Which Lever Should the WHO Pull?

Scott Pledger

The University of Texas at Austin

Computer Science

Department of Natural Sciences

Abstract

Countries around the world are facing rapidly evolving healthcare challenges, and both international and local organizations need to make difficult decisions about where to focus their resources to maximize health outcomes. This paper examines the types of challenges faced in various regions of the world, as informed by K-Means clustering. It then examines the Pearson correlations between 31 different mortality categories and assesses the inherent dimensionality of the data using principal component analysis. Next, it creates a predictive model for life expectancy using the mortality category data and performs sensitivity analysis with SHAP values. Lastly, it performs perturbation analysis on the predictive model to determine the mortality category for which a 25% reduction would have the greatest impact on the life expectancy prediction of the model. The model determines that the category of greatest life expectancy impact after a 25% reduction would be cardiovascular diseases for 100 out of 185 countries.

Introduction

The World Health Organization was founded by the United Nations in 1948, and Article 1 of its Constitution is as follows (World Health Organization, 2020):

> *The objective of the World Health Organization (hereinafter called the Organization) shall be the attainment by all peoples of the highest possible level of health.*

Ultimately, "healthiness" is a difficult metric to measure at scale, but there are numerous proxies to indicate the improvement or decline in health around the world. The largest and most obvious is life expectancy. While it isn't a perfect metric because health is about much more than just life or death, it is the easiest to measure and therefore the most globally available.

This paper investigates the varying health challenges around the world by grouping countries based on their mortality statistics, analyzing the mortality and life expectancy trends over the past 20 years, creating a predictive model for life expectancy, and performing sensitivity analysis on the model. This analysis can be used to broadly understand the trends of certain diseases and ailments spatially and temporally, and it can help make predictions about how improvements in particular sectors of the healthcare space would impact life expectancy.

Materials and Sources

This analysis uses two main data sources:

1. Causes of Deaths around the World (Banerjee, n.d.)
2. Life expectancy at birth (years) (World Health Organization, n.d.)

The Causes of Deaths dataset includes 29 mortality categories for the years 1990 – 2019. The numbers given are the absolute number of people who died from the given ailment per country per year. After calculating the global mortality rate per category in the first and last year of the dataset, the differences are seen below. Most notably, diseases such as cardiovascular disease and neoplasms (cancer) rose 5% globally in the past three decades, and diseases most improved with modern healthcare, such as diarrheal diseases and neonatal disorders, are down 4%. A potential partial explanation that categories like cardiovascular diseases and cancers would increase over time is that as countries' economic conditions and healthcare systems improve, people who would have died from preventable illnesses are more likely to die from causes that are more difficult to treat. Furthermore, economic prosperity comes with socio-cultural and lifestyle changes that could lead to increased risk of different types (Torre et al., 2016).

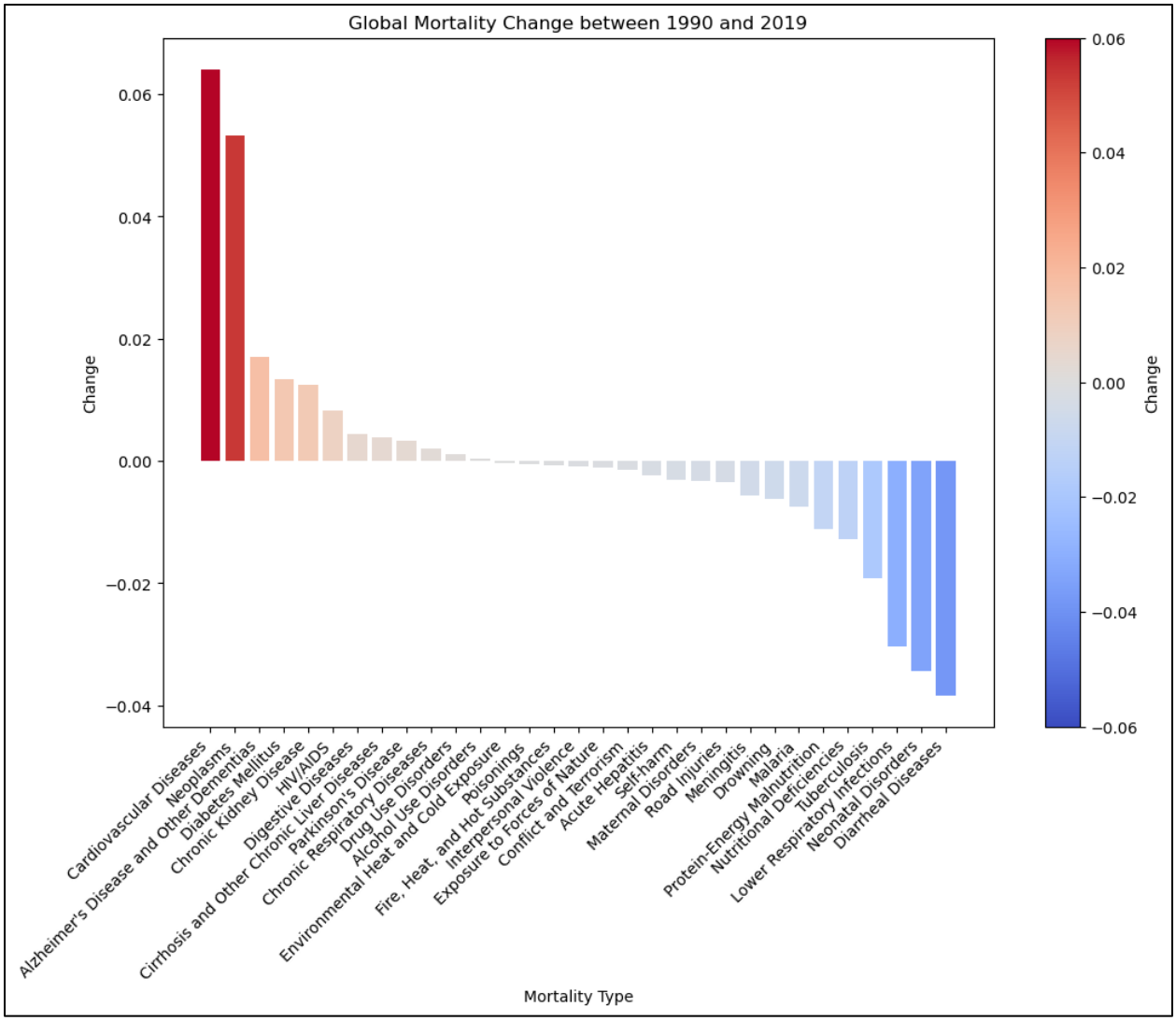

Figure 1: Global Mortality Change between 1990 and 2019

The full trends of each category over the three-decade span are shown below. This view gives context to specific events or health crises such as the AIDS epidemic, which peaked during this three-decade span in 2004, and other causes of death, such as conflict and terrorism, which sharply spiked in 1994 due to the Rwandan genocide. 500,000 deaths were attributed to conflict and terrorism in Rwanda during this year, which accounted for 83.5% of the total deaths in the country.

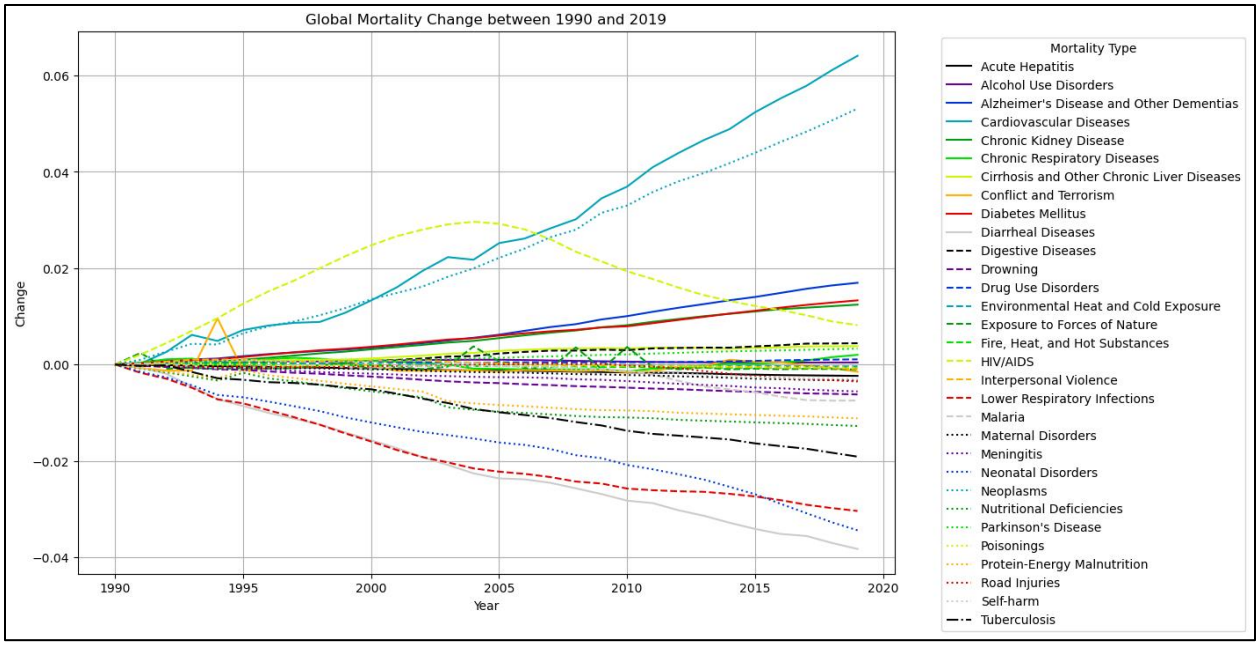

Figure 2: Global Mortality Change between 1990 and 2019

These trends and outliers demonstrate the deep connections between healthcare and world events, and they could also affect predictive models in ways that would be otherwise difficult to decipher.

The second dataset contains the life expectancy estimates by country for the years 2000 – 2021. Below is a plot of the life expectancy by country in 2021.

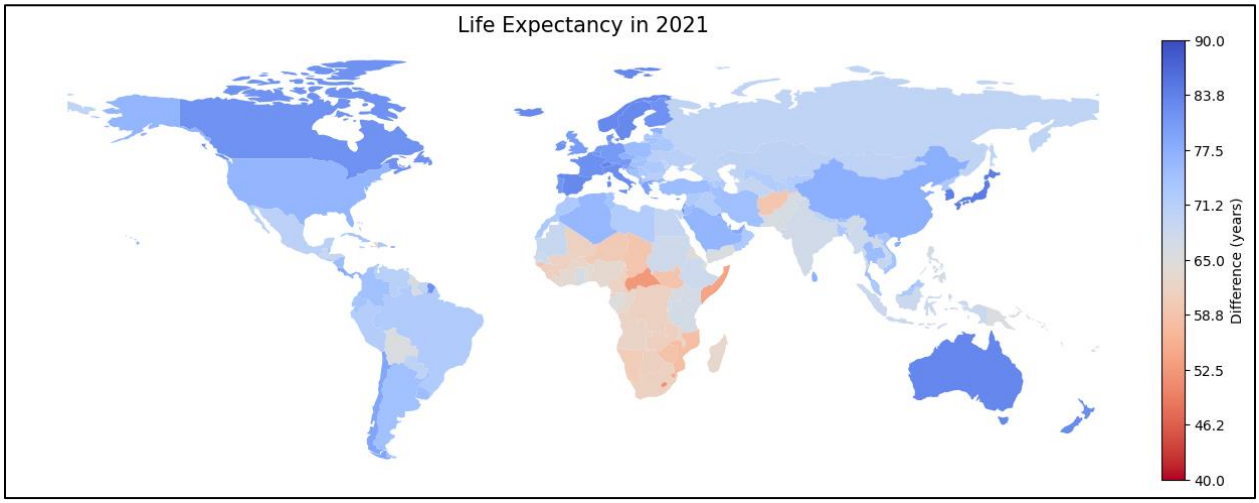

Figure 3: Life Expectancy in 2021

Notably, Sub-Saharan Africa lags most of the world in life-expectancy, but when observing the changes from 2000 – 2021, it has also made the most progress.

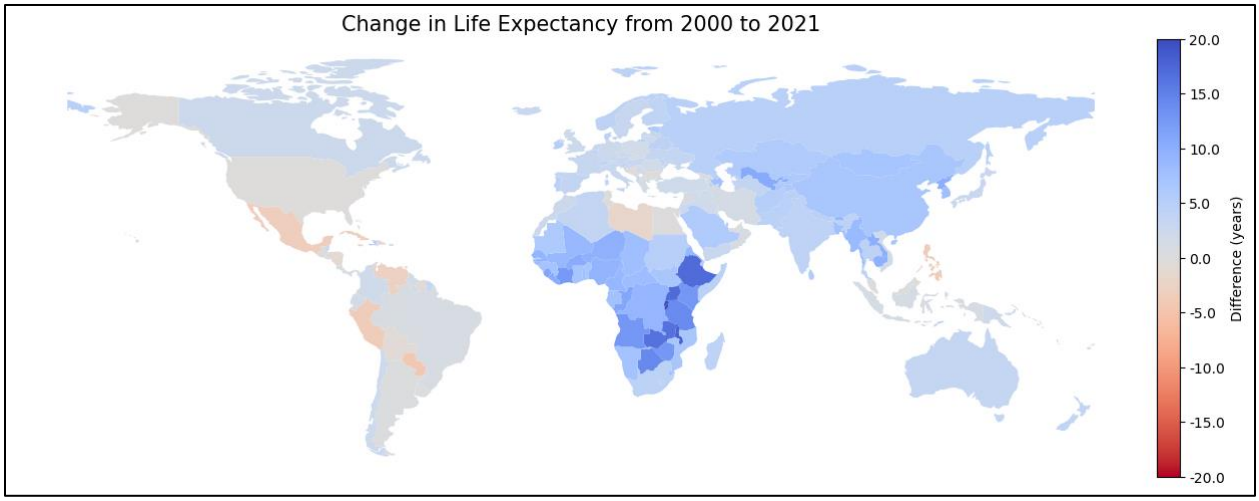

Figure 4: Change in Life Expectancy from 2000 to 2021

Interestingly, Mexico, Venezuela, Peru, Uruguay, Libya, and the Philippines have experienced decreases in life expectancy over this period.

Both the cause of death and life expectancy datasets have columns with the standardized ISO 3166 country codes, making it simple to join these datasets together to establish a relationship between mortality types and life expectancy (IBAN, n.d.). The datasets cover overlapping but different time periods, so for any years not present in both datasets, the data points were excluded from any analysis that requires both metrics.

Research and Methods

Global Health

Prior Global Health research from 1997 showed that "Five of the ten leading killers are communicable, perinatal, and nutritional disorders largely affecting children" (Murray & Lopez, 1997). When childhood mortality rates are high, it has an outsized impact on overall life expectancy calculations by significantly decreasing the mean. This is largely observed in developing countries, where "98% of all deaths in children younger than 15 years" occur (Murray & Lopez, 1997). This is why Sub-Saharan African continues to have a lower average life expectancy than the rest of the world, but it's also why it has been able to make the largest gains in life expectancy over the past 20 years. As these countries develop and gain access to better healthcare, sanitation, and food processing techniques, the childhood mortality rates have dropped, leading to gains in life expectancy up to 20 years in some countries, such as Ethiopia.

On the other side, developed countries have dramatically fewer deaths due to preventable and treatable diseases due to improved access to and quality of healthcare. This has led to life expectancies of approximately 83 to 84 years old in countries such as Japan, Singapore, South Korea, and Switzerland. In the Causes of Death data, however, these countries experience a proportional increase in diseases associated with longer lifespans, such as cancer and dementia.

While the treatments for cancer have improved over the years, the incidences of cancer have increased as well.

The global mortality type distribution is shown below.

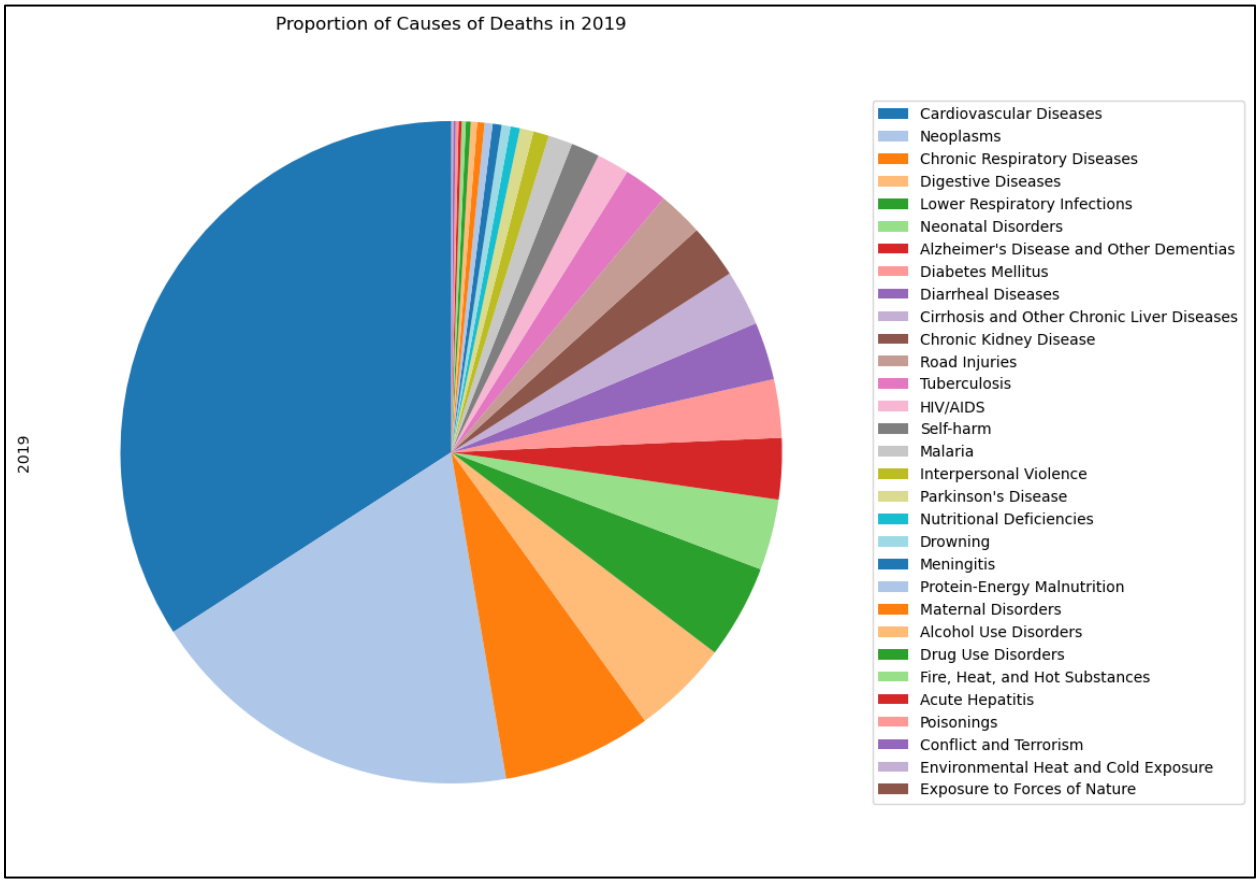

Figure 5: Proportion of Causes of Deaths in 2019

Encompassing over 50% of the deaths in the world are cardiovascular diseases and cancers, and given that these proportions are even higher in developed countries, it would be reasonable to expect continued growth as more countries experience economic development.

Correlation Analysis

Reviewing the mortality type features, we can examine the relationships between the different categories to understand how strongly associated they are with one another. The easiest way to do this is to create a correlation matrix using the pandas .corr() function to calculate the Pearson correlations between each feature (Freedman et al., 2007). One limitation of this approach is that correlation analysis assumes linearity, but in this context, the relationships are less likely to contain complex relationships (Janse et al., 2021).

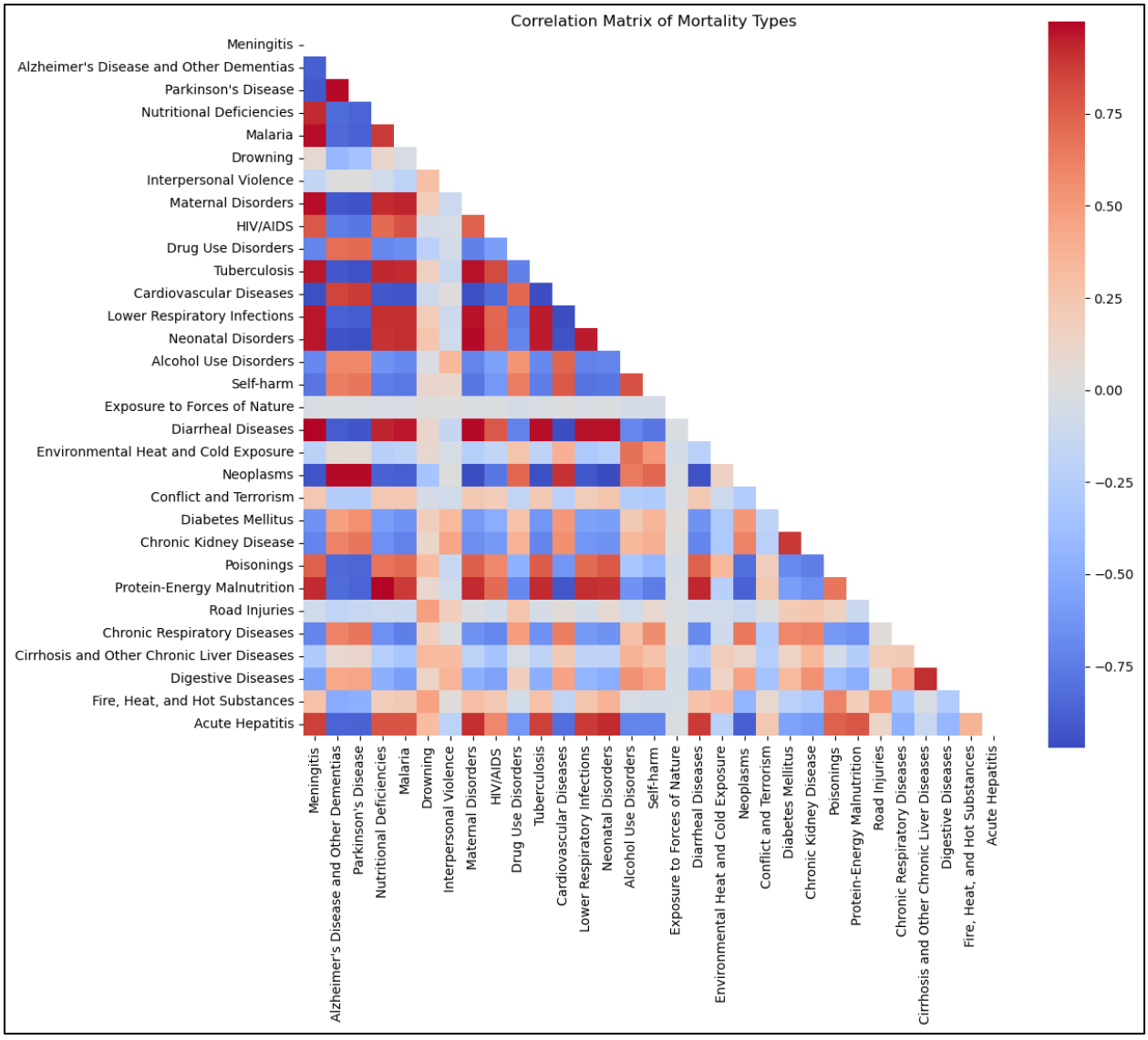

Figure 6: Correlation Matrix of Mortality Types

The first thing that stands out about this matrix are the noticeable gray lines indicating that those mortality types are not strongly associated with any other mortality types. Logically, it makes sense that these categories: Drowning, Interpersonal Violence, Exposure to Forces of Nature, and Road Injuries, are less connected to healthcare and more related to other societal or geological dynamics. Some other notably uncorrelated features are Conflict and Terrorism and Environmental Heat and Cold Exposure. One might also expect Fire, Heat, and Hot Substances and Poisonings to have very weak correlations with other features, but they still show stronger correlations than one might expect.

There are many strong correlations, however, and many of them would tend to be associated with either developed or developing countries, but not both. For example, Maternal Disorders has strong positive relationships with Nutritional Deficiencies, Malaria, Diarrheal Diseases, and Protein-

Energy Malnutrition, and it has strong negative correlations with Alzheimer's, Parkinson's, Cardiovascular Diseases, and Neoplasms. All of this makes sense given that Maternal Disorders are more likely to lead to death without access to modern OBGYN healthcare, and all the diseases it's correlated with are strongly associated with either the developing world or the developed world.

There's a large amount of information to parse on the correlation matrix, so the mortality types are sorted into supertypes below to examine the correlations at a higher level.

Mortality Supertype	Mortality Type
Infectious Diseases	Meningitis Malaria Tuberculosis HIV/AIDS Diarrheal Diseases Lower Respiratory Infections Acute Hepatitis
Chronic Diseases	Alzheimer's Disease and Other Dementias Parkinson's Disease Cardiovascular Diseases Diabetes Mellitus Chronic Kidney Disease Chronic Respiratory Diseases Cirrhosis and Other Chronic Liver Diseases Neoplasms
Nutritional and Neonatal Disorders	Nutritional Deficiencies Protein-Energy Malnutrition Neonatal Disorders Maternal Disorders
Substance Use and Mental Health Disorders	Alcohol Use Disorders Drug Use Disorders Self-harm
Environmental and External Causes	Drowning Fire, Heat, and Hot Substances Exposure to Forces of Nature Environmental Heat and Cold Exposure Interpersonal Violence Conflict and Terrorism Road Injuries Poisonings
Digestive Diseases	Digestive Diseases

Table 1: Mortality Type to Mortality Supertype Mapping

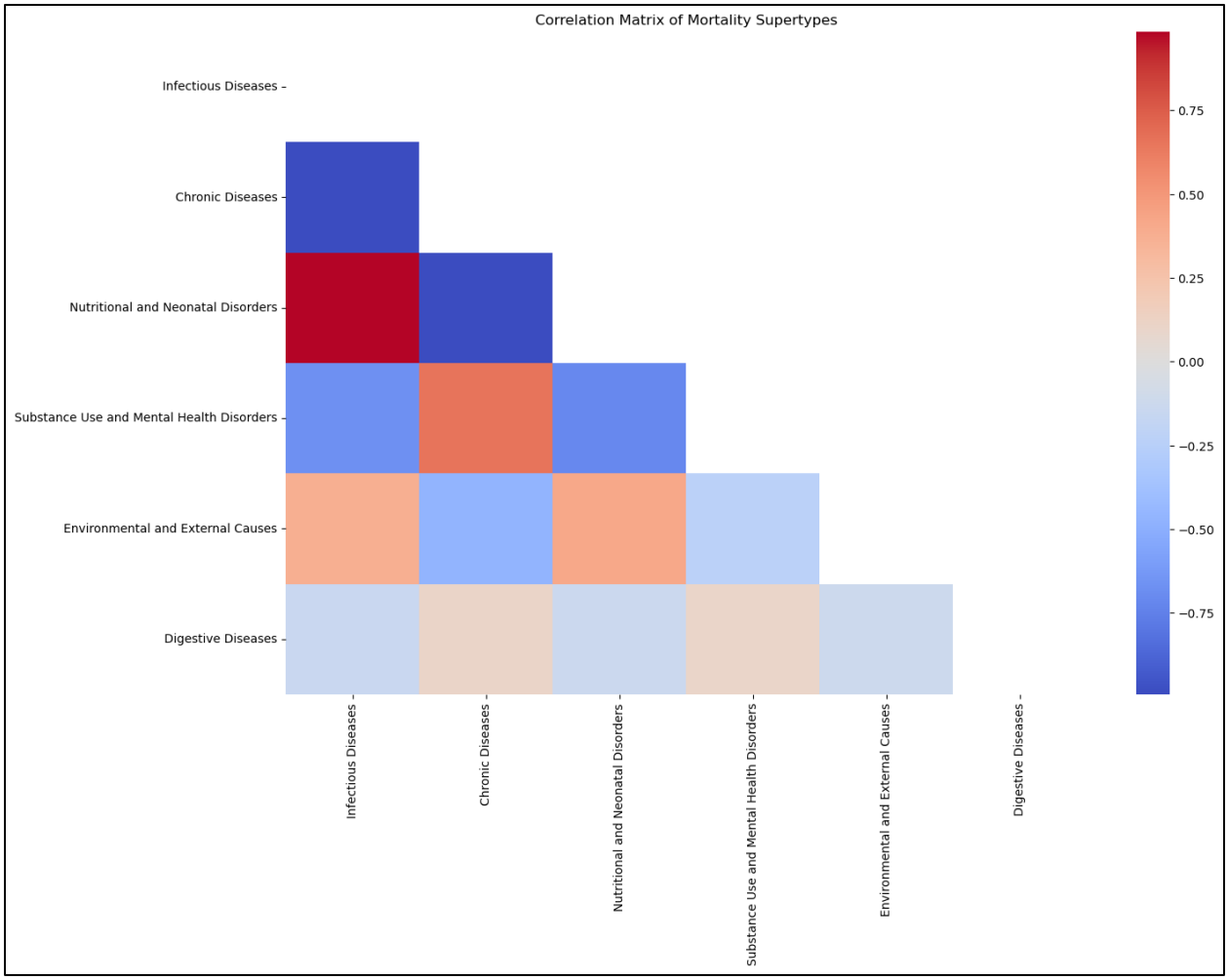

Figure 7: Correlation Matrix of Mortality Supertypes

Unsurprisingly, Infectious Diseases have a strong positive correlation with Nutritional and Neonatal Disorders, and both have strong negative correlations with Chronic Diseases. This further supports the hypothesis that these categories tend to have particularly strong correlations because of their distinct association with countries at different levels of the economic and healthcare development spectrum.

Clustering

Even countries in similar stages of economic development can experience disparate outcomes in life expectancy. Every country employs its own unique approach to healthcare – privatized vs socialized, preventative vs reactive, etc. – and they have unique cultural and societal norms that can reduce or increase risk. Some other factors include:

1. Diet
    a. Access to healthy food
    b. Organic vs GMO

    c. Natural vs processed foods
    d. Cost differential for healthy foods
    e. Access to and cultural acceptance of fast foods
2. Sanitation Systems
    a. Waste treatment
    b. Water filtration
3. Environmental Regulations
    a. Air pollution
    b. Exposure to carcinogens
    c. Microplastic prevalence
4. Safety Regulations
    a. Worker protection
    b. Transportation safety and infrastructure
5. Geography and Climate
    a. Natural disasters
    b. Sunlight exposure
6. Other Cultural Norms
    a. Active vs sedentary lifestyles
    b. Overconsumption

No two countries are exactly alike in all these categories, and there is also significant variation within each country between communities of varying socio-economic statuses and cultural norms. For example, one would expect that Moab, Utah and Brooklyn, New York would differ among many of the categories listed above even though they're both in the United States.

Understanding the similarities between countries can help organizations like the WHO to plan their initiatives to target challenges faced by a number of countries. Creating an individualized plan for every country and every problem can be quite difficult, so it could prove advantageous to identify the most significant problems faced by the largest number of countries and aggressively tackle those issues. In effect, the impact on health outcomes in the aggregate could be best achieved by identifying the points of maximum leverage. This is a commonly discussed point in the business world: don't try to do everything, choose the product or service you are best at and focus intensely on being the best in the world at that thing.

There are many unsupervised clustering techniques that could be used to identify the countries facing similar healthcare and social challenges, and for this analysis, K-Means clustering will be used (Xu & Tian, 2015; Yiu, 2019). It has been found that setting the number of clusters to be slightly larger than the expected number of classes is likely to show better performance (Rodriguez et al., 2019). After experimentation and analysis of cluster uniqueness, the number of clusters for the final analysis was lowered incrementally from 10 to ultimately settle on 5 clusters.

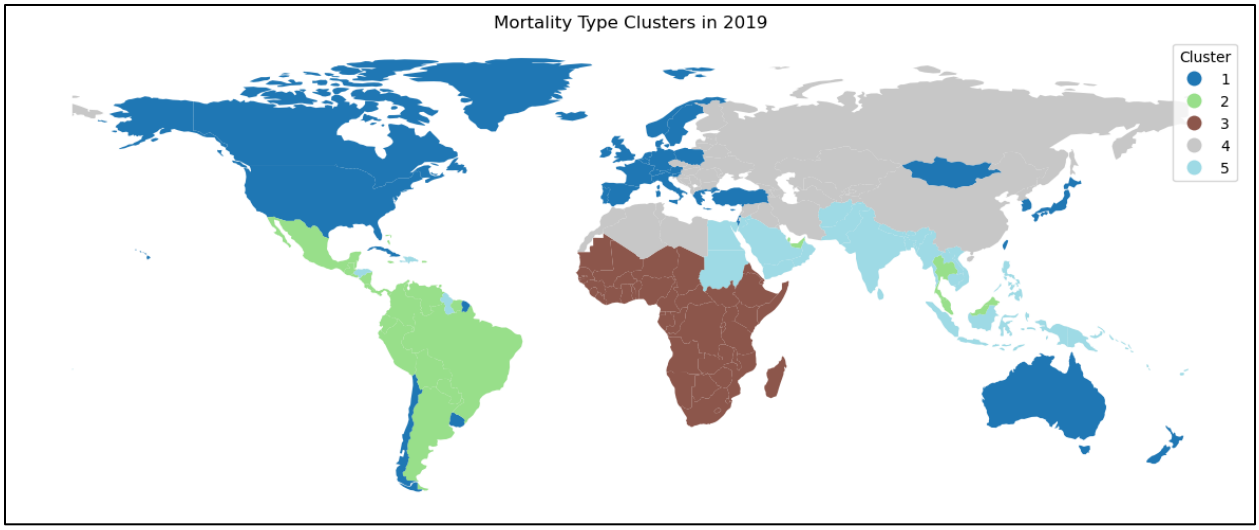

Figure 8: Mortality Type Clusters in 2019

Even though the clustering algorithm gives no consideration to geographical proximity, it nonetheless has strong geographical grouping because neighboring countries tend to have more similar cultures, wealth, and healthcare. There are some notable exceptions to this though, particularly in Cluster 1, which is dominated primarily by Western European and North American countries. Notably, it also identifies "westernized" countries from other regions of the world, such as Japan, South Korea, Australia, New Zealand, and a few countries in South America. Interestingly, the South American countries: Chile, Uruguay, and Guyana have the three highest GDP per capita metrics in South America (International Monetary Fund, 2024). Understanding how the K-Means algorithm identified these clusters can help one diagnose the problems unique to each cluster. Some new techniques use Mixed Integer Optimization to create interpretable tree-based clustering models, but in this case, we'll compare the cluster averages to the global averages (Bertsimas et al., 2020).

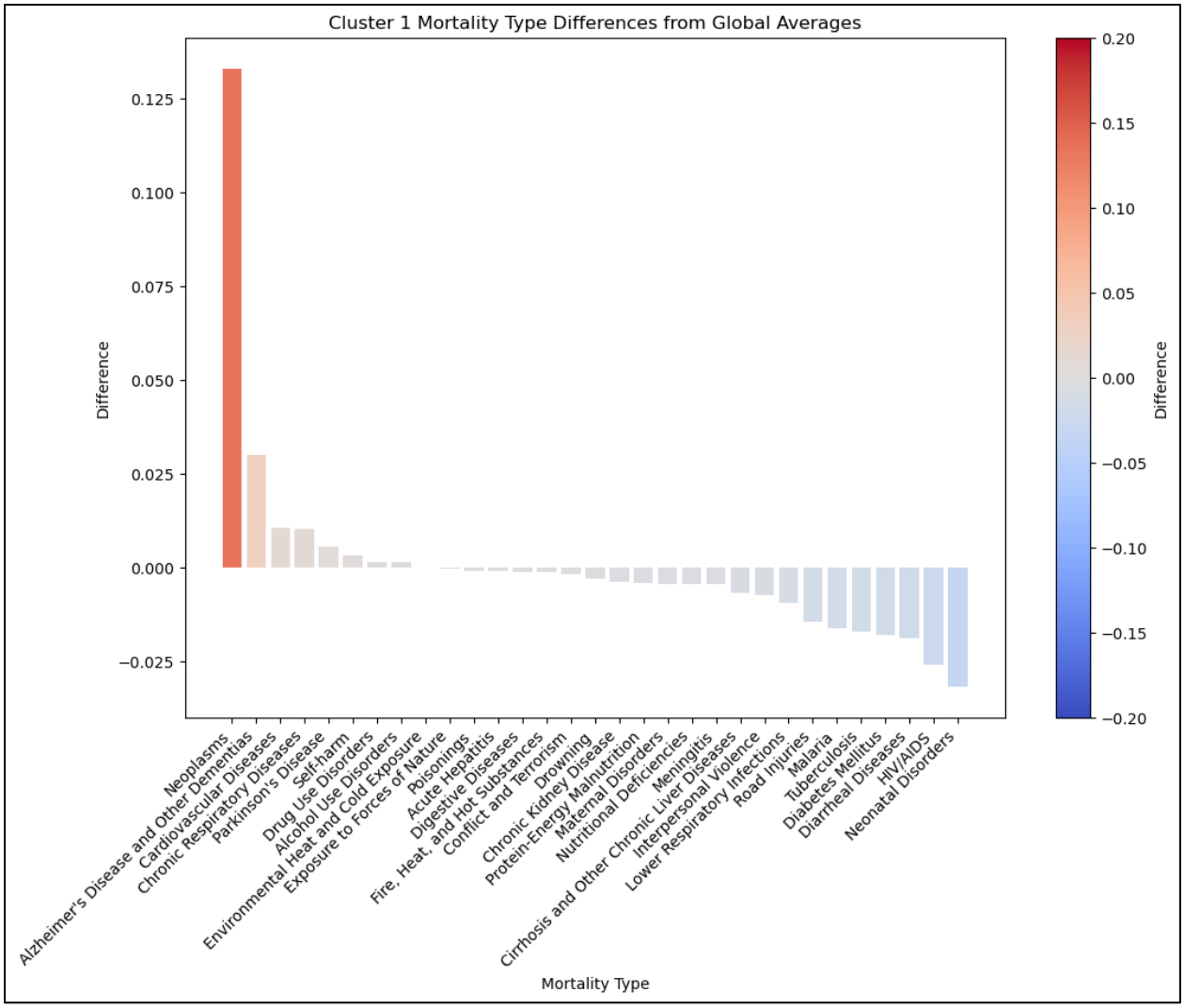

Figure 9: Cluster 1 Mortality Type Differences from Global Averages

Cluster 1 is overwhelmingly defined by its higher rates of cancer deaths. There are many possible explanations ranging from higher rates of plastic use to the fact that the likelihood of cancer increases with age, so countries that have been able to minimize risk of other types of deaths enable their population to live later into life, when cancer is more likely to occur. This is not a medical paper and is not intended to diagnose the causes of these diseases, but it is helpful to connect these statistics with real-world trends to show that the K-Means algorithm is able to identify connections between countries that go deeper than cause of death statistics might demonstrate on the surface.

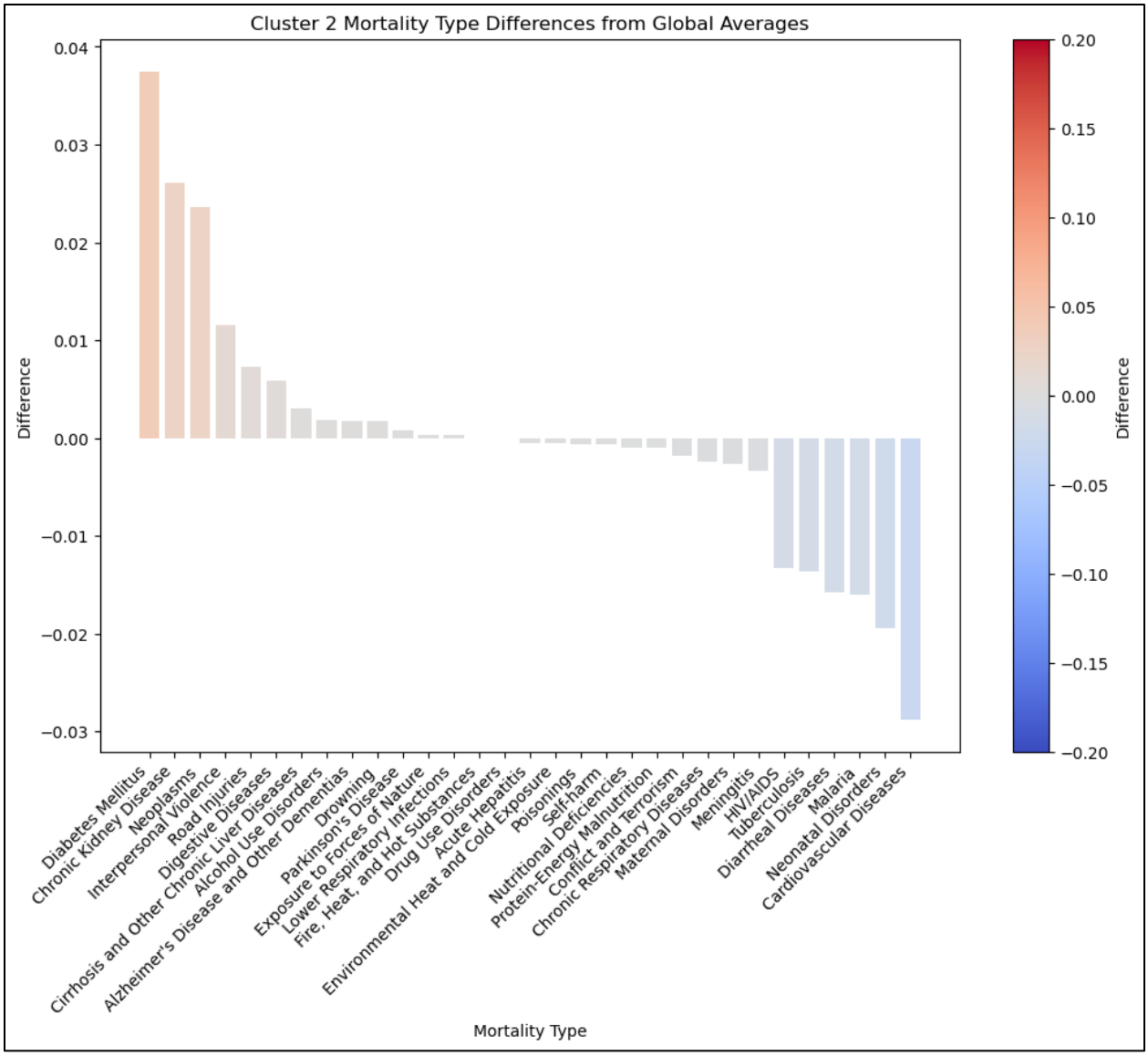

Figure 10: Cluster 2 Mortality Type Differences from Global Averages

Cluster 2 is characterized by a 4% higher rate of Diabetes, a 3% higher rate of Chronic Kidney Disease, and a 3% lower rate of Cardiovascular disease in comparison to the global averages. This cluster is overwhelmingly comprised of Central and South American countries, which begs the question, what is it about these countries that leads to higher rates of Diabetes? Could it be genetic, or might there be certain lifestyle or healthcare differences in these regions?

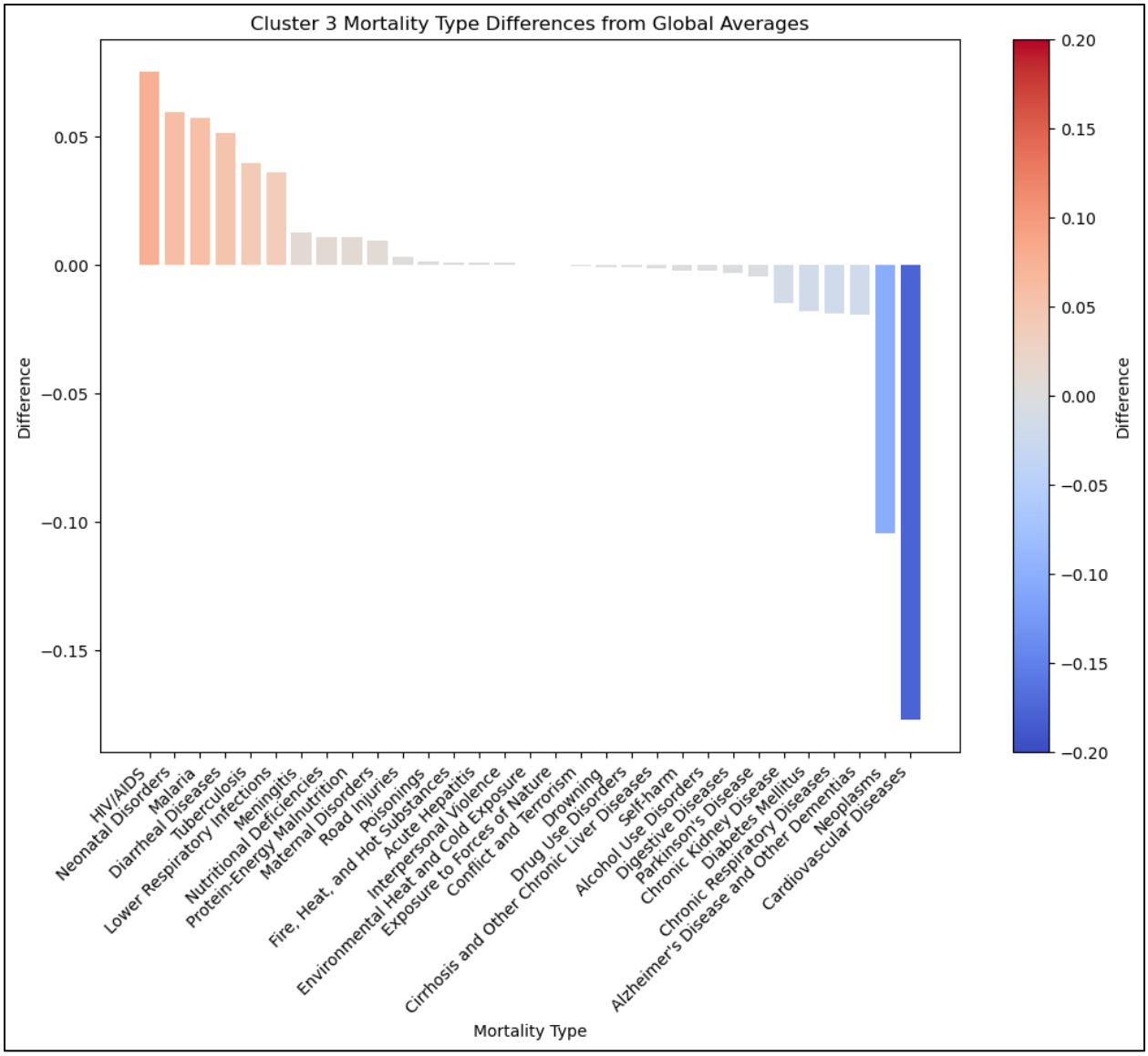

Figure 11: Cluster 3 Mortality Type Differences from Global Averages

Cluster 3 has a 17% lower rate of cardiovascular disease deaths and a 10% lower rate of cancer deaths, which are the top two causes of death globally. This region, however, also has higher rates of HIV/AIDS, Neonatal Disorders, Malaria, and Diarrheal Diseases, none of which are more likely to affect people of older ages. This region is completely comprised of Sub-Saharan African countries, and while they have made great strides over the last 20 years as seen in Figure 4, they still have room for improvement on preventable and treatable diseases.

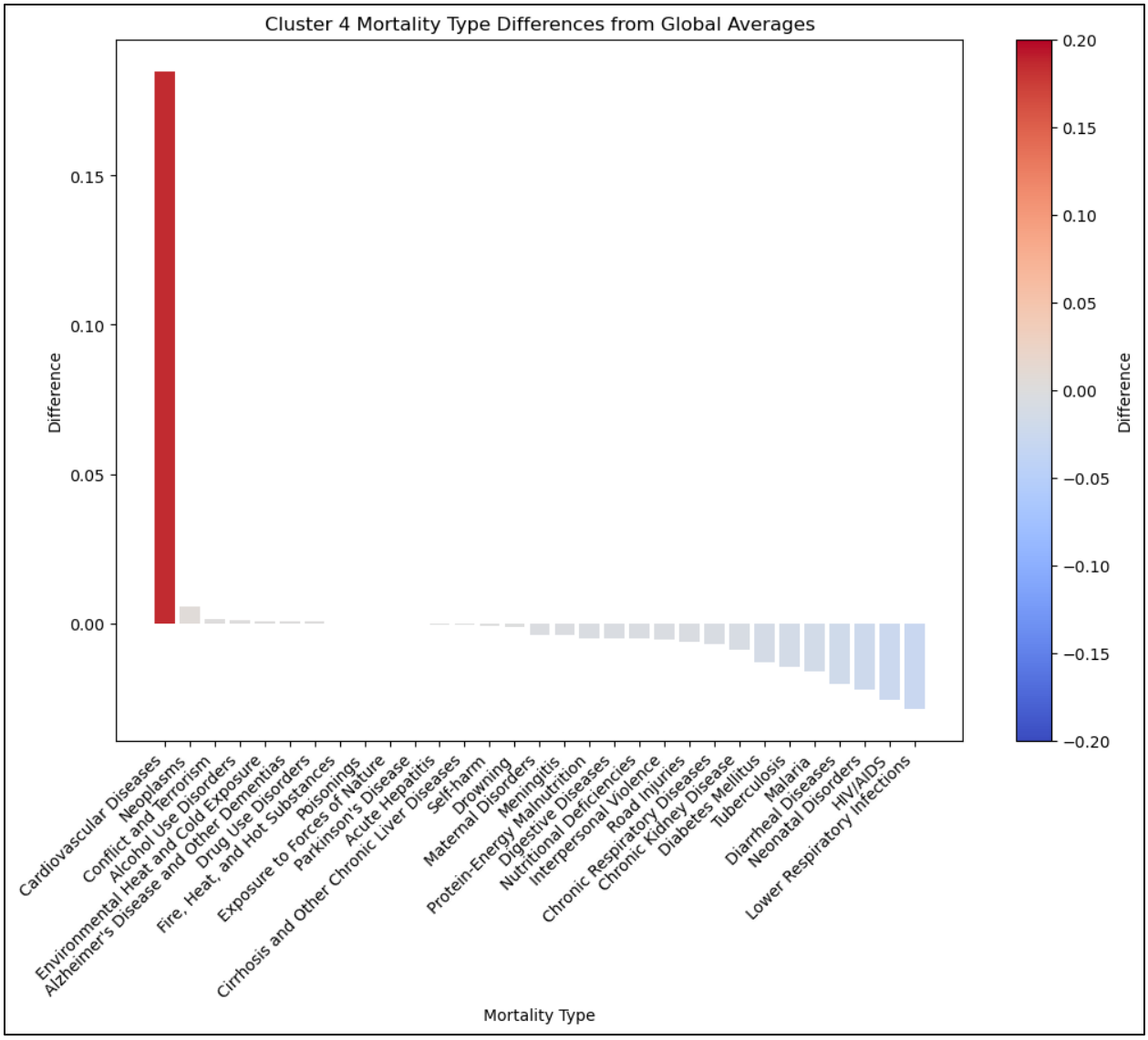

Figure 12: Cluster 4 Mortality Type Differences from Global Averages

In Cluster 4, the rates of Cardiovascular Diseases are a whopping 19% above the global average. This is the largest difference from the global average of any category for all five of the clusters and certainly warrants closer examination by global health professionals. This cluster is primarily comprised of Eastern Europe, East and North Asia, and North Africa.

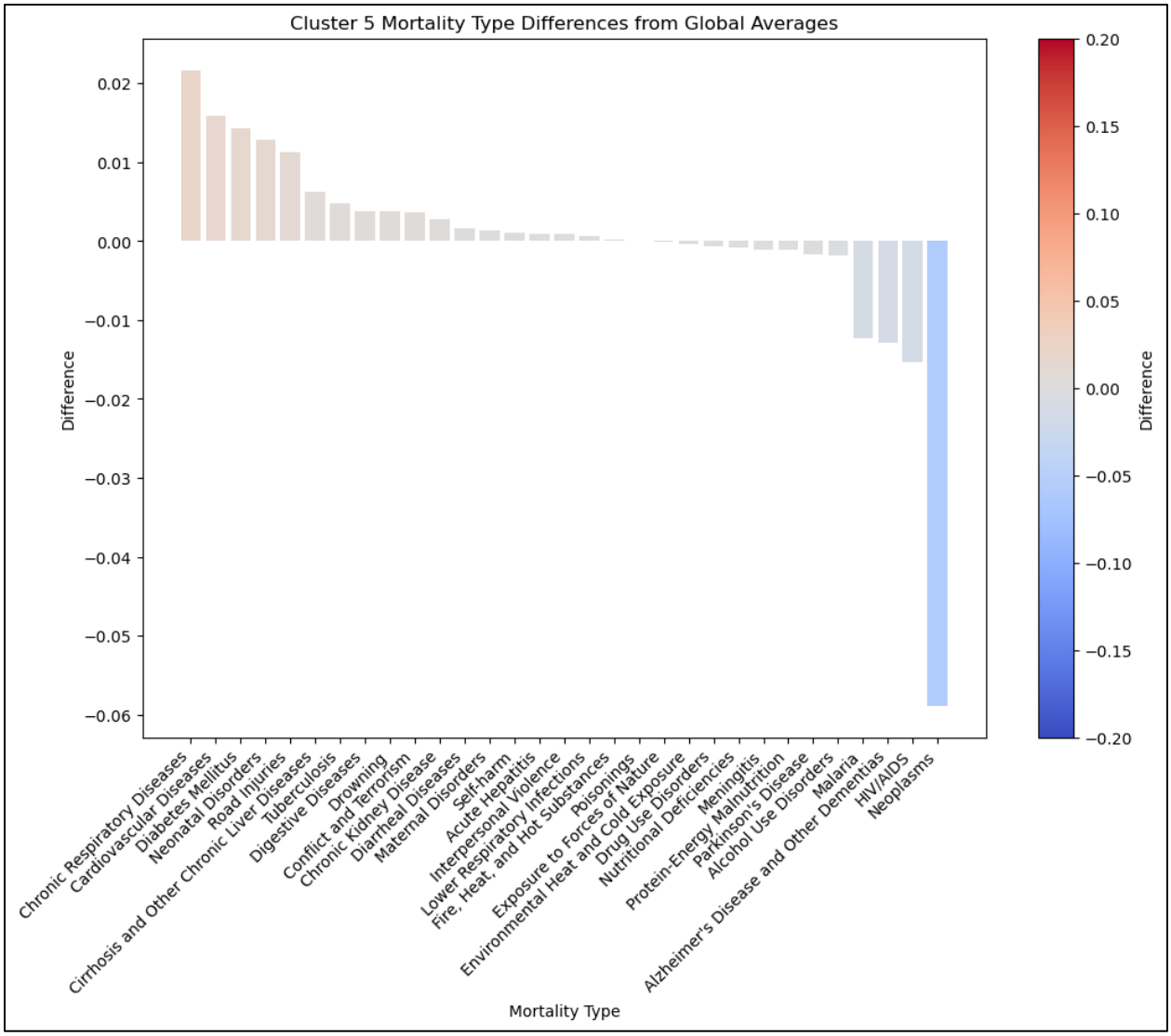

Figure 13: Cluster 5 Mortality Type Differences from Global Averages

Lastly, Cluster 5 is primarily characterized by a 6% lower rate of Neoplasms than the global average, with all other mortality types remaining within 2.5% of the global average. While not as notably unique as the other clusters, there is a good question to ask about why rates of cancer are lower in the Middle East, South Asia, and Southeast Asia.

In addition to clustering the countries based on the full mortality type data, we ran the K-Means algorithm on the grouped data (mortality supertypes), only to find that the clusters were less decisive and interesting when working with the lower-dimensional data.

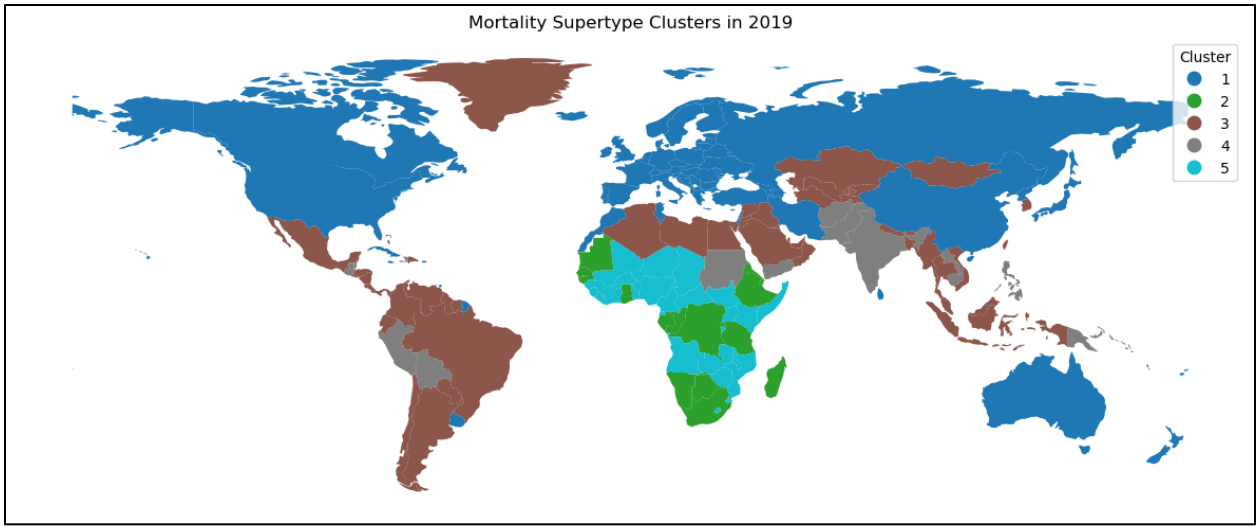

Figure 14: Mortality Supertype Clusters in 2019

Principal Component Analysis

Another way to examine the K-Means clusters is through principal component analysis, where the data is projected onto its top two principal components to observe how separable the it is (365 Data Science, n.d.). PCA can also be used as a pre-processing step for predictive models, which is particularly advantageous when working with high-dimensional data, and it can be used for feature selection (Song et al., 2010). Furthermore, an assumption of linear models is that the features are not collinear because the presence of collinear features could prevent the models from converging. PCA decomposes the data into orthogonal components, effectively eliminating collinearity. In this paper, PCA is solely used for K-Means analysis (Jaadi, n.d.).

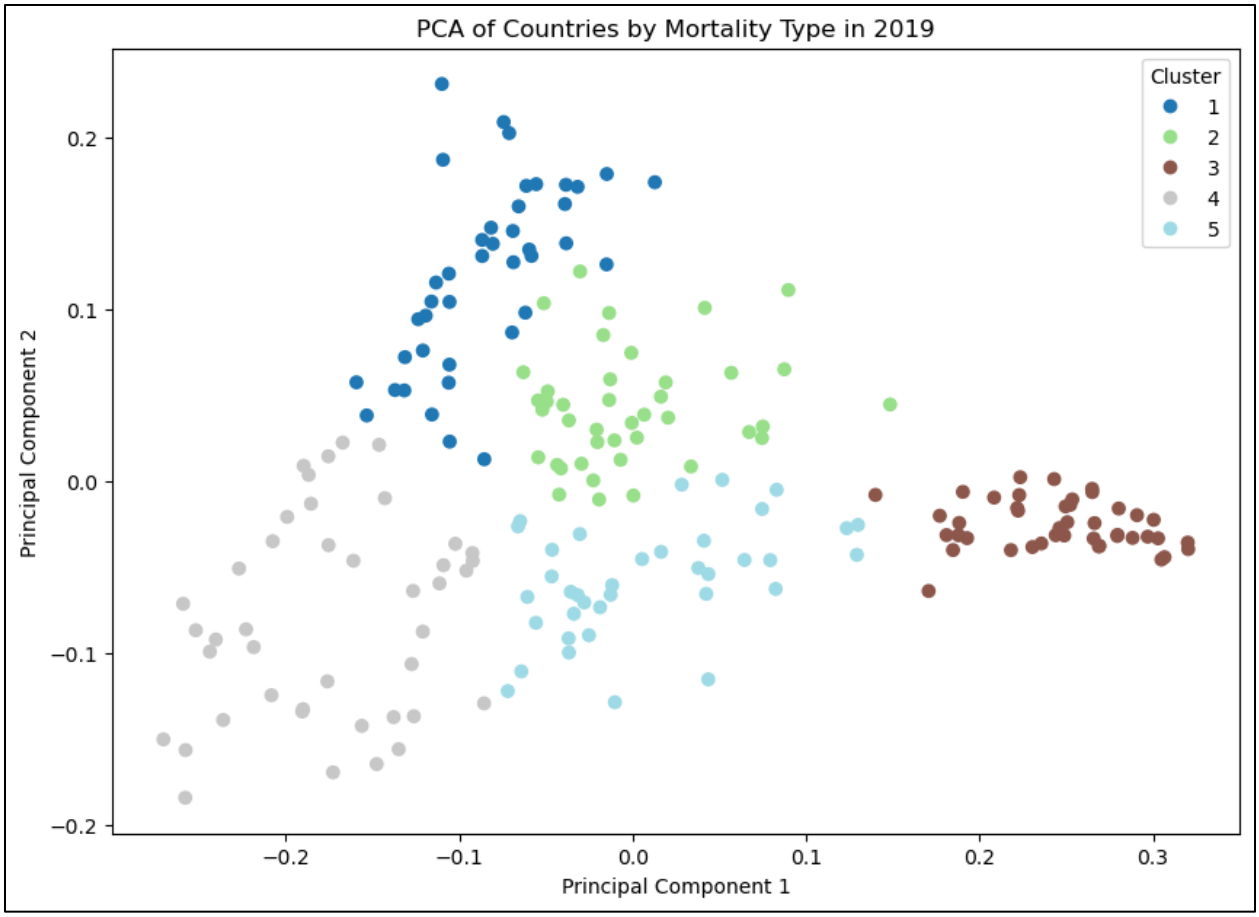

Figure 15: PCA of Countries by Mortality Type in 2019

The plot shows that the data is nearly linearly separable when projected onto its top two components, but not entirely. Much of the variance in the data can be described with only the top two components, meaning that the data is inherently low-dimensional. Corroborated by the correlation matrix, it shows that many of the features could be considered collinear.

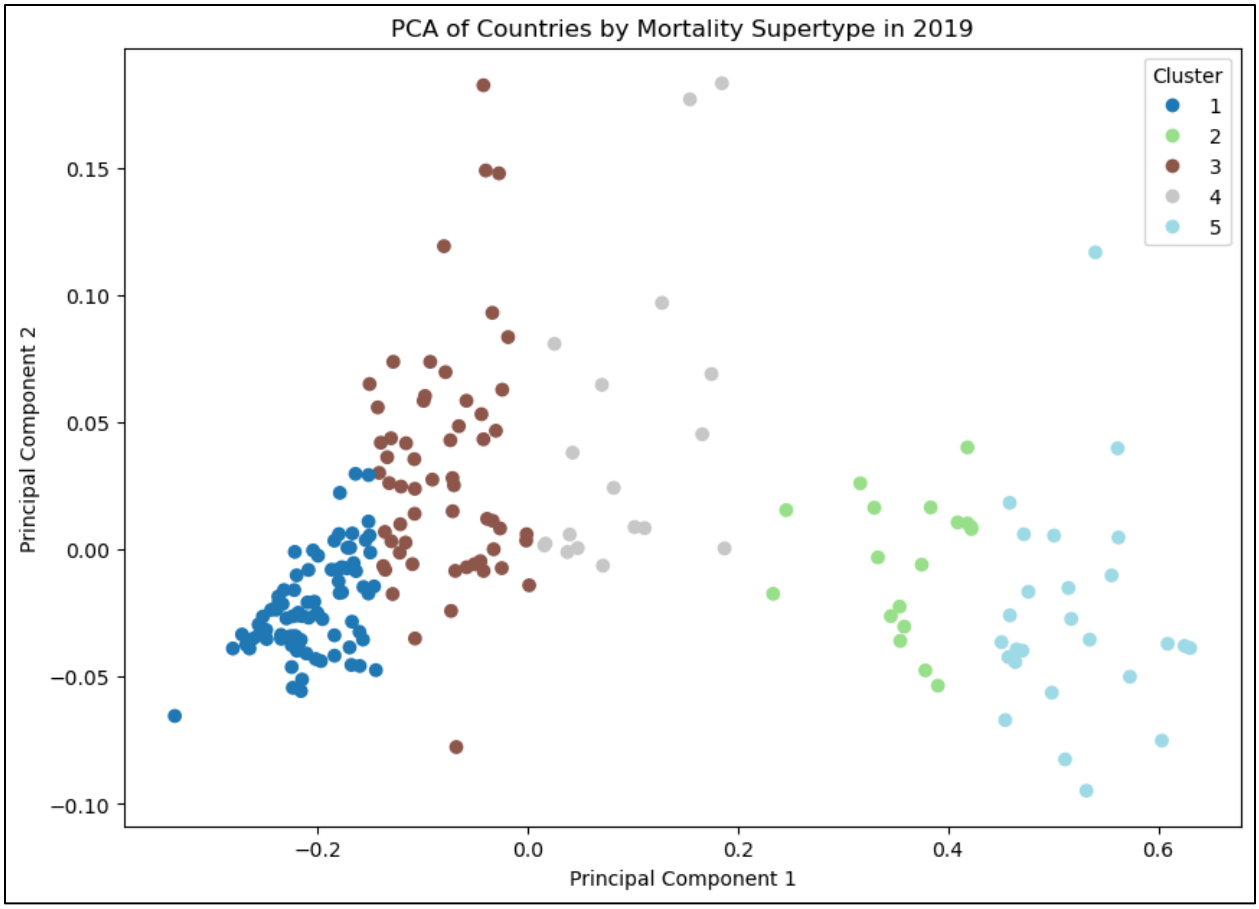

Figure 16: PCA of Countries by Mortality Supertype in 2019

Projecting the mortality supertype data onto its top two components is an even starker example of inherently low-dimensional data, since it is almost fully linearly separable along its first principal component. This supertype data will not be used in the rest of this analysis because it does not contain enough information to provide useful insight.

Random Forest Regression

To understand the relationship between the mortality type statistics and average life expectancy, we can examine the accuracy of a predictive model on this task. Given that the input features don't contain highly complex relationships as established through PCA, a sophisticated neural network model is not necessary to model the relationships between independent and dependent variables.

One option for a predictive model is the simple linear regression model (Legendre, 1805). While this model would likely perform well, it would not be able to account for non-linearities in the data and therefore would likely not have the best accuracy scores.

A Random Forest Regressor model was selected as the initial model for this task (Breiman, 2001). Random forests are a form of ensemble learning, and they work by training a defined number of decision trees on separate subsets of the data and then make predictions by averaging (for

regression) or taking the mode (classification) of the decision tree outputs (Dong, 2020). Another benefit of using a decision tree-based model over a more sophisticated neural network model is that they are more comprehensible, making sensitivity analysis easier to perform (Kotsiantis, 2013). Specifically, the Sci-kit Learn RandomForestRegressor module was used in its default configuration, except that n_estimators was set to be 20. The data was separated for training and testing using an 80/20 split.

Results

Random Forest Model Metrics

After training the model, its accuracy was evaluated on the test dataset, and the results are shown below:

R-squared Score	0.989
Mean Squared Score	0.831
Mean Absolute Error	0.498

Table 2: Random Forest Accuracy Metrics

The overall performance of the model is more than satisfactory, as its average error is only six months.

SHAP

Feature explainability is an important factor in understanding how each input feature affects the model output. Shapley Additive Explanations (SHAP) is a game-theory based method for evaluating the global sensitivity of the model to each independent variable (Kuhnt et al., 2022). Machine learning models, particularly neural networks, are difficult to decipher because of their "black-box" nature (Molnar, 2024). There are many weights involved in the final model, and there can be many complex interactions between the independent variables within the model's structure. Increasing feature A could raise the output value in one scenario and decrease the output value in another.

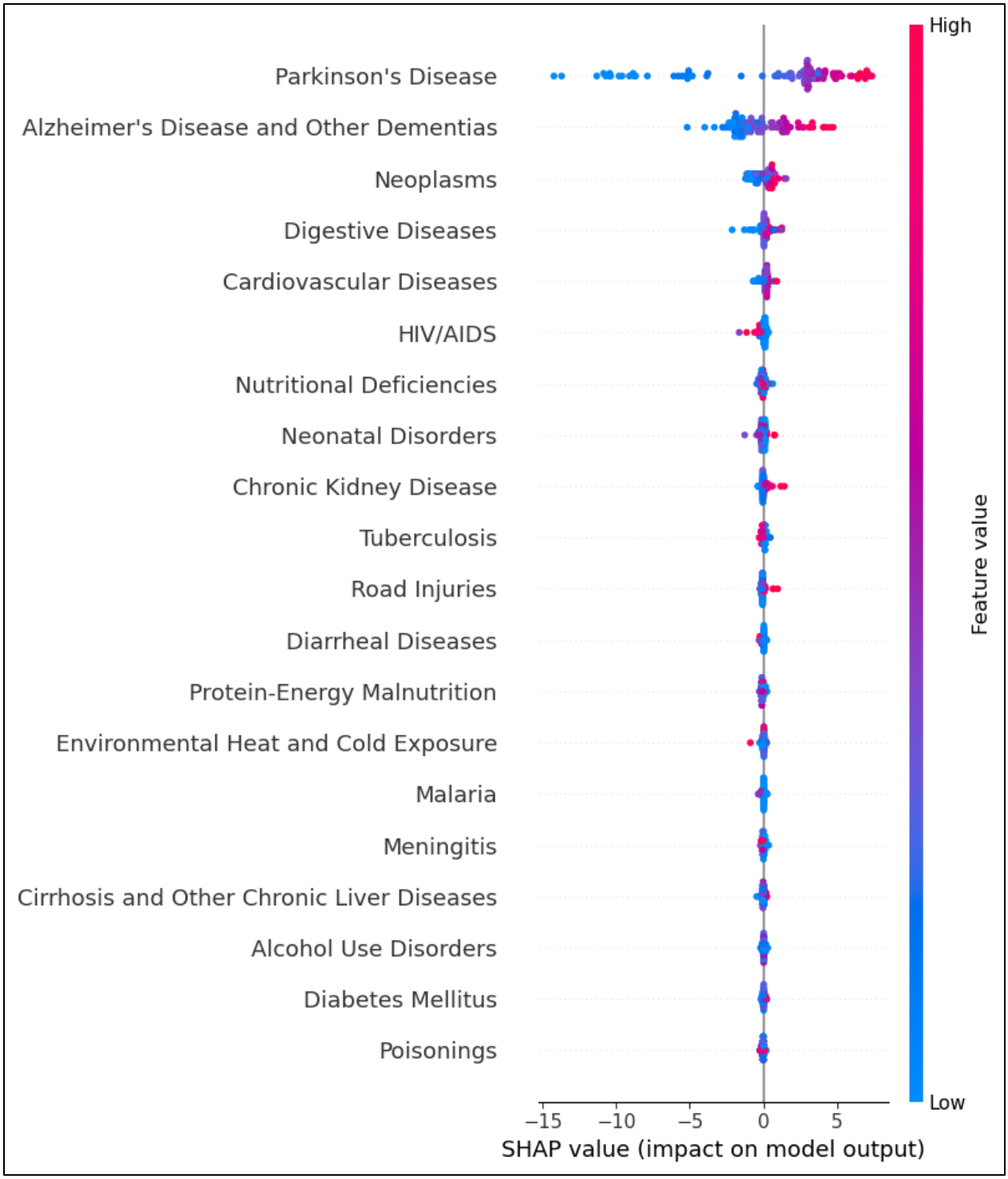

Figure 17: SHAP Plot of Random Forest Regression Model Features

The SHAP values above show that the most influential three features in the model were Parkinson's, Alzheimer's, and Neoplasms (Trevisan, 2022). These also happen to be diseases most commonly affecting older people, as, for example, as only 0.9% developed Parkinson's before age 40, and 5.4% before age 50 (Wickremaratchi et al., 2009).

Perturbation Analysis

SHAP is a form of perturbation analysis, which simply means perturbing the underlying input data to observe the changes in the output. In addition to this analysis, manual perturbation analysis was used in search of an answer to the following question:

If you could drop the proportion of deaths in a given mortality category by 25% in each country, which one would have the greatest positive impact on life expectancy?

The 2019 mortality data was selected for this analysis because it would be most similar to the current state of the world in 2024. Pre-processing for the regression model involved converting the absolute number of deaths in each category into a proportion, so that the data could be compared across countries and was standardized for the model. Normalizing to a standard normal distribution is also a good option, but it would have complicated the perturbation and redistribution.

```
ADJUSTMENT_RATE = 0.25
life_2019_adjusted_data = []
for index, row in life_2019.iterrows():
    for col in mortality_cols:
        new_row = row.copy()
        new_row['Adjusted Column'] = col
        adjustment = new_row[col] * (1 - ADJUSTMENT_RATE)
        new_row[col] = adjustment
        A = (1 - new_row[col]) / (1 - row[col])
        for c in mortality_cols:
            if c != col:
                new_row[c] = new_row[c] * A
        life_2019_adjusted_data.append(new_row)
```

Figure 18: Proportional Redistribution Code Sample

The figure above shows the perturbation and proportional redistribution logic. For each country, one mortality type is decreased by 25%, so the sum of all the proportions would no longer sum to 1. To address this, the amount that was decreased from the perturbed feature is redistributed proportionally to the other features to preserve their proportional relationships and to maintain that the features all sum to 1.

Perturbed Feature:

$$f_j' = f_j \times (1 - p)$$

Proportional Redistribution:

$$A = \frac{1 - f_j'}{1 - f_j}$$

$$f_i' = f_i \times A \ \ \forall i \neq j$$

After this transformation, there are 31 rows per country, with each row having a different feature decreased by 25%. This data is then passed through the model, and the mortality type with the largest positive difference from the unperturbed model prediction was selected as the mortality type of greatest leverage.

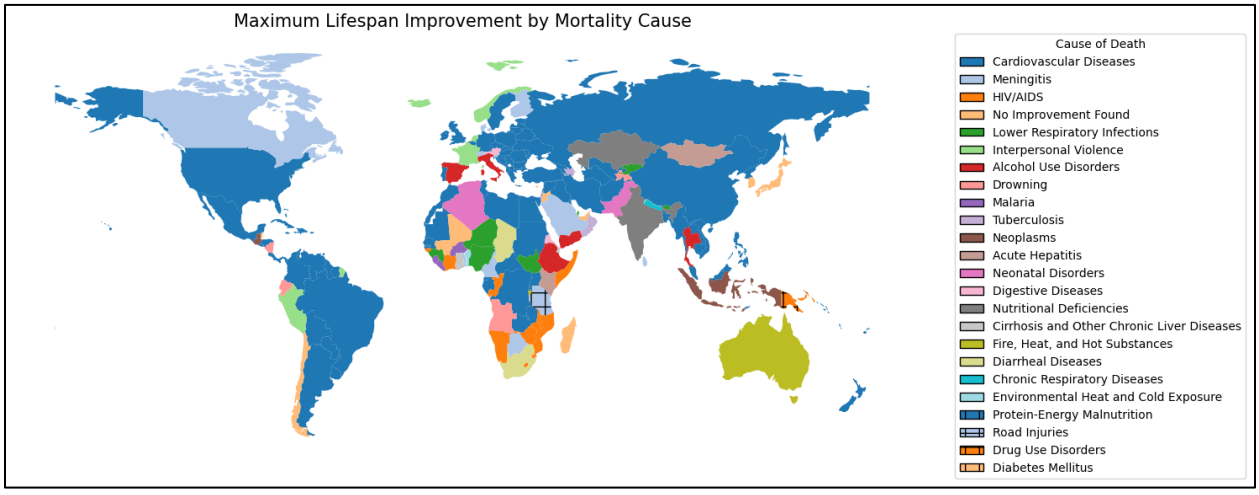

Figure 19: Maximum Lifespan Improvement by Mortality Cause

The cause of death that, if reduced by 25%, would have the most positive impact on the world is Cardiovascular Diseases. Of the 185 countries included in this analysis, exactly 100 had this cause of death selected as the category of maximal impact. Tied for second are HIV/AIDS and Meningitis, each with 10 countries. And the fourth most common category is that No Improvement in life expectancy was found through the perturbation analysis. Cardiovascular diseases have also been found to largely affect working age individuals in low – middle income countries, which could explain why it was so frequently identified as the disease of highest impact on life expectancy (Jagannathan et al., 2019).

Country/Territory	Adjusted Column	Predicted Life Expectancy	Adjusted – Predicted Life Expectancy	Difference
Georgia	Cardiovascular Diseases	73.7615	78.506	4.7445
Eswatini	HIV/AIDS	55.9275	60.383	4.4555

Lesotho	HIV/AIDS	51.705	55.723	4.018
Micronesia	Cardiovascular Diseases	66.16	70.167	4.007
Samoa	Cardiovascular Diseases	70.051	73.973	3.922
Honduras	Cardiovascular Diseases	71.8055	75.538	3.7325
Romania	Cardiovascular Diseases	75.317	78.9925	3.6755
Bulgaria	Cardiovascular Diseases	75.2185	78.842	3.6235
Latvia	Cardiovascular Diseases	75.363	78.9615	3.5985
Jamaica	Cardiovascular Diseases	72.5415	76.0315	3.49

Table 3: Top 10 Life Expectancy Improvements through Perturbation Analysis

The ten countries for which the largest positive impact was identified are displayed above.

Conclusions

What Worked Well

This analysis demonstrates that machine learning and statistics can inform global health policy by creating visibility into the types of problems facing different regions and countries of the world. The K-Means algorithm was used to group countries into five clusters based on the types of challenges they face. And perturbation analysis was conducted on a random forest regression model to determine the varying impact each health challenge has on life expectancy.

Limitations

This type of analysis would not be necessary if more granular data on mortality types were available. In such a case, one would not need to estimate the effect on life expectancy because it could be calculated directly using the age of each individual who died and how they died. However, such data is not available globally, even if it could be procured on a country-by-country basis.

Furthermore, health goes much deeper than simply lifespan because disabilities can significantly alter the quality of one's life without shortening it. There is a metric that can measure this, called "Disability Adjusted Life Years" (DALY), and if there is similar data available about the causes of the of these disabilities, then a similar analysis using that data could be promising (Roser et al., 2016; World Health Organization, 2019).

Cardiovascular Disease

As shown in Figure 5, Cardiovascular Diseases are the leading cause of death globally and have also had the largest increases over the past thirty years. This should cause great alarm and warrant large public health studies to diagnose the causes and to recommend both treatment and

preventative measures to reduce the risk globally. The modern world is changing more rapidly than ever, and along with that come novel challenges. We need to embrace new technology and growing data ecosystems to help us best handle the problems of the 21st century.

References

Banerjee, S. (n.d.). *Cause of deaths around the world* [Data set]. Kaggle. Retrieved November 26, 2024, from https://www.kaggle.com/datasets/iamsouravbanerjee/cause-of-deaths-around-the-world/data

Bertsimas, D., Orfanoudaki, A., & Wiberg, H. (2021). Interpretable clustering: an optimization approach. *Machine Learning*, 110, 89–138. https://doi.org/10.1007/s10994-020-05896-2

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Builtin. (n.d.). Step-by-step explanation of principal component analysis (PCA). Retrieved November 26, 2024, from https://builtin.com/data-science/step-step-explanation-principal-component-analysis

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241–258. https://doi.org/10.1007/s11704-019-8208-z

Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics (international student edition)*. Pisani, R. Purves, 4th Edn. WW Norton & Company, New York.

GeeksforGeeks. (n.d.). Random forest regression in Python. Retrieved November 26, 2024, from https://www.geeksforgeeks.org/random-forest-regression-in-python/

GeoPandas Development Team. (n.d.). *Mapping and plotting shapes.* GeoPandas 0.5.0 documentation. Retrieved November 26, 2024, from https://geopandas.org/en/v0.5.0/mapping.html

IBAN. (n.d.). Country codes. Retrieved November 26, 2024, from https://www.iban.com/country-codes

International Monetary Fund. (2024). World Economic Outlook: GDP per capita. Retrieved November 26, 2024, from https://www.imf.org/external/datamapper/NGDPDPC@WEO/OEMDC/ADVEC/WEOWORLD

Jagannathan, R., Patel, S. A., Ali, M. K., & Narayan, K. M. V. (2019). Global updates on cardiovascular disease mortality trends and attribution of traditional risk factors. *Current Diabetes Reports*, *19*(7), 44. https://doi.org/10.1007/s11892-019-1161-2

Janse, R. J., Hoekstra, T., Jager, K. J., Zoccali, C., Tripepi, G., Dekker, F. W., & Van Diepen, M. (2021). Conducting correlation analysis: Important limitations and pitfalls. *Clinical Kidney Journal*, *14*(11), 2332–2337. https://doi.org/10.1093/ckj/sfab085

Kotsiantis, S. B. (2013). Decision trees: A recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Kuhnt, S., & Kalka, A. (2022). Global sensitivity analysis for the interpretation of machine learning algorithms. In A. Steland & K.-L. Tsui (Eds.), *Artificial Intelligence, Big Data and Data Science in Statistics* (pp. 155–169). Springer International Publishing. https://doi.org/10.1007/978-3-031-07155-3_6

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. Paris: Didot.

Molnar, C. (2024, July 31). Chapter 9.5: Shapley values. In *Interpretable machine learning: A guide for making black box models explainable.* Retrieved November 26, 2024, from https://christophm.github.io/interpretable-ml-book/shapley.html

Murray, C. J., & Lopez, A. D. (1997). Mortality by cause for eight regions of the world: Global Burden of Disease Study. *The Lancet, 349*(9061), 1269–1276. https://doi.org/10.1016/S0140-6736(96)07493-4

Rodriguez M., Comin C., Casanova D., Bruno O., Amancio D., Costa L., et al. (2019) Clustering algorithms: A comparative approach. *PLoS ONE* 14(1): e0210236. https://doi.org/10.1371/journal.pone.0210236

Roser, M., Ritchie, H., & Spooner, F. (2016). *Burden of disease.* Our World in Data. Retrieved November 26, 2024, from https://ourworldindata.org/burden-of-disease

Scikit-Learn (n.d.). *sklearn.ensemble.RandomForestRegressor* (version 1.3.0) [Documentation]. Scikit-learn. Retrieved November 26, 2024, from https://scikit-learn.org/dev/modules/generated/sklearn.ensemble.RandomForestRegressor.html

Song, F., Guo, Z., & Mei, D. (2010) Feature Selection Using Principal Component Analysis. *2010 International Conference on System Science, Engineering Design and Manufacturing Informatization* (pp. 27-30). https://doi.org/10.1109/ICSEM.2010.14

Torre, L. A., Siegel, R. L., Ward, E. M., & Jemal, A. (2016). Global cancer incidence and mortality rates and trends—An update. *Cancer Epidemiology, Biomarkers & Prevention, 25*(1), (pp. 16–27). https://doi.org/10.1158/1055-9965.EPI-15-0578

Trevisan V. (2022). Using SHAP values to explain how your machine learning model works. *Towards Data Science.* Retrieved November 26, 2024, from https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137

Wickremaratchi, M. M., Ben-Schlomo, Y., & Morris, H. R. (2009). The effect of onset age on the clinical features of Parkinson's disease. *European Journal of Neurology*, 16(4), (pp. 450–456). https://doi.org/10.1111/j.1468-1331.2008.02514.x

World Health Organization. (2019). *Global health estimates: DALY methods 2019* [PDF]. Retrieved November 26, 2024, from https://cdn.who.int/media/docs/default-source/gho-documents/global-health-estimates/ghe2019_daly-methods.pdf?sfvrsn=31b25009_7

World Health Organization. (2020). *Basic documents* (49th ed.). Retrieved November 26, 2024, from https://apps.who.int/gb/bd/pdf_files/BD_49th-en.pdf#page=6

World Health Organization. (n.d.). *Life expectancy at birth (years)* [Data set]. Retrieved November 26, 2024, from https://www.who.int/data/gho/data/indicators/indicator-details/GHO/life-expectancy-at-birth-(years)

Xu, D., & Tian, Y. A (2015) Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, 2, (pp. 165–193). https://doi.org/10.1007/s40745-015-0040-1

Yiu, T. (2019). A practical guide on K-means clustering. Towards Data Science. Retrieved November 26, 2024, from https://towardsdatascience.com/a-practical-guide-on-k-means-clustering-ca3bef3c853d

365 Data Science. (n.d.). PCA and K-means clustering in Python. Retrieved November 26, 2024, from https://365datascience.com/tutorials/python-tutorials/pca-k-means/