Clustering in food waste analysis: case study at student cafeteria

Billy Viyanto, Laurence, Agustina Christiani
Department of Industrial Engineering, University of Pelita Harapan, Jl. M.H. Thamrin Boulevard, Tangerang, 15811 Banten, Indonesia
Corresponding author: laurence.fti@uph.edu

Abstract. Food waste is one of the most recent global issues that is getting more attention nowadays. Food waste is all food that is not consumed or thrown away even though it is fit for human consumption. This paper presents a method of direct weighing of edible food that is not consumed in the student canteen. Based on preliminary observations, food is classified into carbohydrates, vegetables, meat and others. The results of weighing data show more accurate results compared to previous study. With a total of 226 data, the research was focused on the carbohydrate and vegetable variables. The computer algorithm using the elbow method shows the unsupervised K-Means clustering of 4 clusters. The quality of the clustering method is good, with an average silhouette coefficient of 0.6. This research can be used as a basis for further studies on food waste and life cycle assessment.

1. Introduction
Food waste is one of the most recent global issues that is getting more attention nowadays. In 2017 Food and Agriculture Organization of the United Nations (FAO) estimated that 30% of food globally that are produced for human consumption turned into food loss or food waste, which cost more than 1 trillion dollars for economic, environmental and social aspects. The significant number should become a deep concern for everyone, knowing that based on a report published by FAO in 2019, there were still 1 of 9 people do not have enough food to eat. It is a global issue, and many countries have a serious commitment to participate in the Sustainable Development Goals (SDG) program, with a vision to reduce half food waste of the total number.

Food waste is a problem in many countries, including Indonesia. Studies on discarded food also predict a significant increase that may occur in 2050 [1]. According to Food Sustainability Index (2018), using food loss and waste index, Indonesia was ranked 51 out of 67 countries that are being measured. This condition showed that there is still homework for every stakeholder to have a responsible consumption. Food waste means money spent on nothing and the additional resources required to handle the effect of post-consumption. Furthermore, it affects the environment by being a significant indirect contributor to global warming.

Despite its impact, food waste is still not a significant issue for people's everyday life. The previous study on quantifying weight of food waste is a questionnaire-based survey for undergraduate students [2]. The research relied on the memory of the respondent to remember some teaspoon of food waste on their consumption, and not everyone will pay attention to the weight, especially on detail such as food category. The indirect measurement will lead to inaccuracy for further analysis. Therefore this study seeks to answer questions about the weight and type of food discarded. In addition, questions that may arise are, what kind of data variations occur and how to group the data.
2. Materials and Method

2.1. Edible and Inedible Part of Food
Food refers to all substances, whether they have been processed, are being treated or are raw, that is intended for human consumption. Food waste is a reduction in the quantity or quality of food as a result of actions or decisions made by retail, food service providers and consumers [3]. Edible parts refer to components of food in a food supply chain that is intended for human consumption. Examples of edible slices of mango and papaya are mesocarp, the rich-nutrient middle layer between epicarp and endocarp. Inedible parts refer to components in meat products that are not designed for human consumptions, such as bones, skin, and feather of animal products [4].

However, waste products of inedible animal products are primary ingredients for cosmetics and pharmaceutical industry. Food waste is a result of the natural process, such as vegetables and fruits [5], and it is common for a business to lengthen period and minimise waste of product for financial purposes.

![Figure 1](image-url)  
**Figure 1.** Steps in Food Loss Waste Protocol(FLW)

The first five steps in figure 1 starting from goal definition, accounting review and principle reporting, establishing scope, deciding how to quantify FLW, data gathering and analysis, and inventory results calculation are the scope of the research [6]. The part of the food chain for this observation is student canteen, which in 2019 initial investigation shows that the most frequent waste type are rice or noodle, vegetables, meat and others. Various menus exist and diverse among food operators. Material type collection is a mere edible part of the foodstuff, but the liquid portion is excluded.

The process of weighing food scraps also does not take into account the weight of the spices. The reason is various spices already dissolved as an integral part of the food mixture. For example, beef that are not consumed is considered to be the weight of the meat itself. It is challenging to accurately predict the detailed composition of each seasoning because usually spices have transformed and dissolved with other ingredients. The only thing a consumer can do is make a rough prediction type of the ingredients in the food, such as the sour aroma might come from the taste of kaffir lime leaves in soup.

Among ten different methods for measuring or estimating food loss and waste, it is settled to use direct weighing method to get a more accurate result. The inedible part of food that contains plates or food containers used as research samples is not taken into account in the weighing process. If there is food that has edible and inedible parts, the parts, as mentioned above, must be separated first. The inedible part that appeared in the study was the bone. The first step in the weighing process is to take the food container containing the food waste to the weighing station.

The selected weighing station is one of the tables in the canteen, which is located closest to the place to put used dishes. The tool used for weighing is a digital scale that can show 2-digit results, paper form and pen. The food waste in the container is then sorted according to a category, weighed and recorded in the blank form. Food loss is divided into four major categories, namely carbohydrates, meat, vegetables and others. Foods that are classified as carbs are rice-based foods such as rice, noodles, rice vermicelli, and kway-teow (flat rice noodles). The researcher then wrote the recording results into a form that had been discussed and prepared before heading to the canteen location.

2.2. Clustering Method
The necessary amount of data meets $2^k$, but is expected to exceed five times that value. With four variables, it means that at least 80 data must be collected. The total number of data collected was 300, well above the minimum threshold required [7]. Data obtained on weighing is the result of
measurement in the form of variable data. With a data model like this, for further data processing is to use clustering, an unsupervised method that divides the data into specific groups. The primary purpose of clustering is to assist in the analysis and grouping of data that have similar characteristic. k-means clustering is an unsupervised method, works by dividing the results of observations with the partition system until the desired number of group forms data groups.

K-means is one of the most comfortable unsupervised clustering algorithms to implement in grouping data by calculating the distance of each observation. The span between two points are computed using the Euclidean Distance formula. First, the algorithm put random centroid and estimate the distance from the centroid to each point. Next, the centroid move to average cluster value, the gap is recalculated, and the iterations are repeated until the distance is stable. To determine the number of clusters is by finding the turning point of the curve in the elbow method with the x-axis is the number of groups and y-axis is the Within-Cluster-Sum-of-Square (WCSS) [8].

Cluster quality is determined by the value of Silhouette measure of cohesion and separation [9]. The coefficient range is between -1 and 1 and the bigger, the better. If the coefficient value is close to 1, it means that the average distance between objects in the cluster and other clusters is extensive and this number what the researcher expected.

3. Results and Analysis

Data processing uses the python programming language. In the initial step of data cleaning, 73 data could not be used due to incomplete data. And after checking the second time using the program, it turns out that there is still one incomplete data.

| Table 1. Post-Cleansing |
|-------------------------|
| Data | Column | Non-null counts | Data type |
|------|--------|-----------------|-----------|
| 0    | Carbohydrate | 226 | float64 |
| 1    | Vegetable | 226 | float64 |
| 2    | Meat | 226 | float64 |
| 3    | Other | 226 | float64 |
| 4    | Valid N | 226 | float64 |

Table 1 describes the data after the cleaning process, where the information has a matrix size of 226 x 4. Gender is a nominal data type, and python recognises it as an object. This type of data cannot be processed because it must be transformed into numbers. The code given for men is one and women are 2, where the two numbers are equivalent levels. The result of FW measurement per day of observation is illustrated in table 2.

All data is valid with minimum weight for all categories is 0 gram, which indicates that there are people who eat all the food on the plate. The value of the standard deviation is higher than the average, and this value shows a significant difference in food waste for each person. It can be inferred that the data is not normally distributed.

| Table 2. Descriptive Statistics (g) |
|-------------------------------|
| N   | Min | Max  | Sum      | Mean | Std. Deviation |
|-----|-----|------|----------|------|----------------|
| Carbohydrate | 226 | 0    | 159.8    | 4,867.4 | 21.54 | 35.35 |
| Vegetable | 226 | 0    | 146.2    | 3,201 | 14.11 | 24.56 |
| Meat | 226 | 0    | 143.9    | 2,284.2 | 10.08 | 19.53 |
| Other | 226 | 0    | 122.4    | 2,927.4 | 12.95 | 22.30 |
| Valid N | 226 | 0    |          |       |      |          |

From table 2, it can be seen that the FW in order from the most to the least are rice and noodle category, other, vegetable, and lastly meat. From table 2, food waste with the highest percentage was
rice and noodle (36.65%), followed by vegetables (24.10%), other (22.04%), and finally meat (17.20%). Figure 2 shows the histogram of vegetable weight does not follow the normal distribution, and the data distribution plot shows that all data is skewed to the right.

With a non-normal distribution in figure 2 shows that people do not eat up the vegetables, but leave a small amount. The maximum recorded weight of uneaten vegetables is 146.2 grams, while the load of rice that is not consumed is 159.8 grams. Because the data are not normally distributed, to standardize the data, the standard scaler method cannot be used, but the minimum-maximum scaler.

**Figure 2.** Histogram of Vegetable Waste

The heatmap in figure 3 shows the strength of the relationship between variables, with dark green color showing the perfect connection between the variables and the variables themselves, forming a diagonal formation. Positive value falls between 0.13 and 0.27, which belongs to the category of weak correlation [10]. Dark red color represents negative values, such as vegetable and other with an amount of -0.11. The correlation coefficient indicates that the maximum value 0.27 between carbohydrate to vegetable.

This study tries to explore more deeply about variable carbohydrate and vegetables. Before further processing, the data need to be transformed using the minimum-maximum scaler. The min-max tool makes data adjustments within a specific range, commonly used is 0 to 1. Figure 4 shows the data after standardization. It can be seen that both the x and y axes have a minimum value of 0 to 1. In general, there is no significant difference, only for the points on the left, which are denser from one point to another.

For the initial estimation, it can be seen visually that the data is in groups of more than 1. Minimal data are forming vertical dots, horizontal information on the x-axis and data that is clustered in the upper right area of the graph. Thus, the value of k is temporarily estimated to be three. However, if only rely on visuals, the clustering output will not be accurate. Therefore, the algorithm to find a turning point is used through WCSS calculations and the location of the elbow. The result is obtained at point four. The search for the number of clusters is initiated with a randomly placed center point. The center point will later move along with repeated iterations. For this research, it was carried out 100 times. Although there is no rule about the minimum limit in this iteration, the rule of thumb is that the more iterations, the better. However, a large number of iterations will consume even more computer resources. The small circle on the left in figure 5 shows the first turning point with the cluster value on the x-axis of 2. The next and also the last turning point is shown in cluster 4, with the WCSS value on the y-axis being lower than the first circle. The iteration results are shown in table 3, between iteration 0 and iteration 99 for centroid and inertia. If the program is rebooted, the iteration values will be different; this change is because the starting center point is not the same every time the program running.

However, the inertia value will be similar in magnitude to around 3.5. The position for the centroid of each cluster also consistently remains the same after 100 iterations. The difference between initial
and final inertia is – 2.66. Among four centroids, only one center point has not changed from the beginning of the iteration to the end of the iteration. The other three points experience a shift towards the x-axis and the y-axis.

Figure 4. Initial Estimation

![Initial Estimation](image)

Figure 5. Elbow Plot

![Elbow Plot](image)

**Table 3.** Comparison of Initial and Final Iteration

|        | Initial | Final |
|--------|---------|-------|
| Inertia| 6.16    | 3.50  |
| Centroid | (0.621, 0.097) | (0.444, 0.057) |
|         | (0.008, 0.096) | (0.041, 0.029) |
|         | (0.237, 0.033) | (0.011, 0.336) |
|         | (0.895, 0.795) | (0.895, 0.795) |

Compared to the initial estimates of grouping the data, figure 6 shows a visual difference in the location of the centre points and the distribution of the data in table 4. It can be seen that there is an unequal distribution of the number of members for each cluster. The most populous group occupies the lower-left area, and the group that has the least number of members occupies the group in the upper right. This area indicates that not many people waste enormous amounts of food on rice and vegetables based on the data in this study. The group that has the most number of members is cluster 1, with a total of 144 data, followed by set 0 with 40 data. The accumulated percentage of the two groups mentioned above is 81.4%.

**Table 4.** Distribution of Data in Each Cluster

| Cluster | N  | %   | C.N | C. % |
|---------|----|-----|-----|------|
| 0       | 40 | 17.7| 40  | 17.7 |
| 1       | 144| 63.7| 184 | 81.4 |
| 2       | 33 | 14.6| 217 | 96.0 |
| 3       | 9  | 4.0 | 226 | 100.0|

It is necessary to measure quality clustering by comparing the number with a standard. Average silhouette calculations are calculated where the greater the number, the better the model quality. With 4 clusters, the average silhouette value is 0.6. For comparison, with the number of sets 3, the average cost is smaller, namely 0.58. For better quality separation, the benefits should reach certain value, with
a number greater than 0.5 [9]. Another recommendation which is lower, that recommended minimum value of 0.2, which is in the temperate zone [11].

4. Conclusion
Food waste case study in student cafeteria has been going well, and the research has obtained total food waste data from measurement, which is 13,280.40 gr. Weight of carbohydrate category is 4.86 kg, vegetable 3.2 kg, meat 2.28 kg, and the other type is 2.92 kg. This study uses tools commonly used in clustering, such as the elbow method to determine the number of clusters and the silhouette coefficient to assess the quality of the clustering method. For data grouping using the k-means clustering algorithm approach, using only variable carbohydrate and vegetable, the number of clusters selected is 4 clusters for 226 data.

The four groups can be called vegetable dominant, carbohydrate dominant, low carbo-veg, and high carbo-veg. The results of this study can be used by food business owners to consider what food will be sold and the portion of food served. If food portions can be adjusted better, then food waste and food prices can be reduced. The results of the study can be utilized by the canteen management to see the pattern of food waste generated to prepare ways to handle food waste. For further research, it can be carried out over a more extended period and in a larger amount of data. Besides, questionnaires can also be distributed to find out the reason not to eat all the cuisine. Furthermore, if other data such as the final destination of waste disposal can be traced, a more comprehensive study on life cycle assessment can be carried out.

Reference
[1] Lopez Barrera E and Hertel T 2020 Food Policy 101874
[2] Mandasari P 2018 Quantifying and analysing food waste generated by Indonesian undergraduate studentsE&ES131(1)012058
[3] Gustavsson J, Cederberg C, Sonesson U, Van Otterdijk R and Meybeck A 2011 Global food losses and food waste
[4] Alao BO, Falowo AB, Chulayo A and Muchenje V 2017 The potential of animal by-products in food systems: Production, prospects and challengesSustainability9(7)1089
[5] Mena C, Adenso-Diaz B and Yurt O 2011 The causes of food waste in the supplier–retailer interface: Evidences from the UK and Spain Resources, Conservation and Recycling55(6) 648-58
[6] Hanson C, Lipinski B, RobertsonK, Dias D, Gavilan I, Gréverath P, Ritter S, Fonseca J, VanOtterdijk R, Timmermans T and Lomax J 2016 Food loss and waste accounting and reporting standard
[7] Dolnicar S 2002 Proc. Australian and New Zealand Marketing Academy Conference
[8] Han J, Kamber M and Pei J 2012 Classification: Basic Concepts, Chapter 8, Data Mining: Concepts and Techniques

[9] Sarstedt M and Mooi E 2019 Cluster Analysis 301-54

[10] Schober P, Boer C and Schwarte LA 2018 Anesth Analg 2018 126 1763-68

[11] Tkaczynski A 2017 Segmentation using two-step cluster analysis In Segmentation in social marketing 109-25 Springer Singapore