Drug Use in the United States
Statistics-152
Matt Ronnau, Willis Silliman

Introduction

"All the advantages of Christianity and alcohol; none of their defects."
-Aldous Huxley, *Brave New World*

Aldous Huxley's description of the drug "Soma" in his dystopian novel *Brave New World* characterizes many individuals' reasons for deciding to take drugs. Throughout our lives, we are taught the dangers of drugs and alcohol, yet people still choose to abuse these substances every year. In college, drug and alcohol use is quite common, whether it be non-prescription stimulant use to prepare for exams or taking hard drugs to enjoy concerts and parties. We seek to analyze this drug use over a few different demographic identifiers to see which segments of the population are using what drugs. We also seek to find out if cigarette use is in fact linked to many leading health problems, as is taught to students from a very young age.

Description of Questions

The goal of our project is to analyze differences in drug usage as well as health disorders based on different demographic factors, such as age, geographic location, and prior drug use history. We will focus in on a few key areas: stimulant usage amongst college aged individuals compared with others; opioid usage amongst individuals in the Midwest of the United States compared with other geographic locations; hard drug usage amongst individuals who have smoked cigarettes and marijuana compared to those who haven't; and health problems amongst smokers and non-smokers of cigarettes.

Our motivation for analyzing stimulant (Ritalin) usage amongst college aged students compared with others lies in the fact that we ourselves are college students. There is immense pressure to succeed on exams and papers, and many students use non-prescription stimulants such as Ritalin and Adderall in order to focus in preparation for their tests. Because of this societal pressure and college culture, we wish to see if this kind of stimulant use is higher amongst college aged individuals than other age groups.

The United States is currently facing an opioid epidemic, and the Rust Belt is one of the regions of the country that has been the hardest hit. We want to see if opioid use is larger in this region compared to others.

We are often taught in school to stay away from drugs. Anti-drug campaigns teach students that

using drugs is a slippery slope, and that the probability of using other drugs is higher once someone has smoked cigarettes or marijuana. We want to see if this is in fact the case, by comparing the usage of hard drugs (cocaine, LSD, ecstasy, heroin, and morphine) amongst people who have tried marijuana compared to those who haven't, and amongst people who have used cigarettes compared to those who haven't.

Finally, we want to see if the rate of common health issues is higher amongst smokers compared with non-smokers. According to the CDC, "smoking is the leading cause of preventable death." Studies have linked smoking to many health issues, and we wish to see if those trends are similar in our data.

Description of the Survey

About the US Department of Human and Human Services

The US Department of Health is a cabinet-level department of the executive branch. The department was first established in 1939 as the Federal Security Agency, and renamed to the HHS in 1979. In 1971, the department established the National Survey on Drug Use and Health (NSDUH) as a way to track the effects of drug use on health. The survey is conducted yearly, surveying the effects of drugs on mental health, blood pressure, asthma, and other health-related issues (NSDUH, 2020). Many federal departments, such as the USDE, the USDOT, and the CDC use the information gathered in this survey to estimate the number of people in the public using each drug, and the health conditions that may be caused by them.

NSDUH and their subsidiary the National Institute on Drug Abuse (NIDA) aim to cast a wide net on our population, gathering as much data as possible. This year, a reported 70,000 people will be interviewed for the survey. This includes members from each geographic region, and peoples of all ethnic backgrounds in the US. The survey covers all peoples ages 12 or older. In this 1990 survey, there was a survey portion, and an interview portion. The survey portion consisted mainly of demographic questions, with some initial questions about drug use. The interview asked participants more detailed questions about their history with drug use and health issues. The NSDUH data sets are released yearly, after a two year grace period, to the public, though without identifiers. The 1990 set was posted with identifiers in 2013 (ICPSR, 2013).

Design Elements

The NSDUH survey used a five stage survey design. More information about their design, variables, and questionnaire design can be found at: https://datafiles.samhsa.gov/study-dataset/national-household-survey-drug-abuse-1990-nhsda-1990-ds0001-nid13795

The survey's target population is the entire population of the United States. To reduce the variance of small groups in the US, the survey used oversampling and poststratification to better

weight these responses. These include Asians, American Indian/ Alaskan Natives, Black, low income whites, younger people, and less educated people. Their sampling estimations relied on data from the 1980 census to sample certain areas.

The survey had six levels with three sampling levels. The survey first designated several regions of the US which form strata. These regions were designated as: New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, and Pacific. These strata were chosen, as they are very different regions of the country, but are fairly self-similar. These strata are of very different sizes, so the next level has more stratum defined for larger populations.

The second level is a division of these regions. These are encrypted in the data set, to protect the identities of the survey subjects, however we know they were chosen geographically, and that each had roughly 100 to 150 PSUs selected within themselves. These PSUs are also encrypted, so we only know that they also represent geographical regions. These PSUs were sampled with probability according to size.

From these PSUs, segments form a final strata. These segments were usually city blocks, though occasionally larger areas were used in smaller cities. These aimed to have a similar number of households within each to sample from.

Finally, each household acted as a PSU from which an individual would be drawn to take the survey. The sampling levels can be found below in a table format

Stage	Strata or Sampling Unit	Sampling Method	Criterion
1	Divisions	-	Geographic
2	Geographic Region (Stratum)	-	Geographic
3	Unnamed Geographic PSU (PSU)	Probability Proportional to Size	Geographic
4	Segments (Blocks, or other defined region)	-	Geographic
5	Household	Equal Probability across all households in segment	Geographic
6	Individual	Equal Probability for those aged 12 or older	

Table 1: *Survey Methodology.*

Description of the Data

Public Data

To protect individual's privacy, NSDUH encrypts their stratum and PSUs, not distinguishing the criterion to create the high levels. The necessary data to perform a survey analysis is then given with weights to help analyse the data without giving away information about the individual's place of residence. Regions are provided, however, to assist in comparison.

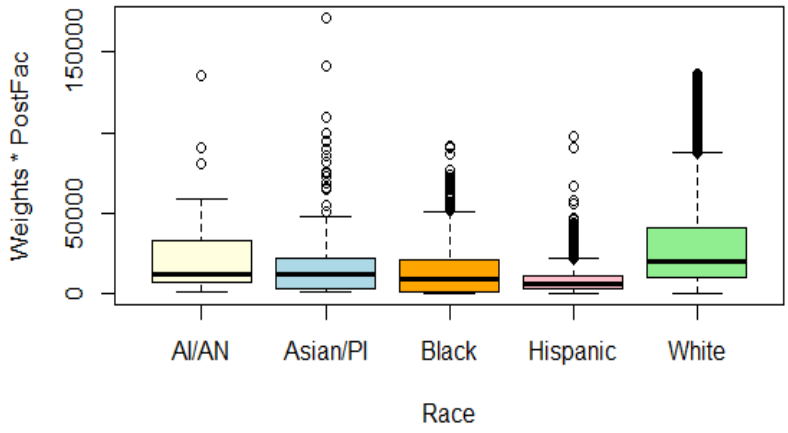
NSDUH provides two different weight variables, of which we have chosen to use the weights per person, and not the scaled weights that count per thousand. NSDUH also provides a poststratification factor to scale the weights by, in order to better represent undersampled subpopulations.

The statisticians at NSDUH have imputed some variables for the public, both logically filled in, and via hot-deck and cold-deck imputation from years past. In order to better estimate values and reduce variance among smaller groups, we have imputed some missing data using the hot-deck method. We have also cleaned and decoded the data to better visualize the data in this project.

Exploration

Our sample weights for our data set range from 70 to 171,282. The lowest is a 16 year old white male from the South, which supports our poststratification, since this group was sampled more than other groups. Our maximum was a 27 year old Asian/ Pacific Islander from the Northeast, so these people were undersampled initially. A weight vs race boxplot can be seen below in Figure 1. There are many outliers in each boxplot, so there must be another factor that also contributes to the weights not seen in this plot.

Figure 1: Weights by Subject Race

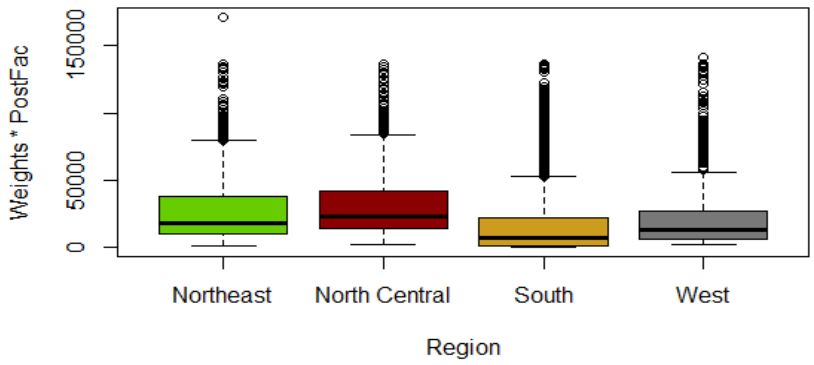


The weights for sex were very similar, as we expected. Others showed similar trends, such as education level weighting higher education people more. The age division showed that older

people were weighted more heavily.

The weights for the region, which contain the divisions, are shown in Figure 2. below. They show that the North Central and North East was under-represented in the sample. The West and South are large areas with lots of people, so they were oversampled, thus their weights are smaller. We can also see the highest and lowest weights in their respective regions.

Figure 2: Weights by Region



Methodology

The data was obtained from the ICPSR website for the 1990 NSDUH report on drug abuse, which can be found at:

https://www.icpsr.umich.edu/web/ICPSR/studies/9833/summary?fbclid=IwAR30t8KImNd5QorUaLWMCtsQydUTweSdFoIJPyNR92hXwBHy9XKtiCnSg1Y

The data from the SAS data was in a .tsv file, which can be read in as a data table in R via the function read_tsv() in the readr package. We then created a new frame with the variables we were interested in, renamed and re-coded to be more informative to the reader. We then imputed the data, and recleaned it to give NAs to those individuals who did not report using a certain drug.

The items were re-coded to be more interpretable, for example, we re-encoded the "WorkStatus" vector to have levels: "Full-Time", "Part-Time", "Away from Work", "Unemployed looking for Work", etc. instead of the original vector, which had entries ranging from 1 to 11.

The item nonreponse was not too high in this survey, and the NSDUH indicates percentages of nonrepsonse in their codebook. This is a sensitive subject, however, and some people did not respond to the questions in the survey. Still, many eagerly said they didn't use drugs, so the

highest nonreponse rate was only around .33%. Still, to fill in this missing data, we used the function hotdeck() in the VIM package, which will do a hot-deck imputation on the desired columns based on given ordinal variables. To fill these in, we used region, demographic, age and education ranges, work status, family income, and insurance as the variables to group by. The age ranges were: minors (< 18), 18 to 29, 30 to 39, 40 to 49, 50 to 59, 60 to 69, and 70 and above. The education levels were split by levels of partially completing elementary, junior high, highschool, or college, and levels in which each was completed.

Once the data was imputed and cleaned, we created a survey design object, using the survey package, with the following line of code:

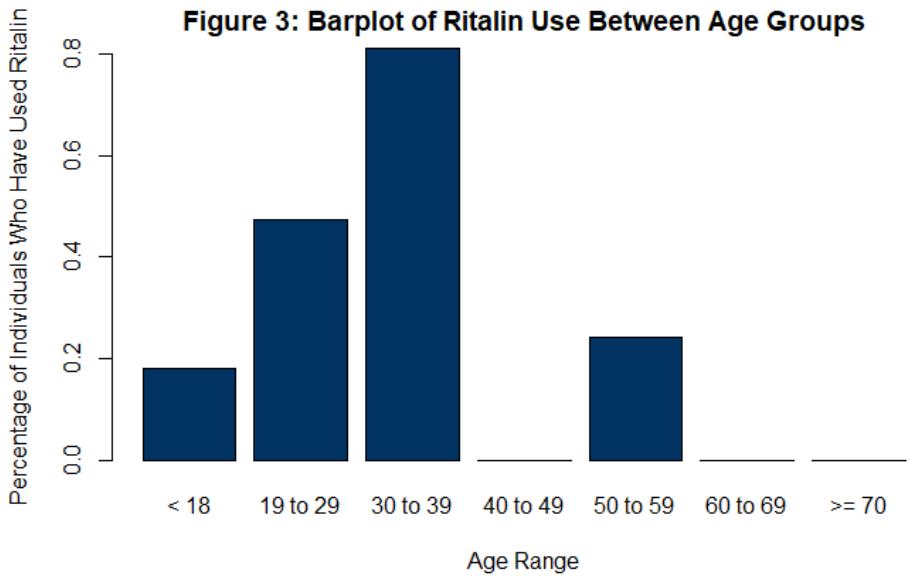
design <- svydesign(ids = ~ PSU + Household, strata = ~ Division + Stratum + Segment, weights = ~ Weights*PostFac, data = imputed_data, nest = T)

The PSUs, and stratums were given, so our estimates should be accurate to the original data analysis done by NSDUH. The also supplied weights and poststratification factors, so our variance estimates should also be accurate.

Analysis of Data

All of the data calculations that we carried out were on proportions; we scaled both the mean and the standard error by 100 to reflect percentages out of 100%.

We first analyzed Ritalin usage amongst the different age groups. We calculated the percentage of individuals in each age class who used Ritalin, along with a standard error of the estimate. The results are shown in Figure 3 and Table 2. We were initially searching to see if Ritalin use was largest amongst college aged students, or those that fell in the 19-29 age group class. However, the results of our analysis shows that age group 30-39 exhibited the highest percentage of individuals who used Ritalin, those 19-29 was the second largest.



Age Group	Percentage of Ritalin Users	Standard Error	95% Confidence Interval
Less than 18	0.001804094	0.001229389	(0, 0.4213652)
19 to 29	0.004724320	0.002138125	(0.05336726, 0.8914968)
30 to 39	0.008117111	0.003024876	(0.21884628, 1.4045759)
40 to 49	0.000000000	0.000000000	(0, 0)
50 to 59	0.002423650	0.001793551	(0, 0.5938945)
60 to 69	0.000000000	0.000000000	(0, 0)
Over 70	0.000000000	0.000000000	(0, 0)

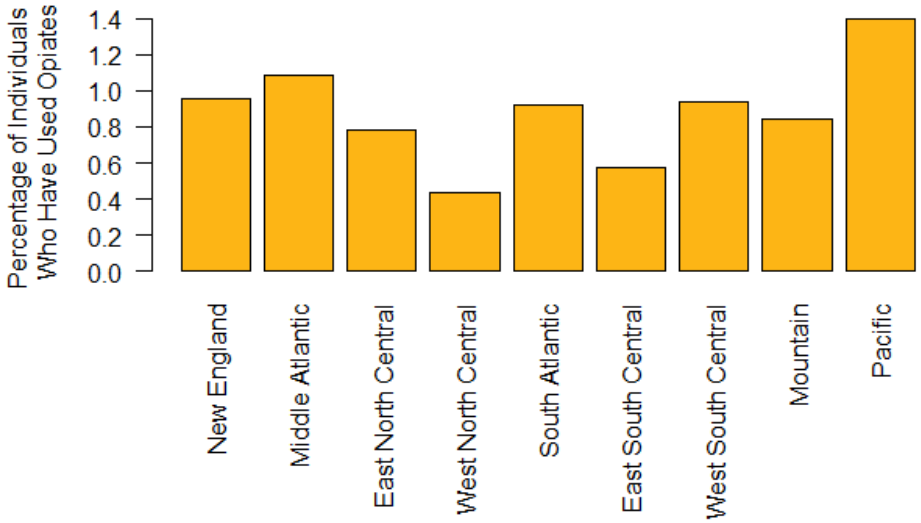
Table 2: *Percentage of Ritalin Users by Age Group.*

We next sought to analyze opiate usage across different geographic regions within the United States. For this analysis, we used the variable “Division”, as it broke the “Region” variable up a bit for a better look at the different regions of the country.

We made a vector of 1s and 0s that combined data from Morphine, Cocaine, Heroin, LSD, and Ecstasy usage to indicate if an individual had ever used any of these five drugs; we took the mean of this vector, called “HardDrugs”, in each division to indicate the proportion of individuals who have used a hard drug. The results are in Figure 4 and Table 3.

It is clear from the barplot that the midwest does not have the largest percentage of opioid users (we are taking East North Central and West North Central as the areas that define the Midwest). In fact, the division defined as the "Pacific" has the largest percentage of opioid users. This is surprising given the population of the Pacific coast. Though the inland parts of the west coast are rural, we thought the cities with large populations would bring down the percentage of users.

Figure 4: Barplot of Opiate Usage Between Geographic Region



Geographic Region	Percentage of Opioid Users	Standard Error	95% Confidence Interval
New England	0.9598709	0.3160123	(0.34049818, 1.5792436)
Middle Atlantic	1.0891996	0.3653772	(0.37307353, 1.8053257)
East North Central	0.7865973	0.2331267	(0.32967723, 1.2435173)
West North Central	0.4327003	0.1913505	(0.05766021, 0.8077403)
South Atlantic	0.9251822	0.2858638	(0.36489943, 1.4854649)

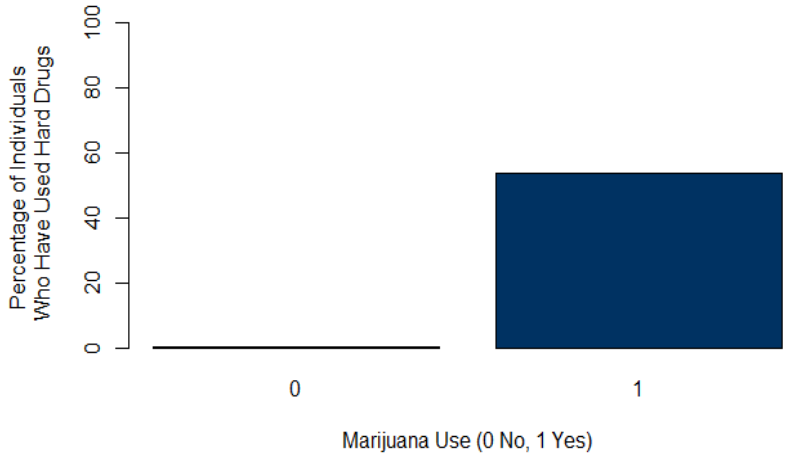
East South Central	0.5762691	0.2374614	(0.11085331, 1.0416850)
West South Central	0.9352729	0.3583785	(0.23286393, 1.6376819)
Mountain	0.8412866	0.2875202	(0.27775743, 1.4048158)
Pacific	1.4034291	0.3866951	(0.64552069, 2.1613375)

Table 3: *Percentage of Opioid Users by Geographic Region.*

We used our vector "HardDrugs" for our next analysis, in which we compared hard drug use amongst people who have smoked marijuana and people who have not, as well as amongst people who have smoked cigarettes and those that have not. The results of our comparison against marijuana can be seen in Figure 5 and Table 4.

It is evident that amongst individuals who have smoked marijuana, more individuals have used hard drugs than amongst individuals who have not smoked marijuana. However, this is purely observational data, and any sort of causal inference between the two variables cannot be deduced from this analysis. Nevertheless, there is a stark difference in the percentage of hard drug users amongst people who have smoked marijuana compared with those who haven't. This can be due to the fact that people who are more likely to do hard drugs are in turn more likely to use "softer" drugs, such as marijuana.

Figure 5: Barplot of Hard Drug Use Between Marijuana Users

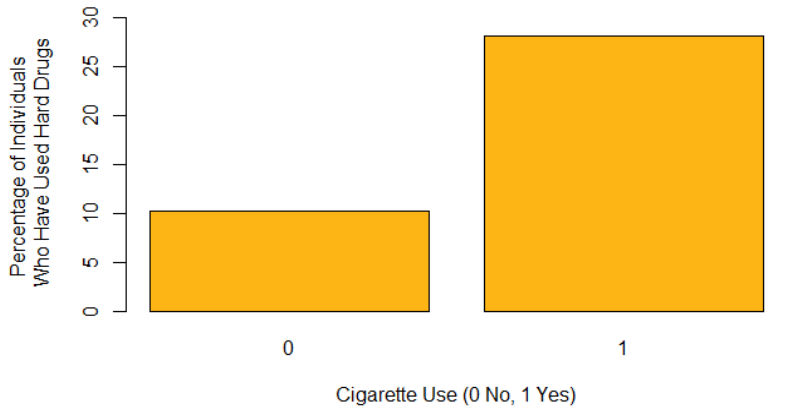


Marijuana Use	Percentage of Hard Drug Users	Standard Error	95% Confidence Interval
No	0.7232273	0.1744608	(0.3812903, 1.065164)
Yes	53.8055847	2.2221215	(49.4503066, 58.160863)

Table 4: *Percentage of Hard Drug Users by whether an individual has smoked marijuana.*

The analysis of hard drug usage amongst people who have and have not smoked marijuana holds true for individuals who have and have not smoked cigarettes. Again, this correlation is strictly observational, but it is interesting to note the difference in percentages amongst those who have used cigarettes and those who haven't.

Figure 6: Barplot of Hard Drug Use Between Cigarette Users

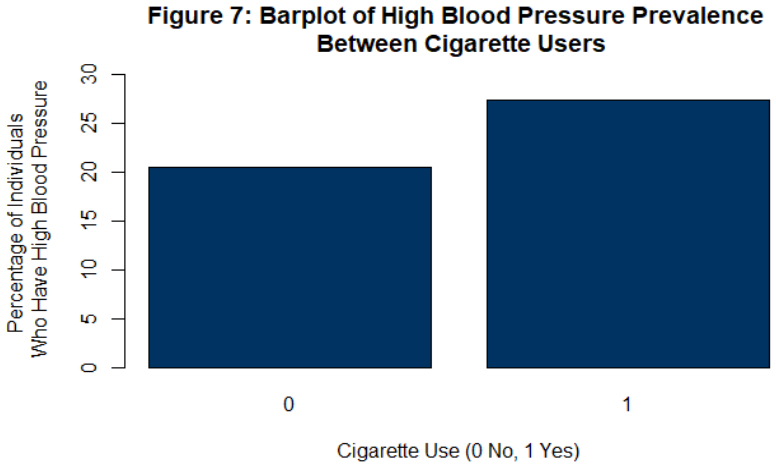


Cigarette Use	Percentage of Hard Drug Users	Standard Error	95% Confidence Interval
No	10.28890	0.8882101	(8.548037, 12.02976)
Yes	28.10971	1.5822424	(25.008567, 31.21084)

Table 5: *Percentage of Hard Drug Users by whether an individual has smoked cigarettes.*

The last of our analysis centered around the prevalence of certain health issues (high blood pressure, asthma, psychological/ emotional disorders, and heart disease) amongst individuals who have and have not smoked.

The first analysis of health problems looked at the percentage of individuals who had high blood pressure, and the results are displayed in Figure 7 and Table 6. As can be seen by the non overlapping confidence intervals in Table 6, there does appear to be a difference between the percentage of individuals who have high blood pressure and who have smoked and the percentage of individuals who have high blood pressure and have not smoked. However, like above, these results are observational, and we cannot conclude that it was because of the smoking habit that these individuals developed their high blood pressure.

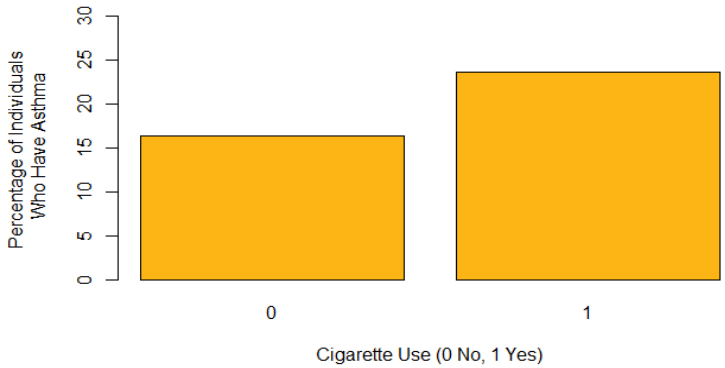


Cigarette Use	Percentage of Individuals With High Blood Pressure	Standard Error	95% Confidence Interval
No	20.48294	0.9261517	(18.66772, 22.29816)
Yes	27.46285	1.0061261	(25.49088, 29.43482)

Table 6: *Percentage of individuals with high blood pressure by whether an individual has smoked cigarettes.*

Our second health analysis compared the percentage of individuals who had asthma between people who have smoked cigarettes and people who have not smoked cigarettes. The results are listed below in Figure 8 and Table 7. Like the results of our analysis on the percentage of individuals who had high blood pressure, the confidence intervals for the prevalence of asthma amongst smokers and non-smokers do not show overlap, indicating that there is a difference in the rates of asthma prevalence between individuals who have smoked cigarettes and individuals who have not.

Figure 8: Barplot of Asthma Prevalence Between Cigarette Users

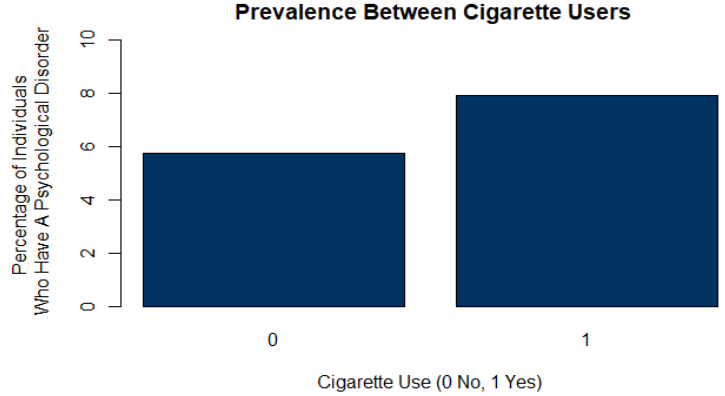


Cigarette Use	Percentage of Individuals With Asthma	Standard Error	95% Confidence Interval
No	16.40044	0.8793951	(14.67686, 18.12403)
Yes	23.65967	1.2877217	(21.13578, 26.18356)

Table 7: *Percentage of individuals with asthma by whether an individual has smoked cigarettes.*

We next looked at the prevalence of psychological disorders amongst individuals who have and have not smoked cigarettes. The results are summarized in Figure 9 and Table 8. The confidence interval for the estimated percentage of individuals who had a psychological disorder and smoked cigarettes did overlap with the confidence interval for the estimated percentage of individuals who had a psychological disorder and did not smoke. Thus, we cannot conclude that there is a difference in these two groups.

Figure 9: Barplot of Psychological Disorder Prevalence Between Cigarette Users

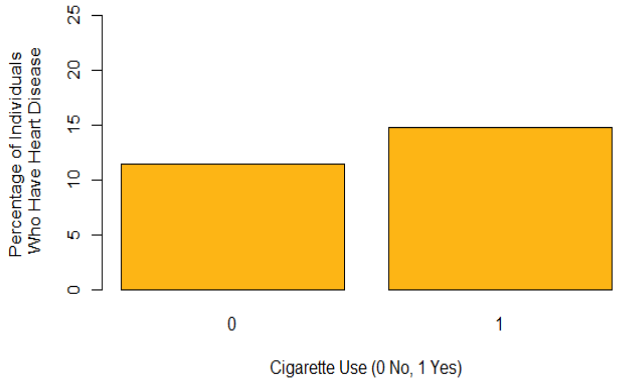


Cigarette Use	Percentage of Individuals With A Psychological Disorder	Standard Error	95% Confidence Interval
No	5.760469	0.5424684	(4.697250, 6.823688)
Yes	7.942871	0.6528159	(6.663376, 9.222367)

Table 8: *Percentage of individuals with a psychological disorder by whether an individual has smoked cigarettes.*

Finally, we analyzed the prevalence of heart disease between cigarette smokers and non cigarette smokers; the results are displayed in Figure 10 and Table 9. Surprisingly, the confidence intervals for each segment do overlap. Because of this, we cannot draw a conclusion about the difference in heart disease prevalence amongst smokers compared with non-smokers.

Figure 10: Barplot of Heart Disease Prevalence Between Cigarette Users



Cigarette Use	Percentage of Individuals With Heart Disease	Standard Error	95% Confidence Interval
No	11.47029	0.8645991	(9.775712, 13.16488)
Yes	14.82575	0.9392891	(12.984778, 16.66672)

Table 9: *Percentage of individuals with heart disease by whether an individual has smoked cigarettes.*

Conclusion

Some of our assumptions going into the project turned out to be incorrect, based on our analysis. Our preconceived notions regarding stimulant and opioid use turned out to be incorrect; we found that college aged individuals did not exhibit the largest percentage of stimulant use, and we found that the Midwest did not exhibit the largest percentage of opioid use. However, this could be in part due to the age of the data we are working with – it is from 1990. Societal pressures and trends have changed in the past 30 years, so a more recent data set may have produced different results. We had also hoped to study Adderall specifically for this purpose, but it was not included as a drug in the 1990 survey.

Regardless of the relevance of the data, we must be careful in how we interpret our results. The data that we analyzed was from a study in which an interviewer asked selected individuals questions about their lifestyle habits. Because of this, all of our analysis must be viewed through the lens of an observational study; no causal inference can be made. However, we can point to certain associations that appear in the data.

Some results we found could likely be due to correlation with other confounding variables, which we could not control for in our data. For example, high blood pressure could be caused by a number of factors, including stress levels and dietary habits. These can each be correlated with smoking, which could be why smokers exhibited a greater proportion of individuals that had high blood pressure than non smokers.

Amongst age groups that did have Ritalin users, we did not find any statistical significance in the percentage of users. All confidence intervals for age groups that showed usage overlapped with each other, indicating that there is not a statistically significant difference in the mean percentage of users.

Likewise, there does not appear to be a difference in opioid use amongst different regions of the United States. All of the confidence intervals produced overlapped with each other, so we could not conclude that any one region exhibited more use than another.

However, our analysis of hard drug use amongst individuals who have smoked marijuana compared with individuals who have not smoked marijuana *did* indicate statistical significance, as did our results of hard drug use amongst individuals who have smoked cigarettes compared with individuals who have not smoked cigarettes. In both cases, individuals who had smoked, whether it be marijuana or cigarettes, showed significantly higher rates of hard drug use than their non-smoking counterparts. We cannot conclude that these drugs are gateways, however, due to the observational status of our data. Nonetheless, the association is still starkly apparent.

High blood pressure and asthma prevalence amongst cigarette smokers also showed statistical significance, while the prevalence of psychological disorders and heart disease did not. These of course can not be causally proved, but there does appear to be a strong association between smoking and the development of high blood pressure or asthma.

It is not surprising that there does not appear to be an association between psychological disorders and smoking, as cigarettes are not a mind altering substance; we also do not expect individuals who have been diagnosed with a psychological disorder to tend to smoke more than those who don't.

The lack of an association between heart disease and cigarette use is surprising, however. The CDC states that smoking *causes* heart disease, but our results do not provide the same conclusion. This could be in part, though, to an individual's lack of knowledge about any heart issues they may have. If a doctor has never told them that they had heart disease, then they would have responded "No" to the survey when asked if they had heart disease; better results could be garnered by having a doctor complete an examination of each individual in conjunction with the survey, but this would be costly and time consuming.

In conclusion, we found strong correlations between "gateway drugs" and hard drug use, as well as between cigarettes and known health problems. We also discovered that stimulant use amongst college aged individuals does not appear to be statistically different from individuals of a different age, perhaps pointing to an over exaggeration of the issue. No matter the results we drew upon here, drug use is a damaging habit, and it is unfortunately a growing problem in the United States. It is comforting to know, though, that there are many individuals and organizations throughout the United States who are fighting back against the dangers of addiction.

Citations

(NSDUH, 2020) "Welcome to the National Survey on Drug Use and Health (NSDUH)." *National Survey on Drug Use and Health*, nsduhweb.rti.org/respweb/homepage.cfm.

(ICPSR, 2013) United States Department of Health and Human Services, et al. "National Household Survey on Drug Abuse, 1990." *National Household Survey on Drug Abuse, 1990*, Inter-University Consortium for Political and Social Research [Distributor], 6 May 2013, www.icpsr.umich.edu/web/ICPSR/studies/9833/summary?fbclid=IwAR30t8KImNd5QorUaLWMCtsQydUTweSdFoIJPyNR92hXwBHy9XKtiCnSg1Y.

(CDC, 2019) "Fast Facts." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 15 Nov. 2019, www.cdc.gov/tobacco/data_statistics/fact_sheets/fast_facts/index.htm.

Appendix - R Code

The data must first be downloaded from the ICPSR website in SAS form. The link can be found in the citations above (ICPSR, 2013)

```
# Loading Required Packages:
library(readr)
library(survey)
library(VIM)

# Read in the Data
data <- read_tsv("09833-0001-Data.tsv")

# Creating the New Data Frame and renaming variables that don't have
missing values
new_data <- data.frame(Stratum = data$VESTR,
                       PSU = data$ENCPSU,
                       Segment = data$ENCSEG,
                       Household = data$ENCCASE,
                       Weights = data$ANALWT,
                       PostFac = data$POSTFAC,
                       Region = factor(data$REGION, levels = 1:4,
labels = c("Northeast", "North Central", "South", "West")),
                       Division = factor(data$DIVISION, levels = 1:9,
labels = c("New England", "Middle Atlantic", "East North Central",
"West North Central", "South Atlantic", "East South Central", "West
South Central", "Mountain", "Pacific")),
                       Age = data$IRAGE,
                       Sex = factor(data$IRSEX, levels = 1:2, labels =
c("Male", "Female")),
                       Race = factor(data$IRRACE, levels = 1:4, labels
= c("AI/AN", "Asian/PI", "Black", "White")),
                       Hispanic = factor(data$IRHOIND, levels = 1:2,
labels = c("Hispanic", "Not Hispanic")),
                       Education = data$IREDUC,
                       WorkStatus = factor(data$IRWORKST, levels =
1:11, labels = c("Full-Time", "Part-Time", "Away from work",
"Unemployed looking for work", "Unemployed not looking for work",
"Full-Time Homemaker Only", "In School Only", "Retired", "Disabled,
```

```
not able to work", "Other, in labor force", "Other, not in labor
force")),
                      Insurance = c(1, 0)[data$IRINSUR],
                      FamilyIncome = factor(data$IRINC47, levels =
c(1:13, 99), labels = c("No personal earnings", "< $5000", "$5000 to
$6999", "$7000 to $8999", "$9000 to $11999", "$12000 to $14999",
"$15000 to $19999", "$20000 to $24999", "$25000 to $29999", "$30000
to $39999", "$40000 to $49999", "$50000 to $74999", "> $75000",
"Legitimate Skip")),
                      HighBP = factor(data$HIGHBP, levels = c(1:2,
96:98), labels = c("Yes", "No", "Multiple Responses", "Refused",
"Blank")),
                      Asthma = factor(data$ASTHMA, levels = c(1:2,
92, 94, 98), labels = c("Yes", "No", "Illegible", "Don't Know",
"Blank")),
                      Psych = factor(data$TRPSYCH, levels = c(1:2,
92, 94, 98), labels = c("Yes", "No", "Illegible", "Don't Know",
"Blank")),
                      Heart = factor(data$HEART, levels = c(1:2,
94:96, 98), labels = c("Yes", "No", "Don't Know", "Bad Data",
"Multiple Responses", "Blank")),
                      CigTry = data$CIGTRY,
                      CigAge = data$CIGAGE,
                      CigYears = data$CIGYRS,
                      AlcTry = data$ALCTRY,
                      AlcAge = data$ALCAGE,
                      Xanax = data$XANAX,
                      Ritalin = data$RITALIN,
                      Morphine = data$MORPHINE,
                      MJAge = data$MJAGE,
                      CocAge = data$COCAGE,
                      LSD = data$LSD,
                      Ecstasy = data$ECSTASY,
                      HeroinAge = data$HERAGE
                      )

# Defining Transformation Functions for Data

#### Needed for this data set, since most values are encoded
```

```r
# Grouping Nonreponse and special variables
age_transformer <- function(x) {
  for (i in 1:length(x)) {
      if (x[i] %in% c(185, 192, 194:198)) {
      x[i] = NA
      }
      else if (x[i] %in% c(181, 191, 193, 199)) {
      x[i] = 100
      }
  }
  return(x)
}

# Transforming Drug Variables
drug_transform <- function(x) {
  for (i in 1:length(x)) {
      if (x[i] %in% c(2, 81, 91)) {
      x[i] = 0
      }
      else if (x[i] %in% c(94, 98)) {
      x[i] = NA
      }
  }
  return(x)
}

# Creating Ranged Variables
age_create <- function(x) {
  for (i in 1:length(x)) {
      if (x[i] < 18) {
      x[i] = 1
      }
      else if (x[i] < 30) {
      x[i] = 2
      }
      else if (x[i] < 40) {
      x[i] = 3
      }
      else if (x[i] < 50) {
```

```
      x[i] = 4
      }
      else if (x[i] < 60) {
      x[i] = 5
      }
      else if (x[i] < 70) {
      x[i] = 6
      }
      else {
      x[i] = 7
      }
  }
  x = factor(x, levels = 1:7, labels = c("< 18", "19 to 29", "30 to
39", "40 to 49", "50 to 59", "60 to 69", ">= 70"))
  return(x)
}

# Adding Hispanic to Race
race <- function(x, y) {
  x = as.character(x)
  for (i in 1:length(x)) {
      if (y[i] == "Hispanic") {
      x[i] = "Hispanic"
      }
  }
  return(x)
}

# Grouping by education
group_education <- function(x){
  for(i in 1:length(x)){
      if(x[i] < 6){
      x[i] = 1
      }
      else if(x[i] < 9){
      x[i] = 2
      }
      else if(x[i] < 12){
      x[i]  = 3
```

```
      }
      else if(x[i] == 12){
      x[i] = 4
      }
      else if(x[i] < 16){
      x[i] = 5
      }
      else if(x[i] == 16){
      x[i] = 6
      }
      else if(x[i] > 16){
      x[i] = 7
      }
  }
  x <- factor(x, levels = 1:7, labels = c("Some Elementary", "Some
Middle", "Some High School", "High School", "Some College",
"Bachelors", "Some Graduate"))
  return(x)
}

# Converting to Binary Response
new_data$HighBP <- c(1, 0, NA, NA, NA)[new_data$HighBP]
new_data$Asthma <- c(1, 0, NA, NA, NA)[new_data$Asthma]
new_data$Psych <- c(1, 0, NA, NA, NA)[new_data$Psych]
new_data$Heart <- c(1, 0, NA, NA, NA)[new_data$Heart]

# Transforming Drugs
new_data$CigTry <- age_transformer(new_data$CigTry)
new_data$CigAge <- age_transformer(new_data$CigAge)
new_data$CigYears <- age_transformer(new_data$CigYears)
new_data$AlcTry <- age_transformer(new_data$AlcTry)
new_data$AlcAge <- age_transformer(new_data$AlcAge)
new_data$Xanax <- drug_transform(new_data$Xanax)
new_data$Ritalin <- drug_transform(new_data$Ritalin)
new_data$Morphine <- drug_transform(new_data$Morphine)
new_data$MJAge <- age_transformer(new_data$MJAge)
new_data$CocAge <- age_transformer(new_data$CocAge)
new_data$LSD <- drug_transform(new_data$LSD)
new_data$Ecstasy <- drug_transform(new_data$Ecstasy)
```

```r
new_data$HeroinAge <- age_transformer(new_data$HeroinAge)

# Creating New Columns
new_data$AgeRange <- age_create(new_data$Age)
new_data$Race <- race(new_data$Race, new_data$Hispanic)
new_data$Race <- as.factor(new_data$Race)
new_data$EduLevel <- group_education(new_data$Education)

# Hot-Deck Imputation
set.seed(251)

imputed_data <- hotdeck(new_data, variable =
colnames(new_data)[17:33], ord_var = c("Region", "AgeRange", "Sex",
"Race", "EduLevel", "WorkStatus", "Insurance", "FamilyIncome"))

# Recoding "Never Used" values to NAs
recode100 <- function(x) {
  for (i in 1:length(x)) {
      if (x[i] == 100) {
      x[i] <- NA
      }
  }
  return(x)
}

imputed_data$CigTry <- recode100(imputed_data$CigTry)
imputed_data$CigAge <- recode100(imputed_data$CigAge)
imputed_data$CigYears <- recode100(imputed_data$CigYears)
imputed_data$AlcTry <- recode100(imputed_data$AlcTry)
imputed_data$AlcAge <- recode100(imputed_data$AlcAge)
imputed_data$MJAge <- recode100(imputed_data$MJAge)
imputed_data$CocAge <- recode100(imputed_data$CocAge)
imputed_data$HeroinAge <- recode100(imputed_data$HeroinAge)

# Creating a Hard Drug and an Opioid Column
Hard <- imputed_data$Morphine +
as.numeric(!is.na(imputed_data$CocAge)) +
as.numeric(!is.na(imputed_data$HeroinAge)) + imputed_data$LSD +
imputed_data$Ecstasy
```

```r
imputed_data$HardDrugs <- Hard

Opioid <- imputed_data$Morphine +
as.numeric(!is.na(imputed_data$HeroinAge))
Opioid <- Opioid > 0
imputed_data$Opioid <- as.numeric(Opioid)

# Plots
# Figure 1
with(imputed_data, boxplot(Weights*PostFac ~ Race, main = "Figure 1:
Weights by Subject Race", col = c("light yellow", "light blue",
"orange", "pink", "light green")))

# Figure 2
with(imputed_data, boxplot(Weights*PostFac ~ Region, main = "Figure
2: Weights by Region", col = colors()[seq(50, 200, 50)]))

# Creating Survey Design Object
# Adjusting PSU Options
options(survey.lonely.psu="adjust")
design <- svydesign(ids = ~ PSU + Household, strata = ~ Division +
Stratum + Segment, weights = ~ Weights*PostFac, data = imputed_data,
nest = T)

# Analysis

#### Ritalin Usage
ritalin_obj <- svyby(~Ritalin, design = design, FUN = svymean, by =
~AgeRange)
ritalin_obj$Ritalin <- ritalin_obj$Ritalin * 100
ritalin_obj$se <- ritalin_obj$se * 100
ritalin_obj

# Figure 3.
barplot(ritalin_obj, xlab = "Age Range", ylab = "Percentage of
Individuals Who Have Used Ritalin", main = "Figure 3: Barplot of
Ritalin Use Between Age Groups", col = "#003262")

confint(ritalin_obj)
```

```r
#### Opioid Use
opioid_obj <- svyby(~Opioid, design = design, FUN = svymean, by =
~Division)
opioid_obj$Opioid <- opioid_obj$Opioid * 100
opioid_obj$se <- opioid_obj$se * 100
opioid_obj

# Figure 4.
par(mar = c(9, 6, 5, 4))
barplot(opioid_obj, las = 2, ylab = "Percentage of Individuals \n Who
Have Used Opiates", main = "Figure 4: Barplot of Opiate Usage Between
Geographic Region", col = "#FDB515")

confint(opioid_obj)

#### Hard Drug Usage
hard_marijuana_obj <- svyby(~HardDrugs, design = design, FUN =
svymean, by = ~as.factor(as.numeric(!is.na(imputed_data$MJAge))))
names(hard_marijuana_obj) <- c("Marijuana", "HardDrugs", "se")
hard_marijuana_obj$HardDrugs <- hard_marijuana_obj$HardDrugs * 100
hard_marijuana_obj$se <- hard_marijuana_obj$se * 100
hard_marijuana_obj

# Figure 5.
par(mar = c(6, 6, 4, 3))
barplot(hard_marijuana_obj, ylim = c(0, 100), xlab = "Marijuana Use
(0 No, 1 Yes)", ylab = "Percentage of Individuals \n Who Have Used
Hard Drugs", main = "Figure 5: Barplot of Hard Drug Use Between
Marijuana Users", col = "#003262")

confint(hard_marijuana_obj)

hard_cig_obj <- svyby(~HardDrugs, design = design, FUN = svymean, by
= ~as.factor(as.numeric(!is.na(imputed_data$CigAge))))
names(hard_cig_obj) <- c("Cigarettes", "HardDrugs", "se")
hard_cig_obj$HardDrugs <- hard_cig_obj$HardDrugs * 100
hard_cig_obj$se <- hard_cig_obj$se * 100
hard_cig_obj
```

```r
# Figure 6.
par(mar = c(6, 6, 4, 3))
barplot(hard_cig_obj, ylim = c(0, 30), xlab = "Cigarette Use (0 No, 1
Yes)", ylab = "Percentage of Individuals \n Who Have Used Hard
Drugs", main = "Figure 6: Barplot of Hard Drug Use Between Cigarette
Users", col = "#FDB515")

confint(hard_cig_obj)

#### Health Issues
bp_obj <- svyby(~HighBP, FUN = svymean, design = design, by =
~as.factor(as.numeric(!is.na(imputed_data$CigAge))))
names(bp_obj) <- c("Cigarettes", "HighBP", "se")
bp_obj$HighBP <- bp_obj$HighBP * 100
bp_obj$se <- bp_obj$se * 100
bp_obj

# Figure 7.
par(mar = c(6, 6, 4, 3))
barplot(bp_obj, ylim = c(0, 30), xlab = "Cigarette Use (0 No, 1
Yes)", ylab = "Percentage of Individuals \n Who Have High Blood
Pressure", main = "Figure 7: Barplot of High Blood Pressure
Prevalence \n Between Cigarette Users", col = "#003262")

confint(bp_obj)

asthma_obj <- svyby(~Asthma, FUN = svymean, design = design, by =
~as.factor(as.numeric(!is.na(imputed_data$CigAge))))
names(asthma_obj) <- c("Cigarettes", "Asthma", "se")
asthma_obj$Asthma <- asthma_obj$Asthma * 100
asthma_obj$se <- asthma_obj$se * 100
asthma_obj

# Figure 8.
par(mar = c(6, 6, 4, 3))
barplot(asthma_obj, ylim = c(0, 30), xlab = "Cigarette Use (0 No, 1
Yes)", ylab = "Percentage of Individuals \n Who Have Asthma", main =
"Figure 8: Barplot of Asthma Prevalence Between Cigarette Users", col
```

```
= "#FDB515")

confint(asthma_obj)

psych_obj <- svyby(~Psych, FUN = svymean, design = design, by =
~as.factor(as.numeric(!is.na(imputed_data$CigAge))))
names(psych_obj) <- c("Cigarettes", "Psych", "se")
psych_obj$Psych <- psych_obj$Psych * 100
psych_obj$se <- psych_obj$se * 100
psych_obj

# Figure 9.
par(mar = c(6, 6, 4, 3))
barplot(psych_obj, ylim = c(0, 10), xlab = "Cigarette Use (0 No, 1
Yes)", ylab = "Percentage of Individuals \n Who Have A Psychological
Disorder", main = "Figure 9: Barplot of Psychological Disorder \n
Prevalence Between Cigarette Users", col = "#003262")

confint(psych_obj)

heart_obj <- svyby(~Heart, FUN = svymean, design = design, by =
~as.factor(as.numeric(!is.na(imputed_data$CigAge))))
names(heart_obj) <- c("Cigarettes", "Heart", "se")
heart_obj$Heart <- heart_obj$Heart * 100
heart_obj$se <- heart_obj$se * 100
heart_obj

# Figure 10.
par(mar = c(6, 6, 4, 3))
barplot(heart_obj, ylim = c(0, 25), xlab = "Cigarette Use (0 No, 1
Yes)", ylab = "Percentage of Individuals \n Who Have Heart Disease",
main = "Figure 10: Barplot of Heart Disease Prevalence \n Between
Cigarette Users", col = "#FDB515")

confint(heart_obj)
```