Motivation: (What is the problem I’m trying to solve?)

Healthcare is a much debated topic today in the United States, and with government mandated healthcare through the Affordable Care Act, there is a lot of talk about what kind of coverage is actually available for low-income individuals who are required to sign up for catastrophic insurance plans. These insurance plans rarely cover any kind of necessary non-emergency treatment and have impossibly high deductibles for low-income adults to meet, leaving them with a reality that is in stark contrast to receiving “Affordable” healthcare. This brings me to the goal of my project. There are clear gaps in the United States healthcare system between patient needs and healthcare services, one of which is the absence of mental health treatment in many insurance plans and a lack of awareness in general of mental illness and what it entails. Through my analysis of the 2014 National Survey on Drug Use and Health, I hope to call attention to the unmet needs of the subset of the U.S. population who had experienced serious psychological distress within the past year but have not received any form of mental healthcare treatment. The primary method I will use is the construction of a logistic regression model that will predict whether or not someone within this subset is likely to receive treatment for their mental health.

Impact: (Who is my client? Why do they care?)

This project has the potential to impact anyone who is concerned with mental health disorders or mental illness in their family or their community. In order to disseminate the results of my analysis, I may present my findings to local hospitals, nonprofit organizations, or the university healthcare system that have the power to reach out to adults who are disproportionately affected by this gap in our healthcare system. The potential clients within my reach are in and around the area of Athens, GA and the University of Georgia, but the results of my survey analysis can be applied to the population of the United States, for which the National Survey on Drug Use and Health was designed to represent. Therefore, my logistic regression models could be useful to individuals around the country who access my work publicly on my GitHub repository (https://github.com/codyw826/Public-Projects), not to mention they could be useful for the survey makers themselves at SAMHSA as this is an annual survey.

Dataset: (Survey Design, Limitations, and Missing Data)

Conducted by the Substance Abuse and Mental Health Services Administration (SAMHSA), I am using the annual National Survey on Drug Use and Health (NSDUH), which is the primary source of information on the prevalence patterns and consequences of drug use and abuse and mental disorders in the U.S. This survey contains thousands of variables, most carried over from year to year. Most of the variables are also binary or categorical responses to multiple choice questions about drug use or mental health or just for acquiring demographic information. Being a nationwide survey, there are samples of around 1000 individuals or more for each state, weighted to gather larger samples from states with higher populations in an attempt to make the survey sample representative of the U.S. population. In particular, the 2014 iteration of the National Survey on Drug Use and Health that I used had a sample size of 55,271 individuals, although the subset used for my regression was only about a tenth of that.

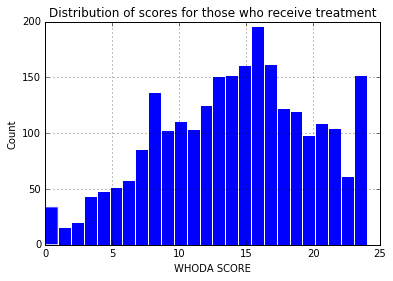
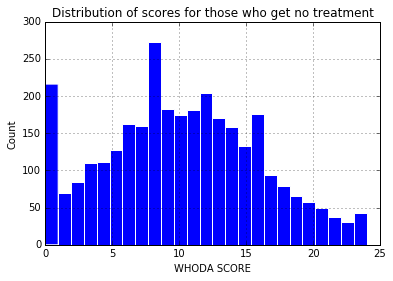
The target population of the survey is defined as the civilian, noninstitutionalized population of the United States who are 12 years old or older, meaning a small percentage (roughly 3%) of adults in the population are excluded. Due to the emphasis of my analysis being placed on mental health, it is worth mentioning that these excluded adults are members of the active duty military and individuals in institutional group quarters (e.g. hospitals, prisons, nursing homes, treatment centers), whom I would personally expect to have a different distribution of mental health than the population measured in this survey. In addition to this, it should also be kept in mind that the survey presently collects data through face-to-face interviews with a representative sample of the population at the respondent's place of residence. This means that homeless individuals not living in a shelter are also excluded from the population of this survey; another group that likely has a different distribution of mental health from that of the U.S. population represented in this survey. These limitations should be kept in mind when interpreting my analysis results and my models should not be applied to individuals who fall in the excluded subsets of the NSDUH.

In order to address my primary goal of predicting whether or not someone experiencing serious psychological distress in the past year will receive mental health treatment, I looked at a number of mental health diagnostic variables, mental health treatment variables, and demographics in an attempt to identify any relevant differences between those who seek out treatment and those who don't. Most if not all of these variables corresponded to survey questions which had a certain number of individuals who either skipped the question because it did not apply to them or they were not relevant to the question for one reason or another. For example, some of the variables I used in my predictive model were from the Adult Mental Health sections of the survey, which excluded respondents aged 12-17. These missing values in the dataset were re-coded at some point during the creation of the survey to a number that did not make sense in the context of the other responses. In order to avoid the problems that missing data causes, I replaced all of the recoded values of -9 and 99 with N/A in the dataset. I also noticed that there were some binary variables coded with a 0-1 response, yet others coded with a 1-2 response. In order to increase the re-usability of my code and make the regression fitting process more intuitive, I changed all binary response variables of interest to be coded 0-1. With these changes, the data I used for my analysis had been coerced to a more useful form, but I later faced another issue with the presence of N/A values. Although histograms could be created while ignoring N/A values in the data, I couldn’t fit a regression model to data that included these values. Therefore, the last change I made before fitting my models was to drop roughly 50 or so rows (individuals) which contained a N/A value.

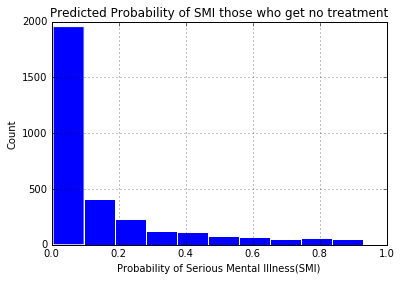
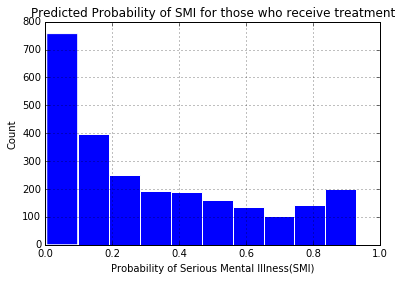
Exploration: (Variable Selection, Histograms)

After spending a long time reading through the survey questions, particularly the Mental Health, Adult Depression, Adult Mental Health Treatment, and Demographics sections of the survey, I had compiled a list of about 30 variables to investigate in relation to adult mental health. Knowing that I wanted to call attention to the difficulty that individuals with mental illness have with getting proper treatment, I first identified two key variables I would use throughout my analysis: SPDYR and AMHTXRC3. SPDYR is a binary variable that is coded 1 for individuals who experienced serious psychological distress within the past year and 0 otherwise. This was evaluated based on a series of 6 questions included in the survey (K6 questionnaire)that measure the degree to which an individual’s mental instability or anxiety affect their daily lives that are summed to measure between 0 and 24. Individuals who score 13 or higher on this K6 questionnaire for the past year were coded 1 for SPDYR. I am looking specifically at the subset of 5,696 individuals who are coded 1 for this variable in making my regression models. The other key variable, AMHTXRC3 is a binary variable that is coded 1 for adults who received any form of mental health treatment in the past year and 0 otherwise. This includes inpatient and outpatient treatment as well as prescription medicine as the forms of treatment received. Within the group of individuals who are coded SPDYR = 1, my logistic regression will predict the probability of an individual receiving mental health treatment in the past year (AMHTXRC3 = 1). The rest of the variables I selected were the list of possible predictor variables that I would come to use for my regression model.

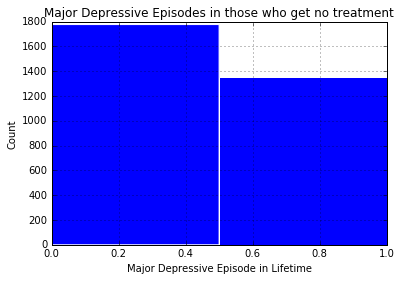
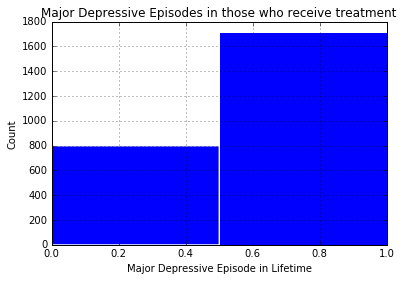
The selection process for useful predictor variables at first consisted of creating two histograms for each variable of interest. Within the subsample of 5,696 psychologically distressed adults that I was now focusing on, I compared the distributions of responses for each variable between the groups of adults who had received treatment in the past year (AMHTXRC3=1), and those who had not (AMHTXRC3=0). In some cases, it also seemed appropriate to compare the distribution of my subsample to that of the whole survey sample to see if the subsample was similarly distributed. Given that almost all of my predictor variables were either binary or categorical, histograms were the quickest and most immediately useful way of assessing relationships between predictors and my response variable. Variables that did not have a distinct difference in distributions between receiving treatment and not receiving treatment were quickly eliminated, but a number of variables had a statistically significant difference in proportions between these groups.



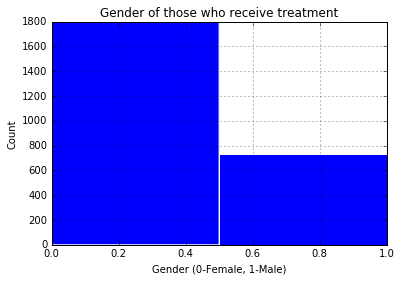
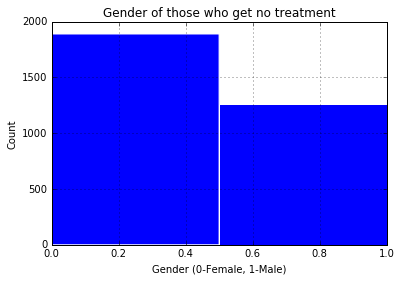
* People who receive mental health treatment tend to score higher on the WHODAS, which is a measure similar to the K6 diagnostic questionnaire, which was also measured on a scale from 0-24. This is presumably because their emotions, nerves, or mental health tend to have a more severe effect on their daily lives compared to those who don’t receive treatment. It makes sense that this similarly scaled variable from the portion of the survey assessing Mental Health would be useful for identifying those with the most severe symptoms, who are most likely to receive treatment.



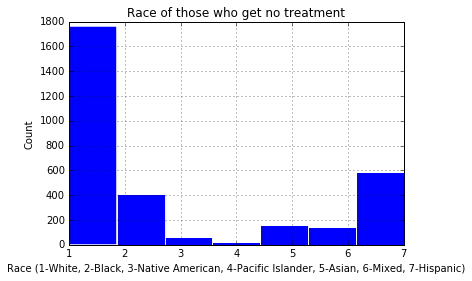
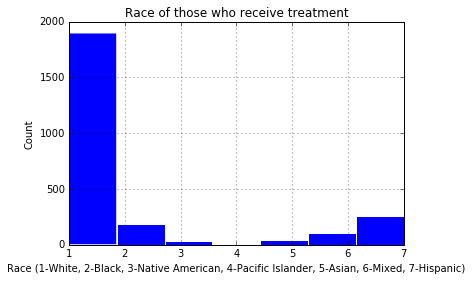
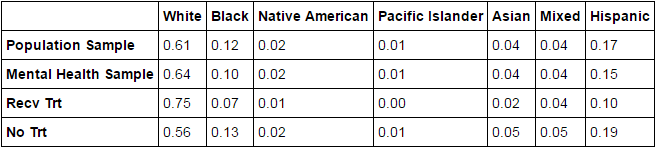
* In the last few years that the NSDUH has been given, survey makers have begun to include the variable SMIPP\_U, which is an individual’s predicted probability of serious mental illness. I was not able to determine what formula is used to calculate this probability so it may not be a useful variable outside of the context of this survey. With that being said, this variable is one of the best predictors for an adult receiving mental health treatment. Naturally, those who receive treatment for mental illness were more likely to have a high predicted probability of serious mental illness.

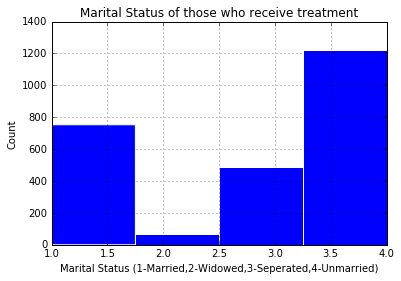


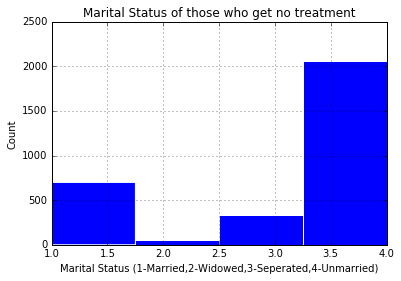
* AMDELT is binary variable coded 1 for an adult who has experienced a major depressive episode in their lifetime and 0 if not. Although it makes sense that people currently experiencing high levels of psychological or emotional impairment are likely to seek treatment, there are also people who have experienced severe trauma only at certain stages in their lives that wish to seek therapy or treatment retroactively. Of those experiencing serious psychological distress in the past year, 68% of those who receive mental health treatment have experienced a major depressive episode in their lifetime as opposed to only 43% of those who have not received treatment, which is a significant difference (z=18.9, p-value<.001).



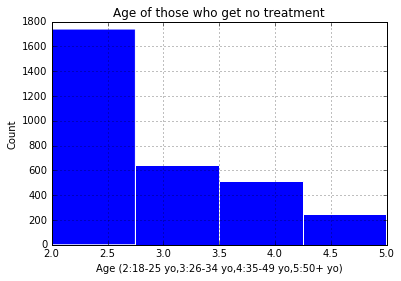
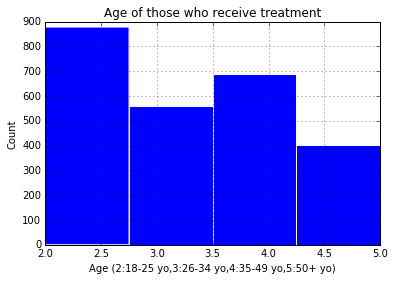
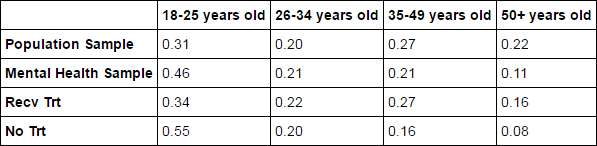
* So people are more likely to seek mental health treatment if they experienced some traumatic events in their lifetime, which seems fairly intuitive. With that in mind, when I took a look at the genders of adults in my sub-sample, I was shocked to see that 65% of those who experienced serious psychological distress in the past year were female. Among these individuals, 71% of those who received treatment for mental health in the past year were female while only 60% of those who get no treatment for mental health were female. A quick two-proportion z-test (z=8.63, p-value< .001) shows that this is a significant difference. My initial thought was that women must be experiencing more traumatic events or consistent psychological distress than men, which may be true. However, the difference in gender proportions seems to indicate that adult men who are experiencing serious psychological distress are significantly less likely to receive mental health treatment than their female counterparts. I can only speculate that this perhaps means men are more likely to under-report mental health symptoms or “tough it out” than women, for whom it is more socially acceptable in the U.S. to show emotions or vulnerability.



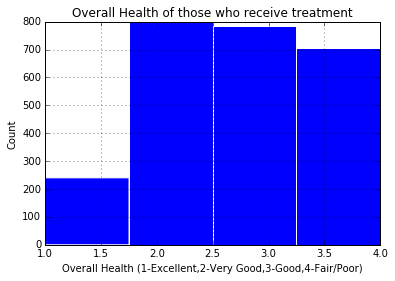
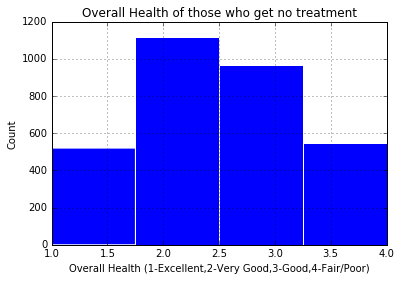
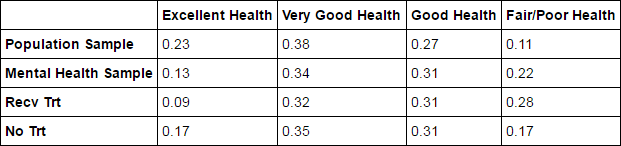
* After finding a gender gap between my groups, it seemed like a good bet that race would produce a similar difference between groups. When looking at the racial makeup of the entire survey population compared to the sample of adults experiencing serious psychological distress, there doesn’t seem to be a significant difference; which is a good sign for the legitimacy of my sample and indicates fairly consistent prevalence of mental health disorders between races as we should expect. However, when comparing the racial makeup between those who receive mental health treatment and those who don’t, there is a substantial difference in proportions. It seems to be the case that a much higher proportion of those who receive mental health treatment are White while a much higher proportion of those who receive no treatment are Black, Hispanic, or Asian. It is difficult to say why this is the case, but it is clear that a racial divide exists in access to treatment. This data speaks to a larger trend in the U.S. Healthcare system that Whites have better access to all forms of treatment than non-Whites.



* When looking through demographics to find possible predictors for my model, I discovered that adults who receive treatment for mental health are more likely to be widows, separated, or married. This means that a large proportion of those who do not get treatment for mental health are unmarried. It seemed a little unusual that all but the unmarried proportions were higher in the receive treatment group, which lead to me think that the age distributions of my survey sample may explain this difference, with younger adults less likely to receive treatment.



* 46% of the adults experiencing serious psychological distress in the past year are between the ages of 18 and 25, which is significantly higher than the 31% of the total survey sample they made up. However, there seems to be a distinct difference between the distributions of ages for those who receive treatment for psychological distress and those who don’t. Those who receive treatment have closer to a representative distribution of ages, while those who don’t get treatment are mostly between the ages of 18 and 25. This could explain the distribution of marital statuses shown before and raises the question: Why are young people less likely to have their mental health problems addressed with treatment? It could have to do with the stigma associated with mental health, high stress that comes with school and work transitionary periods, or simply just a lack of individual resources for affording treatment.



* The last variable I looked at that seemed to have a distinct difference in proportions between groups was the overall physical health of those experiencing serious psychological distress. Compared to the total survey sample, the overall physical health of those experiencing severe psychological distress was typically worse than average. Yet between those who did and did not receive treatment for their mental health, there was still a significant difference in overall physical health. While I was able to understand why a higher proportion of adults who had poor physical health would have received mental health treatment, it surprised me to see that the proportion of adults with excellent health was higher for those who did not get treatment. This could possibly mean that adults with excellent overall physical health are less likely to seek legitimate treatment for mental illness because they are otherwise in good health.

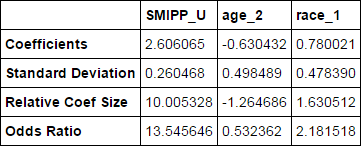
Logistic Regression: (Full Model, Reduced Models)

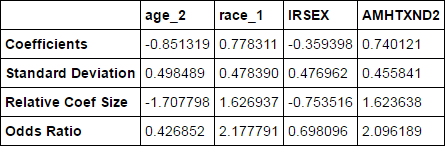
After I finished analyzing my list of predictor variables, I had narrowed down which variables indicated a significant difference between the two groups of my response variable. As I mentioned earlier, before I could fit a regression model I needed to drop all individuals in my subsample with any N/A values, reducing my subsample size from 5696 to 5634 adults. In order to judge the accuracy of my models, every time I fit a new model I used 5-fold cross-validation to separate my data into a training set(80% of subsample) and test set(20% of subsample). The training data is used for fitting the model, and then the model is used to classify the test data, giving me an accuracy score: the percentage of the test data that is correctly classified. The first logistic regression model I created was a full model including all of my significant predictor variables, with categorical variables split into a binary dummy variable for each category. This model contained 17 predictors and achieved an average accuracy score of 70% correct classification with repeated cross-validation of the data. All of the terms in this model were significant, but a model with this many terms is definitely overfitting the sample data, and is unlikely to generalize well to new data. On top of this, it didn’t take long to realize that a similar accuracy score could be replicated with far fewer terms in the model. In the final report, the output and coefficients for this full model will be included in the appendix.

In order to create an effective reduced model, I decided to restart the model building process from the ground up. All of my model terms had given significant p-values, and all together there were three binary variables (AMHTXND2, AMDELT, IRSEX), two continuous variables (WHODASC2, SMIPP\_U), and three categorical variables split(AGE, RACE, HEALTH) into a total of 15 dummy variables. The only variable that was significant in the model and not shown in the previous sections is AMHTXND2, which is coded 1 for individuals who have a self-perceived need for mental health treatment and 0 otherwise. I created a one term logistic regression model using each of these twenty variables to assess their individual predictive power as the first step in building a minimal reduced model (There were 17 terms instead of 20 terms in the full model because all dummy variables for categorical variables cannot be included in a regression model.). This iterative process resulted in my determination that the most powerful predictive variables for receiving mental health treatment for psychological distress are: SMIPP\_U, WHODASC2, AMDELT, AMHTXND2, age\_2 (dummy var. ages 18-25), and race\_1 (dummy var. race=White), in that order.

After looking at the coefficients for the race and age dummy variables in the full model, I noticed that all of the age terms except for age\_2 had a positive effect on the probability of receiving treatment and that all of the race terms except for race\_1 had a negative effect on the probability of receiving treatment. In addition, these two variables were the most powerful predictive dummy variables, so it made sense to include them in my reduced model over the other dummy variables that were used in the full model. The prediction accuracy of the one term models for each of the five variables listed above ranged between 58 and 65 percent accuracy, which is already approaching the 70 percent accuracy given by the full model. The best combination of these variables perhaps along with the inclusion of a term for gender or health should be able to replicate the full model’s accuracy.

With my best predictors identified, I went about testing the best combinations of two variables on model accuracy. Most of these models were ranging between 63-68 percent accuracy, though the best models all included the variable SMIPP\_U. I could not replicate the full model’s accuracy with two terms, but with the addition of a third variable I was able to reach 70 percent using the model predicting whether or not a psychologically distressed adult receives mental health treatment based on SMIPP\_U, age\_2, and race\_1. This model is practically speaking just as useful as the full model for classification.

This is the best reduced model that I could achieve with my selected variables, though it would only be useful within the context of the survey when the value of SMIPP\_U is calculated. Logistic regression models are best interpreted using odds ratios, as shown in the table above, for understanding what these variables are actually predicting. According to this reduced model, a psychologically distressed individual who is within ages 18-25 is about half as likely to receive mental health treatment for their psychological distress as someone older. In other words, 1.88(1/.53236 = 1.88) people who are older than 25 receive needed mental health treatment for each similarly distressed 18-25 year old who receives treatment. As for interpreting the race term, an adult who is white is more than twice as likely to receive treatment as an adult of another race. This means that 2.18 white adults receive needed treatment for mental health disorders for each 1 person of another race that similarly needs treatment. SMIPP\_U is a little more difficult to interpret, but for each .10 higher that SMIPP\_U is (10% more likely to have serious mental illness), that person's odds of receiving treatment increase by 1.35 to 1. Therefore, 13.5 adults with a 100% predicted probability of serious mental illness receive mental health treatment for each person with a 0% predicted probability of mental illness that receives treatment. Unfortunately, SMIPP\_U is a significantly better predictor than many other terms, but is not a practical value that can be understood and applied outside of the annual NSDUH. However, the interpretation of the race and age variables still have practical meaning in this model and can be included in a reduced model that exclusively uses terms that can be relatively easily determined.

With the goal of a practical reduced model in mind, I excluded SMIPP\_U, WHODASC2, and AMDELT from my model experimentation and focused on finding the optimal combination of my remaining model terms. It seemed to me that any model with more than three or four terms would likely fall victim to overfitting, so I limited my possible model size to four terms and found the best combination of predictor variables within that scope. The model with the best accuracy that fulfilled these conditions turned out to be a model predicting whether or not an adult experiencing psychological distress would receive needed treatment based on age\_2, race\_1, gender, and perceived need for treatment.

In my final report, I will analyze these odds ratios in context, comment on the accuracy score of this model in relation to the previous models, and summarize my findings.