ETL Project Report

Healthy Food Nutrition Values

By Saurabh Goyal & Julia Sebring

# Extract

The following source data from was extracted from csv’s, which were downloaded from data.world, and read into a Jupyter Notebook file as pandas dataframes:

* <https://data.world/craigkelly/usda-national-nutrient-db> (Source 1)
* <https://data.world/awram/food-nutritional-values> (Source 2)

## Tables:

Nutritional\_info: Food nutritional measurement values per 100 grams

Food\_groups: Food nutritional measurement values are per 100 grams

# Data Cleanup & Analysis

## Transform:

Post loading the data into different dataframe, unnecessary columns or columns with NaN values were discarded. Few columns were renamed to ensure no discrepancy later point of merging or joining tables. Special characters and capitalization were also considered while filtering and renaming the fields.

4 different dataframes were created from the 2 main dataframes for further data analysis

1. basic\_nutirent
2. Food\_group
3. vitamins
4. minerals

## Joining:

The tables were joined based on food IDs

# Final Tables:

Destination: Postgres MySQL (relational database)

The tables are organized for the end user to explore foods that meet recommended daily value targets for different aspects of nutrition, which have varying ranges. These tables were queried to get recommended daily values for different nutritional needs.

Tables were defined and created in Postgres SQL. Data was loaded into their respective tables from each data frame defined above

Table:

1. Food Groups

Columns:

* 1. ID
  2. Food group
  3. short description

1. Basic Nutrients

Columns:

* 1. ID
  2. Short description
  3. Calories
  4. Protein
  5. Fat
  6. Carb
  7. Sugar

1. Vitamins

Columns

* 1. ID
  2. Short description
  3. Vitamin A
  4. Vitamin B6
  5. Vitamin B12
  6. Vitamin C
  7. Vitamin D
  8. Vitamin E

1. Minerals

Columns

* 1. ID
  2. Short description
  3. Calcium
  4. Phosphorus
  5. Potassium
  6. Sodium
  7. Magnesium

# Querying

We ran multiple queries to identify recommended food with the description and its nutritional value from various tables

We queried to identify nutrition values for Vitamin C, Calcium, Protein in our Jupyter notebook and also merged all the different tables into one for making it as a complete dataset