Vidhu Nath
CS 591 DM

Analysis of WHO Data

The data set I will be utilizing is a comprehensive compilation of healthcare, economic, and societal indicators for every nation in the world. Consisting of 358 columns containing the different indicators, with 203 rows containing every country (listed in alphabetical order without any rearranging), the WHO data set holds a plethora of information. Additionally, while the data set is primarily focused on the health issues of the countries, it was originally a combination of data sets from the WHO health survey as well as from statistics from the World Bank and the IMF; thus, there are many economic indicators which can be used to provides analysis for the reasons behind the disparities in healthcare between countries.

All actual data values are stored as ints or floats; with the exception of the country names, all other data are represented numerically. This is due to the fact that the columns are detailed and lend themselves to a numerical representation. For the purposes of my analysis, I have selected around thirty columns to work with, though some of them may not be used. There are: the names of the countries (string representations), the population of the countries; the median age of the populations; the fertility rates of the countries the mortality rates – overall, for males, for females, and for infants; the literacy rates for the countries – overall, for males, and for females; the life expectancy – overall, for males, and for females; the death statistics for different kinds of cancer – breast, cervical, colon, liver, lung, prostrate, and stomach; the prevalence of several diseases – HIV/AIDS and tuberculosis; the number of physicians; the amount of foreign aid provided; the percentage of the GDP spent on healthcare; the amount of income per person; the prevalence of contraceptives, the amount of internet access for the population, and the amount spent on the military as a percentage of the total government expenditure.

With these columns/attributes selected, there are three main hypotheses that I want to test. First, does a higher literacy percentage of the population affect the prevalence of preventable diseases, primarily HIV/AIDS and tuberculosis? The justification for this that I believe the literacy rate to be the most general, but accurate, representation of a population's average education level. The higher the rate, the more educated the population. Naturally, then, I wold expect that higher literacy rates means smaller incidences of diseases which are preventable due to adequate education about vaccines and the causes of the diseases; this hypothesis could be verified or negated by several plots of the HIV/AIDS prevalence compared to the literacy rates. I would the run a linear regression to extrapolate what correlation would be present between the comparisons.

The second hypothesis that I want to examine is whether more doctors guarantees a lower rate of mortality for a population – for males, females, and infants – and if that also correlates to how much the country spends on its healthcare versus some other expenditure, like the military. This analysis would be interesting because I think it would be indicative of where the country's interests lie: more developed nations would probably have a higher percentage of doctors in their countries, but they might also have a smaller military expenditure (the US being the obvious outlier to that generalization). Also, there are a few other indicators regarding supportive medical staff like nurses and examiners. That could be factored in to give a more comprehensive look at how robust a country's healthcare system is. This could be analyzed with a logistic regression to see how the expenditure on healthcare affects the number of physicians.

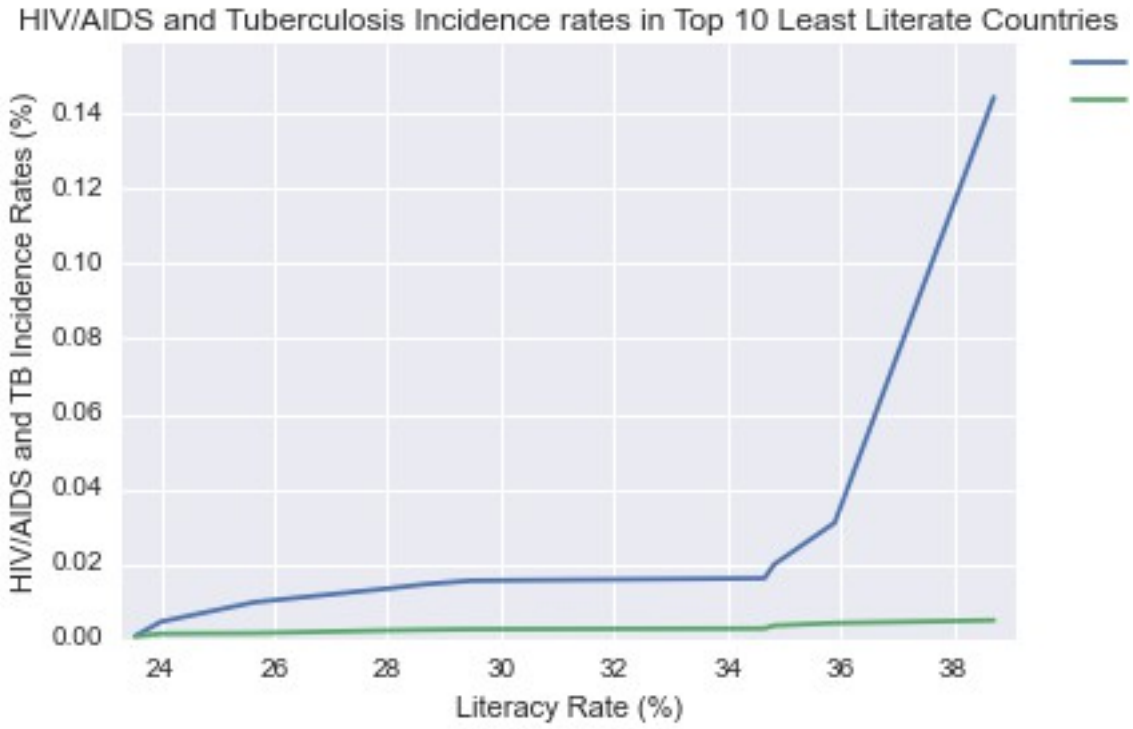
My final experiment is determining which types of cancer cause the highest rates of mortality among the overall population, among males, and among females within the most developed countries, and the least developed countries. Since I have around forty attributes related to the seven cancer types I mentioned previously such as the rates within specific population segments, this experiment would be more comprehensive than the other two, but would provide some of the more interesting results. In

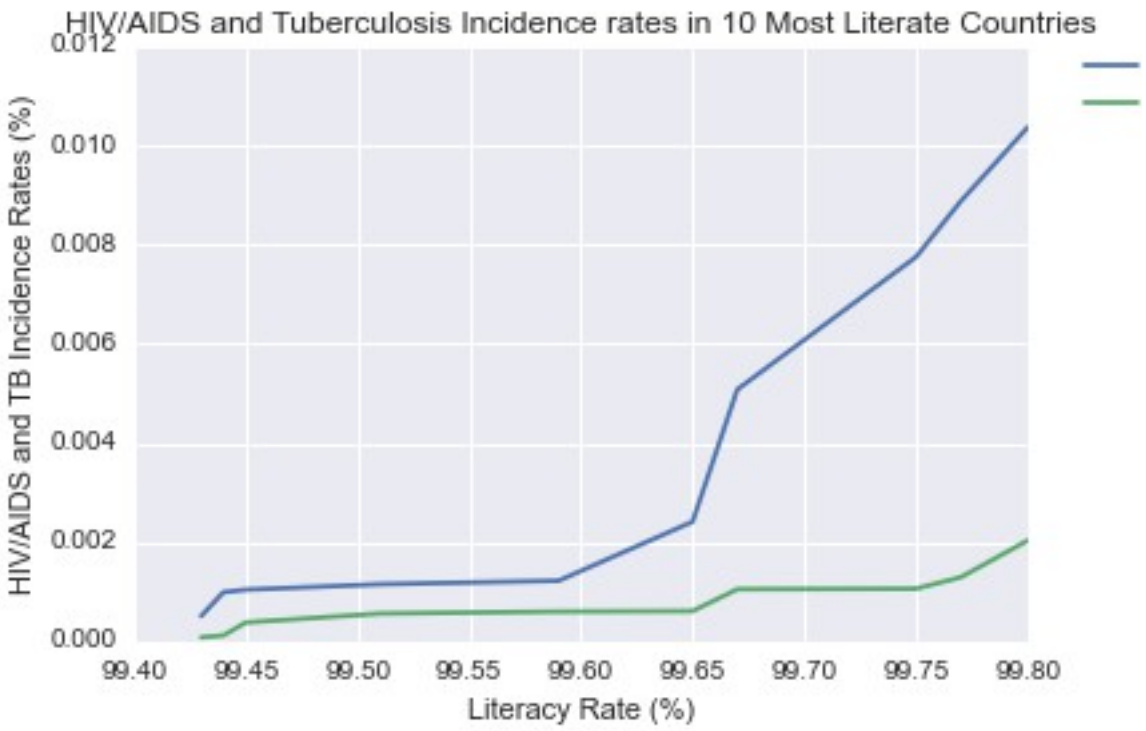
truth, I have very little idea as to which cancer is the most prominent in which country or even region of the world, so this experiment for me would be informative. Using clustering would be most effective, since I want to see which countries are most affected by a certain type of cancer. Possibly, the clustering could even lead to a geospatial map of the prevalence of different types of cancer.

For majority of the attributes I have selected, the data is filled out, so there isn't much of a worry about filling in missing data values. In the instances where I do encounter NaN values, I plan on inserting the median value of the column into those NaN. I would choose the median because it is insensitive to outliers, and with a data set this large (~50,000 values), I can't guarantee that there are no outliers. Using the mean would make the data range sensitive to its outliers, and that would potentially skew the results. Another interesting aspect is that a lot of the data are represented as percentages of the population; this means that if I want to work with the values of the attribute in terms of population counts, I will have to normalize the percentages (since they are not out of the entire population, but instead from a surveyed portion).

Most of the data will be stored in a DataFrame for easy access. Individual columns would be made into Series for manipulation, and then added back into the main DataFrame. I plan on making copies of the original DataFrame so that the integrity of the values remains intact. In the case that one of my hypotheses doesn't work out as well as I would have initially expected, I plan on developing the other two more in depth to make up for that difference. Additionally, other columns that I didn't mention above may be used, since there are a variety of intriguing ones like the amount of carbon emissions generated per person, so there could be more inclusions to the hypotheses for a more fleshed-out analysis. Overall, the testing of the three hypotheses will take priority over any other evaluation.

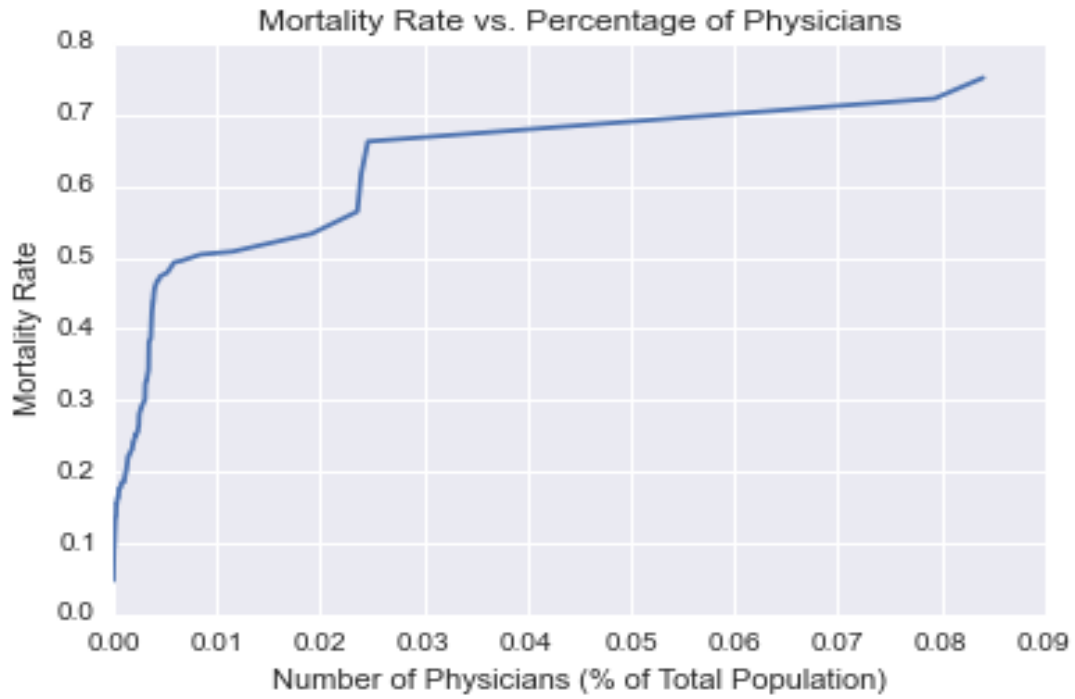
From a preliminary analysis from the first hypothesis, the following two plots were created:





Following this will be a linear regression, but also an updated plot to account for more of the countries. Obviously, with just these two graphs, it's difficult to make any immediate conclusions since there is the high potential of other factors playing a role in these outcomes (i.e. correlation does not equal causation). For now, this is what has been produced.

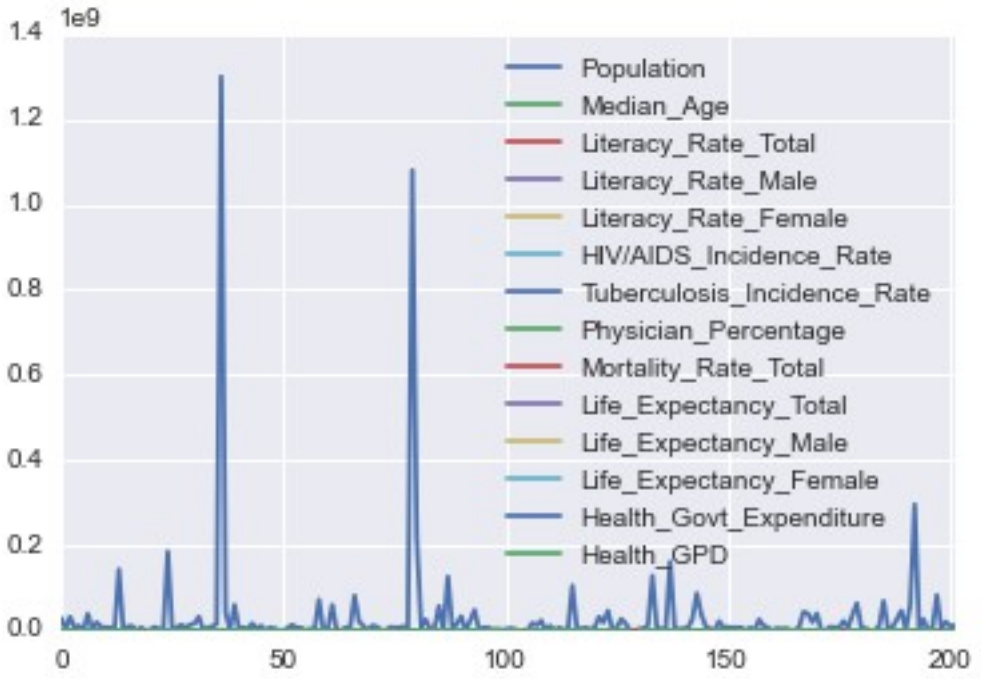
For the second hypotheses, the following has been made:



Again, correlation does not equal causation, even though it seems that higher mortality rates are associated with a higher percentage of doctors in the population. This is why a linear and logistic regression will be so useful, since it will much more clearly delineate the relationship between these two variables. Additionally, the expenditure by the government on healthcare and the military will be compared to find other conclusions.

There is no clear visualization on the cancer hypotheses since the k-means algorithm still has kinks in it. As of now, the clear choice of seven clusters will be used, but in conjunction with highest rates of mortality for the types of cancers, not for the cancers themselves.

Below is a figure representation of the new DataFrame I created to accommodate the specific columns that I want to use. This representation serves no more purpose than to just highlight the specific columns themselves; the actual plot has no meaning.



That is the current state of the project, more emphasis will be placed on achieving results by the end of the week.