IST 687 – GROUP 1

FINAL PROJECT REPORT

Group Members: Austin Beller, Luke Miller, Samuel Rogers, Randall Taylor, Todd Tetreault

Table of Contents

INTRODUCTION

DATA ACQUISITION, CLEANING, AND TRANSFORMATION

UNDERSTANDING AND ANALYZING THE DATA

USE OF MODELING TECHNIQUES

IST 687 – Group 1 Final Project Report

INTRODUCTION

1. Project Background and Description

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.

2. About the Data

The data set we chose was the Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease, and Cancer: https://healthdata.gov/dataset/community-health-status-indicators-chsi-combat-obesity-heart-disease-and-cancer

Community Health Status Indicators (CHSI) to combat obesity, heart disease, and cancer 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 goal of CHSI is to give local public health agencies another tool for improving their community's health by identifying data resources and facilitating the setting of priorities. The CHSI report contains over 200 measures for each of the 3,141 United States counties. Although CHSI presents indicators like deaths due to heart disease and cancer, it is imperative to understand that behavioral factors such as obesity, tobacco use, diet, physical activity, alcohol, and drug use, sexual behavior, and others substantially contribute to these deaths.

3. What are the questions you were looking to answer?

The data set we chose had 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.

The main questions we were looking to answer were the following:

1. What is the make-up of the population (age, ethnicity, etc.)
2. What is the average life expectancy (ALE)? Is there a wide variation in the ALE across the U.S.?
3. What are the leading risk factors that are contributing to premature death?
4. What correlations are the most relevant? The data set has lots of information around vulnerable populations and risk factors, we would like to better understand which correlate the highest with death (as that is what the data is trying to help prevent - premature death)
5. Lastly, would we be able to use these correlations to create models to predict, which would then allow for programs/interventions to be administered and tracked to ensure impact?

DATA ACQUISITION, CLEANING, TRANSFORMATION

4. Describe your data acquisition process

The data was provided via a government website: https://healthdata.gov/dataset/community-health-status-indicators-chsi-combat-obesity-heart-disease-and-cancer

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.

5. What data did you select, all or subset, and why

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.

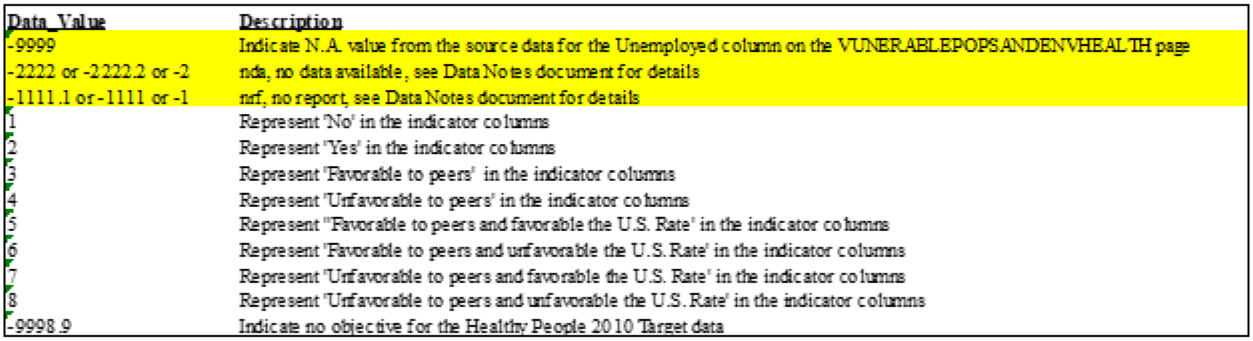
The subset of data we selected included health afflictions, descriptive characteristics, and risk factors. Health afflictions included diseases such as cancer, high blood pressure, and various STIs. Descriptive characteristics in the data included identifiers like poverty, lack of high school education, 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.

6. What was your initial quality assessment

When we first reviewed the data, it was spread across multiple excel documents and the column headers required a key to identify their meaning. The data was free of erroneous punctuation and characters however it contained a mix of characters and numbers with numeric counts and percentages. This required some recalculation when comparing data such as the death count to the percentage of smokers in a county. Using population size this issue was resolved. The numbers were also not recognized as numeric values, initially. Lastly, there were multiple codes littered throughout the dataset which indicated a value was NA. This required some translation code to avoid miscalculations when the values were converted to numeric. The data set did provide a legend that was useful in cleaning these NA or erroneous values (**Table 1**).

Table 1: Legend for Not Available values in our data.



Data_Value	Description
-9999	Indicate N.A. value from the source data for the Unemployed column on the VUNERABLEPOPSANDENVHEALTH page
-2222 or -2222.2 or -2	nda, no data available, see Data Notes document for details
-1111.1 or -1111 or -1	nrf, no report, see Data Notes document for details
1	Represent 'No' in the indicator columns
2	Represent 'Yes' in the indicator columns
3	Represent 'Favorable to peers' in the indicator columns
4	Represent 'Unfavorable to peers' in the indicator columns
5	Represent "Favorable to peers and favorable the U.S. Rate' in the indicator columns
6	Represent 'Favorable to peers and unfavorable the U.S. Rate' in the indicator columns
7	Represent 'Unfavorable to peers and favorable the U.S. Rate' in the indicator columns
8	Represent 'Unfavorable to peers and unfavorable the U.S. Rate' in the indicator columns
-9998.9	Indicate no objective for the Healthy People 2010 Target data

7. Data dictionary

In order to give a sense of the data, included is the header data for each table used in our project. Note the full data dictionary (provided via the link above) would have produced a ~75-page document. Samples of relevant tables used are included in **Appendix 1.**

8. Data Frame Structure

The complete dataset (‘combined’) has 3,141 observations (rows) and 140 variables (columns). The output when examining the “Structure” of our Combined Dataset is included in **Appendix 2.**

9. Data Cleansing

The main cleanup required was specific to how we correctly handle NAs. In our data, NA was stored as a negative number. Within R we ran an lapply to substitute NA for all negative numbers. That gave us a dataset with NA within it.

For some analyses, we used the data including NAs. We created a separate working data frame that replaced the NAs with means of the columns. Some visualizations were better with NAs included, since it shows at a glance both the data we are representing as well as the cases when it wasn’t available.

Similarly, some of our data was numeric (absolute counts), while others were represented as percentages. Since county populations vary wildly, we needed to turn absolute counts into percentages (to compare to percent data). For certain analytics we chose the opposite approach; translate percentages to counts in order to better compare variables.

DESCRIPTIVE STATISTICS

10. Demographic Statistics

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.

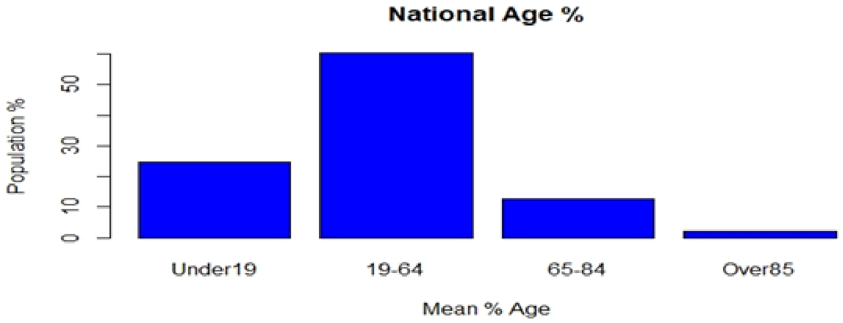


Figure 1: Distribution of age in the United States

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**).

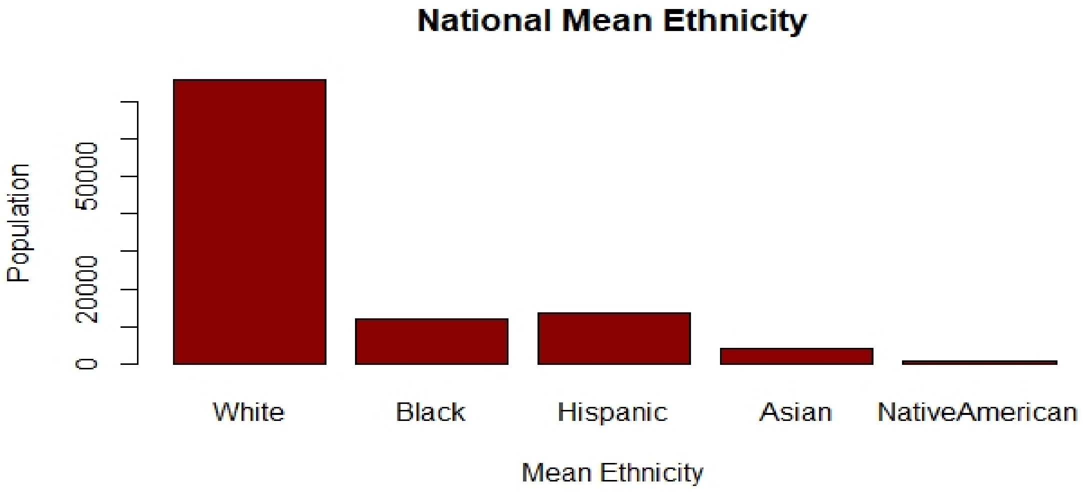


Figure 2: Barchart showing the national distribution of ethnicity by count

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).

11. 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, we imagine, 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).**

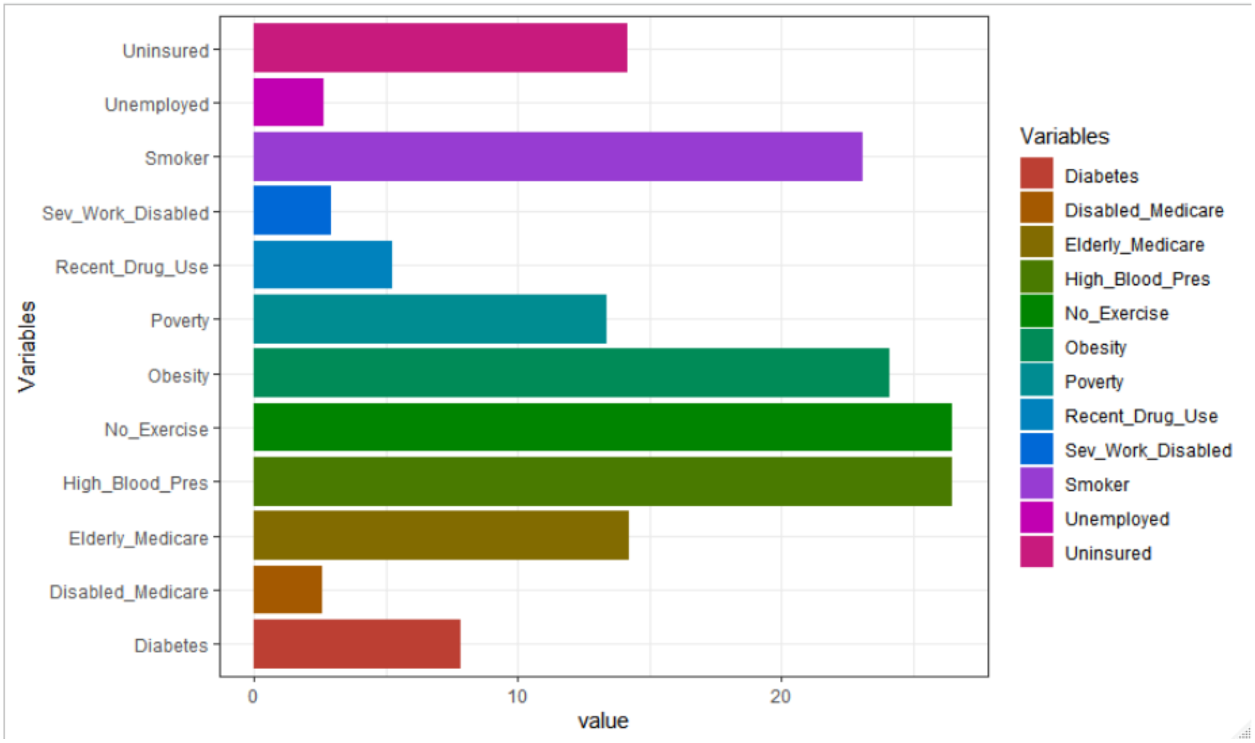


***Figure 3**: Shows the leading risk factors contributing to premature death*

In analyzing the life expectancy of the United States, we took summary statistics to learn more about the break down of each county. A scatter plot of this can be seen in **Figure 4**. The first quartile of the United States Average Life Expectancy (ALE) is 75 years old **(Table 2)**.

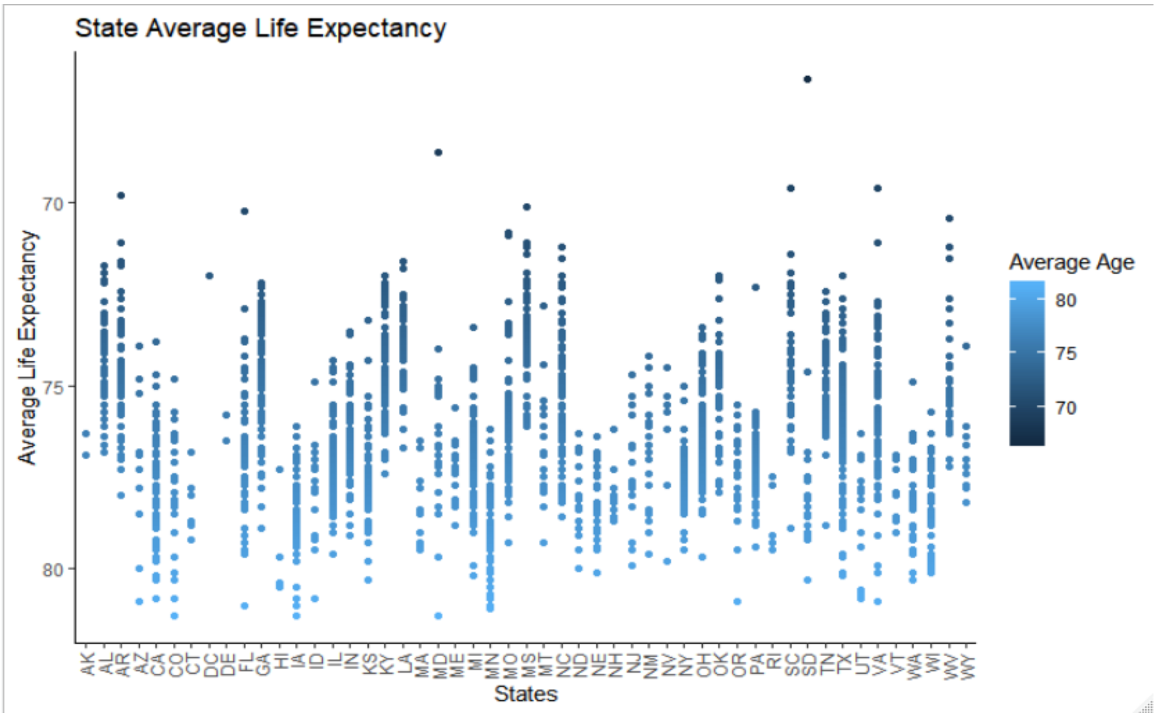


Figure 4*: Shows the average life expectancy of each county, grouped by state*.

Table 2: Summary statistics on the average life expectancy of U.S. counties

United States Average Life Expectancy					
Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
66.60	75.00	76.50	76.32	77.70	81.30

The states shown in **Figure 5** are at the highest risk as those states have counties with an average life expectancy below the nation's 1st quartile.

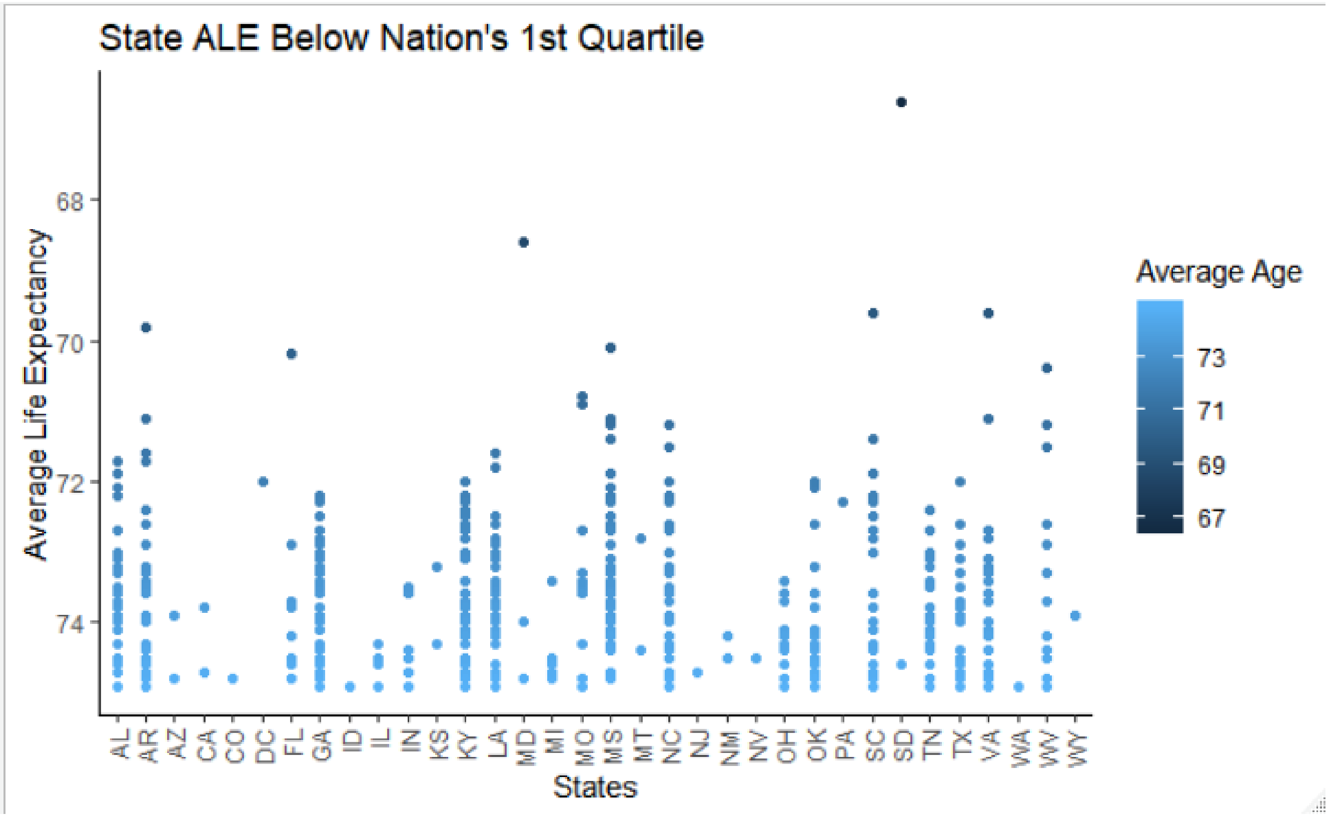


***Figure 5**: Shows average life expectancy of each county, grouped by state, which has an average life expectancy less than the United States' first quartile of 75 years old.*

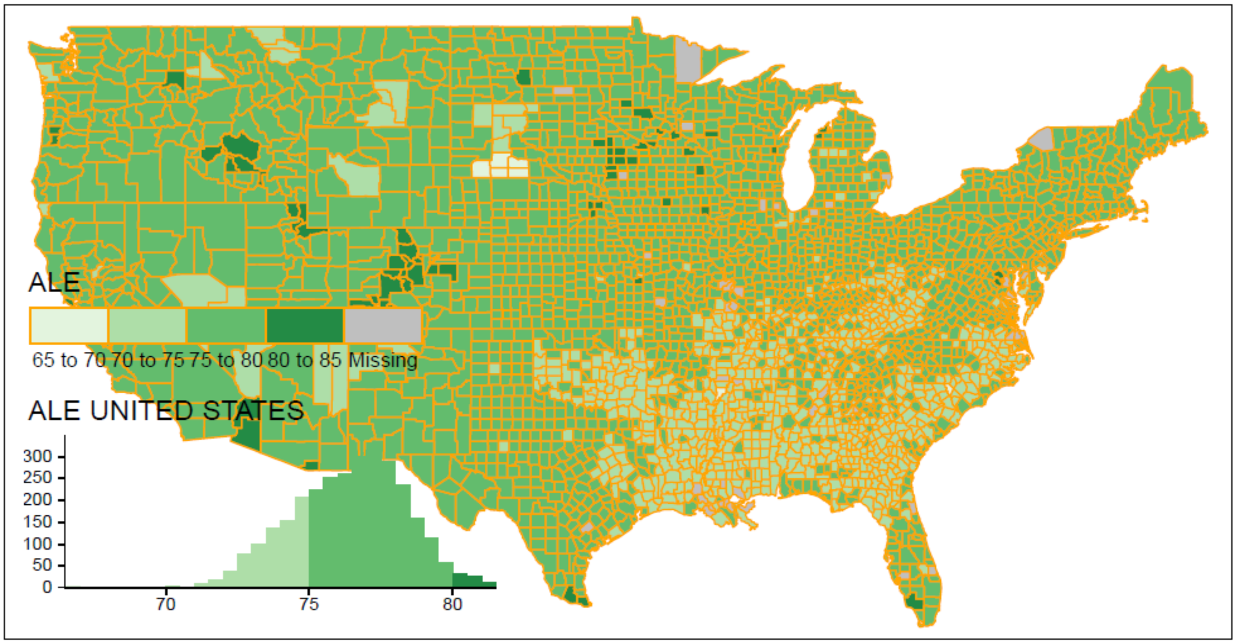
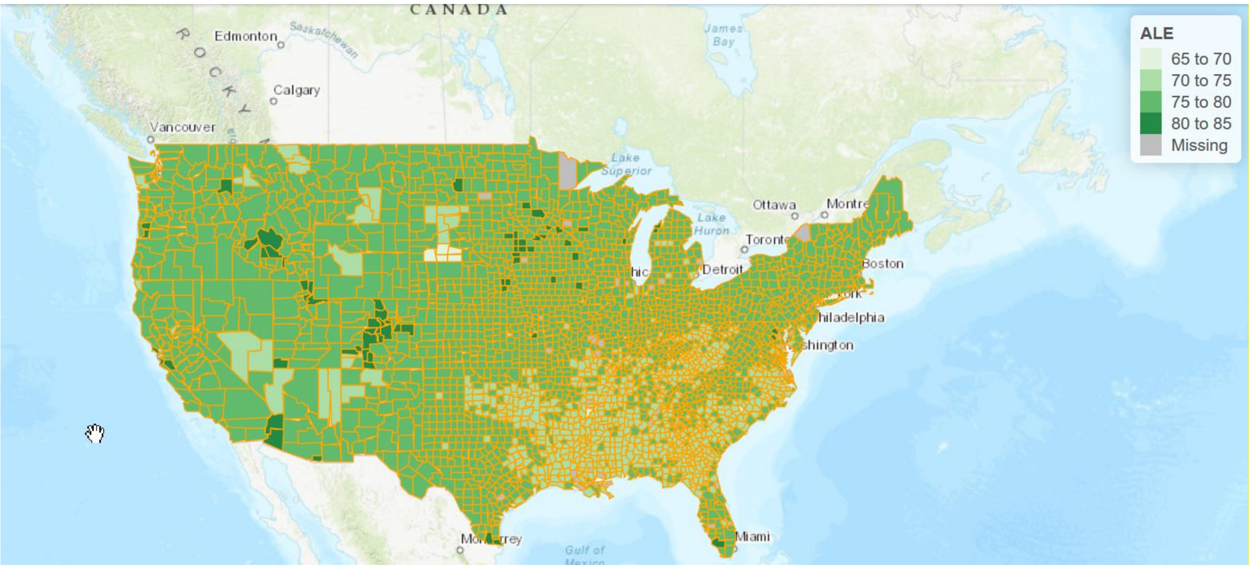


Figure 6 (above) and 6a (below): *Shows a visual of ALE by state/county*



As can be seen above (Figure 6 and 6a), the map allows for a double click/zoom at the state/county level. This gives a better sense of how a county compares nationally.

In **Figure 7** (below), we've zoomed to New York state and shown the ALE of Onondaga County, NY (where Syracuse is located), which is 77.60 years.

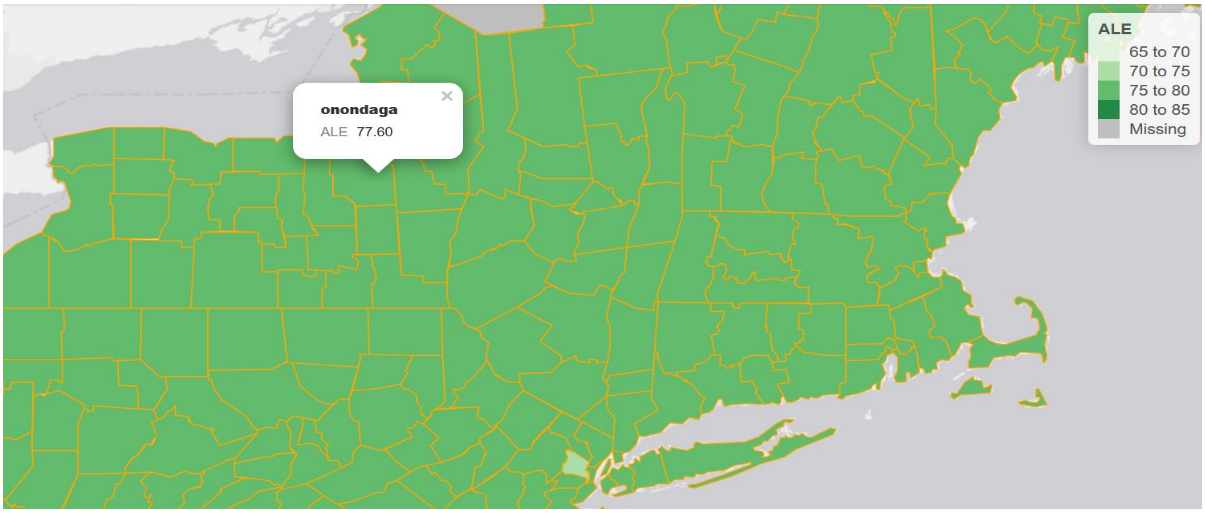


Figure 7: *Average Life Expectancy for Onondaga County (where Syracuse is located)*

Using the zoom tools learned in class we took this all the way down to a local level to better view two counties within NY state **(Figures 8 and 9).**

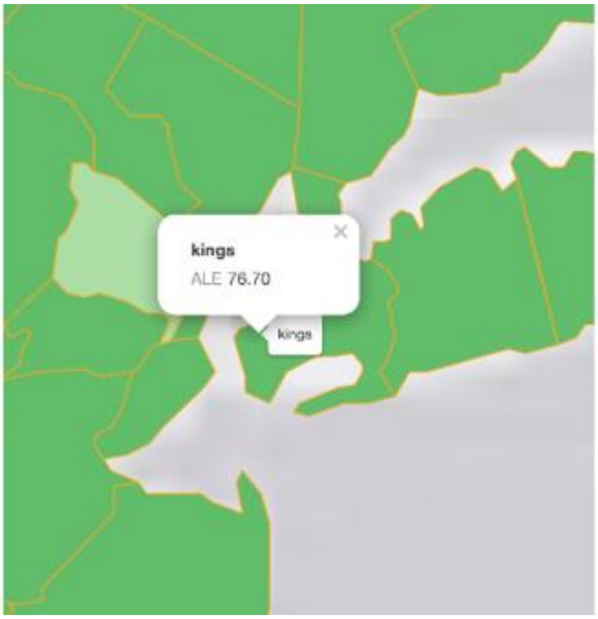


Figure 8

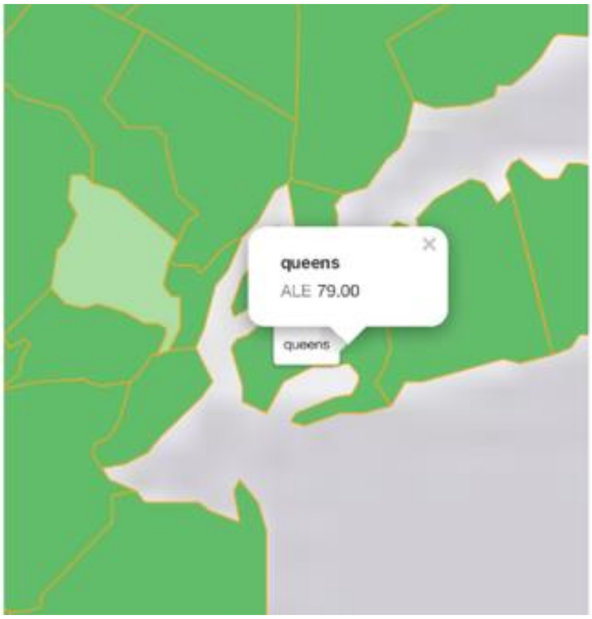
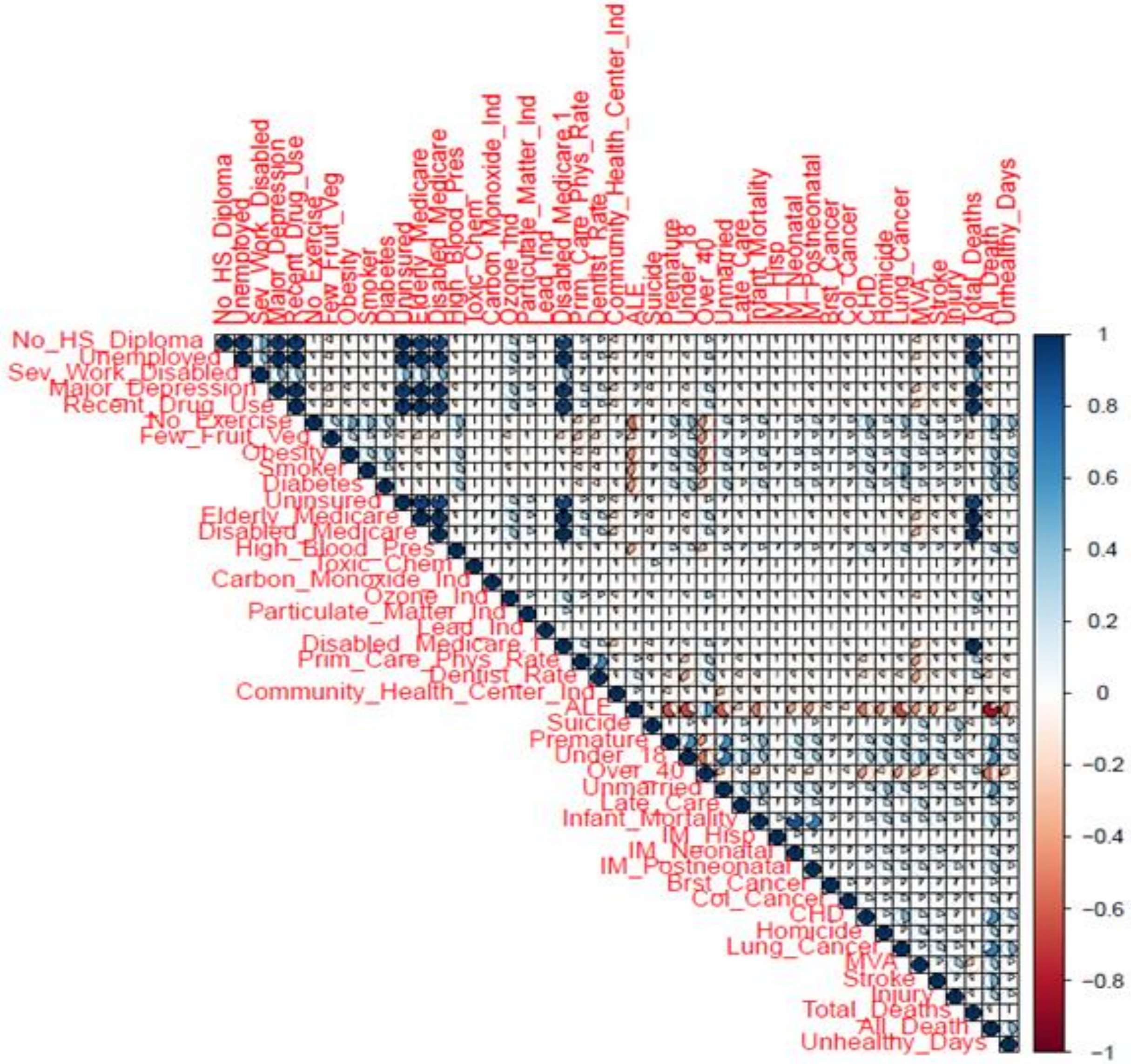


Figure 9

Figures 8 and 9: *State of New York, locations Kings County (Figure 8) and Queens County (Figure 9).*

12. Correlations

First, was to create an overall graphic of all the correlations. We ran a correlation matrix, individual plots and point analytics (a vs b). As you can see in the data set, this was a challenge as there is a great deal of information to bring to life. The team decided to move away from the matrix and go with the following visual **(Figure 10)**, which helped direct some initial hypotheses and queries.



Figures 10: *Correlations across all risk factors and causes of death*

The beauty of working with such a broad data set is that it allowed each member to investigate the data in different ways. The numerous correlations available for analysis are evident in the correlation matrix **(Figure 10)**. At the beginning of this paper in the Descriptive Statistics Analysis, members of the team were charged with finding strong correlation values between factors presented within the data. Within the bounds of those experiments and investigations, some interesting findings were determined about the data.

Figure 11 demonstrates an early interpretation of our data story which dove into the statistics of poverty in the United States. Although not the main focus of our project, the visualizations were interesting to examine. This static map of the United States used the percentage of the population which found itself impoverished. The “Poverty” values from our dataset were merged into a spatial data frame for mapping purposes.

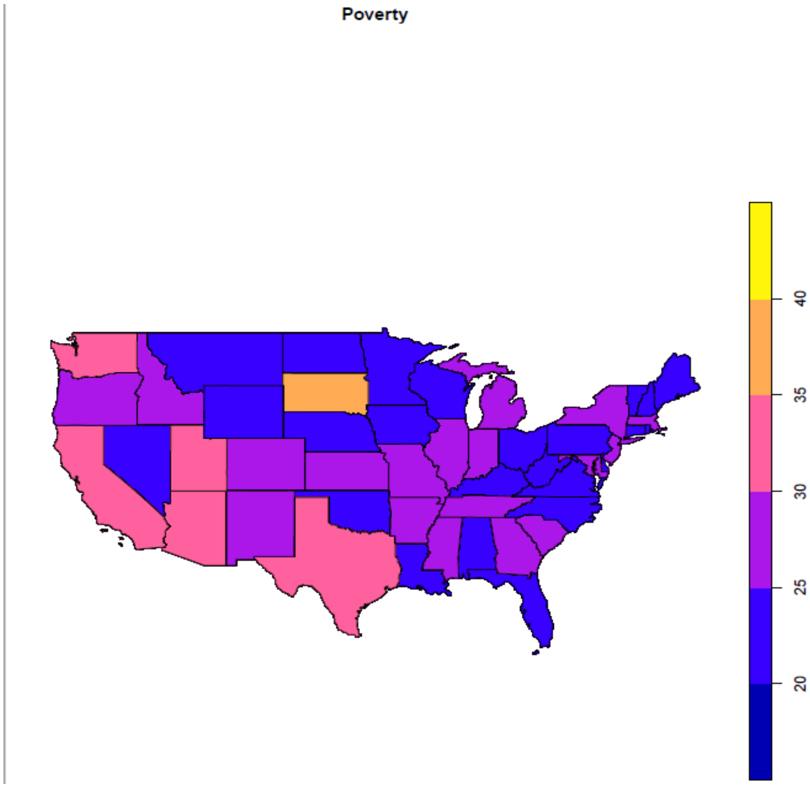


***Figure 11a**: Poverty Visualization (RT)*

The study into poverty shows South Dakota as the greatest poverty-stricken state in the US by population percentage (**Figure 11a**). **Figure 11b** breaks the united states into counties and the dense clustering of dark red counties in South Dakota shows why it has the greatest impoverished population by state.

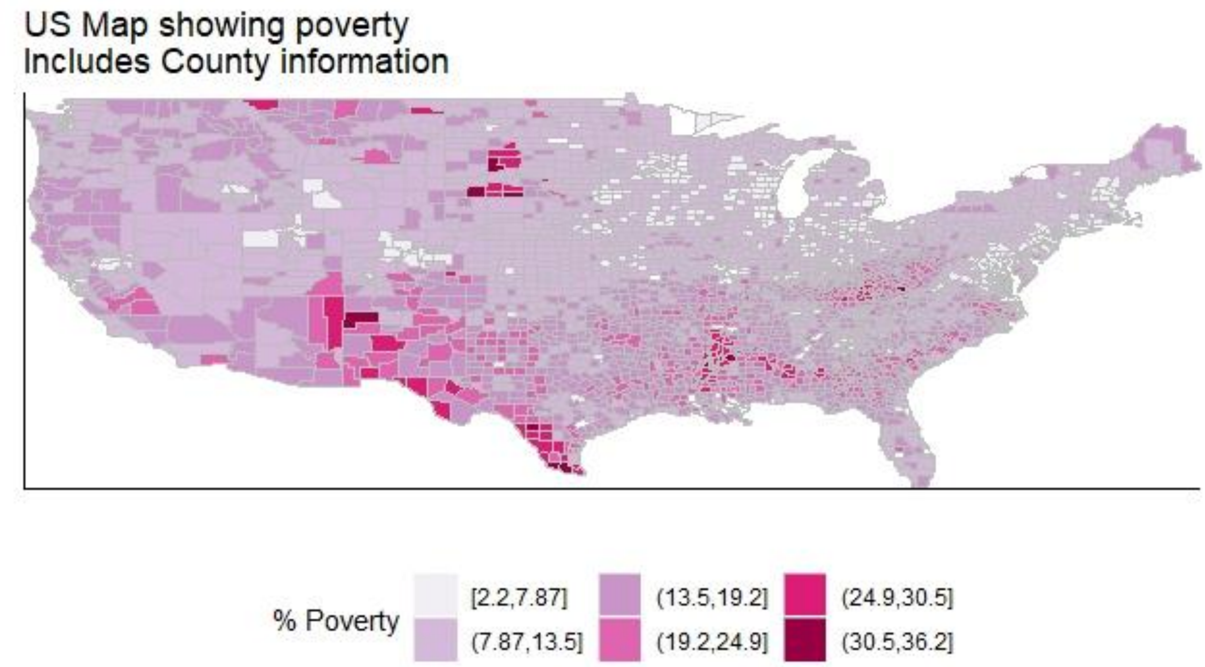


***Figure 11b**: Poverty Visualization county level (AB)*

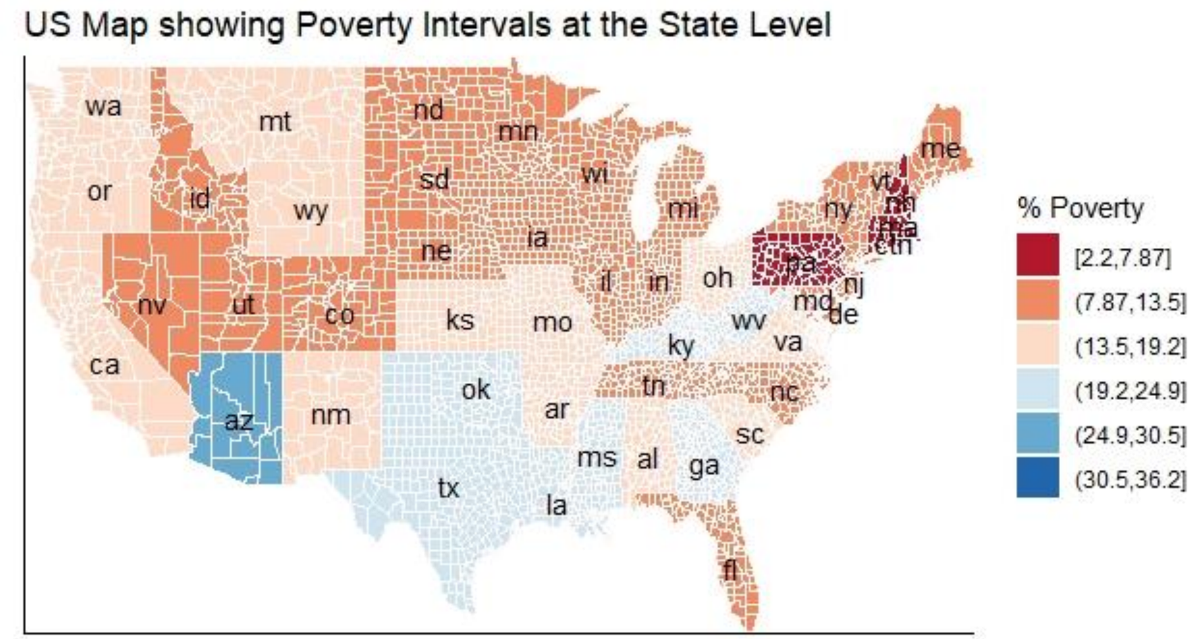


***Figure 11c**: Poverty Visualization state level (AB)*

In our study of **Figure 11b & 11c,** we noticed states were not made up of densely clustered counties of the same poverty percentage.

As our grander data story developed we learned in creating **Figure 12** that there was little direct correlation between Poverty and ALE. A significant correlation was found between ALE and No_Exercise. No_Exercise has the strongest correlation to poverty and No_Excercise has a very strong correlation to High_Bloodpressure. Bringing these all together, High_Bloodpressure and No_Exercise both have strong correlations to ALE, so as a co-variable, Poverty and ALE can be argued that they are still correlated.

As the study continued, and members were attempting to interpret what the data was trying to convey, there were some interesting ‘happy failures’. We call these failures only in that they didn’t directly answer our data questions. They were, however, worth pursuing a deeper analysis into, as they provided valuable insight into the data. One of these surprising failures was the lack of significant correlation Suicide had with the other variables.

There are a couple of different interpretations that could be made from these findings. Given the controversial and often stigmatized nature of suicide, it might be difficult to collect an accurate count of these instances. Laws and regulations change across the United States which can also affect the diagnosis for cause of death to be suicide. It seems intuitive that Major Depression and Suicide would be highly correlated, and yet they’re not. Perhaps when a patient seeks medical attention to be diagnosed as Majorly Depressed, they also tend to seek treatment or therapy to help ‘cure’ them of their mental issues as well. By this we mean, Majorly Depressed people receive this diagnosis and also seek treatment. This could be a reason why the people who tend to be diagnosed with Major Depression don’t tend to commit suicide.

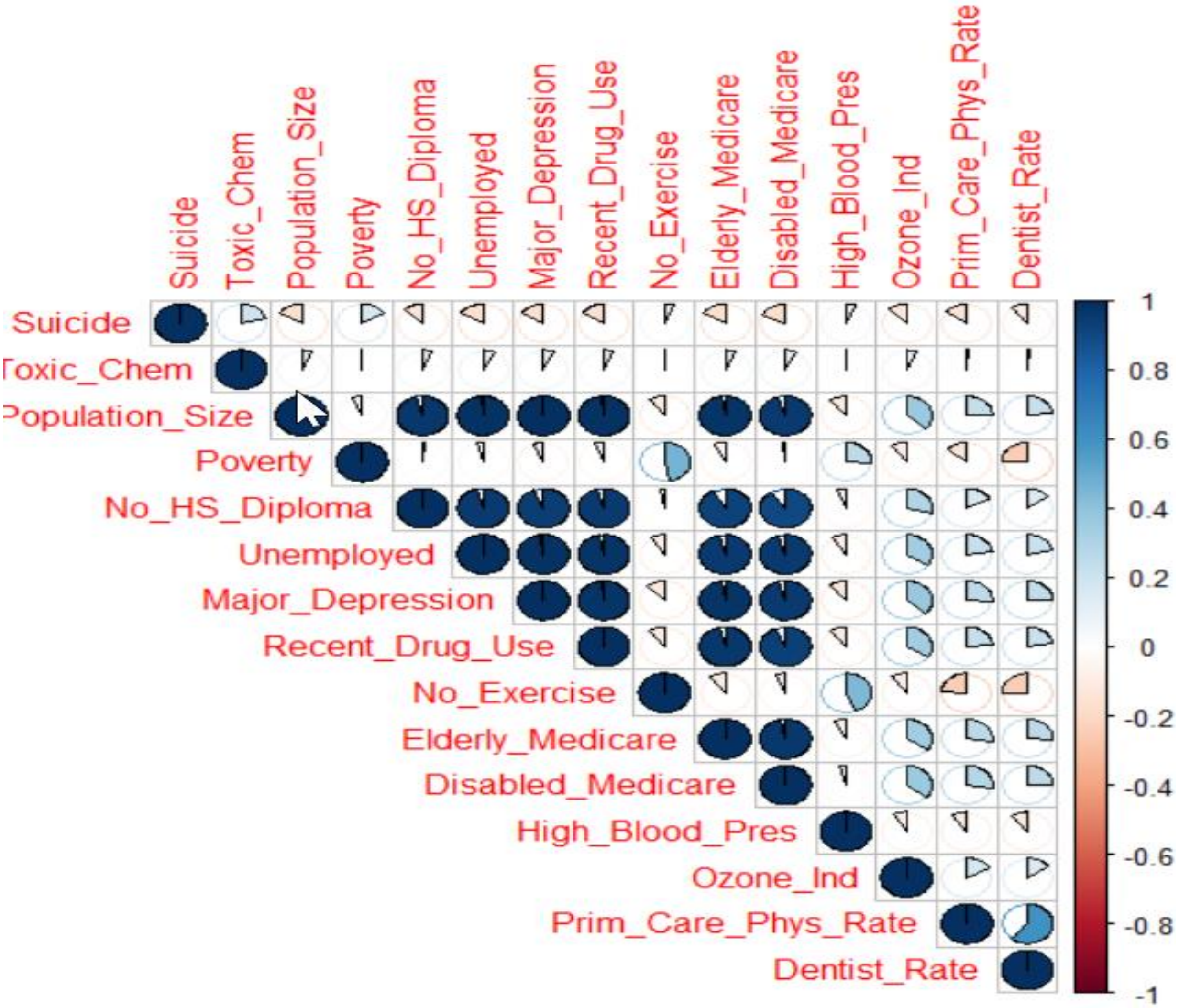


Figure 12: *Cut of correlation data focused on suicide*

Worth noting: while not pinging at a statistical significance of +/- 0.50, suicide's highest correlation was to that of Toxic Chemical and Poverty. Both risk indicators had a positive correlation of about 0.20 as demonstrated in **Figure 13a and Figure 13b**.

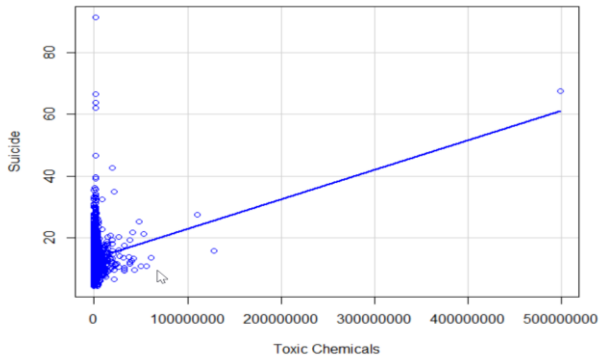


Figure 13a: Suicide~Toxic Chemicals

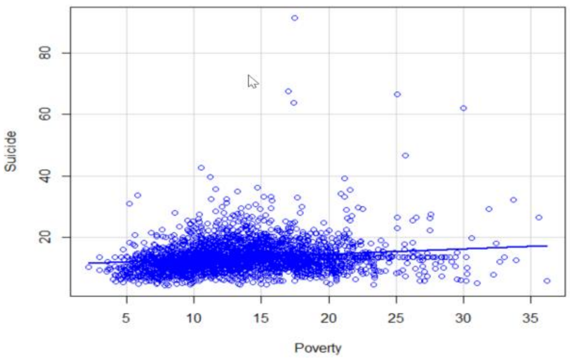


Figure 13b: Suicide~Poverty

Finally, as demonstrated in Figure 12, negatively correlated factors associated with Suicide are as follows: Primary Care Physician Rates, Disabled Medicare, and Elderly Medicare. These all had a negative correlative factor upon suicide. These correlations could be interpreted as the following: having some form of basic healthcare can be a preventative against a person committing suicide. An example of this can be seen by the negative correlation value that Major Depression has upon suicide, -0.20. This admittedly did not make much intuitive sense at the onset of the study, until a possible interpretation of the following was considered; a person reporting Major Depression has a high likelihood of having been diagnosed by a Doctor. Given that we have already seen the same level of correlative values from other health care factors (Primary Care Physician Rates, Disabled Medicare, Elderly Medicare) the team felt that the data was confirming that some form of healthcare, is a preventative against committing suicide.

Based on an early analysis in our data story, we found the top correlations to low life expectancy were eating few fruits and vegetables, high blood pressure, no exercise, obesity, and smoking. **Figure 14** shows the effects of not exercising and smoking on ALE. Notice the positive correlation between higher ALE and greater population percentage that exercise. The slope of the line has an adjusted R-square of 0.404. For aesthetic purposes, I reversed the X-axis to show the positive correlation moving from bottom left to top right in Figure 14. Additionally, the percentage of smokers in the population is colored from green (non-smoker) to blue (smoker). This correlation was scaled to the range of the smoking culture (min to max), thus enhancing the color range. This correlation with ALE is not as strong but still has an adjusted R-Square of 0.240. The top right of Figure 14 is the greenest which also represents the highest life expectancy. P-values for both smoking and not working out have very low p-values, all much less than the standard alpha of 0.05. From these results, the data suggests a longer life could be lived by working out and not smoking.

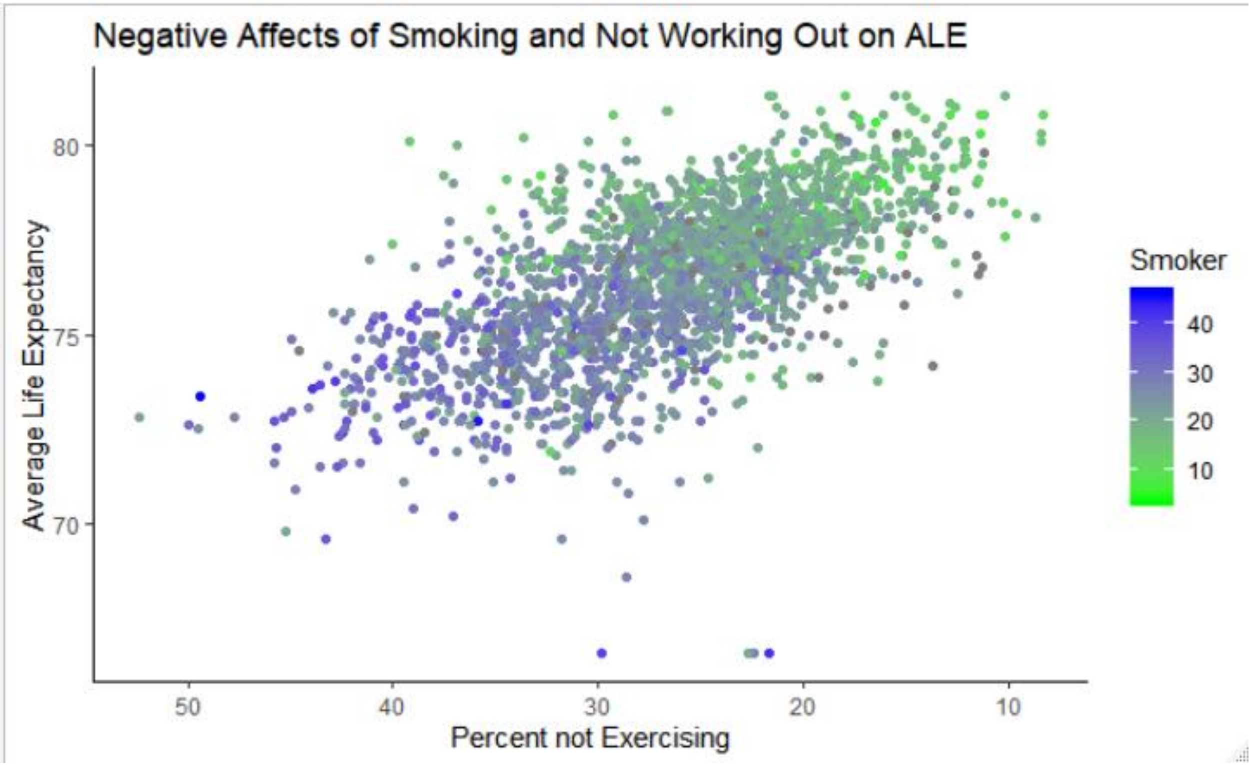


Figure 14: *The effects of not working out and smoking cause the population ALE to decrease*.

We visualized the smoking data across the US, coloring counties based on the percentage of smokers within them.

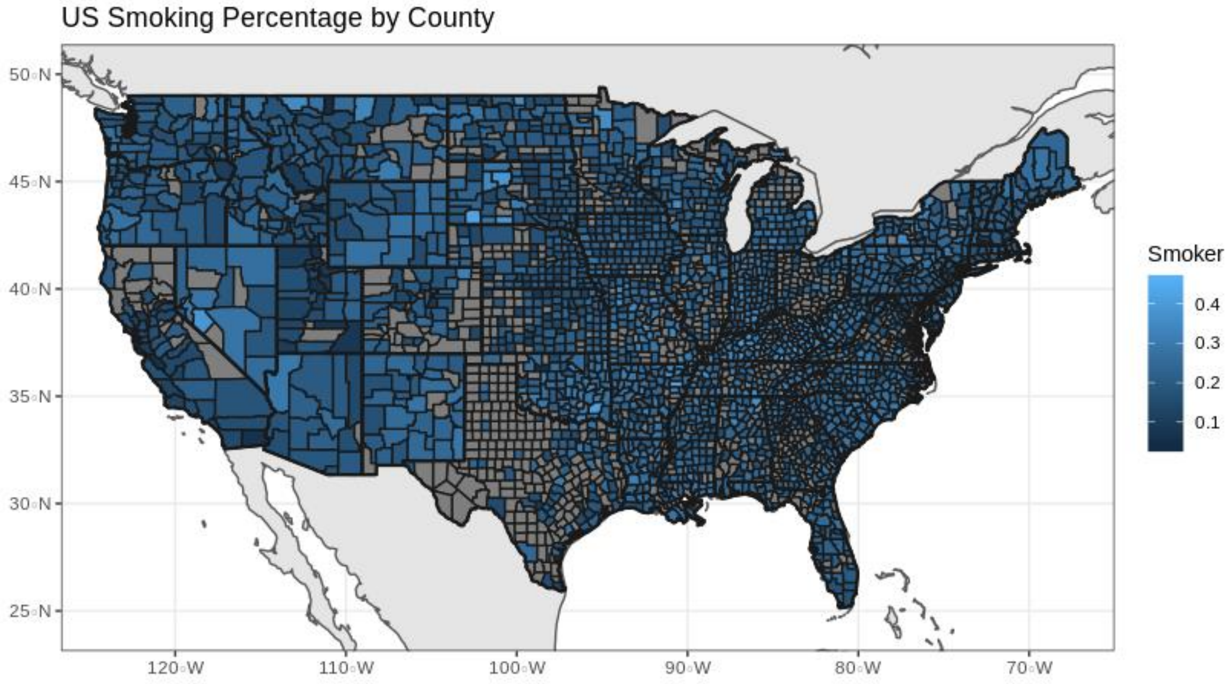


Figure 15: Shows the percentage of the U.S. population that smokes per county

Visually, this smoking map shows a remarkable similarity to Figure 11b (poverty map). Even though there are several counties filled gray for NA, we can get a good sense of where smoking is highest across the country.

Overall observations noted as a result of multiple correlations revealed the following:

- **Vulnerable populations and Death Rate:** positive correlations do exist between vulnerable populations and the Death Rate. Per the correlation matrix in **Figure 8,** the group was able to determine that there are several **positive correlations** that exist between the death rate and the following vulnerable populations:
    - No High School Diploma
    - Unemployed
    - Major Depression
    - Recent Drug Use
    - Elderly and on Medicare
    - Disabled and on Medicare
    - No Health Insurance (not surprisingly)

- **Risk Factors for Premature Death:** Per the initial correlation analysis (shown in **Figure 10)**, the following at-risk communities were identified as populations having a **negative correlative value** to average life expectancy (ALE).
    - No Exercise
    - Eat Few Fruits and/or Vegetables
    - Smokers
    - Obese

These factors all have above a - 0.45 correlation value to ALE, and thus are statistically significant. It can be interpreted that these people who live with these lifestyle choices are indeed at a significant statistical risk for premature death within a degree of certainty, at the minimal correlation value of 48%. Given that these are lifestyle health choices, these negative correlative factors to ALE do indeed contribute to premature death.

USE OF MODELING TECHNIQUES

13. Linear modeling

In order to run a model for risk factors compared to ALE, we needed to turn absolute numbers into percentages. Otherwise, the model would really be showing us population density rather than which risks affect ALE.

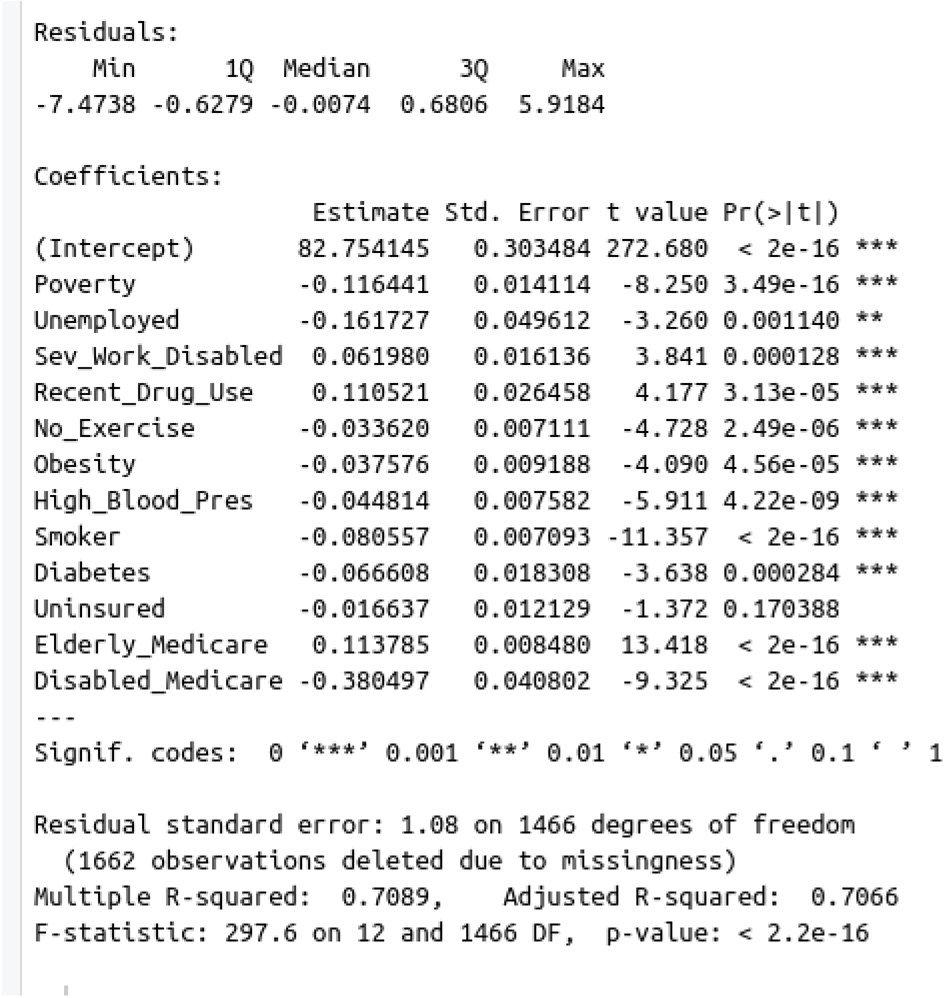
Several variables were considered and we ultimately chose several of interest:

- Poverty
- No_HS_Diploma
- Unemployed
- Sev_Work_Disabled
- Major_Depression
- Recent_Drug_Use
- No_Exercise
- Few_Fruit_Veg
- Obesity
- High_Blood_Pres
- Smoker
- Diabetes
- Uninsured
- Elderly_Medicare

- Disabled_Medicare
- Prim_Care_Phys_Rate
- Dentist_Rate

In the first model, we discovered that all of these were statistically significant except for No_HS_Diploma, Major_Depression, Few_Fruit_Veg, and Dentist_Rate.

A second model after removing these variables showed all still significant, except Uninsured (**Figure 16**). Yet keeping Uninsured in the model gave us a better R-squared than without it:



```
Residuals:
    Min      1Q  Median      3Q     Max 
-7.4738 -0.6279 -0.0074  0.6806  5.9184 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       82.754145   0.303484 272.680  < 2e-16 ***
Poverty           -0.116441   0.014114  -8.250 3.49e-16 ***
Unemployed        -0.161727   0.049612  -3.260 0.001140 ** 
Sev_Work_Disabled  0.061980   0.016136   3.841 0.000128 ***
Recent_Drug_Use    0.110521   0.026458   4.177 3.13e-05 ***
No_Exercise       -0.033620   0.007111  -4.728 2.49e-06 ***
Obesity           -0.037576   0.009188  -4.090 4.56e-05 ***
High_Blood_Pres   -0.044814   0.007582  -5.911 4.22e-09 ***
Smoker            -0.080557   0.007093 -11.357  < 2e-16 ***
Diabetes          -0.066608   0.018308  -3.638 0.000284 ***
Uninsured         -0.016637   0.012129  -1.372 0.170388    
Elderly_Medicare   0.113785   0.008480  13.418  < 2e-16 ***
Disabled_Medicare -0.380497   0.040802  -9.325  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.08 on 1466 degrees of freedom
  (1662 observations deleted due to missingness)
Multiple R-squared:  0.7089,    Adjusted R-squared:  0.7066 
F-statistic: 297.6 on 12 and 1466 DF,  p-value: < 2.2e-16
```

Figure 16*: Linear modelling coefficients are mostly below 0.05 p-value and has an adjusted R-squared of 0.7066.*

With an Adjusted R-squared of **0.7066**, we can explain quite a bit about average life expectancy using this model.

One surprising thing this shows is that ‘Recent_Drug_Use’ has a positive coefficient. Based on our expectations of health and risk factors, we should expect drug use to have a negative impact on life expectancy. Looking into the details on how this was collected, we see that “Estimates are based on state-level prevalence information and adjusted to reflect local demographic characteristics.” In other words, this represents state aggregates distributed to counties based on population. *That means it is not actually a good input for county-level ALE, despite the high statistical significance. It should be dropped from the model.*

The independent variable with the highest coefficient is ‘Disabled_Medicare’ with a -0.38. Unfortunately, this makes sense. This number means that a county has a higher percentage of the population who are considered disabled and are on Medicare. This is a result of earlier health problems, so is a reasonable indicator for poor health.

14. Support vector

We chose to use the same variables from the linear model for a SVM. Using KSVM from ‘kernlab’ we performed cross-fold validation on a training set using 70% of our data. We saw low training error, so continued to check predictions against the testing data.

Testing the predicted numbers against the actual testing ALE, we see the following:

- Actual average: 76.68201 (years)
- Actual deviation: 1.924118 (years)Pngl/gas b/m to life lin time
  ce_2018q4edicted average: 76.6573 (years)
- Predicted deviation: 1.906623 (years) actual/predicted difference: 0.7858914 (years)
- Percent accuracy: 98.96903 % / RMSE is ~ 1.056212

In other words, this is a very good model that predicts life expectancy variation from the inputs we selected. Comparison box plots can be seen in **Figure 17**.

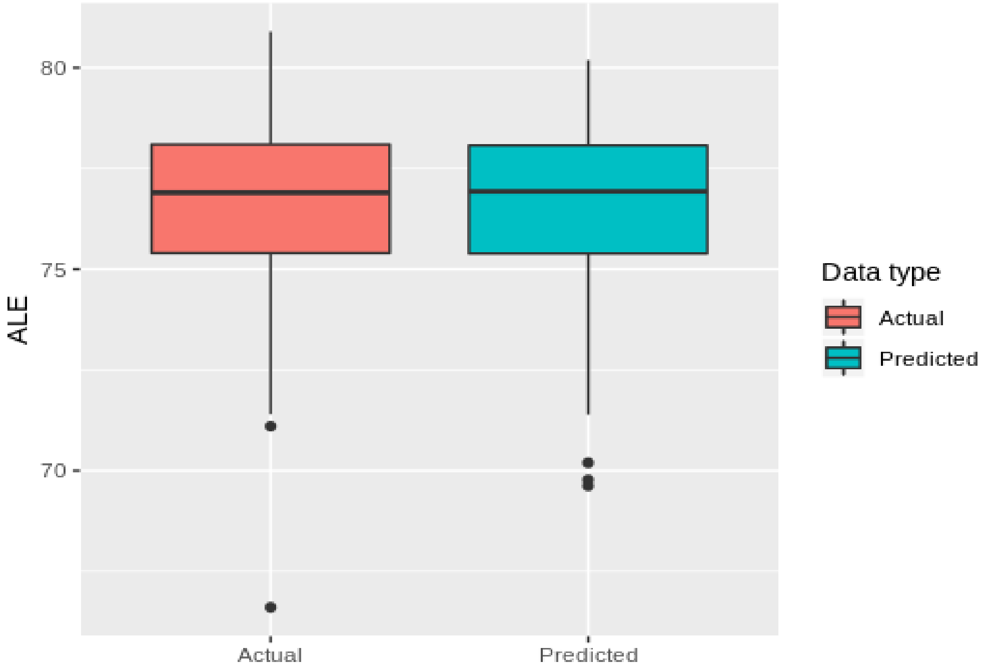


Figure 17:

We also tested a neural net analysis, but this did not give us good predictions at all.

OVERALL INTERPRETATION OF RESULTS/ ACTIONABLE INSIGHTS

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.

REFERENCES

Introduction to Data Science (2016), by Jeffrey S. Saltz & Jeffrey M. Stanton.

Appendix Section

APPENDIX 1 - RSTUDIO CODE

Script

```
######################## Initial Import Data (Below)  #########################
#########################################################################

# load required packages using a bulk check/install function from professor

packages <- c("cowplot", "googleway", "ggplot2", "ggrepel",  "ggspatial", "lwgeom", "sf", 
"rnaturalearth",

              "rnaturalearthdata","maps","dplyr", 
"sqldf","readr","reshape","neuralnet","dplyr","e1071","kernlab",
"googleway","rgeos","rgdal","maptools","scales","shiny","Rcmdr","tidyverse","RcmdrMisc
")

## function to check for libraries and install if necessary.

package.check <- lapply(packages, FUN = function(x) {

  if (!require(x, character.only = TRUE)) {

    install.packages(x, dependencies = TRUE)

    library(x, character.only = TRUE)

  }

})


#This script imports data from the chsi dataset and combines several sheets to

#make it easier to work with. You need to run it from within the folder that

#contains all the individual csv files


# use sqldf for joins

library(sqldf)

library(readr)


#!!change to your own directory where the CSV files are located!!

setwd("/home/luke/Documents/Grad_School/IST687/Project/chsi_dataset")


#this is the one that explains what columns mean

key = read_csv("DATAELEMENTDESCRIPTION.csv")


#Healthypeople is a reference sheet; we shouldn't join it. Instead, we'll use it
```

```
#later for comparison to standards

healthypeople = read_csv("HEALTHYPEOPLE2010.csv")


#These 5 will be combined

demographics = read_csv("DEMOGRAPHICS.csv")

vulnpopsandenvhealth = read_csv("VUNERABLEPOPSANDENVHEALTH.csv")

riskfactors = read_csv("RISKFACTORSANDACCESSTOCARE.csv")

leadingcausesofdeath = read_csv("LEADINGCAUSESOFDEATH.csv")

summarymeasures = read_csv("SUMMARYMEASURESOFHEALTH.csv")


combined <- sqldf('

select

  d.county_fips_code  ,d.state_fips_code  ,d.chsi_county_name  ,d.chsi_state_name
,d.chsi_state_abbr  ,d.strata_id_number  ,d.strata_determining_factors
,d.number_counties  ,d.population_size  ,d.population_density  ,d.poverty
,d.Age_19_Under  ,d.Age_19_64 ,d.Age_65_84, d.Age_85_and_Over, d.White, d.Black,
d.Native_American, d.Asian, d.Hispanic, v.No_HS_Diploma, v.Unemployed,
v.Sev_Work_Disabled, v.Major_Depression, v.Recent_Drug_Use, v.Ecol_Rpt,
v.Ecol_Rpt_Ind, v.Ecol_Exp, v.Salm_Rpt, v.Salm_Rpt_Ind , v.Salm_Exp, v.Shig_Rpt,
v.Shig_Rpt_Ind, v.Shig_Exp, v.Toxic_Chem, v.Carbon_Monoxide_Ind,
v.Nitrogen_Dioxide_Ind, v.Sulfur_Dioxide_Ind, v.Ozone_Ind, v.Particulate_Matter_Ind,
v.Lead_Ind, v.EH_Time_Span, r.No_Exercise, r.Few_Fruit_Veg, r.Obesity,
r.High_Blood_Pres, r.Smoker, r.Diabetes, r.Uninsured, r.Elderly_Medicare,
r.Disabled_Medicare, r.Prim_Care_Phys_Rate, r.Dentist_Rate,
r.Community_Health_Center_Ind, r.HPSA_Ind, l.A_Wh_Comp, l.A_Bl_Comp,
l.A_Ot_Comp, l.A_Hi_Comp, l.A_Wh_BirthDef, l.A_Bl_BirthDef, l.A_Ot_BirthDef,
l.A_Hi_BirthDef, l.B_Wh_Injury, l.B_Bl_Injury, l.B_Ot_Injury, l.B_Hi_Injury,
l.B_Wh_Cancer, l.B_Bl_Cancer, l.B_Ot_Cancer, l.B_Hi_Cancer, l.B_Wh_Homicide,
l.B_Bl_Homicide, l.B_Ot_Homicide, l.B_Hi_Homicide, l.C_Wh_Injury, l.C_Bl_Injury,
l.C_Ot_Injury, l.C_Hi_Injury, l.C_Wh_Homicide, l.C_Bl_Homicide, l.C_Ot_homicide,
l.C_Hi_Homicide, l.C_Wh_Suicide, l.C_Bl_Suicide, l.C_Ot_Suicide, l.C_Hi_Suicide,
l.C_Wh_Cancer, l.C_Bl_Cancer, l.C_Ot_Cancer, l.C_Hi_Cancer, l.D_Wh_Injury,
l.D_Bl_Injury, l.D_Ot_Injury, l.D_Hi_Injury, l.D_Wh_Cancer, l.D_Bl_Cancer,
l.D_Ot_Cancer, l.D_Hi_Cancer, l.D_Wh_HeartDis, l.D_Bl_HeartDis, l.D_Ot_HeartDis,
l.D_Hi_HeartDis, l.D_Wh_Suicide, l.D_Bl_Suicide, l.D_Ot_Suicide, l.D_Hi_Suicide,
l.D_Wh_HIV, l.D_Bl_HIV, l.D_Ot_HIV, l.D_Hi_HIV, l.D_Wh_Homicide, l.D_Bl_Homicide,
l.D_Ot_Homicide, l.D_Hi_Homicide, l.E_Wh_Cancer, l.E_Bl_Cancer, l.E_Ot_Cancer,
l.E_Hi_Cancer, l.E_Wh_HeartDis, l.E_Bl_HeartDis, l.E_Ot_HeartDis, l.E_Hi_HeartDis,
l.F_Wh_HeartDis, l.F_Bl_HeartDis, l.F_Ot_HeartDis, l.F_Hi_HeartDis, l.F_Wh_Cancer,
l.F_Bl_Cancer, l.F_Ot_Cancer, l.F_Hi_Cancer, l.LCD_Time_Span, s.ALE, s.US_ALE,
s.All_Death  ,s.US_All_Death, s.Health_Status, s.US_Health_Status, s.Unhealthy_Days

,s.US_Unhealthy_Days

from demographics d

left join vulnpopsandenvhealth v on v.county_fips_code=d.county_fips_code and
v.state_fips_code=d.state_fips_code

left join riskfactors r on r.county_fips_code=d.county_fips_code and
r.state_fips_code=d.state_fips_code

left join leadingcausesofdeath l on l.county_fips_code=d.county_fips_code and
l.state_fips_code=d.state_fips_code

left join summarymeasures s on s.county_fips_code=d.county_fips_code and
s.state_fips_code=d.state_fips_code

' )


# We know that negative numbers are really NAs; let's replace them

combined[,-1:-7] <-  data.frame(lapply(combined[,-1:-7], function(x){

   as.numeric(gsub("-1*|-2*|-9*",NA,x))

}))


str(combined)

write_csv(combined,'combined.csv')


#replace NAs with mean of column auto
```

```
na_mean_swap <- function(x) {

  replace(x, is.na(x),mean(as.numeric(x),na.rm=TRUE))

}


mean_clean <- cbind(combined[,1:7],replace(combined[,-1:-7],TRUE,
lapply(combined[,-1:-7], na_mean_swap)))

str(mean_clean)

################### Initial Import Data ^ #############################

################################################################


############### Plot Average Life Expectancy by State ####################

# Sam Rogers

#assistance from https://www.datanovia.com/en/blog/ggplot-themes-gallery/

Life_Exp_States <- ggplot(data = mean_clean, aes(x=CHSI_State_Abbr, y = ALE, color
= ALE, group=CHSI_State_Abbr)) +

  geom_point() +

  scale_color_continuous(name = "Average Age") +

  theme_classic() +

  xlab(label = "States") +

  ylab(label = "Average Life Expectancy") +

  scale_y_reverse() +

  ggtitle(label = "State Average Life Expectancy") +

  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

Life_Exp_States

################### plot ALE Under 1st Quartile #######################

#################################################################
```

```
# Sam Rogers

Life_Exp_75 <- ggplot(data = mean_clean[mean_clean$ALE<75,],
aes(x=CHSI_State_Abbr, y = ALE, color = ALE, group=CHSI_State_Abbr)) +

  geom_point() +

  scale_color_continuous(name = "Average Age") +

  theme_classic() +

  xlab(label = "States") +

  ylab(label = "Average Life Expectancy") +

  scale_y_reverse() +

  ggtitle(label = "State ALE Below Nation's 1st Quartile") +

  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))

Life_Exp_75


################ Plot of Smokers relating to ALE and Exercise ################
#####################################################################

# Sam Rogers

# Shows a plot of no exercise to ALE. Less smoking in green, heavy smoking in blue.

#determine the range of smokers

summary(combined$Smoker)

#plot

#assistance from https://www.datanovia.com/en/blog/ggplot-themes-gallery/

gtt <- ggplot(data = combined, aes(x=No_Exercise, y = ALE,   color = Smoker)) +

  theme_classic() +

  geom_point() +

  xlab(label = "Percent not Exercising") +

  scale_x_reverse() + #x_axis reversed to show positive correlation
  ylab(label = "Average Life Expectancy") +
  scale_colour_gradient(limits=c(3.6,46.2), low="green", high="red") + #Smoker Range
  ggtitle(label = "Negative Affects of Smoking and Not Working Out on ALE")
gtt

#Create formula to later take statistics on the plot
#These are all class or personal knowledge
ggtt <- lm(formula= ALE ~ No_Exercise, data = combined)    #ALE(dep)
ggts <- lm(formula= ALE ~ Smoker, data = combined)         #ALE(dep)
# shows adj R square of ALE to smoking and No exercise plots
summary(ggtt)
summary(ggts)

############# Barchart Summary of Variables Most Affecting Death #############
#####################################################################
#Sam Rogers
#this is from the section which takes population and percent of population into
#consideration. It's a true average of the population as values are originally presented
#as percentage of county.
#This calculates the count then takes percent of the US
DF_Bar <- vuln[,c('Poverty','Unemployed','Sev_Work_Disabled','Recent_Drug_Use','
No_Exercise','Obesity','High_Blood_Pres','Smoker','Diabetes','Uninsured','Elderly_Medi
care', 'Disabled_Medicare')]
PLot_Bar <- replace(DF_Bar, TRUE, lapply(DF_Bar, na_mean_swap))
#calculate mean of each column
PLot_Bar <- colMeans(PLot_Bar)
#assistance from Quick-R by DataCamp
```

```
#used melt from library(reshape)

#this makes each row a unique id-variable combination

PLot_Bar <- melt(PLot_Bar)

PLot_Bar$Variables <- rownames(PLot_Bar)

#used Cookbook for R, Colors(ggplot2) for assistance

#toned the colors down to be a bit darker and match the colors of our other graphs

BarPlotFin<- ggplot(data=PLot_Bar) + geom_bar(aes(x =
Variables,y=value,fill=Variables),stat="identity") + coord_flip() +scale_fill_hue(l=45)


###############################################################

######################Suicide to Toxic Chemicals###########

#Randall

scatterplot(Suicide~Toxic_Chem, regLine=TRUE, smooth=FALSE,

        boxplots=FALSE, xlab="Toxic Chemicals", ylab="Suicide",

        data=sContribs.df)

cor(sContribs.df[,c("Suicide","Toxic_Chem")], use="complete")


################Suicide to Poverty#######################

scatterplot(Suicide~Poverty,

        regLine=TRUE, smooth=FALSE,

        boxplots=FALSE,

        xlab="Poverty",

        ylab="Suicide", data=sContribs.df)
```

```
####Correlations Matrix runs, and Visualizations

#First run removed environmental factors that had no correlative value


sContribs.corr3 = cor(ALEdimin.df[,c( "No_HS_Diploma",
"Unemployed","Sev_Work_Disabled", "Major_Depression",
"Recent_Drug_Use","No_Exercise", "Few_Fruit_Veg", "Obesity", "Smoker", "Diabetes",
"Uninsured", "Elderly_Medicare", "Disabled_Medicare",
"High_Blood_Pres","Toxic_Chem", "Carbon_Monoxide_Ind", "Ozone_Ind",
"Particulate_Matter_Ind", "Lead_Ind", "Disabled_Medicare", "Prim_Care_Phys_Rate",
"Dentist_Rate", "Community_Health_Center_Ind","ALE","Suicide", "Premature",
"Under_18", "Over_40", "Unmarried", "Late_Care", "Infant_Mortality", "IM_Hisp",
"IM_Neonatal", "IM_Postneonatal", "Brst_Cancer", "Col_Cancer", "CHD", "Homicide",
"Lung_Cancer", "MVA", "Stroke", "Injury" ,  "Total_Deaths", "All_Death",
"Unhealthy_Days")], use="complete")

#### Second run; removed categorical descriptions: US ALE, US UNHEALTHY DAYS

sContribs.corr3 = cor(ALEdimin.df[,c( "No_HS_Diploma",
"Unemployed","Sev_Work_Disabled", "Major_Depression",
"Recent_Drug_Use","No_Exercise", "Few_Fruit_Veg", "Obesity", "Smoker", "Diabetes",
"Uninsured", "Elderly_Medicare", "Disabled_Medicare",
"High_Blood_Pres","Toxic_Chem", "Carbon_Monoxide_Ind", "Ozone_Ind",
"Particulate_Matter_Ind", "Lead_Ind", "Disabled_Medicare", "Prim_Care_Phys_Rate",
"Dentist_Rate", "Community_Health_Center_Ind","ALE","Suicide", "Premature",
"Under_18", "Over_40", "Unmarried", "Late_Care", "Infant_Mortality", "IM_Hisp",
"IM_Neonatal", "IM_Postneonatal", "Brst_Cancer", "Col_Cancer", "CHD", "Homicide",
"Lung_Cancer", "MVA", "Stroke", "Injury" ,  "Total_Deaths", "All_Death",
"Unhealthy_Days")], use="complete")


#####look at a single correlation between ALE and Suicide

sContribs.corr3.1 = cor(ALEdimin.df[c("ALE", "Suicide")], use='complete')

View(sContribs.corr3.1)


##########Correlations matrix visualization ALE correalting risk factors
```

```
sContribs.corrPlt3 = corrplot::corrplot(sContribs.corr3, method = ('pie'), type = ('upper'),
add = FALSE, col = NULL, bg = "white", title = "ALE correlating risk factors",
is.corr=TRUE, diag = TRUE, outline = FALSE, addgrid.col = TRUE,addCoef.col =
NULL,addCoefasPercent = TRUE, sig.level = 0.05)

plot(sContribs.corrPlt3)

##########Correlations matrix visualization ALE correalting risk factors next iteration

sContribs.corrPlt3.1 = corrplot::corrplot(sContribs.corr3.1, method = ('pie'), type =
('upper'), add = FALSE, col = NULL, bg = "white", title = "ALE correlating risk factors",
is.corr=TRUE, diag = TRUE, outline = FALSE, addgrid.col = TRUE,addCoef.col =
NULL,addCoefasPercent = TRUE, sig.level = 0.05)

######Visualization for interactive maps#####

##Required Libraries are as follows:

packages <- c("readr","sqldf","cowplot", "googleway", "ggplot2", "ggrepel",  "ggspatial",
"lwgeom", "sf", "rnaturalearth",

"rnaturalearthdata","maps","dplyr","raster","stringr","spData","leaflet","sp","tidyr","devtool
s","tmap","tmaptools","corrplot","rgeos","rgdal","maptools","scales","shiny","Rcmdr")


ggplot(data = world) +

  geom_sf() +

  geom_sf(data=noHSDIPLOMA1.map,aes(fill=No_HS_Diploma),color=gray(0.1)) +

  geom_sf(data=states,fill=NA,size=0.8,color=gray(0.1)) #+

  #coord_sf(xlim=xlimit,ylim=ylimit, expand=FALSE)

plot(noHSDIPLOMA1.map)


nohsD.map1 = tm_shape(noHSDIPLOMA1.map) + tm_fill("No_HS_Diploma",
legend.show = TRUE, id = "county", palette = "Greens" ) +  tm_polygons("county",
"orange", style = "cont",  n = 2, alpha = 1 , stretch.palette = TRUE, popup.vars = NA,
convert2density = TRUE, midpoint = FALSE)

tmap_mode("view")
```

```
nohsD.map1

#####Bar Graphs for Demographics######

#Transform demographics for valid columns

# subset data to only include some columns

demographics <- subset(demographics, select =

                  c(CHSI_County_Name:CHSI_State_Abbr, Population_Size,

                   Population_Density, Poverty, Age_19_Under, Age_19_64,

                   Age_65_84, Age_85_and_Over, White, Black, Native_American, Asian, Hispanic))

# clean names of the subset

nms_demo_dat <- c("county.name","state.name","state.abbr","pop.size","pop.density",

           "poverty","age.19_under","age.19_64","age.65_84","age.85_over",

           "white","black","nat.amer","asian","hispanic")

# change col names

names(demographics)<-nms_demo_dat

#This data is a representation at the county level

#we have to do some work to find it at state level

mean(demographics$pop.size)

max(demographics$pop.size)

min(demographics$pop.size)

sd(demographics$pop.size)

#made with help from R Studio Community

library(tidyverse) #load library

# build vectors

county <- c(demographics$county.name)

state <- c(demographics$state.name)
```

```
pop <- c(demographics$pop.size)

# now we assemble into data frame from vectors

df <- data.frame(county, state, pop)

# assembled pipe

stateandmeanpop <- df %>% group_by(state) %>% summarize(mean_pop = mean(pop))

#Now we have a dataframe we can make different charts from

#with each states mean population

#This is the true mean of the population in our dataset

mean(stateandmeanpop$mean_pop)

#This is the range of the population in our dataset

max(stateandmeanpop$mean_pop)

min(stateandmeanpop$mean_pop)

sd(stateandmeanpop$mean_pop)

#B.	Show mean age, and ranges (a distribution would be a good visual for this as well)

#We can do the same with age.

demographics$age.19_underraw = (demographics$pop.size*demographics$age.19_under/100)

demographics$age.19_64raw = (demographics$pop.size*demographics$age.19_64/100)

demographics$age.65_84raw = (demographics$pop.size*demographics$age.65_84/100)

demographics$age.85_overraw = (demographics$pop.size*demographics$age.85_over/100)

xviiii_U <- mean(demographics$age.19_under)

max(demographics$age.19_under)
```

```
min(demographics$age.19_under)


xvii_lxiv <- mean(demographics$age.19_64)

max(demographics$age.19_64)

min(demographics$age.19_64)


lxv_lxxxiv <- mean(demographics$age.65_84)

max(demographics$age.65_84)

min(demographics$age.65_84)


lxxxv_O <- mean(demographics$age.85_over)

max(demographics$age.85_over)

min(demographics$age.85_over)

meanofage <- c(xviiii_U, xvii_lxiv, lxv_lxxxiv, lxxxv_O)

barplot (meanofage,

        main = "National Age %",

        xlab = "Mean % Age",

        ylab = "Population %",

        names.arg = c("Under19", "19-64", "65-84", "Over85"),

        col = "blue",

        horiz = FALSE)

#C.     Show ethnicity nationally (e.g. 87% white, etc.) - bar chart for visual

#How to create raw population numbers for age and ethnicity

#First we have to add new columns in the data.frame with the raw data

demographics$whiteraw = (demographics$pop.size*demographics$white/100)
```

```
demographics$blackraw = (demographics$pop.size*demographics$black/100)
demographics$hispanicraw = (demographics$pop.size*demographics$hispanic/100)
demographics$asianraw = (demographics$pop.size*demographics$asian/100)
demographics$nat.amerraw = (demographics$pop.size*demographics$nat.amer/100)

#Create mean vectors for each ethnicity
w <- mean(demographics$whiteraw)
b <- mean(demographics$blackraw)
h <- mean(demographics$hispanicraw)
a <- mean(demographics$asianraw)
n.a <- mean(demographics$nat.amerraw)
#Create dataframe
meanofrace <- c(w,b,h,a,n.a)
#Create bar chart
barplot (meanofrace,
        main = "National Mean Ethnicity",
        xlab = "Mean Ethnicity",
        ylab = "Population",
        names.arg = c("White", "Black", "Hispanic", "Asian", "NativeAmerican"),
        col = "darkred",
        horiz = FALSE)

##################
# linear model & svm
##################
```

```
# Linear model
# After you've imported CHSI
library(neuralnet)
library(dplyr)
library(e1071)
library(kernlab)

ggplot(data=mean_clean,aes(x='Life expectancy',y=ALE)) + geom_boxplot()

vuln <- sqldf('
        select
                county_fips_code
                ,state_fips_code
                ,chsi_county_name
                ,chsi_state_name
                ,population_size
                ,poverty
                ,No_HS_Diploma
                ,Unemployed
                ,Sev_Work_Disabled
                ,Major_Depression
                ,Recent_Drug_Use
                ,No_Exercise
                ,Few_Fruit_Veg
                ,Obesity
                ,High_Blood_Pres
                ,Smoker
                ,Diabetes
                ,Uninsured
                ,Elderly_Medicare
                ,Disabled_Medicare
                ,Prim_Care_Phys_Rate
                ,Dentist_Rate
                ,ALE
                ,All_Death
                ,US_All_Death
        from combined')


str(vuln)

#Convert raw count data into percentages so we can compare it across counties with different populations

vuln$No_HS_Diploma <- vuln$No_HS_Diploma/vuln$Population_Size*100

vuln$Unemployed <- vuln$Unemployed/vuln$Population_Size*100

vuln$Sev_Work_Disabled <-vuln$Sev_Work_Disabled/vuln$Population_Size*100

vuln$Major_Depression <- vuln$Major_Depression/vuln$Population_Size*100

vuln$Recent_Drug_Use <- vuln$Recent_Drug_Use/vuln$Population_Size*100

vuln$Uninsured <- vuln$Uninsured/vuln$Population_Size*100

vuln$Elderly_Medicare <- vuln$Elderly_Medicare/vuln$Population_Size*100

vuln$Disabled_Medicare <- vuln$Disabled_Medicare/vuln$Population_Size*100

vuln$DeathRate <- vuln$All_Death/vuln$Population_Size*100
```

```
cor(vuln[,c("ALE","Dentist_Rate","Diabetes","Disabled_Medicare",
        "Elderly_Medicare","Few_Fruit_Veg","High_Blood_Pres","Major_Depression",
        "No_Exercise","No_HS_Diploma","Obesity","Population_Size","Poverty",
        "Prim_Care_Phys_Rate","Recent_Drug_Use","Sev_Work_Disabled","Smoker",
        "Unemployed","Uninsured")], use="complete")
```

```
LinearModel.1 <- lm(ALE ~ Dentist_Rate + Diabetes + Disabled_Medicare +
                Elderly_Medicare + Few_Fruit_Veg + High_Blood_Pres +
Major_Depression +
                No_Exercise + No_HS_Diploma + Obesity + Poverty +
Prim_Care_Phys_Rate +
                Recent_Drug_Use + Sev_Work_Disabled + Smoker + Unemployed +
Uninsured,
                data=vuln)
summary(LinearModel.1)
```

```
v_lm <- lm(ALE ~ Poverty +No_HS_Diploma +Unemployed +Sev_Work_Disabled
+Major_Depression +Recent_Drug_Use +No_Exercise +Few_Fruit_Veg +Obesity
+High_Blood_Pres +Smoker +Diabetes +Uninsured +Elderly_Medicare
+Disabled_Medicare +Prim_Care_Phys_Rate +Dentist_Rate
,data=vuln)
summary(v_lm)
```

```
v_lm1 <- lm(ALE ~ Poverty + Unemployed + Sev_Work_Disabled + Recent_Drug_Use
+No_Exercise + Obesity + High_Blood_Pres + Smoker + Diabetes + Uninsured +
Elderly_Medicare + Disabled_Medicare
,data=vuln)
summary(v_lm1)
```

```
######################
# svm
######################
#add ID number so we can easily split into train/test
vuln <- vuln %>% mutate(id=row_number())
vuln <- na.omit(vuln)

#put 70 percent of the data into a training dataset
rtrain <- vuln %>% sample_frac(0.7)

#put the rest into a testing dataset
rtest <- anti_join(vuln,rtrain, by='id')

svmOutput <- ksvm(ALE ~ Poverty + Unemployed + Sev_Work_Disabled + Recent_Drug_Use +No_Exercise + Obesity + High_Blood_Pres + Smoker + Diabetes + Uninsured + Elderly_Medicare + Disabled_Medicare
            , data = rtrain,
            kernel = "rbfdot", # kernel function that projects the low dimensional problem into higher dimensional space
            kpar = "automatic",# kpar refer to parameters that can be used to control the radial function kernel(rbfdot)
            C = 10, # C refers to "Cost of Constrains"
            cross = 10, # use 10 fold cross validation in this model
            prob.model = TRUE # use probability model in this model
)
# check the model
```

```
svmOutput


# 2) Test the model with the testData data set

svmPred <- predict(svmOutput, # use the built model "svmOutput" to predict

                   rtest, # use testData to generate predictions

                   type = "votes" # request "votes" from the prediction process

)




# create a comparison dataframe that contains the exact "Ozone" value and the
predicted "Ozone" value

# use for RMSE calc


compTable <- data.frame(rtest[,c('ALE')], svmPred[,1])

colnames(compTable) <- c("actual","Pred")




compTable$diff <- abs(compTable$actual-compTable$Pred)

compTable$err <- (1 -compTable$diff/compTable$actual)*100

compTable

# compute the Root Mean Squared Error

ksvm_rms <- sqrt(mean((compTable$actual-compTable$Pred)^2)) #A smaller value
indicates better model performance.


cat('Actual average:          ',mean(compTable$actual),' years',

'\nActual deviation:   ',sd(compTable$actual),' years',
'\nPredicted average:   ',mean(compTable$Pred),' years',

'\nPredicted deviation: ',sd(compTable$Pred),' years',

'\nAverage difference:  ',mean(compTable$diff),' years',

'\nPercent accuracy:           ',mean(compTable$err),'%',

'\nRMS is                  ',ksvm_rms)




ggplot(data=compTable,aes(x=1:nrow(compTable),y=actual,color='Actual')) + geom_line() + geom_line(aes(y=Pred,color="Predicted"),alpha=0.5) + xlab('rownum') + scale_color_manual(values=c('darkblue','red')) +labs(color='Type')


ggplot(data=compTable) + geom_boxplot(aes(x='Actual',y=actual,fill='Actual')) + geom_boxplot(aes(x='Predicted',y=Pred,fill='Predicted')) + labs(fill='Data type',y='ALE',x='')


sd(rtest[,c('ALE')])

summary(rtest[,c('ALE')])



########################

# Add neuralnet

########################


#sampling example from https://medium.com/@HollyEmblem/training-and-test-dataset-creation-with-dplyr-41d9aa7eab31

#Assign row number to an id column within dataframe; this allows us to easily

#split the data
```

```
str(rtrain)
#create neural net model using training data
vuln_net <- neuralnet(ALE ~ Poverty + Unemployed + Sev_Work_Disabled + Recent_Drug_Use +No_Exercise + Obesity + High_Blood_Pres + Smoker + Diabetes + Uninsured + Elderly_Medicare + Disabled_Medicare
,data=rtrain, hidden=2,lifesign='minimal',linear.output=FALSE,threshold = 0.5)


#here are coeeficients
vuln_net$result.matrix


#here's what the net looks like
plot(vuln_net,rep='best')


#predict results of the testing dataset using the model
v.results <- predict(vuln_net,rtest[,c('Poverty','Unemployed','Sev_Work_Disabled','Recent_Drug_Use','No_Exercise','Obesity','High_Blood_Pres','Smoker','Diabetes','Uninsured','Elderly_Medicare','Disabled_Medicare' )])
#create comparison dataframe with the real data and the predicted data
v.compare <- data.frame(actual = rtest[,c('ALE')],predicted=v.results)
typeof(v.compare)
# add diff column
v.compare$diff <- abs(v.compare$actual - v.compare$predicted )
# add accuracy column. Since we're comparing percentages, I *think* we just need
# to take 1 minus the diff. Is that right?
v.compare$accuracy <- 100 - abs(v.compare$actual - v.compare$predicted )
mean(v.compare$accuracy)
```

```
sd(v.compare$accuracy)

v.compare


#bar plot for Sam


dgraph <-
vuln[,c('Poverty','Unemployed','Sev_Work_Disabled','Recent_Drug_Use','No_Exercise','
Obesity','High_Blood_Pres','Smoker','Diabetes','Uninsured','Elderly_Medicare','Disabled
_Medicare')]


dgraph <- colMeans(dgraph)


dgraph <- melt(dgraph)

dgraph$variable <- rownames(dgraph)


ggplot(data=dgraph) + geom_bar(aes(x = variable,y=value,fill=variable),stat="identity")
+ coord_flip()

#Code found on Github by Todd and provided by Luke Miller

#Code for Poverty maps

#a copy of the function

lower.df = function(v)

{

  if(is.character(v)) return(tolower(v))

  else return(v)

}

# subset data to only include some columns

demographics <- subset(demographics, select =
                      c(CHSI_County_Name:CHSI_State_Abbr, Population_Size,
                        Population_Density, Poverty, Age_19_Under, Age_19_64,
                        Age_65_84, Age_85_and_Over, White, Black, Native_American, Asian, Hispanic))
# clean names of the subset
nms_demo_dat <- c("county.name","state.name","state.abbr","pop.size","pop.density",
              "poverty","age.19_under","age.19_64","age.65_84","age.85_over",
              "white","black","nat.amer","asian","hispanic")
# change col names
names(demographics)<-nms_demo_dat
head(demographics)
demographics$state.name <- lower.df(demographics$state.name)
demographics$county.name <- lower.df(demographics$county.name)
demographics$state.abbr <- lower.df(demographics$state.abbr)
# get state map
us_state <- map_data("state")
# change us_state_map names
names(us_state)<- c("long","lat","group","order","state.name","subregion")
# merge us_state_map data frame with demographics_dat by state.name
# only include matching records
Po_data<-merge(us_state, demographics, by ='state.name')
# preserve order
Po_data<-Po_data[order(Po_data$order),]
# remove subregion column
Po_data$subregion<-NULL
# split %'s into 6 cuts
```

```
Po_data$poverty <- cut_interval(Po_data$poverty, 6)
# state data
state_df <- map_data("state")
# create dataframe with county information from maps
# Longitude and Latitude information here
county_df<-map_data("county")
# change names of county_df
names(county_df)<- c("long","lat","group","order","state.name","county.name")
# check out state.abb and state.name Datasets
# will add state.abbr to county_df based on match
head(state.abb)
head(state.name)
# add a column with state abbreviations based on matching between
# county_df$state_name and lowercase state.name dataset
county_df$state.abbr<- state.abb[match(x = county_df$state.name, tolower(state.name))]
# remove state_name column since have abbreviations
county_df$state.name<-NULL

# make all names lowercase to match
county_df <- data.frame(lapply(county_df, lower.df))

#------------------------------------------------------------
# will use this to zoom in on % poverty at County level
#------------------------------------------------------------
# merge county_df and demographics_dat by county.name and state.abbr
```

```
Pop_map <- merge(x = county_df, y = demographics, by=c("county.name","state.abbr"))

# retain order

Pop_map <- Pop_map[order(Pop_map$order), ]

# add breaks for ranges in new map risk_map

Pop_map$poverty <- cut_interval(Pop_map$poverty ,6)

#------------------------------------------------------------------------------------------

# plot Poverty across US states - not at county level

#------------------------------------------------------------------------------------------

# create dataframe to add state map abbreviations on map

# state.center - x=long, y=lat. state.abb is a list containing state abbreviations

state.info <- data.frame(state.center, state.abb)

# lower names

state.info$state.abb <- tolower(state.info$state.abb)

# add group info

state.info$group <- Po_data$group[match(x = state.info$state.abb,
Po_data$state.abbr)]

# remove ak and hi (no group)

state.info <- state.info[!is.na(state.info$group),]

# map of poverty at state level - org palette

# doesnt include state names

Pop_map_all_org <- ggplot(Po_data, aes(long, lat, group = group)) +

  geom_polygon(aes(fill = poverty)) +

  geom_polygon(data = state_df, colour = "black", fill = NA, size =0.2) +

  geom_polygon(data = county_df, colour = "snow", fill = NA, size =0.1) +

  geom_text(data = state.info, aes(x=x, y=y, label = state.abb, group = group), colour
='black') +
  theme_classic() +
  theme(legend.position="right") +
  xlab(label = "") +
  ylab(label = "") +
  scale_x_continuous(expand = c(0,0)) + # expand size of map along x axis
  scale_y_continuous(expand = c(0,0)) + # expand size of map along y axis
  theme(axis.ticks = element_blank(),
        axis.text.x = element_blank(),
        axis.text.y = element_blank()) +
  ggtitle(label = "US Map showing Poverty Intervals at the State Level") +
  scale_fill_brewer(type='div', palette = 'RdBu', name = "% Poverty")
Pop_map_all_org

#Poverty Calculated at the county level

poverty_Cmap_purd <- ggplot(Pop_map, aes(long, lat, group = group)) +
  theme(panel.background = element_rect(fill = "snow")) +
  geom_polygon(aes(fill = poverty), colour = alpha("snow", 1/2), size = 0.2) +
  geom_polygon(data = state_df, colour = "black", fill = NA,size =0.2) +
  geom_polygon(data = county_df, colour = "grey", fill = NA, size =0.1) +
  theme_classic() +
  theme(legend.position="bottom") +
  xlab(label = "") +
  ylab(label = "") +
  scale_x_continuous(expand = c(0,0)) + # expand size of map along x axis
  scale_y_continuous(expand = c(0,0)) + # expand size of map along y axis
  theme(axis.ticks = element_blank(),
        axis.text.x = element_blank(),
        axis.text.y = element_blank()) +
  ggtitle(label = "US Map showing poverty \nIncludes County information") +
  scale_fill_brewer(type='seq',palette = 'PuRd', name ="% Poverty")
poverty_Cmap_purd

########################
# SF map plotting groundwork
########################
# Luke Miller
# 2 September 2019
# following map drawing tutorial
# from https://www.r-spatial.org/r/2018/10/25/ggplot2-sf-2.html
# Modifying to visualize project data

#load required packages using a bulk check/install function from professor
packages <- c("cowplot", "googleway", "ggplot2", "ggrepel",  "ggspatial", "lwgeom", "sf",
"rnaturalearth", "rnaturalearthdata","maps",'dplyr','reshape')

package.check <- lapply(packages, FUN = function(x) {
  if (!require(x, character.only = TRUE)) {
          install.packages(x, dependencies = TRUE)
          library(x, character.only = TRUE)
  }
})

theme_set(theme_bw())

#################
# Plot world map
#################

world <- ne_countries(scale='medium',returnclass='sf')
#class(world)

ggplot(data=world) +
  geom_sf() +
  coord_sf()

##################
# Plot state map
##################

#turn map state into a shapefile
states <- st_as_sf(map("state",plot=FALSE,fill=TRUE))
#add coordinates of 'centroid' so we can later plot names
states <- cbind(states, st_coordinates(st_centroid(states)))
#label state names as uppercase
states$name <- toupper(states$ID)
```

```
#add us so we can use the limits

us <- map_data("county")[,c('long','lat')]

xlimit <- c(max(us$long)+2,min(us$long)-2)

ylimit <- c(max(us$lat)+2,min(us$lat)-2)



ggplot(data=world) +

  geom_sf() +

  geom_sf(data=states,fill=NA) +

  geom_label(data=states,aes(X,Y, label=name), size=3) +

  coord_sf(xlim=xlimit,ylim=ylimit, expand=FALSE)


#############################

# Zoom in on chosen location

#############################


zoomAmount <- 5


#knoxville x=-83.99,y=35.82 Pick your favorite city

centerx <- -83.99

centery <- 35.82


ylimit <- c(centery-zoomAmount, centery+zoomAmount)

xlimit <- c(centerx-zoomAmount*2, centerx+zoomAmount*2)
```

```
ggplot() +
  geom_sf(data=states) +
  geom_label(data=states,aes(X,Y, label=name), size=3) +
  coord_sf(xlim=xlimit,ylim=ylimit, expand=FALSE)

#################
# Counties
#################

counties <- st_as_sf(map("county", plot=FALSE, fill=TRUE))

ggplot(data = world) +
  geom_sf() +
  geom_sf(data=states,size=0.8) +
  geom_sf(data=counties,fill=NA,color=gray(0.5)) +
  coord_sf(xlim=xlimit,ylim=ylimit, expand=FALSE)

############################
# Counties with smoking data
############################

#must have loaded chsi already
smokers <- combined[,c('CHSI_State_Name','CHSI_County_Name','Smoker','Poverty')]
smokers$Smoker <- as.numeric(smokers$Smoker)/100
```

```
smokers$state <- tolower(smokers$CHSI_State_Name)
smokers$county<- tolower(smokers$CHSI_County_Name)
smokers$ID <- paste(smokers$state,smokers$county,sep=',')

counties <- st_as_sf(map("county", plot=FALSE, fill=TRUE))

#create new object that combines the counties sf with the smoker dataset
s.map <- left_join(counties,smokers)

zoomAmount <- 3
#knoxville x=-83.99,y=35.82 Pick your favorite city
centerx <- -83.99
centery <- 35.82

ylimit <- c(centery-zoomAmount, centery+zoomAmount)
xlimit <- c(centerx-zoomAmount*2, centerx+zoomAmount*2)

#inherits xlimit and ylimit from above; change as desired
ggplot(data = world) +
  geom_sf() +
  geom_sf(data=s.map,aes(fill=Smoker),color=gray(0.1)) +
  geom_sf(data=states,fill=NA,size=0.8,color=gray(0.1)) +
  coord_sf(xlim=xlimit,ylim=ylimit, expand=FALSE) + labs(title='US Smoking Percentage by County')

##################################
```

```
# ALE

##################################


ALE <- combined[,c('CHSI_State_Name','CHSI_County_Name','ALE')]

ALE$state <- tolower(ALE$CHSI_State_Name)

ALE$county<- tolower(ALE$CHSI_County_Name)

ALE$ID <- paste(ALE$state,ALE$county,sep=',')


counties <- st_as_sf(map("county", plot=FALSE, fill=TRUE))


#create new object that combines the counties sf with the smoker dataset

s.map <- left_join(counties,ALE)


#zoomAmount <- 3

##knoxville x=-83.99,y=35.82 Pick your favorite city

#centerx <- -83.99

#centery <- 35.82

#

#ylimit <- c(centery-zoomAmount, centery+zoomAmount)

#xlimit <- c(centerx-zoomAmount*2, centerx+zoomAmount*2)


#inherits xlimit and ylimit from above; change as desired

ggplot(data = world) +

  geom_sf() +

  geom_sf(data=s.map,aes(fill=ALE),color=gray(0.1)) +
  geom_sf(data=states,fill=NA,size=0.8,color=gray(0.1)) +

  coord_sf(xlim=xlimit,ylim=ylimit, expand=FALSE) + labs(title='US life expectancy by county')
```

Appendix 2: Data Dictionary

Example of Demographic

Page_Name	***Column_Data***	***DATA_ TYPE***	***Is_%_ Data***	***Description***
Demographics	State_FIPS_Code	Text	N	Two-digit state identifier, developed by the National Bureau of Standards
Demographics	County_FIPS_Code	Text	N	Three-digit county identifier, developed by the National Bureau of Standards
Demographics	CHSI_County_Name	Text	N	Name of county
Demographics	CHSI_State_Name	Text	N	Name of State or District of Columbia
Demographics	CHSI_State_Abbr	Text	N	Two-character postal abbreviation for state name
Demographics	Strata_ID_Number	Integer	N	CHSI Peer County Stratum Number
Demographics	Strata_Determining_Factors	Text	N	Listing of strata factors
Demographics	Number_Counties	Integer	N	Number of peer counties
Demographics	Population_Size	Integer	N	County data, population size
Demographics	Population_Density	Integer	N	County data, population density (people per square mile)
Demographics	Poverty	Decimal	Y	County data, individuals living below poverty level

Demographics	Age_19_Under	Decimal	Y	County data, population under age 19

Example of Summary Measures of Health

Page_Name	***Column_Data***	***DATA_TYPE***	***Is_%_Data***	***Description***
SummaryMeasuresOfHealth	State_FIPS_Code	Text	N	Two-digit state identifier, developed by the National Bureau of Standards
SummaryMeasuresOfHealth	County_FIPS_Code	Text	N	Three-digit county identifier, developed by the National Bureau of Standards
SummaryMeasuresOfHealth	CHSI_County_Name	Text	N	Name of county
SummaryMeasuresOfHealth	CHSI_State_Name	Text	N	Name of State or District of Columbia
SummaryMeasuresOfHealth	CHSI_State_Abbr	Text	N	Two-character postal abbreviation for state name
SummaryMeasuresOfHealth	Strata_ID_Number	Integer	N	CHSI Peer County Stratum Number
SummaryMeasuresOfHealth	ALE	Decimal	N	County data, average life expectancy
SummaryMeasuresOfHealth	US_ALE	Decimal	N	Medium for all U.S. counties, average life expectancy
SummaryMeasuresOfHealth	All_Death	Decimal	N	County data, all causes of death
SummaryMeasuresOfHealth	US_All_Death	Decimal	N	Medium for all U.S. counties, all causes of death
SummaryMeasuresOfHealth	Health_Status	Decimal	Y	County data, self-rated health status

Example of Leading Causes of Death

Page_Name	***Column_Data***	***DATA_TYPE***	***Is_%_Data***	***Description***
LeadingCausesOfDeath	State_FIPS_Code	Text	N	Two-digit state identifier, developed by the National Bureau of Standards

LeadingCausesOfDeath	County_FIPS_Code	Text	N	Three-digit county identifier, developed by the National Bureau of Standards
LeadingCausesOfDeath	CHSI_County_Name	Text	N	Name of county
LeadingCausesOfDeath	CHSI_State_Name	Text	N	Name of State or District of Columbia
LeadingCausesOfDeath	CHSI_State_Abbr	Text	N	Two-character postal abbreviation for state name
LeadingCausesOfDeath	Strata_ID_Number	Integer	N	CHSI Peer County Stratum Number
LeadingCausesOfDeath	A_Wh_Comp	Integer	Y	County data, under age 1, complications of pregnancy/birth, White
LeadingCausesOfDeath	CI_Min_A_Wh_Comp	Integer	Y	Confidence interval lower limit, under age 1, complications of pregnancy/birth, White
LeadingCausesOfDeath	CI_Max_A_Wh_Comp	Integer	Y	Confidence interval upper limit, under age 1, complications of pregnancy/birth, White
LeadingCausesOfDeath	A_Bl_Comp	Integer	Y	County data, under age 1, complications of pregnancy/birth, Black
LeadingCausesOfDeath	CI_Min_A_Bl_Comp	Integer	Y	Confidence interval lower limit, under age 1, complications of pregnancy/birth, Black
LeadingCausesOfDeath	CI_Max_A_Bl_Comp	Integer	Y	Confidence interval upper limit, under age 1, complications of pregnancy/birth, Black

Example of Measures of Birth and Death

Page_Name	***Column_Data***	***DATA_TYPE***	***Is_%_Data***	***Description***
MeasuresOfBirthAndDeath	State_FIPS_Code	Text	N	Two-digit state identifier, developed by the National Bureau of Standards
MeasuresOfBirthAndDeath	County_FIPS_Code	Text	N	Three-digit county identifier, developed by the National Bureau of Standards
MeasuresOfBirthAndDeath	CHSI_County_Name	Text	N	Name of county
MeasuresOfBirthAndDeath	CHSI_State_Name	Text	N	Name of State or District of Columbia

MeasuresOfBirthAndDeath	CHSI_State_Abbr	Text	N	Two-character postal abbreviation for state name
MeasuresOfBirthAndDeath	Strata_ID_Number	Integer	N	CHSI Peer County Stratum Number
MeasuresOfBirthAndDeath	LBW	Decimal	Y	County data, birth measures, low birth wt. (<2500 g)
MeasuresOfBirthAndDeath	LBW_Ind	Decimal	Y	Favorable indicator, birth measures, low birth wt. (<2500 g)
MeasuresOfBirthAndDeath	Min_LBW	Decimal	Y	Tenth percentile from peer counties, birth measures, low birth wt. (<2500 g)
MeasuresOfBirthAndDeath	Max_LBW	Decimal	Y	Nintieth percentile from peer counties, birth measures, low birth wt. (<2500 g)
MeasuresOfBirthAndDeath	VLBW	Decimal	Y	County data, birth measures, very low birth wt. (<1500 g)
MeasuresOfBirthAndDeath	VLBW_Ind	Decimal	Y	Favorable indicator, birth measures, very low birth wt. (<1500 g)
MeasuresOfBirthAndDeath	Min_VLBW	Decimal	Y	Tenth percentile from peer counties, birth measures, very low birth wt. (<1500 g)
MeasuresOfBirthAndDeath	Max_VLBW	Decimal	Y	Nintieth percentile from peer counties, birth measures, very low birth wt. (<1500 g)

Example of Relative Health Importance

Page_Name	***Column_Data***	***DATA_TYPE***	***Is_%_Data***	***Description***
RelativeHealthImportance	State_FIPS_Code	Text	N	Two-digit state identifier, developed by the National Bureau of Standards
RelativeHealthImportance	County_FIPS_Code	Text	N	Three-digit county identifier, developed by the National Bureau of Standards
RelativeHealthImportance	CHSI_County_Name	Text	N	Name of county
RelativeHealthImportance	CHSI_State_Name	Text	N	Name of State or District of Columbia
RelativeHealthImportance	CHSI_State_Abbr	Text	N	Two-character postal abbreviation for state name

RelativeHealthImportance	Strata_ID_Number	Integer	N	CHSI Peer County Stratum Number
RelativeHealthImportance	RHI_LBW_Ind	Integer	N	Relative health indicator, low birth wt. (<2500 g)
RelativeHealthImportance	RHI_VLBW_Ind	Integer	N	Relative health indicator, very low birth wt. (<1500 g)
RelativeHealthImportance	RHI_Premature_Ind	Integer	N	Relative health indicator, premature births (<37 weeks)
RelativeHealthImportance	RHI_Under_18_Ind	Integer	N	Relative health indicator, births to women under 18
RelativeHealthImportance	RHI_Over_40_Ind	Integer	N	Relative health indicator, births to women over 40
RelativeHealthImportance	RHI_Unmarried_Ind	Integer	N	Relative health indicator, births to unmarried women
RelativeHealthImportance	RHI_Late_Care_Ind	Integer	N	Relative health indicator, no care in first trimester
RelativeHealthImportance	RHI_Infant_Mortality_Ind	Integer	N	Relative health indicator, infant mortality
RelativeHealthImportance	RHI_IM_Wh_Non_Hisp_Ind	Integer	N	Relative health indicator, White non Hispanic infant mortality
RelativeHealthImportance	RHI_IM_Bl_Non_Hisp_Ind	Integer	N	Relative health indicator, Black non Hispanic infant mortality

Example of Vulnerable Populations

Page_Name	***Column_Data***	***DATA_TYPE***	***Is_%_Data***	***Description***
VunerablePopsAndEnvHealth	State_FIPS_Code	Text	N	Two-digit state identifier, developed by the National Bureau of Standards
VunerablePopsAndEnvHealth	County_FIPS_Code	Text	N	Three-digit county identifier, developed by the National Bureau of Standards
VunerablePopsAndEnvHealth	CHSI_County_Name	Text	N	Name of county
VunerablePopsAndEnvHealth	CHSI_State_Name	Text	N	Name of State or District of Columbia

VunerablePopsAndEnvHealth	CHSI_State_Abbr	Text	N	Two-character postal abbreviation for state name
VunerablePopsAndEnvHealth	Strata_ID_Number	Integer	N	CHSI Peer County Stratum Number
VunerablePopsAndEnvHealth	No_HS_Diploma	Integer	N	County data, no high school diploma (among adults age 25 and older)
VunerablePopsAndEnvHealth	Unemployed	Integer	N	County data, unemployed
VunerablePopsAndEnvHealth	Sev_Work_Disabled	Integer	N	County data, severely work disabled
VunerablePopsAndEnvHealth	Major_Depression	Integer	N	County data, major depression
VunerablePopsAndEnvHealth	Recent_Drug_Use	Integer	N	County data, recent drug users (within past month)
VunerablePopsAndEnvHealth	Ecol_Rpt	Integer	N	County data, E.coli reported cases
VunerablePopsAndEnvHealth	Ecol_Rpt_Ind	Integer	N	Favorable indicator, E.coli
VunerablePopsAndEnvHealth	Ecol_Exp	Integer	N	County data, E.coli expected cases

Example of Risk Factors and Access to Care

Page_Name	***Column_Data***	***DATA_TYPE***	***Is_%_Data***	***Description***
RiskFactorsAndAccessToCare	State_FIPS_Code	Text	N	Two-digit state identifier, developed by the National Bureau of Standards
RiskFactorsAndAccessToCare	County_FIPS_Code	Text	N	Three-digit county identifier, developed by the National Bureau of Standards
RiskFactorsAndAccessToCare	CHSI_County_Name	Text	N	Name of county
RiskFactorsAndAccessToCare	CHSI_State_Name	Text	N	Name of State or District of Columbia
RiskFactorsAndAccessToCare	CHSI_State_Abbr	Text	N	Two-character postal abbreviation for state name

RiskFactorsAndAccessToCare	Strata_ID_Number	Integer	N	CHSI Peer County Stratum Number
RiskFactorsAndAccessToCare	No_Exercise	Decimal	Y	County data, no exercise
RiskFactorsAndAccessToCare	Few_Fruit_Veg	Decimal	Y	County data, few fruits/vegetables
RiskFactorsAndAccessToCare	Obesity	Decimal	Y	County data, obesity
RiskFactorsAndAccessToCare	High_Blood_Pres	Decimal	Y	County data, high blood pressure

Appendix 3: Data Frame Structure

'data.frame':3141 obs. of 140 variables:

```
 $ County_FIPS_Code        : Factor w/ 324 levels "1","100","101",..: 1 134 246 288 318 8 22 35 49 64 ...

 $ State_FIPS_Code          : Factor w/ 51 levels "1","10","11",..: 1 1 1 1 1 1 1 1 1 1 ...

 $ CHSI_County_Name       : Factor w/ 1840 levels "Abbeville","Acadia",..: 82 89 100 150 165 226 236 247 294 316 ...

 $ CHSI_State_Name           : Factor w/ 51 levels "Alabama","Alaska",..: 1 1 1 1 1 1 1 1 1 1 ...

 $ CHSI_State_Abbr           : Factor w/ 51 levels "AK","AL","AR",..: 2 2 2 2 2 2 2 2 2 2 ...

 $ Strata_ID_Number          : Factor w/ 88 levels "1","10","11",..: 22 8 47 37 21 73 74 56 46 61 ...

 $ Strata_Determining_Factors : Factor w/ 4 levels "frontier status, population size",..: 3 3 4 3 3 4 4 2 4 4 ...

 $ Number_Counties            : Factor w/ 37 levels "15","16","20",..: 20 10 16 33 22 20 21 33 10 24 ...

 $ Population_Size :Factor w/ 3088 levels "100018","100086",..: 2153 649 1444 1049 2320 120 1001 143 1718 1233.

 $ Population_Density        : Factor w/ 587 levels "0","1","10","100",..: 548 9 318 341 556 161 274 167 477 407 ...

 $ Poverty                   : Factor w/ 259 levels "10","10.1","10.2",..: 5 3 123 69 20 155 102 65 63 53 ...

 $ Age_19_Under              : Factor w/ 208 levels "1.4","13.4","14",..: 111 77 85 88 87 89 98 83 90 61 ...

 $ Age_19_64                 : Factor w/ 214 levels "47.6","48.8",..: 124 104 126 134 122 133 86 117 96 115 ...
```

```
$ Age_65_84                 : Factor w/ 203 levels "10","10.1","10.2",..: 202 46 17 10 22 1 37 28 36 53 ...

$ Age_85_and_Over           : Factor w/ 64 levels "0.1","0.2","0.3",..: 9 18 16 12 13 22 24 15 22 14 ...

$ White                     : Factor w/ 561 levels "100","13.1","13.4",..: 370 446 137 333 533 32 166 347 195 495 ...

$ Black                     : Factor w/ 491 levels "0","0.1","0.2",..: 89 491 343 146 16 457 313 112 280 362 ...

$ Native_American           : Factor w/ 187 levels "0","0.1","0.2",..: 6 6 5 4 6 5 3 5 2 4 ...

$ Asian                     : Factor w/ 112 levels "0","0.1","0.2",..: 7 5 4 2 3 3 4 9 3 4 ...

$ Hispanic                  : Factor w/ 391 levels "0","0.1","0.2",..: 18 113 182 15 323 296 10 110 13 12 ...

$ No_HS_Diploma  : Factor w/ 2792 levels "100","100009",..: 2286 782 2294 1986 139 1159 1688 749 2622 2223 ...

$ Unemployed : Factor w/ 1817 levels "100","1000","1003",..: 1603 752 1375 1026 1654 955 1329 626 1679 1217 ...

$ Sev_Work_Disabled: Factor w/ 1923 levels "1","10","100",..: 479 1330 248 1832 689 1250 1902 1451 349 124 ...

$ Major_Depression  : Factor w/ 2387 levels "100","10002",..: 1059 2317 513 213 1226 1907 161 1925 741 388 ...

$ Recent_Drug_Use   : Factor w/ 2310 levels "100","1000","10035",..: 888 2058 347 33 991 1745 26 1725 511 140
...

$ Ecol_Rpt                  : Factor w/ 67 levels "0","1","10","102",..: 17 17 1 17 2 1 1 1 1 2 ...

$ Ecol_Rpt_Ind              : Factor w/ 2 levels "3","4": 1 1 1 2 1 1 1 1 1 1 ...

$ Ecol_Exp                  : Factor w/ 59 levels "0","1","10","106",..: 35 35 1 15 43 2 1 27 15 15 ...

$ Salm_Rpt                  : Factor w/ 287 levels "0","1","10","100",..: 210 287 218 275 126 73 114 153 156 230 ...

$ Salm_Rpt_Ind              : Factor w/ 2 levels "3","4": 2 2 2 1 1 1 1 1 2 1 ...

$ Salm_Exp                  : Factor w/ 272 levels "0","1","10","100",..: 146 233 135 148 146 86 183 214 135 161 ...

$ Shig_Rpt                  : Factor w/ 209 levels "0","1","10","100",..: 128 132 2 107 12 174 144 115 2 161 ...

$ Shig_Rpt_Ind              : Factor w/ 2 levels "3","4": 1 2 1 1 2 2 1 2 1 2 ...

$ Shig_Exp                  : Factor w/ 196 levels "0","1","10","100",..: 12 25 38 38 142 165 42 104 188 188 ...

$ Toxic_Chem    : Factor w/ 2194 levels "0","1","10","100",..: 1007 1206 1510 1883 2075 NA 2174 1971 969 1858 ...

$ Carbon_Monoxide_Ind       : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...

$ Nitrogen_Dioxide_Ind      : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...

$ Sulfur_Dioxide_Ind        : Factor w/ 1 level "1": 1 1 1 1 1 1 1 1 1 1 ...

$ Ozone_Ind                 : Factor w/ 2 levels "1","2": 1 2 1 1 1 1 1 1 1 1 ...

$ Particulate_Matter_Ind    : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...

$ Lead_Ind                  : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...

$ EH_Time_Span              : Factor w/ 0 levels: NA NA NA NA NA NA NA NA NA NA ...

$ No_Exercise               : Factor w/ 330 levels "10.2","10.3",..: 164 158 NA NA 221 NA 131 178 233 NA ...
```

```
 $ Few_Fruit_Veg           : Factor w/ 261 levels "63.1","63.3",..: 128 104 NA 208 88 NA NA 161 188 NA ...

 $ Obesity                 : Factor w/ 263 levels "10.2","10.5",..: 129 120 140 NA 126 NA 104 154 NA NA ...

 $ High_Blood_Pres         : Factor w/ 260 levels "10.8","11.4",..: 149 163 NA NA NA NA NA 190 NA NA ...

 $ Smoker                  : Factor w/ 300 levels "10.1","10.2",..: 162 142 73 NA 132 NA 169 151 14 132 ...

 $ Diabetes                : Factor w/ 165 levels "0.5","0.6","0.7",..: 54 138 132 43 150 NA 161 24 76 45 ...

 $ Uninsured               : Factor w/ 2725 levels "10002","1001",..: 2064 778 1935 1398 2490 937 1668 568 2025 1515 ...

 $ Elderly_Medicare        : Factor w/ 2717 levels "10003","1003",..: 1842 926 1386 989 1912 402 1320 565 2031 1294 ...

 $ Disabled_Medicare : Factor w/ 1822 levels "100","1001","1002",..: 178 1090 78 1799 236 1284 1779 1326 345
1696 ...

 $ Prim_Care_Phys_Rate     : Factor w/ 1133 levels "0","10.1","10.2",..: 616 840 621 581 279 710 596 928 702 441 ...

 $ Dentist_Rate            : Factor w/ 704 levels "0","10","10.1",..: 153 238 173 110 10 105 117 354 153 38 ...

 $ Community_Health_Center_Ind: Factor w/ 2 levels "1","2": 1 1 1 1 2 1 1 1 1 2 ...

 $ HPSA_Ind                : Factor w/ 2 levels "1","2": 2 2 2 1 1 1 2 2 2 1 ...

 $ A_Wh_Comp               : Factor w/ 57 levels "14","17","18",..: NA 41 NA NA 18 NA NA 20 NA NA ...

 $ A_Bl_Comp               : Factor w/ 50 levels "25","29","30",..: NA NA NA NA NA NA NA NA NA NA ...

 $ A_Ot_Comp               : Factor w/ 25 levels "14","19","23",..: NA NA NA NA NA NA NA NA NA NA ...

 $ A_Hi_Comp               : Factor w/ 43 levels "24","30","31",..: NA NA NA NA NA NA NA NA NA NA ...

 $ A_Wh_BirthDef           : Factor w/ 42 levels "10","11","12",..: NA 12 NA NA 25 NA NA 5 NA NA ...

 $ A_Bl_BirthDef           : Factor w/ 22 levels "10","11","12",..: NA NA NA NA NA NA NA NA NA NA ...

 $ A_Ot_BirthDef           : Factor w/ 21 levels "12","13","14",..: NA NA NA NA NA NA NA NA NA NA ...

 $ A_Hi_BirthDef           : Factor w/ 30 levels "11","13","14",..: NA NA NA NA NA NA NA NA NA NA ...

 $ B_Wh_Injury             : Factor w/ 59 levels "10","12","13",..: NA NA NA NA NA NA NA NA NA NA ...

 $ B_Bl_Injury             : Factor w/ 38 levels "13","16","17",..: NA NA NA NA NA NA NA NA NA NA ...

 $ B_Ot_Injury             : Factor w/ 17 levels "13","17","19",..: NA NA NA NA NA NA NA NA NA NA ...

 $ B_Hi_Injury             : Factor w/ 27 levels "18","21","24",..: NA NA NA NA NA NA NA NA NA NA ...

 $ B_Wh_Cancer             : Factor w/ 22 levels "10","11","12",..: NA NA NA NA NA NA NA NA NA NA ...

 $ B_Bl_Cancer             : Factor w/ 10 levels "10","11","12",..: NA NA NA NA NA NA NA NA NA NA ...

 $ B_Ot_Cancer             : Factor w/ 8 levels "10","11","13",..: NA NA NA NA NA NA NA NA NA NA ...

 $ B_Hi_Cancer             : Factor w/ 17 levels "10","11","12",..: NA NA NA NA NA NA NA NA NA NA ...

 $ B_Wh_Homicide           : Factor w/ 10 levels "10","11","12",..: NA NA NA NA NA NA NA NA NA NA ...

 $ B_Bl_Homicide           : Factor w/ 16 levels "10","11","12",..: NA NA NA NA NA NA NA NA NA NA ...
 $ B_Ot_Homicide         : Factor w/ 1 level "10": NA NA NA NA NA NA NA NA NA NA ...
 $ B_Hi_Homicide         : Factor w/ 5 levels "10","11","13",..: NA NA NA NA NA NA NA NA NA NA ...
 $ C_Wh_Injury           : Factor w/ 69 levels "14","15","19",..: NA 42 NA 42 43 NA 46 36 NA 47 ...
 $ C_Bl_Injury           : Factor w/ 54 levels "10","11","12",..: NA NA NA NA NA NA 11 NA NA NA ...
 $ C_Ot_Injury           : Factor w/ 36 levels "18","24","28",..: NA NA NA NA NA NA NA NA NA NA ...
 $ C_Hi_Injury           : Factor w/ 53 levels "15","20","21",..: NA NA NA NA NA NA NA NA NA NA ...
 $ C_Wh_Homicide         : Factor w/ 24 levels "10","11","12",..: NA NA NA NA NA NA NA NA NA NA ...
 $ C_Bl_Homicide         : Factor w/ 58 levels "10","11","12",..: NA NA NA NA NA NA 4 NA NA NA ...
 $ C_Ot_homicide         : Factor w/ 18 levels "10","11","12",..: NA NA NA NA NA NA NA NA NA NA ...
 $ C_Hi_Homicide         : Factor w/ 36 levels "10","11","12",..: NA NA NA NA NA NA NA NA NA NA ...
 $ C_Wh_Suicide          : Factor w/ 33 levels "10","11","12",..: NA NA NA 4 NA NA 4 2 NA 3 ...
 $ C_Bl_Suicide          : Factor w/ 13 levels "10","11","12",..: NA NA NA NA NA NA 10 NA NA NA ...
 $ C_Ot_Suicide          : Factor w/ 23 levels "10","11","12",..: NA NA NA NA NA NA NA NA NA NA ...
 $ C_Hi_Suicide          : Factor w/ 16 levels "10","11","12",..: NA NA NA NA NA NA NA NA NA NA ...
 $ C_Wh_Cancer           : Factor w/ 14 levels "10","11","12",..: NA NA NA NA NA NA NA NA NA NA ...
  [list output truncated]
```