Final Project, 2020SP - DATA WRANGLING 16:954:597:01

Elena Novikova

5/04/2020

GitHub repository for the project: https://github.com/elkanovikova/final_project

For my final project I selected a dataset of Community Health Status Indicators from the Data.gov website. Let's read the description of teh dataset from the website:

Table 1: CHSItable

x
Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease and Cancer Metadata Updated: February 26, 2020 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 CHSI report and dataset was designed not only for public health professionals but also for members of the community who are interested in the health of their community. 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.

The file called DATAELEMENTDESCRIPTION.csv contains column names and column descriptions for each file from the source. I will import it and nest it into a dataframe with values in the first column corresponding to file names, and second column containing dataframes with column descriptions for each file.

Here is the list of csv data files in the imported CHSI dataset:

```
## # A tibble: 8 x 2
## # Groups:   PAGE_NAME [8]
##   PAGE_NAME                 data
##   <fct>                     <list>
## 1 Demographics              <tibble [44 x 5]>
## 2 SummaryMeasuresOfHealth   <tibble [28 x 5]>
## 3 LeadingCausesOfDeath      <tibble [235 x 5]>
## 4 MeasuresOfBirthAndDeath   <tibble [141 x 5]>
## 5 RelativeHealthImportance  <tibble [28 x 5]>
## 6 VunerablePopsAndEnvHealth  <tibble [28 x 5]>
## 7 PreventiveServicesUse      <tibble [43 x 5]>
## 8 RiskFactorsAndAccessToCare <tibble [31 x 5]>
```

Let's start with looking at column descriptions in the Demographics file:

Table 2: Demographics

COLUMN_NAME	DATA_TYPE	IS_PERCENT_DATA	DESCRIPTION
State_FIPS_Code	Text	N	Two-digit state identifier, developed by the National Bu
County_FIPS_Code	Text	N	Three-digit county identifier, developed by the National
CHSI_County_Name	Text	N	Name of county
CHSI_State_Name	Text	N	Name of State or District of Columbia
CHSI_State_Abbr	Text	N	Two-character postal abbreviation for state name
Strata_ID_Number	Integer	N	CHSI Peer County Stratum Number

The CHSI dataset is labeling counties with Strata IDs. Here is the information provided about stratas, or "Peer County Groups", on the countyhealthrankings.org website. I used the description given on sheet 1 of the csv file posted on this website:

Table 3: Peer Counties

x
County Health Rankings and Roadmaps and CDC's Community Health Status Indicators (CHSI) have teamed up to offer an enhanced peer county comparison feature. This excel file provides information on the groups of counties that could be considered peers based on key demographic, social, and economic indicators. To utilize this feature, please locate your county (by county name or FIPS code) in the second tab of this file. The number in the "Peer County Group" column indicates the peer cluster for your county. To learn about others in this peer county group, you can sort or filter the spreadsheet for other counties in this group. Once you have identified the counties in your peer group, you can use the CHR&R compare counties feature to explore the health factors and outcomes across counties in your peer group.

Output files County_FIPS.xlsx and County_FIPS.csv are generated. County name and State names are put into two separate columns.

I am a resident of Passaic county, NJ. I will find what Strata my county belongs to, find Peer counties in this Strata, and count how many Peer counties are there:

```
## [1] "Passaic county, NJ belongs to strata 9"
```

Table 4: Peer counties in Strata 9, a total of 34 counties

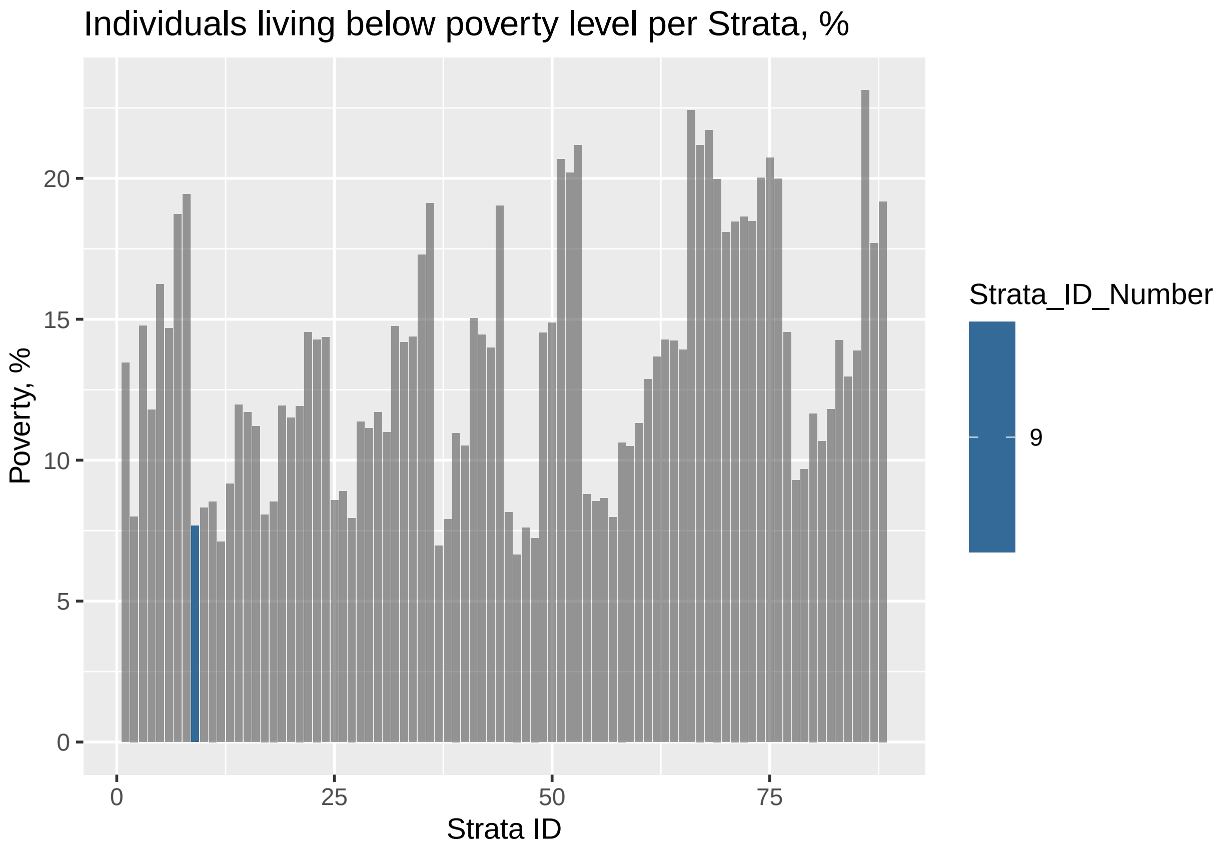
FIPS	PeerCountyGroup	County	State
6069	9	San Benito County	California
6071	9	San Bernardino County	California
6113	9	Yolo County	California
8001	9	Adams County	Colorado
12097	9	Osceola County	Florida
13063	9	Clayton County	Georgia
13089	9	DeKalb County	Georgia
13097	9	Douglas County	Georgia
13135	9	Gwinnett County	Georgia
13151	9	Henry County	Georgia
13217	9	Newton County	Georgia
13247	9	Rockdale County	Georgia
17037	9	DeKalb County	Illinois
17163	9	St. Clair County	Illinois
18089	9	Lake County	Indiana
20209	9	Wyandotte County	Kansas
22051	9	Jefferson Parish	Louisiana
22087	9	St. Bernard Parish	Louisiana
25005	9	Bristol County	Massachusetts
26099	9	Macomb County	Michigan
34007	9	Camden County	New Jersey
34031	9	Passaic County	New Jersey
36071	9	Orange County	New York
37071	9	Gaston County	North Carolina
41071	9	Yamhill County	Oregon
51570	9	Colonial Heights city	Virginia
51630	9	Fredericksburg city	Virginia
51650	9	Hampton city	Virginia
51670	9	Hopewell city	Virginia
51700	9	Newport News city	Virginia
51740	9	Portsmouth city	Virginia
51800	9	Suffolk city	Virginia
53053	9	Pierce County	Washington
55059	9	Kenosha County	Wisconsin

I will now work with the Demographics dataframe from the CHSI dataset, will look at the poverty levels by strata. Let's see how the strata my county belongs to compares to other stratas in the country.

Table 5: Passaic County, NJ

…
Passaic county, NJ belongs to Strata 9 that ranks 6th lowest in Poverty level amongst the total of 88 Stratas in the US. 7.675% of the population of Strata 9 lives below poverty level.

Now that information on Strata 9 ranking is obtained, I will demonstrate it on a bar plot. Strata 9 is shown in blue, and the plot confirms the information received above. We see how most of the stratas have a higher poverty level.



The Population density data is provided by the Demographics file. I will select pertinent columns and clean the data of missing values.

Table 6: County data, population density (people per square mile)

CHSI_County_Name	CHSI_State_Name	Population_Density
Autauga	Alabama	82
Baldwin	Alabama	102
Barbour	Alabama	32
Bibb	Alabama	35
Blount	Alabama	86
Bullock	Alabama	18

I will filter out and print 10 most populated counties in the country.

Table 7: Top 10 counties with highest population density in the US

CHSI_County_Name	CHSI_State_Name	Population_Density
New York	New York	69390
Kings	New York	35211
Bronx	New York	32300
Queens	New York	20520
San Francisco	California	15837
Hudson	New Jersey	12926
Suffolk	Massachusetts	11183
Philadelphia	Pennsylvania	10832
Washington	District of Columbia	8966
Alexandria City	Virginia	8915

Four of the boroughs of New York City are leading the list leaving other counties far behind. No wonder they have the most of COVID-19 cases. Another proof that self isolation is essential to stop spreading COVID-19.

Now I will import and review the SUMMARYMEASURESOFHEALTH.csv file for the average life expectancy data.

Table 8: County data, average life expectancy

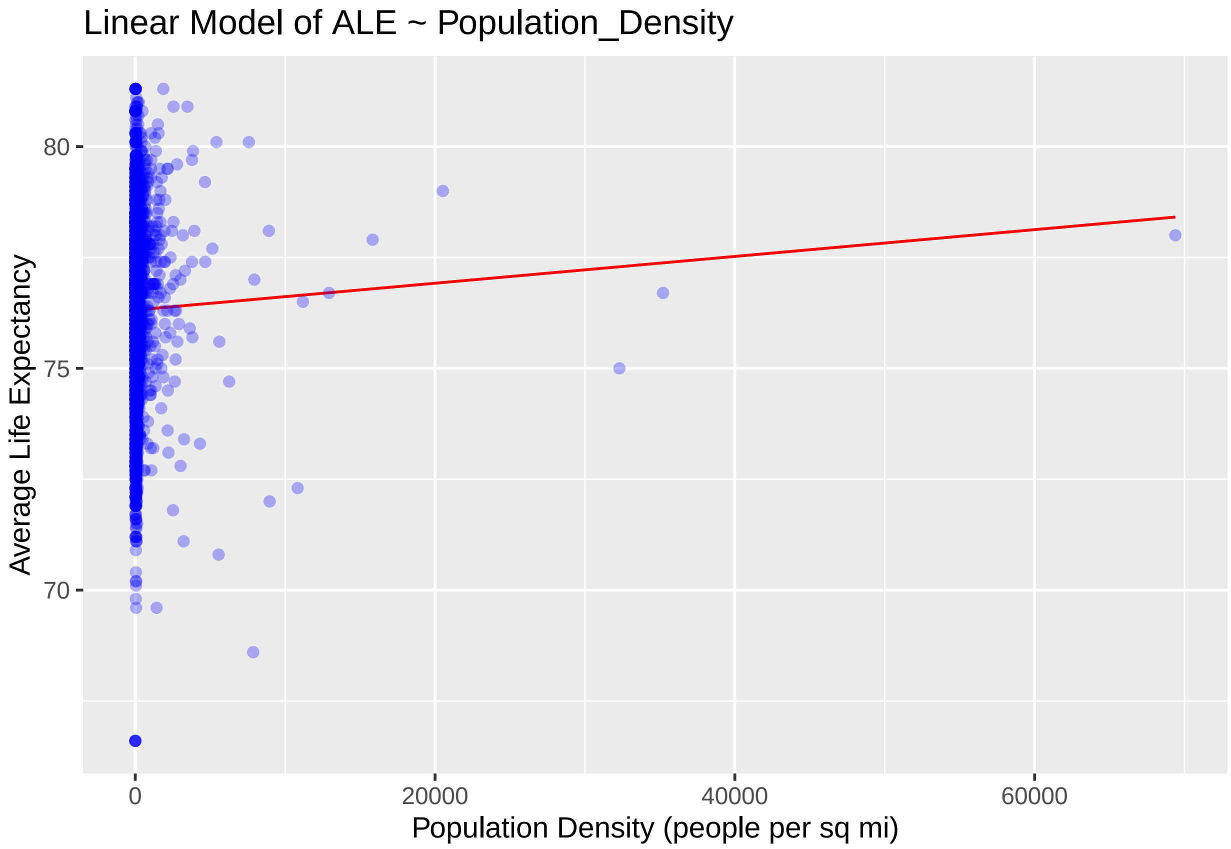
CHSI_County_Name	CHSI_State_Name	ALE
Autauga	Alabama	74.9
Baldwin	Alabama	76.6
Barbour	Alabama	74.5
Bibb	Alabama	73.2
Blount	Alabama	76.1
Bullock	Alabama	71.9

I will left join the Population density and the Life Expectancy data frames to produce a Linear Model of these two variables:

Table 9: Population Density and Average Life Expectancy by county

CHSI_County_Name	CHSI_State_Name	Population_Density	ALE
Autauga	Alabama	82	74.9
Baldwin	Alabama	102	76.6
Barbour	Alabama	32	74.5
Bibb	Alabama	35	73.2
Blount	Alabama	86	76.1
Bullock	Alabama	18	71.9

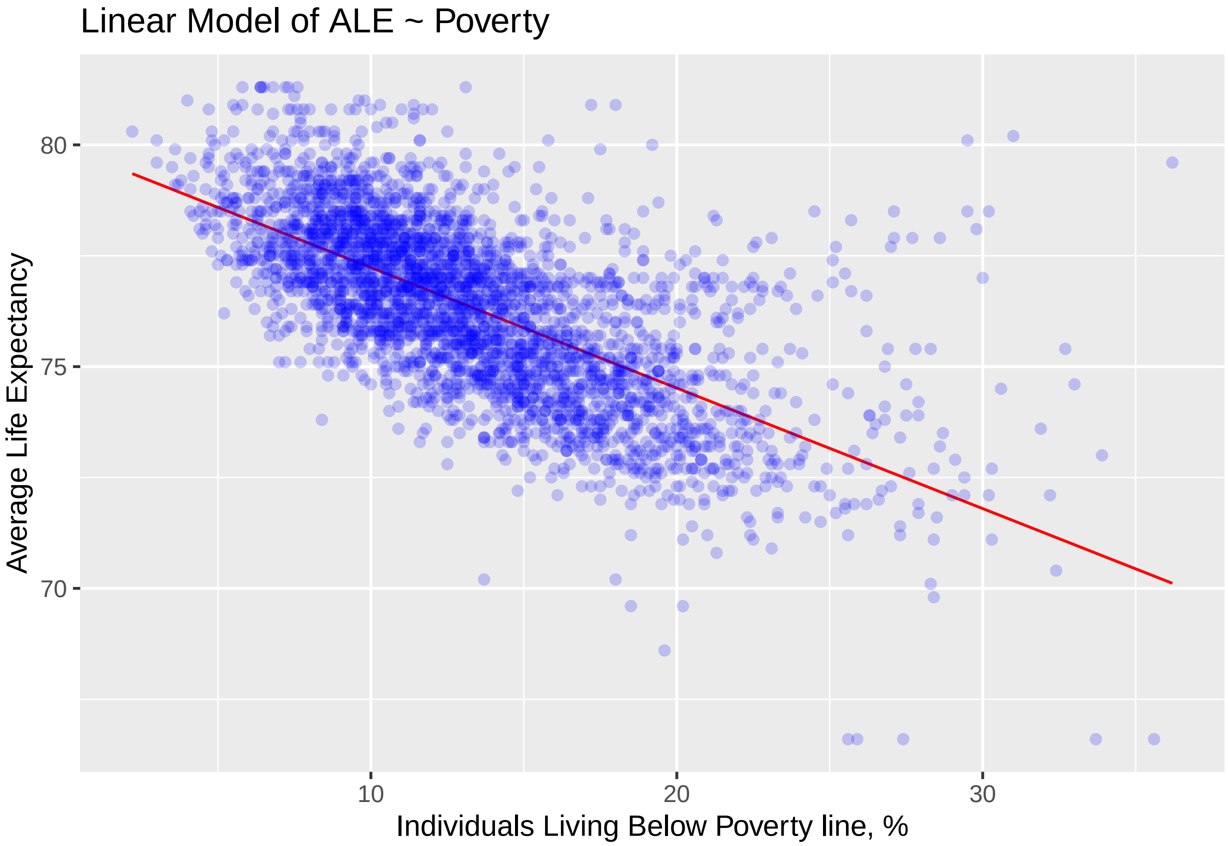
I built a Linear Model of ALE ~ Population_Density graph below, but does not look very informative. The scatter plot is jammed to the left. There are probably very few counties with high population density. I listed them previously in this report in Table 9.



I will build a Linear Model of Average Life Expectancy ~ Poverty and see if there is a better correlation between these two variables. See the LM data printed below:

```
##  State_FIPS_Code County_FIPS_Code   CHSI_County_Name CHSI_State_Name
##  Min.   : 1.00   Min.   :  1.0    Washington:  32    Texas   : 254
##  1st Qu.:19.00   1st Qu.: 35.0    Jefferson :  26    Georgia : 159
##  Median :29.00   Median : 79.0    Franklin  :  25    Virginia: 134
##  Mean   :30.33   Mean   :103.7    Jackson   :  24    Kentucky: 120
##  3rd Qu.:45.00   3rd Qu.:133.0    Lincoln   :  24    Missouri: 115
##  Max.   :56.00   Max.   :840.0    Madison   :  20    Kansas  : 105
##                                   (Other)   :2987    (Other) :2251
##  Strata_ID_Number      ALE           Poverty           fit
##  Min.   : 1.00    Min.   :66.60   Min.   : 2.20   Min.   :70.11
##  1st Qu.:23.00    1st Qu.:75.00   1st Qu.: 9.80   1st Qu.:75.55
##  Median :44.00    Median :76.50   Median :12.60   Median :76.53
##  Mean   :44.68    Mean   :76.32   Mean   :13.35   Mean   :76.32
##  3rd Qu.:66.00    3rd Qu.:77.70   3rd Qu.:16.20   3rd Qu.:77.29
##  Max.   :88.00    Max.   :81.30   Max.   :36.20   Max.   :79.35
##
```

The LM of Average Life Expectancy ~ Poverty plot below provides a visible correlation. People in populations with less percent below poverty line tend to live longer lives.



Next, I will build the Average Life Expectancy data by county on the US map using the ALEdf dataframe I previously extracted from the SUMMARYMEASURESOFHEALTH.csv file. To use the choroplethr library I need the FIPS codes be in a 5-digit format. The original file has them separated into county and state code columns. I will add a new FIPS column.

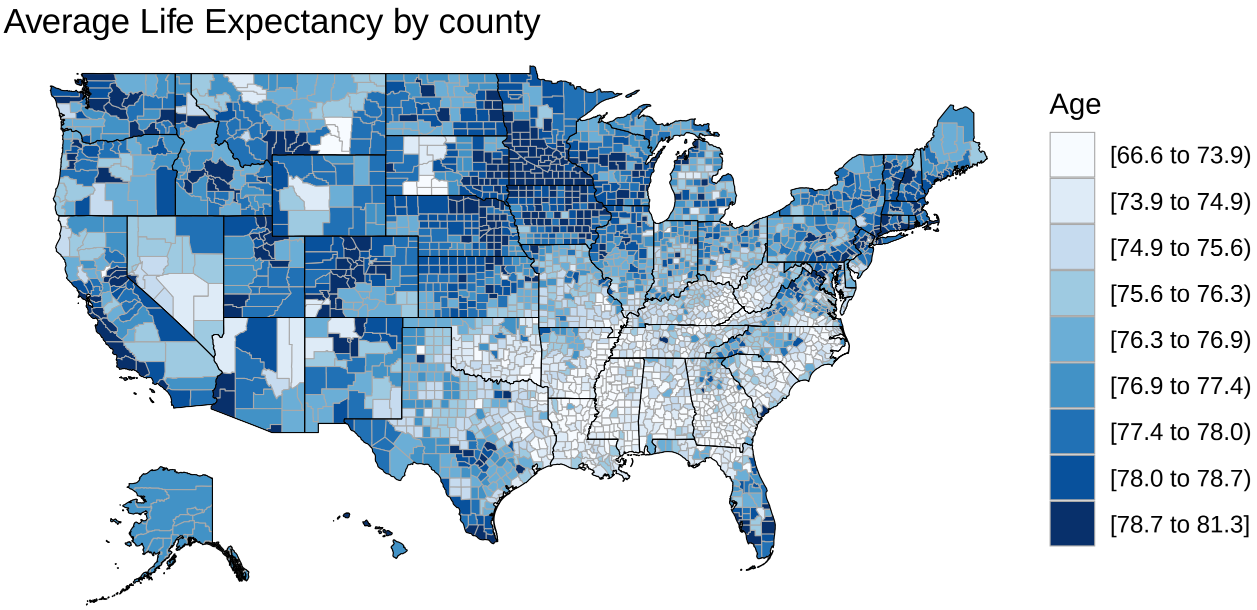
Table 10: FIPS codes added to the Average Life Expectancy by county dataframe

FIPS	CHSI_County_Name	CHSI_State_Name	ALE
1001	Autauga	Alabama	74.9
1003	Baldwin	Alabama	76.6
1005	Barbour	Alabama	74.5
1007	Bibb	Alabama	73.2
1009	Blount	Alabama	76.1
1011	Bullock	Alabama	71.9

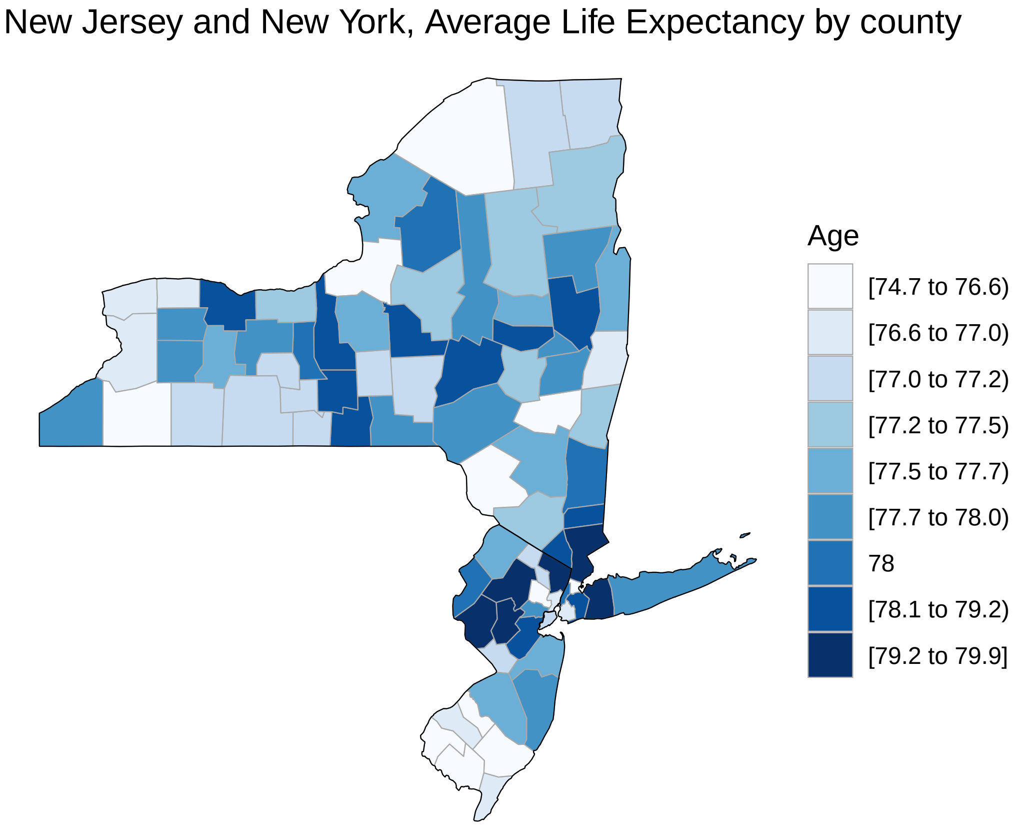
Since the FIPS codes are added, I can plot the ALE data on the US map.

While working with choroplethr I noticed that the fips code have to be in a numeric format, and missing leading zeros for 1-digit state codes are not a problem.

The resulting map has very interesting patterns that could be further explored. We can see the area covering states from Texas to Carolinas where ALE is consistely low. I wonder what factors are causing the ALE being relatively low on such a large area.



I zoomed in to New York and New Jersey. Downstate NY is doing great in terms of ALE, and similar do Bergen, Morris, Hunterdon, and Somerset counties in New Jersey.


New Jersey and New York, Average Life Expectancy by county

An excel file of average life expectancy with FIPS codes named ALE_with_FIPS.xlsx is created.

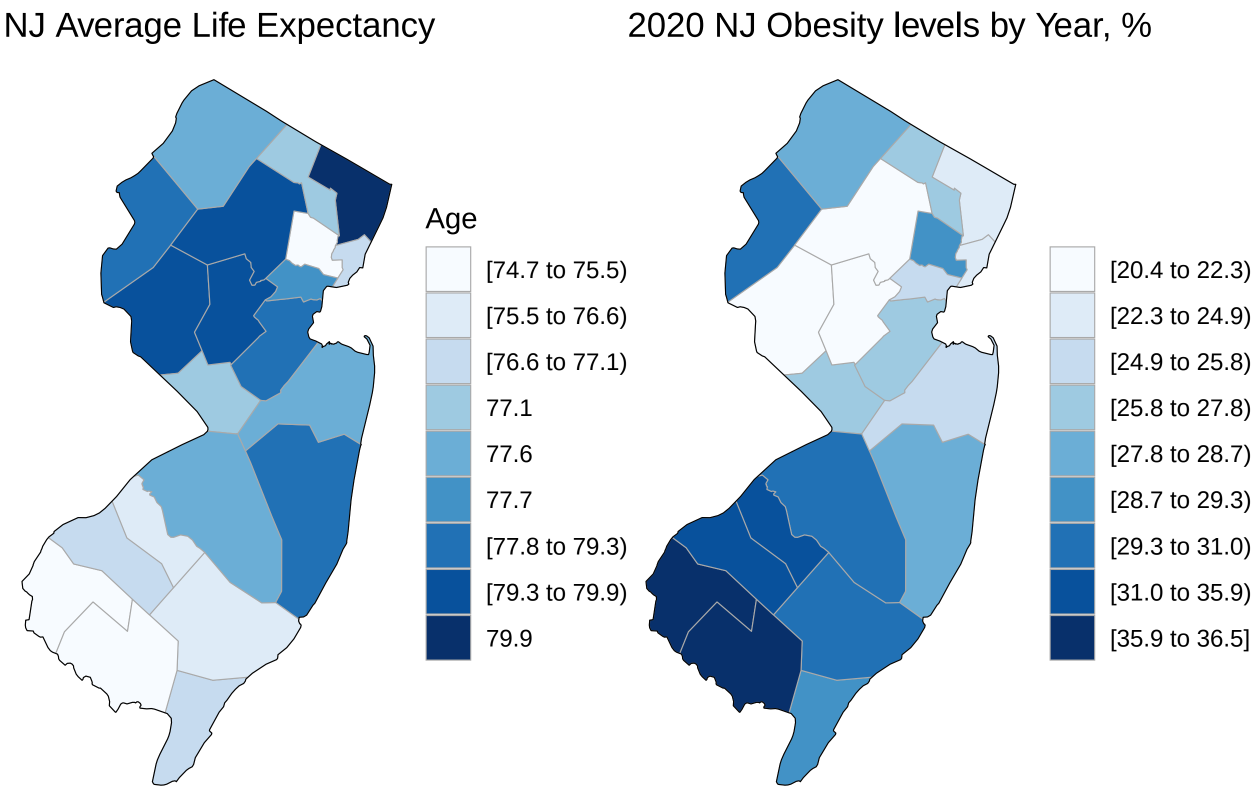
My second source of data, the countyhealthrankings.org website provides yearly health data for each state. I downloded files for New Jersey, years 2010 - 2020, and extracted data on Adult Obesity. I merged data from all years into one tibble to furhter work with it.

Table 11: New Jersey Adult Obesity levels, %

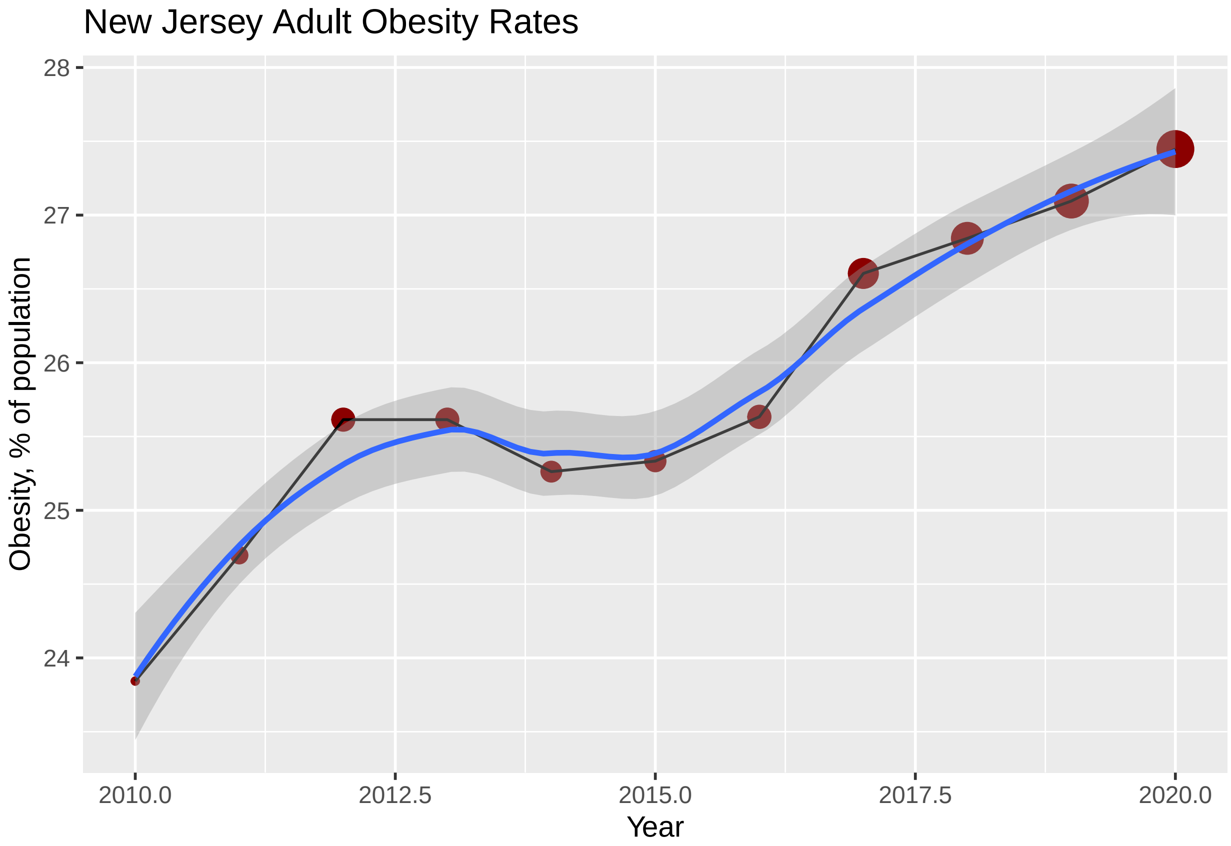
FIPS	County	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
34001	Atlantic	26.5	26.5	28.2	28.2	27.0	26.8	26.8	27.6	27.6	27.2	29.8
34003	Bergen	19.6	20.8	21.8	21.8	21.4	20.8	20.4	21.5	22.1	23.0	22.3
34005	Burlington	26.0	26.2	27.6	27.6	27.1	27.0	27.0	28.0	28.2	28.1	29.3
34007	Camden	25.8	26.8	27.9	27.9	27.5	28.3	29.0	30.2	29.4	29.2	31.4
34009	Cape May	25.1	24.8	25.4	25.4	24.8	26.3	27.1	27.9	28.8	27.8	28.7
34011	Cumberland	27.2	29.6	33.3	33.3	33.4	33.9	33.6	34.5	34.7	35.1	35.9
34013	Essex	25.3	26.1	26.0	26.0	25.9	26.5	27.3	28.7	28.5	28.5	29.1
34015	Gloucester	24.9	25.6	27.0	27.0	28.2	28.6	29.2	30.3	30.9	31.2	31.0
34017	Hudson	22.8	24.1	23.7	23.7	23.9	23.6	23.9	23.9	23.1	23.7	23.9
34019	Hunterdon	18.8	19.8	20.8	20.8	20.0	20.6	21.3	22.3	21.4	20.8	20.4

FIPS	County	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
34021	Mercer	22.2	24.0	25.0	25.0	25.1	23.7	24.3	24.6	25.2	25.2	25.8
34023	Middlesex	23.8	23.7	23.7	23.7	23.4	22.6	24.0	25.0	26.8	25.9	25.9
34025	Monmouth	20.4	21.7	21.9	21.9	22.4	23.0	22.7	23.2	22.6	23.4	25.2
34027	Morris	20.4	20.8	21.9	21.9	20.8	21.4	20.4	21.0	21.4	22.0	21.3
34029	Ocean	26.4	25.8	27.1	27.1	26.4	26.8	26.8	28.1	28.7	29.4	28.1
34031	Passaic	24.7	23.7	24.4	24.4	24.7	23.6	24.1	25.7	28.2	28.7	27.3
34033	Salem	28.7	29.7	34.2	34.2	32.8	32.3	32.0	33.9	33.6	34.3	36.5
34035	Somerset	20.4	22.3	21.6	21.6	21.1	21.3	21.3	22.6	22.3	23.6	22.2
34037	Sussex	23.7	26.9	26.7	26.7	25.8	24.6	25.5	26.2	27.6	28.1	27.8
34039	Union	21.1	22.2	22.3	22.3	23.2	23.7	24.5	24.7	24.7	24.8	24.9
34041	Warren	26.9	27.5	27.4	27.4	25.6	26.6	27.1	28.8	27.9	29.0	29.6

Since I have ALE and Obesity data, I will put it on the NJ map side by side using the grid.arrange function. I used the 2020 Obesity data for this plot. Although expected, still interesting to see how counties with higher obesity rates overlap with lower Average Life Expectancy rates.



And the last step is to display how the obesity rate changed in a span of 11 years in New Jersey. Unfortunately, this rate is consistently increasing according to the plot below.



GitHub repository for the project: https://github.com/elkanovikova/final_project

Bibliography:

1. The County Health Rankings & Roadmaps program. (August 30, 2017). Peer Counties Tool.

Retrieved 30 April 2020, from https://www.countyhealthrankings.org/resources/peer-counties-tool

2. The County Health Rankings & Roadmaps program. (2020). New Jersey Rankings Data.

Retrieved 30 April 2020, from https://www.countyhealthrankings.org/app/new-jersey/2020/downloads

3. U.S. Government's open data. (February 26, 2020). Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease and Cancer.

Retrieved 30 April 2020, from https://catalog.data.gov/dataset/community-health-status-indicators-chsi-to-combat-obesity-heart-disease-and-cancer