Suicide Analysis, Visualization and Predictive Modelling

A Project Report

Submitted by

Sanat Madkar B050

Tanay Maheshwari B053

Mann Merani B060

Rahil Merchant B061

Under the Guidance of

Prof. Ameyaa Biwalkar

in partial fulfilment for the award of the degree of

Bachelors of Technology

Computer Engineering

At

SVKM'S NMIMS Deemed-to-be UNIVERSITY

Mukesh Patel School of Technology Management and Engineering, Mumbai

March, 2020

DECLARATION

We, **Sanat Madkar, Tanay Maheshwari, Mann Merani, Rahil Merchant** Roll Nos. **B050, B053, B060, B061** of B.Tech. (Computer Engineering), VI semester understand that plagiarism is defined as anyone or combination of the following:

1. Un-credited verbatim copying of individual sentences, paragraphs or illustration (such as graphs, diagrams, etc.) from any source, published or unpublished, including the internet.
2. Un-credited improper paraphrasing of pages paragraphs (changing a few words phrases, or rearranging the original sentence order)
3. Credited verbatim copying of a major portion of a paper (or thesis chapter) without clear delineation of who did wrote what. (Source: IEEE, The institute, Dec. 2004)
4. I have made sure that all the ideas, expressions, graphs, diagrams, etc., that are not a result of my work, are properly credited. Long phrases or sentences that had to be used verbatim from published literature have been clearly identified using quotation marks.
5. I affirm that no portion of my work can be considered as plagiarism and I take full responsibility if such a complaint occurs. I understand fully well that the guide of the seminar/ project report may not be in a position to check for the possibility of such incidences of plagiarism in this body of work.

Signature of the Students: __________, ____________, ___________

Names: Sanat Madkar, Tanay Maheshwari, Mann Merani, Rahil Merchant

Roll Nos.: B050, B053, B060, B061

Place: Mumbai

Date:

CERTIFICATE

This is to certify that the project entitled "**Suicide Analysis, Visualization and Predictive Modelling**" is the bonafide work carried out by **Sanat Madkar, Tanay Maheshwari, Mann Merani and Rahil Merchant** of B.Tech. MPSTME (NMIMS), Mumbai, during the VI semester of the academic year 2019-2020, in partial fulfillment of the requirements for the Course Programming Language.

____

Prof. Ameyaa Biwalkar

Internal Mentor

________________________ ________________________

Examiner 1 Examiner 2

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 INTRODUCTION

We have seen an alarming rise in suicide rates across the globe over the past few years with suicide rates in 2017 being the highest they've been since World War II. This is one among many such alarming statistics that only enunciate the importance of having good mental health.

Being the 10th most common cause of death, suicide prevention is a matter of utmost importance. This requires a clear understanding of the reasons that could lead an individual to take such drastic steps. For this to be done, previous data can be visualized in order to interpret it in a better manner.

Predictive modelling can go a long way towards trying to get a better picture of the dependency of the number of suicides on various factors. It can be used to predict the number of suicides for any year based on values of correlative attributes.

Thus, via this project we aim to pin-point the major factors that influence suicides and try to increase awareness through the creation of various visualizations.

1.2 PROBLEM STATEMENT

To determine the factors that influence suicide rates the most. We analyse datasets containing information from 1985 to 2013 on various parameters that could possibly have an effect on suicide rates. We intend to create visualizations that better illustrate the large amounts of data and conduct exploratory data analysis. We also try to correlate different factors with the dependent variable (number of suicides) and use these correlative variables to create predictive models. We then evaluate these models to identify the most effective ones. These models can then be used to predict the number of suicides in a year based on the values of correlative variables.

CHAPTER 2: SOFTWARE AND APIs USED

Software: Python 3.7.3, Jupyter Notebook, Anaconda, Kaggle

2.1 Jupyter Notebook:

- The notebook extends the console-based approach to interactive computing in a qualitatively new direction, providing a web-based application suitable for capturing the whole computation process: developing, documenting, and executing code, as well as communicating the results. The Jupyter notebook combines two components:
  - A web application: a browser-based tool for interactive authoring of documents which combine explanatory text, mathematics, computations and their rich media output.
  - Notebook documents: a representation of all content visible in the web application, including inputs and outputs of the computations, explanatory text, mathematics, images, and rich media representations of objects.
- Main features of the web application:
  - In-browser editing for code, with automatic syntax highlighting, indentation, and tab completion/introspection.
  - The ability to execute code from the browser, with the results of computations attached to the code which generated them.
  - Displaying the result of computation using rich media representations, such as HTML, LaTeX, PNG, SVG, etc. For example, publication-quality figures rendered by the matplotlib library, can be included inline.
  - In-browser editing for rich text using the Markdown markup language, which can provide commentary for the code, is not limited to plain text.
- Notebook Documents:
  - Notebook documents contains the inputs and outputs of a interactive session as well as additional text that accompanies the code but is not meant for execution. In this way, notebook files can serve as a complete computational record of a session, interleaving executable code with explanatory text, mathematics, and rich representations of resulting objects. These documents are internally JSON files and are saved with the .ipynb extension. Since JSON is a plain text format, they can be version-controlled and shared with colleagues.

- Notebooks may be exported to a range of static formats, including HTML (for example, for blog posts), reStructuredText, LaTeX, PDF, and slide shows, via the nbconvert command.
- Furthermore, any .ipynb notebook document available from a public URL can be shared via the Jupyter Notebook Viewer (nbviewer). This service loads the notebook document from the URL and renders it as a static web page. The results may thus be shared with a colleague, or as a public blog post, without other users needing to install the Jupyter notebook themselves

2.2 Libraries Used:

- **Pandas**: Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool built on top of the Python programming language. When working with tabular data, such as data stored in spreadsheets or databases, Pandas will help you to explore, clean and process your data. In Pandas, a data table is called a DataFrame.
- **Matplotlib**: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.
- **Seaborn**: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
- **Plotly**: Plotly is another interactive visualization library.
- **NumPy**: NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- **Scikit-Learn**: Scikit-Learn is a free machine learning and data science library. This was used by us for implementing regression algorithms, model selection and metrics such as RMSE and R^2 score.
- **Scipy**: Scipy is another machine learning library, we used it to calculate Pearson Coefficients and p-values to recognize linear correlation of variables with the number of suicides per year.

2.3 Datasets Used:

We obtained the following datasets from Kaggle and the links mentioned below:

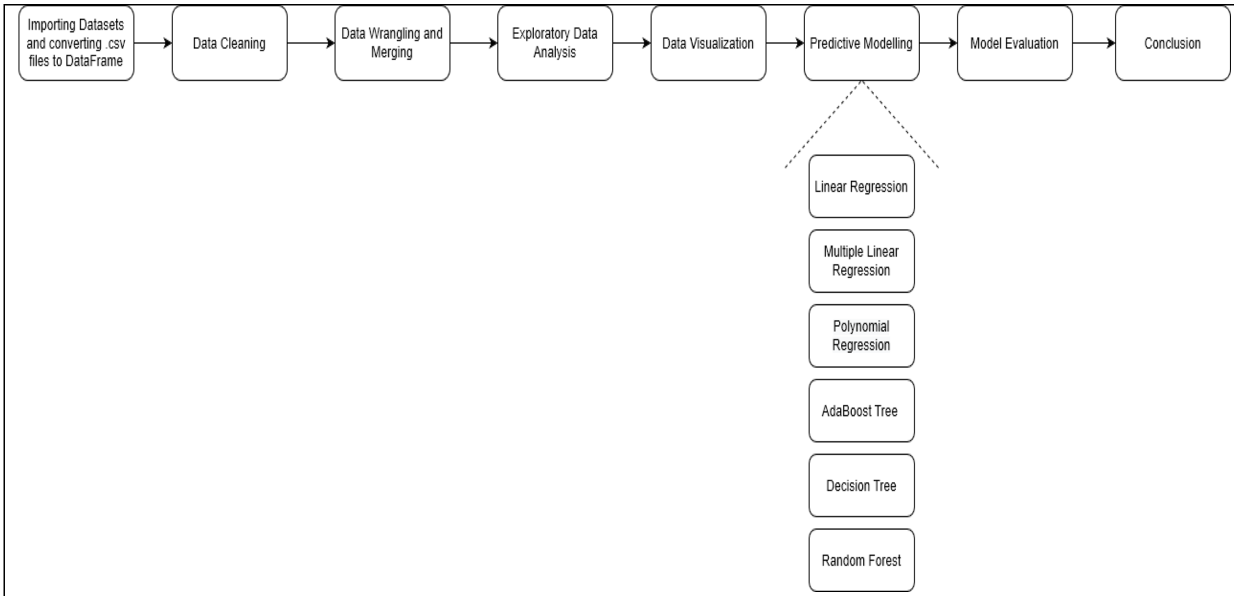
- https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016
- https://www.kaggle.com/dornani/widandsuicide
- Above datasets refer to the following datasets:
    - United Nations Development Program. (2018). Human development index (HDI). Retrieved from http://hdr.undp.org/en/indicators/137506
    - World Bank. (2018). World development indicators: GDP (current US$) by country:1985 to 2016. Retrieved from http://databank.worldbank.org/data/source/world-development-indicators#
    - [Szamil]. (2017). Suicide in the Twenty-First Century [dataset]. Retrieved from https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook
    - World Health Organization. (2018). Suicide prevention. Retrieved from http://www.who.int/mental_health/suicide-prevention/en/
    - The World Bank World Development Indicators. Retrieved from https://datacatalog.worldbank.org/dataset/world-development-indicators

Following were the columns present in the datasets:

```
suicides_no
population
yearlyHDI
GDPpyear
GDPpcapital
Individuals using the Internet (% of population)
Expense (% of GDP)
Compensation of employees (% of expense)
Unemployment, total (% of total labor force) (modeled ILO estimate)
Physicians (per 1,000 people)
Labor force, total
Life expectancy at birth, total (years)
Mobile cellular subscriptions (per 100 people)
Refugee population by country or territory of origin
Contributing family workers, total (% of total employment) (modeled ILO estimate)
Access to electricity (% of population)
Lower secondary completion rate, total (% of relevant age group)
```

CHAPTER 3: METHODS IMPLEMENTED

Flow Diagram:



- **Importing Datasets**: We imported the datasets using the read_csv("*file_path*") method.
- **Data Cleaning:** Operations such as renaming of columns, dropping columns with data that depended on other columns (e.g.: Generation depending on the Age Group column), changing data types of certain columns, dropping of records with incorrect/null values was done.
- **Data Wrangling:** The original dataset was grouped by the following three columns together: Country, Year, Sex and Age Group. In order to conduct proper analysis, we had to convert multiple rows of the same year to one row per year. We then set the index as the year.

	country	year	sex	age	suicides_no	population	suicidesper100k	country-year	yearlyHDI	GDPpyear	...	Unemployment, total (% of total labor force) (modeled ILO estimate)	Physicians (per 1,000 people)	Strength of legal rights index (0=weak to 12=strong)
0	Argentina	1985	male	75+ years	202	363000	55.65	Argentina1985	0.694	8.841667e+10	...	0.0	0.0	0.0
1	Argentina	1985	male	55-74 years	485	1997000	24.29	Argentina1985	0.694	8.841667e+10	...	0.0	0.0	0.0
2	Argentina	1985	male	35-54 years	414	3346300	12.37	Argentina1985	0.694	8.841667e+10	...	0.0	0.0	0.0
3	Argentina	1985	female	55-74 years	210	2304000	9.11	Argentina1985	0.694	8.841667e+10	...	0.0	0.0	0.0
4	Argentina	1985	male	25-34 years	177	2234200	7.92	Argentina1985	0.694	8.841667e+10	...	0.0	0.0	0.0

Above is the original format of the dataset. We converted it to the below format:

year	index	yearlyHDI	GDPpyear	GDPpcapital	Individuals using the Internet (% of population)	Expense (% of GDP)	Compensation of employees (% of expense)	Unemployment, total (% of total labor force) (modeled ILO estimate)	Physicians (per 1,000 people)	Strength of legal rights index (0=weak to 12=strong)	Labor force, total	Life expectancy at birth, total (years)	subsc (
1995	6340	0.772	1.368780e+11	13550	0.749616	44.767234	20.855936	9.062000	3.9000	NaN	4530131	77.585366	2
1996	6352	NaN	1.458620e+11	14330	1.395361	44.169723	19.610644	9.655000	3.9000	NaN	4627651	77.685366	4
1997	6364	NaN	1.431580e+11	13968	1.849635	42.814109	22.234439	9.577000	4.0000	NaN	4654407	78.136585	8
1998	6376	NaN	1.444280e+11	14008	3.221822	43.197267	21.824062	10.839000	4.1000	NaN	4787045	77.839024	18
1999	6388	NaN	1.425410e+11	13756	6.877292	42.304012	22.302908	11.853000	4.2000	NaN	4862414	77.987805	35
2000	6400	0.799	1.301340e+11	12509	9.138837	43.343845	22.297540	11.248000	4.3000	NaN	4896618	77.887805	53

In order to have a more specific analysis, we decided to focus on suicide rates of Greece only. Greece was picked as it suffered from a collapse in economy in the late 2000s and early 2010s. We intended to analyse the impact of the falling GDP on the number of suicides.

Certain missing values were filled by the mean of other values in that column.

For certain values that rose over time (E.g.: Number of Physicians per 1000 people), we carried out linear interpolation to fill the missing values.

- **Data Analysis:**

  Once the data had been prepared, we conducted some analysis to better understand the data. This involved plotting of data (as shown in screenshots below) and the generation of a correlation matrix as shown below:

	suicides_no	population	yearlyHDI	GDPpyear	GDPpcapital	Individuals using the Internet (% of population)	Expense (% of GDP)	Compensation of employees (% of expense)	Unemployment, total (% of total labor force) (modeled ILO estimate)	Physicians (per 1,000 people)	Labor force, total
suicides_no	1.000000	0.137307	0.404149	0.223344	0.229987	0.655167	0.831506	-0.625692	0.903888	0.518174	0.261296
population	0.137307	1.000000	0.456981	0.192652	0.186314	0.327924	0.065879	0.309711	0.204029	0.291057	0.502344
yearlyHDI	0.404149	0.456981	1.000000	0.896257	0.893429	0.923693	0.604293	0.059125	0.368092	0.951200	0.937692
GDPpyear	0.223344	0.192652	0.896257	1.000000	0.999749	0.806295	0.495391	0.059399	0.098184	0.901747	0.831639
GDPpcapital	0.229987	0.186314	0.893429	0.999749	1.000000	0.806734	0.500191	0.047465	0.103812	0.900523	0.823541
Individuals using the Internet (% of population)	0.655167	0.327924	0.923693	0.806295	0.806734	1.000000	0.849840	-0.291170	0.642153	0.975886	0.802904
Expense (% of GDP)	0.831506	0.065879	0.604293	0.495391	0.500191	0.849840	1.000000	-0.671158	0.843877	0.766335	0.430018
Compensation of employees (% of expense)	-0.625692	0.309711	0.059125	0.059399	0.047465	-0.291170	-0.671158	1.000000	-0.676665	-0.170663	0.253937
Unemployment, total (% of total labor force) (modeled ILO estimate)	0.903888	0.204029	0.368092	0.098184	0.103812	0.642153	0.843877	-0.676665	1.000000	0.475852	0.224403

A correlation matrix shows you how different columns of a dataset correlate. Values range from -1 to 1. If the value is close to 1 and positive, it implies a strong positive correlation. On the other hand, if it is close to -1 and negative, it implies a strong negative correlation wherein the increase of one parameter causes the decrease of the dependent variable.

To check for linear correlation of different variables with the number of suicides, we used Pearson coefficient an p-value as shown below:

The P-value is the probability value that the correlation between these two variables is statistically significant. When the

- p-value is < 0.001: we say there is strong evidence that the correlation is significant.
- the p-value is < 0.05: there is moderate evidence that the correlation is significant.
- the p-value is < 0.1: there is weak evidence that the correlation is significant.
- the p-value is > 0.1: there is no evidence that the correlation is significant.

If the Pearson Coefficient is large, it means there is a strong linear relationship.

We got the following results using this:

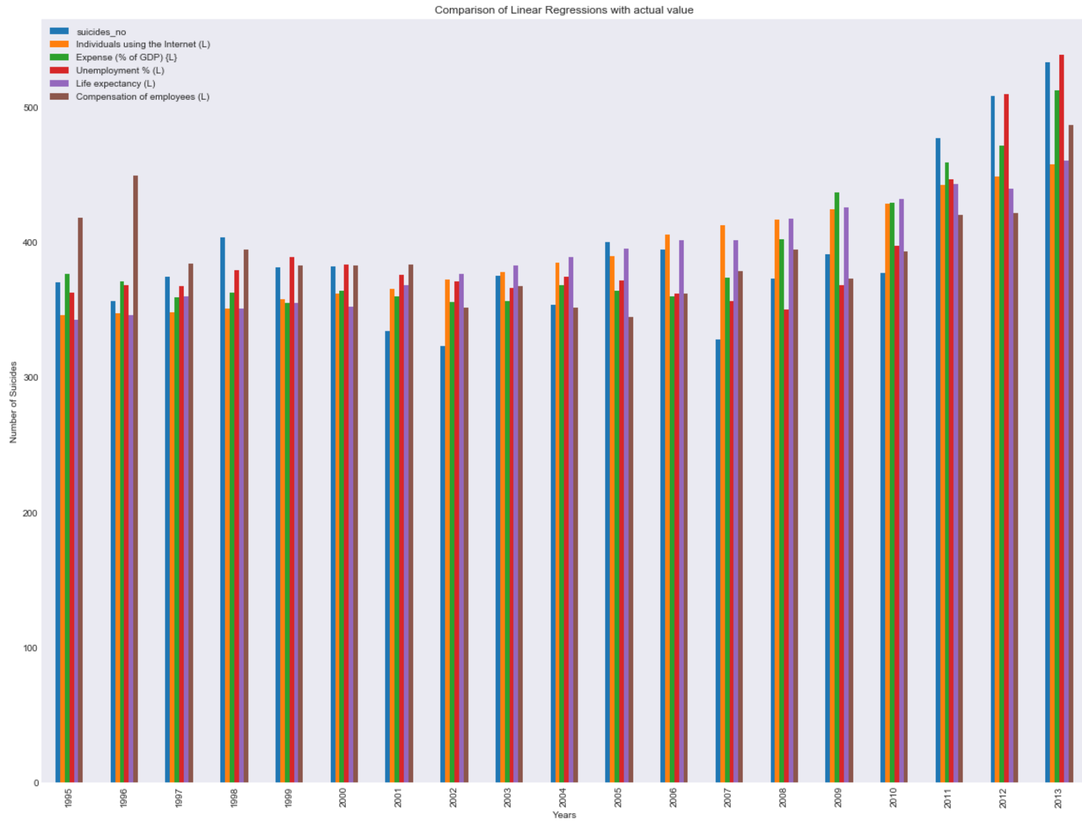
- Individuals using the internet shows a moderate evidence of a significant correlation with the number of suicides and the linear relationship is moderately strong
- Expense (% of GDP) shows a strong evidence of a significant correlation with the number of suicides and the linear relationship is strong
- % Unemployment shows a strong evidence of a significant correlation with the number of suicides and the linear relationship is strong
- Life expectancy at birth, total (years) shows a moderate evidence of a significant correlation with the number of suicides and the linear relationship is moderate
- Compensation of employees (% of expense) shows a moderate evidence of a significant correlation with the number of suicides and the linear relationship is moderately strong

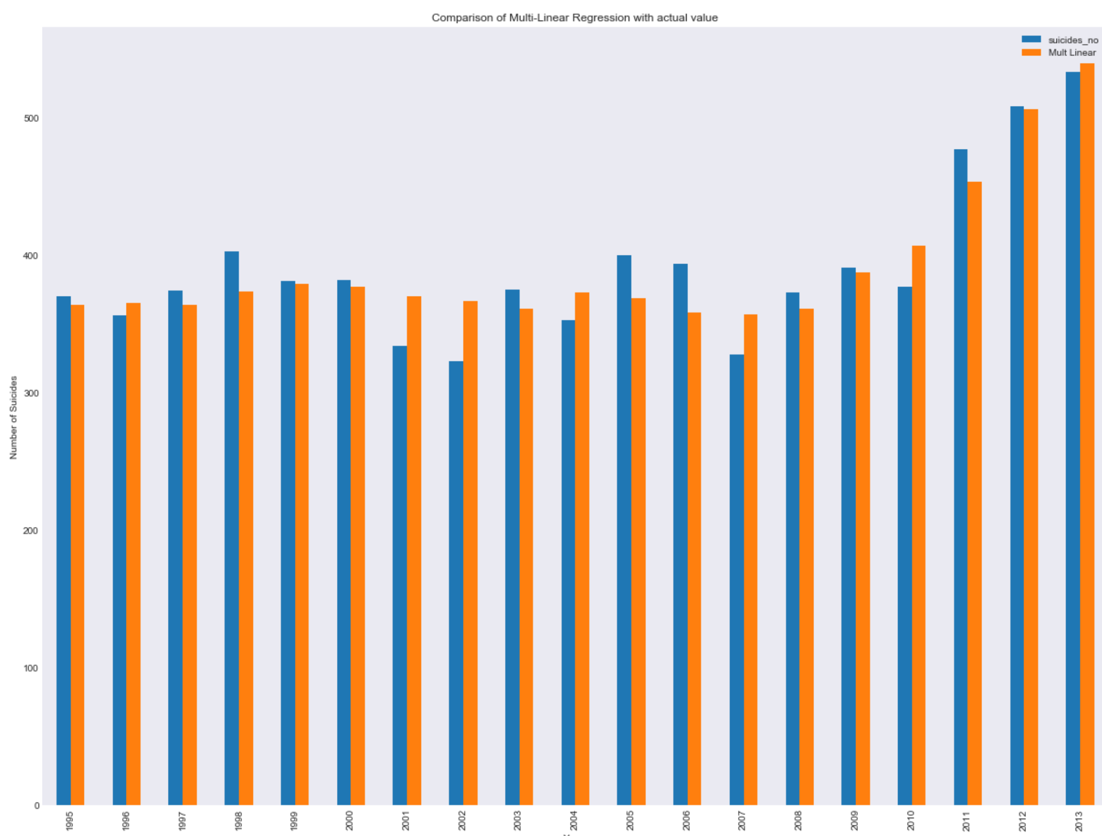
Overall, we observed the following:

```
We find that the following columns have a moderate to strong positive correlation with the number of suicides in Greece:
1. Individuals using the Internet (% of population)
2. Expense (% of GDP)
3. Unemployment, total (% of total labor force) (modeled ILO estimate)
4. Life expectancy at birth, total (years)

Following column has a negative correlation:
1. Compensation of employees (% of expense)
```

- **Predictive Modelling:** In order to develop a model that predicts the number of suicides per year given the values of influencing factors, we used the following algorithms: Linear Regression, Multiple Linear Regression, Polynomial Regression, Decision Trees, AdaBoost Trees, Random Forest. We achieved a fair amount of accuracy using these models. The below bar graphs show the predicted values being compared with actual values of each year:

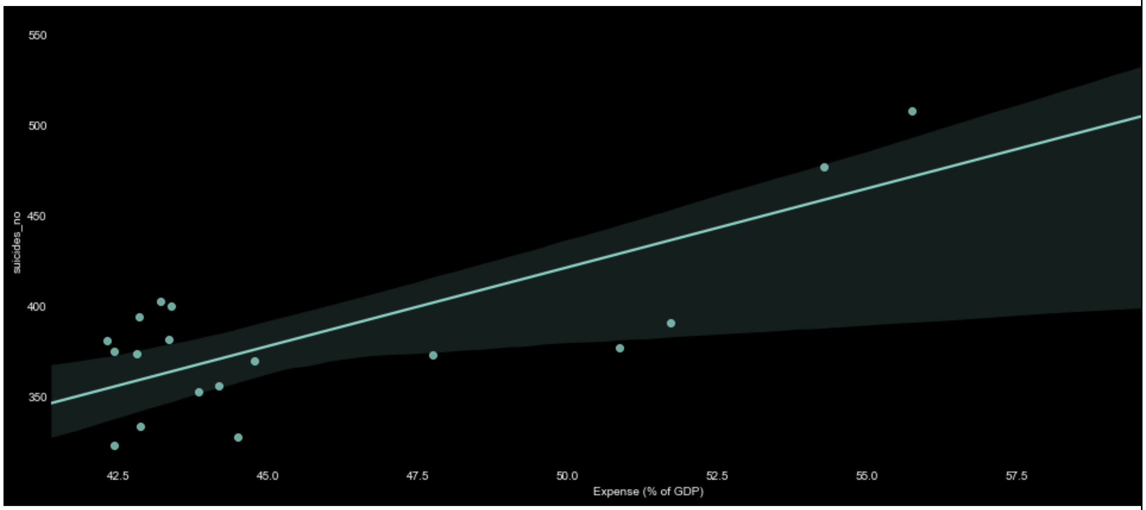


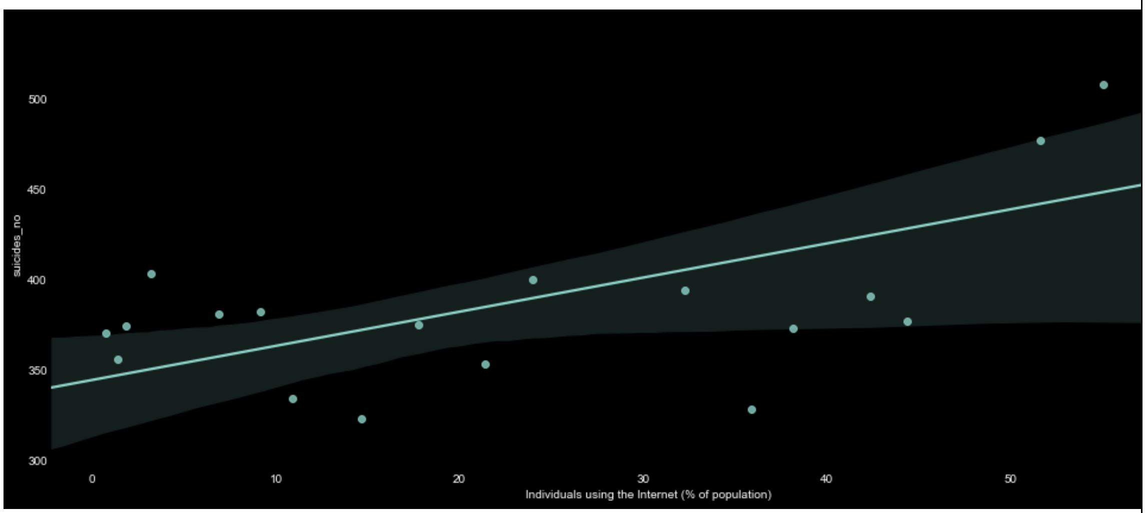


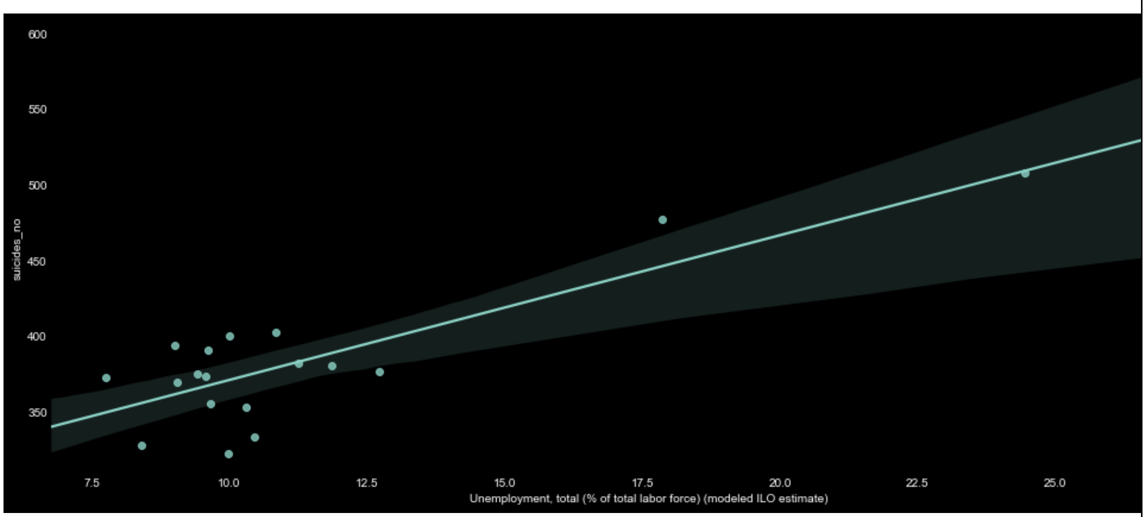
- **Model Evaluation:**

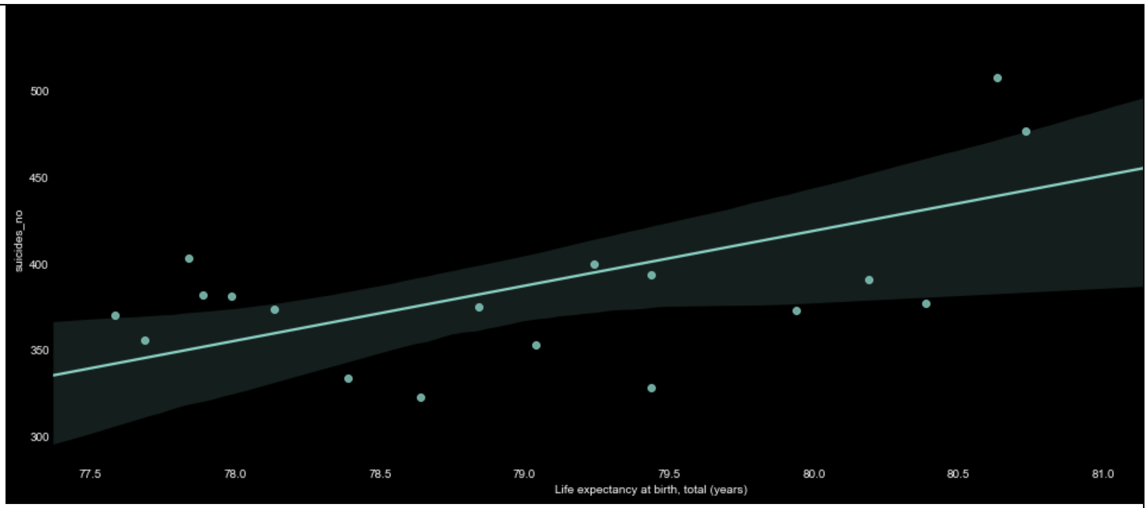
Once the above models were developed, in order test the effectiveness of each model, we conducted some tests:

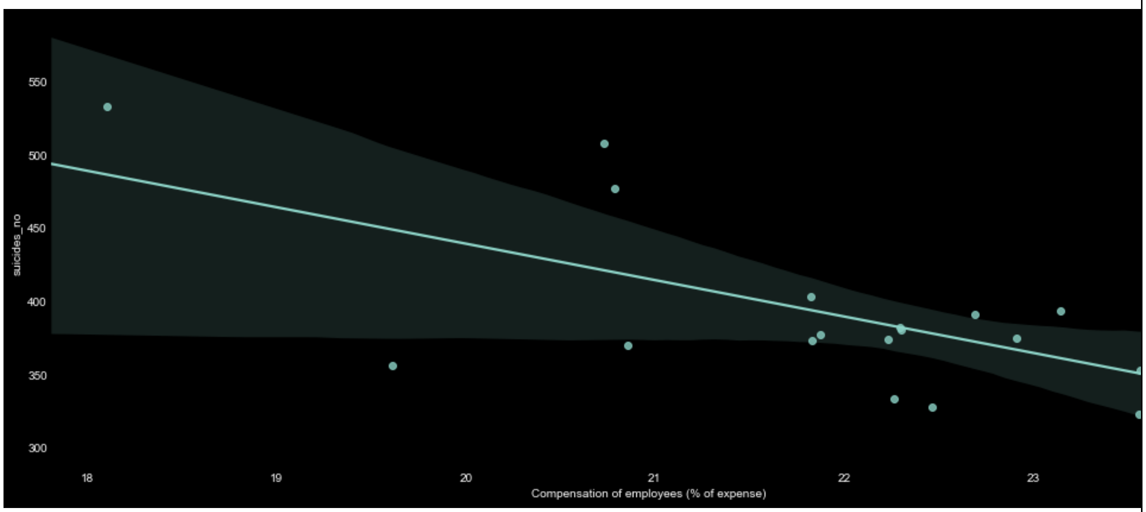
Regression Plots:



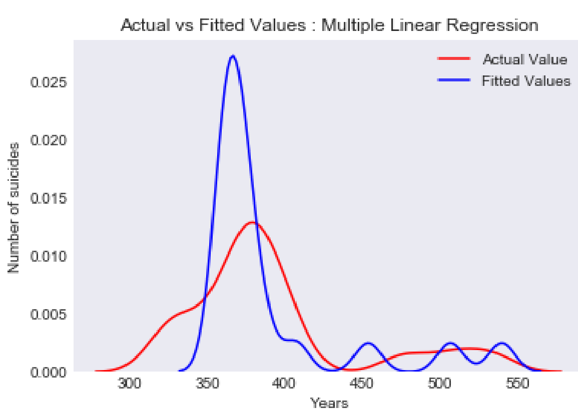


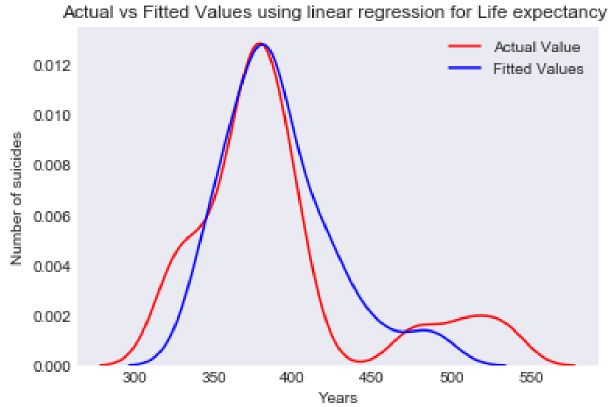


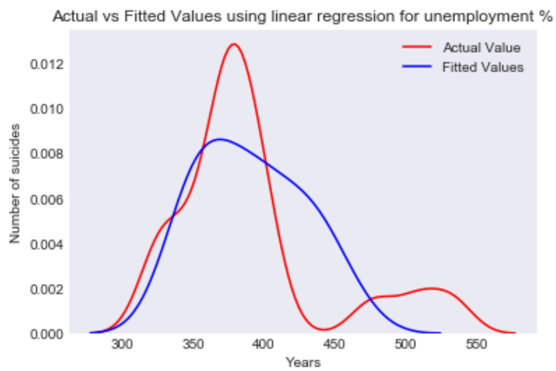


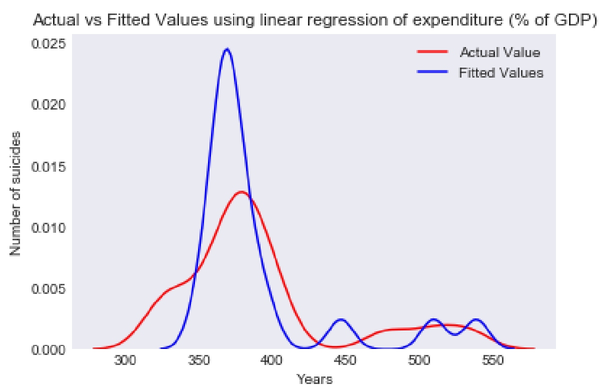


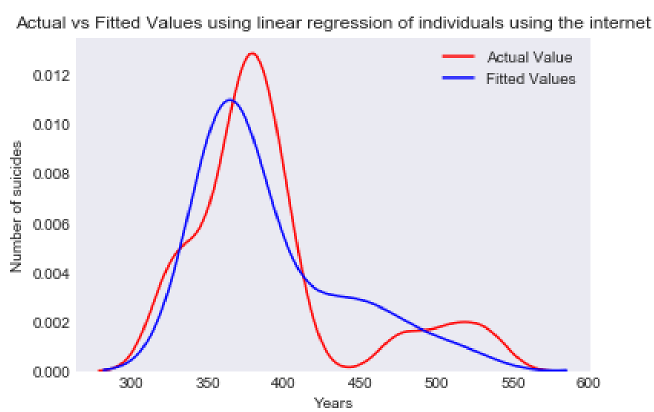
Distribution Plots:











R^2 Scores and Root Mean Square Error Reuslts:

Random Forest:

MSE : 88367.26827639317

RMSE : 297.2663255002039

r2_score : 0.8492420728236623

Decision Trees:

MSE : 149612.27254243137

RMSE : 386.79745674245504

r2_score : 0.7447557616232916

AdaBoost Regressor:

MSE : 262742.37335208926

RMSE : 512.5840158960181

r2_score : 0.5517514984840426

Linear Regression with Individuals using the Internet (% of population):

The R-square value is: 0.4292437980319016

The mean square error is: 1732.7114803958398

Linear Regression with Expense (% of GDP) (% of population):

The R-square value is: 0.6914027403312277

The mean square error is: 936.844860910797

Linear Regression with Unemployment, total (% of total labor force):

The R-square value is: 0.8170139397104347

The mean square error is: 555.5122245239428

Linear Regression with Life expectancy at birth:

The R-square value is: 0.419990876490284

The mean square error is: 1760.8016585263242

Linear Regression with Compensation of employees (% of expense):

The R-square value is: 0.3914901510863291

The mean square error is: 1847.324650193793

Multiple Linear Regression with all the above factors:

The R-square value is: 0.833730907615482

The mean square error is: 504.76256625198

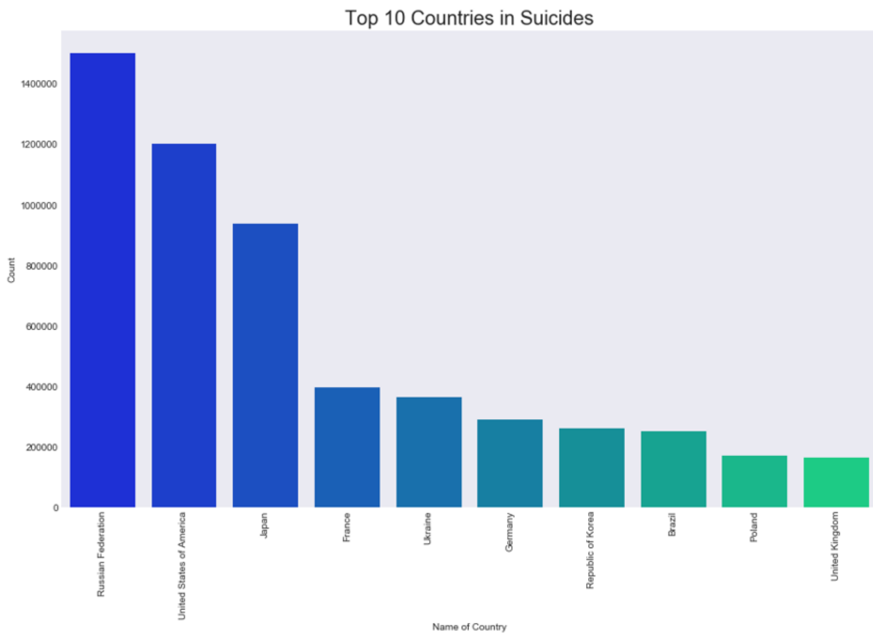
Polynomial Regression:

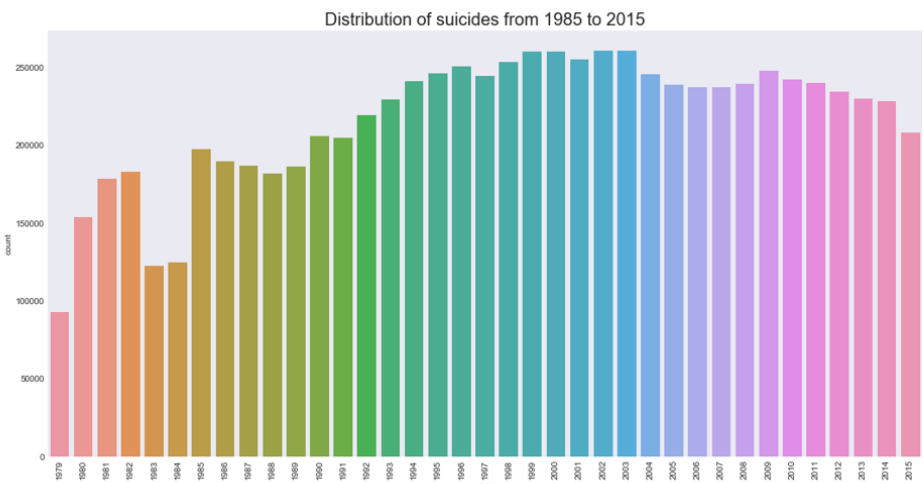
The R-square value is: 0.8270000073159866

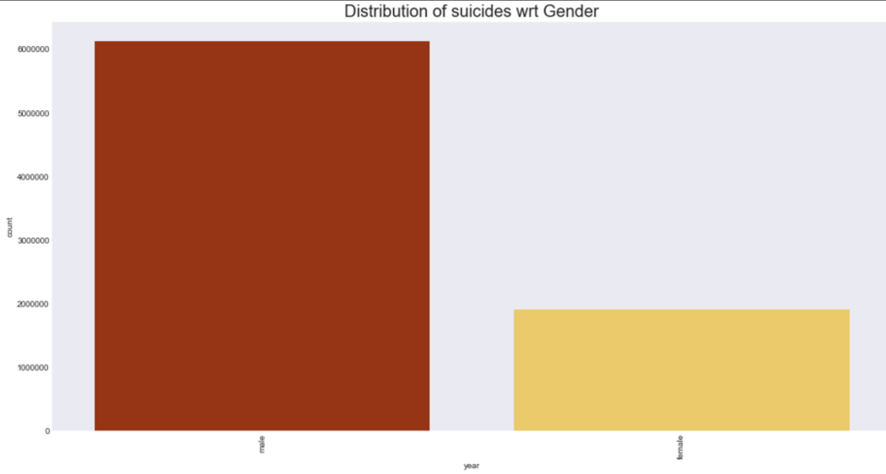
MSE: 525.1963489811382

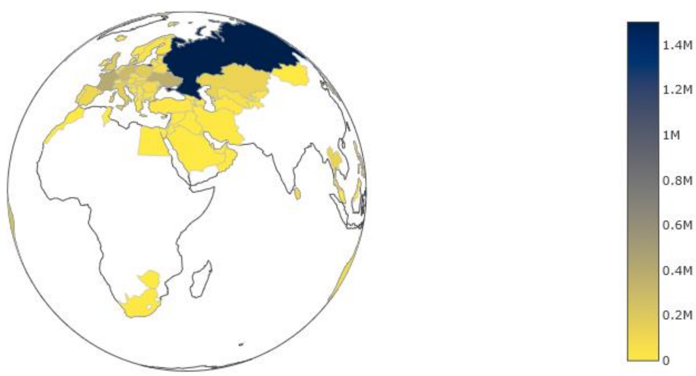
CHAPTER 4: SCREENSHOTS

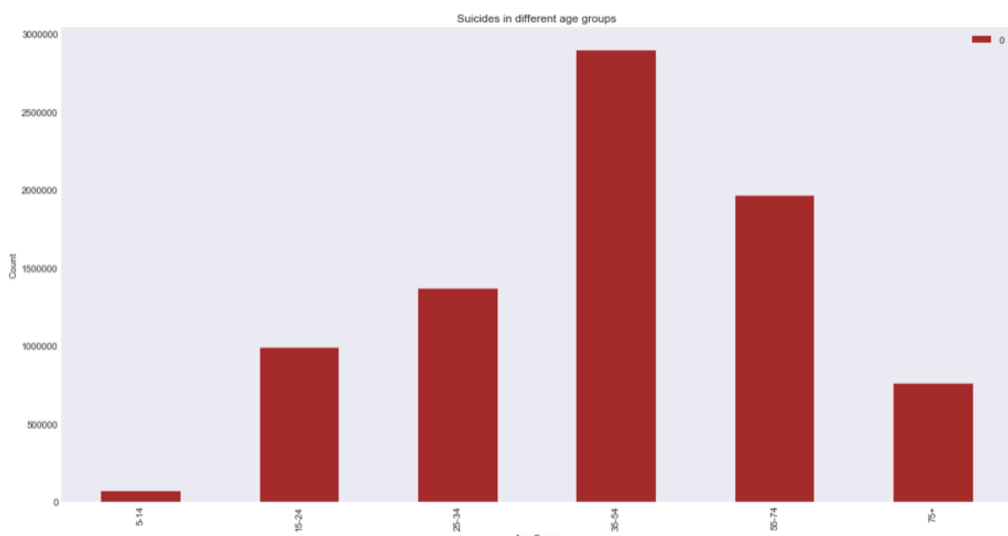
Following are the rest of the screenshots and visualizations:

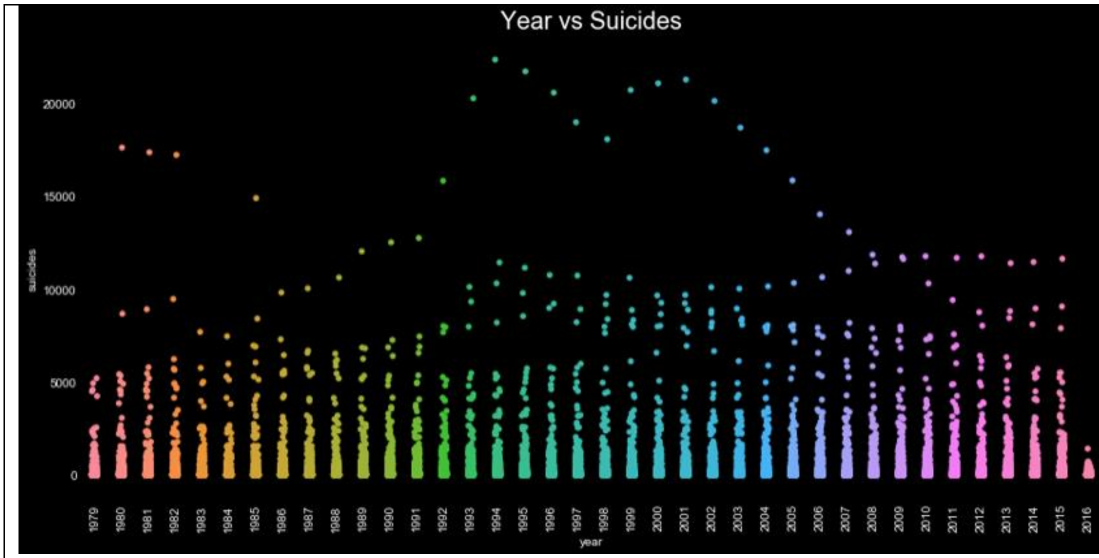


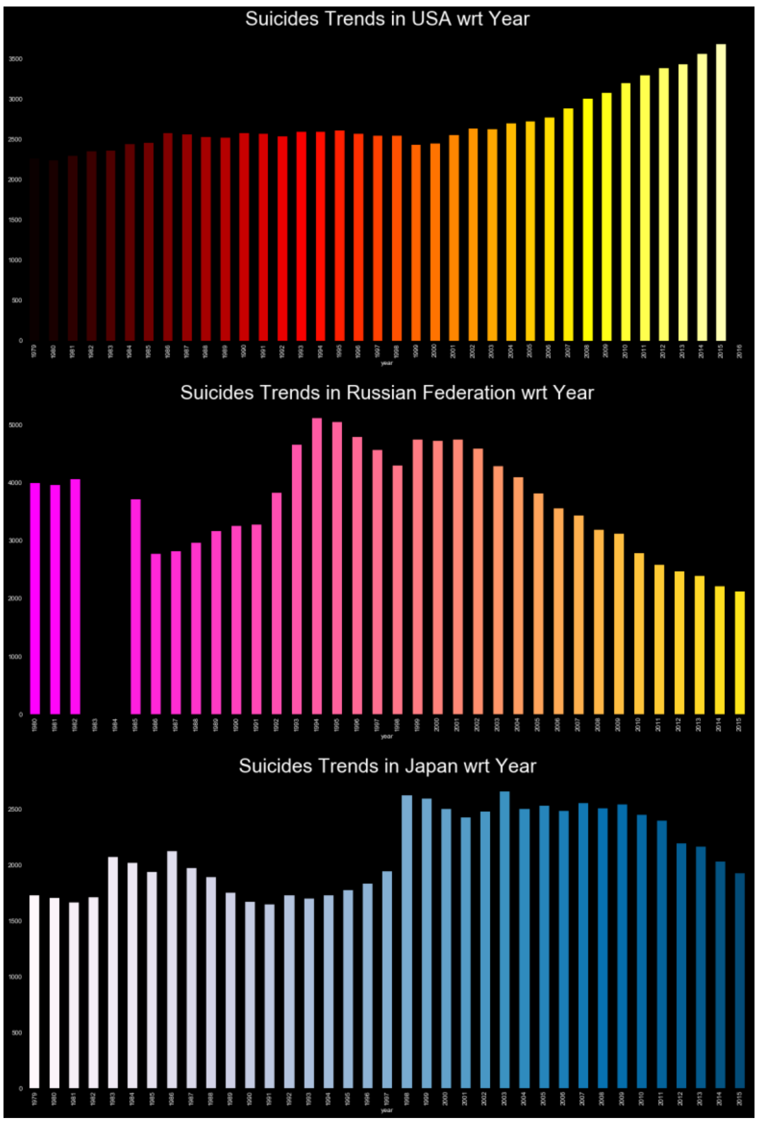


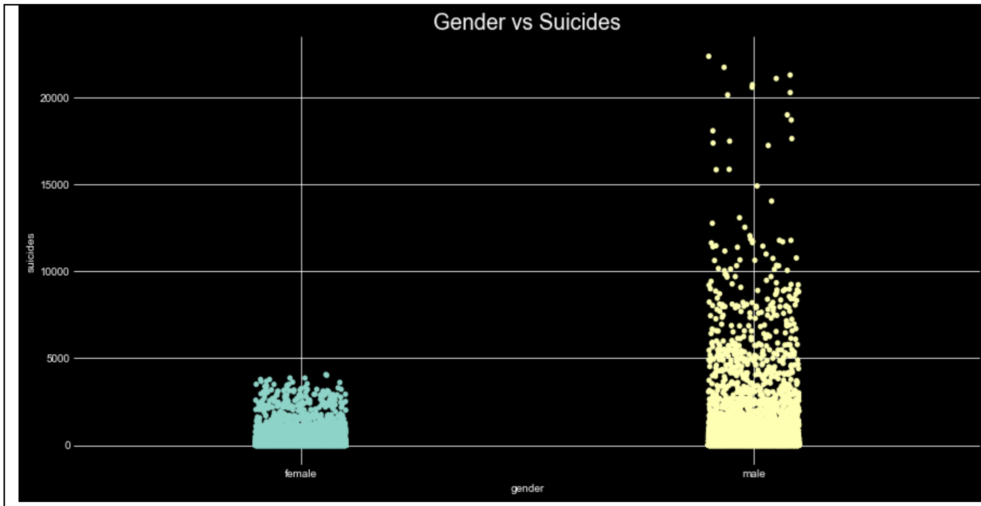

Distribution of suicides wrt Gender
6000000
5000000
4000000
3000000
2000000
1000000
0
count
male
female
year

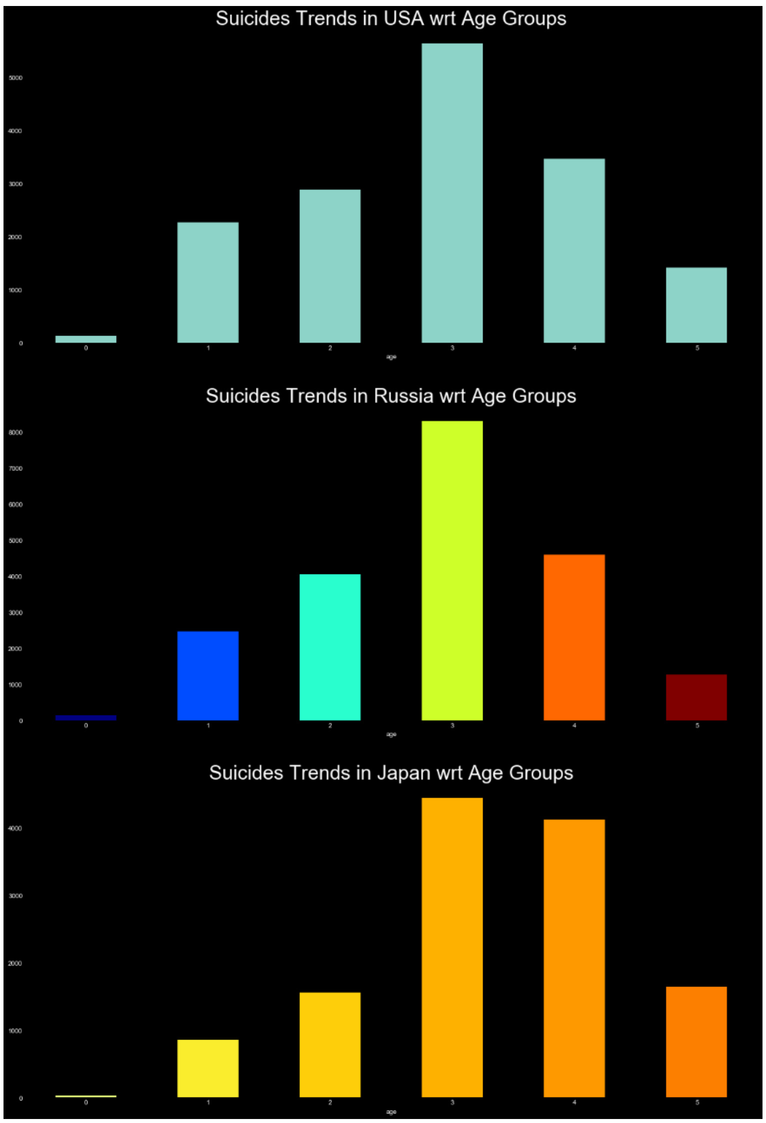

1.4M
1.2M
1M
0.8M
0.6M
0.4M
0.2M
0

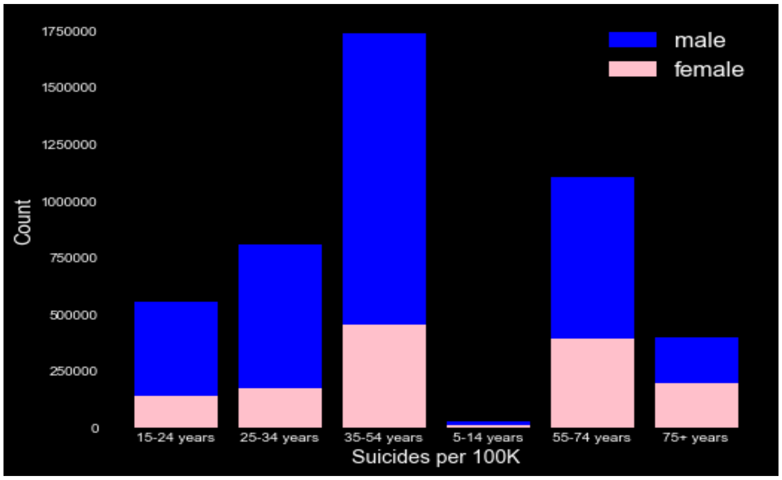

Suicides in different age groups
0
3000000
2500000
2000000
1500000
1000000
500000
0
Count
5-14
15-24
25-34
35-54
55-74
75+
Age Group

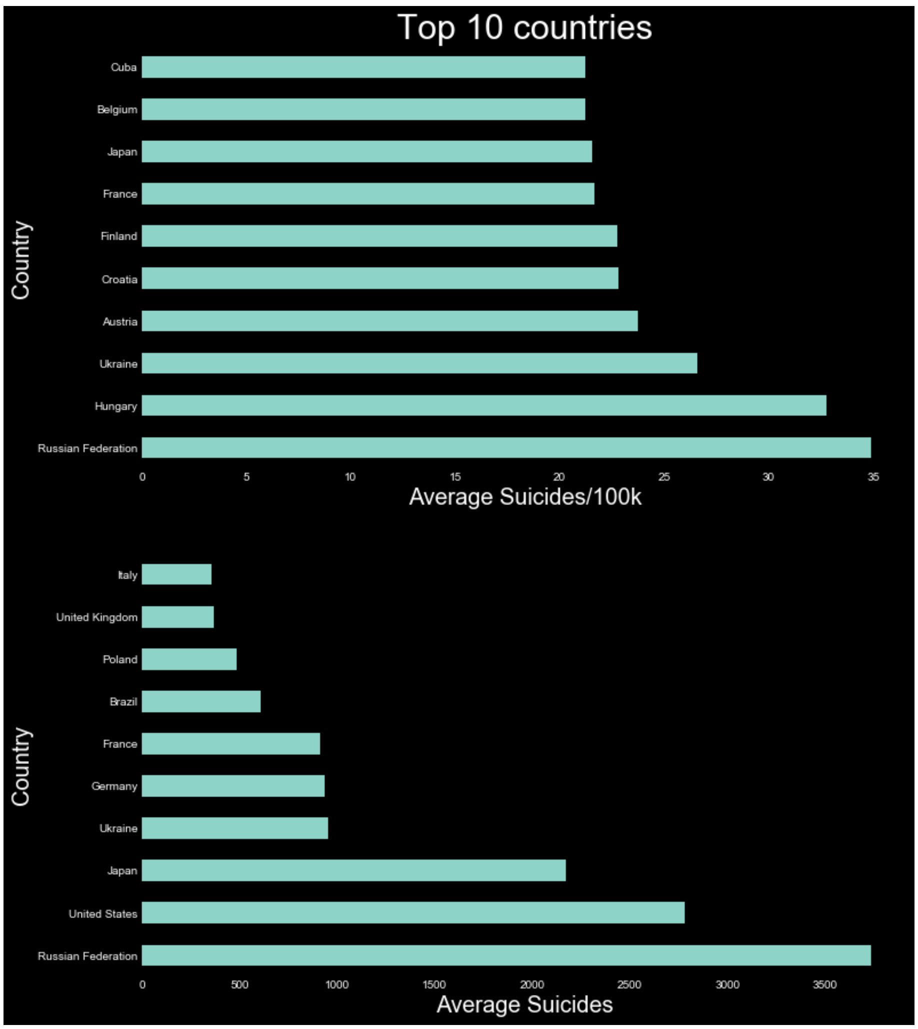

Year vs Suicides
suicides
20000
15000
10000
5000
0
1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016
year

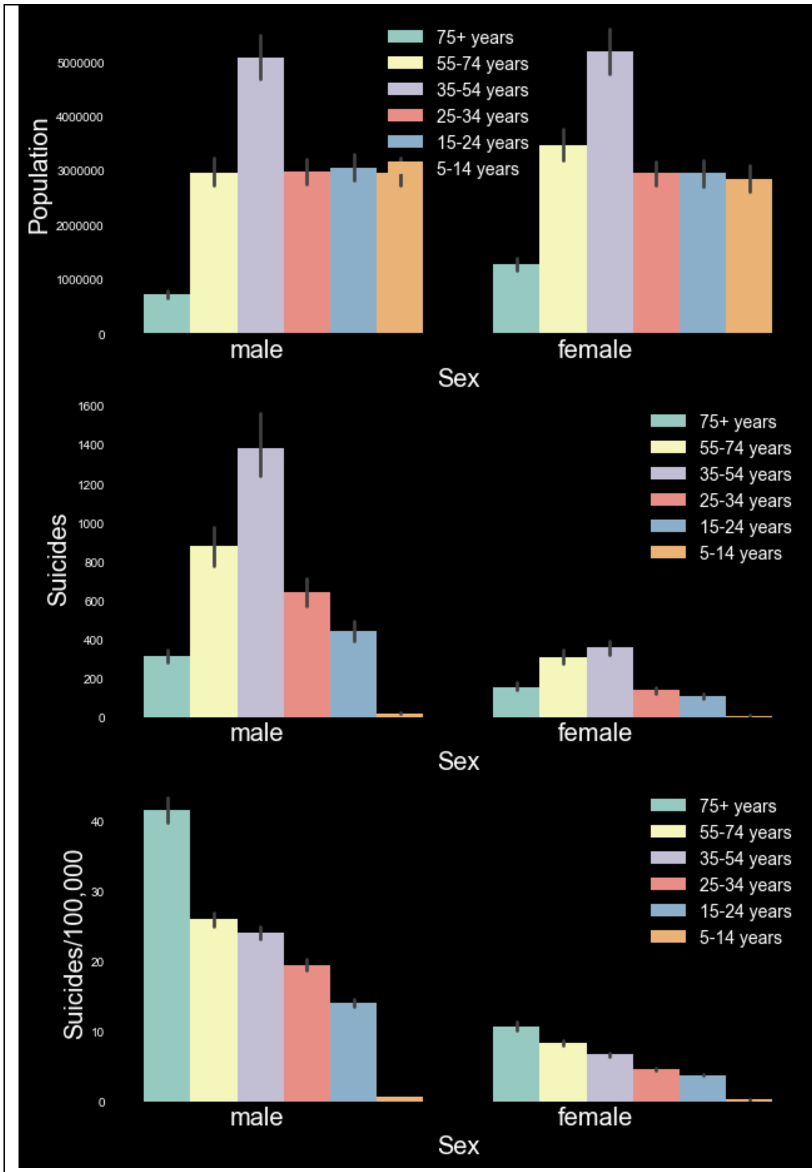

Suicides Trends in USA wrt Year
Suicides Trends in Russian Federation wrt Year
Suicides Trends in Japan wrt Year

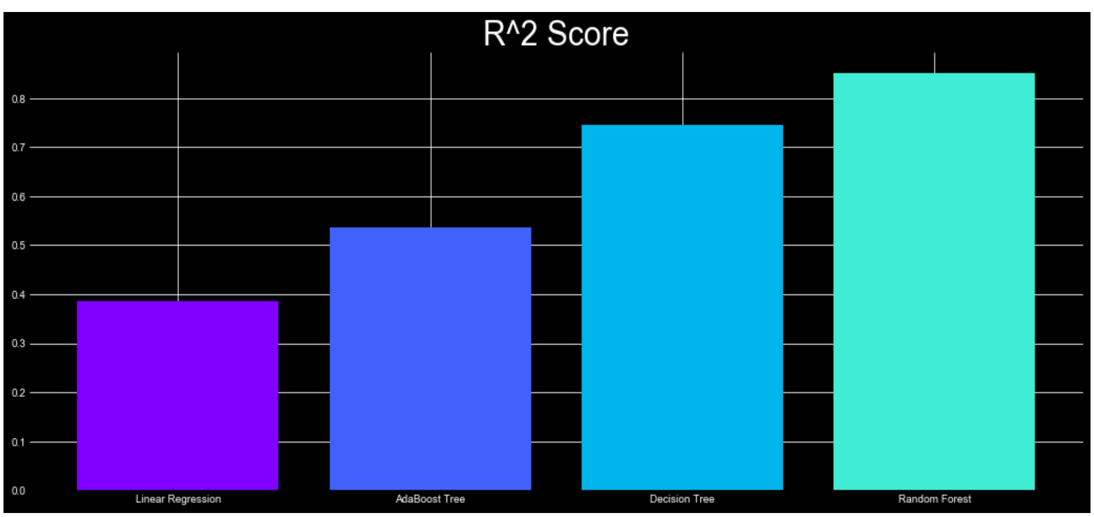

Gender vs Suicides
20000
15000
10000
5000
0
suicides
female
male
gender

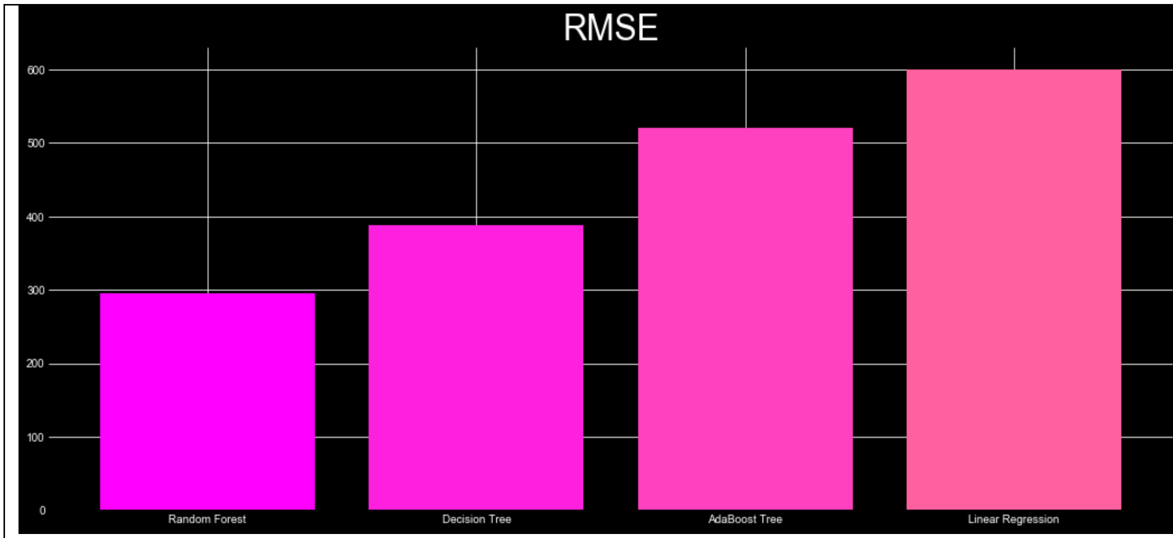

Suicides Trends in USA wrt Age Groups
Suicides Trends in Russia wrt Age Groups
Suicides Trends in Japan wrt Age Groups


male
female
Count
1750000
1500000
1250000
1000000
750000
500000
250000
0
15-24 years
25-34 years
35-54 years
5-14 years
55-74 years
75+ years
Suicides per 100K


Top 10 countries
Country
Cuba
Belgium
Japan
France
Finland
Croatia
Austria
Ukraine
Hungary
Russian Federation
0
5
10
15
20
25
30
35
Average Suicides/100k
Country
Italy
United Kingdom
Poland
Brazil
France
Germany
Ukraine
Japan
United States
Russian Federation
0
500
1000
1500
2000
2500
3000
3500
Average Suicides


75+ years
55-74 years
35-54 years
25-34 years
15-24 years
5-14 years
Population
5000000
4000000
3000000
2000000
1000000
0
male
female
Sex
Suicides
1600
1400
1200
1000
800
600
400
200
0
male
female
Sex
Suicides/100,000
40
30
20
10
0
male
female
Sex


R^2 Score
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Linear Regression
AdaBoost Tree
Decision Tree
Random Forest



CHAPTER 5: CONCLUSION AND FUTURE SCOPE

Conclusion:

- We performed an analysis of existing data on suicides and tried to attain meaningful observations using different visualizations and predictive modelling.
- We believe this could be used to pin-point various factors that drive individuals to commit suicide and try and address these issues.
- The models developed can be used to predict the effect of different correlative factors with suicide rates.
- We found the following models to have sufficient accuracy:
    - Random Forest Regressor
    - Decision Tree Regression
    - Linear Regression with Expense (% of GDP) (% of population):
    - Linear Regression with Unemployment, total (% of total labour force):
    - Multiple Linear Regression
    - Polynomial Regression

Future Scope:

- We can conduct a similar analysis for different countries and figure out different factors affecting the suicide rates there.

- Here, we merged HDI (Human Development Index) data released by the World Bank with suicide statistics released by WHO. We carry out a similar process with other such datasets such as suicides during times of military conflicts
- A dashboard can be created which displays the visualizations created in a user friendly manner.
- In the following years, increase in the amount of data available and additional factors could also play a part in affecting suicide rates. These could be identified and analysed.

CHAPTER 6: SOCIETAL APPLICATION

- We believe this analysis can be used to identify prominent factors that influence the number of suicides.
- Predictive models developed by us can be used to predict effects due to changes in factors affecting suicide rates such as an increase in unemployment rates or the falling of the GDP of a nation.
- Visualizations created can be used to increase awareness and illustrate the magnitude of the issue at hand.
- We believe that analysis of data is necessary in order to control suicide rates. Without having a clear picture of the reasons that drive individuals to such steps, it is not possible for one to take any concrete steps.