TERM PROJECT 2 | *Exploring the relationship between suicides rates and the GDP per capita as well as the unemployment rate among countries*

Data Engineering 1: SQL and Different Shapes of Data

MSc in Business Analytics 2021
Central European University

Team Manila
Haaris Cheema
Muhammad Talha Zahid
Peter Endes-Nagy
Sabina Umarova

Analysis questions

- What is the relationship between suicide rates and economic development (GDP per capita) among countries?
- What is the relationship between suicide rates and unemployment rates among countries?

Data Sources

- World Health Organization
- World Bank API

Indicators Considered

- WHO Suicide rates
- Population - WB API ID: *SP.POP.TOTL*
- GDP per capita - WB API ID: *NY.GDP.PCAP.CD*
- Unemployment rate - WB API ID: *UEM.TOTL.ZS*

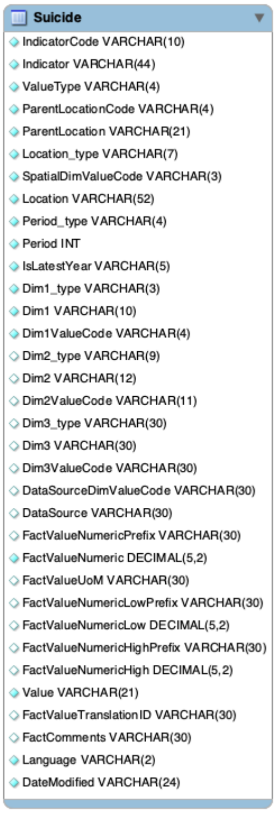
Data Collection

1. Suicide rates

World Health Organization database was used to collect data on suicide rates among countries between 2000 and 2019. The dataset on suicide rates contains information on suicide mortality rates per 100,000 population for a specific region, country, gender and age group.

Data obtained from WHO was transformed into .sql format using Freeware Software to make the project easily reproducible on any computer without being sabotaged by secure-file-priv issues in MySQL. Unnecessary columns were dropped, and variables were renamed into meaningful names. The rest of the data cleaning process was executed in KNIME.

For the EER Diagram please refer below:

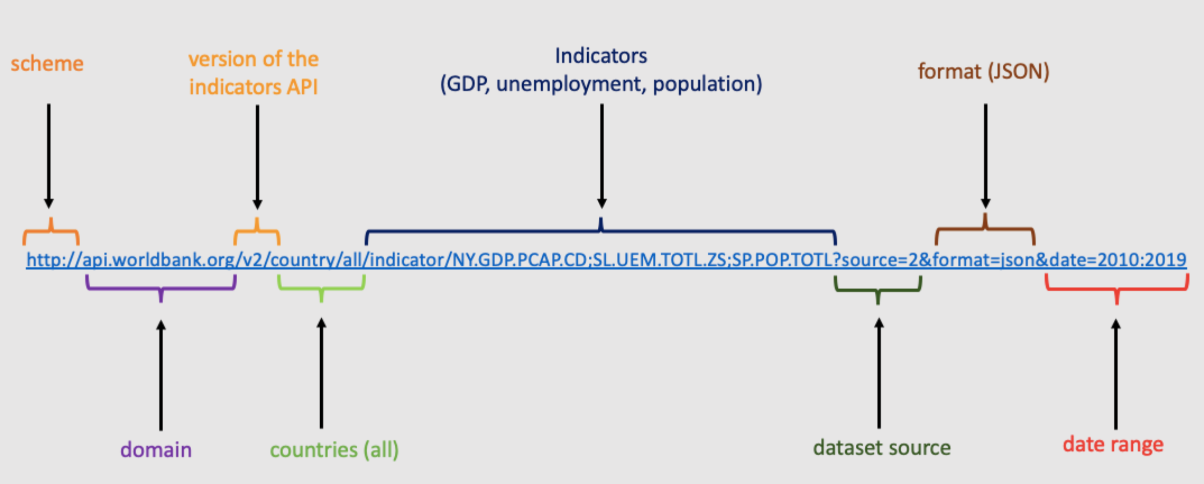


Suicide

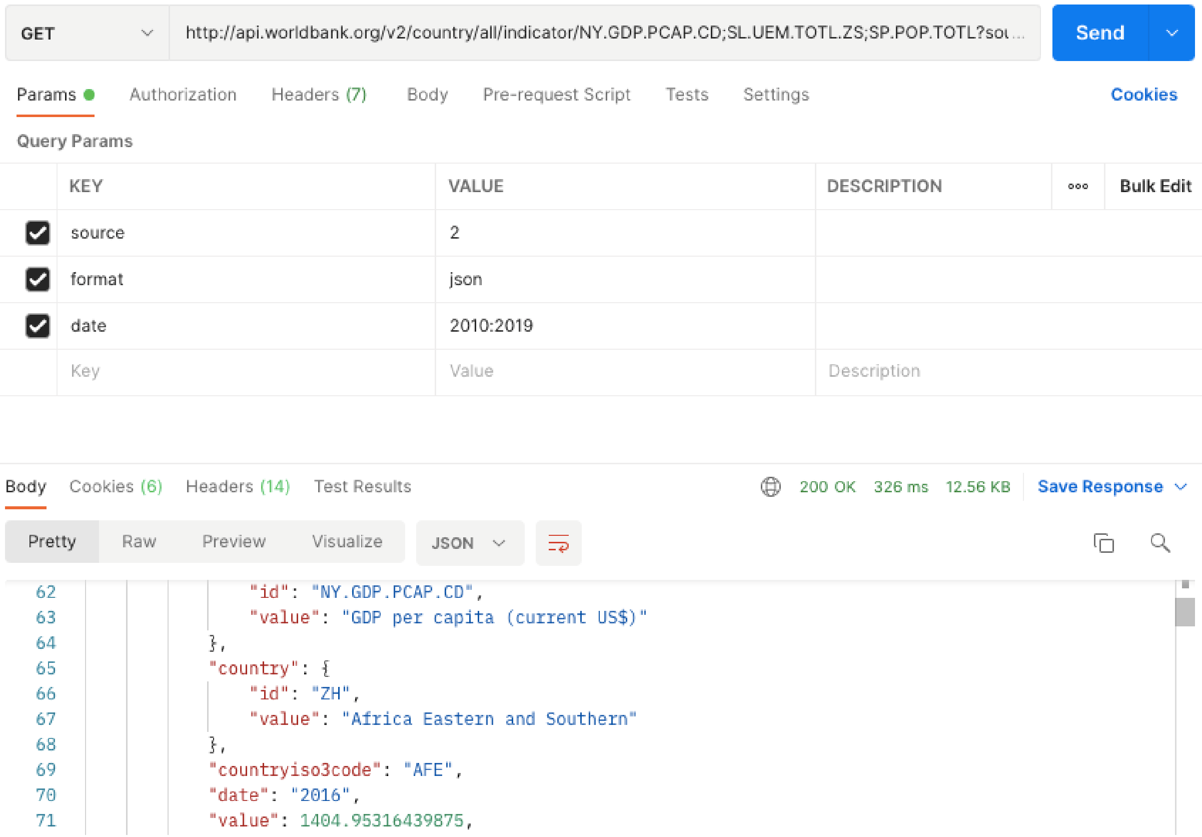
◆ IndicatorCode VARCHAR(10)
◆ Indicator VARCHAR(44)
◆ ValueType VARCHAR(4)
◆ ParentLocationCode VARCHAR(4)
◆ ParentLocation VARCHAR(21)
◆ Location_type VARCHAR(7)
◆ SpatialDimValueCode VARCHAR(3)
◆ Location VARCHAR(52)
◆ Period_type VARCHAR(4)
◆ Period INT
◆ IsLatestYear VARCHAR(5)
◆ Dim1_type VARCHAR(3)
◆ Dim1 VARCHAR(10)
◆ Dim1ValueCode VARCHAR(4)
◇ Dim2_type VARCHAR(9)
◇ Dim2 VARCHAR(12)
◇ Dim2ValueCode VARCHAR(11)
◇ Dim3_type VARCHAR(30)
◇ Dim3 VARCHAR(30)
◇ Dim3ValueCode VARCHAR(30)
◇ DataSourceDimValueCode VARCHAR(30)
◇ DataSource VARCHAR(30)
◇ FactValueNumericPrefix VARCHAR(30)
◆ FactValueNumeric DECIMAL(5,2)
◇ FactValueUoM VARCHAR(30)
◇ FactValueNumericLowPrefix VARCHAR(30)
◇ FactValueNumericLow DECIMAL(5,2)
◇ FactValueNumericHighPrefix VARCHAR(30)
◇ FactValueNumericHigh DECIMAL(5,2)
◆ Value VARCHAR(21)
◇ FactValueTranslationID VARCHAR(30)
◇ FactComments VARCHAR(30)
◆ Language VARCHAR(2)
◆ DateModified VARCHAR(24)

2. World development indicators

Population, unemployment rates and GDP per capita for each country between 2000 and 2019 years were obtained from the World Bank database using World Bank API. Firstly we obtained the structure of the query string and figured out how to send request to API using datahelpdesk.worldbank.org. We requested all three indicators together for the required period in JSON format. Further JSON files were transformed into data tables in KNIME. For the breakdown of World Bank API request please refer below:


scheme
version of the indicators API
Indicators (GDP, unemployment, population)
format (JSON)
http://api.worldbank.org/v2/country/all/indicator/NY.GDP.PCAP.CD;SL.UEM.TOTL.ZS;SP.POP.TOTL?source=2&format=json&date=2010:2019
domain
countries (all)
dataset source
date range

We used Postman in order to check the validity of the request. For the Postman result please refer below:



http://api.worldbank.org/v2/country/all/indicator/NY.GDP.PCAP.CD;SL.UEM.TOTL.ZS;SP.POP.TOTL?sou...

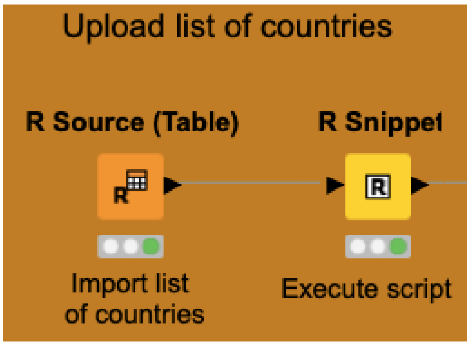
	KEY	VALUE	DESCRIPTION
☑	source	2	
☑	format	json	
☑	date	2010:2019	
	Key	Value	Description

```
            "id": "NY.GDP.PCAP.CD",
            "value": "GDP per capita (current US$)"
        },
        "country": {
            "id": "ZH",
            "value": "Africa Eastern and Southern"
        },
        "countryiso3code": "AFE",
        "date": "2016",
        "value": 1404.95316439875,
```

Data Workflow

1. List of countries

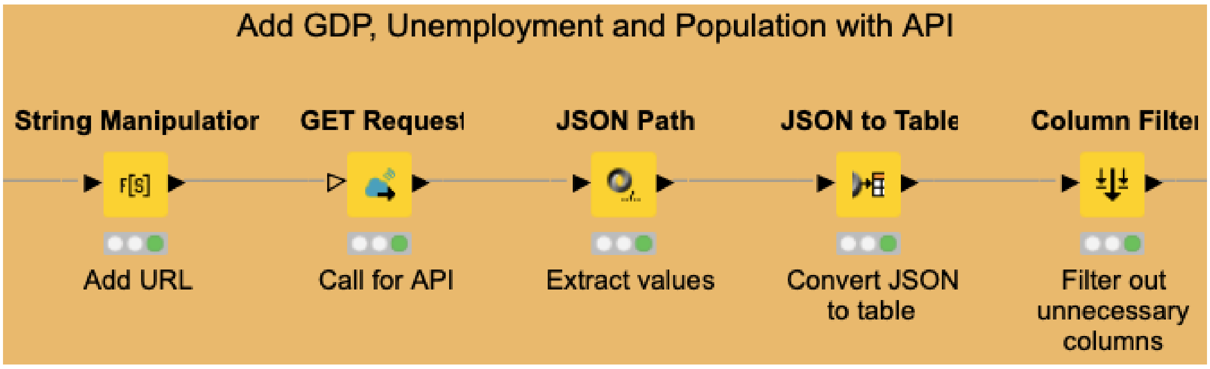
Firstly, we decided to use R script to obtain the list of all countries. Using WDI package in R we obtained the list of countries and excluded aggregated groups of countries such as European Union, Arab World, East Asia and others. For the nodes used in KNIME workflow to load list of countries using R please refer below:



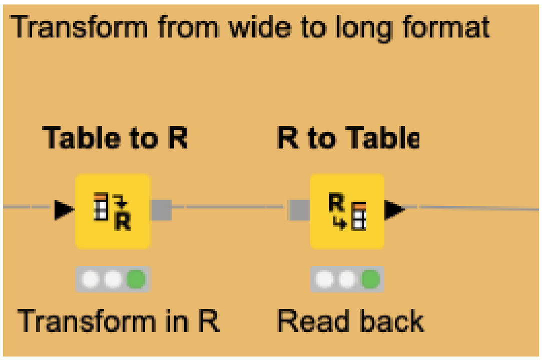
2. API

Further we wrote the URL for World Bank API using String Manipulation node, which is join("http://api.worldbank.org/v2/country/",$knime.in.code$,"/indicator/NY.GDP.PCAP.CD;SL.UEM.TOTL.ZS;SP.POP.TOTL?source=2&format=json&date=2010:2019").

The URLs for each country were added in a new column. A get request was placed and the values of the indicators were extracted using JSON Path. Finally, we converted the JSON format file into a table. We filtered out unnecessary columns, so we have only Country Name, Country Code and data values. For that part of KNIME workflow please refer below:

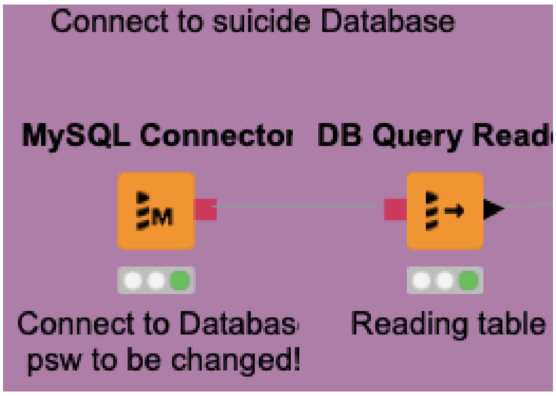


The table was resulted in wide format, so in the next step, we transformed it into a long format using an R Script. A necessary step for joining with suicide data since the latter is in long format. For that part of KNIME workflow please refer below:



3. Suicide database

Using MySQL Connector node in KNIME we loaded the suicide database. DB Query Reader allowed us to execute SQL query and the data was read as a data table in KNIME. For that part of KNIME workflow please refer below:



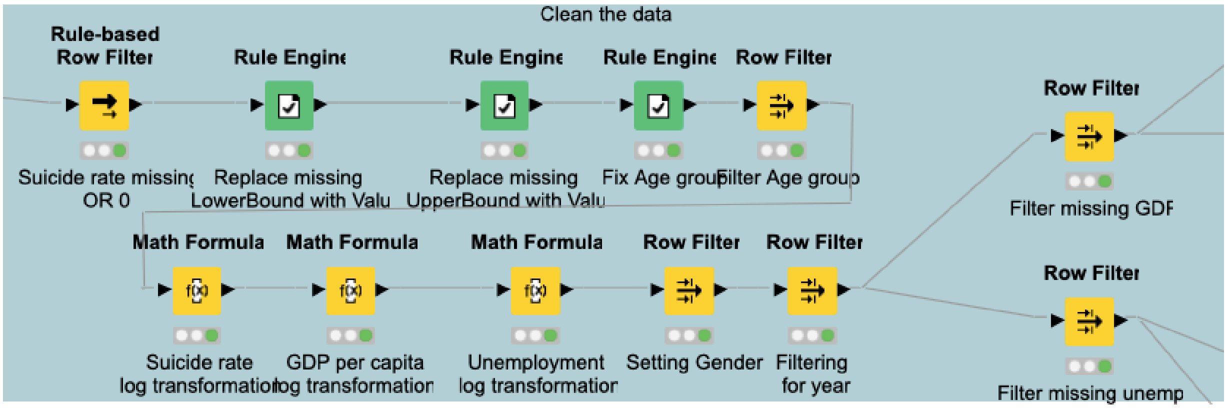
The .sql file shall be run before the KNIME Workflow and the credentials changed according to the local configuration.

4. Join and clean data

Suicide and WDI data were joined with inner join by using year and country. The resulting table was subject to further cleaning:

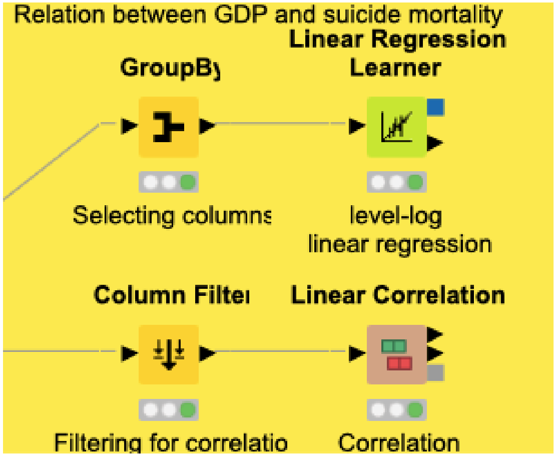
- We dropped every row where suicide rate was missing. We also considered zero as missing value, since most probably a measuring error. Node used: Rule-based Row Filter.
- In many cases where the suicide rate estimates were exact values, the lower and upper bound estimates were missing. We replaced the missing values with the exact values so the analysis can use both lower both upper bound estimates. Node used: Rule Engine.
- For the age group variable there was no data by age groups before 2019. Oddly enough, the Totals were missing values, so we replaced, then we also filtered for total only. Node used: Rule Engine and Row filter.
- Preparing for the regression, we created new variables with log transformation: log_GDP, log_unempl and log_suicide. Node used: Math Formula.
- Filters were also added to set the year and gender. It gives flexibility to the user for choosing for which year and which gender they'd like to run the regressions. For this report, all the analysis will be done for 'both sexes', and the year '2019'. Node used: Row filter
- To find the individual correlations of GDP per capita and unemployment rate with the suicide rate, we filtered out the missing values separately so we can keep more datapoints for the distinct regressions (unemployment rate might be missing for some countries while the GDP data is available).

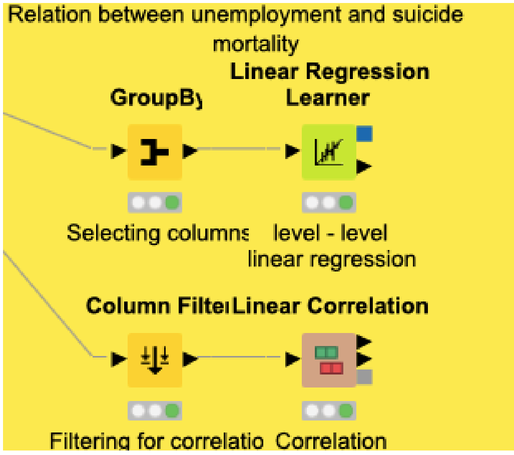
For that part of KNIME workflow please refer below:



5. Analytics

Upon filtering out the missing values, we used the group-by node to select the relevant columns on which the correlations will be calculated, and the regressions will be performed. We performed linear regressions and correlations for the suicide and GDP relationship and the suicide and unemployment relationship separately. For that part of KNIME workflow please refer below:


Relation between GDP and suicide mortality
GroupBy
Selecting columns
Linear Regression Learner
level-log linear regression
Column Filter
Filtering for correlatio
Linear Correlation
Correlation


Relation between unemployment and suicide mortality
GroupBy
Selecting columns
Linear Regression Learner
level - level linear regression
Column Filter
Filtering for correlatio
Linear Correlation
Correlation

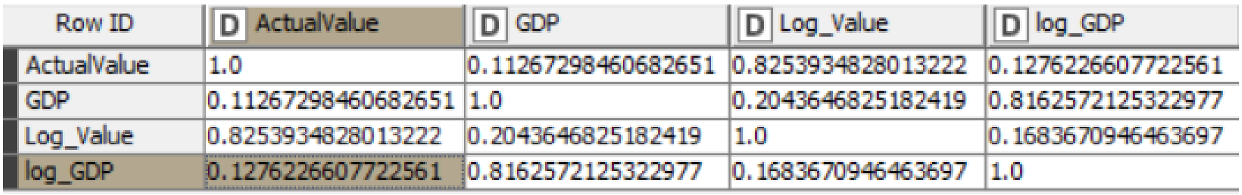
Results of the analysis

We ran the regressions for each possible combinations: level-level, level-log, log-level, log-log. We chose the models that seem to fit the most and also makes sense conceptually.

1. Impact of GDP per capita on the number of suicides (per 100,000 population)

<u>Correlation:</u>

We formed a correlation matrix for the GDP per capita, its log value, the suicide rates (Actual Value) and its log value. The results are summarized below.



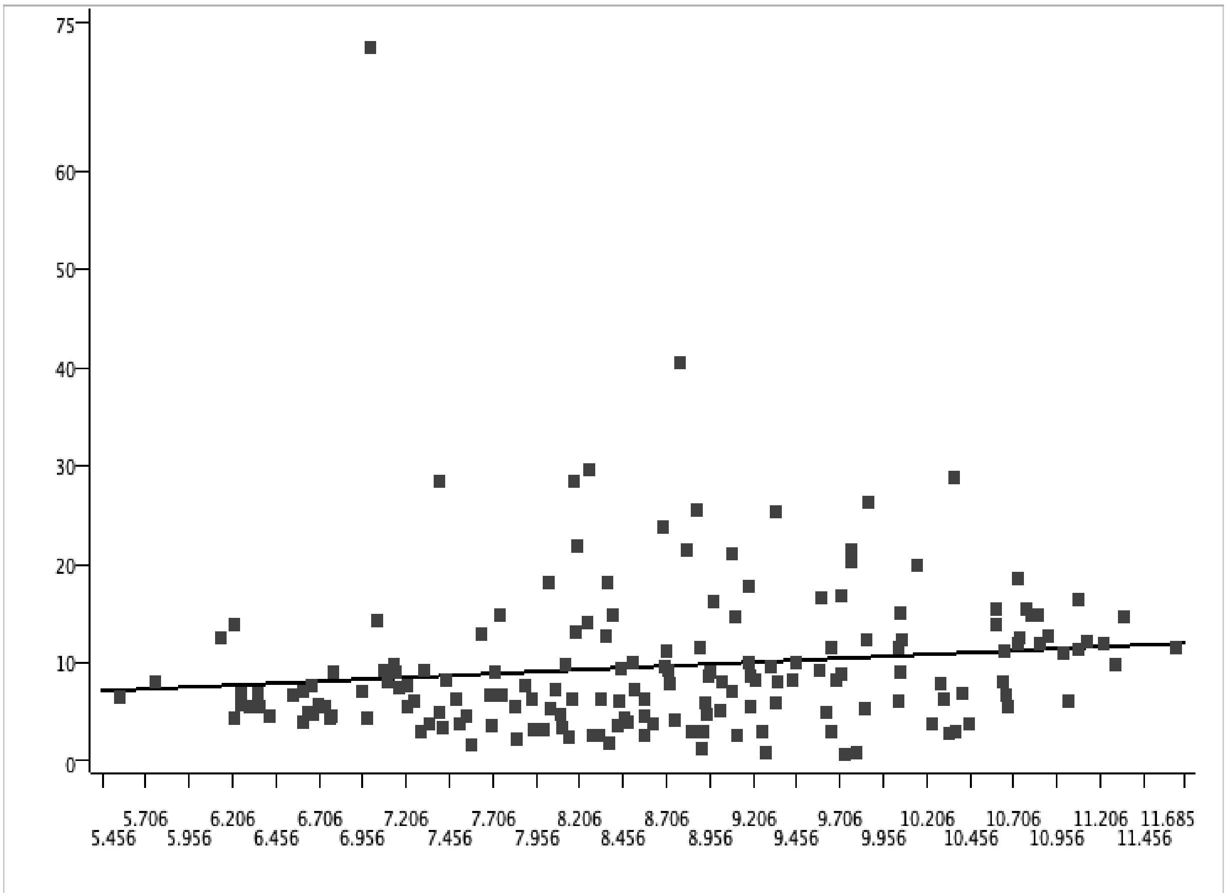
Row ID	D ActualValue	D GDP	D Log_Value	D log_GDP
ActualValue	1.0	0.11267298460682651	0.8253934828013222	0.1276226607722561
GDP	0.11267298460682651	1.0	0.2043646825182419	0.8162572125322977
Log_Value	0.8253934828013222	0.2043646825182419	1.0	0.1683670946463697
log_GDP	0.1276226607722561	0.8162572125322977	0.1683670946463697	1.0

The output indicates that there is a positive correlation between the number of suicides and the GDP per capita as well as the log of the GDP per capita. The magnitude of the correlation though is weak in both the cases.

Linear Regression:

We also ran a linear regression where we regressed the number of suicides on log_GDP as this model had a higher value for the correlation coefficient and makes most sense.

Variable	Coefficient	Standard Error	t - value	p - value
log_GDP	0.79	0.42	1.85	0.067
Intercept	2.70	3.77	0.72	0.475

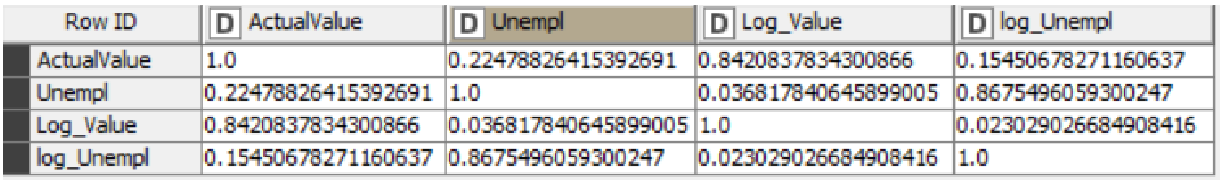


The regression output indicates that the log_GDP is a statistically insignificant variable in explaining the variation in the number of suicides if we test at a 5% level of significance. The scatterplot with the regression line shows the positive correlation between the two variables. However, due to the low strength of the correlation, there is not much variation in the number of suicides in a country as the GDP per capita increases.

2. Impact of unemployment rate on the number of suicides (per 100,000 population)

Correlation:

Like the previous case, a correlation matrix was formed between the unemployment rate, the suicide rates, and their respective log values.



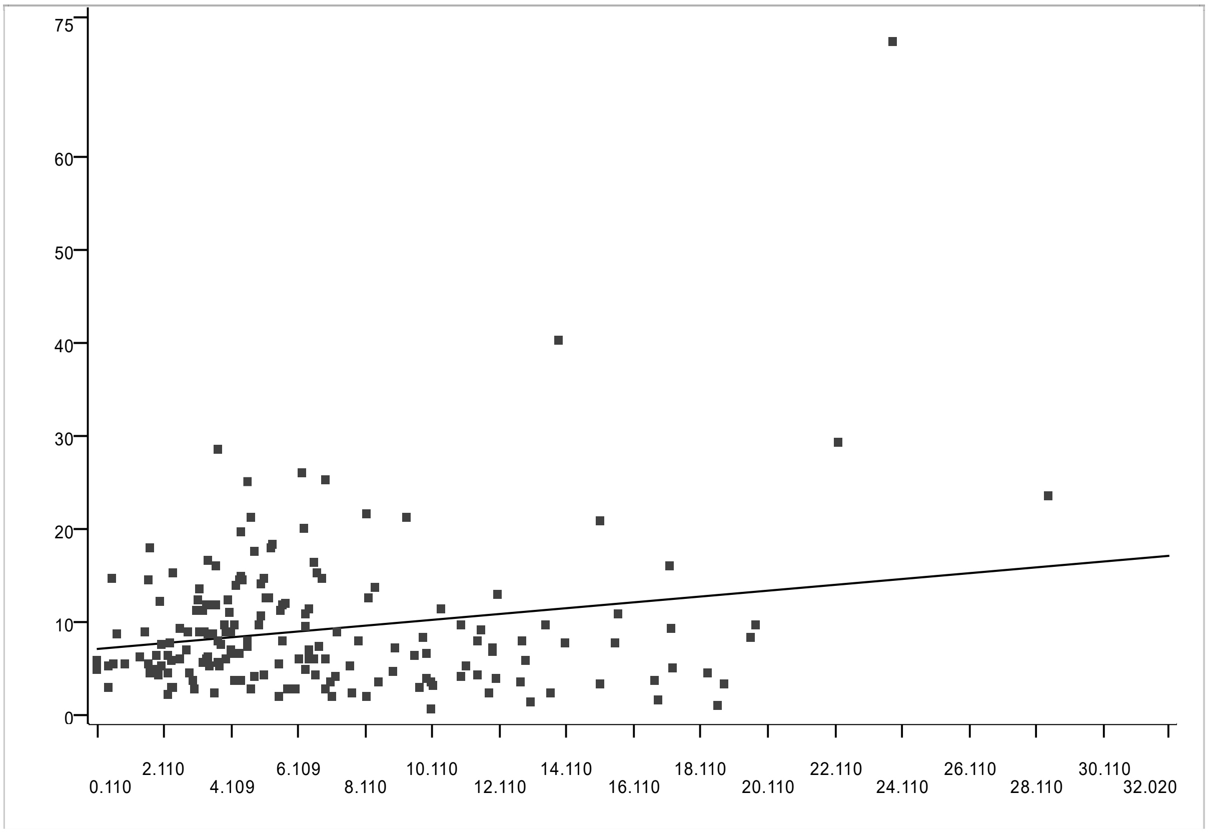
Row ID	D ActualValue	D Unempl	D Log_Value	D log_Unempl
ActualValue	1.0	0.22478826415392691	0.8420837834300866	0.15450678271160637
Unempl	0.22478826415392691	1.0	0.036817840645899005	0.8675496059300247
Log_Value	0.8420837834300866	0.036817840645899005	1.0	0.023029026684908416
log_Unempl	0.15450678271160637	0.8675496059300247	0.023029026684908416	1.0

The output in this case indicates a positive correlation between the unemployment rate as well its log with the actual number of suicides in a country. The correlation is stronger in the case of the of the level-level correlation.

Linear Regression:

Running a level-level model makes most sense. The regression output is shown below.

Variable	Coefficient	Standard Error	t - value	p - value
Unemployment	0.31	0.11	2.79	0.006
Intercept	7.19	0.96	7.46	3.73E-12



The regression output indicates the unemployment is a statistically significant variable in explaining the variation in the number of suicides when tested at a level of significance of 5%. The regression line also reflects the positive correlation between the two variables. As the unemployment rate increases, we see a gradual rise in the number of suicides (per 100,000 population) in a country.

Conclusion

To conclude the project, we analyzed the relationship between suicide rates and GDP per capita as well as unemployment. We used R, MySql and World Bank API to build a database and get data set. We used KNIME to document workflow. We found out positive correlation between suicide rates and unemployment rates and its statistically significant at 5%. We also found out that positive correlation between suicide rates and GDP per capita is statistically insignificant at 5%.

Who did what?

- Haaris Cheema – Knime Workflow, Analytics, Documentation, Presentation
- Muhammad Talha Zahid - MySQL, Documentation
- Peter Endes-Nagy - Knime Workflow, Analytics, Documentation, Presentation
- Sabina Umarova – API, Postman, Knime Workflow, Documentation, Presentation