Mapping Community Need for Mental Health Facilities

Frances Hung, Cheryl Yau, Candice Wang

12/09/2017

StoryMap Presentation

https://arcg.is/1fDKLD All the map visualizations can be found here.

Motivation

According to the CDC, suicide was the 10th leading cause of death in the US in 2015, and the 2nd leading cause of death among adolescents and young adults. Psychological disorders, particularly depression, are a significant risk factor for suicide especially when they go untreated. There is no reliable way to predict who is at risk for committing suicide, because most screening approaches depend on self-report information and people contemplating on suicide would often deny it when asked.

In the first part of our project, we hence aim to build a logistic regression model to identify important variables in predicting suicide rates. Due to the limits of our data, we consider the period from 2004-2015, within the scope of cities in California. One interesting explanatory variable we use is Google search term data (under the product of "Google Trends"). Our hypothesis is that individuals are more likely to tell the truth to Google, than on a questionnaire. In the second part of our project, we build a series of maps using the ArcGIS software. Using suicide rate and mental health treatment facilities data as well as Google search term data, our project aims to map the demand for and supply of mental health treatment in California cities.

Ultimately, we hope to shed some light on important explanatory variables correlated with suicide rates (with the regression model), and to help identify cities where there is a large treatment service gap (with the maps) so that we can address this problem in a more data-driven way.

Variable choice

Ideally, the response variable that we are interested in is the gap between the demand and supply of mental health treatment. Which areas are over/under-served, and why? This would be very useful information to policy makers, mental health service providers, related non-profits and such. However, such a variable does not exist (or we could not find it), and we would have had to create an algorithm to derive this data from other existing variables. We could not decide on an accurate way to code "demand" (and what weights to give each component). Furthermore, even though "supply" is more straightforward, there also exists discrepancies between the size of the facilities, or the affordability of the services that would need to be captured by our variable. In the end, we decided that we would use suicide rate as a response variable, although we agreed that it would be an interesting extension to look at service gap. We also hope that our GIS maps would help our audience to begin to think about and identify areas which are under-served.

Regression Model

We originally intended to look at suicide rate and Google Trends data from one year, eg. 2015, but the logit model returned no significant variables as both suicide rate and depression search fluctuate a lot each year, influenced by factors like celebrity suicides which are not directly relevant to population mental health. Hence, we decided to aggregate suicide rate and Google Trends data over 12 years (constrained by data availability), from 2004 to 2015. Since demographic information is fairly stable over time, we used demographic information from the most recent year to train our model.

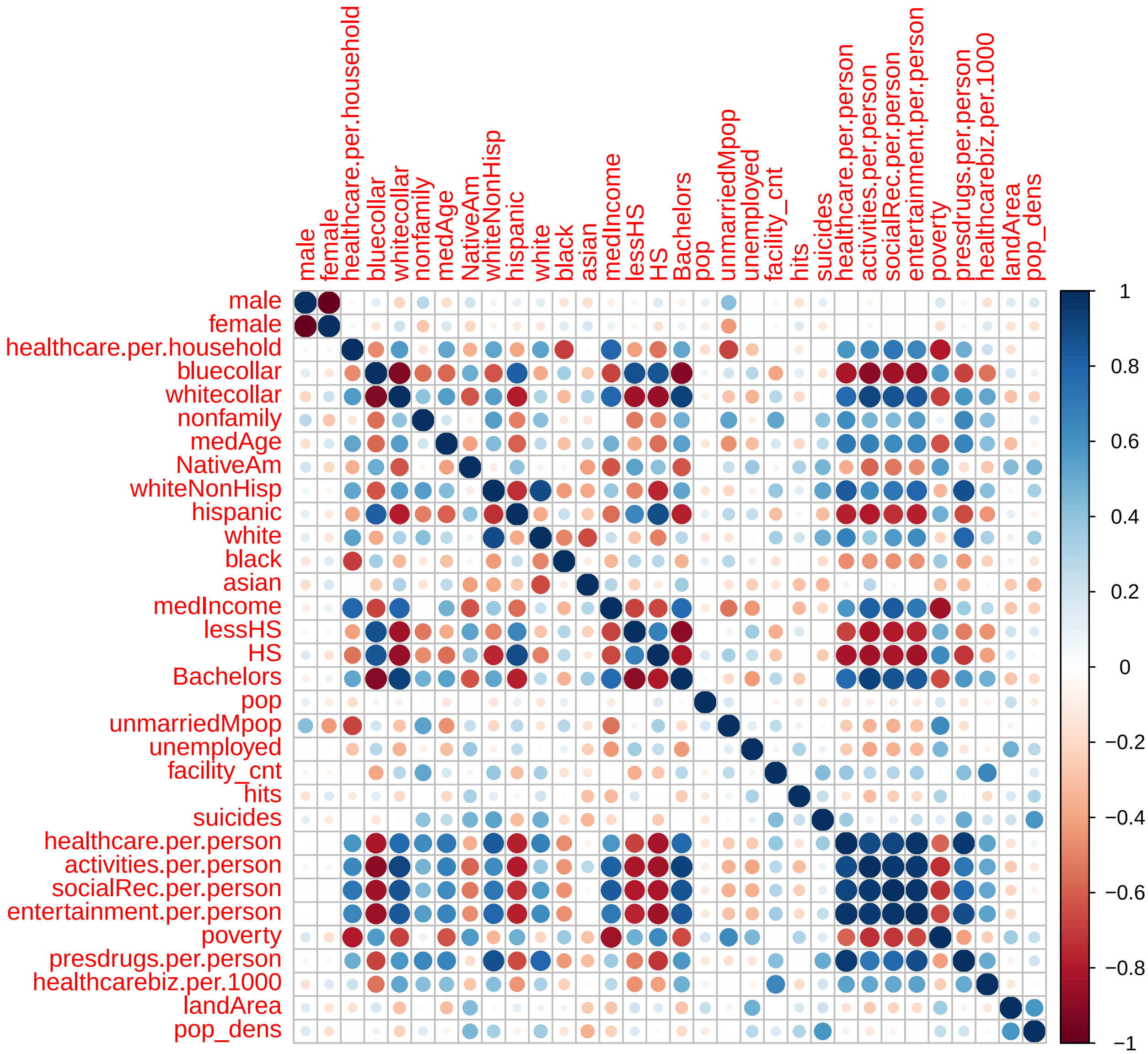
Making Dataframes

- Longitude and latitude of cities (for mapping them later)
- Google search frequency by city on "depression" as a mood (to exclude unrelated searches on economic depression etc) from 2004-2015 (downloaded from Google Trends). The "hits" values are calculated on a scale from 0 to 100, where 100 is the location with the most popularity as a fraction of total searches, where 50 indicates a location which is half as popular and so on.
- List of verified mental health treatment clinics and facilities (downloaded from ReferenceUSA). We only included places with a certified psychiatrist or psychologist, and which focuses on general mental health (excluding substance abuse facilities)
- Number of suicides by zipcode from 2004-2015. We downloaded leading causes of death data from California Health and Human Services Agency and filtered for cause of death is suicide.
- Demographic data downloaded from SimplyAnalytics, including racial and gender makeup, age, marriage, education level, employment, income, healthcare, etc.

Satisfying Conditions for Linear Regression

We use linear regression to find a linear relationship between explanatory variables and our scalar response variable (cumulative suicide rate per 100,000 people). Before we settle on a final model, there are some conditions on our data we have to ensure. Most conditions, like mean of residuals is 0 and homoscedasticity (equal variance) and normality of residuals are easy to check (for example, we can plot our linear models in R to get Q-Q plots). One of these conditions, however(no multicollinearity), must be checked and corrected for more vigorously.

If we want to reliably determine significant variables, we want to ensure that explanatory variables aren't collinear (or significantly correlated). Looking at the correlation plot of variables in our data, we see significant correlation between some variables.



We can use VIFs (Variance Inflation Factors), which measure how much the variance of a variable's coefficient changes if predictors in a model are correlated, to determine which variables to remove. We first split our data into test and training data and make a model with all variables to see the VIFs we start with.

```
##
## Call:
## lm(formula = (suicides) ~ ., data = logtable.train[, -1])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.214 -16.294  -4.159  17.321 172.363
##
## Coefficients: (1 not defined because of singularities)
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              7.536e+02  5.035e+02   1.497  0.13757
## male                     9.091e-01  4.389e+00   0.207  0.83632
## female                          NA         NA      NA       NA
## healthcare.per.household -1.396e-01  6.408e-02  -2.178  0.03170 *
```

```
## bluecollar                2.296e+00  1.931e+00   1.189  0.23720
## whitecollar              -3.063e-02  1.683e+00  -0.018  0.98552
## nonfamily                -3.399e+00  2.413e+00  -1.408  0.16215
## medAge                    1.088e+01  4.230e+00   2.571  0.01158 *
## NativeAm                  4.232e+01  1.291e+01   3.278  0.00143 **
## whiteNonHisp             -2.915e+00  4.520e+00  -0.645  0.52043
## hispanic                 -4.257e+00  4.059e+00  -1.049  0.29678
## white                    -3.617e-01  2.297e+00  -0.157  0.87519
## black                    -1.073e+01  4.793e+00  -2.238  0.02744 *
## asian                    -4.127e+00  4.072e+00  -1.014  0.31320
## medIncome                 1.144e-03  1.241e-03   0.922  0.35867
## lessHS                   -1.162e+00  2.035e+00  -0.571  0.56920
## HS                       -3.162e+00  1.875e+00  -1.686  0.09480 .
## Bachelors                -1.824e+00  1.890e+00  -0.965  0.33670
## pop                      -8.483e-06  1.041e-05  -0.815  0.41718
## unmarriedMpop             6.933e-01  2.076e+00   0.334  0.73913
## unemployed               -9.630e+00  4.074e+00  -2.364  0.02001 *
## facility_cnt              1.790e+00  6.364e-01   2.813  0.00591 **
## hits                      1.193e-01  5.964e-01   0.200  0.84181
## healthcare.per.person    -1.937e+03  1.045e+03  -1.853  0.06674 .
## activities.per.person     1.694e+04  1.631e+04   1.039  0.30131
## socialRec.per.person     -1.085e+04  5.345e+03  -2.030  0.04501 *
## entertainment.per.person  2.026e+03  1.105e+03   1.834  0.06955 .
## poverty                   2.938e+00  1.786e+00   1.645  0.10315
## presdrugs.per.person      1.232e+04  7.479e+03   1.648  0.10251
## healthcarebiz.per.1000   -1.988e+00  2.125e+00  -0.936  0.35175
## landArea                  1.710e-02  4.771e-02   0.358  0.72074
## pop_dens                  1.314e+04  4.349e+03   3.021  0.00319 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.51 on 101 degrees of freedom
## Multiple R-squared:  0.8027, Adjusted R-squared:  0.7441
## F-statistic:  13.7 on 30 and 101 DF,  p-value: < 2.2e-16
```

This model, when tested on the test data, yields a correlation rate of about 81%.

```
##          actuals predicts
## actuals  1.000000 0.810071
## predicts 0.810071 1.000000
```

To determine which variables to remove, we choose variables with the highest VIFs to discount in the final analysis. From the initial model, we remove variables one at a time, testing to see how the predictions and VIFs of our model change.

```
##                     male healthcare.per.household               bluecollar
##                   3.2352                  65.8630                  19.6030
##              whitecollar                nonfamily                   medAge
##                  38.0420                  47.9950                  41.8320
##                 NativeAm             whiteNonHisp                 hispanic
##                   3.9585                 892.5400                 536.5900
##                    white                    black                    asian
##                 140.6400                  85.7240                 371.9200
##                medIncome                   lessHS                       HS
##                  96.6070                  14.5700                  27.8470
##                Bachelors                      pop            unmarriedMpop
```

```
##                 23.8280                1.4132               15.3710
##              unemployed          facility_cnt                  hits
##                  2.2086                2.8123                2.2004
##    healthcare.per.person  activities.per.person  socialRec.per.person
##              10731.0000             1339.9000             2361.4000
## entertainment.per.person               poverty  presdrugs.per.person
##               9629.7000               13.0250             2007.1000
##   healthcarebiz.per.1000              landArea              pop_dens
##                  3.3188                3.2946                2.7977
```

We decide to remove activity, entertainment, social recreation, and healthcare spending per person, female, white/blue collar, white(non-Hispanic), Asian, black, and white population, nonfamily households, median income, % Bachelor's/high school degrees, and healthcare spending per household from the explanatory variables.

We rebuild our model using the remaining variables.

```
##
## Call:
## lm(formula = (suicides) ~ ., data = logtable_crop.train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -96.22 -19.32   0.50  19.34 212.68
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -1.603e+02  1.691e+02  -0.948 0.345158
## male                   5.583e-01  3.256e+00   0.171 0.864157
## medAge                 2.733e+00  1.173e+00   2.329 0.021577 *
## NativeAm               5.668e+01  1.043e+01   5.434 3.08e-07 ***
## hispanic              -1.135e+00  2.994e-01  -3.792 0.000239 ***
## lessHS                 1.951e+00  9.432e-01   2.069 0.040797 *
## pop                   -3.325e-06  1.007e-05  -0.330 0.741744
## unmarriedMpop         -9.158e-01  9.613e-01  -0.953 0.342741
## unemployed            -7.551e+00  3.523e+00  -2.144 0.034158 *
## facility_cnt           2.072e+00  6.073e-01   3.412 0.000890 ***
## hits                   1.930e-01  5.003e-01   0.386 0.700319
## poverty                3.146e+00  1.010e+00   3.115 0.002319 **
## presdrugs.per.person   1.072e+03  3.011e+02   3.561 0.000537 ***
## healthcarebiz.per.1000 -2.372e+00  1.902e+00  -1.247 0.214828
## landArea              -5.547e-02  3.980e-02  -1.394 0.166042
## pop_dens               1.777e+04  3.887e+03   4.571 1.22e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.89 on 116 degrees of freedom
## Multiple R-squared:  0.7688, Adjusted R-squared:  0.7389
## F-statistic: 25.72 on 15 and 116 DF,  p-value: < 2.2e-16
```

Significant variables include: Facility density (#mental health facilities/100,000 people) (+), % population Native American(+), % population Hispanic (-), population density (+), median age (+), prescription drug spending per person (+), unemployment rate (-), and % population in poverty (+).

We test our model on the same test data and get a 84.6% correlation rate, which is better than our original starting model. Not only is our prediction accuracy better in this case, but we're also more sure about which variables are significant and what their coefficients are.

```
##            actuals  predicts
## actuals  1.0000000 0.8463398
## predicts 0.8463398 1.0000000
```

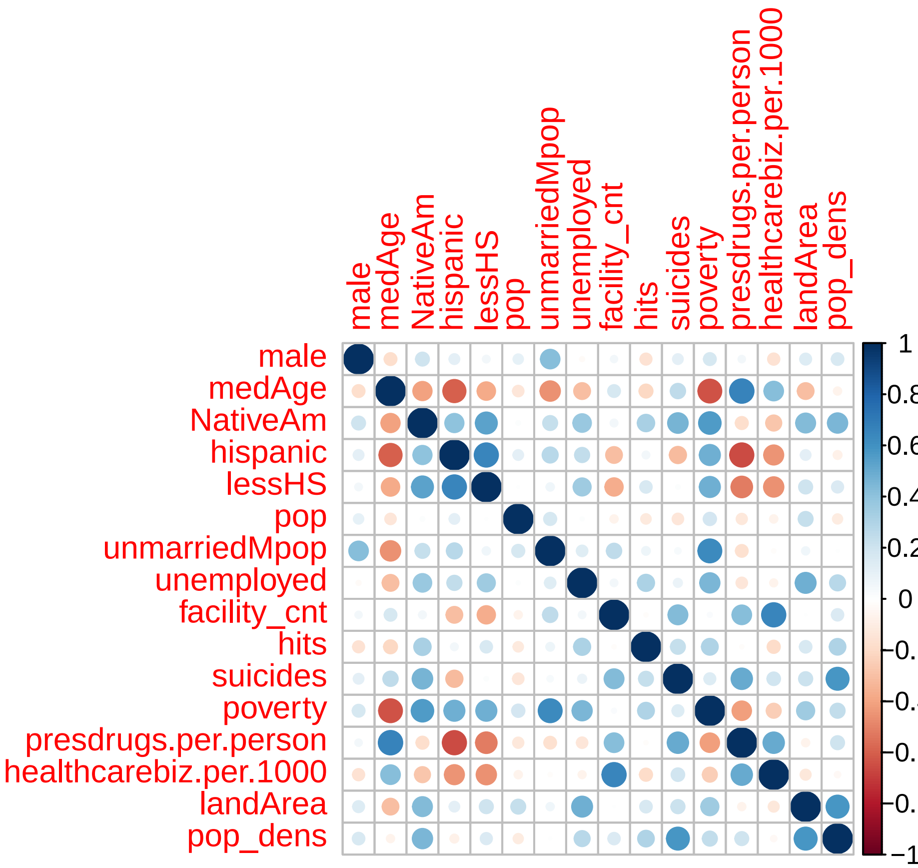
```
##                         actuals  predicts
## San Luis Obispo     176.09694 211.93274
## Atascadero          177.69731 235.55283
## Salinas             109.72449  81.39655
## San Clemente        141.36484 132.78542
## Victorville         135.03007 139.74088
## Redwood City        167.21796 147.35644
## Santa Barbara       250.22724 203.42249
## Gilroy              102.12380 162.49813
## Ontario              75.96052  59.83799
## Livermore           113.60441 162.85641
## Chico               209.40649 246.89824
## Baldwin Park         51.24001  45.35739
## Norwalk              65.21300  88.64938
## Yuba City           157.56539 182.82644
## Fresno              121.46856 156.72774
## South San Francisco  88.13032  93.33945
## Fullerton           120.71311 103.03246
## La Habra            108.84311 110.84235
## Thousand Oaks        93.08074 106.54687
## Lodi                140.70592 170.69612
## Morro Bay           248.27586 315.47898
## Rancho Cordova      115.02847 150.92125
## Turlock             126.77240 107.78019
## Brea                 98.97405 102.24197
## Modesto             160.92483 148.98047
## Richmond             77.26946  97.70358
## Pasadena             99.78980 130.44244
## Moreno Valley        65.46074  76.41203
## San Mateo           132.34030 135.79693
## Compton              73.44360  54.86507
## Whittier            175.52601 124.18885
## Camarillo           167.09797 139.97544
## Chula Vista          78.18340  64.76334
## Santee              144.13225 167.19731
## Novato              174.58814 160.57698
## Montebello           68.01039  87.50931
## San Gabriel         118.80228  97.57707
## Aliso Viejo         112.76187  65.60654
## Carson               56.58717  59.15905
## Redlands            110.27486 141.21389
## Glendale             73.78545 125.86732
## Albany              132.82826 142.75687
```

All VIFs for our explanatory variables are acceptable (< 5).

```
##                  male                 medAge               NativeAm
##                1.7451                 3.1546                 2.5318
##              hispanic                 lessHS                    pop
##                2.8619                 3.0688                 1.2945
```

```
##          unmarriedMpop               unemployed          facility_cnt
##                 3.2291                   1.6184                2.5106
##                   hits                  poverty  presdrugs.per.person
##                 1.5176                   4.0809                3.1879
## healthcarebiz.per.1000                 landArea              pop_dens
##                 2.6069                   2.2475                2.1896
```

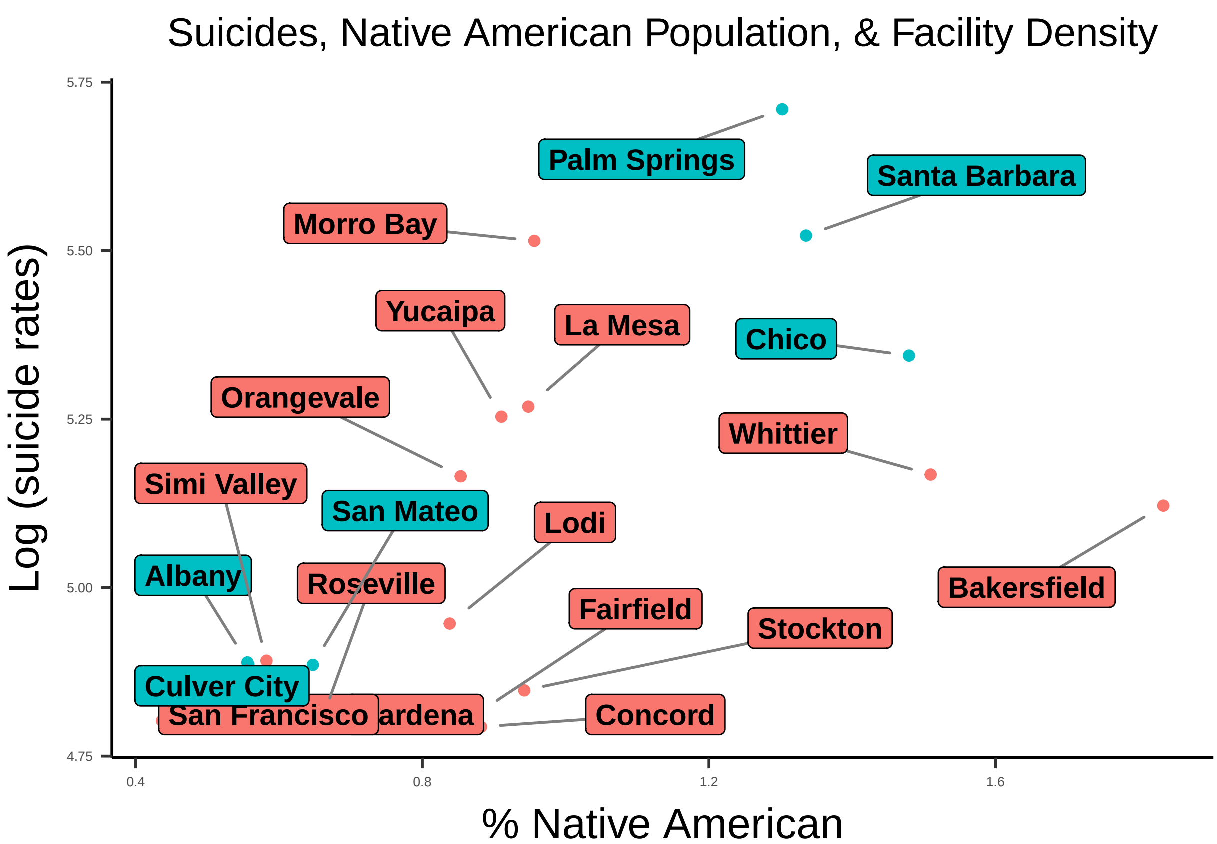
This is supported by our correlation plot, which shows that our variables are a lot less correlated than in the original model.



Visualizing Explanatory Variables

We can visualize explanatory variables two at a time, one on the x-axis and one as a color variable, plotted against log(suicides) on the y-axis. We're interested in the cities with higher suicide rates, so we sample cities from the cities with suicide rates higher than the 50th percentile. In this case, we've plotted the % Native American population on the x-axis and colored the points by facility density (if it's more than 70th percentile, we color it teal; otherwise, it's red).

To explore variable comparisons interactively, check out this Shiny app! https://frances-hung.shinyapps.io/ma154-project5-teammentalhealth/



Mapping in ArcGIS

We made two GIS maps on the state level. http://arcg.is/4Tza5 The first one shows suicide rate for each zipcode normalized by population and locations of mental health facilities. We can see that regions with the highest suicide rates have no nearby facilities that serve them.

The second map plots the depression Google Trends data for different cities on a layer of facilities density. http://arcg.is/1PvOHf We can see that while the facilities are concentrated in coastal metropolitan areas, the high search frequency cities are scattered across the state.

Comparative city-level maps for Inglewood and Santa Barbara with four significant variables (African American population, healthcare spending, health/social club spending, and population density) and facilities locations are embedded in the StoryMap presentation. They can also be found here (toggle the layers to see different variables). https://services.arcgis.com/hVnyNvwbpFFPDV5j/arcgis/rest/services/InglewoodandSantaBabara/FeatureServer

From these maps, we can see that only one of the four significant variables is correlated with suicide rate in the direction suggested by the regression model, if we only compare two cities. Furthermore, facilities in both cities are concentrated in areas with relatively higher population density. This means that while the regression model gives us a generalized view of the bigger picture, maps on a local level provide an additional level of nuance. In the end, both kinds of information could be useful for policy making and bringing mental health facilities to underserved areas.