Patterns and Correlates of Substance Use and Socio-Health Demographic Factors

Tyler Park
Graduate Student
typark

Kun Woo Cho
Graduate Student
kwcho

Sai Srikar Kasi
Graduate Student
skasi

Minsung Kim
Graduate Student
minsungk

Abstract

Use of substances such as cigarette or alcohol is prevalent. Studies implicate that these prevalent substances lead to further use of severe drugs such as heroin. Significant number of reports have demonstrated the adverse effect of these substances on medical conditions. Given such relevance, it is important to identify whether there exists a set of information we can leverage to infer the cause of substance abuse. In this project, we want to find information that is related to cause and effect of substance abuse. We apply models for supervised and unsupervised learning to survey responses that contain information on substance use and mental health. While our models identified key features that are relevant and predictive of substance abuse, the results indicate that additional analyses and processing of the data are necessary due to inherent complexities of the dataset.

1 Introduction

Substance use and mental health issues affect millions of adolescents and adults in the United States and has become a major health crisis of epidemic proportions in recent times [4, 10]. Identifying the underlying cause of substance abuse and its impact on mental health are important subjects of study that can help prevent severe medical damage. The National Survey on Drug Use and Health (NSDUH) [3] is the primary source for statistical information on illicit drug use, alcohol use, substance use disorders (SUDs), and mental health issues for the civilian, non-institutionalized population of the United States. The number of new consumers for substances such as alcohol or marijuana have steadily increased [2]. Furthermore, studies indicate that frequency of deaths from drug overdose is associated with age and their demographics. In this project, we will study whether there is an explicit correlation between these information by analyzing the effect and cause of substance abuse from NSDUH dataset.

Specifically, we will divide our analysis into the following sub-problems: 1) Find features that can predict substance abuse 2) Analyze the effect of substance abuse on mental health 3) Find patterns and correlations among features with respect to substance abuse. We will address the first 2 tasks by applying supervised learning to find and evaluate predictive features for substance abuse and its impact on mental health. For the last task, we will apply unsupervised learning to search for latent structures to gain intuition on patterns in the data. We will evaluate our models using standard metrics and relevance of resulting features to the outcome variables. We expect our work will allow us to discover latent patterns that can enhance our power to predict substance abuse.

Related Work. Research on finding crrelates between substances and other factors has been conducted based on NSDUH dataset. For instance, [8] compares prevalence estimates and assess issues related to the measurement of adult cigarette smoking. [5] examines the pattern of polytobacco use among a large, nationally representative population over an extended period of time. Relationship between substance use and teen pregnancy using population-based samples is analyzed by [9].

2 Methods

2.1 Dataset

We use publicly available dataset [3] of survey responses collected by NSDUH consisting of 56,277 samples (individuals) with over 3,000 features (questions and answers) that include information on substance abuse treatment history, and perceived need for treatment, and questions from the Diagnostic and Statistical Manual (DSM) of Mental Disorders. We note that although the responses have been collected for over a decade, the dataset is not longitudinal as different individuals participated in the survey at each time point.

Feature Engineering. For preprocessing, we first find responses that correspond to 'not applicable', 'blank', 'bad data' (which are initially assigned very high numeric values) and convert these to 'Nan.' We remove features that contain more than 80% 'Nan.' Questions with categorical responses are converted to binary features via one hot encoding, and numeric responses are normalized across samples to range from 0 to 1. Any remaining entries of 'Nan' is assigned a value of 2 (for supervised learning). Then, we apply one-hot encoding since most of missing data are NMAR (this will be further discussed in Section 3.2). To address our subproblems, we partition our preprocessed data into 3 subsets that each contains features relevant to substance abuse, medical, and demographic information. We then divide our dataset into training, validation, and test sets according to the standard 8:1:1 ratio. Unless otherwise specified, this is default setting of our preprocessing. Effect of our preprocessing is shown in Figure 1.

2.2 Supervised Learning

The Cause Problem we address for supervised learning is predicting substance abuse using demographic features from the dataset. In particular, we focus on explaining use of severe, rare drugs such as Cocaine. We hypothesize that two sets of features would be predictive of drug abuse: A) Use of common, mild substance, B) Demographic information. The first hypothesis is motivated by the notion that some substances such as cigarette as well as alcohol may serve as **gateway** drugs that lead to further use of more severe substances. The second hypothesis is based on an assumption that individuals with economic and educational challenges are more likely to be exposed to hostile environments and thus have easier access to severe substances.

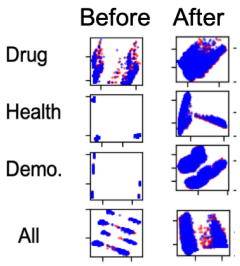


Figure 1: **Preprocessing.** 2 components PCA scatter for different subsets before and after preprocessing.

Testing Hypotheses. We further partition our 'substance abuse' dataset into 2 subsets according to frequency of usage among respondents. A drug that less than 10 % of all respondents have experienced are designated 'rare drugs' and the others are 'common drugs.' Our sets of demographic factors, common substances, and rare drugs consist of 131, 92, and 89 numeric features that contain responses to survey questions such as the first age exposed to substance, frequency of substance use, years at school, or household income. We apply linear and nonlinear regression models to predict responses on rare drugs using features from common drugs and demographics. Specifically, we use ordinary least squares (OLS), lasso, ridge, elastic net (EN), partial least squares (PLS), and random forest (RF) regression models from the SciKitLearn Python libraries [7]. We perform cross validation to find hyperparameters with the best mean squared error.

The Effect Problem we address is how well we can predict the future potential mental health disorders of an individual, given the drug use data? This is performed by first sub-partitioning the data in three ways, first, by train-test splitting in ratio 4:1, second, by a 5-fold cross validation, and third, by a 10-fold cross validation. Several regressor models mentioned below have been applied for each of these three sets of data.

Hyperparameter Tuning. As hyper-parameters(α) are not directly learnt within the estimators, they are passed as arguments to the constructor of the estimator classes. While Linear Regression is a baseline regressor, the hyperparameters for other regressors have been tuned using GridSearchCV, which exhaustively generates candidates from a grid of parameter values specified. Our grid consisted values in {0.001, 0.1, 1, 5, 10, 100, 500, 1000, 1500, 2000, 6000, 10000}. Because of the widely

random and sparse characteristics of the dataset, our models did not converge for lower values of α's, and hence raises a question whether the regression coefficients might be misleading when the errors are not normally distributed. Hence, we investigate each regressor by further bootstrapping it.

The Bootstrapping Approach. Bootstrapping is a general approach to statistical inference based on building a sampling distribution for a statistic by re-sampling from the data at hand. The key bootstrap analogy is as "The user dataset is to the sample as sample is to the bootstrap samples". Bootstrapping is helpful to calculate confidence intervals around predictions required in critical analysis. We performed the following regular bootstrap method.

- Start with dataset D = (X_1,Y_1),(X_2,Y_2),...,(X_n,Y_n).
- Generate (K=1000) bootstrap samples D_1, D_2, ..., D_k, each sample has (n=10000) obsevation pairs randomly sampled from D with replacement.
- Build a model on each bootstrap sample D_i and test using the original dataset D.
- Construct a confidence interval based on the above observations.

The Regressor Models. A brief description of the regressor models used in this project:

- **OLS** Simple linear model that minimizes the mean squared error.
- **Lasso** Adds L_1 penalty to loss function of OLS, resulting in a sparse solution.
- **Ridge** Adds L_2 penalty to loss function of OLS, forcing the solution to be close to 0.
- **EN** Combines both L_1 and L_2. Resulted coefficients for correlated predictors are similar values.
- **PLS** Fits regression after projecting predictor and outcome variables to a reduced dimensionality.
- **RF** An ensemble method that builds multiple decision trees and combines their results.

Evaluation Strategy. For the first task, due to high dimensionality as well as our assumption on low dimensional latent structure of the dataset, we expect that partial least squares regression which finds regression in reduced dimensional space will perform the best. We also anticipate that regularization will be effective since the original dataset is highly sparse. We will evaluate each model using standard metrics for regression mean squared error (MSE) and coefficient of determination (R^2) as well as runtime. We will assess importance of each feature according to their regression coefficients, and consider whether there are specific features and their values that indicate usage of rare substances.

Unlike the first task above, in the second task, the comparison metric of Mean Absolute Error (MAE) is used because few features poorly correlate due to which the squared function of Mean Square Error (MSE) results ramping up of more error. Hence, we opt to investigate MAE for the second task and coefficient of determination (R^2). For the bootstrapping approach, we evaluate the confidence of R^2 by calculating the mean and standard deviation of the bootstrap samples.

2.3 Unsupervised Learning

2.3.1 Latent Variable Models

We use five different latent variable methods from the SciKitLearn Python libraries [7]. All parameters used are the default unless specified.

- **PCA :** transforms a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables. We discuss further below as a main model.
- **LDA :** maps all the documents to the topics in a way, such that the words in documents are mostly captured by those topics. We found 7-component provides the highest average log-likelihood.
- **NMF :** decomposes the multivariate data via factorizing matrix V into two matrices W and H where all three matrices have no negative values. We select the number of component to 8 based on the observation that it achieves the lowest reconstruction error.
- **FA :** describes variability among observed variables in terms of a potentially lower number of unobserved variables called factors. 5-component is used based on the average log-likelihood.
- **GMM :** is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. Using the average log-likelihood as a criterion we select the optimal number of components to 9.
- **t-SNE :** t-Distributed Stochastic Neighbor Embedding (t-SNE) [6,11] is a nonlinear dimensionality reduction method that converts high dimensional data into low dimensional (2 or 3) representation in a way that similarity between pairs of datapoints are preserved. Compared to PCA, t-SNE has

a non-convex objective function. The objective function is minimized using a gradient descent optimization that is initiated randomly.

PCA with K-means clustering : PCA uses an orthogonal transformation to convert a set of observations of correlated variables into a set of values of linearly uncorrelated variables called principal components (PCs).

We hypothesize that the PCA with K-means clustering fits to our application due to following reasons: (1) Our survey consists a high dimensional dataset (approx. 3k). As high dimensionality makes the distance metrics meaningless and data points are more sparse with relatively more uniform distances, clustering would work better after dimension reduction (sparsity of our data further increased after performing one-hot encoding); (2) Previous research has proved that the principal components are the continuous solutions to the discrete cluster membership indicators for K -means clustering [1].

2.3.2 t-SNE-based Feature Analysis

We use t-SNE as another method of pattern analysis and data visualization since t-SNE follows a different type of computational method than other manifold learning methods, which were derived from PCA. Unlike other latent models, t-SNE does not include built-in feature importance function, so we introduced our own algorithm to evaluate the importance of features. After t-SNE visualization, we apply K-Means clustering to t-SNE model with C, the number of clusters, in order to identify centroids of clusters and then K-Nearest Neighbors to select the 50 closest samples to each cluster's centroid. Then, we vertically add up all features across all samples and extract the features that summation is over 45 as important features for each cluster (recall that we apply one-hot encoding). Our generalized algorithm is shown in Algorithm 1.

Algorithm 1 Pseudocode for t-SNE-based Analysis

1: **for** c in C clusters obtained from t-SNE **do**
2: Apply K-Means ($K = C$) clustering to identify centroid point c_p.
3: Apply K-Nearest Neighbors ($K = P$) to find the P closest samples to c_p.
4: Add up P samples' features vertically: $F^{tot}_{1,\ldots,N} = F^{1}_{1,\ldots,N} + \cdots + F^{P}_{1,\ldots,N}$.
5: **Sort** $F^{tot}_{1,\ldots,N}$ and **extract features** that satisfy $F^{tot}_{n} \geq P \cdot r$ where $0 \leq r \leq 1$ (close to 1).
6: **end for**

3 Results

3.1 Supervised learning

The Cause Problem. We note that while ridge regression performed than others, regression models for predicting rare substance use from either sets of predictor features did not perform well overall. We note that runtimes for all models were below less than an hour and therefore considered practically efficient. Interestingly, performances for predicting rare substance use are better when using common substances than demographic features. Combining the two sets of predictive features did not improve prediction performance. Our results call for additional tests to see whether high dimensionality of outcome variables has affected our model performances. We reason that since PLS that fits regression in the reduced dimensions did not work well, outcome features have wide variance with small degree of correlation within themselves. This calls for future work that further subdivides outcome variables and separately predicts use of each rare substance.

We assessed importance of each feature according to results from random forest regression and found top features from each set of predictor variables. Interestingly, features with the highest importance from each set are 'COCEVER', 'COCREC', and 'COCAGE', and 'AGE2' and 'SEXRACE.' The 3 features from common drug predictors are all related to use of cocaine. We note that 12% of all respondents have experienced cocaine at least once. These respondents also correspond to individuals who haev also experienced rare drug substances such as heroin or LSD that 2% and 8% of respondents have used. We observe in the dataset that the 12% of respondents with cocaine use also responded yes to alcohol and cigarette use. Also, the top demographic features of age and sex and race are unexpected. The fact that age, sex, and race of respondents are more highly predictive of rare substance use than other demographic factors such as income or educational status is highly

	Test1		Test2	
	MSE	R^2	MSE	R^2
LASSO	0.030	0.056	0.029	0.056
RIDGE	**0.023**	**0.149**	**0.028**	**0.073**
EN	0.029	0.056	0.029	0.056
PLS	0.025	0.112	0.028	0.066
RF	0.024	0.118	0.029	0.063

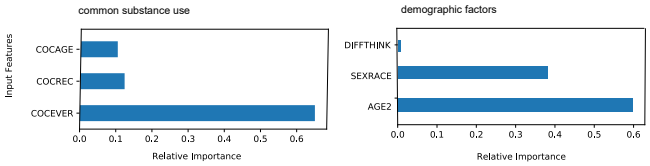


Figure 2: Left is the performance of models for predicting rare substance use from common substances (Test1) and demographic factors (Test2) and right is the importance of features from two predictor sets.

interesting and calls for additional tests to test correlation within demographic factors for potential confounding effects. We conclude that despite low predictive power of regression models, we have discovered top features that are reasonable and interesting.

The Effect Problem. Likewise the cause problem above, ridge regressor performs relatively better than others for the effect problem. However the performance, though not terrible, is fairly acceptable. This is due to the high dimensionality and sparsely distributed data in feature space. Regarding runtimes, when the hyperparameters tuning (mentioned in §4.2) choice was low, the runtimes were long (1.5 hrs - 2 hrs) to output that the models are not converging. While on the other hand, when a converging value of hyperparameters is selected through GridSearchCV, the runtimes were found less than 30 minutes for all the regressors combined. Figure 3 *(Left)* below depicts the regular train-test split results along with evaluation results for k-fold cross validation techniques for k=5 and k=10 for each of the regressors. The purpose of cross validation is clearly not effective here as we observe that cross validating the training set in more folds doesn't follow an increase in the precision of model prediction. Moderate folds like 5 or 10 support a decent curve fitting of training model. However, on a side note, here we also note that k-fold techniques can impose difficulties when it is applied to a time data series as the order of the series is important.

	train-test-split		5-Fold CV		10-Fold CV	
	MAE	R^2	MAE	R^2	MAE	R^2
RIDGE	**0.614**	**0.158**	**0.626**	**0.156**	**0.628**	**0.156**
LINEAR	0.615	0.147	0.622	0.137	0.613	0.146
ELAS_NET	0.873	0.134	0.879	0.067	0.877	0.066
LASSO	1.015	0.004	1.018	0.0019	1.023	0.0037

	R^2 (mean)	R^2 (std)
RIDGE	0.0141	0.2635
LINEAR	Poor	Poor
ELAS_NET	**0.0665**	**0.00136**
LASSO	0.0388	0.00029

Figure 3: ***(Left)*** Performance of models for predicting mental illness disorders raised due to drug consumption. Ridge performs better while Lasso being relatively worst. ***(Right)*** Bootstrap Performance of models for predicting mental illness raised due to drug consumption. Elastic Net being more stable with relatively decent mean, Ridge samples deviate relatively more.

The Bootstrap results are depicted in Figure 3 *(Right)*. Interestingly, bootstrap results now clearly reflect why the cross validation is not effective in Table 2 results. Ridge regressor was better earlier, however upon investigating bootstrap, it is found that Elastic Net regressor is more stable compared to others, while samples of the other regressors deviate a lot from their mean. This reason clearly explains why cross validation doesn't perform as usual as the validation samples are deviating from their mean. Linear Regression we conclude as completely out of confidence ("*Poor*") as their R^2 mean and standard deviation are in the range of 10's which is not in an acceptable range of deviation.

The results of the Effect problem reflect that 18.9% of adults exposed to substances in the U.S. had experienced some sort of mental illness (MI), while 4.5% experienced serious mental illness (SMI) in 2017. Anti-social behavioural trait is the commonly observed among substance users. Specifically, *IMPSOCM*-difficulty participating in social acts for 1mo/12mo is the relatively closely predicted feature among hard-drug consumers.

3.2 Unsupervised learning

3.2.1 Latent Variable Model Performance

Model performance comparison: Here, we compare the performance of the models described in Sec. 2.3.1. Some models are not able to generate the RE and/or ALL and therefore written as NA.

model	train data			test data	
	time	RE	ALL	RE	ALL
PCA	**4.36**	**0.0323**	**0.0136**	**0.0323**	**0.0546**
LDA	510.2	NA	-2246	NA	-2253
FA	52.62	NA	0.0076	NA	-12071
GMM	2297	NA	0.2157	NA	-0.3176
NMF	66.80	0.0346	NA	0.0346	NA

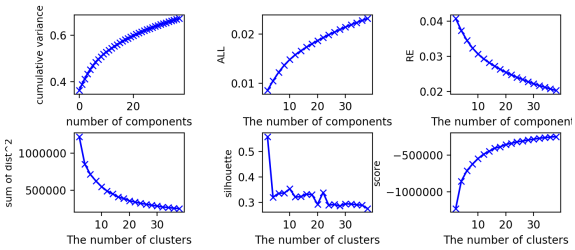


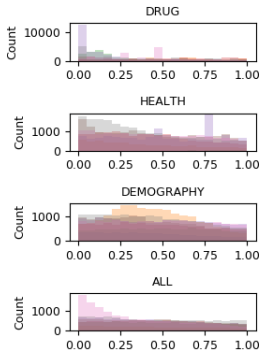
Figure 4: **Model performance comparison and hyperparameter tuning for PCA (top) and K-means (bottom).** RE indicates the reconstruction error, ALL denotes the average log-likelihood. The score is the opposite of the data value on the K-means objective, and silhouette is a measure of how similar a data point is to its own cluster compared to other clusters.

Table.1 shows that both PCA has the lowest RE, second highest ALL, and the least training time. For the ALL, GMM has the largest value with LDA being lowest. However, GMM takes very long time to train. Thus, PCA is used as our main model. For test data, ALL of PCA is still higher than that of LDA and FA, signifying that PCA is more likely a better dimensionality reduction model than LDA and FA for our observed data. It must be noted that the measures with test data were not considered when selecting the model.

Hyperparameter optimization: Selecting the number of component is important for preserving the essence of the original data and discovering new patterns. The number of components for PCA is set to 8 based on the elbow on the arm of varying cumulative explained variance ratio, ALL, and RE. In particular, the cumulative variance ratio reached over 0.5 when the number was 8.

For K-means clustering model, we tuned the number of cluster using the sum of squared distance, silhouette, and score. By definition, high silh. value indicates that the data is well matched to its own cluster and poorly matched to neighboring clusters. K-means score measures how far the points are from the centroids and thus, the model performs better when the score is closer to zero. The number of clusters for K-means is set to 10 as the elbow is approximately at 10. For both the number of components and the number of clusters, we used train data for tuning.

3.2.2 Pattern analysis



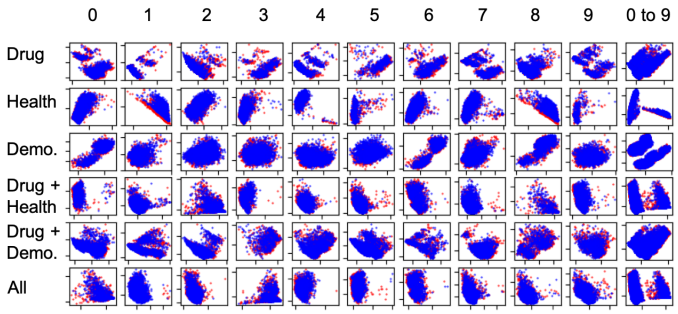
Data	1	2	3	4	5	6	7	8
Drug	CIGEVER_1 MJEVER_1 CIGYFU_9	CIGEVER_2 CIGYFU_91 CIGTPE_91	NA	NA	STMANYREC_1 STMNMES_93	PNRNMLIF_2 ALCEVER_1	TRQANYLIF_5 TRQANYREC_1 PNRANYLIF_2	PNRANYLIF_1 PNRANYREC_2 CIGTPE_5
Health	YIAIL_2 YMHASPTX_2 PNRNMYFU_91 PNRANYLIF_1 TRQANYLIF_97	ADDPREV_1 ADDPDISC_1 INHALEVER_91 INHALREC_91 INHALES_1	HIVAIDSEV_2 KIDNYDSEV_2 CIGMEN_1 CIGMEN_97	**DSTNRV_2** **DSTRST_2** **CIGEVER_2** **ALCEVER_2** **TOBEVER_2**	**DSTEFF_1** **DSTEFF_1** **CIGMEN_1** **ALCEVER_1** **TOBEVER_1**	AMHRX_1 AMHTXRC_1 CIGAREVR_1	IMPWORK_1 IMPSOC_1 IMPCONCN_1 LGAS_1 FELTMARKR_2	IMPWORK_1 IMPCONCN_1 LGAS_1
Demo.	NA	NA	IRPRVHLT_2 WRKDPSTWK_2 CRKEVER_91 CRAKREC_9	IRPRVHLT_2 CRKEVER_91 CRAKREC_9	PREGAGE_4 MJEVER_2 MJEVER_97	IRSEX_2 SEXRACE_2 MJEVER_97	IRSEX_1 SEXRACE_1 CIG30MEN_1	NA
All	12-17 AGE_1	MJEVER_1 HALLUCEVR_1 PEYOTE_2	18-25 AGE_1	WRKHADJOB_2 WRKDPSTWK_2	ADDPREV_1 AMIYRU_1 ADDPDISC_1	CIGYFU_91 CIG30MEN_91 CIG30MLN_91	ETHER_2 INHALEVER_1 INHALOTH_2	**MJEVER_1** **WRKDPSTWK_2** **WRKHADJOB_2**

Figure 5: **Feature analysis using PCA Components.** Left figure is user proportion from latent user 1 to 8 for drug, health, demographic, and all dataset. Right table shows the feature names with the highest pseudocount for each component

Analysis using PCA components: The study participants are mixture of the latent participants. Left of Fig. 5 depicts the participant proportions from latent participants 1 to 8. The interesting observation here is that the pseudo-count for drug dataset is much higher than that of other dataset. Also, certain latent participant is highly unique in terms of count compared to the rest latent participants in the drug dataset. This is because the drug dataset contains a lot of overlapping questions with high ratio of those who refused to answer the questions. Moreover, while the participant proportion of drug, health, and all dataset resemble each other, the participant proportion of socio-demographic data

doesn't have a peak near 0. We presume the reason is that the number of demographic dataset is relatively less than the other dataset.

Right table shows the top features with the highest pseudocount for each component and dataset. Features are written in the form of acronyms. Intuitively, the feature name for drug is formatted as the combination of the drug acronyms, question type, and participant's answer [1]. It could be simply interpreted by drug acronym [2] and the answer (i.e. CIGEVER_2 and CIG30MEN_2 are equivalent to not use CIG). For health and demographic data, we extract not only the health or demographic features with the highest pseudo-count but also the drug features with the highest pseudo-count in the original dataset belonging to the participants in the corresponding component. By doing so, we attempt to find the correlation of substance use and social/health factors. Thus, for health [3] and demographic data [4], the first half row represents its own features while the second half row shows the drug features. Overall, the drug dataset is grouped into people who use CIG and MJ, people not using CIG and MJ, people using STIM, people who drinks ALC, and those who use PRN and not use PRN. The health dataset is partitioned into participants who didn't utilize the youth mental services, have adult depression, do not have physical illness, not have mental disorders, have mental disorder, have received mental health treatment, and those with no difficulties in daily routines. Interestingly, we found that those who have mental disorder are likely to use CIG, ALC, and TOB and vise versa. However, while the demographic dataset is largely divided by gender and the presence of health insurance and job, we had difficulty finding its relationship with the drug use. For all dataset, we are able to find a slight relationship between the employment status and the use of MJ. People abusing MJ are likely to be unemployed.



Data	Silhouette
Drug	0.4502
Health	0.3528
Demo.	0.3170
Drug+Health	0.4081
Drug+Demo.	0.3416
All	0.2955
Male	0.3375
Female	0.3760

Figure 6: **Feature analysis using K-means clusters.** The blue indicates female, and the red represents male. 10th column combined the cluster $0-9$. Ticks are not presented for the visualization purpose.

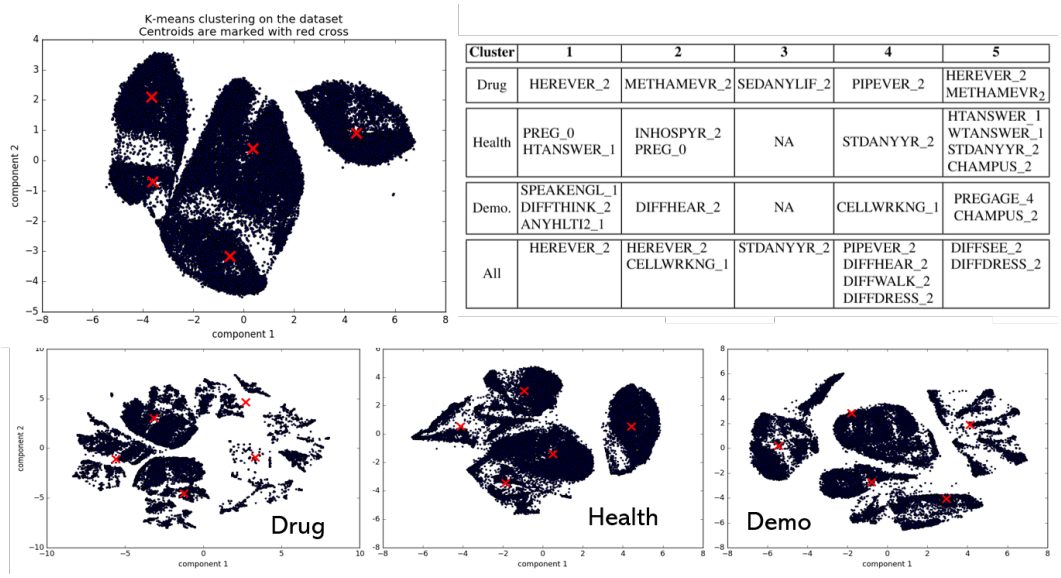
Analysis using K-means clusters: We further used K-means clusters for feature pattern analysis shown in Fig. 6. First 10 columns of X-axis represent cluster 0 to 9 and 11th column illustrates the clusters combined from 0 to 9. Y-axis represent different sets of data. The blue points indicates female while the red points represents male. First observation is that the male subjects are more widely spread out. Correspondingly, the silhouette value of female clusters is higher than that of male clusters. However, the percentage of male (48.04%) is lower than that of female (51.96%). Thus, we cannot presume that the dispersion of male data points are due to the large number but assume that there are more various answers from the male participants. Second, as mentioned above, the clusters with drug data tend to be more separated than the clusters of other dataset because the data is largely divided into those who answered and those refused to answer the questions. Third, the shape of the clusters with all dataset resemble that of drug and health combined data while the form of drug clusters is similar to that of drug and demo. combined data. This denotes that demographic data is less influential, and the possible reason could simply be its small number of features compared to that of drug and health data. Likewise, it must be noted that the silhouette measure of dataset containing demographic data is lower than the silhouette value of dataset not

[1] 1 – use/male, 2 – not use/female, 5 – sometimes, 9 – skip, 91 – never use, 97 – refused

[2] CIG – cigarette, CRK – crack MJ – marijuana, STIM – stimulant, PNR – pain reliever, ALC – alcohol, TOB – tobacco, DRG – drug, HALLUC – hallucination, INHAL – . inhalant, OXC - oxycontin

[3] Y – youth service, ADD – depression, DST – mental disorder, AMH- mental treatment, IMP - daily work

[4] IRP - health insurance, WRK - employment status, PREGAGE - pregnancy age, IRSEX - gender



Cluster	1	2	3	4	5
Drug	HEREVER_2	METHAMEVR_2	SEDANYLIF_2	PIPEVER_2	HEREVER_2 METHAMEVR$_2$
Health	PREG_0 HTANSWER_1	INHOSPYR_2 PREG_0	NA	STDANYYR_2	HTANSWER_1 WTANSWER_1 STDANYYR_2 CHAMPUS_2
Demo.	SPEAKENGL_1 DIFFTHINK_2 ANYHLTI2_1	DIFFHEAR_2	NA	CELLWRKNG_1	PREGAGE_4 CHAMPUS_2
All	HEREVER_2	HEREVER_2 CELLWRKNG_1	STDANYYR_2	PIPEVER_2 DIFFHEAR_2 DIFFWALK_2 DIFFDRESS_2	DIFFSEE_2 DIFFDRESS_2

Figure 7: **t-SNE Analysis.** Figures are t-SNE (component 2) scatter plots of dataset (upper) and three subsets (lower). We assume five separate clusters and use K-means clustering to obtain a centroid of each cluster (red marks). Table shows important features for each cluster based on our own algorithm in Section 2.3.2. Cluster indices are assigned from upper left to upper right counter-clock-wise.

containing the demographic data. On the other hand, the drug data has the highest silhouette. As high silh. value indicates that the object is more similar to its own cluster (cohesion) compared to other clusters (separation), demographic data is less separable.

Analysis using t-SNE: It is observed that the dataset and three subset bring out several separate clusters shown in Figure 7. We assume there are five separate clusters and apply our algorithm to see important features and any correlates between dataset and subsets (also between t-SNE and other models). The most important features of each contriod among 50 nearest neighbors from the dataset are experience of Heroin (cluster 1, 2), sexual disease (cluster 3), pipe tobacco (cluster 4), and difficulty of seeing (cluster 5). It seems that dataset's cluster 1 is affected by drug data, cluster 2 by drug + demo data, cluster 3 by health data, cluster 4 by drug + demo data, and cluster 5 by demo. We observe that the most important features of cluster 1,2,4 are from drug data despite the fact that drug data is sparse compared to other subsets. While previous models present more correlation between drug and health data, in t-SNE demographic data shows more correlation to drug data.

4 Discussion and Conclusion

Our overarching goal was to determine whether there exists a pattern between substance abuse and demographic or medical information obtained from a large scale collective survey dataset. We divided the tasks into two subproblems of prediction and latent structure identification and applied supervised and unsupervised learning to address each. We were able to identify reasonable sets of important, interesting features such as race and cocaine use that predict rare drug abuse. Our results call for future work that would build predictive models for each subset of outcome variables independently as we detected wide variance and high sparsity in the high dimensional outcome features that we were unable to address via dimensionality reduction. With unsupervised learning, we identified such stronger correlation between drug and health compared to drug and demographic information. We assume that, unlike supervised learning, the correlation is present in the latent structure and thus stronger in the reduced dimensions. It would also be interesting to develop the imputation methods even further, possibly extending with fuzzy k-means imputation, or with multiple imputation using a hotdeck or carry-forward method. In addition, it would also be interesting to explore more opportunities with the non-binary variables, since the highly peculiar bias likely leaves the way open for alternative methodologies.

References

[1] Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29. ACM, 2004.

[2] https://www.samhsa.gov/data/sites/default/files/NSDUH-FFR1-2016/NSDUH-FFR1-2016.pdf.

[3] National Survey on Drug Use and Health Webpage, https://datafiles.samhsa.gov/study-dataset/national-survey-drug-use-and-health-2017-nsduh-2017-ds0001-nid17939.

[4] NSDUH. Website, https://datafiles.samhsa.gov/.

[5] Brian V Fix, Richard J O'Connor, Lisa Vogl, Danielle Smith, Maansi Bansal-Travers, Kevin P Conway, Bridget Ambrose, Ling Yang, and Andrew Hyland. Patterns and correlates of poly-tobacco use in the united states over a decade: Nsduh 2002–2011. *Addictive behaviors*, 39(4):768–781, 2014.

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[7] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[8] Heather Ryan, Angela Trosclair, and Joe Gfroerer. Adult current smoking: differences in definitions and prevalence estimates—nhis and nsduh, 2008. *Journal of environmental and public health*, 2012, 2012.

[9] Christopher P Salas-Wright, Michael G Vaughn, Jenny Ugalde, and Jelena Todic. Substance use and teen pregnancy in the united states: evidence from the nsduh 2002–2012. *Addictive behaviors*, 45:218–225, 2015.

[10] Nora D Volkow, Thomas R Frieden, Pamela S Hyde, and Stephen S Cha. Medication-assisted therapies—tackling the opioid-overdose epidemic. *New England Journal of Medicine*, 370(22):2063–2066, 2014.

[11] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 1(10):e2, 2016.