# Limitations of the Dataset[1]:

The NSDUH is the annual estimate of drug use patterns among civilian members of non-institutionalized American society. It is designed to be a highly private and confidential in order to encourage honest reporting of delicate and sensitive issues like drug-use. But, fact is that there are certain limitations to the data sampling that are inherent to the data collection when collecting data about human behavior, that will certainly reflect on the insights that can be gained from this endeavor. The limitations include

## Truthfulness of respondents:

The drug use is self-reported data whose validity is directly proportional to the honesty and the memory of the respondents and there is no way to verify the truthfulness of their claims. Though the surveys were designed to encourage honesty there is a measure of wrong information that would have crept into the data.

## Cross Sectional data collection:

Respondents were interviewed only once and there was no follow up in the subsequent years. Insights that could be made about alcohol were only those based on data obtained in one cross sectional slice of time rather than the changing drug habits of one person over time.

## Target population of the survey:

The respondents were chosen from the non- institutionalized civilian population of the American states. This criterion excluded people serving in the army, people in prisons and considering the data being collected would mean excluding a meaningful subset of data. Close to just less than 2% of population was being excluded and this would have introduced a significant skew to the data.Assumptions:

# Assumptions[1]:

## Assumed Random Sampling:

The data is assumed to be randomly sampled even thou it is not so.

## Weighting by population:

The sampling process is weighted by the population of the state that it is being collected from. Within each State, sampling strata called State sampling regions (SSRs) were formed. These SSR’s were based on their population. The eight large sample States(Top 8 states by population) were divided into 48 SSRs each; the remaining States were divided into 12 SSRs each. Therefore, the partitioning of the United States formed 900 SSRs.

## Oversampling of the young-adult population:

The sampling distribution was not uniform across all age groups. The survey process over sampled the young adult population to derive more insights into them. The surveys were distributed equally among three age groups (12-18 ,18-25 ,26 and higher). This implies that the surveys were taken more by young adults than people above the age of 26.

## Stability

The process of collecting data on these variables is the survey that respondents took to answer these questions. And the survey process changes over time. Till a certain year, the respondents were voluntary but now get paid an amount of $30. Also, the distribution of SSR’s change with changing population distributions across the states. Some variables are added and removed from the dataset based on feedback to encourage honesty and confidentiality. These changes happen year over year and the survey process is kept constant h the year that the survey is taken on and so for this reason we are assuming that the surveying process is stable.

## Normality of errors

The distribution curve of the errors is assumed to be normal across all the possible values of the predictors. This assumption can be verified by doing a standardized residuals test by plotting the standardized plots vs dependent variable, standardized plots vs predictor variables plots and checking for Heteroscedasticity. The normality can also be checked using the Shapiro-Wilk test

# Why a low R^2 does not mean a bad model?

The interpretation of R^2 is the amount of variability in the dependent variable that can be explained by the model. In a simple linear regression, it is a measure of the correlation between the dependent and independent variable, squared to remove the sign.   And in multiple regression it is a cumulative measure that represents the correlation between all the variables involved in the model, ie between the dependent and the independent variables and between the independent variables themselves [2]. Let us consider a clinical trial of a drug as an example to consider how much explanation of the variable is significant. As with most drugs the effectiveness of the drug varies as per the individual who is undergoing the clinical trial. So the amount of variance in effectiveness that can be explained by the model is very low and consequently has a very low R^2. But the model variables might be significant (B values). A result like this could potentially save many lives in the long run and be worth millions of dollars if the model results in the drug’s approval for widespread use [2].  Even when R-squared is low, low P values still indicate a real relationship between the significant predictors and the response variable. The coefficients estimate the trends while R-squared represents the scatter around the regression line. The interpretations of the significant variables are the same for both high and low R-squared models. So, low R-squared values are problematic only when you need precise predictions.

In some cases, it’s possible that additional predictors can increase the true explanatory power of the model. However, in other cases, the data contain an inherently higher amount of unexplainable variability with each added independent variable to the model. And in our case, we are considering human behavior and that too the behavior of alcoholics, so it is normal to expect a degree of unexplained randomness attached to the model [3]. The point of this analysis was to see whether there was a relationship between alcoholism and a multitude of other factors both concerning the present behavior and past demographics of respondents, however small the relationship maybe. Hence we are not placing much emphasis on the low R squared value obtained for the final model obtained in our analysis.

## Refrences

[1]- National Survey on Drug Use and Health,2012 Codebook

[2]-<https://people.duke.edu/~rnau/rsquared.html>

[3]- http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-a-regression-model-with-low-r-squared-and-low-p-values