**TERM PROJECT DOCUMENTATION |   
ANALYSIS OF SUICIDE RATES**

Data Engineering 2

*MS in Business Analytics*

**TEAM**

Julianna Szabo -

Maeva Braeckevelt -

Dominik Gulácsy – 2003374

11/26/2020

**Global Data Flow Outline:**

**Data Sources:**

* Kaggle
* Rest Countries to search for alpha-2 codes of countries
* World Bank API

**Economic indicators worth considering:**

* GINI
* Unemployment Rate
* GDPPC
* HDI
* Alcohol cons | SH.ALC.PCAP.FE.LI and SH.ALC.PCAP.MA.LI
* Smoker | SH.PRV.SMOK.MA and SH.PRV.SMOK.FE

**Analytical Questions:**

* Has number of suicides increased over time?
* What age group tend to have higher suicide rates?
* Which gender has higher suicide rates? What factors maybe behind this
* Does this still stand if we look at genders in the different age groups? (Simpson’s Paradox)
* What if do this for Generations?
* Which countries tend to have higher suicides?
* How can it be related to socioeconomic indicators like GINI, Unemployment Rate, GDPPC, HDI?
* Pick an interesting country and look at it.

**Statistical techniques:**

* Histogram or boxplot
* Bar charts – percentage of total <- stacked bar charts
* Scatterplots
* Correlation Matrix with heatmap / Linear Regression matrix

**Extra:**  
Connect an R snippet to KNIME for advanced analysis

**Questions:**

1. ERD for one table??
2. What if data is quite clean?