The Influence of Population on Disease in Cities

Assignment 6

Christian Wesselborg

Data Analytics BCBP – 4960

**Abstract**

Urban populations are dense. Concentration of living, contact and exposure to chemicals and pollutants occur in higher concentration when humans build cities [Gould, 2018]. Businesses and consumers demand goods, food, cars, and energy. When this happens on a condensed scale pollution accumulates. Los Angeles, Beijing, and far more suffer from poor air quality [Han, 2018]. Poor air quality may have an effect on pulmonary illness rates such as asthma [Pyne, 2002]. Chemical pollutants while controlled by the EPA can still be present under acceptable concentrations. The collective effect of these carcinogens, irritants, and particulates may be a general increase in the prevalence of health conditions in United States cities that have large, dense populations [Lima d.S., 2018]. For this project population will be treated as a proxy for a general exposure factor.

Figure 1



*Smog in Los Angeles. [1]*

By analyzing data on disease rates in cities across the United States an insight into correlations, links, and relationships may be discovered. With this information city planners as well as city officials may make better decisions regarding the government of cities for the benefit of the people who live in them.

**Data**

This project will be exploring the “500 Cities: Census Tract-level Data (GIS Friendly Format), 2018 release” data set from the CDC, Centers for Disease Control and Prevention, (the specific data set version is in the GitHub repository for this project along with related code [2]). The data set contains information on 27 various general health related statistics, as well as a population statistic (2010) for 474 total cities for a total of 27,210 subsect areas (used by the census) of the cities. Data was collected in 2015 and 2016 by the CDC. Information on cancer prevalence, diabetes, stroke, smoking, sleep loss, mental health and more are included. The data set was selected for the breadth of information accounted for, general cleanliness (few missing data points), and the current maintenance (last updated in December, 2018). Having a well maintained data set is important in developing sound models from the data. In addition the CDC provides a ‘.shp’ file that is used for constructing maps with spatial-data from the CDC.

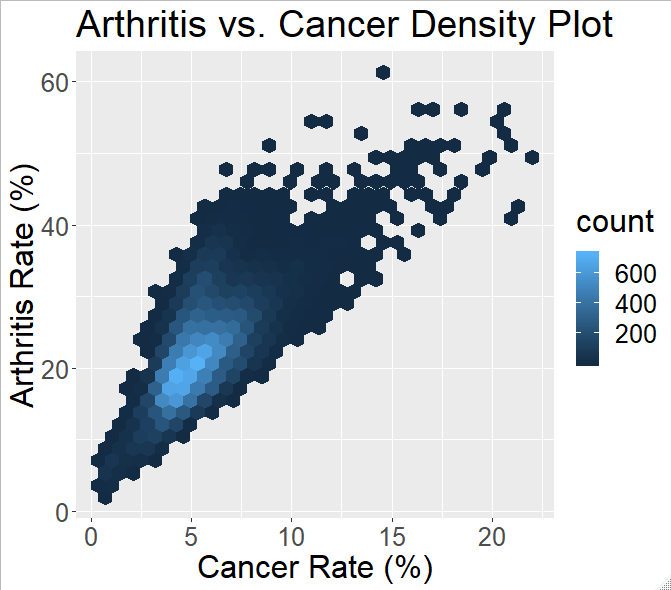
The data set unfortunately does not provide depth of information on population. The data set only includes the population count without information on population density, age, or wealth. These variables would provide a more comprehensive characterization of these areas that population count alone does not provide. Despite missing information, the 500 Cities data set still provides quality information and will be a sound source for this project.

**Analysis**

Initial data cleaning was sparse as the data set was well maintained and had few missing values. Because the data set presented the data in the form of subsect areas from the individual cities, combining the data into total city level data would provide a useful comparison in the project. For example, because the data set does not provide land area information, individual population in an area would not represent that area in the context of a city. A small area with high population may be surrounded by equally high-population small areas, thus being part of a dense population. By comparing results with the total, city-level data provides a look at larger trends that may not be clear in the original form of the data. This was done by taking the population of the area and multiplying that value by the rate of that variable, then calculating the summation for each city and dividing by the total population to return the data to the original rate format, percent values. This process was run in a loop calculating the values for each variable and city then saved that data in a separate file. The end result was a data set representative of all 474 cities with all of the rate variables associated with the whole city.

When beginning to look at the data, it was clear that population was not a suitable predictor for any of the variables in the data set. Even when compared with the total city data the regressions were not indicative of a correlated relationship. Most observations have low populations, typically less than 2,000. This caused most regressions to not represent the high populations well, making them unfit for analysis. However two regressions re-characterized the approach to this project. First a regression of arthritis rates against cancer rates provided a strong positive relationship. A very reasonable explanation being that older populations have higher rates of cancer due to simply being older, cancer tends to appear later in life. Arthritis being an illness which is more prevalent in older populations implicated this age link.

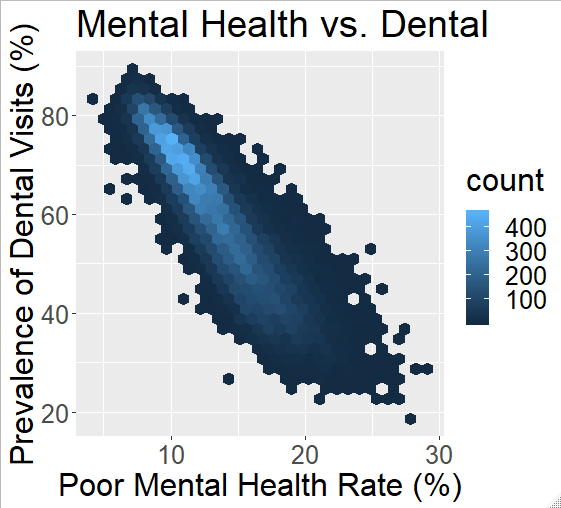
Figure 2



*Arthritis rates against cancer rates represented in a density plot. The two variables are clearly positively correlated. Age is likely the link between these variables.*

What this indicates is that without an age variable the contingency of disease on population cannot be effectively explored using this data set. In order to properly investigate, the affect of age must be corrected for, without that information the results would be unreliable. The second regression was of rates of regular dental visits against poor mental health rates, this provided a distinct negative correlation.

Figure 3

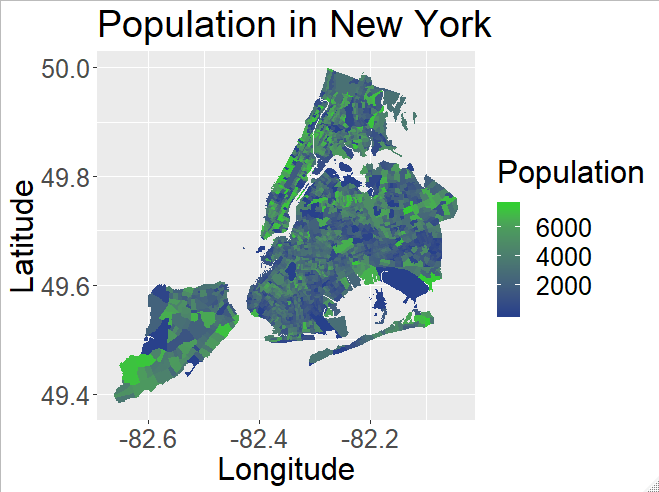


*Dental visit rate against poor mental health rate. The two are negatively correlated.*

The reason for this negative correlation is not immediately obvious and will be explored further in the model development, and conclusions section of this paper.

Regressions were not suitable for this project because they did not represent the observations, segmented city areas, in relation to others in the same city. The common geographic area information would be lost. Constructing maps allows a visual analysis of the data. Maps place observations, segmented city areas, in relation to each other. Using New York City as a representation of a high population city the general relationships explored in this project can be observed.

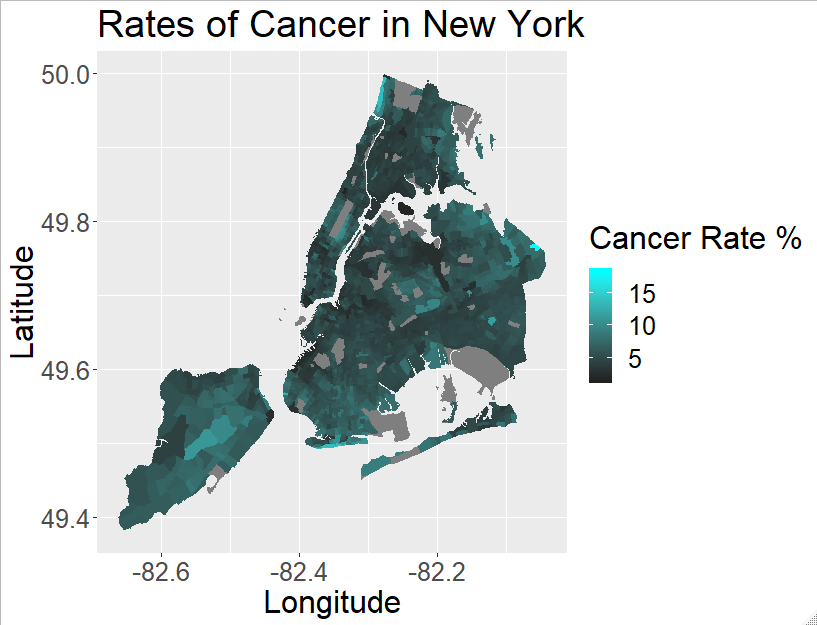
Figure 4



*Population distribution of New York City with higher population areas displayed in green.*

The distribution of population appears to have high populations in Manhattan and Staten Island. Brooklyn and Queens have comparatively less dense populations.

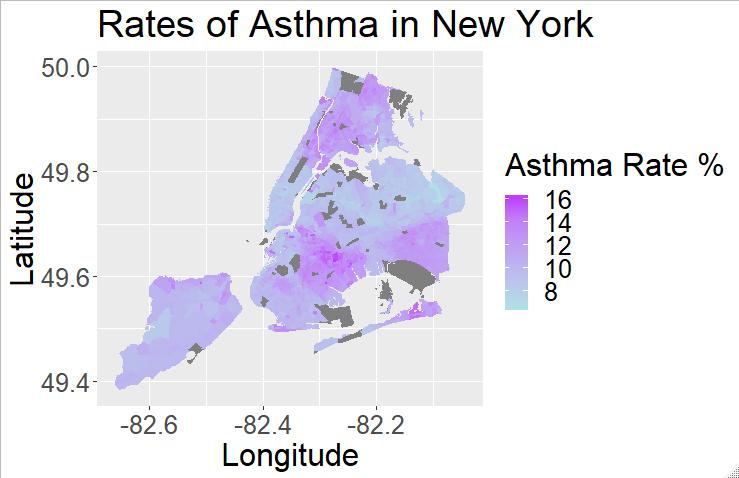
Figure 5



*Cancer rates do not appear to follow the population distribution. Notable centers being in Staten Island and Queens.*

Cancer rates appear do not follow a similar distribution as population. It appears to fairly random where higher areas are. This is seen by a lack of grouping, or concentration, of the cyan color.

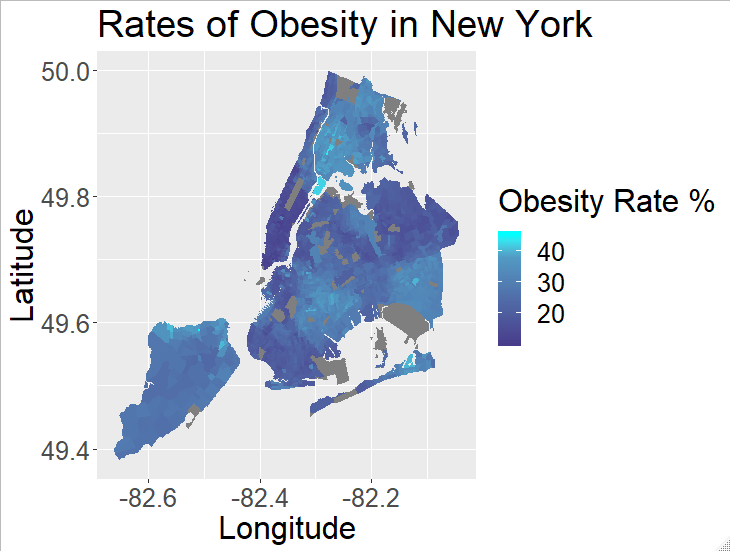
Figure 6



*Asthma rates do not share the same distribution as population. Notable centers are seen in the Bronx and east Brooklyn.*

Similar to the cancer rate map, asthma does not follow the population distribution. Dissimilar to the cancer rate map, asthma has clear and distinct centers. Notably Manhattan and Queens have very low asthma rates, while the Bronx and east Brooklyn are centers of high asthma rates.

Figure 7



*Obesity rates follow a similar distribution to asthma rates.*

The obesity rate map follows a curiously similar distribution to asthma rates. The two maps share clear centers in east Brooklyn and the Bronx. The significance of this correlation is explored in the conclusion section.

**Model Development and Application**

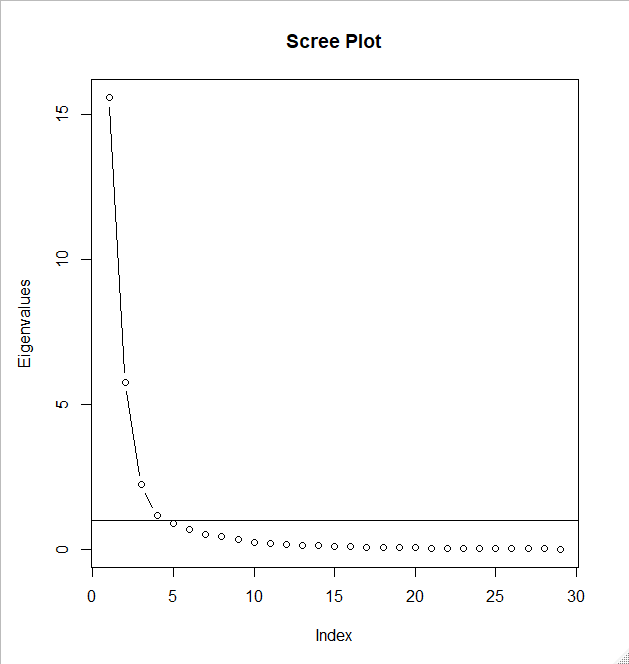
As stated in the analysis section the two graphs, arthritis against age and mental health against dental visits, refocused the approach to this problem. It was clear that this data set, without information on age, population density, pollution, and others, was not suitable to conclude on the influence of population on health statistics. The two graphs prompted the question, if population does not influence disease, what are the underlying factors? In order to investigate this, factor analysis would be used on the data set. Factor analysis groups variables that correlate together. This is a form of data reduction where many variables are grouped into manageable clusters. These groups can be interpreted as common effects of an underlying, latent, variables.

For a factor analysis the number of latent variables extracted is dependent on the analyst. The data must first be examined to determine whether or not factor analysis should be done. In order to answer this the Kaiser-Meyer-Olkin measure (KMO), Bartlett’s test of sphericity, and the determinant of the correlation matrix will be used. Several tests can be done to determine the optimal number of factors to extract. A scree plot and Kaiser’s criterion will be used in this project to determine the optimal number. Additionally. The goal of the factor analysis being to reduce the data, each factor extracted should be meaningful and well separated. Well separated means that the variables that belong to an extracted factor should correlate well with that factor and less so with others.

The KMO test and Bartlett’s test determine whether or not the data being examined should be used in factor analysis. The tests look for correlation between the variables in the data. The determinant of the correlation matrix indicates whether multicollinearity may be an issue. Multicollinearity is the shared correlation between variables. The KMO test returned a value of 0.919, which is a very good result. Note that the total city data returned a value of 0.909 which is equally promising. The KMO test indicates that the data is appropriate for factor analysis. Bartlett’s test also indicates that the data is appropriate, returning very significant results. The determinant of the correlation matrix, however, does indicate that multicollinearity may be an issue. This will be considered later on in this section. This result will not prevent a factor analysis from being conducted for this project.

In determining the optimal number of factors to extract a scree plot will be constructed and evaluated in conjunction with Kaiser’s criterion. The scree plot plots the eigenvalues of the each factor against the index of that factor. The index of a factor is determined by the order it is extracted, a lower index means a more significant factor.

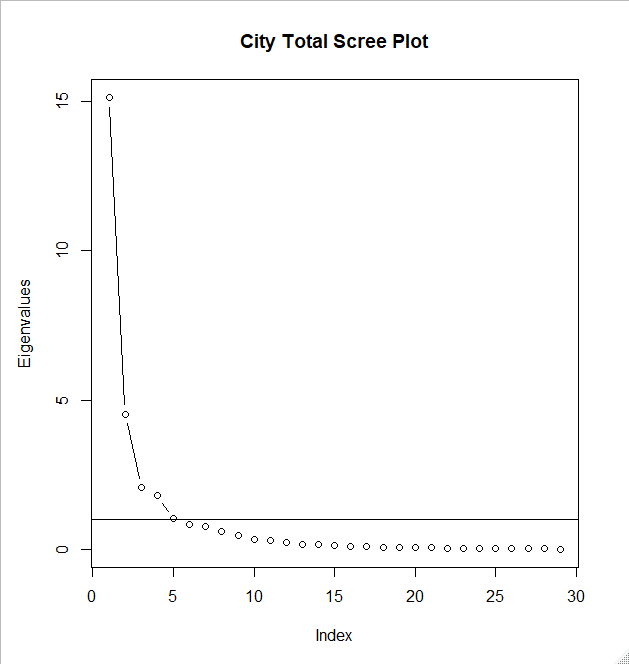
Figure 8



*The scree plot (the name comes from the rock formation off the side of a cliff) indicates the importance of successive factors. The line is placed at the value 1 to indicate Kaiser’s criterion.*

The scree plot above indicates that approximately three factors should be extracted. This is seen where the slope of the plot changes drastically. In addition above the line at a value of 1 indicates Kaiser’s criterion, where factors should only be extracted if they are above that line. These two indicators together point to three factors being an optimal number to extract. The total city data scree plot is shown below.

Figure 9



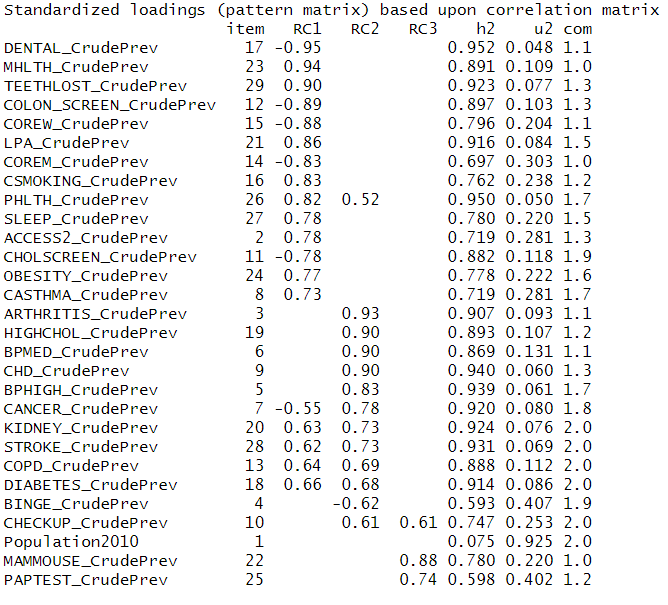
*The scree plot for the total city data gives a similar result to the segmented data.*

Here the optimal number of factors to extract is about three. This is promising that the two data sets are returning similar values.

The factor analysis technique to be used in this project is principal component analysis, PCA. PCA calculates the correlation coefficients of each variable with respect to the extracted factors. The extracted factors are at 90 degree angles to each other and will be rotated using the “varimax” method. This maximizes the variable loadings to the extracted factors but does not change the overall results. Analysis will be done by examining the eigenvalues, and by visual analysis of the biplots from the PCAs. Biplots display the extracted factors as the two axes of a graph and display the variables as arrows. The closer the arrow to 90 degree intervals, the better the separation of that variable. Variables between 90 angles share correlation with both of the factors. Biplots can only be meaningfully interpreted for PCAs of two factors.

The first PCA generated with the segmented data extracted three factors.

Figure 10

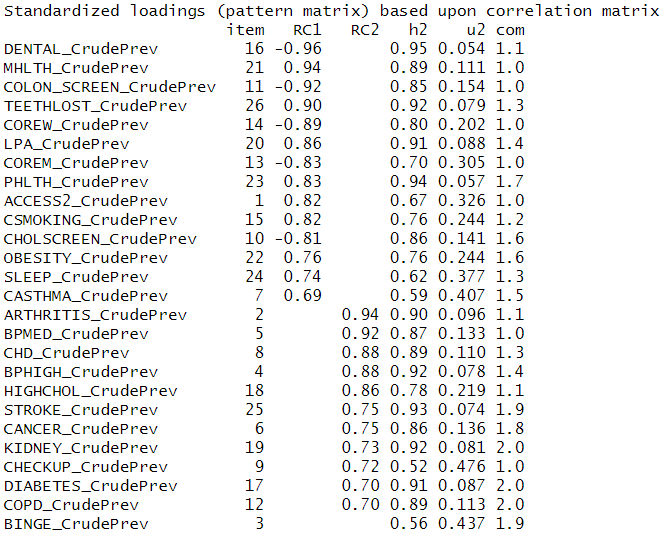


*The result of PCA on the segmented data without removing variables. Values are not shown below 0.5 in order to retain clarity of results.*

The first PCA conducted included all variables in the data set. Note that the population variable was not sorted meaningfully into any factor indicated by the red elipse. Additionally a few variables, cancer, chronic kidney disease, stroke, chronic obstructive pulmonary disease, and diabetes were also correlated with factor 1 as well as factor 2, but more significantly with factor 2. In addition poor physical health was sorted in both factor 1 and 2 but more significantly into factor 1. Regular checkup equally sorted into factor 2 and factor 3. A preliminary guess into the factors could be as follows, factor 1 contains medical screening and health variables, factor 2 contains disease variables, and factor 3 contains women’s reproductive health testing. However there are overlaps between factor 1 and factor 2 where disease variables are correlated with both. Additionally certain physical conditions and lifestyles are sorted into factor 1. For example, teeth loss, no leisure physical activity (LPA), poor physical health, sleep deprivation, obesity, asthma, and various diseases mentioned previously, all sort into this category as well.

The next series of PCAs were conducted after removing one variable at a time. First removing the population variable had, as expected, little effect of the PCA. This was done because at this point, population was no longer significant in this analysis as it had not grouped with any factor. Next removing the variables of the third factor, mammography and pap-test, was done to focus on the first two factors.

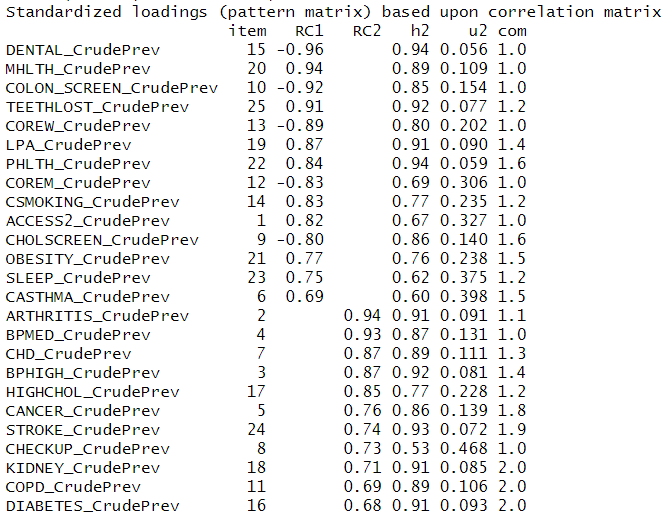
Figure 11



*Result of the PCA, removing population and factor 3 variables. Values under 0.66 are not shown for clarity.*

After reducing the focus to two factors there appears to be a clear separation between the two factors. The variables for stroke, cancer, chronic kidney disease, diabetes, chronic obstructive pulmonary disease, and poor physical health, are correlated with the other factor but with no higher a value then the lowest correlation value in either factor. The binge drinking variable does not sort into any factor however. It is likely that this does not correlate meaningfully with the factors created and will therefore not be included in the model.

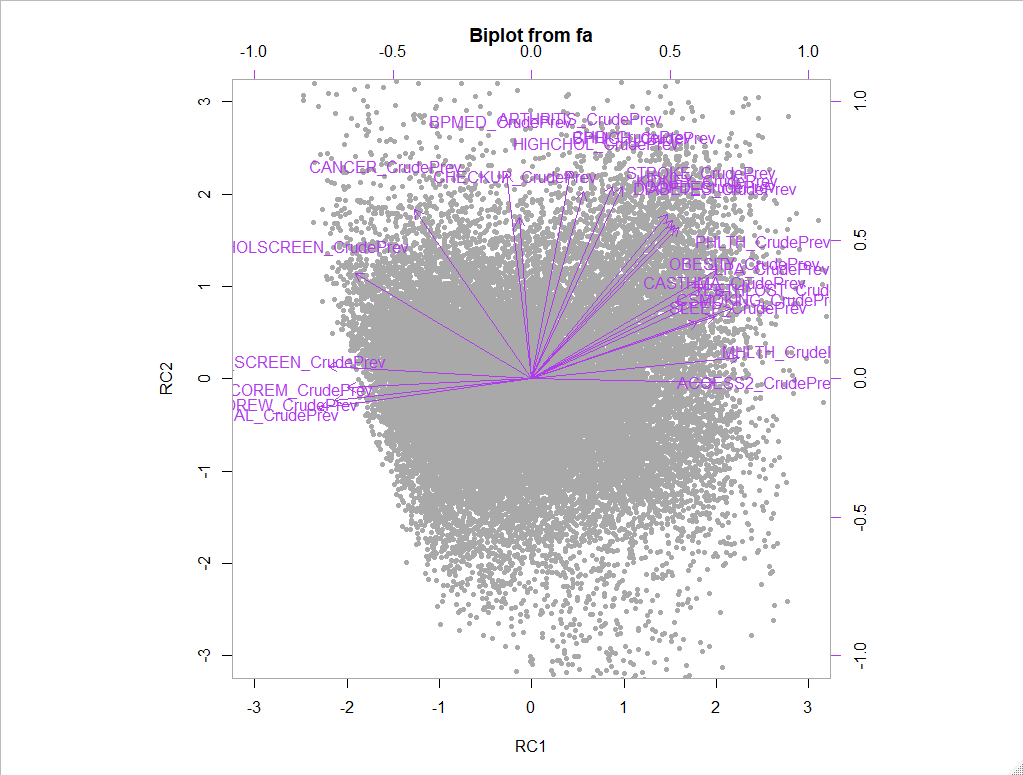
Figure 12



*The final PCA model results. Two distinct factors were extracted. Values under 0.67 are not shown.*

The final model for the segmented data is shown above. The variables: regular dental visits, mental health, colon screening, teeth loss, women’s core health check-ups and preventative services, low physical activity men’s core health check-ups and preventative services, poor physical health, access to health care, smoking, cholesterol screening, obesity, sleep deprivation, and asthma, group together in factor 1. The variables: arthritis, blood pressure medication, coronary heart disease, high blood pressure, high cholesterol, stroke, regular check-ups, cancer, chronic kidney disease, chronic obstructive pulmonary disease, and diabetes, all group together in factor 2. The interpretation of these factors is detailed in the conclusions section. There remain variables that have values greater than 0.5 that indicate that these variables are correlated with each factor.

Figure 13

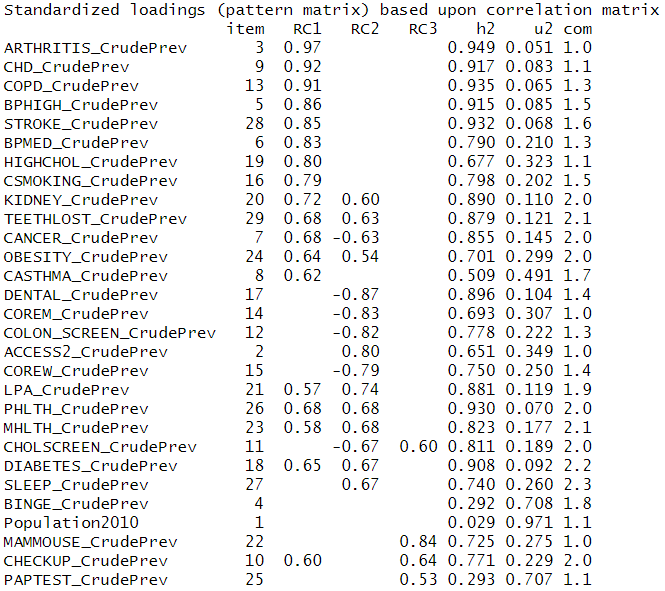


*Biplot for the PCA in figure. Certain variables, shown in circles, correlate with both factors.*

The variables in the red circle are: stroke, chronic kidney disease, and diabetes. The variables in the light blue circle are: cancer, and cholesterol screening. These variables correlate with both factors. Without these factors the perpendicular nature of the factors can be seen. There are a lot of variables on the right side that appear to go in the same direction towards the top right of the graph. This is most likely the reason for the multicollinearity observed by the determinant test. Overall this is a fairly well separated factoring when considering the numeric results, and marginally well separated when considering the visual analysis. There appears to be correlation with both factors from most of the variables.

Comparing the results with the total city data will verify results and observe if this grouping is true city wide. First a PCA with all variables was done.

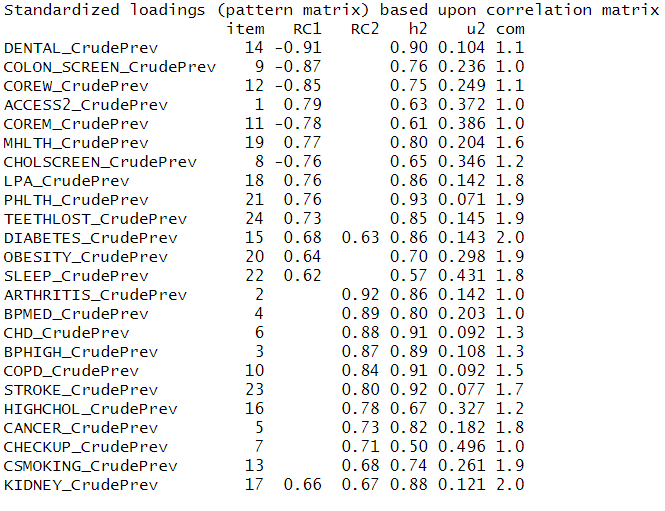
Figure 14



*The results of a PCA constructed with all variables on the total city data. Similar groupings are seen as the segmented data. Values less than 0.5 are not shown.*

The PCA from the total city data returns a similar result as the segmented data. Again population is not included in any factor. Observable is more variables that are sorted into multiple factors. Removing variables as before returned the following result.

Figure 15



*Result of PCA after removing variables of the total city data.*

Removing the variables for mammography, and pap-test were done to reduce the factors to two. Population, and binge drinking were removed because they did not sort into factors. Similarly asthma, which had previously been sorted into a factor, did not significantly correlate and was subsequently removed. The PCA of the total city data is noticeably less distinct than the segmented data. This will be explored in the conclusion section. Despite this, nearly the same variables were sorted into the same factors. Exceptions being smoking, and diabetes being switched. This is very a very interesting validation of the PCA performed on this data.

**Conclusions and Discussion**

The initial hypothesis of this project cannot be satisfactorily concluded on. Without the necessary information about population density, age, pollution statistics, and most likely a host more, the effect of population on illness rates is uncertain. However in the analysis done on the PCA model, it is unlikely that population is a major factor in illness rates or other health related issues. Population did not group in the three most important factors relating to health in cities in either the subsect-city-areas or the total city data. What remain as the most important factors are disadvantaged, under-covered populations, and age. In addition the third factor of women’s reproductive health and cancer is observed. While not a groundbreaking result it is consistent in the data and remains a reasonable interpretation.

The results of the PCA, in relation to the hypothesis, are as follows. General city health statistics will be heavily influenced by age, and socio-economic status. Providing effective care for the elderly as well as disadvantaged populations will be the most effective in contributing to a health city. Area that have similar demographics will suffer from similar disease rates following the factor analysis. Furthermore these two factors have a notable correlation with each other and likely have significant interaction, meaning they are not mutually exclusive. Certain health variables are shared by these two factors. Population does not appear to be a significant part of the health related statistics explored in this project. However this is a preliminary interpretation without the necessary information, mentioned previously, to more accurately model the problem. Population did not appear to correlate well with any variable or latent factor.

The similarity between the segmented data and the total city data corroborate each other and implicate a wider reach for these results. Having similar results means that the factors extracted for small areas of cities also apply to the whole city. The factors extracted are represented in comparisons of total cities. The total city PCA is less distinct than the segmented data results. This might be expected as individual areas tend to be more homogeneous than a whole city. People who live near to each other geographically have similar lives. If a specific area is similarly aged, and similarly wealthy they would likely have similar statistics. This would be reflected in the more distinct separation in factor analysis. While whole cities would be less homogenous. It is surprising that these factors were observed at the city level at all and indicates that these results are sound.

Additionally from the PCA results, insights into the nature of disease can be seen. Disease rates appear to follow population demographics. From the PCA result there was a significant factor that relates to a poorly-covered population that suffers from health issues and unhealthy lifestyles. This group does not have the health coverage or ability to treat their illnesses as a wealthier area. This is clearly a group that is in need of attention. An under-covered population is a serious concern. While age is a natural part of life, a disadvantaged population is something that can be addressed. The effect being a healthier population that suffers less from health issues and is better cared for.

**Works Cited**

Gould, C. F., ⨯ Steven, N. C., Phillips, D., Perzanowski, M. S., & Hernández, D. (2018). Soot and the city: Evaluating the impacts of clean heat policies on indoor/outdoor air quality in new york city apartments.*PLoS One, 13*(6) doi:http://dx.doi.org/10.1371/journal.pone.0199783

Han, L., Zhou, W., Pickett, S. T. A., Li, W., & Qian, Y. (2018). Multicontaminant air pollution in chinese cities.*World Health Organization.Bulletin of the World Health Organization, 96*(4), 233-242,242A-242E. doi:http://dx.doi.org/10.2471/BLT.17.195560

Lima, d. S., Cícero, J. V., Ícaro Fillipe, d. A., de Amorim, É. M., André Severino, d. S., Thiago Moura da, R. B., . . . Rohde, C. (2018). 0RW1S34RfeSDcfkexd09rT2Drosophila melanogaster1RW1S34RfeSDcfkexd09rT2 as model organism for monitoring and analyzing genotoxicity associated with city air pollution.*Environmental Science and Pollution Research International, 25*(32), 32409-32417. doi:http://dx.doi.org/10.1007/s11356-018-3186-5

Pyne, S. (2002). Small particles add up to big disease risk.*Science, 295*(5562), 1994. Retrieved from http://libproxy.rpi.edu/login?url=https://search.proquest.com/docview/213596172?accountid=28525

**References**

1 - <https://www.nlc.org/sites/default/files/field/image/la-smog.jpg>

2- <https://github.com/ChristianWesselborg/Data-Analytics-Project>

**Packages**

MASS\_7.3-51.4

maptools\_0.9-5

psych\_1.8.12

GPArotation\_2014.11-1

corpcor\_1.6.9

forcats\_0.4.0

stringr\_1.4.0

sp\_1.3-1

dplyr\_0.8.3

purrr\_0.3.2

readr\_1.3.1

rgdal\_1.4-7

rgeos\_0.5-2

tidyr\_0.8.3

tibble\_2.1.3

tidyverse\_1.2.1

ggplot2\_3.2.1