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. However, even if someone wouldn't tell the truth on a questionnaire, they will often tell Google. 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.

- use result/visualization as hook

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

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 death which are not directly relevant to population mental health. Hence, we decided to aggregate suicide rate and Google Trends data over 16 years (constrained by suicide rate data availability), from 1999 to 2015. Since demographic information is fairly stable over time, we used demographic information from the most recent year to train our model.

Playground

```r
require(gtrendsR)
```

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

```r
require(ggplot2)
```

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

```r
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
```

```r
require(zipcode)
```

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

```r
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.

```r
trend<-gtrends("suicide",c("US"),time="all")
trend$interest_by_region
```

```
##                location hits keyword geo gprop
## 1            New Mexico  100 suicide  US   web
## 2                Alaska   96 suicide  US   web
## 3               Wyoming   96 suicide  US   web
## 4                Nevada   95 suicide  US   web
## 5         West Virginia   94 suicide  US   web
## 6               Vermont   94 suicide  US   web
## 7               Montana   93 suicide  US   web
## 8              Delaware   92 suicide  US   web
## 9          South Dakota   92 suicide  US   web
## 10              Indiana   92 suicide  US   web
## 11                 Utah   92 suicide  US   web
## 12              Arizona   91 suicide  US   web
## 13        New Hampshire   89 suicide  US   web
## 14             Kentucky   89 suicide  US   web
## 15                Maine   89 suicide  US   web
## 16                Idaho   88 suicide  US   web
## 17         North Dakota   88 suicide  US   web
## 18             Colorado   88 suicide  US   web
## 19             Oklahoma   87 suicide  US   web
## 20         Pennsylvania   87 suicide  US   web
## 21             Arkansas   87 suicide  US   web
## 22             Nebraska   87 suicide  US   web
## 23           Washington   86 suicide  US   web
## 24         Rhode Island   86 suicide  US   web
## 25             Michigan   85 suicide  US   web
## 26             Missouri   85 suicide  US   web
## 27                 Ohio   85 suicide  US   web
## 28                 Iowa   85 suicide  US   web
## 29           New Jersey   84 suicide  US   web
## 30            Tennessee   83 suicide  US   web
## 31             Maryland   83 suicide  US   web
## 32               Kansas   82 suicide  US   web
## 33           California   82 suicide  US   web
## 34        Massachusetts   82 suicide  US   web
## 35          Connecticut   82 suicide  US   web
## 36            Wisconsin   81 suicide  US   web
## 37                Texas   81 suicide  US   web
## 38               Hawaii   81 suicide  US   web
```

```
## 39             Illinois   81 suicide  US   web
## 40              Alabama   80 suicide  US   web
## 41            Louisiana   79 suicide  US   web
## 42 District of Columbia   78 suicide  US   web
## 43            Minnesota   77 suicide  US   web
## 44             New York   77 suicide  US   web
## 45          Mississippi   77 suicide  US   web
## 46       South Carolina   77 suicide  US   web
## 47       North Carolina   76 suicide  US   web
## 48               Oregon   75 suicide  US   web
## 49              Florida   75 suicide  US   web
## 50              Georgia   74 suicide  US   web
## 51             Virginia   67 suicide  US   web
```

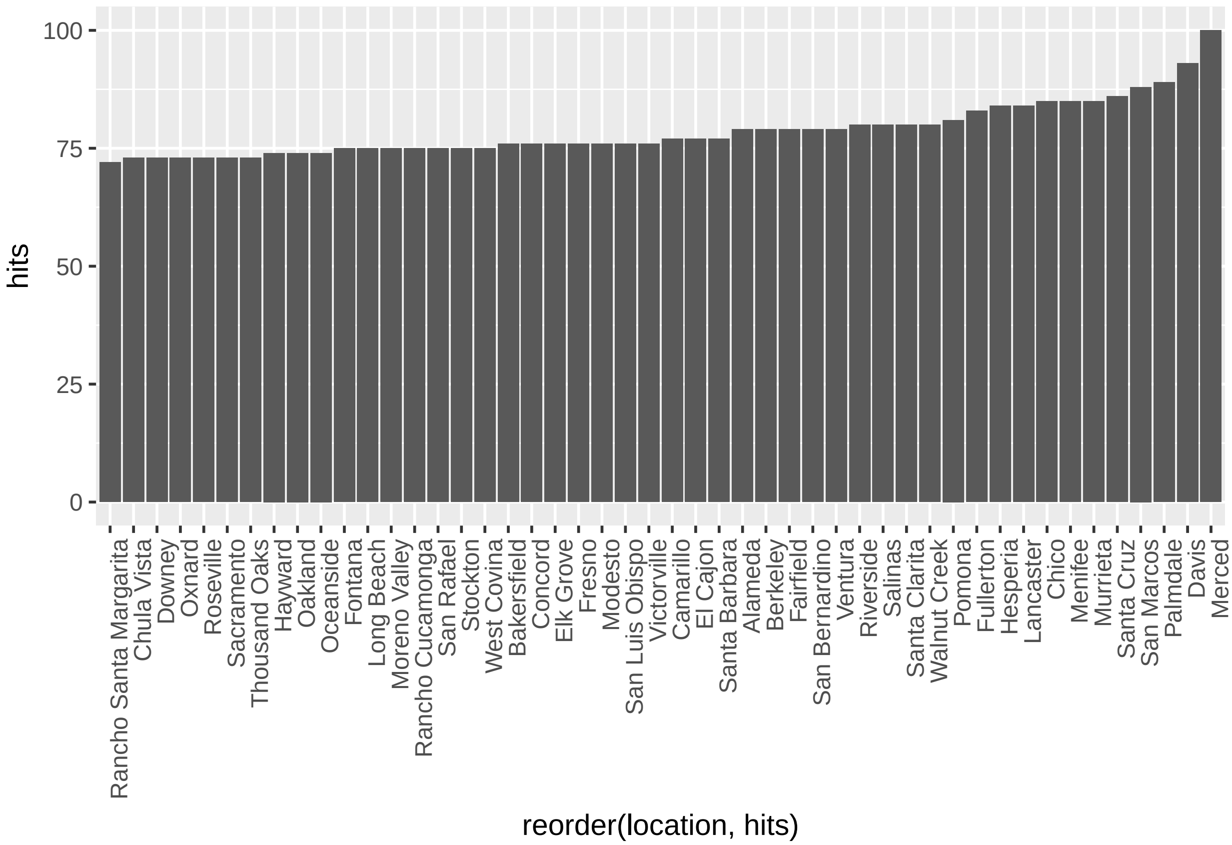
For example, this gives us search frequencies by cities in CA in the U.S.

```r
cities_longlat<-read.csv("cal_cities.csv",header=TRUE) %>% select(c(location,Latitude,Longitude))
cities_dep<-gtrends("depression",c("US-CA"),time="all")$interest_by_city
cities_dep<-cities_dep %>% inner_join(cities_longlat,by="location")

write.csv(cities_dep,file="cities_top49.csv")

##only has 48 cities

# 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
```

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

This plots cities_dep.

```r
ggplot(cities_dep,aes(x=reorder(location,hits),y=hits))+geom_bar(stat="identity")+theme(axis.text.x = el
```




```r
# for (i in 1:length(cities_dep$location)) {
#   place=geocode(cities_dep$location[i],output="latlon",source="dsk")
#   cities_dep$lat[i]=as.numeric(place[1])
#   cities_dep$lon[i]=as.numeric(place[2])
#   print(place)
# }
cities_dep$keyword<-NULL
cities_dep$geo<-NULL
cities_dep$gprop<-NULL
```

```r
facilities<-read.csv("filtered_licensed-healthcare-facility-listing-june-30-2017.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"
```

```r
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
```

```r
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")
```

```r
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)
```

```r
# 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, 54, 55, 13,154,155), ]
# now the table has 186 cities, whereas the full list of gtrends had 188. not a big loss
logtable_full_crop <- logtable_full [,-c(1,24,25)]
```

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

REVISE-Variables significant in this regression are poverty rate (+), density of healthcare businesses(+), and black/Asian population percentwise (-) in the city. Variables which also should be considered according to this regression are percentage of white-collar workers (-), searches of "depression" (+), median income (+), hispanice population (-), and white non-Hispanic population (-).

```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
## -0.62630 -0.09350 -0.00961  0.07559  1.32871
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.398e-16  1.454e-02   0.000 1.000000
## male          2.309e-02  2.269e-02   1.017 0.310529
## female               NA         NA      NA       NA
## healthcare   -1.632e-01  1.170e-01  -1.395 0.164999
## bluecollar   -2.127e-02  6.076e-02  -0.350 0.726752
## whitecollar  -1.432e-01  8.357e-02  -1.713 0.088621 .
## nonfamily    -2.030e-01  7.981e-02  -2.544 0.011938 *
```

```
## medAge         2.119e-01  6.073e-02   3.490 0.000629 ***
## AmInd          1.419e-01  2.608e-02   5.440 2.04e-07 ***
## whiteNonHisp  -4.773e-01  4.022e-01  -1.187 0.237094
## hisp          -6.177e-01  3.190e-01  -1.936 0.054651 .
## white         -1.092e-01  1.558e-01  -0.701 0.484455
## black         -3.532e-01  1.198e-01  -2.949 0.003683 **
## asian         -4.776e-01  2.259e-01  -2.114 0.036085 *
## medIncome      7.307e-02  1.387e-01   0.527 0.599115
## lessHS        -3.618e-02  5.319e-02  -0.680 0.497424
## HS            -1.651e-01  7.807e-02  -2.114 0.036090 *
## Bachelors     -6.798e-02  6.687e-02  -1.017 0.310910
## pop           -2.547e-02  1.565e-02  -1.628 0.105528
## unmarriedMpop  5.968e-02  4.707e-02   1.268 0.206681
## unemployed    -4.662e-02  1.929e-02  -2.417 0.016802 *
## facility_cnt   6.047e-01  1.766e-01   3.423 0.000791 ***
## Hits           1.873e-03  1.914e-02   0.098 0.922174
## healthcarepp  -2.439e+01  1.175e+01  -2.075 0.039652 *
## activities     4.855e+00  3.123e+00   1.555 0.122089
## socialRec     -6.502e+00  3.192e+00  -2.037 0.043369 *
## entertainment  1.692e+01  9.460e+00   1.788 0.075649 .
## pov            1.542e-01  5.029e-02   3.066 0.002561 **
## presdrugs      9.448e+00  4.636e+00   2.038 0.043230 *
## healthcarebiz -4.780e-03  6.601e-02  -0.072 0.942359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1973 on 155 degrees of freedom
## Multiple R-squared:  0.967,  Adjusted R-squared:  0.9611
## F-statistic: 162.4 on 28 and 155 DF,  p-value: < 2.2e-16
```

Significant variables include:

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

```
# library(caret)
# modelrf<-train(suicides~.,method="rf",tuneGrid=data.frame(mtry=c(2,3,4,5,6)),data=whole_norm_logtable
# modelrf$finalModel
# importance(modelrf$finalModel)
```

facilities data

```
# facilities <- read.csv("filtered_licensed-healthcare-facility-listing-june-30-2017.csv")
# facilities <- filter(facilities, LICENSE_CATEGORY_DESC == "Acute Psychiatric Hospital"|LICENSE_CATEGO
# View(facilities)
# write.csv(facilities, file="facilities.csv")
# # more facilities
# facilities.2 <- read_csv("facilities.csv")
# View(facilities.2)
# write.csv(facilities.2, file = "facilities_2.csv")
```

```
# rownames(norm_logtable)<-logtable[,1]
# dist_whole<-dist(norm_logtable)
# cluster_whole<-hclust(dist_whole,method="centroid")
# plot(cluster_whole, labels=logtable[,1])
# groups=cutree(cluster_whole,k=12)
# groups
```

```
# x<-cbind(norm_logtable, groups)
#
# suis_cluster<-function(clus) {
#   sd<-logtable %>% filter(location %in% rownames(subset(x,groups==clus))) %>% .[["suicides"]] %>% log
#   mean<-logtable %>% filter(location %in% rownames(subset(x,groups==clus))) %>% .[["suicides"]] %>% l
#   View(logtable %>% filter(location %in% rownames(subset(x,groups==clus))))
#   return(c(mean,sd))
# }
# suis_cluster(2)
#(lapply(1:12,suis_cluster))
```