**Name: Bhushan Suryawanshi   
Date: 2020-07-22  
Title: Suicide data Analysis**

**Section 1**

* Introduction – We are facing pandemic and whole world is working from home. I am also practicing work from home and I can see that being at home 24hrs and not able to go out does build frustration. Same might be case with others (kids, teenagers, singles etc). They locked themselves home and that’s why we have seen protests emerging around the world to open the work places. If we keep this restricted work for longer it will definitely build depression and can cause “Suicide”. My organization is continuously conducting virtual sessions with psychologists and motivational speakers to ease these type of tensions. This led me to check what is historical data is saying about Suicides. If we know the suicide rate and able to predict which age group is more affected, we can enhance response to depressed individuals.
* Research questions -

1. Is the suicide only be seen in low income house hold because they don’t have income to pay their debt as well as regular day to day expenses?

2. What is the rate in each country?

3. Is Suicide rate affected recently (decreasing/increasing)?

4. Does age affect suicide rate?

* Approach – Plotting data using different visualization methods will provide details on the data. Find relationship between different variables. Create correlation martrix. Identify dependent and independent variables in the dataset. Apply appropriate model as per the EDA. Test the model and predict results using confusion matrix.
* How your approach addresses (fully or partially) the problem.

The approach mentioned above will give insights on the data. It will help to identify the data types and relations in the variables. The exploratory analysis of the data will help in identifying which data set suits the problem statement. It will address if there is any cleaning process required? Also getting more appropriate model will give us higher accuracy on prediction of suicide rate.

* Data

1. <https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016>

Dataset has 12 columns and 27000 rows. This compiled dataset pulled from four other datasets linked by time and place, and was built to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum.

2. <https://www.kaggle.com/szamil/who-suicide-statistics>

Dataset has 6 columns and 43K rows. Basic aggregate numbers covering 1979-2016, by country, year, age groups and sex based on [WHO Mortality Database online tool](http://apps.who.int/healthinfo/statistics/mortality/whodpms/).

3. https://healthdata.gov/dataset/vital-statistics-suicide-deaths-age-group-raceethnicity-resident-county-region-and-gender

Dataset with 8 columns and 2000 rows. This dataset contains suicide death counts by region, race or ethnicity, sex, and age group. For more information, check out: <http://www.health.ny.gov/statistics/vital_statistics/>.

* Required Packages

1. ggplot2 – For plotting graphs.

2. knitr – Rmarkdown documentation.

3. ggm – For pcorr and correlation.

4. car, QuantPsyc – For regression analysis.

* Plots and Table Needs

Histogram will be used to get distribution of data. Scattered plot to see relationship between different variables. Linear regression graph. Confusion matrix. There might be necessity to have other graphs based on further analysis.

* Questions for future steps.

Most of the data collected looks good however, further exploration will prove if the dataset selected is appropriate statistically. Is logistic regression is good model based on the data? Will need to dig little more on the psychology domain just to have better understanding on each data point.

**Section 2**

* How to import and clean my data

For importing data R has function read\_csv() provided for reading data in csv file format. Suicide dataset is in csv format and this will be handy. For cleaning lot of checks need to be done like sensitive data, missing values, variable names, duplicate rows etc. In case of suicide dataset referred here there is not sensitive or personal information. When you summarize data it will show the data types and if there are any NA values. This can help in cleaning process. Data duplication can be verified with duplicate function in dplyr package.

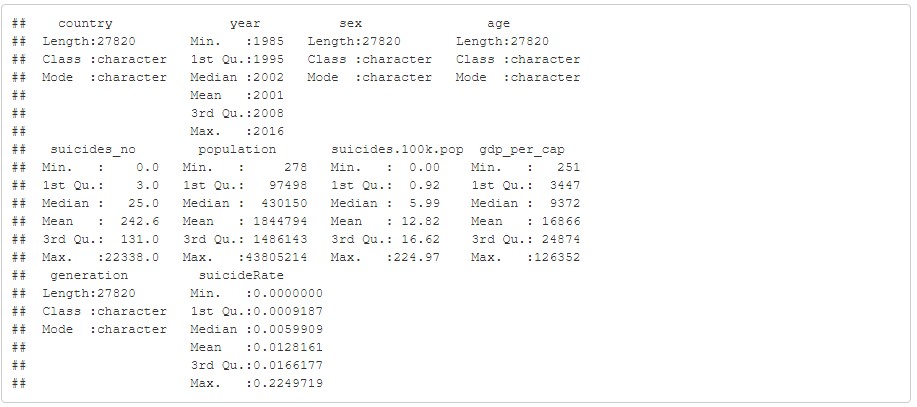
Here is small example how I am cleaning data – This will remove HDI.for.year, country.year and gdp\_for\_year and rename two columns “ ï..country” to country and “gdp\_per\_capita....” to gdp\_per\_cap.



* What does the final data set look like?

Get summary(), str(), head() and tail() functions to know how data looks after cleaning to make sure all the cleaning steps worked as expected. Also need to find additional graphs to check the data.

Here is summary of final data -



* What information is not self-evident?

Suicide rate is not given directly and need to use statistical formula to add this in the dataset. We don’t have individual suicide information to know what was the cause and hence only using country specific suicide rate for the analysis.

* What are different ways you could look at this data?

Summarizing data, using different plots and visualizations, grouping data by different categories can be used. In case od suicide data I will be using scattered plot, histograms, density plot to look at the data. Geographical plots on map.

* How do you plan to slice and dice the data?

For the analysis I will be dividing data into – Full data set with cleaned data. Cleaned dataset for only suicide happened in specific country. This way it will help to look at overall trend. Also looking at each country which is representation of different geographical area will be helpful.

* How could you summarize your data to answer key questions?

As I will be answering questions related to suicide by age, year, economy etc. So having mix of visualization and showing mean and standard deviation will be helpful. Correlating GDP with suicide rate for specific country over the year will also tell if improved GDP rate affects suicide rate positively/negatively. Suicide numbers by age can help to know which age group has more number of suicides. Suicide numbers by the year can bring out information if there was specific time period when suicide rate was high.

* What types of plots and tables will help you to illustrate the findings to your questions? Showing geographical heat map of suicides. Histogram, scatter plot, line graphs. Correlation plot to get correlation between variables.
* Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

I am going to use regression model for the prediction of rate but not yet sure how that will fit. If

it works that can help to predict suicide rate in specific country or age group.

* Questions for future steps.

What is the best model to use? Can I use multiple models to answer my question? What is the accuracy of each model? How to visualize geographical heat map in R?

**Section 3**

* A story / narrative that emerged from your data. Follow this structure.
  + Introduction.

We are facing pandemic and whole world is working from home. I am also practicing work from home and I can see that being at home 24hrs and not able to go out does build frustration. Same might be case with others (kids, teenagers, singles etc). They locked themselves home and that’s why we have seen protests emerging around the world to open the work places. If we keep this restricted work for longer it will definitely build depression and can cause “Suicide”. My organization is continuously conducting virtual sessions with psychologists and motivational speakers to ease these type of tensions. This led me to check what is historical data is saying about Suicides. If we know the suicide rate and able to predict which age group is more affected, we can enhance response to depressed individuals.

* + The problem statement you addressed.

1. Is the suicide only be seen in low income house hold because they don’t have income to pay their debt as well as regular day to day expenses?

2. What is the rate in each country?

3. Is Suicide rate affected recently (decreasing/increasing)?

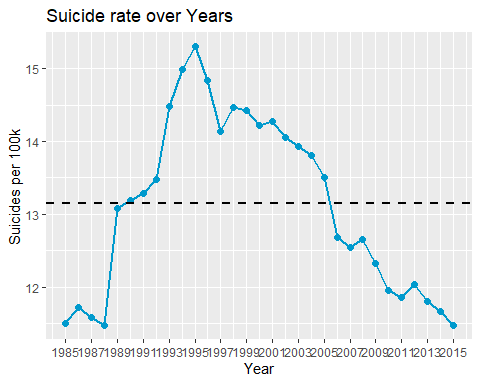
4. Does age affect suicide rate?

* + How you addressed this problem statement

Searched datasets across suggested online data stores. Each dataset found was different in its structure and properties. Finally got 3 dataset and verified data among them to bring it to final dataset. Finalized data is been verified for cleanliness with basic data hygiene checks like missing values, data types, duplicate values. Once data is finalized the statistical verification of the data was important to make sure its good fit for the analysis. Used histograms and basic 5 point summary of data to check mean, median, standard deviation, quartile ranges, etc. Used entire data population to answer questions. Plotting data using different visualizations helped in getting the questions answered with different view point. Correlation test was important in getting the meaningful variables for the regression analysis. Summarizing data by mean, length and visualization by different age group gave the suicide rate by age group. I verified my hypothesis of two similar economies should have similar suicide rate. In this case I used United State and United Kingdom. When verified found that my hypothesis was rejected. In last 10 years US has more suicide than UK. It was great learning through mistakes and hands on experience with driving the data story.

* + Analysis.

**Global Suicide Rate.**



- Average suicide rate for available data from 1985 to 2015 shows 13.5 per 100k of population.

- Highest rate of suicide was observed in 1995.

- This answers our question " Is Suicide rate affected recently (decreasing/increasing)?". Yes suicide rate decreased since 1995. If we compare suicide rate in 1995 and 2015 there is approximately 25% decrease.

**Suicide Rate by GDP** summary(lm(suicide\_sorted\_df$suicideRate~suicide\_sorted\_df$gdp\_per\_cap\_log,data=suicide\_sorted\_df))

##   
## Call:  
## lm(formula = suicide\_sorted\_df$suicideRate ~ suicide\_sorted\_df$gdp\_per\_cap\_log,   
## data = suicide\_sorted\_df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.013313 -0.011844 -0.006840 0.003758 0.212023   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.0111241 0.0008228 13.519 <2e-16 \*\*\*  
## suicide\_sorted\_df$gdp\_per\_cap\_log 0.0004290 0.0002070 2.072 0.0383 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.01897 on 27658 degrees of freedom  
## Multiple R-squared: 0.0001552, Adjusted R-squared: 0.0001191   
## F-statistic: 4.294 on 1 and 27658 DF, p-value: 0.03827

cor.test(suicide\_sorted\_df$suicideRate, suicide\_sorted\_df$gdp\_per\_cap\_log, method = "pearson")

##   
## Pearson's product-moment correlation  
##   
## data: suicide\_sorted\_df$suicideRate and suicide\_sorted\_df$gdp\_per\_cap\_log  
## t = 2.0721, df = 27658, p-value = 0.03827  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.0006736194 0.0242397371  
## sample estimates:  
## cor   
## 0.01245841

- The mean correlation was 0.878, indicates that there is strong positive linear relationship between GDP and number of suicides.

- This shows increase in wealth per person also affect suicide rate positively.

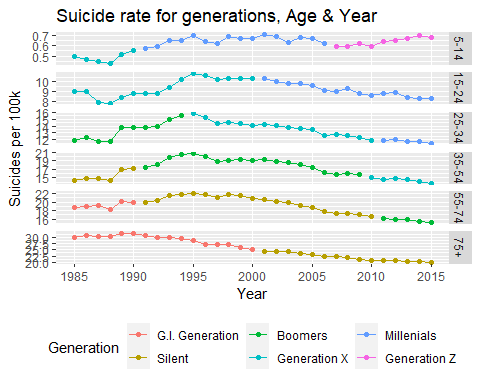
- The regression model shows p-value of 0.0057 which means there is little variance in suicide rate because of GDP.

- This means there is positive relation but is a weak relationship.

- However earlier graph of suicide rate over time shows that the suicide rates are decreasing over time.

- Hence more answers can be found by analyzing each country separately.

**Suicide rate by age**



by(suicide\_sorted\_df$suicides\_no, suicide\_sorted\_df$age, mean)

## suicide\_sorted\_df$age: 5-14  
## [1] 11.33709  
## ------------------------------------------------------------   
## suicide\_sorted\_df$age: 15-24  
## [1] 175.1295  
## ------------------------------------------------------------   
## suicide\_sorted\_df$age: 25-34  
## [1] 243.3497  
## ------------------------------------------------------------   
## suicide\_sorted\_df$age: 35-54  
## [1] 530.667  
## ------------------------------------------------------------   
## suicide\_sorted\_df$age: 55-74  
## [1] 358.7262  
## ------------------------------------------------------------   
## suicide\_sorted\_df$age: 75+  
## [1] 141.2714

- The above number shows that age group 5-14 has lowest number of suicides.

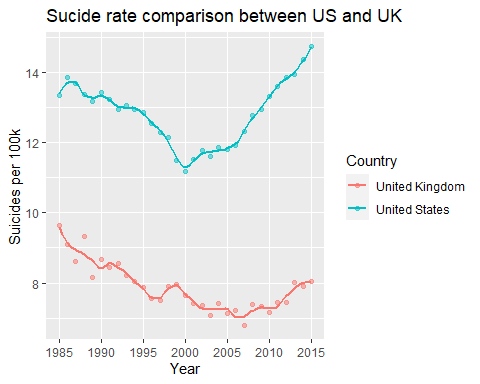
- Age group 35-54 has highest number of suicides globally.

anova <- aov(suicides\_no ~ age, data= suicide\_sorted\_df)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## age 5 7.596e+08 151914991 192.1 <2e-16 \*\*\*  
## Residuals 27654 2.187e+10 790783   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

- The p-value shown above is < 0.05. Hence we can say that there is significant difference in suicide rate across age group.

**Comparing suicide rate between similar economies (US and UK)**.



- UK rates shows lower and also decreased consistently over the years.

- US rate of suicide is quite volatile and shows decrease between 1995 to 2000. However there is equally sharp increase between 2000 to 2015.

* + Implications.

- Suicide rates are decreasing globally.

- There is a weak positive relationship between a countries GDP (per capita) and suicide rate.

- There are approximately 13 suicides per 100k for almost 40% of the countries.

- The above number shows that age group 5-14 has lowest number of suicides.

- Age group 35-54 has highest number of suicides globally.

* + Limitations.

The dataset do not have suicide reason data which will help to identify root cause and provide preventive measure. Also recent data for the suicide if mapped with covid pandemic impact will be good to add. May be we need to wait for another year to see the same. Also using unemployment data and verifying suicide rate against it might unfold how much unemployment affects suicide.

* + Concluding Remarks

The objective of project was to see the suicide rate in different age groups and impact of GDP over suicide rate. I learned a lot while doing this analysis. However I still feel that there is lot can be done with this data. There might be more tools which could help to get yet another insights from this data. What I can conclude for this analysis is -

- Higher living standard doesn’t mean country will have lower suicide rate.

- Middle age people have more suicides globally.

- US and UK being similar type of economies have significant difference in suicide rate between 2000 and 2015.

- Global suicide rate is decreasing over time.

This project could have been improved with other variables such as unemployment, stress level etc which would have improved suicide rate analysis.

As a final take away from this analysis is I am feeling confident to do more data stories than I was few weeks back.

Reference -

1. https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

2. https://www.kaggle.com/szamil/who-suicide-statistics

3. https://healthdata.gov/dataset/vital-statistics-suicide-deaths-age-group-raceethnicity-resident-county-region-and-gender

4. Field, A., J. Miles, and Z. Field. 2012. Discovering Statistics Using R. SAGE Publications.

5. Lander, J. 2017. R for Everyone (Second Edition). Addison Wesley Publications.

**DSC 520 Final Project Template**

This handout is meant to provide you with some structure for your final project. Note that there are two parts to this handout. First is the description of what you will be doing in each section and their general due dates. Second is the template itself with the various headings.

You are free to adjust this as needed. However, given what you’ll be doing, it is advantageous to you to just follow the template as it will help keep you focused. Should you have any questions, please ask!

**Section 1 – Week 8 – Getting Started**

* Provide an introduction that explains the problem statement you are addressing. Why would someone be interested in this?
* Draft 5-10 Research questions that focus on the problem statement.
* Provide a concise explanation of how you plan to address this problem statement.
* Discuss how your proposed approach will address (fully or partially) this problem.
* Do some digging on a dataset that you can use to address the issue.
  + Original source where the data was obtained is cited and, if possible, hyperlinked.
  + Source data is thoroughly explained (i.e. what was the original purpose of the data, when was it collected, how many variables did the original have, explain any peculiarities of the source data such as how missing values are recorded, or how data was imputed, etc.).
* Identify the packages that are needed for your project.
* What types of plots and tables will help you to illustrate the ﬁndings to your research questions?
* What do you not know how to do right now that you need to learn to answer your research questions?

**Section 2 – Week 9 – Cleaning Your Data and Exploratory Data Analysis**

* Data importing and cleaning steps are explained in the text and in the DataCamp exercises (tell me why you are doing the data cleaning activities that you perform) and follow a logical process.
* With a clean dataset, show what the final data set looks like. However, do not print off a data frame with 200+ rows; show me the data in the most condensed form possible.
* What do you not know how to do right now that you need to learn to import and cleanup your dataset?
* Discuss how you plan to uncover new information in the data that is not self-evident.
* What are different ways you could look at this data to answer the questions you want to answer?
* Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.
* How could you summarize your data to answer key questions?
* What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).
* What do you not know how to do right now that you need to learn to answer your questions?
* Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Suggestion from the course professor: Some additional questions you may want to consider asking yourself as you work through this section of the project:

1. What features could you ﬁlter on?
2. How could arranging your data in different ways help?
3. Can you reduce your data by selecting only certain variables?
4. Could creating new variables add new insights?
5. Could summary statistics at different categorical levels tell you more?
6. How can you incorporate the pipe (%>%) operator to make your code more efﬁcient?

**Section 2 – Week 10**

* Overall, write a coherent narrative that tells a story with the data as you complete this section.
* Summarize the problem statement you addressed.
* Summarize how you addressed this problem statement (the data used and the methodology employed).
* Summarize the interesting insights that your analysis provided.
* Summarize the implications to the consumer (target audience) of your analysis.
* Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.
* In addition, submit your completed Project using R Markdown or provide a link to where it can also be downloaded from and/or viewed.