Data Driven Decision Making

Data-Driven Response to High Opioid Abuse Impact: Leveraging Machine Learning, AI, Health, Social and Economic Determinants

Student Name(s):
Ana Maria Calderon
Iliana Joaquin
Zhi Long Li
Justin Shields
Zachary Thomson

School of Graduate Professional Studies
MPS/MS in Data Analytics
DAAN 881 – Data Driven Decision Making

Spring, 2025

Document Control

Work carried out by:

Name	Email Address	Task description
Zachary Thomson	zjt5163@psu.edu	• Deliverable document creation and editing coordinator • Teams discussion and collaboration • Database creation & administration
Ana Calderon	amc9797@psu.edu	• Contributed to the overall report by researching and adding information to various sections • Collaborated with team members to identify information gaps and address them through targeted research and content creation • Data collection, visualization • Formatted and cleanup document • Helped with the data cleaning section • Worked on discussion section to address bias • Analyzed data using Python and RStudio
Iliana Joaquin	ijj5077@psu.edu	• Contributed to the revisions of problem statement, prescriptive questions, and defining the timeline • Teams discussion and collaboration • Data collection • Citation additions/edits
Justin Shields	jjs8575@psu.edu	• Problem statement • Data Source descriptions • Data preparation and potential issues descriptions • Teams discussion and collaboration • Analyzed data using Python and RStudio
Zhi Long Li	zml5452@psu.edu	• Potential timeline creation and CRISP-DM control, creation of Feasibility Estimate. • Teams discussion and collaboration • Graph creation and narrative of results • Database description • Analyzed data using Python and RStudio

Revision Sheet

Release No.	Date	Revision Description
v1	01/26/2025	First draft - deliverable 1
v2	02/02/2025	Second draft – deliverable 1
v3	02/09/2025	Third draft – deliverable 1
v4	02/09/2025	First draft – deliverable 2
v5	02/16/2025	Second draft – deliverable 2
v6	02/22/2025	Third draft – deliverable 2
v7	02/23/2025	Fourth draft – deliverable 2
v8	02/23/2025	First draft – deliverable 3
v9	03/09/2025	First draft – deliverable 4
v10	03/30/2025	First draft – deliverable 5
v11	04/13/2025	First draft – deliverable 6, revisions on deliverable 5

Table of Contents

Executive Summary

The opioid crisis remains a pressing public health issue in Pennsylvania, with disparities in access to harm reduction and treatment services affecting communities in significant ways. While harm reduction strategies like Naloxone distribution and syringe programs have public support, their availability is inconsistent across urban and rural areas. This geographic and socioeconomic imbalance limits treatment accessibility exacerbates overdose risks, and strains healthcare systems. Our research aims to identify key factors influencing access to harm reduction services and develop data-driven recommendations to improve intervention efforts across the state.

This study utilizes publicly available data from various Pennsylvania governmental sources, including the Department of Health, Department of Corrections, and the Pennsylvania Attorney General's Office [1],[3]. By applying Principal Component Analysis (PCA) and predictive modeling, we will assess where service gaps exist and explore how economic and geographic factors influence harm reduction accessibility [5],[6].

Our hypothesis is that regions with lower access to harm reduction services will exhibit higher overdose rates and negative economic impacts, particularly in rural areas. We also anticipate that by identifying patterns in service distribution, we can recommend targeted interventions that maximize the effectiveness of public health resources.

By leveraging advanced data analysis techniques and collaboration with public health stakeholders, this research will provide actionable insights to guide policy recommendations, resource allocation, and community-level harm reduction strategies. Our findings will help pinpoint areas most in need of intervention and propose data-driven solutions to mitigate the ongoing opioid crisis in Pennsylvania.

Problem Statement

Opioid misuse remains a significant public health crisis in Pennsylvania, yet access to harm reduction and treatment services varies significantly across regions. Disparities in availability, particularly between urban and rural areas, can hinder access to care, increase overdose risks, and strain healthcare resources [7]. Understanding the geographic, economic, and social factors influencing access is essential to improving intervention strategies. The study, using publicly available data from Pennsylvania government sources, aims to identify service gaps, analyze the impact of harm reduction efforts, and provide data-driven recommendation to enhance accessibility and reduce opioid-related harm across the state [8].

Objectives

Understanding Disparities in SUD Treatment Access (formulated based on [8],[3])

- Are there areas in the state where SUD reduction and treatment services are harder to access?
- Where are harm reduction services (like Naloxone and syringe programs) located across Pennsylvania?
- How does being in a rural or urban area affect the ability to access harm reduction programs?

Socioeconomic and Geographic Factors Affecting Access (formulated based on [12],[6])

- What social, economic, or geographic factors impact where SUD reduction and treatment services are available?
- How do overdose rate impact GDP in Pennsylvania region (workforce, healthcare cost and production capability)?

The Impact of Harm Reduction on Overdose Rates (formulated based on [5],[7])

- Is there a link between access to harm reduction services and overdose rates in different areas of the state?
- Can we predict how increasing access in underserved areas would impact overdose prevention efforts?
- What are the key factors influencing the relationship between overdose rates and harm reduction efforts?

Developing Data-Driven Recommendations (formulated based on [14],[9])

- What area/region should be focused upon to reduce SUD mortality or increase educational efforts?
- What factors contribute to opioid users declining SUD treatment services, and how can these barriers be addressed?
- What type of harm reduction strategy (e.g., Narcan distribution, syringe programs, overdose prevention centers) would be most effective in specific underserved regions?

Data and Methodology

Resources Available

Much of the data was collected from the Commonwealth of Pennsylvania, specifically:

- Overdose information [3]
- Dispensation data [4]
- Emergency Medical Services (EMS) Naloxone dose data [5]
- Inmate Admissions with Substance Use Disorder [6]
- Pennsylvania State Police opioid seizures and arrests data [7]
- Socioeconomic data [12]

Expert research with the Penn State Harrisburg research team [14]:

- Access to academic databases and research literature
- Potential collaboration opportunities with public health agencies, harm reduction organizations, and community stakeholders

Data Collected

All data collected is a matter of public record and was found on various Pennsylvania governmental websites, including the Department of Health, Department of Corrections, and the Pennsylvania Attorney General's Office. It is being saved both locally on team members' hard drives, and communally in our Microsoft Teams Group. It will be shared alongside the project to compliment the reader's understanding of our research and findings. This data is being reformatted and visualized using KNIME, a data analytics platform, to produce corresponding visualizations of our findings [4].

Due to the non-sensitive nature of this information, the data is not being held confidentially, is not password protected, and is free to share both for this project and outside of it. It can be legally and ethically shared and it is not personally identifiable information. At the conclusion of this project, though the data is not sensitive, it will be purged from all hard drives and databases.

The following datasets were used to create our project database:

1. Drug and Alcohol Treatment facilities:
    a. Source: Open Data PA [8]
    b. Description: Reports the geographical information of Pennsylvania Department of Drug and Alcohol Programs (DDAP) drug and alcohol treatment facilities. Not only will location data be important to our project, but also the type of treatment facility as there are three different types.
    c. Variables: county name, county FIPS code, county code
    d. Records: 892
2. Opioid Seizures and Arrests:
    a. Source: Open Data PA [7]
    b. Description: Contains the summary information on opioid drug seizures and arrests made by Pennsylvania State Police on a quarterly basis. An incident in this dataset is defined as a recorded violation of the Controlled Substance Act and may or may not contain an arrest. Seizures of opioids may have been the result of undercover buys, search warrants, traffic stops, or other investigations.
    c. Variables: county code, year, quarter, drug, sum of incidents, sum of arrests
    d. Records: 3,819
3. Naloxone Administered by County:
    a. Source: Open Data PA [5]
    b. Description: Dataset contains quarterly information on the number of doses given for Naloxone in Pennsylvania by County. Missing data in the incident county code column means it was not recorded by the provider.
    c. Variables: incident county code, year, date, medication given, record count
    d. Records: 25,404
4. Opioid Dispensations by County:
    a. Source: Open Data PA [4]
    b. Description: Contains data for dispensation of all Schedule II-IV opioids in Pennsylvania by County on a quarterly basis.
    c. Variables: county name, year, quarter, type of dispensation/prescription, age group, gender, count of dispensation/prescription
    d. Records: 84,320
5. Overdose Records by County:
    a. Source: Open Data PA [3]
    b. Description: Data contains summary information on overdose responses and naloxone administrations by Pennsylvania criminal justice agencies as well as some third-party first responders. Data is recorded voluntarily, and as such does not represent all overdose incidents in Pennsylvania.
    c. Variables: county name, date, gender, age group, ethnicity, ethnicity description, drug, Naloxone administered y/n, survive y/n
    d. Records: 59,907
6. GDP by County:
    a. Source: Open Data PA [12]
    b. Description: Gross Domestic Product information by County in Pennsylvania, not adjusted for inflation.
    c. Variables: year, county FIPS code, county name, GDP
    d. Records: 1,541
7. Population by County:
    a. Source: Open Data PA, U.S. Census Bureau [13]
    b. Description: Data is sourced from the US Census Bureau using information from an ongoing survey (the American Community Survey) and provides population data by gender and age group for each county in Pennsylvania.
    c. Variables: year, county name, population, male population, female population
    d. Records: 1,742 (two datasets)

Database Description

1. Title: SUD database
2. Sources: PA Attorney General, Centers for Disease Control and Prevention (CDC).
3. Past Usage: All data was exported from the PA Attorney General and CDC websites.
4. Relevant Information:
    a. The SUD database is built by collecting from various data sources, aiming to discover the facts of opioid related information for further modeling and developing analysis reports. Model attributes will consider in the following attributes:
        i. COUNTY Base data of Pennsylvania counties
        ii. YEAR Calendar year of the data
        iii. QUARTER Calendar quarter of the data
        iv. GDP Economic data from counties
        v. OPIOID_INCIDENTS Opioid-related cases reported
        vi. OPIOID_ARRESTS Opioid-related arrests
        vii. TOTAL_POPULATION Population metrics of counties
        viii. OPIOID_DISPENSATIONS Gender-based opioid dispensation data
        ix. OPIOID_PRESCRIPTIONS Gender-based opioid prescription data
        x. NALOXONE_ADMINISTRATIONS Gender-based Naloxone administrations
5. Number of Instances: 1,686
6. Number of Attributes: 15
7. Attribute Values:
    a. CO_YR_QTR Adams 2019 Q1, Forest 2014 Q4, Erie 2017 Q3
    b. COUNTY Allegheny, Erie, Lawrence
    c. YEAR 2000, 2010, 2014
    d. QUARTER Q1, Q2, Q3, Q4
    e. GDP_CATEGORY LOW, LOWER-MIDDLE,HIGH
    f. TOTAL_POPULATION 100839, 1230360, 65867
    g. OPIOID_INCIDENTS 2, 1, 3
    h. OPIOID_ARRESTS 1, 5, 4
    i. NALOXONE_ADMINISTRATIONS 5, 11, 25
    j. OPIOID_DISPENSATIONS_FEMALE 150, 7697, 146645
    k. OPIOID_PRESCRIPTIONS_FEMALE 359, 7859, 14168
    l. OPIOID_DISPENSATIONS_MALE 145, 5427, 1104541
    m. OPIOID_PRESCRIPTIONS_MALE 157, 8744, 11393
    n. NALOXONE_ADMIN_FEMALE 0, 3, 123
    o. NALOXONE_ADMIN_MALE 7, 273, 19
8. Missing Values:

Class	Missing_Count	Missing_Percent
GDP_CATEGORY	0	0.00
TOTAL_POPULATION	0	0.00
OPIOID_INCIDENTS	867	51.45
OPIOID_ARRESTS	867	51.45
NALOXONE_ADMINISTRATIONS	779	46.23
OPIOID_DISPENSATIONS_FEMALE	0	0.00
OPIOID_PRESCRIPTIONS_FEMALE	0	0.00

OPIOID_DISPENSATIONS_MALE	0	0.00
OPIOID_PRESCRIPTIONS_MALE	0	0.00
NALOXONE_ADMINISTRATIONS_OD_FEMALE	315	18.69
NALOXONE_ADMINISTRATIONS_OD_MALE	178	10.56

Data Preparation and Cleaning

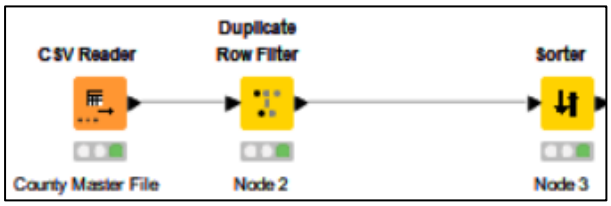
Data Preparation and Analysis of Sources

The next step is to prepare the data for analysis and identify possible areas of concern, including but not limited to missing data, outliers, misspellings, or other inconsistencies that would interfere with the analysis. Below, we'll discuss the possible issues with each dataset and how we plan to resolve them moving forward.

1. Drug and Alcohol Treatment Facilities

This dataset is concerned primarily with the identification and classification of treatment facilities in each county of Pennsylvania. There are no outliers, as the fields we are concerned with are categorical, and all data is relevant to the study. One possible issue of concern is that the date range includes only facilities constructed through 2018, leading to the possibility of missing new facilities that have been constructed in the past seven years. Currently, we plan to use the data as is provided in the document.

To prepare our data for analysis, we began with the "Drug and Alcohol Treatment Facilities" file, a .csv containing 892 records and 10 attributes detailing treatment facilities across Pennsylvania counties in 2018. First, we imported the .csv into our data processing environment. Next, we focused on retaining only three essential attributes: *County Name, County FIPS Code*, and *County Code*, discarding the remaining seven. To address redundancy, we removed duplicate rows based on these three retained attributes. Finally, we sorted the resulting data by *County Code* to ensure alphabetical organization. This cleaning process significantly reduced the dataset, resulting in 69 unique county records with the three core attributes necessary for our project.



2. Opioid Seizures and Arrests

The date range for this dataset is at the quarterly level from 2013 through 2025, however the data is only partially presented after 2023. The data is arranged in such a way that there is a row for each drug and a count of incidents. We can disregard the analysis by drug and instead concentrate on summary analysis of incidents and arrests by county, year and quarter. This will allow each county to have a single row for each quarter, which will match the granularity in other datasets. Care must be taken that no information is lost during this transformation. The data is unlikely to have any outliers, however several fields are irrelevant to the study such as latitude and longitude data, the unit used (only kilograms), redundant county codes, and state code.

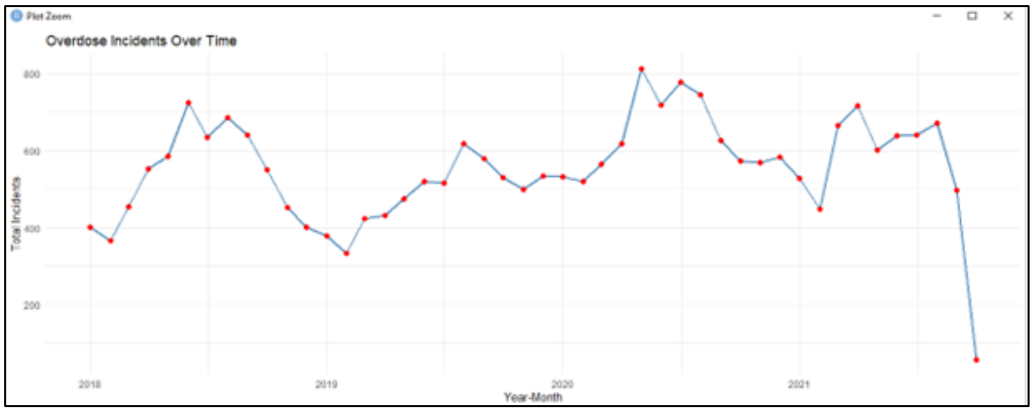
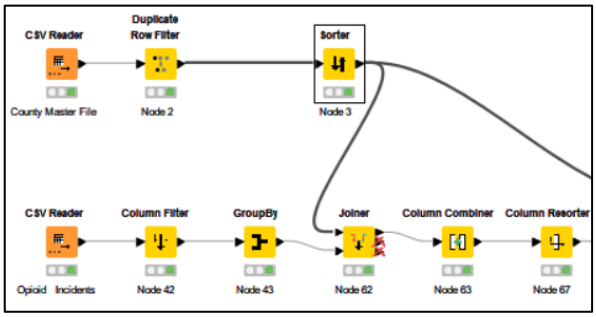


Figure 1. Opioid Seizures and Arrests CY 2013-Current Quarter

As shown in Figure 1, after the year 2022, we observe significant drops of overdose incidents. We will need to take care to analyze our data closely for missing values and outliers.

The "Opioid Seizures Arrest" file, a .csv containing 3,819 records and 16 attributes, detailed opioid incidents and arrests across Pennsylvania counties from 2013 to 2025. This file underwent several cleaning and transformation steps. First, we imported the .csv. The *Drug* attribute was retained though not utilized, and the intention was to allow for potential future filtering based on opioid type without disrupting the current workflow. Next, we grouped the data by *County Code*, *Year*, and *Qtr* to aggregate incident and arrest counts. An inner join was performed with the "County Master File," using *County Code* as the key, to incorporate *County Name* into the dataset. Subsequently, we created a primary key, *Co-Yr-Qtr*, by concatenating *County Name*, *Year*, and *Qtr*, ensuring a unique identifier for each summary of records. Finally, the columns were reordered to place the primary key first. This comprehensive cleaning process resulted in 2,701 records with six retained attributes, providing a structured dataset for further analysis.



3. Emergency Medical Services Naloxone Dose Administered

The date range for this dataset is from 2018 to 2021. The data is collected daily, requiring us to aggregate it up to the quarterly level to match the level of granularity of the other datasets. Additionally, the data is merely a record of the dispensation of naloxone; there are no other factors available, such as sociological gender, age, education level, etc., which limits the usability of the data. There is a small risk of outliers which are unlikely to appear in the fields we are analyzing, and several irrelevant fields such as event time, complaint by dispatch, and multiple county identifiers.

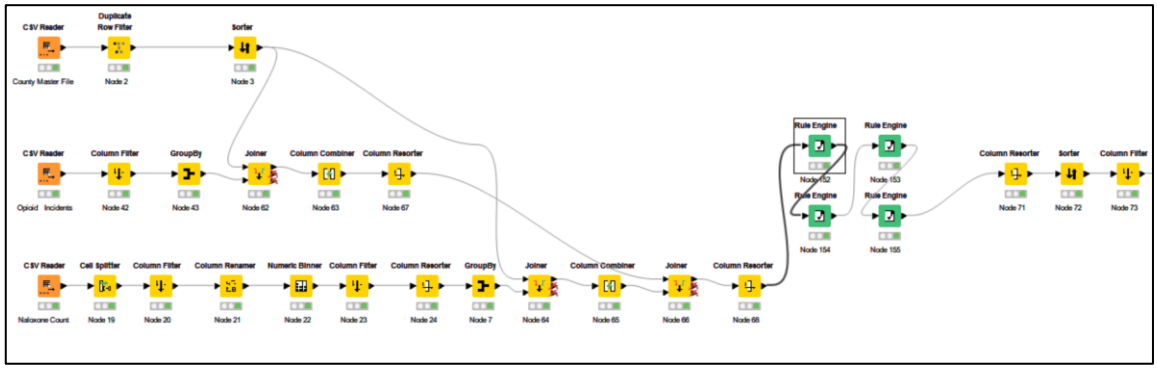
The "Naloxone Administered by County" file details Naloxone administrations in Pennsylvania counties from 2018 to 2021. It was renamed "Naloxone Count." This file contained 25,404 records with 14 attributes. The cleaning process began by importing the .csv and then splitting the *Event Date* column at each "/" delimiter. The original *Event Date* and its day and year components were removed, leaving only

the month value which was renamed *Month*. Using a Numeric Binning node, the *Month* column was transformed into *Quarter* based on defined ranges: Q1 (1-3), Q2 (4-6), Q3 (7-9), Q4 (10-12). The *Month* column was subsequently removed. The data was then sorted by *County Number*, *Calendar Year*, *Quarter*, *Medication Given*, and *Record Count*. Records were grouped by *County Number*, *Calendar Year*, and *Quarter* to aggregate the *Record Count* into a sum for each county, year, and quarter combination. Finally, an inner join with the "County Master File," using *County Code* as the key, was performed to add *County Name* to the dataset. This process streamlined the data to 5 retained attributes and prepared it for further analysis.

With the Naloxone Count and every subsequent file that we will analyze, we must undergo what KNIME labels a "prediction" process using Rule Engine nodes. When joining records, in most cases, we will utilize a full join, so that we get left outer, inner, and right outer records. There will be mismatched records on both sides of the join due to missing primary key records. For instance, the left side of the join may not contain "Adams 2020 Q1," both files may contain "Adams 2020 Q2" and "Adams 2020 Q3," but only the right may contain "Adams 2020 Q4." We will use Rule Engine nodes to compare the two files being joined and utilize then retain value that is not null/missing. We will do this once per join on the *CO-YR-QTR*, *COUNTY NAME*, *YEAR*, and *QUARTER* attributes. The formula for this, inside the Rule Engine node is (using *CO-YR-QTR* as an example):

```
MISSING $CO-YR-QTR$ => $CO-YR-QTR (right)$
MISSING $CO-YR-QTR (right)$ => $CO-YR-QTR$
TRUE => $CO-YR-QTR$
```

Once the data has been properly predicted, the prediction data is sorted to the front of the record, it is resorted in primary key order, and then the superseded *CO-YR-QTR*, *COUNTY NAME*, *YEAR*, and *QUARTER* attributes are eliminated.

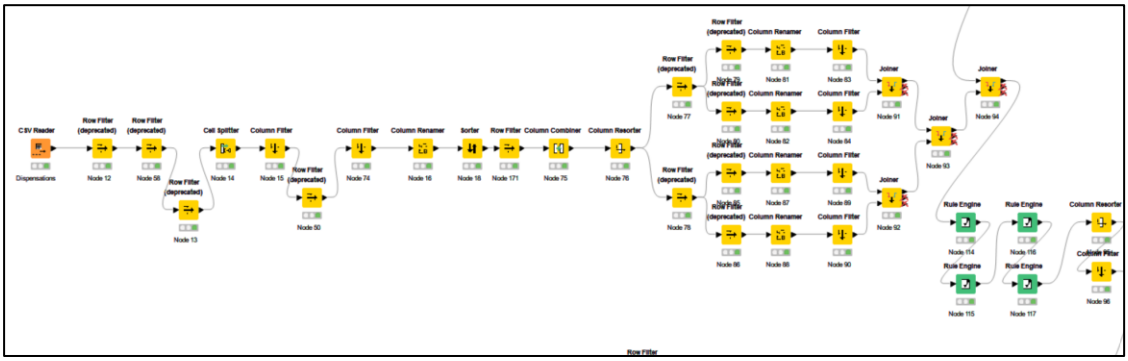


4. Dispensation Data without Buprenorphine

The data range for this dataset is more limited, being only from 2018 to 2024, with the data from 2024 being only partially complete. Additionally, the period is formatted as "Year-Quarter" which does not match our other datasets and will need to be reformatted. A few issues also occur with the data type or groupings found in other fields. The attribute *Type of Rate or Count Measure* mixes two types of data (integers and doubles) that should be separately reported for this kind of analysis. Due to this complication, we will be eliminating all rate measurements from the study and utilizing only the count data. Also, the *Age Group* category, while potentially useful, is grouped differently than other datasets, leading to difficulties matching similar age ranges. Likewise, it will also be eliminated from the study. With the elimination of the rate measurement data, outliers are unlikely. There are, however, multiple irrelevant

fields that will be removed from the final dataset including type of drug class, time measure, quarter date start, notes, latitude, longitude, and three county identifiers.

This dataset includes 84,320 records with 18 attributes. Of these attributes, only 7 are of interest and were kept for analysis – *County Name*, *Year*, *Time Period*, *Type or Rate of Count Measure*, *Age Grouping*, *Gender*, and *Rate or Count*. This dataset required many transformation steps to address the issue identified above. First, we filtered out any rate value from the *Type of Rate or Count Measure*. As mentioned, including both rates and counts in the same field makes analysis or any kind or summation impossible. We determined the counts are the most useful metric and will concentrate on these records only. Next, we filtered out all records with "All Gender" in the *Gender* field and keep only records with a value of "Male" or "Female." Not only do we feel this would lead to more interesting insights than simple aggregation, but it will pair nicely with any other gender related data we may find to supplement the analysis. We also filtered out records in the *County Name* attribute that only contained "Pennsylvania," as this was clearly not a county-based record but rather an aggregation of all county data. For the *Time Period* attribute, we split the value using the first instance of a space as a delimiter to separate the year from the quarter data. Additionally, we filtered out values of *Time Period* and *Time Period (Year)* from the field and renamed the *Time Period (Qtr)* attribute to *Quarter*. This dataset included age groupings that were incompatible with age groupings found in other datasets. With no way to correct this, we were forced to eliminate the attribute from our analysis. We then moved to sorting the data in the order of *County Name*, *Year*, *Quarter*, and *Type of Rate or Count*, then filtered out any data after 2023. This filter was necessary as details from 2024 were only partially present and could potentially skew the analysis. We next needed to perform several steps twice for each gender. We row-split the data between genders, then split again between dispensations and prescriptions, and renamed the split data to include the gender being measured (e.g. *Opioid Dispensations – Female*). We then filtered out the *Gender* and *Type or Rate of Count Measure* attributes and replaced them by adding the gender to the column names, then used a Joiner node to join the respective gender information back together. Finally, we used another Joiner node to join both gender's data together. We then performed a full join with the final table from the Naloxone Count file on the *CO-YR-QTR* column to retain unmatched records from both datasets and then repeated the same Rule Engine/Prediction process as described in the Naloxone Count file.



Finally, we sorted the new prediction columns to the front of each record and filtered the columns to retain the following 11 attributes: *Prediction CO-YR-QTR, Prediction County Name, Prediction Year, Prediction Quarter, Sum of Incident Count, Sum of Arrest Count, Sum of Record Count, Opioid Dispensations – Female, Opioid Prescriptions – Female, Opioid Dispensations – Male, Opioid Prescriptions – Male*. In total we retained 2,087 records from the combined datasets.

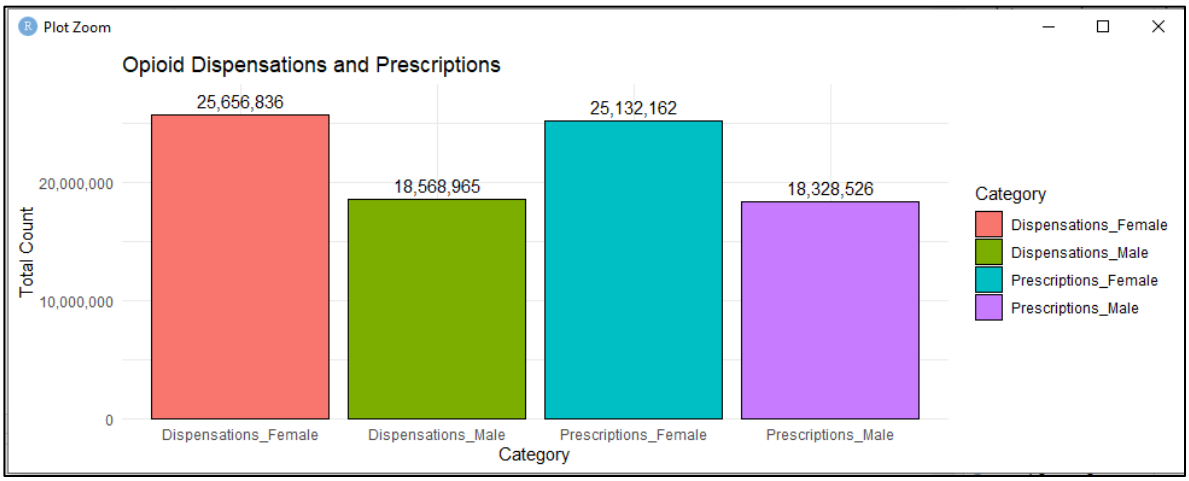


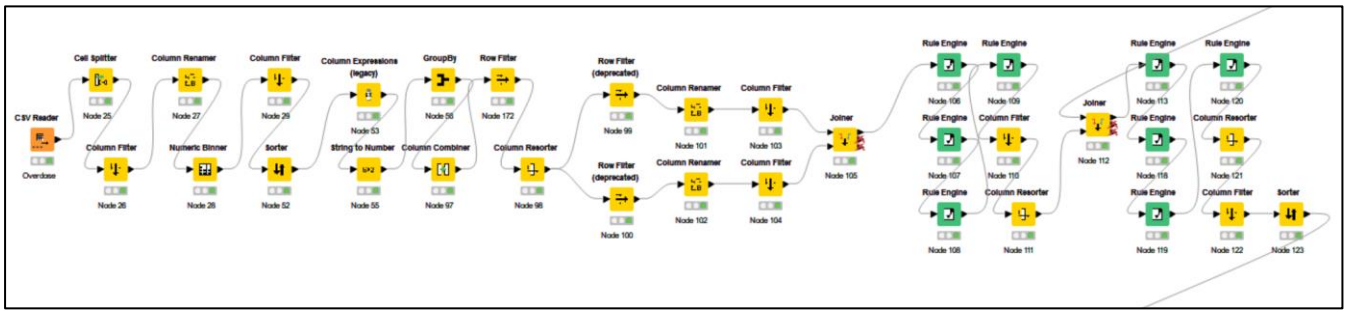
Figure 2. Dispensation Data without Buprenorphine 2016-2023

5. Overdose Information Network Data

The date range for this dataset covers 2018 to 2024, but unlike other datasets, the data for 2024 appears to be complete and usable. The data is presented daily, again requiring aggregation up to the quarterly level of detail. This dataset also contains a field for age group, but it cannot be matched to the groupings used in other datasets. As a result, this attribute needs to be eliminated from the study. There is also a field available for ethnicity, which would be interesting to include in the study, however it does not exist in any other dataset we have available. Unfortunately, this field must also be eliminated. The specific data of interest in this dataset is the Naloxone administrations, which are contained as a binary Y/N, requiring a conversion to an integer type to aggregate and analyze as a total number of administrations. There is a risk of outliers in several fields such as ethnicity description, suspected opioid, and race attributes, however these will be eliminated from the study as they are irrelevant to our analysis.

This dataset includes 59,907 records and 34 attributes, of which we retained 9: *Incident County Name*, *Incident Date*, *Gender*, *Age Range*, *Race*, *Ethnicity Description*, *Suspected Opioid*, *Naloxone Administered* and *Survive*. For our transformation steps, we first split the *Incident Date* field using the "/" character as the delimiter. We did this to extract the month and year values to use the previously described binning procedure on these records for splitting into quarters, and to have the year value separated for the creation of the primary key. We then filtered out the original *Incident Date* field now that the month value has been extracted and reviewed the remaining columns for additional filtering. We ultimately determined that although much of the data was interesting, unfortunately, it was incompatible with our other data sources. This included *Age Range* (the ranges did not match our other data sources), *Race* and *Ethnicity Description* (this data was not found at all in other data sources). We also filtered out the attributes *Suspected Opioid*, *Survive*, and *Incident Date (day)*, as these do not match the aggregation level we are analyzing. Like steps taken with previous datasets, we placed these filters towards the end of the process to make retrieval easier should we find additional datasets that could be used for comparison. Next, we renamed the remaining *Incident Date* fields to their respective levels: *Year* and *Month*, then used a Numeric Binning node to put months into quarters and named the new field as *Quarter* and then finished this step by filtering out the remaining *Month* field. The next step was to use a Column Expression node to translate the binary object of "Y/N" in the *Naloxone Administered* field into a 1 or 0, then converted this new field into an integer to use the data quantitatively. We then grouped the records by *County*, *Gender*, *Year*, and *Quarter* in this specific sequence to aggregate the counts into a sum for each sequence. We then created the primary key from these fields and sorted the primary key field to the front of the records. Next, we filtered all data out after 2023, as the 2024 data was only partially presented and could skew the results. We then performed the row splits similarly to what was done for the Dispensation Data to create new fields for each gender (e.g. *Naloxone Administered – Female*). The dataset at this point was ready to run the same Rule

Engine/Prediction process as described previously. It should be noted here that this process was run twice for this dataset, as there are records that were present for only one gender or the other. For example, Bradford County in the 3rd Quarter of 2019 only had data for females, not males, causing a mismatched record error when rejoining the two gender files back together. After this first run of the Rule Engine, we then performed a full join with the final table from the Dispensations file to again retain all records (mismatched and otherwise) from both sides of the join. We then performed the second run of the Rule Engine and sorted the new prediction columns to the front of each record.



At this point, we have retained 2,087 records and the following 13 attributes throughout the process: *Prediction CO-YR-QTR, Prediction County Name, Prediction Year, Prediction Quarter, Sum of Incident Count, Sum of Arrest Count, Sum of Record Count, Opioid Dispensations – Female, Opioid Prescriptions – Female, Opioid Dispensations – Male, Opioid Prescriptions – Male, Naloxone Administered – Female, Naloxone Administered – Male*.

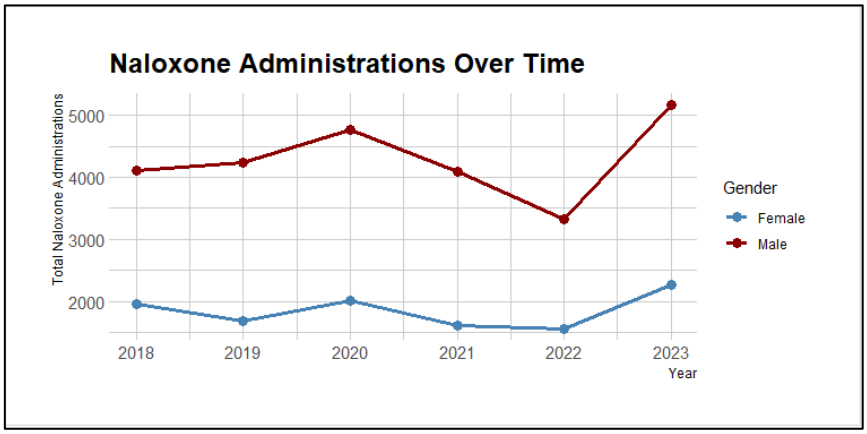


Figure 3. Overdose Information Network Data CY 2018-2023

6. Gross Domestic Product by County

This dataset has the largest date range, spanning from 2001 to 2023 (Figure 2). However, since none of our other datasets go back to 2001, we need to evaluate how much of the GDP data is relevant to keep. The data here is also presented annually, which is at a lower level of granularity than our other datasets. We will adjust this data to the quarterly level. Another issue is the formatting of some of the data. The County FIPS code is provided in a format that does not match the master file and will need to be reformatted for mapping. Also, the County name is provided in a "County, State" format, and will need to be reformatted to provide just the County name alone. Another issue is the GDP value is presented in the thousands, which, while being more precise, may lead to readability issues. We will determine whether to leave it in this

format or aggregate to a higher level. Finally, after analysis of the dataset, it has been concluded that there is no irrelevant data in this dataset.

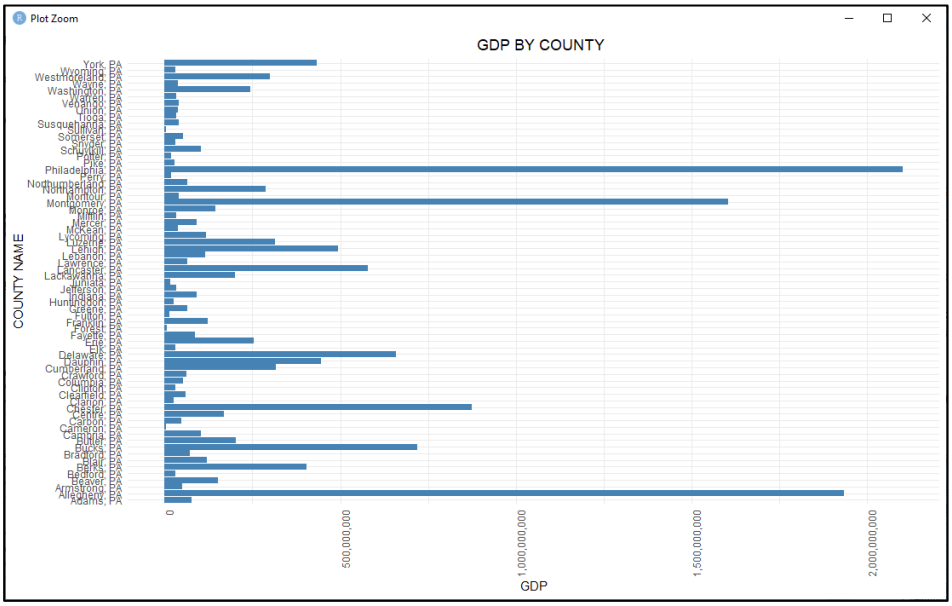
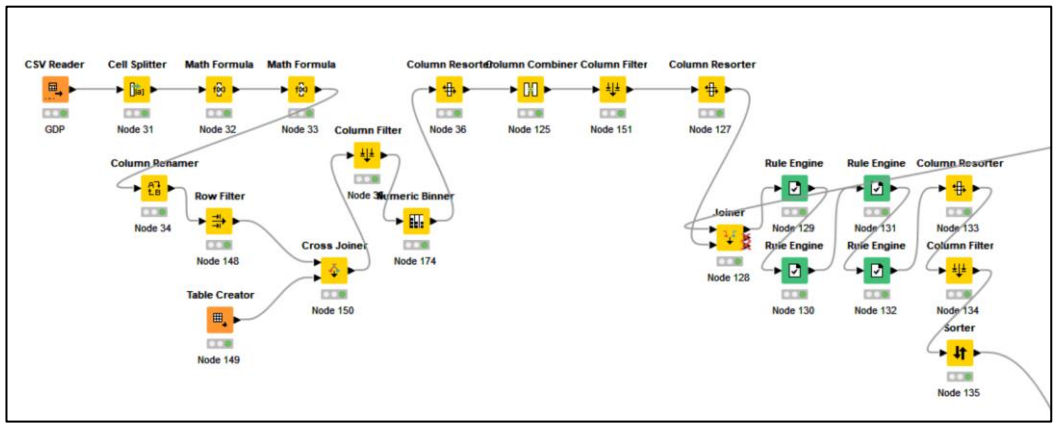


Figure 4. Gross Domestic Product by County 2001 to Current

The dataset includes 1,541 records and 5 attributes, of which 4 are retained for the transformations and joins: *Year*, *County FIPS Code*, *County Name* and *GDP (in thousands)*. One note on the *County FIPS Code* attribute is that this is the first dataset that uses this value instead of the *County Code* to identify county. Although we will use the *County Name* to join with the previous datasets, we will retain this additional field in case it is necessary for future joins. For the transformation and cleaning, we first needed to split the *County Name* attribute by the comma delimiter, as all records were recorded in the form of "Adams, PA" and the state data is not needed. Next, we used a Math Formula node on the *County FIPS Code* to subtract 42000 from the value. This was necessary as the formatting here contained the state code of "42" appended to a three-digit county code (e.g., 42001 for Adams). We then used another Math Formula node on the *GDP* field as we did not want this value aggregated at the thousands level. We renamed the *County Name (County)* field to just *County* and then filtered all rows with a year value less than 2010 as data before this period is not relevant to our other datasets. For our last step before the join, we used a Table Creator node to create a duplicate value for each of the four quarters, so that we could present this data quarterly rather than annually to mirror our other data's granularity. For the join, instead of using the Joiner node as in previous steps, we used a Cross Join node. This is to ensure we have full data for each quarter of the year; however, we acknowledge this method simply duplicates the yearly GDP data for each of the newly created quarter records. Instead of using the numeric values, for the next step we used a Numeric Binning node to classify each county's GDP as "Low," "Lower-Middle," "Upper-Middle," and "High." These values were derived by observing the quartiles for all values and then classifying each county into the corresponding bin. The new column was named *GDP Category*. Finally, we created the primary key using the *County Name*, *Year*, and *Qtr* columns, and performed a full join with the final table from the previous dataset. We then repeated the Rule Engine/Prediction process as with the other datasets and sorted the new prediction columns to the front of each record.



After these steps, we have retained 3,829 records with the following 14 attributes: *Prediction CO-YR-QTR, Prediction County Name, Prediction Year, Prediction Quarter, Sum of Incident Count, Sum of Arrest Count, Sum of Record Count, Opioid Dispensations – Female, Opioid Prescriptions – Female, Opioid Dispensations – Male, Opioid Prescriptions – Male, Naloxone Administered – Female, Naloxone Administered – Male, GDP Category*.

7. Population by Gender and Age & Population 2020-2023

The date range of this dataset is the most limited of all our datasets, covering only the years from 2010 to 2019. We will need to search for more current data for better analysis. Like the previous dataset, the County Name is provided in a different format than our database schema and requires reformatting. The population is also split by gender with many different age categories. We will keep the total gender-based counts, but the age categories again do not match any of the groupings from other datasets, and as such, will be eliminated. With the high number of different categories there is a risk of outliers, however with the elimination of the age categories this becomes less of a concern.

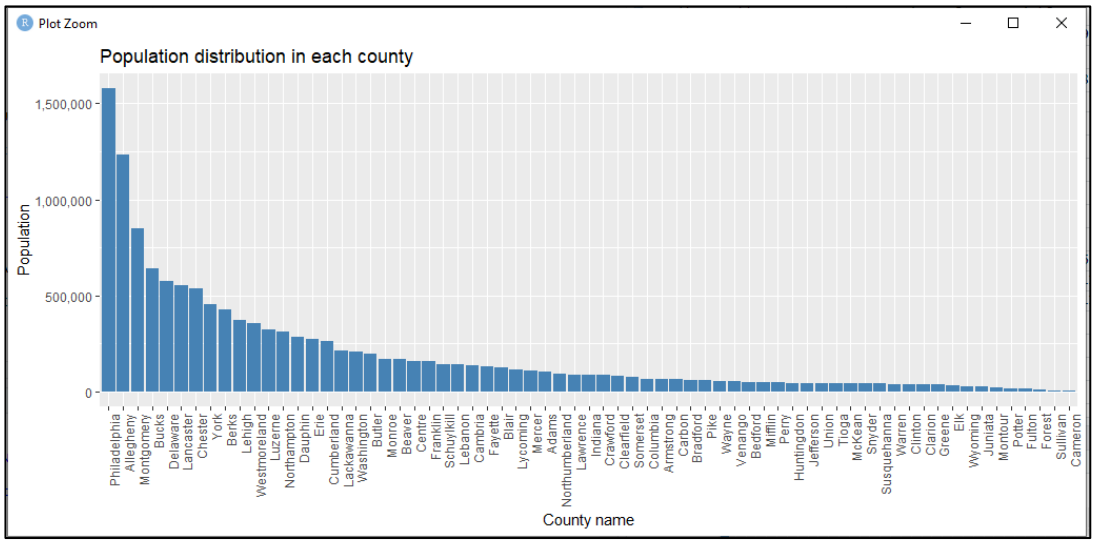
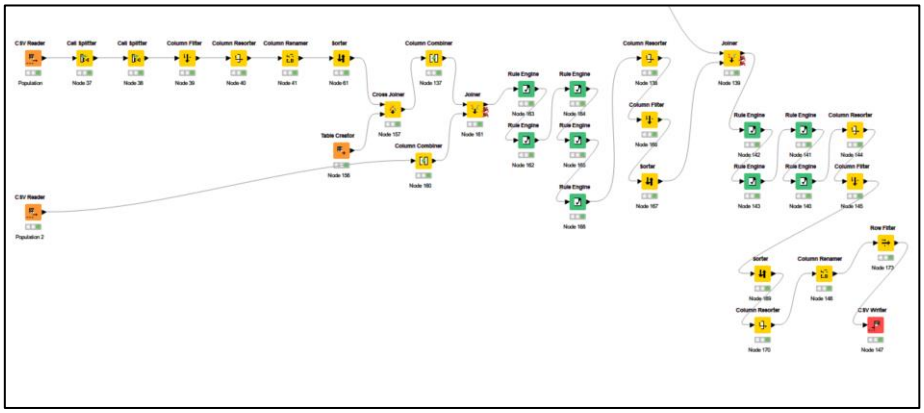


Figure 5. Population (US Census: ACS 5 Year Estimates) by County

The dataset contains only 670 records with 53 attributes, of which we only retained 3: *Period End Year*, *Count Name*, and *Population*. As the *County Name* attribute had the format of "Adams County, Pennsylvania," we had to extract the county name by splitting the column first using the comma delimiter and then again using the space delimiter. We can then filter out all of the extra columns except the one

containing the name of the county in the format we have been using (e.g., *Adams)*. We then sorted the data by *County Name* and *Period End Year* and used a Table Creator node to create duplicate values for each of the four quarters just as we had done with the GDP dataset. We used a Cross Join node to create the quarterly data and created the primary key using the *County Name*, *Year*, and *Qtr* columns. At this point we needed to bring in a second file with the updated population data from 2020 to 2023, which contained 1,072 records. We preprocessed this dataset in Excel as it was sourced much later than the other datasets being used, but the only transformations required were to remove formatting and exclude certain records in the *County Name* column (e.g., Adams, Pennsylvania). We then used a Joiner node to perform the full join with the population tables and repeated the Rule Engine/Prediction process two times. This had to be done due to the mismatched records in both population files and to make a fifth prediction on population between the two population tables.



After these steps, we have finally retained 3,829 records with 15 total attributes: *CO_YR_QTR, County, YEAR, QUARTER, GDP_CATEGORY, TOTAL_POPULATION, OPIOID_INCIDENTS, OPIOID_ARRESTS, NALOXONE_ADMINISTRATIONS, OPIOID_DISPENSATIONS_FEMALE, OPIOID_PRESCRIPTIONS_FEMALE, OPIOID_DISPENSATIONS_MALE, OPIOID_PRESCRIPTIONS_MALE, NALOXONE_ADMINISTRATIONS_OD_FEMALE, NALOXONE_ADMINISTRATIONS_OD_MALE*.

Data Modeling Planning

With our dataset in a cleansed and processed state, we turn back to our original questions to plan the data modeling steps focused on how to answer them. We will look at the questions in turn and describe how we plan to use the data to address each.

Understanding Disparities in SUD Treatment Access

To answer these questions, we need to look at both the availability of services (e.g., treatment centers, naloxone administration) and the prevalence of the issue in the counties (e.g., opioid prescriptions, incident and arrest counts). We can first use Cluster Analysis techniques such as K-Means and DBSCAN to identify similarities across counties based on the factors we've identified. This can help us identify "at-risk" areas where treatment may be harder to access relative to what services are available. To help us with our modeling, we may need to create new features such as the ratio between opioid prescriptions to population or between naloxone administrations to incident count, in order to better understand the differences between counties.

In addition, we want to know where harm reduction services, specifically Naloxone access and syringe service programs, are located across Pennsylvania; our data modeling plan incorporates several key data

sources. Primarily, we will leverage publicly available data from the Pennsylvania Attorney General's website, which identifies locations where Naloxone has been administered by emergency medical services. This dataset will provide a valuable initial layer of information, indicating areas with a high incidence of opioid overdose requiring emergency intervention.

It's crucial to acknowledge the limitations of this single data source. As noted [18], Naloxone is also accessible over the counter at any pharmacy statewide, and numerous Community-Based Organizations (CBOs) distribute it directly to individuals, particularly those without insurance. To create a more comprehensive understanding of Naloxone availability, our data model will aim to integrate supplementary information. This may involve exploring publicly available lists of registered pharmacies, if such data exists, and identifying known CBOs that provide Naloxone distribution. We will investigate potential data sources from the Pennsylvania Department of Health or other relevant agencies that might list or map these CBOs.

The data model will need to accommodate different data formats and levels of granularity across these potential sources. We anticipate the need for data cleaning and standardization to ensure consistent location information (e.g., addresses, zip codes) that can be used for mapping and spatial analysis. The goal is to create a model that can visualize the geographic distribution of both documented Naloxone administrations and the locations of key harm reduction service providers, offering insights into service accessibility across the state.

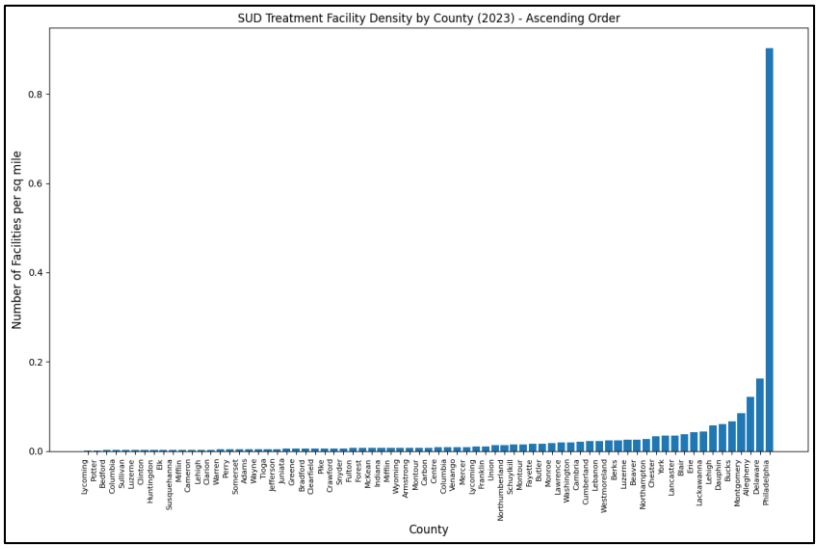


Figure 6. SUD Treatment Facility by County

We used each county land area [19] and the number of treatment facilities by county [8] to calculate the facility density by county (Figure 6) revealing a highly uneven distribution. Counties on the right side of the graph, such as Philadelphia, Delaware, and Allegheny, exhibit a significantly higher concentration of treatment facilities relative to their land area. In contrast, many counties on the left, including Lycoming, Potter, and Bedford, show a very low number of facilities per square mile, potentially indicating challenges in geographic accessibility.

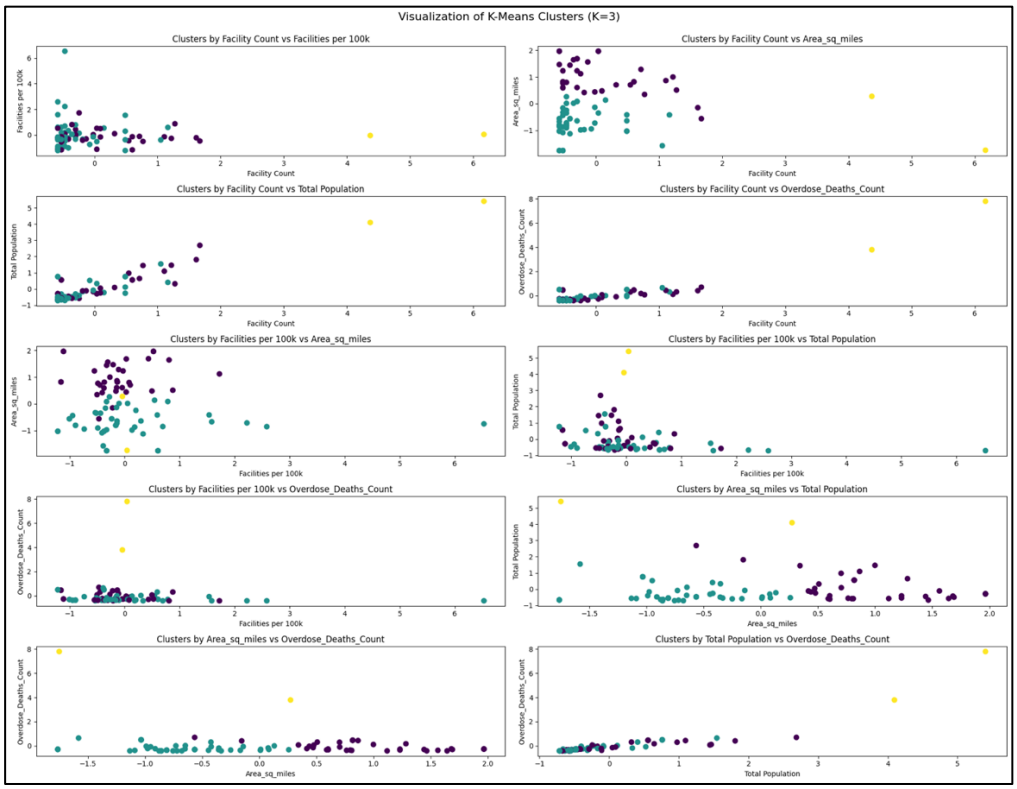


Figure 7. K-Means Clustering Analysis

The K-Means clustering analysis (Figure 7), utilizing an optimal number of three clusters, segmented Pennsylvania counties into distinct groups based on their SUD treatment facility counts and rates, geographic area, population size, and estimated 2023 overdose death counts [20]. The resulting clusters revealed key differences: 1) one group comprises densely populated, smaller counties with a high concentration of treatment facilities; 2) another consists of larger, more rural counties characterized by an average number of facilities but a lower per capita rate; and 3) a third group includes smaller, less populated counties with fewer facilities but a slightly higher treatment rate relative to their population. These descriptive findings offer a valuable segmentation of Pennsylvania counties, highlighting inherent similarities within each group regarding their treatment landscapes and the impact of the opioid crisis, providing a foundation for understanding the diverse needs and characteristics across the state.

KMeans Cluster	Facility Count	Facilities per 100k	Area sq miles	Total Population	Overdose Deaths Count
0 (Purple)	0.063393886	-0.132430623	0.968456147	0.098308588	-0.072052732
1 (Turquoise)	-0.272837498	0.101414566	-0.710385799	-0.277370424	-0.191479956
2 (Yellow)	5.27059124	0.000508918	-0.738144373	4.748650383	5.796728155

Table 1.K-Means Clustering Analysis Results

Socioeconomic and Geographic Factors Affecting Access

To comprehensively address these questions, our data model will integrate information from various sources:

- **Harm Reduction Service Locations:** Data on Naloxone administration by emergency medical services [5], potential lists of pharmacies selling over-the-counter Naloxone, and information on Community-Based Organizations (CBOs) providing Naloxone [18].
- **Socioeconomic and Geographic Factors:** Data at the county or zip code level will be gathered from sources such as the U.S. Census Bureau (American Community Survey), Bureau of Labor Statistics, County Health Rankings & Roadmaps, and Pennsylvania state government websites.
- **Overdose Rates and Economic Indicators:** Historical data on overdose deaths and non-fatal incidents [3][9] will be obtained from relevant health agencies. Economic data, including GDP for Pennsylvania and potentially its regions, labor force participation rates, healthcare expenditure data related to SUD, and indicators of production capability, will be gathered from sources like the Bureau of Economic Analysis and state economic development agencies.

Data integration will involve linking these datasets using common geographic identifiers (e.g., zip codes, county FIPS codes) and aligning temporal data where necessary. Data cleaning and standardization will be crucial for ensuring data quality and enabling effective analysis.

To understand the underlying factors that influence SUD and access to services, and to potentially reduce the dimensionality of our socioeconomic and demographic variables, we will employ factor analysis. This technique will allow us to identify latent constructs (such as underlying economic status, education level, age-related factors, etc.) that explain the patterns of correlation among a larger set of observed variables. By identifying these key underlying factors, we can gain a more parsimonious representation of the data and focus on the most influential dimensions in our subsequent analyses, including regression modeling and spatial analysis. This will also aid in reducing multicollinearity among predictor variables.

To begin our analysis of social, economic, and geographic factors impact treatment services, we first look at the number of opioid incidents per county and compare that to the number of treatment centers available (Figure 8). We converted our metrics to incidents and treatment centers per capita to compare fairly across counties, and this led to an insight that treatment centers may not be distributed well across the state as needed. From the chart below, we can see that many of the counties with the highest rate of opioid incidents do not have the most treatment centers available. In fact, it appears counties with fewer rates of opioid incidents contain the most treatment centers. This hints at an issue with access to treatment for those most in need.

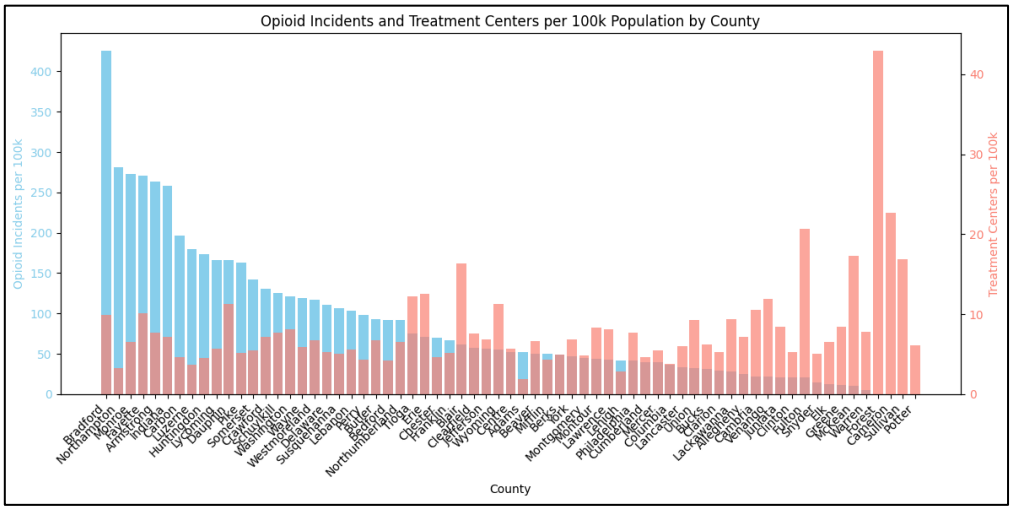


Figure 8: Opioid incidents compared to treatment centers (per capita)

Next, we compare the number of treatment centers available to GDP per capita (Figure 9). Doing a direct comparison we do not see much of a relationship, in fact the correlation between the two is fairly low (R^2 = 0.34).

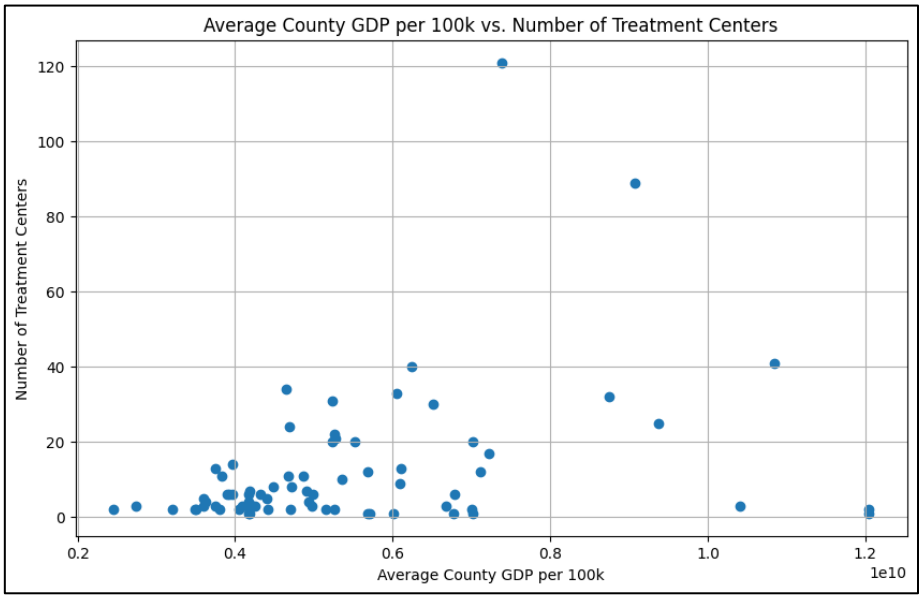


Figure 9: Number of Treatment Centers by County GDP per capita

To check this assumption, we again transform the number of treatment centers to a per capita number, and we see even less of a relationship. The richest counties have some of the fewest number of treatment centers available, but not overly different from the poorest counties.

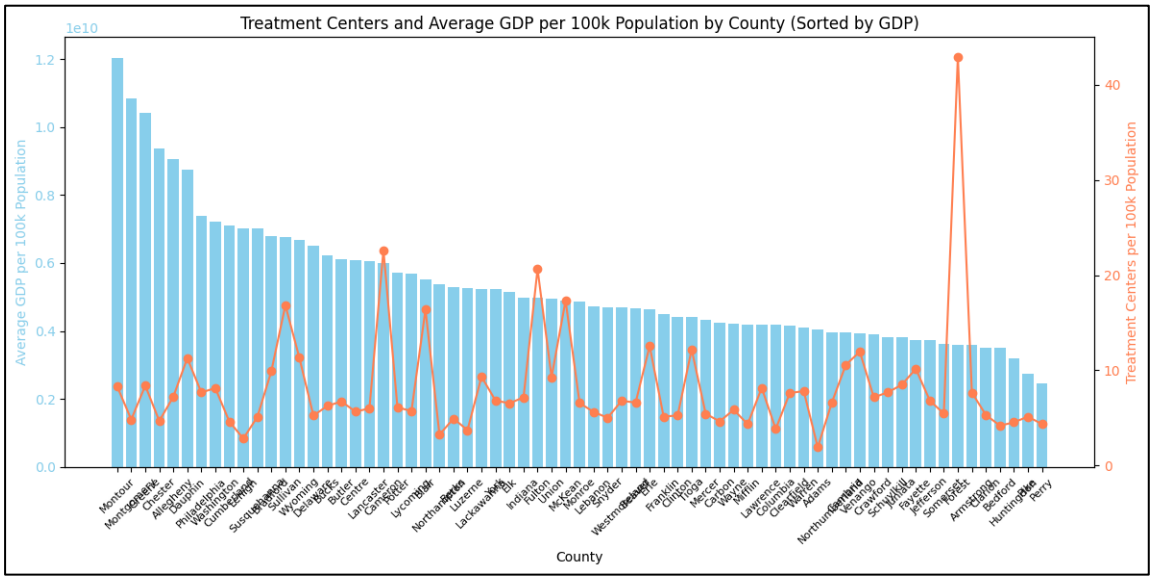


Figure 10: GDP per capita compared to Treatment Centers per Capita

To further explore the relationship between opioid incidents and GDP, we've plotted these against each other first as raw numbers shown in figure 11a and again transformed as per capita in figure 11b. The per capita chart clearly shows some relationship between opioid incidents and GDP, as counties with lower GDP per capita experience more opioid incidents per capita.

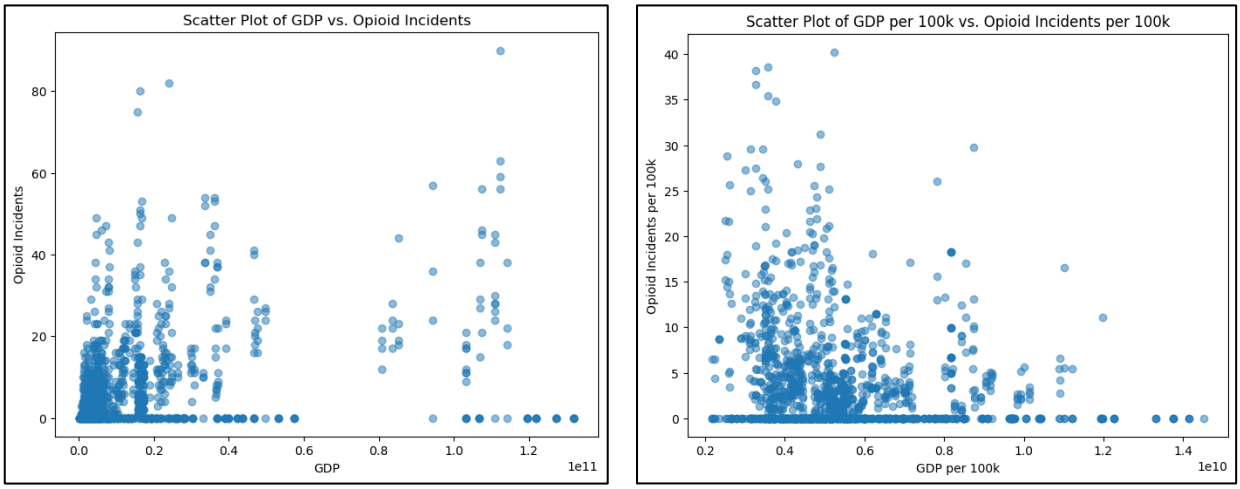


Figure 11a and 11b: GDP vs. Opioid Incidents

The Impact of Harm Reduction on Overdose Rates

We will use regression-based modeling, such as Multiple Linear Regression or Poisson Regression (depending on data distribution), to explicitly measure the relationship between harm reduction service availability (like the number of Naloxone administrations or locations of syringe programs) and overdose rates. Additionally, we'll employ predictive simulations to estimate how increasing these services in underserved areas might influence overdose rates. This clear modeling approach will allow us to quantify the effectiveness of harm reduction programs and provide evidence-based insights for targeted interventions.

"Service Availability" may be tricky to quantify, and answering the questions below may require additional factors to bring into the analysis. For example, evaluating the link between access to hard reduction services and overdose rates may not be easily answered simply by evaluating the number of treatment centers available. We may need to transform our existing factors to check variables such as accessibility. Service density per capita, public transportation options, or distance to the nearest treatment center can all affect the usage of services. Additionally, operating hours can also impact usage, as facilities with broader hours are more likely to be used than ones closing early or unavailable during the weekend.

We will also need to be mindful of potential confounders within the data. There are potential linkages between the number of treatment centers available and socioeconomic factors or law enforcement practices. To this end, we will attempt to account for confounders in our models, particularly Multiple Linear Regression models.

Similar to what we observed with the relationship between opioid incidents and treatment centers, we see the same imbalance between Naloxone Administrations with treatment centers (Figure 12). Northampton County has seen the highest per capita rate of Naloxone administered in the state yet has the third fewest number of treatment centers available per 100k residents.

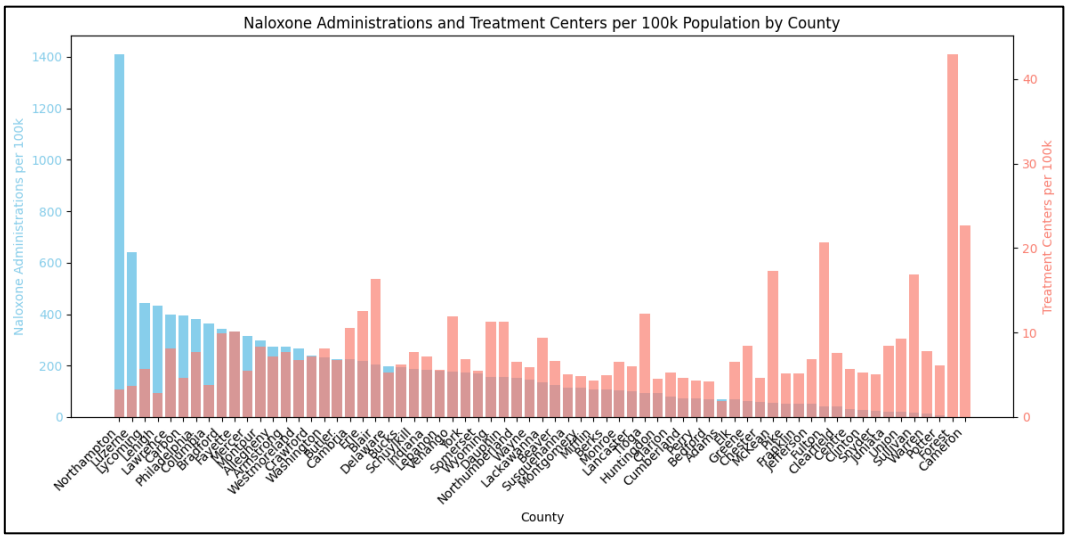


Figure 12. Naloxone Administrations vs Treatment Centers by County

We can also observe a potential relationship between Opioid Dispensations and Naloxone Administrations. Again, Northampton County has the highest count of both metrics, indicating a serious health crisis in the county. Aside from a few exceptions there does not seem to be a direct relationship between the two metrics, and the correlation is low (R^2=0.13).

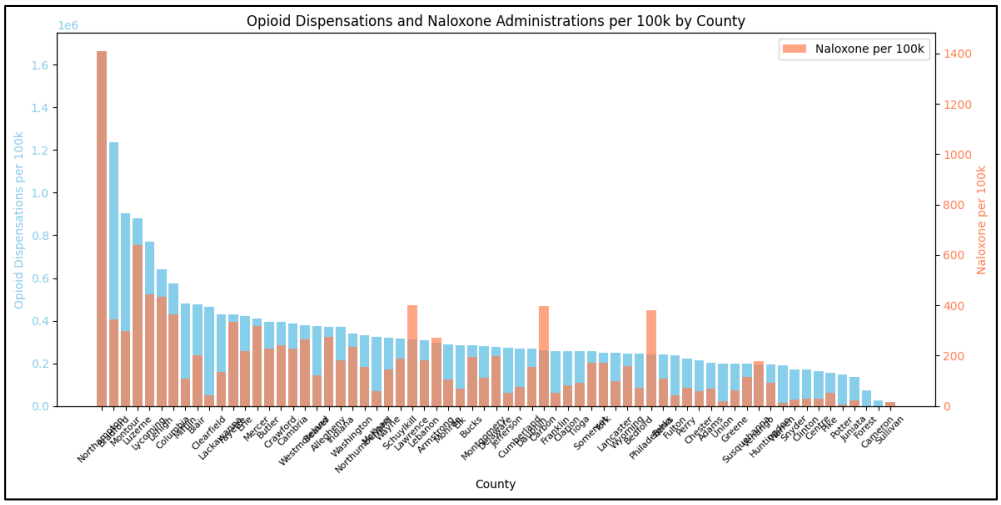


Figure 13. Opioid Dispensations vs. Naloxone Administrations per 100k population

Developing Data-Driven Recommendations

Finally, we'll create scenario-based simulations to show how strategically expanding harm reduction interventions (e.g., Narcan distribution, syringe exchange programs) in identified underserved counties could lower overdose rates and improve economic outcomes, including workforce productivity and reduced healthcare costs. This step provides clear, actionable recommendations that directly support targeted policymaking and public health resource allocation, effectively addressing our core hypothesis.

Drug abuse is commonly concentrated in large cities. As we saw in our GDP analysis, Figure 5 visualizes the top large cities in Pennsylvania are Philadelphia, Allegheny and Montgomery. The large population of cities correlates with established drug markets and distribution channels and a lower population has limited

access to the support program. Hence, harm-reduction services should be delivered to both large populations and lower population cities.

For larger population cities, expanding and delivering more harm-reduction services can save more lives and valuable feedback on services is more effective than in smaller population cities. Also, adjusting based on feedback is vital to better retain users. Increasing educational efforts are beneficial for the long-term SUD harm reduction strategy, and it's crucial for social economic and public health.

Overdose prevention centers are more practical for urban cities, public transportation is available, helping people save more time and money when receiving treatment services compared to those in lower-population cities. On the other hand, people who live in rural regions are facing more challenging situations to receive treatment services, such as long travel distances, which costs them more for the same services and potentially leads to a lower probability of SUD treatment completion. To better address these barriers, it's important to establish telehealth services in inconvenient regions such as Potter, Cameron and Forest counties. Treatment services design should be humanized, especially for the disabled and elderly who live farther distances from treatment facilities. Telehealth services can efficiently provide appropriate guidance on medication.

Variable	Estimate	Std. Error	z value
GDP_CATEGORYLow	3.055779e-08	1.354497e+05	2.256025e-13
QUARTERQ4	2.173023e-08	2.619985e+04	8.294029e-13
GDP_CATEGORYLower Middle	1.563485e-08	1.208762e+05	1.293460e-13
QUARTERQ3	-2.348243e-09	2.590325e+04	-9.065438e-14
QUARTERQ2	2.232413e-09	2.601578e+04	8.580998e-14
OPIOID_DISPENSATIONS_MALE	-1.527855e-11	2.773537e+01	-5.508686e-13
OPIOID_PRESCRIPTIONS_MALE	2.523575e-11	3.689105e+01	6.840616e-13
OPIOID_PRESCRIPTIONS_FEMALE	5.355016e-12	1.588976e+01	3.370104e-13
OPIOID_DISPENSATIONS_FEMALE	-1.905420e-12	2.084859e+00	-9.139320e-13

Table 2. Coefficients for predicting Naloxone administration

GDP_CATEGORYLow : 3.055779e-08 (0.00000003055779)
GDP_CATEGORYLower-Middle : 1.563485e-08 (0.00000001563485)

Based on the logistics model results, we can identify the pattern of Naloxone administration. The higher estimate of the variables is towards the positive correlation of higher Naloxone administration. From an economic aspect, counties with lower GDP levels and more rural regions are experiencing more overdose situations. This might be due to limited access to harm reduction services such as mobile clinics, weekly voluntary medical events and transportation fare discounts.

QUARTERQ4 :2.173023e-08
QUARTERQ3: -2.348243e-09
QUARTERQ2:2.232413e-09

Although Q2 appears to have a higher estimate than Q4, when we convert them into decimals, the estimate for Q4 (0.0000000217) is actually higher than that of Q2 (0.00000000223). This suggests that during Q4 or the fall and early winter season, the overdose rate tends to increase compared to other quarters or seasons. The logistic model makes harm-reduction services predictable and visualizes the patterns of the target group.

OPIOID_PRESCRIPTIONS_MALE :2.523575e-11 (0.0000000000252)
OPIOID_PRESCRIPTIONS_FEMALE: 5.355016e-12 (0.00000000000536)

This indicates that males correlate to higher Naloxone Administration about 5 times greater than females. Education efforts are crucial to achieving harm reduction for long term success. Additionally, to evaluate the effectiveness of opioid harm reduction for long term goal, we can track the increase of services participation and retention rate over time.

Discussion

Bias

While our analysis provides valuable insights into the opioid crisis in Pennsylvania, it's crucial to acknowledge the inherent limitations and potential biases present in the data. This awareness is essential for responsible interpretation and application of our findings.

First, the data on opioid incidents, arrests, and naloxone administrations may not fully capture the true extent of the crisis. Underreporting is a common issue in public health data, particularly with sensitive topics like substance abuse [15]. Stigma, fear of legal consequences, and inconsistent reporting practices can all contribute to underestimation of the problem [16]. Furthermore, our data relies on recorded incidents, meaning it may not accurately reflect the experiences of individuals who do not interact with law enforcement or emergency medical services. This could lead to an underrepresentation of certain populations, such as those experiencing homelessness or those in rural areas with limited access to services.

Second, the "Drug and Alcohol Treatment Facilities" dataset only provides a snapshot of available resources in 2018. The landscape of treatment options may have changed significantly since then, with new facilities opening and others closing. This temporal limitation restricts our ability to assess the current availability and accessibility of treatment services.

Third, relying on county-level aggregations can mask important within-county variations. Opioid use and access to services may differ significantly between urban and rural areas within the same county. This ecological fallacy can lead to misleading conclusions if not considered carefully [17].

Finally, it is important to recognize that correlation does not equal causation. While we may identify associations between economic factors, geographic accessibility, and harm reduction outcomes, further research is needed to establish causal relationships and understand the complex interplay of factors driving the opioid crisis.

Despite these limitations, our analysis provides a valuable foundation for understanding the opioid crisis in Pennsylvania. By acknowledging potential biases and interpreting the findings cautiously, we can use this information to inform targeted interventions and resource allocation strategies. Future research should focus on addressing the identified limitations, incorporating more granular data, and exploring the causal mechanisms underlying observed patterns.

Logistic Regression Model Interpretation

We chose a logistic regression model to test our dataset, using Naloxone administration as the binary response variable. We assigned a value of 1 when Naloxone was administered and 0 when it was not. The predictor variables included a range of factors across socioeconomic, geographic, and temporal dimensions, such as GDP category, county, year, quarter, total population, and gender-specific opioid dispensation and prescription counts.

After successfully fitting the model, we achieved a training accuracy of 92.9% and a testing accuracy of 89.5%, suggesting that the model generalizes well and does not suffer from overfitting. In addition, we implemented confusion matrix with high precision (0.925), recall (0.952) and F1(0.930), this strongly indicates reliability of our logistic model.

To further illustrate the implementation and evaluation of our logistic regression model for Naloxone administration, the following code snippets demonstrate key steps in the process. This includes the model fitting and the generation of the confusion matrix and performance metrics.

#Since logistic model only deals with binary outcomes, and we need to convert variables into with consistency.

```
SUD_Data_Export$CO_YR_QTR <- factor(SUD_Data_Export$CO_YR_QTR)
SUD_Data_Export$COUNTY <- factor(SUD_Data_Export$COUNTY)
SUD_Data_Export$QUARTER <- factor(SUD_Data_Export$QUARTER)
SUD_Data_Export$GDP_CATEGORY <- factor(SUD_Data_Export$GDP_CATEGORY)
SUD_Data_Export$YEAR <- factor(SUD_Data_Export$YEAR)
```

#Our target binary variable NALXONE_ADMINISTRATIONS has different categories, while our model only handles binary outcomes, we need to distinguish them into (0 and 1).

```
SUD_Data_Export$NALOXONE_ADMINISTRATIONS <- trimws(SUD_Data_Export$NALOXONE_ADMINISTRATIONS)
SUD_Data_Export$NALOXONE_ADMINISTRATIONS <- ifelse(
  is.na(SUD_Data_Export$NALOXONE_ADMINISTRATIONS), NA,
  ifelse(SUD_Data_Export$NALOXONE_ADMINISTRATIONS == "overdose reserved", 0, 1)
)
```

#Ensure the model's consistency and no NAs

```
SUD_Data_Export$NALOXONE_ADMINISTRATIONS <- as.numeric(SUD_Data_Export$NALOXONE_ADMINISTRATIONS)
SUD_Data_Export$NALOXONE_ADMINISTRATIONS[is.na(SUD_Data_Export$NALOXONE_ADMINISTRATIONS)] <- 0
```

#Splitting dataset into train and test with p=0.85, which 85% of train and 15% of test dataset.

```
set.seed(987954)
sud_sample_vector <- createDataPartition(SUD_Data_Export$NALOXONE_ADMINISTRATIONS, p = 0.85, list = FALSE)
sud_train <- SUD_Data_Export[sud_sample_vector, ]
sud_test <- SUD_Data_Export[-sud_sample_vector, ]

nzv_cols <- nearZeroVar(sud_train, saveMetrics = TRUE)
sud_train <- sud_train[, !nzv_cols$nzv]
sud_test <- sud_test[, colnames(sud_train)]
```

#Adding all the variables to the model

```
sud_model <-glm(NALOXONE_ADMINISTRATIONS ~ COUNTY + GDP_CATEGORY + YEAR + QUARTER+ TOTAL_POPULATION +
OPIOID_PRESCRIPTIONS_MALE + OPIOID_PRESCRIPTIONS_FEMALE+OPIOID_DISPENSATIONS_FEMALE
+OPIOID_DISPENSATIONS_MALE, data = sud_train, family = binomial("logit"))
summary(sud_model)
```

#train accuracy with: 0.929

```
> train_prediction <- predict(object = sud_model, newdata= sud_train, type= "response")
> head(train_prediction)
        1         2         3         4         5         6
0.9775952 0.9800263 0.9648119 0.6971293 0.8993788 0.9412426
```

```
> train_class_prediction <- as.numeric(train_prediction > 0.5)
> mean(train_class_prediction == sud_train$NALOXONE_ADMINISTRATIONS)
[1] 0.9296978
```

#test accuracy with:0.895

```
> test_prediction <- predict(object= sud_model,newdata= sud_test, type= "response")
> test_class_prediction <- as.numeric(test_prediction >0.5)
> mean(test_class_prediction == sud_test$NALOXONE_ADMINISTRATIONS)
[1] 0.8955224
```

#Implement confusion matrix

```
> confusion_matrix <- table(
+   predicted = train_class_prediction,
+   actual = sud_train$NALOXONE_ADMINISTRATIONS
+ )
> print(confusion_matrix)
         actual
predicted   0   1
        0 591  41
        1  66 824
> precision <- confusion_matrix[2,2]/sum(confusion_matrix[2,])
> print(precision)
[1] 0.9258427
> recall <- confusion_matrix[2,2]/ sum(confusion_matrix[,2])
> print(recall)
[1] 0.9526012
> (f1 = 2*precision*recall/(precision +recall))
[1] 0.9390313
```

References

[1] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, p. 20150202, 2016.

[2] Pennsylvania Open Data, "Pennsylvania opioids," [Online]. Available: https://data.pa.gov/stories/s/Pennsylvania-Opioids/9q45-nckt/. [Accessed: Feb. 2025].

[3] Pennsylvania Department of Health, "Overdose Information Network Data CY January 2018-Current," Pennsylvania Open Data. [Online]. Available: https://data.pa.gov/Opioid-Related/Overdose-Information-Network-Data-CY-January-2018-/hbkk-dwy3/about_data. [Accessed: Feb. 2025].

[4] Pennsylvania Department of Health, "Dispensation Data without Buprenorphine Quarter 3-Current," Pennsylvania Open Data. [Online]. Available: https://data.pa.gov/Opioid-Related/Dispensation-Data-without-Buprenorphine-Quarter-3-/rr54-ur6z/about_data. [Accessed: Feb. 2025].

[5] Pennsylvania Department of Health, "Emergency Medical Services (EMS) Naloxone Dose Administration," Pennsylvania Open Data. [Online]. Available: https://data.pa.gov/Opioid-Related/Emergency-Medical-Services-EMS-Naloxone-Dose-Admin/wst4-3int/about_data. [Accessed: Feb. 2025].

[6] Pennsylvania Department of Corrections, "Inmate Admissions with Substance Use Year 2016-Current," Pennsylvania Open Data. [Online]. Available: https://data.pa.gov/Opioid-Related/Inmate-Admissions-with-Substance-Use-Year-2016-Cur/bvin-4fk2/about_data. [Accessed: Feb. 2025].

[7] Pennsylvania Attorney General's Office, "Opioid Seizures and Arrests CY 2013-Current Quarter," Pennsylvania Open Data. [Online]. Available: https://data.pa.gov/Opioid-Related/Opioid-Seizures-and-Arrests-CY-2013-Current-Quarte/wmgc-6qvd/about_data. [Accessed: Feb. 2025].

[8] Pennsylvania Department of Drug and Alcohol Programs, "Drug and Alcohol Treatment Facilities May 2018-Current," Pennsylvania Open Data. [Online]. Available: https://data.pa.gov/Opioid-Related/Drug-and-Alcohol-Treatment-Facilities-May-2018-Cou/eswt-bam9/about_data. [Accessed: Feb. 2025].

[9] Centers for Disease Control and Prevention, "Dose Dashboard: Nonfatal Surveillance Data," National Center for Injury Prevention and Control. [Online]. Available: https://www.cdc.gov/overdose-prevention/data-research/facts-stats/dose-dashboard-nonfatal-surveillance-data.html. [Accessed: Feb. 2025].

[10] Centers for Disease Control and Prevention, "SUDORS Dashboard: Fatal Overdose Data Accessible," National Center for Injury Prevention and Control. [Online]. Available:

https://www.cdc.gov/overdose-prevention/data-research/facts-stats/sudors-dashboard-fatal-overdose-data-accessible.html. [Accessed: Feb. 2025].

[11] OverdoseFreePA, "Know the Facts," [Online]. Available: https://www.overdosefreepa.org/know-the-facts/. [Accessed: Feb. 2025].

[12] U.S. Bureau of Economic Analysis, "Gross Domestic Product by County 2001 to Current," Pennsylvania Open Data. [Online]. Available: https://data.pa.gov/Census-Economic/Gross-Domestic-Product-by-County-2001-to-Current-B/7ktx-wvbh/about_data. [Accessed: Feb. 2025].

[13] U.S. Census Bureau, "Population by Gender and Age (US Census: ACS 5 Year Estimates) by County," Pennsylvania Open Data. [Online]. Available: https://data.pa.gov/Census-Economic/Population-By-Gender-and-Age-US-Census-ACS-5-Year-/ib65-r5ts/about_data. [Accessed: Feb. 2025].

[14] D. Vaughan, *Analytical Skills for AI and Data Science: Building Skills for an AI-Driven Enterprise*, Sebastopol, CA: O'Reilly Media, 2020.

[15] Johnson T. P. (2014). Sources of Error in Substance Use Prevalence Surveys. International scholarly research notices, 2014, 923290. https://doi.org/10.1155/2014/923290

[16] NIDA. 2022, June 7. Stigma and Discrimination. Retrieved from https://nida.nih.gov/research-topics/stigma-discrimination on 2025, February 21

[17] Pollet TV, Stulp G, Henzi SP, Barrett L. Taking the aggravation out of data aggregation: A conceptual guide to dealing with statistical issues related to the pooling of individual-level observational data. Am J Primatol. 2015 Jul;77(7):727-40. doi: 10.1002/ajp.22405. Epub 2015 Mar 24. PMID: 25810242.

[18] "Naloxone", Pennsylvania Department of Health. [Online]. Available: https://www.pa.gov/agencies/health/programs/opioids/naloxone.html. [Accessed Mar. 28, 2025].

[19] IndexMundi, "Pennsylvania - Land area," [Online]. Available: https://www.indexmundi.com/facts/united-states/quick-facts/pennsylvania/land-area#table. [Accessed: April 9, 2025]

[20] data.pa.gov, "Estimated Drug Overdose Deaths (CY 2012-Current) County," [Online]. Available: https://data.pa.gov/Opioid-Related/Estimated-Drug-Overdose-Deaths-CY-2012-Current-Cou/azzc-q64m/about_data. [Accessed: April 6, 2025]

Appendix A

Potential Timeline

Phase 1: Business Understanding

- Acquiring knowledge on harm reduction strategies for SUD.
- Understanding the groups that we are serving to provide the best data related to legal opportunities for SUD treatments in geographic areas.

Phase 2: Planning and Data Collection

- Literature review on harm reductions strategies, data availability, and relevant research methodologies.
- Identify and access relevant data sources from the Commonwealth of Pennsylvania.
- Develop a data analysis plan.

Phase 3: Data Analysis and Interpretation

- Clean, transform, and analyze data using appropriate statistical methods (e.g. regression analysis, spatial analysis).
- Generate visualizations to communicate key findings.
- Interpret results and draw initial conclusions.

Phase 4: Modeling (Time series or Linear Regression)

- Developing a time series prediction model based on historical data and forecasting future churn rate and retention percentage.
- Developing a linear regression model using continuous data to find insights on recovery efforts for the population.

Phase 5: Evaluation

- Evaluating data of the discontinuation rates of participation in SUD reduction/elimination programs and drawing conclusions thereon.
- Finding success rates of various SUD treatments and their correlations with geographical data.

Phase 6: Dissemination and Reporting

- Prepare a final report summarizing key findings from the analysis, including geographic insights, statistical correlations, and trends in harm reduction accessibility.
- Propose specific policy interventions or resource allocation strategies to enhance harm reduction efforts, such as targeted Narcan distribution or expanded syringe service programs in underserves areas.
- Present findings to the Penn State Harrisburg research team and relevant stakeholders, ensuring the insights are actionable and relevant to public health efforts.
- Explore opportunities for publishing or presenting findings, including academic journals, public health conferences, or community forums.

Appendix B

Feasibility Estimate

Substance Use Disorder is an ever-growing concern in the state of Pennsylvania and everywhere else in this country. Correspondingly, data should be readily available for this project. Geographic data to pair with the SUD research data should be abundantly available from many governmental and educational websites and databases. (U.S. Census Bureau, n.d.; Pennsylvania Open Data, n.d.) Given that fact, it should be feasible to successfully split this project amongst our group members and then form cohesive conclusions to present to all stakeholders and interested parties.

Appendix C

Anticipated Results / Deliverables

- A comprehensive report summarizing the findings of the research, including:
    - An assessment of the availability and accessibility of harm reduction strategies across the Commonwealth of Pennsylvania.
    - An identification of factors associated with the availability of these strategies.
    - Data-driven recommendations for improving local access to harm reduction strategies.
- Data visualizations and presentations to communicate findings to the research team other stakeholders.
- Potential publications or presentations at academic conferences.

Appendix D

Tables and Figures

- Figure 1: Opioid Seizures and Arrests CY 2013-Current Quarter: The graph displays the overall trend of opioid incidents over time. As the description mentioned, after the year 2023 there are significant drops. This is due to the limitation of the dataset. The data for the year 2023 and after was only partially present.

- Figure 2: Dispensation Data without Buprenorphine 2016-2023: The bar graph explores deeper insights into the relationship between opioid dispensations and prescriptions for different genders. Providing different aspects and reinforcement for future opioid reduction strategies.

- Figure 3: Overdose Information Network Data CY 2018-2023: The scatterplot visualizes naloxone administration over time and analyzes the naloxone administration demand by gender.

- Figure 4: Gross Domestic Product by County 2001 to Current: From an economic aspect, the graph demonstrates how different counties' GDP is affected by opioid usage. Also, clarify the relationship between opioid usage and GDP.

- Figure 5: Population by Gender and Age (US Census: ACS 5 Year Estimates) by County: Counties with different population levels impact on assessment of opioid harm reduction services and identifying disparities in service availability.