Exploratory Data Analysis of PLACES: Local Data for Better Health

Xuemao Zhang*

Abstract

PLACES is an extension of the 500 Cities project launched in 2015 which is a collaboration between CDC (Centers for Disease Control and Prevention), the Robert Wood Johnson Foundation, and the CDC Foundation. The purpose of the project is to provide city- and census tract-level small area estimates for 27 chronic disease risk factors, health outcomes, and clinical preventive service use for the largest 500 cities in the United States. In this exploratory data analysis, visualizations, correlation and regression analysis of the health-related variables are presented to better understand the health burden and needs of the United States over the last five years.

Key Words: Correlation, data visualization, exploratory data analysis, longitudinal data analysis, regression analysis.

1. Introduction

PLACES (`https://www.cdc.gov/places/index.html`) is an extension of the 500 Cities project launched in 2015 which is a collaboration between CDC, the Robert Wood Johnson Foundation, and the CDC Foundation. The purpose of the project is to provide city- and census tract-level small area estimates for 28 chronic disease risk factors, health outcomes, and clinical preventive service use for the largest 500 cities in the United States. PLACES provides model-based population-level analysis and community estimates to all counties, places (incorporated and census designated places), census tracts, and ZIP Code Tabulation Areas (ZCTAs) across the United States. In this paper, I focus on the estimates of the variables at city levels only without considering estimating errors.

The 27 variables measured are classified as 13 health outcomes listed in Table 1, 9 clinical preventive service use variables listed in Table 2, and 5 unhealthy behaviors listed in Table 3.

In section 2, the distributions of the 27 variables by years are checked separately. A special interest is the study of the correlation among the variables. The correlations among the 13 health outcomes, 10 use of preventive services, and the 5 chronic disease-related unhealthy behaviors are inspected in section 3. Furthermore, the correlations between prevention and health outcome variables, and the correlations between the unhealthy behaviors and health outcomes are investigated, respectively. All variables are estimated at the state level and the patterns are summarized in Section 4. A brief summary follows in section 5.

*xzhang2@esu.edu, East Stroudsburg University, 200 Prospect Street, East Stroudsburg, PA 18301

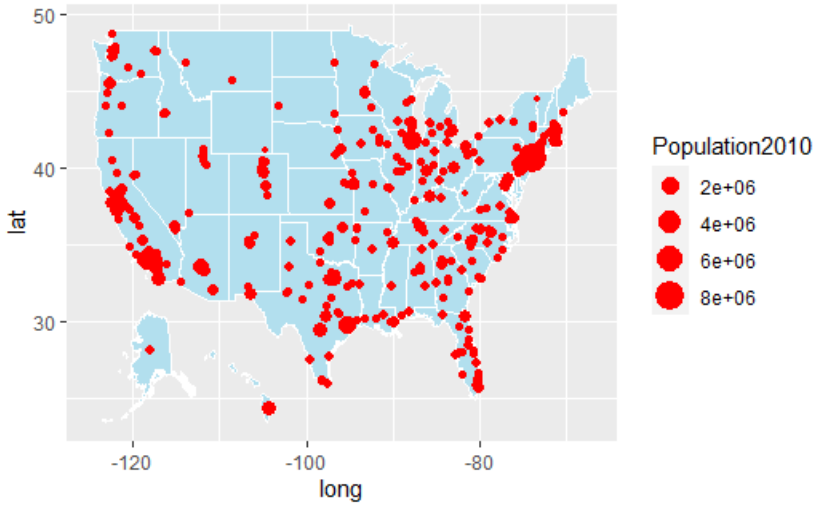

Figure 1: Distribution of the cities 500-Cities project

Table 1: Description of Health Outcome Variables in PLACES

Name	Description
ARTHRITIS	Arthritis among adults aged ≥18 years
CHD	Coronary heart disease among adults aged ≥18 years
CASTHMA	Current asthma prevalence among adults aged ≥18 years
DIABETES	Diagnosed diabetes among adults aged ≥18 Years
BPHIGH	High blood pressure among adults aged ≥18 years
MHLTH	: Mental health not good for ≥14 days among adults aged ≥18 years
CANCER	Cancer (excluding skin cancer) among adults aged ≥18 Years
PHLTH	Physical health not good for ≥14 days among adults aged ≥18 Years
HIGHCHOL	High cholesterol among adults aged ≥18 years who have been screened in the past 5 years
TEETHLOST	All teeth lost among adults aged ≥65 years
KIDNEY	Chronic kidney disease among adults aged ≥18 Years
STROKE	Stroke among adults aged ≥18 years
COPD	Chronic obstructive pulmonary disease among adults aged ≥18 Years

2. Distribution of the variables

The data in the study were summarized data since raw data were not available, and units of the variables are percentage. I used the density plots to check the distributions of the variables and see if the distributions differ significantly over years.

It can be seen that all density plots are uni-modal and about symmetric. For example, the following is the density plot of the variable ARTHRITIS for the year

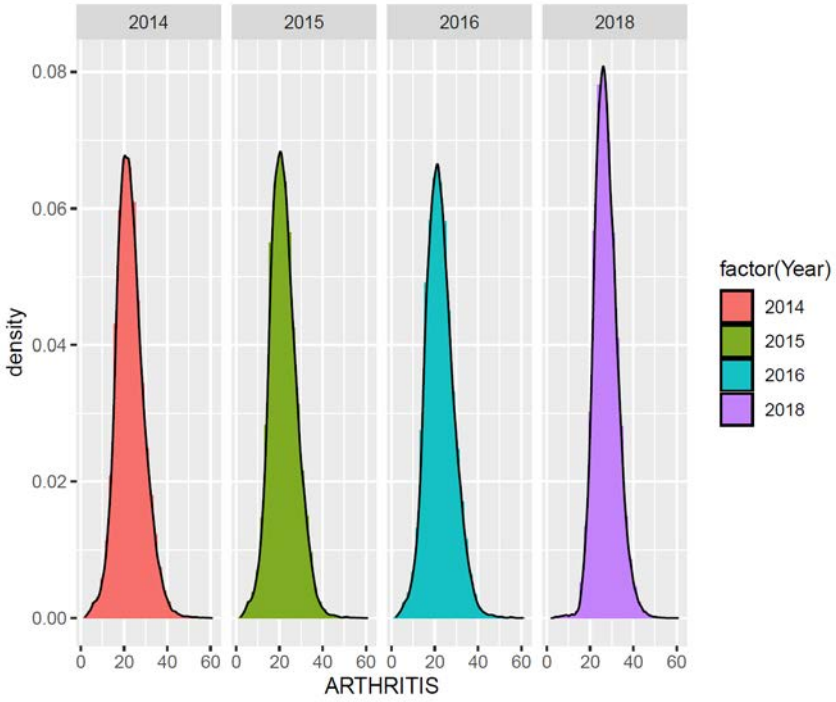

Figure 2: Density plots of the variable ARTHRITIS over years 2014 - 2018

Table 2: Description of Clinical Preventive Service Use Variables in PLACES

Name	Description
ACCESS2	Current lack of health insurance among adults aged 18–64 years
MAMMOUSE	Mammography use among women aged 50-74 years
CHECKUP	Visits to doctor for routine checkup within the past year among adults aged ≥18 years
PAPTEST	Papanicolaou smear use among adult women aged 21-65 years
DENTAL	Visits to dentist or dental clinic among adults aged ≥18 years
COLON_SCREEN	Fecal occult blood test, sigmoidoscopy, or colonoscopy among adults aged 50–75 years
BPMED	Taking medicine for high blood pressure control among adults aged ≥18 Years with high blood pressure
COREM	Older adult men aged ≥65 Years who are up to date on a core set of clinical preventive services
COREW	Older adult women aged ≥65 Years who are up to date on a core set of clinical preventive services
CHOLSCREEN	Cholesterol screening among adults aged ≥18 years

2014, 2015, 2016 and 2018. Because the distributions of other variables show similar patterns, the other graphs are not included in this report.

3. Correlation Investigations

To study the correlation among the variables, I considered city level data only. First, the correlation among the 13 health outcomes, 10 use of preventive services, and 5

Table 3: Description of Unhealthy Behaviors Variables in PLACES

Name	Description
BINGE	Binge drinking among adults aged ≥18 years
OBESITY	Obesity among adults aged ≥18 years
CSMOKING	Current smoking among adults aged ≥18 years
SLEEP	Sleeping less than 7 hours among adults aged ≥18 years
LPA	No leisure-time physical activity among adults aged ≥18 years

chronic disease-related unhealthy behaviors are investigated. Second, I present the correlations between the prevention and health outcome variables, and the correlations between unhealthy behaviors and health outcomes. All the correlations are presented using scatter plot and correlation matrices.

3.1 Correlations Among the Variables

The 13 health outcome variables are denoted by $V1-V13$, the 10 use of preventive services denoted by $V14-V22$, and 5 chronic disease-related unhealthy behaviors are denoted by $V23-V27$ for better clarity of the graphs. Furthermore, the correlations do not differ much over those years. So graphs for the year 2016 only are listed.

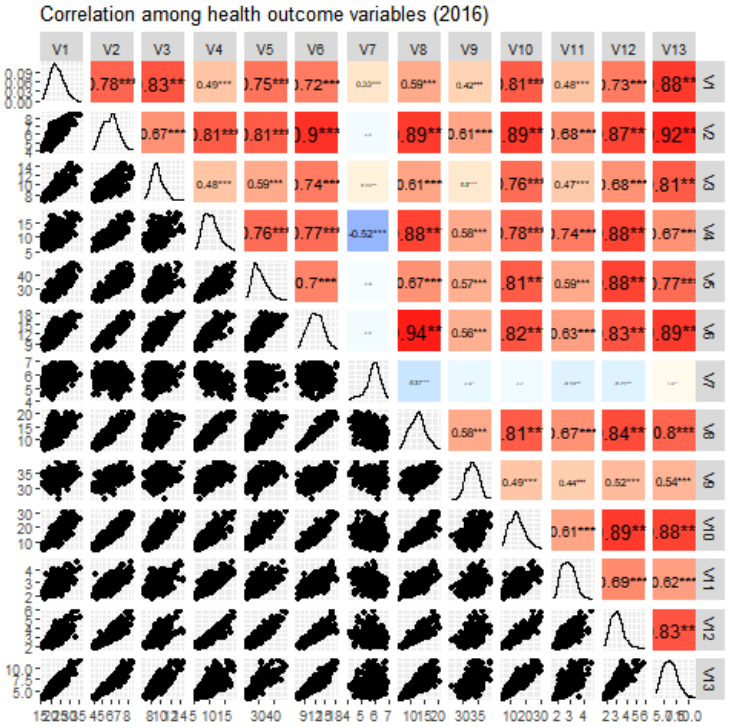

Figure 3: Scatter plot and correlation matrix among the 13 health outcome variables in 2016

It can be seen from Figure 3 that most health outcome variables are strongly positively correlated except 'CANCER' with 'DIABETES', 'PHLTH', 'HIGHCHOL', 'TEETHLOST', 'KIDNEY', 'STROKE' and 'COPD'. The positive correlations show

that areas with higher rate of one or more health problems generally have higher rate of other health problems.

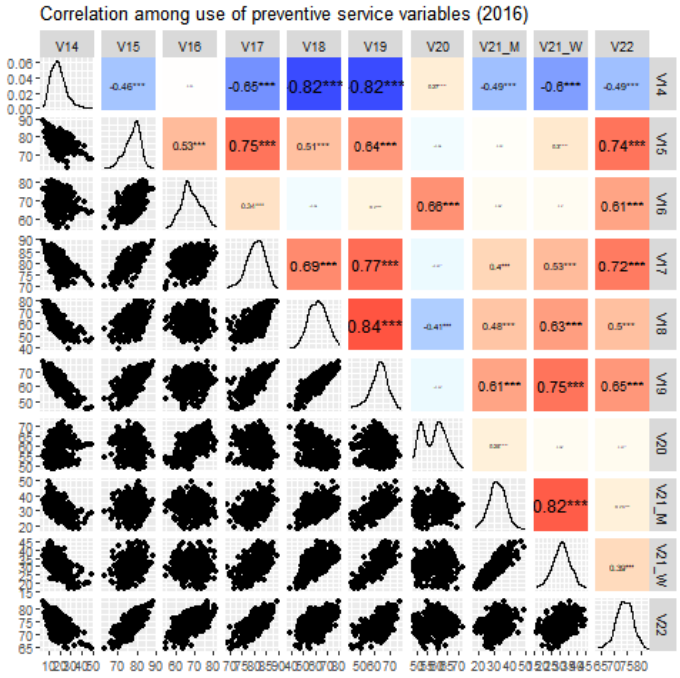

Figure 4: Scatter plot and correlation matrix among the 10 use of preventive services in 2016

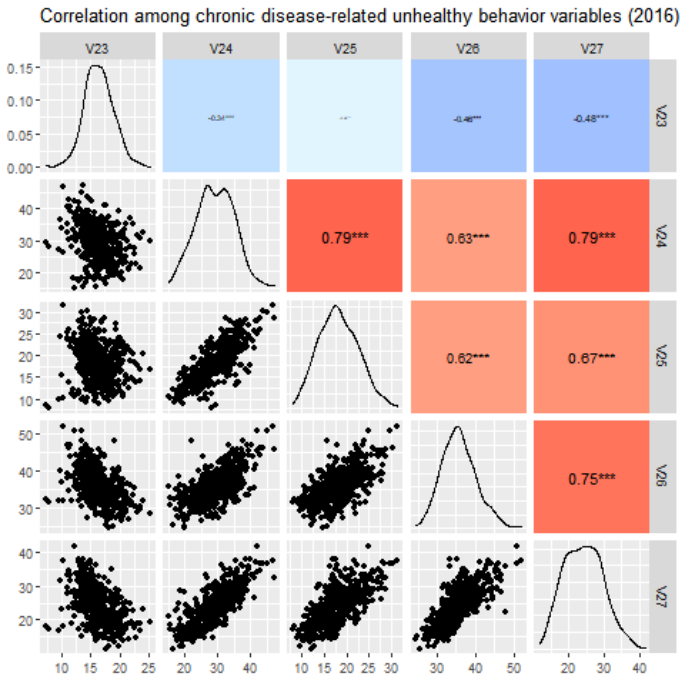

Figure 5: Scatter plot and correlation matrix among the 5 chronic disease-related unhealthy behaviors in 2016

Figure 4 shows that 'ACCESS2' (Current lack of health insurance among adults aged 18–64 years) are negatively correlated with other prevention variables. The negative correlations may result from that people without health insurance in general are unwilling to seek preventive heath services. All other prevention variables are positively correlated, except that 'DENTAL' (Visits to dentist or dental clinic among adults aged $\geq$ 18 Years) and 'BPMED'(Taking medicine for high blood pressure control among adults aged $\geq$ 18 Years with high blood pressure).

Figure 5 shows that all variable are positively correlated except that 'BINGE' (Binge drinking among adults aged $\geq$ 18 Years). The positive correlations show that areas with one of the unhealthy behaviors 'OBESIT', 'CSMOKING', 'SLEEP' and 'LPA' usually have more unhealthy behaviors. It is hard to explain why 'BINGE' is negatively correlated with other unhealthy behaviors. Maybe only healthy people tend to do binge drinking.

3.2 Correlations Between the Variables

It is of interest to check the correlations between prevention and health outcome variables, and the correlations between unhealthy behaviors and health outcomes. In this study, the city data from the year 2016, 2017 and 2018 are combined.

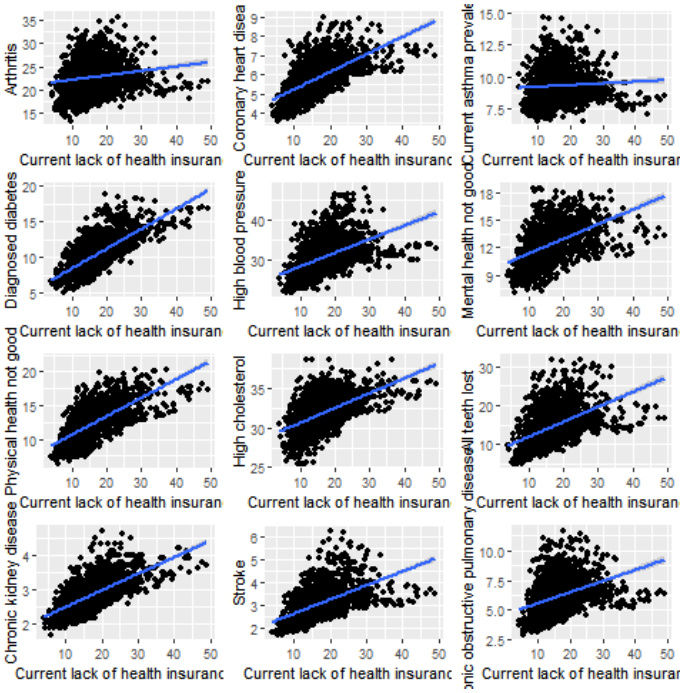

Figure 6: Scatter plot matrix between ACCESS2 and 12 health outcome variables.

It can be seen from Figure 6 that ACCESS2 (Current lack of health insurance among adults aged 18–64 years) is positively correlated with all the health outcome variables except with CANCER (see Figure 7). That is, areas with less health insurance coverage tend to have more health problems except cancer. The negative correlation may due to that patients with cancer in general have health insurance.

It is not surprising to see that CHECKUP (Visits to doctor for routine checkup within the past Year among adults aged $\geq$ 18 Years) is positively correlated with most health outcome variables and most correlations are strong as shown in Figure 8

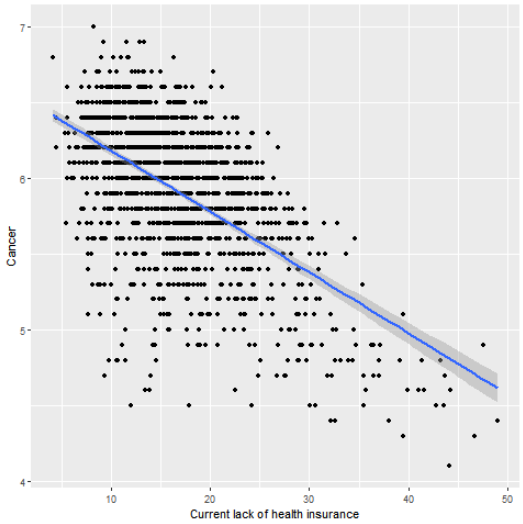

Figure 7: Scatter plot matrix between ACCESS2 and CANCER.

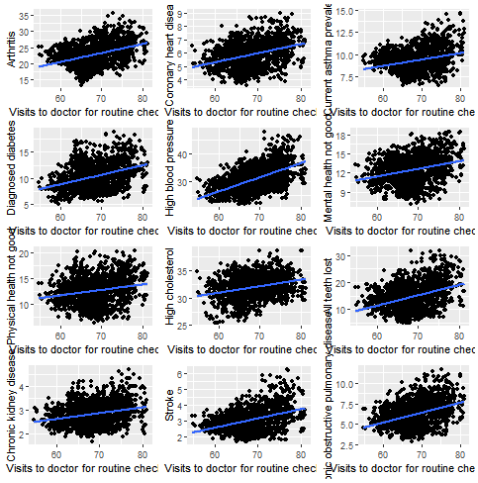

Figure 8: Scatter plot matrix between CHECKUP and 12 health outcome variables.

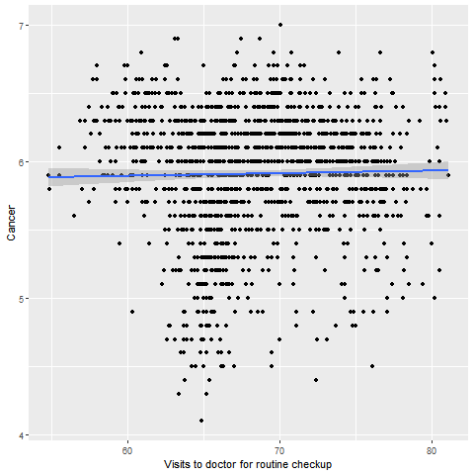

Figure 9: Scatter plot matrix between CHECKUP and CANCER.

because ACCESS2 is negatively correlated with other prevention variables. Again, one exception is the correlation between CHECKUP and CANCER shown in Figure 9. The correlation is positive but very weak. The correlations between health outcome variables and the other prevention variables are similar which are skipped here.

Another interesting investigation is to check the correlations between unhealthy behaviors and health outcome variables. All unhealthy behaviors except Binge drinking are positively correlated to health outcomes. For example, OBESITY are strongly correlated with most health outcome variables except CANCER as shown in Figure 10 and Figure 11. It means that unhealthy behaviors in general are positively correlated with health problems.

It is surprising to see from Figure 12 that Binge drinking and the other 12 health outcome variables have very strong negative correlations. That is, areas with more

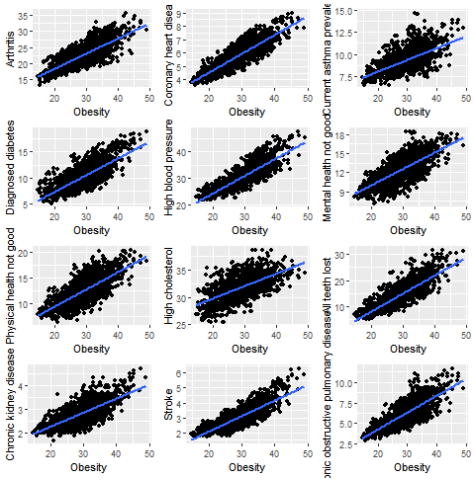

Figure 10: Scatter plot matrix between Binger drinking and 12 health outcome variables.

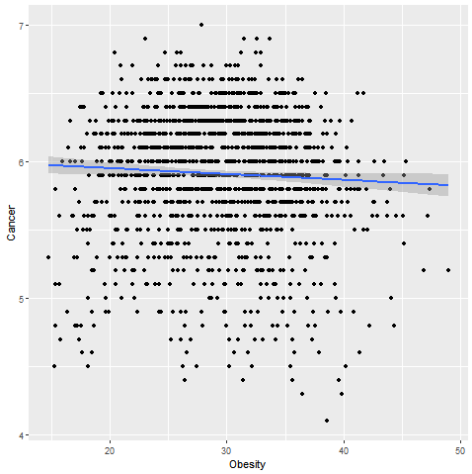

Figure 11: Scatter plot matrix between Binger drinking and CANCER.

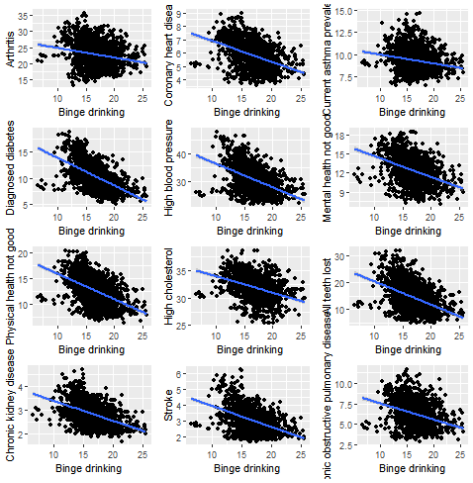

Figure 12: Scatter plot matrix between Binger drinking and 12 health outcome variables.

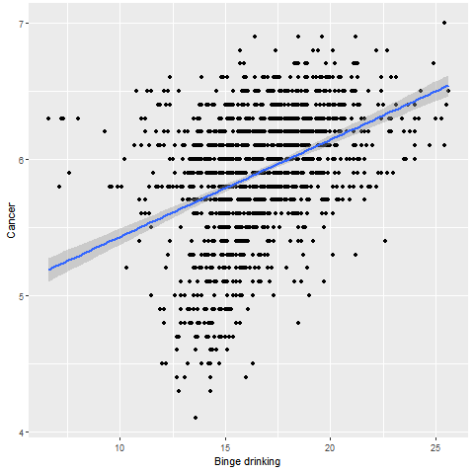

Figure 13: Scatter plot matrix between Binger drinking and CANCER.

rates of Binge drinking generally have less health problems. One explanation may be that only health people do Binge drinking. Again, one exception is the correlation between Binge drinking and CANCER which is positive as shown in Figure 13.

4. Analysis of State-level Data

The state-level data can be estimated using the weighted average of city-level data in each respective state. Figure 14 shows the distribution of the first 8 health outcome variables, 4 prevention variables and 4 unhealthy behavior variables in year 2018. Some patterns can be seen for each variable clearly. For example, the rate of ARTHRITIS, CHD, DIABETES, and PHLTH in Eastern US is much higher than the rates of other regions; Texas has higher rate of ACCESS2 (Current lack of health insurance among adults aged 18–64 years); and Northern US has higher rate of Binge drinking etc. The maps of the distribution for other variables and other

years are not included in this report. There is a region pattern in general for each variable which may be due to different types of economy and culture.

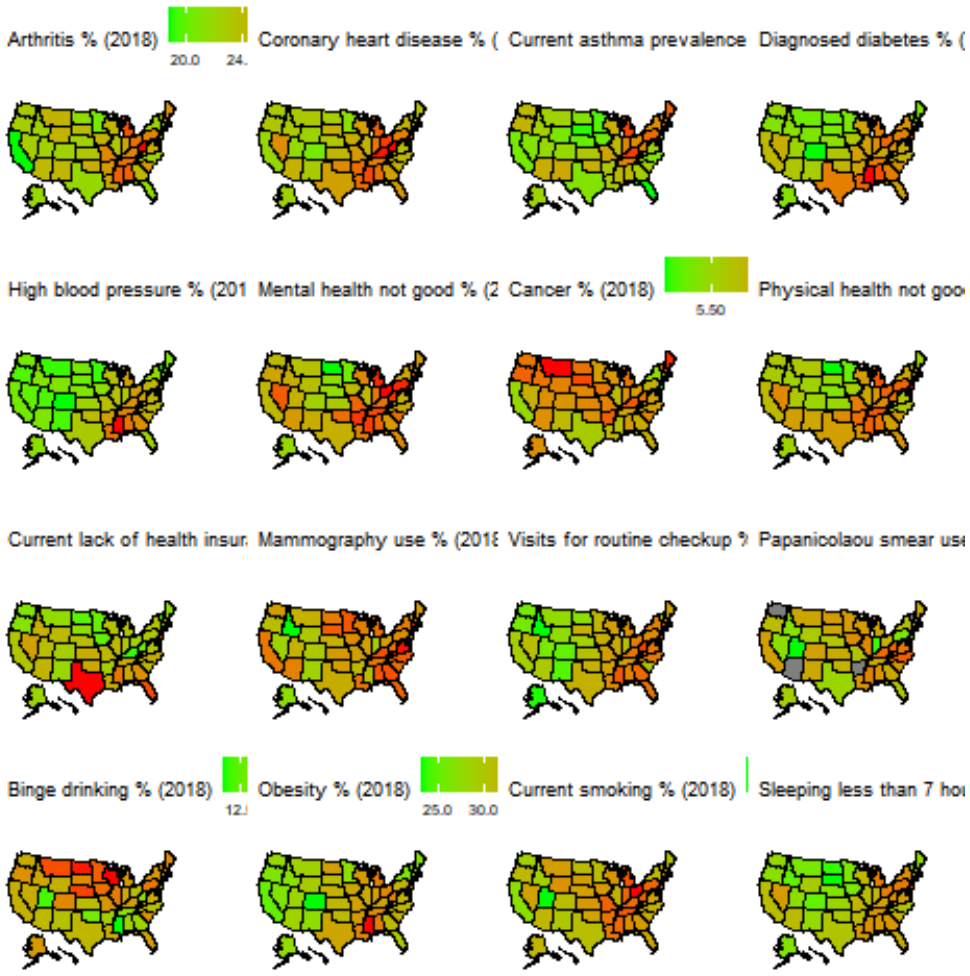

Figure 14: Distribution of state-level data for 8 health outcome variables, 4 prevention variables and 4 unhealthy behaviors.

Another interest is to see how the distribution of the variables vary over the years. Figure 15 lists the time series of the first 8 health outcome variables, 4 prevention variables and 4 unhealthy behavior variables. Other graphs are omitted here. The marginal model (Zeger and Liang, 1992) for longitudinal data analysis is used to check the overall pattern which are displayed by bold blue curves in the graphs. Fluctuations over time can be seen. Especially concerns are on that the rates of health problems and unhealthy behaviors generally did not decrease in the past several years.

5. Summary

This article reports the exploratory data analysis of the project PLACES: Local Data for Better Health. Data visualizations of correlation and regression analysis, longitudinal data analysis, and distribution of the regional differences of the variables are conducted. All data analysis in this report are conducted using software

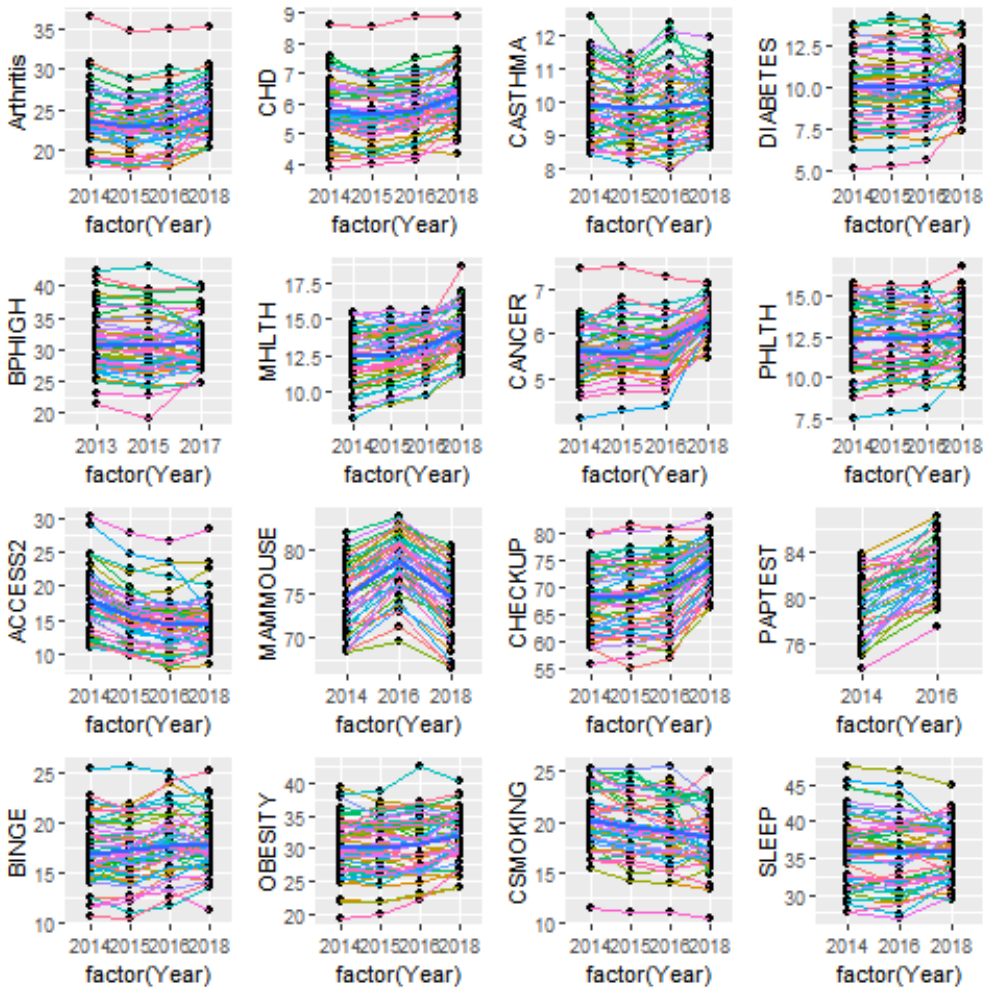

Figure 15: Distribution of state-level data for 8 health outcome variables, 4 prevention variables and 4 unhealthy behaviors.

R (R Core Team, 2021) and some R data manipulation packages like stringr (Wickham, 2019) and dplyr (Wickham, Francois, Henry and Muller, 2021), and data visualization packages including usmap (Lorenzo, 2021), ggplot2 (Wickham, 2016) and gridExtra (Auguie and Antonov, 2017).

REFERENCES

Auguie, B. and Antonov, A. (2017), "gridExtra: Miscellaneous Functions for "Grid" Graphics," `https://cran.r-project.org/web/packages/gridExtra`

Lorenzo, P.D. (2021), "usmap: US Maps Including Alaska and Hawaii," `https://cran.r-project.org/web/packages/usmap`

R Core Team (2021), "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, `https://www.R-project.org/`.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, New York: Springer-Verlag.

Wickham, H. (2019), "stringr: A Simple, Consistent Wrappers for Common String Operations," `https://cran.r-project.org/web/packages/stringr`

Wickham, H., Francois, R., Henry, L. and Muller, K. (2021), "dplyr: A Grammar of Data Manipulation," `https://cran.r-project.org/web/packages/dplyr`

Zeger SL, Liang KY. (1992), "An Overview of Methods for the Analysis of Longitudinal Data," *Statistics in Medicine* ,11,1825–1839.