Data Analysis of Suicide Rates

Name

Institution

**Pre-step**

**Research Question**

The research question which we would like to answer with our data along with our hypothesis for each is listed below.

1. Is there a correlation between the Country Population and the Number of Suicides occurring?

The Population of a country refers to the total number of people in the country. The significance of this question is determining countries higher or lower population have more or fewer suicides.

**Data**

**Dataset**

The dataset chosen for this project is the “Suicide Rates Overview 1985 to 2016” dataset from Kaggle datasets (Kaggle, 2020). The dataset contains information of the top 100 countries with the highest number of suicides ranked by the WHO. This can help in determining which countries have the highest number of reported suicides. The data is obtained through reports from the United Nations Development Program, the WHO, and the World Bank. My dataset has 12 variables. My only categorical data is the country name. The numerical data include the year, the number of suicides (suicides\_no), the population, the number of suicides per a population of 100,000 people (suicides/100k pop), the HDI for a year, the GDP for a year and the GDP per capita. The age variable provides a range of years, for example, 15 – 24 years. The age variable corresponds to the generation which describes the name of the generation falling in the range. The country-year variable is also a combination of the country name variable and the year. For better analysis, the dataset should have had more countries in order to enable there to be comparison between various countries.

**Motivation and Background of the Research Questions**

The choice of data analysis project was brought about by the will to determine the increasing number of suicides. The American Psychological Association recently published an article called “By the numbers: An alarming rise in suicide” (Winerman, 2019). The article explains that there has been an increase in suicide cases at a rate of about 30% between the year 2000 and 2016. Suicide was also recorded as the 10th leading cause of death in the US. My motivation in this project is to analyze the data and focusing on the research questions I have posed to determine the correlations.

**Previous Research**

A research paper titled “Suicide Rates: Age-Associated Trends and Their Correlates” by Ajit Shah covers the relation between age and sex and the rates of suicide. The article uses data for 97 countries from the WHO (World Health Organization) to determine the relationships between aging and suicide rates (Ajit, 2013).

**Initial Analysis**

The research focuses at the relationship of different factors and the number of suicides. There are many statistical methods which can be used to evaluate the effects of various variables on the dataset. In order to determine the statistical difference in the data provided, I will use Python Jupyter Notebook for analysis. Python modules such as matplotlib, pandas and NumPy will be in use to determine the relationships between the data, calculating the correlation coefficient and mapping the values. The use of NumPy is in calculations such as correlation. The use of matplotlib will be used in data visualization of data. This will involve bar graphs, scatter graphs, and line graphs. Pandas is used in importing the data and storing the data in a dataframe. Seaborn is using in generating a heatmap. Python statsmodels is used in doing regression analysis.

**Data Manipulations**

The first step is determining the data:

Input 1:

import pandas as pd

import numpy as np

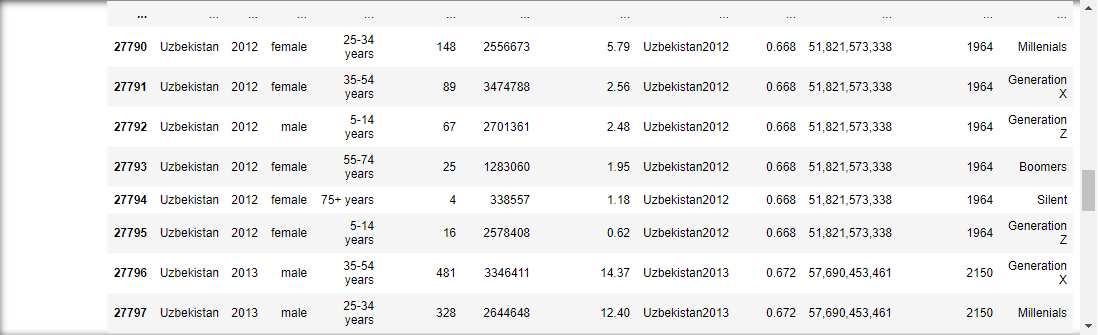
import matplotlib.pyplot as plt

df = pd.read\_csv("master.csv")

df

Output 1:





Code 1:

This python code is used to show a preview of the data present in the dataset. This information can be used in understanding how the data is structured as a table. The data also provides a view of the first and last few rows which provides a change of data view. With this information we can observe that there are 12 columns present and 27820 rows.

Input 2:

corr = df.corr()

corr.style.background\_gradient(cmap='coolwarm').set\_precision(2)

Output 2:



Code 2:

This output provides the correlation between the different numerical variables in the dataset. The color scheme used is coolwarm. This means that the higher the correlation of variables, the more the value will move to a warm color (red). The vice versa is also true whereby the lesser the correlation, the more the value adopts a cool color (blue).

**Graphs**

Input 3:

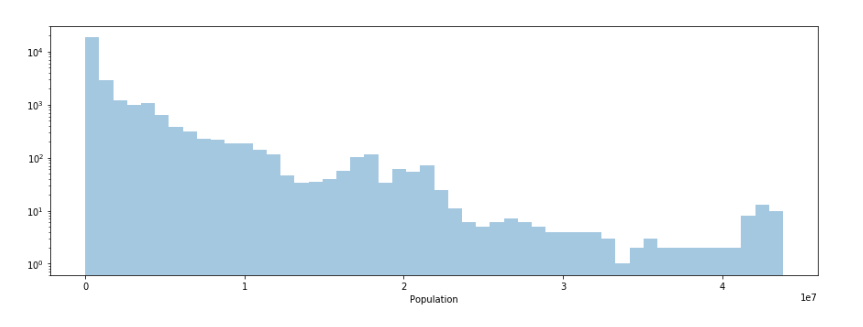
plt.rcParams['figure.figsize'] = (15,5)

sns.distplot(df.population,kde=False,hist\_kws={'log':True})

plt.xlabel("Population")

plt.show()

Output 3:



Code 3:

The output shows that there are many countries which have lower populations as compared to countries with more than 25 million people and above.

Input 4:

q = df["population"]

p = df["suicides\_no"]

plt.scatter(q,p)

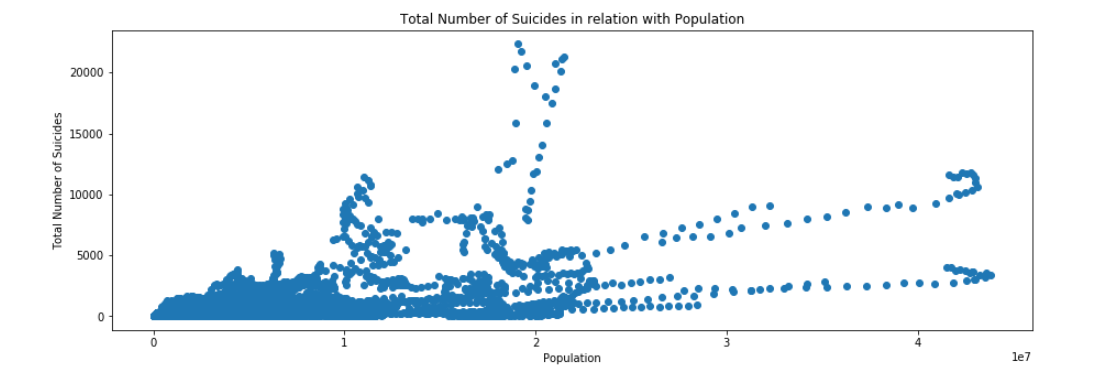
plt.xlabel(‘Population’)

plt.ylabel('Total Number of Suicides')

plt.title('Total Number of Suicides in relation with Population)

plt.show()

Output 4:



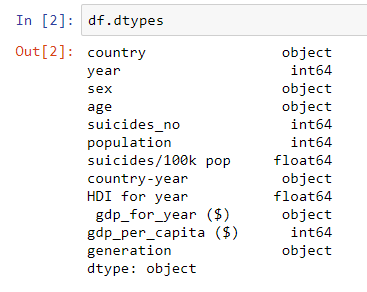
Code 4:

This scatter plot shows the relation between the population and the total number of suicides. This provides a good view of how the two variables are related and the trend in data.

**Hypothesis Formation**

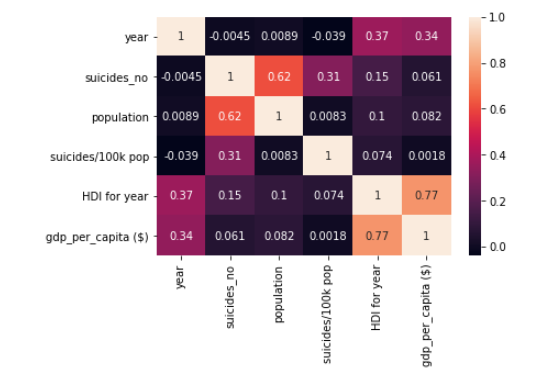
**Dependent and Independent Variables**

The dependent variable in this case is the Number of Suicides (suicide\_no) whereas the independent variable is the Population (population). The two variables are both integers:



**Correlation Coefficient**

The correlation coefficient can be calculated using the correlation matrix from the seaborn module. This can be easily represented by a heatmap as follows:



**Regression Model**

The regression model can be derived by generating OLS Regression Results. The following code is used:

Input

X = df["population"]

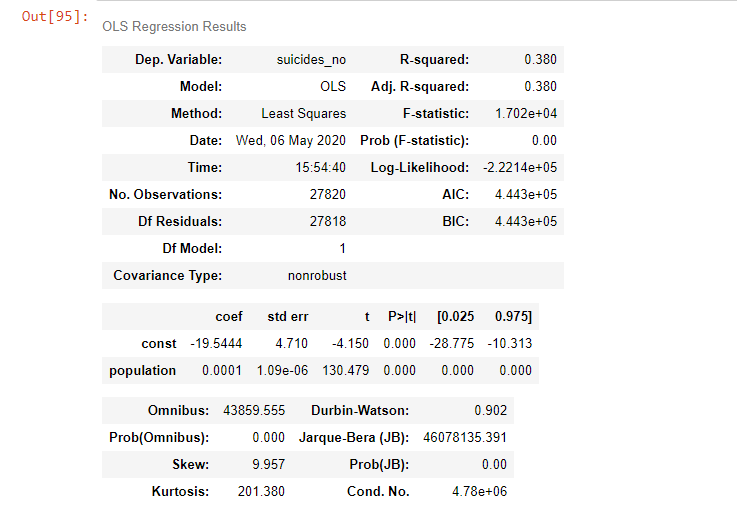
Y = df["suicides\_no"]

X = sm.add\_constant(X)

results = sm.OLS(Y, X).fit()

results.summary()

Output



An equation derived from this data is:

Y = 0.0001X – 19.5444

**Null and Alternative Hypothesis**

The null hypothesis is: all the βs are equal to zero simultaneously.

The alternative hypothesis is: at least one β differs from zero.

**Regression Analysis**

**Showing the Regression Output**

The regression equation is given by:

Y = 0.0001X – 19.5444

When interpreted with the data present, this translates to:

Number of Suicides = 0.0001 (Population) – 19.5444

For example, in a population of 250,000; the number of suicides estimates are:

0.001 (250000) – 19.5444 = 5 people

**Interpreting the Regression Output**

From analyzing the OLS Regression Results, the following data can be interpreted:

*Coefficient*

The constant coefficient is given as -19.5444 whereas the other coefficient is 0.0001

*P-value*

The p-value is given as 0.00

*Confidence Interval for Independent Variable*

The confidence interval is 100% since the p-value is 0.00

**Hypothesis Accepted and Rejected**

The null hypothesis, all the βs are equal to zero is accepted since the F-statistic probability as viewed from the OLS Regression results is 0.00. The alternate hypothesis is rejected.

**Heteroskedasticity**

There is little presence of heteroskedasticity in the regression analysis. This is because the regression equation falls in line with the distribution of data. This can be shown by the following graph:

Input:

plt.scatter(q,p)

yhat = 0.0001\*q - 19.5444

fig = plt.plot(q,yhat, lw=4, c='orange', label = 'regression line')

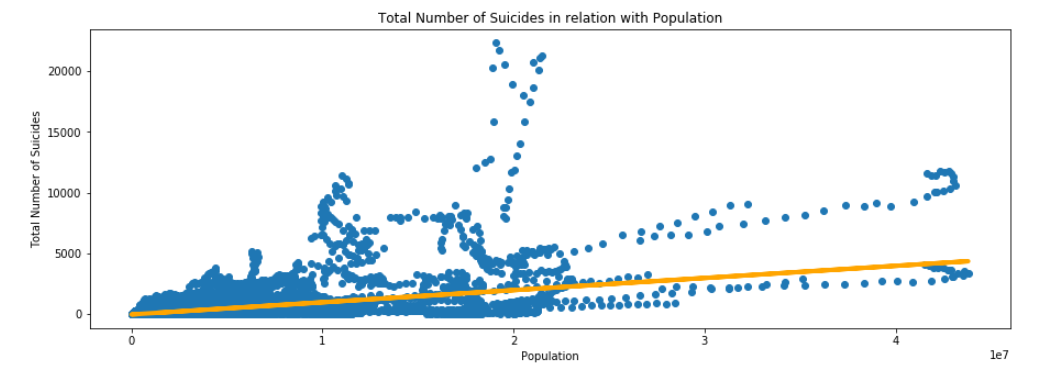
plt.xlabel('Population')

plt.ylabel('Total Number of Suicides')

plt.title('Total Number of Suicides in relation with Population')

plt.show()

Output:



**Results and Conclusion**

The data could be affected by two main forms of bias: selection bias and observer bias. The dataset contains data which has selected countries where suicide is viewed to be very high. The countries left out from the dataset if included could change the distribution of data and could form different values on regression analysis. The observer bias could have resulted in picking of data from research which involves subconsciously inputting their expectations in the data. The research question was “Is there a correlation between the Country Population and the Number of Suicides occurring?” With the data analysis and representation, one can identify that there is high correlation between the two variables. In the correlation diagram in Output 2 and the heatmap shown in the correlation coefficient section, one can view that the variables have a correlation coefficient of 0.62. The regression analysis also shows the relation between the variables. The greatest weakness in this analysis is that the population could change and with multiple suicide awareness prevention programs available, the data such as the regression equation could change drastically with time.

# **References**

Ajit, S. (2013). *Suicide Rates: Age-associated Trends and Their Correlates.* US National Library of Medicine.

Kaggle. (2020). *Suicide Rates Overview from 1985 to 2016*. Retrieved from Kaggle: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

Winerman, L. (2019, January). *By the numbers: An alarming rise in suicide*. Retrieved from American Psychological Association: https://www.apa.org/monitor/2019/01/numbers