MathWorks Math Modeling Challenge

Aneek Barua, Inko Bovenzi, Danyil Blyschak, Chris Chan, Jeremy Kim

1 Summary

We built regression models to predict nicotine use from both cigarette and e-cigarette use in the future. We discovered that cigarette use would decrease linearly, while e-cigarette use would increase exponentially. This led to the conclusion that overwhelmingly more people were being exposed to nicotine from e-cigarettes. To build these models, we used extensive behavioral and microeconomic analysis. Part of this economic analysis included building auxiliary models, such as constructing a demand curve and predicting the future price of e-cigarettes, all in an effort to tune the model to the most important economic influences, making it more accurate.

We also built a model to predict substance use using a logarithmic regression curve. We keep a weighted count of how many characteristics and features a student has (such as education level, time of first exposure to drugs, etc.) and then pair this weighted count with a confidence score. The confidence score is generated by multiplying the percent of teens who have the characteristic or features (such as 12.1% of teens discharged from a hospital reported drug use in the past month) by the weighting score of each feature. Our regression model is able to map this relationship and now all people have to do is complete a binary checklist of what characteristics and features they have and then the model can create a confidence score and use a threshold of 60% confidence to say that a student has used drugs. We fitted our logarithmic regression equation using general drug use data but to see predictions on more specific drugs, we can just refit our model with the necessary datasets.

Finally, we also determined that opioids were the most impactful drug on a per-user basis. The damage that opioid usage had both financially, and on the quality of life were devastating when compared to similar statistics for the usage of nicotine, alcohol, and marijuana. Alcohol was the least impactful drug, which is likely attributable to the fact that unlike the other drugs, Alcohol is typically consumed in a controlled environment like at home or in a restaurant or bar, and people are far more sensible about safe alcohol consumption.

2 The Problem

Substances such as tobacco alcohol, and narcotics can affect the physical and mental health of users. The consequences of substance abuse, both financial (health care, the criminal justice system, workplace productivity, etc) and non-financial (divorce, domestic abuse, etc), ripple through society and affect more than just the user. The effects of substance abuse on individuals and society have come to the forefront recently as opioid addiction has become prominent.

Efforts, such as taxes and regulations on cigarettes and the Drug Abuse Resistance Education program, have been made at the local, state, and national level to educate, control, and/or restrict the consumption of such substances. Such efforts need to start with an understanding of how substance abuse spreads and affects some individuals more than others.

2.1 Darth Vapor

Often containing high doses of nicotine, vaping (inhalation of an aerosol created by vaporizing a liquid) is hooking a new generation that might otherwise have chosen not to use tobacco products. Build a mathematical model that predicts the spread of nicotine use due to vaping over the next 10 years. Analyze how the growth of this new form of nicotine use compares to that of cigarettes.

2.2 Above or Under the Influence?

Like nicotine, the abuse of most substances is correlated with numerous internal and external factors that affect the likelihood of an individual becoming addicted. Create a model that simulates the likelihood that a given individual will use a given substance. Take into account social influence and characteristic traits

(e.g., social circles, genetics, health issues, income level, and/or any other relevant factors) as well as characteristics of the drug itself. Demonstrate how your model works by predicting how many students among a class of 300 high school seniors with varying characteristics will use the following substances: nicotine, marijuana, alcohol, and unprescribed opioids.

2.3 Ripples

Develop a robust metric for the impact of substance use. Take into account both financial aid and non-financial factors, and use your metric to rank the substances mentioned in question 2.

3 Our Proposed Solutions

3.1 Darth Vapor

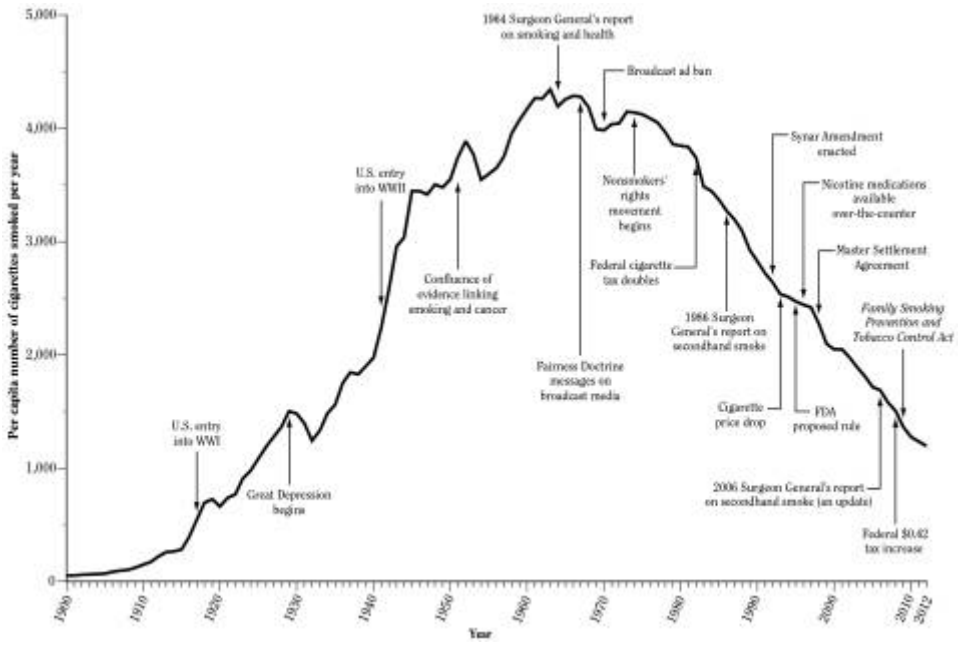
3.1.1 The Models

Since this question seems to be handling both E-Cigarette users and nicotine users at once, we decided to use data from smokers as well as general e-cigarette users to model out predictions. Instead of modeling using data correlating to the individuals who use nicotine and cigarettes, we decided to use sales data, since it more accurately represents the changes in usage in the general population without the complexity of sifting through individualized situations for each person who answers the survey. For the rest of this paper, we will be referring to vaping devices that include nicotine as E-Cigarettes. Since it is a difficult task to consider whether or not daily or weekly, or even monthly smokers belong in the smoking category, as well as classify users based on empirical data, our alternative, or using sales of these devices as functions of time, allows us to analyze the situation using an economics lens, and produce more compelling and realistic trends that follow basic market behavior.

3.1.2 Cigarette Sales

Cigarettes used to be the biggest source of nicotine in the 20th century, at least until the Surgeon General's statement regarding health issues. In order for us to look at future cigarette sales as time continues, we looked to the past to analyze how cigarette sales have been going. But what kind of analysis would we use?

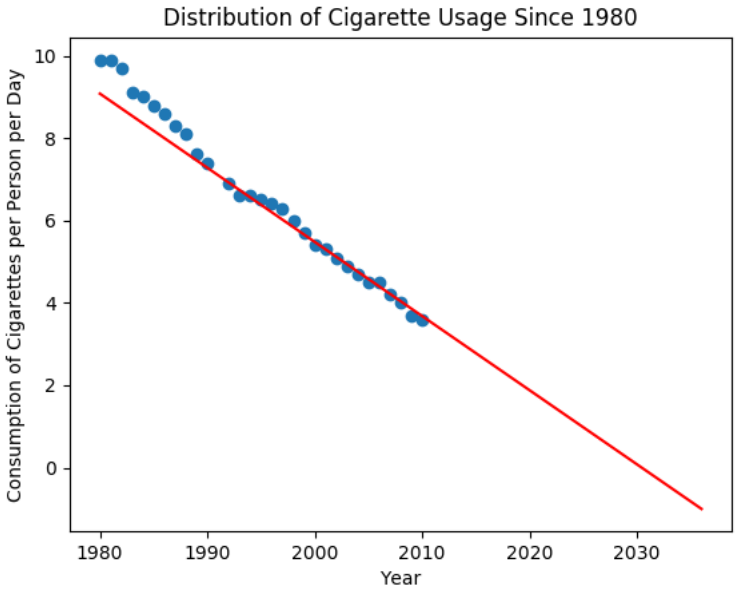
After examining various graphs which depicted the relationships between cigarette usage/sales over time, like the one below:



From the graph above, we can see that the steady decline since 1975 has been quite linear, considering the slope only deviates ever so slightly from its median. As for the increase leading up to 1975, it has a much more exponential characteristic. This makes sense, because of the properties of nicotine. Nicotine is extremely addictive, so as people begin to smoke, they compulsively begin to purchase more, which causes an exponential growth with every new customer. As for the linear decrease, the process to quit and get off the addiction to nicotine is quite long and grueling, and as a result it takes much longer and

its less likely that people do it at the rate they picked up smoking.

In another example, we ran a linear correlation test on another set of data on cigarette consumption, and obtained an R^2 value of 0.96.



This R^2 value suggested a very strong linearity, and afterwards we proceeded to perform a linear regression on the data set and obtain the following model:

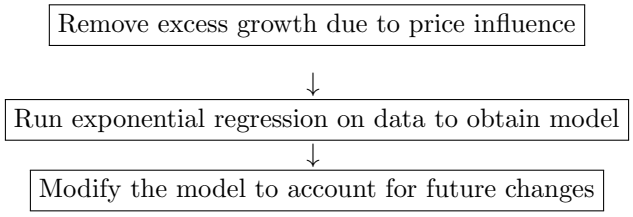
$$C(t) = -0.18t + 365.48$$

$C(t)$ is consumption of cigarettes per person per day in the year t

3.1.3 E-Cigarette Sales

We also decided to use a regression algorithm in this instance. We did this by analyzing the behavioural economics of e-cigarette markets.

One of the important effects that's taken into account is the network effect. This is when an increased number of people using a product improve the value of the good. One good example is the peer-pressure effect, which is the influence one's own social circle has on someone's tendency to do something. The more people who use them, the more those who don't feel left out, and the intrinsic value of owning one because everyone else does drastically increases. This rippling effect that e-cigarettes have, like most viral trends, can be represented by an exponential function with respect to time:

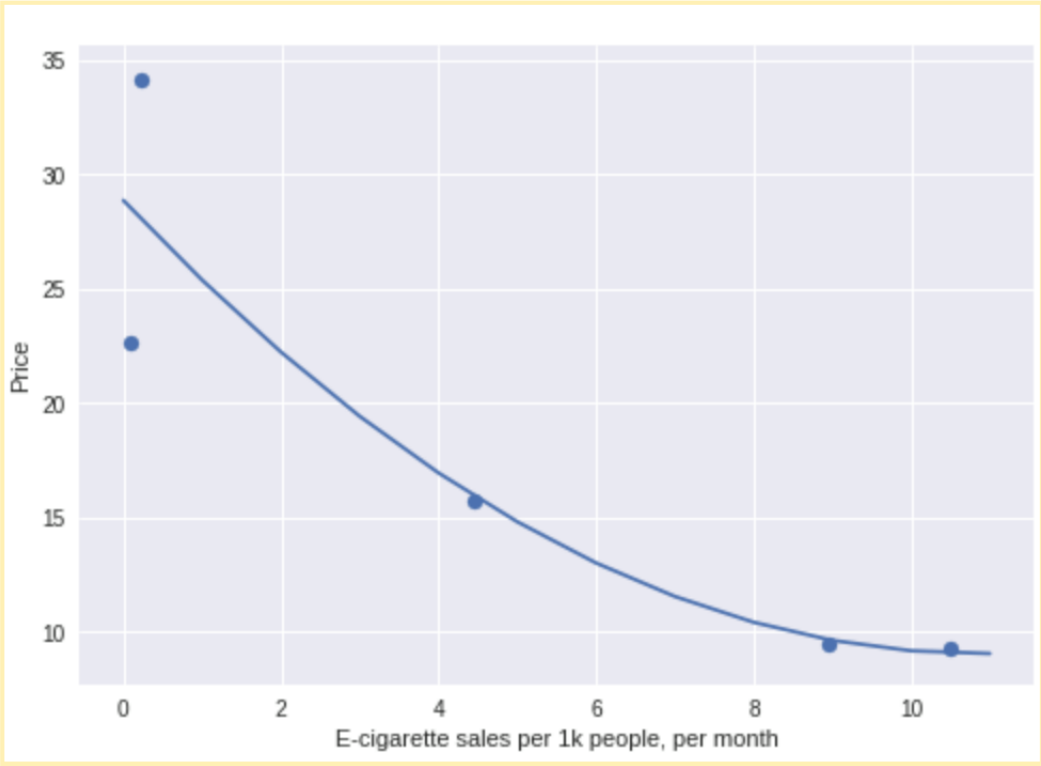


In order to build a successful model, we need to build some auxiliary models first. In order to do so, we must remove the influence of price on the past 5 years of e-cigarette data, by subtracting the excess units sold due to lower prices from the demand curve for e-cigarettes.

Then, we must make a model that predicts past the given data to output predictions for future prices.

3.1.4 Demand Curve Model

In order for us to construct the demand curve model, we need to use the price and demand data from the other countries. These countries are all first world countries with similar control variables such as unemployment, economic growth, and cigarette markets. This means that the final demand curve we find can be used universally as a general demand curve. A quadratic regression was run on the data from these countries, and after the prices were converted from EUR to USD, the following model was obtained:



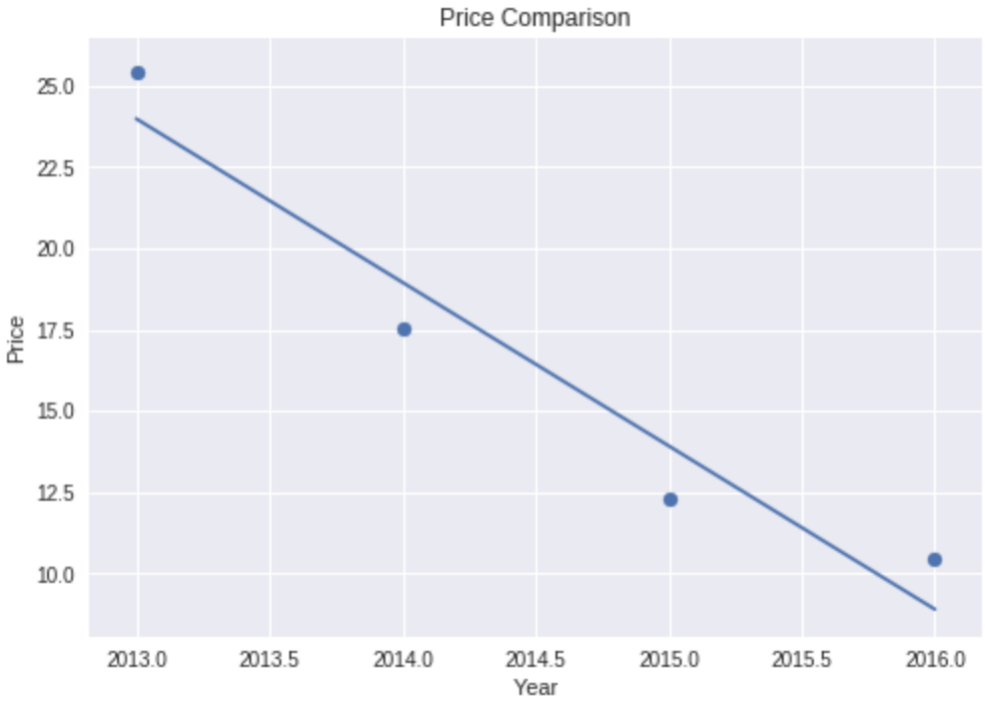
As for the equation:

$$D(P) = 0.16796167867861977P^2 - 3.6467855752006493P + 28.857328720676218$$

3.1.5 Price Prediction Model

To construct the price prediction model, another economic analysis had to be done. E-cigarette technology is new, so there are many innovations to take place. With each innovation, manufacturing becomes more efficient, parts become cheaper, and as a consequence the price of the E-cigarette goes down. Additionally, new companies get involved in the field, bring competition to the market, and lower prices. Through each iteration of this cycle, however, the magnitude of the price decrease goes down. After all, you can only make so many revolutionary changes to E-cigarette technology, and with each company joining the industry, the incentive for other potential firms to join also decreases, as they want to avoid the increased competition. Thus, the rate that the price would decrease would decrease over time. This

conclusion led us to create a logarithmic regression model that aims to predict future price:



4.png

3.1.6 Final E-Cig Model

The natural growth of e-cigarettes (without any influences from price) can be represented as follows, where $G_{overall}(t)$ is the original data set of the number of e-cigarette units sold in the year t

$$G_{natural}(t) = G_{overall}(t) - (D(P(t)) - D(P(t0)))$$

t_0 is the first known year in the $G_{overall}$ dataset .
The excess sales due to price can be calculated as the difference in units demanded from the first price in the data set $G_{overall}$ to the price at time t. Once this quantity is subtracted from $G_{overall}$, the influence of the price has been removed.

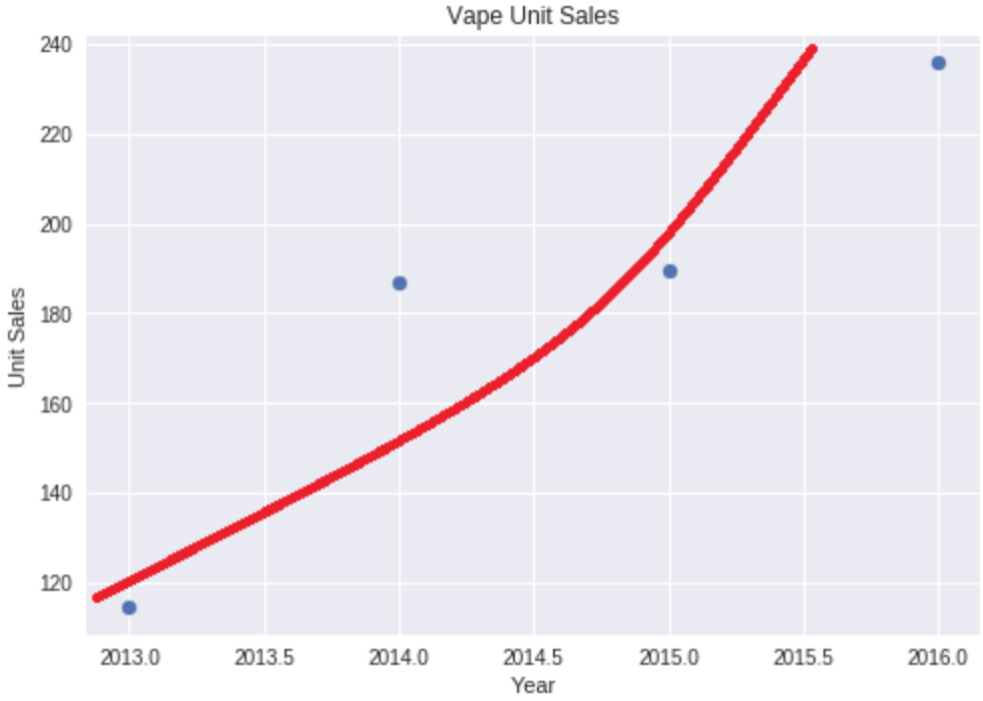
Now, $H_{natural}(t)$ can be found by performing an exponential regression on $G_{natural}(t)$

Then, the final model, $H_{overall}(t)$ can be found by adding the additional units demanded from the further decrease in the future price

$$H_{overall}(t) = H_{natural}(t) + (D(P(t)) - D(P(t_0)))$$

Where t_0 is the last known year in the $G_{overall}$ dataset.

After all is done, the final model should mimic the behaviour of the following graph:



5.png

3.2 Above or Under the Influence?

In order to determine whether a person was likely to use drugs, we used a logarithmic regression model. We used 18 features and gave each feature of weight from -1 to 5. For each feature, or trait, that a person had, we add these weights together to get the regression input. Our prediction goal is a confidence rating. This rating is the average probability that a teenager falls into this feature population multiplied by the features weighting factor. For example, if we know that 22.3% of teenagers had their first drug use at 12-14 years old and that this was given a weight of 2, our confidence score would be 44.6. The reasoning behind using the percentage of teenagers who fall under a certain feature set such as time of first drug use or income level is that if a high percentage of teenagers do a certain act, there is a high probability that a teenager with similar characteristics will act the same way. We add up all these confidence scores then divide by the number of features used to get the regression output.

Feature	Weighting for Feature	Confidence Rating
No Drug Use in the Past Month	-0.5	-14
1-3 in Past Month	1	12.1
1-2 times in Past Week	3	26.4
3-6 times in Past Week	4	45.6
Daily	5	189
First Drug Use 11 and under	1	7.3
First Use at 12-14	2	44.6
First Use at 15-17	3	84
At or Above Poverty Level	1	23.7
Below Poverty Level	2	34.3
Insurance Coverage: Medicaid	1	38
Insurance Coverage: Medicare	2	27.6
Insurance Coverage: Private	-1	-10.7
Insurance Coverage: Uninsured	4	152
Education Level: Still in School	1	29.5
Education Level: Stopped at 8th or Before	3	63
Education Level: Stopped between 9th and 11th	2	73.6
Education Level: Stopped at 12th	1	30.2

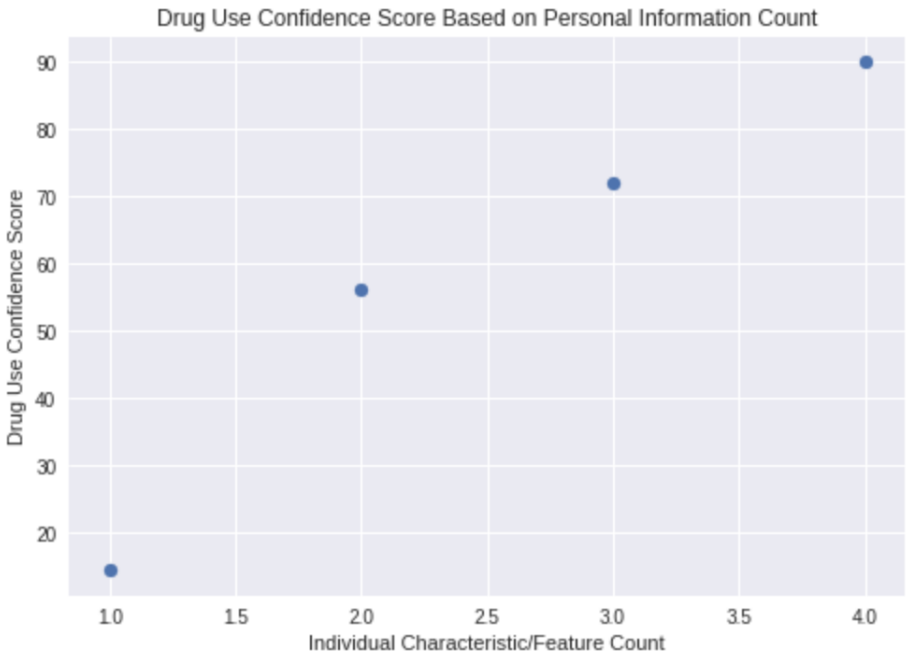
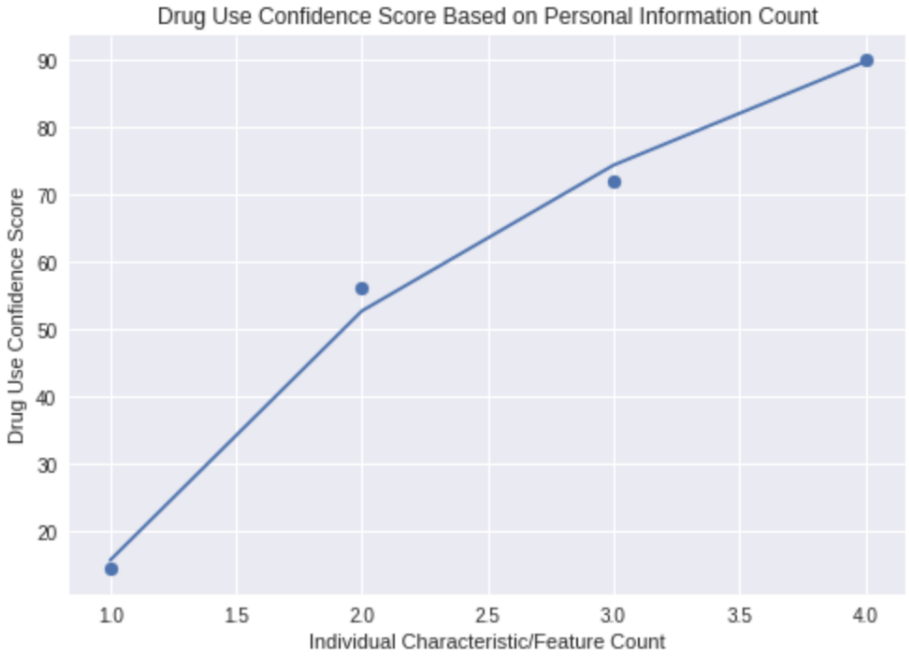


Figure 1: An illustration of a sample set from our data

We fit our data to this dataset using the equation: $a + b * log(x)$ where a is 15.74892141531573 and b is 53.41769633191436.



This model was programmed using data on general drug use from one hospital. To get predictions on more specific drug types, we just need to retrain this model. This flexibility is a key benefit to machine learning models. Our curve fits the data points well and that means all a person needs to do to diagnose drug use is input a yes or no (binary checklist) for the features shown in the above table and then the person gets a confidence score and a yes or no for drug use output is outputted with a thresholding of

60% score (so any confidence score above 60 means the person is using drugs).

3.3 Ripples

We created a model to assess the relative danger of an individual using each drug to society as a whole. Using a variety of variables, we created two indices to assess the financial and social impacts of one person using each drug over the course of one year.

3.3.1 Methods

We decided to assess the potential dangers of each drug in six different categories: three financial and four non-financial, with one shared by both. The financial categories were health care costs, personal costs, and potential imprisonment, the non-financial factors risk of personal death, risk of outside death, the addictiveness of the drug, and potential imprisonment. We incorporated the financial factors into a financial index out of 1, and the non-financial factors in a quality of life index also out of 1. The final index is the sum of both previous indices, and is thus out of 2.

Categories:

Financial:

Health Care Costs:

We calculated the health care costs of each drug per user by taking the total, nationwide health care costs and dividing that price by the number of drug users. This allowed us to find the average health care cost from one user. Even though health care costs are shared between individuals, the government (through programs such as Medicare and Medicaid), and private insurers, a $100 health care cost will always cumulatively cost society $100. Each health care cost is per year.

Personal Costs:

We defined the personal costs of using a drug as the financial burden of drug usage for the average user over the course of a year. To obtain this value, we found the total consumer spending for each drug in America, and divided by the number of users, or found data detailing average weekly/monthly usage for certain drugs.

Non-Financial:

Personal Risk of Death:

This index reflects the probability that the average drug user dies as a direct result of their consumption of that drug in any given year. We found this value by dividing the number of annual deaths from each drug by the total number of users.

External Deaths:

This measure the average number of non-drug users who die as a result of others usage. For example, this would count all motor vehicle deaths involving alcohol-impaired drivers. This number is calculated by dividing the total number of external deaths by the number of users.

Addiction Index:

We used the work of two scientists, Henningfield and Benowitz, to calculate relative addiction indices for each drug. These scientists each ranked the drugs on a scale from 1 to 6, in several different categories to

assess their addictiveness. We dropped the category of intoxication, because the severity of intoxication is already reflected in our other calculations, such as risk of death, and does not impact addictiveness per the scientists. We took the sum of the rankings for each drug, and subtracted them from the maximum score of 48 to assess each drugs addictiveness. A higher score on our index represents a more addictive drug, because before subtraction higher scores (6) corresponded to less danger of addiction. For opioids, we averaged the values given for different specific drugs, which were all non-prescription opioids.

Imprisonment:

We calculated the probability of imprisonment by dividing the total number of drug users by the number incarcerated each year. Thus, each probability reflects the chances of the average drug user being imprisoned in any given year. The financial aspect of imprisonment is reflected in the governments spending on each inmate, on average $33,274. The non-financial aspect came from the loss of future employment opportunities, and lost time while spent in prison. Most convicts unfortunately never fully rebound from their time in prison, and are often re-incarcerated in the future.

Cumulative Indices:

Financial Index:

This index measures the financial cost of one drug user to society. We took the list of the sums of the health care cost and personal costs for each drug, and to scale them, divided each value by the maximum of the list (in this case $174,322 for opioids).

Quality of Life Index:

This index measures the impact of drug use on the long term quality of life of a user, and other members of society. We added the probabilities of deaths (both personal and external) and imprisonment together, because as discussed before imprisonment largely ruins future economic and social success for a person, due to high recidivism rates and job placement issues for ex-convicts. We then multiplied this probability sum by the addiction index. We did so because addictive drugs pose the greatest danger to someones quality of life, because more frequent usage both increases the chances of conviction and death in the short term and long term, especially following release from prison.

Final Index:

To obtain the final values for each drug we summed the financial and quality of life indices, because we thought that large financial impacts can be just as devastating as imprisonment for impoverished Americans, many of which would be pushed into debt by even a small, $300 medical expense. Debt increases the chances of death for entire households, by depriving them of critical resources needed for health and survival. Moreover, a lower quality of life for individuals in a household means they will be financially burdened in the future. Because both indices impact each other so much, we could not and felt that it would be impossible and arbitrary to reconcile one as being more important than the other.

3.3.2 Data:

Drug:	Nicotine	Alcohol	Marijuana	Opioids
Healthcare Cost/Person	4473	123	0**	174,322
Personal Costs	2105	565	2080	0**
Probability of Imprison	0%**	0%**	1.42%	11%
Probability of Death	1.263%	0.005025%	0%**	1.368%
Probability of External Death	0%**	***	0%**	0%**
Addiction Index	29	25	3	32
Financial Index	0.037	0.0038	0.0014	1.0
Quality of Life Index	0.84	0.0029	0.00098	1.0
Final Value	0.88	0.0067	0.015	2.0

*All values are estimates.
**These values are close to 0, dwarfed by other related values (ex. Health care costs vs. personal costs in the case of opioids), or insignificant and thus will not significantly affect the final scores.
***The statistics used to calculate the probability of death resulting from alcohol consumption grouped both user deaths and external deaths together.

These methods produced a ranking system of the four drugs in terms of their severity. In order of decreasing harms to society, the final ranking was non-prescription opioids, nicotine, marijuana, and last alcohol. Opioids come in clear first in terms of their dangers, and are most dangerous in both the financial index and quality of life index.

3.3.3 Discussion:

Overall, the results came as expected. We thought that opioids would be the most dangerous drug for society per user, as they are simply more potent than the other drugs. The largest surprise from our data stemmed from the fact that marijuana ranked higher than alcohol in terms of its impact on society, as many experts have argued that alcohol abuse is far more dangerous than marijuana, especially because alcohol is much more addictive. However, our analysis accounted for two factors beyond the simple danger of each drug:

First, we accounted for the chances of imprisonment. When scientists argue that marijuana is similarly dangerous to alcohol, they are not accounting for the fact that marijuana is currently illegal. In fact, often they are using this as a reason to argue that marijuana should be legal! That gives marijuana a larger cost to society.

Second, most alcohol use involves controlled, limited usage at home, not abuse. Almost 200 million Americans consume alcohol regularly, but few of them abuse the substance or endanger others lives while under its influence. So while alcohol addiction is worse than marijuana use, on average, marijuana is more threatening, because of its higher cost and threat of imprisonment. Ironically, both of these factors likely stem from the fact that marijuana is illegal; if it were not, then the drug would be more available on the market, pushing down prices.

One caveat to our results is that these indices do not in any way represent the harm each drug poses to society. This is because the number of regular users of each drug vary wildly, from 199 million for

alcohol to fewer than one million for opioids. Our results only measure the severity of each drug in an individual, and thus compare the ripple effects of one person using each drug.

3.3.4 Sensitivity Analysis

One of the major obstacles we faced was deciding how to compare the two indices to each other. One of the core weaknesses that our system has is that a large enough outlier in either data pool could render one of the indices insignificant, since each drugs values are scaled with comparison to that of the worst drug in that category. With more data regarding different drugs, we could have come up with a better system for comparing these two values, where the drugs would be ranked in each category with regards to a mean or median. However, with only four data values to work with, we did not feel comfortable using these methods because the mean and median would have been next to meaningless, especially considering the fact that the data points spanned many different orders of magnitude.

3.4 References

https://www.wikileaf.com/thestash/marijuana-price/
https://www.healthline.com/health-news/can-marijuana-kill-you1

https://www.npr.org/sections/health-shots/2018/02/13/585199746/cost-of-u-s-opioid-epidemic-since-2001-is-1-trillion-and-climbing
https://www.marketwatch.com/story/whats-your-net-worth-and-how-do-you-compare-to-others-2018-09-24
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2838492/
https://www.drugabuse.gov/publications/research-reports/heroin/scope-heroin-use-in-united-states
https://www.asam.org/docs/default-source/advocacy/opioid-addiction-disease-facts-figures.pdf
https://www.alcoholrehabguide.org/alcohol/crimes/
https://www.thebalanceeveryday.com/what-lifetime-of-drinking-costs-4142309
https://www.reuters.com/article/us-health-marijuna-us-adults/one-in-seven-us-adults-used-marijuana-in-2017-idUSKCN1LC2B7
https://www.verywellmind.com/how-many-people-drink-alcohol-in-the-us-67305
https://www.scientificamerican.com/article/the-truth-about-pot/
http://www.drugpolicy.org/issues/drug-war-statistics
https://www.ncbi.nlm.nih.gov/pubmed/6514753
https://www.reuters.com/article/us-healthcare-costs-smoking-idUSKBN0JX2BE20141219
https://money.cnn.com/galleries/2011/pf/1105/gallery.money$_w asters$/4.*html*
https : *//whyquit.com/whyquit/*A_H*enningfield*$_B$*enowitz.html*
http : *//www.center4research.org/vaping* − *safer* − *smoking* − *cigarettes* − 2/
https : *//www.cdc.gov/media/releases/*2018/*p*0118 − *smoking* − *rates* − *declining.html*
https : *//www.cdc.gov/media/releases/*2018/*p*0118 − *smoking* − *rates* − *declining.html*
https : *//www.cdc.gov/tobacco/data*$_s$*tatistics/fact*$_s$*heets/fast*$_f$*acts/index.htm*
https : *//www.cdc.gov/mmwr/pdf/wk/mm*6444.*pdfpage* = 1
https : *//www.icpsr.umich.edu/icpsrweb/NAHDAP/studies/*30122/*datadocumentation*
https : *//www.cdc.gov/pcd/issues/*2018/17$_0$555.*htm*
https : *//www.investopedia.com/terms/n/network* − *effect.asp*
https : *//ourworldindata.org/smoking?fbclid* = *IwAR*0*wbOt*$_i$*Y*0*hI*9*aX*$_D$*sLnliQHWW*2*CwV*1*DAcKrAynCNU*04*UJ*
https : *//www.drugabuse.gov/publications/research*−*reports/tobacco*−*nicotine*−*e*−*cigarettes/nicotine*−*addictive*
https : *//sci* − *hub.tw/https* : *//jamanetwork.com/journals/jama/article* − *abstract/*2705175
https : *//m*3*challenge.siam.org/sites/default/files/uploads/Figure*$_A$*dult*%20*per*%20*capita*%20*cigarette*%20*consump*
PricesandE−*CigaretteDemand* : *EvidenceFromtheEuropeanUnionMichalStoklosaMA*1, *JeffreyDropePhD*1, *Fran*

4 Code

Here is a link:
`https://github.com/jrmkim50/megamoody2019`