Stat 515
Mid-Term Project

Redesigning a Bad Graph - Spaghetti to Micromaps

Chaithanya Kasula (G01197109) & Aishwarya Varala (G01206112)

Introduction

The primary focus of the current report is redesigning a bad graph depicting Opioid overdose death rates per 100,000 (Age-Adjusted), in different states of the US. It also discusses detailed approaches and techniques used to obtain meaningful insights form the data. The strengths and weakness of the bad graph and the redesigned graph are annotated. Further, data analysis has been performed to reveal hidden knowledge and such information has been visualized through effective and interactive graphics.

Bad Graph:

Graph Source:

https://datausa.io/visualize?enlarged=c-lineplot-Zpn26u&groups=0-Zpn26u&measure=2sUCF4

The bad graph in discussion has been obtained from 'Data USA' which provides shared US government data for public use.

Description:

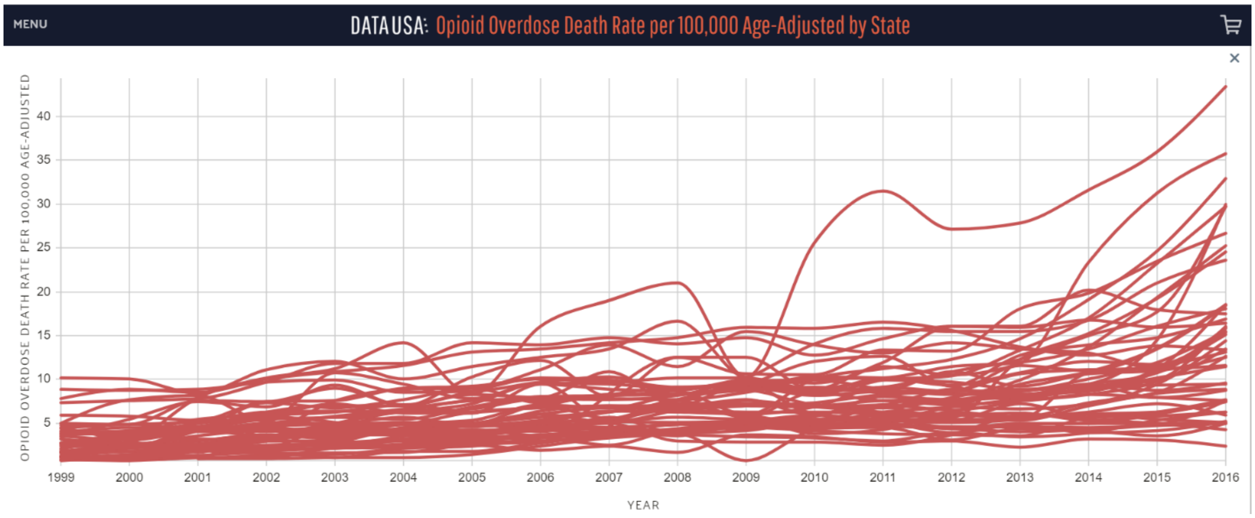


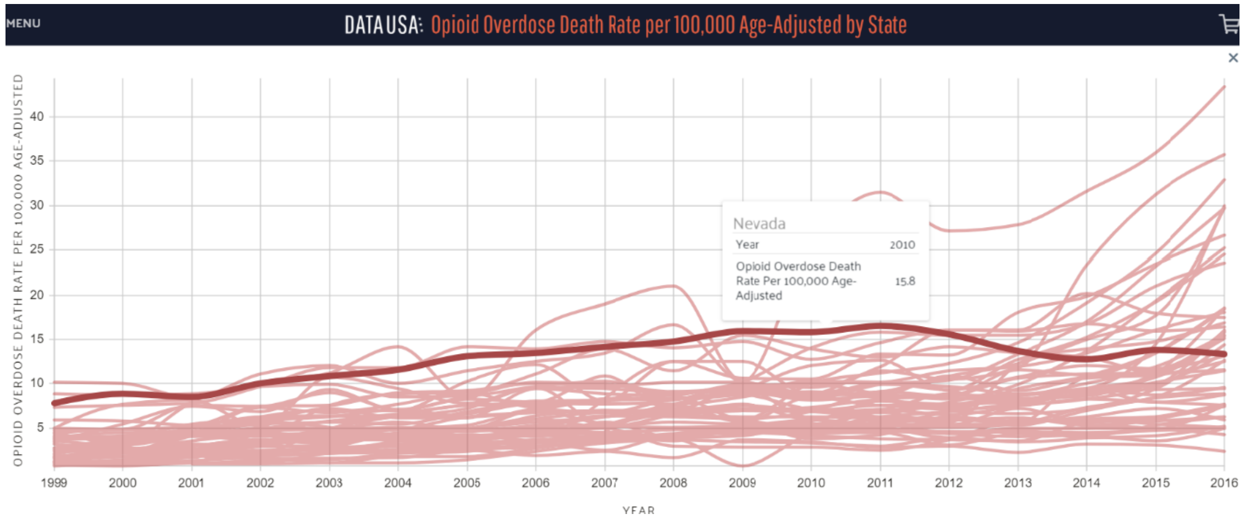
Fig. 1

The above graph quantifies the Opioid overdose death rate per 100,000 (Age-Adjusted) recorded in 51 states of the U.S. from 1999 to 2016. Age-specific death rates were applied to 2000 U.S. standard population age distribution to calculate age adjusted death rates. Each line in the plot represents a state. The graph on hover displays a card detailing the values of the X and Y axes at the hovered point as shown in Fig. 2. The intent of the graph is to visualize time-series change in the Opioid overdose death rate per 100,000 (Age-Adjusted) across 51 states in the U.S.

Strengths:

Accurate Data Representation: The raw data contains missing fields encoded as NSD (Not Sufficient Data) and NR (Not Reported). The graph did not plot any values or try to substitute the unavailable data with any other value. Additionally, an accurate numeric value is represented on hover.

Highlights the line on selection: On selection, it is easy to focus a line that represents a state.



Nevada
Year 2010
Opioid Overdose Death Rate Per 100,000 Age-Adjusted 15.8

Fig. 2

Weakness:

Data Over-plotting: As the data set is large, the lines in the plot have overlapped making it unreadable. To illustrate, due to the presence of large number of lines close to each other, the relationship between the X and Y values of different states cannot be comprehended.

Indistinguishable States: It's hard to identify a state from the others. For example, if a reader wants to spot a state, he needs to search for it in the messy plot by hovering over the lines. This can be a confusing, time consuming and an exhausting process.

Arduous to follow trend line: It's difficult to follow along the line of a state over the years by hovering and moving along with it. Moreover, when the lines are too close and overlap each other, it is very strenuous to perform this process.

Needless Curve Smoothing: Curve smoothing is observed which makes it difficult to recognize and read the peak values of the line graph.

Incomprehensible analysis at a glance: Graph doesn't aid in quick analysis of data in a single view. For example, looking at the graph, simple statistical information such the highest, lowest or median states for Opioid overdose death rates cannot be swiftly inferred.

Doesn't support concurrent analysis: Multiple states cannot be simultaneously analyzed. This is because hovering only highlights the line of a single state at a time. Also, with large number of states in sight, it is difficult to compare/analyze the death rates of different states at once.

Can work only with user interaction: The particulars of the graph heavily rely on user interaction. For example, the graph doesn't make any sense when printed on a paper. Additionally, it is very difficult and time consuming for people who are unfamiliar with interactive graphics.

Suggestions:

- Can make use of geographical context, as data is related to the states of the US.
- Spatial separation can be provided for the lines in the graph.
- Logical sub-groups of data can be constructed instead of visualizing it as a whole.
- Color can be used to distinguish different states.

Data Source:

https://www.kff.org/other/state-indicator/opioid-overdose-death-rates/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D

The data has been obtained from Kaiser Family Foundation (KFF) which is a non-profit organization that focuses on health issues in USA. It has been featured on the 'Mental Health and Substance Use' collection under the 'State Health Facts' section. As stated in their website, KFF had obtained this data from the 'Centers for Disease Control and Prevention', National Center for Health Statistics. The data can be accessed by making a data request to the CDC WONDER data base for 'Multiple Cause of Death, 1999-2017' (https://wonder.cdc.gov/controller/datarequest/D77).

Data Collection:

The data for Opioid overdose death rates can be obtained by clicking on trend graph and choosing the years 1999-2017 under the 'Timeframe' section displayed on the left side of the screen. The downloaded file consists of raw data with 20 columns, out of which the first column is 'Location'(State) and the other 19 columns denote the years from 1999 to 2017. Each row represents a state and its Opioid Overdose death rates across the mentioned timespan. Additionally, it was observed that other information regarding 'All Drug Overdose Death Rate (Age-Adjusted) per 100,000 Population', 'Percent Change in Opioid Overdose Death Rate from Prior Year' and 'Percent Change in All Drug Overdose Death Rate from Prior Year' for 1999 to 2017 were collected, and made a part of the data set by **an R program**. Further, this data set was used to obtain insights that are displayed through **Tableau**. Henceforth, 'per 100,000 Population' for 'Opioid overdose death rates' may not be explicitly mentioned in the document. All references to 'Opioid overdose death rates' can be considered as 'Opioid overdose death rates per 100,000 (Age-Adjusted)'.

Data Exploration:

The complete data set consists of 77 columns and 53 rows (header inclusive). Each row represents a state in the United States. It can be understood as 4 measures ('Opioid Overdose Death Rate (Age-Adjusted)', 'All Drug Overdose Death Rate (Age-Adjusted)', 'Percent Change in Opioid Overdose Death Rate from Prior Year' and 'Percent Change in All Drug Overdose Death Rate from Prior Year') x 19 years (1999-2017) x 51 States (of the USA).

The downloaded data consists of missing values or cells represented as NSD and NR. NSD stands for 'Not Sufficient Data'. It is also mentioned in the data source that NSD data was suppressed to ensure confidentiality. NR stands for 'Not Reported/Not Reliable'. The number of cells with NR are 32 and NSD are 187 in the total data set. In Opioid Overdose death rates from 1999-2017, the number of cells with NR are 24 and NSD are 22. For states such as Alaska, North Dakota, South Dakota and Wyoming, the number of cells containing NSD and NR are high. Hence, the analysis for such states is not accurate.

Data Preprocessing:

As a part of data cleaning, the cells with NSD and NR are replaced with the numeric zero. This was decided as a better option over replacing with mean, median, mode or any other numeric, because, a variable such as death is more impactful in a time series. For example, mean tends to shift towards the higher value and it may not accurately reflect the deaths in a particular year with the cell containing NSD or NR. Therefore, substituting such quantities makes the data unreliable. Moreover, columns such as '1999__Percent Change in Opioid Overdose Death Rate from Prior Year' contain all values as NSD as no previous data of 1998 is available for calculation. Hence, a safe option would be to replace all cells containing NSD and NR with a zero.

Additionally, the first row containing USA has been removed, because, it did not make any significant difference in understanding the data as a whole. As the data available was distributed across different files, **R code** was written to **merge different data frames** into a single data set. For data cleaning, **janitor** (Wickham, 2017), **tidyverse** (Firke, 2019) were used.

Discovery of misrepresentation:

During the process of data exploration, it was observed that values in columns related to 'Percentage Change in Opioid Overdose Death Rate from Prior Year' and 'Percent Change in All Drug Overdose Death Rate from Prior Year' represented only the 'growth rate' from the prior year and not the percentage change. For example, '2000__Percent Change in Opioid Overdose Death Rate from Prior Year' for Alabama has been represented as 0.25 instead of 25% but the column name states 'percent change'.

Feature Engineering:

In order to extract hidden knowledge from data certain columns were created as follows through an **R code**:

avg_opioid_death_growth_rate: This has been computed by calculating the mean of all the growth rates for Opioid overdose death rates for every state across 19 years (1999-2017).

avg_deaths_due_to_opioid: Obtained by performing the mean of Opioid overdose death rates over the time period (1999-2017) for every state.

avg_deaths_due_to_all_drugs: Calculated by taking the average of All drug overdose death rates for the time period (1999-2017) for every state

avg_opioid_death_rate_2009_2013: Computed by performing the mean of Opioid overdose death rate for the time period (2009-2013) for every state.

avg_opioid_death_rate_2013_2017: Computed by performing the mean of Opioid overdose death rate for the time period (2013-2017) for every state.

Redesigned Graph:

The line graph shown in Fig. 1 is inappropriate for the given data. As the data is large and must accommodate all the 51 states, Linked Micromaps are used to redesign the existing bad graph. As stated in (Carr et al., 1998), Micromaps enable the creation of small perceptual groups to simplify the visual appearance and simultaneously point to the geographical location linked to the statistical estimate. Micromaps can also show supplemental information. A sorting mechanism makes the Micromaps more readable. In such cases, the maps are separated by a median contour. The redesigned graph is shown in Fig. 3.

It can be observed that the sorting mechanism that was used is the most recent Opioid overdose death rate in 2017 for all the states. Cumulative Micromaps have been used to plot the time series data for 51 states. The median for the sorted panel occurs at Utah. Moving to the bottom of the plot, it can be noted that the states referenced previously, are outlined with black and colored in yellow. The current states in the group are represented by colors and are displayed in a legend adjacent to it.

The third column of the Micromap contains an x-axis that represents years from 1999-2017. The y-axis denotes Opioid overdose death rates of each state. A total of 51 states have been divided into small perceptual groups with each group containing 5 states denoted by five different colors. The criteria used for grouping is a sorting mechanism for the latest Opioid overdose deaths occurred in 2017 in the decreasing order. The fourth column of the Micromap picturizes the same. The x-axis represents 'Death Rate per 100,000' population. Each dot in the scatterplot is colored in accordance with the color of the state.

Construction of time-series Micromaps:

The redesigned graph was built with the help of R packages named **micromapST** (Carr et al., 2010). The required data was read into a data frame by using 'read.table' function. A time series object was constructed such that the first dimension represents 51 states of the U.S. The second dimension represents the time periods of the time series i.e., from 1999 to 2017. The third dimension consists of the Opioid overdose death rates. The x and y values of the third dimension i.e., TSdata[, , 2] was assigned to a variable called 'temprates'. The 'panelDesc' data frame was constructed by using ‘type’, ‘lab1’, ‘lab2’, ‘lab3’, ‘lab4’, ‘col1’ and ‘panelData’. The title was assigned to a variable (‘ExTitle’). The pdf file name where the Micromap must be stored is provided as an input to the pdf function. Finally, temprates, panelDesc, sortVar, ascend, title were provided to the micromapST function to construct the new graph. ‘sortVar’ determines the column index (year) of Opioid Overdose death rates which is used for sorting the states. ‘ascend’ determines the order of the sort (increasing or decreasing).

Strengths:

Avoids Data Over-plotting: All the 51 states of the US and their Opioid overdose death rates have been grouped into logical perceptual sub-groups, eliminating the problem of over-plotting in a single graphic.

Clear distinction of States: By visualizing the data in a spatial context, clear distinction between the states is achieved. Further, in the third column of the redesigned graph, a specific color has been assigned to each state. Hence, the reader can easily distinguish the trend line and the state associated with it.

Easy to follow trend lines: As there are only 5 states represented in a sub-group, each with a different color, it is relatively easy to follow the trend line along the time series.

Facilitates quick data analysis by sorting and linking: By looking at the Micromap, the reader will be able to get a quick understanding of the data and its patterns. This is made possible by the median value, usage of cumulative Micromaps and a sorting mechanism.

Promotes Concurrent Analysis: All the trend lines of 51 states have been plotted together in a single visual. Hence, concurrent analysis can be performed without much effort.

Supports plotting of additional Information: Apart from the trend lines represented in the bad graph, supplemental representation of the latest Opioid death rates is explicitly shown. Micromaps facilitate such additional rendering of data.

Minimal Memory Burden: It is not a strenuous task to locate different states. With the help of cumulative Micromap, at any given instant, it is simple to identify states that were previously referenced above.

Weakness:

Trend lines with minute variance: States with minute variance in the Opioid overdose death rates suffer from clear distinction but are still better than their representation in the bad graph.

Lacks user interaction: The redesigned graph doesn’t facilitate user interaction. Hence operations such as clicking, hovering cannot be performed.

Does not support custom color assignment: Micromaps do not provide the facility of assigning user defined colors to states or trend lines. Hence, colors meeting certain criteria such as color-blind friendly, print friendly etc. cannot be assigned.

Conclusion:

Keeping in view of the mentioned strengths and weaknesses of both the graphs (Fig. 1 and Fig. 3), it can be concluded that Micromaps is a more suitable technique to represent the data in discussion by providing a geo-spatial context and other supplemental information. However, the application of interactive Micromaps can enhance the current technique and provide a better visualization.

Extracting hidden knowledge from data with Tableau

Visualization aided by simple calculations can help in extracting a wealth of information from data. The below graphs exhibit the power of data visualization in knowledge extraction.

Heatmap:

The heatmap shown in Fig. 4, is plotted by using the ‘avg_opioid_death_growth_rate’ column (specified in Feature Engineering section). It can be observed that **‘New Jersey’** has the highest growth rate in Opioid overdose death rates followed by **‘West Virginia’**. Additionally, with the help of heat maps it is evident that the average growth rate in Opioid overdose death rates **increases from West to East**.

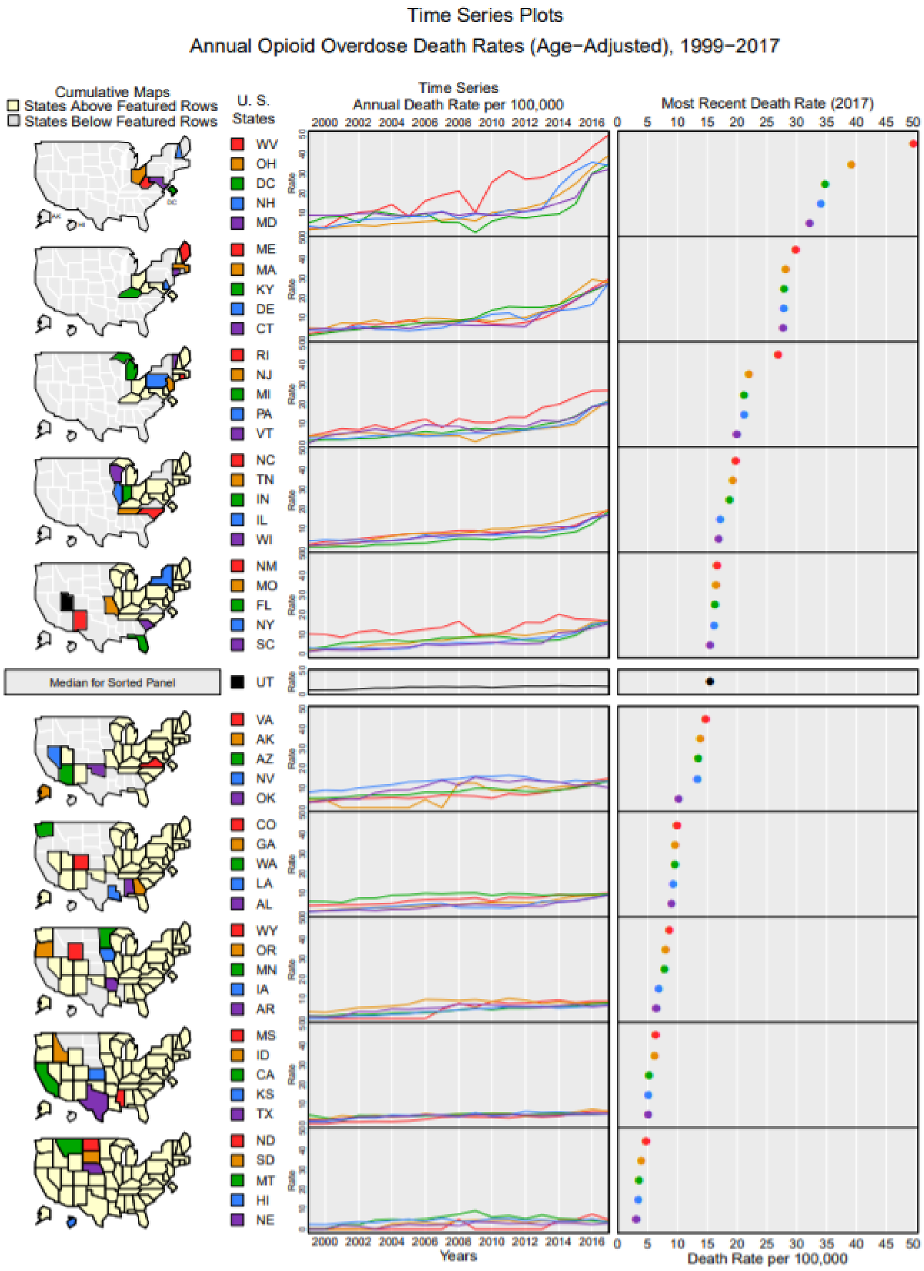


Fig. 3

Fig. 5.1 illustrates a row labelled plot representing the percentage of Opioid death rates amongst all Drug death rates for 100,000 population. It can be inferred that the top five states with the highest percentages are West Virginia, New Hampshire, Maryland, Massachusetts and Rhode Island.

Fig. 5.2 illustrates the time-series plot of the top 3 states (excluding West Virginia) identified in Fig. 5.1. A sudden **spike** is observed from 2013 to 2017 when compared to 1999 to 2013. Further investigation explaining this peculiar spike in the death rates has **shed light over the dangerous induction of synthetic drugs** in the market from 2013. News articles and research papers (Gladden et al., 2016), (Scholl et al., 2018), (Stephens, 2019) confirming the same have been published.

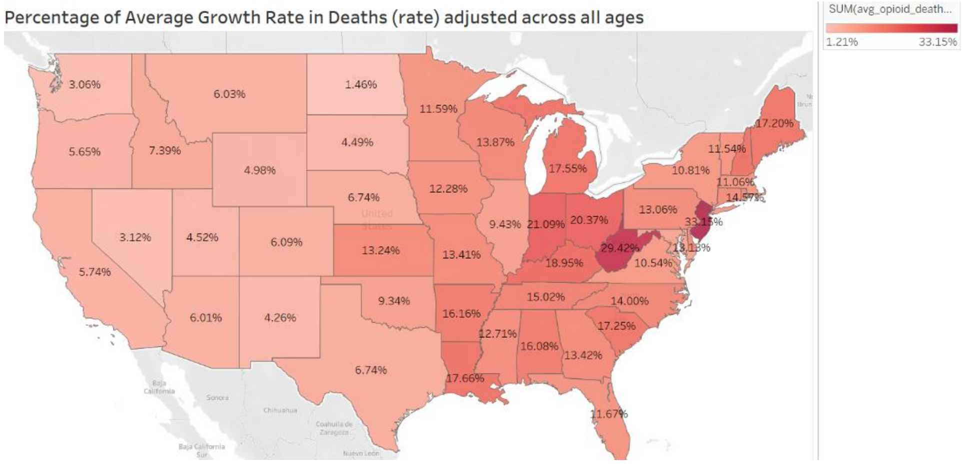


Fig. 4

Out of curiosity, Fig. 5.3 was plotted which depicts the percentage change in the average number of deaths between 2009-2013 and 2013-2017 by using the ‘avg_opioid_death_rate_2009_2013’ and ‘avg_opioid_death_rate_2013_2017’ columns generated during feature extraction. This led to the discovery that ‘Washington D.C’ had the highest percentage change followed by Connecticut, New Hampshire, New Jersey and Maine. Additionally, it was also recognized that Montana, Oklahoma, Nevada, Oregon and Hawaii had the lowest percentage increase between the stated years. This is depicted in Fig. 5.3 and 5.4 (top 3 states).

All the colors represented in the below graphs are **color-blind friendly, B&W friendly, print friendly and laptop friendly.**

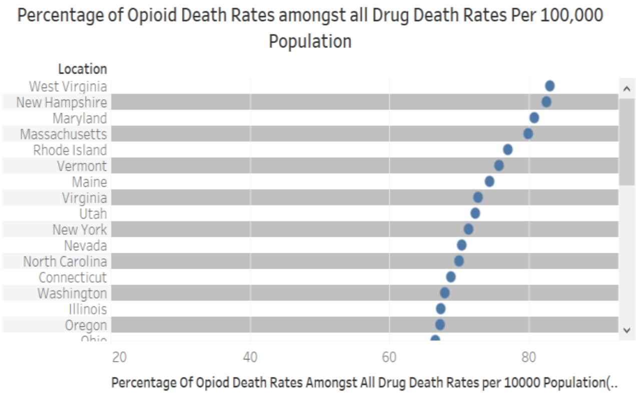
Challenges:

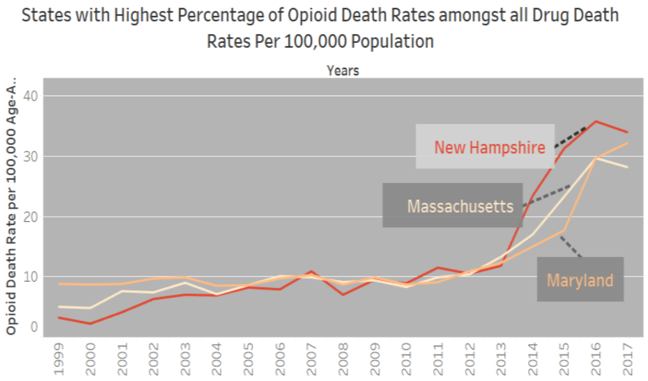
- Comprehending the statistical terms associated with the data set and its domain was a challenge. Acquired the knowledge of statistical concepts linked with the metadata by from various online resources such as ‘Centers for Disease Control and Prevention’ (Drug Overdose Deaths, n.d.).
- Data cleaning and pre-processing was an arduous task. R packages such as ‘tidyverse’ and ‘janitor’ were used to make the process easier.
- The data format for generating time series plots is complex. A great deal of effort had been put into bringing the data into the required format and building the time-series object. The documentation of R package ‘micromapST’ and TSdata (Time Series Example Dataset) (Carr et al., 2010) had been thoroughly studied and analyzed to achieve the perfect object.
- Plotting of time-series data as line plots within the Micromaps is a difficult task. The description and the examples present in the documentation of ‘micromapST’ (Carr et al., 2010) aided in better understanding, development and eventual realization of the redesigned graph (Fig. 3).
- Picking color-blind friendly, B&W friendly and Laptop friendly colors was a challenge. Data has been sorted and colored sequentially to accommodate the criteria. This is depicted in Fig. 5.2 and Fig. 5.4.

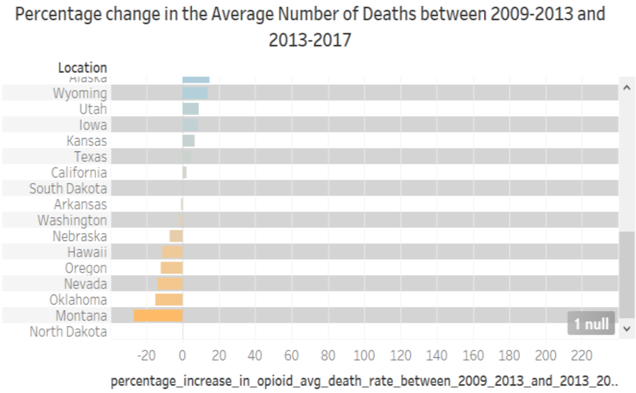
Special Efforts:

- Gathering distributed data and merging them to form a single dataset. Additional data of 2017 was collected and included in the Micromap but was not present in the bad graph.

- Feature Extraction by constructing columns using statistical formulas to gain additional knowledge and represent acquired insights through visualizations.
- Constructing time-series objects to be suitable for building Micromaps in R.
- Searched for supporting documents and research papers to strengthen the findings obtained from the graphics (Fig. 5.2).
- Made significant efforts to find and develop the best suitable graph to present the information. For example, Fig. 5.1 can be represented as bar graphs. However, **with the techniques discussed in class**, it was converted to a dot plot. Also, proximity labelling was provided for Fig 5.2 and 5.4.
- The graph was transformed to accommodate color-blind friendly, B&W friendly and Laptop friendly colors.


Percentage of Opioid Death Rates amongst all Drug Death Rates Per 100,000 Population


States with Highest Percentage of Opioid Death Rates amongst all Drug Death Rates Per 100,000 Population


Percentage change in the Average Number of Deaths between 2009-2013 and 2013-2017

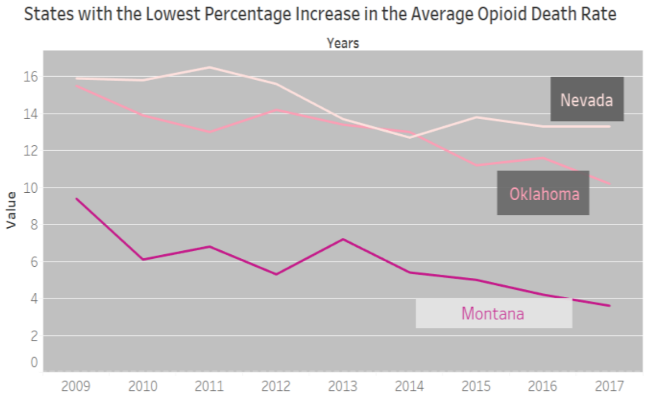

States with the Lowest Percentage Increase in the Average Opioid Death Rate

Fig. 5.1 – Percentage of Opioid Death Rates amongst all Drug Death Rates
Fig. 5.2 – States with Highest Percentage of Opioid Death Rates amongst all Drug Death Rates
Fig. 5.3 – Percentage in the Average Number of Deaths between 2009-2013 and 2013-2017
Fig 5.4 – States with the Lowest Percentage Increase in the Average Opioid Death Rate

Project Log

Chaithanya Pramodh Kasula

- Performed data pre-processing which includes merging, cleaning and reorganization of data using R.
- Construction of new features from the existing data during the process of Feature Engineering using R.
- Preparing data to be compatible with the specifics of the time series object in R.
- Redesigning the bad graph and construction of time series plots by using Linked Micromaps.

Aishwarya Varala:

- Investigation and collection of data source for the identified bad graph.
- Construction of Heatmap for Percentage of Average growth rates in deaths adjusted across all ages in Tableau.
- Generation of dot plot to represent the percentage of Opioid death rates amongst all Drug death rates using Tableau.
- Construction of frequency polygons to represent 'Percentage in the Average Number of Deaths between 2009-2013 and 2013-2017' and 'States with the Lowest Percentage Increase in the Average Opioid Death Rate'.

Collective efforts:

- Identification and analysis of bad graph.
- Data exploration for the collected data.
- Discovery of misrepresentation.
- Analysis of redesigned graph.
- Investigation of spike observed in Fig. 5.2.
- Report writing.

REFERENCES

Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

Sam Firke (2019). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 1.2.0. https://CRAN.R-project.org/package=janitor

Carr D.B., Olsen, A.R., Courbois, J.P., Pierson, S.M and Carr, D.A. (1998). Linked Micromap Plots: Named and Described, Statistical Computing and Statistical Graphics Newsletter 9(1):24-32.

Daniel B. Carr and Linda Williams Pickle (2010), Visualizing Data Patterns with Micromaps.

Gladden, R. M., Martinez, P., & Seth, P. (2016). Fentanyl Law Enforcement Submissions and Increases in Synthetic Opioid–Involved Overdose Deaths — 27 States, 2013–2014. MMWR. Morbidity and Mortality Weekly Report, 65(33), 837–843. doi: 10.15585/mmwr.mm6533a2

Scholl, L., Seth, P., Kariisa, M., Wilson, N., & Baldwin, G. (2018). Drug and Opioid-Involved Overdose Deaths — United States, 2013–2017. MMWR. Morbidity and Mortality Weekly Report, 67(5152). doi: 10.15585/mmwr.mm6751521e1

Stephens, E. (2019, October 2). Opioid Toxicity.

Retrieved from https://emedicine.medscape.com/article/815784-overview

Drug Overdose Deaths. (n.d.).

Retrieved from https://www.cdc.gov/drugoverdose/data/statedeaths.html