**Customer Review Clustering for Marketing Insights Using KMeans**

**SUBMITTED BY : SREEJA SANYAL**

**BTECH. CSE (2026)**

**KIIT UNIVERSITY , BHUBANESHWAR**

**ABSTRACT**

In the current digital era, customer feedback acts as a valuable resource for companies looking to enhance their products and promotional tactics. However the vast availability of unstuctures reviews make manual analysis in tradional method inefficient and less accurate.This project leverages machine learning, specifically clustering algorithms, to analyze Amazon product reviews.

The main objective is to categorize customers according to their review behavior to discover significant patterns and insights.Utilizing unsupervised learning methods on structure review data, this study reveals various customer personas and suggests customized marketing strategies for every segment.

The project segments customers into meaningful groups and provide distinct feedback patterns for each segment.The result affirms better engagement,retention and strategic planning based on data informed approach.

**Keywords** : Customer Segmentation , KMeans Clustering , Unsupervised Learning , Amazon product review , Marketing Analysis , Customer Insights , Behavioral Data Analysis.

**INTRODUCTION**

Analyzing customer feedback on a large scale is essential for making well-informed business decisions.Market segmentation tradionally are based on demographics,psycograpics or purchase history.However, with digital platform review behaviout itself becomes rich dimension for market analysis.

Clustering algorithms like KMeans can automatically categorize customers who exhibit similar behaviors, facilitating targeted actions for specific segments.The approach involves comprehensive data processing , enhanced feature engineering and internal validation using evaluation metrics.

In this project, we process the Amazon review dataset, create new features that reflect customer sentiment and behavior, and apply clustering techniques to identify and examine customer segments.The ultimate goal is to translate these clusters into marketing insights.Each cluster is analyzed to propose targeted strategies.

This approach moves beyond simple review rating analysis and dives deep into the behavioral market analysis, demonstrating the power of machine learning in buisness applications.

**RELATED WORKS**

Several notable works in the field of e-commerce and market analysis that I reffered while working on this project have laid groundwork for applying unsupervised learning to consumer nehavior modelling.

Jain and Ahuja(2014) developed a KMeans based clustering model to categorize online customers int distict profiles.Their work emphisized on understanding pre-purchase behavior pattern for customer engagement.

While their work was grounded in transaction data our project extends behavioural segmentation focusing on engagement with review content,helpfullness and sentiment indicators.

Li and Lee(2024) did a comparitive study between KMeans,SVM and KMA and found out KMeans outperform SVM.This comparision was quite relevent to this project in order to choose the best approach while moving forward and yielded some valuable insights.

Jain (2010) talked about practical advantages of KMeans in customer analytics.Their views aligns with this project’s choice of KMeans over aany other models hence enabaling clear segment profiling for actionable market decisions.

Han,Kamber and Pei(2011) provided the explaination of validation metrics such as Silhoutte Score and Calinski-Harabasz Index are used in this project to determine optimal cluster.

**LITERATURE REVIEW**

In her paper Jain emphasized on creating a model that attempts to study the segmented profile of the online consumers in the context of their pre purchase behaviour(2014). Their work utilized Kmeans clustering to segment online users on e-commerce platform,categorizing them into window shoppers,loyal buyers and infrequent users.Jain’s study validated the core idea of using unsupervised learning for behavioural segmentation.

Li and Lee(2024)’s research integrates clustering algorithms within a big data ecosystem.Their model is capable of real-time adjustment to evolving customer behavior.While talking about comparision of accuracy provided by KMeans and Svm Xaotong Li and Lee.(2024) mentined that while SVM provides 72.36% accuracy KMeans provides 81.08% accuracy.

In the textbook “Data Mining : Concepts and Techniques”Han,Kamber and Pei(2011) defines the theoritical foundation for Kmeans clustering and cluster validation metrics.The use of internal evaluation tools such as silhoutte score and calinski-harabasz index in this project directly follows their recommended practices.

Jain (2010) stated in his paper that KMeans is a reliable algorithm for market segmentation particularly when interpretability and computational efficiency are priorities.

**PROBLEM STATEMENT**

This project aims to address the problem

How can we use unsupervised clustering algorithm to segment customers based on their review behaviour and how can these segments be translated into marketing strategies?

By focusing on clustering Amazon product review data,the projects seeks to go beyond traditional analysis methods and develop a data driven approach for behavioural segmentation.

**OBJECTIVE**

1. Identifying features that represent customer behaviour and engagement best in the data.
2. Choosing appropiate clustering method
3. Validating the quality of clusters
4. Analyzing the behavioural qualities of each segment to recommend personalized marketing approach.

**WORK FLOW**

Collection of data

Cleaning of the data

Engineer features from the reviews that reflect user sentiment,engagement and behavior

Evaluate clustering quality

Apply clustering techiniques to segment customers.

Analyze segments for actionable marketing insights.

**DATA COLLECTION**

**Resource**

Dataset : Amazom Product Review Dataset

Source: Kaggle

Size : 843 X 19(post formating)

843 X 27(actual datset)

Datatype: CSV format

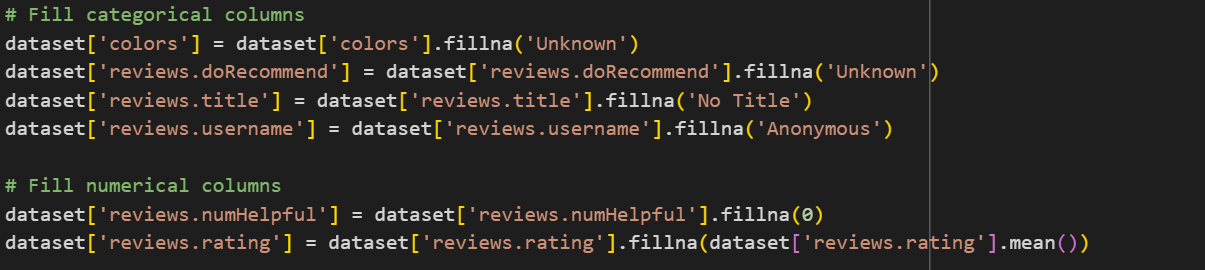
|  |  |
| --- | --- |
| id | Unique identifier for the product |
| asins | Amazon Standard Identification Number |
| brand | Brand of the product |
| name | Product name |
| categories | Product category |
| colors | Color of the product |
| dateAdded | Date the product was added to the dataset |
| dateUpdated | Date the product metadata was updated |
| manufacturer | Manufacturer name |
| prices | Price of the product |
| reviews.date | Date the review was posted |
| reviews.rating | Customer rating |
| reviews.text | Full textual review |
| reviews.title | Title of the review |
| reviews.username | Username of the reviewer |
| reviews.numHelpful | Number of user marked it helpful |
| reviews.doRecommend | Wheter the user recommends the product or not |
| reviews.sourceURLs | URL pointing to the original review |
| keys | Product key information |

**DATA PREPROCESSING**

1. **Handeling missing values**:

Categorical columns like colors filled with ‘Unknown’

Numerical columns like reviews.rating filled with the mean

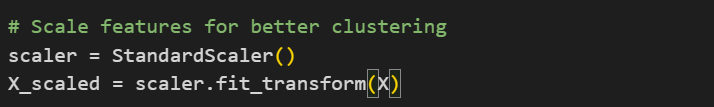


1. **Feature Scaling(Standardization)**:

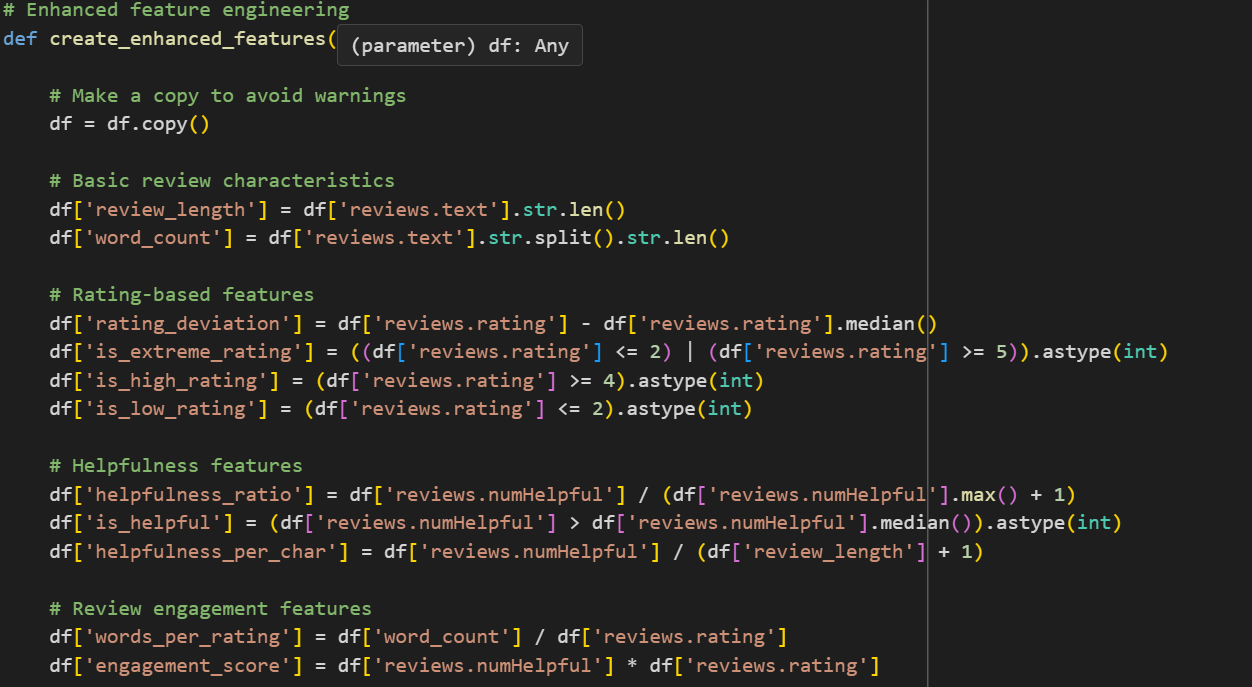
Z = value - mean

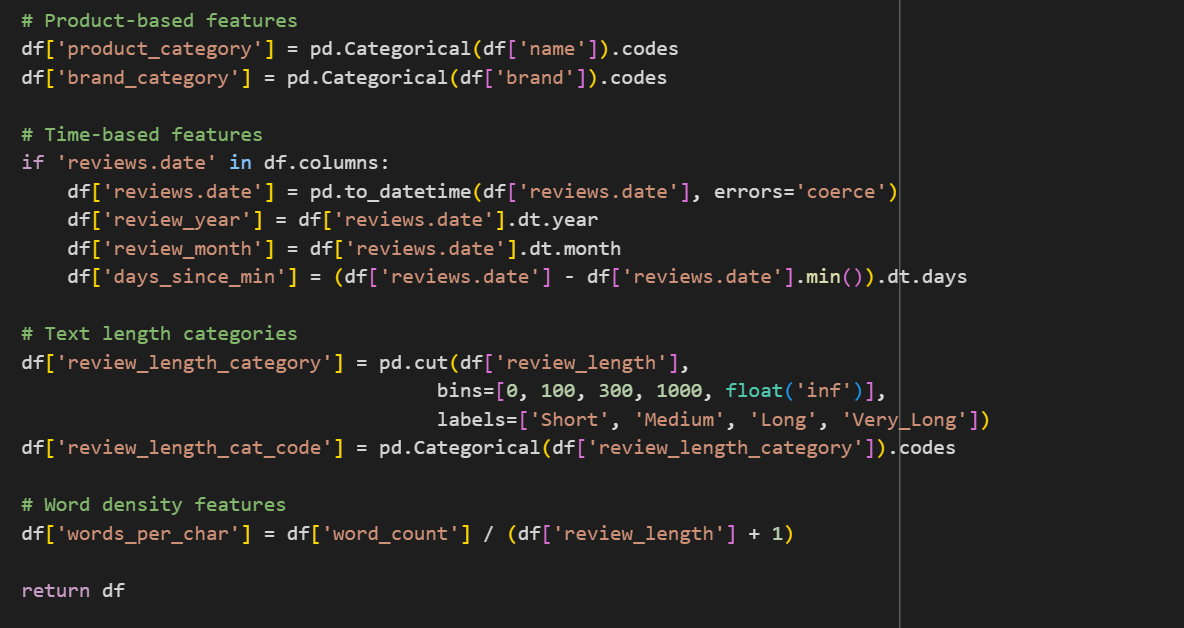
standard deviation

applied to normalize features before clustering

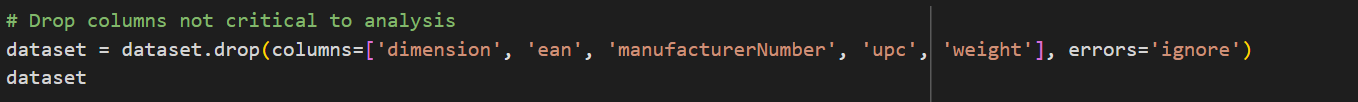


1. **Feature Engineering** : Created new columns such as review\_length,helpfullness\_ratio,is\_high\_rating,engagement\_score,rating\_deviation.





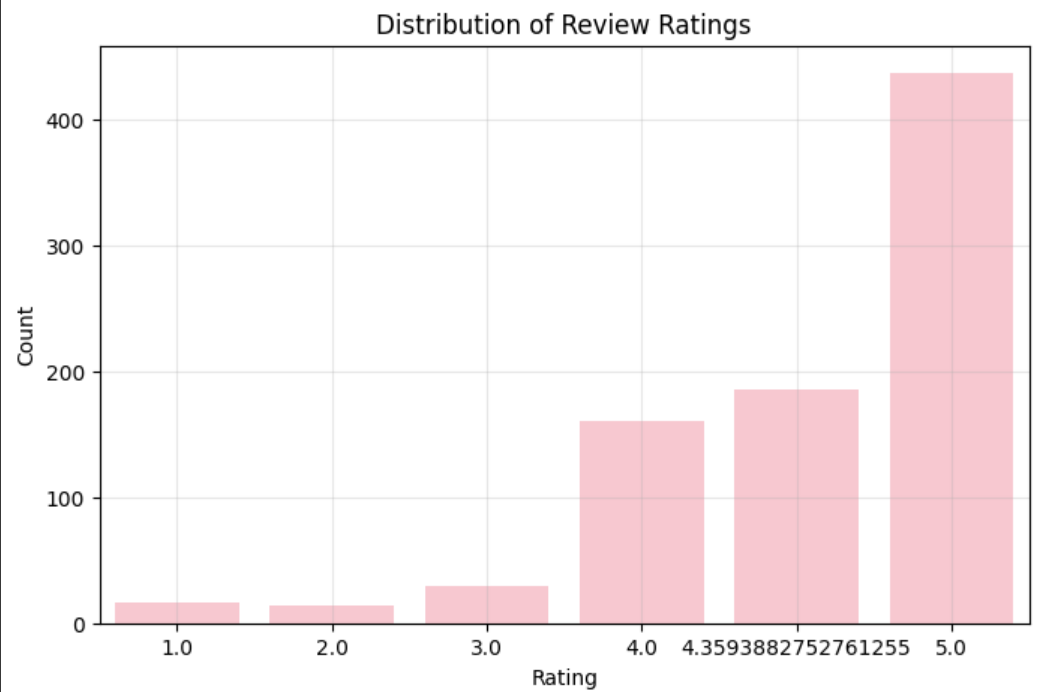
1. **Dropped Columns** : dimension,ean,manufacturerNumber,upc,weight were dropped as they were not critical for analysis.

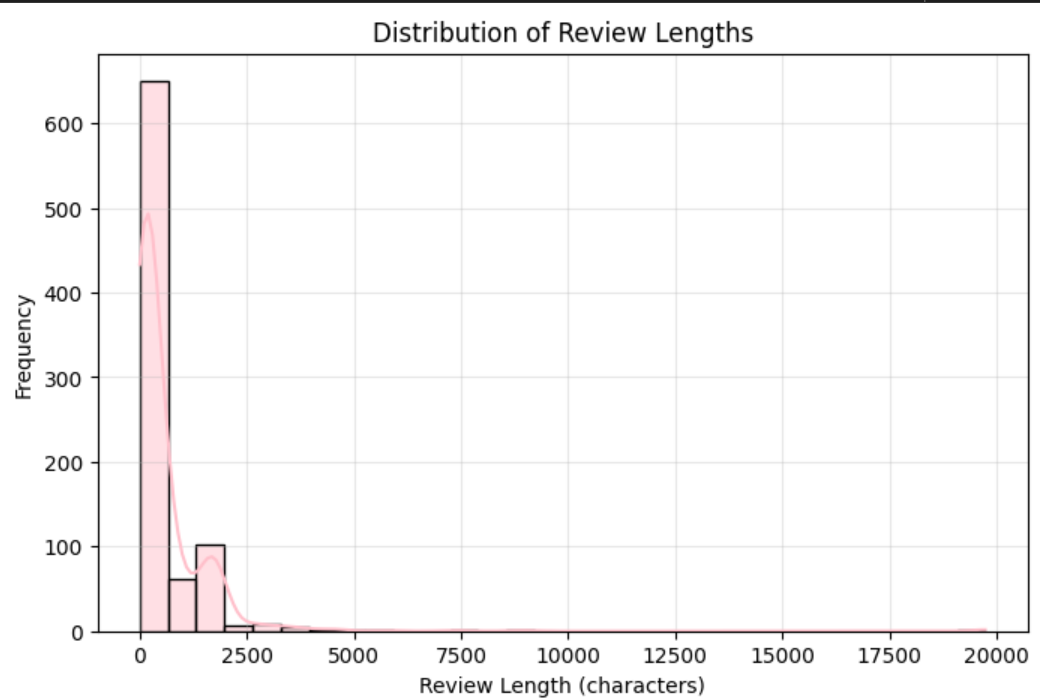


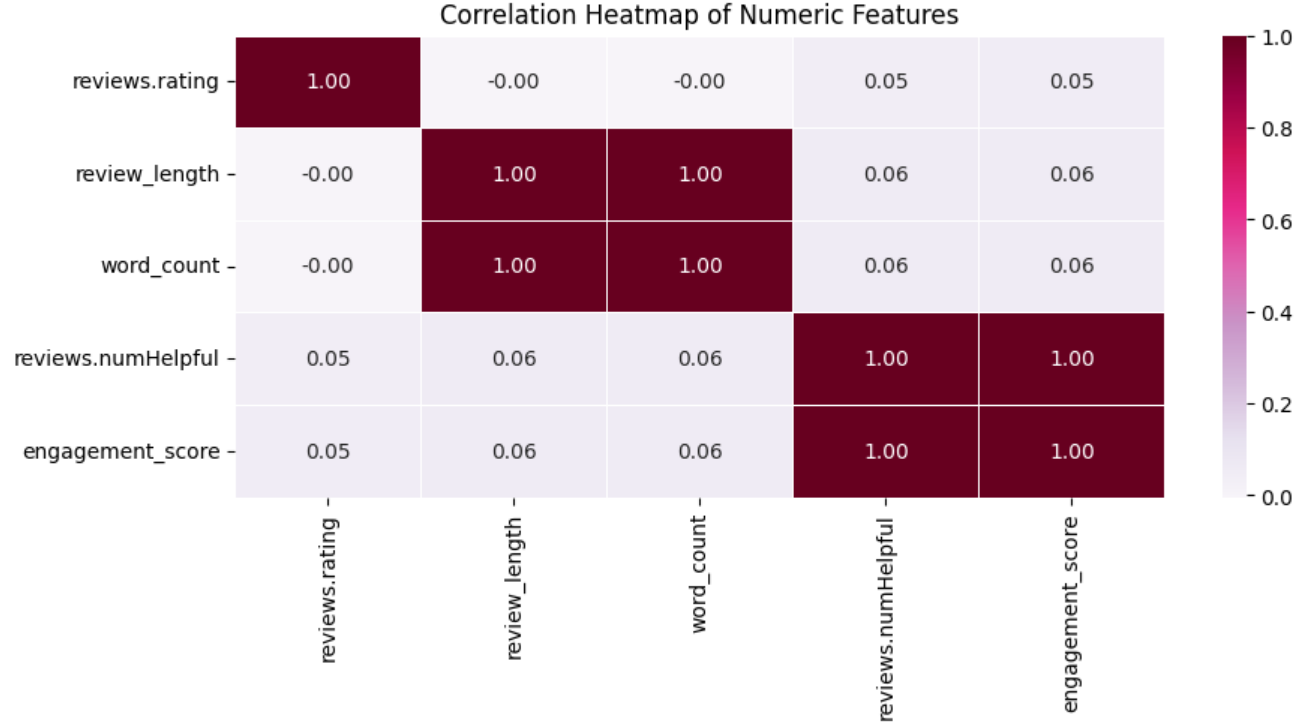
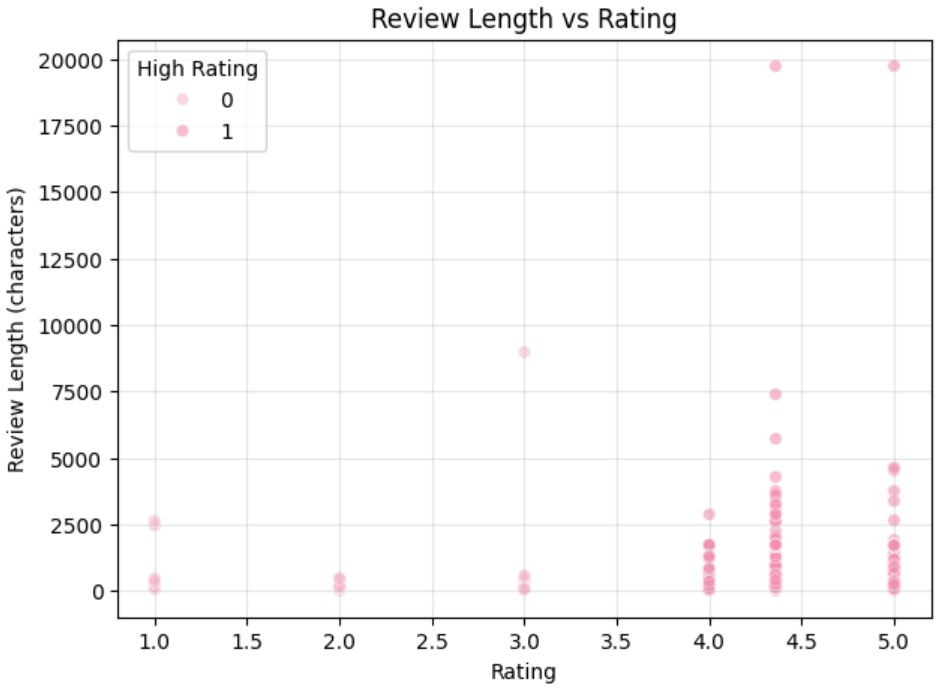
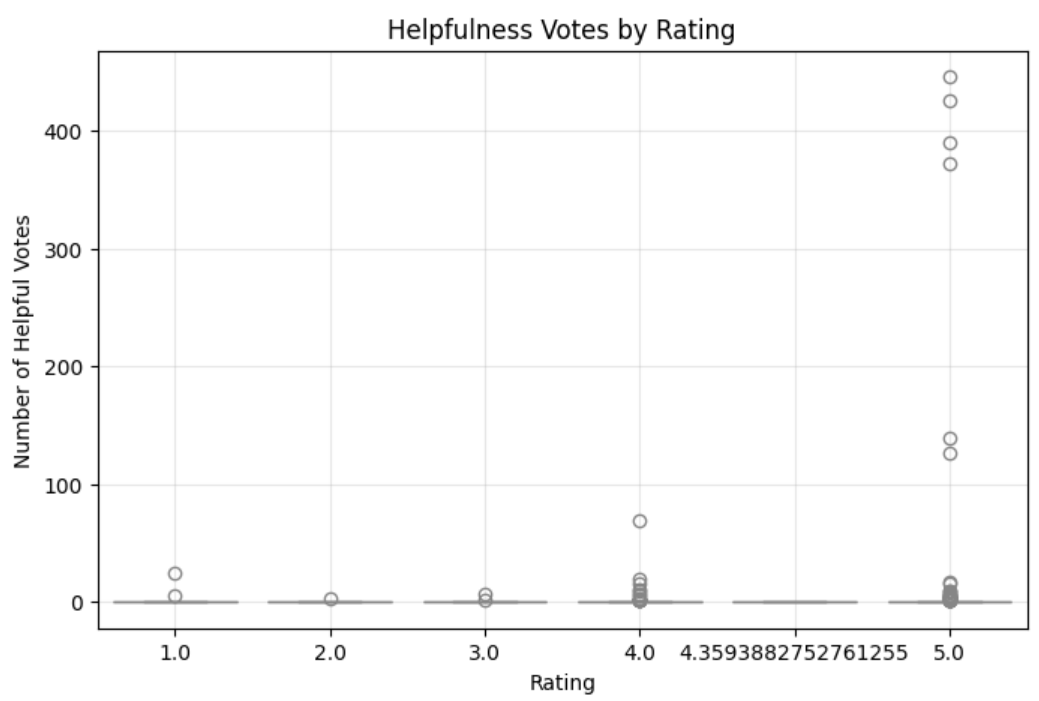
**EXPLORATORY DATA ANALYSIS**

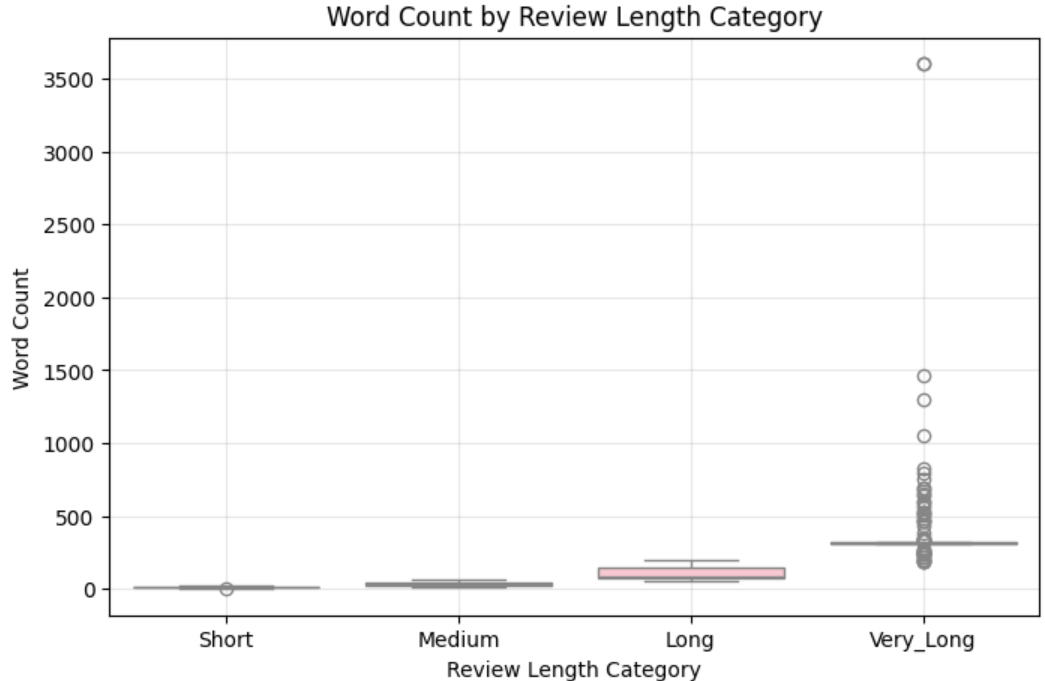
EDA involves visual and statstical analysis to understand data distribution relationship.

1. Most of the reviews are 4 to 5 star reviews indicating positive customer sentiment overall.
2. Detailed reviews tend to receive more upvotes suggesting review length can also be a factor of engagement.
3. High engagement score points to customer who are noth positive and impactful.
4. Plots showing that texual review patterns vary across rating level such as low rating reviews often have high word count suggesting dissatisfied customers explain their issues in more depth and it is also considered helpful for others.
5. Correlation heatmaps revealed that engagement\_score and review\_rating are positively collerated with helpfulness votes highlighting key features for clustering.









**BUISNESS INSIGHTS ANALYSIS**

1. Customers with consistent 5-star ratings are ideal for loyalty programs.
2. Early detection of dissatisfaction allows timely intervention.
3. Detailed reviews often provide valuable feedback and may influence others.
4. Loyal customers but they keep the review short needed to be encouraged for detailed feedback.
5. Some product have dedicated loyal segments.

**ML MODEL CREATION**

Model used : **KMeans Clustering**

KMeans is an insupervised machine learning algorithm used to group data into k distinct clusters based on feature similarity.It tries to minimize the vaariance within each cluster and maximize the variance between clusters.

The objective function

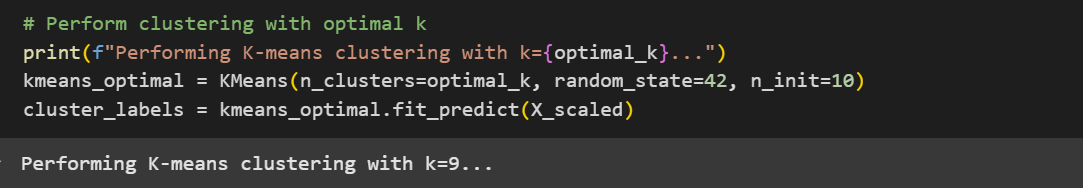
J=

where:

X = {} be the data point

The centroid of the cluster j is

The set of points assigned to cluster j is



Validation Metrics

1. **Silhoutte Score** : The metric used to evaluate the quality of clusters in clustering algorithm.

The Silhouette Score for a single samole is :

s(i) = b(i) - a(i)

max{a(i),b(i)}

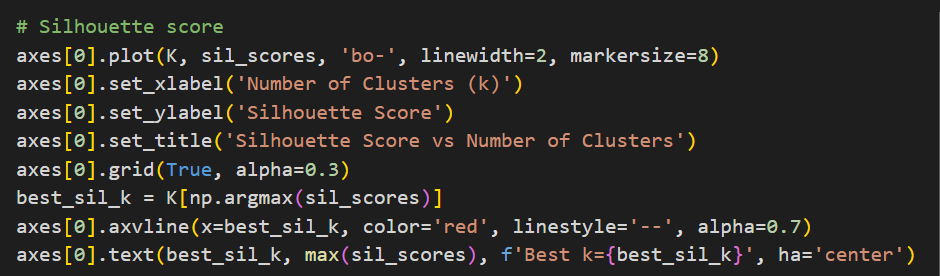
where :

a(i) = average distance of point i to all other points in the same cluster

b(i) = lowest average distance of point i to points in other clusters

Average Silhouette Score over all points gives an overall measure of clustering quality.

* Closer to 1 = better-defined clusters.
* Closer to 0 = overlapping clusters.
* Negative = incorrect clustering.



1. **Calinski - Harabasz Index** : Evaluation metric used to measure the quality of clustering.It assesses how well seperated and dense the clusters are.

CH index = Tr(SSB) . n - k

Tr(SSW) k - 1

where :

k = number of clusters

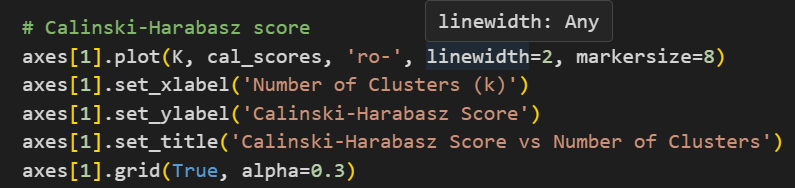
n = number of data points

SSB = between cluster dispersion i.e. how far clusters are from the verall centroid

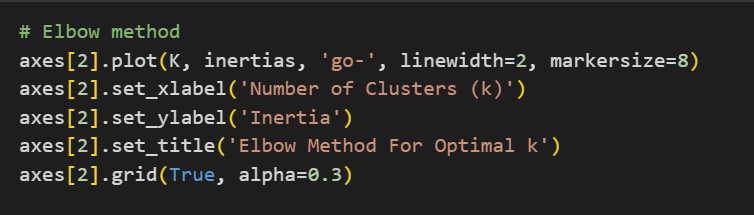
SSW = within cluster dispersion i.e. how tight the points are in each cluster

Tr = trace of a matrix i.e. sum of diagonal elements.

Higher value indicated better clustering.

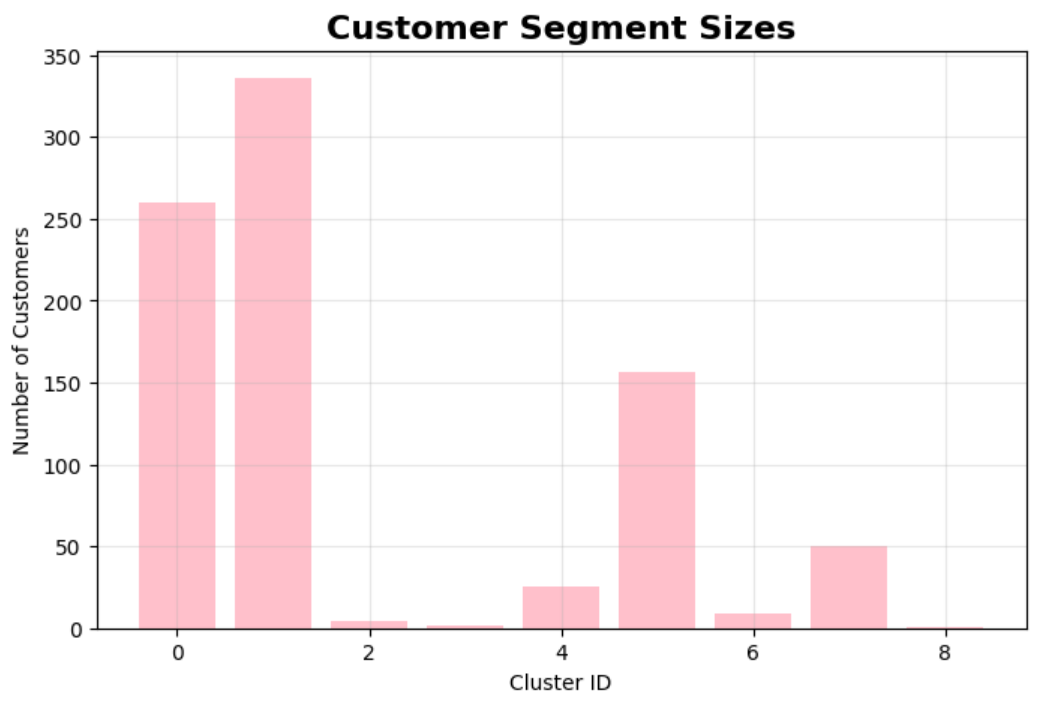


1. **Elbow - Method** : Visual technique used within clustering algorithms like KMeans to determine the optimal number of clusters by analyzing the inertia which is within cluster sum of squared errors.
2. Run KMeans from the range of values of k
3. Calculate Inertia(sum if squared distances from each point to its assigned cluster centroid)
4. Plot the number of clusters k against inertia
5. Look for an “elbow” point- the point agter which the inertia reduction slows down.



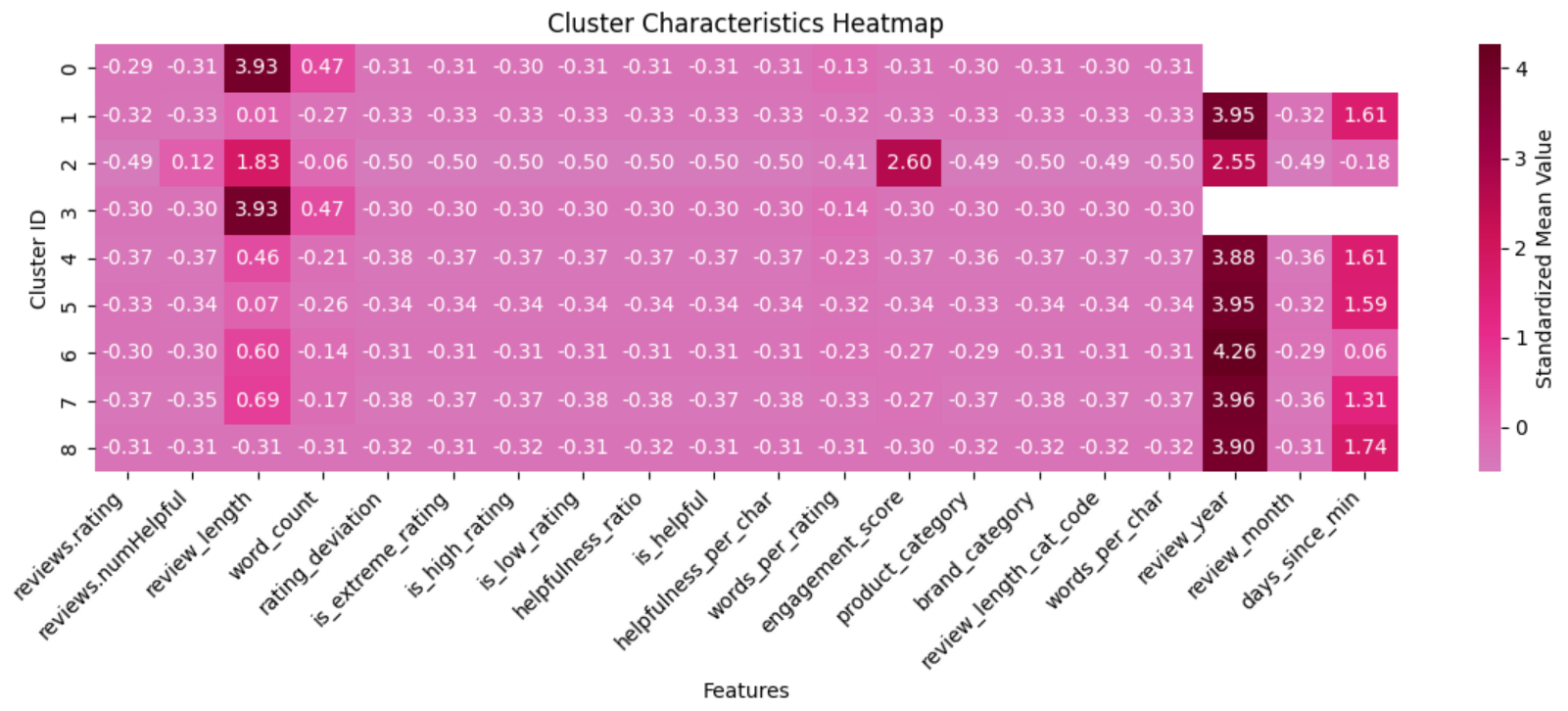
**VISUAL PLOTS**:

1. Cluster size distribution

 2. Rating vs Helpfullness



3.Feature importence map



**RESULT AND EVALUATION** :

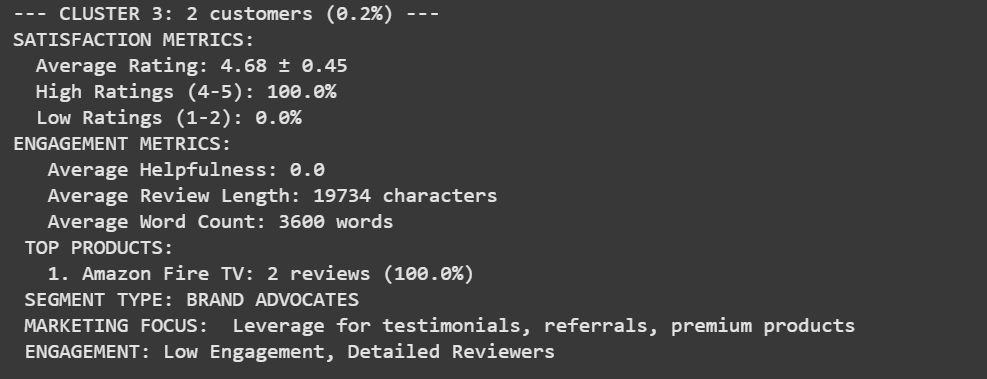
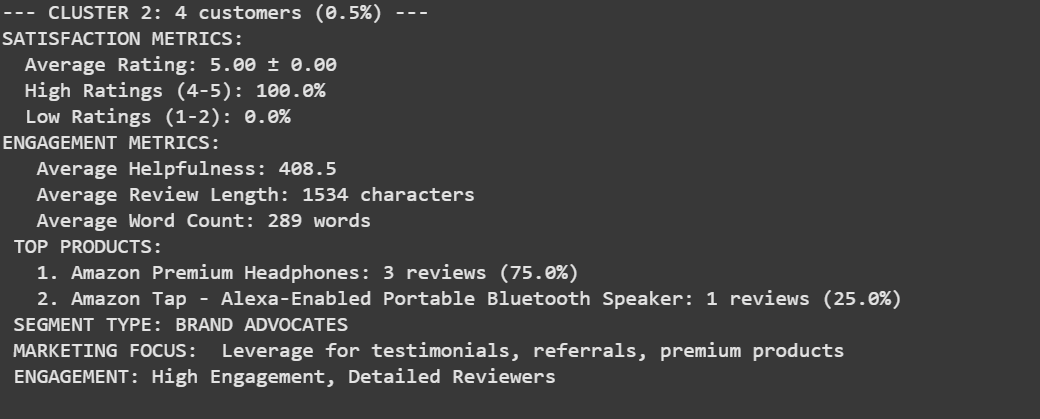
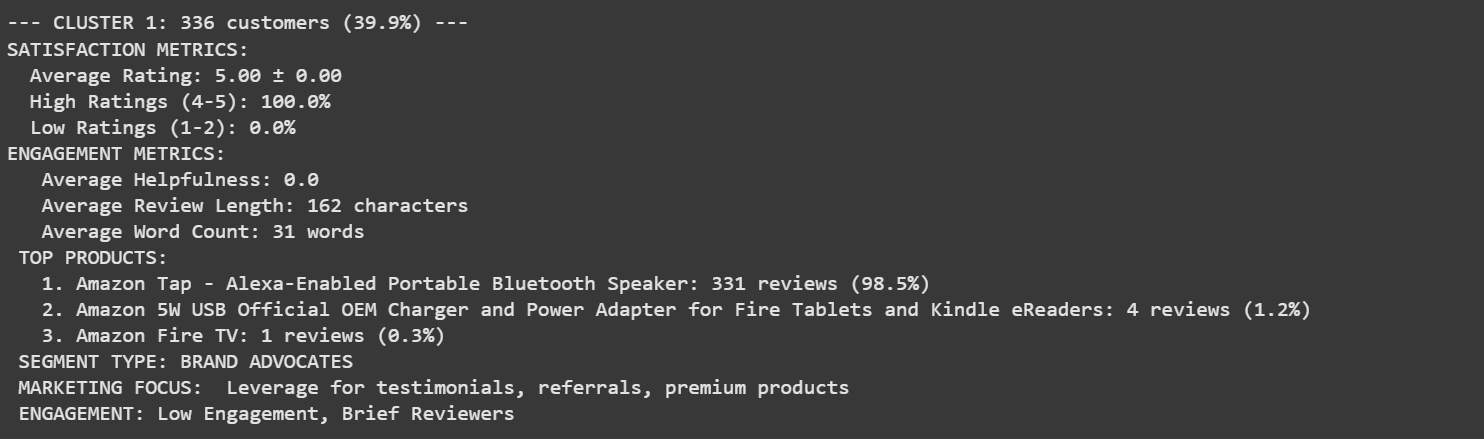
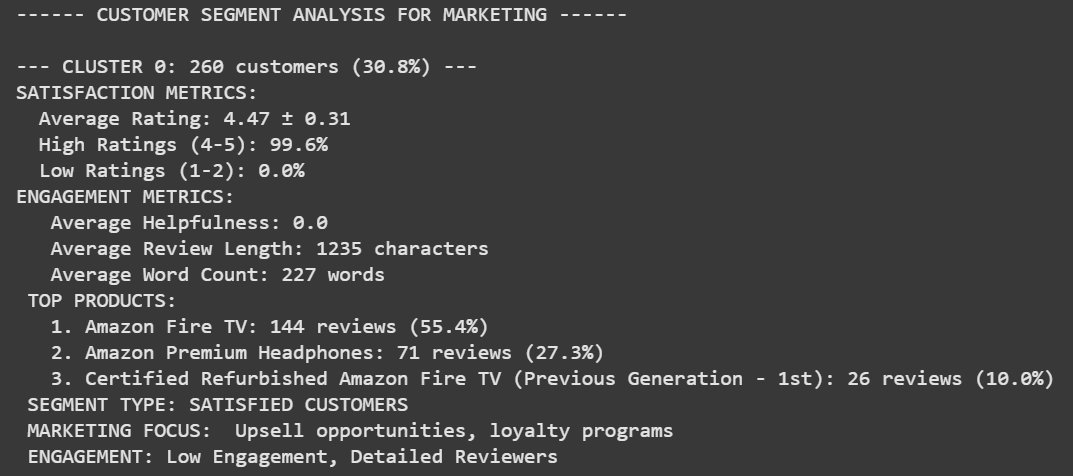
* Using a combination of Silhouette analysis and elbow method the optimal clusters are found to be 9. Thus **k = 9**.
* Silhoutte Score = 0.398 indicating moderate cluster separation.
* Calinski - Harabasz is 283.66 indication that higher value means well seperated clusters with tight internal cohesion
* Each cluster was analyzed based on its

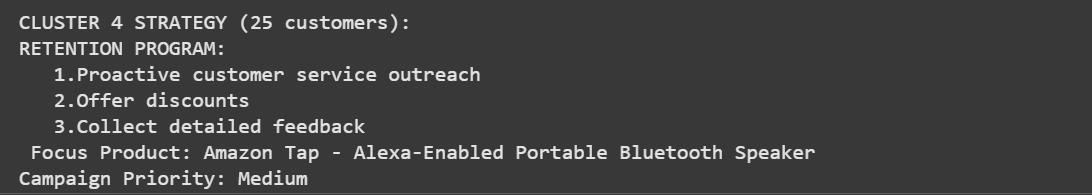
1. Average rating
2. Review length and helpfulness
3. Top reviewed products
4. Customer engagement patterns

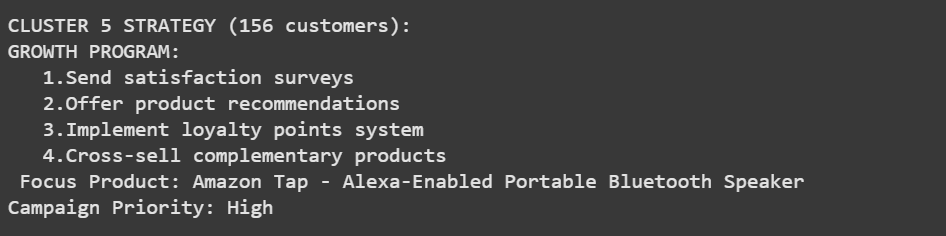
* Segment Type identified

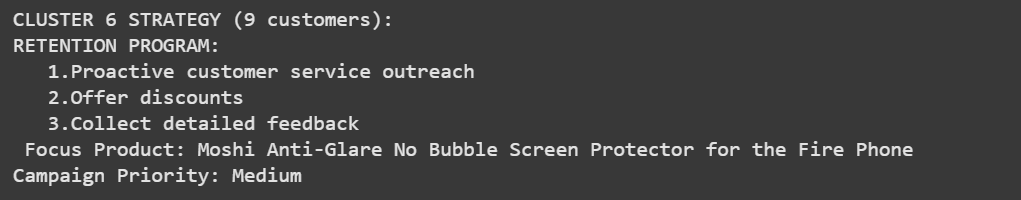
1. Higer ratings,constant praise ,ideal for analysis
2. Positive but less engaged,suitable for upselling a product
3. Frequent low ratings which require retention efforts
4. Indifferent behaviourbut can be helpful for engagement campaigns

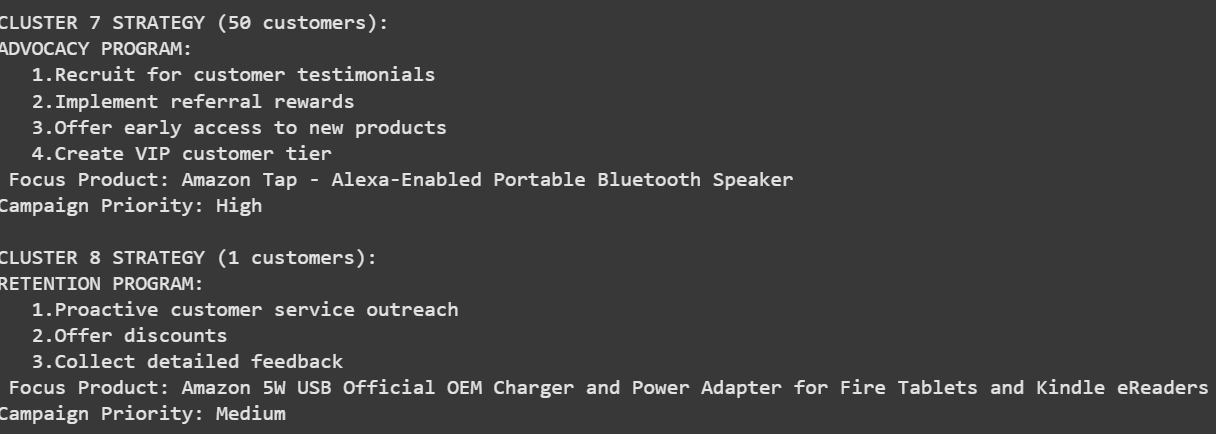
* Each clusters directly informed specific marketing strategies.











**FUTURE SCOPE :**

1.Integrate sentiment analysis via NLP

2.Real time clustering pipeline for new reviews.

3.Personalize offers based on cluster behaviour.

**LIMITATIONS :**

1. No textual sentiment analysis.
2. Missing values led to data loss.
3. Limited to Amazon device-related products only.

**CONCLUSION :**

The project succesfully demonstrated the unsupervised machine learning methodoloy in deriving actionable insights from not so perfect customer feedback.Focused on Amazon product review,we transformed raw textual and numerical data into meaningful segments through feature engineering and clustering for market analysis.

The primary objective was to corelate some patterns in customer behaviour which cannot just be available through review counts or average ratings.Using Kmeans we were able to identify nine distinct customer segments with unique review patterns,engagement levels and sentiment orientations.

The evaluation scores demonstrated that the clustering structure was statistacally valid and meaningful.Additionally visualization tools confirmed the clusters were well seperated and behaviorally diverse.

From a marketing standpoint as well this segemntation provides valuable guidance.

This project bridges the gap between data science and decision making enabaling smarter,data-driven marketing strategies grounded in customer behaviour rather than assumptions.

**BIBLIOGRAPHY**

# Customer Segmentation Marketing Strategy Based on Big Data Analysis and Clustering Algorithm(2024)

By Xaotong Li and Young Sook Lee

<https://www.researchgate.net/publication/378048316_Customer_Segmentation_Marketing_Strategy_Based_on_Big_Data_Analysis_and_Clustering_Algorithm>

# Segmenting online consumers using K-means cluster analysis(2014)

By Neha Jain and Vandana Ahuja

<https://www.researchgate.net/publication/277650903_Segmenting_online_consumers_using_K-means_cluster_analysis>

**Data Mining : Concepts and Techniques (2011)**

By Jiawei Han,Micheline Kamber and Jian Pei

<https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques>

**Data clustering : 50 years beyond K-means(2010)**

By Anil K.Jain

[https://www.sciencedirect.com/science/article/abs/pii/S0167865509002323#:~:text=The%20new%20similarity%20between%20a,et%20al.%2C%202009a).&text=res%20image%20(435KB)-,Fig.,measure%20among%20the%20data%20points.](https://www.sciencedirect.com/science/article/abs/pii/S0167865509002323" \l ":~:text=The%20new%20similarity%20between%20a,et%20al.%2C%202009a).&text=res%20image%20(435KB)-,Fig.,measure%20among%20the%20data%20points.)