Project Report

Suicide rates and risk - Prediction and Analysis

Aishwarya Ponkshe

Abstract :

Suicide and depression is precarious and a serious problem faced by people all over the world. Reducing suicide rates and helping the society in combating depression is the need of the hour. With increasing stress levels due to unemployment, low income and high debts, suicide rates have also started to rise up. In the year 2015 alone, about 788,000 people died by suicide at a global rate of 10.7 per 100,000 person-years. Suicide risk prediction and assessment is extremely necessary and recommended and these huge numbers is an indication of how important it is to determine and analyze suicide rates in advance. This project aims at finding insights and patterns in the historical data about suicide cases by visualizing the data. It also aims at predicting whether a group of individuals is at a risk of suicide or not so that the necessary help can be provided to these people.

Introduction :

The enormous number of suicide rates motivated me to work on this topic for my project. These numbers can be reduced to a great extent if machine learning is used appropriately for the benefit of the society. Following is the link to the dataset that was found on Kaggle. It was very advantageous in attaining the goal of this project - https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

There were four goals that were achieved using this dataset -

1. Predict the number of suicides based on the input predictors to the model.
2. Classify the instances in the dataset based on whether they are at a risk of suicide or not.
3. Clustering the instances in the dataset based on their economic condition.
4. Forecasting the total number of suicides for the future years by training the model on the historical data.

Research Questions :

In order to attain the goals of the project it was important to analyze and visualize the trends and patterns in the data. Below are the research questions that would prove helpful before model building-

- Which countries have the highest average number of suicide cases?
- How has the suicide rate changed over a period of time? Is there an increasing or a decreasing trend?
- Which age group, generation and gender is at a higher risk of suicides?
- Which factors affect the number of suicides?
- What is the distribution of the data with respect to suicides per 100k of the population?
- Depending on the input predictors, which instances in the dataset are at a higher risk of suicides?
- Can this data be used for forecasting the number of suicides for future years so that suitable action can be taken well in advance?
- Is there a correlation among predictors? If yes, can one variable be used to replace the other one?

In order to answer the above research questions, I have used suitable machine learning algorithms and used python visualization libraries.

Methods :

i) Data set description:

The data in the dataset is dated back to 1985 up until 2016 and is still being updated today. This dataset can be accessed via the link provided above. This dataset consists of the following attributes –

1. Country
2. Year
3. Sex
4. Age group
5. Number of suicides
6. Population
7. Suicides/100k population
8. Human Development Index for the year
9. GDP for the year in dollars

An interesting aspect about this data is that it contains predictors such as Human Development Index and Gross Domestic Product which can be used to examine whether the economic condition of a country is affecting the mental health of its citizens and thereby causing an increase in suicide rates.

ii) Data treatment and cleaning:

A parsed version of the dataset from 1990-2014 was used to better interpret plots and save processing time**.** Records where the 'year' was missing was removed from the dataset.

One important observation that was made from the dataset was that 80% of the records had null values for 'HDI' or 'Human development index'. One alternative while performing data cleaning was to remove this column entirely. However, after performing some research, it was found that HDI is very important in analyzing suicide rates. HDI is the geometric mean of life expectancy, education and Gross National Income per capita (GNI). Many studies have shown that these components of HDI help in identifying groups of high and low suicide risk individuals since suicide rate varies significantly between countries with different levels of development. To fill in the exact values for HDI instead of imputing them, a dataset was found that contained country and year-wise HDI values. This dataset was found on the 'The United Nations Development Programme' website under Human Development Reports. Following is the link to this website - Human Development Data (1990-2018) | Human Development Reports . From this link the 'Human Development Index (HDI)' dimension was selected and the dataset was downloaded.

From this dataset, the HDI values for a specific country for a specific year were extracted and loaded into the main dataset by substituting the missing values in the 'Human Development Index' attribute. A tremendous amount of data cleaning, unpivoting, merging and transformation was required for this task since the entire HDI data was highly unstructured. It was stored in the form of multivalued indexes, column names were in the first row, they had to be unpivoted and stored into a new DataFrame and finally merged to match with the values main data set. The Human Development Index (HDI).csv also had a few missing values and those records were removed from this dataset.

Apart from this, another data treatment that was needed was – Converting the data types as required. Most of the predictors were categorical which is why they had to be converted to the type - 'category' for the purpose of model building. Also, before building a model, it is important to encode the categorical predictors in the model. Label Encoder was used to encode the categorical predictors – Country, Sex, Age, Generation and Year. Here the ‘Age’ attribute is not a numerical value in the dataset, but instead it is an Age group range, which is a string such as ‘35-54 years’, ‘15-24 years’ and so on. Hence, ‘Age’ has been treated as a categorical variable. Scaling has been performed using ‘StandardScaler’ since all the different predictors have different ranges and different units of measurement. ‘StandardScaler’ will help scale a feature to zero mean and unit variance.

One of the goals of this project was to classify the instances in the dataset based on whether they are at a risk of suicide or not. This is a classification problem. But the target label is missing in this dataset. To make it a Supervised Machine Learning approach, there was a need to create a target column for this dataset. A new target label column ‘Suicide_risk’ was created using the following logic – if Suicides per 100k population of an instance is greater than the mean of the Suicides per 100k population of the dataset, then its value was set to 1, else it was set to 0. Thus, when Suicide_risk= 1, the instance is at a high risk of suicide whereas if Suicide_risk=0, the instance is at a lower risk of suicide. A cutoff value was not used here to make it dynamic, so that the model would work well for new unseen data as well.

Results :

i) Data visualizations:

The next step after performing data cleaning and processing was to answer the research questions.

Figure(1) is a histogram displaying the data distribution for Suicides per 100k population. It shows that the distribution is left skewed. More data exists for suicide per 100k count in the first 2 bins ranging 0-20.

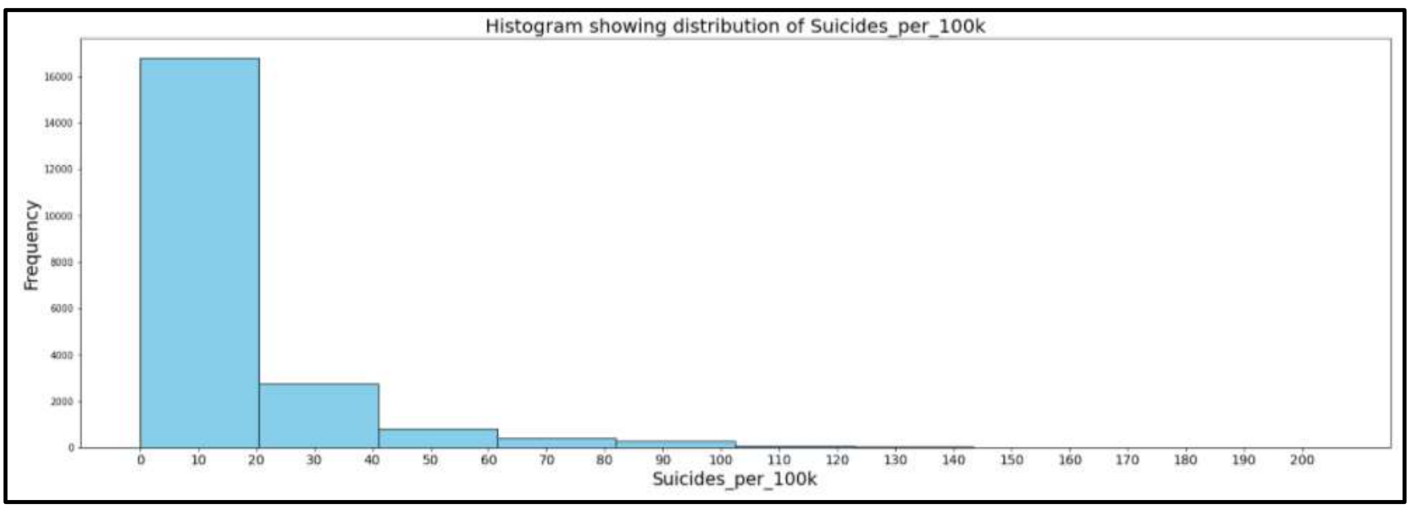

Figure 1

Figure(2) is a line plot showing the trend in the Suicides per 100 persons population over the past 14 years (1990-2014). As seen below, the trend seems to be decreasing in the recent years, whereas suicide rate was highest in 1995.

It is also a fact that Suicide was among the 10 leading causes of death until 1998. It was at number 8. As the years passed by, the data for 2003 showed that suicide was placed at number 11 overall. The below line plot supports this fact.

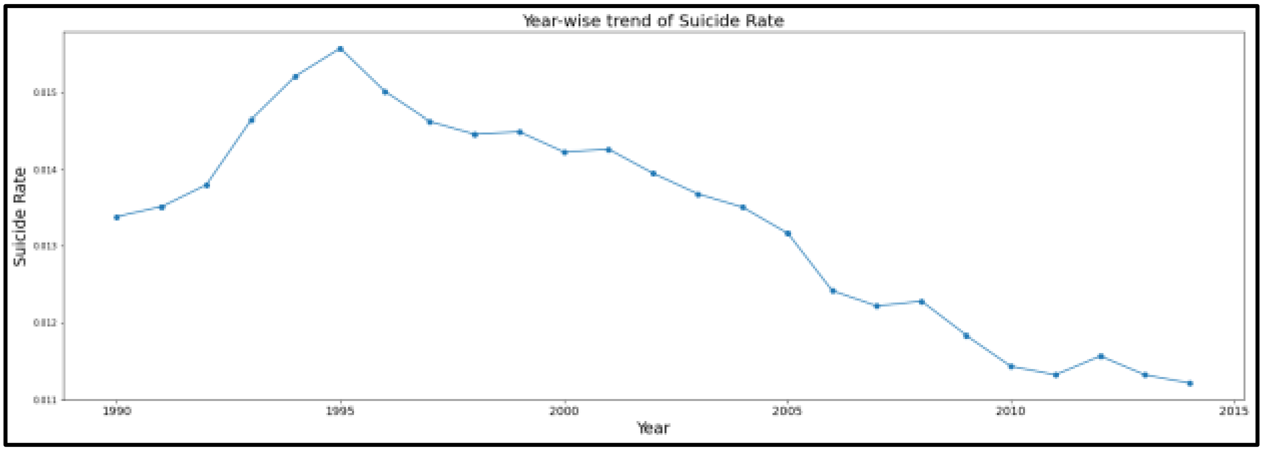

Figure 2

Figure(3) is a stacked bar graph of ‘Count of Number of Suicides’ vs ‘Age group’ for males and females. This graph was plotted to answer the question – which individuals in which age group and gender are at a high risk of suicides?

Looking at the below plot, the historical data shows that highest number of suicide cases have been reported for individuals in the age group 35-54 years and the number of males in this age group is disturbingly high!

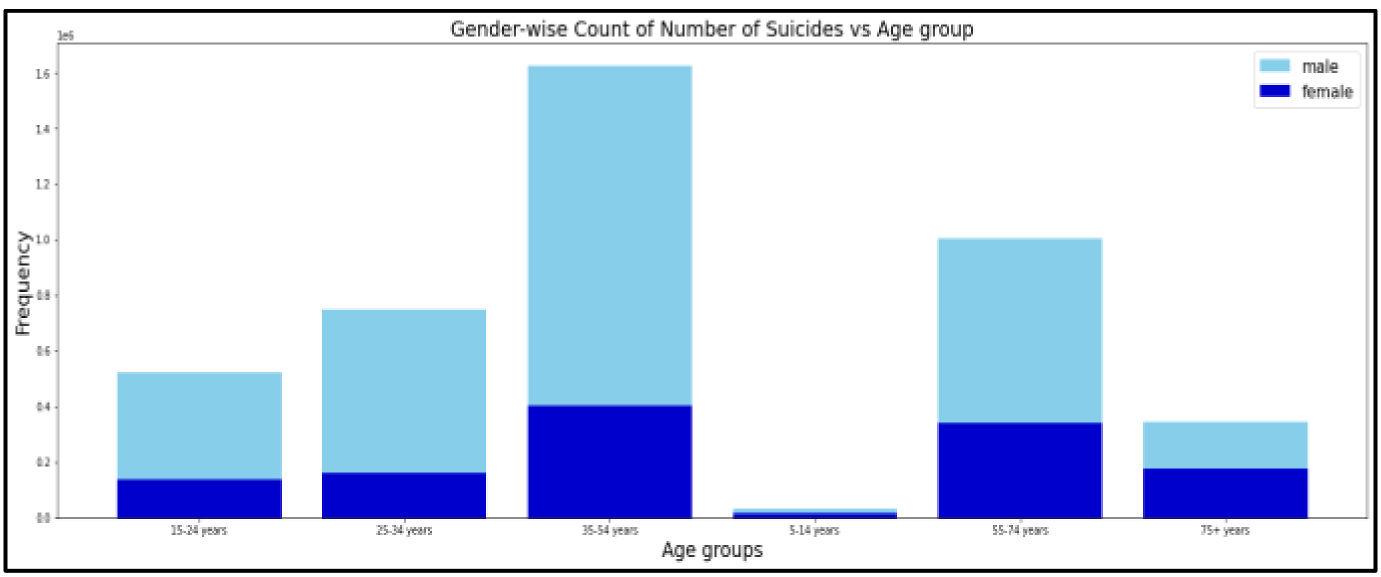

Figure 3

Figure(4) is a pie plot showing the percentage of suicide cases in each age group. 36.86% of the individuals who committed suicides were in the 35-54 year age group!

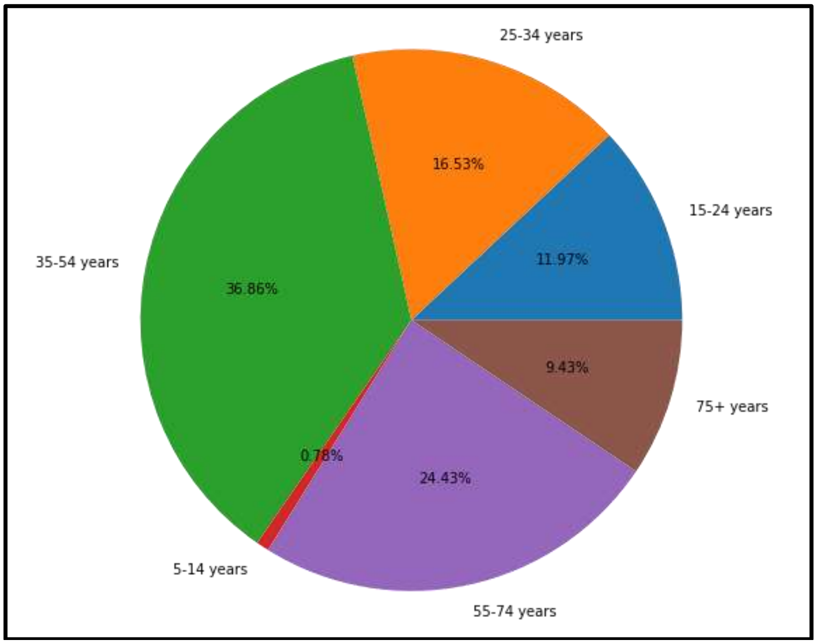

Figure 4

Figure(5) shows the Average Suicides per 100k population of the top 10 countries for which the average was the highest. They are sorted by highest to lowest average suicide rate. Observing this graph it is seen that Lithuania has the highest average of suicides per 100k population.

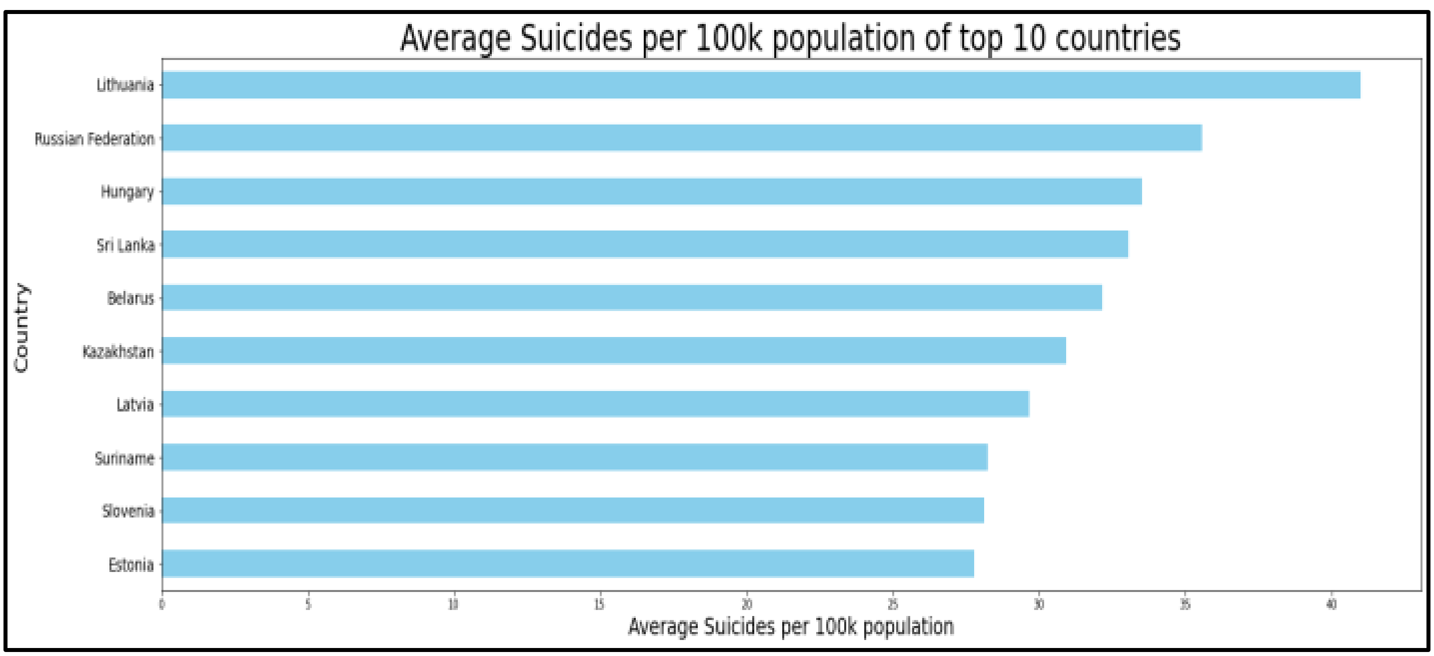

Figure 5

For the top 10 countries which had the highest average Suicides rate, a line plot was plotted to show the trend in the rise/fall of the rate over the years 1990-2014. As seen in the below figure(6), although Lithuania and the Russian Federation were the countries with the highest average suicide rate in 1995, the rate has dropped significantly over the next years. Whereas, the country Suriname had initially a lower average, but the rate has risen tremendously between 2005-2010.

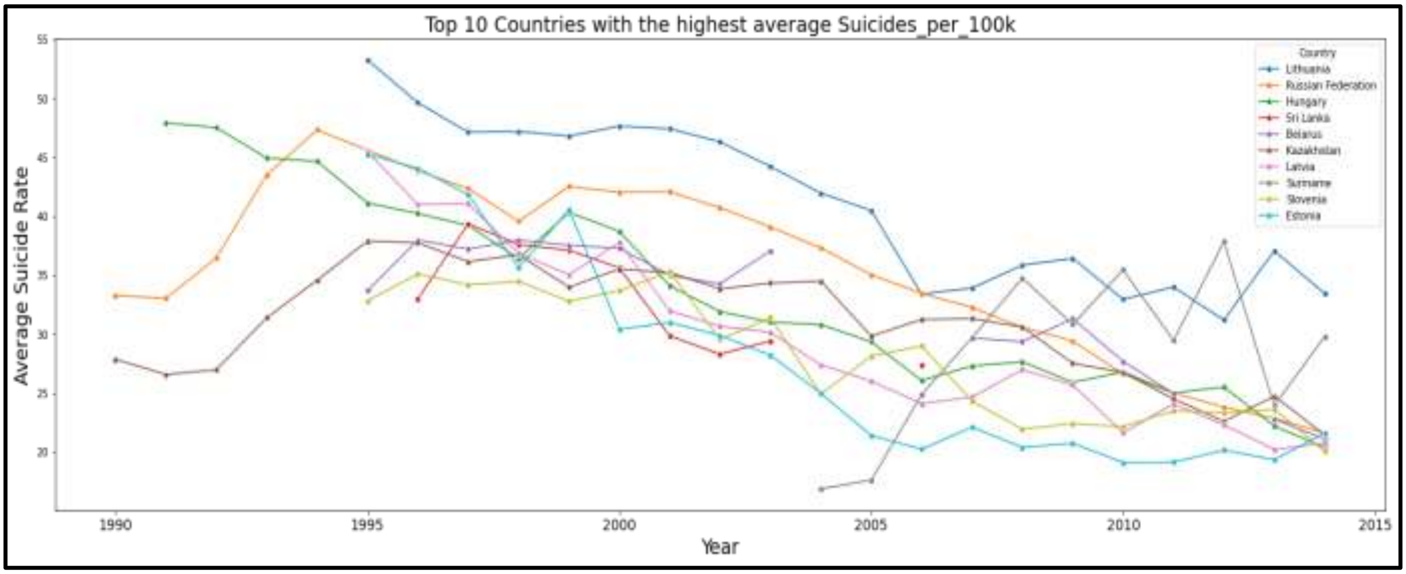

Figure 6

The below figure(7) is a stacked bar chart indicating the Count of Number of Suicides in Year bins for each Generation. The Silent generation are the individuals born between 1927-1946 and their age group is 75+ years. The suicide cases for the Silent generation have risen tremendously after 2000. This is very saddening!

Below is some information about people in each generation -

1990-1995 had most cases for the Boomers Generation.
1995-2000 had most cases for the Generation X.
2000-2005 had most cases for the Silent Generation.
2005-2010 also had most cases for the Silent Generation.
2010-2015 had most cases for the Millennials Generation.

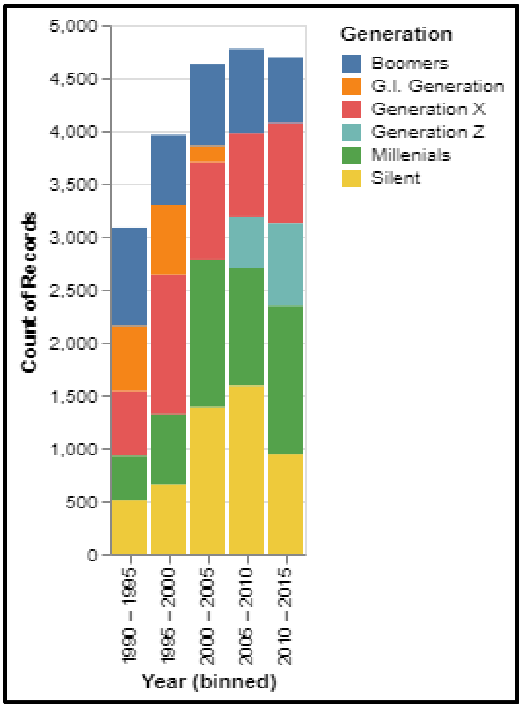

Figure 7

For model building, it was necessary to observe how GDP_per_capita and HDI are related to the Number of Suicides. Hence an interactive scatter plot was plotted using Alteryx library to observe this. A clear relationship is not seen in Figure(8), but it can be said that Number of Suicides are low for very low and very high HDI. When HDI is in the range of 0.7-0.8, a very high count of suicide cases is seen.

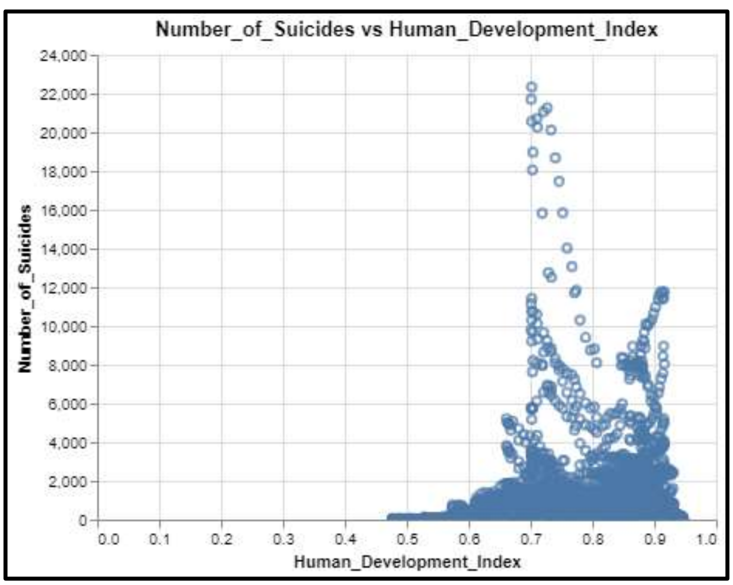

Figure 8

Similarly as seen in Figure(9), for GDP_per_capita as well, it cannot be necessarily said how the number of suicides is affected by increase/decrease in GDP_per_capita.

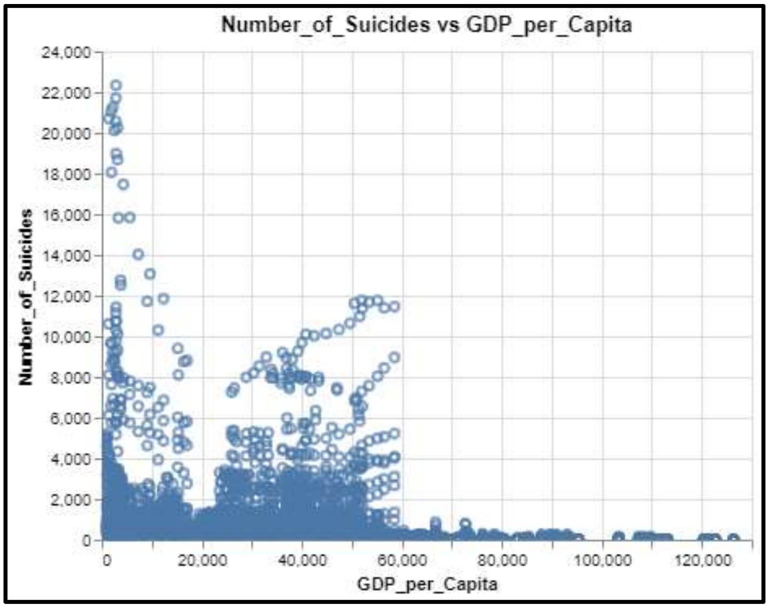

Figure 9

However for the Population and GDP_for_Year column, an increasing trend in the regression line can be seen in figure(10). It is observed that as the Population and the GDP per Year increases, the number of suicides also increase.

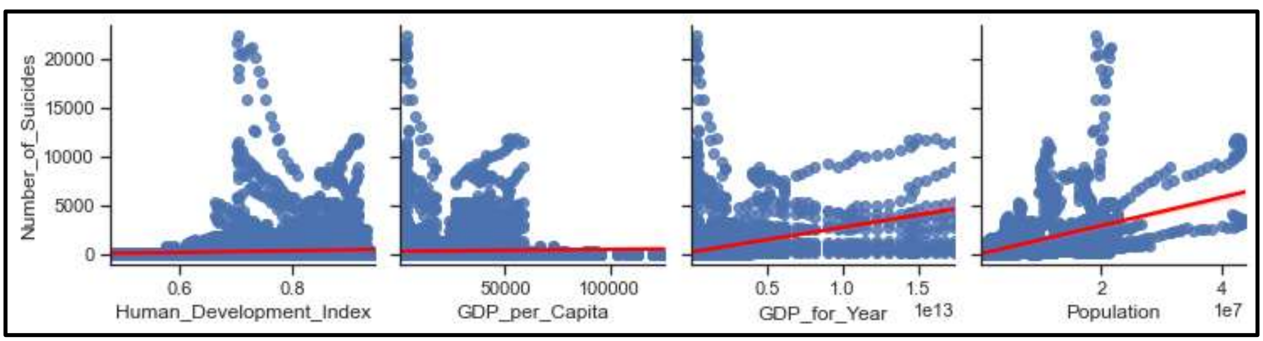

Figure 10

A heatmap has also been demonstrated in the below figure(11) with the purpose of observing the correlations in the data. Population is highly correlated with GDP_for_Year. HDI is highly correlated with GDP_per_capita. This makes sense why both these pairs of predictors were behaving in the same way, i.e. showing the similar trend when plotted against the target variable - Number_of_Suicides in above figure(10).

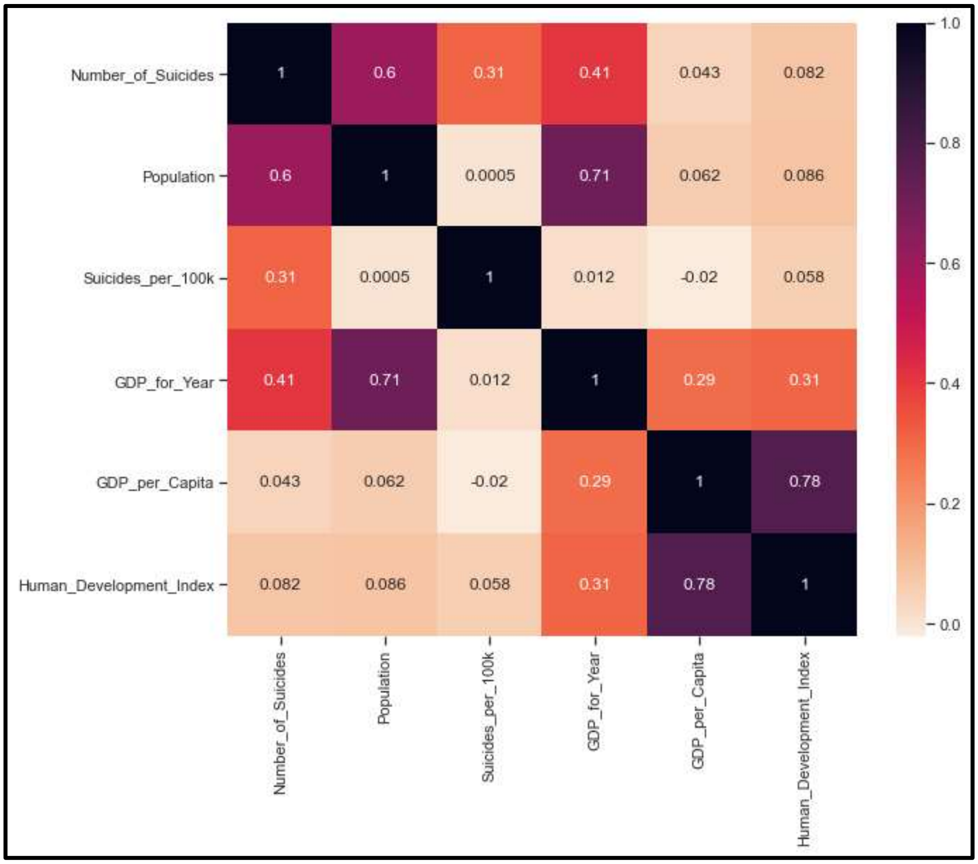

Figure 11

ii) Model validations -

In this section of the report, I have shown how each of the goals of this project has been accomplished.

Goal 1: Predict the number of suicides based on the input predictors to the model

Linear regression was used to predict the number of suicides based on the input predictors to the model. For performing this task, it was necessary to find which predictors in the model were the most useful in predicting the Number_of_Suicides, which is the target variable for this regression model. Four different approaches were attempted and the model which gave the highest R square adjusted value was further used for model building.

Initially, a full model was fit i.e. with all the predictors, it gave an R square adjusted value of 54.71% as seen in figure(12).

```
# Choosing all predictors

model_full = 'Number_of_Suicides ~ Human_Development_Index + GDP_per_Capita + \
GDP_for_Year + Country + Sex + Age + Generation + Population + Year'

result_full = smf.ols(formula = model_full, data = new_df).fit()

print('MSE:',result_full.mse_resid)
print('R2:',result_full.rsquared)
print('R2_adj:',result_full.rsquared_adj)
```

```
MSE: 426189.89259247744
R2: 0.5497678011305825
R2_adj: 0.5471978604054563
```

Figure 12

The second model that was fit, used interactions as seen in the below figure(13). This model was built since dependencies and correlations were found between some predictors in the heatmap in figure(11). However, this resulted in the R^2 adjusted value to drop down to 36.5% and an increase in the MSE.

```
# Building a model with interactions (observed from heat maps)

model_interactions = 'Number_of_Suicides ~ Human_Development_Index + GDP_per_Capita + \
GDP_for_Year + Country + Sex + Age + Generation + Population + Year + Human_Development_Index:GDP_per_Capita + \
Population:GDP_for_Year'

result_interactions = smf.ols(formula = model_interactions, data = new_df).fit()

print('MSE:',result_interactions.mse_resid)
print('R2:',result_interactions.rsquared)
print('R2_adj:',result_interactions.rsquared_adj)
```

```
MSE: 597018.3163942941
R2: 0.3657625296989724
R2_adj: 0.3657025289922602
```

Figure 13

In the third approach, a model was fit by only selecting the significant predictors, i.e. the predictors with p-values< 0.05. The R^2 adjusted value for this model was 52.18%.

```
# Significant predictors only

model_significant = 'Number_of_Suicides ~ Human_Development_Index + GDP_for_Year + Country + Population + Year + Age + Country'

result_significant = smf.ols(formula = model_significant, data = new_df).fit()

print('MSE:',result_significant.mse_resid)
print('R2:',result_significant.rsquared)
print('R2_adj:',result_significant.rsquared_adj)
```

```
MSE: 450061.51304550964
R2: 0.5244138358547412
R2_adj: 0.5218356427541393
```

Figure 14

In the last approach, log transformation of the predictors – and was used to check if the R^2 adjusted value increases. However, this did not improve the R^2 value and it remained 54.7%.

```
model_log = 'Number_of_Suicides ~ np.log(Human_Development_Index) + GDP_per_Capita + np.log(GDP_for_Year) + \
Country + Sex + Age + Generation + Population + Year'

result_log = smf.ols(formula = model_log, data = new_df).fit()

print('MSE:',result_log.mse_resid)
print('R2:',result_log.rsquared)
print('R2_adj:',result_log.rsquared_adj)
```

```
MSE: 426456.6571503161
R2: 0.5496788539985373
R2_adj: 0.5469144384739257
```

Figure 15

Thus, observing the MSE and R^2 adjusted values for the above four models, it was decided to use all the predictors for the model building process as the full model has the highest R square and lowest MSE.

The goal of the Linear Regression model is to predict the number of suicides for the instances in the dataset, so that when new instances come in, the Linear model will be able to perform well. The below figure(16) shows the model predictions for the number of suicides for 3 instances in the dataset. According to the Linear regression model, in the year 2014 in Greece when the GDP per Capita was $22,834 and the Human Development Index was 0.866 for females in the age group 25-34 years, the total number of suicides predicted for this instance was 73.

```
X_new = X_suicide.sample(3, random_state = 1)
X_new
```

	Human_Development_Index	GDP_per_Capita	GDP_for_Year	Country	Sex	Age	Generation	Population	Year
6030	0.905	55354.0	2.828849e+11	Denmark	male	15-24 years	Milenials	312053.0	2006
11471	0.725	2232.0	3.083370e+10	Kazakhstan	female	5-14 years	Milenials	1316237.0	2003
8540	0.866	22834.0	2.370296e+11	Greece	female	25-34 years	Milenials	689746.0	2014

```
pipe.predict(X_new).round()
```

```
array([ 15., 173.,  73.])
```

Figure 16

Goal 2: Classifying the instances in the dataset based on whether they are at a risk of suicide or not.

To attain this goal, a classification model needs to be built. The dataset was imbalanced, meaning there were 14,534 instances with Suicide_risk = 1 and only 6,610 instances with Suicide_risk = 0. Hence, Stratified sampling was used since the dataset was imbalanced. The data was split into train and test data as shown below –

Training data - 80%
Test data – 20%
Random state – 77

Three classification models were built to classify the instances in the dataset based on whether they are at a risk of suicide or not.

1) **Logistic Regression :**
Logistic regression is the most frequently used model used for classification. This model was built using the predictors - 'Human_Development_Index', 'GDP_per_Capita' , 'GDP_for_Year' , 'Country' , 'Sex' , 'Year', 'Age', 'Generation' and 'Population'.
The target predictor was ‘Suicide_risk’.

Below is the Confusion matrix obtained for this model –

```
confusion matrix
[[2459  448]
 [ 584  738]]
```

Figure 17

Accuracy, recall, precision, f1-score for this model is shown below –

```
classification report for Logistic Regression
              precision    recall  f1-score   support

           0       0.81      0.85      0.83      2907
           1       0.62      0.56      0.59      1322

    accuracy                           0.76      4229
   macro avg       0.72      0.70      0.71      4229
weighted avg       0.75      0.76      0.75      4229
```

Figure 18

A Stratified 5-fold cross validation model was also built for the Logistic regression model. Below is the ROC_AUC score obtained –

```
cv = StratifiedKFold(n_splits = 5, shuffle=True, random_state=99)

print('ROC AUC scores for 5 fold Logistic regression model')
cross_val_score(LR, X_scaled, y_log_sui, cv=cv, scoring='roc_auc')

ROC AUC scores for 5 fold Logistic regression model

array([0.80792255, 0.81539031, 0.81722869, 0.81431148, 0.81337975])
```

Figure 19

2) K Nearest Neighbors (KNN) :

Since KNN is also used for classification, a K-nearest neighbor model has also been built for classifying the instances into prone to suicide risk or not. Several values of k were attempted between 1 to 19, but k=3 gave the best results. Below were the results obtained for the Confusion matrix and accuracy, precision, recall and f1-score for the KNN model.

```
confusion matrix
[[2738  169]
 [ 133 1189]]
```

Figure 20

```
knn = KNeighborsClassifier(n_neighbors=3)

knn.fit(X_train, y_train)

y_pred = knn.predict(X_test)

print('Accuracy of KNN:',metrics.accuracy_score(y_test, y_pred)*100,'%')
print('Precision_recall_fscore_support of KNN', precision_recall_fscore_support(y_test, y_pred, average='weighted'))

Accuracy of KNN: 92.85883187514779 %
Precision_recall_fscore_support of KNN (0.9292533474158379, 0.9285883187514778, 0.9288454831497063, None)
```

Figure 21

Finally, a Stratified 5-fold cross validation model was also built for the KNN model. Below is the ROC_AUC score obtained –

```
print('ROC AUC scores for 5 fold KNN model')
cross_val_score(knn, X_scaled, y_log_sui, cv = 5, scoring='roc_auc')

ROC AUC scores for 5 fold KNN model

array([0.8081868 , 0.74840101, 0.74149869, 0.74844785, 0.78322043])
```

Figure 22

It can be clearly seen that KNN performs way better than the Logistic regression model.

3) Support Vector Machines (SVM) :

Lastly, I also tried using a new machine learning algorithm – Support Vector Machine (SVM) for classification. I used this model because this model I was curious to see how this model works on my data. Below were the results obtained for SVM –

```
from sklearn.svm import SVC

svm = SVC().fit(X_train, y_train)
svc_predictions = svm.predict(X_test)

svc_accuracy = svm.score(X_test, y_test)
print('SVM Accuracy:', svc_accuracy)

SVM Accuracy: 0.8576495625443368

svc_recall = recall_score(y_test, svc_predictions)
print('Recall of SVM model:',svc_recall)
svc_precision = precision_score(y_test, svc_predictions)
print('Precision of SVM model:',svc_precision)

Recall of SVM model: 0.716338880484115
Precision of SVM model: 0.8066439522998297
```

Figure 23

In conclusion, among all three classification models, the K nearest neighbor (KNN) classification model with k=3, works really well on this data. It gave the highest accuracy, precision and recall in classifying the instances in the test dataset into high suicide risk and low suicide risk categories.

Goal 3: Clustering the instances in the dataset based on their economic condition

An unsupervised machine learning approach was also tried on the Suicide dataset – K means clustering. Each instance in the dataset is a group of male/female individuals belonging to a certain country and age group in a certain year. For every country, there is a yearly GDP and HDI values for each instance. Clusters are created based on the Human development index (HDI), GDP per capita and Suicides per 100k population. GDP per capita is a country's economic output per person. It is calculated by dividing the GDP of a country by its population. HDI is the geometric mean of Life expectancy, education and Gross National Income per capita (GNI).

Firstly, a scatter plot of Suicides_per_100k vs GDP_per_capita was created using 2 clusters as seen in the below figure(24).

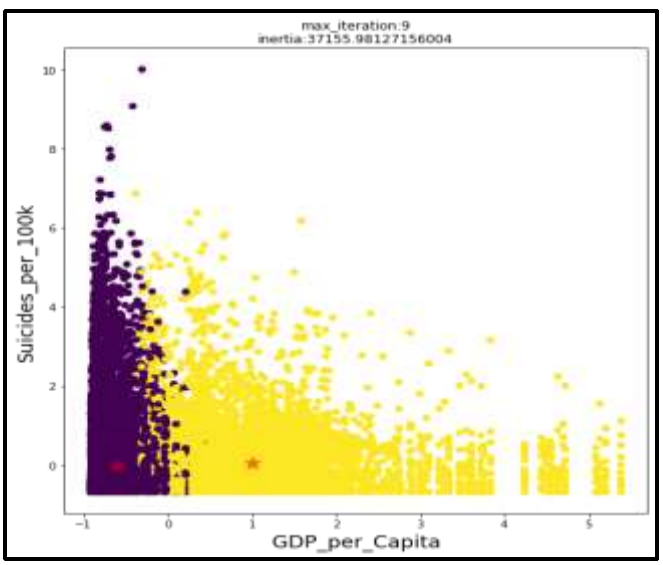

Figure 24

Silhouette analysis was done to select an appropriate value of k for the k means clustering algorithm.

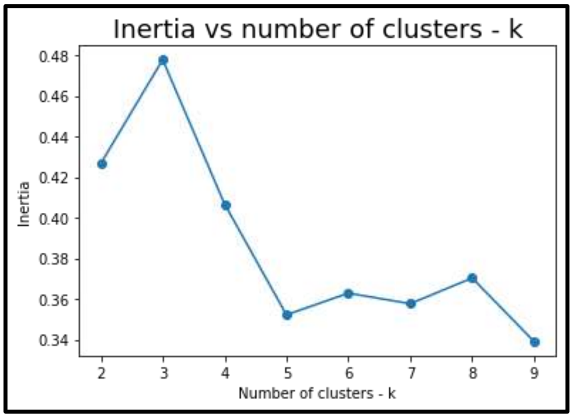

Figure 25

Looking at the results, k=3 seemed like a good choice for k. Hence, using k=3, a k means clustering model was built. In the below figure(26), Suicides_per_100k vs GDP_per_capita and Suicides_per_100k vs Human Development Index has been plotted for k=3.

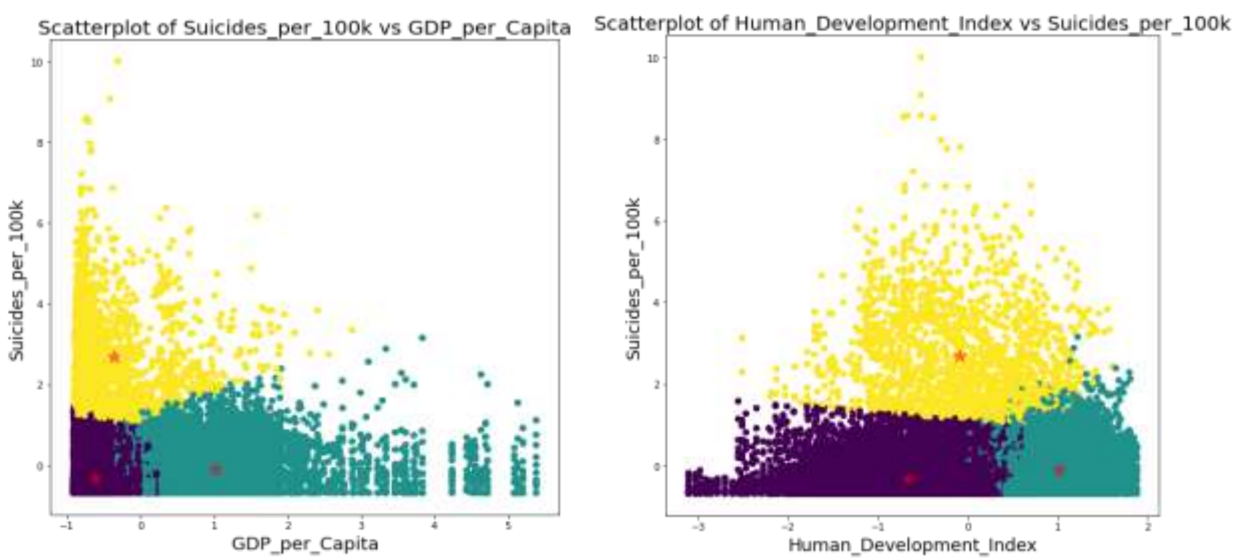

Figure 26

The first cluster (seen in the above image in purple) has the lowest Gross Domestic Product (GDP) per capita, lowest Human development index (HDI) and has the lowest suicide rates.

The second cluster (seen in the above image in yellow) has average Gross Domestic Product (GDP) per capita, average Human development index (HDI) and has the highest suicide rates among the three clusters.

The third cluster (seen in the above image in green) has the highest Gross Domestic Product (GDP) per capita, highest Human development index (HDI) and has low suicide rates.

Looking at the GDP and HDI values and assuming that under-developed countries would have the lowest GDP and HDI, developed countries would have the highest GDP and HDI and developing countries would have average GDP and HDI, it could be supposed that - the first cluster (purple) contains instances of under-developed countries, the second cluster (yellow) contains instances mostly from developing countries, whereas the third cluster (green) contains instances from the developed countries.

Looking at three clusters in the above figure, it can be seen that the second cluster (yellow) has the highest suicide rate. From some literature review (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1414701/#:~:text=There%20are%20marked%20differences%20in,a%20higher%20risk%20of%20suicide) and research that was performed before starting this project, it was witnessed that developing countries are at the highest risk of suicides.

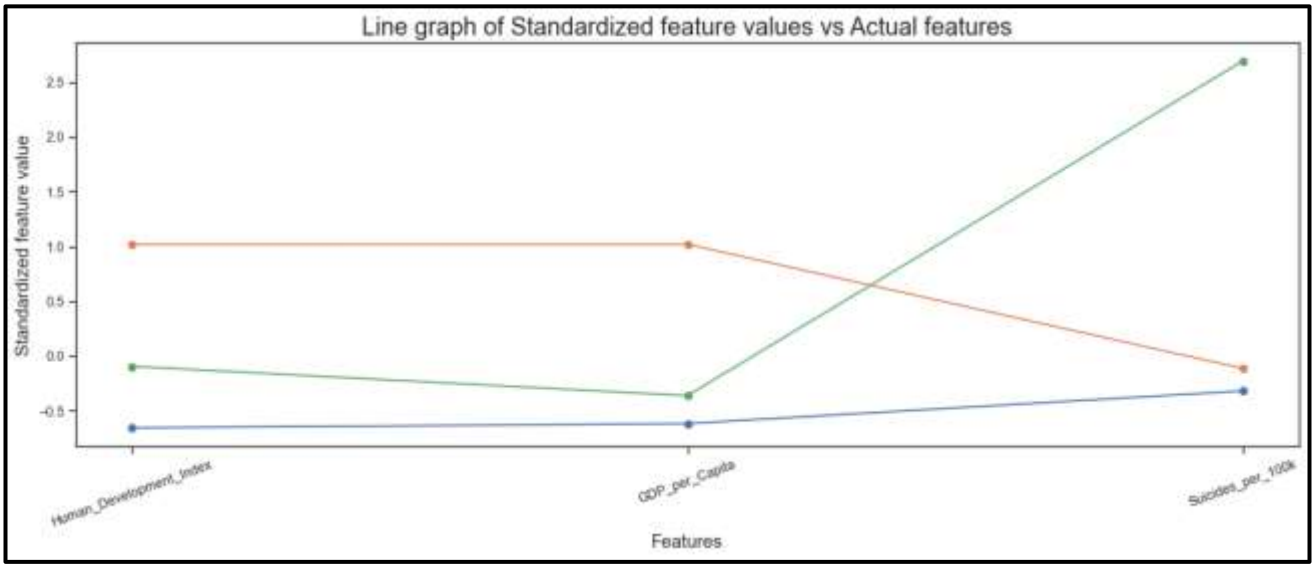

Figure 27

	Human_Development_Index	GDP_per_Capita	Suicides_per_100k
Cluster 1- Underdeveloped countries	0.707390	6199.892927	7.084011
Cluster 3- Developed countries	0.864645	38943.499027	11.017530
Cluster 2- Developing countries	0.760105	11302.477855	64.702745

Figure 28

The above line graph in figure(27) and table in figure(28) gives a summary of this discussion and shows that the developed (Cluster 3) and underdeveloped countries (Cluster 1) have significantly low suicide rate as compared to the developed countries (Cluster 2).

The DataFrame in figure(28) is sorted in increasing order of Suicides per 100k. All the instances that fall in Cluster 2 – Developing countries, are the ones that need the most attention as the rate of Suicides per 100k population is the highest for them.

Goal 4 : Forecasting the total number of suicides for the future years by training the model on the historical data.

For performing this task, a subset of the dataset was created using 2 predictors in the dataset – Year and Total Number of Suicides (group by year). The first few records of the input to the forecasting model is the DataFrame shown in the below figure(29).

Year	Number_of_Suicides
1990-01-01	183645.0
1991-01-01	188476.0
1992-01-01	170714.0
1993-01-01	210549.0
1994-01-01	220598.0
1995-01-01	234729.0
1996-01-01	236779.0
1997-01-01	200151.0

Figure 29

The 'Year' in the DataFrame has values from 1990 to 2014. To perform the task of forecasting, this DataFrame was divided into 2 parts – training and validation. Auto_arima has been used for fitting the model with the training data. Training data contains historical data, which is total count of number of suicides from the Year 1990 – 2011. The validation dataset contains data from 2011-2014. Forecasting has been done on the validation data.

Below is the time series plot showing the forecasted number of suicides for the validation data in orange.

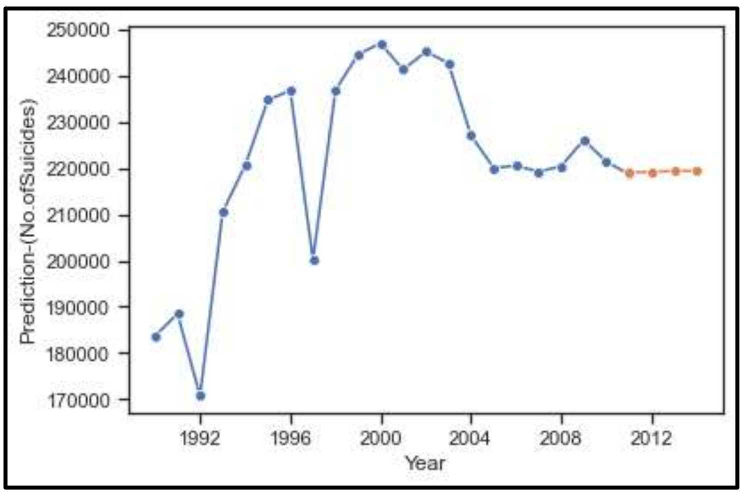

Figure 30

However, in the real world, there will be unseen data that needs to be forecasted. Since the DataFrame contains data from the years 1990-2014, the years 2015 and 2016 are unseen data/test data for this model. Hence, another model was built to forecast the number of suicides for the years 2015 and 2016. (Note- as of 2020 this data is in the past, but for my dataset and model this is new/unseen data).

This DataFrame containing the forecasted count of the number of suicides is shown in the below figure(31). The model forecasts 219037 suicides for the year 2015 and 219232 suicides for the year 2016 worldwide.

Forecasting on new (unseen data) : Test data

```
# Creating the dataset for next years
X_new = pd.DataFrame(columns=validation_forecast.columns,
                     index=['2015', '2016'])
X_new.index = pd.to_datetime(X_new.index, format = '%Y',errors = 'coerce')
```

```
forecast_X_new = model.predict(n_periods=len(X_new))
forecast_X_new.round()
```

```
array([219037., 219232.])
```

```
# Prediction of Number of Suicides made for next 2 years
test = pd.DataFrame(forecast_X_new.round(),index = X_new.index,columns=['Prediction-(No.ofSuicides)'])
test
```

	Prediction-(No.ofSuicides)
2015-01-01	219037.0
2016-01-01	219232.0

Figure 31

Discussions:

Several interesting statistics were observed in this data. Using data visualizations, numerous important questions were answered. From the historical data it was observed that males in the age group of 35-54 years are at the highest risk. The suicide cases among the 75+ year old population or the Silent generation have also risen tremendously after the year 2000.

Performing clustering showed an interesting finding – the fact that developing countries are facing the highest threat of suicides. Countries like Lithuania, Russian Federation, Hungary, Sri Lanka were among the top 10 countries which had the highest average number of suicides per 100k population and all these countries actually do fall in the list of developing countries according to the World Population Review website - https://worldpopulationreview.com/country-rankings/developing-countries

Coming to the study limitations, if this dataset had more predictors, it would lead to better results. This dataset mainly focuses on the economic impact on the suicide rates. However there are other factors too due to which the mental health of an individual is affected such as – depression, mental health disorders, family history, etc. If such a dataset is combined with the economic factors, the model could give more accurate predictions.

A lot of efforts were put into finding these datasets. Similar to the HDI dataset, there were more such datasets on the WHO website, however the data in these datasets was collected for every 5 years which would result in missing data problem due to insufficient information. Too many imputations are also not good for the model which is why these datasets were not included in this project.

Conclusion:

Baylor College of Medicine's Dr. John Oldham says that 'Suicide has been a major concern for the psychiatric mental health field and in the broad field of medicine for decades. We need to know a lot more about unplanned, unexpected suicides, as they may be helpful in improving prevention strategies. The issue here is that not enough initiative is taken to prevent possible suicides. This is what we can improve right now.' The several models built in this project would be very useful in identifying individuals/groups of individuals who are at a high risk of committing suicides. Due to the forecasting model, future suicide numbers could be predicted and preventive measures could be taken well in advance.