Recreational Drug Use and Mental Health

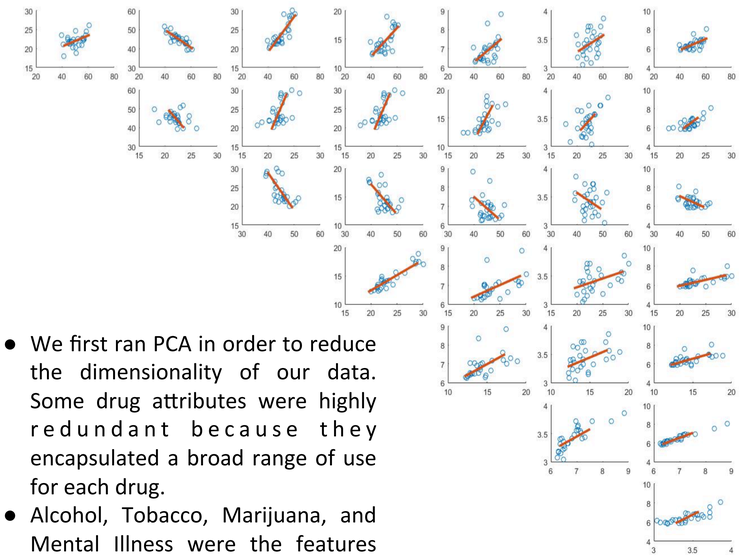
Group 7: Ji Chung, Michael Nguyen, Gabriel Catalano, Sara Brooke
University of California San Diego, Department of Cognitive Science

Abstract

- Our data comes from the National Survey of Drug Use and Health, collected across United States Of America from 2012-2014 and published in 2014.
- This dataset consists of many features about drug use and mental health, such as: use of alcohol in the past month or suicidal thoughts in the past year, recorded for several designated regions within each state.
- Narrowing down our total data, we analyzed observations from the states of California, Michigan, Texas, and New York.
- We aim to find the relationship, or linear regression, for each state's rates of mental illness and the specific drug that most influences those rates in that state.
- This methodology assumes that rates of mental illness are linear combinations of different drug use factors

Preprocessing: PCA

Figure 1: *California PCA subplots of the Alcohol dimensions*



- We first ran PCA in order to reduce the dimensionality of our data. Some drug attributes were highly redundant because they encapsulated a broad range of use for each drug.
- Alcohol, Tobacco, Marijuana, and Mental Illness were the features we extracted and reduced each category to one feature, respectively; Table 1.
- The subplots shown above, figure 1, plots Alcohol's principle components against each other for the state of California in order to compare the variances and find the one with the most variance.

Table 1: *Eigenvalues of each Principal Component derived during PCA*

Alcohol	48.329	4.7323	2.1391	1.5810	0.2041	0.1483	0.010
Tobacco	0.1599	11.7181					
Marijuana	11.166	0.1973	0.0667				

Analysis: Multivariate Regression

- For our principal analysis, we ran a multivariate linear regression algorithm across 5 attributes of drug use rates against mental illness rates for each state with a particular focus on the relative strengths of the weights of the resulting multivariate linear model.
- Our state models were created using cross-validation regression techniques, with the regression model training on 70% of the data points and being tested on the remaining 30%, see Table 3.
- The weights with the greater absolute magnitudes were interpreted as having the larger predictive value for mental illness in that state, see highlighted values Table 2.

Table 2: *Multivariate Regression Model Weights*

State	Alcohol	Marijuana	Tobacco	Cocaine	Prescription	Bias/Error
California	-0.0393	0.0135	-0.1433	0.3011	**1.3934**	-7.4049
Michigan	-0.1796	0.5031	0.0695	**-4.7420**	2.6015	1.7024
Texas	-0.0360	0.4563	-0.2249	0.3093	**3.9884**	17.2767
New York	-0.0214	0.2690	0.0623	-1.2194	**-3.4075**	17.4958

Table 3: *Total Mean Squared Error*

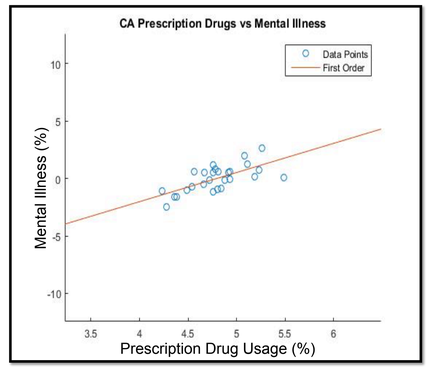
States	Total Mean Sum Squared Error
California	0.5011
Michigan	0.4701
Texas	0.5403
New York	0.6936

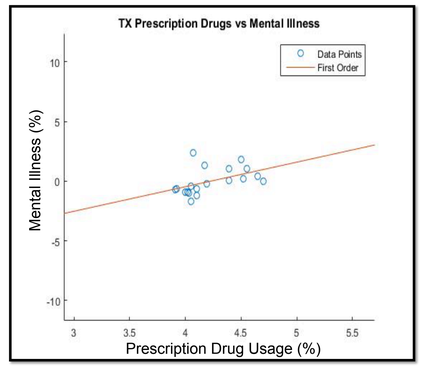
Results

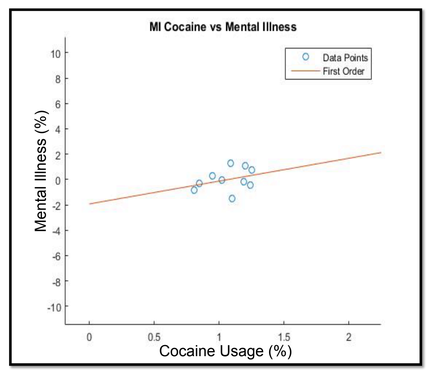
Table 4: *These were the resulting mean sum squared error (MSE) of different orders after having used the leave-one-out model validation on the data regarding the greatest predictive value for a specific state. The highlighted cells and bolded numbers are the least mean sum squared error for the states and is thus the line of best fit.*

Polynomial Linear Regression	Mean Sum Squared Error $MSE = \frac{1}{n}\sum_{n=1}^{n}(y_p - y_a)^2$		
Data Types	**1st Order**	**2nd Order**	**3rd Order**
California	**0.7733**	2.2081	2.4860
Michigan	**0.4701**	0.9109	1.0568
Texas	**1.0402**	1.5928	1.8491
New York	**0.5421**	1.5847	1.5787

- Note: Sometimes the MSE changed, causing different orders to have the least MSE; however, majority of the times, it was the highlighted order that appeared frequently. Therefore, our group has decided to use these orders.







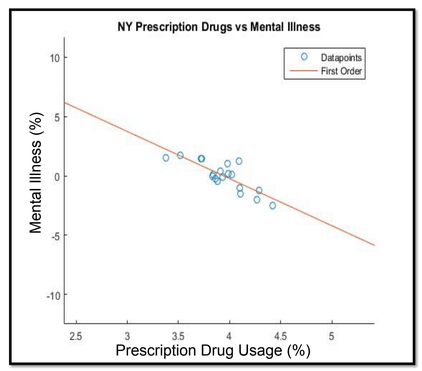


Figure 2-4: *After we found the least mean sum squared error, we represented the line of best fit for each model, finding the relationship between state's highest substance abuse and mental illness.*

Discussion

- The predictive value of drug usage on mental health, while initially promising, resulted in weak at best relationships between the components and their output.
- Looking at the bias terms of the multivariate regressions, we see that a lot of weight is unaccounted for, implying a large error or missing variable.
- Though we see that these drug measures can influence the mental illness rates of any given region, we see that they are not accurate enough or not complete enough to predict mental illness reliably or accurately.
- What we could do next is see if with more data our results would provide more insight, or if accuracy is improved when we look instead at mental illness rates and see if that can predict drug use.

References

SAMHSA. (2016). *2012-2014 Substate Estimates of Substance Use and Mental Illness.* [Survey Data from NSDUH]. Retrieved from http://www.samhsa.gov/data/population-data-nsduh/reports
Olson, Jake (2016) Linear Regression [PDF]. Retrieved from http://documents.scribd.com.s3.amazonaws.com/docs/9gd2n46zy85e8kud.pdf
Olson, Jake (2016) PCA Principal Component Analysis [PDF]. Retrieved from http://documents.scribd.com.s3.amazonaws.com/docs/8dyjpaklxc5eeql0.pdf
Special thanks to Jake Olson, Robert Gougelet, and Michael Hatch for theoretical development support