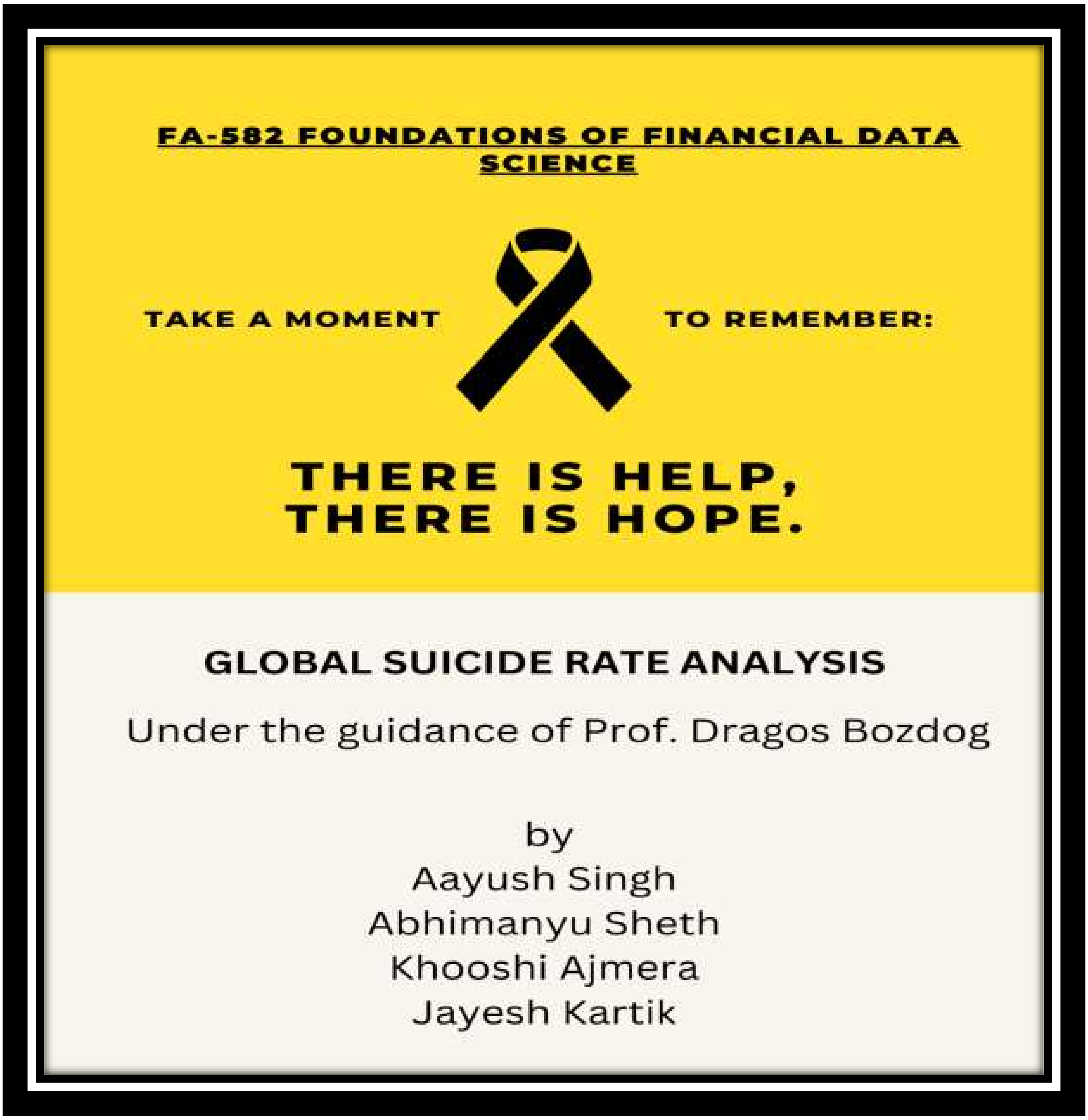


FA-582 FOUNDATIONS OF FINANCIAL DATA SCIENCE

TAKE A MOMENT TO REMEMBER:

THERE IS HELP, THERE IS HOPE.

GLOBAL SUICIDE RATE ANALYSIS

Under the guidance of Prof. Dragos Bozdog

by
Aayush Singh
Abhimanyu Sheth
Khooshi Ajmera
Jayesh Kartik

Table of Contents

Section1-Introduction

Every day, we come across the news that an individual committed suicide due to mental pressure and depression. This is a global problem now and is one of the most debated and heated topics. We wanted to understand how this phenomenon works on a global level, what the reasons behind it are, and how it can be reduced, hence we decided to study this sensitive topic.

The Covid-19 pandemic changed the lifestyle of almost everyone. Life before and after the pandemic was totally different. Though people's tendency to do suicide before the pandemic was also significant but after covid-19, the problem of mental health was more talk of the town. Because of everything reaching our homes and the new hybrid work culture, everyone in their lives got centered on a screen in front of them. Long story short, Covid-19 acted as a catalyst that has worsened the whole situation.

In recent years, due to digitalization in almost every sector, the personal touch, and communication have almost gone and at times we don't find time to even connect with our closest friends and family members.

Threats or comments about killing themselves
Social withdrawal from friends, family, community
Increased alcohol and drug use
Dramatic mood swings
Aggressive behavior
Talking, writing, or thinking about death
Impulsive or reckless behavior

According to WHO-

- Globally 800,000 people commit suicide that's twice the number of people dying due to homicide.
- Suicide is one of the leading causes of death among young people.
- Globally the suicide rate for men is twice as much as that of women.
- The suicide rate due to firearms is particularly high in the USA- almost 60% of USA suicides are because of firearms.
- In low to middle-class families, self-poisoning is the leading way to commit suicide and a ban on some of these pesticides has been effective in reducing the suicide rates in these countries.

Section2-Abstract

Right now, we are considering people suffering from mental issues such as anxiety, depression, and loneliness as our target audience. Though we are not limiting our research to such a particular domain, the major research work of the project focuses on the suicidal tendency of an individual before and after Covid-19. The target audience can also consist of persons having a lot of struggles and hardships in their lives just to make sure they are meeting family needs.

It also includes mental institutions that are working on reducing such problems and educational institutions who has a lot of young students who are expected to perform well.

Some of the important focus points which we are trying to accomplish are as follows:

- Determining the major situations under which people think of suicide.
- Exploring the countries having the highest suicidal rates.
- Segmenting the suicidal rates among different age groups
- Identifying which gender needs more help to avoid these suicidal trends.
- Using regression analysis to identify the number of suicides happening in every country over time and establishing a linear trend of whether the country has an increasing/decreasing trend of suicides.

The current scenario of the world gives a lot of emphasis on the mental health of an individual.

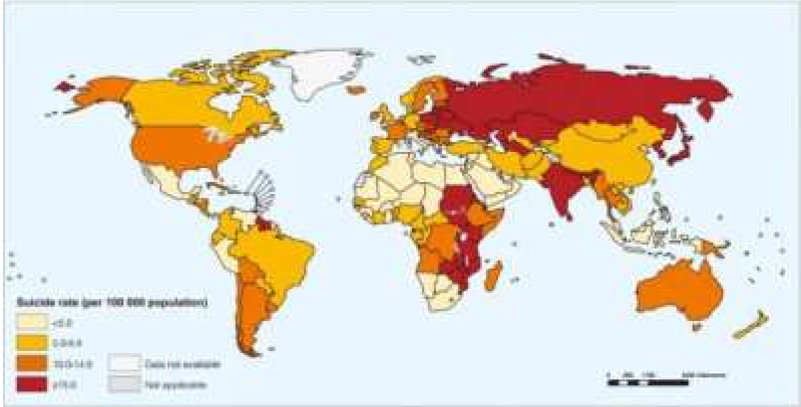
With World Mental Health Day celebrated just around the corner, and everything coming back to normal after Covid-19, people tend to be suffering a lot in their personal or professional lives.

While on the broader aspect, we don't prefer to talk openly about these things, but deep down inside these things boil up to depression, anxiety, and other mental illnesses. Upon reaching the threshold these problems become unbearable and the only way out possible in front of an individual is to give up.

This research work contains global data on suicide rates and our aim is to deeply analyze this data to provide some relevant information.

We are very much interested to find some relevant information on the suicidal rate current trend based on the GDP and population of a country. Though it's very evident that population is one of the most crucial factors when analyzing the data of a country, the GDP of a country will provide us with an idea about the standard of living of a country.

On the dataset, we have done exploratory data analysis and modeling to get the best results. We have considered the GDP and population of the countries as the main variables for the analysis.



Section3-Data and Methods

1. Data collection

- A CSV file containing the data has been attached with the submission- https://drive.google.com/file/d/17qs41jSXhhdPeesz1_hokXxap-nFqBRa/view?usp=drivesdk

2. Source

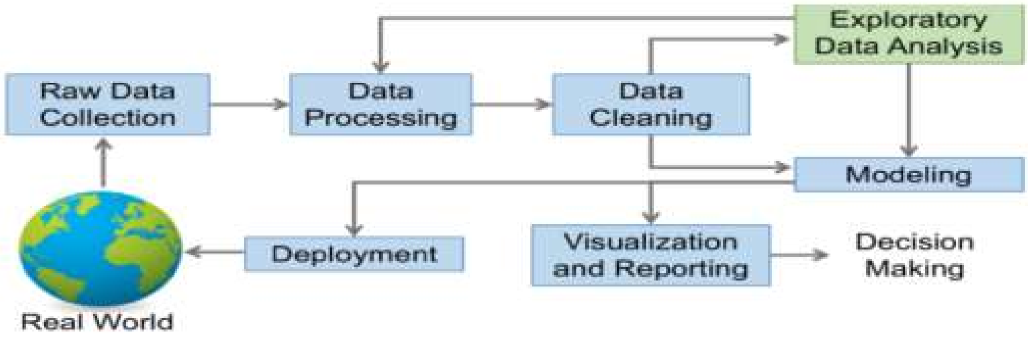
- Link – https://www.who.int/news-room/factsheets/detail/suicide#:~:text=More%20than%20700%20000%20people,15%2D29%20year%2Dolds.

- Bloomberg

Section 4- Methods, Results and Conclusions

Exploratory data analysis

Exploratory Data Analysis has been done in R using the following methodology-



- **RAW DATA COLLECTION**

The first step for every data science project is data collection that is, getting the actual raw data. Raw data is a set of information that was delivered from a certain data entity to the data provider and hasn't been processed yet by machines or humans. This information is gathered from online sources to deliver deep insight into users' online behavior.

- **DATA PROCESSING**

Data processing occurs when data is collected and translated into usable information. Usually performed by a data scientist or team of data scientists, it is important for data processing to be done correctly so as not to negatively affect the product or data output.

Data processing starts with data in its raw form and converts it into a more readable format (graphs, documents, etc.), giving it the form and context necessary to be interpreted by computers and utilized by employees throughout an organization.

- **DATA CLEANING**

Data cleansing or data cleaning is the process of identifying and removing (or correcting) inaccurate records from a dataset, table, or database and refers to recognizing unfinished, unreliable, inaccurate, or non-relevant parts of the data and then restoring, remodeling, or removing the dirty or crude data. Data cleaning may be performed as batch processing through scripting or interactively with data wrangling tools. After cleaning, a dataset should be uniform with other related datasets in the operation. The discrepancies identified or eliminated may have been basically caused by user entry mistakes, corruption in storage or transmission, or by various data dictionary descriptions of similar items in various stores

WHY DATA CLEANING?

It is more important for any organization to have the right data as compared to a large data set. Data cleansing solutions can have several problems during the process of data scrubbing. The company needs to understand the various problems and figure out how to tackle them. Some of the key data cleaning problems and solutions include –

- Data is never static-It is important that the data cleansing process arranges the data so that it is easily accessible to everyone who needs it. The warehouse should contain unified data and not in a scattered manner. The data warehouse must have a documented system that is helpful for the employees to easily access the data from different sources. Data cleaning also further helps to improve the data quality by removing inaccurate data as well as corrupt and duplicate entries.

- Incorrect data may lead to bad decisions-While operating your business you rely on certain sources of data, based on which you make most of your business decisions. If the data has a lot of errors, the decisions you take may be incorrect and prove to be hazardous for your business. The way you collect data and how your data warehouse functions can easily have an impact on your productivity.

- Incorrect data may affect district records-Complete client records are only possible when the names and addresses match. Names and addresses of the client can be poor sources of data. To avoid these mistakes, companies should provide external references which can verify the data, supplement data points, and correct any inconsistencies.

- Developing a data cleaning framework in advance-Data cleansing can be a time-consuming and expensive job for your company. Once the data is cleaned it needs to be stored in a secure location. The staff should keep a complete log of the entire process to ascertain which data went through which process. If a data scrubbing framework is not created in advance, the entire process can become repetitive.

- Big data can bring in bigger problems-Big data needs regular cleansing to maintain its effectiveness. It requires complex computer data analysis of semi-structured or structured and voluminous data. Data cleansing helps in extracting information from such a big set of data and coming up with some data which can be used to make certain key business decisions.

DATA MODELLING

Data Modelling is the process of analyzing the data objects and their relationship to the other objects. It is used to analyze the data requirements that are required for business processes. The data models are created for the data to be stored in a database. The Data Model's focus is on what data is needed and how we have to organize data rather than what operations we have to perform.

Data Model is basically an architect's building plan. It is a process of documenting complex software system design as in a diagram that can be easily understood. The diagram will be created using text and symbols to represent how the data will flow. It is also known as the blueprint for constructing new software or re-engineering any application.

WHY DATA MODELLING?

- The data model makes sure that all the data objects required by the database are represented or not.
- The database at the logical, physical, and conceptual levels can be designed with the help data model.
- The Tools help in the improvement of data quality. Data Model gives a clear picture of business requirements.
- Redundant data and missing data can be identified with the help of data models. In data models, all the important data is accurately represented.
- The chances of incorrect results and faulty reports decreased as the data model reduces data omission.
- The data models create a visual representation of the data. With its help of it, the data analysis gets improved. We get the data picture, which can then be used by developers to create a physical database.
- Better consistency can be qualified with the help of a data model across all the projects. The model is quite time-consuming, but it makes maintenance cheaper and faster.

VISUALIZATION AND REPORTING

Data storytelling goes hand in hand with visual communication and the visual representation of data. When clients need to communicate complex data clearly and intuitively to a targeted audience, we design sleek statistical graphs, charts, information graphics, and more. Automated reporting offers an effective, time-saving layout that can be used over and over. Data visualization can effectively improve audience understanding, resulting in successful campaigning and ROI. You can have all the information in the world, but it means nothing if you can’t make sense of it.

Data visualization is useful for data cleaning, exploring data structure, detecting outliers and unusual groups, identifying trends and clusters, spotting local patterns, evaluating modeling output, and presenting results. It is essential for exploratory data analysis and data mining to check data quality and to help analysts become familiar with the structure and features of the data before them. This is a part of data analysis that is underplayed in textbooks, yet ever-present in actual investigations.

Famous sayings have a way of developing a life of their own. A picture is not a substitute for a thousand words; it needs a thousand words (or more). For data visualization you need to know the context, the source of the data, how and why they were collected, whether more could be collected, the reasons for drawing the displays, and how people with the necessary background knowledge advise they might be interpreted.

DEPLOYMENT

The concept of deployment in data science refers to the application of a model for prediction using new data. Building a model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data science process. In many cases, it will be the customer, not the data analyst, who will carry out the deployment steps. For example, a credit card company may want to deploy a trained model or set of models (e.g., neural networks, meta-learner) to quickly identify transactions, which have a high probability of being fraudulent. However, even if the analyst will not carry out the deployment effort it is important for the customer to understand up front what actions will need to be carried out to actually make use of the created models.

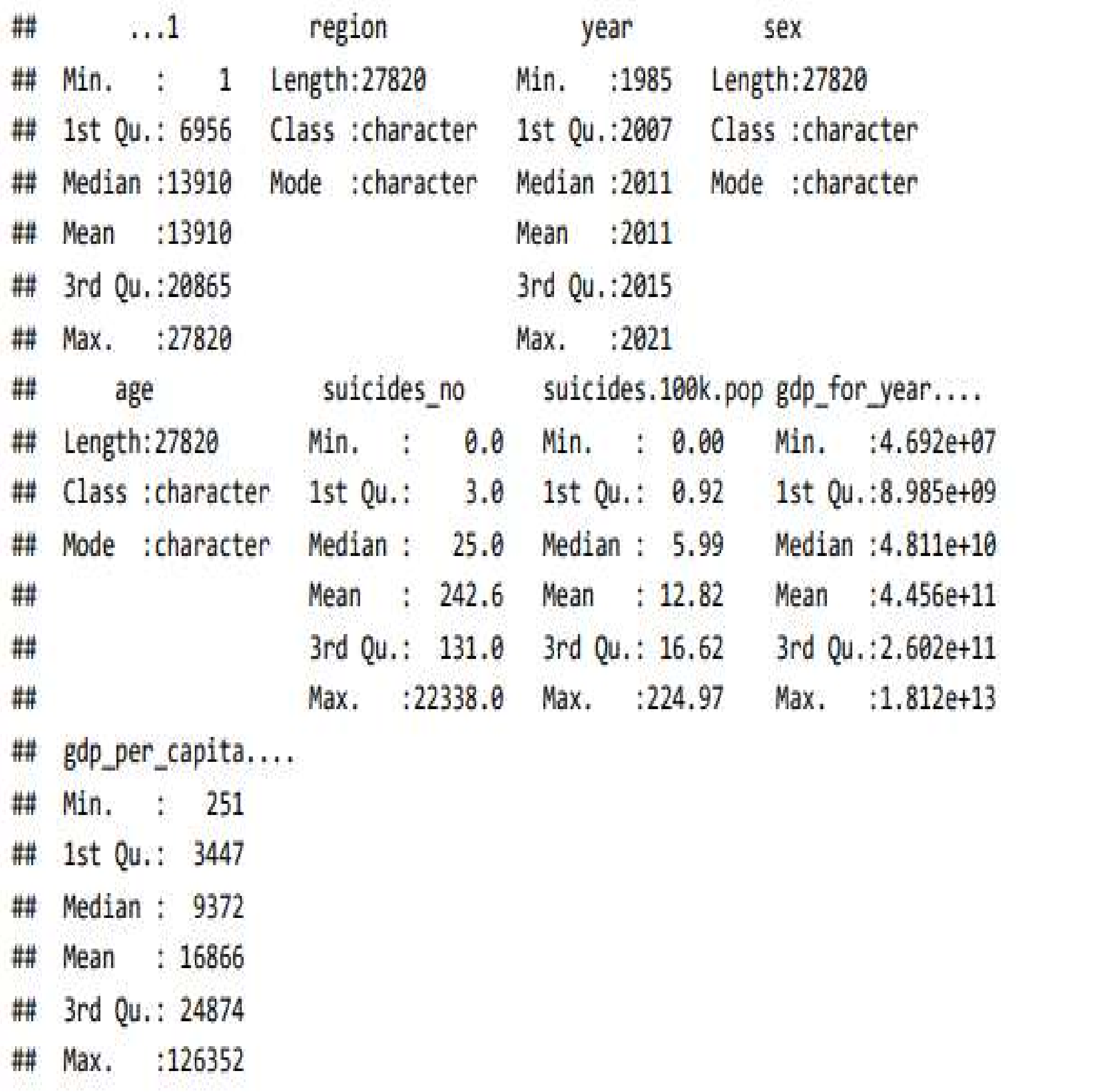
Model deployment is the process of putting machine learning models into production. This makes the model's predictions available to users, developers or systems, so they can make business decisions based on data, interact with their application (like recognize a face in an image) and so on.

Model deployment is considered to be a challenging stage for data scientists. This is because it is often not considered their core responsibility, and due to the technological and mindset differences between model development and training and the organizational tech stack, like versioning, testing and scaling which make deployment difficult. These organizational and technological silos can be overcome with the right model deployment frameworks, tools and processes.

A. Exploratory data analysis

Exploratory Data Analysis has been done in R using the following methodology-

- **Summary of the main dataset**

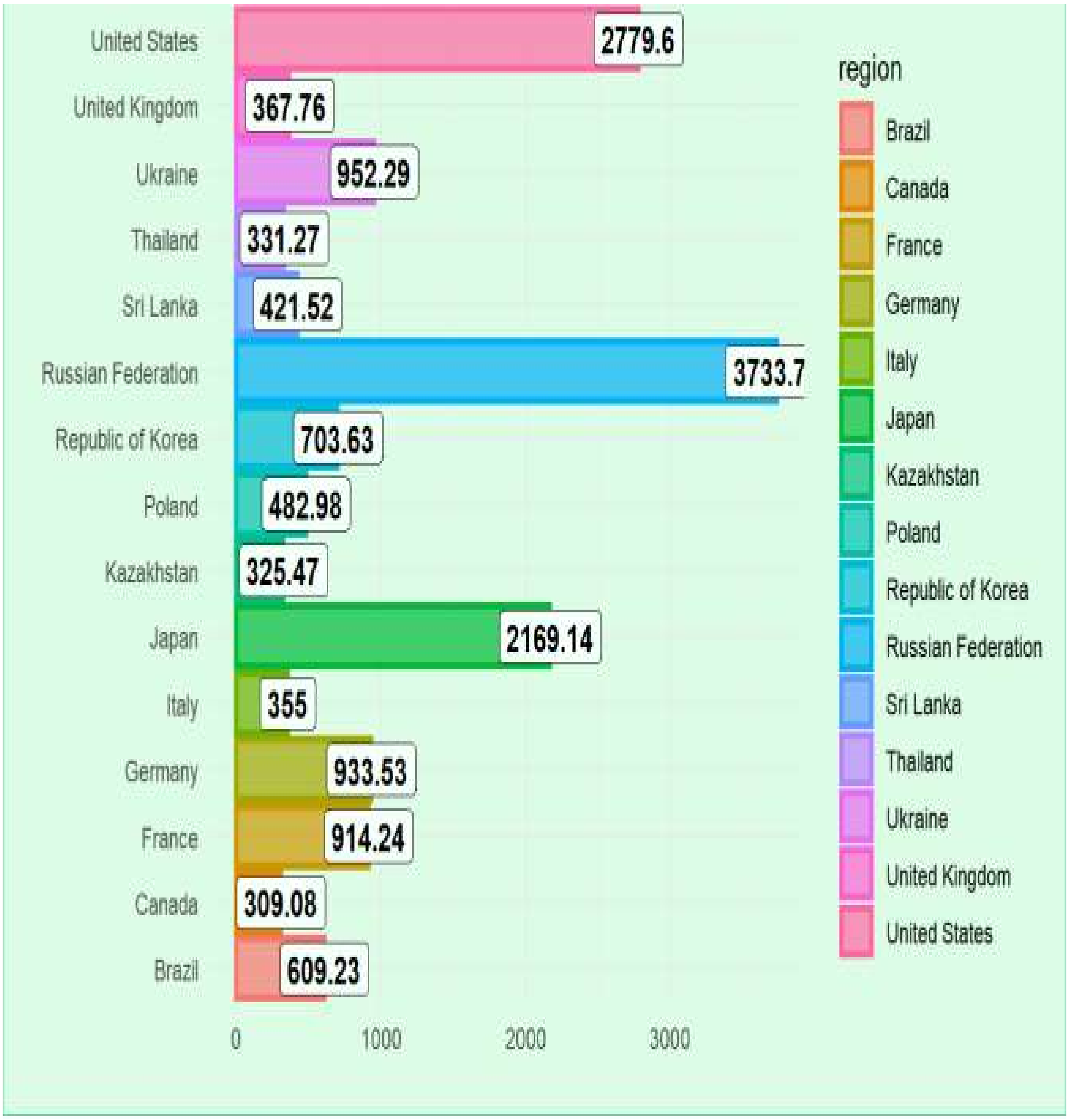


```
##       ...1          region               year           sex
##  Min.   :    1   Length:27820       Min.   :1985   Length:27820
##  1st Qu.: 6956   Class :character   1st Qu.:2007   Class :character
##  Median :13910   Mode  :character   Median :2011   Mode  :character
##  Mean   :13910                      Mean   :2011
##  3rd Qu.:20865                      3rd Qu.:2015
##  Max.   :27820                      Max.   :2021
##      age             suicides_no     suicides.100k.pop gdp_for_year....
##  Length:27820       Min.   :    0.0   Min.   :  0.00    Min.   :4.692e+07
##  Class :character   1st Qu.:    3.0   1st Qu.:  0.92    1st Qu.:8.985e+09
##  Mode  :character   Median :   25.0   Median :  5.99    Median :4.811e+10
##                     Mean   :  242.6   Mean   : 12.82    Mean   :4.456e+11
##                     3rd Qu.:  131.0   3rd Qu.: 16.62    3rd Qu.:2.602e+11
##                     Max.   :22338.0   Max.   :224.97    Max.   :1.812e+13
##  gdp_per_capita....
##  Min.   :   251
##  1st Qu.:  3447
##  Median :  9372
##  Mean   : 16866
##  3rd Qu.: 24874
##  Max.   :126352
```

Analysis:

The focus point to note here is that the highest suicide rate for Russia was 22338 for the year 2013 which clearly depicts the level of mental pressure and anxiety people are going through even though they have high GDP per capita of 2853k.

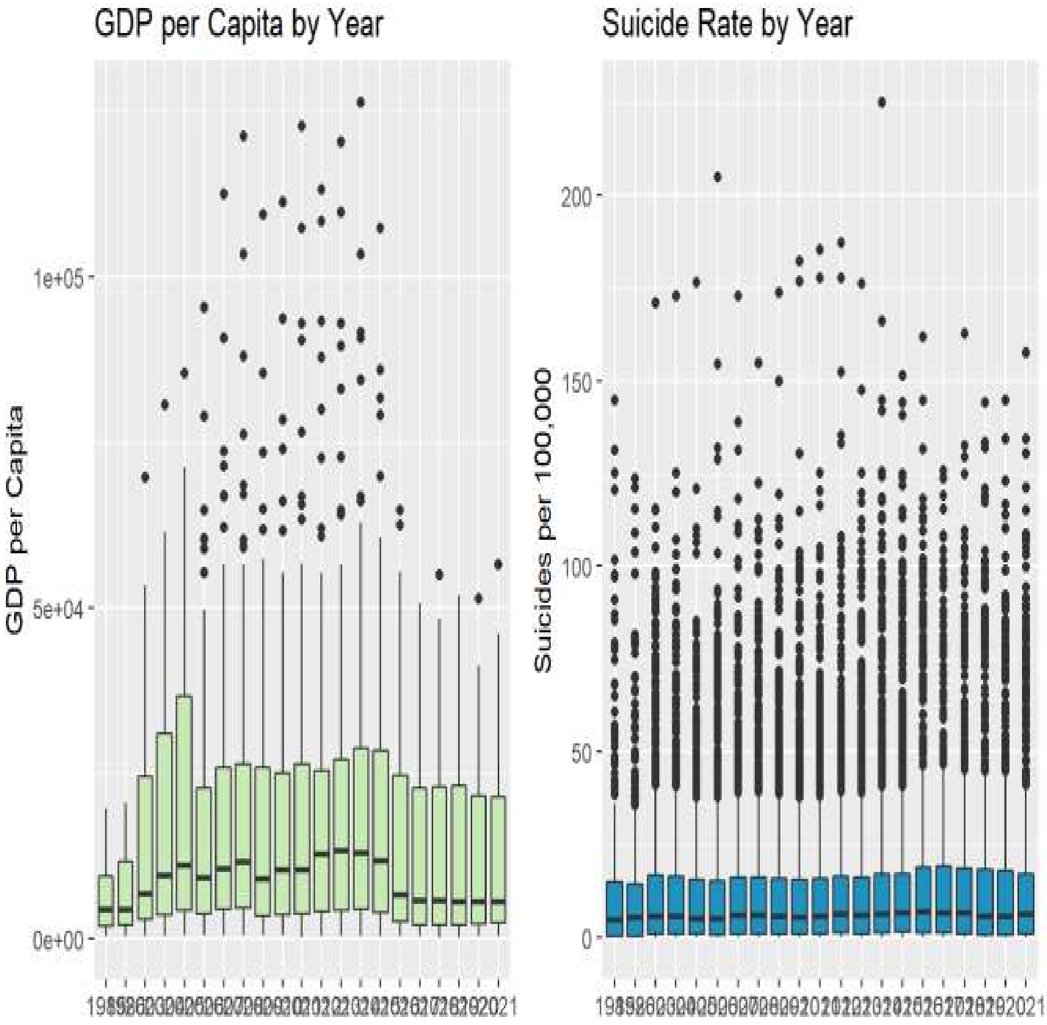
- **Countries with the highest suicide rates on an average basis**



Analysis:

It can be inferred that on an average Russia has the highest suicide rate followed by USA and Japan. In general, when it comes to reality, the standard of living is always considered better in developed countries, but these figures show that even though you have a good income and lifestyle, mental health still can be impacted in a negative way, hence it should be always given a top-notch priority.

- **Comparison between GDPs per capita by year and Suicide Rate by year globally**

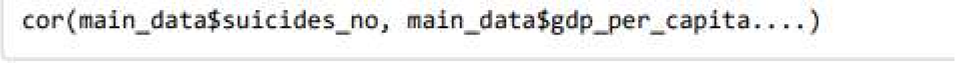


<u>Analysis:</u>

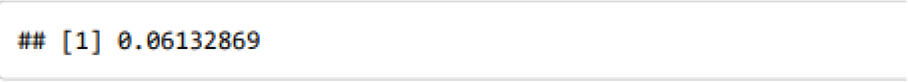
Here we notice that from 1985 till 2021, GDP per capita by year was mostly between 50000 and 100000, while suicides per 100,000 over the same duration was between 50 and 130. We can conclude that even during the years of high GDP per capita, suicides were still high.

- <u>**Checking the correlation between the following variables:**</u>

Suicide number and GDP per capita

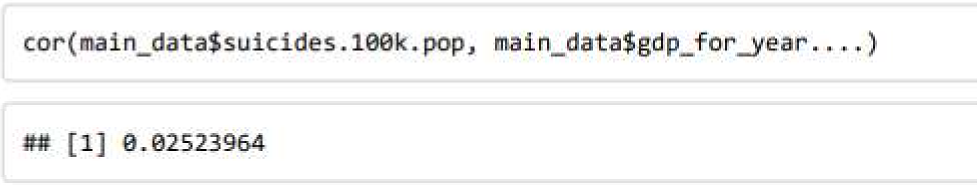


```
cor(main_data$suicides_no, main_data$gdp_per_capita....)
```



```
## [1] 0.06132869
```

Suicides.100k. pop and GDP for the year



```
cor(main_data$suicides.100k.pop, main_data$gdp_for_year....)
```

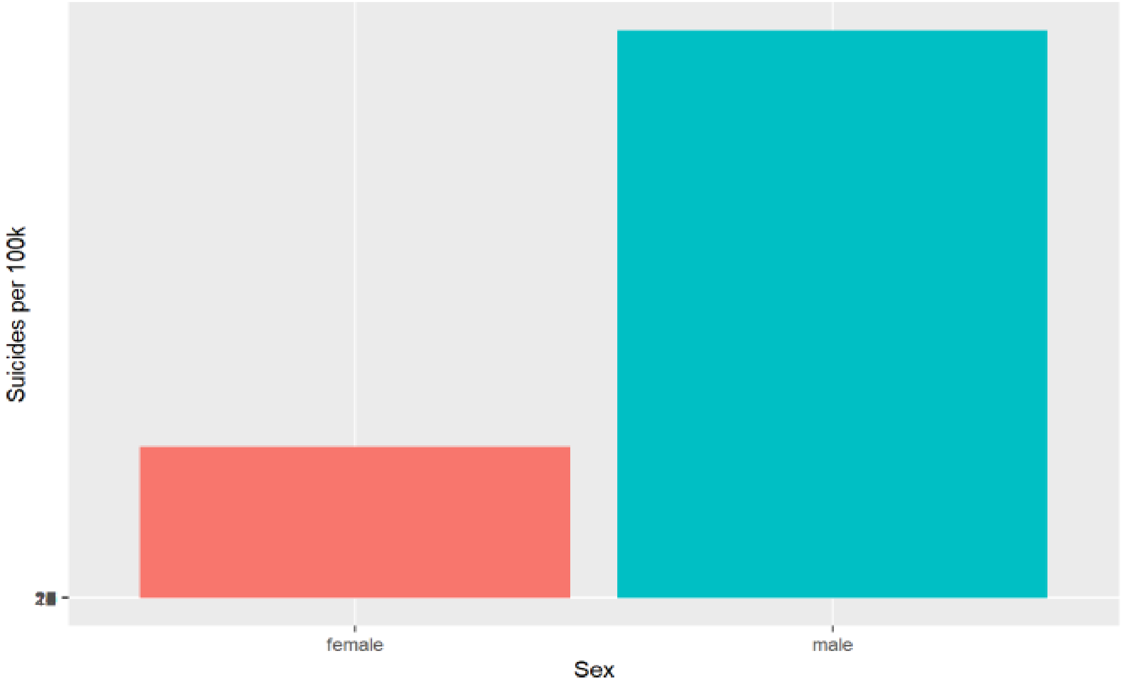
```
## [1] 0.02523964
```

<u>Analysis:</u>

Since the correlation in both cases is close to 0, we can say that there is a weaker correlation between the variables which supports our assumptions that though the GDP might be high, it does not prove that the suicide rate will be low.

It clearly signifies the fact that with the expectation of good lifestyle comes a huge responsibility of maintaining it, while some people fail to do so and hence end up giving up on their lives.

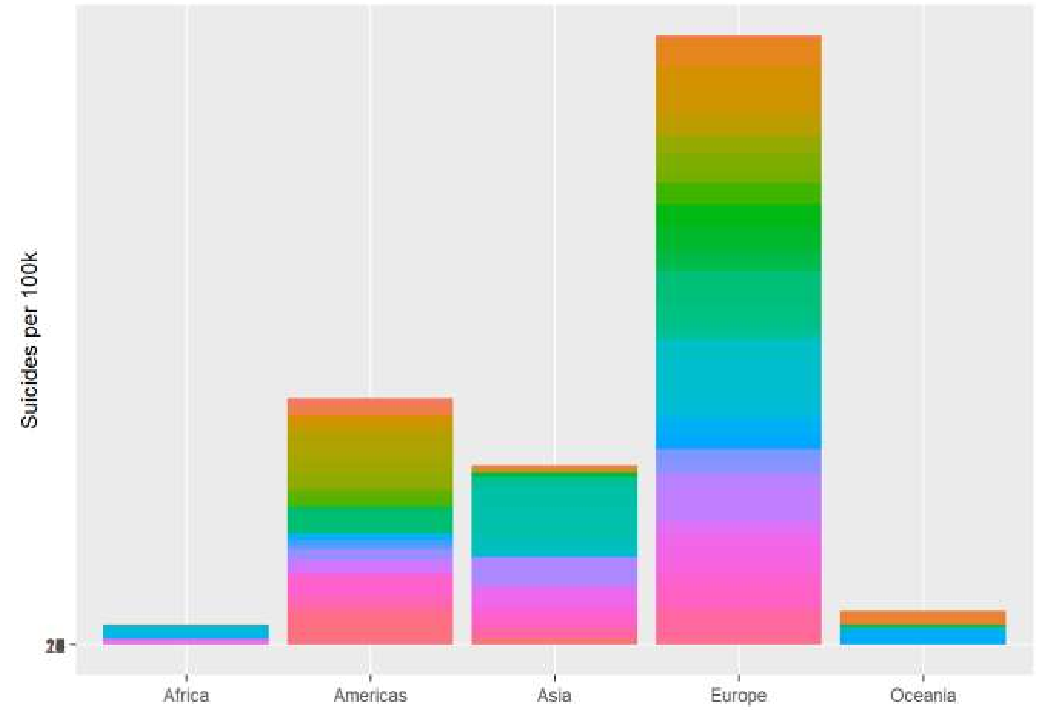
Global Suicides per 100K by sex:

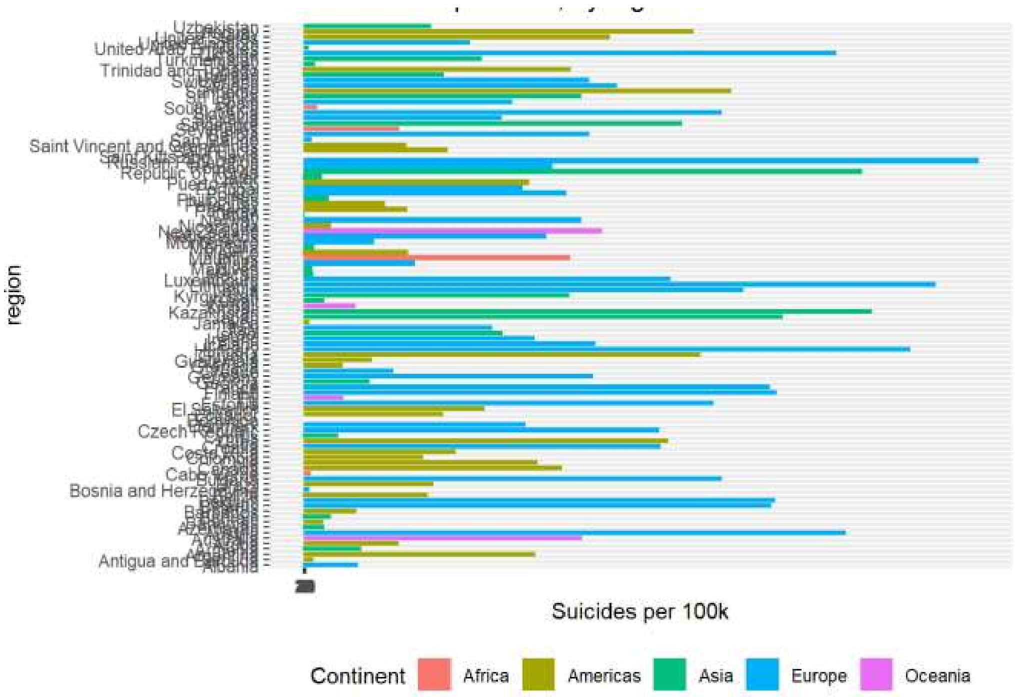


<u>Analysis:</u>

- ➢ The above histogram shows that males have a higher tendency of committing suicide compared to females.
- ➢ There have been various research across the global level which also supports this phenomenon.
- ➢ The reason may be that men tend to use more lethal methods which leaves less opportunity for rescue. Also, men are less likely to seek professional help for depression.

Global Suicides per 100K by continent :

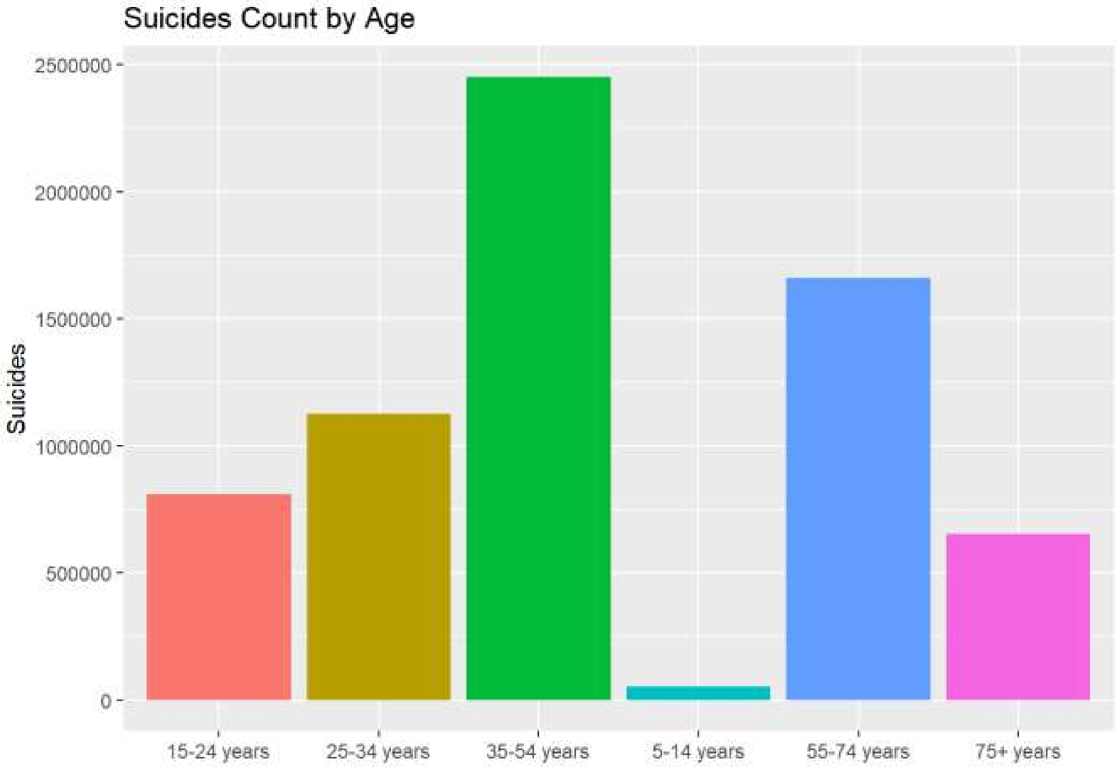




<u>Analysis</u>

- The above graphs depict the highest suicide rates prevalent in Europe.
- We know that most of the developed countries across the globe are in Europe.
- This demonstrates that with luxury comes loneliness, rejection, and, material conflicts which persuade a person to commit suicide.

I. Suicides count by age:



Analysis:

The plot depicts that suicide rates are highest among the age group of 35-54 years. These means there is a high suicide ratio in middle-aged people. A well-known factor that can be considered here is that there are a lot of responsibilities that come when an individual reaches the age of 35. And unfortunately, when he or she is not able to fulfil the same, it causes depression resulting in a thought of suicide.

Regression Analysis

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.

We have used simple linear regression in our project and the same is discussed below.

Simple linear regression

Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable. The simple linear model is expressed using the following equation:

$Y = a + bx + \epsilon$

Where:

Y – Dependent variable

X – Independent (explanatory) variable

a – Intercept

b – Slope

ϵ – Residual (error)

Linear Model Assumptions

Linear regression analysis is based on six fundamental assumptions:

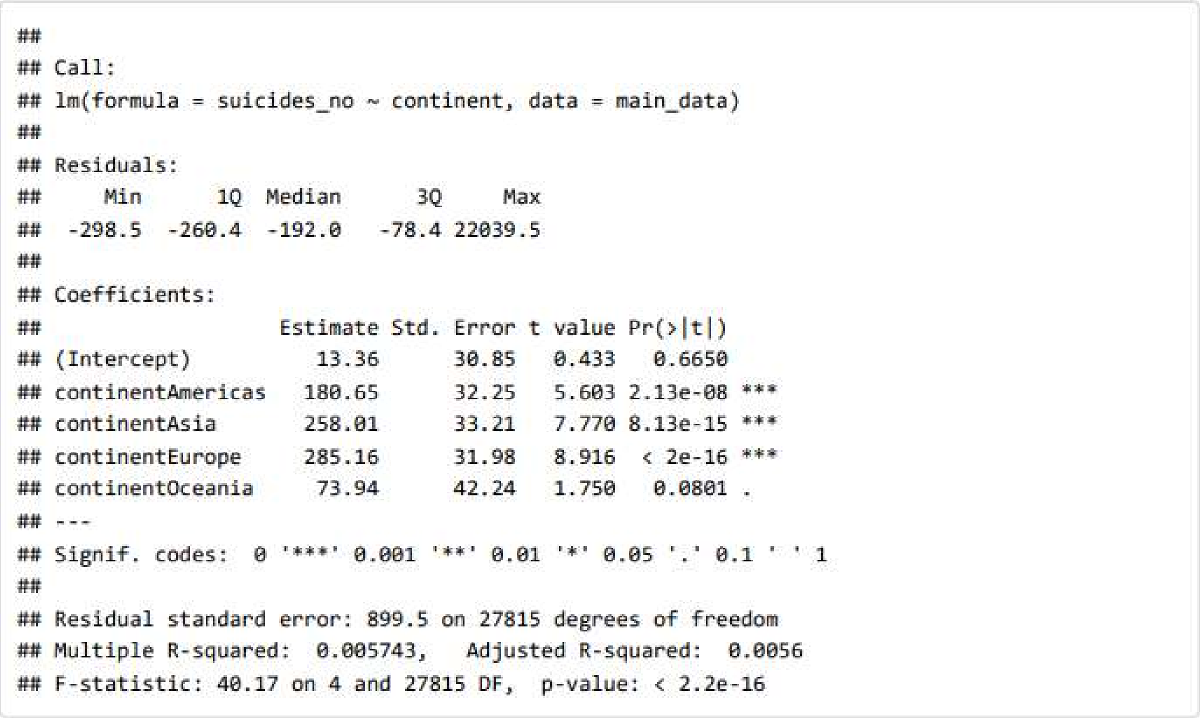
- The dependent and independent variables show a linear relationship between the slope and the intercept.
- The independent variable is not random.
- The value of the residual (error) is zero.
- The value of the residual (error) is constant across all observations.
- The value of the residual (error) is not correlated across all observations.
- The residual (error) values follow the normal distribution.

Benefits of Linear regression

Linear regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.

You'll find that linear regression is used in everything from biological, behavioral, environmental, and social sciences to business. Linear regression models have become a proven way to predict the future scientifically and reliably. Because linear regression is a long-established statistical procedure, the properties of linear regression models are well understood and can be trained very quickly.

Suicides number and continent



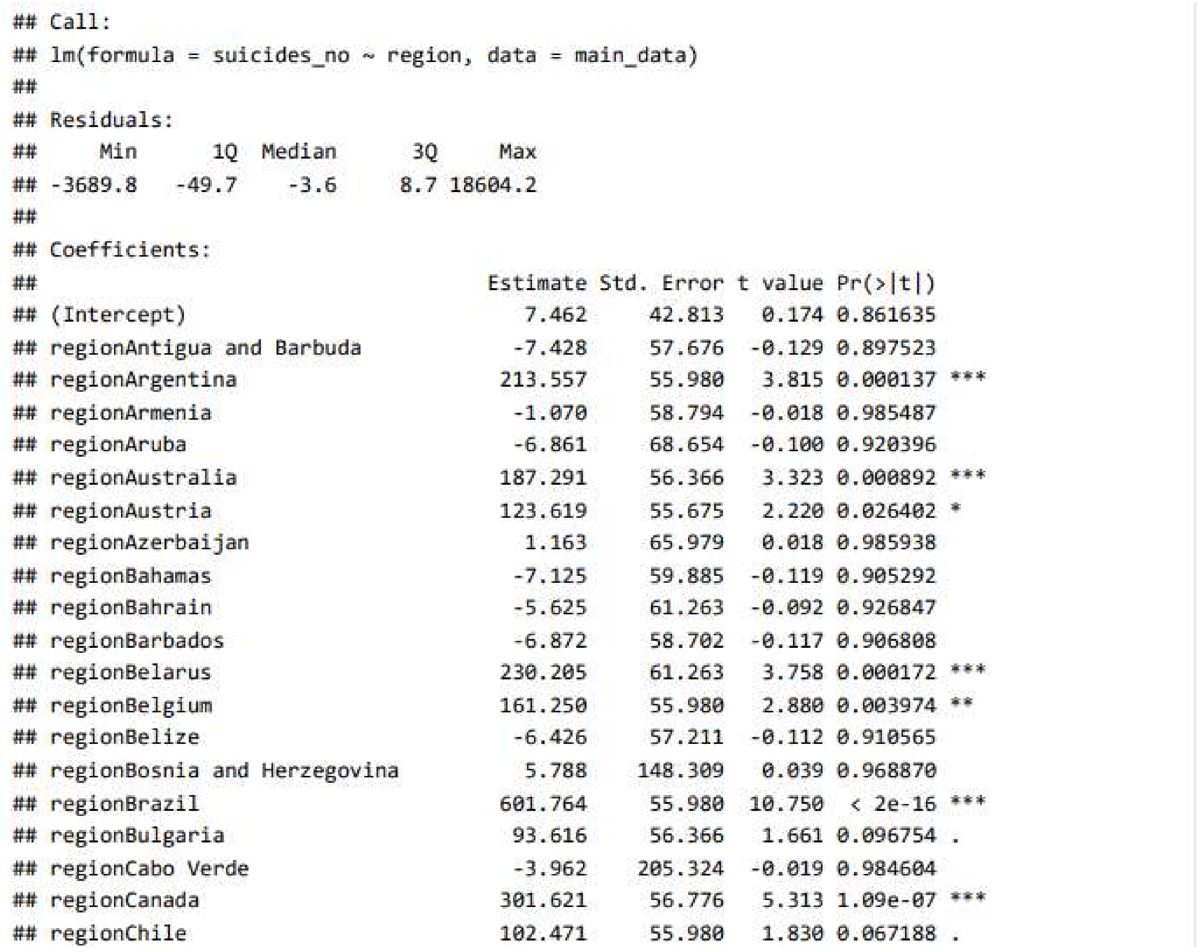
```
##
## Call:
## lm(formula = suicides_no ~ continent, data = main_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -298.5  -260.4  -192.0   -78.4 22039.5
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          13.36      30.85   0.433   0.6650
## continentAmericas   180.65      32.25   5.603 2.13e-08 ***
## continentAsia       258.01      33.21   7.770 8.13e-15 ***
## continentEurope     285.16      31.98   8.916  < 2e-16 ***
## continentOceania     73.94      42.24   1.750   0.0801 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 899.5 on 27815 degrees of freedom
## Multiple R-squared:  0.005743,   Adjusted R-squared:  0.0056
## F-statistic: 40.17 on 4 and 27815 DF,  p-value: < 2.2e-16
```

Analysis:

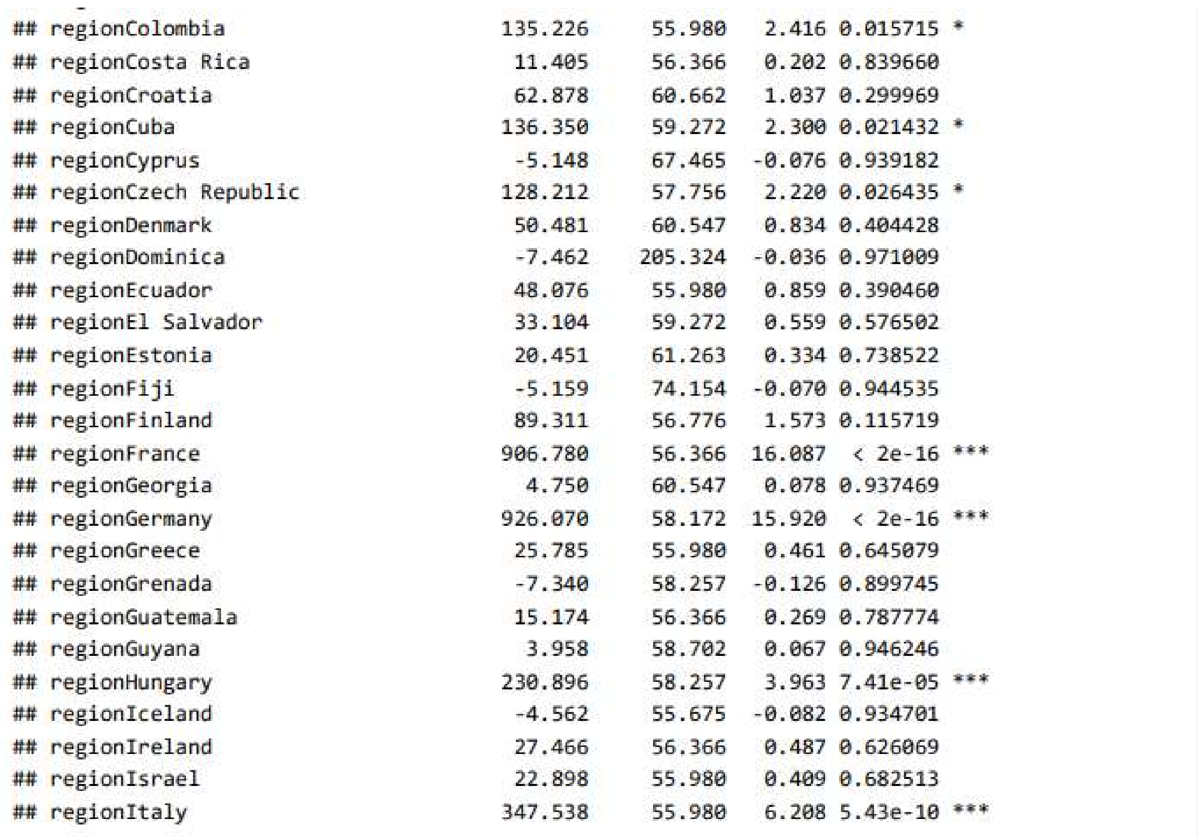
- From the above output, we can conclude that the distribution of the residuals do not appear to be strongly symmetrical. That means that the model predicts certain points that fall far away from the actually observed points.

- Also, we can see that p-values are very small when we selected suicide numbers with each of the continents. This means there exists a good relationship between suicide numbers and continents and we can reject the null hypothesis and accept the alternative hypothesis.

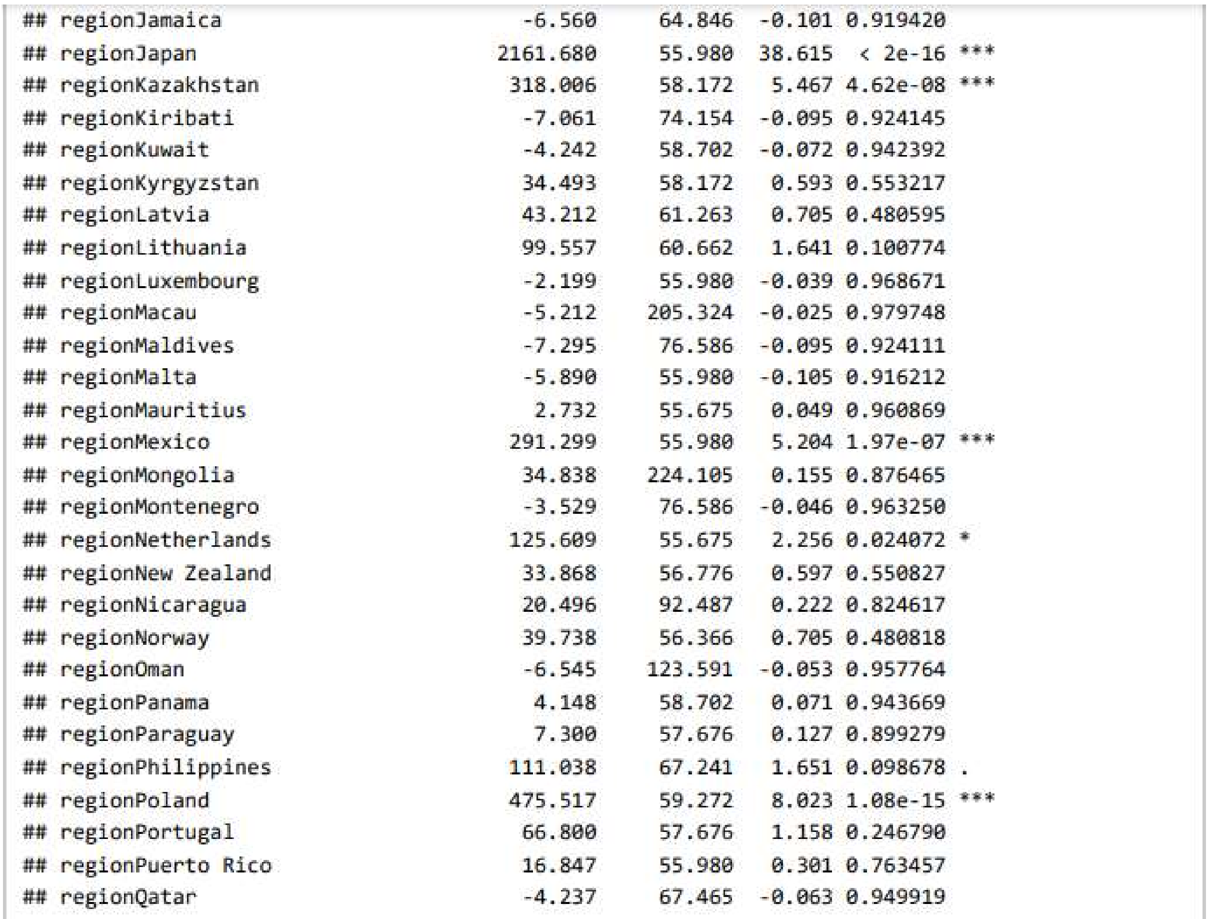
Suicides number and region



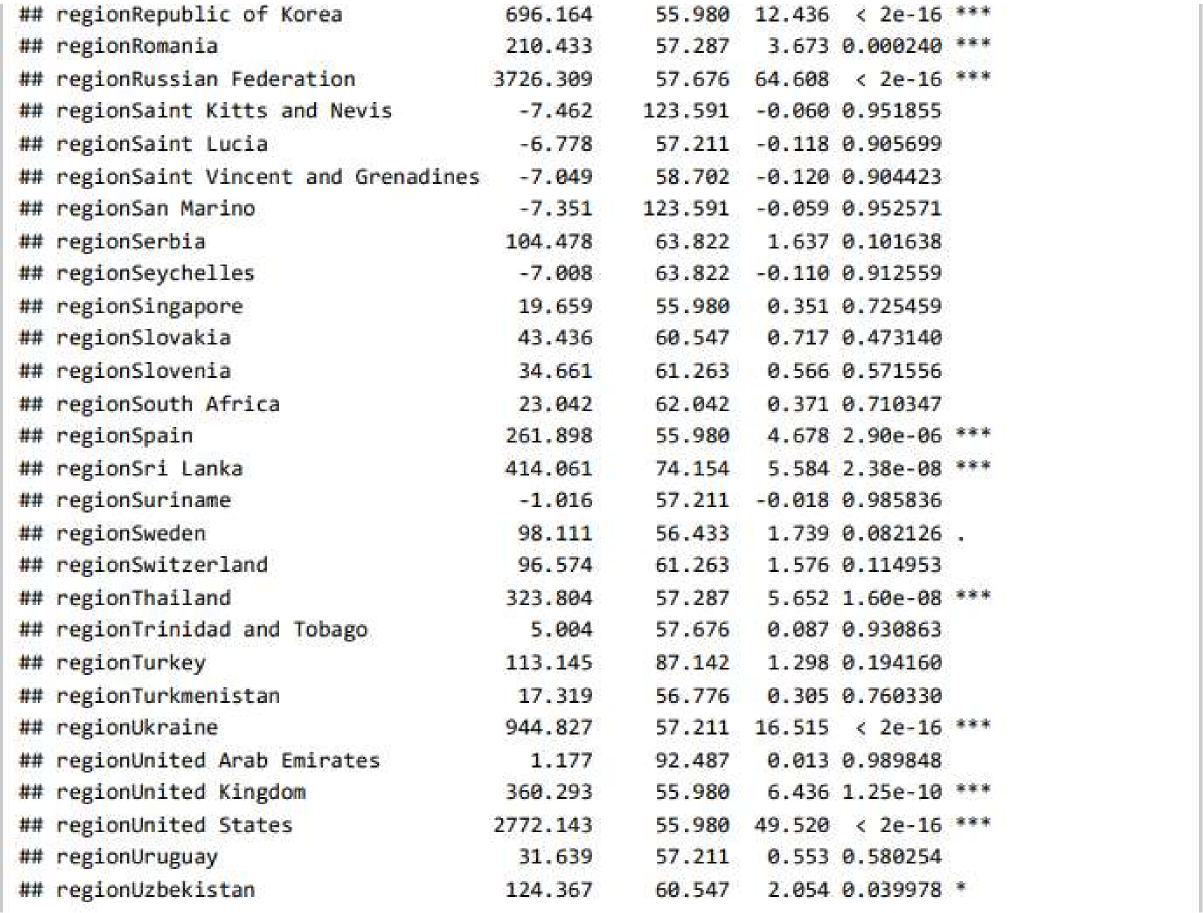
```
## Call:
## lm(formula = suicides_no ~ region, data = main_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3689.8   -49.7    -3.6     8.7 18604.2
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             7.462     42.813   0.174 0.861635
## regionAntigua and Barbuda              -7.428     57.676  -0.129 0.897523
## regionArgentina                       213.557     55.980   3.815 0.000137 ***
## regionArmenia                          -1.070     58.794  -0.018 0.985487
## regionAruba                            -6.861     68.654  -0.100 0.920396
## regionAustralia                       187.291     56.366   3.323 0.000892 ***
## regionAustria                         123.619     55.675   2.220 0.026402 *
## regionAzerbaijan                        1.163     65.979   0.018 0.985938
## regionBahamas                          -7.125     59.885  -0.119 0.905292
## regionBahrain                          -5.625     61.263  -0.092 0.926847
## regionBarbados                         -6.872     58.702  -0.117 0.906808
## regionBelarus                         230.205     61.263   3.758 0.000172 ***
## regionBelgium                         161.250     55.980   2.880 0.003974 **
## regionBelize                           -6.426     57.211  -0.112 0.910565
## regionBosnia and Herzegovina            5.788    148.309   0.039 0.968870
## regionBrazil                          601.764     55.980  10.750  < 2e-16 ***
## regionBulgaria                         93.616     56.366   1.661 0.096754 .
## regionCabo Verde                       -3.962    205.324  -0.019 0.984604
## regionCanada                          301.621     56.776   5.313 1.09e-07 ***
## regionChile                           102.471     55.980   1.830 0.067188 .
```



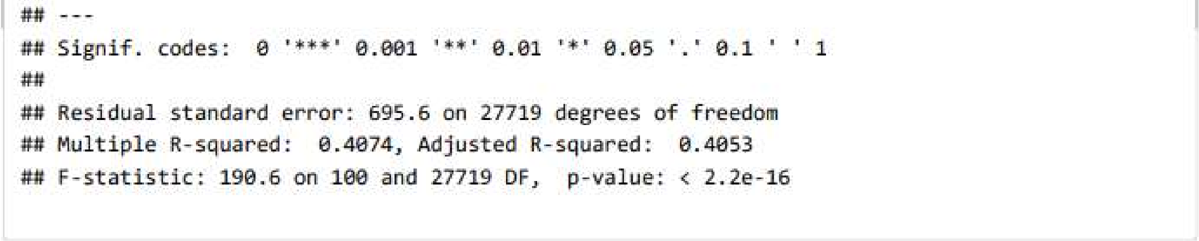
```
## regionColombia                        135.226     55.980   2.416 0.015715 *
## regionCosta Rica                       11.405     56.366   0.202 0.839660
## regionCroatia                          62.878     60.662   1.037 0.299969
## regionCuba                            136.350     59.272   2.300 0.021432 *
## regionCyprus                           -5.148     67.465  -0.076 0.939182
## regionCzech Republic                  128.212     57.756   2.220 0.026435 *
## regionDenmark                          50.481     60.547   0.834 0.404428
## regionDominica                         -7.462    205.324  -0.036 0.971009
## regionEcuador                          48.076     55.980   0.859 0.390460
## regionEl Salvador                      33.104     59.272   0.559 0.576502
## regionEstonia                          20.451     61.263   0.334 0.738522
## regionFiji                             -5.159     74.154  -0.070 0.944535
## regionFinland                          89.311     56.776   1.573 0.115719
## regionFrance                          906.780     56.366  16.087  < 2e-16 ***
## regionGeorgia                           4.750     60.547   0.078 0.937469
## regionGermany                         926.070     58.172  15.920  < 2e-16 ***
## regionGreece                           25.785     55.980   0.461 0.645079
## regionGrenada                          -7.340     58.257  -0.126 0.899745
## regionGuatemala                        15.174     56.366   0.269 0.787774
## regionGuyana                            3.958     58.702   0.067 0.946246
## regionHungary                         230.896     58.257   3.963 7.41e-05 ***
## regionIceland                          -4.562     55.675  -0.082 0.934701
## regionIreland                          27.466     56.366   0.487 0.626069
## regionIsrael                           22.898     55.980   0.409 0.682513
## regionItaly                           347.538     55.980   6.208 5.43e-10 ***
```



```
## regionJamaica                    -6.560     64.846  -0.101 0.919420
## regionJapan                    2161.680     55.980  38.615  < 2e-16 ***
## regionKazakhstan                318.006     58.172   5.467 4.62e-08 ***
## regionKiribati                   -7.061     74.154  -0.095 0.924145
## regionKuwait                     -4.242     58.702  -0.072 0.942392
## regionKyrgyzstan                 34.493     58.172   0.593 0.553217
## regionLatvia                     43.212     61.263   0.705 0.480595
## regionLithuania                  99.557     60.662   1.641 0.100774
## regionLuxembourg                 -2.199     55.980  -0.039 0.968671
## regionMacau                      -5.212    205.324  -0.025 0.979748
## regionMaldives                   -7.295     76.586  -0.095 0.924111
## regionMalta                      -5.890     55.980  -0.105 0.916212
## regionMauritius                   2.732     55.675   0.049 0.960869
## regionMexico                    291.299     55.980   5.204 1.97e-07 ***
## regionMongolia                   34.838    224.105   0.155 0.876465
## regionMontenegro                 -3.529     76.586  -0.046 0.963250
## regionNetherlands               125.609     55.675   2.256 0.024072 *
## regionNew Zealand                33.868     56.776   0.597 0.550827
## regionNicaragua                  20.496     92.487   0.222 0.824617
## regionNorway                     39.738     56.366   0.705 0.480818
## regionOman                       -6.545    123.591  -0.053 0.957764
## regionPanama                      4.148     58.702   0.071 0.943669
## regionParaguay                    7.300     57.676   0.127 0.899279
## regionPhilippines               111.038     67.241   1.651 0.098678 .
## regionPoland                    475.517     59.272   8.023 1.08e-15 ***
## regionPortugal                   66.800     57.676   1.158 0.246790
## regionPuerto Rico                16.847     55.980   0.301 0.763457
## regionQatar                      -4.237     67.465  -0.063 0.949919
```



```
## regionRepublic of Korea          696.164     55.980  12.436  < 2e-16 ***
## regionRomania                    210.433     57.287   3.673 0.000240 ***
## regionRussian Federation        3726.309     57.676  64.608  < 2e-16 ***
## regionSaint Kitts and Nevis       -7.462    123.591  -0.060 0.951855
## regionSaint Lucia                 -6.778     57.211  -0.118 0.905699
## regionSaint Vincent and Grenadines  -7.049     58.702  -0.120 0.904423
## regionSan Marino                  -7.351    123.591  -0.059 0.952571
## regionSerbia                     104.478     63.822   1.637 0.101638
## regionSeychelles                  -7.008     63.822  -0.110 0.912559
## regionSingapore                   19.659     55.980   0.351 0.725459
## regionSlovakia                    43.436     60.547   0.717 0.473140
## regionSlovenia                    34.661     61.263   0.566 0.571556
## regionSouth Africa                23.042     62.042   0.371 0.710347
## regionSpain                      261.898     55.980   4.678 2.90e-06 ***
## regionSri Lanka                  414.061     74.154   5.584 2.38e-08 ***
## regionSuriname                    -1.016     57.211  -0.018 0.985836
## regionSweden                      98.111     56.433   1.739 0.082126 .
## regionSwitzerland                 96.574     61.263   1.576 0.114953
## regionThailand                   323.804     57.287   5.652 1.60e-08 ***
## regionTrinidad and Tobago          5.004     57.676   0.087 0.930863
## regionTurkey                     113.145     87.142   1.298 0.194160
## regionTurkmenistan                17.319     56.776   0.305 0.760330
## regionUkraine                    944.827     57.211  16.515  < 2e-16 ***
## regionUnited Arab Emirates         1.177     92.487   0.013 0.989848
## regionUnited Kingdom             360.293     55.980   6.436 1.25e-10 ***
## regionUnited States             2772.143     55.980  49.520  < 2e-16 ***
## regionUruguay                     31.639     57.211   0.553 0.580254
## regionUzbekistan                 124.367     60.547   2.054 0.039978 *
```



```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 695.6 on 27719 degrees of freedom
## Multiple R-squared:  0.4074, Adjusted R-squared:  0.4053
## F-statistic: 190.6 on 100 and 27719 DF,  p-value: < 2.2e-16
```

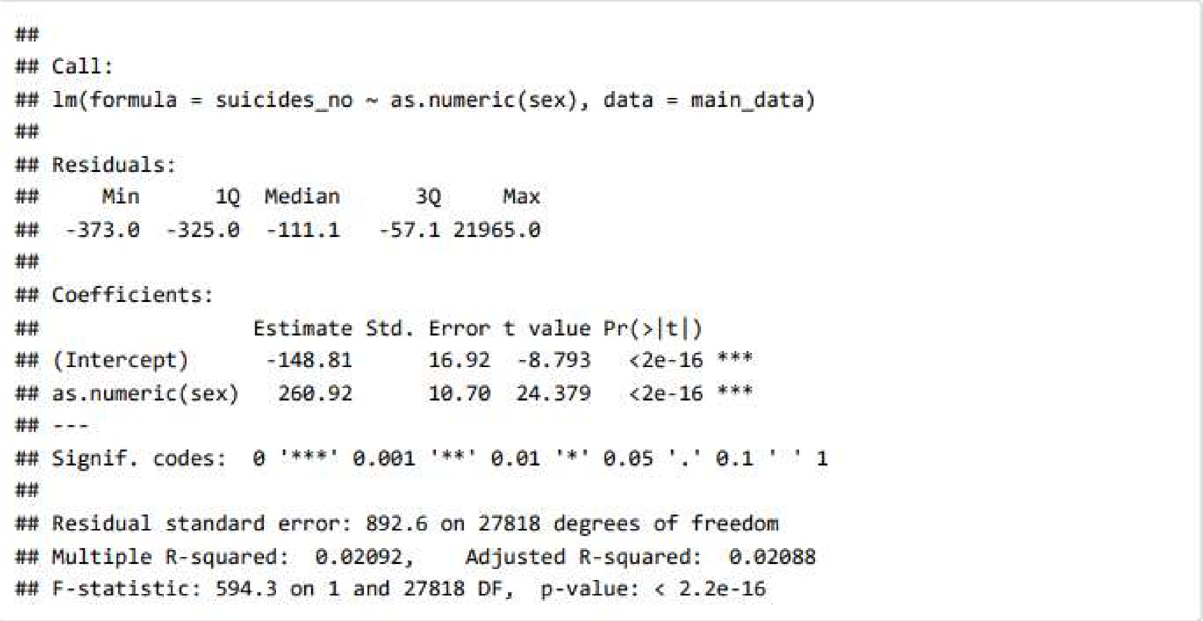
Analysis:

- From the above output, we can see that p-values are very small for many European countries which shows that Europe has a higher number of suicides.

- Moreover, both Multiple R-squared and adjusted R-squared values are around 0.40 which shows that roughly 40% of the variance found in the response variable (suicide numbers) can be explained by the predictor variable (regions).

- Also, the F-statistic is 190.6 which means it is far away from 1 and this is because of the large number of data points present in the dataset.

Suicides number and sex



```
##
## Call:
## lm(formula = suicides_no ~ as.numeric(sex), data = main_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -373.0  -325.0  -111.1   -57.1 21965.0
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -148.81      16.92  -8.793   <2e-16 ***
## as.numeric(sex)   260.92      10.70  24.379   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 892.6 on 27818 degrees of freedom
## Multiple R-squared:  0.02092,    Adjusted R-squared:  0.02088
## F-statistic: 594.3 on 1 and 27818 DF,  p-value: < 2.2e-16
```

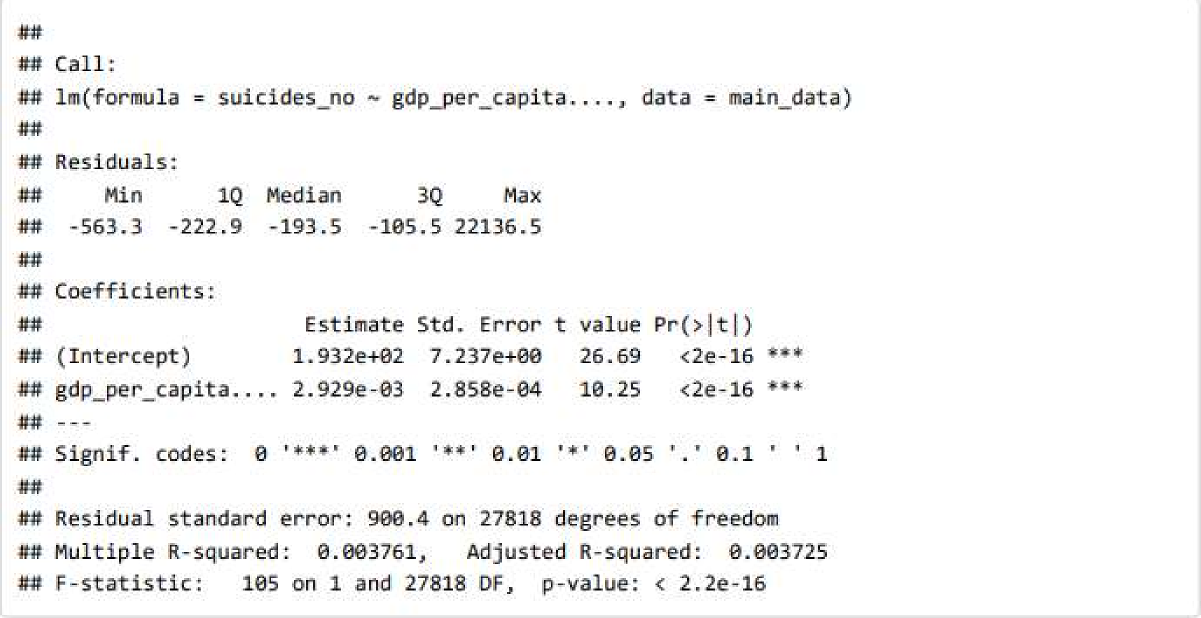
Analysis:

- The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. We want it to be far away from zero as this would indicate we could reject the null hypothesis - that is, we could declare a relationship between sex and suicides number exist.

- In our example, the t-statistic values are relatively far away from zero and are large relative to the standard error, which could indicate a relationship exists.

- Also, the p-value is very small that further justifies the same.

Suicides number and GDP per capita



```
##
## Call:
## lm(formula = suicides_no ~ gdp_per_capita...., data = main_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -563.3  -222.9  -193.5  -105.5 22136.5
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.932e+02  7.237e+00   26.69   <2e-16 ***
## gdp_per_capita.... 2.929e-03  2.858e-04   10.25   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 900.4 on 27818 degrees of freedom
## Multiple R-squared:  0.003761,   Adjusted R-squared:  0.003725
## F-statistic:   105 on 1 and 27818 DF,  p-value: < 2.2e-16
```

Analysis:

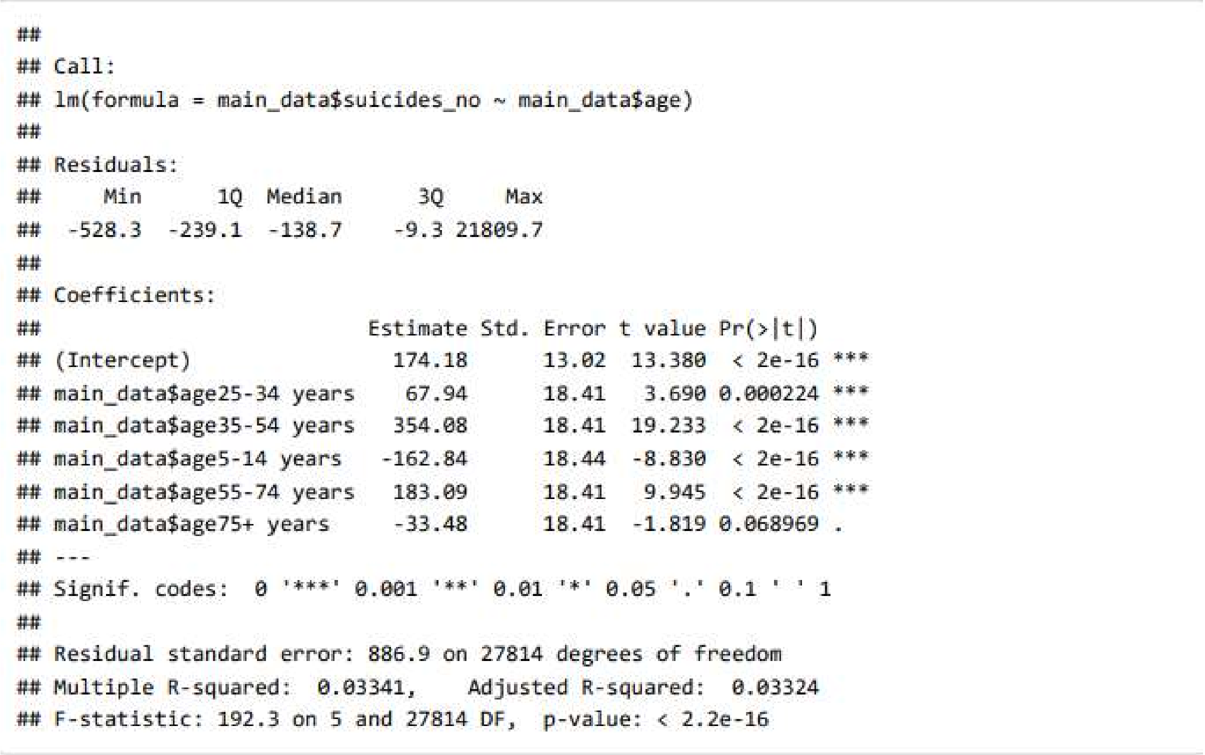
- In our output, it can be seen that p-value of the F-statistic is < 2.2e-16, which is highly significant. This means that the predictor variable (GDP per capita) is significantly related to the outcome variable (suicide number).

- Moreover, the t-statistic value is relatively far away from zero and is large relative to the standard error, which could indicate a relationship exists.

- Also, the p-value is very small that further justifies the same.

- There is a weak but significant positive linear relationship - richer countries are associated with higher rates of suicide, but this is a weak relationship which can be seen from the figures.

Suicides number and age



```
##
## Call:
## lm(formula = main_data$suicides_no ~ main_data$age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -528.3  -239.1  -138.7    -9.3 21809.7
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               174.18      13.02  13.380  < 2e-16 ***
## main_data$age25-34 years   67.94      18.41   3.690 0.000224 ***
## main_data$age35-54 years  354.08      18.41  19.233  < 2e-16 ***
## main_data$age5-14 years  -162.84      18.44  -8.830  < 2e-16 ***
## main_data$age55-74 years  183.09      18.41   9.945  < 2e-16 ***
## main_data$age75+ years    -33.48      18.41  -1.819 0.068969 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 886.9 on 27814 degrees of freedom
## Multiple R-squared:  0.03341,    Adjusted R-squared:  0.03324
## F-statistic: 192.3 on 5 and 27814 DF,  p-value: < 2.2e-16
```

Analysis:

- In the above output, it can be seen that age group of 35-54 years is significantly associated with suicide numbers whereas the rest of the age groups are not significantly associated with suicide numbers.

- The r-squared is 0.03341, so the age group explains very little of the variance in suicide rate overall.

Random Forest Method

➢ What is a random forest?

Random Forest is a powerful and versatile supervised machine learning algorithm that grows and combines multiple decision trees to create a "forest." It can be used for both classification and regression problems

➢ Important Features of Random Forest

1. Diversity- Not all attributes/variables/features are considered while making an individual tree, each tree is different.

2. Immune to the curse of dimensionality- Since each tree does not consider all the features, the feature space is reduced.

3. Parallelization-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.

4. Train-Test split- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.

5. Stability- Stability arises because the result is based on majority voting/ averaging.

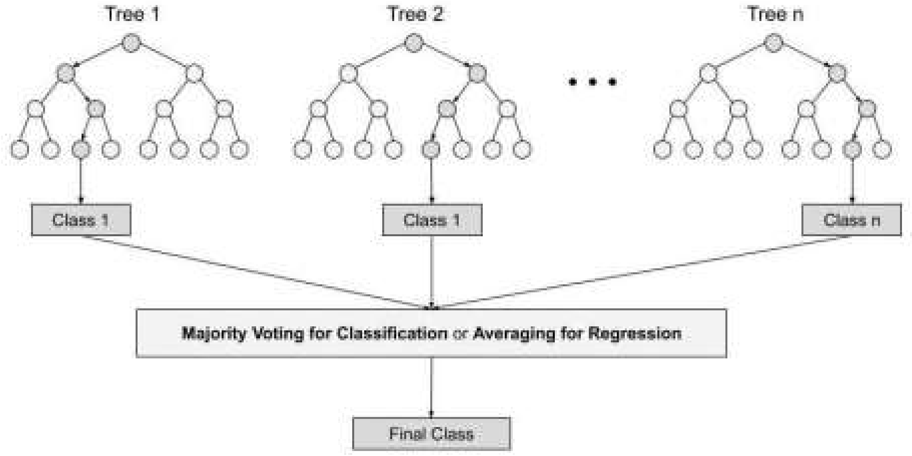
➢ Steps involved in random forest algorithm:

Step 1: In Random forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.


Tree 1
Tree 2
Tree n
• • •
Class 1
Class 1
Class n
Majority Voting for Classification or Averaging for Regression
Final Class

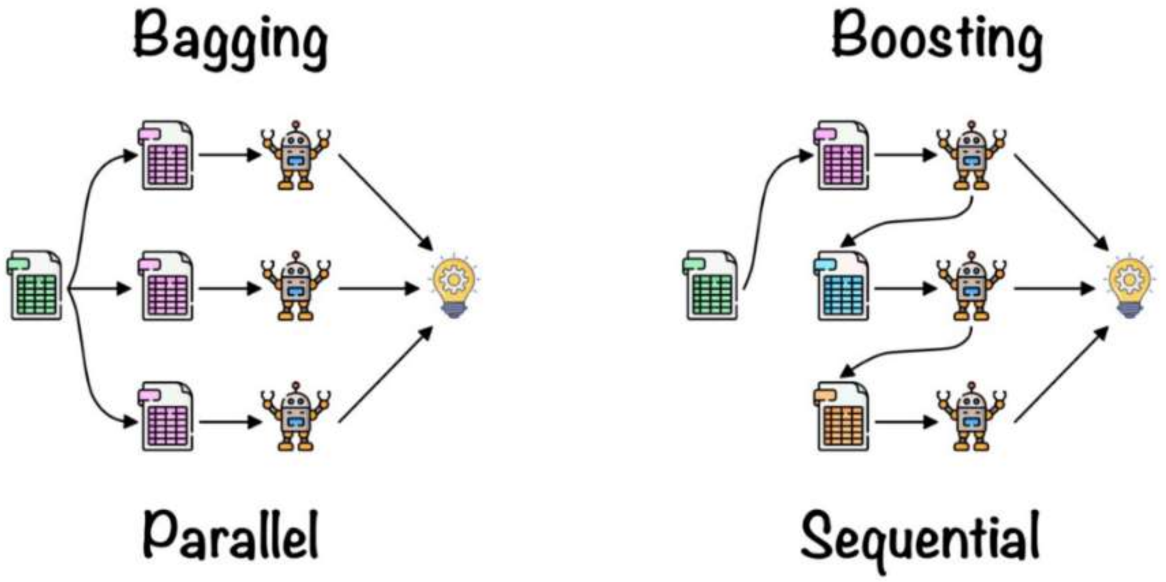
➢ How does random forest works?

Before understanding the working of the random forest algorithm in machine learning, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus a collection of models is used to make predictions rather than an individual model.

Ensemble uses two types of methods:

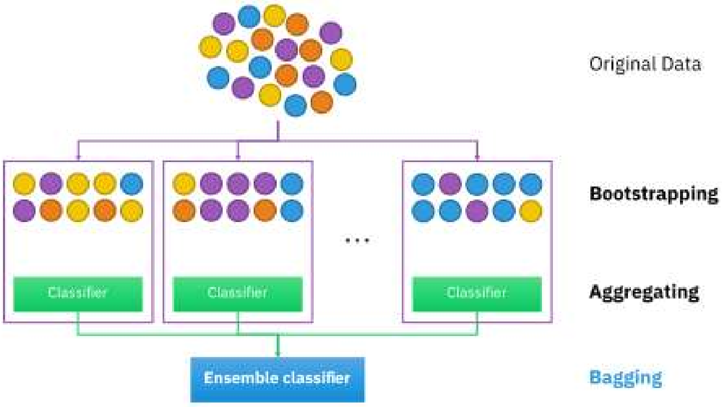
1. Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.

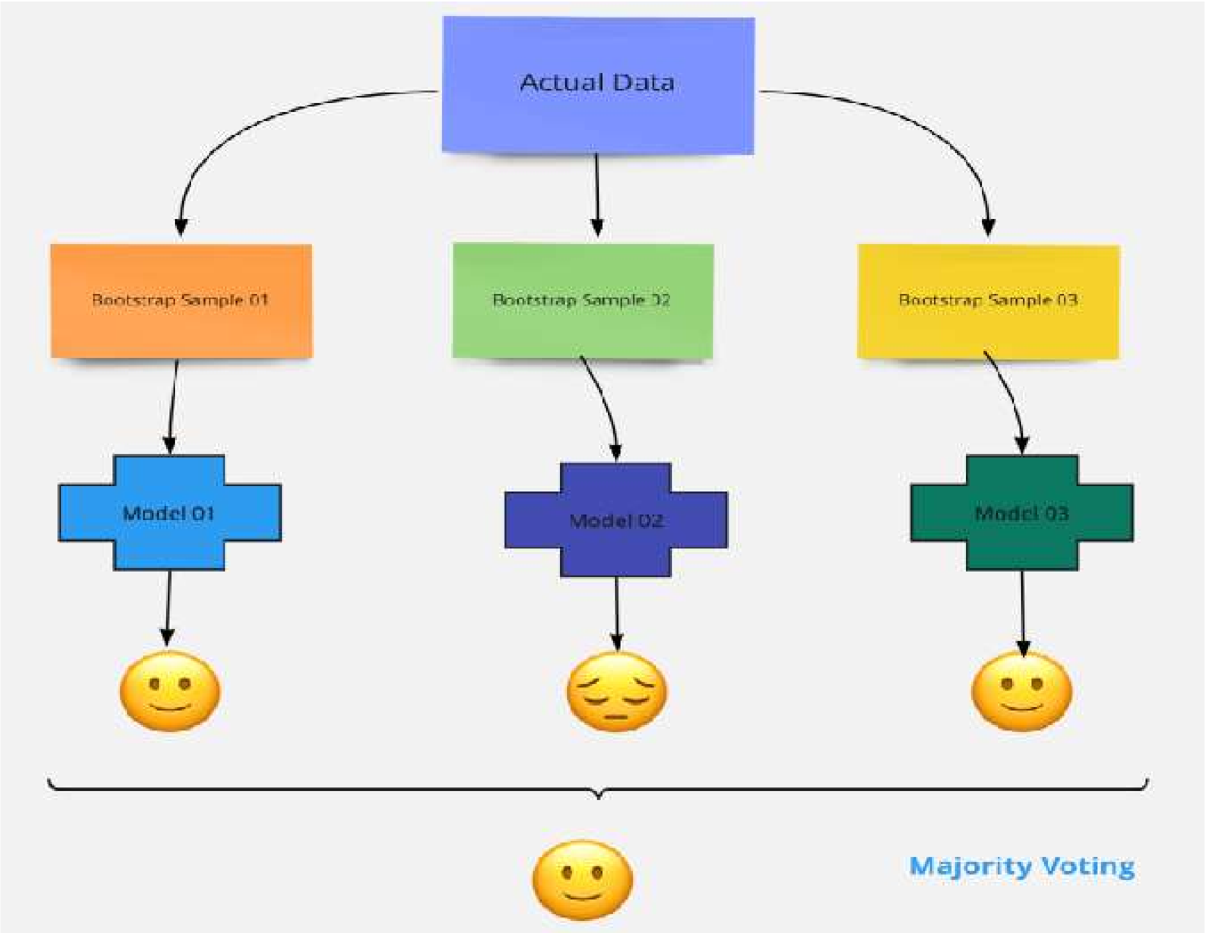
2. Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST



As mentioned earlier, **Random forest works on the Bagging principle**. Now let's dive in and understand bagging in detail.

Bagging, also known as Bootstrap Aggregation is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.





Random Forest grows multiple decision trees which are merged together for a more accurate prediction.

The logic behind the Random Forest model is that multiple uncorrelated models (the individual decision trees) perform much better as a group than they do alone. When using Random Forest for classification, each tree gives a classification or a "vote." The forest chooses the classification with the majority of the "votes." When using Random Forest for regression, the forest picks the average of the outputs of all trees

The key here lies in the fact that there is low (or no) correlation between the individual models—that is, between the decision trees that make up the larger Random Forest model. While individual decision trees may produce errors, the majority of the group will be correct, thus moving the overall outcome in the right direction.

- **Advantages**

1. It can be used in classification and regression problems.

2. It solves the problem of overfitting as output is based on majority voting or averaging.

3. It performs well even if the data contains null/missing values.

4. Each decision tree created is independent of the other thus it shows the property of parallelization.

5. It is highly stable as the average answers given by a large number of trees are taken.

6. It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.

7. It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.

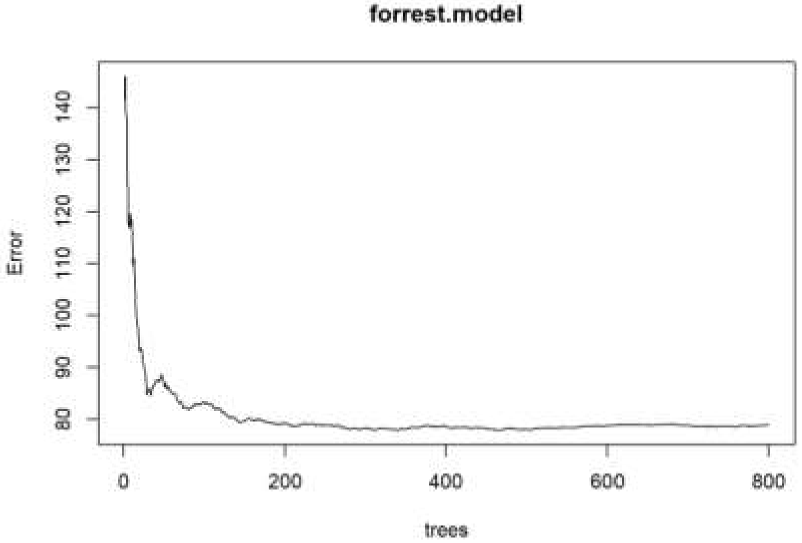
8. We don’t have to segregate data into train and test as there will always be 30% of the data which is not seen by the decision tree made out of bootstrap

- **Disadvantages**

1. Random forest is highly complex when compared to decision trees where decisions can be made by following the path of the tree.

2. Training time is more compared to other models due to its complexity. Whenever it must make a prediction each decision tree has to generate output for the given input data.

Mean Squared Error and Number of trees



<u>Analysis:</u>

From the above graph, it can be concluded that the mean squared error (MSE) is extremely high when the trees are low. But as the number of trees increases, MSE also reduces. This mainly happens because more trees mean more features/parameters have been added to the model. And a high number of features in an ML model always reduces training error.

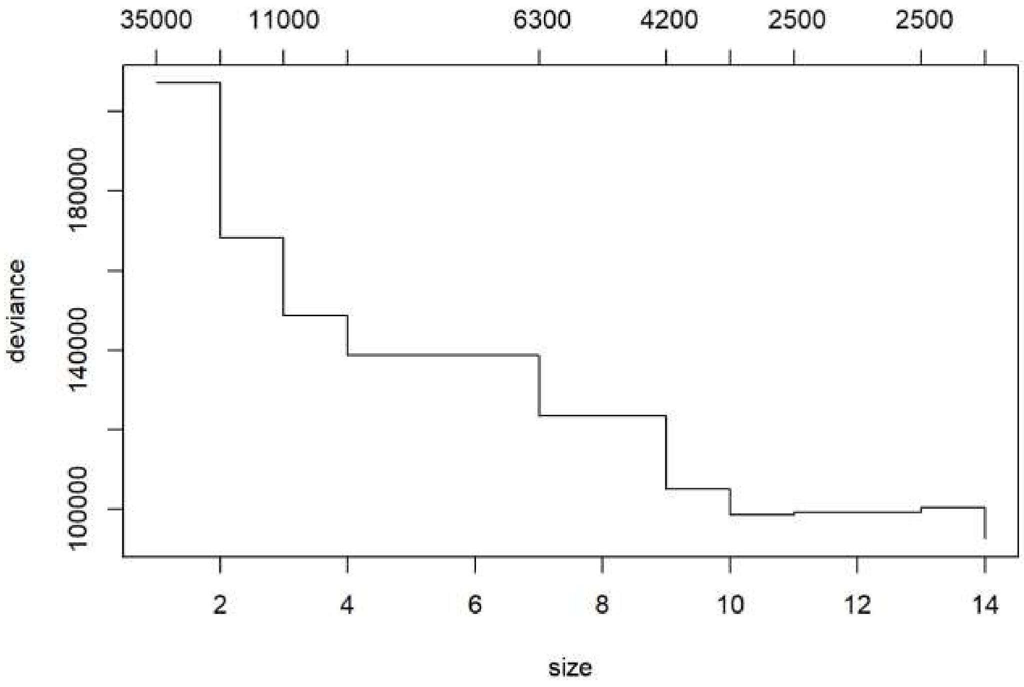
This is simply because if those additional features are unhelpful, then those features will not be used, and the training error will at least remain the same as the model with fewer features.

This however does not mean that adding more features/parameters is always a good idea as a reduction in training error does not imply a reduction in generalization error. In other words, your model could be overfitting on the training data but may not show error reduction on test data. A good approach to finding the ideal number of trees is to plot the test error with an increase in the number of trees and select the number at which the test error starts plateauing.

Thus, beyond a certain point, the tradeoff may not be worth it.

If we see the graph, it is visible that after 300, there is no need of increasing the number of trees as it MSE is not diminishing below that level.

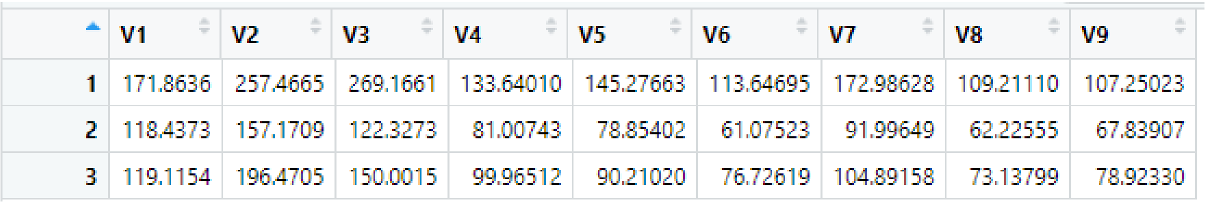
Deviance and Size



Analysis:

It can be inferred from the above graph that as we increase the size of the model, the overall deviance reduces which justifies the basic property of the random forest method. Thus the size at 14 gives the optimal solution.

Actual MSE across all Models (Single tree, bagging, and random forest method)

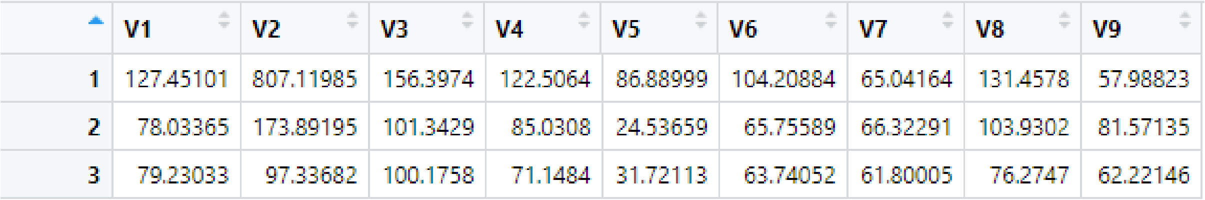


	V1	V2	V3	V4	V5	V6	V7	V8	V9
1	171.8636	257.4665	269.1661	133.64010	145.27663	113.64695	172.98628	109.21110	107.25023
2	118.4373	157.1709	122.3273	81.00743	78.85402	61.07523	91.99649	62.22555	67.83907
3	119.1154	196.4705	150.0015	99.96512	90.21020	76.72619	104.89158	73.13799	78.92330

Analysis:

It can be deduced from the above table that the bagging method gives the lowest MSE for all the years as compared to the single decision tree and random forest method.

Predicted MSE across all Models (Single tree, begging and random forest method)



	V1	V2	V3	V4	V5	V6	V7	V8	V9
1	127.45101	807.11985	156.3974	122.5064	86.88999	104.20884	65.04164	131.4578	57.98823
2	78.03365	173.89195	101.3429	85.0308	24.53659	65.75589	66.32291	103.9302	81.57135
3	79.23033	97.33682	100.1758	71.1484	31.72113	63.74052	61.80005	76.2747	62.22146

Analysis:

In the above table, there is no clear picture of which method is better because for some years MSE is lowest using the bagging method while for others, the random forest method seems to be a good choice.

Section5-Team Members Contribution

The whole project work is done with the continuous collaboration of all the team members. The whole process took us almost 40-45 days and we divided our work according to our understanding and availability. The project dealt with several processes which were not segmented, hence it's not possible to pinpoint the work done by an individual. This project work was done with each one of the groups involved in every step and each member has a detailed comprehensive understanding of every step.