CZ4124 Assignment #1 - Technical Report

Name: Wang Anyi **Title**: Visualizing Mortality

1. Description of dataset

The main dataset used is titled "Age dataset: life, work and death of 1.22M people", which contains information including full name, short description of person, gender, country, occupation, birth year, death year, manner of death and age of death of 1,222,997 unique deceased famous people. The dataset is diverse in terms of geographical origins of people (original data contains 5961 unique values for 'Country'), diverse occupations (9313 unique values in original data), and consists of contemporary and historical persons with birth years ranging from BC 2700 to 2016. However, since data is sparse in earlier years (before 1800), most of the following studies will be based on data from contemporary times. Some attributes in this dataset also contains substantial amounts of empty data (eg. 'Manner of death' column contains 1,169,406 empty values). Since these (categorical) values cannot be appropriately estimated / interpolated, they were removed when parts of the analysis involve these attributes.

Other supplementary datasets were also used when appropriate:

- Countries-Continents.csv : used for mapping unique countries to the continents they belong to. This dataset contains 195 unique countries and 6 unique continents (Africa, Asia, Europe, Oceania, North America, South America).
- Life Expectancy Data.csv : This dataset was collected from the Global Health Observatory (GHO) data repository under World Health Organisation (WHO). It contains the recorded life expectancies of all contemporary countries from year 2000 to 2015, along with additional information of countries such as developed/developing status, population, GDP, average BMI of people, average number of years schooling, alcohol consumption per capita (litres), immunisation rate of various diseases (eg. Hepatitis B). This dataset was chosen since it was considered to have more accurate and holistic information about various countries, and could provide greater depth in the analysis on mortality in different countries.

[Note on terminologies: The terms 'life expectancy' and 'age of death' may be used interchangeably in this report. 'Age of death' is the variable of interest in the main dataset and mean values of 'Age of death' would serve as an estimate for the theoretical 'life expectancy' in various subpopulations.]

2. Objectives

The overarching objective of this project is to study and visualize the effects of different factors on life expectancy (Age of death). Specifically, the following questions will be answered:

- How does life expectancy generally change over time?
- How does life expectancy vary across different genders, and over time for each gender?
- Are there any major changes in the manner / reasons of death over time and how does it vary across different occupations?
- Are there any major events that resulted in a sharp change in life expectancy? If so, which locations / types of people were most affected by it?
- How does life expectancy vary across different countries and continents, and over time in these geographic locations?
- What is the impact of other factors, such as schooling, alcohol consumption and GDP of a country have on life expectancy?

Different visualization methods will be deployed to answer these questions and the effectiveness of the visual designs and techniques will also be discussed.

3. Design Rationale

Why did you choose your particular visualisation techniques and why did you think they were effective in helping you achieve your objectives? How did the nature of the data influence your visualisation design? Was there an overall strategy used in the design of the presentation?

3.1 Choice of plots

- Line plots (figures 1.3, 1.6, 1.7 and 1.14) were used to represent trends and changes in age of death over time, in which age of death are ratio values. The y axis begins at 0 to truthfully convey age of death without making certain ages seem too small. The interactive chart also allows for hovering over the line to see exact values.
- Ridgeline charts (figures 1.8 and 1.16) from the plotly library was chosen over regular violin plots / overlapping histograms since the plots arranged vertically and slightly overlapping each other allows for easier comparison of mean and median values across categories (compared to violin plots where the distance between the violins are much larger). I also found the symmetry in symmetrical violin plots rather redundant while the ridgeline plot is more concise in describing distribution since it only shows half of a typical violin plot. Overlapping histograms, on the other hand, would be result in too much visual occlusion when there are 6 or 7 categories in my case. Compared to boxplots, it was also able to show more insights, such as in figure 1.8, the deaths of people below 50 years old were particularly high in African continent. This would not have been demonstrated as effectively in a boxplot, which only shows median, IQR and outliers.
- The interactive scatter plot in figure 1.9 is used to represent 5 dimensions of information (population, continent, GDP per capita, life expectancy, years). Population (ratio type) is represented using the perceptual channel of size of points, where larger circles represent bigger continent population. One area of improvement would be to use apparent area for size of points instead, since Steven's power law tells us that we perceive areas with some degree of error. Continent (nominal) is encoded using different categorical colours, whereas life expectancy and GDP (ratio) are represented using position of points (y and x axis respectively). Time in years (ordinal value) is adjustable from the slider of the interactive chart.
- The stacked bar graph in figure 1.15 is effective in representing proportions that add up to 1. Proportions of manner of death (ratio value) were represented using the perceptual channel of size (length) in the chart. This is a similar case for the stacked area chart in figure 1.13, which also uses position to encode changes in proportion over time. Since it is difficult to compare the changes in areas of categories located in the middle of the stacked area chart, as well as exact values (proportions) of each category, I have tried another interactive line chart (figure 1.14) to encode the same information.

3.2 Choice of colour and colour palettes

Colours have been deliberately chosen based on the type of data and the intended connotation to bring across. For instance:

- male and female (nominal data) in figure 1.4 were represented using blue and red respectively, to follow the conventional colours associated with the two genders.
- The bars for World War 1 and 2 in interactive line charts (figures 1.3, 1.6, 1.7 and 1.14) have a brick red colour, which connotates danger. This makes it easier to understand the reason for the decline in age of death.

- Line chart in figure 1.3 has a green smoothed line to indicate general increase in mean age of death, since green usually indicates positivity.
- Nominal values of manner of death in figure 1.5 were encoded using a categorical palette with relatively serious colours (dark red and muted colours). The colours used to represent natural causes of death and accident were more muted compared to the other two because the intention was to bring across the fact that 50% of the subpopulation were affected by homicide and suicide.
- Continents (nominal values) were encoded using categorical colour palettes in figure 1.7. Colours were also chosen based on my impression of the continents (eg. Africa in dark green associated with plantations and forested areas, Asia in olive associated with land and mountains in the continent, Europe in royal blue, etc.)
- In the ridgeline charts in figures 1.8 and 1.16, even though continents and occupations are nominal values, I decided to use a sequential palette of darkening reds to highlight the decreasing mean age of death across the categories, especially since the violin plots are already sorted in decreasing order. Since black is commonly associated with death, I used darker red to represent categories with higher mortality.
- A continuous sequential palette was also used to represent life expectancies (ratio values) in countries in the choropleth map (figure 1.10). This palette was useful in representing the relative magnitudes of life expectancies across the world and visualize changes across the years easily, although it isn't effective in representing exact values (since the human eye cannot distinguish minute changes in shades of colours).
- A categorical palette was chosen to represent manners of death (nominal) in figures 1.13, 1.14 and 1.15. On hindsight, perhaps a more serious palette should have been chosen (darker colours instead of the bright ones used).
- As for the correlation heat map in figure 1.11, a diverging palette was used to represent Pearson's correlation values, since the data is interval in nature and has a mid-point at 0.

Colour counts have also been kept to a minimum as much as possible to avoid confusion.

4. Knowledge Application

What specific principles and concepts learnt in the course was used and how were they applied to the design of the visualisation. Describe with examples how you have taken human visual perception constraints and considerations into your design.

4.1 Gestalt principles

- Proximity

  In figure 1.3, the age of death in 1500 and 2020 were labelled close to the head and tail of the line to make the association of the values (51.7 and 80.2) with the line more pronounced.

  In figure 2.5, the bars representing manners of death within each occupation were put right beside each other while leaving gaps in between the occupations. This makes it easier to visually classify the occupations and compare the differences among them.

- Similarity

  In figure 1.15 of the stacked bar graph showing the proportion of manners of death for each occupation, colour similarity was used to associate sections of the bars to their corresponding manner of death, which also makes it easier to compare lengths of sections of the same manner of death among occupations. There were also other charts (figure) related to visualizing

proportions of manner of death. Consistent colour mapping was used across these three charts to encode the manner of deaths (eg. Natural causes was represented in red across all charts), so that it would be easier for viewers to follow through the flow of the presentation without having to reassociate a different palette to the same categories of data when the chart is changed.
In figure 1.12, there were 2 categories of countries, developed and developing. Colour similarity was again used to unify points within the same group. A better way would have been to use redundant encoding of shape to further enhance the separation.
In figure 1.4, the 2 split violin plots were also separated into one for mainstream genders and another for transgenders. Males were consistently denoted in blue while females in red across all charts.

- Simplicity

  Ridgeline plots (figure 1.8 and 1.16) have been sorted by decreasing mean age of death across the groups to ease interpretation. Similarly, the heatmap in figure 1.11 was also sorted by decreasing Pearson's correlation coefficient. Data to ink ratio was minimized when appropriate, such as by not adding grids / enclosures in chart when unnecessary.

  Line charts figures 1.3, 1.6, 1.7 and 1.14 were also smoothed using rolling window with gaussian kernel to accentuate the general trend instead of distract viewers with noise (numerous small fluctuations) in the data.

- Connectedness

  Line charts in figures 1.3, 1.6, 1.7 and 1.14 gives a sense of connectedness of points across many discrete years. It helps in the visualizing of trend and grouping of categories (eg. points belonging to the same gender in figure 1.6 were connected together by lines, so that they are associated as belonging to the same gender). This way viewers can focus their attention on trends and relationships between categories instead of trying to make sense of the categories in which points belong to.

- Enclosure

  Line charts in figures 1.3, 1.6, 1.7 and 1.14 exploits enclosure to highlight the changes in age of death during periods of war by the appearance of the brick red blocks to mark out the time range in which wars took place. This helps us to disassociate these section of the graph from the rest to focus on the drop in mean age of death for a while.

- Symmetry

  While we tend to perceive symmetrical elements as belonging to the same group and a sense of symmetry tends to appeal to the human mind, we also tend to quickly notice dissymmetry with a sense of oddness, and therefore perceive the difference of the 2 sides in elements with some dissymmetry more quickly. The split pair plot in figure 1.4 utilises this idea of asymmetry to draw attention to the differences in KDE distributions of age of death in different gender groups. We can then quickly notice the difference in median values, the difference in extent of skewness and shapes of different genders.

4.2 Human visual perception

Some plots in this project exploited preattentive processing to draw attention to some data points of interest. An example is the pair plot in figure 1.12, where the data points for the developed countries are highlighted in a brighter color of orange while developing countries were in blue. This brings

attention to how developed countries have higher life expectancies and allows viewers to pay attention to the distribution of various variables in developed countries.

In the interactive line plots (figure 1.3, 1.6, 1.7, 1.14), the lines could be muted (decrease alpha value) so that the desired lines stand out, which is also a form of preattentive processing. The 2 bars for world war one and two would also be highlighted in brick red when the legend is clicked to capture attention to those points.

Visual reference grids have been added to all of the line plots, making it easier to compare trends and behaviour of the line graphs. In figure 1.3, for example, it is much easier to tell that there is an increasing trend because of the presence of a horizontal reference line (from the grid) at y=60. It also made it easier to visualize the changes gradient of the line when there is a reference (gradient is much gentler before 1900 and steeper after 1950s). Reference lines makes it much easier to estimate values at certain points without having to hover over points, such as the dip of the second world war (with the help of the grids, we can easily tell that it occurred between 1940-1950 and the mean age of death was in the low 50s). The Weber's Law states that the longer the length, the greater the difference required to notice a change. Especially since the y axis (age of death) was truthfully configured to start at 0, the small fluctuations, the changes in gradient / dips in graph are even more difficult to notice (since the line now takes up only less than half of the vertical space available). Reference grids would therefore assist in the perception.

In figure 1.15, the goal was to visualize how the proportion of manners of death varies among 5 different occupations. Since the proportions summed up to 1, there were mainly 2 choices of charts to choose from, the pie chart and the stacked bar chart. However, Steven's power law tells us that length is an easier stimulus to perceive value compared to area/angle in the pie chart. Therefore the horizontal stacked bar chart was chosen over the pie chart. However, for figure 1.5, also with the intention of showing the proportion of deaths (among transgender female gender group), the pie chart was chosen instead. This is because the goal of this chart was to highlight that homicide and suicide accounted for around 50% of deaths in this subpopulation, and the pie chart is very effective in doing so because 50% would be a 180-degree line cutting across the pie. With the homicide and suicide sections filled with bolder colors of red and blue, it is fairly easy to perceive this semicircle (50%) using the pie chart.

5. Novelty

Highlight what you think were original contributions in your data visualisation that you are particularly proud of. Also provide a brief description of how some of the major reference sources [2] have contributed preliminary ideas and rendered technical help during design and implementation of the project.

While there have been numerous studies done on societal factors contributing to life expectancies and changes in life expectancies over time, I believe these studies do not delve into the manner of death and occupations of people, as well as interactions between occupation and manner of death. This project also provides some insights on the transgender community (figure 1.4, 1.5). In general, this project combines multiple datasets with the intention of providing a more holistic analysis about mortality (including analysis concerning countries, continents, attributes of countries such as GDP percap, events such as the world war, occupations and genders).

The plotly violin plot library [4] has rendered support and inspired me to create the ridgeline plot over traditional violin plots (rationale for this plot is discussed in section 3). I found this plot much more effective than regular violin / boxplots / overlaying histograms in my case. The interactive scatter plot (figure 1.9) was also able to encode 5 dimensions of data into a single plot effectively.

Besides the visualisation plots, I am also particularly proud of being able to uncover some insights through the visualizations, such as which continents were most affected by the war, reasons for low age of death in transgender female community etc.

6. Technical Challenges and Innovation

What visualisation tools did you employed in exploring your data and creating your visualisation and presentation? Describe any noteworthy technical contributions done during the design and implementation of the visualisation. What were the technically challenging aspects in creating the visualisation?

Some of the visualization tools I used for exploration involves the word cloud (figure 2.4) to identify the major categories of manner of death to focus on in the main plots, as well as basic plots for exploratory analysis provided by the seaborn / scipy library, such as histogram, qqplot and pairplots to gain some insights on the distribution and relationships of attributes before working on more sophisticated plots.

One of the main technical challenges was that due to the sheer amount of data, it was difficult to plot effective charts for some of the categorical attributes due to the large diversity of unique values. Hence a lot of pre-processing and cleaning has to be done. Some examples were:

- There were initially 21 unique gender values in the original dataset. Some of which has redundant / confusing labels such as ‘Transgender Female; Female’, and ‘Female; Male’ (details in figure 2.1). Some of the values, for instance ‘Transgender Female; Female’ will be renamed as ‘Transgender Female’, whereas confusing values or those with very low count (1 or 2) will be removed. Eventually, the data will consist of only 4 gender types, ‘Male’, ‘Female’, ‘Transgender Male’ and ‘Transgender Female’.
- Country had 5962 unique values in the dataset. We know that there are currently 190+ countries in the world at the moment, but some of the country values were ancient geographic areas like ‘Tang Empire’ (which was renamed to ‘China’), or former countries before colonisation. There were also quite a number of redundant names such as ‘England’ and ‘Great Britain’ (which were renamed to ‘United Kingdom’). Please refer to figure 2.2 for details on renaming of countries. Other countries that could not be handled were removed. These country names also had to be used to merge with another supplementary dataset [1] to retrieve continents for plotting, hence consistent naming had to be ensured (eg. ‘South Korea’ vs ‘Korea, South’).
- The ‘Manner of death’ column had 207 unique values initially. Similarly, some of them were renamed so that they are grouped into the umbrella of bigger categories (see figure 2.3). Other smaller categories were grouped into ‘Others’, so that there would eventually only be 7 unique manners of death.

These pre-processing steps, although tedious, makes the output plots much more effective since it would be difficult to direct human attention to important elements when numerous unique values are depicted on plots.

Besides that, the dataset also consisted of data of people whose death year dated back to as far as BC 2700s that caused me some difficulties in determining the range of years to extract for graph plotting. Data points in this dataset typically gets more sparse the more historical/dated it is, hence plots including older data would be less accurate and meaningful, especially for plots such as how the proportion of different manners of death (figure 1.14) changes over time, which would require multiple data points from a single year. It would also make fluctuations/ trends in recent years less obvious since space is taken up by older data. I therefore chose cut-off years (after trial and error) such that plots could show reasonable amounts of both detail (such as decrease in life expectancies during war periods) and an indicative general trend.

Some challenges were also faced when deciding the type of plot to use. For instance for the different proportions if manner of deaths in 5 occupations feature, I had some trouble deciding between the stacked bar chart (figure 1.15) and a horizontal bar chart grouped by occupations (figure 2.5). The stacked bar graph was more concise / condensed but it makes comparing lengths of categories in the middle of the bar more difficult since they are not aligned to the same starting / ending positions. The horizontal bar chat, however, has all bars starting at the same location. However, I found it even more difficult to compare between occupations due to the large number of bars present. I eventually chose the stacked bar chart and annotated selected sections that were more crucial to the analysis.

References

List the main reference sources that have contributed preliminary ideas and technical help during design and implementation of the project. The source and weblink where the datasets were taken from must be listed as the first entry in your list of references.

Note: There is no page limit for References but all included references must be cited in your write up.

[1] Main dataset: https://www.kaggle.com/datasets/imoore/age-dataset

[2] Supplementary dataset: https://github.com/dbouquin/IS_608/blob/master/NanosatDB_munging/Countries-Continents.csv

[3] Supplementary dataset: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

[4] Violin plots on plotly: https://plotly.com/python/violin/

[5]

Appendix – Figures and Tables

Put relevant figures and tables in this Appendix and label them with appropriate numbers so that it is clear which figure or table you are referring to in your write-up.

Note: There is no page limit for this Appendix (but do be reasonable, be as succinct as possible and only include relevant figures here).

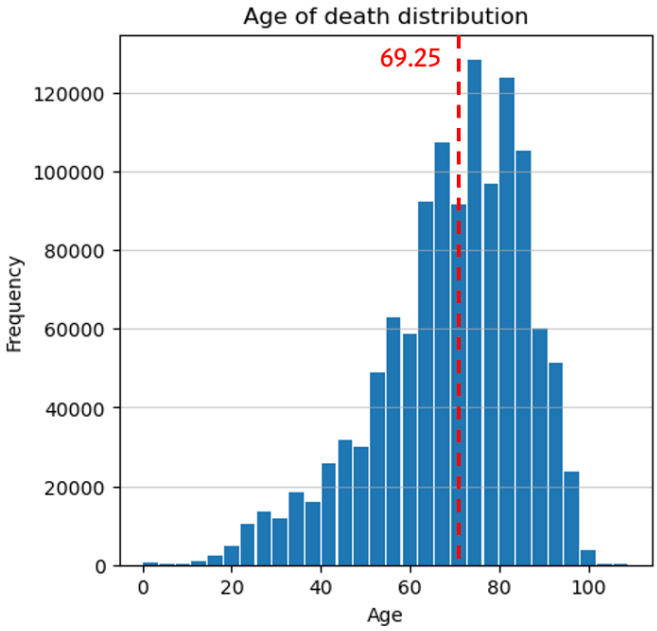

Figure 1.1: Histogram of age of death

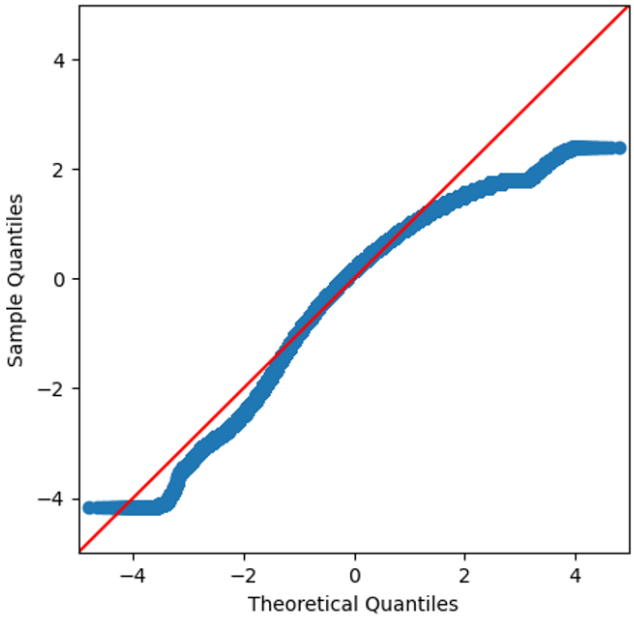

Figure 1.2: qqplot for age of death

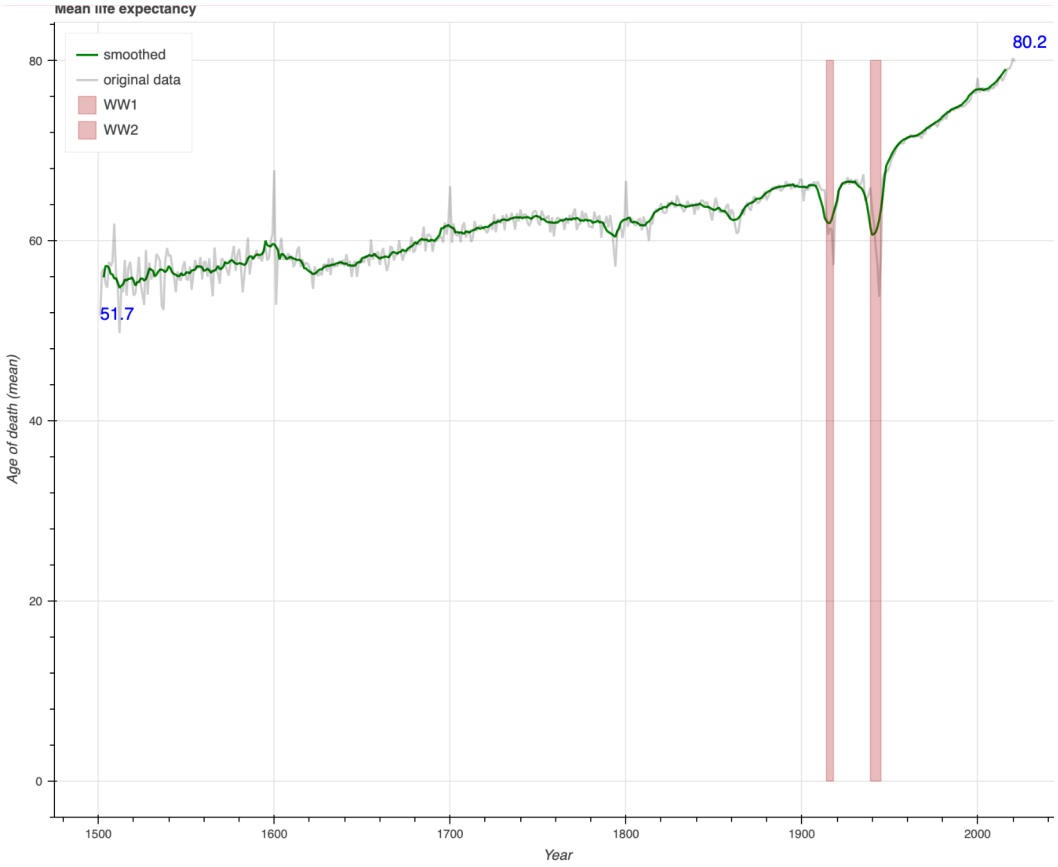

Figure 1.3: Interactive line graph of age of death over time (with smoothed)

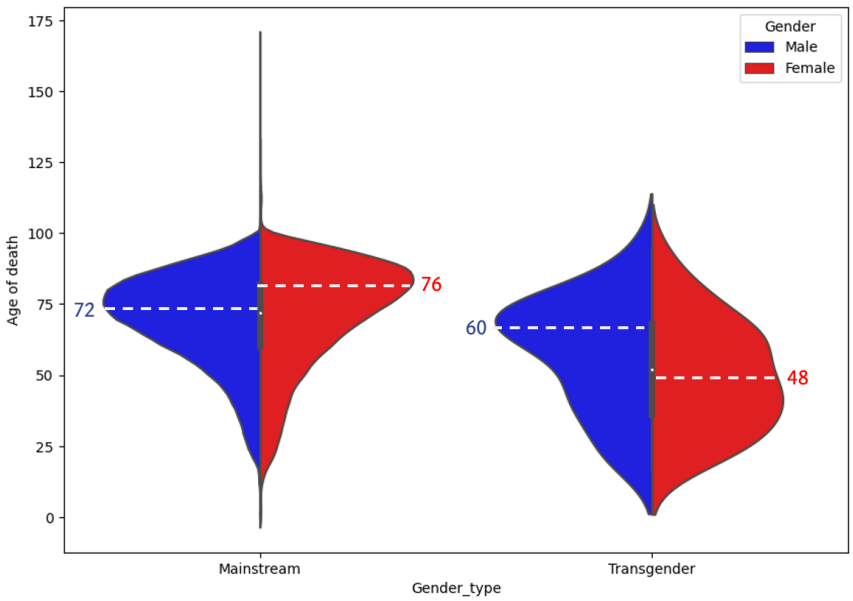

Figure 1.4: Split violine plot of age of death of genders

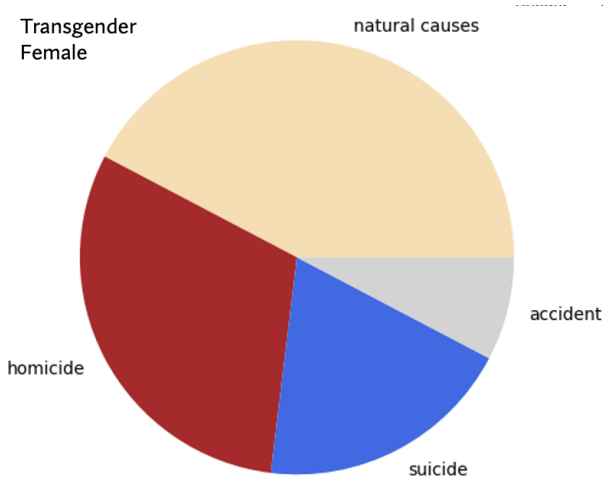

Figure 1.5: Manner of death among transgender female subpopulation

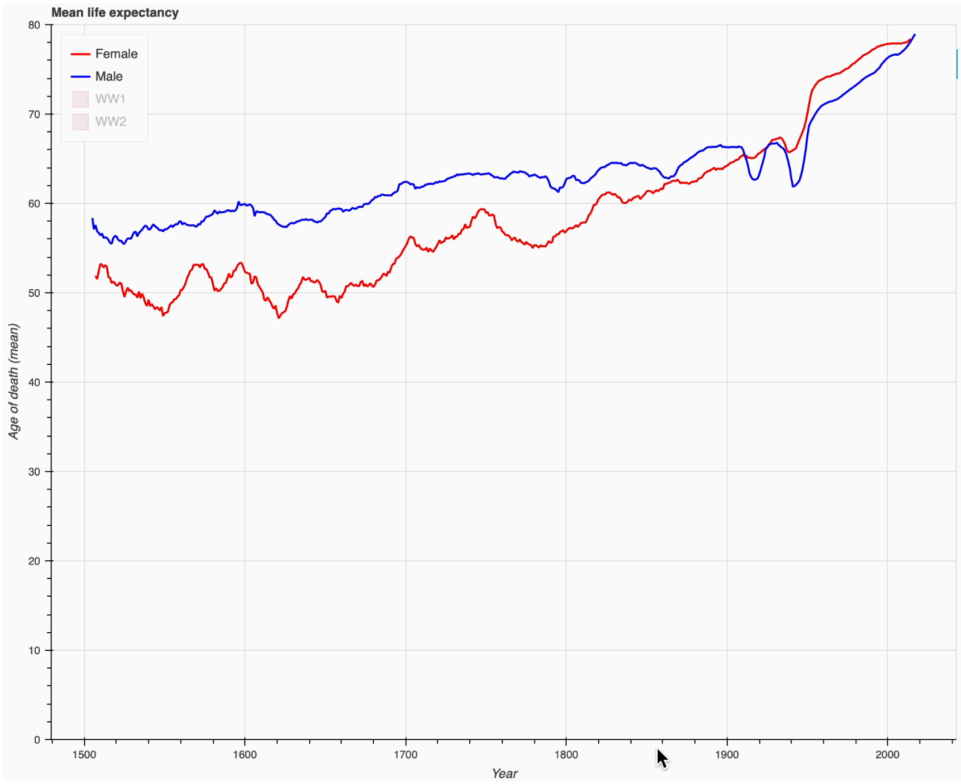

Figure 1.6: Interactive Line graph of age of death of genders

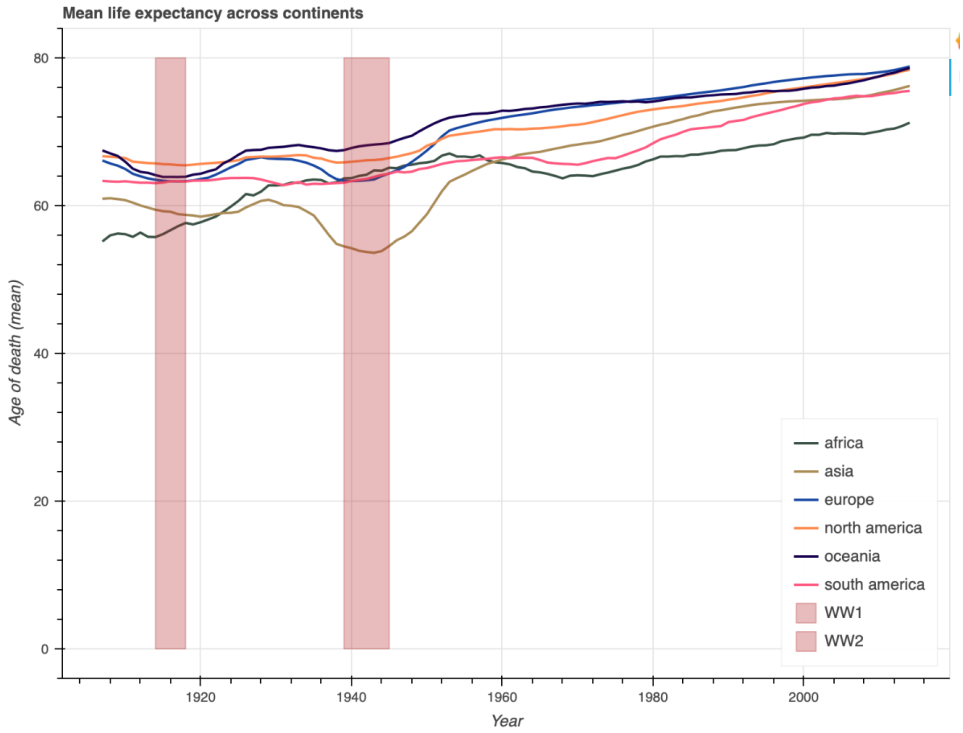

Figure 1.7: Interactive line graph of Age of death in different continents

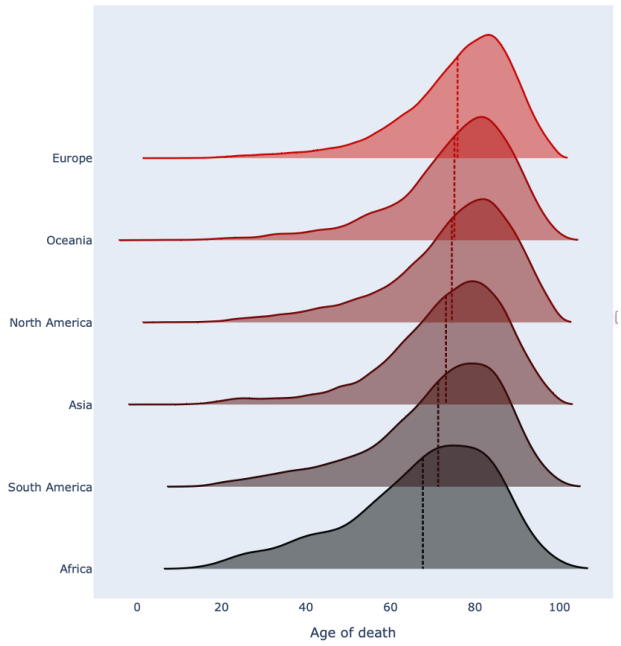

Figure 1.8: Ridgeline plot of age of death grouped by continents

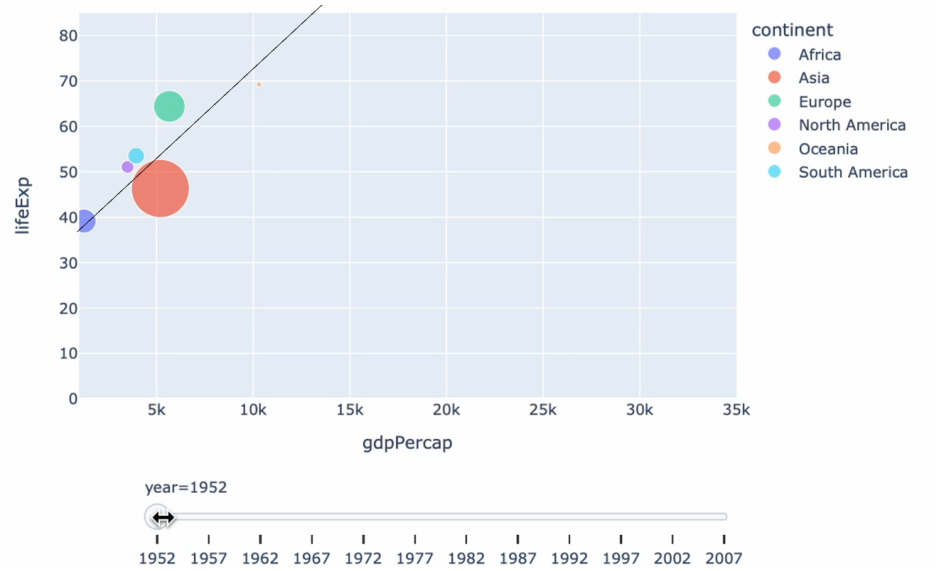

Figure 1.9: Interactive scatter plot of life expectancy vs GDP per capita

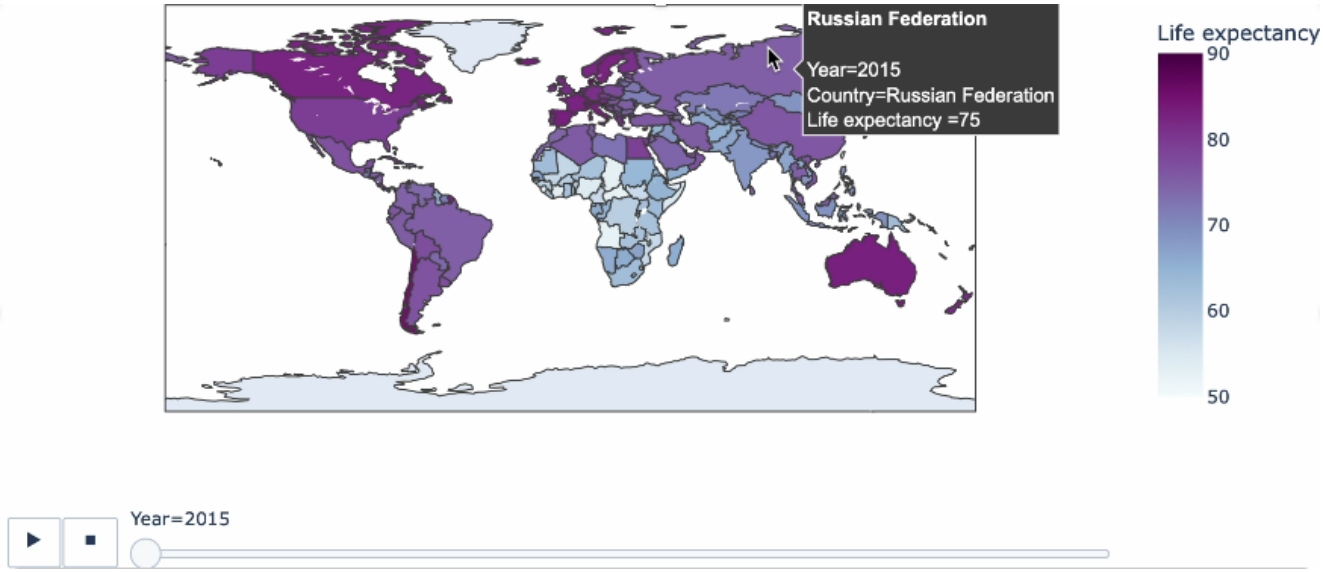

Figure 1.10: Interactive choropleth showing life expectancies in countries

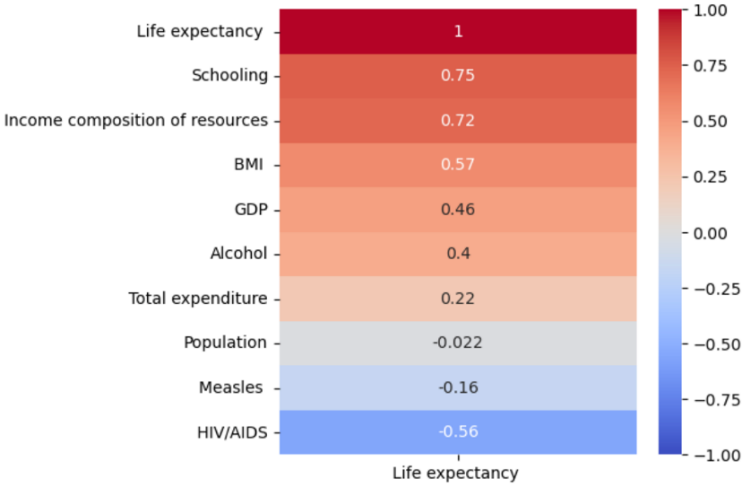

Figure 1.11: Heatmap of correlation of different variables with life expectancy

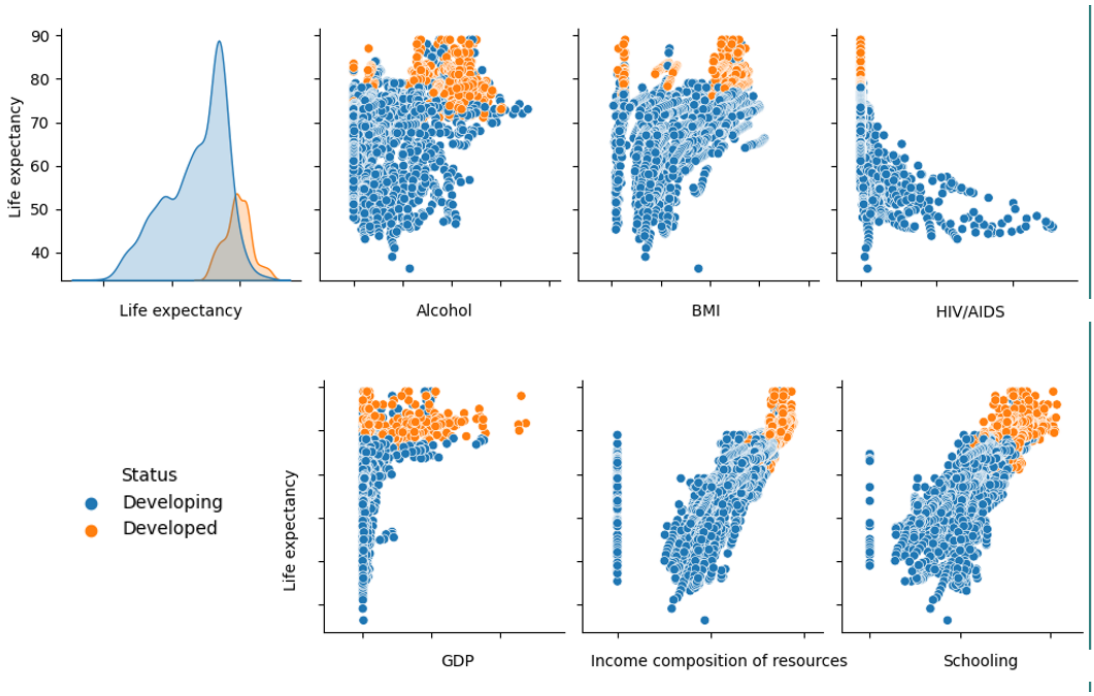

Figure 1.12: Part of pairplot of variables with life expectancy

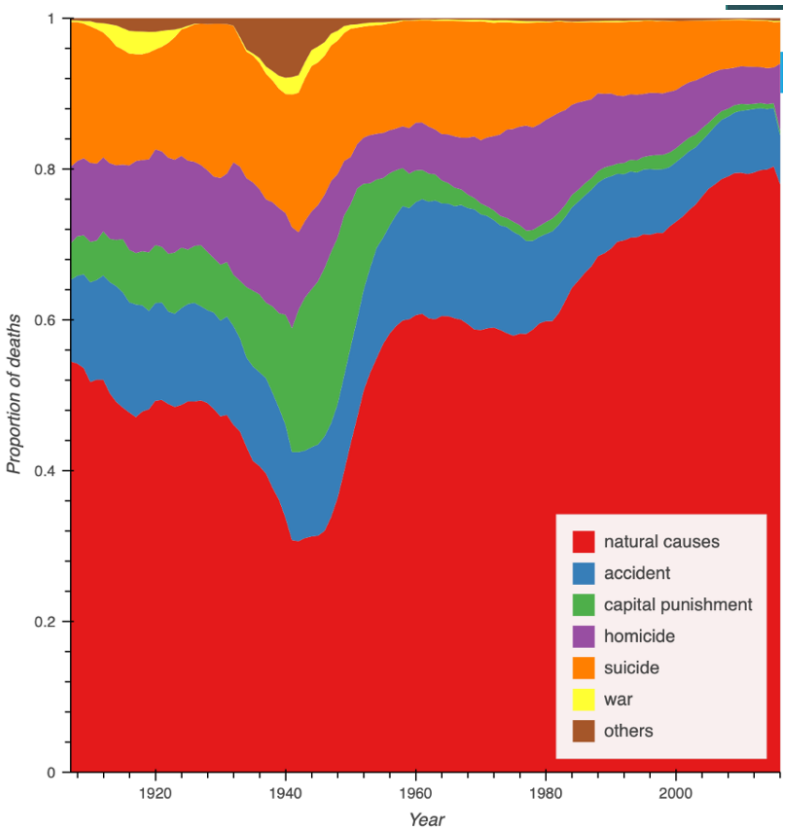

Figure 1.13: stacked area chart of proportion of deaths due to 7 causes over time

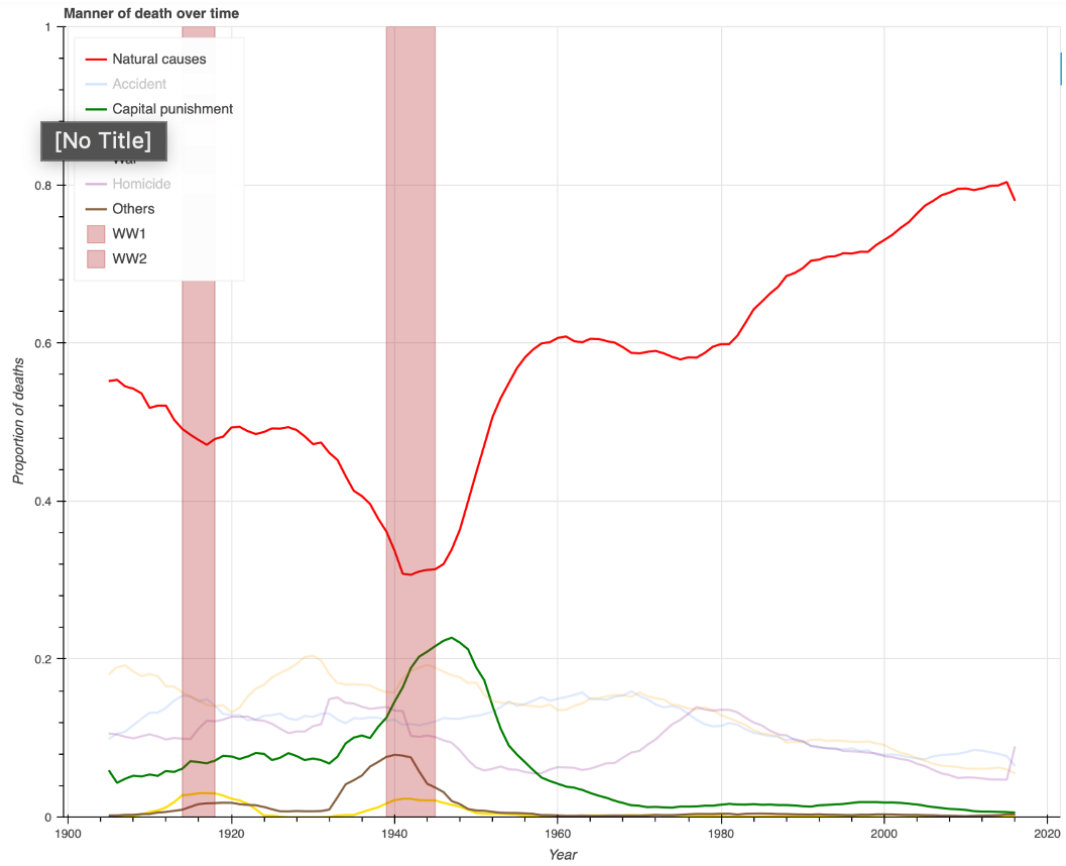

Figure 1.14: Interactive line plot of proportion of deaths due to 7 causes over time

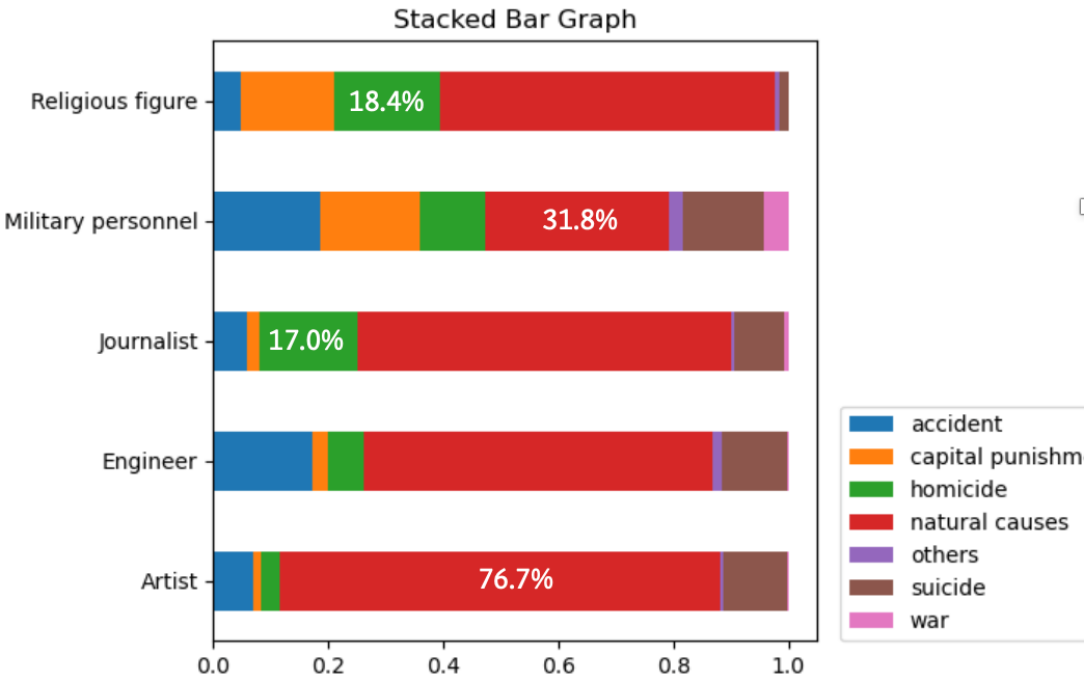

Figure 1.15: Stacked bar chart of manner of death grouped by occupations

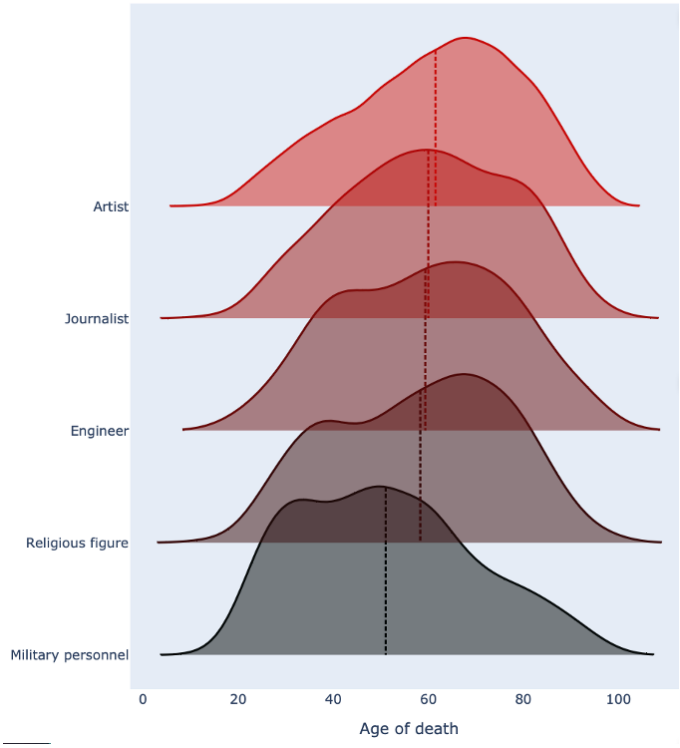

Figure 1.16: Ridgeline of age of death grouped by occupations

Male	981365
Female	107830
Transgender Female	79
Transgender Male	36
Eunuch; Male	18
Intersex	7
Female; Male	7
Eunuch	3
Transgender Male; Female	3
Intersex; Female	2
Non-Binary	2
Intersex; Male	2
Transgender Female; Male	2
Female; Female	1
Transgender Female; Intersex	1
Transgender Female; Female	1
Transgender Male; Male	1
Intersex; Transgender Male	1
Transgender Person; Intersex; Transgender Male	1
Non-Binary; Intersex	1

Figure 2.1: Unique gender values

```
{'United States of America':'US', 'Kingdom of France':'France', 'Nazi Germany':'Germany',
'Kingdom of the Netherlands':'Netherlands', 'Czech Republic':'CZ',
 'German Empire':'Germany', 'Kingdom of Poland':'Poland','Duchy of Milan':'Italy',
 'Kingdom of England':'United Kingdom', 'Grand Duchy of Tuscany':'Italy',
 'Republic of Florence':'Italy', 'Austria-Hungary':'Austria',
 'Kingdom of Great Britain':'United Kingdom',
 'United Kingdom of Great Britain and Ireland':'United Kingdom',
 'Kingdom of Portugal':'Portugal', 'Russian Empire':'Russian Federation', 'British Raj':'United Kingdom',
 'Austrian Empire':'Austria', 'Kingdom of Italy':'Italy', 'Kingdom of Sardinia':'Italy',
 'Republic of Venice':'Venice', 'Russian Soviet Federative Socialist Republic':'Russian Federation',
 'German Democratic Republic':'Germany', 'Papal States':'Italy', 'Kingdom of Prussia':'Germany',
 'Southern Netherlands':'Netherlands', 'Kingdom of Scotland':'United Kingdom', 'Scotland':'United Kingdom',
 'Grand Duchy of Finland':'Finland', 'Ottoman Empire':'Turkey', 'Dutch Republic':'Germany',
 "People's Republic of China":'China', 'Wales':'United Kingdom', 'Republic of Geneva':'Switzerland',
 'German Confederation':'Germany', 'Tang Empire':'China', 'Eastern Han Dynasty':'China',
 'Song dynasty':'China', 'Yuan dynasty':'China', 'Ming dynasty':'China',
 'Democratic Republic of the Congo':'Congo, Democratic Republic of', 'Tsardom of Russia':'Russian Federation',
 'ancient Rome':'Romania', 'Qing dynasty':'China', 'Puerto Rico':'US', 'Joseon':'Korea, North',
 'South Korea':'Korea, South', 'Republic of China (1912–1949)':'China', 'Taiwan':'China',
 'Polish–Lithuanian Commonwealth':'Poland', 'Kingdom of Naples':'Italy', 'Crown of Castile':'Spain',
 'Weimar Republic':'Germany', 'Kingdom of Romania':'Romania', 'Kingdom of Bavaria':'Germany',
 'England':'United Kingdom', 'Irish Free State':'Ireland', 'Northern Ireland':'Ireland',
 'Kingdom of Hawai‘i':'US', 'Great Britain':'United Kingdom', 'British people':'United Kingdom',
 'Americans':'US', 'Republic of the Congo':'Congo', 'Canadians':'Canada', 'North Korea':'Korea, North',
 'Italians':'Italy', 'Russia':'Russian Federation', 'Soviet Union':'Russian Federation',
 'Holy Roman Empire':'Romania', 'Venice':'Italy', 'Kingdom of Ireland':'Ireland', 'German Reich':'Germany',
 'Kingdom of Hungary':'Hungary'
}
```

Figure 2.2: Unique country values

```
replace_dict = {
                 'extrajudicial killing':'homicide',
                 'Eastern Front of World War II': 'war',
                'shipwrecking':'accident', 'unfortunate accident':'accident',
                'femicide':'homicide',
                'World War I':'war', 'Whitechapel murders':'homicide',
                'Spanish Civil War':'war', 'assisted suicide':'suicide',
                'Bombing of Berlin in World War II':'war',
                 'death in battle':'war',
                 'unnatural death':'others'
                }
```

Figure 2.3: Unique manner of death values

Figure 2.4: word cloud for ‘Manner of death’ in raw data

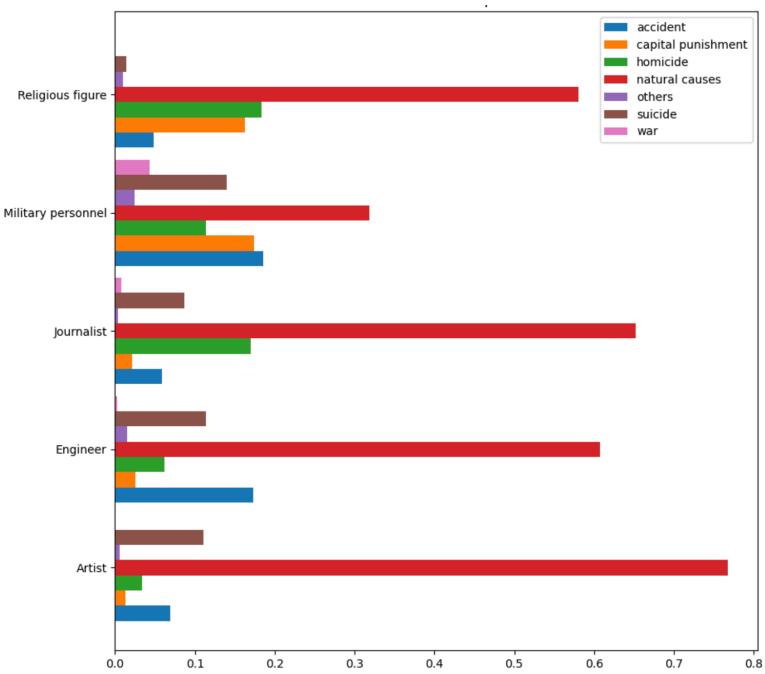

Figure 2.5: Occupation and manner of death horizontal bar chart (Exploration)