BCIIT e-Journal 2026

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707

K-means: Clustering for Customer Segmentation

Manju Arora
Assistant Professor
Department of Information Technology
Jagan Institue of Management Studies
Rohini, Delhi
manju.arora@jimsindia.com

Pari Batra
Department of Computer Applications
Jagannath University
Bhadurgarh,Harayana
paribatra86@gmail.com

Abstract: The customer's behaviour will be essential for the success of the e-commerce personalized marketing and decision-making. This research presents a non-supervised Machine Learning (ML) customer segmentation based on K-means Clustering Algorithm. In this way we perform the customer clustering analysis by applying some of the most important aspects of their purchasing behaviour, i.e., purchase frequency and spending patterns, in order to identify segments of customers that can be considered as useful for businesses. Before applying the clustering procedure, we preprocess our dataset and normalize it in order to improve the accuracy of our ML model. We apply the Elbow Method to find out the optimal number of clusters to use and consequently balance model performances with model complexity. These clusters provide very relevant information regarding different kinds of customers (high-value, medium-value and low-value). Our proposed approach allows businesses to automatically define customer groups without any manual interventions and enables them to take marketing decisions and distribute resources accordingly. Finally, this study shows how ML approaches can contribute to the development of Customer Relationship Management and ultimately to the global improvement of business performances.

Keywords: K-Means Clustering, Customer Segmentation, E-commerce Analytics, Unsupervised Learning, Elbow Method.

INTRODUCTION

Every customer is different. The one walking through your door right now might know your menu, your process, or your product lineup better than half your staff do. Another might be visiting for

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707

the first time, still figuring out whether they even like what you offer.

If we treat every customer the same, it would be like responding to them without listening to their needs and what the customer wants, because the new customer is allowed to say what they want and what they are looking for specifically. Because if we keep on pushing the customers according to our traditional ways, it would be insignificant and lead to pushing the customer away.

And our goal is to bridge the gap between existing system and new technologies that can help us understand customers' behavior about their purchasing and other attributes which will contribute to our further study.

So here, we are talking about customer segmentation, which exactly is not a new idea. Many retail shops, supermarkets and other industries are customer centric as there are always new methods that let us know more about our customers and get to know them better and understand their needs in a clear, concise manner.

The best way to study customer segmentation is to understand both the words customer and segmentation. If we talk about the customer, the primary concern is what the requirements of customer are and knowing all the attributes like their age group, their buying pattern, what they like most help us in knowing them better and what will lead them back again. And if we talk about the segmentation part, which is the most important and crucial because it helps us in dividing a large number of customers into various groups where we can cater their needs and from that we can conclude various inside results that help us in growing our business and also benefit the customer as well.

The segmentation part can include dividing the customer into supervised groups and unsupervised groups where we can keep the track of the customers, we can get their information through feedback forms and also help us in knowing more about what they actually require and this information can help in bridging the gap between age, income range, geographic regions.

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707

And most certainly, we don't need to force the customers into specific groups because each customer is different and there can be a group for each customer because not everyone is the same, but we can categorize the customers according to their locality, their age group, because they have the common things. And our point is not to group the customers into categories, it's to find the patterns from their purchasing so that we can help in deriving the insights that can help us in shaping more relevant marketing, building genuine satisfaction, and supporting growth that lasts. When people feel good and special, they don't just make a purchase and move on, they come back and become something more than a customer.

EXISTING SYSTEM

The foundation of this analysis is a dataset made from many sources and here's a dataset that can help in learning about this more:

Table 1: Showing customer segment dataset for different groups.

Customer_ID	Age	Income	Spending Score	Segment
C1	22	18000	65	Young customer
C2	23	20000	70	Young customer
C3	27	25000	75	Young customer
C4	38	40000	50	Average customer
C5	29	25000	68	Young customer
C6	32	18000	45	Average customer
C7	30	28500	52	Average customer
C8	45	49500	40	High income

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707

Here's how we calculated spending score:

Spending score = (customer spending – min spending / max spending – min spending) * 100

This above table shows that the sample we have taken of the customers from some sources like supermarkets (usually where people shops) , indicates that the segment – Young customer are mostly have high spending score and usually the high income segment has low spending score which clearly tells us about lifestyle of people .

The data we have is not any data. Every single piece of data has a reason for being there. When we put this data into the K-Means Clustering algorithm we can see the groups of customers who are similar to each other. These customers have characteristics and habits and they like similar things. We use these groups to make marketing campaigns that're just for them. We also use this data to make the time they spend in our store better. In the end we have better and longer relationships, with our customers because of the K-Means Clustering algorithm and the data we have.

This project is different from ways of grouping customers because it looks at what is happening now and what might happen next. It uses real time information. Tries to guess what customers will do tomorrow. This means companies can change quickly and stay popular when peoples buying habits change. The thing that really matters here is not about using computers to look at numbers.

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707



Fig. 1: CRM Platform

PROPOSED SYSTEM

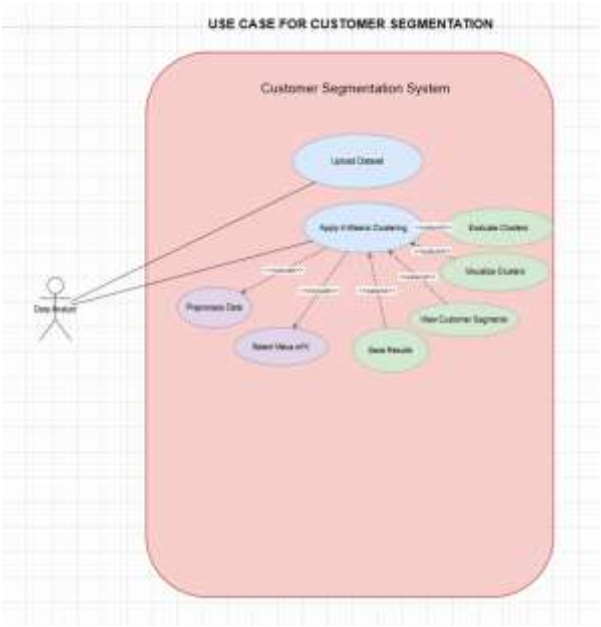


Fig. 2. Use case for Customer segmentation System

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707

The above use case for customer segmentation system shows the following components: an actor named Data Analyst, which communicates with use cases -> Upload Dataset and Apply K-Means Clustering.

In this System boundary names "Customer segmentation system", we have various relationships with Apply K-Means clustering use case, which include preprocess data, select value of K, and has various options like save result, view customer segments, visualize clusters, evaluate clusters, which are in the extend relationship form.

Then K-Means Clustering makes a profile of each group. This profile is, like a plan that tells businesses how to talk to each group of customers in a way that really matters to them. It gives them an idea of what to do to engage each group of customers.

This system is really valuable because it helps businesses in ways. It lets them make their marketing right for the people they want to reach. They can also respond fast when peoples buying habits change. They can make their customers feel like they are getting something special just for them. When there are a lot of stores trying to sell the same things being able to do these things is not just a good thing. It is something that businesses really need to do to stay ahead of the others. The system gives businesses the ability to tailor their marketing, which's a big part of being successful Create customer experiences that feel personal rather, than something that every other store is doing.

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707

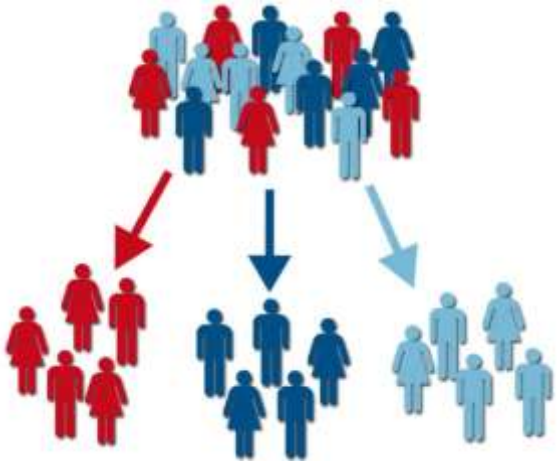


Fig. 3: K-means Clustering

LITERATURE REVIEW

Customer segmentation is a very common topic and widely explored in the field of data mining and marketing analytics because it helps us in understanding how buying behavior of customers affects our decision making and our business.

There are various studies that show how customers have been grouped in the past on basis of their buying, age group, regions etc. K-means clustering is one of the most common techniques that is used by everyone widely because it is very simple and easy to work with, and it also provides us efficiency.

In K-means clustering, the data is divided into a number of clusters where each data point belongs to the cluster within the nearest means. In many studies, it is found that many researchers and research papers have shown that K-means is one of the finest and most repeated techniques that is used to identify the patterns.

From the various studies that have been acquired, there is focus on improving and growing the result through visualization techniques. And in k-means clustering, most of the visualization tool is done by a Python programming language, and the tools are made in the programming language

only using matplotlib, which is used for scatter plots, cluster diagram, which help the technical as well as the non-technical reader to understand how the customer groups are working. Because there are so many existing approaches which focus on diagrams and algorithms rather than going into the system design.

Table 2: Showing comparison of existing technologies from sources

Author (Year)	Technique Used	Dataset Type	Strengths	Limitations
MacQueen (1967)	K-means	General dataset	Simple, efficient	Requires predefined K
Hartigan & Wong (1979)	K-means variant	Statistical data	Improved clustering accuracy	Sensitive to initialization
Kaufman & Rousseeuw (1990)	Hierarchical	Customer data	No need to predefine clusters	High computation cost
Tan et al. (2013)	Clustering + Visualization	Large datasets	Better interpretability	Limited system integration

To know what our customer actually want, we need to understand the depth and know how to do our marketing very well, because when people buy things from the companies, the company need to understand how customer is a really fast group, because there are trends and patterns that are changing day by day, and by understanding the customer on a deeper level, we can

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707

understand how we can improve our decisions so quickly by just noticing the pattern. We can also promote the products on social media through various platforms like Instagram, Facebook, Twitter, etc. because most of the customers are now approaching marketing.

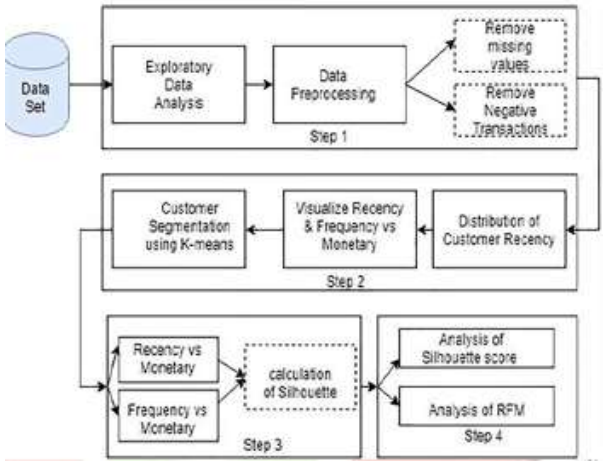


Fig. 4: Architecture

METHODOLOGY

In my methodology, I have taken a structured machine learning pipeline to perform the customer segmentation and show k-means clustering algorithm. Here, the process begins with generating the synthetic dataset that helps us in simulating the realistic customer attributes like age, income, spending score, and purchase frequency.

In this approach, we have taken controlled data while maintaining the practical relevance for analysis and deriving the patterns. The dataset is pre-processed using a feature scaling technique that standardizes the values and here the k-means clustering relies on distance-based calculation.

We have used elbow method which determines the optimal number of clusters by analyzing the error across the different cluster values. After identifying the number of clusters, we have grouped them into distinct segments based on their similar behavior and then using the visualization

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707

libraries in the Python, we have created scatter plot which has helped us in seeing the results and allowing us to identify the key characteristics and differences among the customer group.

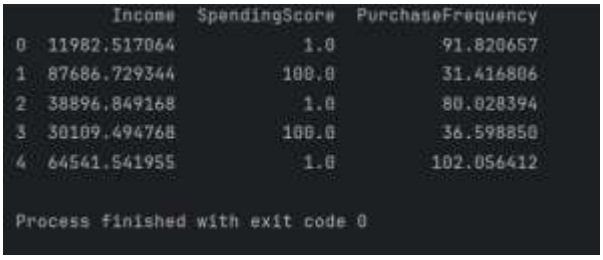
This approach has ensured that the transformation of raw data into meaningful insight and information can help us in increasing our growth and make supportive, efficient decisions.

3.1 Data Collection

The first and foremost step in any research is to collect the data and see if the data is authentic or not. Here, we have taken synthetic data created by python library scikit-learn which generates clustered data. And synthetic data is preferred here due to controlled environment and easy to test and see the result plus the benefit of no privacy issues.

Here we have chosen various features for customers, like their age, spending score, purchase frequency. Because these feature helps us in understanding the concept better about our customers and their patterns.

The dataset consists of 500 data points, and it was generated using make_blobs function which creates clusters based on Gaussian distribution.



	Income	SpendingScore	PurchaseFrequency
0	11982.517064	1.0	91.820657
1	87686.729344	100.0	31.416806
2	38896.849168	1.0	80.028394
3	30109.494768	100.0	36.598850
4	64541.541955	1.0	102.056412

Process finished with exit code 0

Fig 5. Dataset generated using Python library

3.2 Data Preprocessing

The next step is cleaning any noise from the data, handling missing values and errors like duplicates which is very important before we proceed further. It is important step as it helps us in

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707

achieving accuracy and make our data efficient. For K-means clustering, scaling is very important, almost compulsory. We can use Z-score (Standardization) for this.

Here's a snippet of code for better understanding:

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler ()

scaled_data = scaler.fit_transform(df)

print(scaled_data[:5])
```

In the above code, we are importing StandardScaler from scikit-learn library and this tool is very important because it helps us in normalizing our data and it is needed because k-means uses distance and features with large value like income can dominate the smaller ones.

In the next step, we have created an object named scaler where we have called our class and after that we have applied scaling and which is the most important part because it is doing two things at once. It is calculating meaning, standard deviation for each column as well as transforming it for each value and meaning becomes zero and standard deviation becomes one which comes out as the result.

3.3 Applying K-Means Clustering

Here we have used k-means algorithm that was applied to divide or segment customers into distinct groups. This algorithm works by initializing centroid and then increasing and assigning data points to the nearest centroid based on Euclidean distance. The centroids are then updated until convergence is achieved. Here we have used elbow method which has helped us in Determining the optimal number of clusters, here, the error was computed for different values of k and the point where the rate of decrease changes was selected as the optimal number of clusters.

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707

Here's a short snippet of the code for Elbow Method:

```
from sklearn.cluster import KMeans

import matplotlib.pyplot as plt

wcss = []

for i in range (1, 10):

    kmeans = KMeans(n_clusters=i)

    kmeans.fit(scaled_data)

    wcss.append(kmeans.inertia_)

plt.plot(range (1,10), wcss)

plt.title("Elbow Method")

plt.xlabel("Number of Clusters")

plt.ylabel("WCSS")

plt.show()
```

And here's the snippet of code for K-means Model:

```
kmeans = KMeans (n_clusters=3)

clusters = kmeans.fit_predict (scaled_data)

data['Cluster'] = clusters
```

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707

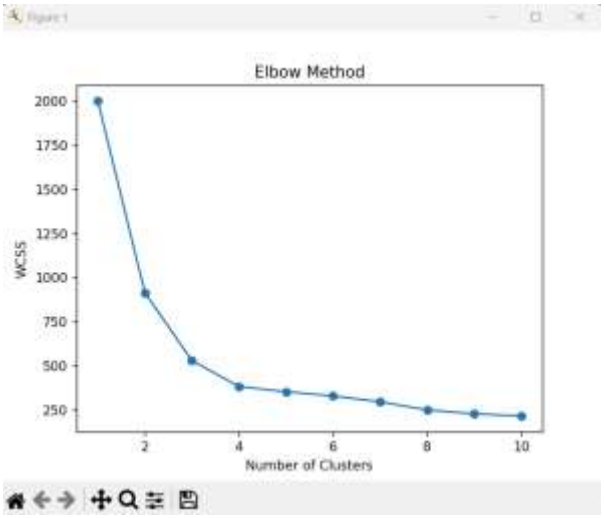


Fig 6: Elbow Graph for Determining Optimal Number of Clusters

In the above diagram for elbow method for determining optimal number of clusters, what is actually happening here is that the x-axis tells us about the number of clusters, while the y-axis tells us about WCSS, which means within-cluster sum of squares. As the K increases, clusters get smaller and Error, that is WCSS decreases. But after a certain point, the improvement becomes very small, and that bend point is called elbow point, which is observed at k is equal to 4, indicating that 4 clusters provide an optimal balance between reducing the error and maintaining the simplicity.

3.4 Visualization and Interpretation

After our programming and getting out the results, we have come to the final output of the k-means, which is shown here as visualization of customer segments using k-means clustering. Here, Figure 7 illustrates how customer segments are obtained using k-means clustering. It also helps us in understanding the cluster separation because it is benefit for us if the clusters are separated clearly. And we have used each color for distinct cluster, indicating group of customers based on their characteristics. Moreover, these results were used and can be used for further analysis that

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707

can be performed by calculating the mean values of the features within each cluster to interpret the customer behaving patterns.

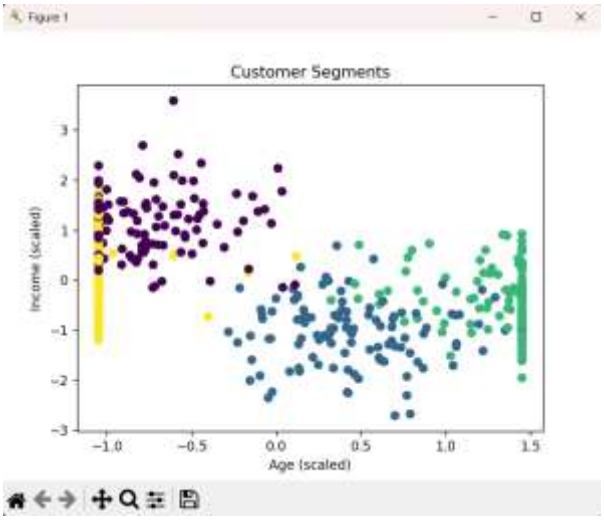


Fig. 7: Visualization of Customer Segments using K-means Clustering

In the above graph, which shows visualization of customer segments, here each dot represents one customer and there are different colors which represent different clusters. As we can see, the axis contains the features of age and income, which helps us in knowing how customers within the same cluster share similar characteristics while those in different have distinct behavioral patterns.

Here's a snippet for cluster visualization:

```
cluster_summary = df.groupby('Cluster'). Mean ()

print ("\nCluster Summary:\n", cluster_summary)
```

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707



Fig. 8: Cluster Visualization

3.5 Tools and Technologies Used

For the research paper, to gather some practical knowledge I have used Python Programming language which is easy to code with and interpretable.

As python provides various libraries like pandas and NumPy to deal with data that we need to handle. I have used NumPy for handling integer-based data and for the tabular, spreadsheet like data I have used pandas.

Python has various libraries which helps us in visualization as well so that we can take insights from it and help make decisions as well that can be understood by non-technical person as well.

These libraries are scikit-learn which is a popular machine learning library. While with pandas and NumPy we handle array type data, this focuses on deriving actual intelligence from the dataset like house price prediction which is the most common example.

CONCLUSION

In this project, the most important part and the main goal, which seemed pretty easy but was very actually hard, was how businesses can get to know their customers when there is so much information and data set out there in the world. The answer to that, we found out, is not to collect more information, but to use that information in a smarter way that can help us in deriving results.

By using various tools and technologies and platforms for gathering out some insight for this project, we have found various patterns in customer behavior that was not obvious before. What

Journal of Innovations in Computer Science and Trends in IT, (October 2026), Vol 3, Issue2, ISSN 3048-4707

we found it before was a bunch of numbers, graphs, illustrations that meant something else. And we were not getting a clear picture of how actually customers think before they spend their money because every customer is different from each other. And in this project, we find out how each unique characteristic of the customer makes it possible to come up with a plan that actually makes the customer comfortable because they are specially designed for them rather than everyone else.

The most important asset for any business is to know their customer. And gathering the customer cluster or the dataset is very important because knowing what kind of customers are there in the world and how we can plan strategies and form them so that we can understand how customers buy products, which help the business increase their profits, is very important. Because by analyzing everything, we can see that business can grow and take precautions in the future by monitoring their stock, doing the right thing at the right time, like discount, sales, sending out messages and emails to everyone. If the businesses focus on customer needs and requirements and they become customer centric in a very strategical manner, they will not miss the chance to make money.

And last but not least, it would be very helpful if the businesses knew about which product is the best-selling and which is popular among the customers that they are willing to pay so much for it and there are repetitive orders. These findings did not just give us a number to look at; they told us what customers really wanted and how much we can engage them in it. This information is very useful because it helps us in getting to know the customer at the right time, which adds value to our business as well.

For the people who are marketing this information is very useful as it helps them to talk to the customers at the right time with correct messages that really mean something to them and add value. They do not have to send out messages to everyone in the reach and hope that someone is out there to listen. Instead, they can send messages to the groups of customers who are grouped as regular, specialized under certain group and know that they will really care about what they are saying and be truthful which will be huge pleasure for us to have such loyalty and authenticity.

REFERENCES

[1] J. R. Vulchi, "Customer Segmentation Using K-Means Clustering Algorithm," 2021.

[2] H. Kilari, S. Edara, G. R. S. Yarra, and D. V. Gadhiraju, "Customer Segmentation using K-Means Clustering," *International Journal of Engineering Research & Technology (IJERT)*, vol. 11, no. 3, 2022.

[3] P. Perapu, "Customer Segmentation Using K-Means Clustering for Personalized Marketing Campaigns," *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, vol. 11, no. 3, pp. 810–815, 2025.

[4] K. Tabianan, S. Velu, and V. Ravi, "K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data," *Sustainability*, vol. 14, no. 12, 2022.

[5] Y. Li, J. Qi, and W. Mu, "Customer Segmentation Using K-Means Clustering and the Hybrid Particle Swarm Optimization Algorithm," *The Computer Journal*, vol. 66, no. 4, pp. 941–962, 2023.

[6] Y. Zheng, "Customer Segmentation Research in Marketing through Clustering Algorithm Analysis," *Journal of Intelligent & Fuzzy Systems*, vol. 45, no. 4, 2023.

[7] T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury, "Customer Segmentation using K-Means Clustering," in *International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 2018.

[8] M. Singh, A. Singh, M. Gupta, and R. Reddy, "Leveraging K-Means Clustering and Hierarchical Agglomerative Algorithms for Scalable AI-Driven Customer Segmentation," *Journal of AI & Machine Learning Research*, 2022.

[9] J. M. John, O. Shobayo, and B. Ogunleye, "An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market," *arXiv preprint arXiv:2402.04103*, 2024.

[10] M. E. Celebi, H. A. Kingravi, and P. A. Vela, "A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm," *arXiv preprint arXiv:1209.1960*, 2012.