STAT 850 Final Project
*** OUR PROJECT TITLE ***

Yutong Liu, Lin Zhao
yutong.liu@huskers.unl.edu, lzhao12@huskers.unl.edu

***Abstract*—** This is our abstract

1 INTRODUCTION

Introduction goes here

2 DATASET DESCRIPTION

FoodData is an integrated database that provides food component and nutrient information. There are several tables in the original database, we only take three of them and merge the variables together by **fdc_id**, which is a unique permanent identifier of a food across tables. We would like to discover potential recycling of non-edible food component which has relatively higher nutrient.

Three tables from the database are chosen for this project:

· food_component.csv
· food_nutrient.csv
· nutrient.csv

By combining three datasets together and selecting several variables, a new dataset called **food_dataset** is generated for further analysis.

Description of variables:

component_name - The kind of component, e.g. bone

pct_weight - The weight of the component as a percentage of the total weight of the food

is_refuse - Whether the component is refuse, i.e. not edible

gram_weight - The weight of the component in grams

nutrient_name - Name of the nutrient

nutrient_amount - Amount of the nutrient per 100g of food. Specified in unit defined in the nutrient table.

min - The minimum amount

max - The maximum amount

median - The median amount

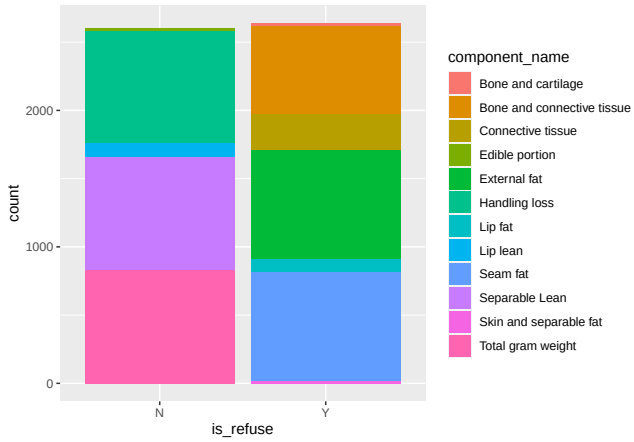
nutrient_unit - The standard unit of measure for the nutrient (per 100g of food)

nutrition_amount_in_component - The true value of nutrition amount in component

3 RESULTS

3.1 Refuse food component

The variable *is_refuse* in original dataset has two levels: Y (is refuse) and N (not refuse). Therefore, we can assign all components into 2 groups: one is refuse, another is not refuse. Since our ultimate goal is to recycle the refuse components and explore the potential value of them, we plot the data by variable *is_refuse* to see what kind of component is refuse.

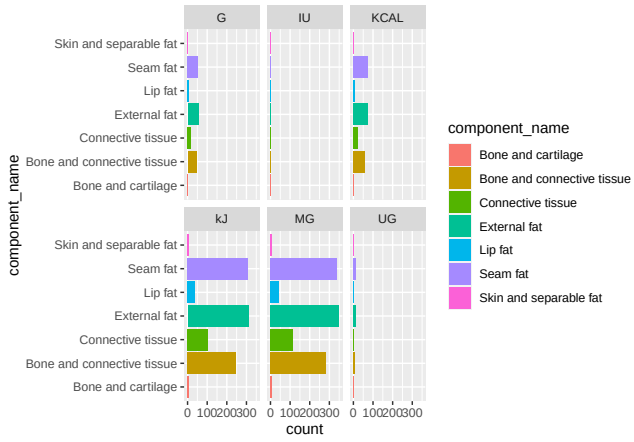


The bar chart above shows the detailed information of whether a component is refuse or not. The left bar is not refuse, contains components such as *separable lean*, *total gram weight*, *handling loss*, *edible portion*, *lip lean*. On the right-side bar,the refuse components are mainly *bone and connective tissue*, *connective tissue*, *external fat*, *seam fat*, *bone and cartilage*, *lip fat*, *skin and separable fat*.

It is easy to see that all of tissues and fats are refuses. Fat, tissue and cartilage are components we can recycle and extract nutrition from. We will only work on refuse components from this point.

3.2 Adjust units of nutrition amount

In order to run statistical analysis on the nutrition amount, it is necessary to have uniform units for all components. First of all, we would like to know how many units exist in dataset and which one is the most commonly used.



We can see from the plot above that most common unit used for nutrition amount in dataset is MG and KJ. We will adjust the mass unit to G, the energy unit to KJ. After adjustment, there exists mass unit *G*, energy unit *KJ* and vitamin unit *IU*.

Mass unit adjustment: 1 g = 1000 mg = 1000000 ug

Energy unit adjustment: 1 KCAL = 4.184 KJ