**DE2 Report Text**

***Introduction***

This report is written in fulfilment of the term project requirements for the Data Engineering 2: Different Shapes of Data course at the 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, 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.

***Data Collection***

We collected data from 2 sources, the Eurostat-, & World Development Indicators’ (WDI) databases, & utilized 3 methods. We detail the data pre-processing for each of these 3 methods in the next section, however we 1st would like to conceptualize how we see these data sources fitted in a single model. We prepared a Mermaid Entity-Relationship diagram to do so:

***Conceptual Data Model – Mermaid EER***

Diagram

Description automatically generatedWhile planning for this assignment, we wanted to ensure various data sources fit together well. Hence we chose to stick to country-level data & only for 2018 to be able to obtain a cross-sectional dataset eventually. We ensured a 1-to-1 match between data sources via country codes, to be able to obtain our final data table in KNIME.

To obtain our *Country Data*, 1st EU countries’ average *Satisfaction* scores were collected directly through the Eurostat API with Postman, 2nd EU countries’ *Greenhouse Emissions*, *Life Expectancy*, *Average weekly working hours* 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***

***Eurostat with MySQL***

***WDI with RStudio***

***Workflow in KNIME***