TERM PROJECT DOCUMENTATION | ANALYSIS: COUNTRIES' LIFE SATISFACTION

Data Engineering 2: Different Shapes of Data

MSc in Business Analytics
Central European University

TEAM
Viktória Mészáros - 2002392
Brúnó Helmeczy - 2003888
Attila Serfőző - 2003629

Prepared for:
Miklós Koren
László Salló

Date of Submission:
11/12/2020

Data Flow Overview:

Data Sources

- Eurostat Database & API
- World Bank API

Indicators Considered

- Population | World Bank ID: SP.POP.TOTL
- GDP per capita | World Bank ID: NY.GDP.PCAP.PP.KD
- Life Expectancy | Eurostat ID: DEMO_MLEXPEC$DV_292
- Employment Rate | Eurostat ID: LFST_HHEREDCH$DV_324
- Greenhouse Emission per capita | Eurostat ID: SDG_13_10
- Average Weekly Working Hours | Eurostat ID: LFSA_EWHUN2
- Life Satisfaction | Eurostat ID: ILC_PW01$DV_528

Analysis Questions:

- What is the association pattern between Countries' Satisfaction Scores and:
    - GDP per capita ?
    - Greenhouse Emission ?
    - Life Expectancy ?
    - Employment Rate ?
    - Weekly Working hours ?
    - Population ?

Abstract

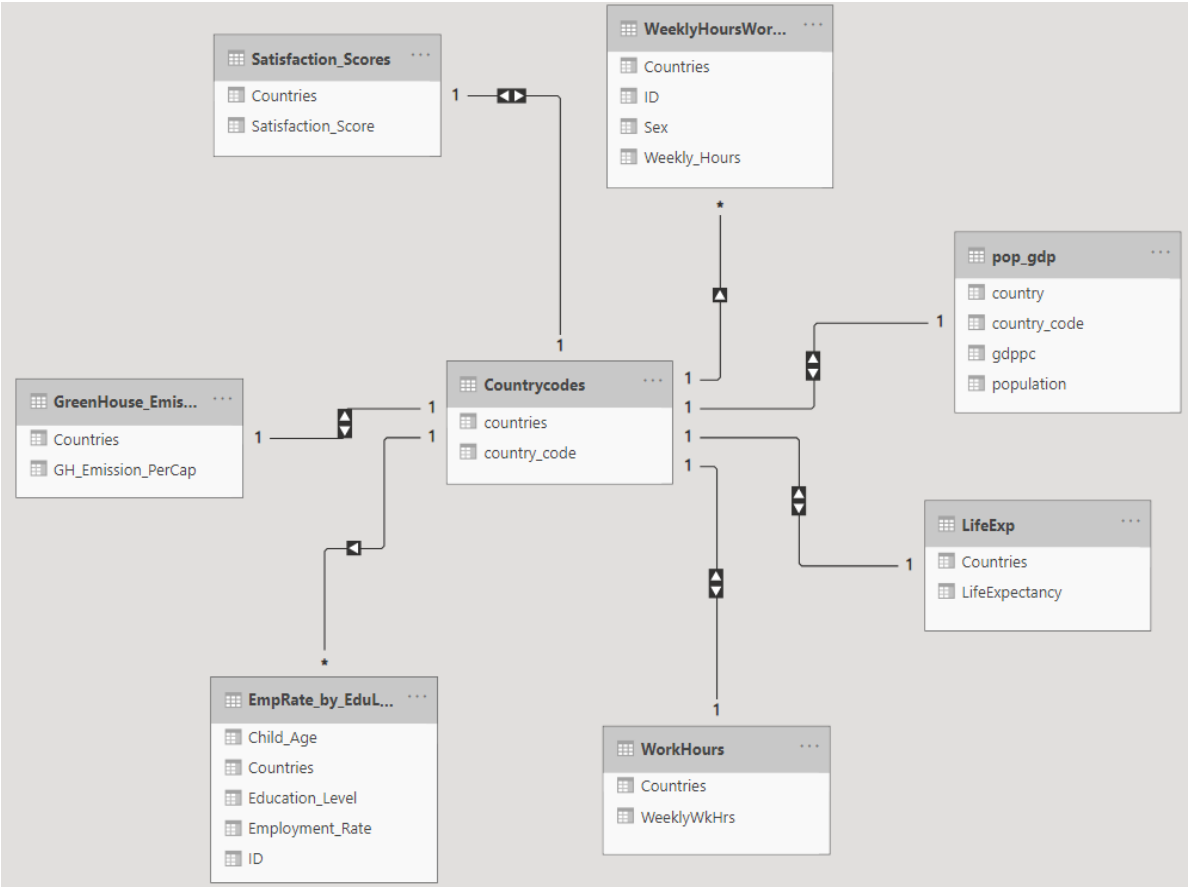
This report is written in fulfilling term-project requirements for the Data Engineering 2: Different Shapes of Data course of Central European University's MSc in Business Analytics program. Our task was to collect data on a topic of interest, create a KNIME-based workflow to engineer our dataset to be compatible for analysis, & prove the former via data visualizations, e.g. visually inspecting relationships among variables.

Our research question of interest is '*How can Life Satisfaction be associated with Greenhouse Emissions, GDP per capita, Employment & Life Expectancy in EU countries in 2018 ?*' As such, below we present an Extract-Transform-Load pipeline built primarily with KNIME, utilizing data extracted from MySQL, the Eurostat API & the World Banks' World Development Indicators platform.

We found that the most strongly correlating variables with Life Satisfaction, are Weekly working hours, GDP per capita & Life expectancy. After running multiple linear regression with the variables above, we may conclude with 90% confidence that there is a positive correlation between GDP per capita, Weekly working hours and Life Satisfaction while there is a negative correlation with Weekly working hours.

Data Collection

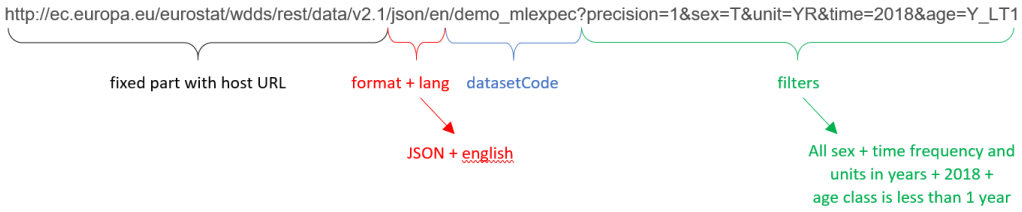
We collected data from 2 sources (Eurostat-, & World Development Indicators' (WDI) databases) & utilized 3 methods. We detail data pre-processing for these methods in the next section, however we would like to conceptualize how we see these data sources fitted in a single model. We prepared a diagram of table connections please see below.



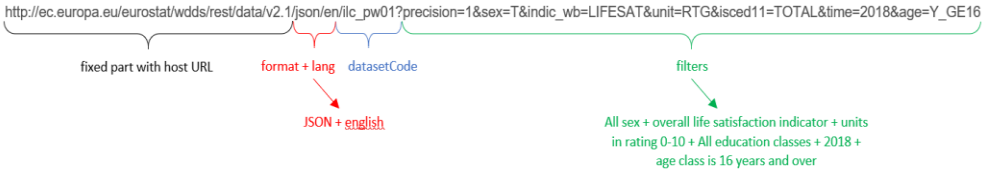
While planning, we wanted to ensure various data sources fit together well. Hence we stuck to country-level data for 2018, to obtain a cross-sectional dataset. We ensured a 1-to-1 match between data sources via country codes, to obtain our final data table in KNIME. To obtain *Country Data*, 1st EU *Country Codes* and EU countries' average *Satisfaction* scores were collected directly via the Eurostat API using Postman, 2nd EU countries' *Greenhouse Emissions*, *Life Expectancy*, *Average weekly working hours, Employment rates were* downloaded from Eurostat & processed into a MySQL relational database, & 3rd *Population & GDP* data were extracted from the WDI database, cleaned & written to comma separated values (CSV) format using RStudio.

Eurostat with Postman

To easily connect various datasets, we downloaded ISO 2 country codes from Eurostat via an API. The API was edited in Eurostat's query builder and is built according to the image below. It extracts country codes from the life expectancy table, containing all needed countries.



We also used another API to extract the average rate of life satisfaction



per country variable. We decided to use an API for this task as this is our main variable in the analysis part (our 'Y'), with which we want to associate other variables. Thus, we wanted to present the skills we acquired during the course on the main variable. Please see the request code for average rate of satisfaction below. As we downloaded JSON files through the APIs, we needed to transform them to data tables in KNIME, to extract valuable information and be able to join with other data sources.

WDI with RStudio

We obtained Population and GDP per capita values for all countries worldwide through the World Banks' World Development Indicators platform, using R scripts, using a library functioning as an API, the WDI package. Thus our initial data had 264 observations. Please see a snippet below. It contained several grouped observations such as World or South Asia. It contained observations related to other classifications instead of countries like "Heavily indebted poor countries" or "Low & middle income".

	Country_code	country	population Population, total	gdppc GDP per capita, PPP (constant 2017 international $)
1	AE	United Arab Emirates	9630959	66968.2699
2	AF	Afghanistan	37172386	2190.2403
3	AG	Antigua and Barbuda	96286	21115.7983
4	AL	Albania	2866376	13601.3034
5	AM	Armenia	2951776	12714.9582
6	AO	Angola	30809762	6933.5093
7	AR	Argentina	44494502	22745.9038
8	AT	Austria	8840521	55687.1893
9	AU	Australia	24982688	49575.9811
10	AZ	Azerbaijan	9939771	14209.6494

The cleaning process 1st excluded all observations NOT belonging to a country. 2nd, all remaining observations missing either population or GDP were dropped. Finally, we renamed variables and removed the year column. Our table thus held 187 observations. We included these steps node-by-node, please see the complete R script here.

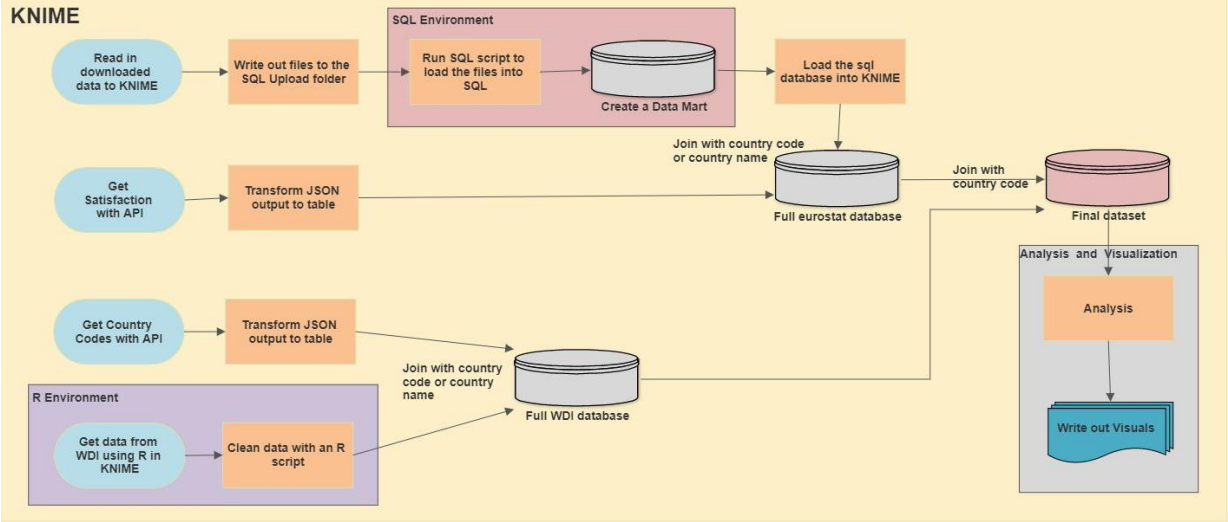
Eurostat with MySQL

In total 5 datasets were downloaded from Eurostat & loaded into MySQL. In all cases, only measurements for 2018 were considered. For each variable of interest, the complete dataset was downloaded in spreadsheet format & selected worksheets were saved into documents for ease of loading into an SQL schema (please see below selected sheet numbers with hyperlinks to respective databases).

- Life Expectancy: Sheet 1
- Average Weekly Work hours: Sheet 1,21,41
- Greenhouse Emissions (Expressed in units of CO2 equivalents): Sheet 2
- Average Satisfaction Rating: Sheet 1
- Employment Rate: Sheets 1-4

The Load_Data.sql script loads all text files & denormalizes the dataset to a single table. This is achieved with a series of (sub)-joins, concatenating each variable as a new column, aiming to keep 1 Country as 1 row. This denormalized data table also contained Average Weekly Work hours by Gender & Employment Rates by Education level, violating requirements for a normalized dataset. Hence, a Main Table was extracted as a View, in addition to an Average Work Hours by Gender; & Employment Rate by Education views.

Workflow in KNIME



Preparatory phase: Ensure MySQL & R interface works

Read text files to MySQL folder: To enable a reproducible workflow, we wanted to avoid readers having to copy sql data sources manually. Thus, the 1st step to be executed is the 'File Reader Part', which reads & saves all text files into C://ProgramData/MySQL/MySQL Server 8.0/Uploads folder path.

Set Username & Password in MySQL Connector node: To connect MySQL with KNIME, users must input their credentials into this node to access their MySQL user interface, as well as select the de2_eurostat database.

Install Rserve in R & set R folder path in KNIME: Finally, to execute R-based code snippets, readers must download the " KNIME Interactive R Statistics Integration" extension in File → Install KNIME Extensions → search for R integration & communicate their R root folders' path to KNIME (File > Preferences > KNIME > R > Browse to R folder (C:\Program Files\R\R-4.0.3)).

Load & Clean Data sources

Eurostat API: As the GET request is in JSON format, the 1st JSON path node extracts Country names & codes and satisfaction scores & codes into separate columns. The 2 subsequent JSON nodes pivot all these values, which are unpivoted by the 2 Unpivoting Nodes. Joiner nodes connect these via country codes, & variable are renamed.

MySQL Connector & Query Reader: Having set up these nodes in the pre-processing stage, the query reader simply selects all fields from the Main Table.

Import WDI data with R Codes: This section loads & cleans the WDI data table introduced in the previous sections, through a series of filtering steps. Once the data is loaded with the R Source (table) node, the 1st R snippet filters observations with numbers in their IDs & the 2nd R snippet filters EU, HK & OE IDs. These remove various aggregated entities from observations, just as the 3rd R snippet does any observation whose ID starts with X or Z, except XK, ZA, ZM & ZW. The last 2 R snippet nodes remove any observation with missing GDP / Population values, & rename variables to 'countrycode','country','population' & 'gdp'.

Joining Data Sources

1) **SQL + Satisfaction from Eurostat API:** Both data sources are from Eurostat, thus their country names & codes should be identical, but we encountered missing observations for inner joins, when using either variable individually. Joining with both columns & allowing either column to match, yielded the desired result however.
2) ***WDI & Country Codes from Eurostat API:*** As above, some countries' WDI names' or codes' did not match Eurostat's. In all cases however, at least 1 of the 2 were correct, therefore an inner join using 2 columns gave desired results.
3) ***1) & 2)***: As we used the same country codes during the previous 2 joins, we were able to use only one column (country_code) for a correct match, using a right join to keep all Eurostat data & keep only relevant ones from WDI.

Visualization & Analysis

Once we obtained the cleaned dataset we could do our analysis on it. First we had a look on all the scatterplots for all our numeric variables. From this we could have an overview if there seems to be a correlation between Satisfaction and another variable (this is visible in the first row and column). See this summary view here. We ran a multiple linear regression using Satisfaction as the outcome variable and the other variables as explanatory variables.

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
GreenHouse_Em_percap	0.0119	0.0442	0.2692	0.7904
LifeExpectancy	0.058	0.047	1.2339	0.2309
Weekly_AvgWkHrs	-0.1025	0.0481	-2.1303	0.0451
Employment_Rate	0.0086	0.0191	0.4506	0.6569
population	-1.75E-11	5.31E-9	-0.0033	0.9974
gdppc	7.35E-6	9.07E-6	0.8104	0.4268
Intercept	5.3586	4.4226	1.2116	0.2391

Multiple R-Squared: 0.5787
Adjusted R-Squared: 0.4584

Beside this, we also created separate scatter plots checking all the regressions of our variables and Satisfaction. We saved them and uploaded them to this folder. If you want to save them to your local computer, you will need to change the "Output location" for the "Image Writer" nodes to the place where you want these images. Regression results & visuals showed the 3 most influential variables affecting countries' average satisfaction level are weekly working hours, GDP per capita & life expectancy. We decided to focus only on the most important ones. As a last step, we created a model which regresses satisfaction on these 3 variables.

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
LifeExpectancy	0.0707	0.0375	1.8872	0.0695
Weekly_AvgWkHrs	-0.1102	0.0368	-2.995	0.0057
gdppc	1.02E-5	5.98E-6	1.704	0.0995
Intercept	5.2769	3.549	1.4868	0.1482

Multiple R-Squared: 0.6474
Adjusted R-Squared: 0.6096

We can conclude with 90% confidence, that there is a positive correlation between average life satisfaction, life expectancy & GDP per capita of European countries, and a negative correlation with average weekly working hours. The model above captured 65% of the variation in Satisfaction. We should take into consideration a simple linear model may not be the best representation of this problem. As the sake of our analysis was to find out which variables have a significant influence on satisfaction, we feel satisfied with these results.

Who Did What ?

Project Execution
- Viktória Mészáros: R Scripts, Visualization & Analysis
- Attila Serfőző: File Reader, Eurostat API, Country Codes, Joining Data Sources, File write out
- Brúnó Helmeczy: File Reader, SQL database, Joining Data Sources

Project Documentation
- Viktória Mészáros: WDI with RStudio, Visualization & Analysis
- Attila Serfőző: Eurostat with Postman, Workflow Visualization
- Brúnó Helmeczy: Abstract, Eurostat with MySQL, Workflow in KNIME text, report editing