**Paul Sujith Rayi**

**IST 652 Mini project 1**

**Report**

**Data and its source:** The dataset on which I am doing analysis for my mini project is a Suicide Rates Overview from 1985 to 2016. This dataset has been downloaded from the website Kaggle. This dataset is a 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.

Here’s the link to the dataset:

<https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016>

I downloaded the dataset and it was named master.csv. I then opened excel and imported the csv file which I downloaded and saved it as suicides.xlsx. I did this as I wanted to work with excel and pandas as opposed to csv and pandas which was taught in class. I wanted to try out something new.

**Data Exploration and Data Cleaning, Description of the program file:**

The total number of columns in the dataset are 12 and they are:

a) country – it is the name of the country

b) year – it is the year number

c) sex – it contains sex values male and female

d) age – it is divided into different age categories or brackets

e) suicides\_no - indicates the number of suicides

f) population - indicates the total population of the country

g) suicides/100k pop - this column tells us the ratio of suicides per hundred thousand population

h) country-year - It indicates the name of the country, followed by the year number

i) HDI for year - HDI stands for human development index

j) gpp\_for\_year ($) - It is the Gross Domestic Product of the country expressed in dollar terms

k) gpd\_per\_capita ($) - It is the Gross Domestic Product of the country divided by the total population of the country and it is expressed in dollars

l) generation - This column contains the generation type of the population. For example – Generation X, Silent, Boomers etc.

The total number of rows in the data set excluding the header row or metadata, or the total number of rows which contain values are 27,820

Now let me explain what I did in the .py file attached along with this report. I used Visual Studio Code to create the .py file and do coding (perform data analysis)

First, I imported all the modules and objects which I will be using for my data analysis.

The modules that I imported, and the description are as follows:

Pandas – data structure to analyze data

DataFrame – pandas object will help us store and perform actions on the data

Matplotlib.pyplot – help us visualize data and see the visualized output in a separate window

Numpy – helps us perform numerical analysis on data

Xlrd – package or module for reading excel data

Xlwt – package or module for writing excel data

I then read in the excel data and looked at the first ten rows/records of the data, to get an idea of how the data is structured and formatted.

Then I looked at the number of rows and columns using the shape function and I used the columns attribute to extract and view the column names. The reason I did this is to look at column names and get an idea or insight as to whether I will be needing all the columns for my analysis and what columns can I delete.

I then deleted the columns that I did not want using the drop function.

Then I used the unique function to look at unique values in relevant columns (columns that do not have too many distinct numerical values and columns who have only strings) so as to understand how I can analyze the data and frame my data questions.

I then used the sort\_values function just to sort the cleaned dataset based on values of one column and see how the output looks like and understand my data.

Then I visualized the first ten number of suicides values from the sorted dataset using plot function and gave the value ‘barh’ to the kind variable so that the diagram comes out as a horizontal plot. The plot opens in a new window and if you look at it, all the bars are beautifully colored in different colors. The y axis tells the value of the suicide number and the x-axis tells us the row/record number of that particular data in the sorted dataset.

Then I used the describe function to get the summary statistics of columns which have numerical values in our cleaned data set.

**Data Questions:**

There are three data questions which I wanted to answer looking at the cleaned data set

1) In each country, over the years, who committed the greatest number of suicides based on sex. And whether this trend persisted among all the countries.

2) In each year, over all the countries, which was the major age group that committed the greatest number of suicides and whether this trend persisted along all the years.

3) In each country, over the years, who committed the greatest number of suicides based on generation type. And whether this trend persisted among all the countries.

**Process Explanation:**

Now let me go about how I aggregated the data and produced the results.

First, I created subset datasets by selecting only the relevant columns for doing analysis and trying to answer each question individually.

Then I created pivot tables using the pivot\_table function and by passing the columns on which the result/values must be aggregated as index columns. The function which I used for all my pilot tables was sum as I wanted to know the total to find the suicides trend across years or countries.

Then I visualized some rows/records of the data from the pivot tables using plot function which is part of the matplotlib.pyplot module.

Then I exported the pivot tables into excel files using the to\_excel function which writes and saves the output produced by the pivot table in excel files based on the name which we specify.

I named the files as output1.xlsx, output2.xlsx and output3.xlsx respectively for ease of readability.

One important point to remember here is that the output files will be saved to the same working directory as the .py file.

**Data Visualization and Output Files explanation:**

Visualization1: I produced a bar chart which depicts country and sex in the x-axis and aggregated values of suicide numbers in the y-axis

Output1: This excel file contains our three columns which we specified as header row i.e. country, sex and suicides\_no.

The values under country column are the names of countries. The values under sex column is male and female and these values are repeated for every country. The values of suicide\_no are the total number of suicides committed by a sex in a country over the years.

Looking at the excel file data, I can clearly say that in any country the number of males committing suicides is greater than the corresponding number for females. And this pattern persists over all countries. Hence, we can safely conclude that for any given country, the number of males committing suicide is higher than the number of females over the years.

Visualization2: I produced a line chart which depicts year and age in the x-axis and aggregated values of suicide numbers in the y-axis

Output2: This excel file contains our three columns which we specified as header row i.e. year, age and suicides\_no

The value under the year column is the number of the particular year. The values under age column are ‘15-24 years’, ‘25-34 years’, ‘35-54 years’, ‘5-14 years’, ‘55-74 years’ and ‘75+ years’ and these are repeated for every year. The values of suicide\_no are the total number of suicides committed by an age group in a year over the countries.

Looking at the excel file data, I can clearly say that in any year the number of suicides committed by ‘35-54 years’ age group is the highest over all the countries. And this pattern persists over all the years. Hence, we can safely conclude that for any given year, the number of 35-54 age group committing suicide is higher than any other age group over all the countries.

Visulaization3: I produced a pie chart which depicts the country and generation type as parts/fractions of the pie and suicide number being the deciding factor for the percentage/size of the part of the pie

Output3: This excel file contains our three columns which we specified as header row i.e. country, generation and suicides\_no

Expand the width of the generation column manually so that the generation type becomes clearly visible.

The value under the country column is the name of the particular country. The values under the generation column are ‘Boomers’, ‘G.I. Generation’, ‘Generation X’, ‘Generation Z’, ‘Millenials’ and ‘Silent’, and these values are repeated for every country. The values of suicide\_no are the total number of suicides committed by a generation type in a country over the years.

Looking at the excel file data, I can say that there is no persistent generation type that committed highest suicides in any country over the years. Hence, we can conclude that generation type does not has a strong correlation with suicides committed in a country over the years.

References:

<https://www.dataquest.io/blog/excel-and-pandas/>