Georgia Tech

Online Master of
Science in Analytics

MGT 6203 Group Project

Final Report

Effects of Demographic and Socioeconomic Factors on Alcohol Abuse in the United States

April 16th, 2023

Team 6:

Brock Mitchell Wilkinson

Pankaj Shrestha

Caroline Elizabeth Stephenson Seay

Nicholas John Rupert

Table of Contents

Background

Choice of Topic and Literature Review

Alcohol abuse is a significant public health concern in the United States, with negative impacts on individuals, families, and communities. Binge drinking is the most common and costly pattern of alcohol abuse in the United States, [1] and has been linked to numerous health problems, including liver disease, cardiovascular disease, and cancer.

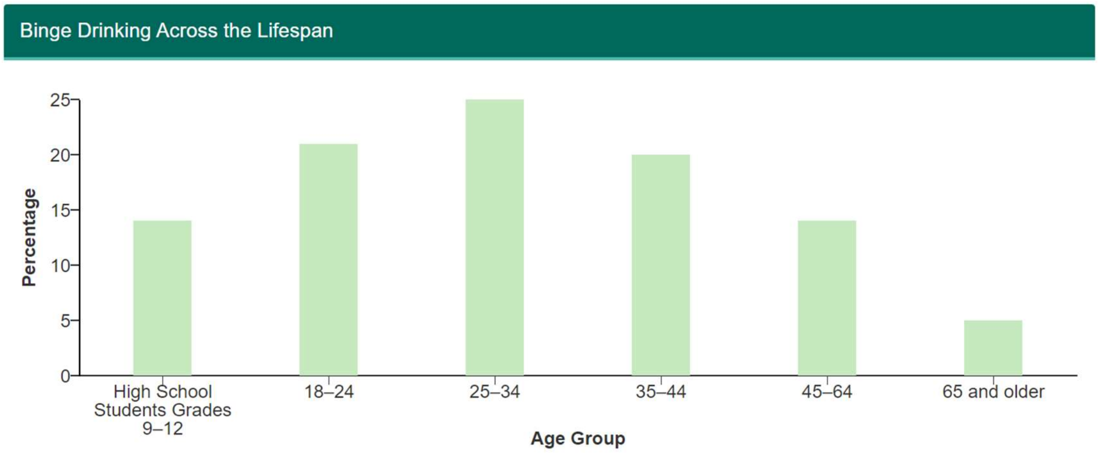


Figure 1: Binge Drinking in the United States by Age Group

Alcohol abuse can also exacerbate mental health issues such as depression and anxiety.[2] The long-term health consequences of alcohol abuse can be severe, leading to reduced quality of life and premature death: it is estimated that around 14 million adults in the United States suffer from alcohol use disorder (AUD), [3] and excessive drinking is responsible for around 95,000 deaths within the US each year.[4] Alcohol abuse can also lead to a range of social problems, including relationship breakdowns, negative impact on work and academic performance, domestic violence, and criminal behavior.

Business Justification and Objective

The economic cost of alcohol abuse in the United States is substantial, with estimates suggesting that it costs the economy around $249 billion per year in lost productivity, healthcare expenses, and criminal justice expenses.[5] For this project our team is consulting for the Centers for Disease Control and Prevention, or CDC. Specifically, we are helping them explore treatment and prevention efforts for alcohol abuse in the United States at the local level. If we can help the CDC reduce the amount of alcohol abuse in the US, we can reduce the financial strain that the disease causes, and those dollars can be better allocated elsewhere within the US economy.

Problem Statement

Alcohol abuse is a complex issue that requires multidisciplinary treatment, including medical, psychological, and social interventions. Targeted intervention efforts have proved successful in community samples of adolescents,[6] however, understanding the extent and distribution of alcohol abuse at the local level can be challenging.[7] US Census data may be able to provide valuable insights into the demographic and social factors that contribute to alcohol abuse in specific US counties, which can inform targeted prevention and treatment efforts.

Client Requirements

In helping the CDC reduce the financial burden that alcohol abuse causes within the United States, they have provided us three guidelines for our analysis:

1. Data analyzed must be owned by the government, and thus available to the CDC for free.
2. Insights drawn should be scalable across the US and should not be focused on one specific community or county.
3. Recommendations should identify specific policy areas (e.g., level of Educational Attainment) where the CDC should focus their targeted prevention efforts.

Data Sources

The US Census collects accurate and comprehensive data about the United States to allocate federal funds for various programs and services, and to help determine political representation in Congress. The 2020 US Census collected over 300 million data points on a wide range of demographic and social factors like poverty, income, and level of education. A subset of these data attributes will be used in our analysis as predictors for our dependent variable of Binge Drinking Among Adults by county.

The CDC PLACES Survey collects data on a range of environmental and social factors that can affect health outcomes at the community level. This includes information on access to healthy food, availability of recreational facilities, exposure to pollution and hazardous materials, and prevalence of crime and violence. The data is used to inform public health policies and interventions aimed at improving health outcomes and reducing health disparities across the United States. Binge Drinking Among Adults by County from this survey will be used as our dependent variable/response variable.

It should be noted that the Census rankings and the rate of binge drinking from the CDC will be from the latest year they were both available, 2020. While same-year Census rankings are not a perfect predictor for the rate of binge drinking in a particular County, we are treating the Census rankings as cumulative figures that have years of political policy, economic, and social impact baked into them. While there is no standard timeline to develop alcohol dependency,[8] alcohol abuse is a condition that often builds over time. As such, 2020 Census rankings serve as an appropriate proxy for factors from previous years that could impact alcohol abuse in the year 2020.

Data Set Creation and Variable Transformations

Our team first used an R Script to federate 15 distinct data sets from the US Census Bureau with County-level data on broad categories such as Age and Sex, Employment Status, and Financial Characteristics. These data sets contained hundreds of variables, and using information from our literature review, we narrowed these down to 48 initial predictor variables. Several new variables were created as part of this initial variable selection, such as the creation of a total proportion of veterans within the population of a particular county rather than breaking that figure out by eras of war veterans.

Using dplyr and sqldf, the selected attributes from the original 15 data sets were merged into a new data set using a unique identifier present in each of the data sets known as the County Federal Information Processing System (FIPS) Code. Our team then turned our attention to the CDC PLACES survey, where we exported and isolated the “Binge Drinking Among Adults” data attribute that will serve as our dependent variable. Using the same County-level FIPS Code, our dependent variable was added to the merged data set for each of the Counties in the United States.

Our final merged csv file has 3,144 rows: one row for the headers for our data attributes, and the remaining 3,143 rows represent each of the 3,143 Counties there were within the United States in 2020. For each of those rows, there are 53 columns: four of these columns are identification variables such as the FIPS Code or State Abbreviation, one of these columns is our dependent variable of Binge Drinking Among Adults, and the remaining 48 are the initial predictor variables.

Approach/Methodology

Research Questions and Initial Hypothesis

Using the data above, the primary research question we aim to answer is: how can US Census data be used to predict the prevalence of Binge Drinking in US Counties? Two secondary research questions will support the primary research question: What demographic and socioeconomic factors are most strongly associated with alcohol abuse at the local level, and how do these factors differ across the US? How does the prevalence of alcohol abuse vary across age groups, racial/ethnic groups, and genders, and how can these patterns be used to target prevention and intervention efforts at the County level?

The null hypothesis is that no data from the US Census can be used to predict the prevalence of alcohol abuse in the United States at the County level. Our initial hypothesis is that a single, or combination of demographic and socioeconomic factors can be used to predict the prevalence of alcohol abuse in the Unites States at the County level.

Analytical Methods and Models

The table below details the analytical methods and models used to answer our research questions:

#	Method	Rationale
1	Exploratory Analysis & Data Visualization	▪ Understand variable skews, distributions and correlations ▪ Outlier and missing value identification
2	Data Preparation	▪ Outlier removal or imputation where appropriate ▪ Creation of Training/Validation/Test Sets
3	KNN Classification/K-Means Clustering	▪ Group similar Counties so the CDC can apply targeted prevention programs at scale
4	Logistic Regression	▪ Feature Selection to narrow data attributes ▪ Determine the probability of Binge Drinking for a County group identified in KNN Classification ▪ Use difference estimators to understand the specific factors driving Binge Drinking in a group of Counties
5	Tuning Methods	▪ Assess Ridge/LASSO Performance ▪ Assess Log Model Performance ▪ Assess Boosting Algorithm Performance
6	Model Comparison and Selection	▪ Comparison of model performance and selection of best model for each County group

Table 1: Analytical Methods and Rationale

Each method in Table 1 is grouped by color, signifying a different step within our analysis. Exploratory Analysis and Data Preparation will be conducted to garner a strong understanding of the different variables we have at play, helping to answer the two supporting research questions. Then, KNN Classification or K-Means Clustering will be used to group US counties and ensure the insights we provide will be scalable across the United States, per the CDC's requirements. Logistic Regression and various Tuning Methods will then be used to determine the driving factors of Binge Drinking within each

County grouping, and finally, our team will select the best performing models for each Country grouping and deliver our recommendations to the Centers for Disease Control and Prevention.

Exploratory Analysis and Data Visualization

In Exploratory Analysis, our team first checked for any missing values, and found only two: the Median House Cost for Jefferson County, TX, and the Median Household Income for Kennedy County, TX. We determined that imputation was appropriate for these Counties and imputed the values using Median Imputation. Next, we turned our attention to variable correlations to garner a preliminary understanding of those independent variables correlated to our dependent variable, and those independent variables that are correlated with one another.

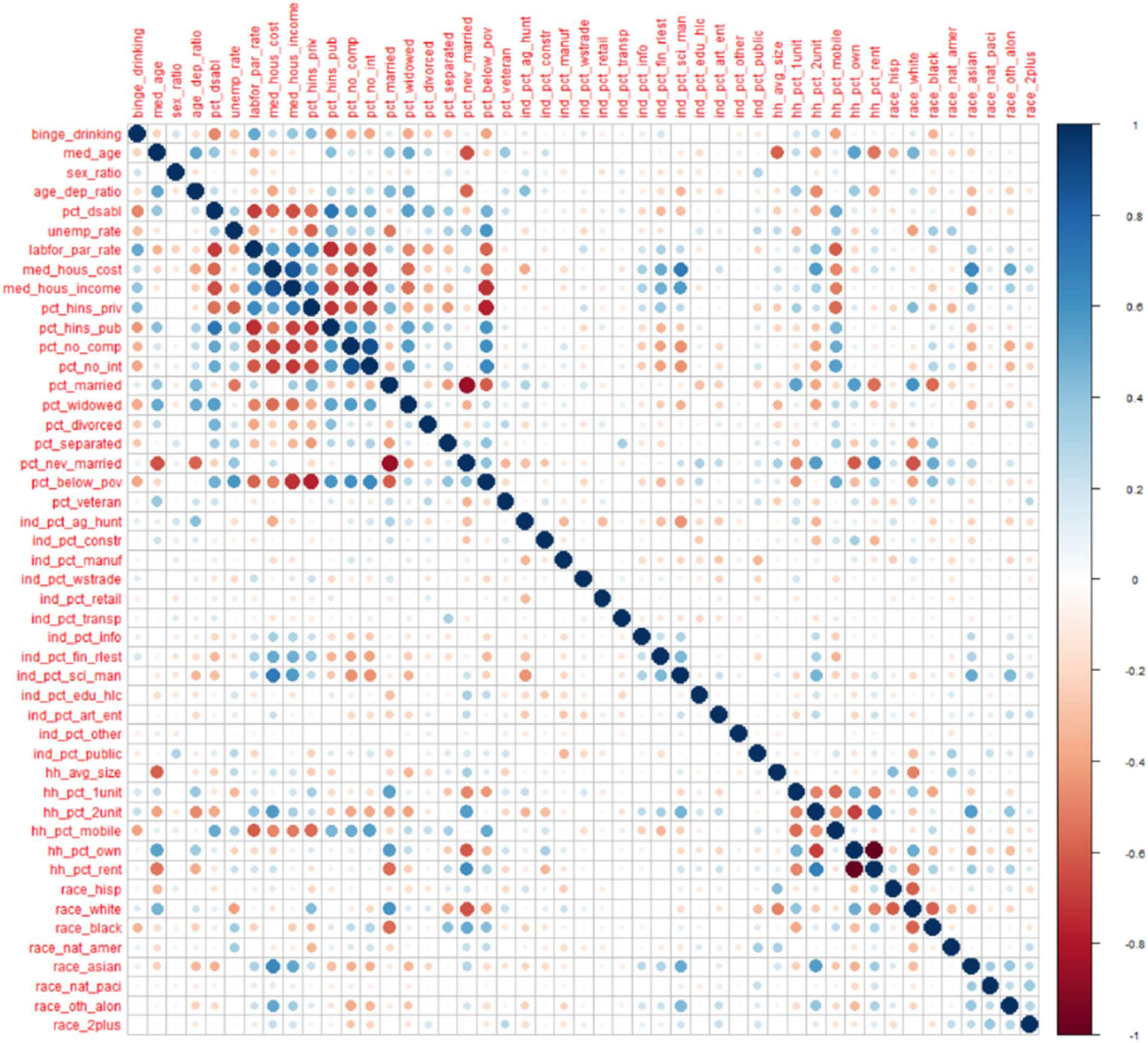
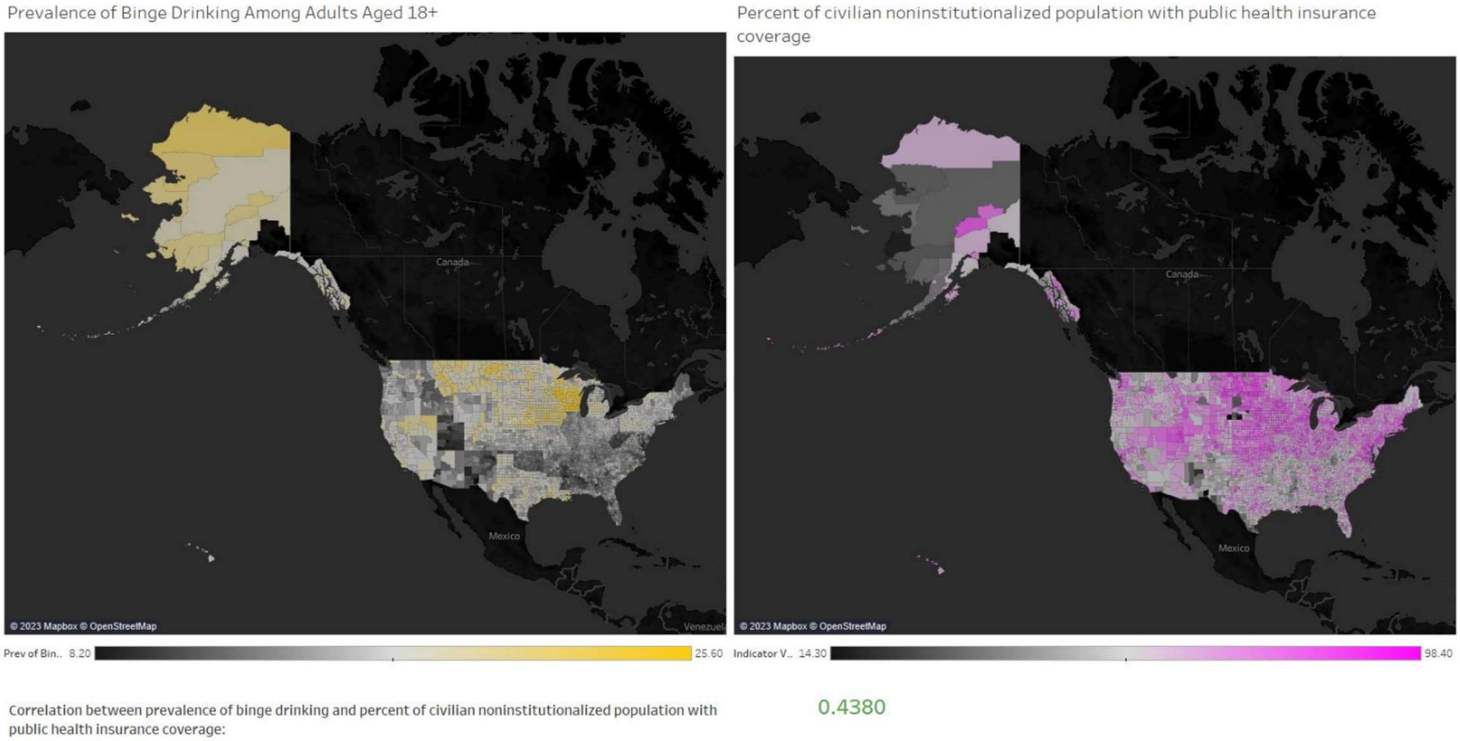


Figure 2: Correlation Plot for Initial Variables

This correlation plot appears to support our initial hypothesis – there are several variables strongly correlated with the rate of Binge Drinking in a county, such as the Labor Force Participation Rate, Percentage of Population with a Disability, and the Median Household Income in each County. This plot will also be useful as we do feature selection for our Logistic Regression Models. Perfectly correlated factors such as the Percentage of the Population Renting and the Percentage of the Population Owning can be consolidated into one factor, and other strongly correlated factors can be removed entirely.

Our team also leveraged Tableau to create heat maps to see how our dependent variable and certain independent variables are concentrated within the United States. Some variables such as the Percent of civilian noninstitutionalized population with public health insurance coverage displayed strong correlation with the prevalence of Binge Drinking Among Adults, which can be seen in Figure 3:


Prevalence of Binge Drinking Among Adults Aged 18+

Percent of civilian noninstitutionalized population with public health insurance coverage

Correlation between prevalence of binge drinking and percent of civilian noninstitutionalized population with public health insurance coverage:

0.4380

Figure 3: Heat Maps for Dependent Variable and Selected Independent Variable

Finally, our team used box plots to check for any outliers present in the data. Our dependent variable of Binge Drinking by County did not appear to have any outliers, however, several predictor variables had significant outliers, such as the proportion of “Native Hawaiian and Other Pacific Islander” present in the population. Further exploration showed the Counties in these scenarios are in Hawaii, and thus these outliers are entirely valid and should not be removed. Ultimately, our team decided against removing any outliers as they could be indicative of trends that should be further explored.

Grouping Similar Counties

K-means Clustering

To ensure the insights we provide the CDC are scalable, our team first leveraged K-means Clustering to group similar Counties that will each have their own Logistic Regression Models. Our team leveraged an Elbow Diagram to determine the optimal number of centers for our K-means Clustering:

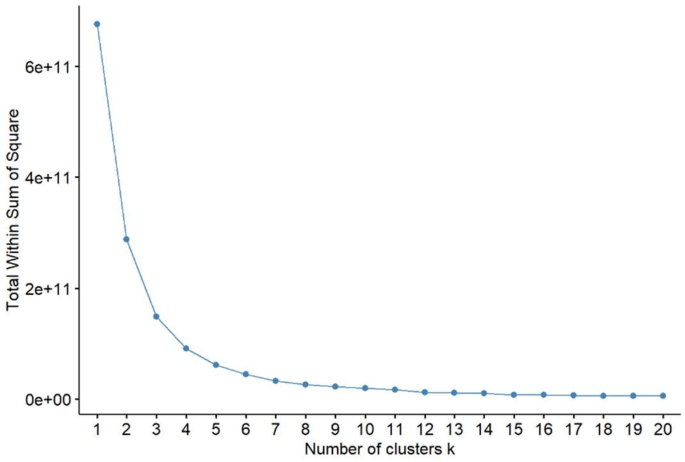


Figure 4: Elbow Diagram for K-means Clustering Analysis

Our team opted to move forward with 4 centers for our K-Means Clustering to balance the reduction of Within Cluster Sum of Squared Errors (WSS) and Explainability so that our analysis can be easily interpreted by the Centers for Disease Control and Prevention.

K-Nearest-Neighbors (KNN) Classification

Our team also explored K-Nearest-Neighbors (KNN) Classification as an alternative to the K-Means Clustering. Unfortunately, the optimal k was 31, meaning there would be over 100 groupings of United States Counties. While this model performed quite well and correctly classified approximately 80% of US Counties, it would be impractical for the CDC to launch over 100 different targeted prevention programs for Alcohol Abuse. Thus, our team will move forward with the 4 Clusters from K-Means Clustering and begin building and interpreting Logistic Regression Models for each of the Clusters.

Logistic Regression

Data Preparation and Initial Logit Model

Since Logistic Regression requires a binary dependent variable, our team first converted the rate of Binge Drinking for each County to a binary variable, highbinge. Any County that had an above average rate of Binge Drinking Among Adults received a value of 1, while those Counties with a below average rate of Binge Drinking received a value of 0. The data for each Cluster were then split into training, validation and test sets, and our team began creating Logistic Regression models. The initial Logit Model for each Cluster was made using the glm function in R with 13 features deemed relevant through exploration of correlation with the dependent variable during Exploratory Data Analysis.

Tuning Methods

Once the initial Logit Model was created for each Cluster, we leveraged four tuning methods to perform hyperparameter optimization – a stepwise selected model, ridge regression, log transformed models, and models made with boosting algorithms. To conduct stepwise selection in R, our team used the step AIC function, which conducts logistic regression for a given set of variables and iteratively adds or removes variables based on their significance until the function selects the model with the best AIC. To run Ridge Regression, our team used the cv.glmnet function to identify the ideal lambda, and then used the glmnet function with a value of alpha = 0 to create the ridge model.

To explore whether Log Transformations were appropriate for any of the four Clusters, our team tested the linearity of the relationship between the Clusters' independent variables and dependent variables. Where nonlinear relationships were present and thus log transformations were appropriate, they would be used to log transform a variable. However, linearity testing showed that independent variables generally displayed linear relationships with their dependent variables, and thus Log Transformations were not appropriate for any of our Clusters. Figure 5 demonstrates the linearity testing for Cluster 2, where each of the independent variables generally had linear relationships with the dependent variable:

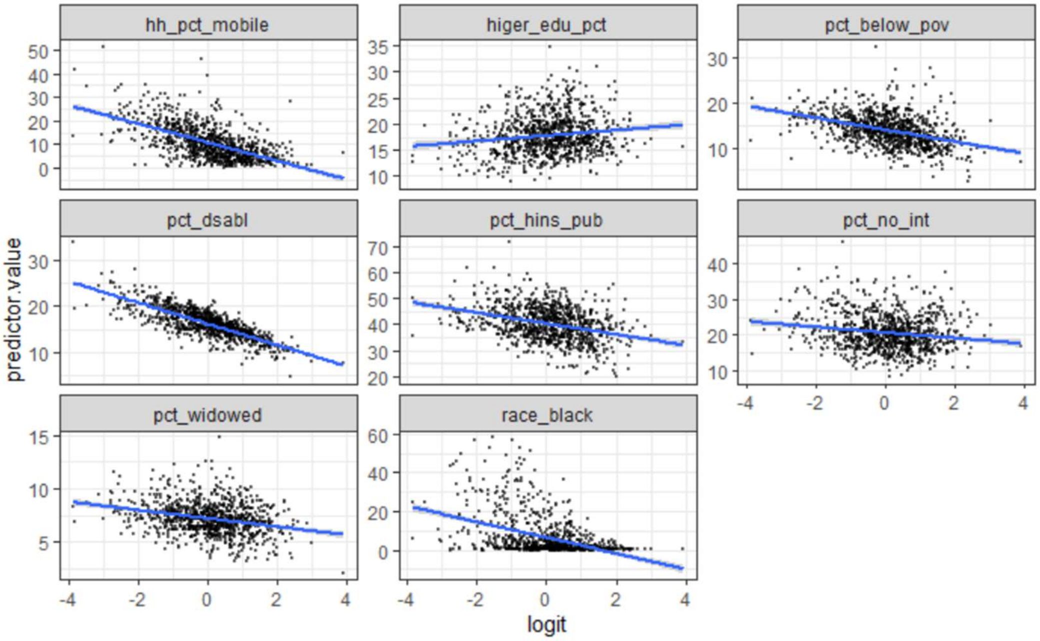


Figure 5: Testing of Linearity Assumptions for Independent Variables in Cluster 2

Finally, our team used boosting, the process of iteratively adding basis functions in a greedy fashion so that each additional basis function further reduces the loss function and improves model performance. To do so, we used the gbm function in R to conduct Generalized Boosted Regression modeling with a Bernoulli distribution. To tune the hyperparameters for this model, a grid search method was used with two options for each hyperparameter optimization which included number of trees, shrinkage, interaction depth, and minimum number of observations in trees.

Model Selection and Performance

Model Selection

Once all tuned models were created using the training data for each Cluster, they were compared using the validation data for each Cluster. Because Logistic Regression does not have an R^2 value that can be used to assess model performance, our team instead made predictions with each of the models using a threshold of 0.5 and compared the Accuracy of each model. Table 2 displays the Accuracy for each model against the validation data for each Cluster, with the best performing model(s) for each Cluster highlighted:

Model Name	Cluster 1 Accuracy	Cluster 2 Accuracy	Cluster 3 Accuracy	Cluster 4 Accuracy
Standard Logit	0.6897	0.7767	0.8527	0.7315
Stepwise Logit	0.7241	0.7718	0.8527	0.7593
Ridge Logit	0.7241	0.7573	0.8682	0.7593
Log Transformed Logit	*n/a*	*n/a*	*n/a*	*n/a*
GBM (Boosted) Logit	0.6552	0.7524	0.8605	0.6944

Table 2: Performance of Logit Models Against Cluster Validation Sets

Using Cluster 2 as an example, the Standard Logit model performed the best against the validation set, followed by the Stepwise Logit, Ridge Logit, and GBM (Boosted) Logit. As mentioned in the Tuning Methods section of this report, the relationships between the independent variables and the dependent variable for Cluster 2 were generally linear, so a Log Transformed Logit was not applicable.

Model Performance

For each of the four Clusters, the best performing Logit model was selected to be assessed against the test set. Since the Standard Logit model for Cluster 2 and the Ridge Logit model for Cluster 3 performed the best against the validation set, they were selected, and will be assessed against the test set. Since model performance was the same for the Stepwise Logit and Ridge Logit models for Clusters 1 and 4, our team selected the Stepwise models as they would be easier to explain to our client, the Centers for Disease Control and Prevention.

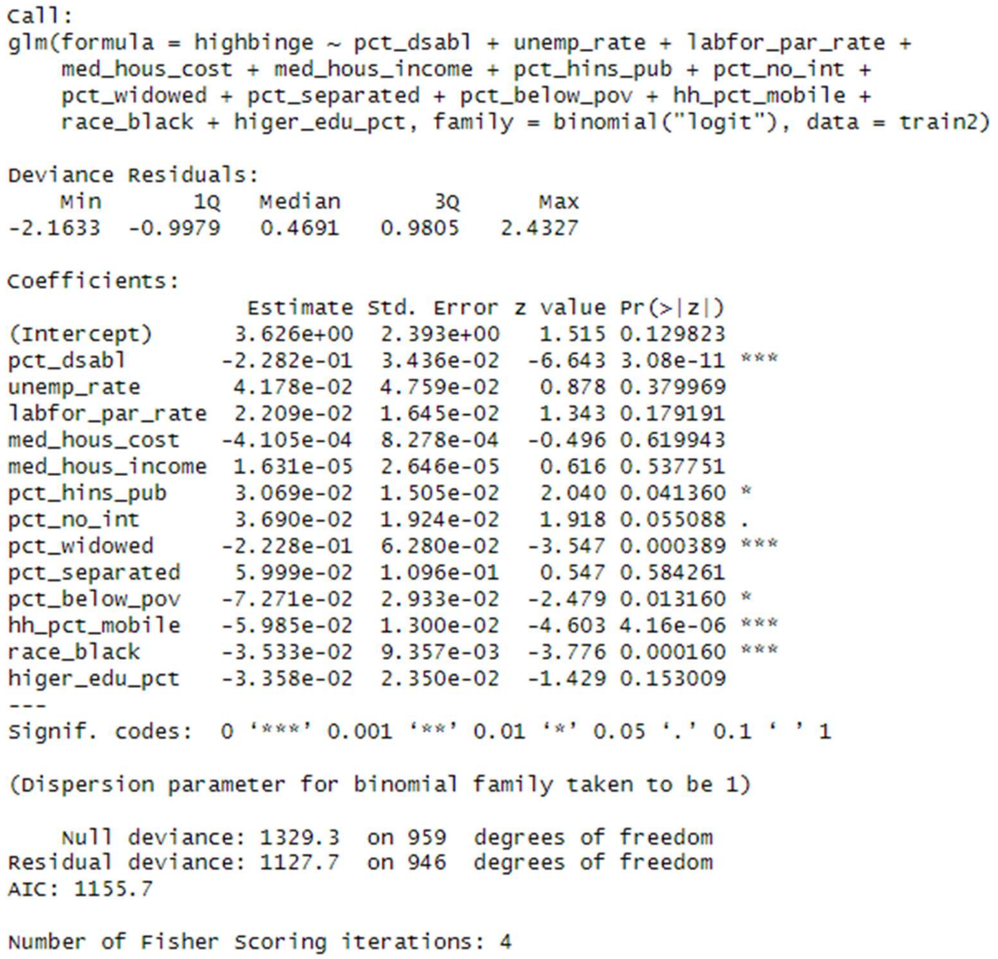
Table 3 demonstrates the final performance of each selected model against their respective test set:

Cluster	Selected Model	Accuracy (Test Set)
Cluster 1	Stepwise Logit	0.7931
Cluster 2	Standard Logit	0.6942
Cluster 3	Ridge Logit	0.8385
Cluster 4	Stepwise Logit	0.7222

Table 3: Performance of Selected Logit Models Against Cluster Test Sets

Model Interpretation

Our team analyzed each selected model’s summary output to understand the difference estimators that increase or decrease the log odds of highbinge for a particular Cluster. Continuing to use Cluster 2 as an example, Figure 6 displays the summary output for the Standard Logit model for Cluster 2:



```
Call:
glm(formula = highbinge ~ pct_dsabl + unemp_rate + labfor_par_rate +
    med_hous_cost + med_hous_income + pct_hins_pub + pct_no_int +
    pct_widowed + pct_separated + pct_below_pov + hh_pct_mobile +
    race_black + higer_edu_pct, family = binomial("logit"), data = train2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1633  -0.9979   0.4691   0.9805   2.4327

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.626e+00  2.393e+00   1.515 0.129823
pct_dsabl       -2.282e-01  3.436e-02  -6.643 3.08e-11 ***
unemp_rate       4.178e-02  4.759e-02   0.878 0.379969
labfor_par_rate  2.209e-02  1.645e-02   1.343 0.179191
med_hous_cost   -4.105e-04  8.278e-04  -0.496 0.619943
med_hous_income  1.631e-05  2.646e-05   0.616 0.537751
pct_hins_pub     3.069e-02  1.505e-02   2.040 0.041360 *
pct_no_int       3.690e-02  1.924e-02   1.918 0.055088 .
pct_widowed     -2.228e-01  6.280e-02  -3.547 0.000389 ***
pct_separated    5.999e-02  1.096e-01   0.547 0.584261
pct_below_pov   -7.271e-02  2.933e-02  -2.479 0.013160 *
hh_pct_mobile   -5.985e-02  1.300e-02  -4.603 4.16e-06 ***
race_black      -3.533e-02  9.357e-03  -3.776 0.000160 ***
higer_edu_pct   -3.358e-02  2.350e-02  -1.429 0.153009
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1329.3  on 959  degrees of freedom
Residual deviance: 1127.7  on 946  degrees of freedom
AIC: 1155.7

Number of Fisher Scoring iterations: 4
```

Figure 6: Summary Output for Cluster 2 Standard Logit Model

We can see that for Cluster 2, increasing the percentage of population with public health insurance increases the log odds of highbinge. As a result, the CDC should seek either to improve the public health insurance offered in Cluster 2 Counties, or help those Counties increase access to private health insurance. Next, we can see that increasing the percentage of the population with no internet access also increases the log odds of highbinge. This means that the CDC should seek to increase access to the Internet for Individuals with No Internet access in Cluster 2 Counties.

Finally, we can see that increasing the percentage of population below the poverty line, and the percentage of households that are mobile homes decreases the log odds of highbinge. While somewhat counterintuitive at first, this indicates that for Counties in Cluster 2, binge drinking is an activity that those with discretionary income partake in. As such, the CDC should focus their efforts on those individuals with discretionary income when crafting a targeted prevention program for Counties in Cluster 2.

Recommendations for the CDC

For each of the selected models for the four County Clusters, our team interpreted the summary output as detailed in the previous section. Since our client requested that we identify specific levers for their targeted prevention programs, Table 4 displays recommended levers for targeted prevention programs for the Centers for Disease Control and Prevention based on each of the summary outputs:

Cluster	Recommended Levers for Targeted Prevention Programs
Cluster 1	▪ Focus prevention efforts on individuals that are employed ▪ Focus prevention efforts on individuals living in mobile homes
Cluster 2	▪ Improve public health insurance, or assist Cluster 2 Counties in increasing access to private health insurance ▪ Increase access to the Internet ▪ Focus prevention efforts on individuals with discretionary income
Cluster 3	▪ Focus prevention efforts on individuals that are unemployed, or below the poverty line ▪ Target prevention efforts towards individuals that are separated
Cluster 4	▪ Focus prevention efforts on individuals that are employed ▪ Increase access to higher education

Table 4: Recommended Levers for Targeted Prevention Programs by County Cluster

Conclusion

By creating several Logit models for each of the four County Clusters, our team can successfully reject the null hypothesis that no US Census factors can be used to predict the prevalence of alcohol abuse at the County level. For each of our four County Clusters, a mix of US Census factors can not only be used to predict the prevalence of alcohol abuse in US counties but can be used to inform targeted prevention strategies for the Centers for Disease Control and Prevention. By launching targeted prevention strategies with the recommended levers our team has detailed, the CDC can reduce the large financial burden alcohol abuse causes within the United States and help to better allocate those dollars elsewhere.

Next Steps and Further Exploration

Given more time and resources, the Centers for Disease Control and Prevention should explore several areas that were outside the scope of this report. First, because census factors are not a perfect predictor for same-year rate of binge drinking, the CDC should consider using time series census data to explore trends over time. Second, while our team whittled down thousands of census variables to our 48 initial predictor variables, the CDC should consider exploring any other Census variables that were outside the scope of this project. Finally, the CDC should consider collecting data at a level even more local than a county – at the district, or township level.

Works Cited

1. “Binge Drinking.” *Centers for Disease Control and Prevention*, 14 Nov. 2022, https://www.cdc.gov/alcohol/fact-sheets/binge-drinking.htm.

2. “Alcohol Use and Your Health.” *Centers for Disease Control and Prevention*, 14 Apr. 2022, https://www.cdc.gov/alcohol/fact-sheets/alcohol-use.htm.

3. “Alcohol Use Disorder Per Year among Persons Aged 12 or Older, by Age Group and Demographic Characteristics.” *Substance Abuse and Mental Health Services Administration*, 11 Sept. 2020, https://www.samhsa.gov/data/sites/default/files/reports/rpt29394/NSDUHDetailedTabs2019/NSDUHDetTabsSect5pe2019.htm#tab5-4a.

4. “Alcohol and Public Health: Alcohol-Related Disease Impact (ARDI).” *Centers for Disease Control and Prevention*, 4 Jan. 2020, https://nccd.cdc.gov/DPH_ARDI/Default/Report.aspx?T=AAM&P=612EF325-9B55-442B-AE0C-789B06E3A8D5&R=C877B524-834A-47D5-964D-158FE519C894&M=DB4DAAC0-C9B3-4F92-91A5-A5781DA85B68&F=&D=.

5. “Excessive Drinking Is Draining the U.S. Economy.” *Centers for Disease Control and Prevention*, 14 Apr. 2022, https://www.cdc.gov/alcohol/features/excessive-drinking.html.

6. Griffin, Kenneth W., and Gilbert J. Botvin. “Evidence-Based Interventions for Preventing Substance Use Disorders in Adolescents.” *Child and Adolescent Psychiatric Clinics of North America*, vol. 19, no. 3, 1 July 2011, pp. 505–526., https://doi.org/10.1016/j.chc.2010.03.005.

7. “Substance Use and Misuse in Rural Areas.” *Rural Health Information Hub*, 9 Dec. 2020, https://www.ruralhealthinfo.org/topics/substance-use.

8. Becker, Howard C. “Alcohol Dependence, Withdrawal, and Relapse.” *Alcohol Research & Health*, vol. 31, no. 4, 2008, pp. 348–361., https://doi.org/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3860472/.

Code Repository

- Merged Data Set
  - https://github.gatech.edu/MGT-6203-Spring-2023-Canvas/Team-6/tree/main/Data/Merged%20Data

- Independent Variable Glossary
  - https://github.gatech.edu/MGT-6203-Spring-2023-Canvas/Team-6/blob/main/Data/Merged%20Data/Metadata.xlsx

- Final Code
  - https://github.gatech.edu/MGT-6203-Spring-2023-Canvas/Team-6/tree/main/Final%20Code