

Suicide Rates Overview

1985 - 2016



Student Name	Exam Number	Student Mail
Denisa Alicia Rissa	190402331	deris22@student.sdu.dk
Vivek Misra	4117637	vimis22@student.sdu.dk

1 Statement of Contribution

Both team members contributed equally to the project.

2 Abstract

This project presents a interactive visualization dashboard implemented in R and RShiny which analyses a dataset that covers records of global suicide rates from 1985 to 2015 and encompasses many factors which can affect suicide rate, such as region, demographic factors like age and sex and socioeconomic factors like population and GDP per capita. The visualization dashboard explores trends across time between each of these variables. The key insights found from this project's visualizations show that there is a steady decline in global suicide rates since the peak in the dataset in 1995. In Europe there are higher rates of suicide with a declining trend, but Oceania and Asia have an increasing trend. There is a universal gender disparity with men being 3.5 times more likely to die by women on average. The visualization does suffer from data quality issues, mostly being the varying amount of countries participating yearly.

3 Background and Motivation

Suicide is a complex issue worldwide and there numerous potential factors that can contribute to suicide rates across different demographics, regions and time periods. Visualisation and analysis of this data can help better understanding of the trends in suicide rates as well as inform prevention strategies by highlighting what characteristics and patterns might appear in the most vulnerable groups to suicide. The dataset for this project was chosen because of the variables present for analysis; its scope that encompasses multiple countries and years, which will allow the project to analyse both geographical and temporal patterns in suicide rates; and demographic factors such as age, gender and generation, as well as socioeconomic indicators such as GDP of the country, which will allow complex analysis and exploration of potential correlations across different variables.

4 Project Objectives

With the dataset, the project will explore the different variables provided in order to find any trends or patterns occurring across time and plot to visualize any potential relationships. Specifically, the project will focus on answering the following questions with the visualizations:

1. How have global suicide rates changed over time?
 - a. Are there any significant trends in the change of global suicide rate over time?
 - b. What do these temporal trends look like across genders and ages?
2. How does the suicide rate vary between different demographics, like gender and age?
 - a. Which demographics are most at risk, and how has this changed throughout time?
 - b. Are there disparities between the genders or between the age groups, and how have they changed throughout time?
3. How does the suicide rate vary geographically?
 - a. Which countries have the highest and lowest suicide rates?
 - b. Is there any regional clusters in terms of suicide rates?
4. Is the suicide rate in a country affected by its economic status and population?
 - a. Does the suicide rate increase or decrease countries during economic development?
 - b. How do suicide rates vary across countries with similar economies or populations?

5 Data

The dataset¹ used in this project is sourced from data collection website *kaggle.com* and created by an independent user, Russel Yates (username 'Rusty') for public use by the community; the dataset is compiled from different sources including the United Nations Development Program (2018)², the World Bank (2018)³ and the World Health Organization (2018)⁴. The dataset contains 27820 records and 12 variables: country, year, age, sex, age group, number of suicides, population, suiciderate per 100k of population, the country-year composite key, Human Development Index (HDI) for the year, GDP for the year, GDP per capita, and generation. The dataset encompasses regions globally and from the years 1985-2016. The rows in this dataset primarily count suicides by demographic, region and time and thus the columns are related to each other across the different variables as well as temporally; for example, for the first record:

```
country,year,sex,age,suicides_no,population,suicides/100k pop,country-year,HDI for year,gdp_for_year,gdp_per_capita,generation
Albania,1987,male,15-24 years,21,312900,6.71,Albania1987,, "2,156,624,900",796,Generation X
```

The number of suicides, population and suicides per 100k of population are per year-country-sex-age level. The age is actually a categorical variable (in this dataset these are discrete ranges: 5-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years, 75+ years) the categories are referring to the age range referred to in this record. For the number of suicides, this should be interpreted as, in Albania year 1987, 21 males of age group 15-24 years old died by suicide. The population refers to 312900 males of age group 15-24 years old in the year 1987 in Albania. Suicides per 100k of population are the reported suicides of this demographic per 100000 of this demographic living in the country at that year (21 reported suicides of males of age group 15-24 years old living in Albania in 1987, divided by 312900 males of 15-24 years old living in Albania in 1987 divided by 100000, which results in 6.17). In comparison, the HDI for the year, GDP for the year and the GDP per capita are for the whole country's population for that year (and are thus the same across records for different demographics in the same countries and same years).

The Human Development Index for the year is missing in a large majority of the records, only appearing in 1/3 of the records; this column will thus be discarded and not be analysed as a part of the project.

Generation is another problematic column in the dataset; this is to be interpreted per year-country-sex-age level like the previously mentioned variable of population, and can be classified because the age range and current year is known, since generation is classified from birth year. However, because the generation can overlap, this can create large spikes in interpreted suicides from where an age band starts and ends being classified as that generation. In addition, not everyone in an age band and year will be of one generation; it is stated as the average of the age band, so this variable is potentially problematic from the start, and will be excluded and not evaluated in the visualizations for the project.

Whilst it will not be excluded, another potentially problematic aspect is the differing numerical ranges of the age ranges; from the first categories 5-14 years, 15-24 years, and 25-34 years there is a 10 year

¹ [1] Kaggle. (2018). Suicide Rates Overview 1985 to 2016. Retrieved from <https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016>

² [2] United Nations Development Programme. (2018). Human Development Index (HDI). Retrieved from <http://hdr.undp.org/en/indicators/137506>

³ [3] World Bank. (2018). World Development Indicators: GDP (current US\$) by country: 1985 to 2016. Retrieved from <http://databank.worldbank.org/data/source/world-development-indicators#>

⁴ [4] World Health Organization. (2018). Suicide prevention. Retrieved from http://www.who.int/mental_health/suicide-prevention/en/

difference but the following ranges 35-54 years, 55-74 years are a 20 year difference, which can be misleading and appear as if there are more suicides occurring in these age ranges. The final category, 75+, is its own age band. The dataset does not state the oldest person who died by suicide so this age range is actually quite nebulous, and can misleadingly lead to the conclusions both ways; that the 75+ age band has less or more suicides compared to the other age bands.

A different amount of countries are a part of the dataset in each year. In data preprocessing 8 countries have been excluded from the analysis for having 3 or less years of data recorded total (countries appended in parentheses): Bosnia and Herzegovina (2), Cabo Verde (1), Dominica (1), Macau (1), Mongolia (1), Oman (3), Saint Kitts and Nevis (3) and San Marino (3). In 2016 only a few countries in comparison to the rest of the years recorded in the dataset have data, and the countries which do also often have data missing, and this year will therefore be excluded from the analysis. The differing amount of countries every year in the dataset should also be taken account when analysing the visualizations; the following preliminary visualization shows how the participation of countries varies per year:

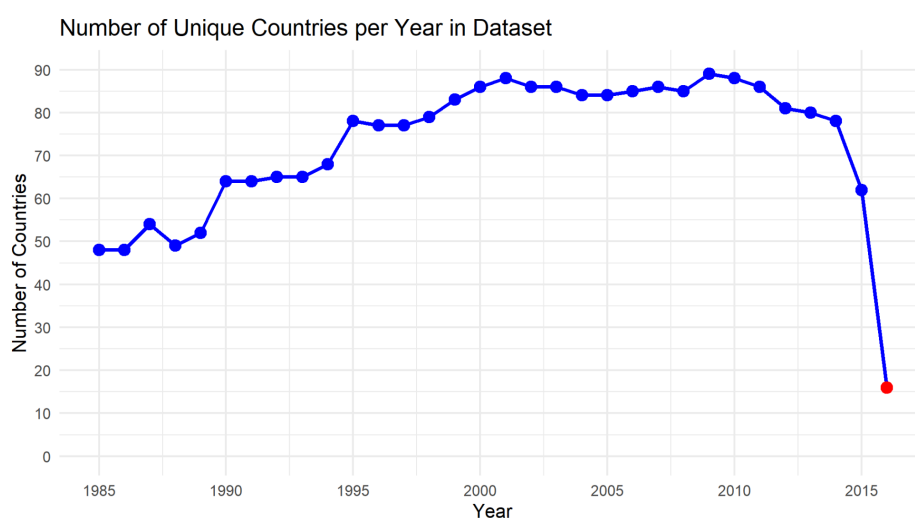


Figure 1: Number of unique countries that have records for each year in the dataset by Russel Yates

The participation of countries increases per year with some stability from the year 2000 and the sharp dip towards the excluded year 2016 is highlighted in red, where it is visually visible as significantly lacking in comparison to the previous years. Since this varying data might display in the visualizations as a misleading increase in suicide rate globally, this should be considered during the analysis.

To create choropleth maps in the project visualizing suicide rates regionally, the *countrycode* library is used. In preprocessing, the dataset had to be manipulated in two ways; firstly, the dataset spells a number of countries differently to the *countrycode* library, and the *countrycode* library classifies countries into the continents Africa, Europe, Asia, Oceania, and the Americas. In this visualization, the project would like to visualize suicide rate statistics in both North America and South America, due to the varying population vs size of country especially in North America, so two different classifications are maintained for each record so that visualizations can be made for both choropleth maps to be displayed correctly, and other plots can group North and South America separately. For the countries that were not excluded due to lacking data, the first step was mainly mapping the “and” in country names to “&” such as in Trinidad and Tobago; mapping Saint to St. as in Saint Lucia; the United Kingdom to UK and United States as USA; and finally the Czech Republic as Czechia.

After reclassifying countries into North and South America, the division of countries that have submitted data overall for each continent are of the following distribution: Africa (4), Asia (24), Europe (38), North America (3), South America (28), Oceania (4). The lower numbers in North America and Oceania are

justified due to the larger landmass and population of country in countries like Canada, America and Australia; while there are less countries in these continents, the land occupied is larger in comparison to, for example, Europe, Asia or South America which have more countries in the dataset. Africa, however, just has a low amount of countries participating in the dataset, and is therefore of low quality and should be considered when analysing regional trends.

6 Visualisation/Dashboard

Due to the records in the dataset being connected at the year-country-sex-age level in terms of number of suicides, smaller frames can be created from the dataset by sorting by year and country and summing up different variables. Summing up the number of suicides in each demographic record for the country and year can give the global total number of suicides each year; this can be done for each country, and this can be flipped with a demographic column to find the total number of males or females who died by suicide each year, or age groups, or population. A useful column in the dataset is also the suicide rate per 100k demographic population; this will be the main variable used in the visualizations in order to compare suicide rates regionally, since this normalizes the number of suicides in varying population sizes regionally.

This project will present the visualizations interactively in an RShiny dashboard, although in this report the visualizations are static. For the visualizations, it will also be a nice-to-have to have subtitles, interactive and informative labels in the dashboard and visual highlighting to maximize storytelling. In accordance to the focus questions introduced in the project objectives section, alongside the global suicide rates the dashboard will provide an exploration of the different relationship between the columns; to create a division of focus, the dashboard could be split into different tabs focusing on different areas of analyses—first of all, an overview, then a tab focusing on geographical and regional analyses, then demographics, and then socioeconomic. These tabs should be accessible via a sidebar. Finally, on the sidebar there should be a button to download the dashboard as a .pdf report. The downloadable report will include the static visualizations as well as descriptions from this document's later section, the story/results.

The project's requirements demand at least 3 different types of plots, in total 9 graphs, one of which is animated and one of which is AI-generated. The overview section of the dashboard can show some key insights or general trends in suicide rates; suicide rates trends by year globally or in each demographic can be depicted by a time series line chart. For demographics like age bands, gender or region groupings like continent this line chart can become a multi-line chart in order to visualize disparities between demographics and how this can evolve over time. A global average temporally, as well as between different demographics, can also be useful; this can be visualized as a bar chart or pie chart.

When moving onto the geographic focus, these can become grouped bar charts by continent. When moving onto countries this can become a large amount of bars, so the chart can be stacked horizontally, which can also visually depict any significant disparities. When the demographics become grouped across each other, such as age range in each continent, it can be useful to colour code the groups. This is the same case for bar charts of countries, which can be colour coded into continents. This project will use the *viridis* R library for accessibility; the colour palettes provided are both uniform so that values close to each other have similar values and further values are colourful, as well as being robust to colourblindness. For male and female, this will be a simple blue and red coding as these colours are typically associated with the genders as well as visually distinct if you are colourblind.

A nice-to-have for the dashboard feature may come in handy for bar charts involving countries; since a bar chart may include many values in order to preserve visual clarity the size should be maintained, so a

nice to have feature in the dashboard would be scrolling to see the whole chart, as well as a filter to show the top selected number of countries for the relevant chart. For both continent and country, choropleth maps will also be made to visualize any regional trends in suicide rates.

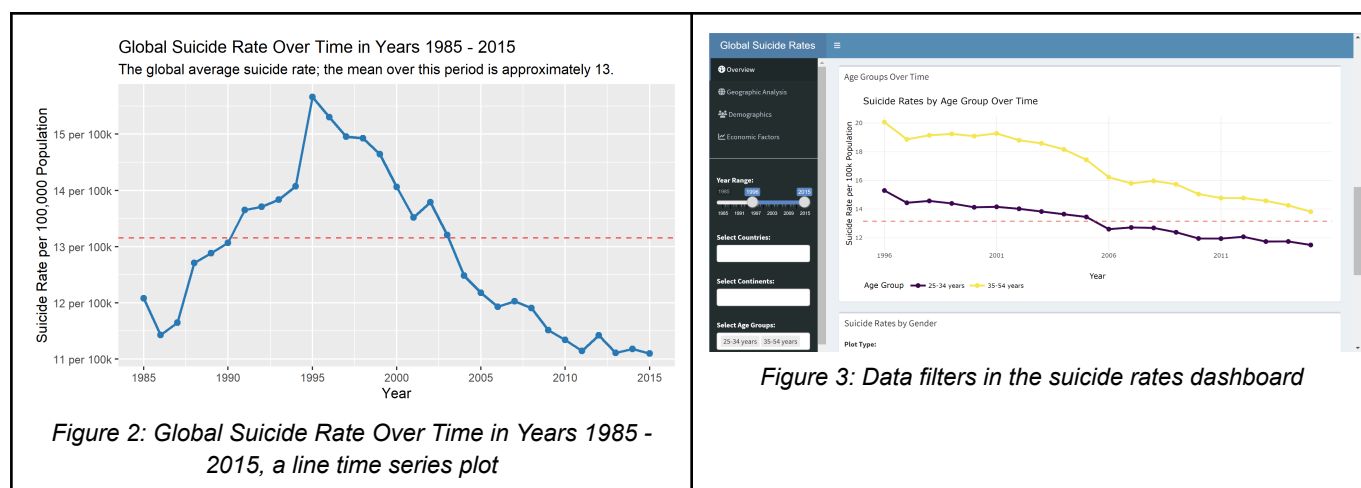
For the age analysis, in addition to the aforementioned plots in the overview, this project would also like to experiment with a box/violin plot for the age ranges and use AI in order to create this plot as well as animate the plot.

For analysing socioeconomic trends, a static and animated graph can be made to visualize the relationship between country and GDP. This can be in the form of a scatter plot, and the project would also like to experiment making it into a bubble plot by also visualizing the points in proportional size to the population of the country. The animation will play the progression of the suicide rate, GDP and the population growth each year. A nice to have in the dashboard will be a play/pause button that allows the viewer to let the plot play out to see the progression of the plot by the year as an animated plot, but also a slider for the year in order to see the status of the country at that particular year. The bubbles can also again be colour-grouped by continent. Here, the x-axis for GDP may need to have a log scale depending on how the different countries' GDPs vary.

Another nice-to-have for the dashboard would be filters on the dashboard sidebar to filter the data by each demographic group or region, and then sliders for year range, so the viewer can focus on specific variables or time periods as they like.

7 Story/Results

7.1 Overview



The dashboard can be found deployed at <https://deris22.shinyapps.io/DataVisualization/>⁵. This report is also downloadable as a manual to the dashboard

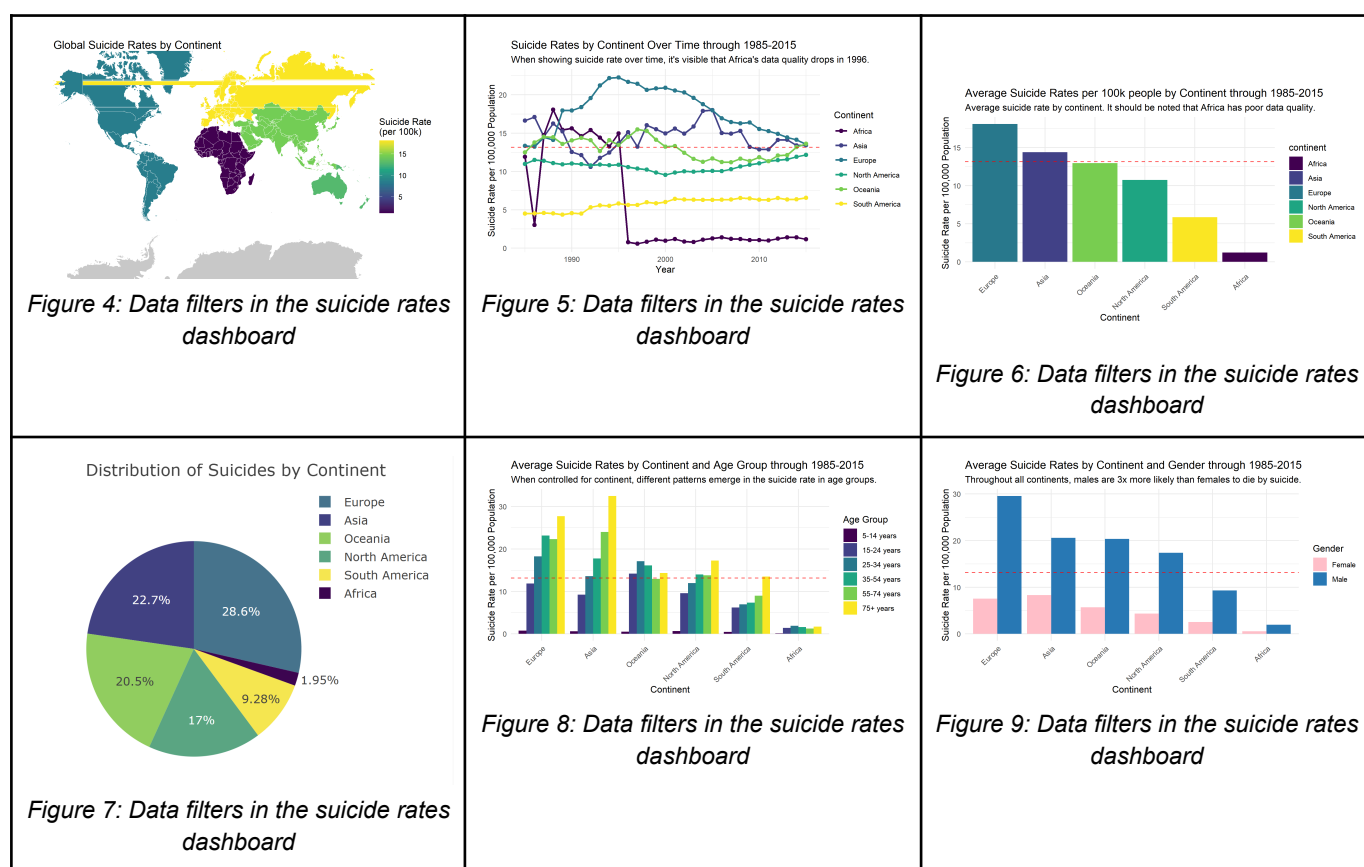
The above *Figure 2* and *Figure 3* show the first chart shown on the dashboard's overview and the look of the dashboard respectively. In the dashboard the viewer is able to filter the data in the plots by different nominal variables and year range; this changes the data so that only the data in the filtered range is plotted. This can give a greater insight without needing to necessarily create new plots, but is limited in that the axes' restructuring is automated. This can mean that the display of the data isn't necessarily the

⁵ [5] Data Visualization. (2024). Retrieved from <https://deris22.shinyapps.io/DataVisualization/>

best way to present it. This is explored later in the demographic analyses with genders as example filtered variables in *Figure 16* and *17*.

For *Figure 2* and other plots later shown, the red dashed line represents the global average suicide rate from 1985-2015, which is 13.2 deaths per 100k, per year. This is a helpful representative when looking at how the suicide rate across different demographics or regions might be at more risk of suicide (like in other bar charts explored in the dashboard), although in a time-series chart like *Figure 2*, a yearly average might be more helpful. The highest global suicide rate was in 1995 at 15.7 average deaths by suicide per 100k. This has been decreasing steadily and as of 2015 is down to 11 which is almost a 25% decrease. However, for both the rates of suicide in the 80s, 90s and 10s when referring to *Figure 1* these years have less country participation and may not be representative of the real global suicide rate; however, the decrease in the 10s was across a fairly stable participation of different countries and can be therefore considered as reliably observed.

7.2 Geographic analysis



For the geographic focus, first starting with continents, a choropleth map in *Figure 4*, bar chart in *Figure 6* and pie chart in *Figure 7* display different ways of visualizing the average suicide rate over the years 1985-2015 by region. Different colour mappings are used for each visualization, and this can be improved by using a consistent colour for each continent, and a different colour mapping than viridian to display different values like highest and lowest range of suicide rates. For the overview of the suicide rate in *Figure 6*, *Figure 7* and *Figure 4*, Europe is the highest rate over all. Since the data quality is similar to Oceania, South America and North America, this trend can be seen as fairly reliable. The extremely low suicide rate in Africa is due to the previously discussed poor data quality, and this can be seen in the time series chart of the suicide rate across continents over time in *Figure 5* where the participation of countries results in a significant and sharp drop in suicide rate. In addition, it can be seen that in *Figure 5* that Europe's rate of suicide has been decreasing since 1995. Despite having slightly

less data in these years, however, Oceania and Asia's rate of suicides display a slight upwards trend and increase in suicide rate.

Further differentiation with demographic age and sex grouping of continent suicide rate in *Figure 8* and *Figure 9* show that suicide rate increases with age for North America, South America, Asia and Europe. This is in line with the pattern of suicide rate increasing with the age range which is explored further later. Oceania and Africa show a different pattern and are actually higher for the 25-34 dataset. It should be noted that Oceania and Africa both have slightly and significantly less data than the other continents. European men are at the highest risk of suicide according to *Figure 9*, and this significant gender disparity is later explored more, as it appears that men are overrepresented in deaths by suicide in all continents.

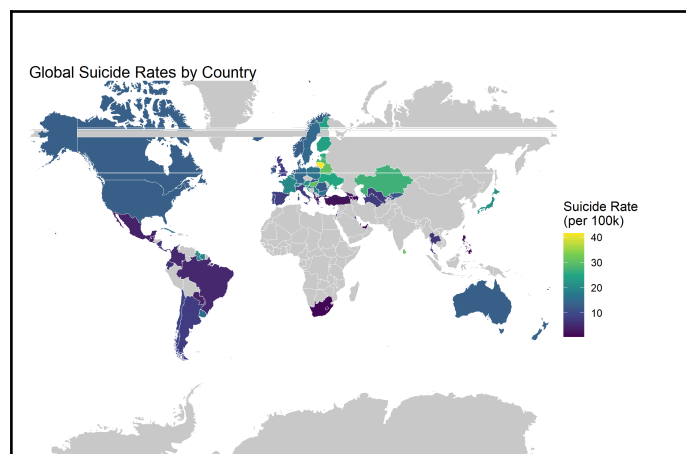


Figure 10: Data filters in the suicide rates dashboard

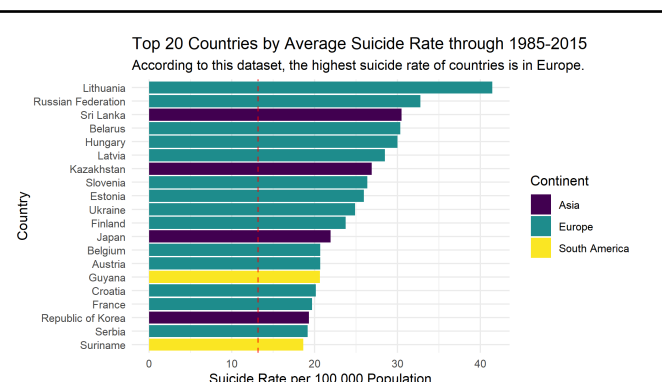


Figure 11: Data filters in the suicide rates dashboard

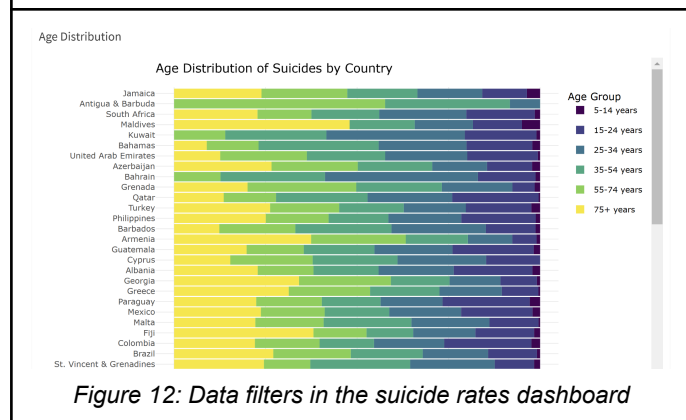


Figure 12: Data filters in the suicide rates dashboard

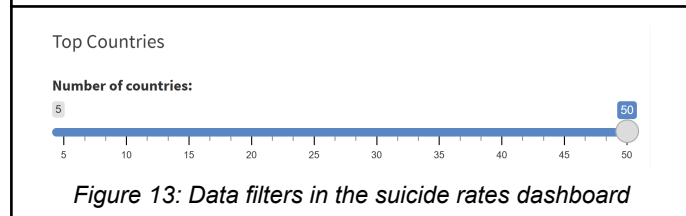


Figure 13: Data filters in the suicide rates dashboard

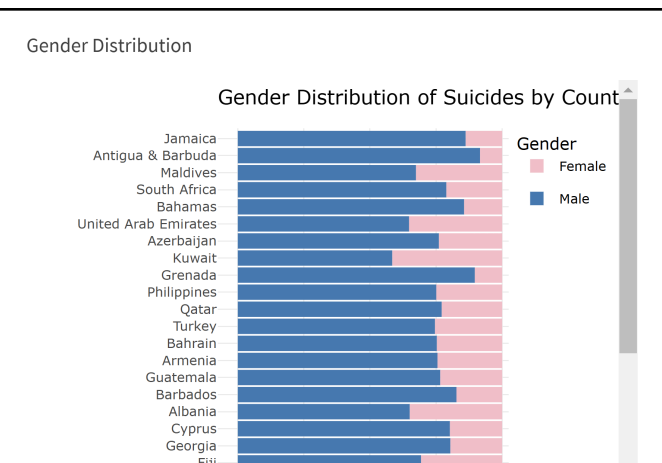


Figure 14: Data filters in the suicide rates dashboard

Figures 10 to 14 narrows the geographical analysis even more down to the country. The country choropleth map and chart of top 20 countries again reaffirms that Europe has the highest average suicide rate per 100k per year. *Figure 14* unfortunately does not show all countries due to its size, but in the dashboard, where it is scrollable, shows that the overrepresentation of men in death by suicide is universal and this phenomena is exhibited in all countries. *Figure 13* shows a filter that can be applied to *Figure 11* to show the top countries in order to vary the graph, and the country plots are scrollable, but this unfortunately limits the field of view of the axis and could be improved. Furthermore, this top

countries filter is currently not applied to every graph. *Figure 12* which shows the age distribution of average suicides in each country can also benefit from being sorted highest to lowest, where a trend here in ages could potentially be more easily seen.

7.3 Demographic analysis

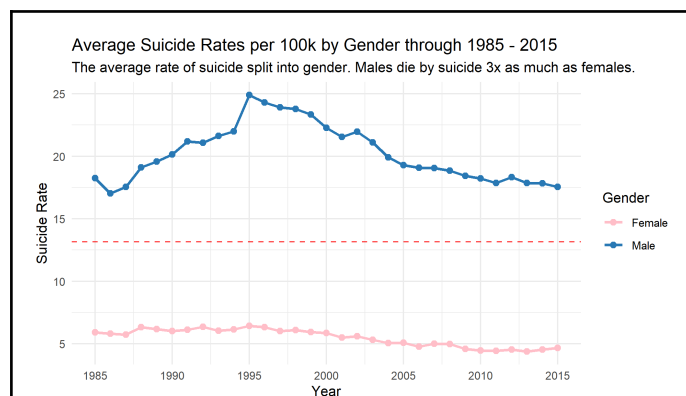


Figure 15: Data filters in the suicide rates dashboard

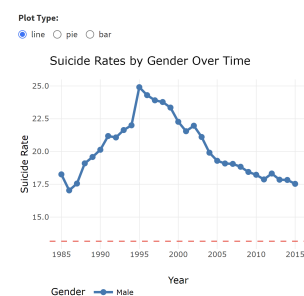


Figure 16: Data filters in the suicide rates dashboard

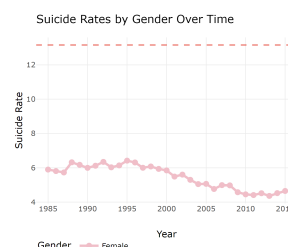


Figure 17: Data filters in the suicide rates dashboard

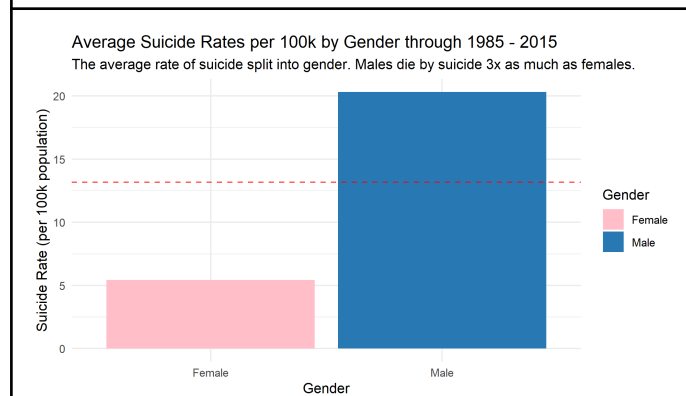


Figure 18: Data filters in the suicide rates dashboard

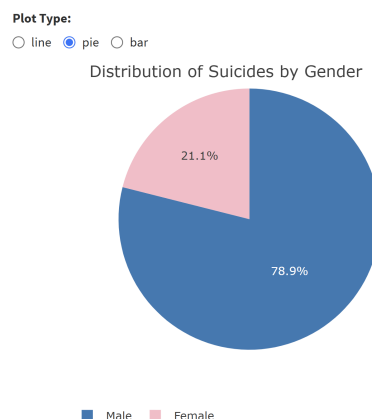


Figure 19: Data filters in the suicide rates dashboard

As seen in the continental and country-level breakdown, men are overrepresented universally in being at risk for death by suicide. The average visualized in the bar chart in *Figure 18* and *Figure 19* show that men are 3.5x as likely than women to die by suicide, and as in the continental bar chart *Figure 9* this can be as high as 4 times more likely as seen in the European male average. *Figure 15*, which is a line chart time series plot of the annual global rate of suicide split into genders show that this disparity remains throughout the years; *Figure 16* and *Figure 17*, which are generated in the dashboard from *Figure 15* by filtering for the male and female gender variable respectively, shows just how big this rift is with the global average per year is significantly lower for the annual female rate and how significantly higher the annual male rate is than the average. Additionally, both breakdowns do show that the rate is slowly decreasing for both genders per year.

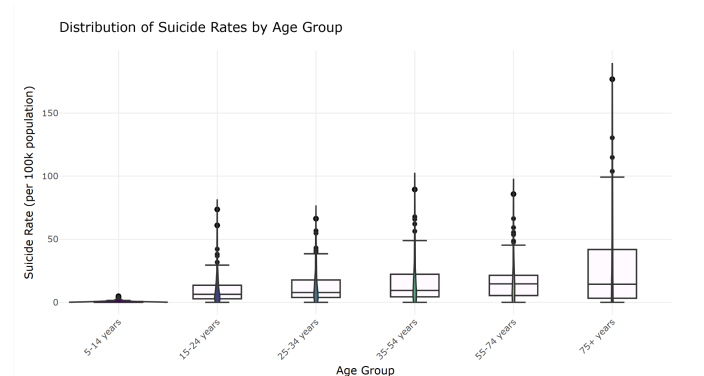


Figure 20: Data filters in the suicide rates dashboard

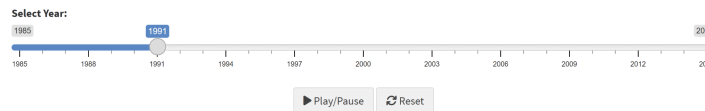


Figure 21: Data filters in the suicide rates dashboard

The AI generated graph experiment is in *Figure 20* and was generated with the help of Gemini 1.5, which was prompted to create an animated violin plot that depicts the change in rate of suicide in different age groups. The question intended to be answered was the focus in 2a about how the demographic of age changes throughout time in being a risk factor. A lot can be improved with this visualization, possibly with some data preprocessing as it seems like data outliers may skew the graph's entire axis and suppress the rest of the graph, and this hard-to-see change is also present when the animation is being played. Additionally, a violin plot may not have been the right visualization to show this information, though the box plot shows the previously mentioned patterns of higher age groups being at more risk of suicide.

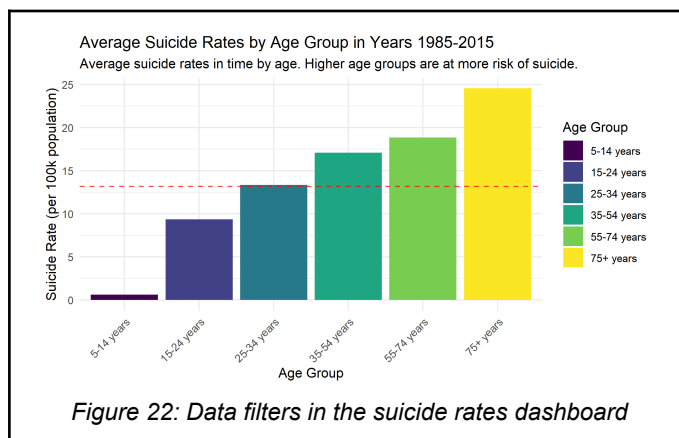


Figure 22: Data filters in the suicide rates dashboard

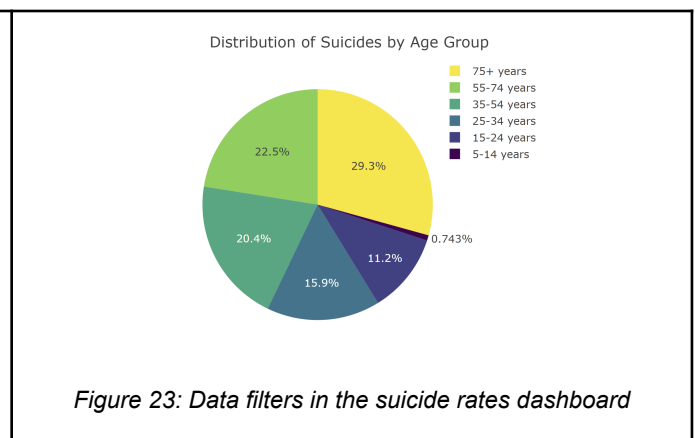


Figure 23: Data filters in the suicide rates dashboard

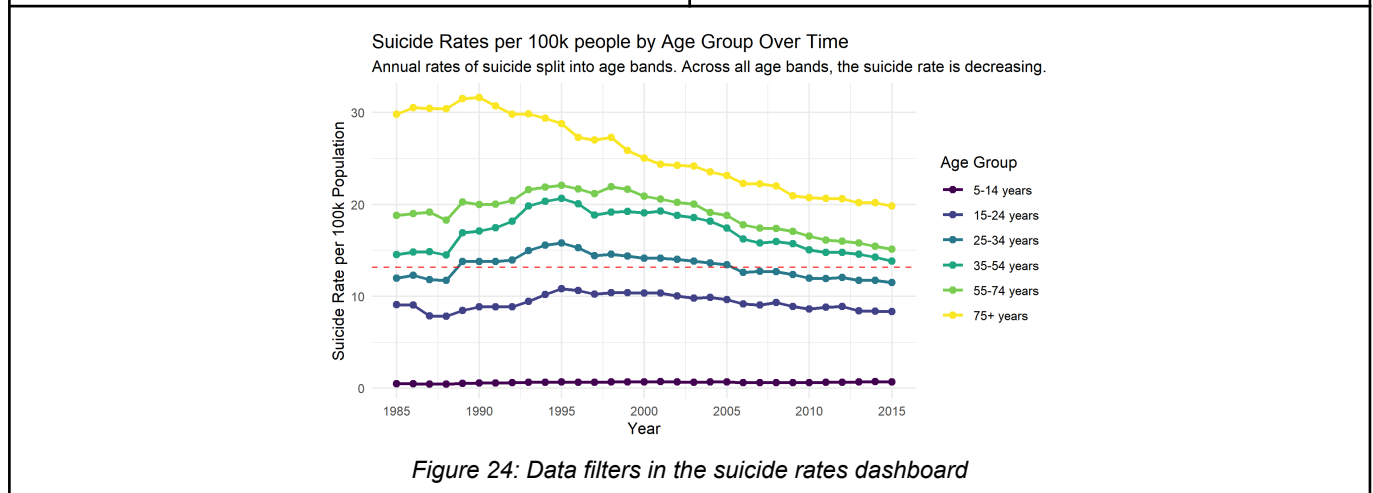


Figure 24: Data filters in the suicide rates dashboard

Figures 22, 23, and 24 is a more in-depth exploration of the trend of suicide rates in individual age groups, and as earlier discussed and observed in the regional patterns, higher rates of suicide can be seen in older age bands. However, as also discussed in data preprocessing, this can be debatable because of the age bands above 35 being 20 years of difference as opposed to 10 years of difference in the age bands 34 and below. The highest risk of suicide at 75+ years is also debatable as this range is nebulous and does not have an end limit. According to this dataset, globally, the likelihood of suicide increases with age according to this dataset, but this would need to be verified with more uniform age bands. The time series line chart in Figure 24 does show that the suicide rate for every age band above 15 is decreasing, and below this remains a negligibly continuously changing small number each year.

7.4 Economic analysis

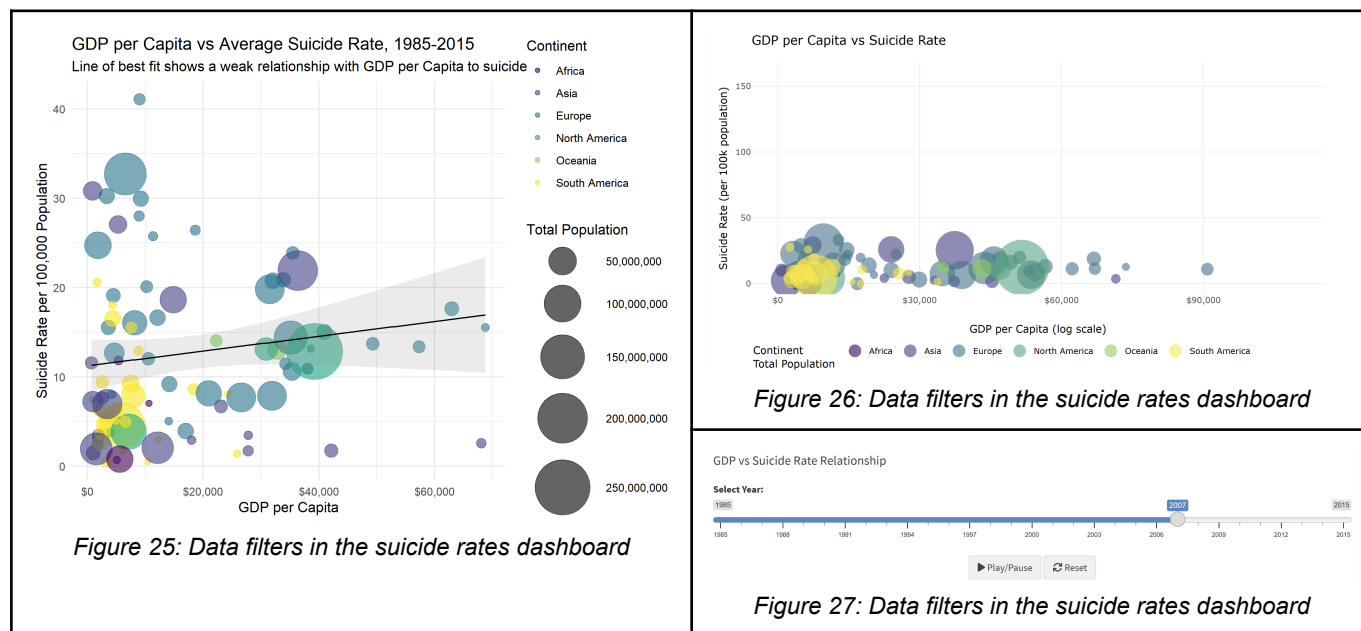


Figure 25 explores the relationship between GDP capita and suicide rate per 100k on an annual average. The bubbles are grouped in a colour map according to the continent, and the bubbles are also proportional in size to the country's population. This way, many relationships and potential clusters can be observed; the relation of suicide rate to population, relation of suicide rate to GDP per capita, and any regional clusters that may show patterns in GDP or suicide rate. In the interactive dashboard, the tooltip shows which country the bubble is. A line of best fit has been plotted on the graph and there appears to be a weak but positive correlation between GDP per capita in a country and increased suicide rate. This can however be skewed by outliers and for a more detailed analysis needs to be verified further with other statistical values; these outliers can be observed in the animated version of this chart in Figure 26.

Figure 26 is an animated version of Figure 25 and on the dashboard just like the animated violin plot Figure 20 also comes with a feature to play and pause the animated graph so that its status in a year can be analysed but also played to see progression. Figure 26 as opposed to its static and averaged across annual values counterpart has a log scaled x axis due to the significant economic development some of the countries with highest GDP values go through throughout the years. Its tooltip showing country name allows the viewer to also track how a suicide rate increases and decreases with GDP rate and there is a different trend for every country. Much of the increase in suicide rate due to development, however, has similar trends to just the population and time increasing and it is inconclusive just observing this graph if there are any strong trends in the increase or decrease of suicide rate.

8 Conclusion/Discussion

This project aimed to analyse trends in global suicide rates across different demographics of age and gender, find potential regional clusters and patterns, analyse socioeconomic impact on suicide rate, all throughout time provided by the limit of the dataset from 1985 to 2015. The visualizations have found the the global suicide rate has shown a decline since its peak in 1995, although its previous levels are not possible to know from this dataset due to a decreased number of countries with records in the dataset in the years before, and the decline should also be interpreted critically due to a smaller rate of participation after the 2010s that may make a significant difference if the data existed.

Europe has the highest suicide rates amongst all continents but has been decreasing steadily since 1995; however, Oceania and Asia has been trending slightly upwards in suicide rates. It is difficult to analyse the suicide rates in Africa due to poor data quality and only 3 countries being included in the dataset. Furthermore, Europe, North America, South America and Asia make up a majority of the database's records, which may mean that there is a regional bias. For regional data, the validity of these observed trends can be improved by analysing information outside of the dataset that can fill in the missing values of countries not in the dataset.

There is a universal significant gender disparity; men are on average 3x more likely to die by suicide than women, and in Europe this is 4x more likely. For age, there appears to be a trend of suicide rate increasing with age, although because of the age band intervals being uneven, with ages above 35 being double in age bands than the age bands below age 54, this should be interpreted critically.

The static and animated plot showed that there is a weak positive correlation between GDP per capita and increased suicide rates, but that this ultimately varies significantly between countries and overtime, and due to gaps in data and undetected outliers, should be interpreted critically. This can be improved again by validating by enhancing the dataset with missing data of countries not present, and further preprocessing of the data to exclude outliers.

For the visualizations, implementing a more consistent colour map, as well as differentiating between other categorical variables, can improve intuitive understanding of the dashboard. For example, the continents can be one colour map, and the age bands can be another, but Europe or the age band 75+ should remain the same across visualizations. The interactive visualizations can be improved by having better and more detailed tooltips instead of its current default labels, like from the GDP bubble plot which shows the country. The plots in the country demographic focus can be significantly improved by a more complete implementation of dynamic axis scaling and scrolling for large charts that include a lot of countries. When filtering, the plots can also be improved by being edited to fit their axis more.

The visualizations can benefit from more cohesive storytelling; although some subtitles were included in the charts and highlighting the mean global average to serve as a comparison for a lot of the data charts were included, this could be improved; for example, the average may have been more useful if it was an annual average instead of a flat average line for time-series charts. In general, more graphs could have used subtitles as well as subtitles that were more informative of the chart. These subtitles were also not visible in the interactive dashboard, though it is visible in the the report's version of the static graphs

Some visualizations may not have been right to visualize the data and answer the questions it wanted to, even though these questions were answered by other visualizations; for example, the violin plot maybe have been better off a dumbbell plot or lollipop plot to show if there are positive or negative trends over time for a regional focus, and may even be applicable to show the disparity or change in disparity in demographics like gender over time.

9 References

- [1] Kaggle. (2018). Suicide Rates Overview 1985 to 2016. Retrieved from <https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016>
- [2] United Nations Development Programme. (2018). Human Development Index (HDI). Retrieved from <http://hdr.undp.org/en/indicators/137506>
- [3] World Bank. (2018). World Development Indicators: GDP (current US\$) by country: 1985 to 2016. Retrieved from <http://databank.worldbank.org/data/source/world-development-indicators#>
- [4] World Health Organization. (2018). Suicide prevention. Retrieved from http://www.who.int/mental_health/suicide-prevention/en/
- [5] Data Visualization. (2024). Retrieved from <https://deris22.shinyapps.io/DataVisualization/>