Relationship between homicide deaths and opioid drug overdoses in the US between 1999-2014

Ansar Kokbayev, Tymur Antonovskyi, Attila Szuts

2020.12.11

Distribution of Work:

ETL in MySQL: Ati

API fetch: Ansar

Data preparation for Analysis: Tymur

Analysis: Ati, Tymur

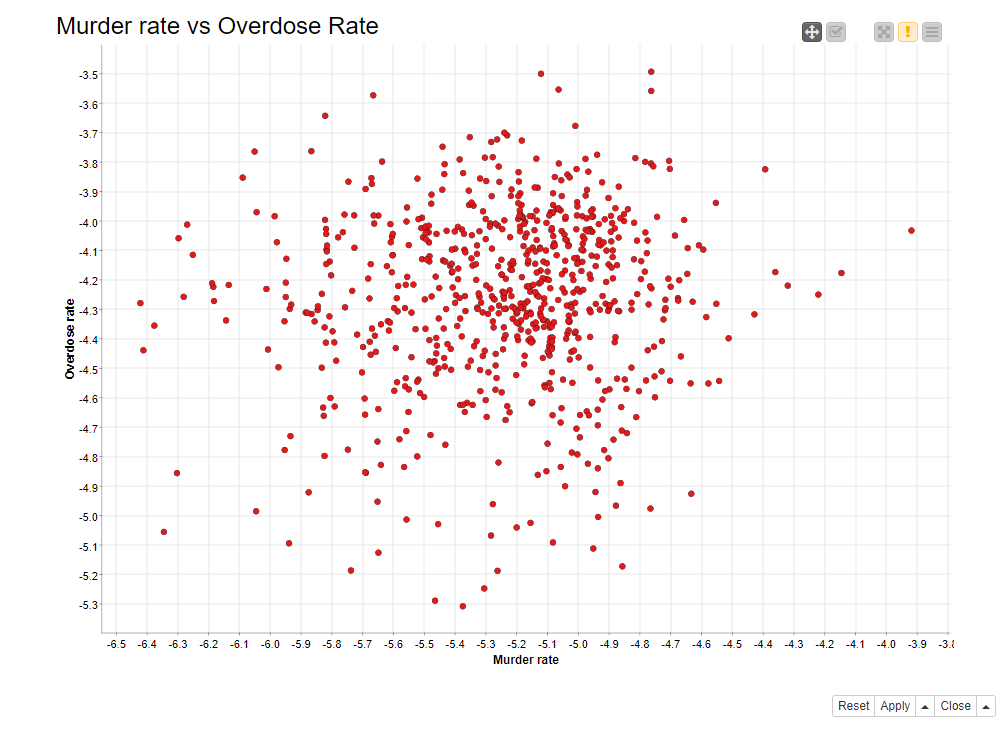
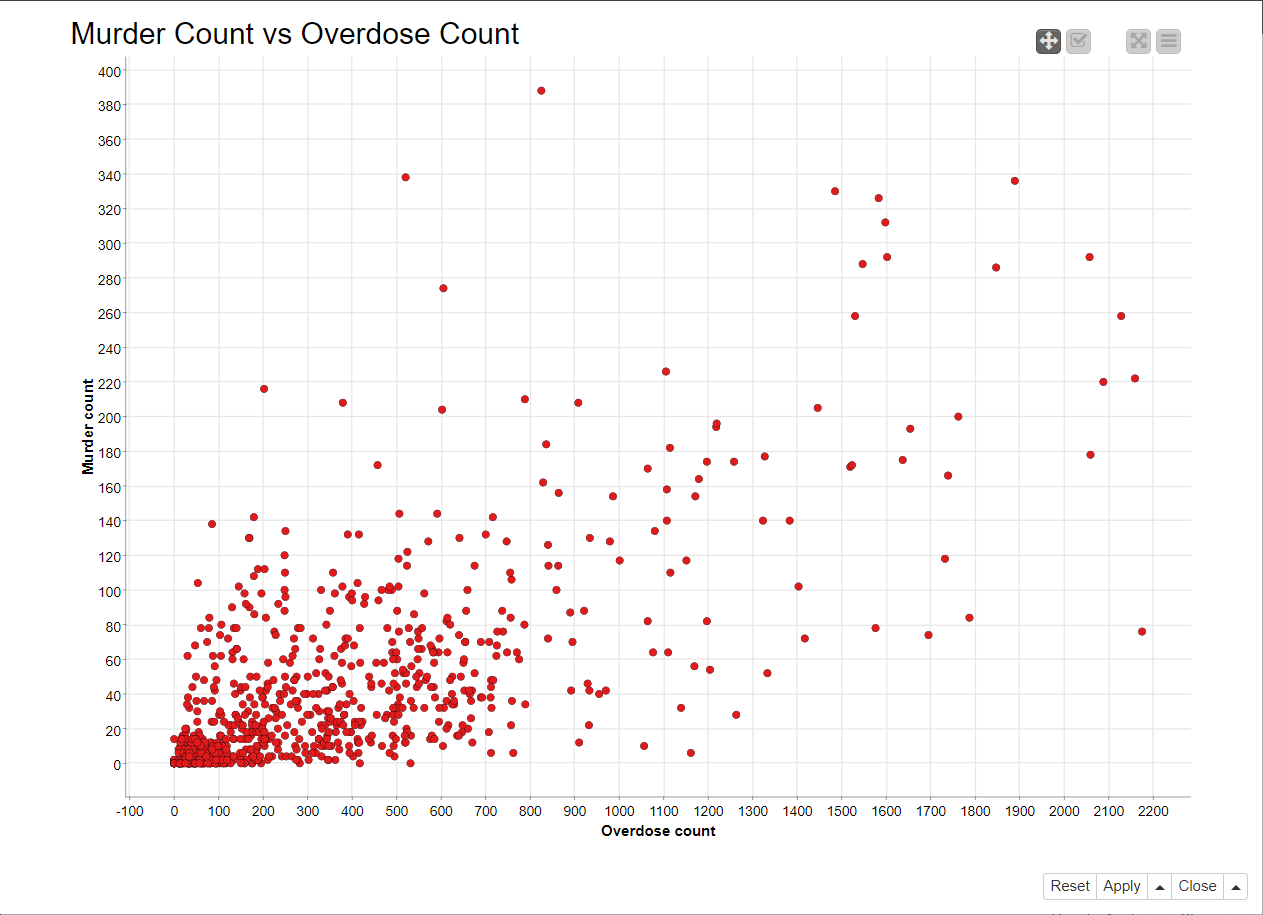
Report: Everyone

Presentation: Ansar

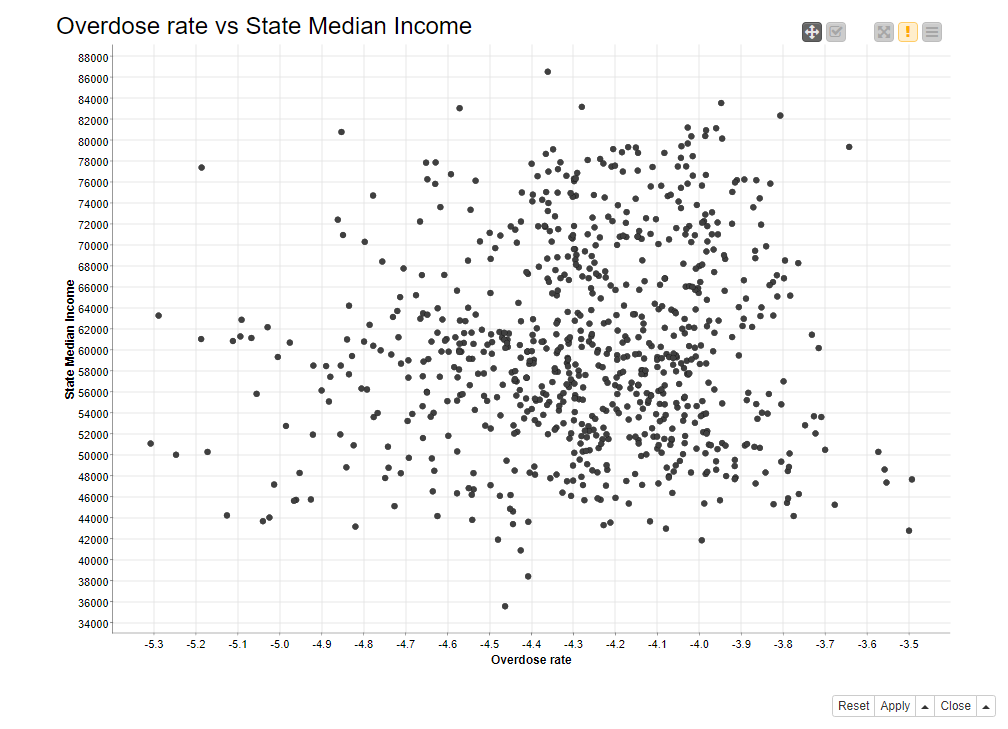
# The purpose of the term project

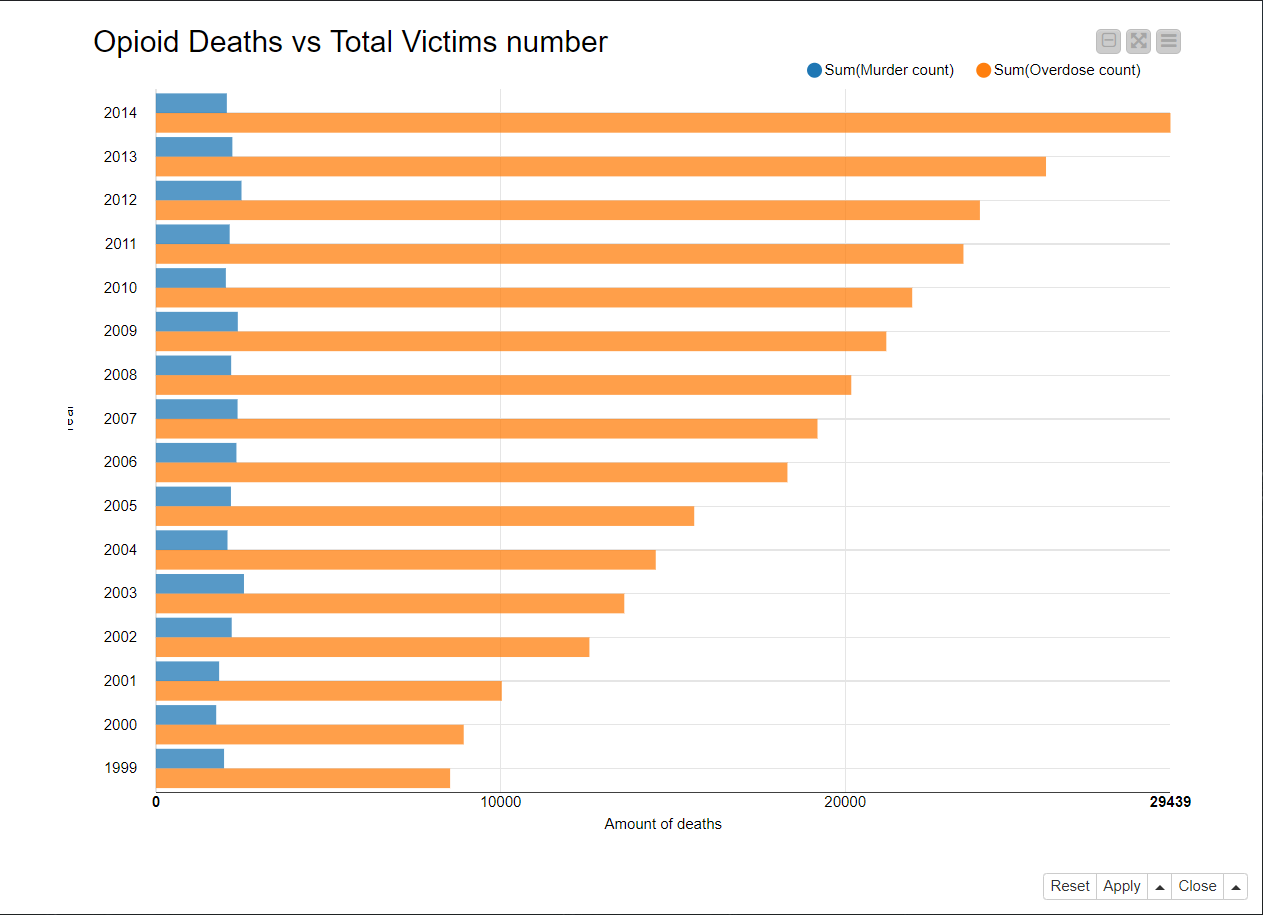
Extract, load and analyze the information on homicide rates and opioid overdose deaths in the US by state between 1999 and 2014. Distinguish any relationships between deaths and annual household income by each state for the same period.

# Key findings

First we created a correlation between Murder count in each state and year, and Overdose count and found a strong correlation between the two. However we realized this is because we did not control for population. Also, we noticed that the distribution of our variables were skewed to the left, so log transforming it would help us find the linear relationship between the two, assuming that it exists. So after this two key insights, we created two new variables: log Overdose rate and log Murder rate. We created these by dividing total murder and overdose counts with population of the state in a given year and then taking the log of it.

But even with these modifications, we haven’t found any meaningful relationship between the two variables. We also investigated correlation between median household income and overdoses. We have found similar correlation coefficients here as well.



Although overdoses and murder rates are mean independent, we can conclude some meaningful insights. Since we can see that overdose rates have been steadily increasing and murder rates have not, we can conclude, that drug use dose not cause violence. On the other hand, not finding correlation between income and overdoses could be simply because our data is too high-level. It could very well be that in poorer neighbourhoods overdose rates are higher than in more well-off ones, but since our data is on a US state level, we can not investigate this.

# How to execute analysis

* Set up database:
  + Move raw data files to the secure file priv location
  + execute code/load-data.sql
* Open knime workflow file in knime/murder-overdose.knwf
  + Use database’s password for knime database connection
  + Execute all nodes

# Our work process

First week:

* Brainstorming research questions. Potential candidates were League of legends matches, Car accidents and Opioid overdoses
* Choosing topic: Relationship between opioid overdoses and homicide rates.
* Data collection and cleaning

Second week:

* Creating analysis and visualisation to investigate our research question.
* Enriching our data with median household income using API

Third week:

* Creating report, presentation and documentation
* Quality and reproducibility control, naming
* Presentation

# Data sources

Real Median Household Income by State, Annual

<https://fred.stlouisfed.org/release/tables?rid=249&eid=259515&od=>

Opioid Overdose Deaths

<https://www.kaggle.com/jazzang/opioid-overdose-deaths>

U.S. Homicide Reports, 1980-2014

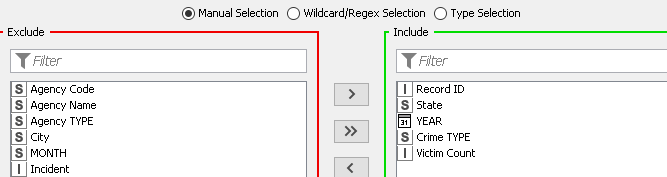
<https://www.kaggle.com/jyzaguirre/us-homicide-reports>

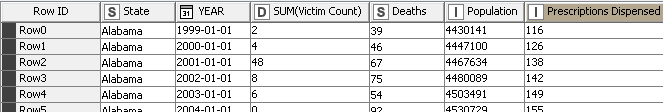
# Data preparation

## Data collection

We downloaded the Opioid Overdose Deaths by state from 1999 to 2014 and U.S. Homicide Reports by county for the same period. Then we used MySQL Workbench to create a database for two tables. Both of the tables were inserted in two separate schemas: murders and opioid\_deaths (638454 and 816 rows respectively).

Later, we used the “MySQL Connector” node in KNIME to import the data there. We connect to our mysql database and with the help of “DB Table Selector” we import both of our tables.

Both of the original tables contain a lot of columns that may not be useful for our research. We only pick the ones we need in “DB Column FIlter”.

Now, since the research is aimed only at homicides (murder), we use “DB Row Filter” to pick only deaths that were ruled as “Murder or Manslaughter”. Later, we group the data by state, as there are a lot of counties. Finally we join all the data in one final table.

Before we go any further, we need the Median Household Income data. We retrieve it from the Federal Reserve Bank of St.Louis website. Every year of the data is on a separate page, unlike the data we had on a .csv file. We decided to use API, as it would be much more efficient. The only problem with the GET request, we needed multiple URLs with different observation\_date(s).

On a sample URL below we needed to change the last number (1990-01-01):

<https://fred.stlouisfed.org/release/tables?rid=249&eid=259515&od=1990-01-01#>

Using a loop that runs through the list, we managed to pull the data accordingly using the “GET Request”.

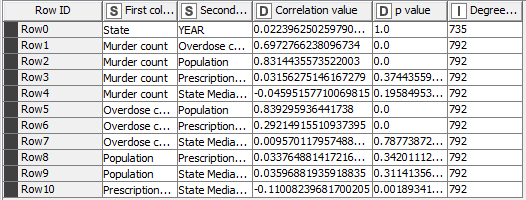


We used two “JSON to Table” to divide the data in a proper table with the state, year and median income as columns. Later we join it with the original data using “Joiner”.

## Data preprocessing

As the first step, we aimed to pre-process our data and make sure that there are no missing values, variables are set in proper form, values are normalized and ready for use. Firstly, we need to substitute “Suppressed” notation by 0 in the Overdose count column,using String Manipulation Note. Besides, the values in some columns were stored as strings so we fixed it by converting them into numeric values.

# Correlation analysis

****

As the next step, we started to explore the relationship between the total number of overdoses and the total number of murders in each US state. Accordingly, the correlation function was applied. Based on features, we identified the relationship between Murder Count and Overdose count variables, equals to 0.7 and can be treated as a strong relationship. As we expected, the analysis revealed the strong relationship between population and the number of murders and overdoses.It means that with increase in population the number of murders and overdoses cases will increase too. In order to eliminate the effect of population, we will create per capita deaths in the following step. Also, in this step we are also going to log transform our variables.



We can see here, that there is a significant 0 correlation, that is there is no relationship between murders and opioid overdoses.