UNIVERSITY OF CALGARY

Project Proposal

Determining Factors of Life Expectancy Across Continents and Countries.

Ali Afkhami (30271805)

Evan Losier (30022571)

Luisa Alejandra Sierra Guerra (30261956)

Ruby Nouri Kermani (30261323)

University Of Calgary

Faculty Of Science

Master In Data Science And Analytics

Course: Data 604

Winter 2025

Introduction

Life expectancy is one of the key indicators of well-being and development of a population. In this project we aim to study and investigate the effect of various broad factors such as **Economy and Politics (Ali), Nutrition (Evan), Air Pollutants (Ruby), and Violence (Alejandra)** on life expectancy across different countries and continents. By doing this, we will learn which factors are most correlated and influential in the calculation of life expectancy. The subject is interesting to us because life expectancy can vary a lot around the world and finding out why people in some areas live longer than people in other areas could help us all live longer and healthier lives.

Datasets

Our primary dataset for life expectancy was found on Our World in Data and includes information such as country, year, and life expectancy at birth. The dataset has 4 columns and 21565 rows, and is stored in one csv file. The main field we're interested in is life expectancy at birth, but we will use country and year to match the life expectancy data with other datasets. We will explore this dataset and other relevant datasets using SQL databases and queries. If any of the individual datasets are missing values for countries or years, we will clean them by either deleting the null rows or filling missing values with appropriate alternatives if removing nulls would be harmful to the analysis. If more datasets are beneficial for analysis, we may add them during later stages of the project.

The datasets being used to investigate **economy and politics** from Our World In Data are going to be used to explore the effect of various economic and political factors on the life expectancy of the population. The target values are going to be GDP per capita, V-Dem index and living above or below the poverty line. The datasets being used are:

- Poverty (3 columns, 2706 rows) - Includes three columns of the countries (as well as their aggregation as continents), the years and percentage of population living below the poverty line.
- Economic Growth (4 columns, 7064 rows) - Includes three columns of the countries (as well as their aggregation as continents), the years and GDP per capita.
- Human rights (4 columns, 33554 rows) - Includes three columns of the countries (as well as their aggregation as continents), the years and civil liberty index.

The datasets being used to investigate **nutrition** were all found on Our World in Data and will be used to explore the overall topic of life expectancy by comparing how a population's nutrition is related to their overall life expectancy. We suspect that all the datasets related to nutrition will be correlated with life expectancy, where countries with high crop yield, food supply, and less malnourishment will generally have higher life expectancies. The datasets being used are:

- Hunger and Undernourishment (4 columns, 4455 rows) - Includes columns for country, year, and prevalence of undernourishment ("Share of individuals that have a daily food intake that is insufficient to provide the amount of dietary energy required to maintain a normal, active, and healthy life.").
- Crop Yields (3 columns, 14577 rows across multiple datasets) - Includes datasets for different crops (currently corn, wheat, oats, and rice), which have columns of country, year, and crop yield (in units of tonnes per hectare).
- Food Supply (3 columns, 12750 rows) - Includes columns for country, year, and food supply (in units of kcal per capita per day).

The dataset being used to investigate **air pollutants** will explore the effect of different types of emissions on life expectancy. We suspect that all different types of pollutants will have a negative impact on life expectancy and will explore if any one is more impactful than the others. The dataset being used is:

- Emissions of Air Pollutants (11 columns, 63883 rows) - Includes columns for country, year, and 9 columns for different pollutants (ammonia, black carbon, carbon monoxide, methane, nitrogen oxides, nitrous oxide, non-methane volatile organic compounds, organic carbon, and sulfur dioxide, all in units of tonnes).

The datasets being used to investigate **violence** will explore the effect of different factors that contribute to overall peace and violence of countries and how they might affect total life expectancy. The datasets being used are:

- Civilian and combatant deaths in armed conflicts based on where they occurred (6 columns, 7176 rows) - This dataset includes the country, country code, year (ranging from 1989 to 2023), deaths of civilians in ongoing conflicts within a country (conflict type: all), deaths of unknown type in ongoing conflicts within a country (conflict type: all), and deaths of combatants in ongoing conflicts within a country (conflict type: all).
- GPI - The Global Peace Index (GPI) data is available in a report containing GPI information by country from 2008 to 2024, published by the Institute for Economics and Peace (IEP).
- Homicide (9 columns, 13180 rows) - This dataset includes information such as ISO, Country, region, homicide data by sex, homicide data by age group, the year (only for 2000, 2010, and 2019), population, and the crude rate.

Guiding Questions

1. Among the factors we identified such as economy, politics, nutrition, air pollution, and violence, which has the strongest impact on the life expectancy of the population?
2. How correlated are these factors with life expectancy and which one is the strongest?
3. How do the correlations vary based on the different continents and countries?

Methodology

1. **Data Collection:** Various datasets from different sources such as Our World in Data, World Health Organization, World Bank Open Data and United Nation Development Program will be utilized.
2. **Data Cleaning:** Each team member will be responsible for cleaning the data in their dataset using pandas or machine learning techniques such as KNN to deal with any missing values.
3. **Data Analysis:** Each group member will perform a variety of the required operations on the data for each specific factor to extract valuable information, analyze and visualize them.
4. **Comparative Study:** The refined data will be aggregated and combined into a single dataset and compared over the countries and continents to observe the overall effect of our separately analyzed factors on the longevity of the population.

Expected Results

1. Determination of the most influential factors on the longevity of the population.
2. Understanding how each factor contributes to life expectancy.
3. Identifying the effect of various policies on life expectancy across the countries and continents.

Conclusion

By investigating the effect of various elements on the life expectancy of a population, this study will provide a data-driven approach on understating the trends of life expectancy in various countries and continents. We will know our project has succeeded if we are able to identify key trends in how different factors affect life expectancy around the world and visualize their impacts.

References

Dattani, S., Rodés-Guirao, L., Ritchie, H., Ortiz-Ospina, E., & Roser, M. (2023). Life expectancy. Our World in Data. Retrieved from https://ourworldindata.org/life-expectancy

Food and Agriculture Organization of the United Nations. (2023). Share of people that are undernourished – FAO [Dataset]. With major processing by Our World in Data. Retrieved from https://ourworldindata.org/hunger-and-undernourishment

Hasell, J., Roser, M., Ortiz-Ospina, E., & Arriagada, P. (2022). Poverty. Our World in Data. Retrieved from https://ourworldindata.org/poverty

Herre, B., & Arriagada, P. (2016). Human rights. Our World in Data. Retrieved from https://ourworldindata.org/human-rights

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., ... & Bond, T. C. (2024). Ammonia emissions from all sectors [Dataset]. Community Emissions Data System (CEDS) – with major processing by Our World in Data. Retrieved from https://ourworldindata.org/explorers/air-pollution

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., ... & Bond, T. C. (2024). Black carbon emissions from all sectors [Dataset]. Community Emissions Data System (CEDS) – with major processing by Our World in Data. Retrieved from https://ourworldindata.org/explorers/air-pollution

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., ... & Bond, T. C. (2024). Carbon monoxide emissions from all sectors [Dataset]. Community Emissions Data System (CEDS) – with major processing by Our World in Data. Retrieved from https://ourworldindata.org/explorers/air-pollution

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., ... & Bond, T. C. (2024). Methane emissions from all sectors [Dataset]. Community Emissions Data System (CEDS) – with major processing by Our World in Data. Retrieved from https://ourworldindata.org/explorers/air-pollution

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., ... & Bond, T. C. (2024). Nitrogen oxides emissions from all sectors [Dataset]. Community Emissions Data System (CEDS) – with major processing by Our World in Data. Retrieved from https://ourworldindata.org/explorers/air-pollution

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., ... & Bond, T. C. (2024). Nitrous oxide emissions from all sectors [Dataset]. Community Emissions Data System (CEDS) – with major processing by Our World in Data. Retrieved from https://ourworldindata.org/explorers/air-pollution

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., ... & Bond, T. C. (2024). Non-methane volatile organic compounds emissions from all sectors [Dataset]. Community Emissions Data System (CEDS) – with major processing by Our World in Data. Retrieved from https://ourworldindata.org/explorers/air-pollution

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., ... & Bond, T. C. (2024). Organic carbon emissions from all sectors [Dataset]. Community Emissions Data System (CEDS) – with major processing by Our World in Data. Retrieved from https://ourworldindata.org/explorers/air-pollution

Hoesly, R. M., Smith, S. J., Feng, L., Klimont, Z., Janssens-Maenhout, G., Pitkanen, T., ... & Bond, T. C. (2024). Sulfur dioxide emissions from all sectors [Dataset]. Community Emissions Data System (CEDS) – with major processing by Our World in Data. Retrieved from https://ourworldindata.org/explorers/air-pollution

Our World in Data. (2023). Civilian and combatant deaths in armed conflicts based on where they occurred. Retrieved from https://ourworldindata.org/grapher/civilian-and-combatant-deaths-in-armed-conflicts-based-on-where-they-occurred

Ritchie, H., Rosado, P., & Roser, M. (2022). Crop yields. Our World in Data. Retrieved from https://ourworldindata.org/crop-yields

Ritchie, H., Rosado, P., & Roser, M. (2023). Hunger and undernourishment. Our World in Data. Retrieved from https://ourworldindata.org/hunger-and-undernourishment

Roser, M., Arriagada, P., Hasell, J., Ritchie, H., & Ortiz-Ospina, E. (2023). Economic growth. Our World in Data. Retrieved from https://ourworldindata.org/economic-growth

Roser, M., Ritchie, H., & Rosado, P. (2013). Food supply. Our World in Data. Retrieved from https://ourworldindata.org/food-supply