World Health Organization Mortality Investigation

Bryan Arment
CSCI
University of Colorado Boulder
Boulder CO USA
brar9262@colorado.edu

Jeff Dank
CSCI
University of Colorado Boulder
Boulder CO USA
jeda7205@colorado.edu

Tony Pearo
CSCI
University of Colorado Boulder
Boulder CO USA
anthony.pearo@colorado.edu

Abstract

We utilized the World Health Organization (WHO) Mortality Database as it was a large multifactor dataset with lots of potential for real world application. As we performed our initial analysis we focused on questions relating to groups of countries and the mortality factors that could be associated with each set. We wanted to see if we could perform a clustering analysis that would segregate the countries in a way that would further allow investigation as to why the factors presented similarly in those countries.

The data chosen was the ICD10, 10th revision of the International Statistical Classification of Diseases and Related Health Problems dataset, from the WHO. Disease codes were condensed into 22 disease categories from the WHO guidelines. The data was then normed by population to give normed mortality rates. A subset of the overall dataset was selected for the analysis. The chosen dataset followed one particular disease coding methodology and was composed of 38 and 42 countries for the years 2005 and 2015 respectively. These years were chosen because they had the most countries of any of the years accounted for within the icd10 dataset. Data was pivoted into data arrays that enabled a K-means clustering analysis and subsequent principal component separation allowing for the production of a two dimensional cluster plot.

The results of the clustering analysis showed 3 separate clusters that had some common features across them. The 3 clusters our analysis identified were former Soviet countries, more traditional European countries and a variety of other well developed non-European first world nations. We then investigated the individual factors that determined where the clusters would form in the data. Normed mortality rates were averaged across disease categories for all countries as well as for the countries composing each cluster. Some of the noticeable qualities of these clusters were the sizable differences in normed all-cause mortality rates as well as the composition of the primary constituent normed disease mortality rates represented by each cluster. Mental diseases were noted as increasing for first world countries in Cluster 3 from 2005 to 2015 as well as increasing rates in circulatory disease across the same time span.

Introduction

Being able to identify the causes of mortality in a country can have important implications for public policy as well as personal health care decisions. Understanding the “water one is swimming in” aids in making healthy choices and the scope and scale of potential risk factors for you and yours. Policy makers can understand how the decisions they make have widespread consequences for both the good and the bad. The biggest problem here is the scope and scale of the problem itself. The world we live in is incredibly diverse and the risk factors mirror this. There are numerous different causes in the WHO data and this information is not easily penetrated by someone without the domain knowledge and expertise required to parse and interpret it. We were able to consolidate this data into a manageable format and performed a clustering analysis.

The question we asked was whether there was a way to split countries into groupings that all had common mortality profiles and what the makeups of these profiles were. The primary work that this analysis leaves one with is the interpretation of the clustering method and cluster distributions. Different numbers of clusters and different methods can create different distributions, and looking across these methods a better understanding of the data can be mined. We consider this an important set of questions because it will allow a few different comparisons to occur. It allows individuals to understand the risk profiles based on their location/ethnic background and it allows policymakers in countries to determine where to allocate scarce healthcare resources most efficiently. Another use case would allow countries with less successful healthcare outcomes to look at the systems in a “better” cluster and see if they can make changes to improve their own healthcare. Finally, countries that straddle cluster classifications can be identified and interpreted as moving to or from one cluster to another. The lifestyles, policies, and risk factors can be better understood to provide a snapshot of what transformation between classifications looks like as well as what it takes to be replicated.

Related Work

An immense amount of research has been done revolving around mortality statistics and other associated health studies. While it is likely that new ground will not be broken with this project, it is an interesting and expansive data set that will allow the team to apply methods and techniques covered in the scope of this course.

Mathers and Loncar did some extensive work to try and predict global mortality out to the year 2020 (Projections of Global Mortality and Burden of Disease from 2002 to 2030). They were able to split the data out by sex, age, and income. Also, they used simple regression equations to make their predictions. This type of analysis lines up well with our initial formulations when first stumbling upon the WHO database. Also, they were able to group some diseases together for their analysis such as “respiratory” and “digestive”. Lumping/clustering together similar causes of death like this is a great idea since the WHO mortality databases causes of death can be very specific.

Another paper focused on solely heart disease related deaths, but failed to make any predictions with generating any kind of models (Mortality From Ischemic Heart Disease). This was more of a survey of various countries and a discussion of the overall trends. One aspect that could be incorporated into the analysis is that they broke down the countries that were investigated into groups based on the countries GDP. Noting that the data may not be as accurate in still developing countries.

One last paper goes into a clustering method for US mortality data: Clustering of 27,525,663 Death Records from the United States Based on Health Conditions Associated with Death: An Example of Big Health Data Exploration. This paper is of interest because of the way that they group similar types of cause of death and then look into the health conditions of the sample in question like age, location, and demographics to better understand the people that died. This shift from a disease oriented approach to a person-centric approach gives a different lens to approach the dataset with and also goes into structuring a method that we ourselves can apply.

Data set

The database has been compiled by the WHO and contains number of deaths by country, year, sex, age group and cause of death as far back from 1950. Data are included only for countries reporting data properly coded according to the International Classification of Diseases (ICD). Our specific data consists of two large sets that have data for the 10th revision of the ICD. The database itself can be downloaded from here: Download the raw data files of the WHO Mortality Database

The 10th revision of the ICD lists over 100 separate causes of death each with numerous sub categories, breaks the age group down into one year buckets for ages five and under and then 5 year buckets for ages above five. The data is taken as reported to the WHO with a warning that it may not be complete for all countries.

Tools

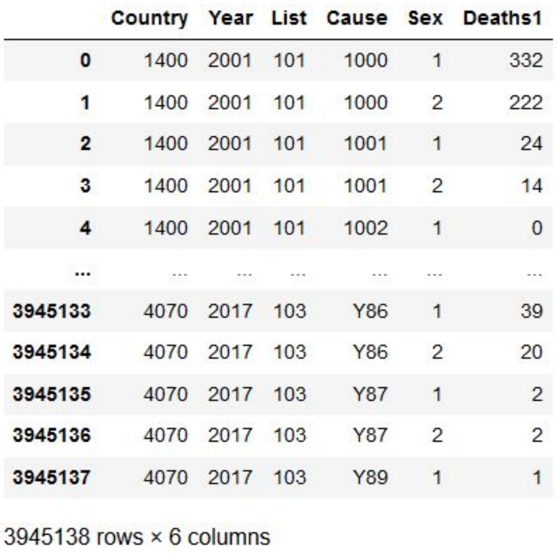
Python is a good candidate to perform the brunt of the analysis. Pandas is a data analysis library available to Python that can read in large datasets, slice and dice them to how we see fit, and create clusters and basic regressions. Python can also be used for its various graphing libraries such as matplotlib to create custom visualizations from our results.

Tableau is a useful tool to perform quick data visualization, and we took advantage of this to put together some basic choropleths detailing the countries that were present in our trimmed down dataset. An example of Tableau output can be seen in the website tracking Covid-19 cases in Colorado: https://covid19.colorado.gov/data/case-data.

Excel can be used for quick sanity checks if needed or to create some simple plots while we are working with our dataset. Tables were created from some of the final data arrays. It's just faster to use before making things more pretty with Tableau.

Main Techniques Applied

As mentioned above we underwent a lot of cleaning and parsing of the dataset. This was a significant amount of work due to the volume of data we were working with (4 million rows by 40 columns to start). We generated a couple different files we worked off of. The first was the combined and reduced ICD10 data frame directly from the WHO site. This ICD10 data frame is shown below:



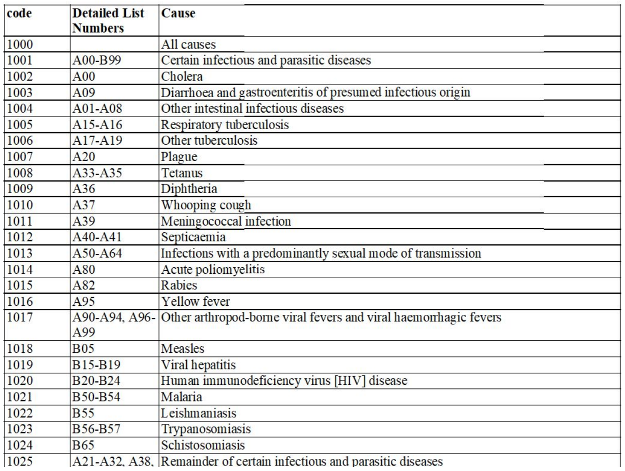
	Country	Year	List	Cause	Sex	Deaths1
0	1400	2001	101	1000	1	332
1	1400	2001	101	1000	2	222
2	1400	2001	101	1001	1	24
3	1400	2001	101	1001	2	14
4	1400	2001	101	1002	1	0
...	...	...	...	...	...	...
3945133	4070	2017	103	Y86	1	39
3945134	4070	2017	103	Y86	2	20
3945135	4070	2017	103	Y87	1	2
3945136	4070	2017	103	Y87	2	2
3945137	4070	2017	103	Y89	1	1

3945138 rows × 6 columns

As you can see we have reduced the initial dataset significantly and now have our key data points in columnar form. This is a good intermediate state for our data as we still have all the countries and all the years but we have consolidated the ages into a total and removed any Nan values up to this point. From here we further sliced and diced in order to answer specific questions.

Next we added in population data by country by year and used that datapoint to norm the death amounts by the total population. This gave a normalized set of data where each cause had a death number that was a percentage of the total population. This removes the noise in the data and the potential for large countries to dominate smaller countries while doing analysis. This was especially relevant for our clustering work as those types of models require normed data.

The biggest issue we had with this data was we needed to come up with a way to consolidate the causes of death across the columns into a manageable amount. There were around 3000 columns which included every value from A00 to Z99 plus a couple other one off formats wel dropped for consistency. Each code was super specific as you can see in the below snip from the data description:



code	Detailed List Numbers	Cause
1000		All causes
1001	A00-B99	Certain infectious and parasitic diseases
1002	A00	Cholera
1003	A09	Diarrhoea and gastroenteritis of presumed infectious origin
1004	A01-A08	Other intestinal infectious diseases
1005	A15-A16	Respiratory tuberculosis
1006	A17-A19	Other tuberculosis
1007	A20	Plague
1008	A33-A35	Tetanus
1009	A36	Diphtheria
1010	A37	Whooping cough
1011	A39	Meningococcal infection
1012	A40-A41	Septicaemia
1013	A50-A64	Infections with a predominantly sexual mode of transmission
1014	A80	Acute poliomyelitis
1015	A82	Rabies
1016	A95	Yellow fever
1017	A90-A94, A96-A99	Other arthropod-borne viral fevers and viral haemorrhagic fevers
1018	B05	Measles
1019	B15-B19	Viral hepatitis
1020	B20-B24	Human immunodeficiency virus [HIV] disease
1021	B50-B54	Malaria
1022	B55	Leishmaniasis
1023	B56-B57	Trypanosomiasis
1024	B65	Schistosomiasis
1025	A21-A32, A38,	Remainder of certain infectious and parasitic diseases

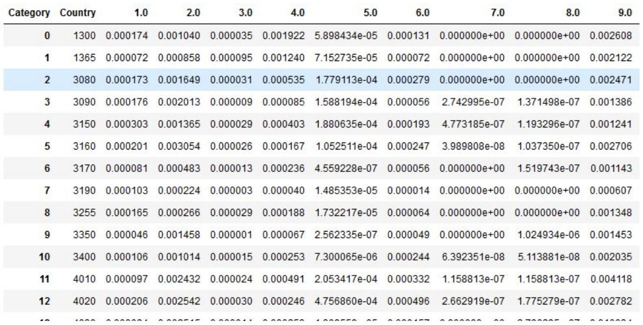
From the above images, you can see that there are many types of causes of death, lists, and country codes. In order to get a handle on what exactly was in the entire database, one of our python Jupyter notebooks was used to append distinct values for columns of interest to an empty list. This technique allowed us to determine that there were five separate values for the column “list”. The WHO uses 5 separate list types that have their own entries for detailed causes of death. For example, one list may use the code “AAA” for all deaths lumped together while another list uses the code “1000” for all deaths. The five liest types are 101,103, 104, 10M, and UE1. From here, we wanted to determine exactly what countries used each of these lists. A small bit of logic was used to append distinct country codes to empty lists corresponding to each of the WHO lists. Upon doing this, we realized that only one country used list UE1, one country used list 101, two countries used list 10M, 15 countries used list 103, and 86 countries used list 104. Since a vast majority of the countries used list 104, we decided to drop the other lists so that we did not have to worry about matching up various lists causes of death to each other.

With just one WHO list, list 104, to work with, we now needed to condense the causes of death into something more manageable. From this website, https://icdlist.com/icd-10/guidelines/ , we followed the grouping strategy based upon chapter. In WHO list 104, using the same list appending strategy for causes of death, it was discovered there were 9,838 distinct causes of death. Grouping these thousands of different causes based upon the icd-10 guidelines, we were able to get a more manageable amount of causes of 22 groups. The first 21 groups corresponding to the 21 chapters from the above website, and the last group we had as total deaths.

At this point the ICD10 data frame contained data that was normed for population and had all the cause codes remapped. This set of data is the “data warehouse” used to complete the rest of the project

We then took our data warehouse and boiled it down further into our “data cubes” in order to have a manageable way of working with the data. In doing so we avoided doing time series analysis that was scoped out due to deadline limitations. The example below shows 2015 only with a row for each country, a column for each cause of death and then the data in the columns is a normed death number calculated by taking the deaths in each cause/ total deaths in the country for the year. We also replaced the

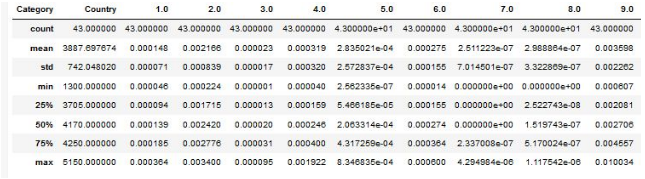
NaNs with zero values for purposes of this initial analysis. See below:



Category	Country	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0
0	1300	0.000174	0.001040	0.000035	0.001922	5.898434e-05	0.000131	0.000000e+00	0.000000e+00	0.002608
1	1365	0.000072	0.000858	0.000095	0.001240	7.152735e-05	0.000072	0.000000e+00	0.000000e+00	0.002122
2	3080	0.000173	0.001649	0.000031	0.000535	1.779113e-04	0.000279	0.000000e+00	0.000000e+00	0.002471
3	3090	0.000176	0.002013	0.000009	0.000085	1.588194e-04	0.000056	2.742995e-07	1.371498e-07	0.001386
4	3150	0.000303	0.001365	0.000029	0.000403	1.880635e-04	0.000193	4.773185e-07	1.193296e-07	0.001241
5	3160	0.000201	0.003054	0.000026	0.000167	1.052511e-04	0.000247	3.989808e-08	1.037350e-07	0.002706
6	3170	0.000081	0.000483	0.000013	0.000236	4.559228e-07	0.000056	0.000000e+00	1.519743e-07	0.001143
7	3190	0.000103	0.000224	0.000003	0.000040	1.485353e-05	0.000014	0.000000e+00	0.000000e+00	0.000607
8	3255	0.000165	0.000266	0.000029	0.000188	1.732217e-05	0.000064	0.000000e+00	0.000000e+00	0.001348
9	3350	0.000046	0.001458	0.000001	0.000067	2.562335e-07	0.000049	0.000000e+00	1.024934e-06	0.001453
10	3400	0.000106	0.001014	0.000015	0.000253	7.300065e-06	0.000244	6.392351e-08	5.113881e-08	0.002035
11	4010	0.000097	0.002432	0.000024	0.000491	2.053417e-04	0.000332	1.158813e-07	1.158813e-07	0.004118
12	4020	0.000206	0.002542	0.000030	0.000246	4.756860e-04	0.000496	2.662919e-07	1.775279e-07	0.002782

The next step was a formal Exploratory Data Analysis to go with some of the more informal analysis that has been discussed above where we made decisions about how to pare down the dataset and what countries/ time periods to explore further. We took our data cubes and ensured that there were no glaring statistical discrepancies that would corrupt the results of our modeling work. The description of the exploratory data analysis below was done on both the 2005 and 2015 data cubes but for brevity only the 2015 data is discussed/shown here. The results and conclusions were very similar.

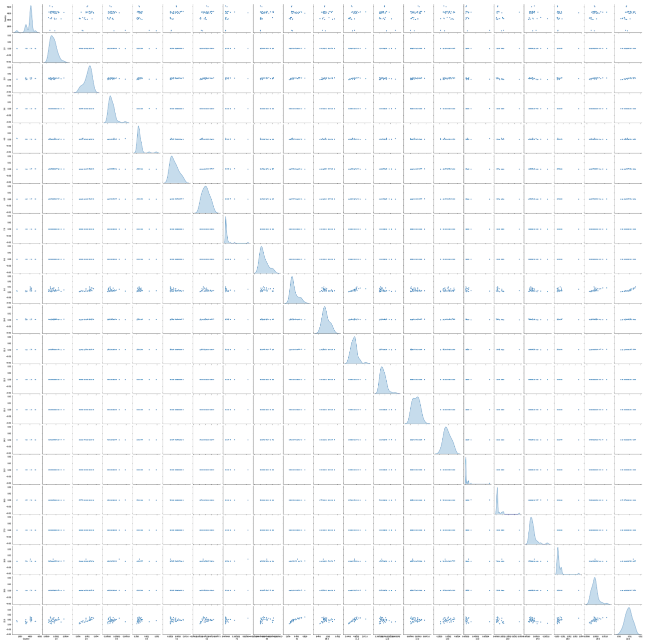
The first step was to generate some simple summary statistics using python. A subset of that is shown here but can be found in detail in our main code repository:



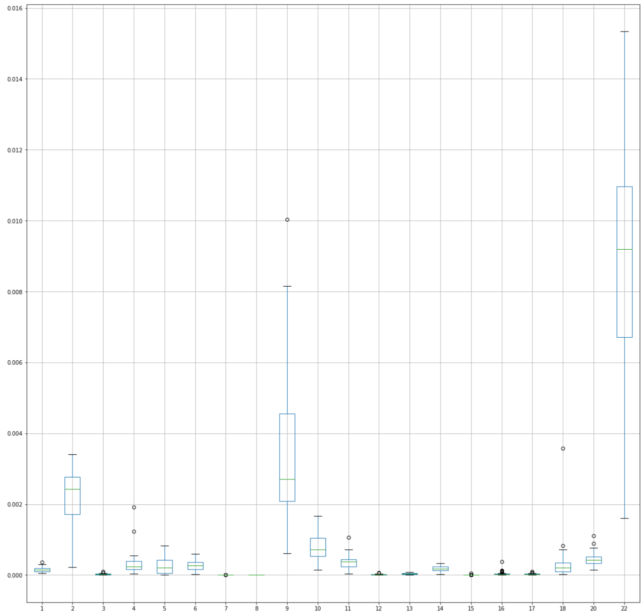
Category	Country	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0
count	43.000000	43.000000	43.000000	43.000000	43.000000	4.300000e+01	43.000000	4.300000e+01	4.300000e+01	43.000000
mean	3887.697674	0.000148	0.002166	0.000023	0.000319	2.835021e-04	0.000275	2.511223e-07	2.988864e-07	0.003598
std	742.048020	0.000071	0.000839	0.000017	0.000320	2.572837e-04	0.000155	7.014501e-07	3.322869e-07	0.002262
min	1300.000000	0.000046	0.000224	0.000001	0.000040	2.562335e-07	0.000014	0.000000e+00	0.000000e+00	0.000607
25%	3705.000000	0.000094	0.001715	0.000013	0.000159	5.466185e-05	0.000155	0.000000e+00	2.522743e-08	0.002081
50%	4170.000000	0.000139	0.002420	0.000020	0.000246	2.063314e-04	0.000274	0.000000e+00	1.519743e-07	0.002706
75%	4250.000000	0.000185	0.002776	0.000031	0.000400	4.317259e-04	0.000364	2.337008e-07	5.170024e-07	0.004557
max	5150.000000	0.000364	0.003400	0.000095	0.001922	8.346835e-04	0.000600	4.294984e-06	1.117542e-06	0.010034

The results showed that the data was pretty consistent and a survey of the max's confirmed our norming worked appropriately as they are all small percentages.

Next a pairplot was generated to identify any correlation amongst the different features and to see how the data was distributed. The chart has each of the causes paired up with each other as well as a diagonal showing the distribution of each cause factor. What this showed us is that in general the data is normally distributed in each category and that there are mostly no correlations between the categories. We decided that where the data was skewed to leave it alone as we ended up not running regression analysis where normal distribution is critical. Similarly we since only a few factors were correlated we left them separate since there wasn't an easily identifiable reason for the correlation and the magnitude of it was quite small.



As part of the pairplot analysis, we did note that a few of the factors had the potential for extreme outliers so we created some box and whisker charts to investigate. The main chart is included below with additional zoomed in versions included in the codebase. What the chart shows us is that there are indeed a few outlier data points, none of them are immediately identifiable as errors as they are all still small percentages (sub 1%) of the countries total populations. While this could be bad data it is more likely to be an extreme result due to some other macro event such as a terrible flu season or natural disaster. Alternatively if a disease (say Ebola) that was consolidated into one of our 20 causes was only rampant in one country it would be a huge outlier in that category but is still a valid datapoint by itself. For these reasons we decided it was more appropriate to leave the outliers in as there are no definitive rules for handling and the outliers seemed to still be appropriate for the dataset. If one of the dots was a full number say 3 indicating that 300% of the population died of that cause then that would have been something we threw out.

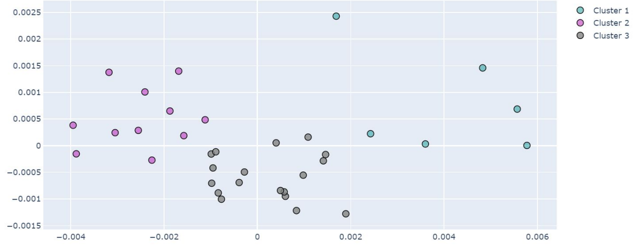


Based on the above analysis we gained more comfort with our methods and choices in paring and consolidating the large mass of data and were comfortable with beginning our clustering analysis.

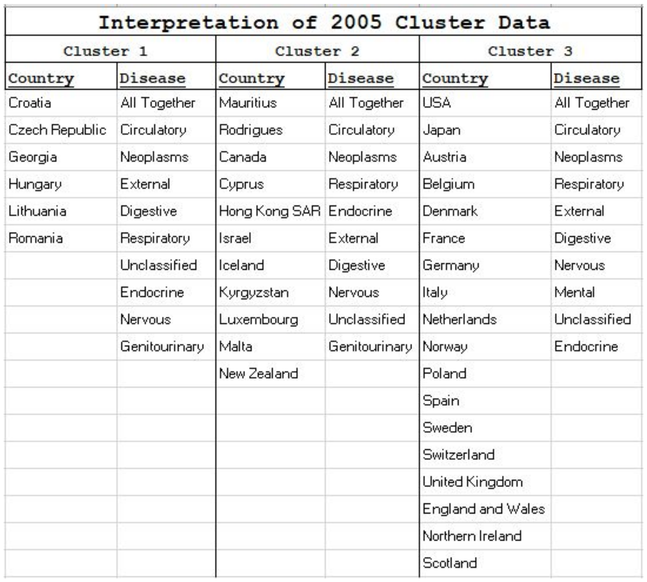
2005 K-Means Clustering Analysis

K-means clustering was chosen as an unsupervised learning method to analyze the data. The 104 list data was pivoted such that countries and categories made up the rows and columns with normed mortality rates in the cells. The functions used for this analysis were from sklearn.cluster (KMeans)

and from sklearn.decomposition (PCA). K was set to 3 for the K-Means function and principal component separation was used to create a 2-dimensional picture to represent these cluster data points. In the figure below the 2005 cluster data is presented:



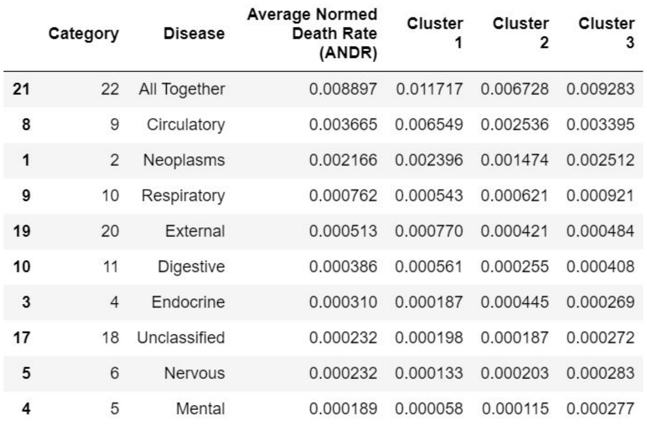
The countries represented in each cluster can be found in the following table as well as a the ten highest average mortality rates for each disease category in subsequent:



Interpretation of 2005 Cluster Data

Cluster 1		Cluster 2		Cluster 3	
Country	Disease	Country	Disease	Country	Disease
Croatia	All Together	Mauritius	All Together	USA	All Together
Czech Republic	Circulatory	Rodrigues	Circulatory	Japan	Circulatory
Georgia	Neoplasms	Canada	Neoplasms	Austria	Neoplasms
Hungary	External	Cyprus	Respiratory	Belgium	Respiratory
Lithuania	Digestive	Hong Kong SAR	Endocrine	Denmark	External
Romania	Respiratory	Israel	External	France	Digestive
	Unclassified	Iceland	Digestive	Germany	Nervous
	Endocrine	Kyrgyzstan	Nervous	Italy	Mental
	Nervous	Luxembourg	Unclassified	Netherlands	Unclassified
	Genitourinary	Malta	Genitourinary	Norway	Endocrine
		New Zealand		Poland	
				Spain	
				Sweden	
				Switzerland	
				United Kingdom	
				England and Wales	
				Northern Ireland	
				Scotland	

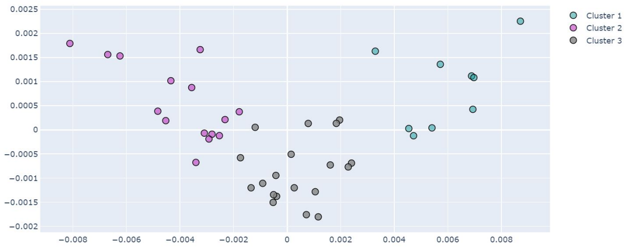
The average normed death rate data is also presented in the following table to get an idea of how each cluster compares to the averages across all data points. It is interesting to note the similarities and differences of each cluster to get an idea of why things may be grouped together in the way that they are.



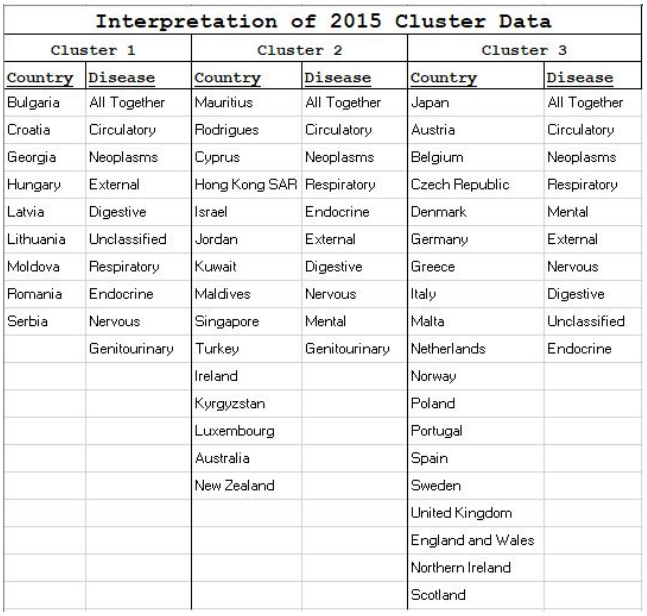
	Category	Disease	Average Normed Death Rate (ANDR)	Cluster 1	Cluster 2	Cluster 3
21	22	All Together	0.008897	0.011717	0.006728	0.009283
8	9	Circulatory	0.003665	0.006549	0.002536	0.003395
1	2	Neoplasms	0.002166	0.002396	0.001474	0.002512
9	10	Respiratory	0.000762	0.000543	0.000621	0.000921
19	20	External	0.000513	0.000770	0.000421	0.000484
10	11	Digestive	0.000386	0.000561	0.000255	0.000408
3	4	Endocrine	0.000310	0.000187	0.000445	0.000269
17	18	Unclassified	0.000232	0.000198	0.000187	0.000272
5	6	Nervous	0.000232	0.000133	0.000203	0.000283
4	5	Mental	0.000189	0.000058	0.000115	0.000277

2015 K-Means Clustering Analysis

In a similar fashion the same analysis was conducted for the 2015 data points. The cluster analysis, country/disease breakdown, and normed death rates can all be found in the following cluster and tables.



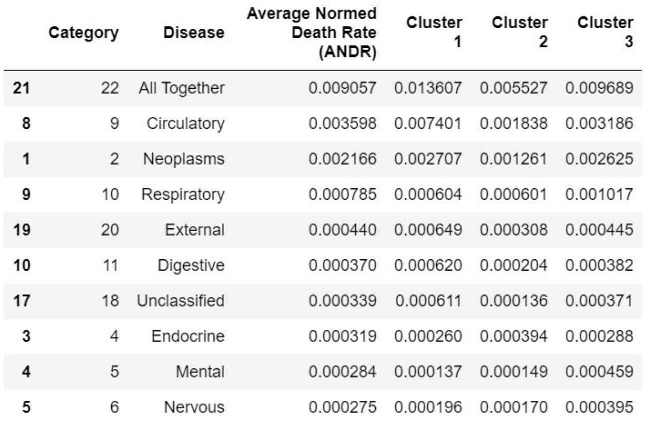
It can be seen that Eastern European countries make up Cluster 1, small high GDP countries make up Cluster 2, and traditional first world countries make up Cluster 3. Cluster 2 and 3 are quite similar in the order in which the top 5 or so average normed death rates present themselves, but Cluster 3 has higher rates of circulatory, neoplasms and respiratory deaths than Cluster 2. Cluster 1 has a much higher all cause mortality rate than the other two, as well as having a slightly different composition of diseases making up its most prominent mortality rates.



Interpretation of 2015 Cluster Data

Cluster 1		Cluster 2		Cluster 3	
Country	Disease	Country	Disease	Country	Disease
Bulgaria	All Together	Mauritius	All Together	Japan	All Together
Croatia	Circulatory	Rodrigues	Circulatory	Austria	Circulatory
Georgia	Neoplasms	Cyprus	Neoplasms	Belgium	Neoplasms
Hungary	External	Hong Kong SAR	Respiratory	Czech Republic	Respiratory
Latvia	Digestive	Israel	Endocrine	Denmark	Mental
Lithuania	Unclassified	Jordan	External	Germany	External
Moldova	Respiratory	Kuwait	Digestive	Greece	Nervous
Romania	Endocrine	Maldives	Nervous	Italy	Digestive
Serbia	Nervous	Singapore	Mental	Malta	Unclassified
	Genitourinary	Turkey	Genitourinary	Netherlands	Endocrine
		Ireland		Norway	
		Kyrgyzstan		Poland	
		Luxembourg		Portugal	
		Australia		Spain	
		New Zealand		Sweden	
				United Kingdom	
				England and Wales	
				Northern Ireland	
				Scotland	

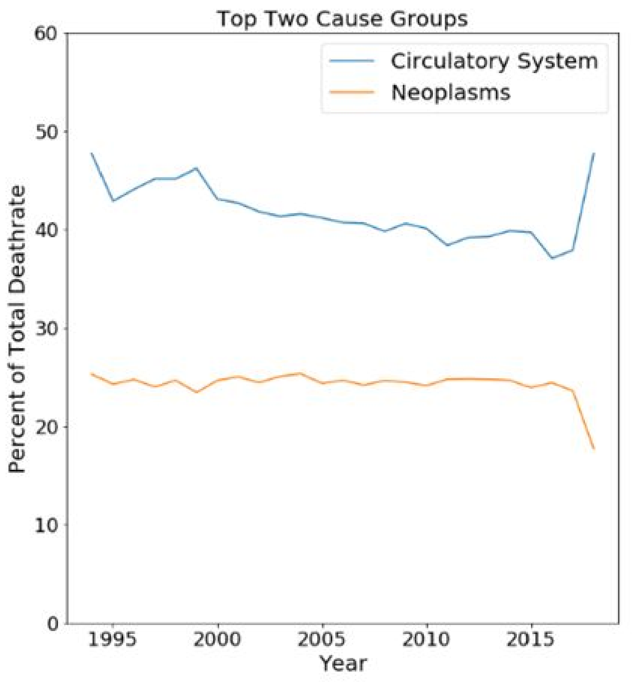
Many of our key results came from identifying patterns and relationships within these clusters and tables. The better we can understand how these clusters are formed, the better we can understand the underlying patterns or causes. Clustering method and number of clusters become important here and a high degree of domain knowledge would be required to make the best decisions. The methods and number of clusters chosen here allow us to survey the data and serve as an introduction to this data mining approach. There are some curious notions that arose in our discussion of this analysis and these notions and their implications will be discussed in the following section.

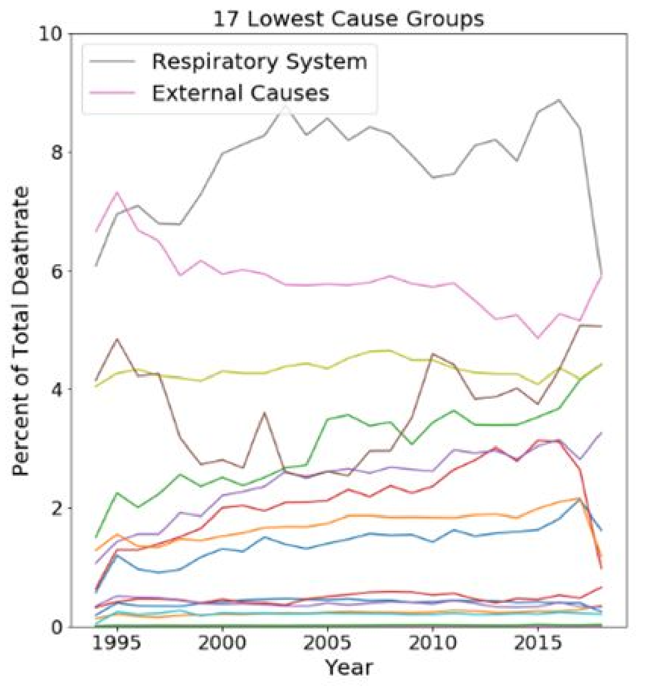


	Category	Disease	Average Normed Death Rate (ANDR)	Cluster 1	Cluster 2	Cluster 3
21	22	All Together	0.009057	0.013607	0.005527	0.009689
8	9	Circulatory	0.003598	0.007401	0.001838	0.003186
1	2	Neoplasms	0.002166	0.002707	0.001261	0.002625
9	10	Respiratory	0.000785	0.000604	0.000601	0.001017
19	20	External	0.000440	0.000649	0.000308	0.000445
10	11	Digestive	0.000370	0.000620	0.000204	0.000382
17	18	Unclassified	0.000339	0.000611	0.000136	0.000371
3	4	Endocrine	0.000319	0.000260	0.000394	0.000288
4	5	Mental	0.000284	0.000137	0.000149	0.000459
5	6	Nervous	0.000275	0.000196	0.000170	0.000395

Key Results

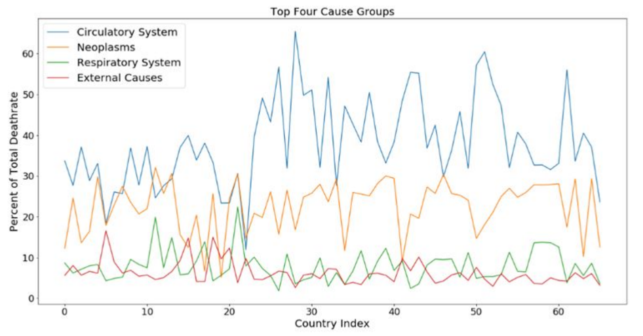
After parsing down the large amount of causes of death into the 22 groups, we then began to perform our real analysis. The final group of overall deaths was used to get a sense of the percentage of death by group and get a plot of that data over time.



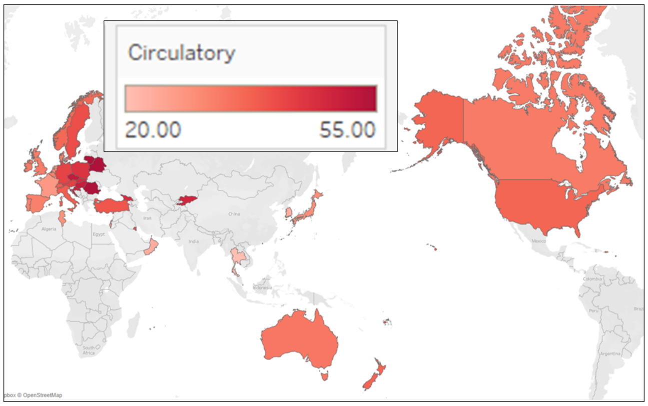


The top two groups, circulatory system and neoplasm (growths/cancers) were plotted in a separate graph because they were so far above the other groups that we were unable to see any trends in the lower groups when having all groups plotted on a single graph. The top two groups in the bottom 17 groups are specifically called out in the legend (respiratory and external). Note that there were not any death groups that showed a great variation from one year to the next. In the years 2005 and 2015, where our clustering analysis was done, the same top four causes of death group were the same four depicted in the charts above. There were two groups of deaths that did not have any entries, or may have been dropped in our earlier cleaning, corresponding to chapters 19 and 21.

The top four causes felt like a good place to start doing an analysis by country instead of by year. We sliced out the top four groups and plotted them against the country index below.



It quickly became clear that we needed to bring in Tableau for doing the country analysis. The WHO provides a country code index vs name list at the end of its documentation that accompanied our dataset. Since there were only 86 countries that used list 104, we manually entered the names of the countries next to the country id and the corresponding percentage data by group. From there, Tableau was able to take the country name and give the entire country a shade of color that corresponded to what percentage of death that particular group was.



The map of the world provides a birds-eye view of the circulatory deaths by country rather than some arbitrary index. Unfortunately, not all countries in the world use the same list to report deaths to the WHO or even provide any data to the WHO, so the only countries we had data for that was not dropped are the ones that have been shaded in. What was interesting about this view, is that you can see that Eastern European countries seemed to have the worst problem with circulatory complications, while nearby France seemed to be doing better than any other country with this cause of death group.

Looking at the data gathered from our clustering analysis there are many interesting things to note. Starting with the average normed death rates for the 22 disease categories, if we rank them from lowest to highest we can see that the largest contributors to overall mortality are circulatory disease and neoplasms across all three clusters of countries. Respiratory and external follow suit in clusters 2 and 3, but not in cluster 1 where external and digestive mortality take a precedence over respiratory. Some of the key differences between cluster 1 and the other two clusters are that the all cause mortality in cluster 1 is quite a bit higher than that of cluster 2 or 3, approximately 20% higher than cluster 3 and nearly double that at 40% greater than cluster 2. Circulatory disease is a much greater risk in cluster 1 than it is in the other two clusters and is likely a significant contributor to the increased overall mortality rate represented in cluster 1.

Comparing the 2015 data to the 2005 data, there are some surprising discoveries as well. One to notice right away is that overall mortality rate actually increases slightly in cluster 1 and 3 while decreasing by about 12% in cluster 2. Circulatory disease in clusters 1 and 3 also increases during this time span by a small percentage and probably makes up a large portion of the increases detected in all cause mortality rate. The category 'Mental' representing the normed death rate for mental diseases actually doubled across this time span as well for cluster 3, moving from position 8 to position 5 in the top ten normed death rates list above. The category unclassified also sees noticeable increases in all three clusters across this timespan, though it is hard to say if this is due to the nature of the categorization system being used now compared to then or if it points to novel health complications that have not yet been included in this documentation.

It is not our role in this paper to attempt to attribute causation to some of these patterns, but it is very interesting to note some of these shifts happening across time . It is worrisome to see health, especially mental health, to be declining in what can be considered the most well established and modern countries in the world. Especially given the recent advances in technology and medicine that one would hope to solve some of these complex health problems. If it is the case that despite our recent advancements in science and technology, that people are living less healthy lives, dying from health consequences that are in fact preventable (and more work would need to be done to verify this), then perhaps we could learn from the lifestyles of those who live in countries and clusters that live the longest, healthiest lives. The progress of modern medicine and technology is incredible, but ultimately it may be a bandaid to what is in reality a lack of self-care and community that makes this self-care possible.

Application

There are many ways in which our analysis could be useful in understanding larger and more complex relationships in mortality across the world.

This analysis could be used to further investigate why some countries fare better than others in regards to particular categories of mortality risk. For example, France could be used as a case study to see if there is anything that their population is doing to lower the risk of circulatory disease. This could be from lifestyle, environmental, or dietary factors. Countries like France that have the lowest risk from particular mortality factors could serve as role models and pave the way for other peoples, whether that be motivated individuals, health network systems, or countries/national groups to make necessary changes in themselves and their environments to better emulate the success that is evident. On the other hand, countries that are failing in regards to ebbing the curve in increasing circulatory illness may be able to identify the problem before it gets too out of hand and begin to take preventative measures.

The work that we have done using our K=3 means clustering analysis can easily be extended to the rest of the data within the WHO mortality set. We only worked with disease codes following the 104 guidelines, but other guidelines such as 103 and 1000 can easily be extended to by mapping the subcategories to the overarching categories. The same holds true for moving into the icd9, icd8, and icd7 datasets that would allow a more in depth analysis stretching back to the 50s. Once these are taken into consideration a larger time series analysis would be possible and it would be fascinating to see the movement of countries between cluster classifications across time. Countries that were capable of shifting categorization could be identified and further understanding what changes had been made that allowed them to cross clusters.

Along similar lines countries that find themselves between clusters or that could be clustered into say cluster 1 or cluster 3 depending on the starting conditions of the model or whether the model using starting points or clustoids. These countries that straddle clusters could provide valuable information. Maybe they are a country that is moving from cluster 1 to cluster 3 and so reasons as to why their mortality profile are changing could be identified and used to help guide the decisions of other countries that would like to emulate or avoid the policies, behaviors, or cultural landscape that provide the impetus to such a transformation.

Another interesting application of our conducted research would be to compare to other clustering techniques to get a better picture of these cluster formations. As it stands, it is inconclusive as to whether or not this is the best way to cluster these countries together. Using another model, say density or hierarchical, could be beneficial to create a better understanding of the broad scope reasons that these clusters are being formed. While considering clustering methods it would also be interesting to consider the results of using a larger number of clusters to conduct the analysis if an expanded dataset was created. Bringing more clusters into the model with more data would surely result in the data being partitioned into many more groups. A higher number of clusters could paint a better picture of mortality rates. Other clusters of say, Caribbean islands, Southeast Asian countries, or Central America could find themselves forming in the same way that Eastern European countries clustered together in the K=3 means clustering analysis used here. The importance of this of course would be to notice the similarities and differences between the average normed mortality rates in each cluster in the same way we presented the information here.

All of this is of course to say that the real insights from clustering methods come from interpreting how and why the clusters were formed. Much more work remains to be done in this aspect. This project serves as an introduction to some of the exploratory analysis and clustering techniques necessary to better understand the health factors and risks that we all should be concerning ourselves with in our day to day decisions. What it is that we should be eating, how much we should be exercising, and the types of lives that we and our children should be living are all very much up for debate. What is fascinating though is how much there is to learn from the diverse world that we find ourselves in. What is the most sustainable lifestyle that we can put together for ourselves, our families, and our neighbors around the world? That is the question we are left to. May our research serve as an introduction to this pursuit.