INFOSYS 722 / RESEARCH REPORT

Analysis of the world suicide population from 1985 to 2016

Name Jinyun Liu

Student ID 246149448

UPI jliu737

Contents

1. The Practical Problem

Suicide is a dangerous behavior that an individual intentionally or voluntarily takes various means to end his life under the action of complex psychological activities.

There are the facts about the situation of suicide:

- It is estimated that between 500000 and 1.2 million people worldwide die by suicide.[8]
- In most European countries, more people die of suicide than traffic accidents each year. Globally, the number of suicide deaths exceeds the combined number of deaths due to murder and war. [16]
- The report from World Health Organization shows that from 1950 to 1995, the number of suicides increased from 10.1 per 100000 to 16 per 100000. [16]
- Suicide has become one of the top ten causes of death in humans.[16]
- Suicide is a global phenomenon that occurs in all regions of the world. The global Suicide Distribution Map in 2012 can be seen in Figure 1.1.[16]

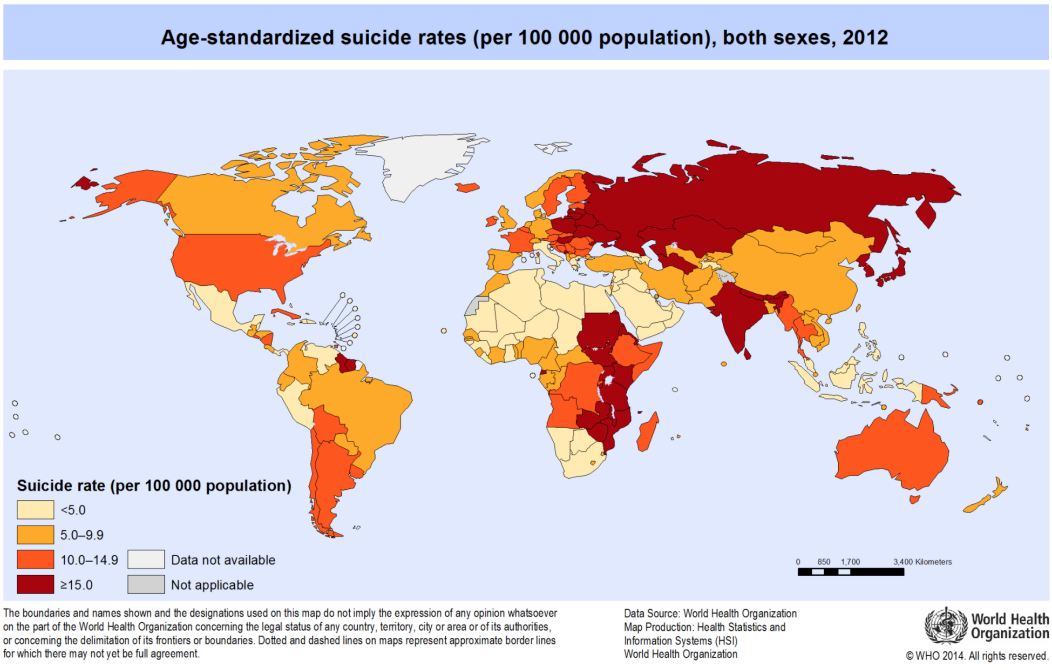

Figure 1.1

These facts show that every year, a large proportion of people lose their lives because of suicide. Suicide has become an important cause of human death, and the suicide rate is still rising. The World Health Organization has listed suicide as a global public health problem and calls on all human society to take measures to prevent suicide tragedies from occurring.

Suicide will have a negative impact on many aspects of society, such as:

- Suicide will have a great impact on the family and society, causing great suffering in the family, and the economic loss of the society due to the loss of life.
- Suicide, suicide attempts will have a series of negative reactions to family, friends, community and society[15]

From these facts and influences, we can know that suicide has become one of the biggest urgent problems in the world.

Therefore, we should actively respond to the call of WHO and take measures to prevent suicide tragedy.

2. The Research Problem.

Every year, more than 800,000 people die by suicide, and many more attempt suicide. Therefore, every year, millions of people experience or are affected by the pain of bereavement caused by suicide. Suicide can occur throughout the life process and is the second leading cause of death for people aged 15-29 years in the world in 2012.[16]

As we have discussed before, suicide is a major social issue, and reducing the suicide rate is the thing we urgently need to do. It is difficult for us to directly prevent people from committing suicide, but we can make reasonable suggestions to relevant departments by analyzing the relationship between the suicide rate and different factors and the future trend of suicide rate. According to the factors affecting suicide and the trend of suicide rate, we can put forward reasonable suggestions to the relevant departments to reduce the suicide rate.

To be specific, by knowing these:

- By knowing the trend of suicide rate, the government and society will pay more attention and take more measures to prevent people from committing suicide.
- According to the relationship between the suicide rate and different factors, the relevant organizations can try to develop different programs to change these factors to help specific groups of people.

Therefore, we need to do a research to conduct a specific data analysis about these. Because a specific data analysis about suicide rates can give us a predict of future trends in suicide rates and find the various factors that affect suicide rates. With the development of data science and technology, big data analysis has become a viable method for predicting accurate results

3. The Research Objectives

In this research, we need to select a suitable dataset to predict. The data source used in this project is "Suicide Rate Overview from 1985 to 2016" provided by Kaggle which is a reliable, freely-available dataset website.

By using data mining and machine learning methods for this dataset, we need to achieve the following research objectives:

- Analyze the future trend of suicide rate.

  By identifying the future trend of suicide, we can give some warning to the world and remind the human society to pay attention to the suicide group.

- Find out the relationship between the suicide rate and different factors and identify the factors that most influence the suicide rate.

  Finding these relationships mean that if we know that these factors affect people's suicide rates, we can try to change these factors to reduce the suicide rate.

Tentatively, the study will be judged a success if:

- Accurate analysis of future suicide rate trends. Through MSE and R-square to evaluate the accurate of the models. The margin of error is less than 15%.
- Through analysis, sum up various factors affecting suicide rate and provide effective advice to relevant organizations. After improving the related factors, the suicide rate can be significantly reduced.
- The project is finished in time.

4. Literature Review.

Some studies have been conducted on the trend of suicide rate and the suicide factors that affect suicide rate, such as gender, age, country, etc. There have been relevant studies on the factors between these factors and suicide rate.

In 2009, KA-YUET LIU and D PHIL presented a research paper aims to find the relationship between country and suicide rate. They selected the trends of suicide rates in 71 countries from 1950 to 2004. Their research shows that suicide rates vary greatly among different countries. Even in regions with similar development levels, there is still a big gap in suicide rates among different countries. They used random effects models to test the stability of suicide rates between and within countries. The results show that more than 90% of the difference in suicide rates is caused by differences between countries, which indicates that the suicide rate exhibits strong temporal stability.[10]

In 2020, Agnus M.Kim presented a research paper applied Pearson correlations and a multiple linear regression to study the relationship between suicide rate and heavy drinking, health care, religion, and demographic characteristics in 251 regions of South Korea. The regression analysis showed that regional income levels were negatively correlated with suicide rates, while the prevalence of heavy drinking and the percentage of people aged 65 and over were positively correlated with suicide rates. When taking measures to prevent suicide, special attention should be paid to excessive drinking and social and economic conditions.[12]

In 2011,Wafaa M.Abdel et al.did a retrospective research on evaluating suicide cases in Assiut from 2005 to 2009. Their research showed men have a higher suicide rate than women, and male victims try to use more violent methods than women when they commit suicide. Since 1987, the suicide rate has been on the rise, which is an urgent problem to be solved.[14]

In 2014, Caroline Coopea and others published a paper to determine the population groups most affected by suicide rates. They found that unemployment and layoffs are key risk factors for unemployed and long-term unemployed men aged 35-44. In addition to unemployment and layoffs, other indicators of economic stress, such as personal debt and house resettlement, may increase the suicide rate of young men.[4]

In 2010, Yip, Paul S F; Caine and Eric D did a research to study the correlation between unemployment rate and suicide rate, and the exact relationship. The result shows that there is a strong correlation between the unemployment rate and the suicide rate (0.86), but the suicide rate among the employed and unemployed groups is contrary to the overall economic trend. In other words, the suicide rate of the unemployed fell during the economic recession and rose during the recovery. According to the result, they suggest that when formulating suicide prevention strategies, it is necessary to be able to accurately distinguish between population-level concepts (such as ratios) and individual-level characteristics (such as employment status).[5]

In 2013, Mok, Pearl L H et al.did a research to study why Scotland has the highest national suicide rate among all British countries. The author focused on the factors related to the suicide rate and society, culture and health. The conclusion shows that 57% of the suicide risk in Scotland is explained by a series of regional measures, including psychotropic drug

prescription, alcohol and drug abuse, socioeconomic deprivation, social differentiation and other health-related indicators. Socioeconomic deprivation and social differentiation have a relatively small impact on suicide rates. Addressing these issues can also reduce the risk of excessive suicide rates in Scotland.[13]

In 2016, Inoue, Ken1 et al.did a long-term research to study the links between various economic factors and suicide in order to determine the link between economic concerns and suicide. This study investigated the relative poverty rate and suicide rate in Japan over the past 30 years, and studied the relationship between the two rates. The results show that the relative poverty rate may be related to the suicide rate of men and women. This connection applies especially to men. Organizations and professionals involved in implementing suicide prevention measures should understand the current findings and consider formulating other specific measures.[9]

In 2012, Ajit Shah presented a paper to investigate factors associated with age-related trends in suicide rates. The authors studied the relationship between socioeconomic status, income inequality, health care expenditure, child mortality and life expectancy with increasing and decreasing suicide rates. The result show that age related trends in suicide rates were significantly associated with socioeconomic status (men) or income inequality (women), per capita medical expenditure, the proportion of gross national product (GDP) devoted to health care, child mortality and life expectancy.[17]

In 2011, Yip, Paul S F, Caine and Eric D presented a paper to study the exact mechanism of the strong interaction between the unemployment rate and the suicide rate. This study examines the relationship between the suicide rate and the regional unemployment rate and personal employment status during Hong Kong's economic recession (2000-3) and recovery (2003-6). The results show that although there is a strong correlation between the unemployment rate and the suicide rate, the suicide rate of the employed and unemployed groups is contrary to the overall demographic trend. During the economic recovery, the unemployment rate and suicide rate are both falling.[19]

In 2018, India State-Level Disease Burden Initiative Suicide Collaborators presented a paper to report the temporal trends and heterogeneity of suicide deaths in Indian states from 1990 to 2016. They estimated suicide death rates for both sexes in each state of India from 1990 to 2016. The report found significant differences in suicide rates between men and women in some regions. In 2016, suicide was the main cause of death in Indians aged 15-39; the proportion of suicide deaths among women and men was 71.2% and 57.7% respectively. India accounts for a high proportion of global suicide deaths and will increase. For women, the range and gender ratio vary widely between states. They called on India to develop suicide prevention strategies that take into account these differences to address this major public health problem.[6]

In 2017, Brazinova et al.presented a paper indicated to analyze the time trends of suicide rates in the Slovak Republic in 1993–2015 for targeted suicide prevention strategies. They used join point regression and negative binomial regression to calculate the trends and relative risks of suicide according to age and sex. The results show the suicide rate in the Slovak Republic is slightly below the average of Organisation for Economic Cooperation and Development (OECD) nations. Highest suicide rate is observed in men of working age and in retirement. Society might benefit from a strategy of education for improving the recognition of suicide risks.[1]

5. Research Methodology Adopted.

Based on the research objectives, this research focus on the following data mining goals:

- Analyze the historical data of the world suicide rate through different data analysis models. Find the lowest error according to the performance of different models, and the best performing model to predict the future development trend of suicide rate.
- Determine the important factors that affect the suicide rate by analyzing the relationship between different factors in the data set and the suicide rate, and rank them according to the degree of impact. Finally, we give the visual effect of data analysis to provide support and suggestions to relevant departments more clearly and intuitively.

Data Mining Success Criteria in this research:

- Prediction: If the accuracy of prediction is more than 70%, we can believe that the prediction is success.
- Explanation: If we get a rank for the important features, the explanation will be considered as success.

The methods of data mining can be roughly divided into two categories. They are Supervised Method and Unsupervised Machine Learning Algorithms. [3]

- Supervised Learning

  Supervised Learning is a method in machine learning, which can learn from training materials or establish a pattern (function / learning model), and infer new examples according to this model.

  There are two types of learning supervision problems, One of them is regression problem and another is classification problem.

  - Regression

    Regression method is a supervised learning algorithm for predicting and modeling numerical continuous random variables. The characteristic of regression task is that the labeled dataset has numerical target variables. Use cases generally include house price forecast, stock trend or test results and other continuous changes.
  - Classification

    Classification method is a supervised learning algorithm for modeling or predicting discrete random variables. Classification algorithms are usually used to predict a category (or the probability of a category) rather than continuous values. Use cases include email filtering, financial fraud, and forecasting employee turnover.

The objective in this project is to predict the trend of suicide rates. The target we want to predict is a numerical variables, so based on the previous discussion, the regression method is more suitable for our project.

- Unsupervised Learning

  The feature of unsupervised learning is that the data of model learning has no label, so the goal of unsupervised learning is to reveal the inherent characteristics and laws of data by learning these unlabeled samples, and its representative is clustering.

  - Clustering

    Clustering is a method of partitioning the dataset into groups. The goal is to split up the data in such a way that points within single cluster are very similar and points in different clusters are different. It determines grouping among unlabeled data

  - Association

    An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.[3]

  Our project is to predict the trend of suicide rate and find the factors which can affect the suicide rate. Our target suicide rate has labeled in the dataset, so this method may not suitable for our prediction. Therefore, I should choose regression model of supervised learning to conduct our objectives.

  There are different ways to build the model, such as SPSS, Python, Pyspark and so on.

  IBM SPSS Modeler is IBM's data mining software application. Users have a visual interface that allows them to use statistical information and data mining algorithms without programming. Make complex predictive models very easy to use.

  Python is an interpreted high-level programming language. Python has a comprehensive standard library, of which PySpark is Spark's Python API. PySpark allows interaction with Spark using the Python programming language. This makes the interaction with Spark simple and effective. We also can use many Python packages to process and analyze datasets, such as Pandas, Sklearn, etc., They can make the process of data mining easier.

  All tools have their own advantages. In this research, I will choose Python and Pyspark to bulid our models. Some packages such as pandas, sklearn, and seaborn will be used.

  The process of this research include Business understanding, Data Understanding, Data preparation and Transformation, Modelling, Evaluation, and Deployment.

  First of all, we should select a suitable datase for this research.

  Then, check the characteristics and quality of the data set in the "data understanding" stage.

  Thirdly, we should check the quality of the dataset to find if there are missing value, outliers in the dataset, which can be solved in the step of Data preparation.

  Next, after preparing the data. We need to use the selected method and algorithm to fit the number. And it is necessary to calculate the accuracy and error of the model to evaluate the performance of each model.

  Finally, explain the results based on the results of each model. If the output of this method can meet our requirements, it should describe its implementation in real life and monitor it.

6. The design of the processes that converts data into insights

6.1 Data Source

The dataset used in this study comes from the free and open source website Kaggele. I downloaded the dataset named Suicide Rates Overview 1985 to 2016 from kaggle by this URL: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-20 This dataset pulled from four other datasets linked by time and place. The data in the dataset comes from trusted data sources like World Bank and United Nations Development Program.

6.2 Data Description

This data set records 27,820 records and 12 attributes of suicide data. The time interval of this data is from 1985 to 2016. The structure of data has been shown below: There

```
df.show()
```

```
+-------+----+------+-----------+-----------+----------+-----------------+------------+------------+-----------------
-+------------------+---------------+
|country|year|   sex|        age|suicides_no|population|suicides/100k pop|country-year|HDI for year| gdp_for_year ($)
|gdp_per_capita ($)|     generation|
+-------+----+------+-----------+-----------+----------+-----------------+------------+------------+-----------------
-+------------------+---------------+
|Albania|1987|  male|15-24 years|         21|    312900|             6.71| Albania1987|        null|     2,156,624,90
0|               796|   Generation X|
|Albania|1987|  male|35-54 years|         16|    308000|             5.19| Albania1987|        null|     2,156,624,90
0|               796|         Silent|
|Albania|1987|female|15-24 years|         14|    289700|             4.83| Albania1987|        null|     2,156,624,90
0|               796|   Generation X|
|Albania|1987|  male|  75+ years|          1|     21800|             4.59| Albania1987|        null|     2,156,624,90
0|               796|G.I. Generation|
|Albania|1987|  male|25-34 years|          9|    274300|             3.28| Albania1987|        null|     2,156,624,90
0|               796|        Boomers|
|Albania|1987|female|  75+ years|          1|     35600|             2.81| Albania1987|        null|     2,156,624,90
0|               796|G.I. Generation|
|Albania|1987|female|35-54 years|          6|    278800|             2.15| Albania1987|        null|     2,156,624,90
0|               796|         Silent|
|Albania|1987|female|25-34 years|          4|    257200|             1.56| Albania1987|        null|     2,156,624,90
0|               796|        Boomers|
|Albania|1987|  male|55-74 years|          1|    137500|             0.73| Albania1987|        null|     2,156,624,90
0|               796|G.I. Generation|
|Albania|1987|female| 5-14 years|          0|    311000|              0.0| Albania1987|        null|     2,156,624,90
0|               796|   Generation X|
|Albania|1987|female|55-74 years|          0|    144600|              0.0| Albania1987|        null|     2,156,624,90
0|               796|G.I. Generation|
|Albania|1987|  male| 5-14 years|          0|    338200|              0.0| Albania1987|        null|     2,156,624,90
0|               796|   Generation X|
```

Figure 6.1: Date structure

are 12 columns in this dataset. They are:

- country (string): names of different countries

- year (integer): year of data in this row

- sex (string): the sex of the person who committed suicide

- age (string): the age range of the person who committed suicide

- suicides_no (integer): number of suicides that fit this line of information
- population (integer): number of population that fit this line of information
- suicides100k pop (double): suicide rate(suicides_no/population)
- country-year (string): record the country's name and year together
- HDI for year (double): this year's Human Development Index
- gdp_for_year ($) (string): the GDP for the country
- gdp_per_capita ($) (integer): per capita GDP
- generation (string): which generation belongs to

One of the purposes of this project is to analyze the future trend of suicide rates, so we need to explore the distribution of suicide rates.

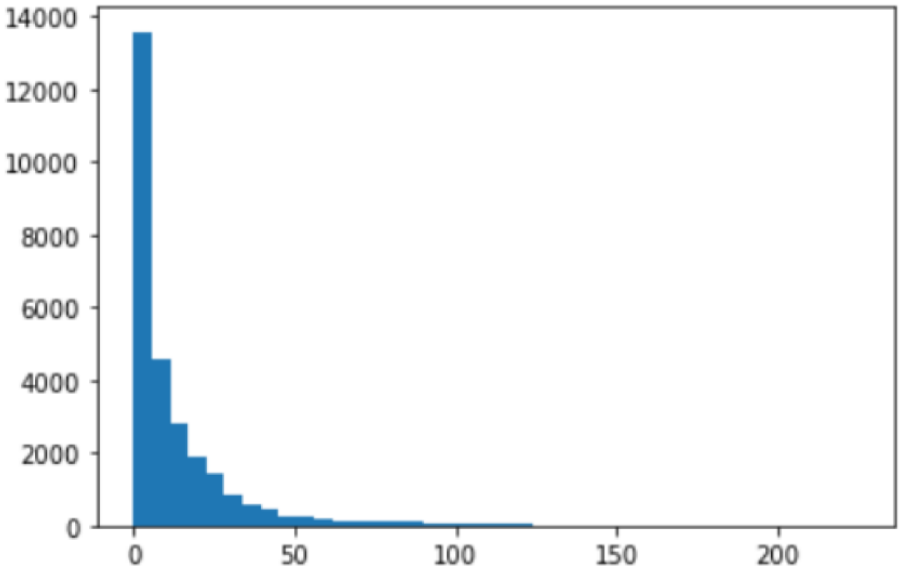

Figure 6.2

The Figure shows the histogram of suicide rate. We can get it is not like normal distribution, so we need to consider to transform it before we creating model.

Because we need to find the relationship between factors and suicide rate, so we choose to plot the relationship between age, gender, country and suicide rate in this step. We combined the five figures in one which has shown in Figure Figure 6.3.

From the figure, we can roughly see that the suicide rate is obviously related to these factors. From the figure, we can roughly see that the suicide rate is obviously related to these factors. Especially age, gender, gdp. The higher the age, the higher the suicide rate, and the higher the suicide rate of male sex. The higher the gdp, the lower the suicide rate. These are just our preliminary estimates. More specific inferences require the model to be used for data analysis.

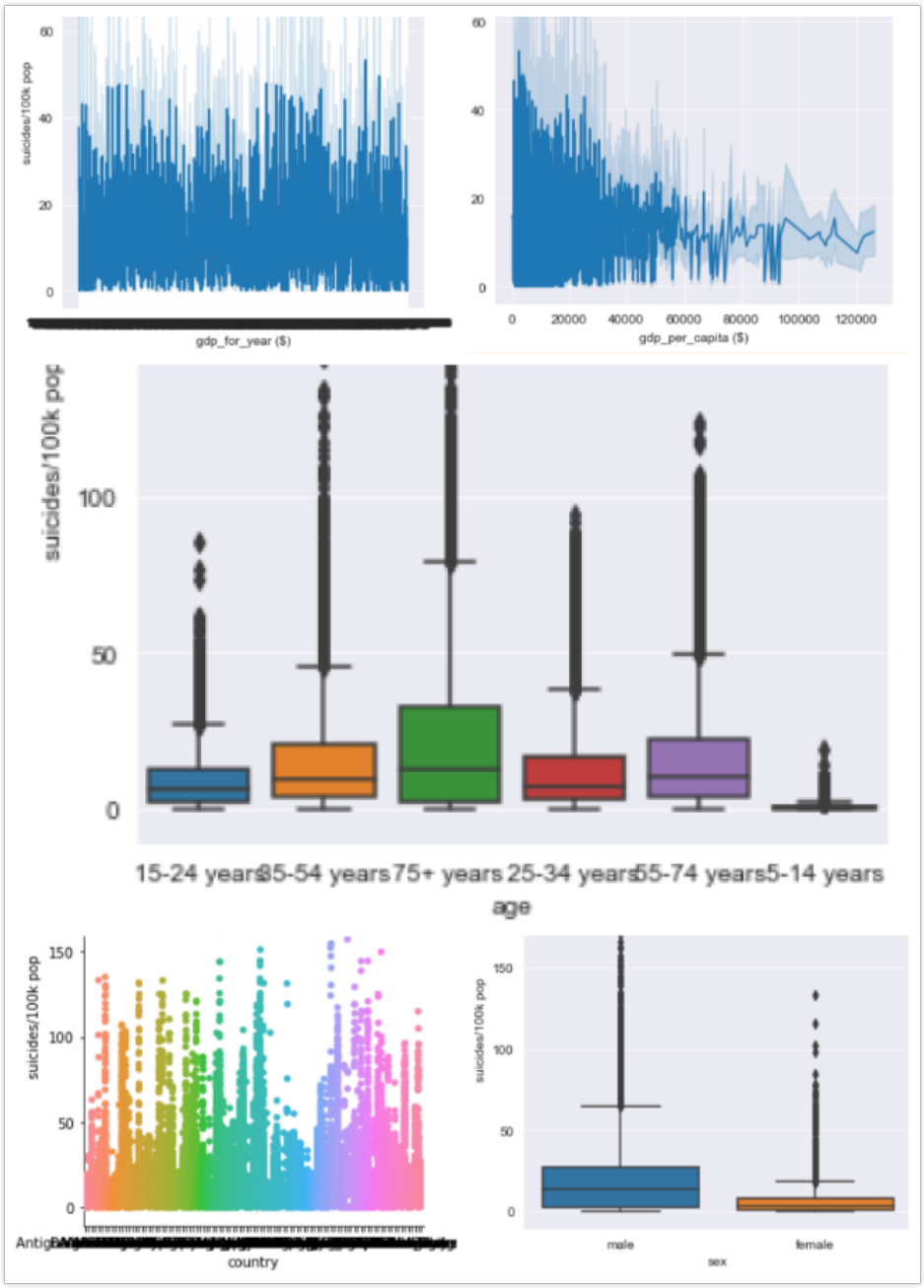
suicides/100k pop
gdp_for_year ($)
gdp_per_capita ($)
suicides/100k pop
15-24 years
35-54 years
75+ years
25-34 years
55-74 years
5-14 years
age
suicides/100k pop
country
suicides/100k pop
male
female
sex

Figure 6.3

6.3 Verify the Data Quality

Data's quality is very important to data mining. It can affect the accuracy of the model. Therefore, we should verify the data quality before data mining. We should check missing, outliers and inappropriate values.

- Missing data:

```
In [63]: datas.isnull().apply(pd.value_counts)
Out[63]:
       country     year      sex      age  suicides_no  population  \
False  27820.0  27820.0  27820.0  27820.0      27820.0     27820.0
True       NaN      NaN      NaN      NaN          NaN         NaN

       suicides/100k pop  country-year  HDI for year   gdp_for_year ($)   \
False            27820.0       27820.0          8364            27820.0
True                 NaN           NaN         19456                NaN

       gdp_per_capita ($)  generation
False             27820.0     27820.0
True                  NaN         NaN
```

Figure 6.4

Through Figure 6.4, we can see that most fields have no miss values. Most of them are 100% completed. However, it can be found that there are null values in the "HDI for year" field, which is a value describing the human happiness index. It can be found in Figure 2.14 the data in the "HDI for year" field is about only 30% completed. Because there are too many missing values in "HDI for year" field, the data in this filed cannot be effective for our prediction, so we are going to delete this column in data cleaning state in this project.

- Outliers and extreme values:

  I used the method named IQR(interquartile range). IQR is a measure of statistical dispersion. In IQR, all the values are arranged from small to large and divided into four equal parts. The calculation formula of the quartile distance is IQR = Q3-Q1. In this project, we determined if the values are greater than Q3+1.5*(Q3-Q1) or less than Q1-1.5*(Q3-Q1), these values are outliers and extremes. After this, we sum up the amount of outliers and extremes, which has clearly listed in Figure 6.5. Outliers and extreme values exist in features suicides_no, population, suicides/100k pop, HDI for year and gdp_per_capita ($). These values may caused by data errors or measurement errors. We need to take care and transform them in following stages.

```
In [117]: datas_outliners = (datas<(Q1 - 1.5*IQR)) | (datas>(Q3 + 1.5*IQR))

In [118]: datas_outliners.sum()
Out[118]:
 gdp_for_year ($)        0
HDI for year            12
age                      0
country                  0
country-year             0
gdp_per_capita ($)    1016
generation               0
population            4180
sex                      0
suicides/100k pop     2046
suicides_no           3909
year                     0
dtype: int64
```

Figure 6.5

6.4 Select the data

In the previous step, we have a deeper understanding of the data set. We will select data according to the following goals:

1. Whether the data is meaningless.

   After careful observation of the data set, we found the field population and suicides_no are meaningless because our aim in this project is to find factors that most influence the suicide rate. The suicide rate(suicides/100k pop) is obtained by dividing the number of suicides(suicides_no) by the total number of people(population), so these two fields are meaningless for our analysis. We need to remove them.

2. Whether there are repetitive items.

   Repetitive items mean that the meaning of the fields is repetitive. For example, the field country-year is repetitive item. It combines the field country and year. It is meaningless for our prediction. Therefore in this project, we will remove the field of country-year. After selection, the dataset is shown as Figure 6.6.

```
In [122]: datas.head()
Out[122]:
   country  year     sex          age  suicides/100k pop  HDI for year  \
0  Albania  1987    male  15-24 years               6.71           NaN
1  Albania  1987    male  35-54 years               5.19           NaN
2  Albania  1987  female  15-24 years               4.83           NaN
3  Albania  1987    male    75+ years               4.59           NaN
4  Albania  1987    male  25-34 years               3.28           NaN

   gdp_for_year ($)   gdp_per_capita ($)       generation
0      2,156,624,900                 796     Generation X
1      2,156,624,900                 796           Silent
2      2,156,624,900                 796     Generation X
3      2,156,624,900                 796  G.I. Generation
4      2,156,624,900                 796          Boomers
```

Figure 6.6

6.5 Clean the data

In order to prepare for modeling, We need to clean the data, because we have found missing value and some outliers and inappropriate values in the dataset.

The most missing value is in the filed of "HDI foryear". There are 19456 rows of missing values in this column, which means this filed is only 30% completed. We decided to use mean value to instead of these missing values. After this operation, there are no missing values in the data set, now.

```
In [127]: datas.isnull().apply(pd.value_counts)
Out[127]:
       country   year    sex    age  suicides/100k pop  HDI for year  \
False    27820  27820  27820  27820              27820         27820

        gdp_for_year ($)   gdp_per_capita ($)  generation
False               27820               27820       27820
```

Figure 6.7

Considering that the number of the outliers and extreme values only accounts for a very small part of the entire data set, so we plan to use IQR to discard them. In this project, we determined if the values are greater than Q3+1.5*(Q3-Q1) or less than Q1-1.5*(Q3-Q1), these values are outliers and extremes. Therefore, we decide to filter these values. After this operation, the rest of the data is valid.

6.6 Construct the data

In this project, I would like to explore the suicide problem of the elderly, so I want to create a new column to distinguish whether the suicide population is the elderly or the young. I will set 55 as the threshold, and people younger than 55 are considered young, and people older than 55 are considered old. Therefore, after operations, we can get a new field named "old_or_young", which has shown in Figure 3.5. "1" in the new field means the elderly and "0" means the young.

```
In [150]: datas_cl.head()
Out[150]:
   country  year     sex          age  suicides/100k pop  HDI for year  \
0  Albania  1987    male  15-24 years               6.71      0.776601
1  Albania  1987    male  35-54 years               5.19      0.776601
2  Albania  1987  female  15-24 years               4.83      0.776601
3  Albania  1987    male    75+ years               4.59      0.776601
4  Albania  1987    male  25-34 years               3.28      0.776601

   gdp_for_year ($)   gdp_per_capita ($)       generation  old_or_young
0      2,156,624,900                 796     Generation X             0
1      2,156,624,900                 796           Silent             0
2      2,156,624,900                 796     Generation X             0
3      2,156,624,900                 796  G.I. Generation             1
4      2,156,624,900                 796          Boomers             0
```

Figure 6.8

6.7 Integrate various data source

In this project, we have all the data in one dataset, so we do not need to merge or append dataset in this step.

6.8 Format the data as required

In this project, our goal is to analyze the data in the dataset and draw the conclusions we want. Therefore, the data type in the data set should be transformed into a state suitable for analysis. We plan to use regression model to try to analyze the data, so the fields we need in the model should be numeric. In this data set, the format of "sex", "age", "generation", "country" and "gdp_for_year ($) " fields are string and we decide to change them to numeric through sklearn package. The preprocessing of sklearn has been imported. We use 0 to instead "female" and 1 to instead "male" inthe field of sex. In the "age" field, we set "5-14 years" as 0, "15-24 years" as 1, "25-34 years" as 2, "35-54 years" as 3, "55-74 years" as 4, "75+ years" as 5. In the "Generation" field, we will use 0 to instead "Boomers", 1 to instead "G.I. Generation", 2 to instead "Generation X", 3 to instead "Generation Z", 4 to instead "Millennials", 5 to instead "Silent Generation". Next, we change the type of "country" field. In the "country" field, there are 101 countries in this dataset and we use 1–101 to instead these countries(in alphabetical order). lastly, in the "gdp_for_year ($) " field, the original data is of string type separated by commas, so I use code to changed it to be integer. After this operation, the data is shown as:

```
In [223]: datas_cl.head()
Out[223]:
   country  year  sex  age  suicides/100k pop  HDI for year  \
0        0  1987    1    0               6.71      0.776601
1        0  1987    1    2               5.19      0.776601
2        0  1987    0    0               4.83      0.776601
3        0  1987    1    5               4.59      0.776601
4        0  1987    1    1               3.28      0.776601

    gdp_for_year ($)   gdp_per_capita ($)  generation  old_or_young
0         2156624900                  796           2             0
1         2156624900                  796           5             0
2         2156624900                  796           2             0
3         2156624900                  796           1             1
4         2156624900                  796           0             0
```

Figure 6.9

6.9 Reduce the data

In the previous steps, when we explored the dataset, we found the field of "HDI for year" is only 30% completed. Although we have dealt with this problem with random numbers, we still decided to delete this column considering that the percentage of missing is too large. We also need to reduce the filed which is meaningless. It can be seen the three columns of suicides_no, population and suicides/100k pop. In this study, we mainly study what factors affect the suicide rate. The suicides/100k pop is

got by suicides_no divide by population. Therefore, we should remove the two factors(population and suicides_no). The field country-year is repetitive item. It combines the field country and year. It is meaningless for our prediction, so we need to reduce them. We have reduced the three fields in Data Selection. Here we are just reiterating why we should reduce these three items. To sum up, in this step, we decide to reduce the fields "HDI for year". After reducing the unimportant features, there are remain nine columns as input data.

```
In [231]: del datas_cl["HDI for year"]

In [232]: datas_cl.head()
Out[232]:
   country  year  sex  age  suicides/100k pop   gdp_for_year ($)   \
0        0  1987    1    0               6.71         2156624900
1        0  1987    1    2               5.19         2156624900
2        0  1987    0    0               4.83         2156624900
3        0  1987    1    5               4.59         2156624900
4        0  1987    1    1               3.28         2156624900

   gdp_per_capita ($)  generation  old_or_young
0                 796           2             0
1                 796           5             0
2                 796           2             0
3                 796           1             1
4                 796           0             0
```

Figure 6.10

6.10 Project the data

In data exploration, we have plot the histogram of suicide rate. The distribution is not a normal distribution, which is inconvenient for us to use regression model. Therefore, we decide to project the data by statistical transformations (the log of a distribution). In order to conduct this process, first we need to remove the number of zero in "suicides/100k pop" field, because zero can not be logged. After transformation, we plot the suicide rate data again. The distribution of it is close to normal distribution now.

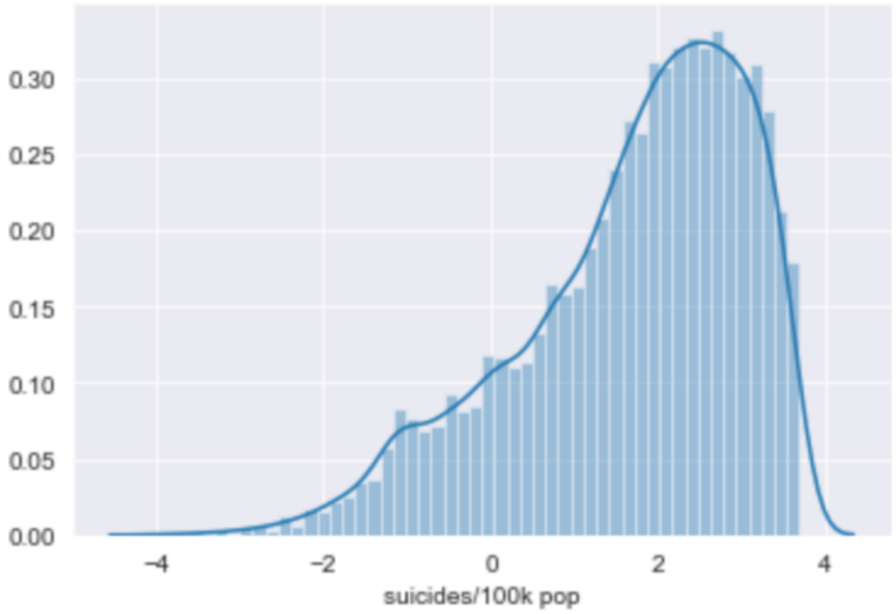

Figure 6.11

7. Algorithms and Technologies of Implementation

7.1 Match and discuss the objectives of data mining to data mining methods.

We have two objectives in this study.

- Analyze the future trend of suicide rate.
- Find out the relationship between the suicide rate and different factors and identify the factors that most influence the suicide rate.

According to the above objectives, our data mining objectives is:

- Analyze the historical data of the world suicide rate through different data analysis models to predict the future trend of suicide rate. Find the lowest error according to the performance of different models, and the best performing model to predict the future development trend of suicide rate.
- Determine the important factors that affect the suicide rate by analyzing the relation-ship between different factors in the data set and the suicide rate, and rank them according to the degree of impact. Finally, we give the visual effect of data analysis to provide support and suggestions to relevant departments more clearly and intuitively.

The methods of data mining can be roughly divided into two categories. They are Supervised Method and Unsupervised Machine Learning Algorithms. [3]

- Supervised Learning

  Supervised Learning is a method in machine learning, which can learn from training materials or establish a pattern (function / learning model), and infer new examples according to this model.

  There are two types of learning supervision problems, One of them is regression problem and another is classification problem.

  * Regression
    Regression method is a supervised learning algorithm for predicting and modeling numerical continuous random variables. The characteristic of regression task is that the labeled dataset has numerical target variables. Use cases generally include house price forecast, stock trend or test results and other continuous changes.
  * Classification
    Classification method is a supervised learning algorithm for modeling or predicting discrete random variables. Classification algorithms are usually used

to predict a category (or the probability of a category) rather than continuous values. Use cases include email filtering, financial fraud, and forecasting employee turnover.

The objective in this project is to predict the trend of suicide rates. The target we want to predict is a numerical variables, so based on the previous discussion, the regression method is more suitable for our project.

- Unsupervised Learning

  The feature of unsupervised learning is that the data of model learning has no label, so the goal of unsupervised learning is to reveal the inherent characteristics and laws of data by learning these unlabeled samples, and its representative is clustering.

  * Clustering
    Clustering is a method of partitioning the dataset into groups. The goal is to split up the data in such a way that points within single cluster are very similar and points in different clusters are different. It determines grouping among unlabeled data
  * Association
    An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.[3]

  Our project is to predict the trend of suicide rate and find the factors which can affect the suicide rate. Our target suicide rate has labeled in the dataset, so this method may not suitable for our prediction.

7.2 Select the appropriate data-mining method based on discussion

Based on the previous discussion, we will choose regression model of supervised learning to conduct our models. Because the objective of this is to predict the trend of suicide rate and find the factors which can affect the suicide rate. The fields in this dataset have been labeled and the target value is real value, so regression model is the most suitable one. We use the models provided by the sklearn package in Python, such as linear regression, random forest regression and KNN regression models.

7.3 Data-mining algorithms selection.

Based on the discussion before, we would like to choose Supervised regression method.We can conduct this kind of model with the pyspark.ml.regression package approach such as Random Forest, Linear Regression and KNN Regression model. These models have advantages and disadvantages so we need to analyze them.

- **Linear Regression Model**

  Linear Regression Model is a common statistical technique for classifying records based on the values of numeric input fields. Linear regression fits a straight line

or surface that minimizes the discrepancies between predicted and actual output values. [18]

Requirement: There must be one target and one input. The target items must be continuous.

Strengths: The time complexity of linear regression is relatively low compared with other data mining models. The principle of linear regression is also easy to understand. Therefore, linear regression is a very easy model to master.

- **Random Forest Model**

  Random Forest is an advanced implementation of a bagging algorithm with a tree model as the base model. In random forests, each tree in the ensemble is built from a sample drawn with replacement (for example, a bootstrap sample) from the training set. When splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features. Instead, the split that is picked is the best split among a random subset of the features. Because of this randomness, the bias of the forest usually slightly increases (with respect to the bias of a single non-random tree) but, due to averaging, its variance also decreases, usually more than compensating for the increase in bias, hence yielding an overall better model.[2] After a series of base learners are obtained, their prediction results are synthesized as the final output of the integration model.

  Requirement: The input value and target value of Random Forest model should be numeric.

  Strengths: It has great advantages in performance compared with other algorithms. A random forest can process high-dimensional data (that is, data with many features) without having to make feature selection. After the training, which characteristics can be given by the random forest are more important.

- **KNN Regression Model** K-nearest neighbor algorithm (KNN) is a supervised learning algorithm, which means that the training data set needs to have label or category. The goal of KNN is to label the unlabeled data points (samples) automatically or predict the category they belong to. KNN can also be used for regression.

  Requirement: KNN Regression Model requires the input should be numeric and the target value can be numeric or categorical value.

  Strengths: One of the biggest advantages of K-NN is that K-NN can be used both for classification and regression problems.

7.4 Select data-mining algorithms based on the discussion.

We have discussed the different data-mining algorithms in the former steps. Our data mining goals are:

- Analyze the historical data of the world suicide rate through different data analysis models to predict the future trend of suicide rate. Find the lowest error according to the performance of different models, and the best performing model to predict the future development trend of suicide rate.
- Determine the important factors that affect the suicide rate by analyzing the relation-ship between different factors in the data set and the suicide rate, and

rank them ac-cording to the degree of impact. Finally, we give the visual effect of data analysis to provide support and suggestions to relevant departments more clearly and intuitively.

Combined with the goal of data mining and the discussion in the previous section, we decide to use the following data mining algorithms:

- Linear Regression Model

  Firstly, we decide to use Linear Regression Model to generate our model. Linear Regression Model is a straightforward and easy-understanding model so we wiil try it first. According to the requirements of the linear regression model listed in the previous discussion, our data set meets this requirement, so we can use Linear Regression Model which is included in pyspark package to predict the trend of suicide rates directly.

- Random Forest model

  The second method we will use is Random Forest model. According to the discussion in Chapter6.1, Random Forest model is very suitable for our project. It can be used to find the fields which affect the suicide rates.

- KNN Regression Model

  Finally, we will useKNN Regression Model. We wil use the output of this model to compare with the other model. Then, conclude the best model and result of our project.

7.5 Build/Select appropriate model(s) and choose relevant parameters.

1. Linear Regression Model

   Now, we have cleaned and projected the dataset. It's time to build models. I build the model by Linear Regression from sklearn package.

   The parameters we used:

   The Linear Regression Model in this project is not complicated, so we do not need special setting. We decided to use default settings in this model.

   - fit_intercept, default=True
     Whether to calculate the intercept for this model. If set to False, no intercept will be used in calculations.
   - normalize, default=False
     This parameter is ignored when fit_intercept is set to False. If True, the regressor X will be normalized before regression by subtracting the mean and dividing by the l2-norm.[7]
   - copy_X, default=True
     If True, X will be copied; else, it may be overwritten.

2. Random Forest model

   We use Random Forest model as the second one to build our model. The input value and target value of Random Forest model should be numeric. Our input and target values meet the requirements. The parameters we used:

- maxDepth = 10
  The maximum depth of the tree. If we set it as a large number, the model will be overfit. If we set it as a small number, the model will be underfit. Therefore, we set it to be 10. It should be a suitable number.[7]
- numTrees = 10
  The number of trees in the forest. We set up 10 trees for this model, because the more trees there are, the higher the accuracy. The more trees there are, the more time it takes. Therefore, we think 10 is a suitable number in our project.
- maxBins = 101
  The DecisionTree algorithm requires maxBins (= 32) to be at least as large as the number of values in each categorical feature, but categorical feature 7 in our data has 101 values. Therefore, we set maxBins = 101 to fit the model.
- minInfoGain
  Minimum information gain for a split to be considered at a tree node.

3. KNN Regression Model We use KNN Regression Model in last step. KNN is K-nearest neighbor algorithm. It can predict the nearest neighbors of the target value.

   The parameters we used: In the parameters of this model, we mostly use the default settings.

   - n_neighborsint: default=5
     Number of neighbors to use by default for kneighbors queries.
   - leaf_size : int, default=30
     Leaf size passed to BallTree or KDTree. This can affect the speed of the construction and query.[7]
   - p: default=2
     Power parameter for the Minkowski metric. When p = 1, this is equivalent to using manhattan_distance (l1), and euclidean_distance (l2) for p = 2.
   - algorithm: default='auto'
     'auto' will attempt to decide the most appropriate algorithm based on the values passed to fit method.

7.6 Create and justify test design

If we use the same data set to fit and test the model, our tests will be meaningless because the model is built on the same data set. We're going to get a pretty good result from this test but it doesn't mean anything. Therefore, we need to split the data into training and testing sets. In machine learning, we generally need to divide the sample into two separate training sets (train set) and test set (test set). The training set is used to construct the model, and the test set is used to test the accuracy of the model. For example, we can divide the data set into training set and data set according to 80% and 20%.

In this project, we will use Pareto principle to divide our dataset, which we will use 80% to be train set and 20% to be test set.[11] This operation has be done by pyspark package in our project.

7.7 Conduct data mining – classfy,regress,cluster,etc.

We have split the data. We have 80% data is training data and 20% data is test data. We will conduct data mining with these data. After running the model, we will use RMSE(root mean squared error) and R-squared to help us to evaluate these model. I have built and run these models by sklearn package and calculate the MSE (The average squared difference between the estimated values and the actual value) and R-square (is a statistical measure of how close the data are to the fitted regression line) of these models.

7.7.1 Linear Regression Model

We will use the default parameters to run the model and then use the model to predict in the test data. After running the model, we check MSE and R-square. The R-square is 0.45785836147432724 and the MSE is 0.9865890031121202. The lower the MSE, the better the model is, the higher the R-square(close to 1), the better the model is. In this model, MSE seems to be relatively high and R-square is relatively low. In order to check the fitting effect of the model more intuitively, we draw the residual plot.

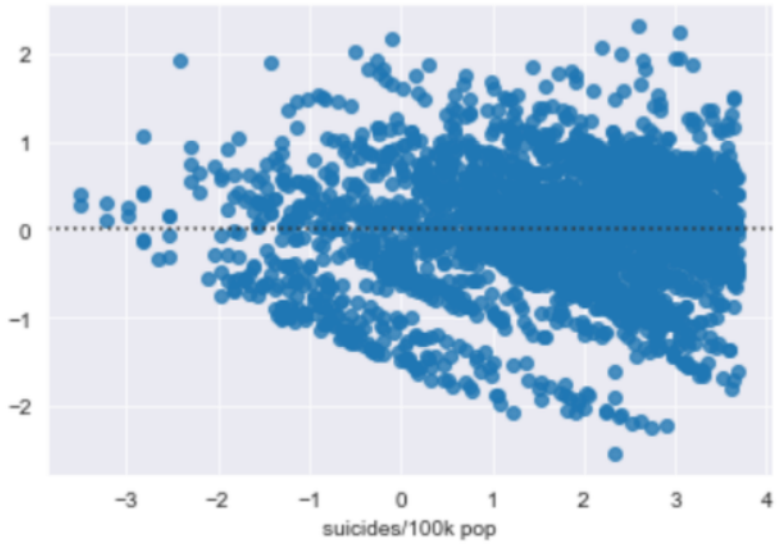

Figure 7.1

If the model fits well, the residuals should change up and down 0, and thechanges will be in a fixed range, and there will be no increase or decrease inthe range of changes. However, there is an increase in the range of changes in the residual plot of Linear Regression Model. According to the RMSE, MSE and R-square of the model, the performance of the model seems to be unsatisfactory.

7.7.2 Random Forest Model

In this step, we used random forest model to fit the model and then use the model to predict in the test data. After running the model, we got the R-square is 0.91574 and the MSE is 0.160315. The higher the R-square(close to 1), the better the model is and the lower the MSE (closer to 0), the better the model is. In this model, MSE is low and the R-square is high, which means the performance of this model is very good.

In order to check the fitting effect of the model more intuitively, we draw the residual plot. As we can see, the residuals change up and down 0, and the changes are almost in a fixed range. Combined with the MSE, and R-square of the model, the performance of this model is fairly good.

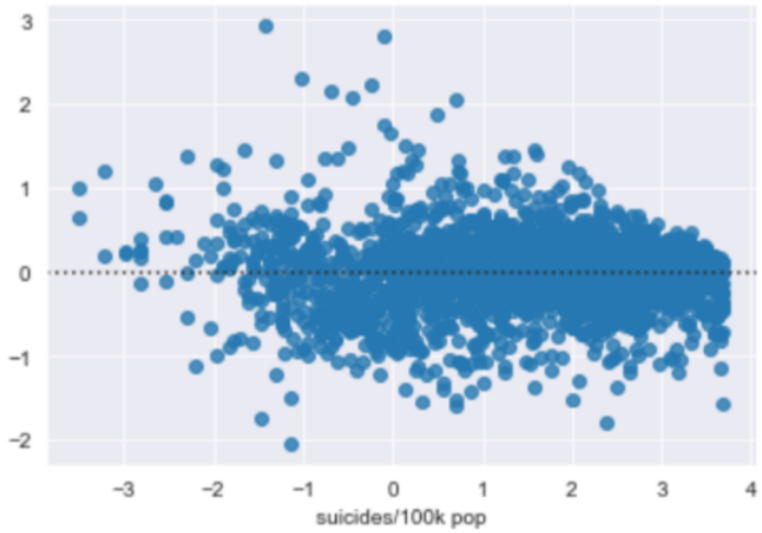

Figure 7.2

7.7.3 KNN Regression Model

The last model is KNN Regression Model, which we have introduced in the previous chapter. We fitted the model and printed the MSE and R-square of it. The MSE of KNN Regression Model is 2.08 and R-square is 0.10. It seems to have a poor performance. In order to check the fitting effect of the model more intuitively, we draw the residual plot.

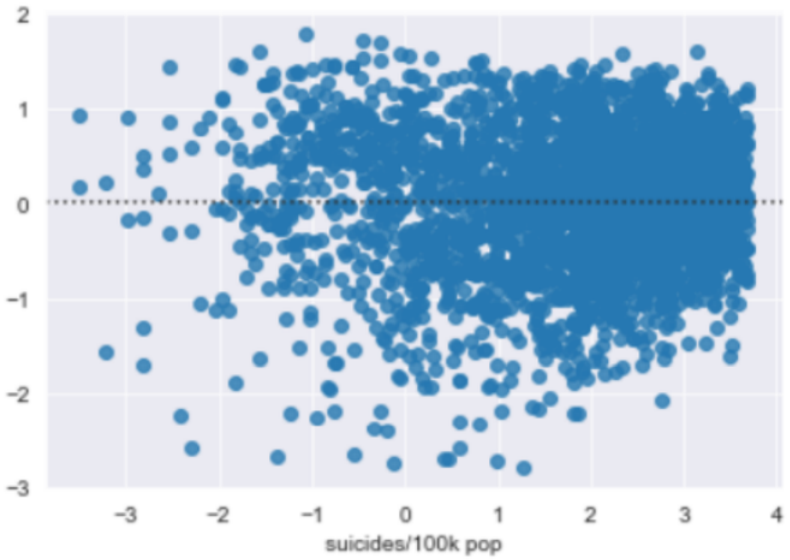

Figure 7.3

The residual plot does not look very good, too. The changes of residuals is not in a fixed range and the range seems to be relatively large in the residual plot. Combined with the MSE and R-square of the model, it should not be a

good model fit.

7.7.4 Assessment

After comparing the performance of the three models(Linear Regression Model, Random Forest Model, KNN Regression Model),we find that the performance of Random Forest Model is the most outstanding. It has the lowest MSE and the highest R-square in these three models.

7.8 Search for patterns

After the discussion in the previous section, we list the performance of the various models which is shown in below.

Table 7.1

Algorithm	**R-square**	**MSE**
Linear Regression	0.457	0.986
Random Forest	0.915	0.160
KNN Regression Model	2.08	0.10

From the table, we can see, Random Forest model has the lowest MSE. It also has highest R-square. In the previous steps, we have explained that the lower the MSE (closer to 0), the better the model is, and the higher the R-square(close to 1), the better the model is. Therefore, in these three models, Random Forest model has the best performance. It shows a high explanation and low mean squared error.

Now, let's list the output of Random Forest model. It can be concluded that the field of age is the most important field which can influence suicide rate in this model and the importance of country, gdp_for_year, gdp_for_capita and sex followed close behind.

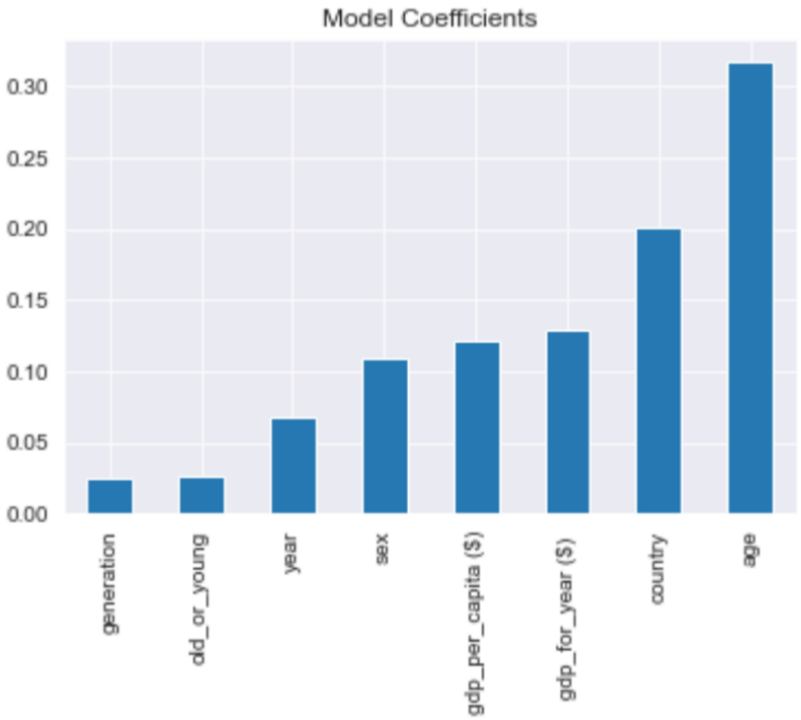
Model Coefficients
0.30
0.25
0.20
0.15
0.10
0.05
0.00
generation
old_or_young
year
sex
gdp_per_capita ($)
gdp_for_year ($)
country
age

Figure 7.4

8. Interpretation of Patterns and Results

The three models has been built and we have reported the outcome of these models. Now, we need to valuate the results and interpret the patterns. The MSE and R-square can be used to evaluate the models' performance.

8.1 Study and discuss the mined patterns.

The data we used in this project is from kaggle which is a open-source website. Firstly, we collect initial data and choose the useful data fields. Then we clean the data which wasn't appropriate. We discarded the extremes and transformed the missing values. After that, We construct a new data named "old_or_young" in the dataset, because we want to explore whether the factor of young or old has an effect on suicide rate. Then, we format the data, we changed the category data into numeric data in order to better fit the model. In the next step, we found the distribution of our target field "suicides/100k pop" is not normal distribution, so we use log to transform it. Because we want to fit model to predict the trend of suicide and find the factors which can influence the suicide rate, so we list some different models and compare them and try to find the best one. We choose three different models to fit and list their performance(R-square and MSE of the models). They are Linear Regression Model, Random Forest model and KNN Regression Model. After comparison, we found Random Forest model performs the best.

Therefore, we decide to discard the other two models and use Random Forest model to predict the trend of suicide rate and use this model to find the important factors which can affect suicide rate. We list the patterns we found:

- Random Forest model has a higher R-square than other models and the MAE(mean absolute error) is the lowest. As we have explained before, a higher R-square means the better the model's performance is. The lowest MSE means the difference between true value and predict value is low.
- In order of importance field to the suicide rate, age, country, sex,generation,old_or_you gdp and year. The most important field is age. In order to discover how

they affect the suicide rate, the best way is to plot them first and then interpret the patterns.

8.2 Visualize the data, results, models, and patterns.

8.2.1 Visualize the data

- The histogram of target value(log_suicide_rate)

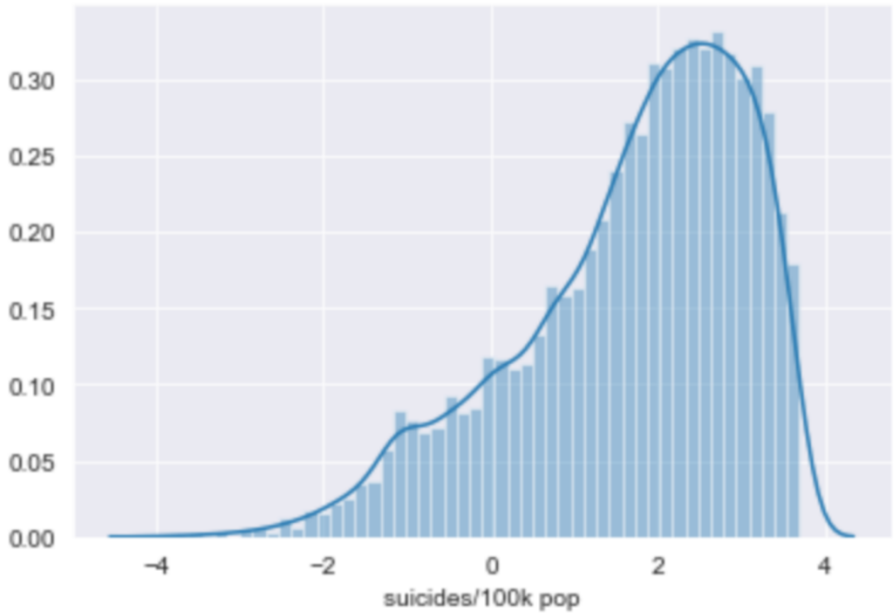

Figure 8.1

- The relationship between each field and the target value

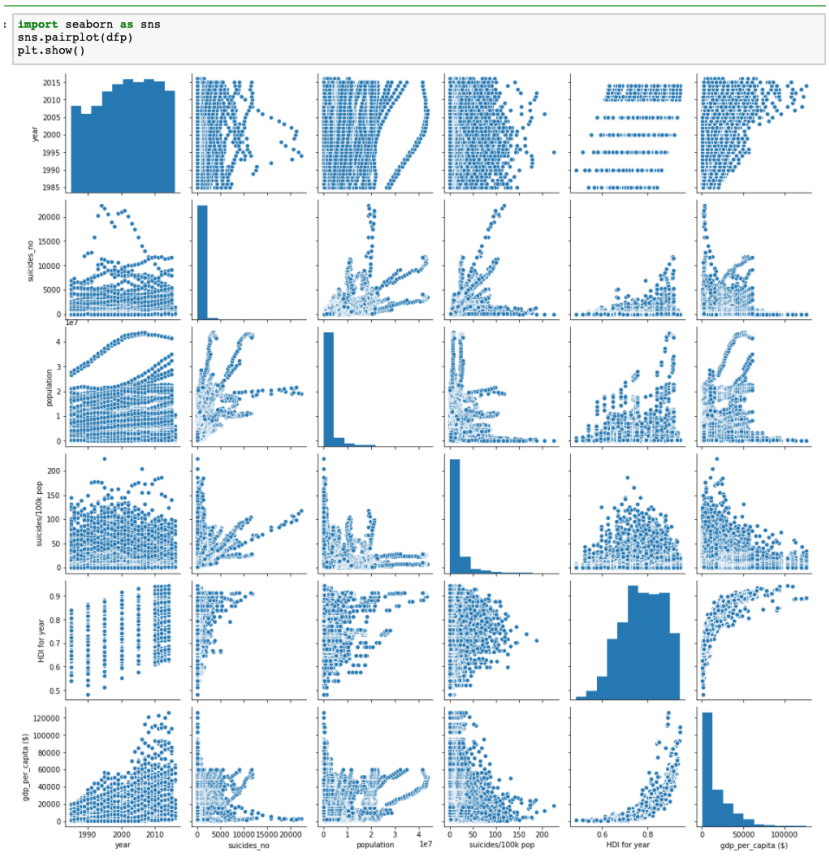

Figure 8.2: The overall plot of correlation coefficient

8.2.2 Visualize the models and result

- Linear Regression Model

  we plot the output of the Linear Regression Model. In Figure 8.3, we plot the coefficients of every fields which can also used to represent the field's importance. In Figure 8.8, we plot the residuals plot of this model.

```
In [390]: framelinear.sort_values(by='coef', ascending=False)
Out[390]:
                fields          coef
7        old_or_young  3.047370e+00
2                 sex  9.283038e-01
1                year  4.849060e-03
0             country  1.921904e-03
5  gdp_per_capita ($)  1.054765e-05
4   gdp_for_year ($)  -2.265120e-14
6          generation -2.415111e-01
3                 age -5.995057e-01
```

Figure 8.3

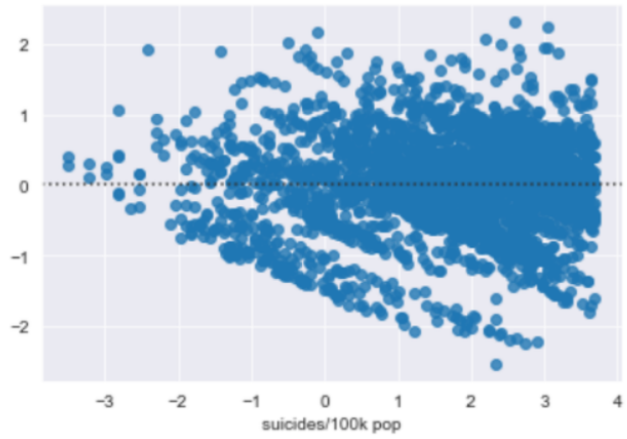

Figure 8.4

- Random Forest Model

  In this section, we will plot the output of the Random Forest Model. In Figure 8.5, we plot the importance of every fields. In Figure 8.6, we plot the residuals plot of this model.

```
In [377]: frame.sort_values(by='importances', ascending=False)
Out[377]:
                fields  importances
3                  age     0.316990
0              country     0.201751
4   gdp_for_year ($)       0.129884
5  gdp_per_capita ($)      0.121429
2                  sex     0.108581
1                 year     0.068536
7         old_or_young     0.027139
6           generation     0.025690
```

Figure 8.5

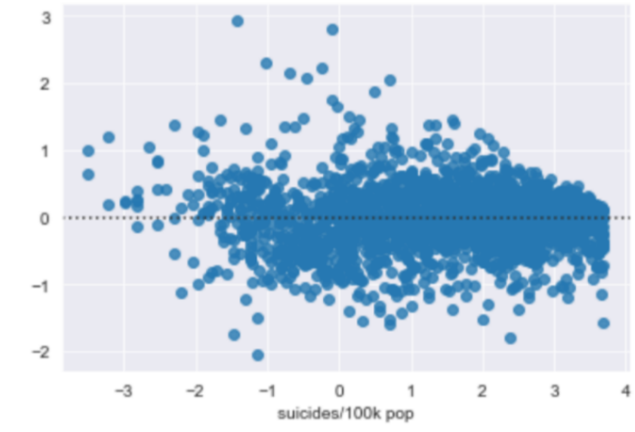

Figure 8.6

- KNN Regression Model

  In this section, we plot the output of the KNN Regression Model. Then, we plot the residuals plot of this model.

	Fields	Importance
0	year	0.0756
1	gdp_for_year ($)	0.0212
2	gdp_per_capita ($)	0.0159
3	old_or_young	0.0137
4	sex_new	0.0865
5	age_new	0.1512
6	generation_new	0.0106
7	country ID	0.6253

Figure 8.7

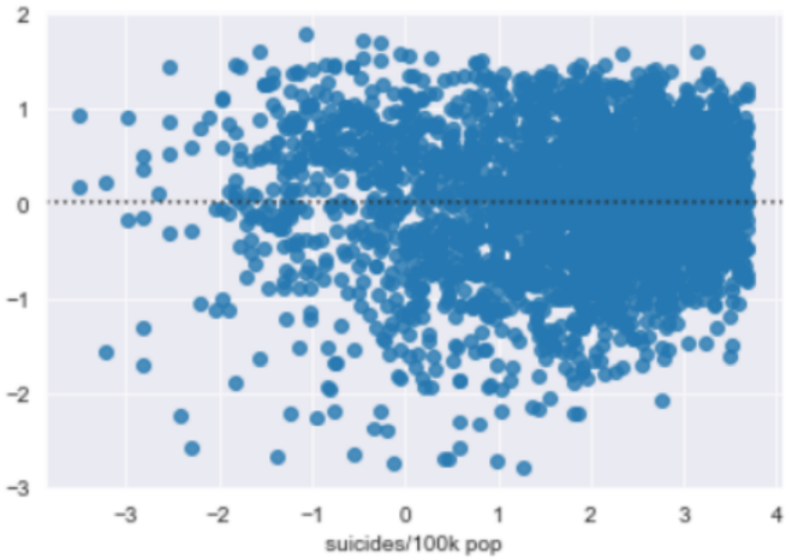

Figure 8.8

We have shown the MSE and R-square before. The MSE and R-square of Linear Regression Model are 0.986 and 0.457. The MSE and R-square of Random Forest Model are 0.160 and 0.915. The MSE and R-square of Random Forest Model are 2.08 and 0.10. After comparing the performance of the three models(Linear Regression Model, Random Forest Model, KNN Regression Model), we find that the performance of Random Forest Model is the most outstanding. It has the lowest MSE and the highest R-square in these three models. Therefore, we decided to plot the first five important factors and target value(suicide rate) based on the output of the random forest model. From the previous discussion, we can know, in random forest model, the field of age is the most important field which can influence suicide rate in this model and the importance of country, gdp_for_year, gdp_for_capita and sex followed close behind. We will plot their relationship with the target value(suicide rate) in order of importance.

8.2.3 Visualize the patterns

- The relationship between age and suicide rate. The higher the age, the higher the suicide rate.

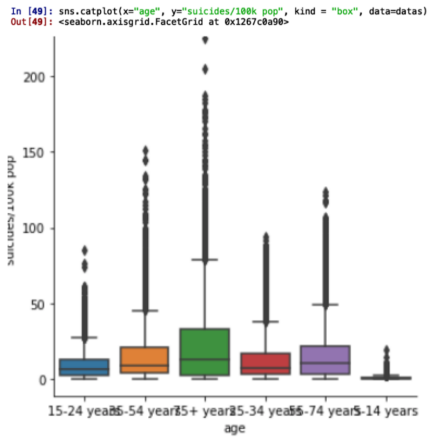

Figure 8.9

- The relationship between country and suicide rate. Suicide rates vary significantly in different countries.

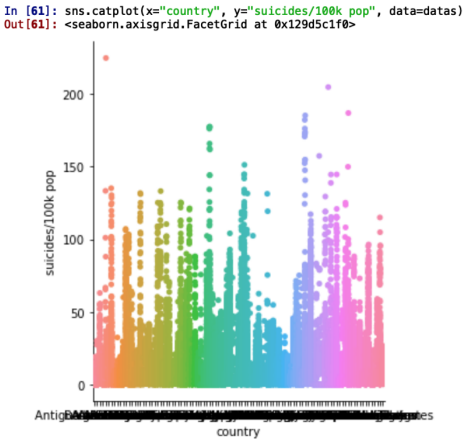

Figure 8.10

- The relationship between gdp_for_year and suicide rate. Suicide rates vary significantly in different gdp situation.

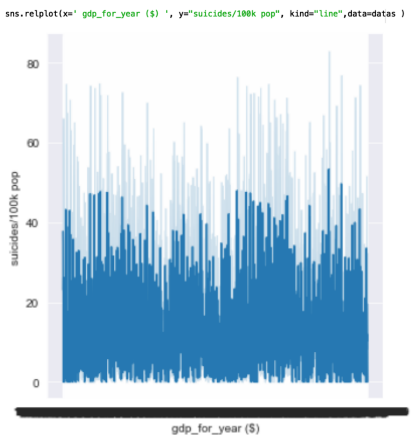

Figure 8.11

- The relationship gdp_for_capita and suicide rate. The higher the gdp_for_capita, the lower the suicide rate.

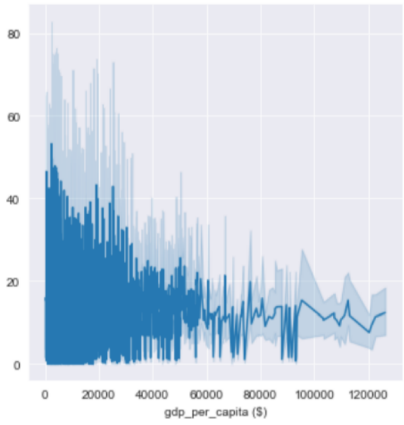

Figure 8.12

- The relationship between sex and suicide rate. The suicide rate of men is higher than that of women.

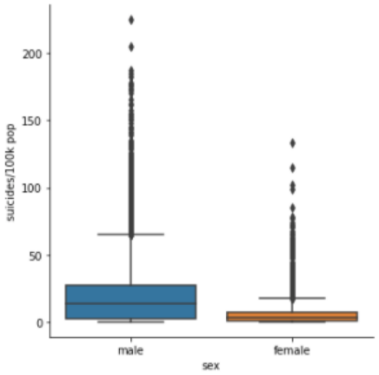

Figure 8.13

8.3 Interpret the results, models, and patterns.

We have used three different model. Through the performance of these models in Table 8.1, we can get the performance of Random Forest Model is the best. It has the highest R-square and lowest MSE. The second best performer is Linear Regression Model and the last is KNN Regression Model.

Table 8.1

Algorithm	**R-square**	**MSE**
Linear Regression	0.457	0.986
Random Forest	0.915	0.160
KNN Regression Model	2.08	0.10

Then we need to interpret the results and patterns of Random Forest Model which is the most suitable model in these models. Let's list the output of the random forest model again.

The R-square of Random Forest Model is 0.91574. R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. The closer the value of a is to 1, the better the mosel is. The R-square of Random Forest Model is 0.91574 which is close to 1, indicating that the the approximately 91.574% of the observed variation can be explained by the model's inputs.

The MSE (mean squared error) of Random Forest Model is 0.160315. MSE (mean squared error) is the error between the actual value and predicted value, and the smaller the value the better the model's performance. 0.160315 is a

small number which is not far from 0, so we can conclude the difference between the value predicted by Random Forest Model and actual value is not large.

We also give the residual plot of Random Forest Model in Figure 8.6. If the model fits well, the residuals should change up and down 0, and the changes will be in a fixed range, and there will be no increase or decrease in the range of changes. From the figure, we can see the residuals change up and down 0, and the changes are almost in a fixed range, which shows the model fits well. Combined with the R-square and MSE of the model, the performance of this model is fairly good. Next, we plot the importance of every fields. Based on the output in Figure 8.14. From the figure, we can know, in random forest model, the field of age is the most important field which can influence suicide rate in this model and the importance of country, gdp and sex followed close behind.

```
In [377]: frame.sort_values(by='importances', ascending=False)
Out[377]:
                fields  importances
3                  age     0.316990
0              country     0.201751
4   gdp_for_year ($)       0.129884
5  gdp_per_capita ($)      0.121429
2                  sex     0.108581
1                 year     0.068536
7        old_or_young      0.027139
6          generation      0.025690
```

Figure 8.14

We have plotted the relationship between age, gender, country and suicide rate in last step.

From the relationship between age and suicide rate, we can clearly see that the suicide rate of young people who is less than 14-year-old is very low, while the suicide rate of the elderly over 55 years old is relatively high.

From the relationship between country and suicide rate, we can clearly see that that suicide rates vary greatly in different countries.

From the relationship between sex and suicide rate, we can see that the suicide rate of men was significantly higher than that of women.

From the relationship between gdp_for_capita and suicide rate, we can see the level of gdp has an impact on the suicide rate, which is more obvious in the relationship between gdp per capita and suicide rate. The lower the gdp_for_capita, the higher the suicide rate.

Based on above patterns discovered, we think We believe that older men with lower GDP have the highest suicide rate, especially in some countries with high suicide rate. We plot the boxplot of age, gender and suicide rate

in Figure 8.15, which proved our guess. The suicide rate of male elderly is relatively high.

```
sns.boxplot(x="age", y="suicides/100k pop", hue="sex", data=dfp);
```

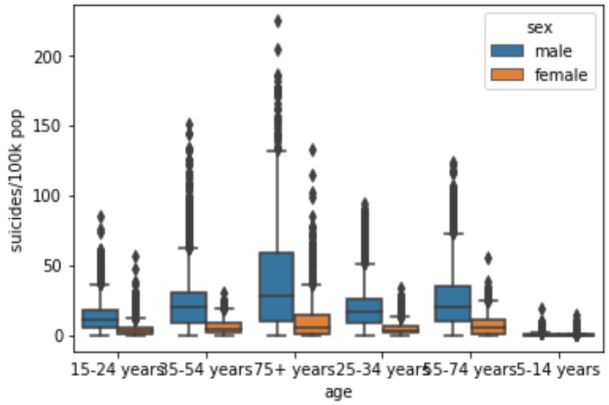

Figure 8.15

8.4 Assess and evaluate results, models, and patterns

- Assess and evaluate results

  We have discussed the result in the previous steps and we concluded Age, country, gdp and sex can have a great influence on suicide rates.

  - The higher the age, the higher the suicide rate.
  - Suicide rates vary significantly in different countries and different generation.
  - The lower GDP, the higher suicide rate.
  - The suicide rate of men is higher than that of women.

- Assessment for Prediction of model

  We have known the performance of Random Forest model is the best in the three models. The MSE of it is 0.160 and R-square is 0.915. The higher the R-square(close to 1), the better the model is and the lower the MSE (closer to 0), the better the model is. In this model, MSE is low and the R-square is high, which means the performance of this model is very good.

Although the performance of Random Forest model is very good, we still need to list more to prove the model is suitable to our project. The regression model has four evaluation indexes. They are MSE, R-square, Mean Absolute Error(MAE) and explained_variance_score. In the previous steps, we have list the MSE and R-square of Random Forest model. Both of them show that the model is performed good. Now, we decide to check the other two evaluation indexes(Mean Absolute Error(MAE) and explained_variance_score). We use RegressionEvaluator to calculate Mean Absolute Error(MAE) and explained_variance_score.

```
: evaluator4 = RegressionEvaluator(labelCol="suicides/100k pop", predictionCol="prediction", metricName="var")
  varrf = evaluator2.evaluate(predictions_rf)
  print("Explained variance of Random Forest Regression Model = %g" % varrf)
  Explained variance of Random Forest Regression Model = 0.916782

: evaluator5 = RegressionEvaluator(labelCol="suicides/100k pop", predictionCol="prediction", metricName="mae")
  maerf = evaluator5.evaluate(predictions_rf)
  print("MAE of Random Forest Regression Model = %g" % maerf)
  MAE of Random Forest Regression Model = 0.285655
```

Figure 8.16

As we can see, Mean Absolute Error(MAE) is 0.285655 Mean Absolute Error (MAE), used to assess the degree to which the predicted results are close to the real data set. The smaller the value is, the better the fitting effect is. 0.285655 is close to 0, which means the degree between the predicted results are close to the real data set.

The explained_variance_score of Random Forest model is 0.916782. Explained_variance_score explains the variance score of the regression model, whose value range is [0,1]. The closer to 1, the more independent variables can explain dependent variables. The smaller the value is, the worse the effect is. 0.916782 is very close to 0.916782, which means the model is very good. Based on the previous discussion, we can conclude Random Forest model is suitable for our project. Therefore, we can use Random Forest model to predict the trend of suicide rate.

- Assess and evaluate of patterns

  The importance fields of Random Forest model shows that age, country,gdp and sex are important factors which can affect the suicide rate. Therefore, we should advice the relevant agencies to pay attention of these features in order to prevent the loss of lives.

9. Proposed actions based on the discovered knowledge

9.1 Apply the knowledge and deploy the implementation

According to the previous chapter of Methodology, our research objectives can be reached by the Random Forest Model due to its well performance.

To be Specific, we can use this Random Forest Model to predict the future development trend of suicide rate and determine the important factors that affect the suicide rate by analyzing the relationship between different factors in the data set and the suicide rate, and rank them according to the degree of impact. Finally, we can apply these information in real life, we can give suggestions to relevant departments more clearly and intuitively.

In order to achieve the purpose of reducing the suicide rate, the prediction of the suicide rate trend generated by the random forest model can be provided to the government and related organizations. This is to warn them to raise awareness of helping suicides and strengthen the protection of high-risk suicide areas, such as studying the installation of guardrails in areas such as bridges and railroad tracks, and showing warm-hearted films in suicide green high-risk areas, etc.

We can also provide the government and related organizations with the important factors affecting suicide rate and the relationship between these factors and suicide rate generated by the random forest model. In order to help them make better plans to help potential suicides to achieve the purpose of reducing the suicide rate. For example, in the previous chapter, we predicted that the suicide rate of elderly men is relatively high. We can suggest that the community carry out warming activities, and the main target population is positioned as elderly men. Through frequent communication with them, heart-to-heart talk, and practical solutions to their life and mental problems can reduce the probability of their suicide.

9.2 Monitor the implementation

After the implementation of suicide rate prediction, the processing track should be monitored. The suicide data from World Health Organization should be checked every year. If the data is updated or in an abnormal state, we will not get accurate output.

Another thing should be concerned is to test model. When new suicide data are available, the errors of the model should be checked regularly. As the amount of data increases, the "random forest" model may increase the error, which will not be the best model for the project. It is a good choice to change the model with higher accuracy or adjust the parameters.

9.3 Maintain the implementation

At the same time, we should consider the running time consumed by the model. When the training data is updated, we should actively test the model with different parameters in time to find the most accurate and the least time-consuming parameter. In this way, we can continuously improve the accuracy of the model and reduce the running time and cost.

9.4 Enhance the solution in the future

Our current data volume is still not large enough, and there may be sample selection bias. In the future, more data should be used to study models and predict the trend of suicide rates, which will be more accurate and effective. At the same time, our data set does not contain enough factors that affect the suicide rate. We did not consider family, culture, education and environment background. These factors will also be important reasons for the suicide rate. Therefore, future models should consider factors such as family, culture, education, and environment to improve the accuracy of our predictions.

In this study, we mainly considered analyzing the data through several regression models. Through evaluation, the results are not bad. However, there may be other models that perform better than the random forest model to predict suicide rates, such as artificial neural networks, clustering models, and time series models. We should test more models

Finally, we can use Spark Streaming to process real-time data so that results can be obtained faster.

References

[1] Alexandra Brazinova. Suicide rate trends in the slovak republic in 1993–2015. *SAGE Journals*, 2017.

[2] L. Breiman. "random forests,". *Machine Learning*, pages 5–32, 2001.

[3] J. Brownlee. Supervised and unsupervised machine learning algorithms. *Retrieved from machinelearning mastery*, 2019.

[4] David Gunnella Caroline Coopea. Suicide and the 2008 economic recession: Who is most at risk? trends in suicide rates in england and wales 2001–2011. *Science Direct Journals*, 2014.

[5] David Gunnella Caroline Coopea. Suicide and the 2008 economic recession: Who is most at risk? trends in suicide rates in england and wales 2001–2011. *Science Direct Journals*, 2014.

[6] India State-Level Disease Burden Initiative Suicide Collaborators. Gender differentials and state variations in suicide deaths in india: the global burden of disease study 1990–2016. *Public Health*, 2018.

[7] G. V. F Pedregosa. Machine learning in python. *Journal of machine learning research*, 2011.

[8] Heeringen K Hawton K. The international book of suicide and attempted suicide. *John Wiley Sons, Ltd*, 2020.

[9] Yasuyuki ; Inoue, Ken ; Fujita. A long-term study of the association between the relative poverty rate and suicide rate in japan. *Journal of Forensic Sciences.*, 2016.

[10] DPHIL KA-YUET LIU. Suicide rates in the world: 1950-2004. *Suicide and Life-Threatening Behavior*, 2009.

[11] Lipovetsky. Pareto 80/20 law: derivation via random partitioning. *International Journal of Mathematical Education in Science and Technology*, 2009.

[12] Agnus M.Kim. Factors associated with the suicide rates in korea. *Psychiatry Research*, 2020.

[13] Alastair H ; Mok, Pearl L H ; Leyland. Why does scotland have a higher suicide rate than england? an area-level investigation of health and social factors. *Journal of Epidemiology and Community Health*, 2013.

[14] Wafaa M.Abdel Moneim. Suicide rate: Trends and implications in upper egypt. *Science Direct Journals*, 2011.

[15] World Health Organization(WHO). Figure and facts about suicide. *Department of Mental Health*, 2008.

[16] World Health Organization(WHO). Figure and facts about suicide. *Department of Mental Health*, 2012.

[17] Ajit Shah. Suicide rates: age-associated trends and their correlates. *Journal of Injury and Violence Research*, 2012.

[18] Chao YanmingYang, LuluSun. Aero-material consumption prediction based on linear regression mode. *Internationa Congress of Information and Communication Technology*, 2018.

[19] Eric D. Yip, Paul S F; Caine. Employment status and suicide: the complex relationships between changing unemployment rates and death rates. *Journal of Epidemiology and Community Health*, 2011.