CS 5010 Semester Project: Mental Health

Kelly Farell - knf7vg@virginia.edu

Samy Kebaish - sak3qf@virginia.edu

Gretchen Larrick - jem37b@virginia.edu

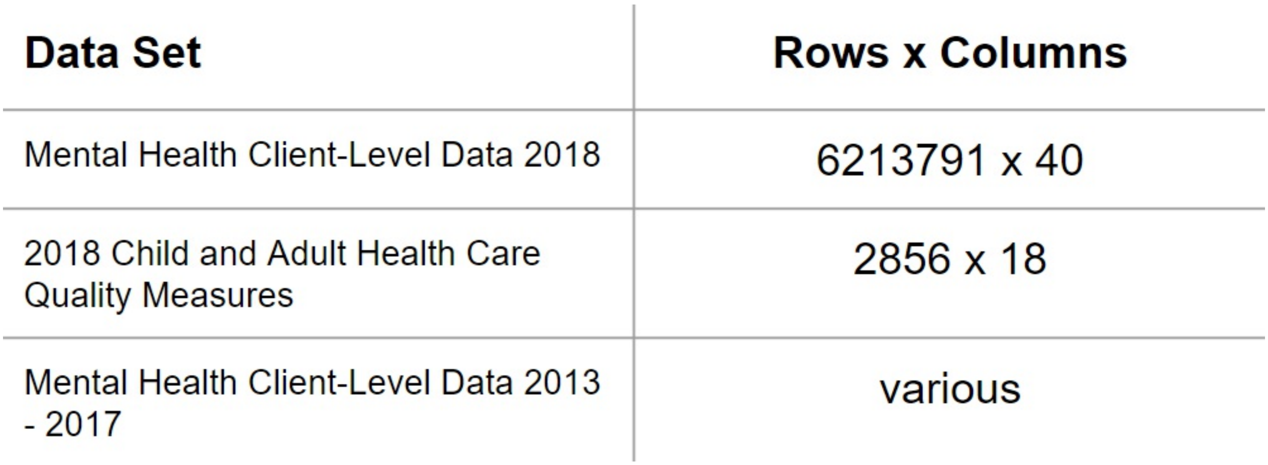
GitHub Repository - https://github.com/samykebaish/cs5010_data_analysis_project

Introduction

Mental health impacts everyone on an individual level, but it can be hard to visualize what the distribution of mental illness and treatment are across the United States. Using descriptive statistics, exploratory analyses, and an interactive website complete with visualizations was created to help patients, treatment providers, and the general public engage with diagnostic and healthcare information regarding mental illness and care. Additionally, trend analyses and traditional and machine-learning (ML) modeling methods were used to observe shifts over time and predict level of care and mental health diagnoses.

Datasets

Our primary datasets include Adult and Child Health Care Quality Measures for fiscal year 2018. Each year, the Centers for Medicare and Medicaid Services (CMS) collects benchmark data from a variety of treatment providers, with the goal of summarizing the quality of care received by adult Medicaid recipients and CHIP beneficieries. We also used the Mental Health Client-Level Data (MH-CLD) from SAHMSA (Substance Abuse and Mental Health Services Administration, a part of the US Department of Health and Human Services). We primarily focused on the year 2018, but trend analyses included data from 2013-2018. For some supplemental information regarding poverty by state, we utilized the Kaiser Family Foundation's State Facts database summarizing poverty rate by race and ethnicity, and geoJson files containing location coordinates for US state boundaries (Story and Fernandez, 2016), and state abbreviation/FIPS (World Population Review (n.d.)) to generate choropleths and merge datasets which utilized different encoding methods for state.



Data Set	Rows x Columns
Mental Health Client-Level Data 2018	6213791 x 40
2018 Child and Adult Health Care Quality Measures	2856 x 18
Mental Health Client-Level Data 2013 - 2017	various

Table 1: Dimensions of the primary datasets used in analysis and modeling mental health and treatment in the United States.

Adult and Child Health Care Quality Measures

The Social Security Act enacted by the federal government requires that an annual report of predefined core measures of healthcare quality for adult Medicaid enrollees. Based on data from electronic health records and reports from treatment centers which accept Medicaid and CHIP, this dataset contains state-level performance rates for many aspects of health, including behavioral healthcare, prenatal and pregnancy-related healthcare, and early childhood care. Reporting rates vary considerably across measures--newly defined or updated measures often have lower participation rates due to the required changes that must be made to a state's health department infrastructure necessary to support the changes to research methodology.

MH-CLD

The Mental Health Client-Level Data contain demographic, diagnoses, and treatment setting and outcome for individuals receiving services through their state mental health agency. The analyses and modeling relied on variables such as education, race, ethnicity, age, gender, primary, secondary, and tertiary mental health diagnoses, and indicators of diagnosis in a variety of illness categories (such as depressive disorders or personality disorders).

These data are somewhat limited--although the year was reported, months and days were excluded. These details would provide supplemental benefit due to the seasonality of depression. For example, in a study by Ayers et al. in 2013, Google mental health queries monitored from 2006 to 2010 revealed seasonal patterns for all mental health queries, with winter peaks and summer troughs (14% difference in the United States; 11% difference for Australia).

KFF Datasets

Based on US Census surveys, the Kaiser Family Foundation estimated the rate of people living at or under the federal poverty line for the year 2018 and grouped that data by state, race, and ethnicity.

Geographic Data

Two different datasets were utilized in order to create interactive choropleths and to join the MH-CLD and health quality measures data. One included the list of state names, abbreviations, and the state FIPS codes, which are used to identify states in the MH-CLD data (). The other contains state boundary data through arrays of coordinates (Story and Fernandez, 2016).

Tech Stack

Code was written in Jupyter Notebook (Kluyver et al., 2016) for reproducibility and easy annotation.

Much of the exploratory data analysis was conducted with functions imported from NumPy (Harris et al., 2020) and Pandas (McKinney et al., 2010).

Visualizations for the exploratory data analysis were produced using Matplotlib (Hunter et al., 2007) and seaborn (Waskom et al., 2017). Plotly (Plotly Technologies Inc, 2015) was used to create further exploratory visualizations, as well as interactive graphics such as choropleths for the website.

Plots related to the logistic regression and advanced ML modeling were created using Plotly (2015) and bokeh (Bokeh Development Team, 2021).

Code for modeling was primarily written using Scikit (Pedregosa et al., 2011), Keras (Chollet et al., 2015), and _Flask (Grinberg, 2018). The website was hosted by AWS.

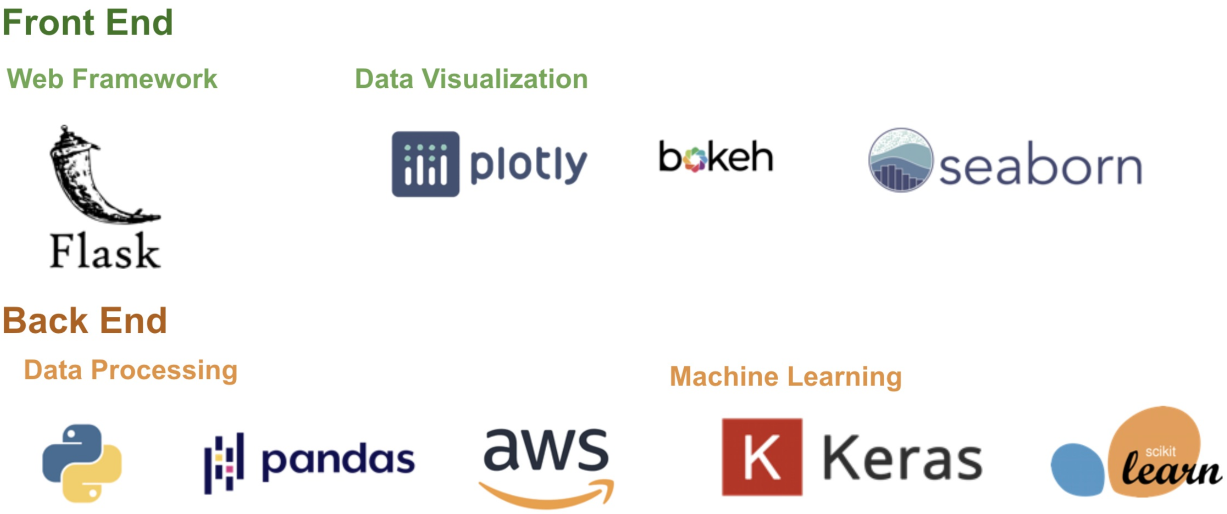


Figure 1. Tech Stack: Front End and Back End Architecture

Preprocessing

The datasets were carefully cleaned by the governmental and nongovernmental research agencies prior to publishing, so little preprocessing was required. However, null/missing values needed to be accounted for and the datasets needed to be merged together for some of the analyses.

Irrelevant Data

Rows from the Adult and Child Health Care Quality Measures dataset which involved measures from treatment domains other than "Behavioral Health" were removed in order to focus on measures related to mental health. The rows were identified by filtering out rows from the column "Domain" which were not equal to "Behavioral Health" and excluding them from the dataset.

Null Data

Null data in the MH-CLD and Adult and Child Health Care Quality Measures datasets were encoded using "#NR" or -9. These cells were removed prior to analysis and modeling. Since in nearly all cases, the data was categorical, some outlier handling techniques such as imputation with median values or upper quartile values was not possible.

Joining Datasets

The Health Care Quality Measures dataset was merged with the folio state abbreviation list from World Population Review (n.d.) using state code/abbreviation as the shared identifier. The MH-CLD dataset was joined with the same state abbreviation list using the USPS state name as the shared identifier.

Finally, the merged Health Care Quality Measures data and merged MH-CLD data were then joined using the state code/abbreviation as the shared identifier.

Exploratory Data Analysis

With the data cleaned, relevant descriptive statistics and other trends could be identified in order to inform the visualizations and prediction models.

MH-CLD

Single Variable Analysis

The MH-CLD data set consists of a large number of categorical variables corresponding to a specific state. To understand the data, the first object was the look at the total number of entries per state, Figure 2.

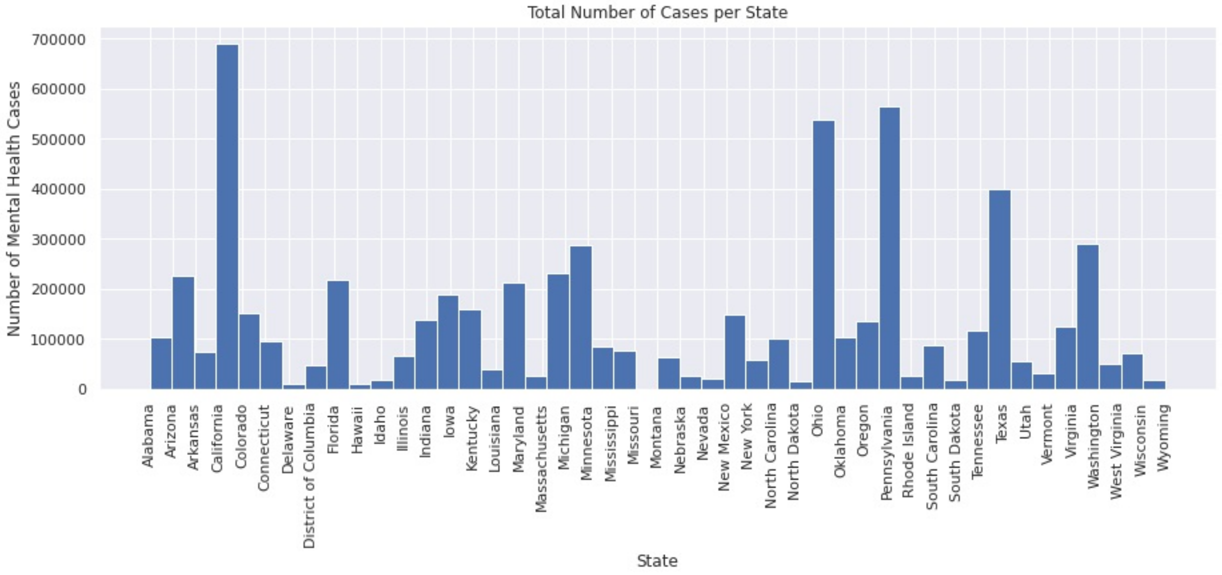


Figure 2: Mental Health Cases per State

The next part of the exploratory analysis w as to sort through the guidebook provided by SAMHA detailing the breakdow n of the variables in the dataset. Age and mental health diagnosis have a high number of populated categories so these w ere the main areas of focus.

The first to explore is age. The variable is not an integer age, but a range of ages. The first goal is to get a visualization of each age group's total count, figure 3. This show s that the 0-11 age group has a significant number of higher cases than the other age groups.

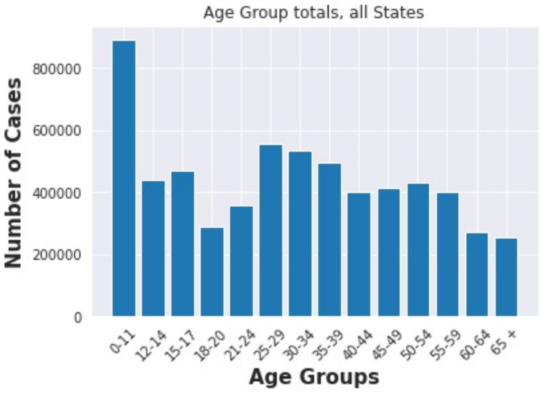


Figure 3: Age Group Totals

With the dataset broken dow n into states, the data is then group into age groups by state, figure 4. Again, the 0-11 age range has the highest number of cases in a majority of the states.

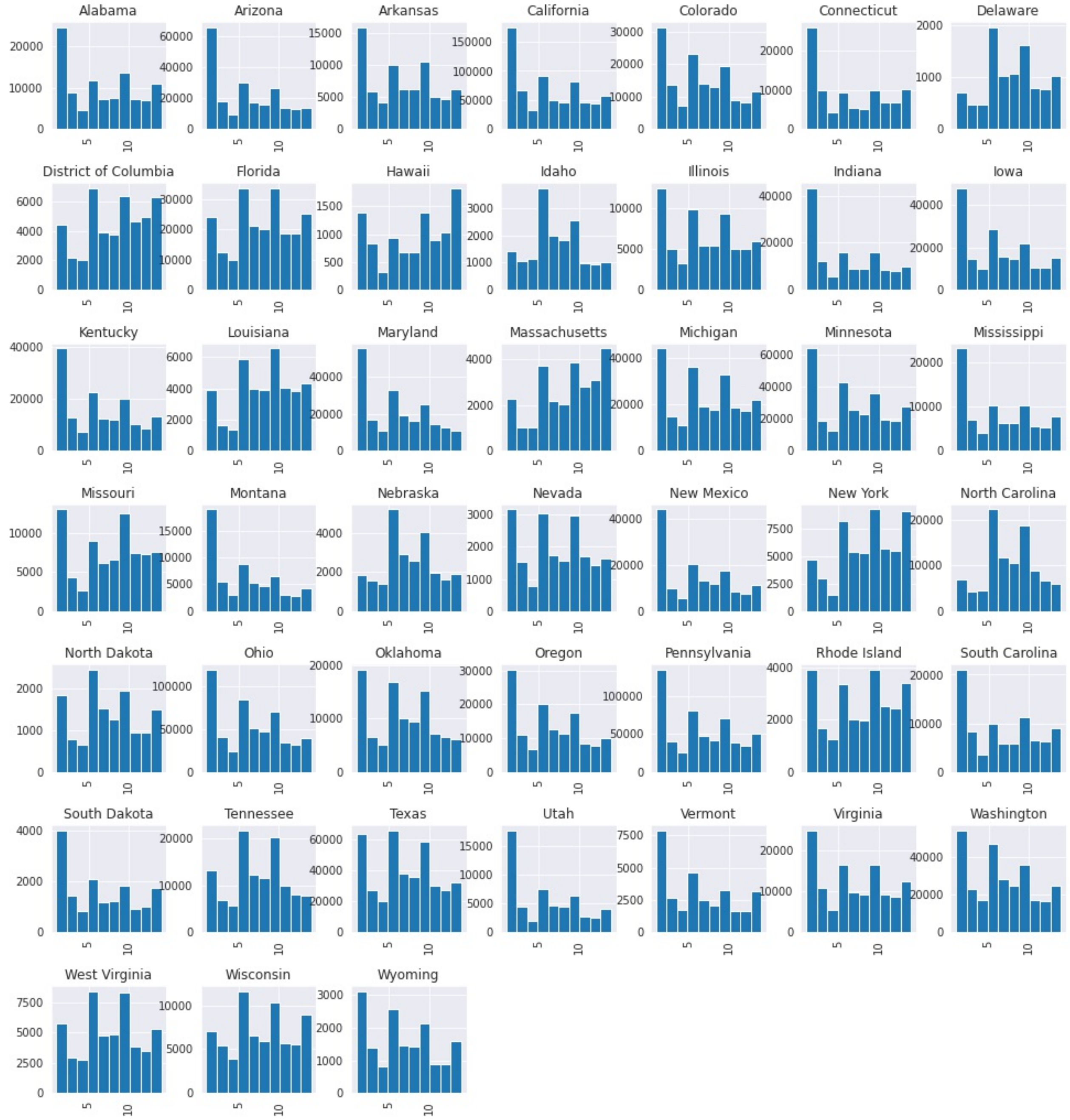


Figure 4: Age Group Totals by State

The same analysis w as completed on the mental health diagnosis variable. The totals for each mental health diagnosis are show n in figure 5. Depression has the highest number of cases in this dataset.

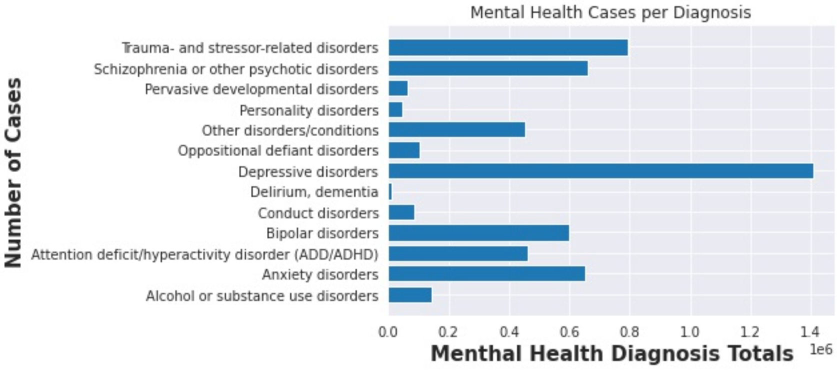


Figure 5: Mental Health Diagnosis Totals

The variable is then broken dow n by state, figure 6, and the conclusion of w hich mental health diagnosis is the highest isn't as clear as w ith the age analysis. The breakdow n of the state mental health diagnosis is much more varied.

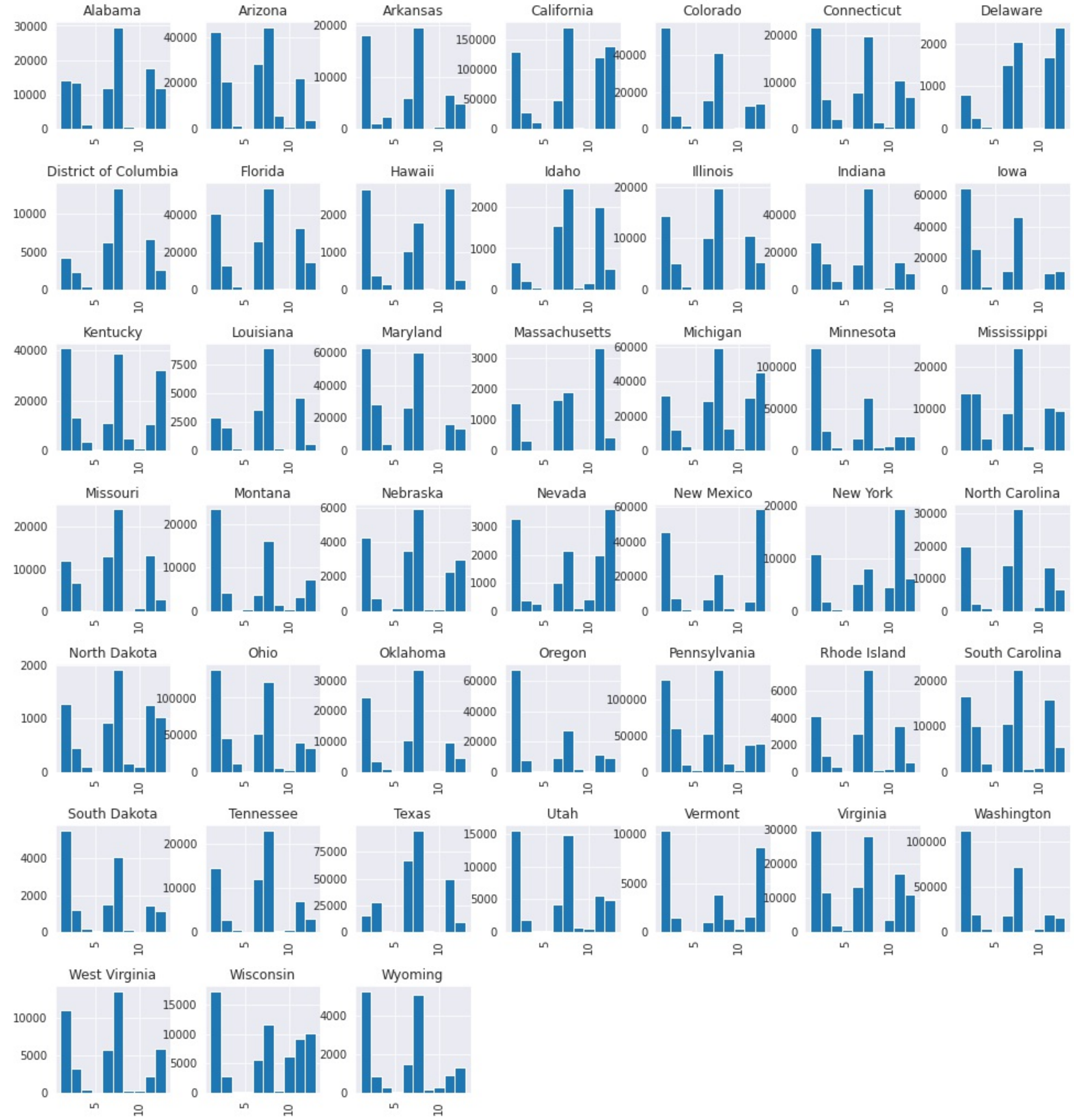


Figure 6: Mental Health Diagnosis Totals by State

Multivariable Analysis

After looking at age and mental health diagnosis separately, how do these tw o variables relate to each other? In figure 7, age ranges are grouped by their mental health diagnosis. From this, it is seen that the depressive mental health diagnosis is prominent in most of the age groups, w ith a higher instance after an individual is over the age of 15.

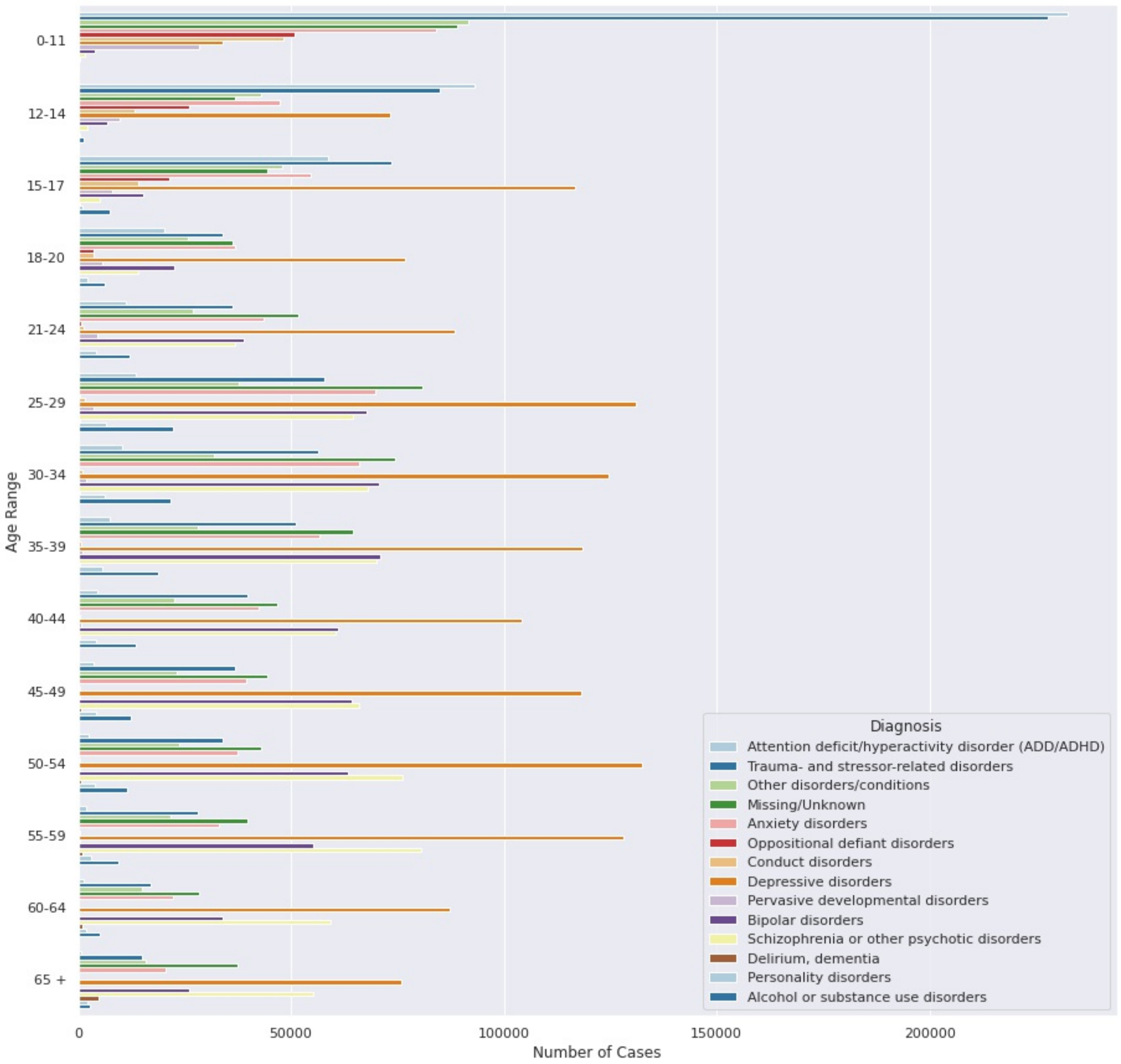


Figure 7: Age groups by Mental Health Diagnosis

There is a difference in diagnoses for individuals w ho are over 18 in relation to individuals under the age of 18. Individuals under the age of 18 have a much higher number of cases in the ADD/ADHD, Trauma/Stressors, and Oppositional Defiant Disorders categories. After the age of 18, there is an increase in depression, but it varies in the over 18 age groups, but each group remains higher individually than the under 18 age groups.

No prediction value w as added by utilizing the poverty dataset. For that reason, w e chose to eliminate it from future consideration in our models. The analyses show ing no meaningful insights w ith the poverty data included can be seen in our code w ithin the team's GitHub repository.

Adult and Child Health Care Quality Measures

Overall, the measures using the adult population had a higher percentage of nonmissing data than the child population. On the documentation for the original dataset, SAHMSA indicates that comparison of the child dataset w ith past years may not be fasible because some measures w ere only created or standardized in recent years. As a result of the changes, many states do not have sufficient data to be included in the analyses.

We chose to analyze follow -up visit information--specifically, outpatient follow -up appointments provided w ithin 30 days after discharge from a hospitalization for mental illness. In order to be hospitalized for mental illness, a person needs to be evaluated by medical professionals (most commonly at an emergency department) and determined to be at an acute risk of harm to themselves or others. This indicates a lot of distress and impact on an individual's life, so outpatient care is needed to help support them as they return to daily life. This variable w as chosen because it is one of the few measurements that can apply to any individual w ho is hospitalized--it is not specific to any diagnosis. We also looked at outpatient follow -up appointments provided w ithin 30 days after an emergency room visit (w hich did not result in a hospitalization).

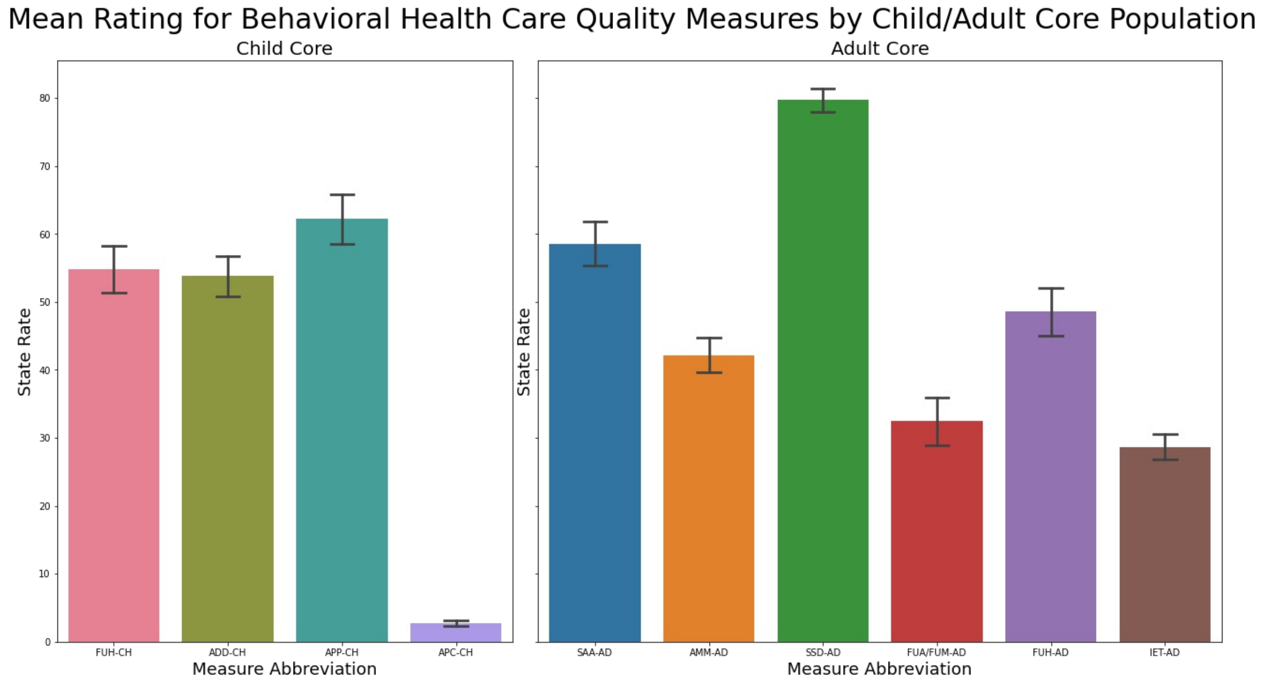
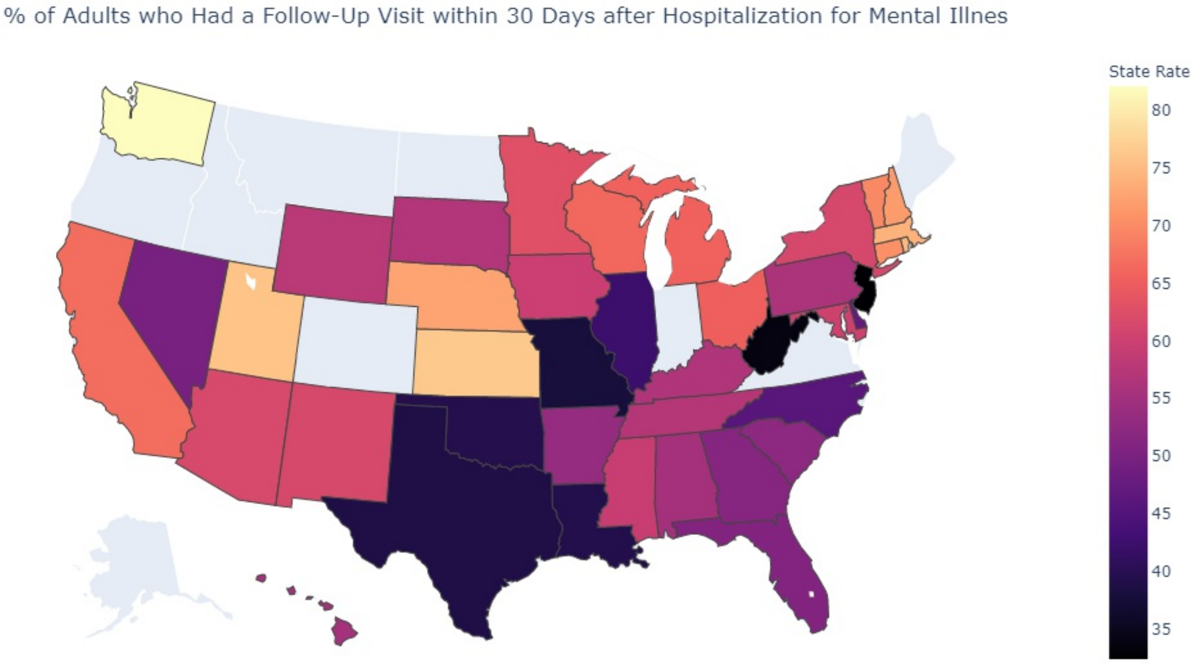
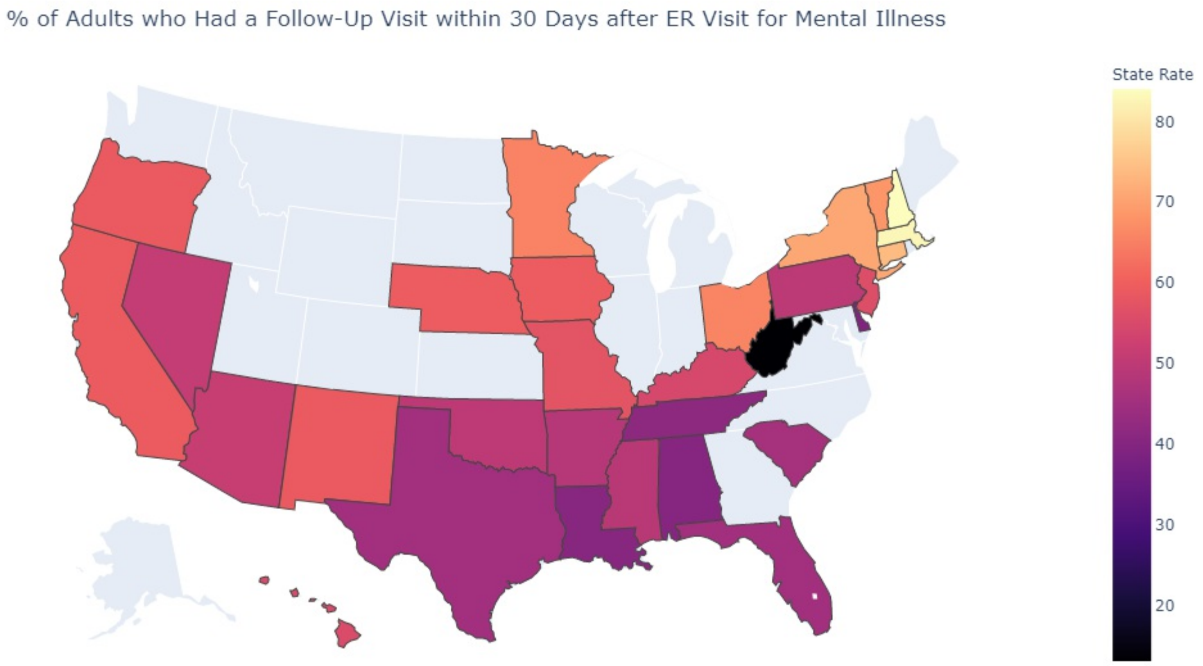


Figure 8: National Average Performance on Health Care Quality Measure (child population - right, adult population - left)


% of Adults who Had a Follow-Up Visit within 30 Days after Hospitalization for Mental Illnes
State Rate
80
75
70
65
60
55
50
45
40
35

Choroleth 1: The percentage of adults who received a follow-up visit with an outpatient treatment provider after discharge from hospitalization for mental illness, by US state


% of Adults who Had a Follow-Up Visit within 30 Days after ER Visit for Mental Illness
State Rate
80
70
60
50
40
30
20

Choropleth 2: The percentage of adults w ho received a follow -up visit w ith an outpatient treatment provider after an emergency room visit for mental illness, by US state

While follow -up visits w ith outpatient treatment providers are more likely to occur w hen a patient is discharging from the hospital, rather than after being screened out in the emergency room, states' performance on this measure had a range of about 35% to 80%.

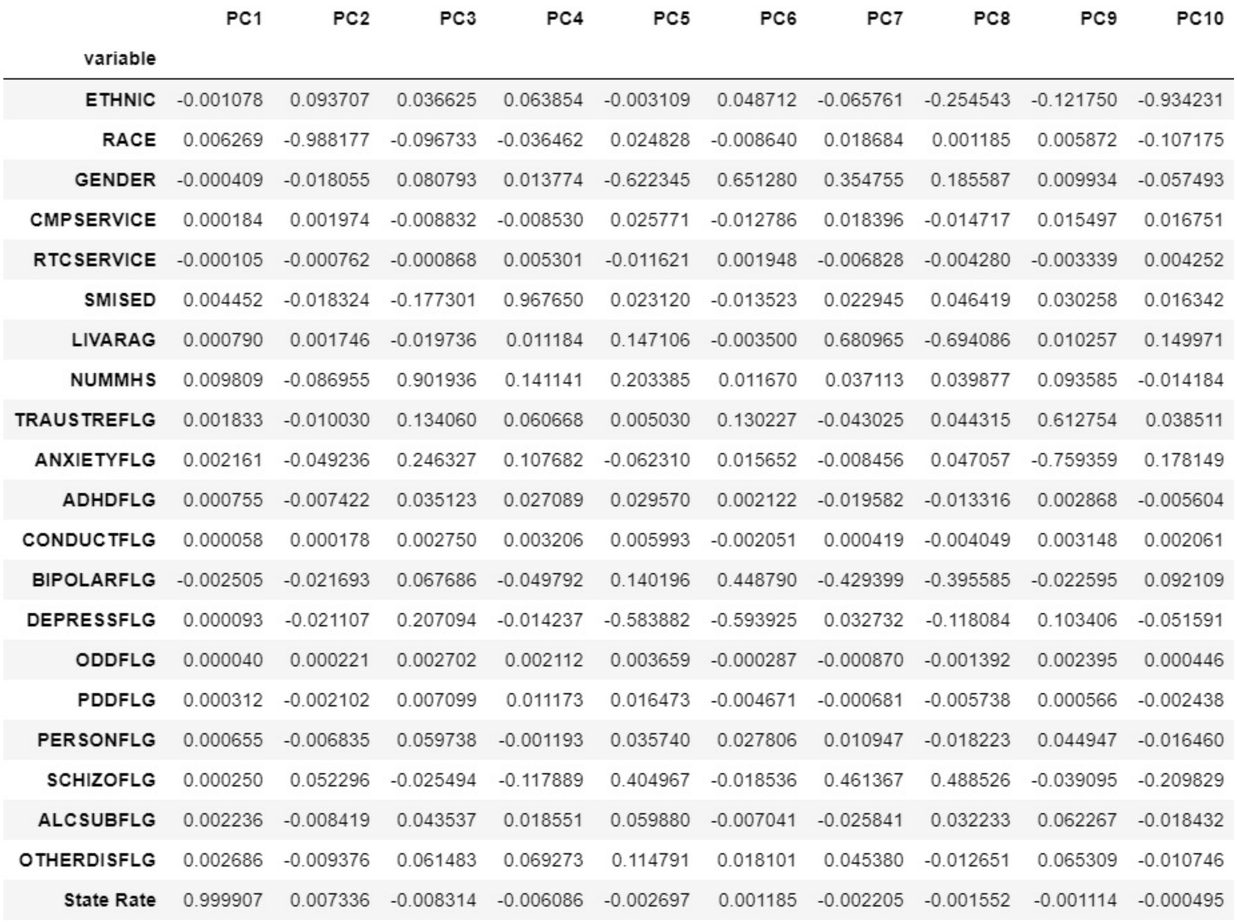
Since the proportion of states providing data for the follow -up measure after emergency room w as substantially low er than the measure for follow -ups after hospitalization, w e chose to incorporate only the post-hospitalization follow -up measure in the logistic regression model. We w ere interested in w hether or not each state's performance on this follow -up measure could predict treatment outcomes--specifically, does a higher rate of outpatient follow ups for patients discharging from a hospitalization for mental illness w ithin a state correlate to the level of care that a patient from that state need? In other w ords, if states provide better follow -up care, do patients from that state require treatment from inpatient or criminal justice-based treatment centers less often than in states w ith poorer follow -up care?

General Linearized Model

In order to study the relationship betw een post-hospitalization follow -ups and treatment outcomes, w e decided to fit a logistic regression model w ith a binary target variable. The target variable w as 'InptJust', w hich combined the flags for inpatient treatment from state-funded facilities, inpatient treatment from non-state-funded facilities, and treatment provided by factilities for individuals w ho are detained or incarcerated. If a patient met any of those criteria, they w ould be considered to be in the higher level of treatment grouping ('InptJust' = 1), w hile patients receiving outpatient or community-based treatment w ould be classified as the low er level of treatment group ('InptJust = 0'). We then needed to determine w hich other variables to include beside 'State Rate', w hich encodes the state's performance on the follow -up visit variable.

Principal Component Analysis

A Principal Component Analysis w as used to evaluate all variables in the combined CLD-Health Care Quality Measures dataset. First, the numeric data w as scaled using a standard scaler, to prevent biased scores that result from having different measurement units for numeric variables. The data w ere split into training and testing sets, w ith 90% of the data used for training the model, and the remaining 10% for testing.



variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
ETHNIC	-0.001078	0.093707	0.036625	0.063854	-0.003109	0.048712	-0.065761	-0.254543	-0.121750	-0.934231
RACE	0.006269	-0.988177	-0.096733	-0.036462	0.024828	-0.008640	0.018684	0.001185	0.005872	-0.107175
GENDER	-0.000409	-0.018055	0.080793	0.013774	-0.622345	0.651280	0.354755	0.185587	0.009934	-0.057493
CMPSERVICE	0.000184	0.001974	-0.008832	-0.008530	0.025771	-0.012786	0.018396	-0.014717	0.015497	0.016751
RTCSERVICE	-0.000105	-0.000762	-0.000868	0.005301	-0.011621	0.001948	-0.006828	-0.004280	-0.003339	0.004252
SMISED	0.004452	-0.018324	-0.177301	0.967650	0.023120	-0.013523	0.022945	0.046419	0.030258	0.016342
LIVARAG	0.000790	0.001746	-0.019736	0.011184	0.147106	-0.003500	0.680965	-0.694086	0.010257	0.149971
NUMMHS	0.009809	-0.086955	0.901936	0.141141	0.203385	0.011670	0.037113	0.039877	0.093585	-0.014184
TRAUSTREFLG	0.001833	-0.010030	0.134060	0.060668	0.005030	0.130227	-0.043025	0.044315	0.612754	0.038511
ANXIETYFLG	0.002161	-0.049236	0.246327	0.107682	-0.062310	0.015652	-0.008456	0.047057	-0.759359	0.178149
ADHDFLG	0.000755	-0.007422	0.035123	0.027089	0.029570	0.002122	-0.019582	-0.013316	0.002868	-0.005604
CONDUCTFLG	0.000058	0.000178	0.002750	0.003206	0.005993	-0.002051	0.000419	-0.004049	0.003148	0.002061
BIPOLARFLG	-0.002505	-0.021693	0.067686	-0.049792	0.140196	0.448790	-0.429399	-0.395585	-0.022595	0.092109
DEPRESSFLG	0.000093	-0.021107	0.207094	-0.014237	-0.583882	-0.593925	0.032732	-0.118084	0.103406	-0.051591
ODDFLG	0.000040	0.000221	0.002702	0.002112	0.003659	-0.000287	-0.000870	-0.001392	0.002395	0.000446
PDDFLG	0.000312	-0.002102	0.007099	0.011173	0.016473	-0.004671	-0.000681	-0.005738	0.000566	-0.002438
PERSONFLG	0.000655	-0.006835	0.059738	-0.001193	0.035740	0.027806	0.010947	-0.018223	0.044947	-0.016460
SCHIZOFLG	0.000250	0.052296	-0.025494	-0.117889	0.404967	-0.018536	0.461367	0.488526	-0.039095	-0.209829
ALCSUBFLG	0.002236	-0.008419	0.043537	0.018551	0.059880	-0.007041	-0.025841	0.032233	0.062267	-0.018432
OTHERDISFLG	0.002686	-0.009376	0.061483	0.069273	0.114791	0.018101	0.045380	-0.012651	0.065309	-0.010746
State Rate	0.999907	0.007336	-0.008314	-0.006086	-0.002697	0.001185	-0.002205	-0.001552	-0.001114	-0.000495

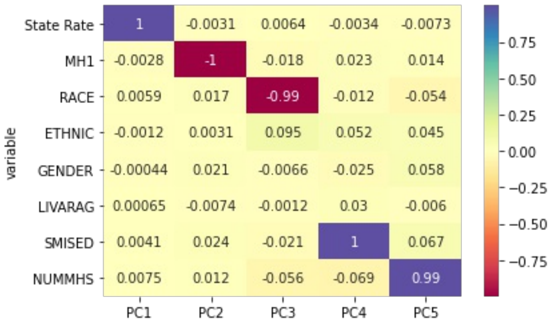
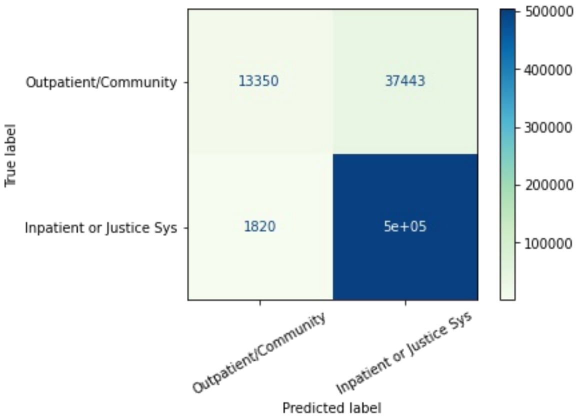
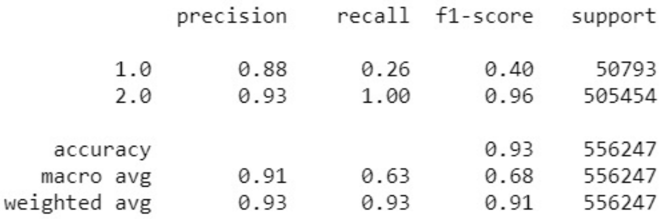


Figure 9. PCA CLD Health Care Quality Measures. Top: All variables. Bottom: variables with highest coefficients (for simplcity)

Based on the results of the principal component analysis, the variables 'State Rate' (of follow-up visits within 30 days of discharge from hospitalization), 'MH1' (primary mental health diagnosis), Race, Ethnicity, Gender, LIVARAG (Living Arrangement), SMISED (flag code for Severe Mental Illness or Serious Emotional Disturbance), and NUMMHS (number of mental health diagnoses) were regressed in order to predict the binary variable

Logistic Regression





```
              precision    recall  f1-score   support

         1.0       0.88      0.26      0.40     50793
         2.0       0.93      1.00      0.96    505454

    accuracy                           0.93    556247
   macro avg       0.91      0.63      0.68    556247
weighted avg       0.93      0.93      0.91    556247
```

Figure 10. Logistic Regression Analysis (Top: confusion matrix, Bottom: precision, recall, and f-1 scores.)

The model's accuracy w as high overall, but this w as mainly due to how large the proportion of low er level of treatment group w as. The model's performance at predicting and differentiating the higher level of treatment group w as w eak, though, w ith only 26% recall.

Since the logistic regression w as not successful, w e decided to investigate trends over time as w ell as machine learning methods to improve the model's accuracy.

Trend Analysis

As data has been garnered since the year 2013 in a consistent fashion, w e decided to conduct analyses on w hether mental illness incidence has been increasing. To our surprise, the rates of mental illness have been fairly consistent throughout the years, and the diagnostic comparison betw een males and females follow ed a similar trend line amongst both groups. Moreover, disease distribution w as also demonstrated to be consistent among years. For example, see Figure 11 for the distribution of mental health illnesses in 2016 as compared to 2018. How ever, in future studies, more extensive analysis is w arranted. Namely, in our analysis, w e did a deep dive into the dataset related to 2018. In order to present a truly robust representation, a similar level of depth w ould be necessary for each year (from 2013 onw ards), w hich then can make comparisons more meaningful.

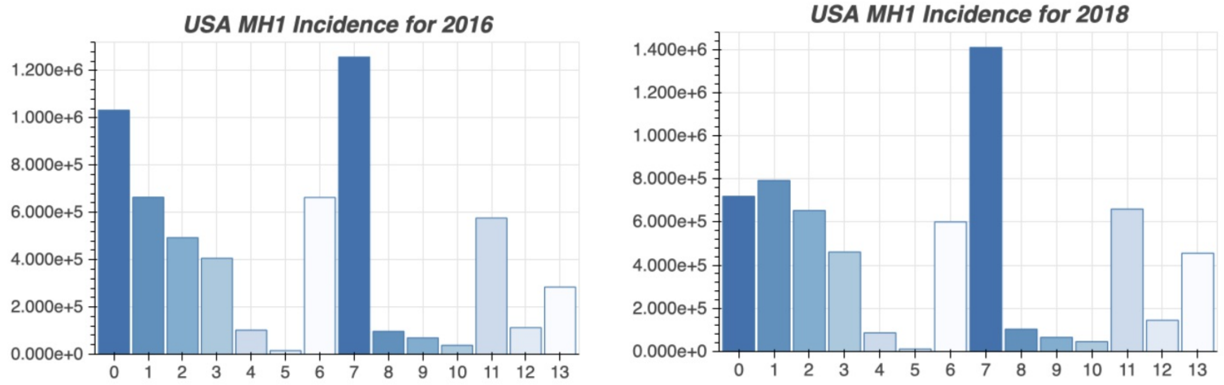


Figure 11. MH1 Histogram Comparison: 2016 vs 2018

Above and Beyond: Advanced ML Methods and Interactive Website

As part of our advanced analysis, w e investigated the potential usage of machine learning algorithms tow ards uncovering trends related to various features and labels of our dataset. The labels w e used included a mix of demographic, w hereas the labels w ere based on multiple disorder diagnoses (Table X). The range of values for the labels and features w ere either binary (0 or 1), polyadic (e.g. -9, 1, 2, 3, 4) and continuous (0-100). Some of the features includes w ere age, education, ethnicity, race, substance abuse history, marital status, veteran status etc. The labels investigated included multiple disorders - such as major depressive disorder, ADHD, anxiety, schizophrenia - w hich w ere coded 0 for absence and 1 for presence thereof. Based on the structure of the data, w e evaluated three separate machine learning models: Naive Bayes, K-Nearest Neighbors and a multilabel classification model using ADAM optimization. The NB and KNN classifiers w ere conducted using the scikit-learn library, w hereas the multilabel classification model w as derived from the keras library.

Naive Bayes

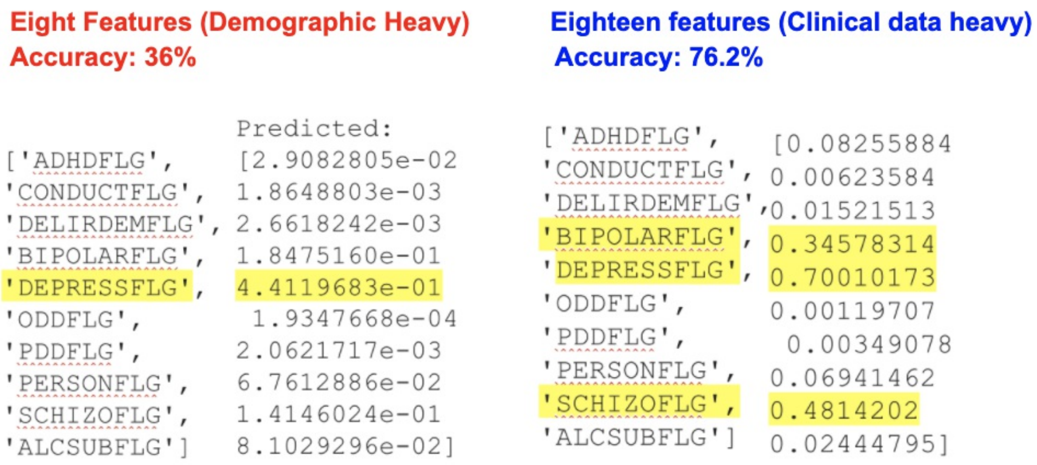
The three models differed in their properties and use cases. In the case of Naive Bayes (NB), as the name suggests, Naive bayes utilizes Bayes theorem (that the probability of an event can be based on prior know ledge of conditions w hich may have a relation to the event) in conjunction w ith the “naive” assumption that the attributes are conditional independent. As a supervised learning technique, Naive Bayes is considered to be a decent classifier, extremely fast compared to sophisticated methods, but a bad estimator, particularly as datasets become more complex. There are multiple types of naive bayes, including Gaussian (Figure 12), Multinomial, Complement, Bernoulli, Categorical, and “Out-of-core” Naive Bayes model fitting. The Naive Bayes algorithm utilized in this case w as GaussianNB. Unlike the KNN and Multilabel classifier (discussed in subsequent paragraphs), the Naive Bayes classifier could handle multiple features, but w as limited to a single label output. Accordingly, our implementation meant that, in order to evaluate multiple labels, each one w ould need to be done individually, and the others follow ed in iterative O(n) fashion.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Figure 12. NB Equation

k-Nearest Neighbors & Multilabel Classification Optimization

K Nearest Neighbors is an algorithm w hich assumes that similar data points are close to or “neighbor” one another. It can be used to solve classification and regression problems. We used a KNN classifier due to its ability to handle multi feature, multilabel data. Additional, w e use keras for similar, albeit more robust, multifeature, multilabel classification. Keras is an intuitive deep learning API w hich acts as an interface for the Tensorflow Library. We implemented a Multilabel Classification using the ADAM optimization algorithm. The Adam optimization algorithm is used, as opposed to stochastic gradient descent, w hich stands for Adaptive Moment Estimation. Unlike stochastic gradients, w hich use single learning rates, the Adam algorithm implements both Adaptive Gradient Algorithm and Root Mean Square Propagation. These tw o mechanisms calculate an exponential moving average of the gradient and the squared gradient. As an example, w e ran a Keras multi label classification model implementing initially 8 demographic variables (features) and 10 mental health diagnoses (labels) for the follow ing representation: A 40-44 year old, High School Educated, Hispanic (other than Mexico or Puerto Rico), White in Ethnicity, Female, from California, w ho is divorced. The initial accuracy w as only 36%, demonstrating the low impact of demographic variables in ascertaining mental health (for this dataset). How ever, after adding information about veteran status, substance abuse, homelessness, admissions to hospital and other clinical features (18 total), the maximum accuracy w as 76.2%. Ultimately, accuracy scores w ere similar for KNN and multilabel classification, although multilabel classification had a low er sensitivity.



Eight Features (Demographic Heavy)
Accuracy: 36%

```
                 Predicted:
['ADHDFLG',      [2.9082805e-02
'CONDUCTFLG',    1.8648803e-03
'DELIRDEMFLG',   2.6618242e-03
'BIPOLARFLG',    1.8475160e-01
'DEPRESSFLG',    4.4119683e-01
'ODDFLG',         1.9347668e-04
'PDDFLG',        2.0621717e-03
'PERSONFLG',     6.7612886e-02
'SCHIZOFLG',     1.4146024e-01
'ALCSUBFLG']     8.1029296e-02]
```

Eighteen features (Clinical data heavy)
Accuracy: 76.2%

```
['ADHDFLG',      [0.08255884
'CONDUCTFLG',    0.00623584
'DELIRDEMFLG',0.01521513
'BIPOLARFLG',    0.34578314
'DEPRESSFLG',    0.70010173
'ODDFLG',        0.00119707
'PDDFLG',         0.00349078
'PERSONFLG',     0.06941462
'SCHIZOFLG',     0.4814202
'ALCSUBFLG']     0.02444795]
```

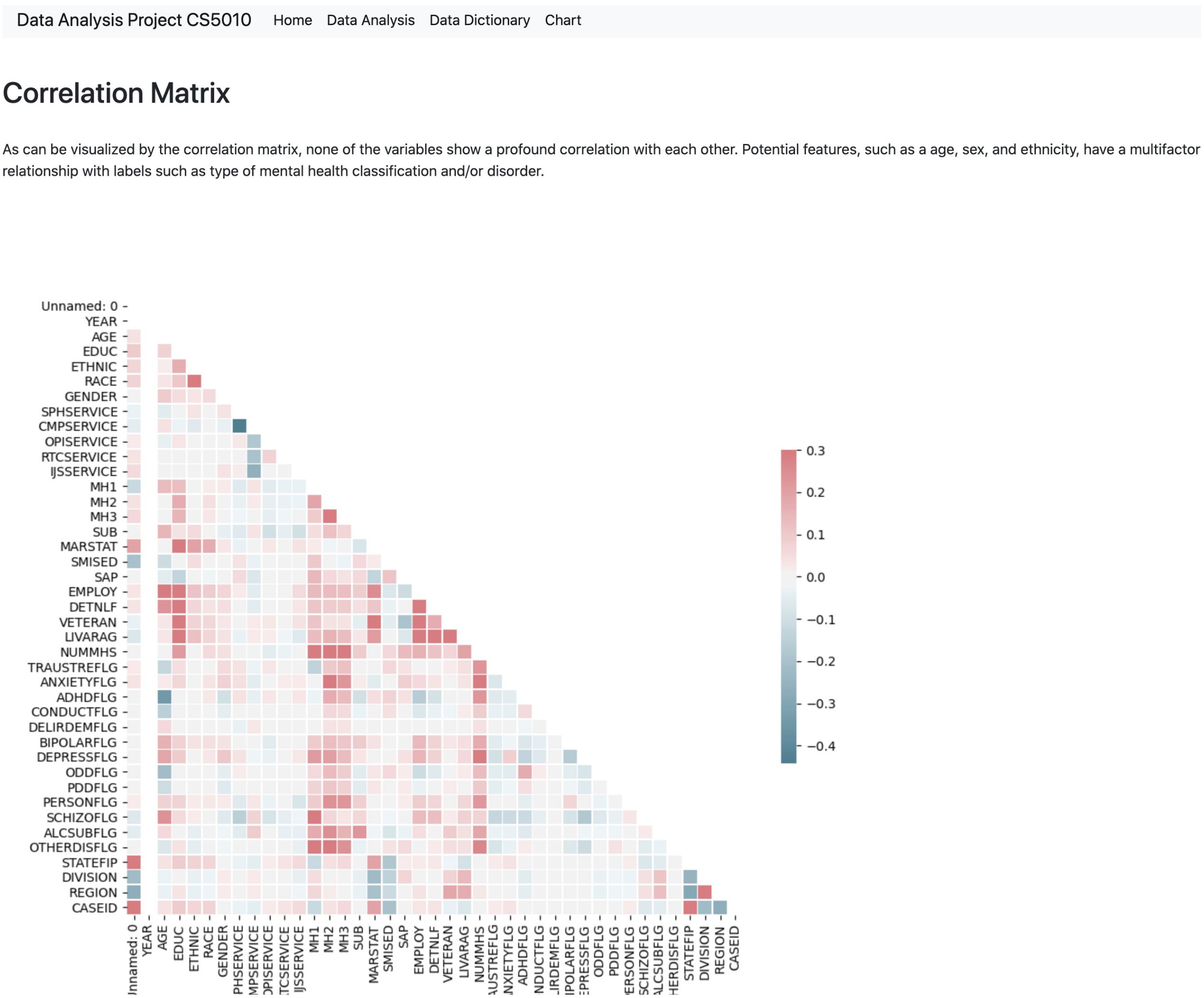
Figure 13. Keras Multilabel Classification Prediction Results

Findings

The findings from the machine learning model revealed intuitive trends behind the modeling output. Specifically, as the number of features w as increased, there w as also an increase in the accuracy of the output. Furthermore, demographic variables being added, such as age, race and sex, had a lesser effect on diagnosing as compared to variables such as substance abuse history, hospital admission, psychiatric w ard admissions, and . Of the models, Naive Bayes had the highest accuracy but this w as to be expected as it w as measuring a single label. In using four features for a single label, NB had an accuracy of around 90% compared to 35% for KNN. Comparatively, K-nearest neighbors and the multilabel classifier had low er accuracy scores, but could handle much more complex labels. This is important depending on the use case. If you w ant to determine if a patient has a particular diagnosis, then NB is a better classifer. How ever, if you are interested in determining if a patient has multiple diagnoses, this is a more complicated affair, as some disorders may be comorbid w ith others. Accordingly, KNN and multilabel classifiers are ideal in those cases.

Interactive Web Application

A w ebsite complete w ith user-interactive visualizations w as created to encourage engagement w ith the datasets and analyses. The choropleths show the raw percentages for the Health Care Quality measure performance of a given state w hen the user hovers their mouse over it. The interactive w eb-based application w as built in Flask. Flask is a Python w eb framew ork w hich implements the Jinja template engine to serve HTML files and the Wekzeug WGSI toolkit. We used three data visualization libraries as samples on this w ebsite, including seaborn, plotly express (to generate choropleth maps) and bokeh. Regarding the latter, w e imported historical data from the years of 2013-2018 to determine the frequency of primary, secondary and tertiary diagnoses over the years. Users could select the year from the dropdow n menu, prompting the redraw function to create a plot based on the attributes.



Data Analysis Project CS5010 Home Data Analysis Data Dictionary Chart

Correlation Matrix

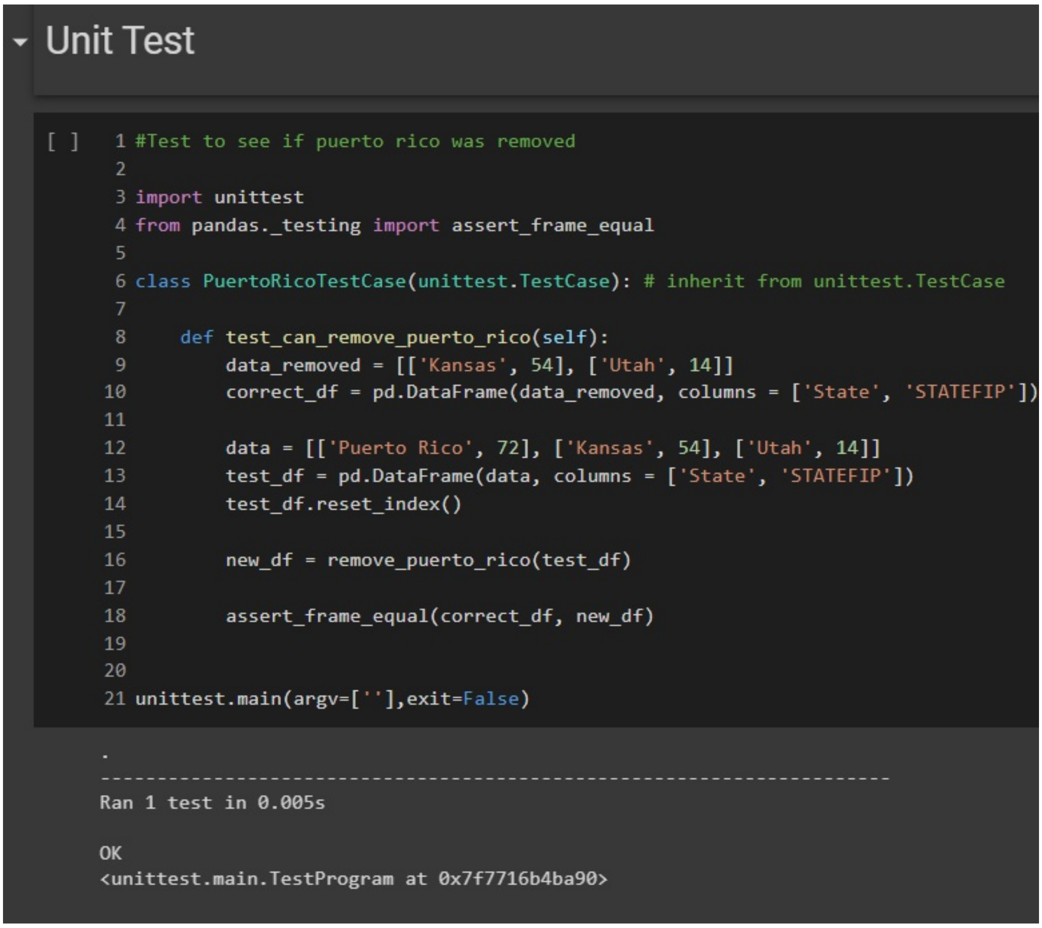
As can be visualized by the correlation matrix, none of the variables show a profound correlation with each other. Potential features, such as a age, sex, and ethnicity, have a multifactor relationship with labels such as type of mental health classification and/or disorder.

Figure 14. Website Sample

Unit Testing

MLH - CLD Dataset

The unit testing for this dataset w as of a function that w as w ritten to remove Puerto Rico from the dataset. Below is a picture of the results. A data frame w ith and w ithout Puerto Rico w as created and tested using the function to remove Puerto Rico.



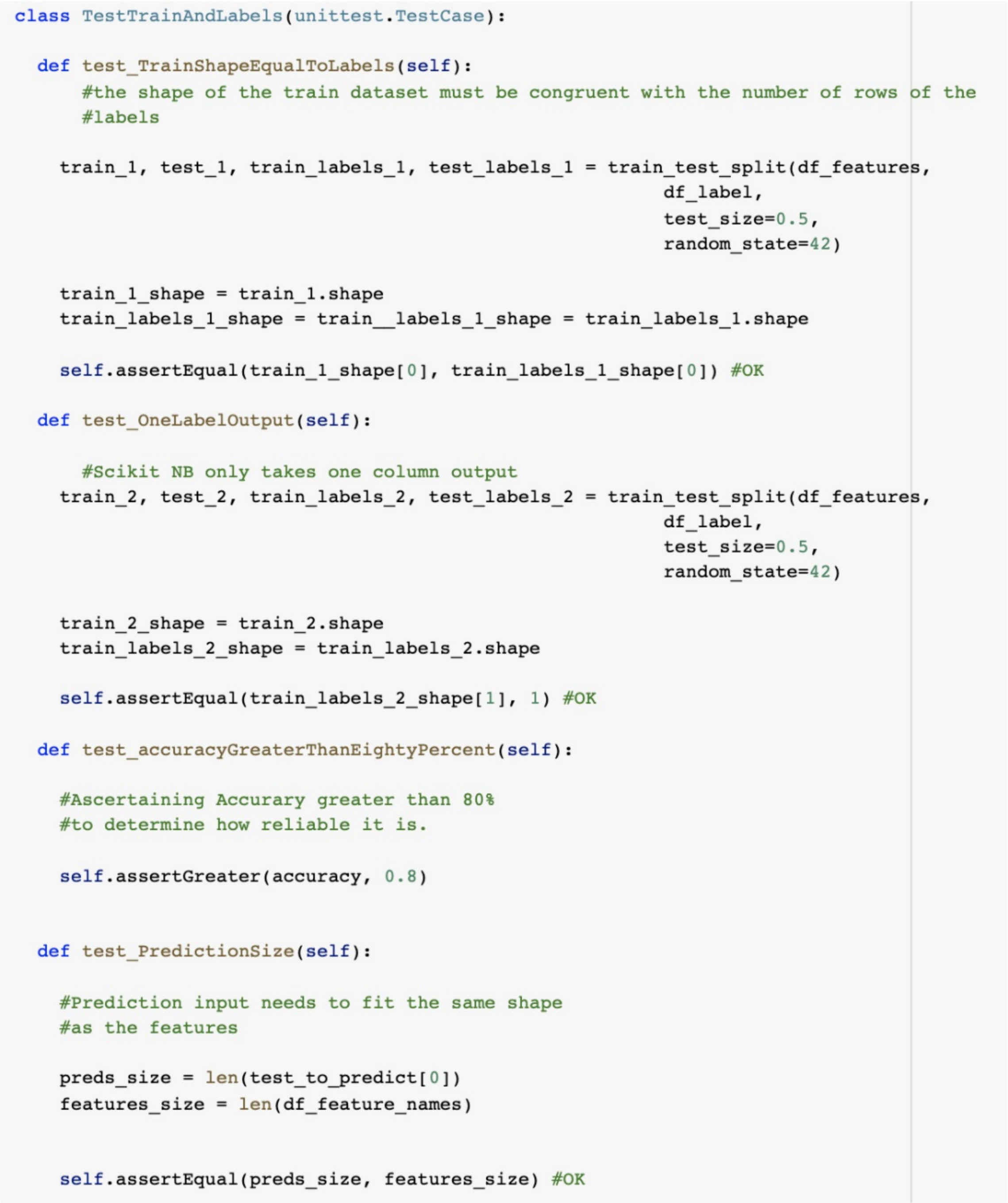
Unit Test

```
[ ]
#Test to see if puerto rico was removed

import unittest
from pandas._testing import assert_frame_equal

class PuertoRicoTestCase(unittest.TestCase): # inherit from unittest.TestCase

    def test_can_remove_puerto_rico(self):
        data_removed = [['Kansas', 54], ['Utah', 14]]
        correct_df = pd.DataFrame(data_removed, columns = ['State', 'STATEFIP'])

        data = [['Puerto Rico', 72], ['Kansas', 54], ['Utah', 14]]
        test_df = pd.DataFrame(data, columns = ['State', 'STATEFIP'])
        test_df.reset_index()

        new_df = remove_puerto_rico(test_df)

        assert_frame_equal(correct_df, new_df)


unittest.main(argv=[''],exit=False)
```

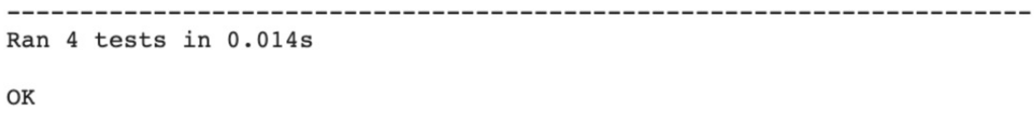
```
.
----------------------------------------------------------------------
Ran 1 test in 0.005s

OK
<unittest.main.TestProgram at 0x7f7716b4ba90>
```

Naive Bayes Classifier

Moreover, w hen using a NB classifier, the parameter sizes must be very precise in order for the model to run properly. Here, w e check the feature and labels to ensure they are of the proper dimensions. We also check if accuracy is above the 80% threshold.



```
class TestTrainAndLabels(unittest.TestCase):

  def test_TrainShapeEqualToLabels(self):
      #the shape of the train dataset must be congruent with the number of rows of the
      #labels

    train_1, test_1, train_labels_1, test_labels_1 = train_test_split(df_features,
                                                                      df_label,
                                                                      test_size=0.5,
                                                                      random_state=42)

    train_1_shape = train_1.shape
    train_labels_1_shape = train__labels_1_shape = train_labels_1.shape

    self.assertEqual(train_1_shape[0], train_labels_1_shape[0]) #OK

  def test_OneLabelOutput(self):

      #Scikit NB only takes one column output
    train_2, test_2, train_labels_2, test_labels_2 = train_test_split(df_features,
                                                                      df_label,
                                                                      test_size=0.5,
                                                                      random_state=42)

    train_2_shape = train_2.shape
    train_labels_2_shape = train_labels_2.shape

    self.assertEqual(train_labels_2_shape[1], 1) #OK

  def test_accuracyGreaterThanEightyPercent(self):

    #Ascertaining Accurary greater than 80%
    #to determine how reliable it is.

    self.assertGreater(accuracy, 0.8)


  def test_PredictionSize(self):

    #Prediction input needs to fit the same shape
    #as the features

    preds_size = len(test_to_predict[0])
    features_size = len(df_feature_names)


    self.assertEqual(preds_size, features_size) #OK
```



```
----------------------------------------------------------------------
Ran 4 tests in 0.014s

OK
```

Conclusions

Using age and mental health diagnosis to explore the mental health dataset allowed us to gain an understanding of who is receiving treatment and which diagnosis is the most prevalent. The age group 0-11 years had the highest number of cases overall and with most of the states. Most of these cases were in the ADD/ADHD and Trauma/Stressor category. As the age increased, the diagnosis with the higher number of cases changed over to depression To increase the understanding and awareness of mental health, adding more variables into the data set and looking more at how socioeconomic conditions impact the type of diagnosis received. Additionally, while states varied a lot in terms of their ability to provide follow-up care after hospitalization for mental illness, we were not able to find a correlation with treatment outcomes. In terms of machine learning models, additional testing can be conducted, such as investigating precision and recall scores. Moreover, investigation into optimizing model parameters is warranted, such as number of features and identity of features, number of neighbors for KNN, number of epochs and layers for the keras multilabel classification model, in addition to researching other models.

Future research opportunities include further time series analyses to investigate whether the Covid-19 global pandemic affected any of the client-level and temporal trends identified during exploratory data and trending analyses. Minority oversampling and/or majority undersampling methods could be used to decrease the bias introduced by the unequal sample sizes for the target variable levels for the logisitc regression model. Additional parameter adjustment and feature engineering may also boost the predictive ability of the machine learning models.

Works Cited

Ayers, J. W., Althouse, B. M., Allem, J. P., Rosenquist, J. N., & Ford, D. E. (2013). Seasonality in seeking mental health information on Google. *American journal of preventive medicine*, 44(5), 520-525. https://www.sciencedirect.com/science/article/abs/pii/S0749379713000809

Bedre, R. (2021). "Performing and visualizing the Principal component analysis (PCA) from PCA function and scratch in Python". *Renesh Bedre Data Science Blog*. https://www.reneshbedre.com/blog/principal-component-analysis.html

Bokeh Development Team (2021). Bokeh: Python library for interactive visualization. https://bokeh.org

Chollet, F., & others. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras

Grinberg, M. (2018). *Flask web development: developing web applications with python.* " O'Reilly Media, Inc."

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hunter, J. D., "Matplotlib: A 2D Graphics Environment," in *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90-95, May-June 2007, doi: 10.1109/MCSE.2007.55.

The Kaiser Family Foundation State Health Facts, Poverty Rate by Race/Ethnicity. Data Source: KFF estimates based on the 2008-2019 American Community Survey (United States Census Bureau), 1-Year Estimates.https://www.kff.org/2d5cbf8/

Kluyver, T., Ragan-Kelley, B., Fernando P'erez, Granger, B., Bussonnier, M., Frederic, J., … Willing, C. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87–90).

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825–2830.

Plotly Technologies Inc (2015). Collaborative data science. Montreal, QC: Plotly Technologies Inc. Retrieved from https://plot.ly

Stojiljković, M. "Logistic Regression in Python." *Real Python*, Real Python, 24 Nov. 2020, realpython.com/logistic-regression-python/#logistic-regression-in-python-with-scikit-learn-example-1.

Story, R., Fernandez, F. us-states.json. folium, 2016. https://raw.githubusercontent.com/python-visualization/folium/master/examples/data/us-states.json

Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality. *Mental Health Client-Level Data 2018*. Rockville, MD: Substance Abuse and Mental Health Services Administration, 2020. https://www.datafiles.samhsa.gov/sites/default/files/field-uploads-protected/studies/MH-CLD-2018/MH-CLD-2018-datasets/MH-CLD-2018-DS0001/MH-CLD-2018-DS0001-info/MH-CLD-2018-DS0001-info-codebook.pdf

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., … Qalieh, A. (2017). *mwaskom/seaborn: v0.8.1 (September 2017)*. Zenodo. https://doi.org/10.5281/zenodo.883859

World Population Review. (n.d.). List of STATE ABBREVIATIONS (DOWNLOAD CSV, JSON). https://worldpopulationreview.com/states/state-abbreviations.