Tesco Grocery 1.0 Dataset vs. Educational Attainment

CM50266-Applied Data Science Coursework 2024-2025

# Task 1: Describe the dataset

In this task you need to describe the dataset. It contains many fields. A good description will find ways to group fields together to summarise the content rather than simply list all the fields. You need to explain the value of the data rather than only provide basic description. You should also address any assumptions or limitations that need to be considered which may affect how the data can or should be used.

Introduction:

In this project I will analyse the Tesco 1.0 Dataset. This dataset contains information about in-store purchases made by clubcard holders in the 411 shops located in London. The data in this dataset allows us to obtain food purchasing patterns in different geographical areas throughout the year and also enables us to associate these patterns with cultural, economic, health-related aspects, or, as we will see later in this presentation, with educational factors.

Summary of the data:

The tesco Dataset contains 202 fields. We can separate them into four different groups:

1. Census Statistics (10 fields in total): contain information about the geographical area and its population in 2015. For example, the area ID, the average age of its habitants, its surface area or the density of residents.
2. Market presence (4 fields): These variables express Tesco’s reach in an area, fields included in this category are the normalized ratio between customers and residents, or representativeness, purchase dates, number of transactions and number of days a transaction was made.
3. Nutrient information (30 fields/156 fields): This category Covers the nutritional properties of the typical product and of the nutrients it contains: such as the weight, volume or energy-density. Added to these fields are their respective standard deviation, 95% confidence interval and some precentiles. As the final fields in this category we have the entropy of the nutrients weight and calories, which give information about the diversity of the nutrients in the typical product.
4. Product category information (32 fields): Food items sold at Tesco can be assigned to 17 different categories. Listed in these fields are the probability distribution of items belonging to the categories, the relative weight of each category, and the entropy of the distribution and the relative weight.

Limitations, biases and assumptions:

With this dataset we can’t estimate the diet of individuals as we don’t know wether the clubcard is used by only one persone or a whole family, we can only get information about the average product purchased in certain area. Furthermore, this dataset does not take into account meals served at restaurants or products shopped online, so it can be used to study trends in food consumption habits rather than daily food intake.

In the data we are only taking into account people that shop at Tesco and have a clubcard, so our sample may not be representative of all the population. In addition to this the data may also be biased because there are more tesco stores in some areas than other. Having this acknowledged, with Tesco being one of the biggest grocery retailers in the UK we can assume that the data can be used to reflect consumption patterns of the population of Boroughs. In order to tackle the geographic bias we can filter out the boroughs with less representativeness norm.

Insights:

The first insight we can extract from the data is that there are some correlations between the average age of the population of a borough and the weight of alcohol, fibre and entropy. This knowledge can be used to conduct more focused advertisement campaings in boroughs with older population to encourage the purchase of alcoholic products, or those rich in fiber whereas in boroughs with more young people fibre or other nutrient.

Secondly, if we check the monthly product category distribution, we can see that in summer more beer is sold whereas in winter more wine is sold across all the boroughs. With this information we can make campaings or special offers of this products in certain seasons in order to try and boost sales.

Dataset(<https://www.compare-school-performance.service.gov.uk/download-data>):

https://www.compare-school-performance.service.gov.uk/download-data?currentstep=datatypes&regiontype=all&la=0&downloadYear=2014-2015&datatypes=ks4underlying

This dataset contains data about educational attainment in key stage 4 at national level. It contains the grades and number of entries for each subject and GCSE qualification, or equivalent, in every school of the UK, whether it is independent or state funded, and the borough in which it is located. If the number of students that entered the qualification was five or fewer the data is suppressed, in this case we will just filter this subjects out. By filtering this data by borough we will be able to relate the data obtained in the Tesco Dataset to the educational attainment in that area. Some assumptions we will be making: GCSE data is enough to represent academic attainment at borough level. We will also be assuming that students study at the borough the school is in, in order to be able to successfully connect the data from this dataset to the one in the Tesco dataset. We will also be assuming that the data from the academic course 2014-2015 will be enough to relate to the 2015 shopping habits. As it doesn’t change abruptly each year.

Analysis of dataset and relation to literature:

<https://sci-hub.lu/10.1080/08870440802460426>

<https://sci-hub.lu/10.1017/S136898000800178X>