# Behavioral Data Mining

# On

# 500 Cities Local Health Data Set

# 

# Under the esteemed guidance of Dr. Zhengxin Chen

# And support from Dr. Yong Shi

#### Venkata Naga Sai Sriram Akella

#### Graduate Student – Management Information Systems

#### University of Nebraska at Omaha

#### Brief Summary of Overall Process

#### The project primarily focuses on the concept of Behavioral Data Mining – which includes understanding of how a business process work based on data and how it is used to predict, understand and analyze pattern from the introspection and mining inti the data. This theoretical approach is sufficed and replicated through a practical approach considering the 500 Cities Local Health dataset.

#### The 500 Cities dataset is considered where the data, business terms, the scope of the data mining variables were initially understood. This is followed by the cleaning of data through Excel and R. The detailed process is described below. Later through SQL Analysis of data in MS SSAS, I have understood the scope and data introspection limits through which I have worked on Web Crawling process to understand and collect similar web content related to variables like Health Outcomes and Prevention measures. To confirm this, I have performed Correlation analysis, that provided a positive relation between the Health Outcomes and Prevention. Where I have included Region variable in Data Cleaning process that enhanced the scope to understand the data. While the visualization of the data is done through R Plots ( with the help of ggplot, Boxplot and Linear regression in Excel.

#### Then I have performed Data Mining through WEKA where I have used Clustering, Classification analysis to produce confusion matrix and variables that impacted the Population Count of a particular region based on Health Outcomes ad Preventive Measures.

#### This behavioral data mining shall help the end-users/ Target Audience including data enthusiasts, Government Health Officers and the US Government in regulating the health care policies and benefits to the people in the particular location.

#### Link for the entire Project is available at:

#### <https://github.com/Sriramakella123/DataMiningProject>

### Data Understanding and Analysis

### Task-1 Description of the data source with citation

This is a complete dataset for the 500 Cities project, available from data.gov.

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. It includes 21 variables and 810103 observations.

Currently, we are using the dataset with respect to the year 2013 and have separated the original dataset containing years 2013 and 2014 into the 2013 and has 116025 observations and 13 variables that need to be cleaned for better data evaluation and analysis. Due to the reason that we are not doing census tract analysis based on the population count, the variable Low confidence limit, High confidence limit have been removed.

Variables include:

\* Year

\* StateDesc

\* CityName

\* CategoryLevel

\* Measure

\* Data value Type

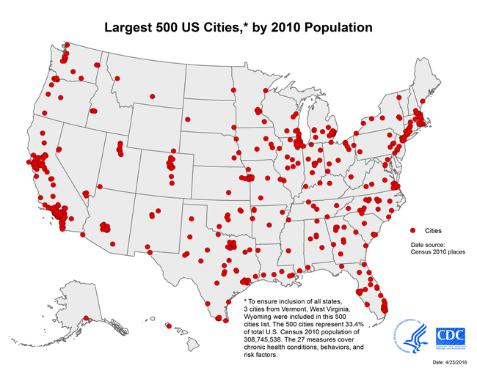
\* Data Value

\* Data Value Footnote

\* Population Count

This dataset is significant to identity emerging health problems and provide information for disease prevention activities.

**Note**: I tend to choose the dataset corresponding to the year 2013 and then try to document the cleaning of data.



**Reference and Citation for Data Source**

500 Cities: Local Data for Better Health. (2016, December 07). Retrieved April 13, 2018, from https://catalog.data.gov/dataset/500-cities-local-data-for-better-health-b32fd

### Task-2 Identifying any intellectual policy constraints (licensing)

This is an open database intended for public use. Open Database License(ODbL). The details of its usage and availability are as follows:

The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Database while maintaining this same freedom for others. The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Database while maintaining this same freedom for others. Many databases are covered by copyright, and therefore this document licenses these rights. Some jurisdictions, mainly in the European Union, have specific rights that cover databases, and so the ODbL addresses these rights, too. Finally, the ODbL is also an agreement in contract for users of this Database to act in certain ways in return for accessing this Database.

Databases can contain a wide variety of types of content (images, audiovisual material, and sounds all in the same database, for example), and so the ODbL only governs the rights over the database, and not the contents of the Database individually. Licensors should use the ODbL together with another license for the contents, if the contents have a single set of rights that uniformly covers all of the contents. If the contents have multiple sets of different rights, Licensors should describe what rights govern what contents together in the individual record or in some other way that clarifies what rights apply.

Sometimes the contents of a database, or the database itself, can be covered by other rights not addressed here (such as private contracts, trade mark over the name, or privacy rights / data protection rights over information in the contents), and so you are advised that you may have to consult other documents or clear other rights before doing activities not covered by this license. More information is available at <https://opendatacommons.org/licenses/odbl/1.0/>

### Task-3 Description of Metadata and Purpose of Project

1. This is the complete dataset for the 500 Cities project. This dataset 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). Data were provided by the Centers for Disease Control and Prevention (CDC), Division of Population Health, Epidemiology and Surveillance Branch. The project was funded by the Robert Wood Johnson Foundation (RWJF) in conjunction with the CDC Foundation.
2. It represents a first-of-its kind effort to release information on a large scale for cities and for small areas within those cities. It includes estimates for the 500 largest US cities and approximately 28,000 census tracts within these cities. These estimates can be used to identify emerging health problems and to inform development and implementation of effective, targeted public health prevention activities.
3. Because the small area model cannot detect effects due to local interventions, users are cautioned against using these estimates for program or policy evaluations. Data sources used to generate these measures include Behavioral Risk Factor Surveillance System (BRFSS) data (2013, 2014), Census Bureau 2010 census population data, and American Community Survey (ACS) 2009-2013, 2010-2014 estimates. More information about the methodology can be found at www.cdc.gov/500cities.

**Purpose of the Project**

1. “This project reports city and census tract-level data, obtained using small area estimation methods, for 27 chronic disease measures for the 500 largest American cities.
2. The data are published through a public, interactive “500 Cities” website that allows users to view, explore, and download city- and tract-level data.
3. Although limited data are available at the county and metropolitan levels, this project represents a first-of-its kind data analysis to release information on a large scale for cities and for small areas within cities. This system complements existing surveillance data necessary to more fully understand the health issues affecting the residents of that city or census tract.

These high-quality, small-area epidemiologic data can be used both by individual cities and groups of cities as well as other stakeholders to help develop and implement effective and targeted prevention activities; identify emerging health problems; and establish and monitor key health objectives. For example, city planners and elected officials may want to use this data to target neighborhoods with high rates of smoking or other health risk behaviors for effective interventions. “(n.d.500 Cities: Local Data for Better Health.)

**Reference**:

500 Cities: Local Data for Better Health. (2016, December 07). Retrieved April 14, 2018, from https://www.cdc.gov/500cities/about.htm

### Task-4 Issues encountered with data

1. **Data validity and relativeness** - The data being collected for understanding the health of 500 cities in Unites States is limited to entire United States, respective city and census tract levels to understand the Health Outcomes and Prevention categories. Assuming the data with respect to 2013, the parameters are hence analyzed.
2. **Missing Values** - In the column CityName , DataValue , Data\_Value\_Footnote , PopulationCount has data range values having ‘Blanks’ that need to be filled with meaningful and reasonable data values that add meaning and value to the dataset and helps in efficient understanding of data.
3. **Unstandardized data** - The column UniqueID has data values ‘59’, ‘107000’ , ‘0107000-01073000100’ etc. which doesn’t follow a specified pattern or a clarity in naming the ID or providing definition for the ID. Moreover, the value or ID details is not necessary. These need to be fixed or deleted based on the mode of data cleaning and the goal of cleansing the data.
4. **Irrelevant data with respect to Data cleaning and Visualization Goals** - The column UniqueID is not necessary with respect to evaluation of data or the pre-determined research goals. So we chosse to delete the column. Also, the Data\_Value\_Unit which is in ‘%’ can be deleted .The CityFIPS, TractFIPS, ShortQuestionText can be deleted

### Task-5 Description of Rationale for Data Remediation

1. We have made our data research questions and goals clear and do not tend to choose the Census Data for tracting the Census values, instead we have limited data with respect to year 2013 and also eliminated the FIPS value.

2. Additionally, we have filled the missing data values for CityName , DataValue , Data\_Value\_Footnote each with ‘NA’ - *Not Applicable* value and PopulationCount with ‘Unknown’ as those columns with the blank or unknown datavalue are not considered for any sort of data analysis and doesn’t provide much detail or clarity on visualization. Instead, naming them appropriately as NA or Unknown helps in rethinking in data collection and analysis steps.

3.Also we have deleted the irrelevant or unnecessary column and double-checked for data duplicities and their existence and ensure their validity with respect to the data evaluation. This also involved redefining and understanding the end-goals in the data cleaning thereby we have deleted the unstandardized data.

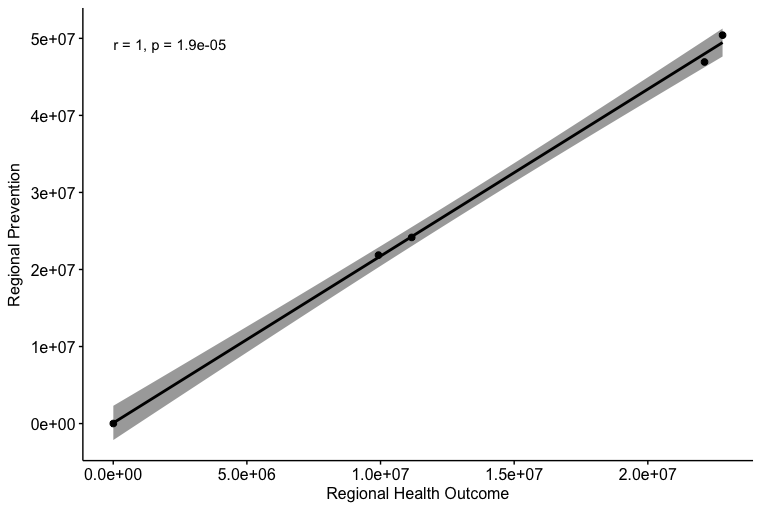
### Task-6 Step-by-step Description of Data Cleaning Process for Replication

* Below is the sequence of steps followed in cleaning the data:

As part of cleaning, I have used R script and that helped in cleaning the data with respect to missing values, data anomalies, unstandardized values and entity matching etc.

1. To get the current working directory: **getwd()**
2. To set the working directory to the desired location: **setwd("C:/Users/sriram/Downloads")**
3. Loading the desired files to the data frame "data": **data<-read.csv("500CitiesLocalDataSetForBetterHealth\_CleanedExcel.csv", header = T , na.strings = c("", "NA"))**
4. To view the loaded data frame : **view(data)**
5. To remove the unwanted columns from the loaded data frame and making a new data frame: **clean\_data<-subset(data, select = -c(6,8,11,12,13,18,19,20))**
6. To view statistics relating the data: **summary(data)**
7. Changing the "NA" values in "PopulationCount" Column to "Unknown" : **clean\_data[["PopulationCount"]][is.na(clean\_data[["PopulationCount"]])] <- "Unknown"**
8. To view the top 6 rows of our working dataframe: **head(clean\_data)**
9. To view the entire dataframe: **View(clean\_data)**
10. To print the required table with set of arguments listed above for clean data:**write.csv(clean\_data, "500CitiesLocalDatasetForBetterHealth\_CleanedExcelv2.csv")**

**Linear Regression Analysis**



Pearson Moment- Correlation analysis as well as Linear regression analysis have confirmed that Region based Prevention and Health Outcome are positively correlated and is shown from the graph as above.

### RPlot - Documentation

This document also explains the steps in creating an R Plot for data visualization for the Research Questions raised in the previous assignments for 500 Local Cities Health Dataset. The R Plot is created based on the R script generated previously as part of Data Preparation.

*#Removed unwanted state rowname with the value "USA"*

mergedHealthOutcome\_Prevention <- mergedHealthOutcome\_Prevention[-8,]

*#Create 3 temporary data frames for creating an R Plot*

a <- data.frame(mergedHealthOutcome\_Prevention$Region)

b <- data.frame(mergedHealthOutcome\_Prevention$Population\_Health\_Outcomes)

c <- data.frame(mergedHealthOutcome\_Prevention$Population\_Prevention\_Category)

*#Merge all the data frames into one*

df <- data.frame(a,b,c)

*#Filter the data frame values based on Region*

df <- melt(df, id.vars = "mergedHealthOutcome\_Prevention.Region")

*#Create an R plot using ggplot library*

ggplot(df, aes(x=reorder(mergedHealthOutcome\_Prevention.Region, value),

y=value/1000000, fill=variable)) +

geom\_bar(stat = "identity") +

scale\_fill\_manual(values = c("#303B41", "#00B2B9"),

labels= c("Health Outcome", "Prevention Category")) +

xlab("Regions") +

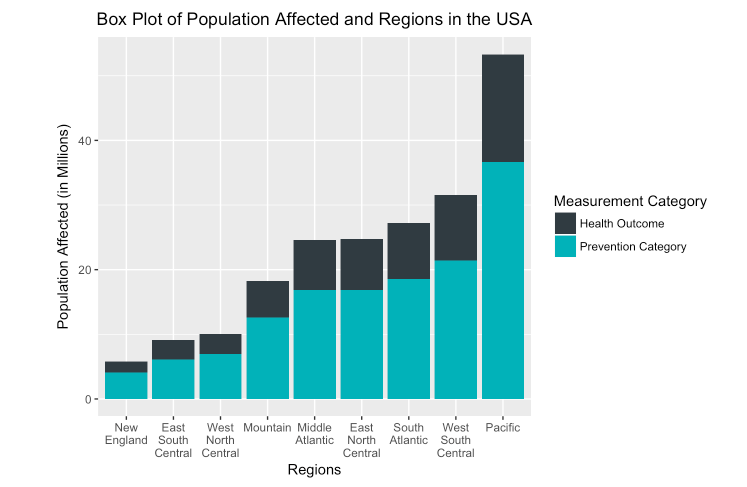
ylab("Population Affected (in Millions)") +

ggtitle("Box Plot of Population Affected and Regions in the USA") +

theme(plot.title = element\_text(hjust = 0.5)) +

scale\_x\_discrete(labels = **function**(x) str\_wrap(x, width=5)) +

labs(fill="Measurement Category")



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 5 Regions in the USA.

*#Create a CSV of the cleaned data and R Script created in the previous assignment*

state\_and\_region <- read.csv("Updated\_US\_States\_Regions\_Health.csv")

*#Create a new data frame creating a relation between Measures and States*

measureStateRelation <- aggregate(list(state\_and\_region$ActualPopulation,

state\_and\_region$PopulationCount),

by=list(Measures = state\_and\_region$MeasureId, State = state\_and\_region$StateDesc), FUN=sum)

*#Rename column name to a meaningful name*

colnames(measureStateRelation)[3] <- "Affected\_Population"

colnames(measureStateRelation)[4] <- "Total\_Population"

measureStateRelation$Percent\_Affected\_Population <-

(measureStateRelation$Affected\_Population /

measureStateRelation$Total\_Population \* 100)

*#Create a data frame for High BP Health Outcome*

measureBPHigh <- subset(measureStateRelation, Measures == 'BPHIGH')

*#Removed unwanted "USA" column*

measureBPHigh <- measureBPHigh[-45,]

measureBPHigh <- measureBPHigh[,-1]

*#Convert column and rownames to lower case for plotting on US Map*

colnames(measureBPHigh)[1] <- "state"

levels(measureBPHigh$state) <- tolower(levels(measureBPHigh$state))

*#Install required libraries*

**library**(fiftystater)

**library**(mapproj)

## Loading required package: maps

data("fifty\_states")

*#Create a plot to show State wise Distribution of High BP in the USA*

p <- ggplot(measureBPHigh, aes(map\_id = state)) +

geom\_map(aes(fill=Percent\_Affected\_Population),

map = fifty\_states) +

expand\_limits(x = fifty\_states$long, y = fifty\_states$lat) +

coord\_map() + scale\_x\_continuous(breaks=NULL) +

scale\_y\_continuous(breaks=NULL) +

labs(x="", y="") +

theme(panel.background = element\_blank())

*#Provide a title to the R Plot*

p + scale\_fill\_gradient(low="#86b2f9", high = "#002266",

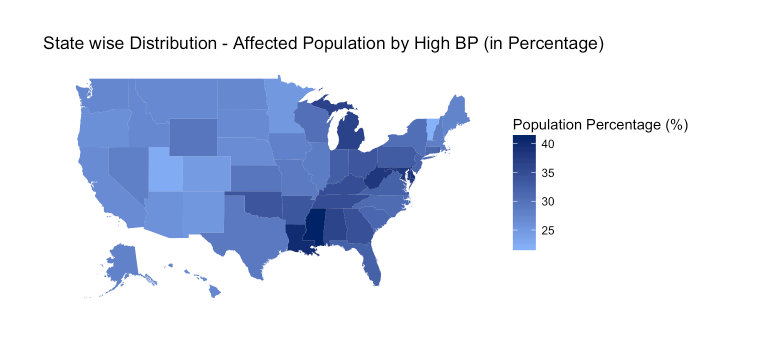
space = "Lab",

guide = "colourbar") +

ggtitle("State wise Distribution - Affected Population by High BP (in Percentage)") +

theme(plot.title = element\_text(hjust = 0.1)) +

labs(fill="Population Percentage (%)")



The above plot shows a visualization of the High BP measure (Health Outcome) in all the states in the USA. It can be seen that a state like California has population affected by High BP

*#Rename column to meaningful name in the main data frame*

colnames(measureStateRelation)[2] <- "state"

*#Create a data frame for BP Prevention Measure*

measureBPMed <- subset(measureStateRelation, Measures == 'BPMED')

measureBPMed <- measureBPMed[,-1]

measureBPMed <- measureBPMed[-45,]

*#Change the state names to lower case for mapping on US map*

levels(measureBPMed$state) <- tolower(levels(measureBPMed$state))

*#Create a R Plot for State Wise Distribution for BP Prevention*

vizMeasureBPMed <- ggplot(measureBPMed, aes(map\_id = state)) +

geom\_map(aes(fill=Percent\_Affected\_Population),

map = fifty\_states) +

expand\_limits(x = fifty\_states$long, y = fifty\_states$lat) +

coord\_map() + scale\_x\_continuous(breaks=NULL) +

scale\_y\_continuous(breaks=NULL) + labs(x="", y="") +

theme(panel.background = element\_blank())

*#Provide a Title and colour to the R Plot*

vizMeasureBPMed + scale\_fill\_gradient(low="#9dfbb0", high = "#006622",

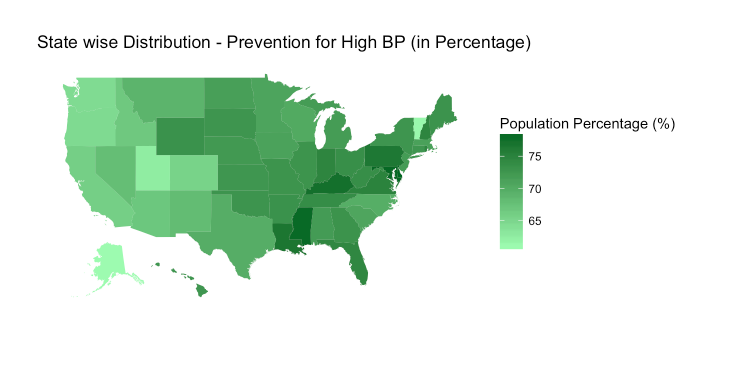
space = "Lab",

guide = "colourbar") +

ggtitle("State wise Distribution - Prevention for High BP (in Percentage)") +

theme(plot.title = element\_text(hjust = 0.1)) +

labs(fill="Population Percentage (%)")



The above plot shows a visualization of a Preventive measure of Medium BP in all the states in the USA. This provides an insight, in comparison with the Health Outcomes of High BP in the previous plot, on how the Preventive measure was carried out.

*#Create a data frame for High Cholesterol Health Outcome in states of USA*

measureHighCholesterol <- subset(measureStateRelation, Measures == 'HIGHCHOL')

measureHighCholesterol <- measureHighCholesterol[,-1]

measureHighCholesterol <- measureHighCholesterol[-45,]

*#Change rownames to lower case for proper mapping*

levels(measureHighCholesterol$state) <- tolower(levels(measureHighCholesterol$state))

*#Create a R Plot for State wise distribution of High Cholesterol Health Outcome in the USA*

vizMeasureHighCholesterol <- ggplot(measureHighCholesterol, aes(map\_id = state)) +

geom\_map(aes(fill=Percent\_Affected\_Population),

map = fifty\_states) +

expand\_limits(x = fifty\_states$long, y = fifty\_states$lat) +

coord\_map() + scale\_x\_continuous(breaks=NULL) +

scale\_y\_continuous(breaks=NULL) + labs(x="", y="") +

theme(panel.background = element\_blank())

*#Provide a title and colour to the R Plot*

vizMeasureHighCholesterol + scale\_fill\_gradient(low="#f1a7b0", high = "#af0418",

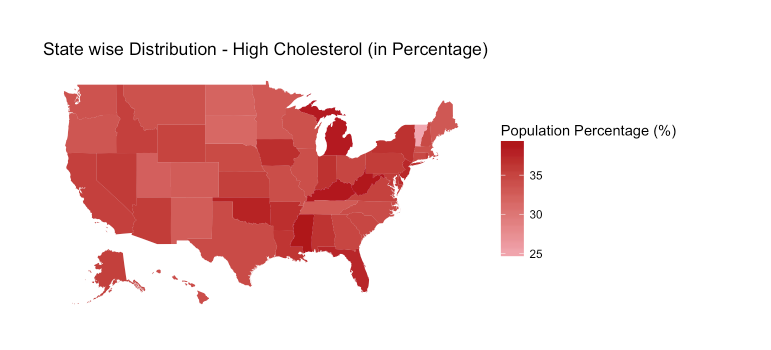
space = "Lab",

guide = "colourbar") +

ggtitle("State wise Distribution - High Cholesterol (in Percentage)") +

theme(plot.title = element\_text(hjust = 0.1)) +

labs(fill="Population Percentage (%)")



The above plot shows a visualization of the 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.

*#Create a data frame for Cholesterol Prevention Measure in USA*

measureCholesterolScreen <- subset(measureStateRelation, Measures == 'CHOLSCREEN')

measureCholesterolScreen <- measureCholesterolScreen[,-1]

measureCholesterolScreen <- measureCholesterolScreen[-45,]

*#Change rownames to lower case for proper mapping*

levels(measureCholesterolScreen$state) <- tolower(levels(measureCholesterolScreen$state))

*#Create R Plot for State wise distribution of Cholesterol Prevention measure*

vizHighCholesterolPrevention <- ggplot(measureCholesterolScreen, aes(map\_id = state)) +

geom\_map(aes(fill=Percent\_Affected\_Population),

map = fifty\_states) +

expand\_limits(x = fifty\_states$long, y = fifty\_states$lat) +

coord\_map() + scale\_x\_continuous(breaks=NULL) +

scale\_y\_continuous(breaks=NULL) + labs(x="", y="") +

theme(panel.background = element\_blank())

*#Provide a title and colour to the R Plot*

vizHighCholesterolPrevention + scale\_fill\_gradient(low="#e6bcf0", high = "#990bbb",

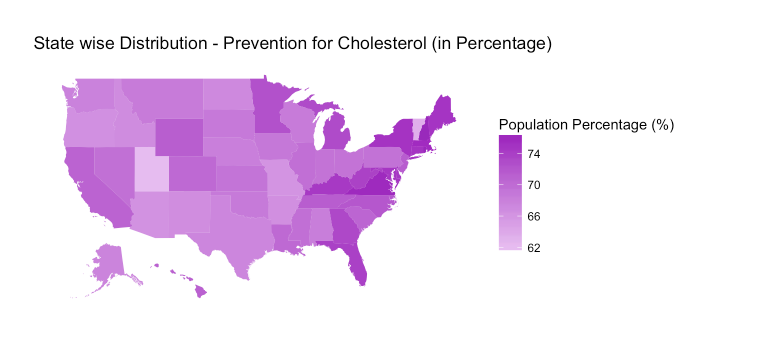
space = "Lab",

guide = "colourbar") +

ggtitle("State wise Distribution - Prevention for Cholesterol (in Percentage)") +

theme(plot.title = element\_text(hjust = 0.1)) +

labs(fill="Population Percentage (%)")



The above plot shows a visualization of the Cholesterol Screening (Preventive Measure) in all the states in the USA.

The dataset has only one quantitative value by which different states or regions of the USA could be compared. Hence the opportunity to use data and create different kinds of visualization was limited.

However, the current visualization created through R Plots provides an insight on how the Health Outcomes and Preventive Measures are distributed across various Regions in the USA and provides a concentration of each Measure in all the states of USA

**R Crawler – Web Crawling for Data Mining**

R Crawler is an R package for web crawling websites and extracting structured data which can be used for a wide range of useful applications, like web mining, text mining, web content mining, and web structure mining. So, what is the difference between Rcrawler and rvest : rvest extracts data from one specific page by navigating through selectors. However, Rcrawler automatically traverses and parse all web pages of a website and extract all data you need from them at once with a single command. For example, collect all published posts on a blog, or extract all products on a shopping website, or gathering comments, reviews for your opinion mining studies. More than that, Rcrawler can help you studies web site structure by building a network representation of a website internal and external hyperlinks (nodes & edges)

**Five Primary Requirements and need for usage of R Crawling**

1.*R-native* : Usually, when R-users need to crawl and scrape a large amount of data automatically, they turn to external tools to collect URLs or carry out the complete task, and then import the collected data into R. Thus, our main reasons for writing RCrawler were to support web crawling and scraping in the R environment. Thus, our solution should be natively implemented in R.

2.*Parallelization* : RCrawler should take advantage of parallelism, in order to gain significant performance enhancements and make efficient use of various system resources, including processors.

3. *Politeness*: The crawler should be able to parse and obey *robots.txt* commands. In addition, the crawler should avoid requesting too many pages in a short interval of time from a given host.

4.*Efficiency* : Our solution should make clever use of resources and be resilient to spider traps. The crawler should be able to detect and avoid duplicate URLs or web pages.

5.*Flexibility and Reconfigurability* : We would like to design a flexible system that can be applied in various scenarios. Below is a summarized list of the settings options for RCrawler:

Project name and directory.

User agent, connection time-out, and request delay.

Filters: Include/exclude content type (MIME), error pages, file extension, and URLs matching a specific regular expression pattern.

Parallel crawling: the user specifies the number of nodes and number of connections (simultaneous requests).

Choose to honor *Robots.txt* file or to ignore it.

Maximum depth level to control the crawling depth.

The crawler should provide the ability to apply some specific functions of content extraction, such as XPath patterns, to each crawled page during the crawling process.

Regarding visualization and results, the crawler should return the crawled data in data structures that are well-organized and ready to use, such as vectors, lists, and data frames. In fact, data returned by other packages is always outspread, and additional effort is required to arrange this into one data structure. We describe the data structures for RCrawler as follows:

A data frame representing the generic URL index, including the list of fetched URLs and page details (content type, HTTP state, number of out-links and in-links, encoding type, and level).

A file repository that contains all downloaded pages.

A vector for scraped contents.

A message including the crawling statistics.

For link analysis, a data structure are needed to represent the connectivity of the web graph (edges).

During the crawling process, the crawler should display the crawling state.

**R Crawler Architecture**

The crawler begins from a given website URL, provided by the user, and progressively fetches this and extracts new URLs (out-links). These in turn are added to the frontier list to be processed. The crawling process stops when all URLs in the frontier are processed.

First, the crawler initiates the working environment, comprising the index structure, the repository that will contain the collection of web documents, and the cluster nodes (workers) for parallel computing. Crawling is performed by multiple worker threads, and the work-pool-handler component prepares a pool of URLs to be processed in parallel. Then, each node executes the following functions for the given URL:

1. Download the corresponding document and its HTTP header using a GET request.

2. Parse and extract all links contained in the document.

3. Proceed with the canonicalization and normalization of URLs.

4. Apply a URL filter, keeping only URLs that match the user-supplied configuration, file type, and domain specific URLs.

As a result, for each URL each worker returns its HTML document, HTTP header details, and the list of discovered out-links. The URL-seen function checks if the URL has already been processed or queued before being added it to the frontier. Before storing the document in the repository, the Is-not-duplicate? function checks that it has not been processed with a different URL. Otherwise, it is discarded. We prefer to write the frontier list and the index to disk in each crawler iteration, to protect them against data loss in the case of a crawler crash. The following are the main RCrawler components:

HTTP requests Handler: handles HTTP requests.

Link Extractor: parses and extracts links from crawled documents.

Content duplicate checker: detects duplicate and near-duplicate documents.

Work-thread manager: handles multi-threading and parallel computing.

Data extraction: a component for parsing and extracting web page contents.

Index: a data structure to store information regarding crawled web pages.

Other components relating to spider traps, URL canonicalization, the robot.txt parser, and other features are discussed in later sections.

**A close up of text on a black background

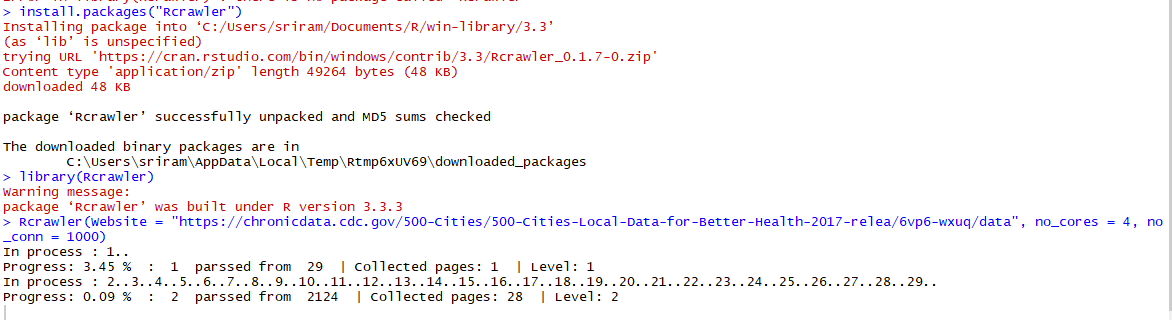
Description generated with very high confidence**

**Steps and Procedure followed -**

1. **Collecting Web Pages from website**

With the documentation and research suggestion from Dr.Salim.K , I have performed web crawling to test and retrieve web pages related to the Health DataSet. More information to perform the R Crawling to any website is available [here](https://github.com/salimk/Rcrawler).

The website link I have considered to initial R Crawler is <https://chronicdata.cdc.gov/500-Cities/500-Cities-Local-Data-for-Better-Health-2017-relea/6vp6-wxuq/data> with number of HTTP requests as 4 and it goes by Levels 1 – 4 and number of connections as 1000.



The package Rcrawler is installed on RStudio environment and the web address to crawl the information i.e. dataset source or address can be provided with the number of HTTP requests to be sent and the number of cores can also be set.

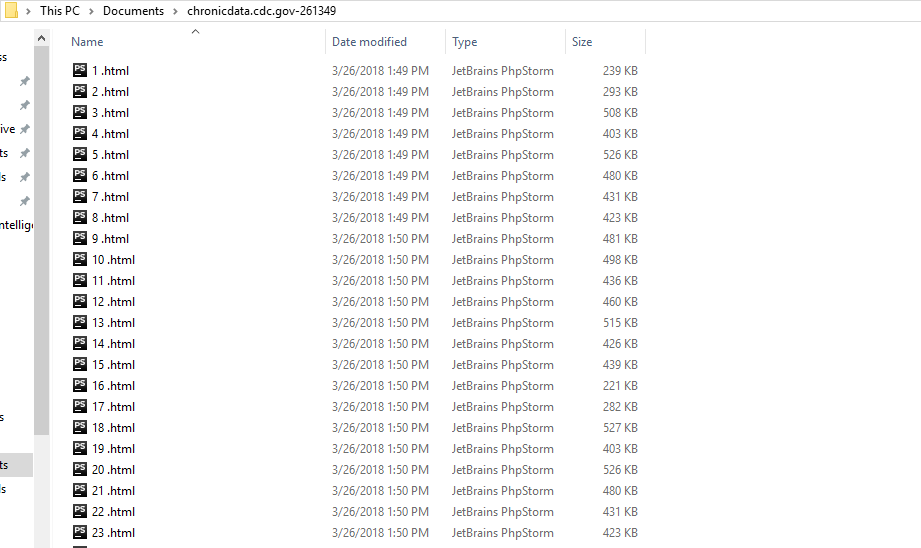
This command allows downloading all HTML files of a website from the server to your computer. It can be useful if you want to analyze or apply something on the whole web page (HTML file).

* no\_cores specify how many processes will execute the task
* no\_conn specify how many HTTP requests will be sent simultaneously (in parallel).

At the end of crawling process this function will return :

A variable named "INDEX" in the global environment: It's a data frame representing the general URL index, which includes all crawled/scraped web pages with their details (content type, HTTP state, the number of out-links and in-links, encoding type, and level). [n.d. Salim, K (April, 2018)]

Upon collection of web pages, the documents can be stored in the system as shown below:



1. **Loading collected HTML files to memory R environment**

After running Rcrawler command, Collected HTML web pages are supposed to be stored on your hard drive, In fact putting downloaded files directly into variables will consume the RAM, So, the crawler creates a folder for each crawling sessions with a name similar this pattern "website-DateTime" . To load collected files into a variable for processing or analysis, you will need to run these two functions: ListProjects and LoadHTMLFiles.

**ListProjects()**

Run this command to list all your crawling project folders. Then you just need to pick-up (copy) the project name you want. Then run the following command which will load all HTML into a vector.

**MyData <- ­­­­­LoadHTMLFiles("chronicdata.cdc.gov-261349", type = "vector")**

You can specify "list" as a type of returned variable. Also, this function has a parameter called (max) useful to limit the number of imported files.

**Keywords – “**Health” and “Indicators” to get list of 100 pages

The next steps can be followed from the documentation and research provided [here](https://github.com/salimk/Rcrawler)

**Data Mining through WEKA**

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

The Documentation and tutorials are as follows:

* + The [**online appendix on the WEKA workbench**](http://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf) for the fourth edition of "Data Mining: Practical Machine Learning Tools and Techniques" by I.H. Witten, Eibe Frank, Mark A. Hall, and Chris J. Pal (2016).
  + The [**Weka Wiki**](http://weka.wikispaces.com/), including [**frequently asked questions (FAQ)**](http://weka.wikispaces.com/FAQ), help on [**trouble-shooting**](http://weka.wikispaces.com/Troubleshooting) Weka, and information on [**using the package manager**](http://weka.wikispaces.com/How+do+I+use+the+package+manager%3F) and [**how to structure packages**](http://weka.wikispaces.com/How+are+packages+structured+for+the+package+management+system%3F) in Weka >= 3.7.
  + Videos and slides for our [**three online courses**](https://www.cs.waikato.ac.nz/ml/weka/courses.html) on data mining with Weka: [**Data Mining with Weka**](https://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/), [**More Data Mining with Weka**](https://www.cs.waikato.ac.nz/ml/weka/mooc/moredataminingwithweka/), and [**Advanced Data Mining with Weka**](https://www.cs.waikato.ac.nz/ml/weka/mooc/advanceddataminingwithweka/).
  + The Weka[**mailing list**](http://list.waikato.ac.nz/mailman/listinfo/wekalist) ([**archive**](http://list.waikato.ac.nz/pipermail/wekalist/)).
  + [**Community documentation**](http://wiki.pentaho.com/display/DATAMINING/Pentaho+Data+Mining+Community+Documentation) for Weka at [**Pentaho**](http://www.pentaho.com/).
  + The Weka manual ([**Weka 3.6.15**](http://prdownloads.sourceforge.net/weka/WekaManual-3-6-15.pdf?download), [**Weka 3.8.2**](http://prdownloads.sourceforge.net/weka/WekaManual-3-8-2.pdf?download), [**Weka 3.9.2**](http://prdownloads.sourceforge.net/weka/WekaManual-3-9-2.pdf?download)), as included in the distribution.
  + The Weka API, extracted from the Javadoc ([**Weka 3.6**](http://weka.sourceforge.net/doc.stable/), [**Weka 3.8**](http://weka.sourceforge.net/doc.stable-3-8/), [**Weka 3.9**](http://weka.sourceforge.net/doc.dev/)), as included in the distribution.
  + [**A list of packages**](http://weka.sourceforge.net/packageMetaData/) for Weka >= 3.7 that can be installed via its package manager. Corresponding Javadoc for the packages is available at http://weka.sourceforge.net/doc.packages/[name of package].
* **Videos, tutorials, blogs, and slides**
  + [**Mark Hall's Weka-related blog**](http://markahall.blogspot.co.nz/)
  + A [**presentation**](http://www.cs.waikato.ac.nz/~eibe/WEKA_Ecosystem.pdf) entitled "WEKA in the Ecosystem for Scientific Computing". Covers how to access the WEKA data mining software from Octave/Matlab, R, and Python. Also discusses how some R functionality can be applied from within Weka and facilities for distributed computation in Weka.

A **[tutorial on connecting Weka to MongoDB using a JDBC driver.](https://github.com/selvinsource/mongodb-datamining-shell/wiki/Weka-MongoDB)**

* + [**A somewhat outdated, but still useful**](https://github.com/selvinsource/mongodb-datamining-shell/wiki/Weka-MongoDB)[**introduction**](http://weka.wikispaces.com/Primer), written by Alex K. Seewald, to using Weka 3.4.6 from the command line.
  + A [**presentation**](http://prdownloads.sourceforge.net/weka/weka.ppt) demonstrating all graphical user interfaces (GUI) in Weka. (Warning: large file.)
  + A [**presentation**](http://prdownloads.sourceforge.net/weka/Weka_a_tool_for_exploratory_data_mining.ppt) which explains how to use Weka for exploratory data mining. (Warning: large file.)
  + A few [**screenshots**](http://www.cs.waikato.ac.nz/~ml/weka/gui_explorer.html) of the Explorer user interface in Weka.
  + An [**introductory video**](http://www.youtube.com/watch?v=m7kpIBGEdkI) by Brandon Weinberg.
  + An [**introductory video on text mining with Weka**](https://www.youtube.com/watch?v=IY29uC4uem8) by Brandon Weinberg.
  + A Weka [**tutorial**](http://www.metaemotion.com/diego.garcia.morate) in Spanish.
  + A [**presentation**](http://weka.sourceforge.net/presentations/KDD05.ppt) about Weka's history, current state and future plans. (June 2005).
* **Technical information**
  + A [**page**](http://weka.wikispaces.com/ARFF) documenting the ARFF data format used by Weka.
  + Documentation on the [**XRFF data format**](http://weka.wikispaces.com/XRFF) used by Weka.
  + A [**page**](http://weka.wikispaces.com/tag/view/ide) listing all the IDEs with step by step instructions on how to setup WEKA as a project.
  + A [**page**](http://weka.wikispaces.com/GenericObjectEditor) describing how to make new classifiers, filters, etc. available in WEKA (using the GenericObjectEditor).
  + Documentation on accessing [**Windows Databases**](http://weka.wikispaces.com/Windows+Databases) like MS Access or other [**Databases**](http://weka.wikispaces.com/Databases) in general (MySQL, PostgreSQL, MS SQL Server, Oracle, etc.) from Weka using JDBC.
  + A [**page**](http://weka.wikispaces.com/Stemmers) describing the stemmer support in Weka.
  + A [**page**](http://weka.wikispaces.com/ARFF+Syntax+Highlighting) where you can find syntax highlighting of the ARFF format for various editors.
  + A [**page**](http://weka.wikispaces.com/XML) describing the use of XML in WEKA.
  + A [**page**](http://wiki.pentaho.com/display/DATAMINING/PMML+Support+in+Weka) with news and documentation on Weka's support for importing [**PMML**](http://www.dmg.org/) models.

1. **ZeroR method Classification**

=== Run information ===

Scheme: weka.classifiers.rules.ZeroR

Relation: 500CitiesLocalHealthData-weka.filters.unsupervised.attribute.Remove-R9

Instances: 116024

Attributes: 16

Year

StateDesc

CityName

GeographicLevel

Category

Measure

Data\_Value\_Type

Data\_Value\_Footnote

PopulationCount

CategoryID

MeasureId

ActualPopulation

Region

HealthOutcomes

Prevention

Test mode: split 75.0% train, remainder test

=== Classifier model (full training set) ===

ZeroR predicts class value: Pacific

Time taken to build model: 0.18 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.77 seconds

=== Summary ===

Correctly Classified Instances 6838 23.5744 %

Incorrectly Classified Instances 22168 76.4256 %

Kappa statistic 0

Mean absolute error 0.1715

Root mean squared error 0.2928

Relative absolute error 100 %

Root relative squared error 100 %

Total Number of Instances 29006

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

0.000 0.000 ? 0.000 ? ? 0.500 0.000 USA

0.000 0.000 ? 0.000 ? ? 0.500 0.044 East South Central

1.000 1.000 0.236 1.000 0.382 ? 0.500 0.236 Pacific

0.000 0.000 ? 0.000 ? ? 0.500 0.097 Mountain

0.000 0.000 ? 0.000 ? ? 0.500 0.146 West South Central

0.000 0.000 ? 0.000 ? ? 0.500 0.033 New England

0.000 0.000 ? 0.000 ? ? 0.500 0.131 South Atlantic

0.000 0.000 ? 0.000 ? ? 0.500 0.137 East North Central

0.000 0.000 ? 0.000 ? ? 0.500 0.058 West North Central

0.000 0.000 ? 0.000 ? ? 0.500 0.119 Middle Atlantic

Weighted Avg. 0.236 0.236 ? 0.236 ? ? 0.500 0.143

=== Confusion Matrix ===

a b c d e f g h i j <-- classified as

3 0 0 0 0 0 0 0 0 0 | a = USA

0 1268 0 0 0 0 0 0 0 0 | b = East South Central

0 0 6838 0 0 0 0 0 0 0 | c = Pacific

2 0 0 2811 0 0 0 0 0 0 | d = Mountain

0 0 0 0 4230 0 0 0 0 0 | e = West South Central

0 0 0 0 0 9630 0 0 0 0 | f = New England

0 0 0 0 0 0 3795 0 0 0 | g = South Atlantic

0 0 0 0 0 0 0 3956 0 0 | h = East North Central

0 0 0 0 0 0 0 0 1289 0 | i = West North Central

0 0 0 0 0 0 0 0 0 2698 | j = Middle Atlantic

**Health Outcomes (Cleaned DataSet)**

=== Run information ===

Scheme: weka.classifiers.rules.ZeroR

Relation: 500CitiesLocalHealthData-weka.filters.unsupervised.attribute.Remove-R9

Instances: 116024

Attributes: 16

Year

StateDesc

CityName

GeographicLevel

Category

Measure

Data\_Value\_Type

Data\_Value\_Footnote

PopulationCount

CategoryID

MeasureId

ActualPopulation

Region

HealthOutcomes

Prevention

Test mode: split 75.0% train, remainder test

=== Classifier model (full training set) ===

ZeroR predicts class value: 595.1234587600

Time taken to build model: 0.18 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.9 seconds

=== Summary ===

Correlation coefficient 0

Mean absolute error 634.912

Root mean squared error 755.8457

Relative absolute error 100 %

Root relative squared error 100 %

Total Number of Instances 29006

**Decision Table Rule based Classification**

=== Run information ===

Scheme: weka.classifiers.rules.DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"

Relation: 500CitiesLocalHealthData-weka.filters.unsupervised.attribute.Remove-R9

Instances: 116024

Attributes: 16

Year

StateDesc

CityName

GeographicLevel

Category

Measure

Data\_Value\_Type

Data\_Value\_Footnote

PopulationCount

CategoryID

MeasureId

ActualPopulation

Region

HealthOutcomes

Prevention

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Decision Table:

Number of training instances: 116024

Number of Rules : 47

Classification value : 540.8547322221

Non matches covered by Majority class.

Best first.

Start set: no attributes

Search direction: forward

Stale search after 5 node expansions

Total number of subsets evaluated: 82

Merit of best subset found: 100

Evaluation (for feature selection): CV (leave one out)

Feature set: 1,14

Time taken to build model: 45.34 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 115976 99.9586 %

Incorrectly Classified Instances 48 0.0414 %

Kappa statistic 0.9995

Mean absolute error 0.0009

Root mean squared error 0.0096

Relative absolute error 0.5047 %

Root relative squared error 3.2678 %

Total Number of Instances 116024

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class

1.000 0.000 0.889 1.000 0.941 0.943 1.000 0.889 USA

0.999 0.000 0.999 0.999 0.999 0.999 1.000 0.999 East South Central

1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 Pacific

0.999 0.000 0.999 0.999 0.999 0.999 1.000 0.999 Mountain

1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 West South Central

0.998 0.000 0.998 0.998 0.998 0.998 1.000 0.999 New England

0.999 0.000 1.000 0.999 1.000 0.999 1.000 1.000 South Atlantic

1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 East North Central

0.999 0.000 0.999 0.999 0.999 0.999 1.000 0.999 West North Central

1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000 Middle Atlantic

Weighted Avg. 1.000 0.000 1.000 1.000 1.000 1.000 1.000 1.000

=== Confusion Matrix ===

a b c d e f g h i j <-- classified as

8 0 0 0 0 0 0 0 0 0 | a = USA

1 5108 0 0 0 0 0 0 3 0 | b = East South Central

0 1 27331 1 1 0 2 0 0 0 | c = Pacific

0 0 3 11240 1 0 0 1 2 1 | d = Mountain

0 2 0 1 16851 1 0 1 0 0 | e = West South Central

0 0 0 2 1 3762 2 0 0 1 | f = New England

0 0 1 0 0 4 15396 0 2 1 | g = South Atlantic

0 0 0 1 0 0 1 15793 1 0 | h = East North Central

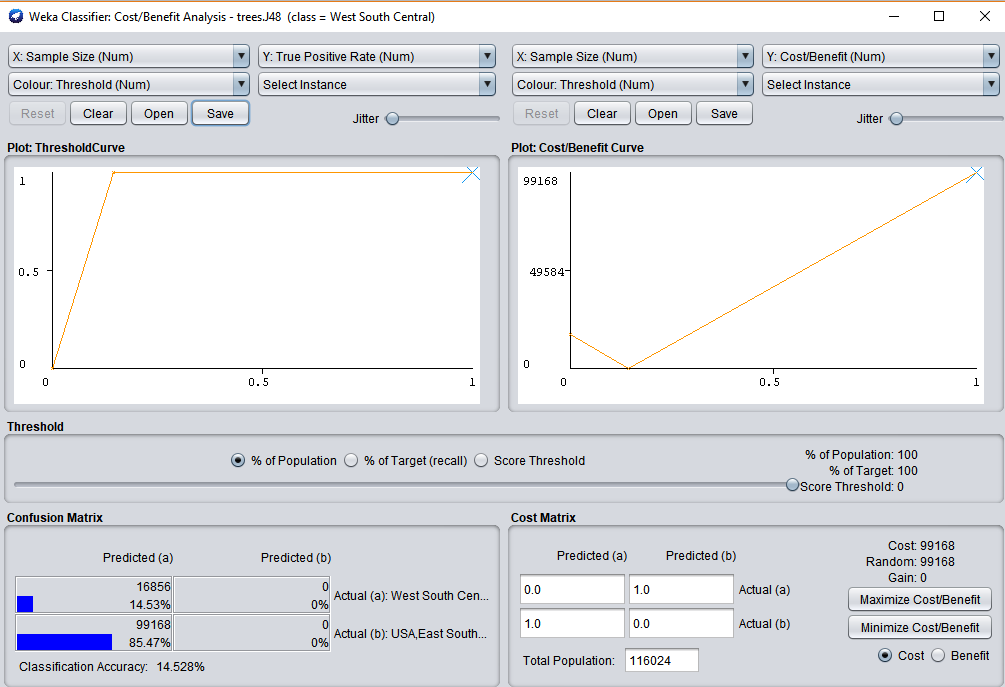
0 1 0 0 0 0 2 2 6539 0 | i = West North Central

0 0 1 1 0 2 0 0 0 13948 | j = Middle Atlantic

**Comparison of Mining Results before and after RCrawling for Cleaned Data**

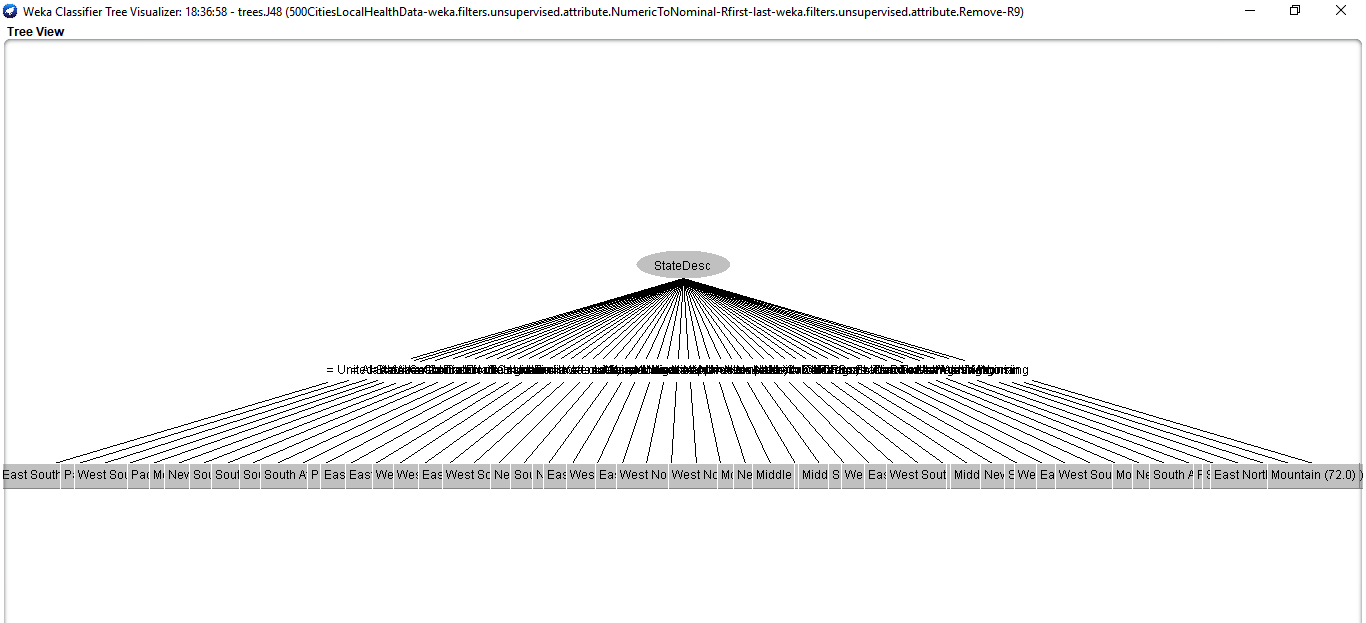
Before web crawling the results only showed the Pacific Region population count being affected by the Health Outcomes variables which lead to a misconception that other regions are least affected. While R Crawling on the variables “Health and “Indicators” got the content distributed in Gaussian manner and there is a well-established confusion matrix distributed and effectively represents the Health Outcomes and Prevention variables for all the regions’ population count which include USA, East South Central, Pacific, Mountain, West South Central, New England, South Atlantic, East North Central, West North Central and Middle Atlantic. California is the state that is most affected by Health Outcomes and requires more preventive measures. While New England has good healthcare facilities and moderate supervision and government health regulation policies would suffice.

**Cost Benefit Analysis (Sample – West South Central Region)**



The Prediction accuracy for maximum Cost/Benefit is 85.47% which equivalents 99168 and it corresponds to USA, East South Central Region. Similarly, the prediction accuracy for the classification for Wet South Central region is 14.53% corresponding to 16856 of the population of the Total Population instance 116024.

**J48 Prune Decision Tree**



J48 Prune Decision Tree Classification shows the decision tree for StateDesc variable providing states and regions for the entire 500 Cities Dataset

**K-Means Clustering**

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 10 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: 500CitiesLocalHealthData-weka.filters.unsupervised.attribute.Remove-R9-weka.filters.unsupervised.attribute.NumericToNominal-Rfirst-last

Instances: 116024

Attributes: 16

Year

StateDesc

CityName

GeographicLevel

Category

Measure

Data\_Value\_Type

Data\_Value\_Footnote

PopulationCount

CategoryID

MeasureId

ActualPopulation

Region

HealthOutcomes

Prevention

Test mode: evaluate on training data

=== Clustering model (full training set) ===

kMeans

======

Number of iterations: 5

Within cluster sum of squared errors: 711229.0

Initial starting points (random):

**Cluster 0**: 58658,2013,Michigan,Detroit,'Census Tract',Prevention,'Cholesterol screening among adults aged equal and larger than 18 Years','Crude prevalence',NA,6295,PREVENT,CHOLSCREEN,4859.74,'East North Central',0,4859.74

**Cluster 1:** 34677,2013,Florida,'Fort Lauderdale','Census Tract',Prevention,'Cholesterol screening among adults aged equal and larger than 18 Years','Crude prevalence',NA,2224,PREVENT,CHOLSCREEN,1779.2,'South Atlantic',0,1779.2

**Cluster2:** 86332,2013,Oklahoma,'BrokenArrow','Census Tract',Prevention,'Cholesterol screening among adults aged equal and larger than 18 Years','Crude prevalence',NA,1350,PREVENT,CHOLSCREEN,1058.4,'West South Central',0,1058.4

**Cluster3:** 90809,2013,Pennsylvania,Philadelphia,'Census Tract',Prevention,'Cholesterol screening among adults aged equal and larger than 18Years','Crude prevalence',NA,3996,PREVENT,CHOLSCREEN,2901.096,'Middle Atlantic',0,2901.096

**Cluster 4:** 93347,2013,'South Carolina','North Charleston','Census Tract','Health Outcomes','High cholesterol among adults aged equal and larger than 18 Years who have been screened in the past 5 Years','Crude prevalence',NA,2649,HLTHOUT,HIGHCHOL,1003.971,'South Atlantic',1003.971,0

**Cluster 5:** 59078,2013,Michigan,Detroit,'Census Tract','Health Outcomes','High cholesterol among adults aged equal and larger than 18 Years who have been screened in the past 5 Years','Crude prevalence',NA,1899,HLTHOUT,HIGHCHOL,725.418,'East North Central',725.418,0

**Cluster 6:** 68474,2013,'New Mexico',Albuquerque,'Census Tract','Health Outcomes','High blood pressure among adults aged equal and larger than 18 Years','Crudeprevalence',NA,3170,HLTHOUT,BPHIGH,763.97,Mountain,763.97,0

**Cluster 7**: 100051,2013,Texas,'El Paso','Census Tract','Health Outcomes','High blood pressure among adults aged equal and larger than 18 Years','Crude prevalence',NA,4542,HLTHOUT,BPHIGH,1380.768,'West South Central',1380.768,0

**Cluster 8:** 40486,2013,Georgia,'Johns Creek','Census Tract',Prevention,'Taking medicine for high blood pressure control among adults aged equal and larger than 18 Years with high blood pressure','Crude prevalence',NA,287,PREVENT,BPMED,200.039,'South Atlantic',0,200.039

**Cluster 9**: 99170,2013,Texas,Dallas,'Census Tract',Prevention,'Taking medicine for high blood pressure control among adults aged equal and larger than 18 Years with high blood pressure','Crude prevalence',NA,7167,PREVENT,BPMED,4880.727,'West South Central',0,4880.727

Missing values globally replaced with mean/mode

Final cluster centroids:

Cluster#

Attribute Full Data 0 1 2 3 4 5 6 7 8 9

(116024.0) (17410.0) (5470.0) (4238.0) (5998.0) (24431.0) (5319.0) (23446.0) (4817.0) (20681.0) (4214.0)

==================================================================================================================================================================================================================================================================================================================================================================================================================================================================================================================================================================================================================================

1 5 4867 3219 29989 7 2084 1 6009 3 6028

Year 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013

StateDesc California California Florida Texas New York California Illinois California Texas California Texas

CityName New York Los Angeles Baltimore Houston New York New York Chicago New York Houston Los Angeles Houston

GeographicLevel Census Tract Census Tract Census Tract Census Tract Census Tract Census Tract Census Tract Census Tract Census Tract Census Tract Census Tract

Category Health Outcomes Prevention Prevention Prevention Prevention Health Outcomes Health Outcomes Health Outcomes Health Outcomes Prevention Prevention

Measure

High blood pressure among adults aged equal and larger than 18 Years Cholesterol screening among adults aged equal and larger than 18 Years Cholesterol screening among adults aged equal and larger than 18 Years Cholesterol screening among adults aged equal and larger than 18 Years Cholesterol screening among adults aged equal and larger than 18 Years

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

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

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

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

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

**Data\_Value\_Type** Crude prevalence Crude prevalence Crude prevalence Crude prevalence Crude prevalence Crude prevalence Crude prevalence Crude prevalence Crude prevalence Crude prevalence Crude prevalence

**Data\_Value\_Footnote** NA NA NA NA NA NA NA NA NA NA NA

**PopulationCount**  1394 2529 1 1394 1754 1394 1 1394 1394 1394 1394

**CategoryID**  HLTHOUT PREVENT PREVENT PREVENT PREVENT HLTHOUT HLTHOUT HLTHOUT HLTHOUT PREVENT PREVENT

**MeasureId** BPHIGH CHOLSCREEN CHOLSCREEN CHOLSCREEN CHOLSCREEN HIGHCHOL HIGHCHOL BPHIGH BPHIGH BPMED BPMED

**ActualPopulation**  0 0 0 0 0 0 0 0 0 0 0

**Region**  Pacific Pacific South Atlantic West South Central Middle Atlantic Pacific East North Central Pacific West South Central Pacific West South Central

HealthOutcomes 0 0 0 0 0 0 0 0 0 0 0

Prevention 0 0 0 0 0 0 0 0 0 0 0

Time taken to build model (full training data) : 9.88 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 17410 ( 15%)

1 5470 ( 5%)

2 4238 ( 4%)

3 5998 ( 5%)

4 24431 ( 21%)

5 5319 ( 5%)

6 23446 ( 20%)

7 4817 ( 4%)

8 20681 ( 18%)

9 4214 ( 4%)

**Data Mining Analysis Conclusion and Recommendations**

Based on my analysis, I believe that there are still many states which are not completely covered with Preventive measures for the health risk faced by the population. The Population affected by the High Blood Pressure is covered with measures than the states facing High Cholesterol.

For Population affected by High Cholesterol, more preventive measures need to be taken covering more states and more cities. I would recommend the Data decision makers/ government officers/ Chief Medical officers should create strategy Plan to solve the problem of High Cholesterol in various states. The information also needs to be updated and researched based on other heath aspects and vision that could benefit more population in the United States.

**References and Bibliography**

* + 1. 500 Cities: Local Data for Better Health. (2016, December 07). Retrieved February 6, 2018 from <https://catalog.data.gov/dataset/500-cities-local-data-for-better-health-b32fd>
    2. R Crawler, Salim. K: Retrieved March 25th, 2018 from <https://github.com/salimk/Rcrawler/blob/master/README.md>
    3. 500 Cities: Local Data for Better Health. (2016, December 07). Retrieved April 13, 2018, from <https://catalog.data.gov/dataset/500-cities-local-data-for-better-health-b32fd>