**DSA3101**

**Assignment 2 Report**

**Group 2**

**Members**

* Ahmad As-Shodiqqul Amin B M T

(Survey Analysis, Report, Slides)

* Dione Lim Yee Sze

(Media Plan EDA, Report, Slides, App Design)

* Kellie Chin Shu Wen

(EDA, Feature Engineering, Clustering, Media Plan, Report)

* Keith Tay Xiang Rui

(Survey Analysis, MBA, Report, Slides)

* Lua Jun An

(Survey Analysis, Feature Engineering, Recommendation System, MBA, Media Plan, Report, Slides)

* Ni Hui Ling

(Survey Analysis, Report, Slides)

* Timothy Wong Hoey Pheen

(Survey Analysis, Report, Media Plan)

* Tu Zhehao

(Survey Analysis, Clustering Analysis, Report, Slides)

Table of Contents

[**1. Context**](#_mvgvm4t3g8fn) **4**

[**2. Objectives**](#_1cgqnfrrxiyz) **5**

[**3. Exploratory Data Analysis**](#_v9pc9wuvihm5) **6**

[Transaction Dataset](#_n6dqxy4fgcsx) 6

[Categories Information](#_kby0385hbse) 8

[Panelists Demographics](#_gnrorilfaagz) 9

[Media Habit Survey](#_5gcis7nf9he7) 9

[**4. Feature Engineering**](#_lsib7zpcwuhv) **16**

[**5. Clustering**](#_pfkcty5lzmdi) **18**

[**6. Recommender System**](#_egknntha6u4a) **36**

[**7. Media Plan**](#_m0t7o0ff4ca1) **40**

[Mobile Application](#_1j37rmbmvq76) 40

[Loyalty / e-payment System](#_4waafbdrt1xm) 42

[Coupons](#_5scixchpccub) 43

[Media Publicity](#_tp7u02fz00tc) 44

[**8. Evaluation**](#_6zz8upv6xifw) **46**

[**9. Future Works**](#_sjm7aeof7u0y) **48**

[**10. References**](#_m2dzsjerbtpw) **49**

# 

# 1. Context

## Defining ‘Healthier Choice’

As the Hackathon’s objective is to ‘encourage healthy eating’, it is crucial that we first define ‘healthy eating’ or more specifically, what is considered to be a healthier choice. To answer this, we based the definition on nutritional aspects that are recognised under Singapore’s Healthier Choice Symbol (HCS). These nutritional aspects include both macronutrients like fats and proteins as well as micronutrients like calcium and sodium.

## Our 6 Aspects

For our group, we chose 6 main nutritional categories which are **wholegrains, calcium, sugar, saturated fats, sodium and trans fat**. These six were chosen because calories itself is not a good measure of how healthy a food is and furthermore, it is more appropriate that we look into other aspects of nutrition.

## Nutritional Information of the Food Items in the Dataset

To supplement calorie information provided, we found and added the 6 additional nutritional information from external online sources, namely, My Fitness Pal and Fatsecret, the latter being culturally-contextualised.

Our assumption is that the data we have garnered from these sources are accurate and would hence, prove fruitful in our analysis.

# 

# **2. Objective**s

The main objective of the Hackathon is to encourage the population to minimise their spending on packaged foods while maintaining an optimal level of calories consumption based on previous habits. In other words, it is to promote healthier purchasing and consumption habits.

To achieve this, our group’s sub-objectives are as follows:

1. To identify different consumers segments based on their personal details and overall lifestyle such as spending habits and media consumption
2. To identify suitable media channels through which we can maximize our campaign’s media outreach
3. To incentivise and encourage different consumer segments to continually purchase healthier substitutes
4. To design a mobile application which utilises a recommendation system for the different consumer segments to purchase healthier substitutes

More importantly, our overall plan focuses on a **long-term implementation period** because we believe the adoption of a healthier lifestyle requires time. Using this approach, the future potential of our strategies will ensure that the effects of these measures are not temporary and truly encourage change in mindset and lifestyle among consumers.

# **3.** Exploratory Data Analysis

To begin, we have been provided with four data resources:

1. DACC\_Hackathon\_Categories\_Information\_V2
2. DACC\_Hackathon\_Media\_Habit\_Survey
3. DACC\_Hackathon\_Panelists\_Demogs
4. DACC\_Hackathon\_TransactionData

An important assumption we have made is that the panelists in this dataset were randomly sampled from the general Malaysian population. Thus, we will treat this dataset as an accurate representation of the general Malaysian population.

We then explore each dataset individually to gain useful insights which we will capitalise on for our media plan.

## Transaction Dataset

We first plot the number of transactions per food category to get an overview of all the transactions made. From Figure 1 below, we can observe that there is an uneven distribution of items purchased.

|  |
| --- |
| Figure 1: Bar Plot of Number of Transactions for each Category |

As we plan to use the panelist’s spending to gauge their grocery shopping habits later on, we plot the number of transactions which involve no money i.e. spend is 0. From Figure 2 below, we can observe that the number of transactions which involve no money (1,193) is relatively low as compared to the total number of transactions we have (1,389,941). The distribution of items involved in these transactions are also similar to the overall distribution of items involved in all transactions as shown in Figure 1.

|  |
| --- |
| Figure 2: Bar Plot of Number of Transactions with no monetary value for each Category |

These transactions could possibly have been erroneous data or items which were received by the panelists from Buy-1-Get-1-Free promotions. Hence, we removed these transactions as we assume the panelists had no intention of purchasing that item.

## 

## Categories Information

As for DACC\_Hackathon\_Categories\_Information\_V2, we want to explore and understand the nutritional information of the categories of the products we have seen in the transactions.

|  |
| --- |
| Figure 3: Histogram of Calories/100g of All Products |

From Figure 3, we can see that, generally, there is greater availability of products with lower calorie content than those with higher calorie content. However, just using calories as a measure yielded little information from this dataset.

## 

## Panelists Demographics

To attain an overview of our panelists’ health status, we plot the count of each BMI type, as indicated in Figure 4.

|  |
| --- |
| Figure 4: Bar Plot of BMI of Panelists |

We consider over weight, obese and under weight as unhealthy BMI types. Alarmingly, among the panelists who provided their BMI, more than half have unhealthy BMI.

## Media Habit Survey

Lastly, we compile the results of the media survey conducted on our panelists from the dataset, DACC\_Hackathon\_Media\_Habit\_Survey.

The media survey contains 7 questions which are centred around the different types of media, their ownership or adoption rate of media devices as well as usage habits. Some forms of the media include television, social media sites, newspapers and e-wallets. We will assess each question individually.

### Question 1: Television

|  | |
| --- | --- |
|  | |
| Figure 5: Viewing Habits across the day for the Entire Week | |

Based on Figure 5 and 6, we notice that even on weekends, the majority of the panelists watches television during ‘Prime Time’ and in the ‘Evening’. Additionally, ‘TV3’ is also the most watched channel on television.

|  |
| --- |
| Figure 6: Box Plot of Watch Times of Different Channels |

### Question 2: Internet

|  | Figure 7:  Frequency Of  Internet Access |
| --- | --- |

As can be seen from Figure 7, most panelists surveyed use the Internet daily.

### 

### Question 3: YouTube

|  | Figures 8:  Viewing Habits on Youtube across the day for the Entire Week |
| --- | --- |
|  |
|  |

Generally, most people spend either less than 1 hour or none at all. People spend time watching in the afternoons and evenings. Saturday appears to be the most watched day for YouTube, followed by Sunday, and finally the weekdays.

### Question 4: Mobile Phone

|  | Figure 9:  Ownership of personal mobile phone |
| --- | --- |

Most panelists surveyed own a mobile phone.

|  | Figure 10:  Internet Service on Mobile Phones |
| --- | --- |

Adding on to the previous point, most panelists surveyed also have access to internet service (mobile data) on their mobile phones.

### 

### Question 5: Radio

|  |
| --- |
| Figure 11: Frequency of Listening to Radio |

Most panelists who listen to radio do so on a daily basis.

|  |
| --- |
| Figure 12: Time Periods that users listen to radio on the Weekdays |

Radios have the highest listenership on weekday mornings.

### Question 6: Newspaper

|  |
| --- |
| Figure 13: Type of Content Read in Newspapers |

Panelists tend to read newspapers for the local, international and entertainment news.

### Question 7: E-transactions

|  |
| --- |
| Figure 14: E-Wallet Types |

Touch N’Go’s e-wallet is the preferred choice for the panelists if they used e-wallet services.

# 4. Feature Engineering

## 6 Main Food Groups

In the ‘DACC\_Hackathon\_Categories\_Information\_V2’ dataset provided, there were 62 different food categories. To allow for the data to be more interpretable, we further group these 62 different categories of food into 6 main food groups - *Alcohol, Beverages, Cooking Essentials, Carbohydrates, Dairy/Eggs* and *Snacks*.

## High or Low Calorie

We also label foods with more than 400 calories per 100g as *high calorie foods* and foods with less than or equal to 400 calories per 100g as *low calorie foods*.

## Percentage of Spending on Each Type of Food

To better understand each panelist’s grocery shopping habits, we look at their spending on different types of food. For each panelist, we calculate the total amount spent on each of the 6 main food groups, high calories foods and as a whole. Then, we compute the percentage spent on each type of food.

The dataframe below shows in each row, a unique panelist and each column, the panelist’s percentage of spending on that type of food.

|  |
| --- |
| Figure 15: Dataframe of the percentage of spending on each type of food |

## Important Features Extracted from Media Habits Survey

These are the important features that we extracted from the Media Habits Survey which will help us to design a robust media plan across all the platforms.

For television users, we focus on:

* Frequency of television consumption

For internet users, we focus on :

* Availability of Internet access
* Frequency of use of Internet and Youtube

For mobile phone users, we focus on:

* Possession of mobile phone
* Accessibility to mobile data

For radio users, we focus on:  
  
 - Frequency of radio consumption

For the panelist who read newspapers, we focus on:

* Frequency of reading newspapers
* Content read in newspapers

For eWallet users, we focus on:

* Usage of eWallet
* Usage of Touch N’Go

# 

# 5**.** Clustering

## K-means Clustering

We segment the panelists based on their grocery shopping habits using k-means clustering analysis. This segmentation creates smaller groups of panelists with similar shopping habits. Hence, we can ensure that we are able to design better media plans which suit their needs.

|  |
| --- |
| Figure 16: Plot of WCSS vs k |
|  |
| Figure 17: Plot of Average Silhouette Width vs k |

The within cluster sum of squares (WCSS) and average silhouette width plots above suggest that the optimal number of clusters could be between 3 to 6 clusters. Thus, we proceed to perform k-means clustering using k = 3, 4, 5 and 6.

|  |
| --- |
| Figure 18: Silhouette Plot for 5 clusters |
|  |
| Figure 19: Mean of each feature for each cluster |

Comparing the silhouette plots and cluster means from the different k-means clustering done prior, our choice was to go with 5 clusters. This is because k-means with 5 clusters created sizable clusters with distinct characteristics. Its average silhouette score was also similar to that of having 3, 4 and 6 clusters.

## 

## The Clusters

Based on the transactions, we can describe the 5 clusters in the following way:

* 0: People with the highest alcohol consumption
* 1: People who seldom cook at home but consume high amounts of beverages
* 2: People who spend large amounts on cooking essentials
* 3: People with a high carbohydrate diet
* 4: People who snack high amounts and drink alcohol

## Demographic Analysis

Based on these 5 clusters, we now analyse the demographics of each cluster.

### BMI

|  | |  |
| --- | --- | --- |
|  | |  |
|  | Figure 20:  Proportion of each BMI type  in Clusters 0 to 4 | |

As shown above, Clusters 0 and 4 each have the highest proportion ofhealthyindividuals, forming approximately 30% for each cluster. In contrast, Clusters 2 and 3 seem to have high proportions of unhealthy individuals; around 45% of Clusters 2 and 3 belong to the “*Obese*”, “*Over Weight*”, and the “*Under Weight*” categories.

### Ethnicity

|  | |
| --- | --- |
|  | |
|  | |
| Figure 21: Proportion of each Ethnicity type in clusters 0 to 4 | |

Malays form the majority of all clusters but Clusters 2 and 3 have the highest proportion of Malays, constituting around 80% in each of the clusters.

### Lifestage

|  | |
| --- | --- |
|  | |
|  | |
| Figure 22: Proportion of each Lifestage type in clusters 0 to 4 | |

Cluster 1 has an overwhelmingly large proportion of families with young kids as ‘Nesting Families’ and ‘Teensand Toddlers’ make up about 80% of the cluster.

### Income

|  | |
| --- | --- |
|  | |
|  | |
| Figure 23: Proportion of each Income type in clusters 0 to 4 | |

For Cluster 0, the majority of individuals are ‘low to middle income earners’, with approximately 30% of the cluster earning between RM3000 and RM4999 (*Mid Low*) and about 40% of the cluster earning below RM3000 (*Low + Low High*).

For Cluster 1, the majority of the individuals are ‘middle income earners’, with about 55% of the cluster earning between RM3000 and RM7999 (*Mid Low + Mid High*). Additionally, Cluster 1 has the highest proportion of ‘high income earners’. About 13% of the cluster earns more than RM7999 (*High*).

For Cluster 2, the majority of the individuals are ‘low to middle income earners’, with around 50% of the cluster earning below RM3000 (*Low + Low High*). The ‘high income earners’ only make up less than 10% of this cluster.

For Cluster 3, the majority of the individuals are ‘low income earners’, with roughly 60% of the cluster earning below RM3000 (*Low + Low High*). Akin to Cluster 2, the high income earners make up a very small proportion of the cluster (~5%).

For Cluster 4, the majority of the individuals are ‘middle-low income earners’, with 45% of the cluster earning below RM4999 (*Low + Mid Low*). Similar to Cluster 1, Cluster 4 has a high proportion of ‘high income earners’ too. About 12% of the cluster earns more than RM7999 (*High*).

## 

## Market Basket Analysis

We will perform market basket analysis on each cluster to get the top 10 itemsets with highest supports to identify transaction trends.

# 

Figure 24: Top 10 itemsets with highest supports for Cluster 0

As seen from the figure above, the top supports for Cluster 0 include *eggs, biscuits, rice, instant noodles* and *cooking oils*.

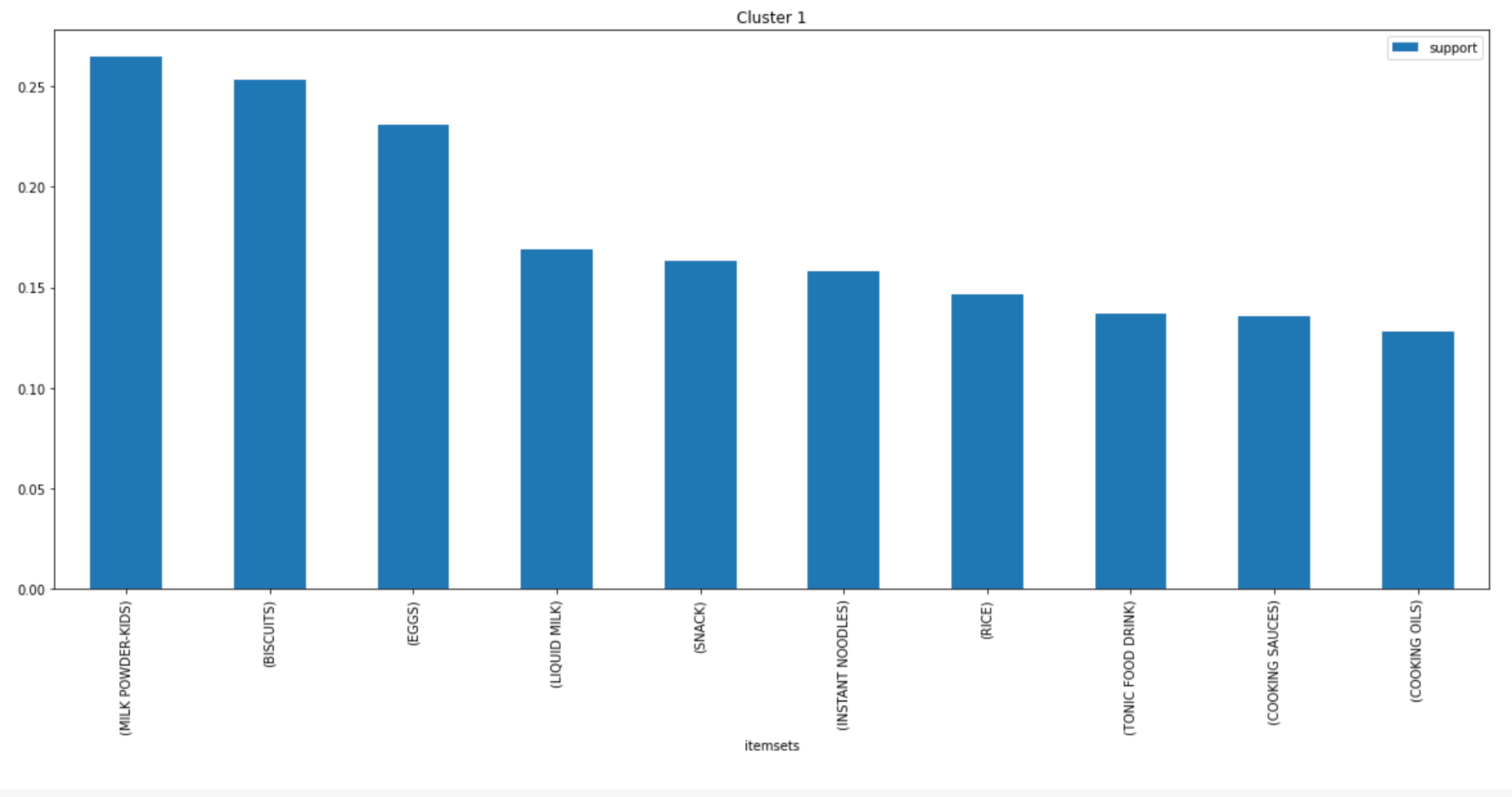
****

Figure 25: Top 10 itemsets with highest supports for Cluster 1

For Cluster 1, the support for *Milk Powder - Kids* is the highest at ~0.27 and the support for *Liquid Milk* is ~0.18, both of which are highest amongst all the clusters. This could be because Cluster 1 has an overwhelmingly high proportion of young families with kids as we have observed earlier in Figure 23. For this group, we also witness a low support for Cooking *Oils, Cooking Sauces* and *Rice*, each approximately at 0.14.

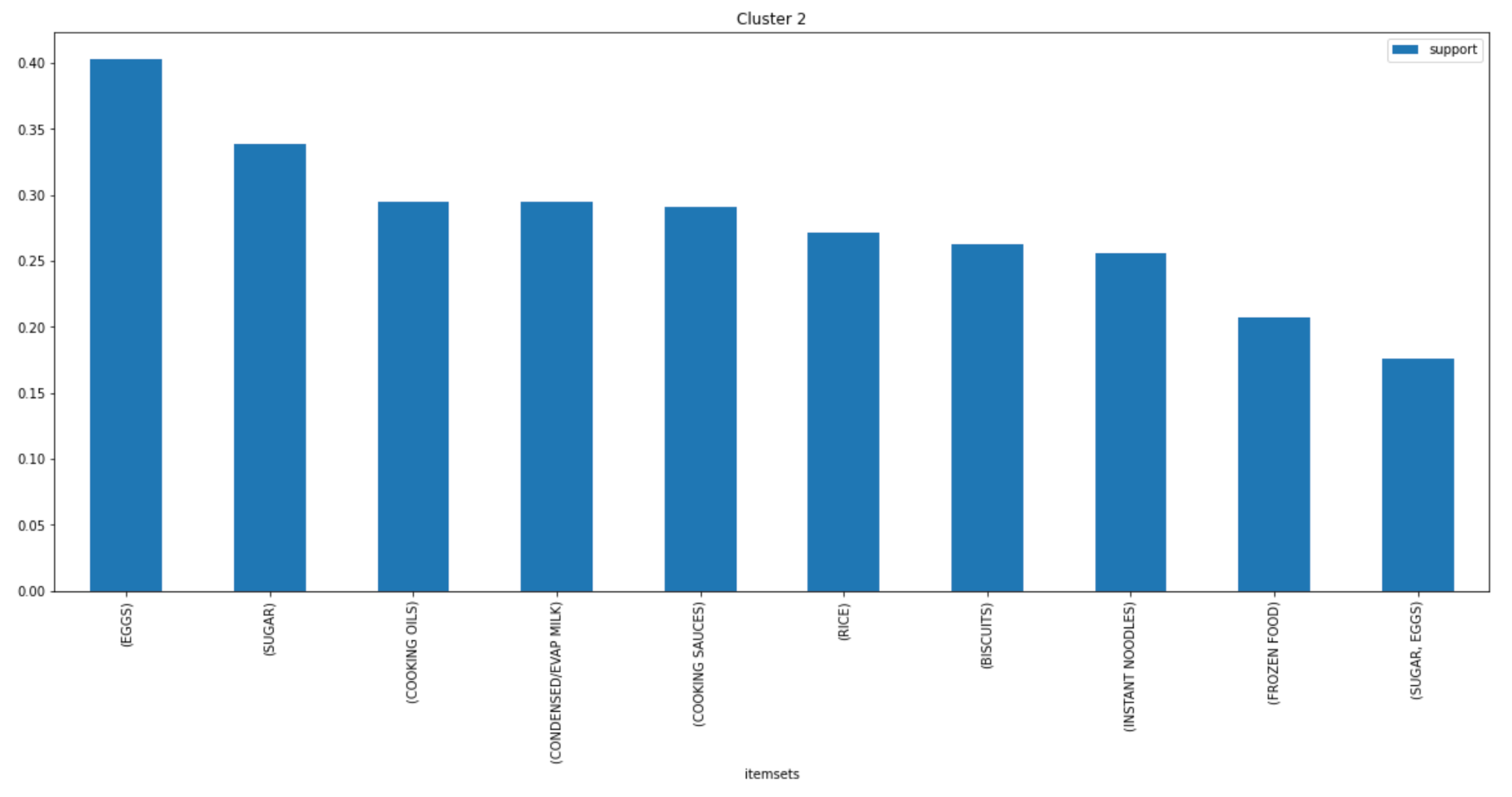
****

Figure 26: Top 10 itemsets with highest supports for Cluster 2

For Cluster 2, we observe high support for *Egg, Sugar, Cooking Oils* and *Cooking Sauces*. The support for *Condensed Milk* is also quite high, having a support of ~0.3.

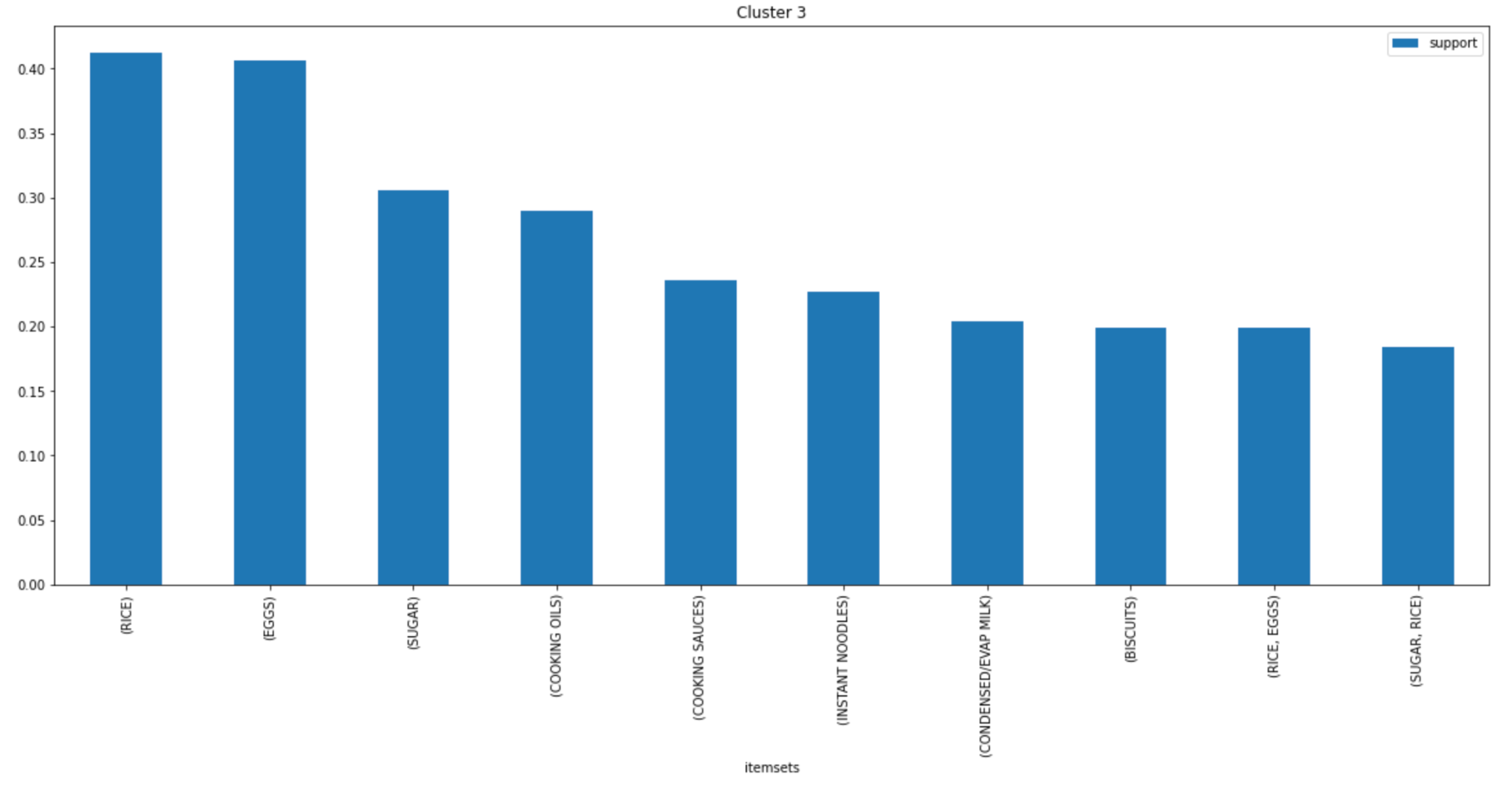
****

Figure 27: Top 10 itemsets with highest supports for Cluster 3

Cluster 3 is very similar to Cluster 2. Four out of the top 5 itemsets with highest support: *Eggs, Sugar, Cooking Oils* and *Cooking Sauces*, appear in the top 5 itemsets of Cluster 2 as well. Furthermore, for Cluster 3, the support for *Rice* and *Eggs* are exceptionally high, at ~0.4 each.

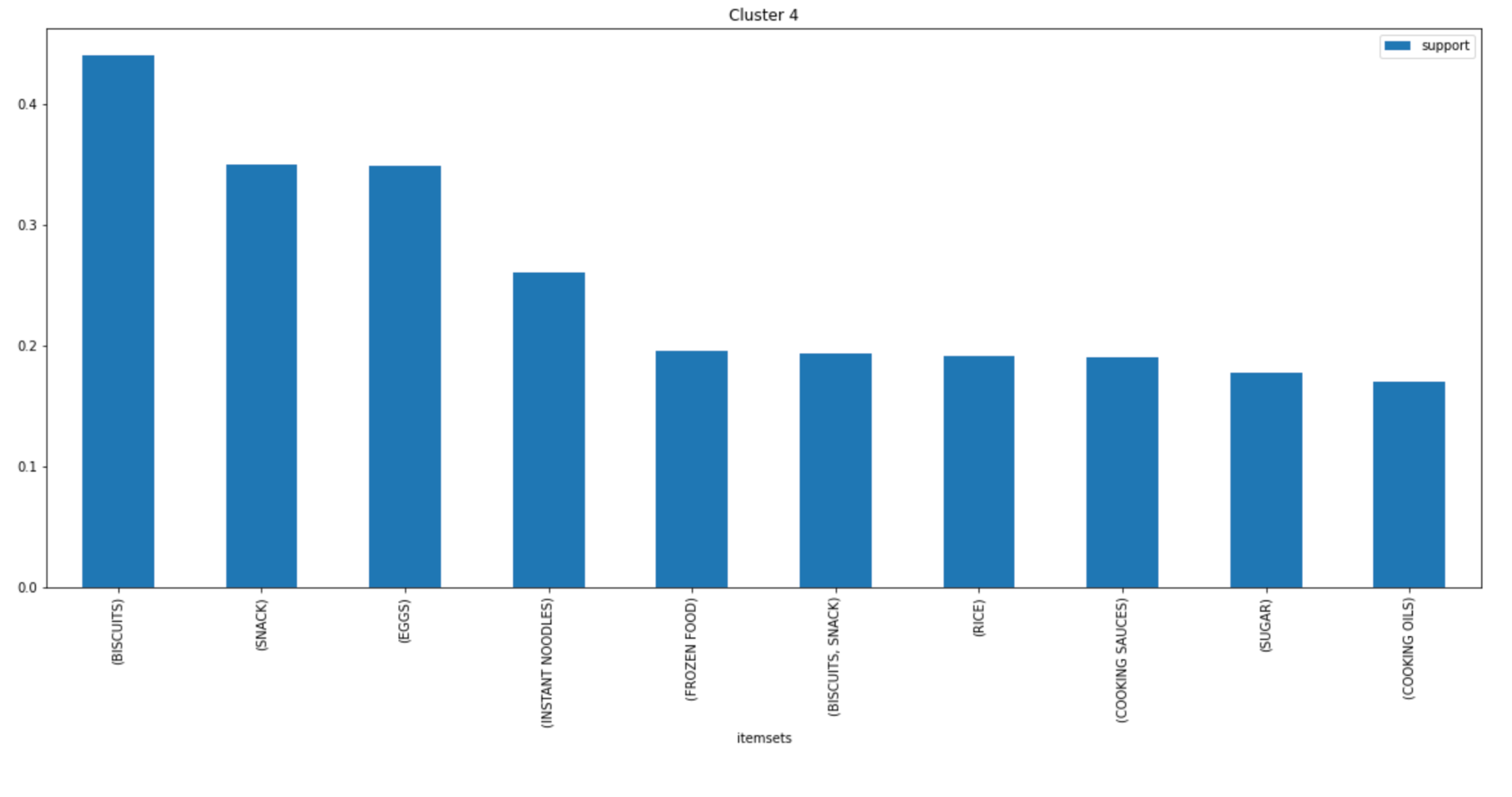
****

Figure 28: Top 10 itemsets with highest supports for Cluster 4

Cluster 4 has the highest support for *Biscuits* and *Snacks* as compared to the other clusters, with supports of ~0.45 and ~0.35 respectively. This cluster also has low support for *Cooking Oils, Sugar* and *Cooking Sauces* of ~0.2 each.

## Characteristics of Each Cluster

With the insights from our MBA analysis, we are able to understand each cluster better and assign distinct characteristics for each.

**Cluster 0:** The top 10 itemsets indicate that this group usually has more *domesticated baskets*. Since most readings from the indicators are average, this group can be denoted to represent the **average consumer demographic**.

**Cluster 1:** As can be deduced from the lifestage breakdown in Figure 22, an overwhelming majority of this cluster belong to *Nesting Families*. This would probably indicate **families with many young children**, hence, explaining the high support for *milk powder* - *kids* and *liquid milk*. Moreover, the group probably also **does not cook as often** as the other clusters judging from the low values in support for *cooking essentials,* which is even lower than Cluster 4. This information would be proven significant, upon reading the description for Cluster 4.

**Cluster 2**: Evident from the income breakdown, this group consists of mainly l*ow to middle* income earners. Furthermore, from the high support values of *egg, sugar, cooking oils* and *cooking sauces*, this group can be said to **cook often** and eat a high amount.

**Cluster 3:** This particular group is similar to Cluster 2, differing only by having a generally lower income. From the BMI breakdown, 45% of this group also has unhealthy BMI. In addition, the group has high support values for *rice* and *instant noodles*. The itemsets *(rice, eggs)* and *(sugar, rice)* are also within the top 10 highest support itemsets for this group. Hence, this group likely consists of **individuals with high carbohydrate diets**.

**Cluster 4**: From the BMI breakdown, this final group has the highest proportion of *healthy* people, and the lowest proportion of *underweight* individuals (only ~10%). This might be due to snacking as shown in the high support values for *biscuits* and *snacks*. When we also look at the income breakdown, we see this group has a higher proportion of higher income individuals, which might be a reason why they are able to afford eating out and eating healthy. The higher income indicates the high possibility that this group consists of many white-collared office workers. Hence, we denote this group as **working individuals who seldom cook**.

Since Cluster 2 and Cluster 3 have the highest proportions of panelists with unhealthy BMI, we will focus more on these two clusters.

## Media Usage Analysis

Given the previous analysis on general media habits, to cater to our 5 clusters, further analysis was done.

### Media Usage

|  |  |
| --- | --- |
|  |  |
| Figure 29: Histogram of Time Spent on Facebook, Instagram, Twitter & Youtube respectively | |

As can be seen above, the majority of panelists used Facebook and Youtube daily.

### Phone Usage

|  | Figure 30:  Devices used to  access the Internet across the different clusters |
| --- | --- |

Across the different clusters, phones are the most widely used device to access the Internet. This may be due to high phone ownership coupled with internet and mobile data access, which can be shown in the combined bar plot on the next page.

|  |  |
| --- | --- |
|  |  |
|  | |

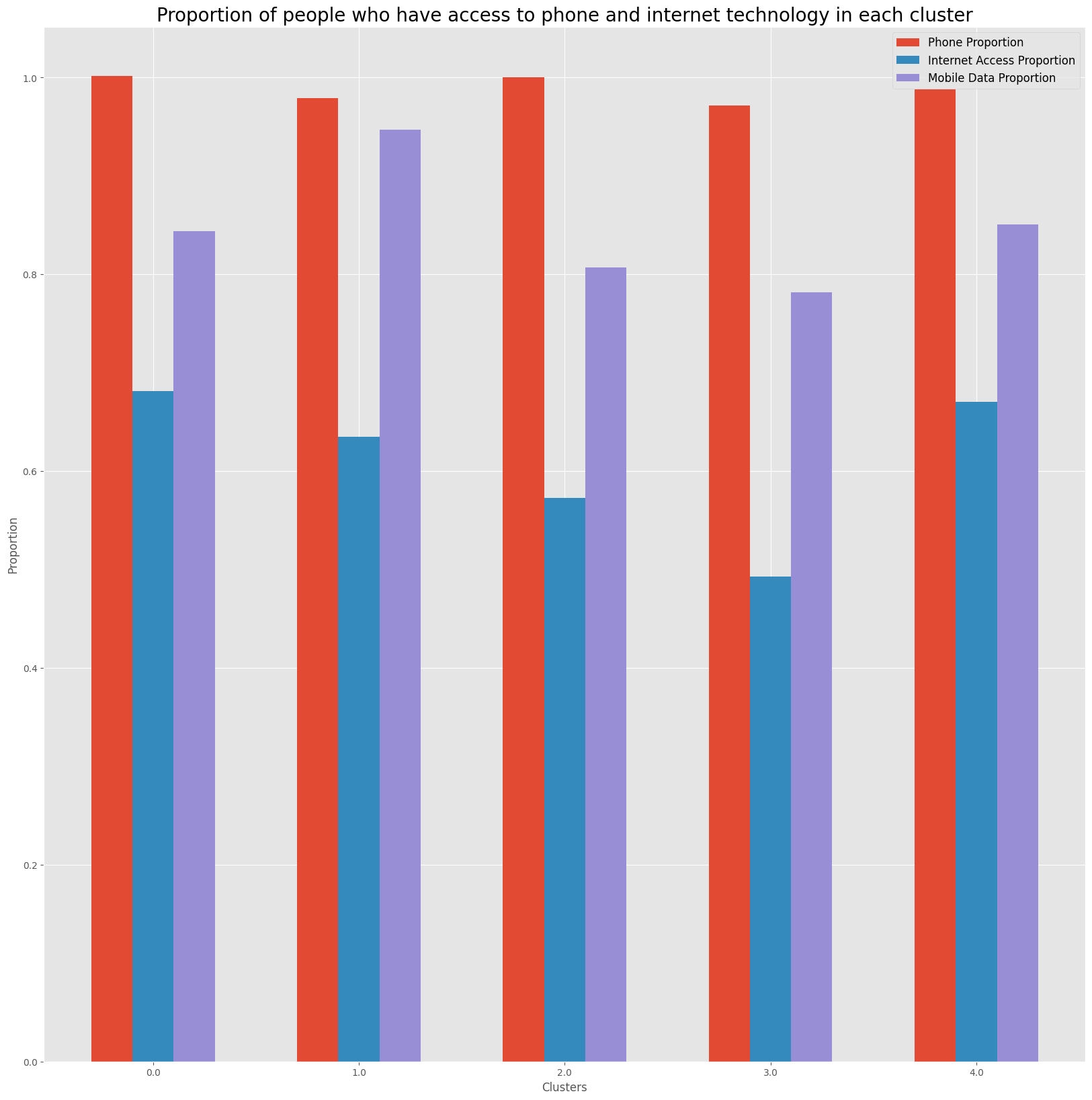


Figure 31: Bar Plot of Internet Access, Mobile Data Endowment and Phone Ownership

Examining the barplot, we can see that more than 95% of panelists in each cluster own a phone. However, clusters 2 and 3 came in lower for internet access and mobile data ownership, while clusters 0, 1 and 4 have higher mobile data and internet access.

### Radio, TV, Newspaper and Internet Frequency

Analysing the different media types across the 5 clusters, these are two sets of plots produced.

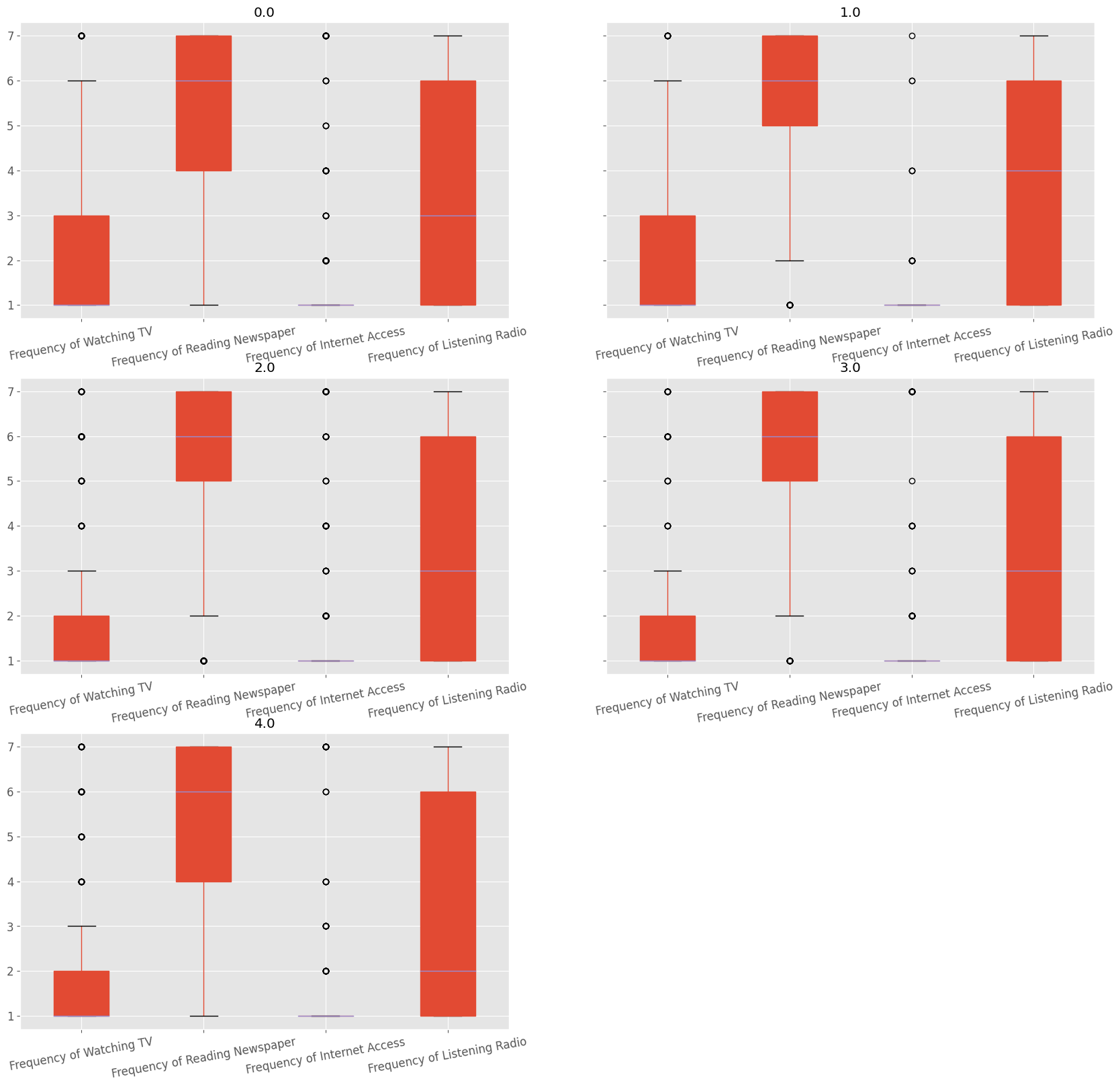


Figure 32: Average media frequency in each cluster

Figure 32 displays the average media frequency in each cluster on a scale of 1 to 7 with 1 being most frequent and 7 being least frequent.

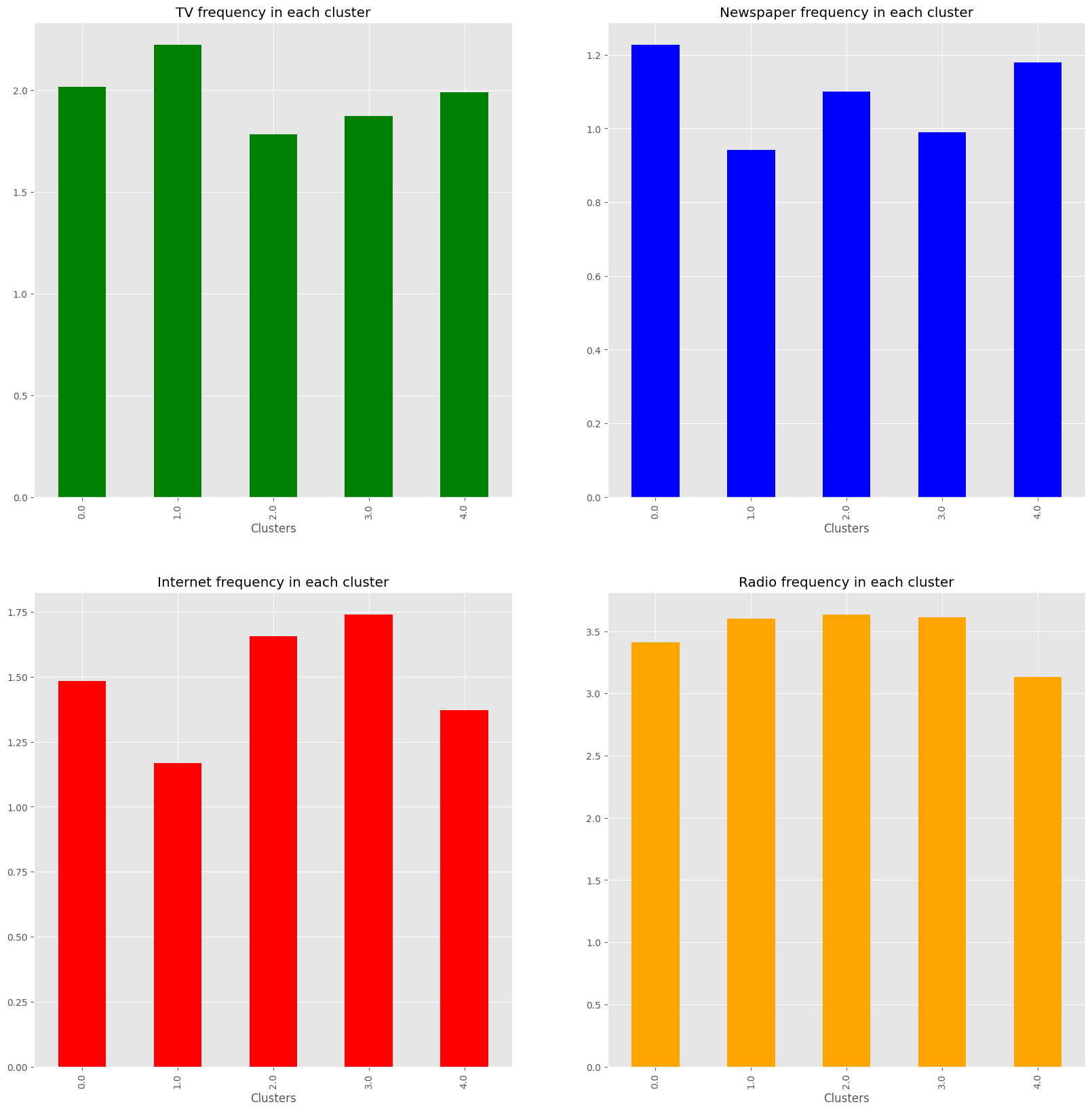


Figure 33: Media frequencies across the clusters

From Figure 32, it is observed that TV is the most frequently used media type, followed by radio and newspapers. There is no distinct quantile for the internet frequency.

This can be more clearly seen in Figure 33, which compares across the clusters faceted by the media type. Radio frequency was the highest for clusters 0, 1 and 4. TV frequency was the highest for clusters 2 and 3. Meanwhile, internet frequency for clusters 2 and 3 was the lowest. This supports our earlier findings, with cluster 2 and 3 having the lowest endowment of technology across phone access, internet access and mobile data access, which would hence, be related to their internet frequency. Mobile data was also found to have high correlation to internet frequency with Point Biserial correlation. Newspaper frequency for clusters 0, 2, 4 was the lowest, and 1 and 3 had the highest.

### Electronic Transactions

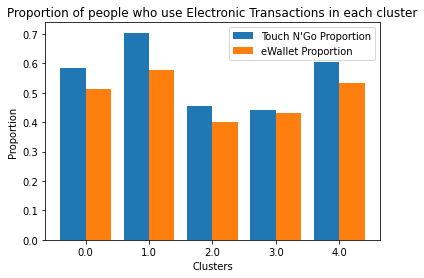


Figure 34: Proportion of Panelists Using Each Type of Electronic Transaction

Using Cramer's V correlation, we found that most panelists who used eWallet also used Touch N' Go. Figure 34 shows the proportion of Touch N' Go and eWallet users in each cluster following very similar patterns.

In general, electronic transactions are more popular amongst clusters 0, 1 and 4, while clusters 2 and 3 were found to have the least electronic-transaction usage in their clusters. This may be explained by the demographics of the clusters, since clusters 2 and 3 are of lower income ranges and have lower internet access.

From our Media Usage Analysis, we also see a clear distinction in the media habits of panelists in clusters 2 and 3 as opposed to those in clusters 0, 1 and 4. This will be leveraged in our media plan.

At this stage, we have achieved our first and second sub-objectives of identifying and understanding different consumer segments as well as their media consumption habits through clustering.

# 6**.** Recommender System

Our group developed a Recommender System based on BERT. BERT is a computational model which converts words into vectors of numbers and is widely used in the field of Natural Language Processing (NLP).

To better facilitate the explanation of our methodology, we have presented it in a flowchart in Figure 35.

|  |
| --- |
| Figure 35: Workflow of our Recommender System |

The items in our dataset will now be passed into the Bidirectional Encoder Representation from Transformers (BERT) model with the aim of finding semantically similar items. By converting the items into word embeddings, we simplify the problem by calculating the cosine similarity between two words to decide if they are semantically similar.

| Cosine similarity - Wikipedia |
| --- |
| Figure 36: Cosine Similarity Formula |

To better visualise our idea of ‘closeness’, we have included Figure 37 of the 2-dimensional space our words exist in.

| Word Mover&#39;s Embedding: Universal Text Embedding from Word2Vec | IBM  Research Blog |
| --- |
| Figure 37: 2D Visualization of Word Embedding Space |

We can see that in the 2D space, “science” and “research” are semantically similar. With this in mind, we hope the NLP model pre-trained upon BERT would be able to provide us with the same function of gathering food that is similar.

After running the NLP model, the list of similar items for each given item would be generated. We then proceed to assess whether these recommendations are indeed healthier based on our check\_healthierV2 function.

As mentioned in our Context, we denote healthiness based on Health Promotion Board’s standards, judging by 6 aspects: Wholegrains​, Calcium, Sugar, Sodium, Saturated Fats and Trans Fat.

| def check\_healthierV2(item1, item2, cal1, cal2):  #We give the trans-fat categorical variable a weight of 10  weight = 10  score = 0  for i in range(len(item1[:2])):  #we reward nutritional value for higher wholegrains and calcium  score += (item1[i] - item2[i])  for i in range(len(item1[2:5])):  #we penalise heavier for food with high sodium, saturated fats and sugar  score -= (item1[i] - item2[i])\*weight  #we penalise the food if it contains trans-fat  score += (item1[5] - item2[5])\*weight  #we penalise the food if it has higher calories  score -= (cal1 - cal2)/10  if score >= 0:  return True  else: |
| --- |
| Figure 38: Definition of check\_healthierV2 function |

From Figure 38, the *check\_healthierV2* function takes in the nutritional values of the 6 columns before denoting the relative healthiness of item1 against item2, penalising and rewarding nutritional aspects accordingly. By knowing the healthier option, we recommend the best substitute to the buyers based on their choice.

Now that we have a robust Recommender System to recommend healthier food options to our panelists, we are able to proceed.

In the next section, we will focus on the media channels we should use to maximize our media outreach and how to incentivize different customer segments to purchase healthier foods.

# 7**. Media Plan**

Our campaign is called ‘Utamakan Kesihatan Anda’, loosely translated to ‘Prioritise Your Health’. Given that the vast majority of our population are Malays and Bahasa Malaysia is the primary language used in Malaysia, we have decided to give our campaign a Malay name.

There are 5 main elements to our campaign, each purposefully designed to maximise our reach and promote healthy eating. These 5 elements are a mobile application, coupons, TV and radio publicity, discounts and a point system.

## Mobile Application

From our media usage analysis, we learnt that across all 5 clusters, almost all panelists owned a mobile phone and more than 75% of panelists subscribed to a mobile data plan. We can safely conclude that more than 75% of the general population would be able to use a mobile application as long as they are within telecom boundaries. Therefore, utilising a mobile application as a media channel can help us maximise our reach.

|  |
| --- |
| Figure 39: Logo of our application |

We have created an application called *Sihat*, literally translated to “healthy” in English. *Sihat* acts as a place for users to create a grocery list before they head to the grocery store. The tagline, ‘Sihat apa? Sihat App-ah!’, which loosely translates to ‘Healthy what? Sihat App!’, is catchy and will definitely get the message across to our intended consumers. Additionally, we hope to be able to partner with a delivery service provider and allow our users to have the option of having their groceries delivered right to their doorstep.

| *Sign-in page Grocery catalogue* |
| --- |
| *Item Recommendation Basket* |
| Figure 40: Application Interface |

The typical steps a user would take while using *Sihat* is as follows:

1. Sign-in to their account
2. Select the item they want to buy
3. Add recommendation or original item to basket
4. Repeat steps 2 to 3 until all the items they need are added into their basket
5. Head to the *Basket* page to view grocery basket

(Check and edit accordingly)

1. Make payment and schedule delivery (For those with epayments enabled)

Our recommender system will play a vital role in steps 2 and 5. In step 2, based on the food item selected, they will be recommended a healthier food item. In step 5, based on the items in their basket, they will also be recommended a healthier food item. We intend to make this feature offline to accommodate individuals with low internet access and mobile data.

## Loyalty / e-payment System

Meanwhile, for clusters 0, 1 and 4, they have distinct characteristics, *i.e*. heavy usage of electronic payment and high internet frequency. We will hence tap on this to promote healthy living in these clusters by incorporating electronic payments into the application. It is worth mentioning this will be an online feature due to the nature of the electronic payment system, which would not be a problem for our target group.

| *E-Payments with Loyalty Points* |
| --- |
| Figure 41: Application Interface |

The catch is that this payment option would include a points system whereby consumers will receive points upon purchase of a healthier choice item. This would be translated to a loyalty card system whereby points can be accumulated and used to redeem items with the Healthier Choice Symbol (HCS). There will also be some discounts on foods with the HCS labels to incentivise consumers to try healthier alternatives.

Although our main purpose is to introduce healthier options, we would like to leave a lasting impression on our consumers. The loyalty system was thus introduced to incorporate healthy eating habits in their daily lives. From this, we firmly believe that we can make a difference for the better.

## Coupons

Previously, our demographic analysis of each cluster also revealed that panelists with *low* and *low high* income types made up at least 40% in all clusters. This percentage is even higher in our focus clusters - Clusters 2 and 3. It is generally known that healthier food choices can be more expensive than their unhealthier counterparts. Therefore, we must actively cater to the lower income population. We trust providing coupons to offset additional costs would be a strong incentive.

|  |
| --- |
| Figure 42: Example of a coupon |

These coupons will give users a discount if they purchase a certain amount of healthy food items. A key feature of the coupon would also be the recommendation of a healthy food item. For example, the coupon sample in Figure 42 recommends oat milk, giving consumers an idea of a healthier food option.

Through the coupon system, we hope to enable and encourage those with lower income to give healthier foods options a try and slowly change their consumption habits. It would also ensure inclusiveness of those with lower usage of the app.

We will be disseminating the food coupons through the neighbourhood community internet centres. ID proof will be required to allow only low income individuals and prevent abuse of this system. This collection would be done once a month to minimise the costs of administration.

## Media Publicity

Furthermore, our media usage analysis (Figure 32) revealed that television and radio are the most frequently used media channels. Hence, leveraging on these channels is an efficient way to maximise our reach. Broadcasts will be made through these media to promote the mobile application and coupons, as well as, suggestions on how to eat healthily.

Based on our market basket analysis in the sections prior, clusters 2 and 3 often cook at home and are high consumers of carbohydrates and cooking essentials. In order to cater to these clusters, additional content about healthier food options under the carbohydrate and cooking essential food groups will also be featured.

Radio broadcasts will typically take place in the morning due to the high listenership. Since each ethnicity has clear preferences for radio stations, we ensure that a radio station for each is chosen to feature the media content; THR Raaga for Others, ERA for Malay, MyFM for Chinese. The application will also be advertised through television on TV1 and TV3 during the Prime Time or evening, which has the most viewership per channel per timing.

There will also be sponsored television and radio contests by the government to promote healthier living and cost saving, featuring trivia on healthy foods and guessing the prices of cheaper and healthier packaged foods. Prizes will then be provided by the companies who distribute these packaged goods. Hence, a greater number of individuals will also get to know new alternatives which would be cheaper and healthier for their consumption.

Through the comprehensive media strategy developed by capitalising on the insights from our cluster analysis, we have confidence that our media strategy will be successful in encouraging the general population to pick up healthier consumption habits and hence, live a healthier life.

# 

# 8**. Evaluation**

## Evaluation of Recommender System

Understandably, it is unrealistic to think that everyone would use our application or take up our recommendations. We shall conservatively **assume that 20% of the targeted panelists use the application and are receptive** **towards the recommendations** for the purpose of evaluating our strategy.

If the 20% of the panelists do take up our recommendation and make changes to their basket and purchases, the total change in percentage for each of the five macronutrients will be as shown below in Table 1. We exclude the calculation of trans-fat as we only have information on the presence of trans-fat but not the specific amount present.

Wholegrains (Fibre) Per Serving 4.76464

Calcium Per Serving (Scaled by 10) -47.3271

Sugar Per Serving (Scaled by 10) -32.3425

Sodium Per Serving (Scaled by 100) 0.326619

Saturated Fats Per Serving (g) -64.2958

| Table 1: Percentage Change of the 5 Macronutrients |
| --- |

As can be seen in the table above, by following the recommended food items from our recommendation system, we have successfully reduced the sugar intake by 32% and the saturated fats intake by 64%.

Since the majority of the unhealthy individuals from Clusters 2 and 3 are either "Obese" or "Over Weight", by reducing the amount of sugar and saturated fat consumed by these clusters, a positive effect on the BMI can be expected with high likelihood, shifting these consumers towards the "Healthy" category.

Based on our earlier assumption that the panelists accurately represent the general Malaysian population, these results can be similarly expected in the general Malaysian population.

## Evaluation of Media Plan

To further enhance the adoption of our *Sihat* app and the coupons, we can conduct a period of discount for nutritionally healthier goods. This will generate hype and create an impetus for consumers to adopt our solutions.

# 9. Future Works

## Scalability of Recommender System

The dataset only has 62 unique items and this is not an ideal representation of an actual mart. Realistically, in bigger marts, there will be a greater number of specific items which can range from 1,000 to 10,000. In such cases, our recommender system becomes a scalable solution as it allows the application to increase the recommendations returned by adding the necessary inputs for the system to work. Due to the small number of unique items we are working with, not many items have direct substitutes within the dataset.

To illustrate the potential of our recommender system, we have decided to gather additional unique items to show how well it would perform in such cases.



Figure 43: Output of Recommendation System with additional items

From Figure 43, we can see that the recommendations are much more sensible when we have actual substitutes. With this, we believe that our solution would be able to perform far better on a larger database of items, while ensuring scalability and accuracy.

# 

# 10**. References**

1. *Guidelines on healthier choice logo malaysia*. (n.d.). Retrieved October 19, 2021, from https://myhcl.moh.gov.my/assets/doc/guidelines.pdf.
2. *Word mover's embedding: Universal text embedding from word2vec*. IBM Research Blog. (2019, February 7). Retrieved October 19, 2021, from https://www.ibm.com/blogs/research/2018/11/word-movers-embedding/.