Data Visualization Final Project

Xinran (Grace) Zhang

Github Link

https://github.com/xinran-zhang/suicides_rate_data_viz

Dataset Description - Introduction

- What is this dataset?

  The dataset I chose is the Suicides Rates Overview from 1985 to 2016.
- Where did you get it from?

  I got the datasets from Kaggle.

  https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016
- Why did you choose this particular data?

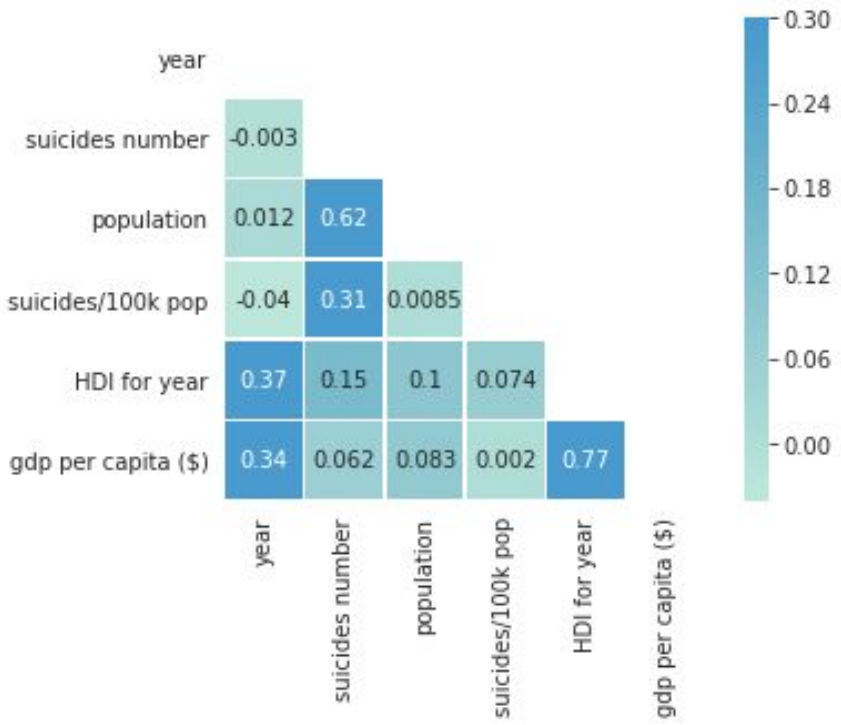
  I chose the datasets because I am interested in seeing what factors are important on the mental health of people, especially in the people who committed suicides.
- What types of questions were you hoping to explore with this data?

  I would like to see what are the factors impacting the suicides rate and try to explore possible solutions.

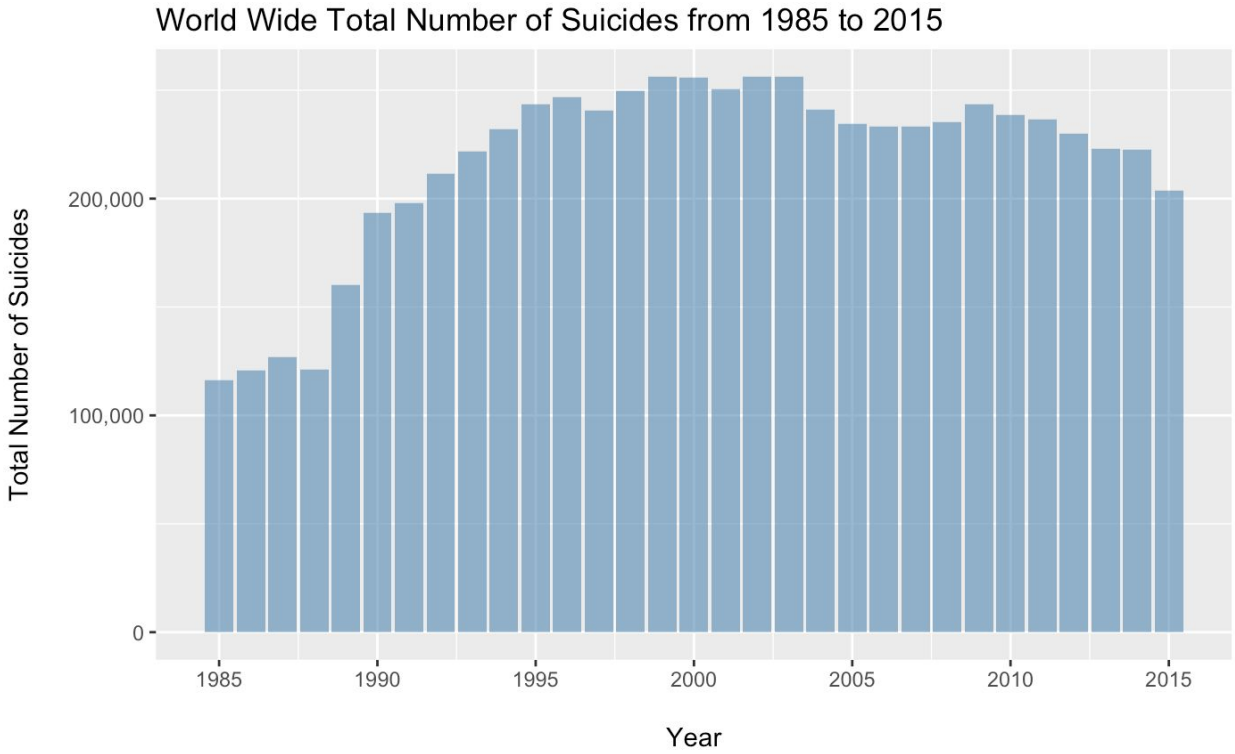
Data Preprocessing

- Removed all data from 2016 as data included in 2016 are not complete
- Removed countries with less than or equal to 3 years of data
- Added continent information

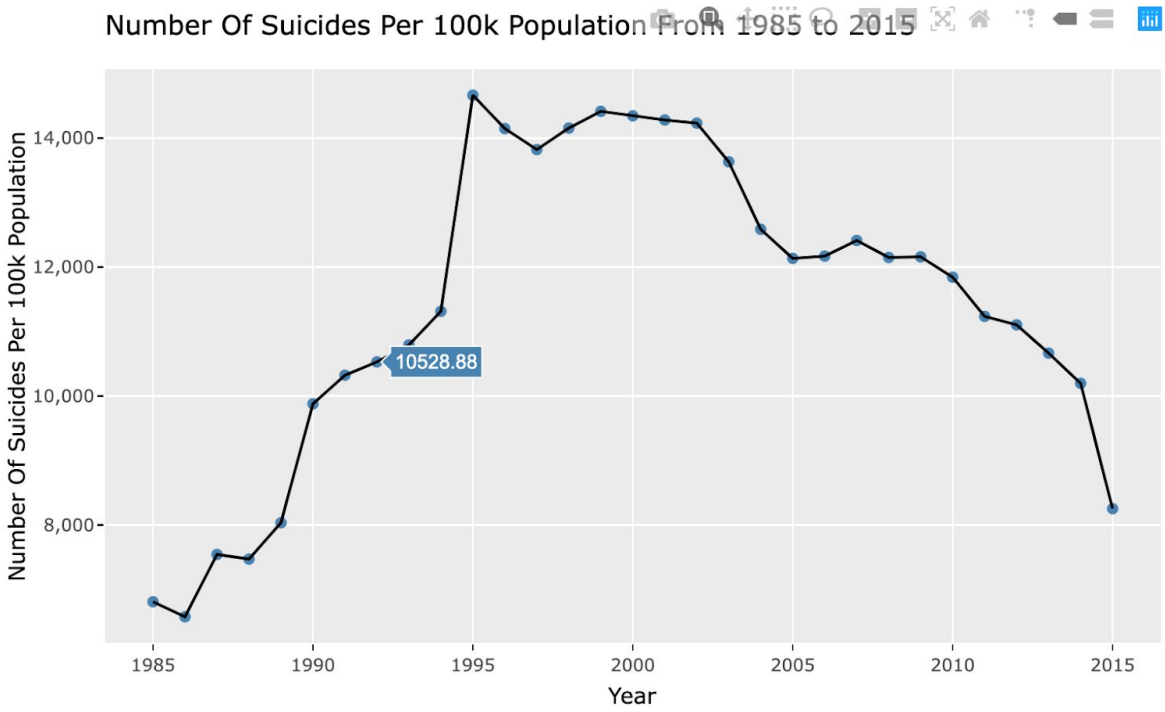
Exploratory Data Analysis - Summary of Data



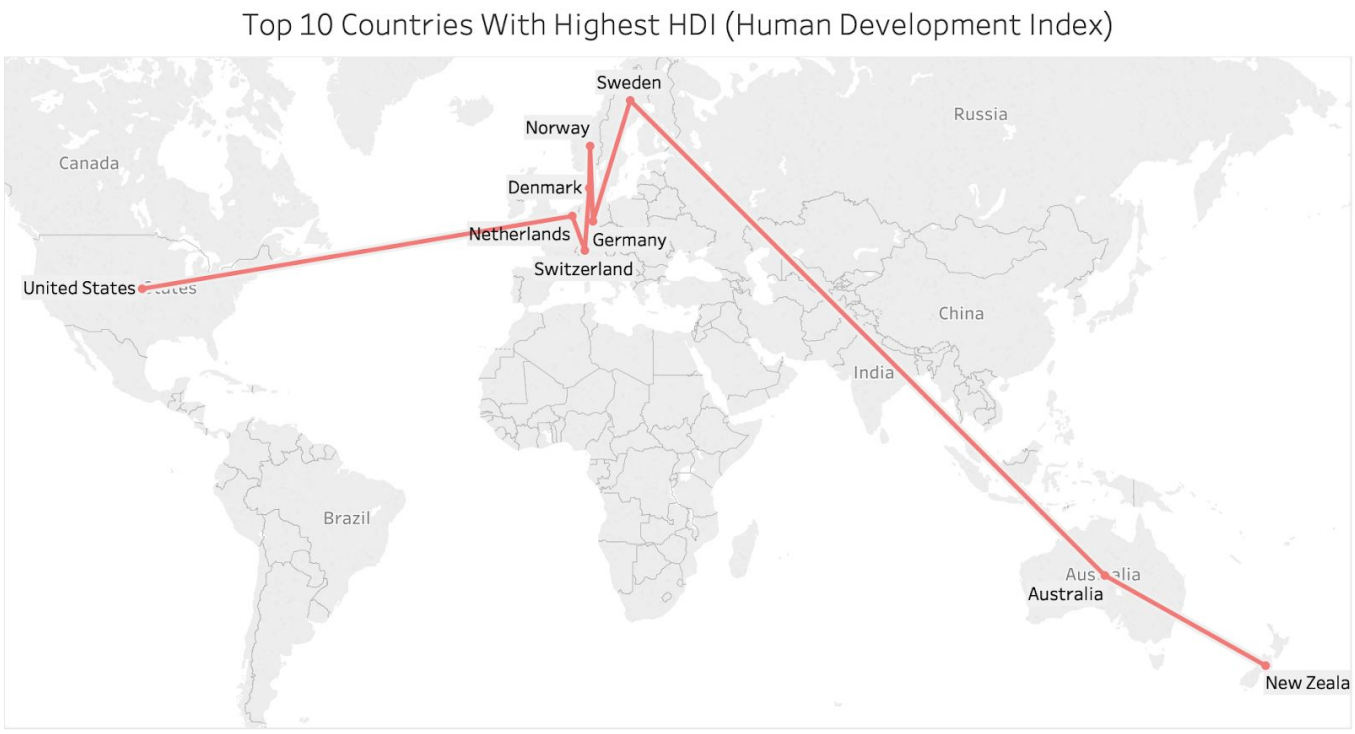
The above heat plot shows the correlation between different variables in the dataset. Note that population is highly correlated with number of suicides. Also, GDP Per Capita is correlated with year and HDI for Year.



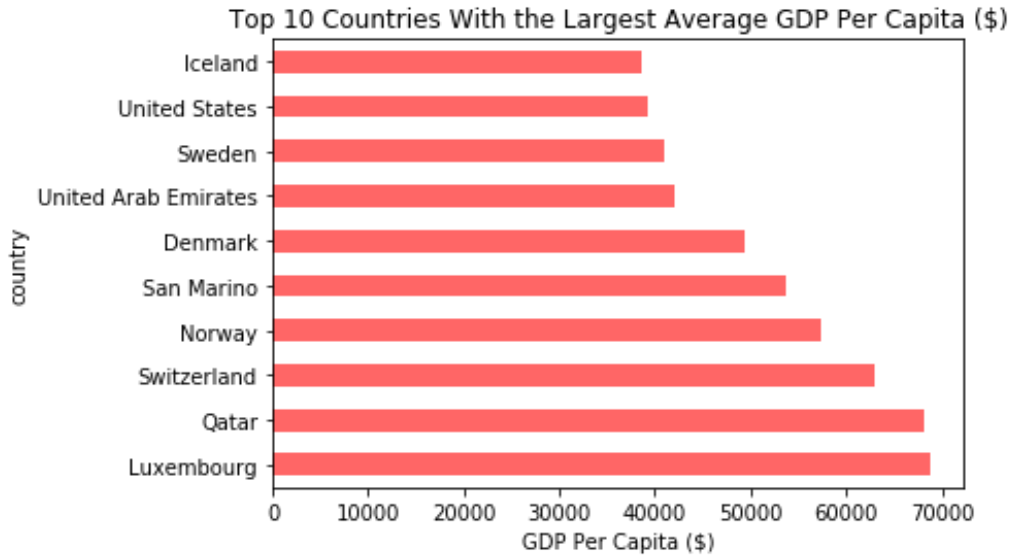
This histogram shows the number of total suicides from 1985 to 2015 throughout the world. Notice that the number increased a lot from 1985 but started to decrease from around 2003.



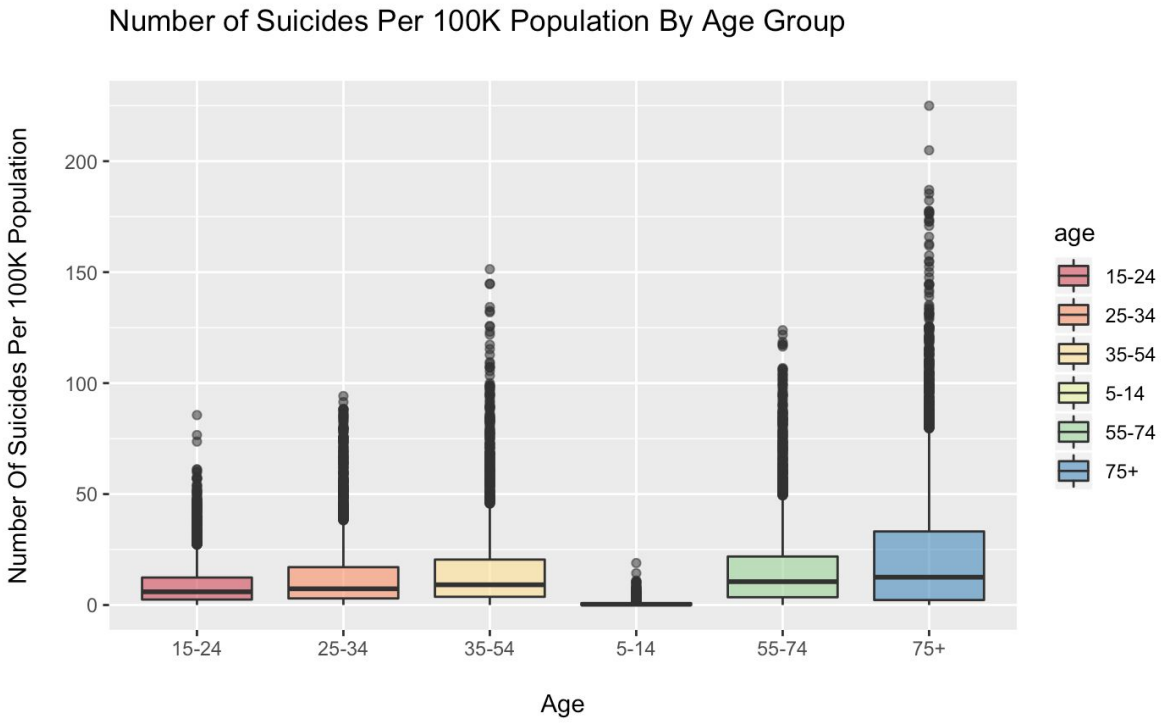
The histogram does not give all the information because the population of the world is increasing as well. Therefore, I plotted the number of suicides per 100k population from 1985 to 2015. Notice that the line plot basically follow a similar trend.



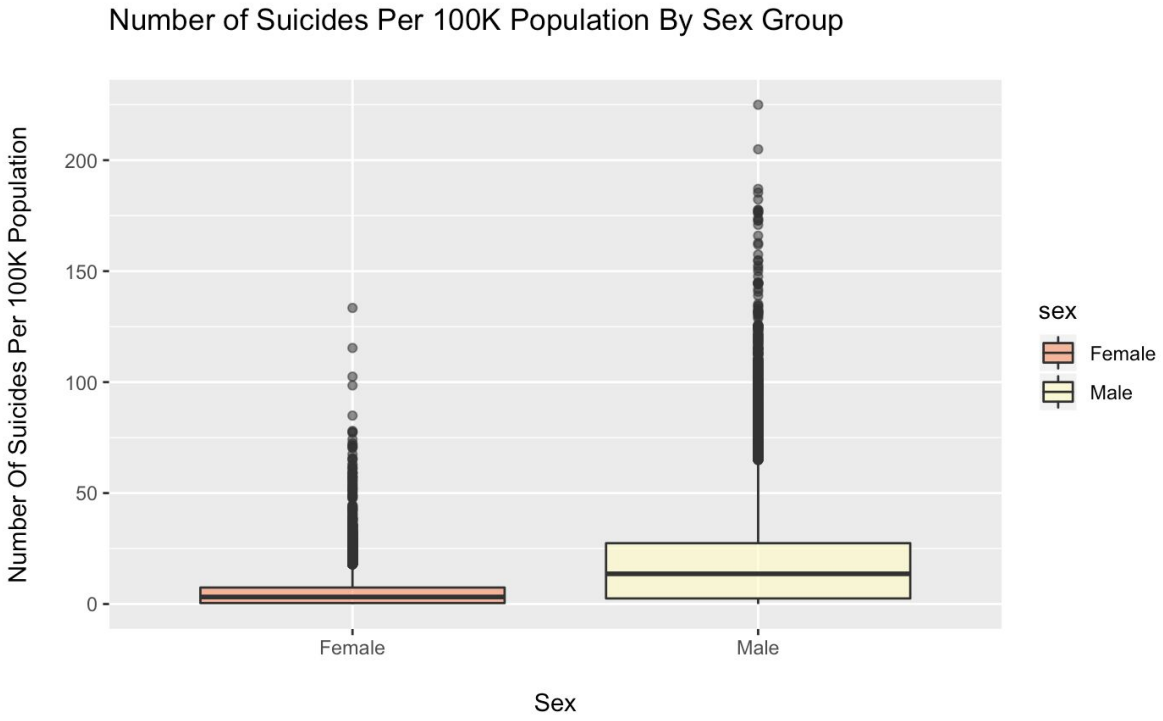
The above connection map shows the top 10 countries with the highest average HDI. Noticed that most of the top countries are in Europe, along with the United States, Australia, and New Zealand.



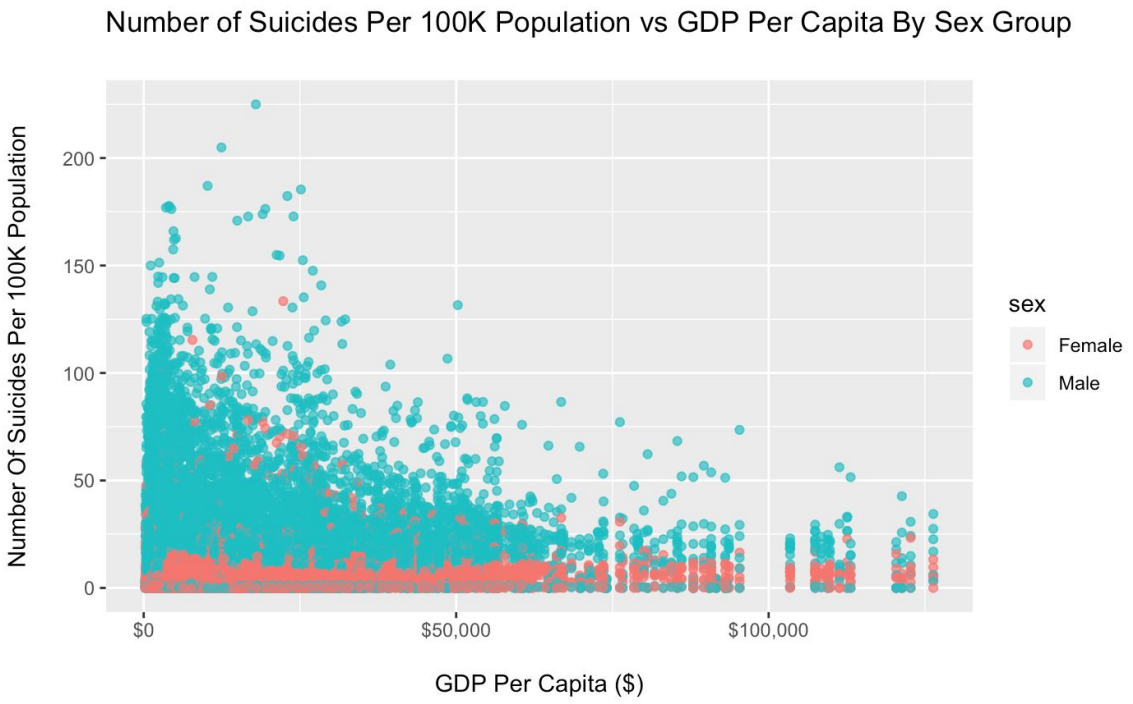
The bar plot shows the top 10 countries with the largest average GDP Per Capita ($) from 1985 to 2015. Most of the countries are in Europe as well.



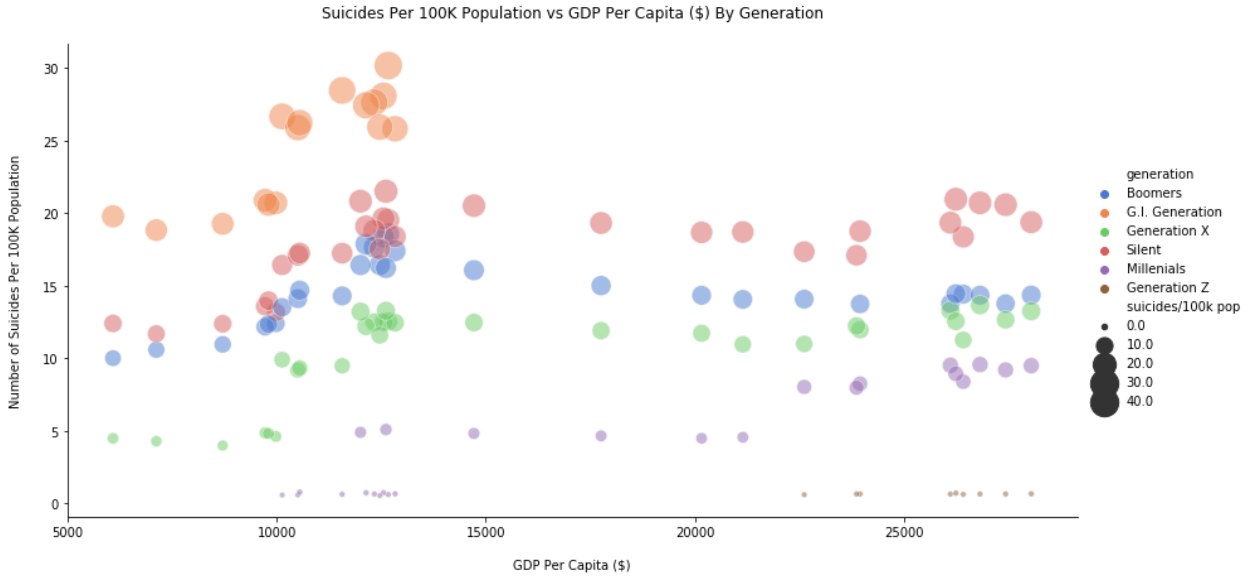
Furthermore, I would like to see how the number of suicides per 100k population changes by different age group. Surprisingly, the group of people who are over 75 years old has the largest median number of suicides per 100k population, followed by the age group 55-74 years old and 35-54 years old.



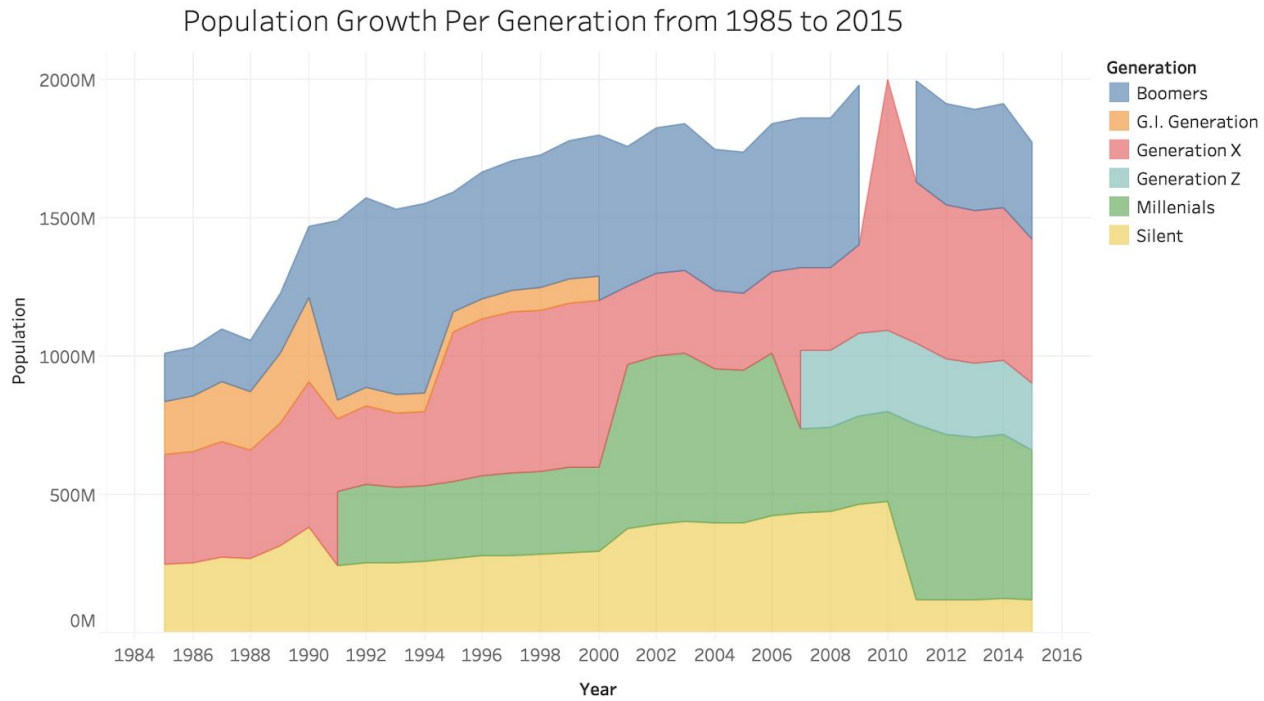
This boxplot shows the number of suicides per 100k population changes by different sex group. Notice that there are more males than females who committed suicides per 100k population, in general.



This scatter plot shows the relationship between the number of suicides per 100k population and GDP per capita ($) per sex group. Noticed that overall, the higher the GDP per capita is, the lower the suicides rate is. Also, the males, in general, have higher suicides rate than the females.



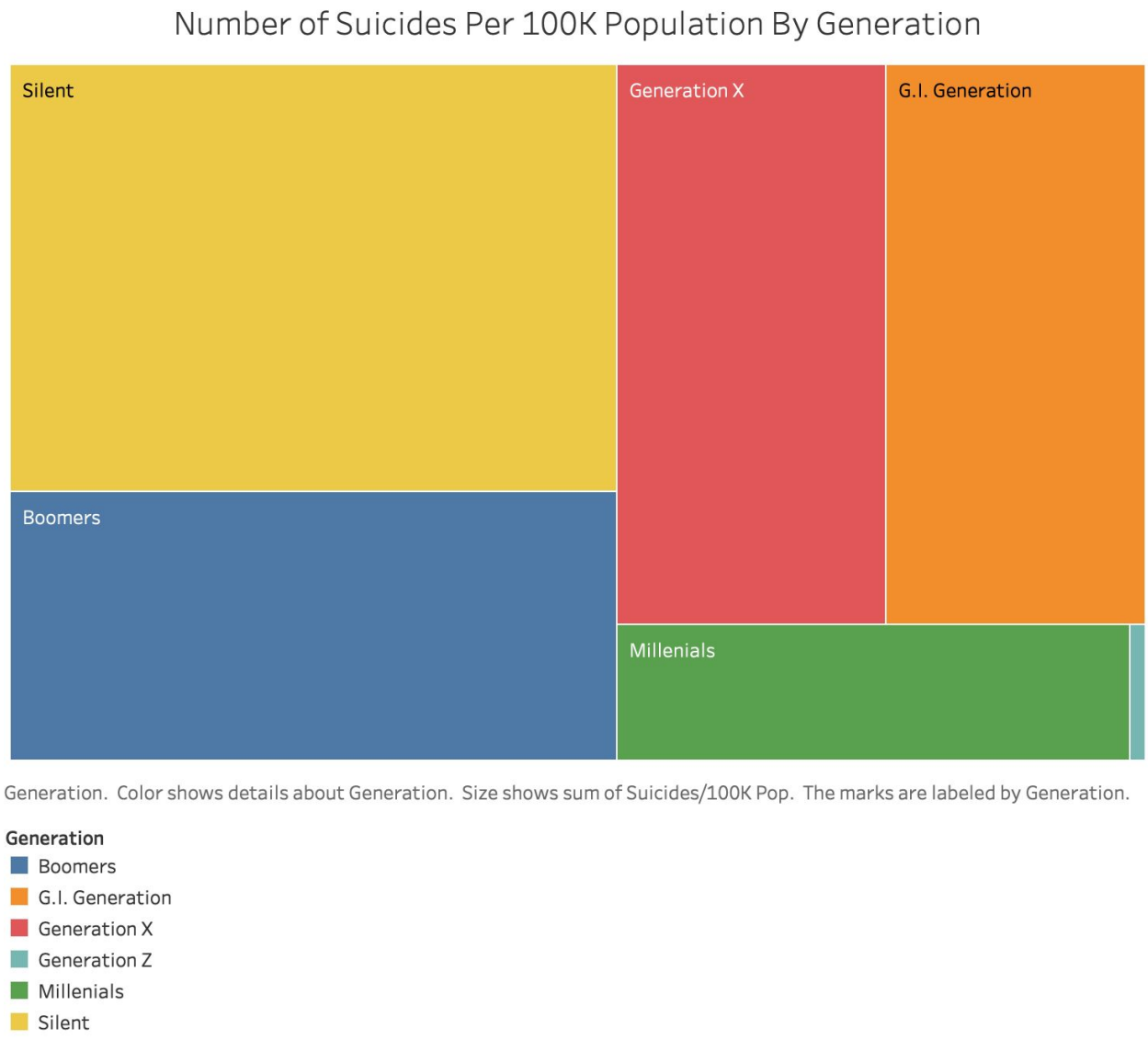
The above bubble plot shows the suicides per 100k population vs GDP per capita ($) by generations (definitions of different generations are below). Note that the smaller the GDP per capita is, the larger the number of suicides per 100k population. In general, the G.I. Generation has the largest number of suicides per 100k population.



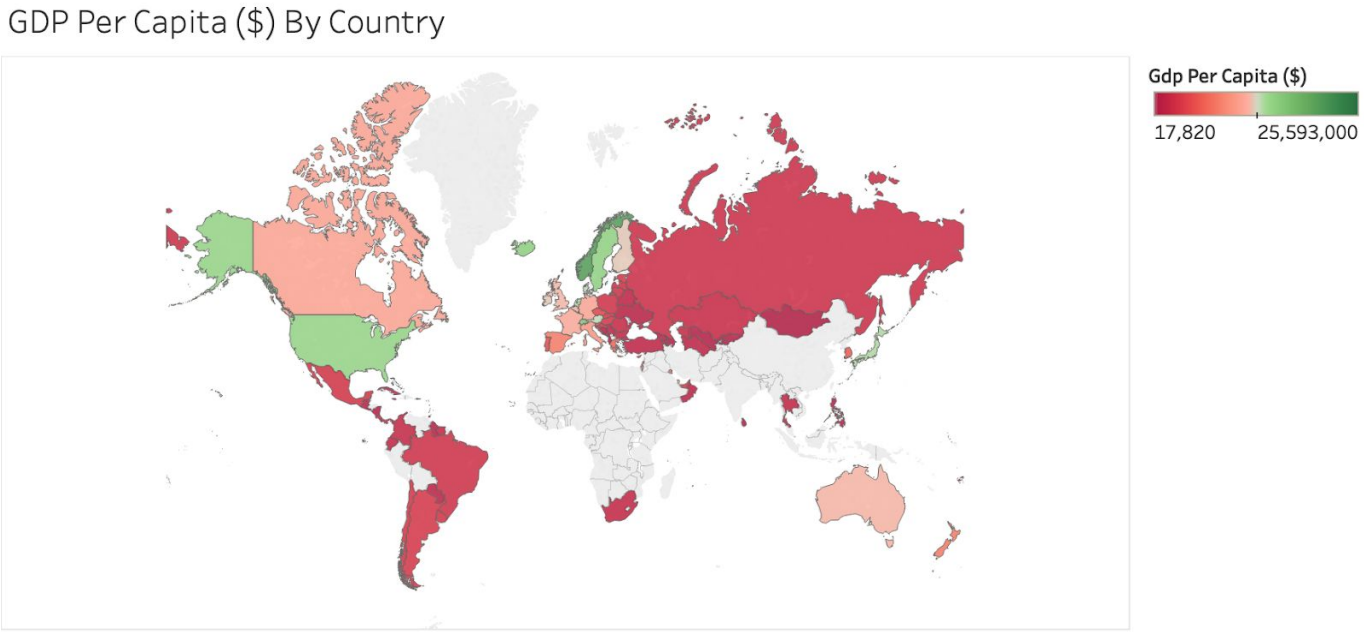
The plot of sum of Population for Year. Color shows details about Generation. The view is filtered on Year, which ranges from 1985 to 2015.

This stacked area plot shows the population growth per generation from 1985 to 2015. Notice that throughout the years, the Boomers and the Generation X have been increasing steadily. However, there is some decrease in the population for the Silent, the Generation Z, and Millenials. Definition for these different generation groups are as below:

- **The Greatest Generation, also known as the G.I. Generation** and the World War II generation, is the demographic cohort following the Lost Generation. Demographers and researchers typically use the early 1900s as starting birth years and ending birth years in the mid to late 1920s. The generation was shaped by the Great Depression and were the primary participants in World War II.
- **The Silent Generation** is the demographic cohort following the Greatest Generation. Demographers and researchers typically use mid-to-late 1920s as starting birth years and early-to-mid 1940s as ending birth years for this cohort.
- **Baby boomers (also known as boomers)** are the demographic cohort following the Silent Generation and preceding Generation X.
- **Generation X or Gen X** is the demographic cohort following the baby boomers and preceding the Millennials. Demographers and researchers typically use birth years ranging from the early-to-mid 1960s to the early 1980s.
- **Millennials, also known as Generation Y or Gen Y**, are the demographic cohort following Generation X and preceding Generation Z.
- **Generation Z or Gen Z**, also known by a number of other names, is the demographic cohort after the Millennials. Demographers and researchers typically use the mid-1990s to mid-2000s as starting birth years.



The above treemap gives us a better visual idea of how the suicides rate per 100k population by different generations looks like.



Green indicates a higher GDP per capita and red indicates a lower GDP per capita.

This choropleth map shows the average GDP Per Capita ($) by countries. Note that generally, the countries in western Europe have a higher GDP per Capita along with the United States.

Story

The above bar plot shows the top 10 countries, age, sex groups which have the largest average number of suicides per 100k population. Surprisingly, most of the people who committed

suicides are of the 75+ years old group and all of them are male. More surprisingly to me is that Hungary is the country with the largest average number of males committed suicides who is over 75 years old. According to a Hungarian author Noemi Szecsi, “the frustration comes from loss after loss in the international political realm, dating back to the break-up of the Austro-Hungarian Empire that saw Hungary lose much of its land”. She also mentioned that this also relates to “the insufferable living conditions in the refugee and migrant camps”.

In addition, the suicides happening to the elder generation is a serious issue. After researching for a while, I realized that this might relate to a lot of different factors, including physical health issues, and employment and financial factors.

From the above EDA and analysis, it is not hard to tell that, there are a lot of factors that influence the suicides rates, including the economy of a country, the mental and physical health conditions, and the living and working conditions. In order to decrease the suicides rate and help people living a better life, we should pay more attention to people’s mental health and offer help whenever noticing any non-normal signs.

Code

R

```
# data preprocessing
data <- data %>%
   filter(year!=2016)
minimum_years <- data %>%
 group_by(country) %>%
 summarize(rows = n(),
           years = rows / 12) %>%
 arrange(years)

data <- data %>%
 filter(!(country %in% head(minimum_years$country, 7)))
data$age <- gsub(" years", "", data$age)
data$sex <- ifelse(data$sex == "male", "Male", "Female")

# histogram
data %>%
 group_by(year) %>%
 summarise(total_suicides_no=sum(suicides_no)) %>%
 ggplot() +
```

```
  geom_histogram(aes(x=year,y=total_suicides_no), stat = "identity", fill="steel blue", alpha=0.6)
+
  xlab("\nYear") +
  ylab("Total Number of Suicides\n") +
  ggtitle("World Wide Total Number of Suicides from 1985 to 2015") +
  scale_y_continuous(labels = comma) +
  scale_x_continuous(breaks = seq(1985, 2017, 5))

# interactive plot
lineplot <- data %>%
  group_by(year) %>%
  summarise(suicides_per_100k=sum(`suicides/100k pop`)) %>%
  ggplot() +
  geom_point(aes(x=year, y=suicides_per_100k, text = suicides_per_100k), color="steel blue") +
  geom_line(aes(x=year, y=suicides_per_100k)) +
  xlab("\nYear")+
  ylab("Number Of Suicides Per 100k Population\n") +
  ggtitle("Number Of Suicides Per 100k Population From 1985 to 2015\n") +
  scale_y_continuous(labels = comma) +
  scale_x_continuous(breaks = seq(1985, 2017, 5))
ggplotly(lineplot, tooltip = c("text", "size"))

#box plot
data %>%
  group_by(age) %>%
  ggplot() +
  geom_boxplot(aes(x=age, y=`suicides/100k pop`, fill=age), alpha=0.6) +
  xlab("\nAge") +
  ylab("Number Of Suicides Per 100K Population\n") +
  scale_fill_brewer(palette="Spectral") +
  ggtitle("Number of Suicides Per 100K Population By Age Group\n")

data %>%
  group_by(sex) %>%
  ggplot() +
  geom_boxplot(aes(x=sex, y=`suicides/100k pop`, fill=sex), alpha=0.6) +
  xlab("\nSex") +
  ylab("Number Of Suicides Per 100K Population\n") +
  scale_fill_brewer(palette="Spectral") +
  ggtitle("Number of Suicides Per 100K Population By Sex Group\n")


# scatter plot
```

```
data %>%
  group_by(year) %>%
  ggplot() +
  geom_point(aes(x=`gdp_per_capita ($)`, y=`suicides/100k pop`, color=sex), alpha=0.7) +
  xlab("\nGDP Per Capita ($)") +
  ylab("Number Of Suicides Per 100K Population\n") +
  ggtitle("Number of Suicides Per 100K Population vs GDP Per Capita By Sex Group\n") +
  scale_x_continuous(labels = dollar)
```

Python

```
# heatmap
mask = np.zeros_like(data.corr())
mask[np.triu_indices_from(mask)] = True
with sns.axes_style("white"):
    ax = sns.heatmap(data.corr(), vmax=.3, center=1, square=True, linewidths=.5,annot=True,
mask=mask)
plt.show()

# bar plot
data.groupby(['country'])["gdp_per_capita ($)"].mean().nlargest(10).plot(kind='barh', color="r",
alpha=0.6)
plt.title("Top 10 Countries With the Largest Average GDP Per Capita ($)")
plt.xlabel("GDP Per Capita ($)")

data.groupby(['country','age', "sex"])["suicides/100k pop"].mean().nlargest(10).plot(kind='barh',
                                                                color='blue', alpha=0.5)

data.groupby(['country','age']).suicides_no.sum().nlargest(10).plot(kind='barh', color='orange',
alpha=0.5)
plt.title("Top 10 Country,Age Groups With the Highest Suicides Number 1985-2015")
plt.xlabel("Suicides Number")

# bubble plot
sns.relplot(y="suicides/100k pop", x="gdp_per_capita ($)", hue="generation",
size="suicides/100k pop",
            sizes=(20, 500), alpha=.5, palette="muted", aspect=2,
            height=6, data=data_2)
plt.title(f"Suicides Per 100K Population vs GDP Per Capita ($) By Generation\n")
plt.ylabel("Number of Suicides Per 100K Population\n")
plt.xlabel("\nGDP Per Capita ($)")
plt.show()
```

Reference

Wikipedia
https://www.kaggle.com/lmorgan95/r-suicide-rates-in-depth-stats-insights
https://www.aginginplace.org/elderly-suicide-risks-detection-how-to-help/
https://www.aginginplace.org/elderly-suicide-risks-detection-how-to-help/