For office use only
T1 _____
T2 _____
T3 _____
T4 _____

Team Control Number
1922676

Problem Chosen
C

For office use only
F1 _____
F2 _____
F3 _____
F4 _____

2019

MCM/ICM
Summary Sheet
(Your team's summary should be included as the first page of your electronic submission.)
Type a summary of your results on this page. Do not include the name of your school, advisor, or team members on this page.

Opioid drug use in the United States is a growing epidemic that has reached the level of a national crisis. In our paper we discuss the opioid epidemic, it's causes, how it will behave in the future, and what aspects we need to analyze before we can make meaningful change. We answered these questions using our various models which examined five US states: Ohio, Kentucky, Tennessee, Virginia, and West Virginia.

These models deduced the socioeconomic factors associated with the people and places affected by opioid usage. It also viewed how substance abuse affects the usage rates of related opioids and the potential implications of the situation if left unchecked. Our analysis of the situation was executed in three different models where we critically analyzed the data supplied by the US census Bureau, and NFLIS.

The first model we developed to predict how the various opioids influenced the growth and decay of the other prevalent drugs. The model extrapolated by taking the pairwise correlation of every drug from our data. From that we were able to discern various groups of drugs based on these correlations. Within these groups the correlation values were upwards of 0.95 for almost every pairing. For other drugs our model gave us negative correlations which tended to be drugs from separate groups. These insights provided us to further analysis the data even more precisely and evaluate each individual state.

A portion of our first model was based on regional differences. The main difference we noticed was in the opioid data for Kentucky where we found a strong variation between counties. Through further analysis, we were able to identify that Kentucky had a great deal of dry counties. When accounting for geographic proximity, we concluded that dry counties and their neighbours had worse drug rates when compared to wet counties only surrounded by other wet counties. By modifying the definition of a "dry county" in this way we found there to be a difference that was about 10% between the dry counties and wet counties. This is somewhat small but it's paired with far more extremes in the dry counties especially in 2010.

Our model that compared dry versus wet counties also used socioeconomic models as part of its analysis. Our model for the socioeconomic demographic used nearly a hundred different parameters. Each of these parameters described a different factor in the lives for the people across all counties in the states we analyzed. Our model then used linear regression to find which factors were correlated with high opioid use across all counties. From this method we found a few demographic groups that were the most highly correlated with high drug abuse. These groups involved individuals who were married in a relationship where they were either divorced or widowed, grandparents caring for grandchildren on a long-term basis, and disabled people living independently. All these groups are generally considered to be underprivileged.

Contents

1 Introduction

Millions of Americans are suffering from pain and as a result for the treatment of the management of pain(prescription use) or for recreational purposes (illicit use) have turned to using a variety of opioids. However, with the use of opioids there has been an evident increase in misuse, opioid use disorder, and overdose resulting in there being declared a "national opioid crisis". As a result of the high levels and addictive usage of opioids, government organizations like the Centers for Disease Control and Drug Enforcement Agency, who are struggling to save lives of opioid users and prevent the harmful effects from substance abuse.

Due to the severity of the epidemic, the ordinary laws and enforcement techniques have been unable to control the misuse of opioid abuse. In this case, we will investigate methods that the American government should implement in order to deal and control the crisis. Our investigation will begin with analyzing the situation and data and, from there, answering the following questions, which could prove to be a resolution for/to the opioid crisis.

First, How does drug use grow over time? What counties currently have the highest drug rates, and how will that change into the future?

How does the use of some opioid drugs affect others? What is the effect of dry counties on illicit drug use? What socioeconomic demographics are most correlated with drug abuse?

The importance of these questions and the effect of their answers vary wildly. However, what exactly these questions mean needs and explanation.

1.1 Significance: Why What We're Asking is Important and Relevant

The effect of drug use over time is an obvious first question to figure out the impact of drug use on a population. A population that is more exposed to illicit drugs is a less productive population; however, once illicit drugs are introduced into a population their effects can be sporadic, and hard to remove from the population. As a result, time may be the most important factor when considering the spread of drugs through a population. This is also relatively easy to measure due to the volume and granularity of the data widely available. The most important aspect of this data is the widely available information and statistics that have been stored over many years. This lends well to the common types of modelling and gives a basis for further questions.

One of the most important questions from there is to ask where precisely the drugs have been actively used in the past years, and where they can potentially spread. This factor has vast implications in prediction and mitigation. By nature, the effect of drugs are a community problem, so analyzing the large scale can only get us so far. If we know where the drugs are and where the drugs will spread, education and law enforcement methods can be directed to precise counties to support the communities that need help the most. This makes it so that the efforts of drug mitigation are targeted not just where the drugs currently are, but so other mitigation efforts can make a difference to local communities even before a drug crisis occurs. By analyzing the state of drugs in a county, and the movement of drugs between counties we gain the ability to help everyone both locally in counties, and across states.

Another insight we get from looking at the state of drug abuse in different counties is the common factors between the counties. One of the common factors that affects the level of drug use is the legality of other drugs in the five states that were analyzed. The two major drugs that have effects this large are marijuana and alcohol. In Ohio, Kentucky, West Virginia, Virginia, and Tennessee marijuana laws are roughly the same, so that is not a large factor to consider. The large factor we do need to consider is alcohol. Across the United States, there are still many dry counties i.e. counties where alcohol is prohibited. Many of these

counties exist in two of the states we analyze namely Kentucky, and Tennessee. In Kentucky, there's a large amount of variety in both dry counties, wet counties, and "moist" counties where there are only some restrictions on alcohol. This contrasts with Tennessee where the majority of counties are moist, and alcohol laws are generally consistent throughout the state, which makes it harder to make distinctions between dry and wet counties under similar circumstances. Due to this factor, Kentucky was the primary subject of our analysis. It's obvious that legality of drug use affects all the use of all drugs, which provides a singular, yet significant and valid reason for variation between counties, but it doesn't account for everything.

In an attempt to account for as much of the variation as possible, concerning the spread and presence of drug use- we must also look at socioeconomic status. This status gives insight into all sorts of factors that are commonly considered when looking at drug use. Some of the factors include country of origin, family status, number of occupants in the household, age of the occupants, marital status, etc. How all these factors affect drug use, in general, is a hard question to answer. However, with large amounts of data and hundreds of different variables, much of the variation can be taken into account.

With all those indicators, predictors states and counties we will be able to more accurately look at what can be changed within their local jurisdictions and make meaningful impact. We believe that possible changes made, can result in the beginning of the end of the opioid crisis in the five states.

1.2 Model Assumptions and Constraints

1.2.1 Assumptions

- *No new opioids will be developed in the future.* If they did show up, we assume that their effect on the overall illicit drug scene would remain constant. The data to support this is the fact that a small number of drugs account for the vast majority of the supply on the drug market.
- *The number of drug reports is proportional to the real number of users of that drug.*
- *Correlation between two drugs is constant over time.* The relationship between the drugs is then used as a parameter rather than a variable.
- *Geographic proximity between counties and states isn't a large factor when deciding their effect on each other.* This avoids the questions of what borders count and others don't as well as the question of population distributions within the counties themselves.
- *Non-opioid drug use has little effect on opioid drug use.* Whether a county is dry or not is the only exception to this.

1.2.2 Constraints applied by assumptions

- The data in use must be entirely accurate, complete, and representative to have any form of working model
- The counties must be in roughly the same geographic region

2 Methodology

2.1 Approach

We needed a method of extrapolating the data provided, to predict drug usage, analyze socio-economic trends, and model possible solutions. We considered the following three different strategies of mapping data

onto a function.

2.1.1 Polynomial Fitting

Originally, we considered Polynomial Fitting, which is a conventional technique in this type of situation. The merits of this strategy were obtaining a nice curve that can be extrapolated indefinitely, are extremely precise on the given data points, and is relatively cheap in a computational sense for small amounts of data ($O(n^2)$ efficiency). However, the data was fragmented enough so that the curve was much too steep for precise extrapolation, past the immediate vicinity of the provided points.

2.1.2 Bicubic Splines

This technique is a modified version of polynomial fitting that considers cubic splines on small intervals along the data set. The algorithm is quite fast and almost never gives steep slopes on a uniform data set, but there is no meaningful way of extrapolating the returned function.

2.1.3 Recursive Neural Networks

The concept of using RNNs to find non-linear relationships in a set of data has been steadily gaining popularity over the last few years, and we tried to apply an existing general-purpose implementation to this problem. However, in using this strategy, we quickly ran into the impassable roadblock of the ratio of dependent-to-independent variables being too high. That is, we only have data for seven points in time and hundreds of values that we are trying to express as a function of time. For that reason, there was no way to train the network to identify any useful patterns in the given data.

2.1.4 Modified Linear Regression

Linear regression is a method for finding a linear function that approximates the given data set. It can depend on one or more factors and generally has the following form:

$$f(x) = b_0 + \sum_{i=0}^{n} b_i x + \epsilon$$

As now we had the opportunity to include as many weighted parameters as we want, we decided it would be prudent for the other parameters to be a function of the rate of change of the other drugs. So, our modified linear regression function took the form of:

$$f(t+1) = f(t) + f'(t) + c\sum_{D} A_D D'(t)$$

where t is the time in years, f' and D' are rates of change of the drugs (i.e.) slope of line of best fit for the data points for the particular drug from years 0 to $t-1$, A_D is the impact of use of drug D to drug A, and c is the coefficient deciding how impactful the rate of change of the other drugs is on any given drug. This c can be reverse-engineered from existing data of the previous year.

2.2 Drug Use Prediction Model

This model was made with the purpose of predicting the future drug usage across all the counties in the given states. This was done by the modified linear regression technique seen above. The model would first measure the correlations in between the uses of different drugs in all the counties. We employed the following formula:

$$cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(\sum_{i=1}^{n}(x_i - \bar{x}))^2(\sum_{i=1}^{n}(y_i - \bar{y}))^2}$$

Where X and Y is the amount of recorded uses for drugs A and B. The above correlation would be measured separately for every county. The correlation of drug A to drug B (A_B) would be always be in the

interval $[-1, 1]$ and signify how related the usage of the two drugs is (i.e. $A_B = 1$ would mean the usage of drug A is historically directly proportional to that of drug B, and $A_B = -1$ would signify an inverse relationship. Of course, $A_B = B_A$). So, for any arbitrary drugs A and B:

$$A_B = cor(\{A_i, B_i \,|\, i \in \mathbb{N}, t_0 \leq i \leq t\})$$

Where t_0 is the earliest year for which we have data ($t_0 = 2010$) and t is the latest year we are considering (usually $t = 2017$). Then, the model would calculate the slope of the line of best fit of drug use over time for every drug for every county. The following formula was used:

$$slp(X, Y) = \frac{\sum_{i=1}^{\dim X}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{\dim X}(x_i - \bar{x})^2}$$

Where X is the time (in years) and Y the amount of recorded cases for the required drug. The rate of change of drug A at year t ($a'(t)$) is calculated from t_0 to the current year t:

$$A'_t = slp(\{i, A_i \,|\, i \in \mathbb{N}, t_0 \leq i \leq t\})$$

As mentioned in the section on Modified Linear Regression, we took advantage of the opportunity to let A_t depend on multiple variables, those variables being the rate of change of other drugs. So, combining the above formulas we have:

$$A_{t+1} = A_t + A'_t + c_t \sum_D (cor(A, D) \cdot D'_t)$$

The coefficient c_t is necessary to account for the different possible number of drugs in the system. We can reverse-engineer it from the previous year's data:

$$c_t = \frac{A_t - A_{t-1} - A'_{t-1}}{\sum_D (cor(A, D) \cdot D'_{t-1})}$$

So, expanding the above formula for A_{t+1}, we have:

$$A_{t+1} = A_t + \frac{\sum_{i=t_0}^{t}(A_i - avg(A))(i - \frac{t+t_0}{2})}{\sum_{i=t_0}^{t}(A_i - avg(A))^2} +$$

$$+c_t \sum_D \left(\frac{\sum_{i=t_0}^{t}(A_i - avg(A))(D_i - avg(D))}{\sqrt{\sum_{i=t_0}^{t}(A_i - avg(A))^2}\sqrt{\sum_{i=t_0}^{t}(D_i - avg(D))^2}} \cdot \frac{\sum_{i=t_0}^{t}(D_i - avg(D))(i - \frac{t+t_0}{2})}{\sum_{i=t_0}^{t}(D_i - avg(D))^2}\right)$$

Where $avg(X)$ is the function that takes the average drug usage from t_0 to t, i.e.:

$$avg(X) = \frac{\sum_{i=t_0}^{t} x_i}{t - t_0}$$

However, in its base form the model failed capture was prone to drawing false correlations and calculating meaningless derivatives when it came to drugs with few or only occasional users. A correction on how the model perceives drug usage and popularity was necessary; so we made the following correction:

$$A_t = A(t) \sum_D \frac{1}{D(t)}$$

Which defines A_t as the ratio of the users of drug A to total drug users (market share of year t), $A(t)$ being the users of drug A in year t, and D being all the drug types used in the year t. All of the above formulas still hold – but output the result in terms of the previous year's market share instead of total users. So, the final formula to get the predicted amount of users of drug A in year t is:

$$A(t) = (\sum_D D(t-1))(A_{t-1} + A'_{t-1} + c \sum_D (cor(A, D) \cdot D'_{t-1}))$$

2.3 Socio-economic Trends Model

Our second model was made not to predict future drug usage, but discover trends and patterns in the existing substance and socio-economic data. Here we used similar techniques as for the previous model, most notably the correlation function. In this case we considered the correlation between the average number of drug incidents per household in each county and the given socio-economic factors. This was done with the following formula:

$$C_F = cor(\{D_c, F_c \,|\, \forall c \in \mathbb{U}\})$$

Where C_F is the correlation between a certain socio-economical factor F and average drug usage per household, D_c is the average drug use per household in county c, F_c is the prevalence of factor F in county c, and $\mathbb{U}$ is the set of all counties. As the drug use we are measuring is not total but scaled to the population, we must also use factors that have been similarly scaled. Therefore, we only considered factors that are measured in percent (i.e. keys of type HC03_XXXX).

3 Results and Analysis

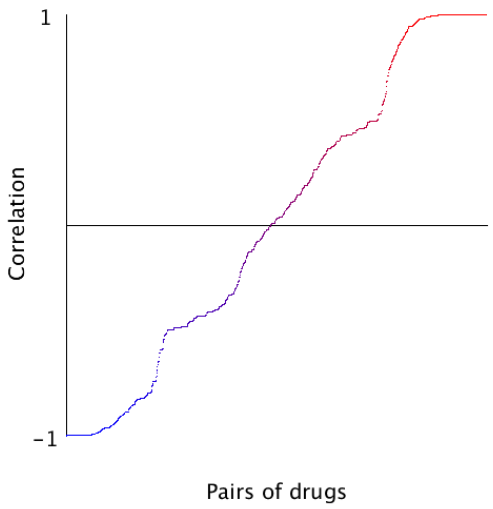
3.1 Drug Prediction Model

The results gathered from the drug prediction model mainly illustrate the growth of several, heavily-mutually correlated clusters of drugs and the decline of most other ones. The most notable of these are the groups for fentanyl-related drugs(Fentanyl, Acetyl fentanyl, Furanyl fentanyl, p-Fluorobutyryl fentanyl, 3-Methylfentanyl, Butyryl fentanyl, ANPP) and the oxycodone-related drugs (Oxycodone, Methadone, Hydrocodone, Codeine). Heroin, on the other hand, (as well as the one drug related to it, Buprenorphine), is predicted to be in steady decline over the next few years.

The first two groups of are extremely highly mutually correlated, with correlations in the range of 0.989 to 0.996 for the fentanyls and in the range of 0.996 to 0.999 for the oxycodone. Heroin, however, has the highest correlation of 0.57 which is significantly less useful. It appears our model predicts that clusters of similar drugs will dominate the market share of unique drugs, and heroin, which in 2017 has a stunning market share of 38% is predicted to lose it's crown to Fentanyl in 2020.

The negative correlations are just as intriguing. The lowest correlations are between members of groups of drugs (such as Furanyl fentanyl and Buprenorphine at -0.995 or Furanyl fentanyl and Oxycodone at -0.983), but even more so between members of prominent clusters of drugs and morphine (with Furanyl fentanyl and Morphine having a correlation of -0.9997). This, however, seems to have little effect on the future – all tightly-correlated groups of drugs seem to be destined to prosper at the cost of more unique drugs, and which one will ultimately emerge victorious is impossible to predict accurately.

Figure 3.1: Graph of all drug correlations



3.2 Case Study: Dry Counties in Kentucky

Modelling future drug growth in individual states and counties had similarly curious results, with the average drug use (and future growth) of a certain area being strongly tied to that area's alcohol laws. It is fairly obvious that the legality of alcohol would affect how drugs overall are used within counties – what is surprising is the effect that it had. The exploration of this data will be broken down into a few parts. Each of these parts will give a slightly different definition of what it means to be a dry county vs a wet county.

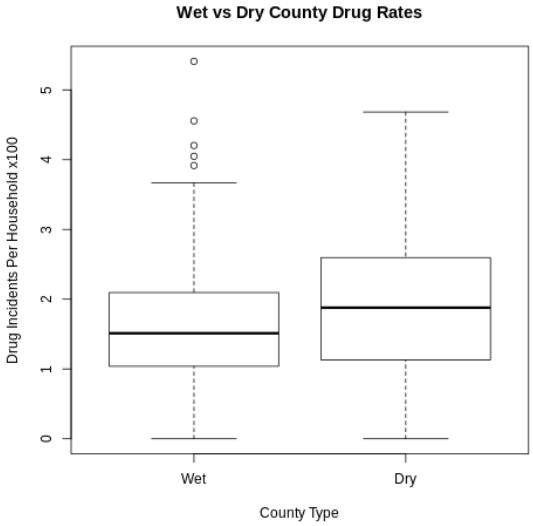
The first most obvious way we define a dry county is to look at the counties that prohibit all alcohol sales with no exceptions. Looking at this data, we find two important results and a few caveats. The first result comes from when we solely look at the year 2010 when the laws of Kentucky were the most strict. The results of this are shown in figure 1 with a box plot. The wet counties have more extremely high outliers, however, the median and upper quartiles of drug use is worse in the dry counties. This is a result of the simple flaw of counting very large counties with high populations equally with very sparsely populated counties. It is especially strong since we only aggregate data from a single year, and don't account for that in any other way.

The second result uses the same definition of a dry county as above, but we change how the data is aggregated, by instead averaging all the years we have data for. This yields the null result where the dry and wet counties are nearly indistinguishable, but this has a very strong reason and thus, is not contradictory to other results. Over the course of time, when the data was collected for the alcohol laws in Kentucky changed drastically. In 2010 the alcohol laws were very strict, but as time went on the laws got more relaxed making the distinction between dry and wet counties much less considerable. Since the laws changed in many of the counties, we analyzed it was hard to draw any conclusion from them. This also lies in the fact that our initial definition of dry was so strict. In order to have meaningful results we would need to change our definition of what it means to be dry every year that the county laws change. Because of this a more robust and looser definition is required.

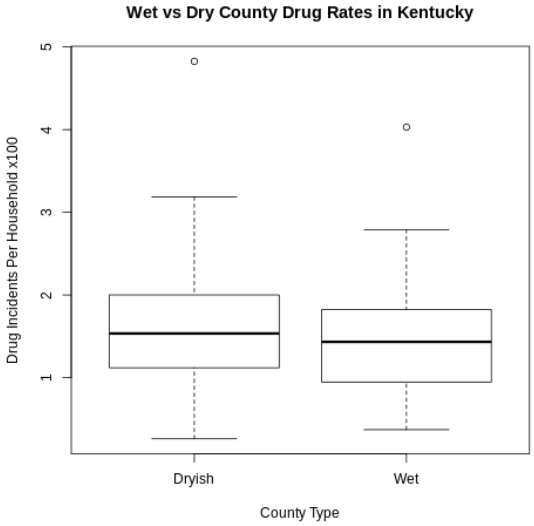
The last result we have for this comes from when we loosen our definition of a dry county. This part of the analysis needs to be detailed to have meaningful results, so we must account for geographic proximity, in just this case. We now redefine a dry county to be any county that is either a dry county or a wet (or partially wet) county bordered by at least two other dry counties. This was done to support or disregard our hypothesis that being close to dry counties made drug rates higher than they would be otherwise. It was a consideration to take into account other moist counties, however going through the county laws for

all Kentucky counties is a time-consuming task that would require a lot of subjectivity. When comparing those counties, the summary results are that for dry counties the 1st quartile, median, and 3rd quartile are 1.148%, 1.5345%, and 2.0009% of drug incidents per household. For the wet counties, the same results are 0.9526%, 1.4727%, and 1.9945%. This is a small difference, but not a sweeping one.

After this analysis, another one of our findings was that there were many extreme outliers, whereas there were less for the dry counties. We found many possible reasons for most of the outliers. We decided that Bell, Harlan, and Rowan counties were closer with there laws to the dry counties than the wet. Since this is ambiguous by the definition of a dry county, we then left these counties out of the tally and recalculated. After the recalculation, the results for the dry counties didn't change, but the wet counties summary data lowered to 0.9472%, 1.4333%, and 1.8224% for the first quartile, median, and third quartile (the data is shown in figure 2). We believe that if we took into account more counties that were ambiguous in definition, these results would be even stronger.

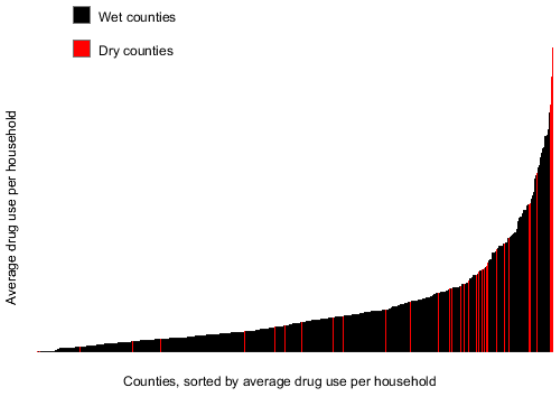


(a) Wet vs dry counties of Kentucky in the year 2010



(b) Wet vs Dry counties over time when using a modified definition of dry county

Figure 3.3: Graph drug use per household per county in 2010



3.3 Socioeconomic Model

After taking into account how the drugs affect each other, we must also take into account the situation of the people. This was done with the socioeconomic model. With the model described above, we have 88 different parameters with which we have data and tells us how much it correlates with drug use.

We found relatively few negative correlations. The ones that we did find were small but certainly non-zero. The largest group of demographics that were negatively correlated with drug use were women who were recently pregnant. There were several sub-categories, but the most correlated gave us a correlation of -0.174 which accounts for roughly 3% of the variance. The other interesting negative correlation is that veteran status is slightly negatively correlated, but that only yielded a very small correlation and an even smaller amount of the variance. Overall the negative correlations are interesting, but the positive correlations are much stronger predictors.

The positive correlations with strong correlations can be grouped into several categories, each of which have relatively strong correlations. Those categories are: divorced or widowed households, children sharing household with relatives not in immediate family, and households with a disabled person. The above categories account for all of the correlations ¿0.2 except for two. The only two other predictors are people with associate's degrees and total American born population in the county. Those may have an explanation and more predictable effect, but the more interesting results come from the groups that are closely linked together.

3.4 Grouping by Socioeconomic Parameters

The groups of parameters that correlate the most highly with high drug use are worth looking into as it's likely that the cause of the correlation is the same for all the parameters within a group. The first group to consider is the divorced and widowed group. This group has two parameters: divorced males, and widowed males with correlations of 0.221, and 0.214 respectively. This was surprising to us as none the equivalent female correlations were nearly as high.

The next parameter group was children living with non-immediate family. The parameters for this with correlation ¿0.2 are grandparents taking care of grandchildren without parents present both long and short term, as well as grandparents living with children with the parents present as well. These have correlations of 0.242, 0.219, and 0.204. Other factors in this group appear further down just below the 0.2 correlation threshold as well. There is outside evidence to support that this is because when parents use drugs the grandparents have to look after the children both on a short scale, and sometimes on a longer scale.

The last parameter group we can look at is households with a disabled person. These parameters specifically deal with disabled people who are living independently without the care of the state. The three

parameters available to us here are the different age groups: 64 and under, 65 and over, and total population. These have correlations of 0.224, 0.200, 0.215 respectively. These factors themselves may seem uninteresting because they're inherently tied together; however this supports accuracy of the data.

Socioeconomic parameter groups like this are very useful for finding broad categories that correspond to drug use. Some give insight into what the root cause is, which is not unique to measuring parameters about the demographics of the population. This sort of analysis can also be applied to the characteristics about the states and counties themselves, and what sort of laws they have in place within them. Analysis of this nature gives even more direction into what meaningful change can be made.

4 Conclusion

4.1 Summary of Results

Opioid use in the United States has been a burden on every aspect of American life. A variety of solutions to the problem have been proposed with varying levels of success; however, none of them have been able to eliminate the issue.

Over the course of our paper, we introduced a model primarily using the tools of linear regression and correlation to outline, deduce, and illustrate what more of the causes of the opioid crisis are. The various models we used approached the issue in several different ways. These ways sought out to answer several different questions that we started with. The questions varied in nature, but they all had relatively common sense answers, that is "All drugs are connected". If we can mitigate how the public uses a particular drug, it will have wide-scale impacts on all similar drugs, as there are only a few major groups within opioids. This leads us to believe that focused efforts on drug use are better suited to the problem than broad solutions.

The next conclusion we came to was that the prohibition of alcohol in particular counties leads to higher drug use. The increase in drug use wasn't just in its own counties, but also in surrounding counties. This conclusion came about under analysis of the last model where we looked at the trends of the different drugs. This also gives us a simple more focused way to look at the problem. Instead of large scale nationwide change, meaningful change can be made simply by addressing this issue in county government chambers. We've already seen the start of this impact across Kentucky where prohibition laws have slowly been repealed over the course of the last decade.

The last question we answered was about the individual people. We explained what demographic parameters were closely related to opioid drug use. The highest demographics we found were based on disability, instability, unhappiness in marriages and children not being cared for by their parents. Some of these demographics we can mitigate the growth of, but in situations like widowed partners, there's no easy solution.

If we want large scale change as a society, this is the change that must happen. The policy and social changes that are needed to mitigate opioid use across the nation aren't obvious. However, with continued modelling research, and action based on that research, we can make meaningful results and make strides to reduce opioid use across the nation.

4.2 Strengths and Weaknesses

4.2.1 Strengths

- The algorithms we employed to create our models can we applied not only to this specific problem, but to any problem that includes extrapolating a set of data by the independent variable where the dependent variables are dependent not only on the independent variable, but also on each other,
- The multilinear approach to extrapolation gives rather accurate results while running somewhat quickly,

- The socio-economic trends found by our model hold for multiple different methods of calculating them.

4.2.2 Weaknesses

- Our algorithm is prone to finding false correlations (P-Hacking), between socio-economic factors and drug use due to the large volume of possible correlations.
- We assume that the correlation betweens drugs remains constant forever. Although it holds on the given dataset, it is an approximation.

4.3 Future Improvements

Our models were built on non-trivial assumptions that could be improved upon:

- An upgraded model could account for the emergence of new drugs on the market and find warning signs of what type of drug will emerge and where,
- Socio-economic factors could be considered in predicting the future values of drug use,
- With access to more data, a more accurate and complete model can be built that considers the impact of non-opioid drugs on the system.

5 Memo to the DEA

Dear Chief Administrator,

The United States is facing a national Opioid Crisis which threatens to continue spreading and result in even more casualties. Currently, the epidemic threatens to further devastate the lives of American citizens and various industries as opioids and other drugs reach largely untouched portions of the population (post-secondary school graduates). Preventive measures and action must be taken, in order to reduce and eventually stop the misuse of opioids. We investigated and analyzed the data of the five U.S. states, Kentucky, Ohio, West Virginia, Virginia, and Tennessee which was provided by the U.S. Census and DEA/NFLIS for socioeconomic demographics and drug incidence reports in these states.

From the socioeconomic data, we were able to identify a list of social factors that wer highly correlated with drug users. There were three primary socioeconomic characteristics that were highly linked with drug usage rates:

Factors:

- The most popular factor that was linked to drug usage were individuals involved in a relationship where there is a grandparent who is the legal guardian of a child. This factor was the best predictor of drug use, sharing a correlation of 0.2416 with opioid drug cases,
- The second socio-economic factor that was found to be linked with the second-highest number of drug users was disabled individuals who have not been institutionalized or otherwise cared for by the state between the ages of 16 and 64. The correlation was approximately 0.2244 with opioid drug cases,
- Third, we found that the amount of men who were over the age of 18 and had been divorced or widowed correlate with opioid drug cases by 0.2212.

Linkage-Between drugs:

- Another insight we found and utilized was through correlation. We used correlation to help identify and model two distinct types of relationships: drugs that are highly positively correlated and drugs that are highly negatively correlated.
- Our results very evidently show that similar drugs are highly correlated in cases of reported drug use, with oxycodone, methadone, and hydrocodone showing a highly positive correlation of 0.997, while for drugs that were different in effect we found they would be highly negatively correlated, such as fentanyl and morphine which had a correlation of -0.997
- The correlation insights provide valuable information into the forecasting of the growth or decline of different drug groups. It also provides the ability to predict the usage and "popularity" of other major drug groups.

Dry and Wet:

- The last result we found was the difference between dry and wet counties where dry counties had higher opioid cases at 1.5345% of households while wet counties had a rate of 1.4333% of households. The shocking part of this isn't necessarily the percentage, but the number of extreme outliers in dry counties compared to wet counties.
- Additionally, from our data, we can find that dry counties that neighbour a wet county, increase the usage within their neighbouring wet counties. An example of this insight is Powell county (wet) which is bordered by three different dry counties in Kentucky, and subsequently has a drug rate to match.
- We hypothesize that in this case, the three neighbouring dry counties affect Powell's drug usage, as they naturally border each other. This insight can be used to analyze what wet counties have done legally, economically, and socially to have significantly lower drug usage rates.

From our results and insights, we have identified three major socioeconomic factors that are correlated with high rates of opioid usage. We also highlighted the correlation between dry counties and a higher opioid drug usage rate. Lastly we have discovered that through the correlation of different drugs and their grouping we can forecast the usage of a variety of drugs. We can use our data and insights to develop potential strategies to reduce and combat the ongoing national opioid crisis. For example, we can investigate the reasons of what makes a wet county have a lower drug usage rate. While from our socio-economic data we could potentially develop accurate prescription monitoring programs for each county/state or invest in help strategies for opioid users facing addiction in the form of health service. That is just a few of the many ways the opioid epidemic can be averted. With continued modelling research efforts by the US government the drug crisis will eventually be mitigated, and we'll all see a happier and healthier America.