**Summary**

**Data Preprocessing:**

I the collected data, we have so many rows of data and some of them are unknown. So, I started cleaning it by checking the no. of unknown values included in the data. I found almost negligible count in the almost all columns of data except in one column. That is, “User ID”. As we cannot assign any random number to User ID, I just remove those rows of unknown User ID. But, I am still trying to do some alternatives without just removing them.

After cleaning the data, I found some unusual deviation. And yes, this is because of some outliers present in the data. Using the z-score method I have removed some of the outliers so that, the insights will be focused to what we need.

While cleaning data, I have removed all the unknown values in the data. Also we have sufficient known data which may be enough to conclude. But still, the removed 1 percent of data may change the conclusions.

**Analyzing Data:**

**List of segments identified and User IDs belonging to that segment:**

According to the collected data, I am able to segment the customers into two types.

1. Geographical Segmentation.
2. Behavioral Segmentation.

**1. Geographical Segmentation:**

This type of segmentation purely depends upon the location of the transaction. As the one and only location provided in data is country. So, on the basis of country we can cluster customers of each country into a segment. Then, I get 37 segments as there are 37 unique countries. But, it does not look like well. As a country itself is more diversified and people in it too.

So, I started doing behavioral segmentation as the given data suggested their money spending for the products.

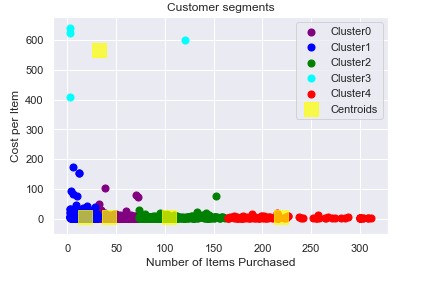
**2. Behavioral Segmentation:**

This type of segmentation depends on how many products they buy and cost of per unit product they buy.

I use KMeans clustering algorithm to segment the customers. Using the elbow method I found the efficient value of “k”. It suggests me that “k=5”.

So, there are 5 segments of customers

**Visualization:**

.

Using the above visualization, I have concluded the clusters of customers.

|  |  |
| --- | --- |
| **Cluster** | **Description of Customers** |
| *Cluster 0* | *Customers who buy less number of items which are less in cost* |
| *Cluster 1* | *Customers who buy very less numbers of items which moderate in cost* |
| *Cluster 2* | *Customers who buy moderate number of items which are very less in cost* |
| *Cluster 3* | *Customers who buy less numbers of items which are very high in cost* |
| *Cluster 4* | *Customers who buy large number of items which are very less in cost* |

|  |  |  |
| --- | --- | --- |
| **Cluster** | **Number of Items** | **Cost per Unit Item** |
| *Cluster 0* | *Less* | *Less* |
| *Cluster 1* | *Very Less* | *Moderate* |
| *Cluster 2* | *Moderate* | *Less* |
| *Cluster 3* | *Less* | *Very High* |
| *Cluster 4* | *Large* | *Very Less* |

* **P.V. Hem Charan**