t-SNE Analysis of Cause of Death Data Across Continents

Keaton Raymond
kraymond@uccs.edu

Abstract

This analysis aims to discover a relationship between the cause of death and the six continents that have a noteworthy human population (i.e., excluding Antarctica). The results overall showed a loose correlation between cause of death being disease/infection related and Africa, which is the most impoverished continent [5]. Besides this loose correlation, other conclusions cannot be reliably extrapolated; however, with further analysis, the data would suggest that there could be correlation discoveries regarding the relationship between the less impoverished continents and specific types of violence leading to death.

Introduction

For this assignment, I was tasked with finding a high dimensional dataset (at least 10 variables and thousands of observations) and performing a t-SNE analysis on the data. t-SNE analysis is a method of visualizing high dimensional data in a low dimensional form. I had to perform the analysis on the base set of data, then remove some variables and run the analysis again. This process was then performed again, running the analysis a total of three times. In this paper I will explaining the data itself, the analysis process, and what that analysis might suggest.

Data

The process of finding the data proved to be the most difficult aspect of this project overall. Nearly all the datasets I found either did not have enough observations, did not have enough variables, did not have a variable that I could separate the clusters by, or did not have any correlation. This meant that there was no way to run the t-SNE analysis and provide any insights, as it would just look like a randomized scatterplot.

The data that I settled on was a dataset that I found on Kaggle [1] that came from Our World in Data [3] that has data for the years 1990-2019 detailing the annual number of deaths by cause for 293 countries or regions. When I initially tried running the analysis, I used the “Entity” variable, which held the value for the country or region that observation was for. However, I found that the graph simply had far too many points and was unreadable. I then contemplated running the analysis on only five or ten countries, but that did not provide nearly enough observations. Eventually, I settled on performing the analysis based on continent. To do this, I found a second dataset [2], also from Our World in Data, that had a list of countries and what continent they belonged to. I then removed the unnecessary “Code” and “Year” variables from that dataset and merged the two datasets on the “Entity” variable. This gave each of the observations an additional variable that contained information regarding the observation’s continent. I could now run the analysis on all the usable observations, while also having a small enough number of categories that the data was interpretable.

To clean the data, I had to remove all NA values present in the dataset so that I was able to perform the t-SNE analysis. I tried doing this four separate ways. The first way I tried was simply dropping all the observations that contained an NA value; unfortunately, so many of the observations contained an NA value that this left me with less than 300 usable observations. The second idea that I had was simply putting a zero into all the cells containing an NA value. However, this caused the analysis to become very inaccurate and there was no correlation that was able to be found. Additionally, even if a correlation could be found, the results and the conclusions drawn from the correlation would likely not be representative of reality. The third idea that I had was to try and ignore the individual values that were NA while keeping the overall row intact; however, there was not a method that I was able to find to accomplish that that worked with the t-SNE function in R. The fourth way that I tried to remove these

NA values was by manually inspecting the data, seeing what variables had many NA values, and dropping those variables. From there, I was able to remove the observations containing NA values. This method ended up working the best in the end and I don't believe that it had much of a negative effect on the analysis because these variables weren't present in most observations. The variables that I remove were terrorism, malaria, exposure to forces of nature, and conflict and terrorism.

Results and Analysis

To decide what variables to remove for the second and third t-SNE analyses, I first tried removing five random variables, then an additional five random variables (for a total of ten). However, I found that while that did change the distribution of the datapoints, I was unable to interpret the results because what was removed was random and there was no rhyme or reason to the change. Due to that outcome, I tried a new approach that did allow me to interpret the results. I first took all the causes of death that I was looking at and created three categories: disease/infection (e.g., Parkinson's, HIV, or tuberculosis), violence/external causes/self-inflicted (e.g., drowning, poisoning or self-harm), and a miscellaneous category with the values that did not necessarily fit into either of the two previous categories. I then ran the first analysis on the base dataset that included all the categories other than those I removed due to their high volume of null values. For the second analysis, I only removed the variables that were in the first category, disease/infection. Then, for the third analysis, I only removed the variables that were in the second category, violence/external causes/self-inflicted. This separation allowed me to view the t-SNE plots dependent on different categories, which allowed me to draw some conclusions based on which categories of cause of death have a stronger relationship with specific continents.

The first analysis (fig. 1) showed a large grouping of Africa in the center with a scatter of other data points surrounding it. There was not much separation of the categories outside of Africa other than Oceania (Australia).

The second analysis (fig. 2) (where the disease/infection variables were removed) caused a tighter grouping of Europe and North America. However, Africa was no longer as tightly grouped as in the first analysis, turning instead into several smaller, more spread-out groups, rather than one large group.

The third analysis (fig. 3) (where the violence/external causes/self-inflicted variables were removed) caused an even tighter grouping of Africa in the center. Additionally, the continents Asia, Europe, and North America were less random and had tighter groupings, as well forming a more uniform outline around Africa. This was the most clustered result of the three and provided the highest quality insights.

Conclusion

Overall, none of these clusters gave results that were particularly concrete. Nonetheless, there are still a few conclusions that are able to be drawn. The analysis that I found to be the most insightful was the third one (fig. 3), which highlighted causes of death related to disease/infection. This analysis showed Africa as having the strongest correlation to the cause of death of disease/infection compared to other continents. This result seems to make logical sense, since typically a higher rate of poverty brings with it a higher rate of illness as well as death from preventable/curable illnesses [4]. Since Africa is the continent with the highest rate of poverty [5], it makes sense that it would also have the strongest correlation with disease/infection being the cause of death.

The second analysis (fig. 2) I performed, which focused on the cause of death being violence/external causes/self-inflicted, also highlighted some interesting correlations within the non-African continents. This analysis showed stronger groupings within the other five continents, compared to the other two t-SNE analyses performed. With further analysis, a stronger correlation may become apparent between the specific types of violence-caused deaths and geographical location at a continental level.

I did not find the first analysis (fig. 1) performed to be insightful, other than being able to say that there is a clear correlation and grouping of the cause of death in Africa, and that the other continents did not seem to have a strong correlation. I believe that it is difficult to extrapolate a conclusion from the first analysis because there are so many variables that could affect the final result. However, the second and third analyses, which focused on a more specific cause of death, provided more valuable information that made it easier to extrapolate results. Unfortunately, all three analyses were relatively scattered and do not illustrate strong enough correlations for me to confidently state any certain conclusions. After performing t-SNE analysis on this data set, I believe that higher quality conclusions could be extrapolated utilizing a different statistical analysis method.

References

[1] Chavez, I. (2022, March 29). *Causes of death - our world in data*. Kaggle. Retrieved March 24, 2023, from https://www.kaggle.com/datasets/ivanchvez/causes-of-death-our-world-in-data?select=20220327%2Bannual-number-of-deaths-by-cause.csv

[2] *Continents according to our world in Data*. Our World in Data. (n.d.). Retrieved March 24, 2023, from https://ourworldindata.org/grapher/continents-according-to-our-world-in-data?tab=table

[3] Ritchie, H., Spooner, F., & Roser, M. (2018, February 14). *Causes of death*. Our World in Data. Retrieved March 24, 2023, from https://ourworldindata.org/causes-of-death

[4] *Poverty and health - the family medicine perspective (position paper)*. AAFP. (2019, December 12). Retrieved March 24, 2023, from https://www.aafp.org/about/policies/all/poverty-health.html

[5] ISSAfrica.org. (2022, July 13). *Africa is losing the battle against extreme poverty*. ISS Africa. Retrieved March 24, 2023, from https://issafrica.org/iss-today/africa-is-losing-the-battle-against-extreme-poverty

Appendix

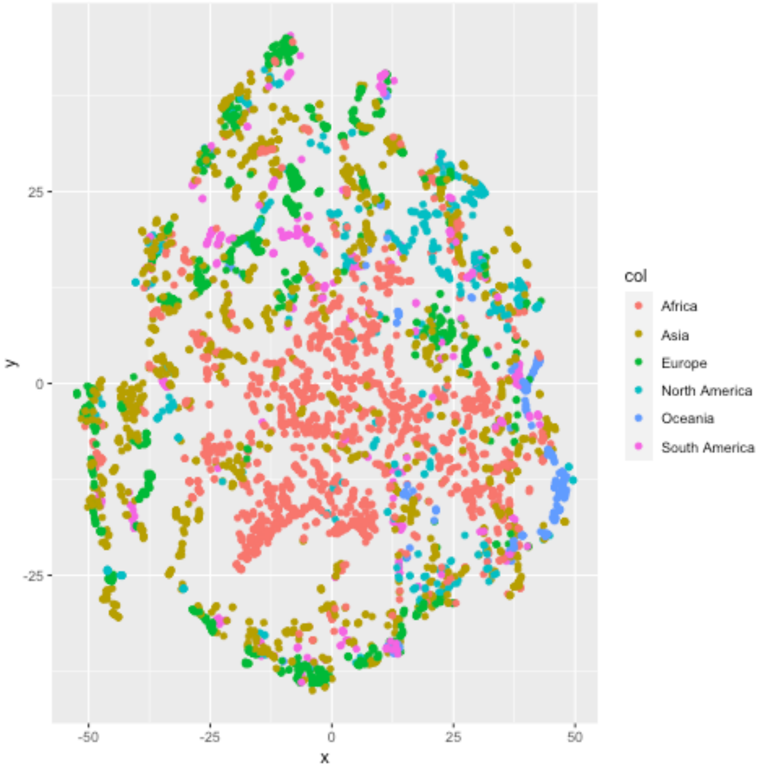

Figure 1: t-SNE plot with all variables

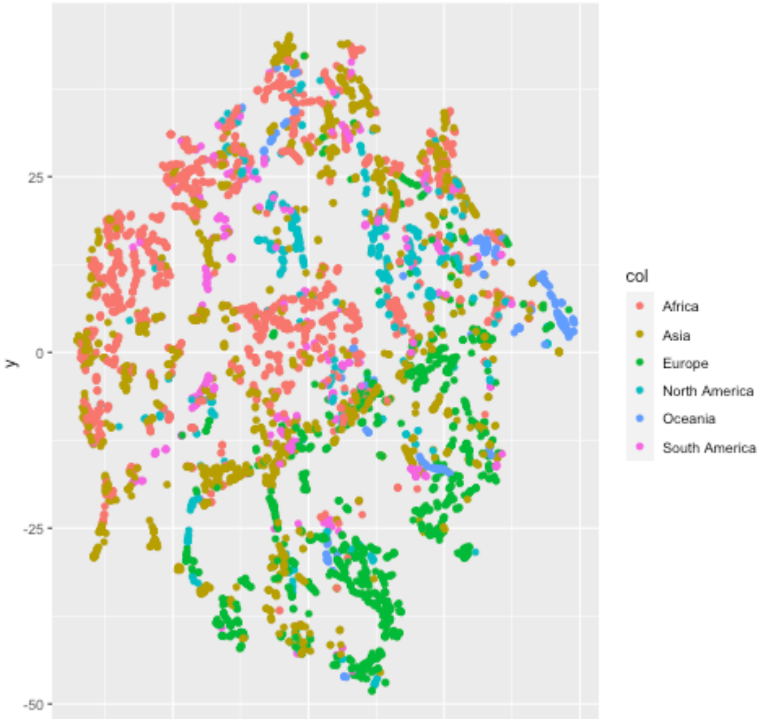

Figure 2: t-SNE plot with disease/infection variables removed

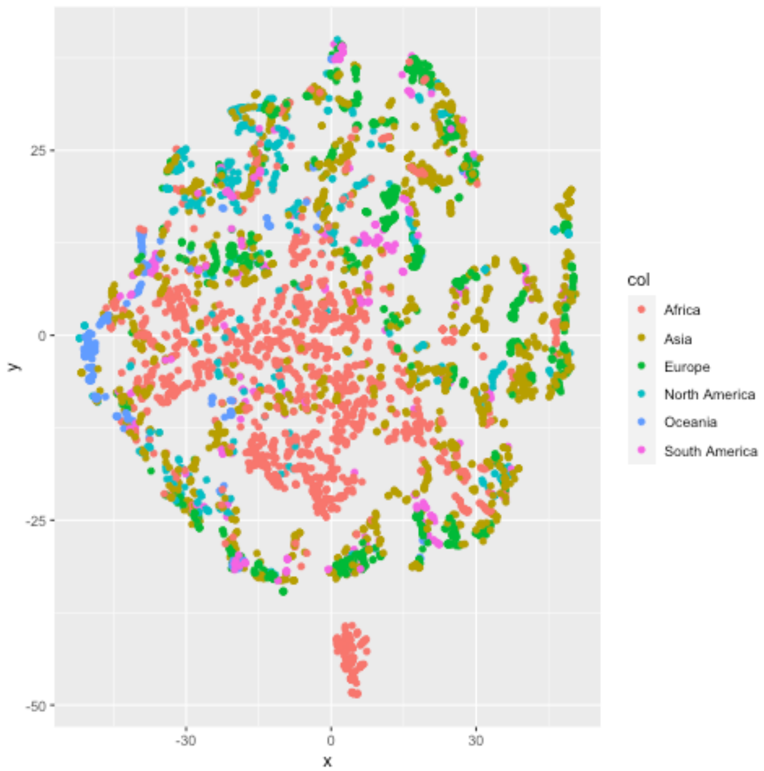

Figure 3: t-SNE plot with violence/external cause/self-inflicted variables removed