Annika Pavlovich
Carmela Monis
Dana Walker
Susan Xia
Thaddeus Gray
Vasu Manikarnika

ETL Project Report

For our project, we chose to transform four datasets comparing data on diet and food availability for 170 different countries globally. We found the following datasets:

1. Percentage of fat intake in overall diet per food category by country.
2. Percentage of kcal intake in overall diet per food category by country.
3. Percentage of protein intake in overall diet per food category by country.
4. Percentage of total food intake (kg) per food category by country.

Each dataset also included obesity rates, undernourished rates and total population for each country in the dataset, as well as Covid 19 rates related to confirmed cases, deaths, and recoveries for each country.

To begin our analysis, we first extracted the data from each of the four csv's and created tables in PostgreSQL. New data frames were then created using Pandas to be uniquely transformed by each member of our team. For food categories we chose to only look at:

- Animal Products: includes butter
- Cereals: barley, maize, rice, oats, etc.
- Eggs
- Fruits
- Meat
- Milk: excludes butter
- Fish and Seafood
- Vegetables

We first extracted CSV files (Fat_Supply_Quantity_Data, Food_Supply_kcal_Data, Food_Supply_Quantity_kg_Data, Protein_Supply_Quantity_Data) into DataFrames, created a filtered DataFrame from specific columns (country & food categories) for each table, and renamed them.

We then combined all the data frames together by using the pd.concat function so that a deeper analysis comparing the four datasets side-by-side could be further explored.

The final tables were loaded into PostgreSQL, where we ran a number of different queries based on our own unique tables. Dana's queries selected data based on the "type" column in the covid_19_diet table, where fat_supply_quantity=1 , food_supply_quantity=2, food_supply_kcal=3 and protein_supply_quantity=4. It also selected data from the covid-19-diet table based on country.

We also transformed the data by first dropping columns with irrelevant information and then renaming the food category columns in each dataset to differentiate them from each other (fat, kcal, protein, quantity). These were then merged into one dataframe, joining on Country, Population, Obesity Rate, Undernourished Rate, and the Covid 19 rates, ensuring a single row for each country's relevant data.

These were then loaded in PostgreSQL as one merged table and 4 smaller tables specific to fat, kcal, protein, and quantity. Queries could be done to look at potential relationships between COVID 19 deaths, recoveries, and cases, and the types of diets being eaten in each country.

Through each of our table transformations and unique queries, a further analysis could be completed by grouping data by country or food category to assess where relationships may exist. For example, we were able to query the food supply quantity, food supply kcal, fat supply quantity, and protein supply quantity of just the vegetable category for comparison.