ETL Report – Aishwarya Damodharan & Edward Joseph

Extraction: Data sets of interest: Food based datasets that may show actionable insights into consumer interest, market food trends, and general nutrition on an international scale.

Source 1 - FAO.csv Link - <https://www.kaggle.com/fauzantaufik/fao.csv> A CSV file containing over 50 years of food/feed purchase data from multiple countries. Data includes food identifiers, countries, and amounts purchased in tonnes in a given year. The included columns are: Data Source, Country Name, Product name, Element name, units of production, year of production, and volume of production.

Source 2 - en.openfoodfacts.org.products.tsv Link - <https://www.kaggle.com/openfoodfacts/world-food-facts> A TSV file currently containing information on over 700,000 different food products logged by consumers world-wide. The organization providing the data is self-described as a food products database made by everyone, for everyone. The data was collected from buyers around the world adding data from previously purchased products. This data set has over 700,000 individual records and over 100 columns with varied product descriptions such as nutrition, brand, and manufacturing country.

Transform: The first step after retrieving the data was to read the csv files into a jupyter notebook. This was done using the pandas read\_csv function then previewed with a .head(). At first look the FAO data was ordered well. But lacked any obvious columns to merge databases on. To fix this issue a standard categorical bin was created based on the FAO.Item and OpenFoodFacts.main\_ingredient columns. This bin used a more centralized base food groups and then brute forced set of key-value pairs to create a functional column for table merging. During this phase of the project a categorical standard was developed and integrated into each data set. Once the data was sufficiently cleaned and standardized the tables were merged, then cleaned again to ensure data quality.

For this project, Sql was chosen for its speed and simplicity when dealing with large datasets, and because there was little need to scale up space on this database.

Reproducible ETL steps:

1. Choose data to extract.
2. Read in the data and place it into a dataframe.
3. Group the Data Frame by most relevant data.
4. If there is no obvious choice for a data table merge. Create a normalized category (bins) for joins, merges and comparisons.
5. Merge the normalized category into the data tables.
6. Repeat steps 1-5 for every imported data set.
7. Clean the data: a. Remove duplicate data. b. Drop incomplete data records. c. Standardize data types. (Language, date/time, encoding, etc.) d. Remove any files with a count beneath your threshold of interest to lessen data static.
8. Create new data frame(s) with the desired columns in the desired order.
9. Compare new data frame(s) to original to compare category lists.
10. Merge tables and set normalized categories as primary keys.
11. Remove any missing values (again), organize data into logical order.
12. Save the final dataframe to a csv (or prefered file option).
13. Load the file into a relational SQL database for future business or analytical use.
14. Visualize data and present as needed.