SUMMARY

* **Statistical/Hypothetical Question**

Questions I am planning to focus on (initial list, may change as I progress):

1. Is there any relation between the suicide rates and gdp (developed/developing countries)?
2. Does the above relation change if we consider age group?
3. Which age group is experiencing highest suicide rates in the recent times?
4. What are the various factors affecting the suicide number?

* **Outcome of your EDA**

From the analysis I could find out that there is not much correlation between suicides number and GDP. Suicide rates were high in the age group of 35-54. I have tried to find out the correlation between different variables and suicides number. Population and HDI\_for\_year have positive correlation with the suicides number. I have conducted the regression analysis with the above variables and the result is satisfactory.

* **What do you feel was missed during the analysis?**

I wanted to find out the outliers and try remove them for the analysis part. But I couldn’t find out any for further action. The suicide numbers are mainly dependent on population, which is why at this stage we cannot say that smaller numbers or higher numbers are outliers. Because population is dependent on country size. Which is why I have tried to plot the histogram for suicides\_per\_100k\_pop. No visible outliers from that plot as well.

* **Were there any variables you felt could have helped in the analysis?**

What are the other factors affecting the suicides number was unknown at this time. Particularly I want to give more emphasis on which concludes the HDI.

* **Were there any assumptions made you felt were incorrect?**

I thought GDP will have larger impact on suicides number. Because people in developed countries will have much more facilities compared to under developed countries and they will be happy. But my assumption was wrong and there is very little correlation between GDP and suicides number.

* **What challenges did you face, what did you not fully understand?**

I have tried so many analytical distributions for my data. But couldn’t find exact fit for this. Finally I have tried lognormal distribution. I want to apply log transformation to reduce the skewness of the feature but it is giving an error value because feature has zero values. Because log (0) is undefined. So I have used log(x+1) transformation. Which eliminated the problem of zero’s from the data and the result was satisfactory to me.