Predicting Alcohol Use Behaviors in the United States: A Complex Survey Analysis

by

Benjamin Minardi V

BS, Mathematics, Applied Economics, Ursinus College, 2020

Submitted to the Graduate Faculty of the

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Benjamin Minardi V

It was defended on

April 25, 2022

and approved by

Ada Youk, PhD, Associate Professor and Vice Chair of Education, Department of Biostatistics

Christina Mair, PhD, MPH, Associate Professor, Department of Behavioral and Community Health Sciences

Jenna Carlson, PhD, Assistant Professor, Department of Biostatistics

Thesis Advisor: Jeanine Buchanich, MEd, MPH, PhD, Research Associate Professor and Vice Chair for Practice, Department of Biostatistics

Predicting Alcohol Use Behaviors in the United States: A Complex Survey Analysis

Benjamin Minardi V, MS

University of Pittsburgh, 2022

Introduction: Risky alcohol use behaviors create a heavy toll on the health of the United States population. While many studies have attempted to understand the true underpinnings of alcohol use, alcohol consumption and drinking behaviors are multifaceted issues that have both risk factors and consequences. This thesis intended to study the relationship between alcohol use behaviors and their potential predictors in order to answer the question: what influences best predict alcohol use behaviors in the general United States population?

Methods: Data from the National Health and Nutrition Examination Survey (2013-2018) were utilized to create and test models that predicted alcohol use behaviors. Logistic regression and classification and regression tree models were built with complex survey weighting to produce estimates generalizable to the population of the United States. Accuracy and receiver operating characteristic curves were used to assess prediction ability of the models.

Results: Age and sex were the strongest predictors of alcohol use behaviors. The final logistic regression model resulted in odds ratios of 0.971 per each one-year increase in age and 0.380 for females compared to males (p-value < 0.001 for each). Other statistically significant and marginally significant characteristics included being a college graduate, belonging to the Mexican American race/ethnicity group, living with a smoker, self-described health condition, score of a PHQ-9 depression screener, and marital status. Both the logistic regression and classification and regression tree models predicted alcohol use behaviors well with accuracies of 0.702 and 0.660, respectively.

Conclusions: Findings show that the covariates age, sex, education, race, smoking status, marital status, mental health, overall health, and living with a smoker are all important predictors of alcohol use behaviors. These results are generally consistent with the literature and provide evidence that advocates for further exploration of certain characteristics like living with a smoker.

Public Health Significance: Understanding these risk factors or potentially uncovering new risk factors holds a large public health impact. It would provide public health officials with intuition about where to direct research and where to apply interventions in attempt to reduce the burden of risky alcohol use behaviors on the health of the public.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

Between 2011 and 2015, unhealthy alcohol consumption caused approximately 95 thousand deaths per year in the United States, contributing 29 years of life lost each (Centers for Disease Control and Prevention, 2021). Not only can unhealthy drinking habits cause death, but it can also lead to a number of other serious issues. These potential issues range from physical disease to social or familial complications.

Drinking behaviors that cause such severe health risks can be defined in various ways and can differ between men and women. A moderate drinking level is defined by one drink a day for women and two drinks a day for men. Consequently, risky drinking is denoted by consumption of alcohol above the daily recommended value. There are other classifications of even more dangerous alcohol consumption behaviors. For example, the CDC describes binge drinking as four or five drinks in a single occasion for women and men, respectively (O'Connor, et al., 2018). Understanding risk factors related to unhealthy alcohol consumption is pertinent because it helps advise researchers and public health officials on how and where to properly intervene or introduce policy in order prevent future burden. The following characteristics as well as others suggested by institutional knowledge have been included in this study in attempt to predict alcohol use behaviors.

It is widely known that a person's demographics are highly associated with alcohol use behaviors. This relationship is well documented in the literature. Of all demographics, age and sex are some the most important predictors—not just in the United States but throughout the entire world. In general, alcohol consumption is higher for men as compared to women and it increases with age with a potentially quadratic relationship (Chaiyasong, et al., 2018). In addition, the

literature shows that individuals with low socioeconomic status and individuals that belong to racial minority groups experience more negative outcomes as a result alcohol use ranging from economic losses to alcohol dependency (Collins, 2016). While demographics are among the most important predictors of alcohol use, there are various other important variables to be explored as well.

Smoking status, depression, and anxiety are further examples of commonly accepted strong predictors of alcohol consumption. It is a generally known fact that smoking status and drinking behaviors are highly correlated. One study examined this relationship over a decade and found strong evidence linking drinking and smoking, especially for those who start smoking in adolescence (Paavola, Vartiainen, & Haukkala, 2004). This suggests that the age an individual starts drinking or smoking may be an important predictor of their behavior—not only if the individual drinks or smokes. Literature also cites a strong relationship between anxiety disorders, mental illness, and alcohol use disorders. However, even though the relationships are known to be strong, there is little evidence for a direction of the causality between the two (Morris, Stewart, & Ham, 2005). For any study of alcohol use, it is important that these characteristics be considered.

Additionally, some lesser established relationships between certain features and alcohol use are explored. This includes, but is not limited to, physical activity and peer substance use. Interestingly, multiple studies find a positive association with alcohol and physical activity. This means that the more an individual exercises or exerts some physical effort, the more alcohol they consumed on average (Piazza-Gardner & Berry, 2012). One study of women even found that drinking was associated with about a 10% increase in the probability of working out vigorously and those who drink work out approximately 10 minutes per day more than their counterparts (French, Popovici, & Maclean, 2009). Peer or familial substance use is also studied as a predictor

of alcohol use. Many of these studies focus on adolescence, however the results are overwhelmingly similar—people with substance using peers, whether it is drugs or alcohol, are more susceptible to alcohol use (Kelly, et al., 2012).

This paper aimed to study the relationship between unhealthy alcohol use behaviors and its potential risk factors in the general population of the United States. Cross-sectional data was used from the National Health and Nutrition Examination Survey to build generalizable models and test their prediction abilities. Three iterations of the survey were utilized ranging from 2013 to 2018. The major objective of this study was to be hypothesis generating. In other words, there was no one specific individual factor of interest, instead the goal was to identify multiple factors related to unhealthy alcohol use and those that predict it well. Understanding these risk factors or discovering new ones has a great public health impact since it would provide researchers and public health officials insight for future research and for appropriately placing interventions in attempt to lower the burden of alcohol use on the public.

2.0 Methods

2.1 National Health and Nutrition Examination Survey

The National Health and Nutrition Examination Survey (NHANES) is a cross-sectional survey conducted within the United States with the goal of evaluating the nation's health. It is composed of multiple parts, including an interview section and a physical examination section, that result in the measurement of various characteristics for each participant. The physical exam takes place in a mobile examination center that travels to selected counties. The survey is conducted yearly on a representative sample of approximately 5,000 individuals and the data are available to the public in the form of two year cycles. Specifically, NHANES is designed to be representative of the civilian, non-institutionalized subpopulation of the United States. NHANES helps the public health officials in the United States by determining the incidence of health conditions and health risk factors throughout the country (National Center for Health Statistics, 2017). Because NHANES utilizes a complex survey design, and data are not collected via simple random sample, for it to be properly representative of the United States population each observation must receive a survey weight when calculating statistics.

The sampling procedure takes place at four levels: the county level, the county segment level, the household level, and ultimately the individual level. The county level is the first and largest stage of the sampling process. Selected counties are referred to as primary sampling units (PSUs). For this first stage, PSUs with larger population counts receive a higher probability of being selected. The selected PSUs are broken up along preexisting divisions into segments, such as by census tract. Similar to the first stage, the county segments with larger populations receive

higher selection probability. Households, and subsequently individuals, are then randomly selected from the chosen county segments. The sampling process is illustrated in the figure below (National Center for Health Statistics, 2022).

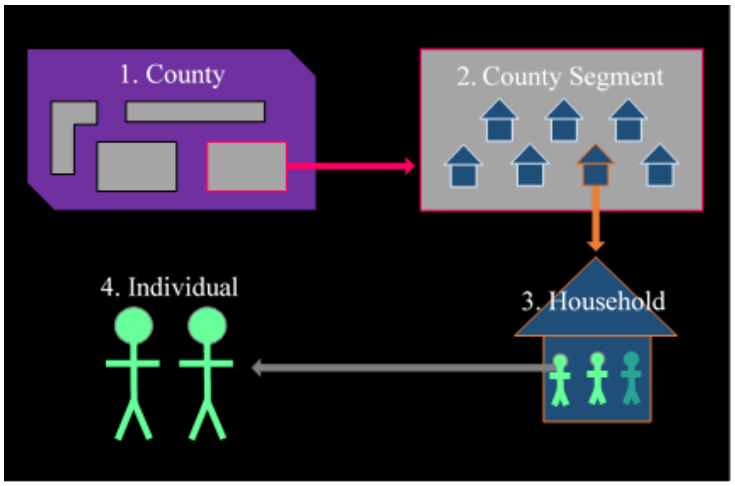


Figure 1 NHANES Sampling Procedure

Once the sampling is complete, each observation is assigned a sample weight. This weight is a variable that accounts for the number of people each observation represents. Sample weights are calculated by taking the inverse of the participant's probability of being selected, adjusted for nonresponse, survey completion, and the oversampling of certain demographics (National Center for Health Statistics, 2022). The weighting process results in two types of weights for each iteration of NHANES—one accounting for those who completed just the interview portion of the survey and one accounting for those who completed the survey and the mobile examination. For this analysis of alcohol use behaviors, the mobile examination weights were utilized since it was during the mobile physical exam that the alcohol related questions were asked of participants.

It is of note that NHANES oversamples certain demographics that literature and public health officials deem to be of interest. Populations that are oversampled differ year to year. This

oversampling allows for more reliable estimates within these specific subgroups and is accounted for when calculating sample weight values. In addition, in order to properly de-identify study participants, NHANES does not provide the true PSUs used in the original survey design. Instead, masked PSUs that closely represent the real PSUs and produce similar statistical estimates are provided (National Center for Health Statistics, 2022).

2.2 Data Management and Processing

2.2.1 Variable Definitions

The following is a list of 22 variables and their brief definitions that were utilized in this study. It is of note that the data collected through NHANES is mostly self-reported, except for a select few characteristics measured by physicians. This paper consists of primarily self-reported data. The list of 22 variables displayed in Table 1 includes two dependent variables about alcohol use behaviors and 20 independent variables. The two outcome variables (risky and drink number) were both created from the question, “During the past 12 months, on those days that you drank alcoholic beverages, on the average, how many drinks did you have?”.

The predictors were selected for this paper based on the following parameters: institutional knowledge, suggestion from literature, and sample size. Because the question of interest involves alcohol use, the data were subset to only include those age 20 to age 79. Moreover, the variable PHQ Score was created from the PHQ-9 depression screener (Kroenke, Spitzer, & Williams, 2001). While the PHQ-9 is not a depression diagnosis, it has excellent predictive capabilities of true depression. The PHQ-9 depression screener is a nine item questionnaire that gauges

depression symptoms. Each item has a Likert scale response ranging from zero to three where a higher score means worse symptoms. The results of the nine items are summed up to create a cumulative score between zero and 27. The cumulative score can be treated as dichotomous when specific cutoffs that are well documented in the literature are applied or it can be treated as continuous (Kroenke, Spitzer, & Williams, 2001). For this study, the PHQ-9 cumulative score was treated as a continuous covariate. The variables shown in Table 1 were cleaned from their original NHANES state to reflect their supplied definitions.

Table 1 Variable Definitions

Variable Name	Data Type	Brief Definition
Risky	Binary	Moderate (two or less drinks) or risky (3 or more)
DrinkNum	Continuous	Number of drinks on days that you drank alcohol
Age20to79	Continuous	$20 \leq \text{Age} \leq 79$
White	Binary	White or other race
Black	Binary	Black or other race
MexicanAm	Binary	Mexican American or other race
Sex	Binary	Male or female
College	Binary	College graduate or not
PHQScore	Continuous	Value 0 to 27 that assesses depression symptoms
Peer	Binary	Smoker in household (HH Smoker) or not
IncUnder20	Binary	Income under $20,000 or over
Married	Binary	Married or any other relationship status
Sed	Continuous	Typical time spent sedentarily in a day

Smoker	Binary	Being a smoker or a never smoker
Diabetes	Binary	Being diabetic/borderline diabetic or not
Insure	Binary	Having insurance or not
Health	Binary	Self-described fair/poor or good health
Home	Binary	Owning/buying a home or renting/other
Work	Binary	Working in some capacity or not working
Sleep	Continuous	Average hours of sleep during weekdays
BMI	Continuous	Body mass index
Children	Binary	Having no children or having at least one child

2.2.2 Survey Weighting

Prior to any analysis, it is required to apply the sample weights discussed above to create generalizable estimates. In order to properly apply these weights all observations must be retained in the dataset. Therefore, when the data were cleaned all coded missing values, top coded values, or unwanted observation values were kept in the dataset but converted to a missing value. If the observations had been removed from the dataset before applying weights, it would have created biased estimations. Once the dataset was appropriately cleaned, without dropping any observations, each participant was assigned its weight value as given by the mobile examination weight variable and the data were ready to be analyzed. Hence, the models are meant to be representative of the civilian, non-institutionalized, age 20 to 79, subpopulation of the United States. All data cleaning, weighting, and analyses were completed through the programming language R.

2.2.3 Training and Testing Sets

The data used to build every model comes from the 2017-2018 NHANES cycle. While the 2017-2018 cycle was weighted and used to build proper generalizable models, an unweighted testing set was also created solely to evaluate prediction ability. Model training data utilizes weighted NHANES data from the 2017-2018, and model testing data combines unweighted data from two NHANES iterations: 2013-2014 and 2015-2016. Variables across all three survey iterations were coded consistently, thus cleaned the same way. Since the testing data is not weighted, all missing and coded missing observations were subsequently removed leaving testing sample size of 6,269 observations.

2.3 Model Types and Prediction Assessment

2.3.1 Logistic Regression

Logistic regression was used to model the dichotomized outcomes of interest: whether an individual is a risky drinker or moderate drinker. Logistic regression is used when modeling dichotomous outcome variables, however predictors can be continuous or categorical. Linear regression is not appropriate with binary outcome variables because the estimates likely extend much past the upper and lower limits of one and zero, respectively (LaValley, 2008). Thus, logistic regression, where predicted values are sandwiched between one and zero, is used for binary dependent variables.

Logistic regression is a type of generalized linear model that utilizes a logit link and estimates log odds of an event. The below equation details the specification of a logistic regression with N covariates, modeling the probability that some variable Y equals one, $p_i = P(Y_i = 1)$, where variable Y only takes the values one and zero.

$$logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_N X_{Ni} + \varepsilon$$

Predicted log odds, odds ratios, and probabilities can all be calculated from logistic regression models. Logistic models have two key assumptions: linearity and independent observations (LaValley, 2008).

2.3.2 Classification and Regression Trees

Classification and regression tree (CART) models were also utilized in this paper to analyze drinking behaviors. CART models are a type of supervised statistical learning method used for prediction. In other words, CART models require the true outcome variable to be supplied in order for them to be properly trained. Typically, CART models can handle both binary and continuous outcome variables. However, for this study, a CART model was used to estimate only the continuous outcome—the average number of drinks a person consumed on days they drank alcohol. This is because the R package for weighted CART models can currently only work properly with continuous dependent variables (Toth, 2017). CART models recursively split the data and then fit regression models within each created partition. In order to classify predictions, CART models analyze trends in the data related to the true outcome of each observation and then split the data where the greatest differences occur in attempt to create the best grouping. While

CART models are one of the simplest statistical learning methods, they have very few assumptions and are extremely interpretable (Loh, 2011).

2.3.3 Accuracy and Receiver Operating Characteristic Curves

To assess prediction performance of the models, predictive accuracy and receiver operating characteristic (ROC) curves were used. Both of these methods require a model to be built on some training set and tested on a separate set of data that was not used in the construction of the model whatsoever. Accuracy is the humbler of the two methods. The predictive accuracy of a model is simply the number of correctly predicted classifications from a test set divided by the number of observations in the test set. Total observations can also be defined as the sum of all of the positive cases and all of the negative cases. More formally, accuracy is the sum of the true positive classifications and the true negative classifications over the total sample size. Similarly, it can also be interpreted as the ratio of correctly predicted observations to total observations. Accuracy is bounded between zero and one, where one, or 100%, represents the model correctly classifying every observation and zero, or 0%, represents the model incorrectly classifying every observation. Hence, the higher the predictive accuracy, the better the model has performed. The formula for predictive accuracy is as follows.

$$Accuracy = \frac{True\ Positive\ Classifications + True\ Negative\ Classifications}{Total\ Postive + Total\ Negative}$$

The values to calculate accuracy can be pulled directly from a confusion matrix which is illustrated in Figure 2. For example, the denominator in the accuracy formula—the number of total observations—is easily calculated by taking the sum of the two column totals. A confusion matrix

maps the true classifications from a testing set against the predicted classifications and is a useful tool in calculating accuracy and other similar types of measures (Fawcett, 2006).

		True Classification	
		Yes	No
Predicted Classification	Yes	True Positive	False Positive
	No	False Negative	True Negative
		Total Positive	Total Negative

Figure 2 Confusion Matrix

An ROC curve is a method that helps visualize the classification performance of a model and is frequently used for logistic regression. ROC curves take the predicted probabilities and true classifications of a test set, and then graphs the true positive rate on the vertical axis against the false positive rate on the horizontal axis for various cutoff points of the predicted probabilities. A cutoff point signifies where the predicted probabilities are cut and assigned to a specific group. For example, a cutoff point of 0.5 would signify everyone with a predicted probability of 0.5 and above would be assigned to one group, else they would be assigned to the other group. Both the false positive rate and the true positive rate can be calculated from a confusion matrix. Their formulas are detailed below (Fawcett, 2006).

$$True\ Positive\ Rate, \qquad TP = \frac{True\ Positive\ Classifications}{Total\ Positive}$$

$$False\ Positive\ Rate, \qquad FP = \frac{False\ Positive\ Classifications}{Total\ Negative}$$

Once the ROC curve is constructed, it is possible to summarize its prediction performance by calculating the area under the ROC curve (AUC). AUC will never be greater than one because it is bounded by the true positive rate and the false positive rate. However, it will also never be less than 0.5 since this value signifies the classification ability of randomly guessing, like flipping a coin. Similar to accuracy, higher AUC values mean better classification ability. Furthermore, when evaluating AUC, one should always use predicted probabilities and not predicted groups. This is because the idea behind ROC curves and AUC is to not use a single cutoff but all possible cutoffs. ROC curves detail the change in the true positive rate and the false positive rate as the cutoff goes from zero to one. Thus, supplying predicted groups would result in a fixed true positive rate and false positive rate, rendering AUC uninformative. Therefore, if only predicted groups are available, and predicted probabilities are not, it is more reasonable to measure predictive performance using accuracy.

2.4 Weighted Logistic Regression Model

2.4.1 Model Building

A purposeful selection method with the following steps was chosen to build the weighted logistic regression model that predicted the probability of exhibiting risky drinking behaviors. First, univariate models for each predictor were fit with the intention of including variables significant at the 0.20 alpha level. The multivariate model ran with all variables significant at the 0.20 alpha level had significant collinearity issues. In other words, the predictors included in the model were highly correlated with one another. This is an issue because collinear variables create

biased estimations and sometimes prevent the model from running altogether. To assess collinearity, variance inflation factors (VIFs) were calculated. Those with VIF greater than 10 were deemed problematic and removed. VIFs are a measure of the relationship between predictors—the higher the VIF, the more collinear that predictor is with the others. VIFs do not just measure the relationship of two predictors, instead it accounts for all possible linear combinations of the predictors included in the model.

Then variables with p-values greater than 0.5 were removed from the model. Once variables that were not statistically significant were removed from the multivariable model, the percent change in the coefficient estimates was assessed to see if the removal of the covariates created any meaningful change. Any percent change in coefficient estimate greater than 20% was considered a meaningful change. The model at this point is referred to as the preliminary main effects model.

With the preliminary main effects model identified, the functional form of age and an age-sex interaction term were separately assessed. The quadratic functional form of age was assessed because institutional knowledge suggests that alcohol use may have a nonlinear association with age. In other words, it is feasible to see the likelihood of risky drinking to increase with age at first, but then begin to decrease as age increases. Institutional knowledge also informed the assessment of the age-sex interaction. It is viable for the impact of age on alcohol use behaviors to vary by sex. Testing the preliminary main effects model for the above interaction and functional form of age resulted in the final weighted logistic regression model.

2.4.2 Model Diagnostics

Once the final weighted logistic regression model was fit, diagnostics were performed to assess model validity. First the assumptions were checked. The first assumption, independence among observations is met through the design of NHANES. The linearity assumption was verified using a Pearson residual plot.

Overall fit of the model was assessed using the Hosmer-Lemeshow Goodness-of-Fit Test. The null hypothesis for this test is that the model fits the data well and the alternative is that it does not. The hypotheses are assessed using a Chi-squared distribution. Collinearity was also assessed using VIFs as discussed above. Outliers were also evaluated using the same Pearson residual plot used to verify the linearity assumption. Finally, since there was a presence of potential outliers, a sensitivity analysis was completed by refitting the model only with observations with standardized Pearson residuals between two and negative two. The intent being, if the regression without the outliers yields similar estimates to the regression with the outliers, we can conclude that the model is not sensitive to these outlying observations.

2.4.3 Prediction Assessment

To assess the performance of the logistic regression model's ability to predict risky drinking behavior an ROC curve was created and AUC was measured. The ROC curve was created from predicted probabilities calculated by applying the model to a test data set which is discussed above. In addition, variable importance in predicting drinking behavior was deduced from the remaining covariates in the model following the purposeful selection method.

2.5 Weighted CART Model

2.5.1 Model Building

A weighted CART model was then fit using the covariates suggested by the logistic regression model building process. For this model, instead of predicting a dichotomous drinking status, the specific value of the continuous version of the outcome variable was predicted. Therefore, when applying the model to the test dataset, the result is not predicted probability of being in the risky group, but rather the predicted average number of drinks on days where alcohol was consumed. This result allows predicted groups to be assigned using a cutoff somewhere between two and three, inclusive. The cutoff must be between two and three because moderate drinking behavior is defined as two or less drinks and risky drinking behavior is defined as three or more. Thus, the cutoff for the continuous-discrete version of the outcome can only fall between two and three, inclusive.

2.5.2 Prediction Assessment

Because we have predicted groups and not predicted probabilities, ROC analysis is not ideal with the CART model utilized in this paper because predicted groups provide a fixed sensitivity and specificity as discussed above. Therefore, the best way to analyze model performance is to calculate accuracy for various cutoff points between two and three. In addition, variable importance was inferred from the covariates ultimately selected by the final weighted CART model.

2.6 Comparison of Model Performance and Variable Importance

Because an ROC analysis was only appropriate for the logistic regression model, predictive accuracy was used to compare the two final models. For each model, the best possible accuracy was assessed and used to compare model performance. The best possible accuracy was found by creating plots that graphed cutoff points against the resulting accuracy. In addition, variables selected from each model and their impact on prediction was used to assess variable importance in predicting alcohol use behaviors.

3.0 Results

3.1 Weighted Logistic Regression Model

3.1.1 Univariate Models

Simple weighted logistic regression models were fit to the dichotomized outcome of interest, whether an individual exhibits moderate or risky drinking behaviors on average, using the 20 predictors listed in Table 2. Of the 20 models fit, 19 were statistically significant at the 0.2 alpha level. The dummy variable for being for being white was the only covariate not significant at the 0.2 alpha level. BMI and time sedentary were the only variables not significant at the 0.05 alpha level.

Table 2 Univariate Models

Covariate	Group	Coefficient	95% Lower CL	95% Upper CL	p-value
Age (Age20to79)	(Continuous)	-0.032	-0.039	-0.025	< 0.001*
Sex	Female	-0.946	-1.132	-0.761	< 0.001*
White	White	-0.121	-0.335	0.093	0.290
Black	Black	-0.288	-0.510	-0.066	0.026**
Mexican Am.	Mexican Am.	0.733	0.466	0.999	< 0.001*
College Graduate	Yes	-0.778	-0.977	-0.579	< 0.001*
PHQ Score	(Continuous)	0.044	0.021	0.067	0.002*

HH Smoker	Yes	0.880	0.635	1.125	< 0.001*
Income	≥ $20,000/year	-0.401	-0.704	-0.097	0.021*
Marital Status	Married	-0.704	-0.990	-0.419	< 0.001*
Time Sedentary	(Continuous)	-0.001	-0.001	0.000	0.085**
Smoking Status	Smoker	0.788	0.621	0.955	< 0.001*
Diabetic Status	Diabetic	-0.376	-0.709	-0.043	0.044*
Insurance Status	Insured	-0.883	-1.104	-0.663	< 0.001**
General Health	Fair or Poor	0.433	0.235	0.630	0.001*
Home Ownership	Owned	-0.523	-0.771	-0.275	0.001*
Work Status	Working	0.375	0.104	0.645	0.017**
Avg. Sleep Time	(Continuous)	-0.072	-0.119	-0.024	0.010**
BMI	(Continuous)	0.008	-0.003	0.019	0.174**
Have Children	Yes	0.338	0.093	0.584	0.017*

*Significant at the 0.2 alpha level **Significant at 0.2 alpha level but removed from first model for collinearity issues

3.1.2 Initial Multivariable Model

Covariates simply significant at the 0.2 alpha level were included in an initial multivariable logistic regression model also predicting the dichotomous outcome of interest. However, because of collinearity issues as measured by VIFs, six predictors had to be removed from this first multivariable model. These covariates were BMI, average sleep time, insurance status, Black, work status, and sedentary time. Thirteen predictors were included the initial multivariable model which utilized 2,907 observations—the model is detailed in Table 3. Once the collinear variables were removed, each predictor had a VIF of less than 10. The variables age, sex, Mexican

American, and smoking status had significant log odds ratio coefficient estimates. The smallest p-value was observed for the sex covariate with a p-value of 0.021. Multiple covariates, including income, having any children, home ownership status, and diabetic status, had insignificant log odds ratio coefficient estimates with p-values > 0.5.

Table 3 Initial Multivariable Model

Covariate	Group	Coefficient	SE	t	p-value	VIF
Intercept	N/A	0.514	0.287	1.788	0.216	N/A
Age (Age20to79)	(Continuous)	-0.029	0.004	-6.407	0.024*	3.202
Sex	Female	-0.938	0.137	-6.863	0.021*	1.974
Mexican Am.	Mexican Am.	0.712	0.165	4.318	0.050*	1.905
College Graduate	Yes	-0.345	0.118	-2.936	0.099	1.763
PHQ Score	(Continuous)	0.021	0.017	1.219	0.347	3.122
HH Smoker	Yes	0.442	0.158	2.792	0.108	4.416
Income	≥ $20,000/year	0.114	0.168	0.680	0.566	4.628
Marital Status	Married	-0.386	0.180	-2.139	0.166	6.010
Smoking Status	Smoker	0.694	0.121	5.730	0.029*	6.813
Diabetic Status	Diabetic	-0.140	0.229	-0.612	0.603	4.048
General Health	Fair or Poor	0.213	0.132	1.609	0.249	4.073
Home Ownership	Owned	0.053	0.157	0.336	0.769	3.828
Have Children	Yes	0.004	0.164	0.023	0.984	3.741

*Significant at the 0.05 alpha level

3.1.3 Covariate Removal

Covariates with the highest p-values were removed one at a time as suggested by the purposeful selection method. This resulted in the removal of four variables: income, having any children, home ownership status, and diabetic status to form the preliminary main effects model—the preliminary main effects model is detailed in Table 4. Changes in coefficient estimates were assessed before and after the covariate removal. None of the remaining nine covariates' coefficient estimates had a meaningful point estimate change (i.e. no coefficient estimate changed by greater than 10%). VIFs only decreased as more covariates were removed.

Table 4 Preliminary Main Effects Model

Covariate	Group	Coefficient	OR	SE	t	p-value
Intercept	N/A	0.704	2.023	0.219	3.218	0.018*
Age (Age20to79)	(Continuous)	-0.030	0.971	0.003	-9.072	< 0.001*
Sex	Female	-0.968	0.380	0.135	-7.177	< 0.001*
Mexican Am.	Mexican Am.	0.684	1.982	0.162	4.211	0.006*
College Graduate	Yes	-0.331	0.718	0.118	-2.809	0.031*
PHQ Score	(Continuous)	0.021	1.022	0.016	1.295	0.243
HH Smoker	Yes	0.433	1.542	0.150	2.885	0.028*
Marital Status	Married	-0.382	0.683	0.171	-2.230	0.067
Smoking Status	Smoker	0.666	1.947	0.117	5.701	0.001*
General Health	Fair or Poor	0.228	1.257	0.095	2.395	0.054

*Significant at the 0.05 alpha level

3.1.4 Functional Form of Age and Age-Sex Interaction Assessment

To assess if age should be included with a quadratic functional form, a squared age term was added to the preliminary main effects model. The coefficient estimate on the linear age term and on the squared age term were not statistically significant with p-values of 0.223 and 0.055, respectively. The coefficient estimate for the squared age term was very small at -0.0008. Thus, treating age as linear is more appropriate. Similarly, an interaction term between age and sex was included in the preliminary main effects model to test if the effect of age on drinking status varied by sex. The coefficient estimate for the interaction term was equal to 0.991 with a p-value of 0.207, and the inclusion of this interaction term caused the sex coefficient estimate to become not statistically significant at the 0.05 alpha level (p-value = 0.099). With no relevant functional forms or interactions found significant, the model specified in Table 4—the preliminary main effects model—was used as the potential final model to be assessed for validity and checked for potential issues.

3.1.5 Model Diagnostics

Several diagnostics were used to assess model validity and assumptions. A Hosmer-Lemeshow Goodness-of-Fit Test was used to assess how well the model fit the data ($\chi^2 = 10.670$, p-value = 0.221). With p-value = 0.221, there is not sufficient evidence to reject the null hypothesis that the model fits the data well. There were no VIFs higher than 2.5. To evaluate linearity and the presence of outliers, a standardized Pearson residual plot was created. The data appear to be generally linear with a handful of outliers present.

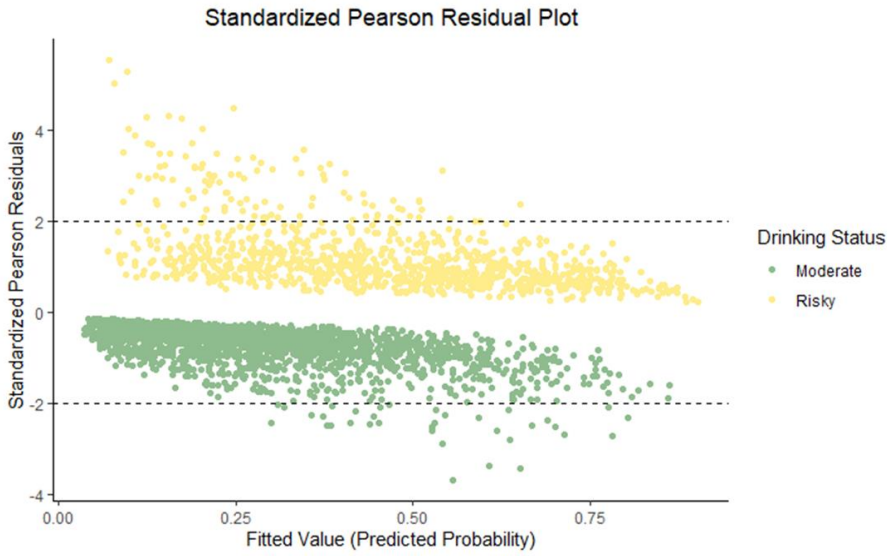


Figure 3 Standardized Pearson Residual Plot

3.1.6 Sensitivity Analysis

As a sensitivity analysis, the model was refit with only the observations with standardized Pearson residuals between two and negative two. The resulting model removed 147 outliers as defined by the Pearson residuals, and is detailed in Table 5. There were no meaningful changes in the coefficient estimates nor the p-values.

Table 5 Sensitiviy Analysis Model

Covariate	Group	Coefficient	OR	SE	t	p-value
Intercept	N/A	0.718	2.050	0.227	3.163	0.019*
Age (Age20to79)	(Continuous)	-0.031	0.970	0.003	-9.584	< 0.001*
Sex	Female	-0.931	0.394	0.135	-6.903	< 0.001*
Mexican Am.	Mexican Am.	0.720	2.055	0.161	4.465	0.004*

College Graduate	Yes	-0.386	0.680	0.137	-2.812	0.031*
PHQ Score	(Continuous)	0.026	1.026	0.017	1.557	0.171
HH Smoker	Yes	0.433	1.542	0.148	2.919	0.027*
Marital Status	Married	-0.396	0.673	0.160	-2.480	0.048*
Smoking Status	Smoker	0.672	1.958	0.116	5.814	0.001*
General Health	Fair or Poor	0.233	1.262	0.105	2.220	0.068

*Significant at the 0.05 alpha level

3.1.7 Final Model and Prediction Assessment

Because the model specified in Table 4 did not require additional functional forms nor interactions terms and was not sensitive to outliers, it was used as the final weighted logistic regression model. Using this model, an ROC curve was generated resulting in an AUC of 0.744 and can be seen in Figure 4.

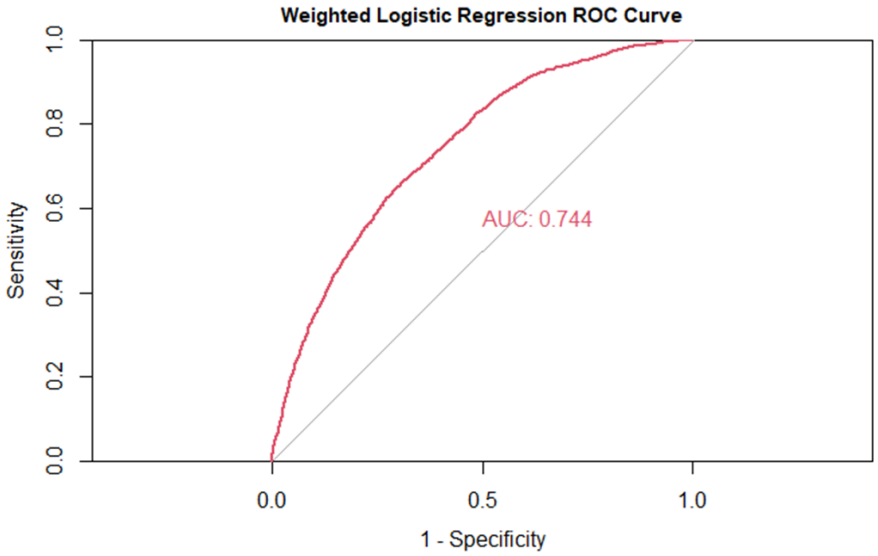


Figure 4 Logistic Regression ROC Curve (AUC = 0.744)

Consequently, to acquire the best possible accuracy, the cutoffs between zero and one were graphed against the resulting accuracy. This graph for the weighted logistic regression model is visible in Figure 5. The cutoff yielding the highest accuracy for this model was 0.49—any predicted probability greater than or equal to 0.49 was classified as someone predicted to exhibit risky drinking behaviors. This best cutoff resulted in an accuracy of 0.702.

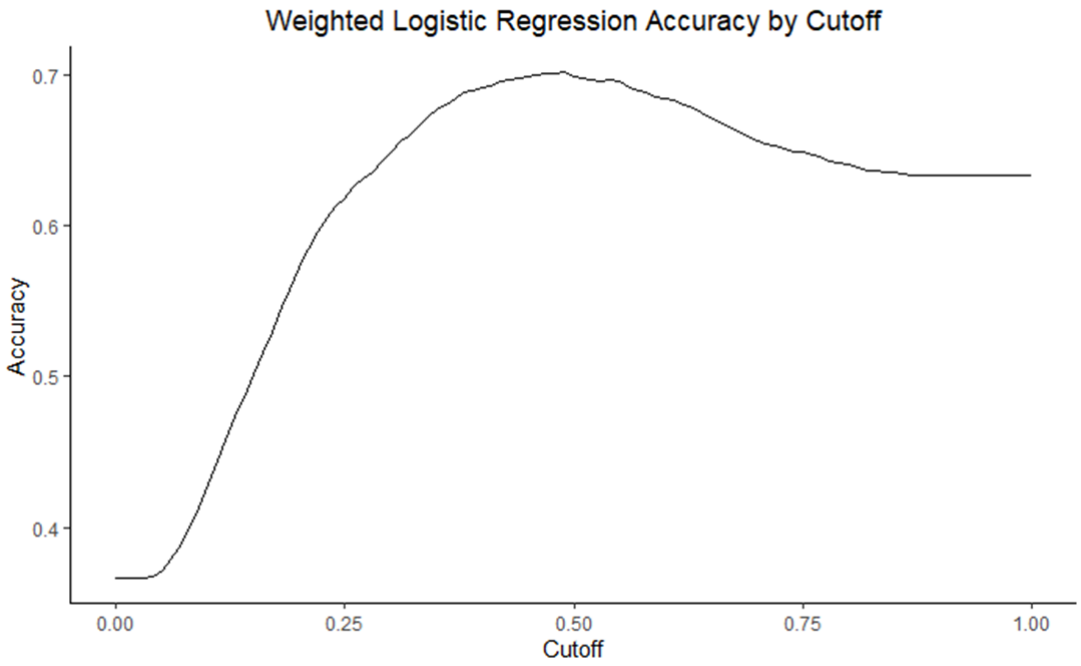


Figure 5 Logistic Regression Accuracy vs. Cutoff

3.2 Weighted CART Model

3.2.1 Model Building

With the variables suggested by the weighted logistic regression model, a CART model was fit. Figure 6 displays the resulting model in tree form. Sex is the primary split of the model. Age appears to be more important for females since it appears twice in that branch. In addition, age is the only variable to appear in each sex branch. Marital status and exposure to secondhand smoke through living with a smoker only appear in the male branch, where smoking status and PHQ-9 score only appear in the female branch. Unmarried men with smokers living in the household have the highest average drinking estimate of 4.24 drinks on average. Woman aged 54.5 and older that do not smoke have the lowest estimate with 1.39 drinks on average. The model completely removed health status, education, and race/ethnicity from its decision making and predictions.

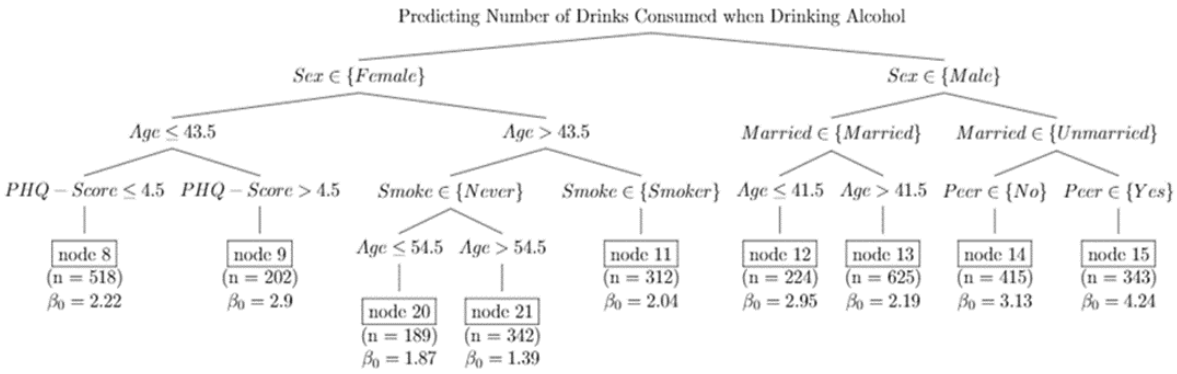


Figure 6 Weighted CART Model

3.2.2 Prediction Assessment

Accuracy of predicting the correct group was calculated. The accuracy by cutoff graph was generated for the CART model between the valid cut points of two and three. This graph for the weighted CART model is displayed in Figure 7. Due to the nature of CART models this graph is less smooth than its logistic regression counterpart. The cutoff yielding the highest accuracy for this model was exactly 3 drinks—any predicted number of drinks greater than or equal to 3 drinks was classified as an individual predicted to exhibit risky drinking behaviors. This best cutoff resulted in an accuracy of 0.660.

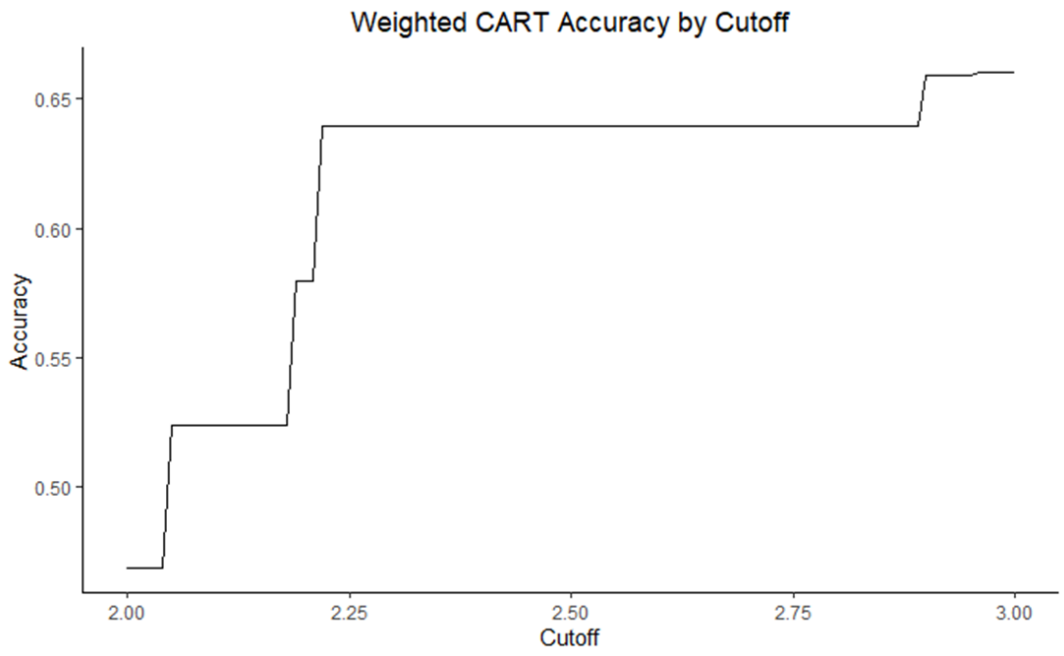


Figure 7 CART Accuracy vs. Cutoff

3.3 Comparison of Model Performance and Variable Importance

The final weighted logistic regression model ultimately used three more predictors than the final CART model, and its largest possible accuracy was 0.702, or 70.2%. The weighted CART model did not include health status, education level, and race/ethnicity, and had a largest possible accuracy of 0.660, or 66.0%. While the models performed similarly at predicting alcohol use behaviors, the final weighted logistic regression accuracy slightly outperformed the weighted CART model.

4.0 Discussion

4.1 Interpretation of Results

The goal of this nationally representative cross-sectional analysis was to identify relationships between unhealthy alcohol use behaviors and its potential risk factors. This objective is hypothesis generating in nature with not one, but many, characteristics explored for an association with alcohol use behaviors. The results show that the covariates age, sex, education, race, smoking status, marital status, mental health, overall health, and living with a smoker were all found to be significant predictors of unhealthy alcohol use. These results corroborated many findings within the literature as well as provided evidence that suggests further exploration of certain characteristics, for example, living with a household smoker—which could be considered a proxy estimate to having substance using peers.

Sex is clearly an important predictor throughout both of the models presented in this paper. Sex represents the primary split in the final CART model, with men being associated with generally higher predicted number of drinks and women being associated with generally lower number of drinks. Moreover, the final logistic regression model estimated the odds ratio of risky drinking for females to be 0.380 (p-value < 0.001). This means that the odds of being a risky drinker for men is 2.63 times higher than the odds of being a risky drinker for women, controlling for the other covariates in the model. The level of significance and direction of the association of the effect of sex on drinking behaviors are consistent with the findings in literature.

Education and age were also important predictors of alcohol use behaviors. Age was treated as a continuous variable where education was dichotomous. The two education groups were those

who graduated from college and those who did not. Both education and age had statistically significant odds ratios of 0.718 (p-value = 0.031) and 0.971 (p-value < 0.001), respectively. While these significant odds ratios are both consistent with the literature, they are both in the opposite direction that the literature primarily cites. In the literature, age is usually cited to have a differential effect on alcohol consumption as it changes but increase linearly, on average, and those with higher levels of education are usually expected to drink more, on average. For age, this difference may be observed because age was looked at as a whole from age 20 to age 79. Many studies either bin age or look at subgroups because the effect of age on alcohol use varies. Therefore, if age were binned instead of being used as continuous there would potentially be different trends. For education, this difference may be because how the variable is dichotomously defined. The education group that the odds ratio is estimated for only included those who have finished college, thus it excludes people who are currently in college and it is comprised of an older population on average.

In addition, smoking status and the Mexican American race/ethnicity group were also significant predictors for alcohol use behaviors with odds ratios of 1.947 (p-value = 0.001) and 1.982 (p-value = 0.006), respectively. Therefore, the odds of being a risky drinker is 1.947 times higher for smokers than non-smokers and 1.982 times higher for Mexican Americans than another race, controlling for all other covariates in the model. Even though the final CART model did not select race as a covariate to make its prediction, this does not exclude it from being an important predictor of alcohol use behaviors. The direction and significance of these covariates are consistent with the literature.

Furthermore, literature suggests the peer substance use may be a strong predictor of alcohol use behavior. In this study, living in a household with someone who smokes was viewed as a way

to approximate the relationship of having substance using peers and an individual's alcohol use behaviors. Living with a smoker was associated with the highest amounts of average alcohol consumption in the CART model and had a significant odds ratio of 1.542 (p-value = 0.028). This means that the odds of being a risky drinker for those who live with someone who smokes is 1.542 times higher than the odds of being a risky drinker for those who do not, controlling for the other covariates in the model. While these results agree with the literature, peer substance use is measured in various ways throughout it. This result provides evidence that the relationship of alcohol use behaviors and having substance using peers should be further investigated. Future research may want to use a more comprehensive measure of substance use in order to appropriately understand the true relationship among these characteristics.

4.2 Strengths and Limitations

One of the major strengths of this analysis was the ability to use a complex survey design and weight the data to properly represent the civilian, non-institutionalized subpopulation of the United States. Thus, all estimates were generalizable with a high external validity. In addition, statistical learning techniques, such as CART modeling, are seldom used when complex survey weights are required. This analysis demonstrates how complex survey weights can be used in such statistical learning analyses.

Conversely, this analysis does have several limitations. One of the major limitations was variable availability. NHANES collects a wide variety of data, but the participants do not have to respond to every question. Consequently, more controversial questions, such as questions related to drug and alcohol use, tend to have lower response rates. There were many other covariates that

would have been interesting to explore in this analysis, however their inclusion would cause the sample size to approach zero and the results to become less meaningful. Similar to this, certain outcome variables were not able to be explored because of lack of responses and extreme imbalance among those who did respond. It would be wise for future research to investigate the association of the characteristics included in this study and a more dangerous form of drinking, like binge drinking. Although that was not possible for this analysis because of low response rates and high data imbalance in those variables. The dichotomous outcome used in this analysis—whether a participant was on average a risky or moderate drinker—had a high response rate relative to other alcohol related questions and was balanced between the two groups. This allowed for the results to be much more reliable.

Furthermore, it is of note that this analysis was cross-sectional. Therefore, there is a higher chance that the results may possibly be an artefact and not actually true. Literature also suggests that when an individual begins drinking can impact their current alcohol use behaviors. However, NHANES did not collect this type of data. Hence, future research should consider longitudinal data that accounts for when alcohol or substance consumption began.

Other limitations include how age was treated in the analysis and the unweighted testing set. It is common in the literature for researchers to solely look at age subgroups rather than everyone as a whole—using age bins may have helped better determine the relationship of age and the outcomes. In this analysis, a quadratic functional form of age was assessed and deemed inappropriate for the data, however, age bins may have been more suitable since they provide a more flexible functional form. Moreover, the testing set used to assess prediction ability was unweighted, unlike the data used to build out the model. While this does not affect the generalizability of the estimates, it possibly introduces bias into the prediction assessment.

4.3 Public Health Significance

The overall goal of this thesis was to find risk factors of unhealthy alcohol use behaviors. The resulting hypothesis generating analysis showed that the covariates age, sex, education, race, smoking status, marital status, mental health, overall health, and living with a smoker are all potentially important in predicting alcohol use behaviors. These findings provide researchers and public health officials with information about risk factors associated with unhealthy drinking behaviors that can be used to guide interventions and future research in attempt to reduce the burden of unhealthy and risky alcohol use on the health of the public.

Appendix A Survey Weighted Descriptive Statistics

In this appendix, weighted descriptive statistics of characteristics used in this thesis are displayed. The statistics were created strictly from the survey weighted training dataset which consisted of the 2017-2018 NHANES iteration. Appendix Table 1 displays the survey weighted mean and corresponding standard errors for the continuous covariates in this analysis. The mean number of drinks was 2.48 which falls directly in the middle of the moderate drinking and risky drinking threshold. The average participant was 46.77 years old, has a moderately low score on the PHQ-9 depression screener, is sedentary for just under six hours a day, and sleeps about 7.62 hours on weeknights.

Appendix Table 1 Numerical Summary for Continuous Variables

Covariate	Mean	SE
Drink Number	2.48	0.0807
Age	46.77	0.5436
PHQ Score	3.16	0.0780
Time Sedentary	350.90	6.7182
Avg. Sleep Time	7.62	0.0336

Appendix Table 2 displays the proportion of the specified group for each dichotomous covariate in this study. These weighted proportions can be used to check if the sample used in the study is representative of the true national proportions. The weighted sample consisted of 33.22% risky drinkers and 66.78% moderate drinkers. In addition, 51.12% was female, 30.86% were

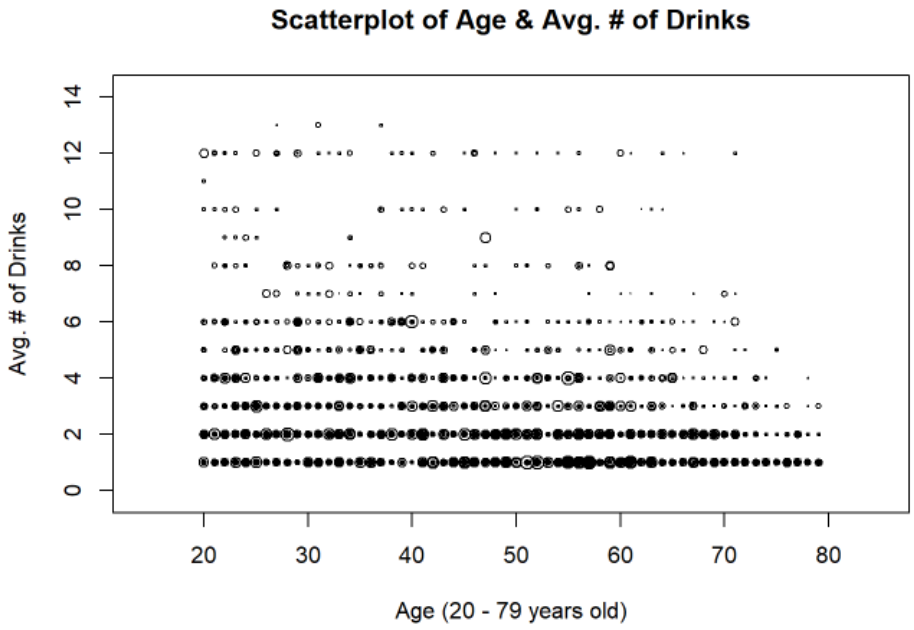
college graduates, and 53.20% were married. Only 30.01% lived in a house with someone who smoked cigarettes and only 15.24% considered their overall health to be just fair or poor.

Appendix Table 2 Numerical Summary for Dichotomous Variables

Covariate	Group	Proportion
Drinking Status	Risky	0.3322
Sex	Female	0.5112
White	White	0.5913
Black	Black	0.1185
Mexican Am.	Mexican Am.	0.1084
College Graduate	Yes	0.3086
HH Smoker	Yes	0.3001
Income	≥ $20,000/year	0.8689
Marital Status	Married	0.5320
Smoking Status	Smoker	0.4113
Diabetic Status	Diabetic	0.1091
Insurance Status	Insured	0.8824
General Health	Fair or Poor	0.1524
Work Status	Working	0.5943
Have Children	Yes	0.2536

Appendix Figure 1 shows a scatterplot graphing on of the most important predictors, continuous age, against the continuous outcome variable, average number of drinks consumed. The larger the point on the graph is, the higher the weight it received in the analysis. In other

words, larger points represent more people and smaller points represent less people. Note that both variables are recorded as integers in the NHANES data collection process, thus losing some of the information when assessing an overall trend.



Appendix Figure 1 Weighted Scatterplot of Age and Average Number of Drinks

Appendix B Analysis R Code

Included below is the R code used to analyze the data presented in this thesis. This code can be copied and pasted directly into an R markdown script.

````
# Load Libraries

```{r load_libraries, message=FALSE, include=FALSE}
library(tidyverse)
library(jtools)
library(SASxport)
library(gridExtra)
library(lsmeans)
library(expss)
library(mlbench)
library(rpms)
library(pROC)
library(haven)
library(survey)
library(caret)
library(performance)
```

# Data

## Reading in Data

```{r}
alc <- read.xport("...File Path.../ALQ_J.XPT")
demo <- read.xport("...File Path.../DEMO_J.XPT")
bmi <- read.xport("...File Path.../BMX_J.XPT")
diabetes <- read.xport("...File Path.../DIQ_J.XPT")
phq <- read.xport("...File Path.../DPQ_J.XPT")
insurance <- read.xport("...File Path.../HIQ_J.XPT")
house <- read.xport("...File Path.../HOQ_J.XPT")
health <- read.xport("...File Path.../HUQ_J.XPT")
job <- read.xport("...File Path.../OCQ_J.XPT")
activity <- read.xport("...File Path.../PAQ_J.XPT")
sleep <- read.xport("...File Path.../SLQ_J.XPT")
smoking <- read.xport("...File Path.../SMQ_J.XPT")
peers <- read.xport("...File Path.../SMQFAM_J.XPT")
```

## Joining Data

We use full join as to not drop any observations.


```{r}
# First I'm joining pairs of data

join1.1 <- full_join(alc, activity, by=c("SEQN"))
join1.2 <- full_join(bmi, demo, by=c("SEQN"))
join1.3 <- full_join(diabetes, health, by=c("SEQN"))
join1.4 <- full_join(house, insurance, by=c("SEQN"))
````
````

```
join1.5 <- full_join(job, peers, by=c("SEQN"))
join1.6 <- full_join(phq, sleep, by=c("SEQN"))

### Note: leaving smoking out till later to keep joins even numbers

# Now joining the joined pairs

join2.1 <- full_join(join1.1, join1.2, by=c("SEQN"))
join2.2 <- full_join(join1.3, join1.4, by=c("SEQN"))
join2.3 <- full_join(join1.5, join1.6, by=c("SEQN"))

# Joining 2.1 and 2.2 to create a semi-final dataset

join3.1 <- full_join(join2.1, join2.2, by=c("SEQN"))

# Joining 2.3 with smoking to create the other semifinal dataset

join3.2 <- full_join(join2.3, smoking, by=c("SEQN"))

# joining the previous two steps (3.1 & 3.2) to create the full dataset

data <- full_join(join3.1, join3.2, by=c("SEQN"))
```

Selecting Variables

```{r}
analysis <- data %>% select(SEQN, WTMEC2YR, SDMVPSU, SDMVSTRA,  ALQ130,
                            RIDAGEYR, RIDRETH1, RIAGENDR, DMDEDUC2, INDHHIN2,
                            SMD460, PAD680, SMQ020, DIQ010, HIQ011, HUQ010, DMDMARTL,
                            OCD150, SLD012, BMXBMI, DMDHHSIZ, DMDHHSZA, HOQ065,
                            DPQ010,  DPQ020,  DPQ030,  DPQ040,  DPQ050,  DPQ060,  DPQ070,
DPQ080,
                            DPQ090)
```

Data Cleaning

Outcome: Number of Drinks

```{r}
analysis$Risky <- ifelse(analysis$ALQ130 == 1 | analysis$ALQ130 == 2, "Moderate",
                         ifelse(analysis$ALQ130 == 777 | analysis$ALQ130 == 999, NA,
                                "Risky"))
analysis$DrinkNum <- ifelse(analysis$ALQ130 < 15, analysis$ALQ130, NA)
```

```{r}
analysis <- analysis %>% mutate(Risky = factor(Risky, levels=c("Moderate", "Risky"),
                                               labels=c("Moderate", "Risky")))
```

```{r}
table(analysis$ALQ130)
table(analysis$DrinkNum)
table(analysis$Risky)
```

Age

```{r}
table(analysis$RIDAGEYR)
```

```{r}
analysis$Age20to79 <- ifelse(analysis$RIDAGEYR >= 20 & analysis$RIDAGEYR <= 79,
                             analysis$RIDAGEYR, NA)
```

```{r}
table(analysis$Age20to79)
```

Race

```{r}
table(analysis$RIDRETH1)
```

```{r}
analysis$white <- ifelse(analysis$RIDRETH1 == 3, "White", "Other Race")

analysis$black <- ifelse(analysis$RIDRETH1 == 4, "Black", "Other Race")

analysis$mexicanAm <- ifelse(analysis$RIDRETH1 == 1, "Mexican American", "Other Race")
```

```{r}
analysis <- analysis %>% mutate(white = factor(white, levels=c("Other Race", "White"),
                                               labels=c("Other Race", "White")))

analysis <- analysis %>% mutate(black = factor(black, levels=c("Other Race", "Black"),
                                               labels=c("Other Race", "Black")))

analysis <- analysis %>% mutate(mexicanAm = factor(mexicanAm,
                                                   levels=c("Other Race",
                                                            "Mexican American"),
                                                   labels=c("Other Race",
                                                            "Mexican American")))
```

Sex

```{r}
analysis$sex <- analysis$RIAGENDR

analysis <- analysis %>% mutate(sex = factor(sex, levels=c(1, 2),
                                             labels=c("Male", "Female")))
```

Education

```{r}
table(analysis$DMDEDUC2)
```

```{r}
analysis$college <- ifelse(analysis$DMDEDUC2 == 5, "College grad",
                           ifelse(analysis$DMDEDUC2 == 7 | analysis$DMDEDUC2 == 9, NA,
                                  "Not college grad"))

analysis <- analysis %>% mutate(college = factor(college, levels=c("Not college grad",
                                                                   "College grad"),
                                                 labels=c("Not college grad",
                                                          "College grad")))
```

```{r}
table(analysis$college)
```

PHQ-9 Screener

```{r}
analysis$score1 <- ifelse(analysis$DPQ010 == 7 | analysis$DPQ010 == 9, NA,
analysis$DPQ010)
analysis$score2 <- ifelse(analysis$DPQ020 == 7 | analysis$DPQ020 == 9, NA,
analysis$DPQ020)
analysis$score3 <- ifelse(analysis$DPQ030 == 7 | analysis$DPQ030 == 9, NA,
analysis$DPQ030)
analysis$score4 <- ifelse(analysis$DPQ040 == 7 | analysis$DPQ040 == 9, NA,
analysis$DPQ040)
analysis$score5 <- ifelse(analysis$DPQ050 == 7 | analysis$DPQ050 == 9, NA,
analysis$DPQ050)
analysis$score6 <- ifelse(analysis$DPQ060 == 7 | analysis$DPQ060 == 9, NA,
analysis$DPQ060)
analysis$score7 <- ifelse(analysis$DPQ070 == 7 | analysis$DPQ070 == 9, NA,
analysis$DPQ070)
analysis$score8 <- ifelse(analysis$DPQ080 == 7 | analysis$DPQ080 == 9, NA,
analysis$DPQ080)
analysis$score9 <- ifelse(analysis$DPQ090 == 7 | analysis$DPQ090 == 9, NA,
analysis$DPQ090)
```

```{r}
analysis$phqScore <- analysis$score1 + analysis$score2 + analysis$score3 +
analysis$score4 + analysis$score5 + analysis$score6 + analysis$score7 + analysis$score8
+ analysis$score9
```

```{r}
table(analysis$phqScore)
```

Peer Substance Use

```{r}
table(analysis$SMD460)
```

```{r}
analysis$peer <- ifelse(analysis$SMD460 == 1 | analysis$SMD460 == 2 | analysis$SMD460
== 3,
                        "Yes",
                        ifelse(analysis$SMD460 == 0, "No", NA))

analysis <- analysis %>% mutate(peer = factor(peer, levels=c("No", "Yes"),
                                              labels=c("No", "Yes")))

table(analysis$peer)
```

Income

```{r}
table(analysis$INDHHIN2)
```

```{r}
analysis$incUnder20 <- ifelse(analysis$INDHHIN2 <= 4 | analysis$INDHHIN2 == 13,
                              "Less than 20k",
                              ifelse(analysis$INDHHIN2 == 77 | analysis$INDHHIN2 == 99,
NA,
                                     "20k or more"))

analysis <- analysis %>% mutate(incUnder20 = factor(incUnder20, levels=c("Less than
20k",
                                                                         "20k       or
more"),
                                                    labels=c("Less than 20k",
                                                             "20k or more")))
```

```{r}
table(analysis$incUnder20)
```

Married

```{r}
table(analysis$DMDMARTL)
```

```{r}
analysis$married <- ifelse(analysis$DMDMARTL == 1, "Married",
                           ifelse(analysis$DMDMARTL == 77 | analysis$DMDMARTL == 99, NA,
                                  "Not Married"))

analysis <- analysis %>% mutate(married = factor(married, levels=c("Not Married",
                                                                   "Married"),
                                labels=c("Not Married", "Married")))
```

```{r}
table(analysis$married)
```

Sedintary Time

```{r}
table(analysis$PAD680)
```

```{r}
analysis$sed <- ifelse(analysis$PAD680 == 7777 | analysis$PAD680 == 9999, NA,
                       analysis$PAD680)
```

Smoking

```{r}
table(analysis$SMQ020)
```

```{r}
analysis$smoke <- ifelse(analysis$SMQ020 == 7 | analysis$SMQ020 == 9, NA,
                         ifelse(analysis$SMQ020 == 1, "Smoker",
                                "Never Smoker"))

analysis <- analysis %>% mutate(smoke = factor(smoke, levels=c("Never Smoker",
"Smoker"),
```

```
                                labels=c("Never Smoker", "Smoker")))
```

```{r}
table(analysis$smoke)
```

Diabetes

```{r}
table(analysis$DIQ010)
```

```{r}
analysis$diabetes <- ifelse(analysis$DIQ010 == 1 | analysis$DIQ010 == 3, "Yes",
                            ifelse(analysis$DIQ010 == 2, "No", NA))

analysis <- analysis %>% mutate(diabetes = factor(diabetes, levels=c("No", "Yes"),
                                labels=c("No", "Yes")))
```

```{r}
table(analysis$diabetes)
```

Health Insurance

```{r}
table(analysis$HIQ011)
```

```{r}
analysis$insure <- ifelse(analysis$HIQ011 == 1, "Yes",
                          ifelse(analysis$HIQ011 == 2, "No", NA))

analysis <- analysis %>% mutate(insure = factor(insure, levels=c("No", "Yes"),
                                labels=c("No", "Yes")))
```

```{r}
table(analysis$insure)
```

General Health Condition

```{r}
table(analysis$HUQ010)
```

```{r}
analysis$health <- ifelse(analysis$HUQ010 == 1 | analysis$HUQ010 == 2 |analysis$HUQ010
== 3,
                          "Good or better",
                          ifelse(analysis$HUQ010 == 4 | analysis$HUQ010 == 5,
                                 "Fair or poor", NA))
analysis <- analysis %>% mutate(health = factor(health, levels=c("Good or better",
                                                                 "Fair or poor"),
                                labels=c("Good or better", "Fair or poor")))
```

```{r}
table(analysis$health)
```

Home Own

```{r}
table(analysis$HOQ065)
```

```{r}
analysis$home <- ifelse(analysis$HOQ065 == 1, "Owned",
                        ifelse(analysis$HOQ065 == 2 | analysis$HOQ065 == 3,
                               "Not owned", NA))

analysis <- analysis %>% mutate(home = factor(home, levels=c("Not owned", "Owned"),
                                labels=c("Not owned", "Owned")))
```

```{r}
table(analysis$home)
```

Work Status

```{r}
table(analysis$OCD150)
```

```{r}
analysis$work <- ifelse(analysis$OCD150 == 2 | analysis$OCD150 == 3 | analysis$OCD150
== 4,
                        "Not working",
                        ifelse(analysis$OCD150 == 1, "Working", NA))

analysis <- analysis %>% mutate(work = factor(work, levels=c("Not working", "Working"),
                                labels=c("Not working", "Working")))
```

```{r}
table(analysis$work)
```

Sleep

```{r}
table(analysis$SLD012)
```

```{r}
analysis$sleep <- ifelse(analysis$SLD012 == 2 | analysis$SLD012 == 14, NA,
                         analysis$SLD012)
```

```{r}
table(analysis$sleep)
```

BMI

```{r}
analysis$bmi <- analysis$BMXBMI
```

Children

```{r}
table(analysis$DMDHHSZA)
```

```{r}
analysis$children <- ifelse(analysis$DMDHHSZA == 0, "No", "Yes")
analysis <- analysis %>% mutate(children = factor(children, levels=c("No", "Yes"),
                                labels=c("No", "Yes")))
```

```{r}
table(analysis$children)
```

New Data Structure

```{r}
str(analysis)
```

Survey Design

```{r}
nhc <- svydesign(id = ~SDMVPSU,
                 weights = ~WTMEC2YR,
                 strata = ~SDMVSTRA, nest=TRUE,
                 survey.lonely.psu = "adjust",
                 data = analysis)
nhc
```

Descriptive Statistics

Missingness

```{r echo=FALSE}
analysis %>%
  tibble::rowid_to_column("obs_id") %>%
  tidyr::gather(key = "key", value = "value", -obs_id) %>%
  group_by(key) %>%
  summarise(`Observations` = n(),
            `Missing Observations` = sum(is.na(value))) %>%
  knitr::kable()
```

Outcome

Avg. # of Drinks (DrinkNum)

```{r echo=FALSE}
svymean(~DrinkNum, nhc, na = T)
```

```{r echo=FALSE}
svyhist(~DrinkNum, nhc, col = "darkseagreen", xlab = "Avg. # of Drinks",
        main = "Histogram of Avg. # of Drinks")
```

```{r echo=FALSE}
svyby(~DrinkNum, ~sex, nhc, svymean, na = T)
```

```{r echo=FALSE}
svyboxplot(~DrinkNum~sex, nhc, all.outliers = TRUE, col = c("cornflowerblue",
"indianred1"),
          ylab = "Avg. # of Drinks", main = "Box Plot of Avg. # of Drinks by Sex")
```

```{r echo=FALSE}
svyplot(~Age20to79+DrinkNum, nhc, style="bubble", col = "lightgoldenrod1",
        ylab = "Avg. # of Drinks", xlab = "Age (20 - 79 years old)",
        main = "Scatterplot of Age & Avg. # of Drinks")
```

Dichotomous Drinking Outcome (Risky)

```{r echo=FALSE}
svymean(~Risky, nhc, na = T)
svytable(~Risky, design = nhc)
```

```{r echo=FALSE}
barplot(svymean(~Risky, nhc, na = T), names.arg = c("Moderate", "Risky"),
        col = "lightgoldenrod1", horiz = TRUE,
        main = "Percentage of Moderate and Risky Drinkers")
```

Age

```{r echo=FALSE}
svymean(~Age20to79, nhc, na = T)
```

```{r echo=FALSE}
svyhist(~Age20to79, nhc, col = "darkseagreen", xlab = "Age (20 - 79 years old)",
        main = "Histogram of Age")
```

```{r echo=FALSE}
svyboxplot(~Age20to79~sex, nhc, all.outliers = TRUE,
          col = c("cornflowerblue", "indianred1"),
          ylab = "Age", main = "Box Plot of Age by Sex")
```

```{r echo=FALSE}
svyplot(~Age20to79+DrinkNum, nhc, style="bubble",
        ylab = "Avg. # of Drinks", xlab = "Age",
        main = "Scatterplot of Age & Avg. # of Drinks")
```

Sex

```{r echo=FALSE}
svymean(~sex, nhc, na = T)
```

```{r echo=FALSE}
barplot(svymean(~sex, nhc, na = T), names.arg = c("Male", "Female"),
        col = "lightgoldenrod1", horiz = TRUE,
        main = "Percentage of Participant's Sex")
```

Education

```{r echo=FALSE}
svymean(~college, nhc, na = T)
```

```{r echo=FALSE}
barplot(svymean(~college, nhc, na = T), names.arg = c("Not College Grad", "College Grad"),
        col = "lightgoldenrod1", horiz = TRUE,
        main = "Percentage of Participant's Education")
```

BMI

```{r echo=FALSE}
svymean(~bmi, nhc, na = T)
```

```{r echo=FALSE}
svyhist(~bmi, nhc, col = "darkseagreen", xlab = "BMI",
        main = "Histogram of BMI")
```

```{r echo=FALSE}
svyboxplot(~bmi~1, nhc, all.outliers = TRUE, col = "cornflowerblue",
           ylab = "BMI", main = "Box Plot of BMI")
```

```{r echo=FALSE}
svyplot(~bmi+DrinkNum, nhc, style="bubble",
        ylab = "Avg. # of Drinks", xlab = "BMI",
        main = "Scatterplot of BMI & Avg. # of Drinks")
```

Peer Substance Use

```{r echo=FALSE}
svymean(~peer, nhc, na = T)
```

```{r echo=FALSE}
barplot(svymean(~peer, nhc, na = T), names.arg = c("No", "Yes"),
        col = "lightgoldenrod1", horiz = TRUE,
        main = "Percentage of Participant's with a Smoker in their Household")
```

PHQ-9 Screener

```{r echo=FALSE}
svyplot(~phqScore+DrinkNum, nhc, style="bubble",
        ylab = "Avg. # of Drinks", xlab = "PHQ Score",
        main = "Scatterplot of PHQ & Avg. # of Drinks")
```

Model Building

Step 0: Univariate Models

Age

```{r}
logit1 <- (svyglm(Risky ~ Age20to79,
```

```
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Sex

```{r}
logit1 <- (svyglm(Risky ~ sex,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Race

```{r}
logit1 <- (svyglm(Risky ~ white + black + mexicanAm,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = T,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

College

```{r}
logit1 <- (svyglm(Risky ~ college,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

PhqScore

```{r}
logit1 <- (svyglm(Risky ~ phqScore,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
```

```
     ci.width = getOption("summ-ci.width", 0.95))
```

Peer

```{r}
logit1 <- (svyglm(Risky ~ peer,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

IncUnder20

```{r}
logit1 <- (svyglm(Risky ~ incUnder20,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Married

```{r}
logit1 <- (svyglm(Risky ~ married,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Sed

```{r}
logit1 <- (svyglm(Risky ~ sed,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Smoker

```{r}
logit1 <- (svyglm(Risky ~ smoke,
```

```
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Diabetes

```{r}
logit1 <- (svyglm(Risky ~ diabetes,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Insure

```{r}
logit1 <- (svyglm(Risky ~ insure,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Health

```{r}
logit1 <- (svyglm(Risky ~ health,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Home

```{r}
logit1 <- (svyglm(Risky ~ home,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
```

```
     ci.width = getOption("summ-ci.width", 0.95))
```

Work

```{r}
logit1 <- (svyglm(Risky ~ work,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Sleep

```{r}
logit1 <- (svyglm(Risky ~ sleep,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

BMI

```{r}
logit1 <- (svyglm(Risky ~ bmi,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Children

```{r}
logit1 <- (svyglm(Risky ~ children,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 5), exp = FALSE, vifs = FALSE,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", TRUE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Step 1: First Multivariable Model

```{r}
logit1 <- (svyglm(Risky ~ Age20to79 + sex + mexicanAm + college + phqScore + peer
```

```
                  + incUnder20 + married + smoke + diabetes + health
                  + home + children,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 3), exp = FALSE, vifs = T,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", FALSE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Step 2: Removing Covariates

```{r}
logit1 <- (svyglm(Risky ~ Age20to79 + sex + mexicanAm + college + phqScore + peer
                  + incUnder20 + married + smoke + diabetes +  health,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 3), exp = FALSE, vifs = F,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", FALSE),
     ci.width = getOption("summ-ci.width", 0.95))
```

```{r}
logit1 <- (svyglm(Risky ~ Age20to79 + sex + mexicanAm + college + phqScore + peer
                  + incUnder20 + married + smoke +  health,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 3), exp = FALSE, vifs = F,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", FALSE),
     ci.width = getOption("summ-ci.width", 0.95))
```

```{r}
logit1 <- (svyglm(Risky ~ Age20to79 + sex + mexicanAm + college + peer
                  + married + smoke +  health + phqScore,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 3), exp = FALSE, vifs = F,
     model.fit = getOption("summ-model.fit", TRUE),
     confint = getOption("summ-confint", FALSE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Step 3: Important Change in Covariates?

Percent change in coefficent estimate:

- Age20to79 = (-0.0298 + 0.0286)/(-0.0286) = 0.042
- sexFemale = (-0.9676 + 0.9381)/(-0.9381) = 0.031
- mexicanAmMexican American = (0.6842 - 0.7120)/(0.7120) = -0.04
- collegeCollege grad = (-0.3315 + 0.3454)/(-0.3454) = -0.04
- peerYes = (0.4333 - 0.4417)/(0.4417) = -0.02
- marriedMarried = (-0.3820 + 0.3859)/(-0.3859) = -0.01

```
- smokeSmoker                    =  (0.6661 - 0.6943)/(0.6943)   = -0.04
- healthFair or poor             =  (0.2284 - 0.2129)/(0.2129)   =  0.073
- phqScore                       =  (0.0213 - 0.0213)/(0.0213)   =  0
```

```
## Step 4: Preliminary Main Effects Model (Latex Code)

$                                        RISKY                                          =
\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\
beta_8X_8+\beta_9X_9 + \epsilon $

## Step 5: Functional Form

```{r}
logit1 <- (svyglm(Risky ~ Age20to79 + I(Age20to79^2) + sex + mexicanAm + college + peer
                  + married + smoke +  health + phqScore,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 3), exp = F, vifs = F,
     model.fit = getOption("summ-model.fit", FALSE),
     confint = getOption("summ-confint", FALSE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Step 6: Interactions

```{r}
logit1 <- (svyglm(Risky ~ Age20to79 + sex + mexicanAm + college + peer
                  + married + smoke + health + phqScore
                  + sex*Age20to79,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 3), exp = T, vifs = T,
     model.fit = getOption("summ-model.fit", F),
     confint = getOption("summ-confint", FALSE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Step 7: Model Diagnostics

```{r}
logit1 <- (svyglm(Risky ~ Age20to79 + sex + mexicanAm + college + peer
                  + married + smoke +  health + phqScore,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 3), exp = T, vifs = F,
     model.fit = getOption("summ-model.fit", F),
     confint = getOption("summ-confint", FALSE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Linearity and Outliers

```{r}
spr <-  rstandard(logit1) # standardized Pearson residual
phat <- logit1$fitted.values # p-hat (prediced probs)
y <- logit1$y # Risky group (of the training data)
```
```

```
resid <- data.frame(spr, phat, y)
resid <- resid %>% mutate(y = factor(y, levels=c(0, 1),
                                     labels=c("Moderate", "Risky")))

ggplot(resid, aes(x = phat, y = spr, color = y)) +
  geom_point() +
  geom_hline(yintercept = 2, linetype = "dashed") +
  geom_hline(yintercept = -2, linetype = "dashed") +
  ggtitle("Standardized Pearson Residual Plot") +
  xlab("Fitted Value (Predicted Probability)") + ylab("Standardized Pearson Residuals")
+
  labs(color = "Drinking Status") +
  scale_color_manual(values = c("darkseagreen", "lightgoldenrod1")) +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))
```

Sensitivity

```{r}
sensitivity <- subset(nhc, rstandard(logit1) > -2 & rstandard(logit1) < 2)
```

```{r}
logit2 <- (svyglm(Risky ~ Age20to79 + sex + mexicanAm + college + peer
                  + married + smoke +  health + phqScore,
                  family = quasibinomial,
                  design = sensitivity, na.action = na.omit))

summ(logit2,
     digits = getOption("jtools-digits", default = 3), exp = F, vifs = F,
     model.fit = getOption("summ-model.fit", F),
     confint = getOption("summ-confint", FALSE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Partial Residual Plot (Linearity for cont. IVs only)

```{r}
car::crPlot(logit1, variable = "Age20to79")
car::crPlot(logit1, variable = "phqScore")
```

Overall Fit

```{r}
performance::performance_hosmer(logit1, n_bins = 10)
```

VIFs

```{r}
logit1 <- (svyglm(Risky ~ Age20to79 + sex + mexicanAm + college + peer
                  + married + smoke +  health + phqScore,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 3), exp = T, vifs = T,
     model.fit = getOption("summ-model.fit", F),
     confint = getOption("summ-confint", FALSE),
     ci.width = getOption("summ-ci.width", 0.95))
```

Final Logistic Model

```{r}
logit1 <- (svyglm(Risky ~ Age20to79 + sex + mexicanAm + college + peer
                  + married + smoke +  health + phqScore,
                  family = quasibinomial,
                  design = nhc, na.action = na.omit))

summ(logit1,
     digits = getOption("jtools-digits", default = 3), exp = T, vifs = F,
     model.fit = getOption("summ-model.fit", F),
     confint = getOption("summ-confint", FALSE),
     ci.width = getOption("summ-ci.width", 0.95))
```

CART Model

Model

In figure:
- Smoke = Smoking status {Never Smoker, Smoker}
- Peer = Substance using peer in household {Yes, No}
- PHQ   = PHQ-9 screener score {0-27}

```{r}
CART1 <- rpms(DrinkNum ~ Age20to79 + sex + mexicanAm + college + peer
              + married + smoke +  health + phqScore,
              data = analysis,
              weights = ~WTMEC2YR,
              strata = ~SDMVSTRA,
              clusters = ~SDMVPSU,
              pval = 0.01)
```

```{r}
# In Latex:
#\usepackage{qtree} \usepackage{lscape} \usepackage{tikz-qtree}

qtree(CART1, title = "Predicting Number of Drinks Consumed when Drinking Alcohol",
      #label = "label", #caption = "caption",
      digits = 2,
      #s_size = FALSE,
      scale = .5, lscape = F, subnode = 1)
```

Test Data

```{r}
load("C:/Users/beminardi/Documents/2020 Pitt Grad/2022 Spring/Thesis/Data/test.RData")
```

```{r}
table(test$sex)
```

Removing NAs from Test Data

- 'test' will be the cleaned testing data with all obsevations.
- 'testF' is the dataset with NO missing used to check AUC and build ROC curve.
- 'testF$Risky' is the vector of true drinking status.

``

```{r}
testF <- test %>% select(SEQN, Risky, DrinkNum, Age20to79, sex, mexicanAm, college,
                         married, phqScore, health, smoke, peer)
testF <- na.omit(testF)
```

ROC/AUC and Accuracy Analysis

Logistic Model

```{r}
yhat <- predict(logit1, newdata = testF,
                type = "response")

roc <- roc(testF$Risky, as.numeric(yhat))
plot(roc, legacy.axe=TRUE, plot=TRUE, print.auc=TRUE,
     col=2, print.auc.y=0.6, xaxs="i", yaxs="i")
title(main = "Weighted Logistic Regression ROC Curve", cex.main = .92)
```

LR Cutoff Graph

```{r}
LRg = data.frame(Cutoff = numeric(), Accuracy = numeric(), stringsAsFactors = FALSE)

for (i in seq(0, 1, by = 0.01)) {
    yhat <- predict(logit1, newdata = testF, type = "response")

    yhat <- data.frame(yhat)
    yhat$Riskpred <- ifelse(yhat$response >= i, "Risky", "Moderate")
    yhat <- yhat %>% mutate(Riskpred = factor(Riskpred, levels=c("Moderate", "Risky"),
                                              labels=c("Moderate", "Risky")))

    A <- table(yhat$Riskpred, testF$Risky)
    CM <- confusionMatrix(A)
    Acc <- CM$overall['Accuracy']

    tmp <- c(i, Acc)

    LRg <- rbind(LRg, tmp)
}

LRg = LRg %>% rename(Cutoff = X0, Accuracy = X0.366725155527197)
LRg
```

```{r}
ggplot(LRg, aes(x = Cutoff, y = Accuracy)) + geom_line() + theme_classic() + labs(title
=  "Weighted  Logistic  Regression  Accuracy  by  Cutoff")  +  theme(plot.title  =
element_text(hjust = 0.5))
```

CART Model

CART Cutoff Graph

```{r}
CARTg = data.frame(Cutoff = numeric(), Accuracy = numeric(), stringsAsFactors = FALSE)

for (i in seq(2, 3, by = 0.01)) {
    yhat <- predict(CART1, newdata = testF)

    yhat <- data.frame(yhat)
```

```
    yhat$Riskpred <- ifelse(yhat$yhat >= i, "Risky", "Moderate")
    yhat <- yhat %>% mutate(Riskpred = factor(Riskpred, levels=c("Moderate", "Risky"),
                                              labels=c("Moderate", "Risky")))

    A <- table(yhat$Riskpred, testF$Risky)
    CM <- confusionMatrix(A)
    Acc <- CM$overall['Accuracy']

    tmp <- c(i, Acc)

    CARTg <- rbind(CARTg, tmp)
}

CARTg = CARTg %>% rename(Cutoff = X2, Accuracy = X0.468495772850534)
```

```{r}
ggplot(CARTg, aes(x = Cutoff, y = Accuracy)) + geom_line() + theme_classic() + labs(title
= "Weighted CART Accuracy by Cutoff") + theme(plot.title = element_text(hjust = 0.5))
```

Appendix C Testing Data R Code

Included below is the R code used to create the testing data used to assess model prediction presented in this thesis. This code can be pasted and utilized directly into an R markdown script.

```
# Load Libraries

```{r load_libraries,message=FALSE}
library(tidyverse)
library(SASxport)
library(gridExtra)
library(lsmeans)
library(expss)
library(mlbench)
library(haven)
library(survey)
```
Reading in Data

```{r}
# 2013-14

alc_2013 <- read.xport("...File Path.../ALQ_H.XPT")
demo_2013 <- read.xport("...File Path.../DEMO_H.XPT")
phq_2013 <- read.xport("...File Path.../DPQ_H.XPT")
health_2013 <- read.xport("...File Path.../HUQ_H.XPT")
smoking_2013 <- read.xport("...File Path.../SMQ_H.XPT")
peers_2013 <- read.xport("...File Path.../SMQFAM_H.XPT")

# 2015-16

alc_2015 <- read.xport("...File Path.../ALQ_I.XPT")
demo_2015 <- read.xport("...File Path.../DEMO_I.XPT")
phq_2015 <- read.xport("...File Path.../DPQ_I.XPT")
health_2015 <- read.xport("...File Path.../HUQ_I.XPT")
smoking_2015 <- read.xport("...File Path.../SMQ_I.XPT")
peers_2015 <- read.xport("...File Path.../SMQFAM_I.XPT")
```

Joining Data

2013 Data

```{r}
join1 <- full_join(alc_2013, demo_2013, by=c("SEQN"))
join2 <- full_join(phq_2013, health_2013, by=c("SEQN"))
join3 <- full_join(smoking_2013, peers_2013, by=c("SEQN"))
join_semi <- full_join(join1, join2, by=c("SEQN"))
join_final_2013 <- full_join(join_semi, join3, by=c("SEQN"))
```

```{r}
data_2013 <- join_final_2013 %>% select(SEQN, WTMEC2YR, SDMVPSU, SDMVSTRA,  ALQ130,
                                        RIDAGEYR, RIDRETH1, RIAGENDR, DMDEDUC2,
                                        SMD460, SMQ020, HUQ010, DMDMARTL,
```
```

```
                                        DPQ010, DPQ020, DPQ030, DPQ040, DPQ050,
                                        DPQ060, DPQ070, DPQ080, DPQ090)
```

2015 Data

```{r}
join1 <- full_join(alc_2015, demo_2015, by=c("SEQN"))
join2 <- full_join(phq_2015, health_2015, by=c("SEQN"))
join3 <- full_join(smoking_2015, peers_2015, by=c("SEQN"))
join_semi <- full_join(join1, join2, by=c("SEQN"))
join_final_2015 <- full_join(join_semi, join3, by=c("SEQN"))
```

Select the proper variables

```{r}
data_2015 <- join_final_2015 %>% select(SEQN, WTMEC2YR, SDMVPSU, SDMVSTRA,  ALQ130,
                                        RIDAGEYR, RIDRETH1, RIAGENDR, DMDEDUC2,
                                        SMD460, SMQ020, HUQ010, DMDMARTL,
                                        DPQ010, DPQ020, DPQ030, DPQ040, DPQ050,
                                        DPQ060, DPQ070, DPQ080, DPQ090)
```

Meshing the two

```{r}
test  <-  full_join(data_2013,  data_2015,  by  =  c("SEQN",  "WTMEC2YR",  "SDMVPSU",
"SDMVSTRA",
                                                     "ALQ130",    "RIDAGEYR",    "RIDRETH1",
"RIAGENDR",
                                                     "SMD460",      "SMQ020",      "HUQ010",
"DMDMARTL",
                                                     "DPQ010", "DPQ020", "DPQ030", "DPQ040",
                                                     "DPQ050", "DPQ060", "DPQ070", "DPQ080",
                                                     "DPQ090", "DMDEDUC2"))
```

Check

```{r}
table(test$ALQ130)
```

```{r}
table(test$RIAGENDR)
```

Data Cleaning

Outcome: Number of Drinks

```{r}
test$Risky <- ifelse(test$ALQ130 == 1 | test$ALQ130 == 2, "Moderate",
                     ifelse(test$ALQ130 == 777 | test$ALQ130 == 999, NA,
                            "Risky"))

test$DrinkNum <- ifelse(test$ALQ130 < 15, test$ALQ130, NA)
```

```{r}
test <- test %>% mutate(Risky = factor(Risky, levels=c("Moderate", "Risky"),
```

```
                                                labels=c("Moderate", "Risky")))
```

```{r}
table(test$ALQ130)
table(test$DrinkNum)
table(test$Risky)
```

Age

```{r}
table(test$RIDAGEYR)
```

```{r}
test$Age20to79 <- ifelse(test$RIDAGEYR >= 20 & test$RIDAGEYR <= 79,
                         test$RIDAGEYR, NA)
```

```{r}
table(test$Age20to79)
```

Race

```{r}
test$mexicanAm <- ifelse(test$RIDRETH1 == 1, "Mexican American", "Other Race")

test <- test %>% mutate(mexicanAm = factor(mexicanAm,
                                           levels=c("Other Race",
                                                    "Mexican American"),
                                           labels=c("Other Race",
                                                    "Mexican American")))

table(test$RIDRETH1)
table(test$mexicanAm)
```

Sex

```{r}
test$sex <- test$RIAGENDR

test <- test %>% mutate(sex = factor(sex, levels=c(1, 2),
                                     labels=c("Male", "Female")))
```

Education

```{r}
test$college <- ifelse(test$DMDEDUC2 == 5, "College grad",
                       ifelse(test$DMDEDUC2 == 7 | test$DMDEDUC2 == 9, NA,
                              "Not college grad"))

test <- test %>% mutate(college = factor(college, levels=c("Not college grad",
                                                           "College grad"),
                                                 labels=c("Not college grad",
                                                          "College grad")))
table(test$DMDEDUC2)
table(test$college)
```

PHQ-9 Screener

```{r}
test$score1 <- ifelse(test$DPQ010 == 7 | test$DPQ010 == 9, NA, test$DPQ010)
test$score2 <- ifelse(test$DPQ020 == 7 | test$DPQ020 == 9, NA, test$DPQ020)
test$score3 <- ifelse(test$DPQ030 == 7 | test$DPQ030 == 9, NA, test$DPQ030)
test$score4 <- ifelse(test$DPQ040 == 7 | test$DPQ040 == 9, NA, test$DPQ040)
test$score5 <- ifelse(test$DPQ050 == 7 | test$DPQ050 == 9, NA, test$DPQ050)
test$score6 <- ifelse(test$DPQ060 == 7 | test$DPQ060 == 9, NA, test$DPQ060)
test$score7 <- ifelse(test$DPQ070 == 7 | test$DPQ070 == 9, NA, test$DPQ070)
test$score8 <- ifelse(test$DPQ080 == 7 | test$DPQ080 == 9, NA, test$DPQ080)
test$score9 <- ifelse(test$DPQ090 == 7 | test$DPQ090 == 9, NA, test$DPQ090)
```

```{r}
test$phqScore <- test$score1 + test$score2 + test$score3 + test$score4 + test$score5 +
test$score6 + test$score7 + test$score8 + test$score9

table(test$phqScore)
```

Peer Substance Use

```{r}
test$peer <- ifelse(test$SMD460 == 1 | test$SMD460 == 2 | test$SMD460 == 3,
                    "Yes", ifelse(test$SMD460 == 0, "No", NA))

test <- test %>% mutate(peer = factor(peer, levels=c("No", "Yes"),
                                      labels=c("No", "Yes")))

table(test$SMD460)
table(test$peer)
```

Married

```{r}
test$married <- ifelse(test$DMDMARTL == 1, "Married",
                       ifelse(test$DMDMARTL == 77 | test$DMDMARTL == 99, NA,
                              "Not Married"))

test <- test %>% mutate(married = factor(married, levels=c("Not Married",
                                                           "Married"),
                                         labels=c("Not Married", "Married")))

table(test$DMDMARTL)
table(test$married)
```

Smoking

```{r}
test$smoke <- ifelse(test$SMQ020 == 7 | test$SMQ020 == 9, NA,
                     ifelse(test$SMQ020 == 1, "Smoker", "Never Smoker"))

test <- test %>% mutate(smoke = factor(smoke, levels=c("Never Smoker", "Smoker"),
                        labels=c("Never Smoker", "Smoker")))

table(test$SMQ020)
table(test$smoke)
```

```
## General Health Condition

```{r}
test$health <- ifelse(test$HUQ010 == 1 | test$HUQ010 == 2 | test$HUQ010 == 3,
                      "Good or better",
                      ifelse(test$HUQ010 == 4 | test$HUQ010 == 5,
                             "Fair or poor", NA))

test <- test %>% mutate(health = factor(health, levels=c("Good or better",
                                                         "Fair or poor"),
                                        labels=c("Good or better", "Fair or poor")))

table(test$HUQ010)
table(test$health)
```

Structure Check

```{r}
str(test)
```

Saving Dataset

```{r}
save(test, file = "test.RData")
```
```

Bibliography

Centers for Disease Control and Prevention. (2021, December 29). *Alcohol Use and Your Health*. Retrieved from CDC: https://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm

Chaiyasong, S., Huckle, T., Mackintosh, A.-M., Meier, P., Parry, C. D., Callinan, S., . . . Casswell, S. (2018, June 13). Drinking patterns vary by gender, age and country-level income: Cross-country analysis of the International Alcohol Control Study. *Drug and Alcohol Review, 37*(2), 53-62. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1111/dar.12820

Collins, S. E. (2016). Associations Between Socioeconomic Factors and Alcohol Outcomes. *Alcohol Research, 38*(1), 83-94. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4872618/

Fawcett, T. (2006, June). An Introduction to ROC Analysis. *Pattern Recognition Letters, 27*(8). Retrieved from https://doi.org/10.1016/j.patrec.2005.10.010

French, M. T., Popovici, I., & Maclean, J. (2009, September). Do Alcohol Consumers Exercise More? Findings from a National Survey. *American Journal of Health Promotion, 24*(1), 2-10. Retrieved from https://doi.org/10.4278/ajhp.0801104

Kelly, A. B., Chan, G. C., Toumbourou, J. W., O'Flaherty, M., Homel, R., Patton, G. C., & Williams, J. (2012, April). Very young adolescents and alcohol: Evidence of a unique susceptibility to peer alcohol use. *Addictive Behaviors, 37*(4), 414-419. Retrieved from https://doi.org/10.1016/j.addbeh.2011.11.038

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001, December 20). The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine, 16*(9), 606-613. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1046/j.1525-1497.2001.016009606.x

LaValley, M. P. (2008, May 6). Logistic Regression. *Circulation, 117*(18), 2395-2399. Retrieved from https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.106.682658#d3e142

Loh, W.-Y. (2011, January 6). Classification and Regression Trees. *WIREs, 1*(1), 14-23. Retrieved from https://doi.org/10.1002/widm.8

Morris, E. P., Stewart, S. H., & Ham, L. S. (2005, September). The relationship between social anxiety disorder and alcohol use disorders: A critical review. *Clinical Psychology Review, 25*(6), 734-760. Retrieved from https://doi.org/10.1016/j.cpr.2005.05.004

National Center for Health Statistics. (2017, September 15). *National Health and Nutrition Examination Survey Overview*. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/nchs/data/nhanes/nhanes_13_14/NHANES_Overview_Brochure.pdf

National Center for Health Statistics. (2022, March). *NHANES Sample Design*. Retrieved from Centers for Disease Control and Prevention: https://wwwn.cdc.gov/nchs/nhanes/tutorials/Module2.aspx

O'Connor, E. A., Perdue , L. A., Senger, C. A., Rushkin, M., Patnode, C. D., Bean, S. I., & Jonas, D. E. (2018, November 13). Screening and Behavioral Counseling Interventions to Reduce Unhealthy Alcohol Use in Adolescents and Adults: An Updated Systematic Review for the U.S. Preventive Services Task Force. *JAMA, 320*(18), 1910-1928. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK534919/table/ch1.tab1/

Paavola, M., Vartiainen, E., & Haukkala, A. (2004, September). Smoking, alcohol use, and physical activity: A 13-year longitudinal study ranging from adolescence into adulthood. *Journal of Adolescent Health, 35*(3), 238-244. Retrieved from https://doi.org/10.1016/j.jadohealth.2003.12.004

Piazza-Gardner, A. K., & Berry, A. E. (2012, January 1). Examining Physical Activity Levels and Alcohol Consumption: Are People Who Drink More Active? *American Journal of Health Promotion, 26*(3), 95-104. Retrieved from https://journals.sagepub.com/doi/abs/10.4278/ajhp.100929-lit-328

Toth, D. (2017). rpms: An R Package for Modeling Survey Data with Regression Trees. *U.S. Bureau of Labor Statistics*. Retrieved from https://mran.microsoft.com/snapshot/2018-08-19/web/packages/rpms/vignettes/rpms_2018_01_22.pdf