# Journey of Over a Million Lives
## Project Report
## PV251: Visualization

### Karmazin Vasilii, učo 540500

### Faculty of Informatics, Masaryk University Brno

### January 5, 2024

## 1 Introduction

In this project, I aim to create an interactive visualization that allows users to explore the data by asking questions within the provided dataset.

To achieve these goals, I decided to use flexible D3.js. Additionally, since my data includes information about various countries, I used Globe.gl to visualize the data on a 3D model of the Earth, which enhances information perception.

The visualization was tested in the Google Chrome browser with Full HD resolution (1920x1080 pixels).

The project is available at GitHub: `https://github.com/e1four15f/AgeDatasetVisualization`

## 2 Data

I chose the Age Dataset from the Kaggle [1], which contains over a million records of various people, their lifespans, and occupations. I used country boundary data from the Natural Earth website [2], which is used in the Earth visualization.

The dataset processing included the following steps:

- Removal of missing fields.

- Normalization of columns, merging of similar data.

- Exclusion of the long tail of records that were difficult to process.

As a result of preprocessing, the total data count was reduced to approximately 722,000 records.

This preprocessing stage was carried out semi-automated, which may include some errors. A complete manual data preprocessing would have been too time-consuming. The data shows a strong rightward skew because the number of people grows exponentially.

| Column | Nulls % | Type | Nulls % | Type |
|---|---|---|---|---|
| Id | 0 | Id | 0 | Id |
| Name | 0 | Text | 0 | Text |
| Description | 5 | Text | 0.02 | Text |
| Gender | 11 | Factor (20) | 0 | Factor (2) |
| Country | 27 | Factor (5961) | 0 | Factor (173) |
| Occupation | 16 | Factor (9313) | 0 | Factor (6) |
| Birth year | 0 | Numeric | 0 | Numeric |
| Death year | 0 | Numeric | 0 | Numeric |
| Manner of death | 95 | Factor (206) | 0 | Factor (5) |
| Age of death | 0 | Numeric | 0 | Numeric |

Table 1: Dataset before and after preprocessing.

# 3 Design

I used the following types of charts for data visualization:

1. **Choropleth Globe Map**: This chart displays the number of famous personalities living in different countries. It also serves as an interactive filter for country selection.

2. **Crosstable/Heatmap**: Used to visualize the number of records with *Occupation* and (*Manner of death.*

3. **Boxplot**: Shows statistical analysis of the age of people by occupations, including parameters such as median, quartiles, and outliers.

The visualization includes filters for various parameters: gender, occupation, years of life, and country. There is also an option to view a person's page on Wikipedia.

# 4 Results

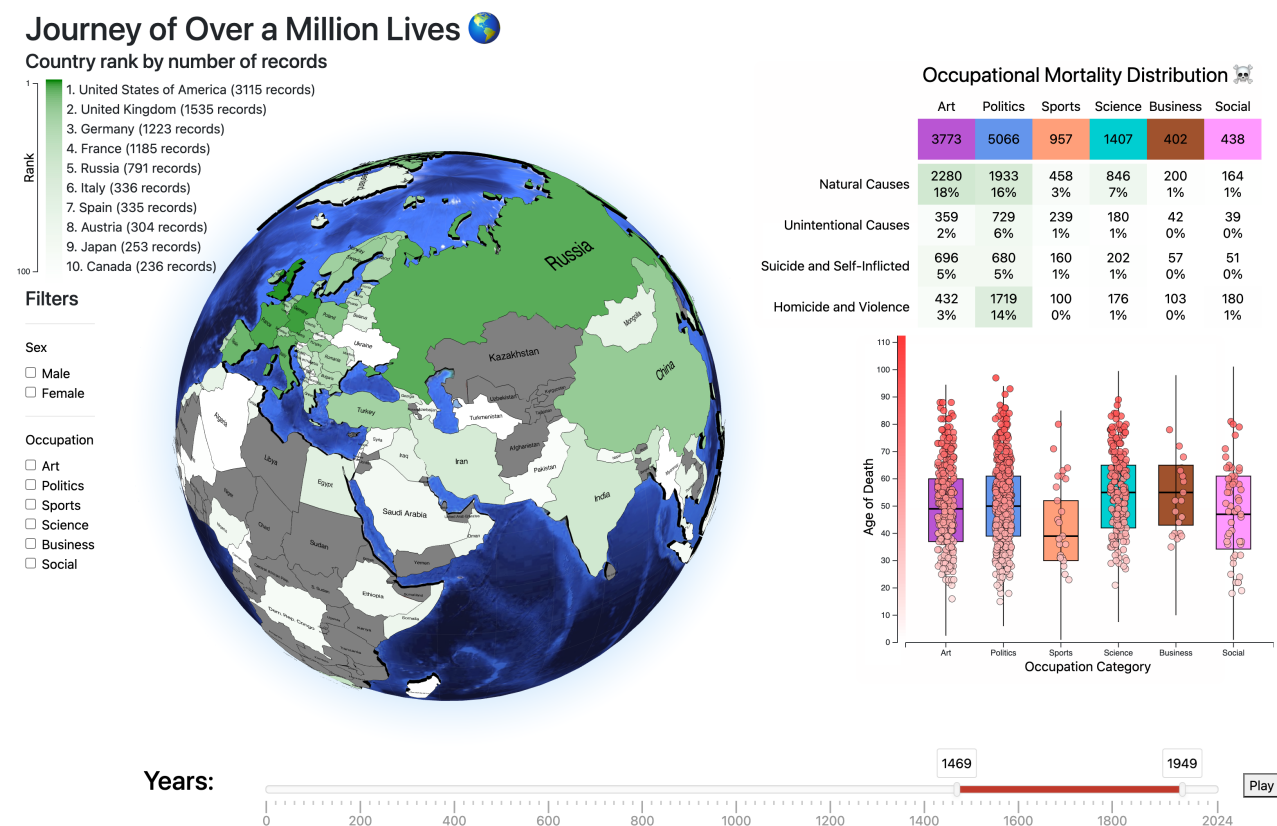The results of the visualization are shown in Figure 1.



Figure 1: Visualization screenshot.

1. The average human lifespan has significantly increased over the last centuries. It is observed that sportspeople tend to have a shorter lifespan on average, whereas scientists and business people show a longer lifespan.

2. There is a noticeable shift in the causes of death: from *Homicide and Violence* to *Natural Causes.*

3. Changes in the leadership in the number of famous personalities: Initially, China was a leader, but in subsequent years, the United States of America took the lead.

4. Before the beginning of the 20th century, most famous personalities were associated with politics.

5. Artists are the most common in most European countries, but political figures are more noticeable in Germany.

# 5 Conclusion

1. The interactive visualization turned out to be engaging and fun to interact with.

2. Looking at the data distribution is a good idea. In my case, most countries have only a few records, or there are insufficient records in the range of 0-1500.

3. D3.js is a well-suited tool for finished designs but is complex for quick experiments.

# References

[1] "Age dataset," 2022. [Online]. Available: `https://www.kaggle.com/datasets/imoore/age-dataset`.

[2] "Natural earth countries data," 2009. [Online]. Available: `https://www.naturalearthdata.com/%20downloads/110m-cultural-vectors/110m-admin-0-countries/`.