UNIVERSITI MALAYA

-

**COURSE**

**SIT3017 STATISTICAL LEARNING AND DATA MINING 2023 / 2024**

**WORK TITLE**

**GROUP PROJECT 1 (GROUP 6)**

**REPORT TITLE**

# THE ANALYSIS OF FOOD PRICES IN FEB 2022 USING K-MEANS CLUSTERING METHOD

**LECTURER**

**PROFESOR DR. IBRAHIM BIN MOHAMED**


| No. | NAME | MATRIC NO. |
|---|---|---|
| 1. | NUR NAJLA NABILA BT AZMAN | U2001004 |
| 2. | MUHAMMAD IMRAN BIN MOHD ISA | U2000717 |
| 3. | CHIN YEE WEI | U2103394 |
| 4. | KOO HOONG KHEN | U2103676 |
| 5. | PANG WILSON | S2132422 |

**TABLE OF CONTENTS**

**LIST OF TABLES**

**LIST OF FIGURES**

## 1.0 Introduction

There has been a lot of debate about whether the food or grocery prices in Malaysia are getting more expensive for the past few years. According to the Global Food Security Index (GFSI), Malaysia is ranked number 30 out of the whole world for the affordability of food prices (Economist Impact, 2022). This statistic is a really unbelievable number considering that there are another 112 countries that we are being compared to. However, we had dropped 6 positions from 2019 which we were ranked 24th (KPKM, 2022). This shows that the price of food is getting more expensive throughout the years. However, if we look at the food prices deeper into states or districts, we can actually observe that there is a significant difference in the food prices, some of the places have a higher than average while some of the have a lower food price. There must be an underlying factor that causes this to happen.

## 2.0 Background of The Data

This report used the dataset retrieved from the OpenDOSM website. It contains over a million food prices of item categories by districts and states in Malaysia for February 2022. The data was obtained from the PriceCatcher mobile app created by The Ministry of Domestic Trade and Cost of Living (KPDN, formerly KPDNHEP), to assist consumers in comparing the costs of essential commodities in their neighbourhood. This data comes along with the lookup tables to be left joined with the primary PriceCatcher data's item code and premise code. The whole data can be retrieved here. The early six observations of the coded data by states and districts can be seen in Appendix A and Appendix B respectively.

There are total of 13 states and 3 Federal Territory in Malaysia which are Johor, Kedah, Melaka, Negeri Sembilan, Pahang, Perak, Pulau Pinang, Kelantan, Perlis, Terengganu, Selangor, Sabah, Sarawak, Kuala Lumpur, Federal Territory of Putrajaya and Federal Territory of Labuan. Besides that, there are 8 item groups that contain the respective categories as shown in Table 1.

**Table 1 : The Items' Categories**

| Item Groups | Item Category | Item Groups | Item Category |
|---|---|---|---|
| 1.Packaged foods | Rice, Vermicelli, Esen and yeast, Sugar, Canned fish, Soy sauce and sauce, Creamer and powdered milk, Butter, Instant noodles, Oil and fat, Spices (packaged), Spices (unpackaged), Bread, Coconut milk (box), Spreads, Flour | 2.Convenience store items | Stationery and reading materials, Toothbrushes, Biscuits, Chocolates, Magazines, Snacks, Fast food, Drinks, Mouth wash, Mosquito repellent, Home fragrance, Body soap, Shampoo, Tissues, Towels, Toothpaste, Medicines |
| 3.Dried foods | Onion, Dried Chili, Dried seafood, Nuts, Potatoes | 4.Fresh goods | Chicken, Seafood, Fruits, Meat, Landfish, Coconut, Noodles/Kuetiau, Vegetables, Tofu and tempeh, Eggs |
| 5.Ready-to-eat | Rice, Drinks, Side Dishes, Noodles/ Vermicelli/ Kueytiaw, Others | 6.Beverages | Beverage ingredients, Ready to drink |
| 7.Cleaning products | Self-care, Home care | 8.Baby products | Disposable diapers, Baby food, Baby milk |

## 3.0 Result

### 3.1 The K-means Clustering by States

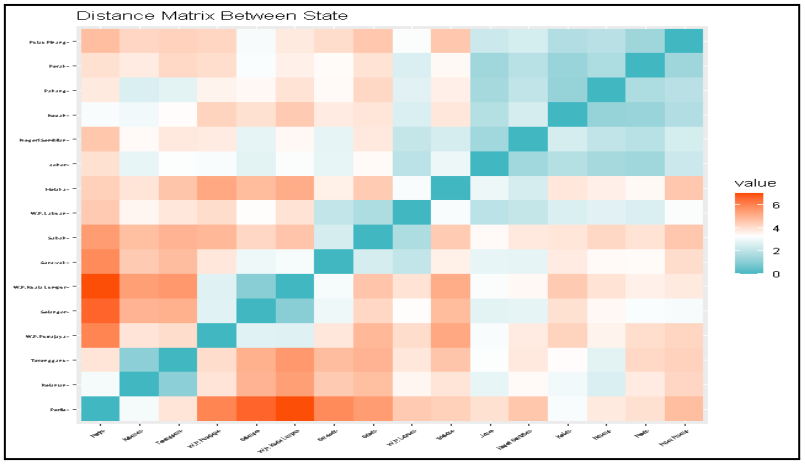



**Figure 1 : Distance Matrix between States**

Figure 1 shows the distance matrix heatmap between 16 states in Euclidean distance. From the figure, we can see that there are three colours in the graph which are blue, white and red. When the colours are near to blue, it shows that the distance between two observations are near (high similarity). When the colour is near to red, this means that the two observations have a very long distance (low similarity). From Figure 1, we know that the Federal Territory of Kuala Lumpur and Perlis have a long distance which indicates low similarities while the Federal Territory of Labuan and Sarawak show low distance (high similarities). We can also see that Johor and Terengganu show white colour which indicates that Johor and Terengganu have neutral similarities.

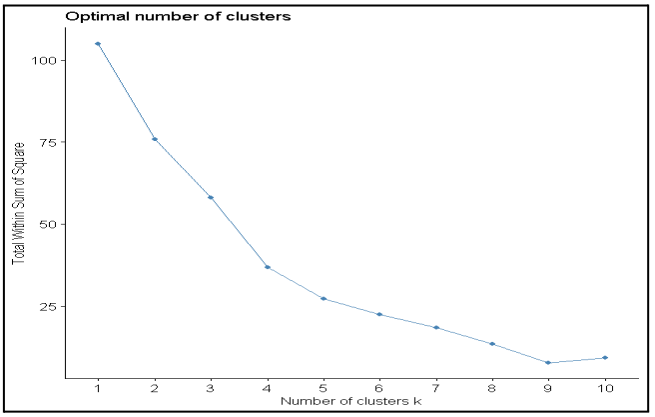



**Figure 2 : Elbow Method Plot for States**

Figure 2 shows the elbow method plot for clustering by states. The elbow method is a graphical illustration of the process of determining the best 'k' in a K-means clustering. From Figure 2 , we cannot claim any value for k is an optimal number of clusters since the elbow shape cannot be observed obviously. However, we observed that the plot between the total within sum of squares for k=3 and k=4 has decreased more steeply than the plot between k=4 and k=5. Hence, we might expect that the optimal number of clusters, k for this K-means clustering method is k=4 or k=5.

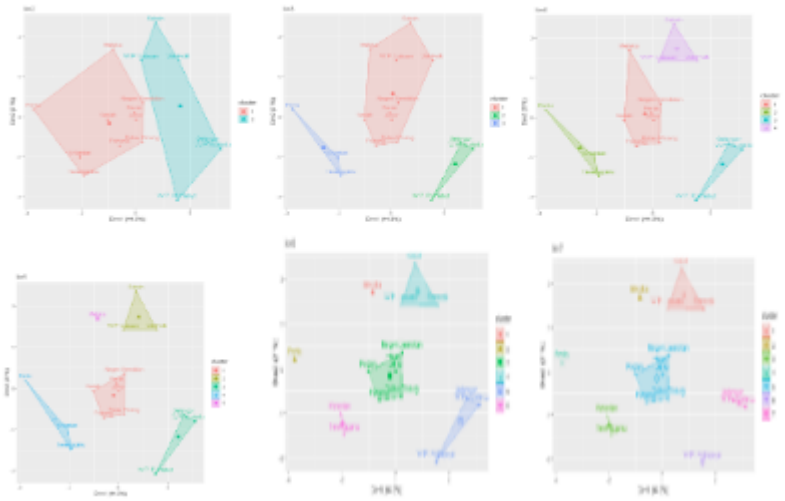


**Figure 3 : The Clustering Plot by States**

Figure 3 illustrates the comparison of clustering plots for k equal to two until seven. From the figure, we observe that there is no overlapping for all of the number of clusters, k. However, starting from k=5, the clustering plot shows a condition or a ‘dot’ where there is just one state belonging to a cluster. In this case, the clustering plot with k=5 has one ‘dot’ which is Melaka, the clustering plot with k=6 shows two ‘dots which are Melaka and Perlis and the clustering plot with k=7 shows three ‘dots’ which are Melaka, perlis and Federal Territory of Putrajaya. We expect that the number of ‘dots’ will increase as the number of clusters increases, and hence the number of clusters, k=5 onwards are no longer the optimal number of clusters for this K-means Clustering method. So, we construct the test to observe the variation to choose the optimal number of clusters for k=2, k=3 and k=4.

**Table 2 : Proportion of Variance,$R^2$ values for each clusters**

| Cluster | $R^2$ (%) |
|---|---|
| 2 | 29.1 |
| 3 | 49.3 |
| 4 | 64.9 |

Table 2 shows the $R^2$ value in percentage for k=2, k=3 and k=4. From Table 1, the $R^2$ for k=2 is 29.1%, the $R^2$ for k=3 is 49.3.1% and the $R^2$ for k=4 is 64.9%. Since the number of clusters, k=4 has the highest $R^2$ compared to the $R^2$ for k=2 and k=3 which indicates the best fit, hence, we decided to take the optimal number of clusters, k is 4.

**Table 3 : States's Cluster Means for 7 Variables**

| Cluster | Packaged Goods | Dried Goods | Fresh Goods | Ready-to Eat Food | Beverages | Cleaning Products | Milk and Baby Products |
|---|---|---|---|---|---|---|---|
| 1 | -0.350 | 0.006 | -0.043 | -0.868 | -0.205 | -0.108 | -0.018 |
| 2 | -0.996 | -0.710 | -1.003 | 0.326 | 1.524 | -1.066 | -1.378 |
| 3 | 0.023 | -0.293 | 1.543 | 1.179 | -0.750 | 1.483 | 1.022 |
| 4 | 1.789 | 0.988 | -0.440 | 0.520 | -0.297 | -0.166 | 0.399 |

Table 3 shows the cluster means for 7 different prices variables which are Packaged Goods, Dried Goods, Fresh Goods, Ready-to-eat Food, Beverages, Cleaning Products and lastly Milk and Baby Products for 4 different clusters. From Table 3, the first cluster is the most moderate or average states amongst all as it has almost the average prices for most of the categories, despite their cheapest ready-to-eat food. The second cluster has the lowest prices in almost all categories, except the averagely-priced ready-to-eat food and unexpectedly most expensive beverages. The third cluster has the highest prices in essential needs categories such as fresh goods, ready-to-eat food, cleaning products, milk and baby products, the lowest beverages price, and other categories being averagely-priced. The prices in the fourth cluster are extremely expensive for packaged and dried goods, others are moderate.

**Table 4 : The State Classifications for 4 Different Clusters**

| Cluster | States |
|---|---|
| 1 | Johor, Kedah, Melaka, Negeri Sembilan, Pahang, Perak, Pulau Pinang |
| 2 | Kelantan, Perlis, Terengganu |
| 3 | Selangor, Kuala Lumpur, Federal Territory of Putrajaya |
| 4 | Sabah, Sarawak, Federal Territory of Labuan |

Table 4 shows the state classifications for 4 different clusters. There are 7 states in the first cluster which are Johor, Kedah, Melaka, Negeri Sembilan, Pahang, Perak and Pulau Pinang. The second cluster has 3 states which are Kelantan, Perlis and Terengganu. The third cluster can be categorised as the Klang Valley since the states located in the third cluster are Selangor, Kuala Lumpur and Federal Territory of Putrajaya. Lastly, cluster 4 can be categorised as the Borneo states since the states located in cluster 4 are Sabah, Sarawak and Federal Territory of Labuan.

### 3.2 The K-means Clustering by Districts

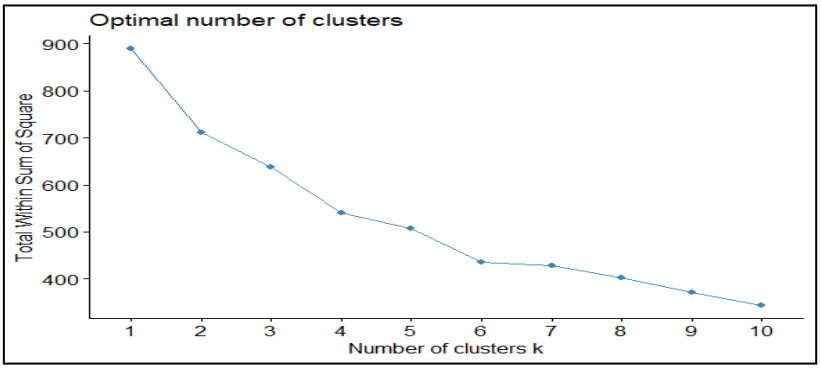



**Figure 4 : Elbow Method Plot for Districts**

Figure 4 shows the elbow method plot for clustering by districts. The elbow method is a graphical illustration of the process of determining the best 'k' in a K-means clustering. From Figure 5 , we cannot claim any value for k is an optimal number of clusters since the elbow shape cannot be observed obviously. However, we observed that the plot between the total within sum of squares for k=3 and k=4 has decreased more steeply than the plot between k=4 and k=5. Hence, we might expect that the optimal number of clusters, k for this K-means clustering method is k=4 or k=5.

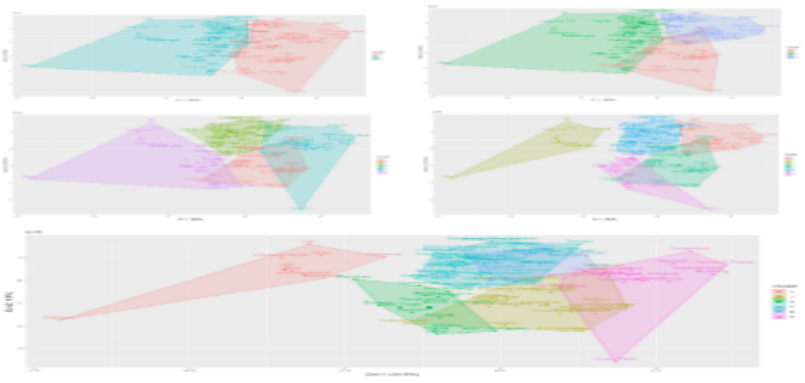


**Figure 5 : The Clustering Plot by Districts**

Figure 5 shows the cluster plots by districts for five different values of k. From the plots, we can notice that slight overlapping exists for k = 2 and k = 3, while the other plots show more obvious overlapping. However, from our previous analysis of the total within sum of squares for the different values of k, we found that the likely optimal number of clusters for the data is k=4 or k=5 by using the elbow method. Some overlapping between the clusters might be unavoidable as there exist several outliers in our dataset as well as the fact that points near the boundary of each cluster might be similar to those in other clusters. Thus, we should aim to minimize the total within sum of squares of the clusters and so despite the overlapping, we should choose k = 5 as the optimal number of clusters as it has a lower total within the sum of squares when compared to k=4.

**Table 5 : Districts's Cluster Means for 7 Variables**

| Cluster | Packaged Goods | Dried Goods | Fresh Goods | Ready-to-eat Food | Beverages | Cleaning Products | Milk and Baby Products |
|---|---|---|---|---|---|---|---|
| 1 | -0.010 | 0.566 | 1.446 | 0.349 | -0.678 | 1.240 | 0.808 |
| 2 | -1.485 | -0.834 | -0.951 | -0.421 | 0.740 | -1.770 | -2.008 |
| 3 | 1.231 | 0.908 | -0.452 | 0.450 | -0.175 | -0.174 | 0.411 |
| 4 | -0.275 | -0.414 | -0.016 | -0.639 | -0.161 | -0.036 | -0.112 |
| 5 | -0.305 | -0.441 | -0.490 | 1.253 | 1.362 | -0.104 | -0.156 |

Table 5 shows the cluster means for 7 different prices variables which are Packaged Goods, Dried Goods, Fresh Goods, Ready-to-eat Food, Beverages, Cleaning Products and lastly Milk and Baby Products for 5 different clusters. From Table 5, the first cluster has the most expensive fresh goods, cleaning products, milk and baby products, slightly expensive dried goods, cheapest beverages and

average packaged goods and ready-to-eat food. The second cluster has the lowest price in almost all categories, except a higher price for beverages. The third cluster has the most expensive packaged goods and dried goods, moderately-priced beverages and cleaning products, slightly higher price for ready-to-eat food, milk and baby products, and a slightly lower price for fresh goods. The fourth cluster has the cheapest ready-to-eat food, with other categories priced lower than the average price by a tiny margin. Significantly, the fifth cluster has the highest price for ready-to-eat food and beverages, others being priced slightly lower than the average market price.

## 4.0 Discussion

### 4.1 The K-means Clustering by States

There are many factors that influence the price clustering such as the economic activities, the population and the Consumer Price Index (CPI) for states in Malaysia. States in the third cluster, such as Selangor and Kuala Lumpur, are economic hubs with a strong focus on manufacturing and services (The Malaysian Reserve, 2023). Plus, Kuala Lumpur recorded the highest population density with 8157 persons per square kilometre followed by 2215 persons per square kilometre in Putrajaya in February 2022. Meanwhile, the urbanisation rate for Selangor is 95.8% compared to other years (Nur Hanani, 2022).These factors contribute to higher prices for essential goods, as there's a higher demand for quality products in these areas. Other than that, the food and beverages ' CPI for Selangor and Putrajaya are among the highest compared to other states due to the inflation, resulting in higher food prices in these states compared to the other states in Malaysia. This information can be referred to in Appendix C.

The second cluster states are more rural, agricultural and almost no metropolitan area. These states are known for fishing, cultivating rice and cultivating fruits like Harum Manis in Perlis due to the similar soil types existing in some of these regions which might explain their lower prices for most categories. Plus, the CPI for food and beverages for these states are on the average indicating low average foods and goods' price for this cluster. Before the discovery of oil and gas reserves, the main Terengganu economic activities used to be fishing and farming. Because it's a holiday destination, the tourism industry is also a major economic bolster in the state. The economy of Perlis depends largely on forestry, agriculture and fishing industries. It is one of the largest sugarcane producers in Malaysia (IIM, n.d.). Kelantan's main economic activities are centered on agriculture, urban tourism, ecotourism, arts, culture and heritage tourism as well as manufacturing activities to further promote economic growth (ECERDC, 2020).

Even so, the states in the first cluster show lower Ready-to-Eat Food and Beverages' price than the states in the second cluster even though the states in the first cluster do have metropolitan areas, higher populations and the fact that the states in the first cluster have higher food and beverages CPI compared to the states in second cluster. This is due to the monsoon season that occurs from November 2021 until March 2022 in most of the states in the second cluster that causes the sellers to increase their selling price due to insufficient stocks (Wan Zahirah, 2021).

The fourth cluster, consisting of states in East Malaysia, faces additional transportation costs due to their geographic isolation and may lead to higher prices for packaged and dried goods, as these products often need to be transported over longer distances. This is due to the Cabotage law that required the imported goods to be sent in West Malaysia first before being sent to East Malaysia to fulfil the condition where the movement of goods between two ports must be handled by Malaysian-Owned Shipping Companies (*What Is Cabotage and What Are Cabotage Laws? | Teleroute*, n.d.). The moderate prices in other categories may be due to a mix of local production and imports, as a result from the lowest overall CPI, indicating lowest inflation for states locating in this cluster.

In summary, diverse economic activities, population density, and regional factors such as the monsoon season and Cabotage laws collectively contribute to distinct pricing patterns across different clusters of states in Malaysia, highlighting the intricate interplay of local conditions on consumer goods and services.

### 4.2 The K-means Clustering by Districts

When we perform a deep dive into the districts rather than having an overlook at the states, we notice some similarities between the districts in each cluster. Most notably, Cluster 3 which has the most expensive packaged goods and dried goods mainly consists of districts from East Malaysia with the exception of Melaka Tengah and Tampin. As mentioned in the discussion on states above, this is likely because of the additional costs associated with transporting goods to East Malaysia. (Kota Marudu, Sri Aman, Tampin, Melaka Tengah, Kawasan Luar Bandar, Putatan).

Cluster 2 is the only cluster that has no overlapping with the other clusters, It mainly consists of districts which are known for their agricultural activity. This cluster has the lowest price in almost all categories, except a higher price for beverages. This is likely due to the fact that these districts with high agricultural activity produce many of their goods locally, which reduces transportation costs and leads to a large supply of goods, ultimately resulting in a lower cost for consumers in those districts (PlanMalaysia, 2022).

Cluster 1 consists of major cities and economic hubs in both Western and Eastern Malaysia. This cluster is categorised by very high prices for fresh goods, cleaning products and milk-and-baby products due to the high inflation that occurs in almost all of the districts. However, from Figure 6, this cluster shows an overlap with cluster 3 due to the outlier which is the Federal Territory of Putrajaya. This district is believed to have higher prices or packaged goods and dried goods compared

to the other districts in Cluster 1. One of the reasons may be that this district is built mainly for government administrative purposes, and it represents the face of the country (Sarah M., 2010). Despite the expensive prices for other categories, Putrajaya tends to have more expensive packaged goods and dried goods since it has few factories there to process them (ProQuest, 2022).

Cluster 4 encompasses most districts, characterised by average prices across all categories. This cluster represents a middle ground in terms of pricing, indicating a balance between various economic activities and consumer demands. The districts in this cluster may not face extreme challenges in terms of transportation costs, and they exhibit a relatively stable pricing structure across different categories. Pantai Tengah is an outlier in this cluster that exhibits the same characteristics as in Cluster 1, which are the major cities and economic hubs. The district in Langkawi is mainly a tourist spot, locals tend to sell things more expensive than other places as tourism is their main income source.

Cluster 5 exhibits a unique pricing profile, featuring below-average costs for essential goods like packaged and dried goods, fresh goods, and cleaning products. However, it stands out with above-average prices for ready-to-eat food and beverages, suggesting a market dynamic shaped by specialised preferences or a higher demand for convenience. This cluster may represent districts with a focus on specialised cuisines or gourmet options, leading to premium prices in the ready-to-eat and beverage categories.

In conclusion, a detailed examination of districts reveals distinct pricing patterns within each cluster, with notable factors such as transportation costs influencing higher prices in Cluster 3, agricultural self-sufficiency contributing to lower prices in Cluster 2, major cities experiencing inflationary pressures in Cluster 1, a middle-ground pricing balance in Cluster 4, and unique consumer preferences or demand dynamics shaping the pricing profile of above-average prices for ready-to-eat food and beverages in Cluster 5.

## References


Awati, R. heat map (heatmap). https://www.techtarget.com/searchbusinessanalytics/definition/heat-map

*Global Food Security Index (GFSI)*. (n.d.). impact.economist.com https://impact.economist.com/sustainability/project/food-security-index/

"Kelantan - ECERDC." *ECERDC*, 27 Aug. 2018, www.ecerdc.com.my/about-ecer/kelantan/#:~:text=Kelantan. Accessed 15 Nov. 2023.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, *1*(6), 90-95. https://www.researchgate.net/profile/Trupti-Kodinariya/publication/313554124_Review_on_Determining_of_Cluster_in_K-means_Clustering/links/5789fda408ae59aa667931d2/Review-on-Determining-of-Cluster-in-K-means-Clustering.pdf

Marzuki, A., & Jais, A. S. (2020). URBANISATION AND THE CONCERNS FOR FOOD SECURITY IN MALAYSIA. *PLANNING MALAYSIA*, *18*(13). https://doi.org/10.21837/pm.v18i13.786

Mengyao Cui, Introduction to the K-Means Clustering Algorithm Based on the Elbow Method. Geoscience and Remote Sensing (2020) Vol. 3: 9-16. DOI:http://dx.doi.org/10.23977/geors.2020.030102.

Moser, S. (2010). Putrajaya: Malaysia's new federal administrative capital. *Cities*, *27*(4), 285–297. https://doi.org/10.1016/j.cities.2009.11.002

Murdad, R., Muhiddin, M., Osman, W. H., Tajidin, N. E., Haida, Z., Awang, A., & Jalloh, M. B. (2022). Ensuring Urban Food Security in Malaysia during the COVID-19 Pandemic—Is Urban Farming the Answer? A Review. *Sustainability*, *14*(7), 4155. https://doi.org/10.3390/su14074155

Nur Hanani, A. (2022, February 14). *Malaysia's population stands at 32.4m* [Review of *Malaysia's population stands at 32.4m*]. The Malaysia Reserve. Retrieved from https://themalaysianreserve.com/2022/02/14/malaysias-population-stands-at-32-4m/

*Perlis - Info Malaysia (IIM) Leading Industrial, Commercial, Tourism & Information in Malaysia.* (n.d.). Www.iim.com.my. Retrieved November 15, 2023, from https://www.iim.com.my/state-of-malaysia/about-perlis.html#:~:text=Perlis%20econoy

*Portal Rasmi Kementerian Pertanian dan Keterjaminan Makanan (KPKM)*. (n.d.). www.kpkm.gov.my https://www.kpkm.gov.my/en/gfsi-2022

*Portal Rasmi PLANMalaysia - Utama*. (n.d.). Www.planmalaysia.gov.my. Retrieved November 15, 2023, from https://www.planmalaysia.gov.my/index.php/pages/view/520?cats=10

*Terengganu - Info Malaysia (IIM) Leading Industrial, Commercial, Tourism & Information in Malaysia.* (n.d.). https://www.iim.com.my/state-of-malaysia/about-terengganu.html#:~:text=Economy%20of%20 Terengganu Text=Before%20the%20discovery%20of%20 oil

Wan Zahirah, W. Z. I. (2021, December 6). *Musim tengkujuh: Harga ikan naik di Terengganu* [Review of *Musim tengkujuh: Harga ikan naik di Terengganu*]. Berita Malaysia; Astro Awani.https://www.astroawani.com/berita-malaysia/musim-tengkujuh-harga-ikan-naik-di-terengganu-334740

*What Is Cabotage and What Are Cabotage Laws? | Teleroute*. (n.d.). Teleroute.com. https://teleroute.com/en-en/resources/glossary/cabotage/

## Appendices

Appendix A

**Table 6 : First 6 Observations from Coded Data by States**

| | Packaged Goods | Dried Goods | Fresh Goods | Ready-to Eat Food | Beverages | Cleaning Products | Milk and Baby Products |
|---|---|---|---|---|---|---|---|
| Johor | -0.363 | 0.246 | 0.006 | -0.121 | -0.196 | 0.333 | -0.237 |
| Kedah | -0.064 | -0.902 | -0.762 | -1.02 | -0.338 | -0.295 | -0.623 |
| Kelantan | -0.944 | -0.602 | -0.603 | 0.780 | 1.853 | -0.804 | -0.955 |
| Melaka | -0.503 | 2.395 | 0.606 | -0.931 | 0.698 | -1.199 | -0.389 |
| Negeri Sembilan | -0.502 | 0.662 | 0.305 | -0.621 | -0.063 | -0.320 | 0.881 |
| Pahang | -0.269 | -0.95 | -0.085 | -0.794 | 0.503 | 0.113 | -0.138 |

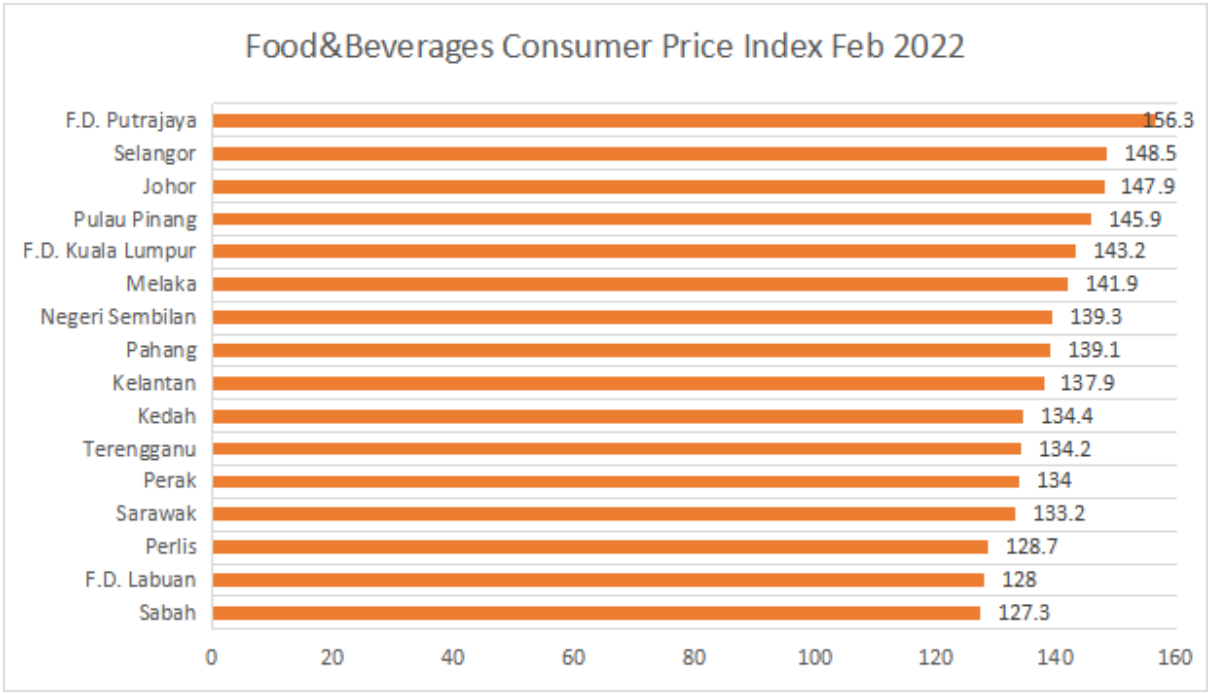
Appendix B

**Table 7 : First 6 Observations from Coded Data by Districts**

| | Packaged Goods | Dried Goods | Fresh Goods | Ready-to Eat Food | Beverages | Cleaning Products | Milk and Baby Products |
|---|---|---|---|---|---|---|---|
| Alor Gajah | -1.315 | 2.640 | 0.726 | -1.163 | 1.279 | -1.050 | 0.490 |
| Arau | -0.761 | -0.074 | -1.275 | -1.351 | -0.294 | -1.990 | -1.964 |
| Bachok | -0.380 | -0.789 | -1.570 | 0.915 | 1.280 | 0.190 | -0.446 |
| Bagan Datuk | 1.425 | -2.072 | -0.263 | -1.545 | -0.585 | 2.462 | -0.361 |
| Bahau | 0.178 | -0.367 | 0.067 | -0.853 | -0.036 | -0.873 | -0.723 |
| Baling | 0.464 | -1.080 | -0.402 | -0.818 | -0.849 | -0.054 | 0.136 |

Appendix C

**Figure 6 : Food and Beverages' Consumer Price Index in Feb 2022**




Source : OpenDOSM Website