HIS project - Analysis of death by suicide

Youssef El fadi - 727582

Introduction

In this report, I will analyze multiple datasets related to three key concepts: **deaths by suicide, mental health disorders among the population, and life expectancy.** Each concept is supported by various datasets, providing valuable insights into different aspects of these phenomena. The ultimate objective is to identify potential relationships and correlations among the data collected within each concept's respective datasets.

Library used

```
suppressPackageStartupMessages({
  library(reshape2)
  library(tidyverse)
  library(data.table)
  library(gridExtra)
  library(maps)
})
```

Import and cleaning of datasets

Suicide - datasets

Suicide: total deaths per year The datasets about **suicide** are taken from the site https://www.gapminder.org/data/. Let's try to take a look at the dataset "suicide_total_deaths":

```
suicide_deaths <- read.csv("./datasets/suicide/suicide_total_deaths.csv", stringsAsFactors = F)
head(suicide_deaths)
```

```
##                country X1990 X1991 X1992 X1993 X1994 X1995 X1996 X1997 X1998
## 1          Afghanistan   696   751   855   943   993  1030  1070  1100  1110
## 2               Angola   973   991  1020  1070  1110  1120  1080  1090  1160
## 3              Albania   125   136   132   131   125   133   141   158   170
## 4              Andorra  5.96  6.46  6.81  7.25  7.24  7.11  7.09     7  6.97
## 5 United Arab Emirates   114   121   130   140   147   161   170   181   190
## 6            Argentina  2760  2870  3070  3190  3340  3500  3710  3970  4160
##   X1999 X2000 X2001 X2002 X2003 X2004 X2005 X2006 X2007 X2008 X2009 X2010 X2011
## 1  1120  1140  1180  1190  1230  1280  1300  1300  1310  1330  1340  1370  1390
## 2  1200  1220  1230  1220  1250  1310  1310  1350  1370  1400  1450  1500  1550
## 3   163   151   148   154   165   170   169   162   158   163   160   160   160
## 4   6.9  6.89  6.87  6.91  7.29  7.68  7.74  7.95  8.04  8.12   8.1  8.08  7.46
## 5   200   204   213   223   222   224   231   265   339   432   516   541   562
## 6  4330  4390  4550  4600  4460  4290  4280  4310  4500  4550  4570  4580  4680
##   X2012 X2013 X2014 X2015 X2016 X2017 X2018 X2019
```

```
## 1  1410  1430  1440  1470  1500  1550  1580  1610
## 2  1600  1640  1650  1690  1720  1800  1870  1930
## 3   158   157   158   160   158   156   154   152
## 4  7.28  7.29  7.49  7.56  7.63  7.81  7.98  8.13
## 5   580   610   636   645   648   648   655   664
## 6  4780  4840  4810  4770  4930  4980  5010  5030
```

```
names(suicide_deaths)
```

```
##  [1] "country" "X1990"   "X1991"   "X1992"   "X1993"   "X1994"   "X1995"
##  [8] "X1996"   "X1997"   "X1998"   "X1999"   "X2000"   "X2001"   "X2002"
## [15] "X2003"   "X2004"   "X2005"   "X2006"   "X2007"   "X2008"   "X2009"
## [22] "X2010"   "X2011"   "X2012"   "X2013"   "X2014"   "X2015"   "X2016"
## [29] "X2017"   "X2018"   "X2019"
```

In the dataset, the rows represent different countries, while the columns represent different years. Each data point represents the number of deaths by suicide registered in a specific year for a particular country.

Furthermore, it seems that when using the **read.csv(...)** function, the letter "X" was added to the column names. To address this issue, we can re-import the CSV file, this time specifying the attribute **check.names=FALSE** to prevent the addition of "X" to the column names.

```
suicide_deaths <- read.csv("./datasets/suicide/suicide_total_deaths.csv", stringsAsFactors = F, check.na
head(suicide_deaths)
```

```
##                country 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000
## 1          Afghanistan  696  751  855  943  993 1030 1070 1100 1110 1120 1140
## 2               Angola  973  991 1020 1070 1110 1120 1080 1090 1160 1200 1220
## 3              Albania  125  136  132  131  125  133  141  158  170  163  151
## 4              Andorra 5.96 6.46 6.81 7.25 7.24 7.11 7.09    7 6.97  6.9 6.89
## 5 United Arab Emirates  114  121  130  140  147  161  170  181  190  200  204
## 6            Argentina 2760 2870 3070 3190 3340 3500 3710 3970 4160 4330 4390
##   2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
## 1 1180 1190 1230 1280 1300 1300 1310 1330 1340 1370 1390 1410 1430 1440 1470
## 2 1230 1220 1250 1310 1310 1350 1370 1400 1450 1500 1550 1600 1640 1650 1690
## 3  148  154  165  170  169  162  158  163  160  160  160  158  157  158  160
## 4 6.87 6.91 7.29 7.68 7.74 7.95 8.04 8.12  8.1 8.08 7.46 7.28 7.29 7.49 7.56
## 5  213  223  222  224  231  265  339  432  516  541  562  580  610  636  645
## 6 4550 4600 4460 4290 4280 4310 4500 4550 4570 4580 4680 4780 4840 4810 4770
##   2016 2017 2018 2019
## 1 1500 1550 1580 1610
## 2 1720 1800 1870 1930
## 3  158  156  154  152
## 4 7.63 7.81 7.98 8.13
## 5  648  648  655  664
## 6 4930 4980 5010 5030
```

Given the fact that we have only numbers in the dataset, except for the column *country*, let's check if the data type used is *numeric* for the column 1990. We can verify this by examining the class of the values in the 1990 column using the *class()* function in R.

```
class(suicide_deaths$"1990")
```

```
## [1] "character"
```

The datatype recognized for the data is **character**, this can be caused by NA value or by the presence of characters within the data. In fact it can be caused by the notation used: for example number can be expressed in scientific notation (which is handled and interpreted correctly by R) or for example it can be

used the characters **"k,K"** for expressing **thousands**, or **"m,M"** for expressing **millions** and so on. The last case needs to be taken care of. Let's see if the first column has any problem (NA value):

```
na_1990 <- sum(is.na(suicide_deaths$"1990"))
print(na_1990)
```

```
## [1] 0
```

We can see that there are no NA values in this column, let's check if an explicit coercion to numeric gives us any error or warning:

```
coercion_1990 <- sum(is.na(as.numeric(suicide_deaths$"1990")))
```

```
## Warning: NAs introduced by coercion
```

```
print(coercion_1990)
```

```
## [1] 9
```

These tests confirm the presence of characters in 9 rows of the first column **1990**. Note that the coercion in the numeric format of numbers expressed in scientific notation within a string does not give any errors, so we proceed to test the presence of **k, M,...** and we will substitute them respectively with **e3** and **e6**. Finally we convert the new value in numeric using **as.numeric(...)** and then check if the substitution+conversion gives any error or NA values.

```
copy_suicide_deaths <- cbind(suicide_deaths)

copy_suicide_deaths[, -1] <- lapply(copy_suicide_deaths[, -1], function(x) {
  gsub("k", "e3", gsub("m", "e6", tolower(x)))
})

copy_suicide_deaths[, -1] <- lapply(copy_suicide_deaths[, -1], as.numeric)

sum(is.na(copy_suicide_deaths$"1990"))
```

```
## [1] 0
```

Another observation we can make is that some numbers in the dataset, which were displayed at the beginning, are in floating-point format rather than integers. This might seem peculiar considering that the dataset pertains to the number of deaths per year. However, it is important to note that the dataset provider explicitly states that the values represent the '*Total number of estimated deaths from self-inflicted injury*'.

Given this information, we will proceed to convert the values to integers. Since rounding these estimated values does not significantly impact our analysis and introduces a maximum error of ± 1, we can confidently round them to the nearest integer without compromising the overall validity of our findings.

Suicide: annual total number of deaths by age group I found another dataset which contains data about the death by suicide collected according some range of age. The dataset is taken from https://ourworldindata.org/suicide.

```
suicide_deaths_by_age <- read.csv("./datasets/suicide/suicide_deaths_by_age.csv", stringsAsFactors = F)
names(suicide_deaths_by_age)
```

```
## [1] "Entity"
## [2] "Code"
## [3] "Year"
## [4] "Deaths...Self.harm...Sex..Both...Age..70..years..Number."
## [5] "Deaths...Self.harm...Sex..Both...Age..50.69.years..Number."
## [6] "Deaths...Self.harm...Sex..Both...Age..15.49.years..Number."
```

```
## [7] "Deaths...Self.harm...Sex..Both...Age..5.14.years..Number."
```

Looking at the dataset and the column names, we can determine that we are not interested in the 'code' column, which represents the abbreviation of the country. Therefore, we will drop this column from our analysis. Additionally, we will rename the first column from 'Entity' to 'country' to better represent its content.

Furthermore, we will modify the names of the last columns to have the following format: '*age_70*', '*age_50_69*', '*age_15_49*', and '*age_5_14*'.

```
suicide_deaths_by_age <- subset(suicide_deaths_by_age, select = -c(2))

names <- c("country", "year", "age_70", "age_50_69", "age_15_49", "age_5_14")

colnames(suicide_deaths_by_age) <- names
```

```
names(suicide_deaths_by_age)
```

```
## [1] "country"   "year"      "age_70"    "age_50_69" "age_15_49" "age_5_14"
```

The result of these modifications are now showed:

```
head(suicide_deaths_by_age)
```

```
##       country year age_70 age_50_69 age_15_49 age_5_14
## 1 Afghanistan 1990     35       167       482       12
## 2 Afghanistan 1991     35       168       535       12
## 3 Afghanistan 1992     36       171       634       14
## 4 Afghanistan 1993     37       176       716       15
## 5 Afghanistan 1994     38       180       759       15
## 6 Afghanistan 1995     38       183       795       16
```

We can notice that from this dataset we can easily obtain the information about the total death per year which we had in the first separate dataset. So let's proceed with the creation of a new column that we will call **total_death** and the value will be the sum of each row. The new dataset will be displayed.

```
suicide_deaths_by_age <- suicide_deaths_by_age %>%
  mutate(total_death = select(., age_70:age_5_14) %>%
  apply(1, sum, na.rm=TRUE))

suicide_deaths <- suicide_deaths_by_age
rm(suicide_deaths_by_age)
head(suicide_deaths)
```

```
##       country year age_70 age_50_69 age_15_49 age_5_14 total_death
## 1 Afghanistan 1990     35       167       482       12         696
## 2 Afghanistan 1991     35       168       535       12         750
## 3 Afghanistan 1992     36       171       634       14         855
## 4 Afghanistan 1993     37       176       716       15         944
## 5 Afghanistan 1994     38       180       759       15         992
## 6 Afghanistan 1995     38       183       795       16        1032
```

```
na_values_suicide_by_age <- sum(is.na(suicide_deaths))
na_values_suicide_by_age
```

```
## [1] 0
```

There are no NA values present in the dataset.

Suicide: male to female ratio of suicide rates This dataset is also taken from https://ourworldindata.org/suicide.

```
suicide_male2female <- read.csv("./datasets/suicide/Male-Female-Ratio-of-Suicide-Rates.csv", stringsAsFa
head(suicide_male2female)
```

```
##        Entity Code Year Male.female.suicide.ratio
## 1 Afghanistan  AFG 1990                      2.70
## 2 Afghanistan  AFG 1991                      2.73
## 3 Afghanistan  AFG 1992                      2.75
## 4 Afghanistan  AFG 1993                      2.78
## 5 Afghanistan  AFG 1994                      2.79
## 6 Afghanistan  AFG 1995                      2.81
```

In these datasets, we will drop the column Code and rename *Entity* to *Country*, as well as rename *Male.female.suicide.ratio* to *Male_to_female_ratio*.

Upon examining the dataset, we can observe that the data is provided for the years 1990-2017, whereas the previous dataset had data available from 1990-2019.

It is important to note that the dataset does not include information about the population breakdown by gender for each country. To facilitate comparisons and to analyze the growth of the male-to-female ratio, we will import another dataset from https://ourworldindata.org/world-population-growth.

```
population_dataset <- read.csv("./datasets/population/population-male.csv", stringsAsFactors = F)
names(population_dataset)
```

```
##  [1] "Country.name"
##  [2] "Year"
##  [3] "Male.population"
##  [4] "Male.population.of.children.under.the.age.of.1"
##  [5] "Male.population.of.children.under.the.age.of.5"
##  [6] "Male.population.of.children.under.the.age.of.15"
##  [7] "Male.population.under.the.age.of.25"
##  [8] "Male.population.aged.15.to.64.years"
##  [9] "Male.population.older.than.15.years"
## [10] "Male.population.older.than.18.years"
## [11] "Male.population.at.age.1"
## [12] "Male.population.aged.1.to.4.years"
## [13] "Male.population.aged.5.to.9.years"
## [14] "Male.population.aged.10.to.14.years"
## [15] "Male.population.aged.15.to.19.years"
## [16] "Male.population.aged.20.to.29.years"
## [17] "Male.population.aged.30.to.39.years"
## [18] "Male.population.aged.40.to.49.years"
## [19] "Male.population.aged.50.to.59.years"
## [20] "Male.population.aged.60.to.69.years"
## [21] "Male.population.aged.70.to.79.years"
## [22] "Male.population.aged.80.to.89.years"
## [23] "Male.population.aged.90.to.99.years"
## [24] "Male.population.older.than.100.years"
```

```
population_dataset <- subset(population_dataset, select = c("Country.name", "Year", "Male.population"))

names <- c("country", "year", "male_population")

colnames(population_dataset) <- names
```

```
na_values_male2female <- sum(is.na(suicide_male2female))
na_values_male_population <- sum(is.na(population_dataset))

na_values_male2female
```

```
## [1] 0
```

```
na_values_male_population
```

```
## [1] 0
```

There are no NA values present in both of the datasets.

The next step will be to join these two datasets, and then the resulting dataset will be joined with the previous one containing the data regarding deaths by suicide.

To perform the join, we will use the **inner_join(...)** function, with the join made on the columns *country* and *year*. Subsequently, we will assess how many data points are lost due to non-matching between the two datasets.

```
suicide_male2female <- inner_join(x = suicide_male2female, y = population_dataset, by=c("country", "yea
```

```
old_nrows_suicide_deaths = nrow(suicide_deaths)
suicide_deaths <- inner_join(x = suicide_deaths, y = suicide_male2female, by=c("country", "year"))
```

```
percentage_data_lost <- (1-(nrow(suicide_deaths)/old_nrows_suicide_deaths))*100
percentage_data_lost
```

```
## [1] 20.5848
```

We lost approximately **20.5%** due to the join. We still have **5432 observations** which correspond to approximately **194 countries** (for each country we have data for 28 years).

The final dataset that we are going to use for the plots and for our analyses is:

```
head(suicide_deaths)
```

```
##       country year age_70 age_50_69 age_15_49 age_5_14 total_death
## 1 Afghanistan 1990     35       167       482       12         696
## 2 Afghanistan 1991     35       168       535       12         750
## 3 Afghanistan 1992     36       171       634       14         855
## 4 Afghanistan 1993     37       176       716       15         944
## 5 Afghanistan 1994     38       180       759       15         992
## 6 Afghanistan 1995     38       183       795       16        1032
##   male_to_female_ratio male_population
## 1                 2.70         5348392
## 2                 2.73         5372964
## 3                 2.75         6028498
## 4                 2.78         7003645
## 5                 2.79         7733464
## 6                 2.81         8219472
```

Mental health - datasets

Mental health: population with anxiety by country In this section we are going to import a dataset downloaded from https://ourworldindata.org/mental-health. The dataset is about the total number of people for whom an anxiety disorder was registered.

```
anxiety_country <- read.csv("./datasets/mental health/number-with-anxiety-disorders-country.csv", string
names(anxiety_country)
```

```
## [1] "Entity"
## [2] "Code"
## [3] "Year"
## [4] "Prevalence...Anxiety.disorders...Sex..Both...Age..All.Ages..Number."
```

In this dataset, we will once again drop the column *Code*, rename *Entity* to *country*, and rename *Prevalence... Anxiety.disorders... Sex..Both... Age..All.Ages..Number* to *people_with_anxiety_disorders*.

The data provided for each country spans from the year **1990** to **2019**. Let's check if there are any NA values:

```
total_NA <- sum(is.na(anxiety_country))
total_NA
```

```
## [1] 0
```

There is no other cleaning needed for this dataset.

Mental health: population with depression by country The last dataset concerning **mental health** is also imported from https://ourworldindata.org/mental-health and is about the total number of people with depression by country from the year **1990** to **2019**.

```
depression_country <- read.csv("./datasets/mental health/number-with-depression-by-country.csv", string
names(depression_country)
```

```
## [1] "Entity"
## [2] "Code"
## [3] "Year"
## [4] "Prevalence...Depressive.disorders...Sex..Both...Age..All.Ages..Number."
```

Once again, we will drop the column *Code*, rename *Entity* to *country*, and rename *Prevalence... Depressive.disorders... Sex..Both* to *people_with_depression*.

Finally we join the two datasets:

```
old_nrows_anxiety_country = nrow(anxiety_country)
anxiety_depression_country <- inner_join(x = anxiety_country, y = depression_country, by=c("country", "
```

```
percentage_data_lost <- (1-(nrow(anxiety_depression_country)/old_nrows_suicide_deaths))*100
percentage_data_lost
```

```
## [1] 0
```

We didn't lose any data!

Life expectancy - dataset

The last dataset that we will need for our analyses is taken from https://ourworldindata.org/life-expectancy and it is about the life expectancy in every country for the last 72 years.

```
life_expectancy <- read.csv("./datasets/life expectancy/life-expectancy.csv", stringsAsFactors = F)
head(life_expectancy)
```

```
##        Entity Code Year Life.expectancy.at.birth..historical.
## 1 Afghanistan  AFG 1950                                  27.7
## 2 Afghanistan  AFG 1951                                  28.0
## 3 Afghanistan  AFG 1952                                  28.4
```

```
## 4 Afghanistan  AFG 1953                                        28.9
## 5 Afghanistan  AFG 1954                                        29.2
## 6 Afghanistan  AFG 1955                                        29.9
```

We are going to drop the column *Code*, change *Entity* to *country* and *Life.expectancy.at.birth..historical.* to *life_expectancy*.

We will also check for any NA values in the dataset.

```
total_NA <- sum(is.na(life_expectancy))
total_NA
```

```
## [1] 0
```

Complete dataset

Given the nature of our analysis, which involves examining data distributed across countries and years, it will be crucial to have information about the total population for each country. This information will enable us to make meaningful comparisons and draw relevant conclusions.

The dataset is taken from https://ourworldindata.org/world-population-growth.

```
population <- read.csv("./datasets/population/population-and-demography.csv", stringsAsFactors = F)
names(population)
```

```
##  [1] "Country.name"
##  [2] "Year"
##  [3] "Population"
##  [4] "Population.of.children.under.the.age.of.1"
##  [5] "Population.of.children.under.the.age.of.5"
##  [6] "Population.of.children.under.the.age.of.15"
##  [7] "Population.under.the.age.of.25"
##  [8] "Population.aged.15.to.64.years"
##  [9] "Population.older.than.15.years"
## [10] "Population.older.than.18.years"
## [11] "Population.at.age.1"
## [12] "Population.aged.1.to.4.years"
## [13] "Population.aged.5.to.9.years"
## [14] "Population.aged.10.to.14.years"
## [15] "Population.aged.15.to.19.years"
## [16] "Population.aged.20.to.29.years"
## [17] "Population.aged.30.to.39.years"
## [18] "Population.aged.40.to.49.years"
## [19] "Population.aged.50.to.59.years"
## [20] "Population.aged.60.to.69.years"
## [21] "Population.aged.70.to.79.years"
## [22] "Population.aged.80.to.89.years"
## [23] "Population.aged.90.to.99.years"
## [24] "Population.older.than.100.years"
```

We will keep only *Country*, *Year* and *Population*.

```
population <- population %>%
    select(Country.name, Year, Population)
names <- c("country", "year", "population")
colnames(population) <- names
```

Finally we merge all the datasets:

```
list_df = list(suicide_deaths,anxiety_depression_country,life_expectancy, population)
complete_dataset <- list_df %>%
  reduce(inner_join, by=c("country", "year"))

head(complete_dataset)
```

```
##       country year age_70 age_50_69 age_15_49 age_5_14 total_death
## 1 Afghanistan 1990     35       167       482       12         696
## 2 Afghanistan 1991     35       168       535       12         750
## 3 Afghanistan 1992     36       171       634       14         855
## 4 Afghanistan 1993     37       176       716       15         944
## 5 Afghanistan 1994     38       180       759       15         992
## 6 Afghanistan 1995     38       183       795       16        1032
##   male_to_female_ratio male_population people_with_anxiety_disorders
## 1                 2.70         5348392                         486787
## 2                 2.73         5372964                         532646
## 3                 2.75         6028498                         609249
## 4                 2.78         7003645                         662644
## 5                 2.79         7733464                         685973
## 6                 2.81         8219472                         709523
##   people_with_depression life_expectancy population
## 1                 439837            46.0   10694804
## 2                 478890            46.7   10745168
## 3                 545449            47.6   12057436
## 4                 593044            51.5   14003764
## 5                 613915            51.5   15455560
## 6                 634196            52.5   16418911
```

Since we already have data on the total population and male population for each country, we can easily calculate and add a new column containing the *female population*.

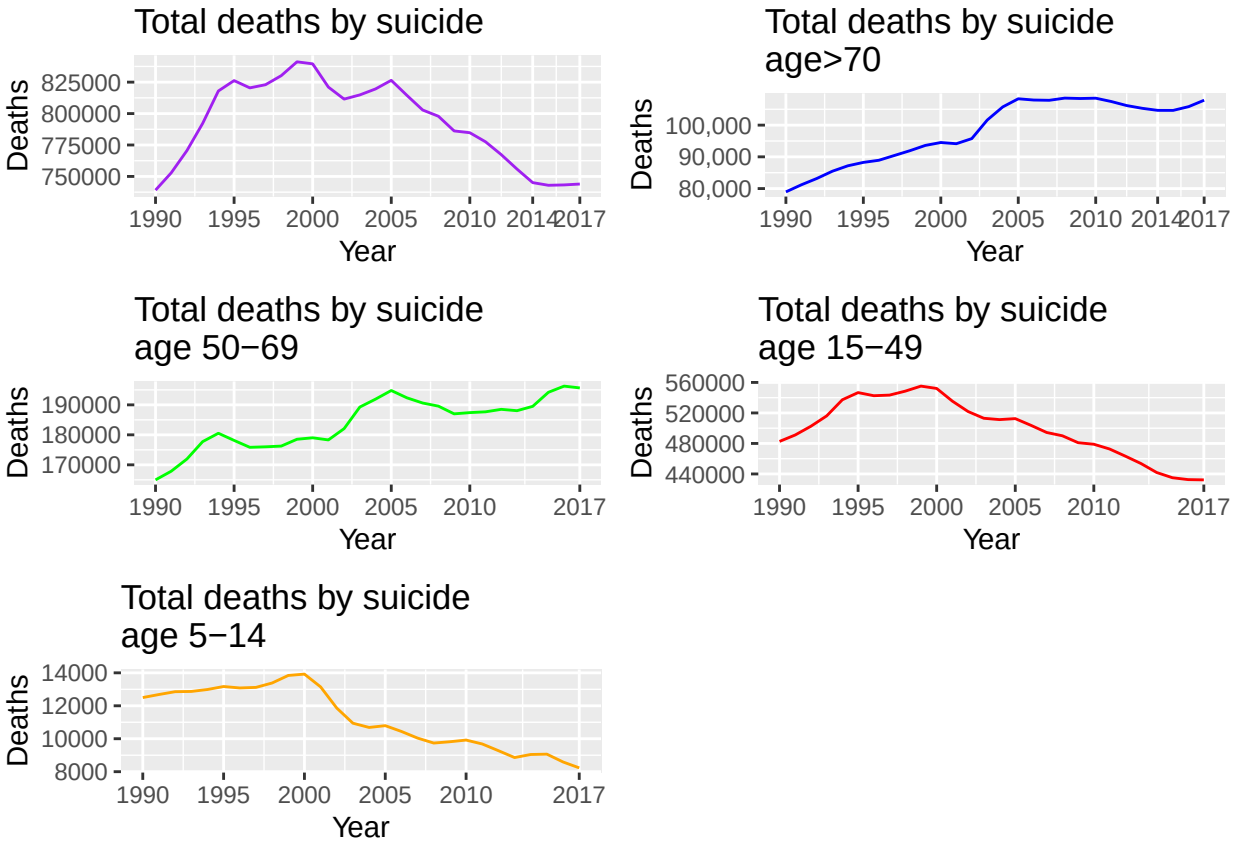
NOTE: in our dataset there is already the data about the "World" with all the attributes for each year from *1990 to 2017*.

Analyses

Analysis by age ranges over the years 1990-2017

We start our analysis by looking at how the total deaths by suicide are distributed by age range over the years.

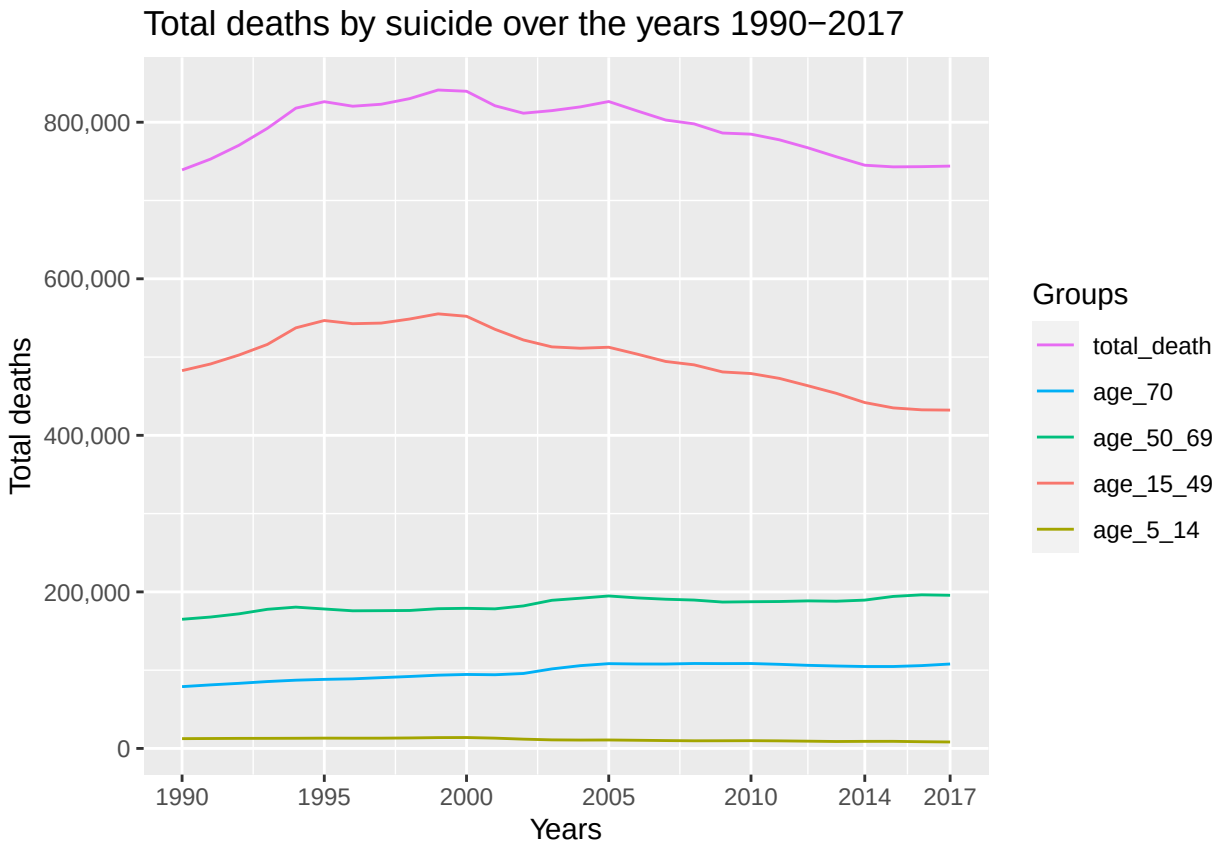
```
grid.arrange(total_death, total_death_over_70, total_death_50_69, total_death_15_49, total_death_5_14, n
```



From this plot, we can observe that the number of deaths by suicide has been consistently decreasing since **2005**. Additionally, it is interesting to note that the number of deaths has been steadily increasing since **1990** for individuals aged **over 50**. However, for individuals between the ages of **5 and 49**, the number of deaths has been decreasing since **2000**.

Now, let's combine all the plots together to get a comprehensive view of the data.

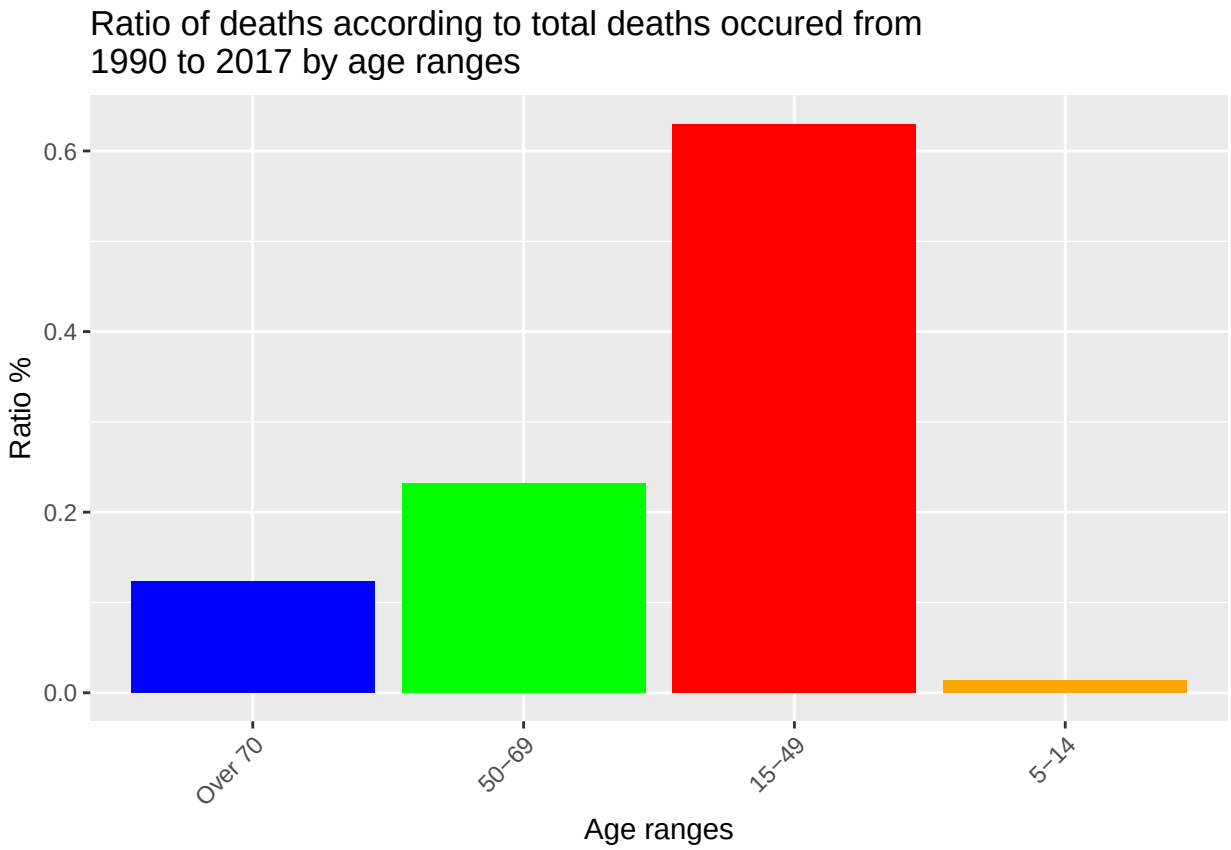
```
print(plot_combined)
```



Based on this plot, we can already observe that the majority of deaths occurred among individuals aged **15-49**.

To gain a better understanding of the relative contribution of each age range to the total deaths, we will calculate the sum of deaths for each age range from 1990 to 2017. We will then calculate the ratios of these sums to the total deaths that occurred during the same period. Finally, we will create a bar chart using the **geom_bar(...)** function to visualize these ratios.

```
print(plot_ratio)
```

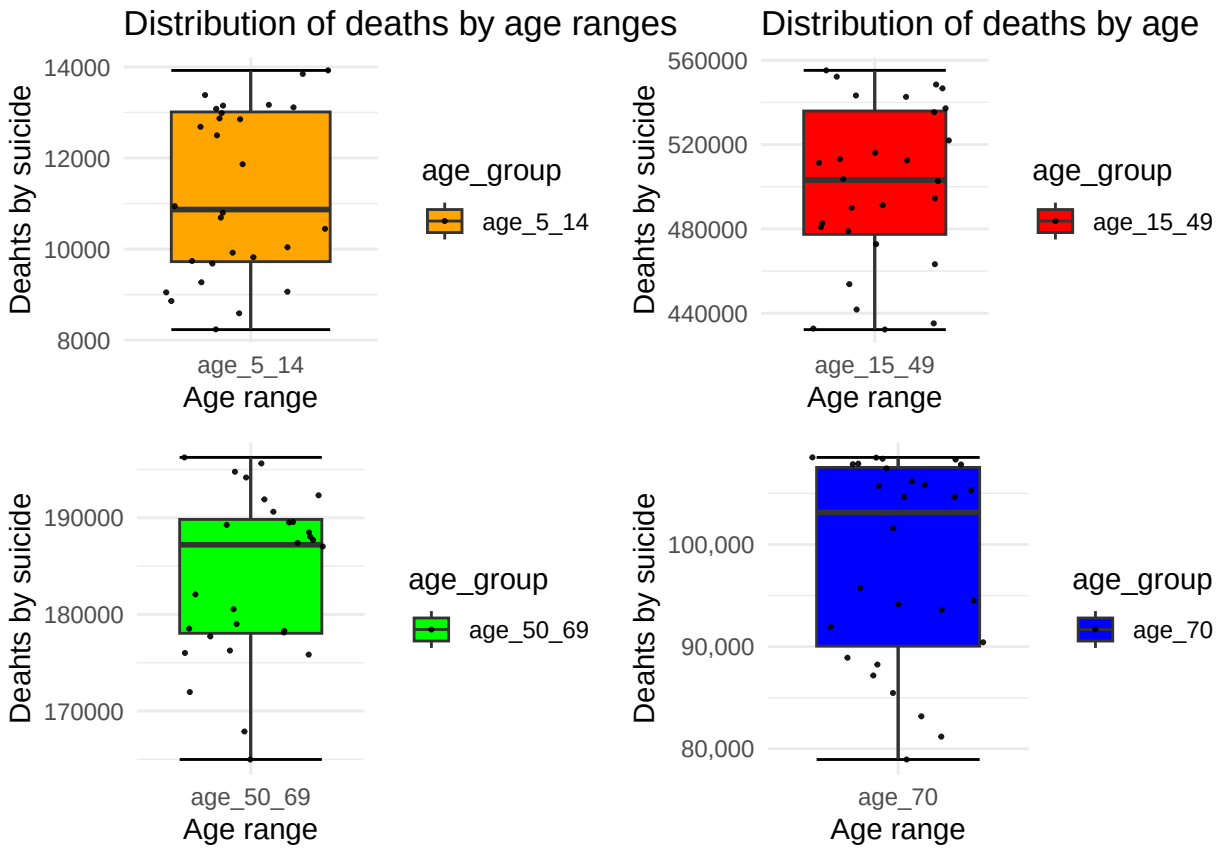


As expected from the previous plots, more than **60%** of the total deaths that occurred from 1990 to 2017 are attributed to the age range **15-49**. On the other hand, less than 3% of the total deaths fall within the age range **5-14**.

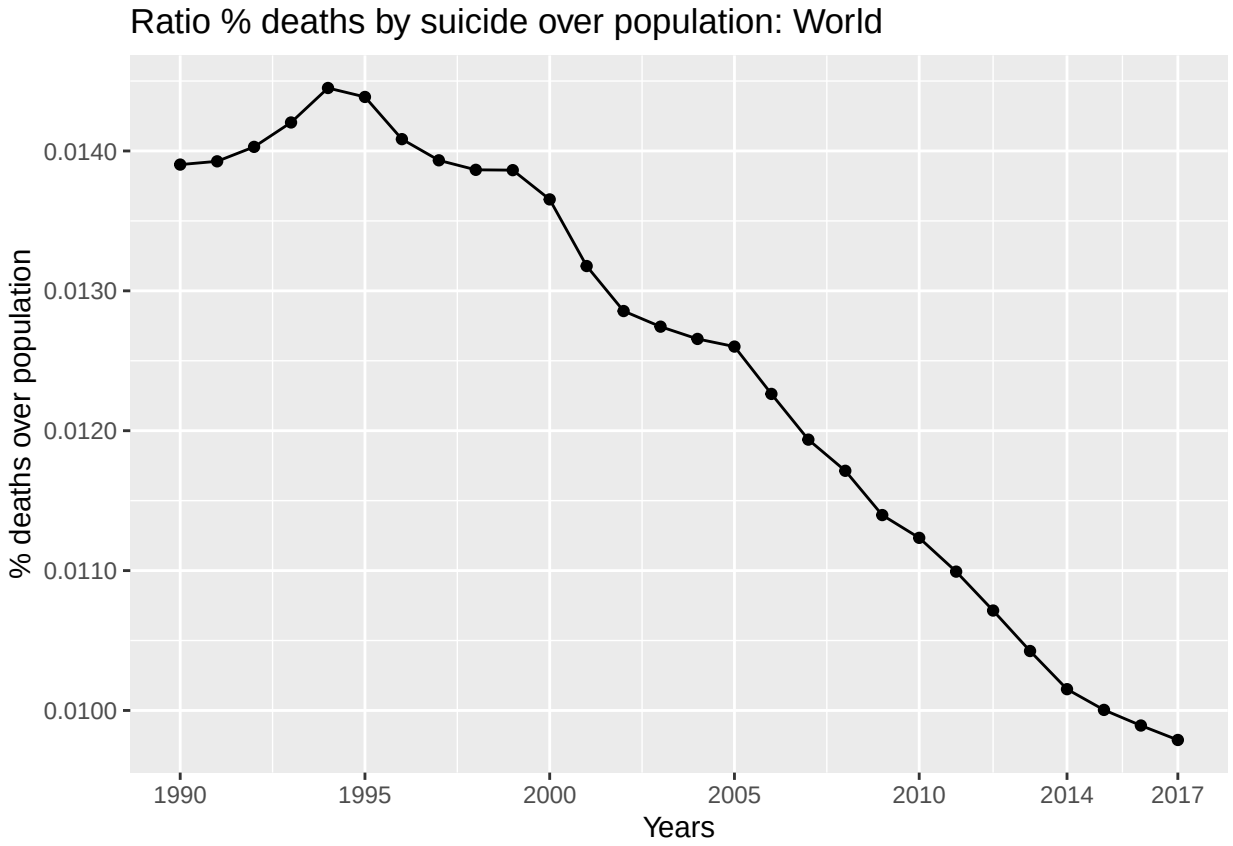
It would be interesting to have a more specific breakdown of age groups within the 15-49 range, such as distinguishing between young individuals (under ~30 years old) and adults (between ~30 and 49 years old). However, unfortunately, we do not have this more detailed breakdown available in our dataset.

Additionally, we can use a **box plot** to visualize how the values of total deaths by suicide per year are distributed across the different age ranges.

```
grid.arrange(boxplot_5_14, boxplot_15_49, boxplot_50_69, boxplot_70, ncol = 2)
```




```
print(plot_combined)
```

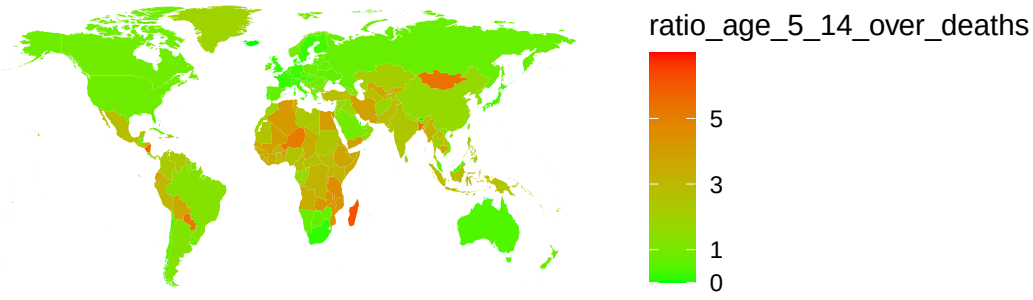


From the first plot, we observed that the total number of deaths by suicide in recent years has returned to a level similar to that of 1990. However, when we examine the last plot, we can see that the ratio of deaths by suicide to the world population has been steadily decreasing since 1995.

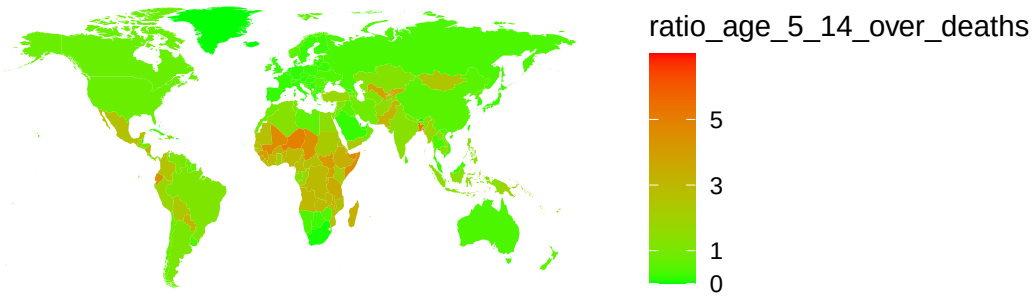
In the next section, we will plot the ratios of total deaths by suicide for specific age ranges relative to the total deaths by suicide for each country in the years 1990 and 2017 on a map.

```
grid.arrange(world_ratio_5_14_to_total_1990, world_ratio_5_14_to_total_2017, ncol = 1)
```

Ratio % deaths by suicide age 5 to 14 over total death by suicide of the country 1990

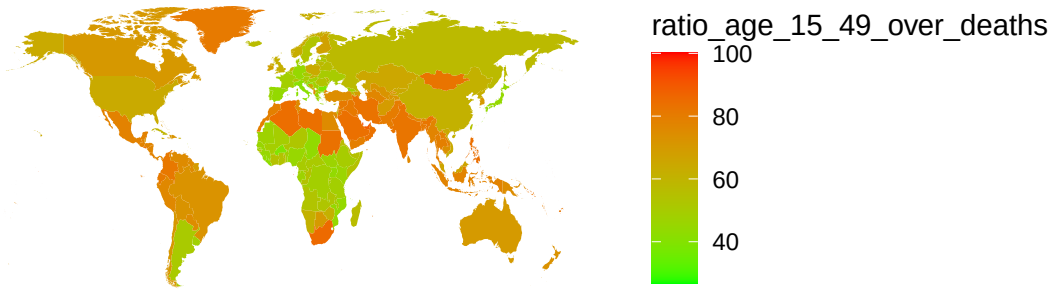


Ratio % deaths by suicide age 5 to 14 over total death by suicide of the country 2017

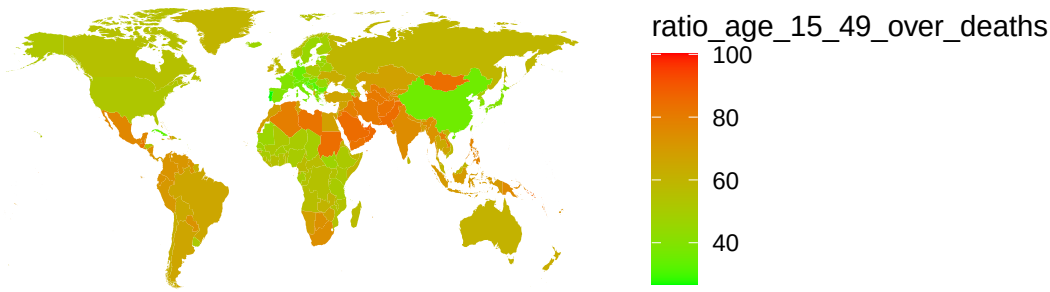



```
grid.arrange(world_ratio_15_49_to_total_1990, world_ratio_15_49_to_total_2017, ncol = 1)
```

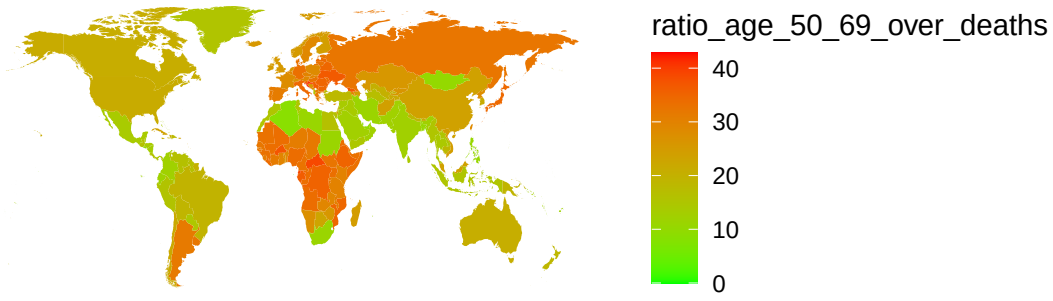
Ratio % deaths by suicide age 15 to 49 over total death by suicide of the country 1990

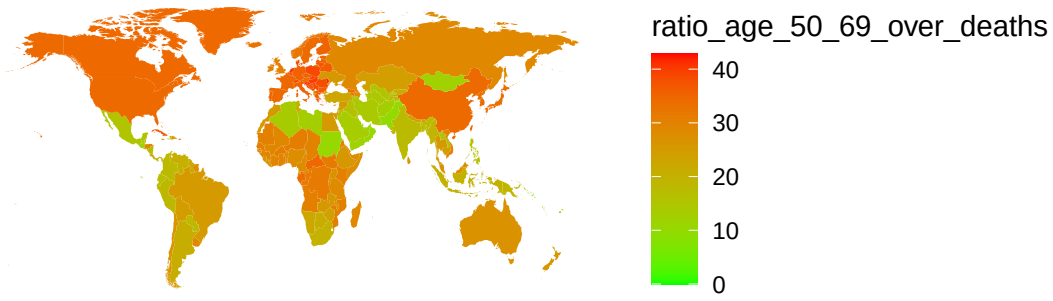


Ratio % deaths by suicide age 15 to 49 over total death by suicide of the country 2017



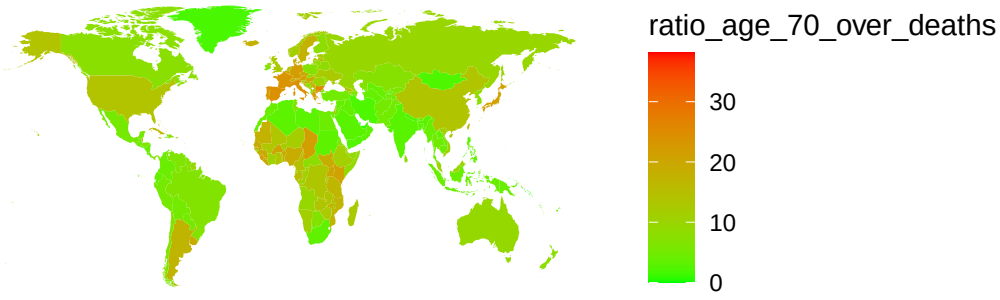
```
grid.arrange(world_ratio_50_69_to_total_1990, world_ratio_50_69_to_total_2017, ncol = 1)
```



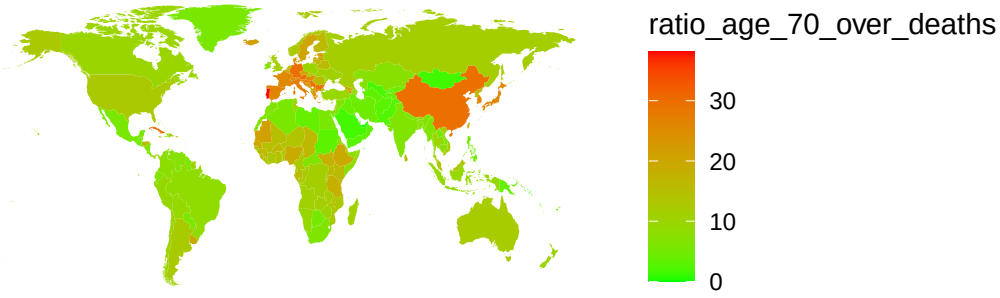



```
grid.arrange(world_ratio__70_to_total_1990, world_ratio__70_to_total_2017, ncol = 1)
```

Ratio % deaths by suicide age > 70 over total death by suicide of the country: 1990



Ratio % deaths by suicide age > 70 over total death by suicide of the country: 2017



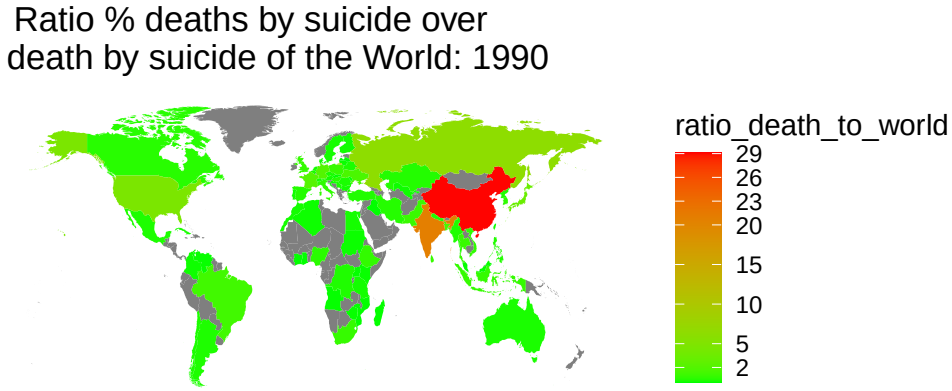
- **Age range between 5 and 14:** we can see that the ratio got lower for most of the countries, the countries with the highest value are in Africa. We can notice that there are no countries with ratio equal to *100%*.
- **Age range between 15 and 49:** the deaths of individuals in the 15-49 age range are prevalent in every country, with only a few countries having a minimum ratio of around **30%**. Arab countries did not show significant differences between 1990 and 2017. Countries in central Europe, China, USA, and Canada experienced a notable decrease in the ratio.
- **Age range between 50 and 69:** this time the Arab countries had a "better performance" compared to the rest of the World in terms of the ratio for this age range. However, central Europe, China, USA, Canada and Greenland had a notable increase (up to *20%*!).
- **Age over 70:** most countries did not show significant changes in the ratios for this age range, except for central Europe and China, which experienced a notable increase in the ratio.

In conclusion we can say that for the age range *15-49* and *50-69* the ratios are similar for most of the countries, while for the age range *5-14* and *over 70* the ratio is higher for few and specific countries.

Analysis of distribution of death by suicide over the World

Now we are going to analyse which are the countries where the ratio *total deaths / population of the world* is high, we will compare the situation in 1990 with 2017.

```
grid.arrange(world_ratio_death_to_total_1990, world_ratio_death_to_total_2017, ncol = 1)
```



Note: the countries colored with grey have a ratio $<$**0.1%**.

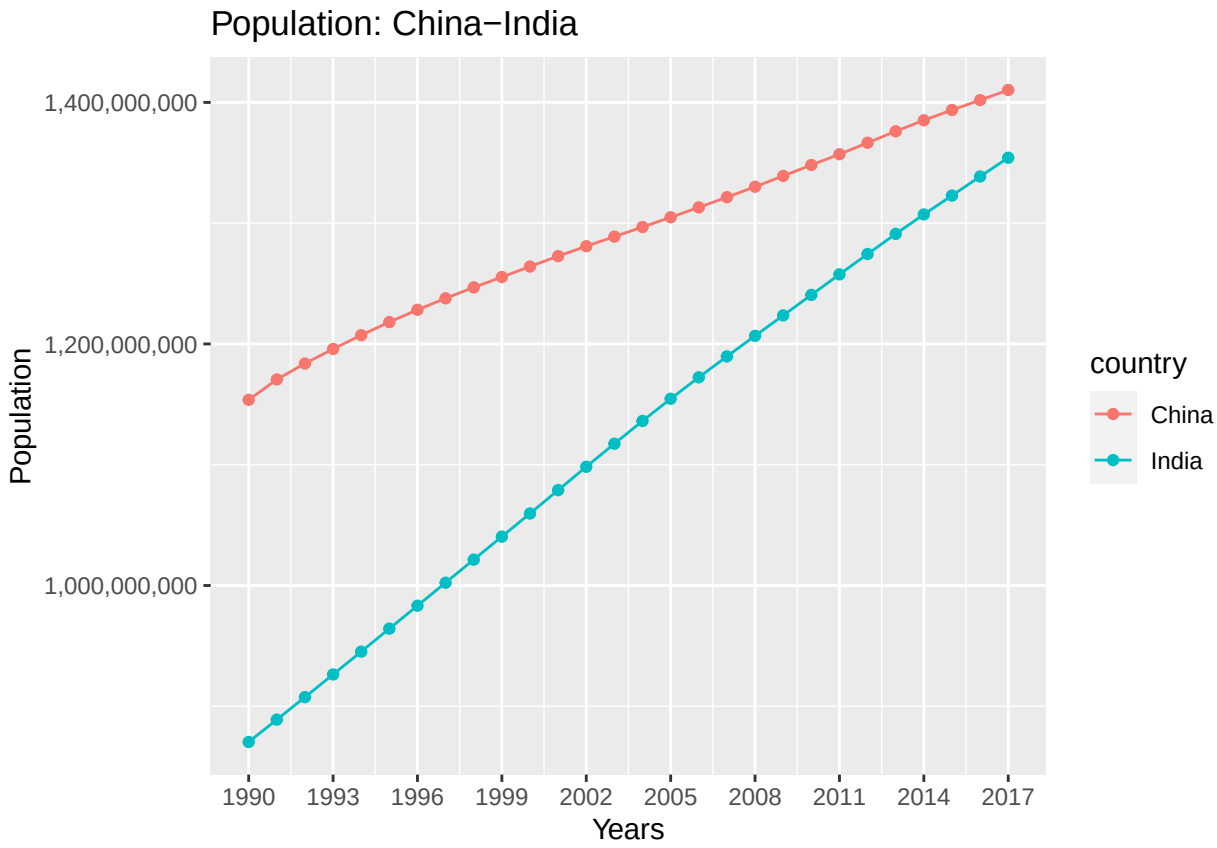
As expected, countries with high populations such as **USA, Russia, China, and India** tend to have higher ratios of deaths by suicide. A larger population size can contribute to higher absolute numbers of suicide deaths.

It is interesting to note that while most countries did not experience significant changes in the ratio over the years, **India and China** show a different trend. This suggests that there may be specific factors or dynamics influencing suicide rates in these countries that deviate from the overall pattern.

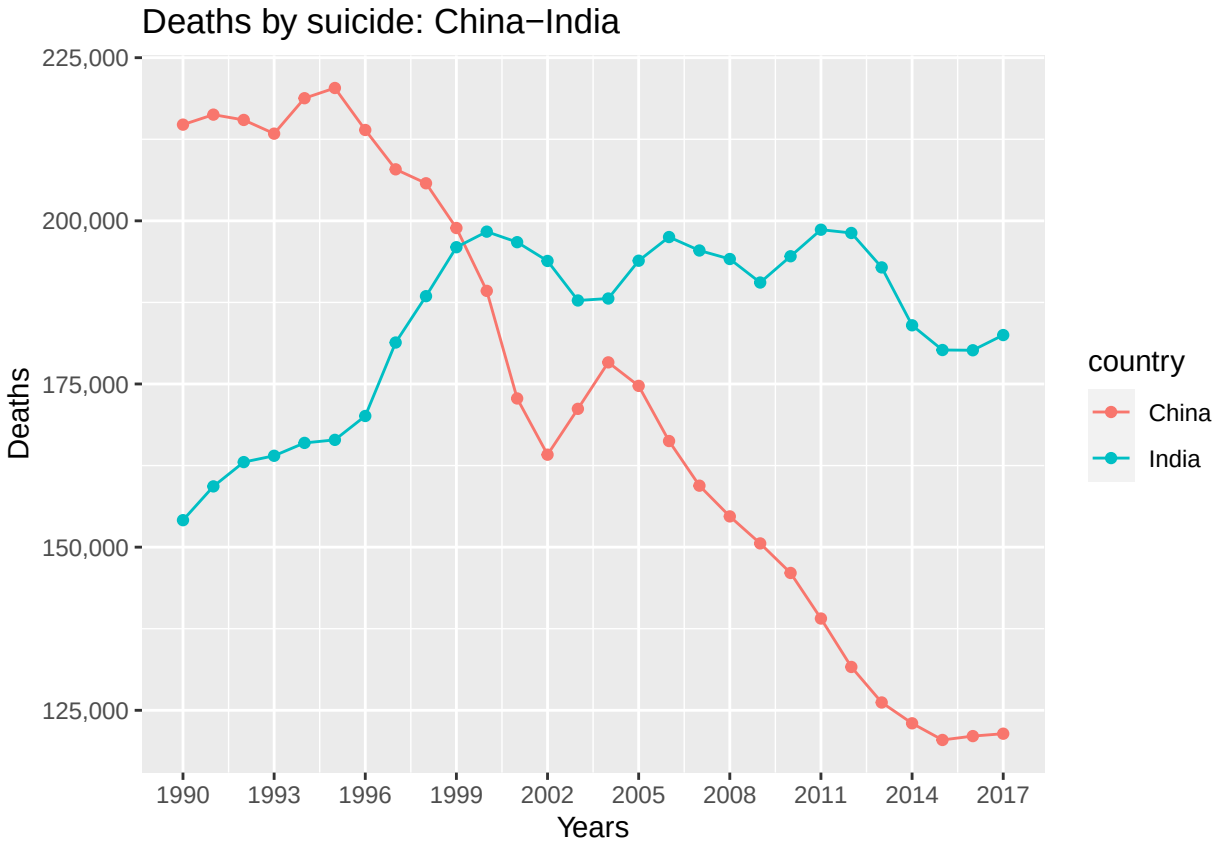
Furthermore, we can notice that for some central Africa countries they had an increase of the ratio from less than **0.1% to ~2%**!

For the next plots, we will further investigate and provide a valid interpretation for the unexpected trend observed in India and China.

```
print(plot_population_china_india)
```




```
print(plot_total_death_china_india)
```

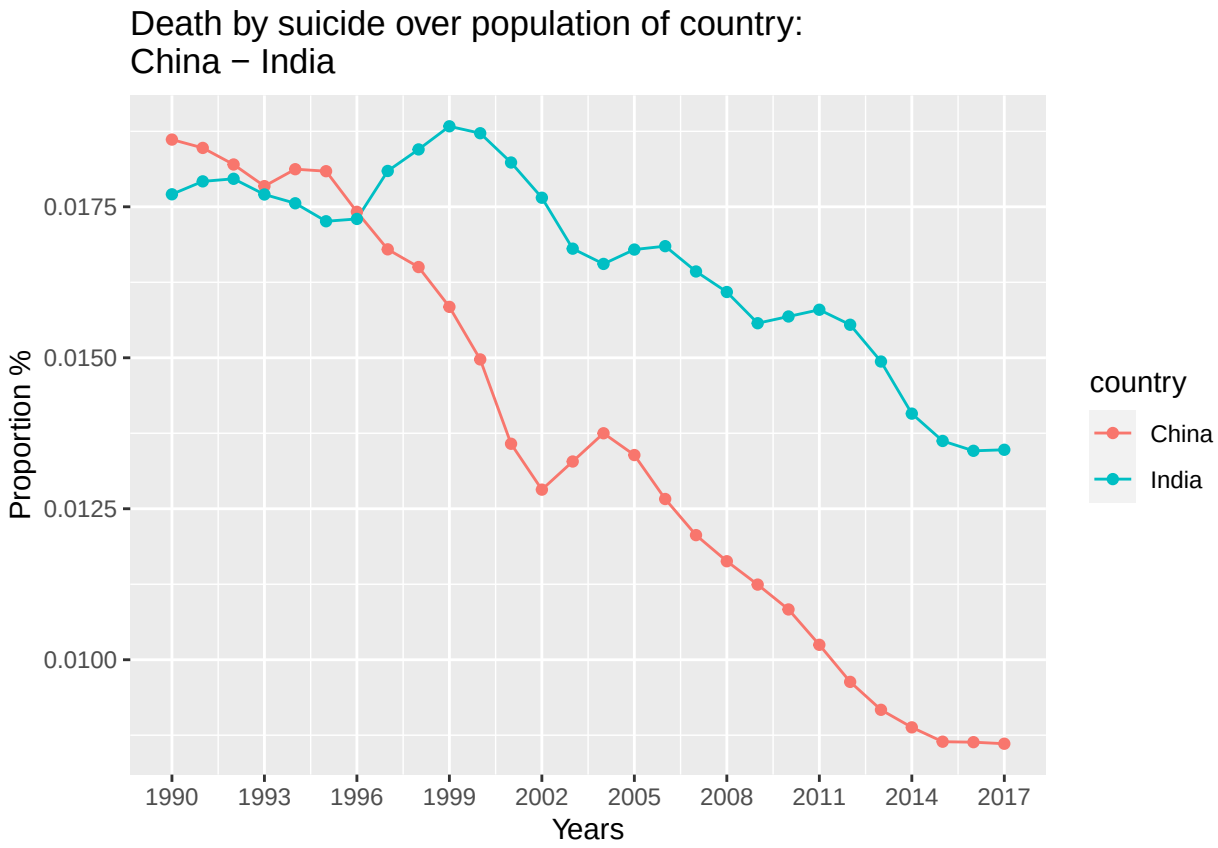


From the plot, we can observe that the populations of **China and India** have been steadily increasing over the years. However, the trends in deaths by suicide differ between the two countries.

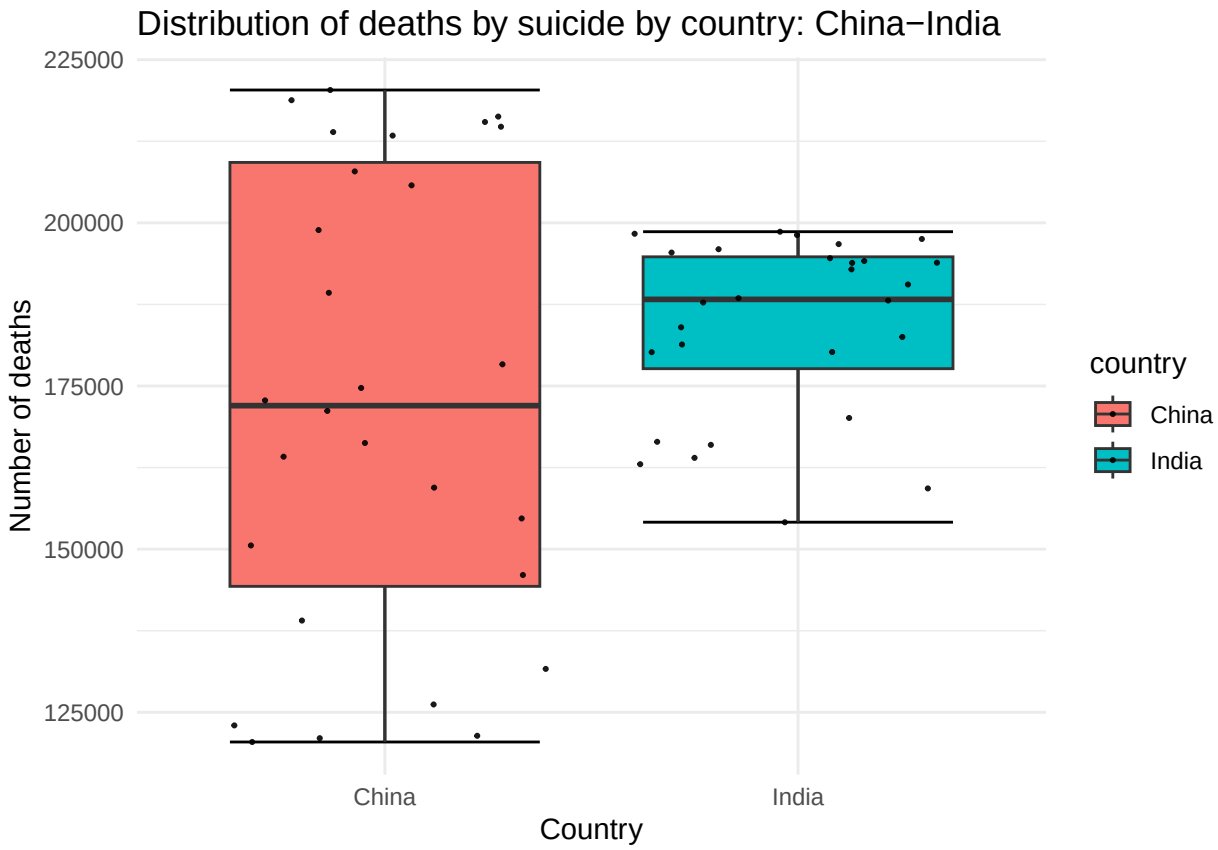
From this plot we can see that the populations of China and India are constantly increasing, the deaths by suicide in India increased during **1990-2000** and has since remained relatively constant. In China, the number of deaths by suicide has been decreasing since **1995**, with a slight increase observed during **2002-2004**. Since **2014**, the number of deaths has remained relatively constant. This indicates a significant decline in suicide deaths over time in China, even though the population continues to grow.

To gain a better understanding of the trends, it would be helpful to examine the ratio of total deaths to the population of each country. This can be achieved by calculating the **ratio (total deaths / population of the country) x 100**. Additionally, plotting a boxplot will allow us to visualize the distribution of these ratios and identify any variations or outliers.

```
print(plot_ratio_china_india)
```

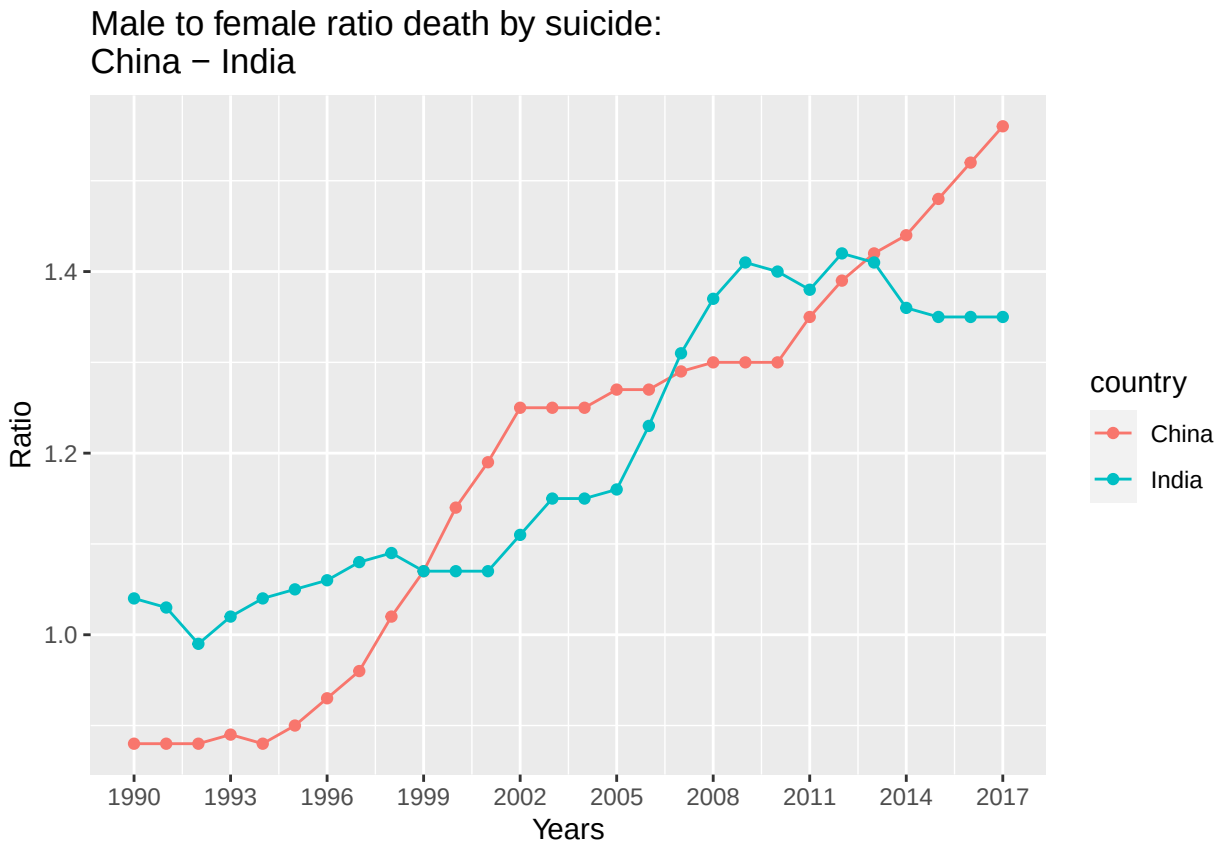



```
print(boxplot_china_india)
```

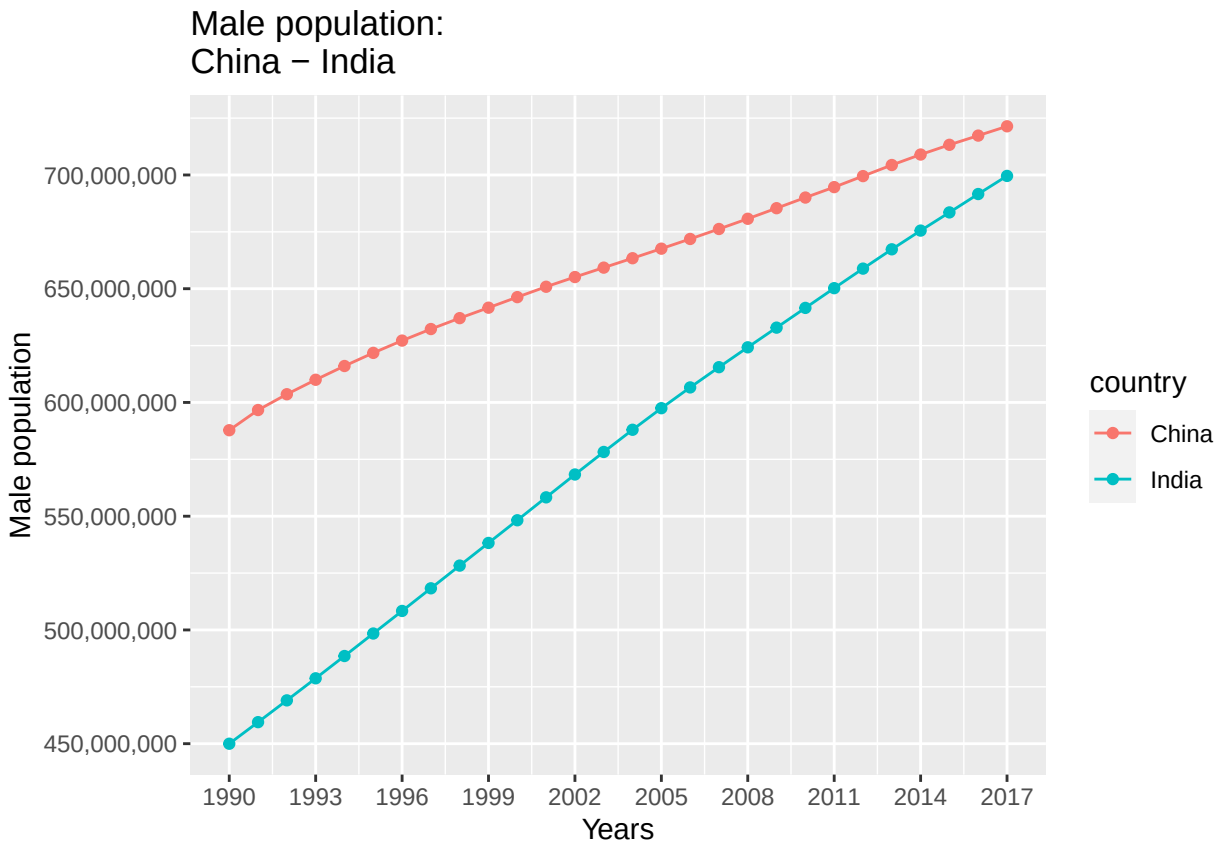


As we expected from the previous plots, we can see that China had a more significant decrease compared to India. It would be interesting to see how the **ratio of male to female** deaths by suicide changed over the years. A value *>1* means that there were more deaths of male compared to female. We also plot the trends of male population and female population over the years.

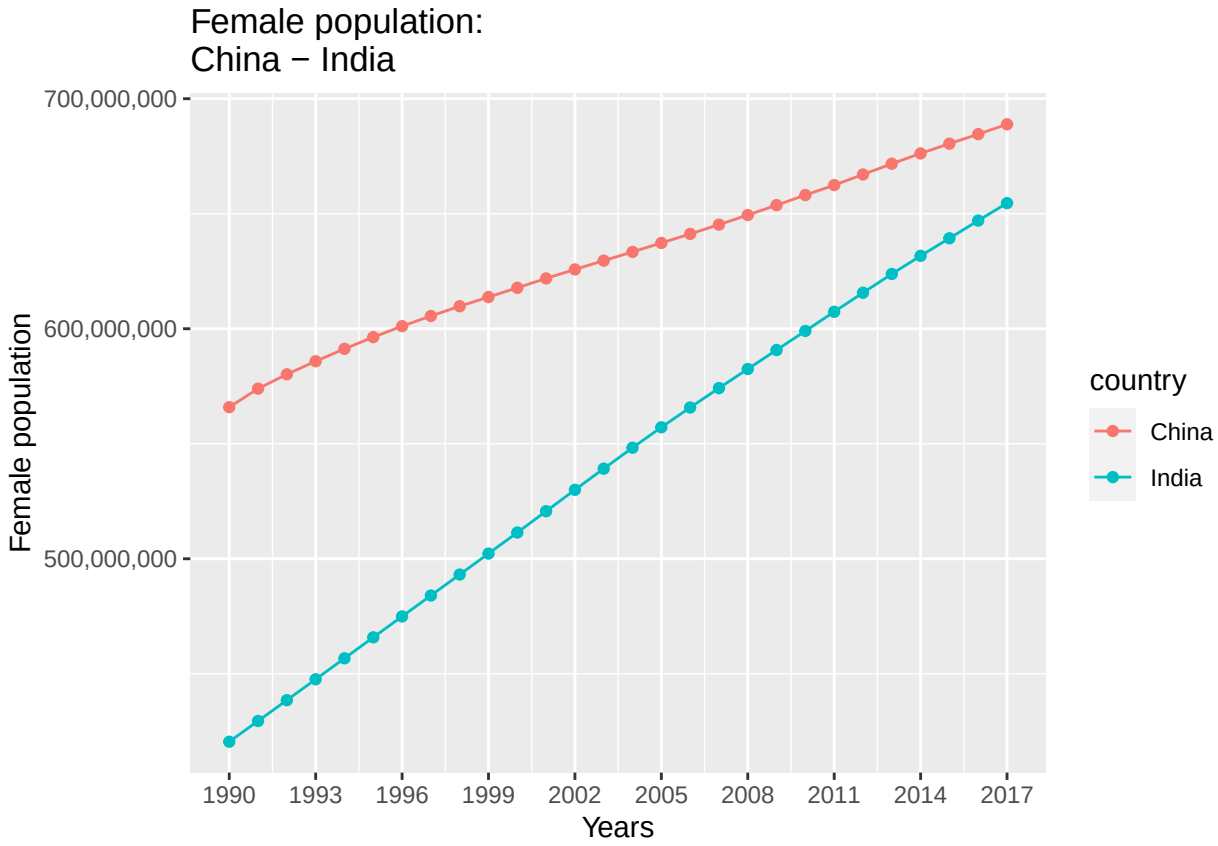
```
print(plot_male_to_female_ratio_china_india)
```




```
print(plot_male_population_china_india)
```




```
print(plot_female_population_china_india)
```



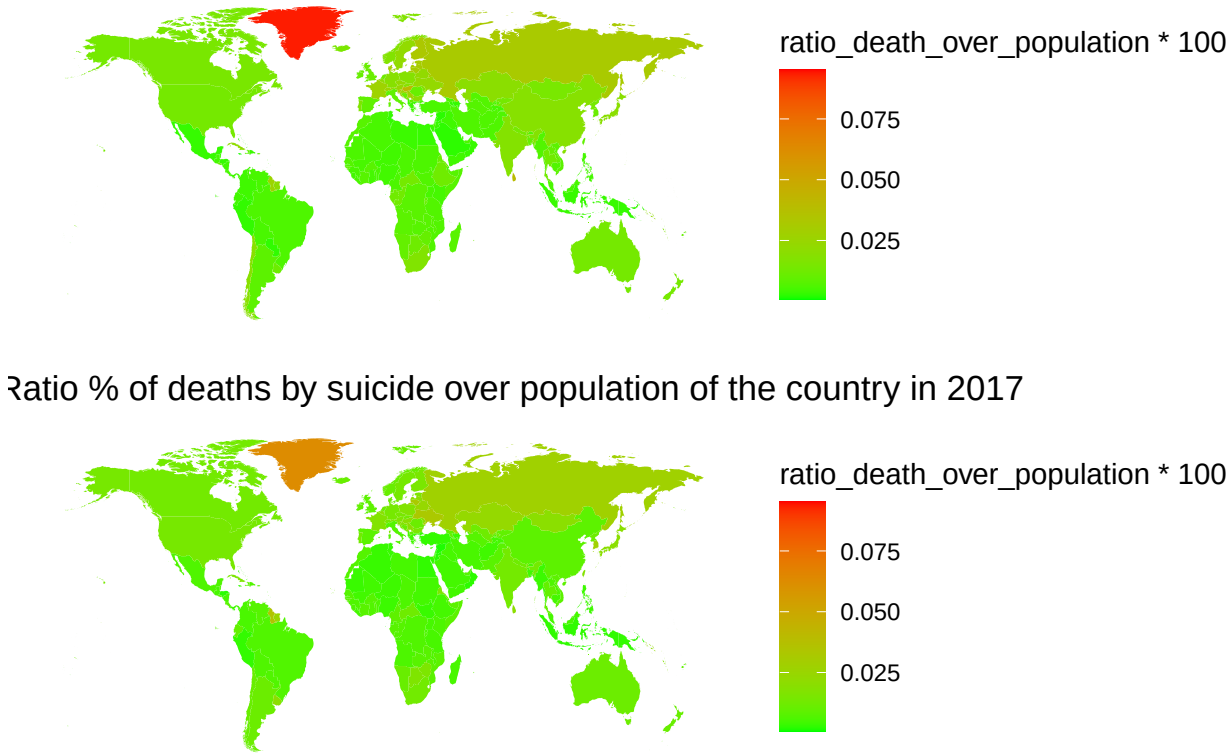
In **India**, the **ratio of male to female** deaths by suicide has consistently been greater than 1, indicating a higher prevalence of suicide deaths among males. This pattern has remained relatively stable since **2008**, with only slight variations.

In **China**, the **ratio of male to female** deaths by suicide has shown a different trend. Until **1997**, the ratio was generally below 1, indicating a higher prevalence of suicide deaths among females. However, since then, the ratio has steadily increased and crossed the threshold of 1, indicating a higher prevalence among males. This change suggests that China has experienced a *significant decrease in suicide deaths among females, despite the growing population for both males and females.*

For both the countries, the male and female population *always* kept increasing.

Now we are going to plot the ratio *(deaths by suicide / population of country) x 100* for every country of the World, we will compare the situation in 1990 with the one in 2017.

```
grid.arrange(world_ratio_death_to_country_1990, world_ratio_death_to_country_2017, ncol = 1)
```

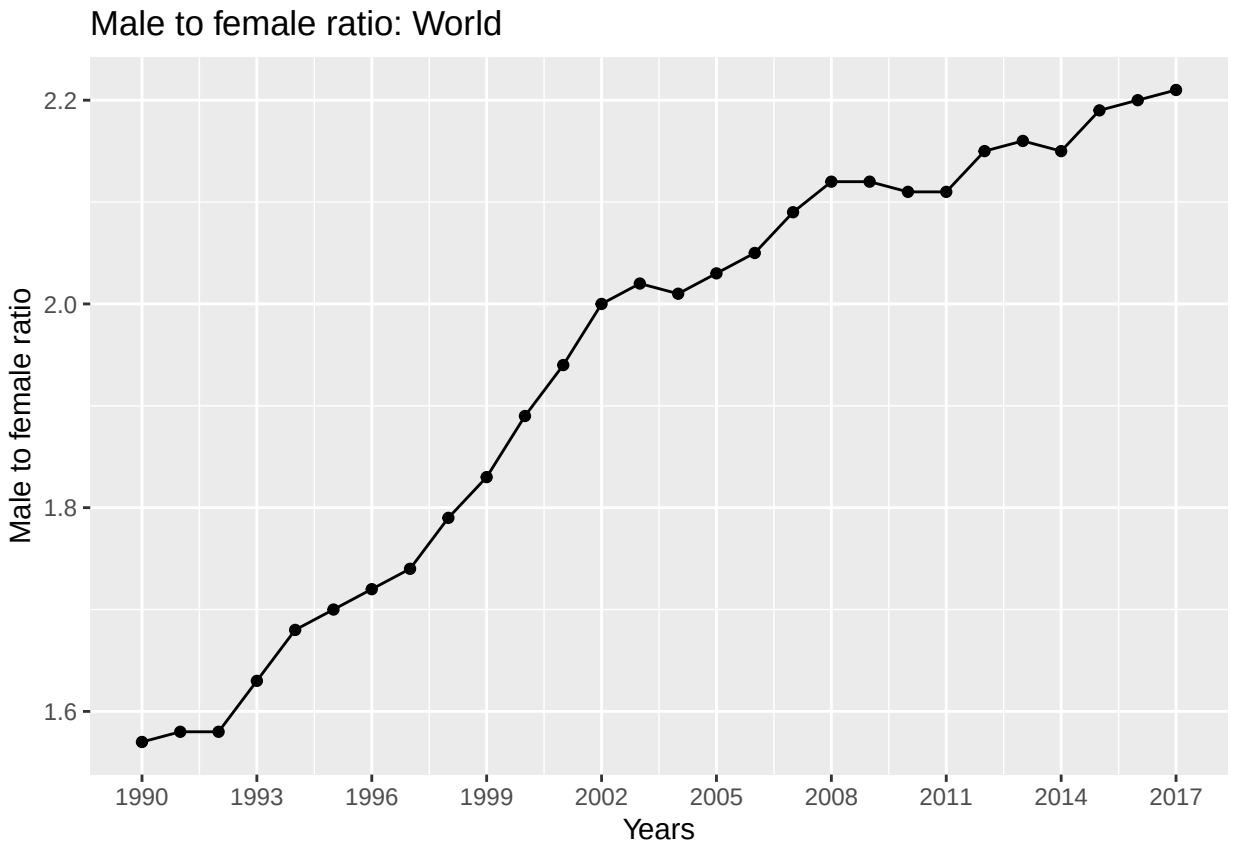


Indeed, the situation seems to be relatively stable for most countries in terms of the ratio deaths by suicide over the population of the country. However, **Greenland** (and China, but we already analyzed it) stands out as an exception, with a notable difference in the ratio compared to other countries. Greenland's unique characteristics, such as its small population size (less than *56 000*!) may contribute to this distinct pattern.

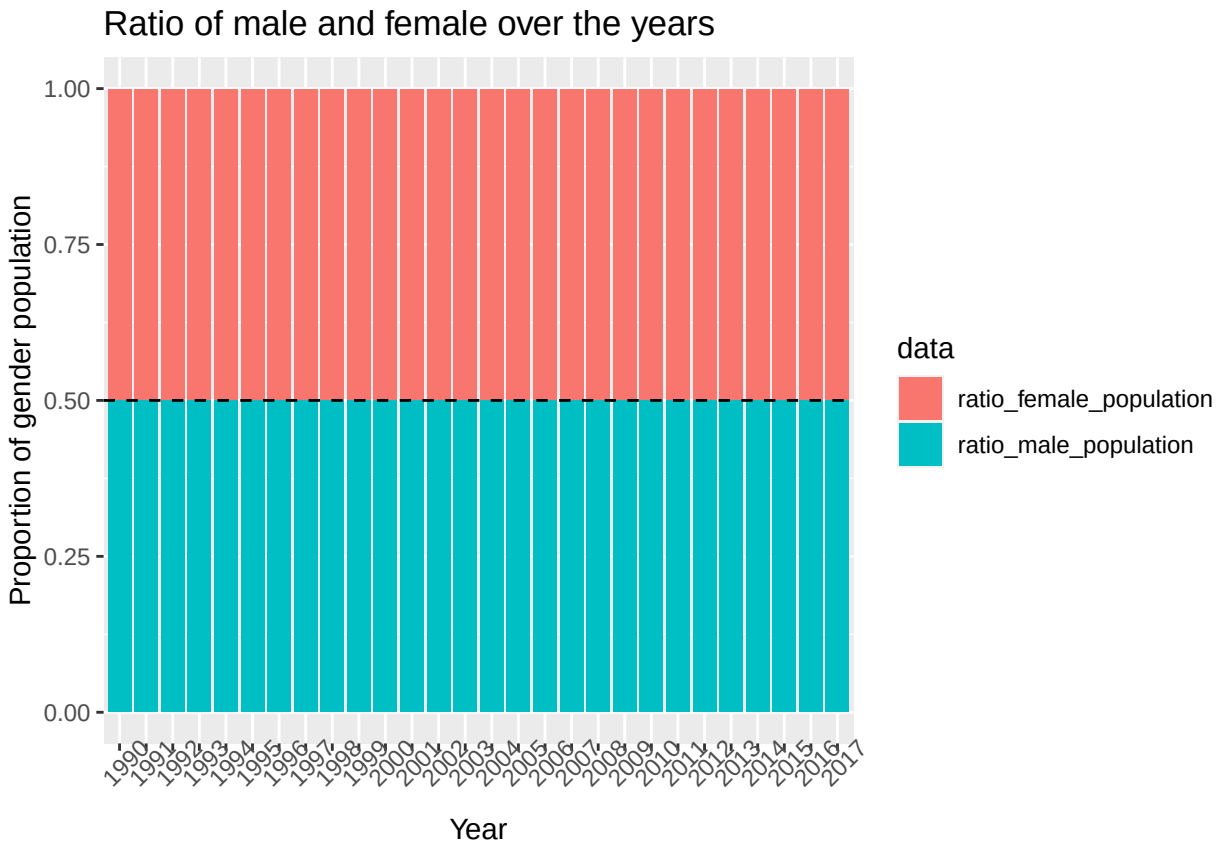
Analysis of death by suicide: male to female ratio

In this section we analyse the male to female ratio, its trend over time and the situation of every country.

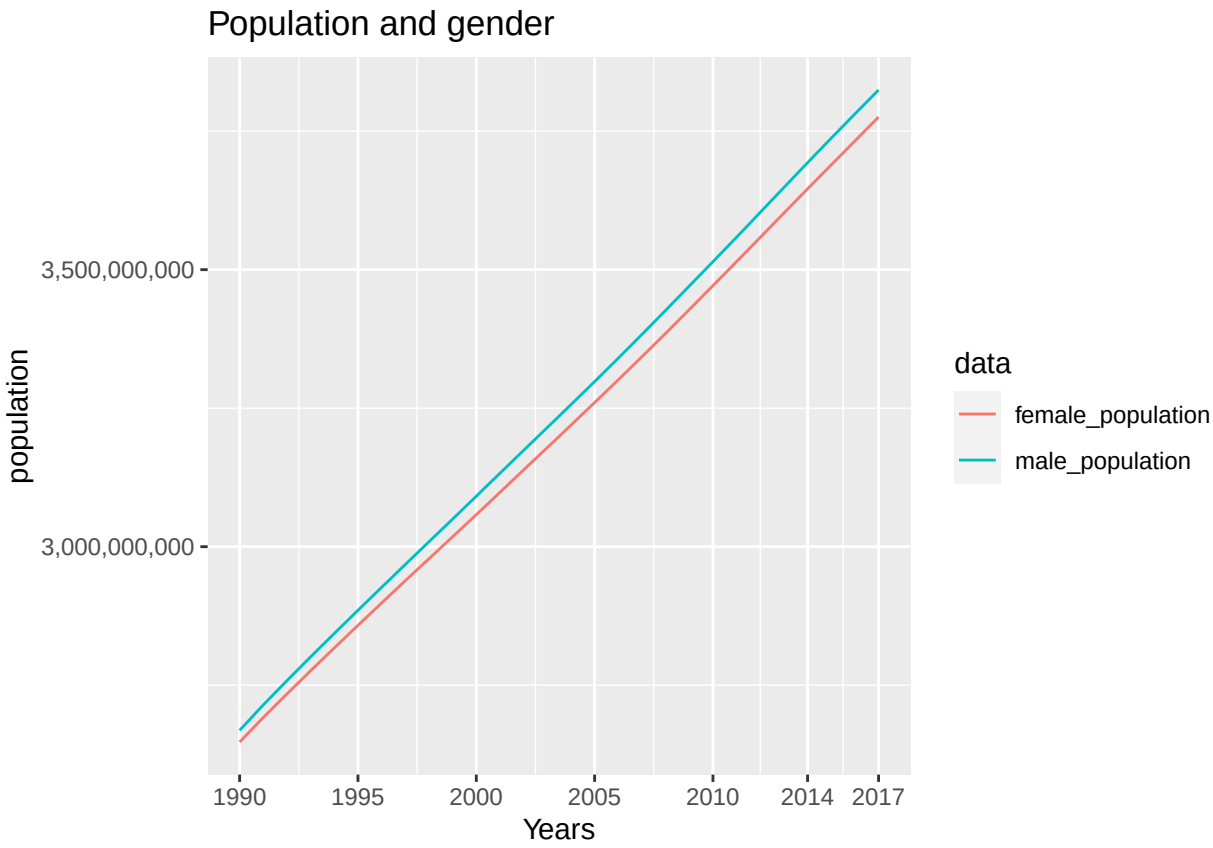
```
print(male_to_female_ratio)
```




```
print(plot_bar)
```



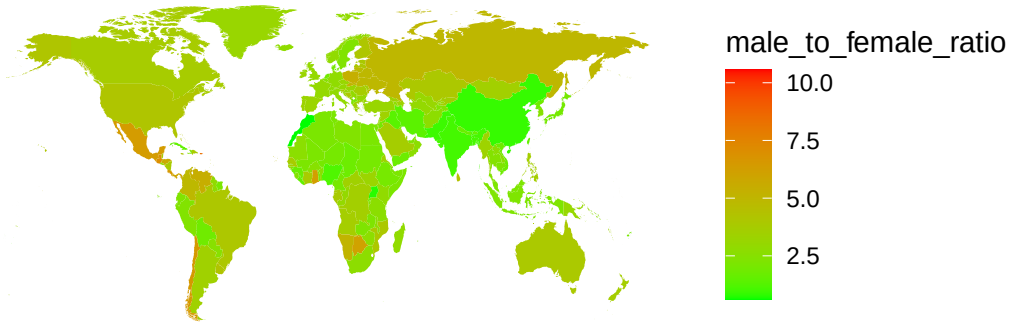

```
print(plot_population_gender)
```



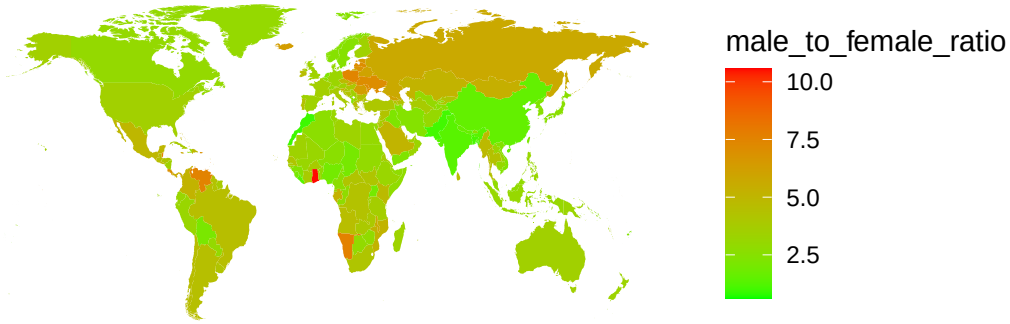
From the first plot we can see that since **1990** the deaths by suicide are higher in the male population compared to female population, and the ratio kept constantly increasing. From the last two plots we see that the proportion of male population is slightly equal to female population, with a proportion equal to *~49%-51%* over the years. Now we plot the male to female ratio in all the countries in 1990 and 2017.

```
grid.arrange(world_ratio_male_to_female_1990, world_ratio_male_to_female_2017, ncol = 1)
```

Ratio male to female death by suicide: World 1990



Ratio male to female death by suicide: World 2017

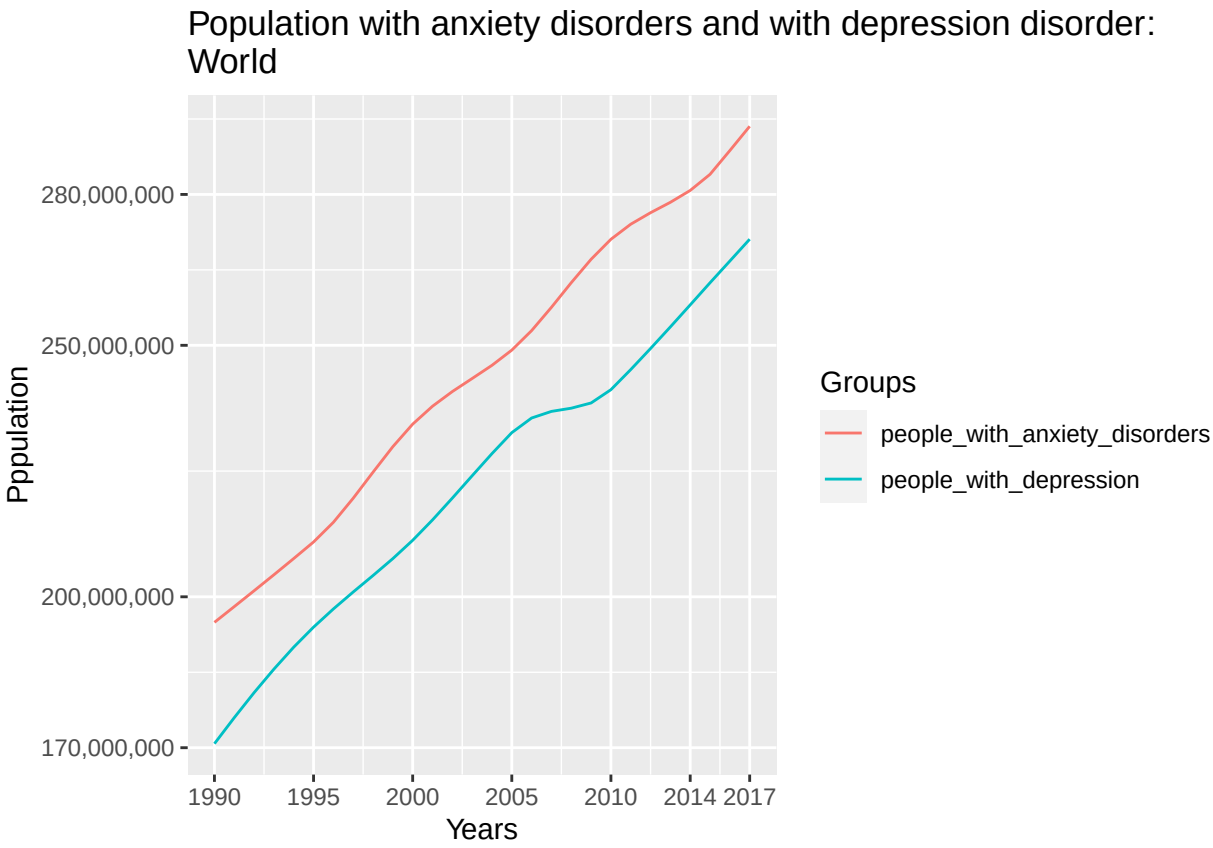


It is indeed interesting to note that the *ratio of male to female* deaths by suicide has either increased or remained unchanged for all the countries. There is no significant decrease!

Analysis of people with anxiety disorders and people with depression

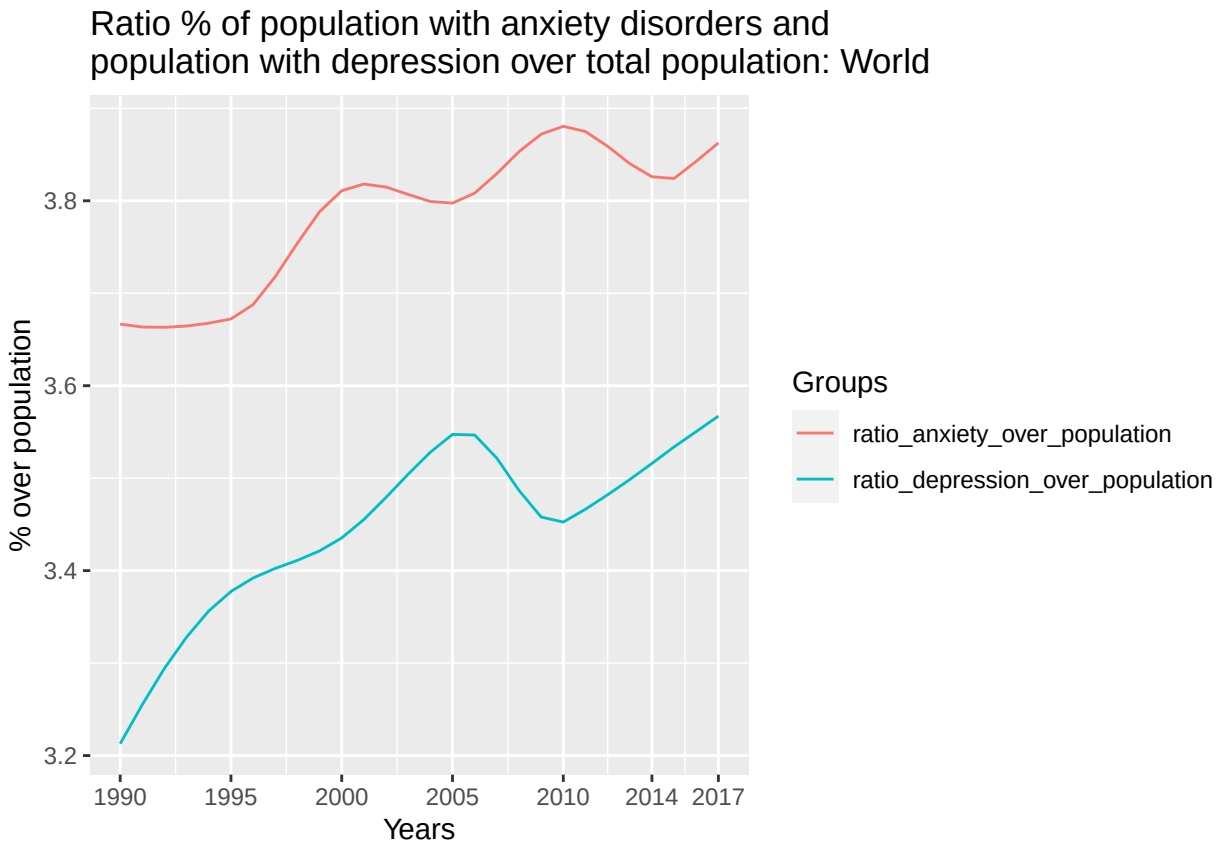
In this section, we will analyze the trends in the **population with anxiety disorders and depression disorders** over the years. We will plot the data to visualize any changes or patterns.

```
print(plot_combined)
```



From the plot we can see that there were always more people with anxiety disorders compared to people with depression. Both kept increasing over the years with the same trend. Let's see how the **ratio over the total population** changed over the year.

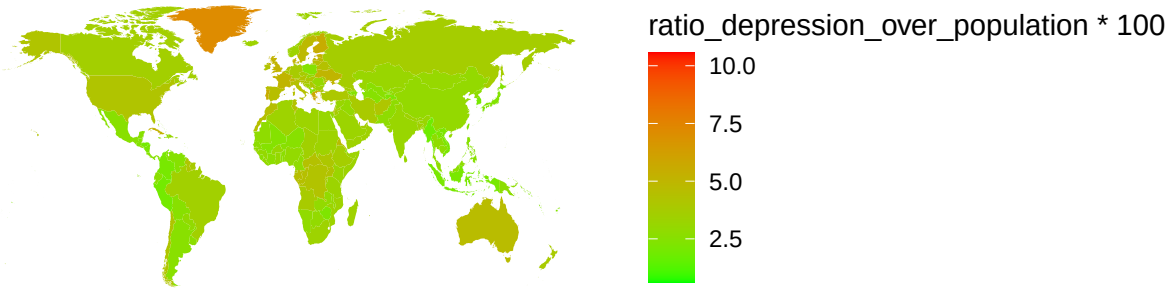
```
print(plot_combined)
```



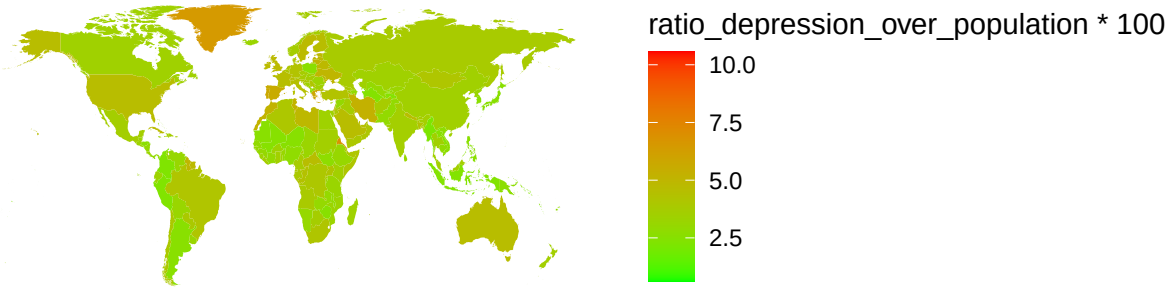
With this plot we can see that even if the ratios over the total population kept increasing, both had some "drops" but not in the same years! Let's see now how are the **ratios in the countries** in 1990 and 2017.

```
grid.arrange(world_ratio_depression_to_country_1990, world_ratio_depression_to_country_2017, ncol = 1)
```

Ratio % people with depression over population of the country:
World 1990

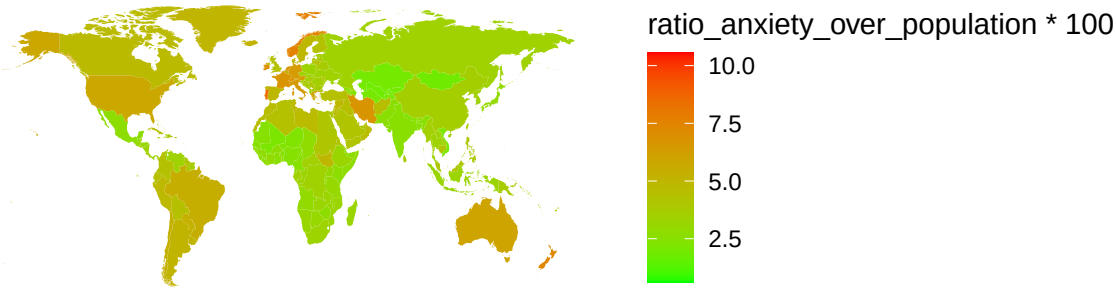


Ratio % people with depression over population of the country:
World 2017

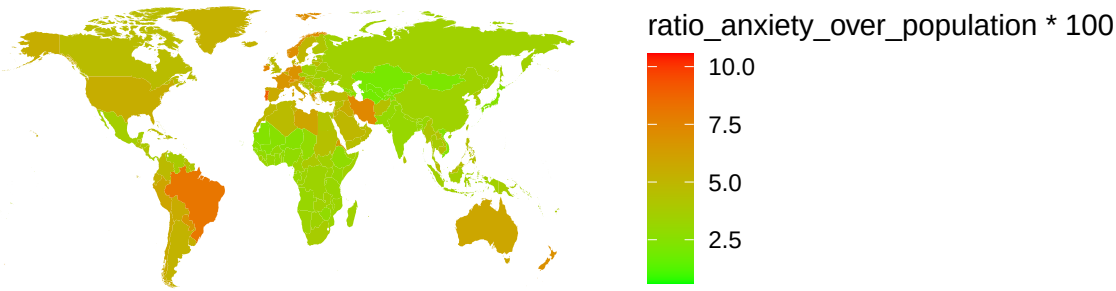



```
grid.arrange(world_ratio_anxiety_to_country_1990, world_ratio_anxiety_to_country_2017, ncol = 1)
```

Ratio % people with anxiety disorders over population of the country:
World 1990



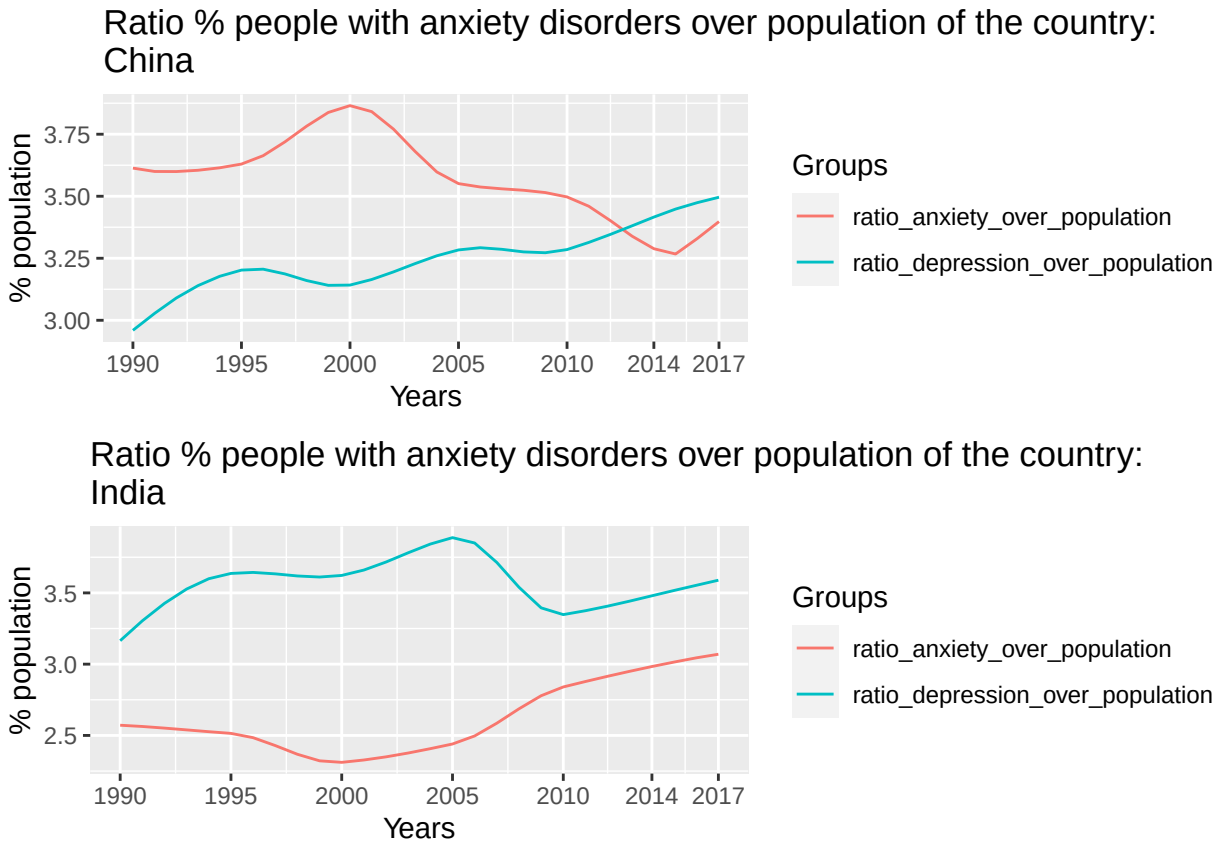
Ratio % people with anxiety disorders over population of the country:
World 2017



We observe that the prevalence of depression disorders remained relatively stable across countries, with no significant variations except for a worsening trend in **Eritrea**. Similarly, the prevalence of anxiety disorders showed little change overall, except for notable increases in **Brazil and Libya**.

Now, let's examine the specific trends in **China and India** regarding anxiety and depression disorders.

```
grid.arrange(plot_ratio_anxiety_and_depression_china, plot_ratio_anxiety_and_depression_india, ncol = 1)
```

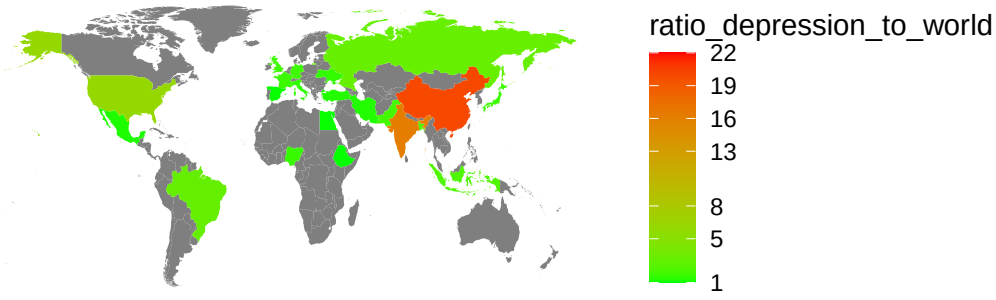


We can observe that the trends in the two countries, China and India, are contrasting. In **China**, there has consistently been a higher prevalence of people with anxiety disorders, except for the recent years. On the other hand, in **India**, there has consistently been a higher prevalence of people with depression.

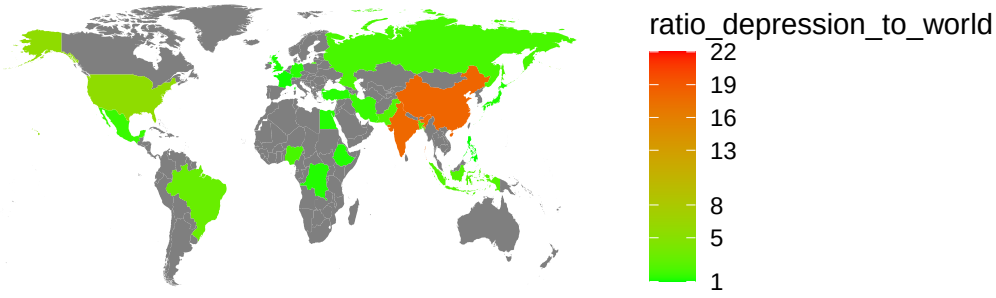
Now, let's explore which countries contribute the most to the total number of people with anxiety and depression disorders. We will examine the data for the years 1990 and 2017.

```
grid.arrange(world_ratio_depression_to_total_1990, world_ratio_depression_to_total_2017, ncol = 1)
```

Ratio % people with depression in a country over
people with depression of the World: 1990

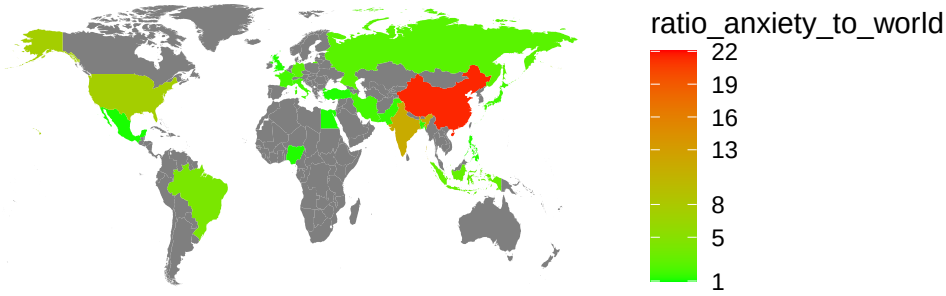


Ratio % people with depression in a country over
people with depression of the World: 2017

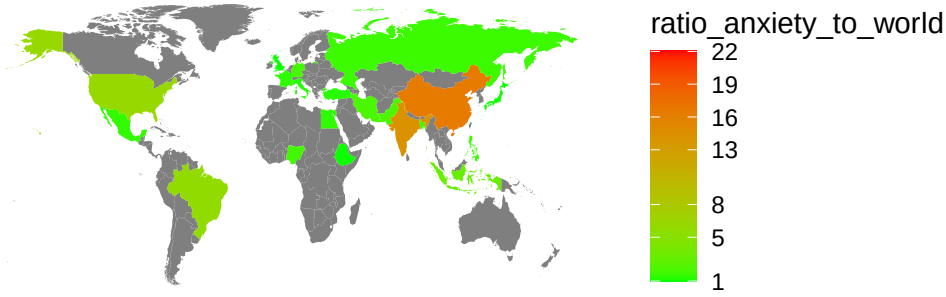



```
grid.arrange(world_ratio_anxiety_to_total_1990, world_ratio_anxiety_to_total_2017, ncol = 1)
```

Ratio % people with anxiety disorders in a country over
people with anxiety disorders of the World: 1990



Ratio % people with anxiety disorders in a country over
people with anxiety disorders of the World: 2017



Note: the countries colored with grey have a ratio **<1%**.

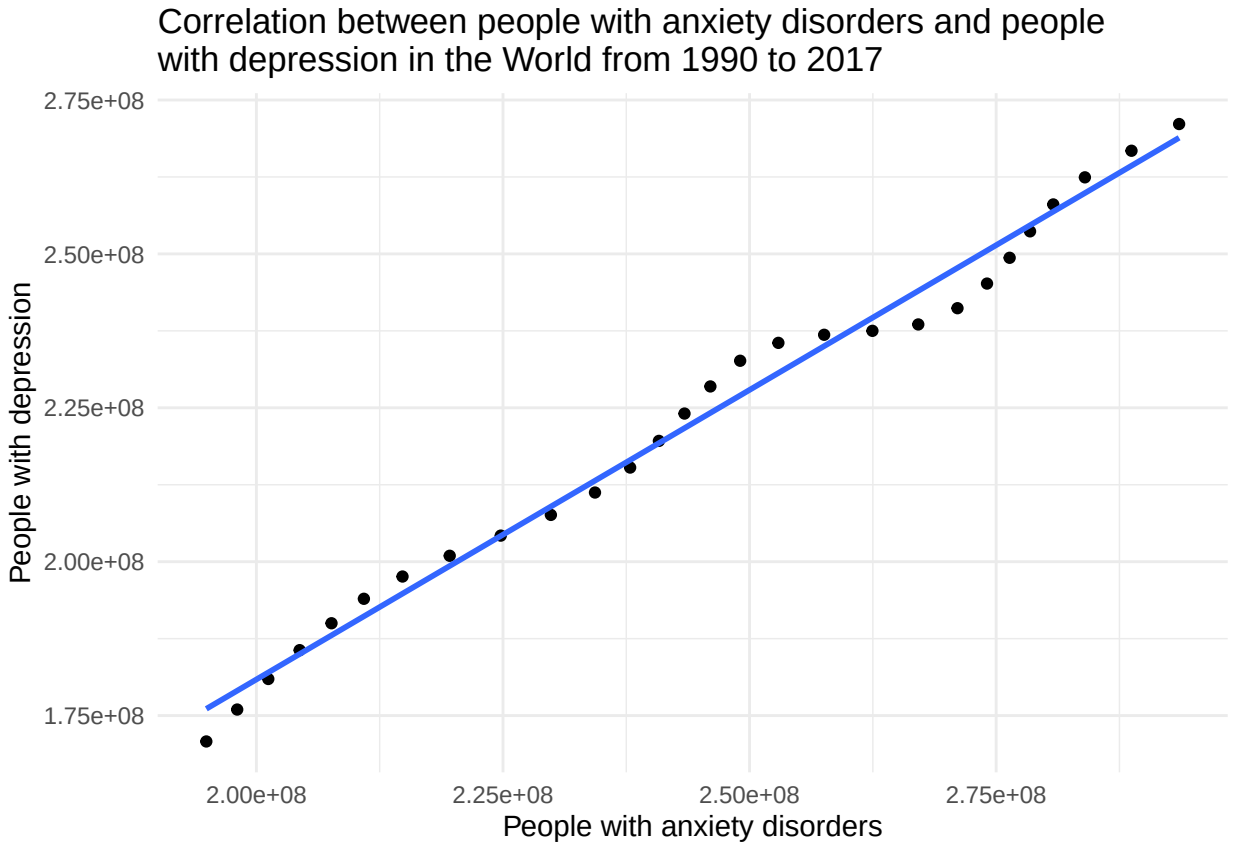
From the previous analysis, we can observe that the countries contributing the most to the total number of people with anxiety disorders and people with depression are the same. Specifically, **China and India** stand out with significant changes between **1990 and 2017**, being the only countries with notable differences.

Additionally, we can see that there are fewer countries in **Africa** with a substantial number of countries contributing at least **1%** compared to other continents. This could be attributed to the lack of reporting by the population in those regions.

Moving forward, we will compute the correlation coefficient to measure the strength of the linear relationship between the variables "**people with anxiety disorders**" and "**people with depression**" using aggregated data for the "World" from 1990 to 2017. Furthermore, we will create a **scatterplot** to visualize this relationship.

```
print(scatter_plot)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
print(paste("Correlation: ", correlation))
```

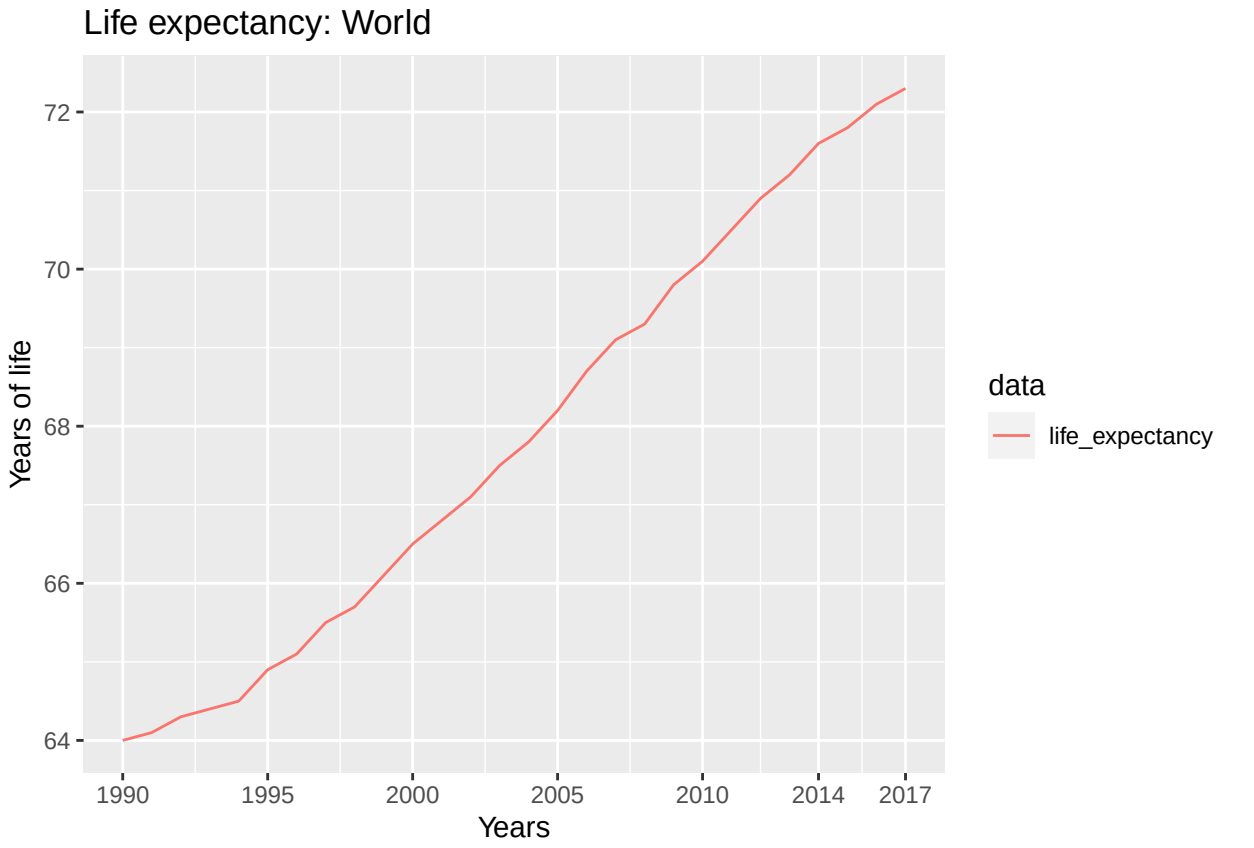
```
## [1] "Correlation:  0.993520500538508"
```

The value of the correlation is ~**1**! This means that if the value of *people with anxiety disorders* increase, the value of *people with depression* will also increase. We should remember that the correlation doesn't give any explicit information about *cause-effect* relation! It just gives the linear relationship between these two variables.

Analysis of life expectancy

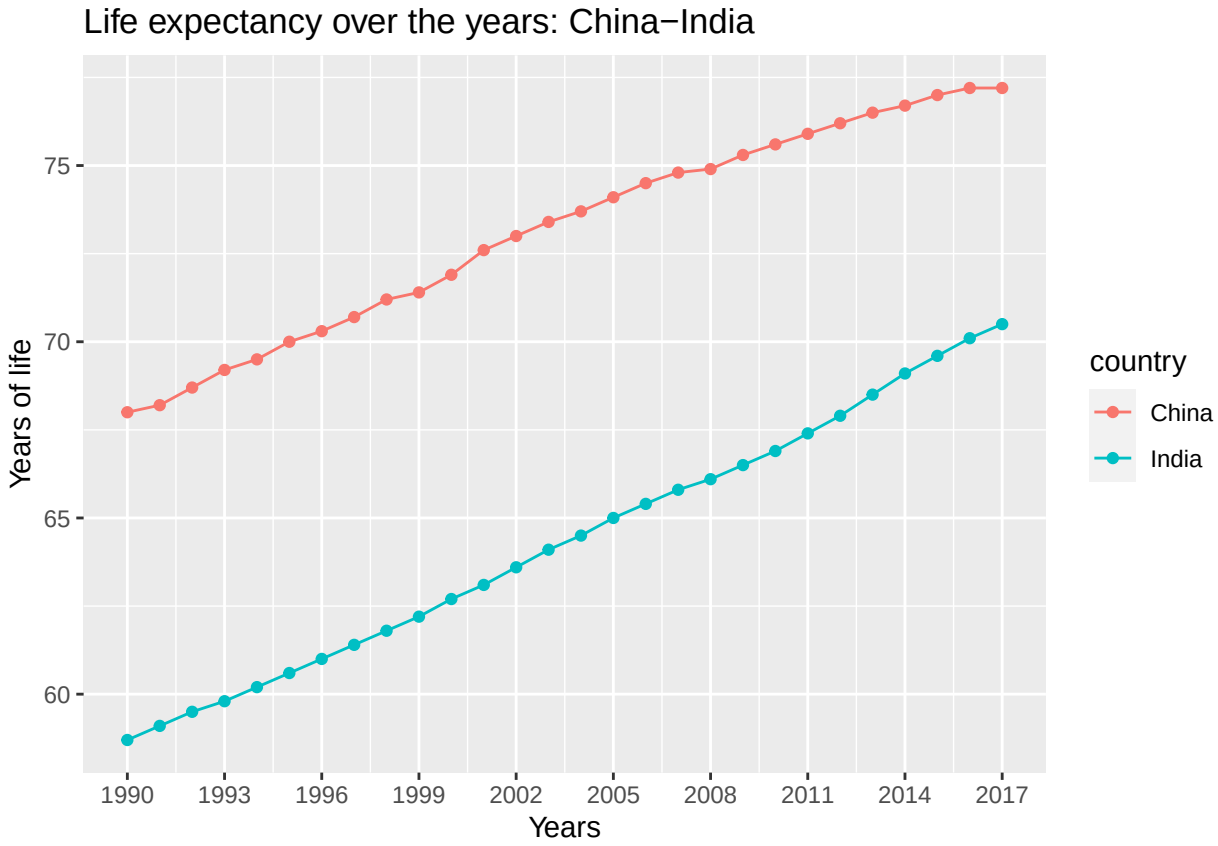
Let's see how the life expectancy for the World changed over time.

```
print(plot_combined)
```



As we see from the plot the life expectancy always kept increasing since 1990, starting from the value **64** (years old) to **72**. For China and India the trend is shown in the next plot:

```
print(plot_life_expectancy_china_india)
```



We can observe that **China** consistently had a higher life expectancy compared to **India** and the rest of the **World**!

Conclusions

Through this analysis, we have discovered that total suicide deaths have returned to the levels of 1990, indicating a continuous decrease in the suicide rate relative to the total population. However, we have found that suicide deaths are decreasing among individuals under the age of 49, while they are increasing among those aged 50 and above. Nevertheless, deaths among individuals aged 15 to 49 still represent a significant portion of the total suicide deaths.

Furthermore, we have observed a notable decline in suicide deaths in China since 1995, with a decrease of approximately 100,000 individuals. This decrease is primarily driven by a decline in female suicide deaths. We conducted a comparison between China and India, two countries with large populations, and discovered contrasting trends over the years in terms of suicide deaths, prevalence of anxiety disorders, and depression.

Additionally, we have found that globally, the population is roughly divided between males and females (approximately 49-51% each). Despite this gender balance, there are approximately twice as many male suicide deaths compared to female suicide deaths.

If we consider the data collectively, for the entire world, we find a strong correlation between the number of people with anxiety disorders and the number of people with depression. It is important to note that high rates of anxiety or depression among the population of a country do not necessarily correlate with high suicide rates, with the exceptions being Greenland and Russia.