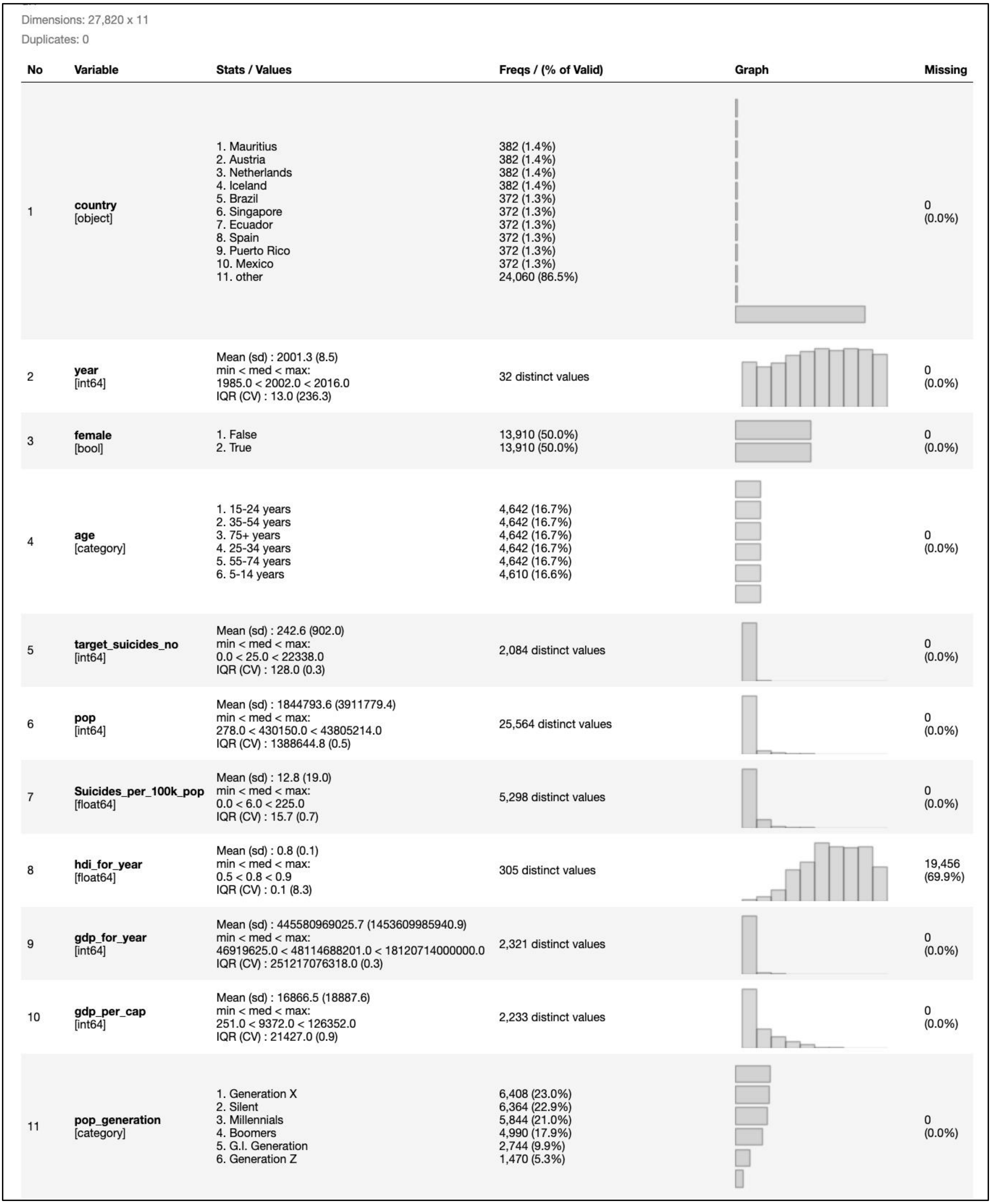
Project 1: Exploratory Data Analysis (EDA)

Task 1

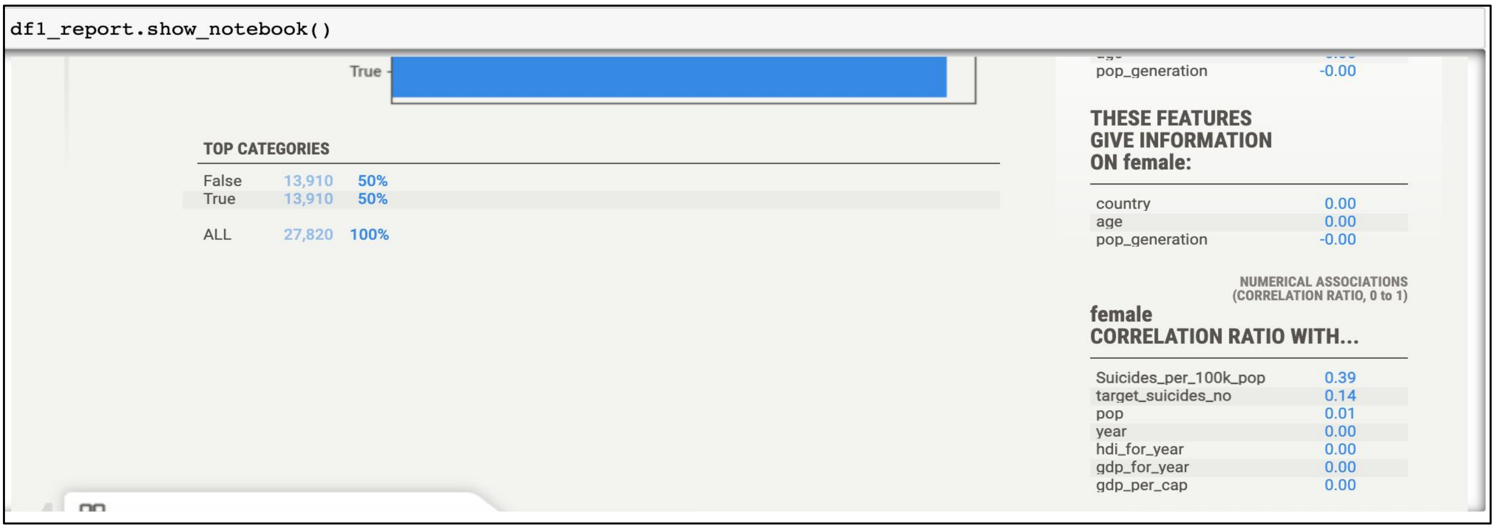
Provide statistical summaries of the data. Please show at least five different types of summaries, and analysis your findings. Use these observations to learn and communicate (in your reports) something about the data. Simple examples include number of samples, attributes, column names, if missing values are present, quartiles etc.



Dimensions: 27,820 x 11

Duplicates: 0

No	Variable	Stats / Values	Freqs / (% of Valid)	Graph	Missing
1	**country** [object]	1. Mauritius 2. Austria 3. Netherlands 4. Iceland 5. Brazil 6. Singapore 7. Ecuador 8. Spain 9. Puerto Rico 10. Mexico 11. other	382 (1.4%) 382 (1.4%) 382 (1.4%) 382 (1.4%) 372 (1.3%) 372 (1.3%) 372 (1.3%) 372 (1.3%) 372 (1.3%) 372 (1.3%) 24,060 (86.5%)		0 (0.0%)
2	**year** [int64]	Mean (sd) : 2001.3 (8.5) min < med < max: 1985.0 < 2002.0 < 2016.0 IQR (CV) : 13.0 (236.3)	32 distinct values		0 (0.0%)
3	**female** [bool]	1. False 2. True	13,910 (50.0%) 13,910 (50.0%)		0 (0.0%)
4	**age** [category]	1. 15-24 years 2. 35-54 years 3. 75+ years 4. 25-34 years 5. 55-74 years 6. 5-14 years	4,642 (16.7%) 4,642 (16.7%) 4,642 (16.7%) 4,642 (16.7%) 4,642 (16.7%) 4,610 (16.6%)		0 (0.0%)
5	**target_suicides_no** [int64]	Mean (sd) : 242.6 (902.0) min < med < max: 0.0 < 25.0 < 22338.0 IQR (CV) : 128.0 (0.3)	2,084 distinct values		0 (0.0%)
6	**pop** [int64]	Mean (sd) : 1844793.6 (3911779.4) min < med < max: 278.0 < 430150.0 < 43805214.0 IQR (CV) : 1388644.8 (0.5)	25,564 distinct values		0 (0.0%)
7	**Suicides_per_100k_pop** [float64]	Mean (sd) : 12.8 (19.0) min < med < max: 0.0 < 6.0 < 225.0 IQR (CV) : 15.7 (0.7)	5,298 distinct values		0 (0.0%)
8	**hdi_for_year** [float64]	Mean (sd) : 0.8 (0.1) min < med < max: 0.5 < 0.8 < 0.9 IQR (CV) : 0.1 (8.3)	305 distinct values		19,456 (69.9%)
9	**gdp_for_year** [int64]	Mean (sd) : 445580969025.7 (1453609985940.9) min < med < max: 46919625.0 < 48114688201.0 < 18120714000000.0 IQR (CV) : 251217076318.0 (0.3)	2,321 distinct values		0 (0.0%)
10	**gdp_per_cap** [int64]	Mean (sd) : 16866.5 (18887.6) min < med < max: 251.0 < 9372.0 < 126352.0 IQR (CV) : 21427.0 (0.9)	2,233 distinct values		0 (0.0%)
11	**pop_generation** [category]	1. Generation X 2. Silent 3. Millennials 4. Boomers 5. G.I. Generation 6. Generation Z	6,408 (23.0%) 6,364 (22.9%) 5,844 (21.0%) 4,990 (17.9%) 2,744 (9.9%) 1,470 (5.3%)		0 (0.0%)



```
df1_report.show_notebook()
```

True

TOP CATEGORIES

False	13,910	50%
True	13,910	50%
ALL	27,820	100%

pop_generation	-0.00

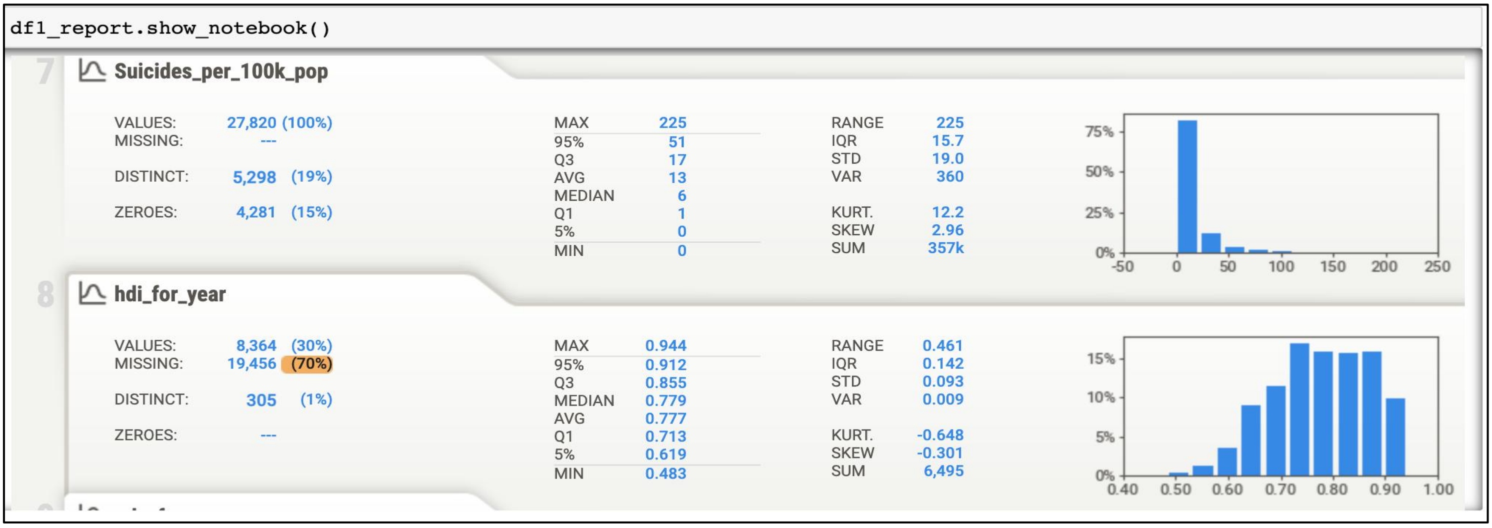
THESE FEATURES GIVE INFORMATION ON female:

country	0.00
age	0.00
pop_generation	-0.00

NUMERICAL ASSOCIATIONS (CORRELATION RATIO, 0 to 1)

female CORRELATION RATIO WITH...

Suicides_per_100k_pop	0.39
target_suicides_no	0.14
pop	0.01
year	0.00
hdi_for_year	0.00
gdp_for_year	0.00
gdp_per_cap	0.00



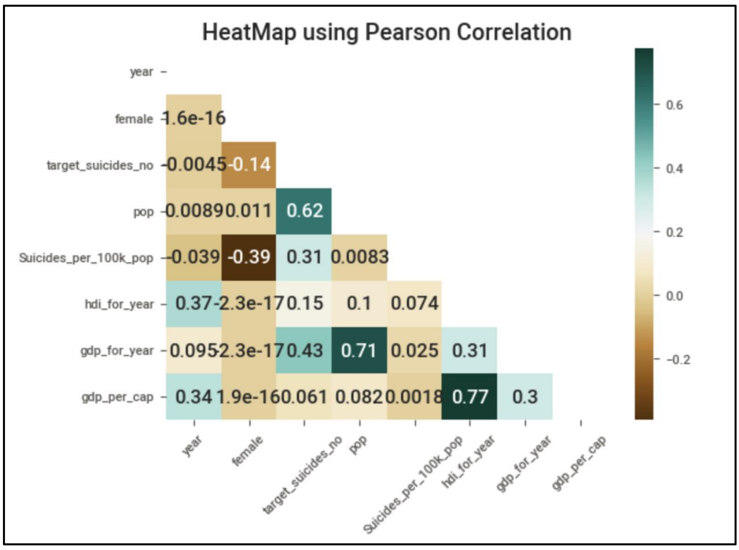
```
df1_report.show_notebook()
```

7 Suicides_per_100k_pop

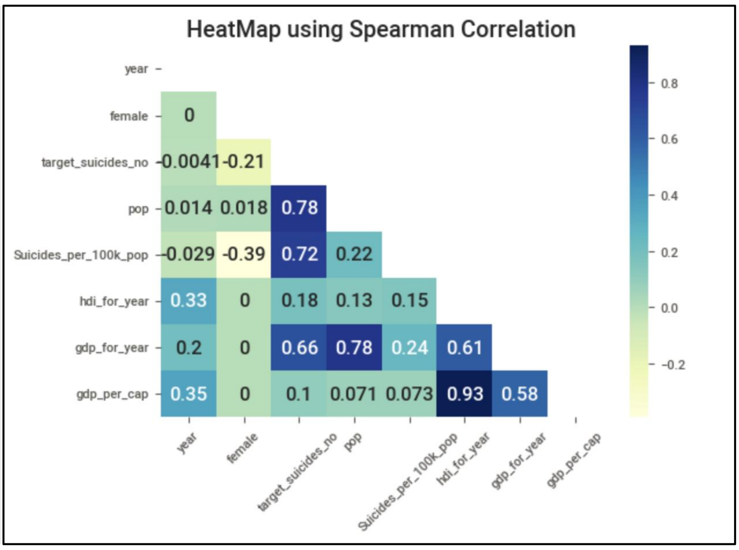
VALUES:	27,820 (100%)	MAX	225	RANGE	225		
MISSING:	---	95%	51	IQR	15.7		
		Q3	17	STD	19.0		
DISTINCT:	5,298 (19%)	AVG	13	VAR	360		
		MEDIAN	6				
ZEROES:	4,281 (15%)	Q1	1	KURT.	12.2		
		5%	0	SKEW	2.96		
		MIN	0	SUM	357k		

8 hdi_for_year

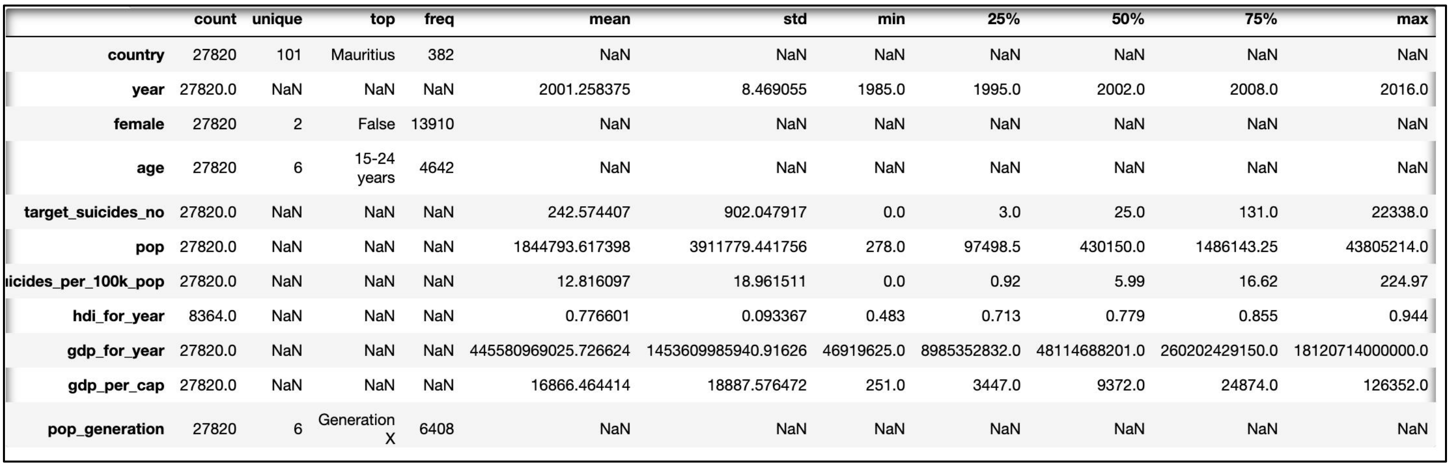
VALUES:	8,364 (30%)	MAX	0.944	RANGE	0.461
MISSING:	19,456 (70%)	95%	0.912	IQR	0.142
		Q3	0.855	STD	0.093
DISTINCT:	305 (1%)	MEDIAN	0.779	VAR	0.009
		AVG	0.777		
ZEROES:	---	Q1	0.713	KURT.	-0.648
		5%	0.619	SKEW	-0.301
		MIN	0.483	SUM	6,495



HeatMap using Pearson Correlation



HeatMap using Spearman Correlation



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
country	27820	101	Mauritius	382	NaN	NaN	NaN	NaN	NaN	NaN	NaN
year	27820.0	NaN	NaN	NaN	2001.258375	8.469055	1985.0	1995.0	2002.0	2008.0	2016.0
female	27820	2	False	13910	NaN	NaN	NaN	NaN	NaN	NaN	NaN
age	27820	6	15-24 years	4642	NaN	NaN	NaN	NaN	NaN	NaN	NaN
target_suicides_no	27820.0	NaN	NaN	NaN	242.574407	902.047917	0.0	3.0	25.0	131.0	22338.0
pop	27820.0	NaN	NaN	NaN	1844793.617398	3911779.441756	278.0	97498.5	430150.0	1486143.25	43805214.0
icides_per_100k_pop	27820.0	NaN	NaN	NaN	12.816097	18.961511	0.0	0.92	5.99	16.62	224.97
hdi_for_year	8364.0	NaN	NaN	NaN	0.776601	0.093367	0.483	0.713	0.779	0.855	0.944
gdp_for_year	27820.0	NaN	NaN	NaN	445580969025.726624	1453609985940.91626	46919625.0	8985352832.0	48114688201.0	260202429150.0	18120714000000.0
gdp_per_cap	27820.0	NaN	NaN	NaN	16866.464414	18887.576472	251.0	3447.0	9372.0	24874.0	126352.0
pop_generation	27820	6	Generation X	6408	NaN	NaN	NaN	NaN	NaN	NaN	NaN

We have generated descriptive statistics to get an overall idea of the distribution of each variable/feature such as their data types, average, max, min, quantile, IQR information, etc. Before generating the statistics the data went through a pre-processing stage which included deletion of a column, renaming of columns, and text processing which reduced the dataset memory space by 28%.

Based on the presented descriptive statistics we can infer that there is a correlation between our features such as population vs. total suicide cases, population vs. total suicide with GDP for the given year, suicides per 100K population vs. GDP per Capita, suicides per 100K population vs. gender, etc. We have a dataset of ~27K rows and 11 features (after pre-processing) and the data expands from 1985 to 2016.

The data consists of features mentioned below

Features Description

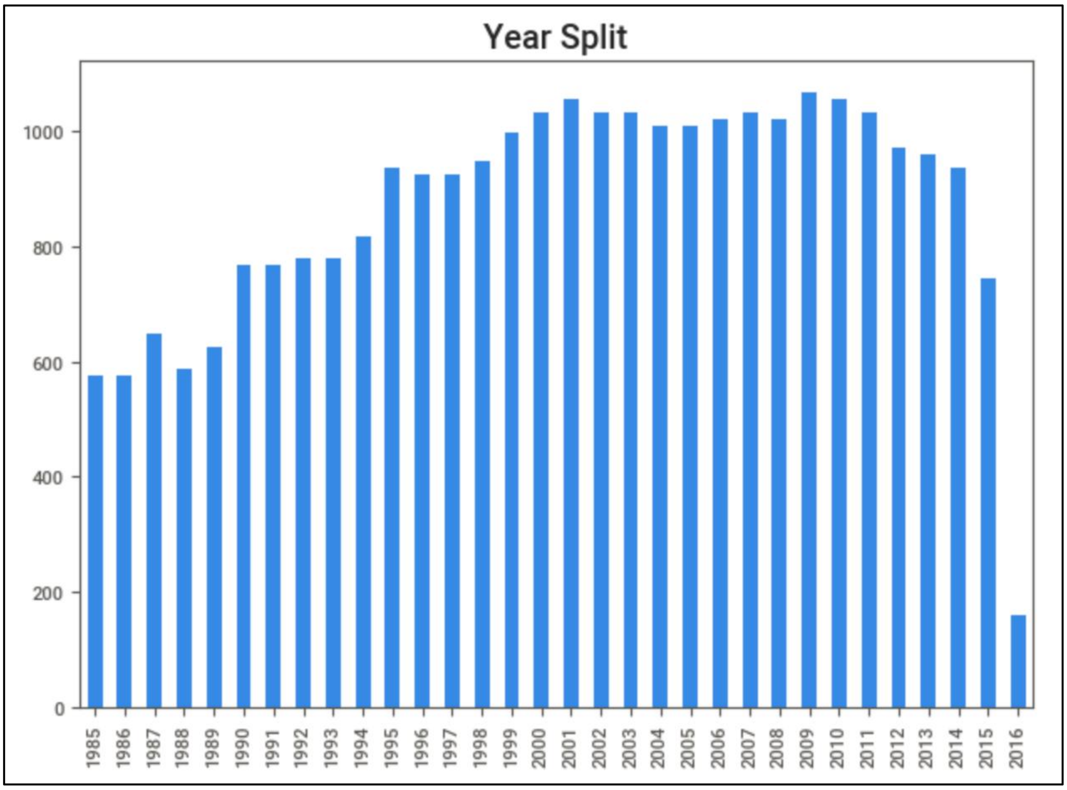
country -> place of suicide
year -> when the suicide happened
female -> gender of the person who committed suicide. True for females, False for male
age -> age of the person who committed suicide grouped in the bin (6 unique with 15 to 24 being the most frequent group)
target_suicides_no -> total number of suicides
pop -> the population of the country where the person lives
suicides_per_100k_pop -> The ratio of suicide rate to 100 thousand of the population.
hdi_for_year -> Human Development Index (~70% missing)
gdp_for_year -> Gross domestic product (a monetary measure of the market value of all the final goods and services produced in a specific time period) for the year
gdp_per_cap -> Gross Domestic Product per Capita shows a country's GDP divided by its total population. (Gross domestic product/total population)
pop_generation -> generation of the person who committed suicide grouped in the bin (6 unique with Generation X being the most frequent group)

pop_generation group details
1. G.I. Generation -> 1901-1927
2. Silent -> 1928-1945
3. Boomers -> 1946-1964
4. Generation X -> 1965-1980
5. Millennials -> 1981-1996
6. Generation Z -> 1997-2012 (since we have data till 2016, we raise the upper bond from 2012 to 2016 for our analysis)

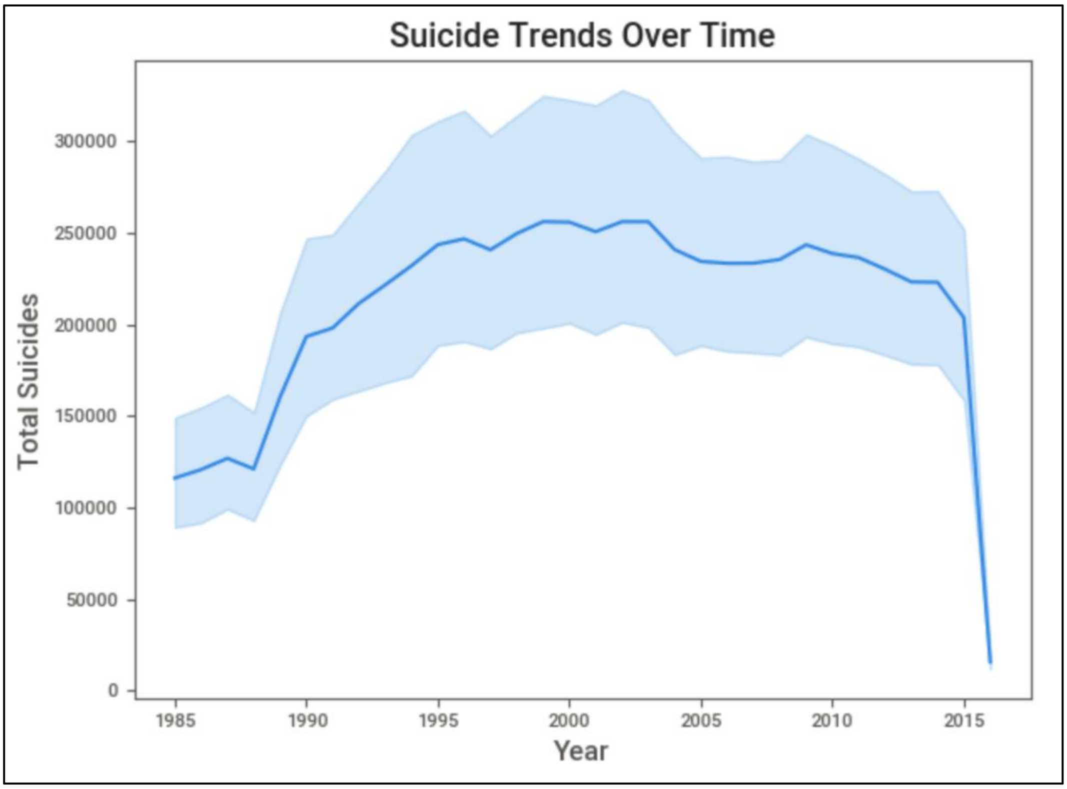
Task 2

Perform exploratory analysis on the data. Research online for ideas, and then show analysis on at least five different aspects of the dataset. Analyze your findings.

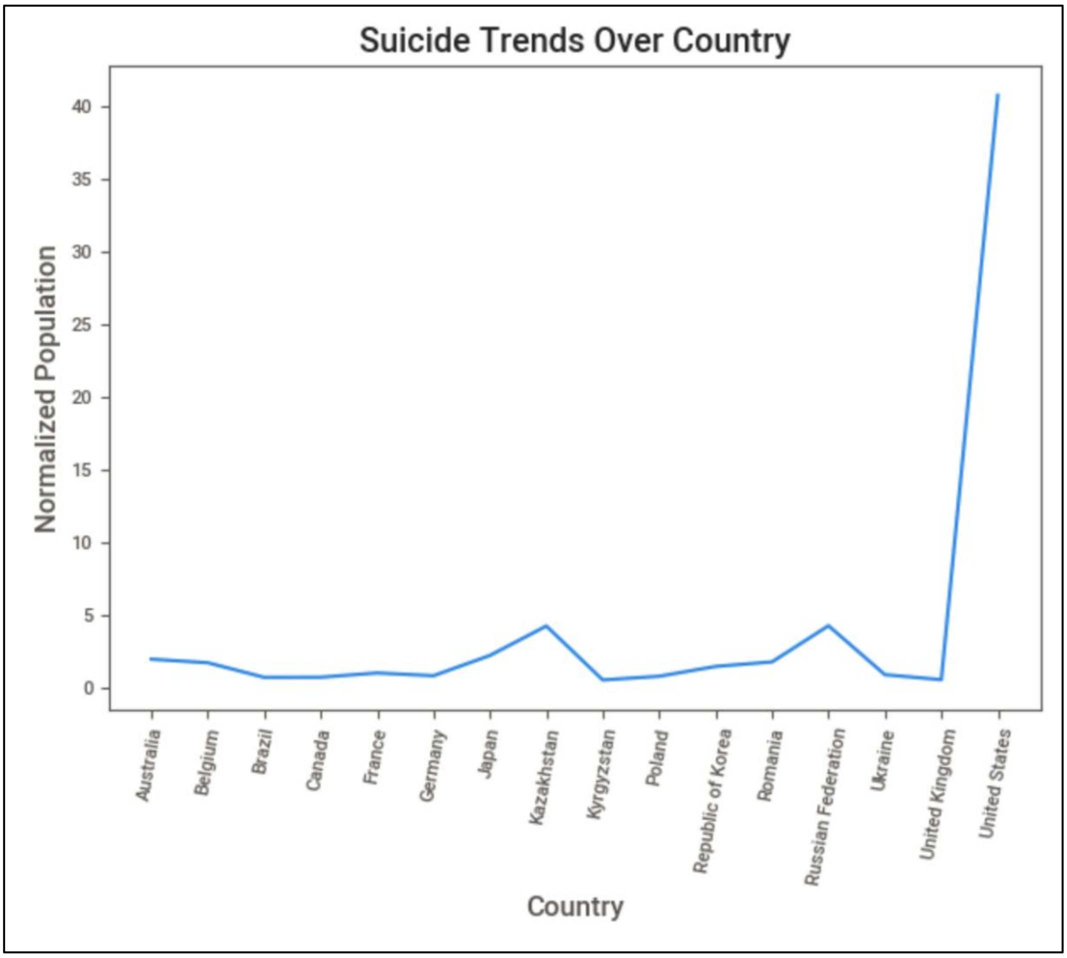
Note: This part is open-ended. Any valid analysis is fine. Use visualizations when necessary.

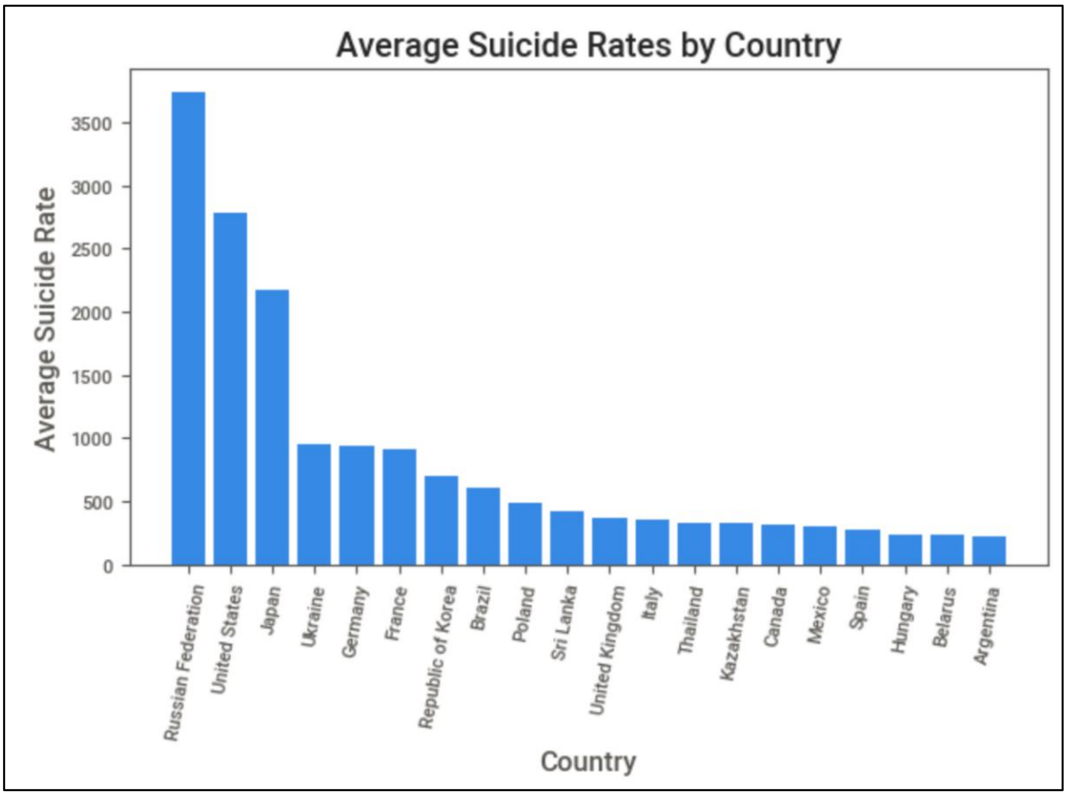


1. The above graph shows the year count and we can see a dip in the data being collected over time. The sharp decline comes after 2014. The highest number of suicides reported were after 1994.

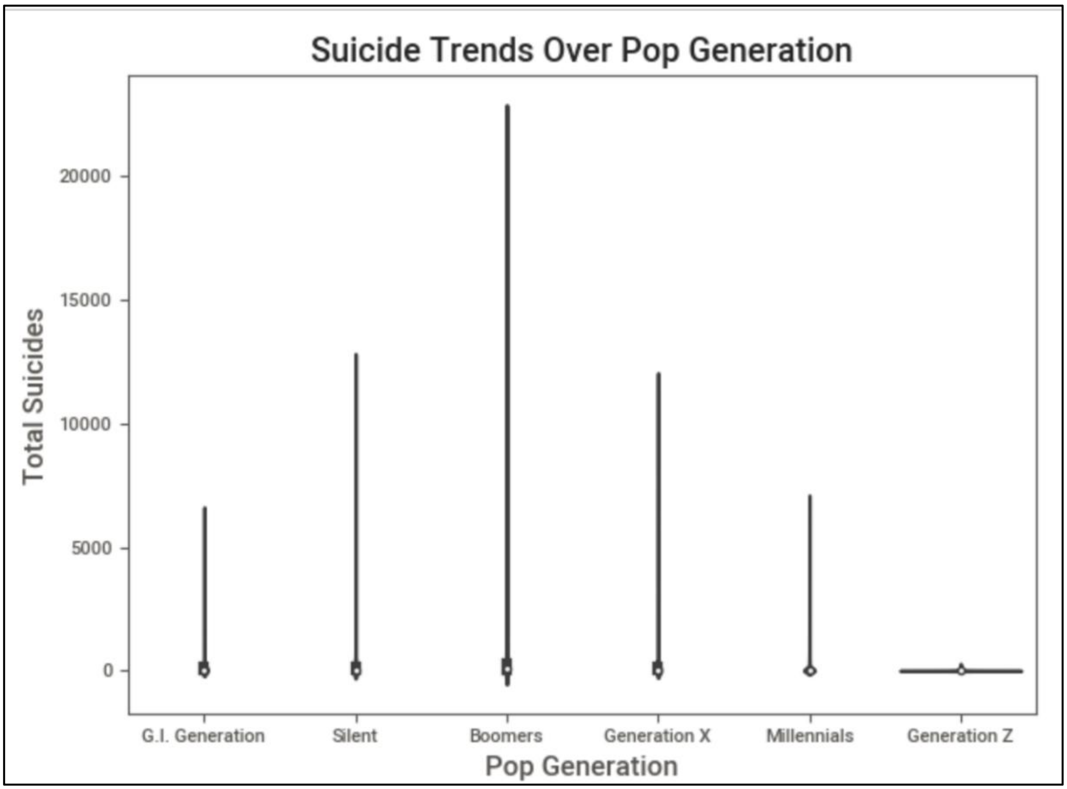


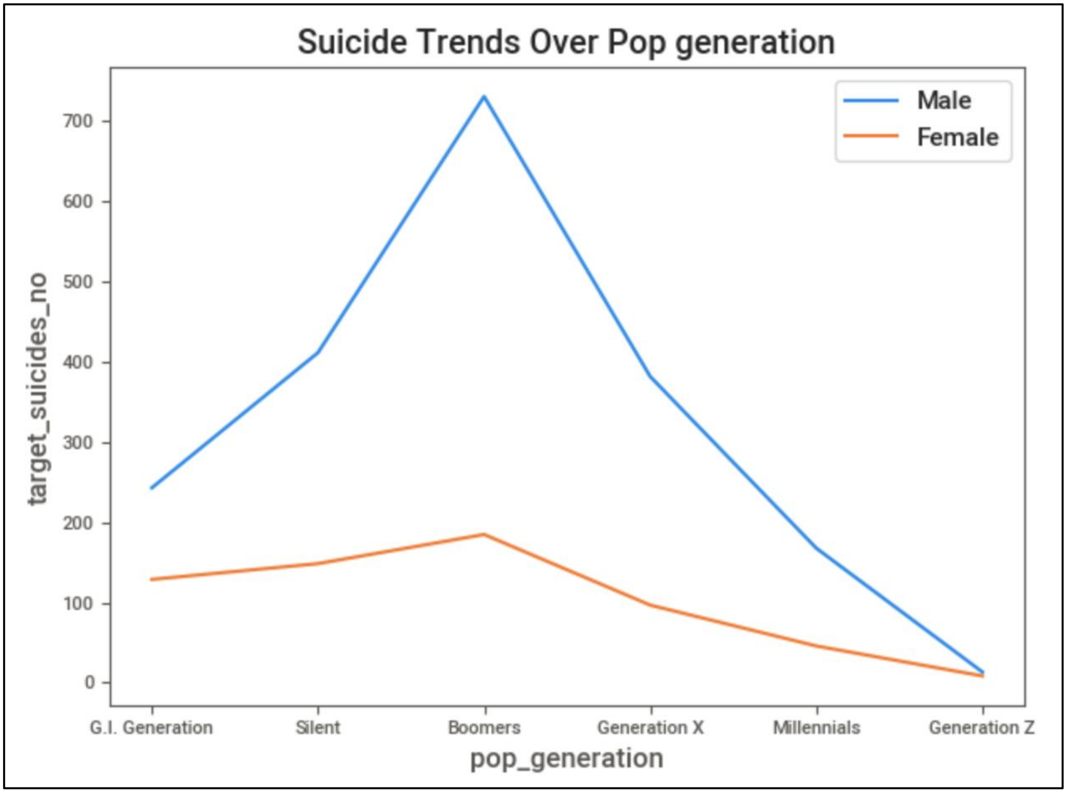
2. The suicide trend took a sharp rise after 1998 and seems to be decreasing after 2008. There may be some additional factors playing a role in the trend.



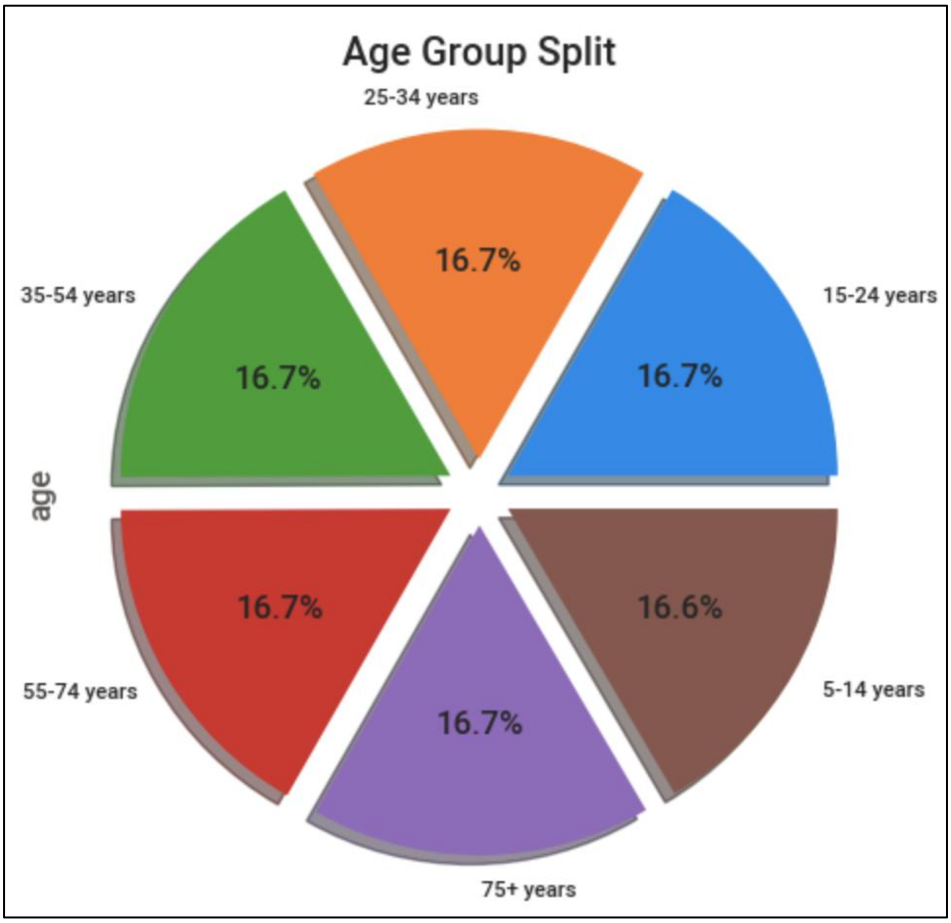


3. There are specific countries where there seems to be a sharp rise in the number of suicide rates such as Kazakhstan, the Russian Federation, the United Kingdom, and the United States.

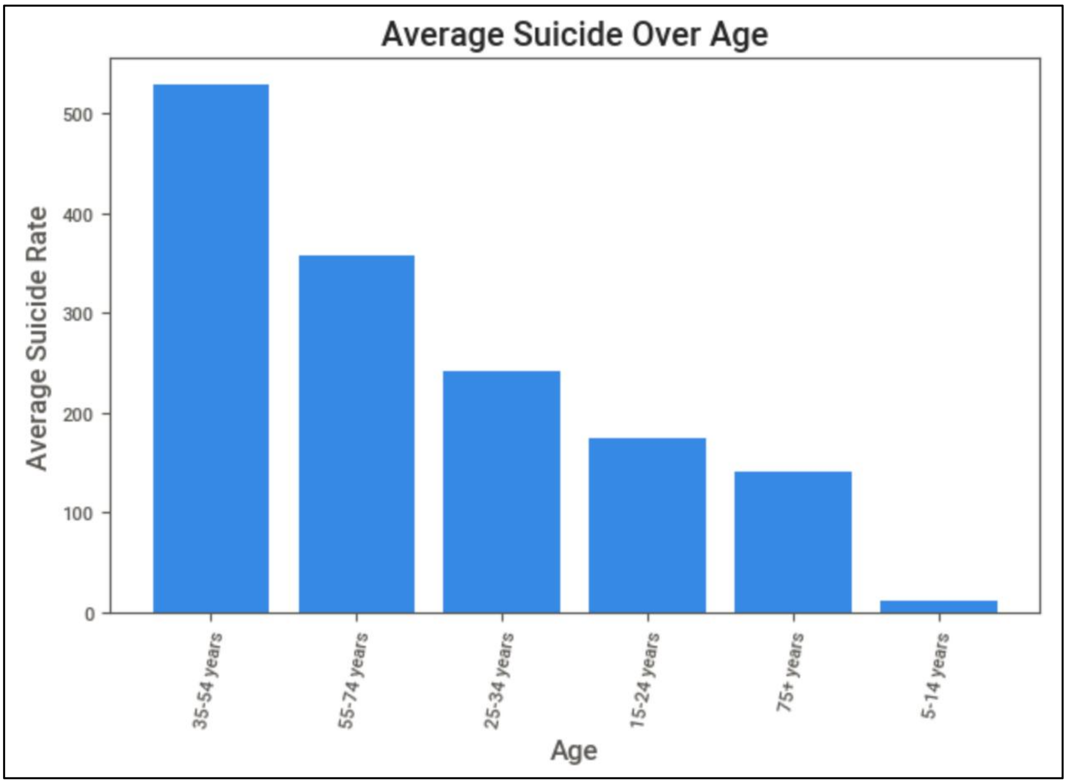




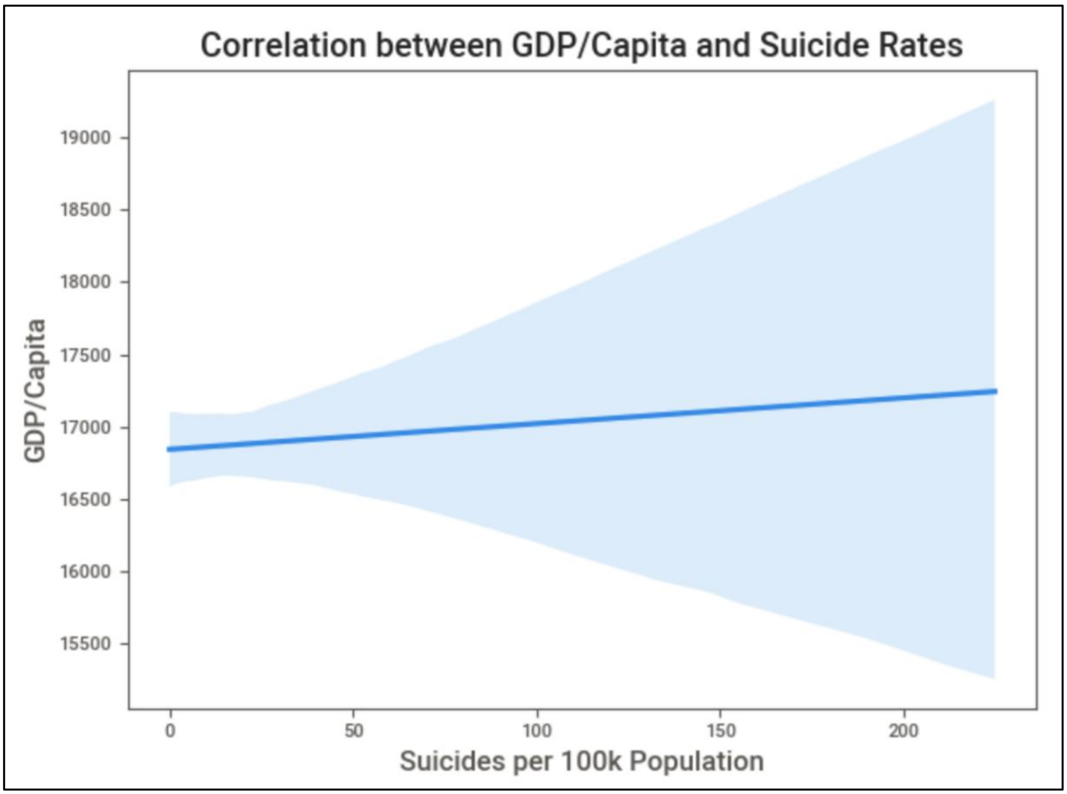
4. The above graph shows that Boomers, Silent, and Generation X were top among the generation groups that committed suicide. After adding the gender dimension we can clearly see the rate of suicide of males was always higher than females.

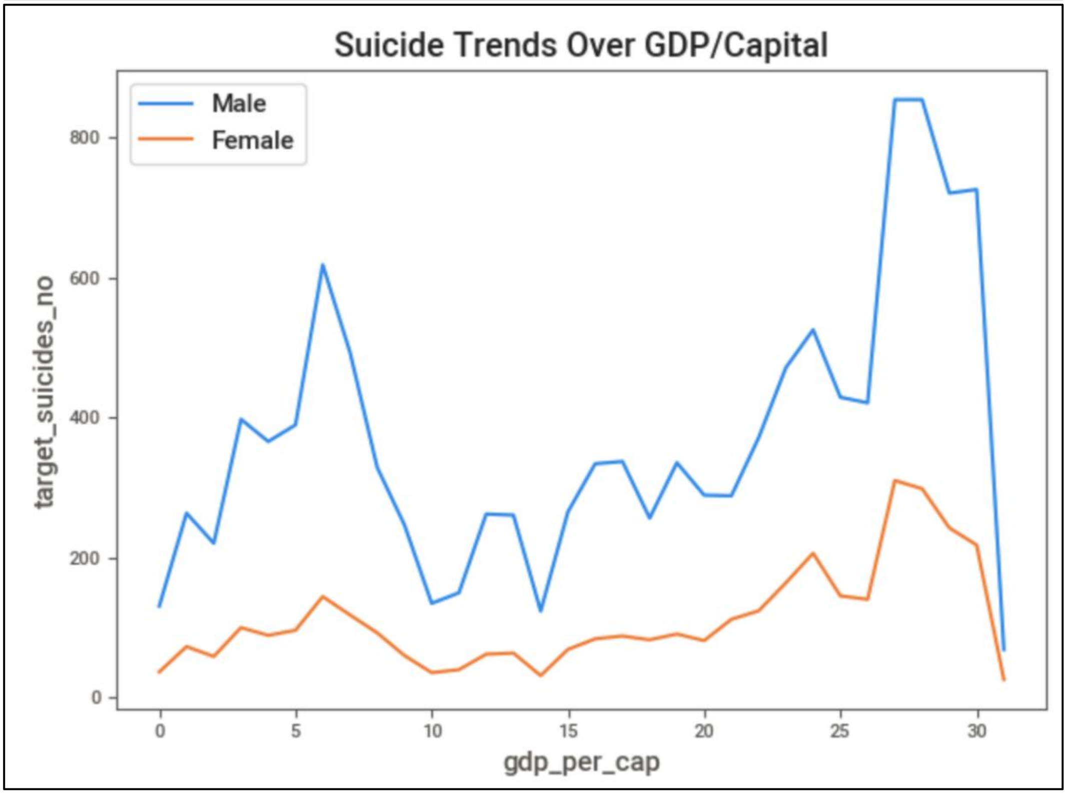


5. There is an almost equal distribution of the count of suicides based on age group.



6. The graph clearly shows that the age between 35-54 years was the highest who committed the most number of suicides from a span of 31 years (1985 - 2016) and the least was the 5-14 years age group.





7. The graph shows the correlation between GDP per Capita vs. Suicides per 100K population. When the GDP per Capita is on a lower bond there seem to be more Males committing suicides than females but there seems to be a rise in the number of females as well even when the GDP per Capita is on a higher bond.

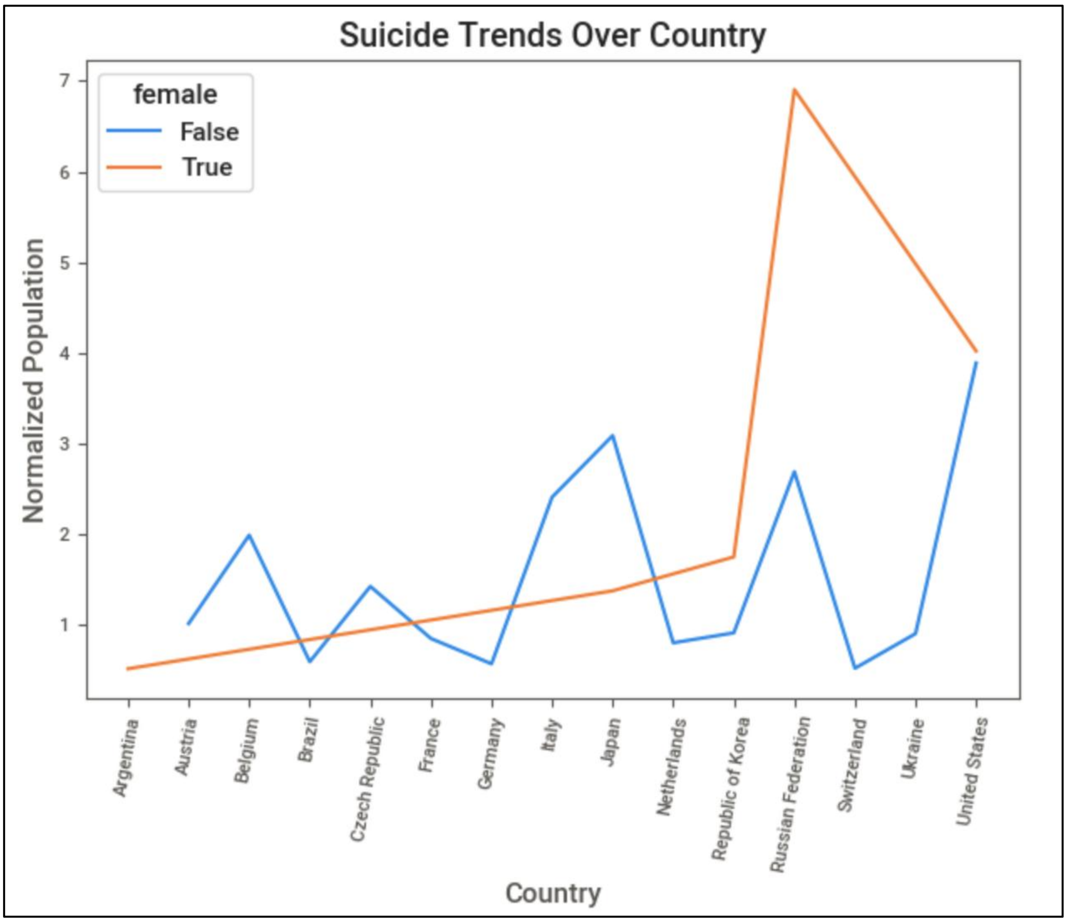
Task 3

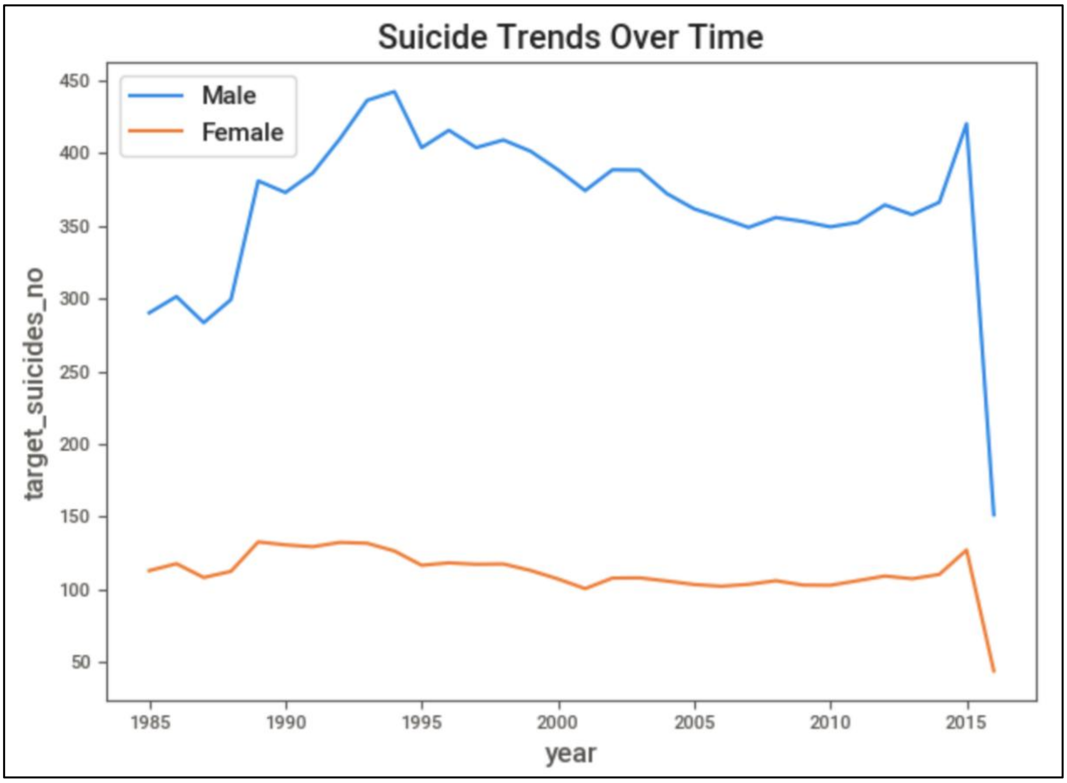
Perform exploratory analysis to answer questions. Design at least three questions by yourself, and use the data to answer. For example:

1. Socio-economic Factors and Their Relation to Suicide Rates: Analyze the relationship between suicide rates and economic indicators in the dataset, such as GDP and GDP per Capita.
2. Geographical Analysis of Suicide Rates: Combine the dataset with other sources of geographical information to determine if there's a correlation between suicide rates and factors like geography, climate, or culture.

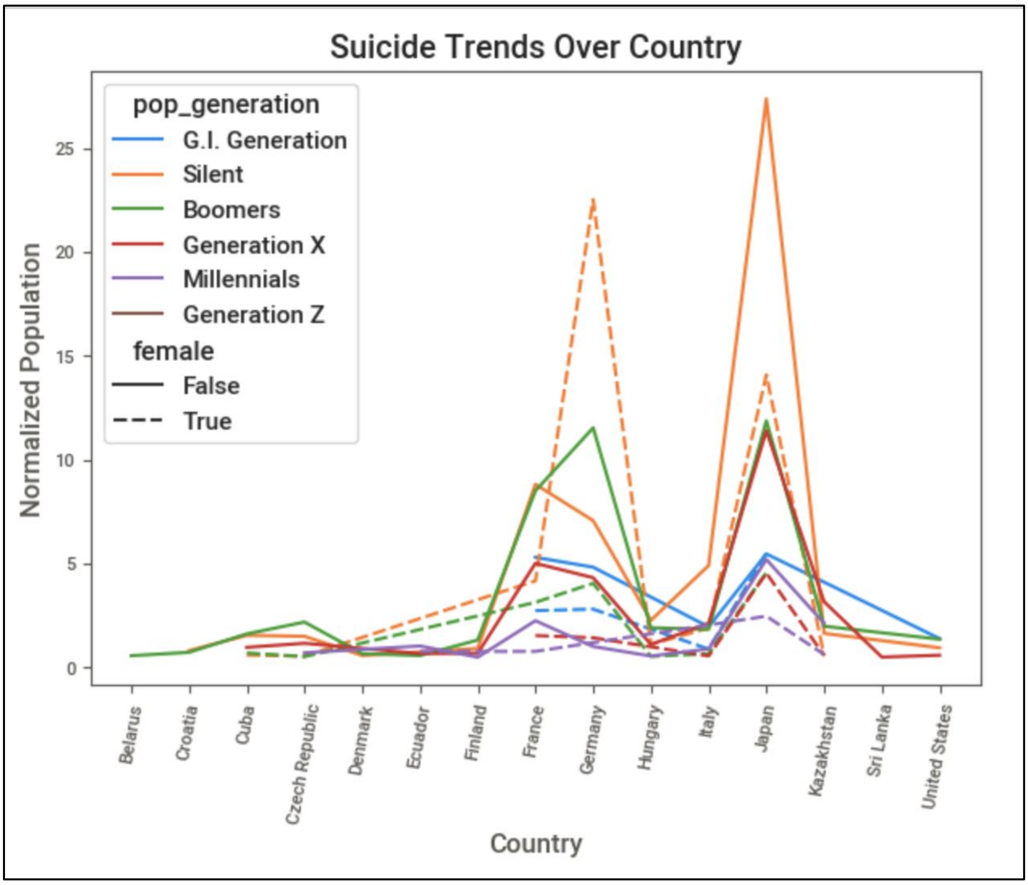
1. How has gender affected the rate of suicide?

The below graph shows the trend which clearly shows that gender played a major role in the rate of suicides in the countries where the rate of suicide was higher among the rest of the members. The Population is normalized by dividing the total rate of suicides by the total population for each country and gender combination.

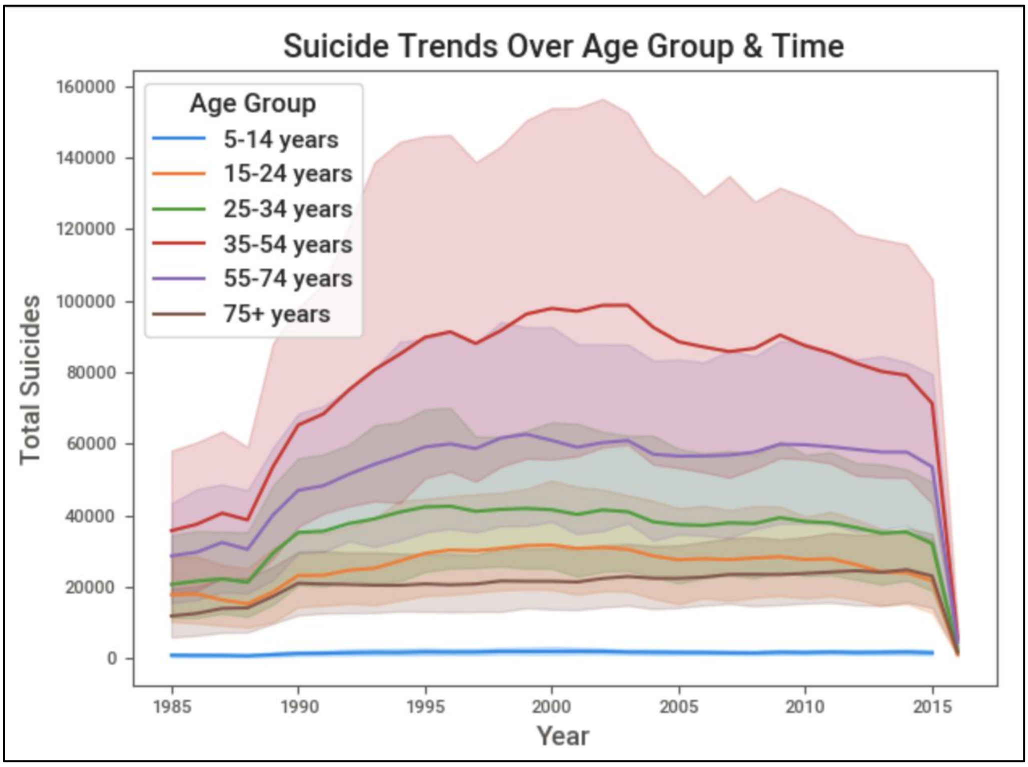




2. How have gender and generation affected the rate of suicide?
The below graph shows the trend which clearly shows that gender coupled with generation played a major role in the rate of suicides in the countries where the rate of suicide was higher among the rest of the members. The Population is normalized by dividing the total rate of suicides by the total population for each country, gender, and generation combination. It shows that the Boomers generation of Males and the Silent generation of Females were among the highest contributors to the total rate of suicide. In Japan, it was the Silent generation of Males with the highest amount of suicides.



3. What was the suicide trend based on age group over a period of 31 years (1985 - 2016)?
The area graph shows the age group and the rate of suicide over a period of 31 years. The age group "31-54" years was the highest of all time with "55-74" being the second highest. There was a steady drop in the rate of suicide after the year 2003 from all the age groups except the "75+" years group as this group was on a uniform path without any dip. The last dip in 2016 seems unnatural as there is a high chance that the data is not completely populated as it was for other years.



Task 4

Explore biases in this dataset. Raise at least three concerns related to bias, and write summaries of your findings. For example:
1. Demographic Bias: All age groups, genders, and other demographics within countries should be represented equitably.
2. Data Collection Methods Bias: If different countries or regions use different methods or standards for data collection, this can lead to inconsistencies or biases in the reported figures.

1. Bias based on Age Group -> The dataset may have age group bias if certain age groups are overrepresented or underrepresented, leading to potential inaccuracies in suicide rate calculations. For example, there are relatively fewer data points for the “5-14 years” age group compared to the “35-54 years” group. This imbalance could lead to biased conclusions when analyzing age-specific suicide rates.

2. Bais based on Gender -> The dataset may have gender bias if there is an unequal representation of males and females or overrepresentation or underrepresentation of a specific gender, leading to potential inaccuracies in suicide rate calculations. For example, there are relatively more male suicides populated than female suicides. This imbalance could lead to biased conclusions when analyzing gender-specific suicide rates.

3. Bias based on Country-Level Reporting -> Based on data collection and reporting practices across different countries there can be a possibility of country bias as some countries might have more robust and accurate reporting systems, while others may have less reliable data management systems.

Additional Features could have been added to the dataset for an in-depth analysis of this critical issue. Some of the additional features are given below
1. Socioeconomic Status
2. Economic Status
3. Urban / Sub-Urban / Rural Status
4. Ethnicity
5. Marital Status
6. Profession / Occupation
7. Education
8. Number of members in the household
9. Number of dependents in the household
10. Month / Yearly Income
11. Month / Yearly Expenditure
12. Credit Score