Jiayuan Shi
MA881-Midterm
CHSI Report

Introduction

This is a report of exploration and modeling data from the Relative Health Importance dataset for the Worcester, MA community. The dataset identifies the data elements and values in the Relative Health Importance indicator domain. It is original from the CHSI(Community Health Status Indicators) report, which contains over 200 measures for each of the 3141 United Stated counties. From analyzing the data, the community leaders of Worcester can know how they compare with other communities in MA State and how they compare with other communities in the United States with regard to the health measures included in the dataset.

Method & Results

In order to prepare for the report, I read the dataset, clean it, and organize it for exploration and simple modeling.

1. Most of the measures are Relative health indicator. I find the data of Worcester contains 11 of the 22 health measures with indicator 5, 1 measure has indicator 6, 6 measures has indicator 7 and 4 measures have indicator 8. The meanings for the indictors are,

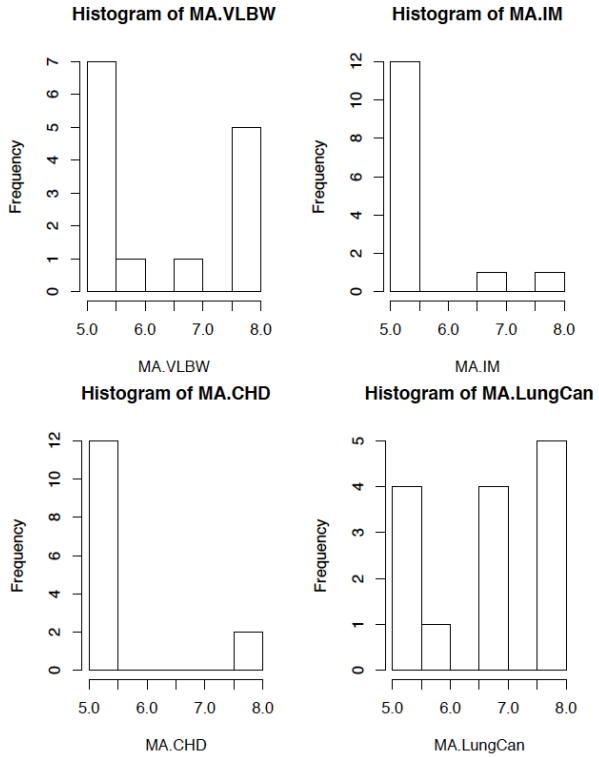
Indictors	Description
5	Represent 'Favorable to peers and favorable the U.S. Rate'
6	Represent 'Favorable to peers and unfavorable the U.S. Rate'
7	Represent 'Unfavorable to peers and favorable the U.S. Rate'
8	Represent 'Unfavorable to peers and unfavorable the U.S. Rate'

2. Therefore, those favorable measures make up of more than 80% of the total ones, so I think Worcester does better in health status than most of the other peer communities and communities in the whole U.S.
3. There are totally 28 measures in the dataset, and 22 are about those communities' health conditions. I chose 6 measures of interest to compare Worcester with other communities as followings,

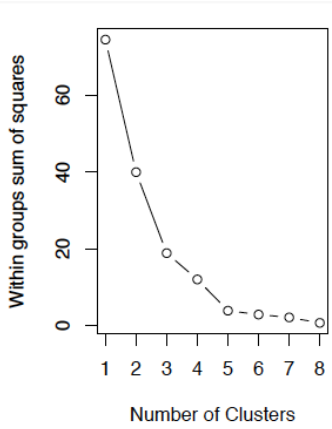
Abbreviation	Column Name	Data Type	Description
SC	State_FIPS_Code	Text	Two-digit state identifier, developed by the National Bureau of Standards
CC	County_FIPS_Code	Text	Three-digit county identifier, developed by the National Bureau of Standards
VLBW	RHI_VLBW_Ind	Integer	Relative health indicator, very low birth wt. (<1500 g)
IM	RHI_Infant_Mortality_Ind	Integer	Relative health indicator, infant mortality
CHD	RHI_CHD_Ind	Integer	Relative health indicator, coronary heart disease
LungCan	RHI_Lung_Cancer_Ind	Integer	Relative health indicator, lung cancer

4. I combine the above measures into a new matrix named *"rhi"* and do the data cleaning. I drop the whole rows with values -2 or -1, because they mean no data available(-2) and no report(-1). I make the new data a data frame with 2860 communities, 6 measures and still call it *"rhi"*.
5. First, I want to compare Worcester with other communities in MA.

- I subset all the 14 counties in MA with the name *"MA"*, and Worcester with name *"Worcester"*. In *"Worcester"* data, *"VLBW"*, *"IM"* and *"CHD"* measures all have indicators 5(representing 'Favorable to peers and favorable the U.S. Rate'), and *"LungCan"* measure has indicator 7(representing 'Unfavorable to peers and favorable the U.S. Rate'). Therefore, for these health measures, Worcester is favorable.
- Because the first 2 columns of *"MA"* data frame are texts, not integers, we only talk about the last 4 columns with measures *"VLBW"*, *"IM"*, *"CHD"* and *"LungCan"*. So I construct a new data frame called *"MAdata"* with only the 4 measures. I write a function *"mysummary"* to summarize the mean, standard deviation, length, min, 25% quantile, median, 75% quantile, max, NAs in each measure. Here I use mean as a way to compare. For *"VLBW"*, *"IM"*, *"CHD"* and *"LungCan"*, the mean for all the counties in MA is 6.29, 5.36, 5.43, and 6.71, compared with Worcester's indicators are 5, 5, 5, 7. So Worcester is more favorable to peers and to the U.S. Rate for *"VLBW"*, more favorable to the U.S. Rate for *"IM"* and *"CHD"*, and less favorable to peers but more favorable to the U.S. Rate for *"LungCan"*.
- I also make stem-plots, plots, histograms, boxplots and qplots for the *"MAdata"* data frame. For example, these histograms below show the distribution for the 14 communities in MA for each measure.



- Finally, I use the K-means clustering method to put the 14 communities, which share similar descriptions in this dataset into groups. I use The Elbow Method to look at the percentage of variance explained as a function of the number of clusters, and plot to determine the number of clusters, which is chosen at the elbow point. In this situation of MA communities, the location of the elbow in the resulting plot suggesting a suitable number of clusters for the k-means is 3.



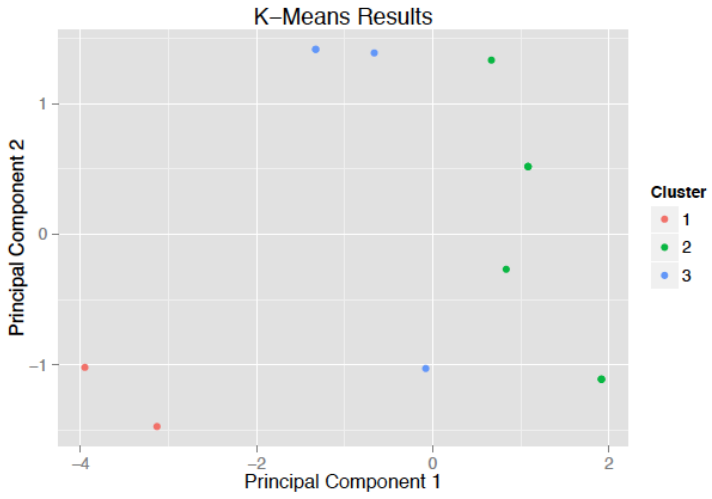
- So I use the *"kmeans"* function in R to divide the 14 counties into 3 groups and plot them. From the output, we can see that,
  a. K-means clustering with 3 clusters of sizes 2, 8, 4.
  b. Worcester is in the 2nd clustering vector with cluster means:

     VLBW IM CHD LungCan
     2 5.125 5 5 7.25

     Worcester's "*VLBW"*(5) and "*LungCan"(7)* is less than the cluster mean of "*VLBW"* and "*LungCan"*, but it has the same "*IM"(5)* and "*CHD"(5)* as the mean of cluster mean "*IM"* and "*CHD"*. This means for "*VLBW"* and "*LungCan"*, Worcester is more favorable to the U.S. Rate than the other counties in MA; for "*IM"* and "*CHD"*, Worcester and other counties in MA are all favorable to peers and favorable to the U.S. Rate.
  c. Within cluster sum of squares by cluster:

     [1] 1.00 10.38 7.50

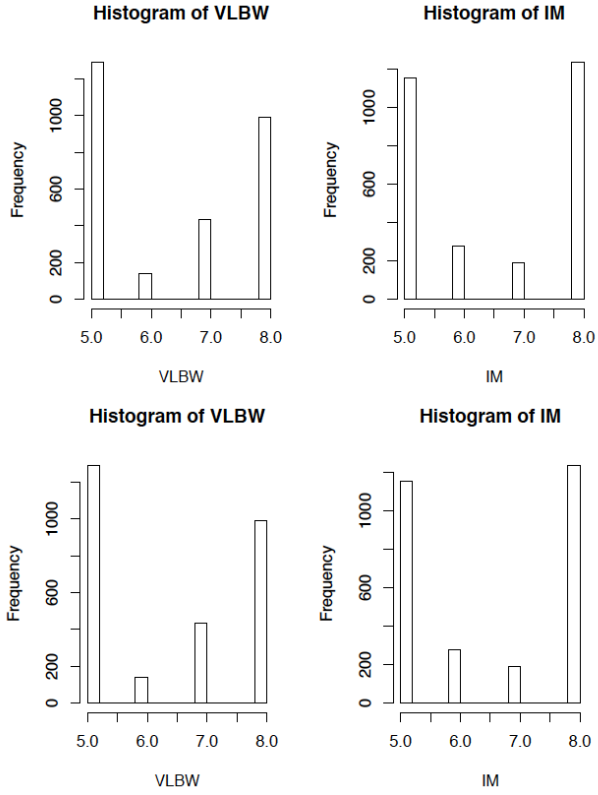
     So the 2nd cluster has the biggest within cluster sum of squares, which means the individual indicators in Worcester's group has the biggest differences. This can also be seen in the plot of the K-means results.



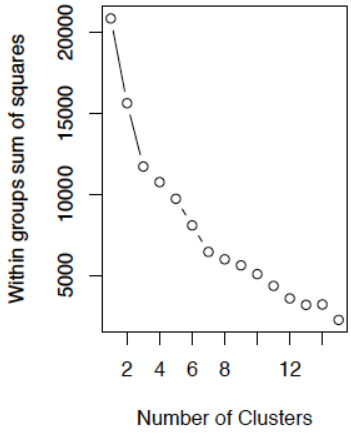
6. Second, I want to compare Worcester with other communities in the U.S. with the similar processes.
   - For the cleaned data frame *"rhi"* with 2860 communities I get before, I also construct a new data frame called *"rhidata"* with only the 4 integer-measures *"VLBW"*, *"IM"*, *"CHD"* and *"LungCan"* and put them into *"mysummary"* function to summarize the mean, standard deviation, length, min, 25%

quantile, median, 75% quantile, max, NAs in each measure. For *"VLBW"*, *"IM"*, *"CHD"* and *"LungCan"*, the mean for all the counties in U.S. is 6.39, 6.53, 6.64 and 6.65, compared with Worcester's indicators are 5, 5, 5, 7. So Worcester is more favorable to peers and to the U.S. Rate for *"VLBW"*, *"IM"*, and *"CHD"*, and less favorable to peers but more favorable to the U.S. Rate for *"LungCan"*.

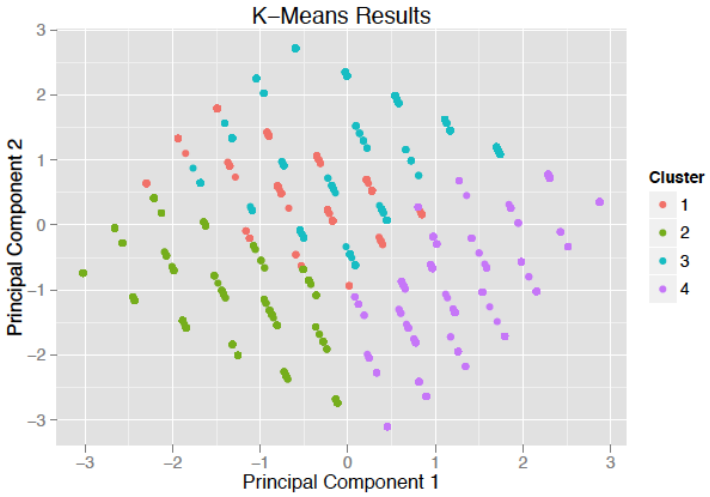
- I also make stem-plots, plots, histograms, boxplots and qplots for the *"rhidata"* data frame. For example, these histograms below show the distribution for the 2860 communities in U.S. for each measure.



- Finally, similar with the process I do K-means for communities in MA, I use the K-means to group the 2860 communities which share similar descriptions in this dataset. With The Elbow Method, I find the location of the elbow in the resulting plot suggesting a suitable number of clusters for the k-means is 4.



- So I use the *"kmeans"* function in R to divide the 2860 counties into 4 groups and plot them. From the output, we can see that,
  a. K-means clustering with 3 clusters of sizes 373, 740, 802, 945.
  b. Worcester is in the 2nd clustering vector with cluster means:
     VLBW IM CHD LungCan
     2 5.661 5.816 5.245 5.364
     Worcester's "*VLBW"*(5) "*IM"(5)* and "*CHD"(5)* is less than the cluster mean of "*VLBW"*, "*IM"* and "*CHD"*, and Worcester's "*LungCan"(7)* is larger than the cluster mean of "*LungCan"*. This means for "*VLBW"*, "*IM"* and "*CHD"*, Worcester is more favorable to the U.S. Rate than the other counties in U.S.; for "*LungCan"*, Worcester is less favorable to peers, but more favorable to the U.S. Rate than the other counties in U.S.
  c. Within cluster sum of squares by cluster:
     [1] 1069 2287 3494 3231
     So the 2nd cluster has the second smallest within cluster sum of squares, which means the individual indicators in Worcester's group has the differences not too big. This can also be seen in the plot of the K-means results.



Reference

Jared P. Lander. (2013). *R for Everyone: Advanced Analytics and Graphics*. Boston: Addison-Wesley Professional.

Community Health Status Indicators (CHSI) to Combat Obesity, Heart Disease and Cancer. Retrieved from https://catalog.data.gov/dataset/community-health-status-indicators-chsi-to-combat-obesity-heart-disease-and-cancer