5011CEM

Big Data Programming Project

Part of Coursework Report (Descriptive Analysis and Cluster Analysis)

Name: Chan Khai Shen

Student ID: 11323943

Contents

Chapter 1: Introduction

Overview

The project is the about the study of the distribution of food and drinks purchases of Londoners. The study was done based on the data provided by Tesco Grocery 1.0, a large-scale dataset about grocery purchases of Londoners in 2015 at Tesco outlets. The study can be divided into two parts, descriptive analysis and cluster analysis. Descriptive analysis was to illustrate the average food and drinks purchase based on category, food purchase based on category and food purchase by weight based on category in all lower super output areas (LSOAs), as well as the range of the fraction of food purchase by weight of each category in all LSOAs. Cluster analysis was to classify LSOAs based on the patterns of food purchase of each category by weight. After classifying the LSOAs into clusters, the food purchase of each category by weight of the LSOAs in each cluster was analysed using boxplots.

Why some categories are combined

The Tesco Grocery 1.0 dataset comes with 17 food and drinks categories. However, some categories only stand for a very small fraction. So, to make the smaller categories to look more obvious and more comparable to other categories on the pie chart, some categories were combined. The following is the list of affected categories:

1. f_oils_sauces (combination of f_fats_oils and f_sauces)
2. f_dairy_eggs (combination of f_dairy and f_eggs)
3. f_fish_meat (combination of f_fish, f_meat_red and f_poultry)
4. f_alcoholic_drinks (combination of f_spirits, f_beer and f_wine)
5. f_non_alcoholic_drinks (combination of f_soft_drinks, f_tea_coffee and f_water)

Report structure

Chapter 2: Data analysis. Describes the data analysis part, i.e. descriptive analysis and cluster analysis.

Chapter 3: Findings and discussion. Describes the findings from cluster analysis.

Chapter 4: Conclusion. Concludes the project.

Chapter 5: Bibliography.

Chapter 2: Data analysis

Objective

The data analysis is related the food and drinks purchase of different lower super output areas (LSOAs) based on the data from Tesco Grocery 1.0. The objective of the data analysis is:

1. To study the eating habit of people based on the data by:
   (a) splitting all LSOAs into different clusters based on food purchase by weight.

Descriptive analysis: Average food and drinks purchase of each food category

The average food and drinks purchase by category was visualized using 3 pie charts depicting food and drinks purchase, food purchase and food purchase by weight.

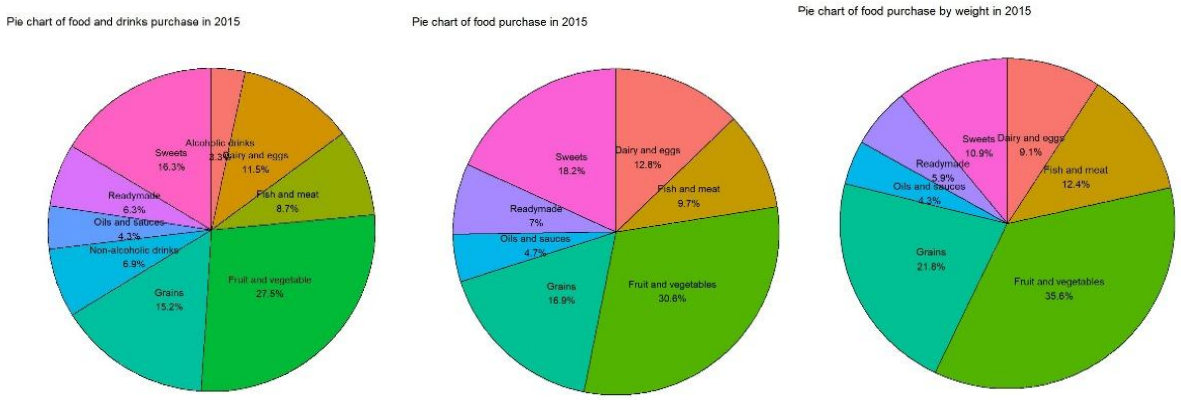

Diagram 1. **Pie charts of average food and drinks purchase by category of all LSOAs in 2015. Food and drinks by count (*left*). Food by count (*centre*). Food by weight (*right*).**

The biggest portion of food and drinks purchase is fruit and vegetable, which accounts for almost one third of all food and drinks purchases (food and drinks: 27.5%; food: 30.6%; food by weight: 35.6%). In general, fruit and vegetable, sweets, fish and meat, dairy and eggs, and grains, are the top five categories of food and drinks purchase. If looking at count, the second to fifth biggest categories are sweets, grains, dairy and eggs, and finally fish and meat. However, if looking at weight, the second to fifth biggest categories are grains, fish and meat, sweets, and

finally dairy and eggs. These four categories jointly represent about half of all food and drinks purchases (food and drinks: 51.2%; food: 57. 6%; food by weight: 54.2%).

The smallest portion of food and drinks purchase is alcoholic drinks (3.3%), followed by oils and sauces (4.3%) and readymade (6.3%). Among all food and drinks purchases, only 10.2% are drinks and the other 89.8% are food.

Descriptive analysis: Statistical distribution of fraction of each food category

The statistical distributions of different food categories are visualised and compared on a boxplot.

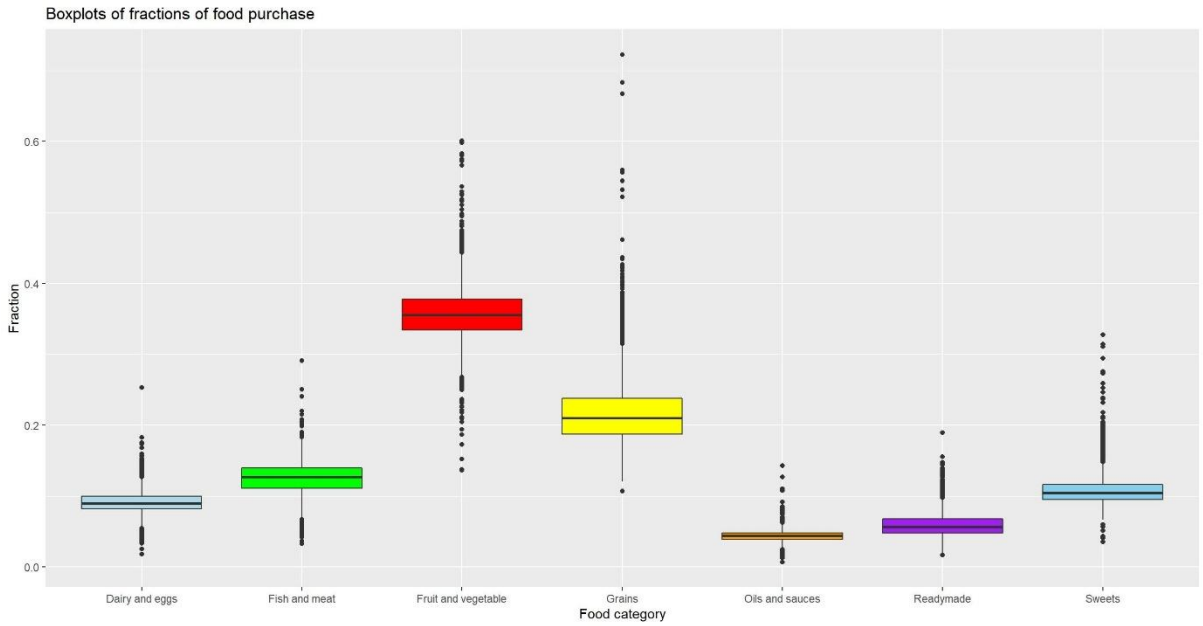

Diagram 2. **Boxplots of fraction of food purchase by weight.**

Among all categories, fraction of grains purchase has the longest range whereas fraction of oils and sauces has the shortest range. In general, fraction of fruit and vegetable has the highest value, followed by grains, fish and meat, sweets, dairy and eggs, readymade and oils and sauces.

Category	Mean	Median	Maximum	Minimum	First quartile	Third quartile
Fruit and vegetable	0.3562	0.3551	0.6006	0.1360	0.3337	0.3776
Grains	0.2177	0.2099	0.7225	0.1071	0.1872	0.2380
Sweets	0.1085	0.1045	0.3271	0.0349	0.0953	0.1165

Fish and meat	0.1266	0.1108	0.2909	0.0323	0.1108	0.1397
Dairy and eggs	0.0911	0.0817	0.2533	0.0181	0.0817	0.0998
Oils and sauces	0.0435	0.0437	0.1428	0.0071	0.0389	0.0482
Sweets	0.1085	0.0983	0.3271	0.0350	0.0953	0.1165

Table 1. **Statistical distribution of fraction of food purchase by weight.**

The statistical distribution of each food category is visualised in details using histograms.

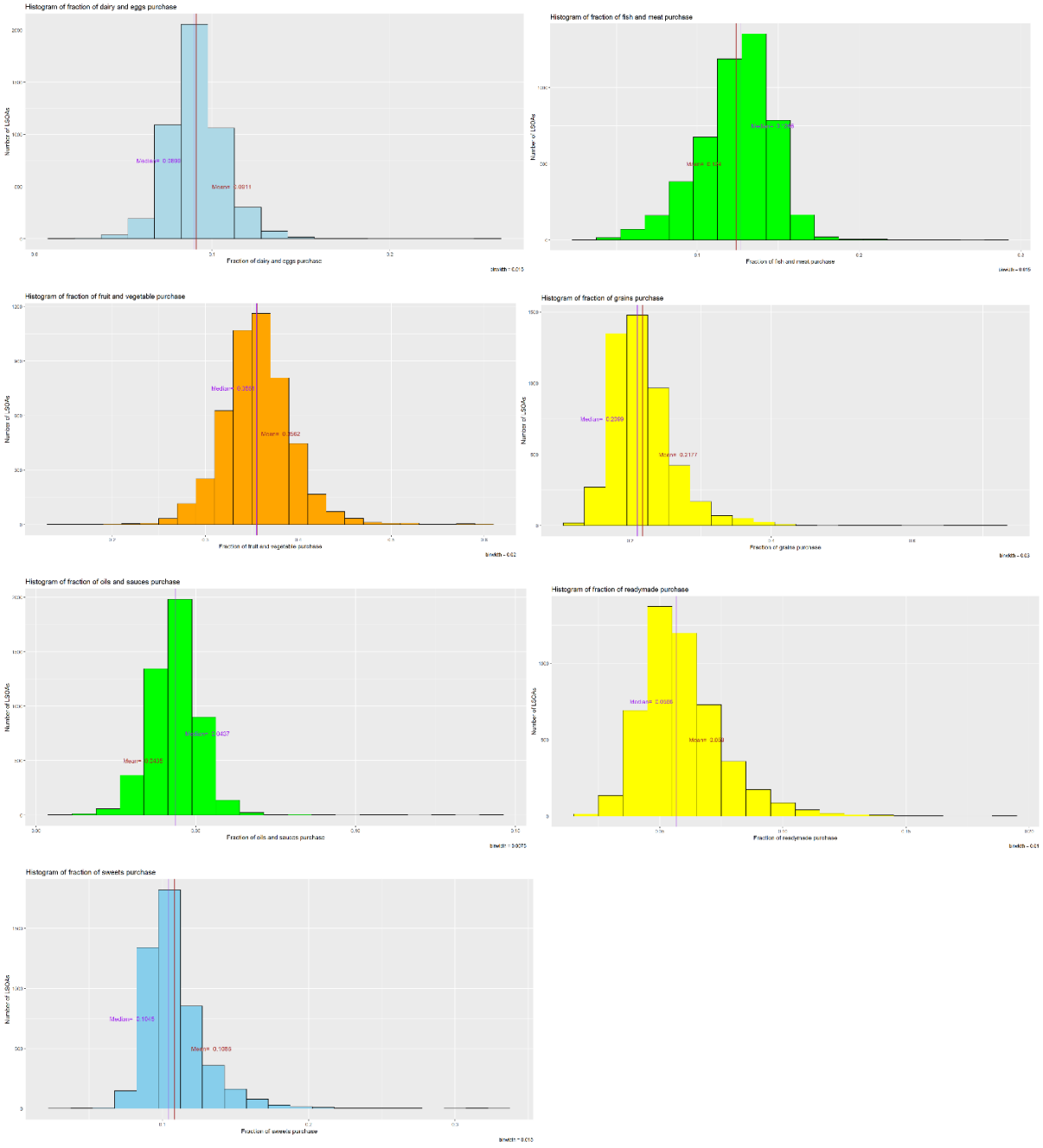

Diagram 5. **Histograms of fraction of food purchase by weight. Dairy and eggs. Fish and meat. Fruit and vegetable. Grains. Oils and sauces. Readymade. Sweets. (*from top left*).**

In general, the statistical distribution of every food category is symmetry. For every food category, the mean is also located near to the centre of distribution.

Data analysis: Clustering of LSOAs based on food purchase by weight

This analysis is trying to classify the LSOAs into a few clusters which has different patterns of food purchase. This analysis can discover the hidden sub-groups of LSOAs that have different eating habits.

This analysis had used k-means clustering[1] to split the LSOAs into 4 clusters. The parameters used are fraction of fruit and vegetable, grains, fish and meat, dairy and eggs, oils and sauces, readymade and sweets purchase by weight.

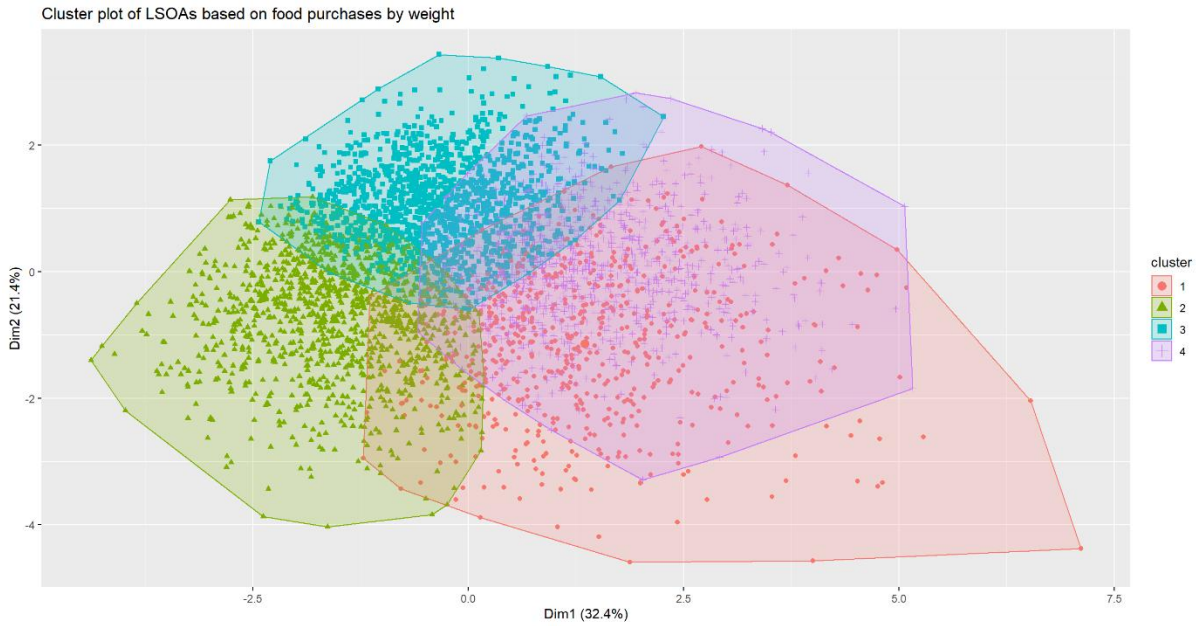

Diagram 3. **Cluster plot of LSOAs based on food purchase by weight.**

[1] Procedure of k-means clustering (Fakhitah Ridzuan, 2022):
First, for each cluster, choose an initial centroid. Then, assign each point to the nearest centroid and recompute the centroid for each cluster. Repeat the previous step until the centroid does not change. The distance between points is calculated based on sum of squared error, with the following formula (Fakhitah Ridzuan, 2022):

$$SSE = \sum_{i=1}^{K} \sum_{x \epsilon C_i} dist^2(m_i, x)$$

k is the number of clusters, x is a data point in cluster C_i and m_i is the centroid (mean) of the cluster C_i.

Why select 4 as the k value?

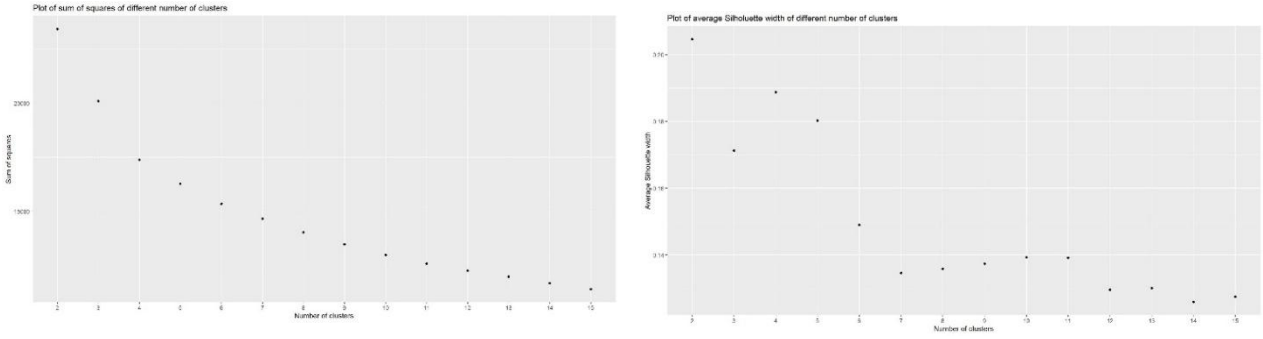

Diagram 5. **Total sum of squared error within cluster vs number of cluster (*left*). Average Silhouette width vs number of clusters (*right*).**

From the plot of total sum of squared error within cluster, the "elbow" position is at 4 because the total sum of squares within cluster drops drastically from 3 to 4 and the total sum of squares drop mildly from 4 onwards. This means that the sum of squared error, or in other words, the sum of distance between points within cluster, decreases significantly from 3 clusters to 4 clusters. This means that the points are classified in a more correct cluster when there are 4 clusters, compared to 3 clusters.

This tallies the plot of average Silhouette width[2] because there is also a peak at 4. This means that the Silhouette width at increases from 3 clusters to 4 clusters and decreases from 4 clusters to 5 clusters, which means that the points are classified more correctly when there are 4 clusters, compared to 3 clusters and 5 clusters.

[2] Silhouette width of a point, say p, is calculated by the following steps (Rick Wicklin, 2023):
First, calculate the average distance between point p and all other points in the same cluster, $AvgDistIn(p)$. Then, calculate the average distance between point p and all points in each other clusters. So, there will be $k-1$ values for k clusters. Take the minimum value as $AvgDistOut(p)$. The Silhouette width of point p, $s(p)$, is calculated using the following formula:

$$s(p) = \frac{AvgDistOut(p) - AvgDistIn(p)}{\max\left(AvgDistIn(p), AvgDistOut(p)\right)}$$

The value of Silhouette width is between -1 and 1. If the Silhouette width is near to 1, then the point is in the correct cluster; else if the Silhouette width is near to 0, then the point is in the border line of the cluster; else if the Silhouette width is near to -1, then the point is in the wrong cluster.

Cluster validation

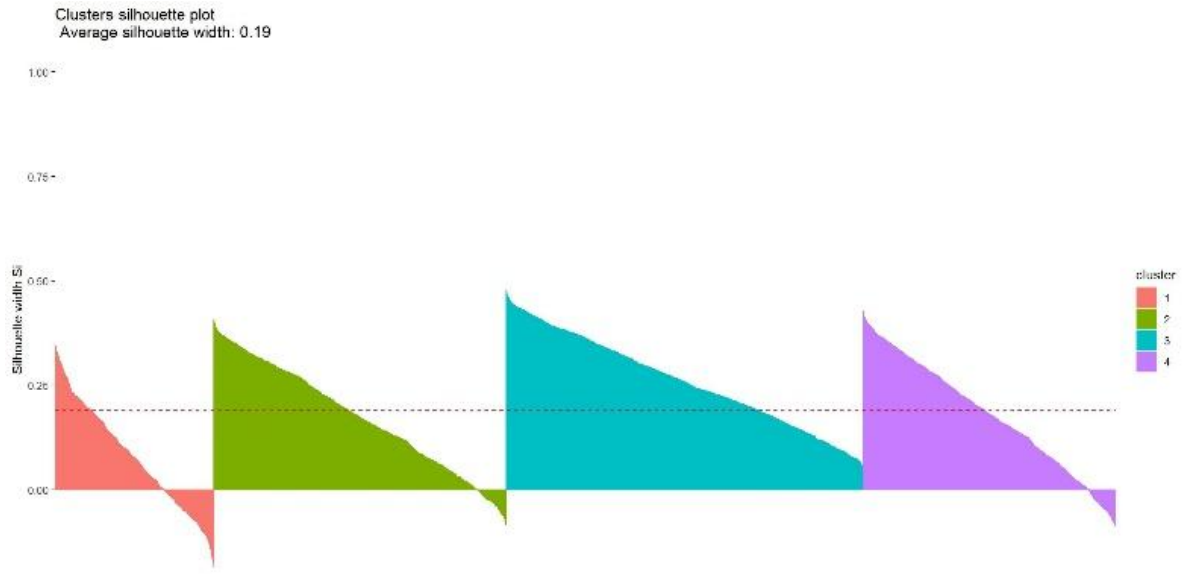

Diagram 6. **Clusters Silhouette plot**

The average Silhouette width is 0.19. This means that the clusters are valid because the Silhouette score is larger than 0. However, the clusters are not very good in terms of difference between clusters because the Silhouette score is still far away from 1.

Chapter 3: Findings and discussion

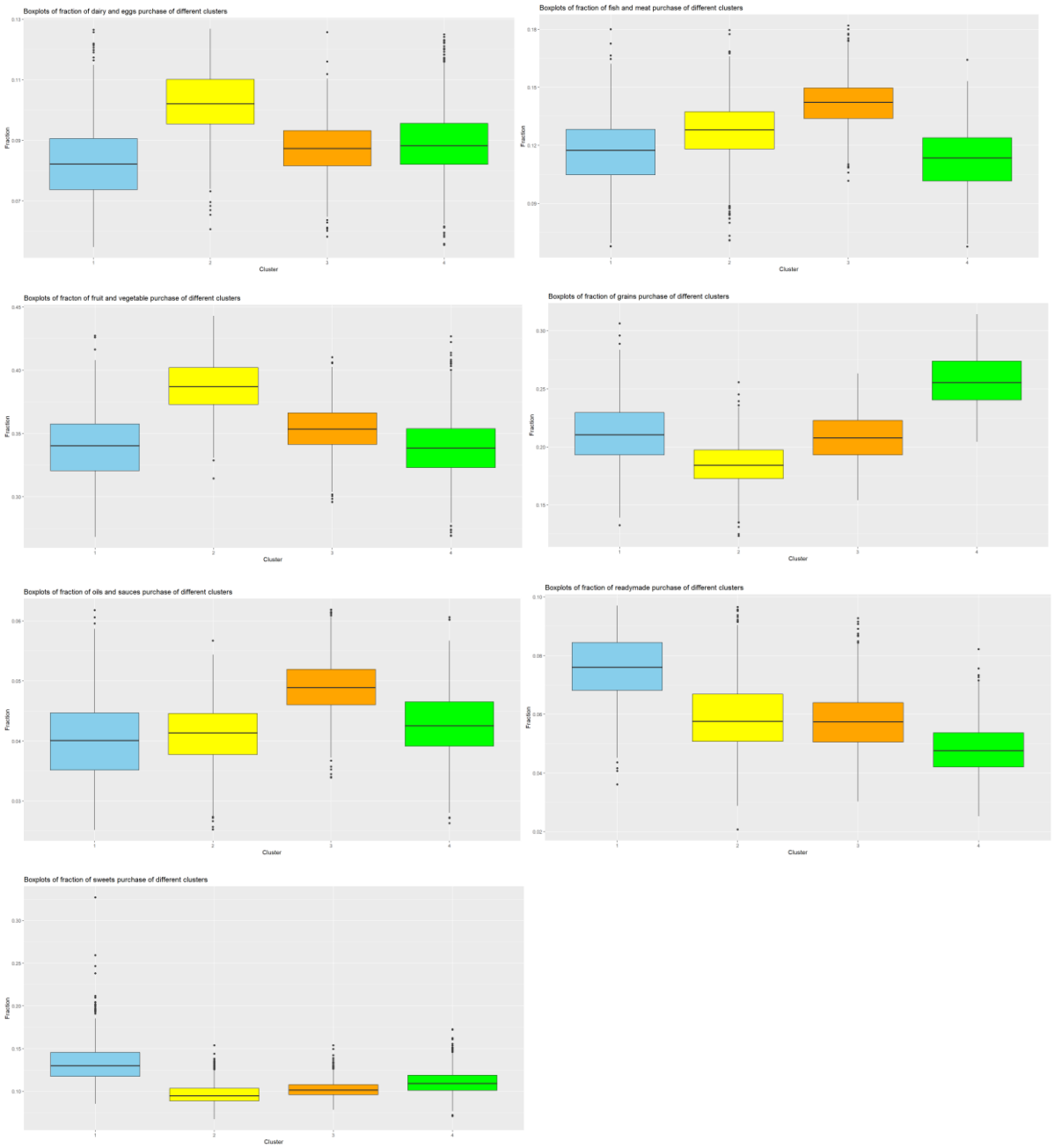

Diagram 7. **Boxplots of fraction of each food category of different clusters. Dairy and eggs. Fish and meat. Fruit and vegetable. Grains. Oils and sauces. Readymade. Sweets. (*from top left*).**

The characteristics of each cluster is summarized as follows:

1. *Cluster 1*: High consumption of readymade and sweets and low consumption of dairy and eggs.
2. *Cluster 2*: High consumption of fruit and vegetable and dairy and eggs and low consumption of grains.

3. *Cluster 3*: High consumption of oils and sauces and fish and meat.
4. *Cluster 4*: High consumption of grains and low consumption of readymade.

Chapter 4: Conclusion

The project managed to carry out descriptive analysis of the grocery purchase data by means of pie chart, boxplot and histogram. Besides that, the project also managed to split the customers into 4 clusters based on their behaviour of buying food and drinks. The distinctive characteristics of each cluster was then studied using boxplot.

Chapter 5: Bibliography

Aiello, L.M., Schifanella, R., Quercia, D., & Del Prete, L. (2020(a), February 5). Tesco grocery 1.0. *Fig share.* Retrieved from https://doi.org/10.6084/m9.figshare.c.4769354.v2

Aiello, L.M., Schifanella, R., Quercia, D., & Del Prete, L. (2020(b), February 18). Tesco grocery 1.0, a large-scale dataset of grocery purchases in London. *Nature.* Retrieved from https://www.nature.com/articles/s41597-020-0397-7

Fakhitah Ridzuan. (2022). Clustering. (Lecture note for Big Data Programming Project subject). *INTI International College Penang in collaboration with Coventry University, UK.* Penang, Malaysia.

Jeon, T. (2018, October). Merge datasets in R. *Data camp.* Retrieved from https://www.datacamp.com/tutorial/merging-datasets-r

Rick Wicklin. (2023, May). What is the Silhouette statistics in cluster analysis? *SAS blogs.* Retrieved from https://blogs.sas.com/content/iml/2023/05/15/silhouette-statistic-cluster.html

Schork, J. (n.d.). Add mean & median to histogram in R. *Statistics globe.* Retrieved from https://statisticsglobe.com/add-mean-and-median-to-histogram-in-r

Soettewey, A. (2020, August 11). Outliers detection in R. *Stats and R.* Retrieved from https://statsandr.com/blog/outliers-detection-in-r/

W3 Schools. (n.d.(c)). R tutorial. *W3 schools.* Retrieved from https://www.w3schools.com/r/default.asp

Zach. (2019, March 9). How to plot multiple boxplots in one chart in R. *Statology.* Retrieved from https://www.statology.org/multiple-boxplots-r/