Business Process Analysis

For

World Suicide Rates

(1950 to 2016)

Submitted to:

Dr. Valarie Bell

Department of ITDS

University of North Texas - UNT

Prepared by:

Samuel Castilla

Satvika Marrapu

Suhail Bari

May, 2020.

Contents

Executive Summary

This project is about analyzing the worldwide suicide data from 1985 to 2016. Based on the obtained results from this analysis, the project includes the study of possible correlated factors that contribute to increasing the suicide rates and the alternatives to face this social problem.

Death by suicide is an extremely complex issue that causes pain to hundreds of thousands of people every year around the world. Globally, 1.4% of deaths were from suicide in 2017.

The project follows the DMAIC (Define, Measure, Analyze, Improve and Control) cycle, that is the core tool for six sigma.

The Define phase establishes that the project purpose is to increase the probabilities of suicide prevention. In addition, it defines the scope of the project focused on some aspects that may affect suicides, such as generational difference, age, sex, and countries.

The current state of the process, or Measure phase, describes the dataset and the preprocessing tasks. The accuracy, completeness, timeliness, believability, and interpretability were demonstrated. In the data cleaning we found that the dimension “HDI for year” that stands for Human Development Index per year, included about 70% of missing values, for this reason it was excluded from the analysis. There was no null value.

To make the required measurements, we used Tableau as data analysis tools. The polynomial regression models were statistically significant to explain the data. The regression model for Avg. Suicides explains the data in 91%, which is statistically significant.

The Analyze phase shows that the GDP, country, generation, and gender influence the suicide rate. Also, this phase shows a highest rate in the mid 90’s.

The Improve phase determines recommendations related with countries involved in the World War II, other countries like the US, Japan, Korea, Brazil and Mexico, and income disparity. Also, these improvements stablish that countries must consider as a priority the mental health, retirement benefits and governmental policies regarding suicide.

Finally, the Control phase establishes the best way to ensure the success in the implementation of the recommendations.

.

1.0 Improvement Opportunity: Define Phase

To increase and alleviate the understanding relating to suicides.

The purpose of this project is to analyze the existing worldwide suicide data from 1985 to 2016 and find out signals correlated to increased suicide rates among different cohorts globally and across the socio-economic spectrum. This compiled dataset is secondary data pulled from Kaggle.

We will examine the suicide rate comparing them to aspects that may affect in its increase, such as generational difference, age, sex, and countries.

1.1 Problem Statement/Discussion of the process being examined

In this project, we look at the world suicides data reported by country from the year 1985 to the year 2016. In order to understand the data, we need to analyze the trend of suicide rates as per country/gender/gdp and to develop an understanding as to what is contributing to these numbers.

1.2 Identification of key measures used to evaluate the success of your project

The key measures used in this project are GDP, No of suicides and the avg. suicide rate per 100,000 people. Our focus is to find out how these measures will stand against dimentions (age/sex/generation/country)

1.3 Discussion of project scope

The scope of this project is very limited. Due to a difference in policies across the world, the data we need is not perfect. Only 101 countries have come up with their numbers. In tandem to that, very few countries have reported the 2016 numbers.

2.0 Current State of the Process: Measure Phase

2.1 Current Performance Level

The data collected in this project comes from the online source Kaggle. The information comes from sources such as The United Nations Development Program, World Bank, and World Health Organization. This second-hand dataset provides information about the suicide rates overview compared to socio-economic info, by year and country. It includes information from 1985 to 2016, and has 27,820 records in total from 101 countries. The dimensions of the Suicide Rate dataset are country, year, sex, age group, count of suicides, population, suicide rate, country-year composite key, HDI for year, GDP_for_year, GDP_per_capita, and generation (based on age grouping average).

This dataset is licensed under: The World Bank. The source of this secondary dataset is https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016.

The following is the Data dictionary:

Dimensions:
Age – Age range of the population from 5 years to 75+ years.
Country – 101 countries
Country-Year – Essentially a combination of country and year
GDP for year – GDP of the country based on year ($ value)
Generation – A set of 6 generations based on factual representation of age
Sex – Male and female only
Year – Year from 1985 to 2016

Measures:
GDP per capita – GDP of relative country as per population
HDI for year – Human development index for the relative year
Population – total population of relative country
Suicides no – Number of suicides
Suicides/100K Pop – Number of suicides per 100,000 people in the relative country

Missing values:
HDI for year – Most of these values are missing and we will not be using these values in our visualizations.
Also, not all the countries have data in every year. Moreover, for the year 2016 only 16 countries have data and the information for the 5-14 years old people is missing.

Data Preprocessing is presented as follows:

Ø Data quality issues:

- Accuracy: the dataset accurately represents the behavior of suicide rates in the analyzed countries.
- Completeness: the files include data about the suicide rate for years from 1985 to 2016. However, not all the countries have data in every year. Moreover, for the year 2016 only 16 countries have data and the information for the 5-14 years old people is missing.
- Timeliness: the data was collected in the correspondent periods, so it is plenty useful.
- Believability: the data can be trusted, due to it comes from organizations like United Nations Development Program, World Bank, and World Health Organization. Also, all values make sense.

- Interpretability: the data shows the information described in the variable head. Likewise, the values are accurate.

Ø Data cleaning:

The dataset included missing values for the dimension “HDI for year. Due to these missing values are about 70% of the records for this variable, we excluded the dimension from our analysis.

Null Values: there was no null value.

Ø Data Integration:

The data was integrated from the source.

Ø Data reduction:

There is no smaller subset of data that can produce the same or similar analytical results.

Data processing is one of the longest tasks in data analysis. However, this phase took about 2 hours to study and get the final cleaned dataset to make the required analysis.

- **Current State of Key Y Outputs**

With the dataset cleaned and preprocessed, we used Tableau to get the following analysis:

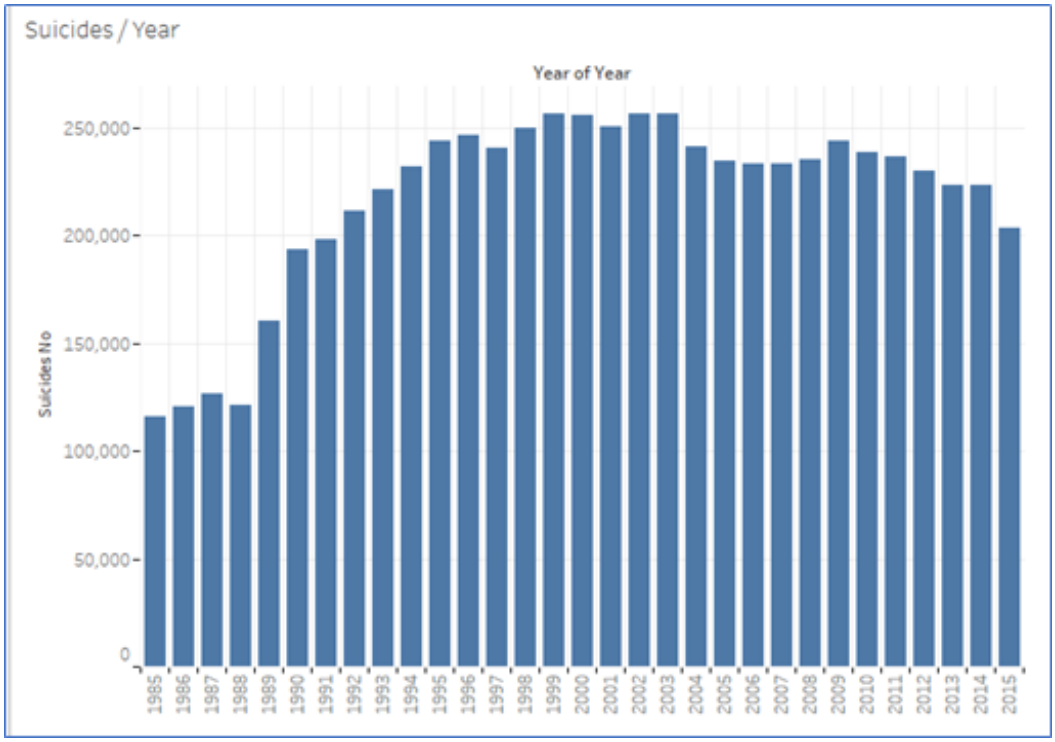

Figure 1. Suicides per Year.

The highest suicide rates have occurred in the years 1999, 2000, 2002, and 2003

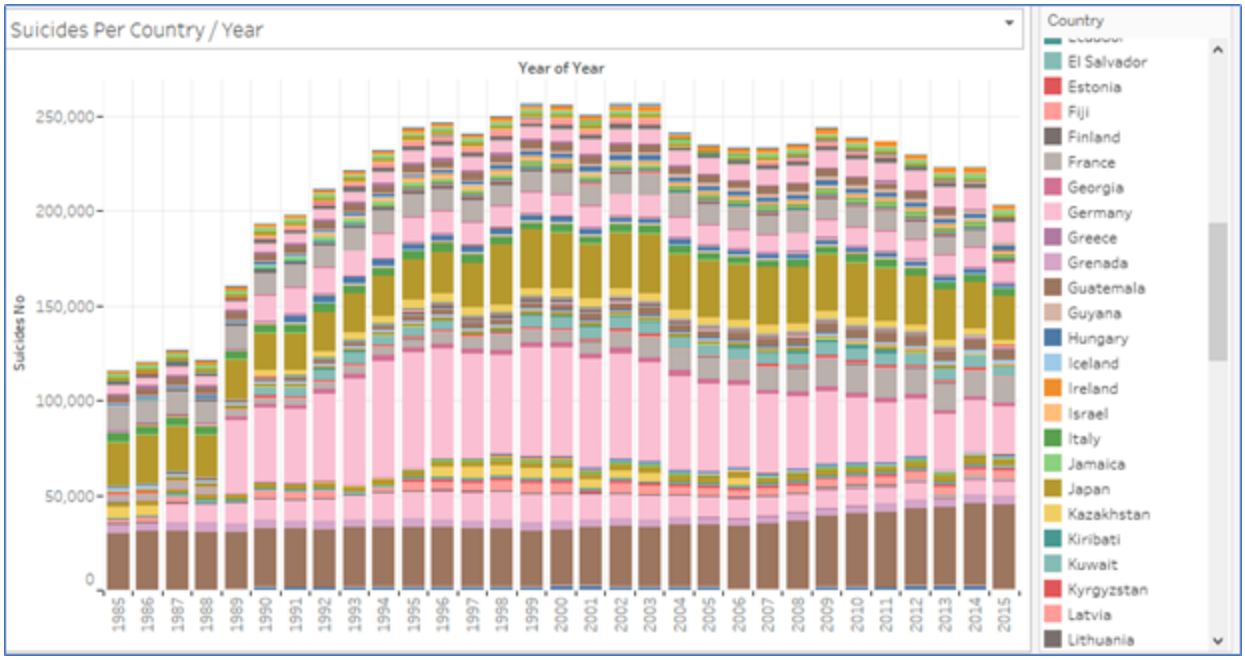

Figure 2. Suicides per Country / Year.

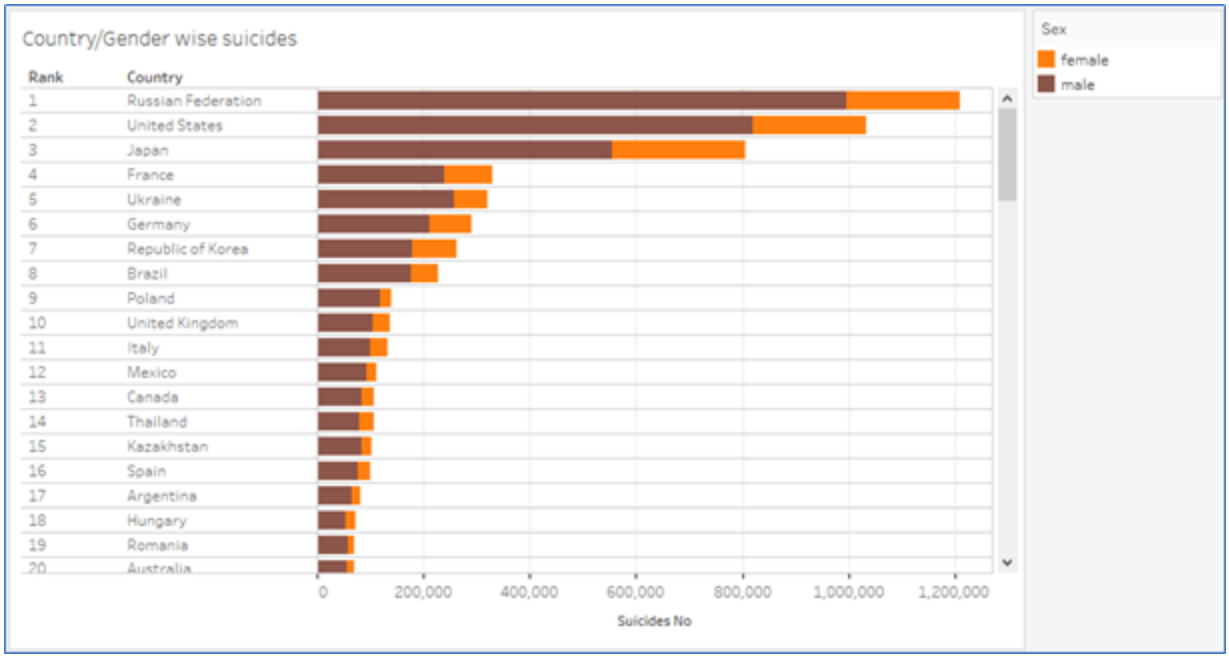

Figure 3. Country / Gender Wise Suicides.

The top five countries in the Rank by suicide rate are Russia, US, Japan, France, and Ukraine. This dataset does not include data from some countries such as China.

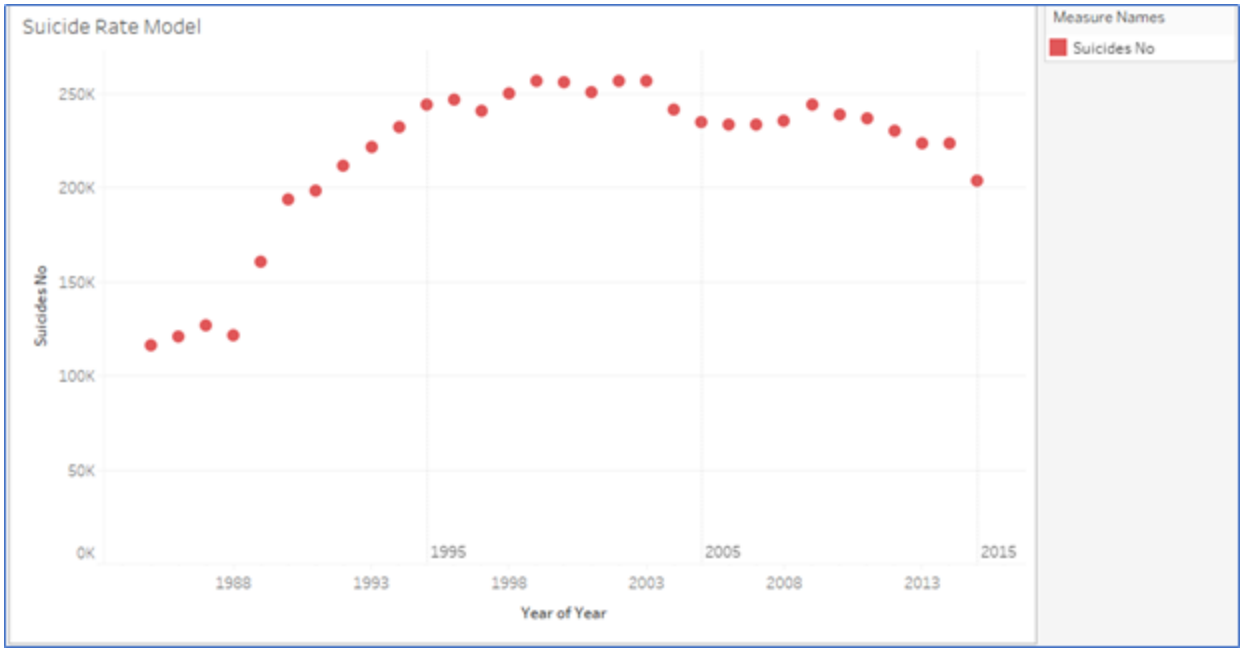

Figure 4. Time Series Plot Suicide Rate.

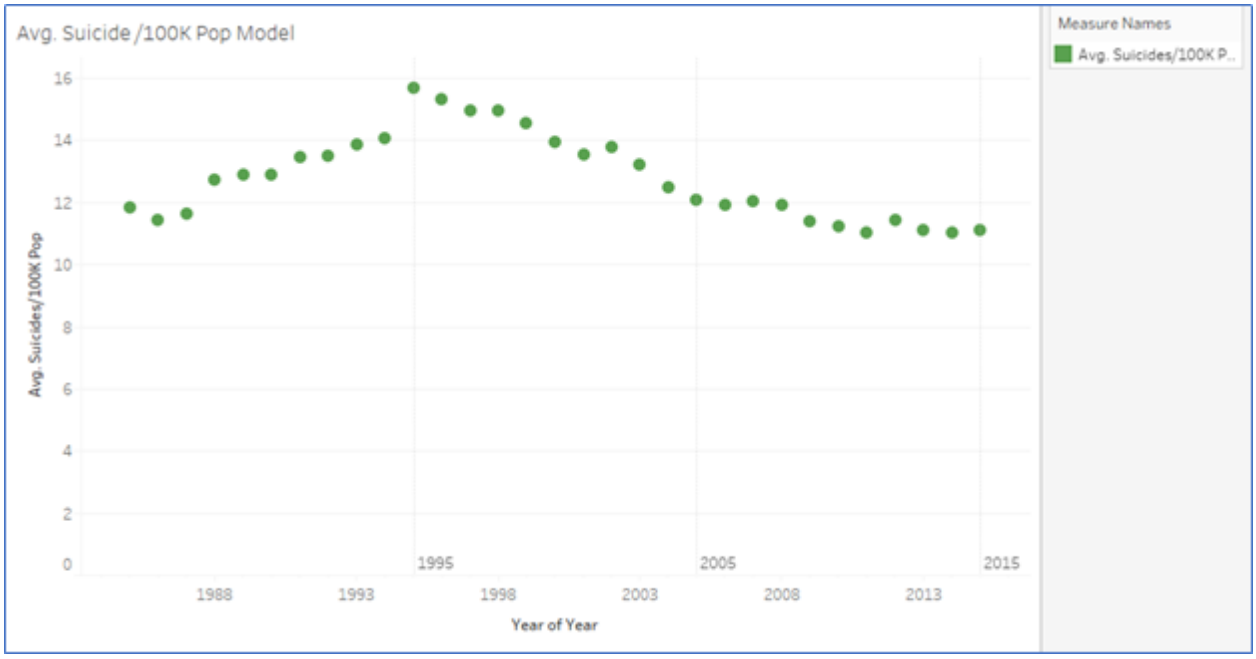

Figure 5. Time Series Plot Avg. Suicide / 100K Pop.

This time series plot for Suicide Rate and Avg. Suicide/100K Pop show a nonseasonal behavior. Also, there are no outliers.

Distribution/Data Patterns of Key Y Outputs

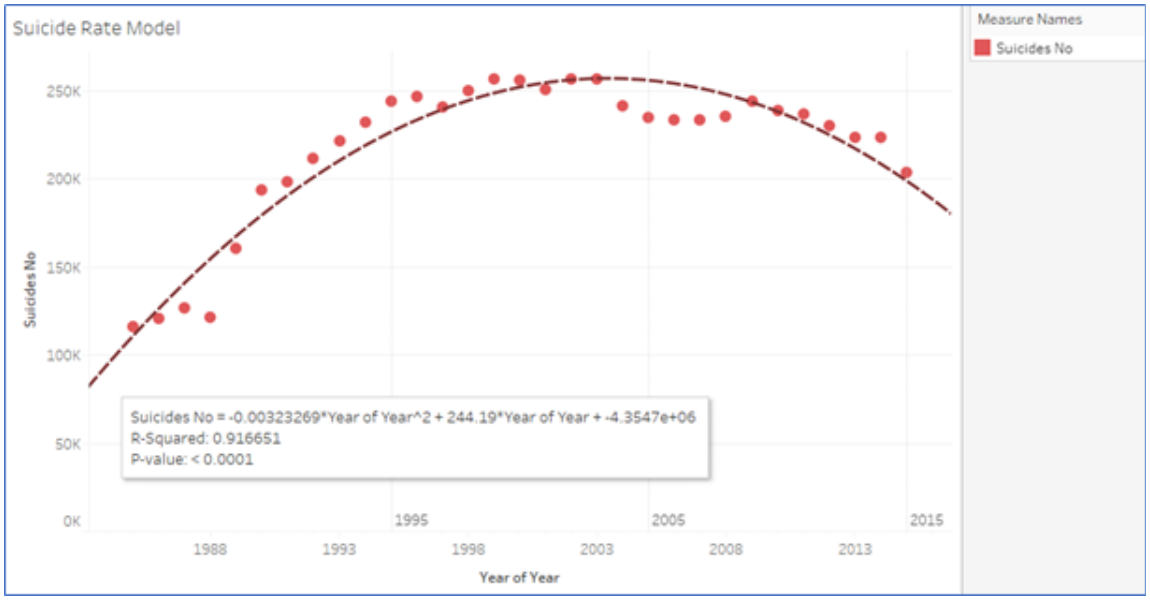

Figure 6. Suicide Rate Plot Polynomial Regression Tendency.

The Regression model for this data has a Pvalue < 0.0001, the R-Squared is 0.9166, so this model explains the data in about 92%. Therefore, this model is significant.

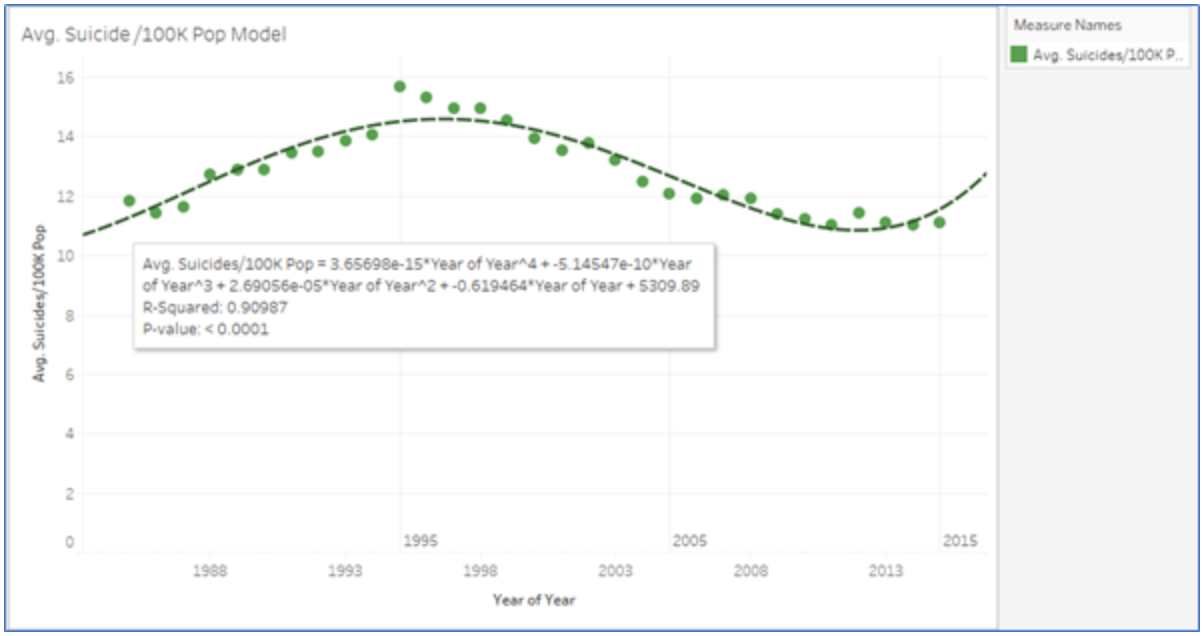

Figure 7. Avg. Suicide /100K Pop Plot Polynomial Regression Tendency.

The polynomial regression model for this data has a Pvalue < 0.0001, the R-Squared is 0.90987. This model is better than the LR model, it explains the data in about 91%. Therefore, this model is significant.

In conclusion, the Polynomial Regression model seems to be adequate for both, Suicide Rate and Avg. Suicide/100K Pop.

3.0 Analysis and Findings: The Analyze Phase

1. **Suicides/Year (Figure 1)**

   This visualization gives us an overview of the years which had the maximum number of deaths due to suicides. Contrary to many beliefs, the years closing to the millenia shot up suicidal rates. The highest suicide rates have occurred in the years 1999, 2000, 2002, and 2003.

2. **Country/Year (Figure 2)**

   We can tweak the first visualization in order to understand which countries contribute the most to these alarming numbers.

3. **Country/Gender wise suicides (Figure 3)**

The bar graph ranks all the countries based on the total number of suicides that it's seen from 1985 to 2016. The INDEX function was used to fill the rank rows. The total number of suicides were color coded by gender.

This view gives us an overview of the total number of suicides along with a gender-based divide. We can see that the male suicides are substantially more than their female counterparts. The ratio of male to female suicides is 3.3 : 1.

4. **Suicidal rate model (Figure 6)**

The suicidal rate model can be obtained by plotting the total number of suicides year on year. The polynomial regression tendency of this model is at a Pvalue < 0.0001, the R-Squared is 0.9166, so this model explains the data in about 92%.

5. **Avg suicide rate per 100K population (Figure 7)**

The suicidal rate model can be modified slightly taking into view the countries which have a lower population but higher number of suicides.The polynomial regression model for this data has a Pvalue < 0.0001, the R-Squared is 0.90987. This model is better than the LR model, it explains the data in about 91%

4.0 Recommendations: The Improve Phase:

These are the improvements we consider to control the suicidal rates:.

The suicide rate has increased exponentially in countries like the US, Japan, Korea, Brazil and Mexico.We need to look into those countries age groups and the people carefully.
The Income disparity is a constant variable as well as a very important factor to maintain the GDP of the country. So, we need to always look after the Income needs even when the GDP of a country is booming.
This data only comprises 101 countries. Countries with high populations like India, China are not included yet.When these countries are also taken in control then we can have more data which leads to improvement.
We need to look after the mental health, retirement benefits and governmental policies regarding these issues must be a priority.

5.0 Monitoring and Controlling:

To take control over the improvisation phase of different factors there are different control plans. But in this case the most important thing we do is that we need to look after the people's health, their retirement benefits and policies. Most people who are committing sucide are over the age of 60 years in most of the countries. So, we need to take care of these people if we want to have control over the suicidal rates.

6.0 Conclusions

Therefore the conclusions we obtained from this project are that more cases are on men compared to women. The people with more than 60 and 70 years are committing more suicide compared to other age people. By seeing all these different factors few conclusions were obtained like there is always an inverse correlation between GDP and the suicide rate. It is always true that Losing even one life to suicide is way too much and it is not acceptable.

Work Cited

Appendix A.

Statistical Analysis Plan.

View the following pages.

Statistical Analysis Plan

For

World Suicide Rates

(1950 to 2016)

DSCI 5260 Business Process Analytics

University of North Texas

Version: Final Project

Author: Samuel Castilla
Satvika Marrapu
Suhail Bari

Date: 08-May-2020

Table of Contents

CONTENTS

Death by suicide is an extremely complex issue that causes pain to hundreds of thousands of people every year around the world. Globally, 1.4% of deaths were from suicide in 2017 .

This project is about analyzing the worldwide suicide data from 1985 to 2016. Based on the obtained results from this analysis, the project includes the study of possible correlated factors that contribute to increasing the suicide rates and the alternatives to face this social problem.

The dataset provides information on suicide rates worldwide, including the socioeconomic spectrum. This dataset contains information from 1985 and 2016. It has 27,820 records in total from 101 countries. The Suicide rates dataset includes country, year, sex, age group, count of suicides, population, suicide rate, country-year composite key, HDI for year, gdp_for_year, gdp_per_capita, generation (based on age grouping average).

According to Kaggle, “This compiled dataset pulled from four other datasets linked by time and place”, and its license is from The World Bank.

The source of this secondary dataset is https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

Dimensions:

Age – Age range of the population from 5 years to 75+ years.

Country – 101 countries

Country-Year – Essentially a combination of country and year

GDP for year – GDP of the country based on year ($ value)

Generation – A set of 6 generations based on factual representation of age

Sex – Male and female only

Year – Year from 1985 to 2016

Measures:

GDP per capita – GDP of relative country as per population

HDI for year – Human development index for the relative year

Population – total population of relative country

Suicides no – No of suicides

Suicides/100K Pop – No of suicides per 100,000 people in the relative country

Null values:

HDI for year – Most of these values are null and we will not be using these values in our visualizations.

3 ANALYSIS OBJECTIVES

The overall scientific objectives of the analysis are the following:

- Determine the correlations between variables within the dataset.
- Draw the Suicide Rate data to analyze its behavior in the time frame.
- Determine relevant factors/signals such as trends and correlations.
- Determine the model that best explain the dataset.

- Based on the identified model, study of some possible business processes such as reduce the suicide rates, and to design alternatives to mitigate the signals related to rising suicide rates

4 ANALYSIS SETS/ POPULATIONS/SUBGROUPS

All the variables included in the dataset are relevant, for this reason, the project includes all of them. However, data will be analyzed according to some factors, such as GDP, gender, and countries.

5 ENDPOINTS AND COVARIATES

The project does not include endpoints nor covariates.

6 HANDLING OF MISSING VALUES, OUTLIERS AND OTHER DATA CONVENTIONS

To handle this data cleaning, we will consult the following references:

Newton, R. R., & Rudestam, K. E. (2013). Your statistical consultant: Answers to your data analysis questions. Thousand Oaks: SAGE Publications.

Konasani, V. R., & Kadre, S. (2015). Practical business analytics using SAS: A hands-on guide.

In addition, to handle missing values the project will use the “most probable value” approach. Therefore, we will estimate these missing values using regression, decision tree, etc.

For example, using SAS Enterprise miner, the SAS Impute button is used to impute missing values. The Tree surrogate option gives the best outcome when compared with Andrews waive, Distribution, Tukey's Bi-Weight, Mean and Huber. The Tree surrogate option enables SAS to first understand existing values and build a fitted predictive model to predict missing values.

About noisy data, the project includes the following possibilities:

- Binning: sort and adjust the value based on those of its neighbors (mean, median, boundary).
- Regression: Use predicted rather than actual values.
- Outlier analysis: Identify and exclude "odd" records.

To analyze the variables, we will use box plot to check for abnormal values. Also, SAS enterprise miner Filter option brings a filter for all outliers. Likewise, we will plot the dataset to find any noisy data.

7 STATISTICAL METHODOLOGY

A standard statistical procedure that will concur a relationship between two years. The Null hypothesis theory will be tested to determine any relationship based on variables between the data sets using cluster analysis.

For a particular year, we plan to analyze the properties of the distribution and find out any outliers and the form theories as to why they became such outliers.

7.1 STATISTICAL PROCEDURES

Depending on the variables, we plan to use correlation, regression analysis and ANOVA between two years of data. The method or procedure can be replicated from the SPSS tutorials website referenced below:

- Degree of Relationship:

To determine the relationship and correlation between the quantitative variables included in the dataset, this project uses the statistical test named Multiple Regression.

- Significance of Group Differences:

To determine the causal relationship between the dependent and the independent variables in the project, we may use the t test, or the analysis of variance (ANOVA) test.

- Prediction of Group Membership:

To identify the independent variables that best predict the dependent variable the project may use the discriminant analysis

7.2 MEASURES TO ADJUST FOR MULTIPLICITY, CONFOUNDS, HETEROGENEITY, ETC.

As of now, this point is invalid. If the data gives out these symptoms, we will adhere to standard statistical research practices.

For Multiplicity, we can refer to multiplicity issues in clinical trials: the what, why, when and how taken from an article from the International Journal of Epidemiology. A methodological guidance review from BMC Med Res Methodol can be used to study the clinical aspects of heterogeneity.

The following are the references:

Guowei Li,1,2,* Monica Taljaard,3,4Edwin R. Van den Heuvel,5,6Mitchell AH. Levine,1,2,7Deborah J. Cook,1,2,7George A. Wells,3,8Philip J. Devereaux1,7,9and Lehana Thabane1,2,9 - An introduction to multiplicity issues in clinical trials: the what, why, when and how.

https://www.researchgate.net/publication/311947626_An_introduction_to_multiplicity_issues_in_clinical_trials_the_what_why_when_and_how

BMC Med Res Methodol. 2012; 12: 111 - Investigating clinical heterogeneity in systematic reviews: a methodologic review of guidance in the literature
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3564789/

8 SENSITIVITY ANALYSES

In this particular project, the data is secondary and no new/uncertain inputs have to be given. Hence, we don't need to perform a sensitivity analysis.

9 RATIONALE FOR ANY DEVIATION FROM PRE-SPECIFIED ANALYSIS PLAN PERFORMED

The data will not be changed except for additions of a few columns for real world calculation purposes.

10 PLANS TO ENSURE QUALITY AND ETHICS

Please refer to Appendix A "Our Firm's Ethical Code".

11 PROGRAMMING PLANS (USE OF PYTHON, R, etc.)

To analyze and process data, we will use the following coding using Python:

```
#visualising numeric data using a histogram
```

```
%matplotlib inline

import matplotlib.pyplot as plt

data.hist(bins=20, figsize=(20,15))


from pandas.plotting import scatter_matrix

attributes = ["Variable name", "Variable name"]

scatter_matrix(data[attributes], figsize=(12,8))

#plotting the data (for visualization purposes only)

%matplotlib inline

import matplotlib

import matplotlib.pyplot as plt

plt.rcParams['axes.labelsize'] = 14

plt.rcParams['xtick.labelsize'] = 12

plt.rcParams['ytick.labelsize'] = 12

plt.plot(X, y, "b.")

plt.xlabel("$x_1$", fontsize=18)

plt.ylabel("$y$", rotation=0, fontsize=18)

plt.axis([0, 1, 4, 12])

plt.show()

#imputing missing values

#scaling numeric data

import numpy as np

from sklearn.preprocessing import StandardScaler
```

```
from sklearn.preprocessing import Imputer

imputer= Imputer(strategy="median")

scaler=StandardScaler()

col_names=list(df_num)

num_col=np.array(df_num)

num_col_imp=imputer.fit_transform(num_col)

num_col_scaled=scaler.fit_transform(num_col_imp)

num_col_scaled

df_num_scaled=pd.DataFrame(num_col_scaled, columns=col_names)

df_num_scaled.head()


#fitting linear regression by solving the normal equation using simple numpy

#.dot is a numpy method for matrix multiplications

X_b = np.c_[np.ones((N, 1)), X] # add x0 = 1 to each instance

theta_best = np.linalg.inv(X_b.T.dot(X_b)).dot(X_b.T).dot(y)

theta_best


#making predictions

X_new = np.array([[0], [1]])

X_new_b = np.c_[np.ones((2, 1)), X_new] # add x0 = 1 to each instance

y_predict = X_new_b.dot(theta_best)

y_predict
```

12 REFERENCES USED IN THE COURSE OF YOUR ANALYSES CAN BE FOUND IN APPENDICES UNDER "APPENDIX B"

13 APPENDICES

APPENDIX A

OUR FIRM'S ETHICAL CODE

We pledge in writing to abide by the American Statistical Association's (ASA) and INFORMS' Codes of Ethics. Our adherence to these Codes signifies voluntary assumption of self-discipline. As the professional associations for our firm in the United States, the ASA and INFORMS requires adherence to their Codes of Ethics as a condition of membership. The standards of conduct set forth in these Codes provide basic principles in the ethical practice of data analysis consulting. The purpose of these Codes is to help us maintain our professionalism and adhere to high ethical standards in the conduct of providing services to clients and in our dealings with our colleagues and the public. Our individual judgment requires we apply these principles. We are liable to disciplinary action under the ASA's and INFORMS' Rules of Procedure for Enforcement of this Code if our conduct is found by the ASA's or INFORMS' respective Ethics Committees to be in violation of their respective Codes or to bring discredit to the profession or to ASA and INFORMS .

Our Commitment to Our Clients

1) We will serve our clients with integrity, competence, independence, objectivity, and professionalism.

2) We will mutually establish with our clients realistic expectations of the benefits and results of our services.

3) We will only accept assignments for which we possess the requisite experience and competence to perform and will only assign staff or engage colleagues with the knowledge and expertise needed to serve our clients effectively.

4) Before accepting any engagement, we will ensure that we have worked with our clients to establish a mutual understanding of the objectives, scope, work plan, and fee arrangements.

5) We will treat appropriately all confidential client information that is not public knowledge, take reasonable steps to prevent it from access by unauthorized people, and will not take advantage of proprietary or privileged information, either for use by ourselves, the client's firm, or another client, without the client's permission.

6) We will avoid conflicts of interest or the appearance of such and will immediately disclose to the client circumstances or interests that we believe may influence my judgment or objectivity.

7) We will offer to withdraw from a consulting assignment when we believe my objectivity or integrity may be impaired.

8) We will refrain from inviting an employee of an active or inactive client to consider alternative employment without prior discussion with the client. Our Commitment to Fiscal Integrity

9) We will agree in advance with a client on the basis for fees and expenses and will charge fees that are reasonable and commensurate with the services delivered and the responsibility accepted.

10) We will not accept commissions, remuneration, or other benefits from a third party in connection with the recommendations to a client without that client's prior knowledge and consent, and will disclose in advance any financial interests in goods or services that form part of such recommendations. Our Commitment to the Public and the Profession

11) If within the scope of my engagement, we will report to appropriate authorities within or external to the client organization any occurrences of malfeasance, dangerous behavior, or illegal activities.

12) We will respect the rights of consulting colleagues and consulting firms and will not use their proprietary information or methodologies without permission.

13) We will represent the profession with integrity and professionalism in my relations with our clients, colleagues, and the general public.

14) We will not advertise our services in a deceptive manner nor misrepresent or denigrate individual consulting practitioners, consulting firms, or the consulting profession.

15) If we perceive a violation of the Code, we will report it to the APA and INFORMS and will promote adherence to the Code

by other member consultants working on our behalf.

APPENDIX B

REFERENCES

Newton, R. R., & Rudestam, K. E. (2013). *Your statistical consultant: answers to your data analysis questions*. Thousand Oaks, CA: SAGE.

Konasani, V. R., & Kadre, S. (2015). *Practical business analytics using Sas: a hands-on guide*. New York: Apress.

Li, Guowei & Taljaard, Monica & Van den Heuvel, Edwin & Mitchell, Alexandra & Cook, Deborah & Wells, George & Devereaux, Philip & Thabane, Lehana. (2016). *An introduction to multiplicity issues in clinical trials* ... (n.d.). Retrieved from

https://www.researchgate.net/publication/311947626_An_introduction_to_multiplicity_issues_in_clinical_trials_the_what_why_when_and_how

Gagnier, J. J., Moher, D., Boon, H., Beyene, J., & Bombardier, C. (2012, July 30). Investigating clinical heterogeneity in systematic reviews: a methodologic review of guidance in the literature. Retrieved March 15, 2020 from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3564789/.

An introduction to multiplicity issues in clinical trials ... (n.d.). Retrieved March 15, 2020, from https://www.researchgate.net/publication/311947626_An_introduction_to_multiplicity_issues_in_clinical_trials_the_what_why_when_and_how