Playground

Frances Hung

10/21/2017

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 exaplantory 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 they 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.

```
require(gtrendsR)
```

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

```
require(ggplot2)
```

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

```
require(dplyr)
```

```
## 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
```

```
require(zipcode)
```

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

```
data("zipcode")
require(ggmap)
```

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

Making Dataframes

This gives us a master dataframe of search frequencies of "depression" over the past 12 months in the US which relate for sure to mental health. We can take different dataframes using "$": see the dataframe for details.

longitude and latitude of cities

```
cities_longlat<-read.csv("cal_cities.csv",header=TRUE) %>% select(c(location,Latitude,Longitude))

# updated gtrends data
# I used the one with top 50, instead of the one including cities with low search volume . Though I act

gtrends <- read.csv("gtrends_20042015_top50.csv")  %>%inner_join(cities_longlat,by="location")

## gtrends only has 49 cities. Stanford (gtrends hit = 99) got lost
```

```
# full gtrends data for all cities
gtrends_full <- read.csv("gtrends_20042015_full.csv") %>%
  inner_join(cities_longlat, by="location")
```

prepping facilities data to find number per city

```
facilities<-read.csv("facilities_final.csv",header = TRUE)
colnames(facilities)[7]<-"zip"
facilities$zip<-as.character(facilities$zip)
filtered_facs<-inner_join(zipcode,facilities,by="zip")
city_facs<-filtered_facs %>% group_by(city) %>% summarise(facility_cnt=n())
colnames(city_facs)[1]<-"location"
```

prepping suicide data to find number per city, 2004-2015

```
suis<-read.csv(file="death.csv",header=TRUE) %>% filter(Causes.of.Death=="SUI") %>% filter(Year >= 2004
colnames(suis)[2]<-"zip"
suis$zip<-as.character(suis$zip)
suis2 <-inner_join(zipcode,suis,by="zip")
# aggregate suicide data across all the years for each city
city_suis<-suis2 %>% group_by(city) %>% summarise(suicides=sum(Count))
colnames(city_suis)[1]<-"location"

# wrangled the data for the purpose of GIS. Need to join to population by zipcode data (same source as
# p/s also need to ensure that the other dem data (esp. those we are going to plot) exist at the zipcod
zip_suis <- suis2 %>% group_by(zip) %>% summarise(suicides=sum(Count))
gis_suis <-suis2 %>% filter(Year == 2015) %>% select(1:3) %>% left_join(zip_suis,by="zip")
```

Adding in city demographic data for 2017

```
citydem<-read.csv("citydems.csv",header=TRUE)
citydem2<-read.csv("citydems2.csv",header=TRUE)
citydem2$Name<-gsub(",.*","",citydem2$Name)
citydem$Name<-gsub(",.*","",citydem$Name)
citydem$FIPS<-NULL
citydem2$FIPS<-NULL
colnames(citydem)<-c("location", "male", "female","healthcare","bluecollar","whitecollar","nonfamily","
citydem<-inner_join(citydem,city_facs,by="location")
citydem$facility_cnt<-citydem$facility_cnt*100000/citydem$pop
colnames(citydem2)<-c("location","healthcarepp","activities","socialRec","entertainment","pov","presdru

# I joined it with the new gtrends data. Not sure why two cities disappeared (meaning the citydem data

## citydem data doesn't have Ventura (gtrends hit=93)
```

```
# explanatory data table using full gtrends data (over 180 cities)
logtable_full<-inner_join(citydem,gtrends_full,by="location") %>% inner_join(city_suis,by="location") %
# viewing the data frame reveals that Burbank and Mountain View are repeated 4 times somehow. remove th
logtable_full <- logtable_full [-c(11,12, 52,53,10,54, 55, 13,141,154,155), ]
# now the table has 164 cities, whereas the full list of gtrends had 188. not a big loss
logtable_full_crop <- logtable_full [,-c(1,24,25)]
```

```
# normalize the explanatory variables data frame
logtable_full_crop_normalized <- scale(logtable_full_crop) %>% data.frame()
```

```r
# try the model again using full gtrends data
set.seed(47)
model_full<-lm((suicides)~.,data=logtable_full_crop_normalized)
summary(model_full)
```

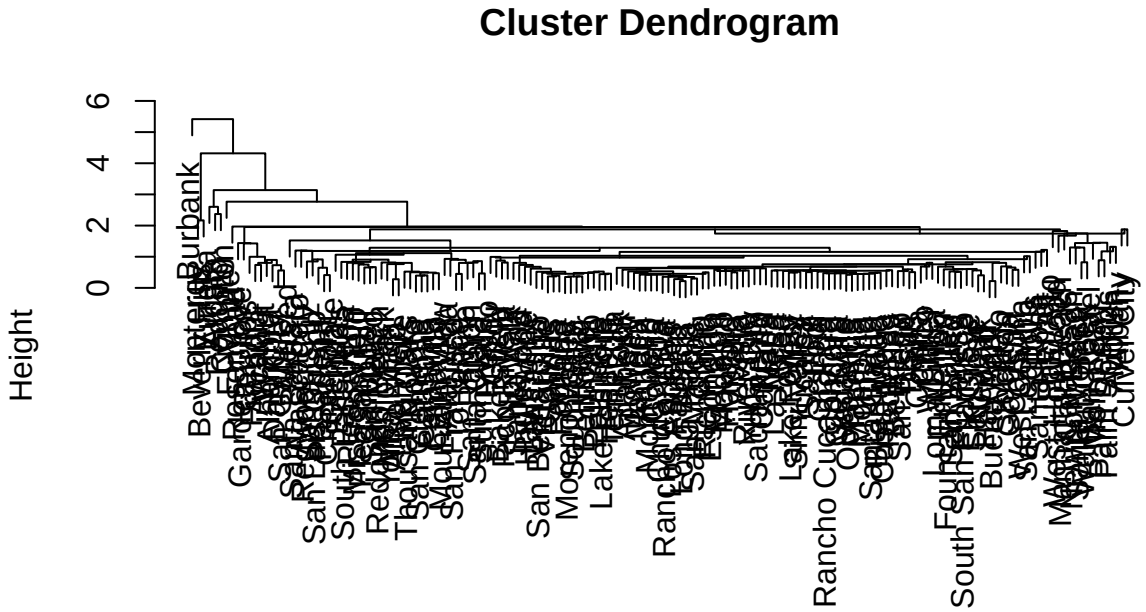
```
##
## Call:
## lm(formula = (suicides) ~ ., data = logtable_full_crop_normalized)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1795 -0.2806 -0.0144  0.2319  3.4673
##
## Coefficients: (1 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.073e-15  4.183e-02   0.000 1.000000
## male          -1.585e-02  6.982e-02  -0.227 0.820755
## female                NA         NA      NA       NA
## healthcare    -5.382e-01  3.390e-01  -1.588 0.114715
## bluecollar    -6.879e-02  1.757e-01  -0.392 0.696044
## whitecollar   -3.577e-01  2.375e-01  -1.506 0.134424
## nonfamily     -5.016e-01  2.518e-01  -1.992 0.048336 *
## medAge         6.053e-01  1.776e-01   3.408 0.000864 ***
## AmInd          3.759e-01  7.314e-02   5.139 9.46e-07 ***
## whiteNonHisp  -2.983e-01  1.134e+00  -0.263 0.792950
## hisp          -7.715e-01  8.933e-01  -0.864 0.389297
## white         -1.229e-01  4.259e-01  -0.289 0.773280
## black         -7.207e-01  3.565e-01  -2.021 0.045212 *
## asian         -5.385e-01  6.491e-01  -0.830 0.408279
## medIncome      1.041e-01  3.856e-01   0.270 0.787620
## lessHS         1.712e-01  1.619e-01   1.058 0.292132
## HS            -3.589e-01  2.141e-01  -1.676 0.096023 .
## Bachelors     -9.713e-02  1.973e-01  -0.492 0.623381
## pop           -6.409e-02  4.539e-02  -1.412 0.160225
## unmarriedMpop  1.306e-01  1.308e-01   0.999 0.319801
## unemployed    -1.488e-01  5.552e-02  -2.680 0.008269 **
## facility_cnt   2.498e-01  6.762e-02   3.695 0.000319 ***
## Hits           3.633e-02  5.895e-02   0.616 0.538826
## healthcarepp  -1.091e+01  4.971e+00  -2.195 0.029874 *
## activities     3.073e+00  1.591e+00   1.932 0.055427 .
## socialRec     -4.575e+00  2.123e+00  -2.155 0.032902 *
## entertainment  8.569e+00  4.437e+00   1.931 0.055568 .
## pov            3.131e-01  1.508e-01   2.077 0.039722 *
## presdrugs      4.344e+00  2.185e+00   1.988 0.048823 *
## healthcarebiz -1.236e-01  7.489e-02  -1.651 0.101160
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5356 on 135 degrees of freedom
## Multiple R-squared:  0.7624, Adjusted R-squared:  0.7131
## F-statistic: 15.47 on 28 and 135 DF,  p-value: < 2.2e-16
```

Significant variables include: Facility density (#mental health facilities/100,000 people) (+), poverty rate (+), % population black (big -), % population native (+), % population asian(big -), median age (+), %

population with up to HS education (-), unemployment rate (-), healthcare spending per person (big -), social/recreation/gym club spending (big -), and prescription drug spending (big +). Other variables to consider would be entertainment (excluding movies and museums) spending (big +), % population not in a family household (-), and % population hispanic (big -).

```
#ggplot(logtable_crop,aes(x=whitecollar,y=log(suicides)))+geom_point()
```

```
clustering_table<-logtable_full_crop_normalized %>% select(presdrugs,socialRec,healthcarepp,facility_cn
rownames(clustering_table)<-logtable_full[,1]
dist_whole<-dist(clustering_table)
cluster_whole<-hclust(dist_whole,method="centroid")
plot(cluster_whole, labels=logtable_full[,1])
```




```
groups=cutree(cluster_whole,k=20)
groups
```

```
##        Palm Springs      West Hollywood           Vacaville
##                   1                   2                   3
##     San Luis Obispo              Eureka               Chino
##                   3                   4                   3
##           Santa Ana         Costa Mesa       Mountain View
##                   3                   3                   3
##          Atascadero       San Francisco             Salinas
##                   3                   3                   3
##           San Diego         Santa Clara              Orange
##                   3                   3                   3
##           Sunnyvale               Vista       San Clemente
##                   3                   3                   3
##            Monterey       Lake Elsinore           El Monte
```

```
##                 5                 3                 6
##          San Jose  Manhattan Beach         Lancaster
##                 3                 7                 3
##       Santa Cruz           Pomona       San Rafael
##                 3                 3                 8
##       Victorville      Redwood City           Arcata
##                 3                 3                 9
##     Santa Barbara           Newark      Los Angeles
##                 3                 3                 3
##      Garden Grove         Escondido     Redondo Beach
##                10                 3                 3
##            Gilroy           Ontario      Lake Forest
##                 3                 3                 3
##             Napa           Anaheim          Hesperia
##                 3                 3                 3
##          Concord           Fontana           Fremont
##                 3                 3                10
##        Livermore             Chico           Norwalk
##                 3                 3                 3
##        Yuba City         Riverside     Newport Beach
##                 3                 3                11
##        Oceanside           Hayward         Encinitas
##                 3                 3                 3
##  Rancho Cucamonga             Poway    San Bernardino
##                 3                 3                 3
##         Petaluma            Merced            Fresno
##                 3                12                 3
##         Rosemead    Santa Clarita South San Francisco
##                13                 3                 3
##          Manteca           Corona       Buena Park
##                 3                 3                 3
##        Fullerton         Fairfield       Simi Valley
##                 3                 3                 3
##        Placentia           Yucaipa          La Habra
##                 3                 3                 3
##          Hanford         Daly City          Stockton
##                 3                10                 3
##        San Bruno         San Ramon       Santa Rosa
##                 3                 3                 3
##           Walnut       South Gate          Carlsbad
##                10                 3                 3
##       Bakersfield    Thousand Oaks          Temecula
##                 3                 3                 3
##             Lodi       Long Beach       Pleasanton
##                 3                 3                 3
##       San Marcos    Mission Viejo          El Cajon
##                 3                 3                 3
##        Morro Bay           Visalia    Rancho Cordova
##                 1                 3                 3
##         Campbell         Palo Alto           Turlock
##                 3                14                 3
##       Pico Rivera          Palmdale       Yorba Linda
##                 3                 3                 3
##             Brea          Berkeley           Modesto
```

```
##                   3                   3                   3
##            Richmond            Pasadena              Rialto
##                   3                   3                   3
##       Moreno Valley           San Mateo              Irvine
##                   3                   3                   3
##             Antioch             Redding     Fountain Valley
##                   3                  15                   3
##          Sacramento            Torrance             Compton
##                   3                   3                  16
##           Elk Grove              Tustin             Vallejo
##                   3                   3                   3
##          Bellflower           Brentwood              Downey
##                   3                   3                   3
##           Calabasas             Oakland            Whittier
##                  11                   3                   3
##           Camarillo         Chula Vista             Santee
##                   3                   3                   3
##              Clovis           Hawthorne         West Covina
##                   3                   3                   3
##           Glendora       Pleasant Hill             Novato
##                   3                   3                   3
##             Burbank             Cypress             Covina
##                  17                   3                   3
##             Upland          Montebello         San Gabriel
##                   3                   3                  10
##        Santa Monica            Cerritos          Menlo Park
##                  11                  10                   3
##             Gardena         San Leandro         Aliso Viejo
##                   3                   3                   3
##              Carson              Auburn             Alameda
##                   3                   1                   3
##            Monrovia            Redlands             Arcadia
##                   3                   3                  10
##            Glendale               Davis             La Mesa
##                   3                   3                   3
##              Albany            Inglewood      South Pasadena
##                  18                   3                   3
##         Palm Desert               Hemet            Alhambra
##                   1                   3                  10
##         Culver City           Claremont          Loma Linda
##                  19                   3                   3
##        Walnut Creek       Beverly Hills
##                  11                  20
```

```
x<-cbind(clustering_table, groups)

suis_cluster<-function(clus) {
  sd<-logtable_full %>% filter(location %in% (rownames(subset(x,groups==clus)))) %>% .[["suicides"]] %>%
  mean<-logtable_full %>% filter(location %in% (rownames(subset(x,groups==clus)))) %>% .[["suicides"]] %
  #View(logtable_full%>% filter(location %in% (rownames(subset(x,groups==clus)))))
  return(c(mean,sd))
}
 suis_cluster(5)
```

```
## [1] 5.015808       NA
```

```
#(lapply(1:12,suis_cluster))
```

```
# set.seed(10)
# kcluster<-kmeans(clustering_table,20, nstart=20)$cluster
# kcluster
# y<-cbind(clustering_table,kcluster)
# logtable_full %>% filter(location %in% (rownames(subset(y,groups==17))))
```