**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 these 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. We chose to also add the population from WDI for data quality purpose. In Kaggle, it is not said where the population data was coming from, so we opted for a universal and reliable source such as World Bank is.

<data citation for Kaggle and World Bank API>

# Data Infrastructure

*Kaggle Suicide Rates Dataset*

First, we downloaded the csv file (data/static\_kaggle/suicide\_rate.csv) from [Kaggle](https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016). We wrote an SQL file (data/static\_kaggle/import\_suicide\_rates.sql) in which we imported the csv data file into a local MySQL database to check if it loads properly and explore the data. We wanted to make our workflow as reproducible and dynamic as we can, so we decided to host a database. We created a MySQL database instance using the RDS service of AWS. We used the AWS account provided by CEU. Our julmaedom-mysql-de2 named DB instance was configured as a db.t2.micro instance with 20 GiB storage. We set up a connection to the instance in MySQL Workbench. We exported the data and structure of our locally hosted DB in a Self-Contained sql file (data/aws/dump\_suicide\_rates.sql). Then we imported the data using this dump to our MySQL on AWS. We added a new user named grader that has only read access to simulate a real-life setup. Credentials for this user and connectivity details can be find in a text file (data/aws/grader\_credentials.txt).

*REST Countries API*

In order to properly and universally identify countries we have to use country codes, more specifically alpha2 country codes. To have this data we used the REST Countries API to search for the country codes of countries in the suicide rates data. This functionality is called name and available through https://restcountries.eu/rest/v2/name/{name}.

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!”

Further cleaning of the data

Once we had a clean data table, we needed to clean It further for visualization. First we filtered out all the missing values from any of the columns. Once that was done, we realized that most observations were for the year 2010. While we originally wanted to do a time series analysis, at this point we realized we would not have the data to do so. Therefore, we decided to switch to a cross-sectional analysis using data from only the year 2010, after we confirmed the significant difference in number of observations for that year using a histogram.

# Visualization

Once we got the data joined and cleaned and could start the visualization. The main challenge here were the different variables that needed to be visualized to give us a good overview of the factors that may affect suicide rates. There were two main groups of variables. The first one was about the whole population, such as GDP per Capita, while others, such as smoking habits, were differentiated between male and female. Therefore, it required two different paths in Knime.

For the ones containing the whole population the aggregation process was relatively simple, since it only needed to be grouped by one variable. I used the sum function for the number of suicides (since this is a flow variable), and the mean for the GDP per Capita as well as the population.

For the population, we created a simple scatter plot at first to see if there was any correlation. The scatter plot shows that there is a general positive correlation between population and the number of suicides. The smaller the population the less suicides, however larger populations do not necessarily have nigher suicides rates. Overall, it seems there is not much correlation between these two variables.

For GDP per Capita, we first created a scatter plot using level-level analysis but decided that a log-log transformation would best represent this data. It shows how one percent change in GDP per Capita could affect the percentage number of suicides. It shows that there seems to be no correlation between percentage change in GDP per Capita has no effect on the percentage change of the number of suicides.

For the variables that have a division by sex, we needed to take a couple of extra steps. First, we aggregated based on country and sex. We then needed to create two paths that filtered for the two sexes, since they were represented in different variables. We again decided to take the sum for the number of suicides again, but the mean for all the other variables. This was done, due to the repetition of the variables, since we were combining different age groups for the suicide rates that were not separate categories for the variables from the World Bank Database.

For the unemployment and the smoking rates both sexes showed that there was not a huge positive correlation to suicide rates. There was a tendency towards the center, meaning that people with an average smoking habit were most likely to commit suicide, but no clear correlation.

For alcohol consumption we decided to do a log-log analysis. This showed significantly more interesting results. There is a cluster around the higher percentage of alcohol consumption and percentage of suicides. Therefore, there is a correlation between the percentage change in alcohol consumption and the percentage change in number of suicides. We cannot claim causality, but there seems to be a connection between these two variables, much more than any others.