Playground

Frances Hung

10/21/2017

StoryMap

https://arcg.is/1fDKLD

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.

- use result/visualization as hook

Variable choice (to be moved to preceding corresponding R chunks)

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.

The original datasets we start with include: - 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) - 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. - 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. - Cities long lat data (if possible, could we find a dataset which has a more exhaustive list, or I could ask Warren..)

Playground

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.

```
## Loading required package: gtrendsR
```

```
## Loading required package: ggplot2
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: zipcode
```

```
## Loading required package: ggmap
```

```
## Loading required package: caret
```

```
## Loading required package: lattice
```

Making Dataframes

Longitude and latitude of cities (for mapping them later)

Full gtrends data for all cities for depression search

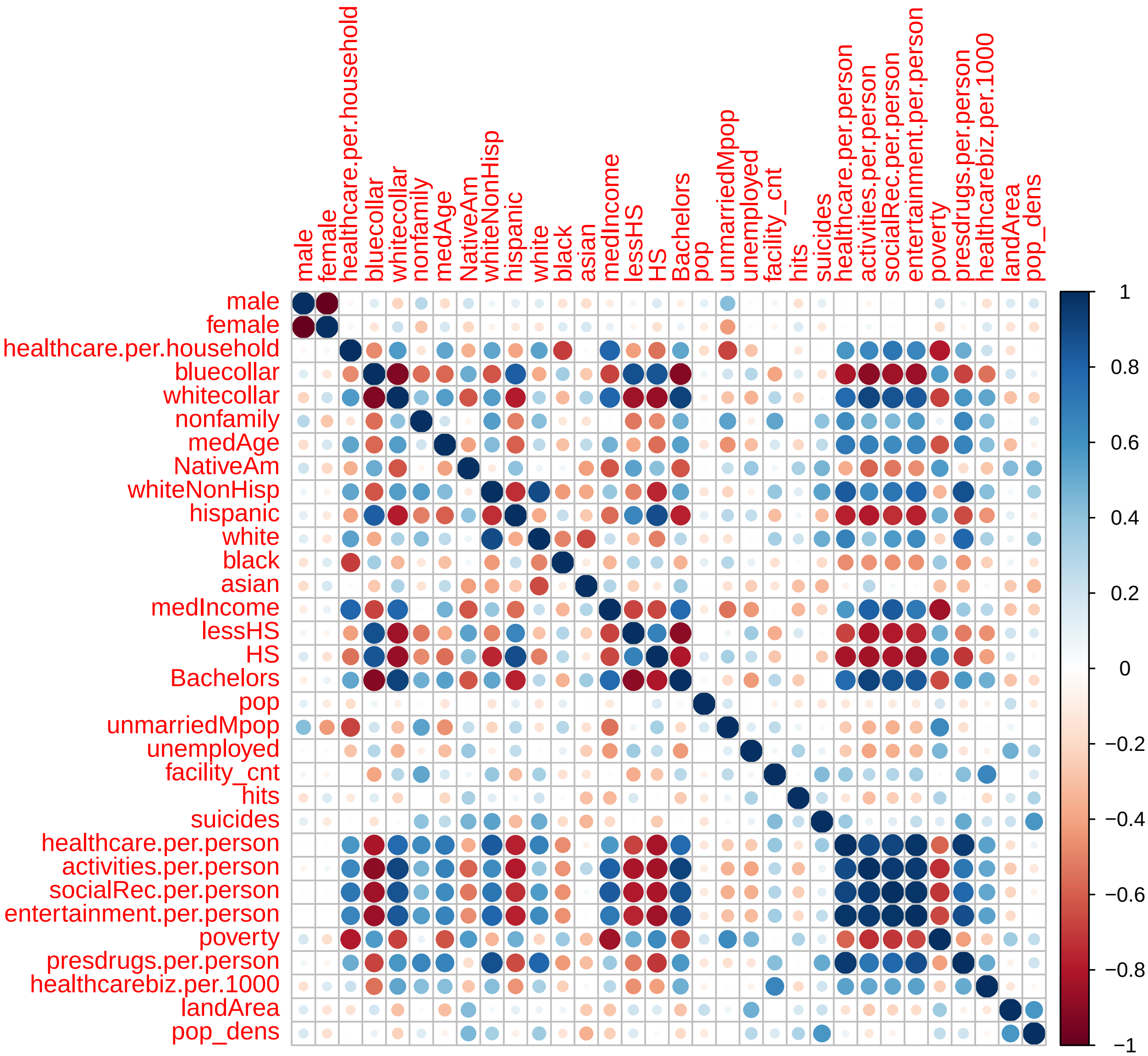
Prepping facilities data to find number of facilities per city

Prepping suicide data to find number of suicides per city, 2004-2015

Adding in city demographic data for 2017

There are two different models we can use. If we want to reliably determine significant variables, we want to ensure that variables aren't collinear. Looking at the correlation plot of variables in logtable, we see significant correlation between some variables.

```
## corrplot 0.84 loaded
```



To determine which variables to remove, we look at the VIF and choose variables with the highest VIFs to discount in the final analysis. We use a linear model using all variables to look at the initial VIF, and then remove variables one at a time, testing to see how the predictions and VIFs of our model change.

```
##
## Call:
## lm(formula = (suicides) ~ ., data = logtable.train[, -1])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.510 -17.623  -1.025  16.305 187.334
##
## Coefficients: (1 not defined because of singularities)
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              6.666e+02  5.453e+02   1.222 0.224383
## male                     2.328e+00  4.652e+00   0.500 0.617880
## female                          NA         NA      NA       NA
```

```
## healthcare.per.household -1.103e-01  6.666e-02  -1.655 0.101092
## bluecollar               -9.259e-02  2.010e+00  -0.046 0.963346
## whitecollar              -1.016e+00  1.782e+00  -0.570 0.569879
## nonfamily                -5.401e+00  2.463e+00  -2.193 0.030606 *
## medAge                    7.686e+00  4.367e+00   1.760 0.081439 .
## NativeAm                  4.481e+01  1.282e+01   3.495 0.000707 ***
## whiteNonHisp             -5.516e+00  4.647e+00  -1.187 0.238010
## hispanic                 -6.273e+00  4.135e+00  -1.517 0.132393
## white                     6.575e-01  2.258e+00   0.291 0.771506
## black                    -1.018e+01  4.896e+00  -2.080 0.040045 *
## asian                    -3.916e+00  4.025e+00  -0.973 0.332933
## medIncome                 5.430e-04  1.254e-03   0.433 0.665882
## lessHS                    1.182e+00  2.212e+00   0.534 0.594225
## HS                       -2.008e+00  1.933e+00  -1.039 0.301275
## Bachelors                -9.811e-01  1.818e+00  -0.540 0.590540
## pop                      -1.308e-05  1.093e-05  -1.197 0.234191
## unmarriedMpop             2.508e+00  2.168e+00   1.157 0.249892
## unemployed               -1.028e+01  3.914e+00  -2.625 0.009998 **
## facility_cnt              1.722e+00  6.459e-01   2.666 0.008931 **
## hits                      3.560e-01  7.119e-01   0.500 0.618094
## healthcare.per.person    -2.166e+03  1.192e+03  -1.816 0.072279 .
## activities.per.person     1.462e+04  1.795e+04   0.815 0.417065
## socialRec.per.person     -1.136e+04  5.706e+03  -1.991 0.049179 *
## entertainment.per.person  2.423e+03  1.172e+03   2.068 0.041221 *
## poverty                   2.676e+00  1.934e+00   1.383 0.169633
## presdrugs.per.person      1.420e+04  8.560e+03   1.659 0.100157
## healthcarebiz.per.1000   -2.121e+00  2.284e+00  -0.929 0.355096
## landArea                  3.094e-02  5.131e-02   0.603 0.547895
## pop_dens                  1.374e+04  4.671e+03   2.943 0.004037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.61 on 101 degrees of freedom
## Multiple R-squared:  0.7919, Adjusted R-squared:  0.7301
## F-statistic: 12.81 on 30 and 101 DF,  p-value: < 2.2e-16
```

```
##            actuals  predicts
## actuals  1.0000000 0.8122788
## predicts 0.8122788 1.0000000
```

```
##                     male healthcare.per.household               bluecollar
##                   2.5820                  59.5610                  20.3300
##              whitecollar                nonfamily                   medAge
##                  38.1090                  38.4450                  40.4380
##                 NativeAm             whiteNonHisp                 hispanic
##                   3.7325                 943.3100                 574.3500
##                    white                    black                    asian
##                 124.5400                  66.1160                 309.2700
##                medIncome                   lessHS                       HS
##                  84.5370                  15.9470                  27.8590
##                Bachelors                      pop            unmarriedMpop
##                  20.1410                   1.4374                  12.6550
##               unemployed             facility_cnt                     hits
##                   2.2957                   2.9691                   2.2113
##    healthcare.per.person    activities.per.person     socialRec.per.person
```

```
##               12023.0000                 1379.4000               2394.1000
## entertainment.per.person                   poverty    presdrugs.per.person
##                9516.4000                   15.4720               2235.7000
##   healthcarebiz.per.1000                  landArea                pop_dens
##                   3.6578                    3.2010                  2.9308
```

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 partition the data into a training and test set, then build our model.

```
##
## Call:
## lm(formula = (suicides) ~ ., data = logtable_crop.train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64.04 -18.12  -0.23  17.15 229.94
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.167e+02  1.587e+02  -1.365  0.17478
## male                   3.862e-01  3.078e+00   0.125  0.90036
## medAge                 2.937e+00  1.151e+00   2.553  0.01198 *
## NativeAm               3.283e+01  1.164e+01   2.821  0.00563 **
## hispanic              -6.389e-01  2.905e-01  -2.199  0.02985 *
## lessHS                 2.476e+00  8.882e-01   2.788  0.00621 **
## pop                   -9.833e-06  9.469e-06  -1.038  0.30122
## unmarriedMpop          5.818e-02  1.009e+00   0.058  0.95411
## unemployed            -8.471e+00  3.081e+00  -2.749  0.00693 **
## facility_cnt           1.889e+00  5.652e-01   3.342  0.00112 **
## hits                   3.948e-01  4.673e-01   0.845  0.39992
## poverty                2.422e+00  9.218e-01   2.628  0.00976 **
## presdrugs.per.person   1.375e+03  3.196e+02   4.303 3.54e-05 ***
## healthcarebiz.per.1000 -2.107e+00  1.762e+00  -1.196  0.23422
## landArea               5.347e-03  3.916e-02   0.137  0.89164
## pop_dens               1.413e+04  3.433e+03   4.117 7.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.09 on 116 degrees of freedom
## Multiple R-squared:  0.7183, Adjusted R-squared:  0.6819
## F-statistic: 19.72 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 (+), and % population in poverty (+). A borderline significant variable is density of healthcare businesses per 1000 people (-).

We test our model on the test data and get a 87% correlation rate, which is the better than our original model (which had an 81% correlaton rate).

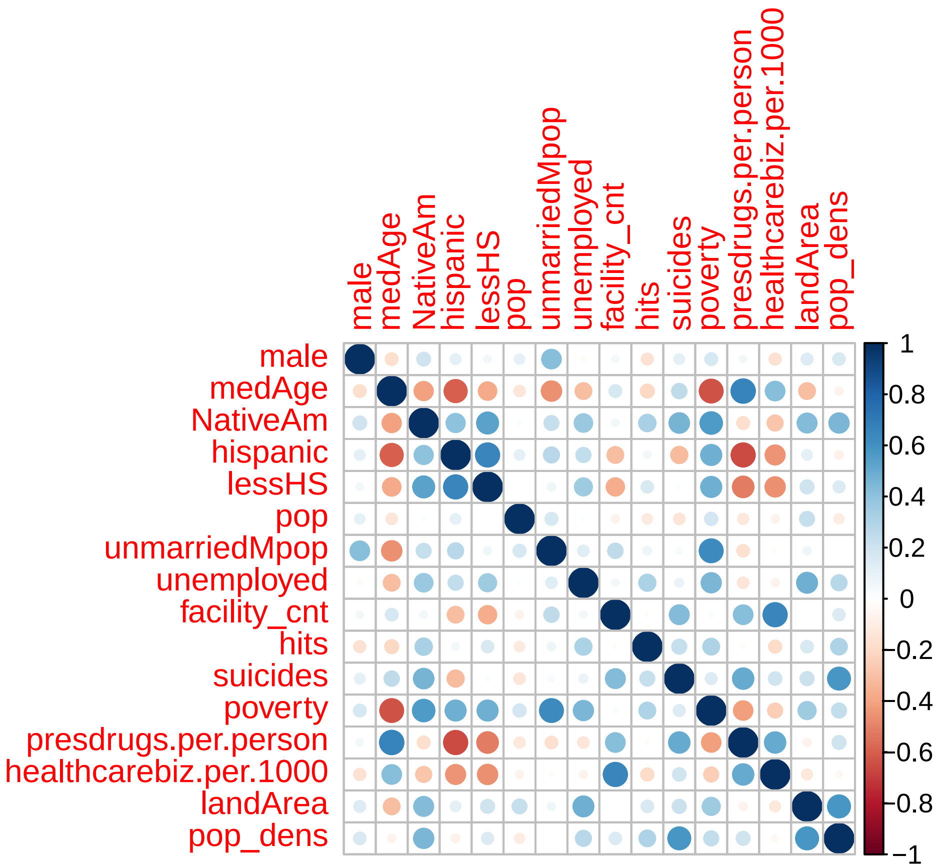
```
##            actuals  predicts
## actuals  1.0000000 0.8697342
## predicts 0.8697342 1.0000000
```

```
##                        actuals  predicts
## West Hollywood      123.11480 222.15449
## Eureka              572.02475 390.09065
## Mountain View        91.32077  88.96332
## Oxnard               93.82788  66.90143
## Atascadero          177.69731 224.43685
## San Francisco       121.80783 123.14461
## El Monte             63.92554  86.68493
## Lancaster           129.32528 182.44697
## Santa Cruz          264.95020 145.54499
## Arcata              260.37245 272.76136
## Anaheim              94.43216  88.83862
## Hesperia            140.22163 151.50400
## Fremont              74.99414  73.04211
## Chico               209.40649 235.68889
## Baldwin Park         51.24001  57.51560
## Norwalk              65.21300  92.35764
## Riverside           133.51537 106.63633
## Newport Beach       118.44538 183.73864
## South San Francisco  88.13032  93.19371
## Fullerton           120.71311 105.86615
## Fairfield           123.20074 121.14479
## Yucaipa             191.26789 160.89759
## La Habra            108.84311 121.77998
## Carlsbad            130.47235 135.57915
## Campbell            143.17568 142.95780
## Yorba Linda         101.50044 104.45694
## Modesto             160.92483 146.35437
## Pasadena             99.78980 135.68026
## San Mateo           132.34030 139.38424
## Fountain Valley     114.74764 106.40660
## Castro Valley       106.95346 142.30844
## Downey               70.44642  90.73334
## Camarillo           167.09797 148.00358
## West Covina          80.81285 106.90381
## Montebello           68.01039 102.60970
## Cerritos             78.59009  99.94798
## Arcadia             115.38991  90.02817
## Inglewood            67.05665  72.13017
## Palm Desert         225.44525 239.02388
## Carmichael          205.11892 182.35737
## Alhambra             85.69576  89.38788
## Walnut Creek        173.09701 185.71432
```

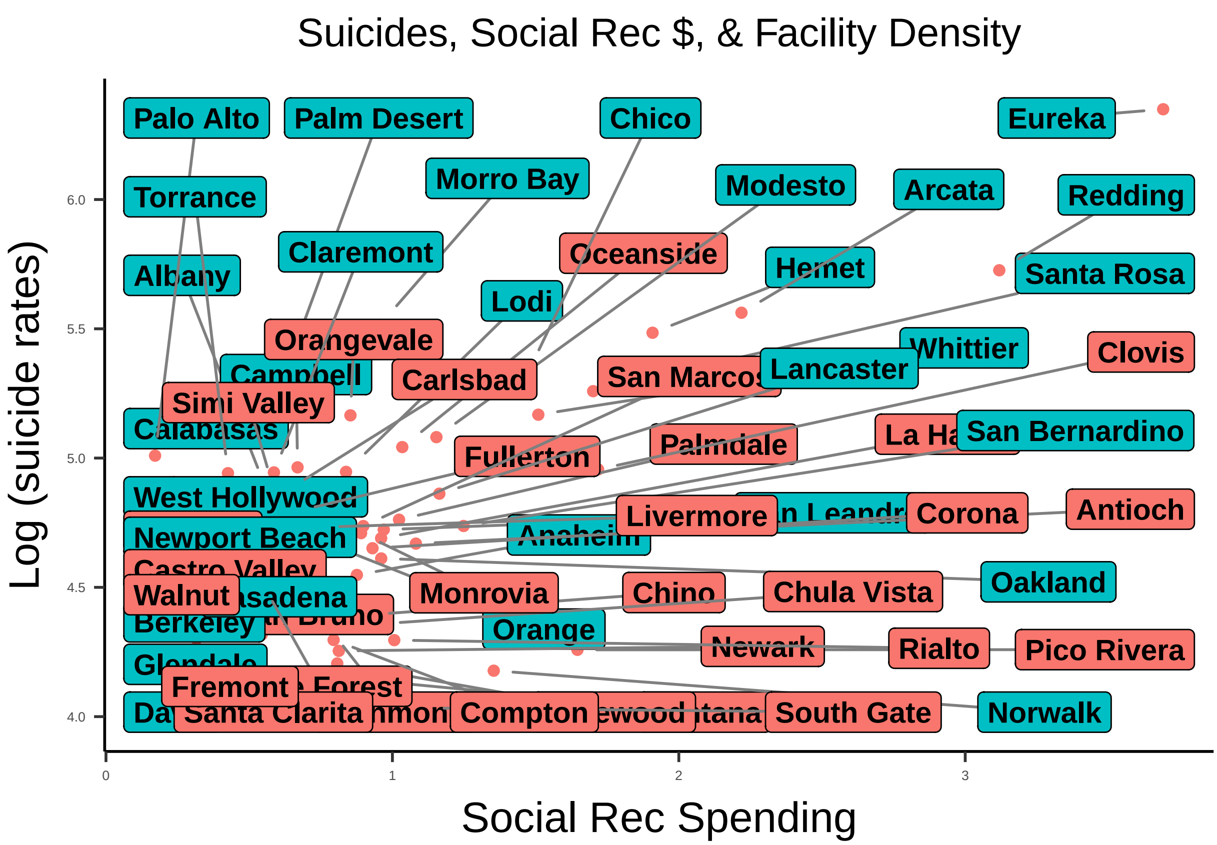
All VIFs for our explanatory variables are acceptable (< 5).

```
##                  male                medAge              NativeAm
##                1.3801                3.0150                2.7216
##              hispanic                lessHS                   pop
##                3.0338                3.0640                1.3040
##          unmarriedMpop            unemployed          facility_cnt
##                2.8295                1.8187                2.4059
##                  hits               poverty  presdrugs.per.person
##                1.3997                4.0036                3.5569
## healthcarebiz.per.1000              landArea              pop_dens
```

```
##                 2.5298                  2.5502                  2.0930
```




```
## Loading required package: ggrepel
```


Suicides, Social Rec $, & Facility Density
Log (suicide rates)
Social Rec Spending
6.0
5.5
5.0
4.5
4.0
0
1
2
3
Palo Alto
Palm Desert
Chico
Eureka
Morro Bay
Modesto
Arcata
Redding
Torrance
Claremont
Oceanside
Hemet
Santa Rosa
Albany
Lodi
Orangevale
Whittier
Clovis
Campbell
Carlsbad
San Marcos
Lancaster
Simi Valley
Calabasas
Fullerton
Palmdale
San Bernardino
West Hollywood
Livermore
Corona
Antioch
Newport Beach
Anaheim
Castro Valley
Walnut
Monrovia
Chino
Chula Vista
Oakland
Berkeley
Orange
Newark
Rialto
Pico Rivera
Fremont
Santa Clarita
Compton
South Gate
Norwalk

}}