

Comparative Analysis of Clustering Techniques for Personalized Food Kit Distribution

Jude Francis¹, Rowan K Baby², Jacob Abraham² and Ajmal P.S¹

¹ College of Engineering Trivandrum, Dept. of Mechanical Engineering, Kerala, India

² College of Engineering Trivandrum, Dept. of Computer Science and Engineering, Kerala, India

Abstract— The Government of Kerala had increased the frequency of supply of free food kits owing to the pandemic, however, these items were static and not indicative of the personal preferences of the consumers. This paper conducts a comparative analysis of various clustering techniques on a scaled-down version of a real-world dataset obtained through a conjoint analysis-based survey. Clustering carried out by centroid-based methods such as k means is analyzed and the results are plotted along with SVD, and finally, a conclusion is reached as to which among the two is better. Once the clusters have been formulated, commodities are also decided upon for each cluster. Also, clustering is further enhanced by reassignment, based on a specific cluster loss threshold. Thus, the most efficacious clustering technique for designing a food kit tailored to the needs of individuals is finally obtained.

Keywords— Conjoint- Analysis, Clustering, K-means, Singular Value Decomposition, singular vectors, SVD signs method

I. INTRODUCTION

The most delicate sections of society often comprise individuals or families ridden with malnutrition, hunger, and those that are inextricably entangled within the vicious circle of poverty^[1]. Hence, most governing authorities employ poverty thresholds to assess whether subsidies have to be sanctioned to such deserving candidates.^[2] And one state that is spearheading the emancipation of individuals below the poverty line in India, is Kerala^[3].

With one of the highest literacy rates in the entire country-Kerala^[4] had recently caught the eye of the global stage through the exceptional way in which it had managed the challenges imposed by the first wave of the pandemic through certain ingenious strategies^[5]. And amidst the plethora of health and safety-related actions taken, one such commendable effort was the government's decision to provide kits containing necessary food items, free of cost,^[6] to all ration card holders. Herein a solution has been devised to improve the disbursement of provisions in a more effective manner by employing a specific type of clustering algorithm that suits the task at hand.

II. MOTIVATION

Prior to the pandemic, the food distribution system was solely focused on providing a means to establish a steady and reliant channel of the food supply through government-intervened subsidization. The rate of subsidies was classified into various categories according to the financial capability

of the individuals in each category^[7]. Thus, the underprivileged, who were the major beneficiaries of this system could easily access sufficient food.

However, food kits containing additional commodities were mostly given only during traditional festive seasons. And when the livelihoods of people were threatened due to the advent of COVID, the government increased the frequency of kits supplied, thereby ensuring that they would act as a safeguard for the common mass, assuring them that their basic necessities would be met.

However, the problem with the implementation of such a system was that the commodities included in these kits were not decided upon by carrying out any statistical analysis, nor were there any mechanisms set into motion, to dynamically alter its constituents according to the tastes and trend patterns of its consumers.

Moreover, another inherent problem was that the consumption data of such food kits could never reflect the innate preferences of its recipients, as they would rather accept food kits offered freely than simply refuse them. Hence, the administration could do nothing but impose a set of items that were commonly thought of as necessary, without any underlying studies or analysis to back up their decisions. Thus, the food kits offered were not able to satisfy the demands of the people, and that is why a new solution had to be devised in order to address this issue.

III. PROBLEM STATEMENT

In order to bridge the gap between the consumer and the food kit facilitator, one needs to ensure that the choices made by the facilitator align with the preferences of the consumer. Though it may seem that such a task could be easily carried out by just asking the consumer directly for his preference,

the difficulty in actualizing this is mainly due to the following reasons: Firstly, it is impossible to allocate a unique kit to each individual, as it would lead to innumerable combinations. Thus, it must be conceded that only a limited number of kits can be allocated, which inevitably causes a mismatch between the interests of the individual and possible kit combinations. Secondly, the preferences of the consumers alone would not yield a result capable of solving this crisis, as the data collected would be spread out depending on multiple factors ranging from geographical to climatic conditions. Therefore, in order to accommodate the multifarious nature of the consumption pattern of various communities, one would inevitably conclude that a personalized kit should be designed from an optimal number of combinations. And once this has been fixed the next question would be to decide upon the level of personalization that the government has to adopt, which is not only viable from a theoretical but also an economical viewpoint.

So in this study, a system has been proposed to address the aforementioned problems, by carrying out a replica of the scaled-down version of the whole situation using a sampled population, and analyzing how various methods fare in providing satisfactory results.

IV. SOLUTION APPROACH

a. Constraints

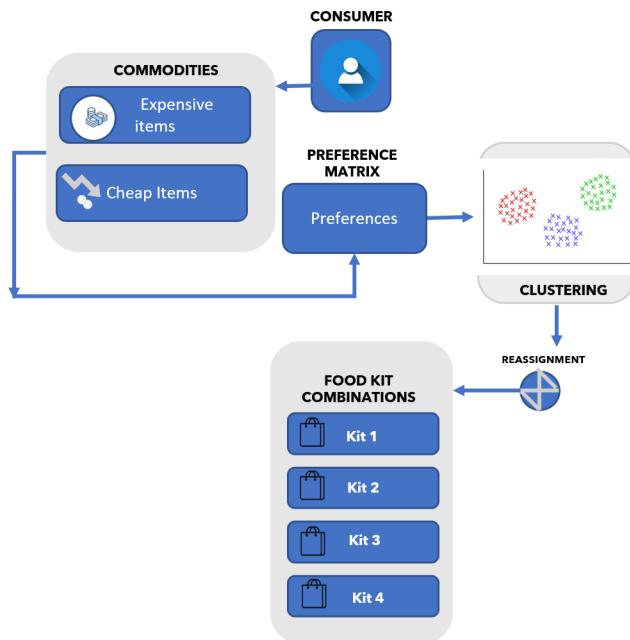


Fig. 1: Solution Flowchart

The first stage of the solution was to assess the preference of individuals so as to ensure that the decisions taken were actually fine. So a random convenience sampling of about 200 people from Kerala was taken, who willingly volunteered for a choice-based conjoint analysis of various potential commodities that the government was willing to offer. The conjoint analysis would effectively help in gauging the importance of items based on psychological trade-offs taken by the consumer in order to maximize the net benefits ob-

tained.

The choices for the scaled-down study were kept at 20 essential items out of which these items were further added to two categories expensive and cheap, in order to prevent people from selecting just the expensive items, rather than what they actually needed. The participants were allowed to select only 6 items from the expensive category and 4 from the cheap category, resulting in a total of 10 items.

After the data collection was carried out, two choices had to be made whether an item-based approach or a user-based approach had to be adopted.

b. Clustering Algorithms

The various types of clustering algorithms are as follows:

Hierarchical Clustering

- Hierarchical clustering creates a tree of clusters.
- It is well suited to hierarchical data, such as taxonomies.
- Any number of clusters can be chosen by cutting the tree at the right level.

Centroid-based Clustering

- Centroid-based clustering organizes the data into non-hierarchical clusters, in contrast to hierarchical clustering. k-means is the most widely-used centroid-based clustering algorithm.
- Centroid-based algorithms are efficient but sensitive to initial conditions and outliers

Density-based Clustering

- Density-based clustering connects areas of high example density into clusters. This allows for arbitrary-shaped distributions as long as dense areas can be connected.
- These algorithms have difficulty with data of varying densities and high dimensions. Further, by design, these algorithms do not assign outliers to clusters.

Distribution-based Clustering

- This clustering approach assumes data is composed of distributions, such as Gaussian distributions.
- As the distance from the distribution's center increases, the probability that a point belongs to the distribution decreases.

And for this problem statement, since the requirement was not to form any hierarchical patterns, partition-based methods were the next best choice. Hence,k-means (a centroid-based algorithm) was chosen as the appropriate clustering algorithm

V. ALGORITHM USED

a. K-means algorithm

The K- means algorithm has gained quite a popularity over the years owing to its ease and simplicity of implementation, scalability, speed of convergence, and adaptability to sparse

data. Basically, it can be thought of as a gradient descent procedure, which begins at starting cluster centroids, and iteratively updates these centroids to decrease the objective function. Some of the various merit measures for evaluating the techniques applied:

1. Graphical aid to the interpretation and validation of cluster analysis.
2. The Silhouette width

Here, the **k**-means algorithm is considered along with Euclidean distance to measure the similarity of the customer's preference. The dataset is represented as n-points, with each point corresponding to whether a customer wishes a particular item to belong to the food kit. And the end goal is to ensure that the designed kits yield a maximum number of customers with a minimum level of dissatisfaction.

To attain this, '*k*' clusters are considered. Their value is altered until a satisfactory value is reached. '*K*' points ($m_1, m_2, m_3, \dots, m_k$) have to be found out, that act as the cluster centroids and the points (x_1, x_2, \dots, x_n) belonging to any arbitrary cluster '*i*' will be such that the Euclidean distance amongst the points is always less than the distance with any point outside that particular cluster.

Algorithm 1 K-Means

```

shuffleRow = randperm(size(input,1));
data = input(shuffleRow,:);           ▷ Randomize the data
Y = data(:,2:end);
K = 15;                                ▷ Centroid limit
lambda = 0.3;                          ▷ Learning Rate
while iter ≠ K do
    initial_centroids = Y(1:iter,:);
    max_iters = 100;
    idx = findClosestCentroids(Y,initial_centroids,lambda);
    centroids = computeCentroids(Y,idx,iter);
    runkMeans(Y,initial_centroids,max_iters);
    bestCentroid(Y,idx,iter);          ▷ Find the best kits
    silhouette(Y,centroids,idx,iter); ▷ Evaluate Error
    iter = iter + 1                  ▷ Evaluate Error
end while

```

For Algorithm 1 defined it takes input data. The input will contain a list of users with their corresponding preferences for the items available for the kit. The user is required to choose 10 items out of a list of *m* items. In the algorithm, the value of *K* can vary according to the needs and is analyzed as shown in the next section. The value of lambda will range from 0 to 1. The most suitable value after various trials was found to be 0.3 and it has been used here. Similarly, after various trials, it was found that the value for maximum iterations (max_iter) should have a value of 100. The methods to compute the closest centroid, running the K-Means, and finding the best centroid corresponds to the standard algorithm for K-Means.

1. Result

A preset condition of a minimum of 4 kits was already set as per the design constraint, hence the value of $k < 4$ is not considered at all. This condition was set in order to ensure

Kits	Trials		
	0.128095	0.1612425	0.1611875
5	0.139614	0.168186	0.167112
6	0.1509433333	0.2312133333	0.1858016667
7	0.1504685714	0.22687	0.1879557143
8	0.15819375	0.21523	0.1954125
9	0.1823155556	0.2282477778	0.2058988889
10	0.189969	0.222137	0.196728
11	0.1881945455	0.1992954545	0.1760672727
12	0.2241175	0.2268658333	0.1781225
13	0.2242584615	0.2309376923	0.1892176923
14	0.2435464286	0.234595	0.2099157143
15	0.2665806667	0.237452	0.2598893333

TABLE 1: SILHOUETTE WIDTH

that the people will be provided with a minimum variety to choose from as the taste of the general public will vary depending on their geographical and cultural diversity.

The given **Table 1** contains the plot for the various trials and the performance evaluation for each of the trials. This table has the first column as the number of kits and in the algorithm, it corresponds to the number of clusters. Each of the columns after that corresponds to three separate trials conducted using the K-Means algorithm. The values in these columns correspond to the Silhouette width, here the average silhouette width for the kits is taken for each iteration of the trial. This was done by taking the sum of the errors for each kit and dividing it by the total number of kits. **Fig 2** shows the performance of the K-Means in a graphical format with respect to the three trials given in Table 1.

Collectively, from both the table and the graph, one reaches the conclusion that the results obtained from this algorithm are a bit ambiguous and can be clearly seen by the variation in the lines in the **Fig 2**. We can attribute this to an error in the data that is being used or it can be the insufficiency of the algorithm used. To perform further analysis in this respect the number of kits considered in the trial is increased to a very large number.

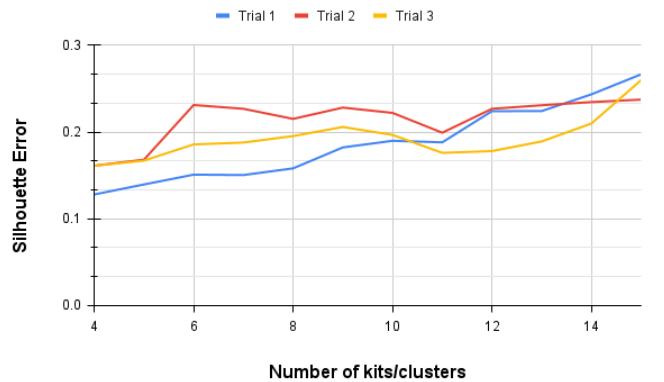


Fig. 2: Performance of K-Means

A trial is run by taking 50 kits or clusters as the limit to the number of clusters to further prove or disprove the inconsistency of the data or the algorithm that has been used. Fig 2 shows the variation in the performance of the algorithm with respect to the number of clusters used. Here from the graph, the performance is varying in a consistent manner and this

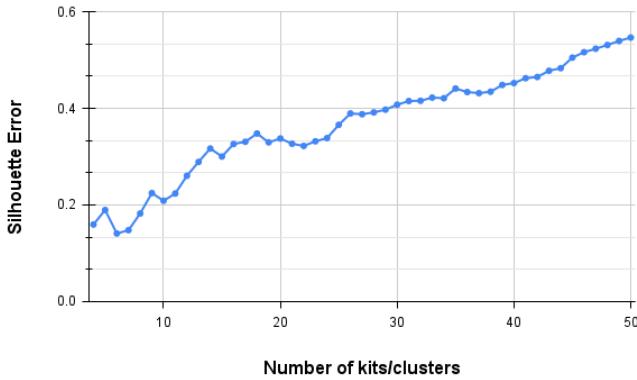


Fig. 3: Performance of K-Means

shows how the data used is consistent for this experiment.

It is noticed that as the number of clusters drops, the errors formed also reduce, but such a conclusion is naive and simply not true. Moreover, the values of the number of users in each cluster keep on varying each time the trial is run, resulting in a highly erroneous and erratic output.

To make a more conclusive and strong statement on the kits and the various contents, one has to resort to other clustering means, in order to obtain a more stream-lined and convergent conclusion

VI. SINGULAR VALUE DECOMPOSITION (SVD)

The singular value decomposition of a matrix $\mathbf{A}_{p \times q}$ with rank r is an orthogonal decomposition of a matrix into three matrices such that $\mathbf{A}_{p \times q} = \mathbf{U}_{p \times r} \Sigma_{r \times r} \mathbf{V}_{r \times q}^T$. The SVD of a matrix is an exact decomposition, thus, the original matrix can be re-obtained by multiplying \mathbf{U} , Σ , and \mathbf{V}^T

However, we can also approximate the given matrix \mathbf{A} to a rank r by multiplying the r columns of \mathbf{U} , the first r values of σ matrix and r rows of the \mathbf{V}^T matrix, resulting in a truncated SVD matrix.

In order to understand the application of **SVD** intuitively related to the problem statement at hand, it is inevitable to understand what each of the resulting 3 matrices obtained from the eigen decomposition on the dataset procured through conjoint analysis implies, and it is stated as follows:

1. The left singular matrix denoted by \mathbf{U} represents the user-latent space relationship matrix.
2. The singular value matrix denoted by Σ represents the extent of scaling.
3. The right singular matrix denoted by \mathbf{V}^T represents the item-latent space relationship matrix

And in terms of its eigen decomposition, they can be defined as follows:

1. The columns of \mathbf{V} (right-singular vectors) are eigenvectors of $\mathbf{A}^T \times \mathbf{A}$.
2. The columns of \mathbf{U} (left-singular vectors) are eigenvectors of $\mathbf{A} \times \mathbf{A}^T$.

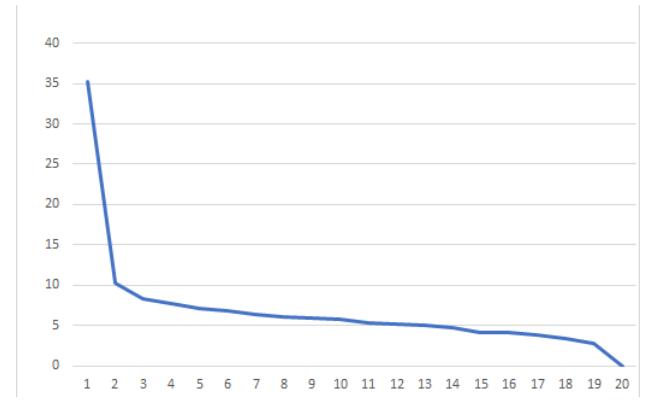


Fig. 4: Variation of σ (Y axis) with rank r (X-axis)

3. The non-zero elements of Σ (non-zero singular values) are the square roots of the non-zero eigenvalues of $\mathbf{A}^T \times \mathbf{A}$ or $\mathbf{A} \times \mathbf{A}^T$

The Σ matrix is a diagonal matrix with σ values arranged in descending order along the leading diagonal, which gives a clear indication of the strength of the scaling. And on plotting these sigma values, it is seen from **figure 4** that the difference in the consecutive values is significant only up to the range between 2- 4, and thereafter the values do not change much.

a. The reason why the item-based approach does not work?

Clustering can be done either in an item-based manner or a user-based manner. The item-based clustering is done by the analysis of the first r rows of the \mathbf{V}^T Matrix and then applying the SVD signs method, whereby the columns having the same sign patterns are clustered together. The results of such an item-based clustering are shown in **Fig.5**. It is observed that as the rank increases, a larger number of items are included in that particular cluster. However, such a pattern is not logically correct, because the increase in the number of clusters points to a failure in effectively grouping the items. Hence, as seen from **Fig 5**, once the number of clusters reaches 17 out of the 20 available items each cluster would comprise mostly of one item, which is why an item-based approach is not suitable for this problem statement.

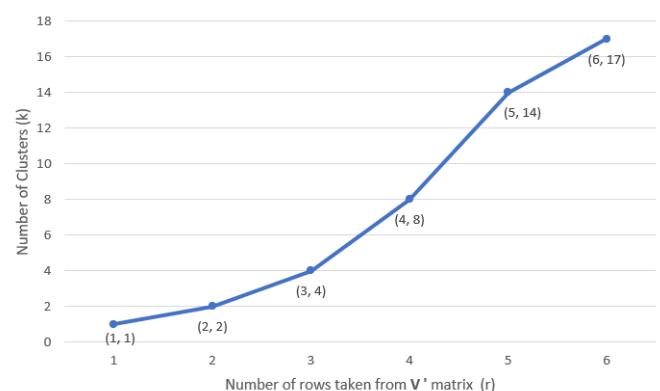


Fig. 5: Variation of Number of Clusters with Rank r

b. User-based approach

Hence, the second method - i.e. the user-based approach, where clustering of the users was carried out to obtain different smaller groups or clusters where the individual within each cluster would show similar characteristics and preferences in terms of the items chosen.

The first r columns of the left singular matrix, representing the user-latent space matrix, are considered and the rest are truncated, and using the *SVD signs method*^[8], the corresponding clusters are obtained. As inferred from **fig 3**, the value for k is taken as 4 and the following steps were carried out to simplify the whole clustering process.

1. The positive values were replaced by 1 and the negative numbers were replaced by 0.
 - (a) Thus, each row gave a binary representation corresponding to a unique number. The rows having the same binary forms would belong to the same cluster.
2. Once the clusters have amalgamated, the next step was to design a kit for a particular cluster and that can be done by counting the frequency of occurrence of each item in a particular cluster.
3. The top 10 items will decide the items assigned to that particular kit.
 - (a) Here, using the dataset, 8 clusters were obtained by taking the r -value as 4, and if $r=2$ were taken, then 16 clusters would've been obtained, and since that is highly uneconomical, we discarded the latter case and stuck to $r=4$ and the subsequent 8 clusters thus obtained.

Value of r	Number of clusters obtained (k)
2	2
3	4
4	8
5	16
6	32

TABLE 2

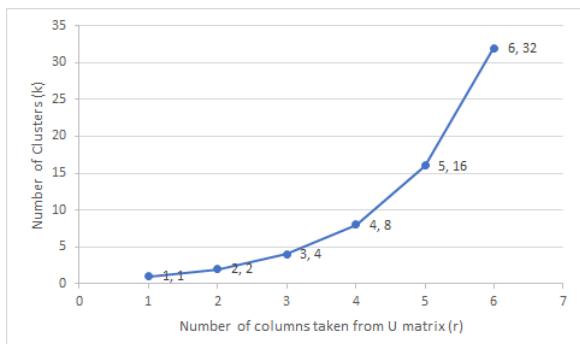


Fig. 6: Variation of Number of Clusters with Rank r

4. Once the clusters and the kits have been obtained, the next step is to optimize the results obtained and a variety

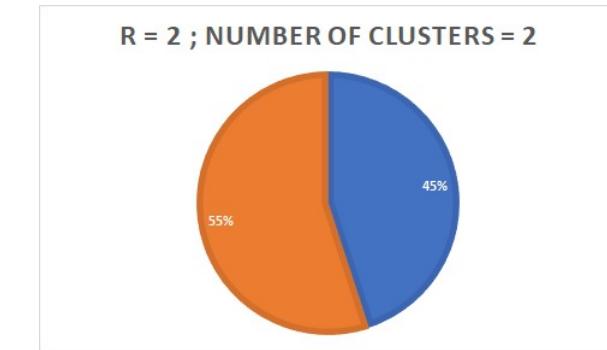


Fig. 7

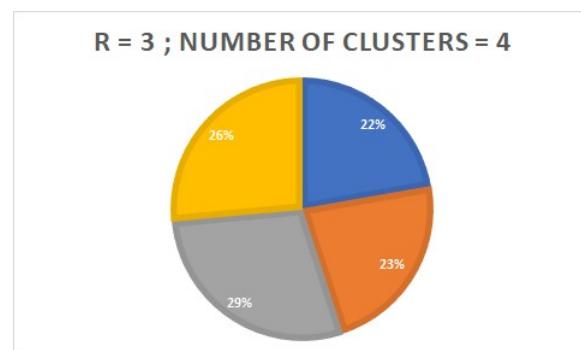


Fig. 8

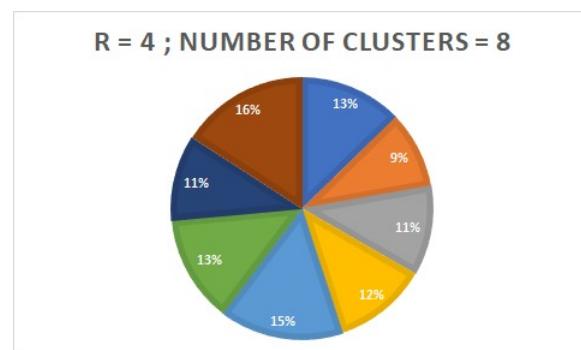


Fig. 9

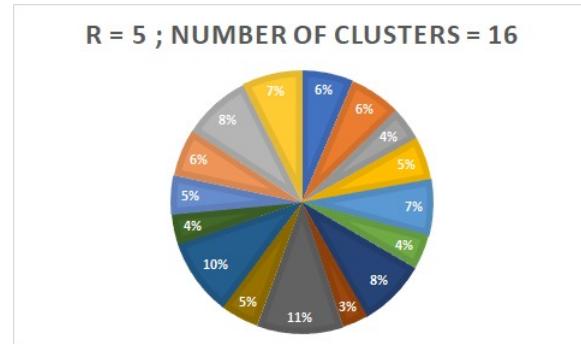


Fig. 10

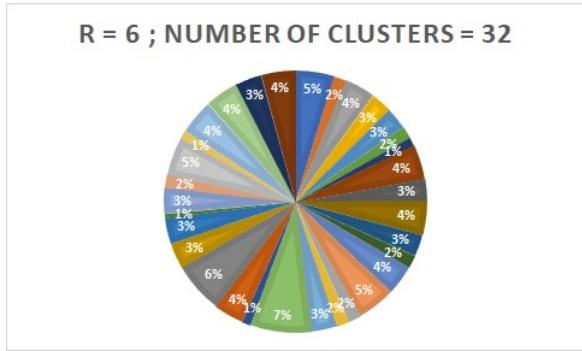


Fig. 11

of metrics were employed to ensure that the solution obtained best suited the preferences of the individuals.

- (a) For the aforementioned reason, the items that each user preferred as marked in the initial survey analysis was compared with the corresponding elements of each of the resultant 8 kits obtained, and for each kit, if these elements did not match, then a loss variable was defined to keep track of the number of mismatches and the kit having the lowest level of dissimilarity or error would be the kit to which the user would be assigned to. This is the main crux of the reassigning that is done to optimize the solution.
- 5. As a final means of validation, the overall cluster loss of each cluster was calculated through two different means(normal and exponential average) to adjudge the efficacy of the reassignment process. The cluster loss of each individual before the reassignment process, as found out earlier, is again computed, but this time, the values obtained are averaged in order to compute the overall cluster loss of each cluster.(fig 12 and fig 13)
- 6. Similarly, the overall cluster loss after reassignment is also computed,fig 13 based on the exponential average amplifies the cluster losses of every cluster, and it can be inferred how the losses have decreased except for the second one. Thus by accepting a slight trade-off for the second cluster, the overall objective can be easily attained by carrying out the reassignment.

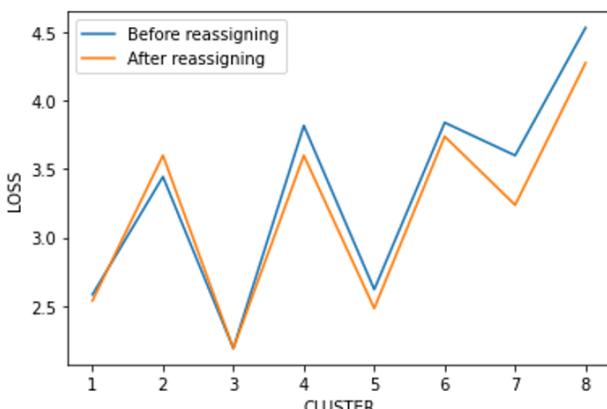


Fig. 12: Normal average loss of cluster loss on reassignment

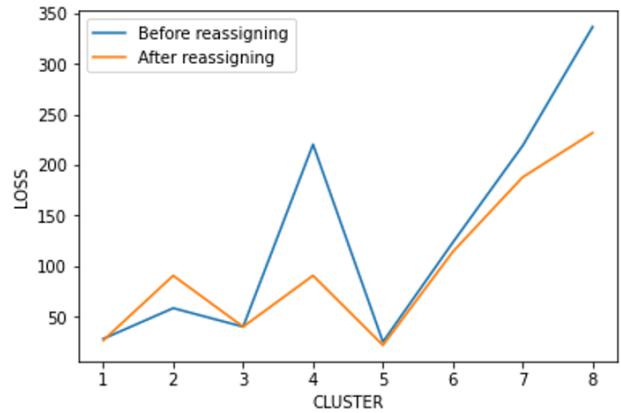


Fig. 13: Exponential Averaging to magnify cluster loss on reassignment

VII. CONCLUSION

The food security of a nation as large as India can be ensured only if appropriate safety measures and precautionary steps are taken right from the grassroots levels itself. Percolating down to the state level, government bodies have been eager to implement a public distribution system(PDS) early on in India. However, during the pandemic, the infrastructure of this system was not sufficient to satisfy the needs of the commoners living below the poverty threshold. With the loss of livelihoods and associated issues, malnutrition was on the rise along with heavy strain imposed on the existing healthcare system due to the infected patients. Perceiving this threat, the government of Kerala introduced an essential grocery food kit that could be collected from nearby ration shops. However, the contents of those food kits were a random assortment of edible items selected by the government without taking into consideration the preferences of its consumers. Hence, this paper sought to carry out a comprehensive analysis of various clustering algorithms to group different consumers into clusters and allocate a kit design based on their personal preferences. Among the different clustering algorithms, SVD-based clustering showed the most promising result, owing to the ease of dimensionality-reduction associated with it. This framework was implemented in a sampled consumer group, and future scope could extend to clustering based on a larger feature space taking into consideration factors such as geographical location, crop distribution, etc. Through this, one can infer that it is possible to develop such a personalized food distribution system, satisfying the consumers at the same time without leading to any food wastage.

VIII. ACKNOWLEDGEMENT

This work was initially implemented at REBOOT Kerala Hackathon 2020, Finals organized by the Additional Skill Acquisition Program (ASAP), Kerala, and the Department of Higher Education, Government of Kerala. Later on, a detailed dataset of consumer preferences was collected by surveying a sampled population in Thiruvananthapuram, Kerala, India. which helped improve the findings. The authors are indebted to Dr. Ajeeesh Ramanajun from the Computer Science Department of the College of Engineering Trivandrum and Prof. Tony S, from the Department of Mathematics, College

of Engineering Trivandrum for their invaluable suggestions and guidance.

REFERENCES

- [1] Anjana Verma, Dhriti Chugh, Ashish Patyal, Jitendra Kumar Meena, and Medha Mathur. Socio-epidemiological study of malnutrition and associated risk factors among under five children in rural rajasthan. *International Journal of Research in Medical Sciences*, 9(3):774, 2021.
- [2] Rashmi Umesh Arora. Food subsidies for the poor in india: are they working? *Journal of Asian Public Policy*, 6(3):217–230, 2013.
- [3] Silvia Masiero. Redesigning the indian food security system through e-governance: The case of kerala. *World Development*, 67:126–137, 2015.
- [4] Udayan Rathore and Upasak Das. Explaining the superior education outcomes of kerala: The role of state activism and historical endowment. *Oxford Development Studies*, 47(2):205–221, 2019.
- [5] KM Sulaiman, T Muhammad, Rishad Muhammad AP, and K Afsal. Trace, quarantine, test, isolate and treat: a kerala model of covid-19 response. *medRxiv*, 2020.
- [6] Business Line. Kerala government to supply food kits to 87 lakh families. <http://bit.ly/3ih8kgM>, 2020. Accessed: 2020-03-29.
- [7] Deepa Ravi and Ambili S Nair. The impact of national food security act-2013 on the public distribution system: A case study of kerala state. *International Journal of Engineering Research & Technology (IJERT)*, 8(06), 2019.
- [8] Emmeline P Douglas. *Clustering datasets with singular value decomposition*. PhD thesis, College of Charleston, 2008.