DSA 5013

ANALYZING DRUG ABUSE IN AMERICA FROM 2012-2015

# Alexandra Beattie

# Executive Summary

## Problem Statement

Currently, the United States is experiencing a drug epidemic with rising drug-related death rates. Within the last two years, drug overdoses have become the leading cause of adult deaths in America. For this exercise, I will be developing an extensive data set by pooling together resources from the CDC, American Census, and Substance Abuse and Mental Health Data Archive. I will then use this data to observe what factors contribute to different states developing higher drug-related death rates.

## Major Concerns

My primary concerns are data availability, data cleanliness, and validation from outside sources. Since the data I am collecting comes from different sources and is obtained by the government through surveys, it is highly unlikely that the sources will fit nicely together. The majority of my work will be to clean the data and create a unified data set based on state. I am also concerned with the fact that surveys change from year to year, this makes it difficult to compare survey answers that may not have existed the year before or after. Lastly, there has not been a great deal of publicity dedicated to this type of research (combining factors to observe the landscape). I will need to do further research on available studies to validate my findings with others.

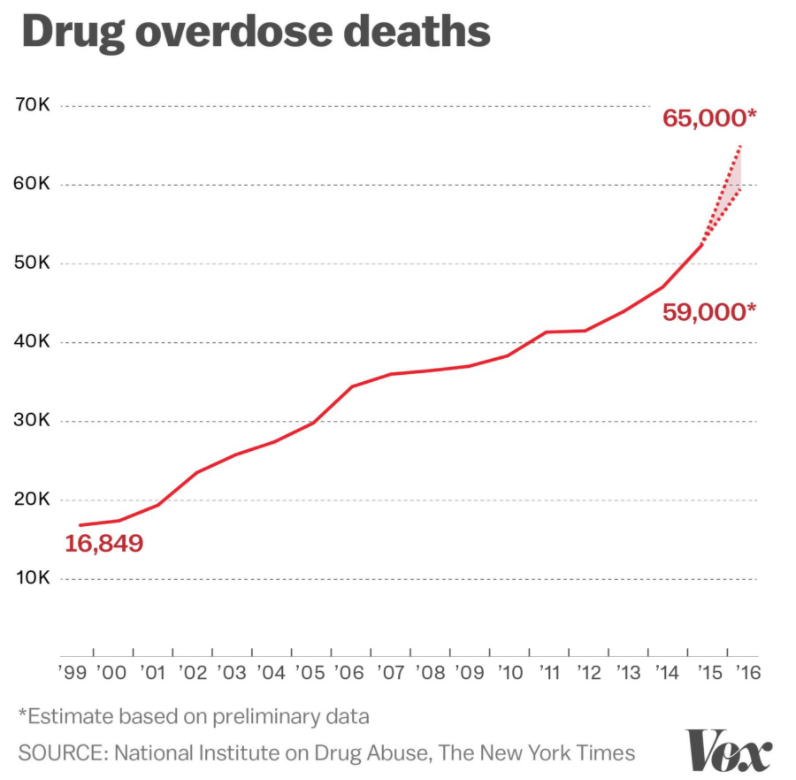
## Summary & Recommendation

My analysis proved that there are differences between communities with high and low drug-related death rates. It also proved that many communities experiencing high drug-related death rates are underfunded by the government and have little access to help focused on long-term recovery. These metrics tend to stay the same for communities with high rates of drug use. This causes certain states to worsen from year to year continuously.

I would recommend that we take action to invest in the macroeconomic landscapes of these states to try and attack the problem at its' core. These actions could involve investing in the state’s education, access to government funding, and providing proper health care or long-term solution plans for drug addicts.

# Problem Statement & Background

Currently, states across America are experiencing increasing rates of deaths due to overdose. Between 2015 and 2016, we saw an increase in drug overdose deaths by 19% since 2015.



This has resulted in drug overdoses to be the leading cause for American adults. As these staggering statistics continue to rise, the issue of why this is happening remains unanswered. Political figureheads speak of solving the opioid crisis without follow up on how or why specific communities are battling this epidemic.

My goal for this project is to develop a dataset from multiple sources to create a robust and engineered set for exploratory analysis as well as modeling for feature importance. This dataset will be used to identify similarities in communities that experience differing levels of drug overdoses within their population. After identifying these trends, I will produce a list of recommendations based on these similarities. In this analysis, I will provide the following:

* Low vs. High Death Rate States – Exploratory & Outlier Analysis
* Multiple Linear Regression Analysis
* Exploration of Decision Trees for Variable Importance
* Random Forest Model Analysis

# Data Creation & Engineering

One of the most significant hurdles for this project was attaining and forming the dataset. SAMHSA (Substance Abuse and Mental Health Services Administration) houses the majority of data concerning drug trends in America. From their datasets, I chose to focus on their substance abuse facilities and client level datasets.

The substance abuse facilities dataset (NSSATS) is developed from a national survey of substance abuse facilities across the United States. This survey captures the types of facilities and programs available to the community. The client level table (TEDS) has two different table options, admissions, and discharges. From this table, I used data about the number of hospitalizations and releases and the type of problem being treated. Variables vary from year to year as these surveys change. For this exercise, I decided only to use shared variables across the yearly tables.

Next, I obtained census data to describe the macroeconomic situation of each state. To analyze this data, I used the acs package in R. I was able to use tables from 2012 to 2015 with a personal key from the American Census Survey. This process created some unwanted limitations to the amount of data that I was able to use for this analysis.

Lastly, I used data from the CDC for deaths each year partitioned by state due to a drug overdose. This variable is my target variable for the exercise.

My final table was formed by creating aggregates by state. Each state has a row for the years 2012, 2013 and 2014. The table included 152 rows (including Puerto Rico for 2013 and 2014) with 143 variables. Since the majority of growth in drug overdoses has happened in the last five years, I believe this data will be able to shed some light on the current state of the drug epidemic in the United States. Unfortunately, the most prominent growth in overdoses was from 2015 to 2016. Currently, the data for 2016 is not available across all three of these data sites.

The features were engineered to reflect the base variable for each data table. Census data was based on the population (ex. government assistance per person). Admission data was based on the number of admits (ex, the percentage of admits concerning a heroin problem). Services data was based on the number of substance abuse facilities reported (ex. portion of facilities that offer Vivitrol).

In regards to missingness, in these surveys, a null value was representative of the question not about the facility or person (either unanswered or skipped). Given the dataset being aggregated, we could safely treat these missing values as zero. There was also the issue of surveys changing from year to year. In the interest of time, I chose only to use questions that occurred each year. This action affected my data significantly because 2015 data was utterly different from the years before. Due to this, I chose to analyze this dataset on its' own. It proved to have similar results as the data from years before so this paper will focus on 2012-2014. If I were to expand on this project, I would dedicate more time to engineering 2015 to work with the prior years.

For this exercise, I used four uniquely created tables:

* 2012-2015 data from the CDC, Census, and TEDS Admissions
* 2012-2014 data from the CDC, Census, TEDS Admissions, and NSSATS
* 2012-2013 data from the CDC, Census, and TEDS Discharges
* 2015 data from NSSATS

# Exploratory Analysis

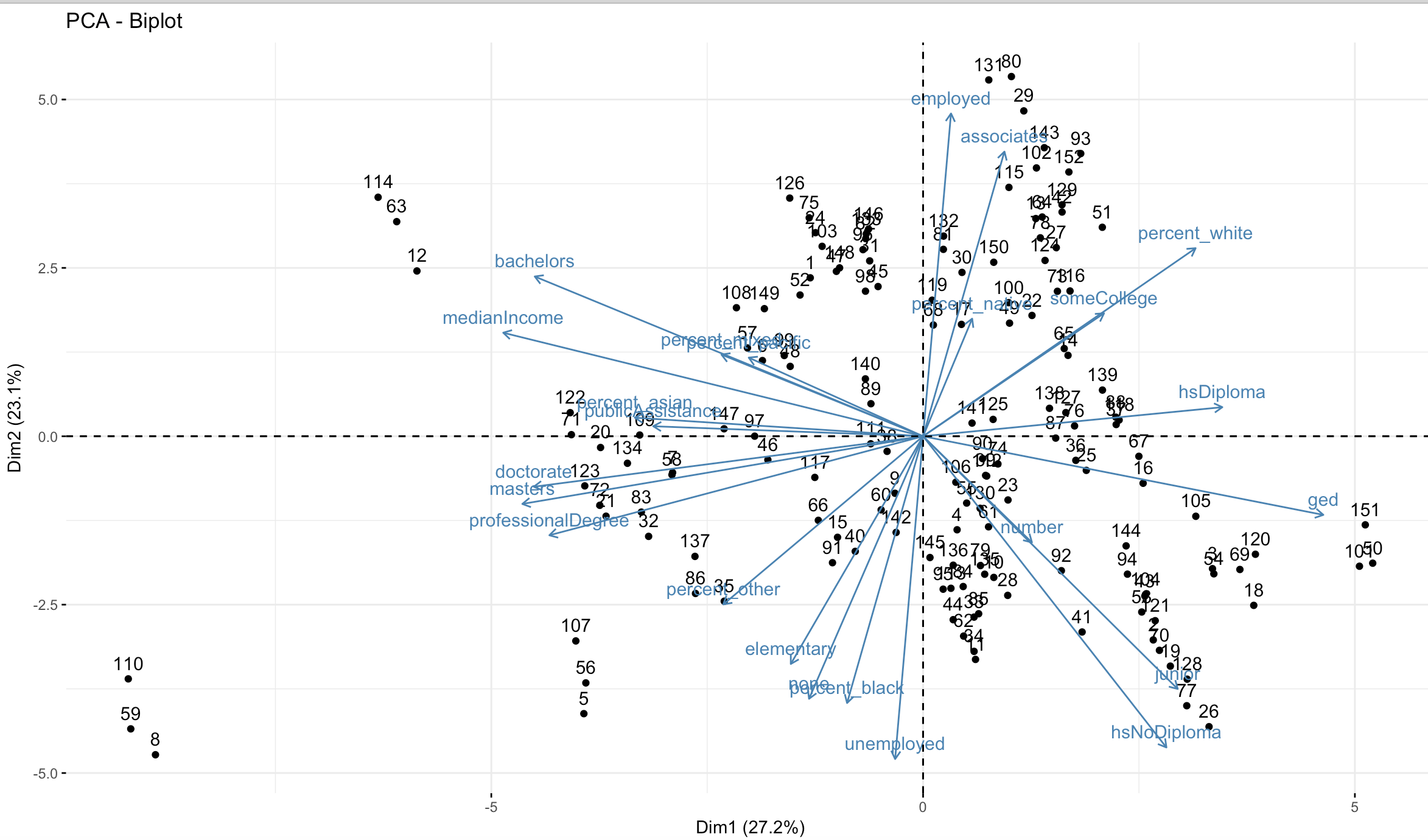
To better understand the data, I took the following steps:

* PCA Analysis of Macroeconomic factors
* Comparison Analysis between Low and High Death Rate States by Variable

In this report, I will provide the most important details of the analysis. Any further graphs or charts detailing the data will be found in the appendix. Please note that the split between low and high death rate states will be based on standardized data concerning deaths per population. Scaled death rates are as follows: low (-1.5>), medium (-1.5 < x < 1.5), and high (1.5<).

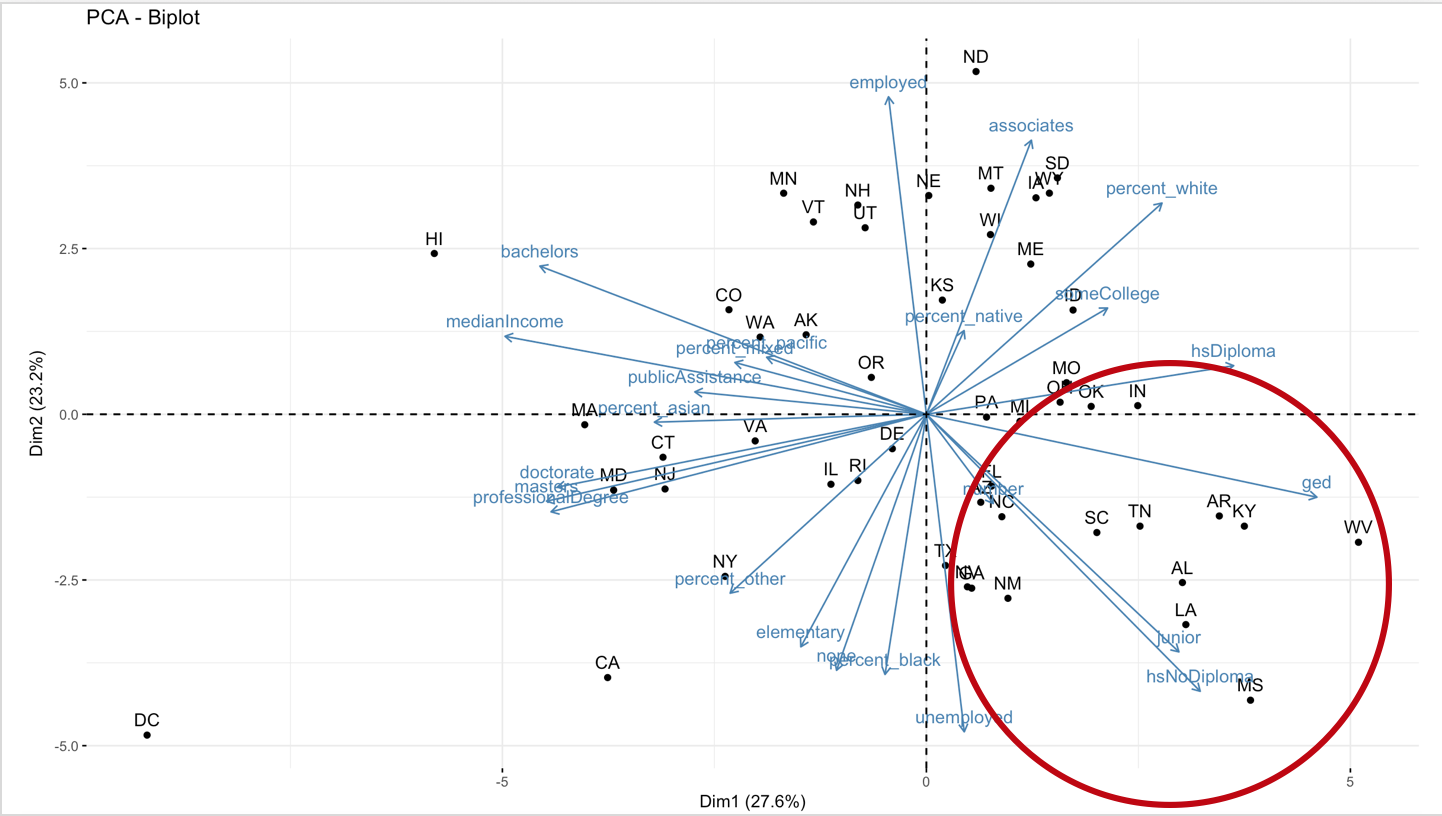
## PCA Analysis of Macroeconomic Factors

To thoroughly understand the landscape of the data, I found it to be essential to analyze the data that all states had available for each year from the census. To do so, I looked at a PCA Analysis of the data from 2012 through 2014. This first round of analysis produced the following:



As you can see, from the two principal components, a higher number of deaths is related to a population that has a lower education. We recognize that the direction of fatalities (labeled as a number) is directly in line with the variables for "junior" and "hsNoDiploma". This finding can be interpreted as states that have higher proportions of their population that do not officially finish high school. We can also observe that the direction in deaths is close to the direction of unemployment and is opposite of higher levels of median income, public assistance, and more diverse populations. This could give an interesting interpretation that a less diverse, more impoverished, uneducated population with less support from local or federal government is more likely to experience drug-related deaths.

In this biplot for 2015 data, we can also see that states which are experiencing a higher rate of drug-related deaths tend to be southern states:



We can also see that the 2015 vs. the three-year biplot does not change dramatically. The lack of change in the macroeconomic environment of the communities could be contributing to why individual communities continue to be top states for drug-related deaths, such as West Virginia or Florida.

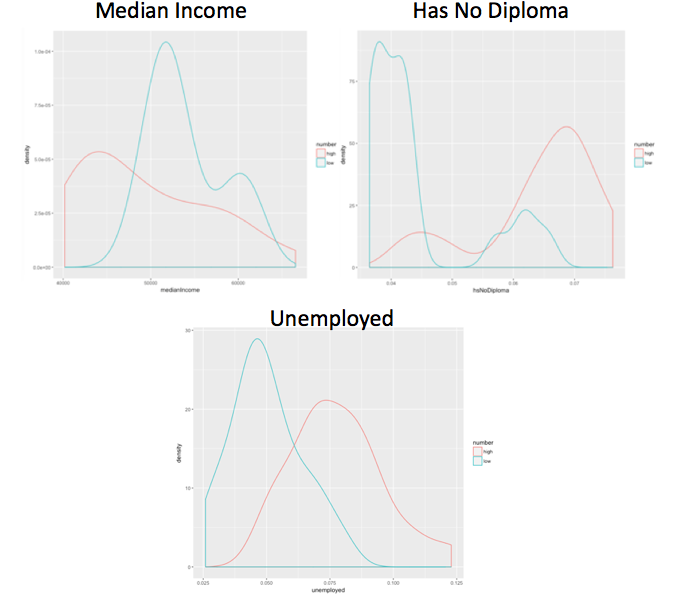
Taking the PCA Analysis a step further, I added in admission statistics into the components. This addition created similar results but also shed interesting light on trends in drug use. While communities were shown to be related to uneducated populations, I found that more educated and diverse communities tended to have more casual drug use reported as well as higher reports of cocaine and heroin. They also frequently had more public assistance resources dedicated to their communities.

In conclusion, the PCA analysis shows that higher drug-related death rates tend to cluster with populations that are poorer, less diverse and have less public assistance dollars allotted per person.

## Comparison Between States with Low and High Death Rates

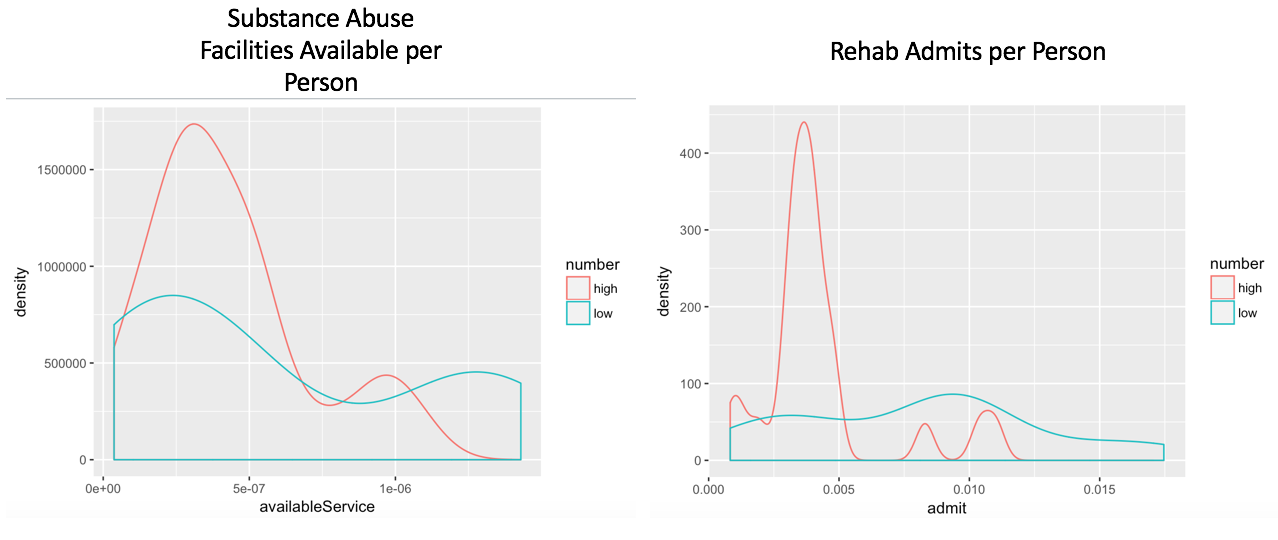
## Census 2012-2015

To confirm my findings from the PCA Analysis, I decided to conduct a cross comparison between states with low and high rates of drug-related deaths.



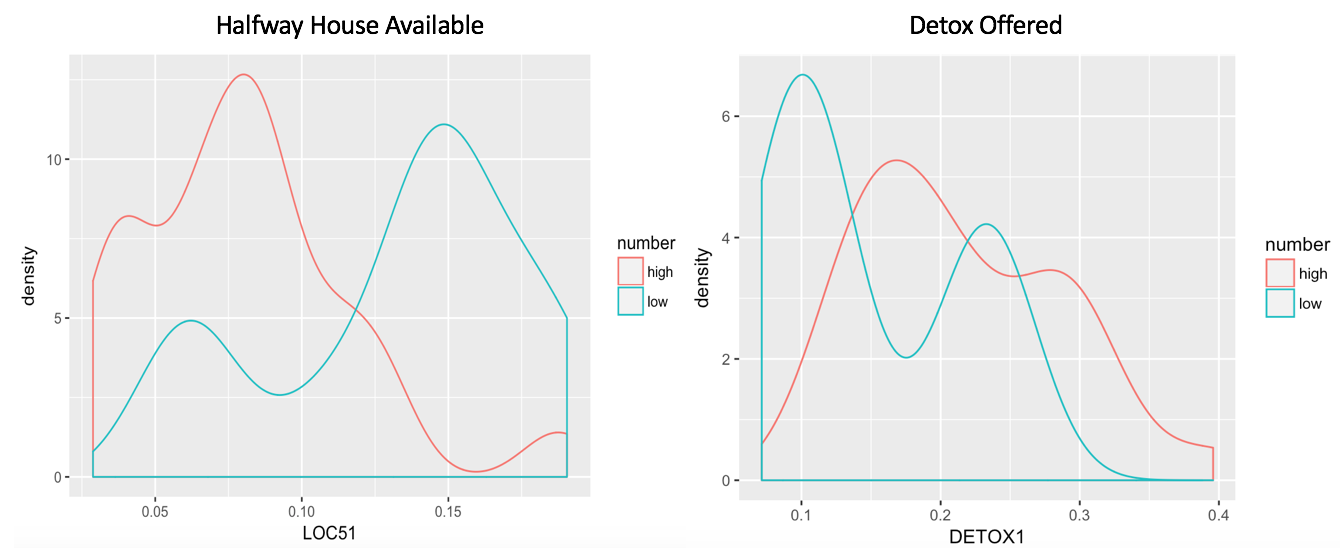
As expected, we can see a significant discrepancy between high and low rate states. The communities experiencing this drug epidemic are communities that are less developed. This finding could potentially be a flag that these communities need more help from the government to invest in education and the creation of jobs.

### Substance abuse services



As shown above, we can see that even though some communities have higher rates of overdose-related deaths, these same populations have lower numbers of available services and lower rehab admission rates. This is opposite of what I expected to see for high rate states. A statistic like this could show why those communities experience a higher rate of deaths. If people are unable to access help, it will be more difficult for them to combat their drug addiction.

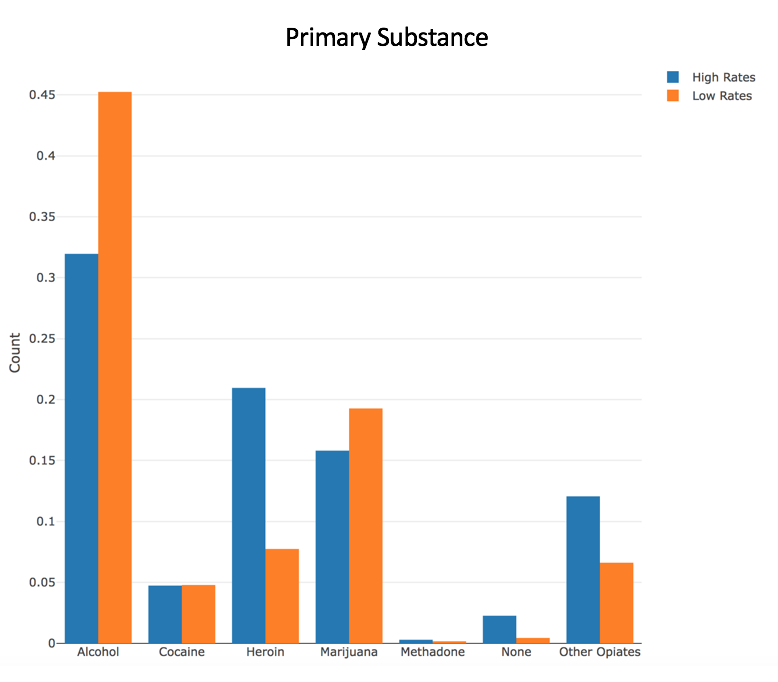
After seeing this statistic, I decided to investigate if there are discrepancies in the types of services offered within the community. I found that the care available for lower rate states varied with the higher rate states regarding the quality and length of care.



Higher rate states had more instances of detox or opioid treatments being available, but they had less availability of halfway houses or intensive outpatient treatment. This type of long-term treatment is critical to the success of an addict. If it is not available, it is challenging for an addict to transition from a controlled environment to real life. This lack of availability could cause more relapses and perhaps end in death.

### Types of Substances Abused

The substances of choice for higher risk communities that stood out were heroin and other opiates. This discrepancy between high and low-risk populations is reflective of the current opioid epidemic in these states. We can also link this back to the higher percentage of facilities offering detox and other opioid-related services.



The data was also reflective of lower risk communities admitting more casual drug users. For a higher risk community, it was more likely for an admitted patient to be using their primary substance on a daily basis.

# Model Analysis

For this exercise, I chose to conduct the following analysis:

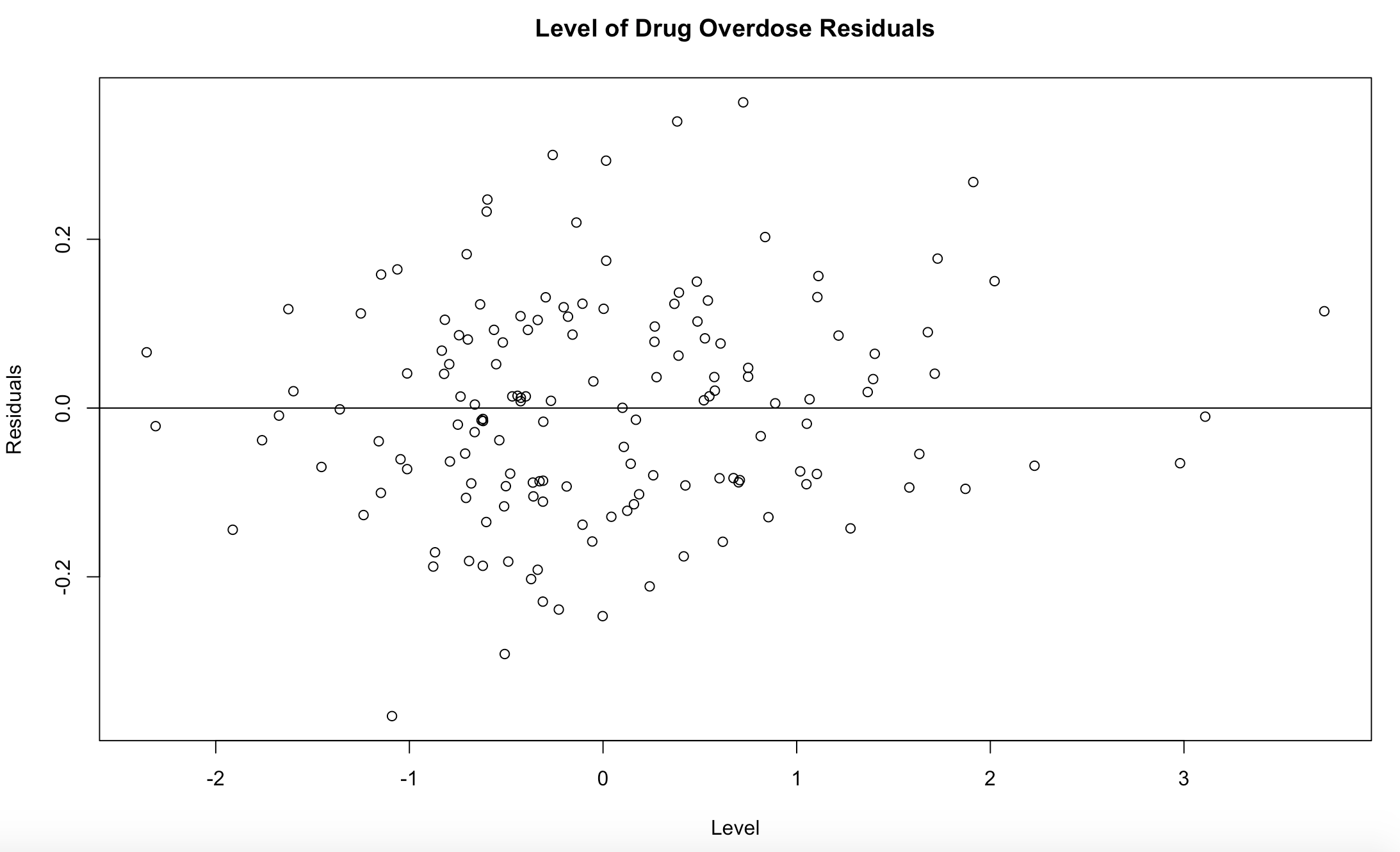
* Multiple Linear Regression Analysis
* Exploration of Decision Trees for Variable Importance

Instead of focusing on prediction, I focused on the discovery of new information in regards to states with higher drug-related death rates. I chose these models because they produce comprehendible results about variable importance. Each model was used to determine variable significance from the data sets that I created.

I validated these models based on a 10-fold cross-validation approach. To properly train the models, I randomly shuffled the 152 data points and chose the first 122 points for training and last 30 points for testing.

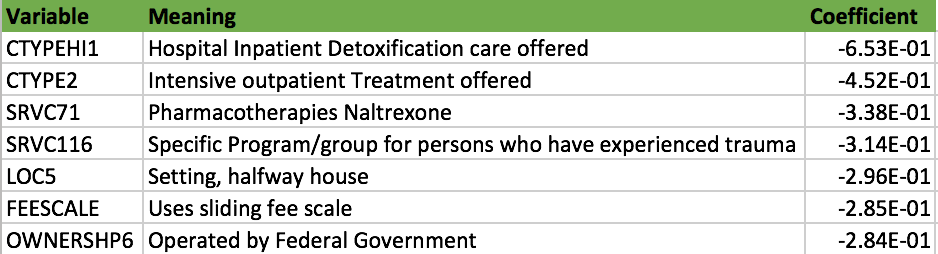
## Multiple Linear Regression Analysis

For my multiple linear regression analysis, I used a step AIC model to search through the 143 variables available. After using this, I focused on what variables related to services offered that significantly decreased death rates. The model evaluated with an adjusted r-squared of 0.96 and a test RMSE of 0.11 when predicting the scaled number of drug-related deaths. This translates to an error of 0.11 standard deviations. I also decided to observe the residuals of the model:



The residuals of the model don't show any signs of a pattern which could point to a problem in the dataset or the model itself. The residuals are relatively equal along the zero line and well dispersed. This relationship displays that the data and model are functioning well. We can also see that the majority of the residuals are between 0.2 and -0.2. Showing that the majority of predictions are off by only a scale of 0.2, we can interpret this as the model not significantly over or under predicting drug-related overdoses.

I found the following variables to have the most negative and significant effect on death rates within states:



Finding these variables is reflective of my exploratory analysis that supports having beneficial services available can significantly help a community in battling a drug crisis. From this analysis, we can see that a society that invests in creating the resources necessary for preventing drug-related deaths will be more successful in having a lower death rate. This type of observation is not surprising, but it does highlight how much investing in drug-related services can help the epidemic. If we continue to underfund communities with high drug-related death rates, those communities will continue to suffer.

It is especially helpful if the community receives government funding for substance abuse services. The multiple linear regression analysis showed that amongst all types of ownership, state, local, and federal government services had a beneficial effect on death rates within states. The least effective owners were private for-profit services. One could hypothesize that government-funded initiatives are more successful than private efforts due to a higher investment in bettering the community and less focus on creating revenue.

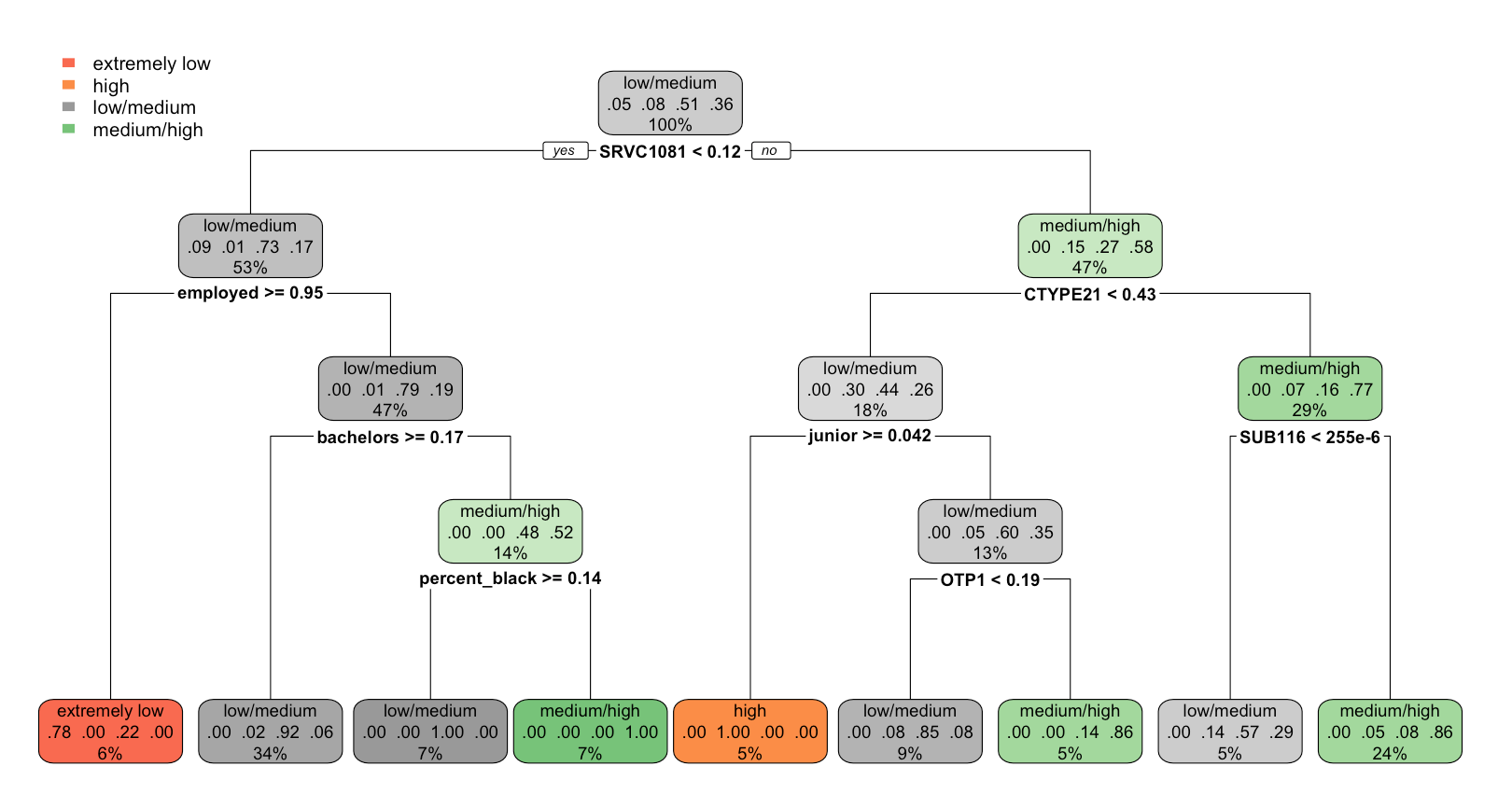
## Exploration of Decision Trees for Variable Importance

For this portion of the analysis, death rates were separated into the following categories:

* Extremely low (<-1.5)
* Low/medium (<0 & >-1.5)
* Medium/high (<1.5 & >0)
* High (>1.5)

### Census, Services, & Admissions

For my decision tree, I chose to use a simple CART decision tree using rpart() and Gini index as the feature selection criterion. I utilized a simple CART tree so we could easily visualize results and explore possible trees via pruning. For this process, I chose to use the entire dataset to use this as a descriptive exercise.



As we can see from this one decision tree, the prediction of a community's death rate is highly dependent on the services offered and the overall well-being of the community. The primary node is based on whether the drug Vivitrol is provided. Vivitrol is a once-a-month treatment that is used along with counseling to treat opioid dependence. One can also see that if intensive outpatient care is available, the probability of a high death rate is lower. Finding that this combination affects the rate of drug-related deaths could show that more in-depth long-term treatment could be beneficial to a community.

The forest also shows us that even with Vivitrol, if the community is less diverse, poorly educated, and unemployed, it is more likely to have a higher drug-related death rate. This highlights that even with the proper treatment, we may need to make a change in the macroeconomic environment of a high death rate community. We can see this finding solidified by the tree showing that without Vivitrol offered and a lower rate of higher education, the community has a higher probability of a high death rate.

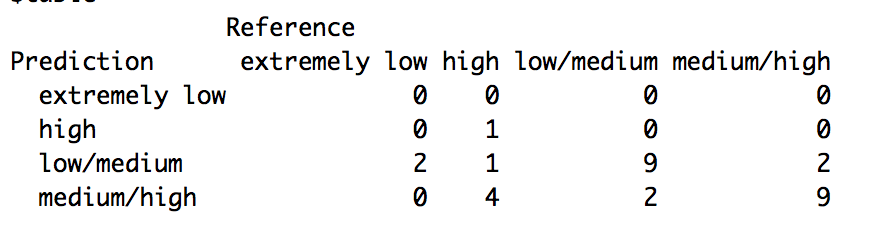
The decision tree above proves to be interesting because it is extraordinarily simple for a dataset with over 100 variables. This further cements my hypothesis that there are overarching factors that contribute to drug overdose rates such as quality of treatment and community education.

I validated this approach by running a random forest algorithm using 1000 possible trees on the data for further accuracy. I also used the training and testing sets to determine the accuracy of the random forest and how this process reflected the variable importance. The training set produced the following:

## 

From this output, we can see that the top three variables are related to the services offered and the education of the community. It is important to note that SRVC711 is related to the drug Naltrexone which is another drug used for the long-term treatment of opioid addiction. Based on this variable importance output, one can observe that obtaining long-term treatment, local funding (OWNERSHP4) and higher education affects the community's death rate.

To validate the model, I used it to predict the test data specified above. The model scored a 0.63 accuracy rate and a kappa of 0.426. This result was not particularly satisfying, but the model results were impressive:



As we can see, the model was decent at predicting middle ground death rates, but not as successful at predicting more extreme death rate communities. This could be a reflection of either the binning used for the model or that the dataset needs more feature engineering to identify more extreme behaviors better. For future iterations of this project, it would be interesting to explore why the model performed poorly for extreme values.

## Validation

Since the majority of the observations from this analysis already proved observed correlations, it was relatively easy to validate the model. The data supported current findings that poor areas and communities experiencing opiate abuse were more prone to higher drug-related deaths. The following article gives a detailed observation of how not investing in drug addiction treatment is a leading cause of the current drug climate:

<https://www.vox.com/identities/2017/10/2/16328342/opioid-epidemic-racism-addiction>

Other materials can be found in the appendix.

## Conclusion

After conducting my exploratory and model analysis, I found that I could support my hypothesis that communities with lower levels of education, government investment and access to aide are more likely to have issues with drug-related deaths. This was supported by both models showing evidence that the mix of a better macroeconomic environment and access to long-term intensive care was beneficial to the drug overdose death rate in a community. My hypothesis was also supported by the overwhelming trend that poor, uneducated, and less diverse communities tended to experience higher levels of drug overdoses. The poster child for this type of community was West Virginia.

Based on these findings, I would recommend that we invest in our forgotten communities that are not thriving in current economic environments. The data shows that for many of these states, they experience very little change in their environment and continuous growth in drug overdoses. We also see that communities with higher drug-related deaths typically have less access to help. If we continue to abandon these communities, this epidemic could become a cyclical effect on their communities. I would recommend on advocating for government's creating more funding for education and substance abuse services focused on solving the long-term problem by providing long-term solution drugs (i.e., Naltrexone and Vivitrol) and counseling.

For future analysis and modeling, I would suggest continuing with my work to build an extensive data set for similar study. This work would involve feature engineering to include more years of data and potentially splitting data from the state level to a city or county level. The main problem that I faced for this project was lack of data. I was able to build an extensive feature library, but I was unable to create a lot of data from this process (ending with 152 data points). If I were to continue on this path, I would work towards building a data set with more instances to observe.

I would also allow more time for exploring collinearity and the correlation between variables. The r-squared results from the multiple linear regression could be a flag for collinearity causing favorable model results. It may be a good idea to address this in further iterations of this data analysis.

# Appendix

# Datasets:

U.S. Census Bureau (2012). *Selected housing charactersitcs, 2007-2011 American Community Survey 1-year estimates*.

U.S. Census Bureau (2013). *Selected housing charactersitcs, 2007-2011 American Community Survey 1-year estimates*.

U.S. Census Bureau (2014). *Selected housing charactersitcs, 2007-2011 American Community Survey 1-year estimates*.

U.S. Census Bureau (2015). *Selected housing charactersitcs, 2007-2011 American Community Survey 1-year estimates*.

CDC - National Center for Health Statistics - Homepage. http://www.cdc.gov/nchs/. November 25- December 15, 2017.

SAMHDA (2012). *Treatment Episode Data Set: Admissions 2012.*

SAMHDA (2013). *Treatment Episode Data Set: Admissions 2013.*

SAMHDA (2014). *Treatment Episode Data Set: Admissions 2014.*

SAMHDA (2012). *Treatment Episode Data Set: Discharges 2012.*

.

SAMHDA (2013). *Treatment Episode Data Set: Discharges 2013.*

SAMHDA (2014). *Treatment Episode Data Set: Discharges 2014.*

SAMHDA (2012). *National Survey of Substance Abuse Treatment Services 2012.*

SAMHDA (2013). *National Survey of Substance Abuse Treatment Services 2013.*

SAMHDA (2014). *National Survey of Substance Abuse Treatment Services 2014.*

SAMHDA (2015). *National Survey of Substance Abuse Treatment Services 2015.*

# Drug Abuse in America: Articles for Validation:

Lopez, G. (2017, April 04). When a drug epidemic's victims are white. Retrieved December 16, 2017, from <https://www.vox.com/identities/2017/4/4/15098746/opioid-heroin-epidemic-race>

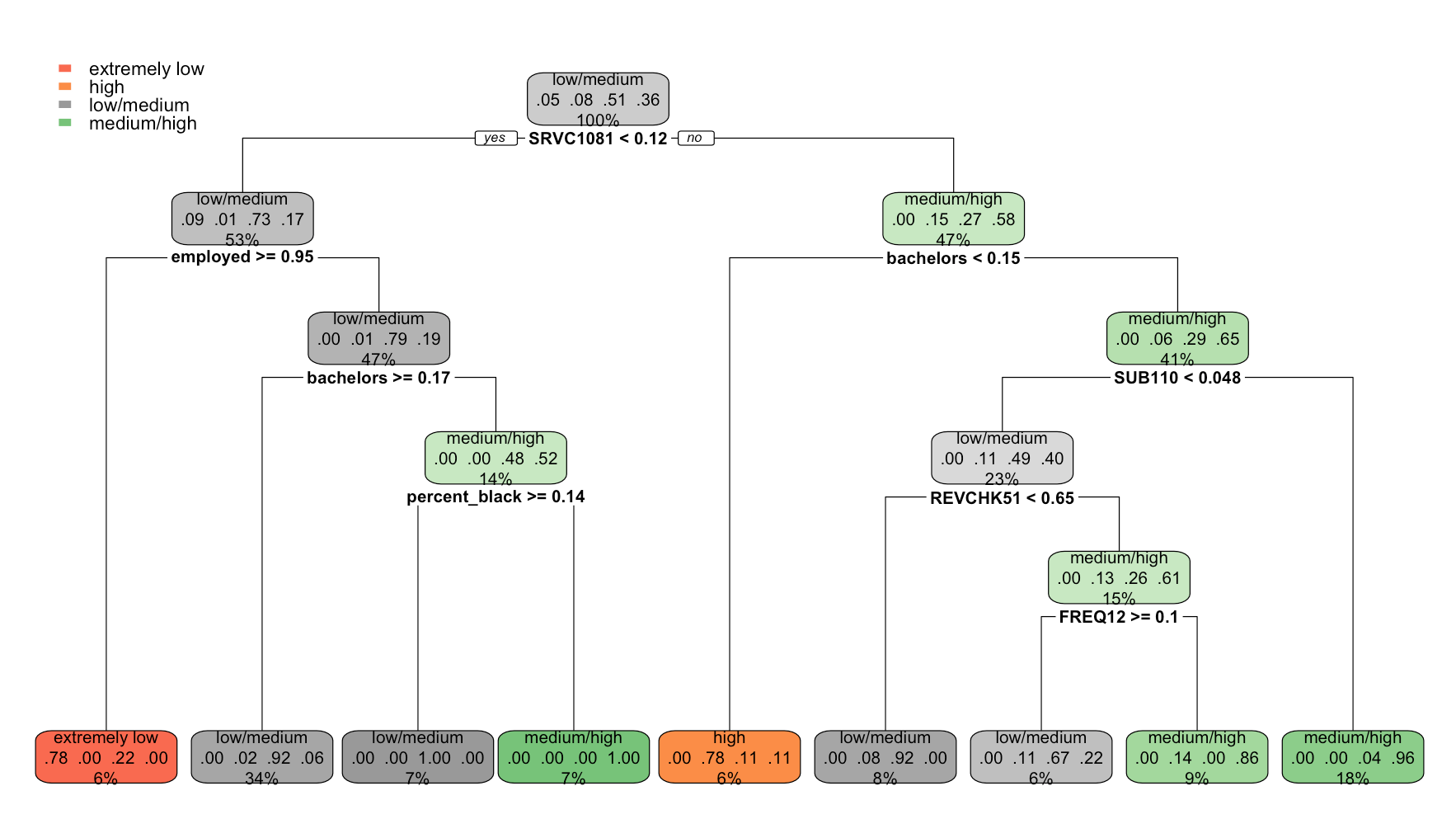
McGreal, C. (2015, November 12). America's poorest white town: abandoned by coal, swallowed by drugs. Retrieved December 16, 2017, from <https://www.theguardian.com/us-news/2015/nov/12/beattyville-kentucky-and-americas-poorest-towns>

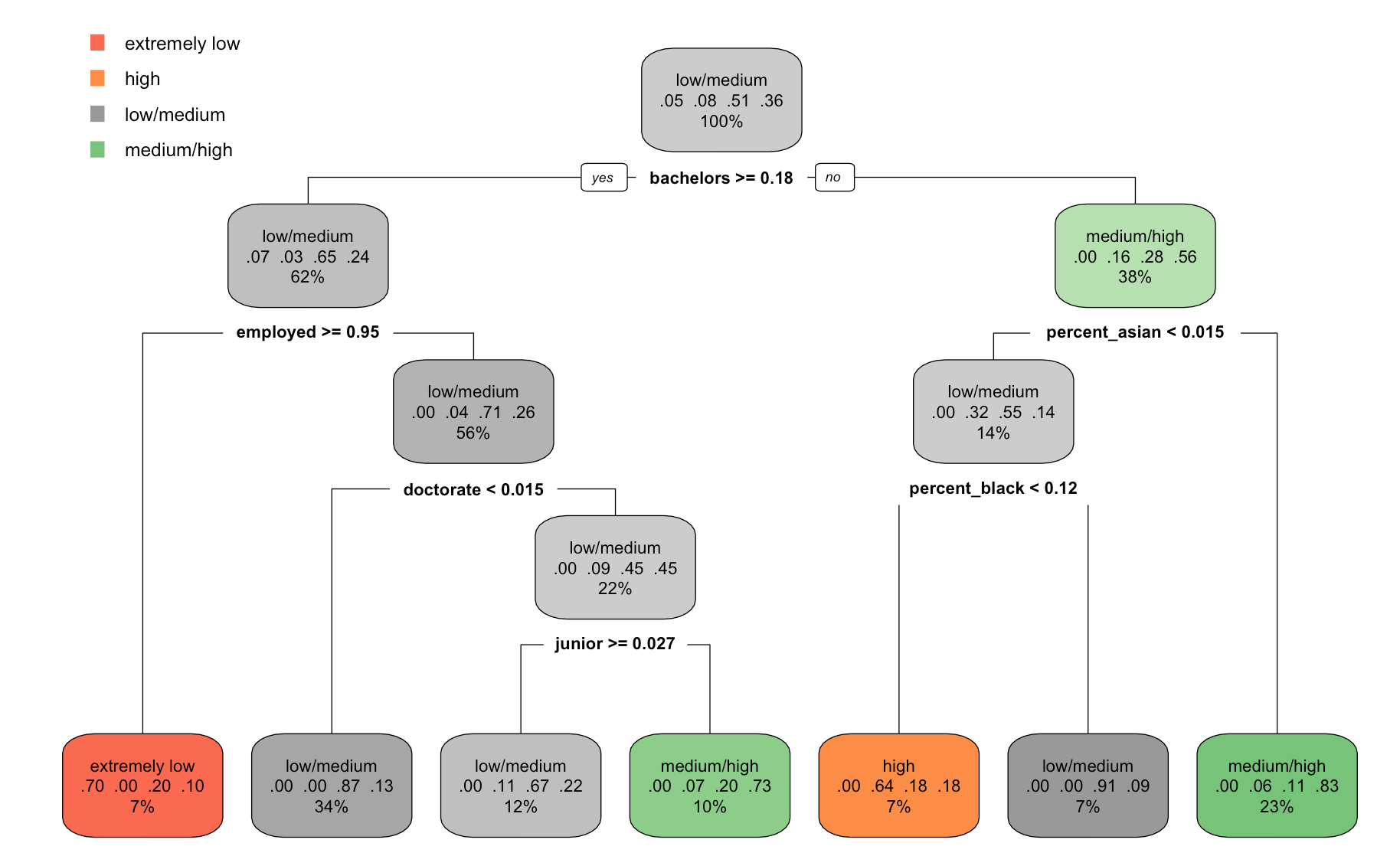
Rothman, J. (2017, June 19). The Lives of Poor White People. Retrieved December 16, 2017, from https://www.newyorker.com/culture/cultural-comment/the-lives-of-poor-white-people

Talbot, M. (2017, June 18). The Addicts Next Door. Retrieved December 16, 2017, from https://www.newyorker.com/magazine/2017/06/05/the-addicts-next-door

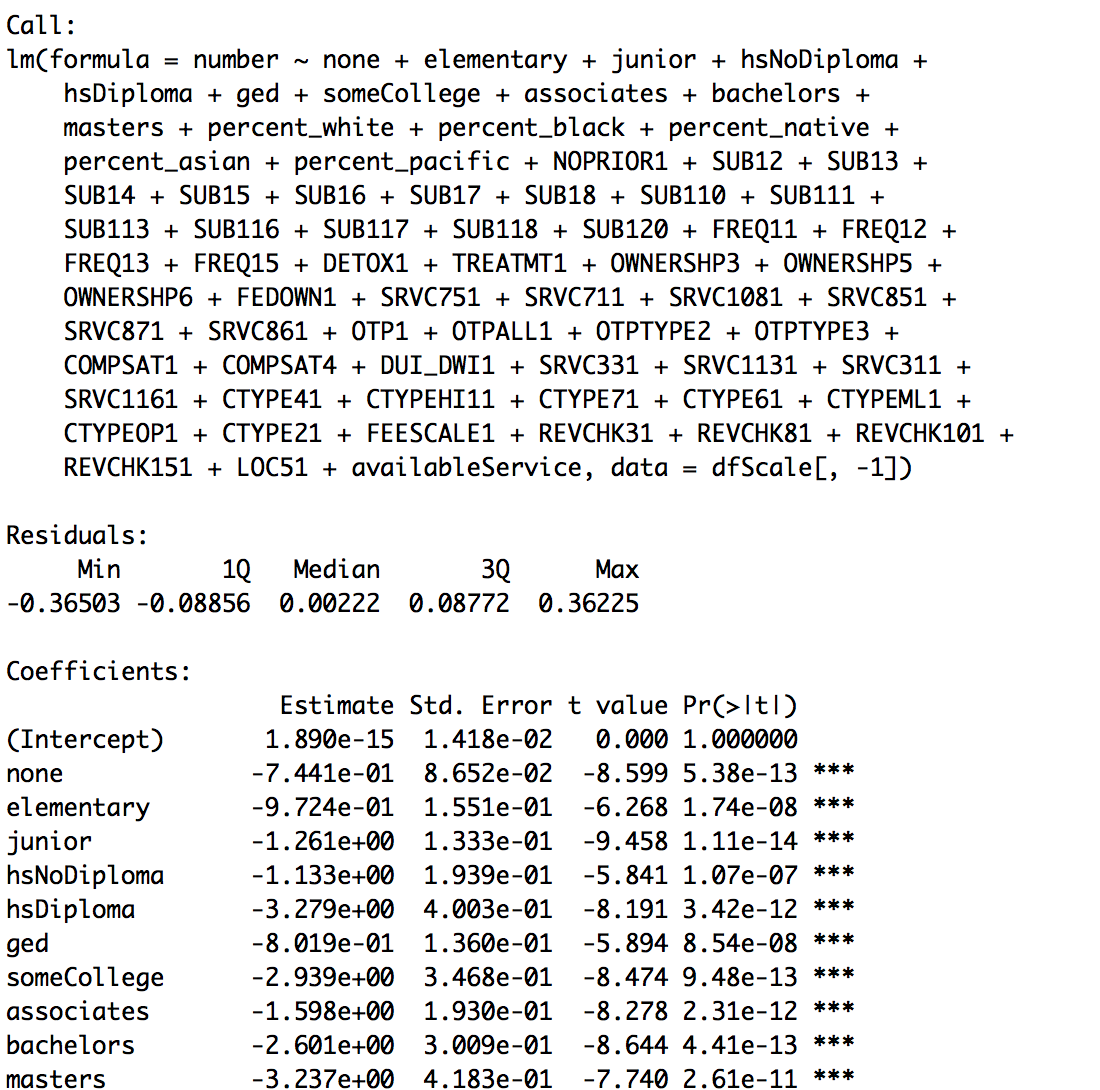
Addiction in West Virginia

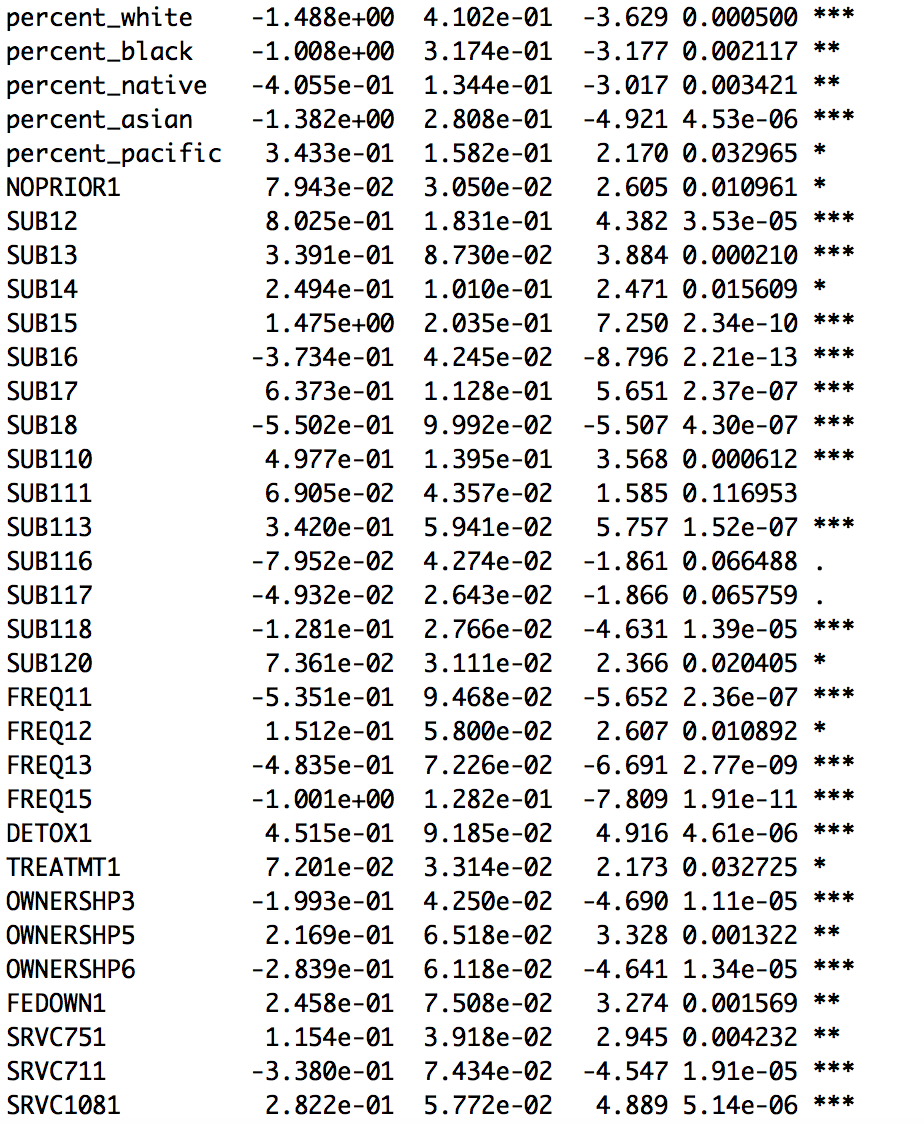
# Decision Tree Output

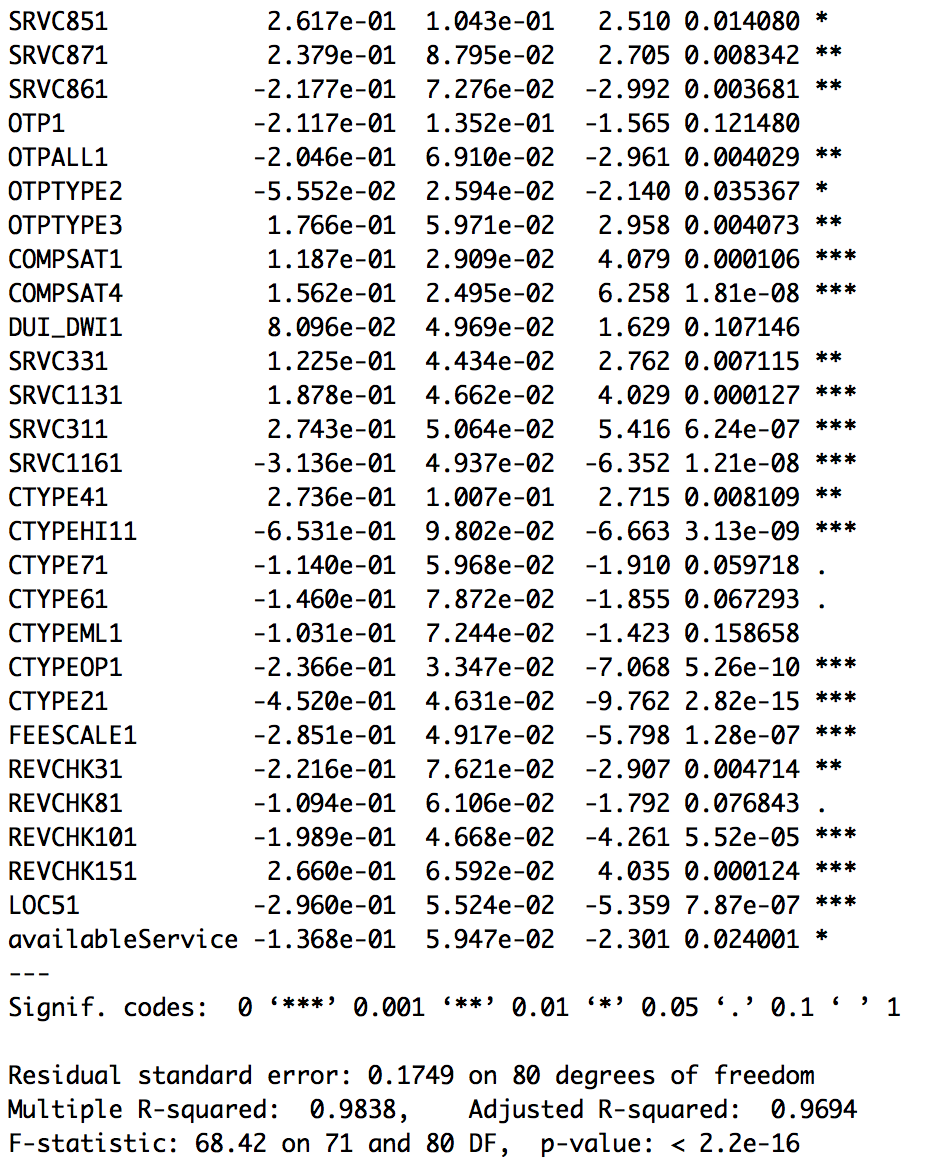




# Model Outputs



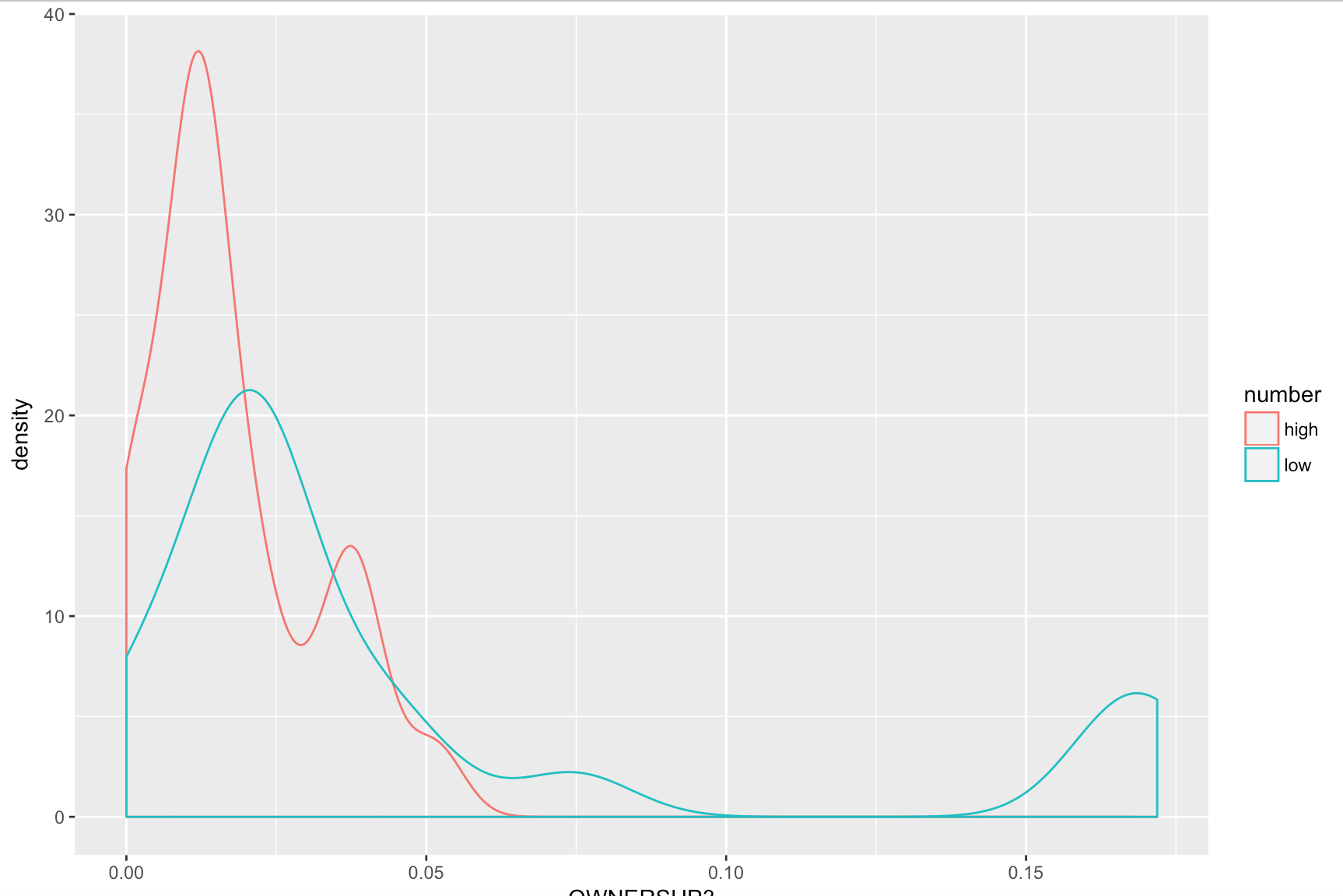




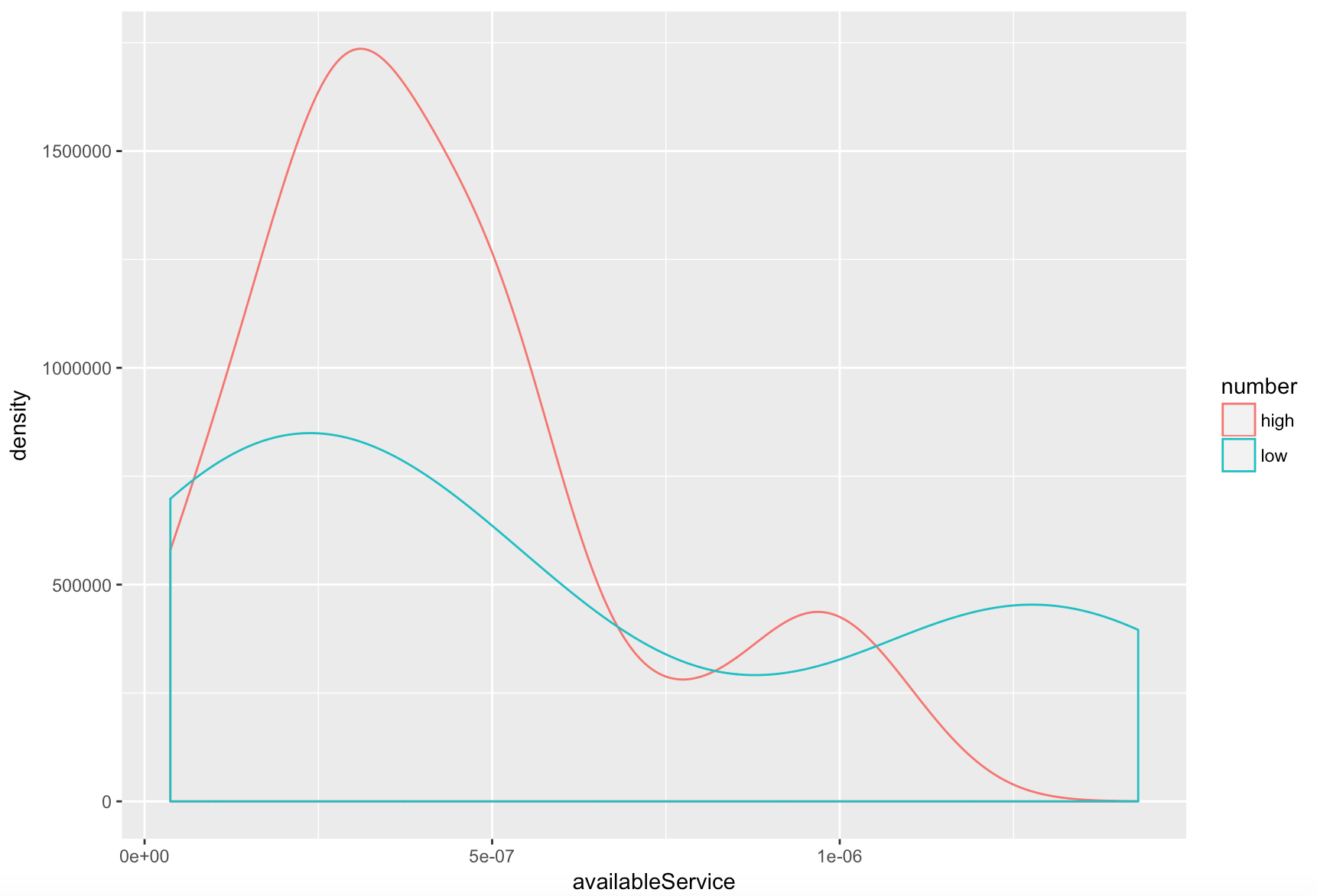


# Density Plots

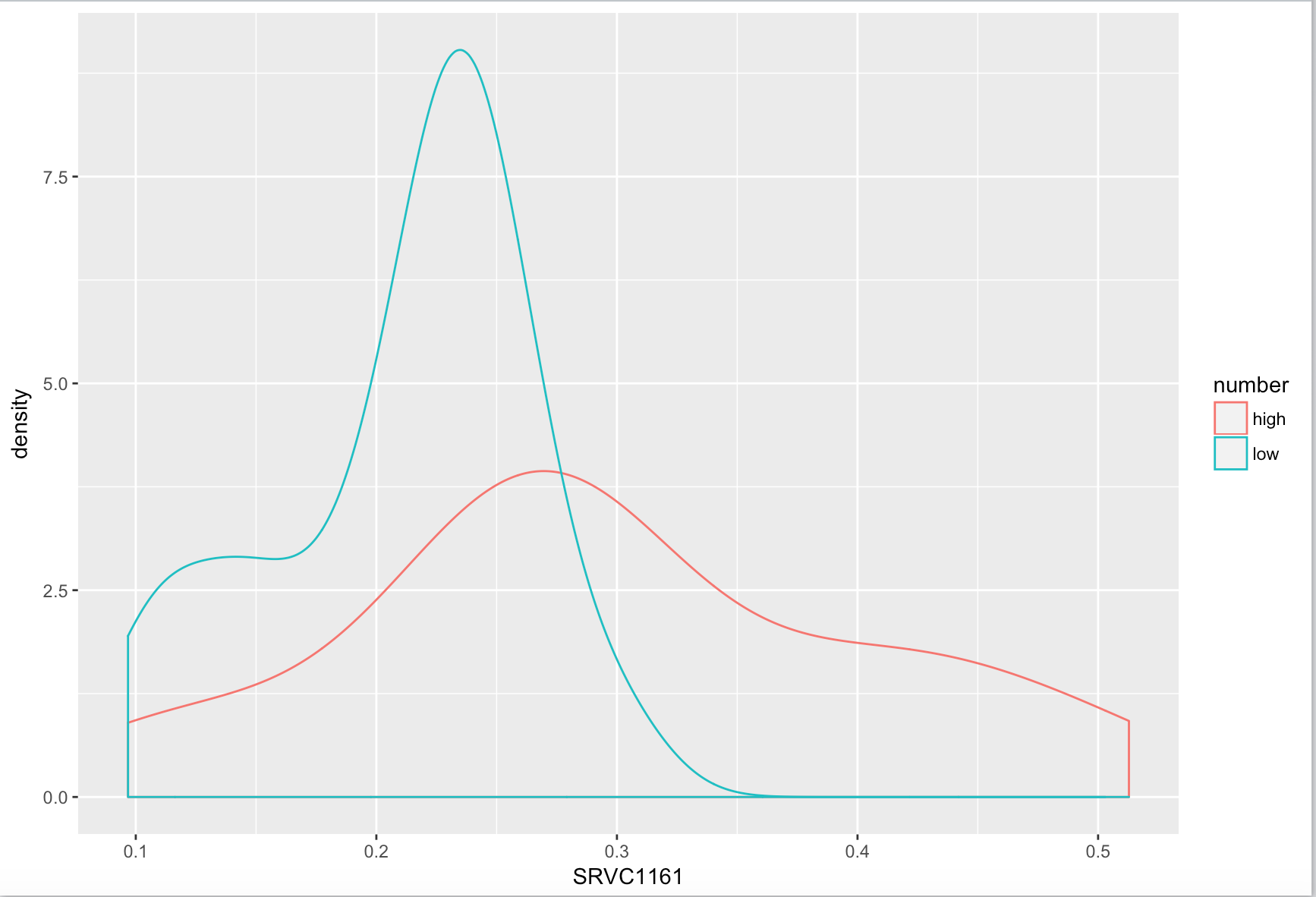
FEDERAL OWNERSHIP



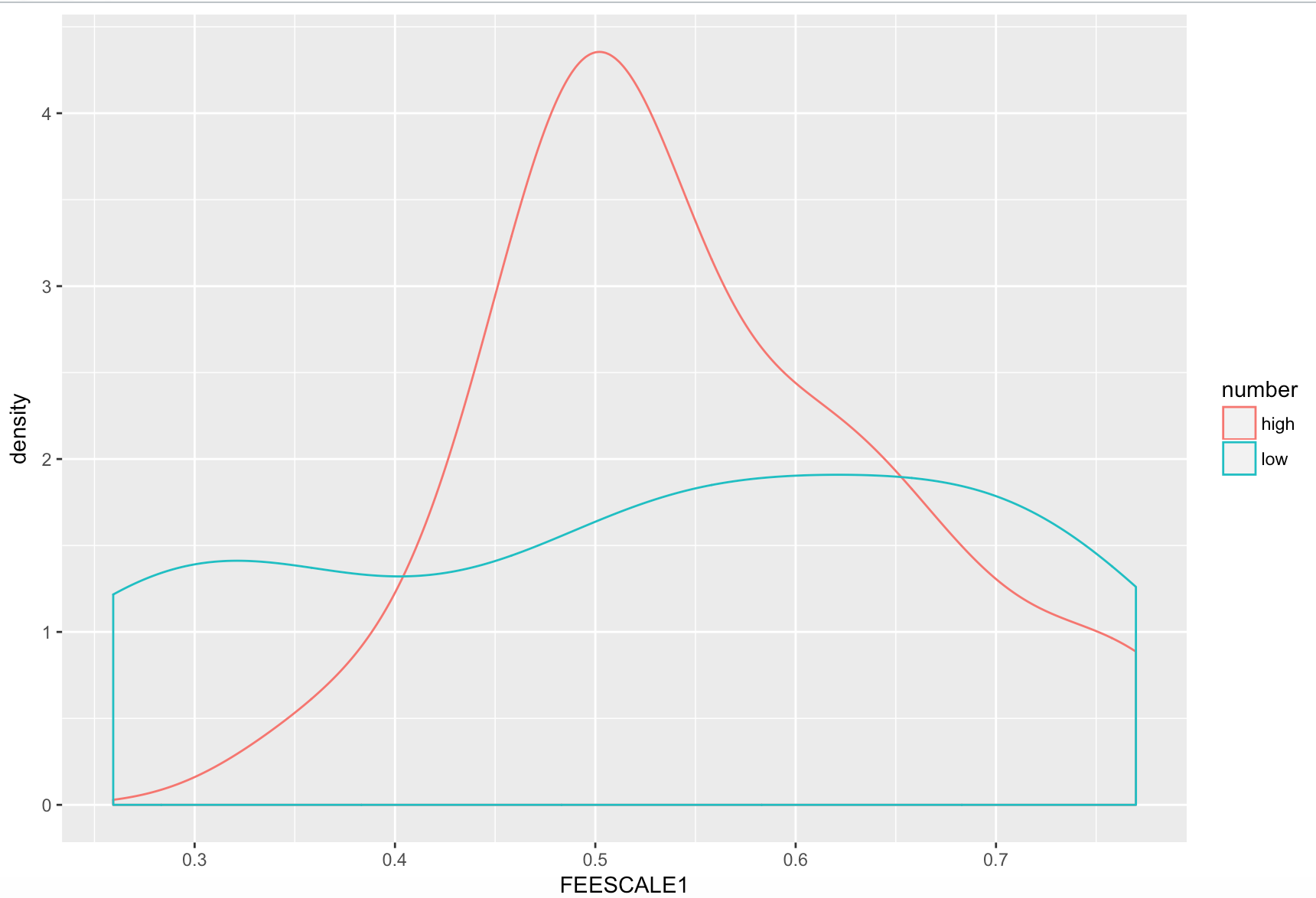
AVAILABLE CLINICS



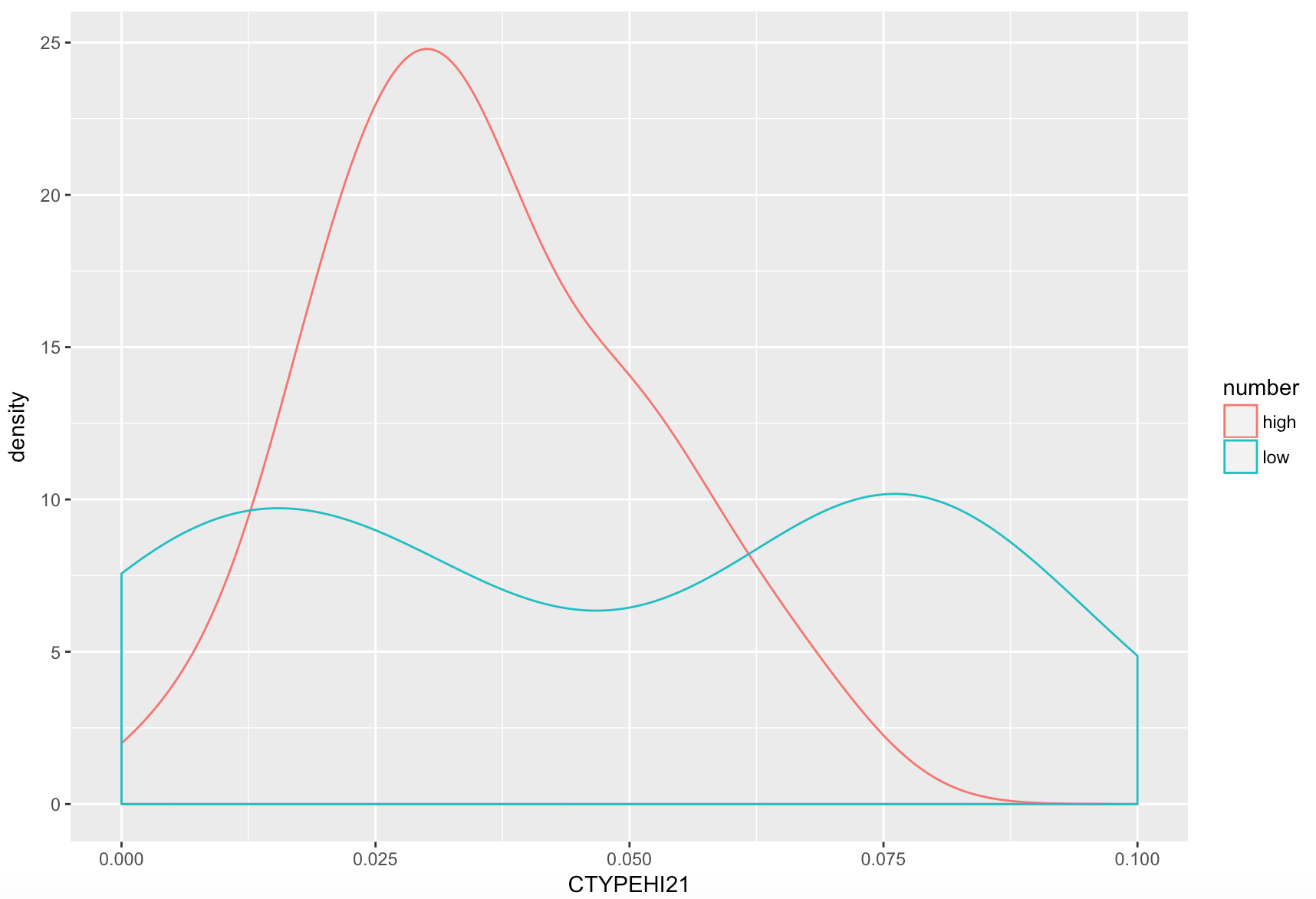
PROGRAM FOR TRAUMA SURVIVORS



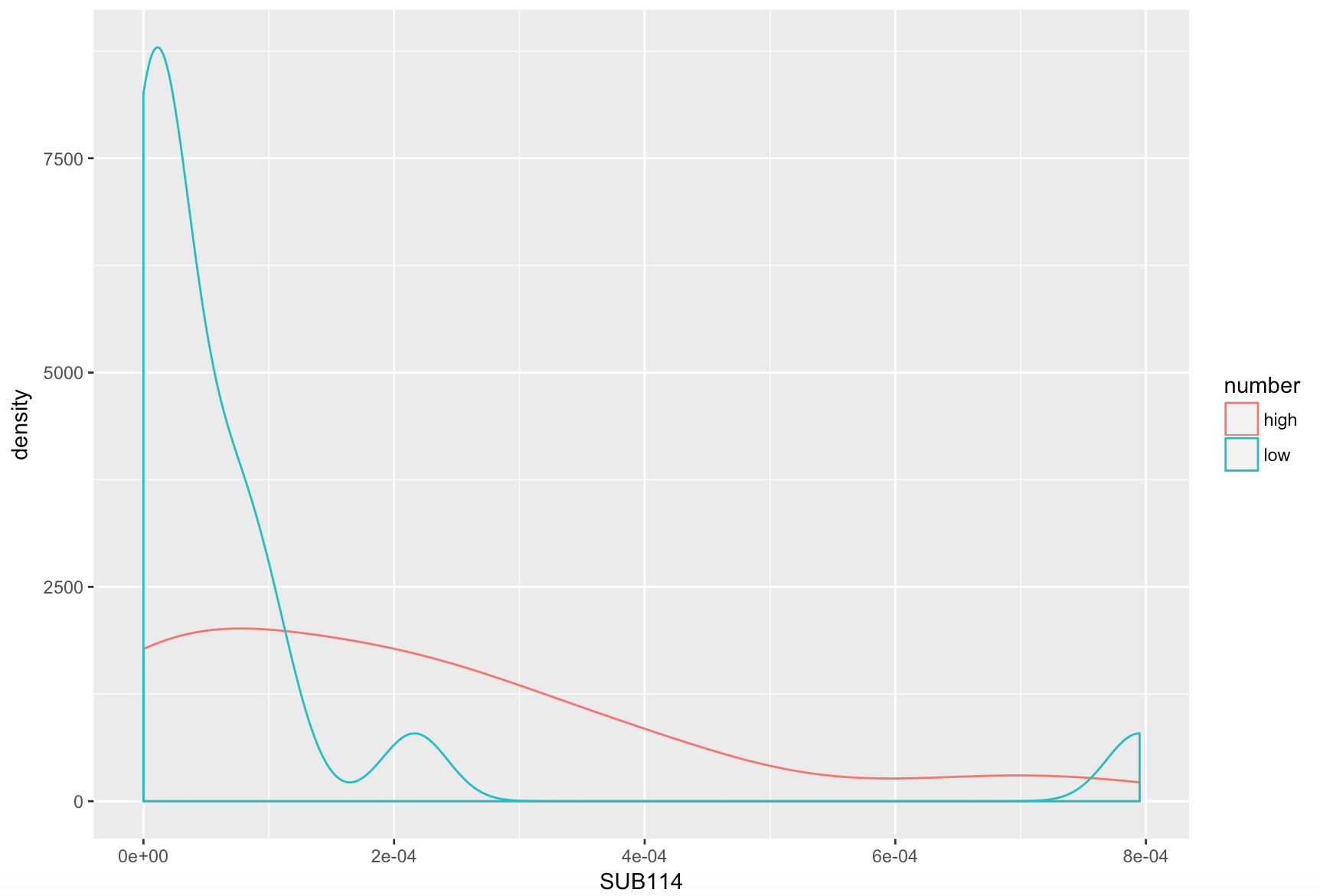
SLIDING PAYSCALE



DETOX OFFERED



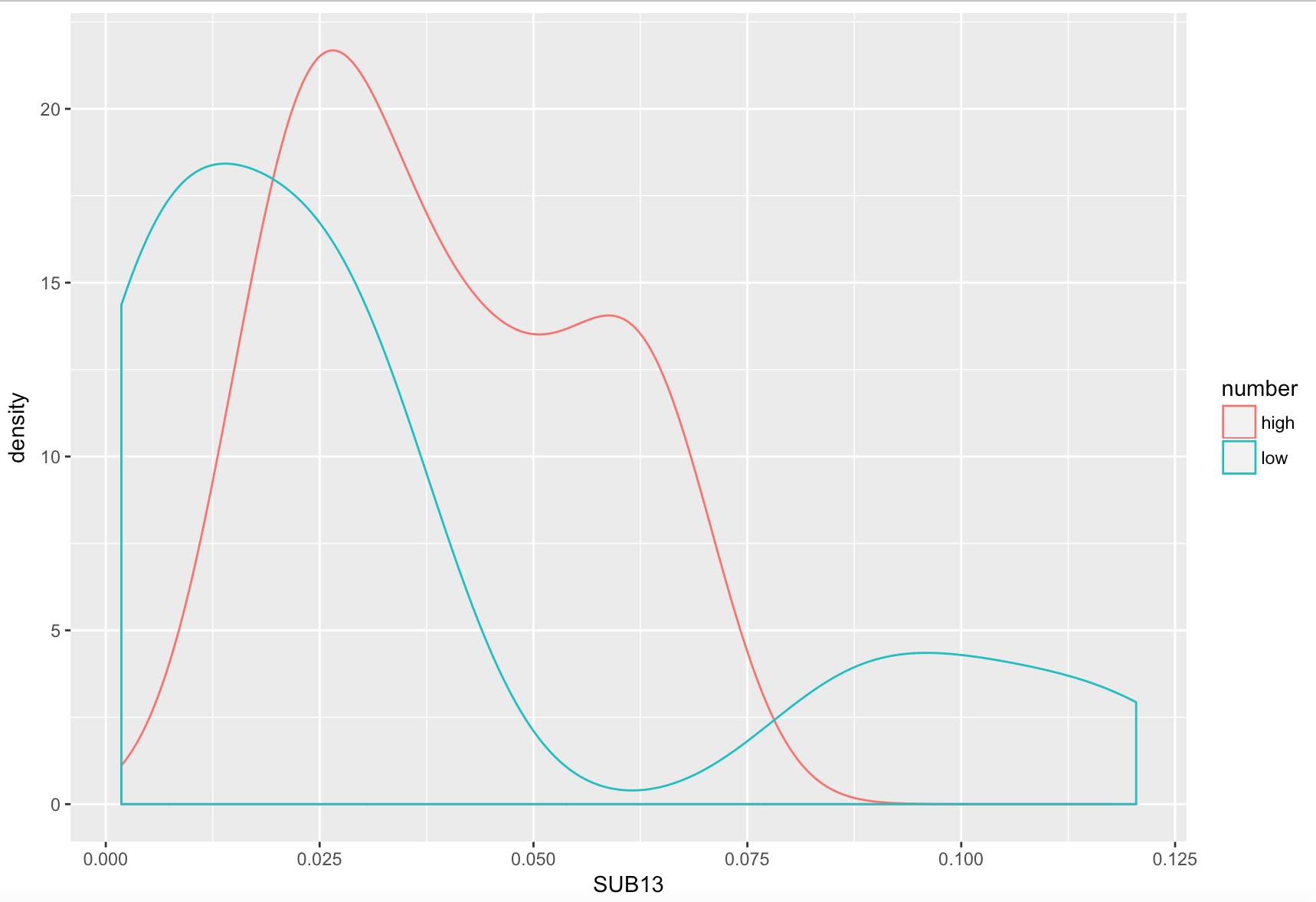
OTHER NON-BENZODIAZEPINE TRANQUILIZERS



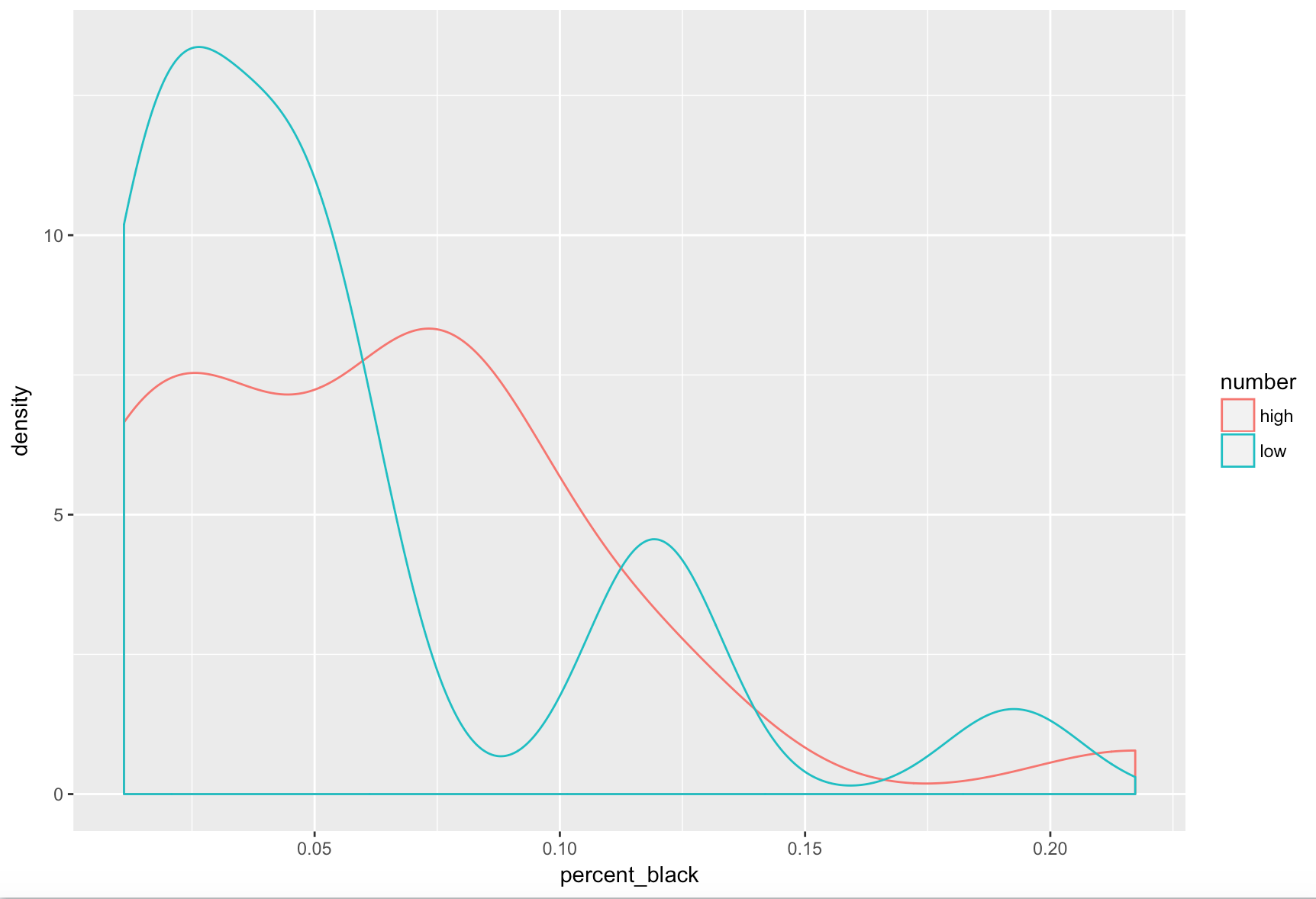
HEROIN



BENZODIAZEPINES



PERCENT BLACK



PUBLIC ASSISTANCE AVAILABLE

