MathWorks Math Modeling Challenge 2019

High Technology High School–

Team # 12038 Lincroft, New Jersey

Coach: Raymond Eng

Students: Eric Chai, Gustav Hansen, Emily Jiang, Kyle Lui, Jason Yan

MathWorks Math Modeling Challenge Champions

$20,000 Team Prize

MathWorks Math Modeling Challenge

***Note: This cover sheet has been added by SIAM to identify the winning team after judging was completed. Any identifying information other than team # on a MathWorks Math Modeling Challenge submission is a rules violation.

***Note: This paper underwent a light edit by SIAM staff prior to posting.

Substance Use and Abuse

Executive Summary

In recent years, substance abuse has intensified to an alarming degree in the United States. In particular, the rise of vaping, a new form of nicotine consumption, is dangerously exposing drug abuse to a new generation. With the need to understand how substance use spreads and impacts individuals differently, our team seeks to provide a report with mathematically founded insights on this prevalent issue.

We first strove to predict the spread of nicotine use due to both vaping and cigarettes over the next decade. By comparing the spread of nicotine use to an infectious disease, we modified the SIRS epidemiology model to create our adapted SIRI model in which individuals are divided into four compartments: infected (drug users), recovered (users who quit drugs), susceptible (potential drug users), and nonsusceptible (those who will never use drugs). People progress from susceptible to infected to recovered, but may relapse into their old habits, causing them to re-enter the infected population. Birth and death rates of our designated population were modeled with linear equations. We solved a system of differential equations to determine e-cigarette and cigarette use in 2029: 26.63% of the American population will vape and 6.45% will smoke cigarettes. These results align with the expectation that vaping will increase in popularity while cigarette smoking will decline.

Substance abuse is associated with numerous social factors and personal attributes. We incorporated those determinants to create a second mathematical model that computes the probability that an individual will use nicotine, marijuana, alcohol, and unprescribed opioids. A binary multivariate logistic model was used to assess the effects of age, gender, ethnicity, income, parental status, friendship, opinion about school, overall health, weapon possession, and bullying on substance use. To demonstrate our model, we coded and executed a Monte Carlo simulation that created 300 high school seniors with varying attributes. We found that 46.3% of the students would use nicotine, 17.3% would use marijuana, 66.0% would use alcohol, and 0.0% would use opiates.

Substance use has far-reaching implications in personal and societal spheres. It is crucial to rank substances based on their overall impact in order to assess necessary government action regarding drug abuse. To address this issue, we developed a robust metric to rank the effects of nicotine, marijuana, alcohol, and opioid abuse. Our model and ranking considers physical harm, dependence, social harm, and economic impact of the drugs. The former three factors were measured on a scale of 0 to 3 based on psychiatrist surveys. Then economic impact was defined as GDP loss from the decrease in life expectancy caused by drug abuse. After applying risk factors obtained from the amount of people that use each drug, the four substances were ranked. From highest to lowest individual impact, the ranking was opioids, alcohol, cigarettes, and marijuana. From highest to lowest total societal impact, the ranking was alcohol, cigarette, marijuana, and opioids.

The repercussions of substance abuse are reverberating and remain with an individual for life. However, drugs not only severely affect the user but also cause extensive societal harm. Increased understanding of the projected spread and impact of substance abuse, as well as the underlying factors that lead to poor judgment, are needed to optimize measures to restrict consumption. Ultimately, we believe that our models provide novel insight into the nationwide issue of substance use and abuse.

Contents

1 Introduction

This section delineates the components of the modeling problem and their objectives. Global assumptions applying to the entire modeling process are also listed.

1.1 Restatement of the Problem

The problem we are tasked with addressing is as follows:

1. Build a mathematical model that predicts the spread of nicotine use due to vaping over the next 10 years. Analyze how this growth compares to that of cigarettes.

2. Create a model that simulates the likelihood that a given individual will use a given substance, accounting for social influence, characteristic traits, and properties of the drug itself. Demonstrate the model by predicting how many students among a class of 300 high school seniors with varying characteristics will use nicotine, marijuana, alcohol, and unprescribed opioids.

3. Develop a metric for the impact of substance use, considering both financial and nonfinancial factors. Use the metric to rank the substances listed in Part II.

1.2 Global Assumptions

1. *The current drug scene remains constant.* We assume that there will be no radical changes in the recreational drug industry, such as new drugs or drug products. This assumption is imperative because attempting to account for unpredictable and volatile factors would make model development virtually impossible.

2. *All vapes count as e-cigarettes.* Some people distinguish between e-cigarettes and vaping. For the purposes of this model, e-cigarettes and vapes will be considered synonymous.

3. *People respond honestly to surveys.* Our model is dependent on survey results to calculate weight constants. Because we have no way of determining the accuracy of the survey responses, we will assume that they are accurate and without bias for simplicity.

2 Part 1: Darth Vapor

First commercialized in 2003, electronic cigarettes have become an increasingly popular product among youth [1]. Although they are advertised as safer alternatives to traditional cigarettes, e-cigarettes contain high doses of nicotine and have introduced a new generation to tobacco products. This section outlines a mathematical model for predicting the change in nicotine use in the United States due to vaping compared to the change due to cigarettes.

2.1 Assumptions

1. *Nicotine use can be modeled as an infectious disease.* Like an epidemic, nicotine use is prevalent and contagious, reflected in the surge in popularity of smoking due to peer pressure, advertisements, and social media. Additionally, the U.S. Surgeon General declared youth vaping a nationwide epidemic in 2018 [2].

2. *Individuals can smoke from age 11 until death.* Peak years for first trying nicotine products is 6th or 7th grade [3].

3. *Rate of entry into pre-adolescence in the U.S. is 0.00103.* [4] Our model defines "birth" as reaching an age at which substance use becomes possible—around 11 years. Thus, we assumed the current birth rate to be constant for the past 11 years, assuming no children die before they turn 11. The current birth rate is 1.03 people/month/person.

4. *Death rate in the U.S. is constant and equal to 0.0007 people per month per person.*[4] Our model assumes that individuals have the capacity to use drugs until their death.

5. *Individuals can only start smoking due to influence from other smokers.* To model substance use as an infectious disease, we must assume that susceptible individuals can become infected only from contact with the already infected. This assumption is valid because peer influence and social media presence are the driving factors behind the popularity of smoking [5].

6. *Individuals are either not susceptible to, susceptible to, infected by, or recovered from substance abuse.* As in the SIR epidemiology model, we assume that people are either unwilling to smoke (not susceptible), open to smoking (susceptible), regular smokers (infected), or past smokers who have quit (recovered).

7. *The infection rate is constant over time.* Because we are assuming that the drug industry does not drastically change, it is reasonable to assume that the infection rate will also not drastically change.

8. *The percentage of susceptible people will stay constant over time.* Because we are assuming that the drug industry does not drastically change, it is reasonable to assume that the number of people susceptible to it will also not drastically change.

9. *Nobody starts as recovered.* At the start of the model, we do not consider any individuals to be former smokers who have quit.

10. *The recovery and relapse constant for cigarette and e-cigarette users are the same.* The two contain similar amounts of nicotine, which acts as the addictive agent. Thus, the recovery and relapse constants are assumed to be the same.

2.2 Model Development

The surge in popularity of conventional cigarettes in the mid-20th century, as well as the current boom of vaping among American youth, is comparable to the spread of an infectious disease during an epidemic. As stated in assumption 1, we model nicotine use as a disease because it rapidly spreads as a result of interpersonal communications (in-person peer pressure to try a drug as well as social media prevalence); additionally, substance use is a condition from which individuals can recover (by quitting smoking).

Our model is a derivation of the SIRS epidemiological model, a technique used to map the spread of infectious diseases such as influenza. We also consider birth and death rate, since population naturally changes over time. The model separates individuals in a population into four categories: NS for Not Susceptible, S for Susceptible, I for Infected, and R for Recovered. At the start of the model, individuals are either in NS, S, or I, since nobody starts off as recovered. While those in NS remain there permanently, individuals in S can move to I, who can then move to R.

The additional S in SIRS represents the possibility of returning to the Susceptible compartment—in this case, a regular user quitting but relapsing. However, we modified the classic SIRS model by recognizing that a relapsing individual would re-enter the Infected category rather than Susceptible, since they will once again become smokers rather than people merely open to smoking. Thus, we renamed the traditional epidemiology model as SIRI to represent this adjustment. Figure 2.2.1 diagrams the aforementioned movement of individuals between categories, while Table 2.2.1 defines and details values for variables and constants used in the SIRI model for both e-cigarette and cigarette smoking.

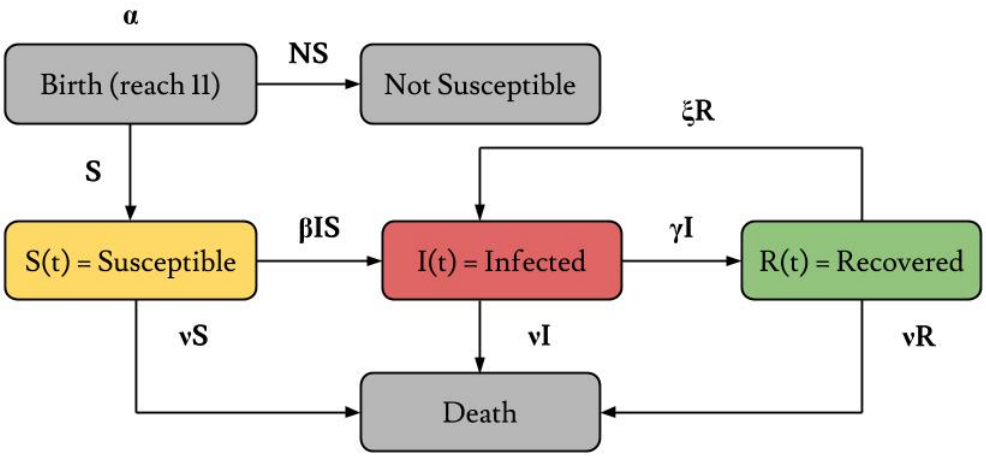

Figure 2.2.1: Diagram of the SIRI Model for Spread of Nicotine Use

2.2.1 Parameters in SIRI Model

Proportion of infected people (I_0). The total number of people that currently vape is approximately 10.8 million [6]. Dividing by the total population of America, 325.7 million

[7], results in an I_0 value of 0.0332 for e-cigarettes. The total number of people that currently smoke cigarettes is approximately 34.3 million [8], resulting in an I_0 value of 0.1053.

Proportion of recovered people (R_0). As per assumption 9, without loss of generality, R_0 was assumed to be 0 at time = 0.

Proportion of susceptible people (S_0). Because I, R, and S are proportions of the total population, their sums must add to 1. Thus, $S_0 = 1 - R - I$, resulting in 0.9667 for e-cigarettes and 0.8947 for cigarettes.

Susceptibility (S). A 2016 Surgeon General report stated that 32% of people are considered susceptible to e-cigarette use [5], while a 2012 report stated that 20% of people are susceptible to cigarettes, which correspond to the S values [9].

Infection constant (β). This was determined based on responses to the survey question "If one of your best friends offered you a cigarette, would you smoke it?" For e-cigarettes, the chance of infection was taken from a 2016 U.S. Surgeon General report that indicated that 18% of young adults responded "yes" to the question [5]. For cigarettes, we obtained β by adding the percentages of the responses "Definitely Yes" and "Probably Yes," from the 2014 National Survey on Drug Use and Health, to get 0.3%, which represented the infection constant [10].

Recovery constant (γ). In a given year, around 40% of smokers attempt to quit [11]. Therefore, in a month, $1.40^{1/12} = 1.0284$ recover, so the recovery rate is 0.0284.

Relapse constant (ξ). In a given year, approximately 6% of attempts to quit smoking succeed and 94% of attempts failed and the person relapsed [12]. Therefore, in a month, $1.94^{1/12} = 1.0568$ fail, so the relapse constant is 0.0568.

Infection rate (y_{inf}). In accordance with assumption 4, we assume that people will only start smoking if they are influenced by a current smoker. In other words, a susceptible person can only become infected if they come into contact with an infected person, which occurs at a rate proportional to $I \cdot S$. The infection constant β represents the likelihood that a susceptible person becomes infected when influenced by a smoker. Thus, infection rate is as follows:

$$y_{inf} = \beta \cdot I \cdot S \tag{1}$$

Recovery rate (y_{rec}). Unlike infection rate, the recovery rate is dependent only on the average probability of an individual quitting. The recovery constant γ multiplied by the proportion of people that currently are infected gives the recovery rate:

$$y_{rec} = \gamma \cdot I \tag{2}$$

Relapse rate (y_{rel}). The relapse rate is dependant only on the average probability of an individual relapsing. The relapse constant is much higher than the infection rate, which is logical because an individual who was previously a regular smoker will be more likely to succumb to the addictive cycle again [12]. Designating ξ as the relapse constant, relapse rate is given by

$$y_{rel} = \xi \cdot R \tag{3}$$

Birth rate (α). The birth rate, as defined by assumption 3, is 1.03 people/month/person.

Death rate (μ). From assumption 4, the death rate is assumed to be constant and equal to 0.0007 people per month per person. Therefore, the number of people dead for each category will be the death rate multiplied by the proportion of the people in each category.

$$\mu_S = v \cdot S \tag{4}$$

$$\mu_I = v \cdot I \tag{5}$$

$$\mu_R = v \cdot R \tag{6}$$

Table 2.2.1 Variables and Constants of SIRI Model for E-Cigarettes and Cigarettes

Variable	Definition	E-Cigarette Values	Cigarette Values
I	Proportion of infected people	$I_0 = 0.0332$	$I_0 = 0.1053$
R	Proportion of recovered people	$R_0 = 0$	$R_0 = 0$
S	Proportion of susceptible people	$S_0 = 0.9667$	$S_0 = 0.8947$
N	Proportion of total individuals in SIR cycle	$N_0 = 0.32$	$N_0 = 0.20$
α	Birth rate	0.00103	0.00103
β	Infection constant	0.18	0.003
γ	Recovery constant	0.0284	0.0284
ξ	Relapse constant	0.0568	0.0568
μ	Death rate	0.0007	0.0007

2.2.2 Differential Equations for SIRI Model

The change in each of the dependent variables S, I, and R is equal to the sum of the input of the respective category minus the sum of its output, as diagrammed by the arrows entering and leaving each box in Figure 2.2.1. Thus, our SIRI model is summarized by the set of ordinary differential equations below:

$$\frac{dS}{dt} = \alpha - \beta \cdot I \cdot S - \mu \cdot S \tag{7}$$

$$\frac{dI}{dt} = B \cdot I \cdot S - \gamma \cdot I + \xi \cdot R - \mu \cdot I \tag{8}$$

$$\frac{dR}{dt} = \gamma \cdot I - \xi \cdot R - \mu \cdot R \tag{9}$$

2.3 Results

With the SIRI model established, we utilized it to predict the change in nicotine use due to e-cigarettes and cigarettes in the next decade. We coded and executed a Python program to solve the system of differential equations, with appropriate constants for each product, and graph the proportion of compartments over time. Figures 2.3.1 and 2.3.2 graph the proportion of the total population falling under each of the SIR categories for both tobacco products, respectively, over a 10-year time period. Table 2.3.1 enumerates

the proportion of the population that is susceptible, infected, and recovered for vaping and cigarettes in 2029.

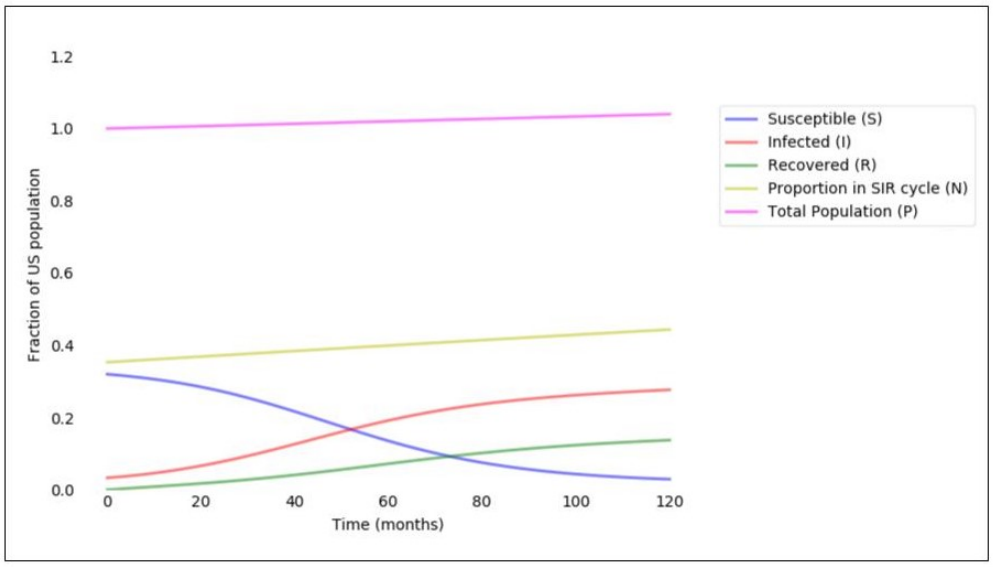

Figure 2.3.1: Graph of SIRI Compartments for E-Cigarettes over Ten Years

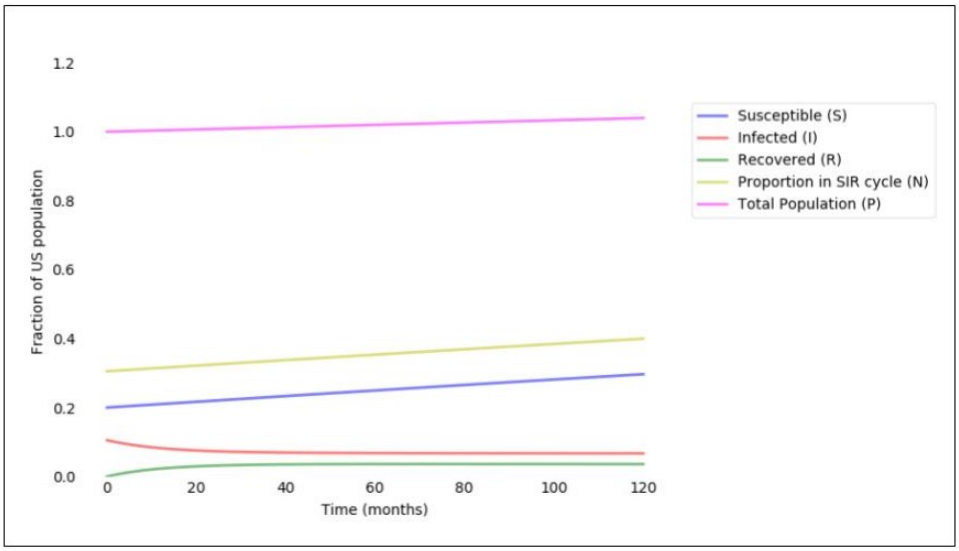

Figure 2.3.2: Graph of SIRI Compartments for Cigarettes over Ten Years

Table 2.3.1 SIR Distribution of 2029 Population for E-Cigarettes and Cigarettes

	Susceptible	Infected	Recovered
E-Cigarettes	2.82%	26.63%	13.21%
Cigarettes	28.53%	6.45%	3.45%

Our model concludes that in 2029, 26.63% of the population will use e-cigarettes, while 6.45% will use cigarettes. This disparity is consistent with previously researched trends,

which suggest that as e-cigarettes gain popularity amongst teens, regular cigarettes decrease in popularity [13].

2.4 Sensitivity Analysis

Table 2.4.1 shows the sensitivity analysis for our SIRI model based on an independent increase and decrease of 10% of the infection constant β, recovery constant γ, and relapse constant ξ.

Table 2.4.1 Sensitivity Analysis for Part I

Constant	% Change in Constant	% Change in Vaping (I)	% Change in Cigarette Use (I)
β	10%	1.014%	0.6202%
β	-10%	-1.615%	-0.6202%
γ	10%	-3.492%	-3.566%
γ	-10%	4.018%	3.721%
ξ	10%	3.098%	3.367%
ξ	-10%	-3.496%	-3.905%

Positive changes in the infection or relapse constants resulted in positive changes in the percentage of infected people for both vaping and cigarette use. This is consistent with our predictions because the rate of infection for susceptible and recovered people is increasing. In contrast, a positive change in recovery constant resulted in a decrease in percent infected because the rate at which people are leaving the infected population is increasing.

2.5 Strengths and Weaknesses

Our model is resilient to small changes and outputs sensible results. As demonstrated in the sensitivity analysis, a 10% change in each of the infection, recovery, and relapse constants accounts for less than 5% change in final vaping and cigarette use after a decade. Changes in the model's output due to shifts are consistent with expected trends as well. SIRS is also an established mathematical modeling technique that we adapted to fit our own aims, lending credence to the validity of our model. Additionally, our model is comprehensive, accounting for many contributing factors such as population change, nonsusceptible individuals, and the possibility of relapse for smokers who have attempted to quit.

The model's weaknesses lie in its inability to account for the introduction of new forms of drugs or rapid changes in popularity of existing forms, as stated in global assumption 1. Specifically, a surge in use of a particular drug would likely impact vaping and cigarette use in unforeseen ways that our model will not accurately predict. Furthermore, our model does not consider the association between vaping and cigarette use, and how the growth or decline of one product would influence the other. This is unrealistic because the popularity of e-cigarettes among youth has led many to smoke traditional cigarettes and prompted cigarette smokers to transition to vaping [13]; however, the opposite effects of these two phenomena can reasonably counterbalance each other.

3 Part II: Above or Under the Influence

Numerous internal and external factors, such as age, gender, health, family background, and behavior, affect the likelihood of an individual becoming addicted to a substance [14]. We incorporated these determinants to develop a second mathematical model that simulates the probability that an individual will abuse nicotine, marijuana, alcohol, and opioids. Then we created a Monte Carlo simulation to predict the frequency of drug abuse in a class of 300 high school seniors.

3.1 Assumptions

1. *High school seniors are 17 years old.* This is the average age of a 12th grader in the United States [15]; assuming this simplifies the demonstration of our model.

2. *The 2005–2006 HBSC survey sample is representative of U.S. high school seniors* [16]. This is the dataset used in our model demonstration, so we must assume that students answered the questions honestly and that the population sampled is representative of the entire population of U.S. 12th-grade students.

3. *All tobacco products contain nicotine.* The HBSC survey question asks students only if they have ever used any tobacco product (e.g., vape, cigarettes) [16]. This relates the survey to our model's aim of gauging probability of nicotine use and is a logical assumption since nicotine is derived from tobacco plants.

4. *The only opioids considered in the model are heroin and morphine.* This is in accordance with the HBSC line of questioning, which is the only dataset we used to train our model [16].

3.2 Model Development

The 2005–2006 Health Behavior in School-Aged Children survey samples approximately 9200 students using a series of 80 questions. Four of these ask whether a student had used nicotine (tobacco), marijuana, alcohol, and opioids use in the past year. Comparing student responses to these four questions to their responses to the other questions helped us identify the relationship between certain social factors and substance use.

Specifically, the questions listed in Table 3.2.1 were used to analyze factors and characteristics we deemed to significantly influence the likelihood of substance use. We assessed the impact of age, gender, ethnicity, income, parental status, friendship, opinion about school, overall health, weapon possession, and bullying in our model.

Table 3.2.1 HBSC Survey Questions for Independent Variables

Question Number	Question
1	What is your age?
2	What is your gender?
3	What ethnicity are you?
4	How well off is your family?
5	Does your mother live with you?
6	Does your father live with you?
7	How many evenings per week do you spend out with your friends?
8	How do you feel about school?
9	What is your overall health?
10	How many days have you carried a weapon during the past month?
11	How often do you bully another student?

Table 3.2.2 HBSC Survey Questions for Dependent Variables

Question Number	Question
1	Have you ever smoked tobacco?
2	Have you used marijuana in the past 12 months?
3	Have you drank alcohol in the past 12 months?
4	Have you used any unprescribed opioids in the past 12 months?

3.2.1 Model Training

We used a binary multivariate logistic model, given by equation $P(Y)$, where b_0 is the intercept, b_n is the weight given to the nth parameter, x_n is the nth parameter, and $P(Y)$ is the probability of Y occurring.

$$P(Y) = \frac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_nx_n}}{1+e^{b_0+b_1x_1+b_2x_2+\cdots+b_nx_n}} \tag{10}$$

Data from the 2005–2006 HBSC survey was preprocessed in the following ways to make it more suitable for the regression and to decrease accidental bias:

1. Data with missing dependent variables were dropped as they would not be useful for training our model.

2. Missing data in the independent variables were flagged and interpolated using existing values. Interpolation is necessary due to the small size of the dataset and the use of L2 regularization to distribute weights more evenly, the latter of which reduces the volatility of the model to slight inaccuracies.

3. Question 7 (evenings spent with friends) was split into 3 evenly distributed bins to decrease the spread of the data and normalize it closer to 0, improving accuracy.

4. All questions that asked if the student had used a drug in the past 12 months was modified to reflect only whether they had taken the drug at all (yes or no).

5. Data for an individual's race were transformed into dummy variables, where each variable represents the absence or presence of that race. The data was split into 2/3 for training the model, and 1/3 for assessing accuracy.

A machine learning algorithm was programmed and executed to determine the b constants, which are the respective weights of the answers for each of the questions. Our use of L2 regularization decreased insignificant weights close to 0, thus generalizing the model and more uniformly distributing the significance of each parameter. Given the loss function below, which represents the mean squared error to be minimized, where y_i represents the actual value and $h_\theta(x_i)$ represents the current test value,

$$L(x, y) = \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2 \tag{11}$$

a complexity term was added to penalize larger weights, as follows, where λ is a constant to control regularization and θ_i is the weight:

$$L(x, y) = \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2 + \lambda \sum_{i=1}^{n} \theta_i^2 \tag{12}$$

Along with the question weights in Table 3.2.3, the algorithm determined the bias values, b_0 for each of the four drugs as shown in Table 3.2.4. The bias value represents the default probability that an individual will use a substance when all features are set to zero.

Table 3.2.3 Weights for 11 Survey Questions

Question Number	Nicotine Weight	Marijuana Weight	Alcohol Weight	Opioids Weight
1	1.972	−2.215	−1.737	−0.708
2	1.716	−2.107	−1.431	−0.573
3	0.941	−0.877	−0.712	−0.483
4	0.555	−0.746	−0.447	0.237
5	0.926	−0.912	−1.024	0.086
6	0.101	−0.021	−0.232	−0.017
7	0.419	−0.667	−0.071	−0.484
8	0.213	−0.462	−0.387	−0.38
9	0.532	−0.638	−0.295	−0.24
10	1.638	−1.978	−1.681	−0.617
11	2.049	−2.345	−1.487	−0.664

Table 3.2.4 Bias Values for Four Substances

Substance	Bias Value b_0
Nicotine	3.688
Marijuana	−4.323
Alcohol	−3.169
Opioids	−1.281

3.2.2 Model Demonstration

A Monte Carlo simulation was developed to determine the percentage of high school seniors in the United States that would use nicotine, marijuana, alcohol, and opioids. Probability distributions of various parameters were obtained from the 2005–2006 Health Behavior in School-Aged Children survey. Those probability distributions were then inputted into a Python program, as detailed in Appendix A.2, to randomly generate a sample of 300 seniors with attributes selected according to the distributions. Those students were then run through the model developed above to determine the percentage of seniors that used nicotine, marijuana, alcohol, and opioids.

3.3 Results

Using 1/3 of the data for testing, the model's accuracies of prediction for nicotine, marijuana, alcohol, and opioid use were determined.

Table 3.3.1 Accuracy of Substance Use Prediction Model

Substance	Percent Predictions Correct
Nicotine	78.16%
Marijuana	86.73%
Alcohol	73.30%
Opioids	97.80%

Table 3.3.2 Substance Use among Simulated High School Seniors

Substance	Percent of High School Seniors Using Substance
Nicotine	46.33%
Marijuana	17.33%
Alcohol	66.00%
Opioids	0.00%

In Table 3.3.1, the model's especially high accuracy of predicting opioid use is due to the small sample size and our classification of opioids as heroin and morphine only, limiting the number of total users.

Additionally, Table 3.3.2 reports virtually no use of opioids among high school seniors. This is again because our model uses only heroin and morphine to represent all opioids, an assumption that reduces the chance of an individual using that drug in the simulation. However, since the opioid epidemic is a recent development, the model output is logical considering the time of data collection, which was 2005–2006.

Table 3.3.3 Comparison of Simulation Output and Actual Substance Use in 2005

Substance	Simulated % of Seniors Using Substance	% of Seniors Using Substance in 2005	% Error
Nicotine	46.3%	50.0% [17]	−7.34%
Marijuana	17.3%	16.4% [18]	5.67%
Alcohol	66.0%	68.6% [17]	−3.79%
Opioids	0.0%	0.8% [17]	−100%

Above, Table 3.3.3 compares our simulated percentage of seniors using a particular substance to other surveys to gauge its consistency. The relatively low percent errors between simulation output and actual survey values indicate that our model is sensible and sound. Although the opioids row has 100% error, this is because the simulation predicted 0% opioid use among seniors. The actual value is 0.8%, a reasonably close percentage. Otherwise, small variations in percentages can be attributed to different sample populations.

3.4 Strengths and Weaknesses

One of the greatest strengths of our model is the accuracy of its machine learning algorithm to predict usage of a substance. The preprocessing of the survey data and use of L2 regularization made the model resistant to occasional inaccuracies in survey responses or to slight changes in its parameters. These features support the resilience of our model. Moreover, our substance use prediction model has high accuracy and adaptability to new data sources, making it reusable for future datasets. Furthermore, our simulation results very closely match actual proportions of drug use in 2005 obtained from other surveys.

The main drawbacks of the model are the outdated data (from 2005–2006) used to develop it and its inaccurate classification of opioids as only heroin and morphine. The former attribute makes our model applicable only to the year for which it was given data. However, as aforementioned, the method we used to develop our model can be repeated on any dataset containing the appropriate information. The latter flaw resulted in a small sample size with which to train the model as well as an underestimation of opioid usage among high school seniors in the simulation.

4 Part III: Ripples

Substance abuse has far-reaching implications, both personal and societal. Considering financial and other factors, we created a metric that quantifies the impact of drug abuse and ranks nicotine, marijuana, alcohol, and unprescribed opioids.

4.1 Assumptions

1. *Each of the four categories of drug harm are weighted equally.* For the purposes of this model, since determining the relative importance between the categories was subjective, we assigned all categories equal weight.

2. *The model applies only to substance abusers.* Since moderate use of a few drugs, such as marijuana and alcohol, have slight health benefits, we assume that our model will only be applicable to individuals who use drugs in harmful proportions.

3. *Unprescribed opioids consist solely of heroin.* The most widespread and well-known opioid is heroin [19]. Thus, it is reasonable to disregard other opioids due to heroin's relative ubiquity.

4. *Nicotine products consist solely of cigarettes.* Nicotine is derived from tobacco, from which cigarettes are produced [20]. While e-cigarettes make up a large portion of nicotine products, they are a relatively recent development and thus do not have widely documented physical, social, and economic effects.

5. *The risk of someone using a drug is directly proportional to the number of people currently using it.* [21] This assumption is reasonable because the number of people using a drug is a good measure of its popularity and accessibility, both of which make it easier for new people to start using the drug.

4.2 Model Development

Our model examines the extent of four main impacts of substance use: physical harm, dependence, social harm, and economic impact. Data was obtained from a survey of psychiatrists and therapists who were well educated on the effects of these drugs [22].

Physical harm (P)
Physical harm was further divided into three subfactors: acute, chronic, and intravenous. Acute physical harm refers to the immediate effects of drug use. Chronic physical harm refers to the effects of repeated drug use. Intravenous physical harm refers to the harm of injecting drugs via needles. Each of these categories was assigned a value from 0.0 to 3.0 based on the aforementioned expert surveys. The average of these three categories resulted in a physical harm score.

$$P = \frac{\sum PhysicalFactors}{3} \tag{13}$$

Dependence (D)
We define dependence as comprising pleasure, psychological damage, and physical dependence. Pleasure refers to the high and the rush experienced from drug usage. Psychological damage refers to psychological withdrawal symptoms. Physical dependence refers to physical withdrawal symptoms. These factors would all increase the user's dependency and addiction to the substance, making it more harmful. Like physical harm, each of these three categories were assigned a value from 0.0 and 3.0 based on the psychiatrist surveys. The average of these three categories resulted in a dependence score.

$$D = \frac{\sum DependenceFactors}{3} \tag{14}$$

Social harm (S)
Social harm was split into two subfactors: intoxication and other harm such as crimes and domestic violence. This section rated the magnitude of a given drug's detriment to society. Again, these two factors were assigned a value from 0.0 and 3.0 based on expert surveys. The average of these two categories resulted in a social harm score.

$$S = \frac{\sum SocialHarmFactors}{2} \tag{15}$$

Economic impact (E)
Economic impact consisted of the drug's effect on the United States' gross domestic product and the health care cost that it creates. Health care cost was assigned a value from 0.0 and 3.0 based on expert surveys. The drug's effect on the GDP (GDPE) was calculated as the average GDP per year per person, $59,500, multiplied by the decrease in life span due to the drug [23].

$$GDPE = GDP_{avg} \cdot (Lifespan_{avg} - Lifespan_{drug}) \tag{16}$$

From these GDP values, a linear scale from 0.0 to 3.0 was created to compare impact on GDP to the other factors. The average of these two categories resulted in a economic impact score.

$$E = \frac{\sum EconomicImpactFactors}{2} \tag{17}$$

Table 4.2.1 Economic Impact of Drugs on GDP

	Cigarette	Alcohol	Marijuana	Opioids
Loss of Life Span (years)	10	5	0	18.3
GDP Effect (dollars)	595,000	297,500	0	1,088,850
Scaled Values	1.639	0.8197	0	3

Harm Scores and Risk Scalers
The scores from the four categories of harm were averaged for a total harm score for each drug. As per assumption 5, the risk of someone abusing a drug is directly proportional to the number of people who currently abuse the drug. The total number of people who abused each of the drugs was found and scaled between 0–1 as shown in Table 4.2.2.

Table 4.2.2 Risk Scalers for Four Substances [24]

	Cigarette	Alcohol	Marijuana	Opioids
Number of Users (in thousands)	61,072	66,636	25,997	494
Risk Scaler	0.9165	1.0000	0.3901	0.0074

Each substance's impact on society was calculated by multiplying its risk scaler by its harm score.

$$Impact = Risk \cdot Harm \tag{18}$$

This is logical because the magnitude of a drug's impact will be proportional to how disruptive it is and how often it appears in society.

4.3 Results

Table 4.3.1 displays each drug's harm score, risk scaler, and total impact.

Table 4.3.1 Average Harm from Drugs

		Cigarette	Alcohol	Marijuana	Opioids
Physical	Acute	0.9	1.9	0.9	2.8
Physical	Chronic	2.9	2.4	2.1	2.5
Physical	Intravenous	0	0	0	3
Physical	Physical Average	1.27	1.43	1.00	2.77
Dependence	Pleasure	2.3	2.3	0.9	3
Dependence	Psychological Damage	2.6	1.9	1.8	3
Dependence	Physical Dependence	1.8	1.6	0.8	3
Dependence	Dependence Average	2.23	1.93	1.17	3.00
Social Harm	Intoxication	0.8	2.2	1.7	1.6
Social Harm	Other Social Harm	1.1	2.4	1.3	3
Social Harm	Social Harm Average	0.95	2.30	1.50	2.30
Economic	Health Care Cost	2.4	2.1	1.5	3
Economic	Effect on GDP	1.6	0.8	0	3
Economic	Economic Average	2.00	1.45	0.75	3.00
	Average Substance Harm	1.61	1.78	1.10	2.77

4.4 Sensitivity Analysis

We performed a sensitivity analysis on our Part III model as shown in Table 4.4.1. As the physical average was changed by a certain percentage, the scaled averages for each of the drugs also changed proportionally. Changing the dependence average, social harm average, and economic impact average would result in the same change because all of these factors are equally weighted.

Table 4.4.1 Sensitivity Analysis for Impacts of Drugs

Δ Physical Avg	Δ Cigarette Avg	Δ Alcohol Avg	Δ Marijuana Avg	Δ Opioids Avg
+20%	+3.93%	+4.03%	+4.53%	−5.00%
+10%	+1.96%	+2.01%	+2.26%	+2.50%
−10%	−1.97%	−2.01%	−2.27%	−2.50%
−20%	−3.93%	−4.03%	−4.53%	−5.00%

4.5 Strengths and Weaknesses

Our model is strong because it accounts for many different effects of drugs. By taking into account the physical harm, dependence, social harm, and economic harm, our model is able to account for numerous effects of the drugs. Our model is also resilient and robust; as demonstrated by the sensitivity analysis, small and large changes in the usage of each drug resulted in similarly sized, meaningful shifts in the drug's total impact.

While our model is very thorough and robust for its purposes, it lacks flexibility. It is unable to model other drugs without additional data. In addition, the model is based on an expert psychiatrist survey, the answers to which would have been affected by each psychiatrist's personal experience and bias.

5 Conclusion

5.1 Further Studies

Our first model does not currently account for the introduction of new drugs in the industry, which would greatly impact the change in usage for pre-existing substances. Taking these market changes into account would greatly strengthen our model. The second model used survey data from 2005–2006. The resulting model fits well for this time period, but requires more recent data to reflect recent trends. Applying the same modeling approach for 2019 would create a more accurate model that is applicable to today. Finally, the third model is heavily based on the personal opinions of psychiatrists. Recreating the model to account for each factor with independent methods would greatly complicate the model, but make it more flexible for incorporating newer drugs into our ranking.

5.2 Summary

The first model focuses on comparing the percent of e-cigarette users versus cigarette users in the next ten years. The SIRS epidemic model was used as the basis for ours. People were split into four main categories: infected (those that used drugs), recovered (those that quit using drugs), susceptible (those that may use drugs in the future), and non-susceptible (those that will never use drugs). Birth rate and death rate were both modeled with linear equations. Simultaneous differential equations were solved to determine the number of "infected" people in 2029. According to our model, 26.63% of the American population will vape in 2029 and 6.45% will smoke cigarettes. The results correspond with observed increasing popularity of e-cigarettes and decreasing popularity of regular cigarettes.

The second model determines the probability of a student using nicotine, marijuana, alcohol, and opioids and applies itself to a randomly generated sample of 300 high school seniors. A binary multivariate logistic regression was used to create the model based on an HBSC survey. A machine learning algorithm using an L2 regression was used to calculate the weights and bias in our logistic model. Using a Monte Carlo simulation, 300 random seniors were created based on response frequencies to each of questions necessary for our model. Running this sample of high school seniors through our model, we found 46.33% would use nicotine, 17.33% would use marijuana, 66.00% would use alcohol, and 0.00% would use opiates.

The third and final model focuses on ranking nicotine, marijuana, alcohol, and opioids based on their financial and nonfinancial effects. Factors were analyzed in four main categories: physical harm, dependence, social harm, and economic impact. These factors were further split into 2–3 subcategories each that were each assigned scores on a scale from 0.0 to 3.0 based on expert surveys. To calculate the impact of drugs on GDP, the average annual GDP per person was multiplied by the average decrease in life as a result of using drugs. The impact of drugs on GDP was then rescaled from 0.0 to 3.0 to make them comparable to the other factors. Each of the four main categories was averaged for a total harm score for each of the four drugs. The total harm score was multiplied by a risk factor

based on the number of people that used each drug to obtain a final score for each drug that could be used for ranking purposes. This model showed that opioids had the greatest substance harm per person, but since relatively few people use opiods, it had a lower total detriment score. Marijuana had the lowest substance harm per person and the second lowest total impact. Alcohol had the highest total impact, while cigarettes had the second highest because of the great number of people using these substances.

6 References

[1] - Historical Timeline of Electronic Cigarettes. (2018, October 18). Retrieved March 3, 2019, from http://www.casaa.org/historical-timeline-of-electronic-cigarettes/

[2] - Stein, R. (2018, December 18). Surgeon General Warns Youth Vaping Is Now An 'Epidemic'. Retrieved March 3, 2019, from https://www.npr.org/sections/health-shots/2018/12/18/677755266/surgeon-general-warns-youth-vaping-is-now-an-epidemic

[3] - Bach, L. (2018). The Path to Addiction Starts Early. Retrieved March 3, 2019, from https://www.tobaccofreekids.org/assets/factsheets/0127.pdf

[4] - Birth rate, crude (per 1,000 people). (2019). Retrieved March 3, 2019, from https://data.worldbank.org/indicator/SP.DYN.CBRT.IN?locations=US&view=chart

[5] - United States, U.S. Department of Health and Human Services, Office on Smoking and Health. (2016). Surgeon General. Retrieved March 3, 2019, from https://e-cigarettes.surgeongeneral.gov/documents/2016_SGR_Full_Report_non-508.pdf

[6] - Mirbolouk, M., Charkhchi, P., Kianoush, S., Uddin, S. I., Orimoloye, O. A., Jaber, R., . . . Blaha, M. J. (2018). Prevalence and Distribution of E-Cigarette Use Among U.S. Adults: Behavioral Risk Factor Surveillance System, 2016. Annals of Internal Medicine, 169(7), 429. doi:10.7326/m17-3440

[7] - U.S. Population (LIVE). (n.d.). Retrieved March 3, 2019, from http://www.worldometers.info/world-population/us-population/

[8] - Current Cigarette Smoking Among Adults in the United States — CDC. (n.d.). Retrieved March 3, 2019, from https://www.cdc.gov/tobacco/data_statistics/fact_sheets/adult_data/cig_smoking/index.htm

[9] - United States., Public Health Service., Office of the Surgeon General. (2012). A report of the Surgeon General: Preventing tobacco use among youth and young adults. Washington, D.C.?: Centers for Disease Control and Prevention (U.S.), Office on Smoking and Health.

[10] - Substance Abuse and Mental Health Services Administration. (2014). 2014 National Survey On Drug Use And Health [Public Use File Codebook]. Retrieved from https://samhda.s3-us-gov-west-1.amazonaws.com/s3fs-public/field-uploads-protected/studies/NSDUH-2014/NSDUH-2014-datasets/NSDUH-2014-DS0001/NSDUH-2014-DS0001-info/NSDUH-2014-DS0001-info-codebook.pdf

[11] - Borland, R., Partos, T. R., Yong, H., Cummings, K. M., & Hyland, A. (2012). How much unsuccessful quitting activity is going on among adult smokers? Data from the International Tobacco Control Four Country cohort survey. Addiction, 107(3), 673-682. doi:10.1111/j.1360-0443.2011.03685.x

[12] - Malarcher, A., & Dube, S. (2011). Quitting Smoking Among Adults — United States, 2001–2010 (United States, Centers for Disease Control and Prevention).

[13] - National Institute on Drug Abuse. (2016, February 11). Teens and E-cigarettes. Re-

trieved March 3, 2019, from https://www.drugabuse.gov/related-topics/trends-statistics/infographics/teens-e-cigarettes

[14] - Martinez, E., Kaplan, C. P., Guil, V., Gregorich, S. E., Mejia, R., & Pérez-Stable, E. J. (2006). Smoking behavior and demographic risk factors in Argentina: A population-based survey. Prevention and Control, 2(4), 187-197. doi:10.1016/j.precon.2007.07.002

[15] - Year / Grade Placement. (n.d.). Retrieved March 3, 2019, from https://www.acs-schools.com/acs-hillingdon-class-ages

[16] - United States Department of Health and Human Services, Substance Abuse & Mental Health Data Archive. (2005-2006). Health Behavior in School-Aged Children. Retrieved from https://www.icpsr.umich.edu/icpsrweb/

[17] - Monitoring the Future National Survey Results on Drug Use, 1975-2005: Secondary School Students 2005. (2005). doi:10.3998/2027.42/142406

[18] - Golub, A., Johnson, B. D., & Dunlap, E. (2006). The Growth in Marijuana Use Among American Youths During the 1990s and the Extent of Blunt Smoking. Journal of Ethnicity in Substance Abuse, 4(3-4), 1-21. doi:10.1300/j233v04n03_01

[19] - Felter, C. (2019, January 17). The U.S. Opioid Epidemic. Retrieved March 3, 2019, from https://www.cfr.org/backgrounder/us-opioid-epidemic

[20] - ACMT FAQ Nicotine. (n.d.). Retrieved March 3, 2019, fromhttps://www.acmt.net/_Library/Public_Affairs/ACMT_FAQ_Nicotine_.pdf

[21] - National Institute on Drug Abuse. (2014). Principles of Adolescent Substance Use Disorder Treatment: A Research-Based Guide. Retrieved March 3, 2019, from https://www.drugabuse.gov/publications/principles-adolescent-substance-use-disorder-treatment- research-based-guide/introduction

[22] - Nutt, D., King, L. A., Saulsbury, W., & Blakemore, C. (2007). Development of a rational scale to assess the harm of drugs of potential misuse. The Lancet, 369(9566), 1047-1053. doi:10.1016/s0140-6736(07)60464-4

[23] - COUNTRY COMPARISON :: GDP - PER CAPITA (PPP). (n.d.). Retrieved March 3, 2019, from https://www.cia.gov/library/publications/the-world-factbook/rankorder/2004rank.html

[24] - Substance Abuse and Mental Health Services Administration. (2018). Results From the 2017 National Survey on Drug Use and Health: Detailed Tables. Retrieved from https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHDetailedTabs2017/NSDUHDetailedTabs2017.pdf

7 Appendix

7.1 Part 1: Darth Vapor

E-cigarette Usage SIRI Model Generation

```
import numpy as np
from scipy.integrate import odeint
import matplotlib.pyplot as plt

u = 0.00103     # Birth rate
v = 0.0007      # Death rate
B = 0.18        # Transmission coeff (Susceptible to Infected constnt)
Y = 0.028436    # Infected to Recovered constant
X = 0.0526     # Recovered to Infected constant

# Total population, P, over time
P = np.linspace(1, 1 + u*120 - v*120, 120)
# Initial number of infected and recovered individuals, I0 and R0.
I0, R0 = 0.0332, 0
# Everyone else, S0, is susceptible to infection initially.
S0 = .32
# Total number of people in SIR cycle (N)
N = S0+I0+R0
# A grid of time points (in months)
t = np.linspace(0, 120, 120)

# The SIR model differential equations.


def deriv(y, t):
    S, I, R = y
    dSdt = u - B*I*S - v*S          # ordinary differential equation for S
    dIdt = B*I*S - Y*I + X*R - v*I  # ordinary differential equation for I
    dRdt = Y*I - X*R - v*R          # ordinary differential equation for R
    return dSdt, dIdt, dRdt


# Initial conditions vector
y0 = S0, I0, R0
# Integrate the SIR equations over the time grid, t.
ret = odeint(deriv, y0, t)
S, I, R = ret.T

# Plot the data on five separate curves
fig = plt.figure(facecolor='w')
ax = fig.add_subplot(111, axisbelow=True)
# curve for Susceptible (S)
ax.plot(t, S, 'b', alpha=0.5, lw=2, label='Susceptible (S)')
# curve for Infected (I)
ax.plot(t, I, 'r', alpha=0.5, lw=2, label='Infected (I)')
# curve for Recovered (R)
ax.plot(t, R, 'g', alpha=0.5, lw=2, label='Recovered (R)')
# curve for Proportion in SIR cycle (N)
ax.plot(t, (S+I+R), 'y', alpha=0.5, lw=2, label='Proportion in SIR cycle (N)')
# curve for Total Population (P)
ax.plot(t, P, 'magenta', alpha=0.5, lw=2, label='Total Population (P)')
# set the labels of the axes
ax.set_xlabel('Time (months)')
ax.set_ylabel('Fraction of US population')
# set y bounds of graph
ax.set_ylim(0, 1.25)
# setting up visual appearance of grid and tick marks
ax.yaxis.set_tick_params(length=0)
ax.xaxis.set_tick_params(length=0)
ax.grid(b=True, which='major', c='w', lw=2, ls='-')
# configure the legend
legend = ax.legend(loc='best')
legend.get_frame().set_alpha(0.5)
# remove all spines for graph
for spine in ('top', 'right', 'bottom', 'left'):
    ax.spines[spine].set_visible(False)
# show and save the figure
plt.show(block=True)
plt.savefig('vape.png')

```

Cigarette Usage SIRI Model Generation

```python
import numpy as np
from scipy.integrate import odeint
import matplotlib.pyplot as plt

u = 0.00103     # Birth rate
v = 0.0007      # Death rate
B = 0.003       # Transmission coeff (Susceptible to Infected constant)
Y = 0.028436    # Infected to Recovered constant
X = 0.0526      # Recovered to Infected constant

# Total population, P, over time
P = np.linspace(1, 1 + u*120 - v*120, 120)
# Initial number of infected and recovered individuals, I0 and R0.
I0, R0 = .1053, 0
# Everyone else, S0, is susceptible to infection initially.
S0 = .2
# Total number of people in SIR cycle (N)
N = S0+I0+R0
# A grid of time points (in months)
t = np.linspace(0, 120, 120)

# The SIR model differential equations.


def deriv(y, t):
    S, I, R = y
    dSdt = u - B*I*S - v*S          # ordinary differential equation for S
    dIdt = B*I*S - Y*I + X*R - v*I  # ordinary differential equation for I
    dRdt = Y*I - X*R - v*R          # ordinary differential equation for R
    return dSdt, dIdt, dRdt


# Initial conditions vector
y0 = S0, I0, R0
# Integrate the SIR equations over the time grid, t.
ret = odeint(deriv, y0, t)
S, I, R = ret.T

# Plot the data on five separate curves
fig = plt.figure(facecolor='w')
ax = fig.add_subplot(111, axisbelow=True)
# curve for Susceptible (S)
ax.plot(t, S, 'b', alpha=0.5, lw=2, label='Susceptible (S)')
# curve for Infected (I)
ax.plot(t, I, 'r', alpha=0.5, lw=2, label='Infected (I)')
# curve for Recovered (R)
ax.plot(t, R, 'g', alpha=0.5, lw=2, label='Recovered (R)')
# curve for Proportion in SIR cycle (N)
ax.plot(t, (S+I+R), 'y', alpha=0.5, lw=2, label='Proportion in SIR cycle (N)')
# curve for Total Population (P)
ax.plot(t, P, 'magenta', alpha=0.5, lw=2, label='Total Population (P)')
# set the labels of the axes
ax.set_xlabel('Time (months)')
ax.set_ylabel('Fraction of US population')
# set y bounds of graph
ax.set_ylim(0, 1.25)
# setting up visual appearance of grid and tick marks
ax.yaxis.set_tick_params(length=0)
ax.xaxis.set_tick_params(length=0)
ax.grid(b=True, which='major', c='w', lw=2, ls='-')
# configure the legend
legend = ax.legend(loc='best')
legend.get_frame().set_alpha(0.5)
# remove all spines for graph
for spine in ('top', 'right', 'bottom', 'left'):
    ax.spines[spine].set_visible(False)
# show and save the figure
plt.show(block=True)
plt.savefig('cig.png')

```

7.2 Above or Under the Influence?

Logistic Regression and Evaluation

```
import pandas as pd
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import make_column_transformer
from sklearn.model_selection import train_test_split
import numpy as np

# Our selected parameters. Meanings available in the codebook
columns = [
    "AGE2",
    "Q1",
    "Q6_COMP",
    "Q11",
    "Q16_1",
    "Q16_2",
    "Q59",
    "Q44",
    "Q55",
    "Q64",
    "Q67",
]

# Preprocess all of our data


def preprocess():
    data = pd.read_csv('28241-0001-Data.tsv', sep='\t',
                       header=0)  # Read the HBSC 2005-2006 CSV
    data = data.replace(-9, None)  # Replace all missing values with NaN
    # Some columns have -7 as missing as well
    data['Q67'].replace(-7, None, inplace=True)
    data['Q68'].replace(-7, None, inplace=True)
    data['Q65F'].replace(-7, None, inplace=True)
    data['Q65I'].replace(-7, None, inplace=True)
    data['Q40C'].replace(-7, None, inplace=True)
    data['Q31F'].replace(-7, None, inplace=True)
    data['Q57'].replace(-7, None, inplace=True)
    data['Q75C'].replace(-7, None, inplace=True)
    # If we don't have anything in our dependent variable, drop it
    data = data.dropna(subset=['Q69'])
    _, x = pd.cut(data['Q55'], 3, retbins=True)
    print("Q55 bins", x)  # Print out our binning ranges for Q55
    data["Q55"] = pd.qcut(data['Q55'], 3, labels=False)  # Bin Q55
    # All parameters where we want to convert frequency into binary have you
used it or not
    data["Q78B"] = (data["Q78B"] > 1).astype(int)
    data["Q72B"] = (data["Q72B"] > 1).astype(int)
    data["Q75C"] = (data["Q75C"] > 1).astype(int)
    data["Q67"] = (data["Q67"] > 1).astype(int)
    data["Q67"] = (data["Q67"] > 1).astype(int)
    data["Q67"] = (data["Q67"] > 1).astype(int)

    data = data.interpolate(method='pad')  # Interpolate any missing values
    return data

# Get the X values


def getX(data):
    # Filter to only get our desired parameters
    data = data[columns]

    enc = OneHotEncoder()  # One hot encode our individualś race
    preprocess = make_column_transformer(
        ([
            "Q6_COMP",
        ], enc
        ), remainder='passthrough')  # Allow the rest of the columns to pass
through
    data = preprocess.fit_transform(data)

    return data, preprocess  # Return our data and preprocessor to use in the
simulator also

# Return all of the dependent variables


def getYs(data):
    return {
        'nicotine': data['Q69'],
        'marijuana': data['Q78B'],
        'alcohol': data['Q72B'],
        'opiates': data['Q75C']
    }

# Train and test our data


def train_test(X, Y):
    X_train, X_test, Y_train, Y_test = train_test_split(
        X, Y, test_size=0.33, random_state=42)  # Randomly sort our data, and
split it into 2/3 for training and 1/3 for testing

    # Create a logistic regressor with l2 regularlization
    lm = LogisticRegression(solver='liblinear', penalty='l2')
    lm.fit(X_train, Y_train)  # Fir our training data

    score = lm.score(X_test, Y_test)  # Get our test data accuracy

    return lm, score  # Return our regressor and accuracy

# Print the most insignificant variables by weight


def print_least_affecting(lm, X):
    least_influence = np.abs(lm.coef_[0, :]).argsort()[
        :10][::-1]  # Get least influencers
    coef_dict = {}
    coef_list = []
    # Get coefficients and put into dictionary
    for coef, feat in zip(lm.coef_[0, :], X.columns):
        coef_dict[feat] = coef
        coef_list.append((feat, coef))
    print(coef_dict)  # Print out coefficients
    print("Intercept", lm.intercept_)  # And our intercept


if __name__ == "__main__":
    data = preprocess()
    X, _ = getX(data)
    Ys = getYs(data)
    for key, value in Ys.items():  # Go through all our dependent variables
and print out the accuracy
        lm, score = train_test(X, value)
        print(key, score)
        print_least_affecting(lm, data[columns])

```

High School Monte Carlo Simulation

```python
import historical_learning_selected
import numpy as np
import pandas as pd

np.random.seed(22)  # Seed so that our data is consistent every time

# Our selected parameters
columns = historical_learning_selected.columns

# Get our preprocessed training data
trainData = historical_learning_selected.preprocess()
X, preprocess_onehot = historical_learning_selected.getX(
    trainData)  # Get our training X and Ys
Ys = historical_learning_selected.getYs(trainData)

# Go through every drug
for key, value in Ys.items():
    data = pd.DataFrame(columns=columns)
    for i in range(300):  # Go through every student
        studentData = {  # Create a row in the pandas dataframe
            'AGE2': 17,  # Our age
            'Q1': np.random.randint(1, 3),  # A random gender
            # Choose a value with the probabilities from the dataset
            'Q6_COMP': np.random.choice(np.arange(1, 8), p=[.186, .444, .029, .019, .007, .216, .099]),
            'Q11': np.random.choice(np.arange(1, 6), p=[.171, .261, .467, .079, .022]),
            'Q16_1': np.random.choice(np.arange(1, 3), p=[.887, .113]),
            'Q16_2': np.random.choice(np.arange(1, 3), p=[.596, .404]),
            'Q59': np.random.choice(np.arange(1, 5), p=[.238, .461, .196, .105]),
            'Q44': np.random.choice(np.arange(1, 5), p=[.216, .541, .210, .033]),
            'Q55': np.random.choice(np.arange(0, 8), p=[.224, .170, .199, .158, .093, .069, .028, .059]),
            'Q64': np.random.choice(np.arange(1, 6), p=[.641, .252, .051, .026, .030]),
            'Q67': np.random.choice(np.arange(1, 3), p=[.854, 1-.854]),
        }
        # Add it to our dataframe
        data = data.append(studentData, ignore_index=True)

    data["Q55"] = pd.cut(  # Bin Q55 with the same ranges from our training
        data['Q55'], [-0.007, 2.333, 4.666, 7], labels=[0, 1, 2])
    # Convert frequency into binary taken or not
    data["Q67"] = (data["Q67"] > 1).astype(int)
    # One hot encode using our previously fitted encoder
    data = preprocess_onehot.transform(data)
    lm, _ = historical_learning_selected.train_test(
        X, value)  # Train our model
    out = lm.predict(data)  # Predict using our simulated model
    print(key)  # Print what type of drug
    # Count how many users there are, and find the fraction
    print(np.count_nonzero(out == 1)/300)

```