**DASHBOARD DEVELOPMENT USING POWER BI VISUALIZATION TOOL AND STATISTICAL ANALYSIS USING RSTUDIO**

**A Project report**

***Submitted by***

***ATITEBI ZAYNAB MORENIKEJI***

**(@00669741)**

***In fulfilment of***

**Applied Statistics and Data Visualization Module course work**

***For***

**MASTER OF SCIENCE**

***In***

**DATA SCIENCE**

**School of Computing, Science & Engineering**

**University of Salford, Manchester.**

**December, 2022**

Table of Contents

Part One: Interactive Dashboard Development………………………………………………………………..3

1. Introduction………………………………………………………………………………………………………….3

1.1 Aims and Objectives………………………………………………………………………………….3

1. Background Research……………………………………………………………………………………………3
2. Exploration of Data Set…………………………………………………………………………………………5
   1. Data Transformation and cleaning…………………………………………………………….6
   2. Data Modelling………………………………………………………………………………………….6
3. Investigation of Data Workflows & Proposal for Design of Dashboard………………….7
4. Discussion…………………………………………………………………………………………………………….15
5. Conclusions………………………………………………………………………………………………………….17

6.1 Recommendations……………………………………………………………………………………17

Part Two: Statistical Analysis……………………………………………………………………………………………18

1. Introduction…………………………………………………………………………………………………………18

1.1 Aim and Objectives…………………………………………………………………………………..18

1. Background Research…………………………………………………………………………………………..19
   1. Descriptive Analysis………………………………………………………………………………….19
   2. Correlation Analysis………………………………………………………………………………….20
   3. Regression Analysis…………………………………………………………………………………..20
   4. Time series Analysis………………………………………………………………………………….21
   5. Comparative Analysis……………………………………………………………………………….22
2. Exploration of Data set…………………………………………………………………………………………22
3. Analysis………………………………………………………………………………………………………………..25
4. Discussion…………………………………………………………………………………………………………….39
5. Conclusions…………………………………………………………………………………………………………..42

Part Three: References and Appendices…………………………………………………………………………...43

References………………………………………………………………………………………………………………………..43

**PART ONE: INTERACTIVE DASHBOARD DEVELOPMENT**

1. **Introduction**

In the recent years, the rate of suicide has increased, and this have raised a cause of concern to the world. In fact, suicide is one of the topmost leading causes of death in the world. It was reported that up to 800,000 people die due to suicide every year. This means about 1 person die due to suicide every 40 seconds (Ritche et al., 2015). This is even an under reported figure since some countries make it illegal and wouldn’t release data for analysis. Hence, the need for us to analyze the trend in the suicide rate over the years.

This report contains analysis of suicide rate of 5 developed and 5 underdeveloped countries. A total of 100 observation with 5 indicators were used, and it spanned over a period of 20 years. It is worth noting that Power Bi visualization tool was used for this project, and the choice of using it was because its more flexible for me and has a better user interface.

Also, after analyzing the suicide rate of the countries over the years, its effect on population growth was analyzed, as this is important to be able to report its effect on the economic growth.

**1.1 Aim and Objectives**

The aim of this project is to make a comparative analysis of the suicide rate between some selected developed and underdeveloped countries and its effect on population growth. To achieve this aim, the following objectives were set out.

1. plotting the total population of each selected countries.
2. Plotting the suicide rate between the total male and female populations.
3. Projecting the suicide rate over the years and forecasting into the future.
4. Viewing the topmost suicidal countries and the least suicidal countries.
5. Estimating the population growth in percentage.

**2.0 Background Research**

A Data Analyst should be aware of the data analytics processes which are Data gathering, Assessment, cleaning, Modelling and Visualization. Most times, Analysts rush to design dashboards without paying close attention to the processes involve in data analysis. We should be aware that visualization is the last stage, and every other process must be completed before visualization starts. And this is where the dashboard design comes in.

But before going further, we need to define what a dashboard is. According to Subotin (2017), a dashboard is a visual representation of the most crucial data required to accomplish one or more goals, gathered, and arranged on a single screen for easy monitoring. As a data scientist or analyst, you must understand the purpose of building the dashboard and this is reflected in the above definition. Also, your dashboard must have an aim, put the target audience into consideration and focus on key metrics to convey information better. Hence, there are 4 c’s that an effective and interactive dashboard must have.

1. Choice of chart: An analyst must choose a chart that will convey the necessary information. In fact, the selection of the best chart type for a given dataset may influence how viewers interpret it in the future. When readers have varying levels of data literacy, it is extremely crucial to use the chart type that most effectively communicates insight (Zubiaga & MacNamee, 2015).
2. Clarity: Clarity of purpose is important in a dashboard design. This will guide an analyst in removing any irrelevant data from the data set as too much information might take away from what's vital and make things more difficult to find (Geckoboard, 2022). Moreover, removing irrelevant information makes the data less bulky and easy to use.
3. Color: Analyst should select a vibrant hue for their illustrations. But the hue shouldn't undermine the goal of the images. Color blinded audience should also be put into consideration as we never can tell how far the dashboard would go. Information from Insightsoftware (2021) reiterated that many colors have connotations that can be used positively or negatively. The appropriate colors can help you convey meaning to your audience without utilizing additional labels and indicators in your dashboards.

In addition to the above, McGuigan (2020) stated that there are three types of color schemes you can use on your charts to make it more appealing. They are the divergent, the sequential, and the categorical schemes. When a center value is shared across both ends of a visual, use the divergent color schemes. Also, sequential colors should be used with ordered or numeric data. While categorical colors should be used with different variables without any sort of sorting. She further stated that this can help prevent chart confusion and keep your story on track. Hence, Colors are essential for conveying your message.

1. Context: Most dashboards are usually categorized as Exploratory and Explanatory. When dealing with an explanatory dashboard most especially, make sure your visuals engage your audience with a fascinating story. You want to make it as simple as you can for users to explore the data and draw their own conclusions when in the exploratory mode (Lu, 2022). This would keep them glued to the dashboard without making them bored.

It is also important to note that before designing a dashboard, you should have a dashboard plan. This could be in white and black or on a screen. It will make your work easier and give you directions. Earlier, it was stated that a dashboard should have a purpose. This purpose would also be reflected in your dashboard plans. Hence, your dashboard visuals should be able to answer the questions What? When? How? Where? etc. depending on the data you have available for your analysis.

In addition to this, Key performance indicators should be included in your dashboard. Metrics and KPIs serve as the foundation for many dashboard visualizations since they are the best tools for informing users of their progress toward their goals. Choosing the appropriate metrics and KPIs to present is the most crucial step in developing an effective dashboard. Regardless of how beautiful or brilliant a visual design is, if it isn't presenting insightful data and insights that are pertinent to its audience, it will just wind up being a gorgeous show that nobody notices (Gonzalez, 2019).

As such, all of these and more should be considered when building an effective dashboard. In addition, all the above guidelines were strictly followed when building the dashboard for this project.

**3.0 Exploration of Data Set**

The data set used in this project was gotten from <https://databank.worldbank.org/source/world-development-indicators>. The data set is a time series collected over the years, and I was careful with selecting data that is free of missing observation. This is to ensure there is an accurate result. For this project, a total of 10 years was evaluated, which spanned from 2000 to 2019, and a total of 4 indicators were used, together with key indicators created by me. In addition, 10 countries were carefully selected for the dashboard development. Below are tables showing the list of those countries evaluated (Table 1), and the indicators used (Table 2).

**Table 1: List of Countries in the Dataset**

|  |  |  |
| --- | --- | --- |
| **S/N** | **DEVELOPED** | **UNDERDEVELOPED** |
| 1. | Germany | Afghanistan |
| 2. | Japan | Angola |
| 3. | United States | Bangladesh |
| 4. | United Kingdom | Ethiopia |
| 5. | France | Sudan |

**Table 2: List of Indicators**

|  |  |
| --- | --- |
| **S/N** | **INDICATORS** |
| 1. | Suicide mortality rate (per 100,000 population) |
| 2. | Suicide mortality rate, female (per 100,000 female population) |
| 3. | Suicide mortality rate, male (per 100,000 male population) |
| 4. | Population, total |
| 5. | Year on Year Population growth (Calculated Measure) |

Kindly note that a couple of indicators were imported into the software but only the above listed were used for building the dashboards.

**3.1 Data Transformation and cleaning**

After the data set was downloaded from the site, it was imported into power BI and transformed through power query.

Some of the steps taken to achieve a clean and tidy data set includes:

* Using first row as the header.
* Changing data type to the right ones.
* Removed unwanted columns.
* Unpivoted the year columns.
* Renamed some columns, including the year columns.
* Pivoted the series names, so we can have them as columns rather than rows.
* Final check on the data types and other things to ensure the data set was tidy.
* Closed and applied changes so they can be imported into the power bi fields pane.

**3.2 Data Modelling**

The act of establishing a connection between common columns of various tables in Power BI is known as Power BI data modelling. In this project, before the modelling was done, I created a “Calendar” table. A date column was randomly generated by Power BI using the dates available in the data set with the following Data Analysis Expression (DAX).

Calendar = CALENDARAUTO() … (1).

I also created the Month, Quarter, and Year columns, using the autogenerated date, with the help of DAX.

Month = FORMAT('Calendar'[Date], "mmm yyyy") … (2).

Quarter = YEAR('Calendar'[Date])&"-Q"&FORMAT('Calendar'[Date],"q") … (3).

Year = FORMAT('Calendar'[Date],"yyyy") … (4).

Kindly note that Power BI date tables are tables that solely include data linked to dates in them. It is a typical dimension table that can be used to evaluate data based on dates and serve as a reference for dates in a model. They are especially helpful when developing reports that need exact date information and doing time intelligence calculations which was also done during this project. Hence, the need for creating them.

After this process, the date column, which happens to be the lowest granularity in the Calendar table was dragged to the Year column in the Suicide Data table. And a one-to-one relationship was formed between the 2 tables. Then, the dashboard design process commenced after this stage.

**4.0 Investigation of Data Workflows and Proposal for Design of Dashboard**

Like the popular adage says that if you fail to plan, then you plan to fail. And like I stated in the earlier discussion, it is important to have a dashboard plans before the design starts. Base on my design. I made a dashboard plan based on my aim and objectives. Below is the template of the plan.

Country Slicer Button 1 Button 2 Year Slicer

Objective 1 Objective 4 Objective 2

Objective 5 Objective 3

**Figure 1: Dashboard Plan**

With the help of the dashboard plan, I knew I needed to have another table with key measures, that would have the year-on-year measure. This would make it easier to achieve the 5th objective.

Hence, I created a new table named “Key measures”. In this table, I was able to create the total population measure, the population last year and the year-on-year population, using the following DAX respectively.

Total Population = Sum (‘Suicide data’ [Population, total]) … (5).

Population LY = Calculate ([Total Population], Previous year (‘Calendar’[Date])) … (6).

YOY Population = Divide ((Total Population – Population LY), Population LY) … (7).

Since we needed to get the year-on-year population growth, we need to first ensure the total population we will be using for this is also a calculated measure. Hence, the need to sum up all the total population we initially have in the suicide data to get this calculated measure. For the 6th equation, I calculated the population last year using the calculate in-built function. This function also needed us to use the initially calculated total population and the date that was autogenerated during the modelling process. This is one of the usefulness of having a calendar table. Finally, using the calculated total population and population last year, I was able to use the in-built function ‘Divide’ to get the YOY population measure.

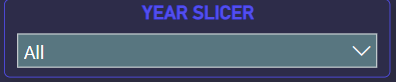
After these calculations, then the visualization started proper. Below are the detailed report of the visuals.

**1. Slicers**: The first visual that was created was the country slicer. This was created to be able to filter the visuals based on the 10 countries present in the data. I dragged the slicer from the visualization pane to the report canvas and then dragged the country name from the field pane to the visual. Then, I renamed the title as country slicer, formatted the visual, and changed the orientation of the slicer to horizontal (drop down), to create more space for other visuals. I also repeated the same process for the year slicer, except that the year field was dragged to the report canvas rather than the country name field. Below is the outcome of the 2 slicers created.

Graphical user interface

Description automatically generated

**Figure 2: Country Slicer**



**Figure 3: Year Slicer**

**2. TreeMap Visual:** Based on the objectives, I needed to show the sizes of each country by plotting the total population by countries. To represent this better, I needed a visual that could show the countries as a part-to-whole. Hence, the reason for settling for Treemap. A tree-structured set of data can be represented using a treemap, which is a rectangle-based visual (Weitz, 2020). The rectangle is divided into sections on a treemap to depict a part-to-whole relationship and as a result, the area is used to represent the size of the entire.

In this visual, I plotted the country names against the total population in the Suicide data table, then formatted the chart appropriately. Each country occupied each rectangle, and the size of the rectangle depends on how large the population is. This means, the biggest rectangle (United States) has the largest population, while the smallest rectangle (Angola) has the smallest population. Below is a screenshot of the final visual from the report.

Chart, treemap chart

Description automatically generated

**Figure 4: Total Population by countries**

**3. Bar Chart and Buttons:** Objective 4 demands that I determine which country has the highest suicide rate and lowest suicide rate. As such, I decided to use bar charts to represent them on the report canvas because bar chart helps to display comparison of various values in categories, which is the case we have here.

Hence, I picked the first bar chart, plotted country names against the total suicide rate and filtered them based on the topmost 5 suicidal countries, then named it “Top 5 suicidal countries”. I repeated the same process for the “Bottom 5 suicidal countries”. But in this case, I filtered based on the countries with the least total suicide rate. But putting these two charts on the report canvas separately would take so much space. As such, to make the charts more interactive and while still creating space for other visuals, I decided to use buttons to reveal each chart as at when needed.

Microsoft (2022) reported in one of their articles that you can design Power BI reports with buttons that function like apps, allowing users to linger over, click on, and engage with the material further. As such, I created two blank buttons “Top 5” and “Bottom 5” respectively, then connected each of them to the respective charts with the help of the Selection and Bookmark tabs (these can be found under the view tab). I was able to achieve this by creating the “Top 5” bookmark on the bookmark page, while the “Bottom 5 suicidal countries” chart was turned off. The same process was also done with the “Bottom 5” bookmark. Then after this, on the format page for each button, I linked each of the created bookmarks to their respective button. I also ensured I place each of the bar charts below the other to create more space for other charts. So, on the report canvas, while holding down on the “ctrl” key on the keyboard, when you click on the “Top 5” button, only the “Top 5 suicidal countries” comes up, while the “Bottom 5 suicidal countries” chart is hidden. This also applies to the “Bottom 5” button and chart too. Figures 5, 6, and 7 below shows the bar charts and the buttons.

Icon

Description automatically generated with low confidence

**Figure 5: Top 5 and Bottom 5 Buttons.**

N.B: The image in-between the button was an imported image to beautify the report.

**A screenshot of a computer

Description automatically generated with medium confidence**

**Figure 6: Top 5 suicidal countries**

**Graphical user interface, application

Description automatically generated**

**Figure 7: Bottom 5 suicidal countries**

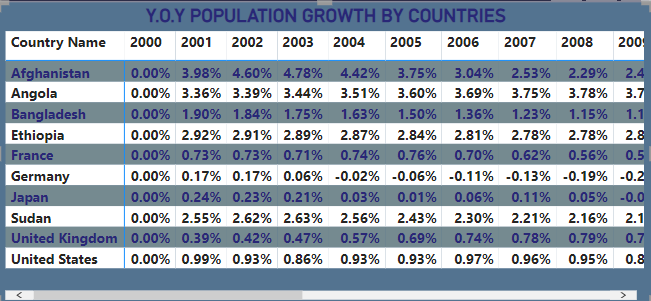
1. **Doughnut Chart:** Just like the tree map, the doughnut chart is a proportion chart which is used to show a part-to-whole visual. As my objective requires that I show the suicide rate proportion between the male and female population, the best chart that could be used to show this is the doughnut chart. Hence, in this project, I dragged the doughnut chart from to the report canvas and selected the suicide rate for male and suicide rate for female. The chart displayed the total figures with each section. So, I formatted the chart so it can display the figures as a percentage of the total inside each section of the doughnut. And here I have it.

Chart

Description automatically generated

**Figure 8: Suicide Rate by Gender**

1. **Matrix Table:** The is like the function of the pivot table in Excel. Power BI's standard table is simply two dimensions. The displayed data are flat, not aggregated, and display duplicate values. A matrix table, however, provides the ability to drill down, aggregate data, support numerous dimensions, and have a stepped layout (**Langmann, 2022)**. Therefore, a Matrix table is used in this project, rather than the standard table. The country names were placed on the rows, years on columns and the year-on-year population as values.

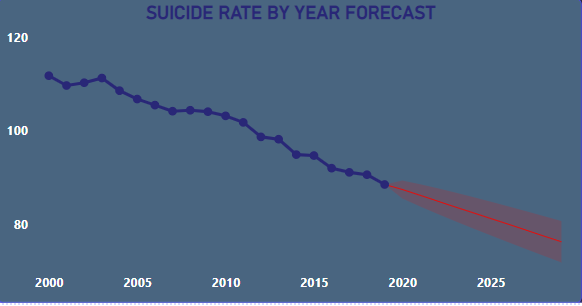


**Figure 9: YOY Population growth table**

N.B: The above only shows between year 2000 and 2008. The scroll bar must be used on the visual to view for 2009 to 2019.

1. **Line Chart:** Line charts are generally used to project time series data. These charts can be used for purposes other than observing change through time. They also aid in bringing out variations and connections in your data. A line chart can also assist a viewer in forecasting potential future events (Tableau, 2022).

For this project, in addition to plotting the total suicide rate for the 10 years period, I also forecasted the suicide rate for the next 10 years so we can plan the next line of action. Below is the line chart plotted in this project.



**Figure 10: Suicide rate by year forecast**

In the figure above, the lines with markers are the actual values spanning from 2000 to 2019. While the red lines are the projected values, spanning from 2020 to 2029. These values also have their upper and lower bound values based on the forecast. The final report was similar to the dashboard plan and the final look can be seen below.

Graphical user interface

Description automatically generated

**Figure 11: Final Report structure.**

This report however has a homepage. The home page comprises of the school logo, the year slicer showing the years we are considering in the report, a button which when clicked takes you to the report page, and KPI’s showing the total population, total suicide rate per 100,000 population and the total number of countries used in the analysis. The report was also given the title; “Comparative analysis of suicide rate between selected developed and Underdeveloped countries”.

A screenshot of a video game

Description automatically generated

**Figure 12: Home page of the report**

**5.0 Discussion of Result**

Starting with the home page, with the help of the card visuals, a total population of 19 billion, 293 million, 757 thousand, and 504 were analyzed in this project. And out of every 100 thousand of this population, the suicide rate was 2030.40 for a period of 10 years.

The first visual on the report page shows that the United States has the highest population of about 6 billion, 136 million, 998 thousand and 594, while Angola has the least population of about 467 million, 260 thousand, and 151, within these 10 years period. That is why the rectangle of the United States looks bigger than every other country in the analysis. Using the year slicer, we can decide to check for the population of a particular year. For instance, if we need the population of these countries in the year 2010, we can easily see there that the United States has a total of 309 million, 327 thousand, and 143 persons, while Angola has 23 million, 356 thousand, and 247 persons in the country in this selected year. We can also filter based on other years too and get the respective values for other countries.

The second visual shows the top and bottom suicidal countries. From these charts, we can see that Japan has the highest number of suicide rates of about 439, based a 100, 000 population, next to France (351), Germany (266), United States (264) and the United Kingdom (164). And for the least suicidal countries, Sudan happens to be the lowest, with a total rate of 81 per 100,000 population, next to Bangladesh (88), Afghanistan (89), Ethiopia (142) and Angola (146). Also, the year slicer here can be used to filter the result based on the preferred year.

Based on these results, it was observed that all the developed countries considered in this project had more suicide rates than the underdeveloped countries, with Japan topping the list.

The fourth visual shows the percent of male and female suicidal rate based on a 100,000 thousand population, within the 10 years period. It was observed that the male gender has the highest suicide rate of about 72.51% within this year, while the female gender has about 27.49% of the suicide rate. The difference is high and the reason for this needs to be further researched on. The country slicer and year slicer can be used on this chart based on the information needed.

The fifth visual shows the year-on-year population growth of each country. These are rates of change expressed over the same period, (depending on how frequently the data are collected) in the prior year (OECD, 2005). Hence, for our visual, we can see that we have the YOY population growths in percentage. That is, it is showing us how much the population has increased between the last year and the current year, in percentage. We can also observe that some countries have a decreased population in comparison to the previous years. For instance, Afghanistan has an increased population of about 4.60% in 2002 compared to a 3.98% in 2001. That’s about 0.62% increase in population. However, Japan on the other hand has been having a negative population growth over this 10 years period. This is not surprising has it topped the list of the most suicidal countries in our previous visual.

And for the final visual, the line chart shows a positive slope in the suicide rate over the years. This means from the year 2000 to the year 2019, the suicide rate has generally been reducing.

I also extrapolated this rate over ten years (2020 to 2029), and observed that a further improvement in the suicide rate would happen, as the suicide rate would decrease from a total of 88.5 in 2019 to about 87.43 in 2020. This decline will continue until 2029 with a total rate of about 76.24. This result shows an improvement in the suicide rate and with time, the rate would have been at its lowest value.

**6.0 Conclusions**

After series of analysis done on this data, the following inference and conclusions can be made.

1. Based on the selected 10 countries and the 10 years of analysis, the United States have the highest population of about 6 billion plus, while Angola has about 467 million plus population.
2. Japan has the highest suicide rate among all the countries while Sudan has the lowest suicide rate.
3. The male gender tends to attempt suicide than the female gender.
4. The year-on-year population growth of each country varies. However, the poor population growth in Japan could be linked to the high rate of suicide in the country.
5. There would be less attempt to commit suicide in the coming years as the total rate continues to decline in the forecast.

**6.1 Recommendation**

Further analysis should be carried out on why the male gender has a high suicide rate than the female gender. In addition, I would also recommend that we find out why there is high suicide rate in Japan.

**PART TWO: STATISTICAL ANALYSIS**

**1.0 Introduction**

A brief explanation about the effect of suicide on the economic growth have been made in the part one of this write up. However, in this part of the project, I used the R analytical tool to explore the same suicide data set that was used in the part one of the projects. Here, I did a descriptive analysis of the whole data set and that of each country by reporting the mean, median, mode, standard deviation, skewness, and kurtosis of the data. I also did correlation analysis to see if there is any relationship between the variables and Regression analysis to determine mathematically which of those variables has an effect.

As for the time series analysis, I used a population growth data set, where I investigated the population growth of Japan, which happens to be the most suicidal country in my previous report. This was done for a period of 25 years, then I forecast into the next 5 years to see the trend of the population growth.

And finally, using the first data set, I did a comparative analysis between the developed and underdeveloped countries, to see if there is a significance difference between the suicide rate of the two groups. This was done using the t-test

* 1. **Objectives**

The aim of this project is to compare the suicide rate between some selected developed and underdeveloped countries. objectives of this project include.

1. To explore the descriptive properties of the Suicide data set.
2. To perform correlation analysis between the variables of the data set.
3. To determine which of this variable has more significance using the Regression analysis.
4. To forecast the trend in population growth using the time series analysis.
5. To compare the suicide rate of the developed and underdeveloped countries using t-test.

**2.0 Background Research**

The gathering and evaluation of data for statistical analysis is done to find patterns and trends (TechTarget, 2022). It is a part of data analytics that is beneficial for businesses that need to work with big amounts of data and business intelligence. Farnsworth (2019) stated 7 statistical tools that could be used to achieve a good analysis. They are SPSS, R, MATLAB, Microsoft Excel, SAS, GraphPad Prism, and Minitab. Of all the above listed, my focus in this report will be on R analytical tool. R has in-built functions that helps ease the process of our analysis. And based on the objectives of my project, I will streamline my report based on the descriptive statistics, Correlation Analysis, Regression analysis, time series analysis and using the t-test, all of which are available on this tool for the analysis of my data set.

**2.1 Descriptive analysis**

A subset of statistics called descriptive statistics aims to summarize, describe, and present a dataset or set of values (Soetewey, 2020). In every statistical investigation, descriptive statistics are frequently the first step and a crucial component. By providing a clear overview of the data, it enables quality control of the data and aids in understanding the data. Descriptive statistics are useful place to start for additional analysis if they are presented correctly. These statistics can be described numerically or graphically, and are broadly categorized two categories, the measure of central tendency and the measure of dispersion.

The center point or typical value of a dataset is represented by measures of central tendency (Frost, 2018). On the other hand, Dispersion measures provide insight into how the data are distributed (Soetewey, 2020). Common examples of the measures of central tendency are the mean, median and mode, while an example of the measure of dispersion is the standard deviation.

The mean is the sum of values of all variables divided by the total number of values, the middle value is median, while the mode is the most occurring value.

Another measurement that could be considered is the distribution characteristics of the data set. Examples of this is the Skewness and Kurtosis of the data set. Skewness is a metric for a distribution's asymmetry. Asymmetry is when the left and right sides of a distribution are not mirror reflections of each other (Turney, 2022). Kurtosis on the other hand is a metric that indicates how heavy-tailed or light-tailed the data are in comparison to a normal distribution. In other words, data sets with a high kurtosis tend to have large outliers or heavy tails. Data sets with low kurtosis frequently lack outliers or have light tails. All these measures were considered in this report.

**2.2 Correlation Analysis**

This studies the relationship between quantitative or categorical data. There are different methods of describing relationship between variables, however, in this report, our focus is on the Pearson correlation method. Pearson evaluates the linear relationship between two variables, and it is the most used method. How closely the observations resemble a line of best fit is shown by the coefficient (r).

Also, correlation is said to be positive if its linear relationship goes uphill, while it is negative if it goes downhill (Rumsey, 2021). And according to Statistics solutions (2022), the degree of correlation could fall into any of this category.

1. If it has an absolute value of 1, it is perfectly correlated.
2. If the coefficient lies between an absolute value of 0.5 and 1, the correlation is high.
3. If the coefficient lies between absolute value of 0.3 and 0.45, the correlation is medium
4. If the correlation is less than 0.29, then it is low
5. If the value is zero, then there is no correlation

However, in this project, I assumed that any value greater than 0.8 is highly correlated. Hence, I dropped those values.

**2.3 Regression Analysis**

Understanding how dependent variables change when one of the independent variables changes while the other independent variables remain constant is made easier using regression analysis. There are different types of linear regression. In these studies, we are focusing on the Multiple Linear regression (Ordinary Least Square-OLS). OLS is employed to determine how two or more independent variables relate to one dependent variable. There are certain assumptions that must be in place to be able to carry out OLS.

1. The residuals should be normally distributed.
2. The residuals are independent (that is, not autocorrelated).
3. Relationship between the dependent and independent variables are linear.
4. Residuals should have a constant variance (Homoskedaskicity).

After creating a linear regression model, we have created a new fact about the relationship between the dependent variable and one or more independent variables. One or more tests are required to support the truth. The act of testing the new fact is called hypothesis tests. Hence, to carry out the hypothesis test, we must follow the following steps (Kumar, 2022).

1. Define the null hypothesis H0: This is usually set that there exists no relationship between the dependent and independent variables.
2. Define the Alternative hypothesis H1: This is usually set that atleast one dependent variable is significantly associated with the independent variable.
3. Define the level of significance (α): the level of significance is the probability of committing Type I error. Type I error occurs when you reject the null hypothesis when it should not be rejected. The default value is usually 5%.
4. Decision Rule: this states that the null hypothesis should be rejected if the p-value <= (chosen level of significance). Otherwise, do not reject.
5. Make Conclusion about the test.

Some terms related to the results needs to be defined.

**Residuals**: This is the difference between the observe values and the predicted values. When we have a regression line, not all points fall on this line. The vertical separation between the data point and the regression line is known as a residual, and each data point has a residual. It is positive if the data point is above the regression line, negative if its below it and zero if its on it.

**Standard Error of Regression:** an estimation of the error term's standard deviation in a regression model.

**T-value**: it is a measure of an independent variable's statistical importance in explaining the dependent variable. It can be calculated by dividing the estimated regression coefficient by its standard error.

**P-value**: In statistical tests, the P value is a probability score that is used to determine the statistical significance of an observed effect (Prabhakaran, 2019).

**Signif. Codes**: States the level of significance of the variables. Anyone that has at least one asterisk is significant.

**R-squared**: In a regression model, R-Squared (coefficient of determination) is a statistical metric that quantifies how much of the variance in the dependent variable can be accounted for by the independent variable. R-squared, thus, displays how well the data match the regression model (the goodness of fit).

**F-statistics**: This is a type of hypothesis test that tests the significance of a model. It tells whether the model is good or not.

**2.4 Time Series Analysis**

Data acquired over time for a single entity is known as a time series. Analysis can be done on such data to forecast into the future with it. Remember one of the assumptions of our correlation is that residuals cannot be autocorrelated. If they are, they can’t be used for F-test, and a load of other problems arises, including making it useless for time series analysis. Hence, to fix this autocorrelation problem, a model called ARMA (Auto-Regressive moving average) was introduced. But ARMA cannot be used if the time series has a trend, that is, it is not stationary. If the time series shows this tendency, the trend is eliminated by applying one or more orders of differencing to the time series. This is what brought about ARIMA. The I stand for integrated.

After this is done, a formal test is carried out to confirm the stationarity of the data this is called ADF (Augmented Dickey-Fuller). If a series is stationary but uncorrelated, this type of time series is called a Random walk. And to test for this, the Ljung-Box test can be used.

To know if the ARIMA model fits,

1. The residuals will be normally distributed.
2. The autocorrelation of the residuals would be zero for all possible lags (that is, white noise).

A Jarque bera test could be done to confirm this. If the residual is normally distributed, then the Jarque bera test should be insignificant. This means the p-value would be greater than 0.05.

Then a forecast can be done on the model once the model is good.

**2.5 Comparative Analysis**

Depending on what the objective of a project is, one can compare two or more groups in a data set using different types of tests to see the significant difference between them. The most common types of tests done are 2. These are the sample t-test and ANOVA (analysis of Variance). To determine whether there is a statistically significant difference between the means of two unrelated groups using an inferential statistical test, the student t-test is used while ANOVA is a method of comparing the means of three or more groups statistically (Mishra et al., 2019). In this report, the independent t-test was used.

Before the t-test can be done, a test of equal-variance should be carried out. There are different types of tests that could be done. But here, I deployed the use of Bartlett test. If the P value of the test is greater than 0.05, it means the variance is homogenous and it should set to true in the t-test. However, if the p value is less than 0.05, it means the variance is non-homogeneous and it should be left at the default state (false).

All the above analysis were carried out in this project.

**3.0 Exploration of Data Set**

Two sets of Data sets were used throughout this analysis. The first is the suicide data, which was used in the previous report. This data was carefully selected to ensure there are no missing data. It has a total of 10 countries, with 4 indicators. And it has 100 observations in all. For each of the countries, I extracted the data between the year 2000 and 2019. This data was used to achieve Objectives 1, 2, 3 and 5. Below is the list of the indicators, and their definitions.

**Table 1: list of Indicators 1**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S/N** | **Indicator** | **Definition** | **Time frame** | **Source** |
| 1. | Suicide mortality rate (per 100,000 population) | This is the rate of death due to suicide. The value was based on a 100,000 population. | 2000-2019 | <https://databank.worldbank.org/source/world-development-indicators> |
| 2. | Suicide mortality rate, female (per 100,000 female population) | This the rate of death of females due to suicide. The value was based on a 100,000 female population. | 2000-2019 | <https://databank.worldbank.org/source/world-development-indicators> |
| 3. | Suicide mortality rate, male (per 100,000 male population) | This the rate of death of males due to suicide. The value was based on a 100,000 male population | 2000-2019 | <https://databank.worldbank.org/source/world-development-indicators> |
| 4. | Population, total | The total number of inhabitants of a country. | 2000-2019 | <https://databank.worldbank.org/source/world-development-indicators> |
| 5. | Population, female | The total number of female inhabitants of a country. | 2000-2019 | <https://databank.worldbank.org/source/world-development-indicators> |
| 6. | Population, male | The total number of male inhabitants of a country. | 2000-2019 | <https://databank.worldbank.org/source/world-development-indicators> |
| 7. | % Pop growth | The rise in a population's size expresses in percentage. | 2000-2019 | <https://databank.worldbank.org/source/world-development-indicators> |

No special preparation was done to the data as it has been done in the previous report. However, I took a quick glimpse of the data set to see its specifications using equation one below and double checked if there are no missing observations in the data set using equation 2. I also removed the country name column and the year column as they were not needed for this analysis.

glimpse(suicide\_data) … (1).

colSums(is.na(suicide\_data)) … (2).

The data set used for the time series analysis however has just 3 variables: the time, country and the population growth. I specifically picked 25 years (between 1997 – 2021) data of the population growth of Japan, as it is the most suicidal country based on the previous report. And I am interested in seeing the effect of this high suicide rate on the population growth. Below is a list of the indicators used.

**Table 2: List of Indicators 2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S/N** | **Indicator** | **Definition** | **Time frame** | **Source** |
| 1. | Country Name | The country being considered in this analysis (Japan) | 1997-2021 | <https://databank.worldbank.org/source/world-development-indicators> |
| 2. | Time | The year being considered. This is between 1997 and 2021 | 1997-2021 | <https://databank.worldbank.org/source/world-development-indicators> |
| 3. | % Pop growth | This is the population growth of Japan between these years in percentage. | 1997-2021 | <https://databank.worldbank.org/source/world-development-indicators> |

Using the power query editor in Microsoft excel, I renamed the population growth column, unpivoted & edited the year column, and deleted irrelevant data before importing it into R studio using the read.csv command. In addition to this, in the R studio, I took a quick glimpse of the data just as done in equation one and selected the 2 columns I am interested in working with by dropping out the country column and named it “pop\_data” using equation 3 below.

Pop\_data<-

Population\_data %>%

select (Time, X..Pop.growth) … (3)

Then the analysis began after the data preparation.

**4.0 Analysis**

Based on the 5 objectives I intend to achieve in this project, I performed 5 different analyses using the R statistical tool. I will be reporting each of the analysis one after the other below.

**1. Descriptive Statistical Analysis**

Before any other analysis was done, I described the sample of the data to be used. Here, I used the numerical method by describing the data using measure of central tendency (mean, median and mode), and measure of dispersion (standard deviation). Also, using the central moments, I examined the sample's distributional characteristics (skewness and kurtosis). This was also done for each country too. Below is the result of the descriptive statistics of the whole data and for just a country out of the 10 countries.

**Table 3: General Descriptive Statistics**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Indicators / Measures*** | **% Pop growth** | **Population, Female** | **Population, Male** | **Population, Total** | **Suicide rate, total** | **Suicide rate, Female** | **Suicide rate, male** |
| **n** | 100.0000000 | 1.000000e+02 | 1.000000e+02 | 1.000000e+02 | 100.0000000 | 100.0000000 | 100.0000000 |
| **Mean** | 1.4406426 | 5.111335e+07 | 5.030258e+07 | 1.014159e+08 | 9.5370000 | 5.2340000 | 13.9810000 |
| **Median** | 0.9474496 | 3.780085e+07 | 3.619037e+07 | 7.376195e+07 | 7.4500000 | 3.8500000 | 11.5500000 |
| **Mode** | NA | NA | NA | NA | 4.0000000 | 2.8000000 | 5.1000000 |
| **Standard devaition** | 1.2718441 | 4.188161e+07 | 4.111890e+07 | 8.298762e+07 | 5.6210042 | 3.1067935 | 8.4424935 |
| **Skewness** | 0.2230890 | 1.670985e+00 | 1.646796e+00 | 1.659423e+00 | 0.6405889 | 1.0336429 | 0.4652077 |
| **Kurtosis** | -1.1160765 | 2.002820e+00 | 1.884747e+00 | 1.946477e+00 | -0.7800015 | 0.2245032 | -1.0721435 |

**Table 4: Descriptive statistics for the United Kingdom**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ***Indicators / Measures*** | **% Pop growth** | **Population, Female** | **Population, Male** | **Population, Total** | **Suicide rate, total** | **Suicide rate, Female** | **Suicide rate, male** |
| **n** | 10.0000000 | 1.000000e+01 | 1.000000e+01 | 1.000000e+01 | 10.0000000 | 10.0000000 | 10.0000000 |
| **Mean** | 0.70666296 | 3.290189e+07 | 3.195204e+07 | 6.485393e+07 | 8.32000000 | 3.9800000 | 12.8200000 |
| **Median** | 0.71590856 | 3.290077e+07 | 3.195849e+07 | 6.485926e+07 | 8.25000000 | 4.0000000 | 12.9000000 |
| **Mode** | NA | NA | NA | NA | NA | 4.1000000 | NA |
| **Standard devaition** | 0.07825871 | 6.485171e+05 | 7.430732e+05 | 1.391580e+06 | 0.34896673 | 0.2149935 | 0.6811755 |
| **Skewness** | -0.48399996 | -2.787575e-02 | -4.180897e-02 | -3.530443e-02 | 0.04743932 | 0.2922270 | 0.1073702 |
| **Kurtosis** | -1.27593099 | -1.614343e+00 | -1.609231e+00 | -1.611628e+00 | -1.51410810 | -0.8879789 | -1.2922919 |

It is worth noting that for all the 10 countries, the report for the United Kingdom alone was shown here as they all have similar interpretations.

From the tables above, n represents the number of samples from the population. Reporting the % pop growth indicator for the 10 countries, table 3 shows the mean of the whole data as 1.4406426, the median as 0.9474496, and there was no most occurring value. The standard deviation of the sample was 1.2718441, Skewness was 0.2230890 and Kurtosis was -1.1160765. On the other hand, table 4 shows the descriptive statistics for the United Kingdom alone. The total population indicator of this country has a mean of 6.485393e+07, median of 6.485926e+07, with no mode as well. The standard deviation was 1.391580e+06, Skewness was -3.530443e-02 and Kurtosis reads -1.611628e+00.

**2. Correlation Analysis**

Here, we want to see how all the variables are related to each other. Using the in-built R function “cor”, I got the correlation matrix for all variables, and I also checked if the correlation is significant using the psych package in R. Below is the correlation matrix for the data.

**Table 3: Correlation matrix**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **% Pop growth** | **Population, Female** | **Population, Male** | **Population, Total** | **Suicide rate, total** | **Suicide rate, Female** | **Suicide rate, male** |
| **% Pop growth** | 1 | -0.44 | -0.42 | -0.43 | -0.74 | -0.7 | -0.74 |
| **Population, Female** | -0.44 | 1 | 1 | 1 | 0.39 | 0.26 | 0.42 |
| **Population, Male** | -0.42 | 1 | 1 | 1 | 0.36 | 0.24 | 0.39 |
| **Population, Total** | -0.43 | 1 | 1 | 1 | 0.37 | 0.25 | 0.41 |
| **Suicide rate, total** | -0.74 | 0.39 | 0.36 | 0.37 | 1 | 0.96 | 0.99 |
| **Suicide rate, Female** | -0.7 | 0.26 | 0.24 | 0.25 | 0.96 | 1 | 0.93 |
| **Suicide rate, male** | -0.74 | 0.42 | 0.39 | 0.41 | 0.99 | 0.93 | 1 |

The above matrix clearly shows that the diagonals are perfectly correlated. We can also see some variables with a perfect correlation, such as the population of male & female, population of female & total, and population of male and total. For others, some are highly correlated, medium correlated and low correlation. None of the variables has no correlation.

Also using the corrplot, I plotted the correlation which can be found in figure 1 below.

Chart, bubble chart

Description automatically generated

**Figure 1: Correlation plot**

As we can see on the plot above, the deeper the color, the more correlated the variables are to each other. The diagonal is perfectly correlated and that is why we can see the color matrix is the deepest.

**3. Regression Analysis**

The aim of the regression analysis was to see which among all independent variables was significantly associated with the dependent variable (% pop growth). But before starting the linear regression, I did a scatter plot matrix just as shown in figure 2 below.

Diagram

Description automatically generated

**Figure 2: Scatter plot matrix**

This matrix openly showed me further that some of the indicators have perfect correlation. And if we go on with our regression with such variables, we won’t have a true result. Hence, I defined the high correlated variables and set the cut off to 0.8. Then I selected the assumed “uncorrelated” variables, that is, variables less than 0.8, then column bind it with the dependent variable. Hence, I had just three variables with hundred observations for the regression analysis. The variables are % pop growth, the population of male and the total suicide rate.

Using the “lm” in built function, I built a linear regression model between the new selected independent variables named df\_regression and the % pop growth. Then I did a summary of the new model just as seen below.

**Table 4: Regression Summary**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Residuals:** | | | | | | | | | |
| **Min** | | **1Q** | **Median** | | | **3Q** | | **Max** | |
| -2.79735 | | -0.67346 | -0.05882 | | | 0.63506 | | 1.63082 | |
| **Coefficients:** | | | | | | | | | |
|  | **Estimate** | | | **Std. Error** | | | **t value** | | **Pr(>|t|)** | |
| **Intercept** | 3.170e+00 | | | 1.755e-01 | | | 18.062 | | < 2e-16 | |
| **Population, Male** | -5.522e-09 | | | 2.206e-09 | | | -2.503 | | 0.014 | |
| **Suicide Rate, Total** | -1.522e-01 | | | 1.614e-02 | | | . -9.429 | | 2.29e-15 | |
| **Residual standard error** | | | | | 0.842 on 97 degrees of freedom | | | | |
| **Multiple R-squared** | | | | | 0.5705 | | | | |
| **Adjusted R-squared** | | | | | 0.5617 | | | | |
| **F-statistic** | | | | | 64.43 on 2 and 97 DF | | | | |
| **p-value** | | | | | < 2.2e-16 | | | | |

In the table above, we can see the difference between the observed values and the predicted values (residuals). Also, the standard error, t-value and P-value for each indicator is in the table. The adjusted R-square tells us the amount of variation in % population growth that is jointly explained by other variables. Also, the F-statistics is also stated in the table above.

Below is the linear model plot which properly explains the assumptions of the linear regression.

Chart

Description automatically generated

**Figure 3: Regression plot**

The figure above shows the residual vs fitted plot, Normal Q-Q plot, scale-location, and residual vs leverage plot. On the Q-Q plot, close alignment of some data points with the dotted line denotes a normal distribution. Also, the residuals vs fitted plot demonstrates if residuals are distributed similarly across the predictor ranges. The scale-location demonstrates the Homoskedaskicity assumptions. And finally, the residuals vs leverage reported how independent the residuals are.

**4. Time Series Analysis**

The objective of this analysis is to forecast the population growth of Japan as it happens to be the most suicidal countries of all the 10 countries assessed. As such, I loaded the forecast, tseries and the xts libraries, then loaded the population growth data set and finally took a glimpse of the data. I also selected just the time variable and the population variables, as those are the only variables needed here.

Before this data set was used, I defined the date structure, by specifying the start and end dates, and the unit of consideration, which was in years. Then with the help of the xts library, I converted the data to a time series object. I also did an autoplot of the new object just as seen below.

Chart, line chart

Description automatically generated

**Figure 4: Time series Object plot.**

The plot above shows a downward trend of the population growth. But we can’t use it to forecast as it is because it is not stationary (that is, the mean, covariance and autocovariance aren’t stable). This obvious in the trend here. To make it stationary, I introduced the ARIMA model. Using the ndiff function, I calculated the number of differencing as 1. Hence, I differenced once, and I plotted the differenced data to confirm if the series is now stable.

Chart, line chart

Description automatically generated

**Figure 5: Stable Time series**

The Autocorrelation plot (ACF) was also checked for significant lag. The third confirmatory test for stationarity was the Augmented Dickey-Fuller (ADF) test, while setting the lag order, k, to zero. This was because we had no significant lag in the ACF. ACF plot can be found in figure 6.

Having observed that the shape of the series is random, I realized I might be in the presence of a random walk time series. Hence, I confirmed this with the Ljung-Box and the auto.arima test.

Before the forecast, I checked the ARIMA model for goodness of fit using the “checkresiduals” function. A normally distributed residual and zero autocorrelated residual tells if the model is good or not. 3 plots and a Ljung-Box test result was gotten from the residual check. A Jarque Bera Test was also used to confirm the normality of the residual.

Chart, line chart

Description automatically generated

**Figure 6: Residual plots**

**Table 5: Ljung-box test**

|  |
| --- |
| **Ljung-Box test** |
| data: Residuals from ARIMA (0,1,0) with drift |
| Q\* = 1.8108 |
| df = 5 |
| p-value = 0.8747 |
| Model df: 0 |
| Total lags used: 5 |

**Table 6: Jarque Bera Test**

|  |
| --- |
| **Jarque Bera Test** |
| data: fit$residuals |
| X-squared = 3.4404 |
| df = 2 |
| p-value = 0.179 |

After confirming the goodness of the ARIMA, I went ahead to forecast for a period of 5 years and plotted the forecast, which shows a downward trend of the time series.

Chart, line chart

Description automatically generated

**Figure 7: Forecast of the Population growth time series.**

**5. Comparative Analysis**

In this analysis, the same suicide data used in part one of this project was also used. My aim was to check if there is a significance difference between the underdeveloped countries group and the developed countries. I started by getting a unique list of all countries in this data and then grouping them into developed and underdeveloped countries. Then I classed them and merged the classification with the existing suicide data.

In this project, the independent t-test was used instead of the ANOVA because we are comparing just 2 groups. But before the test commenced, I performed a Bartlett test to check for the homogeneity of variances. Here, the null hypothesis is that the variances are equal while the alternative hypothesis is that the variance are not equal. And based on the result of the test, when performing the t-test, the var.equal was left at its default, which is false.

**Table 7: Barlett test Result**

|  |
| --- |
| **Bartlett test of homogeneity of variances** |
| data: suicide\_data$Suicide.Rate..per.100k.Total.Pop. by suicide\_data$classification |
| Bartlett's K-squared = 59.478 |
| df = 1 |
| p-value = 1.237e-14 |

For the t-test, the dependent variable was the suicide rate while the independent variable was the new classification column (Table 8). I confirmed how significantly different the 2 groups were by summarizing them using the mean, standard deviation, and a 95% confidence interval (Table 9). And finally, a plot of the mean values and the classification was done as a graphical representation helps to properly explain a data (Figure 8).

**Table 8: Sample t-test result.**

|  |
| --- |
| **Welch Two Sample t-test** |
| **data**: suicide\_data$Suicide.Rate..per.100k.Total.Pop. by suicide\_data$classification |
| t = 15.89 |
| df = 57.424 |
| p-value < 2.2e-16 |
| **alternative hypothesis**: true difference in means between group dev and group underdev is not equal to 0 |
| 95 percent confidence interval:   |  |  | | --- | --- | | 8.297782 | 10.690218 | |
| **sample estimates:**   |  |  | | --- | --- | | mean in group dev | mean in group underdev | | 14.284 | 4.790 | |

**Table 9: Summary of significant difference**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification** | **n** | **mean** | **sd** | **ci** |
| Dev | 50 | 14.3 | 4.05 | 1.15 |
| undev | 50 | 4.79 | 1.19 | 0.339 |

**Chart, line chart, scatter chart

Description automatically generated**

**Figure 8: Plot of mean vs classification**

**5.0 Discussion of Result**

I will be discussing the results based on each objective.

**1. Descriptive Analysis**

From table 3, it was observed that the mean of the data ranges from 1.44 to 1.01 e+08, the median from 0.95 to 7.38 e+07 and about 4 out of the 7 indicators has no mode. The standard deviation, which shows how evenly distributed numbers are ranges from 1.27 to 8.30 e+07. It can also be observed that the male population, female population, total population, and female suicide rate are substantially skewed as their skewness is greater than positive 1. Also, analyzing the Kurtosis of male population, female population, and total population, their distribution is too peak as they have a kurtosis greater than positive 1. While the total suicide rate has a flat distribution because it has a kurtosis less than. Negative of 1. There was none of the values that have both skewness and kurtosis to be equal to 0. Hence, none of the indicators single handedly showed a normal distribution. However, the total suicide rate has its skewness and kurtosis ranging between -1 and 1. Hence, it is near normal distribution.

This same explanation goes for each of the countries too as they have a similar result.

**2. Correlation Analysis**

Looking at the correlation matrix in table 3, all diagonals are perfectly correlated. Other values such as the male & female populations, female & total populations, and male & total populations are also perfectly correlated to each other. For the medium correlation, total suicide rate, female suicide rate, and male suicide rate are all negatively correlated to the population growth. This means that as the population growth reduces, while keeping other values constant, the total suicide rate, female suicide rate and male suicide rate increases. This is expected as there are more deaths, the population of a country would reduce. And for the low correlation, we can see that the female population, male population, and total population are also negatively correlated with the population growth, but it Is a low correlation. This means these indicators are not strongly correlated with the population growth.

The above explanation also applies to the correlation plot. The deeper the color, the highly correlated the indicators are to each other, and vice versa.

**3. Regression Analysis**

The scattered plot matrix further confirmed how linearly correlated these indicators are. The essence of the linear regression is to know which variables exactly contributes to the population growth, not removing the perfectly correlated variables would give me a biased result. Hence, the reason for cutting them out of the data used.

From the result of the linear regression on table 4, the regression coefficients indicate the increase in the dependent variable for a unit change in a predictor variable, holding all other predictor variables constant. This further means that if the total suicide rate increases by 10, the population growth would reduce by -1.522e-01 \* 10 which is -1.522. This further buttress how significantly related the suicide rate is to the population growth. In addition to this, the signif. Codes tells how highly related the indicators are, with the 3-asterisk showing a very high correlation and one asterisk showing a low correlation.

Also reporting the adjusted R-squared, bout 56% variation in population growth is jointly explained by the independent variables. And for the p value, since value is less than 0.05, it means the 2 indicators are significantly related to the population growth.

I also tested for hypothesis, where I checked whether the model I built was significant or not.

1. I defined my null hypothesis that the regression model I built wasn’t significant.
2. I defined the alternative hypothesis that the regression was significant.
3. I chose my level of significance as 0.05.
4. My F-statistics based on my model was 64.43 on 2 variables.
5. The decision rule is that I would reject the null hypothesis if p-value was less than or equal to the level of significance.
6. Since the p-value of my model was less than the level of significance (< 2.2e-16), hence, I have the statistical reason to reject the null hypothesis and conclude that my regression model was significant.

And finally, to confirm the assumptions of linear regression, the Q-Q plot in figure 3 tells how normally distributed the residuals are. All points close to the straight line are normally distributed while points 11, 13 and 52 are outliers. That is, their predicted value is far from the true value. This eventually means that, values above 1 and values below -3 are considered as outliers in this plot.

And for the residual vs fitted plot, we can see that the values above and below the zero dotted line are approximately equal. This shows the linearity in the relationship.

The scale-location plot shows the homoskedaskicity of the residuals. As we can see that the plot has roughly same number of points above and below the line, this means the data set meets the assumption of constant variance. And the last lot confirms that the residuals are not autocorrelated.

**4. Time Series Analysis**

Figure 4 shows a downward trend of the population growth. This means the population rowth of Japan was decreasing throughout the 25 years of the data used. This also buttress the point that the suicide rate was highly correlated to the population growth, as Japan has the highest suicide rate.

Figure 5 on the other hand shows a stationary but uncorrelated series due to the upward and downward trends. Hence the reason I carried out tests to confirm if its truly a random walk time series or not. With the p-value of the Ljung-box test showing a value of 0.8747, which is greater than 0.05 means the data is uncorrelated. The auto.arima test also shows an ARIMA of (0, 1, 0), which is also a sign of a random walk time series.

While checking the normality of the residuals as this tells us if we have a good model or not, the Jarque bera test shows that the model was insignificant. This was because the p-value (0.179) was greater than 0.05. Hence, with this, we can say our model is normally distributed. This is evident on the residual plot in figure 6. Also, the p-value of the Ljung-box test shows the model was uncorrelated as the p-value was greater than 0.05.

In addition to this, the lag image on figure 6 (the autocorrelation of the residuals) has no significance as there is no peak. Meaning, none of the lines crossed the dotted blue lines above and below the graph.

And finally, the 5 years forecast of this time series still shows a downward trend of the population growth. This implies that from 2022 to 2026, there will still be a decline in the population growth of Japan.

**5. Comparative Analysis**

From the Barlett test result, we can see that the p-value was far below 0.05, hence, the reason we have a non-homogeneous variance. And this means that the variance for the two groups is not equal.

And from the t-test result in table 8, the p-value was less than 0.05. this means the difference in the suicide rate between the developed country group and underdeveloped country group is significant. This is also evident in their mean values as the underdeveloped group has a mean value of 4.790 and the developed group has a mean value of 14.284. Their standard deviation respectively was 1.19 and 4.05, while their confidence interval was 0.339 and 1.15 respectively.

Looking at the plot of their mean, it is so evident as the mean of the developed group stays at the top of the graph while that of the underdeveloped group stays at the bottom. This means that there are more suicides being committed in the selected developed countries compared to the underdeveloped countries.

**6.0 Conclusion**

Based on the above results and discussions, the following conclusions was made.

1. Based on the descriptive analysis done, the kurtosis and the skewness of the overall data shows none of the indicators is normally distributed as none of them had a value of zero for both skewness and kurtosis. However, total suicide rate is near normal distribution.
2. Negatively high correlation between the suicide rates and the population growth shows that as the suicide rate increase, the population growth reduces.
3. We can confidently conclude that suicide rate in particular has a significant effect on population growth as all indicators points toward this.
4. Despite the large number of years considered, the population growth of Japan, being the most suicidal country keeps declining. This was also evident in the 5 years forecast.
5. There is a significant difference between the suicide rate in developed countries compared to the underdeveloped countries.

**PART THREE: REFERENCES**

**References**

Farnsworth, B. (2019). The top 7 statistical tool you need to make your data shine. <https://imotions.com/blog/learning/best-practice/statistical-tools/#:~:text=SPSS%20(IBM)&text=SPSS%20offers%20the%20ability%20to,graphical%20user%20interface%20(GUI)>.

Frost, J. (2018). Mean, Median and Mode: Measures of Central tendency. <https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/>.

# Geckoboard (2022). How to design and build a great dashboard. <https://www.geckoboard.com/best-practice/dashboard-design/>.

# Gonzalez, T. (2019). Dashboard design: key performance indicators and metrics. https://towardsdatascience.com/dashboard-design-key-performance-indicators-and-metrics-2b13745f5b2f .

Insightsoftware (2021). 5 Design Tips to Enhance Your Dashboard Colors. <https://insightsoftware.com/blog/5-design-tips-to-enhance-your-dashboard-colors/>

# Kumar, A. (2022). Linear regression hypothesis testing: Concepts, Examples. <https://vitalflux.com/linear-regression-hypothesis-testing-examples/>

# Langmann, K. (2022). **How to Visualize Data in a Matrix in Power BI (Easy Guide).** <https://spreadsheeto.com/power-bi-matrix/>**.**

Lu, S. (2022). Storytelling in Dashboards. <https://www.susielu.com/data-viz/storytelling-in-dashboards>.

# McGuigan, C. (2020). **Designing compelling dashboards: 10 tips for more powerful designs**. <https://uxplanet.org/designing-compelling-dashboards-10-tips-for-more-powerful-designs-298d94db1b13>.

Microsoft (2022). Create Buttons in PowerBI Reports. <https://learn.microsoft.com/en-us/power-bi/create-reports/desktop-buttons?tabs=powerbi-desktop>.

# Mishrah, P., Singh, U., Pandey, C. M., Mishra, P., & Pandey, G. (2019). Application of Student's t-test, Analysis of Variance, and Covariance. Annals of Cardiac Anaesthesia. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6813708/#:~:text=The%20Student's%20t%20test%20is,mean%20difference%20was%20statistically%20significant>.

Organisation for Economic co-operation and development (2005). Glossary of Statistical Terms. <https://stats.oecd.org/glossary/detail.asp?ID=6673>.

# Prabhakaran, S. (2019). What is P-Value? – Understanding the meaning, math and methods. <https://www.machinelearningplus.com/statistics/what-is-p-value/>.

Ritchie, H., Roser M., & Ortiz-Ospina E. (2015). Suicide. https://ourworldindata.org/suicide [Online Resource].

Rumsey, D.J. (2021). How to interpret a correlation coefficient r. Statistics for Dummies. <https://www.dummies.com/article/academics-the-arts/math/statistics/how-to-interpret-a-correlation-coefficient-r-169792/>.

Soetewey, A. (2020). Descriptive Statistics in R. <https://statsandr.com/blog/descriptive-statistics-in-r/>.

Statistics Solutions (2022). Pearson’s Correlation Coefficient. Complete Dissertation. <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/pearsons-correlation-coefficient/>.

Subotin, S. (2017). Dashboard designs: consideration and best practices. toptal.com/designers/data-visualization/dashboard-design-best-practices.

Tableau (2022). Understanding and Using Line Charts. <https://www.tableau.com/data-insights/reference-library/visual-analytics/charts/line-charts>.

TechTarget (2022). Statistical Analysis. <https://www.techtarget.com/whatis/definition/statistical-analysis#:~:text=Statistical%20analysis%20is%20the%20collection,or%20designing%20surveys%20and%20studies>.

Turney, S. (2022). Skewness: Definition, examples, and formula. scribbr.com/statistics/skewness/.

Weitz, D. (2020). Treemaps, Why and How. Towards Data Science. <https://towardsdatascience.com/treemaps-why-and-how-cfb1e1c863e8>.

Zubiaga, A., & MacNamee, B. (2015). Knowing What You Don’t Know: Choosing the Right Chart to Show Data Distributions to Non-Expert Users. <http://hdl.handle.net/10197/7347>.