BM Knowledge Graphs

Project Proposal

Philip Salomons
i6154933

February 17, 2023

For a long time there has been the knowledge that health and business activities are correlated such as shown by Olivier et al where they find that business owners' health is directly affected by their business performance[3]. Whole cities and countries are affected by this factor, the driving economic activities of a region and the population's health are also linked. The OECD (Organisation for economic cooperation and development) has many papers and articles illustrating this, such as Burton et al [2] who show that population health increases business performance and output, suggesting political actions to improve both these areas. Magali et al. [1] show with a goal-oriented approach that the opposite effect also happens, where the well-being of the population increases when there is an increase in business productivity. The business level of a region is therefore an important contributing factor to its inhabitants well being and knowing how these two factors interact is very important for politicians looking to improve their region and for businesses who are looking to expand into certain regions.

The goal of this project is to expand on this knowledge. I will attempt to create indicators that show this relation and expand on previous work by selecting specific diseases and analysing how they affect and are affected by the business activities of the regions. Furthermore, I will try to infer using data from preceding years if the business is increasing health or the opposite, looking at the effect of specific diseases.

There are free and reliable data sources available online that will be used in this project. The World Bank's databases will be used as the business side. The available data sets include world development indicators, with over 200 indicators spread in 15 major databases. The indicators will be narrowed down to a handful of important ones during the project, but may include ones such as GDP per capita and business owner rates. This data is freely available and may be downloaded for projects as these. For the health side of this equation, I will use the data from the world health organization. This data is also freely available, and include over 50 data sets, for various indicators of health and well being. These will again be further selected during the project but may include HIV precense per region, child malnutrition or tobacco control in specific countries.

In this project, these data set will be turned into one by using data integration. I will use a resource description framework (RDF) to create a centralized set and attempt to answer these questions using the data. As the data are apart from each other and have different sources this will be the first step of the project and is expected to be done by March 4. After the data set is joined I can start answering questions that are important to the community, they will be answered by techniques such as correlation and r-squared values, these techniques are very simple and will be used only to get an estimate of potential. Then, more advanced techniques may be used such as machine learning to predict these effects and how changing them increases or decreases the value of the health indicators. These steps are expected to be done by march 15. Furthermore, this project will aim at predicting business performance using only health data. This will again be done using machine learning techniques. The benefit of this will be to allow business owners to make an informed decisions when expanding their operations. An informed decision on business prospects based on a population's health may save a company millions of euros and valuable time. These final steps are expected to be done by March 17. And finally, a report on the findings will be written by March 19 and delivered on March 21st.

According to the papers cited above, I expect there to be a correlation between these indicators. The goal is thus to find the important ones. Thus, I expect to find these important indicators and use them in combination with machine learning to expand the knowledge and decision-making process for important individuals and organizations in the society.

There, can be many things that can not go as planned, and therefore some contingency plans are in order. First of all I address the data. The data that I will use is freely available, but may be insufficient or not unable to be used, for this there are other data sets available online that can be used. For the business side, there are more information on the OECD data bases, whhich are also freely available, similarly, there are other health indicators freely online. The health data is more difficult to find on a global level if the World Health Organisations databases fail, multiple sets available on Kaggle and other sources may have to be combined for this goal. Finally, there may be difficulties storing and reading these large datasets and creating code that can handle this. The initial goal is to use Python to create these sets and extract useful information but a faster language such as C++ may be a better option and using PyBind I can integrate these two languages to improve the speed of the programs. Github will be used to track the progress and allow for backup in case of accidental deletion or a new function breaking the code. If the goals presented here turn out to be unrealistic or not related, a further step may be taken to find other relations, using a similar approach. These may include the relationship between the free personal time of the labour force and the business performance and/or health in a given region.

References

[1] Delmas, M. A., Delmas, M. A., and Durand, R. Measuring business impacts on well-being: A goal oriented approach.

[2] Tony, B., Suzy, M., and Tim, N. Improving economic policy advice.

[3] Torrès, O., and Thurik, R. Small business owners and health. *Small Business Economics 53* (08 2019).