predicted Value
April	458.1
August	467.1
December	534.6
February	457.2
January	441.5
July	459.9
June	453.9
March	459.1
May	459.4
November	519.0
October	506.8
September	485.7

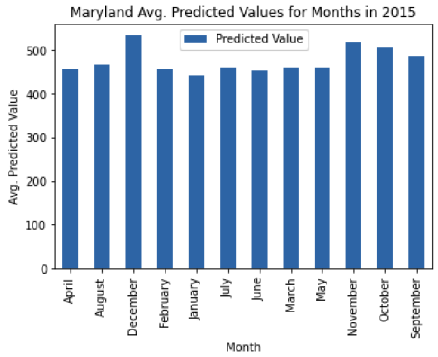
From this calculated data, users can then identify which month in Maryland has the highest average of predicted overdose deaths, and which month has the lowest. In this case,

December and November hold the highest averages, at 534.6 and 519.0, respectively. The month with the lowest average predicted value is January, with an average of 441.5. Users can see a histogram of this data to better visualize the differences in means (*see screenshot below*).

```
# let's try visualizing that:
# note that we may need to reset index if we want to use an index as a graph variable

MD1_y2015 = MD_y2015.reset_index()
graph = MD1_y2015.plot.bar(x='Month', y='Predicted Value')
graph.set_ylabel('Avg. Predicted Value')
graph.set_xlabel('Month')
graph.set_title('Maryland Avg. Predicted Values for Months in 2015')
```

```
Text(0.5, 1.0, 'Maryland Avg. Predicted Values for Months in 2015')
```



The analytics for finding the amount of each drug overdose per month in 2016 starts with using the ‘y2016’ data frame, since this analysis is only concerned with data from year 2016. The ‘Month,’ ‘Indicator,’ and ‘Predicted Value’ columns are the columns in interest and are grouped by ‘Indicator’ and ‘Month,’ where the sum/counts of ‘Predicted Value’ are made. This is defined to a variable called ‘pvt_y2016.’ A preview of what ‘pvt_y2016’ looks like shows a table (*see screenshot below*), where each indicator is followed with the months January to December, and then followed by a column of sum/counts of predicted overdose deaths.

Finding the amount of each drug overdose per month in 2016

```
# let's try to see what the count for each drug overdose was for each month in 2016

pvt_y2016 = y2016[['Month', 'Indicator','Predicted Value']].groupby(['Indicator','Month']).a
pvt_y2016
```

		Predicted Value
Indicator	Month	
Cocaine (T40.5)	April	12400
	August	14517
	December	16916
	February	11680
	January	11266
...	...	...
Synthetic opioids, excl. methadone (T40.4)	March	19189
	May	21771
	November	30487
	October	28705
	September	27375

To dive deeper into a specific month, for instance, January, ‘y2016’ is filtered to only show data where ‘Month’ is ‘January.’ Similar grouping/aggregation rules are applied from the

previous paragraph to show the predicted value sum/counts for each indicator in January, 2016 (*see screenshot below*).

```
# filtering down to a certain month (January)

janpvt_y2016 = y2016.loc[y2016['Month']=='January']
janpvt_y2016
janpvt1_y2016 = janpvt_y2016[['Month', 'Indicator','Predicted Value']].groupby(['Indicator',
janpvt1_y2016
```

		Predicted Value
Indicator	Month	
Cocaine (T40.5)	January	11266
Heroin (T40.1)	January	21169
Methadone (T40.3)	January	5091
Natural & semi-synthetic opioids (T40.2)	January	20059
Natural & semi-synthetic opioids, incl. methadone (T40.2, T40.3)	January	24009
Natural, semi-synthetic, & synthetic opioids, incl. methadone (T40.2-T40.4)	January	36922
Number of Drug Overdose Deaths	January	107370
Opioids (T40.0-T40.4,T40.6)	January	53374
Psychostimulants with abuse potential (T43.6)	January	8282
Synthetic opioids, excl. methadone (T40.4)	January	16753

To make analysis easier, the table will be sorted by 'Predicted Value,' in descending order.

```
# sorting January indicator/predicted values

janpvt1_y2016sort= janpvt1_y2016.sort_values(by=['Predicted Value'], ascending=False)
janpvt1_y2016sort
```

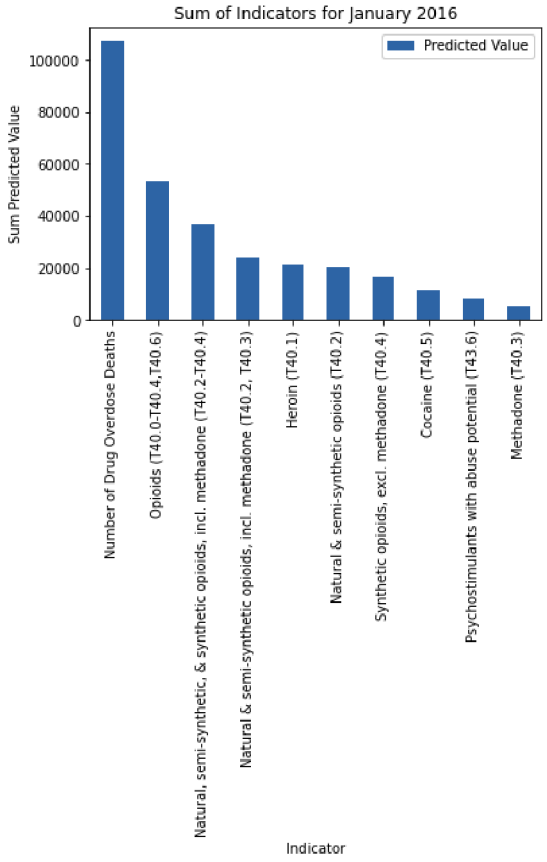
		Predicted Value
Indicator	Month	
Number of Drug Overdose Deaths	January	107370
Opioids (T40.0-T40.4,T40.6)	January	53374
Natural, semi-synthetic, & synthetic opioids, incl. methadone (T40.2-T40.4)	January	36922
Natural & semi-synthetic opioids, incl. methadone (T40.2, T40.3)	January	24009
Heroin (T40.1)	January	21169
Natural & semi-synthetic opioids (T40.2)	January	20059
Synthetic opioids, excl. methadone (T40.4)	January	16753
Cocaine (T40.5)	January	11266
Psychostimulants with abuse potential (T43.6)	January	8282
Methadone (T40.3)	January	5091

From here, users can see that the total number of predicted drug overdose deaths in January 2016 is 107,370. The highest drug-overdose type is opioids, at a predicted value sum of 53,374. The lowest drug-overdose type is Methadone, at a predicted value sum of 5,091. Users can see a histogram of this data to better visualize the differences in counts/sums (*see screenshot below*).

```
# let's try visualizing that:
# note that we may need to reset index if we want to use an index as a graph variable

janpvt1_y2016sort1 = janpvt1_y2016sort.reset_index()
graph = janpvt1_y2016sort1.plot.bar(x='Indicator', y='Predicted Value')
graph.set_ylabel('Sum Predicted Value')
graph.set_xlabel('Indicator')
graph.set_title('Sum of Indicators for January 2016')
```

```
Text(0.5, 1.0, 'Sum of Indicators for January 2016')
```



Version history

Version	Month and Year
1.0	November 2020
2.0	November 2022
3.0	November 2024