**Insight into the Opioid Epidemic from Rates of Prescription and Overdose Deaths**



**Team Members:**

* [Chahnaz Kbaisi](https://github.com/Chahnaz-Kbaisi)
* [Susan Thomas](https://github.com/SusanCThomas)
* [Prarthna Ashutoshmunidottir](https://github.com/Prarthna-design)

**Table of Contents:**

1. Project Description
2. Research Areas
3. Project Key Areas of Investigation  
   - Part 1  
   - Part 2  
   - Part 3
4. Extract
5. Transform
6. Load
7. Process Limitations
8. Data Analysis
9. Languages and Libraries

**Project Description:**

This project assesses the prescribing of opioid drugs across the US from 1999 to 2014, focusing on changes in prescription rates at Veterans Affairs sites from 2012 to 2018, and details relevant information on the types of medical facilities and dispensers that provide them. It describes the rate of usage by race, education and family income. The analysis also shows the rates of overdose that have led to death in the US. Using claims for medically-assisted treatment and deaths from overdose from counties in California, we will determine if there is an association between seeking treatment for opioid abuse and subsequent death.

**Research Areas:**

After picking a topic, we developed some key questions to inform our data search. Then, we narrowed down the unfiltered data to a few datasets for each part to then take a closer look and pick the sets that could potentially be analyzed to answer our key questions. We used a team approach by which every member was able to engage in each of the steps so that learning, usage and application would be as comprehensive as possible. Each team member studied their respective dataset and then downloaded the files to begin the extraction step.

Each team member's part has a separate jupyter notebook which contain the respective three ETL steps. Below are the three links to for each part:

* [Part 1 Jupyter Notebook](https://github.com/Chahnaz-Kbaisi/ETL-Project/blob/main/Part_1/Opioid_Pres.ipynb)
* [Part 2 Jupyter Notebook](https://github.com/Chahnaz-Kbaisi/ETL-Project/blob/main/Part_2/average_dispensed.ipynb)
* [Part 3 Jupyter Notebook](https://github.com/Chahnaz-Kbaisi/ETL-Project/blob/main/Part_3/usage_overdose_us.ipynb)

**Project Key Areas of Investigation:**

**Part 1.** Assessing Opioid Prescriptions across the US from 1999 to 2014

Key Question Part 1 - What is the rate of opioid prescriptions across the US from 1999 to 2014?

Data Sources Part 1 - [Part 1 Resources with all datasets](https://github.com/Chahnaz-Kbaisi/ETL-Project/tree/main/Part_1/Resources)

[Opioid Prescribing Rates at VA Facilities 2012 - 2018](https://catalog.data.gov/dataset/opioid-prescribing-rates-at-va-facilities-2012-2018)

[Opioid Prescriptions Dispensed](https://data.world/jessicayung/opioid-prescriptions-dispensed)

**Part 2.** Information on Opioid Dispensers in the US

Key Question Part 2 - What is the relevant information about dispensers and medical facilities?

Data Sources Part 2 - [Part 2 Resources with all datasets](https://github.com/Chahnaz-Kbaisi/ETL-Project/tree/main/Part_2/Resources)

[Opioid Overdose Deaths](https://data.world/health/opioid-overdose-deaths)

[U.S. Opiate Prescriptions/Overdoses](https://www.kaggle.com/apryor6/us-opiate-prescriptions)

**Part 3.** Demographic information on opioid usage and overdose death across the US

Key Questions Part 3 -

- What is the prevalence of overdose deaths across the US from 1999 to 2014?

- What are the rates of usage organized by demographics across the US?

- Using California as a model, is there a relationship between enrollments in medically  
 assisted facilities and rates of overdose deaths?

- What are the total number of opioid overdose deaths for each state between 1999 and  
 2018?

Data Sources Part 3 - [Resources with all datasets](https://github.com/Chahnaz-Kbaisi/ETL-Project/tree/main/Part_3/Resources)

[Opioid Overdose Deaths](https://data.world/health/opioid-overdose-deaths)

[Drug Use, Employment, Work Absence, Income, Race, Education](https://data.world/balexturner/drug-use-employment-work-absence-income-race-education)

[Medication-Assisted Treatment in Medi-Cal for Opioid Use](https://data.world/chhs/8329a339-ab77-4d05-ab7a-405d0ae5765c)

[Opioid related deaths by state with demographics](https://wonder.cdc.gov/controller/datarequest/D76)

Note: Usage of the CDC Wonder database requires acceptance of the sites rules. Once in the form, user must provide search criteria for the specific data types and the categories of inclusion. The data herein was built with the following inclusion criteria: state, year, gender, race, and mortality due to opioid overdose (ICD-40).

**Extract:**

* Extraction of data was from csv source files
* Deep exploration of data content
* Data retrieval for extraction and writing code
* Read csv files and saved into pandas dataframe

**Transform:**

* Transformed data to accommodate project objectives
* Cleaned the data and removed the null
* Selected and renamed the needed columns
* Removed columns of data that were irrelevant
* Created the new dataset with selected columns
* Checked the data types to prevent errors during loading

**Load:**

* Selected postgres to load the datasets into a database
* Created three different databases, one for each project part
* Adjusted ETL pipeline for ease of use and navigation to the database for analysis
* Created connection to postgres during loading
* Created tables in pgAdmin localhost

**1.** [Part 1 Tables](https://github.com/Chahnaz-Kbaisi/ETL-Project/blob/main/Part_1/part_1_tables.sql)

**2.** [Part 2 Tables](https://github.com/Chahnaz-Kbaisi/ETL-Project/blob/main/Part_2/opioid_query.sql)

**3.** [Part 3 Tables](https://github.com/Chahnaz-Kbaisi/ETL-Project/blob/main/Part_3/usage_overdose_us.sql)

* Loaded the new datasets into the tables

**Process Limitations:**

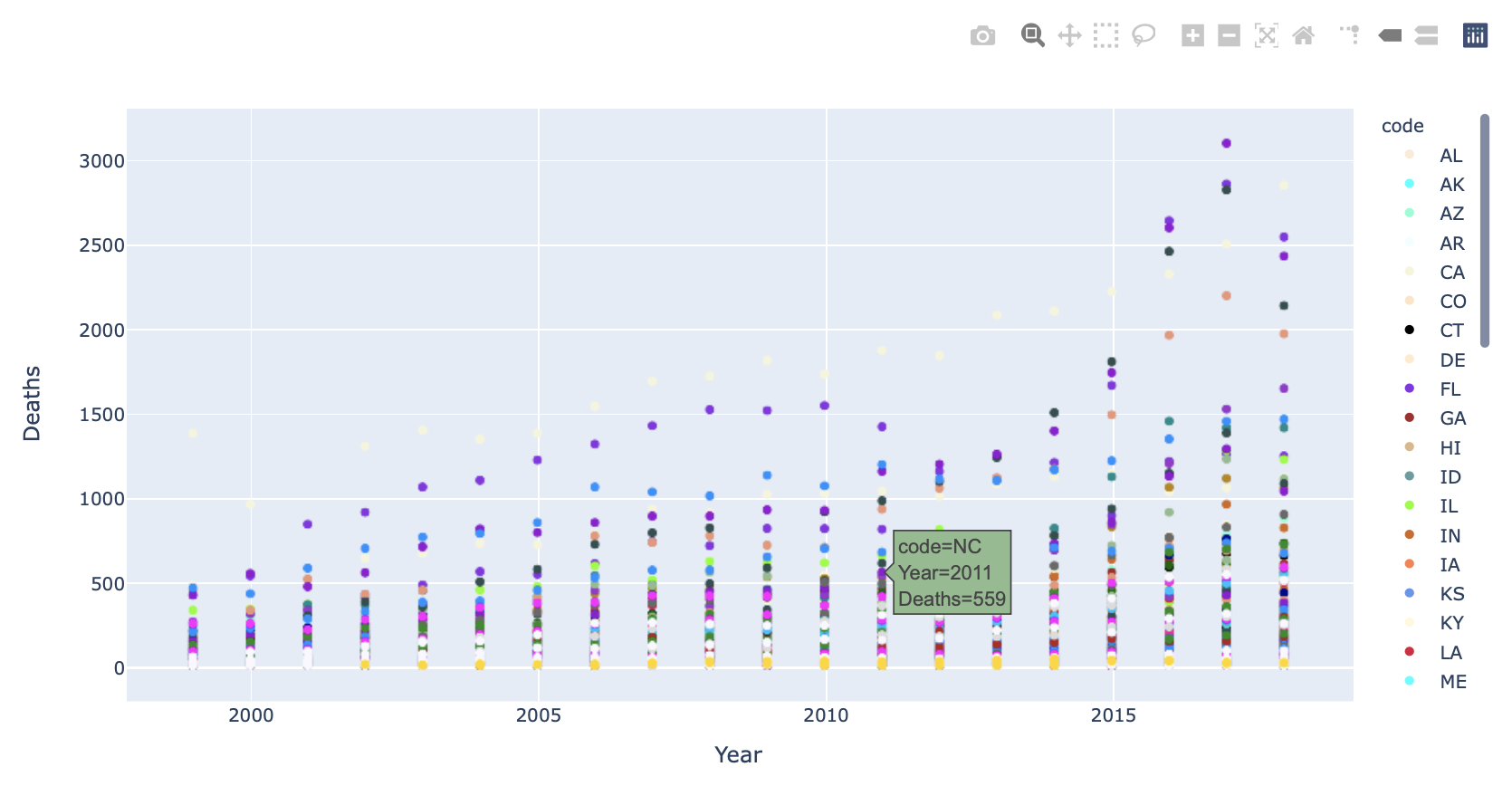
* Identifying datasets that contained appropriately identified opioid usage classification
* Reconciling variations in mortality indicators
* Dealing with missing data and nominal data parameters that needed to transformed to Boolean

**Data Analysis:**

* Analyzed data to answer parts of key questions
* Combined data from different datasets that were misaligned
* Reconciled variations in data forms to ensure proper joining
* Created plots based on joined data and built dataset

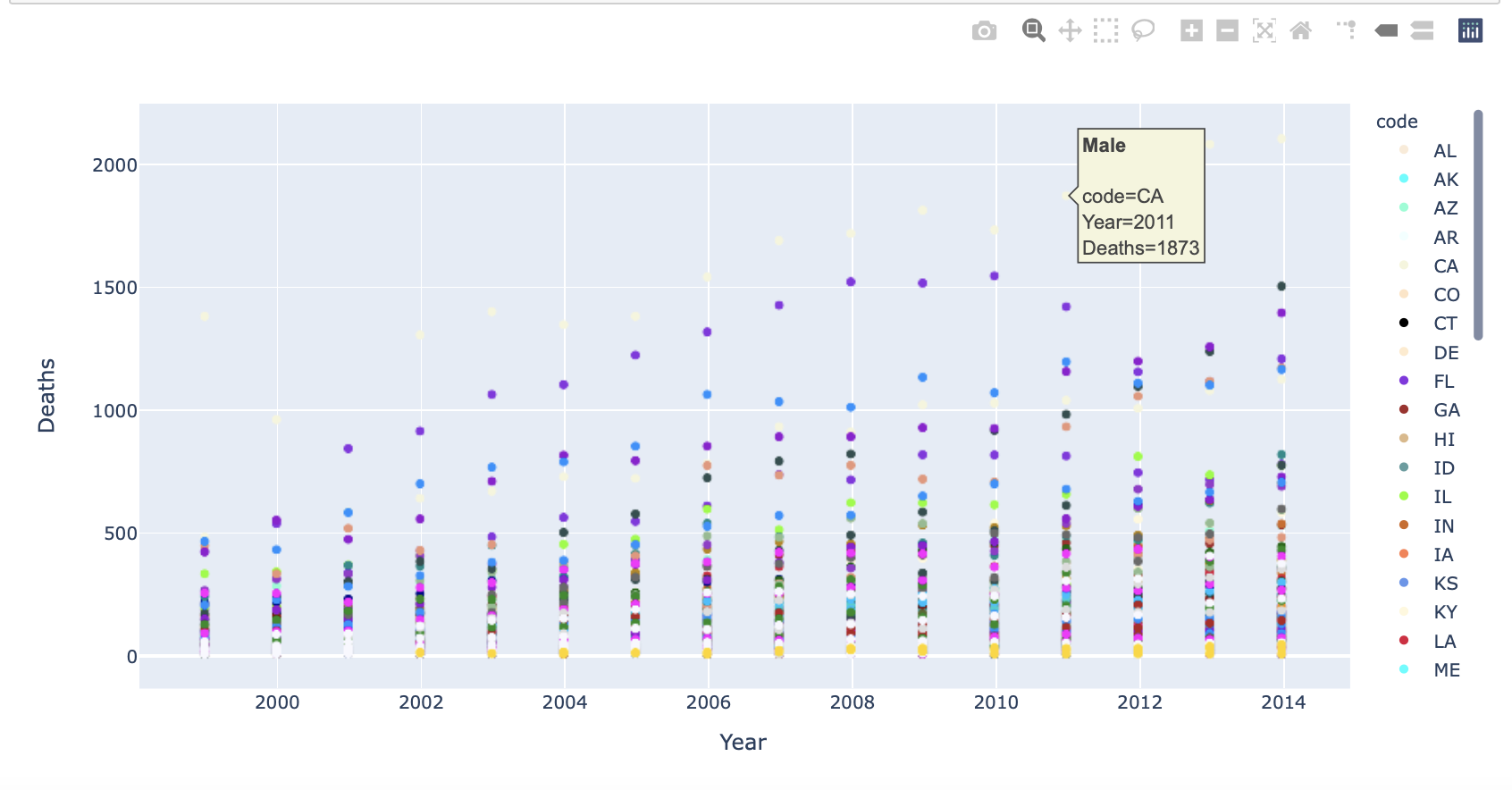
**Part 3 Analysis:**

**Plot 1.** The below Plotly Express shows the number of annual deaths between 1999 & 2018 for all states. This data was acquired from the CDC's wonder database by querying specific parameters and using the ICD-10 codes for opioid related deaths.



Plot Summary: In this plot, hovering over individual data points provides details on each state’s overdose death by year. Between 1999 and 2018, California had the highest number of death and the greatest increases over that time span. On the other hand, Colorado appears to have the lowest death rates over approximately the same period.

**Plot 2.** The below Plotly Express shows the number of annual deaths between 1999 & 2014 for all states by gender.



Plot Summary: The plot above shows the number of deaths by gender in each state from 1999 to 2014. Males in both California and Florida had the most deaths over that period with a doubling of the numbers.

[For further analysis of part 3 please refer to Part 3 Jupyter Notebook](https://github.com/Chahnaz-Kbaisi/ETL-Project/blob/main/Part_3/usage_overdose_us.ipynb)

**Languages and Libraries:**

* Jupyter Notebook
* PostgresSQL
* SQLAlchemy
* Python
* Pandas
* Plotly Express