Investigation into Whether Suicide and Homicide Rates Are Influenced by GDP Per Capita

GROUP-A

Mohammed Alghazi X18208495

Andew Kelly X18212158

Jordan Lee-Graham X19103310

Abstract

— The goal set out for this assignment is to investigate whether Suicide, Homicide and GDP Per capita have any correlations when the economy is struggling versus when the economy is in a good state. The dataset was obtained from Kaggle and contained a list of 111 countries between 2000-2018. The dataset was cleansed and interpreted through Python and R. The dataset was used in variety of ways, such as, comparing GDP against homicides and GDP against suicides and finally all three together to predict an outcome in relation to GDP for a certain year.

Upon the completion of all the tasks that were posed, a summary and plots of the results should also be obtained and should include all the relevant information that is needed. This report will explain in detail all the outcomes and how each result was obtained along with a justification of each test carried out and why it was carried out.

# **Introduction**

Post Covid-19 pandemic, which has affected a lot of people differently, some people benefited greatly from being under a nation-wide lockdown, but the vast majority of ordinary people were negatively affected. Our investigation started by attempting to identify the core issues on what could spark a sharp increase in crime rates. Prior to the Pandemic, the global economy was in a good state, People had jobs and lively hoods. Over the past two years, the pandemic deepened our finical struggles which could lead to a higher rate of homicide and suicide in society. Even though the pandemic is not officially over, economies worldwide are slowly recovering in the aftermath of the pandemic.

The ongoing war in Ukraine had shed light on atrocities across the world and the financial ruins that wars cause. One can only assume the adverse effects that this would have on society such as, homicide and suicide rates. With ongoing conflicts, everyday essentials become extremely sought after which may lead to individuals changing their morals breaking the law in order to obtain.

From our investigation, the decision was made to explore the effects of GDP per Capita in contrast with the homicide rates and suicide rates across the 111-country dataset in which we obtained from Kaggle. The objective of this particular investigation was to highlight the dataset starting with GDP per Capita along with homicide and suicide rates, our first task was to essentially go through the data and determine what data is relevant to our investigation and what data was not of use in this instance. Once this was achieved, the dataset was uploaded into visual studio using python as our core programming language, data cleansing of the dataset commenced. Upon the completion of the dataset cleansing which consisted of removing special characters that could affect the skew our results and taint our visualization.

The question that was posed are as follow

Does the rate of GDP per capita have a direct impact with the suicide and homicide rates? Additionally, to investigate how homicide and suicide rates have changed over the years and could we predict the GDP along with these rates.

To achieve these findings, we implemented a series of machine learning models in an attempt to extract insights. Through single linear regression models, multi linear regression models and K-Means clustering techniques, we have outlined our findings and compared them with the relevant literature available today.

This report will include all of the data tested along with the outcome of each report which justified the selected data for each test.

**Literature Review**

There are various different journals and articles available online, while researching different topics and questions that could be possible to answer in this assessment, The articles below helped to gain an understanding into which elements of the dataset that was obtained that we could make use of for our study.

A paper written in 2016 by Jukka Hintikka, Pirjo I Saarinen and Heimo Viinamaki explored suicide mortality rates in Finland during the economic cycle of 1985-1995. This paper found that during a recession that suicide mortality rates in men went down while they went up during 1985 to 1990 where the economy was doing well. “Age-adjusted suicide mortality in men increased by 20% during the economic upswing from 1985 to 1990. During the economic recession, with its mass unemployment, suicide mortality in men decreased by 13% from 1990 to 1995.” This was an interesting finding and something to note when carrying out our research. [1]

A 2018 article produced by the Cambridge University Press detailed a similar study with similar data using a mixture of variable types to conduct the analysis. In the Cambridge study, ‘GDP’, ‘GDP per capita’ and ‘year’ were all used as rates as opposed to numbers as dependent variables. Their reasoning for this was stated that "the use of rates would put all countries at a similar level and would not give any advantage to the big ones", as well as mentioning how "numbers of suicides did not fit any of the basic distributions and the models derived had poor goodness-of-fit. [2]

A journal written 2011 by Alfred Barth, PhD, Leopold Sogner, PhD, Timo Gnambs, PhD, Michael Kundi, PhD, Andreas Reiner, PhD, and Robert Winker, MD investigated Socioeconomic Factors and Suicide: An Analysis of 18 Industrialized Countries for the Years 1983 Through 2007. They sought out to see if there was a correlation between suicide rates and socioeconomic factors. This paper found that socioeconomic factors do have a correlation with the rate of suicides however it differed between opposing sex. “Socioeconomic factors are associated with suicide rates. However, this relationship differs by sex. [3]

A journal written in 2002 by Alexander Butchart1 & Karin Engstro, titled Sex- and age-specific relations between economic development, economic inequality, and homicide rates in people aged0–24 years: a cross-sectional analysis. The analysis was set out to test whether relations between economic development, economic inequality, and child and youth homicide rates are

sex- and age-specific, and whether a country’s wealth modifies the impact of economic inequality on homicide rates. [4]

There was a paper written in 2016 by researchers in China, investigating the relationship between suicide rates and economic growth in China between the years of 2004 and 2013. This study found that overall, there was a decrease in suicides in China during this time while GDP per capita also steadily increased each year during this time showing that higher GDP leads to lower suicides. “Suicide rates of each age-group, sex, urban/rural area, and region were generally decreased from 2004 to 2013, while annual GDP per capita and rural and urban income per capita were generally increased by year.” [5]

In 2020 there was a paper written by Christopher M. Doran and Irina Kinchin, looking in to Economic and epidemiological impact of youth suicide in countries with the highest human development index. This study was investigating the impact that suicides had on the economy of countries with a high human development index. This study found that suicide rates within a country can have a massive impact on an economy and the development of a country. “The results are staggering–almost 7,000 young lives are lost each year to suicide representing a loss of 406,730 years of life at a cost of over $5.53 billion. Reducing youth suicide requires a multifaceted approach and significant investment by governments.” [6]

In 2019 a report was conducted by the United Nations Office on Drugs and Crime. The topic of this report was a Global Study on Homicide - Homicide, development, and the Sustainable Development Goals. This report found that overall, across the board for countries, that there is a link between the homicide rate in a country and the level of development within a country. These two factors then have a knock-on effect making it harder for either situation to improve. “Macroanalysis of the social and economic factors most closely associated with homicide in different countries reveals that, overall, there is a strong link between a country’s homicide rate and its level of development.” [7]

A journal from 1999 by Pablo Fajnzylber Daniel Lederman Norman Loayza University of Minas Gerais the World Bank Central Bank of Chile and The World Bank conducted a report on Inequality and violent crime. “Paper investigates empirically some of the key issues raised by the apparent connection between income inequality and violent crime. These issues are, first, the robustness of this link to alternative measures of income inequality” [8]

A 2017 journal article on the relationship between economic growth and crime rates in Namibia outlined the findings of an Augmented Dickey-Fuller and Phillips Perron tests. The article states how "the Granger causality results showed a bidirectional causality running from economic growth to crime rate and vice versa." Meaning the variables can be used to predict one another.

[9]

Different studies considered different factors and variables into their analysis. An analysis by Benjamin Northrup and Jonathan Klaer performed a single regression model made use of GDP per capita and the number of violent crimes by state. This data was taken from 2003-2011. [10]

**Methodology**

The analysis process was performed in following the KDD methodology. Reasons for choosing this methodology as opposed to other alternatives such as CRISP-DM or SEMMA which are very similar processes with similar steps involved, was partly due to the focus being on ‘research’ as opposed to a business analysis problem which would be better defined in CRISP-DM’s ‘business understanding’ beginning stage. The final stage of CRISP-DM focuses on deployment of the solution to the business problem whereas KDD focused on ‘problem identification’ and allowed steps for evaluation and interpretation in the latter stages alongside ‘post KDD’ which provided an opportunity to share the knowledge gained from the study.

There are five steps involved in a KDD methodology and these are, Selection, Preprocessing, Transformation, Data Mining, Evaluation.

The first stage of the KDD methodology is the ‘**Selection**’ process which involves identifying a relevant set of data containing fields relevant to our proposed research question. Based on relevant available research such as the journal article by Northruf et al. [10] the main variables used for the research where ‘GDP’ values and ‘GDP per capita’ values as well as number of violent crimes. For the analysis on homicides, suicides in contrast to GDP it was important to identify a dataset containing these values in order to implement a similar analysis. The dataset which was most appropriate for this study was discovered on Kaggle and comprised of the relevant fields required for the research. The dataset itself was created by multiple contributors using the individual values of GDP (in US $), homicide numbers and suicide numbers concatenated from the data.worldbank.org website. The compiled data was engineered by the original creators to contain extra variables such as ‘per capita’ based on 100,000 of a population as well as categorizing different nations into income levels based on this GDP value per capita. These added components became useful for the visualization of data when not entirely necessary for use in a model meaning the full extent of the data was useable for the research.

For the next stage **Preprocessing,** the dataset was explored manually through Excel in order to identify any standout features. One such feature identified early on in this process was the appearance of interrogation points and special characters within some of the columns. Certain rows of data also contained blank rows. These features of the dataset needed to be handled in order to prevent ‘NaN’ values disrupting models or an instance of graphs being skewed by unrecognizable dates. Using Python to automatically remove cases which matched special characters or contained ‘NaN’ values the dataset was cleansed from over 13,000 records to a much more condensed and complete 1,488. Once the raw data cleansing had concluded the exploration of appropriate models for this data began, which leads into the next stage in the KDD process of transformation.

Step three in the KDD methodology is **Transformation**, this stage in the process involves manipulating and maneuvering the data in order to prepare the data for use in possible machine learning algorithms which were being explored. One element of consideration was the use of linear regression and multi linear regression, with these algorithms it is important to identify a possible dependent variable and independent variable(s). Through the use of the Pandas library the functionality of creating separate blocks of data in the form of ‘data frames’ was available in order to create isolated dependent and independent variable sets which could be called upon for the relevant models. Extracting variables such as ‘year’ and ‘intentional homicides’ for use in a simple linear regression model by creating testing and training sets of these isolated variables sets without the need to handle additional variables which were not to be used in the model such as ‘suicides. The isolated variables could simple be split into training/test sets once separated from the other variables.

**Data Mining** is a vital step in the KDD process, throughout this stage algorithms are performed on the prepared data in order to extract insights and findings. The first data mining process implemented during this stage was the use of regression analysis in order to perform a prediction of homicides based on the years previous values. The dependent variable used in this scenario was the ‘homicides’ variable with the independent variable being the ‘GDP\_per\_capita’ meaning the number of homicides altered depending on the GDP per capita provided. The regression equation “*GDP\_per\_capita = 11.663244885095063 + -0.000158 \* N of homicides*” was applied to years post data recorded. This is performed similar to the research provided by Northrup et al. [10] where this same application was applied to the GDP of individual states in the US and their violent crime rates.

Regression was used again although in the form of multiple linear regression. This time around the dependent variable GDP was calculated from two independent variables ‘homicides per 100k’ and ‘suicide mortalities per 100k’. The assumption here was that a combination of using the trends of homicides, suicides, and GDP per capita would allow an input value of homicide and suicide values to predict a countries GDP per capita.

To evaluate the accuracy of the regression model, a formula can be used to see if the model is accurate. This formula takes an initial prediction that is accumulated from homicides and suicides per 100k against GDP. This prediction is giving an expected GDP when homicides and suicides are at a certain rate. Also inputted into this formula is the coefficient of the regression model. The final part of this formula is adding in a number to change one of the variables. This formula will output another prediction for the GDP after the variable change. If the model is accurate the two outputs should be similar numbers to each other. An example of this formula would be 6997.17 + (10\*-79.33).

Graphs were also accumulated for the increase/decrease in homicides, suicides, and GDP year on year. These will be presented in bar chart form.

Through clustering analysis and using the silhouette method to find the best K value (recommended number of clusters) which returned a silhouette score of 7. This value of 7 indicates that the dataset has a potential of 7 different data clusters. These data points would be separated by homicides, suicides, and GDP combined.

**Evaluation –** Graphs will have been plotted at this stage as well as formulas that will have been used to gauge the accuracy of the models. The nature of some of the graphs will make it easier to interpret the results and draw conclusions from them while other results will require the use of formulas and interpreting values such as the R squared values and K values. Plots such as bar graphs will be the easiest to interpret while scatter plots will also be relatively easy to draw conclusions from. The data will also be presented in forms such as Pearson correlation matrices, boxplots, and line graphs. Formulas are very important to this process as you want to be able to accurately evaluate how good the model is and whether correlations are being formed between the variables that are being tested.

Diagram

Description automatically generated

**Implementation**

This section will cover the code base along with how the project was constructed. Topics such as the application workflow will be covered, technologies that were used to develop the report and how one would go about acquiring and interpreting the results. Below is a diagram of the application workflow:

**Purpose Identification** - The first step in the process is to find a purpose behind the report. Find something of interest where conclusions and correlations can be identified. The topic of interest in this case was to see if there was a link between homicide and suicide rates relative to how the economy was performing in a country.

**Data Acquiring** - In this instance the dataset chosen contained information about homicides, suicides, GDP, and the GDP per capita of countries around the world from the years 2000 to 2018. The dataset was obtained from Kaggle Datasets and contained all the variables needed for a thorough investigation into this issue. This dataset required cleansing to remove null values and duplicates of any rows which was completed through the Python library called NLTK. Once this cleansing process was completed the dataset was ready for statistical analysis.

**Modelling** - The next step in the process is the modelling phase. The models that would be used in this process are pre analysis, single linear regression, multi-linear regression, and K means clustering. For pre analysis the goal was to compare how homicides, suicides, and GDP per capita changed year on year. Matplotlib was used to present this data in order to see how they changed each year. Within matplotlib the visualisations that were used were bar charts as they were an easy way to present and interpret the change each year. For the first bar chart, homicides on the Y-axis were plotted against the year on the X-axis. For the second bar chart, suicides on the Y-axis were plotted against the year on the X-axis. For the third bar chart GDP per capita on the Y-axis were plotted against the year on the X-axis.

For single linear regression the Python library sklearn was used along with matplotlib. We separated the data into training and test splits using a split of 80%/20%. Meaning 80% of the data was used for training while the remaining 20% was used for testing. The mean squared error was then calculated between the training and test split. After the mean squared error was acquired, exploratory analysis was carried out on the data including boxplots, histograms, pair plot, and Pearson correlation matrix. Single linear regression was run a second time and in this instance a training and test split of 70%/30% was used. The model will then print an intercept and coefficient which can be used to plot the line of best fit. The single linear regression model was then tested manually to get a brief idea of its accuracy. Once this is complete run the sklearn accuracy score to get the R squared value of the model along with the mean absolute error, mean square error and root mean square error. The statsmodels.api library was then used to output the OLS regression results.

The next step was to carry out multi-linear regression. For multi-linear regression the dependent variable was GDP per capita while the independent variables were homicides per 100k and suicides per 100k. Train and test splits were once again created with a 70% training split and 30% testing split. Random state was also applied for reproducibility. The multi linear regression was then run. The test and training result were produced on a table showing actual value and predicted value. The model was then tested manually using two random variables for the independent variable’s homicides and suicides. The r squared value, the mean absolute error, mean square error and root mean square error were produced. The statsmodels.api library was then used to output the OLS regression results.

K means clustering was carried out in R. In R the libraries corrplot, cluster, and factoextra were imported. The data was read from the same csv file used before in Python to maintain consistency. The data frame was created in order to isolate the important variables and remove unnecessary variables. The goal is to find the best K value, the optimum number of clusters. Another correlation matrix is produced. The silhouette method was then used to find the best K value.

**Reporting** – In this phase of the project, results are gathered from the statistical analysis and compiled into the report along with the completion of the rest of the sections required for a suitable report.

**Results**

This section will cover the results and findings from the statistical analysis carried out in Python and R.

The image below is a sample of the information contained within the dataset once data cleansing was completed, the information below is represented.

Graphical user interface

Description automatically generated with medium confidence

The image below shows the number of suicides from 2000 to 2018 and we can see that in the more recent years that the rate of suicides has gone down following a steady increase from 2006 to 2013.

Chart, bar chart, histogram

Description automatically generated

The image below shows how GDP has changed from 2000 to 2018. GDP has been gradually rising since 2000 with a sharp increase occurring from 2015 to 2018. The study by Northrup [10] used variables such as GDP in the form of rates as opposed to solid values. This may have been an interesting consideration for the research as GDP has a consistent increase over time.

Chart, bar chart, histogram

Description automatically generated

The image below shows how the rate of homicides has changed from 2000 to 2018. We can see that the rate of homicides has remained relatively consistent over the period of time with a sharp increase in 2015 and a decline ever since then.

Chart, bar chart

Description automatically generated

The image below is a Pearson correlation matrix.

gdp\_per\_capita and intentional homicides has a correlation of -0.1 meaning a -10% correlation in these values.

gdp\_per capita and suicide mortality have a correlation of .06% gdp\_per capita and suicide mortality have a correlation of .06%.

The low correlation between these values means there is no need to perform standardization on the variables in order to put them through a machine learning model

A picture containing graphical user interface

Description automatically generated

The images below are two different scatter plots. The first one is plotting GDP per capita to year while the second is plotting homicides to the year.

Chart, scatter chart

Description automatically generated

The plot below is a scatter plot showing results from the single linear regression model

Chart, scatter chart

Description automatically generated

**Conclusions & Future Work**

created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting