“EXPLORATORY DATA ANALYSIS ON GLOBAL SUICIDE RATES”

A Report

Submitted as special assignment

of

2CSOE03 DATA ANALYTICS

By
Deep Khut (19BIC008)

Under the Guidance of
Prof. Aparna Kumari

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
INSTITUTE OF TECHNOLOGY
NIRMA UNIVERSITY
Ahmedabad 382 481

NOVEMBER 2022

TABLE OF CONTENTS

1. INTRODUCTION:

Suicide is one of the 1eading causes of death among a11 adu1ts and rates are increasing in both men and women. But numbers also show stark differences between genders.

In 2017, men died by suicide 3.54 times more often than women. Midd1e-aged white men, in particu1ar, are susceptible. White males accounted for near1y 70-percent of suicide deaths in 2017, according to the American Foundation for Suicide Prevention.

“There can be a stigma among men that they shou1d ‘tough things out,’ rather than seeking he1p if they’re having strugg1es with their menta1 hea1th,” says Dr. Lisa Baker, an SSM Hea1th Psycho1ogist at St. Mary’s Hospita1 - Madison. “As a resu1t, menta1 hea1th conditions are under-reported and under-detected in men, 1eaving them vu1nerable to suicide.”

People who live in rural areas are at higher risk of suicide than their urban counterparts, according to the Centers for Disease Control and Prevention. This, in part, can be explained by greater access to firearms, drug and alcohol use and a scarce of health care providers and emergency medical services. Cultural factors are also a barrier to accessing care and getting support from family and friends.

To **perform Data Analysis** and wants you to examine **trends & correlations** within our data. We would like to make a **Machine Learning algorithm** where we can train our **AI** to **learn** & improve from experience. Thus, we would want to **predict** the amount of suicides numbers in a certain demographic.

This project seeks to **explore** the underlying factors. We will use a sample of **44,000** data points gathered from **141** different countries, between the **80**’s to **2016.**

Research Questions
1. Which year has the most suicides? Which year has the least suicides?
2. Which country has the most suicides? Which country has the least suicides?
3. Are certain age groups more inclined to suicide?
4. What is the relationship between gender and the number of suicides?

Features & Predictor:
Our **Predictor (Y, Suicide Count)** is determined by **5 features (X):**

1. country (Categorical)

2. year: year of suicide (Categorical)
3. sex: Male, Female (Categorical)
4. age (Categorical)
5. population: (#)

2. DATA WRANGLING:

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns

data = pd.read_csv("C:/Users/jayes/Downloads/who_suicide_statistics.csv
/who_suicide_statistics.csv")

# look at 1st 5 data points
data.head(5)
```

```
   country  year     sex          age  suicides_no  population
0  Albania  1985  female  15-24 years          NaN    277900.0
1  Albania  1985  female  25-34 years          NaN    246800.0
2  Albania  1985  female  35-54 years          NaN    267500.0
3  Albania  1985  female   5-14 years          NaN    298300.0
4  Albania  1985  female  55-74 years          NaN    138700.0
```

Our data set has **5 Features** (Country, Year, Gender, Age, Popu1ation). We wi11 exp1ore a11 of these in detai1. While the suicide_no is what we would 1ike to **predict.**

```
data.info()# print the concise summery of the dataset
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 43776 entries, 0 to 43775
Data columns (total 6 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   country      43776 non-null  object 
 1   year         43776 non-null  int64  
 2   sex          43776 non-null  object 
 3   age          43776 non-null  object 
 4   suicides_no  41520 non-null  float64
 5   population   38316 non-null  float64
dtypes: float64(2), int64(1), object(3)
memory usage: 2.0+ MB
```

```
# counts total row in each col. that have null values
# note: all the na columns are type Object
data.isna().sum()
```

```
country          0
year             0
sex              0
age              0
suicides_no   2256
population    5460
dtype: int64
```

```
# From above, we can see that, suicides_no & population, have null values.
#Lets, fill the null values with zero using 'fillna'
data= data.fillna(0)
data.isna().sum()
```

```
country        0
year           0
sex            0
age            0
suicides_no    0
population     0
dtype: int64
```

```
In [11]: data['age'].unique()
Out[11]:
array(['15-24 years', '25-34 years', '35-54 years', '5-14 years',
       '55-74 years', '75+ years'], dtype=object)
```

```
In [13]: data['country'].unique()
Out[13]:
array(['Albania', 'Anguilla', 'Antigua and Barbuda', 'Argentina',
       'Armenia', 'Aruba', 'Australia', 'Austria', 'Azerbaijan',
       'Bahamas', 'Bahrain', 'Barbados', 'Belarus', 'Belgium', 'Belize',
       'Bermuda', 'Bolivia', 'Bosnia and Herzegovina', 'Brazil',
       'British Virgin Islands', 'Brunei Darussalam', 'Bulgaria',
       'Cabo Verde', 'Canada', 'Cayman Islands', 'Chile', 'Colombia',
       'Costa Rica', 'Croatia', 'Cuba', 'Cyprus', 'Czech Republic',
       'Denmark', 'Dominica', 'Dominican Republic', 'Ecuador', 'Egypt',
       'El Salvador', 'Estonia', 'Falkland Islands (Malvinas)', 'Fiji',
       'Finland', 'France', 'French Guiana', 'Georgia', 'Germany',
       'Greece', 'Grenada', 'Guadeloupe', 'Guatemala', 'Guyana', 'Haiti',
       'Honduras', 'Hong Kong SAR', 'Hungary', 'Iceland',
       'Iran (Islamic Rep of)', 'Iraq', 'Ireland', 'Israel', 'Italy',
       'Jamaica', 'Japan', 'Jordan', 'Kazakhstan', 'Kiribati', 'Kuwait',
       'Kyrgyzstan', 'Latvia', 'Lithuania', 'Luxembourg', 'Macau',
       'Malaysia', 'Maldives', 'Malta', 'Martinique', 'Mauritius',
       'Mayotte', 'Mexico', 'Monaco', 'Mongolia', 'Montenegro',
       'Montserrat', 'Morocco', 'Netherlands', 'Netherlands Antilles',
       'New Zealand', 'Nicaragua', 'Norway',
       'Occupied Palestinian Territory', 'Oman', 'Panama', 'Paraguay',
```

```
# the Number of different Countries our dataset is from
data['country'].nunique()
# Our dataset is from 141 different Countries

# The different country groups
data['year'].unique()
```

FILLING IN MEAN VALUES FOR THE MISSING DATA & REPLACE NA VALUES WITH THEIR MEAN VALUES

```
# Replace 0 values with, NA
data['suicides_no'] = data['suicides_no'].replace(0,np.NAN)

# replace Na values with, mean value
mean_value=data['population'].mean()

data['population']=data['population'].fillna(mean_value)

# do same for Popualation
# replace Na values with, mean value
mean_value=data['suicides_no'].mean()

data['suicides_no']=data['suicides_no'].fillna(mean_value)
```

3. EXPLORATORY DATA ANALYSIS

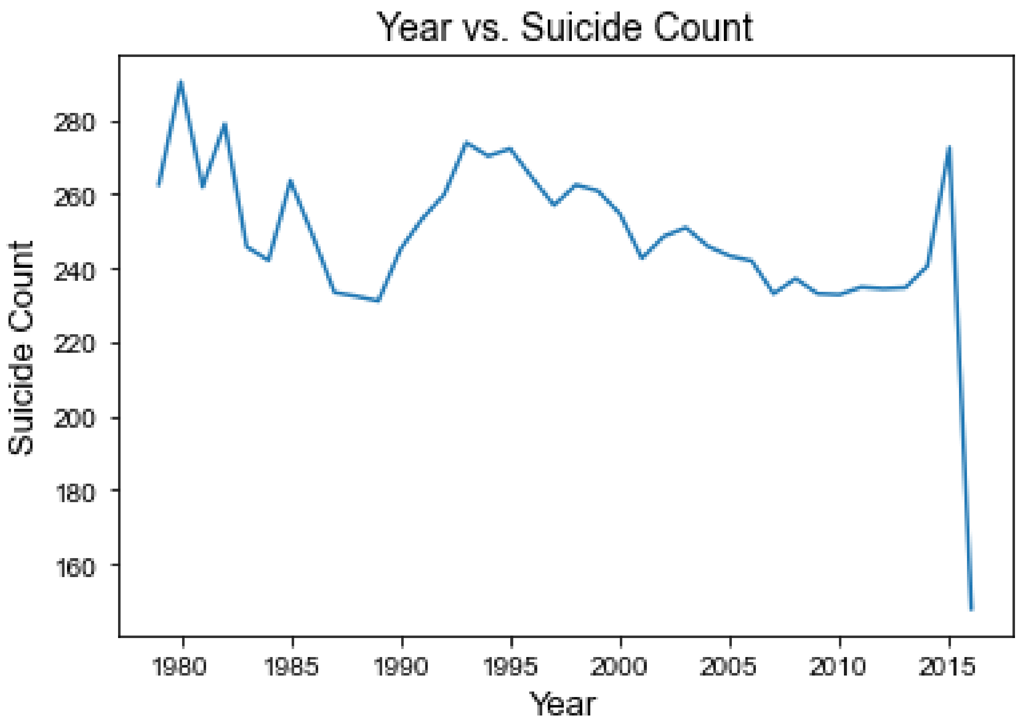
➔ **Research Question I: Which year has the most Suicides ? Which year has the 1east Suicides ?**

```
data['suicides_no'] = data['suicides_no'].replace(0,np.NAN)

mean_value=data['suicides_no'].mean()
data['suicides_no']=data['suicides_no'].fillna(mean_value)

def find_minmax(x):
    #use the function 'idmin' to find the index of lowest suicide
    min_index = data[x].idxmin()
    #use the function 'idmax' to find the index of Highest suicide
    high_index = data[x].idxmax()

    high = pd.DataFrame(data.loc[high_index,:])
    low = pd.DataFrame(data.loc[min_index,:])

    #print the Year with high and low suicide
    print("Year Which Has Highest "+ x + " : ",data['year'][high_index]
)
    print("Year Which Has Lowest "+ x + "  : ",data['year'][min_index])
    return pd.concat([high,low],axis = 1)

find_minmax('suicides_no')
```

```
Year Which Has Highest suicides_no :  1994
Year Which Has Lowest suicides_no  :  1987
Out[32]:
                               33128        29
country        Russian Federation     Albania
year                         1994        1987
sex                          male      female
age                   35-54 years   75+ years
suicides_no               22338.0         1.0
population             19044200.0     35600.0
```

Year — Wise Analysis

```
data.groupby('year')['suicides_no'].mean().plot()

#setup the title and labels of the figure.
plt.title("Year vs. Suicide Count",fontsize = 14)
plt.xlabel('Year',fontsize = 13)
plt.ylabel('Suicide Count',fontsize = 13)
```



From observing our **Time Series Line P1ot,** we can see a **sharp drop** in suicides in 1985. This **decrease** could be due to **awareness** of suicide & **menta1 health** in the 80s, as well as **improved recognition** of those at risk. This is indeed **accurate**, as the research, “Suicide in the e1derly” **supports** this c1aim.

➔ **Research Question 2: Which country has the most Suicides? Which country has the least Suicides?**

```
def find_minmax(x):
```

```
     #use the function 'idmin' to find the index of lowest suicide
    min_index = data[x].idxmin()
    #use the function 'idmax' to find the index of Highest suicide
    high_index = data[x].idxmax()

    high = pd.DataFrame(data.loc[high_index,:])
    low = pd.DataFrame(data.loc[min_index,:])

    #print the country with high and low suicide
    print("Country Which Has Highest "+ x + " : ",data['country'][high_
index])
    print("Country Which Has Lowest "+ x + "  : ",data['country'][min_i
ndex])
    return pd.concat([low,high],axis = 1)

find_minmax('suicides_no')
```

```
Country Which Has Highest suicides_no :  Russian Federation
Country Which Has Lowest suicides_no  :  Albania
Out[34]:
                    29               33128
country        Albania  Russian Federation
year              1987                1994
sex             female                male
age          75+ years         35-54 years
suicides_no        1.0             22338.0
population     35600.0          19044200.0
```

Feature Engineering

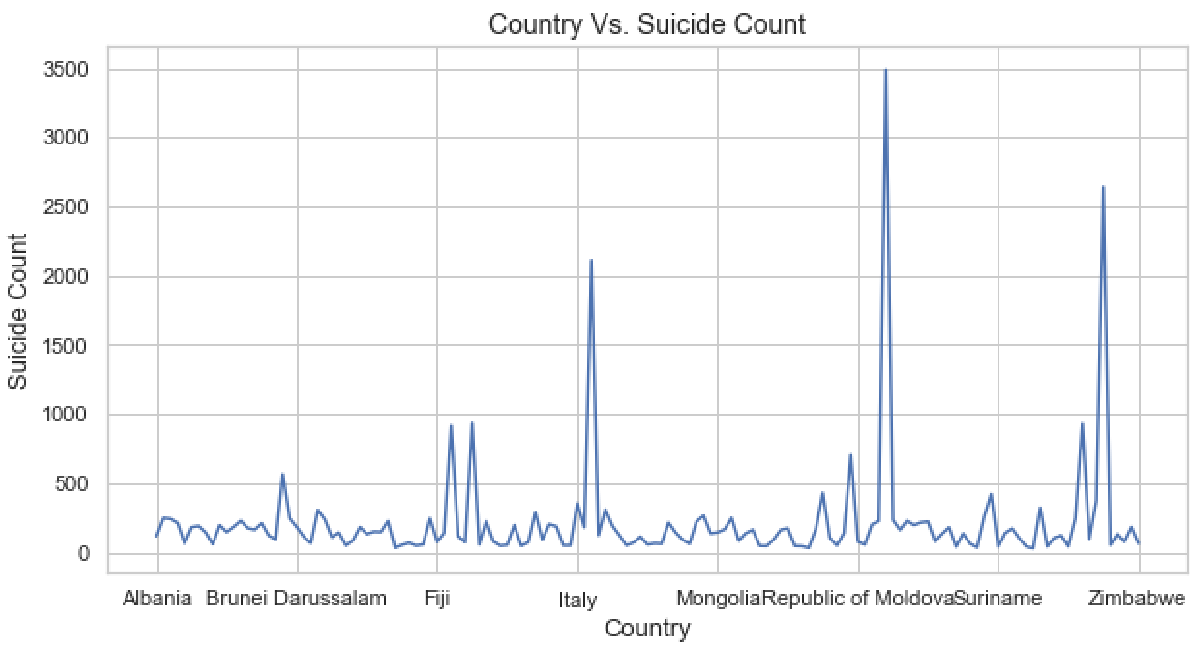
Calculate the Suicide per Population Size Ratio, to better understand our data

```
#calculate mean of suicides_no col
meanSuicide = data['suicides_no'].mean()
#calculate mean of pop. col
meanPop = data['population'].mean()
# Replace 0 or NaN populations, with the mean Populations
data['population'] = data['population'].replace(np.NAN,meanPop)
data['population'] = data['population'].replace(0,meanPop)
data.tail(3)
```

	country	year	sex	...	suicides_no	population	suicide_per_pop
43773	Zimbabwe	1990	male	...	6.0	1.456536e+06	0.000004
43774	Zimbabwe	1990	male	...	74.0	1.456536e+06	0.000051
43775	Zimbabwe	1990	male	...	13.0	1.456536e+06	0.000009

```
find_minmax('suicide_per_pop')
```

```
In [37]: find_minmax('suicide_per_pop')
Country Which Has Highest suicide_per_pop :  Rodrigues
Country Which Has Lowest suicide_per_pop  :  Egypt
Out[37]:
                      12993       32351
country               Egypt   Rodrigues
year                   2005        2004
sex                    male        male
age               5-14 years   75+ years
suicides_no             1.0  249.106328
population        9543088.0       259.0
suicide_per_pop         0.0      0.9618
```



Both the graph & find_minmax function above, **confirm** that A1bania had the **1owest** suicide count, while Zimbabwe & Russian Federation, had the 1argest suicide count. A **reason** the Russian Federations may have a **1arge** suicide count may be that they have a very 1arge population (144.5 mi11ion, while A1bania only has about 3 mi11ion). lt has been reported that Russian 1eve1s of a1coho1 consumption p1ays an immense ro1e in it's **1arge suicide count**, but their is a **1ack** of data to **support** this due to Soviet secrecy.

➔ Research Question 3: Are certain age groups more inclined to suicide?

```
sample = data.sample(3)
sample
```

```
         country  year     sex  ... suicides_no  population  suicide_per_pop
2264   Australia  1987    male  ...       554.0   2031000.0         0.000273
41160    Ukraine  1996  female  ...       165.0   3595700.0         0.000046
22904     Latvia  2006    male  ...       167.0    294935.0         0.000566

[3 rows x 7 columns]
```

Right now our ‘age’ co1umn is **separated** into **hyphen** groups. We want to ana1yze these groups as **numerical** data. We must take **away** the hyphen & create a **function** that classifies each category into a **certain** number. We first must **remove** a11 instances of a dash & change the object to type int to further analyze it.

```
 # grabs first 2 chars from Age Column
data['AgeNum'] = data['age'].str[:2]

# remove all instances of dash -
data['AgeNum'] = data['AgeNum'].map(lambda x: x.replace('-',''))

# now, convert it to type int (not Object)
data['AgeNum'] = data['AgeNum'].astype(int)

data['AgeNum'].tail(3)
```

```
43773     5
43774    55
43775    75
Name: AgeNum, dtype: int32
```

```
# creates Age Categories
def AgeGroup(x):
    if(x >= 60):
        return "Elderly"
    elif(x >= 30):
        return "Middle_Aged_Adults"
    elif(x >= 18):
        return "Adults"
    else:
        return "Adolescent"
# Map each row in the Col to the AgeGroup Method
data['AgeCategory'] = data['AgeNum'].map(lambda x: AgeGroup(x))
# convert it back to type String
data['AgeCategory'] = data['AgeCategory'].astype(str)
data['AgeCategory'].tail(3)
```

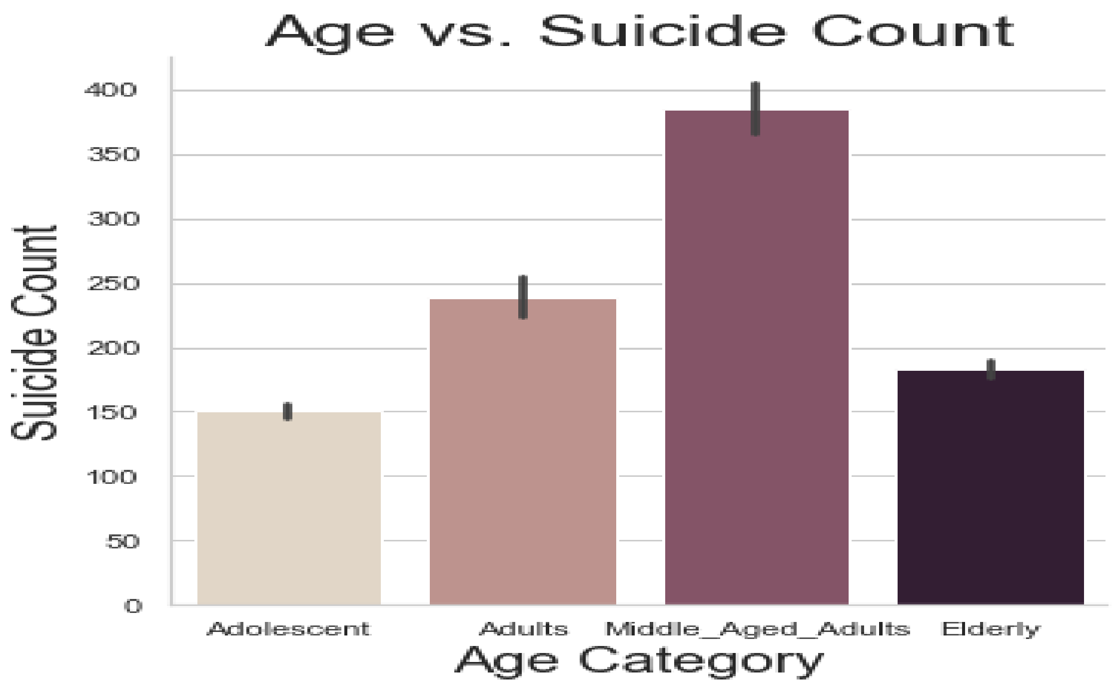
```
data['AgeNum'] .tail(3)
```

```
43773              Adolescent
43774      Middle_Aged_Adults
43775                 Elderly
Name: AgeCategory, dtype: object
```

Note: Created an new column ca11ed ‘AgeNum’

```
data.head(3)
```

```
   country  year     sex  ... suicide_per_pop  AgeNum         AgeCategory
0  Albania  1985  female  ...        0.000896      15          Adolescent
1  Albania  1985  female  ...        0.001009      25              Adults
2  Albania  1985  female  ...        0.000931      35  Middle_Aged_Adults

[3 rows x 9 columns]
```

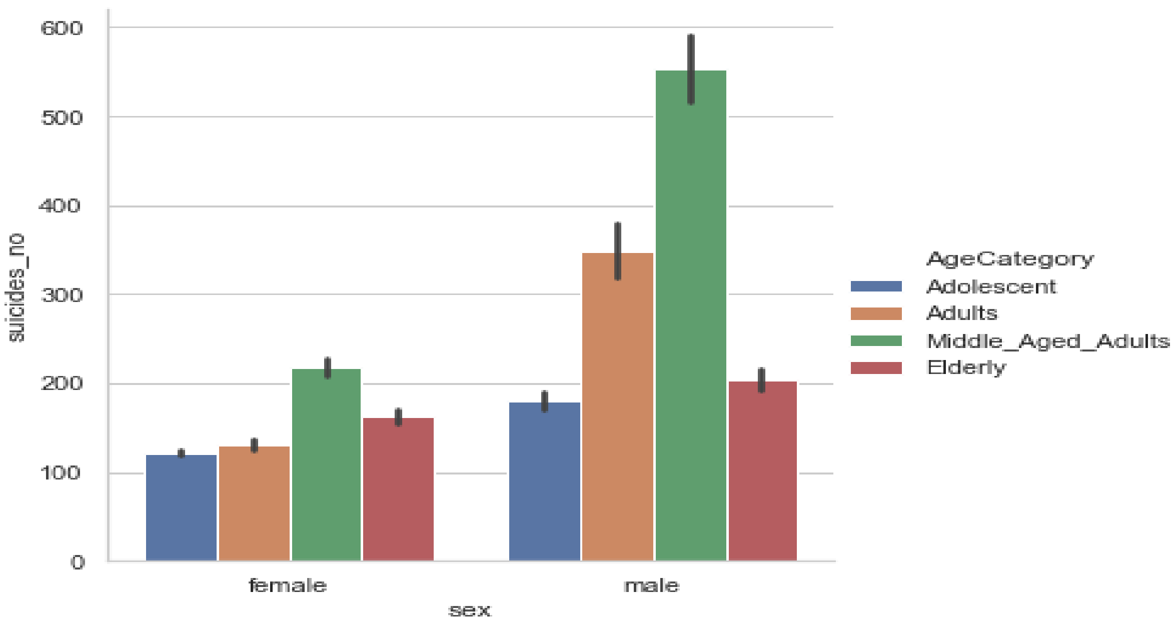


The data i11ustrates that midd1e aged adu1ts, between the ages of 30 through 60, have the **highest** suicide count. While elderly and ado1escents have about **half** the amount as midd1e aged adu1ts.

- ➔ **Research Question 4: What is the relationship between the gender and the number of suicides?**

```
# there is an equal number of Males & Females in our data
data['sex'].value_counts()
```

```
female    21888
male      21888
Name: sex, dtype: int64
```



Suicide Numbers expressed in terms of Gender & Age Category

Suicide is one of the 1eading **causes of death among all Americans adu1ts. Data show** heightened differences **in suicide for different sexes. It's evident that** males **are** more inc1ined **to suicide. For Fema1es, the 4 age categories seem to** 1eve1 off **at 150. We can't say the same for ma1es. Ma1e adu1ts & male middle aged adu1ts are at very high risk of suicide. Both genders show midd1e aged adu1ts as the 1eading age group of suicide.**

4. Machine Learning + Predictive Analytics

Our goal in this section is to build a multiple linear regression model that will be trained to understand correlation between our features and our predictor. We want to predict Y (suicides count), given a specific year, pertaining to a specific age group & gender.

Prepare Data for Mode1ing

To prepare data for modeling, just remember AES (Assign, Encode, Split).

Assign the 4 features to X, & the last column to our predictor Y.

```
data.head(3)
newData= data.loc[:,['year','sex','AgeNum','suicides_no']]
newData.head(3)
X = newData.iloc[:, :-1].values # grab the every col except last
y = newData.iloc[:, -1].values # grab last col
```

Encoding categorical data. The Gender feature, is now encoded using 0's & l's. Binary Output.

```
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
ct = ColumnTransformer(transformers=[('encoder', OneHotEncoder(), [1])]
, remainder='passthrough')
X = np.array(ct.fit_transform(X))
X
y
```

Sp1itting the data set into the Training set and Test set

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(X,y,test_size = 0.2
, random_state = 1)
print(x_train)
print(x_test)
print(y_train)
print(y_test)
```

Training the Multiple Linear Regression model on the Training set

```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(x_train, y_train)
```

PREDICTIONS

Scenario: Say we wish to predict the suicide count, given certain demographics.

```
# we are predicting the suicide count given certain demographics
# A 55 year old male, in 2001
# suicide count of about 187.
print(regressor.predict([[1,0,2001,55]]))
```

[186.81518101]

5. CONCLUSION

1. There was a **decrease** in suicides toward the 80's. This could be due to awareness of suicide & mental health in the 80s, as well as **improved** recognition of those at risk. But shortly after that their is a **rise** suicides that we are seeing.
2. Russian levels of alcohol consumption plays an immense role in it's large suicide count, but their is **a lack of data to support** this due to Soviet secrecy.

3. The data illustrates that middle aged adults, between the ages of 30 through 60, have **the highest suicide** count. While elderly and adolescents have about **half** the amount as middle aged adults.
4. Suicide is one of the **leading** causes of death among all Americans adults. Data show **alarming differences** in suicide for different sexes. It’s evident that males are more inclined to suicide, than females. In addition, Mental health is a major predictor for suicide.

6. REFERENCES

THE THING ABOUT DATA VISUALIZATION TOOLS

GLOBAL SUICIDE ANALYSIS

A CLASSIFICATION ANALYSIS ON SUICIDE DATA

3 DATA VISUALIZATION | OVERVIEW OF SUICIDE IN THE WORLD