Randall Scott Taylor

Data Analytics

Fall 2019

Syracuse University

Final Project Report: Community Health Status Indicators

Project-CHSI

Introduction

CDC(Centers for Disease control and prevention) helps protect America from health, safety and security threats, both foreign and in the U.S. whether diseases start at home or abroad, are chronic or acute, curable or preventable, human error or deliberate attack, CDC fights disease and supports communities and citizens to do the same.

CDC increases the health security of our nation. As the nation's health protection agency, CDC saves lives and protects people from health threats. To accomplish our mission, CDC conducts critical science and provides health information that protects our nation against expensive and dangerous health threats and responds when these arise. Centers for Disease Control and Prevention (CDC) Community Health Status Indicators is a website that provides health profiles for all U.S. counties, including health outcomes, population health status, healthcare access and quality, health behaviors, social factors and the physical environment

CDC's Role: * Detecting and responding to new and emerging health threats * Tackling the biggest health problems causing death and disability for Americans * Putting science and advanced technology into action to prevent disease * Promoting healthy and safe behaviors, communities and environment * Developing leaders and training the public health workforce, including disease detectives * Taking the health pulse of our nation

CDC produces Community Health Status Indicators (CHSI) for all 3,143 counties in the United States. Each profile includes key indicators of health outcomes, which describes the population health status of a county and factors that have the potential to influence health outcomes, such as health care access and quality, health behaviors, social factors, and the physical environment.

Project Objective

Our team was interested in finding a data set in the area of public health. We were looking for a data set that would help us better understand a broad set of health conditions, populations, and potential correlations. This project is an exercise in taking a large pool of data across a variety of metrics and generating relevant questions. Through the use of tools and techniques learned in this course, we will use those insights, which could be used to drive actions for specific populations.

Exploratory Data Analysis

Overview of Data

The data set we chose was the Community Health Status Indicators (CHSI) to combat obesity, heart disease, and cancer that are major components of the Community Health Data Initiative. This dataset provides key health indicators for local communities and encourages dialogue about actions that can be taken to improve community health (e.g., obesity, heart disease, cancer).

The data set has a broad array of health metrics as well as statistics around vulnerable populations, life expectancy and death rates. In reviewing the raw data we felt the goal was to give local public health agencies a set of tools that could help improve the health of their community by identifying root causes and at-risk populations. Website: https://healthdata.gov/dataset/community-health-status-indicators-chsi-combat-obesity-heart-disease-and-cancer

Data Acquisition, Cleaning, Transformation

The file was a zip file that contained several CSV files:

* DATA_ELEMENT_DESCRIPTION.csv defines each data element and indicates where its description is found in Data Sources, Definitions, and Notes.

* DEFINED_DATA_VALUE.csv defines the meaning of specific values (such as missing or suppressed data).
* HEALTHY_PEOPLE_2010.csv identifies the Healthy People 2010 Targets and the U.S. Percentages or Rates.

* DEMOGRAPHICS.csv identifies the data elements and values in the Demographics indicator domain.

* LEADING_CAUSES_OF_DEATH.csv identifies the data elements and values in the Leading Causes of Death indicator domain.

* SUMMARY_MEASURES_OF_HEALTH.csv identifies the data elements and values in the Summary Measures of Health indicator domain.

* MEASURES_OF_BIRTH_AND_DEATH.csv identifies the data elements and values in the Measures of Birth and Death indicator domain.

* RELATIVE_HEALTH_IMPORTANCE.csv identifies the data elements and values in the Relative Health Importance indicator domain.

* VULNERABLE_POPS_AND_ENV_HEALTH.csv identifies the data elements and values in the Vulnerable Populations and Environmental Health indicator domain.

* *PREVENTIVE_SERVICES_USE.csv identifies the data elements and values in the Preventive Services indicator domain.* RISK_FACTORS_AND_ACCESS_TO_CARE.csv identifies the data elements and values in the Risk Factors and Access to Care indicator domain.

In order to provide a robust dataset for our project, we chose a large health dataset containing 573 unique columns for every county in the United States. This broad scope of our dataset was so large that we needed to reduce it in order to focus on key health indicators.

The subset of data we selected included health afflictions, descriptive characteristics, and risk factors. Health afflictions included diseases such as cancer, high blood pressure, and various STIs. Descriptive characteristics in the data included identifiers like poverty, lack of high school education, unemployment, and depression rates. Other data included risk factors such as a lack of healthy eating–defined as few fruits and vegetables–lack of exercise, smoking rates, and frequent drug use. Many of the indicators in our dataset included measures related to ethnicity and age. For example, we can look at the number of white people under 18 with cancer in each county.

Ultimately we selected a subset that helped us understand the factors that represent and influence US county health.

Exploratory Data Analysis

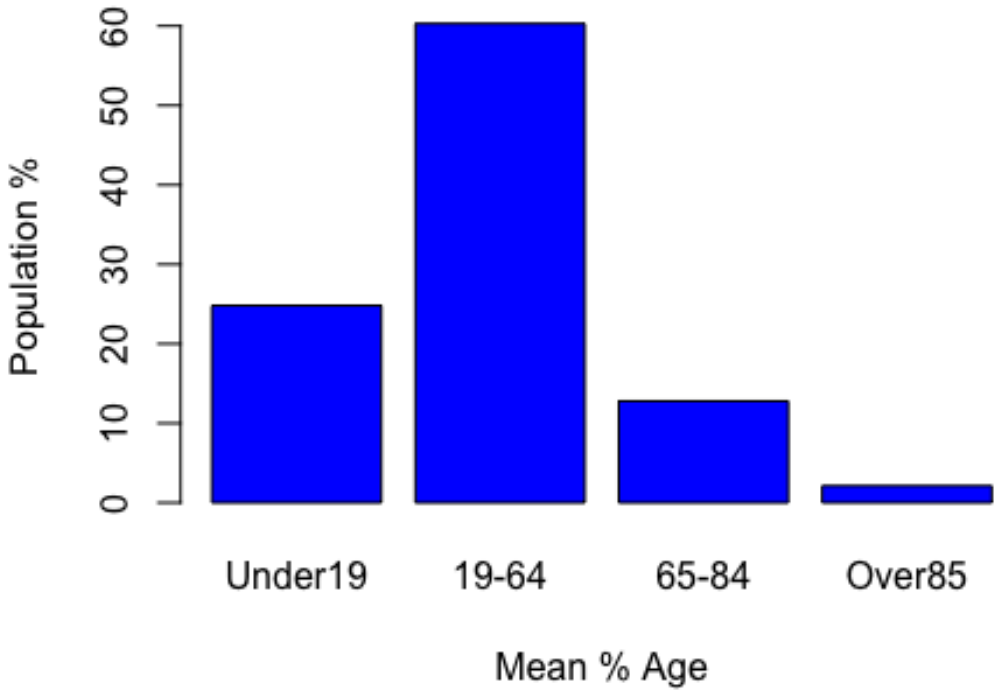
Demographics

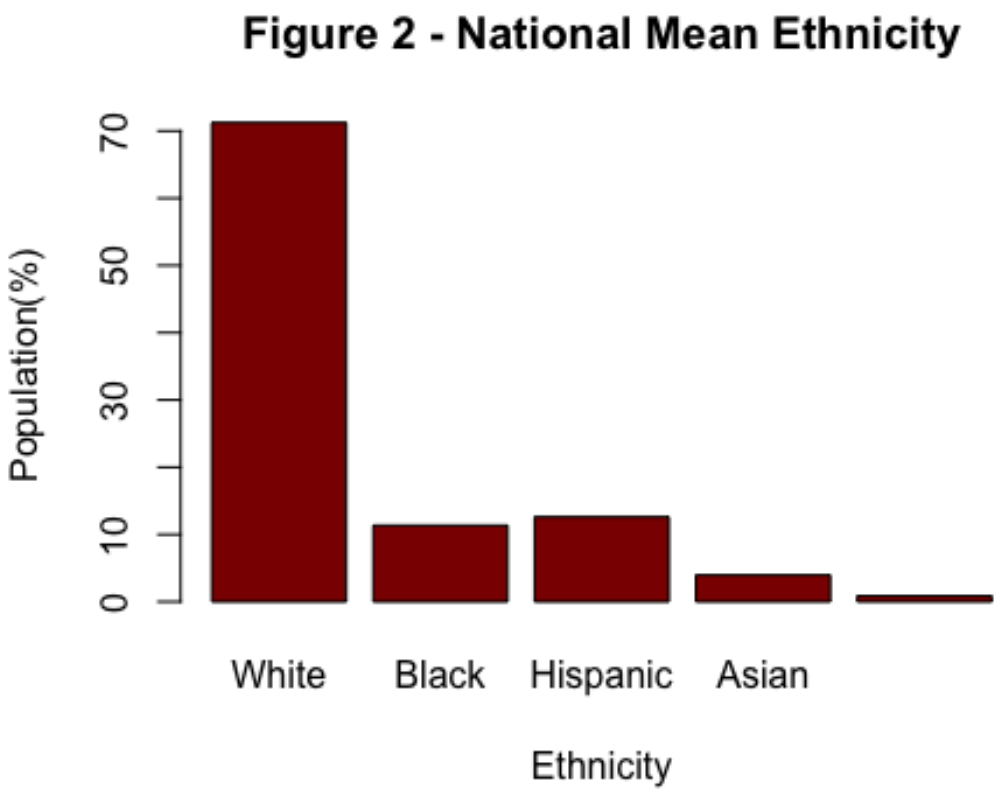
At the beginning of the project, we ran simple analytics to better understand the data and distribution. The bar chart (Figure 1) highlights the average age of the US population and the distribution. The largest bucket is 19-64 (+50%) of the population. It would have been more beneficial if this data had been broken down further; we would have changed the buckets to showcase by ~10-year increments

Next, we looked at another demographic variable, ethnicity. The combination of age/ethnicity would become important for the design of any health program or intervention that is targeted within a certain community. As such, we looked to better understand the national averages and distribution before we dove into a region or county (Figure 2).

So far our statistical analysis has highlighted a largely white population in the age range of 19-64, which is not all that surprising (data ~ 2010).

Figure 1 - National Age %





Population Health Statistics

The data set contained rich information pertaining to life expectancy and major risk factors contributing to the reduction in life expectancy. The goal is to use this data to better understand how to reduce the risk of premature death. In looking at the major contributors to premature death; High Blood pressure, No Exercise, Obesity, and Smoking led the way (Figure 3).

It was analyzed to see what percent of the population of the death was caused due to Suicide and homicide (Figure 5). This actually shows the mental state of the population to a certain extent. ~25% of death is caused due to suicide and homicide which is a high number of the population. Below states seem to have a higher rate in 2015 * Texas * New England * Montana * Colorado

Figure 3 - Population Health Statistics

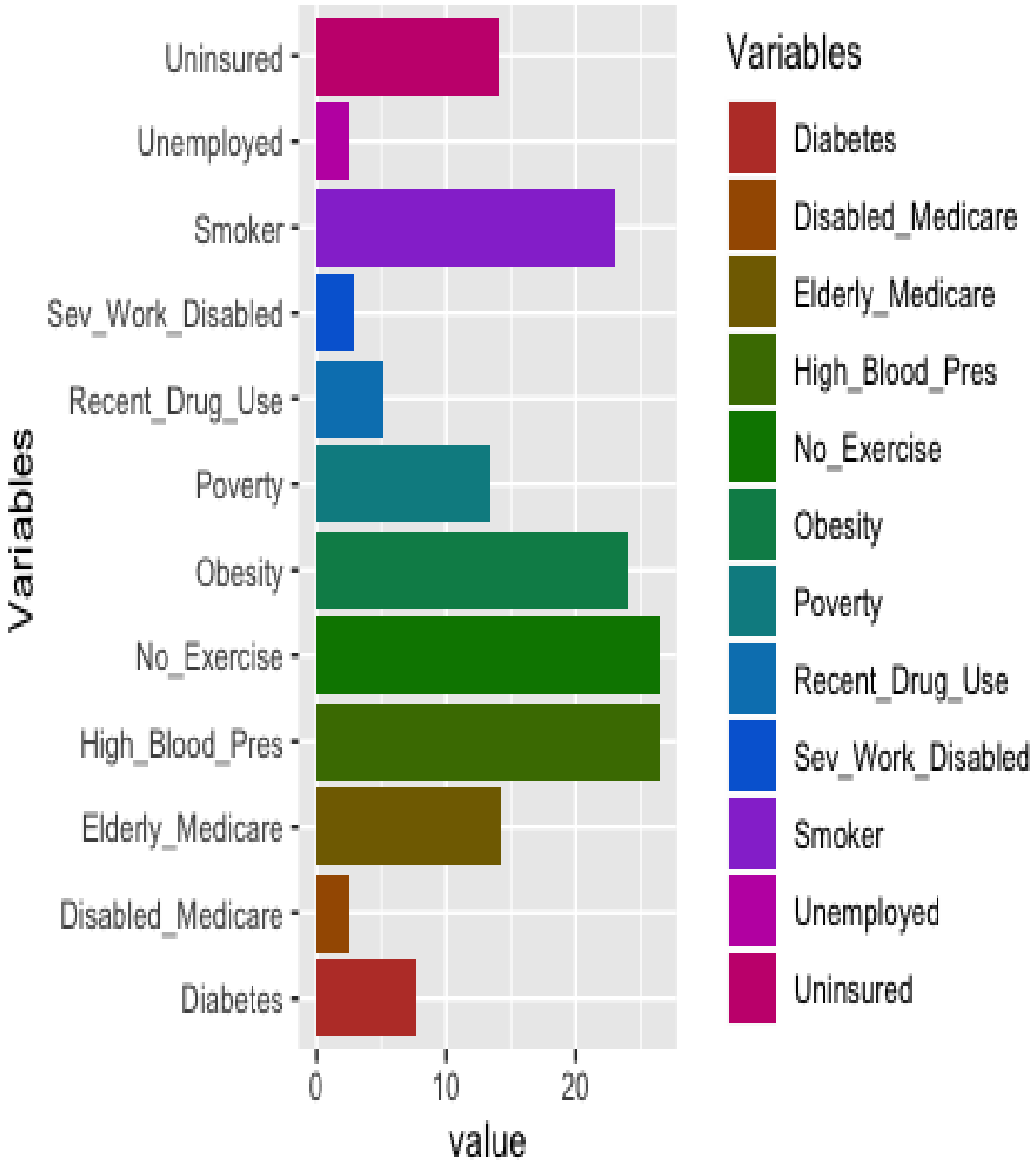
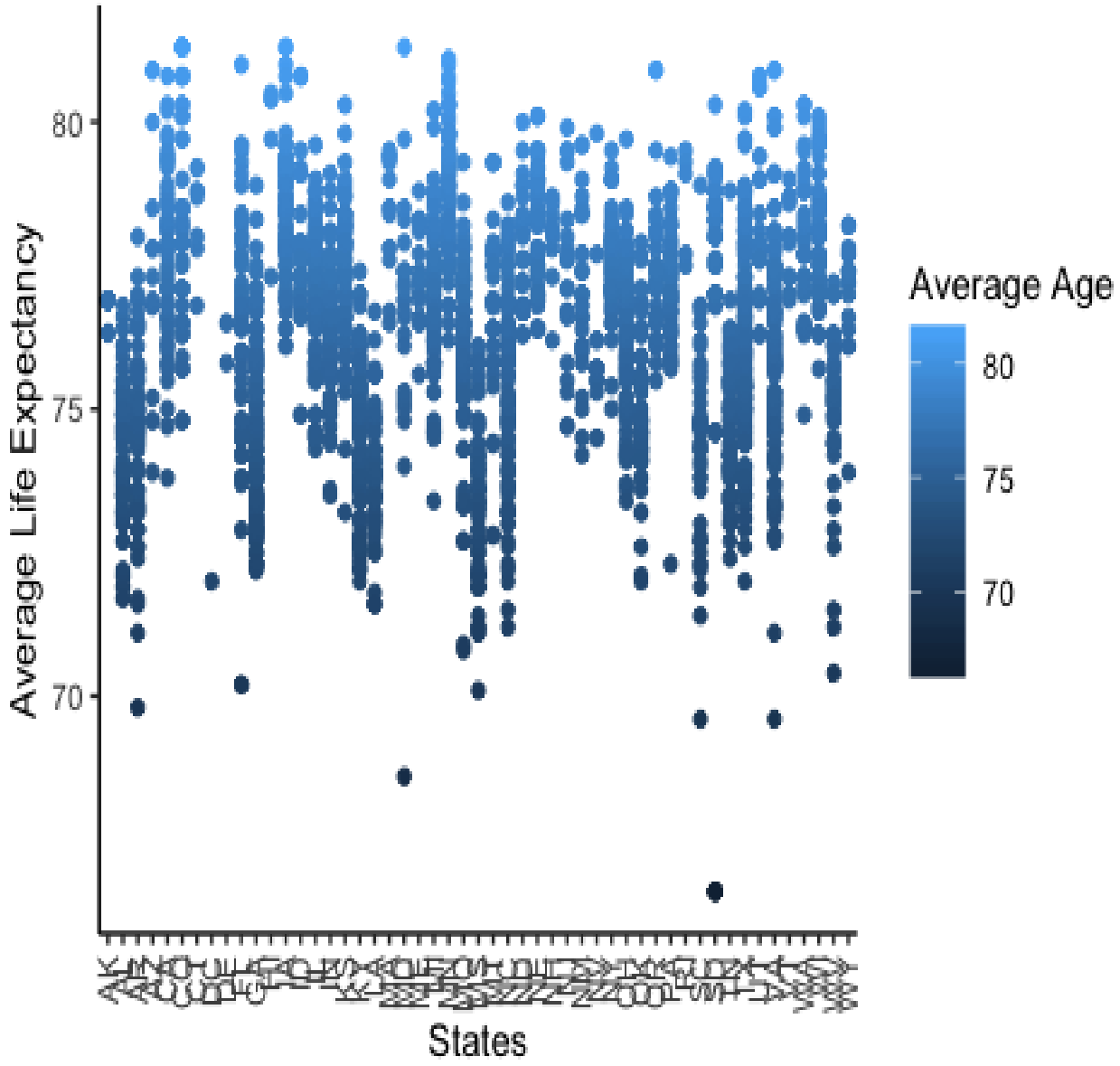


Figure 4 - State Average Life Expectancy



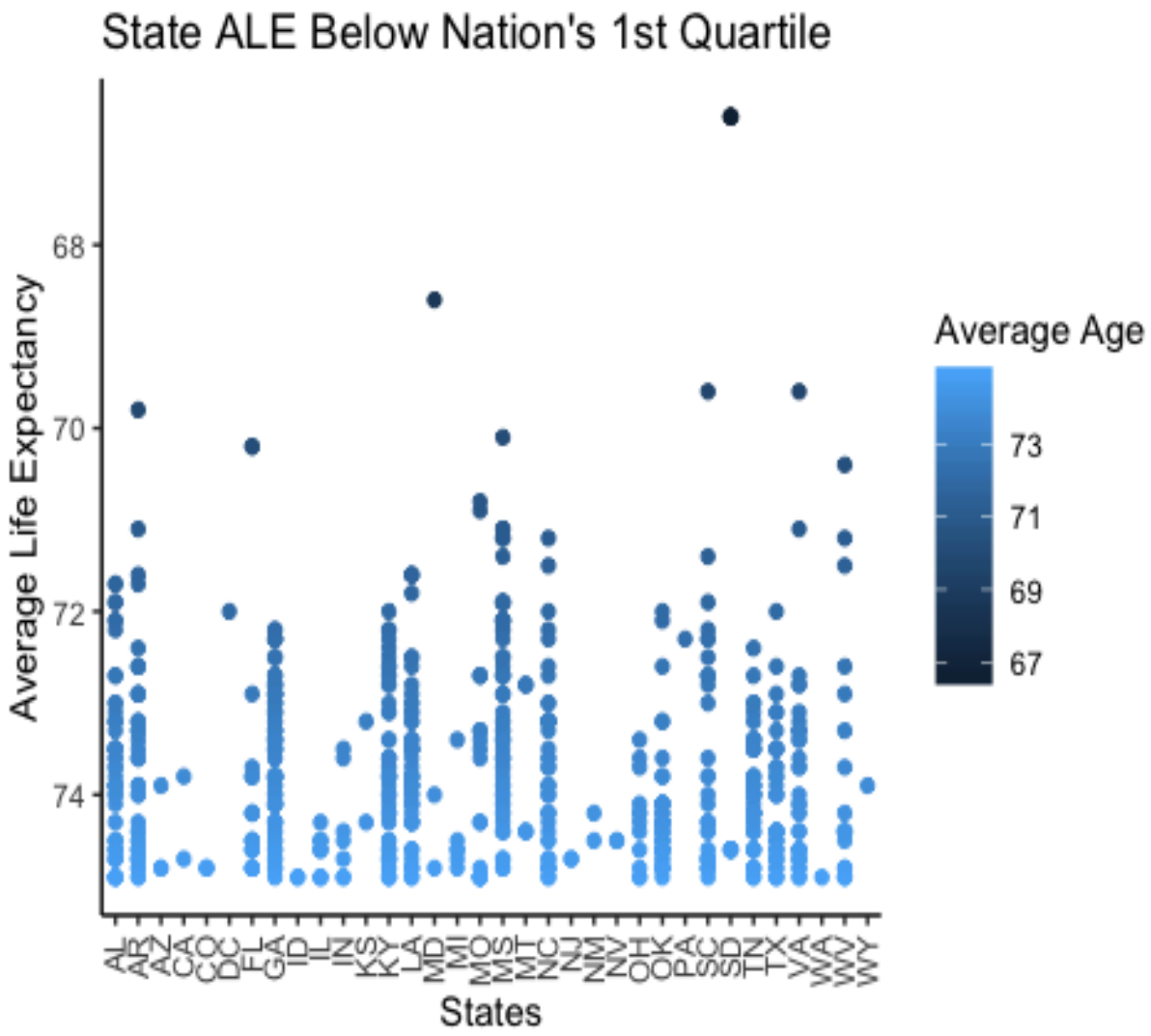
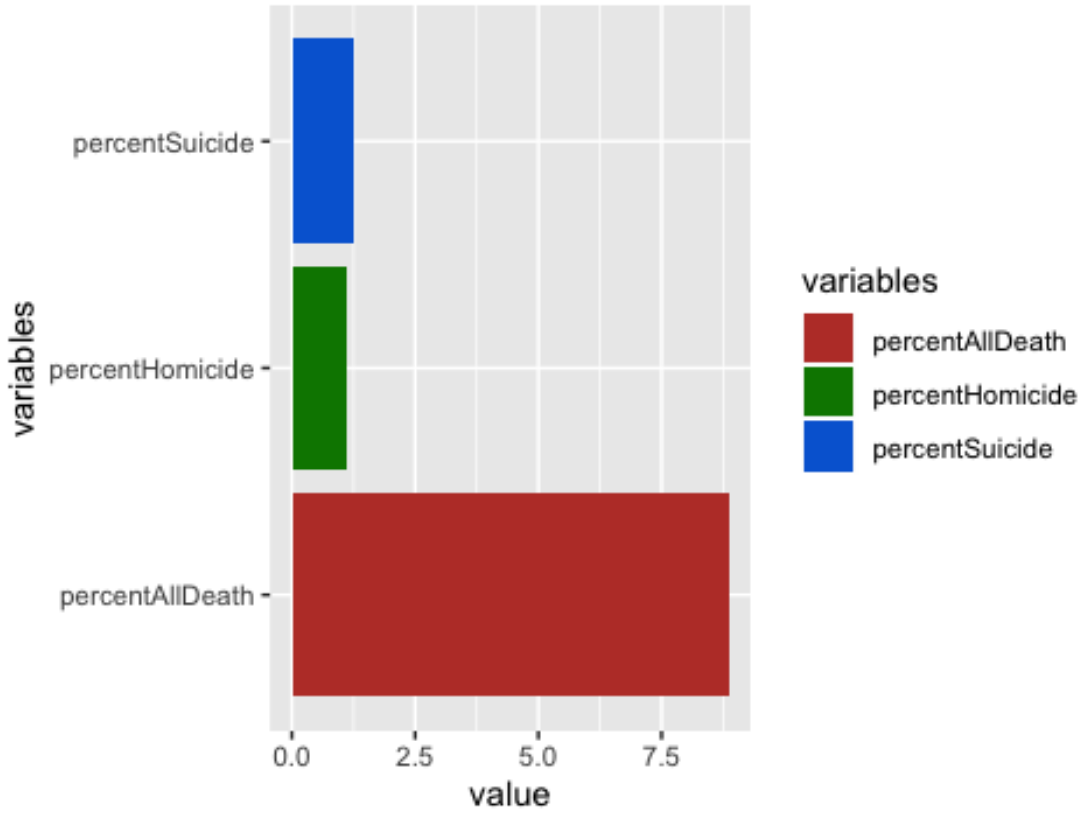

State ALE Below Nation's 1st Quartile
Average Life Expectancy
68
70
72
74
AL
AR
AZ
CA
CO
DC
FL
GA
ID
IL
IN
KS
KY
LA
MD
MI
MO
MS
MT
NC
NJ
NM
NV
OH
OK
PA
SC
SD
TN
TX
VA
WA
WV
WY
States
Average Age
73
71
69
67

Figure 5 - Suicide/Homicide Statistics



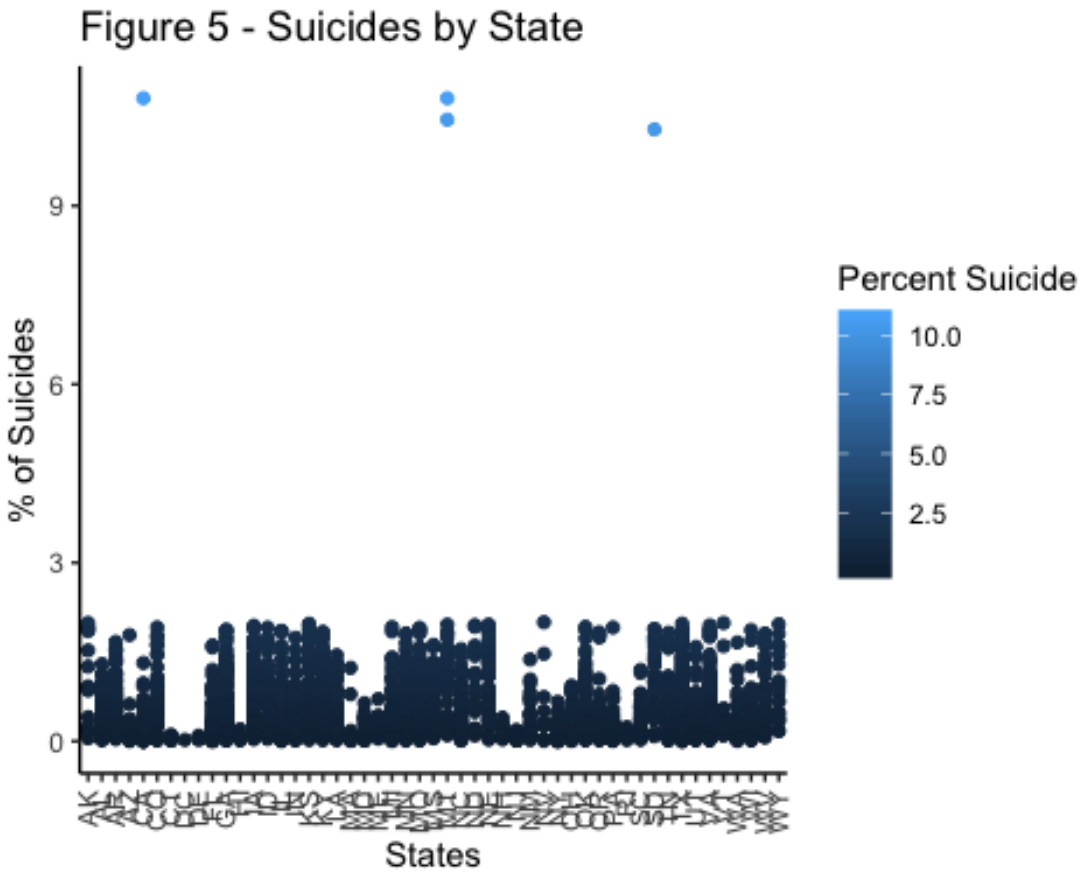
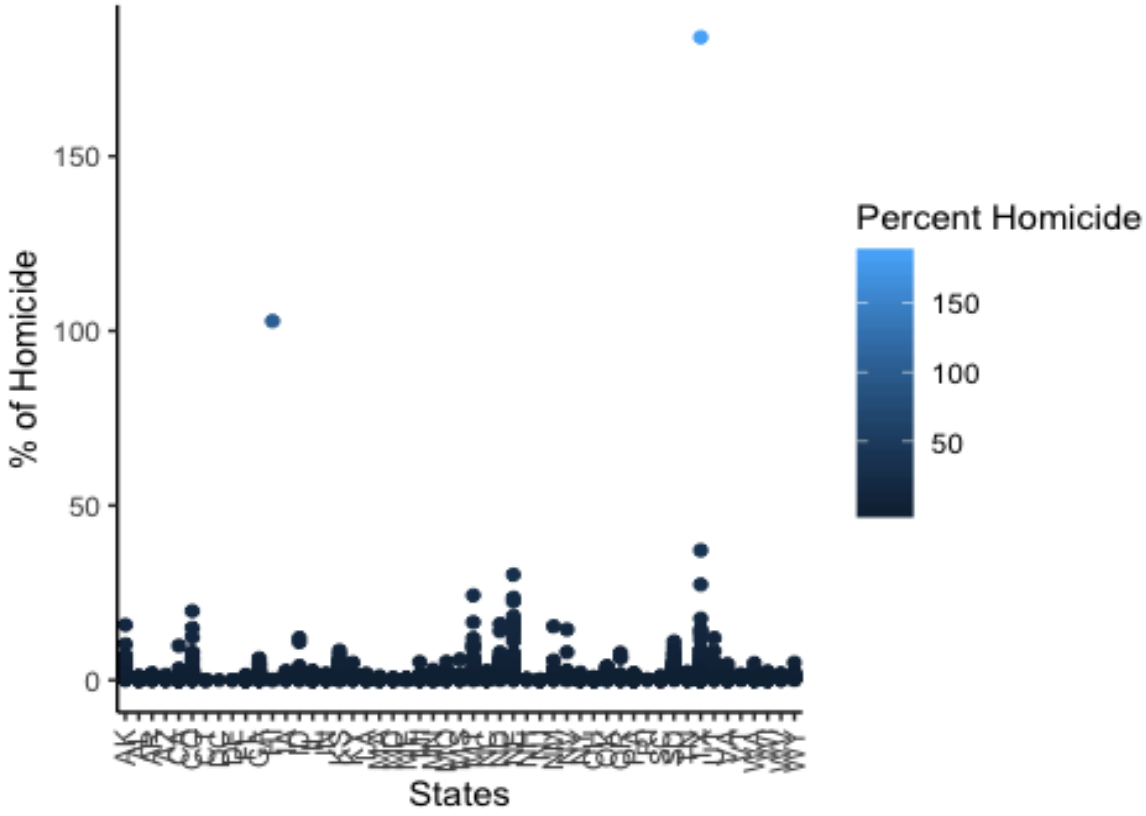
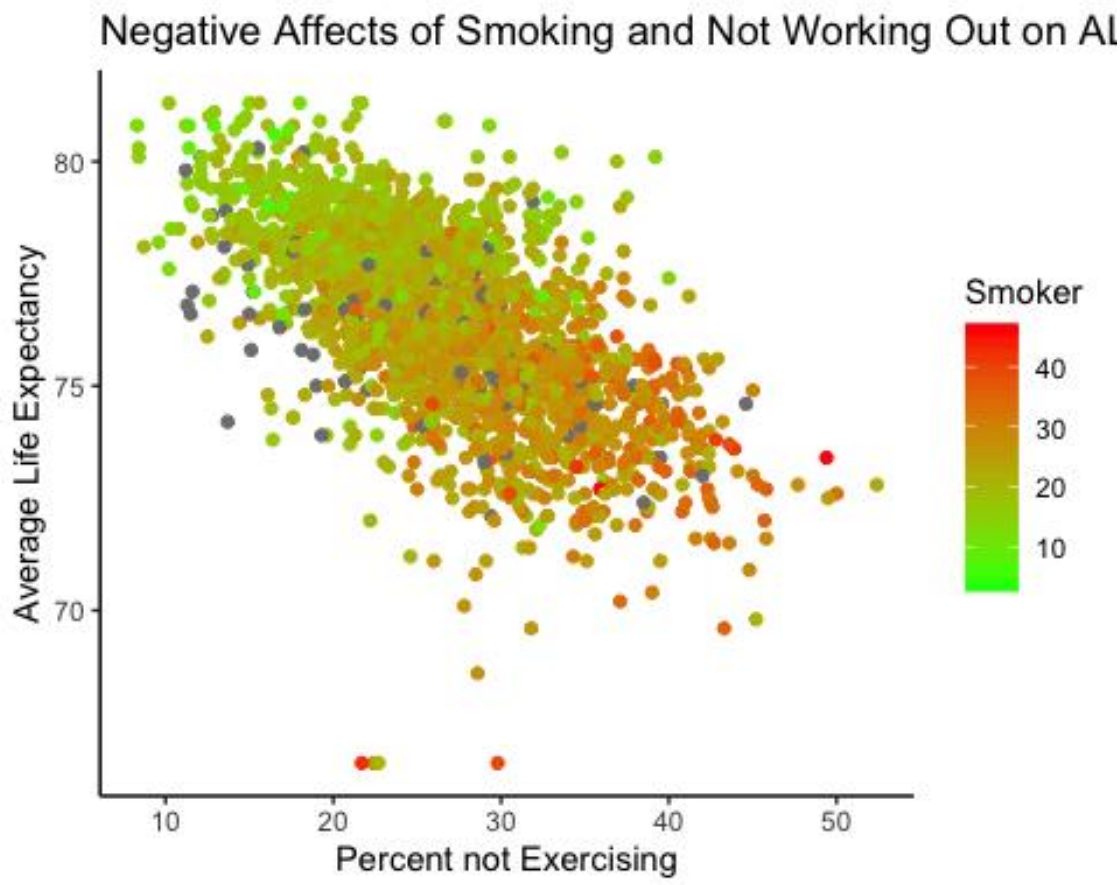


Figure 5 - Suicides by State
% of Suicides
9
6
3
0
States
Percent Suicide
10.0
7.5
5.0
2.5

Figure 6 - Homicides by State





Maps for Demographics and At risk data

AT Risk Clusters MAP

http://rpubs.com/randallscott25/CHSIclustersSouthEast

http://rpubs.com/randallscott25/CHSIclustersGreatLakes

http://rpubs.com/randallscott25/CHSIclustersCentral

http://rpubs.com/randallscott25/CHSIclustersWestCoast

http://rpubs.com/randallscott25/CHSIclustersNortheast

AT Risk Demographic, Categories from EDA with similiartiy clustering

http://rpubs.com/randallscott25/chsiDemoPOV

http://rpubs.com/randallscott25/chsiVULNunemploy

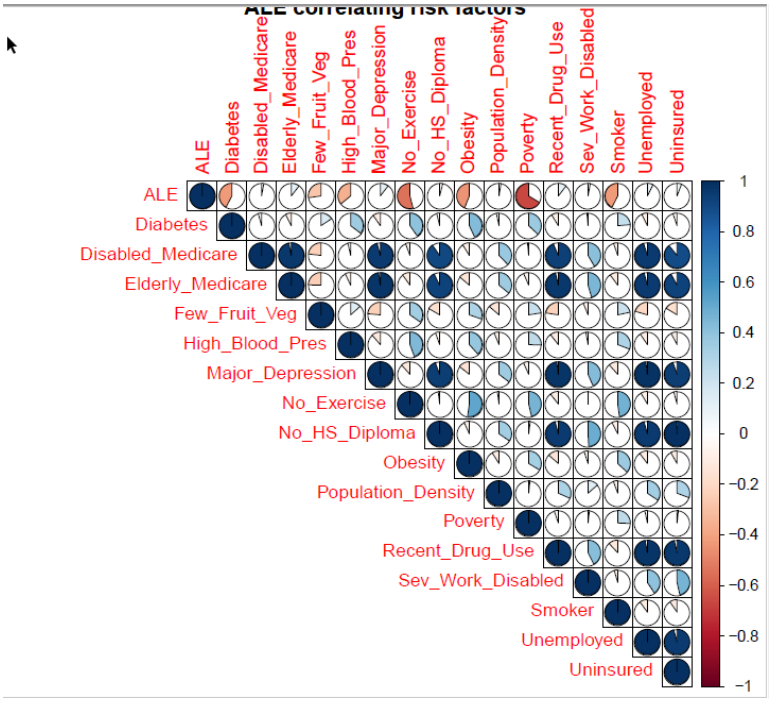
http://rpubs.com/randallscott25/chsiDemoABV19

http://rpubs.com/randallscott25/526493

Data Models

Linear regression

The first step that was taken in the process of transition from “EDA,” by the researcher(s) was to identify: the basic health variables within the data that had to due with negative correlations to life expectancy, to identify and therein see if there was any basic linearity between them. The researcher(s) chose the following correlation matrix visualization to demonstrate the chosen correlation groups; from the CORRPLOT package:



A linear model was built using variables like Age, Demographics, Drug Usage, Use of Toxic Chem, No Exercise, Fruit intake, Blood Pressure, Diabetes and an R-squared value of 71.2% was obtained that explained the variability on Average life expectancy in a county.

* Ethnicity does not impact ALE much but it is seen that impact of longer ALE is in the order of Native American, White, Hispanic, Black and Asian It is seen that as the number of people having HIV, Cancer, Diabetes, Heart Disease increases ALE decreases. The order of negative impact on ALE is Heart disease, Smoking, Diabetes, BP, Obesity, HIV

* It is seen that as the number of people having HIV, Cancer, Diabetes, Heart Disease increases ALE decreases. The order of negative impact on ALE is Heart disease, Smoking, Diabetes, BP, Obesity, HIV

* A strong relationship is seen between Unemployed, Drug Usage and Uninsured Being Uninsured has a strong relationship with unemployment, drug usage, receiving elderly Medicare, and disabled Medicare

* Diabetes is having a moderate to strong relationship with obesity, No-exercise, Blood Pressure

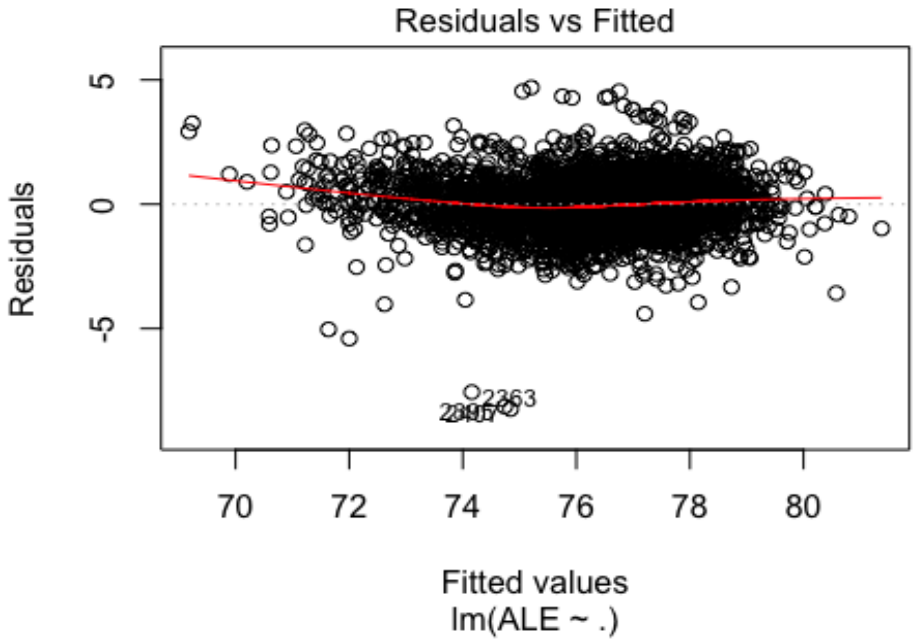
* Major Depression and Drug Use are very strongly related. It is noticed that less Fruit intake is moderately related to no-exercise and obesity

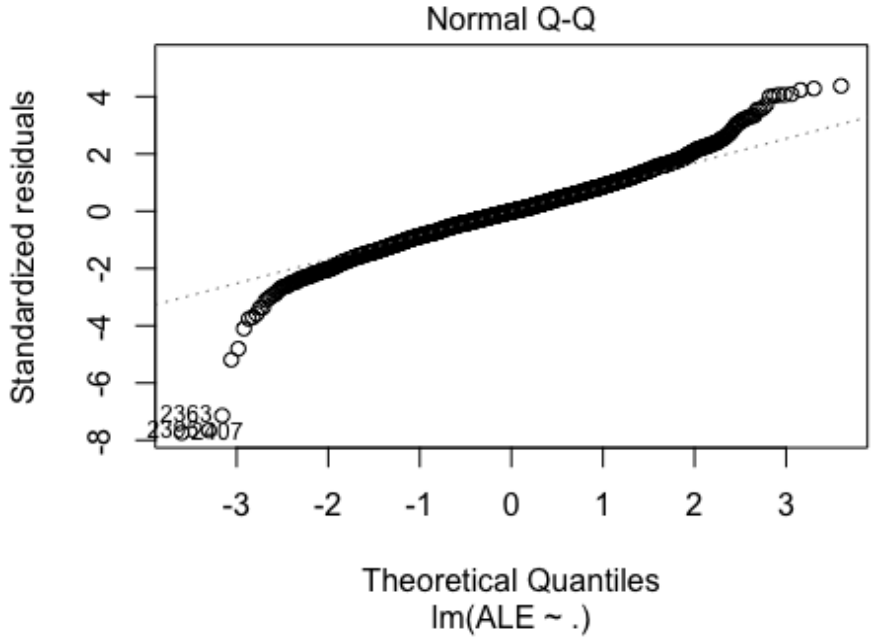
Linear Regression Plots

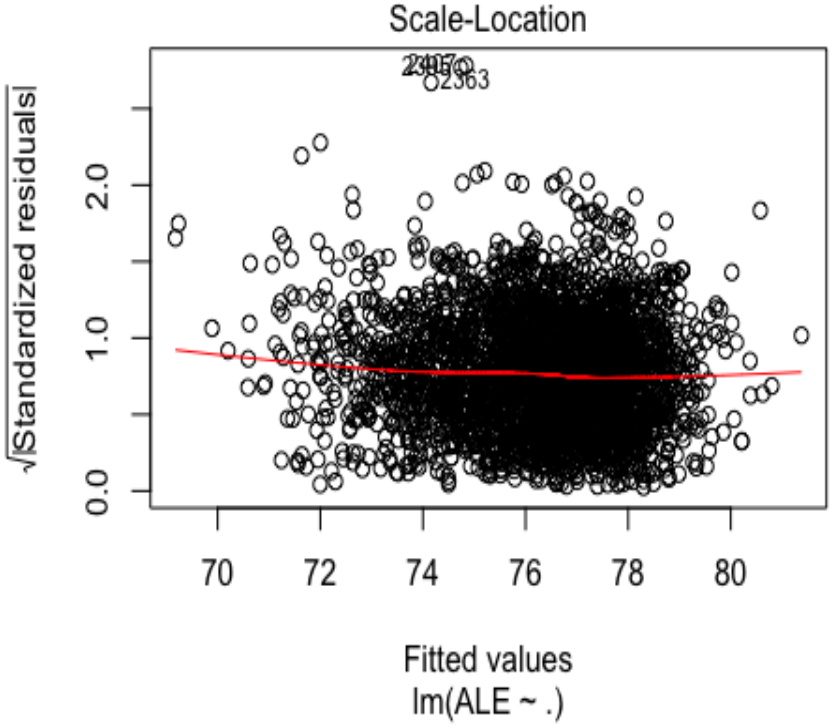
1. Residual vs. Fitter Plot - We see there is no specific pattern in the residuals predicted by our model so we can eliminate the possibility of heteroskedasticity and also possible outliers.

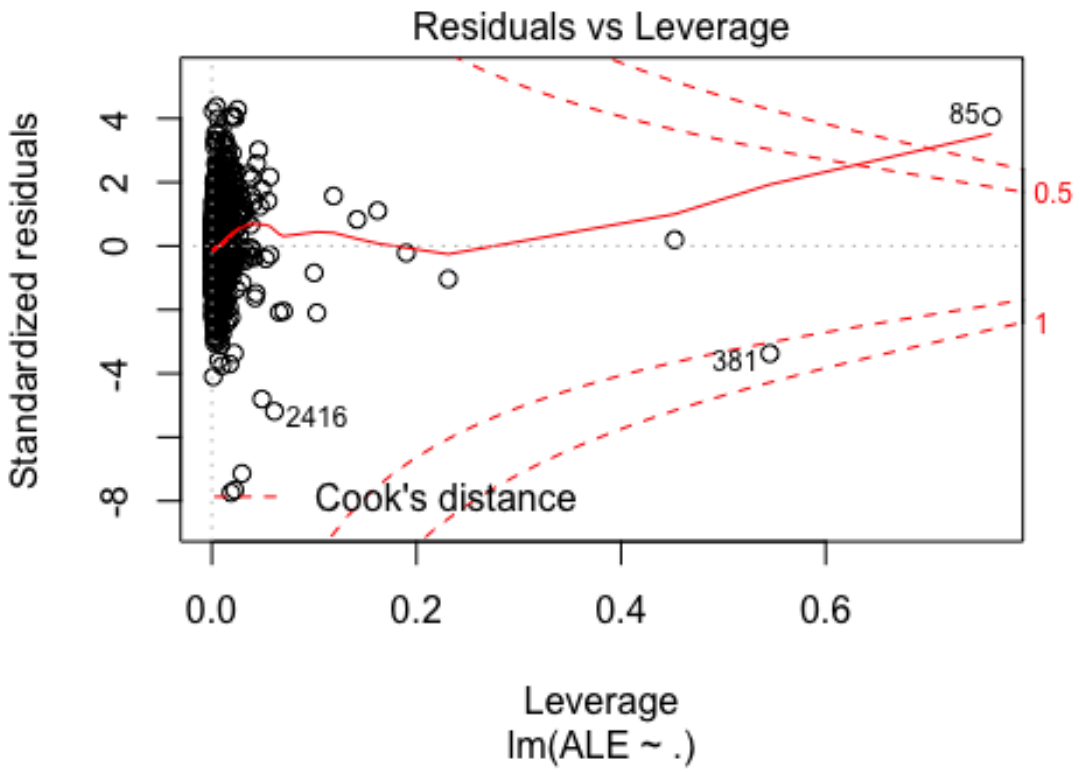
2. QQ Plot - Though most of the points seem to fall on the line which indicates that our residuals come from a normal distribution, there are some points that stray from the line in the lower and upper quantiles of the plot. It is possible that these points do not come from a normal distribution, but most of our points seem to come from a normal distribution so there is not a lot to worry about here.

3. Leverage Plot- This plot graphs the standardized residuals against their leverage. It also includes the Cook's distance boundaries. Any point outside of those boundaries would be an outlier in the x direction. Since we cannot even see the boundaries on our plot, we can conclude that we have no outliers.




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(ALE ~ .)


Scale-Location
√|Standardized residuals|
Fitted values
lm(ALE ~ .)



K-means Cluster Analysis

Cluster analysis is a popular classification technique frequently used to analyze market research data which divides the data into groups. Data appears in rows, purchase intent scores for example, and columns, sales concepts for instance. Rows can then be clustered with respect to columns or columns with respect to rows. For example, clustering techniques can be used to identify demographic or psychographic characteristics of consumers with similar purchasing histories, or to isolate differences between groups of products. In an attempt to understand the data story better, our research team chose to understand the data points relationship to one another utilizing unsupervised K-means clustering. To find these clusters, we utilized Lloyd's Algorithm in the following manner: we start out with k random centroids. A centroid is simply a datapoint around which we form a cluster. For each centroid, we find the datapoints that are closer to that centroid than to any other centroid. We call that set of datapoints its cluster, as demonstrated below:

```
Cluster sizes:

[1] "644 309 433 501 311"

Data means:

   diabetes no.exercise     obesity     poverty      smoker
  0.3618741   0.4352160   0.5543834   0.3113270   0.4502348

Cluster centers:

   diabetes no.exercise   obesity   poverty    smoker
1 0.4029115   0.4268866 0.5810342 0.2222742 0.4479542
2 0.4664020   0.6463567 0.6838661 0.4284614 0.5904513
3 0.2459209   0.2805143 0.4302326 0.1877765 0.3447293
4 0.2890275   0.4380441 0.5389260 0.3248473 0.4721674
5 0.4518314   0.4535139 0.5683005 0.5295878 0.4272040

Within cluster sum of squares:

[1] 18.60534 19.75330 20.76333 14.39275 14.19160
```

Then we take the mean of the cluster and let that be the new centroid. We repeat this process (using the new centroids to form clusters, etc.) until the algorithm stops moving the centroids.

We do this in order to minimize the total sum of distances from every centroid to the points in its cluster — *that is our metric for how well the clusters split up the data*.

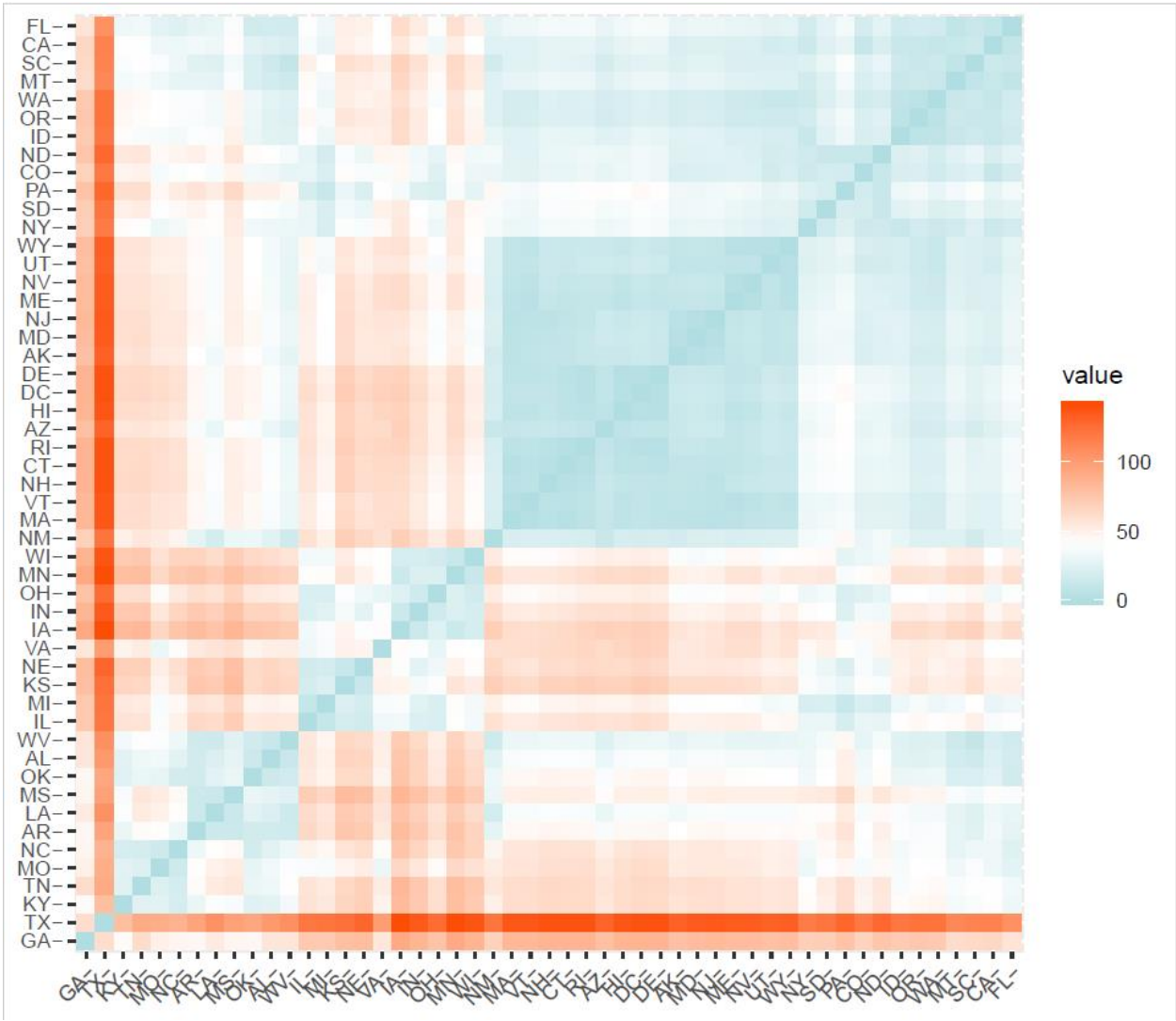
General cluster statistics, identified by R analysis:

```
============================================================================
General cluster statistics:

$n
[1] 2198

$cluster.number
[1] 5

$cluster.size
[1] 644 309 433 501 311

$min.cluster.size
[1] 309

$noisen
[1] 0

$diameter
[1] 28.02858 34.25872 30.77295 26.02153 40.75496

$average.distance
[1]  7.624265 11.758470 10.488787  7.809530  8.543890

$median.distance
[1]  7.300710 11.123848 10.000578  7.623948  7.990162

$separation
[1] 0.200000 1.414214 1.284523 0.100000 0.100000
```

Distance Measure:

The choice of distance measures is very important, as it has a strong influence on the clustering results. For most common clustering software, the default distance measure is the Euclidean distance.

Within R it is simple to compute and visualize the distance matrix using the functions get_dist and fviz_dist from the factoextra R package. This starts to illustrate which states have large dissimilarities (red) versus those that appear to be fairly similar (teal).



Finding ‘k’: number of clusters using the elbow method

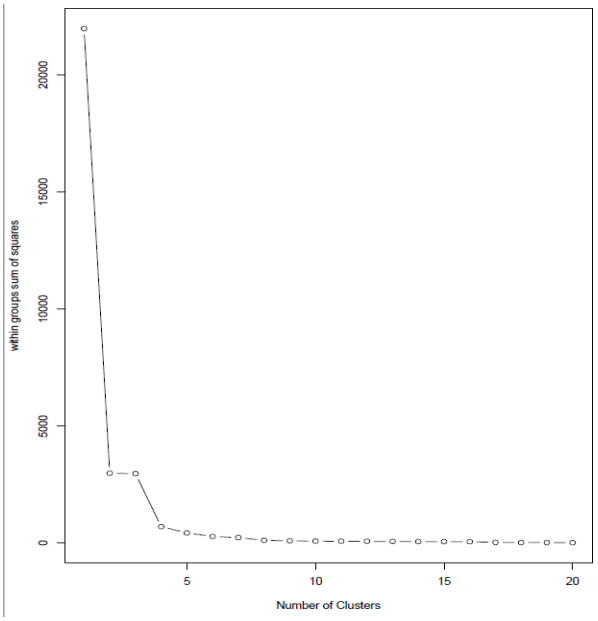
The initial cluster analysis was done utilizing a tool within R, the program rattle. The research team then needed to validate the clusters, to ensure that the initial research regarding cluster ability was sound.

To validate the total clusters of **5**, the research team started to conduct validation utilizing the following method:

Chosen variables from the dataset were established as atRisk, and created into a data.matrix. The atRisk data having been established as a data.matrix was then scale the data, to normalize the data by subtracting the mean and dividing by the standard deviation, to take out the effect of different variables, being measured on different scales, and to eliminate the occurrence of NA in our dataset, as this will interfere with the ability of the algorithm to set, and measure, centroids in the k-means clustering.
Having scaled the data we run an initial, quick function

```
wssplot = function(Test1, nc=20, seed=123){
  wss = (nrow(Test1)-1)*sum(apply(Test1, 2, var))
  for (i in 2:nc){
    set.seed(seed)
    wss[i] = sum(kmeans(Test1, centers = i)$withinss)}
 plot(1:nc, wss, type='b', xlab='Number of Clusters', ylab = "within groups sum of squares")}
```

to decide how many clusters to use.

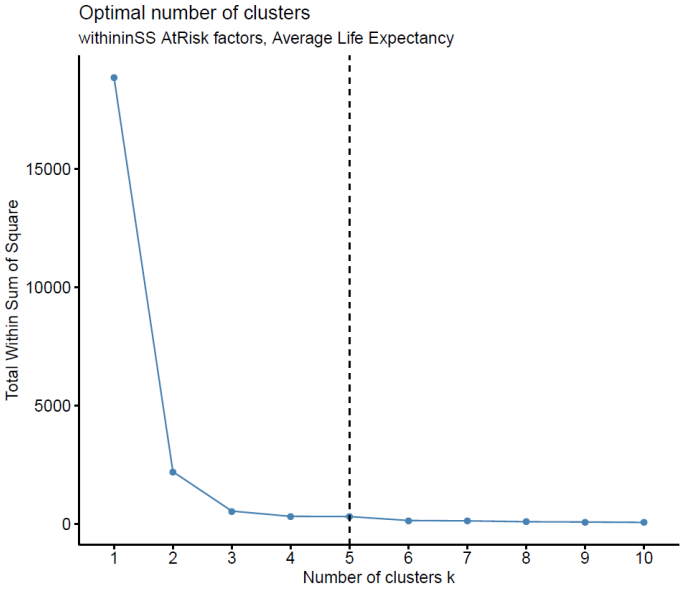


Initial wss plot function

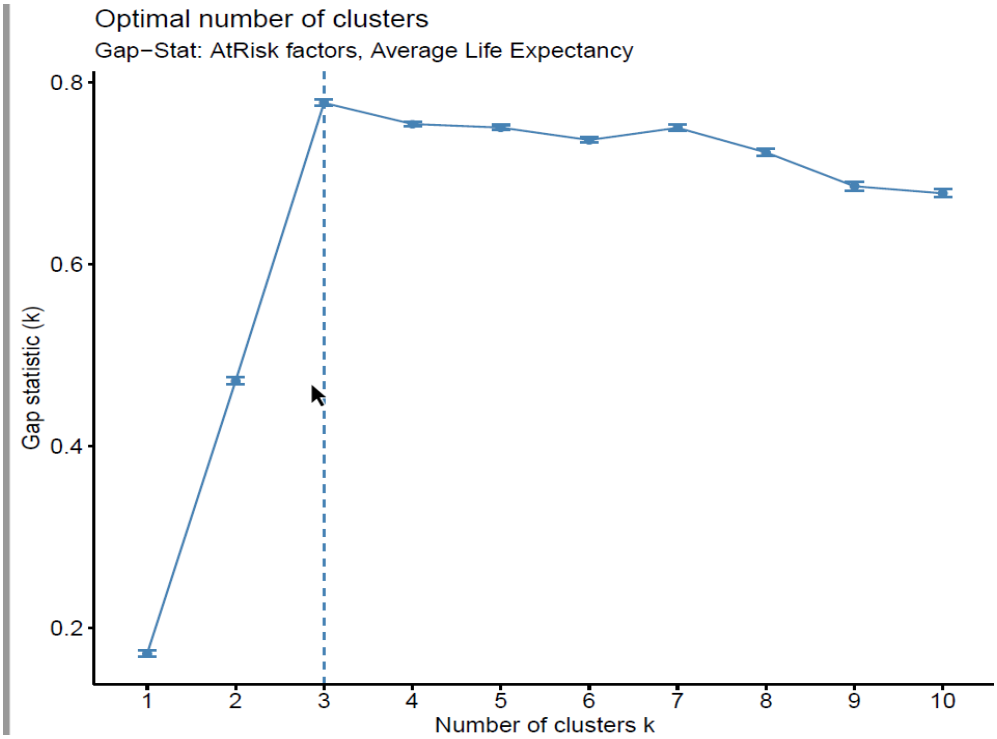
Optimization plots for 3 methods; based upon the NbClust packages

Method 1: A plot of the total within-groups sums of squares against the number of clusters in a K-means solution can be helpful. A bend in the graph can suggest the appropriate number of clusters, as demonstrated --->

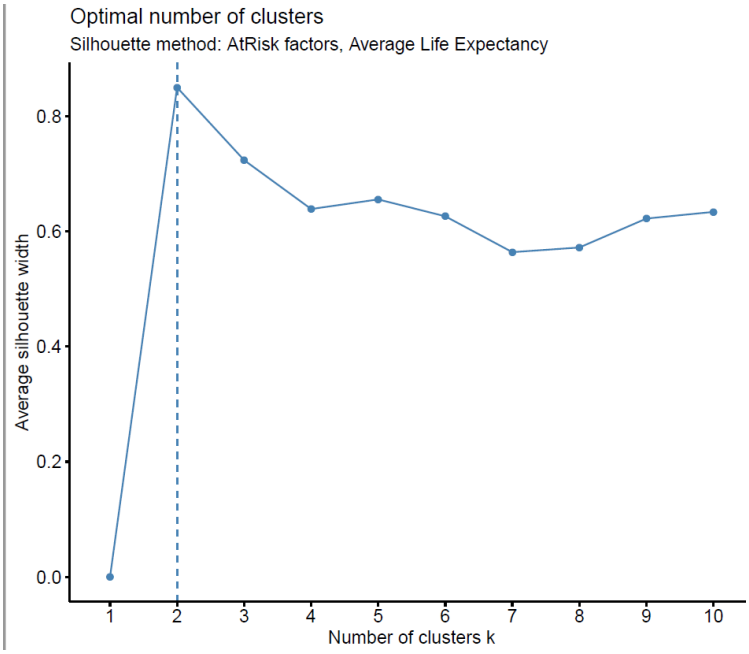
One should choose a few clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the "elbow criterion"



Method 2: A plot of the Gap-stat tests against the number of clusters in a K-means solution can be helpful. The Gap Statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (i.e., that yields the largest gap statistic). This means that the clustering structure is far away from the random uniform distribution of points. A bend in the graph can suggest the appropriate number of clusters, as ←demonstrated.



Method 3: A plot of the silhouette method was chosen as the final optimization plot. Average silhouette method computes the average silhouette of observations for different values of k.

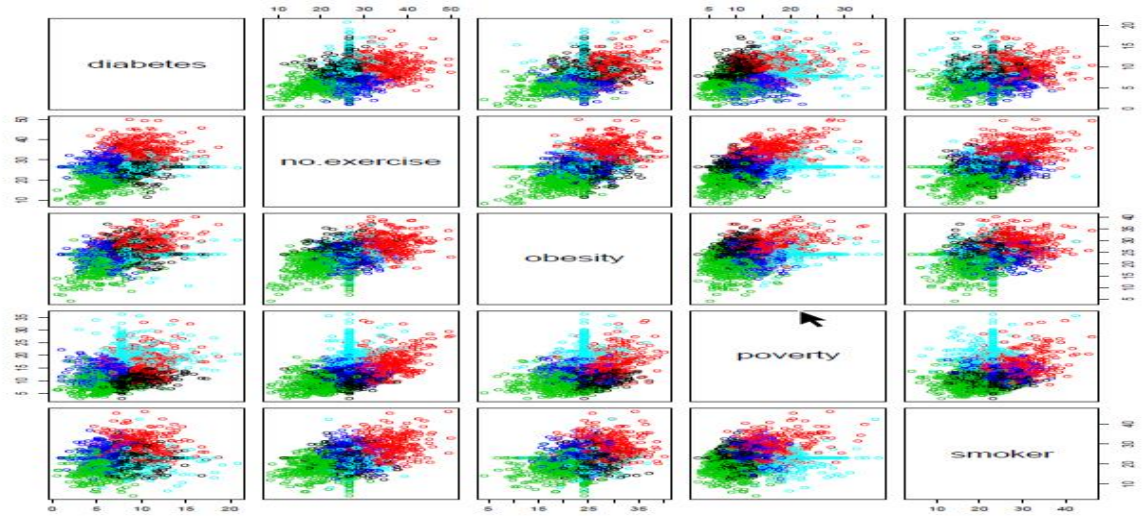


The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for **k**. A bend in the graph can suggest the appropriate number of clusters, as

←demonstrated.

K-means analysis

Having utilized the above three methods, the research team decide to chose 5 as the total amount of clusters to express in the k-means analysis. Given the previously ran correlation matrix, and initial linear regressions ran; the following visualization shows the initial cluster visualizations ran in R:

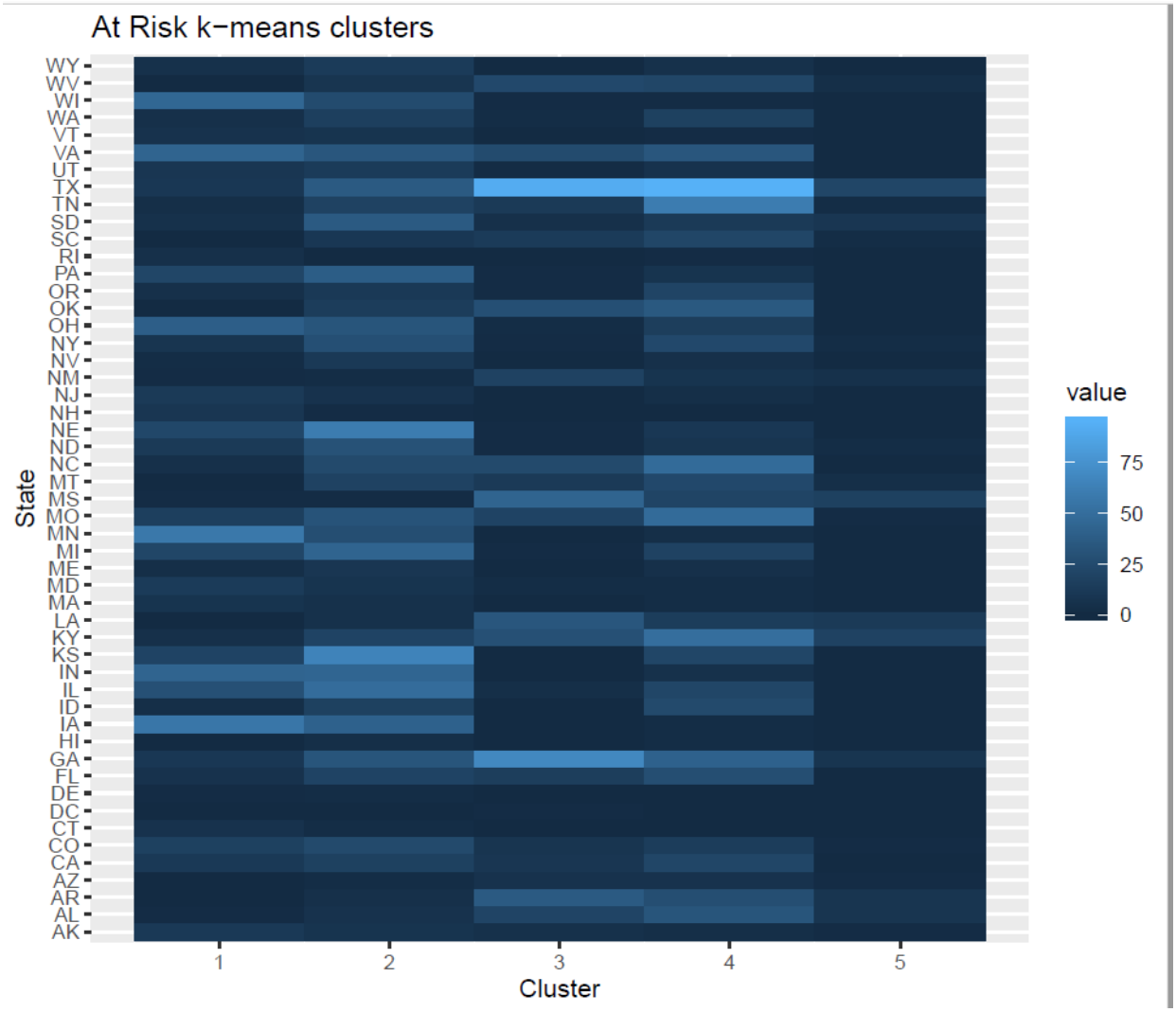



```
      1  2  3  4  5
AK 11  7  5  3  1
AL  1  6 20 32  8
AR  0  4 36 27  8
AZ  0  2  6  6  1
CA 11 17  9 21  0
CO 17 24  8 14  1
CT  6  2  0  0  0
DC  0  0  1  0  0
DE  1  2  0  0  0
FL  6 19 15 27  0
GA 10 32 68 41  8
HI  0  3  0  2  0
IA 56 41  0  2  0
ID  3 17  0 24  0
IL 29 49  3 21  0
IN 43 44  0  5  0
KS 19 65  0 21  0
KY  4 19 29 50 18
LA  0  5 33 14 12
MA  7  5  0  2  0
MD 13  7  2  2  0
ME  3  9  0  4  0
MI 21 43  1 18  0
MN 57 27  0  3  0
MO 16 31 19 48  1
MS  1  1 43 20 17
MT  0 18 12 23  3
NC  4 24 24 48  0
ND 12 30  1  8  2
NE 22 60  1 10  0
NH  9  1  0  0  0
NJ 12  6  0  3  0
NM  1  2 18  7  5
NV  2 11  0  4  0
NY  8 28  1 23  2
OH 39 32  2 15  0
OK  0 14 28 35  0
OR  4 11  1 20  0
PA 21 38  1  7  0
RI  4  0  0  1  0
SC  0 10 13 21  2
SD  5 36  4 12  9
TN  3 18 12 60  2
TX 10 36 92 95 21
UT  9 13  1  6  0
VA 43 33 25 33  0
VT  6  7  0  1  0
WA  4 16  2 17  0
WI 45 25  1  1  0
WV  0  8 23 21  3
WY  5 13  0  5  0

============================================================
 randIndex: measure between state and cluster partitions.

 Note: values vary between -1 to 1
============================================================
        ARI
0.02352161
```

After the assignment step, the algorithm computes the new mean value of each cluster. The term cluster “centroid update” is used to design this step. Now that the centers have been recalculated, every observation is checked again to see if it might be closer to a different cluster. All the objects are reassigned again using the updated cluster means. The cluster assignment and centroid update steps are iteratively repeated until the cluster assignments stop changing (i.e until convergence is achieved). That is, the clusters formed in the current iteration are the same as those obtained in the previous iteration.

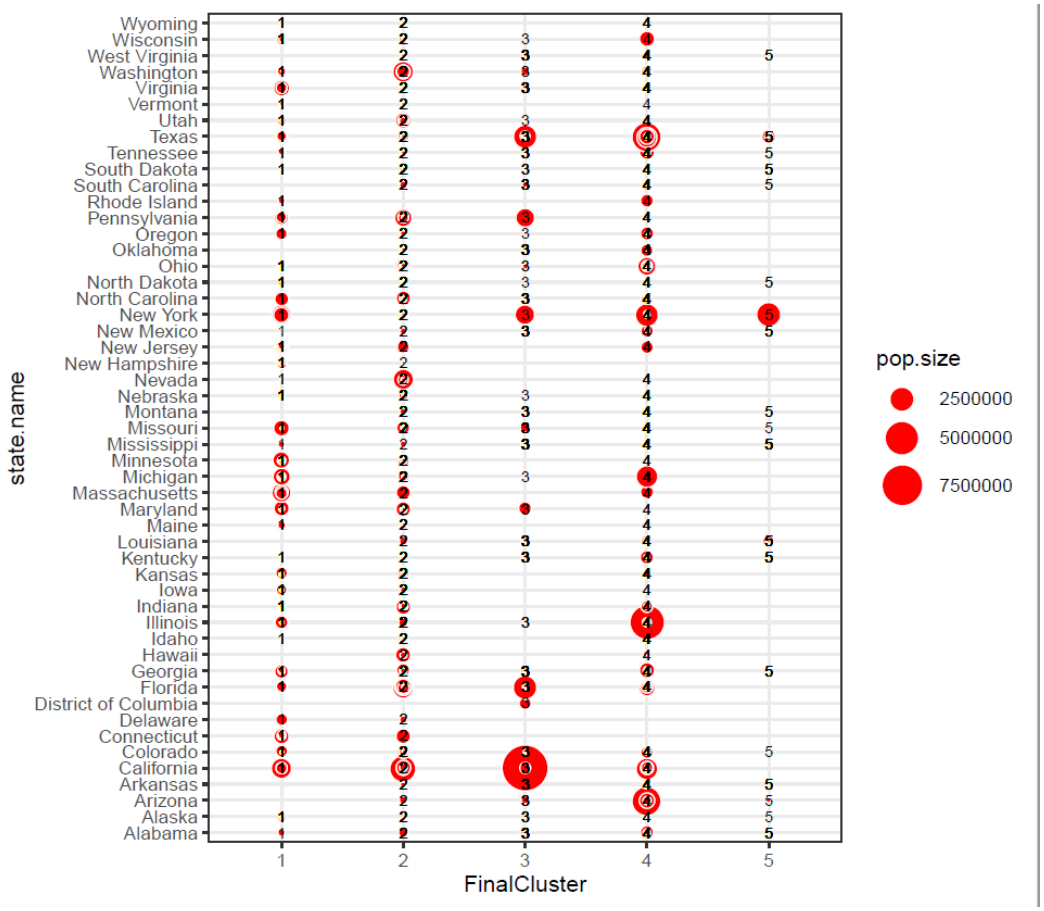
Computing k-means clustering in R

We can compute k-means in R with the kmeans function. Here will group the data into two clusters (centers = 5). The kmeans function also has an nstart option that attempts multiple initial configurations and reports on the best one. For example, adding nstart = 20 will generate 20 initial configurations



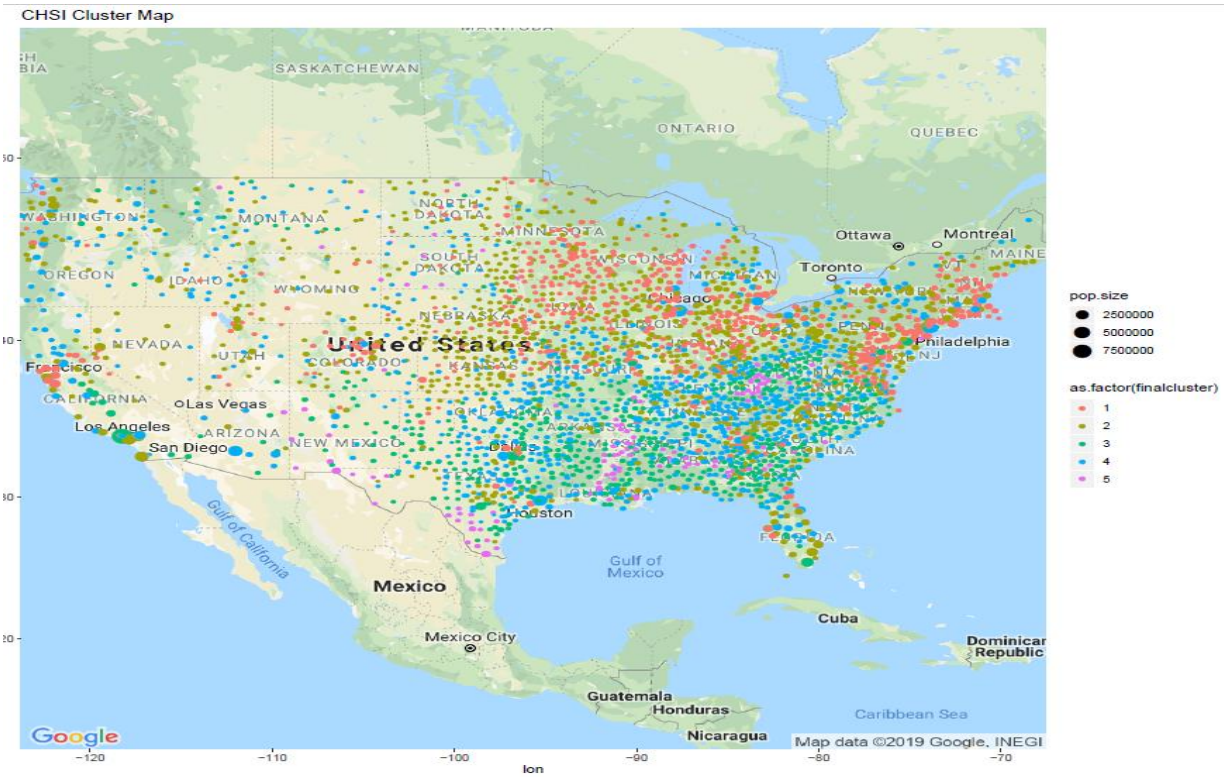
Extracting Results

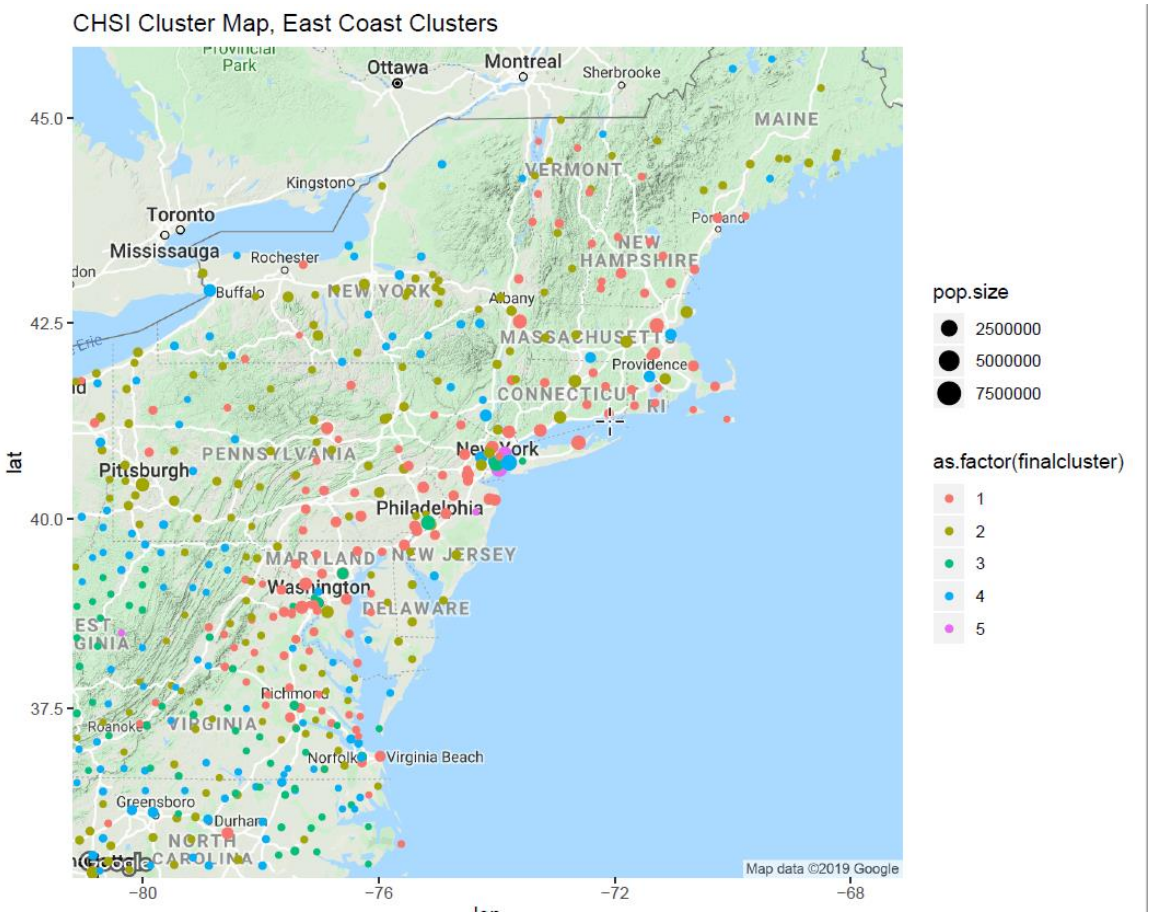
With most of these approaches suggesting 5 as the number of optimal clusters, we can perform the final analysis and extract the results using 5 clusters. In order to better understand these clusters, the research team utilized the fviz_cluster package, the ggplot2 package, and ggmap r packages for visualization.

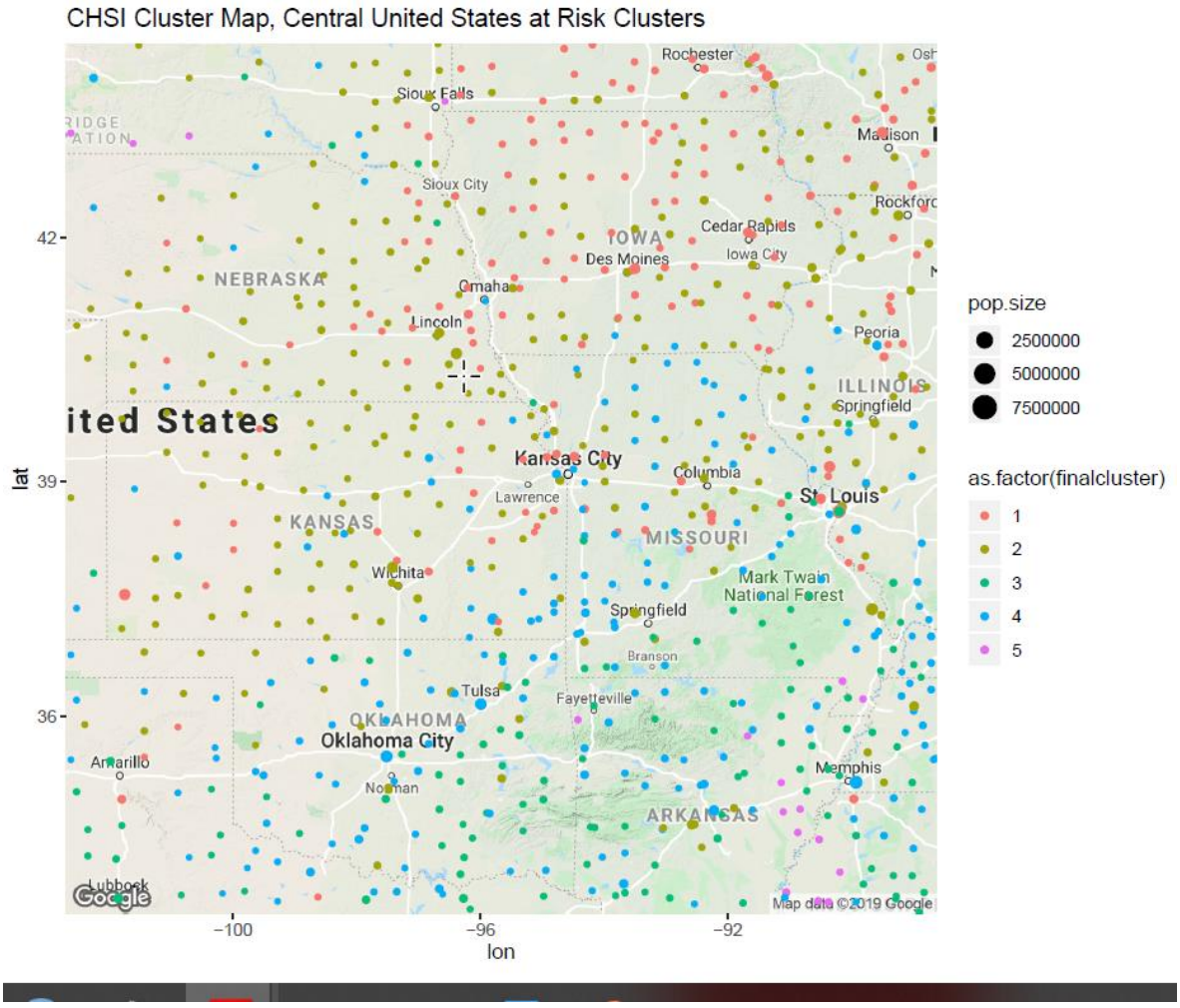


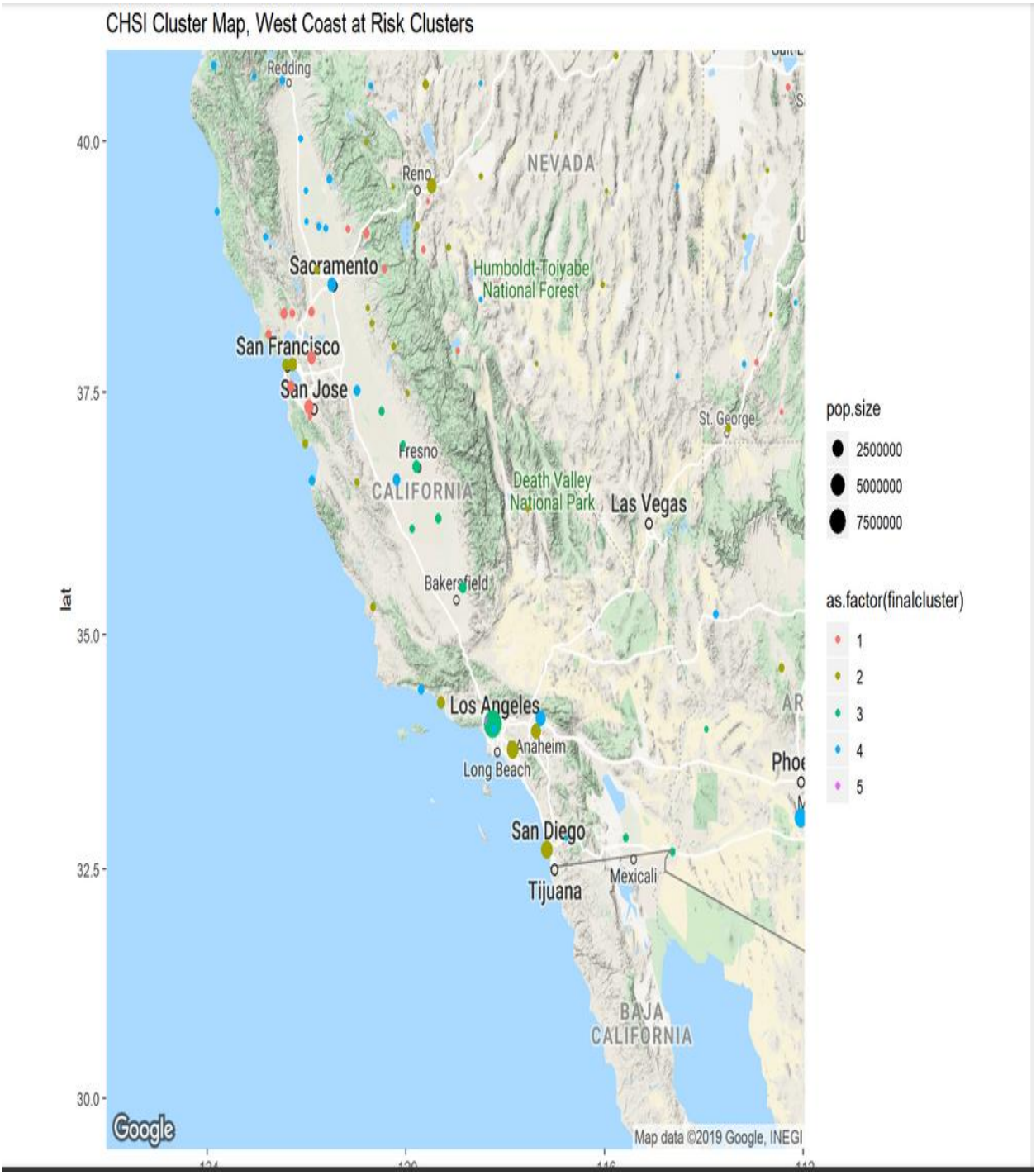
Final cluster chart, At Risk factors, CHSI

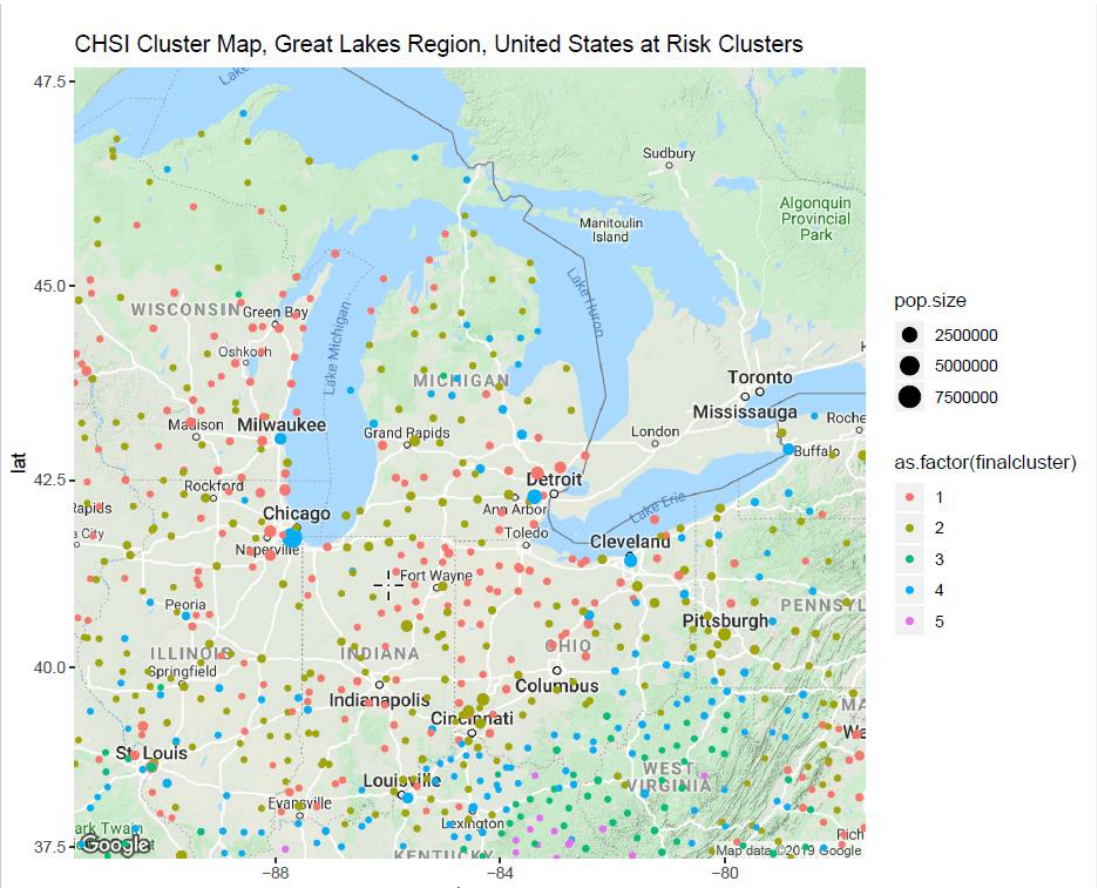
Given that the nature of our dataset is based upon a geographic plane, the research team found it expedient to demonstrate the 5 clusters of the data set on the map, as to better understand the distribution of these clusters via there geographic location, and to determine whether or not there was any explanation to that distribution in this unsupervised method. Utilizing the ggmap package; the following maps demonstrate the 5 clusters distribution; Whole, East Coast, West Coast, Great Lakes, South East Region, New York City Metro area.

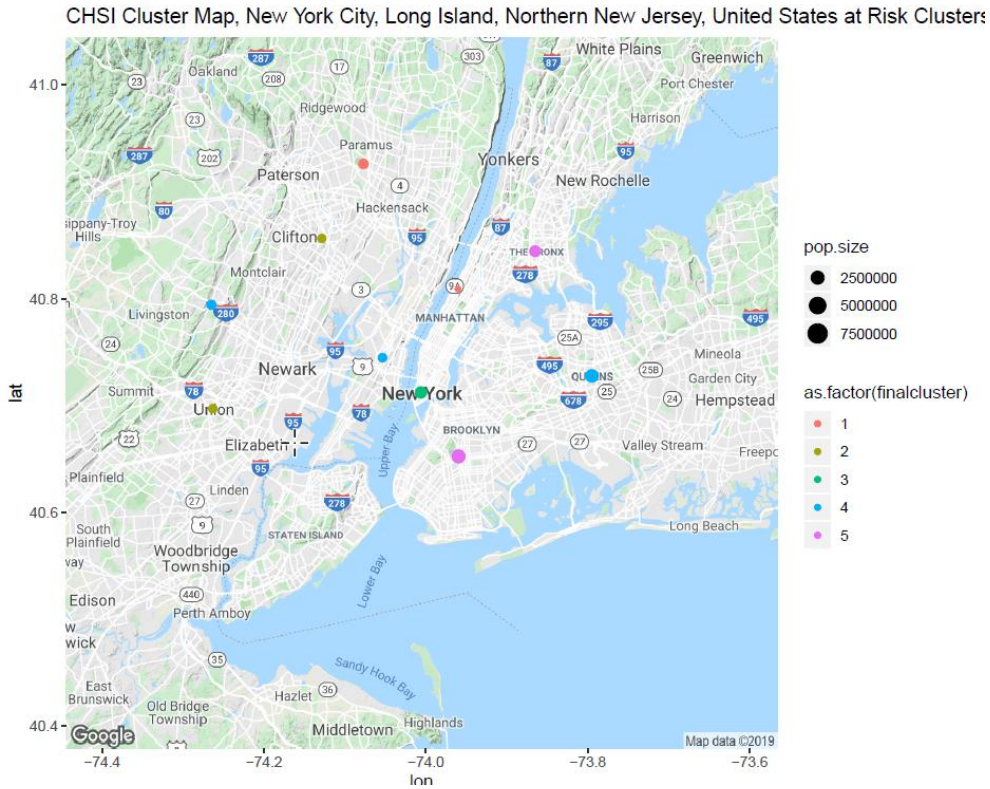

CHSI Cluster Map
SASKATCHEWAN
ONTARIO
QUEBEC
MONTANA
NORTH DAKOTA
MINNESOTA
SOUTH DAKOTA
WISCONSIN
MICHIGAN
OREGON
IDAHO
WYOMING
NEBRASKA
IOWA
ILLINOIS
NEVADA
UTAH
COLORADO
KANSAS
MISSOURI
Francisco
CALIFORNIA
Las Vegas
Los Angeles
ARIZONA
NEW MEXICO
San Diego
Dallas
Houston
Gulf of California
Gulf of Mexico
Mexico
Mexico City
Cuba
Guatemala
Honduras
Nicaragua
Caribbean Sea
Ottawa
Montreal
Toronto
Chicago
Philadelphia
MAINE
NJ
United States
Google
Map data ©2019 Google, INEGI
−120
−110
−100
−90
−80
−70
lon
pop.size
2500000
5000000
7500000
as.factor(finalcluster)
1
2
3
4
5


CHSI Cluster Map, East Coast Clusters
Provincial Park
Ottawa
Montreal
Sherbrooke
MAINE
VERMONT
Kingstono
Toronto
Portland
Mississauga
Rochester
NEW HAMPSHIRE
Buffalo
NEW YORK
Albany
MASSACHUSETTS
Providence
CONNECTICUT
RI
PENNSYLVANIA
New York
Pittsburgh
Philadelphia
MARYLAND
NEW JERSEY
Washington
DELAWARE
Richmond
Roanoke
VIRGINIA
Norfolk
Virginia Beach
Greensboro
Durham
NORTH CAROLINA
Map data ©2019 Google
45.0
42.5
40.0
37.5
lat
−80
−76
−72
−68
pop.size
2500000
5000000
7500000
as.factor(finalcluster)
1
2
3
4
5

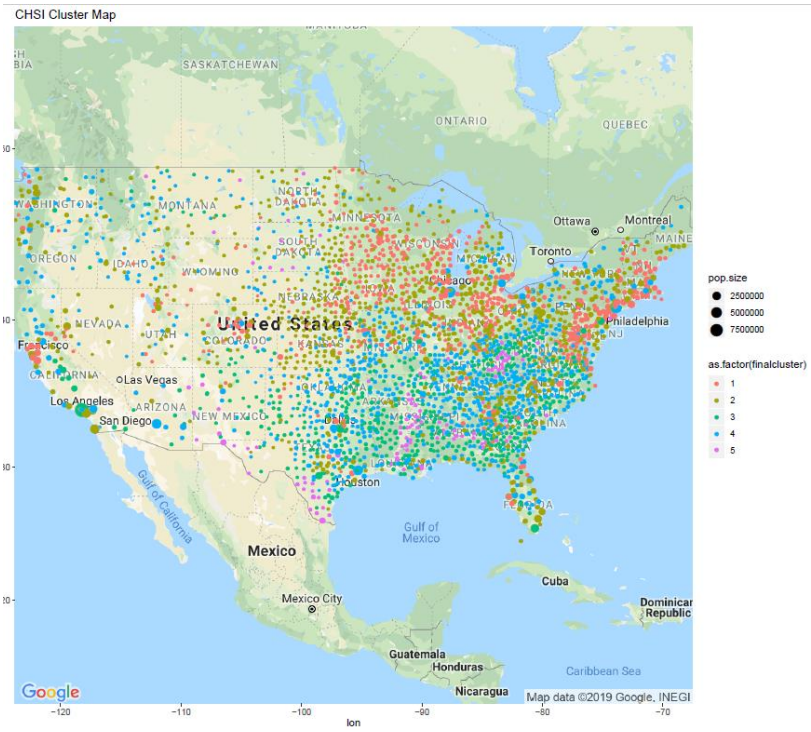

CHSI Cluster Map, Central United States at Risk Clusters
Rochester
Sioux Falls
Madison
Sioux City
Rockford
Cedar Rapids
IOWA
Des Moines
Iowa City
NEBRASKA
Omaha
Lincoln
Peoria
ILLINOIS
Springfield
ited States
Kansas City
Columbia
St. Louis
Lawrence
KANSAS
MISSOURI
Wichita
Mark Twain
National Forest
Springfield
Branson
Tulsa
Fayetteville
OKLAHOMA
Oklahoma City
Amarillo
Memphis
Norman
ARKANSAS
Lubbock
Google
Map data ©2019 Google
lat
42
39
36
−100
−96
−92
lon
pop.size
2500000
5000000
7500000
as.factor(finalcluster)
1
2
3
4
5


CHSI Cluster Map, West Coast at Risk Clusters
Redding
Reno
NEVADA
Sacramento
Humboldt-Toiyabe
National Forest
San Francisco
San Jose
St. George
Fresno
CALIFORNIA
Death Valley
National Park
Las Vegas
Bakersfield
Los Angeles
Anaheim
Long Beach
San Diego
Mexicali
Tijuana
BAJA
CALIFORNIA
Google
Map data ©2019 Google, INEGI
lat
40.0
37.5
35.0
32.5
30.0
pop.size
2500000
5000000
7500000
as.factor(finalcluster)
1
2
3
4
5


CHSI Cluster Map, Great Lakes Region, United States at Risk Clusters
lat
lon
pop.size
2500000
5000000
7500000
as.factor(finalcluster)
1
2
3
4
5
Sudbury
Algonquin Provincial Park
Manitoulin Island
Lake Huron
Lake Michigan
WISCONSIN
Green Bay
Oshkosh
Madison
Milwaukee
MICHIGAN
Grand Rapids
Toronto
Mississauga
London
Buffalo
Rockford
Chicago
Detroit
Ann Arbor
Toledo
Lake Erie
Cleveland
Naperville
Fort Wayne
Peoria
Pittsburgh
ILLINOIS
Springfield
INDIANA
OHIO
Columbus
Indianapolis
Cincinnati
St. Louis
Louisville
WEST VIRGINIA
Evansville
Lexington
Google
Map data ©2019 Google

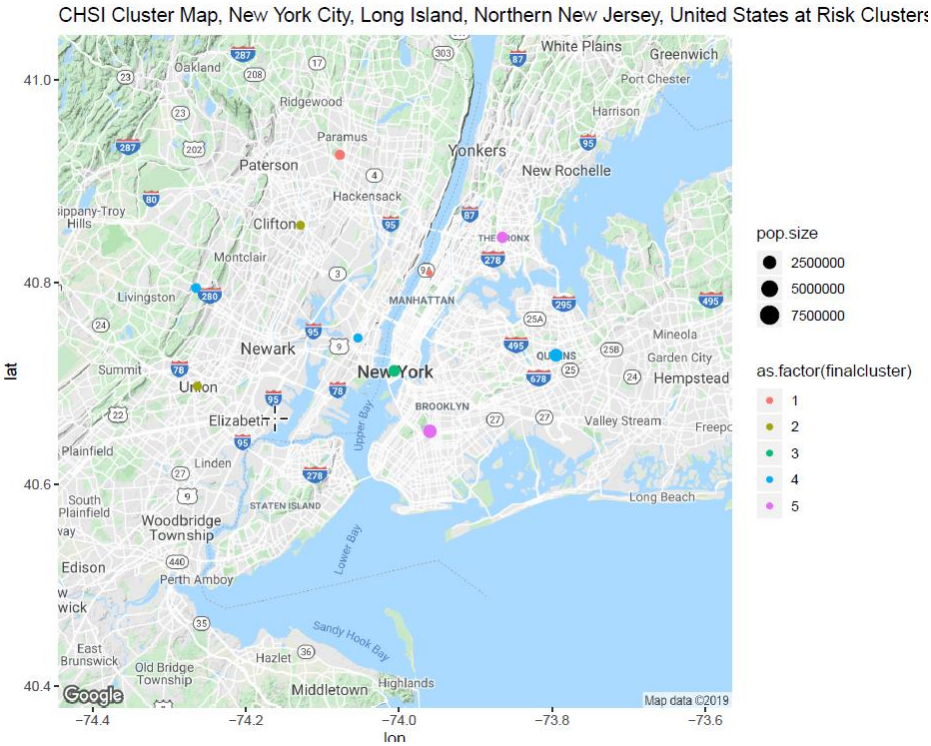

CHSI Cluster Map, New York City, Long Island, Northern New Jersey, United States at Risk Clusters
lat
lon
pop.size
2500000
5000000
7500000
as.factor(finalcluster)
1
2
3
4
5
White Plains
Greenwich
Port Chester
Harrison
Oakland
Ridgewood
Paramus
Yonkers
New Rochelle
Paterson
Hackensack
Clifton
Montclair
Livingston
MANHATTAN
THE BRONX
Mineola
Garden City
Hempstead
Newark
New York
QUEENS
Summit
Union
Elizabeth
BROOKLYN
Valley Stream
Upper Bay
Plainfield
Linden
Long Beach
South Plainfield
STATEN ISLAND
Woodbridge Township
Lower Bay
Edison
Perth Amboy
Sandy Hook Bay
East Brunswick
Old Bridge Township
Hazlet
Middletown
Highlands
Google
Map data ©2019

What are the clusters telling the research team?



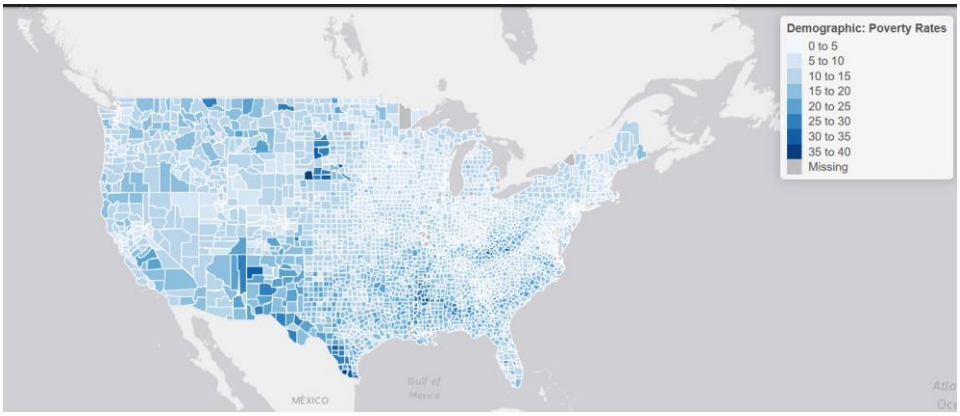
There is a patten that has been identified within the data. The at risk factors that contribute negatively to average life expectancies have cohesion of the data points within the clusters and separation between clusters.

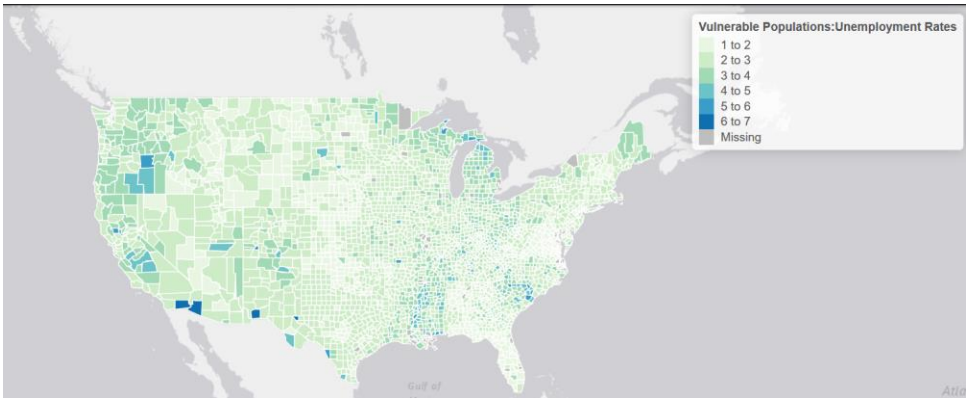
The nature of the clustering is not being determined by urban versus rural areas. In examination of the cluster's distribution around the nation, high urban areas like the west coast and east coast, the clusters contained within areas such as San Jose, San Francisco, and, Los Angeles – all belong to different cluster groups. Indeed, in examination of the New York City Metro area, this is expressed profoundly in the cluster differences found in Manhattan, compared to all the other boroughs and metropolitan surrounding areas, and indeed could be a benchmark as to the interpretation of the results. Look at the below example:

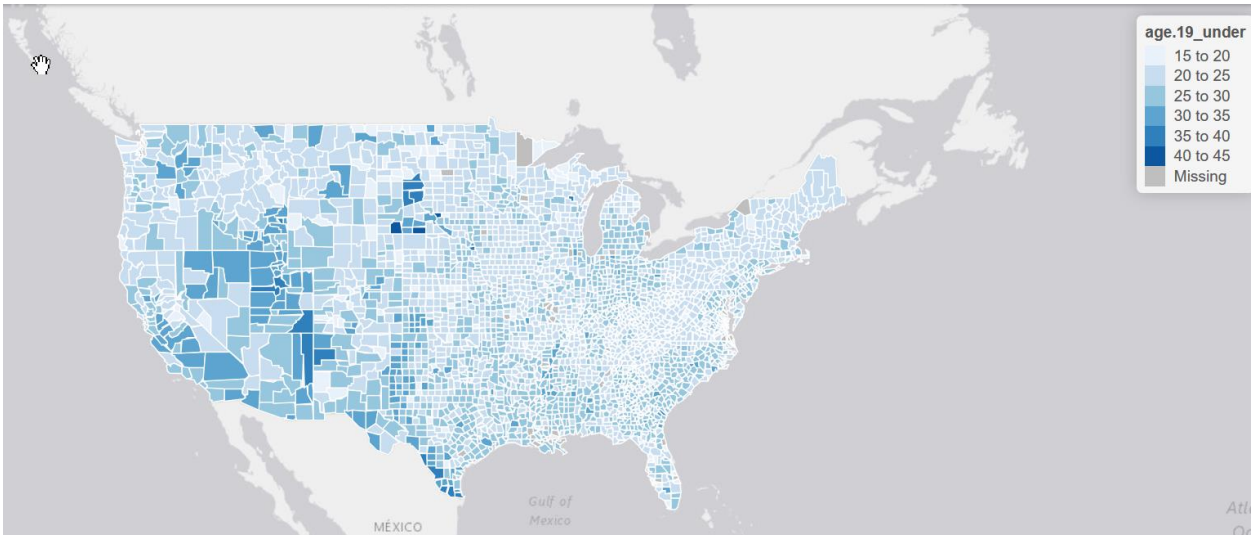


Manhattan is listed in Cluster 3, whereas The Bronx and Brooklyn are listed in Cluster 5. Queens is listed within Cluster 4. Northern New Jersey communities of Paramus and Clifton are all listed in different clusters than New York City, with the highland New Jersey Region near Livingston, having a correlation in clustering with Queens.

Finally, given the nature of the clustering, and, that one of the at risk community factors to the diminishment of Average Life Expectancy is Poverty, the research team found some interesting similarities in the nature of the distribution of poverty within the nation, versus, the nature of the clustering algorithm as determined by the unsupervised k-means analysis upon the data, as shown by a couple of at risk and demographic maps:

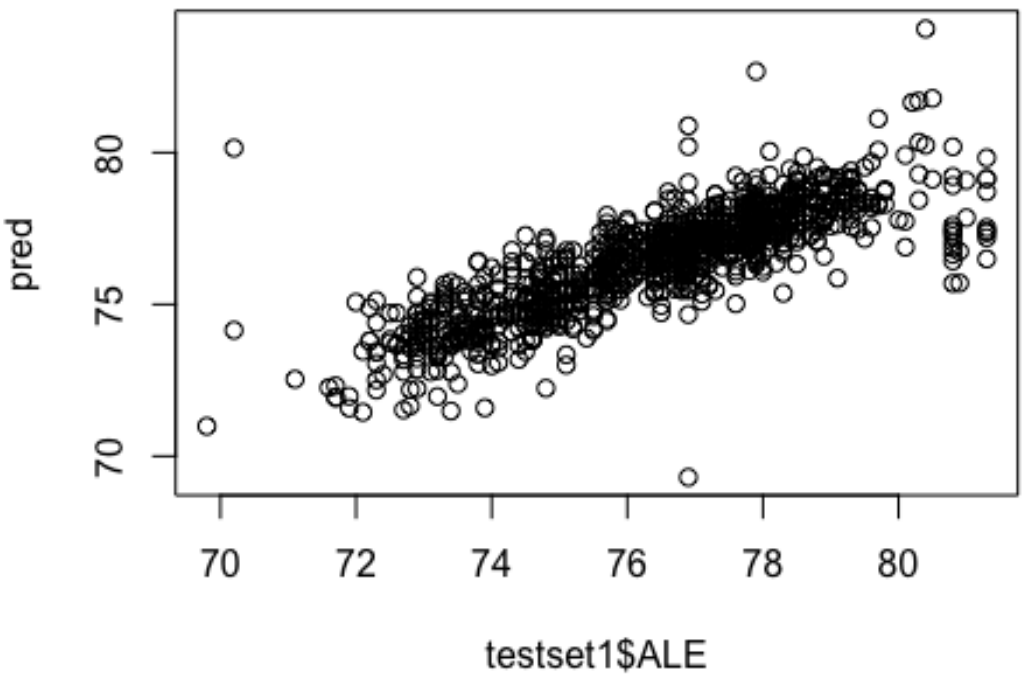






svm model

Several svm models were tried by varying kernel and cost attributes. Linear and Polynomial kernal gave an error of 0.03% for a cost of 1 and 50. So we decided to go with a cost of 1 with the linear model.




```
plot(svm1,trainset1)

# svm-linear, cost = 20
testset1 <- ndf[1:1000,]
trainset1 <- ndf[1001:3142,]
svm1 <- svm(ALE ~., data = trainset1, kernel = "linear", type = "eps", cost =
20)
pred=predict(svm1, testset1, type=c("eps"))
library(caret)
conf.linear <- data.frame(pred,testset1$ALE)
sqrt((mean(conf.linear$pred-conf.linear$testset1.ALE)^2))

## [1] 0.03056059
```

```
# svm-polynomial, cost = 50-0.03
testset1 <- ndf[1:1000,]
trainset1 <- ndf[1001:3142,]
svm1 <- svm(ALE ~., data = trainset1, kernel = "polynomial", type = "eps", co
st = 50)
pred=predict(svm1, testset1[,-38], type=c("eps"))
library(caret)
conf.polynomial <- data.frame(pred,testset1$ALE)
sqrt((mean(conf.linear$pred-conf.linear$testset1.ALE)^2))
```

```
## [1] 0.03056059
```

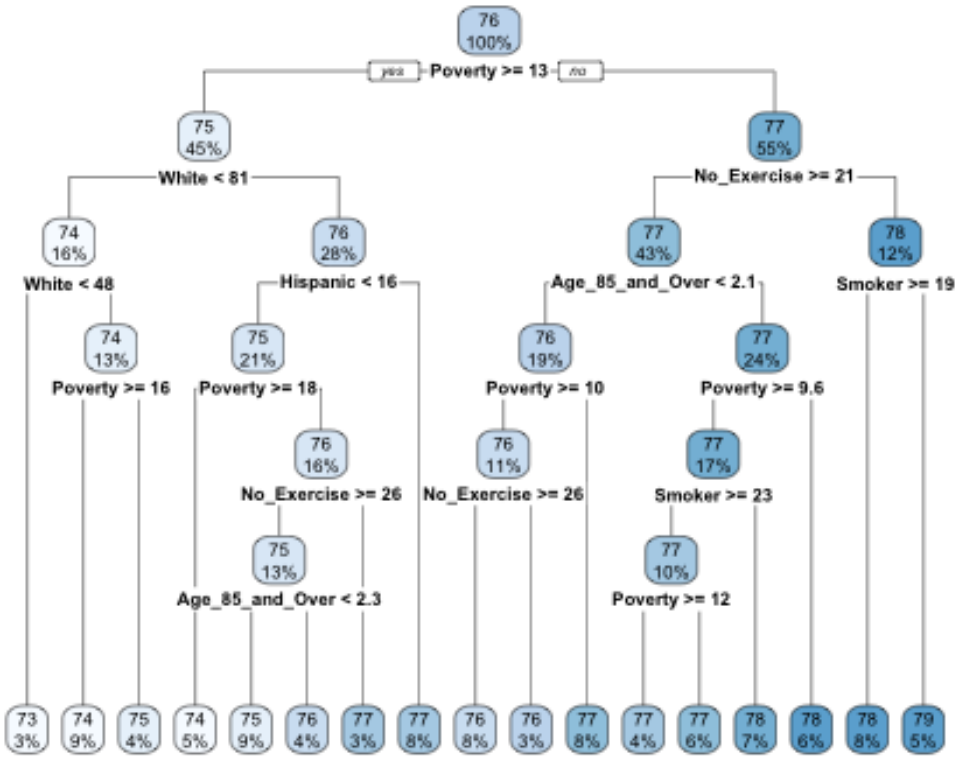
```
# svm-polynomial, cost = 50- 0.039
testset1 <- ndf[1:1000,]
trainset1 <- ndf[1001:3142,]
svm1 <- svm(ALE ~., data = trainset1, kernel = "radial", type = "eps", cost =
50)
pred=predict(svm1, testset1[,-38], type=c("eps"))
library(caret)
conf.radial <- data.frame(pred,testset1$ALE)
sqrt((mean(pred-testset1$ALE)^2))
```

```
## [1] 0.03981619
```

Decision Tree

A decision tree model was built using different rpart paremeters like maxdepth and minsplit. A model with maxdepth of 7,minsplit of 200 yeilds the best result with least error of 8.2% and an accuracy of 91.8%

```
#install.packages('rpart')
#install.packages('rpart.plot')
library(rpart)
library(rpart.plot)
set.seed(50)
dt <- rpart(ALE~.,trainset1,method  = "anova", control = rpart.control(cp = 0
, maxdepth = 7,minsplit = 200))
rpart.plot(dt)
```



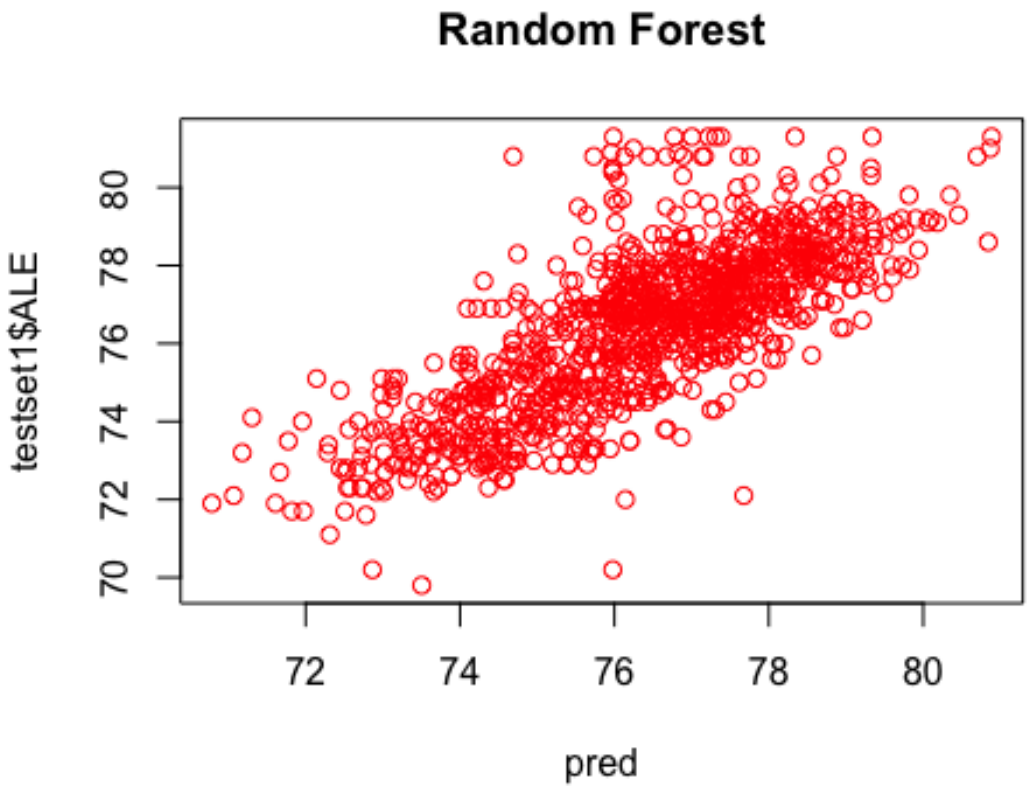

```
pred_dt <- predict(dt, testset1[,-38])
library(caret)
conf.dt <- data.frame(pred,testset1$ALE)
sqrt((mean(pred_dt-testset1$ALE)^2))

## [1] 0.08064625
```

Random Forest

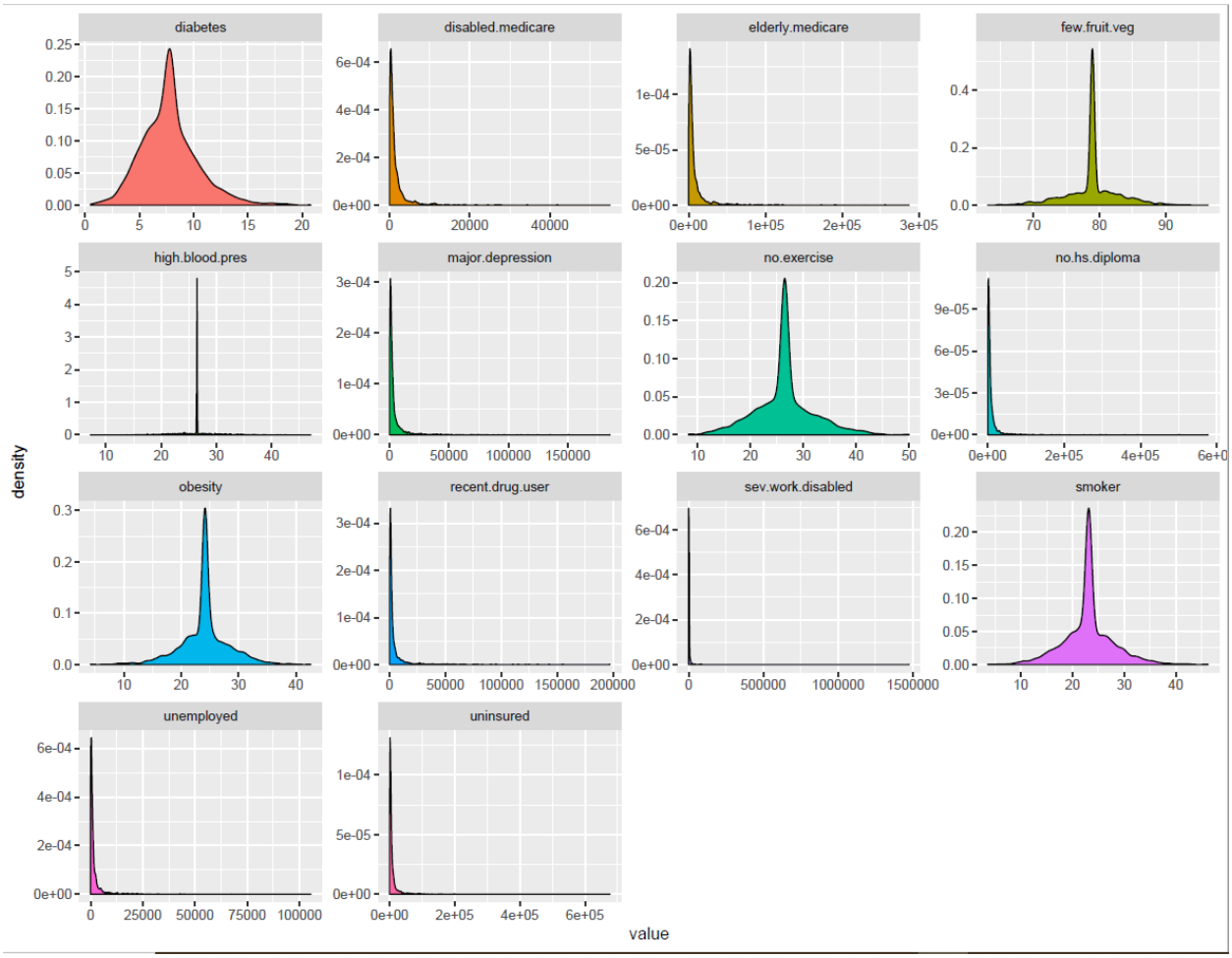
A random forest model was built using different number of trees 5,20,17. A model with 11 trees yeilded the best result with least error of 8.3% and an accuracy of 92%

```
#install.packages("randomForest")
#install.packages("caret")
library(randomForest)

library(caret)
set.seed(50)
rfm_chsi <- randomForest(ALE~., data=trainset1, ntree=11, na.action=na.exclud
e)
print(rfm_chsi)
```

```
##
## Call:
##  randomForest(formula = ALE ~ ., data = trainset1, ntree = 11,      na.act
ion = na.exclude)
##                Type of random forest: regression
##                      Number of trees: 11
## No. of variables tried at each split: 12
##
##           Mean of squared residuals: 1.393254
##                     % Var explained: 63.78

predRF <- predict(rfm_chsi, testset1[,-38])
conf.RF <- data.frame(pred,testset1$ALE)
sqrt((mean(predRF-testset1$ALE)^2))

## [1] 0.08227802

plot(pred,testset1$ALE, col='red', main="Random Forest")
```



Naïve-Bayes



Our Naïve-Bayes analysis proved more difficult to implement due to several factors (please refer to R file). While training, this author's R resulted in a "fatal error" and most all the quarter's variables were lost due to an unsaved workspace. This was a good learning moment, however, and showed us the necessary computing power for such a large analysis.

```
>
> #Predict on the dataset without passing the target feature
> predictions_mlr = as.data.frame(predict(NB_mlr, newdata = atRiskFinal))
>
> ##Confusion matrix to check accuracy
> table(predictions_mlr[,1],atRiskFinal$ale)

      66.6 68.6 69.6 69.8 70.1 70.2 70.4 70.8 70.9 71.1 71.2 71.4 71.5 71.6 71.7 71.8 71.9   72 72.1 72.2 72.3
66.6   330    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
68.6     0   68    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
69.6     0    0  138    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
69.8     0    0    0   69    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
70.1     0    0    0    0   70    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
70.2     0    0    0    0    0  140    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
70.4     0    0    0    0    0    0   70    0    0    0    0    0    0    0    0    0    0    0    0    0    0
70.8     0    0    0    0    0    0    0   70    0    0    0    0    0    0    0    0    0    0    0    0    0
70.9     0    0    0    0    0    0    0    0   70    0    0    0    0    0    0    0    0    0    0    0    0

      72.4 72.5 72.6 72.7 72.8 72.9   73 73.1 73.2 73.3 73.4 73.5 73.6 73.7 73.8 73.9   74 74.1 74.2 74.3 74.4
66.6     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
68.6     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
69.6     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
69.8     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
70.1     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
70.2     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
70.4     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
70.8     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
70.9     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0

      74.5 74.6 74.7 74.8 74.9   75 75.1 75.2 75.3 75.4 75.5 75.6 75.7 75.8 75.9   76 76.1 76.2 76.3 76.32287353
66.6     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0            0
68.6     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0            0
69.6     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0            0
69.8     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0            0
70.1     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0            0
70.2     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0            0
70.4     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0            0
70.8     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0            0
70.9     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0            0

      76.4 76.5 76.6 76.7 76.8 76.9   77 77.1 77.2 77.3 77.4 77.5 77.6 77.7 77.8 77.9   78 78.1 78.2 78.3 78.4
66.6     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
68.6     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
69.6     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
69.8     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
70.1     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
70.2     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
70.4     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
70.8     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
70.9     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0

      78.5 78.6 78.7 78.8 78.9   79 79.1 79.2 79.3 79.4 79.5 79.6 79.7 79.8 79.9   80 80.1 80.2 80.3 80.4 80.5
66.6     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
68.6     0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
```

Result

Our objective was to find factors that impact ALE so that the findings will aid health agencies or government to assist population in those areas. To do this we derived many models to understand the important factors.

Below is a summary of the models built to understand ALE and factors affecting it

# models	model	accuracy	error	cost
model 1	Linear regression	71.20%		-
model 2	svm-linear kernel	96.50%	4%	1
model 3	svm-linear kernel	96.94%	3%	20
model 4	svm-ploynomial	96.95%	3%	50
model 5	Random Forest	92.00%	8%	-
model 6	Decision Tree	91.80%	8%	-

- The results from **linear regression model** gave an understanding that 71.2% of the variability in ALE could be explained by the below factors which were used to build a model. The maximum impact being in the order of Ethnicity, Poverty, HIV, Obesity and Blood Pressure.

Poverty	-1.7E-01
White	5.6E-02
Black	2.3E-02
Native_American	6.8E-02
Asian	1.1E-01
Hispanic	3.4E-02
Recent_Drug_Use	1.6E-07
Toxic_Chem	-2.5E-09
No_Exercise	-4.2E-02
Few_Fruit_Veg	-1.2E-02
Obesity	-2.1E-02
High_Blood_Pres	-2.5E-02
Smoker	-5.0E-02
Diabetes	-4.6E-02
HIV	-1.7E-02
E_HeartDis	-9.9E-03
F_HeartDis	-2.2E-03

- **Correlation** analysis revealed that having No_highschool diploma is highly related to depression, being uninsured, unemployed, use of drugs that impact ALE strongly.
- Kmeans Result helped us understand segments in our population data
- Random Forest and Decision Tree models provided almost the same accuracy of ~92% with major decision making attributes being Poverty, Ethnicity, Exercising and Smoking.
- SVM with different kernels tried almost consistently gace a good result of ~97% accuracy with the factors being Ethnicity, Poverty, Exercising and Smoking.

Conclusions

1. The data set we chose really does fit it's intended purpose, which was to assist local health agencies with assessing the needs of their communities. In addition, armed with this data they would be able to create programs and services that would directly impact the overall health of their communities.
2. Irrespective of how you cut the data, we saw that a lack of education (defined as no HS diploma) had the single largest impact on overall health. Though it wasn't directly significant in the linear model, it had a high correlation to things like unemployment, drug use, and depression. These in turn contribute to a lower Average Life Expectancy (ALE). Programs that target education and/or gainful employment, especially in rural counties, would seem to have the largest impact.
3. In addition, communities with adverse behavioral or lifestyle choices (most notably those who don't exercise, those who eat few fruits/vegetables, and those who smoke) are statistically more at risk for premature death. These correlations (negative correlative

value to life expectancy) are within the individual's control and would benefit from additional support within the community.

4. A general observation of the project is that while we chose a data set that allowed for each individual to learn something or probe in a different direction (e.g. some thinking about cancer, some looking at mental health and others suicide rates) it created a challenge in focusing in on a cohesive data story. The team was often caught between applying things we had learned in class (tools - ensure we "check all the right boxes") and really understanding what the data was telling us. It was a great exercise to highlight the challenges in translating business needs (what do you want to know, how do you want to use the data) and the data side (coding) to ensure they are aligned.