Mental Health Group

Final Presentation

Auliya Fitri
Honglin Ju
Raghav Sharma
Shivendra Singh

Outline

1. Introduction
2. Descriptive Mining
3. Predictive Mining
4. Conclusion and Lessons Learned

Notes

Flow Suggestion (To be discussed):

- Brief introduction in figure (rough statistic of mental health patient)
- Explain our dataset (columns and highlighted features)
- Motivation (mention our prediction result as suggestion to mental health services, for example. This mental health services will be our potential customer. Other recommendation are welcomed)
- Go with the scenario, the virtual characters, to explain the data in a more humane way

Introduction

Mental health: [noun]

“A person’s condition with regard to their psychological and emotional well-being”

--oxford dictionary

Mental Health Facts
IN AMERICA

Consequences

10.2m

Approximately 10.2 million adults have **co-occuring** mental health and addiction disorders.[1]

26%

Approximately 26% of **homeless** adults staying in shelters live with serious mental illness.[1]

Impact

1st

Depression is the leading cause of disability worldwide, and is a major contributor to the global burden of disease.[1]

-$193b

Serious mental illness costs America $193.2 billion in lost earning every year.[3]

1 in 100 (2.4 million) American adults live with schizophrenia.[1]

2.6% (6.1 million) of American adults live with bipolar disorder.[1]

6.9% (16 million) of American adults live with major depression.[1]

18.1% (42 million) of American adults live with anxiety disorders.[1]

SAMHDA

Substance Abuse & Mental Health Data Archive

Treatment Episode Data Set Admission

contains

demographic and substance abuse characteristics
of admissions to alcohol or drug treatment in facilities

From 2010 - 2014
~6,5M rows
65 features

… and cross study from …

National Mental
Health Services

+

United States™
Census
Bureau

for economic data per metro area

Treatment Episode Data Set Admission

YEAR	AGE	GENDER	RACE	ETHNIC	STFIPS	DAYWAIT	SUB1	ROUTE1	FREQ1	FRST USE1	SUB2	DSMCRIT	PSYPROB
2010	9	2	5	5	29	1	2	1	5	3	4	4	1
2012	3	1	21	2	29	3	4	2	2	2	2	7	2
2014	6	2	4	5	31	0	4	2	5	3	2	16	2
2012	6	2	5	5	8	0	2	1	4	2	7	4	1
Year admitted	Patient's background				US States	How long patient have to wait	Patient's primary substance problem: Which substance, how and how often they consume it, when they first use it				Secondary Substance	Mental Disorder Diagnosis	Has Psychiatric Problem? 1 for Yes

Why this dataset?

In some states the health services are not that good

To find the key factors which could improve the situation

Mental health illness is hard to be noticed thus get treated late

To find out what are the strongest predictors of mental health illness

The descriptive analysis could be useful to health related organisations including hospitals and insurance companies and US government

To enable clinicians to tailor treatment on meaningful medical indicators

Virtual Characters

To Build a story around our dataset we introduce virtual characters and mould them with our analysis.

1. Alice is 28 years old, lives in the state of Nebraska
2. Bob is also 28 years old comes from Colorado
3. Carol is 16 years old comes from Missouri
4. Denis is 30 years old comes from Delaware ← HEROIN ADDICT

We want to see the where they will go and what will happen next using the technology ...

This story was taking place at 2012.
Alice and Bob are 28 years old at that time, meaning that he belongs to the majority group of patient based on their age group.

Bob comes from Colorado, where the number of admitted patients is the fifth highest for these 5 years.

We can see how long they have to wait in order to be treated. Figure shows that patients from both Colorado and Nebraska have waiting time 2 days in average. It means that Alice and Bob have to wait for 2 days before getting first treatment whereas Carol have to wait longer (8 days) to be treated in Missouri.

Virtual Characters

			
Alice	Bob	Carol	Denis
28 yo	28 yo	16 yo	30 yo
Nebraska	Colorado	Missouri	Delaware

Reported primary substance

Marijuana	Alcohol	Marijuana	Heroin

Technology Stack

SciPy

NumPy

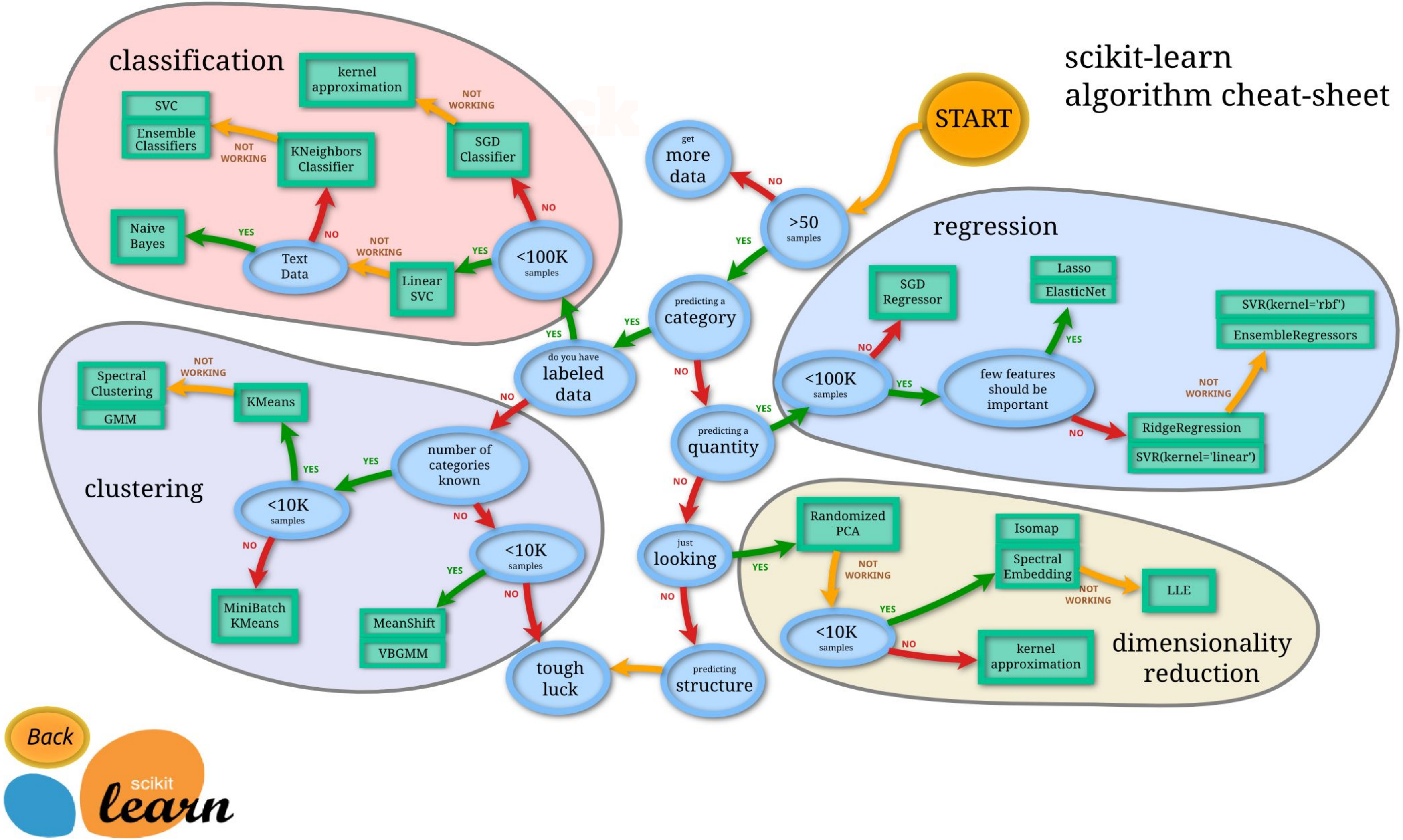
matplotlib

IP[y]: IPython
Interactive Computing

pandas

scikit learn

Google Cloud

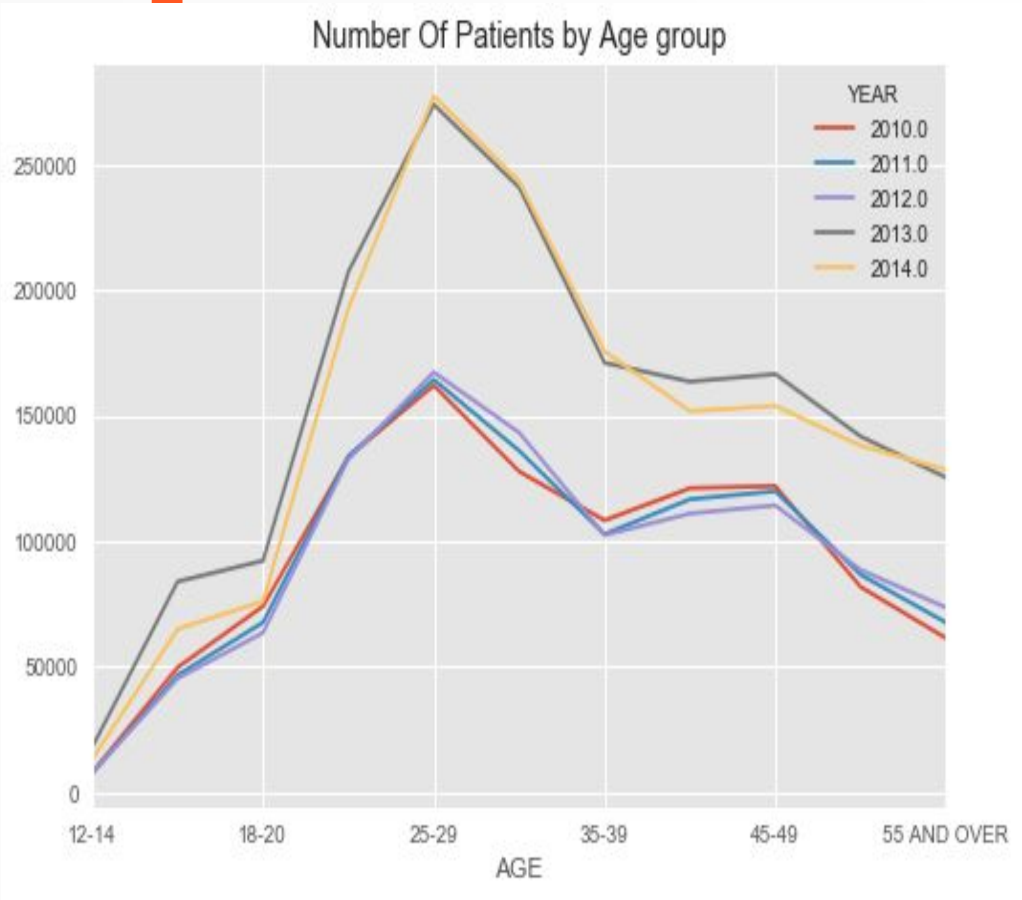

scikit-learn
algorithm cheat-sheet
START
>50 samples
get more data
predicting a category
do you have labeled data
<100K samples
Linear SVC
Text Data
Naive Bayes
KNeighbors Classifier
SVC
Ensemble Classifiers
SGD Classifier
kernel approximation
classification
predicting a quantity
regression
SGD Regressor
few features should be important
Lasso
ElasticNet
RidgeRegression
SVR(kernel='linear')
SVR(kernel='rbf')
EnsembleRegressors
clustering
number of categories known
<10K samples
KMeans
Spectral Clustering
GMM
MiniBatch KMeans
MeanShift
VBGMM
tough luck
just looking
predicting structure
Randomized PCA
Isomap
Spectral Embedding
LLE
kernel approximation
dimensionality reduction
YES
NO
NOT WORKING
Back
scikit
learn

Descriptive Mining

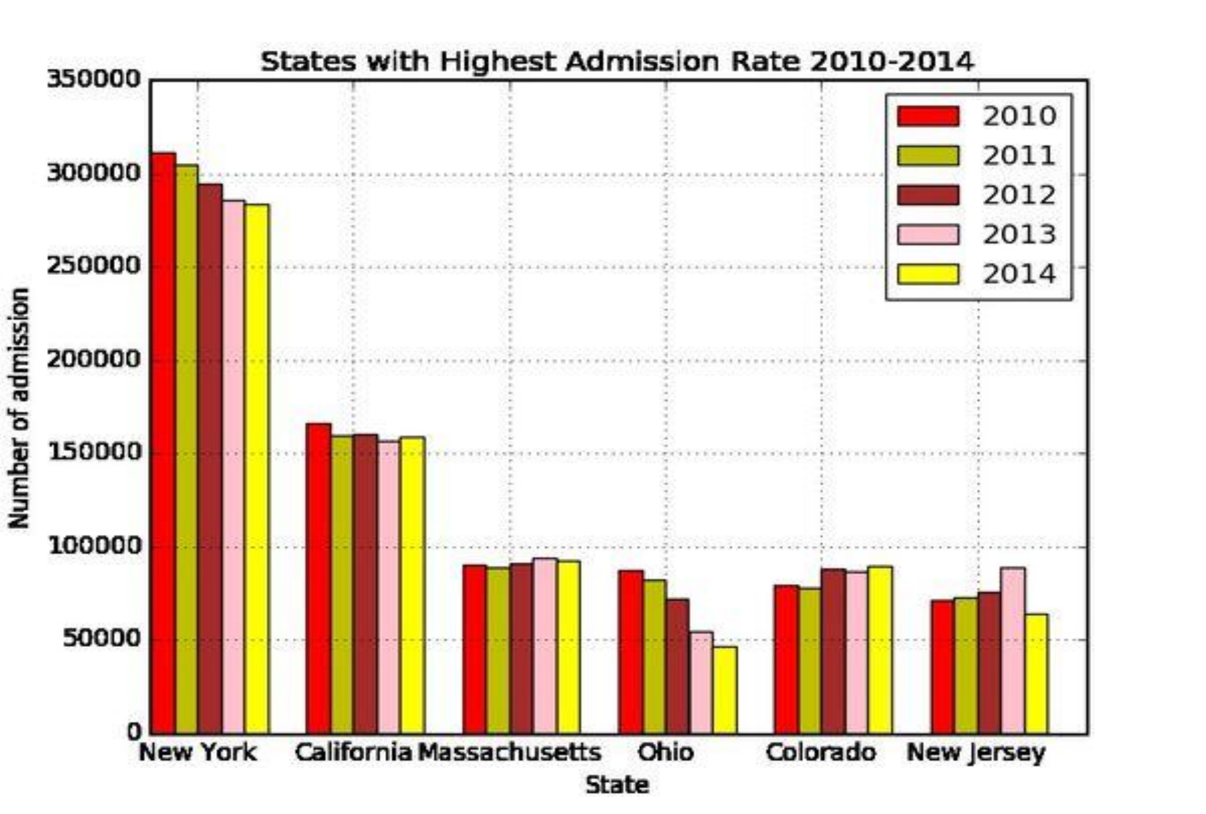
Patients and Age Group

Patients Trend by Age Group over the years.

Highest Number of Patients are in the age group from 25-29.



Admission Rate

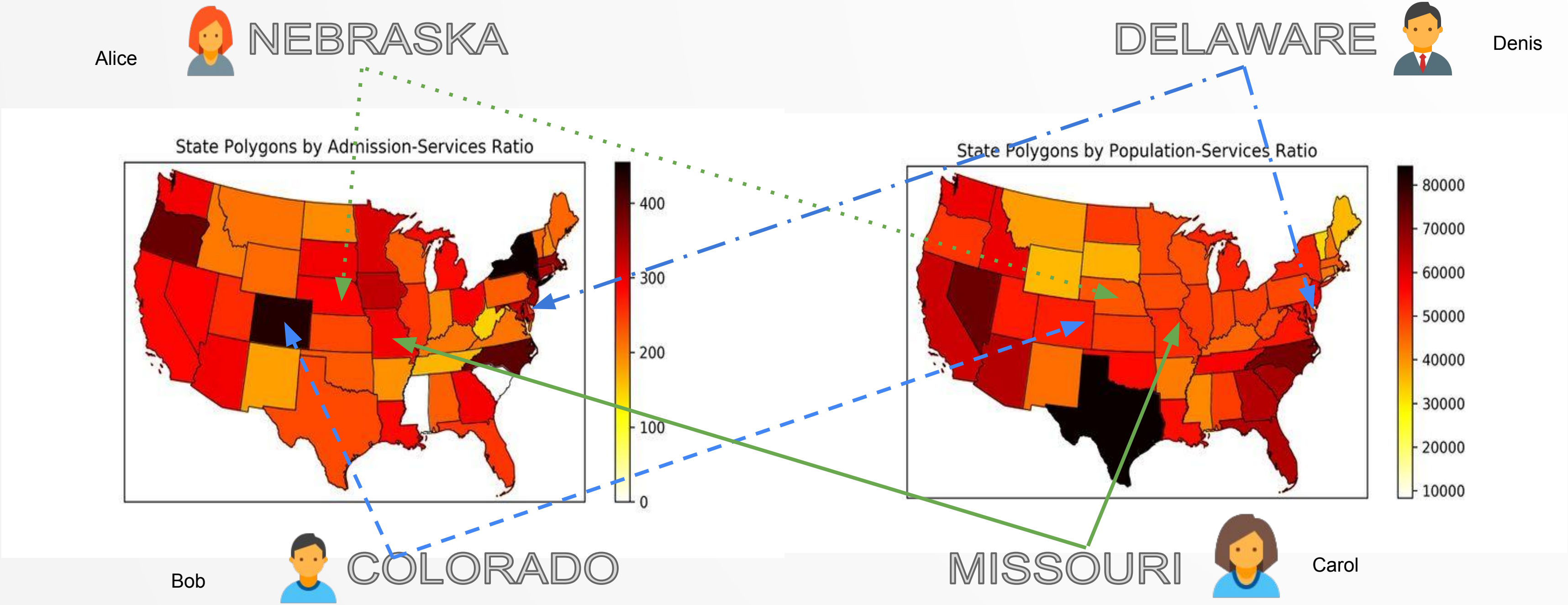


Inferences:

- Number of admission does not differ much
- Some states has decreasing number.

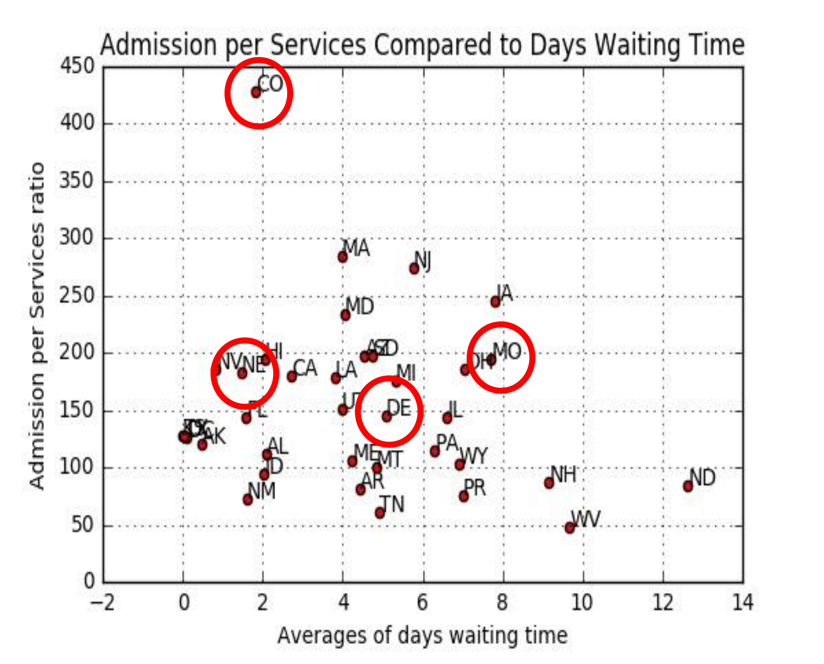
Bob comes from Colorado, where the number of admitted patients is the fifth highest for these 5 years.

Mental Health Services (Hospitals) Density Map



- Density map derived from number of admitted patients per mental health services state-wise (left) and population per services (right).
- Admission-Services Ratio : New York, Colorado, Oregon, North Carolina and Massachusetts
- Population-Services Ratio : Texas, Nevada, North Carolina, South Carolina and Florida
- Darker color means number of patients and population is high while number of services is low

Days Waiting Time

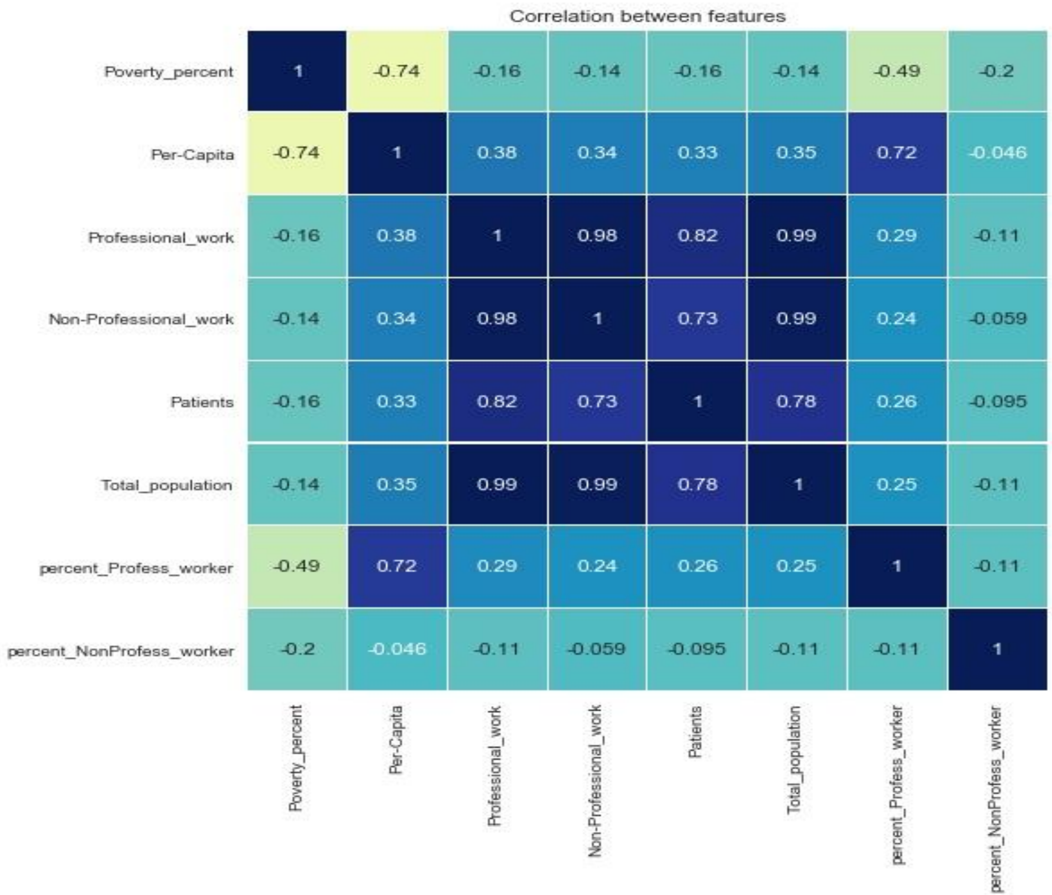


- Average of days waiting time plotted with density of admission per mental health services.
- The expected good value is it has low days waiting time and also not too dense.
- Most of the plots have proportional number, meaning that if the admission per services density is high, the average waiting time is also high.
- Colorado has many alcoholic patients, it might be the reason why days waiting time is low
- Patients from both Colorado and Nebraska have waiting time 2 days in average.
- It means that Alice and Bob have to wait for 2 days before getting first treatment.
- Carol has to wait longer (8 days) to be treated in Missouri.
- Denis has to wait 4-6 days.

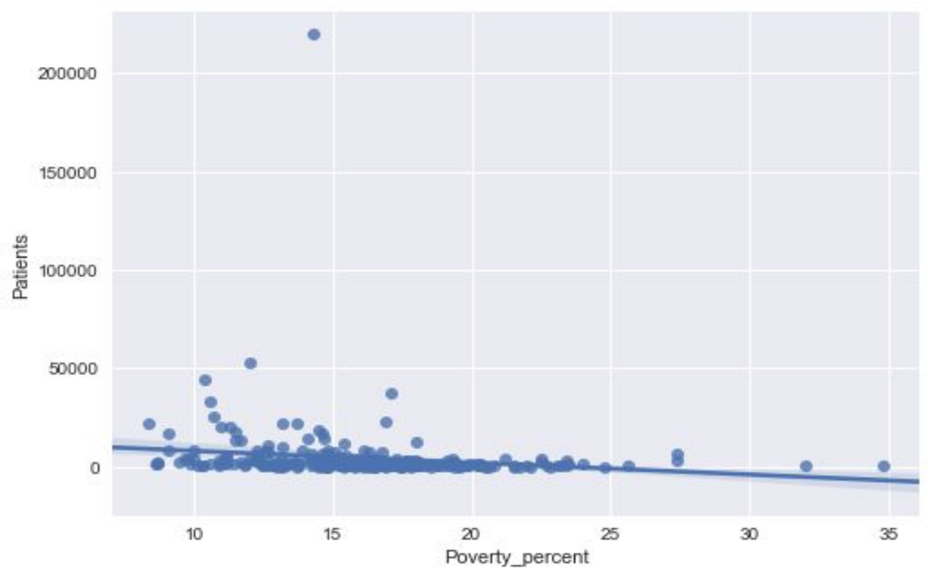
Population Analysis with Census Data

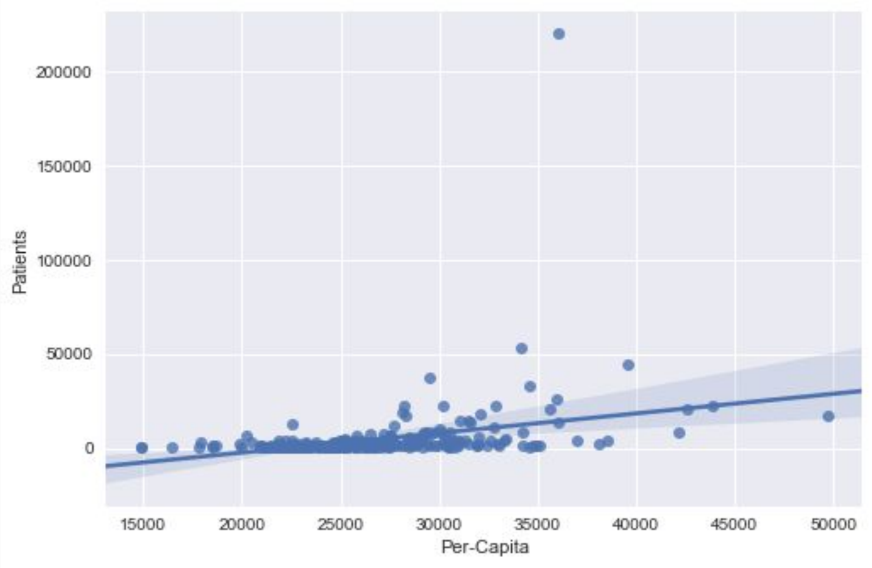
We got data about Core Based Statistical Area which is a collective term for metro area and micro area (where each micro or metro area consist of counties) from US census and employment data and then try to plot some graphs to see variations and relationship with Mental Health data.

Number of patients have a strong direct correlation with Total Population , Professional workers, Non-Professional workers which is obvious as number of people are increasing.



Poverty and Per Capita effect





Exception is New York-Newark-Jersey City Metro Area as it has the highest population of 20.1 million .

- As the poverty Percent increases the number of patients are decreasing but the correlation is very weak.
- The Metro areas with higher Per-Capita has a higher number of patients.

Top 10 metro areas with lowest patients

Poverty_percent	Per-Capita	Patients	Total_population	Prop	Metro_Areas
17.5	26639	73	814805	0.000090	Baton Rouge, LA Metro Area
18.5	21157	59	617323	0.000096	Lakeland-Winter Haven, FL Metro Area
17.8	25781	86	475457	0.000181	Lafayette, LA Metro Area
16.5	24469	39	209402	0.000186	Houma-Thibodaux, LA Metro Area
13.2	30813	149	722784	0.000206	North Port-Sarasota-Bradenton, FL Metro Area
19.4	24833	107	445305	0.000240	Shreveport-Bossier City, LA Metro Area
15.4	25199	119	462339	0.000257	Pensacola-Ferry Pass-Brent, FL Metro Area
14.8	25734	91	313450	0.000290	Evansville, IN-KY Metro Area
15.7	28880	1792	5455053	0.000329	Atlanta-Sandy Springs-Roswell, GA Metro Area
24.8	21814	62	177908	0.000348	Monroe, LA Metro Area

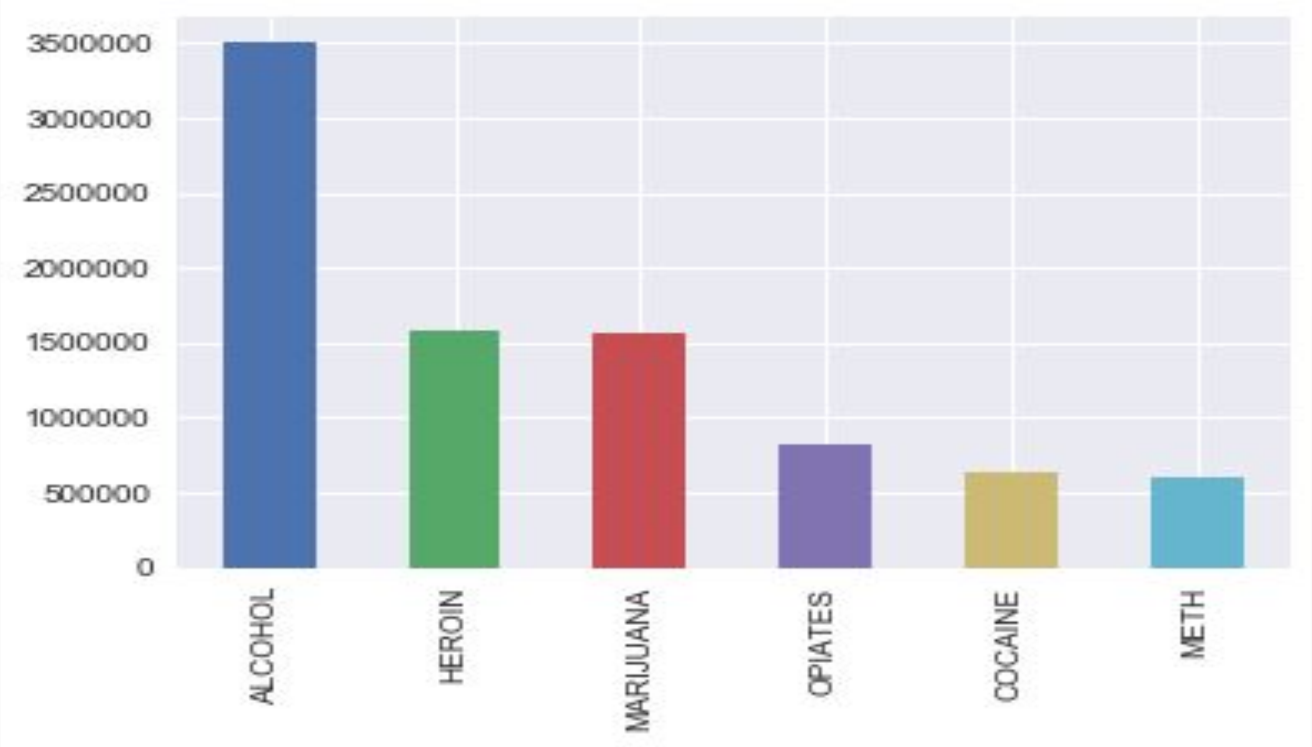
Most of the lowest patients density Metros are in the southern states Of US including Louisiana,Texas,Florida and Georgia.
Reason: The patients are actually high but the mental health facilities are poor in the southern US

Source:
http://www.theadvocate.com/baton_rouge/news/article_54398874-30df-5157-b537-3355192e8d65.html

Overview of Substance Abuse

First we get an overview of the total count of abuse by this chart. We can see alcohol is the most abused substance by a wide margin, but other substances also have a large number of recipients.

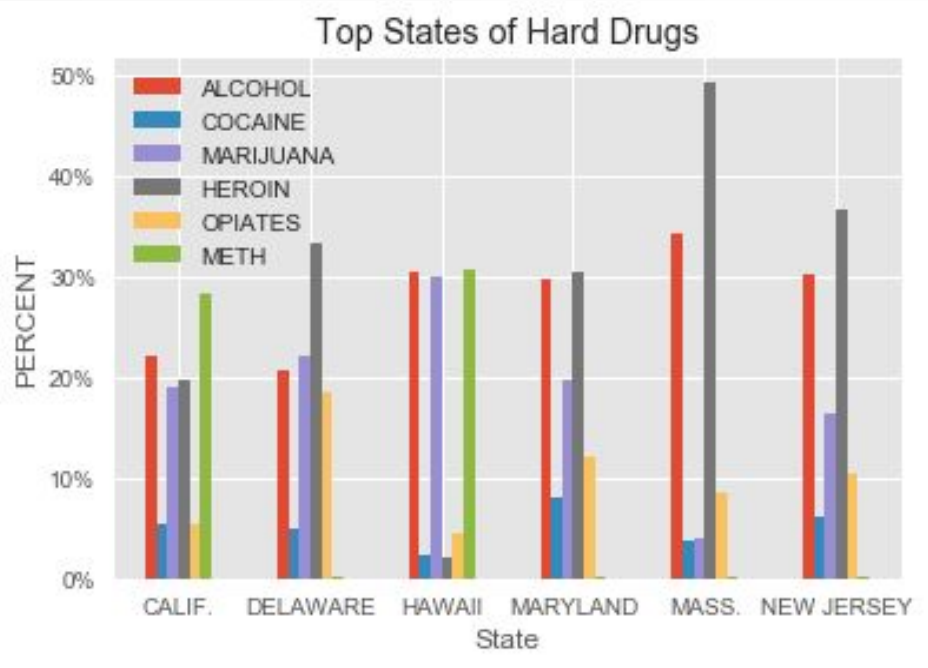
We found that after METHamphetamine there is a great drop in number, so we take top 6 substances for better visualization.



Besides Alcohol

6 states are there where Heroin and Methamphetamine abuse was even higher than alcohol

Denis is from Delaware. He reports heroin as his primary substance



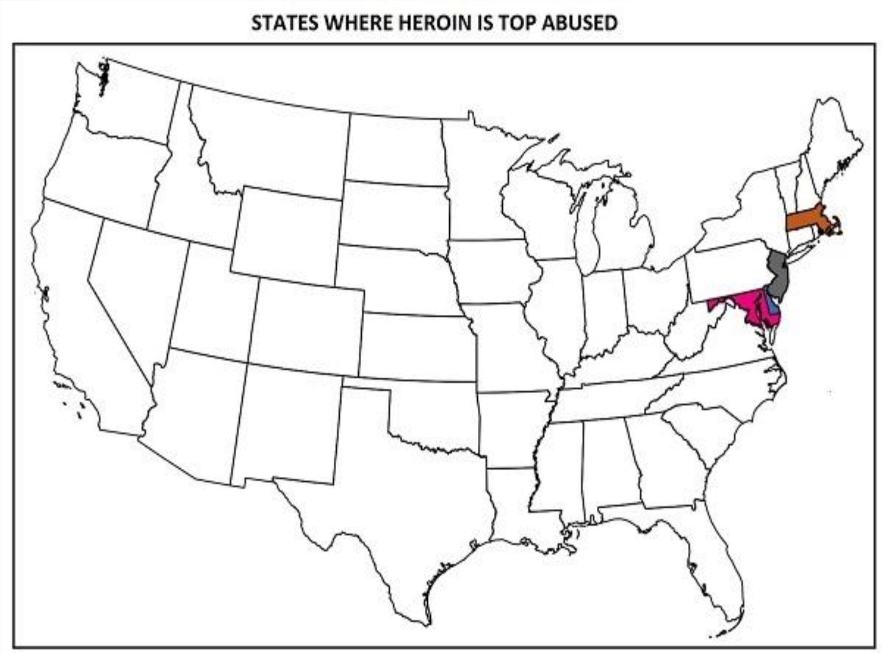
Heat Map for Heroin

We analyzed and saw that among the admitted patients Heroin, was the top abused substance in northeastern US, the states included DELAWARE, MARYLAND, MASSACHUSETTS, NEW JERSEY.

There is a good explanation for these states as they are close to West Virginia, which was once known for its Coal mines, People started using heroin in order to minimize pain caused due to injuries as well as due to Loss of well paying Blue Collar Jobs in that region.

Source:

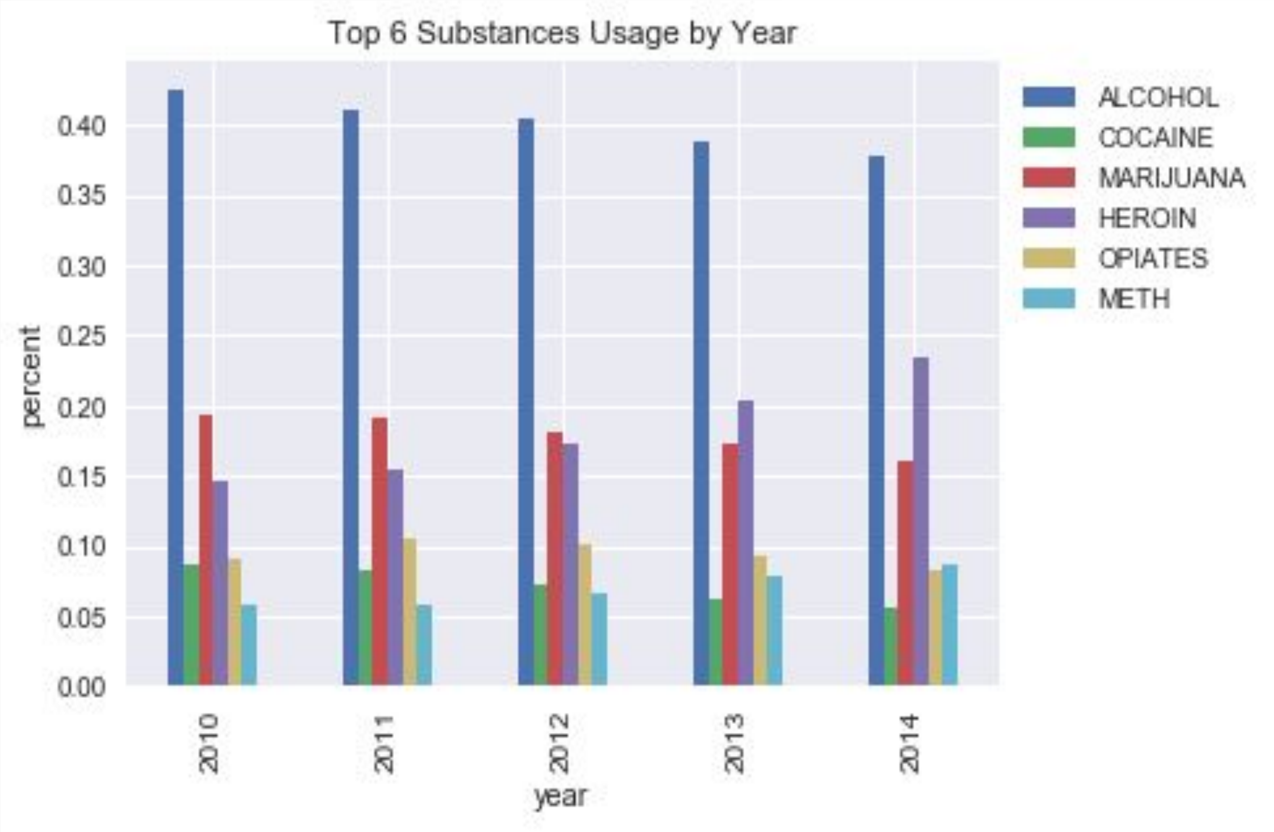
http://www.dailymail.co.uk/news/article-4472758/Inside-America-s-worst-heroin-epidemic.html



Annual Change in Substance Abuse

The abuse of alcohol is dropping sharply. The abuse of marijuana is also decreasing while the usage of heroin and meth are the only two that are increasing. We can conclude that more and more recent patients tend to abuse hard drugs.

Afterwards we found 'YEAR' is actually an important label for classifying.

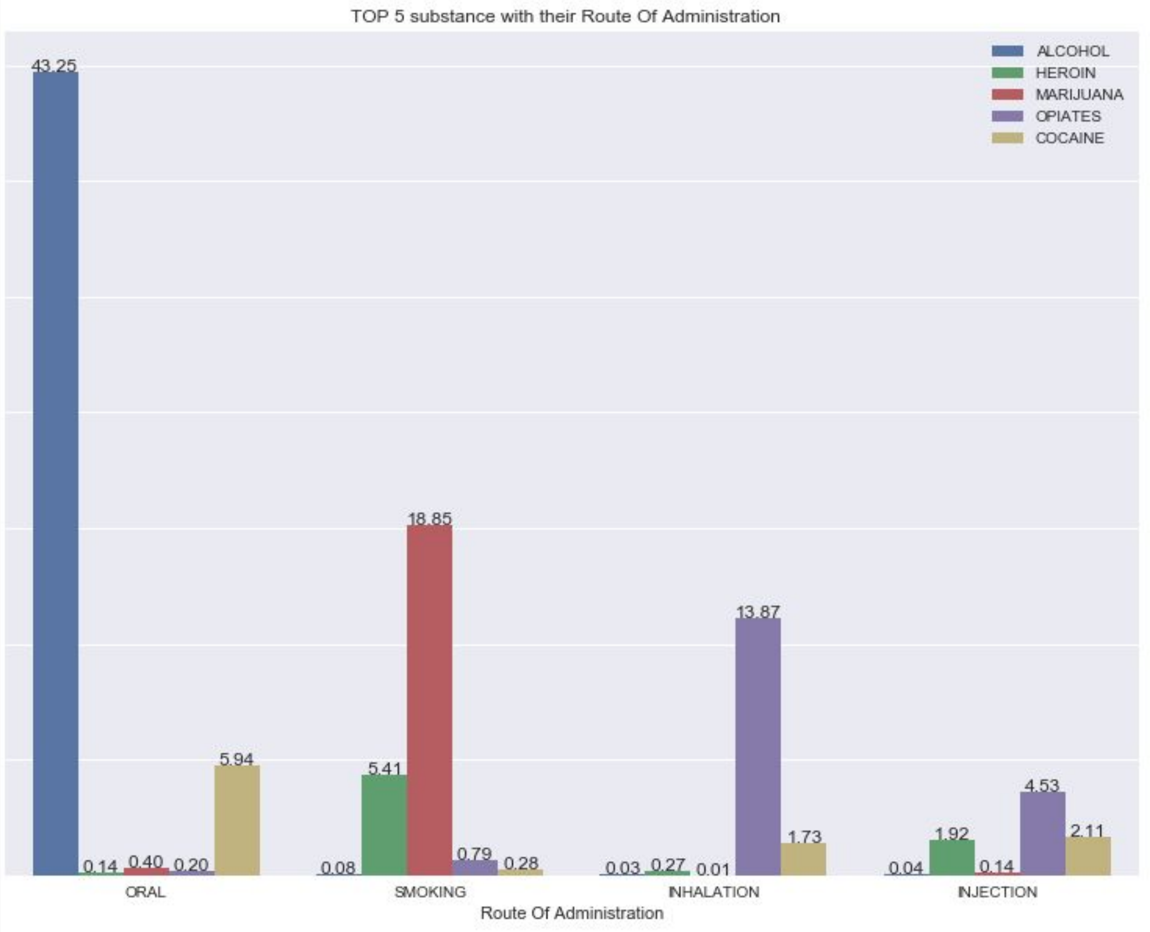


Route of Administration

Substances Abuse and Route of Administration

Observations:

- **Cocaine, heroin, and opiates are taken by all four means.**
- **It's interesting that approx .08% people smoke alcohol.**



DSM Diagnosis Explained

ALCOHOL INTOXICATION
ALCOHOL DEPENDENCE
OPIOID DEPENDENCE
COCAINE DEPENDENCE
CANNABIS DEPENDENCE
OTHER SUBSTANCE DEPENDENCE
ALCOHOL ABUSE
CANNABIS ABUSE
OTHER SUBSTANCE ABUSE
OPIOID ABUSE
COCAINE ABUSE

substance addicted disorder

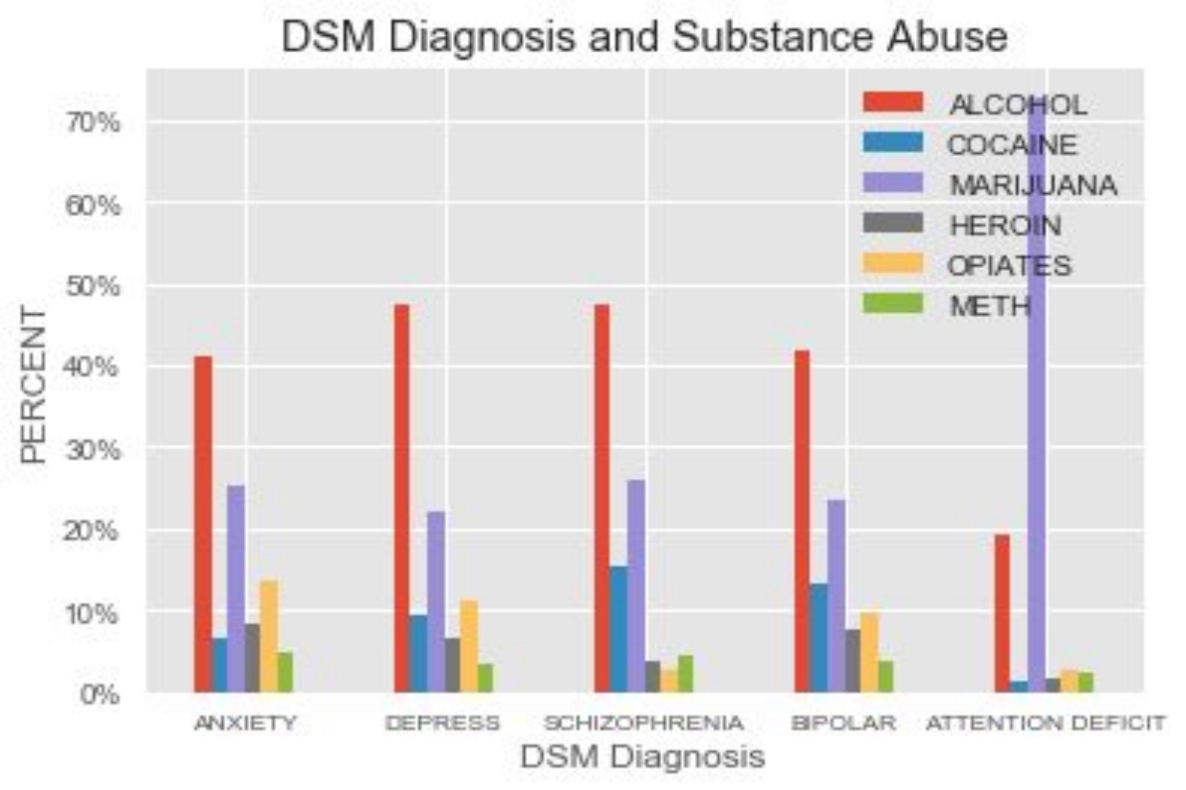
ANXIETY DISORDERS
DEPRESSIVE DISORDERS
SCHIZOPHRENIA / OTHER PSYCHOTIC DISORDERS
BIPOLAR DISORDERS
ATTENTION DEFICIT / DISRUPTIVE BEHAVIOR DISORDERS
OTHER MENTAL HEALTH CONDITION

mental health disorder

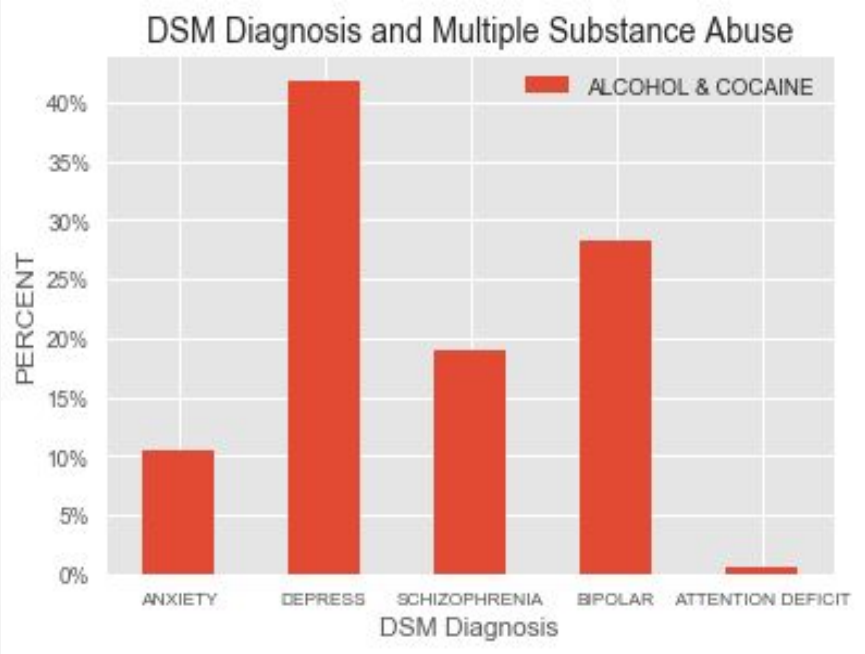
Mental Diagnosis and Substance Abuse

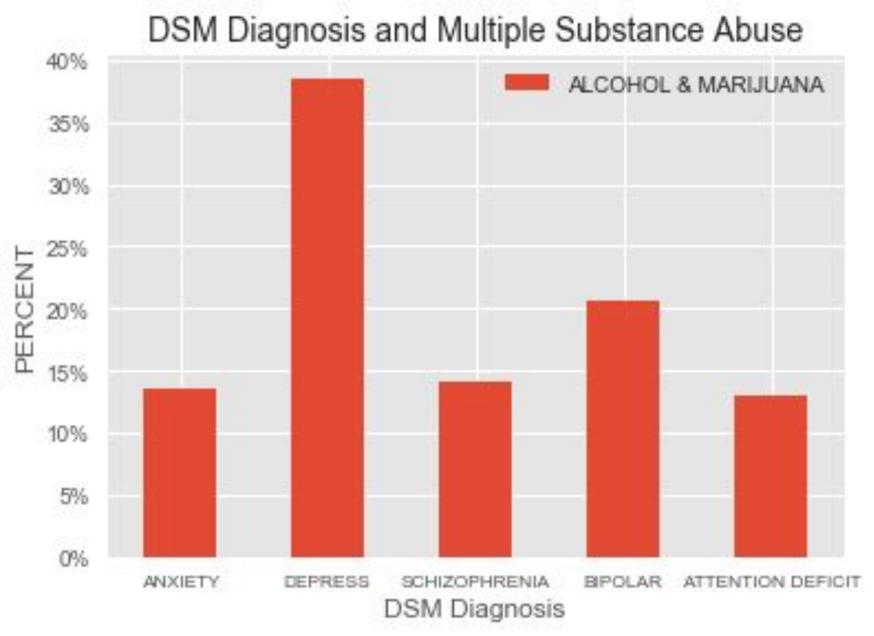
65% patients related to alcohol and marijuana.

But for attention deficit, **90% of its patients are abusing alcohol and marijuana.**



DSM Diagnosis vs SUB1 and SUB2





We try to find the relationship of DSM Diagnosis with both the substance.

- Alcohol and cocaine
- Alcohol and marijuana

Are among the top abused substances.

Predictive Mining

Prediction Task

Predict Secondary Substance

- Most of the people report at least one substance like alcohol, marijuana, inhalants fairly easily.
- But when it comes to reporting substances like cocaine, Heroin, Meth etc which are illegal or hard substances then they may not report about it.

Notes: About 42 % do not take any secondary substance.

may help hospitals to diagnose patient properly

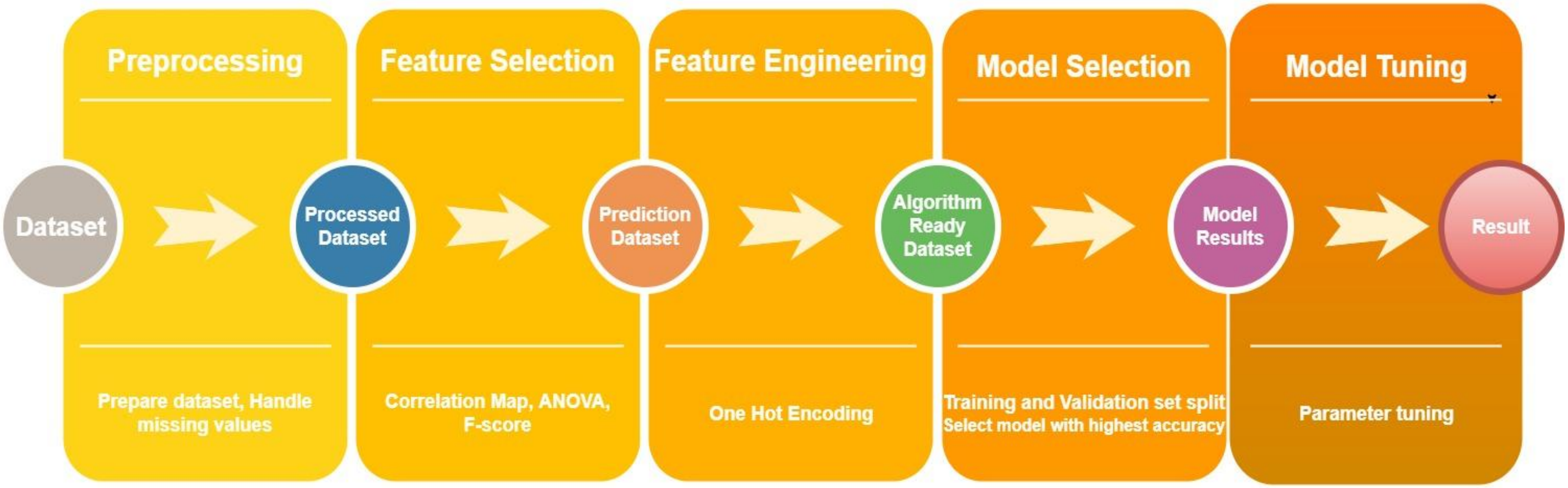
Predict Mental Health Diagnosis

DSMCRIT attribute gives information about patient's mental health diagnosis which has around 60 percent data missing

may be used as reference to help doctors to understand and know about mental health of a patient

Both predictions are aimed to give better mental health treatment to patients

Prediction Process Flow Diagram



Preprocessing

Fetch attributes

1. Fetch dataset with relevant attributes which are important for prediction task.

2. Make dataframes from these attributes.

Concat

1. Concat, merge and join the relevant dataframes together to create one dataframe.

Drop missing values

1. Check the dataset and find the missing values.

2. Impute some of the missing values.

3. Drop the rest of missing observations to create one dataframe.

Final Dataset

1. Take the final dataframe to create csv file.

2. This file can be later used for all the prediction task,

Feature Selection

Correlation Map

1. To Select relevant features for prediction task. We create Correlation Map with different attributes.

2. Based on Correlation map we choose the attributes which can affect our prediction task.

ANOVA

1. Analysis of variance (ANOVA) can determine whether the means of three or more groups are different.

2. ANOVA uses F-tests to statistically test the equality of means. The higher the F-score the better.

3. We use F-score to find the attributes which affects our prediction results.

Feature Selection: ANOVA

F-score for SUB2

```
F-score: 384.64t for feature GENDER
F-score: 398.89t for feature EMPLOY
F-score: 409.23t for feature FREQ1
F-score: 511.42t for feature PSYPROB
F-score: 538.45t for feature YEAR
F-score: 539.85t for feature EDUC
F-score: 576.25t for feature CASEID
F-score: 659.02t for feature PSOURCE
F-score: 777.69t for feature REGION
F-score: 808.08t for feature NOPRIOR
F-score: 817.57t for feature SERVSETA
F-score: 863.48t for feature DETCRIM
F-score: 944.47t for feature DIVISION
F-score: 976.85t for feature STFIPS
F-score: 1106.89t for feature DSMCRIT
F-score: 1229.30t for feature ROUTE1
F-score: 1625.03t for feature AGE
F-score: 1780.06t for feature SUB1
F-score: 2684.00t for feature IDU
F-score: 3434.18t for feature SUB3
F-score: 3526.41t for feature ROUTE3
F-score: 5327.66t for feature FREQ3
F-score: 6027.79t for feature FRSTUSE3
F-score: 35630.31t for feature FREQ2
F-score: 58734.03t for feature FRSTUSE2
```

F-score for DSMCRIT

```
F-score: 2347.10t for feature YEAR
F-score: 2414.86t for feature PSOURCE
F-score: 2435.33t for feature CASEID
F-score: 2529.56t for feature DETCRIM
F-score: 2530.45t for feature DIVISION
F-score: 2560.02t for feature REGION
F-score: 2615.98t for feature NOPRIOR
F-score: 3278.82t for feature STFIPS
F-score: 3331.53t for feature NUMSUBS
F-score: 3424.96t for feature ALCDRUG
F-score: 4740.91t for feature METHUSE
F-score: 5008.03t for feature FRSTUSE1
F-score: 5157.01t for feature AGE
F-score: 5379.46t for feature COKEFLG
F-score: 5725.27t for feature OPSYNFLG
F-score: 5984.62t for feature IDU
F-score: 6952.38t for feature SERVSETA
F-score: 8271.91t for feature ROUTE1
F-score: 8340.87t for feature MARFLG
F-score: 10180.29t for feature MTHAMFLG
F-score: 14283.57t for feature HERFLG
F-score: 16353.05t for feature ALCFLG
F-score: 34193.82t for feature SUB1
```

Feature Selection Result

SUB2 (Secondary Substance)

Age
Gender
Race
Ethnicity
Education
Employment
Living Arrangement
Number of Arrests
SUB1 (Primary Substance) Related
SUB2 (Secondary Substance) Related*

*we exclude it when in binary prediction for using second substance or not

DSMCRIT (Mental Health Diagnosis)

Age
Gender
Race
Ethnicity
Education
Employment
Living Arrangement
Number of Arrests
Number of Substance Reported
SUB1 (Primary Substance) Related
SUB2 (Secondary Substance) Related
Substance Flags

Sanity Check

- Can there be some leaking features that tell the model which secondary substance is using?

- Shuffling: take one feature out every time then use the list left to do the prediction.
- The accuracy doesn't change much.
- -> no leaking label

```
drop FEATURE 'EMPLOY': 69.85
drop FEATURE 'GENDER': 69.88
drop FEATURE 'FREQ1': 69.85
drop FEATURE 'EDUC': 69.69
drop FEATURE 'PSYPROB': 69.78
drop FEATURE 'PSOURCE': 69.55
drop FEATURE 'SERVSETA': 69.62
drop FEATURE 'DETCRIM': 69.86
drop FEATURE 'NOPRIOR': 69.59
drop FEATURE 'DSMCRIT': 69.37
drop FEATURE 'ROUTE1': 69.60
drop FEATURE 'SUB1': 65.14
drop FEATURE 'AGE': 68.73
drop FEATURE 'IDU': 69.11
drop FEATURE 'YEAR': 69.79
drop FEATURE 'REGION': 69.93
drop FEATURE 'DIVISION': 69.25
drop FEATURE 'SUB3': 68.18
drop FEATURE 'ROUTE3': 69.81
drop FEATURE 'FREQ3': 69.84
drop FEATURE 'FRSTUSE3': 69.78
drop FEATURE 'FREQ2': 69.77
drop FEATURE 'FRSTUSE2': 65.69
```

Feature Engineering

One-hot Encoding

1. We use one-hot encoding to convert our categorical data into one of k integer numbers.

2. One hot encoded attributes can be fed to machine learning algorithms for model building.

Oversampling & Undersampling

1. Oversampling and Undersampling in data analysis are techniques used to adjust the class distribution of a data set.

2. Since our dataset has different class distribution, we use algorithms like RandomUnderSampler and SMOTE.

Feature Engineering

- After applying One-hot encoding, dataset dimension changed from (1929971 records, 22 columns) to (1929971 records, 217 columns)
- The dataset has high fluctuation in class distribution of different observation.
- Oversampling increases the observations to adjust class distribution.
- Undersampling decreases the observations to adjust class distribution.

Model Selection

Train and Validate Set Split

1. The final created dataset is split into Train and Validate set in the ratio 2:1 respectively.

Model Building

1. Once we have dataset which in ready for prediction, we build our model for prediction.

2. We use different algorithms like Random Forest Classifier, Decision Tree, Logistic Regression, XGBoost etc.

Model Selection

9 Different classification algorithms were used to build the model

Random Forest, Logistic Regression and XGBoost are chosen since they are always the best three

Models

Secondary Substance

Mental Health Diagnosis

Binary Prediction

Multiclass Prediction

Substance Addicted Disorder

Predict using secondary substance or not

If predicted using it, predict which substance

6 classes
Random probability is 20%

Accuracy score is around 67%

10 classes
Random probability is 10%

Mental Disorder

6 classes
Random probability is 16,67%

Model Validation

K-fold

1. The dataset is partitioned into k equal size.

2. A single sub-sample is retained as the validation data for testing the model, and the remaining k-1 as training data.

3.Cross-validation is repeated k times with each of the k sub-samples.

Principal Component Analysis

One-hot encoding → PCA → Model building and result

Got 187 attributes

The plot shows almost 90% variance by the first 125 components.

Hence, top 125 Principal components with high variance are taken to train the model

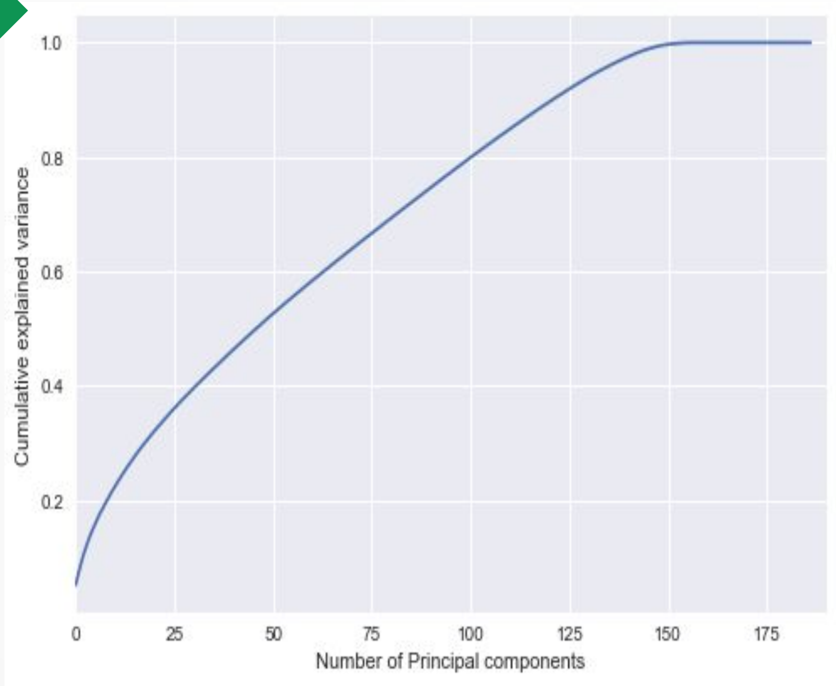
For second substance:

Random Forest: 67,5%

Logistic Regression: 59,8%

Decision tree: 57,2%

>> still lower than models without PCA



Model Tuning

Hyperparameter Tuning

1. Parameter tuning is process where we tune the parameters of Prediction Algorithms to improve results.

2. Every algorithm has different set of parameters.

3. We choose set parameters which may give better results for our dataset and keep changing them for better results.

Logistic Regression

```
- C = 100
```

[68.25 → 69.85 %]

Random Forest Classifiers

```
- n_estimators = 200
- max_depth = 50
```

[68.08 → 72 %]

XGBoost

```
- learning_rate = 0.1
- n_estimators = 5000
- max_depth = 6
- min_child_weight = 1
- gamma = 0
- subsample = 1.0
- colsample_bytree = 0.8
- nthread = 4
- scape_pos_weight = 1
- seed = 27
```

[67.00 → 68.3 %]

Results

Secondary Substance

Binary Prediction

Logistic Regression
77,44%

Random Forest
77,33%

Multiclass Prediction

6 classes
Random probability is 20%

Random Forest
72%

Logistic Regression
69,85%

XGBoost
68,3%

Mental Health Diagnosis

Substance Addicted Disorder

10 classes; Random probability is 10%

XGBoost	87,3%
Random Forest	86,2%
Logistic Regression	85,6%

Mental Disorder

6 classes; Random probability is 16,67%

Random Forest	49,8%
XGBoost	47,9%
Logistic Regression	46,1%

DSM results

	Cross Validation	Cross Validation Mean	Model
2	[0.874, 0.877, 0.868]	0.873	XGBoost
1	[0.866, 0.863, 0.857]	0.862	Random Forest
0	[0.859, 0.861, 0.848]	0.856	Logistic Regression

	Cross Validation	Cross Validation Mean	Model
1	[0.511, 0.493, 0.490]	0.498	Random Forest
2	[0.489, 0.469, 0.480]	0.479	XGBoost
0	[0.444, 0.470, 0.469]	0.461	Logistic Regression

Back to our characters...

Using secondary substance?

Bob

28 yo

Colorado

Alcohol

YES 26%

NO 74%

Carol
16 yo
Missouri

Marijuana

Using secondary substance?

- **YES** **94,8%**
- NO 5,2%

Which Substance?

- **Alcohol**
  73%
- Opiates
  19%

Mental health diagnosis:

- Cannabis use disorder
  87,22%
- Depressive disorder
  22,85%

Using secondary substance?

Which Substance?

Mental health diagnosis:

Alice
28 yo
Nebraska

Marijuana

YES **89,3%**

NO 10,7%

Alcohol
73,5%

Meth
13,5%

Cannabis use disorder
77,59%

Bipolar disorder
33,66%

What did not work

Descriptive part:

- Days waiting time : too many missing value and gibberish data.
- US census data : no good findings

Predictive part:

- PCA
- Ensemble and stacking algorithm
- Multilevel Classification

Conclusion & Lessons Learned

- Spend more time in feature study, feature selection and feature engineering before building model.
- When working on descriptive task, keep looking for prospective prediction tasks. This way we can have good findings related to what we want to predict
- Build simple model first e.g. binary prediction then into more complicated model

Thanks!