Comparative Analysis of Health Care Risk Factors and Economic Influences on Life Expectancy Across Different Countries

David Khachikov
computer science department
Innopolis University
Innopolis, Russia
d.khachikov@innopolis.university

***Abstract*—This study delineates the relationships between distinct healthcare challenges across various countries and offers recommendations for enhancing the quality of life for citizens.**

***Index Terms*—Healthcare, factors of risk, data science, economics and health**

I. Introduction

In the realm of global health, the interplay between health care systems, socioeconomic factors, and public health policies plays a pivotal role in determining the life expectancy and overall health outcomes of populations. This study aims to delve into the intricate relationship between health care issues, economic factors, and the broader socioeconomic landscape, with the goal of identifying key risk factors that significantly impact life expectancy across various country classes. By examining the health care issues prevalent in these countries and their economic underpinnings, we seek to uncover the multifaceted influences that contribute to health disparities and inequities. Recent research has highlighted the significant role of socioeconomic and related social factors in shaping health across a wide range of health indicators and settings [1]. For instance, the Scottish physician Thomas McKeown's study on death records for England and Wales demonstrated that improvements in living conditions, including nutrition, sanitation, and clean water, had a more profound impact on health outcomes than advances in medical care alone [1]. This underscores the importance of considering a broad range of factors, including economic and social, in our analysis of health care risk factors and their impact on life expectancy. This case study will draw on a variety of sources, including health care, economic factors, to provide a comprehensive understanding of the main risk factors within different classes of countries and their economic influences. By comparing these factors and investigating correlated or multi-correlated features, I aim to gain a deeper understanding of the root causes of health disparities and to identify the risk factors that should be prioritized in different regions to achieve the highest possible life expectancy.

II. Data

I utilized a total of four data sets. The first and primary one is a dataset [2], [3] from the World Health Organization (WHO) on deaths from 1990 to 2016, categorized by risk factors. This dataset is quite comprehensive, as it covers a wide range of countries and nearly all cells are populated. However, there is a limitation regarding the column for deaths attributed to elevated cholesterol levels, as the data is only available for 1990, 1995, 2000, 2005, 2006 and 2010. Therefore, it was determined to fill in all missing values through the use of linear interpolation between existing values. Additionally, another value in the 'Outdoor air pollution' column is also missing, which has also been filled in using linear interpolation. There is an issue with respect to country names, as they can vary for the same nation in each dataset. For instance, in the initial dataset containing information on risk factors, my country is referred to as Russia, whereas in the dataset containing population data, the country is identified as the Russian Federation. Due to these concerns, I will be converting country names into their respective codes.

The remaining datasets are supplementary. The second dataset [4] is responsible for converting informal country names to their unambiguous code equivalents. This dataset stores country code data in various formats, as well as the corresponding informal names. The third [5] and fourth [6] datasets include data on Gross Domestic Product (GDP) and population, respectively.

III. Theory, statistical techniques

Several statistical techniques have been employed in this work. The most challenging aspect of the analysis involves the K-means [7] clustering algorithm. In summary, the algorithm begins by selecting cluster centers as follows: a random point is selected, which becomes the first centroid. Subsequently, for each data point, the distance to the closest centroid $D(x)$ is calculated. Afterwards, a new centroid is selected with probability proportional to $(D(x))^2$. This process is repeated until k initial centroids have been identified. The K-means++ [8] algorithm provides an average approximation ratio of $O(\log k)$, where k is the number of clusters being used. This differs from vanilla K-means [7], as the latter can result in clustering that are arbitrarily worse than an optimal solution.

The algorithm itself [7] is relatively straightforward, requiring only the finding of the nearest centroids for each data point. The number of centroids corresponds to the cluster to which a point is assigned at each iteration. Next, for each cluster, the center of mass is calculated, and these steps are repeated until either the centers change (which is certain to occur at some point), or the changes become slow.

The T-SNE algorithm was also used to reduce the dimensionality of the dataset, as there is a significant number of correlated features within it. This will be demonstrated through the correlation matrix. A t-test, Shapiro test was also used for hypothesis testing.

IV. Statistical tools, other software

The work makes use of the Python NumPy libraries for array processing and various metrics calculations [9]. The Scikit-learn library is used to implement algorithms [10], while pandas is utilized to handle datasets [11]. GeoPandas is employed for visualizing clusters of countries according to similar risk factors [12], while Matplotlib is utilized for creating and displaying graphs [13]. Pandas for visualizing heatmap of correlational matrix. Yandex Translate editor was used to improve readability of text, I hope it helped [14].

V. Results

A. Basic data observations

The analysis of mortality rates in various countries, particularly Russia, revealed a significant decline due to various factors. However, it was noted that assuming a constant population may not accurately reflect the current situation, as the population size varies significantly between countries. This discrepancy was particularly evident when comparing countries with vastly different populations, such as India and Russia. To address this problem, it was determined that incorporating population data for each year in the dataset would be essential. Also, information about GDP has been added to initial dataset to analyze connection between economics and medical factors.

Firstly for me was interesting to consider statistics for Russia. How it changed in 25 years? Let's watch.

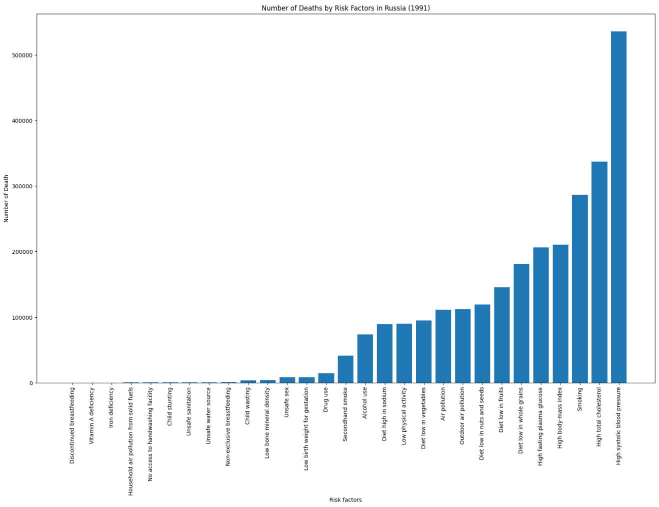

Fig. 1. Russia, 1991

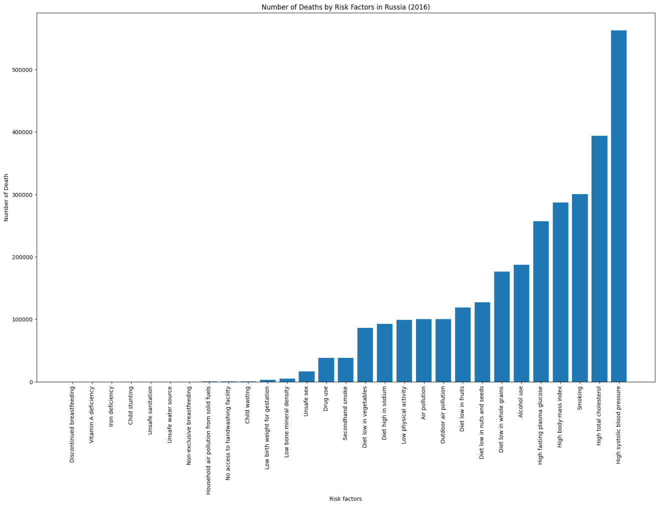

Fig. 2. Russia, 2016

In Figures 1 and 2, you can see that the risk factors have changed very little over the past 25 years, with each other. This means that the medical system has become stable and continues to address the same medical issues year after year.

However, if we plot difference between number of death in 1991 and 2016 years we will see interesting picture

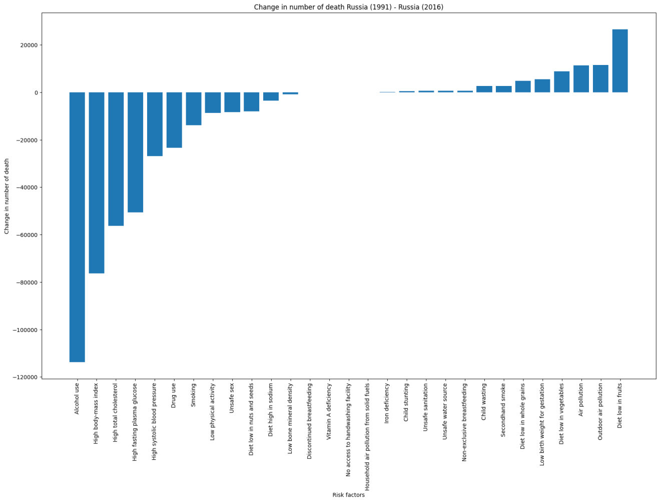

Fig. 3. Difference between number of death in Russia 1991–2016

According to Figure 3, it can be seen that mortality in Russia from various factors has increased significantly, however, an implicit assumption is made that the population remains constant, but this is not the case. Also, a visual comparison of countries with different numbers of people, such as India and Russia, will cause big problems. Therefore, it is necessary to add information about the population by year to the initial data.

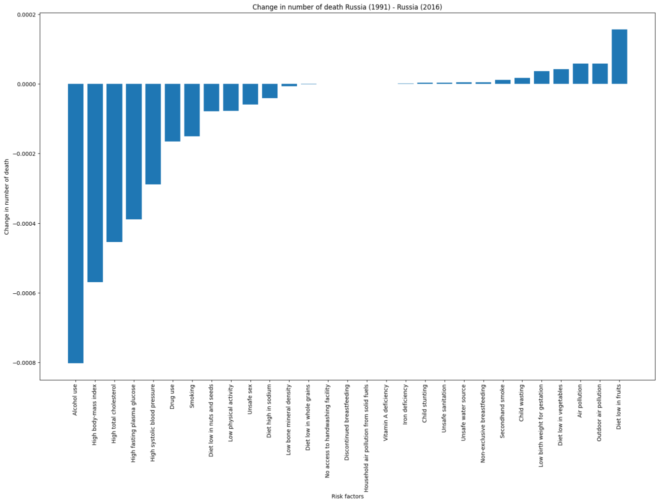

Fig. 4. Difference between number of death (per person) in Russia 1991–2016

You can see that Figure 3, 4 have the same shape. Assumption about equal population holds.

B. Analysis via clustering

The goal was to classify countries based on their exposure to various risk factors associated with death. To accomplish this, I identified a suitable metric and implemented it. Two vectors were generated, representing the number of deaths attributed to different risk factors. There were 29 dimensions initially. These vectors were then normalized to convert them into a probability of death from specific diseases, which does not linearly depend on population size. Please note that the normalization of the data occurs relative to the number of deaths in the country, rather than by population. This is because the purpose of this case study is to precisely determine the ratio, without directly referencing the percentage of deaths in the population.

For metrics there were two major choices - Euclidean and cosine metrics. I decided to use Euclidean distance between the vectors instead of dot product (the same as cosine with vectors of length 1), as it would help us to avoid over-representing the more prevalent risk factors. This approach was tested against both methods in order to evaluate its effectiveness. The study focused on the year 2005, as it had a relatively low number of missing data points. Before Kmeans clustering I applied T-SNE to reduce dimensionality, because initial dataset includes too many correlated features. To demonstrate distribution of countries presented Figures show point distribution in projected 2D space.

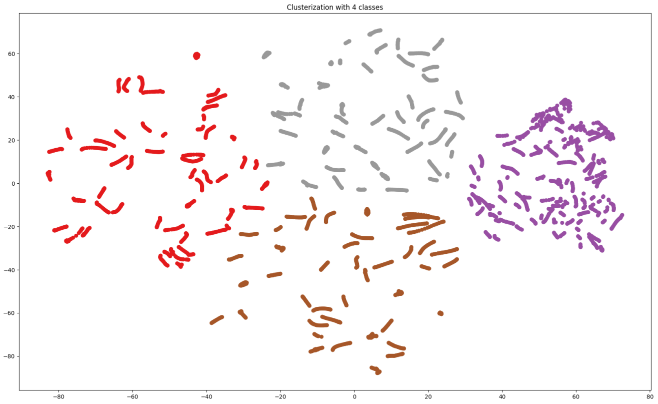

Fig. 5. Four classes on 2D

On Figure 5 present all points for all years projected to the plane. We can observe purple class, which separates very simply and three other, which distributed a bit more randomly on the plane. However, if we project data not on two-dimensional plane, but on three-dimensional separability will be more obvious. We can also note that there are minor bars on the chart that represent the main causes of death in a country. The longer a bar, the more significant the change in the situation. And we can see that many of these bars are directed towards purple, which represents the highest number of dots. This indicates that all countries on average strive to improve healthcare, and purple represents the most developed countries in terms of medical care.

It's convenient to analyze four classes, and you can see that they are separable. After the initial preparation, selecting the number of clusters to be four and the dimension onto which we project to be three, we will obtain a map of countries according to class.

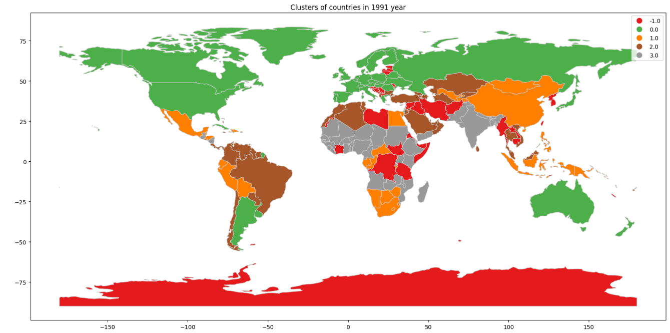

Fig. 6. Clustered world (1991)

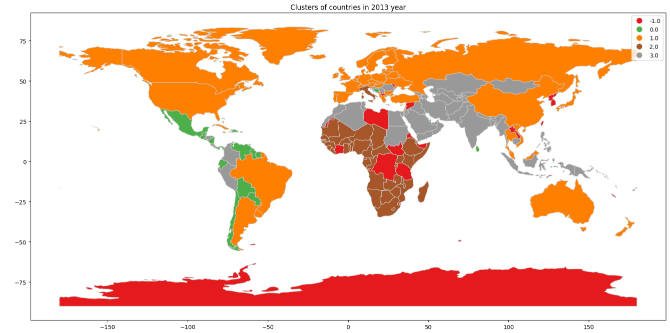

Fig. 7. Clustered world (2013)

An analysis identified several clusters on the world map in 2013 year.

0) Central and part of South America
1) Russia, Europe, North America, part of South America, China
2) India, Middle East, North Africa
3) South and Central Africa

However, these clusters were found to be dynamic and change from year to year. To enhance clarity, graphs for 1991 and 2013 were also presented.

A detailed examination of countries within different and same clusters, such as South Africa, Russia, and Belarus in 2013, revealed significant differences in the issues they faced. The analysis compared two vectors, each containing the number of deaths attributed to various risk factors in each country after normalization.

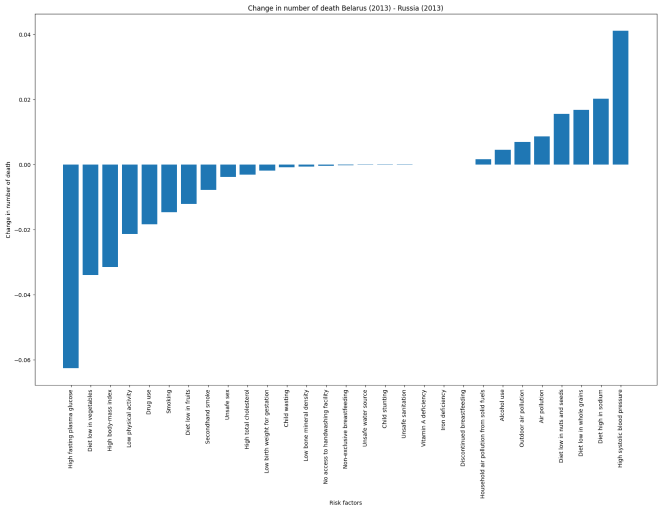

Fig. 8. Difference in percentage of death for different reasons between Russia and Belarus in 2013

Interestingly, Russia and Belarus exhibit similarities in terms of disease causation.

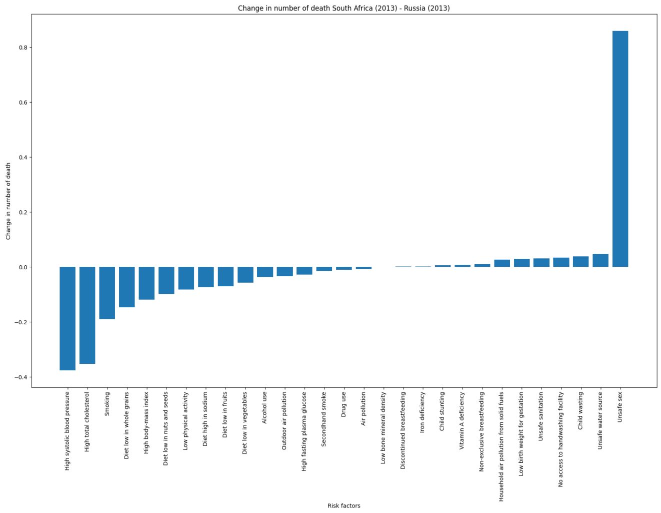

Fig. 9. Difference in percentage of death for different reasons between Russia and South Africa in 2013

It can be seen that the issues faced by people in Russia differ from those faced by individuals in South Africa.

C. Testing change from year to year

It has previously been mentioned that the small strips in Figure 5 represent the same country at different points in time. However, this has not yet been conclusively verified. To verify this, we need to understand how these strips or points can be statistically interpreted.

I propose the following approach to interpreting these points. If we consider the normalized number of deaths in successive years and calculate the average value, this average has a mathematical expectation of zero. However, the band differs in that its mean has an expectation value that is not zero, but rather if its direction is consistent from year to year there is a higher concentration of points closer to the zero line. Or we have some unimodal distribution with non-zero mean and small variance. I have an assumption that Germany is a point on Figure 5 and South Africa is a stripe.

A hypothesis test was conducted to determine whether the situation in Germany is changing from year to year. Basic assumption about data is independence and normal distribution of mean. As we can see on the picture, our distribution has one mode. That means, that CLT fast approaches population mean to sample mean. Independence also can be assumed on low time interval (one year).

H_0: In Germany, the situation does not change from year to year (zero mean).

H_1: The situation changes from year to year (non-zero mean).

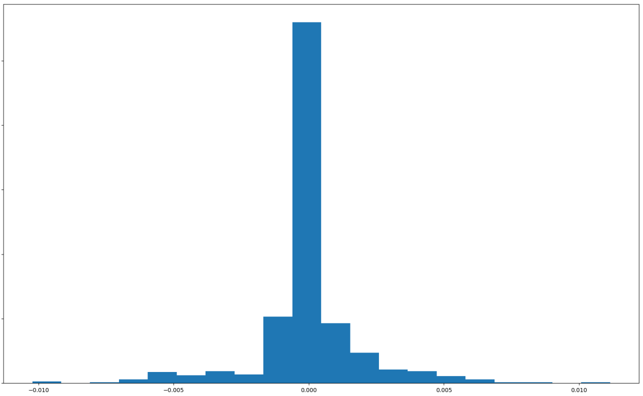

Fig. 10. Visualization of sample for Germany

The $p_{value} \approx 0.74$ found was greater than 0.05, suggesting that the null hypothesis cannot be rejected. However, a more in-depth analysis of the changes in risk factors in South Africa over the past five years reveals a significant difference, indicated by a p-value less than 0.05.

H_0: In South Africa, the situation does not change from year to year (zero mean).

H_1: The situation changes from year to year (non-zero mean).

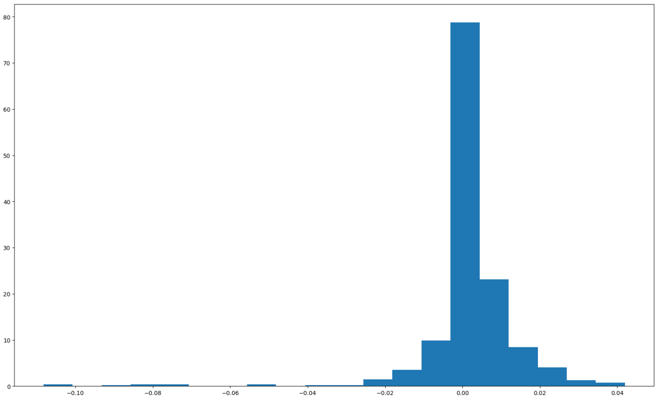

Fig. 11. Visualization of sample for South Africa

The $p_{value} \approx 0.0002$ found was less than 0.05, suggesting that the null hypothesis rejection. The mean is ≈ 0.0017, standard deviation is ≈ 0.0124. The reason of high standard deviation is heavy left tail. So, changes happen, but they directed to different sides. We can additionally check that distribution is not normal using Shapiro test.

H_0: Distribution shown on Figure 11 is normal.

H_1: Distribution shown on Figure 11 is not normal.

$p_{value} \approx 10^{-36}$ and we can reject hypothesis about normality of distribution.

D. *Analysis of correlation*

The study aimed to determine the correlation between a country's purchasing power parity and class using data from 2013, and linear correlation due to the categorical nature of the class variable.

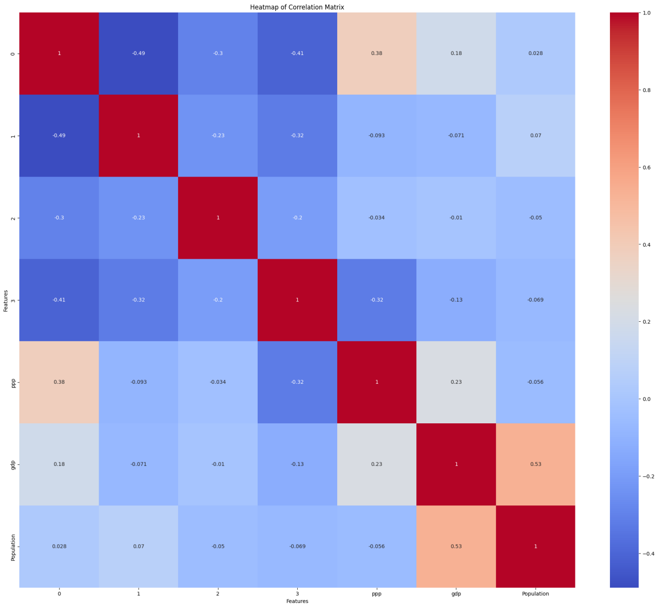

Fig. 12. Correlation between GDP, PPP, population size, and classes

The zero class showed a positive correlation with the PPP, suggesting that countries with a higher level of economic development were more likely to belong to this class. All classes exhibited a negative correlation, suggesting that they could be well separated. Next Figure 13 presents whole correlation between all features.

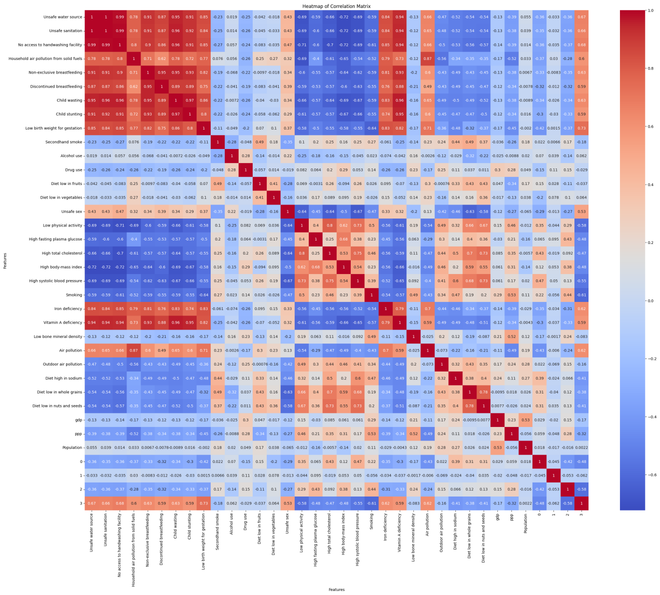

Fig. 13. Correlations

Analysis revealed common problems that often occur together, such as lack of access to basic goods, vitamin A deficiency and iron deficiency, among people in the 'third' class. Countries in the 'first' class were considered to be developing countries that have overcome the challenges faced by countries in the 'third' class with low aged population. The first and second classes faced its own unique set of challenges, including elevated blood pressure and cholesterol levels. Based on the research findings, strategies have been proposed to increase global life expectancy. It has been suggested that providing access to basic needs to countries in the lowest two categories would be the most effective approach for impoverished regions in Africa. For more developed countries, promoting a healthy lifestyle and the development of advanced medical techniques are recommended.

VI. Conclusion

This study has conducted a comprehensive analysis of the interaction between health care systems, socioeconomic conditions, and public health policies, in order to determine the life expectancy and overall health status of populations within various countries. By examining health care concerns, economic aspects, and the wider socioeconomic environment, we have identified significant risk factors that have a significant impact on life expectancy among different country groups. The analysis of mortality rate data, specifically focusing on Russia and South Africa, has revealed minor changes over the last 25 years. While there have been significant alterations in South Africa's health care system, this highlights the dynamic nature of these challenges. The incorporation of population and GDP data into our analysis has allowed us to gain a more in-depth understanding of the economic factors influencing health outcomes. This information has been critical in identifying the most significant health care concerns and their underlying economic factors. We have identified countries with relatively low purchasing power parity (PPP), but which still face significant health challenges. Despite the normalization of our statistics, we have found a strong correlation between certain classes and PPP values. This highlights the fact that every country can achieve a similar distribution by enhancing its healthcare system. The use of statistical techniques, such as Kmeans++ and T-SNE, has allowed us to classify countries according to their exposure to various risk factors related to death. This classification provides insights into the various health care challenges facing countries at different stages of economic development. The analysis also did not reveal significant differences in health care issues among countries within the same class, reinforcing the accuracy of the classification. Furthermore, hypothesis testing conducted on changes in risk factors over time provided evidence of significant shifts in the health care challenges experienced by countries like South Africa. These findings underscore optimistic predictions about the future of our world in 50 years. The correlation analysis between a country's purchasing power parity and its class, as well as the linear correlation due to the categorical nature of the class variable, further highlights the economic influences on health outcomes.

In conclusion, the study has demonstrated the complex influences contributing to health disparities across different countries. Through a comparative analysis of these factors and an investigation of correlated or multi-correlated characteristics, we have achieved a more comprehensive understanding of the underlying causes of health inequalities.

VII. Acknowledgment

I would like to extend my gratitude to the lecturer of the course, Mr. Vladimir Ivanov, for the instruction and the recommendation to utilize T-SNE. Additionally, I am greatly indebted to Mr. Vladimir Bazilevich for the hands-on exercises he led and for checking our work carefully, making sure that we were doing things correctly. Furthermore, I would like to thank Zlata and Zamira for all their previous courses and notebooks from the class sessions.

References

[1] P. Braveman and L. Gottlieb, "The social determinants of health: it's time to consider the causes of the causes," *Public Health Rep.*, vol. 129, no. Suppl. 2, pp. 19–31, jan 2014.

[2] *Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019*. Geneva: World Health Organization, 2020.

[3] A. Verma, "Worldwide deaths by country/risk factors," https://www.kaggle.com/datasets/varpit94/worldwide-deaths-by-risk-factors/data, 2021.

[4] Geof, "Iso country codes with alternative country names," https://www.kaggle.com/datasets/gbertou/iso-country-codes-with-alternative-country-names, 2020.

[5] B. Tunguz, "Country, regional and world gdp," https://www.kaggle.com/datasets/tunguz/country-regional-and-world-gdp, 2021.

[6] R. Pollock, "population," https://github.com/datasets/population?ysclid=lth1wz7isi105780917, 2023.

[7] H. Steinhaus *et al.*, "Sur la division des corps matériels en parties," *Bull. Acad. Polon. Sci*, vol. 1, no. 804, p. 801, 1956.

[8] D. Arthur, S. Vassilvitskii *et al.*, "k-means++: The advantages of careful seeding," in *Soda*, vol. 7, 2007, pp. 1027–1035.

[9] "Numpy," https://numpy.org/, accessed: 2024-03-11.

[10] "Scikit-learn," https://scikit-learn.org/, accessed: 2024-03-11.
[11] "Pandas," https://pandas.pydata.org/, accessed: 2024-03-11.
[12] "Geopandas," https://geopandas.org/, accessed: 2024-03-11.
[13] "Matplotlib," https://matplotlib.org/, accessed: 2024-03-11.
[14] "Yandex translate editor," https://translate.yandex.ru/editor, accessed: 2024-03-11.