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

Data Engineering 2

*MS in Business Analytics*

**TEAM**

Julianna Szabo - 2003870

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 considered:**

* Unemployment Rate | SL.UEM.TOTL.MA.ZS and SL.UEM.TOTL.FE.ZS
* Population | SP.POP.TOTL
* 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:**

* What age group tend to have higher suicide rates?
* Which gender has higher suicide rates?
* Which countries tend to have higher suicides?
* How can it be related to socioeconomic indicators like, Unemployment Rate, GDPPC, alcohol consumption and cigarette consumption?

**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  
API

Data

The dataset we chose to work with, is dataset that collected data on suicide. We found it on Kaggle. We decided to work with this dataset mainly because it had a lot of data as well as a large of variety. Countries, years, sex, ages, HDI, generation, suicide per 100k habitants, gdp per capita were the variable available that could help us for a powerful analysis (details information on the variable is on the variable.xlsx file on Github). However, to understand better the suicide rate, we thought that other variables could be interesting, that’s why, we chose among the world bank data website : alcohol consumption, cigarette consumption and unemployment rate.  
All the engineering work we have done were aim to give us a visualization to answer to thoses question : What age group tend to have higher suicide rates? Which gender has higher suicide rates? Which countries tend to have higher suicides? How can it be related to socioeconomic indicators like, Unemployment Rate, GDPPC, alcohol consumption and cigarette consumption?  
For the alcohol and cigarette consumption, we decided to separate the male to the female. The choice was made because their consumption is usually not equal.

Data loading

For the dataset, we chose to load it on SQL from the csv. It was an easy action such as we learn how to do it in DE1. However, to facility the reproduction of our analysis, we decided to host it on AWS, procedure that just learned to do in in DE3.

Concerning the WDI’s variable, an API extraction was used in Knime. We used a xlsx hosted on onedrive for the same practical reason.

ETL data pipeline

The cleaning and and the analytics were done in Knime. After connecting Knime to the sql, we cleaned the dataset. The main difficulty that occurs was the “Dominik I need you help for this part!”

Visualization

Once we got the data joined and cleaned and could start the visualization.