Data Analysis
On Tesco Dataset

Outline of contents

Task 1

Overview
of the Tesco dataset

Task 2

Exploratory Data Analysis
And data visualisation

Task 3

Combine
With income dataset

Overview of the Tesco dataset

Overview of the dataset

- Tesco Dataset 1.0
- Collected from Tesco’s customers who own Clubcard

- Data collected from Tesco stores in the Greater of London area in 2015
- The greater of London can be divided into 33 areas, using borough scale

- 33 rows and 202 columns, representing each borough area and all features, respectively
- Features can be divided into 3 groups, including fields related to nutrition, fields related to food categories, and others

Nutrition Related columns

1 *Grouped by weight*

Description: **Weight of nutrition in average products** measured in grams along with their percentiles, standard deviation, and confidence interval at 95% for each nutrition

Example: **Fat: 8** ,means there are eight grams of fat in average product

2 *Grouped by energy*

Description: **Energy of each nutrition in average products** measured in kcals along with their percentiles, standard deviation, and confidence interval at 95% for each nutrition

Example: **energy_fat: 76**, means fat in average products contain 76 kcals

3 *Grouped by proportion of energy*

Description: **Proportion of energy from each nutrition**

Example: **f_energy_fat: 0.25**, means fat contributes 25% of total calories in average products

Nutrition Related columns

4 *Entropy of all nutrition by weight*

Description: Represent **how variety of nutrition** by their weights

Example: **h_fat_weight:** the higher entropy, the more variety

5 *Entropy of all nutrition by calories*

Description: Represent **how variety of nutrition** by their calories

Example: **h_fat_calories:** the higher entrpy, the more variety

Food categories Related columns

Food Categories

Food

- Dairy
- Eggs
- Fats_oils
- Fish
- Fruit and vegetable
- Grains
- Red meat
- Poultry
- Readymade
- Sauces
- Sweets

Beverage

- Water
- Wine
- Spirits
- Beer
- Soft drinks
- Tea and coffee

1 ***Grouped by proportion of numbers of each category***

Description: **Proportion of number of each food category** purchased

Example: **f_grains: 0.15**, means 15% of total food and beverage purchased are grains related product.

2 ***Grouped by proportion of weights of each category***

Description: **Proportion of weight of each food category**

Example: **f_grains_weight: 0.10**, means 10% of total weight from all food and beverage purchased are from grains related product.

Food categories Related columns

Food Categories

Food	Beverage
Dairy	Water
Eggs	Wine
Fats_oils	Spirits
Fish	Beer
Fruit and vegetable	Soft drinks
Grains	Tea and coffee
Red meat	
Poultry	
Readymade	
Sauces	
Sweets	

3 *Entropy of all categories by number of purchase*

Description: **Represent how variety of food categories** using number of products purchased

Example: **h_items: 3.2**, the higher entropy, the more variety

4 *Entropy of all categories by weights*

Description: **Represent how variety of food categories** using weights of all categories

Example: **h_items_weight: 2.8**, the higher entropy, the more variety

Demographics and other columns

Demographic

1. **Population (Number)**
2. **Gender (Number)**
   - Male
   - Female
3. **Age (Number)**
   - Average age
   - 0 – 17 years old
   - 18 – 64 years old
   - Above 65 years old

Area

1. **Area (Sq.km.)**
   - Total area of each borough
2. **People/sq.km**
   - Total number of people in each borough area
3. **Representative**
   - How much data can represent the population in each area

Transaction

1. **Transaction days**
   - Number of days that transaction occurs
2. **Number of transaction**
   - Total number of transactions
3. **Man day**
   - The accumulated number of all individual customers in different days

Limitation and Bias

Customer Representativeness Limitation

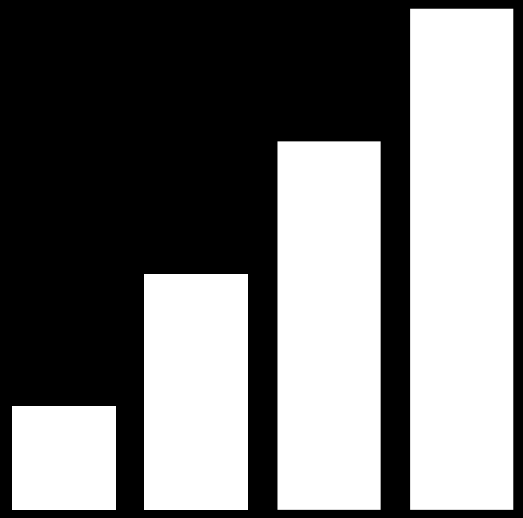
1. Only customers who own Clubcard are included in the dataset
2. Only customers who choose to shop at stores, excluding online customers

Consumption Behaviour Limitation

1. This dataset doesn't include who shop at other chains of grocery store
2. This dataset doesn't reflect who choose to eat at restaurant

Individual Representation Limitation

1. This dataset can only represent at geographical scale, not at individual level

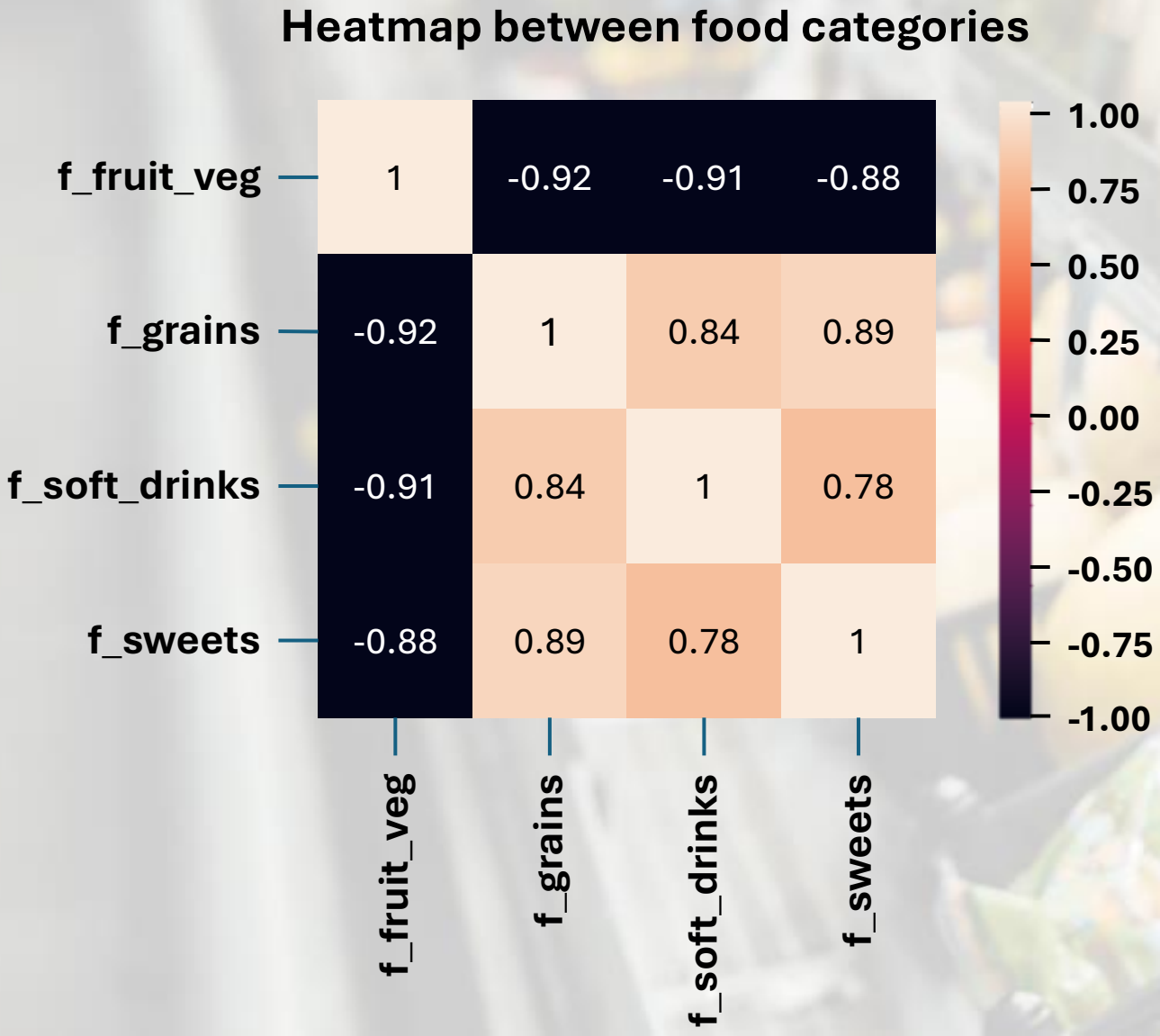


Exploratory Data Analysis and Data Visualisation

Topic 1

Correlation between food categories

Correlation between food categories



Heatmap Correlation

- Four strongest correlations, showing both positive and negative correlation

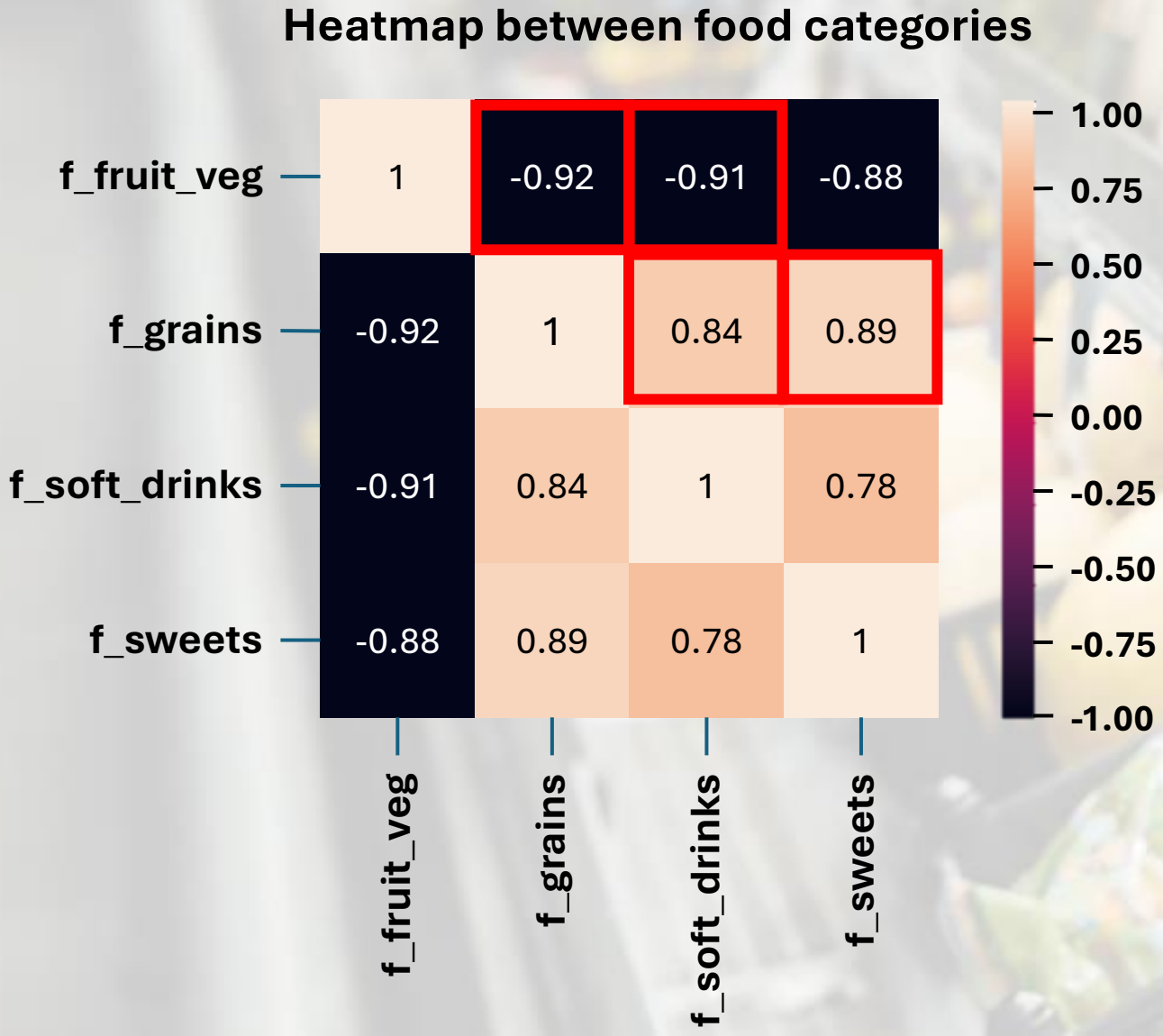
Fruit and vegetable

Sweets

Soft Drinks

Grains

Correlation between food categories



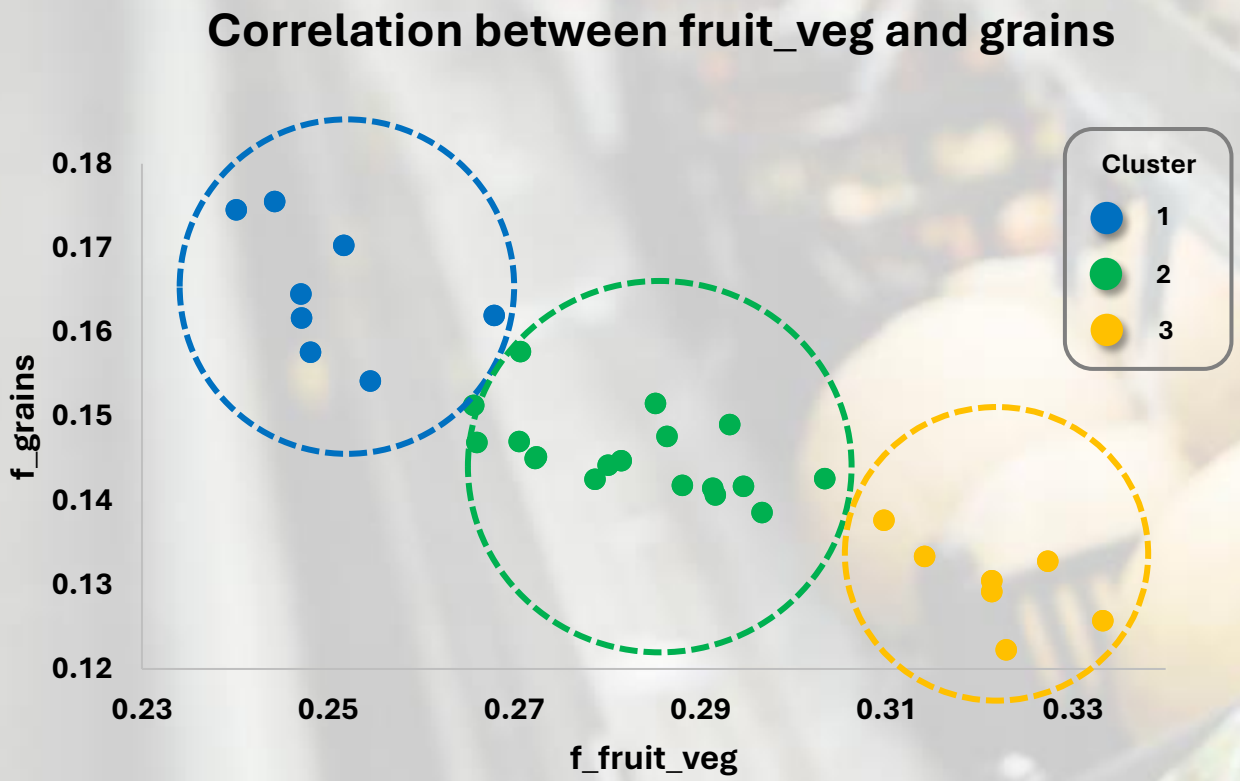
Positive Correlation

- f_grains and f_soft_drinks
- f_grains and f_sweets

Negative Correlation

- f_fruit_veg and f_grains
- f_fruit_veg and f_soft_drinks

Correlation between food categories



K-Means Clustering

- **Classify into 3 groups**

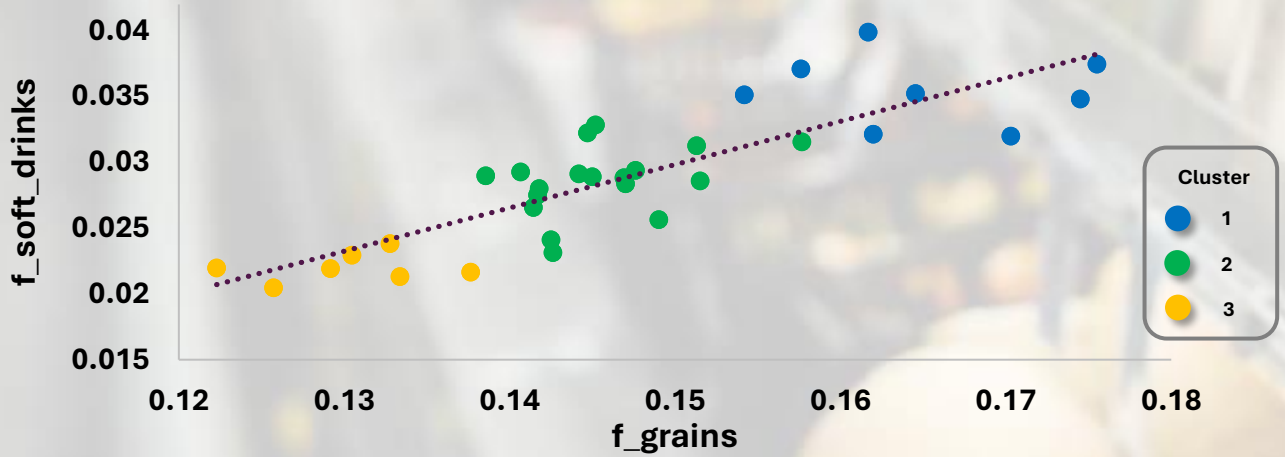
Cluster 1: ***High*** purchase on ***grains***, ***low*** purchase on ***fruit and vegetable***

Cluster 2: ***Medium*** purchase on ***grains***, ***medium*** purchase on ***fruit and vegetable***

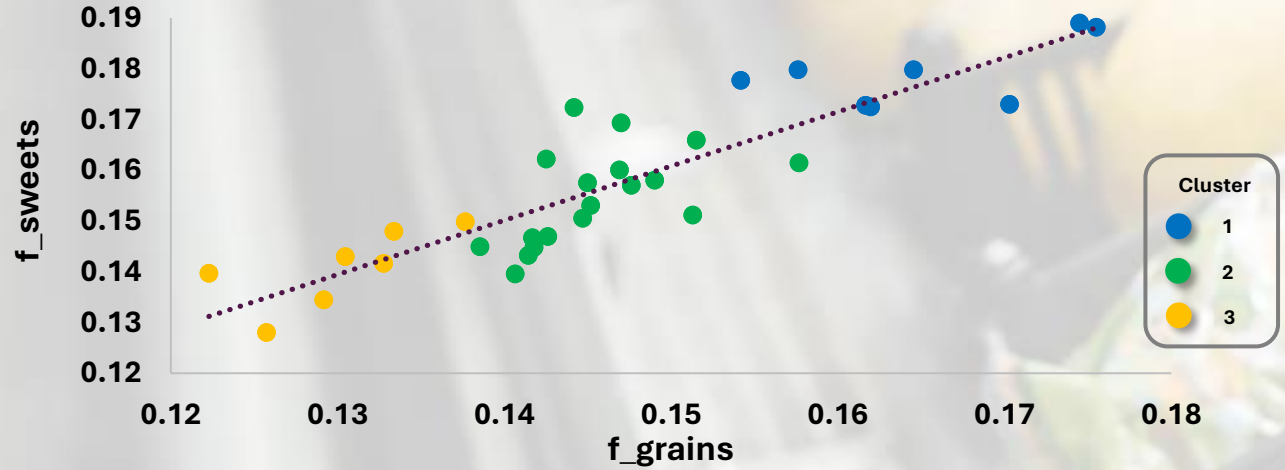
Cluster 3: ***Low*** purchase on ***grains***, ***high*** purchase on ***fruit and vegetable***

Correlation between food categories

Correlation between grains and soft drinks



Correlation between grains and sweet



Positive Correlation

Insight

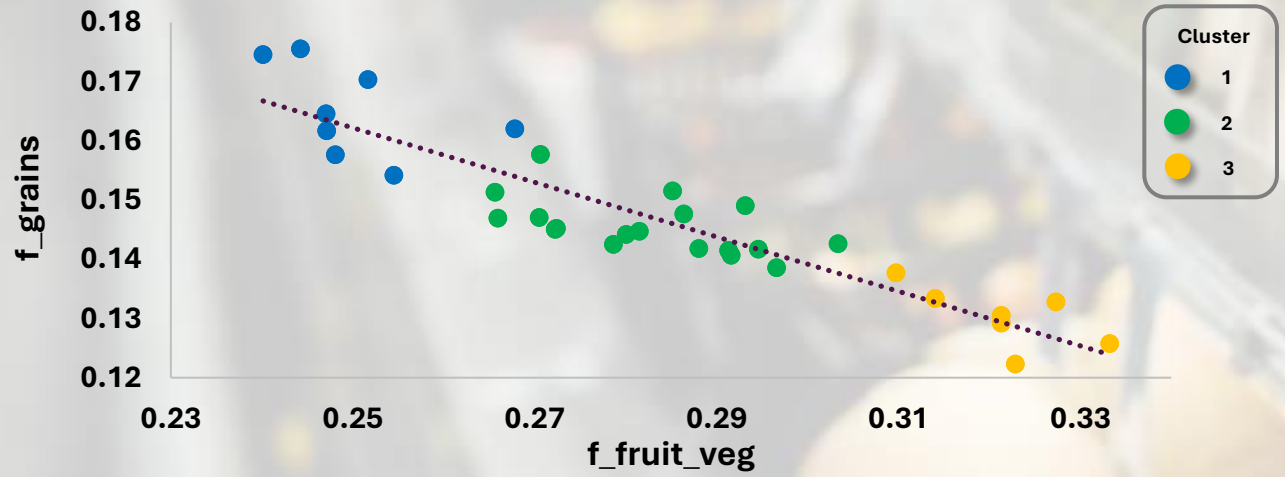
Areas in the cluster 1 consumed more grains related products along with soft drinks and sweets compared to the other two groups, while cluster 3 being the group that consumed the least

Recommendation

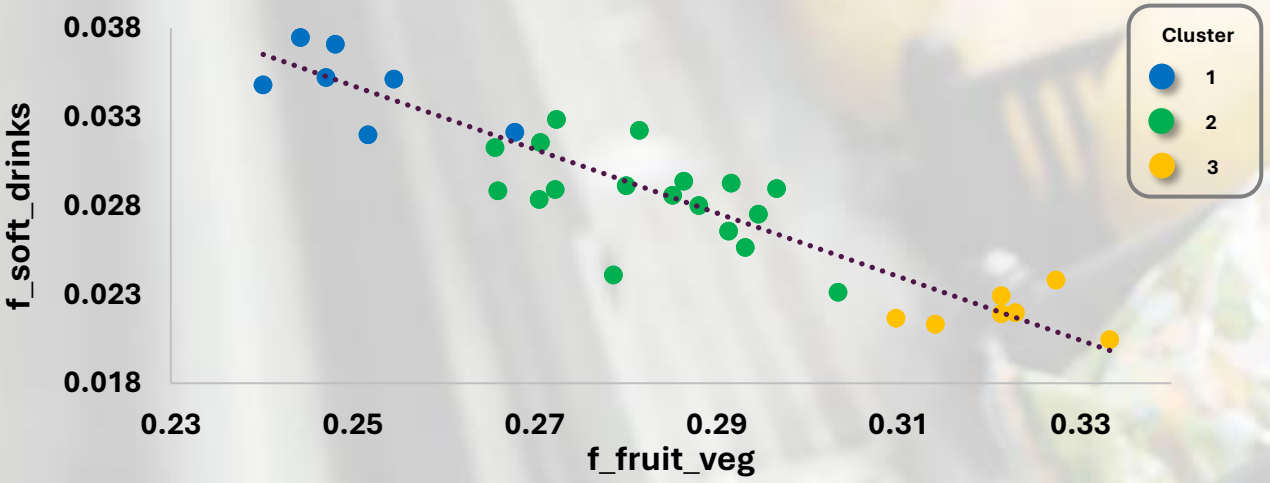
- Allocate more of these three types of product to areas in cluster 1, instead of areas in cluster 3.
- Introduce a promotion for buying grain with soft drinks, or grains with sweets for cluster 1 to boost sales

Correlation between food categories

Correlation between fruit_veg and grains



Correlation between fruit_veg and soft drinks



Negative Correlation

Insight

All areas in the cluster 3 consumed more vegetable and fruits than the other two clusters, while consuming less of soft drinks and grains.

Recommendation

- Allocate more of vegetable and fruits to areas in cluster 3, instead of areas in cluster 1.
- Allocate more space at stores for vegetable and fruits

Maps of Greater London by borough area

Borough area by each cluster

Cluster 1: 8 areas

Cluster 2: 18 areas

Cluster 3: 7 areas

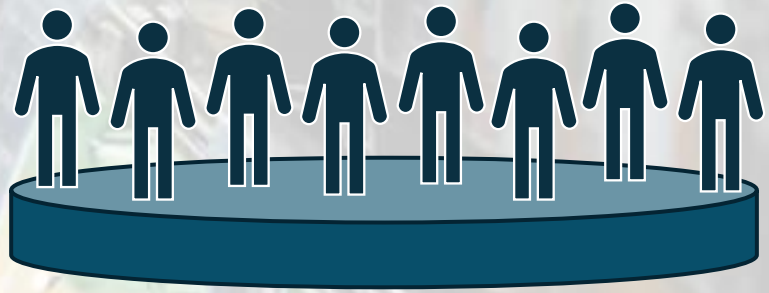
Maps of Greater London by borough area

Cluster 1

Average age: 37 years old

Average total area: 70 sq.km.

Average people per sq.km.: 4607



Maps of Greater London by borough area

Cluster 2

Average age: 35 years old

Average total area: 43.9 sq.km.

Average people per sq.km.: 8173

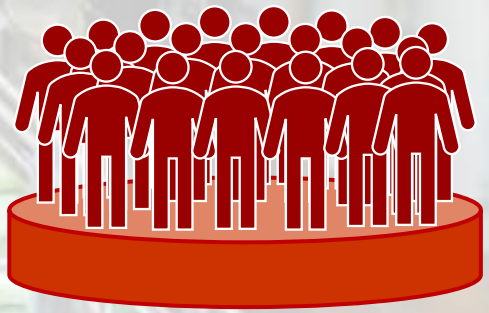
Maps of Greater London by borough area

Cluster 3

Average age: **38** years old

Average total area: **31.2** sq.km.

Average people per sq.km.: **8069**



Conclusion

Insight

- Area in **cluster 1** consumed **more soft drinks, grains, and sweets.**
- Area in **cluster 3** consumes **more vegetable and fruits**
- Larger areas in eastern and southern part of London show a **clear different consuming** behaviour compared to smaller areas in other parts of London

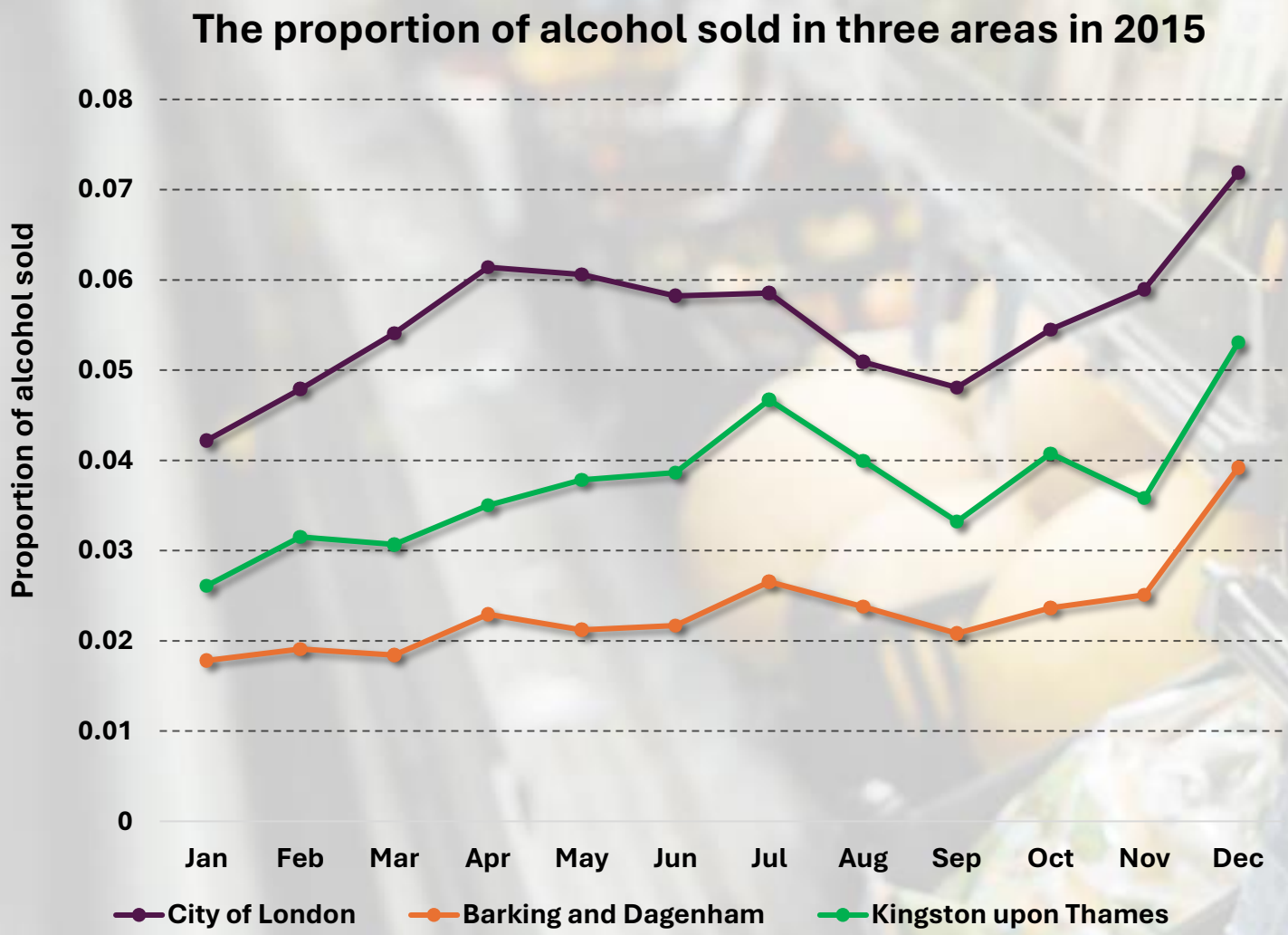
Recommendation

- **Allocate products more wisely** by giving more fruits and vegetable to areas in cluster 3, and soft drink, grains and sweets to areas in cluster 1
- **Conduct a promotion** based on consuming preferences for each cluster to boost sales
- **Conduct a further study** on how the geographic affect consuming behaviour for cluster 1

Topic 2

Seasonal drinking behaviour

Drinking behaviour on Alcoholic beverage



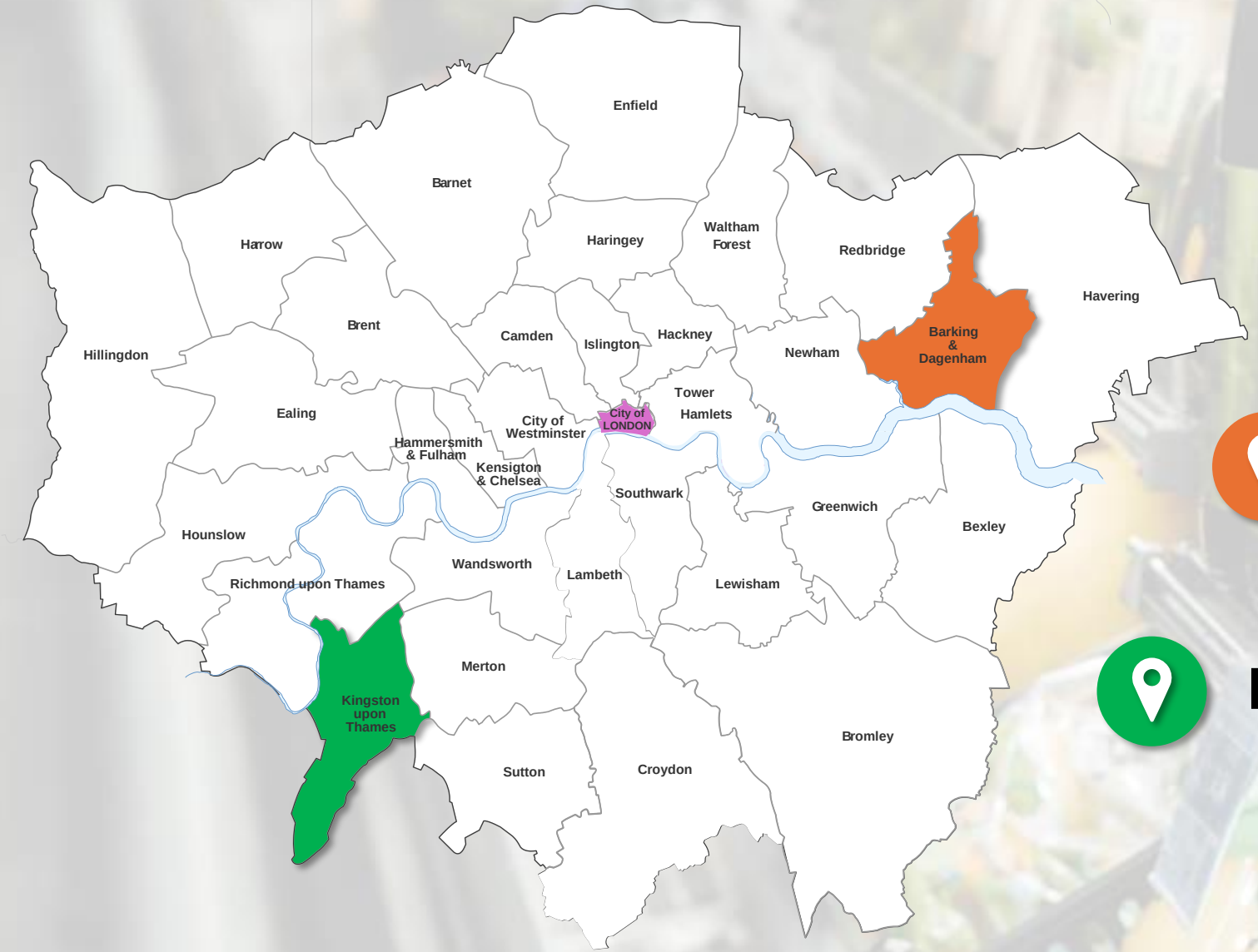
Alcoholic beverage

- Wine
- Beer
- Spirits

Highlighted Area

- City of London
- Barking & Dagenham
- Kingston upon Thames

Maps of Greater London by borough area



City of London

Barking and Dagenham

Kingston upon Thames

Drinking behaviour on Alcohol beverage

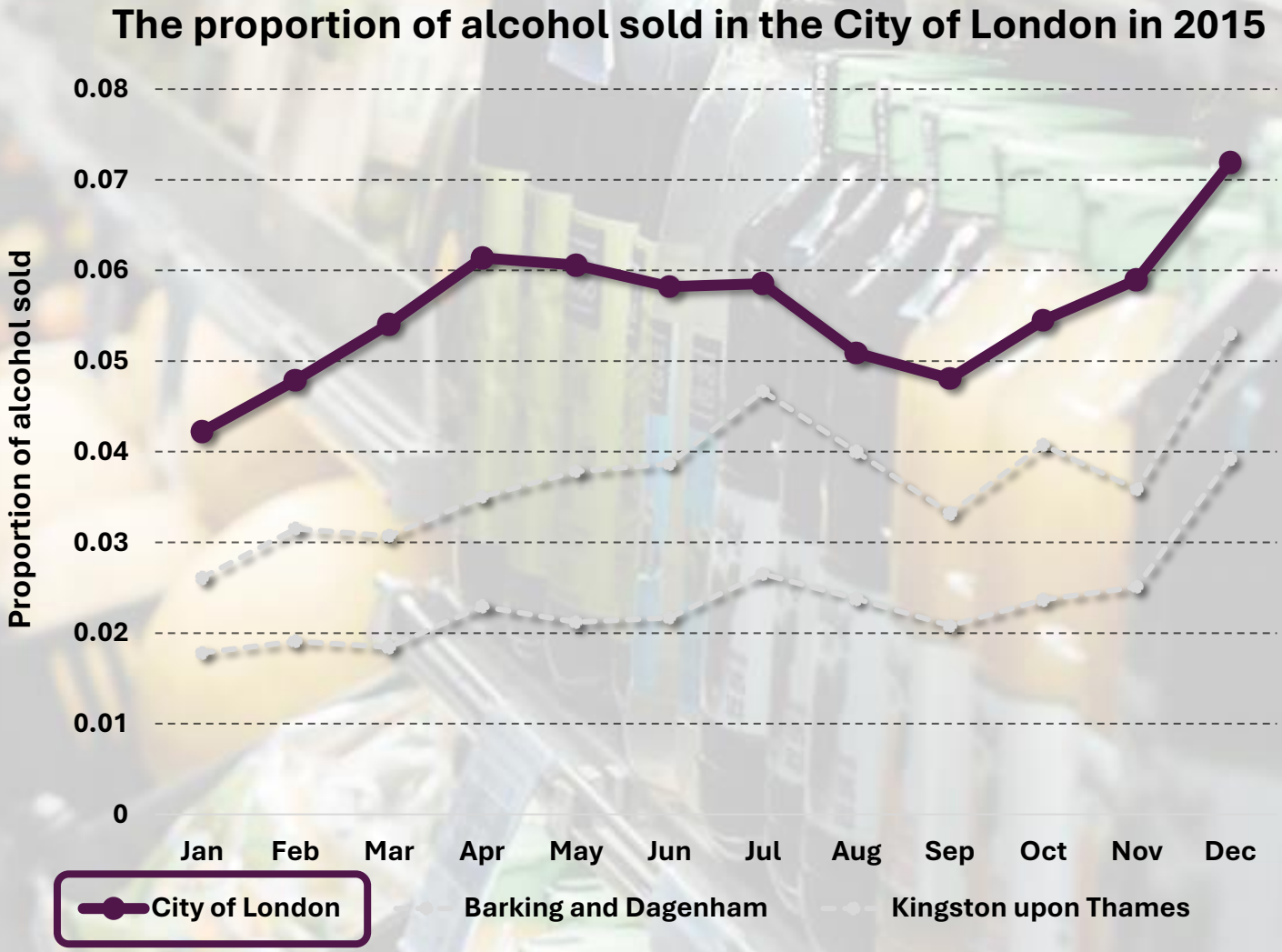
The average age: 43.9 years old

Insight

- ***Oldest average age*** among three areas
- ***Highest alcohol consumption*** area throughout the year
- Constantly ***increase during the first quarter***

Recommendation

- ***Allocate more alcoholic beverage to this area, especially during the first quarter***



Drinking behaviour on Alcohol beverage

The average age: 33 years old

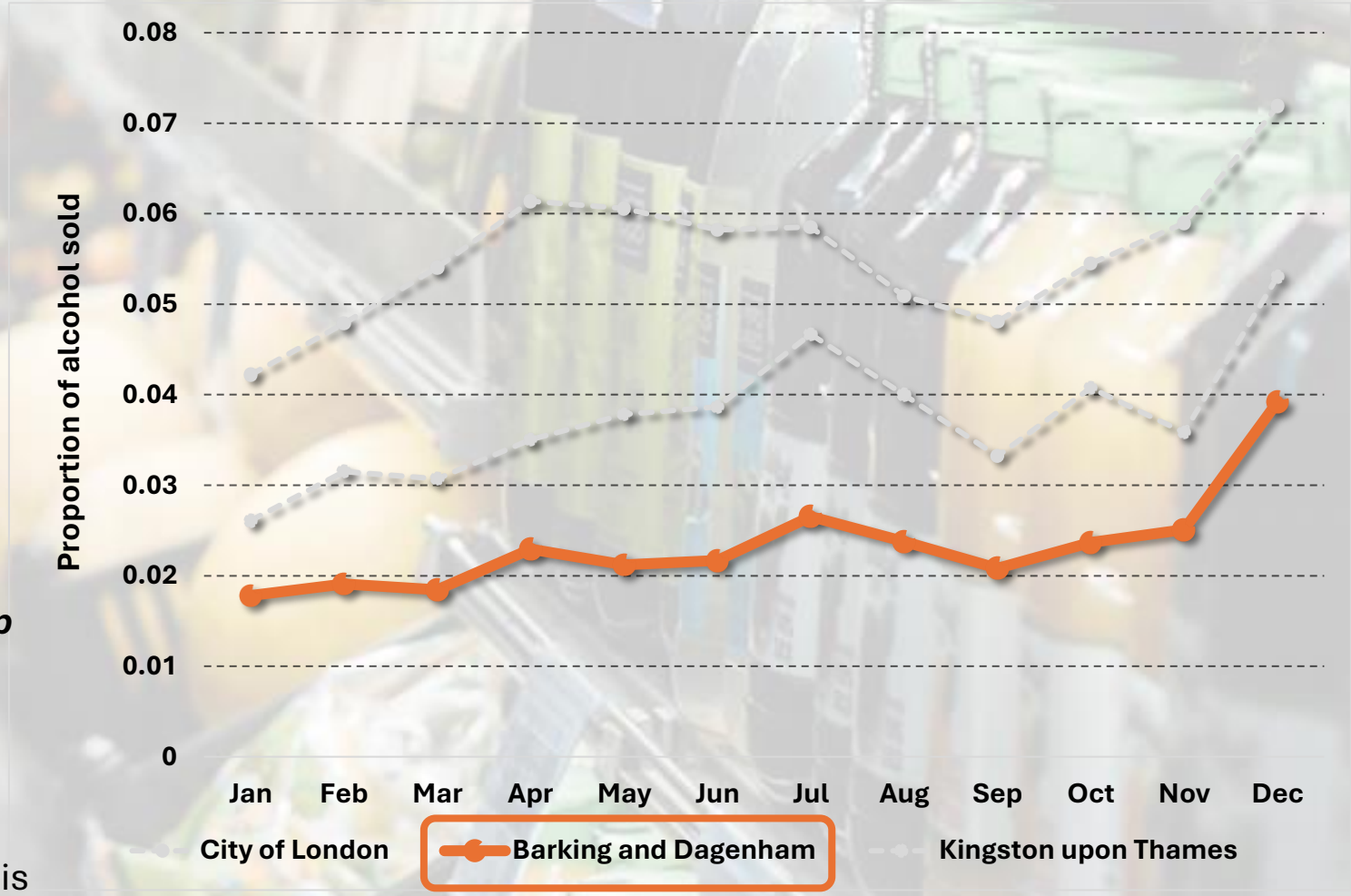
Insight

- ***Youngest average age*** among three areas
- ***Lowest alcohol consumption*** area
- Slightly change over the year despite ***a jump in December***

Recommendation

- Only allocate more alcoholic beverage to this area ***in December***

The proportion of alcohol sold in the Barking and Dagenham in 2015



Drinking behaviour on Alcohol beverage

The average age: 37 years old

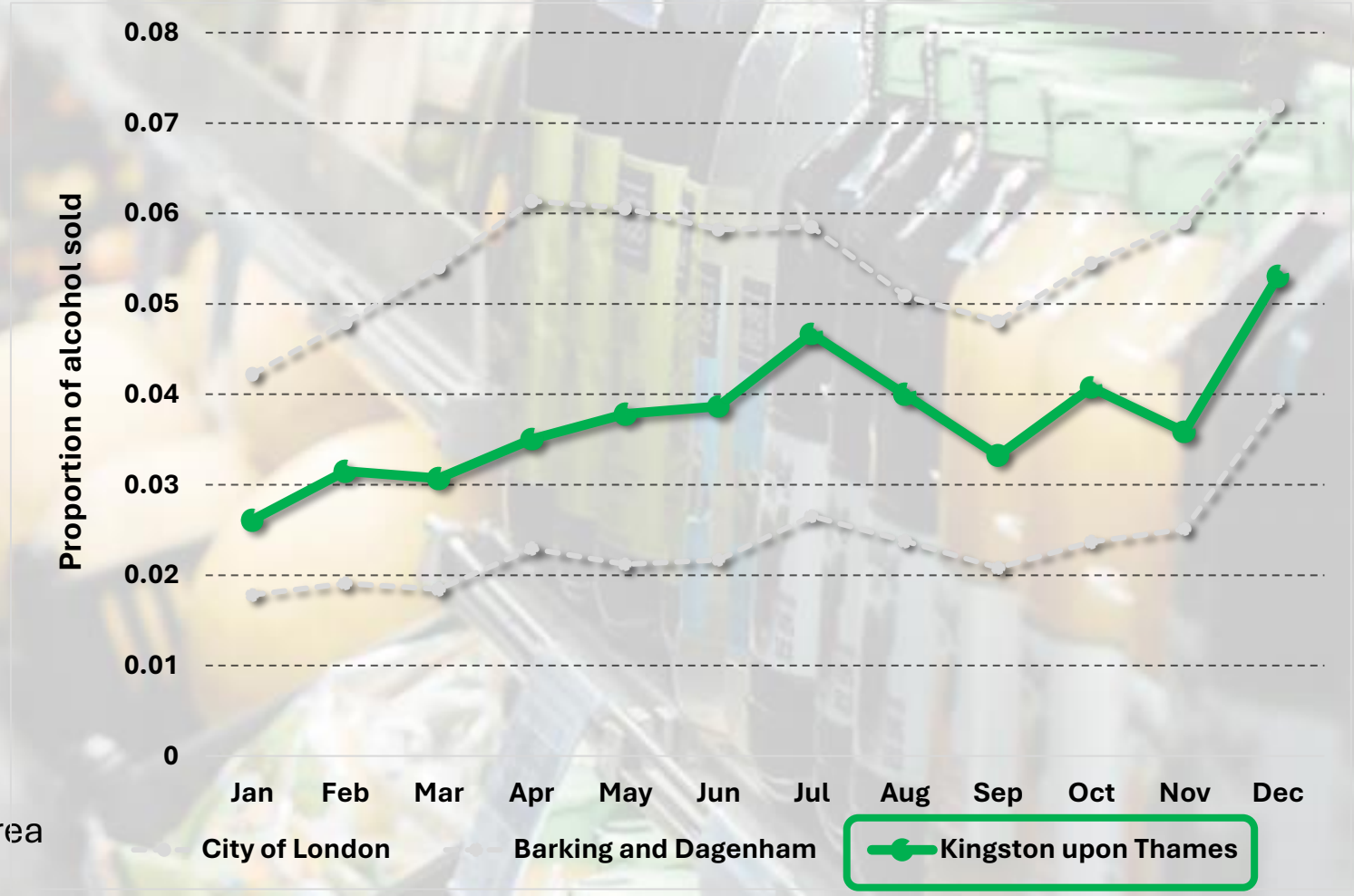
Insight

- Show the ***upward trends*** in the first half of the year with ***a rise in July***
- Notable ***drop in November***

Recommendation

- Allocate ***more alcoholic*** beverage to this area during ***July,*** and ***less during November***

The proportion of alcohol sold in the Kingston upon Thames in 2015



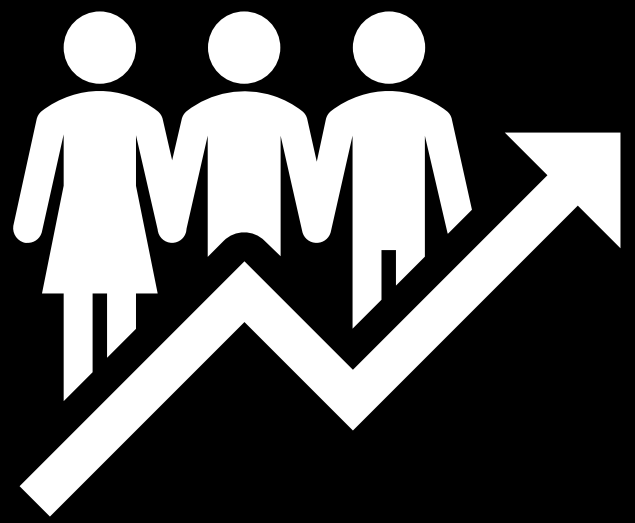
Conclusion

Insight

- **City of London** shows the **highest alcohol consumption** with an outstanding rise during the first quarter
- **Barking and Dagenham** shows the **lowest alcohol consumption** with a jump in December
- **Kingston Upon Thames** shows a clear upward trend for the first half of the year, but **surprisingly drop in November.**
- Younger people are likely to drink less alcohol

Recommendation

- During the first quarter of the year, **allocate more alcohol to City of London** as demand is still existing, while the others show marginal increase.
- During the second quarter, **reduce the number of alcoholic beverage in City of London**, promoting other food categories.
- During the fourth quarter, increase the number of alcohol beverage at all stores as the demand rises



Combine with income dataset

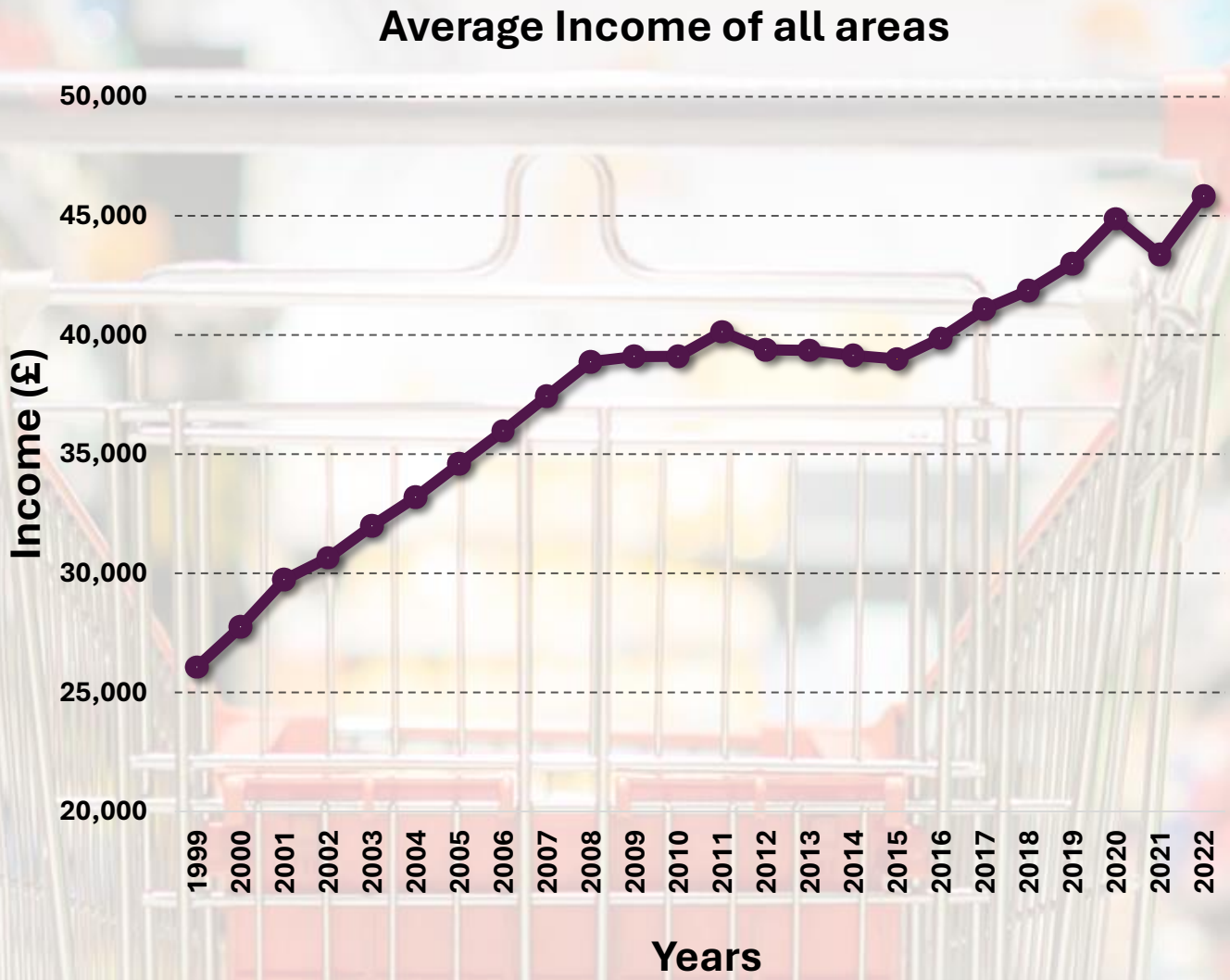
Overview of the dataset

- Sample from full-time worker who work in workplaces
- Not including self-employing worker

- Annual income data for each borough area over 23 years
- Collected from 1999 to 2022
- 23 columns and 33 rows, representing years and areas, respectively

- ‘#’ is used to indicate the statistical unreliability
- Before 2006, the data is represented by mean
- After 2006, the data is represented by median

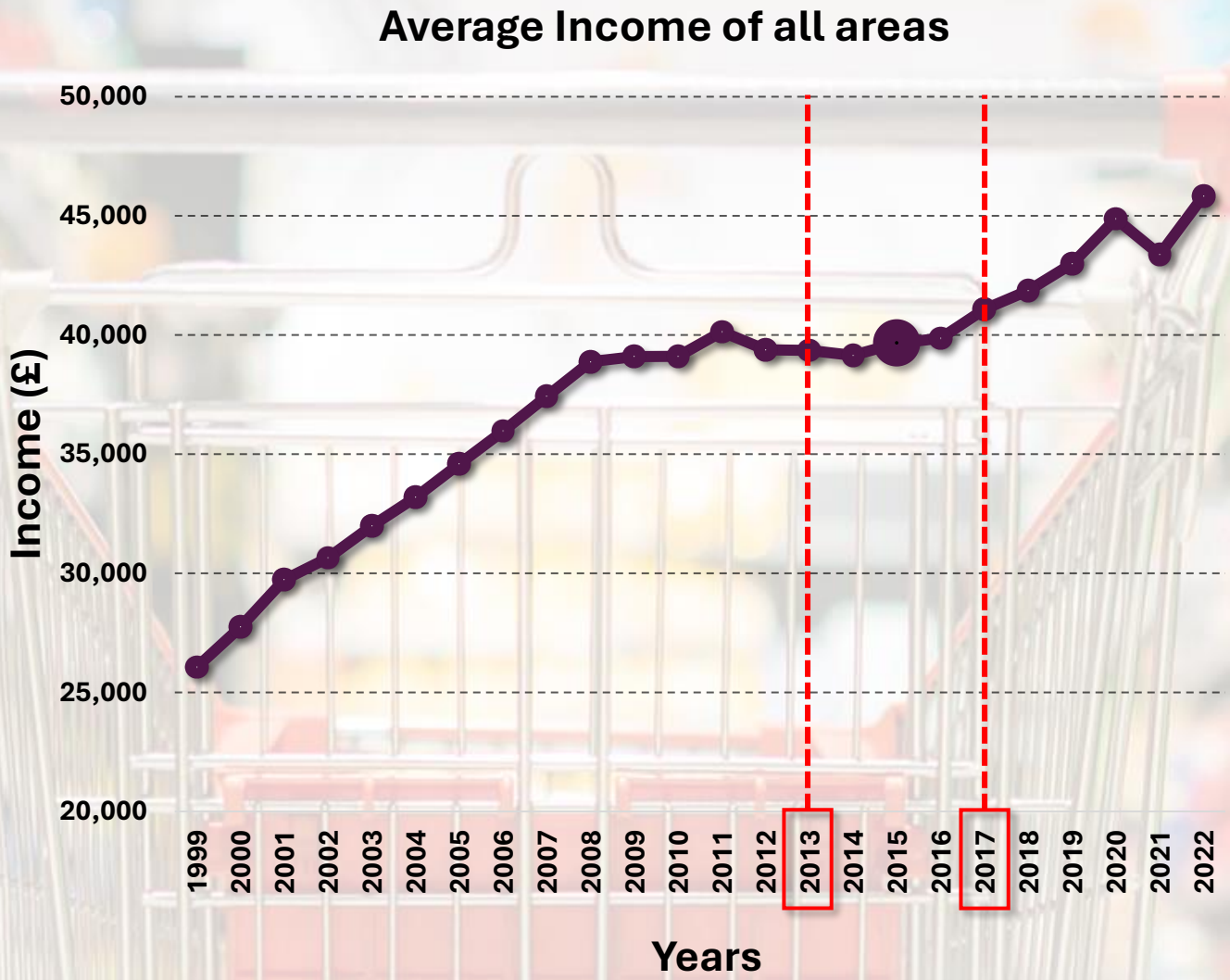
Preliminary Analysis



Key points

- ***Upwards Trend*** throughout the period
- ***Steady increase*** since 1999 until 2008
- Remain ***almost unchanged*** between 2009 and 2015, despite the spike in 2011.
- Increase since 2016 until 2022 with a ***significant drop*** in 2020.

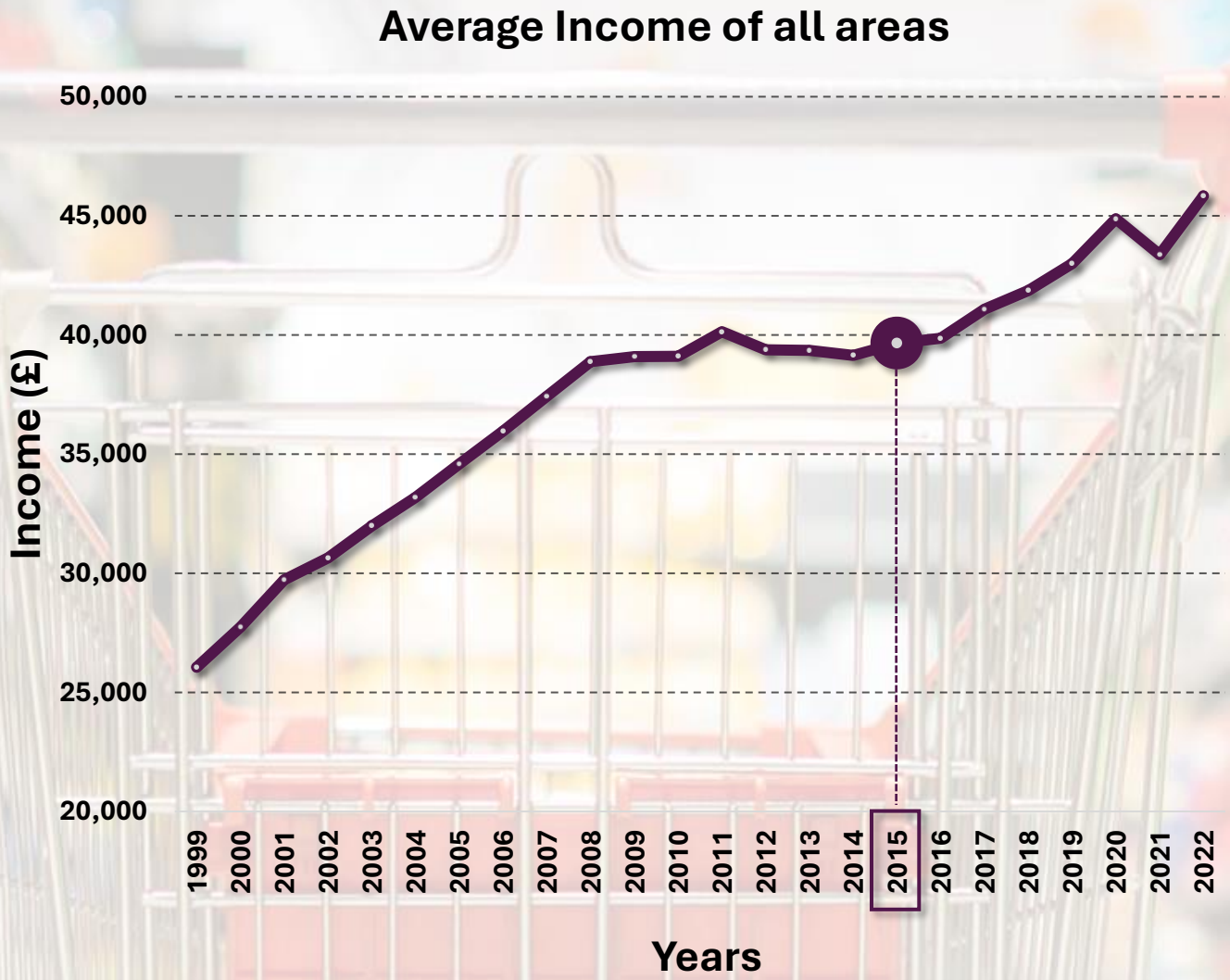
Preliminary Analysis



Preprocessing

- To cover both the steady income before 2015 and the sudden rise after 2015, the data in 2015 used for the analysis is the ***average between 2013 and 2017***.
- ***All missing values*** are replaced by the middle value between the former and later year ***using interpolation***

Preliminary Analysis



Preprocessing

- To cover both the steady income before 2015 and the sudden rise after 2015, the data in 2015 used for the analysis is the ***average between 2013 and 2017***.
- ***All missing values*** are replaced by the middle value between the former and later year ***using interpolation***

Income by each borough area in 2015

Group 1

- *Range:* ***Above 50,000 £***
- *Average income:* ***64,940 £***
- *Total Area:* ***4 areas***

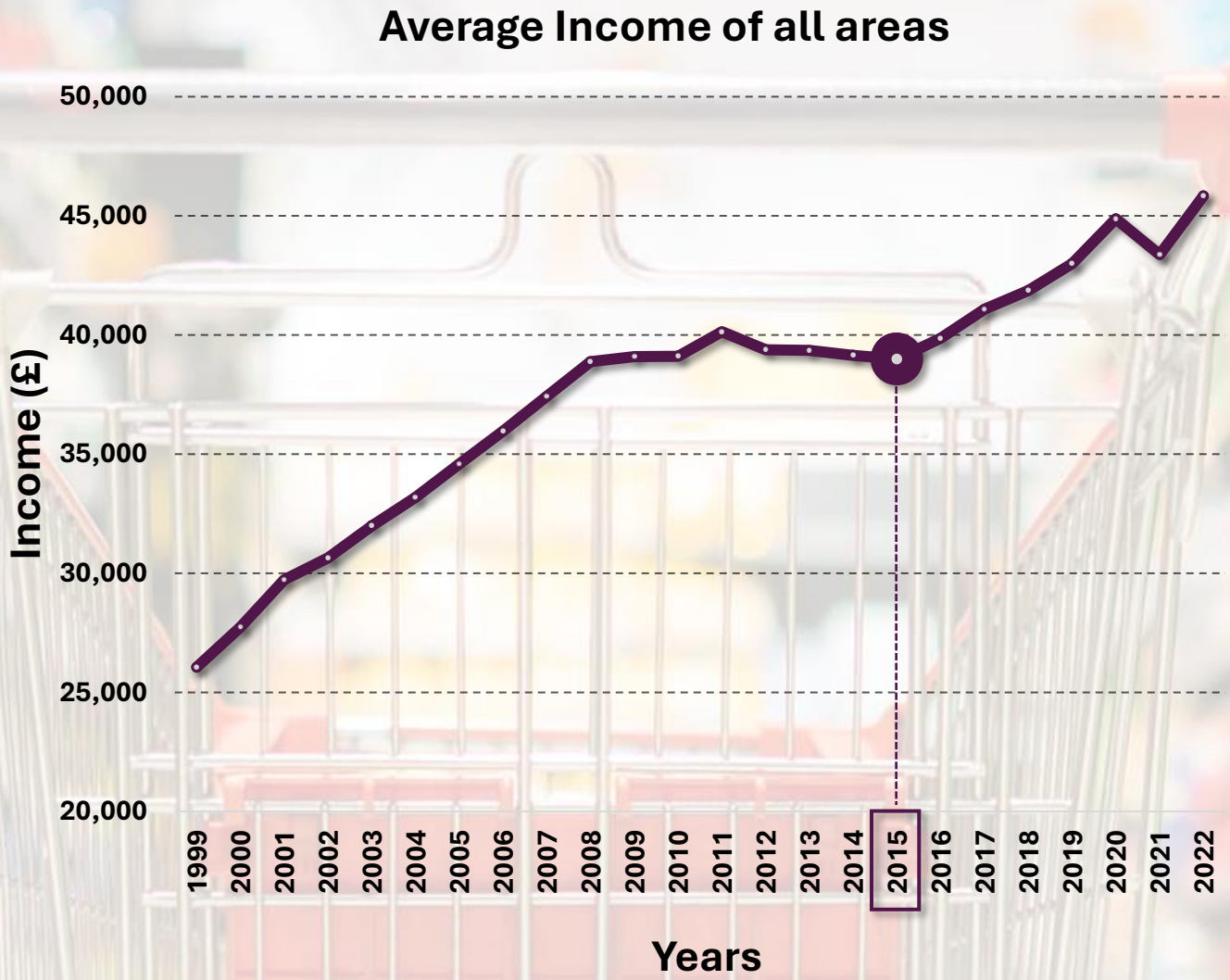
Group 2

- *Range:* ***35,000 – 50,000 £***
- *Average income:* ***39,252 £***
- *Total Area:* ***14 areas***

Group 3

- *Range:* ***25,000 -34,999 £***
- *Average income:* ***33,293 £***
- *Total Area:* ***15 areas***

Preliminary Analysis



Preprocessing

- To cover both the steady income before 2015 and the sudden rise after 2015, the data in 2015 used for the analysis is the ***average between 2013 and 2017***.
- ***All missing values*** are replaced by the middle value between the former and later year ***using interpolation***

Literature Review

TITLE

" Socio-economic dietary inequalities in UK adults: an updated picture of key food groups and nutrients from national surveillance data "

Description

Sampling from **4595 households**

1491 adults were selected to analyse

Use data from **NDNS during 2008-2011**

The paper was **published in 2014**

Sample Group

Divided into **5 groups** based on income

- Below 14,999 £
- 15,000 – 24,999 £
- 25,000 – 34,999 £
- 35,000 – 49,999 £
- Above 50,000 £

Objectives

To show the socio-economic differences in **diet based on household income** by focusing on **three food groups**, including

- Fruit and vegetables
- Red and processed meat
- Oily Fish

Literature Findings

Fruit and Vegetable

"The lowest-income participants consumed 97·1 g/d fewer fruit and vegetables than those with the highest incomes, with an increase in consumption across the income groups."

Red and processed meat

"Participants in the lowest-earning households consumed 15·7 g/d more red and processed meat than the highest-earning households"

Correlation between food purchased and household incomes

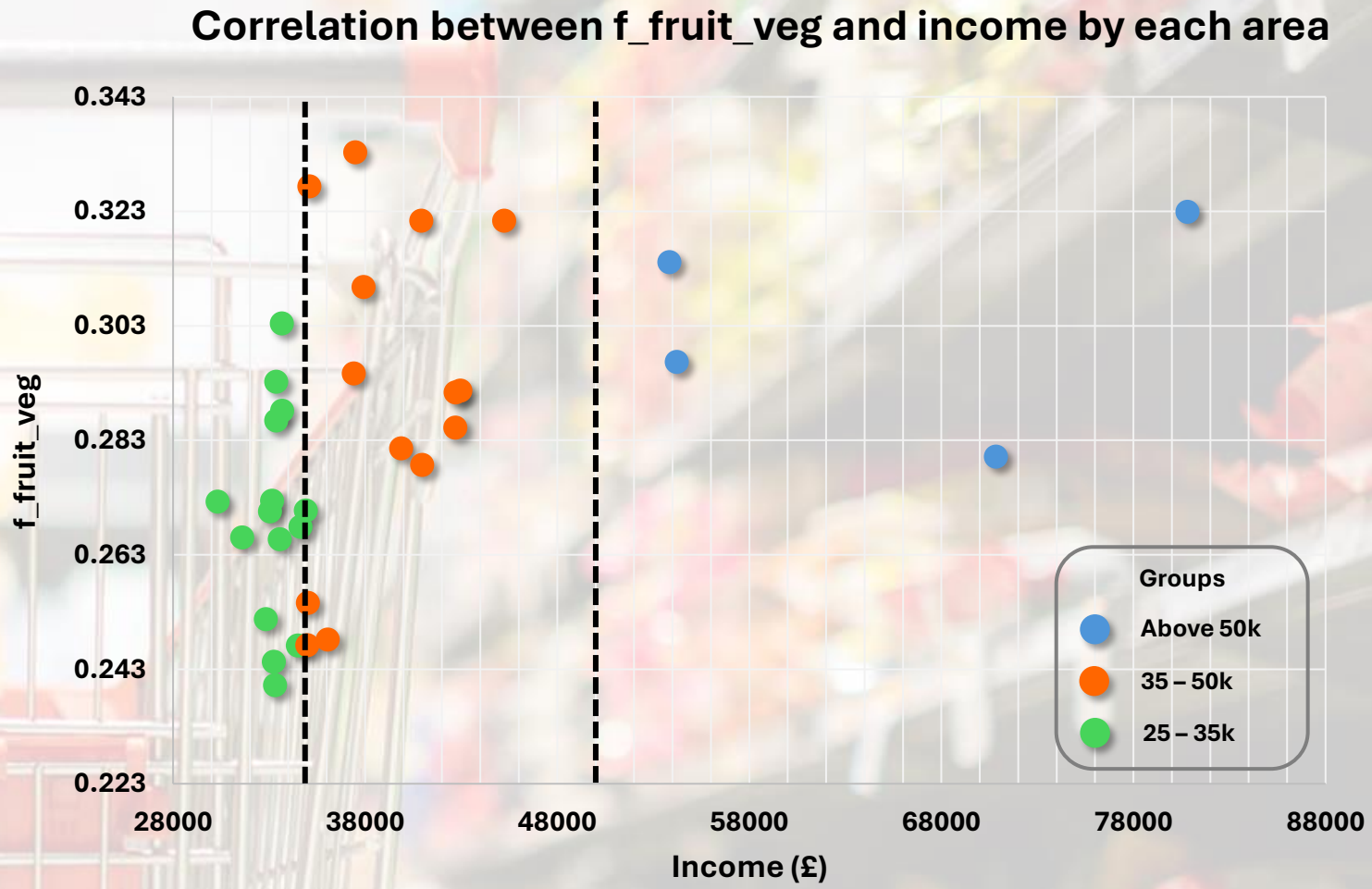
Fruit and Vegetable

Statistics

- **Correlation: 0.4183**
- **P-value: 0.0153**

Conclusion

- **Summary:** It **aligns** with the conclusion of the literature as the p-value is below 0.05, which is **statistically significant**.
- **Reason:** Because higher income household might **be able to afford** additional food, which are **fruits and vegetable.**



Correlation between food purchased and household incomes

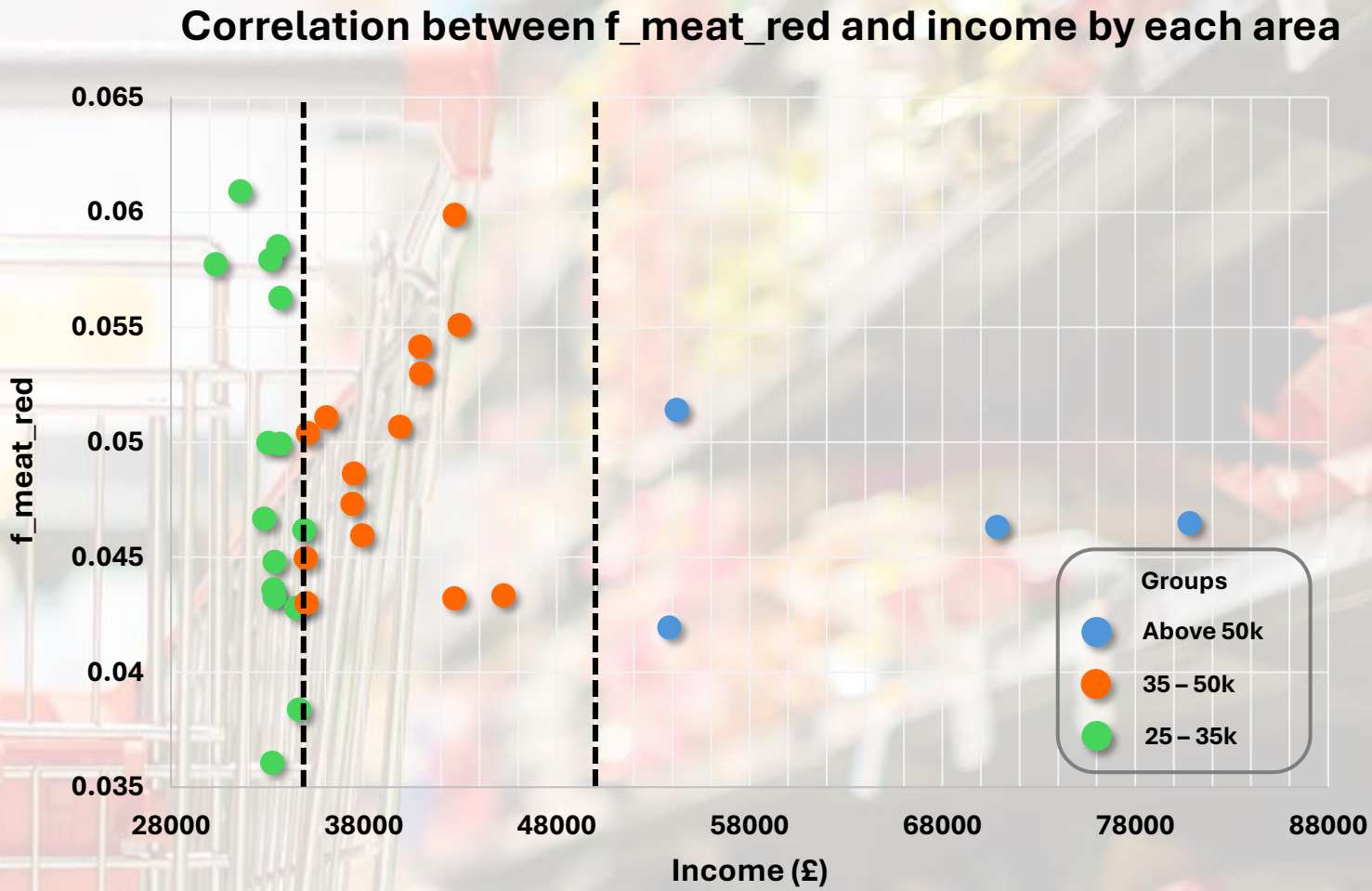
Red and processed meat

Statistics

- **Correlation: -0.1128**
- **P-value: 0.5319**

Conclusion

- **Summary:** Despite a **high value of p-value and a weak correlation, it cannot be concluded** as the lowest income group in this dataset is different from the one in the literature.
- **Reason:** The lowest income group in this dataset earned between **29,000 and 35,000£,** while the lowest group in the literature made **below 14,999£**




Thank you !!