The Datasets represents a record of 420M food items purchased by 1.6M Clubcard owners. This is for the 411 Tesco stored in the Greater London Area over the course of the year 2015. Tesco on itself is the 9th largest grossing retailer in the world with 81B in global revenue, and is the biggest grocery retailer in the UK.

The datasets are split into:

* Food\_categories.csv
* Area Levels which is split into:
  + Lower Layer Super Output Area (LSOA)
  + Middle Layer Super Output Area (MSOA)
  + Ward
  + Borough

Which are in the order of smallest to largest in a geographical sense.

These datasets are listed as {aggregation}\_grocery.csv files that contain:

* **Area\_id**: identifier of area
* **weight**. Weight of the average food product, in grams (Eq. ([3](https://www.nature.com/articles/s41597-020-0397-7#Equ3)))
  + A mathematical equation with numbers and symbols

    Description automatically generated
* **volume**. Volume of the average drink product, in liters (Eq. ([4](https://www.nature.com/articles/s41597-020-0397-7#Equ4)))
  + A mathematical equation with numbers and symbols

    Description automatically generated
* **energy**. Nutritional energy of the average product, in kcals (Eq. ([5](https://www.nature.com/articles/s41597-020-0397-7#Equ5)))
  + A mathematical equation with black text

    Description automatically generated
* **energy\_density**. Concentration of calories in the area’s average product, in kcals/gram (Eq. ([6](https://www.nature.com/articles/s41597-020-0397-7#Equ6)))
  + A mathematical equation with black text

    Description automatically generated
* **{nutrient}**. Weight of {nutrient} in the average product, in grams (Eq. ([7](https://www.nature.com/articles/s41597-020-0397-7#Equ7))). Possible nutrients are: carbs, sugar, fat, saturated fat, protein, fibre. The count of carbs include sugars and the count of fats includes saturated fats.
  + A black text on a white background

    Description automatically generated
* **energy\_{nutrient}**. Amount of energy from {nutrient} in the average product, in kcals (Eq. ([8](https://www.nature.com/articles/s41597-020-0397-7#Equ8)))
  + A math equations and formulas

    Description automatically generated with medium confidence
* **h\_nutrients\_weight**. Entropy of nutrients weight (Eq. ([11](https://www.nature.com/articles/s41597-020-0397-7#Equ11)))
  + A black and white math equation

    Description automatically generated
* **h\_nutrients\_weight\_norm**. Entropy of nutrients weight, normalized in [0,1]. (Normalised eqn 11)
* **h\_nutrients\_calories**. Entropy of energy from nutrients (Eq. ([12](https://www.nature.com/articles/s41597-020-0397-7#Equ12)))
  + A close-up of a word

    Description automatically generated
* **h\_nutrients\_calories\_norm**. Entropy of energy from nutrients, normalized in [0,1]. (normalized eqn 12)
* **f\_{category}**. Fraction of products of type {product\_type} purchased (Eq. ([13](https://www.nature.com/articles/s41597-020-0397-7#Equ13))).
  + A mathematical equation with numbers and symbols

    Description automatically generated
* **f\_{category}\_weight**. Fraction of total product weight given by products of type {category} (Eq. ([15](https://www.nature.com/articles/s41597-020-0397-7#Equ15))).
  + A white background with black text

    Description automatically generated
* **h\_category**. Entropy of food product types (Eq. ([14](https://www.nature.com/articles/s41597-020-0397-7#Equ14))).
  + A black and white math equation

    Description automatically generated
* **h\_category\_norm**. Entropy of food product categories, normalized in [0,1]. (normalized eqn 14)
* **h\_category\_weight**. Entropy of weight of food product categories (Eq. ([16](https://www.nature.com/articles/s41597-020-0397-7#Equ16))).
  + A close-up of a math equation

    Description automatically generated
* **h\_category\_weight\_norm**. Entropy of weight of food product categories, normalized in [0,1]. (normalized eqn 16)
* **representativeness\_norm**. The ratio between the number of unique customers in the area and the number of residents as measured by the census; values are min-max normalized in [0,1] across all areas (Eq. ([2](https://www.nature.com/articles/s41597-020-0397-7#Equ2))).
  + A black text on a white background

    Description automatically generated
* **transaction\_days**. Number of unique dates in which at least one purchase has been made by one of the residents in the area.
* **num\_transactions**. Total number of products purchased by Clubcard owners who are resident in the area.
* **man\_day**. Cumulative number of man-days of purchase (number of distinct days a customer has purchased something, summed all individual customers).
* **population**. Total population of residents in the area according to the 2015 census.
* **male**. Total male population in the area.
* **female**. Total female population in the area.
* **age\_0\_17**. Total number of residents between 0 and 17 years old.
* **age\_18\_64**. Total number of residents between 18 and 64 years old.
* **age\_65 +**. Total number of residents aged 65 years or more.
* **avg\_age**. Average age of residents according to the 2015 census.
* **area\_sq\_km**. Surface of the area (km2).
* **people\_per\_sq\_km**. Population density per km2.
* Where applicable, measures are accompanied with their standard deviation (fields with suffix **\_std**), the 95% confidence interval for the mean (suffix **\_ci95**), and the values of the 2.5*th*, 25*th*, 50*th*, 75*th*, and 97.5*th* percentiles (suffix **\_perc{value}**).
* Validation Data set:
  + **child\_obesity\_london\_borough\_2015-2016.csv**
  + **child\_obesity\_london\_ward\_2013-2014.csv**
  + **diabetes\_estimates\_osward\_2016.csv**
  + **london\_obesity\_borough\_2012.csv**
  + **london\_pcd2geo\_2015.csv**
  + **obesity\_hospitalization\_borough\_2016.csv**
    - Validation of health outcomes
    - We assess the ecological validity of the data by comparing the grocery purchases with metabolic syndrome conditions that are strongly linked to food consumption habits. Specifically, we consider data about the prevalence of obesity and type-2 diabetes.
    - **Prevalence of overweight and obese children**. The fractions of overweight and obese primary school children in Reception class (aged 4 to 5) and year 6 (aged 10 to 11), sampled across wards. This data has been collected by the English National Health Service (NHS) in the 2013–2014 school year[46](https://www.nature.com/articles/s41597-020-0397-7#ref-CR46).
    - **Prevalence of overweight and obese adults**. The fractions of overweight and obese individuals among a statistical sample of borough residents. This data has been collected by the Active People Survey (APS) in 2012[47](https://www.nature.com/articles/s41597-020-0397-7#ref-CR47).
    - **Diabetes prevalence**. The fraction of adults among those registered at a GP practice in England who are affected by type-2 diabetes[48](https://www.nature.com/articles/s41597-020-0397-7#ref-CR48). This data has been collected by the NHS for year 2015 at ward level.
* The pcd2geo….. datasets tells us what postcode belongs to which data entry in the Arealevel datasets

We can use the age data to calculate the overall correlation between the number of people a certain age with the weight of the products bought and volume of drinks.

We can also use it to check the correlation between the age groups and the Kcals.

Also consumption of fats, sugar, salts, saturated fats, protein, carbs etc.. and how that has affected the obesity in children in lndn as we alr have that data for boroughs and oswards. But essentially focus on the one with the same year.

Task 1: Describe the dataset **12%**

Task 2: Vasualise Dataset, we take 2 useful insights from the dataset **22%**

Task 3: Get income dataset, use that to reference the trends with the dataset for example how it affects the trends in buying **28%**

Task 4: Presentation **12%**

Task 5: Summary **16%**

**Gonna use wieight and volume for income data, compare with certain area level**

**Fats, saturated, sugar, and energies of each for obesity and child obesity**

**Compare with protein, fibre and carb maybe**

**Maybe do, correlation between intake of fats, ect… with obesity**

**NORMALISE DATA?????**

**CLEAN CODE**

**Task 2:**

* **Start by calculating the % of 0-17 in osward**
* **Use that to calc estimate for intake of fats, saturated and sugar**
* **Use that for correlation with osward child obesity**

Task 3:

* Use income data to plot corr of income