INTRODUCTION

Suicide is a global public health problem, with millions of people around the world dying by suicide every year. According to the World Health Organization (WHO), an estimated 800,000 people die by suicide each year, representing one death every 40 seconds. This makes suicide the second leading cause of death among people aged 15-29 years, and the 14th leading cause of death overall.

Suicide rates vary widely around the world, with some countries experiencing higher rates than others. In general, suicide rates are higher in countries with lower incomes and less developed health systems. In high-income countries, the rates of suicide are generally lower, but they have been increasing in recent years.

1. APPROACH

Analyzing global suicide trends is a very critical thing as it involves a lot of socioeconomic factors as well as facilities for mental health care of the entire population. The World Health Organization has released a report on suicide rates around the world, and it shows that suicide rates are the highest in Europe. The report also found that the global suicide rate has decreased by 9.8% from 2000 to 2016. However, there are still some regions where the suicide rate is high, and this is mainly due to the lack of mental health services and support. We have tried to explore each of the factors or parameters that are properly quantifiable. We have tried to exploit every public dataset that is available and has relevance to our aims. Several libraries, like Pandas for Data Handling, Matplotlib & Seaborn for plotting - have been used throughout the project. The entire analysis has been divided into four different iPython Notebooks. The first notebook explores 2 different datasets on suicide mortality rate (per 100,000 population) and income groups by country. They are joined before analysis. The second notebook explores 3 different datasets. The first table contains the Health Development Index by country. The second table contains the highest salary of every country in a time series manner. The third dataset contains unemployment rate in a time series manner as well. The third notebook contains data from two different regions NYC and ROS for various causes. The final notebook explores datasets from WHO on various socio-economic parameters. All of these datasets were joined for final implementation of Machine Learning models like Principal Component Analysis, Simple Linear Regression, Random Forest Regression. They can help us to get the over trend that is being seen in the global context. In the fifth notebook, we try to culminate all of the data and visualize them on a global level. This gives us a larger picture. Categorization by age groups, genders and country is done to understand global trends. Further, a time series analysis for each category is also explored. In the final notebook, correlations between various parameters like HDI and GDP per capita with suicide rates are explored and visualized using a heatmap. Further, overall suicide

numbers are explored using colorful barplots after grouping data by age group, generation, country and gender.

2. DATA

The first iPython notebook contains data from The World Bank (https://data.worldbank.org/indicator/SH.STA.SUIC.P5). This has two different tables the first contains time series data on number of suicides per 100,000 people. The second table contains Country Code, Region and Income Group. Both of these tables are joined using the country code. Its combined size is nearly 160 KB. In the second notebook, 3 different datasets are used. The first dataset contains health indicator scores for each country (https://data.world/hdro/47395724-468d-4552-b1ba-d37e7ef48e8a). The second dataset contains time series data on the highest salary for each country by year (https://www.kaggle.com/datasets/mathurinache/highestsalary). The third dataset time series data on unemployment rates of each country by year(https://www.kaggle.com/datasets/pantanjali/unemployment-dataset?resource=download). The sizes of each of them are 280 KB, 4 KB and 40.7 KB respectively. In the third notebook, data on vital statistics about suicide deaths by age, group, race, ethnicity, resident, county, region, and gender - is explored. But this time it is solely focused on two cities that are known for their diversity and social security combined with the highest living standards. The size of the dataset is nearly 80 KB. The final notebook contains four different datasets. The first is from WHO and contains basic aggregate numbers covering 1979-2016, by country, year, age groups and sex. There is only one file, with only a few columns. It can be accessed from https://www.kaggle.com/datasets/szamil/who-suicide-statistics and its size is nearly 1809 KB. The second 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. It can be accessed from https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016 and its size is nearly 2643 KB. The third dataset contains information on population, region, area size, infant mortality and more. Data from the World Factbook is public domain. The website says "The World Factbook is in the public domain and may be used freely by anyone at any time without seeking permission". This can be accessed from https://www.kaggle.com/datasets/fernandol/countries-of-the-world and its size is nearly 38 KB. The final dataset contains happiness scores and rankings using data from the Gallup World Poll. The scores are based on answers to the main life evaluation question asked in the poll. This question, known as the Cantril ladder, asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale. The scores are from nationally representative samples for the years 2013-2016 and use the Gallup weights to make the estimates representative. It can be accessed from https://www.kaggle.com/datasets/unsdsn/world-happiness and its size is nearly 17 KB.

4. Project AIM

4.1 AIM 1

To check trends emerging in advanced western economies or the global north (ie. United States of America and Canada) compared to the developing country Mexico. The time series data will be plotted to tell the difference between them.

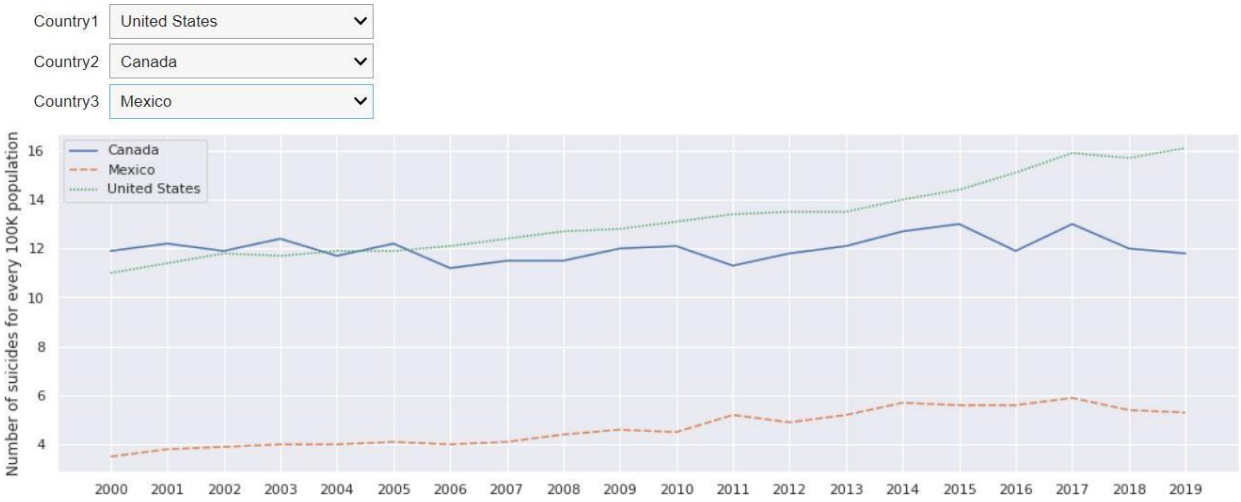
To achieve this, a generalized interactive plotting method was divided so that comparison between other countries can be done with relative ease. The drop down contains the list of country names populated from the Country Name column.

```
option_list = list(df['Country Name'].unique())

country1 = widgets.Dropdown( options
                    = option_list,
                    description='Country1'
                   )
country2 = widgets.Dropdown( options
                    = option_list,
                    description='Country2'
                   )
country3 = widgets.Dropdown( options
                    = option_list,
                    description='Country3'
                   )
```

The selected country names could then be passed onto the plot_by_country function to plot the final time series line plot. And is later invoked to produce the plot that will follow:

```
def plot_by_country(country1, country2, country3): df_country =
 df.loc[df['Country Name'].isin([country1,country2,country3])] df_data =
 df_country.loc[:, years_to_be_studied].T df_data.columns =
 df_country.loc[:,"Country Name"].values plt.figure(figsize=(17,5))
 sns.lineplot(data = df_data)
 plt.ylabel("Number of suicides for every 100K population") plt.show()
```

```
interactive(plot_by_country, country1 = country1, country2 = country2, country3 =
country3)
```



This clearly shows that although Canada and United States have larger economies and Mexico is smaller in comparison, the gap in the overall number of suicides is large. Therefore, it can be concluded that the overall economy of the country has little to do with suicides in the country.

4.2 AIM 2

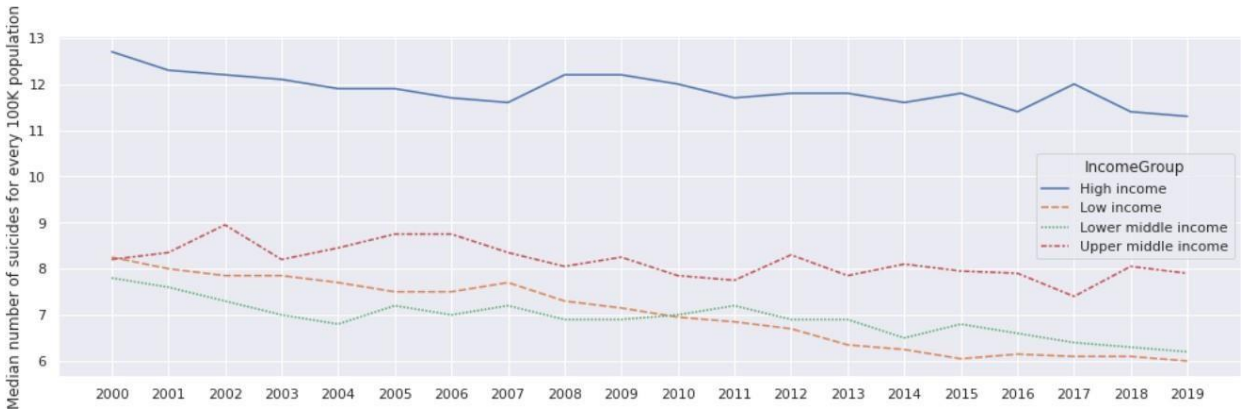
To check how incomes of a country affect the rates of suicide. This is a similar time series plot that will be used. But this time, the data needs to be first sliced into the required columns and then be grouped by income group.

```
cols = years_to_be_studied
cols.append("IncomeGroup") df3 =
df[df.columns[df.columns.isin(cols)]]
```

```
df3 = df3.groupby(["IncomeGroup"])\
        .median()\
        .T
```

```
plt.figure(figsize=(17,5))
sns.lineplot(data = df3)
plt.ylabel("Median number of suicides for every 100K population") plt.show()
```

Following is the plot that we got at the end of this:



After aggregating data on the basis of income group it can be clearly seen that suicides constitute a major percentage of deaths in countries that lie in Higher Income Groups. There is a significant gap in case of high income and others, while the gap between Upper Middle and others is a bit less. Possible reasons:

1. Highest social security in high income countries
2. Better trauma care and other critical care facilities in high income countries.
3. Death by other means like accidents and wars could be a reason in comparatively low income countries.

4.3 AIM 3

To find which specific countries among the high income ones contribute maximum number of suicide cases. We need to shortlist two worst performing countries among them.

First of all, we need to slice the required columns.

```
cols = years_to_be_studied cols.append("Country
Name")

df_higher_income = df[df.IncomeGroup=="High income"]\
                .loc[:,cols]
```

Then we need to find the bottom five countries in terms of number of suicides. This can be generated from the following piece of code.

```
highest_suicides_country_in_high_income = dict() for year in years_to_be_studied:
highest_suicides_country_in_high_income[year] = df_higher_income[[year,"Country
Name"]].sort_values(year)[-5:]["Country Name"].values[::-1] df5 =
pd.DataFrame(highest_suicides_country_in_high_income)
```

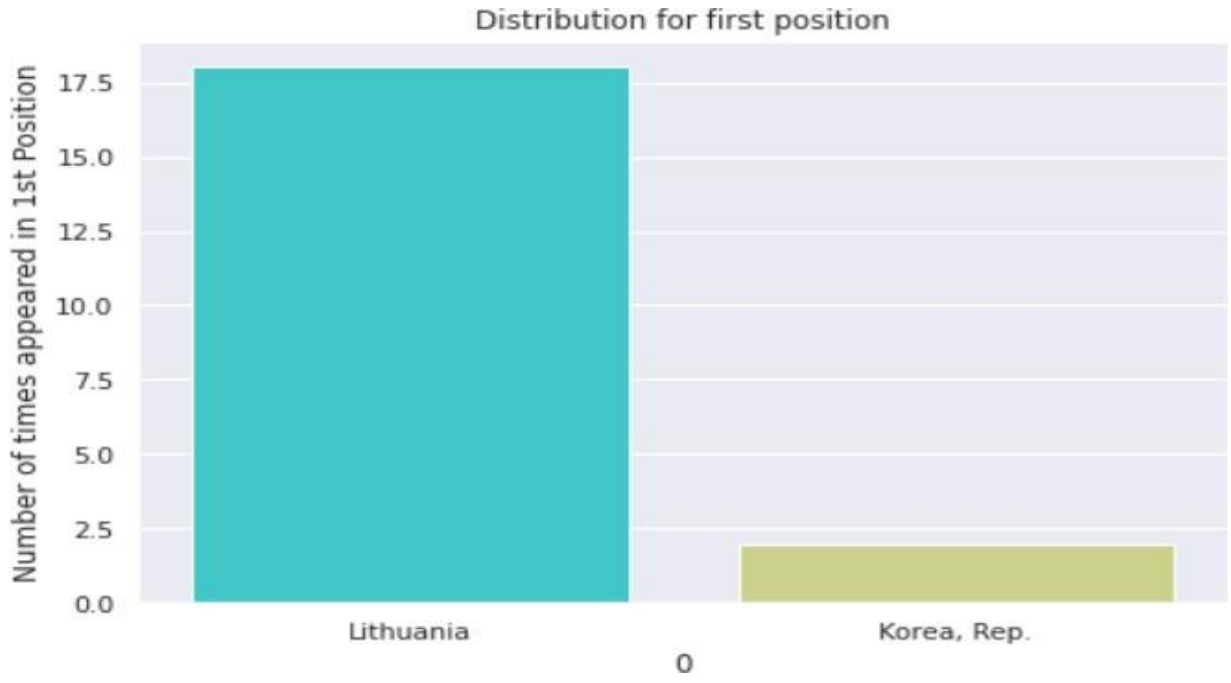
```
df5.head()
```

Following was a the table that was generated from the above code:

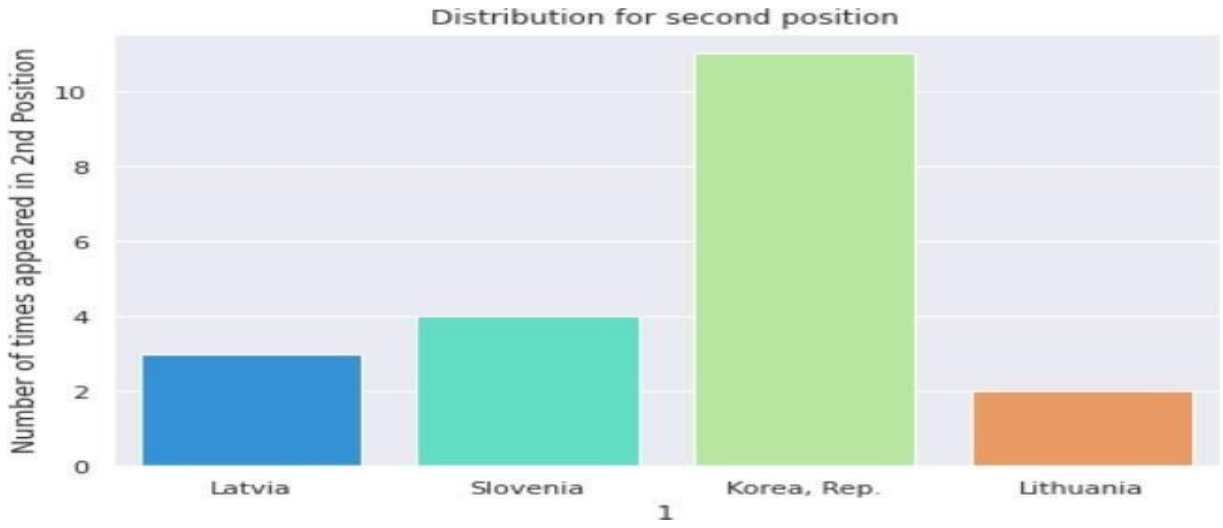
2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Lithuania	Lithuania	Lithuania	Lithuania	Lithuania	Lithuania	Lithuania	Lithuania	Lithuania	Lithuania	Lithuania	Lithuania	Korea, Rep.	Korea, Rep.
Slovenia	Korea, Rep.	Korea, Rep.	Korea, Rep.	Korea, Rep.	Korea, Rep.	Korea, Rep.	Korea, Rep.	Korea, Rep.	Korea, Rep.	Korea, Rep.	Korea, Rep.	Lithuania	Lithuania
Hungary	Japan	Latvia	Latvia	Hungary	Hungary	Hungary	Slovenia	Latvia	Latvia	Latvia	Latvia	Uruguay	Uruguay
Japan	Hungary	Hungary	Hungary	Japan	Japan	Latvia	Hungary	Hungary	Slovenia	Uruguay	Slovenia	Belgium	Latvia
Korea, Rep.	Slovenia	Japan	Japan	Latvia	Latvia	Slovenia	Japan	Japan	Hungary	Belgium	Uruguay	Latvia	Slovenia

We can see that the first position (from the bottom) is almost always occupied by Lithuania except for the last few years. The distribution for the last and and the second last position can tell us more about the rankings and the countries that held them.

```
plt.figure(figsize=(8,5))  sns.countplot(x=0,data=df5.T, palette='rainbow')  plt.ylabel("Number of times appeared in 1st Position") plt.title("Distribution for first position")
```



```
plt.figure(figsize=(8,5))
sns.countplot(x=1,data=df5.T, palette='rainbow')
plt.ylabel("Number of times appeared in 2nd Position") plt.title("Distribution for second position")
```



Following countries were most frequently appearing in bottom 5 of the list:

1. Lithuania
2. Korea

4.4 AIM 4

To determine the region wise distribution of high income countries. This will give us an idea as to which particular areas are high income and which are low income. Based on the distribution, we need to determine the region wise trend in deaths by suicide from the years 2000 to 2019. This will give us an idea as to which places in particular have more suicides per 100,000 population. Since this is a follow up to the economic indicators by region, this can help us to reach a lot of conclusions.

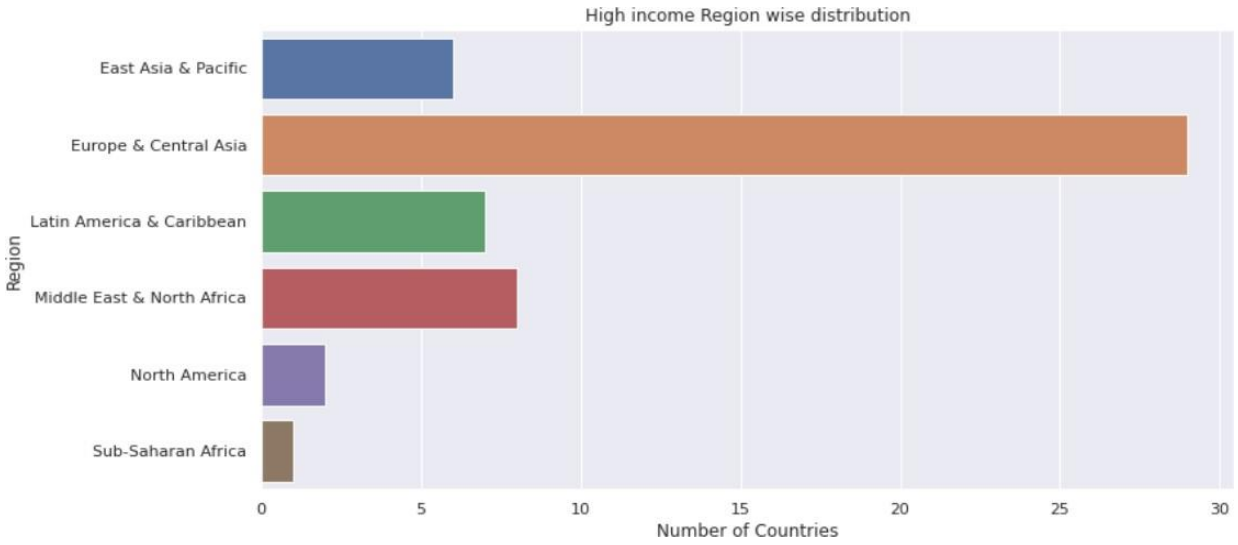
First we need to slice the columns as required.

```
region_wise_income_distribution = pd.DataFrame(region_wise_income.groupby(["IncomeGroup","Region"])["CountryName].count())
region_wise_income_distribution.columns = ["Count"] region_wise_income_distribution
```

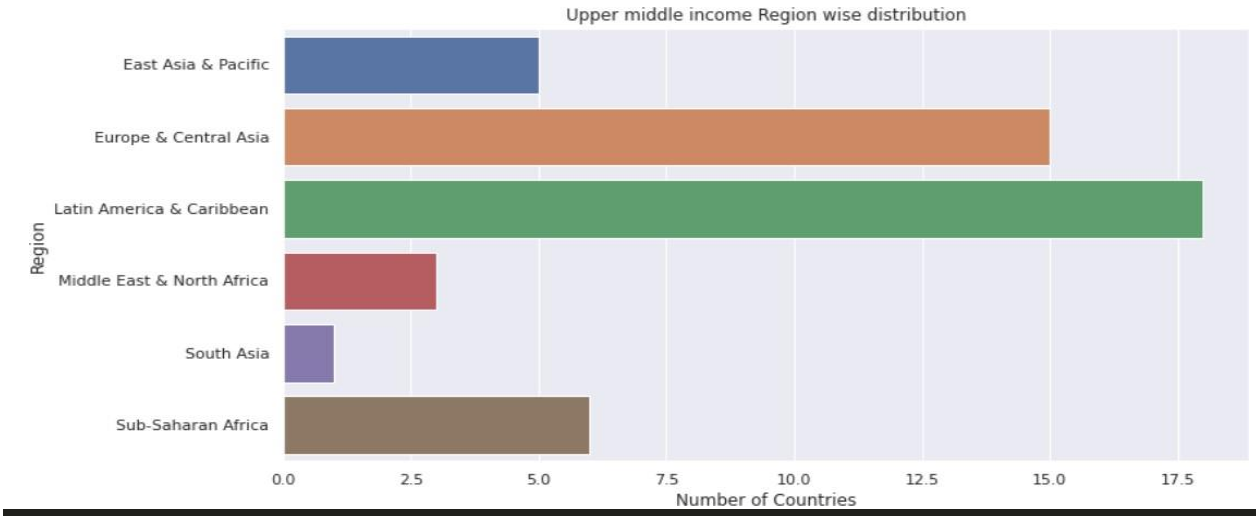
The data is then plotted using a few horizontal bar charts. The function income_wise_distribution takes the income category as input and outputs the required bar plot.

```
def income_wise_distribution(income_category):
    data = region_wise_income_distribution.loc[income_category]
    plt.figure(figsize=(12,6))
    plt.title(f'{income_category} Region wise distribution')
    sns.barplot(data=data, x=data.Count, y=data.index, orient = 'h')
    plt.xlabel("Number of Countries")
    plt.show()
```

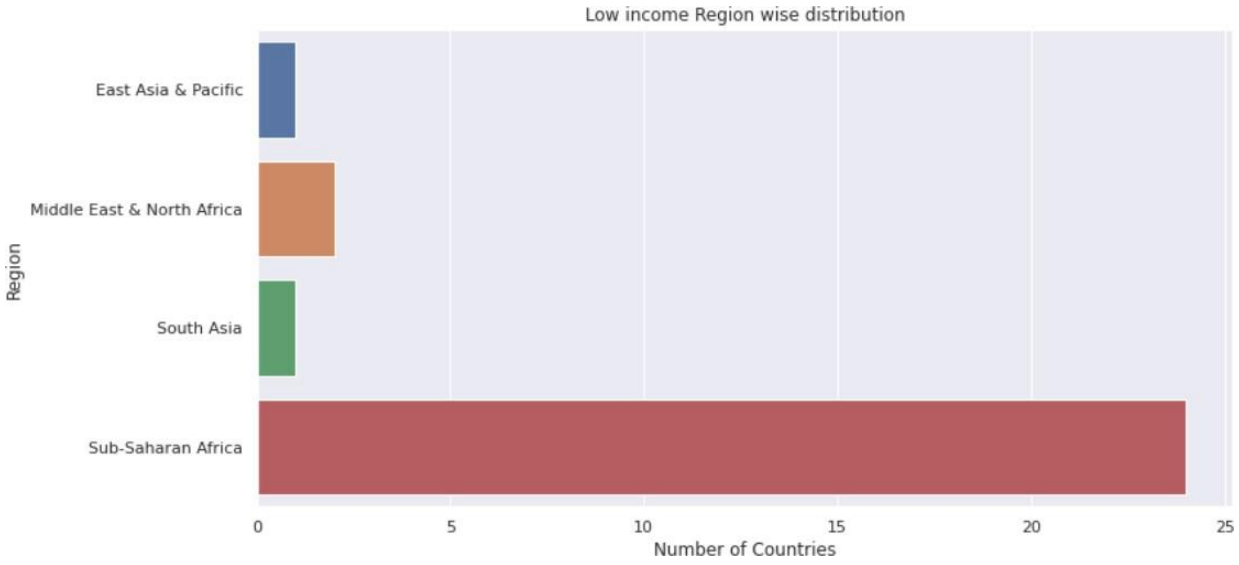
```
income_wise_distribution("High income")
```



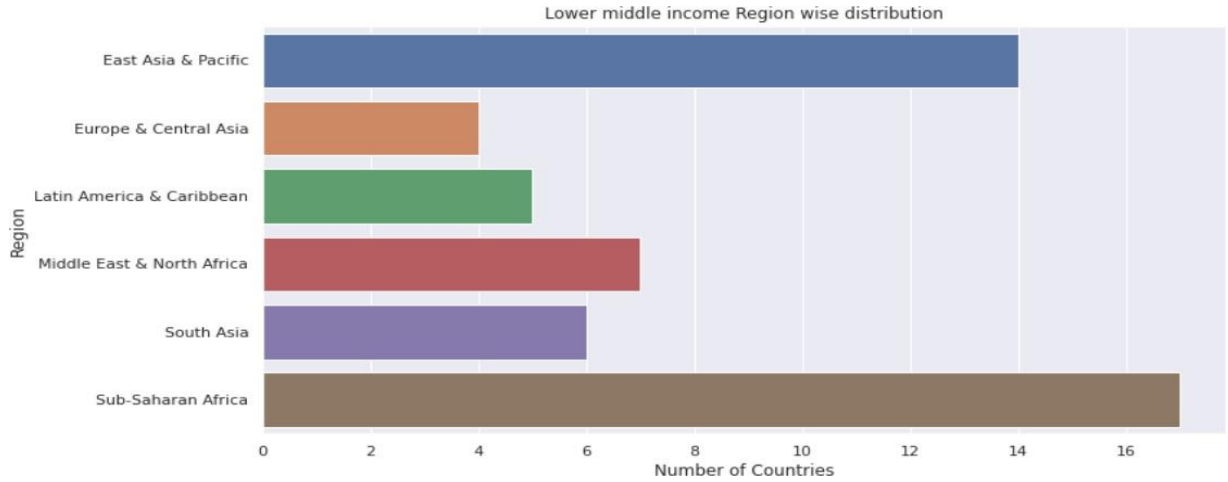

```
income_wise_distribution("Upper middle income")
```




```
income_wise_distribution("Low income")
```




```
income_wise_distribution("Lower middle income")
```



Hence, we can conclude from the above horizontal bar plots that most of the high or higher middle income countries are from Europe & Central Asia or from the Americas & Caribbean. Whereas the lower and lower income countries are mostly from Eastern Asia or from Sub-Saharan Africa.

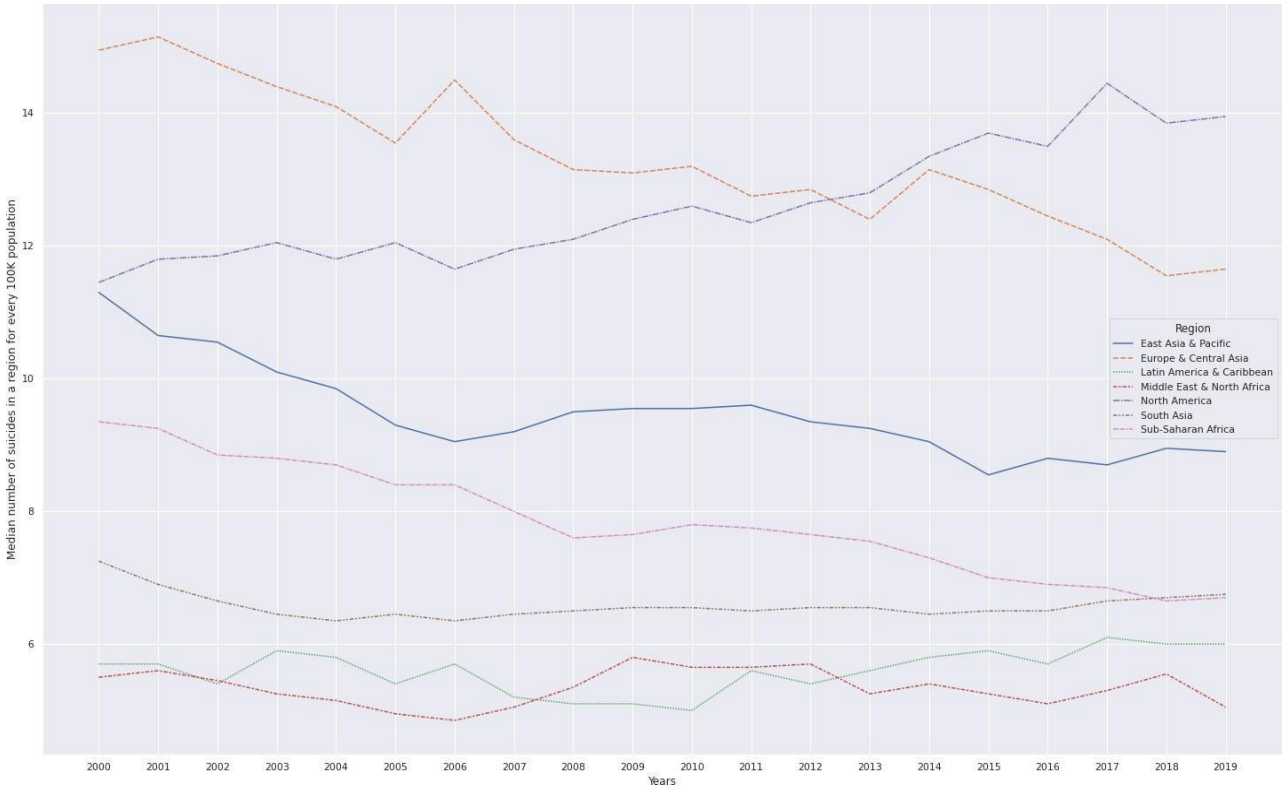
Now we can double down on the analysis by diving deep into the suicide rates for each region. But before all of these, we need to first group by the data on the basis of the region column.

```
region_wise_timeseries_data = df.groupby(["Region"])[years_to_be_studied].median()
```

For the time series line plot we need to use the seaborn library.

```
plt.figure(figsize=(25,15))
plot = sns.line plot(data = region_wise_timeseries_data.T) plt.xlabel("Years")
plt.ylabel("Median number of suicides in a region for every 100K population") plt.show()
```

The result is the following line graph:



We can therefore safely conclude that the two regions ‘Europe & Central Asia’ and ‘North America’ - historically contribute largely to numbers of deaths by suicide. From the previous bar plots we found that these two regions topped in both higher and upper middle income groups. So we can say that the overall economy of a country is inversely proportional to the suicidal deaths in them.

4.5 AIM 5

Does health indicator, salary and unemployment rates have anything to do with the suicides in the country. Since health indicators determine the overall well being of the population of a country in terms of their health and ability to fight endemics from time to time. This can also be the result of some of the schemes introduced by the regimes in power. We also calculate Pearson's correlation value to quantitatively determine the exact levels of correlation between the two.

We can also see how gdp_per_capita affects HDI every year.

	suicides_no	population	suicides/100k pop	HDI for year	gdp_per_capita
suicides_no	1.000000	0.616162	0.306604	0.151399	0.061330
population	0.616162	1.000000	0.008285	0.102943	0.081510
suicides/100k pop	0.306604	0.008285	1.000000	0.074279	0.001785
HDI for year	0.151399	0.102943	0.074279	1.000000	0.771228
gdp_per_capita	0.061330	0.081510	0.001785	0.771228	1.000000

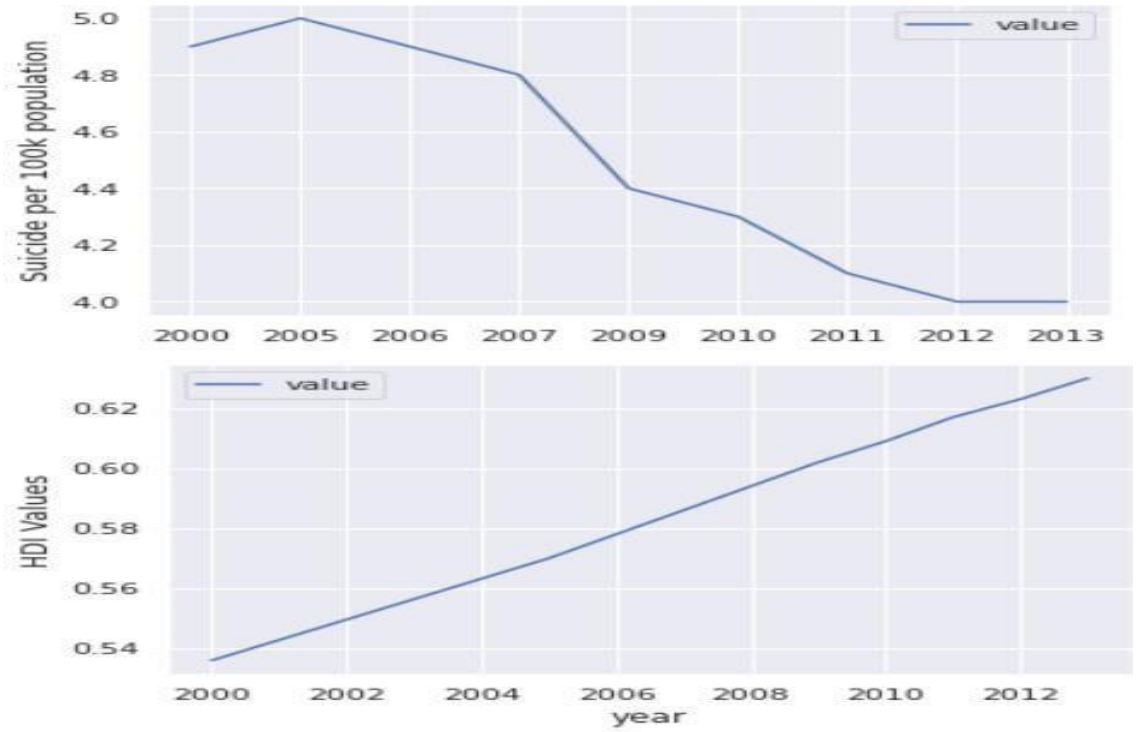
We can clearly see almost all of the parameters positively correlate with each other. The HDI correlates heavily with gdp_per_capita, while the remaining shows positive but too less to determine any relationship.

But before that, we need to create a function that takes the country name as input and outputs two different time series line plots. It should also determine the degree of correlation between the two.

```
def draw_correlation_plot(country_name): afg_health_index =
  df1.loc[df1.country==country_name][["year", "value"]] afg_suicides =
  df.loc[df["Country Name"]==country_name][years_to_be_studied]
  afg_health_index.index = afg_health_index.year
  afg_health_index.drop("year",axis=1, inplace=True) afg_health_index =
  afg_health_index.loc[2000:] years_where_data_is_available =
  afg_health_index.index
  afg_suicides = afg_suicides[[str(year) for year in years_where_data_is_available]].T
  afg_suicides.columns = ["value"]

  sns.line plot(data=afg_suicides) plt.ylabel("Suicide
  per 100k population") plt.show()

  sns.line
  plot(data=afg_health_index)
  plt.ylabel("HDI Values") plt.show()

  corr,  _  =  pearsonr(afg_suicides.loc[:,"value"],  afg_health_index.loc[:,"value"])
  print('Pearson's correlation: %.3f' % corr)
```

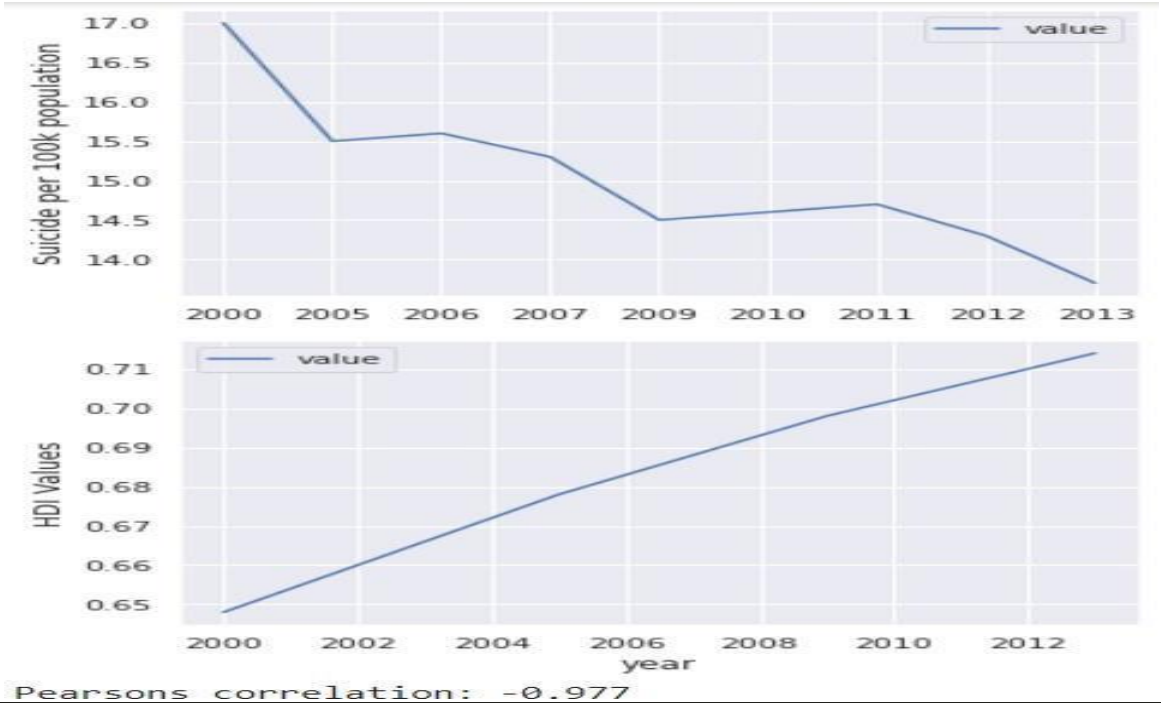
We then plot the same for 3 different countries namely - Afghanistan, India and Lithuania.

```
draw_correlation_plot("Afghanistan")
```



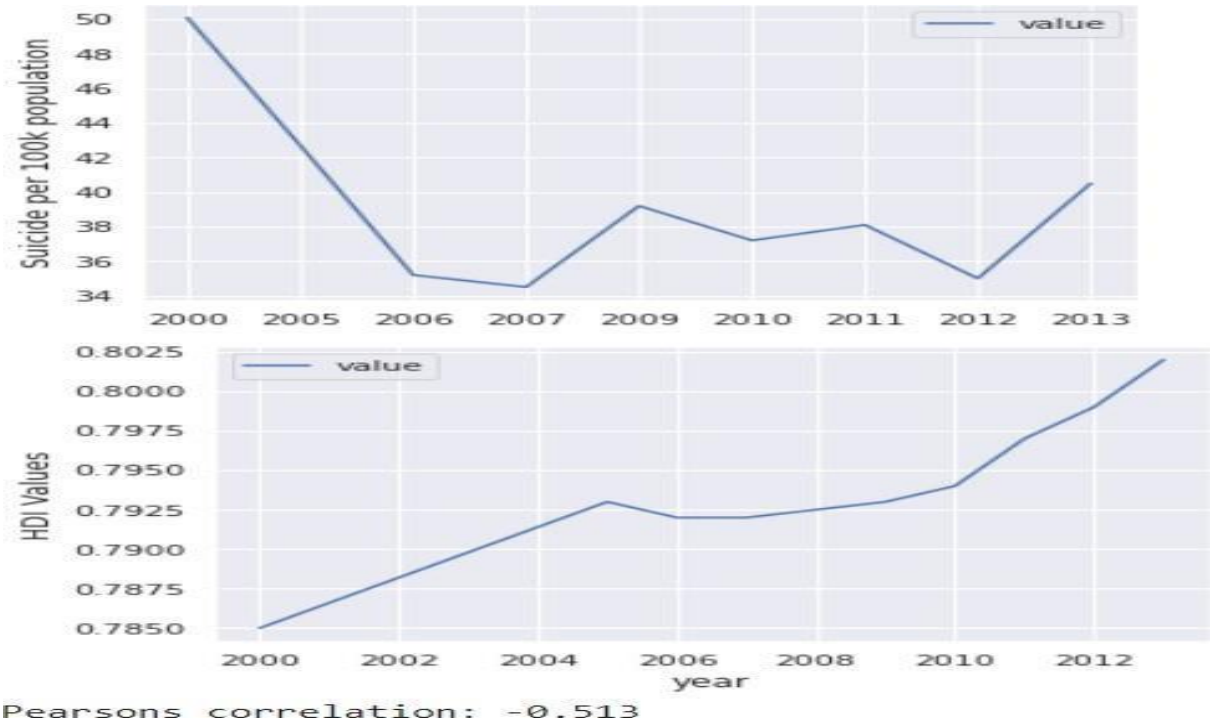

```
Pearsons correlation: -0.900
```

```
draw_correlation_plot("India")
```




```
Pearsons correlation: -0.977
```

```
draw_correlation_plot("Lithuania")
```



We can clearly see that HDI correlates negatively with suicide rates. This implies that countries with weaker health systems are prone to many more deaths by suicides. The trauma care is needed for cases where suicide has been committed and is taken to hospital in the final stages, then emergency trauma care can save a lot of patients.

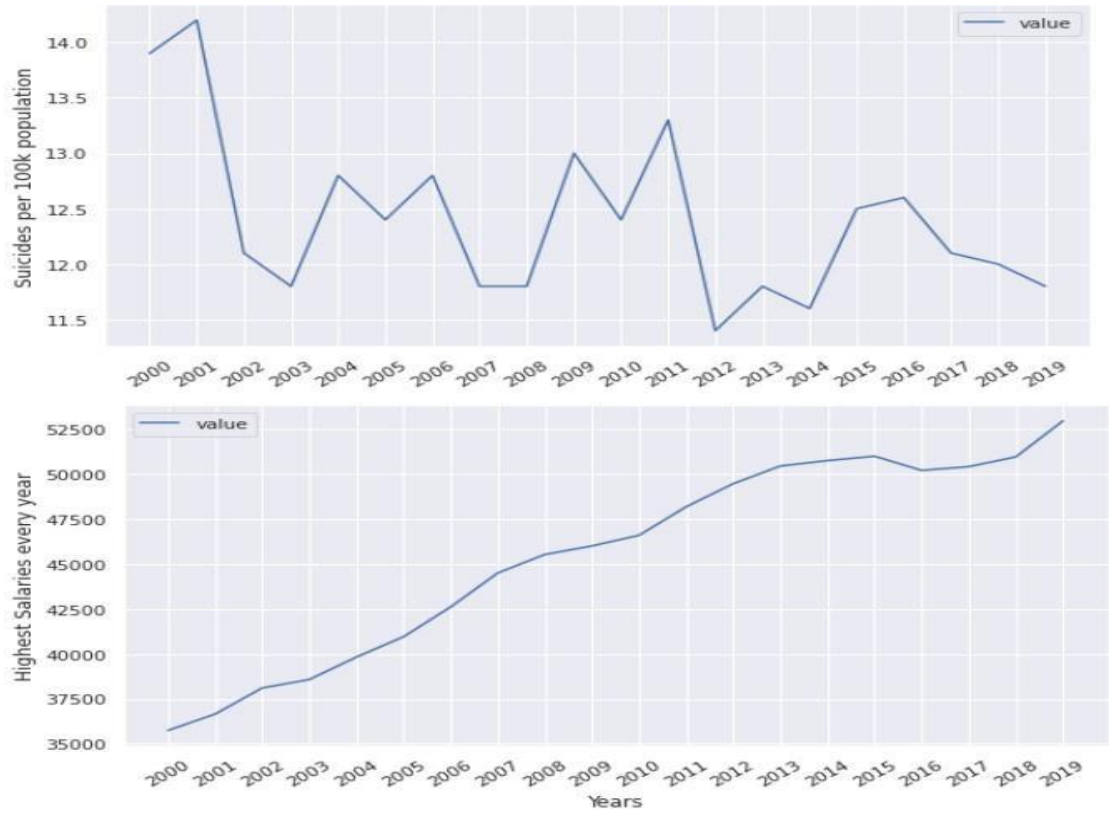
To compare the highest salaries, we need to tweak the earlier plotting function according to our requirements.

```
def draw_correlation_plot_salary(country_name): high_salary =
  df2.loc[df2.Country==country_name][columns].T
  high_salary.columns  =  ["value"]  high_salary.drop("Country",
  inplace=True)
  afg_suicides = df.loc[df["Country Name"]==country_name][years_to_be_studied]
  afg_suicides = afg_suicides[[str(year) for year in years_to_be_studied]].T
  afg_suicides.columns = ["value"]

  plt.figure(figsize=(10,5))              sns.line
  plot(data=afg_suicides)
  plt.ylabel("Suicides per 100k population")
  plt.xticks(rotation=35) plt.show()

  plt.figure(figsize=(10,5))
```

```
sns.line plot(data=high_salary) plt.xlabel("Years")
plt.ylabel("Highest Salaries every year")
plt.xticks(rotation=35) plt.show()

corr, _ = pearsonr(afg_suicides.loc[:,"value"], high_salary.loc[:,"value"])
print(Pearson's correlation:{corr}')
```

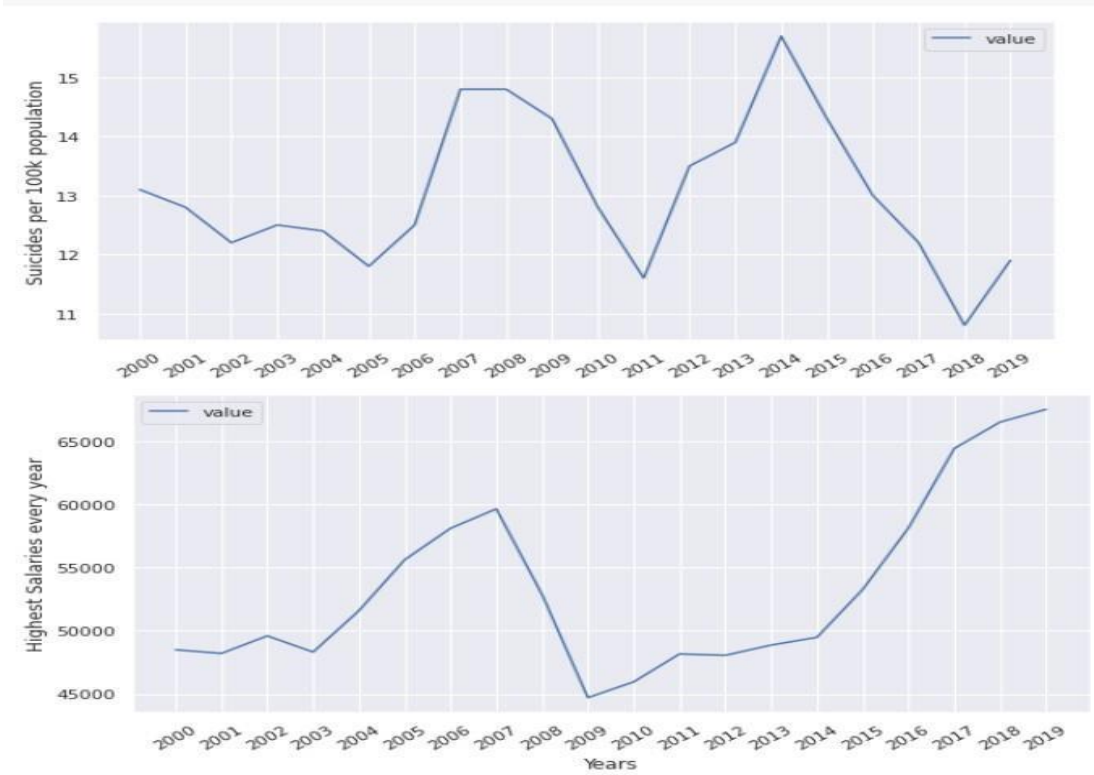
We plot this from 3 different countries, namely - Norway, Iceland, Austria.

```
draw_correlation_plot_salary("Norway")
```



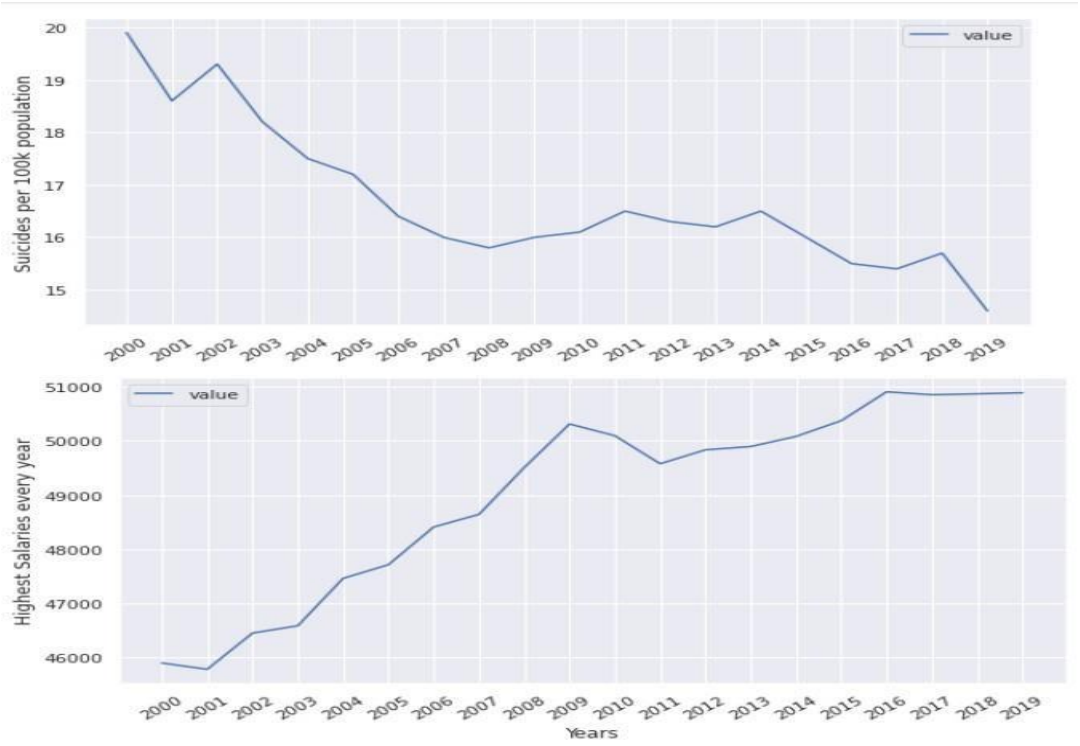

```
Pearsons correlation:-0.5330609198126223
```

```
draw_correlation_plot_salary("Iceland")
```




```
Pearsons correlation:-0.3704373779263976
```

```
draw_correlation_plot_salary("Austria")
```




```
Pearsons correlation:-0.9258268949799148
```

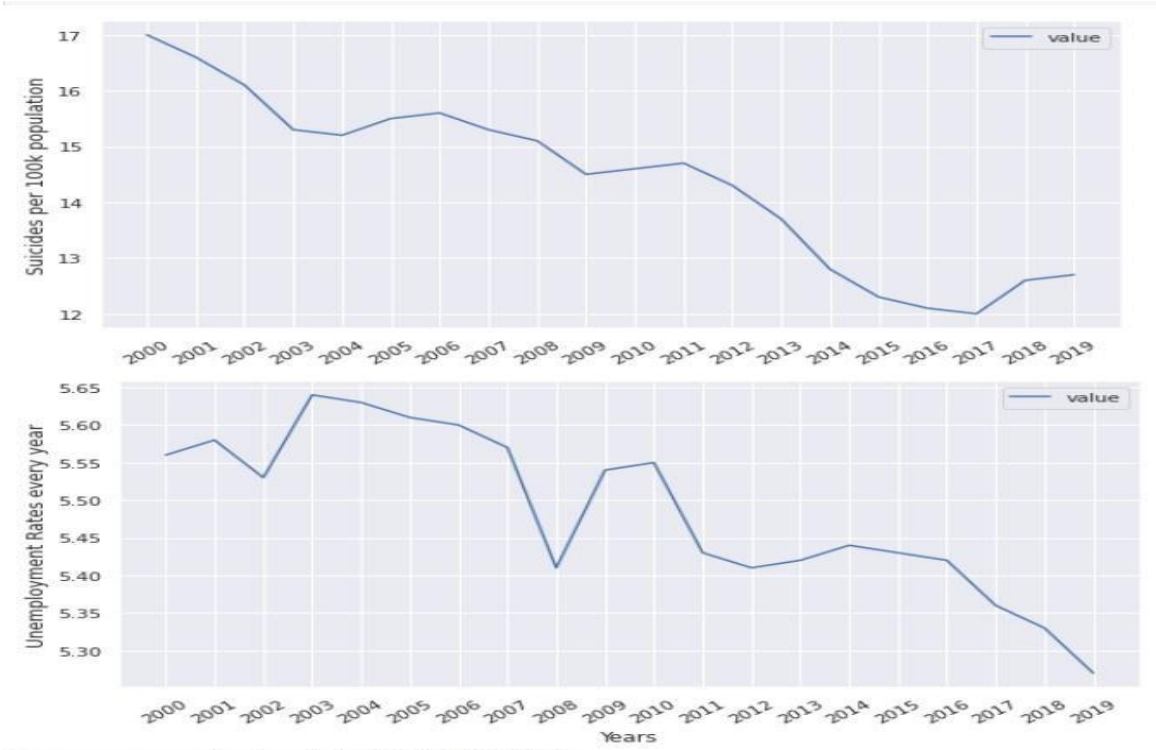
We can safely conclude that slides negatively correlate with the highest salaries. Higher the salaries are lower are the suicide rates. Since, suicides are a result of depression and anxiety stemming from lack of affordability. With higher salaries people can afford a better lifestyle and are not dependent anymore on government schemes that may be critical for their life.

To understand similar trends for unemployment rates we need to first of all tweak the plotting function according to this new requirement. Then we need to invoke it by passing the country name as a parameter. It returns two different plots along with Pearson's correlation value to quantitatively determine the degree of correlation between them.

```
def draw_correlation_plot_unemployement(country_name):
  unemployed = df3.loc[df3["Country Name"]==country_name][cols].T
  unemployed.columns = ["value"]
  unemployed.drop("Country Name", inplace=True)
  afg_suicides = df.loc[df["Country Name"]==country_name][years_to_be_studied]
  afg_suicides = afg_suicides[[str(year) for year in years_to_be_studied]].T
  afg_suicides.columns = ["value"]
  afg_suicides.drop("Country Name", inplace=True)
  plt.figure(figsize=(10,5))                    sns.line
  plot(data=afg_suicides)   plt.ylabel("Suicides   per
  100k        population")        plt.xticks(rotation=35)
  plt.show()
  plt.figure(figsize=(10,5)) sns.line
  plot(data=unemployed)
  plt.xlabel("Years")
  plt.ylabel("Unemployment Rates every year")
  plt.xticks(rotation=35) plt.show()
  corr,    _    =    pearsonr(afg_suicides.loc[:,"value"],    unemployed.loc[:,"value"])
  print(Pearson's correlation:{corr}')
```

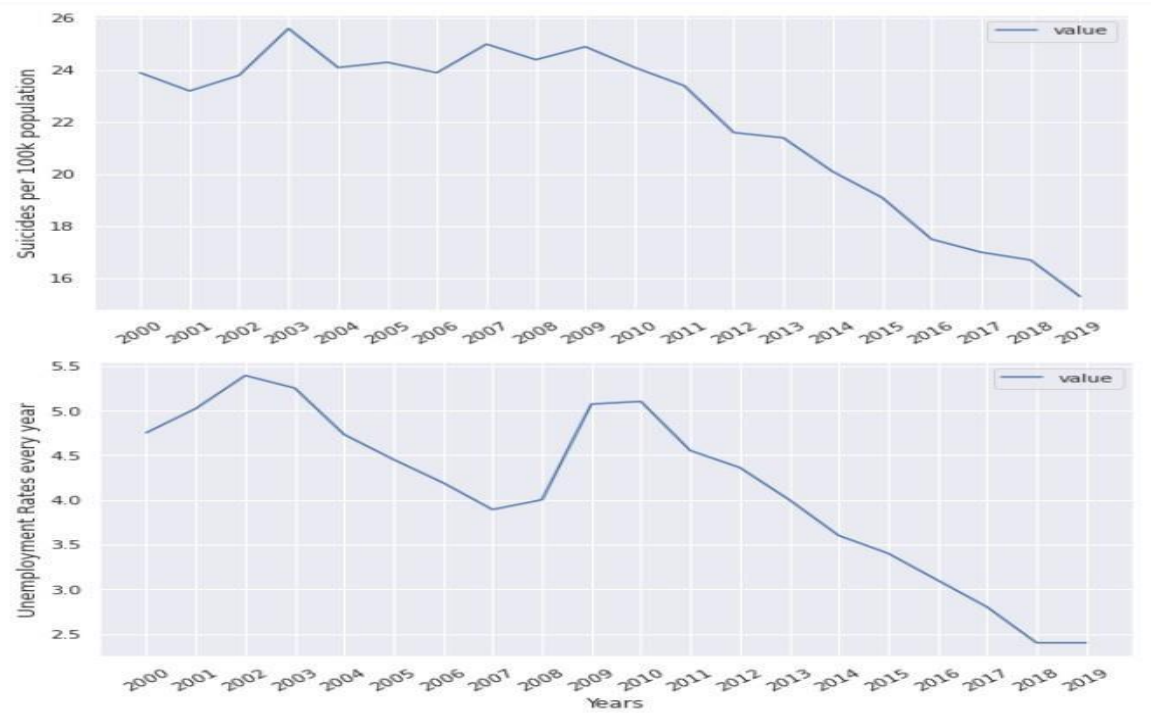
We then call this function thrice for 3 different countries, ie. India, Japan, and Lithuania.

```
draw_correlation_plot_unemployement("India")
```



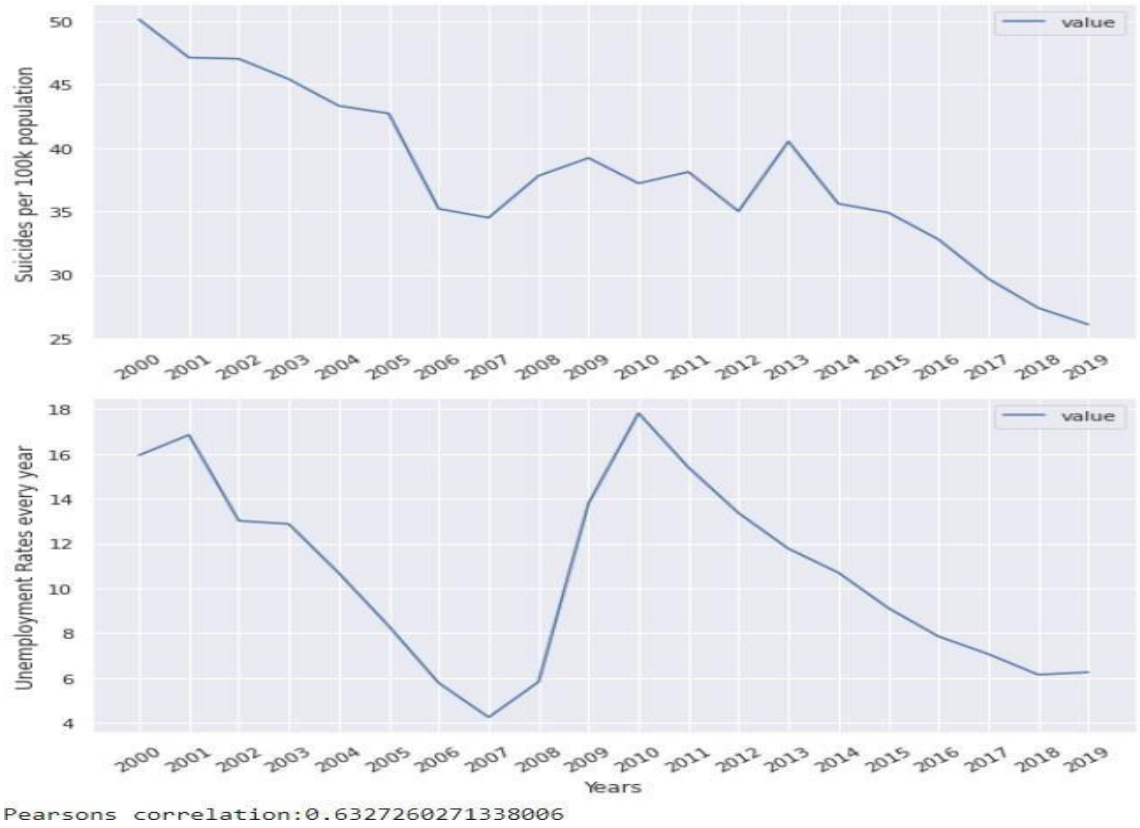

```
Pearsons correlation:0.7588892415313762
```

```
draw_correlation_plot_unemployement("Japan")
```




```
Pearsons correlation:0.8905817757971703
```

```
draw_correlation_plot_unemployement("Lithuania")
```




```
Pearsons correlation:0.6327260271338006
```

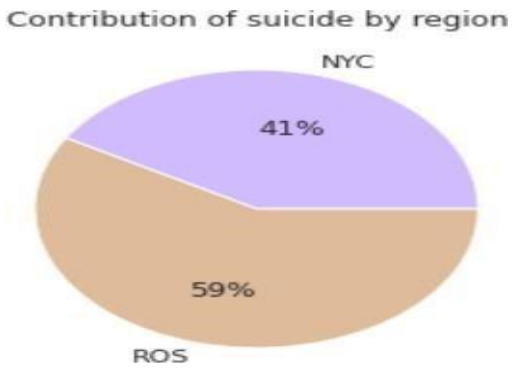
4.6 **AIM 6**

Now we will be dealing with a dataset that focuses only on two bustling metropolitan cities, NYC and ROS. This time series data will be used to explore several different questions. Firstly, we want to understand the contribution in terms of percentage that each of the city has in terms of suicides over the years. Secondly, we want to explore the ethnicity or race of people who are commiting suicides. Thirdly, we want to explore the number of suicides chronologically. We try to understand how different ethinc group have been contributing to the overall toll of suicidal deaths through the years.

For this we need to first group the data by Region column.

```
count_by_region = df2.groupby(["Region"])["Region"].count()
```

Then we use matplotlib to plot the pie chart being used.



We can safely conclude that the contribution of total suicides for both the cities is dominated by ROS.

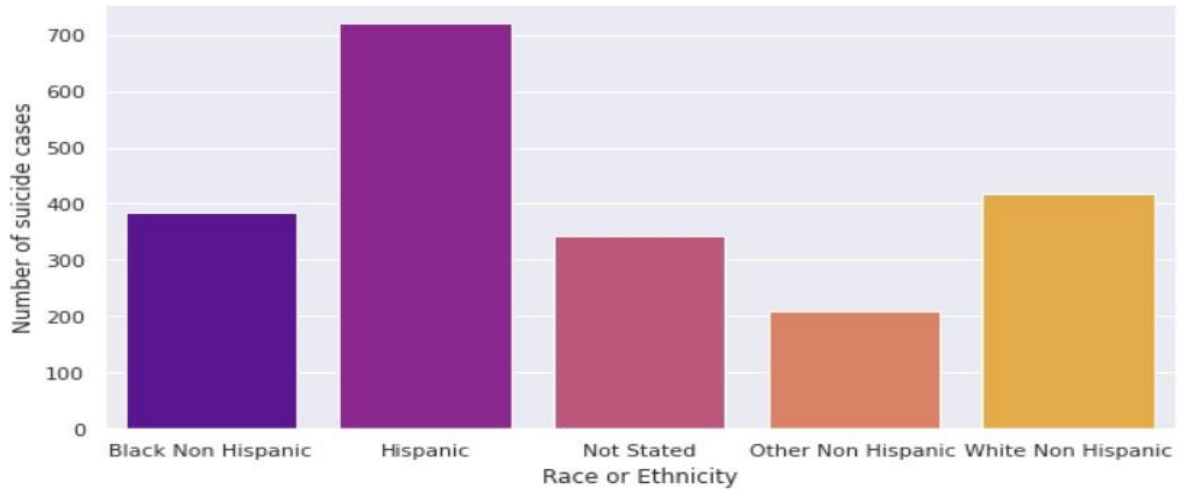
To explore the ethnicity or race of people who are commiting suicides we need to first group by both of the data in terms of ethnicity and then count the entries.

```
count_by_race_or_ethnicity = df2.groupby(["Race or Ethnicity"])["Race or Ethnicity"].count()
```

We then use seaborn's barplot to display the grouped data.

```
plt.figure(figsize=(10,5))
sns.barplot(x = count_by_race_or_ethnicity.index,
            y = count_by_race_or_ethnicity.values, palette
            ='plasma')
plt.show()
```

This results in the following bar plot.



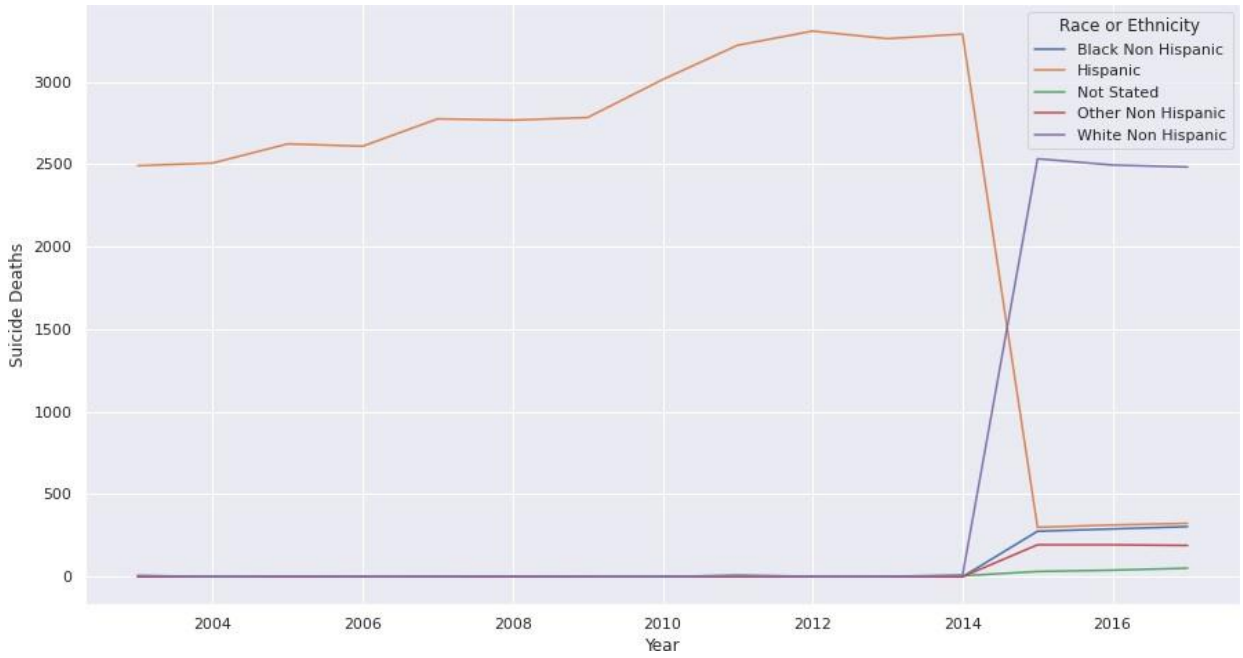
It can be clearly concluded from the given barplot that Hispanics contribute to the maximum number of suicides.

To explore the number of suicides chronologically we need to group the data both by year as well as race/ethnicity. Then we can sum over the total count of deaths by suicide.

```
year_wise_suicides_by_ethnicity = pd.DataFrame(df2.groupby(["Year", "Race or Ethnicity"])["Suicide Deaths"].sum())
```

Then we use a three dimensional seaborn plot for visualizing this newly transformed data. The X axis is years whereas the Y axis is sum of suicide cases. The hue determines the race/ethnicity of the graph.

```
plt.figure(figsize=(15,8))
sns.line plot(data=year_wise_suicides_by_ethnicity, x="Year", y="Suicide Deaths",
        hue="Race or Ethnicity")
plt.show()
```



It can be safely concluded that hispanics have been contributing to the overall cases.

4.7 **AIM 7**

We now want to explore the overall trend of suicides but this time the data will be categorized or grouped by age groups. We also consider the overall deaths by suicide. For this we need to calculate total deaths over the years too.

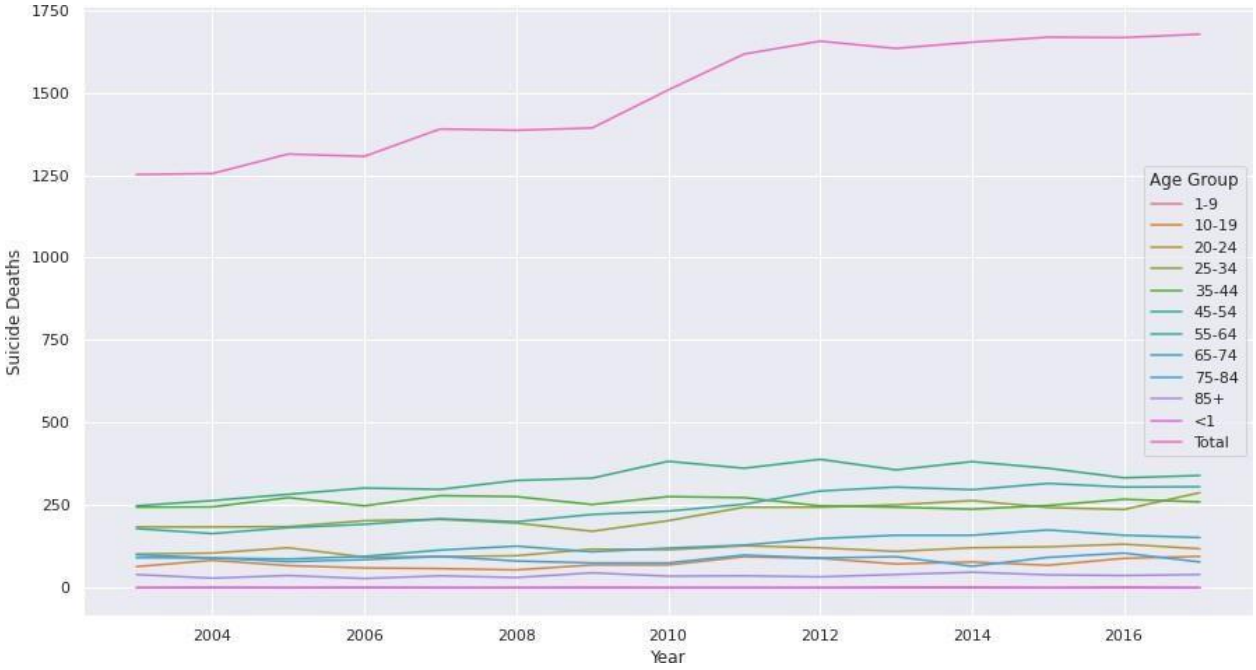
For this we need to group by the data according to Year and Age Group.

```
year_wise_suicides_by_age_group = pd.DataFrame(df2.groupby(["Year", "Age Group"])["Suicide Deaths"].sum())
```

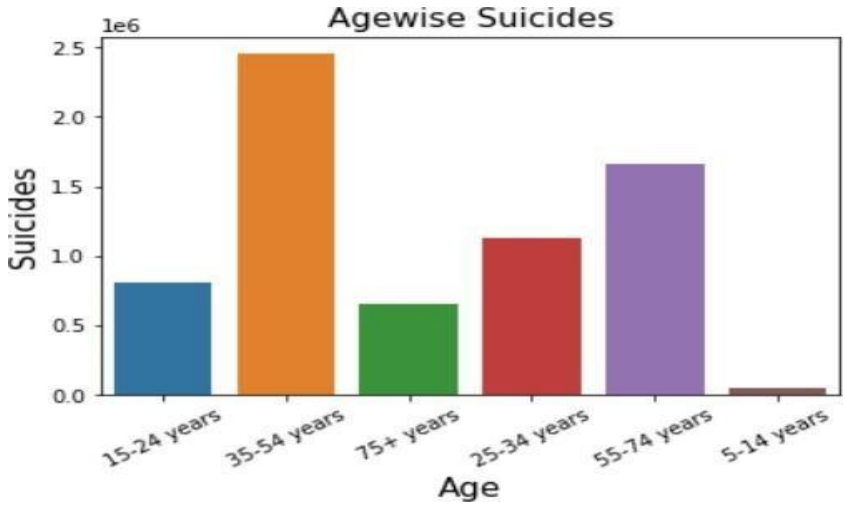
We can then plot this data using seaborn's line plot function.

```
plt.figure(figsize=(15,8)) sns.color_palette("hls", 8)
sns.line plot(data=year_wise_suicides_by_age_group,
        x="Year", y="Suicide Deaths", hue="Age Group")
plt.show()
```

This results in the following line plot.



We can conclude that age groups of 35-44 and 45-54 contributes to the maximum cases of deaths by suicides. The overall deaths also do not see a rapid growth in cases after 2011 which means the situation was always under control by the authorities there.
The same conclusion can also be drawn by visualizing global data.



Globally it can be seen that people from the age group of 35-54 commit a lot more suicides than any other age groups.

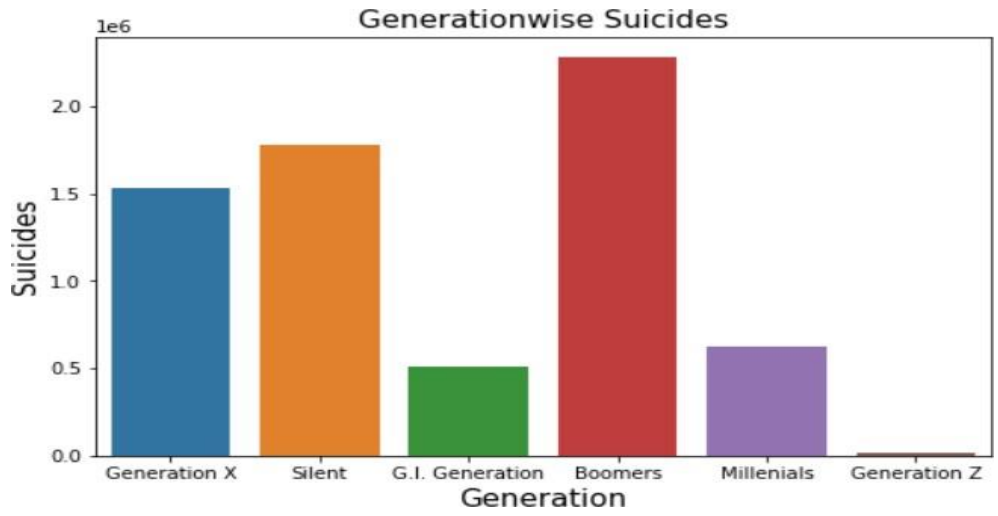
4.8 **AIM 8**

Here we want to determine which generations in particular contribute to the highest number of suicides in the country.

For this we need to plot a barplot to determine the overall distribution of suicide cases among various generations. The following is the code required to achieve the objective.

```
suicides_no_gen = [] for g in
df['generation'].unique():
    suicides_no_gen.append(sum(df[df['generation'] == g]['suicides_no']))
plt.figure(figsize=(8,5))
sns.barplot(x = df['generation'].unique(), y = suicides_no_gen)
## Set the tick and label fontsize
plt.tick_params(labelsize = 10)
## Set the title
_ = plt.title("Generationwise Suicides", fontsize = 15)
## Set the y-label
_ = plt.ylabel("Suicides", fontsize = 15)
## Set the y-label
_ = plt.xlabel("Generation", fontsize = 15)
```

The plot that comes as output for the given code is as follows:



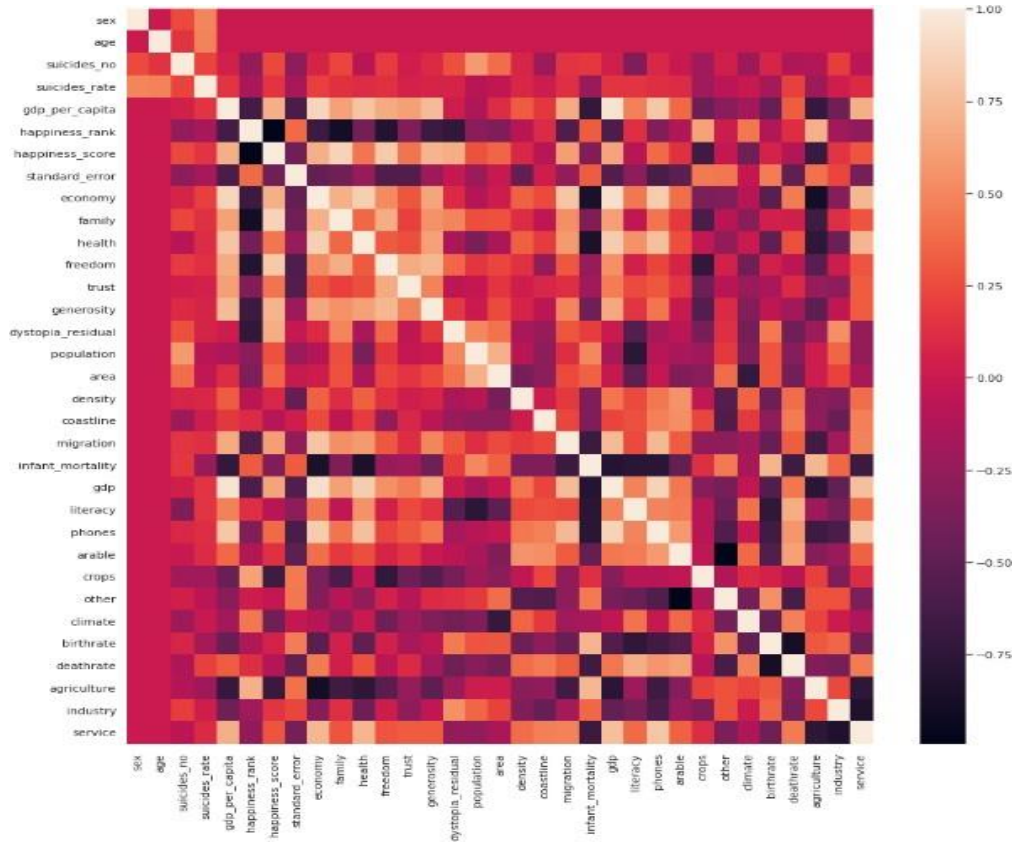
From this we can safely conclude that Boomers have topped the entire tally followed by Silent and Generation X. These 3 contribute to the maximum number of suicides world wide. The prime reasons for this may be: lack of social life, money problems, health issues, over burdened with responsibilities.

4.9 **AIM 9**

From the dataset that we generated by joining several smaller ones in SRA-IV, we try to subset a few quantitative columns that could be used for training the regression models. Following are the columns that were initially subsetted:

- sex
- age
- suicides_no
- suicides_rate
- gdp
- gdp_per_capita
- generation
- region
- happiness_rank
- happiness_score
- standard_error
- economy
- family
- health
- freedom
- trust
- generosity
- dystopia_residual
- region
- population
- area
- density
- coastline
- migration
- infant_mortality
- gdp
- literacy
- phones
- arable
- crops
- other
- climate
- birthrate
- deathrate
- agriculture
- industry
- service

We can then slice these columns and plot a correlation plot using heat maps to find which of the columns correlates with each other to the maximum. First of all, we need to calculate the correlation matrix. Then, on the basis of the correlation matrix we need to plot the heatmap.



From this heatmap we can easily determine the columns that correlate with each other to the maximum, only those who positively correlate(lighter shade) and those who negatively correlate(darker shade) will be retained for the machine learning models. The remaining column that shows no correlation(reddish hue) can therefore be eliminated from further analysis.

The following are the columns that need to be shortlisted based on the above correlation heatmap:

- suicides_rate
- sex
- age
- population
- migration
- infant_mortality
- deathrate
- climate
- literacy
- gdp_per_capita
- happiness_score

4.10 AIM 10

We need to perform regression analysis on subsetted data. Columns like gender, age_band, population, net_migration, infant_mortality, death_rate, climate, literacy, gdp_per_capita, happiness_score are used for calculating the suicide_rate.

We use Scikit learn's simple linear regression and random forest regression for the task. After training each of the models, we check the mean absolute error for each of them. We can get the Simple Linear Regression model using the following code:

```
self.lr = LinearRegression().fit(self.x_train, self.y_train)
y_pred_lr = self.lr.predict(self.x_test)
mean_absolute_error(y_pred_lr, self.y_test)
print(f'Mean Absolute Error is {mean_absolute_error(y_pred_lr, self.y_test)} for linear regression model.')
```

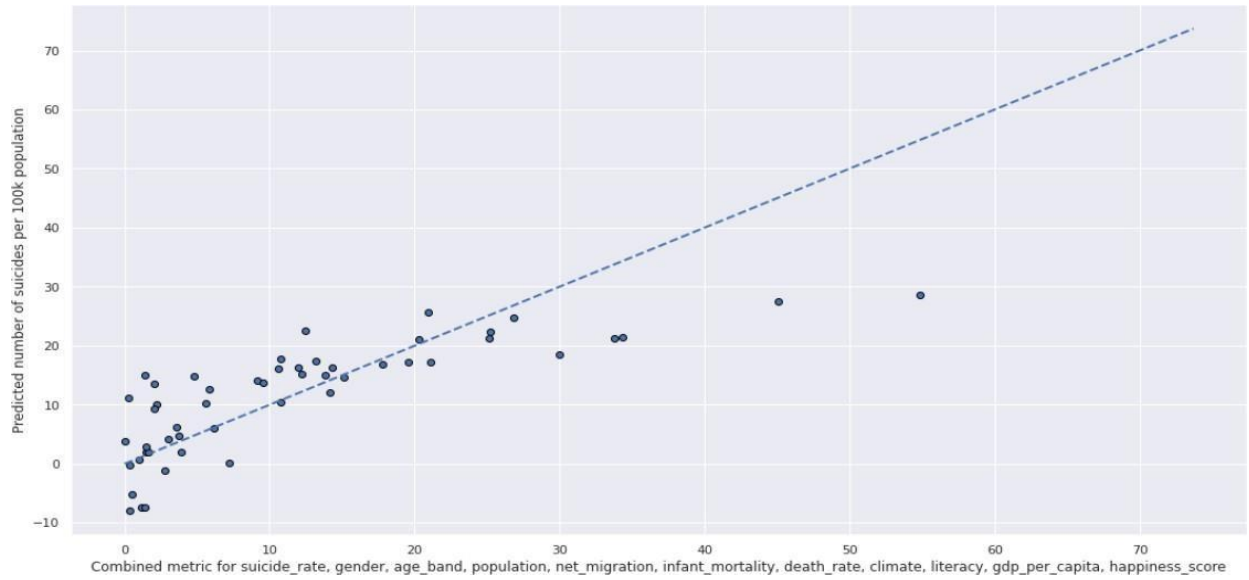
Similarly we can train a random forest regression model using the following:

```
self.rf_model = RandomForestRegressor(random_state=8).fit(self.x_train, self.y_train)
y_pred_rf = self.rf_model.predict(self.x_test)
print('Mean Absolute Error is %2f for RandomForest Regression model after Train-Test split.' %(mean_absolute_error(y_pred_rf, self.y_test)))
```

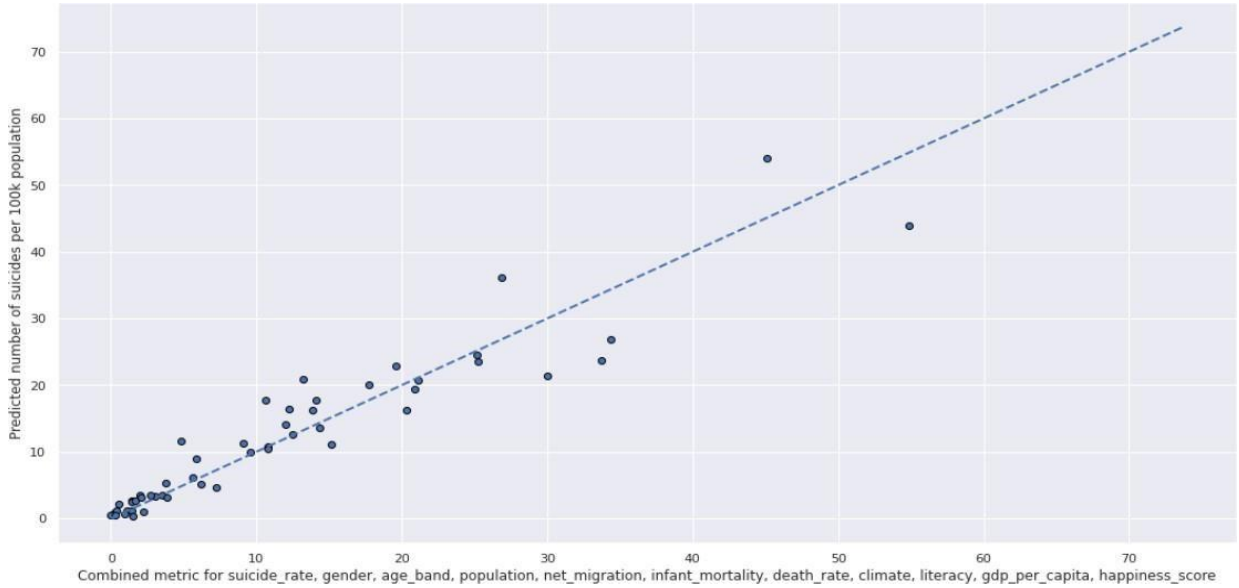
The output of both of them can be plotted using scatter and line plots all on one graph.

```
fig, ax = plt.subplots(figsize=(18,9))
ax.scatter(self.y_test, y_pred, edgecolors=(0, 0, 0))
ax.plot([self.y.min(), self.y.max()], [self.y.min(), self.y.max()], 'b--',lw=2)
ax.set_xlabel(f"Combined metric for {', '.join(self.col_names)}") ax.set_ylabel('Predicted number of suicides per 100k population') plt.show()
```

Following is the resultant plot for simple linear regression:



Following is the result for Random Forest Regression:



With both of them having very low mean absolute errors of 5.48 and 2.63, we can safely say that all of these parameters combined can definitely predict the number of suicides in 100k population. Although each of them may or may not positively correlate with the suicide numbers, but the combined metric containing all of them definitely correlates positively.

5. DISCUSSION

In the initial exploration we deal with a lot of time series data. We tried to plot various parameters against time. For example, in the first plot we try to understand the overall suicides per 100,000 people in 3 different countries. This gives us the freedom to choose between three different countries of our choice and see the trends emerging in them. Here

we choose three different countries, namely - United States, Canada and Mexico. Two among them are developed while the third is a developing economy. Contrary to the general belief it was Mexico, a developing country with several sociopolitical issues stood better than its developed counterparts. We further explore topics like how income groups affect suicide rates. We created a time series line chart to finally conclude that it is actually the high income and the upper middle income groups that contribute to the maximum cases for suicides. We then try to zero down on the few countries that specifically perform worse even when they are on the higher income side. The two countries that performed severely worse were the Republic of Korea and Lithuania. We then tried to determine the region wise distribution of high income countries to give us an idea as to which particular areas are high income and which are low income. This indeed gave us a lot of ideas as to how different regions have people belonging to the higher income category. We found that most of the sub saharan african region was poorest amongst all. While the regions of Europe and North America were the richest. To double down on the exploration spree we further decide to plot the time series data containing suicide rates of all the countries grouped by regions. This showed that the richer european and the northamerican countries occupies the larger chunk in the overall suicides distribution. We then try to find the correlation between various parameters like Human Development Index, Highest Salaries and Unemployment Rates. We then try to calculate Pearson's correlation value between both of those data. We found that there is a negative correlation between that of the Human Development Index and Highest salaries but there was a positive correlation for Unemployment Rates. We then deep dive into regional data of two different bustling metropolitan cities known for their high economic prospects and higher standards of living. A relative comparison is drawn using Pie charts and Bar Charts. The Bar Charts were being used to determine the distribution of various ethnic groups that contributed to the suicides. We then do a time series analysis of them. Further before conducting the regression analysis on the joined data we use the correlation heatmap to come up with a list of columns that correlates to the maximum and can therefore contribute to the overall function. We then conclude our analysis with two different Regression analysis plots.

CONCLUSION

Following are the conclusions that can be drawn from the above analysis:

- Overall economic parameters of the country can not strongly tell anything about the rates of suicide. Moreover, it is the economically developed countries that contribute to the maximum number of deaths by suicide.

- The higher income countries especially the north americas and the european region countries contribute to more cases than any other. Which means higher income is not a solution to depression leading to suicide.

- Among the high income countries, the Republic of Korea and Lithuania perform the worst in terms of number of suicides. Since there is the combined data

available for both the North as well as South Korea we can say the combined data may be skewed.

- The sub saharan african region has the highest number of countries that fall in the lower income category, while the countries from europe and north america contribute to the higher and upper middle income tally.

- Health Index and Highest Salaries correlate negatively means higher the salary and better healthcare lower are the chances that anyone will die by suicide.

- Unemployment Rates correlate positively, which means higher the unemployment rate more are the people susceptible to depression leading to suicide.

- The Hispanic ethnic group/race contributes to more suicide cases in the bustling metropolitan cities of NYC and ROS.

- Age groups of 35-44 and 45-54 contributes to the maximum cases of deaths by suicides.

- Boomers, Silents and Generation X contribute to the maximum number of suicidal cases. This may be because people from these generations are responsible for the well being of the entire family. Hence they deal with a lot of mental pressure, leading to taking their lives as the last resort.

REFERENCES

[1] https://ourworldindata.org/suicide#:~:text=Globally%20800%2C000%20people%20die%20from,of%20death%20in%20young%20people.

[2] https://www.who.int/news-room/fact-sheets/detail/suicide

[3] https://en.wikipedia.org/wiki/List_of_countries_by_suicide_rate

[4] https://worldpopulationreview.com/country-rankings/suicide-rate-by-country

[5] https://www.iasp.info/references/

[6] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3367275/

[7] https://www.washingtonpost.com/wp-srv/world/suiciderate.html

[8] https://groundreport.in/top-10-countries-with-the-highest-suicide-rate/