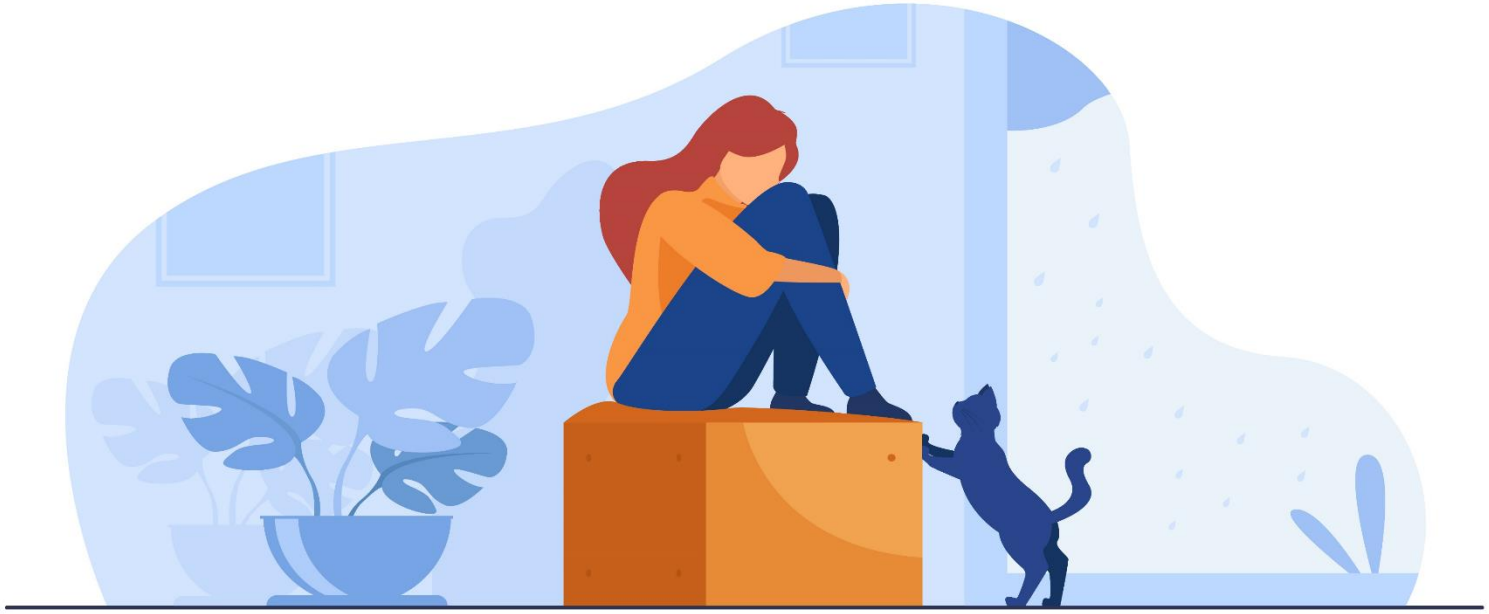
ASSIGNMENT
DATA ANALYSIS
& LINEAR REGRESSION

TUYEN NGUYEN
STUDENT ID: BS19BDS001

TOPIC CHOSEN

Suicide is the act of intentionally **causing one's own death**. Mental disorders, including depression, bipolar disorder, schizophrenia, personality disorders, anxiety disorders, and substance abuse—including alcoholism and the use of benzodiazepines—are risk factors. *(Cambridge Dictionary)*



*Close to **800 000** people die due to suicide every year

*One person dies **every 40 seconds**

*World suicide situation has become **worse than ever**

ABSTRACT

The project is not only to *get to deep understanding about the trend in suicide issue* but also to *figure out the way to prevent the increase in Suicide Rate* – one of hundred hard challenges of the modern world nowadays. In terms of exploring the dataset, with the helping hand of IBM application, I take the descriptive statistics of each field (records about minimum, maximum, the average number and the amount of variation or dispersion of a set of values about the number of suicide as well as its rate in 100,000 citizens.) Then figure out the prediction based on linear relationship between No of suicide/Suicide Rate with Population, GDP and HDI.

$$\text{Suicide Rate (cases per 100k people)} = \frac{Suicidal\ cases\ confirmed * 100,000}{recent\ population}$$

The following project is conducted on a dataset compiled by Nguyen Le Kim Tuyen with data retrieved about how the suicidal situation happening in 5 countries along with many factors that might affect the suicide rate as GDP, HDI, Population, Year, Age Group, Sex...etc. There exist a total of 300 rows and 12 columns. (300 datapoints)

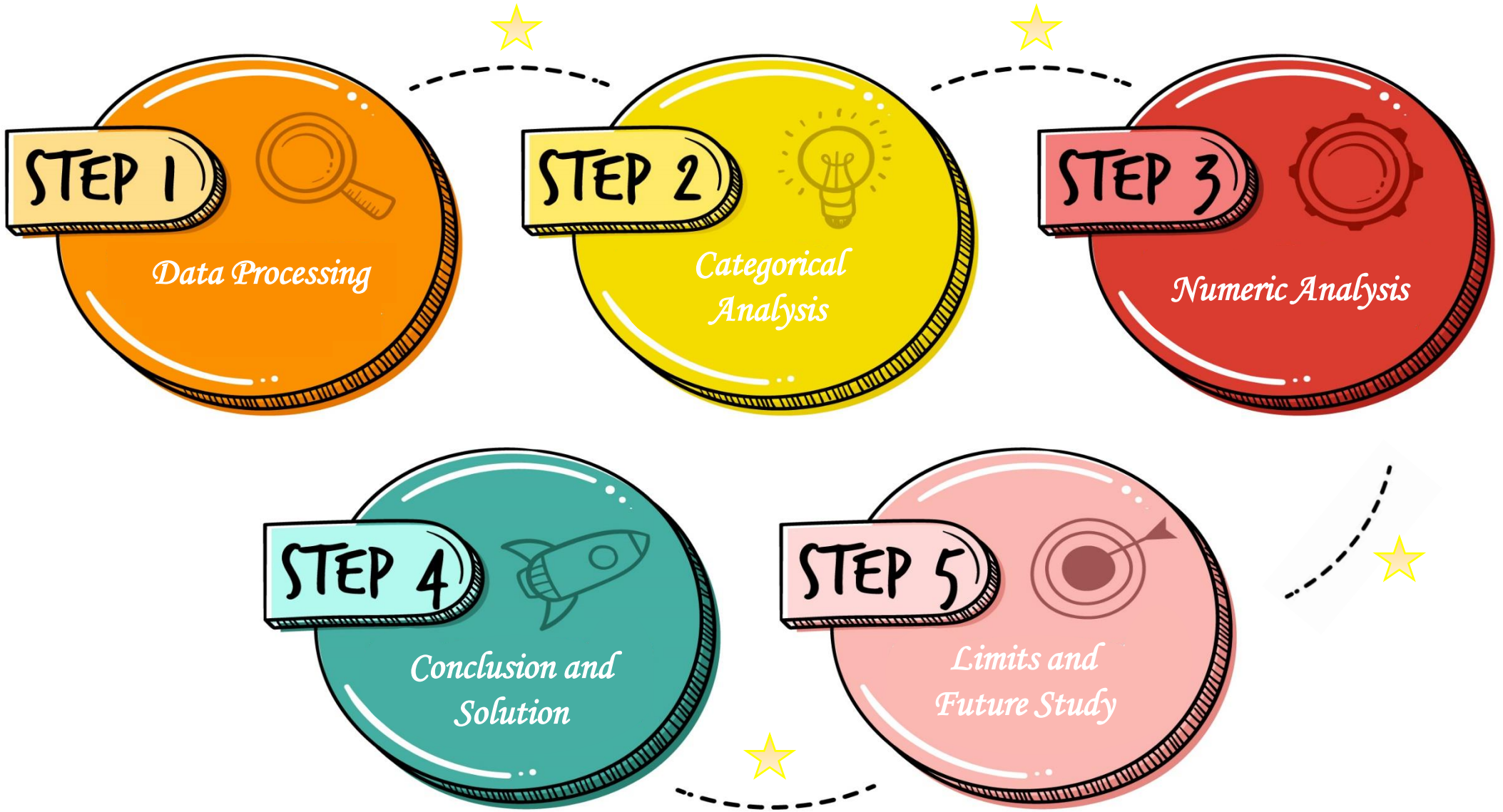
ACKNOWLEDGEMENT

I want to give out my special thanks to many sources that contribute to the success of this project, as this compiled dataset pulled from four other datasets linked by time and place, and was built to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum.
The list of references is shown below:

1) United Nations Development Program. (2018). Human development index (HDI). Retrieved from http://hdr.undp.org/en/indicators/137506
2) World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016. Retrieved from http://databank.worldbank.org/data/source/world-development-indicators#
[Szamil]. (2017).
3) Suicide in the Twenty-First Century [dataset]. Retrieved from https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook
4) World Health Organization. (2018). Suicide prevention. Retrieved from http://www.who.int/mental_health/suicide-prevention/en/

Disclaimer: all the artwork used in this presentation do not belong to me, I want to retribute Freepik Organization as they give out those free designed templates and stocks.

PROJECT STRUCTURE

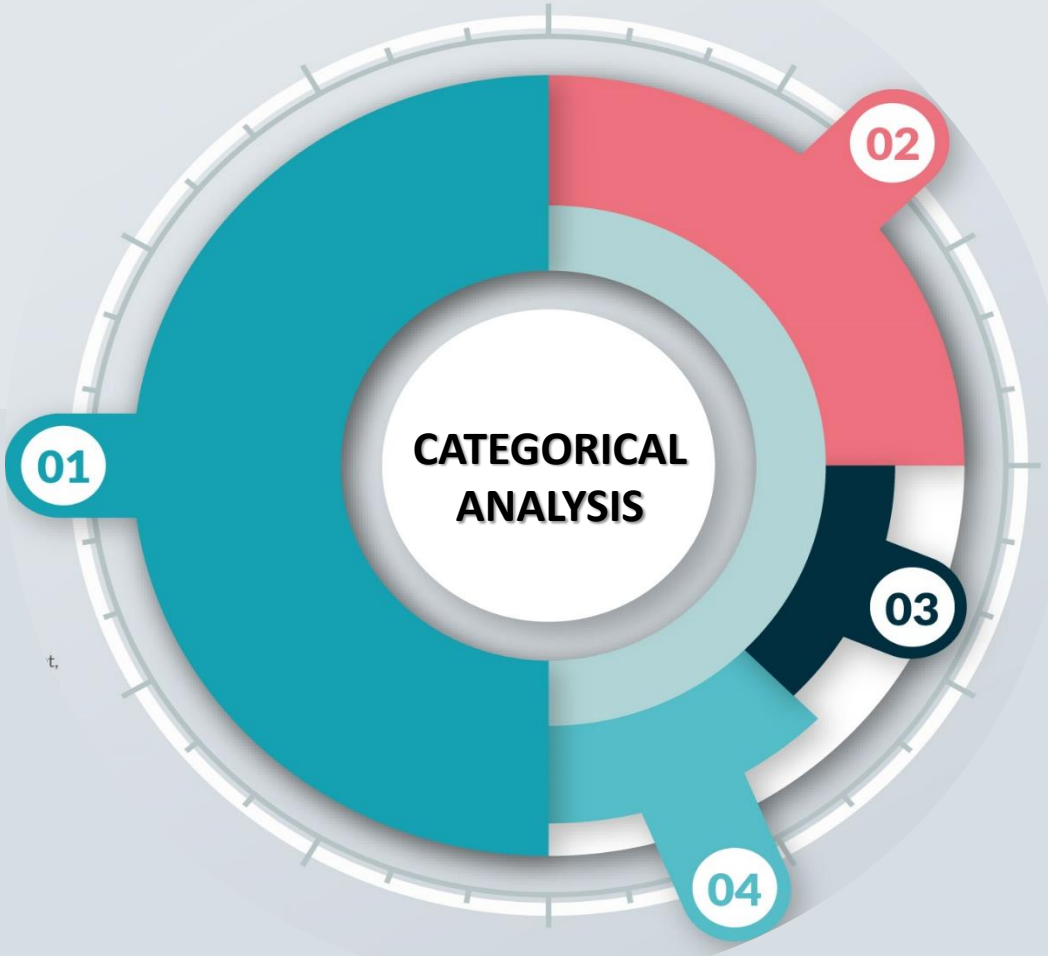


DATA PROCESSING

- The original dataset has total more than 2500 datapoints. Since there are a lot of data points at which, the HDI is missing, I clear them out. And in order to build a good prediction model for research, I reduced the size of dataset into 300 retrieved from: 6 groups of age, 2 groups of gender, 5 observing years (2010-2014) in 5 representative countries from 5 continents: Africa, America, Australia, Asia, Europe. *(6*2*5*5=300)*

CATEGORICAL ANALYSIS

AGE GROUP

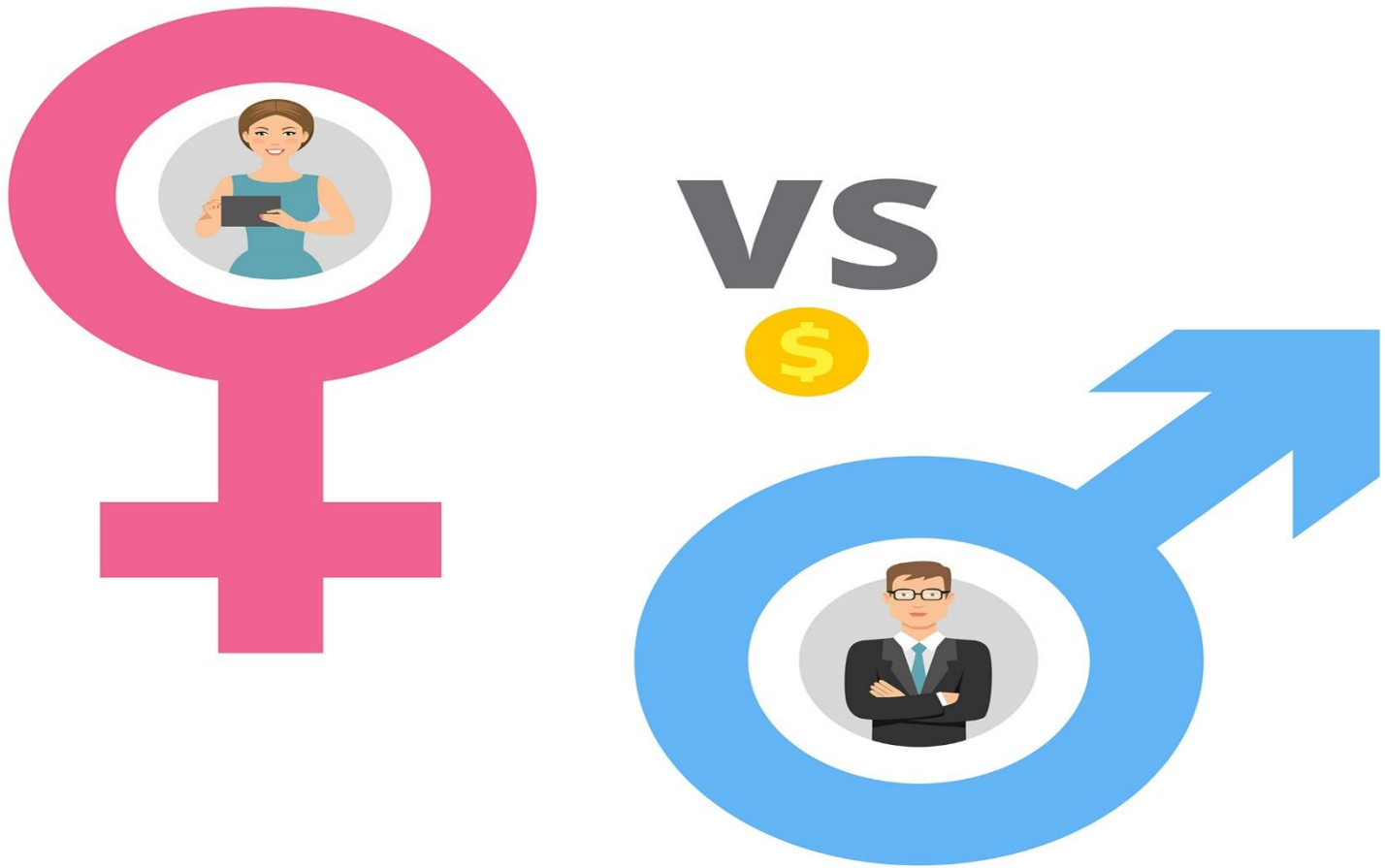


GENDER

YEARS

GENERATIONS

GENDER



FEMALE

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	150	0	2883	315.88	682.723
suicides/100k pop	150	.00	20.75	5.3443	5.70089
Valid N (listwise)	150				

The range of no of cases committed suicide is 2883 cases whereas this figure of suicide rate is 20.75.

Regularly, in 100.000 people, there are approximately 5.3 women committed suicide regardless of age, generation or year.

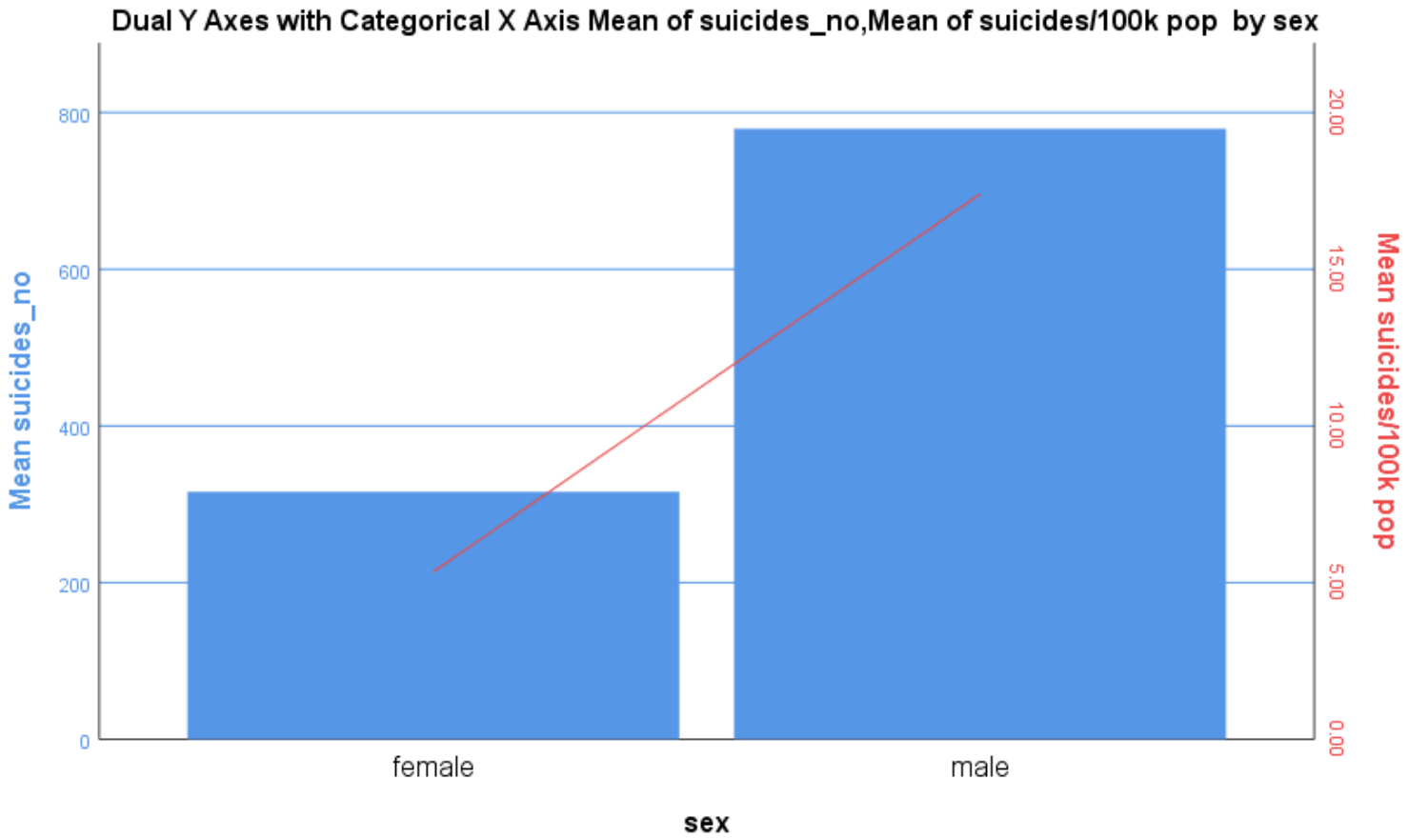
MALE

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	150	0	7466	779.39	1637.734
suicides/100k pop	150	.00	86.69	17.3979	17.46981
Valid N (listwise)	150				

The range of no of cases committed suicide is 7466 cases whereas this figure of suicide rate is 86.69.

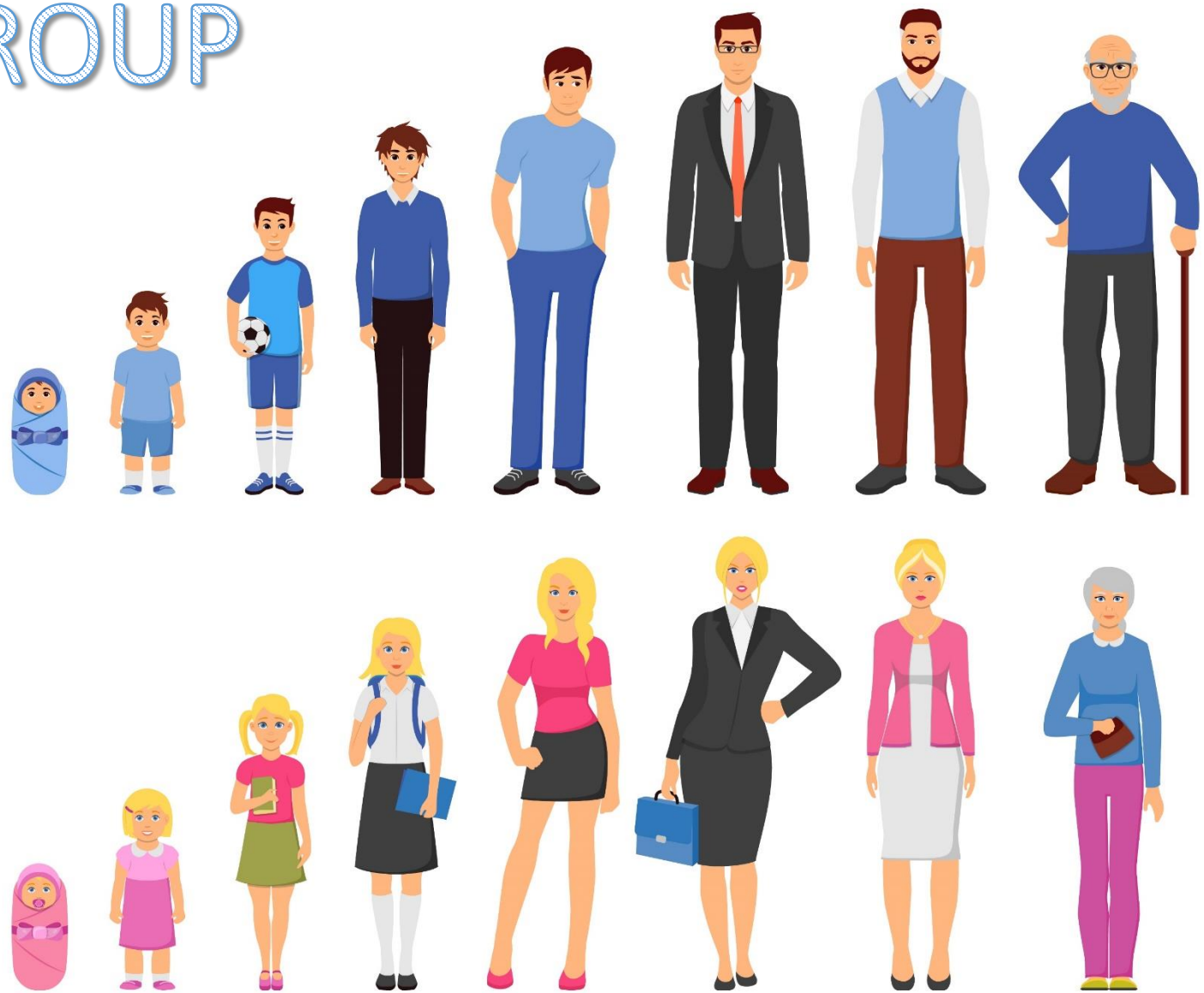
Regularly, in 100.000 people, there are approximately 17.4 men committed suicide regardless of age, generation or year.

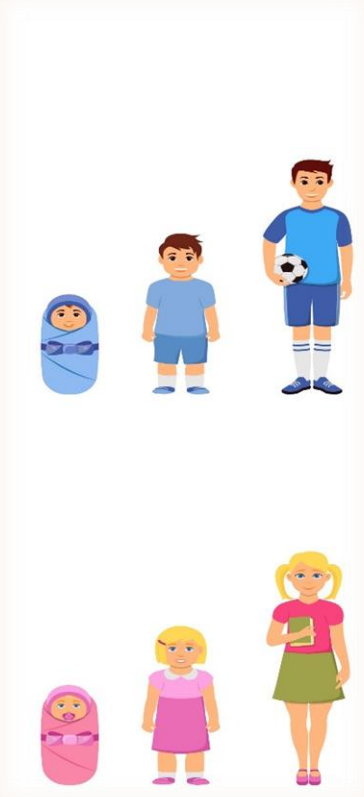
COMPARISON IN TOTAL



From the graph, we could see that both no of suicide cases and its rate in male are 3 times bigger than females.

AGE GROUP



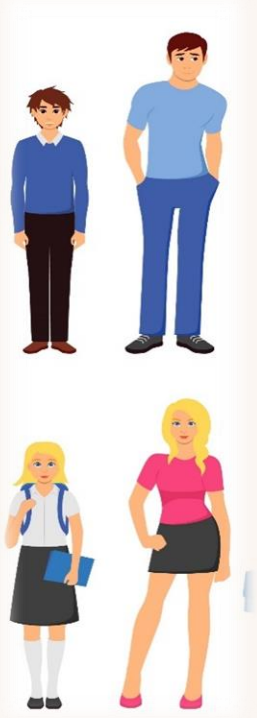


05-14

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	50	0	68	18.46	18.860
suicides/100k pop	50	.00	1.37	.5508	.38767
Valid N (listwise)	50				

The range of no of cases committed suicide is 68 cases whereas this figure of suicide rate is 1.37.

Regularly, in 100.000 people, there are approximately 0.55 individual committed suicide regardless of gender, generation or year.



15-24

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	50	5	1315	291.26	370.254
suicides/100k pop	50	.63	20.92	8.5060	6.30239
Valid N (listwise)	50				

The range of no of cases committed suicide is 1310 cases whereas this figure of suicide rate is 20.29.

Regularly, in 100.000 people, there are approximately 8.5 individuals committed suicide regardless of gender, generation or year.

25-34

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	50	7	2543	438.10	686.230
suicides/100k pop	50	.33	32.51	10.7444	9.58149
Valid N (listwise)	50				

The range of no of cases committed suicide is 2536 cases whereas this figure of suicide rate is 32.18.

Regularly, in 100.000 people, there are approximately 10.7 individuals committed suicide regardless of gender, generation or year.

35-54

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	50	13	7351	1083.58	1968.528
suicides/100k pop	50	.22	42.91	13.4474	12.67476
Valid N (listwise)	50				

The range of no of cases committed suicide is 7338 cases whereas this figure of suicide rate is 42.69.

Regularly, in 100.000 people, there are approximately 13.4 individuals committed suicide regardless of gender, generation or year.

55-74

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	50	7	7466	1019.96	1976.281
suicides/100k pop	50	.26	46.55	14.6166	14.45358
Valid N (listwise)	50				

The range of no of cases committed suicide is 7459 cases whereas this figure of suicide rate is 46.29.

Regularly, in 100.000 people, there are approximately 14.6 individuals committed suicide regardless of gender, generation or year.

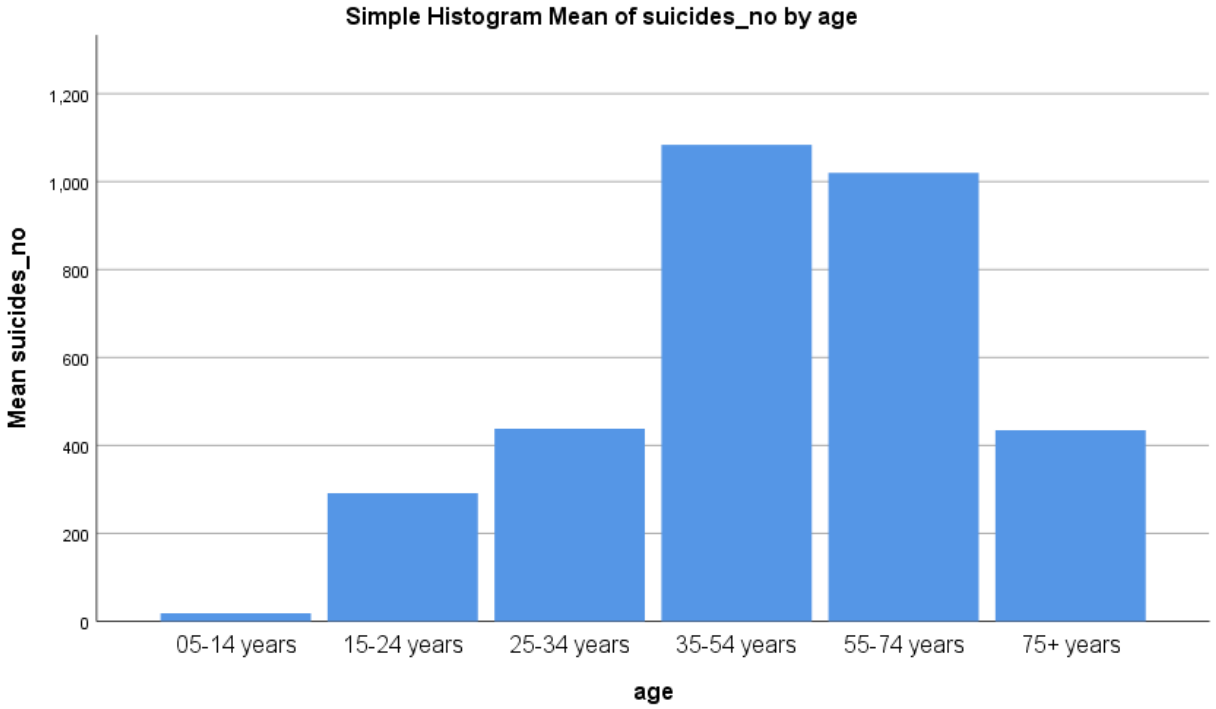
75+

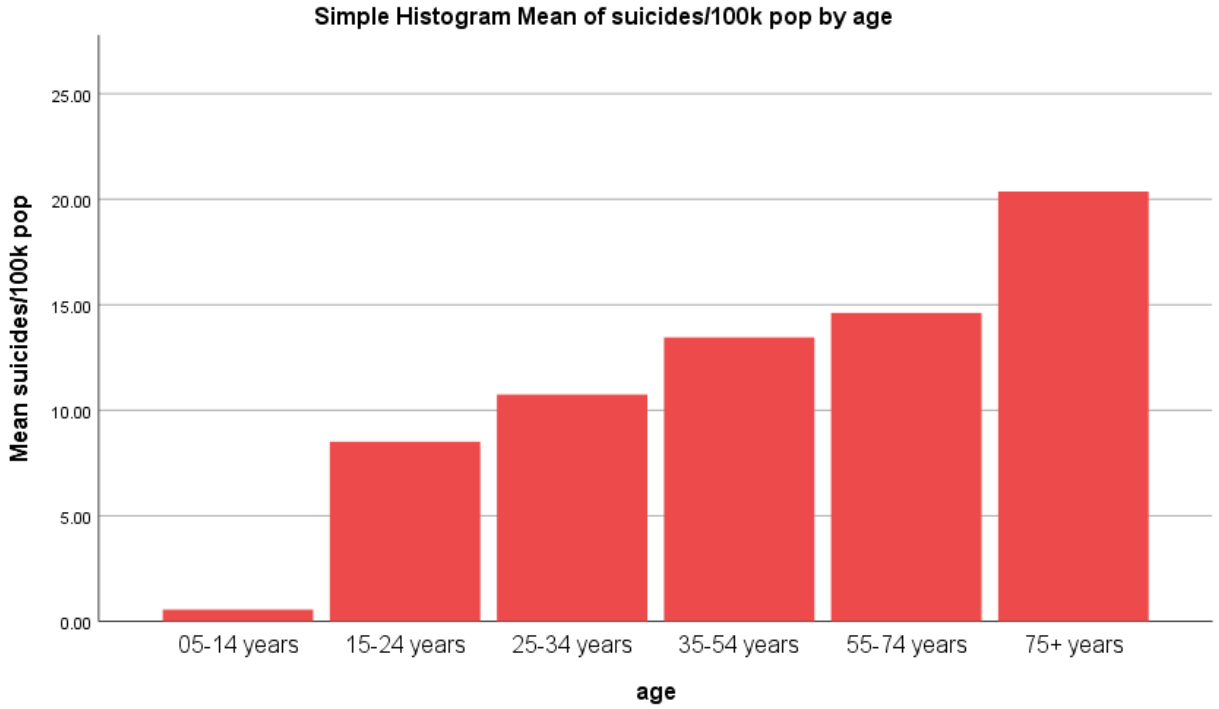
Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	50	1	2362	434.46	783.701
suicides/100k pop	50	.18	86.69	20.3614	22.85056
Valid N (listwise)	50				

The range of no of cases committed suicide is 2361 cases whereas this figure of suicide rate is 86.51.

Regularly, in 100.000 people, there are approximately 20.4 individuals committed suicide regardless of gender, generation or year.

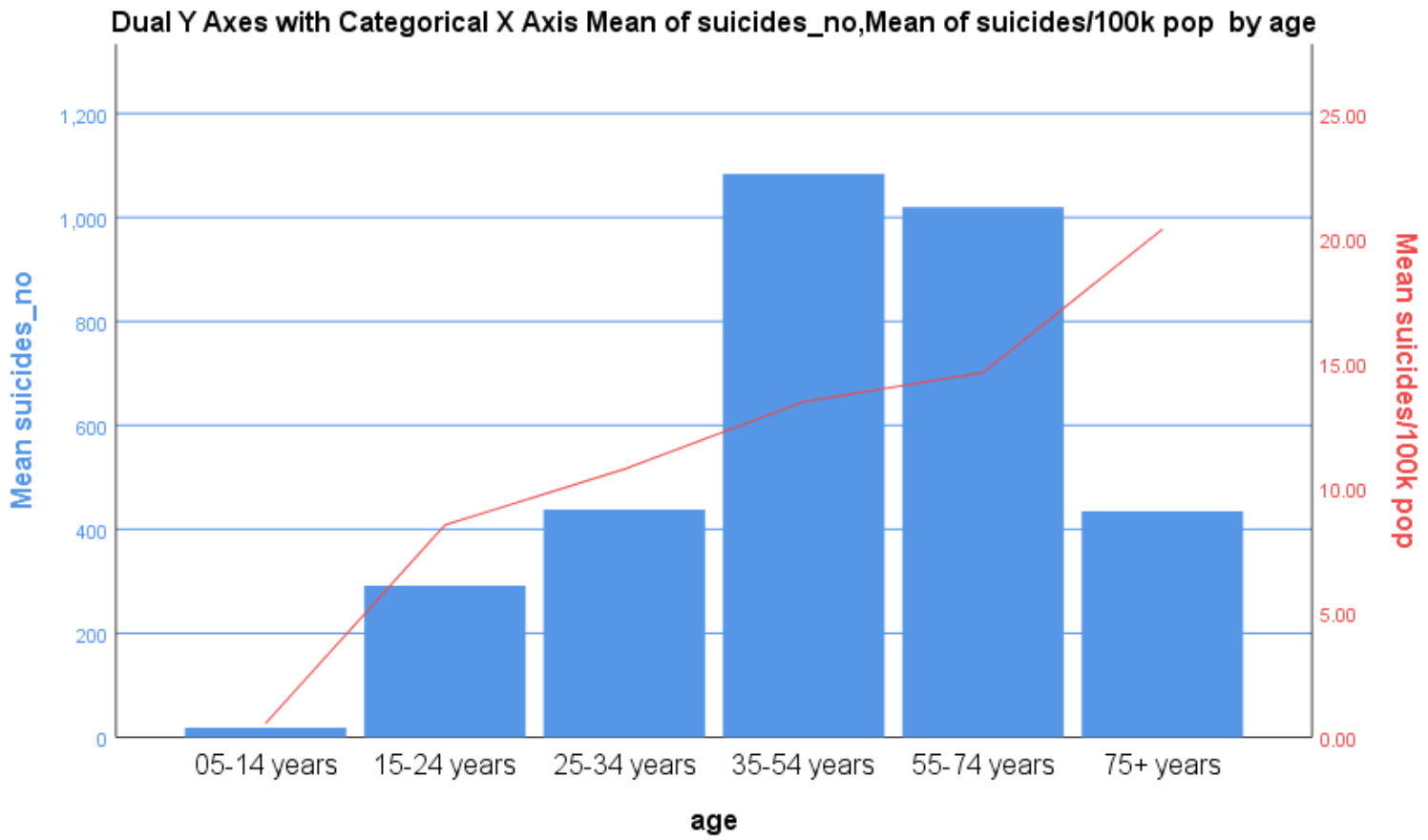
COMPARISON IN TOTAL





In those 2 below histograms, it seems like the trend in suicide recorded takes the normal distribution, where the group of tricenarian to mid-quinquagenarian is more likely to commit suicide than other age groups. However, the trend in rate of suicide illustrates that the chance of getting suicide increase regarding ages. The older people get, the probability to commit suicide gets higher.

COMPARISON IN TOTAL



(This graph here is combined from 2 previous histograms, which helps us understand the trends better.)

YEARS
2010 - 2014

2010

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	60	0	7466	585.12	1439.870
suicides/100k pop	60	.00	86.69	11.8403	15.72584
Valid N (listwise)	60				

The range of no of cases committed suicide is 7466 cases whereas this figure of suicide rate is 86.69.

Regularly, in 100.000 people, there are approximately 11.8 individuals committed suicide regardless of age, gender, generation or year.

2011

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	60	0	6975	571.03	1364.534
suicides/100k pop	60	.00	80.49	11.3155	14.62308
Valid N (listwise)	60				

The range of no of cases committed suicide is 6975 cases whereas this figure of suicide rate is 80.49.

Regularly, in 100.000 people, there are approximately 11.31 individuals committed suicide regardless of age, gender, generation or year.

2012

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	60	1	6413	538.33	1240.941
suicides/100k pop	60	.13	80.75	11.5382	14.65738
Valid N (listwise)	60				

The range of no of cases committed suicide is 6412 cases whereas this figure of suicide rate is 80.62.

Regularly, in 100.000 people, there are approximately 11.54 individuals committed suicide regardless of age, gender, generation or year.

2013

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	60	0	6333	532.43	1217.393
suicides/100k pop	60	.00	67.38	10.9778	13.22691
Valid N (listwise)	60				

The range of no of cases committed suicide is 6333 cases whereas this figure of suicide rate is 67.38.

Regularly, in 100,000 people, there are approximately 10.98 individuals committed suicide regardless of age, gender, generation or year.

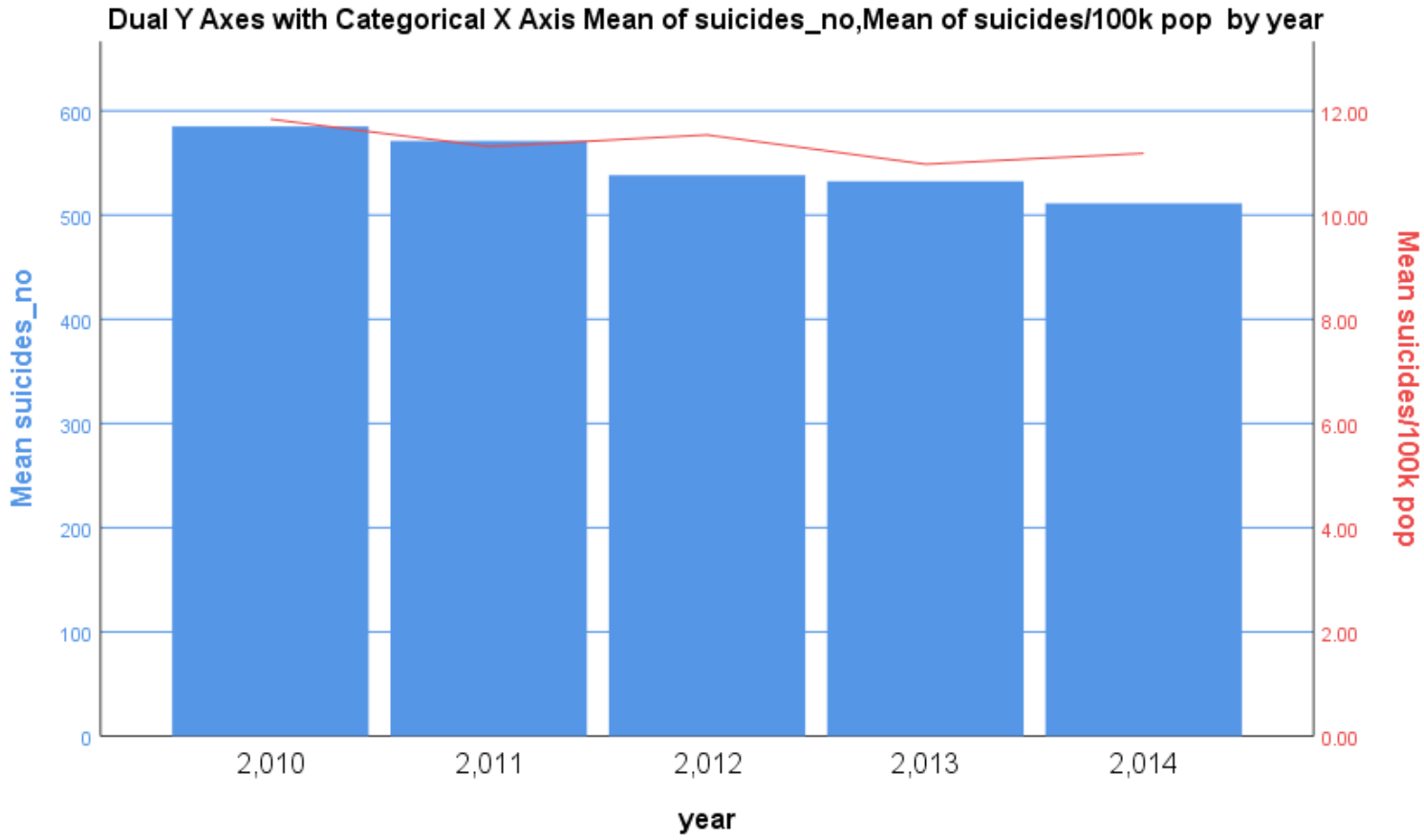
2014

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	60	0	5710	511.27	1123.853
suicides/100k pop	60	.00	75.07	11.1837	13.64311
Valid N (listwise)	60				

The range of no of cases committed suicide is 5710 cases whereas this figure of suicide rate is 75.07.

Regularly, in 100,000 people, there are approximately 11.18 individuals committed suicide regardless of age, gender, generation or year.

COMPARISON IN TOTAL



The graph below illustrates the slowly decrease in number of suicide cases each year from 2010 to 2014. However, the suicide rate variates through years.

GENERATIONS

Silent (The cohort is defined as individuals born between 1928 and 1945)

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	60	1	7466	551.77	1204.092
suicides/100k pop	60	.18	86.69	19.5852	21.95093
Valid N (listwise)	60				

The range of no of cases committed suicide is 7465 cases whereas this figure of suicide rate is 86.51.

Regularly, in 100.000 people, there are approximately 19.58 individuals committed suicide regardless of age, gender or year.

Boomers (The generation is most often defined as individuals born between 1946 and 1964, during the post–World War II baby boom)

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	40	7	6858	990.38	1893.894
suicides/100k pop	40	.26	44.46	14.3448	13.92182
Valid N (listwise)	40				

The range of no of cases committed suicide is 6851 cases whereas this figure of suicide rate is 44.2.

Regularly, in 100.000 people, there are approximately 14.34 individuals committed suicide regardless of age, gender or year.

Generation X (Researchers and popular media typically use birth years around 1965 to 1980 to define Generation Xers)

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	60	7	7351	980.43	1835.205
suicides/100k pop	60	.22	42.91	12.9345	12.25593
Valid N (listwise)	60				

The range of no of cases committed suicide is 7344 cases whereas this figure of suicide rate is 42.69.

Regularly, in 100.000 people, there are approximately 12.93 individuals committed suicide regardless of age, gender or year.

Millennials(also known as Generation Y, are the Researchers and popular media use the early 1980s as starting birth years and the mid-1990s to early 2000s as ending birth years, with 1981 to 1996 a widely accepted defining range for the generation.)

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	90	5	2404	353.57	525.540
suicides/100k pop	90	.33	31.68	9.5424	7.96925
Valid N (listwise)	90				

The range of no of cases committed suicide is 2439 cases whereas this figure of suicide rate is 31.35.

Regularly, in 100.000 people, there are approximately 9.54 individuals committed suicide regardless of age, gender or year.

Generation Z(Researchers and popular media use the mid-to-late 1990s as starting birth years and the early 2010s as ending birth years. Birthyear: 1997-2012)

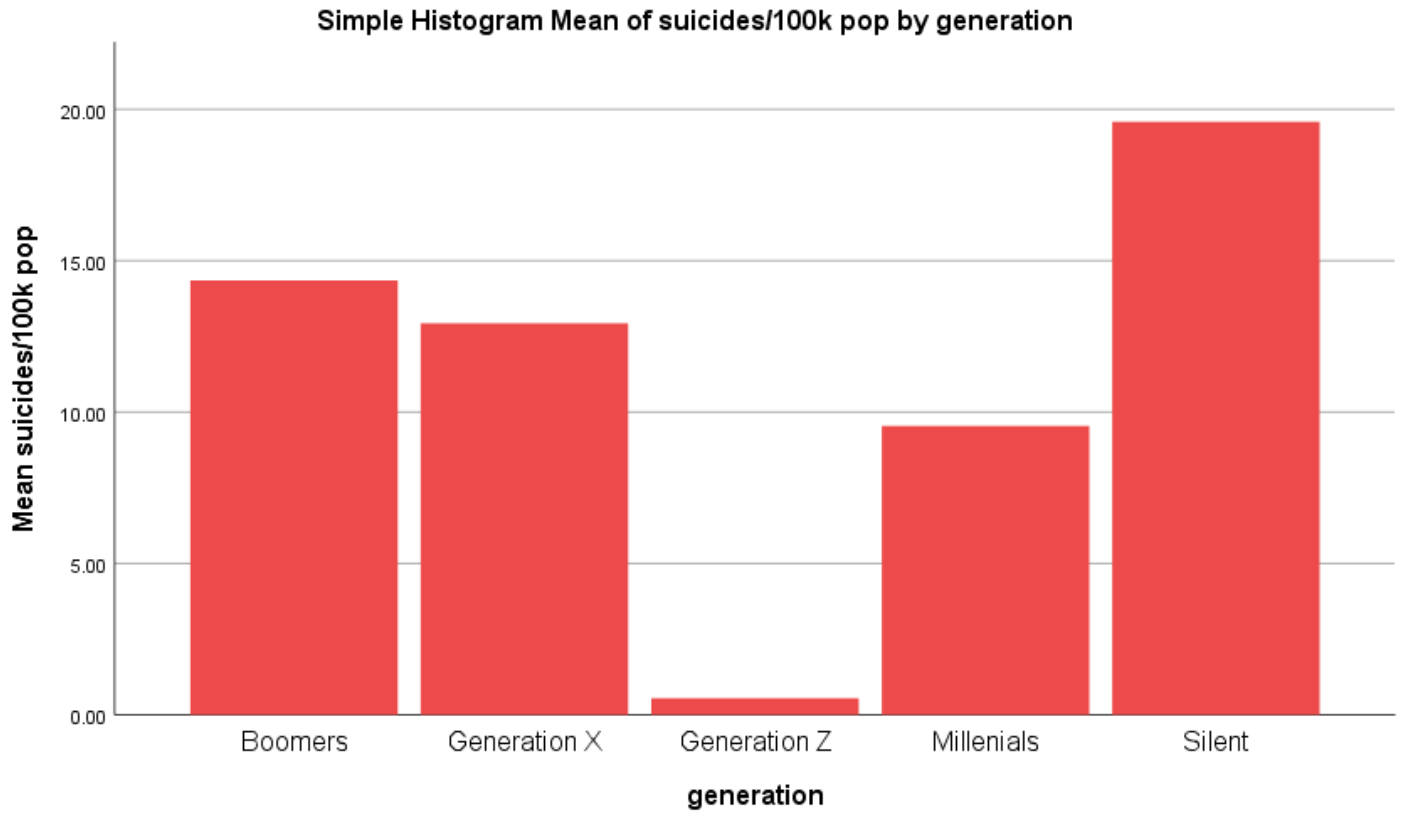
Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	50	0	68	18.46	18.860
suicides/100k pop	50	.00	1.37	.5508	.38767
Valid N (listwise)	50				

The range of no of cases committed suicide is 68 cases whereas this figure of suicide rate is 1.37.

Regularly, in 100,000 people, there are approximately 0.55 individuals committed suicide regardless of age, gender or year.

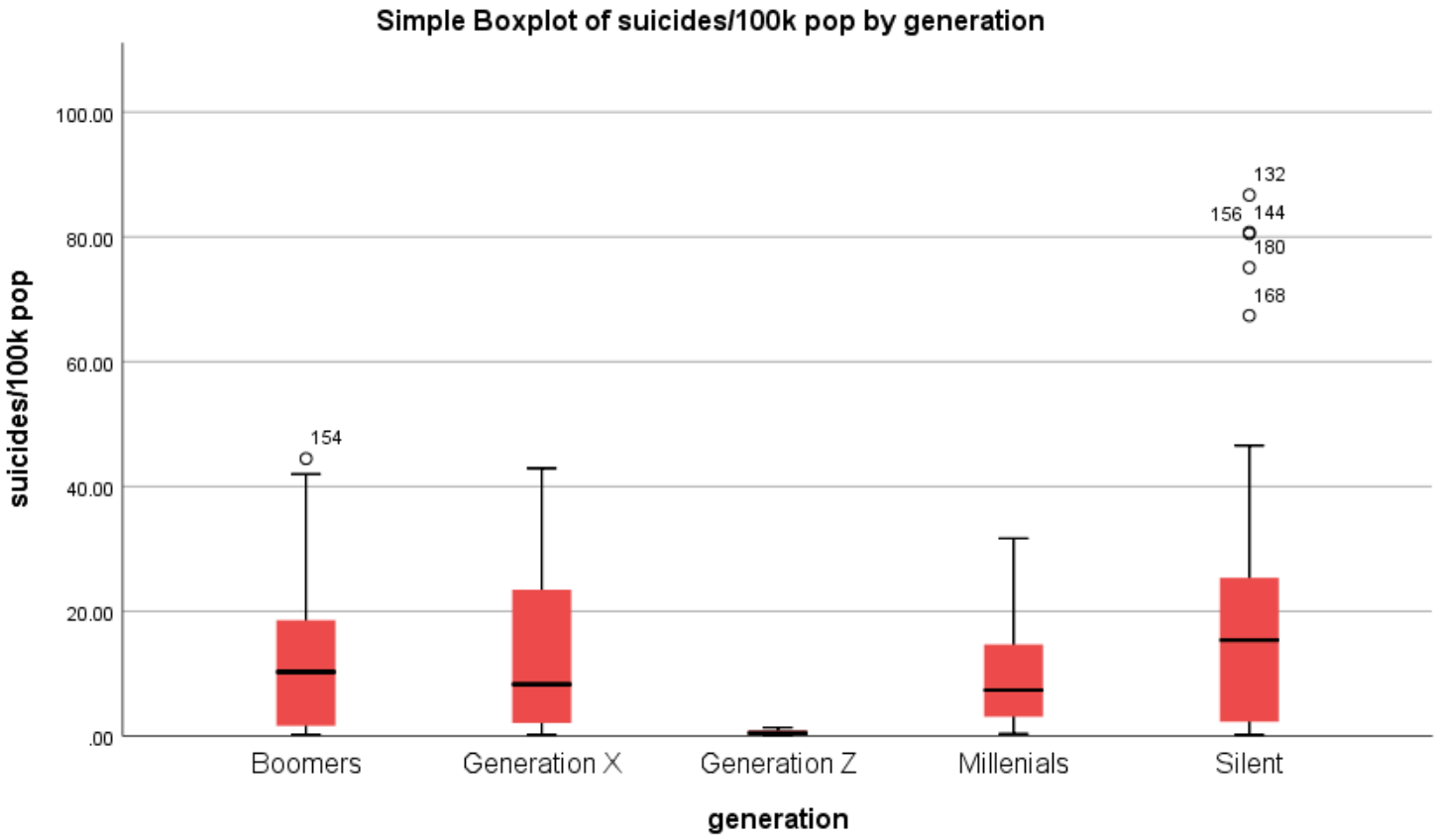
COMPARISON IN TOTAL

As the number of citizens committing suicide in each generation is not equal, we only take the visualization of suicide rate. From the histogram below, we can see people from Silent generation have the highest chance of killing themselves than later generations.



COMPARISON IN TOTAL

However, when we consider the boxplot of the same values, we can see many mild outliers in figure for silent generation. If we remove them all, the trend are mostly similar to X generation's citizens.



NUMERIC ANALYSIS

5 REPRESENTATIVE COUNTRIES

Australia (Australia)

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	60	4	866	214.85	214.660
suicides/100k pop	60	.30	27.66	11.5547	8.86977
Valid N (listwise)	60				

The range of no of cases committed suicide is 862 cases whereas this figure of suicide rate is 27.36.

Regularly, in 100.000 people, there are approximately 11.6 individuals committed suicide regardless of age, gender, generation or year.

Colombia (America)

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	60	4	498	174.40	172.126
suicides/100k pop	60	.59	17.07	5.4453	4.98700
Valid N (listwise)	60				

The range of no of cases committed suicide is 494 cases whereas this figure of suicide rate is 16.48.

Regularly, in 100.000 people, there are approximately 5.4 individuals committed suicide regardless of age, gender, generation or year.

Croatia (Europe)

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	60	0	215	61.13	63.408
suicides/100k pop	60	.00	86.69	18.9072	21.74652
Valid N (listwise)	60				

The range of no of cases committed suicide is 215 cases whereas this figure of suicide rate is 86.69.

Regularly, in 100.000 people, there are approximately 18.9 individuals committed suicide regardless of age, gender, generation or year.

Japan (Asia)

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	60	20	7466	2247.72	2108.535
suicides/100k pop	60	.37	45.52	19.8493	13.64576
Valid N (listwise)	60				

The range of no of cases committed suicide is 7446 cases whereas this figure of suicide rate is 45.15.

Regularly, in 100.000 people, there are approximately 19.8 individuals committed suicide regardless of age, gender, generation or year.

South Africa (Africa)

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
suicides_no	60	1	164	40.08	44.573
suicides/100k pop	60	.02	3.60	1.0990	.97285
Valid N (listwise)	60				

The range of no of cases committed suicide is 163 cases whereas this figure of suicide rate is 3.58.

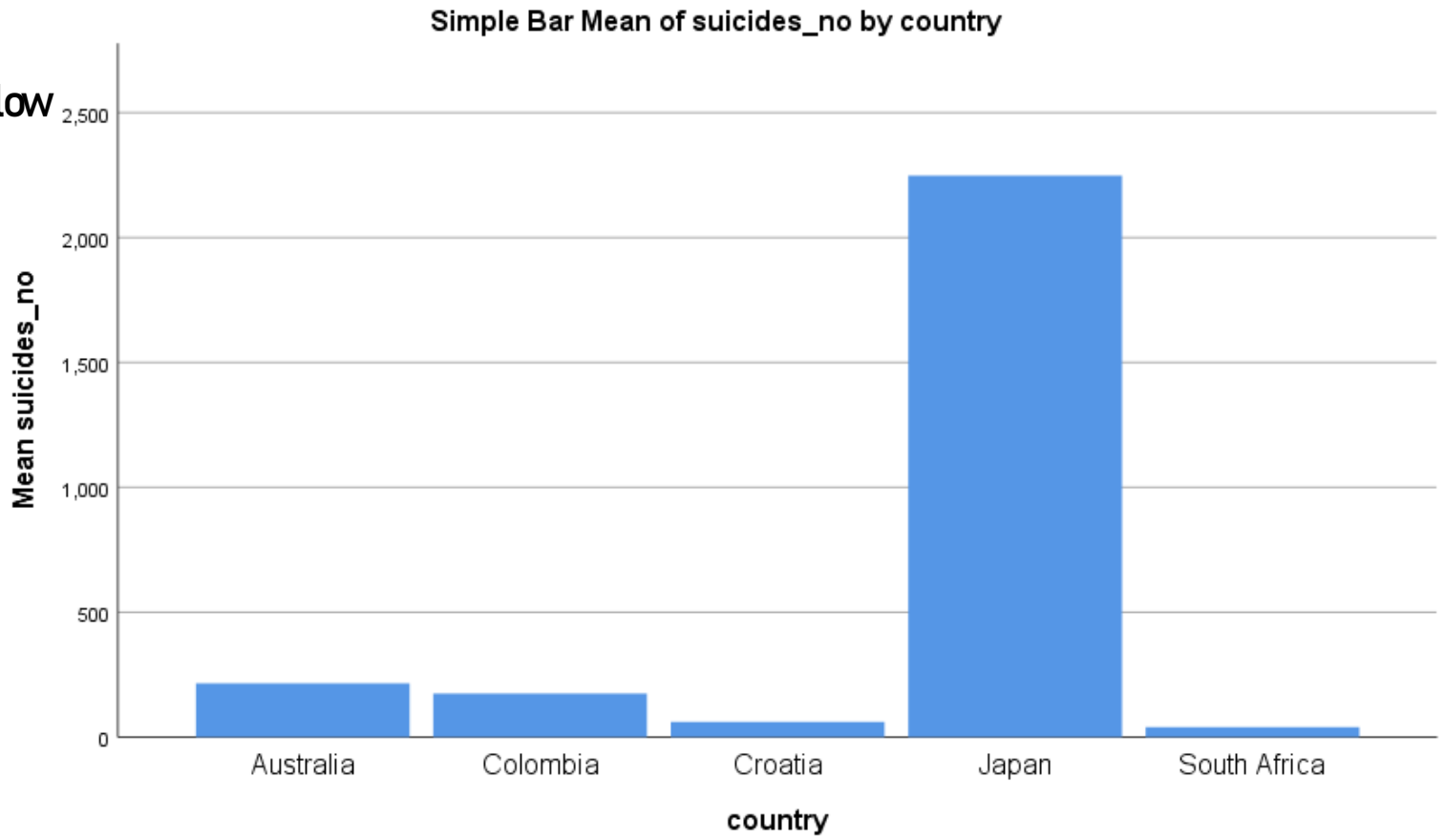
Regularly, in 100.000 people, there are approximately 1 individual committed suicide regardless of age, gender, generation or year.

COMPARISON IN TOTAL

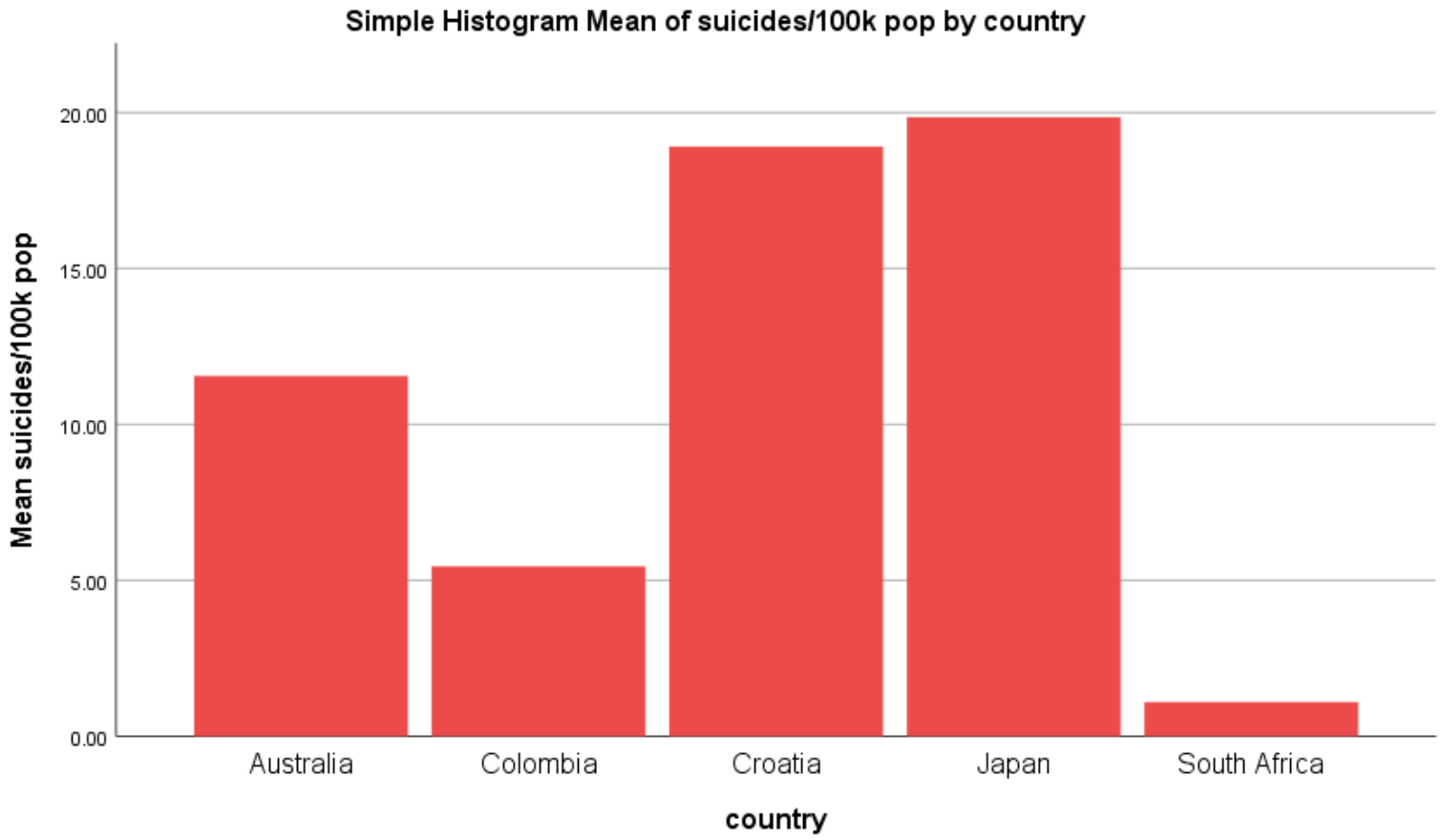
According to the summary we get, whereas the trend witnessed in other countries shows that Croatia did a good job in control their suicide situation (some datapoints have 0 case which leads to rate of suicide is also equal to 0), the country has the lowest rate of suicide in 100,000 people is South Africa with merely 1 suicide.

We can have a brief look at the histogram below

The rank in average number of suicides recorded in 5 countries from least to most is: South Africa, Croatia, Colombia, Australia and Japan. More specifically, all other countries have below 500 cases of suicide during time of research, whereas Japan peak more than 2,000 cases, more than 4 times those of others. It should be noted for the later analysis on HDI and population.



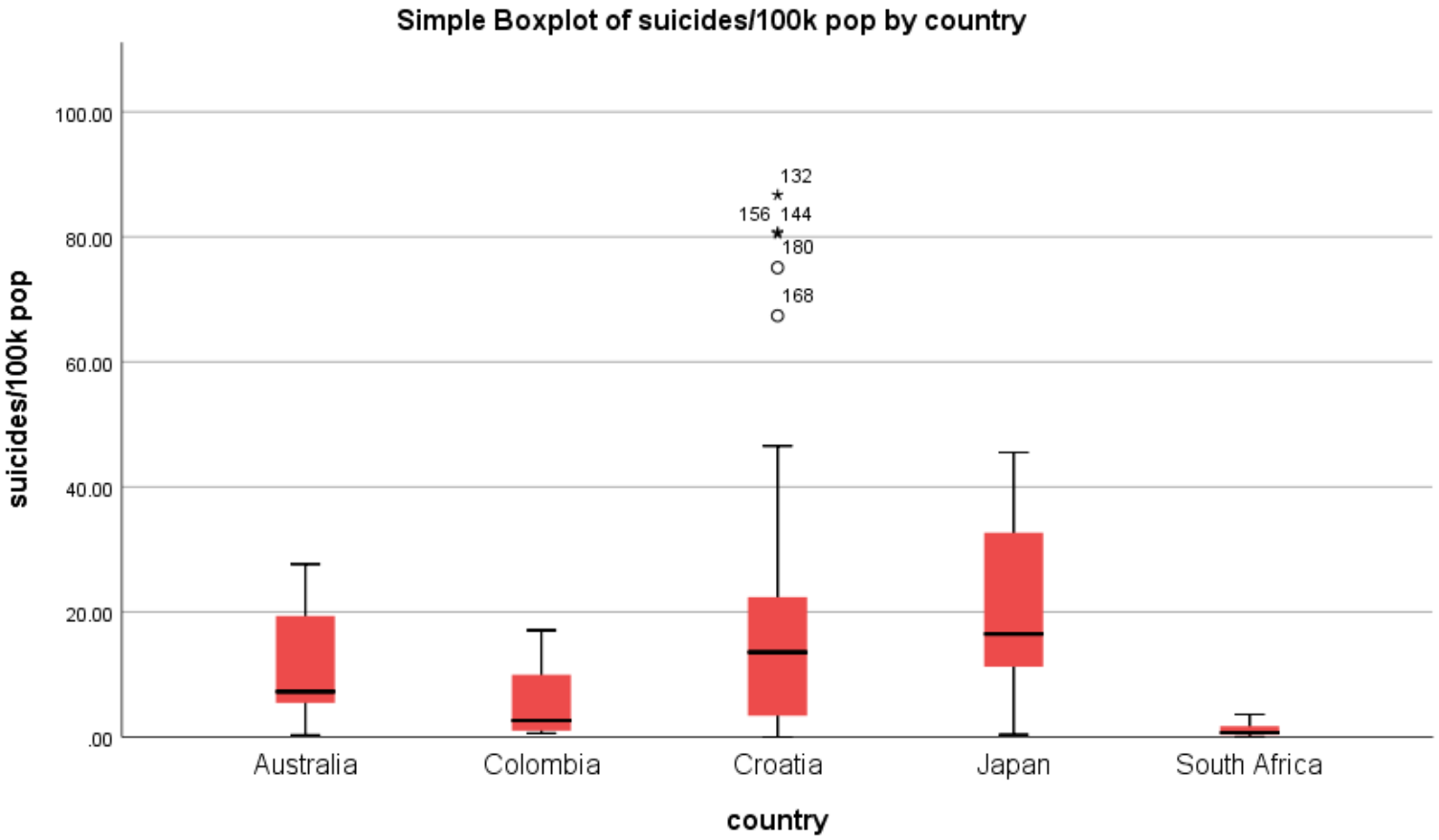
COMPARISON IN TOTAL



However, the rank in average suicidal rate counted in 5 countries from least to most is: South Africa, Colombia, Australia, Croatia and Japan. It's evitable that the rate of suicide in 100,000 people in Croatia is just slightly less than Japan's one. (1 case in 100,000 citizens). What happened in that strange characteristic?

COMPARISON IN TOTAL

As we plot the box plot of suicide rate, we can clearly see it now.



While Croatia is still doing great in dealing with suicide issue, some datapoints (year, generation, age or gender) are remarkably higher than usual trend in this country, which become outliers (2 mild outliers and 3 extreme one).

We have to remove those 5 datapoints to take the better visualization of suicide issue in these 5 countries.

WHAT REALLY DID HAPPEN IN THOSE COUNTRIES?

01

POPULATION

02

GDP/CAPITA

(Gross Domestic Product/head)

03

HDI

(human development index)

LINEAR REGRESSION MODELS
(no of suicide-dependent variables)

No of suicide depends on GDP/capita(X1), population(X2)	
FORMULA	**Y=0.008X1+0.00024X2-612.702**
R SQUARED	66.16%

No of suicide depends on GDP/capita(X1), HDI(X2)	
FORMULA	**Y=0.0016X1+3483.5X2-2287.63**
R SQUARED	9.5%

No of suicide depends on population(X1), HDI(X2)	
FORMULA	**Y=0.00024X1+2796.532X2-2624.15**
R SQUARED	69.02%

No of suicide depends on GDP/capita(X1), population(X2), HDI(X3)	
FORMULA	**Y=-0.014X1+0.000242X2+5725.5X3-4586.28**
R SQUARED	70.4%

BEST LINEAR REGRESSION MODEL

No of suicide depends on GDP/capita(X1), population(X2), HDI(X3)	
FORMULA	**Y=-0.014X1+0.000242X2+5725.5X3-4586.28**
R SQUARED	70.4%

The number of people committing suicide has a positive relationship with number of population and HDI whereas its relationship with GDP/capita is negative.

INDEPENDENT VARIABLE	P-value
GDP/Capita (X1)	0.000224
Population(X2)	8.46E-74
HDI(X3)	2.81E-10

All *p-values < 0.05*
=> all independent variables have a strong relationship with the number of suicide cases

LINEAR REGRESSION MODELS
(rate of suicide-dependent variables)

Rate of suicide depends on GDP/capita(X1), population(X2)	
FORMULA	Y=0.00012X1+(4.5*10^-7)X2+6.12
R SQUARED	6.8%

Rate of suicide depends on GDP/capita(X1), HDI(X2)	
FORMULA	Y=-0.0004X1+128.4433X2-81.176
R SQUARED	22.31%

No of suicide depends on HDI(X1), population(X2)	
FORMULA	Y=51.8X1+(4.33*10^-7)X2-31.82
R SQUARED	16.6%

No of suicide depends on GDP/capita(X1), population(X2), HDI(X3)	
FORMULA	Y=-0.0004X1+(6.06x10^-7)X2+134.05X3-87
R SQUARED	25.32%

BEST LINEAR REGRESSION MODEL

No of suicide depends on GDP/capita(X1), population(X2), HDI(X3)	
FORMULA	**Y=-0.0004X1+(6.06x10^-7)X2+134.05X3-87**
R SQUARED	25.32%

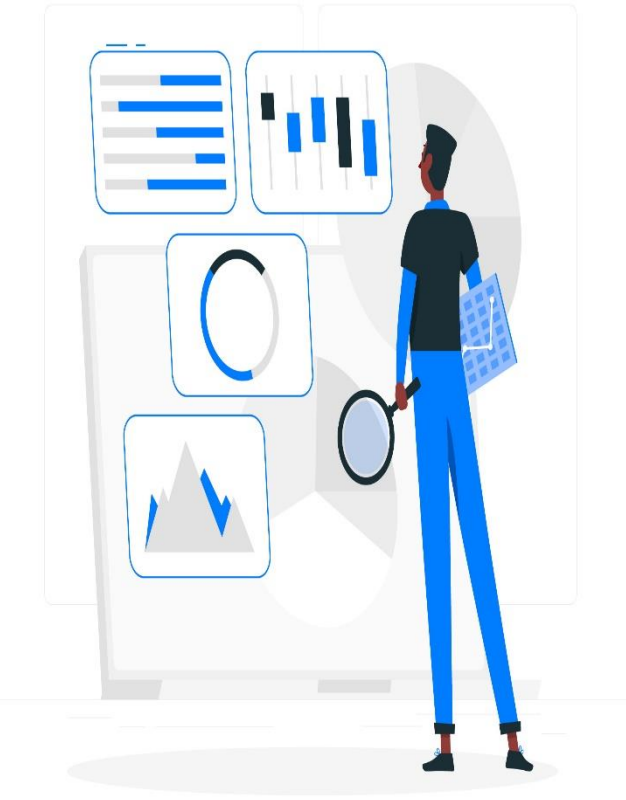
The rate of people committing suicide has a positive relationship with number of population and HDI whereas its relationship with GDP/capita is negative.

INDEPENDENT VARIABLE	P-value
GDP/Capita (X1)	1.12E-08
Population(X2)	5.81E-16
HDI(X3)	0.000625

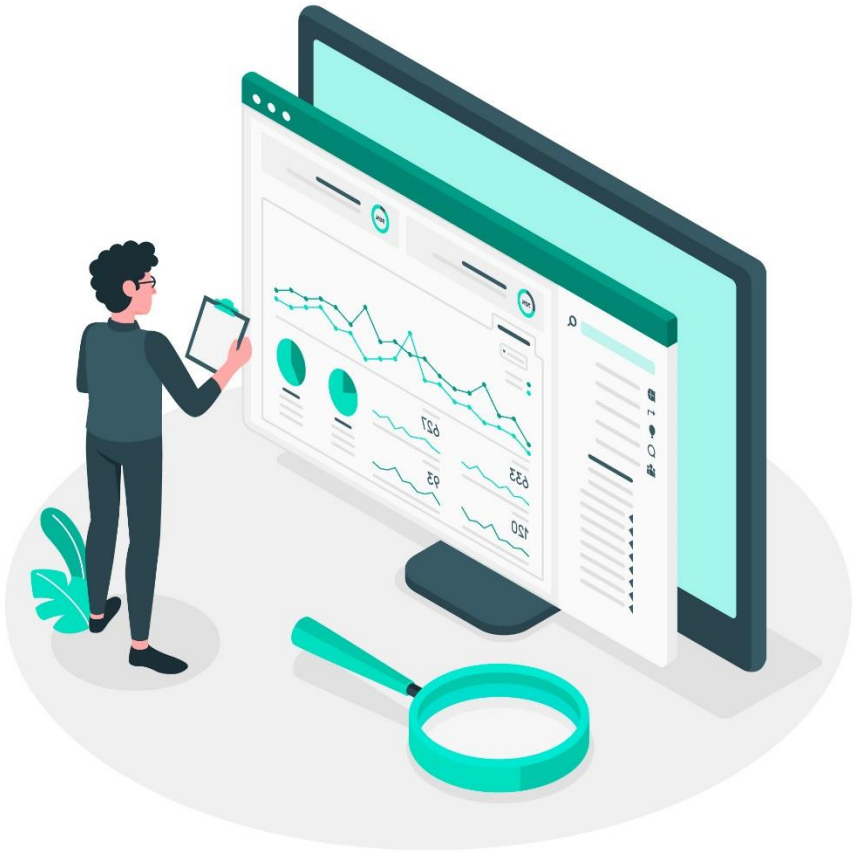
All *p-values < 0.05*
=> all independent variables have a strong relationship with the number of suicide cases

CONCLUSION AND SOLUTION

CATEGORICAL VARIABLES	NUMERIC ANALYSIS
1) Men have a significantly higher number and rate of committing suicide than women. 2) Number of suicides are witnessed the highest in people's mid-life however the its rate shows an upward trend according to the increase of age. 3) Year contributes least to the study on suicide number/rate as it randomly variates with no clearly trend. 4) People from old generations experiencing WWII and Silent time are the most likely one killing themselves than any other generations.	1) Number of population, GDP/capita, HDI of each country contribute hugely to both number and rate of a suicide happening. 2) Number of suicides: **Y=-0.014X1+0.000242X2+5725.5X3-4586.28** Suicide Rate: **Y=0.014X1+0.000242X2+5725.5X3-4586.28** **X1,X2,X3 is no of population, GDP/capital, HDI, respectively** 3) Good GPD/capita can help controls the situation. Countries with crowded population and best HDI seem to have more issues with suicide.



LIMITS AND FURTHER STUDY



LIMIT	FURTHER STUDY
1) The number of countries analyzed are small. 2) Linear model only applies on population, HDI and GDP/capita. 3) Dataset used is on period 2010-2014	Use bigger dataset (2000 datapoints) Take more useful index/data such as HI (Happy Index), average income,… Use more updating data

THANKS EVERYONE FOR WATCHING