Group 2. Data warehouse

To implement our data warehouse, we investigated the available data from the three different sources. We identified that the WDI data indicators table had the most data, so we decided to drop all indicator values we were not going to use from our data warehouse and leave the indicators we saw useful to perceive corruption in a country. We also concluded that we would not use the measure for generosity data from the world happiness report.

During the cleaning of the data we discussed if we should clean unnecessary years and countries out of the data. We decided not to do this, as we didn't have a clear consensus what data we would be using and where and we could make any necessary cleaning with PowerBI later on before making visuals.

Through testing and considerations, we ended up working with star schema for our data warehouse. We had consensus that the task would be easier with simpler schema, and we also considered that PowerBI, which we wanted to use for our visualizations, worked best with a star schema.

Our process for this data warehouse is to investigate corruption, economics and happiness measures between different countries and years. The grain for our fact table is one year for each country, with measurement for all measures. Due to this, our fact table is a periodic snapshot fact table. In our data, there were many measures which we identified being key measures for our process. This can be seen in our data warehouse schema as larger fact table. Country and Date dimensions are identified, as the context for the measures in the fact table are from either of these sources. We decided to keep the date dimension as only yearly, as the measures were recorded each year. We also decided to include any need for a range of years to be implemented in our visuals and not have more values in the date dimension table (for example decade).

To clean the data, we separated each indicator we found useful and combined the information into our fact table. We created surrogate keys for country dimension and date dimensions and used these surrogate keys as the primary keys and as the foreign keys in the fact table. We also created identification keys for the fact table and used it as primary key.

The relationship between country dimensions and the fact table is one to many from country to fact table. The relationship between the date dimension and the fact table is also one to many.