Summary of 500 Cities Health Data Analysis Project

By Team Combat

Venkata Naga Sai Sriram Akella

Venkata Sravani Kakaraparthi

Yanhe Wu

Narahari Sundaragopalan

Audience Analysis:

Our 500 Cities Project aims at focusing on two main audience which includes Mr. Jacob Williams, the Chief Medical Officer and the researchers from the Center for Disease Control and Prevention.

Primary audience: Mr. Jacob Williams, Chief Medical Officer(CMO)

This project is primarily intended for Mr. Jacob Williams who is a senior government official designated as head of medical services. The CMO will serve to advise and lead a team of medical experts on matters of public health importance. Mr. Jacob Williams is a Doctor/Ph.D. and is associated with CHI Hospital group. He is involved in designing medical policies and assists in maintenance of health standards. His main interests lie in research on methods to prevent diseases, study the latest trends and make predictions, ensure effective medical service by providing training. He is also involved in promoting public seminars.

Mr. Williams also has excellent executive leadership skills, communication skills and problems solving skills. He understands how to read data analysis report and create strategic plan base on the health report outcomes. He is more concerned with current health awareness programs and campaigns. The major challenges as viewed by Mr. Williams include predicting the trends of diseases and ensure effective health measures and medical services are made available to all states of USA. The current analysis would help Mr. Williams in understanding the health status and developing effective measures for preventive services. Our analysis of the dataset in hand, is from “500 cities of United States” and the project objective is to provide professional analysis to assist Mr. Williams in planning public health intervention.

Secondary audience: Researchers from Center for Disease Control and Prevention / Physician.

With high education level on disease control and prevention, the secondary audience has encyclopedic medical knowledge that can be recalled at a moment’s notice. Usually they work as

a team member in a team of medical experts and provide specific solution on disease control and prevention. They possess high intelligence, inquisitiveness and have very good communication skill as well as medical problems solving skill. From the analysis, they will need our project outcome for deeper investigation on the health risks, such as why a certain health issue or dieases occurs at a higher rate in some areas, and what innovative solution can be provided to keep people safe.

Brief description of the data source

It's the complete dataset for the 500 Cities project, available at 500 cities data with 21 variables and 810103 observations, it includes 2013, 2014 model-based small area estimates for 27 measures of chronic disease related to unhealthy behaviors (5), health outcomes (13), and use of preventive services (9). It also includes estimates for approximately 28,000 census tracts within 500 largest US cities. After our data cleaning process, we are using the dataset with respect to the year 2013 for data analysis. Dataset includes 2013 model-based small area estimates for 4 measures of chronic disease related to health outcome and prevention with 13 variables and 116024 observations. This dataset is significant to identity emerging health problems and provide information for disease prevention activities.

Variables include: Year, StateDesc, CityName, GeographicLevel, Category, Measure, Data value Type, Data Value, Data_Value (in%), Data Value Footnote, Population Count, CategoryID, MeasureID, Region

The variables Region, Population_Count, Data_Value (in %) are of importance.

Categories: Health outcomes, Prevention Measures.

Measures of each Category:

A. Health outcome

a. High blood pressure among adults aged equal and larger than 18 Years

b. High cholesterol among adults aged equal and larger than 18 Years who have been screened in the past 5 Years

B. Prevention Measures

a. Cholesterol screening among adults aged equal and larger than 18 Years

b. Taking medicine for high blood pressure control among adults aged equal and larger than 18 Years with high blood pressure

Data analysis processing

We consider our analysis in four steps to summarize the processing of data analysis:

Inspection, Data Cleaning, Data Analysis and Data Visualization. Firstly, inspection of data. In this process, our main goal was to understand the dataset and query the data to meet end user data requirements. Then we have defined the research questions that the dataset may be able to answer through brainstorm. Secondly, cleaning of the raw dataset. Maintaining excellent quality data is essential to ensure the data reliability and deliver accuracy in the data analysis process. During this process, we have encountered four issues with the data:

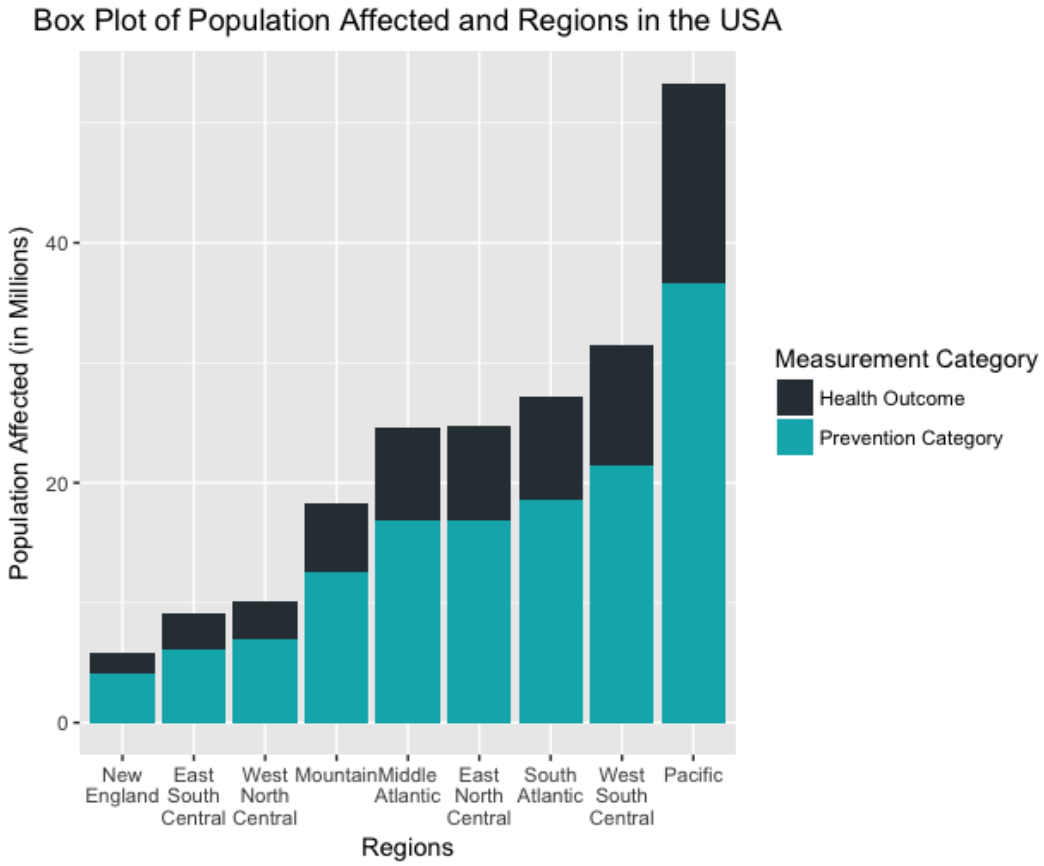
1.Data validity and relativeness,
2.Missing values, unstandardized data,
3. Irrelevant data with respect to data cleaning and
4. Visualization goals.

Then we have used Excel and R as tools to solve these 4 issues. Thirdly, analysis of data. This process is to discover the useful information in the dataset. In order to answer the research questions, an R script was developed by applying descriptive analysis method. Last but not the least, data visualization. It's a process of presenting analyzed data in a pictorial and graphical format. We have developed several R plots that shall answer the research questions and helps

better visualizing and decision making based on the outcome of R Scripts. In the following paragraph, you will see the outcome of the data analysis processing.

Plots from Analysis

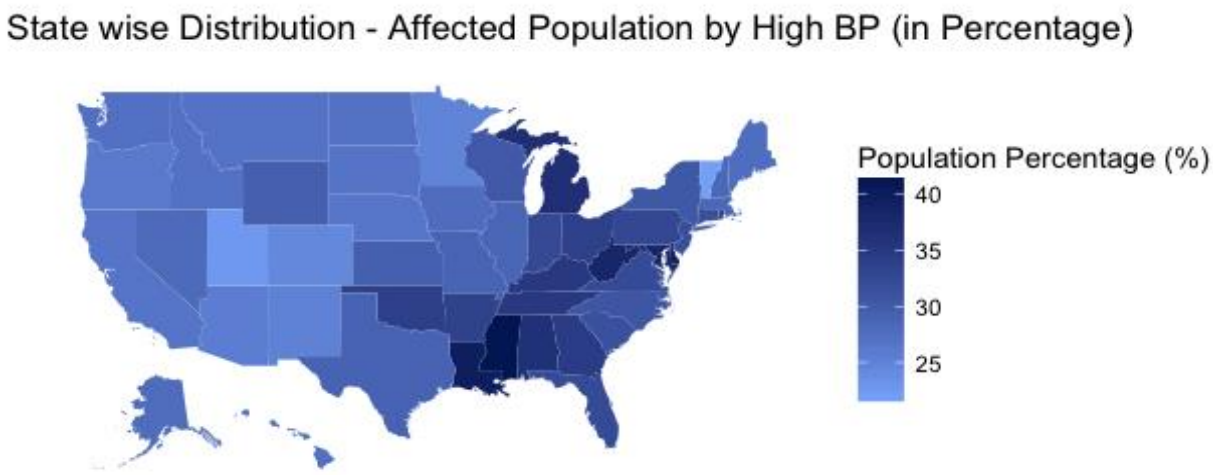
Plot 1 (Population Affected vs Regions)



The above plot intends to provide a visualization on the different measure of Categories by which the 500 Cities Local Health Dataset is based upon. The categories, Health Outcomes and Preventive Measures, are combined to be depicted on a single bar graph divided based on Regions in the USA the visualization depicts that there has been a linear relation with the number of Preventive Measures with respect to Health outcomes in the 9 Regions in the USA. From the above plot, we could know that Pacific region, which include Alaska, California, Hawaii, Oregon, and Washington, has more population who is affected by chronic disease of category of health

outcome and prevention category than other regions. Further, we can also interpret that there is a linear relation between health outcomes and prevention measures for all regions in the USA. That is, for regions with higher health risks and outcomes, the preventive measures are also higher. However, this plot provides an overall view and higher perspective from the regional point of view.

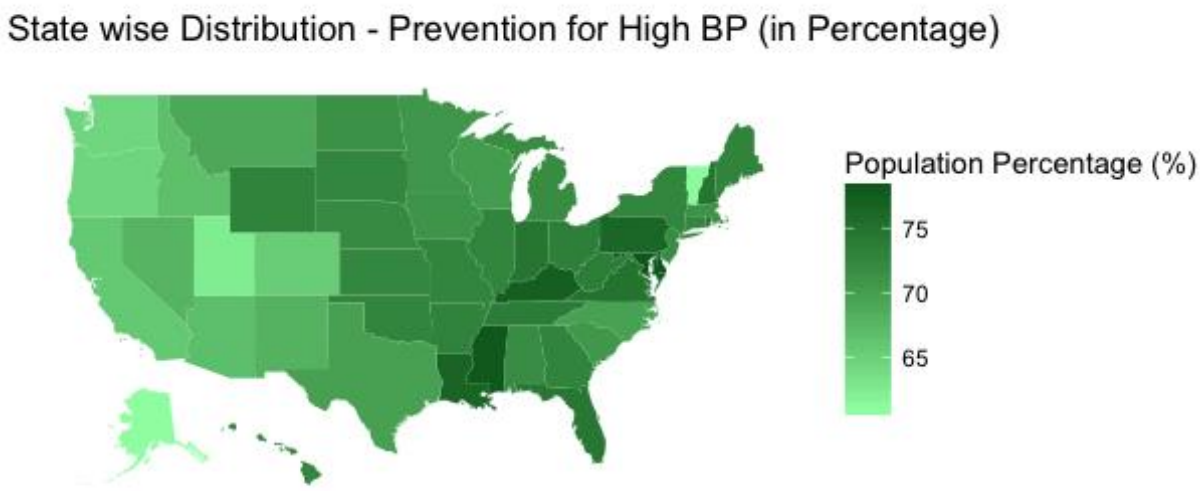
Plot 2 (State wise Distribution – Affected Population by High BP)



The above plot shows a visualization of the High blood pressure among adults aged equal and larger than 18 Years (High BP) measure (Health Outcome) in all the states in the USA. It can be

seen that a state like Mississippi and Louisiana has high population affected by High BP. And state in South and Northeast region has high population affect by high blood pressure.

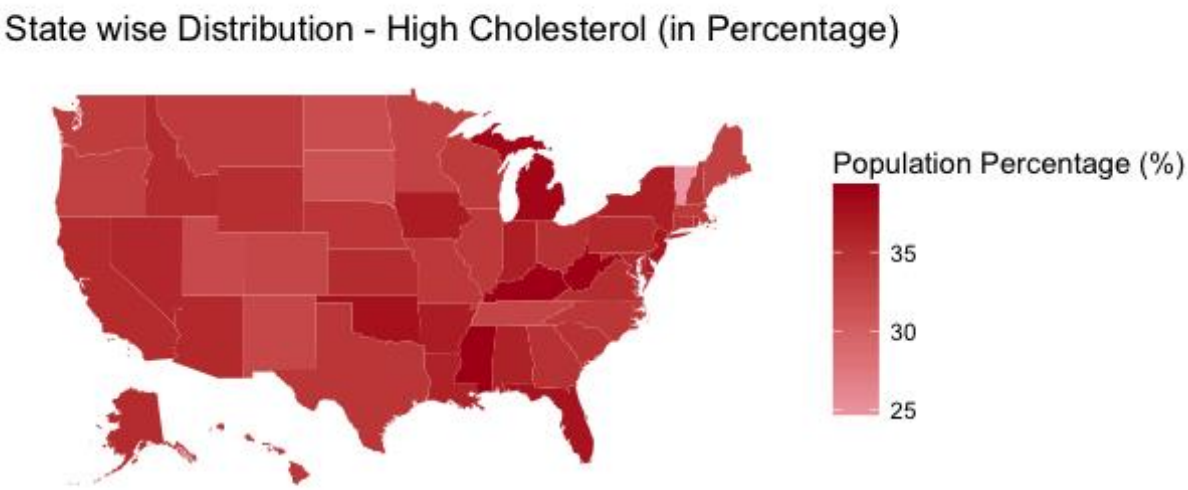
Plot 3 State wise Distribution – Prevention for High BP



The above plot is a measure of preventive methods taken to control population suffering from high blood pressure in various states of USA. This plot is viewed in comparison with plot depicting high blood pressure health outcome in the states of USA. The comparison is summarized below Further, it can be observed that 75% percent of the population affected by high blood pressure are provided with some level of preventive measure in the states part of the southern regions

Also, states with less than 50% of the population affected by high BP have about 25% of the population covered by preventive measures (states part of the Pacific region)

Plot 4 (State wise Distribution – High Cholesterol (in Percentage))

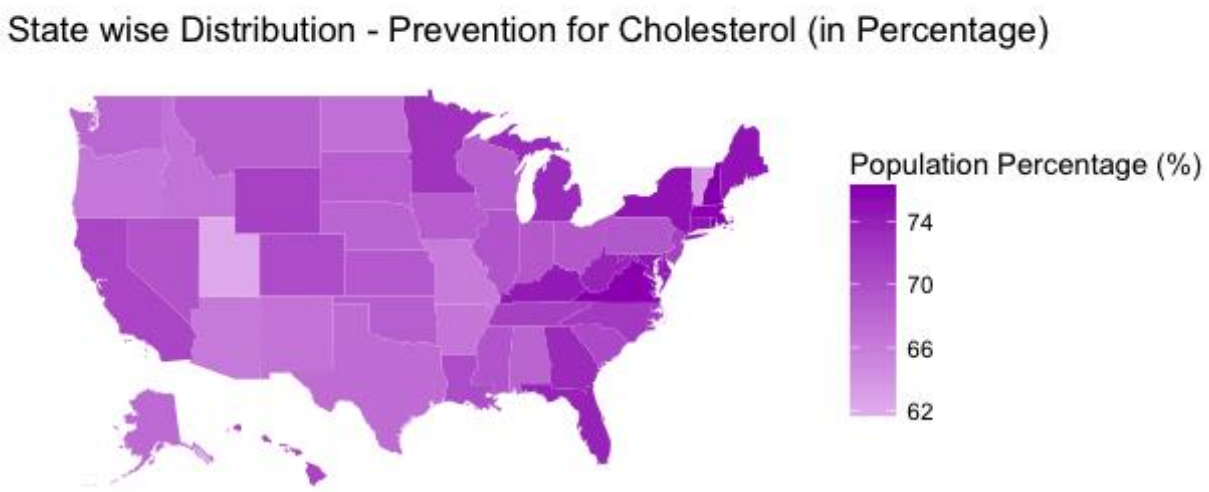


The above plot shows a visualization of the *High cholesterol among adults aged equal and larger than 18 Years who have been screened in the past 5 Years* (High Cholesterol) measure (Health Outcome) in all the states in the USA. It can be seen that a state like California has population affected by High Cholesterol.

Southern Region states have more than 35% population affected by high cholesterol

However, less than 60% of the affected population has received preventive measures for high cholesterol in these states under the southern region

Plot 5 (State wise Distribution – Prevention for Cholesterol)



The above plot shows a visualization of the *Cholesterol screening among adults aged equal and larger than 18 Years* (Preventive Measure) in all the states in the USA. Both states New Hampshire and Virginia represent the deep purple color which means they have the highest percent of population whose Cholesterol screening among adults aged equal and larger than 18 Years.

States with 25 – 30% of population affected by high cholesterol has more than 70% of the population covered by preventive measures
For example, the state of Iowa in which about 35% of the population is affected by high cholesterol, around 65% of the affected population have received preventive measure
But the state of Maine which less than 30% of the affected population, has more than 70% of the population covered

Persuasive Argument

Based on our analysis, we believe that there are still many states which are not completely covered with preventive measures for the health risks faced by them. The population affected by high blood pressure is better covered with measures, than the states facing high cholesterol. For population affected by high cholesterol, more preventive measures must be taken, covering more states.
We recommend that CMO should create strategy plan to solve the problem of High Cholesterol in various states And CMO should also lead their team to investigate why some states with lower levels of high cholesterol have better coverage of preventive measures, than other states with higher level of high cholesterol affected population.

Word Count: 1579 words

References:

1. 500 Cities: Local Data for Better Health. (2016, December 07). Retrieved October 14, 2017, from https://catalog.data.gov/dataset/500-cities-local-data-for-better-health-b32fd
2. 500 Cities: Local Data for Better Health. (2016, December 12). Retrieved December 04, 2017, from https://www.cdc.gov/500cities/definitions/prevention.htm
3. Breakdown of states to region done based on - https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf