Susan Lee
July 27, 2023

Name of Dataset: **Cause of Deaths Around the World**

https://www.kaggle.com/datasets/iamsouravbanerjee/cause-of-deaths-around-the-world

Data Source

- Data Sourcing
  - The dataset is provided by Kaggle. Since it is an open-source website, there is a possibility of errors. However, Kaggle has a "usability" score for each dataset, and this dataset was rated the highest score of 10 which means it has 100% for credibility and completeness. For this reason, the dataset is trustworthy.
- Data Collection
  - The data is gathered from "Our World in Data", an open-source website that provides research and data to make progress against the world's largest problems. The dataset on Kaggle was updated 9 months ago which means the data has good timeliness.
- Data Contents
  - The dataset contains historical data of different cause of deaths for all ages around the world, such as meningitis and Tuberculosis. There are 204 countries/territories included for years 1990-2019.

Reasons for choosing this dataset

- My background is in the sciences, so I was curious about the top causes of death worldwide and if there are differences between countries.
- Good breadth of data because it looks at 31 causes of death for 204 countries/territories over 30 years.
- The most recent year is 2019 so the dataset is relatively up to date.
- There are a variety of causes of death, such as communicative and non-communicative diseases.
- The data is recorded using disability-adjusted life years (DALYs) which assesses health outcomes by both mortality and morbidity. This provides a more encompassing view on health outcomes.

Data Profile

- Shape of Dataset
  - 34 columns
  - 6120 rows
- Consistency Checks
  - 0 missing values
  - No duplicates
  - No mixed-type data
- Variable Types
  - 2 columns with string data (Country/Territory and Code)
  - 32 columns with numeric data (Year and various causes of death)
- Descriptive Statistics
  - Cardiovascular Diseases (73,160) and Neoplasms (37,542) have the highest mean values by far
  - For almost all of the causes of death, the minimum is 0
- Limitations
  - Since the data is from Kaggle which is an open-source website, there is a possibility of errors.

Questions to Explore

- What are the top causes of death worldwide? Are there exceptions for certain countries?
- What differences exist between countries? Can it be explained by the population size, wealth, or another factor?
- What are some trends either globally or within countries?
- Which countries have zero DALYs for certain causes of death? Is it sustained over time?
- What are some outliers in the data? What caused them?