Substance Abuse in Adolescents

HBSC 2001-2002

Stat 152: Sampling Surveys

Introduction

"I want to be able to look back and say, 'I've done everything I can, and I was successful.' I don't want to look back and say I should have done this or that. I'd like to change things for the younger generation of the swimmers combing along."

---Michael Phelps

The above quote is said by Michael Phelps, an American swimmer who has won 28 Olympic medals. The quote reflects how important it is that one spends their youth meaningfully and without regrets. Nowadays in the United States, increasing youth violence and substance abuse has become a major concern in the nation with rising rates in juvenile arrests. Such trend not only endangers the safety of the nation, but also significantly influences the education, growth, and future of the younger generation, physically and mentally, in a dreadful way. Therefore, the research on factors that affects substance abuse rate in younger generation is urgent. With statistically significant and meaningful result, specific programs and laws can be enacted to help our younger generation stay away from illegal overindulgence issues. With this purpose in mind, we will be examining factors and trends in juvenile substance abuse using the HBSC data in this report.

Questions we are addressing

We seek to explore the large question of what factors influence or determine the rate and frequency of adolescent substance abuse (marijuana, tobacco, alcohol, and inhalant). Under this larger background question, we focus on gender, race, grade, family background, personality in terms of happiness level, and the relationship between each of the above variables with substance abuse rate of middle school and high school students.

Do girls and boys differ in their rate and frequency of smoking tobacco, marijuana, inhalant and drinking alcohol? Do students from different grades differ? What about family background: are kids from low-income family more likely become addicted or kids from high-income family background? Does the level of happiness influence students' decision on whether to smoke and drink? And finally, does racial identity play a role in substance abuse?

By answering the above questions, we want to explore in depth the patterns in substance abuse rate and frequency among teenagers in the United States.

Summary of Analysis and Results

Our analysis involved using 18 variables, for which we will explain in further detail in the methodology part. We found that identity variables, such as gender, race, grade, and family background all are statistically correlated with rate and frequency of adolescent substance abuse. Boys are particularly more likely to be addicted than girls to tobacco, marijuana, inhalant and alcohol. Similarly, higher-grade students, students coming from low-income families are more likely to be addicted. And in terms of racial identity, the White population has the highest rate of substance overindulgence. Our result shows that patterns of substance abuse in the young generation are rather obvious and evident.

Survey Design

About ICPSR

ICPSR, the Inter-University Consortium for Political and Social Research, was established in 1962 and has been providing "access to a vast archive of social science data for research and instruction (over 8,000 discrete studies/surveys with more than 65,000 datasets)"[1] ever since. ICPSR maintains many primary research findings and supports students, instructors, researchers and others on conducting secondary research and other study or teach purposes.

The dataset we obtained for use in this report, the HBSC survey, is from the ICPSR data archive on social science.

About HBSC

The HBSC—abbreviated for "Health Behavior in School-Aged Children" study, was initiated in 1982 and became affiliated with World Health Organization shortly afterwards. While initially, surveys were only conducted in founder countries, such as Finland, Norway and England and a small number of other countries, since 1985, HBSC study has been conducted every four years in a growing number of countries around the world. The data set we use in this report is from 2001-02 HBSC study for US. sample only.

The focus of the HBSC study is on school-aged children between 11 to 15 years old. The information collected about these teenagers include family composition, school and after-school activities, nutrition intake, financial background, and substance abuse. With the data collected, the international HBSC study aims at expanding understanding of young people and the reasons behind their health behaviors, and accelerating creation of health and social programs that can efficiently address specific issues to young people.

Design Elements

The HBSC adopted a two-stage stratified cluster sampling method. A full description of the sampling method can be found at
https://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/4372#

[1] https://en.wikipedia.org/wiki/Inter-university_Consortium_for_Political_and_Social_Research

The universe in this study consisted of "public, Catholic and other private school students in grades 6, 7, 8, 9, and 10 or their equivalent in the 50 states and District of Columbia"[2]. Some of the extremely small schools with less than 14 classes were excluded in this data set. We don't expect this exclusion to cause any under-coverage problem because those small schools only "comprise about 1 percent of the enrollment in U.S."

The sample design follows a stratified, two-stage cluster method of classes at grades 6 through 10. The sampling frame used for the United States sample is a comprehensive list of schools from Quality Education Data, Inc. In the first stage, the primary sampling units (PSUS) are the individual district or groups of districts under geographic stratification. Very small districts are excluded from the sample and large districts are split in a way that preserves the probability of selection of each PSU. The selection of PSUs follows proportional-to-size method and a random systematic sampling procedure using aggregate enrollment as the measure of size. It is important to note that all the PSUs are selected independently for our calculation.

Next, HBSC employs different methods to sample private and public schools from the stratum. The private school sample is selected with a single –stage geographic stratification and probability of each private school being selected is proportional to grade enrollment. On the other hand, although public schools are also selected with probabilities proportional to the grade enrollment, the public-school samples are not independent. Schools are selected with a "Permanent Random Number method to maximize the overlap between school samples selected for each grade."

Finally, within the schools selected from sampling in previous stages, whether public or private, classes are selected suing simple random sampling from each target grade. Then all students in the selected class are asked to participate in the study in the form of answering questionnaires. The survey took approximately 45 minutes to complete and was administered in a regular classroom setting. There were two versions for the questionnaires: one for high school students with 92 multiple-choice and one for middle school students with 77 multiple-choice questions. Then, auxiliary questionnaires were sent to the administrators of the participating schools to obtain school level information such as physical activity, violence rate, and school code regarding tobacco use.

Table 1:
Summary of Sampling Stages

Stage	Sampling Unit	Sampling Method	Stratification
1	Districts or groups of districts	Proportional to size & random	Geographic

[2] United States Department of Health and Human Services. Health Resources and Services Administration. Maternal and Child Health Bureau. Health Behavior in School-Aged Children, 2001-2002 [United States]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2008-07-24. https://doi.org/10.3886/ICPSR04372.v2

		systematic sampling	
2	Private schools: one school	Proportional to grade enrollment	geographic
	Public schools: one school	Permanent Random Number method	
3	One class	Simple random sampling	NA

Response Rate

In the U.S., because of the lack of State and local governmental support to enforce the HBSC, there is a relatively low expectation on the response rate. In the original sample frame, 548 schools were selected to participate in the survey of which only 344 schools responded. To achieve the required number of schools, another set of selection into the original sampling frame without replacement were included to reach the balance of 465 schools. In the end, the number of schools responding was 340 resulting in a school participation rate of 73.2 percent. Of the 18620 students in the 340 participating schools, 15245 students returned the questionnaire yielding a student response rate of 81.9 percent. Even though there are still limitations to estimation because non-responses, the "participation rates were sufficient to achieve the target precision levels and confidence intervals for the sub-populations of interest."

Here is a table of the response rate and reasons for non-response:

Table 2: Nonresponse information

Reasons for student-level non-response	Number of non-response	Percentage of the total non-response
Absence to school	637	39.32%
Did not return consent form	600	37.04%
Parents declined to participate	518	31.97%
Students declined to respond	1620	100%

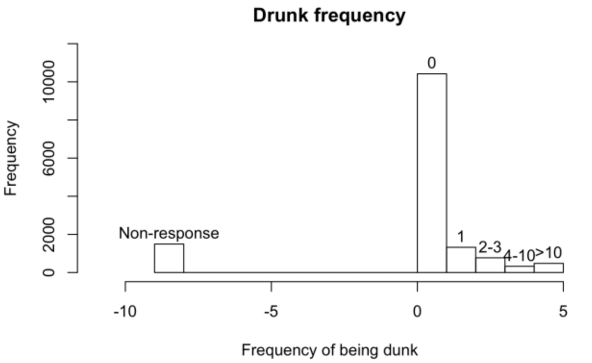
For the auxiliary questionnaires for school administrators, 329 were completed out of 340 sent. For those send to lead health educators, 320 forms were returned out of 340. In total, 317 schools returned both the administrator and health educator surveys for a total response of 93.2 percent participation rate.

Exploration of the Design Element

The main response variables that we want to analyze are the frequencies of tobacco, alcohol, marijuana, and inhalant use among high school students. After exploring the variables, we found that all the response variables are either qualitative variables indicating whether the students engage in certain kind of substance abuse, or factor variables presenting the frequency levels (never, 1-2times, 3-5 times). Specific details about our original response variables can be found in Table 3.

Selected graphs showing distribution of our response variables are presented below. In the histogram for “Drunk Frequency”, we can see that most of the values are zero consisting of more than half of the observation points. The distribution is highly skewed to the right indicating that only a small fraction of students get drunk frequently. Similarly, the graph displaying inhalant use shows that most of the students never had inhalant. The rest of the diagrams of our response variables show similar result. Therefore, we can conclude that while there are students consistently involving in substance abuse, most of the students in the survey are not susceptible to substance abuse.

Methodology



The data were obtained from the ICPSR website for the 2001-2002 HBSC dataset. The original dataset has 14817 observations and 563 variables. For our analysis purpose, we cleaned the dataset and only included 18 variables as our variables of interest, creating another dataset, “dat”, with all variables renamed. Since there are many missing values in the marijuana use column because only high school students were surveyed on this question rather than all grades, we first differentiate the nonresponse to variable “marijuana-freq_hs”, using “-8” to denote that nonresponse is resulted from student not qualified for answering a specific survey question. Next, we calculated the nonresponse rate for each variable. Since all the nonresponse rates are no larger than 15%, we conclude that it is acceptable for us to proceed with the analysis.

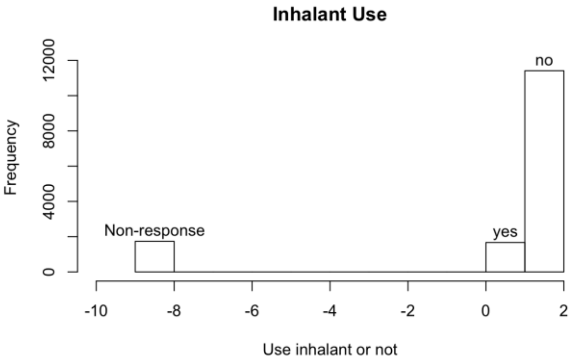


Table 3: variables of interest

Variable Name from HBSC	Variable Name in our analysis	Description	Values, as coded in the data set
Q1	Gender	Answer to question “Are you a boy or a girl?”	1 = Boy 2 = Girl

			-9 = missing data
Q4	Grade	Answer to question: “what grade are you in?”	1, 2, 3, …, 8 denotes 5th grade to 12th grade respectively. 9 = ungraded
Q5, Q6	Race	Answer to the question: “what is your race?”	1 = Native or Alaska Native, 2 = Asian, 3 = African, 4 = Hawaiian or another Pacific Islander, 5 = White, 6 = two or more, -9 = nonresponse
Q7	Urbanicity	“what kind of place do you live in?”	1 = urban, 2 = suburban, 3 = rural, -9 = nonresponse
Q17	Mothereduc	What is your mother’s highest level of education?	1 = some high school or less, 2 = high school, 3 = some college, 4 = college or more, 5 = don't know, -9 = nonresponse
Q18	fathereduc	What is your father’s highest level of education?	Same as mother’s education.
Q76	familywellof	How well off do you think your family is?	1 = very, 2 = quite, 3 = average, 4 = not very, 5 = not at all, -9 = nonresponse
Q80	senseofsafety	Generally Speaking, I feel safe in the area where live…	1 = always, 2 = mostly, 3 = sometimes, 4 = rarely or never, -9 = nonresponse
Q82	Tobacco-yn	Have you ever smoked tobacco? (at least one cigarette, cigar, or pipe)	1 = yes, 2 = no, -9 = nonresponse
Q83	Tobacco_freq	How often do you smoke tobacco at present?	1 = daily, 2 = daily to weekly, 3 = fewer than weekly, 4 = nonsmoker, -9 = nonresponse
Q85_COMP	Alcohol_yn	Have you ever drunk anything alcoholic, such as beer, wine, liquor/spirits, or alcopops?	1 = yes, 2 = no, -9 = nonresponse

Q86	Alcohol_freq	Have you ever had so much alcohol that you were really drunk?	1 = daily, 2 = daily to weekly, 3 = fewer than weekly, 4 = nonsmoker, -9 = nonresponse
Q88A	Marijuana_freq-hs	Have you ever used or taken one or several of these drugs in your life? [high school students only] Marijuana (pot, weed)	1 = yes, 2 = no, -9 = nonresponse
Q88A_COMP	Marijuana_yn_hs	Have you ever used to take one or several of these drugs in your life? Marijuana (pot, weed) [high school only]	1 = never, 2 = once, 3 = 2-3 times, 4 = 4-10 times, 5 = more than 10 times, -9 = nonresponse
Q88B	Inhalant_freq	Have you ever used to take one or several of these drugs in your life? Inhalants (including huffing or sniffing glue, aerosol cans, or paint to get high)	1 = never, 2 = 1-2 times, 3 = 3-5 times, 4 = 6-9 times, 5 = 10-19 times, 6 = 20-39 times, 7 = 40 times or more, -9 = nonresponse of high school students, -8 = nonresponse of middle school students
Q88B_COMP	Inhalant_yn	Have you ever used to take one or several of these drugs in your life? Inhalants (including huffing or sniffing glue, aerosol cans, or paint to get high)	1 = yes, 2 = no, -9 = nonresponse of high school students, -8 = nonresponse of middle school students
STU_WT	Student weight	Student weight measured	Qualitative values
BMI	BMI	What is your BMI? Computed based on the respondent's weight and height using the following formula: [Weight(lbs) / [Height (inches) * Height (inches)]] * 703	Quantitative Variable, -9 = nonresponse

Since our response variables are not continuous quantitative variables, we combined the frequency level data from each substance (tobacco, alcohol, marijuana, inhalant) into one quantitative variable named "abuse propensity index", which is a numeric value that is calculated using a specific mechanism and is intended to showcase the extent to which an individual is susceptible to substances. The formula we used to calculate this index is the following:

(Overall abuse propensity) index = 14 + tobacco_yn + tobacco_freq + alcohol_yn - alcohol_freq + inhalant_yn - inhalant_freq + marijuana_yn_hs - marijuana_freq_hs
(Tobacco abuse propensity) index = -2 + tobacco_yn + tobacco_freq
(Alcohol abuse propensity) index = 4 + alcohol_yn - alcohol_freq
(Inhalant abuse propensity) index = 6 + inhalant_yn - inhalant_freq
(Marijuana abuse propensity) index = 6 + marijuana_yn_hs - marijuana_freq_hs

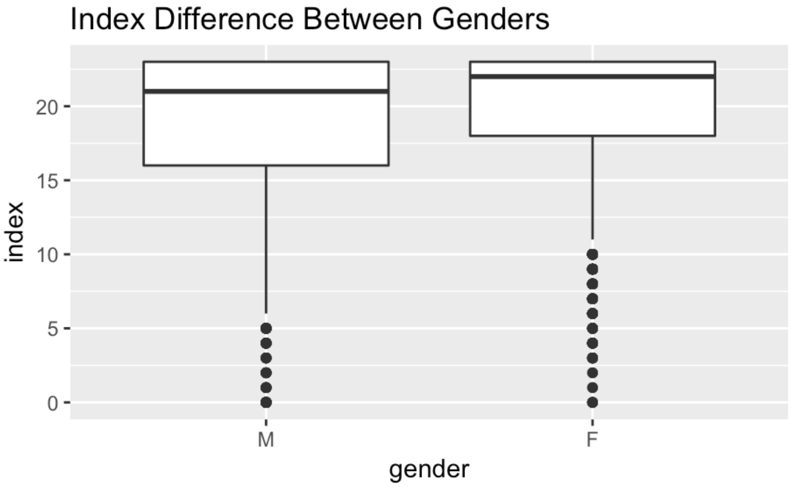
(Note: Constant terms are introduced to **adjust the lowest scores to 0**. All the variables in the equation correspond exactly to the variable names in table 3)

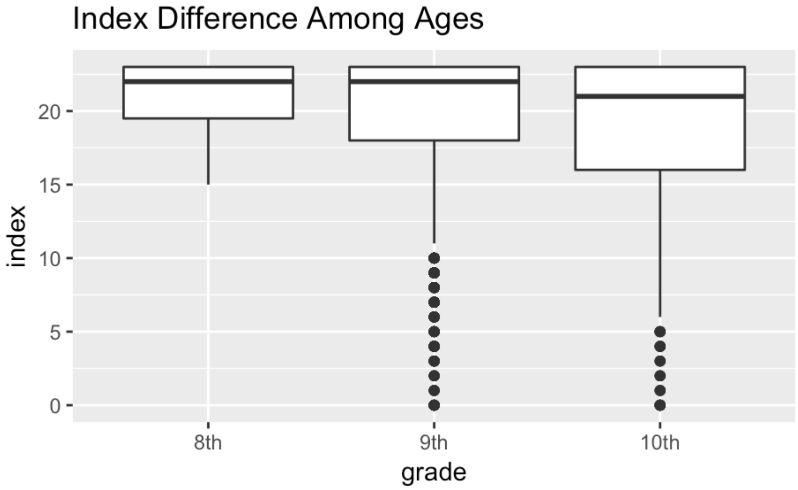
The abuse propensity index is computed in a way that the higher it is, the less likely an individual is susceptible to substances. In other words, the abuse propensity index and likelihood of an individual engaging in substance abuse is negatively related.

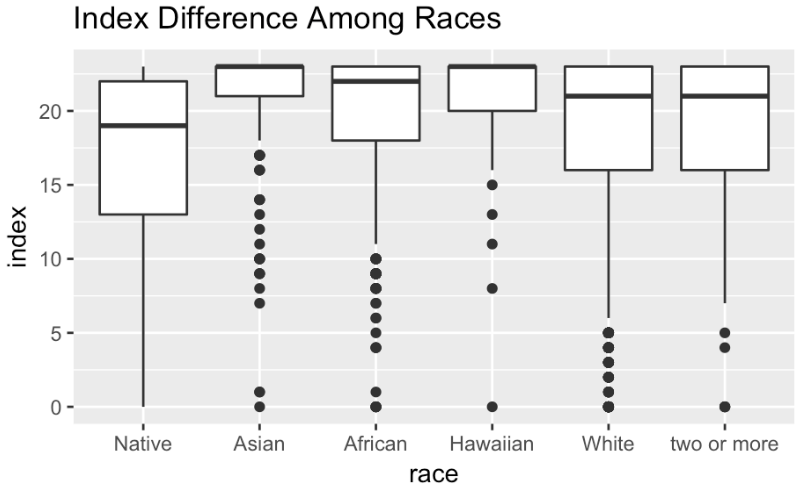
Results

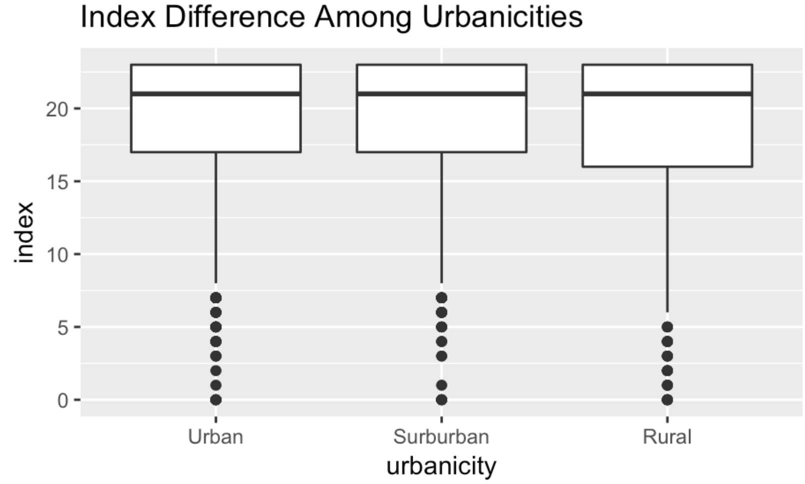
We first look at the **unweighted** comparisons of **abuse propensity scores** among different genders, grades, races, urbanicity, parents' education levels, family well-of, and sense of safety.

From the eight boxplots (see Figure 1-4 for demonstrating examples), we can see that in each of the comparisons, there is a clear difference in the index depending on gender, grade, ethnicity, parents' education level, family wellness, and sense of safety. Particularly, female average index is higher than that of male's, suggesting lower substance abuse likelihood; 10th graders have index significantly lower than that of 8th and 9th graders on average; the indexes for different racial groups appear to be distinctively different from each other too. It is suggested that females, younger students, Asians, children with parents of higher education, children from wealthier families and children with greater sense of safety have a higher index and are less susceptible to substance abuse. While on the other hand, urbanicity does not matter in this case. It only has some effects on the threshold of the lower quartile.


Index Difference Between Genders


Index Difference Among Ages


Index Difference Among Races


Index Difference Among Urbanicities

We then use the survey package to **make weight adjustment**. The variable “student_weight” demonstrates the weight of every observation unit (i.e. student).

```
svy = svydesign(id = ~1, weights = ~student_weight, data = hs_resp, nest = T)
```

From Figure 5 with weight adjustment, the side-by-side histograms all skew to the right, indicating that most of our data points cluster around the lower scale of the histogram, whichever index we are using, most students do not have a high substance abuse propensity level, thus are less susceptible to substance abuse. This is especially the case for inhalant and marijuana usage. Only a very small fraction of students is exposed to these substances.

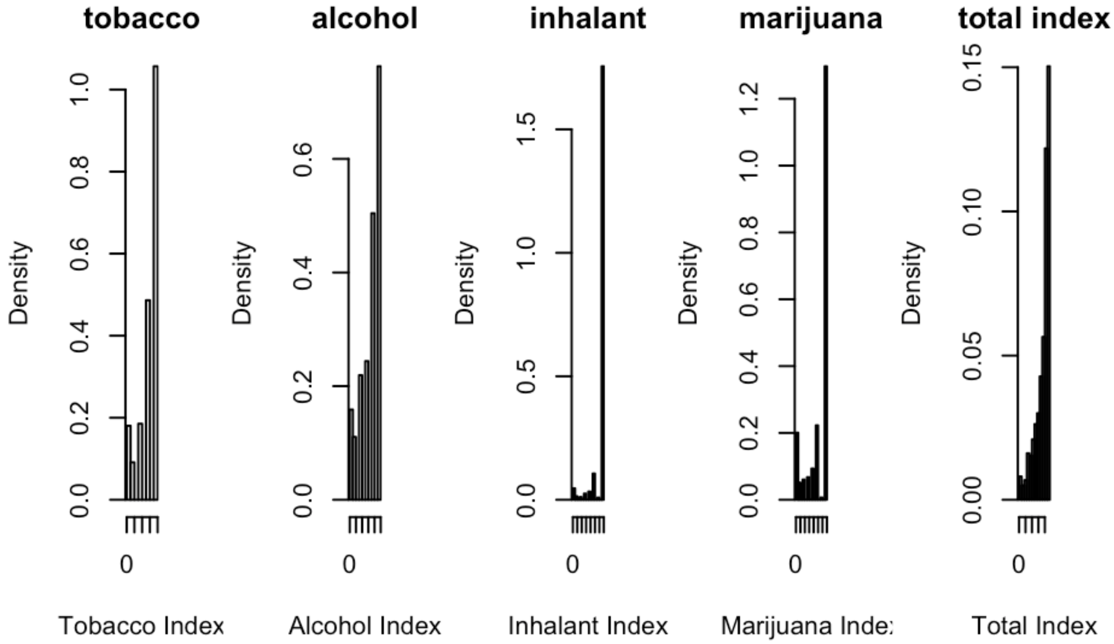


Figure 5

As for Figure 6 with the weighted side-by-side boxplots, we can see that the difference between genders mainly occurs in the lower quartiles, meaning that roughly the number of students less exposed to substances are similar regardless of the gender, while those who are more exposed might be more likely to be male.

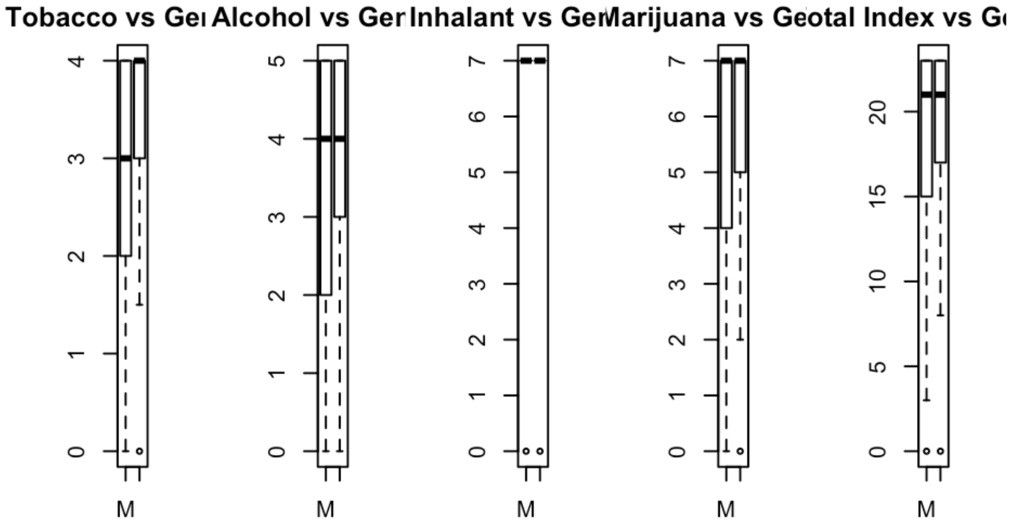


Figure 6

We also generated 6 boxplots with genders as strata and showed the relationships between the abuse index and other explanatory variables. From Figure 7-12, it is demonstrative that for all the explanatory variables (grades, races, urbanicity, education of parents, family wellness, sense of safety), it generally follows the pattern that boys are more susceptible to substance abuse than girls since they generally have a larger distribution towards the lower quartile. It is also notable that within some of the variables, the indexes might significantly differ depending on the level of that variable. For example, for students who live in a family that is not well off at all (Figure 11), the red bar representing boys in this category is especially much more spread out towards the lower end. Similarly, in Figure 12, the red bar that represent boys who rarely or never have a sense of safety is also much more spread out towards the lower end.

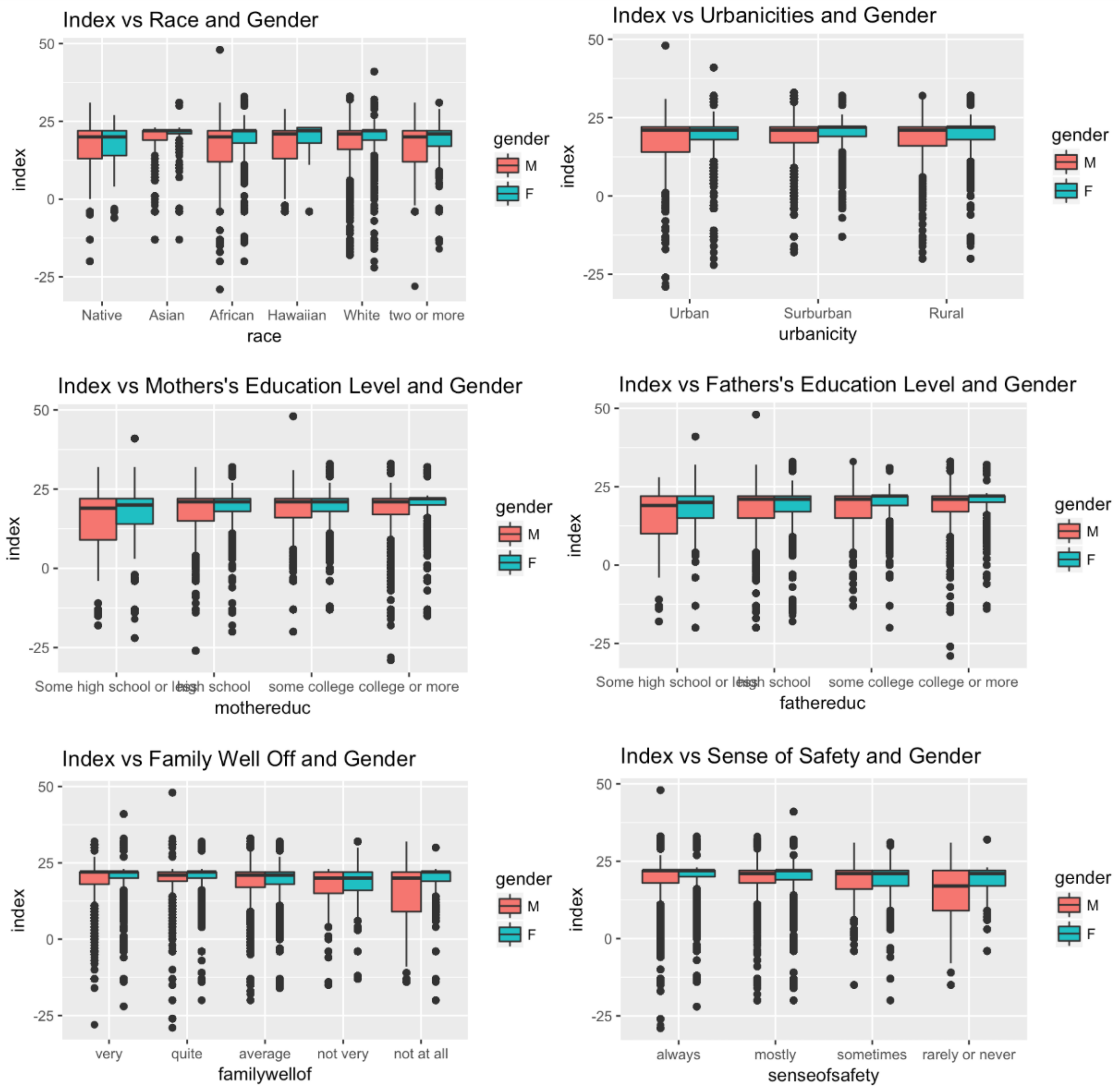


Figure 7-12

Regression analysis

Besides the tests and survey analysis we already conducted, we could also verify our results using multiple-linear regression of substance abuse propensity index on all the variables of interest. In this case, since most of our variables are qualitative and factor variables, instead of the usual regression model, we will incorporate a lot of dummy variables while adhering to the Principle of Marginality.

The result of our regression analysis corresponds to our previous tests to some extent: gender, grades, racial group, parents' education background, and the sense of safety are the variables with the most statistical significance with respect to our response variable—abuse propensity index. Also, the urbanicity factor is not statistically significant on 90% confidence interval, which verifies our analysis in the beginning, that urbanicity is not a major factor influencing substance abuse.

However, there are some concerns with the regression model:

1. Since there are many dummy variables, some of them might be highly correlated in the sense that some variables capture similar information (such as parents' education background and family well-off), therefore influencing the accuracy of coefficients.
2. The model may suffer from omitted variable bias since we only included 18 out of 563 variables from the original survey, not to say the many other variables that might have a potential influence on substance abuse are not included even in the original survey.

Therefore, the validity of our regression model is difficult to verify and we decide to just use the model as an intuition check. Since the results of the regression model mostly follow our test results, we will adhere to our original analysis.

Discussion and Conclusion

The original survey conducted by HBSC employed a 2-staged stratified cluster sampling to collect information on teenagers aged 11, 13, 15 years old on information include but not limited to, identity, background, daily activities, emotional condition, and substance abuse frequencies. This survey targeted teenagers in the U.S. and aimed at supporting analysis on adolescent substance abuse issues.

After our analysis using survey package and multiple regression method, we concluded that the dataset of 2001 to 2002 U.S. HBSC presents high correlation between the level of students' susceptibility to substance abuse and variables such as gender, race, age, and family background. Particularly, "male", "Caucasian", "higher graders", "lower-income family" are tags associated with higher rate and frequency of substance abuse including alcohol, marijuana, tobacco, and inhalants. After confirming with multiple linear regression models, we concluded that our results are statistically and intuitively valid based on the dataset we obtained from HBSC.

Since we only used selected variables from the survey to conduct the analysis, there might be other variables also influencing the susceptibility of adolescents' substance abuse level. Therefore, we suggest that as next step, more research should be conducted on determining a more comprehensive model to predict and control for substance abuse among teenagers in the United States. All in all, we strongly believe in the validity of our analysis, as well as the importance of further study on teenager substance abuse in the United States.

Citations

United States Department of Health and Human Services. Health Resources and Services Administration. Maternal and Child Health Bureau. Health Behavior in School-Aged Children, 2001-2002 [United States]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2008-07-24. https://doi.org/10.3886/ICPSR04372.v2

https://en.wikipedia.org/wiki/Inter-university_Consortium_for_Political_and_Social_Research

Appendix: R Code

```
install.packages("ggplot2")
library(ggplot2)
install.packages("survey")

rawdata <- read.table(file = "~/Desktop/04372-0001-Data.tsv", sep = '\t', header = TRUE)
dat <- data.frame(rawdata$Q1, rawdata$Q4, rawdata$Q6_COMP, rawdata$Q7, rawdata$Q17,
rawdata$Q18, rawdata$BMI, rawdata$Q76, rawdata$Q80, rawdata$Q82, rawdata$Q83,
rawdata$Q85_COMP, rawdata$Q86, rawdata$Q88A, rawdata$Q88A_COMP, rawdata$Q88B,
rawdata$Q88B_COMP, rawdata$STU_WT)
coltitle <- colnames(dat) <- c('gender', 'grade', 'race', 'urbanicity', 'mothereduc', 'fathereduc', 'BMI',
'familywellof', 'senseofsafety', 'tobacco_yn', 'tobacco_freq', 'alcohol_yn', 'alcohol_freq', 'marijuana_freq_hs',
'marijuana_yn_hs', 'inhalant_freq', 'inhalant_yn', 'student_weight')
for (i in 1:nrow(dat)){
  if (dat$marijuana_yn_hs[i] == -8){
    dat$marijuana_freq_hs[i] <- -8
  }
}
# The questions for marijuana use were designed for "high school students only". Hence, for Q88A
(marijuana_freq_hs), we need to differentiate the nonresponse by middle school students from the
nonresponse by high school students. '-8' means that the respondent is not qualified for answering a specific
question.

# Standard summary of variables
# General rule: xxx_yn means yes-no question, and xxx_freq means question with options sorted by
frequency.
# 1. gender (1 = Boy, 2 = Girl)
# 2. grade (1 = 5th, 2 = 6th, 3 = 7th, 4 = 8th, 5 = 9th, 6 = 10th)
# 3. race (1 = Native or Alaska Native, 2 = Asian, 3 = African, 4 = Hawaiian or other Pacific Islander, 5 =
White, 6 = two or more, -9 = nonresponse)
# 4. urbanicity (1 = urban, 2 = suburban, 3 = rural, -9 = nonresponse)
# 5. mothereduc (Mother's highest education: 1 = some high school or less, 2 = high school, 3 = some
college, 4 = college or more, 5 = don't know, -9 = nonresponse)
# 6. fathereduc (Father's highest education: 1 = some high school or less, 2 = high school, 3 = some college,
4 = college or more, 5 = don't know, -9 = nonresponse)
# 7. BMI (Quantitative Variable, -9 = nonresponse)
# 8. familywellof (1 = very, 2 = quite, 3 = average, 4 = not very, 5 = not at all, -9 = nonresponse)
# 9. senseofsafety (1 = always, 2 = mostly, 3 = sometimes, 4 = rarely or never, -9 = nonresponse)
# 10. tobacco_yn (1 = yes, 2 = no, -9 = nonresponse)
# 11. tobacco_freq (1 = daily, 2 = daily to weekly, 3 = fewer than weekly, 4 = nonsmoker, -9 = nonresponse)
# 12. alcohol_yn (1 = yes, 2 = no, -9 = nonresponse)
```

```r
# 13. alcohol_freq (1 = never, 2 = once, 3 = 2-3 times, 4 = 4-10 times, 5 = more than 10 times, -9 = nonresponse)
# 14. marijuana_freq_hs (Question only for high school students: 1 = never, 2 = 1-2 times, 3 = 3-5 times, 4 = 6-9 times, 5 = 10-19 times, 6 = 20-39 times, 7 = 40 times or more, -9 = nonresponse of high school students, -8 = nonresponse of middle school students)
# 15. marijuana_yn_hs (Question only for high school students: 1 = yes, 2 = no, -9 = nonresponse of high school students, -8 = nonresponse of middle school students)
# 16. inhalant_freq (1 = never, 2 = 1-2 times, 3 = 3-5 times, 4 = 6-9 times, 5 = 10-19 times, 6 = 20-39 times, 7 = 40 times or more, -9 = nonresponse)
# 17. inhalant_yn (1 = yes, 2 = no, -9 = nonresponse)
# 18. student weight: unique value

# Nonresponse check: calculate the frequencies of "-9".
nonresponse_rate <- c()
for (i in 1:ncol(dat)){
  nonresponse_rate <- c(nonresponse_rate, c(nrow(dat[dat[,i] == -9,])/nrow(dat)))
}
nonresponse_rate = as.data.frame(nonresponse_rate)
row.names(nonresponse_rate) <- coltitle
nonresponse_rate[,1]
# The nonresponse rates are stated above. It turns out that nonresponse is acceptable for the survey questions we are interested in.

hist(dat$gender, breaks=seq(0,2,by=1), freq = TRUE, ylim = c(0, 9000), xlab = "Gender", main = "Gender counts", labels = c("male", "female"))
hist(dat$alcohol_freq, ylim = c(0, 12000), xlim = c(-11,5), freq = TRUE, xlab = "Frequency of being dunk", main = "Drunk frequency", labels = c("Non-response", " ", " ", " ", " ", " ", " ", " ", " ", "0", "1", "2-3", "4-10", ">10"))
hist(dat$inhalant_yn, ylim = c(0, 13000), xlim = c(-10,2), freq = TRUE, xlab = "Use inhalant or not", main = "Inhalant Use", labels = c("Non-response", " ", " ", " ", " ", " ", " ", " ", " ", "yes", "no"))
hist(dat$tobacco_yn, ylim = c(0, 12000), xlim = c(-11,5), freq = TRUE, xlab = "The use of smoking tobacco", main = "Tobacco use Yes/No")
hist(dat$mothereduc, ylim = c(0, 12000), xlim = c(-11,5), freq = TRUE, xlab = "Education levels", main = "Mother education")
hist(dat$student_weight, ylim = c(0, 6000), xlab = "Student weight", freq = TRUE, main = "Student weight")

dat$gender = factor(dat$gender, levels = 1:2, labels = c("M", "F"))
dat$grade = factor(dat$grade, levels=1:6, labels = c("5th", "6th", "7th", "8th", "9th", "10th"))
dat$race = factor(dat$race, levels=1:6, labels = c("Native", "Asian", "African", "Hawaiian", "White", "two or more"))
dat$urbanicity = factor(dat$urbanicity, levels=1:3, labels = c("Urban", "Surburban", "Rural"))
dat$mothereduc = factor(dat$mothereduc, levels=1:4, labels = c("Some high school or less", "high school", "some college", "college or more"))
dat$fathereduc = factor(dat$fathereduc, levels=1:4, labels = c("Some high school or less", "high school", "some college", "college or more"))
dat$familywellof = factor(dat$familywellof, levels=1:5, labels = c("very", "quite", "average", "not very", "not at all"))
dat$senseofsafety = factor(dat$senseofsafety, levels=1:4, labels = c("always", "mostly", "sometimes", "rarely or never"))

# Unweighted boxplots
# "Index" is a combined addiction index that denotes the extent to which a student is addicted to tobacco, alcohol, inhalant and marijuana.

# Only high school students are sampled.
```

```
library(ggplot2)
hs_resp = dat[dat$tobacco_yn != -9 & dat$tobacco_freq != -9 & dat$alcohol_yn != -9 &
dat$alcohol_freq != -9 & dat$inhalant_yn != -9 & dat$inhalant_freq != -9 & dat$marijuana_yn_hs > -8 &
dat$marijuana_freq_hs > -8,]
hs_resp$index = 14 + hs_resp$tobacco_yn + hs_resp$tobacco_freq + hs_resp$alcohol_yn -
hs_resp$alcohol_freq + hs_resp$inhalant_yn - hs_resp$inhalant_freq + hs_resp$marijuana_yn_hs -
hs_resp$marijuana_freq_hs
hs_resp$index_tob = hs_resp$tobacco_yn + hs_resp$tobacco_freq - 2
hs_resp$index_alc = hs_resp$alcohol_yn - hs_resp$alcohol_freq + 4
hs_resp$index_inh = hs_resp$inhalant_yn - hs_resp$inhalant_freq + 6
hs_resp$index_mar = hs_resp$marijuana_yn_hs - hs_resp$marijuana_freq_hs + 6
ggplot(data = hs_resp) + geom_boxplot(aes(x = gender, y = index)) + ggtitle("Index Difference Between
Genders")
ggplot(data = hs_resp) + geom_boxplot(aes(x = grade, y = index)) + ggtitle("Index Difference Among
Ages")
ggplot(data = hs_resp[!is.na(hs_resp$race),]) + geom_boxplot(aes(x = race, y = index)) + ggtitle("Index
Difference Among Races")
ggplot(data = hs_resp[!is.na(hs_resp$urbanicity),]) + geom_boxplot(aes(x = urbanicity, y = index)) +
ggtitle("Index Difference Among Urbanicities")
ggplot(data = hs_resp[!is.na(hs_resp$mothereduc),]) + geom_boxplot(aes(x = mothereduc, y = index)) +
ggtitle("Index Difference With Respect to Mothers' Education Level")
ggplot(data = hs_resp[!is.na(hs_resp$fathereduc),]) + geom_boxplot(aes(x = fathereduc, y = index)) +
ggtitle("Index Difference With Respect to Fathers' Education Level")
ggplot(data = hs_resp[!is.na(hs_resp$familywellof),]) + geom_boxplot(aes(x = familywellof, y = index)) +
ggtitle("Index Difference Based on Family Well Off")
ggplot(data = hs_resp[!is.na(hs_resp$senseofsafety),]) + geom_boxplot(aes(x = senseofsafety, y = index))
+ ggtitle("Index Difference Based on Sense of Safety")

# Weighted graphes with survey package
library(survey)
svy = svydesign(id = ~1, weights = ~student_weight, data = hs_resp, nest = T)
# Histograms (weighted)
par(mfrow = c(1,5))
svyhist(~index_tob, svy, main = "tobacco", xlab = "Tobacco Index")
svyhist(~index_alc, svy, main = "alcohol", xlab = "Alcohol Index")
svyhist(~index_inh, svy, main = "inhalant", xlab = "Inhalant Index")
svyhist(~index_mar, svy, main = "marijuana", xlab = "Marijuana Index")
svyhist(~index, svy, main = "total index", xlab = "Total Index")

# Boxplots of Index vs Gender (weighted)
par(mfrow = c(1,5))
svyboxplot(index_tob~gender, svy, main = "Tobacco vs Gender", cex = 0.5)
svyboxplot(index_alc~gender, svy, main = "Alcohol vs Gender", cex = 0.5)
svyboxplot(index_inh~gender, svy, main = "Inhalant vs Gender", cex = 0.5)
svyboxplot(index_mar~gender, svy, main = "Marijuana vs Gender", cex = 0.5)
svyboxplot(index~gender, svy, main = "Total Index vs Gender", cex = 0.5)

# Boxplot of Total Index vs Race (weighted)
par(mfrow = c(1,1))
svyboxplot(index~race, svy, main = "Total Index vs Race (weighted)")

# Mean, SE, CI
index_mean = svymean(~index, svy)
index_SE = SE(index_mean)
index_CI = svyquantile(~index, svy, c(0.25,0.5,0.75), ci = TRUE)
index_mean
```

```
# Create a pseudo population based on weights. Need 1 minute runtime.
dat$new_weight = round(dat$student_weight/100)
pseudo = list(0)
for(i in 1:nrow(dat)){
pseudo[[i]] = replicate(dat$new_weight[i], dat[i,])
}
popn = data.frame(matrix(unlist(pseudo), nrow = sum(dat$new_weight), byrow = T), stringsAsFactors =
FALSE)
colnames(popn) = c("gender", "grade", "race", "urbanicity", "mothereduc", "fathereduc", "BMI",
"familywellof", "senseofsafety", "tobacco_yn", "tobacco_freq", "alcohol_yn", "alcohol_freq",
"marijuana_freq_hs", "marijuana_yn_hs", "inhalant_freq", "inhalant_yn", "student_weight", "new_weight")

popn$gender = factor(popn$gender, levels = 1:2, labels = c("M", "F"))
popn$grade = factor(popn$grade, levels=1:6, labels = c("5th", "6th", "7th", "8th", "9th", "10th"))
popn$race = factor(popn$race, levels=1:6, labels = c("Native", "Asian", "African", "Hawaiian", "White",
"two or more"))
popn$urbanicity = factor(popn$urbanicity, levels=1:3, labels = c("Urban", "Surburban", "Rural"))
popn$mothereduc = factor(popn$mothereduc, levels=1:4, labels = c("Some high school or less", "high
school", "some college", "college or more"))
popn$fathereduc = factor(popn$fathereduc, levels=1:4, labels = c("Some high school or less", "high
school", "some college", "college or more"))
popn$familywellof = factor(popn$familywellof, levels=1:5, labels = c("very", "quite", "average", "not
very", "not at all"))
popn$senseofsafety = factor(popn$senseofsafety, levels=1:4, labels = c("always", "mostly", "sometimes",
"rarely or never"))
popn$index = 14 + popn$tobacco_yn + popn$tobacco_freq + popn$alcohol_yn - popn$alcohol_freq +
popn$inhalant_yn - popn$inhalant_freq + popn$marijuana_yn_hs - popn$marijuana_freq_hs
popn$index_tob = popn$tobacco_yn + popn$tobacco_freq - 2
popn$index_alc = popn$alcohol_yn + popn$alcohol_freq + 4
popn$index_inh = popn$inhalant_yn + popn$inhalant_freq + 6
popn$index_mar = popn$marijuana_yn + popn$marijuana_freq + 6
library(ggplot2)

# Addiction index of each gender with respect to race, urbanicity, parents' education, family well of, and
sense of safety.
ggplot(data = popn[!is.na(popn$race),]) + geom_boxplot(aes(x = race, y = index, fill = gender)) +
ggtitle("Index vs Race and Gender")
ggplot(data = popn[!is.na(popn$urbanicity),]) + geom_boxplot(aes(x = urbanicity, y = index, fill = gender))
+ ggtitle("Index vs Urbanicities and Gender")
ggplot(data = popn[!is.na(popn$mothereduc),]) + geom_boxplot(aes(x = mothereduc, y = index, fill =
gender)) + ggtitle("Index vs Mothers's Education Level and Gender")
ggplot(data = popn[!is.na(popn$fathereduc),]) + geom_boxplot(aes(x = fathereduc, y = index, fill = gender))
+ ggtitle("Index vs Fathers's Education Level and Gender")
ggplot(data = popn[!is.na(popn$familywellof),]) + geom_boxplot(aes(x = familywellof, y = index, fill =
gender)) + ggtitle("Index vs Family Well Off and Gender")
ggplot(data = popn[!is.na(popn$senseofsafety),]) + geom_boxplot(aes(x = senseofsafety, y = index, fill =
gender)) + ggtitle("Index vs Sense of Safety and Gender")

# Addiction index of each grade with respect to race, urbanicity, parents' education, family well of, and
sense of safety.
ggplot(data = popn[!is.na(popn$race),]) + geom_boxplot(aes(x = race, y = index, fill = grade)) +
ggtitle("Index vs Race and Grade")
ggplot(data = popn[!is.na(popn$urbanicity),]) + geom_boxplot(aes(x = urbanicity, y = index, fill = grade))
+ ggtitle("Index vs Urbanicities and Grade")
```

```
ggplot(data = popn[!is.na(popn$mothereduc),]) + geom_boxplot(aes(x = mothereduc, y = index, fill =
grade)) + ggtitle("Index vs Mothers's Education Level and Grade")
ggplot(data = popn[!is.na(popn$fathereduc),]) + geom_boxplot(aes(x = fathereduc, y = index, fill = grade))
+ ggtitle("Index vs Fathers's Education Level and Grade")
ggplot(data = popn[!is.na(popn$familywellof),]) + geom_boxplot(aes(x = familywellof, y = index, fill =
grade)) + ggtitle("Index vs Family Well Off and Grade")
ggplot(data = popn[!is.na(popn$senseofsafety),]) + geom_boxplot(aes(x = senseofsafety, y = index, fill =
grade)) + ggtitle("Index vs Sense of Safety and Grade")

# Addiction index of each gender with respect to BMI
ggplot(data = popn) + geom_point(aes(x = BMI, y = index, color = gender), size = 0.1, alpha = 1) + labs(x
= "BMI", y = "index") + ggtitle("BMI of Index vs Gender")
ggplot(data = popn) + geom_point(aes(x = BMI, y = index, color = grade), size = 0.1, alpha = 1) + labs(x =
"BMI", y = "index") + ggtitle("BMI of Index vs Grade")

# Multiple linear regression
svy0 = svydesign(id = ~1, weight = ~ student_weight, data = hs_resp)
lm <- svyglm(index ~ gender + grade + race + urbanicity + mothereduc + fathereduc + familywellof +
senseofsafety, design = svy0)
summary(lm) # Multiple linear regression model of the addiction index versus all categorical variables.

# Weight versus gender/race
svyboxplot(student_weight~gender, svy, main = "Student Weight vs Gender")
svyboxplot(student_weight~race, svy, main = "Student Weight vs Race")
```