الجامعة السعودية الالكترونية
Saudi Electronic University
2011 1432
College of Computing and Informatics

2020/2021 First Semester

Course Code	DS510
Course Name	Statistics for Data Science
CRN	14042
Assignment type	Project
Module	All
Assignment Points	10

Student ID	G200002614	G200007615
Student Name	Enad S. Alotaibi	Abdulaziz M. Alqumayzi

Introduction

Suicide is a global phenomenon and occurs throughout the year at any place. According to World Health Organizations statistics, nearly 800,000 people die due to suicide in every year ("Suicide across the world (2016)", 2020). Effective and evidence-based initiatives can be carried out if we can identify the underline reasons and correct statistics behind the data. According to the research, there are at least 20 suicide attempts for every one suicide. This study was conducted in order to identify the factors associated with suicides in each country in different times. The data was obtained from WHO and consist of 43776 observations with 6 variables. The description of the variables is as follows.

a) Country: Name of the country

b) Year: Year in which the suicide took place

c) Sex: It indicates the gender i.e., male or female

d) Age: indicates the Age group

e) No_of_suicides: Total number of suicide

f) Population: Total number of people

1) Descriptive analysis

The dataset consisted of 6 variables where 4 of them are categorical in nature and only two variables were numerical variables.

Variable	Mean	Minimum	Maximum	Number of missing values
No_of_suicides	193.3	0	22338.0	2256
Population	1664091	259	43805214	5460

The average number suicides happened all the time is 193.3. The average population is 1664091. The country with the highest population in a country on a certain year is 43805214

and the highest number of suicides which have been reported in a year in a certain country is 22338. There are is a one country or countries with no suicides in a certain year.

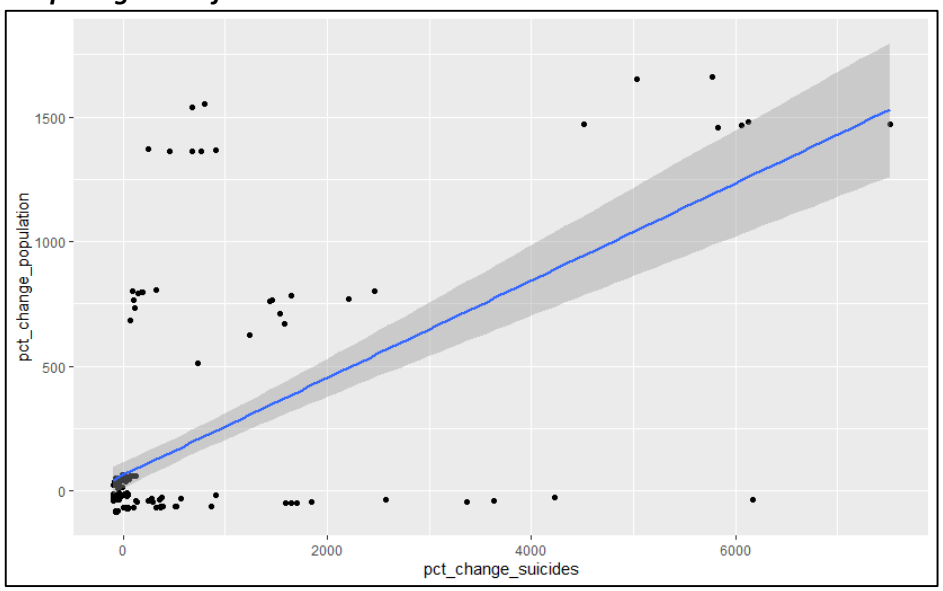
The observation has been obtained considering 38 years and most of the observations belong to year 2002. Thus, the data on many countries have been recorded in that year. There are 6 types of age groups considered in the data and a similar number of individuals are belonging to each age group. Also, there is similar number of individuals in both genders.

2) Dealing with missing values

There are 7716 missing values in the dataset and 5460 belongs to the variable population where other missing observation have been recorded under the variable number of suicides. Altogether, the data of 7716 records (observations) is missing from the data. This 17.62 of the data when considered as a percentage. There are several ways to deal with missing values as imputing or removing them from the analysis. Since the population and number of suicides is highly volatile in nature and the less number of missing values, it was decided to remove the observations with missing records.

3) percentage change in population and percentage change in suicides

The percentage change in the population can be calculated by multiplying the rate of change by 100. The percentage change in both population and suicides were calculated by similar manner and added to the dataset as two new variables. The following graph was plotted considering only ‘Sri Lanka’ as the country. According to the above plot, the number of suicides has been increased with the increase of the population denoting a positive correlation between the two variables. The calculated correlation coefficient between the two variables is 0.59 which is positive and averagely strong.



4) Number of suicides and population by age group

The number of suicides and the population can be denoted according to the age group as follows.

Age group	Population	No of suicides	Ratio
5-14	10451016008	62320	0.0000059
15-24	10587850825	975700	0.000092
25-34	10180794953	1360780	0.00013
35-54	16822867999	2887740	0.00017
55-74	10250167417	1955150	0.00019
75+	3009148664	756777	0.00025

The population of adults aged more than 75 is lower compared to the other age groups. The number of suicides is lower among the children aged 5-14 years. However, the population of children aged 5 to 14 years is similar to other age groups except the adult population aged more than 75 years. The number of suicides is lower next in the line among the adults aged more than 75 years. The smaller number of living adults might also be the reason for it. The highest

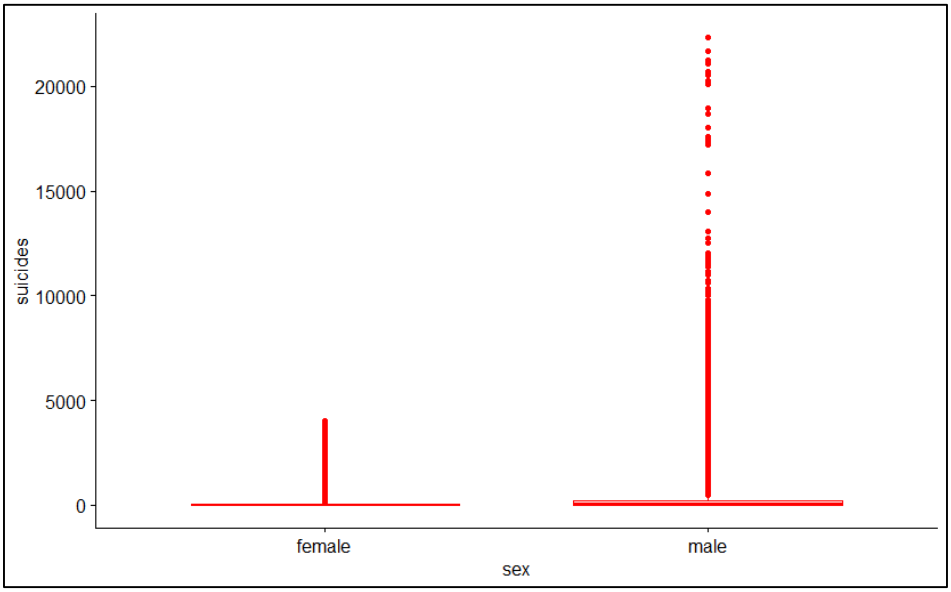
number of suicides are found among the individuals aged between 35 and 54. The correlation coefficient between the population and number of suicides is 0.65 which is positive and greater than 0.5. Thus, there is a positive and somewhat stronger relationship between the population and the number of suicides. The number of suicides increases with the population according to the positive association between the two variables.

Then the ratio between the number of suicides and the population was calculated in order to examine the association of age groups with the ratio of the suicides. The significance of the ratio relevant to each group can be measured by using the Kruksal-Wallis test which is a non-parametric test. The p-value of the test is 0.41 which provides evidence to accept the null hypothesis of independence. Thus, the ratio between the suicides and the population does not differ according to the age group.

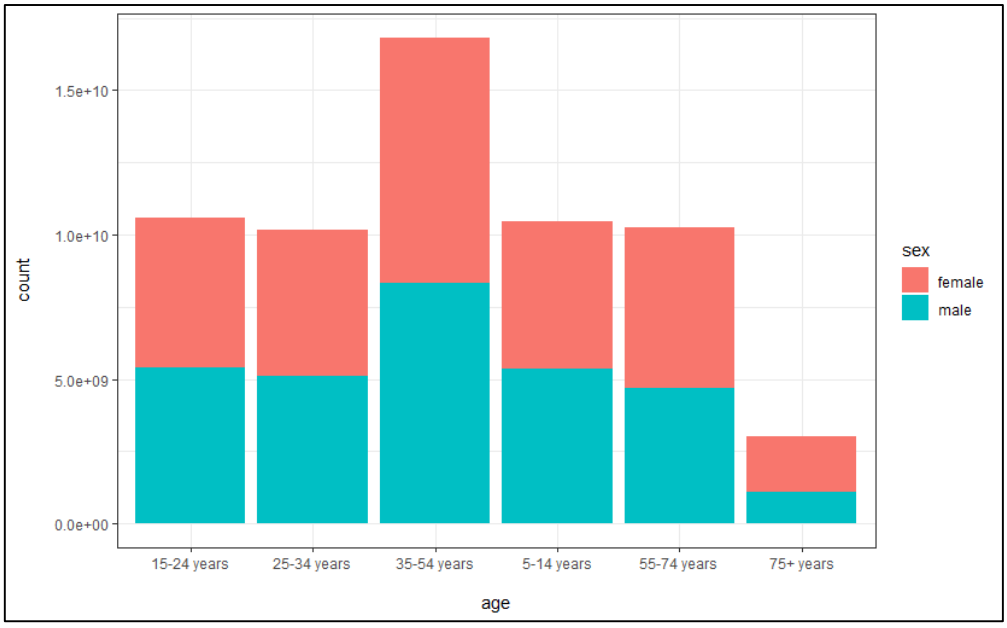
5) Number of suicides by age group for male and female

The below graph denotes the number of suicides according to the gender of the individuals. According to that, the male individuals have been suicided more number of times than female individuals. Thus, the males are more prone to suicide. A t-test can be conducted to statistically confirm the above results. According to the test results, the null hypothesis of the independence can be rejected under 5% significance level with a p-vale of 2.2e-16. Thus, the number of suicides is associated with the gender. According to the group means, it is evidence to say that the males are more prone to suicide than females.

mean of males	mean of females	t-statistic	p-value
338.55	105.06	-26.16	2.2e-16

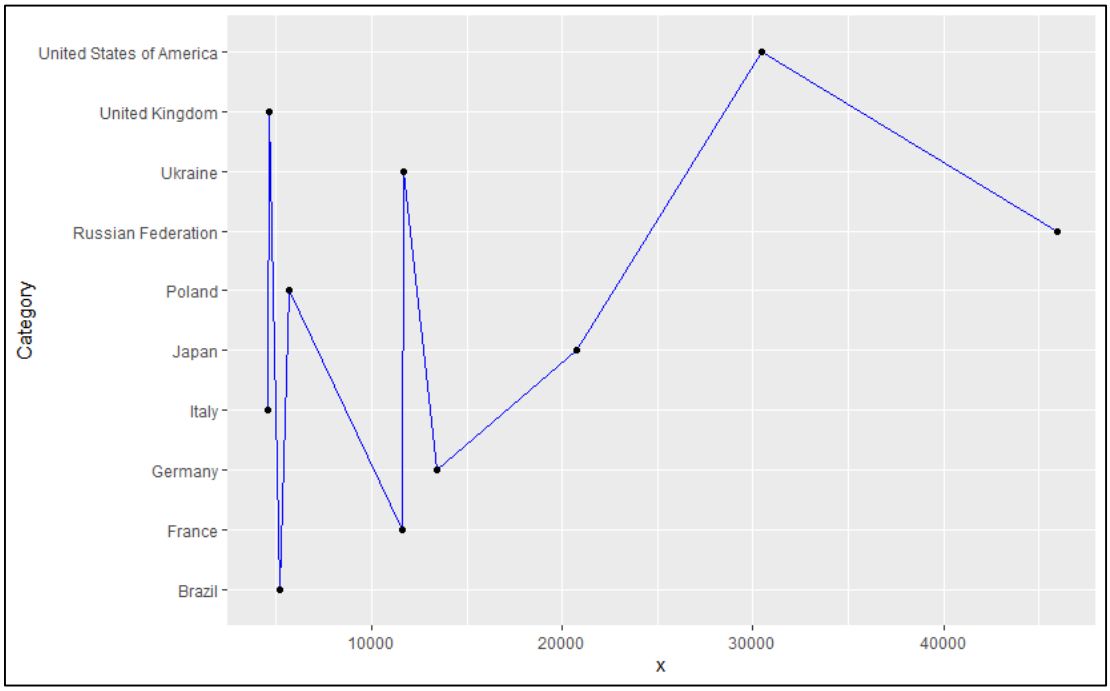


The below graph represents the graph of number of suicides grouped by age and the sex of an individual. The x axis indicates the age group of the individuals and the bars have been stacked according to the gender of the individuals. The females are denoted b the orange color and the males are denoted by blue color. The number of suicides are less among the adults aged more than 75 years and there are approximately a similar number of males and females in each age group.



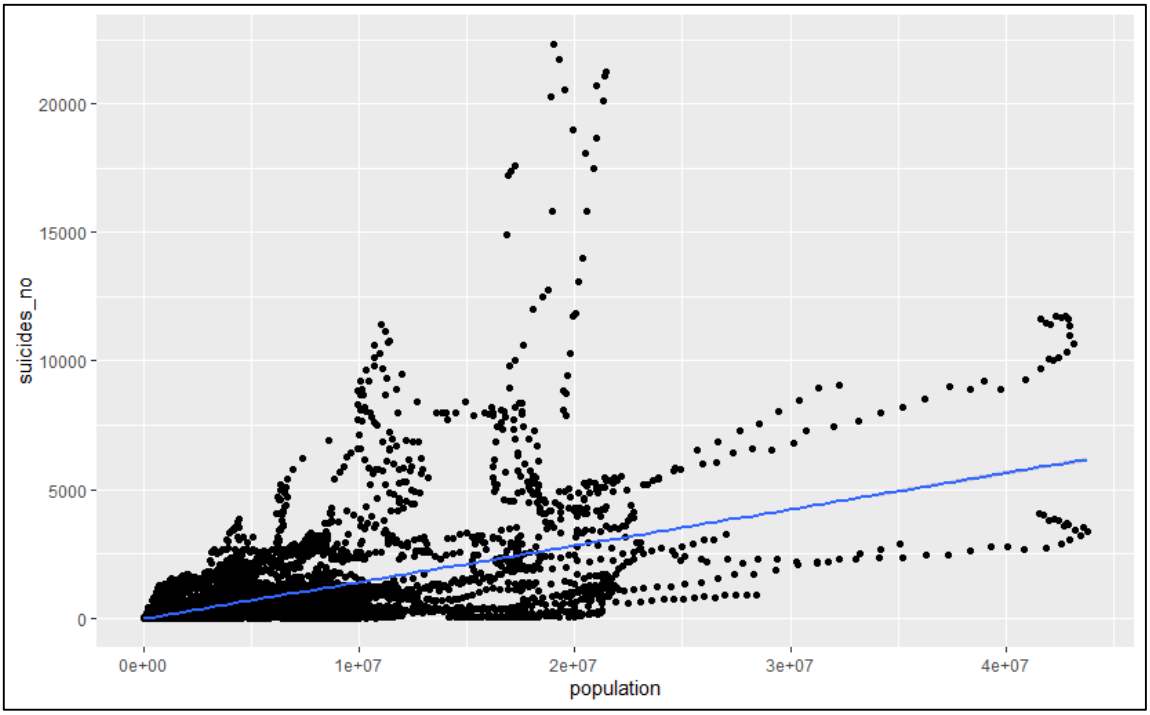
6) Countries and number of suicides

The following statistics were calculated considering the year 1992. The below graph represents the 10 countries which had the highest number of suicides in the year 1992. The Russian federation had the greatest number of suicides which is more than 50000 and the United Kingdom and Italy have taken the 9th and 10th position respectively.



7) Association between the population and the number of suicides

The below graph represents the scatterplot between the population and the number of suicides. The two variables seem to have a positive association where the number of suicides increases with the increase of the population. The correlation coefficient between the two variables is 0.611 which is greater than 0.5. Thus, the two variables have a somewhat strong positive association which conforms with the association shown in the visualization.



8) Estimating the number of suicides

The below model was fitted in order to estimate the number of suicides in coming years by considering five countries. Those countries are "Romania", "South Africa", "Sri Lanka", "United Arab Emirates" and "Turkey".

The least square model can be defined as follows.

No of suicides=(slope)year+(intercept)

The following table represents the values predicted by the least square model considering each country for next 10 years.

country	Romania									
year	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026
No of suicides	2529	2079	2161	2650	2774	2887	2793	2828	2859	2838

country	South Africa									
year	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
No of suicides	143	104	180	279	267	384	236	256	383	458

country	Sri Lanka									
year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
No of suicides	4269	4401	5076	5591	5769	5668	5345	4898	5887	5518

country	Unites Arab Emirates									
year	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
No of suicides	96	90	101	100	124	111	NA	NA	NA	NA

country	Turkey									
year	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
No of suicides	1050	1524	1148	1450	1810	1617	1532	NA	NA	NA

Reference

- Field, A., Miles, J. & Field, Z. (2012). Discovering Statistics Using R
- Suicide across the world (2016), (2020). Retrieved 9 November 2020, from https://www.who.int/mental_health/prevention/suicide/suicideprevent/en/
- Bangdiwala, S. (2018). Regression: simple linear. International Journal of Injury Controls and Safety Promotion, 25(1), 113-115.
- Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology, 70*(4), 407-411.
- Pinder, J. P. (2017). Introduction to Business Analytics using Simulation (pp. 151-195), London: Academic Press.