**ETL Project – Group 13**

**How Food Supply affects Obestity Rates by Country for the Years 1975-2016**

**Extract**

The first dataset we selected was Obesity among adults by country, 1975 – 2016 from kaggle.

The second dataset we used we pulled from the Food and Agriculture Organization of the United Nations (FAOTA) at www.fao.org. The FAOTA website allowed us to select certain fields to download the required data from four categories of data. The categories are countries, items (types of animal based food product), food supply quantity and years. We selected our data as follows:

* Countries – all
* Items
  + Animal Fats
  + Fish, Seafood
  + Meat
  + Milk – Excluding Butter
* Food Supply Quantity – tonnes
* Years – 1975 – 2016

We chose to narrow our groupings of items to those reflected above because they included majority of animal based food products available in different countries. We included meat, milk and animal fats which are often linked to obesity and seafood as a food type that is often associated with healthy eating and lower rates of obesity.

We selected the years 1975 – 2016 to match the years of data available in the dataset on obesity obtained from kaggle.

**Transform**

We loaded both datasets into pandas and narrowed down the database information to only include data relevant to our project.

**FAOTA Data**

The data obtained from FAOTA included several columns of data that were not required for our analysis. We narrowed down this dataset to include the following columns:

* Country
* Year
* Item (food type)
* Value (tonnes)

We dropped fields for Domain Code, Domain, Country Code, Element Code, Element, Year Code, Unit, Flag and Flag Description.

Our data had two datapoints that were common to both datasets, country and year. We elected to use SQL because the data could be related on common datapoints. We chose the country as our primary key for our database schema.

Once the data from FAOTA website was cleaned we aggregated the data into a table that with columns for the country, year, and the value for each item type. The steps we used to complete this process was to sort the original dataframe by country.

We used the function pd.get\_dummies to convert the Item column to a 0 or 1 depending on whether data exists in the dataframe for that item. Next we added columns to the dataframe for each of the item types: Animal fats, Fish, Seafood, Milk – Excluding Butter and Meat and multiplied the Value column by the new amount in Item (either 0 or 1) to reflect the tonnes by item for each country and year.

Next we used the pd.concat function with the axis=1 in order to create a dataframe with the Country, Item, Year, Value, Animal fats, Fish, Seafood, Meat and Milk – Excluding Butter. We used groupby to group the results by country and year and reset the index to get our final dataframe for our FAOTA data. The resulting table showed the country, year, and the value for each of the food items.

**Obesity Data**

We used a similar approach to address the obesity data and get it into a format that would work in our database. Our first step was to convert the obesity % field in our database to a numeric value using pd.to\_numeric. Next We used the .get\_dummies function to determine which rows had values for the fields Both sexes, Female and Male. We multiplied the result for each row by the obesity % and used the pd.concat function to create a dataframe with columns for Country, Year, Obesity (%), Sex, Obesity Ratio, Both sexes, Female and Male.

We used groupby by Country and Year and applied the sum function to the numeric fields. The result was a dataframe with the Country, Year, Obesity Ratio, and the Obesity % for Both sexes, Female and Male.

Finally we created a merged dataframe with both tables using pd.merge and dropping duplicate country and year columns. The resulting dataframe represents the data by Country, year and tonnes of each item (Animal fats, Fish, Seafood, Meat, Milk-Excluding Butter), Obesity Ratio and Obesity % for each sex (Both sexes, Female and Male).

**Load**

We created the following schema using two tables. One for the obesity dataset and one for the FAOTA dataset.

|  |  |
| --- | --- |
| FAOTA Table | |
| Country - PK | str |
| Year | str |
| Animal fats | float |
| Fish, Seafood | float |
| Meat | float |
| Milk – Excluding Butter | float |

|  |  |
| --- | --- |
| Obesity Table | |
| Country - PK | str |
| Year | str |
| Obesity Ratio | float |
| Both sexes | float |
| Female | float |
| Male | float |