Mental Health in US Counties

Stat 471 Final Project

Source: Getty Images/iStockphoto

Emily Kanter and Katie Rush

December 20, 2021

Github Repository: https://github.com/kar4200/kanter-rush-final.git

Executive Summary

In 2019, it was reported that approximately 19.86% of American adults experienced some form of mental illness.[1] Past research shows that mental health is strongly related to demographic factors and social determinants of health.[2] While support exists to improve mental health on the individual level, we wanted to expand on research to address mental health on a larger scale. Therefore, for our final project, we attempted to predict the average number of mentally unhealthy days per month on a county level, using a set of selected features.

We gathered data from the US Decennial Census[3] and the 2019 County Health Rankings National Dataset.[4] After cleaning and merging the datafiles, our tidy dataset comprised 2,371 observations (i.e. counties) and 65 columns. Our explanatory variables fell into six major categories: (1) health outcomes, (2) health behaviors, (3) clinical care, (4) social and economic environment, (5) physical environment, and (6) demographic information. Our response variable of interest was "mentally unhealthy days," the average number of mentally unhealthy days reported by a specific county in the United States.

After we split the data into a training and a test set, we explored the correlations between the features and the relationship between the features and the response. With the data exploration in mind, we fit six predictive models: (1) an ordinary least squares regression, (2) a ridge regression, (3) a lasso regression, (4) a decision tree, (5) a random forest, and (6) a boosted tree.

We then compared the six models on account of method performance using their mean-squared test error. We found that boosting and random forest yielded the highest predictive performance, while the decision tree had the worst predictive performance. It appeared that the averaging and aggregating of the boosted tree and random forest drastically improved the instability of the decision tree, leading to lower test errors. The lasso model performed better than the ridge model, suggesting that a smaller subset of the total features were actually predictive of the number of mentally unhealthy days.

Interestingly, we found that all models pointed to the percentage of smokers in the county as the strongest feature in predicting the number of mentally unhealthy days. Other predictive features tended to fall into the *health behavior* and the *social and economic environment* subcategories. Some examples include: "percent excessive drinking", "percent insufficient sleep", "percent free or reduced lunch", and "household income". For most features, the relationship with the response followed our intuition. The higher the percentage of detrimental health behaviors (i.e. smoking, insufficient sleep, diabetes), the higher the predicted number of mentally unhealthy days. We hope to use the information gained from this analysis to better inform policies targeted at improving mental health across all counties in the United States.

---

[1] Mental Health America (2021). "The State of Mental Health in America."
https://www.mhanational.org/issues/state-mental-health-america.
[2] Office of Disease Prevention and Health Promotion (2020). "Mental Health.
"https://www.healthypeople.gov/2020/leading-health-indicators/2020-lhi-topics/Mental-Health/determinants
[3] Guide to publicly available Demographic Data. (n.d.).
https://demographics.coopercenter.org/guide-to-publicly-available-demographic-data
[4] County Health Rankings and Roadmaps. (2019). "2019 Measures."
https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation/national-data-documentation-2010-2019

Introduction

Background

The World Health Organization defines health as "a state of complete, mental, and social well-being and not merely the absence of disease or infirmity." They go on to say that "the enjoyment of the highest attainable standard of health is one of the fundamental rights of every human being without distinction of race, religion, political belief, economic or social condition."[5] Yet, health, both physical and mental, is influenced by a variety of these factors. Among these are: demographic factors, physical environment, social and economic environment, clinical care, and health behaviors. Therefore, in order to improve health in the United States, it is essential to understand the relationship between these factors and both types of health (physical and mental). However, for our data exploration, we chose to focus solely on mental health for two reasons: personal interest and experience with mental health research.

Analysis Goals

Both of us are interested in understanding the predictive factors of poor mental health. Past research shows that an individual who has several risk factors like the use of alcohol or drugs, poor nutrition, lack of sleep, traumatic life situations, or too few healthy relationships, could have a higher likelihood of developing mental health problems.[6] There are also demographic factors that contribute to mental health, which Emily explored this summer at Girls Inc., a national nonprofit, where she studied the current state of mental health in schools. She found that poor mental health often stemmed from demographic issues – where someone lives, a family's income level, or access to resources both in and out of school.

Therefore, both personal interest and experience led us to ask if the same individual risk factors and demographic factors can be analyzed on a per-county basis to improve mental health outcomes on a larger scale. Therefore, we will use a set of features (demographic factors, physical environment, social and economic environment, clinical care, health behaviors, and health outcomes) to predict the response variable, the number of mentally unhealthy days. This will allow us to analyze if there are similarities at the county-level to conceptualize if mental health can be better widely addressed, rather than addressed solely at an individual level. We will use the mean-squared test error to evaluate and compare model performance.

Significance

We hope that our analysis will contribute to a growing body of research on mental health by expanding our understanding of prevalent factors on a county-level. Our research will shed light on how leaders in our communities and nationwide can address the social determinants of health or demographic disparities, in efforts to improve mental health outcomes for all people in the United States.

---

[5] World Health Organization. (n.d). "Constitution of the World Health Organization.". https://www.who.int/about/governance/constitution

[6] American Mental Wellness Association (2021). "Risk and Protective Factors." https://www.americanmentalwellness.org/prevention/risk-and-protective-factors/

Data

Data Sources

Our dataset merged data from two sources, but used three total datasets. Two of the original datasets were pulled from the 2019 County Health Rankings National Data, collected by the County Health Rankings and Roadmaps.[7] County Health Rankings and Roadmaps is a program at the University of Wisconsin's Population Health Institute. The data collected provides a snapshot of a community's health and a starting point for discussing ways to improve health. From the County Health Rankings National Data, we used the "Ranked Measure Data" dataset and the "Additional Measures" dataset. Each dataset contained health information for each of the 3,142 counties across the 50 US states. Variables in both datasets spanned six major categories: health outcomes, health behaviors, clinical care, social and economic environment, physical environment, and demographic information.

The third dataset used for the project was pulled from the US decennial census, the census that is conducted every ten years in the United States, located on the University of Virginia's Demographics Research Group webpage.[8] This dataset contained more specific demographic information for the 3,142 counties found in the County Health Rankings and Roadmaps datasets. Variables included a percentage of homes in the county that have television or broadband internet and per-capita income information.

It is important to note that we focused solely on data from 2019 for two reasons. First, 2019 data for health and demographic features was more widely accessible than data from later years. It is much harder to locate 2020 or 2021 data that contains all the features we wanted to observe in our data exploration. Second, our response variable is "mentally unhealthy days." Essentially, our main research question was to investigate which health or demographic features are best at predicting this response. We know COVID-19 had a strong impact on mental health over the last year and a half. In fact, in June 2020, younger adults, racial minorities, essential workers, and unpaid adult caregivers reported having experienced disproportionately worse mental health outcomes.[9] Therefore, for our data exploration, we wanted to look at other underlying issues outside of COVID-19 for policymakers to address.

Data Cleaning

The central task to clean the data included merging the three sources and subsetting out variables that we did not want or that contained too many NA values. Data cleaning occurred in four distinct parts.

In the first part, we cleaned the demographic dataset. We first selected "fips", "state", "name", and any column that ended in "2019". From this subset of columns, we then selected the 2019 variables of interest for the data exploration ("household has broadband internet", "household

[7] County Health Rankings and Roadmaps.
[8] Guide to publicly available Demographic Data.
[9] Czeisler MÉ , Lane RI, Petrosky E, et al. (2020). Mental Health, Substance Use, and Suicidal Ideation During the COVID-19 Pandemic — United States, June 24–30, 2020. MMWR Morb Mortal Wkly Rep 2020;69:1049–1057. DOI: http://dx.doi.org/10.15585/mmwr.mm6932a1.

has television", "number of mobile homes", "per capita income", "persons per household", and "number of veterans"). The cleaned demographic dataset provides us with a snapshot of key demographic features for 3,142 counties in the United States.

In the second and third parts, we cleaned the "Ranked Measure Data" and the "Additional Features." For the "Ranked Measure Data," we started by creating a dummy variable for the "presence of violation" feature, which was coded as 1 if the county has a water violation and 0 if the county does not. We then removed any variable that reported a 95% confidence interval, quartiles, and most ratio variables (except income ratio, which measures income inequality). We wanted to look at rates and overall percentages of specific variables to extrapolate if a change in the rate or percentage impacts the reported mentally unhealthy days. The cleaning process for "Additional Features" followed a similar process to the cleaning process for "Ranked Measure Data." In both processes, we removed variables with a large number of NA values.

The fourth step to obtain the final, tidy dataset included joining the three datasets together using the fips code, a 5-digit unique geographic identifier, the state name, and the county name. We then removed variables with the largest number of NA values. Ultimately, we decided it was more important to remove some variables with many NA values in order to maintain a larger number of overall observations.

Data Description

Observations

The final, tidy dataset includes 2,371 rows and 65 columns. The 2,371 observations represent specific counties. If certain counties had not been removed due to the removal of NA values in the data cleaning process, we would have 3,142 observations as there are 3,142 total US counties.

Response Variable

Our continuous response variable is "mentally unhealthy days". This variable is the average number of mentally unhealthy days per month reported for each county. The mean number of mentally unhealthy days is 3.95. The minimum number reported is 2.5 days, and the maximum number reported is 6 days. Although these statistics appear low (as this is monthly data), it is important to remember that we are exploring county-level data, rather than individual-level data. This measure is an average for all individuals in a specific county – therefore, as individuals are less likely to report an extreme number of mentally unhealthy days, the average is on the lower side.

In our original exploration, we included a binary response variable ("mentally unhealthy") instead of a continuous response variable. We believed that a binary variable that classified a county as "mentally healthy" or "mentally unhealthy" would allow stakeholders to more easily determine which counties were mentally unhealthy. However, there were issues with using a binary response because the range of mentally unhealthy days is so small (range = 3.5). For example, if we decided on a cutoff of 4.5 days, what really distinguishes a county with an average of 4 mentally unhealthy days from a county of 5 mentally unhealthy days? Therefore, we decided to use the continuous variable, "mentally unhealthy days", instead for our response.

Explanatory Variables

There are 61 total explanatory variables. These variables can be divided into six categories: 1) demographic information, 2) health outcomes, 3) health behaviors, 4) clinical care, 5) social and economic environment, and 6) physical environment. For a detailed description of all features, refer to the Appendix.

Allocation of Data for Training and Testing

Before building our predictive models, we split our dataset into two subsets: the training dataset to build the predictive models, and the test dataset to evaluate the models. We used an 80-20 split. The training dataset contains 80% of our observations (1,897 observations), and the test dataset consists of 20% of our observations (474 observations). To ensure the 80-20 split leads to the same results each time, we set the seed before performing the split.

Exploratory Data Analysis: Response Variable

To better visualize the variation in our response variable, "mentally unhealthy days", across the different counties, we built a histogram and created heat-map visualizations.

Histogram of Mentally Unhealthy Days

We first illustrated the distribution of mentally unhealthy days among the 2,371 counties in the cleaned dataset. As seen in Figure 1, the data does not appear to be strongly right-skewed or left-skewed. In fact, most of the data is clustered around the mean (mean = 3.95), represented by the red-dashed line. The minimum number of mentally unhealthy days is 2.5 and the maximum number of mentally unhealthy days is 6. The range is 3. 5 days.

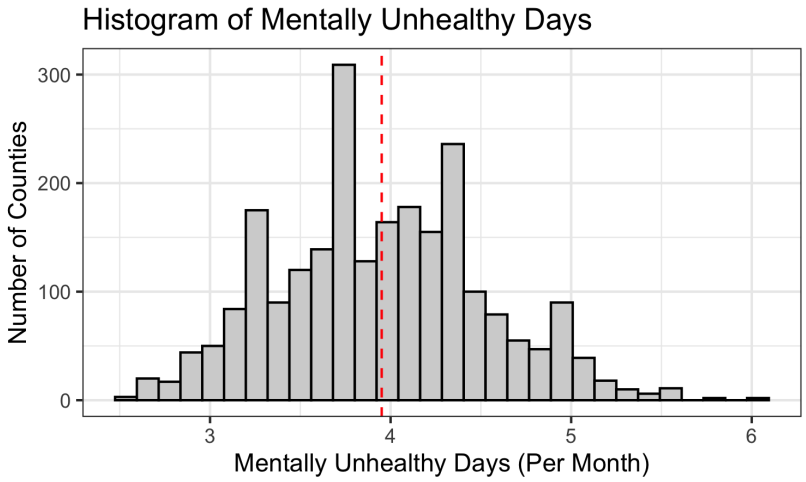


Figure 1: Histogram of Mentally Unhealthy Days

Heat Map of the United States - 3,142 Counties

Figure 2 displays the counties with the highest and lowest number of reported mentally unhealthy days across the United States in 2019. This heat map uses the "mentally unhealthy days" variable from the original "Ranked Measure Data," because we wanted to observe the overall distribution of mentally unhealthy days across every region in the United States. Each county is shaded a specific color, ranging from red to blue, to represent the average number of

mentally unhealthy days for that area. Red corresponds to higher values of mentally unhealthy days, and blue represents lower values.

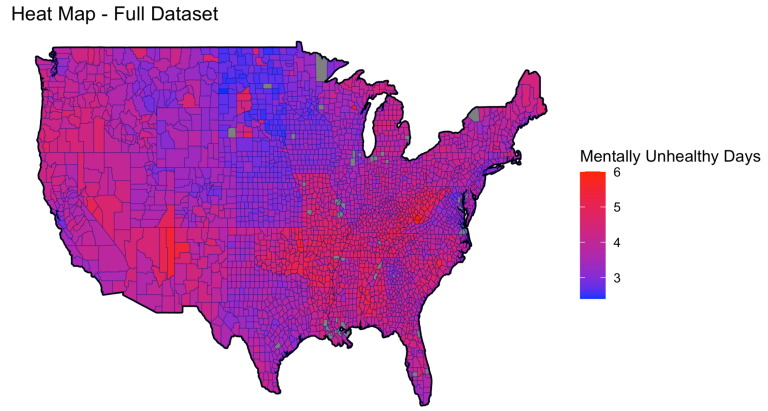


Figure 2: Heat Map of Full Dataset

We observe that counties in the South report a higher number of mentally unhealthy days on average, as seen by the clusters of red-shaded counties. In contrast, counties in the Midwest report a lower number of mentally unhealthy days, seen by the clusters of blue-shaded counties. Clusters of blue and red counties suggest that counties in similar locations (and in the same states) have similar features that contribute to mental health (whether that be environmental, social, or health related).

Heat Map of the United States - 2,371 Counties

Figure 3 also displays the counties with the highest and lowest number of reported mentally unhealthy days across the United States in 2019. This heat map uses the counties from the final, tidy dataset. The gray regions correspond to counties not present in the dataset, as we removed some counties due to NA values. The original observations hold, despite the removal of 771 counties. This visualization shows that a large portion of counties in the middle of the US were removed from our dataset.

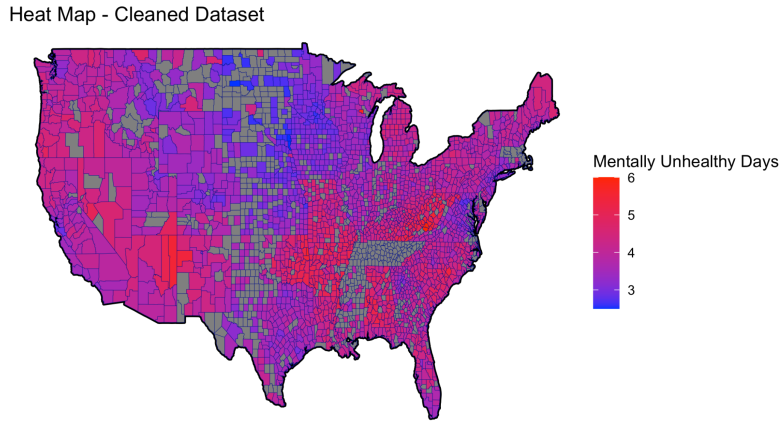


Figure 3: Heat Map of Cleaned Dataset

Healthiest and Unhealthiest Counties in the United States + Associated Heat Maps

We want to further contextualize which counties had the highest or lowest number of mentally unhealthy days from the final, tidy dataset. See below for the findings.

Highest Number of Mentally Unhealthy Days

We find that the counties with the highest number of reported average mentally unhealthy days are: McDowell, West Virginia (6.0 days) and Wyoming, West Virginia (6.0 days) (see Table 1). Additionally, we find that six of the top ten counties with the highest number of mentally unhealthy days are located in West Virginia. See Figure 4 for the distribution of mentally unhealthy days in West Virginia.

State	County	Mentally Unhealthy Days
West Virginia	McDowell	6.0
West Virginia	Wyoming	6.0
Oklahoma	Adair	5.8
Wisconsin	Menominee	5.8
West Virginia	Mingo	5.6
West Virginia	Nicholas	5.6
West Virginia	Raleigh	5.6
West Virginia	Summers	5.6
Arizona	Apache	5.5
Kentucky	Bell	5.5

Table 1: Highest number of mentally unhealthy days

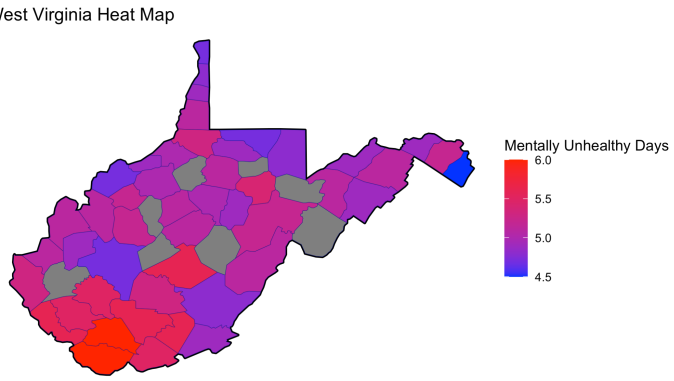


Figure 4: Heat Map of West Virginia

Lowest Number of Mentally Unhealthy Days

We find that the counties with the lowest number of reported average mentally unhealthy days are: Stark, North Dakota (2.5 days) and Lincoln, South Dakota (2.5 days) (see Table 2). Additionally, we find that nine of the top ten counties to report low mentally unhealthy days are located in the Dakotas (five in North Dakota and four in South Dakota). See Figure 5 for the distribution of mentally unhealthy days in North Dakota. Note, we observe many missing values.

State	County	Mentally Unhealthy Days
North Dakota	Stark	2.5
South Dakota	Lincoln	2.5
South Dakota	Union	2.5
Minnesota	Carver	2.6
North Dakota	Morton	2.6
North Dakota	Richland	2.6
North Dakota	Ward	2.6
North Dakota	Williams	2.6
South Dakota	Hutchinson	2.6
South Dakota	Lake	2.6

Table 2: Lowest number of mentally unhealthy days

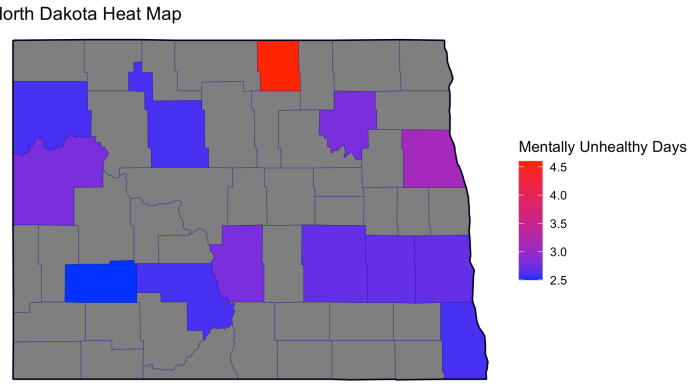


Figure 5: Heat Map of North Dakota

Exploratory Data Analysis of the Training Dataset

In the next part of our exploration, we looked at the correlation between features within specific subcategories, and the relationship between a set of selected features and the response using the training dataset.

Feature Exploration

We seeked to obtain a high-level view of the correlations that may exist among our explanatory variables. We looked at the correlations between variables for each category of features (excluding *demographic information*). We decided not to include the *demographic information* subcategory because we already know these features are highly correlated. For instance, the higher the percentage of females in a certain county, the lower the percentage of males in a county. However, it was important to explore correlations between features from other subcategories, as highly correlated variables may lead to issues with model interpretation. For instance, if variable x and variable y are highly correlated, and variable x strongly predicts the response, we can probably deduce that variable y will also strongly predict the response.

Health Outcomes

In Figure 6, we find that "physically unhealthy days" and "percent diabetic" are negatively correlated with "life expectancy". This follows our intuition as sicker individuals tend to have a shorter predicted life expectancy. "Physically unhealthy days" is positively correlated with "percent diabetic."

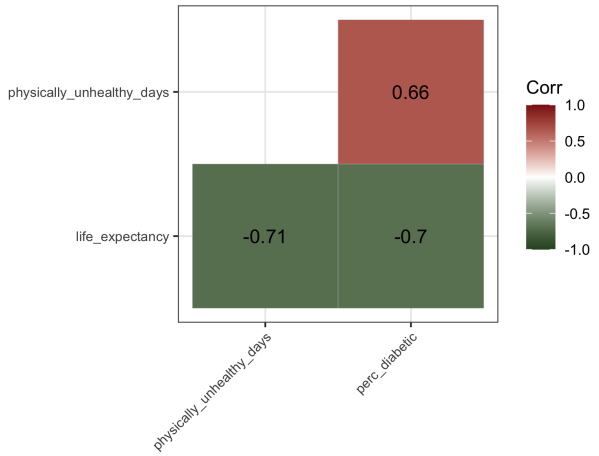


Figure 6: *Health Outcome* Variable Correlations

Health Behavior

Figure 7 shows that many negative health behaviors (i.e., smoking, physical inactivity, and sleep deprivation) are positively correlated with one another. For example, "percent obese" has a strong, positive correlation with "percent physical inactive" (correlation = 0.7). Interestingly, we observe that "percent excessive drinking" is *negatively* correlated with other negative health behaviors– it seems that higher percentages of binge drinking are correlated with lower percentages of negative health behaviors in a specific county. We would expect the opposite result, yet the data does not display this.

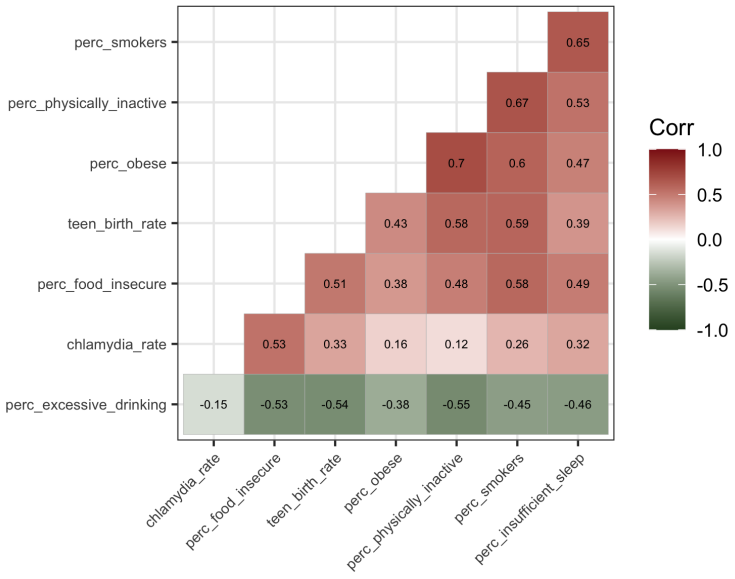


Figure 7: *Health Behavior* Variable Correlations

Clinical Care

In Figure 8, we observe that higher percentages of health-related services (such as vaccinations, mammography screenings, and access to mental health providers) are positively correlated with one another. “Percent uninsured” and “preventable hospital rates" are negatively correlated with the same features listed prior. Essentially, in counties with more uninsured individuals or higher preventable hospitalization rates, there are lower rates of health-related services.

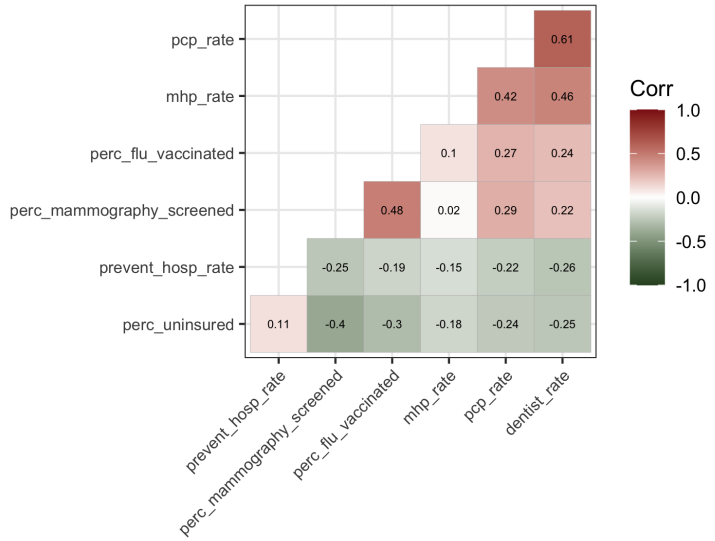


Figure 8: *Clinical Care* Variable Correlations

Social and Economic Environment

Among the majority of the social and economic features in Figure 9, there are positive correlations between income, poverty, unemployment, and crime-related features. Additionally, there are negative correlations between these same features and homeownership, education, and technological features. This follows our intuition as less affluent counties, with higher rates of

poverty, crime, and unemployment, probably have community members with lower likelihoods of attaining an education or owning a home. The strongest, positive observed correlation is between “household has computer” and “household has broadband” (correlation = 0.9), as households need broadband internet access if they own a computer. “Homeownership” and “housing mobile homes” have the strongest, negative observed correlation (correlation = -0.97). Owning a mobile home might not be categorized as being a homeowner.

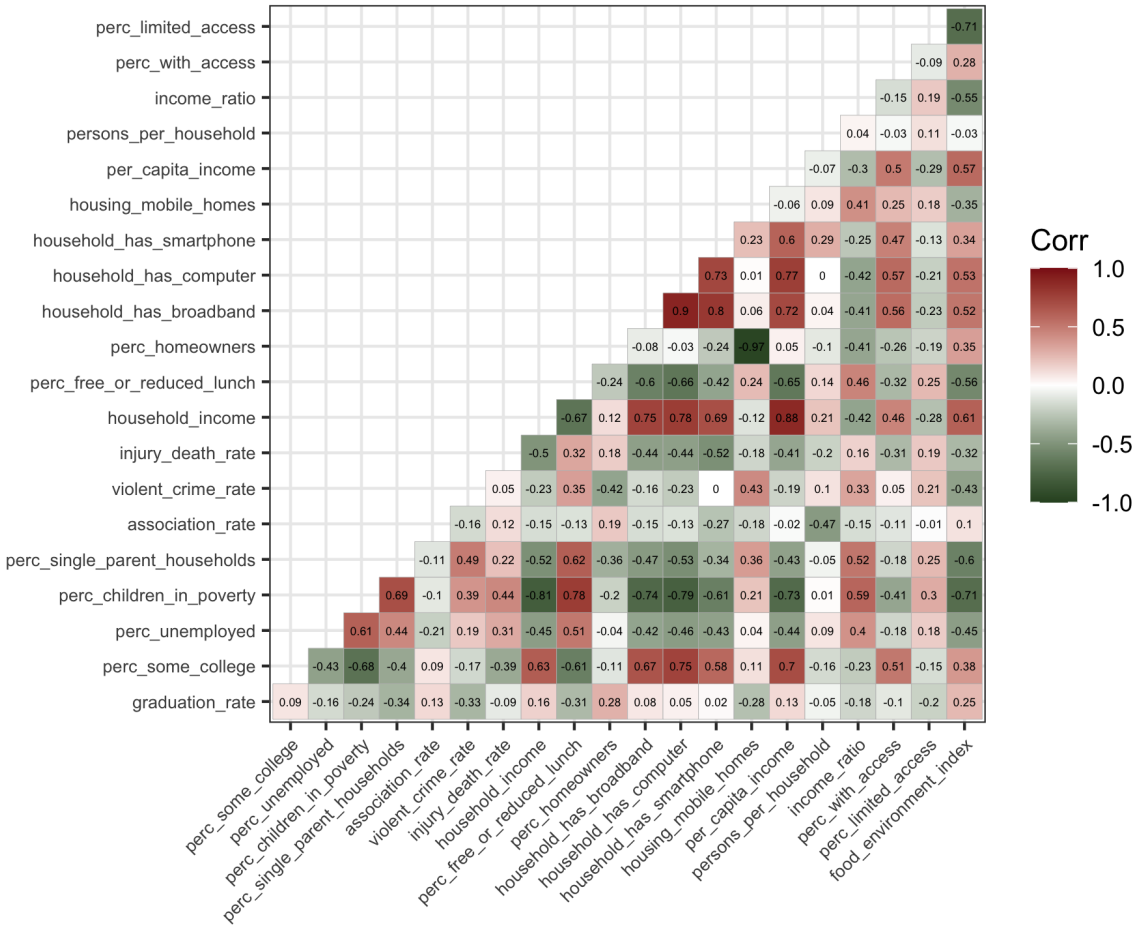


Figure 9: *Social and Economic Environment* Variable Correlations

Physical Environment

The positive and negative correlations in Figure 10 confirm our intuition. For example, “percent long commutes drive alone” is positively correlated with “percent rural”. Individuals need to travel further to get to their jobs if they live in a rural county (more jobs are located in metropolitan areas). The strongest, positive correlation is between “percent severe housing cost burden” and “percent severe housing problems” (correlation = 0.84). “Percent severe housing problems” is defined as the percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, or lack of kitchen or plumbing facilities. Essentially, “percent severe housing problems” encompasses “percent severe housing cost burden.”

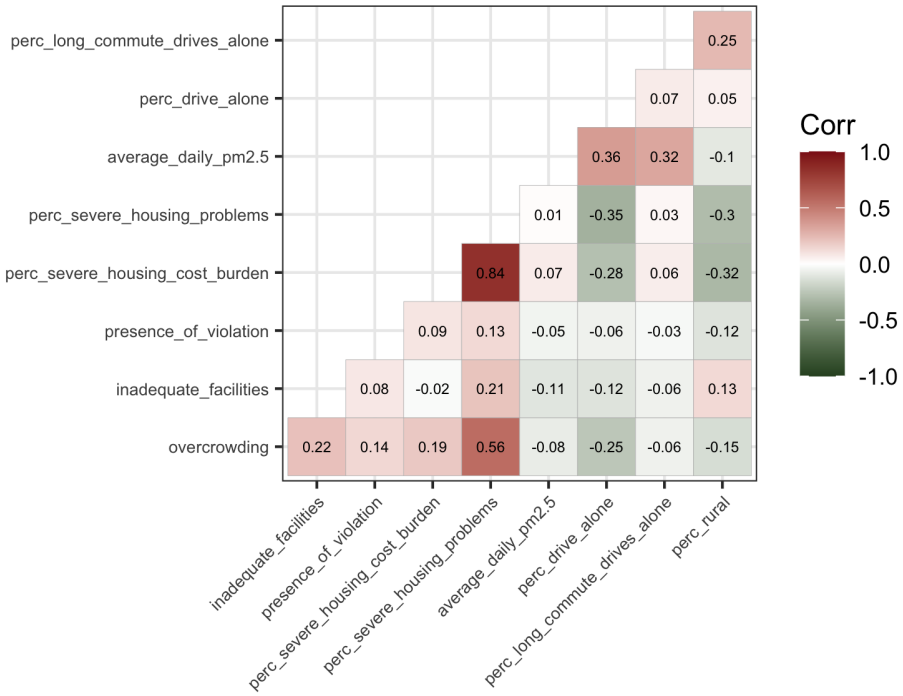


Figure 10: *Physical Environment* Variable Correlations

Feature and Response Exploration

For the final part of our data exploration, we observe the relationship between one feature from each of the feature subcategories and the response variable. We also choose to include the relationship between excessive drinking and the response variable as we observe different expected results when comparing excessive drinking with other negative health behaviors.

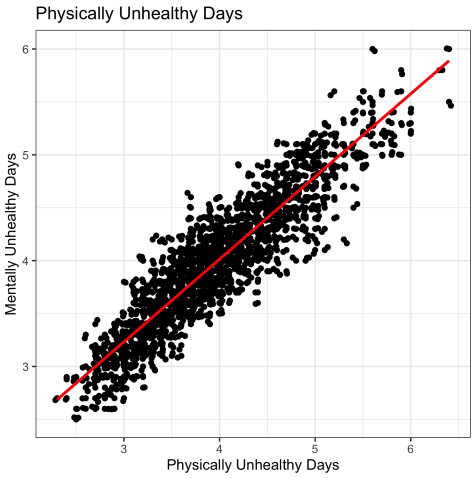


Figure 11: Physically Unhealthy Days

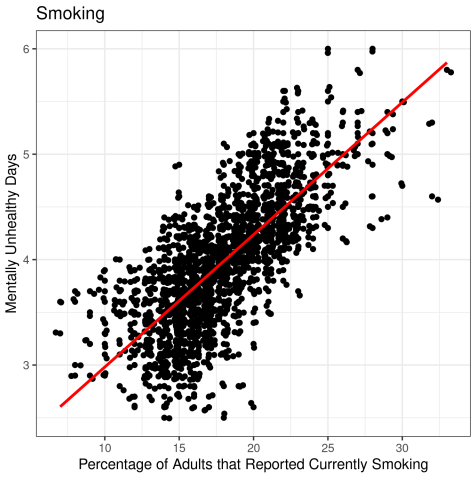


Figure 12: Smoking

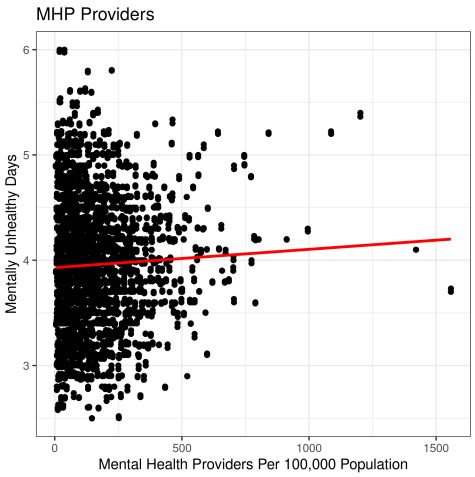


Figure 13: MHP Providers

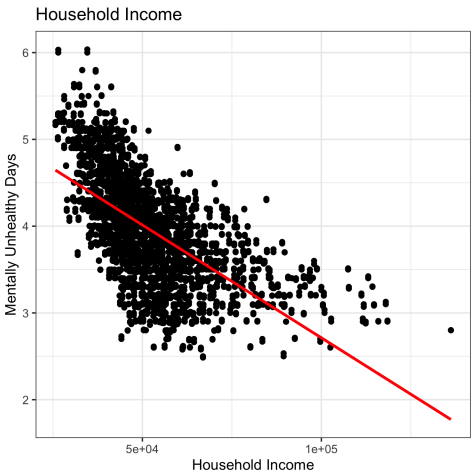


Figure 14: Household Income

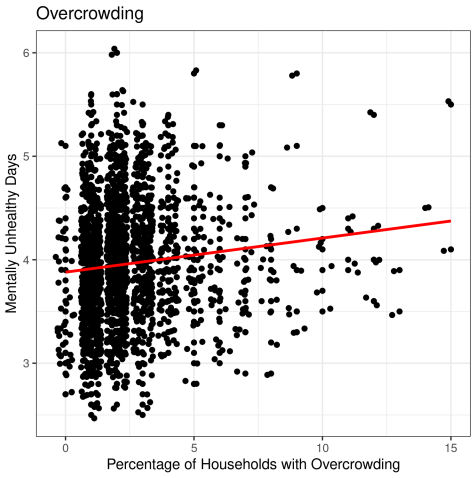


Figure 15: Overcrowding

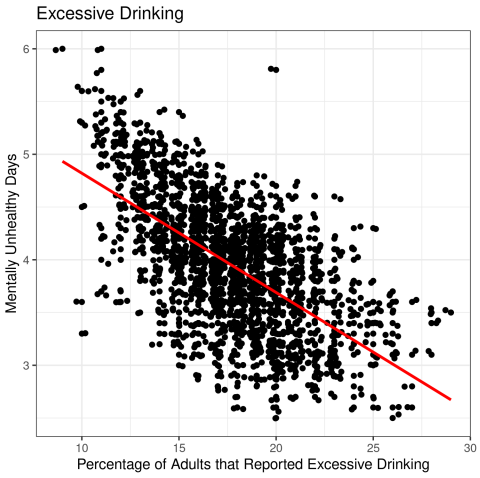


Figure 16: Excessive Drinking

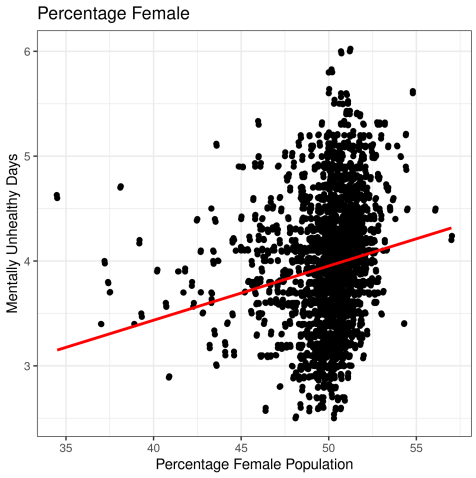


Figure 17: Female

Physically Unhealthy Days

"Physically unhealthy days" is a part of the *health outcomes* subgroup. Observed in Figure 11, there is a strong relationship between "physically unhealthy days" and "mentally unhealthy days". In research, physically and mentally unhealthy days are often taken together to determine quality of life,[10] showing that they are very highly correlated. To avoid the possibility that the number of physically unhealthy days may include responses from the number of mentally unhealthy days (essentially making it another way to measure the response in disguise), we remove "physically unhealthy days" from our explanatory variables in our predictive models.

Percentage of Adults that Report Smoking

"Percentage smokers" is part of the *health behavior* subgroup. Observed in Figure 12, there is a strong, positive relationship between smoking and the number of mentally unhealthy days. In other words, the higher the percentage of smokers for a specific county, the more likely a county will report a higher number of mentally unhealthy days. Therefore, we expect smoking to be a prevalent predictive factor when we run our predictive models.

Mental Health Providers Per 100,000 Population

"Mental health providers" is part of the *clinical care* subgroup. Observed in Figure 13, there is a weak, positive relationship between the number of mental health providers and the number of mentally unhealthy days. This suggests that as the number of mental health providers per 100,000 increases, the more likely a county reports a high number of mentally unhealthy days. While this relationship is slightly surprising (mental health normally improves with support), it appears that the correlation is very weak. Additionally, it appears that there is much more data for counties with a lower number of mental health providers. This might skew the true relationship between this feature and the response.

Household Income

"Household income" is part of the *social and economic environment* subgroup. Observed in Figure 14, there is a strong, inverse relationship between household income and the number of mentally unhealthy days. As household income increases, counties report a lower number of mentally unhealthy days, on average. We expect household income to be a prevalent predictive factor when we run the predictive models.

Percentage of Households With Overcrowding

"Percentage of households with overcrowding" is part of the *physical environment* subgroup. Observed in Figure 15, there is a moderate relationship between the percentage of households with overcrowding and the number of mentally unhealthy days. As the percentage of households with overcrowding increases in a county, the higher the number of mentally unhealthy days.

Percentage of Adults that Reported Excessive Drinking

"Percentage excessive drinking" is part of the *health behavior* subgroup. Observed in Figure 16, there is a strong, inverse relationship between excessive drinking and the number of mentally

[10] Centers for Disease Control and Prevention. (2000). Measuring Healthy Days - Population Assessment of Health-Related Quality of Life. https://www.cdc.gov/hrqol/pdfs/mhd.pdf

unhealthy days. As the percentage of adults that report excessive drinking increases, the number of mentally unhealthy days decreases. This relationship is surprising: research shows that drinking should have a negative impact on mental health.[11] However, the County Health Rankings and Roadmaps data shows the opposite effect. This is something to be aware of as we interpret the results of the predictive models.

Percentage Female Population

"Percentage female" is part of the *demographic information* subgroup. Observed in Figure 17, there is a weak, positive relationship between the percentage of females in a county and the number of mentally unhealthy days that are reported. This plot is interesting as the majority of the data is clustered around populations that are evenly divided between male and female. However, it appears that there are some counties with a lower percentage of females, and those counties tend to have a lower number of mentally unhealthy days. Therefore, the red trend line slopes upward and suggests that a higher number of females in a county is associated with a higher average number of mentally unhealthy days.

Model Building, Evaluation, and Interpretation

With the data exploration in mind, we fit six models: (1) an ordinary least squares regression, (2) a ridge regression, (3) a lasso regression, (4) a decision tree, (5) a random forest, and (6) a boosted tree.

Regression and Shrinkage Methods

Ordinary Least Squares Regression

Using the training data, we began our analysis with an ordinary least squares (OLS) regression of "mentally unhealthy days" on 60 of the 61 explanatory variables (excluding physically unhealthy days). We did not include "physically unhealthy days" in our model, as this feature was too highly correlated with the response (see Feature vs. Response Exploration: Figure 11).

The OLS regression revealed many interesting results. 34 out of the total 60 features were at least marginally significant, and 23 out of the 34 marginally significant coefficients were extremely significant at the 0.001 level. The top three most significant features with a positive coefficient were: "percent smokers", "percent female", and "percent diabetic". Essentially, counties with higher values of these features are more likely to report a higher number of mentally unhealthy days.

It appeared that the *demographic information* subgroup had very significant, negative features, which included: "percent African American", "percent Asian", and "percent non-Hispanic White". Surprisingly, we observed significant, negative coefficients for "percent excessive drinking" and "percent unemployed". The coefficient for "percent excessive drinking" follows what we observed in Figure 16, the scatterplot of excessive drinking versus the number of mentally unhealthy days, where an increase in the percentage of adults that reported excessive drinking corresponds to a decrease in the number of mentally unhealthy days reported by a county. Regardless, we continue on in our analysis.

[11] American Mental Wellness Association

Using the ordinary least squares model, we calculated predictions for the number of mentally unhealthy days for the counties in the test dataset. From these predictions, we found that the training mean-squared error was equal to 0.05973 and the test mean-squared error was equal to 0.06005. Refer to the Appendix for the summary of the OLS regression model.

Lasso, Ridge, and Elastic Net

Fitting an ordinary least squares model with so many explanatory variables may incur a large variance cost and lead to suboptimal predictions. Therefore, we turned to using more parsimonious models to decrease model complexity and hopefully increase prediction accuracy. In this section, we run three predictive models: (1) ridge, (2) lasso, and (3) elastic net.

Ridge

The main goal of ridge regression is to shrink the size of the coefficients through penalizing complex models. The first step in fitting the ridge regression was to determine the optimal value of lambda, the shrinkage parameter, using a ten-fold cross-validation. Based on the one-standard-error rule, we found the optimal value of lambda to equal 0.076. See Figure 18 for the ridge trace plot, which outlines the top ten features in the model.

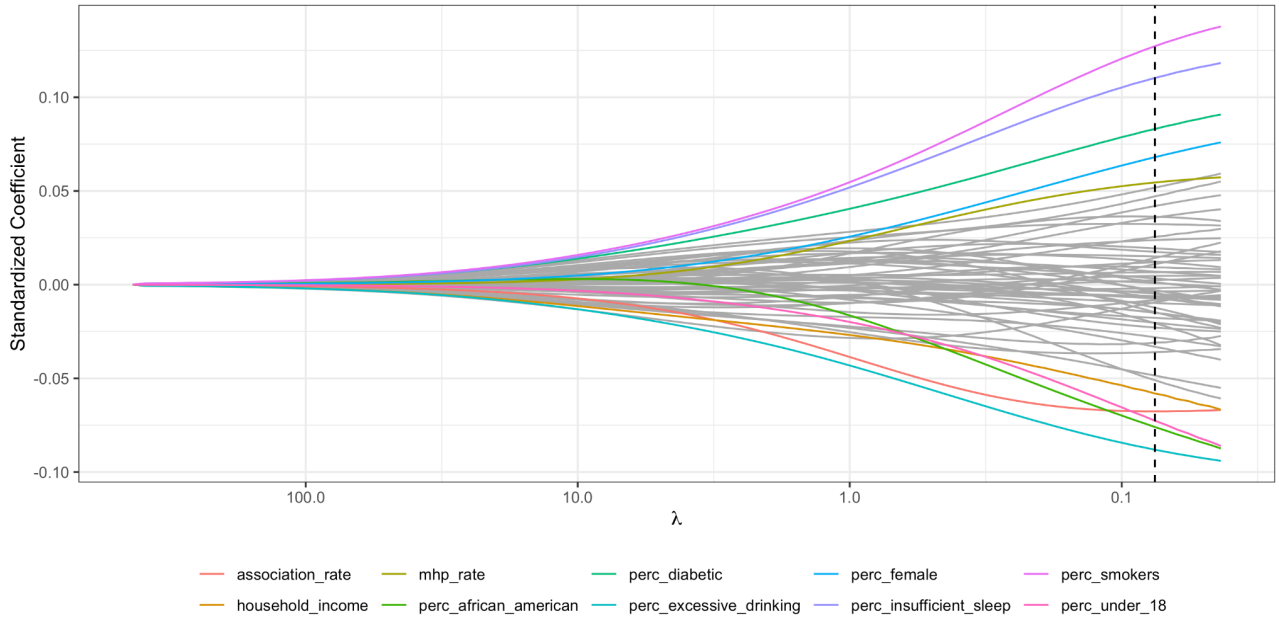


Figure 18: Ridge Trace Plot

The top three features with the largest positive coefficients were: “percent smokers”, “percent insufficient sleep”, and “percent diabetic”. The top three features with the most negative coefficients were: “percent excessive drinking”, “percent African American”, and “percent under 18”. Many of these features overlapped with the most significant features we found from the OLS regression model. Looking at the standardized feature coefficients for ridge (see Appendix), we found that “percent smokers” has the highest magnitude coefficient, equal to 0.127. We can interpret this coefficient in the following way: an increase in the percentage of smokers by one

standard deviation (3.44%) leads to a predicted increase in the number of mentally unhealthy days by 0.127.

Using our ridge regression, we calculated predictions for the number of mentally unhealthy days for the counties in the test dataset. We found that the training mean-squared error was equal to 0.06490 and the test mean-squared error was equal to 0.06411.

Lasso

The main goal of a lasso regression is to select coefficients through penalizing complex models. In a similar fashion to the ridge regression, we used a ten-fold cross-validation to build the lasso model. Using the optimal value of lambda based on the one-standard-error rule (lambda = 0.00259), we found that 46 features are selected. See Figure 19 for the lasso trace plot, which outlines the top eight features in the model.

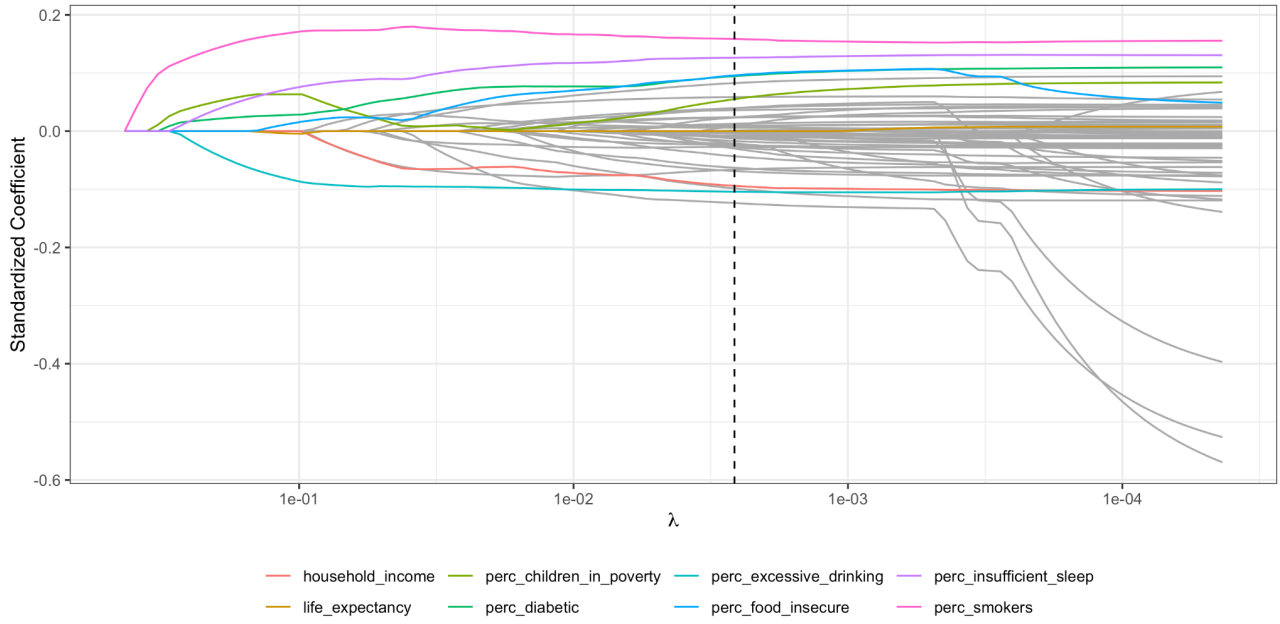


Figure 19: Lasso Trace Plot

The first feature to enter the model with a positive coefficient is “percent smokers”, and the first feature to enter the model with a negative coefficient is “percent excessive drinking”. The top three features with the largest positive coefficients are: “percent smokers”, “percent insufficient sleep”, and “percent food insecure”. The top three features with the most negative coefficients are: “percent African American”, “percent excessive drinking”, and “percent under 18”. These features are consistent with our findings from the ridge and OLS regressions. “Percent smokers” has the highest magnitude coefficient of 0.159, which means that an increase in the percentage of smokers by one standard deviation (3.44%) leads to an increase in the predicted number of mentally unhealthy days by 0.159. Refer to the Appendix for the top ten features selected by the lasso model.

Using the lasso regression, we calculated predictions for the number of mentally unhealthy days for each of the counties in the test dataset. We found that the training mean-squared error was equal to 0.06219 and the testing mean-squared error was equal to 0.06139.

Elastic Net

We next attempted to fit an elastic net predictive model, but did not complete the analysis because the optimal value of alpha that was selected by the model was equal to one. If we set alpha equal to one, this would be the same as fitting a lasso model. This result shows us that the lasso regression is the most effective parsimonious model, as only a certain subset of the features are actually effective in predicting the response. Therefore, we did not complete the elastic net predictive model as it would output the same results as the lasso model.

Tree-Based Methods

We then proceeded to fit tree-based methods to the data to explore any nonlinearity that may exist. As these methods are easily interpretable, we can observe and illustrate the features that are most important in predicting the number of mentally unhealthy days. We use three tree-based methods: (1) a decision tree, (2) a random forest, and (3) a boosted tree. Since decision trees are somewhat unstable and do not give the best predictive performance, we turn to random forests and boosting to eliminate the extra variance and instability in the model.

Decision Tree

We found the optimal decision tree, by first fitting the deepest possible tree. We then pruned back the deepest possible tree to find the optimal tree, which included 17 splits and 18 terminal nodes.

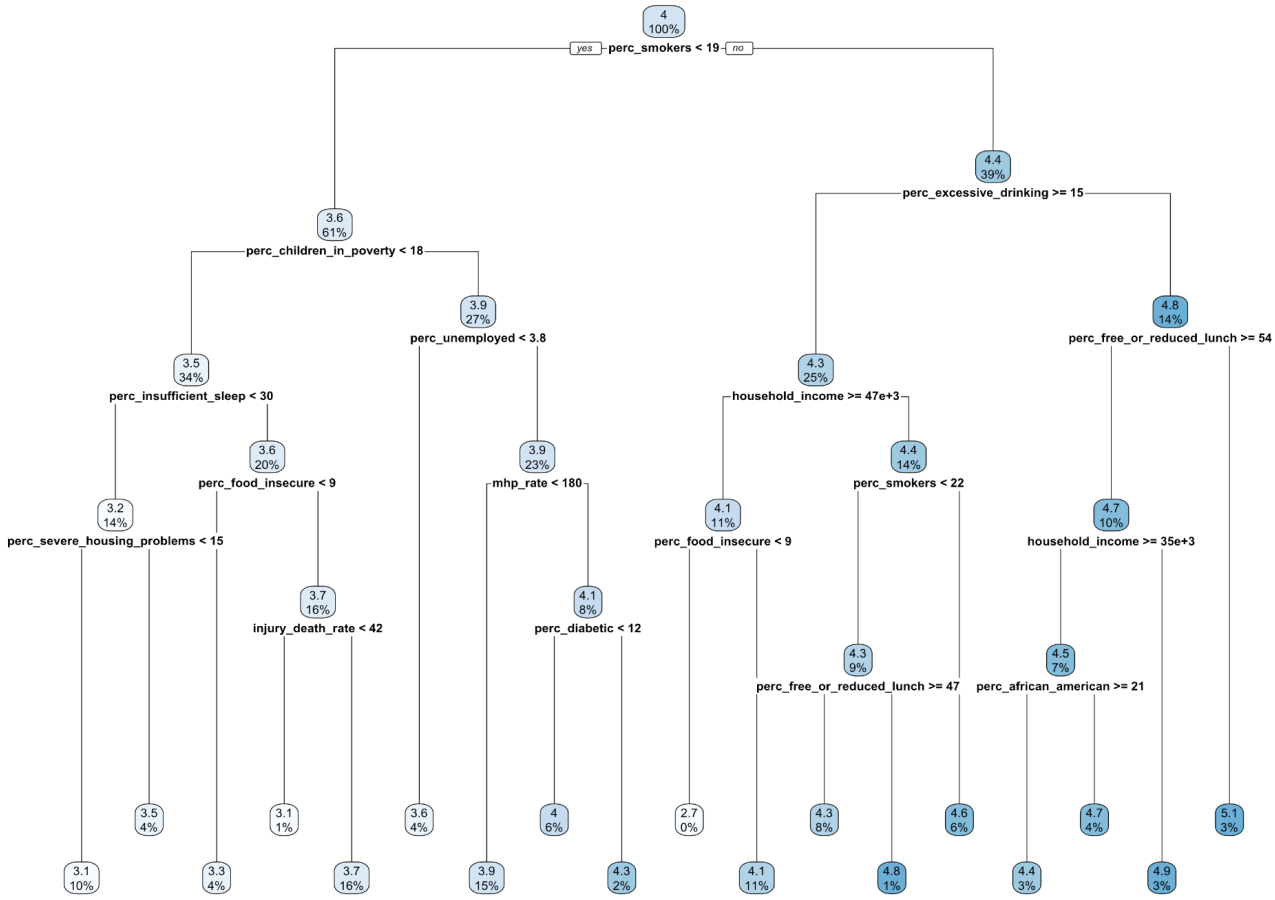


Figure 20: Optimal Decision Tree

As seen in Figure 20, the 10th terminal node has the lowest predicted number of mentally unhealthy days (prediction = 2.7 days). While this terminal node (supposedly) has 0% of the training data, this is probably just due to rounding. To get to this node, we follow a set of decisions based on specific features: “percent smokers” is greater than 19%, “percent excessive drinking” is greater than 15%, “household income” is greater than or equal to $47,000, and “percent food insecure” is less than 9%.

The right-most terminal node has the highest predicted number of mentally unhealthy days (prediction = 5.1 days). This terminal node contains 3% of the training data. To get to this node we follow a set of decisions based on specific features: “percent smokers” is greater than 19%, “percent excessive drinking” is less than 15%, and “percent free or reduced lunch” is less than 54%. Two of the three feature splits follow our intuition. As the percentage of smokers increase, the number of mentally unhealthy days reported also increases, and as the percentage of free or reduced lunch in schools decreases, the number of mentally unhealthy days reported increases. The split on the “percent excessive drinking” is less than 15% does not follow our intuition of the relationship between mental health and alcohol consumption, however it does follow the previous findings of binge drinking and mental health earlier in the data exploration.

Using this model, we find that the training mean-squared error is equal to 0.08986 and the test mean-squared error is equal to 0.09927.

Random Forest

Next, we chose to build a random forest to reduce the variance of the decision tree. We start by fitting a random forest with default parameters. We next tune the random forest by cross-validating using various values of *m,* the number of selected features to split on. In varying *m* from 1 to 60 (the total number of features included in the model), we find that the optimal value of *m* is equal to 31 features. The cross-validation plot is found in Figure 21.

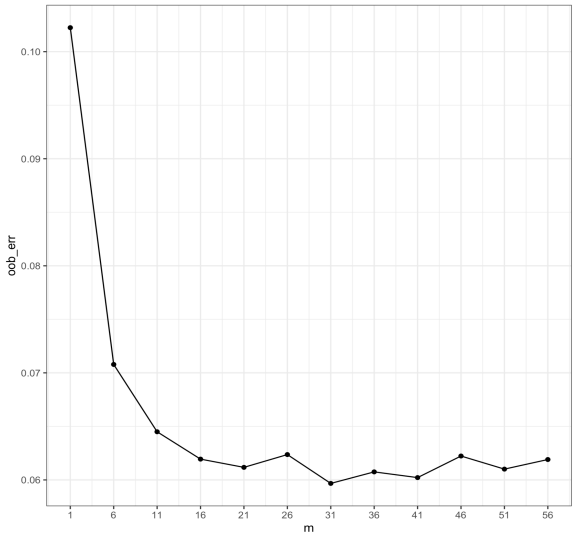


Figure 21: Cross-Validation for Number of Selected Features (m)

We then build a random forest with the optimal value of *m* (m = 31) and 500 trees using the training data. From the optimal random forest, we build Figure 22, the variable importance plot.

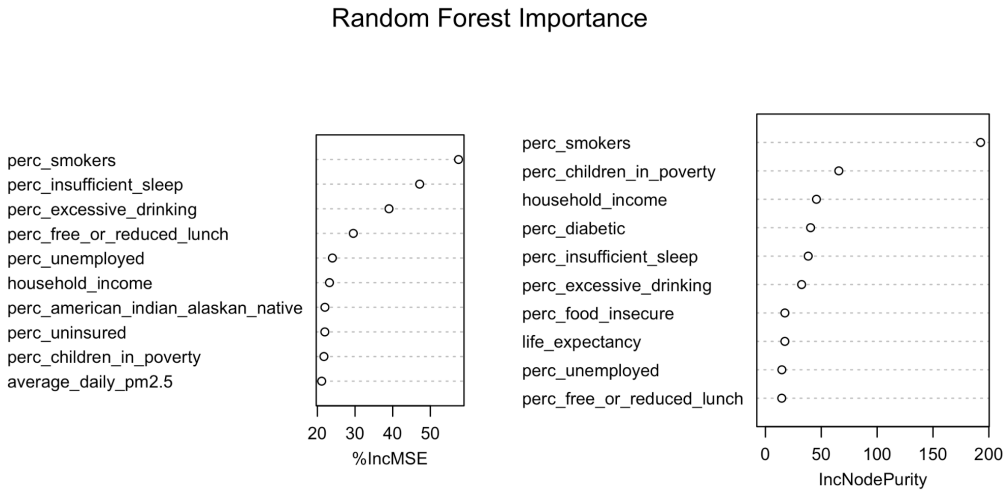


Figure 22: Random Forest Variable Importance Plot

Looking at the variable importance plot, we find that the top features are ranked according to (1) the *% MSE* metric and (2) the *Node Purity* metric. The three most important features according

to the *% MSE* metric are: “percent smokers”, “percent insufficient sleep”, and “percent excessive drinking”. The three most important features according to the *Node Purity* metric are: “percent smokers”, “percent children in poverty”, and “household income”. We find that out of the ten total features listed, seven are listed in both.

For similar reasons to the other models, these predictive features follow our intuition. For example, smoking is strongly and positively correlated with an increase in mentally unhealthy days. It is also important to note that many of the important variables from the random forest model are the same as the variables from the OLS regression, the lasso and ridge models, and the decision tree. We find that in all models, these variables are highly predictive of how many mentally unhealthy days a county will report.

Using the random forest model, we find that the training mean-squared error is equal to 0.00918 and the test mean-squared error is equal to 0.05250.

Boosted Model

We fit three boosted tree models with interaction depths of one, two, and three, using a shrinkage factor of 0.1, 1000 trees, and 5-fold cross-validation. We find that the optimal interaction depth is equal to three, as this depth has the minimum cross-validation error (as pictured in Figure 23).

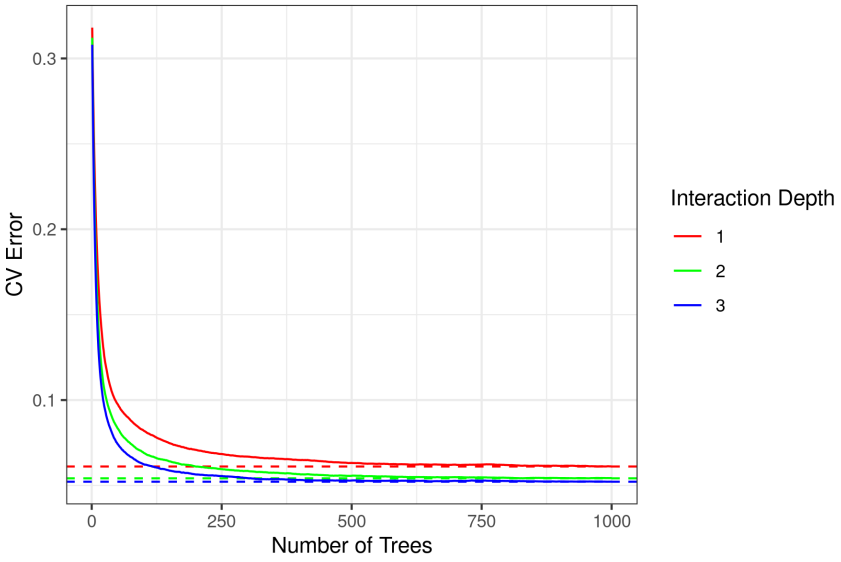


Figure 23: Cross-Validation for Number of Trees, for Each Interaction Depth

The optimal number of trees using an interaction depth of three is equal to 992. Using these optimal values, we look to the relative influence table (Table 3) to determine the top three most influential features. The top three features are “percent smokers” (influence = 29.81), “household income” (influence = 6.74), and “percent insufficient sleep” (influence = 6.46).

Variable	Relative influence
perc_smokers	29.81
household_income	6.74
perc_insufficient_sleep	6.46
perc_children_in_poverty	5.65
perc_excessive_drinking	5.62
perc_food_insecure	5.55
perc_diabetic	2.21
association_rate	1.97
injury_death_rate	1.81
perc_unemployed	1.72

Table 3: Relative Influence Table for Boosted Model

Interestingly, it appears that the relative influence of "percent smokers" is over four times larger than the relative influence of any other feature. This shows that the percentage of adults who smoke has the greatest predictive influence on the number of mentally unhealthy days reported in a county, which is a consistent finding from almost every other model. We also observe that the top three features from Table 3 directly align with the top three features from the random forest importance plot, whether that be based on the *% MSE* metric or the *Node Purity* metric (see Figure 22).

Next, we present partial dependence plots for the two most important variables: "percent smokers" (Figure 24) and "household income" (Figure 25). These plots are only an approximation of the relationship between each feature and the response, as our optimal interaction depth is equal to three.

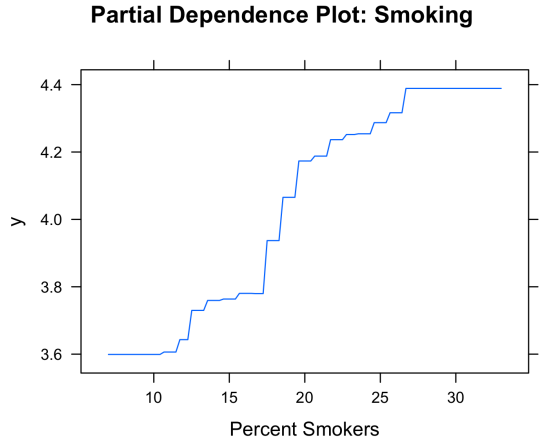


Figure 24: Partial Dependence Plot - Smoking

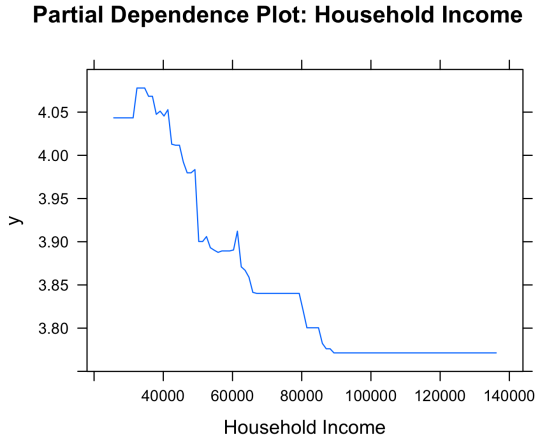


Figure 25: Partial Dependence Plot - Income

From Figure 24, we see that the percentage of smokers shows a positive (i.e. increasing) relationship with the number of mentally unhealthy days. This follows our intuition as smoking is inversely related to mental health. Figure 25 shows a negative relationship between household income and the number of mentally unhealthy days. Essentially, the higher a county's average

household income, the lower the number of predicted mentally unhealthy days. This also follows intuition as financial stability might have a direct influence on mental health.

Finally, we find that the training mean-squared error for the boosted model is equal to 0.00922 and the test mean-squared error is equal to 0.04913.

Conclusion

Comparison of Method Performance

Table 4 displays the mean-squared errors (training and test) for each model: ordinary least squares, ridge, lasso, decision tree, random forest, and boosted tree. Table 4 allows us to compare our six models on account of method performance and predictive accuracy.

Model	Train MSE	Test MSE
OLS	0.05973	0.06005
Ridge	0.06490	0.06411
Lasso	0.06219	0.06139
Decision Tree	0.08986	0.09927
Random Forest	0.00918	0.05250
Boosting	0.00922	0.04913

Table 4: Training and Test Errors for Each Model

The model with the lowest test error is the boosted tree (error = 0.04913), followed by the random forest (error = 0.05250). After boosting and random forest, the best performances are the OLS regression (error = 0.06005), lasso regression (error = 0.06139), and ridge regression (error = 0.06411). The model with the highest test error is the decision tree (error = 0.09927). The order from best to worst test error is understandable as boosted trees and random forests tend to have high predictive accuracy. However, it is important to note that for both boosting and random forest, the training MSE is much smaller than the test MSE, which suggests a level of overfitting in the models. Lasso also appears to be a good predictor which suggests that a smaller subset of the total number of features are predictive of the number of mentally unhealthy days.

While there are definitely differences in the test errors of the above models, it appears that each model identifies similar features as important for predicting the response variable ("mentally unhealthy days").

Overall Conclusions, Recommendations, and Takeaways for Stakeholders

A number of key features that impact the average number of mentally unhealthy days reported by a county can be seen across all models. These features primarily fall into the *health behaviors* and the *social and economic environment* subcategories. The most commonly selected features from the *health behaviors* subgroup are: "percent smokers", "percent insufficient sleep", and "percent excessive drinking". The most commonly selected features from the *social and*

environment subgroup are: “household income” and “percent free or reduced lunch”. In summary, we conclude that regardless of which model yields the lowest test error, all models find that social determinants and health-behavior factors are highly predictive of the average reported number of mentally unhealthy days.

It is important to note that although the features in the *health behavior* subcategory display the average responses within the county, the actual factors address individual behavior and personal choices. For example, the decision to smoke is an individual decision, rather than a reflection of county inequity (like income inequality). In contrast, social and economic determinants comprise the conditions in the environments where people are born, live, learn, and age. With this distinction in mind, recommendations and takeaways for stakeholders fall into two groups: 1) addressing individual action and behavior, and 2) addressing county-based inequities.

For individual action, counties should promote healthy living and behaviors. For example, given smoking was very predictive of reporting a higher number of mentally unhealthy days, counties can create programs in schools or among specific communities that have higher smoking rates to discourage people from smoking. If our conclusions are indeed correct, we can direct counties to extract more specific data for the prevalence of certain unhealthy behaviors within their communities to create better targeted healthy living and behavior programs.

To address social and environmental factors (i.e., household income or resources in schools), stakeholders should make increased efforts to support policies that advocate for a better distribution of wealth across the county, state, or the United States. Additionally, we found that the percentage of those who receive free or reduced lunch in schools predicts the number of reported mentally unhealthy days. From this information, stakeholders can advocate for better funding for disadvantaged school districts to ensure equal access to food and resources.

Limitations

Data Limitations

As detailed in the Data Description section, our primary response variable reports the average number of mentally unhealthy days, rather than the specifics on mental health disorders for each county (i.e. depression, anxiety, etc.). Thus, we must be aware that the models select features that are important *solely* to the prediction of mentally unhealthy days, and not to the prediction of specific mental health disorders. Therefore, we can only use this analysis for general conclusions about mental health in counties across the United States.

We recognize that our final dataset contained only a subset of the total number of counties in the United States due to removing counties with NA values. Overwhelmingly, the counties removed from the original dataset were located in the middle of the United States. As seen in Figure 2, the heat map of the 3,142 US counties, a majority of the counties with the lowest number of mentally unhealthy days were clustered in that middle region. Therefore, after removing these counties, we run the risk of underrepresenting all counties in the United States.

The use of data from 2019 is the final limitation of our dataset. While we did this to avoid the strong influence of COVID-19, data from 2019 might be too outdated to conduct an accurate

analysis of mental health among US counties. Separate from the impact of COVID-19, there have been drastic changes in the world since 2019 (technological improvements, changes in government structure, etc). Analysis performed with more recent data might find new features to be even more predictive of the number of mentally unhealthy days than the features currently in our dataset. Regardless, data from 2019 is still recent, so, in theory, our analysis probably would not deviate that much even if we used more recent data. Instead, we should be aware of this in making recommendations to stakeholders.

Analysis Limitations

We built predictive models, so conclusions from the models indicate correlations, rather than causation. Therefore, we must be careful when reporting that a feature in the data *causes* an increase or decrease in the number of mentally unhealthy days. These are general observations on the relationship between a specific feature and the response, rather than concrete results. Additionally, from the Exploratory Analysis of the Training Dataset section, there is evidence of correlation among some of our explanatory variables. The true predictive importance of a feature may be reported inaccurately if that feature is highly correlated with another. Therefore, this is something to keep in mind when discussing feature importance.

We must also acknowledge the features excluded from our data analysis. We removed a number of features due to NA values during the data cleaning process. Many of these features broke down existing features into demographic categories. For example, we removed the demographic breakdown of "percent children in poverty" for Black, Hispanic, and white subgroups. The overall "percent children in poverty" feature was important to predict the number of mentally unhealthy days (see Figure 22 and Table 3). Including demographic breakdowns for the features could have yielded a different analysis.

Recommended Follow-Up Analysis

To address the above limitations, we should replicate the analysis with a more extensive dataset. In other words, our dataset should be more recent and contain a larger number of the total United States counties. This data would be more representative of all counties, and it would likely have a higher predictive accuracy than our current models. However, it is important to recognize that collecting this type of data on such a large scale (every county in the United States) might be a difficult accomplishment.

To better predict specific mental health disorders (i.e. depression, anxiety, etc.), rather than the average number of mentally unhealthy days, we should use a different response variable. Perhaps in future iterations of this analysis, it would be interesting to predict the percentage of individuals in a specific county with a certain mental health diagnosis. This would allow stakeholders to better understand the factors that have a strong relationship with mental health, and to target those factors to prevent or limit certain disorders. In even further iterations of this analysis, it might be interesting to look into how these features, whether they be social, economic, or health related, *cause* mental health related issues. This would allow policymakers to better target the factors that cause mental health issues across the United States.

Appendix

Explanatory Variables

Below are the 60 explanatory variables found in the final, tidy dataset. All variables, except physically_unhealthy_days, are used in the R analysis. Unless otherwise stated, all variables are continuous.

Health Outcomes:

1. physically_unhealthy_days: average number of reported physically unhealthy days per month. (Note: variable is NOT included in predictive modeling).
2. life_expectancy: life expectancy
3. perc_diabetic: percentage of adults with diabetes prevalence (from the CDC Wonder mortality data)

Health Behavior:

4. perc_smokers: percentage of adults that reported currently smoking
5. perc_obese: percentage of adults that report BMI >= 30
6. *perc_physically_inactive*: percentage of adults that report no leisure-time physical activity
7. perc_excessive_drinking: percentage of adults that report excessive or binge drinking
8. chlamydia_rate: Chlamydia cases per 100,000 population
9. teen_birth_rate: births per 1,000 females ages 15-19
10. perc_insufficient_sleep: percent of adults who report fewer than 7 hours of sleep on average (from the Behavioral Risk Factor Surveillance System)
11. perc_food_insecure: food insecurity (from the Map the Meal Gap)

Clinical Care:

12. perc_uninsured: percentage of people under age 65 without insurance
13. *pcp_rate*: Primary Care Physicians per 100,000 population
14. dentist_rate: dentists per 100,000 population
15. mhp_rate: Mental Health Providers per 100,000 population
16. prevent_hosp_rate: discharges for Ambulatory Care Sensitive Conditions per 100,000 Medicare Enrollees
17. perc_mammography_screened: percentage of female Medicare enrollees having an annual mammogram (age 65-74)
18. perc_flu_vaccinated: percentage of annual Medicare enrollees having an annual flu vaccination

Social and Economic Environment:

19. perc_with_access: percentage of the population with access to places for physical activity
20. perc_limited_access: limited access to healthy foods
21. food_environment_index: an indicator of access to healthy foods - 0 is worst, 10 is best
22. graduation_rate: graduation rate
23. perc_some_college: percentage of adults age 25-44 with some post-secondary education
24. perc_unemployed: percentage of population ages 16+ unemployed and looking for work

25. perc_children_in_poverty: percentage of children (under age 18) living in poverty
26. perc_single_parent_households: percentage of children that live in single-parent households
27. association_rate: social associations per 10,000 population
28. violent_crime_rate: violent crimes per 100,000 population
29. injury_death_rate: injury mortality rate per 100,000
30. income_ratio: ratio of household income at the 80th percentile to income at the 20th percentile
31. household_income: median household income (from the Small Area Income and Poverty Estimates)
32. perc_free_or_reduced_lunch: children eligible for free or reduced-price lunch (from the National Center for Education Statistics)
33. perc_homeowners: homeownership (from the American Community Survey)
34. household_has_computer: percent of households that have desktop or laptop computer
35. household_has_broadband: percent of households that have broadband internet access
36. household_has_smartphone: percent of households that have a smartphone
37. housing_mobile_homes: percent of housing units in mobile homes and other types of units
38. per_capita_income: per capita money income in the past 12 months
39. persons_per_household: persons per household

Physical Environment:

40. average_daily_pm2.5: the average daily amount of fine particulate matter in micrograms per cubic meter
41. presence_of_violation: county affected by a drinking water violations (categorical: yes = 1, no = 0)
42. perc_severe_housing_problems: Percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, or lack of kitchen or plumbing facilities
43. perc_severe_housing_cost_burden: percentage of households with high housing cost
44. overcrowding: percentage of households with overcrowding
45. inadequate_facilities: percentage of households with lack of kitchen or plumbing facilities
46. perc_drive_alone: percentage of workers who drive alone to work
47. perc_long_commute_drives_alone*:* among workers who commute in their car alone, the percentage that commute more than 30 minutes
48. perc_rural: percentage of the population within that county that lives in a rural area

Demographic Information

49. perc_under_18: percentage of the population under 18 years of age
50. perc_65_and_over: percentage of the population over 65 years of age
51. perc_african_american: percentage of the population that is Non-Hispanic African American
52. perc_american_indian_alaskan_native: percentage of the population that is American Indian and Alaskan Native
53. perc_asian: percentage of the population that is Asain

54. perc_native_hawaiian_other_pacific_islander: percentage of the population that is Native Hawaiian or other Pacific Islander
55. perc_hispanic: percentage of the population that is Hispanic
56. perc_non_hispanic_white: percentage of the population that is Non-Hispanic white
57. perc_not_proficient_in_english: percentage non-proficient in English
58. perc_female: percentage of the population that is female
59. population*:* population in each county
60. veterans: percent among civilian population 18 and over that are veterans

Summary Statistics for Ordinary Least Squares

Part #1:

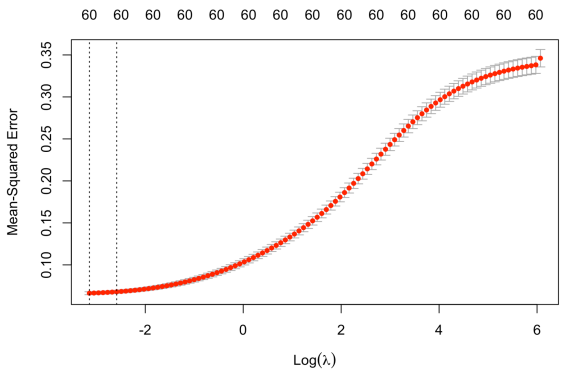
```
Residuals:
     Min       1Q   Median       3Q      Max
-0.98226 -0.17894  0.00691  0.16613  0.85927

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        7.710e+00  1.410e+00   5.466 5.22e-08 ***
perc_smokers                       4.561e-02  3.887e-03  11.734  < 2e-16 ***
perc_obese                        -5.431e-03  2.300e-03  -2.361 0.018325 *
food_environment_index            -1.149e-01  8.797e-02  -1.306 0.191852
perc_physically_inactive          -5.100e-03  2.232e-03  -2.285 0.022436 *
perc_with_access                  -4.258e-04  4.288e-04  -0.993 0.320794
perc_excessive_drinking           -3.031e-02  3.043e-03  -9.961  < 2e-16 ***
chlamydia_rate                    -1.095e-04  5.303e-05  -2.066 0.038993 *
teen_birth_rate                   -1.410e-03  9.834e-04  -1.434 0.151657
perc_uninsured                    -1.594e-02  2.308e-03  -6.909 6.73e-12 ***
pcp_rate                           2.077e-04  2.692e-04   0.771 0.440577
dentist_rate                       2.263e-04  3.519e-04   0.643 0.520136
mhp_rate                           3.601e-04  5.320e-05   6.768 1.75e-11 ***
prevent_hosp_rate                 -1.732e-05  4.431e-06  -3.908 9.63e-05 ***
perc_mammography_screened         -3.335e-03  1.264e-03  -2.638 0.008399 **
perc_flu_vaccinated               -8.399e-04  9.130e-04  -0.920 0.357774
graduation_rate                   -1.706e-03  1.038e-03  -1.644 0.100433
population                        -8.718e-07  7.494e-07  -1.163 0.244858
perc_some_college                 -6.921e-03  1.180e-03  -5.867 5.25e-09 ***
labor_force                        4.532e-07  4.335e-07   1.045 0.295965
perc_unemployed                   -3.014e-02  5.888e-03  -5.120 3.38e-07 ***
perc_children_in_poverty           9.576e-03  2.122e-03   4.513 6.80e-06 ***
perc_single_parent_households     -7.751e-04  1.219e-03  -0.636 0.525027
association_rate                  -1.113e-02  1.528e-03  -7.284 4.78e-13 ***
violent_crime_rate                 2.401e-04  4.309e-05   5.572 2.89e-08 ***
injury_death_rate                  6.722e-04  4.685e-04   1.435 0.151497
average_daily_pm2.5                1.248e-02  4.762e-03   2.620 0.008863 **
presence_of_violation              2.094e-02  1.244e-02   1.684 0.092422 .
perc_severe_housing_problems       2.773e-03  4.581e-03   0.605 0.545081
overcrowding                       9.916e-03  6.875e-03   1.442 0.149370
inadequate_facilities             -1.152e-02  7.827e-03  -1.471 0.141359
perc_drive_alone                  -2.002e-03  1.505e-03  -1.331 0.183515
perc_long_commute_drives_alone    -2.880e-04  8.394e-04  -0.343 0.731579
income_ratio                      -5.545e-04  1.424e-02  -0.039 0.968939
life_expectancy                    2.596e-03  5.694e-03   0.456 0.648487
```

Part #2:

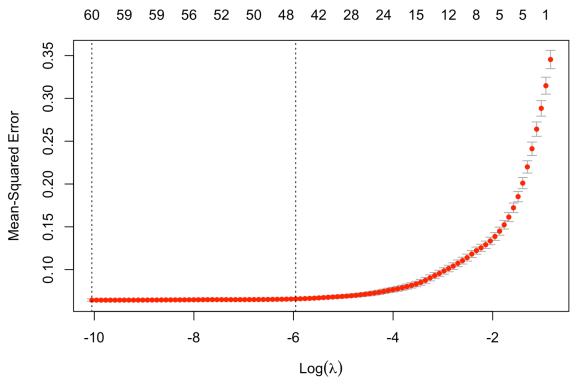
```
perc_diabetic                                4.193e-02  5.204e-03   8.059 1.38e-15 ***
perc_food_insecure                           7.806e-03  1.591e-02   0.491 0.623738
perc_limited_access                         -1.190e-02  8.057e-03  -1.477 0.139823
perc_insufficient_sleep                      3.220e-02  2.927e-03  11.000  < 2e-16 ***
household_income                            -7.294e-06  1.723e-06  -4.234 2.40e-05 ***
perc_free_or_reduced_lunch                  -4.308e-03  6.614e-04  -6.513 9.45e-11 ***
perc_homeowners                             -1.045e-02  3.397e-03  -3.075 0.002139 **
perc_severe_housing_cost_burden             -3.822e-03  4.822e-03  -0.793 0.427998
perc_under_18                               -3.592e-02  4.182e-03  -8.590  < 2e-16 ***
perc_65_and_over                            -1.158e-02  3.937e-03  -2.940 0.003318 **
perc_african_american                       -4.943e-02  7.534e-03  -6.560 6.97e-11 ***
perc_american_indian_alaskan_native         -3.271e-02  7.915e-03  -4.133 3.75e-05 ***
perc_asian                                  -5.606e-02  8.815e-03  -6.360 2.55e-10 ***
perc_native_hawaiian_other_pacific_islander  8.134e-02  3.628e-02   2.242 0.025098 *
perc_hispanic                               -3.979e-02  7.336e-03  -5.423 6.62e-08 ***
perc_non_hispanic_white                     -4.038e-02  7.633e-03  -5.290 1.37e-07 ***
perc_not_proficient_in_english               1.526e-02  5.346e-03   2.853 0.004374 **
perc_female                                  5.064e-02  4.705e-03  10.763  < 2e-16 ***
perc_rural                                   3.468e-04  4.173e-04   0.831 0.406076
household_has_broadband                      6.814e-04  1.945e-03   0.350 0.726134
household_has_computer                      -2.347e-03  2.161e-03  -1.086 0.277651
household_has_smartphone                     5.262e-03  1.807e-03   2.913 0.003626 **
housing_mobile_homes                        -1.465e-02  3.370e-03  -4.347 1.46e-05 ***
per_capita_income                           -4.106e-07  3.060e-06  -0.134 0.893270
persons_per_household                        1.758e-01  5.146e-02   3.416 0.000649 ***
veterans                                    -2.898e-03  3.349e-03  -0.865 0.387075
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2484 on 1836 degrees of freedom
Multiple R-squared:  0.8274,    Adjusted R-squared:  0.8218
F-statistic: 146.7 on 60 and 1836 DF,  p-value: < 2.2e-16
```

Shrinkage and Selection Methods: Cross-Validation Plots

Ridge Regression:



Lasso Regression:



Shrinkage and Selection Methods: Top 10 Coefficients

Ridge Regression:

Feature	Coefficient
perc_smokers	0.1272467
perc_insufficient_sleep	0.1103368
perc_excessive_drinking	-0.0880900
perc_diabetic	0.0830614
perc_african_american	-0.0759787
perc_under_18	-0.0726349
perc_female	0.0680099
association_rate	-0.0676078
household_income	-0.0580969
mhp_rate	0.0545223

Lasso Regression:

Feature	Coefficient
perc_smokers	0.1585937
perc_insufficient_sleep	0.1262111
perc_african_american	-0.1236434
perc_excessive_drinking	-0.1041143
perc_under_18	-0.0985940
perc_food_insecure	0.0957729
perc_diabetic	0.0945663
household_income	-0.0940767
perc_female	0.0825900
perc_free_or_reduced_lunch	-0.0686537