**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. 1st EU countries’ average satisfaction scores were collected directly through the Eurostat API using Postman, 2nd EU countries’ greenhouse emissions, life expectancy, employment rates & average weekly working hours downloaded from Eurostat & processed into a relational database in MySQL, & 3rd Population & GDP data were extracted from the WDI database, cleaned & written to comma separated values (CSV) format using RStudio. Below we detail the data pre-processing for each of these 3 methods.

***Eurostat with Postman***

***Eurostat with MySQL***

***WDI with RStudio***