Shopping Hard or Hardly Shopping: Revealing Consumer Segments Using Clickstream Data

Increasing technologies making human to perform all manual activities to virtual activities and one such activity is online shopping where user can visit online application and make desired shopping. Online shopping applications like UK ASOS or any other applications are utilizing their customers browsing or click stream data to understand their customer behaviour. Based on behaviour application owners can know which customers are more revenue able and to such customers they can advertise more apparels or any other product.

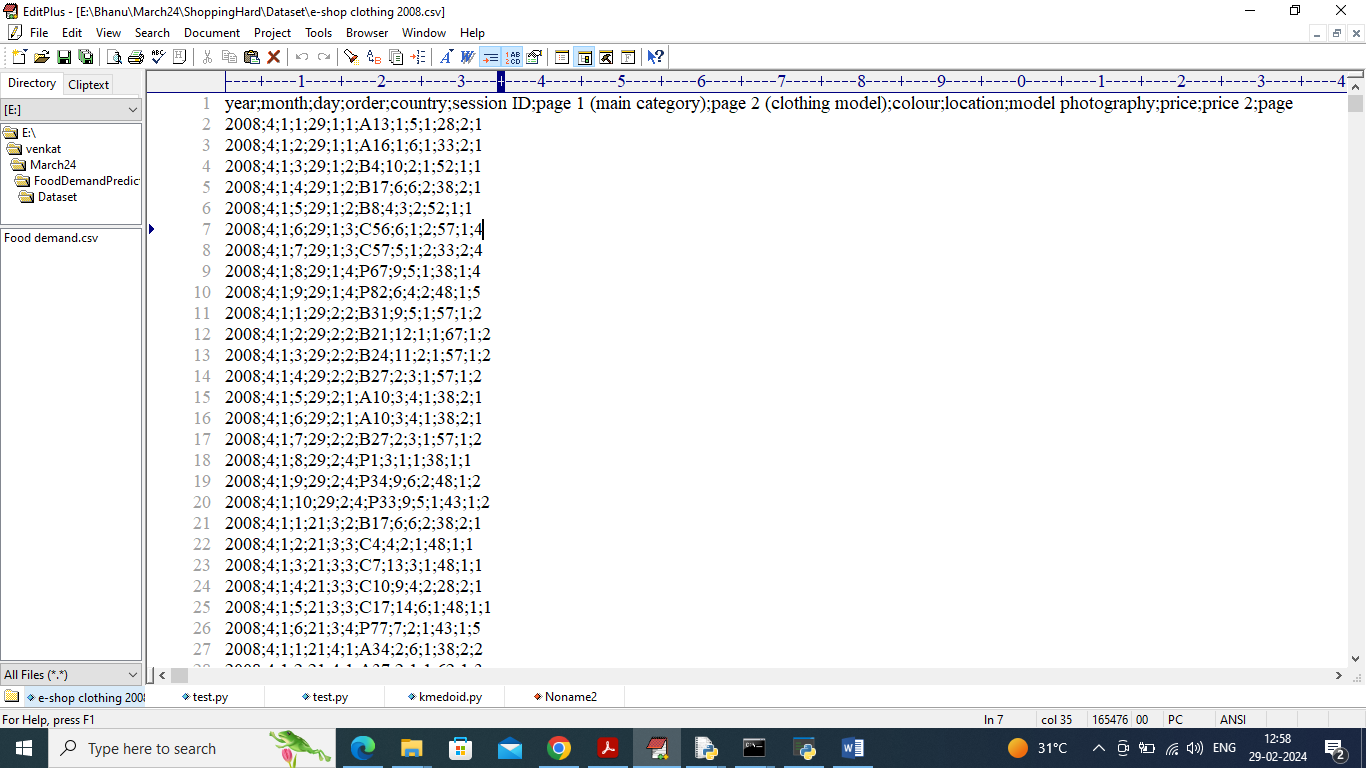
All existing applications were using machine learning algorithms to predict customer behaviour but all those algorithms unable to segment least and high profit customers. So author of this paper combining multiple clustering based techniques such Partition based K-MEDOID, Gower Distance Matrix and KRUSKAL WALLIS.

PAM based provides the greatest insight into the clusters in terms of explain ability due to the use of medoids.

Dataset often contains binary and categorical values and all distance metrics handle this data equally which will degrade clustering quality so author of this paper employing GOWER matrix which will calculate manhattan based distance for binary and categorical values. Gower distance uses a combination of distance metrics that satisfy each of the variables, namely range-normalized Manhattan distance for continuous variables and dice distance for nominal data, which can be calculated after turning each category into a binary variable.

KRUSKAL WALLIS will be used to calculate revenue from different clusters.

In propose paper author has collected data from UK real company and this dataset is not available on internet so we downloaded available click stream data from KAGGLE repository which is quite similar to this. In below screen showing dataset details



In above dataset screen first row contains dataset column names and remaining rows contains dataset values. In above dataset we will be using ‘order’ columns to calculate revenue. More orders mean more revenue, dataset can be anything but procedure of algorithms running will be same

Above dataset can be downloaded from below URL

<https://archive.ics.uci.edu/dataset/553/clickstream+data+for+online+shopping>

Extension Concept

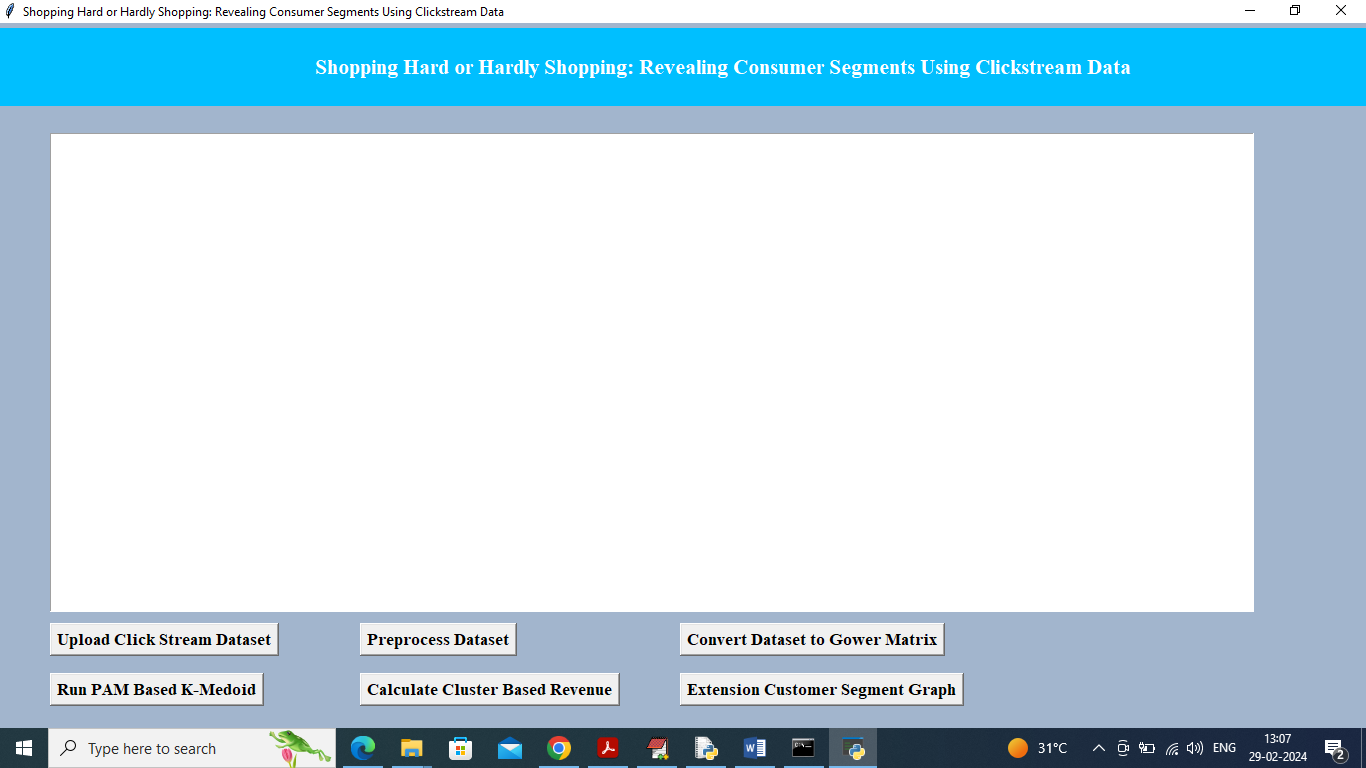
In propose paper author saying using clustering concept he is segmenting customers based on revenue but not showing any graph or output which clearly shows how customers from dataset are segmenting. So as extension we are plotting graph of customer segmentation based on orders in clear visualization graph.

To implement this project we have designed following modules

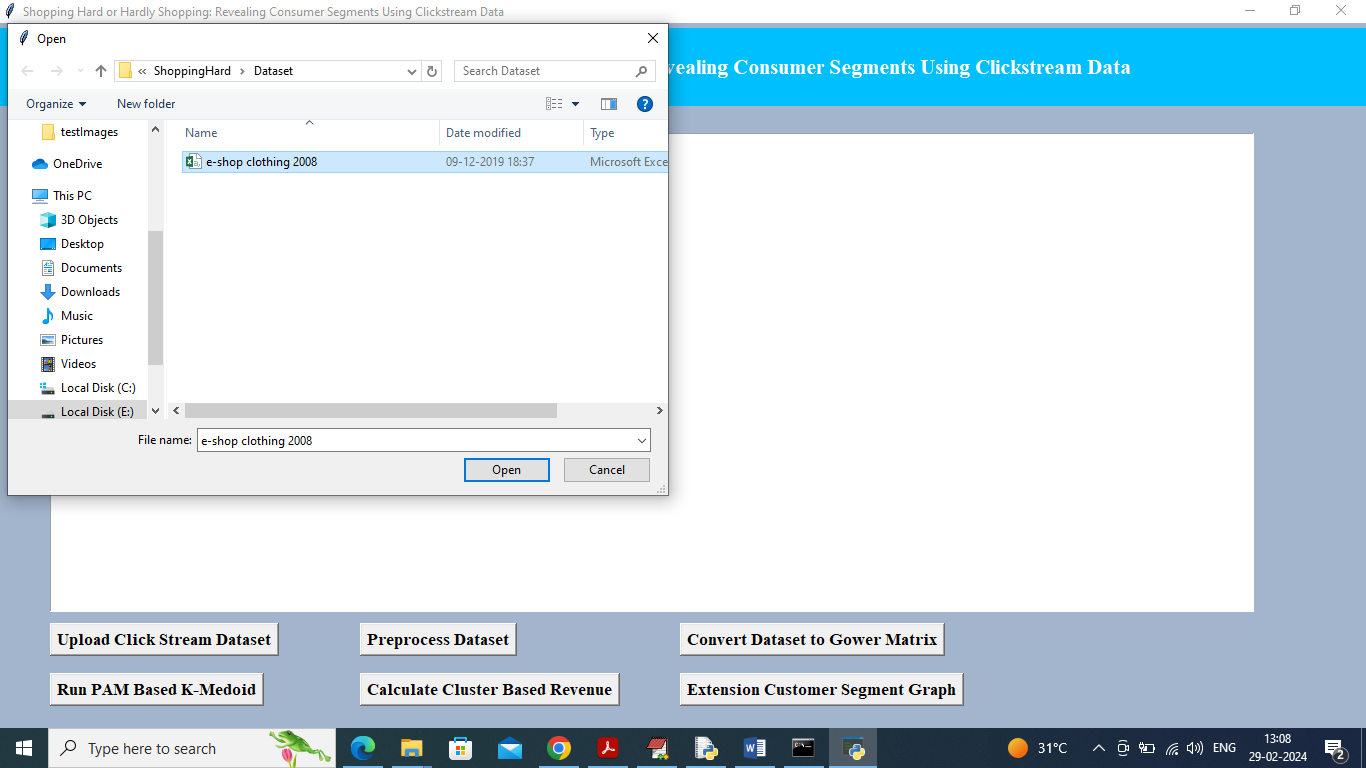
1. Upload Click Stream Dataset: using this module will upload and display dataset values and then perform dataset descriptive analysis like mean, median etc.
2. Pre-process Dataset: using this module will remove missing values and then clean and normalize dataset values
3. Convert Dataset to Gower Matrix: processed values will be input to Gower matrix to convert them into distance matrix
4. Run PAM Based K-Medoid: Gower values will be input to KMEDOID algorithm to cluster dataset and then calculate silhouette score
5. Calculate Cluster Based Revenue: this module will calculate revenue from each generated cluster
6. Extension Customer Segment Graph: this module will display all segmented customers in graph format

SCREEN SHOTS

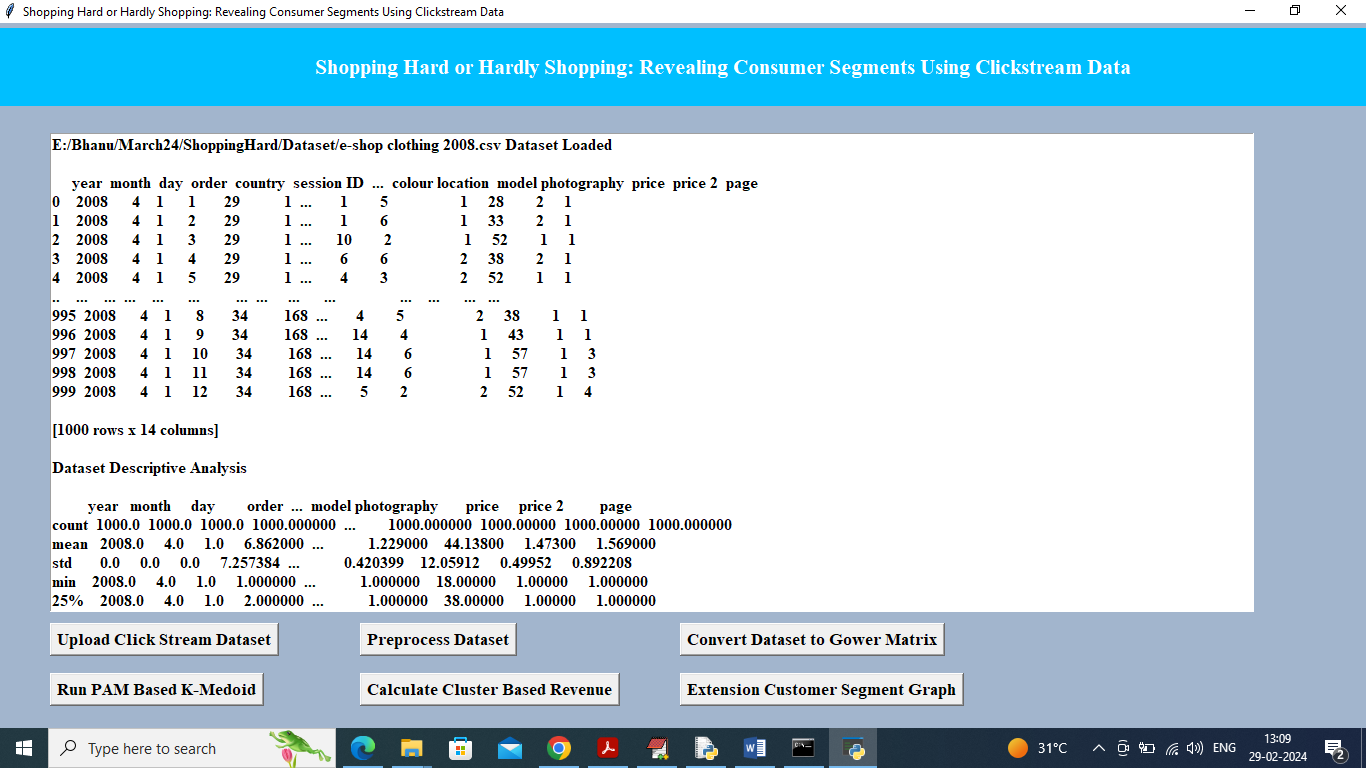
To run project double click on run.bat file to get below screen



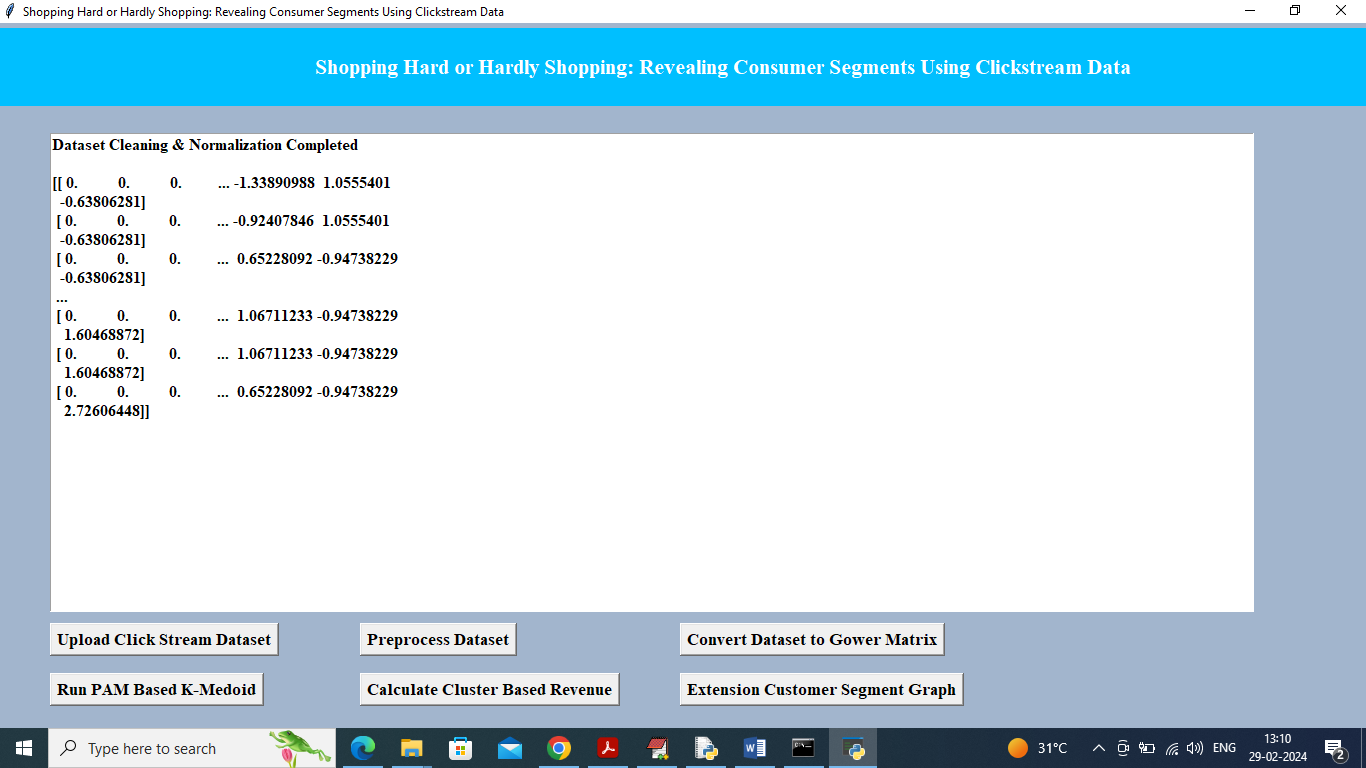
In above screen click on ‘Upload Click Stream Dataset’ button to upload dataset and get below output



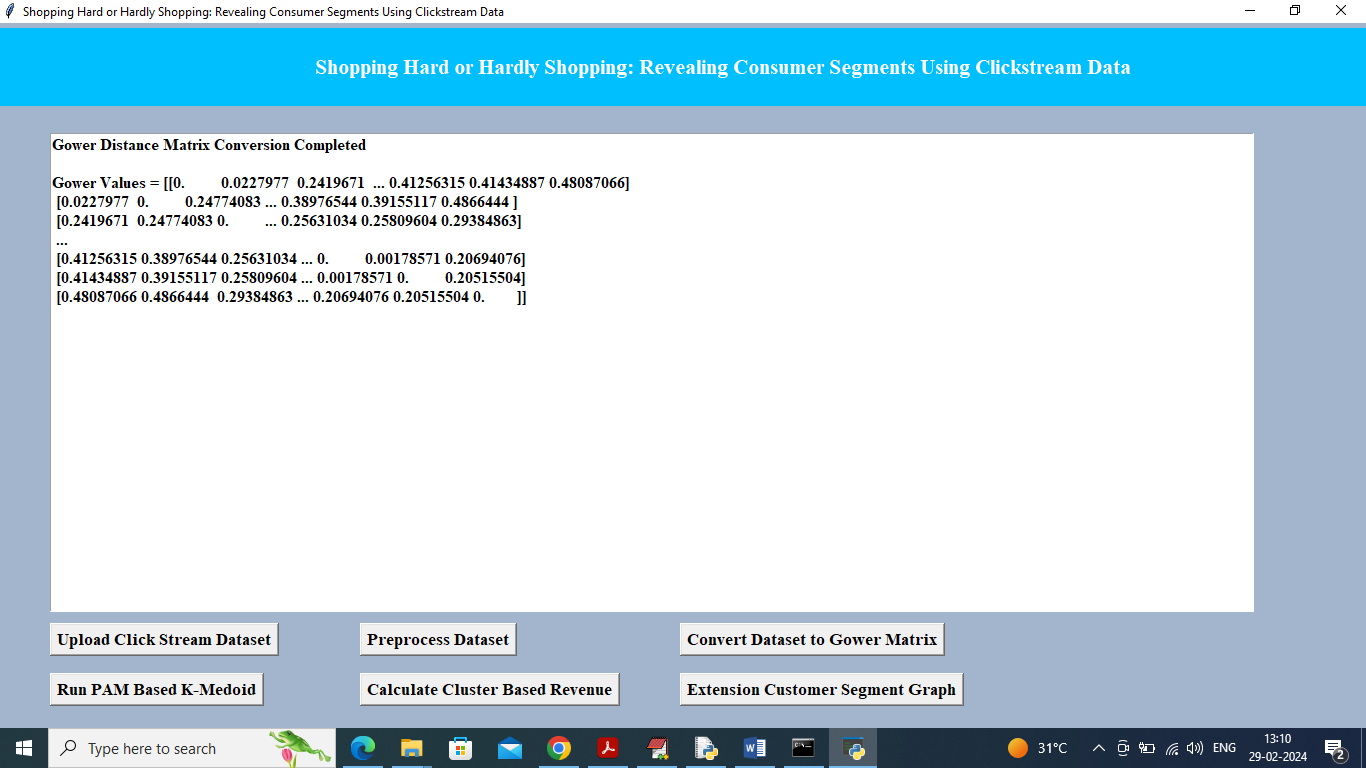
In above screen selecting and uploading ‘e-shop’ dataset and then click on ‘Open’ button to get below page



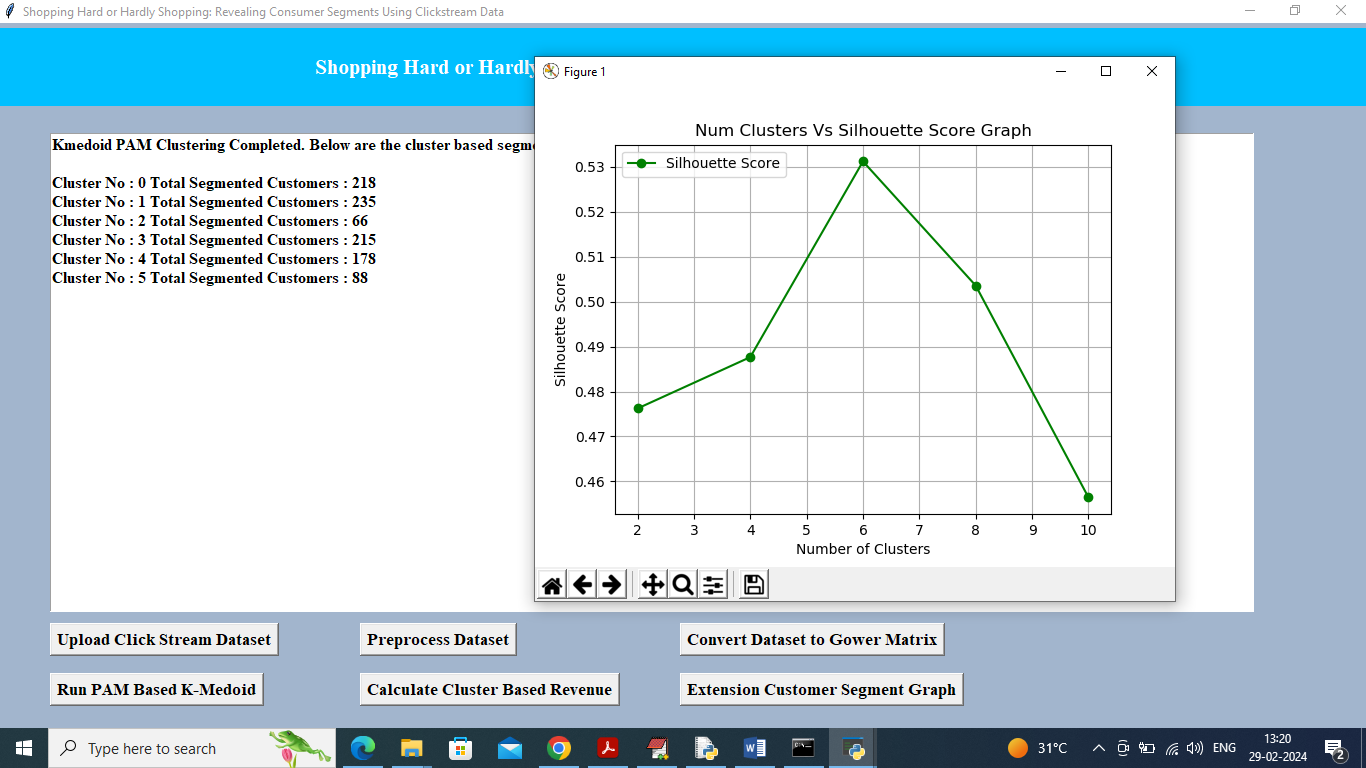
In above screen can see dataset values loaded and can see descriptive analysis like Count, Mean, STD and other descriptive values and now click on ‘Pre-processing’ button to clean dataset as above dataset contains both numeric and non-numeric values



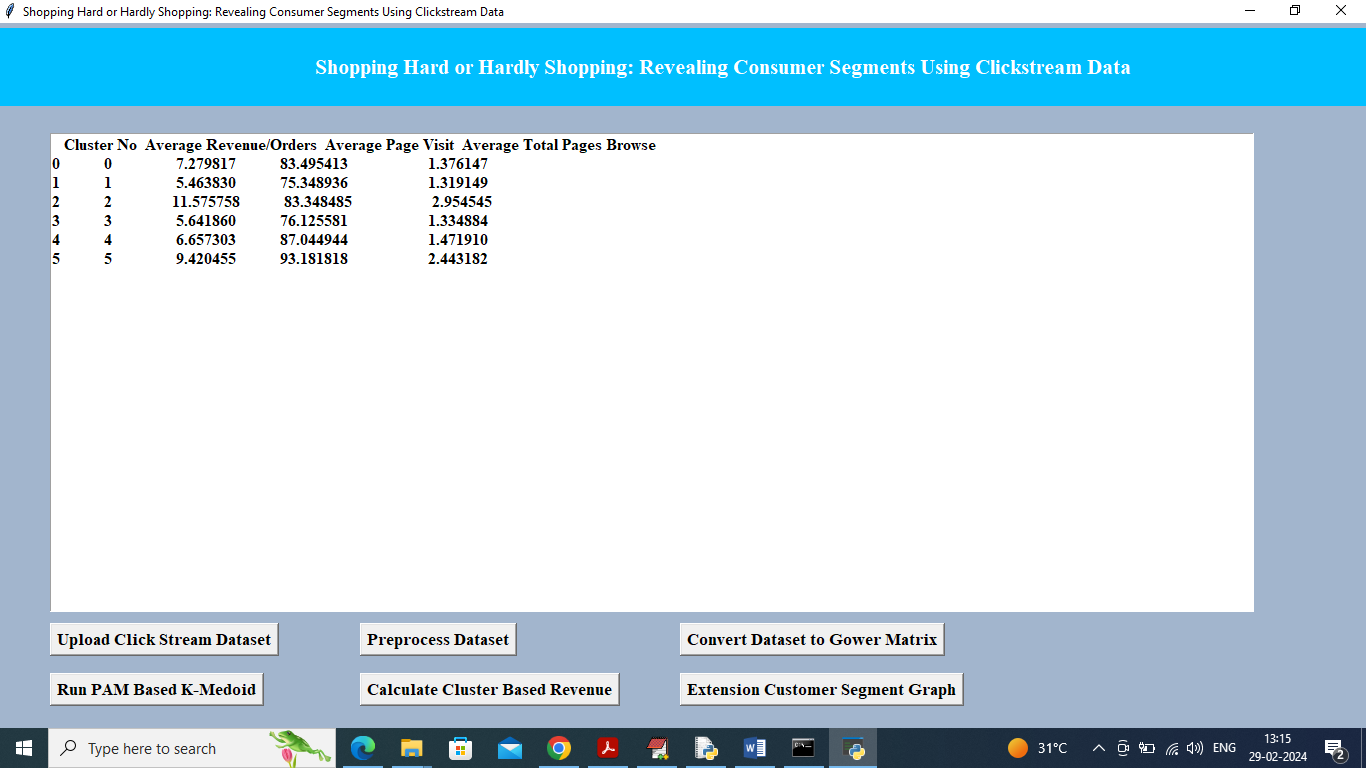
In above screen all data values converted to numeric format and now click on ‘Convert Dataset to Gower Matrix’ button to get below Gower distance values



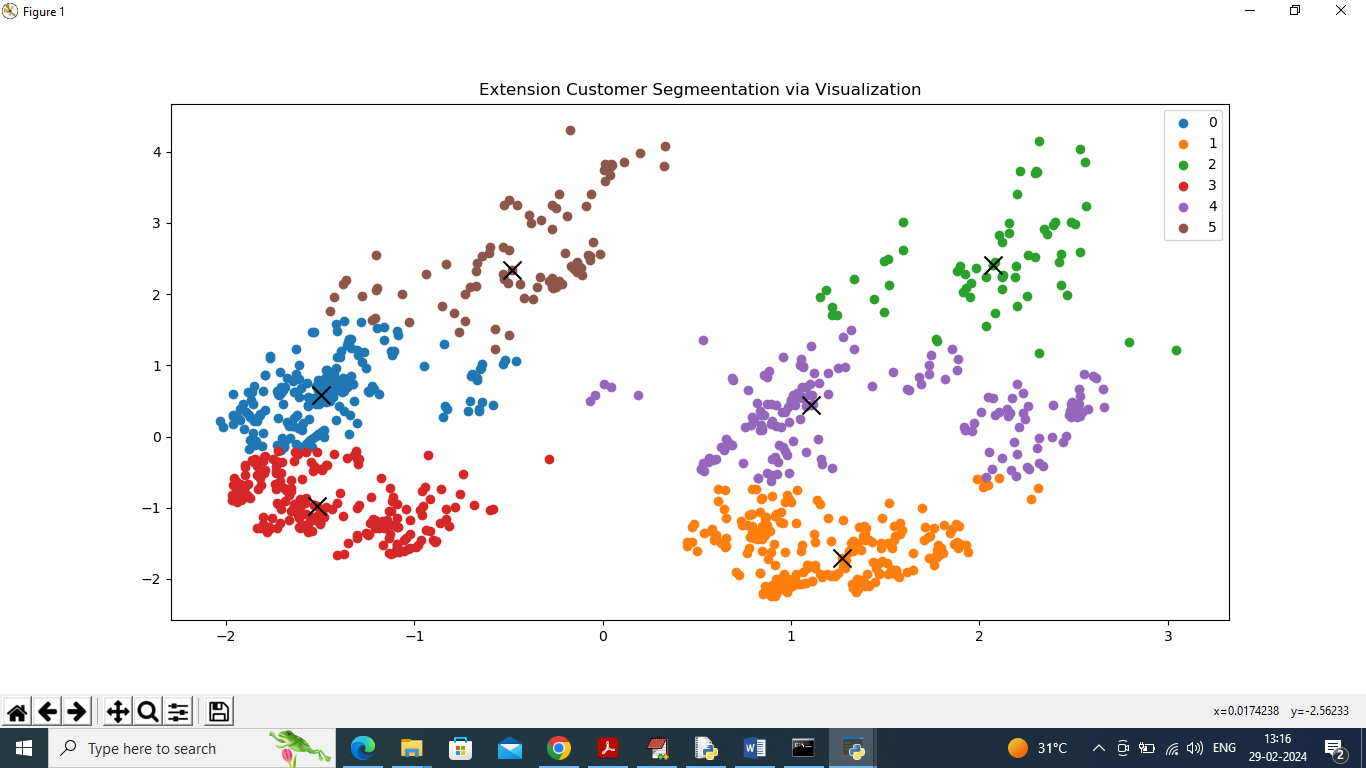
In above screen can see Gower values and now click on ‘Run PAM Based K-Medoid’ button to cluster dataset and get below output



In above screen displaying each cluster number and count of segmented customer in each cluster and in graph x-axis represents number of cluster and then can see silhouette score for each cluster number. We took clusters as 2, 4, 6, 8 and 10 and in above graph centre value is for cluster no 6 and there we got high silhouette score so 6 will be consider as best cluster. Now click on ‘"Calculate Cluster Based Revenue’ button to calculate average revenue and get below output



In above screen can see cluster no, average revenue, average page visit and total pages browser and in above table can cluster 2 and 5 has high number of orders so high revenue will be generated from clusters 2 and 5 customers. Now click on ‘Extension Customer Segment Graph’ button to view segmented customers in graph



In this applications we took 6 clusters and in above graph we have 6 different colour dots and each colour dot refers to 1 cluster and number of dots in that cluster refers to number of customers. In each cluster dots can see ‘X’ mark as cluster Centroid. So above graph is not exists in paper and we are displaying as extension and from above graph we can easily segment or understand number of customer in clusters.