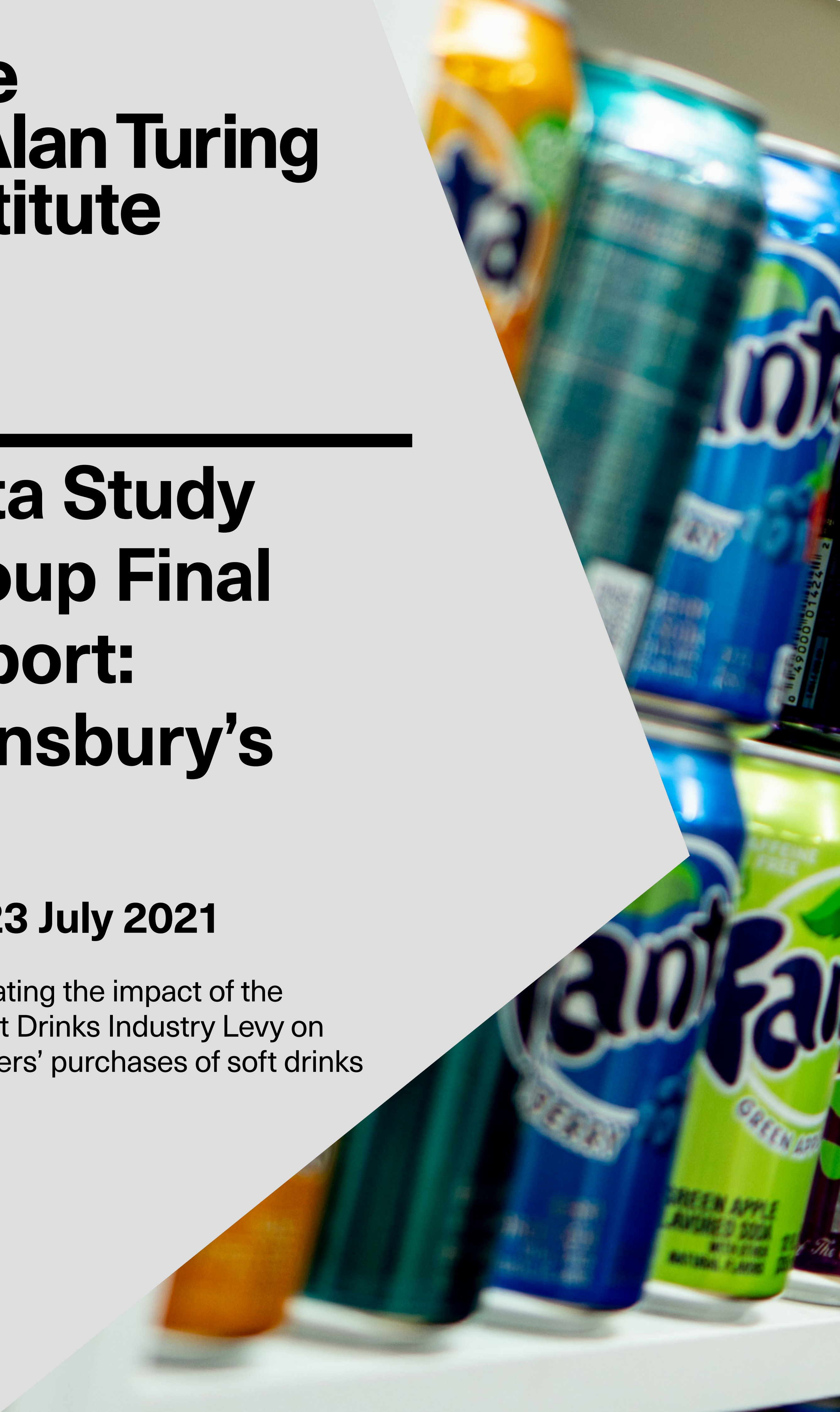
The Alan Turing Institute

Data Study Group Final Report: Sainsbury’s

12 – 23 July 2021

Investigating the impact of the UK’s Soft Drinks Industry Levy on consumers’ purchases of soft drinks



https://doi.org/10.5281/zenodo.6498023

Contents

1. Executive summary

1.1 Challenge overview

Evidence from consumer panel data has already shown that the UK Soft Drinks Industry Levy (SDIL) has been effective at the population level in reducing the volume of soft drinks purchased, and the amount of sugar they contribute to the diet. However, less is understood about how the levy impacted individual customers.

Using customer-level transaction data from Sainsbury's, the challenge was to explore how different types of customers responded to the levy. This work contributes to our understanding of whether fiscal policy is an effective and equitable approach to dietary improvement. Thus, the approaches and findings in this report are expected to be of interest to policy-makers. Understanding customer behaviours will also help Sainsbury's to better understand their customers' propensity to change in response to different market interventions. This will enable Sainsbury's to tailor their offering, ensuring nobody is left behind in the goal to provide healthy sustainable diets for all.

This report describes the analysis undertaken during a 2-week online Data Study Group (DSG) in partnership with the Turing Institute and the Leeds Institute for Data Analytics. A team of 9 data scientists participated in the DSG. Biographies for each participant, and the challenge leaders can be found in section 14 at the end of this report.

1.2 Data overview

The challenge used transactions collected over 3.5 years (January 2016 – end of June 2019), starting just before the levy was announced, until 1-year post-levy implementation. The data follows a cohort of ~300K loyalty card holders who purchased regularly with Sainsbury's during 2016. This sample provides a cohort for tracking responses over time. Data is aggregated at the monthly level and summarises transactions made by each customer for non-alcoholic beverage products, including the amount purchased and calorie and sugar contents of the drink. Finally, customer demographic data from the loyalty card database, and area-level census data are available.

1.3 Main objectives

The challenge aimed to:

1. Categorise customers based on their pre-levy purchase behaviours and understand whether these determined response to the levy.

2. Characterise types of response to the levy

3. Understand whether response to the levy was determined by demographic characteristics of the customer or the area in which they live.

1.4 Approach

Behavioural clusters were generated using the K-means clustering algorithm to characterise pre-levy purchase behaviours. By re-assigning customers to these clusters after the announcement and implementation of the levy, we could quantify how many customers changed their behaviours and describe the nature of the change. A time-series clustering approach (using the K-shape algorithm) was used to cluster customers based on the shape of their overall purchase trend. Thus clusters represent customers who moved in the same direction, rather than those who displayed similar baseline behaviours.

By using unsupervised machine learning algorithms to classify customers based on behavioural parameters only, rather than their demographic characteristics, we aimed to generate behaviourally homogeneous clusters. This was based on the theory that, as behaviours are 'sticky', pre-levy behaviours are likely to influence post-levy behaviours. Demographic variables were then applied post-hoc to explore if any distinct characteristics defined clusters and may predict cluster assignment.

Simple multiple linear regression was used to explore which factors influenced the sugar content of chosen drinks. While multivariate time-series regression was used to understand whether responsiveness to price, reformulation, and volume were demographically and/or geographically determined. Finally, a case study for customers of a single brand of sugary drink used a combination of these methods to explore how behaviours applied to a single product.

1.5 Main considerations

Previous research has revealed socio-demographic determinants of diet, with poorer diets apparent in younger people and those living in more socially deprived circumstances, in general. Through our regression analyses we observed small preferences for higher sugar beverages among younger adults and males, and smaller responsiveness to price increases among people in deprived communities, suggesting they are more resistant to change. However, these demographic traits did not translate to our clustering approaches We did not observe any statistically significant differences between our behaviour or time-series clusters. It is possible that this may be a result of the variables used to generate the clusters or the number of clusters.

The UK Government appears more focused than ever on intervening to tackle poor diet-related health. With new legislation coming in October 2022 to restrict in-store promotions for 'less healthy' products [1], and a recommendation for an added sugar and salt tax from the National Food Strategy report 'The Plan'[2], it is important to understand how population-level interventions affect not just populations, but people. This work proposes methodologies to explore customer-level differences in response to fiscal policy and offers some preliminary findings which could carve a path for future research.

1.6 Limitations

We are limited in our ability to determine if products fall within the SDIL, as 'added sugars' content is unavailable. 'Total sugars', in combination with product sub-category, were used to assign the SDIL status of products. Despite manual checking, it is probable that errors persisted. Customer attrition is also a limitation; we cannot say if reduced purchase volumes are the result of a true change in behaviour, or a switch to shopping elsewhere. Furthermore, without food purchases, we cannot comment on overall dietary quality. Large data volumes and computational capacity also limited our ability to run some analyses.

1.7 Recommendations and further work

Clustering approaches have provided useful insights into different patterns of purchasing and response to the levy. Further exploration of the possible demographic determinants of behavioural patterns is warranted, given that no clear distinction between clusters was observed. Future research could survey customers to understand more about their household composition, employment status, education level and income, for example. We also recommend incorporating category, brand and product information to unpick behaviours between clusters, as it is possible that not controlling for these factors may mask some nuance which accounts for socio-demographic determinants of purchase behaviours.

In the future, it may be possible to develop a model to predict future purchase behaviours and compare these with more recent transaction data. This aligns with the stretch goal to better understand propensity to change and develop tailored approaches to ensure healthier diets are accessible for everyone.

Finally, using learnings from this work as a starting point, it would be interesting to model the impacts of replacing the SDIL with the National Food Strategy's proposed added sugar tax, among different groups of people. This would likely require a sector-wide approach as a single retailer does not cover all purchases.

2. Introduction

Obesity, defined by the World Health Organization as "an abnormal or excessive fat accumulation"[3], is a major worldwide health issue of the twenty first century. In the United Kingdom, one out of four adults and one out five children at the end of the primary school are living with obesity.[4] Obesity is a risk-factor for non-communicable diseases such as Type 2 diabetes mellitus, cardiovascular diseases and some cancers, and also psychological health issues such as a low self-esteem and depression.[5] Although obesity is caused by a combination of several factors (our genes, our physiological and psychological state, and our environment) that are difficult to modify, it is possible to improve our behaviours increasing our physical activity and improving our dietary habits.[6] Indeed, obesity is due to unbalanced energy (i.e. energy intakes exceed energy expenditures).[7]

Currently, except for in the most severe cases where bariatric surgery may be recommended,[8] there is no formal treatment for obesity. Instead, the focus is on prevention of overweight and obesity through the consumption of healthy food and drink. However, given our innate preference for high energy foods and drinks, especially those with a sweet taste,[9] it is difficult for individuals to moderate consumption within an obesogenic environment. In recognition of the role of the food environment, governments across the world have begun to establish policies which aim to modify the nutritional quality, availability and affordability of products on offer, in a bid to improve the health of their citizens.

The United Kingdom recommends that 'free sugars' provide less than 5% of the daily energy intake.[10] However, according to the latest data from the National Diet and Nutrition Survey (2016 - 2019 combined), free sugars intake among children (11-18 years old) and adults (19 – 64 years) accounted for 12.3% and 9.9% of total energy respectively.[11] Despite evidence of decline compared with the previously reported period (2014 – 2016), consumption of free sugars clearly remains too high, particularly among children.[11]

The Soft Drinks Industry Levy

In April 2018 the UK Government introduced a Soft Drinks Industry Levy (SDIL), known colloquially as the "sugar tax" and often referred to in this report as "the levy".[12] The main goal of the levy was to decrease the consumption of sugar in the UK by applying a tax of 18 pence/litre on drinks containing more than 5g of sugar/100 mL and a tax of 24 pence/litre on drinks containing more than 8g of sugar/100 mL 9. The SDIL was designed to incentivise manufacturers to reformulate soft drinks and to discourage excessive purchases among customers. Evidence to date points to the effectiveness of the levy in reducing intake of sugar from soft drinks at the population level,[13] however less is known about how different groups of people responded to the levy.

2.1 Research questions

The aim of this secondary data analysis is to investigate, using purchase data from a sample of loyalty card customers at leading supermarket Sainsbury's, the effect of the SDIL policy on purchases of soft drinks among different types of customers.

1. How did Sainsbury's loyalty card holders respond to the SDIL?
2. Did response to the levy differ according to customer demographic factors (age and gender)?
3. Did response to the levy differ according to area-level demographic factors (deprivation, geodemographic characteristics)?
4. Was difference in response to the levy determined by pre-levy purchase behaviours?

This investigation aims to increase our understanding of Sainsbury's customers' propensity to change their dietary purchase behaviours. This will support their broader aim of ensuring a healthy and sustainable diet is accessible to all, which would see more of their customers benefiting from healthy eating patterns. Furthermore, the analysis will support policy-makers in understanding the impact of fiscal policies on different types of people. This is particularly relevant in the context of Part 2 of the National Food Strategy report, which calls for the SDIL to be replaced with an expanded manufacturers' levy on added sugar and salt, which would impact both foods and beverages.[2]

3. Methods

3.1 Data overview

Data were provided by Sainsbury's and concerned all transactions performed by around 300K customers living in the Yorkshire and Humber region of England, between the first of January 2016 and thirtieth June 2019 (3.5 years). This covers 3-months prior to announcement of the SDIL (16 March 2016), 2 years before implementation (6 April 2018) and 1 year 3 months post-implementation.

Customers in the sample were selected as 'typical' shoppers based on their purchases during the 2016 calendar year. The sampling frame, described by Clark et al[14] in more detail, was designed to include loyalty card customers for whom Sainsbury's is likely to represent the majority of their diet. That is, they purchased from 7 out of 15 categories (or purchased a ready meal and three other categories) on at least 10 occasions throughout the year. These criteria are designed to exclude customers who purchase infrequently with Sainsbury's or from a limited range of categories, e.g. people who only ever purchase a lunchtime meal deal. The customer sample used for this secondary data analysis was originally identified for a different purpose, thus it should be acknowledged that the exclusion of 'meal dealers' is not ideal as it is likely to exclude some regular purchasers of soft drinks.

Selected customers were followed up over the study period and may thus be considered a cohort. It is possible for customers to be lost to follow up if they stopped purchasing at Sainsbury's, but new customers were not able to enter the sample after 2016.

3.2 Data governance

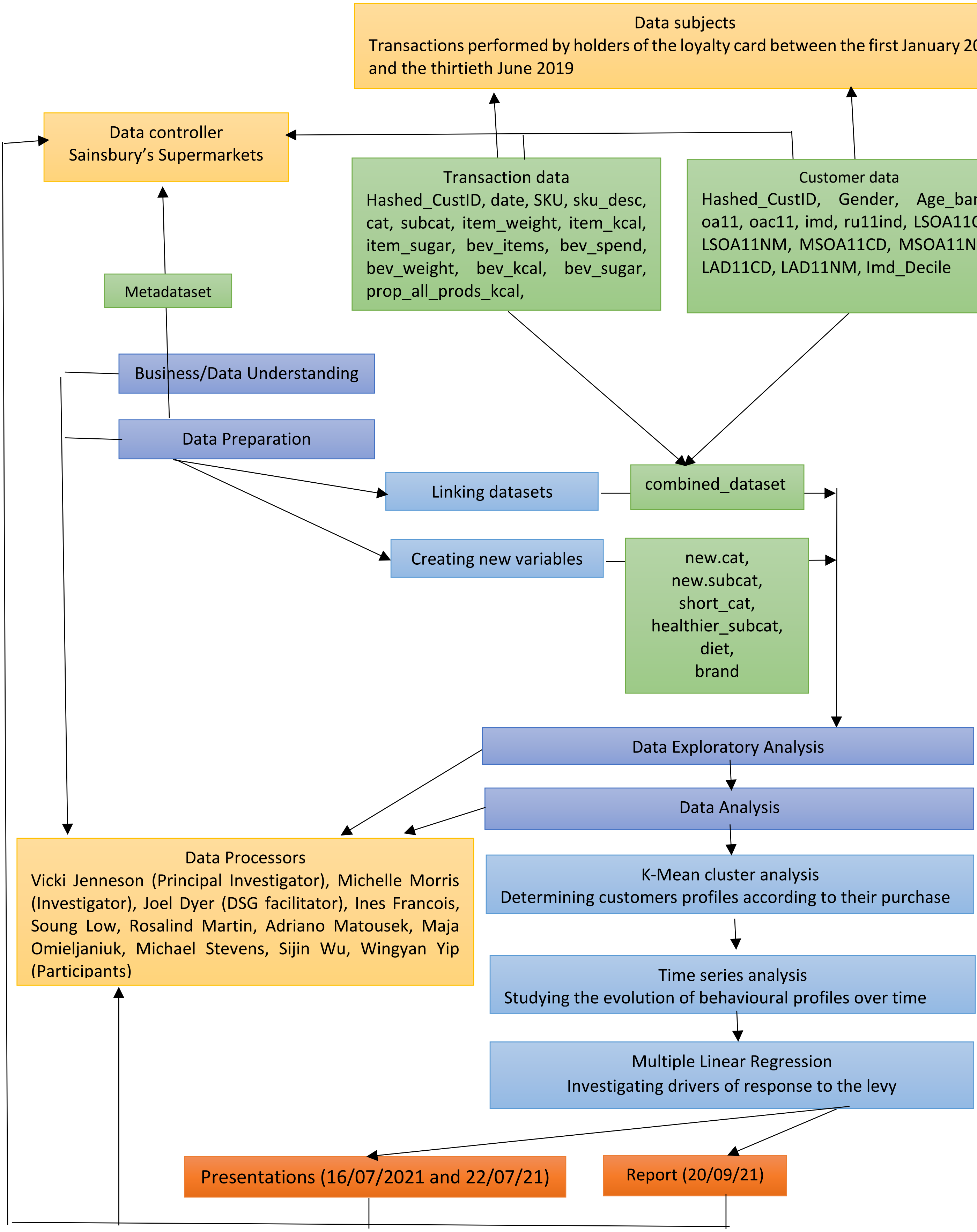
As these data are commercially sensitive, they were stored on LASER,[15] a secure data platform managed by the University of Leeds, to avoid risk of disclosure. All members of the research team undertook two training modules prior to access: Information Security Essentials and Information Security Advanced, to demonstrate competency in safe research practices. In addition, customer identities are pseudononymised using a unique customer ID. In line with statistical disclosure procedures, analyses and descriptive statistics presented exclude groups containing N<10 customers. To minimised exclusions and maximise insights, maps are presented to the Middle Layer Super Output Area (MSOA) level, a UK census geography containing a minimum of 5,000 residents.[16]

3.3 Data set description

The Data provenance chart (Figure 1) describes the whole process of this secondary analysis: data (green), project stages (dark blue), method (light blue), outputs (orange) and stakeholders (yellow).

The data consisted of two csv files; transactions and customers, for which each of the originally provided variables is described in Appendix 1 and Appendix 2 respectively. The transaction dataset contains 17 product purchase variables and ~43 million rows. Transactions represented all drinks purchased during the study period (1 January 2016 – 30 June 2019), where each row represents a product purchased by one loyalty card holder, aggregated by year. There are two weeks of missing data at the beginning of 2017. The customer dataset contains ~300,000 rows (where each row represents a unique loyalty card) and 13 demographic variables associated with the loyalty card owner or their area of residence. These two datasets were linked via the unique customer identification key, "Hashed_CustID". The creation of new variables is described in section 3.4 'Data preparation'.

Figure 1. Data Provenance Chart


Data subjects
Transactions performed by holders of the loyalty card between the first January 20
and the thirtieth June 2019
Data controller
Sainsbury's Supermarkets
Transaction data
Hashed_CustID, date, SKU, sku_desc, cat, subcat, item_weight, item_kcal, item_sugar, bev_items, bev_spend, bev_weight, bev_kcal, bev_sugar, prop_all_prods_kcal,
Customer data
Hashed_CustID, Gender, Age_bar
oa11, oac11, imd, ru11ind, LSOA11C
LSOA11NM, MSOA11CD, MSOA11N
LAD11CD, LAD11NM, Imd_Decile
Metadataset
Business/Data Understanding
Data Preparation
Linking datasets
combined_dataset
Creating new variables
new.cat,
new.subcat,
short_cat,
healthier_subcat,
diet,
brand
Data Exploratory Analysis
Data Analysis
K-Mean cluster analysis
Determining customers profiles according to their purchase
Time series analysis
Studying the evolution of behavioural profiles over time
Multiple Linear Regression
Investigating drivers of response to the levy
Data Processors
Vicki Jenneson (Principal Investigator), Michelle Morris (Investigator), Joel Dyer (DSG facilitator), Ines Francois, Soung Low, Rosalind Martin, Adriano Matousek, Maja Omieljaniuk, Michael Stevens, Sijin Wu, Wingyan Yip (Participants)
Presentations (16/07/2021 and 22/07/21)
Report (20/09/21)

3.4 Data preparation

3.4.1 Product categories

The product categorisation and sub-categorisation approaches are derived from business needs, and reflects a product's location in store and the organisational structure within the business, rather than its nutritional value. Categories are therefore not always well suited for nutritional research. A new categorisation approach was developed (detailed below) to describe products from a nutrition perspective. The popularity of each category- sub-category combination is shown in Table 1, according to ranked total items purchased over the study period. Plain milk, chilled juices and cola are the most popular beverage products while uncategorised coffee products, ethnic drinks and baby foods are the least popular.

1) The names of some categories were changed to aid more meaningful interpretation (e.g. Socialising into Fancier Drinks, empty subcategory in Coffee to Coffee Unknown).

2) Heterogeneous categories (e.g. Front of Store Juice, Soft Drinks Chiller), were removed and drinks reallocated to more nutritionally meaningful sub-categories (e.g. Water and Cola).

3) Some categories/sub-categories were merged (e.g. combined individual ethnic sub-categories into larger ethnic sub-categories – motivated by the small number of items).

4) New sub-categories were created, e.g. 'Flavoured Non Carbs' (non-carbonated) within Soft Drinks, to capture iced teas etc.

5) Irrelevant categories (cooking ingredients and alcoholic beverages) were removed.

6) A 'concentrates' category was created for drinks which are diluted before consumption (powdered drinks & squashes), as nutrients are incomparable with ready to drink beverages.

7) An additional 'short_cat' field was also added to provide a more aggregated grouping which captures the less relevant items in an 'Other' category.

8) A ('healthier_subcat') flag was added to represent a 'healthier option' according to public perception and nutritional guidelines e.g. fruit juices are considered 'healthier' than carbonated soft drinks.

'categories_clean.R', 'name_mining.py', 'preprocessing_product_name_mining.ipynb', 'preprocessing_reformulated_products.ipynb', 'eda_changes_in_sugar_level.ipynb', 'eda_products_reformulation.ipynb'

Table 1. New category and sub-category combinations ranked by popularity (from most to least-purchased)

	new.cat	**new.subcat**	**% of total_items**
1	Milk	Own Label Milk	32.20
2	Soft Drinks	Chilled Juice	9.33
3	Soft Drinks	Cola	6.31
4	Soft Drinks	Soft Drinks - Mixers	4.75
5	Coffee	Coffee - Instant	4.63
6	Milk	Branded Milk	4.26
7	Soft Drinks	Spark&Flav Water	4.26
8	Soft Drinks	Still Water	3.60
9	Soft Drinks	Flavoured Carbs	2.80
10	Milk	UHT Milk	2.76
11	Soft Drinks	UHT Fruit Juices	2.63
12	Soft Drinks	Chilled Smoothie and Niche	2.58
13	Tea	Normal Tea	2.36
14	Concentrates	Squash	2.33
15	Soft Drinks	Lemonade	2.10
16	Milk	Dairy Alternatives	2.07
17	Coffee	Coffee - Pure	1.67
18	Concentrates	Sweet Hot Beverages	1.30
19	Tea	Herbal Tea	1.14
20	Soft Drinks	Fancier Soft Drinks	1.08
21	Milk drinks	Flavoured Milk	1.05
22	Soft Drinks	Sport & Energy	1.03
23	Soft Drinks	Lunchbox	0.83
24	Concentrates	Coffee Machine Pods	0.81
25	Tea	Speciality Tea	0.58
26	Soft Drinks	Flavoured Non Carbs	0.57
27	Ethnic Grocery	Afro Crbbean Grocery	0.21
28	Milk drinks	Breakfast drinks	0.20
29	Baby Food	Follow On Milk	0.14
30	Ethnic Grocery	Asian Grocery	0.13
31	Baby Food	Infant Milk	0.08
32	Ethnic Grocery	Polish, Irish, American, Kosher	0.08
33	Milk drinks	Kefir	0.06
34	Concentrates	Pure Cocoa	0.03
35	Baby Food	Fruit Pouch	0.01
36	Ethnic Grocery	Ethnic Milk Drinks	0.01
37	Coffee	Coffee Unknown	0.00

‘categories_clean.R’, ‘name_mining.py’, ‘preprocessing_product_name_mining.ipynb’, ‘preprocessing_reformulated_products.ipynb’, ‘eda_changes_in_sugar_level.ipynb’, ‘eda_products_reformulation.ipynb’

3.4.2 SDIL categories

Prior to the data study group challenge, a flag was added to products indicating their SDIL status i.e. whether they are in an eligible drinks category and contain added sugar at or above the SDIL level.[12] As added sugar content is not reported (only total sugars), product category and sub-category were used to determine eligibility e.g. Milk and powdered drinks are out of scope, while juice is considered in scope as it is unclear if sugars are added or naturally occurring. Products are labelled as follows:

- Out of scope products are labelled 'Blank'.
- Eligible product categories are labelled 'No' if their total sugars content <5g.100ml,
- 'SDIL1' if total sugars ≥5g/100ml
- 'SDIL2' if total sugars ≥8g/100ml

This was performed for each unique product and sugar level combination, enabling SDIL status to change if sugar content changes.

During the challenge, manual correction for errors was performed by checking the ingredients lists of products. For example, a number of milk products which were incorrectly labelled as in scope, were reassigned as out of scope (Blank).

3.4.3 Missing data

In January 2017, two weeks of recorded transactions are missing due to a change of system. In the analysis sections below we describe how missing data were treated. Where no description is given, no adjustment for missing data was made.

3.4.4 Creating new variables

Three new variables ('brand', 'diet' and 'reformulated) were created (as described in Table 2) to analyse how customers' spending behaviours differed across brands and product type. In particular, it is interesting to know if customers switched to healthier products after the implementation of SDIL. Extraction of the 'brand' variable from the products' names allows us to analyse customers' spending across different brands at different time points as well as the willingness of manufacturers to reformulate their products. With the 'diet' variable, we could examine whether or not customers switched to products that are labelled as 'healthier' after the implementation of SDIL.

Table 2. Definition of new variables – 'brand', 'diet' and 'reformulated'

Variable	Description
brand	The manufacturer of the product extracted from the product's name. **Note that a manufacturer may be represented under different names across products. For instance, the names for Sainsbury's Taste The Difference are 'ttd' and 'js ttd'.*
diet	Whether or not the product features low calories.

	**This is determined by whether or not the following key words appear in products' name: 'diet', 'zero', 'sugar free', 'sugarfree', '100%', '7 up free', 'semi', 'pepsi max', '7up free', 'nas', 'skimmed', '50%', 'unsweetened', 'light', '1%', 'sprite z'.*
reformulated	A binary indicator showing if a product has been reformulated or not. **This is determined based on the presence of different sugar contents over the time period for the same SKU.*

Determining product 'brand'

To generate the 'brand' variable, we used text mining to gather the first word of each product's name after tokenization, based on an assumption that the first word is likely to be the brand of a product. This approach captures most brands accurately. For example, the most common brands in our data is 'js' (Sainbsury's own brand), followed by 'twinings' and 'nescafe'. However, we recognise that brand names could contain more than one word, such as 'irn bru' and 'yeo valley'. Additionally, not every product name starts with the brand, such as 'diet coke'. For better accuracy, we manually examined the brands collected by this approach, corrected those that were invalid, and assigned them into products' categories in which they are found, using the following steps:

1) Identify the category of the target product.
2) Match the first word of the product with brands in its category.
3) If Step 2 does not yield a result, tokenize the product name and match each word with brands.
4) If multiple matches occur, take the longest match.

Assigning the 'diet' variable

For the 'diet' variable, we list out the terms indicating the 'healthy' feature of a product and label a product as 'diet' if the product name contains any of the following terms: 'diet', 'zero', 'sugar free', 'sugarfree', '100%', '7 up free', 'semi', 'pepsi max', '7up free', 'nas', 'skimmed', '50%', 'unsweetened', 'light', '1%', 'sprite z'.

Identifying reformulated products

Reformulated products are those with identical Stock-Keeping Units (SKUs) (a number used by a business to identify unique products) but different sugar contents over the course of the data period. Using the transaction data, we are able to identify these products by comparing their sugar contents. A binary indicator was used to signify is a product was reformulated or not. Note that there are some products have a higher sugar level after reformulation, which might reflect a true increase (e.g. a change from added to natural sugars which are exempt from the levy), or result of a change in analytical method.

3.5 Exploratory data analysis results

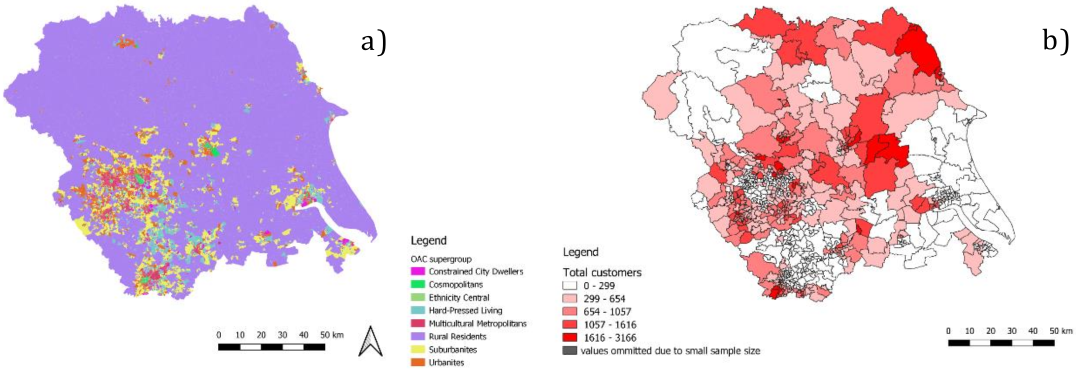
3.5.1 Data summary

In this section, we present descriptive statistics for the customer sample and describe the exploratory data analysis performed. This secondary data analysis presented in this report were performed in R studio (v.4.0.2) and in Python using Jupyter Notebook. Some graphs were generated on TABLEAU (v. 2020.3.2), maps were created using QGIS (v.3.2.1-1).

3.5.2 Descriptive statistics

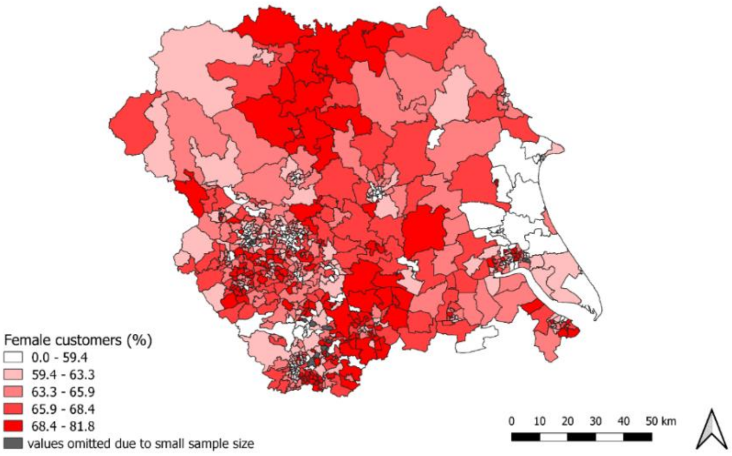
Figure 2a) shows the Output Area Classification (OAC) of Output Areas,[17] as determined by the Office for National Statistics Area Classification for Output Areas (a small neighbourhood census geography containing around 100 households)[16], across the Yorkshire and Humber study region. By area, the region is dominated by areas classified under the ‘Rural Residents’ supergroup, yet these areas are comparatively sparsely populated, with the majority of people tending to reside in more urban areas. Customers are concentrated in rural northern areas of the region, as well as the city suburbs, while the East coast and North West of the region have little customer coverage (Figure 2b)).

Figure 2.a) Output Area Classification Supergroups across the Yorkshire and the Humber region. b) Number of customers across the Yorkshire and Humber region by Middle Layer Super Output Area



Three Middle Layer Super Output Areas (MSOAs) were removed from this visualisation in line with statistical disclosure control procedures.

Figure 3. Percentage of female customers by Middle Layer Super Output Area across the Yorkshire and Humber region

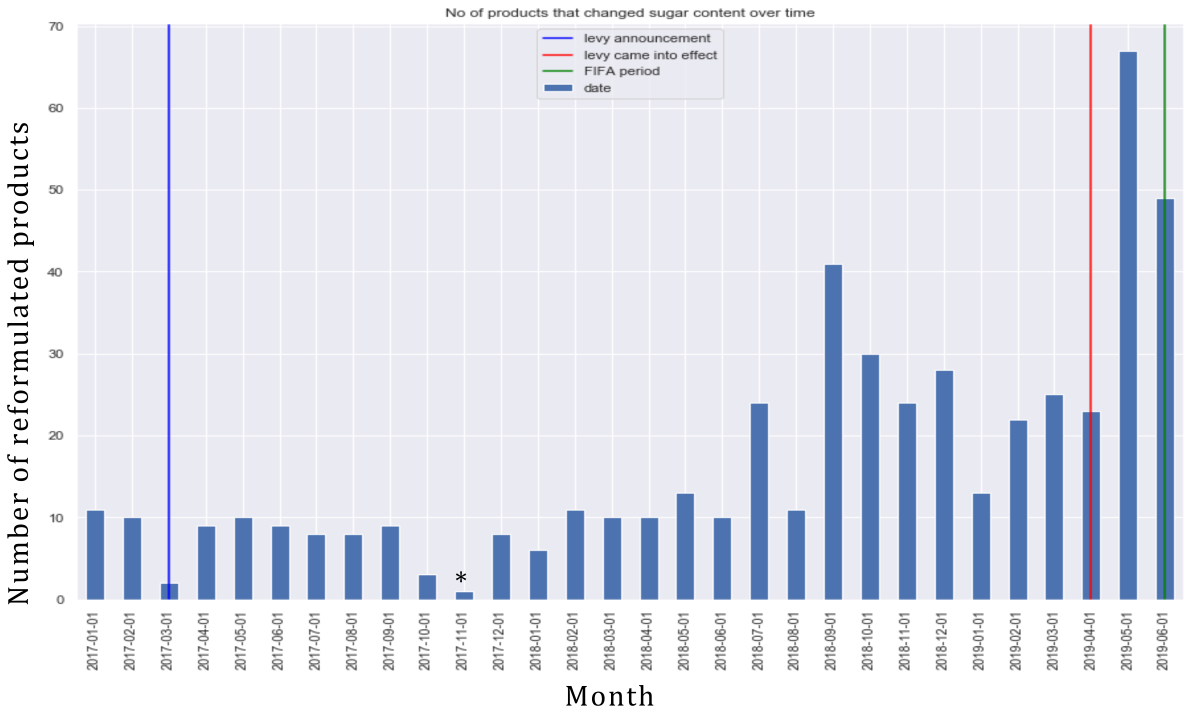


It is clear that in all MSOAs across the study area, loyalty card customers are predominantly females (Figure 3). The percentage of female customers were omitted in 16 MSOAs due to low numbers.

3.5.3 Manufacturer trends

We hypothesise that any effect of the SDIL was driven by 1) response by manufacturers, and 2) response by customers to changes in the market offering. Here we explore trends in manufacturer and customer responses separately. Using the label for reformulated products described in section 3.3.4 Figure 4 shows the number of reformulated products increased in the months preceding the SDIL, reaching its highest level in the month after implementation.

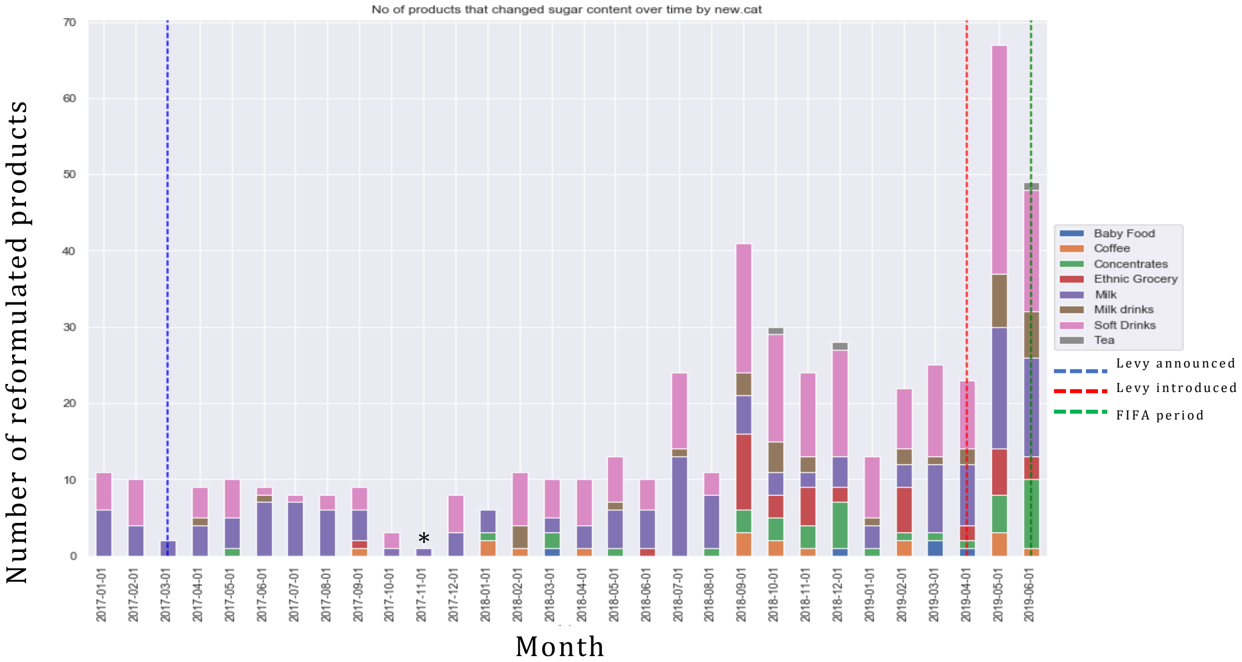
Figure 4. Number of reformulated products in each month



* two weeks of missing data in Janurary 2017. FIFA period refers to start of the FIFA world cup (known to be associated with an increase in sales of soft drinks)

If we break this down by category (Figure 5), we find that in general, soft drinks (the most common product type) represent the category that has the most reformulated drinks. In the month after the levy, we see highest number of reformulated products in soft drinks, followed by milk and milk drinks.

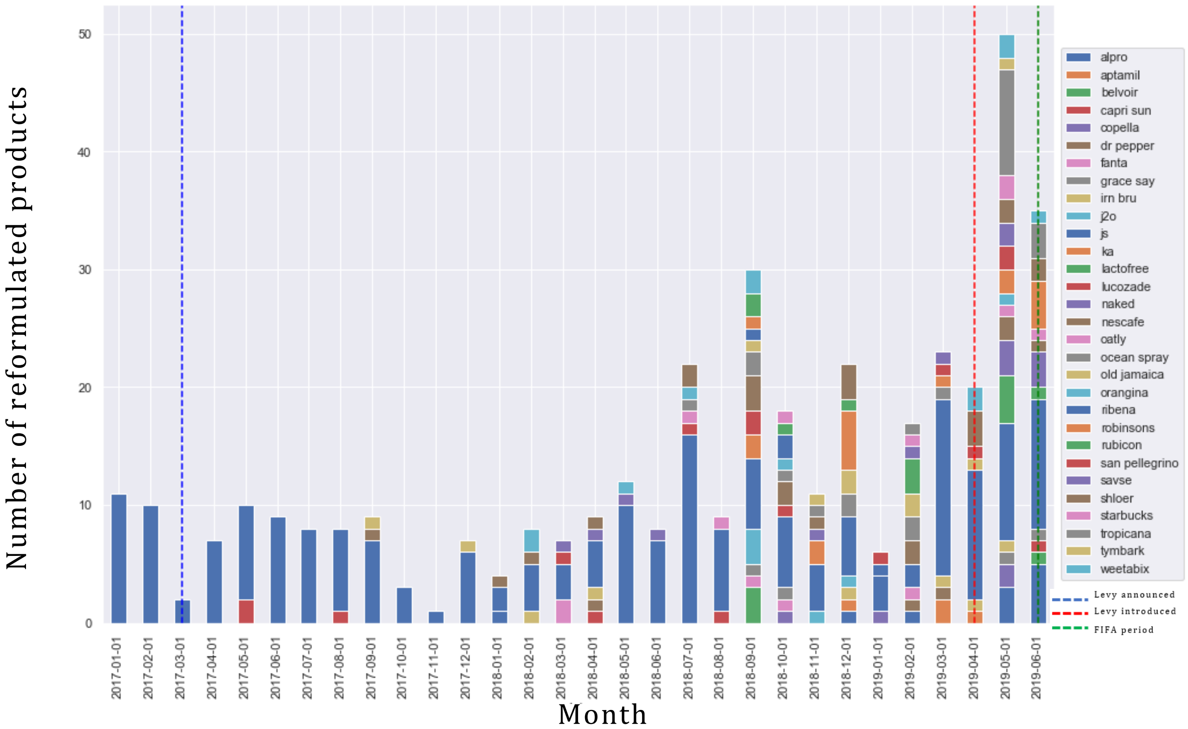
Figure 5. Number of reformulated products in each month by category



* two weeks of missing data in Janurary 2017. FIFA period refers to start of the FIFA world cup (known to be associated with an increase in sales of soft drinks)

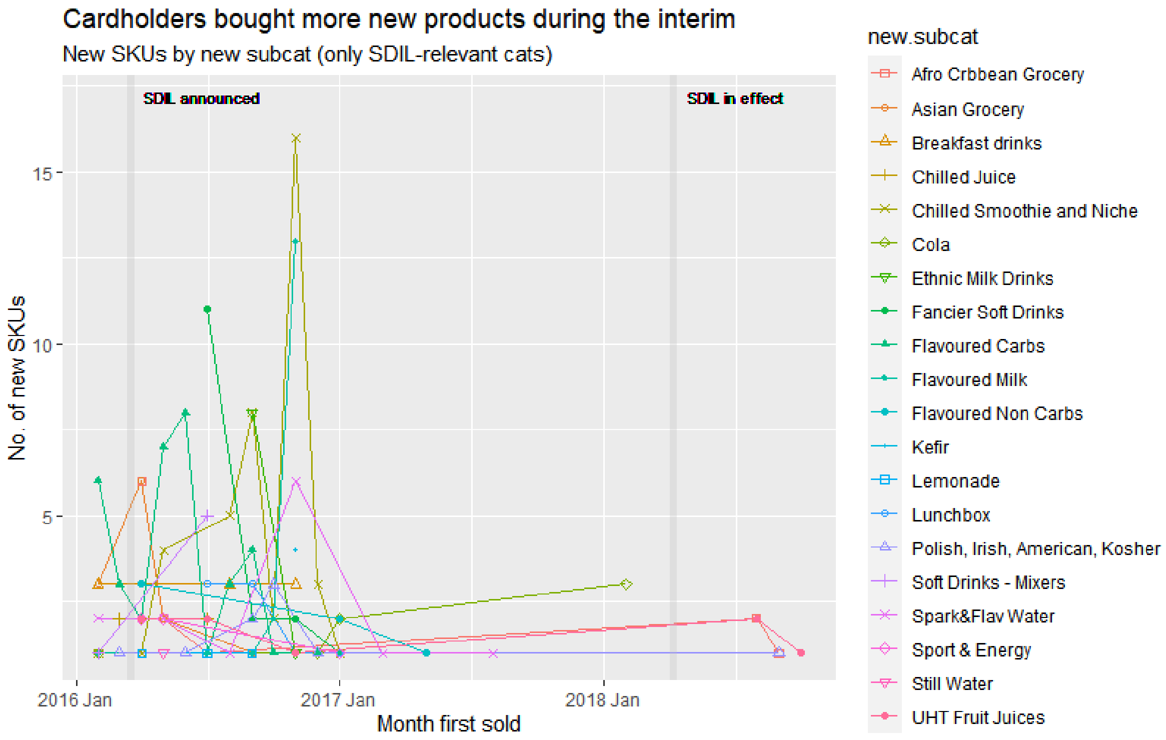
Figure 6 shows that around the announcement of the levy, reformulation mainly occurred within Sainsbury's brands. As the levy implementation became closer, the number of brands bringing reformulated products to market increased, peaking in the month after the levy. It should be noted that JS was the most common brand in the dataset.

Figure 6. Number of reformulated products in each month by top 30 brands



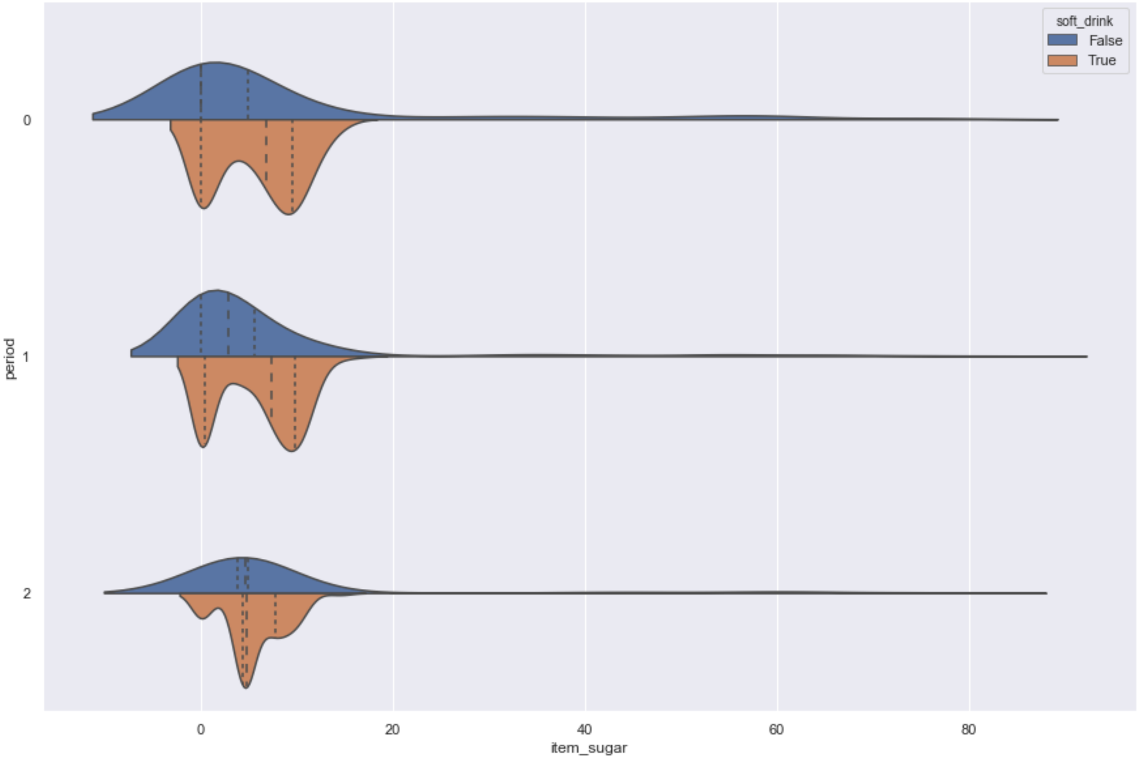
It would appear that the increase in focus on reformulation by drinks manufacturers was to the detriment of new product development. A product is considered ‘new’ when its SKU appears in the dataset for the first time after January 2016. Despite a post-announcement peak in new SKUs coming to market, there was a marked decline in the period leading up to the levy implementation, and after implementation (Figure 7).

Figure 7. Number of new products (Stock Keeping Unit) over time, by subcategory



Following the analysis of the reformulated products, we explored how the overall sugar density (g/100ml) of products changed as a result of the levy. Figure 8 shows that while the overall sugar level of non-soft drinks remained fairly stable, that of soft drinks (shown in orange) dropped significantly over the study period, as evidenced by the fall in the mean (represented by the centre dotted line in the violin plot). Interestingly, we also see that the distribution of sugar levels for soft drinks changed from a bimodal distribution to a unimodal distribution, indicating that sugar levels across different soft drinks are converging. This may be due to reformulation and new product innovations.

Figure 8. Distribution of sugar density (g/100ml) by unique products in three periods.
Period 0: pre-announcement period; Period 1: post-announcement period; Period 2: implementation period.



3.5.4 Customer trends

To see whether or not the fall in sugar levels of soft drinks translates to a fall in sugar purchased by customers, we examined the quantity of sugar purchased from soft drinks over time in Figure 9. Despite a general decline in the sugar consumptions from soft drinks, the steepest fall occurred in the post-announcement period. This coincides with an increase in the number of reformulated products on the market (Figure 6**Error! Reference source not found.**), suggesting that reformulation of soft drinks began to have an impact on levels of purchased sugar even before the levy was implemented. The continued fall in purchased sugar from soft drinks in the post-implementation period may also be due to the sharp increase in price per 100ml of levy eligible drinks (SDIL1 and SDIL2) after implementation of the levy (Figure 10).

Figure 9. Average monthly sugar purchased (g) per customer from carbonated soft drinks

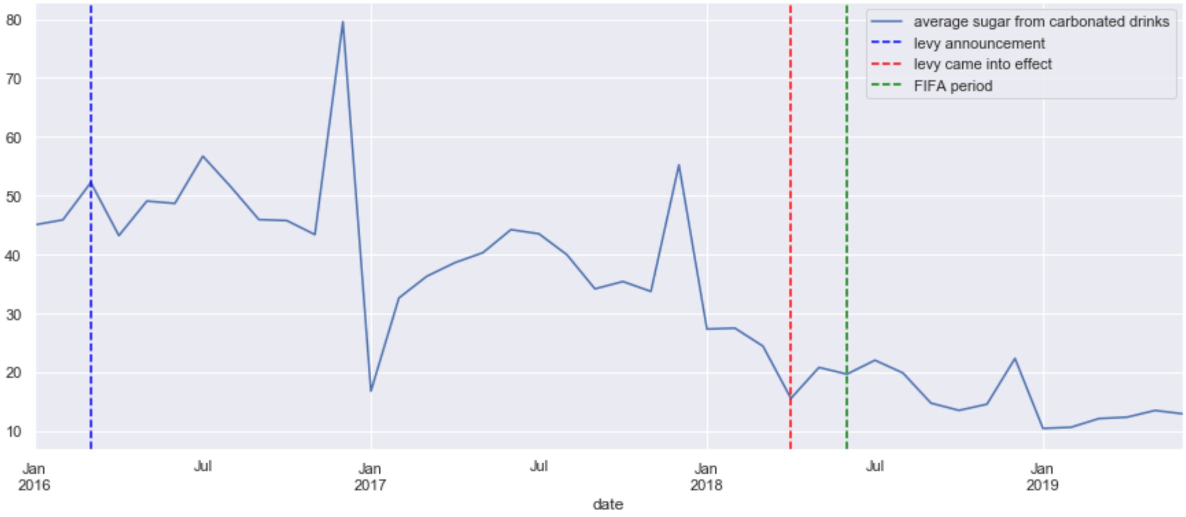
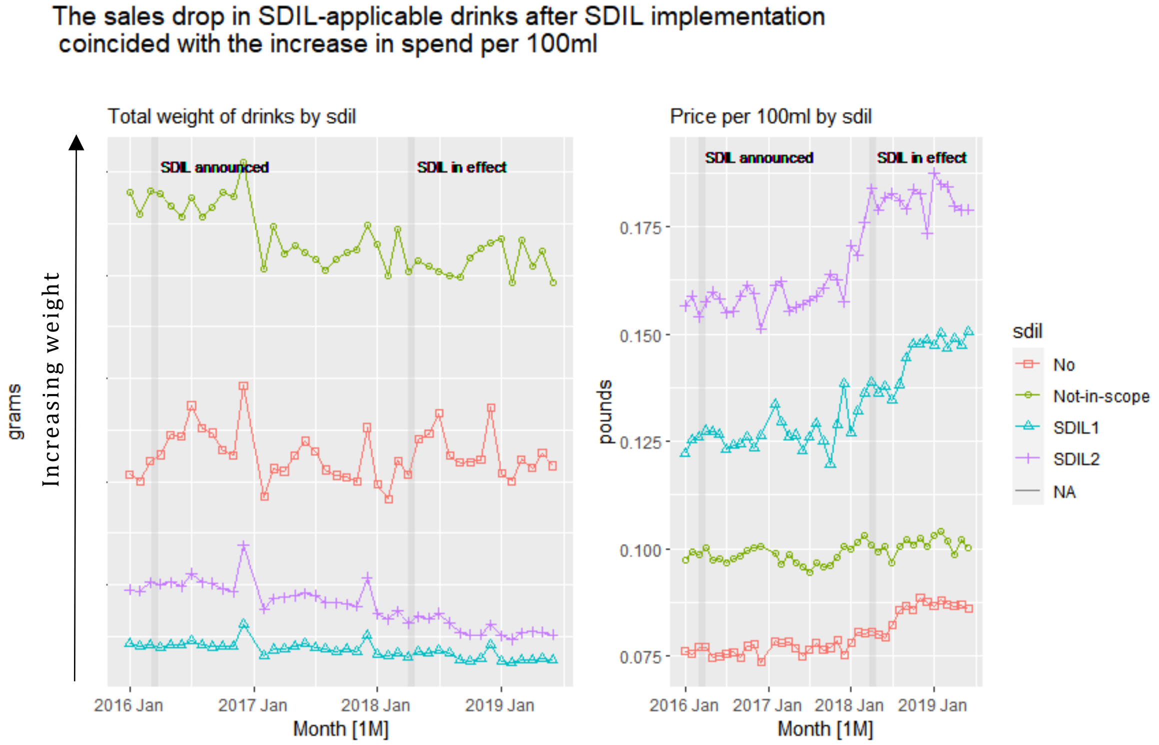
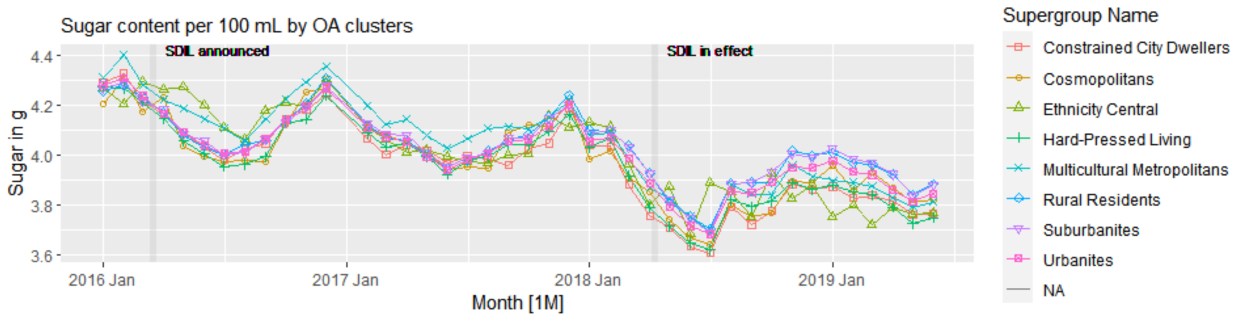


Figure 10. Time trend showing sales and spend on beverages by SDIL status



We hypothesised that different groups of customers might respond differently to the levy. Figure 11 indicates that there may be some geographic component to determination of customer behaviour. While all OA supergroups follow a general trend of decline in the average sugar content of purchased beverages, there are insightful differences by supergroup. For example, customers living in Ethnicity Central and Multicultural Metropolitan areas started with a preference for the highest sugar drinks, but showed the greatest decline over the period, suggesting greatest sensitivity to the levy. This may indicate a willingness to switch to lower sugar options, though volume purchased is not accounted for. Contrastingly, customers living in rural and suburban areas appear to be the least sensitive to the levy, demonstrating the smallest declines in the period immediately post-levy, and the greatest 'bounce back' during the post-levy period. These geodemographic differences may be indicative of differing price sensitivity due to affluence, or differences in shopping habits due to availability and distance from store, for example.

Figure 11. Average sugar content (g/100ml) of purchased drinks by Output Area Supergroup



Our exploratory data analysis supports our hypothesis, indicating that different customer groups appear to have responded differently to the levy. During the remainder of the report, we present a range of clustering approaches which were explored to reveal customer-level behavioural insights.

4. Classification using K-means clustering

K-means clustering was implemented to answer the question: *"Do pre-levy beverage purchasing behaviours determine response to the SDIL?".* Clustering was used to identify groups of customers with similar beverage purchase behaviours, irrespective of demographic features. K-means is an unsupervised clustering algorithm which identifies distinct groups of customers who share similar characteristics; it is commonly applied to customer segmentation problems.[18] The K-means algorithm iteratively moves through the process of clustering data by assigning each data point to one of a pre-specified number of clusters, then calculating the mean within each cluster. Based on the cluster means, data are then re-assigned to clusters and the process repeats until cluster means converge on a set of values.

4.1 Methods: K-means

A sample of 121,284 customers who had all purchased soft drinks in at least six months during the four-year time period were included. Two customers with duplicated demographic records were removed.

Purchase data were split into three distinct time points (T); T1) pre-announcement (January 2016 – February 2016), T2) pre-implementation (March 2016 – March 2018), T3) post-implementation (April 2018 – July 2019). The k-means algorithm was applied to customers at T1 (January 2016 to February 2016), representing baseline soft drink purchasing habits prior to the public announcement of SDIL.

Centroids of the generated T1 clusters were then used to assign customers to a cluster in T2 and T3, independently of their T1 cluster assignment. Applying the clusters to all three time points allowed us to assess whether customers migrated between profiles in response to the SDIL or, at the very least, what customer profiles looked like prior to and after the SDIL was implemented. This approach differs from the other clustering methods applied (described later), as it operates on a heavily discretised temporal domain.

4.1.1 Running the K-means algorithm

Seven variables were calculated from customer transaction data and identified as candidates for inclusion in the K-means clustering algorithm:

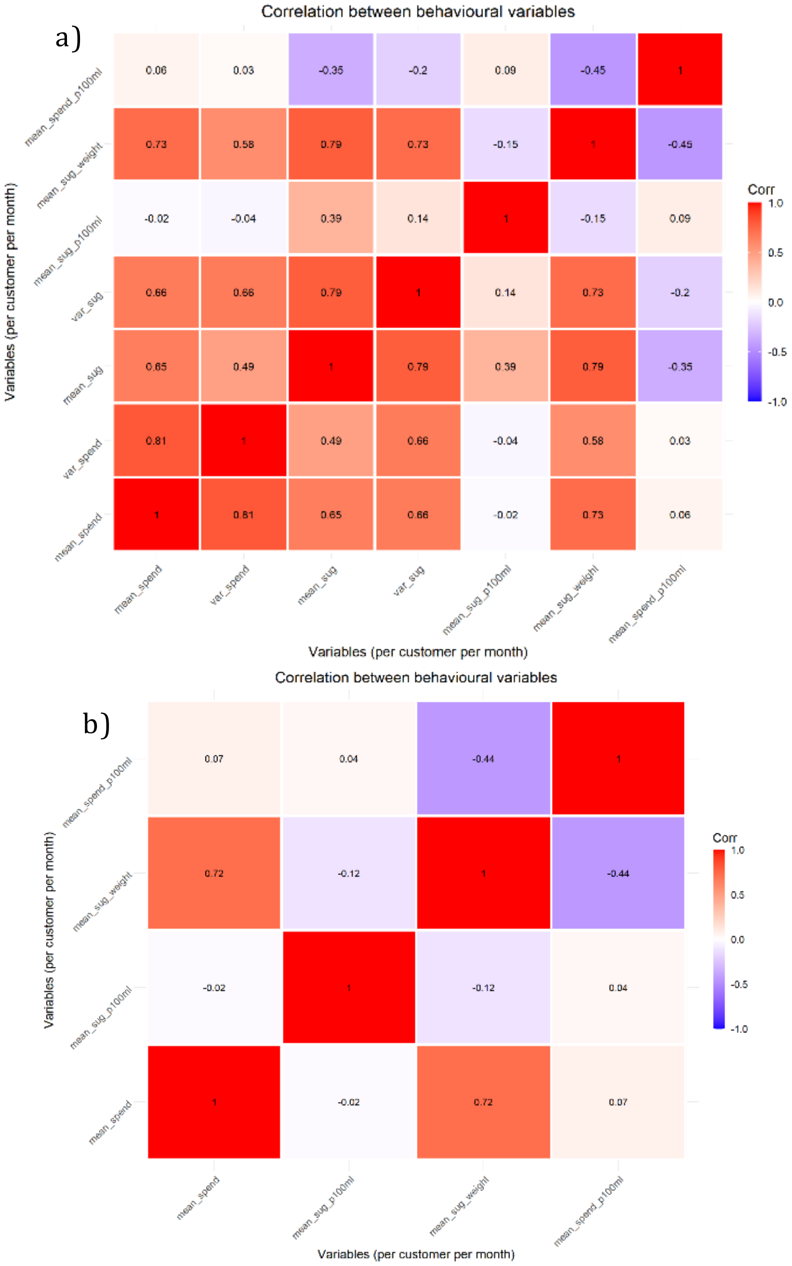
1) Mean monthly spend per 100ml per customer
2) Mean monthly sugar weight (g) per customer
3) Mean sugar (g) per 100ml per customer
4) A measure of a customer's variance of sugar purchased
5) Mean monthly sugar purchased (total) per customer
6) A measure of a customer's variance of spend
7) Mean monthly spend per customer

Candidate variables were firstly transformed (square-rooted) to reduce skewness, as performed by Clark et al[14]. Then, they were standardized by calculating z-scores as follows.

$$z - score = (data - mean\ of\ the\ \ data)\ /\ standard\ deviation\ of\ the\ data.$$

Standardised variables were tested for correlation, as K-means clustering can be skewed by highly correlated variables. For the purpose of variable selection, a Pearson's correlation coefficient over 0.75 was considered high, above this level variables were excluded. Figure 12 shows the correlation coefficients of candidate and selected variables. The four selected variables for k-means clustering were; mean spend per 100ml, mean total sugar (g), mean sugar per 100ml, and mean total spend (£).

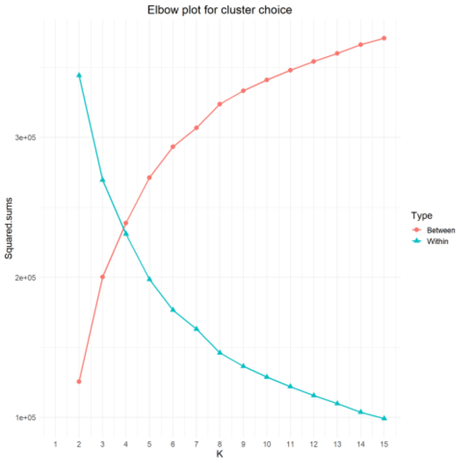
Figure 12. Pearson's correlation heatmaps 16a) Correlation for 7 candidate variables, 16b) Correlation for 4 selected variables



Code for reproducing K-means analysis – "Clustering.R"

The K-means algorithm (base R function kmeans()) ran iteratively for between 2 and 15 pre-specified clusters. The standard procedure of minimising the sum of squared (SOS) distances within, and maximising SOS distances between clusters was used to determine the optimal cluster number as 8, as visualised on the elbow plot in Figure 13. The between cluster distance (red) increased dramatically until around K = 8, at which point, the within distances (blue) decreased at a steady rate, indicating the optimal point at which each cluster is distinct from any other.

Figure 13. Elbow plots illustrating the sum of squared distances between (red) and within (blue) clusters for each pre-specified number of clusters (2 – 15)

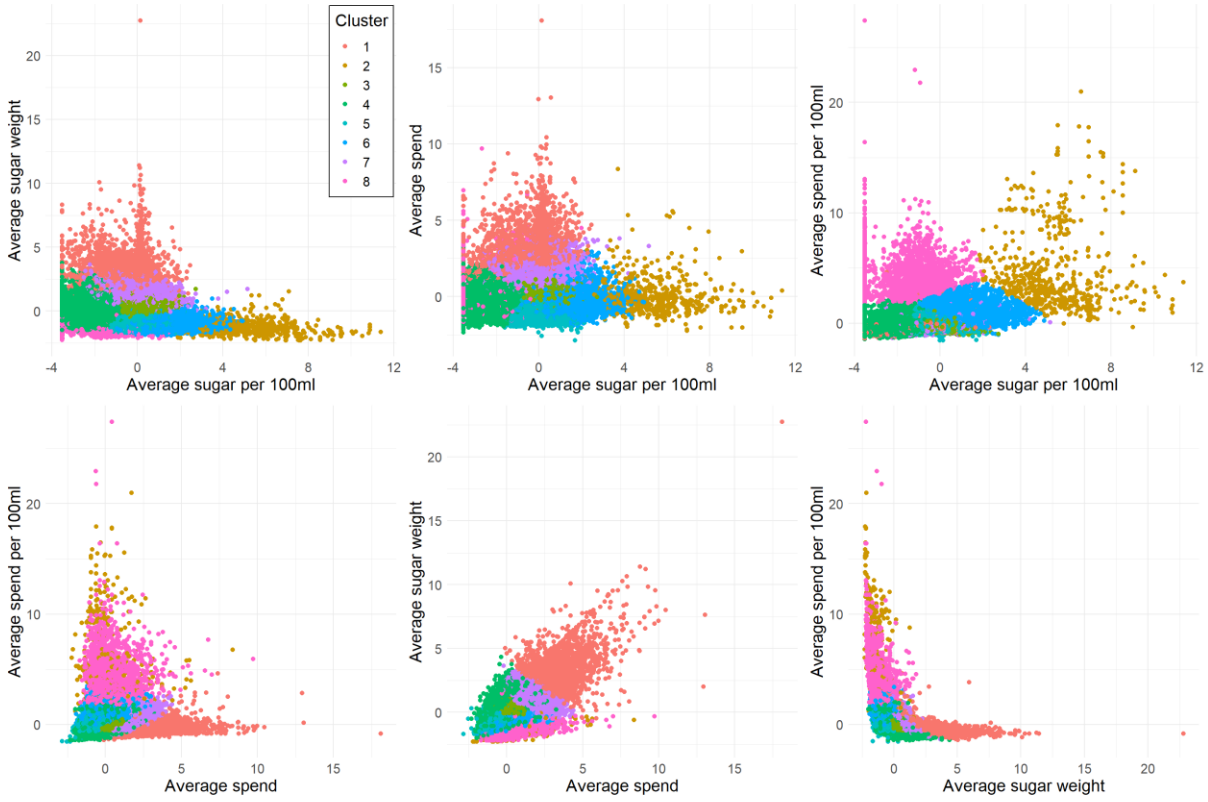


4.3 Results: K-means

Scatter plots of each variable pair combination show how the clusters differ in their behaviours (Figure 14). Each point represents a single customer, where points are colored based upon their cluster assignment. The bottom right graph indicates that customers in cluster 1 (‘Volume bargain buyers’) purchase a high amount of sugar from inexpensive soft drinks. The scatterplots and means for each cluster (Table 4) were then described in terms of volume, spend and sugar density of drinks to generate pen portraits to describe each cluster (Table 3).

Code for reproducing K-means analysis – “Clustering.R”

Figure 14. Scatterplots for each combination pair of behaviour variables for the pre-announcement time period.



Using the centroids of T1 clusters, customers were assigned to the most appropriate cluster in T2 and T3, to assess how customers move between clusters over time. Movement between clusters is visualised in the alluvial plot in

Figure 15. From T1 – T3, the Cheap Diet Drinkers (cluster 4) and Occasional Originals (cluster 6) clusters became larger, while the Average Joes (cluster 3) and Rare Treat (cluster 5) clusters became smaller, suggesting a move towards lower sugar options (Table 3. Pen portraits for each cluster

Cluster	Volume (Weight)	Cost (Spend/100ml)	Sugar (Mean Sugar/100ml)	N customers (T1)	Pen portrait	Description
1	High	Low	Low	3430	Volume bargain buyers	These customers buy large volumes of relatively low sugar drinks, adding up to a high spend and lots of sugar overall. The lowest spend/100ml, indicating a relatively high

Code for reproducing K-means analysis – “Clustering.R”

						price sensitivity.
2	Low	High	High	599	Occasional sugar splurge	These customers don't buy often, but when they do they splash out on expensive sugary drinks. With the highest spend/100ml, they’re unlikely to be sensitive to price.
3	Medium	Low	Medium	32108	Average Joes	Fairly regular purchasers of mid-priced sugary drinks.
4	Medium	Low	Low	11394	Cheap diet drinkers	These customers choose low sugar drinks and buy them fairly regularly, they’re not prepared to spend much.
5	Low	Low	Medium	32819	Rare treat	The least frequent purchasers of soft drinks, these customers appear not to be bothered by sugar content when they do choose to purchase.
6	Low	Medium	High	17974	Occasional originals	This group buys infrequently but when they do they're prepared to pay a bit more and don't mind a bit of sugar - probably choosing original over diet versions of their favourite drinks. This group may be unwilling to

Code for reproducing K-means analysis – “Clustering.R”

						change their behaviours.
7	High	Low	Medium	16238	Regular sugar fix	These customers buy lots of relatively inexpensive sugary drinks. This group may be sensitive to price but less willing to change.
8	Low	High	Low	2878	Low sugar big spenders	These customers don't buy beverages often, but they're prepared to pay more for the occasional low sugar premium drink.

Table 4). Interestingly, the Cheap Diet Drinkers (cluster 4) and Low Sugar Big Spenders (cluster 8) both saw an increase in their average sugar content/100ml, suggesting that reformulation may have seen low-medium sugar drinks entering their repertoire. This translated to an increase in overall sugar purchased by Low Sugar Big Spenders (cluster 8), yet their group mean remained low compared with other clusters.

Summary statistics (Table 5) show that the majority of change occurred between T1 and T2, while around a third of customers did not change cluster. Following implementation of the levy (T2 – T3), 13.7% of the population moved to a cluster with a lower sugar content while 57.1% remained within the same cluster and thus, the same sugar level. The remainder of the population (29.2%) moved to a cluster with a higher sugar content.

Code for reproducing K-means analysis – “Clustering.R”

Table 3. Pen portraits for each cluster

Cluster	Volume (Weight)	Cost (Spend/100ml)	Sugar (Mean Sugar/100ml)	N customers (T1)	Pen portrait	Description
1	High	Low	Low	3430	Volume bargain buyers	These customers buy large volumes of relatively low sugar drinks, adding up to a high spend and lots of sugar overall. The lowest spend/100ml, indicating a relatively high price sensitivity.
2	Low	High	High	599	Occasional sugar splurge	These customers don't buy often, but when they do they splash out on expensive sugary drinks. With the highest spend/100ml, they're unlikely to be sensitive to price.
3	Medium	Low	Medium	32108	Average Joes	Fairly regular purchasers of mid-priced sugary drinks.
4	Medium	Low	Low	11394	Cheap diet drinkers	These customers choose low sugar drinks and buy them fairly regularly, they're not prepared to spend much.
5	Low	Low	Medium	32819	Rare treat	The least frequent purchasers of soft drinks, these customers appear not to be bothered by sugar content when they do choose to purchase.
6	Low	Medium	High	17974	Occasional originals	This group buys infrequently but when they do they're prepared to pay a bit more and don't mind a bit of sugar - probably choosing original over diet versions of their favourite drinks. This group may be unwilling to change their behaviours.
7	High	Low	Medium	16238	Regular sugar fix	These customers buy lots of relatively inexpensive sugary drinks. This group may be sensitive to price but less willing to change.
8	Low	High	Low	2878	Low sugar big spenders	These customers don't buy beverages often, but they're prepared to pay more for the occasional low sugar premium drink.

Code for reproducing K-means analysis – "Clustering.R"

Table 4. K-means cluster means and number of customers

	Mean spend on beverages (GBP £)				**Mean sugar g/100ml**				**Mean sugar weight (Kg)**				**Mean spend/100ml (GBP £)**				**N Customers (%)**			
Cluster	T1	T2	T3	D T1-T3	T1	T2	T3	D T1-T3	T1	T2	T3	D T1-T3	T1	T2	T3	D T1-T3	T1	T2	T3	D T1-T3
1. Volume bargain buyers	12.84	8.99	9.13	-3.71	3.93	3.65	3.58	-0.35	16.97	12.43	12.34	-4.63	0.09	0.09	0.09	0.00	3,430 (2.9)	3,513 (2.9)	3,421 (2.8)	-9
2. Occasional sugar splurge	3.63	3.84	3.71	+0.08	29.11	15.69	16.53	-12.58	0.82	1.83	1.08	+0.26	2.04	1.18	1.1	-0.94	599 (0.5)	522 (0.4)	536 (0.44)	-63
3. Average Joes	3.37	3.15	3.22	-0.15	4.83	4.7	4.48	-0.35	3.80	3.57	3.54	-0.26	0.10	0.10	0.10	0.00	32,108 (27.0)	30,235 (24.9)	30,661 (25.3)	-1,447
4. Cheap diet drinkers	2.78	2.83	2.91	+0.13	1.39	2.01	1.75	+0.36	3.45	3.41	3.30	-0.15	0.11	0.12	0.12	+0.01	11,394 (9.7)	13,737 (11.3)	13,949 (11.5)	+2,555
5. Rare treat	1.70	1.89	1.93	+0.23	4.52	4.39	4.14	-0.38	1.85	2.01	1.97	+0.12	0.10	0.12	0.12	+0.02	32,819 (27.9)	31,552 (26.0)	31,891 (26.3)	-928
6. Occasional originals	2.83	2.72	2.83	0.00	6.93	6.01	5.73	-1.2	1.47	1.70	1.67	+0.20	0.24	0.25	0.25	+0.01	17,974 (15.3)	22,240 (18.3)	21,087 (17.4)	+3,113
7. Regular sugar fix	6.21	5.15	5.26	-0.95	4.54	4.46	4.32	-0.22	7.13	5.95	5.90	-1.23	0.10	0.10	0.10	0.00	16,238 (13.8)	16,661 (13.7)	16,774 (13.8)	+536
8. Low sugar big spenders	3.90	3.50	3.68	-0.22	1.89	3.31	2.94	+1.05	0.83	1.12	1.10	+0.27	1.51	0.89	0.88	-0.63	2,878 (2.5)	2,778 (2.3)	2,927 (2.4)	+49
																TOTAL	117,440	121,208	121,246	8,700

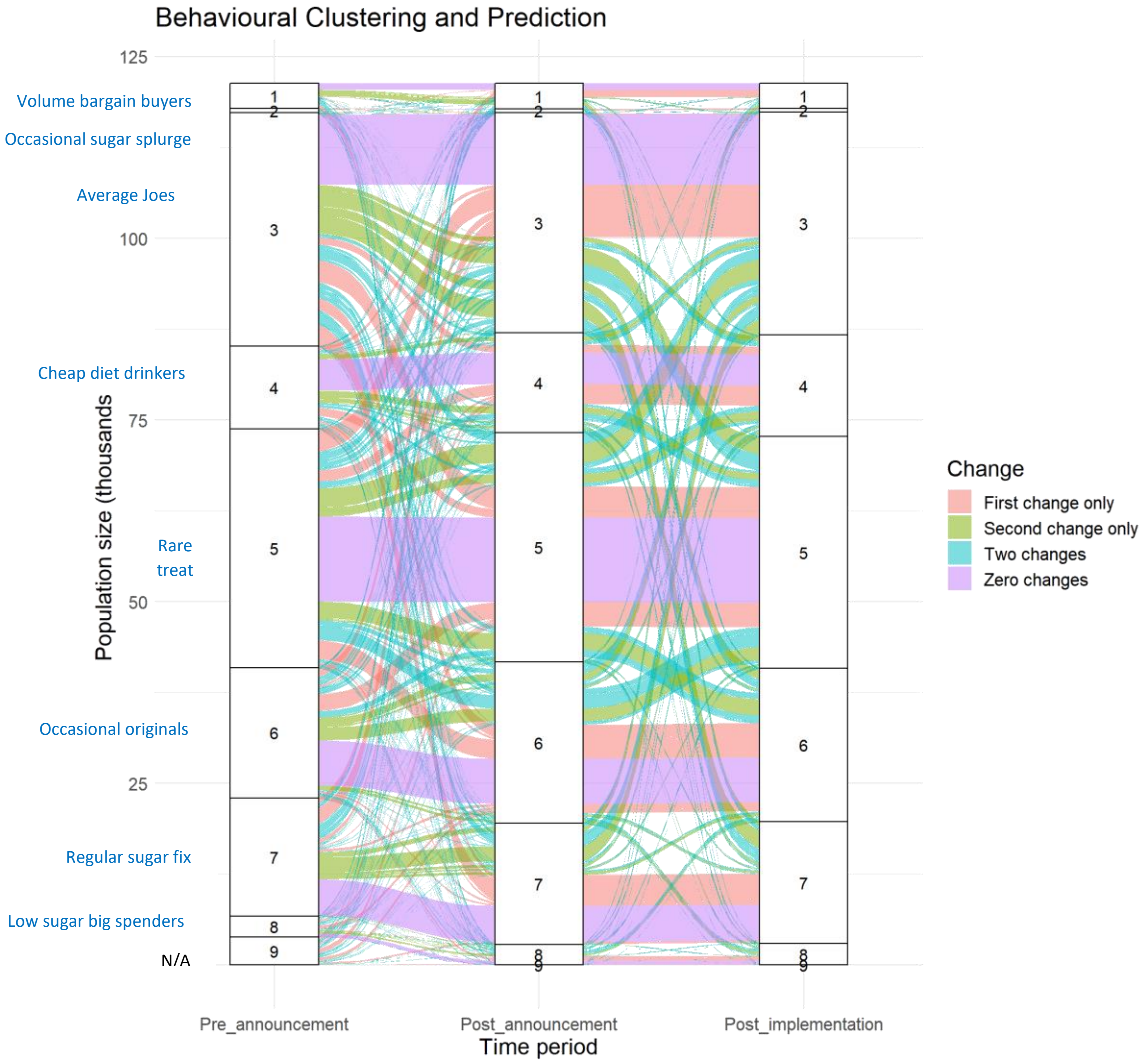
D = Difference. Red = increase, Green = decrease, Yellow = no chang

Code for reproducing K-means analysis – “Clustering.R”

Table 5. Summary of customer movement between clusters

Number of changes between the cluster	Customers	%
First change only	30945	25.5%
Second change only	24414	20.1%
Two changes	27658	22.8%
Zero changes	38267	31.6%

Figure 15. Alluvial plot showing customer movement between clusters between timepoints.



4.4 Conclusions: K-means

We identified eight customer clusters based on beverage purchase behaviours. Around a third of customers were behaviourally sticky, remaining in the same cluster at all three time points, while the majority changed their beverage purchasing habits. It is possible, given the short time-period in T1 (2 months), that seasonal trends may have influenced the original cluster assignments, particularly as there appears to be a consistent seasonal decline in beverage purchases in January. This could explain the unexpected finding that 30% of customers moved into higher sugar clusters.

Code for reproducing K-means analysis – “Clustering.R”

Methods for assessing cluster suitability, or extension of the T1 period should be explored.

Our approach used T1 cluster centres to assign clusters at T2 and T3, so it was not possible to examine if new clusters emerged or disappeared over time. Future research could explore this by creating new clusters at each time point. Unexpectedly, our analysis indicated there was little difference in the demographic characteristics of clusters (data not shown). This may be due to the limited degree of demographic data available, but further demographic/spatial exploration of differences between clusters and determinants of cluster switching is warranted. Our analysis was unable to account for customer loyalty, purchases of other (non-beverage) products, and promotional offers. Finally, alternative clustering methods may be employed.

Code for reproducing K-means analysis – "Clustering.R"

5. Network analysis

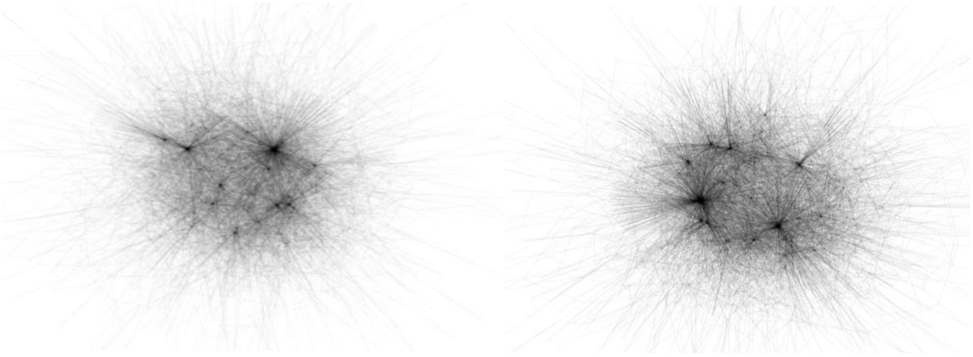
Network Analysis was considered to identify and understand trends in customer purchases. This section documents our thinking, preliminary findings, and suggests applications for future exploration. However, exploration was abandoned early due to time constraints of the 2-week DSG and concerns over suitability.

5.1 Methods and results: Network analysis

'Community Detection' involves evaluating the structure of large complex networks by looking at the division of groups or sets of nodes. Network Analysis Community-Detection (NACD) focuses on the relationships between nodes, rather than node attributes, unlike other clustering approaches explored in this report (K-means and Time-series clustering). It should be mentioned that late-fusion approaches[19] to NACD do allow for consideration of node attributes, but this was not explored.

Figure 16 demonstrates dynamic network graphs (DNGs)[20] constructed using NACD using 2016 (pre-levy) and 2019 (post-levy) transaction data. The DNGs are composed of customer nodes, linked based on their top three favourite beverages (by spend). Darker lines and closer positioning represent stronger relations between customers, thus dark areas on the DNG represent clusters of customers with similar beverage preferences. These may be detected formally using a community-detection algorithm. A similarity coefficient (e.g. Jaccard) could indicate the mapping of clusters between timepoints, while changes in purchase-profile over time may be visualised through an animated graph.

Figure 16. A network graph of customers connected by their top three favourite beverages (left 2016; right 2019)



5.2 Conclusions: Network analysis

Network analysis was proposed to explore the relationship between customers based on shared beverage preferences, but was abandoned early. One reason for this is that the relation links between nodes (customers) do not occur naturally and must therefore be engineered. Thus, the appropriateness of network analysis is dependent upon the ability to generate valid feature links. Additionally, technical challenges contributed to

Code for Network analysis – 'network_monthly_clustering.ipynb' our decision to halt analysis. Firstly, the large data volume proved computationally challenging, requiring a lot of run time. Secondly, the main python package for dynamic community detection, ‘CDlib’, was not readily available within our closed virtual research environment. Considering the time requirements of requesting the package, familiarising ourselves with the methods, and running the analysis, we were not confident we could complete generate sufficient insight within the two-week data study group.

We are intrigued by the potential of network analysis for exploring customer behaviours. Future work should spend substantial effort on the up-front decisions around node relation link features, backed by theoretical reasoning and domain knowledge. Further exploration should also consider late-fusion approaches which can better leverage the vast number of features available in the customer data.

Code for Network analysis – ‘network_monthly_clustering.ipynb’

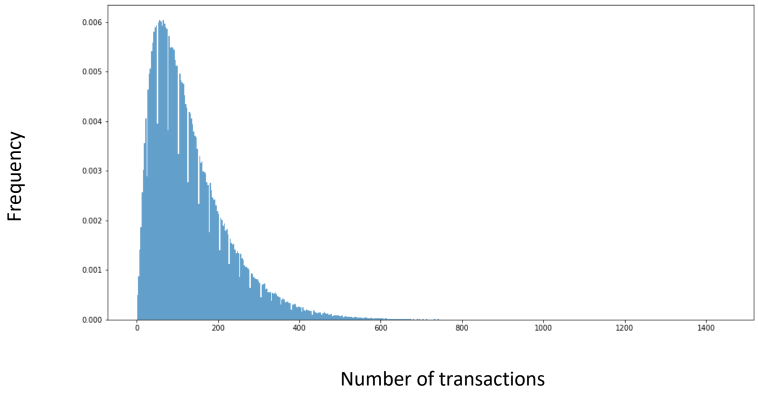
6. Time-series clustering

Time-series clustering was applied to investigate common patterns of purchase trends among customers during the whole time-period (January 2016 – June 2019). Unlike K-means, which clusters customers based on purchases at a discrete time-point, the time-series clustering approach clusters customers based on their overall purchase trend (or shape of their time-series curve). We propose that the overall time-series trend presented in Figure 9 (section 3.4.3) is an average for all customers and actually represents a series of underlying time-series patterns for different customer groups. Here we employ unsupervised machine learning to identify these clusters according to time-series trends.

6.1 Methods: Time-series clustering

The majority of customers made fewer than 200 purchases (Figure 17). Therefore, a sample of (~60,000) 'reliable customers' who shopped in each of the 42 months was selected to reduce noise from customers with few transactions, due to attrition or forgetting to use their loyalty card.

Figure 17. Distribution of number of customer transactions over the 42-month study period



New features (described in Table 6) were used to construct a univariate time-series based on SDIL weight (total weight [g] of levy drinks [SDIL1 + SDIL2]/customer/month). Time-series clustering was performed on SDIL weight using the KShape algorithm in the python package 'tslearn'. KShape captures time series' shape characters as the measurement and is very efficient for large datasets.[21]

Table 6. Data dictionary for new features

Variable	Description
Average sugar content per 100ml	Sum(bev_sugar)/sum(bev_weight)*100 - By month - By consumer per month
Spend per 100ml	Sum(bev_spend)/ sum(bev_weight)*100 - By month - By consumer per month

Code for time-series analysis; (insert Sijin's work), 'Time series clustering for SDIL weight.py', ' multivariate time series.ipynb' and 'Building the selected_data csv file.ipynb', 'clustering_basket_time_series.py' (uses Python packages; Pandas: 1.01, sklearn: '0.22.1', matplotlib: 3.13, scipy:1.4.1, plotly: 4.14, tslearn: '0.5.1.0'

Spending on SDIL 1 and SDIL 2 drinks	Sum(bev_spending) if sdil in ('SDIL1', 'SDIL2') - By consumer per month (average) - By month (total)
Weight of SDIL 1 and SDIL 2 drinks purchased (SDIL weight)	Sum(bev_weight) if sdil in ('SDIL1', 'SDIL2') - By consumer per month (average) - By month (total)
Weight of sugar	Sum(bev_sugar) - By consumer per month (average) - By month (total)
Weight of drinks	Sum(bev_weight) - By consumer per month (average) - By month (total)

6.2 Results: Time-series clustering

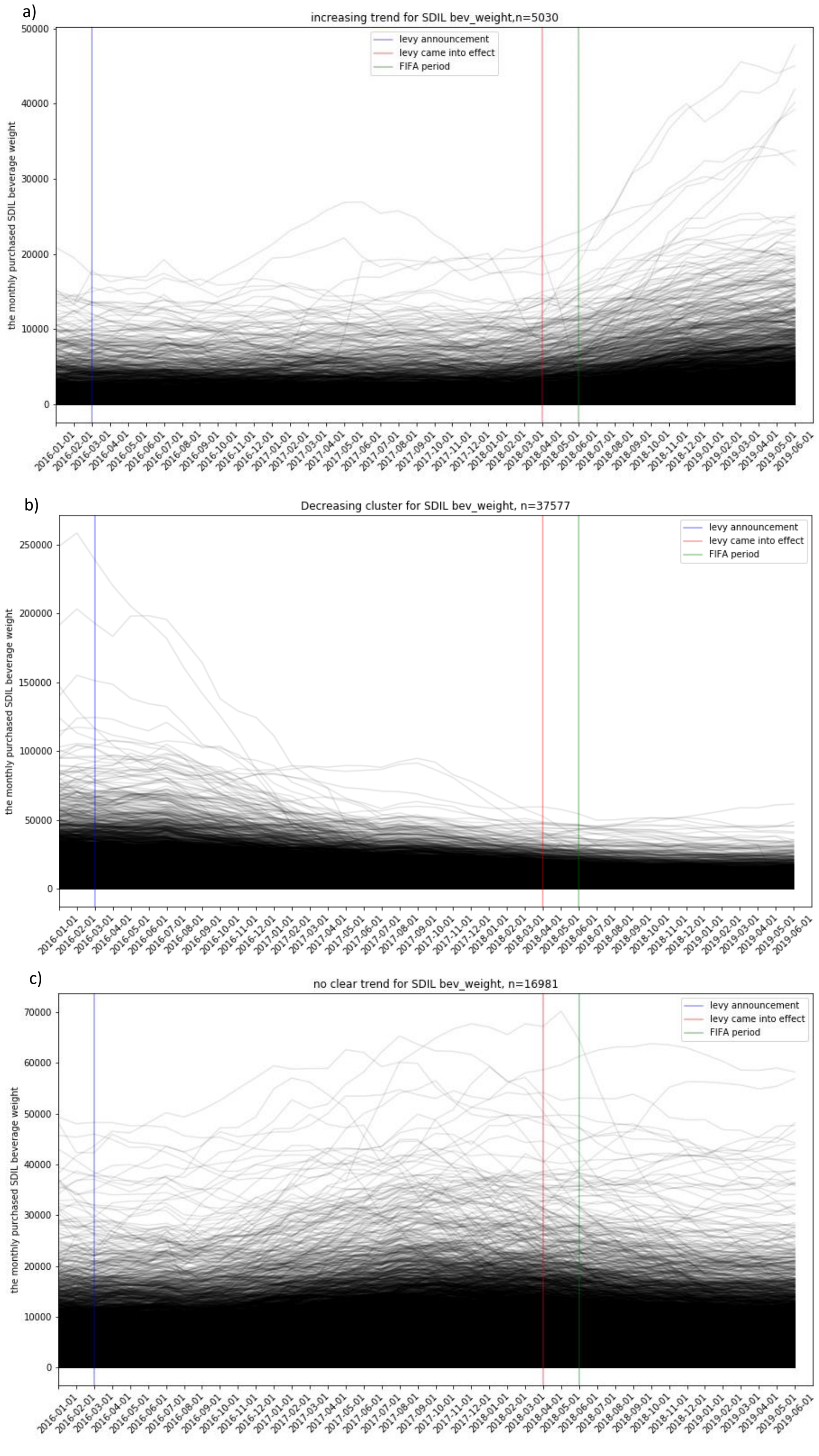
6.2.1 Univariate time-series clustering

Based on total weight of SDIL beverages purchased each month during the 42-month period, the univariate KShape algorithm divided the customer sample into three distinct trend clusters (Figure 18); a) Increasing after levy announcement (N=5,030); b) Decreasing after levy announcement (N=37,577); and c) No clear trend (N=16,981).

Cluster b) is the dominant cluster, with more than 60% of customers in the sample showing a decline in volume of SDIL drinks. There were no clear differences in the customer demographic characteristics across each of these clusters, which led us to hypothesise that univariate time-series clustering, based on volume of SDIL drinks is not sensitive enough to identify customers with similar demographic traits.

Code for time-series analysis; (insert Sijin's work), 'Time series clustering for SDIL weight.py', ' multivariate time series.ipynb' and 'Building the selected_data csv file.ipynb', 'clustering_basket_time_series.py' (uses Python packages; Pandas: 1.01, sklearn: '0.22.1', matplotlib: 3.13, scipy:1.4.1, plotly: 4.14, tslearn: '0.5.1.0'

Figure 18. Time-series clusters resulting from K-Shape algorithm applied to SDIL weight



Code for time-series analysis; (insert Sijin's work), 'Time series clustering for SDIL weight.py', ' multivariate time series.ipynb' and 'Building the selected_data csv file.ipynb', 'clustering_basket_time_series.py' (uses Python packages; Pandas: 1.01, sklearn: '0.22.1', matplotlib: 3.13, scipy:1.4.1, plotly: 4.14, tslearn: '0.5.1.0'

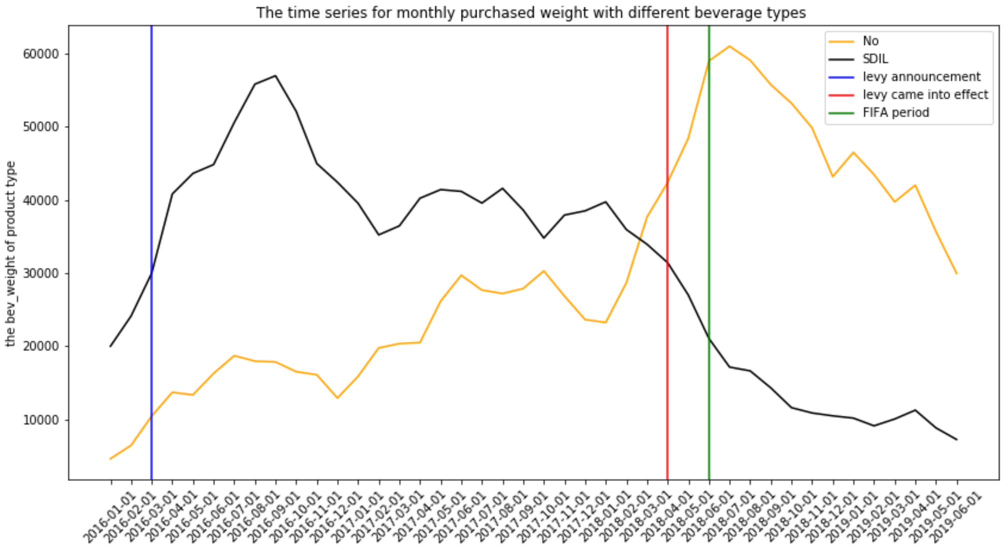
6.2.2 Predictive classifier model

A sample of data labelled with cluster allocations was used to train a predictive classifier. Despite accounting for class imbalance and attempting various classification models (e.g. Decision Trees, XGBoost) we were unable to extract identifying cluster traits. As a result, our classifiers continuously allocated all test samples to the dominant cluster b) ‘decreasing after levy announcement’.

6.2.3 Multivariate time-series classification

While we were unable to operationalise a multi-variate time-series clustering during the DSG time-frame, we report some interesting exploratory findings. In addition to weight of SDIL drinks, we explored time-series trends for non-SDIL drinks, which could reveal switching behaviours. Thus, each customer has two time-series, as seen in the example plot for a single customer in Figure 19. Here we see an example of a customer appearing to switch their purchases of SDIL drinks (black line) for untaxed non-SDIL drinks (yellow line), a trend beginning just before levy implementation.

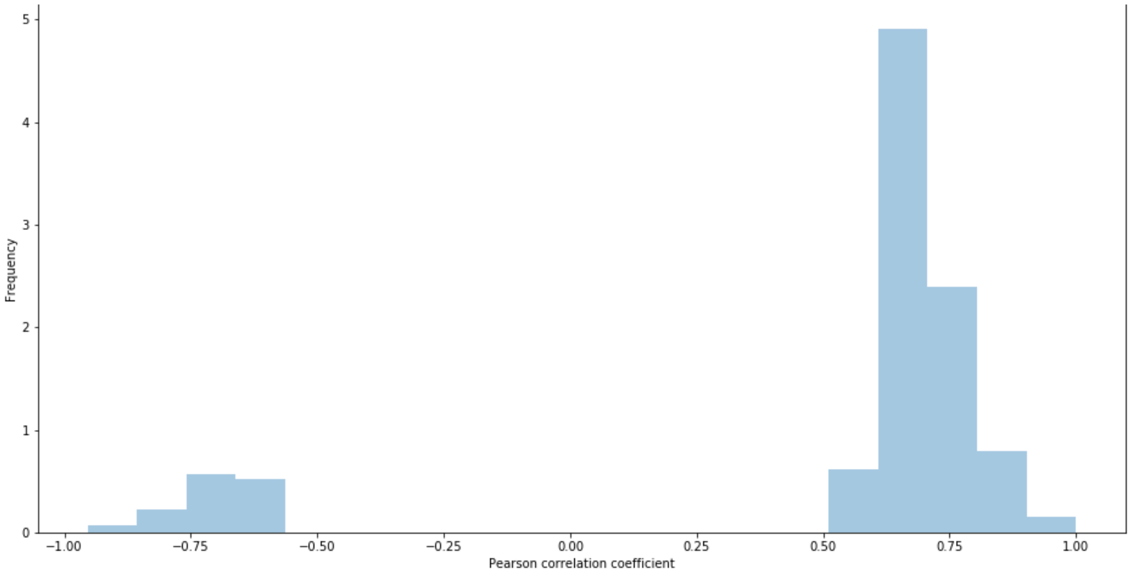
Figure 19. Example of a multivariate time-series plot for one sample customer.



The relationship between the two time-series (SDIL weight, and non-SDIL weight) was quantified by computing Pearson correlation coefficients (r). We considered a relationship between these variables to exist for customers displaying a high correlation (r>0.6) (positive or negative) with a statistical significance at the 95% confidence level. Just over 4% of customers had a high correlation between time-series for SDIL and non-SDIL drinks volumes, with the majority of these positive (Figure 20).

Code for time-series analysis; (insert Sijin’s work), ‘Time series clustering for SDIL weight.py’, ‘ multivariate time series.ipynb’ and ‘Building the selected_data csv file.ipynb’, ‘clustering_basket_time_series.py’ (uses Python packages; Pandas: 1.01, sklearn: ‘0.22.1’, matplotlib: 3.13, scipy:1.4.1, plotly: 4.14, tslearn: ‘0.5.1.0’

Figure 20. Distribution of high correlation coefficients [r]>0.6, p<0.05, N= 2655 customers



6.2.4 Category-level time-series clustering

We explored potential explanations for the unexpected phenomenon that some households increase their purchasing of sugar from soft drinks, using a sample of the first 5,000 customers. The two weeks of missing transaction data in January 2017 was considered by doubling the above computed total volume of beverages per category for that month. We considered the total volume of beverages purchased by customers across different categories and SDIL rating, focusing on three test categories for initial exploration; 'Juice', 'Carbonated ', and 'Non-Carbonated'. A 12 month centered rolling average weight of sugar from beverages purchased within each category-SDIL rating combination was computed for each customer. This was subsequently normalized to have mean 0 and standard deviation 1.

The KShape clustering algorithm was applied to the normalized weights of total sugar from beverages within each short category. This exploratory analysis has not been optimised in terms of number of clusters or included variables, and therefore warrants further exploration. Initial results yielded 5 time-series clusters for each category, visualised in Figure 21 (showing milk and juice) and Figure 22 (showing carbonated and non-carbonated soft drinks). For each drink type, we observe clusters where the total weight of sugar from that category has a net increase, net decrease or remains approximately constant over time (summarised in Table 7).

Code for time-series analysis; (insert Sijin's work), 'Time series clustering for SDIL weight.py', ' multivariate time series.ipynb' and 'Building the selected_data csv file.ipynb', 'clustering_basket_time_series.py' (uses Python packages; Pandas: 1.01, sklearn: '0.22.1', matplotlib: 3.13, scipy:1.4.1, plotly: 4.14, tslearn: '0.5.1.0'

Table 7. Number of customers in each category-based time-series cluster.

		CATEGORY			
		Milk	Juice	Carbonated	Non-carbonated
CLUSTER	1	🟩 9717	🟩 17733	🟧 18130	🟧 12776
	2	🟩 6946	🟩 7040	🟩 8539	🟩 11677
	3	🟥 13309	🟩 12968	🟥 4720	🟧 11473
	4	🟥 11654	🟥 13967	🟩 16698	🟩 11037
	5	🟩 17993	🟥 7911	🟧 11352	🟥 12656

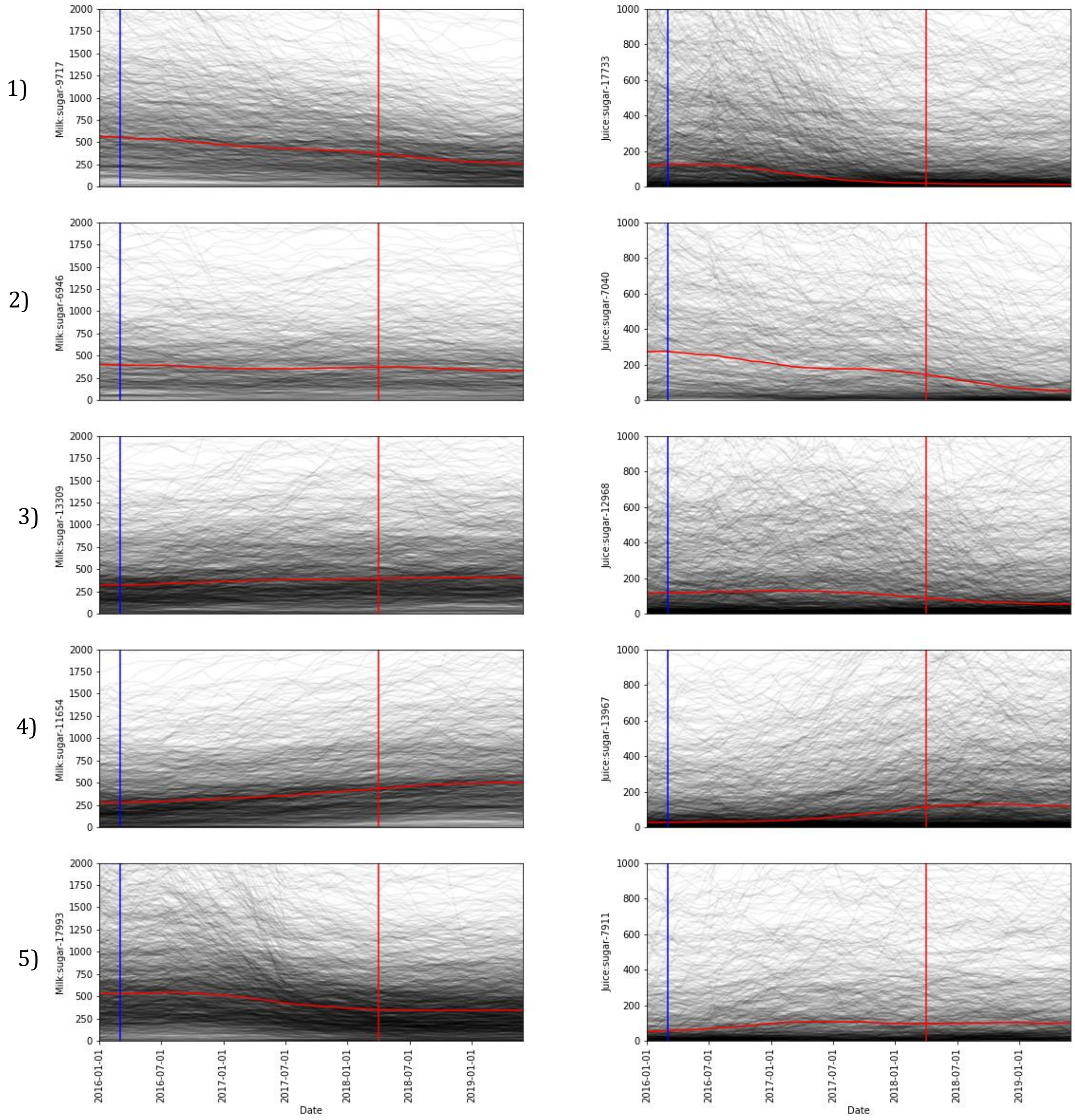
The temporal behaviour of the median is denoted by the colour of the cell: red (increase in sugar), green (decrease in sugar) and orange (sugar remains constant)

The median curves for the time series clusters for milk and juice (Figure 21) are relatively smooth and stable, with some trending upwards (e.g. milk clusters 3 and 4, juice clusters 4 and 5). This suggests that the levy may have resulted in some switching into un-taxed pure fruit juice products containing natural sugars, which may account for some of the overall increase in sugar from drinks among some customers.

Clusters 2 and 4 for carbonated soft drinks (Figure 22) show a clear downward trend in total sugar content either when the levy was implemented (cluster 4) or in the run-up to it being implemented after the announcement (cluster 2). The median of the third cluster displays interesting behaviour, corresponding to sugar content increasing before and decreasing after the levy, which may indicate a ‘stocking up’ behaviour in anticipation of price increases. Similarly for non-carbonated soft drinks, weight of sugar tends to decrease for customers in clusters 2 and 4, whereas the weight increases for cluster 5. We might suppose that this increase in sugar purchased in 2018 arises as these consumers change the types of drinks purchased to non-carbonated drinks.

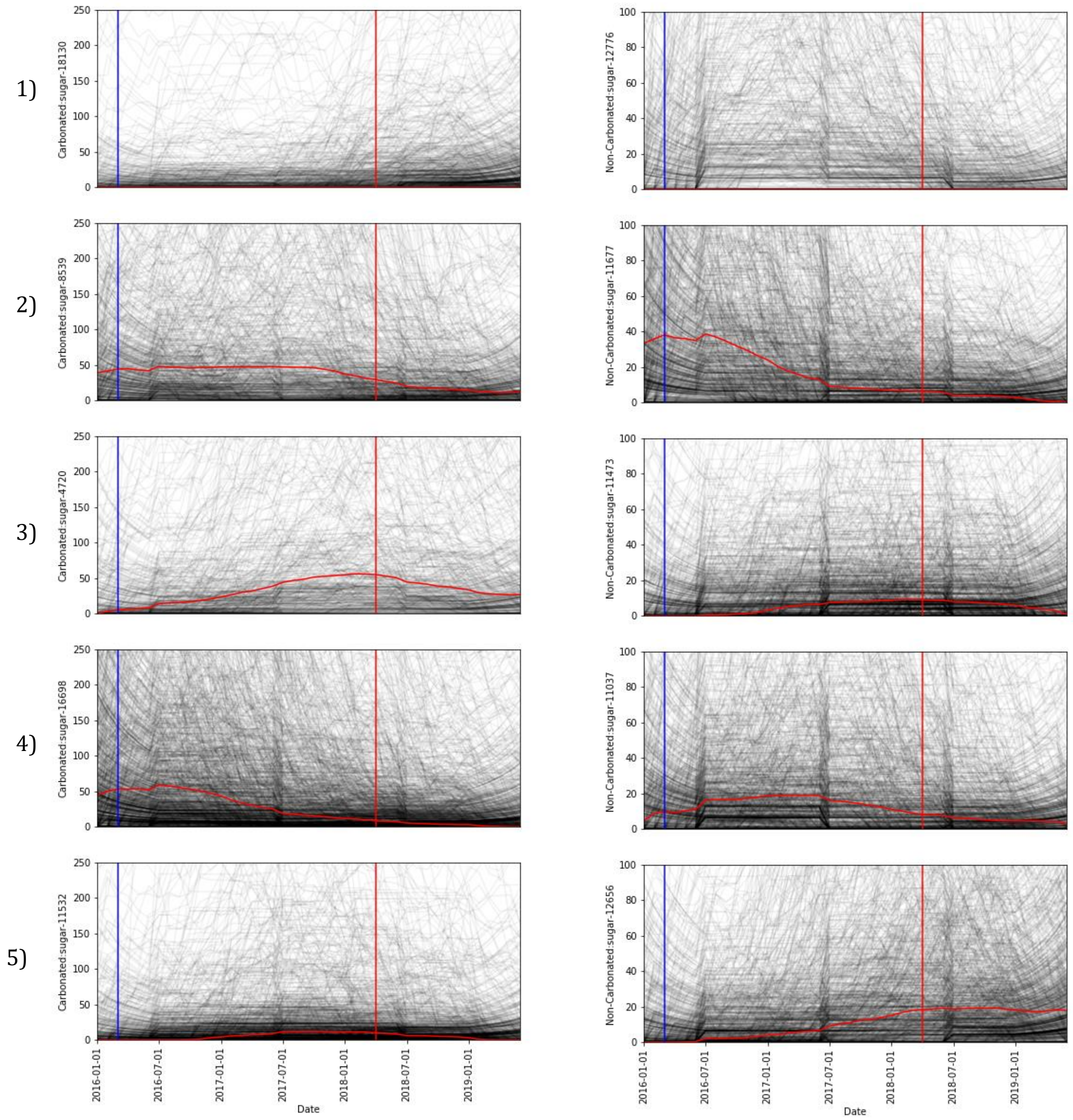
Code for time-series analysis; (insert Sijin’s work), ‘Time series clustering for SDIL weight.py’, ‘ multivariate time series.ipynb’ and ‘Building the selected_data csv file.ipynb’, ‘clustering_basket_time_series.py’ (uses Python packages; Pandas: 1.01, sklearn: ‘0.22.1’, matplotlib: 3.13, scipy:1.4.1, plotly: 4.14, tslearn: ‘0.5.1.0’

Figure 21. Time series clustering on total monthly weight of sugar from milk (left) and juice (right) for the 5 clusters identified. The time series for the first 5000 customers are shown along with the median time series curve (red).



Code for time-series analysis; (insert Sijin’s work), ‘Time series clustering for SDIL weight.py’, ‘ multivariate time series.ipynb’ and ‘Building the selected_data csv file.ipynb’, ‘clustering_basket_time_series.py’ (uses Python packages; Pandas: 1.01, sklearn: ‘0.22.1’, matplotlib: 3.13, scipy:1.4.1, plotly: 4.14, tslearn: ‘0.5.1.0’

Figure 22. Time series clustering on total monthly weight of sugar from carbonated drinks (left) and non-carbonated drinks (right). The time series for the first 5000 customers are shown along with the median time series curve (red).



6.3 Conclusions: Time-series clustering

The majority of customers (60%) decreased sugar from beverages over the period, while around 10% displayed increased sugar from beverages. Analysing trends by category may help to explain this behaviour. Preliminary analysis showed a general decline in sugar from categories most likely affected by the levy (carbonated and non-carbonated soft drinks), while sugar from categories more likely to be exempt (milk and juice) increased for some customers. This suggests switching into naturally sweet untaxed beverages, however, as time-series’ were examined independently, this is not conclusive. Finally, there was some evidence of stocking up on carbonated soft drinks before the levy.

Code for time-series analysis; (insert Sijin’s work), ‘Time series clustering for SDIL weight.py’, ‘ multivariate time series.ipynb’ and ‘Building the selected_data csv file.ipynb’, ‘clustering_basket_time_series.py’ (uses Python packages; Pandas: 1.01, sklearn: ‘0.22.1’, matplotlib: 3.13, scipy:1.4.1, plotly: 4.14, tslearn: ‘0.5.1.0’

Similar to the K-means approach, we did not find any evidence for demographic differences between clusters. It may be that demographic variance is so slight that we did not detect it. We recommend expanding the work on multivariate time-series clustering to see whether inclusion of additional variables may generate more distinct clusters in terms of demographic profiles. In particular, we anticipate that exploring trends across multiple categories in tandem may help us to understand why sugar from beverages increased post-levy for some customers.

Code for time-series analysis; (insert Sijin’s work), ‘Time series clustering for SDIL weight.py’, ‘ multivariate time series.ipynb’ and ‘Building the selected_data csv file.ipynb’, ‘clustering_basket_time_series.py’ (uses Python packages; Pandas: 1.01, sklearn: ‘0.22.1’, matplotlib: 3.13, scipy:1.4.1, plotly: 4.14, tslearn: ‘0.5.1.0’

7. Basket segmentation

In this section, we cluster customers according to the beverages they regularly purchase, to determine whether certain subgroups are likely to undergo any changes. For example, do customers who regularly purchase a large proportion of sugary drinks pre-Levy announcement continue to do so, or do we find that they switch to a lower sugar alternative? This section brings together ideas from both the K-means and time-series clustering approaches described previously.

7.1 Methods: Basket segmentation

To perform this analysis, we clustered customers according to the proportion of beverage volume per subcategory compared to the total at three distinct time points.

7.1.1 K-Means clustering

We split the data into three yearly intervals: post announcement and pre-levy implementation (T1: April 2016 - March 2017 and T2: April 2017 - March 2018) and post-levy implementation (T3: April 2018 - March 2019). These time periods were selected to ascertain typical monthly transactions for each customer and minimize seasonal variation. Data was filtered to only customers who made beverage transactions in at least six months of each year (n = 59,587)

We then introduced monthly totals of spend, weight, kcal and sugar per subcategory-SDIL combination (e.g. Soft Drinks-Cola-SDIL2) and overall beverage transactions. The monthly proportions of spend, weight, kcal and sugar per subcategory-SDIL combination were computed by dividing each subcategory total by the monthly total for overall beverages. The mean of these monthly subcategory proportions was separately taken over all months within the three time intervals. The proportions per subcategory-SDIL combination were subsequently normalized so that each of the total spend, weight, kcal or sugar across all beverage categories sums to 1. Mathematically, this can be expressed as:

$$prop\_cat\_f = \frac{\left(\frac{1}{12}\sum_{month}\frac{\sum_{puchases\ in\ subcategory-SDIL\ in\ month} f}{\sum_{purchases\ in\ month} f}\right)}{\sum_{subcategory-SDIL}\left(\frac{1}{12}\sum_{month}\frac{\sum_{puchases\ in\ subcategory-SDIL\ in\ month} f}{\sum_{purchases\ in\ month} f}\right)},$$

for each customer, where f is one of bev_weight, bev_spend, bev_kcal or bev_sugar. For each customer, we therefore have a single value for each subcategory-SDIL combination (0-1) corresponding to the normalized proportion of monthly sugar purchased from beverages within each category out of the total sugar purchased from all beverages.

Code for basket segmentation analysis - 'preprocessing_clustering_basket.py', 'clustering_basket_time_seire.py', 'kmeans_cluster_validation.py', 'analyse_clusters.py'. Python packages used: Pandas: 1.01, sklearn: '0.22.1', matplotlib: 3.13, scipy:1.4.1, plotly: 4.14.

Since there are 80 subcategory-SDIL categories, the number of features was reduced to improve performance of the clustering algorithm. Firstly, the proportion of sugar content within beverages in the short categories '-', 'Hot beverages', 'Milk', 'Milk Drinks', 'Squash' and 'Juice', are excluded, leaving beverages within the short categories: 'Water', 'Flavoured Water', 'Non-Carbonated', 'Carbonated' and 'Energy', within the dataset. This minimises the influence of drinks that are out of scope for SDIL, including milk (making up over 29% of monthly beverage transactions for the customer sample), and high sugar drinks like juice (making up over 17% of monthly transactions), where it was not clear from the provided data whether sugars are natural or added. This reduced the number of features to 29 subcategory-SDIL combinations, upon which a principal component analysis was performed for the first time period (Apr. 2016-Mar. 2017). The first 7 principal components account for approximately 71% of the total variance (Table 8).

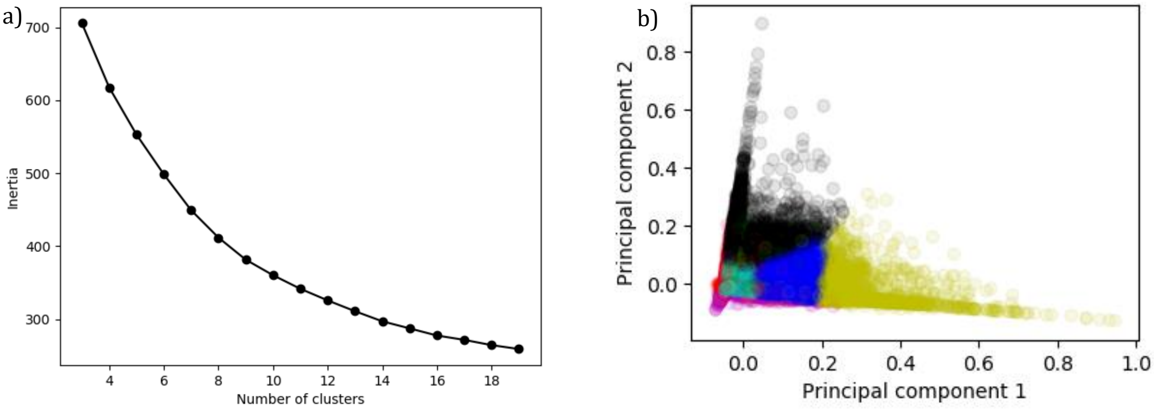
Table 8. Explained variance ratio and cumulative values for the first 10 principle components of the monthly proportion of sugar purchased per selected subcategory-SDIL combinations per customer

PRINCIPAL COMPONENT	1	2	3	4	5	6	7	8	9	10
EXPLAINED VARIANCE RATIO	0.256	0.117	0.093	0.075	0.061	0.056	0.050	0.049	0.043	0.040
CUMULATIVE EXPLAINED VARIANCE RATIO	0.256	0.373	0.466	0.542	0.603	0.658	0.708	0.760	0.800	0.841

The data were transformed to the first 7 principal components and a K-means clustering was applied on the reduced feature set. The optimal number of clusters was determined to be 9, using the elbow method (Figure 23(a)) as the inertia (sum of squared distances of points within each cluster to the centroid) starts to plateau with increasing number of clusters. Figure 23(b) presents the clustering of the T1 data in the first two principal components using 9 clusters. This clustering has a silhouette score of 0.417 which is fairly low owing to the overlap that we can see between clusters so could be improved by incorporating further independent variables into the K-Means algorithm.

Code for basket segmentation analysis - 'preprocessing_clustering_basket.py', 'clustering_basket_time_seire.py', 'kmeans_cluster_validation.py', 'analyse_clusters.py'. Python packages used: Pandas: 1.01, sklearn: '0.22.1', matplotlib: 3.13, scipy:1.4.1, plotly: 4.14.

Figure 23. (a) Elbow plot using inertia to determine the optimal number of clusters. (b) Clusters in relation to the first two principal components found using K-Means clustering. The 9 clusters are represented using different colours.



The corresponding segmentation for customers in time periods Apr. 2017-Mar. 2018 (T2) and Apr. 2018-Mar. 2019 (T3) were obtained, by first transforming the proportions of sugar content per subcategory-SDIL combination per customer onto the principal components found in the first time period (T1; Apr. 2016-Mar. 2017). Customers were then clustered using the centroids of the clusters in the first time period. This separates customers in the time periods T2 and T3 into 9 clusters, which have silhouette scores of 0.415 and 0.535 respectively.

7.2 Results: Basket segmentation

Table 9 presents summary statistics of the nine customer clusters. With the exception of cluster 1, where customers tended to purchase low proportions of total beverage sugar from soft drinks, which will be referred to as 'Low sugar soft drinks', and cluster 5, where customers purchase sugary cola (SDIL2) among a variety of other soft drinks, which will be referred to as 'Cola plus', other clusters will be referred to by the soft drink which provides the highest proportion of sugar within each cluster.

Code for basket segmentation analysis - 'preprocessing_clustering_basket.py', 'clustering_basket_time_seire.py', 'kmeans_cluster_validation.py', 'analyse_clusters.py'. Python packages used: Pandas: 1.01, sklearn: '0.22.1', matplotlib: 3.13, scipy:1.4.1, plotly: 4.14.

Table 9. Descriptions and statistics of the 9 groups of customers. The clusters are ordered in increasing proportion of sugar from soft drinks

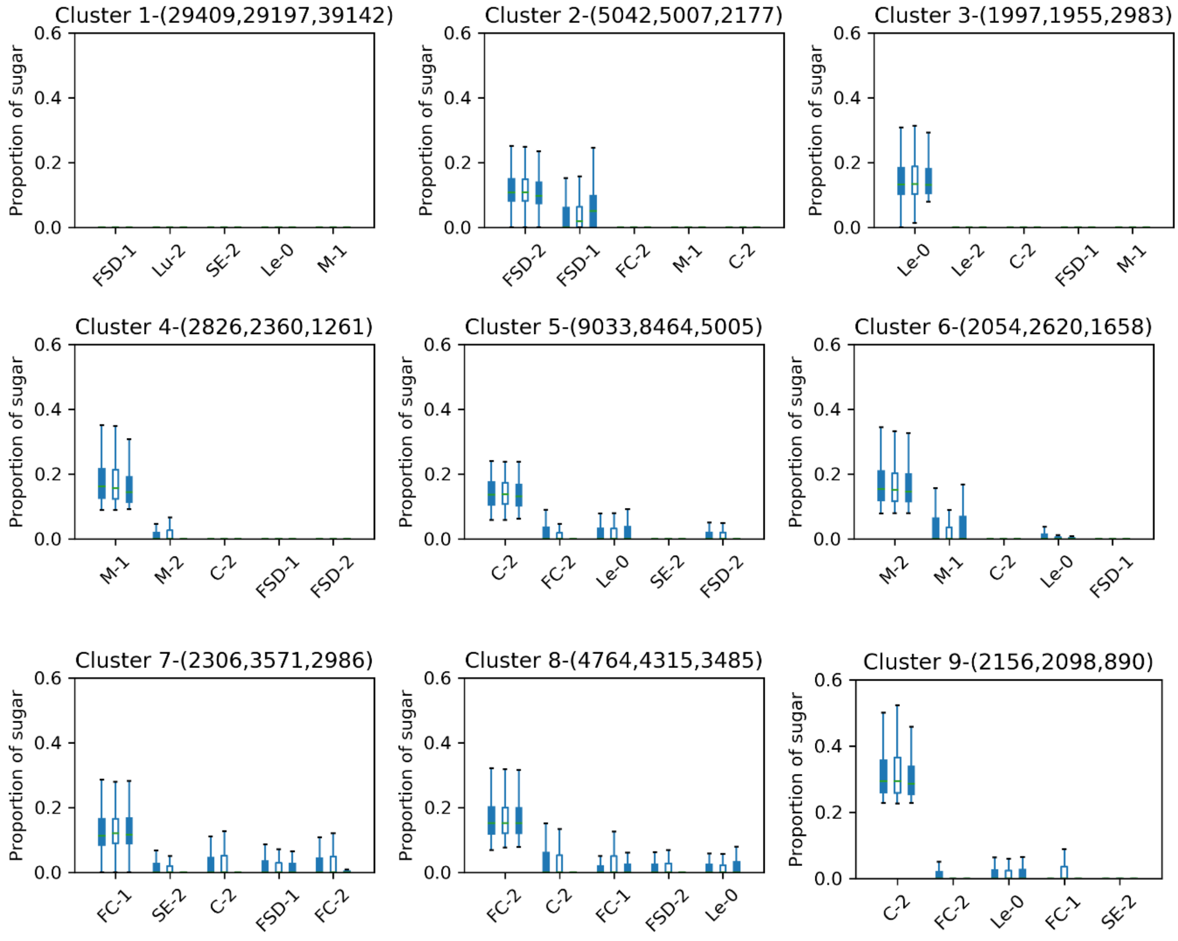
Cluster		*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Preferred Drink		Low sugar soft drinks	Fancy soft drinks (2)	Lemonade (0)	Mixers (1)	Cola (2) and other soft drinks	Mixers (2)	Flavoured Carbs (1)	Flavoured Carbs (2)	Cola (2)
Number in cluster (%)	**2016**	29409 (49.4%)	5042 (8.46%)	1997 (3.35%)	2826 (4.74%)	9033 (1.52%)	2054 (3.87%)	2306 (3.87%)	4764 (8.00%)	2156 (3.62%)
	2017	29197 (49.0%)	5007 (8.40%)	1955 (3.28%)	2360 (3.96%)	8464 (1.42%)	2620 (4.40%)	3571 (5.99%)	4315 (7.24%)	2098 (3.52%)
	2018	39142 (65.7%)	2177 (3.65%)	2983 (5.01%)	1261 (2.12%)	5005 (8.40%)	1658 (2.78%)	2986 (5.01%)	3485 (5.85%)	890 (1.49%)
Proportion of sugar from soft drinks (Mean ±STD)	**2016**	0.088 ± 0.104	0.264 ± 0.126	0.295 ± 0.155	0.313 ± 0.147	0.320 ± 0.133	0.328 ± 0.155	0.356 ± 0.155	0.360 ± 0.148	0.484 ± 0.155
	2017	0.086 ± 0.102	0.269 ± 0.131	0.293 ± 0.160	0.338 ± 0.157	0.320 ± 0.130	0.322 ± 0.158	0.358 ± 0.160	0.364 ± 0.151	0.490 ± 0.158
	2018	0.100± 0.113	0.300 ± 0.138	0.284 ± 0.153	0.306 ± 0.145	0.313 ± 0.134	0.324 ± 0.152	0.367 ± 0.160	0.367 ± 0.148	0.467 ± 0.157
Mean total spend on soft drinks	**2016**	7.08	9.20	6.77	7.67	10.76	9.00	14.59	11.34	18.99
	2017	6.44	8.29	5.93	7.38	9.37	8.65	13.03	10.39	17.84
	2018	6.67	8.64	6.49	8.96	10.49	9.80	13.28	11.60	15.61
Mean total sugar from soft drinks	**2016**	73	172	273	257	392	276	347	394	1222
	2017	61	151	216	223	327	235	325	336	1101
	2018	59	134	184	107	246	195	277	283	653
Mean total volume of soft drinks	**2016**	9272	9962	10873	10628	12393	9934	15278	13835	20843
	2017	8207	8474	9057	9516	10548	8996	14510	12685	19338
	2018	8063	7926	9252	7787	10310	8213	14296	13042	14097
Mean total kcal on soft drinks	**2016**	379	789	1217	1136	1664	1211	1573	1698	4900
	2017	323	697	970	973	1398	1045	1476	1464	4431
	2018	331	638	867	530	1102	898	1317	1291	2653

Numbers in brackets in the cluster names refer to the SDIL category of the predominant drink category; 0 = non-SDIL, 1 = SDIL1, 2 = SDIL2

Code for basket segmentation analysis - ‘preprocessing_clustering_basket.py’, ‘clustering_basket_time_seire.py’, ‘kmeans_cluster_validation.py’, ‘analyse_clusters.py’. Python packages used: Pandas: 1.01, sklearn: ‘0.22.1’, matplotlib: 3.13, scipy:1.4.1, plotly: 4.14.

Figure 24 highlights trends within the clusters. The range of proportions of sugar from the dominant drink remain fairly similar across the years, which indicates this is likely to be the dominant feature of the clustering. This could be checked using a feature selection algorithm.

Figure 24. Boxplots for T1 (left), T2 (middle) and T3 (right) of the proportion of sugar from the five soft drink categories which contribute the highest average sugar proportion for customers in each cluster.



Soft drinks codes: FC = Flavoured Carbs, SE = Sports and Energy, C = Cola, FSD = Fancier Soft Drinks, Le = Lemonade, M = Mixers, Lu = Lunchbox. Numbers represent SDIL classification; 0 = non-SDIL, 1 = SDIL1, 2 = SDIL2. Number of customers in each cluster in sub-plot titles.

7.2.1 Cluster insights

Cluster 1 contains a large number of customers (~50%, rising to 66% post-levy) who obtain a very small proportion of total beverage sugar and kcal from soft drinks (whiskers not visible).

Cluster 2 contains the second largest number of customers (~8%) and is distinguished by a high proportion of sugary Fancy soft drinks (FSDs). These customers are sensitive to the levy, reducing the proportion of SDIL2 FSDs, replacing them with more SDIL1 FSDs in T3.

Code for basket segmentation analysis - ‘preprocessing_clustering_basket.py’, ‘clustering_basket_time_seire.py’, ‘kmeans_cluster_validation.py’, ‘analyse_clusters.py’. Python packages used: Pandas: 1.01, sklearn: ‘0.22.1’, matplotlib: 3.13, scipy:1.4.1, plotly: 4.14.

Cluster 3 initially contained 3% of customers, rising to 5% post levy. Customers in this cluster tend to purchase low sugar lemonade and show little change across time points.

Cluster 4 contains customers who predominantly purchase high sugar mixers (SDIL1), with a small proportion of very high sugar mixers (SDIL2). Over the three years, the upper quartile and upper whisker of the proportion of sugar from SDIL1 mixers decreased, as the proportion of SDIL2 mixers purchased increased. This cluster underwent the largest decrease in mean total sugar from soft drinks (58.4%) as a result of the levy.

Cluster 5 contains customers who obtain between 5-25% of the total beverage sugar from very sugary cola (SDIL2), amongst other soft drinks. Pre-levy, this cluster as small, but grew to 8.4% of regular customers post-levy. The proportion of sugar from SDIL2 Flavoured Carbonated drinks (FCs) and FSDs decreased with time, while the proportion of sugar from low sugar lemonade increased.

Cluster 6 contains customers who buy very high sugar mixers (SDIL2) and a smaller proportion of high sugar (SDIL1) mixers. The proportion of SDIL2 mixers decreased between T1 and T3, while proportion of SDIL1 mixers increased. Purchasing of low sugar lemonade also decreased.

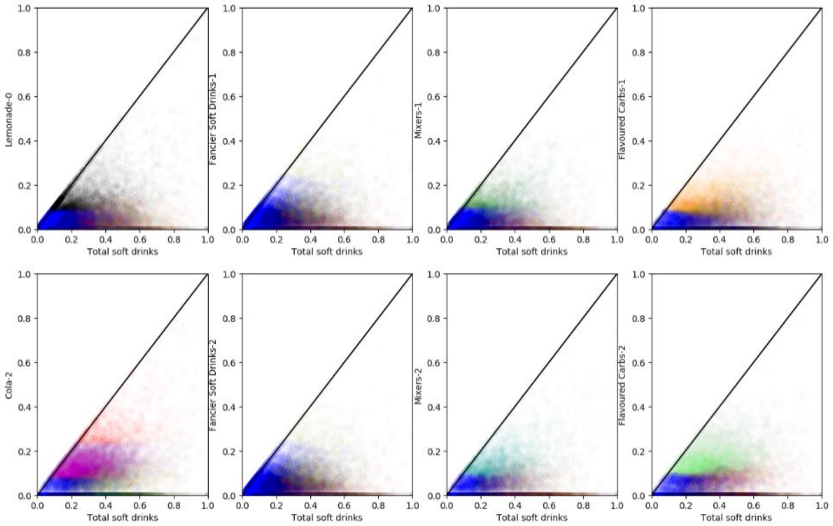
Cluster 7 customers purchased the second highest soft drinks volume, across a variety of types, mostly high sugar FCs (SDIL1). While the distribution of FCs remains similar over time, the proportion of sugar from other drinks (sports and energy (SDIL2), FCs (SDIL2), cola (SDIL2) and FSDs (SDIL1)) decrease. The number of customers in this cluster increased from 3.87% to 5.01%. The reduction in SDIL2 FCs decreased most markedly post-implementation, indicating that price may have had a greater impact than reformulation for these customers.

Cluster 8 customers purchase a similar range of soft drinks as cluster 7 customers, with very high sugar FCs (SDIL2) contributing to the highest proportion of beverage sugar. The proportion of sugar from SDIL2 drinks decreases, while that from low sugar lemonade increases, suggesting switching. The percentage of customers decreased from 8% to 5.85%.

Cluster 9 consumers obtained a high proportion of beverage sugar from very sugary cola. This group was most affected by the levy, as mean monthly total sugar reduced by 47.6% and kcal by 46% after the levy. Additionally, the upper quartile and upper whisker on the boxplot for cola SDIL2 move down. Customers within this group spent the most per 100ml of soft drink.

Figure 25 shows scatter plots for the proportion of beverage sugar by soft drink category against total beverage sugar. Customers on the diagonal line purchase all of their beverage sugar from that category, whereas customers on the horizontal axis do not purchase that type of soft drink. The separation between coloured clusters indicates clusters distinguish well customers who buy large amounts of a particular soft drink type, while those who purchase a broad range of soft drinks are less well distinguished. Consumers who infrequently purchase sugary soft drinks are located in the bottom left corner of each subplot.

Figure 25. Scatter plots showing % sugar from selected soft drink categories at T1.



Clusters are colour coded: 1-blue;2-yellow,3-black,4-green,5-magenta,6-cyan,7-orange,8-lightgreen,9-red.

Figure 26. Sankey diagram showing how customers within the 9 clusters change behaviour over the three time periods. The colour of the nodes indicate the SDIL rating of the preferred drink within each cluster: green (0 or low sugar), orange (SDIL1) and red (SDIL2).

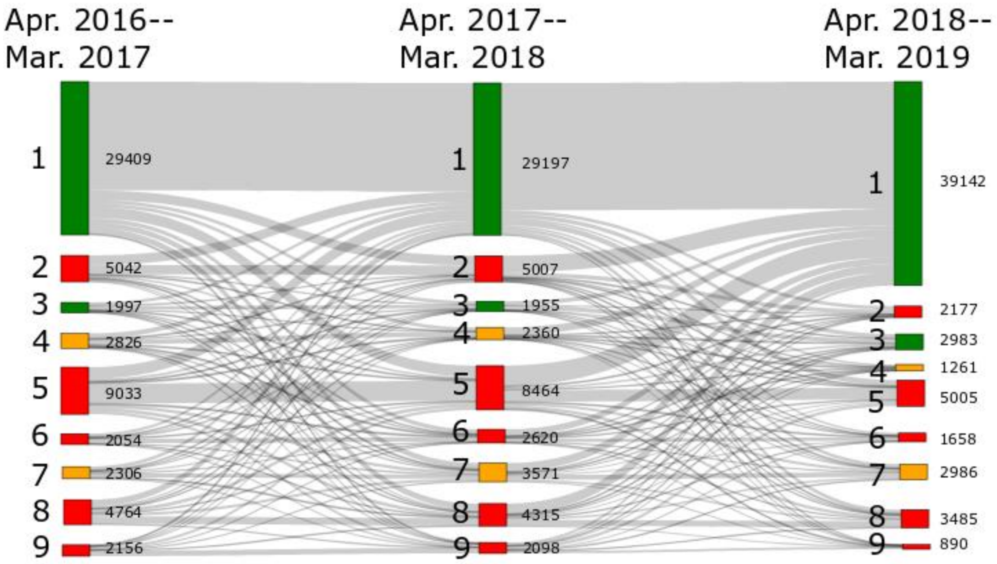


Figure 26 illustrates transitions between clusters, enumerated in Table 10.

Code for basket segmentation analysis - ‘preprocessing_clustering_basket.py’, ‘clustering_basket_time_seire.py’, ‘kmeans_cluster_validation.py’, ‘analyse_clusters.py’. Python packages used: Pandas: 1.01, sklearn: ‘0.22.1’, matplotlib: 3.13, scipy:1.4.1, plotly: 4.14.

Table 10. Cross tabulation of the number of customers transitioning between clusters between time points.

		Apr. 2017---Mar. 2018									*% Out*	
	Cluster	1	2	3	4	5	6	7	8	9	+	-
Apr. 2016---Mar. 2017 (T1 – T2)	1	**20872**	1859	620	691	2101	822	1036	1189	219	29.0	0.00
	2	1858	**1765**	98	154	403	171	239	325	29	28.1	36.9
	3	597	100	**696**	64	207	71	118	109	35	30.2	34.9
	4	887	187	79	**922**	234	257	90	141	29	26.6	40.8
	5	2262	445	207	214	**3849**	209	631	631	585	22.8	34.6
	6	558	135	36	111	152	**838**	83	115	26	10.9	48.3
	7	633	165	57	41	322	62	**769**	214	43	11.1	55.5
	8	1290	310	124	125	662	140	530	**1487**	96	2.02	66.8
	9	240	41	38	38	534	50	75	104	**1036**	0.00	51.9
% In	+	28.5	27.6	27.7	22.4	19.7	9.62	16.9	2.41	0.00		
	-	0.00	37.1	36.7	38.5	34.8	58.4	61.5	63.1	50.6		

		Apr. 2018---Mar. 2019									*% Out*	
	Cluster	1	2	3	4	5	6	7	8	9	+	-
Apr. 2017---Mar. 2018 (T2 – T3)	1	**24451**	673	767	414	1040	402	634	723	93	16.3	0.00
	2	3168	**819**	161	118	190	107	196	224	24	20.4	63.3
	3	797	42	**799**	31	92	23	73	81	17	16.2	42.9
	4	1473	86	164	**236**	126	95	59	105	16	17.0	73.0
	5	3869	223	415	116	**2241**	221	555	610	214	18.9	54.6
	6	1241	74	154	197	128	**643**	64	99	20	6.98	68.5
	7	1622	89	204	53	277	42	**898**	361	25	10.8	64.0
	8	1947	142	202	68	308	94	380	**1155**	19	0.44	72.8
	9	574	29	117	28	603	31	127	127	**462**	0.00	78.0
% In	+	37.5	31.5	42.1	36.6	26.3	10.1	17.0	3.64	0.00		
	-	0.00	30.9	31.1	44.6	28.9	51.1	52.9	63.2	48.1		

Bold = customers remaining within the same cluster. Dark grey shading = transition for each cluster with the largest customer number. Red shading = the number of customers moving towards 'less healthy' clusters while green shading = proportion moving to a 'healthier' cluster (according to % total sugar). The bottom two rows indicate the percentage of customers that move into the cluster from healthier (red) and less healthy (green) clusters.

Between the first two time periods, in all clusters except for 2 (FCD-2), most customers stayed in the same cluster. In cluster 2, slightly more customers move into cluster 1 (low sugar from soft drinks) (n=1858) than stay in cluster 2 (n=1765). Each cluster has a considerable number of customers moving into a different cluster between T1 and T2, though the effect on cluster size for clusters 2, 3 and 9 is minimised by a similar number of customers move into the cluster. Clusters 1 (low sugar from soft drinks), 4 (Mixers 1), 5 (Cola 2-plus) and 8 (Flavoured Carbs 2) decrease in size, whereas clusters 6 (Mixers 2) and 7 (Flavoured carbonated drinks 1) increase between T1 and T2. More customers are found to change cluster assignment between T2 and T3 (n=27883) than between the T1 and T2 (n=27353). The main transition is towards the low sugar cluster 1; this is the dominant transition out for all clusters, bar cluster 3 (lemonade-0) and cluster 9 (cola-2). For cluster 3, a similar (but slightly higher) proportion of customers stay in the same cluster as move to cluster 1, while for cluster 9, the dominant transition is to cluster 5 (smaller % of sugar from SDIL2 cola).

Code for basket segmentation analysis - 'preprocessing_clustering_basket.py', 'clustering_basket_time_seire.py', 'kmeans_cluster_validation.py', 'analyse_clusters.py'. Python packages used: Pandas: 1.01, sklearn: '0.22.1', matplotlib: 3.13, scipy:1.4.1, plotly: 4.14.

7.3 Conclusions: Basket segmentation

By segmenting customers according to the category of soft drink that they tend to purchase, we gained insight into how customers changed their purchasing habits as a result of the levy. In particular, we saw an increase in the number of customers assigned to the ‘healthier’ groups: clusters 1 (low sugar consumption from soft drinks) and 3 (low sugar lemonade), and a decrease in the number of customers within the ‘less healthy’ groups, where the preferred soft drink was eligible for the levy. The trend towards a low proportion of sugar from soft drinks was found across all customer groups. However, the effect appeared smaller for customers who originally purchased a large proportion of beverage sugar from high sugar cola. The method also allows identification of customers who did not change their purchase habits, those who change their drink preferences, or who switch to lower sugar versions of similar drinks, for example, between T2 and T3, 197 customers move from cluster 6 (Mixers SDIL2) to cluster 4 (Mixers –SDIL1)

A note of caution in this analysis is that since the weight of beverage sugar was used as the feature to cluster the customers with, we were unable to explore whether customers transitioned to zero sugar alternatives or stopped purchasing soft drinks altogether. The clustering method should be further validated, with greater consideration of outliers, which may have resulted in unexpected transitions towards a less healthy cluster.

There are a number of ways in which this work can be extended to gain further insight into the customer behaviour and more formal statistical methods should be applied to analyse the significance of the changes observed. Firstly, we could incorporate geodemographic information about the customers to see which were more likely to change their purchasing patterns as a result of the levy or not. Secondly, we could explore traits of customers who made the same transitions, particularly after the levy had been implemented. Thirdly, we could consider the changes in customer behaviour in a probabilistic way. For example, are customer trends independent of availability or price of products of each type (in which case, customers may be making a conscious choice to stick with same product), or are these trends dependent?

Code for basket segmentation analysis - ‘preprocessing_clustering_basket.py’, ‘clustering_basket_time_seire.py’, ‘kmeans_cluster_validation.py’, ‘analyse_clusters.py’. Python packages used: Pandas: 1.01, sklearn: ‘0.22.1’, matplotlib: 3.13, scipy:1.4.1, plotly: 4.14.

8. Multiple Linear Regression

How people were affected by the introduction of the SDIL is likely dependent upon the type of drinks they choose. We expect that customers who purchase the highest sugar drinks are likely to be most affected by price increases. One hypothesis is that people purchasing the highest sugar drinks are likely to benefit most from the levy, as the tax will have the biggest hit, while another hypothesis suggests that the sugariest drink purchasers are the least likely to change their behaviour. This may be because they like the taste and are unwilling to switch to products containing artificial sweeteners, for example.

Here, we used Multiple Linear Regression (MLR) to identify predictors of mean sugar density (g/100ml) for purchased products, at the customer level. Two MLR models were performed to in order to identify factors that predict the amount of total sugar per 100 mL of purchased products within the SDIL1 and SDIL2 levy brackets separately.

8.1 Methods: Linear regression

The dataset used contained 43,436,861 rows and 35 variables, and was split according to levels of SDIL (2,848,752 cases for SDIL1 and 8,053,707 cases for SDIL2). Using item sugar density (g/100ml) as the dependent variable, the same predictive model was generated for both the SDIL1 and SDIL2 levy drinks independently. Data for the whole 42-month time period was used. We hypothesised that demographic and area characteristics are likely to influence sugar density of chosen drinks, and controlled for drink type. Thus, predictive variables for the model were; beverage sub-category (new.subcat), customer age band, gender, output area of residence subgroup, index of multiple deprivation (IMD), local authority district (LAD11NM) and rural/urban classification (RUC11). The regression formula is as follows. Models were run after one-hot encoding.

lm(formula = item_sugar ~ imd + new.subcat + Gender + AgeBand +

LAD11NM + Subgroup.Name + RUC11, data = SDIL1)

8.2 Results: Linear regression

Due to the categorical nature of input variables, the regression model outputs composed of 131 coefficients excluding the intercept. For ease of interpretation, only the intercept and those which reached statistical significance at the 95% confidence level are presented in Table 11 and Table 12.

Code for Multiple Linear Regression – 'Multiple_Linear_Regression_commented_code.R'

Table 11. Statistically significant coefficients of regression for sugar density of SDIL1 beverages

	coefficients	p-value
Intercept	6.09300	
Age 0-16 yo	0.05372	0.01728
Age 17-29 yo	0.04033	0.00349
Age 30-44 yo	0.04616	0.00329
Age 45-64 yo	0.02936	0.00325
Age 65 yo +	-0.01202	0.00329
Lemonade	1.77600	0.03266
Fruit juice	1.67600	0.00898
Chilled juice	1.55200	0.00895
Milk	1.28600	0.00941
Cola	1.15500	0.00916

Coefficients represent sugar density (/100ml) of chosen drink. Coefficients presented are statistically significant at the 95% confidence level. Adjusted R-squared = 0.520, Residual standard error = 0.5199

For SDIL1 beverages 52% of the variation is explained by the model (Table 11) (p < 0.001) and the residual standard error is equal to 0.52. Sugar density of chosen drink decreased with age. Controlling for all other factors, customers aged 0-16 years purchased beverages containing around 0.05g more sugar per 100ml, while the eldest customers (65 years+) purchased beverages containing around 0.01g of sugar less per 100ml, vs the other age groups (Table 11). Sugar density of chosen SDIL1 drinks was also affected by the category; choosing lemonade increased sugar density by around 1.78g/100ml (Table 11). No other demographic variables influenced sugar density of chosen SDIL1 drinks.

For SDIL2 drinks, age was a less convincing predictor of sugar density. Only the 30-44 year age band yielded a significant result, yet the increase in sugar density chosen by this group was very small (just 0.0002g/100ml) and unlikely to be meaningful (Table 12). Unlike for SDIL1 beverages, some gender differences appear to exist for sugar density of chosen SDIL2 drinks, with males choosing drinks on average 0.07g/100ml higher sugar drinks than females and customers with unreported genders (Table 12). Unlike for SDIL1 drinks, IMD was a statistically significant negative predictor of sugar density for SDIL2 drinks (Table 12), that is as IMD decile increases (indicating increasing affluence), the sugar density of chosen beverages decreases. This is the expected direction of effect, yet the effect size is so small it is unlikely to be meaningful, just a 0.000002g decrease in sugar/100ml for every one point increase in the IMD decile. The R-squared value suggests just 39.5% (Table 12) of the variation in sugar density of chosen SDIL2 drinks is explained by the model, indicating a lower model performance.

Code for Multiple Linear Regression – 'Multiple_Linear_Regression_commented_code.R'

Table 12. Statistically significant coefficients of regression for sugar density of SDIL2 beverages

	coefficients	p-value
Intercept	13.21000	
Age 30-44 yo	0.00023	0.01253
Females	0.03860	0.01242
Males	0.07050	0.01253
Coffee machine pods	15.91000	0.01826
IMD	-0.00000	<0.0001

Coefficients presented are statistically significant at the 95% confidence level. Adjusted R-squared = 0.395, Residual standard error = 3.1940

8.3 Conclusions: Linear regression

This investigation suggests there are some differences in the influences of sugar content of chosen drinks within the SDIL1 and SDIL2 tax bands. Age may have more of a bearing on the sugariness of chosen SDIL1 drinks; younger people favour higher sugar beverages. Gender appears to be more important for choice of SDIL2 drink; men purchase higher sugar drinks. Choice of drink category influences sugar content, with lemonade purchasers buying the highest sugar SDIL1 drinks. Area-level demographics appeared to have little influence on the sugar content of chosen drinks, except for IMD; living in a more deprived area may slightly increase sugar density of chosen SDIL2 drink.

The sugar content of chosen SDIL1 and SDIL2 drinks cannot be fully explained by the simple demographic, area and beverage category variables included in the models. Further work could explore other variables to improve model performance. For example, the lack of relationship between sugar content and area-level demographics may suggest a greater importance of individual factors. IMD insights may suggest that customer income, employment, and education may influence the sugar content of chosen drinks. Factors such as household composition should also be considered. We might hypothesise for example, controlling for age and gender, that a single young professional would make different beverage choices compared with a parent of young children. It is also worth considering the influence of point of choice factors such as time of day, season/temperature, and consumption occasion, as well as the influence of beverage purchase volume and purchases of food items.

Finally, this analysis was performed across the whole dataset. Future analysis could look to repeat the regression at different time points to understand whether influences of sugar content of purchases drinks change over time as a result of the introduction of the levy.

Code for Multiple Linear Regression – 'Multiple_Linear_Regression_commented_code.R'

9. Multivariate Time-series Regression

Before the time series regression, simple trends were plotted to understand patterns in consumer behaviour/ product composition. We hypothesised that the SDIL could affect the amount of purchased sugar from beverages via three key mechanisms:

1) **Price** – we expect higher prices per/100ml to lead to lower purchased sugar from drinks
2) **Reformulation** - we expect customers who buy more reformulated products to buy less sugar
3) **Product choice** – we expect customers who choose fewer levy-eligible high sugar products to purchase less sugar

A multivariate time series regression was then performed to explore consumption trends and the interaction of different demographic and behavioural drivers on purchased sugar from beverages.

9.1 Methods: Multivariate time-series regression

The dataset follows a cohort of cardholders, sampled in 2016. It is possible that any observed decrease in purchased beverages/sugar could be due to cardholder attrition (reduced shopping at Sainsbury's over the period), rather than a true reflection of changing diet preferences. Only cardholders who purchased at Sainsbury's during at least 6 months in a year throughout the period were retained. Two customers with the same ID but different demographic data were excluded.

Data was aggregated at the Local Authority District (LAD) – Middle Super Output Area (MSOA) – Output Area Cluster (OAC) Super Group level. Combinations with less than 20 cardholders were eliminated to prevent model results from being based on the eccentricity of a few cardholders. Five broad categories of drinks were removed; Baby Food (out of scope for SDIL), Milk (plain milk is unlikely to contain added sugar, note that milk drinks and flavoured milk are retained), Coffee, Tea and Concentrates (coffee, tea and concentrates, we cannot be sure whether sugar content from the back of pack nutrient panel is for the product as sold, or as consumed). This left 907 unique combinations and 91,375 rows of data where each observation is a product-month level activity.

Feature engineering

Five customer-level variables were engineered, all of which were standardised before the regression.

total_sugar, sums the beverage sugar column (representing total sugar purchased over the period from included beverages) for each LAD-MSOA-OAC combination and is the dependent variable.

Avg_sdil_spend measures the mean expense (GBP£) on drinks per 100ml. This variable captures changes in the purchase price of drinks. It is non-inflation-adjusted but we do not expect inflation to impact the results drastically, as inflation has been low during the data period (0-3%).

Reform_prop measures the proportion of reformulated products in the consumer's total beverage purchase (reformulated products defined earlier), a proxy for reformulation effect.

Sdil_weight records the total weight of SDIL-eligible drinks (containing ≥5g sugar/100ml).

Sdil2_weight_prop measures the quantity of SDIL 2 applicable drinks as a proportion of total drinks purchased. Sdil_weight and Sdil2_weight examine the relative strength of quantity effect (buying less soft drinks) vs. substitution effect (buying less soft drinks and substituting with less sugary alternative).

The model

We used a dynamic regression model for estimation. The model is essentially a linear combination of predictor variables, where the error series is assumed to follow an ARIMA (Auto-Regressive Integrated Moving Average) model.[22]

$$Total\_sugar_{m,l,o,t} = \beta_0 + \beta_1 avg\_sdil\_spend_{m,l,o,t} + \beta_2 reform\_prop_{m,l,o,t} + \beta_3 sdil\_weight_{m,l,o,t} + \beta_4 sdil2\_weight\_prop_{\ m,l,o,t} + \eta_{m,l,o,t}$$

Where m signifies the MSOA, l signifies the LAD, o is the OA Cluster Super Group, t is month.

$\eta_{m,l,o,t}$ is assumed to follow an ARIMA model.

The dynamic regression model was used because it allows for the combination of time series and ordinary least squares methods. Dietary behaviours are notably sticky, so modelling serial autocorrelation is necessary. As past behavior is not entirely predictive of future behavior, a dynamic regression model allows for adding time-variant variables that are likely to affect sugar consumption. A mixed hierarchical and grouped structure is used. Time series are organised in a Local Authority District (LAD) – Middle Super Output Area (MSOA)[16] hierarchy, and grouped with Output Area Classifciation (OAC) [17] Super Groups (all 2011). Since the cardholder population is quite skewed in a few neighbourhoods, MSOA allows fewer areas to be dropped while reducing the 'averaging effect' of aggregating variables on an area level.

The 'fable' package in R was used to train a unique ARIMA model for every LAD-MSOA-OAC combination. Due to imbalanced data, we did not use the model to predict future sugar purchases; each MSOA must contain all OAC Super Groups (e.g. Suburbanites, Hard-pressed Living, etc), such that the model can predict at the MSOA-Super Group level and 'build up' the LAD forecasts. Prediction at MSOA-Super Group level is possible, but there are

''merging_datasets.R', 'product_data_over_time.R', 'time series plots.R', 'time series regression.R'

too many combinations to make predictions meaningful. The training accuracy was computed for each model.

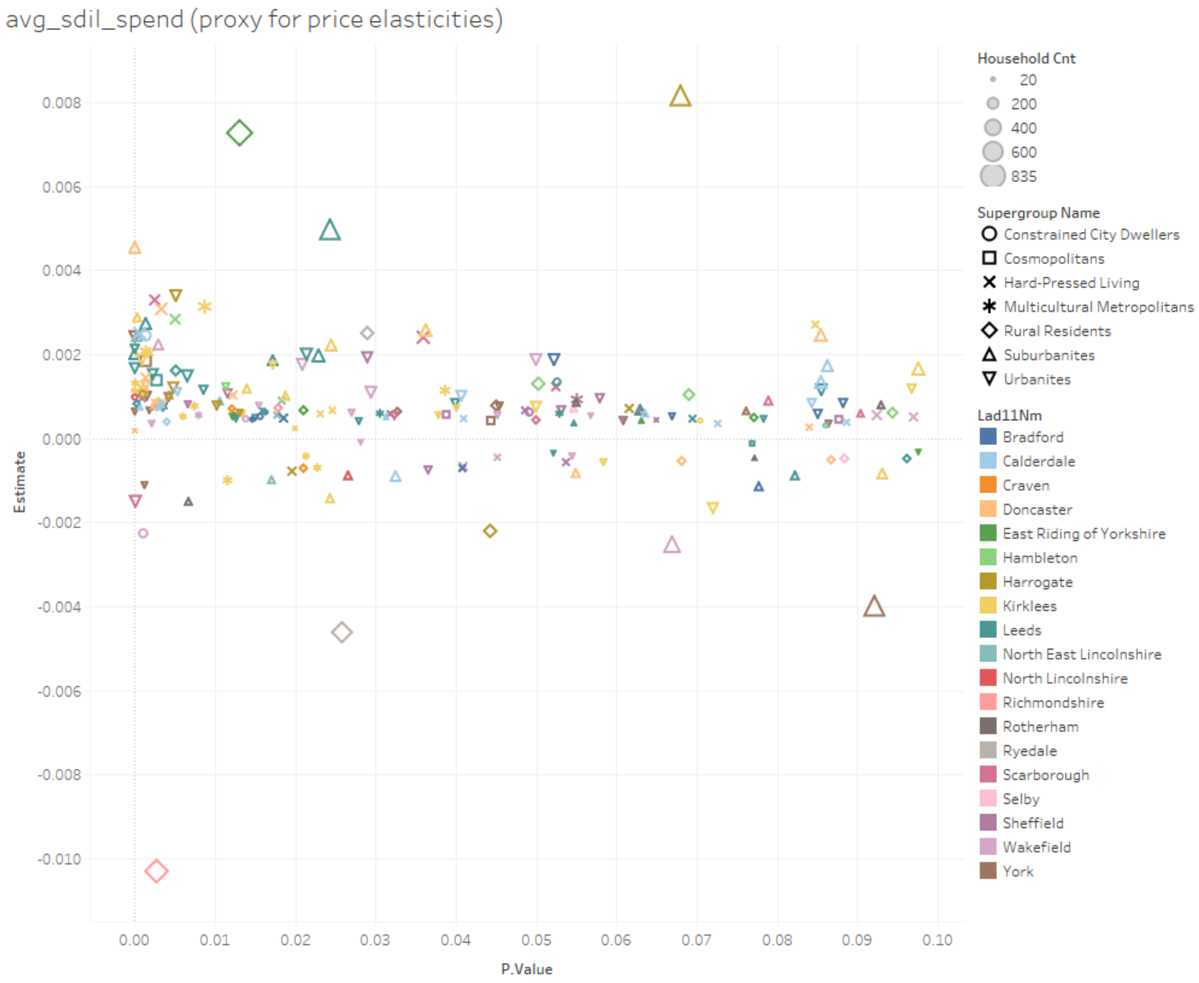
9.2 Results: Multivariate time-series regression

Since the number of coefficients for one variable can be up to three digits, we plot the coefficient values on a scatter plot against its p-value. To focus only on reliable results, only coefficients significant at 10% level in a model with ≤10% mean absolute percentage error (MAPE) are plotted. MAPE is calculated as the average of 100*(actual-observed)/actual for all observations in one dataset. In addition, the LAD and Super Group are visualised by colour and shape correspondingly. The size of the shape denotes the number of households in that combination.

Spend/100ml (β1)

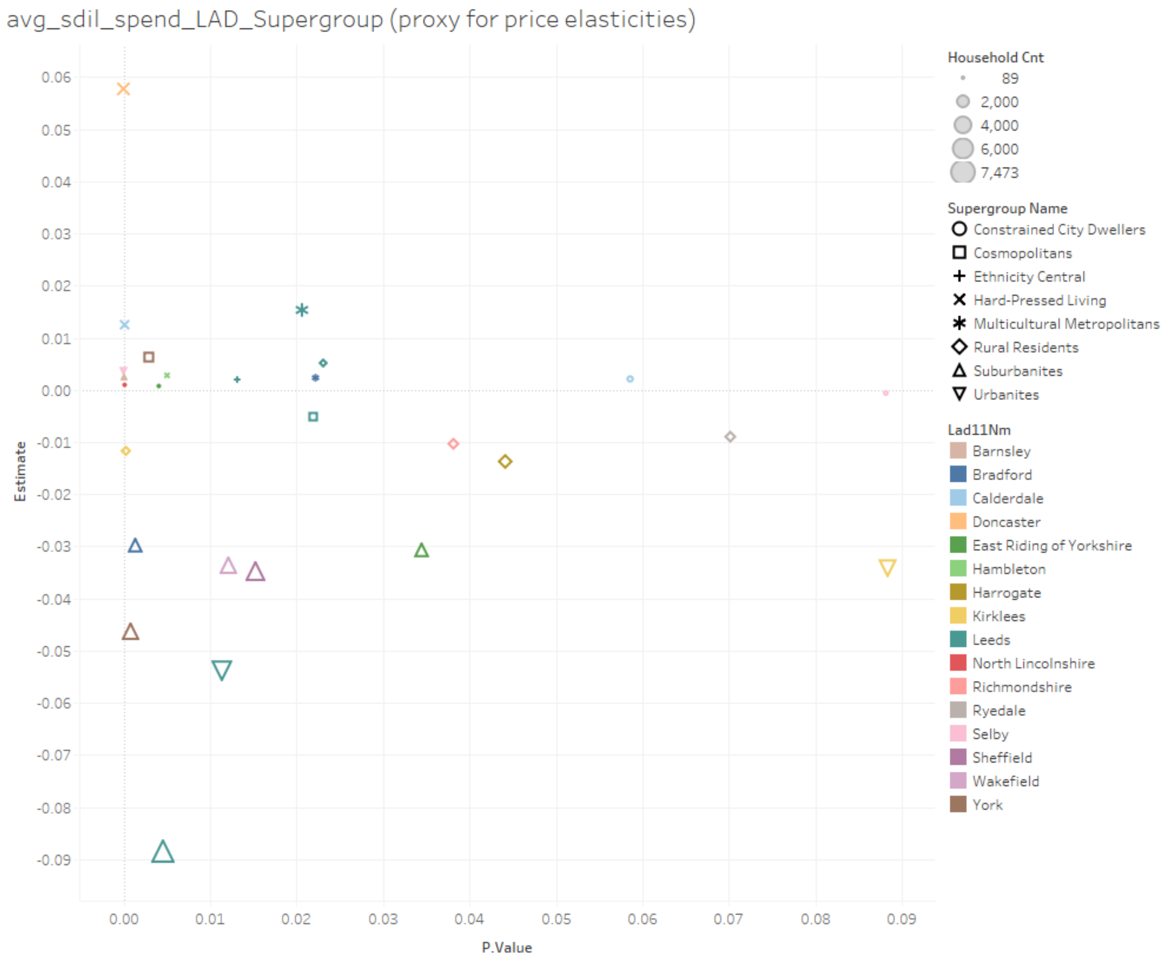
First we consider average spend on SDIL-applicable drinks (£/100mL) (β_1), where we regard β_1 as signifying willingness to pay, a proxy for price elasticity. There is a wide range of $\hat{\beta}_1$ values among LAD-MSOA-OAC combinations (Figure 27) and no clear geographic/demographic pattern can be seen. When plotted at the LAD-Supergroup level (Figure 28), we see a clearer pattern beginning to emerge, suggesting there is high variation within LADs/Supergroup levels and that consumption behaviour correlates with a combination of both geography and demography. Figure 28 shows that significant coefficients for Suburbanites (upward triangles) and Urbanites (downward triangles) tend to be negative, suggesting that they respond negatively to price increases and reduce purchased sugar. Rural Residents (diamonds) display a smaller negative response, while coefficients for Multicultural Metropolitans (Stars) and Hard-Pressed Living (X's) are positive, suggesting that they increase purchased sugar from drinks as prices increase. Significant coefficients for Cosmopolitan (squares) and Constrained City Dwellers (circles) areas tend to sit around zero, suggesting they show behavioural response to price increases.

Figure 27. $\hat{\beta}_1$'s at LAD-MSOA-Super Group level



''merging_datasets.R', 'product_data_over_time.R', 'time series plots.R', 'time series regression.R'

Figure 28. $\hat{\beta}_1$'s at LAD-Super Group level



One caveat is that the coefficient estimates are extremely small. As an example, the coefficient value for 'Rural Residents, East Riding of Yorkshire 009' is only 0.007 (Figure 27), the second-highest in the dataset. The group has 384 households. The standard deviation of average spend for this group throughout the sample period is 0.896p per 100mL. This means if the average spend per 100mL SDIL-applicable drinks is to increase by 2.4p (the SDIL 2 tax), this will roughly increase the sugar consumption by 1.28g per cardholder per month, a minimal effect.

$$\frac{\frac{price\ increase}{s.d.of\ average\ spend\ in\ sample\ group}\times \hat{\beta}_{1,m,l,o}\ \times s.d.of\ total\ sugar\ per\ month\ in\ sample\ group}{No.of\ cardholders\ in\ sample\ group}$$

$$=\frac{\left(\frac{0.024}{0.00896}\right)\times 0.007\ \times 26292}{384}$$

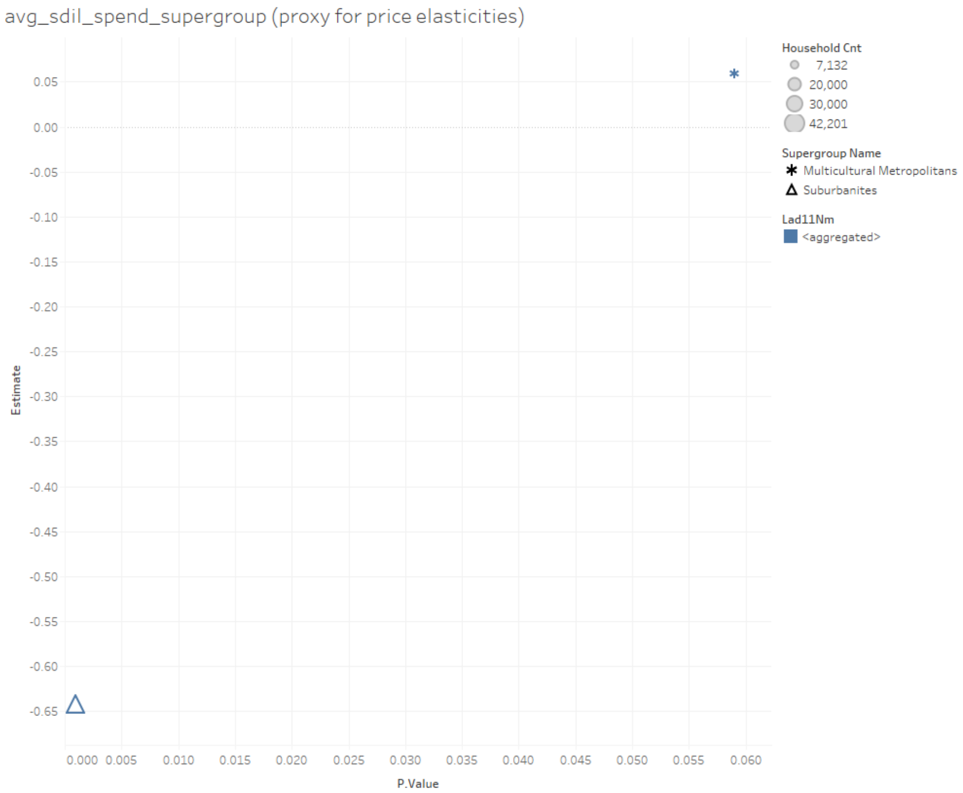
$$= 1.28\,g\ \text{per cardholder per month}$$

The small coefficient values on both sides of the zero line indicates that soft drinks price has limited correlation with purchased sugar from beverages. Additionally, the high number of positive coefficients in Figure 27 suggests that behavioural patterns are persistent for many population groups which

''merging_datasets.R', 'product_data_over_time.R', 'time series plots.R', 'time series regression.R'

are little affected by the price increase. However, one should be cautious to draw the conclusion that lower income groups are less responsive to price, as $\hat{\beta}_1$ for hard-pressed living becomes insignificant when data is aggregated at supergroup level, while $\hat{\beta}_1$ for suburbanites remains significant and negative (Figure 29).

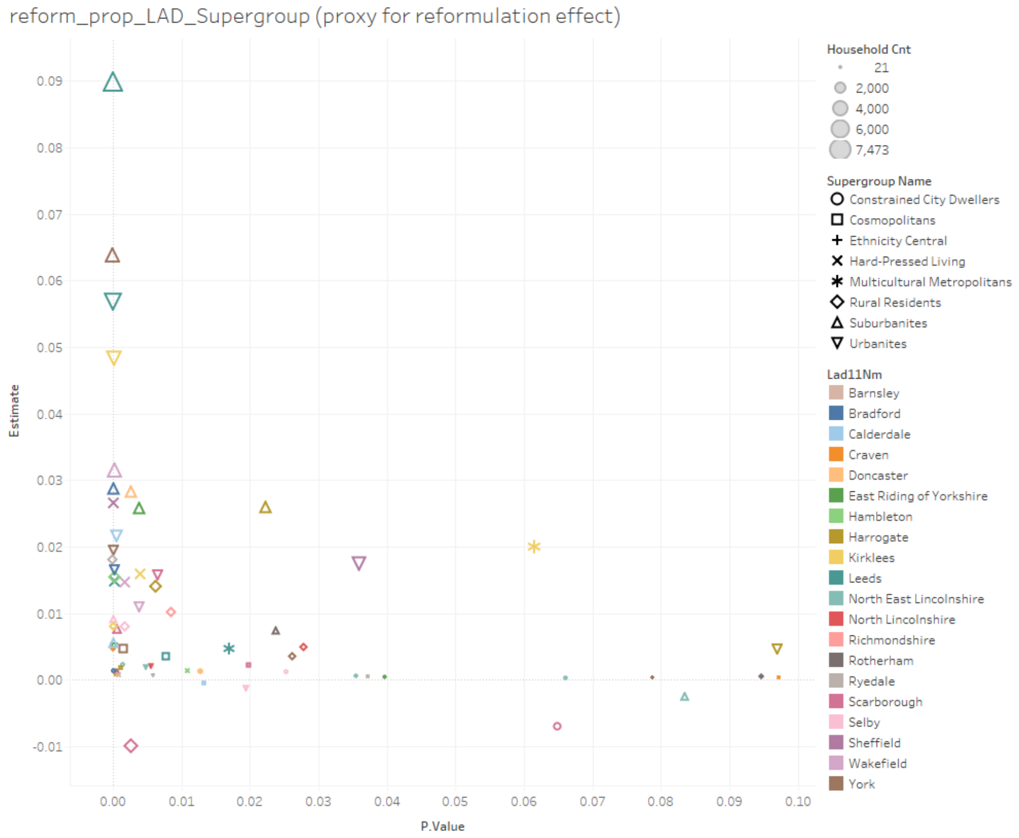
Figure 29. $\hat{\beta}_1$'s at OAC Super Group level



Proportion of reformulated products (β_2)

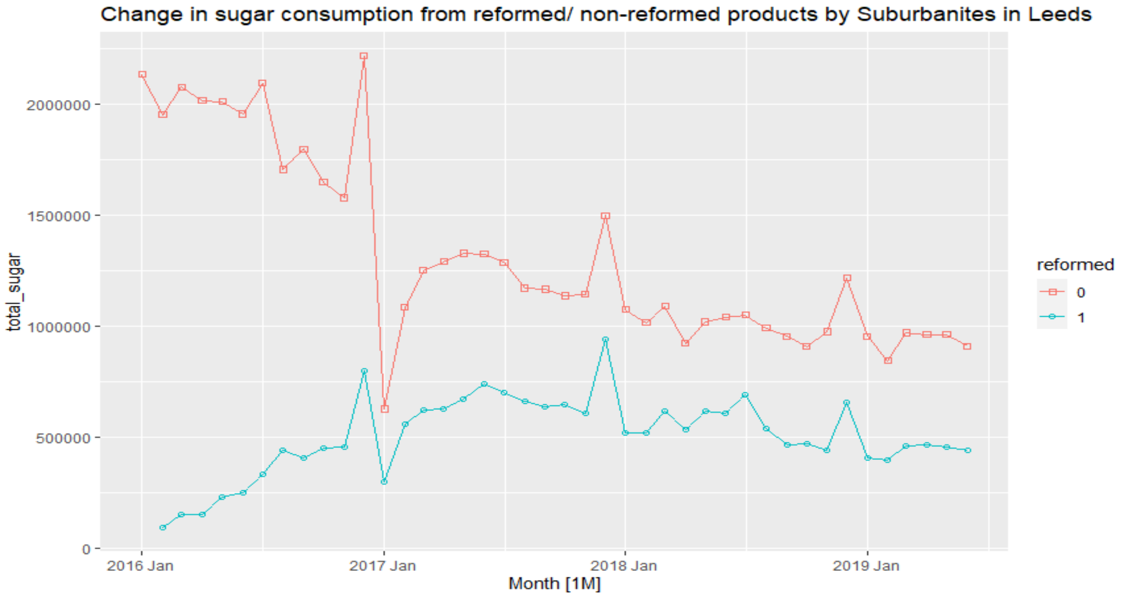
We expected that a bigger proportion of reformulated products in the drinks basket will lower consumers' sugar consumption. However, the majority of $\hat{\beta}_2$'s we estimated are positive. In other words, a higher proportion of reformulated products (of total drinks) in the drinks basket is correlated with higher total sugar from drinks. Figure 30 shows the coefficient values at the LAD-Super Group level.

Figure 30. $\hat{\beta}_2$'s for each LAD--OAC Super Group combination



Taking Suburbanites in Leeds (the highest turquoise triangle on Figure 30) as a case study (Figure 31), we can see that purchased sugar from non-reformulated products (red line in Figure 31) fell substantially over time, while purchased sugar from reformulated products (blue line) rose between the announcement and implementation period (2016 March-2018 April), before declining slowly.

Figure 31. Sugar from reformulated (blue line) and non-reformulated (red line) products among Suburbanites in Leeds

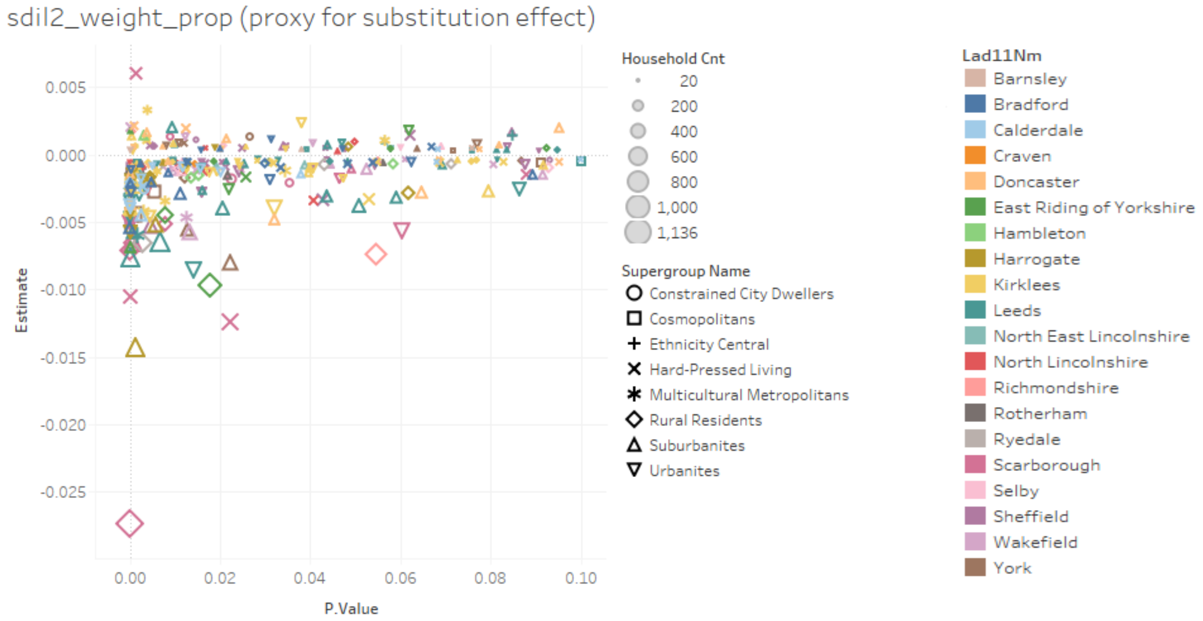


''merging_datasets.R', 'product_data_over_time.R', 'time series plots.R', 'time series regression.R'

Weight proportion of SDIL-2 applicable drinks (β_3)

We expected this variable to correlate positively with sugar consumption, as the more SDIL-2 applicable drinks make up the drinks basket should lead to more sugar consumption. While this is the case for some population groups, more than half of population groups in the data have a negative correlation between SDIL-2 applicable drinks proportion and sugar consumption (Figure 32).

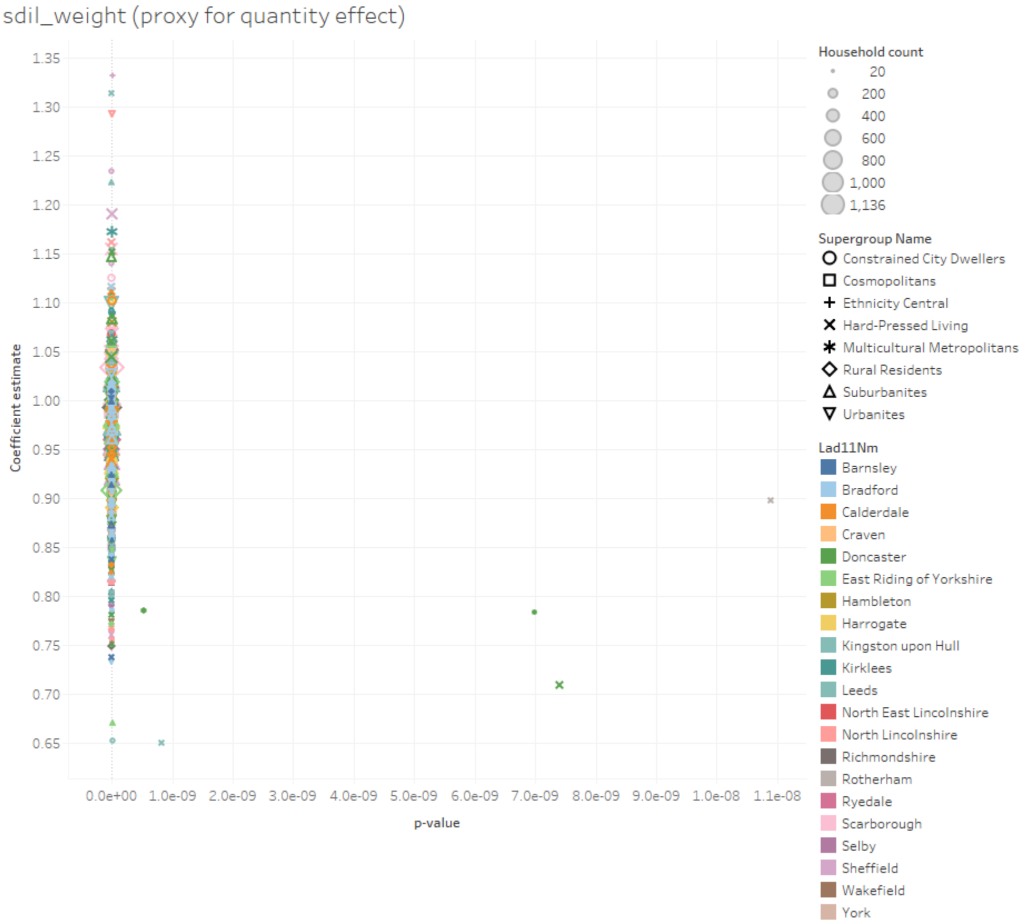
Figure 32. β_3 coefficients at LAD-MSOA-OAC Super Group level



Weight of SDIL-applicable drinks (β_4)

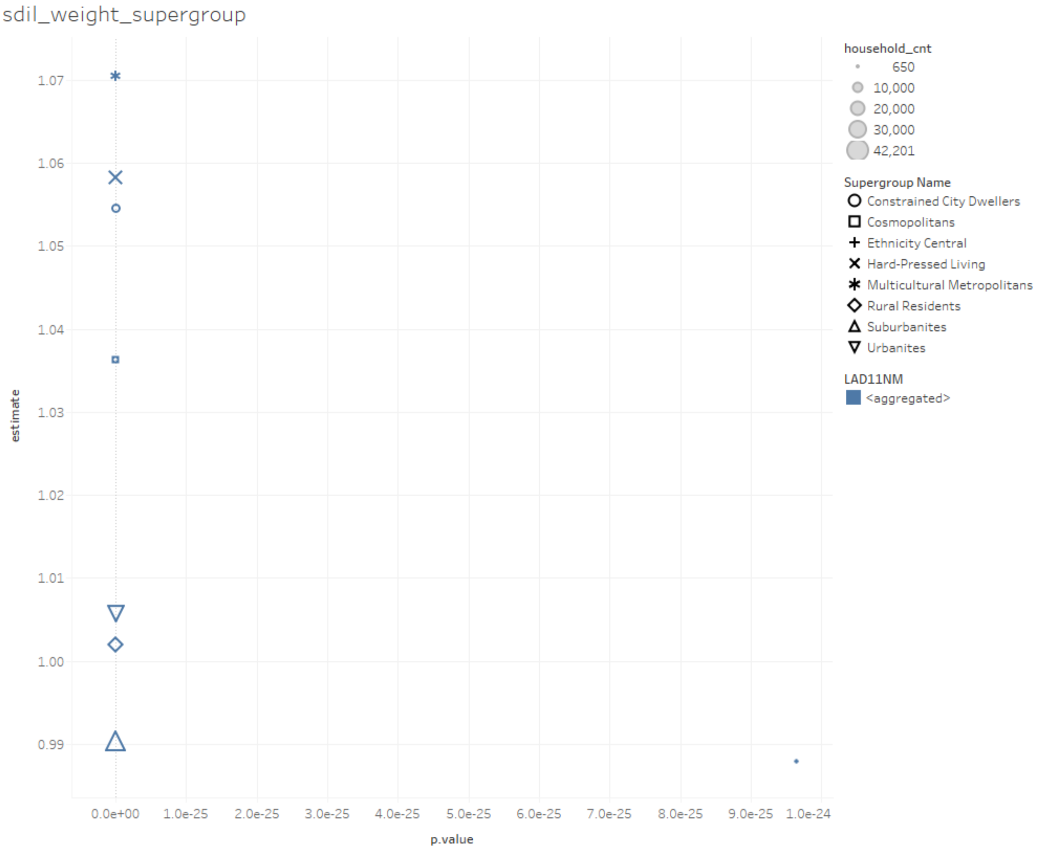
As expected, all significant coefficient estimates for this variable are positive (Figure 33), with the majority clustering around $\hat{\beta}_4$ = 0.9-1.05, and have very low p-values. That is, the higher the weight of SDIL drinks purchased, the higher the amount of purchased sugar from drinks. To put the estimates in context, we use a rough calculation similar to formula 1. Using Calderdale Suburbanites, a population group with 185 households and $\hat{\beta}_4$ = 0.9726, a 1L reduction (supposing 1ml = 1g) in SDIL-applicable drinks is correlated with a 0.49g sugar decrease on average per household per month. The fact that it is not a one-to-one relationship suggests that the reduction of sugar purchased from SDIL drinks is likely to have been compensated by the increase in other kinds of drinks, evidence of a substitution effect.

Figure 33. β_4 coefficients at LAD-MSOA-OAC Super Group level



Another observation is that the effect of SDIL weight decrease is relatively less for suburbanites (Figure 34). Suburbanite groups have the highest coefficients for reformulated proportion of drinks basket, which suggests that they are more likely to switch to reformulated products rather than reducing purchases altogether, dampening the effect of reducing weight of purchased SDIL drinks.

Figure 34. β_4 observations for Super Group



9.3 Conclusions: Multivariate time-series regression

Generally, soft drinks price (/100ml) has limited correlation with total beverage sugar. We speculate that changes in price /100ml are hardly felt by consumers and that price changes per unit may be a greater driver of choice (i.e. shrinkflation where product volume reduced but unit price remained constant). Positive responses to price (/100ml) were observed in Multicultural Metropolitan and Hard-Pressed Living areas, suggesting that lower income groups may have been more adversely affected by shrinkflation. There is a wide range of price sensitivities among groups. Suburbanites in some LADs, such as Bradford, Sheffield and York responded more negatively to price. Rural Residents, on the other hand, appear to have more persistent consumption patterns that are less affected by price.

The quantity effect is positive across the majority of population groups, however, the reduction in sugar is not one-to-one, as reduction in weight of levy drinks is sometimes accompanied by substitution with non-levy drinks. Among some population groups the increase in volume from reformulated drinks seems to more than compensate for reduction in SDIL drinks, thus increasing overall beverage sugar. Some groups are notable for further exploration; Suburbanites appear sensitive to price of SDIL drinks and replace them with low sugar drinks, while prone to over-substitution. Multicultural cosmopolitans are less sensitive to price, but less likely to substitute sugary drinks with low sugar ones, hence reducing SDIL purchases has a more negative impact on total sugar.

Limitations include; 1) we did not have enough data before the announcement or after the implementation to see a longer trend in sugar consumption. 2) Since the tax is implemented unilaterally across the UK, it is not possible to use methods such as difference-in-differences/ synthetic control which (the standard approaches for programme evaluation). 3) We did not have a clear identification method for reformulated products to measure the impact of reformulation on consumers. 4) We represent area demographic characteristics using OAC Super Group instead of the IMD, giving a more holistic demographic picture with features such as rural-urban settlements, age, and ethnic diversity. The same regression method could also be applied to IMD, which may yield additional insights. 5) Demographic variables were limited by those available, we anticipate that additional demographic data could identify consumer traits that correlate with purchased sugar.

10. Case study

Sugar sweetened beverages containing ≥8g sugar/100ml were subjected to the highest level of the Sugar Levy (SDIL2) when it was introduced. Manufacturers were forced to decide whether to reformulate their products to avoid the tax (replacing some or all of the sugar with artificial sweeteners), or to continue with the same product formulation and absorb the tax/pass some or all of it on to the consumer. From our observations of reformulated products, there was a divide in approaches. While Sainsbury's was active in reformulating a large number of their own brand beverages, the larger soft drinks companies opted to keep their original product on the market, which meant the price (/100ml) increased following the levy.

This case study explores the responses of customers of one of the most popular soft drinks brands. It allows us to examine brand loyalty and see how a consumer reacts when they have no choice but to pay more for the sugary variant of the drink they like, or switch to an alternative product (no/low sugar variant of the same drink, or alternatives provided by another brand).

10.1 Methods: Case study

10.1.1 Exploratory data analysis

15 full sugar SKUs of the same branded drink (representing different formats e.g. cans, bottles, multipacks) were identified in the transaction dataset. Throughout the rest of this section, those 15 SKUs will be referred to collectively as 'the target product'. The transaction dataset was then filtered to around 12k customers who; 1) were 'regular' customers - purchasing in Sainsbury's in each of the 42 months of the period captured in the transactions data (to minimize probability that a decrease in consumption will be attributed to attrition), and 2) bought at least one of the target product on 3 or more occasions during the data collection period, to exclude people who only bought once (a lot of customers) but keep those who only buy occasionally.

Initial data exploration was conducted in Tableau to explore 5 questions:

1) How has the price and purchased volume of the target product changed over time?
2) Did total volume of all drinks decrease? (This may indicate a shift away from drinks in general, rather than just the target product)
3) Has total sugar purchased changed?
4) How has the price of the target product changed?
5) How did customers respond to smaller pack sizes of the target product? I.e. Did they keep buying the same amount of units without noticing they were smaller?

'targetproduct_clusters_data.R', 'targetproduct_clusters.R', 'Target product demographics classification.ipynb'

10.1.2 K-means clustering

Following this initial exploration, K-means clustering was applied to identify behavioural subgroups of consumers. We hypothesised that customers would fall into 4 clusters of customers; 1. Increase in consumption – consumers who decided to switch to the target product in an attempt to avoid artificial sweeteners when most other drinks on the market reformulated to include artificial sweeteners. 2. People who are prepared to pay more so didn't change their habits. 3. People who decreased their consumption - they still want to drink the target product but are not ready to pay more. 4. People who eliminated the target product from their basket. We allowed the algorithm to run unsupervised and compared the observed clusters with our hypothesised clusters.

Five groups of variables for each target product consumer were used to feed into the clustering algorithm. For the K-means algorithm to be effective, the data were scaled using the scale function in R. It deducts mean feature value from each observation and divides by feature standard deviation.

Variable groups:

1. **Weight of target product** – total weight and average monthly weight for each year. When calculating monthly averages, missing data in January 2017 was taken into account by summing all the values and dividing by 11 to get a monthly average. In 2019, we only have 6 months of data so the sum was divided by 6. This is the main variable that describes how much of the target product someone purchases.

2. **Weight of diet SKUs of the target product** – 27 Diet SKUs were found in the transaction data and used to get monthly average for each year as described above. This can identify people who also like diet drinks.

3. **Number of transactions involving the target product** - total number as well as monthly average for each year as described above. This could help to distinguish people who buy regularly (and have transactions on most months) vs those who buy seasonally (e.g. one big transaction for Christmas).

4. **Number of target product items monthly average** for each year as described above. This could help to distinguish people who buy big packs (large bottle) vs those who prefer multiple small packs (e.g. cans). Note, multipacks, e.g. 6x330ml are counted as one item so this is not a perfect measure.

5. **Proportion of drinks basket made by the target product**, monthly average for each year as described above. This identifies people whose main drink is the target product vs those who buy a range of soft drinks.

10.1.3 Classification model

'targetproduct_clusters_data.R', 'targetproduct_clusters.R', 'Target product demographics classification.ipynb'

Two classification models were trained on clustered consumers of the target product:

1. Using the same behavioural variables used for clustering, to assign customers to clusters and reveal feature importance to aid understanding of factors determining cluster assignment.

2. Using demographic variables as predictors of cluster assignment.

A multiclass classification XGBoost model (xgboost library in Python) was trained on a subset of the target product clustering dataset to assign each customer to a cluster, based on the same variables used for clustering. F1 score is a balance between precision (the ability of a model to allocate only correct customers, and not more, into the right cluster) and recall (the ability of the model to put all the cluster members into the correct group and not miss any) is used as a metric in this classification. This is better than accuracy (% of correctly assigned members) for imbalanced datasets like this one because if 80% of members fall into one cluster, assigning all the members into that one cluster would give accuracy of 80% - a good result without any work.

The multiclass model (model that assigns each instance to one out of more than two classes) used F1 score with weighted average – F1 score was calculated for each cluster and the average score was weighted on the size of the cluster to create a final score. Permutation importance feature was used to learn which variables were the most important in deciding which cluster a customer belongs to. This function essentially runs models with each feature being assigned randomly - so that age, and all the other variables, one at a time, are not associated with the correct person. If the model performance (F1 weighted average score) goes down when the feature wasn't giving correct information, then it means that it was important to cluster assignment. Negative permutation importance means the feature makes the model performance worse, and zero values indicate it makes no difference. The permutation importance function was run on both the training and test set to control for overfitting (overfitting happens when the model learns the training data so that it is not capable of generalising beyond that). If a feature had a high importance in the training set and not in the test set, it could mean it causes overfitting in the model.

To explore which features are most important for determining assignment to each individual cluster, 4 binary classification xgboost models (one for each cluster) were run; model 1 predicts if a customer will be in cluster 1 or not, model 2 predicts members of cluster 2, and so forth. Finally, an xgboost model was trained on demographic data for the target product customers using the following variables to determine their importance on cluster allocation:

'imd': deprivation level

'Age: midpoint from AgeRange in Demographics dataset

'F', 'M', 'U': one hot encoded gender, U for unknown

'urban': extracted from 'ru11ind', marked urban areas as 1 and 0 for the rural ones

'oac11_cat': factorized to be numerical

'ru11ind_cat': factored to be numerical

'LAD11NM_cat': factorized to be numerical

10.2 Results: Case study

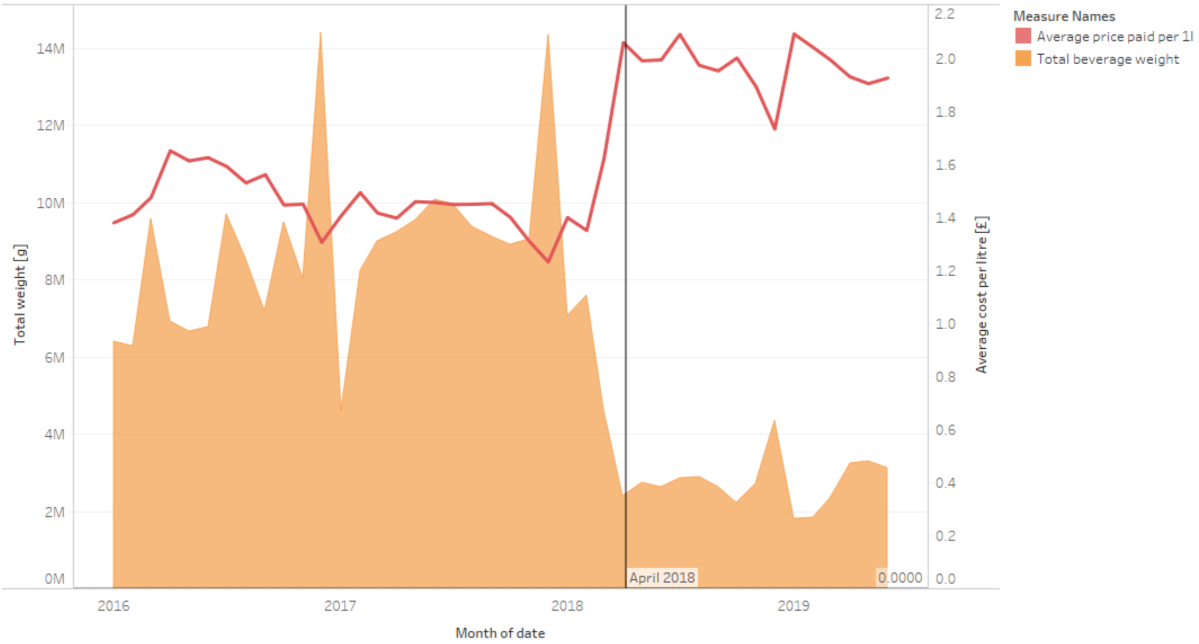
10.2.1 Exploratory data analysis

Here we report answers to each of the 5 exploratory data analysis questions.

1. **How has the price and volume of the target product sold changed over time?**

Figure 35 shows the total weight purchased by our sample of target product purchasers, of the 15 target product SKUs, and the average price per litre of product, over the 42-month study timeframe. A 40% increase in price per litre was implemented for target product SKUs just before the introduction of the levy. This was mirrored by a simultaneous reduction in the target product purchased (by volume) by our cohort, indicating a strong price response.

Figure 35. Purchase volume and price of the target product 2016 – June 2019

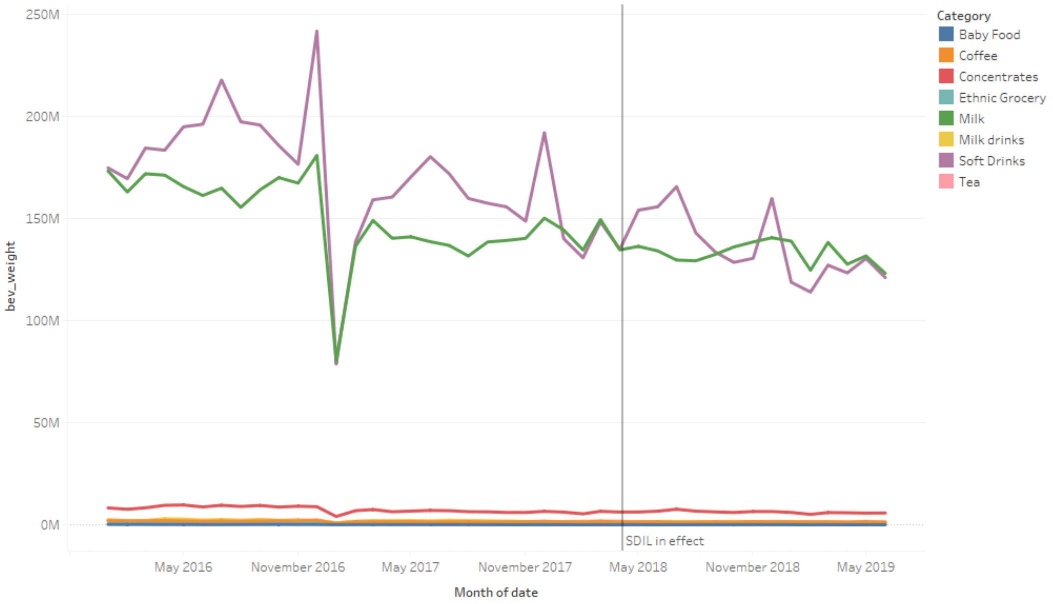


'targetproduct_clusters_data.R', 'targetproduct_clusters.R', 'Target product demographics classification.ipynb'

2. Did total volume of all drinks decrease?

Figure 36 shows that for target product consumers, the trend in purchase weight for all other categories of beverages is not as dramatic as that seen for the target product. Considering milk (green line) as a staple product unaffected by the levy, we can see that purchased levels remain relatively stable apart from the dip due to missing data in early 2017, which can serve as a control that our sample of consumers continued their shopping in Sainsbury's. We can therefore have greater confidence that trends seen in other categories are likely to be due to the levy rather than attrition from the sample.

Figure 36. Weight purchased by beverage category for target product purchasers



3. Has total sugar purchased changed?

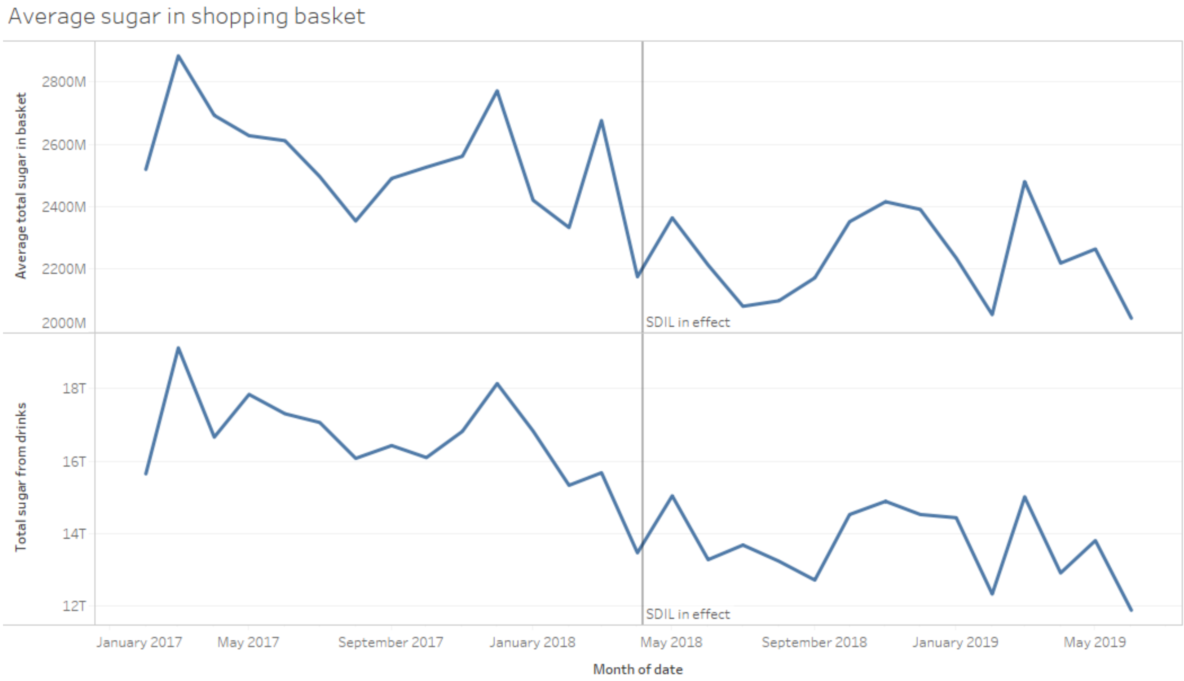
The total sugar content of a monthly basket (including both food and beverages) was calculated using the data available as:

*Bev_sugar [sugar/100ml] *bev_weight [ml] * 100 / prop_all_prods_sugar [sugar/all sugar]*

An approximation of average basket sugar content was calculated (this is not an exact average as records were not grouped by consumer ID, meaning that each consumer contributed their monthly total sugar content N * the number of SKUs they bought that month) and plotted over time (Figure 37). The downward trend observed in both basket sugar content and sugar coming from drinks suggests a correlation implying sugar removed from the diet when target product purchases were limited was not substituted with sugar from food items.

'targetproduct_clusters_data.R', 'targetproduct_clusters.R', 'Target product demographics classification.ipynb'

Figure 37. Average sugar from drinks vs total sugar from food and drinks in shopping basket for target product customers



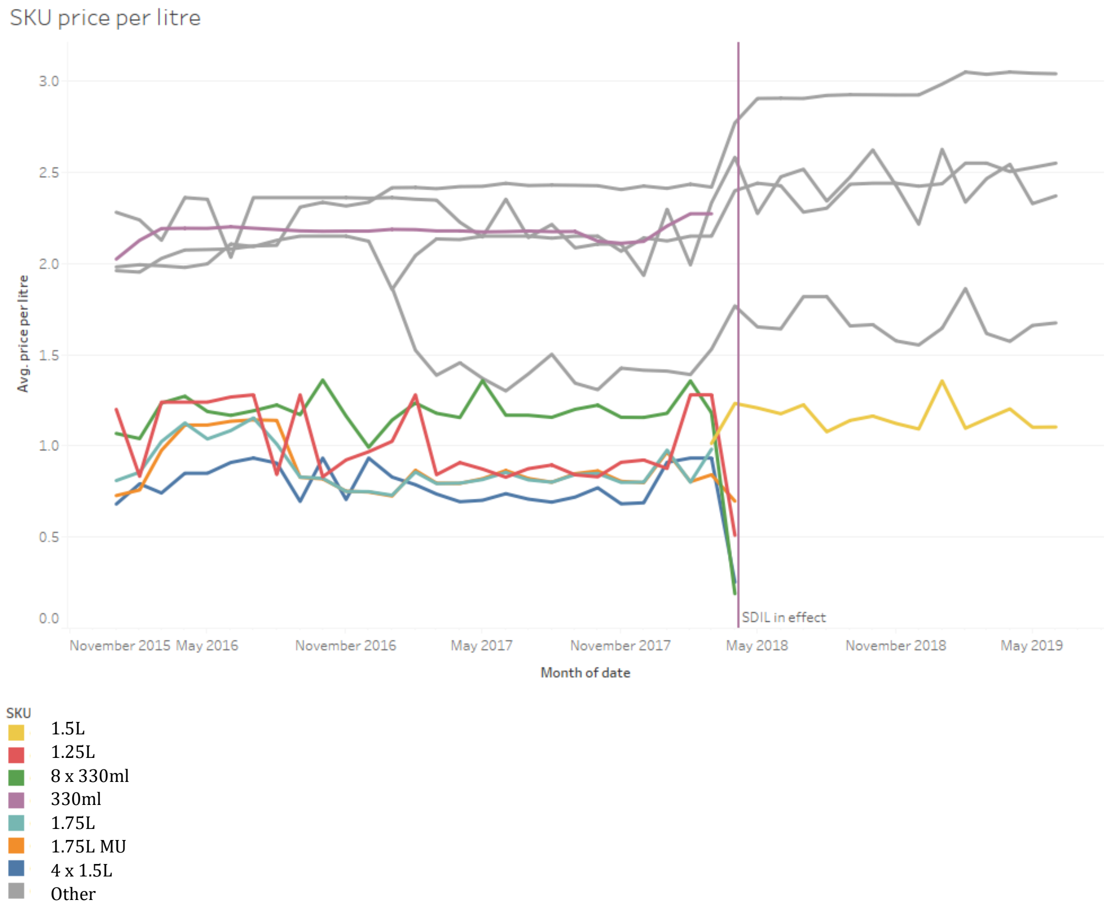
The trends of average of sugar in basket and sum of sugar from drinks for date Month. The view is filtered on average of sugar in basket and date Month. The average of sugar in basket filter keeps non-Null values only. The date Month filter ranges from February 2017 to June 2019.

4. How has the price of the target product changed?

The average price per litre for each target product SKU is plotted in Figure 38. Here four products were removed for which data wasn't available for a while after levy (maybe they were temporarily discontinued until consumers get used to higher prices). Incidentally, two of them were also the most popular products (by total weight); after the levy the 6x330ml SKU replaced them to become the most popular SKU.

From the plot above, we see that after the levy the five cheapest SKUs (per litre) were discontinued. One new SKU was introduced - 1.5L (yellow line), likely to replace 1.75L bottle (light blue line) and 4x1.5L (dark blue line). This is known as shrinkflation – a process in which a producer decreases the size of a pack so that the consumer is less likely to notice the price difference (because unit price can remain the same while price per litre increases). It would appear that customers may not have noticed the decrease in volume, as they continued to buy smaller packs for a similar price.

Figure 38. Price per litre for each SKU of the target product

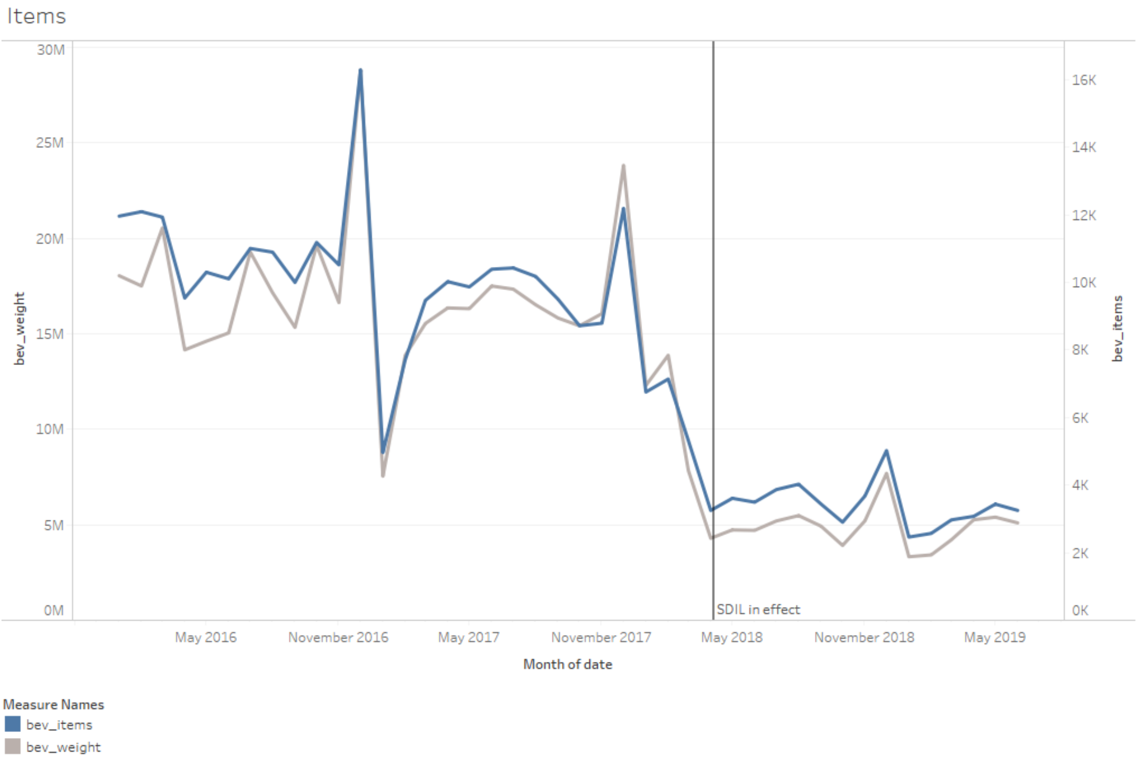


5. How did customers respond to smaller pack sizes of the target product?

In Figure 39 we show the total amount of items purchased over time, alongside the volume (by weight) of the target product purchased. If the amount of items (bottles, cans, multipacks) remained stable, we could suspect consumers may have unconsciously reduced their weight of the target product as they continued to buy the same amount of items at a smaller pack size. Increasing distance between the lines would indicate that customers increased the number of packs purchased to compensate for smaller packs.

In the plot, it appears the number of items followed a similar trend to the weight indicating that people on the whole continued to buy the same amount of items, just with smaller volume. The lack of compensation for reduced pack sizes suggests that shrinkflation may have contributed to the overall reduction in purchased sugar.

Figure 39. Purchases of the target product by number of items and weight

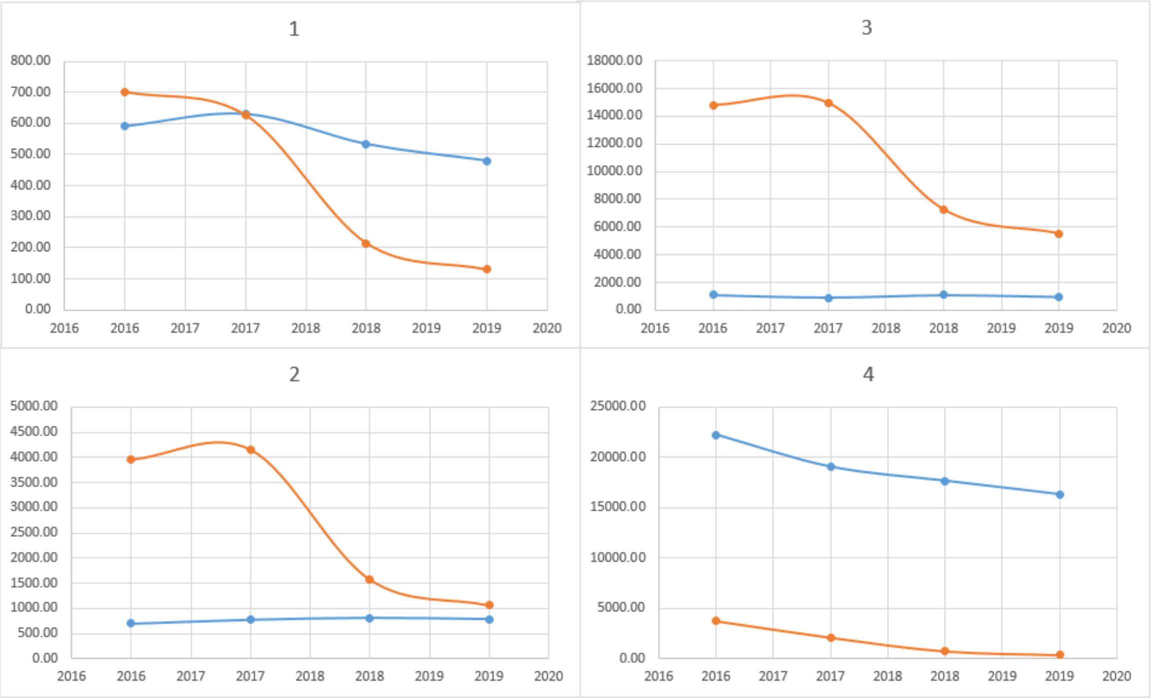


10.2.2 K-means clustering of target product customers

As predicted, four clusters of target product customers were found by the K-means algorithm, as described in Table 13 and shown on the plots in Figure 40. As hypothesised, one cluster contains customers who make little change to their habits (cluster 3, Committed sugar drinkers) and another contains customers who decreased their consumption but still drank the target product (cluster 2, Stickers). We did not see a cluster of customers who increased consumption of the target product to avoid artificial sweeteners, nor a cluster of customers who stopped purchasing the target product. Instead, the largest cluster we observed were customers who switched from the target product to the diet equivalent (cluster 1, Switchers), and a cluster of customers who purchased mainly the diet variant with the occasional target product drink (cluster 4, Dieters). All clusters showed a downward trend of target product purchases, while the diet variant appeared to be more acceptable to the target product customers than we hypothesised.

‘targetproduct_clusters_data.R’, ‘targetproduct_clusters.R’, ‘Target product demographics classification.ipynb’

Figure 40. Trends in monthly volume (ml) of the target product (orange line) and the diet variant (blue line) for each cluster identified by K-means



‘targetproduct_clusters_data.R’, ‘targetproduct_clusters.R’, ‘Target product demographics classification.ipynb’

Table 13 Description of clusters of target product buyers, determined by K-means classification

Cluster number and name	Number of customers	Short description	Detailed description
1. Switchers	9,897	Occasional buyers, significantly decreased their target product consumption and now drink more of the diet variant than the target product.	This cluster started with an average of 2 big bottles of the target product three times a year in 2016 and ended up buying just 1 per year in 2019. At the same time, their consumption of diet version went to surpass the consumption of the sugary version in 2018 with an average of 500ml per month. However, as the diet consumption showed a slight downward trend, it seems like these people reduced their consumption of the target product altogether. Considering all Soft Drink subcategories, no products that followed an upward trend were identified. This group may be too big to see specific trends, or they have decreased their purchases of Soft Drinks overall.
2. Stickers	1923	Relatively frequent buyers, decrease their sugary consumption over time but didn't substitute with the diet variant.	This cluster remained a relatively low, occasional buyer of diet drinks while reducing their consumption of the target product from 4 litres per month to just 1 litre. Their contribution of the target product weight to total weight of beverages went down a lot in 2018 to show a sign of rebound in 2019 suggesting they perhaps switched to other drinks in 2018 but they started giving them up in 2019. They continue to drink more of the target product than the diet variant.
3. Committed sugar drinkers	196	Biggest buyers, limited their sugary consumption but didn't substitute much.	Sugar from drinks contributes around 25% of total basket sugar for these customers, more than in any other cluster. They also purchase much larger volumes of the target product than any other cluster. The target product remains to be among their favourite drinks but the diet variant shows a slight increase over time, but not enough to replace lost purchases of the target product.
4. Dieters	154	Diet drinkers with occasional purchases of the target product. They didn't really change.	This cluster buys both types of the target product and its diet variant, and are the only cluster which started the period buying more diet variant than the target product. Both the consumption of diet and original showed a slow decline. Over the timeframe, they consumed on average 1.7l of the target product per month and 18l of the diet variant on an average month.

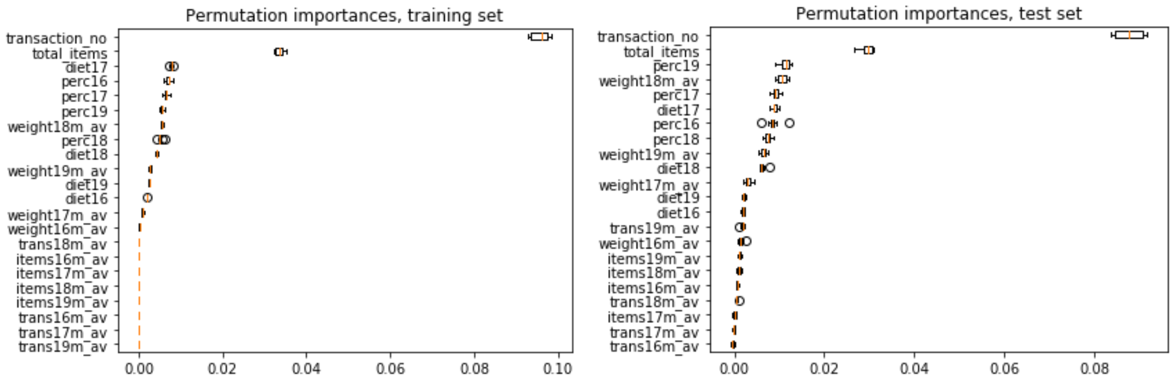
'targetproduct_clusters_data.R', 'targetproduct_clusters.R', 'Target product demographics classification.ipynb'

10.2.3 Classification model

Model 1, trained using the variables used for clustering (listed above), achieved an F1 score of 0.988 (perfect assignment gives a score of 1). From the plots in Figure 41, it is clear that similar features contributed to allocation both in training and test set, indicating a high-quality model.

Total number of transactions looks like the most important feature. Because the data was aggregated by month, one transaction corresponds to buying one SKU at least once in one month. Therefore, more transactions means either shopping frequently throughout the year or shopping for many SKUs. Total transactions are more important than total items which suggests it is perhaps not the volume itself but rather frequency of visits and/or variety of target products in the basket which drives cluster assignment (more work would need to be done to find which one of these is more important). Two of the most important features, total transactions and total items, are related with variables that describe items and transactions per month. It seems like the model could get the information it needed just from the total, without focusing on the yearly values, indicating the yearly total could be removed to force the model to decide which year was important for determining cluster allocation.

Figure 41. Classification Model 1 permutation importance for training and test data sets



Percentage of target product drinks in a basket was informative every year but it seems like the weight of diet drinks was particularly important in 2017, just before the introduction of the levy. However, the values on Y axis (decrease in score when the feature was randomised) for diet features are too small and too similar to each other to conclude that one year was more important than the others.

When the model was run separately for each cluster, it appears to do better on clusters which are driven by purchase frequency (people who buy only occasionally, people who buy a lot), rather than by another feature (e.g. purchase of the diet variant in cluster 4) (Table 14). This confirms the earlier observation that purchase frequency drives model performance.

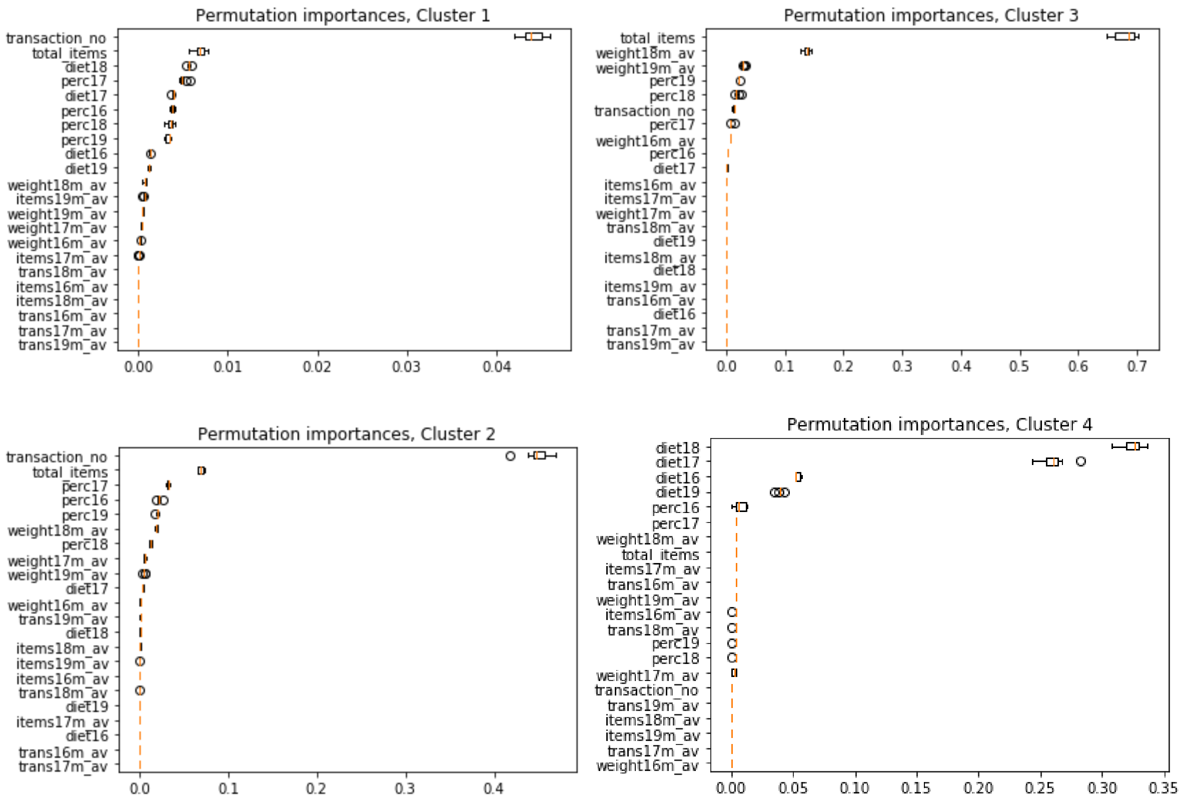
‘targetproduct_clusters_data.R’, ‘targetproduct_clusters.R’, ‘Target product demographics classification.ipynb’

Table 14. Performance scores for classification model run separately for each cluster

Model	F1 score
Cluster 1	0.992
Cluster 2	0.971
Cluster 3	0.909
Cluster 4	0.895

Next, feature importance was assessed for each of the models, to identify which features distinguish each cluster most. These are visualised in the permutation plots in Figure 42, and described in more detail below.

Figure 42. Classification Model permutation importance for each cluster



The total number of transactions made by a customer is the most distinguishing feature of Cluster 1. Customers who were placed in this cluster had many less transactions involving the target product than people in other clusters. That value ranged from 1 to 30 in this cluster. Similarly, people in this cluster bought less items in total than any other target product customers – from the minimum set as 3 to 197 items. There may be quite a large variability within this large cluster as it includes people that bought 4 units once, all the way through to quite frequent customers. Volume of the diet variant, and percentage of the target product added to their total drinks basket are further determining features for this cluster. Cluster 1 customers drink less of the diet variant, and the target product is less important in their drinks basket. They are the only target product cluster for which Soft Drinks is not the biggest (by weight) category – they

‘targetproduct_clusters_data.R’, ‘targetproduct_clusters.R’, ‘Target product demographics classification.ipynb’

buy roughly as many soft drinks as milk. Customers in this cluster buy roughly as much of the target product as still water, indicating that basket variety is an important feature for this cluster. This breadth of drinks portfolio could explain why they appeared more willing to switch from the target product to the diet variant indicating a good potential target for behaviour change, but possibly not a top priority given that soft drink purchases are comparably low.

Cluster 2 customers were differentiated from other clusters by the number of transactions they had. They were buying frequently (but not as frequently as cluster 3) with their transaction number ranging from 20 to 81. Their total items variable was also standing out as the second biggest, but significantly smaller than for cluster 3, similarly with percentage of drink that the target product accounted for. An important difference from cluster 1 is that customers in cluster 2 drink much less of the diet variant so 'diet' features are not particularly informative here.

Cluster 3 customers are determined by the total amount of items bought – an average cluster 3 customer bought over twice as many items as an average member of cluster 1, 2 and 4 summed together. The target product also has a very high contribution to the drinks basket and 'perc' features at the top of feature importance chart. The target product is their main soft drink, with no other drink reaching a similar volume (looking at sum for all members of the cluster over time), indicating limited basket variety.

Cluster 4's permutation importance plot looks different to the other ones. The only variables that count are how many diet drinks someone buys. No other cluster can compare with the weight of the diet variant bought by customers in cluster 4. The diet variant is their favourite soft drink, bottled water comes second in terms of weight and their third and fourth choice of drink is Diet Lemonade and Orange Juice.

Demographic cluster determinants:

Throughout work on this project, we haven't found a clear link between demographics and cluster allocation. The link between demographic data and cluster assignment was checked for the target product customers using the xgboost model. The model achieved an F1 score of 0.712, indicating fair performance at face value, but it allocated most customers in the majority cluster 1 without identifying any members of cluster 3 and 4 suggesting demographics are not a reliable determinant of cluster allocation.

10.3 Conclusions: Case study

This case study investigation allowed us to look more closely at customer responses to brand specific product changes as a result of the levy. Here we found clear evidence of switching for some customers, from the full sugar target product to the diet variant. This appeared to be more likely in customers who already shopped a wider variety of soft drinks in their

'targetproduct_clusters_data.R', 'targetproduct_clusters.R', 'Target product demographics classification.ipynb'

basket. Customers displaying a clear preference for the target product (high volumes and high proportion of total drinks basket) did not show switching behaviours, but still reduced their purchase volume. We did not find a cluster of customers who did not change their target product purchasing behaviour during the period, however it may be that this cluster was simply too small to detect.

We consider that perhaps the decision to include people who bought more than 3 items was not strict enough and people who bought very infrequently (based on number of transactions) could be discarded for another analysis considering only non-occasional customers of the target product. Further analysis could also focus on other categories that target product purchasers may have switched into.

Methods for dealing with imbalanced datasets were explored when working on classification algorithms for general clusters. However, they did not bring much improvement in that case and we haven't had enough time to try them here. It may be that cluster imbalance is still influencing classification models and masking true determinants of classification, so this warrants further exploration. From this, and other analysis undertaken, it seems like the clusters don't follow any demographic pattern, further suggesting that behavioural features are more important for understanding customer response to the SDIL.

'targetproduct_clusters_data.R', 'targetproduct_clusters.R', 'Target product demographics classification.ipynb'

11. Summary of key findings from the challenge

- The majority of customers changed their beverage purchase behaviours during the period.
- Most customers reduced the quantity of purchased sugar from beverages. A small number increased their overall sugar from beverages, but it is unclear why.
- Around a third of customers were ‘sticky’ in their behaviours. They are more likely to be customers who purchase high-sugar beverages and live in lower income areas.
- Behavioural clusters showed little association with customer demographics, suggesting that preferences are better predictors of levy response than demographic characteristics (this may be due to a lack of demographic information).
- Younger people prefer higher sugar SDIL1 drinks, while men prefer higher sugar SDIL2 drinks. Customers in more deprived areas have a slight preference for higher sugar SDIL2 drinks.
- Different groups display different price sensitivities – those living in deprived and ethnically diverse areas appear to be least responsive to price increases.
- There is evidence of switching from high sugar to low/no sugar alternatives, but this is most likely among customers who already bought both types of drinks before the levy.

12. Future work and research avenues

Our results did not find any association between beverage purchase behavioural clusters/patterns and customer/area demographic variables. Further investigation is warranted to understand if this is a real finding, or due to a lack of demographic information. For example, it would be useful to include additional data on household size, household composition, income, employment and education, at the customer or area level. Additional attention could also be paid to the spatial distribution of findings, and the Index of Multiple Deprivation as a potential explanatory variable in future analysis. Proxies for socioeconomic status (e.g. purchase of branded vs own brand products) may also be used.

Further work could focus on a product category level to unpick behaviours within clusters or broader trends observed. While we explored a number of clustering approaches, these could be improved upon by assessing the quality and suitability of clusters through statistical means. Some clustering approaches such as DBSCAN and network analysis were computationally infeasible given the time frame for analysis and data volume available, but could warrant further exploration. Additional information on promotions, temperature, time of day etc. may be useful to improve the accuracy of models by accounting for currently uncaptured point of choice factors. Furthermore, it would be interesting to include food item purchases, which may add value to behavioural clustering, and to understand how changes in beverage purchases may have influenced the diet overall.

An understanding of the impacts of the SDIL and implications for estimating propensity for dietary change are highly relevant in the context of changing food policy. Firstly, we acknowledge that without information on added sugars content, we cannot accurately assess which products are in or out of scope of the SDIL. A more accurate indication of SDIL status is required to confirm the reported findings. The National Food Strategy's report, 'The Plan' outlined a recommendation to replace the SDIL with a manufacturer's levy on added sugar used in food and beverage production. Future work could explore the potential impact of replacing the SDIL with an added sugar tax, or alternative bases for taxation such as the UK's Nutrient Profiling Model.[23]

Finally, the use of supermarket purchase records for the monitoring of population diet and policy evaluation is an emerging science. We acknowledge that purchases from a single supermarket chain do not represent the whole diet, and capture what is purchased rather than what is consumed. Furthermore, a supermarket's loyalty card customer-base is a self-selected sample of individuals, unlikely to be representative of the general population[14; 24]. Additional research is needed to understand the generalisability of findings from a supermarket loyalty card cohort, to assess the coverage of overall dietary purchases, and to understand the agreement between household purchase and individual intake.

13. References

1. DHSC (2019) Consultation on restricting promotions of products high in fat, sugar and salt [Department of Health and Social Care., editor]. London: Assets Publishing

2. National Food Strategy. (2021) *The Plan: National Food Strategy, Independent Review*. London, UK.

3. WHO (2021) Obesity. *Health Topics*. https://www.who.int/health-topics/obesity#tab=tab_1 (accessed 10.09.2021

4. NHS (2019) Obesity. https://www.nhs.uk/conditions/obesity/ (accessed 10.09.2021

5. Sahoo K, Sahoo B, Choudhury AK *et al.* (2015) Childhood obesity: causes and consequences. *J Family Med Prim Care* **4**, 187-192.

6. Jacob JJ, Isaac R (2012) Behavioral therapy for management of obesity. *Indian J Endocrinol Metab* **16**, 28-32.

7. Hill JO, Wyatt HR, Peters JC (2012) Energy balance and obesity. *Circulation* **126**, 126-132.

8. Di Figlia-Peck S, Feinstein R, Fisher M (2020) Treatment of children and adolescents who are overweight or obese. *Curr Probl Pediatr Adolesc Health Care* **50**, 100871-100871.

9. Ventura AK, Worobey J (2013) Early influences on the development of food preferences. *Current biology : CB* **23**, R401-408.

10. Scientific Advisory Committee on Nutrition. (2015) *Carbohydrates and Health*. London.

11. Public Health England. FSA (2020) National Diet and Nutrition Survey: Rolling programme Years 9 to 11 (2016/2017 to 2018/2019). . https://www.gov.uk/government/collections/national-diet-and-nutrition-survey

12. HMRC (2018) Check if your drink is liable for the Soft Drink Industry Levy. https://www.gov.uk/guidance/check-if-your-drink-is-liable-for-the-soft-drinks-industry-levy (accessed 13.08.19

13. Pell D, Mytton O, Penney TL *et al.* (2021) Changes in soft drinks purchased by British households associated with the UK soft drinks industry levy: controlled interrupted time series analysis. *BMJ* **372**, n254.

14. Clark SD, Shute B, Jenneson V *et al.* (2021) Dietary Patterns Derived from UK Supermarket Transaction Data with Nutrient and Socioeconomic Profiles. *Nutrients* **13**, 1481.

15. LIDA (2021) What is LASER? Leeds Analytic Secure Environment for Research. https://lida-data-analytics-team.github.io/laserdocs/docs/laser_info/laser.html (accessed 17.09.2021

16. Office for National Statistics. (2021) Census Geography. https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography (accessed 10.09.2021

17. Gale C, Singleton A, Bates A *et al.* (2016) Creating the 2011 area classification for output areas (2011 OAC). *Journal of Spatial Information Science* **12**.

18. Kansal T, Bahuguna S, Singh V *et al.* (2018) Customer Segmentation using K-means Clustering. *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 135-139.

19. Liu C, Largeron C, Zaïane OR *et al.* (2020) A Late-Fusion Approach to Community Detection in Attributed Networks, 300-312.

20. Rossetti G, Cazabet R (2018) Community Discovery in Dynamic Networks: A Survey. *ACM Comput Surv* **51**, Article 35.

21. Paparrizos J, Gravano L (2015) k-Shape: Efficient and Accurate Clustering of Time Series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pp. 1855–1870. Melbourne, Victoria, Australia: Association for Computing Machinery.

22. Hyndman R. AG (2021) Chapter 10. Dynamic Regression Models. In *Forecasting: Principles and Practice 3rd edition*, 3 ed. Melbourne, Australia: OTexts.com/fpp3.

23. DH (2011) Nutrient Profiling Technical Guidance. London: Crown copyright.

24. Rains T, Longley P (2021) The provenance of loyalty card data for urban and retail analytics. *Journal of Retailing and Consumer Services* **63**, 102650.

14. Team biographies and contributions

Challenge leaders

Vicki Jenneson (Principal Investigator) is a PhD researcher in the Data Analytics and Society Centre for Doctoral Training at the Leeds Institute for Data Analytics. Vicki has a background in Nutrition and Public Health and undertakes research exploring the use of supermarket transaction records as a means to monitor diet at the population level. Her experience working with large transaction datasets in collaboration with a retail data partner were beneficial to her role as PI. Vicki worked with the challenge owner to design the challenge, prepare the data and communicate key information to the participants. She led the team and was the key contact for liaison between the challenge organisation, challenge participants, challenge leaders, and the DSG organising team.

Michelle Morris (Investigator) is an Associate Professor in the Faculty of Medicine and Health at the University of Leeds, where she leads the Nutrition and Lifestyle Analytics Team. Michelle's research interests are in the use of novel big data sources to measure diet and lifestyle (particularly physical activity) behaviours. She has played an important role in the development of the formal research partnership between Sainsbury's and the Leeds Institute for Data Analytics, which enabled this challenge to go ahead. Michelle contributed to the DSG in an advisory capacity, offering support to PI Vicki Jenneson in the design, preparation and running of the challenge.

Joel Dyer (Facilitator) is a PhD student at the University of Oxford's Mathematical Institute and Institute for New Economic Thinking, where he uses and develops mathematics to model social systems. As Challenge Facilitator, Joel acted as part-participant-part-leader to enable the participants to exploit their own and each other's skillsets by guiding discussions and by overseeing the team's research.

Data Study Group participants

(Listed alphabetically by forename)

Adriano Matousek is studying for an MPhil in Population Health Data Science at the University of Cambridge, UK. He was responsible for exploring network analysis as a potential clustering approach, with the aim of understanding the interactions between customers and products through temporal community detection. He also contributed to the project by experimenting with classification models to uncover demographic traits in customer clusters and further understanding of the drivers of behaviour. Additionally, he undertook time-series analysis to identify customers who show a pattern over time, based on various behavioural variables.

Inès François, is based at the University of Leeds where she is undertaking an integrated MSc and PhD in Data Analytics and Society. Inès' work investigates children's eating behaviours in primary schools in Leeds, using automatically collected food data. Thanks to her skills in data and in food research (food industries, biochemistry and molecular biology and physiology and psychology food choice determinants), Inès contributed to the introduction to report, elaborated the data chart provenance and built multiple linear regression models to investigate drivers of response to the levy.

Joanna Tumelty is a PhD researcher in Applied Mathematics at the University of Leeds, UK. Joanna contributed to the project primarily through customer segmentation via variations in customer purchases per category. She took the lead on the basket analysis section of the report.

Maja Omieljaniuk works as a Food and Nutrition Data Scientist at the Quadram Institute in Norwich, UK. Here she combines her interests in food and data science and uses data to improve nutrition and health. Maja contributed to this project by using her domain knowledge to prepare data for further work and support team members with data understanding. She created classification algorithms for assigning customer to clusters based on their demographics, and explored how consumers of a single brand were affected by the levy. Maja particularly enjoyed working on the case study and wishes she had more time to explore this.

Michael Stephens completed his PhD at Queen Mary University of London, UK, where he studies spatial analysis within environmental science. His contribution to the project focused on segmentation of customers into different profile categories dependent on their shopping behaviours. Through this process, Michael was able to identify customers who altered their spending habits as the SDIL came into effect.

Rosalind Martin works at the University of Leeds as a Data Science Intern at the Leeds Institute for Data Analytics. With a background in Geography and Geographic Information Systems, Rosalind contributed to the exploratory data analysis phase through the production of maps. She also worked on consumer clustering based on purchase behaviours, and contributed to the implementation of the K-means clustering algorithm.

Sijin Wu is a student at the University of Leeds, based in the Institute for Transport Studies. He contributed to the project by taking the lead on the time-series clustering of customers based on their purchase behaviours. This involved Sijin learning and applying a new method, as well as passing on his knowledge to other DSG participants.

Soon Yung Low is originally from Malaysia and is currently pursuing an MSc in Applied Social Data Science with a focus on textual analysis at the London School of Economics. In this project, Soon led on the feature engineering using text mining, as well as data manipulation. Additionally, he conducted exploratory data analysis and clustering of customers using time-series methods.

Wingyan Yip is originally from Hong Kong and now works as a Business Intelligence Analyst at Soldo Ltd in London. Wingyan took a lead role in exploratory data analysis, where she explored demographic and product trends. She also ran a time series regression and engineered features that describe consumption behaviour to support other team-mates’ analysis.

Appendices

Appendix 1. Data Dictionary – Transaction data file

Data field name	Description
Hashed_CustID	Unique Customer number for each loyalty card holder. Note that a loyalty card may represent an individual or it may represent a household (household size unknown)
date	Month and year of purchase
SKU	Stock Keeping Unit – Unique product ID
sku_desc	Product name (may contain additional information such as brand, volume etc)
cat	Product category (assigned by retailer)
subcat	Product sub-category (assigned by retailer)
item_weight	Weight of product (unit) in grams. Here we assume 1g = 1ml (this is the case for water, we do not account for density)
item_kcal	Energy density of product, calories (kcal)/100ml of product
item_sugar	Grams of total sugar /100ml of product
bev_items	Number of units of an item purchased per customer per month
bev_spend	Spend (GBP £ sterling) on an item per customer per month (number of units purchased x price per unit)
bev_weight	Weight of an item purchased per customer per month (item_weight x bev_items)
bev_kcal	Number of calories purchased per item (item_kcal x bev_items)
bev_sugar	Amount of sugar (grams) purchased per item (item_sugar x bav_items)
prop_all_prods_kcal	Proportion of total food and beverage calories from the product per customer per month
prop_all_prods_sugar	Proportion of total sugar from food and beverages coming from the product per customer per month
sdil	Flag indicating eligibility for soft drinks industry levy. Blank = ineligible beverage category (e.g. milk) No = Eligible beverage category but total sugar below 5g/100ml threshold SDIL1 = Eligible for low levy threshold (total sugar ≥5g <8g/100ml) SDIL2 = Eligible for high levy threshold (total sugar ≥8g/100ml)

Appendix 2. Data Dictionary – Customer data file

Data field name	Description
Hashed_CustID	Unique Customer number for each loyalty card holder. Note that a loyalty card may represent an individual or it may represent a household (household size unknown)

Gender	M = Male
	F = Female
	U = Unknown ("Prefer not to say")
	Blank = customer did not answer this question

Age band (years)	0 - 16 17 – 29 30 - 44 45 – 64 65+
oa11	Output area of residence for customers (small neighbourhood geography). Derived from customer postcode given at loyalty card sign up. Smallest neighbourhood level census geography.
oac11	Output area classification (developed in 2011). A hierarchical geodemographic classification describing neighbourhoods based on census characteristics of the people who live there. Guide to classification found in N/Incoming/2019-03-13/2011 OAC Clusters and Names Excel v2 (sheet 1)
imd	Index of multiple deprivation for LSOAs. A national ranking indicating relative affluence of areas. Based on income, education, employment, health, barriers and living environment domains. Ranked from 1 – 32844 (1 = most deprived LSOA, 32844 = least deprived LSOA in England). Can be used to construct IMD deciles (decile 1 = most deprived, decile 10 = least deprived LSOAs in England).
ru11ind	Rural/urban index 2011 – index describing the urbanity of areas. Guide to classification found in N/Incoming/2019-03-13/2011 OAC Clusters and Names Excel v2 (sheet 2)
LSOA11CD	Lower layer super output area 2011 code. Neighbourhood census geography containing Output Areas.
LSOA11NM	Lower layer super output area 2011 name.
MSOA11CD	Middle layer super output area 2011 code. Neighbourhood census geography containing LSOAs.
MSOA11NM	Middle layer super output area 2011 name.
LAD11CD	Local Authority District 2011 code
LAD11NM	Local Authority District 2011 name
Imd_Decile	Categorical variable showing which deprivation decile (whole of England) each LSOA belongs to. Calculated from IMD rank = (IMD/3285)+1 Decile 1 = most deprived Decile 10 = least deprived

turing.ac.uk
@turinginst