**Global Suicide Rate Analysis using Hive and PySpark**

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

**Bachelor of Technology/Master of Technology**

In

**Computer Science and Engineering**

**School of Engineering and Sciences**

Submitted by

Candidate Name

**Rohith Kumar AP21110011423**

**Lakshmipathi AP21110011457**

**Tony Babu AP21110011488**

****

Under the Guidance of

Sriramulu Bojjagani

**SRM University–AP**

**Neerukonda, Mangalagiri, Guntur**

**Andhra Pradesh – 522 240**

**[November, 2024]**

**Table Contents**

[1 Introduction 4](#_Toc181531701)

[2 About Dataset 5](#_Toc181531702)

[3 Methodology 6](#_Toc181531703)

[1. Creating database and table in hive 6](#_Toc181531704)

[2. Preprocessing of data in table 7](#_Toc181531705)

[3. Analysis of data 8](#_Toc181531706)

[4. Visualization of data 19](#_Toc181531707)

[4 Conclusion 31](#_Toc181531708)

[5 References 32](#_Toc181531709)

**Abstract**

The Global Suicide Rate Analysis project focuses on understanding suicide trends worldwide from 1985 to 2016, a period marked by significant socio-economic changes in many countries. During these years, global suicide rates exhibited varied patterns across regions, with some countries experiencing notable increases and others observing declines. Economic shifts, such as recessions, rapid industrialization, and transitions in political structures, often had a profound impact on national suicide rates. Additionally, mental health awareness, healthcare infrastructure, and support systems for vulnerable populations evolved differently across nations, leading to disparities in suicide prevention effectiveness.

By analyzing data from this 30-year span, the project aims to identify the impact of factors like economic conditions, human development index (HDI), and gross domestic product (GDP) on suicide rates, offering a nuanced perspective on the socioeconomic influences linked to mental health challenges worldwide.

This project aims to explore suicide patterns worldwide by examining various socio-economic indicators. Using a dataset containing country-wise data on suicides, human development index (HDI), and GDP per capita, the project aggregates data in a Hive table within Ambari to derive meaningful insights.

Analyzing these data points allows for an in-depth investigation of possible correlations between the economic health of a nation and its suicide rates. For instance, examining averages of HDI and GDP per capita alongside suicide rates provides insight into how economic hardship, quality of life, and other social factors might influence mental health outcomes at a national level. By comparing these rates across different years and countries, the project seeks to uncover trends that can guide policy recommendations and interventions.

# Introduction

The Global Suicide Rate Analysis project is designed to investigate the patterns and factors associated with suicide rates around the world, covering data from 1985 to 2016. This period includes diverse economic, political, and social transformations that have shaped national policies, public health strategies, and quality of life. By studying this data over three decades, the project aims to explore how various socioeconomic indicators—such as GDP per capita, human development index (HDI), and suicide rates—interrelate and influence public health outcomes. This analysis will use Hive and Ambari for data processing, enabling efficient handling and querying of extensive datasets for reliable aggregation and insights.

The core of this project lies in examining how economic stability, quality of life, and social conditions correlate with suicide rates at the country and yearly levels. By creating an aggregated Hive table, the analysis consolidates suicide counts, average suicide rates, HDI, and GDP per capita, providing a structured view of these critical factors across countries and years. This allows for a systematic investigation of trends and anomalies, including the role of economic hardship, access to resources, and social support systems in mental health outcomes.

By grouping data by country and year, the project can uncover patterns within regions or time periods that may not be evident in isolated data points, offering a richer understanding of the complex factors that contribute to suicide rates.

This analysis aims to use insights from the data to help shape public health strategies and policies that can address mental health challenges across societies. By better understanding how economic factors, like income levels and development, relate to suicide rates, policymakers and organizations can identify areas where more support could help reduce suicide risks. For example, countries with high suicide rates and lower economic conditions might benefit from programs that improve mental health resources, social services, and economic aid. On the other hand, the analysis may also show cases where a high HDI or GDP does not necessarily mean lower suicide rates, suggesting the need to explore other social or cultural factors that could impact mental health.

# About Dataset

The Global Suicide Rate dataset, sourced from Kaggle, provides a detailed snapshot of suicide statistics across diverse countries and demographics, covering 27,820 entries with data from 1987 to 2016. This dataset includes 12 columns with key demographic attributes like age, sex, and generation, allowing insights into suicide rates across various population segments and time frames.

Each row specifies a unique demographic group within a country for a particular year, enabling year-by-year trend analysis. Age groups are split into six ranges, from 5-14 to 75+ years, revealing how suicide rates might differ across life stages. Gender-specific data helps to examine variations in suicide rates between males and females, while generational categories, including Millennials, Generation Z, and Boomers, provide insights into how societal shifts and experiences impact mental health across generations.

Quantitative measures include suicides\_no and population, along with suicides/100k pop, which standardizes data, allowing comparisons across population sizes. Socioeconomic factors are represented by HDI, GDP for the year, and GDP per capita, supporting analysis of economic conditions alongside suicide rates. Although the HDI column has some missing values, it provides a valuable index of social and economic development.

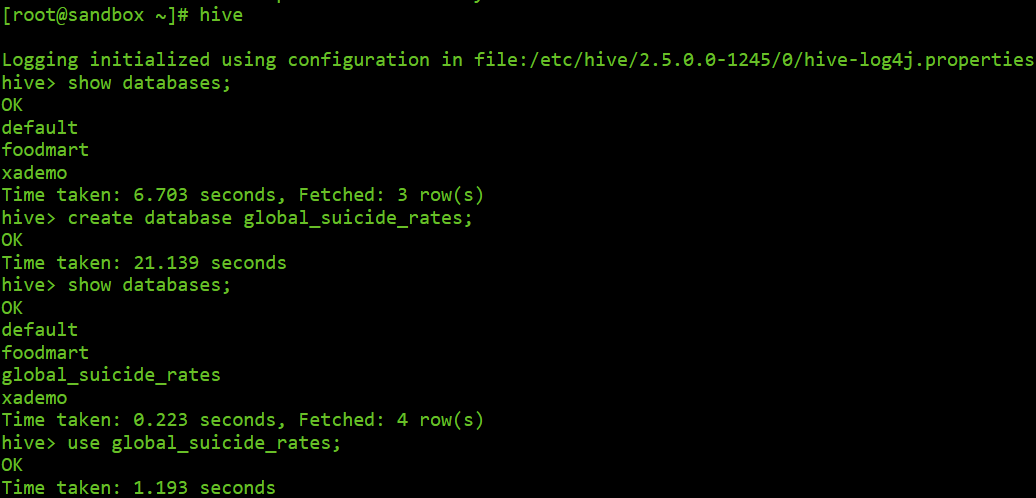
The dataset includes data from 101 countries, spanning both high-income nations like the United States and Australia, as well as lower-income countries like Cabo Verde. This diversity enriches the dataset, allowing researchers to explore potential links between economic status and suicide rates across varied global contexts. With nearly three decades of data, the dataset offers a foundation for investigating long-term trends, identifying shifts tied to economic changes, policy shifts, and public health initiatives.

Ultimately, this dataset is a powerful tool for understanding how demographic, economic, and social factors converge to shape suicide trends worldwide, guiding informed interventions and policies.

# Methodology

## Creating database and table in hive

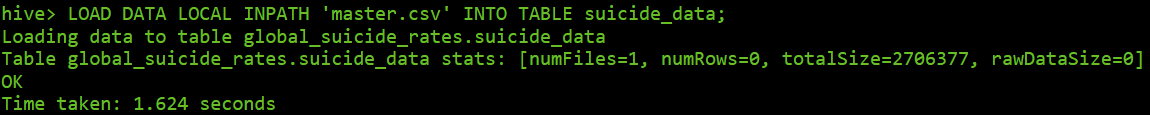
We created a **database name global\_suicide\_rates**

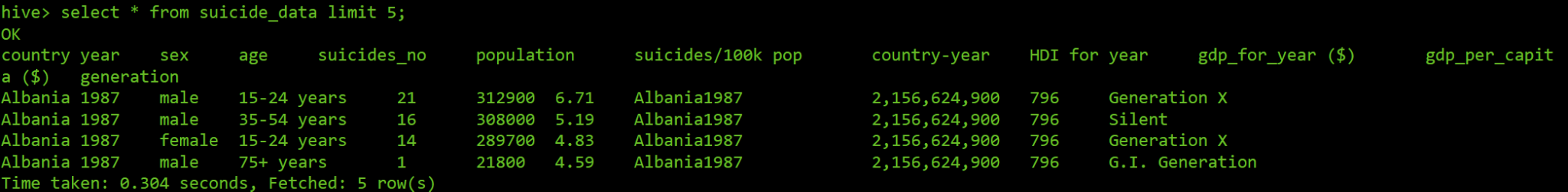


After that we created a **table name suicide\_data**



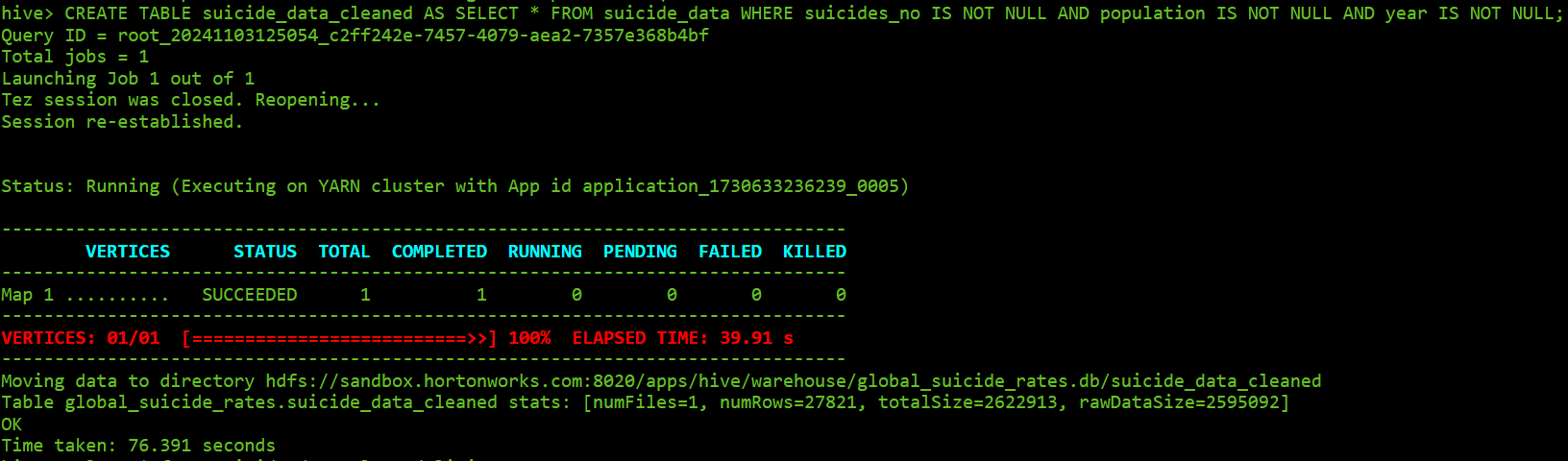
After that **loading data into table from hdfs**



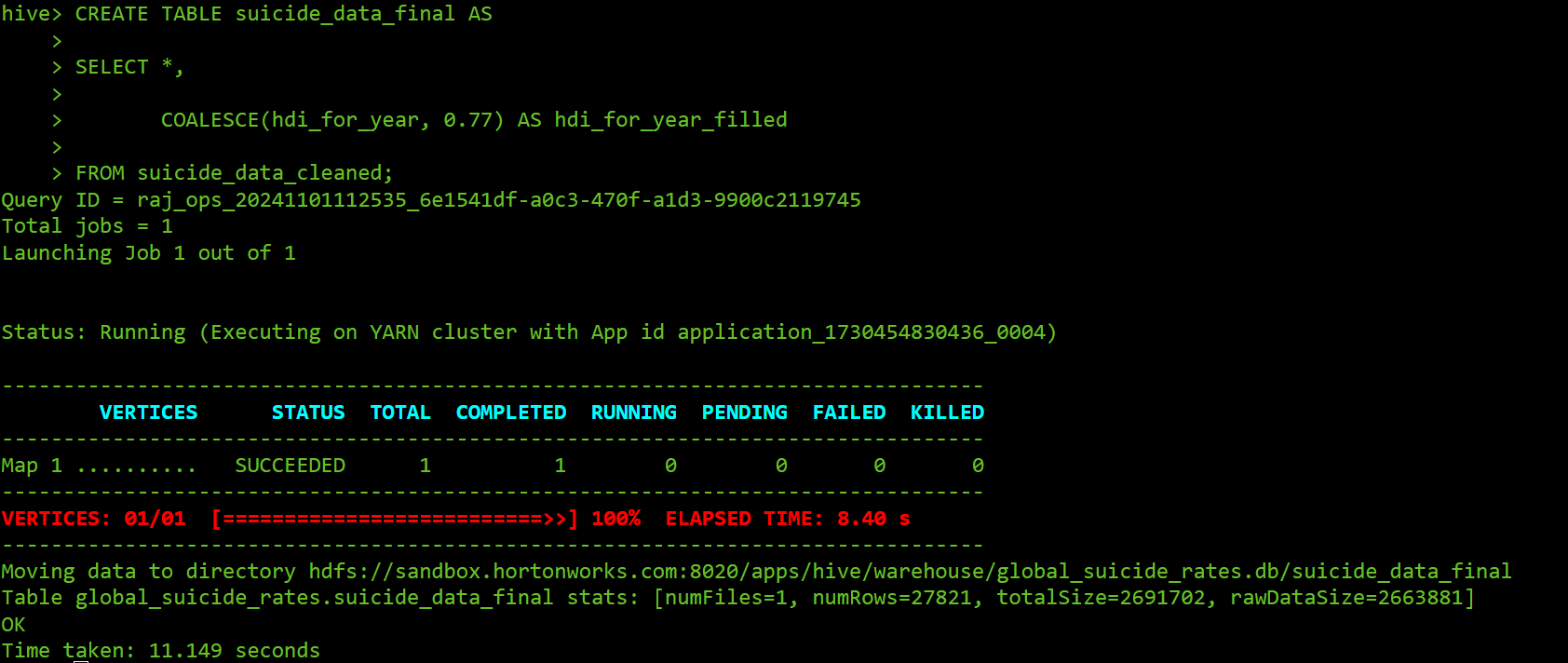


## Preprocessing of data in table

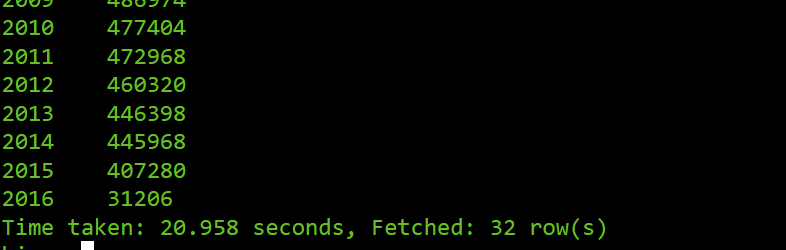
Creating a **new table sucide\_data\_cleaned** with no null values in it.

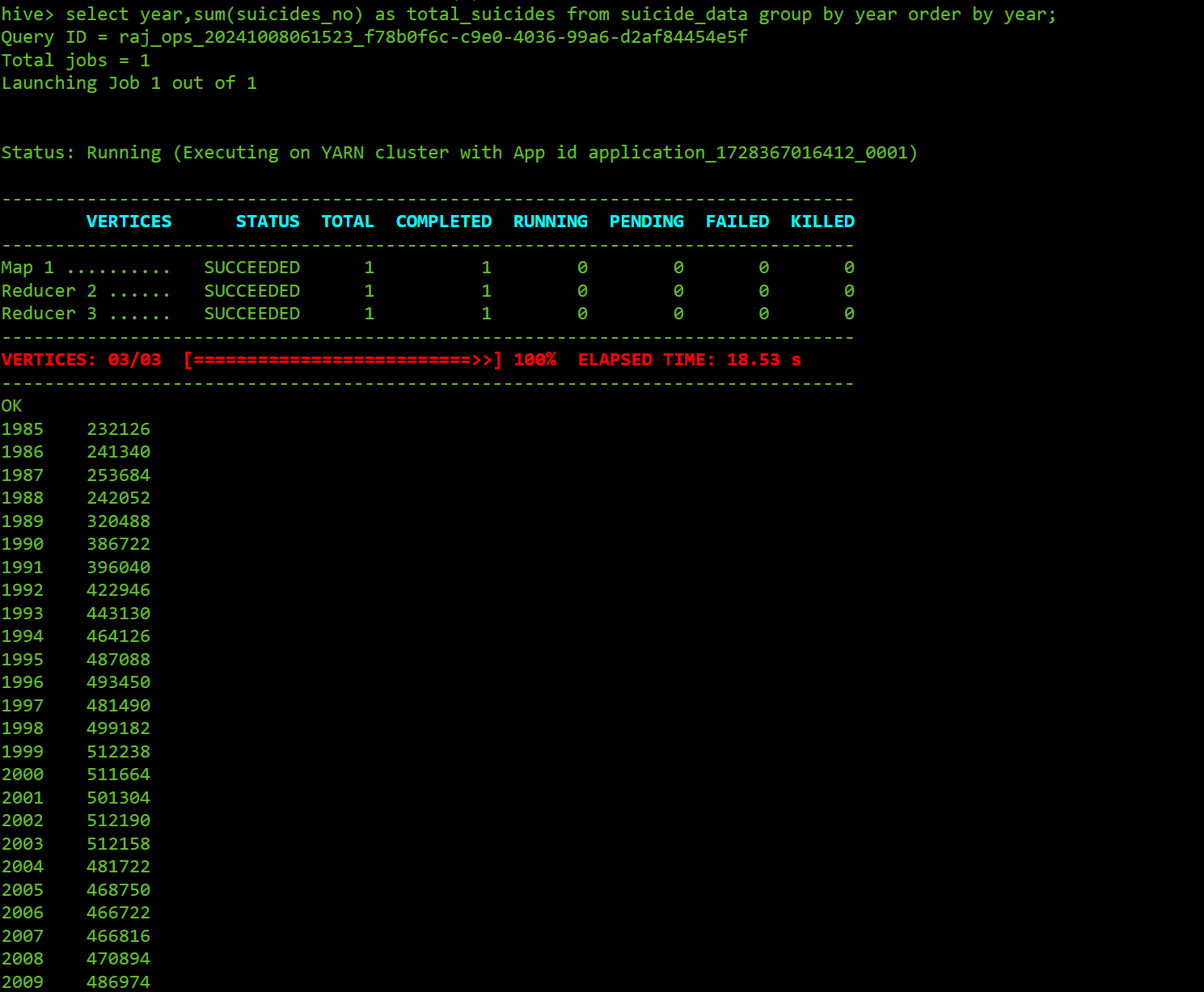
****

**Replacing missing values with average** of hdi which is 0.77

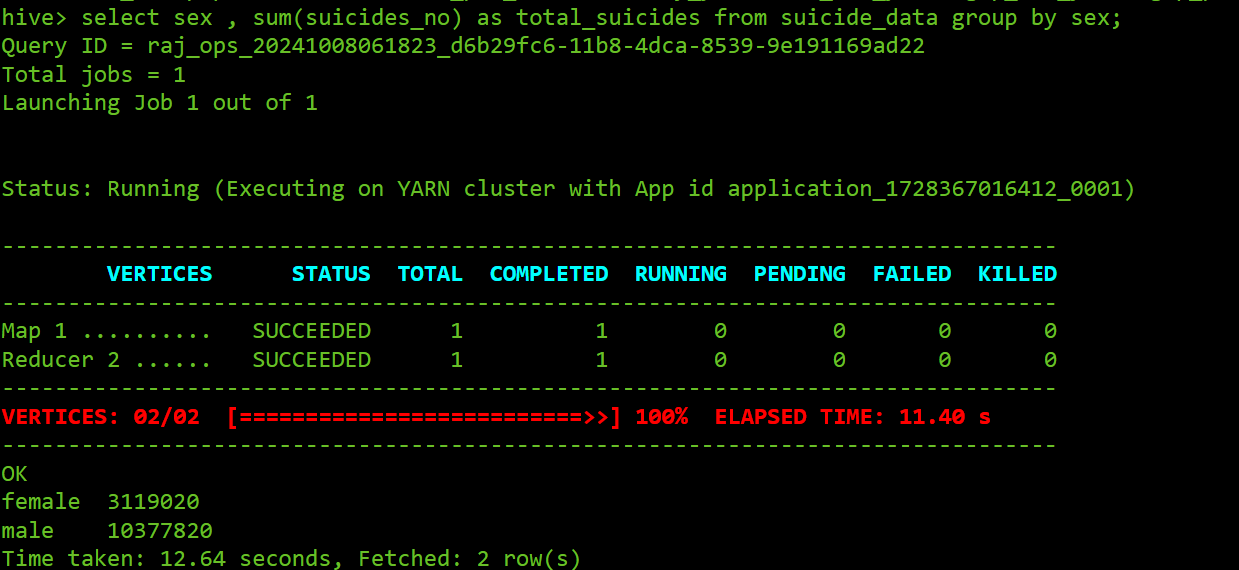


## Analysis of data

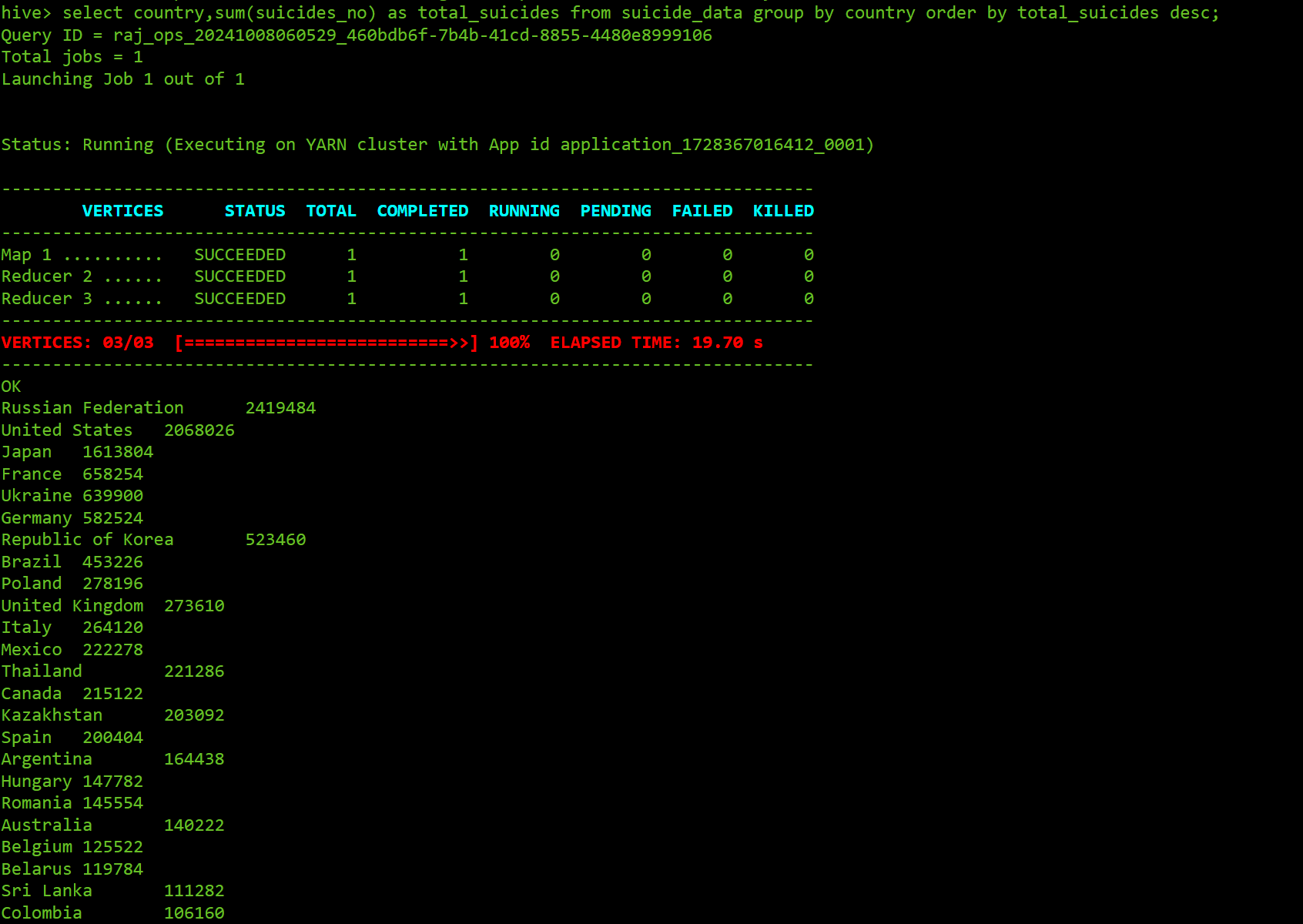
****Year v/s suicide count:

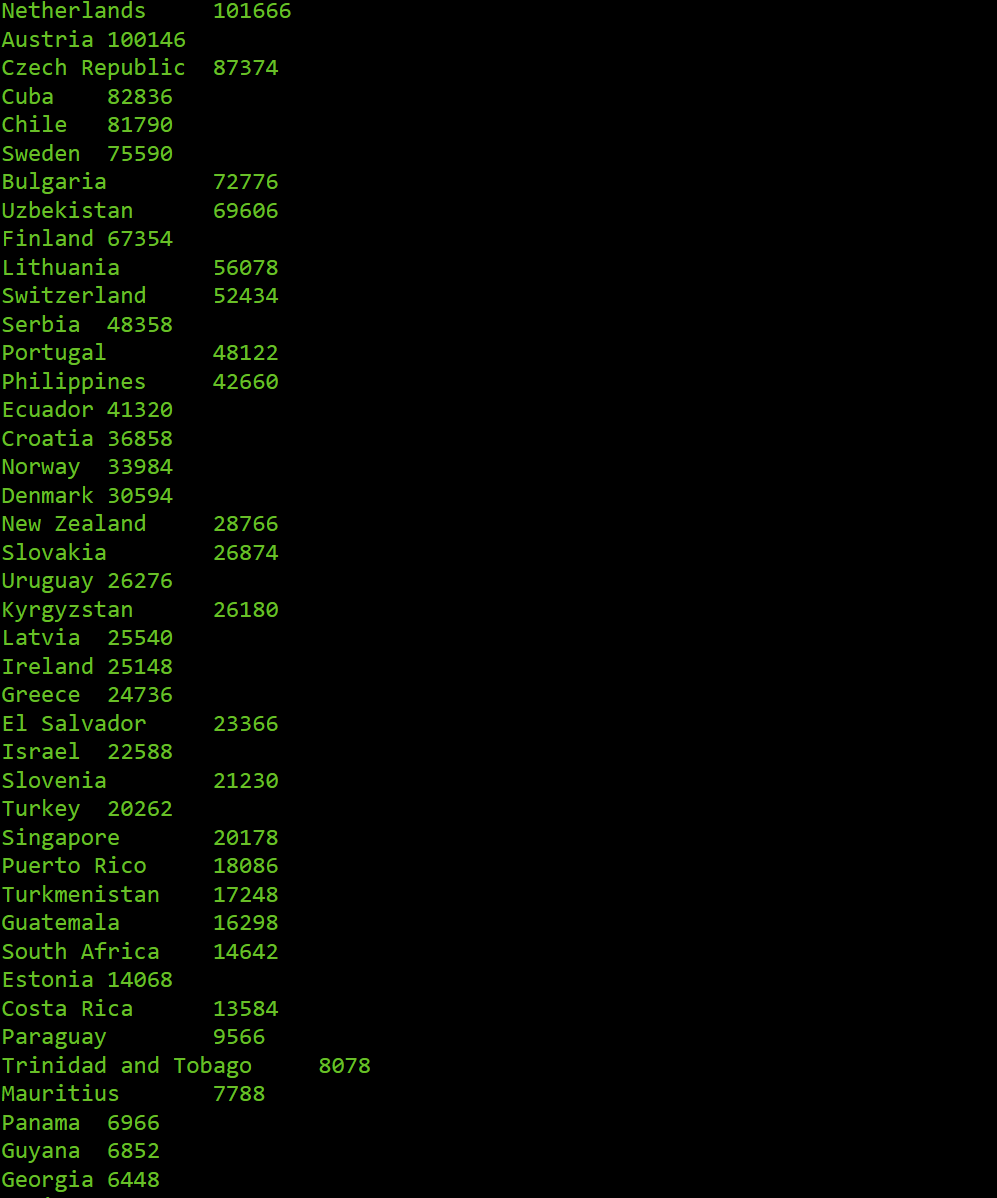
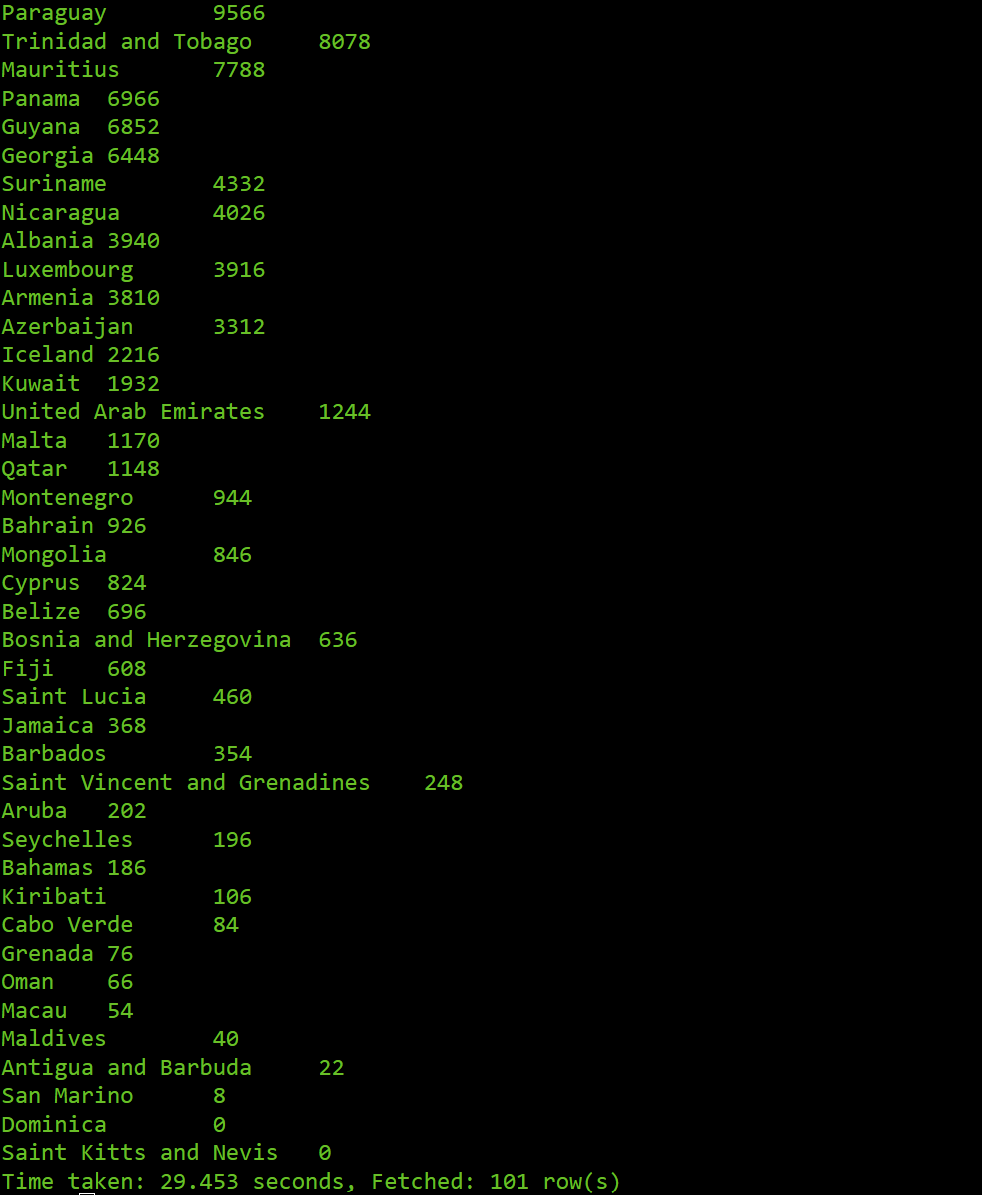


This image shows the output of a Hive query executed in a Hadoop environment. The query retrieves the yearly total number of suicides by summing up entries in a column labeled "suicides\_no" from a dataset called "suicide\_data." The query groups the data by year and sorts it in ascending order. The results display the total suicides for each year from 1985 to 2009, showing a generally increasing trend over the years. From the above image we can see that **Highest suicide count is in 1999 with 512238**. And **Least suicide count is in 2016 with 31206.**

Gender v/s suicide count:

This image displays the total number of suicides based on gender from a dataset called "suicide\_data." The query groups the data by the "sex" field and sums the values in the "suicides\_no" column for each gender. It shows a total of 3,119,020 suicides for females and 10,377,820 suicides for males. This indicates a significant **gender disparity** in suicide rates within the dataset.

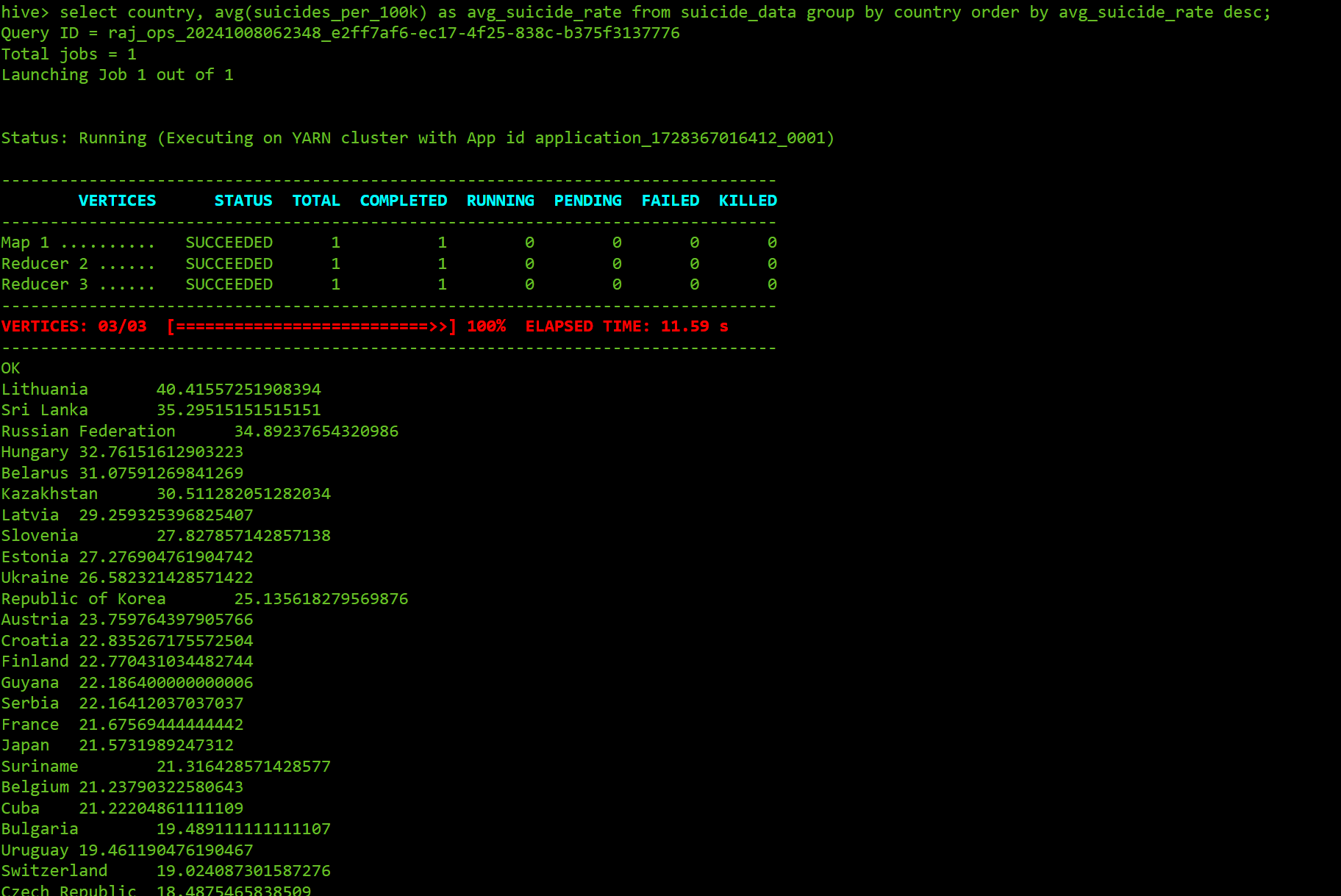
Country v/s suicide count:

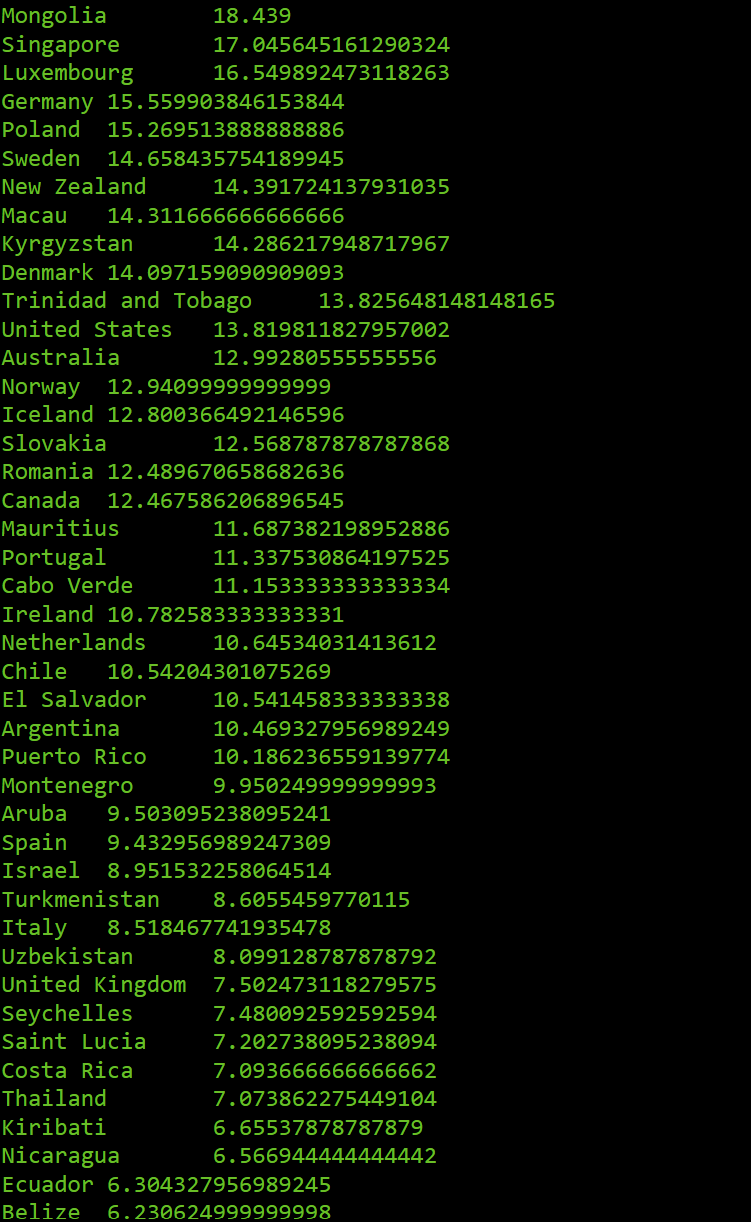
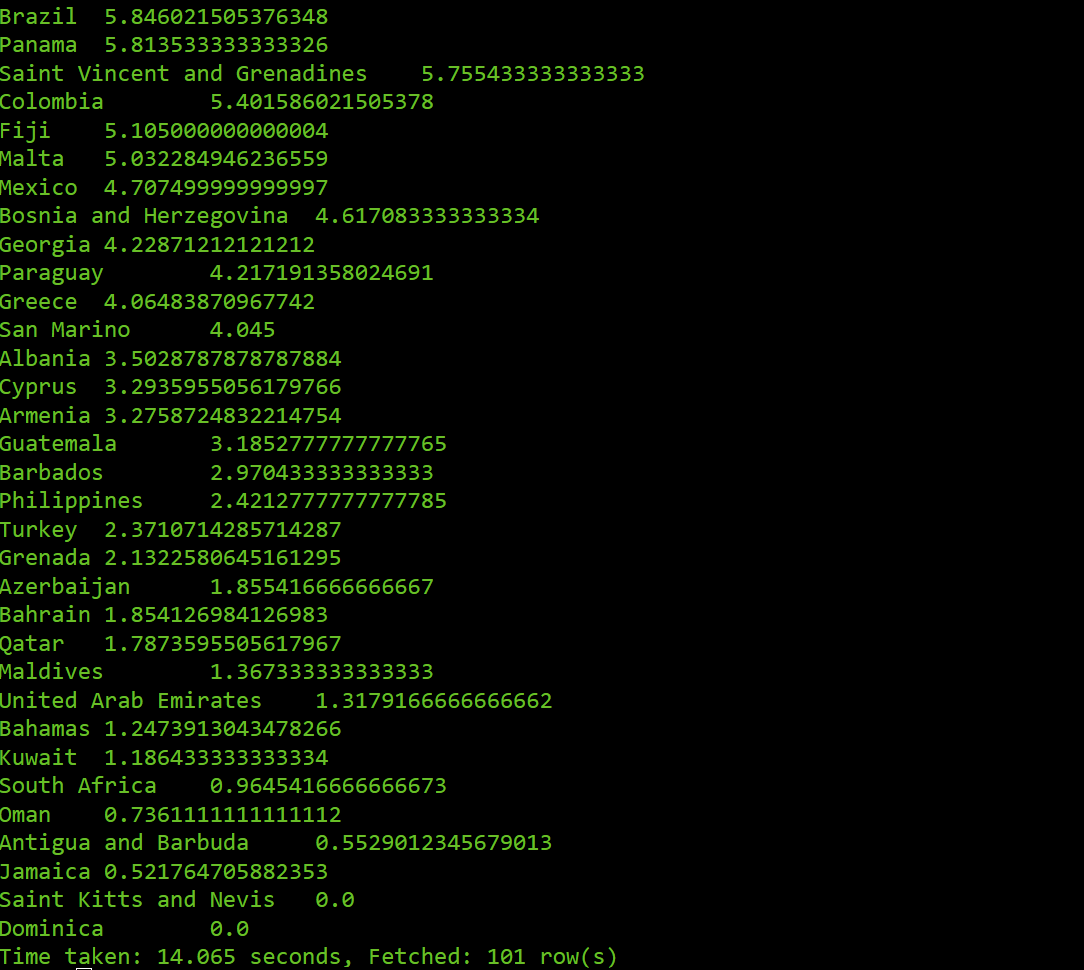


This image shows suicide statistics by country. The query calculates the total number of suicides for each country by summing up values in the "suicides\_no" column from the "suicide\_data" table, grouped by the "country" field and sorted in descending order of total suicides.

The output displays the total suicide counts per country, with the **Russian Federation at the top** (2,419,484 suicides) followed by **the United States** (2,068,026 suicides) and **Japan** (1,613,804 suicides). The countries with **least suicide rates in Dominica and Saints Kitts & Nevis** with 0 suicides.

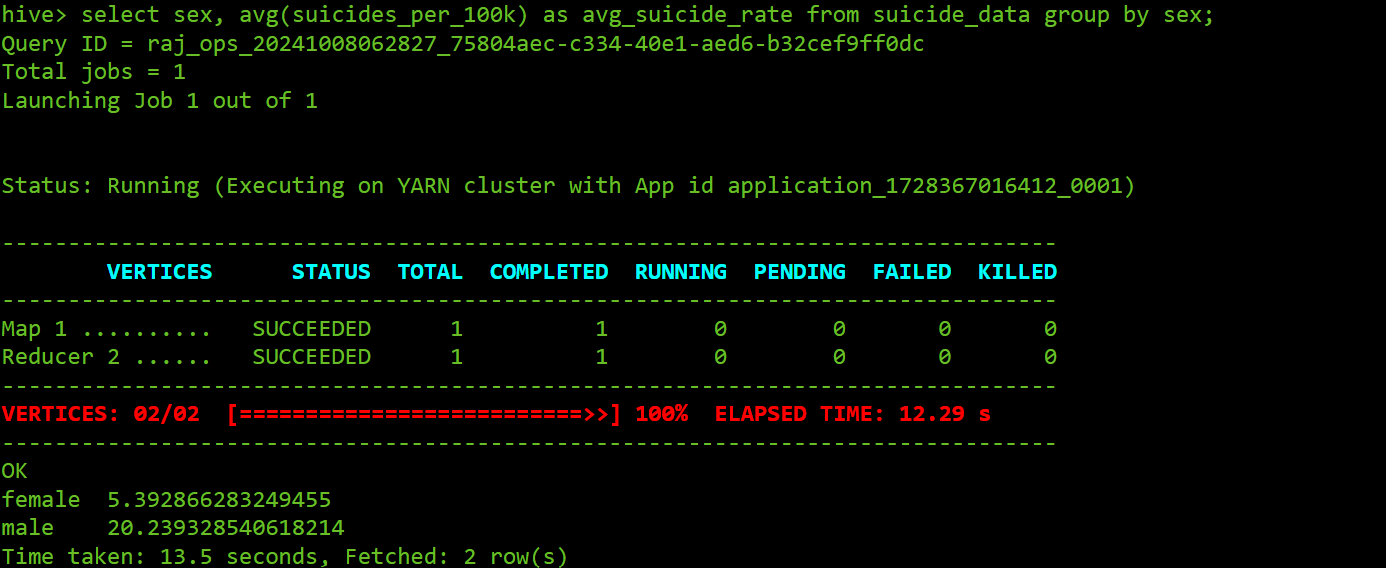
Country v/s suicides rate per 100k:





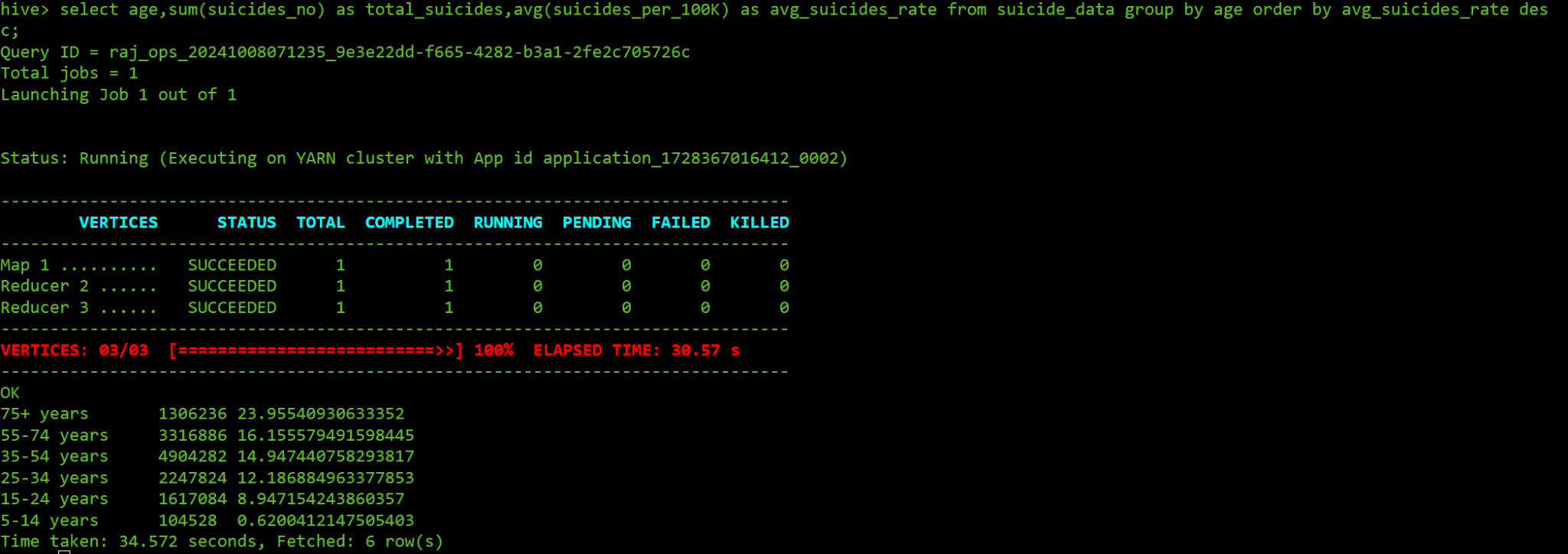
The image you provided shows the average suicide rate per 100,000 people for each country in the dataset. The query groups data by country and orders the results by the average suicide rate in descending order.

In this output, countries are listed along with their average suicide rates, highlighting those with the highest rates. According to the results, **Lithuania**, **Sri** **Lanka**, and the **Russian** **Federation** have some of the **highest** **average** **suicide** **rates** **per** **100,000 people**, **Dominica** and **Saints** **Kitts & Nevis** have **least** **average** **suicide** **rates per 100,000 people**. This data helps to identify countries where suicide is a significant public health issue and can be used for further analysis on contributing factors and prevention strategies.

Gender v/s Average Suicide rate per 100k population:

This image shows the average suicide rate per 100,000 people for each gender from a dataset. The output reveals that the **average suicide rate for females** is approximately **5.39 per 100,000**, while for **males**, it is significantly higher at around **20.24 per 100,000.** This analysis indicates a notable gender difference in suicide rates, with males having a much higher average rate than females in the dataset.

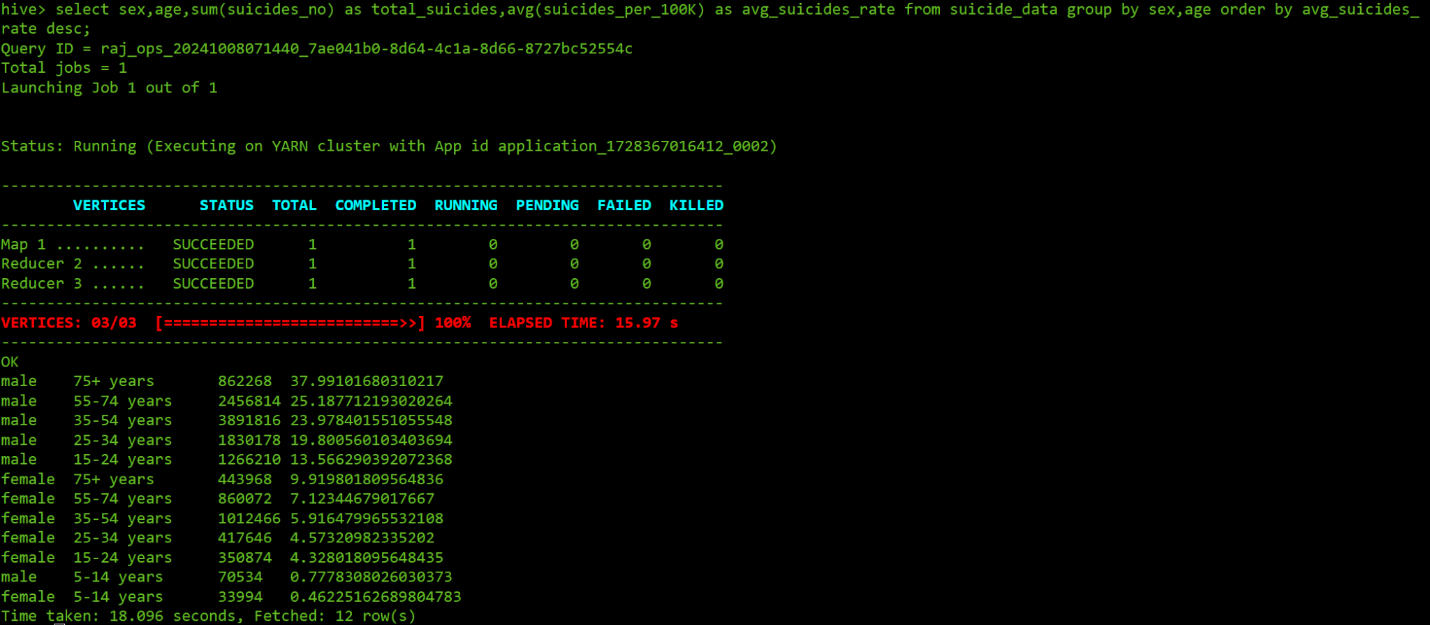
Age v/s Total Suicides v/s Average Suicide rate per 100k population:



This image displays the average suicide rate per 100,000 people for different age groups. The query groups the data by age and orders the results by the average suicide rate in descending order. And the **75+ age group has the highest average suicide rate** at approximately **23.96 per 100,000**, followed by the **55-74** **age group with an average rate of about 16.16.** The suicide rates then gradually decrease for younger age groups, with the lowest rate observed in the **5-14 age group**, at just **0.62 per 100,000.**

The age group between **35-54 years has a greater number of suicides than any other age** group.

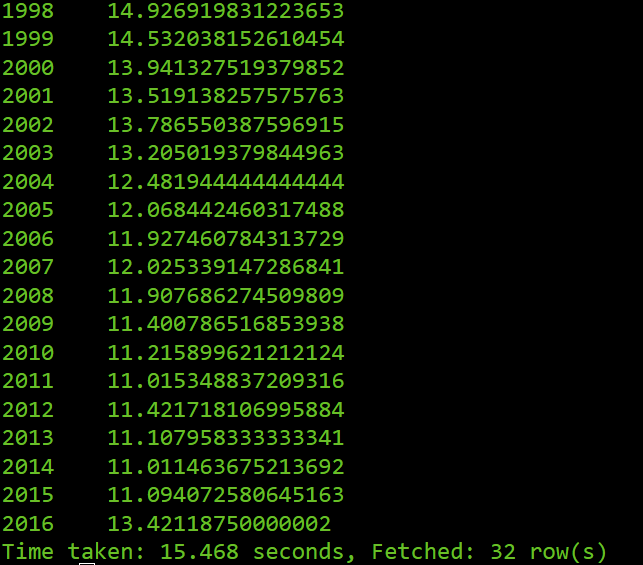
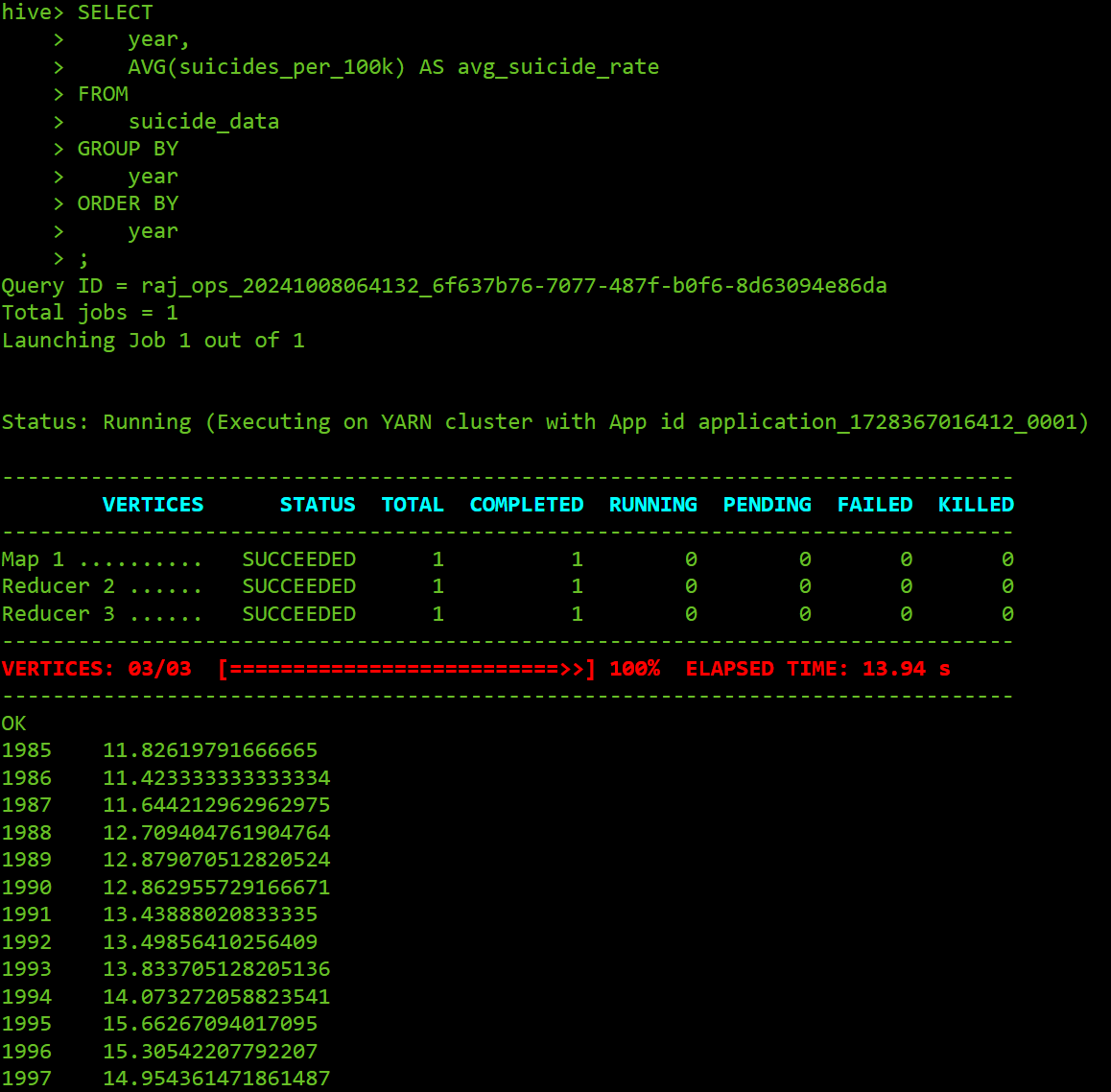
The results provide insight into how suicide rates increase significantly with age, particularly in older demographics. This could suggest that older individuals may face unique risk factors, such as loneliness, physical health challenges, and other mental health concerns, which contribute to higher suicide rates. In contrast, younger age groups exhibit much lower rates, potentially due to greater social support, educational opportunities, and reduced physical health challenges in comparison.

Gender v/s Age v/s Total Suicides v/s Average Suicide rate per 100k population:

The query groups suicide data by gender and age, calculating the total number of suicides and the average suicide rate per 100,000 people for each group. The results are sorted to show the groups with the highest suicide rates first. The findings reveal that older males, especially those **aged 75 and above, have the highest suicide rate**, around **38 per 100,000**. **Men aged 55-74 and 35-54 also show higher rates,** with **averages of about 25 and 20 per 100,000, respectively.**

This data highlights a clear difference in suicide rates between men and women, especially as they age. While suicide rates for older women also rise, they are much lower than those for men. For example, women aged 75 and above have a suicide rate of around 10 per 100,000, which is less than one-third of the rate for men in the same age group. Younger groups, especially females aged 5-14, have the lowest rates, indicating lower risk in these groups.

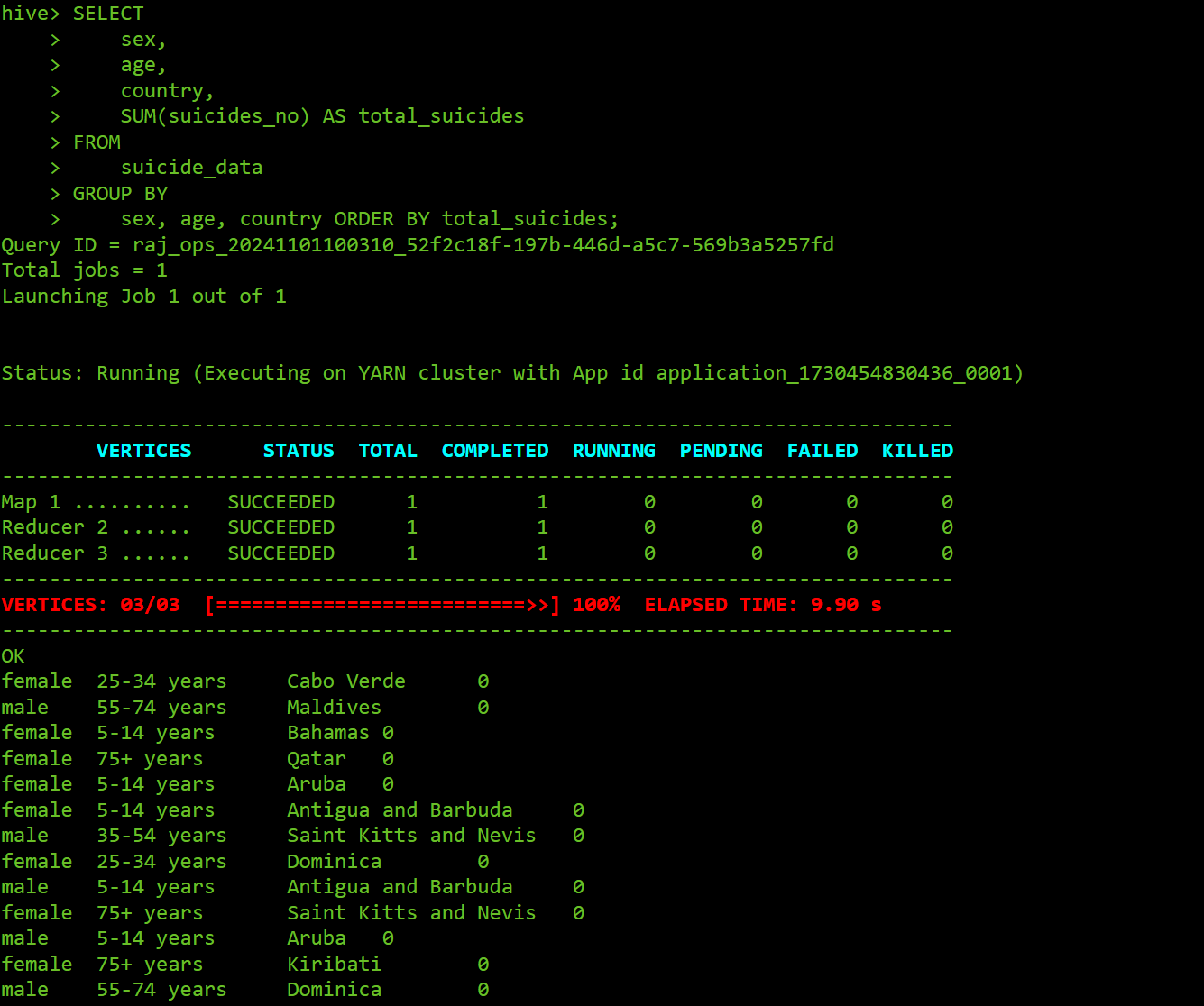
These results suggest that suicide prevention efforts need to be focused more on older men, who are at a higher risk. Providing mental health support tailored to age and gender could help reduce suicide rates, especially among elderly men.

Year v/s Average Suicide rate per 100k population:

The query in this image calculates the average suicide rate per 100,000 people for each year, ordering the results by year to show trends over time. This allows for a clear view of how suicide rates have fluctuated year by year. Starting from **1985**, **the average suicide rate** is around **11.83 per 100,000 people**. Over the years, the rate gradually increased, peaking in the mid-1990s with values reaching above **15 per 100,000 in 1995 and 1996**.

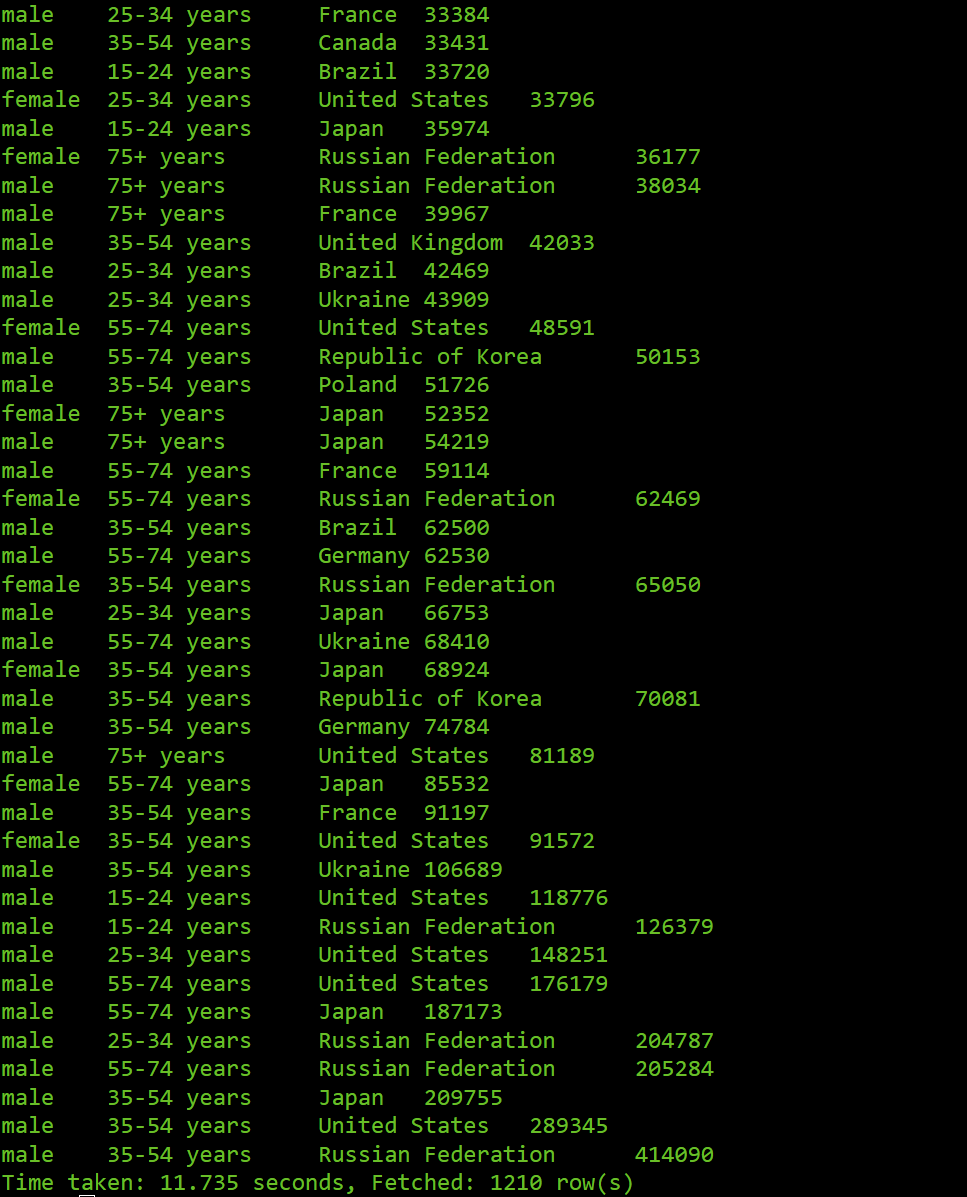
These results suggest that suicide rates had a noticeable upward trend throughout the 1980s and early 1990s, with a high point in the mid-90s. The year **1995** is having the **highest average** **suicide rate per 100k** and **2014** has **lowest average suicide rate per 100k.**

Gender v/s age v/s country v/s total\_suicides:

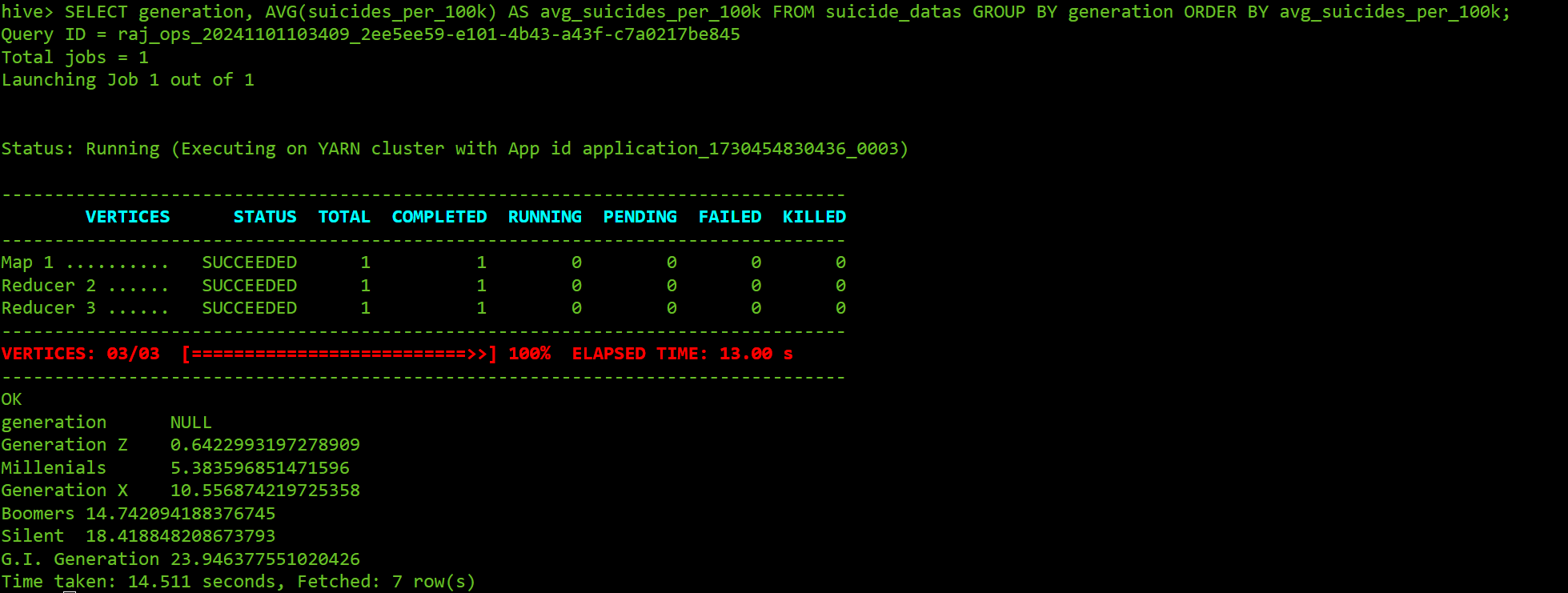


The query calculates the total number of suicides for each combination of sex, age group, and country, based on data stored in the suicide\_data table. The results are sorted by the total number of suicides in ascending order.

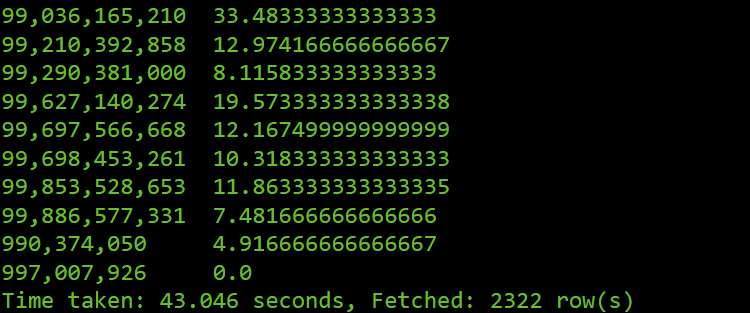
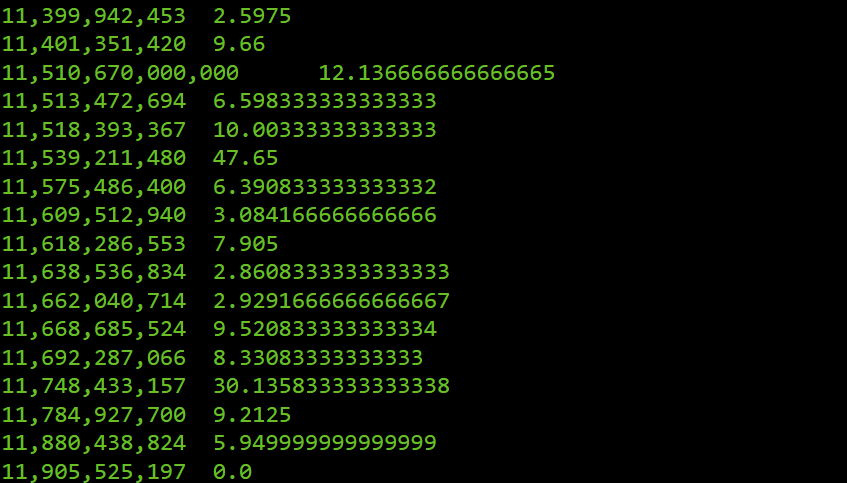
**Russian Federation has more suicides in male at age 35-54 years**. **Cabo verdes, maldives, qatar etc. has zero suicides** in different age groups.



Generation v/s Average Suicide rate per 100k population:

The query calculates the average suicide rate per 100k people for each generation, based on data stored in the suicide\_datas table. The results are sorted by the average suicide rate in ascending order. The results show that **Generation Z has the lowest average suicide rate per 100k people**, followed by Millennials, Generation X, Boomers, Silent Generation, and the G.I. Generation has the highest average suicide rate.

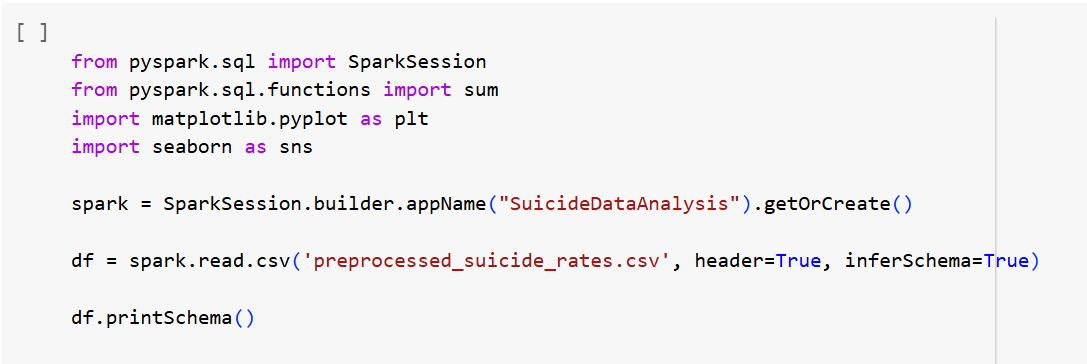
GDP\_Per\_Year v/s Average Suicide rate per 100k population:



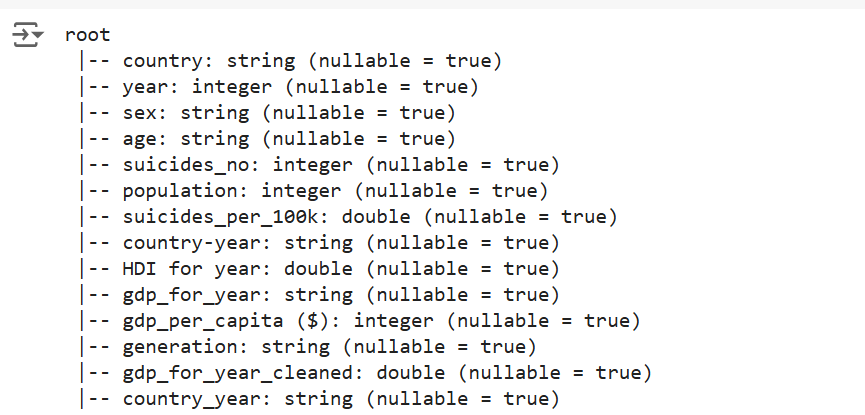
This query will produce a list of GDP years along with the average suicide rate per 100,000 people for each year. The output allows us to observe how the average suicide rate changes across different GDP years, which could help identify any trends or correlations between GDP over time and suicide rates.

As gdp per year increasing the average suicide rate is decreasing vice versa. They both have negative correlation. But there are some fluctuations.

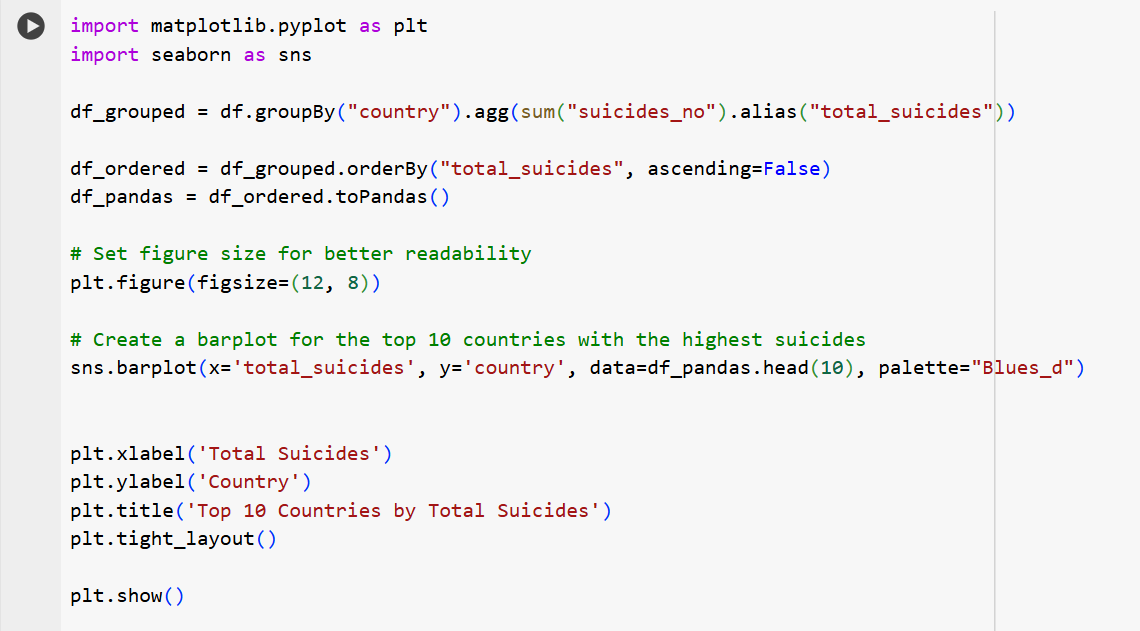
## Visualization of data

****

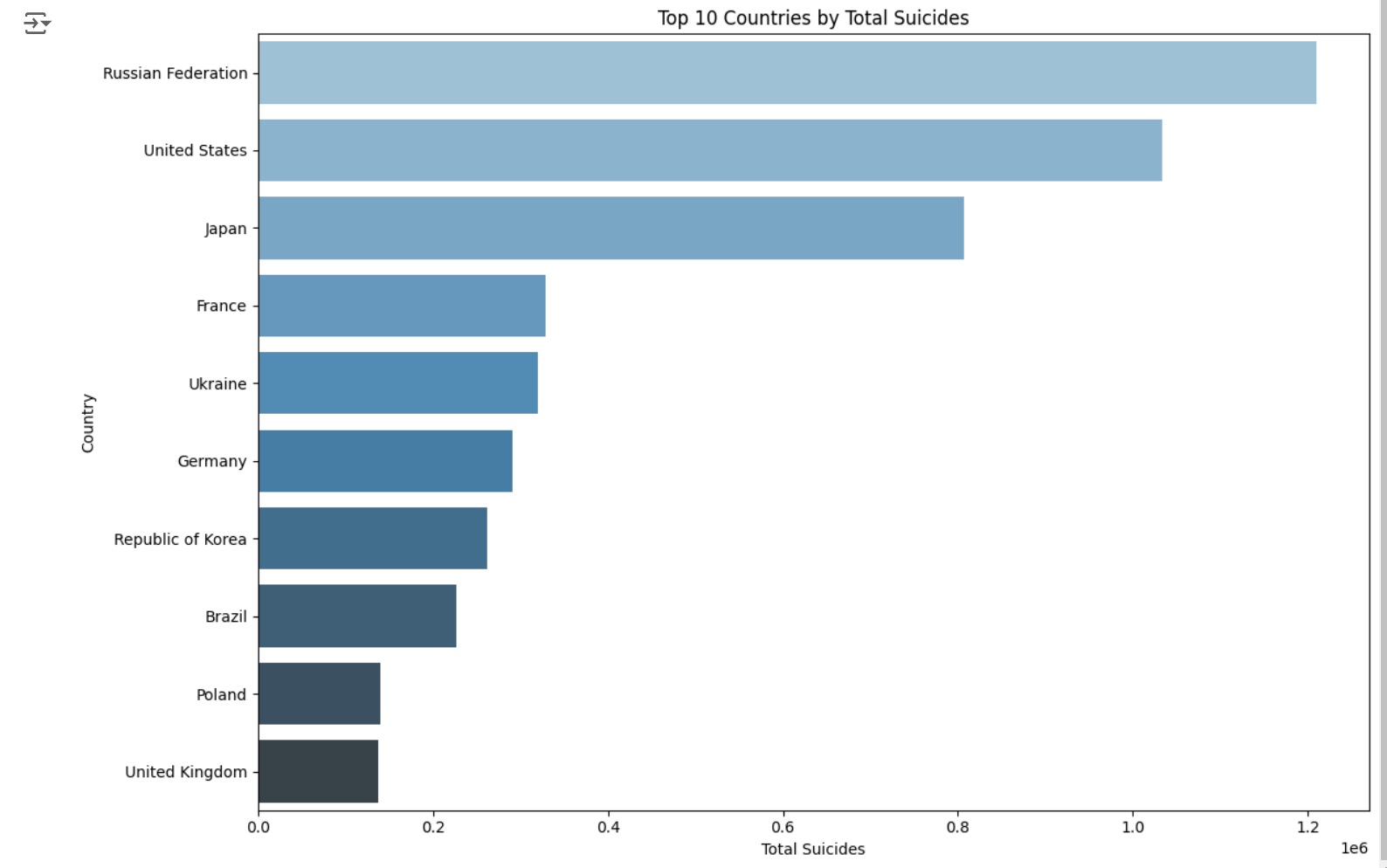
**Output:**

****

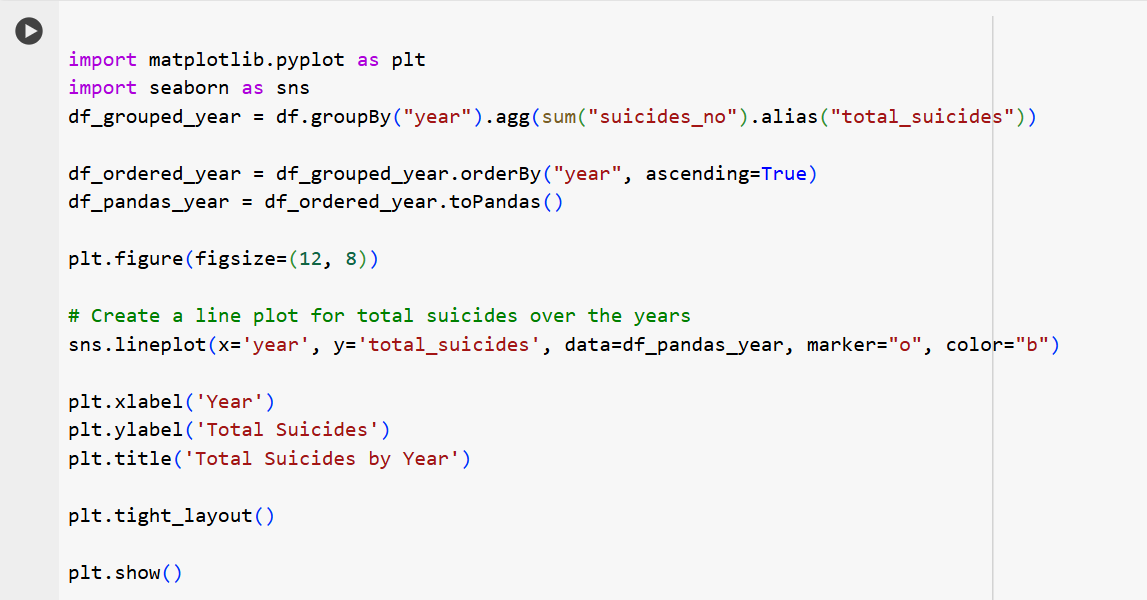
This code initializes a Spark session named "SuicideDataAnalysis" to work with a dataset of suicide rates. After importing necessary libraries for data processing and visualization, it loads a CSV file called **preprocessed\_suicide\_rates.csv** into a Spark DataFrame. The **header=True** option treats the first row as column names, and **inferSchema=True** automatically detects the data types for each column. Finally, **df.printSchema()** displays the **structure of the DataFrame**, showing the names and types of each column, which helps in understanding the dataset's layout before further analysis.

****

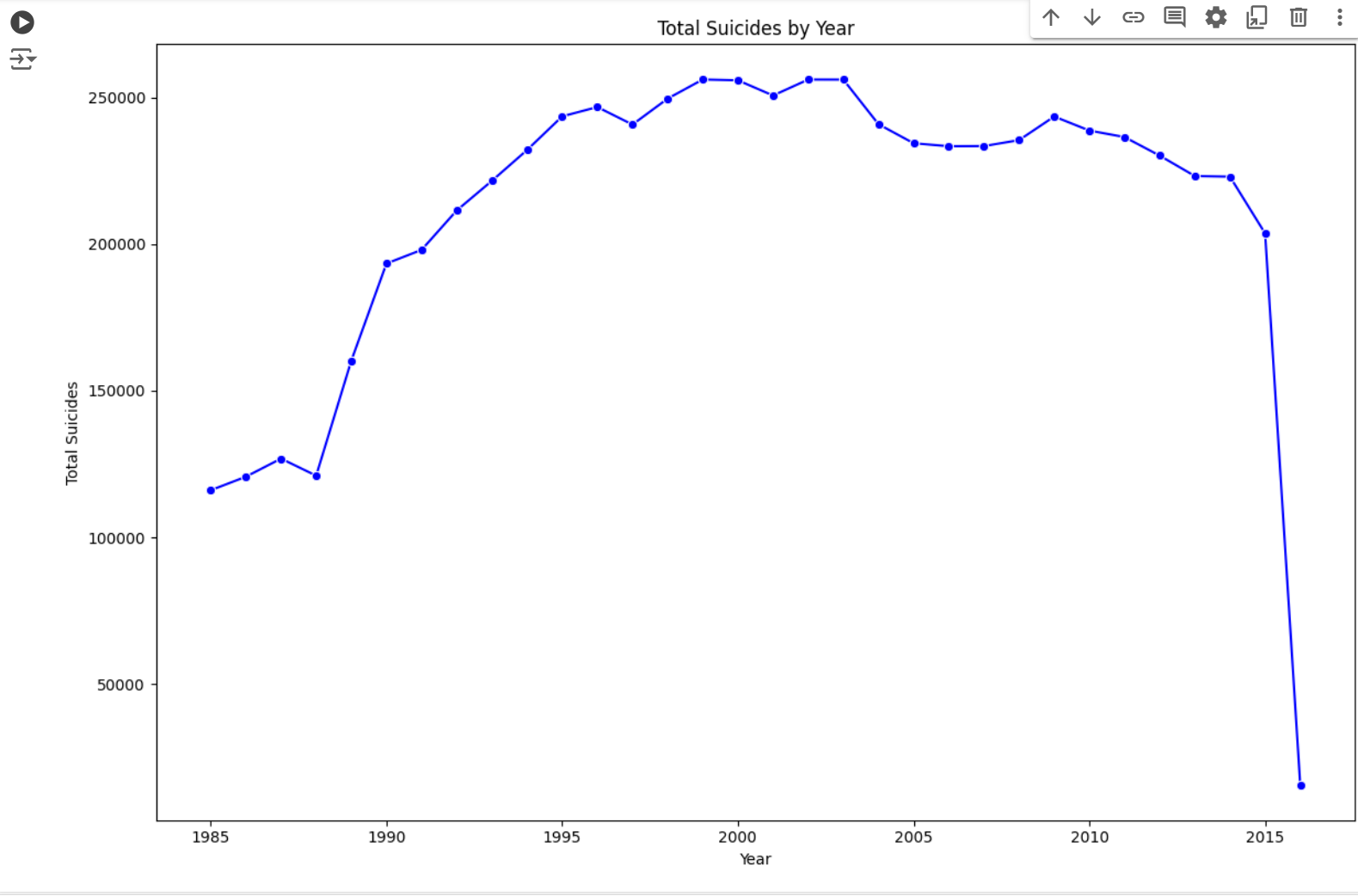
**Output:**

****

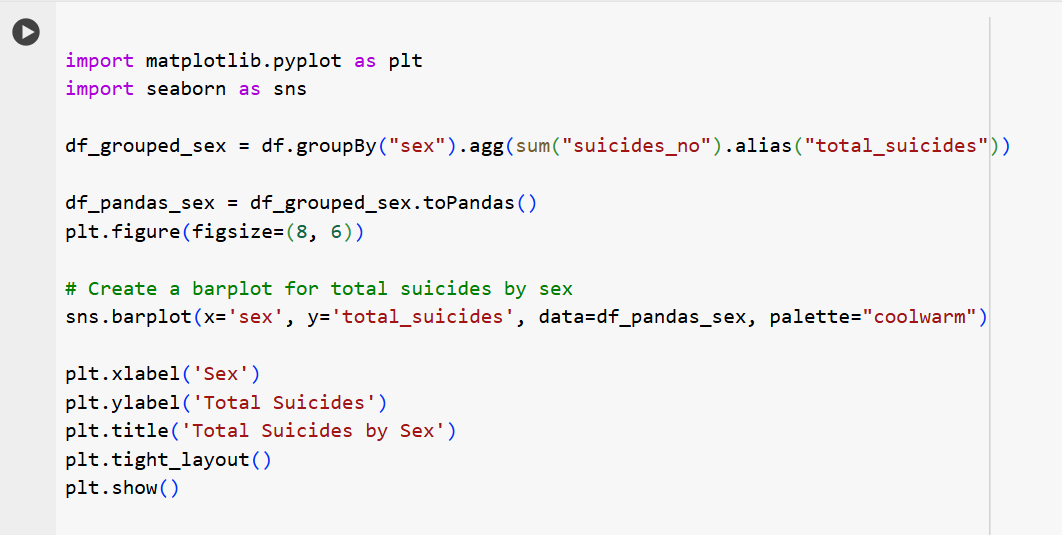
* **Russian Federation**: This country has the highest number of total suicides among the top 10 listed.
* **United States and Japan**: These two countries have the next highest numbers of total suicides.
* **France, Ukraine, Germany, Republic of Korea, Brazil, Poland, and United Kingdom**: These countries have progressively lower numbers of total suicides, with the United Kingdom having the lowest among the top 10 listed.

****

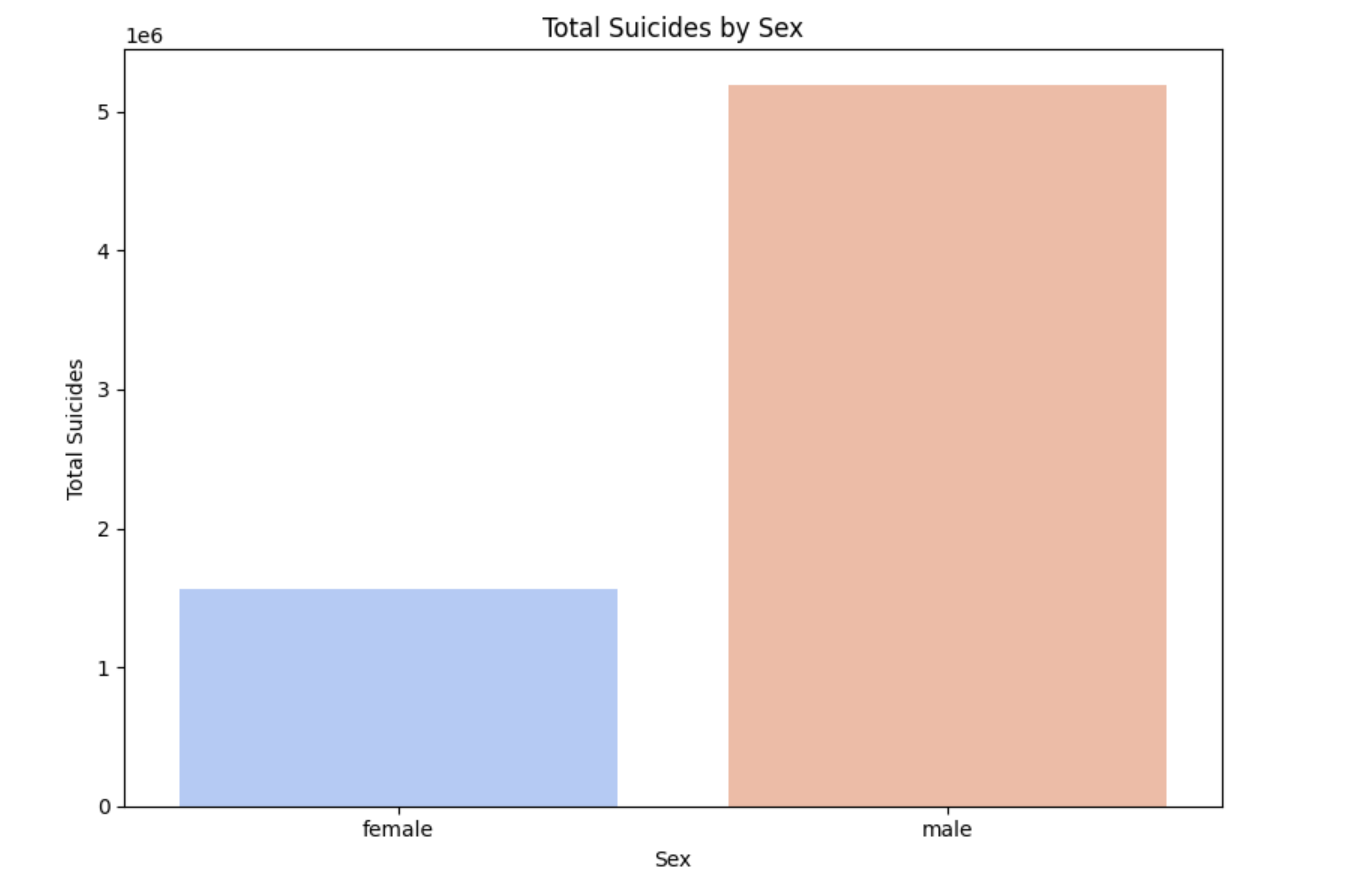
**Output:**

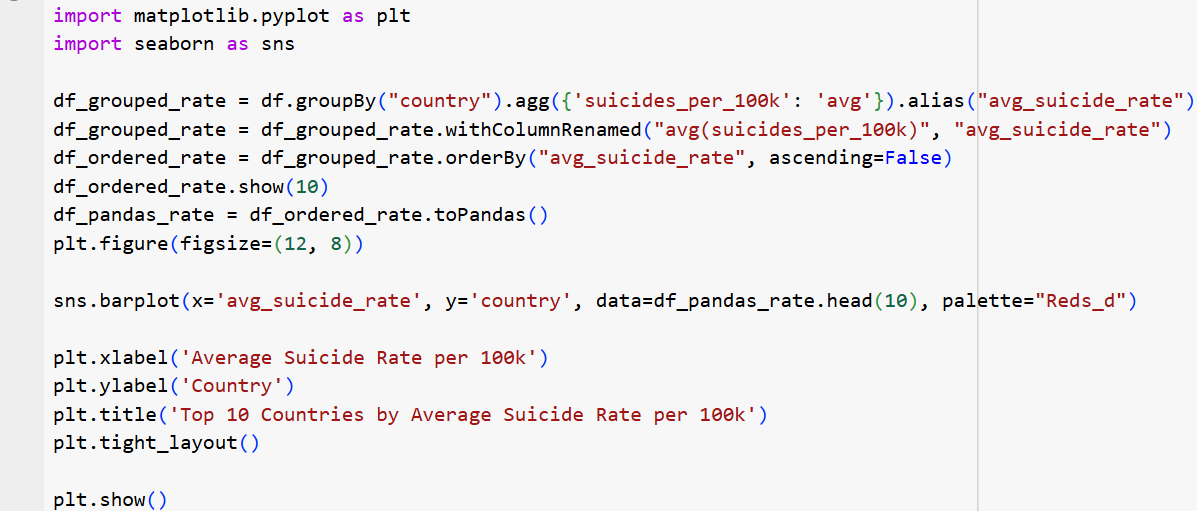
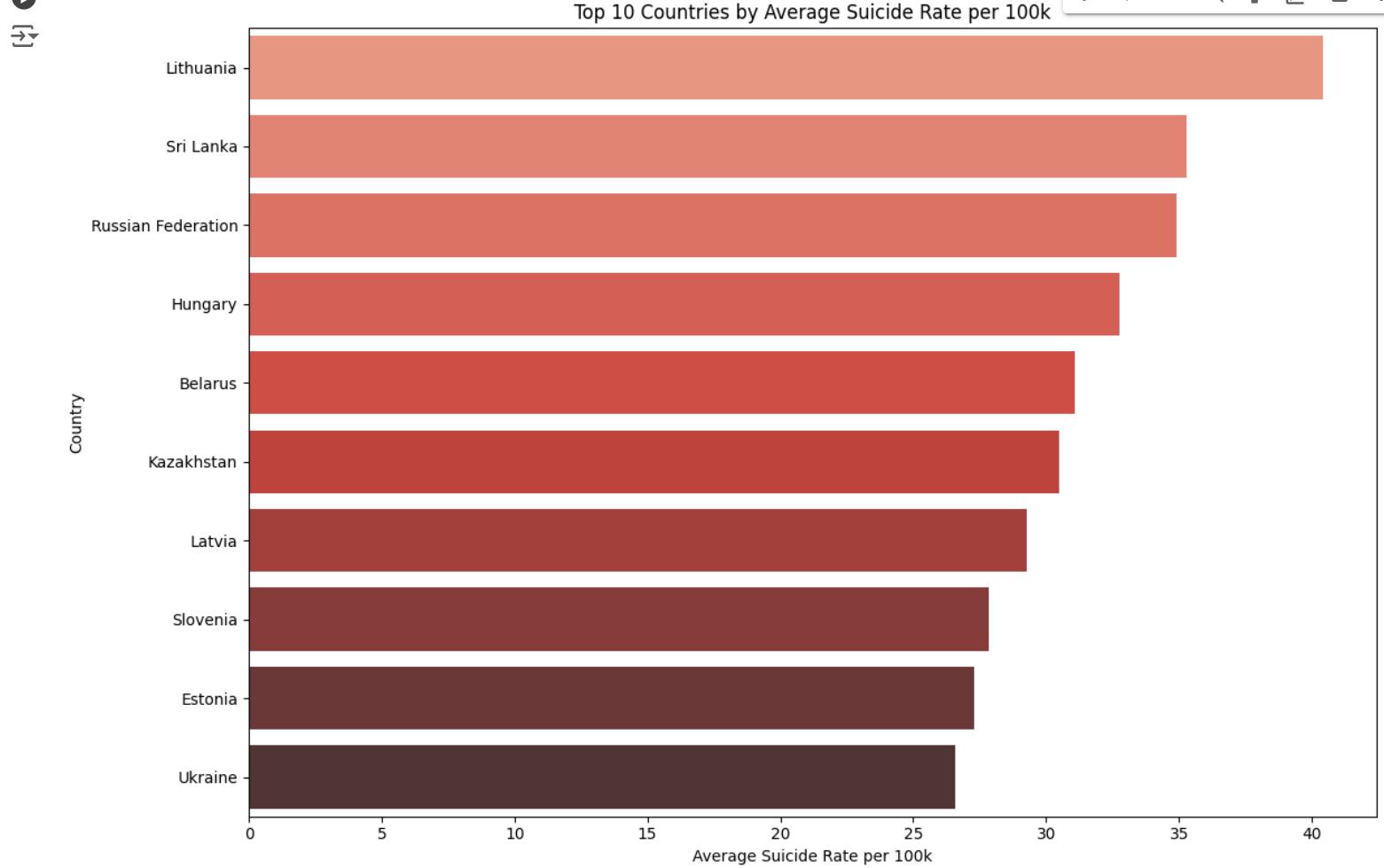
****

* Overall Trend: There is a **general increase in the number of suicides from 1985 to around 2000**, followed by a period of relative stability, and then a **significant decrease in the later years**.
* Fluctuations: The graph shows some fluctuations within this overall trend, with years of higher and lower suicide rates.
* **Peak Years**: The peak years with the highest number of suicides appear to be around **2000**.
* **Sharp Decline**: A sharp decline in suicides is observed **after** **2010**.

****

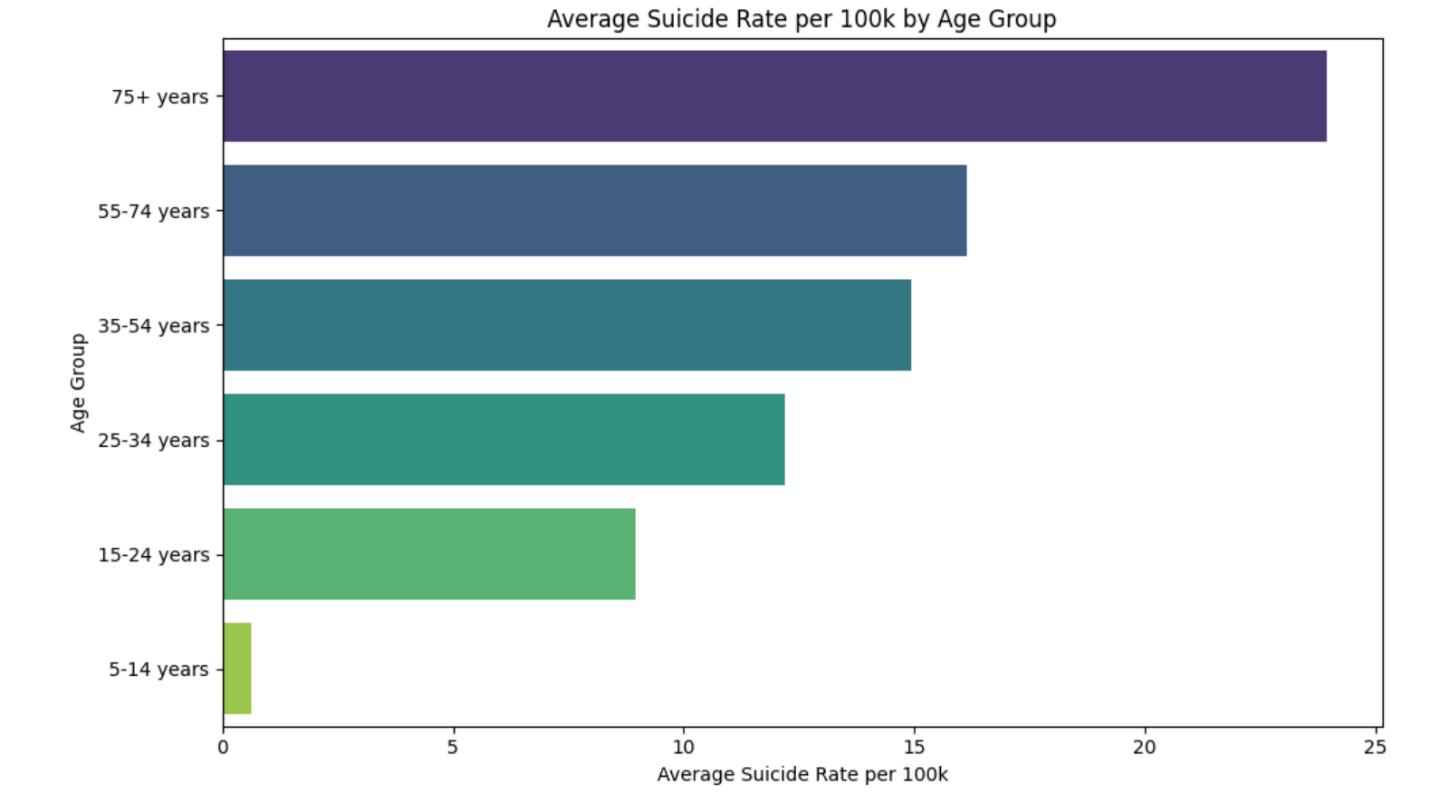
**Output:**

****

* The chart clearly shows that the number of **male suicides is significantly higher than the** number **of female suicides**.

**Output:**

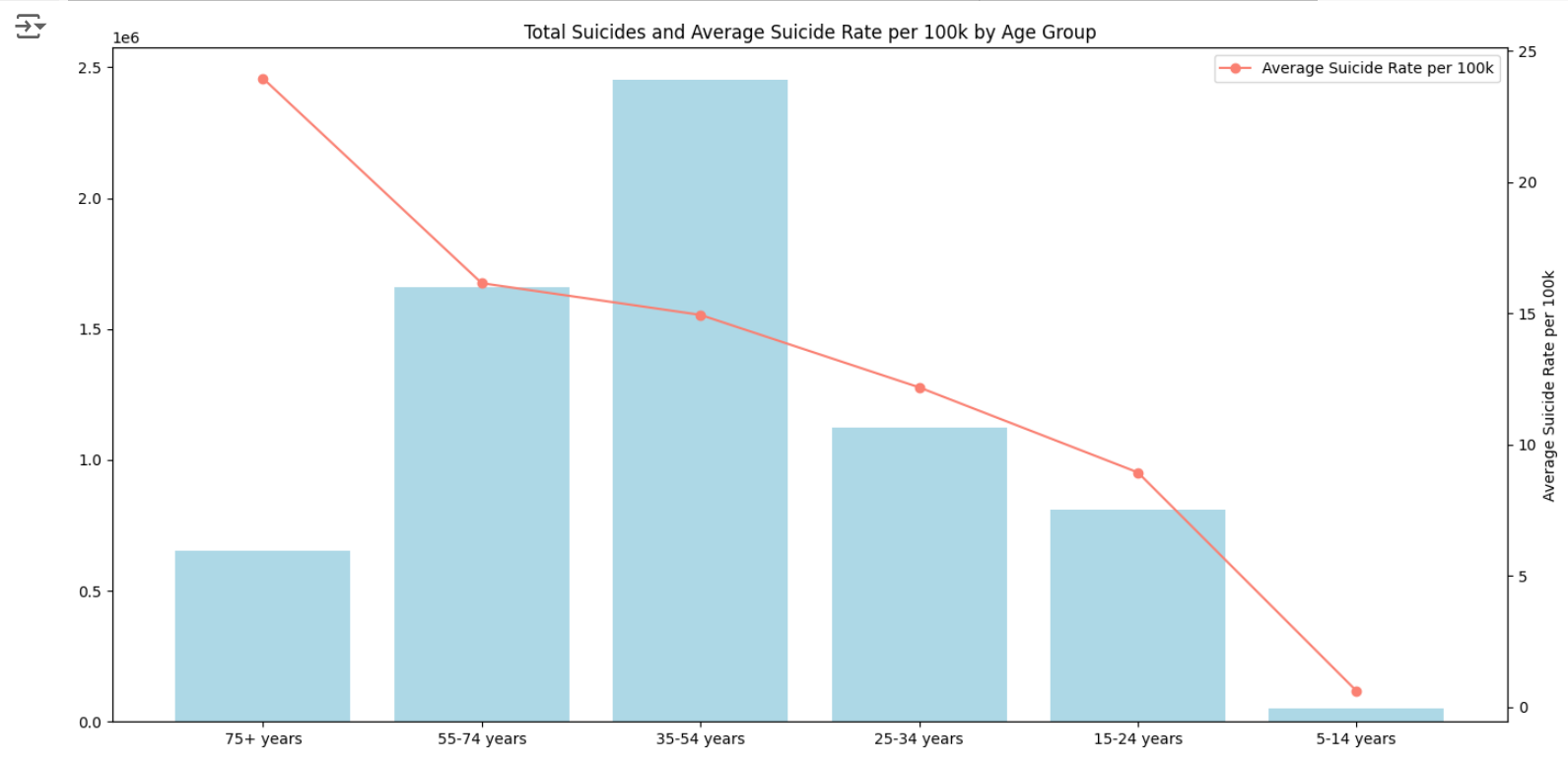
* **Lithuania** has the **highest average suicide rate per 100,000** people among the top 10 listed.
* **Sri Lanka, Russian Federation, Hungary, Belarus, Kazakhstan, Latvia, Slovenia, Estonia, and Ukraine** are countries have progressively **lower average suicide rates per 100,000 people**, with Ukraine having the lowest among the top 10 listed.

**Output:**

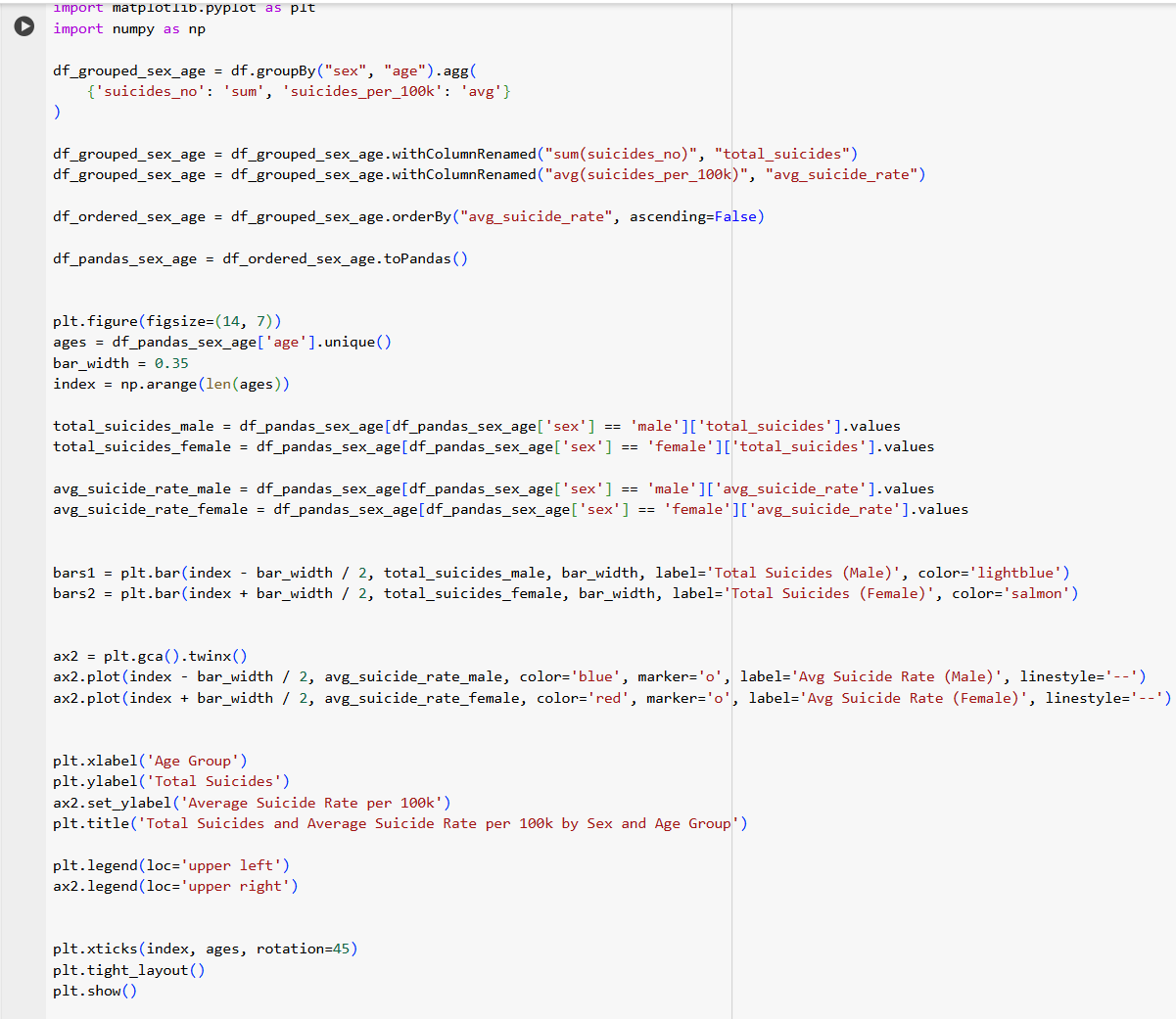
* The chart shows the average suicide rate per 100k people for different age groups, with **the highest rate** observed among the **75+ age group** and the **lowest rate among the 5-14 age group.**



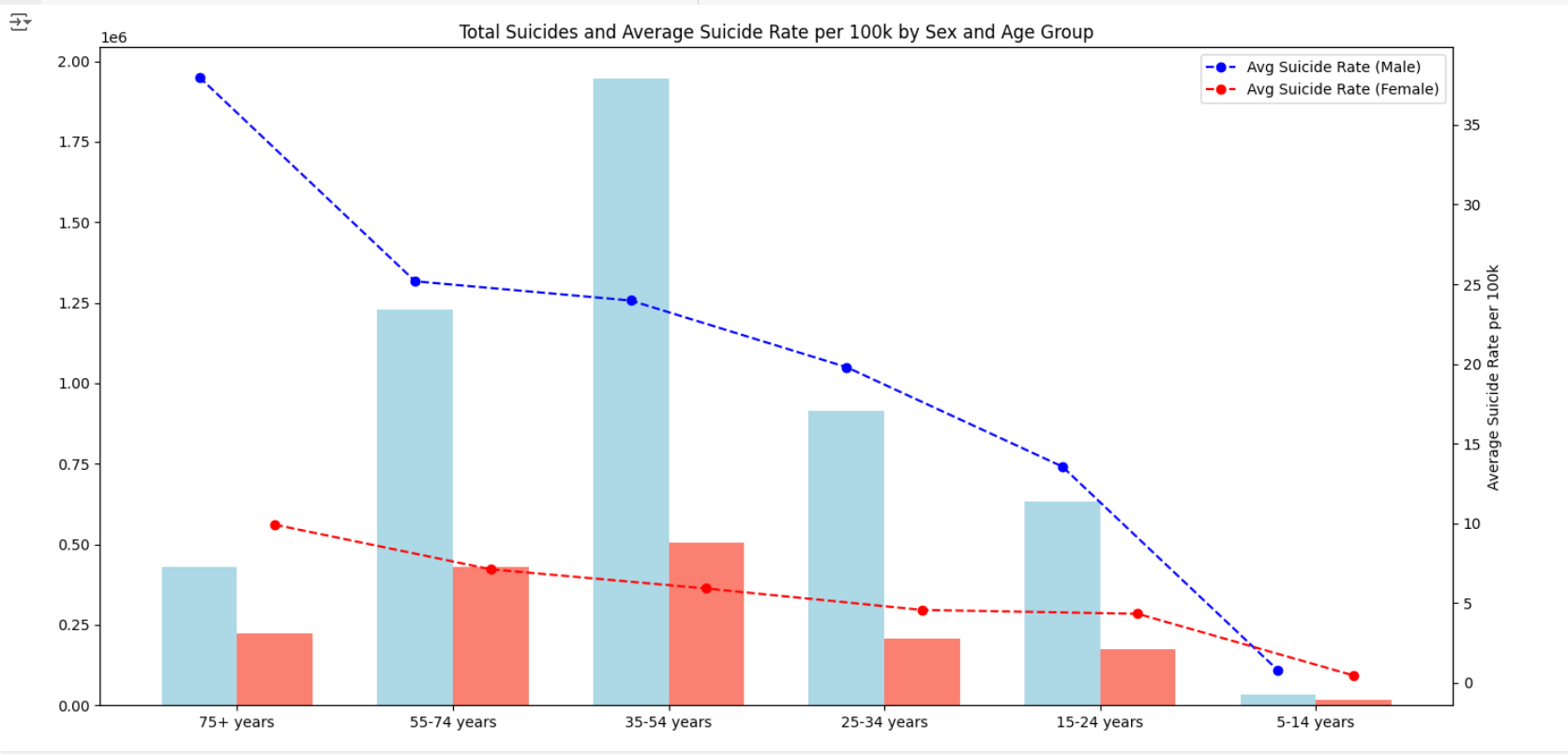
**Output:**

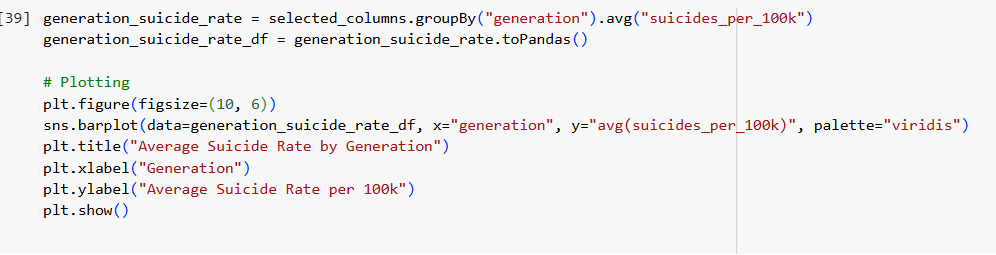


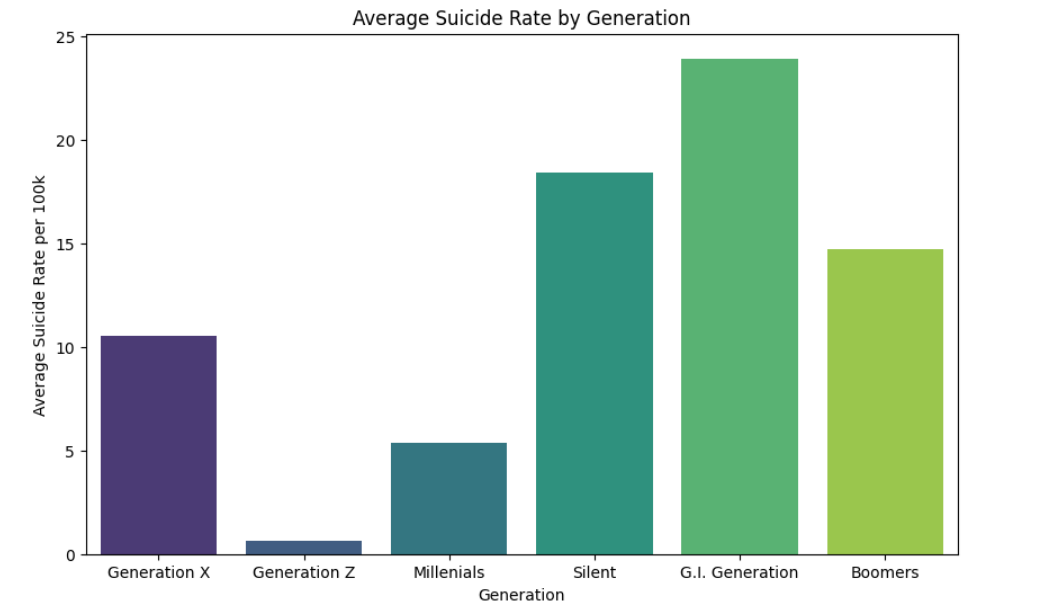
* The chart shows the total number of suicides and the **average suicide rate per 100k people for different age groups**. The **75+ age group has the highest** **total** **number of suicides**, while the **5-14 age group has the lowest**.
* The **average suicide rate per 100k people decreases with age**, with the highest rate among the 75+ age group and the lowest among the 5-14 age group.



**Output:**

****

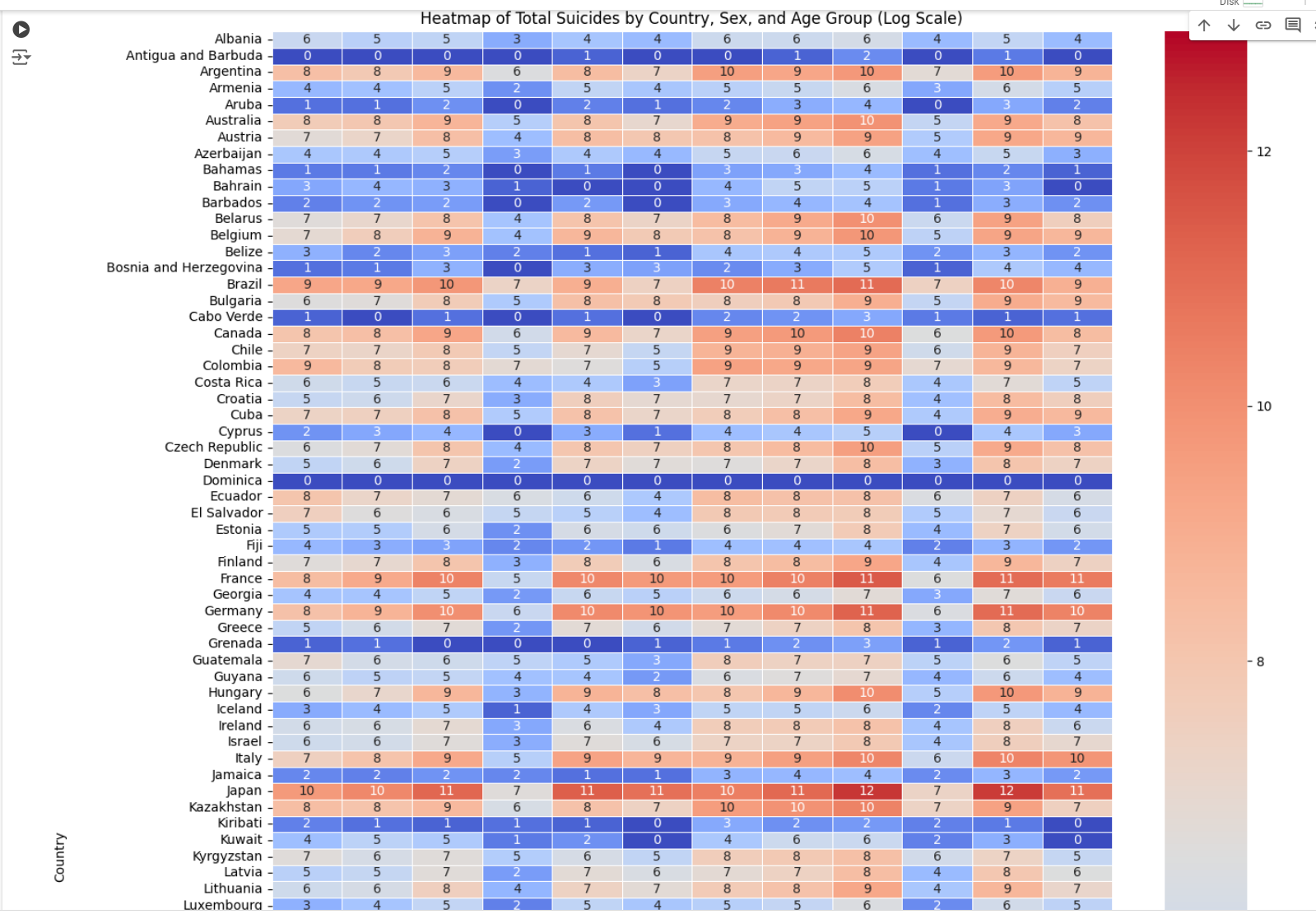
* The chart shows the total number of suicides and the average suicide rate per 100k people for different age groups, broken down by sex.
* The 75+ age group has the highest total number of suicides, while the 5-14 age group has the lowest.
* The **average suicide rate per 100k people** is higher for males than females across all age groups, with the **highest rate observed among older males**.

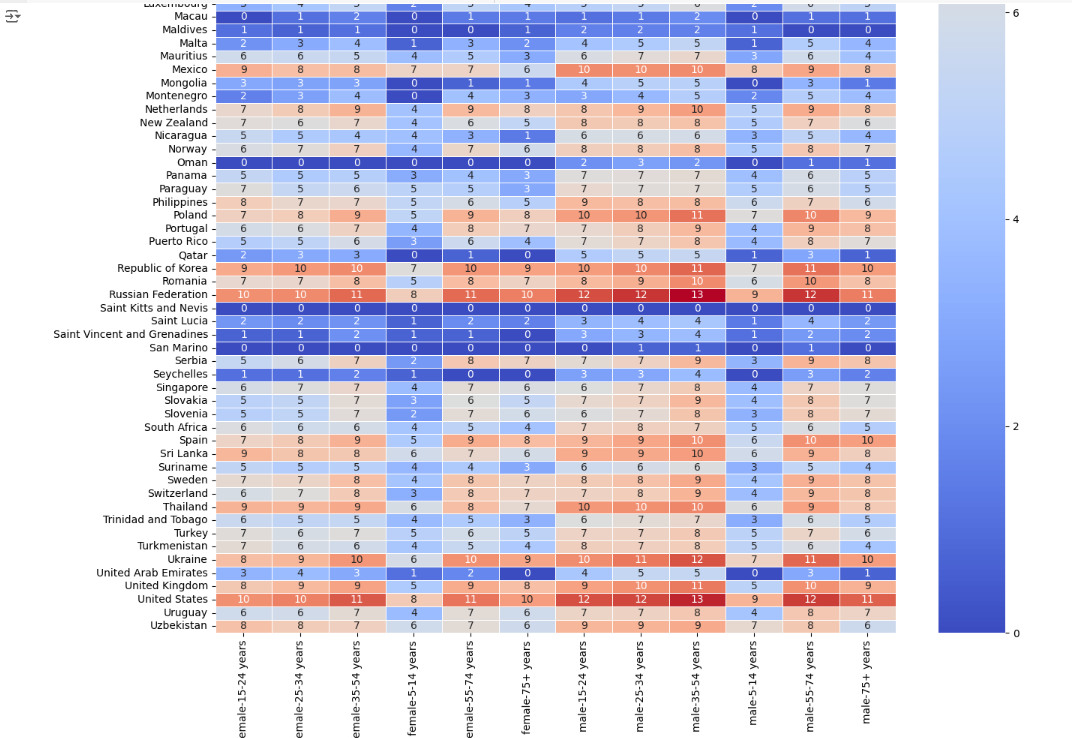
**Output:**

* The chart shows the **average suicide rate per 100k people for different generations**, with the highest rate observed among the G.I. Generation and the lowest rate among Generation Z.



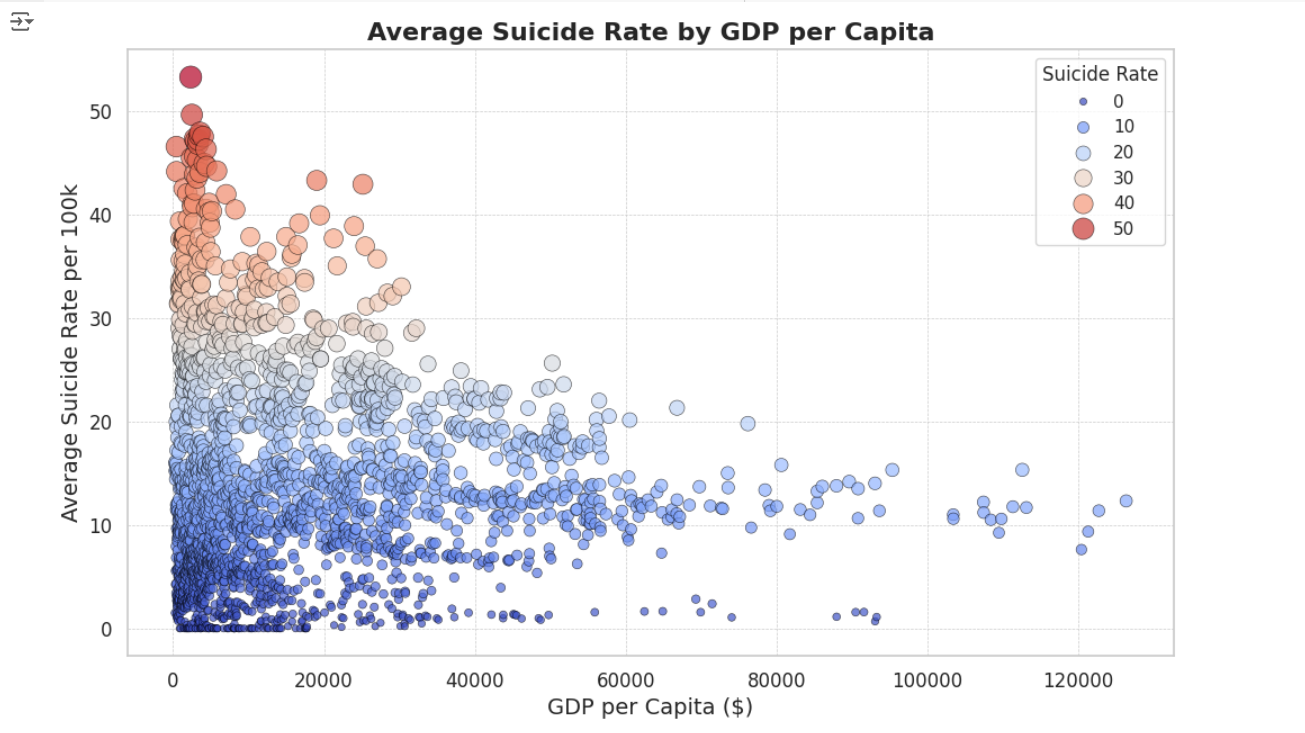
**Output:**





* The heatmap provides a visual representation of the **changes in suicide rates per 100k people for various countries and age groups between two specific time periods**. Each cell in the heatmap corresponds to a particular country-age group combination, and the color intensity indicates the magnitude of change in the suicide rate.
* **Red** shades represent **an increase in the suicide rate**, while **blue** shades indicate a **decrease**.
* In **San Marino** and **Saint Kitts** and **Oman** and (Barbuda & Antigua) has **least suicide** **rates** **in female at all ages.**
* **Dominica** has **zero suicide rates** in both **male** and **female** at **all ages**.
* **Russian Federation** and **United states** and **Korea** and **Japan** have **more suicide rates**, but male suicide rate has slightly more than female.



**Output:**

* The scatter plot shows the relationship between the **average suicide rate per 100k people and the GDP per capita** for different countries. The size and color of the points represent the suicide rate, with larger and darker points indicating higher suicide rates.
* There seems to be a **negative correlation** between the two variables, suggesting that countries with higher GDP per capita tend to have lower suicide rates. However, there is also a significant amount of variation within each GDP level, indicating that other factors besides GDP play a role in suicide rates.

# Conclusion

The analysis of global suicide data gives a clear picture of trends over the years, differences between genders, and variations across countries and age groups. Overall, the data shows that suicides generally increased from 1985 to the late 1990s, peaking around 1995, and then started to gradually decrease. The highest number of suicides was in 1999, with 512,238 cases, while 2016 had the lowest count, with only 31,206 suicides. This trend might be influenced by changes in social conditions, mental health awareness, and prevention efforts over time.

Looking at gender, the data shows a large difference between male and female suicide rates. Males account for over 10 million suicides, while females account for about 3 million. On average, men have a much higher suicide rate (around 20 per 100,000 people) than women (around 5 per 100,000 people). This suggests that men may be at higher risk, and prevention programs might need to focus more on their specific needs.

In terms of countries, some regions have much higher suicide counts and rates. The Russian Federation, the United States, and Japan have the highest total suicides, while countries like Lithuania, Sri Lanka, and Russia have the highest average suicide rates. On the other hand, countries like Dominica and Saint Kitts & Nevis report very low suicide counts and rates, which could be due to differences in cultural, social, or economic factors.

Age also affects suicide rates. The 75+ age group has the highest rate, at nearly 24 suicides per 100,000 people, followed by the 55-74 age group. Rates are much lower for younger age groups, with the lowest in children aged 5-14. However, people aged 35-54 have the highest total number of suicides, indicating that middle-aged and older adults may need more support to reduce suicide risk.

# References

<https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016>

<https://www.who.int/news-room/fact-sheets/detail/suicide>

<https://pmc.ncbi.nlm.nih.gov/articles/PMC9594371/>