SIGMA

Data Consulting

ETL Project - **Documentation**

Documentation

Extraction :

After carefully reviewing the data sources from Kaggle website, we choose the information of the average time expenditure by year on the OECD countries from 2015[1]

The idea was to relate time spent in different categories with other data sources that would enrich the information and give some context and value added to the comparison of how people spent the time in the aforementioned time frame.

We normalize the data to divide into tables the original sources and stream line consulting and eventual change on the sources.

The additional data sources to relate to the main information were the average scores on PISA tests[2] and the average yearly alcohol consumption[3].

Transformation:

The complete Data Transformation process was performed using Python code, with the use of “jupyter notebook”.

The main changes we did to the data frames were:

1. Time Spent DataFrame : We found unique values to the country and categories data and created new tables for each.
2. PISA scores: Data filtered to show only the most recent data (2015)
3. Alcohol consumption data: We averaged the yearly data by country

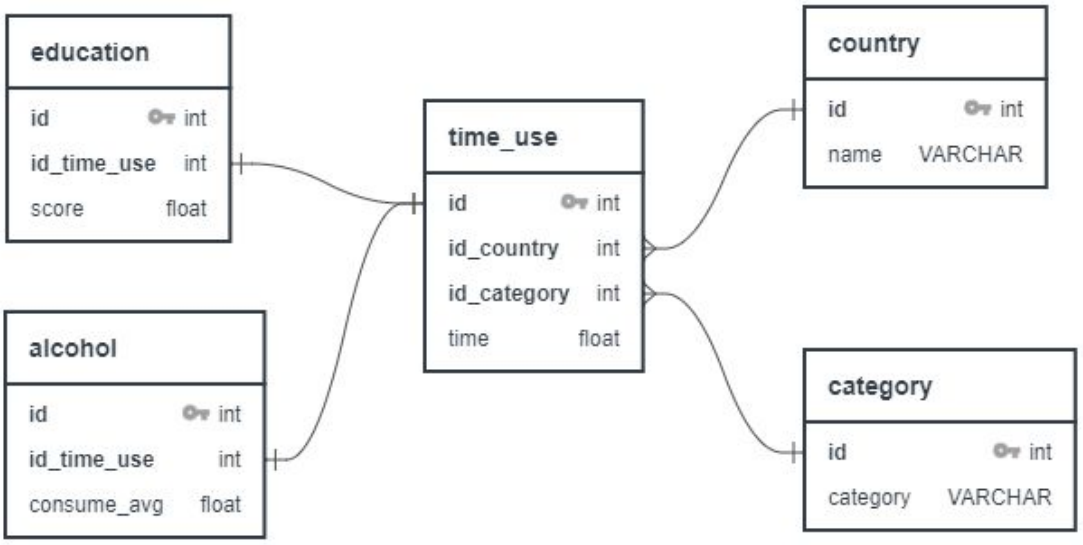
After transforming the original sources we created a new table with filtering time spent on categories Education, Eating and drinking, and afterwards created joins to add the Pisa Scores and Alcohol consumption data respectively

Loading:

After reviewing the process of incorporating the create tables using MongoDB, it was brought to our attention that the best way to do it was to create a data model using QuickDB Diagrams and SQL in order to incorporate the model in a normal and robust way, that could support the introduction on new data into the model structure. SQL database aligns with the type of connection we are intending to create between the various data inputs, and with this achieve the model outputs the team is intending. For the creation of the SQL structure, as a first approach the team decided to use the tool of “Entity Relation Diagram” in order to create the model structure and establish the relationship between the tables data, after this, the team recreated this structure into SQL code and it was executed on a pgAdmin query.
The Data retrieved from the “Data Transformation” section was exported to the SQL database using Python code.

ERD - Entity Relation Diagram

The following diagram shows the relationship between the existing tables using primary and foreign keys



Referenced Sites

Data Sources

1.Esteban Ortiz-Ospina, Charlie Giattino and Max Roser (2020) - "Time Use". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/time-use' [Online Resource]

2.Max Roser, Mohamed Nagdy and Esteban Ortiz-Ospina (2013) - "Quality of Education". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/quality-of-education' [Online Resource]

3.Hannah Ritchie (2018) - "Alcohol Consumption". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/alcohol-consumption' [Online Resource]