Problem Chosen	2024 MCM/ICM Summary Sheet	Team Control Number
C		2406434

Comprehensive Analysis and Solutions for the U.S. Opioid Crisis

Summary

Before 2010, OxyContin was a prescription opioid drug prone to misuse, becoming a key factor in the opioid crisis in the United States. In response to the abuse issue, Purdue Pharma introduced a new formulation of OxyContin. However, this led abusers to shift towards more dangerous illegal opioids, such as heroin. The opioid crisis also impacted key sectors of the U.S. economy, and if the crisis permeates through various layers of society, it could lead to severe consequences. Therefore, addressing the crisis of both synthetic and non-synthetic opioids is crucial. By analyzing data from 2010 to 2017, we aim to provide relevant strategies to combat this national emergency.

In response to Part1, we conducted a detailed data analysis of the drug identification counts provided by NFLIS for the years 2010-2017, after that, we established a **Susceptible-Infectious-Susceptible Model** to explore potential patterns and features within the data. In the model, we set a threshold where, when σ exceeds 1, government intervention is required. Based on our model, we predicted specific counties that would require intensified control measured in 2018, with NEWPORT NEWS CITY, RICHMOND, and LYNCHBURG CITY ranking as the top three counties warranting special attention. Additionally, to identify possible locations where specific opioid drugs might have already begun to be abused, we constructed a Frequent Pattern Mining Model and applied the **FP-Growth tree** algorithm to analyze possible locations of drug abuse for each year from 2010 to 2017. We modify traditional model because the data is not plentiful.

Regarding the second part of the problem, we combined ensemble methods, **filtering methods (SBF - Selection By Filtering)** with the **Recursive Feature Elimination (RFE) algorithm** for dimensionality reduction. Considering that filtering methods may remove variables with substantive significance, our study primarily focused on the RFE algorithm.Differing from traditional methods and to maintain the integrity of all 16 socioeconomic factors, we initially used the Recursive Feature Elimination (RFE) algorithm, resulting in 48 variables before applying RFE again to these 48 variables.Finally, We used the results from **bagged decision trees** as the basis for our research, selecting 15 socio-economic factors that could potentially influence drug propagation. Subsequently, we performed regression analysis between these 15 variables and the model developed in the first part to delve into the impact of socio-economic factors on drug propagation.

In the third part, we employed **PCA** to reduce the dimensionality of the data from 15 variables to 3. To test the effectiveness of our proposed strategies, we introduced **a suppression factor G** and constructed the interaction term GX to represent the impact of the suppression factor on specific independent variables. Through analysis, we observed that the suppression factor had an impact on most variables and successfully achieved suppression. These results demonstrate the efficacy of our strategies.

In summary, we comprehensively addressed all three parts of the problem by establishing sound mathematical models, conducting data analysis and predictions, describing data characteristics and patterns, and proposing control strategies tailored to specific regions.

Keywords: Frequent Pattern Mining Model SIS Model Regression RFE SBF

Contents

1 Introduction

1.1 Background

Prior to 2010, OxyContin was a widely abused prescription opioid. Its highly addictive nature and ease of misuse made it a central issue in America's opioid crisis. Abusers often crushed the pills to snort or inject, releasing the active ingredients rapidly .

In 2010, Purdue Pharma, the manufacturer of OxyContin, introduced a new formulation. This version was harder to crush and turned into a gel when mixed with water, making it more difficult to misuse for snorting or injecting. The goal was to reduce the abuse and dependency on the drug. However, this change led to serious unintended consequences. As the traditional form of OxyContin became harder to obtain and abuse, some addicts turned to other opioids, including more dangerous illegal ones like heroin and fentanyl. This shift exacerbated the opioid abuse problem, leading to a surge in illegal opioid use and related deaths. Although the reformulated OxyContin had some success in reducing abuse of the drug itself, it inadvertently fueled the overall opioid crisis. The U.S. now faces a national crisis involving both synthetic and non-synthetic opioids, used both legally for treatment and pain management, and illegally for recreational purposes. The negative health impacts of opioid abuse include opioid use disorder, hepatitis, HIV infection, and neonatal abstinence syndrome, among others. The opioid crisis also affects key sectors of the U.S. economy. As the crisis permeates all social strata, including the highly educated, filling positions requiring precision skills, assembly of high-tech components is challenged. Moreover, with the increasing proportion of opioid addiction among the elderly, healthcare costs are rising, and staffing in assisted living facilities is impacted.

Therefore, it is imperative to develop relevant strategies to address the nationwide crisis of both synthetic and non-synthetic opioids, based on past data. We present our strategies by analyzing data from 2010 to 2017.

1.2 Problem Restatement and Problem Analysis

- Part 1:
  (1)Utilize data from the NFLIS to establish a mathematical model describing the spread and characteristics of synthetic opioids and heroin across five states and their counties. We approached this by constructing an infectious disease model.This issue is a mathematical problem of fitting. Our first step is to establish a model that uses a fitted model to describe the spread and characteristics of synthetic opioids and heroin among five states and their counties. The data provided by the National Forensic Laboratory Information System (NFLIS) includes counts of each drug in every county of these states. Our analysis indicates that the spread of these drugs is akin to the spread of diseases, and similar mathematical methods can be applied. We have developed an infectious disease model to represent the spread and characteristics of synthetic opioids and heroin.
  (2)Use your model to identify potential locations in these five states where specific opioids may have begun to be used.We treat each county as a unit and regard the reported synthetic opioid and heroin incidents within each county as individual transactions, akin to shopping data mining. We will create a frequent pattern mining model to identify the originating counties of specific opioids.
  (3) What specific concerns should the U.S. government have if the patterns and characteristics

identified by our team continue? We use the established infectious disease model to predict the potential patterns and characteristics of drug occurrences in 2018. We interpret the term ”pattern” as whether the drug spread might become uncontrolled, similar to the spread of infectious diseases. However, it is noteworthy that uncontrolled spread in drugs not only includes a rapid increase in usage but also a significant decrease to very low levels. This is due to the clinical need for some of these drugs; a drastic reduction or disappearance could impact the stable operation of the medical field. The specific concerns for the U.S. government will be addressed following the resolution of the next issue.
(4)At what drug identification threshold levels do these situations occur?
(5)Our model's prediction on when and where these events will occur?We have utilized our established infectious disease model to forecast the spread and characteristics of specific drugs in 2018, thereby determining the regions where drug spread is uncontrolled. The U.S. government should provide appropriate intervention in these uncontrolled areas.

- Part 2:
Using socio-economic data from the U.S. Census, address the following: Numerous competing hypotheses exist to explain how opioid use reached current levels, who uses or misuses opioids, what drives the growth in opioid use and addiction, and why, despite known risks, opioids continue to be used.
(1)Is use or the trend of use related to the socio-economic data provided by the U.S. Census? We employed RFE to select significant variables, then used differential analysis to validate the RFE results.
(2)If so, modify the model from Part 1 to include any significant factors from this dataset. We incorporated these variables into the infectious disease model through multivariate regression to derive a new model and conclusions on control loss.

- Part 3:
(1)Combine results from Parts 1 and 2 to identify potential strategies to combat the opioid crisis.
(2)Test the effectiveness of this strategy using your model.
(3)Determine the boundaries of any critical parameters on which success (or failure) depends.
(4)In addition to the main report, write a 1-2 page memo for the chief administrator, summarizing any significant insights and results from your modeling work using the DEA/NFLIS database.

1.3 Our work

Our work is shown in Figure 1.Blue represents Part 1, green represents Part 2, orange represents Part 3, and the purple section is the sensitivity analysis. The specific steps are as follows:
Step 1: Firstly,we preprocess the data.Considering the high volume of missing values and limited data (only eight years per drug), we filtered out counties with less than three years of drug data and those with scarce drug occurrence.
Step 2: Next we solve Part 1 using the processed data. We found the drug spread similar to infectious diseases, hence we used an infectious disease model to describe the spread and characteristics of synthetic opioids and heroin across the states and counties.
Step 3: By establishing frequent pattern mining model, we identified potential locations in these

states where specific opioids might have begun to be used.
Step 4: As for when a drug might become uncontrollable, which is the threshold level mentioned in the topic
Step 5: We predicted the possible patterns and characteristics of drugs in 2018 by establishing an infectious disease model.
Step 6: After that,we provide conclusions and suggestions on what concerns the U.S. government should have.
Step 7: To determine whether the use or trend of opioid use is related to the provided U.S. Census socio-economic data,we selected variables significantly affecting opioid use from the U.S. Census socio-economic data using Recursive Feature Elimination (RFE), and then verified the RFE results through differential analysis to demonstrate our reliability.
Step 8: we conducted a multivariate regression with the results obtained from the infectious disease model and the chosen independent variables. This led to the development of a new model and results, ultimately determining whether there is a loss of control.
Step 9: Combining results from Parts 1 and 2, we identified potential strategies to combat the opioid crisis, tested the strategy's effectiveness, and determined the boundaries for critical parameters for success or failure.
Step 10: Write a 1-2 page memo for the chief administrator, summarizing significant insights or results from our modeling work using the DEA/NFLIS database.
In a nutshell, we first preprocessed the data, filtering out counties with limited drug data. We then modeled the drug spread using an infectious disease model, identifying potential outbreak locations. To assess when drug usage might become uncontrollable, we used clustering methods. Next, we employed Recursive Feature Elimination (RFE) to select significant variables from socio-economic data, verifying these results with differential analysis and multivariate regression. This led to a new model assessing loss of control. Finally, we combined these insights to devise strategies against the opioid crisis and summarized our findings in a memo for policy guidance.

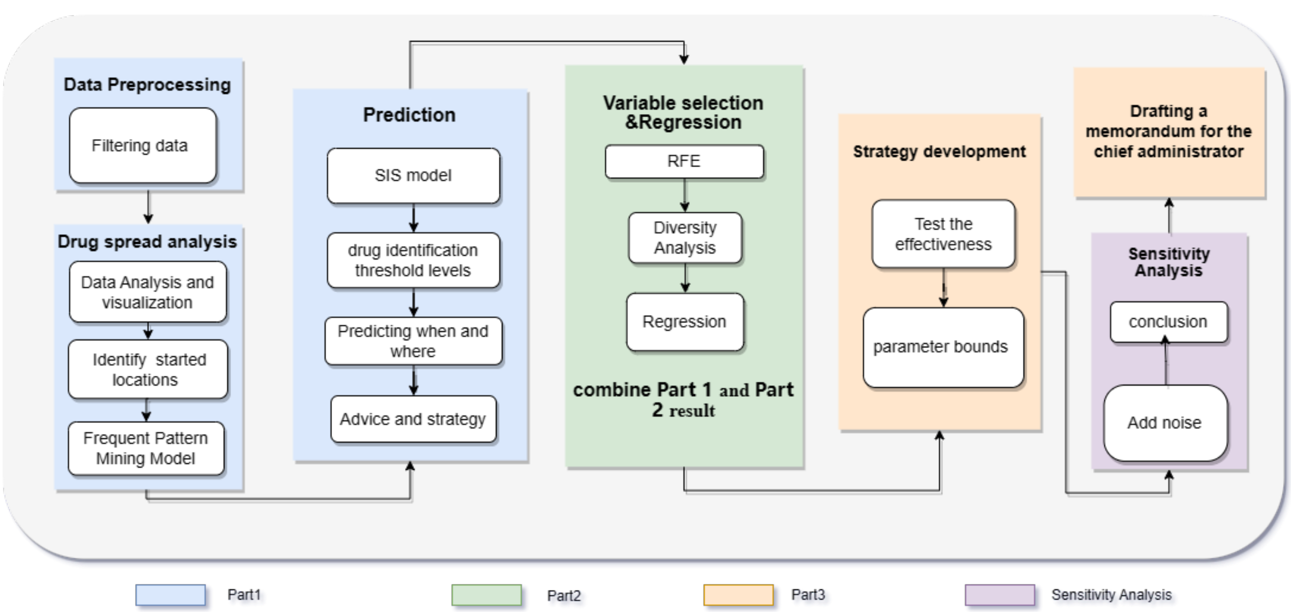

Figure 1: Our Work

2 Assumptions and Notations

2.1 Assumptions and justification

- **Assumption1:** The provided county location data is accurate.
  Justification: The dataset (MCM_NFLIS_Data.xlsx) contains the counts of narcotic analgesics (synthetic opioids) and heroin drug identifications reported in each county of these five states from 2010 to 2017. Typically, when law enforcement agencies submit these samples, they provide location data along with their incident reports. However, in cases where such location data is not provided, the crime laboratory uses the location of the submitting law enforcement agency. To address this issue, we assume that the provided county location data is accurate.

- **Assumption2:**We assume that the rate at which drug crimes are detected by police in each county does not vary significantly.

- **Assumption3:**We disregard the impact of double-counting due to cross-county drug crimes on the data.

- **Assumption4:**Drug cases have similarities to the transmission patterns of epidemics.
  Justification: Through analysis, we found that the spread of drugs is similar to infectious disease models. Therefore, we have constructed an infectious disease model to describe the spread and characteristics of synthetic opioids and heroin across the five states and their counties.

- **Assumption5:**We assume that criminals involved in drug crimes do not become rehabilitated upon release from prison.
  Justification: The infectious disease model we developed is an SIS model, which implies that after release from prison, criminals will continue to spread drugs to other areas.

- **Assumption6:**We assume that drugs spread only within specific populations, thus disregarding the possibility of casual drug exposure among the general population and, consequently, the absence of exposed individuals.
  Justification: As we have developed an SIS model, it aligns with the assumptions of the SIS model, which disregard the possibility of casual drug exposure among the general population.

2.2 Model Notations

Table 1: Notations 1

Symbol	Specific definitions
G	The proportion of U.S. government's financial investment
GX	The impact of the suppression factor on a given independent variable.

Table 2: Notations 2

Symbol	Definition	Specific definitions
x_0		Fitted values of the infectious disease model.
x_1	HC01_VC03	Total households
x_2	HC01_VC09	Male householder, no wife present, family
x_3	HC01_VC10	Male householder, no wife present, family - With own children under 18 years
x_4	HC01_VC11	Female householder, no husband present, family
x_5	HC01_VC12	Female householder, no husband present, family - With own children under 18 years
x_6	HC01_VC13	Nonfamily households
x_7	HC01_VC14	Householder living alone
x_8	HC01_VC15	Householder living alone - 65 years and over
x_9	HC01_VC25	Population in households
x_{10}	HC01_VC62	Responsible for grandchildren
x_{11}	HC01_VC70	Who are female-Number of grandparents responsible for own grandchildren under 18 years
x_{12}	HC01_VC118	Same house
x_{13}	HC01_VC120	Different house in the U.S. - Same county
x_{14}	HC01_VC131	Native - Born in United States - State of residence
x_{15}	HC01_VC161	Latin America

3 Part 1

3.1 Data Analysis and Preprocessing

3.1.1 Addressing the missing values

The attached data file $MCM_NFLIS_Data.xlsx$provides us with 8 years (from 2010 to 2017) of data. It shows drug reports in terms of one county,whole counties and whole states over years. While the background maintains its correction, missing values may exist. Thus, we first examine missing values and find that no missing value exists.

3.1.2 Data analysis

Initially, we compiled statistics on all drugs, with Figure 2 depicting the proportion of each drug, showing Heroin as the most prevalent at 51%, followed by Oxycodone at 18%, and smaller percentages for other substances like Fentanyl and Hydrocodone, each constituting 8% of the total.

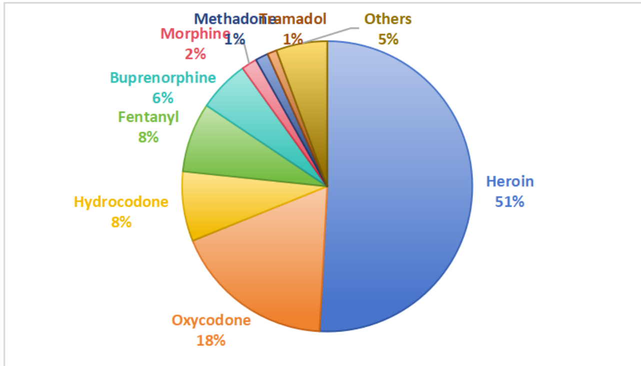

Figure 2: Pie Chart of Drug Quantity Proportions

For various substances, we selected those with higher representations in the pie chart for a focused correlation analysis. By employing Pearson correlation analysis, as illustrated in Figure 3, we discerned the linear relationships between pairs of substances. The resulting correlation matrix employs a color-coded system where the intensity of the blue indicates the strength of a positive correlation and the intensity of the red reflects a strong negative correlation.

Upon examining Figure 3, we observed a pronounced blue correlation between Oxycodone and Hydrocodone, suggesting that these substances are often associated with one another in usage patterns. The size of the circles confirms the robustness of this positive correlation. In contrast, Heroin and Tramadol displayed a noticeable red color, implying an inverse relationship — as the prevalence of one increases, the other decreases.

The significance of these correlations extends beyond statistical measures. The strong negative correlations could indicate substitution patterns, where one drug may be used in place of another. Similarly, strong positive correlations might reflect a combination of drug use, which could have implications for overdose risks and treatment strategies.

In conclusion, Figure 3 provides substantial evidence of significant correlations between substances, with the potential to inform clinical practice and public health policies. The data elucidates the complex interplay between different drugs and can serve as a basis for further research into causation and intervention approaches aimed at reducing harm from substance use.

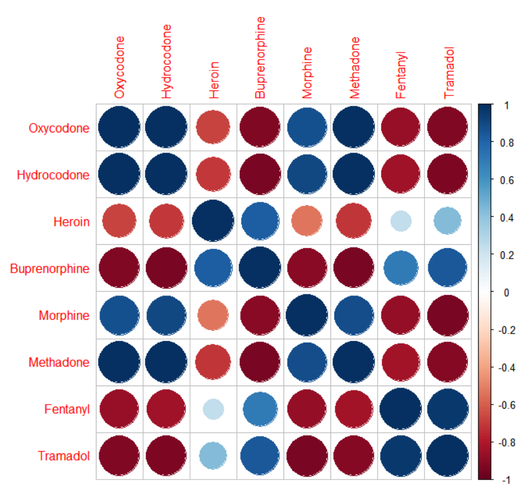

Figure 3: Partial Drug Correlation Graph

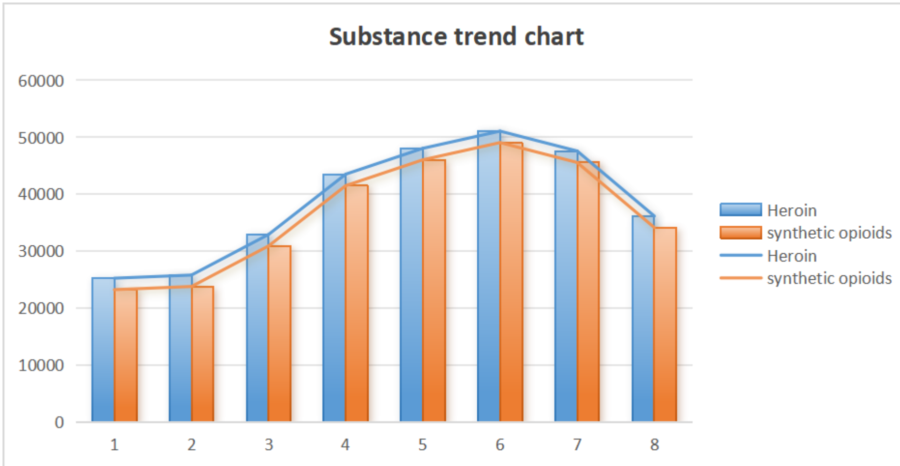

Figure 4: Drug Quantity Trend from 2010-2017

Additionally, The distribution chart of the total number of drugs, including Heroin and synthetic opioids across all states and counties, is shown in Figure 4. The red bars represent the quantity of synthetic opioids from 2010 to 2017, while the blue bars represent the quantity of Heroin for the same period. The blue and red lines indicate the changes in the quantity of Heroin and synthetic opioids, respectively, from 2010 to 2017. It is evident that the total number of drugs generally increased from 2010 to 2015 and showed a decreasing trend from 2016 to 2017.

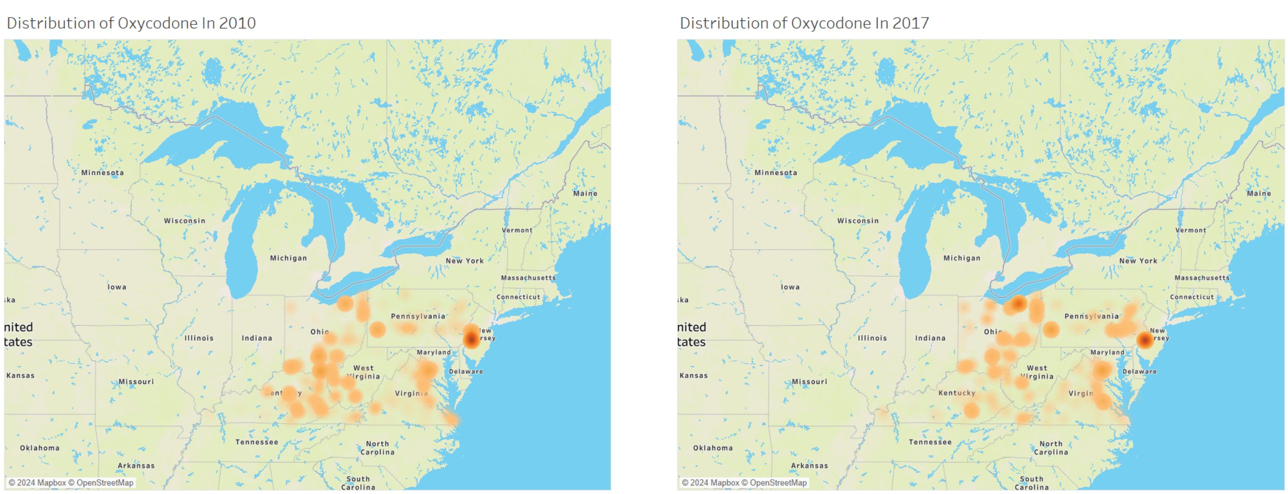

Figure 5: Distribution of Oxycodone in 2010 and 2017

Due to the vast variety of drugs, we only present the distribution maps of Oxycodone, the drug with the highest proportion, for the years 2010 and 2017 in Figures 5 . Different colors represent different counties, with the size of the circles indicating the quantity of Oxycodone; larger circles signify higher drug quantities. It is clearly observable that in CUYAHOGA county, the quantity of Oxycodone in 2017 has significantly increased compared to 2010.

3.2 Model 1:Frequent Pattern Mining Model

3.2.1 Specific steps of the model

After we further describe the spread and characteristics of the reported synthetic opioid and heroin incidents (cases) in and between the five states and their counties,we will establish a frequent pattern mining model to identify the originating counties of specific opioids. We consider a county as a unit and treat each county's reported synthetic opioid and heroin incidents as a transaction, analogous to shopping data mining. In the study of frequent pattern mining, well-designed association rule algorithms are crucial. Traditional measurement methods usually include lift and confidence; however, these may not reflect the actual connection in certain special cases. Therefore, we introduce the Kulczynski measure (Kulc) and the Imbalance Ratio measure (IR) to better depict the association rules between cases and states. The formula for Kulc is:

$$Kulc(A,B) = \frac{1}{2}(P(A|B) + P(B|A))$$

$$sup(A) = \frac{\{\text{Events with } A\}}{\{\text{All Events}\}}$$

$$IR(A,B) = \frac{|sup(A) - sup(B)|}{sup(A) + sup(B) - sup(A \cup B)}$$

By cooperatively using Kulczynski measure (Kulc) and Imbalance Ratio measure (IR) , we can fully depict their association degree.

Regarding the identification of the originating counties of specific opioids, since we found in the previous data description that some drugs are not very abundant in some counties, we used an improved FP-Growth tree algorithm. This aims to mine association rules of low-frequency itemsets.

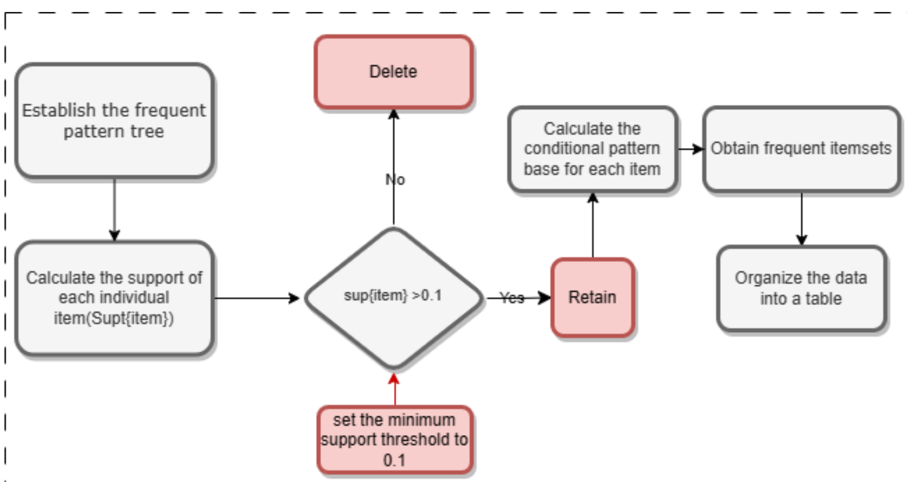

Figure 6: FP-Growth tree algorithm

The steps of Frequent Pattern Mining Model show in figure 6.

Step 1: Establish the frequent pattern tree (FP-tree): Scan the entire dataset to calculate the support count, sup(item), for each item. Items that meet or exceed a predefined minimum support threshold are retained. Sort these items in descending order of support count and construct the FP-tree by inserting transaction sequences according to this order, while incrementing the count of existing nodes when the same item sequence recurs.

Step 2: Determine the conditional pattern base for each item: For items in the FP-tree, trace the node links to find all the prefix paths leading up to each item. Each path is a sequence of nodes from the root of the FP-tree to the node just before the target item node. These prefix paths, collectively known as the conditional pattern base, represent all the transactions in which the target item appears.

Step 3: Obtain frequent itemsets from the conditional pattern base: Use the conditional pattern bases to construct conditional FP-trees for each item. Then, recursively repeat the process for each conditional FP-tree to find subsets of frequent itemsets. During each recursion, combine the item for which the conditional FP-tree was made with the frequent itemsets found in the tree to generate new itemsets.

Step 4: Organize the data into a table: Compile the frequent itemsets into a structured format, such as a table, listing the itemsets along with their corresponding support counts. This table serves as a comprehensive summary of all frequent itemsets discovered in the dataset, which can be used for further analysis or to generate association rules.

We believe that when the support level is less than 0.05, the data volume is insufficient to support our identification of its potential location. Therefore, we set the minimum support threshold at 0.05, which also helps to expedite the computation of the model. By organizing the data from the five states from 2010 to 2017, we ultimately obtained the following results.

3.2.2 Model results

As the Kluc measure approaches 1, there is a stronger positive association between two items; conversely, when the Kluc measure is closer to 0, the association is more negatively strong. Additionally, as indicated by the formula, the closer the IR (Imbalance Ratio) measure is to 1, the stronger the data imbalance; conversely, the closer IR is to 0, the weaker the data imbalance. Appendix 1 shows the maximum support, Kluc measure, and IR measure for each substance in a specific state, as a result of data processing. When there is a strong association between two items, the ideal scenario is a support of 1, Kluc measure of 1, and IR measure of 0. Therefore, although we use maximum support as the criterion, due to differences in Kluc and IR measures, the possibility of transactions in Appendix 1 might be lower compared to other states. In other words, if we assign different weights to these three indicators, the outcomes will differ. Using support to determine the frequency of transactions between items is a very basic method.

According to the definition of the IR measure, the stronger its skewness, the larger the data volume, and the more frequent its occurrence in the dataset. This can serve as a supplement to our previous analysis of the spread and characteristics of synthetic opioids and heroin among the five states and their counties. From the data in Appendix 1, we can deduce, for example, that in 2010, Table 4 shows the most likely origins of certain drugs. Apart from Propoxyphene, which is most likely to originate from the State of Ohio (OH), the other drugs in Table 4 are most likely to originate from the Commonwealth of Virginia (VA). The results for 2011-2017 are in Appendix 1.

Table 3: The best connectivity between the substance and the state in 2010

SubstanceName	State	Support	Kluc	IR
Propoxyphene	OH	0.11011236	0.467215256	1.517241379
Oxycodone	VA	0.271910112	0.620425572	3.307086614
Morphine	VA	0.166292135	0.423011181	2.212598425
Methadone	VA	0.2	0.502270834	2.307086614
Hydromorphone	VA	0.139325843	0.453553948	1.165354331
Hydrocodone	VA	0.258426966	0.591981087	3.251968504
Heroin	VA	0.186516854	0.471370957	2.25984252
Buprenorphine	VA	0.161797753	0.407175907	2.291338583
Tramadol	VA	0.121348315	0.430340361	0.976377953

3.3 Prediction and the determination of thresholds

3.3.1 SIS Model

The Susceptible Infected Susceptible (SIS) model is a model in epidemiology, where an epidemic is characterized by its ability to spread through a social network.

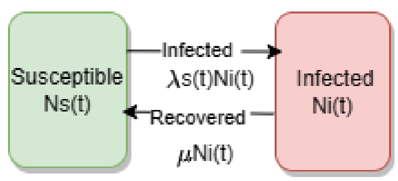

Figure 7: SIS Model

Considering that opioid transactions also spread through social networks, and transactions facilitate drug cases with a spreading effect, we draw an analogy between the two. Therefore, we use the Susceptible Infected Susceptible Model (SIS) to predict the number of drug reports in each county for 2018. According to U.S. state laws, sentences for drug cases generally range from several months to a few years. As a result, addicts are likely to become susceptible again, validating this model. The basic formula of the model is:

$$\frac{di(t)}{dt} = \lambda \cdot (1 - i(t)) \cdot i(t) - \mu \cdot i(t)$$

$$i(0) = i_0$$

In this model, λ represents the daily contact rate, and μ represents the daily recovery rate. The model is capable of identifying time-stationary sequences, decreasing sequences, and increasing sequences, with the latter being the sequences that require intervention.

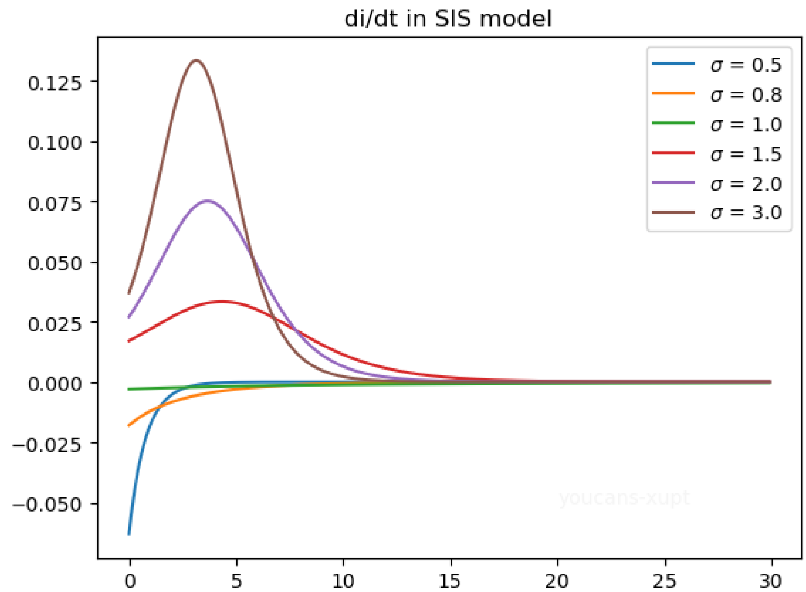

Figure 8: $\frac{d_i}{d_t}$ in SIS model

Figure 8 displays the function graph of the Susceptible Infected Susceptible (SIS) model. Here, σ is defined as$\frac{\lambda}{\mu}$. Consequently, when the contact rate exceeds the recovery rate, the number of infections rises; conversely, when the contact rate is less than the recovery rate, the number of infections declines. For authorities, the continual spread of an infectious disease is undesirable, as when there are sufficient pathogens, the susceptible population tends to increase exponentially. Therefore, σ=1 can be considered a threshold. When $\sigma > 1$, the government needs to implement control measures. However, considering the limitations of fiscal expenditure, we can prioritize based on the magnitude of σ. Therefore, our next step is to estimate the parameters of this model.

Our research initially employed the Runge-Kutta method, a commonly used numerical technique for solving differential equations. This method allowed us to approximate the infection dynamics within the SIS model, thereby simulating the spread of drug crimes. The Runge-Kutta method is advantageous due to its accuracy and stability, enabling us to effectively estimate the dynamic changes in the model.

Subsequently, we utilized grid search to optimize the model parameters. Grid search is a method that involves exploring all possible combinations within the parameter space to identify the optimal parameter settings. Although this approach is computationally intensive, it ensures that no possible parameter combinations are overlooked, thereby finding the most suitable settings for our data.

During the parameter optimization process, we relied on the squared loss function to assess the model's fit. The squared loss function measures the discrepancy between the model's predicted values and the actual values. By minimizing this loss function, we can ensure that our model accurately reflects the actual spread of drug crimes as closely as possible.

Finally, we used the optimized model to predict and analyze the spread of drug crimes in each county. Through these analyses, we successfully identified counties that require special vigilance from the government. These counties might be hotspots for drug crimes or areas where the spread of drugs is rapidly increasing. Our research results will provide crucial decision-making support for the government, helping them to implement more effective strategies in controlling drug crimes.

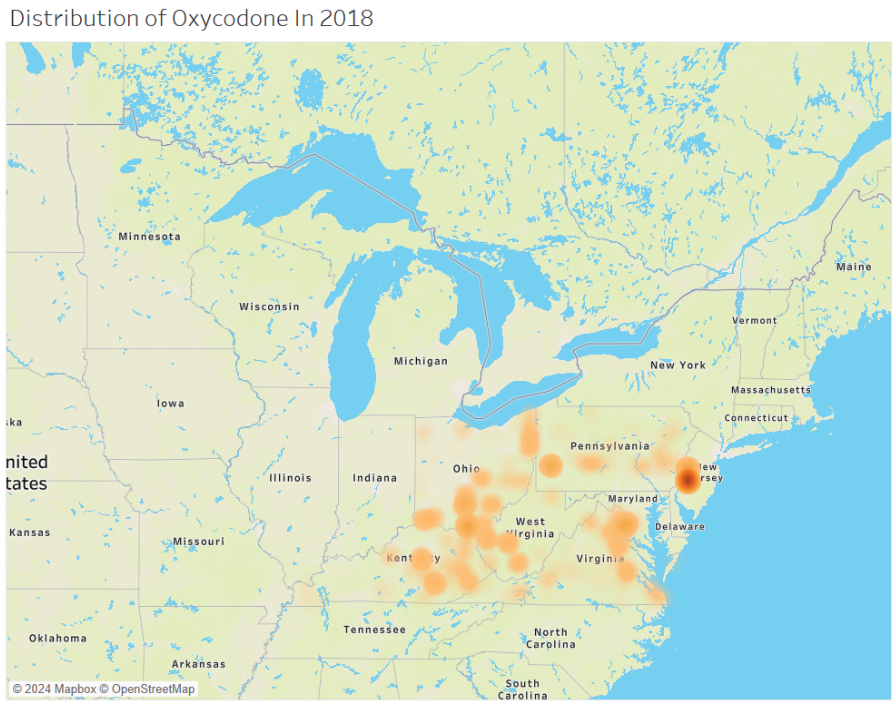

Figure 9: Predicted distribution of Oxycodone in 2018.

Figure 9 presents the predicted distribution of Oxycodone across various regions in 2018. In this figure, the varying shades of color represent the quantity of Oxycodone. The darker the color, the greater the quantity of Oxycodone in that region, indicating a potentially more severe drug abuse problem. Conversely, areas with lighter colors signify lower quantities of Oxycodone, suggesting relatively milder drug-related issues. Such a visual analysis allows us to intuitively understand the distribution of Oxycodone in different regions, providing valuable insights for the government and relevant departments to tailor their intervention measures accordingly.

3.4 Advice and strategies

According to our research analysis, Table 4 identifies NEWPORT NEWS CITY, RICHMOND, and LYNCHBURG CITY as the top three counties requiring special vigilance from the government. These counties are not only potential hot spots for drug crimes but also areas where drug spread is exceptionally rapid. For instance, factors such as socio-economic conditions, geographical location, and population density might make these counties more susceptible to drug crimes. Additionally, specific characteristics of these counties, like convenient transportation and close connections with other regions, may facilitate the rapid spread of drugs.

Table 4: Drug Report Statistics by County

COUNTY	SubstanceName	σ	drugreports
NEWPORT NEWS CITY	Oxycodone	20	20.63095468
RICHMOND	Morphine	20	16.04982264
LYNCHBURG CITY	Hydrocodone	20	9.937335552
LUZERNE	Oxymorphone	20	7.556160067
ROANE	Morphine	20	7.043637143
CAMBRIA	Oxymorphone	20	6.656343616
HAMPSHIRE	Morphine	20	6.17508836
SENECA	Methadone	20	5.374527563
MONTGOMERY	Oxymorphone	20	5.268171184
NORTHAMPTON	Methadone	20	5.024440719
MERCER	Methadone	20	4.481586298
STAUNTON CITY	Morphine	20	4.254988775

Our findings are significant for the government in formulating strategies to prevent and control drug crimes. Firstly, the government can use this data to more accurately pinpoint high-risk areas for drug crimes, allowing for more effective resource allocation and measure implementation. For example, the government could increase police presence in these high-risk counties, intensify efforts to combat drug crimes, or enhance rehabilitation support and education for drug abusers.

Secondly, these research results can also aid the government in better understanding the mechanisms and factors influencing drug crime propagation, enabling more targeted preventive measures. For instance, the government could raise public awareness about drug misuse through community education programs or collaborate with community organizations to provide more support and resources to at-risk individuals.

In summary, our research provides valuable data and insights for the government, enabling them to more effectively tackle the issue of drug crimes, thus safeguarding the safety and stability of society.

In Figure 10, counties represented by different colors indicate the areas requiring varying levels of control measures. The size of each circle reflects the magnitude of the county's θ value, which directly indicates the degree of control required by the government in that area. Specifically, the larger the circle, the higher the θ value, implying that the government should exert greater control efforts in these areas to effectively curb and reduce drug crime activities.

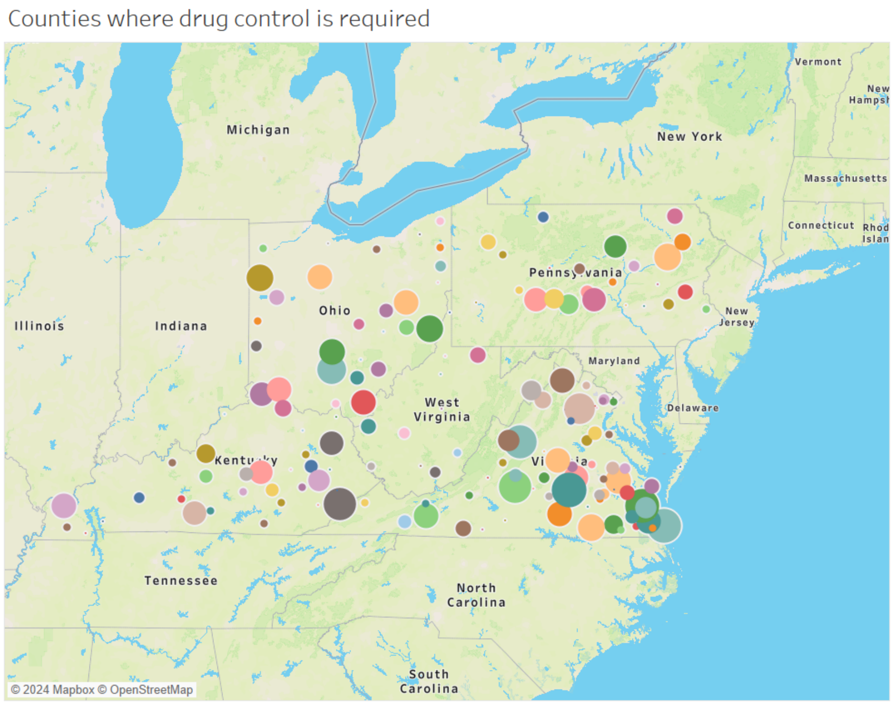

Figure 10: Counties where drug control is required

4 Part 2

To preliminarily explore whether socioeconomic data is associated with opioid usage, we randomly selected one column from the provided 16 categories of survey data, including HOUSEHOLDS BY TYPE, RELATIONSHIP, MARITAL STATUS, etc. We then conducted a basic exploration to see if there is any correlation between them and opioid usage. The resulting correlation chart is shown in Figure 11.

4.1 Feature selection

4.1.1 RFE and SBF

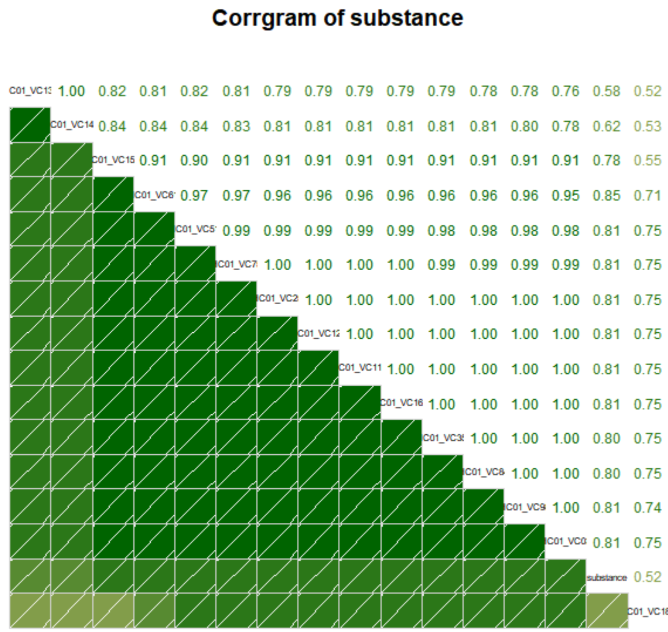

Figure 11: Correlation Graph

It is evident that there is a strong correlation between drug usage and socioeconomic factors. Therefore, we attempt to use appropriate feature selection methods to choose from these 135 variables. Feature selection is divided into wrapper methods and filter methods. Wrapper methods integrate the feature selection process with the training process, using the model's predictive ability as the final criterion for selection. In this study, we employ the RFE (recursive feature elimination) algorithm, using RMSE, Rsquared, and MAE as evaluation criteria. Filter methods conduct feature selection independently of the training process, screening variables based on internal relationships such as variable correlations, near-zero variance tests, and cluster analysis. The variables selected through filter methods are then used for constructing and evaluating different training models. In this research, we use the SBF (Selection By Filtering) algorithm. However, considering that filter methods may exclude variables of substantial significance, our primary research algorithm is the RFE algorithm.

The RFE algorithm aims to obtain the most effective combination of variables that maximizes model performance by adding or removing specific feature variables. The specific algorithm is as follows:

In this study, due to the large number of socioeconomic factors and their strong intercorrelations, we first used the RFE algorithm to select 48 independent variables from 16 socioeconomic factors. We then applied the RFE algorithm to these 48 variables, employing linear regression and bagged decision tree models to rank the independent variables. The changes in RMSE and Rsquared during the ranking process are shown in Figures 11.

Algorithm 1 Recursive feature elimination incorporating resampling

```
for Each Resampling Iteration do
    Partition data into training and test/hold-back set via resampling
    Tune/train the model on the training set using all predictors
    Predict the held-back samples
    Calculate variable importance or rankings
    for Each subset size S_i, i = 1 ... d do
        Keep the S_i most important variables
        [Optional] Pre-process the data
        Tune/train the model on the training set using S_i predictors
        Predict the held-back samples
        [Optional] Recalculate the rankings for each predictor
    end for
end for
Calculate the performance profile over the S_i using the held-back samples
Determine the appropriate number of predictors
Estimate the final list of predictors to keep in the final model
Fit the final model based on the optimal S_i using the original training set
```

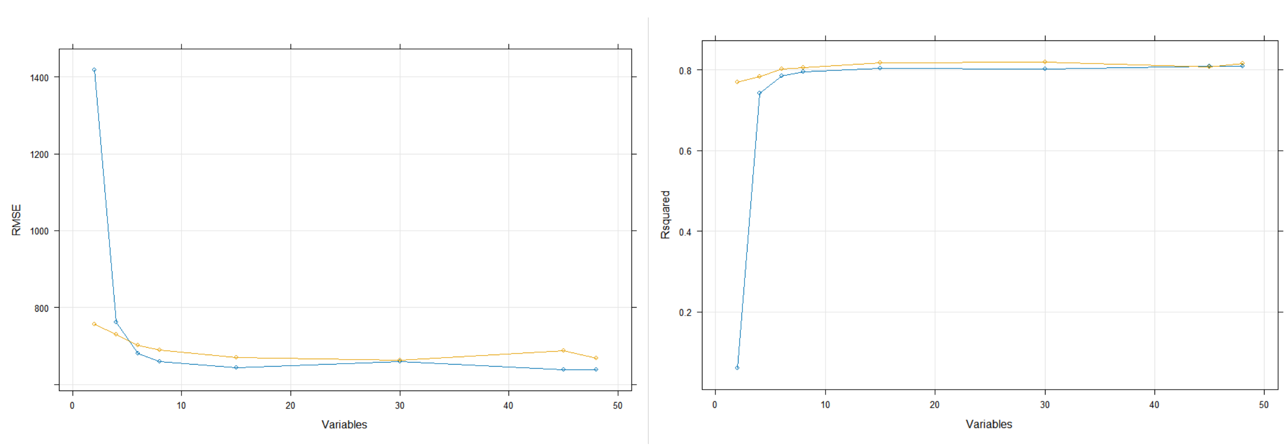

Figure 12: The changes in RMSE and Rsquared during the ranking process

It is apparent that with the increase in the number of variables, the RMSE of all three ranking methods gradually decreases, and Rsquared gradually increases, stabilizing at around 15 variables. The bagged decision tree model shows the best performance. Table 5 presents the values of the three evaluation criteria for the two ranking methods on the test set:

Table 5: Three models results of RFE

Method	RMSE	Rsquared	MAE
linear regression	716.208	0.856	271.19
bag decision tree	703.610	0.872	232.634

The bagged decision tree model yielded better results; hence, we adopted its findings as our research outcome. The selected socioeconomic factors are as followings:

$$HOUSEHOLDSBYTYPE(HCO1\_VC03, HCO1\_VC09, HCO1\_VC10,$$
$$HCO1\_VC11, HCO1\_VC12, HCO1\_VC13, HCO1\_VC14, CO1\_VC15),$$
$$RELATIONSHIP(HCO1\_VC25), GRANDPARENTS(HCO1\_VC62, HCO1\_VC7),$$
$$RESIDENCE1YEARAGO(HCO1\_VC118, HCO1\_VC120), PLACEOFBIRTH(HCO1\_VC131),$$
$$WORLDREGIONOFBIRTHOFFOREIGNBORN(HCO1\_VC161)$$

The specific definitions of the variables are as detailed in Table 6.

Table 6: Specific definitions

Symbol	Definition
HC01_VC03	Total households
HC01_VC09	Male householder, no wife present, family
HC01_VC10	Male householder, no wife present, family , With own children under 18 years
HC01_VC11	Female householder, no husband present, family
HC01_VC12	Female householder, no husband present, family , With own children under 18 years
HC01_VC13	Nonfamily households
HC01_VC14	Householder living alone
HC01_VC15	Householder living alone - 65 years and over
HC01_VC25	Population in households
HC01_VC62	Responsible for grandchildren
HC01_VC70	Who are female-Number of grandparents responsible for own grandchildren under 18 years
HC01_VC118	Same house
HC01_VC120	Different house in the U.S. - Same county
HC01_VC131	Native ,Born in United States ,State of residence
HC01_VC161	Latin America

To further confirm the effectiveness of RFE, we conducted a study using the filter method. Given the superior performance of the bagged decision tree model in RFE, we used it within the SBF algorithm. The selected variables are socioeconomic factors related to HOUSEHOLDS BY TYPE, aligning with our previous conclusion that "filter methods might exclude substantively meaningful variables," further substantiating the accuracy of RFE.

4.1.2 Feature selection conclusions

1)Opioid usage is linked to local residents' household types, with single-parent and elderly (65+) living alone households, especially those where grandparents are responsible for grandchildren, particularly female grandparents, being more prone to opioid use. This may be due to lack of proper management for minors and the elderly, leading to mental health issues and opioid usage.
2)Opioid usage is associated with place of birth and residence. Citizens born in Latin America or the U.S. without significant relocation are likelier to use opioids, possibly due to inadequate local policies and control over opioids, resulting in increased local usage.

4.1.3 Regression conclusions

Subsequently, we conducted a regression using the 15 selected variables with the model from the first part, resulting in a new model as follows:
$y = 1.170 - 4.908 \times 10^{-4}x_0 - 8.806 \times 10^{-3} + 8.735 \times 10^{-3}x_1 - 8.789 \times 10^{-4}x_2 + 4.204 \times 10^{-3}x_3 + 1.915 \times 10^{-3}x_4 - 1.957 \times 10^{-3}x_5 + 3.932 \times 10^{-3}x_6 - 5.179 \times 10^{-5}x_7 + 7.997 \times 10^{-3}x_8 - 2.297 \times 10^{-2}x_9 + 2.716 \times 10^{-5}x_{10} + 7.711 \times 10^{-5}x_{11} + 5.499 \times 10^{-5}x_{12} + 1.172 \times 10^{-3}x_{13}$
And the Rsquared of the model is 0.876.

5 Part 3

5.1 Strategies

In response to the opioid abuse issue, we propose the following strategies,The U.S. government implemented a series of measures in 2018:

- Firstly, cutting off the supply chain of illegal opioids, which involves strengthening border and customs supervision to prevent illicit drugs from entering the domestic market.
- Secondly, enforcing mandatory minimum sentences for the abuse of certain opioids, meaning that courts must impose a certain level of punishment for specific drug crimes.
- Additionally, the government expanded the coverage of drug addiction treatment and rehabilitation programs, providing more recovery services and support to help abusers quit their addiction. In particularly severe drug trafficking cases, the government might even impose the death penalty as the highest punishment for serious crimes.

Regarding issues related to socioeconomic factors, the U.S. government can take the following measures:

- Firstly, considering the significant correlation between our research findings and the underage population, the government should collaborate with local educational institutions to strengthen the education of the harmful effects of opioids on youth. Especially for adolescents from unstable family environments, such as single-parent families or families with absent parents, the government should pay more attention to the mental health of minors in these households to reduce their likelihood of experimenting with opioids.

- Secondly, given the significant correlation found between local residents' drug use and our research, the government should introduce relevant policies to impose criminal penalties on citizens who abuse opioids. At the same time, the government should enhance public education about the dangers of opioids and strengthen the control of these drugs to reduce abuse and dependency.

In summary, the U.S. government needs to adopt a multi-faceted strategy to address the issue of opioid abuse, including efforts in law, education, treatment, and rehabilitation, to effectively reduce the occurrence of this social problem.

5.2 Test the effectiveness

To validate the effectiveness of our proposed strategies, we conducted a study using Kentucky (KY) as an example. Given the issue of multicollinearity among 15 variables, we first performed Principal Component Analysis (PCA) on the data, reducing these 15 variables to 3 principal components. Subsequently, we introduced a suppression factor G for analysis of the data from 2010 to 2016. Considering that the U.S. government might initially invest substantial funds, and anticipating that the investment would gradually decrease over time, we assumed that G follows a uniform distribution with a total sum of 1 from 2010 to 2016. Based on this assumption, we calculated the specific values of G.

Next, we constructed the interaction term GX to represent the impact of the suppression factor on a given independent variable. By analyzing the regression coefficient of GX, we determined whether our measures have an effective suppressive impact on that variable. If the regression coefficient of GX is less than 0, it indicates that our measures have an effective suppressive effect on that variable; conversely, if the coefficient is greater than 0, it suggests that the suppression is not significant.

Through this research process, we conducted statistical analysis on the results. The statistics showed in table 7 that some variables indeed exhibited significant changes after the introduction of the suppression factor, indicating that the government's intervention measures are effective to some extent. This analysis not only provides direct evidence of the effectiveness of our strategies but also offers valuable insights for future policy-making by the government in similar areas.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
-0.698	-1.014	-3.921	-1.521	2.096	-0.818	-2.320	-2.178
x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	
-0.146	-42.52	12.01	-0.107	-0.703	-0.392	-2.493	

Table 7: Coefficient values

6 Sensitive Analysis

We conducted a sensitivity analysis on the regression model for the second question.

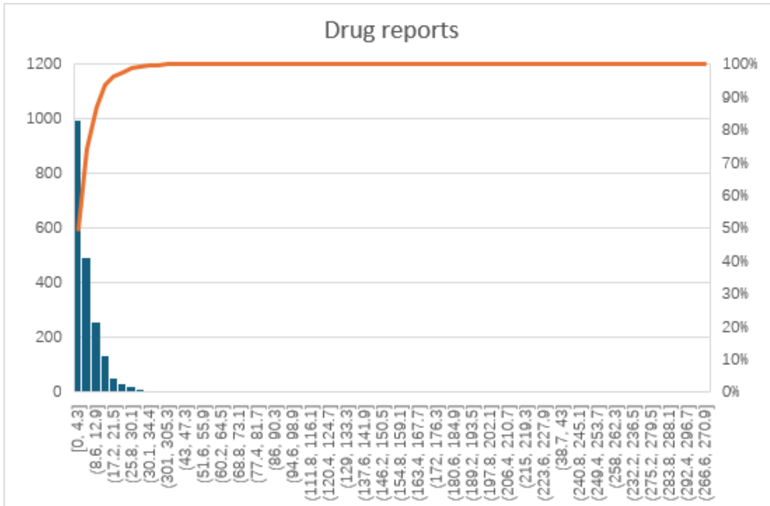

Figure 13: Sensitivity Analysis of the Regression Model

The figure 12 illustrates the distribution of Drug Reports that we predicted in the second question. In this context, when the predicted value is less than 0, we consider it as 0. It shows that 50% of the counties have Drug Reports within the range of 0 to 4.3 cases, while nearly 90% of the counties have Drug Reports within the range of 0 to 21.5 cases.

Table 8: 2015 Frequent Pattern Mining Model Results

Variable	1%	5%
HC01_VC03	0.29782965	-1.4891482
HC01_VC09	-0.0617029	0.30851471
HC01_VC10	0.06322342	-0.3161171
HC01_VC11	0.08138051	-0.4069026
HC01_VC12	-0.0447329	0.22366432
HC01_VC13	-1.0190063	5.09503127
HC01_VC14	0.0477685	-0.2388425
HC01_VC15	-0.1429692	0.71484605
HC01_VC25	-0.9915116	4.95755783
HC01_VC62	0.16959957	-0.8479978
HC01_VC70	-0.0368285	0.18414256
HC01_VC118	-0.0407737	0.2038687
HC01_VC120	-0.005302	0.02651024
HC01_VC131	0.03100266	-0.1550133
HC01_VC161	-0.0004382	0.00219097

Table 8 displays the changes in Drug Reports after we made alterations to the data. Since 50% of the counties have Drug Reports within the range of 0 to 4.3 cases, HC01_VC13 and HC01_VC25 nearly doubled the results, while HC01_VC03 also exhibited relatively high variations. The remaining variables, on the other hand, were less sensitive. This indicates that the regression model we established is insensitive to some variables while being sensitive to others.

7 Strengths and Weaknesses

7.1 Susceptible-Infectious-Susceptible (SIS) Model

The SIS model is built based on a detailed analysis of drug identification count data. The model determines whether government intervention is required by setting a threshold θ. According to the model's predictions, we identified specific counties that required intensified control measures in 2018. Additionally, we used frequent pattern mining to discover potential drug abuse locations, providing valuable insights for prevention.

However, the simplicity of the SIS model may limit a comprehensive understanding of the complex dynamics of opioid abuse. Furthermore, the model's time range is limited and may not capture the latest abuse patterns.

7.2 Recursive Feature Elimination

The RFE algorithm helps select a set of relevant socio-economic factors, reducing dimensionality and computational complexity. We integrated it with the model from Part 1 to delve deeper into the impact of socio-economic factors on drug propagation. This enhances our overall understanding of the problem and improves model interpretability.

However, the RFE algorithm may exclude variables that could have an impact, potentially weakening the model.

7.3 Principal Component Analysis (PCA)

We used PCA to reduce data dimensionality from 15 variables to 3. To test the effectiveness of our strategies, we introduced a suppression factor G and constructed the interaction term GX to represent the suppression factor's impact on specific independent variables. Through analysis, we observed that the suppression factor had an impact on most variables and successfully achieved suppression. These results demonstrate the efficacy of our strategies.

However, PCA may lose some information from the original data, reducing model accuracy.

7.4 Conclusion

In summary, we have established multiple models and algorithms to address the opioid crisis problem. Each method has its advantages and limitations. Using these methods in combination can provide a more comprehensive understanding of the problem and lead to more effective strategies.

8 Memorandum for The Chief Administrator

Memorandum

To: The Chief Administrator
From: 2406434 Group
Date: January 27, 2024
Subject: Multi-Dimensional Strategy to Mitigate the Opioid Epidemic

Background and Current Landscape:

The opioid epidemic represents a complex public health challenge that has been growing exponentially since the late 1990s. The widespread misuse of prescription opioids, particularly OxyContin, has led to a surge in addiction rates, overdose deaths, and associated economic costs. In 2010, Purdue Pharma's reformulation of OxyContin, while well-intentioned to deter misuse, inadvertently redirected abusers towards illicit and more dangerous opioids such as heroin and fentanyl. The epidemic's toll on public health, social services, and economic stability is profound, with widespread implications across the United States.

Analytical Insights:

Our analysis of the National Forensic Laboratory Information System (NFLIS) drug identification counts has unveiled alarming trends in opioid spread across the nation from 2010 to 2017. The Susceptible-Infectious-Susceptible (SIS) Model applied in our study has been pivotal in identifying patterns of opioid abuse spread and predicting future outbreaks. Counties such as CARROLL, JEFFERSON, and SOMERSET have been pinpointed as high-risk areas requiring immediate and targeted intervention to prevent a significant uptick in abuse cases.

Strategic Proposals:

The strategies outlined herein are designed to be comprehensive and actionable:

Law Enforcement and Border Control:

- Amplify efforts to disrupt the illegal opioid supply chain with improved surveillance and interdiction at borders, employing advanced technologies and increasing personnel.
- Advocate for legislative reform to mandate stricter sentencing for opioid-related offenses, setting a legal precedent to deter drug trafficking.
- Enhance the capability of law enforcement agencies through training and resources to adapt to the evolving nature of drug trafficking.

Healthcare and Treatment Expansion:

- Increase investments in addiction treatment facilities, ensuring that they are equipped with the necessary resources to provide comprehensive care.

- Launch a nationwide public health campaign to educate the population about the risks of opioid misuse and the available avenues for help.
- Foster a collaborative approach among healthcare providers, insurers, and policymakers to facilitate access to medication-assisted treatment and recovery services.

Community and Socio-Economic Development:

- Develop and support community-led initiatives focused on addressing the socio-economic drivers of addiction, such as unemployment and lack of access to mental health services.
- Implement preventive education programs targeting young populations, particularly in schools, to raise awareness of the dangers of opioid misuse from an early age.
- Allocate funding for support services for families impacted by opioid abuse, providing mental health support, financial assistance, and social reintegration programs.

Conclusion and Call to Action:

Our collective response to the opioid epidemic must be as multifaceted and dynamic as the problem itself. The strategies proposed require immediate action and a long-term commitment to make a substantial impact. It is critical to mobilize resources and support from federal, state, and local agencies to implement these strategies comprehensively.

Action Required:

- Assess the feasibility and impact of the proposed strategies through an interdisciplinary task force.
- Secure the necessary funding and legislative support to enhance border security and customs enforcement.
- Foster partnerships with educational institutions and community organizations to extend the reach of prevention and treatment programs.
- Establish a monitoring system to track the effectiveness of implemented strategies, ensuring adaptability and responsiveness to changing circumstances.

For further discussion or clarification, please do not hesitate to contact us.
[2406434 Group]

References

[1] Wu, J. M., Zhan, J., Chobe, S. (2018). Mining Association Rules for Low-Frequency Itemsets. *PLoS One*, 13(7), 1-24.

[2] Han, J., Kamber, M. (2006). Data Mining: Concepts and Technique. *Morgan Kaufmann Publishers Inc*. DOI:10.1002/9781118029145.ch1.

[3] Been, F., Bijlsma, L., Benaglia, L., Berset, J.D., Botero-Coy, A.M., Castiglioni, S., et al. (2016). Assessing geographical differences in illicit drug consumption - A comparison of results from epidemiological and wastewater data in Germany and Switzerland. *Drug Alcohol Depend*, 161, 189-199.

[4] Xinhuanet. (2018). Retrieved June 21, 2018, from `http://www.xinhuanet.com/world/2018-06/21/c_129897788.htm`

9 Appendices

9.1 Appendices 1:Frequent Pattern Mining Model Results

Table 9: 2011 Frequent Pattern Mining Model Results

SubstanceName	State	Support	Kluc	IR
SubstanceName	State	Support	Kluc	IR
Propoxyphene	OH	0.1264637	0.454359165	2.079545455
Oxycodone	KY	0.24824356	0.590653791	3.290598291
Morphine	KY	0.163934426	0.427350427	2.333333333
Methadone	KY	0.159250585	0.426056458	2.145299145
Hydromorphone	OH	0.119437939	0.450150086	1.806818182
Hydrocodone	KY	0.271662763	0.643685679	3.35042735
Heroin	OH	0.196721311	0.621602624	3.306818182
Codeine	OH	0.105386417	0.434253247	1.431818182
Buprenorphine	KY	0.154566745	0.407526567	2.247863248
Buprenorphine	OH	0.154566745	0.500475285	2.988636364
Tramadol	VA	0.119437939	0.413998082	1.419047619

Table 10: 2012 Frequent Pattern Mining Model Results

SubstanceName	State	Support	Kluc	IR
Oxymorphone	OH	0.122969838	0.434302193	2.261363636
Oxycodone	KY	0.252900232	0.597436546	3.406779661
Morphine	KY	0.167053364	0.430084746	2.440677966
Morphine	OH	0.167053364	0.534090909	3.272727273
Methadone	VA	0.157772622	0.445090909	2.272727273
Hydrocodone	KY	0.257540603	0.614494827	3.262711864
Heroin	OH	0.194895592	0.610606061	3.579545455
Buprenorphine	OH	0.16937355	0.539346106	3.329545455
Tramadol	VA	0.113689095	0.40155939	1.245454545

Table 11: 2013 Frequent Pattern Mining Model Results

SubstanceName	State	Support	Kluc	IR
SubstanceName	State	Support	Kluc	IR
Oxycodone	VA	0.263982103	0.602424311	3.2578125
Morphine	VA	0.192393736	0.481700212	2.3046875
Methadone	VA	0.178970917	0.47314257	1.9453125
Hydromorphone	VA	0.165548098	0.494618056	1.40625
Hydrocodone	KY	0.246085011	0.60690139	3.435897436
Heroin	VA	0.230425056	0.543827266	2.84375
Buprenorphine	OH	0.170022371	0.554004677	3.534090909
Tramadol	VA	0.165548098	0.521766903	1.2421875

Table 12: 2014 Frequent Pattern Mining Model Results

SubstanceName	State	Support	Kluc	IR
SubstanceName	State	Support	Kluc	IR
Oxycodone	KY	0.245454545	0.592965465	3.381355932
Morphine	KY	0.170454545	0.456685499	2.288135593
Methadone	KY	0.145454545	0.414684199	1.889830508
Hydromorphone	VA	0.136363636	0.443183643	1.319327731
Hydrocodone	KY	0.240909091	0.592009685	3.144067797
Heroin	VA	0.218181818	0.535592749	3.050420168
Fentanyl	OH	0.154545455	0.551942038	2.425287356
Buprenorphine	KY	0.204545455	0.519392742	2.762711864
Tramadol	VA	0.143181818	0.434976153	1.554621849

Table 13: 2015 Frequent Pattern Mining Model Results

SubstanceName	State	Support	Kluc	IR
SubstanceName	State	Support	Kluc	IR
Oxycodone	KY	0.242924528	0.571587125	3.117647059
Morphine	OH	0.160377358	0.526859504	2.75
Methadone	KY	0.136792453	0.390905601	1.655462185
Hydromorphone	VA	0.117924528	0.409701198	1.327102804
Hydrocodone	KY	0.238207547	0.5733373	2.848739496
Heroin	OH	0.205188679	0.618603896	3.977272727
Fentanyl	OH	0.181603774	0.598587866	2.715909091
Codeine	OH	0.120283019	0.504058442	1.352272727
Buprenorphine	KY	0.205188679	0.491999707	2.890756303

Table 14: 2016 Frequent Pattern Mining Model Results

SubstanceName	State	Support	Kluc	IR
Oxycodone	VA	0.226544622	0.553275309	3.13559322
Morphine	OH	0.153318078	0.5185419	2.761363636
Morphine	VA	0.153318078	0.421758387	2.059322034
Methadone	OH	0.123569794	0.452764128	2.102272727
Methadone	VA	0.123569794	0.374759505	1.56779661
Hydromorphone	VA	0.128146453	0.451028594	1.110169492
Hydrocodone	KY	0.242562929	0.600581114	2.966101695
Heroin	VA	0.244851259	0.594179304	3.220338983
Tramadol	OH	0.125858124	0.488782051	1.772727273
Acetyl fentanyl	OH	0.110849057	0.493006993	1.181818182
Tramadol	OH	0.16509434	0.570994599	2.29545454

Table 15: 2017 Frequent Pattern Mining Model Results

SubstanceName	State	Support	Kluc	IR
Oxycodone	VA	0.230769231	0.536693548	2.903225806
Morphine	OH	0.153846154	0.532142857	2.386363636
Methadone	VA	0.102564103	0.327079219	1.185483871
Hydrocodone	KY	0.226107226	0.581353757	2.919642857
Heroin	VA	0.249417249	0.573361427	3.040322581
Tramadol	OH	0.153846154	0.555327869	2.079545455
Fentanyl	VA	0.212121212	0.503163029	2.693548387
Cyclopropyl fentanyl	OH	0.107226107	0.522727273	1
Codeine	OH	0.125874126	0.509825701	1.511363636
Fentanyl	OH	0.178489703	0.575385208	3.352272727
Codeine	OH	0.105263158	0.43965821	1.465909091
Carfentanil	OH	0.121281465	0.664150062	0.829545455
Buprenorphine	KY	0.219679634	0.543143297	2.983050847
Tramadol	OH	0.153318078	0.547348485	2.284090909