**BA815 Clustering Homework**

**Please find the link to our colab notebook below:**

<https://colab.research.google.com/drive/1Bog-yRaPTkZR4s-OBqTSj3CGLa7dKHvl?usp=sharing>

**Team 32:**

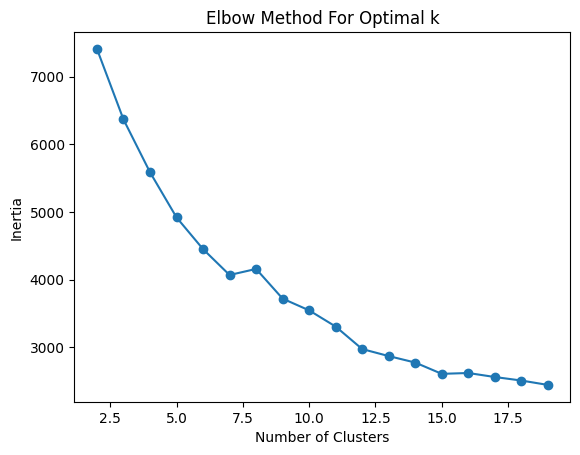
Sneha Ekka, Jasmine Gohil, Gunjan Sharma, Jenil Shah

**Question 1**

**Using the sample data, suggest the optimal number of clusters using an elbow plot.**

#### **Answer 1**

Elbow plot (as obtained using the code in the notebook):



#### We suggest k=15 as the optimal number for this data. At k=15, there is a noticeable bend or "elbow" in the plot, indicating a significant reduction in the within-cluster sum of squares (WCSS). Choosing any k greater than 15 doesn't yield more significant improvements in the model; instead, it prevents it from becoming overly complex or fitting the data too closely.

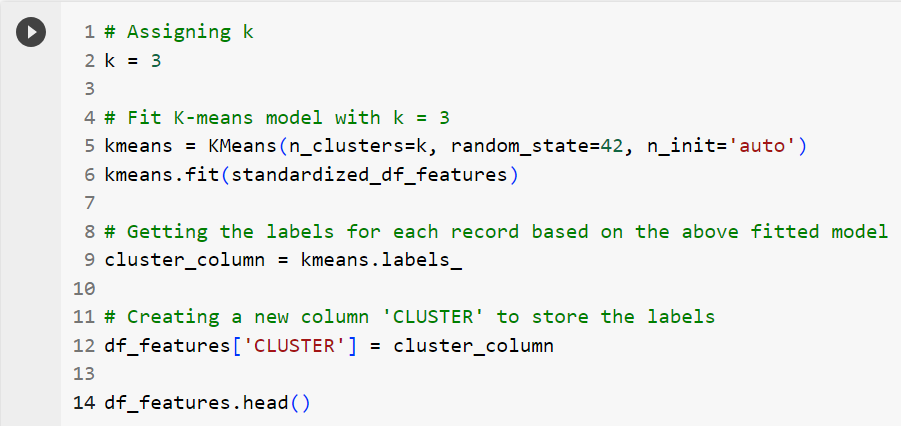
#### In a business case, narrowing down clusters to 3 or 4 can boost efficiency by grouping together similar clusters. This strategic reduction in clusters aims to simplify the model, making it more practical while maintaining its effectiveness. This approach ensures a focused and efficient analysis of the data.

### Question 2

### **Run a k-means clustering algorithm with k = 3 with this sample data.**

**Answer 2**

Below is a snapshot of the code used to fit the above-specified model on the sample data:



### Question 3

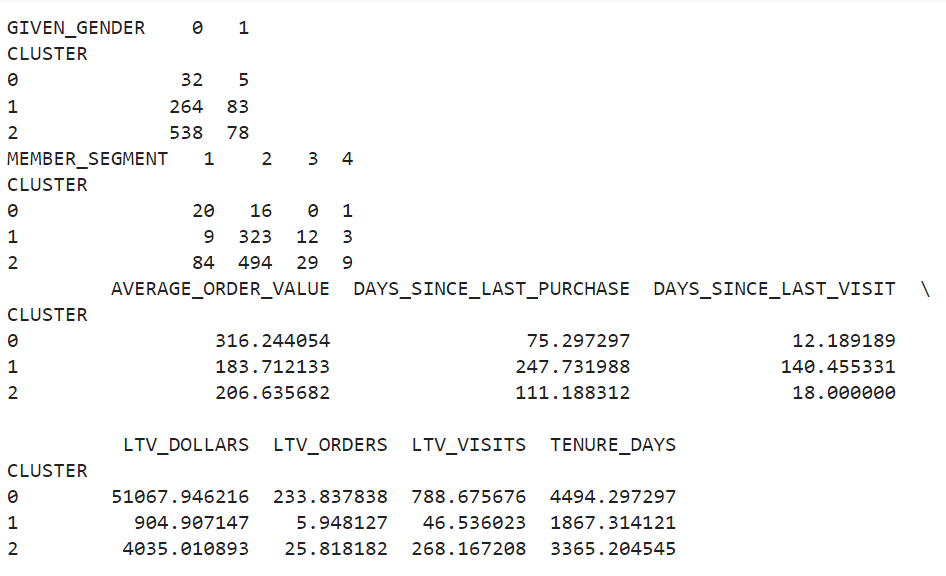
### Describe each segment’s characteristics, i.e., proportion of male/female customers, income buckets, average LTV dollars, Average Order Value, etc. How similar and dissimilar are these clusters?

**Answer 3**

* **Cluster 0**: Cluster 0 is characterized by its exclusive female clientele, comprising dedicated and high-value consumers. Despite a seemingly low average order value of USD 750, their significant lifetime dollar value and order count reveal their loyalty and substantial purchasing behavior. This insight is derived from the cluster's remarkable average tenure of approximately 4500 days which is the highest. Furthermore, their frequent engagement with the website, reflected in a lifetime visit range of 250-1800, emphasizes their active and consistent patronage.
* **Cluster 1**: Cluster 1 indicates a higher representation of females compared to males. The member segments predominantly consist of regular buyers, with few bargainers and big spenders. Most customers have visited within the last 350 days, and the time since their last purchase falls within the 50–350-day range. Customer lifetime visits are less than 750, with orders below 100. Additionally, both the average order value and lifetime dollar value are between 1500 and 20000, respectively. The tenure days for customers are generally less than 2000 days.
* **Cluster 2**: Cluster 2 indicates a predominantly female customer base, with around 500 buyers, followed by approximately 90 big spenders, and smaller groups of bargain shoppers and returners. Most customers made their last purchase within the past 350 days, and their last visit within the past 150 days. The lifetime value (LTV) in terms of visits falls between 0 and 1500, while LTV in orders is in the range of 0 to 150. The Average Order Value is below 1500, and the LTV Dollar Value is less than 30000. The average customer tenure is around 3300 days.

Looking at the plots and metrics for each cluster, we notice that Cluster 1 and Cluster 2 are very similar in the kind of customers they hold. Both clusters include a high proportion of B-BUYERs and fairly the same distribution of other member segments with the exception of high B-BIGSPENDERs in Cluster 2. Across the GENDER feature, all three clusters have a similar distribution of females and males, and no significant distinction can be made between the clusters. Besides the above-mentioned features, all three clusters are quite distinct from each other which can be seen in the pivot tables shown below. It is also noteworthy that Cluster 0 stands out as the most distinct one out of all three clusters and separates out the loyal customers with the highest TENURE and LTV values from the entire customer base.

Summary (pivot) tables of cluster-wise counts/averages of all the features:

****

### Question 4

### Would you recommend different promotional discount amounts for each segment or the same for all? Justify your reasoning.

#### ***Answer 4***

#### ***Business Strategy for each cluster***

* **Cluster 0:** To enhance customer engagement and sales, a tiered discount strategy is proposed. This involves offering substantial discounts of 20-25% to new or less active customers, encouraging initial engagement and exploration of the brand. Meanwhile, a smaller discount of 15-20% targets moderately engaged customers, nudging them towards becoming more regular buyers. This approach is tailored to the different levels of customer familiarity and engagement with the brand, aiming to boost sales and cultivate a loyal customer base.
* **Cluster 1:** This specific cluster comprises newly acquired customers, representing the segment with the least expenditure, minimal order frequency, and a relatively low number of lifetime visits. Given this insight, we propose a targeted marketing approach for these customers, offering a USD 20 discount. The rationale behind providing a moderate discount is rooted in the fact that these customers are in the early stages of their relationship with our brand, making their purchasing behavior somewhat unpredictable. However, by offering an enticing discount, we aim to cultivate their loyalty and encourage them to transition into becoming more regular and engaged buyers over time. Our strategy for this cluster is centered around attracting these new but relatively inactive customers with a USD 20 discount, with the goal of fostering increased engagement and long-term loyalty.
* **Cluster 2:** Given the distinct characteristics of this customer cluster, offering them the maximum discount of USD 30 is strategically sound. The time metrics, with days since the last purchase ranging from 0 to 350 days and days since the last visit below 150 days, indicate an active and recent engagement with the platform. This group of customers indicates a good response to promotions considering the range of LTV visits and LTV orders, along with an average order value below USD 1500 and LTV Dollar Value below USD 30,000. Consequently, offering this cluster the maximum discount of USD 30 aligns with their buying behaviors and potentially boosting overall sales.

### Question 5

### How would you test your campaign to ensure the segmentation helped?

**Answer 5**

To test if the clustering helps inform our campaign strategy and yields better results as opposed to not clustering the customer base, we can perform A/B Testing on them as follows:

1. **Dividing the Dataset:** We start with a dataset that likely contains information about customers or some relevant metrics for our marketing campaign. The first step is to divide this dataset into two parts:

* Control Dataset: This is a subset of our data that serves as a baseline or control group. It represents a group that is not exposed to the changes or interventions brought about by our marketing campaign. It's like the "before" scenario
* Testing Dataset: This is the other subset of our data that will be subjected to the changes or interventions introduced by our marketing campaign. It's like the "after" scenario. The goal here is to have a fair comparison between a group that experiences the campaign and a group that does not

1. **Equal Division of the Dataset:** This means that we are ensuring an equal number of observations or samples in both the control and testing datasets. This is important for a fair comparison, as an unequal division might introduce bias into the evaluation.
2. **K-Means/Clustering and Test Dataset:** This refers to the process of clustering, where we group similar data points in the test dataset together based on certain features or characteristics. The idea is to identify different clusters within our testing group that may react differently to the targeted marketing campaign.
3. **Campaign Testing:** With the testing dataset clustered, we can now use the insights gained from K-Means clustering to inform our marketing campaign strategy for the testing dataset. For example, you might tailor different aspects of the campaign for each identified segment. Meanwhile, the control dataset remains untouched by any specific marketing interventions.
4. **Evaluation:** After the campaign, we can evaluate its effectiveness by comparing the results of the testing dataset with the control dataset. This involves analyzing key performance indicators (KPIs) or metrics relevant to our marketing goals

In summary, this process involves creating a fair and controlled environment for testing a marketing campaign by dividing the dataset into control and testing groups, clustering the testing group using K-Means clustering, and then applying the insights gained to the testing dataset. This allows for a more targeted and informed approach to the campaign and facilitates the evaluation of its impact.

We meticulously analyze the results of our targeted marketing campaigns in comparison to the untouched control dataset. A significant divergence in outcomes serves as an indicator of the efficiency and effectiveness of our clusters. Conversely, if the results show minimal variance, it signals that our current clustering approach may be less effective. In such cases, a reevaluation of our features is necessary to enhance the clustering efficacy of our dataset.