DATA ENGINEERING 1: DIFFERENT SHAPES OF DATA

Team Hong Kong

11/15/2021

A use case study on the relationship between suicide rate and GDP per capita (2010-2016)

In this project we examine the distributions of the suicides committed during the 2010 and 2016 time period. We found that the number of suicides in relation to the unemployment rate of a country shows a normal distribution, if we exclude the outlier values. Also most suicides are committed in the 35-74 age range and more than 75% are committed by males regardless of age. Lastly the number of suicides committed per year shows a steady declining tendency between 2010 and 2015.

Introduction

Due to the COVID-19 pandemic, significant amount of people lost their jobs and possibly even more people got depressed between the seemingly endless lockdowns and mandatory home offices. The 4th wave is upon us, this gave us inspiration to take a deep dive into the suicide numbers of the world. The primary goal of the project was to see if there is a connection between the suicides committed and various economic indicators. Is there a visible trend, are the countries with more unemployment tend to have residents more willing to end their own lives? Perhaps it's the other way around and more work causes more stress which may lead to higher numbers of suicide? Is there a relationship between these variables at all?

To answer these questions we collected data on suicides from all countries between the time period of 1985 and 2016. The unemployment rate is obtained from the World Bank Data via the World Bank API. This API requires country codes as an identifier, which is joined with the suicide data in Knime. The suicide data is obtained from Kaggle, loaded into SQL before joining it with the country codes, as mentioned before. We imported both the country codes and the suicide data in CSV format. We used a scatter plot, bar charts and a sunburst chart to visualize the data.

Technical choices

We chose to build our data pipeline in Knime containing our complete workflow. As a first step we imported the country data from the World Bank API into Knime and did the cleaning there (Figure 1).

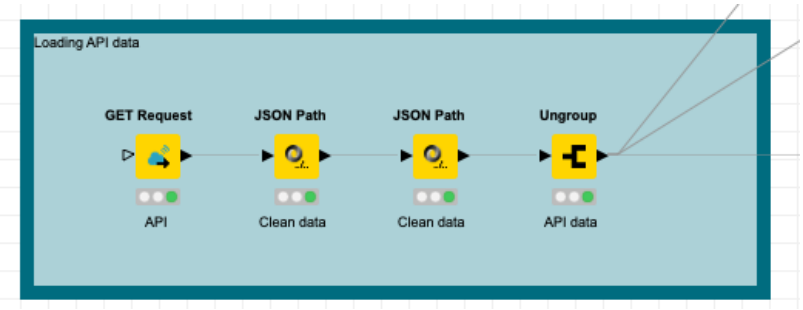


Figure 1: Loading data from Word Bank API

Through the API we obtained the unemployment rate, literacy and GDP per capita of each country. The breakdown of the API can be seen in Figure 2 and this is the node GET Request to the API:

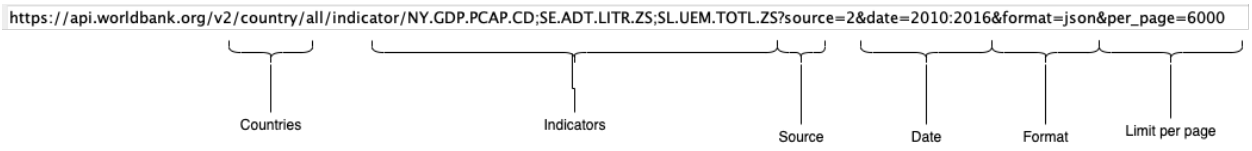


Figure 2: Word Bank API indicators

Here country/all stands for country code, indicator (NY.GDP.PCAP.CD) stands for GDP per Capita, indicator (SE.ADT.LITR.ZS) stands for adult literacy, indicator (SL.UEM.TOTL.ZS) stands for unemployment rate, source stands for source of the data set, date is the time period between 2010 and 2016, format=json means data is in JSON format and per_page=6000 means the request is limited in 6000 results.

The GET Request node sends a Request to the API, as a result we get data in JSON format (Figure 1). Container Output (JSON) gives us a formatted view of the result. Then we use the JSON Path nodes to clean the data and get rid of unnecessary header and page numbers. The ungroup node was later used to convert the data into table format. However the table needed further cleaning before we could run our analysis on it. This concludes the extract part of the ETL process.

After transforming the data into table format with the ungroup node, we split the data into three separate tables and joined them on the countries (Figure 3). The three main variables called by the API were the GDP per capita, Unemployment rate and Adult literacy rate. The use of nominal value filter along with rule engine was done precisely to extract value out of each observation and convert it to a variable.

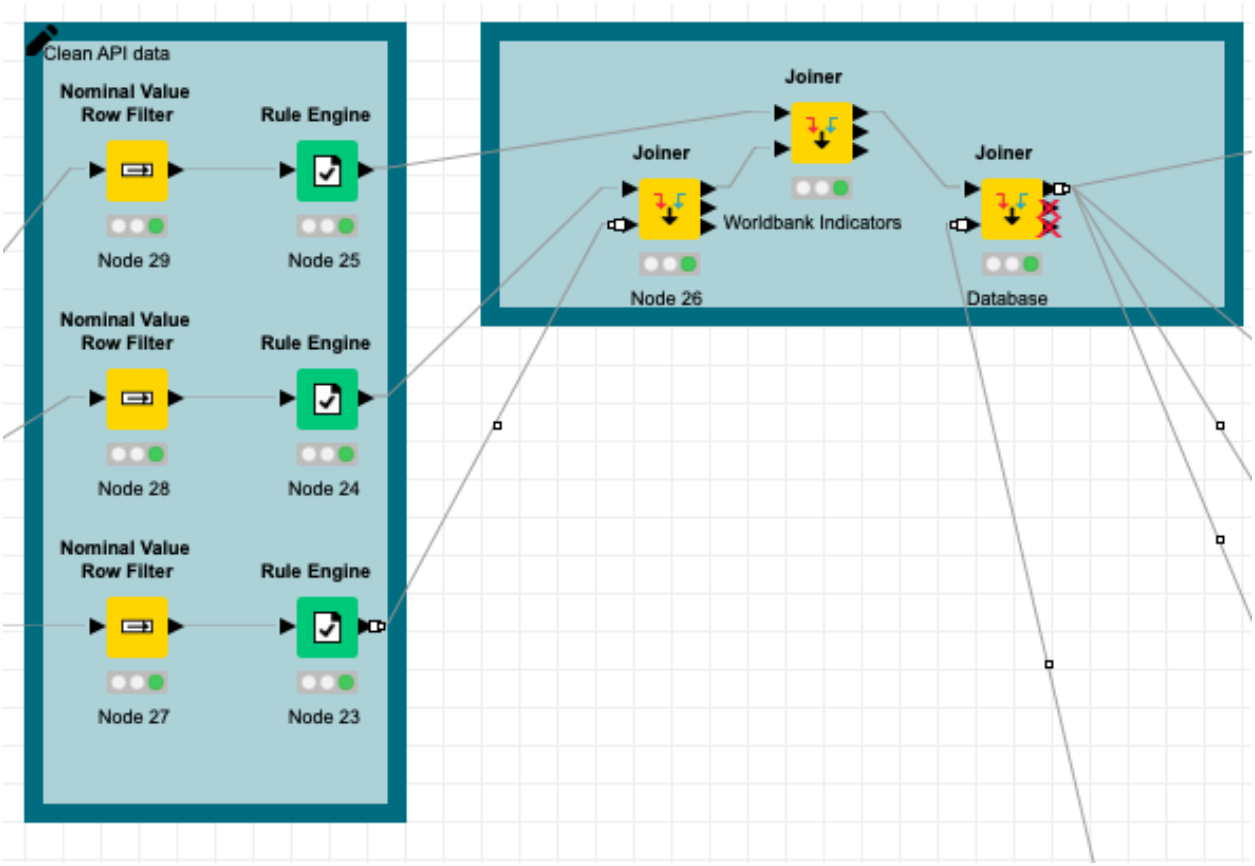
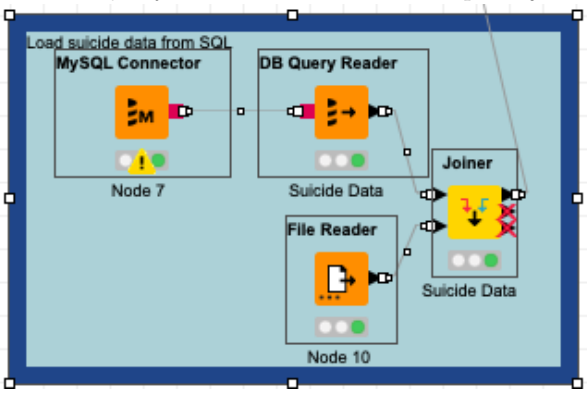


Figure 3: Cleaning the API Data

In the meantime we loaded the suicide data from SQL into the Knime workflow using MySQL connector and DB Query Reader. In the next step we joined it with the three-letter country code table (figure 4).



As the last step before we began the data visualization, we made our Data Warehouse by merging the suicide data with the transformed World Bank API data (Figure 3).

Data Model

In this section we present our data model for the project. Figure 5 shows the EER diagram of our data layer. There are two separate data tables, one contains the suicide data and the other contains the GDP per capita, unemployment rate and adult literacy per country. These two tables were joined on country codes. Initially the suicide table didn't contain the country codes, so we joined it with a table containing the country codes for World Bank.

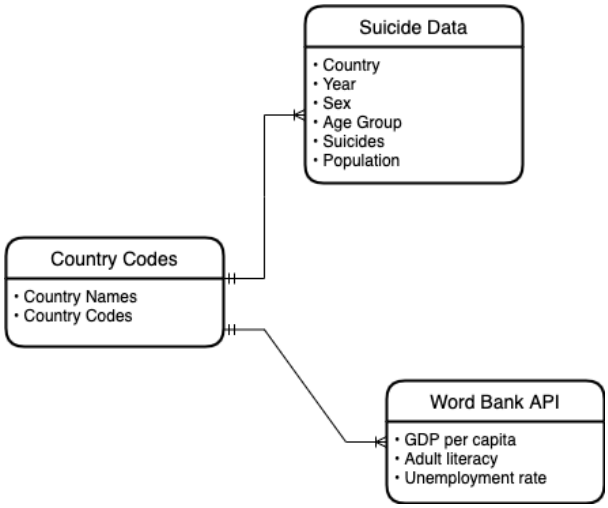


Figure 4: EER Diagram

Visualization

During the visualization we filtered the data for 2015, except for one barchart, where we compared the suicides committed word wide between 2010 and 2016. For the unemployment - sum of suicides scatter plot we grouped the data by the countries. For the sunburst chart we grouped data by sex and then by age groups. Lastly for the sum of suicides barchart we grouped the data by the age groups. Figure 6 displays visualization part in the worlflow.

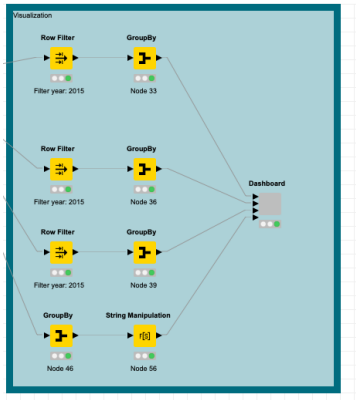


Figure 5: Visualization workflow

The charts can be seen below:

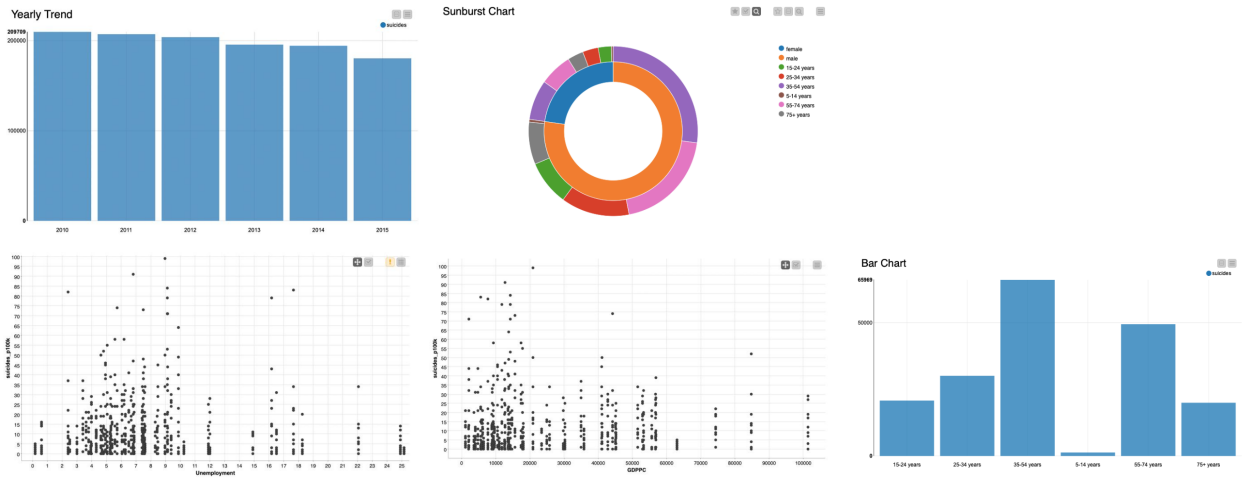


Figure 6: Charts

Conclusion

According to our analysis we concluded that male suicide rate were higher as compare to female. Moreover, based on our result we witness that major occurrence of suicide belong to generation X (35-54 years).

In our comparison with economic indicators and suicide rate the graph illustrated that countries with higher employment rate had more reported suicide cases.

Whereas in our analysis with GDP per capita we found that countries with lower GDPPC had higher number of suicides. Depicting negative correlation between # of suicide and their standard of living.

Below is the job descriptions on each team member:

1. Fatima Arshad: Knime workflow
2. Maryam Khan: Finding dataset loading data in SQL & API
3. Khawaja Hassan: Visualisation and Presentation
4. Oszkar Egervari: Documentation