

REPORT
ON
CUSTOMER SEGMENTATION USING RFM MODEL

BY
VAIBHAV VINIL KUMAR
2022A7PS0253U
COURSE NUMBER: CS F376

AT



BITS Pilani, Dubai Campus
Dubai International Academic City, Dubai
UAE

First Semester 2022-23

**BITS Pilani, Dubai Campus
Dubai International Academic City, Dubai
UAE**

Course Name: DESIGN PROJECT

Course No: CS F376

Duration: 23.08.2024 - 10.01.2025

Date of Start: 23.08.2024

Date of Submission: -

Title of the Project: Customer segmentation using RFM model

ID Number: 2022A7PS0253U

Name of the student: Vaibhav Vinil Kumar

Discipline of Student: Computer Science (CS)

Name of Project Supervisor: Dr. Vilas Gaidhane

Key Words: Customer Segmentation, RFM Model, K-Means Clustering, RM K-Means, Silhouette Index, Box-Cox Transformation, Cluster Compactness, Elbow Method, Skewness Correction

Project Area(s): Clustering Analysis, Skewness Correction

Abstract:

Efficient customer segmentation helps with a better understanding of the needs of customers and assists with retaining more customers. It is said that retaining customers is far more valuable than finding new customers as a company can deploy specific marketing strategies to target specific groups of customers which can help with increasing profits. This project aims to analyze the K means and RM K means clustering algorithm. Then a new approach (Data Tuned RM K means) is proposed and tested against the already existing algorithms. The approach and algorithms is explained and finally the results obtained will be compared by their cluster compactness by analyzing the cluster evaluating metrics such as the Silhouette index and Davies Bouldin index.



Signature of Student

Signature of Supervisor

Date:

Date:

ACKNOWLEDGEMENTS

I express my sincere gratitude to Prof Souri Banerjee, Director of BPDC for providing me with an opportunity to enhance my learnings through a well-planned and intuitive program design of this university. Next, I would like to extend my heartfelt gratitude to Dr. Vilas Gaidhane, my Project Supervisor, for believing in me that I can do this project under him and inspiring me. His constant guidance, support and motivation helped me understand not only the concepts of segmentation by clustering but also understand how research work is done at a professional level. I would also like to thank my family for their constant support and encouragement that helped me with the completion of the project.

Student Name:

Vaibhav Vinil Kumar

ID Number:

2022A7PS0253U

LIST OF FIGURES

- Figure 1: The elbow point in an inertia vs K graph
- Figure 2: Architecture diagram of K means clustering
- Figure 3: Frequency distribution graph of F score
- Figure 4: Frequency distribution graph of F score after Box Cox transformation
- Figure 5: Architecture diagram of Data tuned RM K means
- Figure 6: Silhouette plot of K Means clustering
- Figure 7: Silhouette plot of RM K Means clustering
- Figure 8: Silhouette plot of Data tuned RM K Means clustering
- Figure 9: Silhouette plot of data tuned RM K Means clustering
- Figure 10: Average Silhouette width
- Figure 11: Average Davies Bouldin width

CONTENTS

Abstract

Acknowledgements

List of Figures

1. Introduction.....	6
2. Literature Review	7
3. Data Description	8
Recency (R):.....	8
Frequency (F):	9
Monetary (M):	9
4. Existing Methodologies for Clustering.....	9
4.1 Elbow Method	9
4.2 K Means Clustering	10
5 Proposed Methodology: Data Tuned RM K-Means	11
5.1 Skewness of the data.....	11
5.1.1 Positive Skew	11
5.1.2 Negative Skew	11
5.1.3 Zero Skew	11
5.1.4 Calculations and Interpretations.....	12
5.2 Limitations of using skewed data.....	12
5.3 Box-Cox Transformation.....	13
5.3.1 Log-Likelihood.....	14
5.4 Data tuned RM K Means Clustering	15
5.5 Implementation and testing environment	16
6 Clustering Evaluation Metrics.....	16
6.1 Silhouette Index.....	16
6.2 Davies Bouldin Index	17
7 Results and Conclusions.....	17

1. Introduction

Customer segmentation is the process of segregating consumers into distinct clusters to tailor marketing strategies and improve customer engagement. It helps us in identifying and addressing the unique needs of a customer. In this project, the focus is on segmenting customers based on their RFM values where RFM stands for recency, frequency and monetary using the RM K means algorithm and comparing it to the regular K-means algorithm. RFM (recency, frequency and monetary) values are used to segment customers, where we calculate customers RFM scores, then rank them based on it, which divides them into groups of customers from that of a high valued customer to low valued customers.

The K means algorithm is a machine learning algorithm used for segregating data into distinct groups based on their common features. It helps us form groups of similar customers, where the behavior of each of the clients in the same cluster is similar to another customer from the same cluster but different compared to the behavior of a customer from a different cluster.

The RM K-means algorithm is repeated median K-means algorithm where instead of using the regular mean to calculate the centroid of a cluster it uses the median of the points in the cluster. Therefore, these algorithms help us gain an insight on customer behavior and help us in understanding customers purchasing patterns.

The first section of this paper provides a simple introduction by introducing the topic of the paper and giving a detailed explanation of each of the other sections.

The second section reviews the existing literature on customer segmentation, focusing on the application of the Recency, Frequency, Monetary (RFM) model and K means clustering. While numerous papers display the credibility of K means and RM K means their limitations like skewness of data are rarely discussed. This gap in literature motivates the present study, which directly addresses this issue.

The third section provides a short description of the data used in this project, obtained from Kaggle. The dataset comprises of customer transaction records, characterized by a significant number of attributes. However, this analysis focuses specifically on Recency, Frequency, and Monetary Value (RFM) metrics.

The fourth section analyzes the methodology and discusses the algorithms that are used (K-means and RM K-means). The fifth section proposes the data tuned RM K-means method. It also discusses the skewness of data, limitation of using skewed data and Box-Cox transformation. The sixth section discusses the various evaluation metrics where the Silhouette index and Davies Bouldin index are discussed. This section discusses how the effectiveness of each of the algorithms is measured and how they are being compared.

Finally, we conclude the paper with the results and conclusions where we compare each of the algorithms and show the result of the comparison.

2. Literature Review

Christy et al. (2021) underlined the usefulness of RFM ranking for real-world applications while putting forth a novel method for client segmentation. The study created a ranking system that enhanced consumer insights and decision-making procedures by utilizing the RFM model. By classifying consumers into meaningful groups according to their transactional data, the authors concluded that using RFM ranking improves segmentation [1].

Kansal et al. (2018) explored the benefits of K means clustering for customer segregation. The study showed how the algorithm could classify clients according to their purchase patterns, giving companies useful information to target certain clientele. Although successful, the authors pointed out certain drawbacks when dealing with outliers and proposed hybrid approaches to increase robustness. K means is a commonly used algorithm for clustering due to its straightforwardness and effectiveness. It is considered a standard algorithm that, by utilizing RFM values and defining the amount of clusters to be created beforehand. It is an iterative algorithm that calculates centroid values repeatedly until the sum of squared distances between each data point and the centroid of its designated cluster (often called the within-cluster sum of squares (WCSS) or inertia) is minimized. Additionally, the RFM variables are normalized using min-max normalization to enhance accuracy [2].

Namvar et al. (2010) introduced a two-phase clustering method combining statistical and computational techniques for intelligent customer segmentation. Customer data was preprocessed in the first stage, and then clustering methods were applied. Although the authors acknowledged computational complexity as a barrier, they found improved segmentation accuracy, particularly in identifying specialized client groups [3].

Aryuni et al. (2018) analyzed K means and K medoids clustering for customer segregation in the banking sector. According to the study, K medoids was resilient to outliers, whereas K means worked well on big datasets. The authors used these techniques to pinpoint important client groups that may be the focus of customized marketing campaigns [4].

Kumar (2022) applied K-means clustering to segment shopping mall users, showcasing its effectiveness in identifying distinct customer segments. The elbow technique and silhouette index were used for validation, and the study underlined the significance of choosing the ideal number of clusters. The results demonstrated how useful K-means is for enhancing retail customer engagement tactics [5].

The user Sarahm has provided the appropriate dataset of an RFM model for customer segmentation which had been used by various other studies for customer segmentation purposes [6]. Through the use of clustering algorithms, the analysis helped organizations prioritize their marketing efforts by grouping clients into actionable segments ultimately increasing lifetime value and client retention.

A study published in *Entropy* (2021) examined the use of silhouette analysis for evaluating clustering performance. A reliable indicator for evaluating cluster separation and compactness was found to be the silhouette index. To ensure correct customer segmentation, the authors emphasized its use in verifying K-means and other clustering algorithms [7].

Abbasi et al. (2022) implemented customer segmentation via the RFM model along with data mining techniques. The work improved segmentation accuracy by adding cluster compactness metrics and skewness correction. This strategy was suggested by the authors as a way to identify important clients and enhance marketing results [8].

Von Hippel (2011) explored the statistical concept of skewness and its implications for customer segmentation. The study underlined how crucial it is to fix transactional data skewness in order to increase the precision of clustering algorithms such as K-means. For skewness correction, the Box-Cox transformation was suggested as a workable solution [9].

Leung (2022) discussed maximum likelihood estimation in the context of actuarial principles, emphasizing its relevance for parameter estimation in clustering models. The study demonstrated how MLE (Maximum Likelihood Estimation) could optimize the selection of clustering parameters [10].

The reviewed papers collectively address customer segmentation using various approaches but face several challenges. The susceptibility of clustering algorithms like K-means to outliers and the challenge of figuring out the ideal number of clusters—which frequently calls for validation approaches like the elbow method or silhouette index—were highlighted in numerous papers, including Kansal et al. (2018) and Aryuni et al. (2018). The problem of skewed transactional data was highlighted by Christy et al. (2021) and Abbasi et al. (2022), who emphasized the necessity of skewness correction methods such as Box-Cox transformation to improve clustering accuracy. Namvar et al. (2010) and Kumar (2022) identified computational complexity as a limitation, especially with large datasets or when integrating multiple clustering phases. Studies using the RFM model, such as Sarahm (2023) and Abbasi et al. (2022), faced difficulty balancing the weights of recency, frequency, and monetary indicators to achieve meaningful segmentation.

3. Data Description

The data set is obtained from Kaggle [6], which contains roughly five hundred thousand records with various data attributes. For the purpose of segmentation and obtaining the RFM score, attributes such as Quantity, Invoice Date, Unit Price and unique ID are used. By performing data preprocessing algorithms on the dataset available and omitting the invalid records, the RFM score is obtained in a scale from 1 to 5, 5 being the most ideal scenario for all cases. The paper presented makes use of the RFM model. It is considered as one of the simplest yet powerful models to analyze the customers spending behaviors and segregate them. These variables are defined below.

TABLE 1
RFM Sore calculator

Score	Recency (Days)	Frequency (number of transactions)	Monetary (in dollars)
5	7	15	Above 12,000
4	30	12	9000-12,000
3	90	9	6000-9000
2	180	6	3000-6000
1	365	3 and less	Below 3000

Recency (R):

Recency is a measure to understand the customer's latest purchase. The customers that have purchased items more recently are likely to return sooner. A shorter time indicates a good sign for the seller whereas a higher value shows that the customers are potentially lost and less likely to return to the store.

Frequency (F):

Frequency relates to the measure of transactions that a consumer makes in the course of a fixed period of time. A larger value makes the customer more valuable to the store as such customers can be considered as loyal customers.

Monetary (M):

Monetary value signifies the amount of funds that a customer has spent over a fixed interval of time. It helps us understand the profitability of a customer and target them in ways to make more profit out of them. Understanding such customers allows sellers to make decisions that keep this group of customers satisfied and keep returning for more purchases to their store.

4. Existing Methodologies for Clustering

We analyzed the widely used method of customer segmentation that makes use of the K Means algorithm, it identifies its strengths and weaknesses. Then it approaches another clustering algorithm developed over the K Means algorithm, The RM K means algorithm (Repeated Median K Means). The silhouette index of both clustering methods is analyzed finally to determine the more optimal algorithm. To decide the optimal number of clusters to be created, the elbow method is performed.

4.1 Elbow Method

The elbow method is an effective method to decide the optimum amount of clusters to segment the dataset. While making use of large number of clusters may seem ideal, after a certain point, it doesn't help analyze data better. Hence the elbow method is implemented.

Algorithm

- Apply the clustering algorithm (K means/ RM K means) for a reasonable range of integers such as 1 to 10.
- For each value of K, find the within-cluster sum of squares (WCSS also known as inertia).
- Plot WCSS against K (number of clusters) graph.
- Identify the elbow point. It denotes the point at which rate of reduction in inertia reduces drastically. The graph looks similar to that of the arm and the "elbow" of the graph signifies the elbow point.

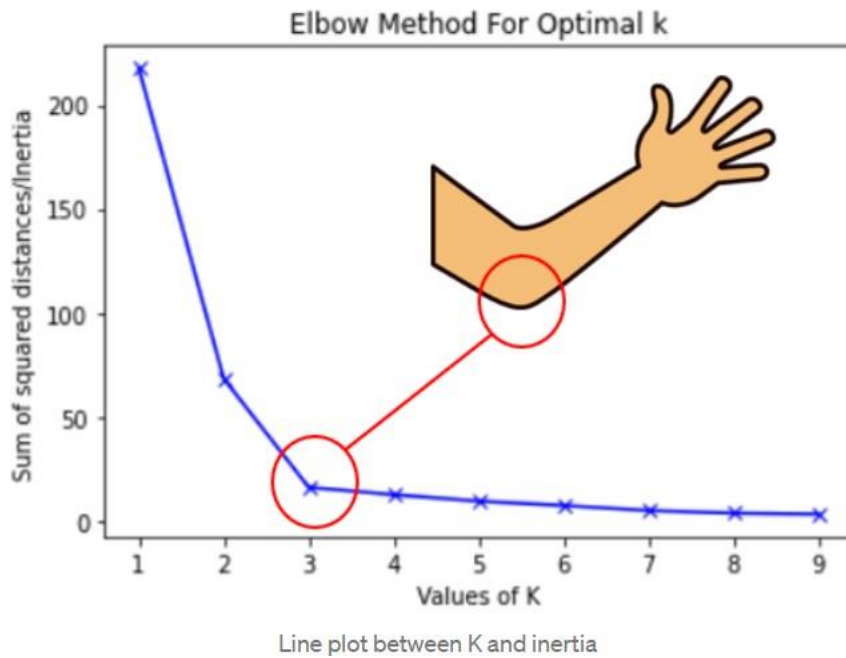


FIGURE 1

The elbow point in an inertia vs K graph

4.2 K Means Clustering

K means clustering is considered as a simple algorithm. Making use of the RFM values and defining the amount of clusters to be created beforehand, it creates a partition. K means is an iterative algorithm that calculates position of centroids before each iteration repetitively until the sum of squared distances between all data point and the centroid of its allocated cluster (often called the within-cluster sum of squares (WCSS) or inertia) is minimum [8]. The value of RFM variables may be normalized via min-max normalization.

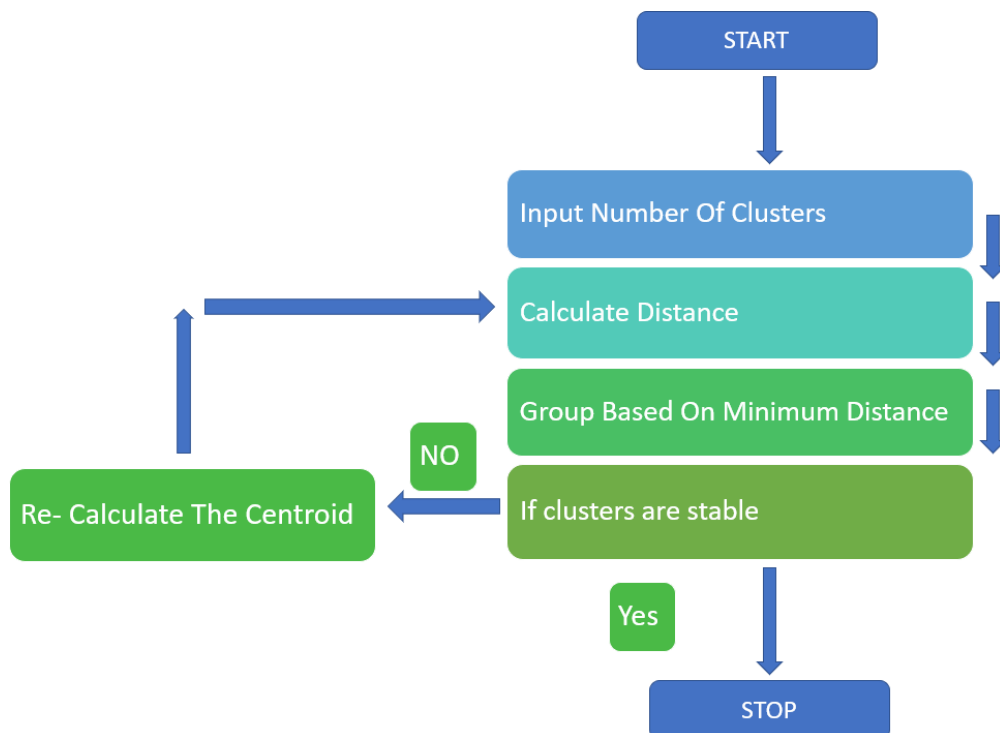


FIGURE 2

Architecture diagram of K means clustering

Algorithm

Input:

- N records of customers.
- K number of clusters.

Outputs:

- Customers segmented into K clusters.

Algorithm:

- K random points are initialized as centroids.
- Assign clusters similar to the K means method.
- Calculate the new centroids, this time, by using median of the points assigned to the cluster. This is an improvement over the regular K means algorithm and reduces the impact of outliers in the dataset.
- Check the inertia and repeat the steps if not minimum or significant change still occurs on iteration.

5 Proposed Methodology: Data Tuned RM K-Means

Upon analyzing the existing methodologies and the dataset, it is observed that they have potential to give better defined clusters.

5.1 Skewness of the data

It is observed on plotting the histogram of density plot of each parameter, that the data is significantly skewed. Skewness refers to the degree of the lack of symmetry in the data [9]. It can be expressed visually by histograms that project the frequency distribution of the data. There are three types of graphs.

5.1.1 Positive Skew

Positive skew (also referred to as right skew) is observed when the distribution has a longer right tail and comparatively a greater number of data points on the left side of the graph. The mean value is greater than the median value in such distributions. The value of skewness is greater than zero.

5.1.2 Negative Skew

Negative skew (also referred to as left skew) is observed when the distribution has a longer left tail and comparatively more data points on the right side of the graph. The median is greater than the mean in such distributions. The value of skewness is less than zero.

5.1.3 Zero Skew

Zero skew (also referred to as symmetrical distribution) is observed when the data is distributed evenly around the mean. The mean and median are roughly indifferent in value. The value of skewness is roughly zero.

5.1.4 Calculations and Interpretations

Mathematically, the skewness of an attribute can be calculated using the formula:

$$Skewness = \frac{n}{(n-1)(n-2)} \times \sum_i^n \left[\frac{(x_i - \bar{x})}{\sigma} \right]^3$$

n : the sample size

x_i : the value of the i^{th} observation

\bar{x} : the sample mean (average of all observations)

σ : the sample standard deviation

The following can be inferred from the obtained value of skewness:

$0 < \text{Skewness} < 0.5$: Slight right skew, almost normal.

$0.5 < \text{Skewness} < 1$: Moderate right skew.

$\text{Skewness} > 1$: Highly right-skewed (strong positive skew).

$-0.5 < \text{Skewness} < 0$: Slight left skew, almost normal.

$-1 < \text{Skewness} < -0.5$: Moderate left skew.

$\text{Skewness} < -1$: Highly left-skewed (strong negative skew).

The skewness of the F score is expressed and calculated. It is observed that the data is heavily right skewed.

5.2 Limitations of using skewed data

- Clustering algorithms rely on distance metrics to group similar points. Skewed data points can heavily dominate the results of calculation causing the clusters to be formed based on the skewed attribute's features of the dataset rather than the meaningful patterns found within the dataset.
- The centroids of clusters end up being misplaced and closer to the extreme values than they are supposed to be.
- Skewed data tends to cause overfitting or underfitting due to the high influence of outliers in distributions.
- Clustering algorithms are generally highly sensitive to initial centroid placements when the data is significantly skewed.
- Increased number of iterations in the algorithms. The clustering algorithms requires more number of iterations to achieve convergence.

To combat this, certain transformations may be applied to the data.

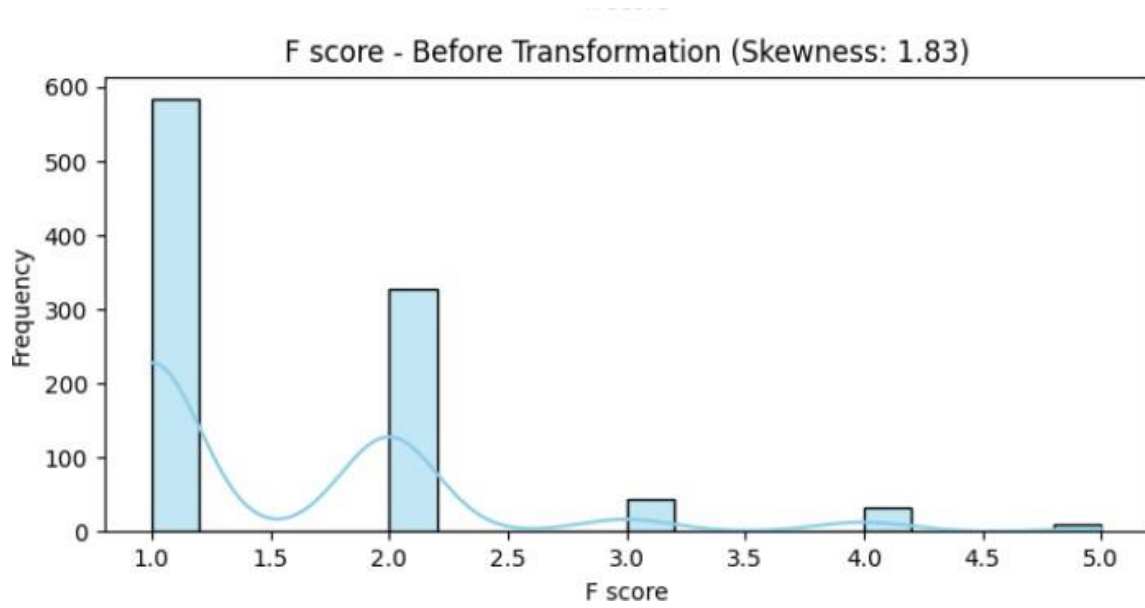


FIGURE 3

Frequency distribution graph of F score

5.3 Box-Cox Transformation

The Box-Cox transformation is a method for statistical data preprocessing. It aims to make the data normally distributed to improve the performance of certain algorithms. The equation of the transformation is explained below in Figure 5.

$$Y(\lambda) = \begin{cases} \frac{X^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \ln(X), & \text{if } \lambda = 0 \end{cases}$$

Box-Cox Transformation equations

where:

X : original data value

λ : transformation parameter

$Y(\lambda)$: Transformed value

Here, λ is the transformational parameter that determines how the transformation occurs. The optimal λ is found by maximizing the log-likelihood function. A range of λ values ranging from -5 to 5 is considered, and for each value, the box-cox transformation is applied.

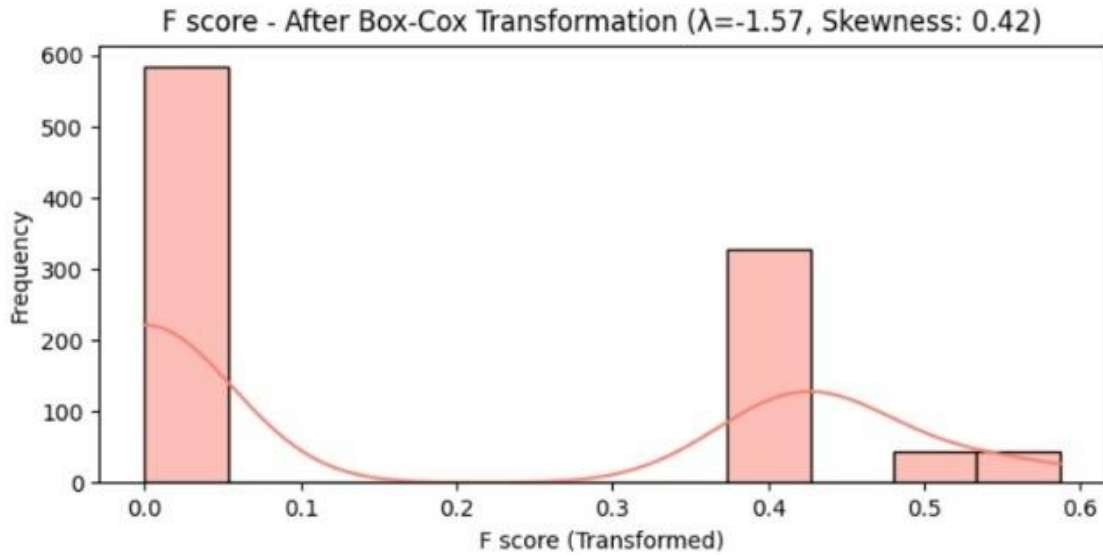


FIGURE 4

Frequency distribution graph of F score after Box
Cox transformation

The log-likelihood of the transformed data is calculated and the λ value which maximizes the log-likelihood is chosen as the optimal value. This is done by software tools available in python.

5.3.1 Log-Likelihood

The likelihood method is a measure of how well a particular model fits the data.

$$\log_L(\lambda) = -\frac{n}{2} * \log(\sigma^2) + (\lambda - 1) * \sum(\log(y_i))$$

where:

n: the number of observations .

σ^2 : is the variance of the transformed data.

Λ : is the transformation parameter.

y_i : is the original data values.

It is simpler to maximize the natural logarithm of the likelihood function [10]. Likelihood values are generally either extremely small or large in value, making it harder to graph. Using log makes computations simpler and optimization easier.

5.4 Data tuned RM K Means Clustering

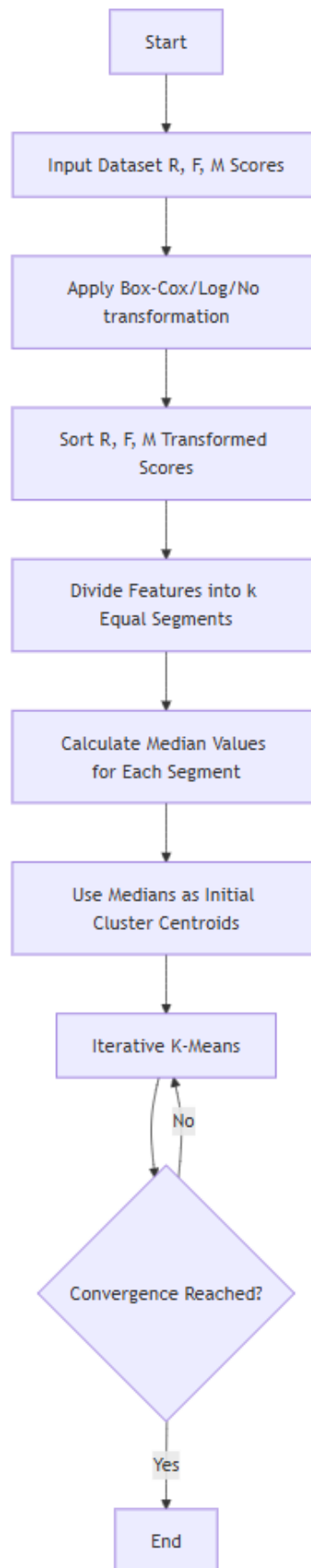


FIGURE 5

Architecture diagram of Data tuned RM K means

Algorithm

The RM K means algorithm also identified as Repeated Median K means algorithm was proposed specifically to tackle the limitations of the K Means algorithm, the randomness centroid initialization [10]. Random selection of centroids faces the possibility of initializing the centroids close to one another which would ultimately lead to less meaningful clusters being formed. This can also lead to a larger number of iterations required to reach an optimal segmentation. RM K means finding the median of the RFM values, each variable separately. These median values are assigned as the centroids. Setting these values as centroids is more meaningful and reduces the chances of centroids being initialized close to each other due to the random factor of the regular K Means algorithm.

In this paper, we propose a data tuned RM K-Means method in which we try various combinations of transformations on each attribute to find the most optimal clusters. The transformations used include box-cox transformation, log transformation, or no transformation at all. This way we find the best possible scenario to obtain clusters that perform significantly better in terms of our evaluating metrics. We also compare the output of data tuned RM K means with regular RM K means in which all attributes are transformed using the Box-Cox transformation. We will refer to this as Box-Cox transformed RM K means in this paper. This is to understand whether simply applying Box-Cox transformation on all attributes would outperform the data tuned RM K means algorithm.

5.5 Implementation and testing environment

The experiments were performed on the Google Colab environment, making use of cloud-based resources for clustering and storing data. The codes used were written and executed in the python language of version 3.10.12. Various libraries such as NumPy, Pandas, Matplotlib and SciKit-learn were made use of for data manipulation, analysis and visualization. The Intel Xeon (2) @ 2.200GHz CPU was primarily used. The code uses GPU acceleration provided by Google Colab, using a NVIDIA Tesla T4. The operating system used was the Ubuntu 22.04.3 LTS x86_64 which was hosted by the Google Compute Engine.

For storage and access of the dataset, the Google Drive was mounted to the Google Collab environment providing ease in access to the dataset. 12.7 GB of Ram was available out of which 1.5 GB was utilized.

6 Clustering Evaluation Metrics

6.1 Silhouette Index

The silhouette index is the measure of how closely related an object is to its own cluster in comparison to the other clusters. Its value confines within -1 to 1 where -1 denotes that the object doesn't match the cluster it is a part of and 1 indicates that the object matches the cluster it is part of.

A silhouette score of above 0.5 denotes credible clustering whereas a score below 0.25 denotes poor clustering. A score within 0.25 and 0.5 signifies decent clustering. The silhouette index for a single point i is denoted by $s(i)$ and it is calculated as shown in the formula below.

$$s_{(i)} = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Silhouette Index Formula

where $a(i)$ is the average distance between i and all of the other points in its own cluster, $b(i)$ is the distance from i to its next nearest cluster centroid [7].

6.2 Davies Bouldin Index

Davies Bouldin Index (DBI) is the metric of the average similarity measure of each cluster with the cluster most similar to it. In this context, similarity is determined by the ratio between inter cluster and intra cluster distances. The DBI ranges from 0 to infinity, where a lower score indicates better clustering as they indicate more compact and well separated clusters.

$$DB = \frac{1}{N} \sum_{i=1}^N \max_{i \neq j} \left(\frac{S_i + S_j}{d(C_i, C_j)} \right)$$

where S_i is the standard deviation of class " i ", $d(C_i, C_j)$ is the Euclidean distance between the centroids.

The evaluation metrics used to evaluate the effectiveness of algorithms are the Silhouette index and the DBI, where a greater Silhouette index and a lesser DBI indicates an algorithm to be more effective in clustering data.

7 Results and Conclusions

On comparing the clusters formed by each algorithm, it is understood that the RM K Means algorithm stood out from the K Means algorithm in terms of the reduced number of iterations and a slight increase in silhouette score. While the K Means algorithm gives a silhouette index of 0.42, the RM K means gives a score of 0.48 which is higher there signifies better clustering of the data.

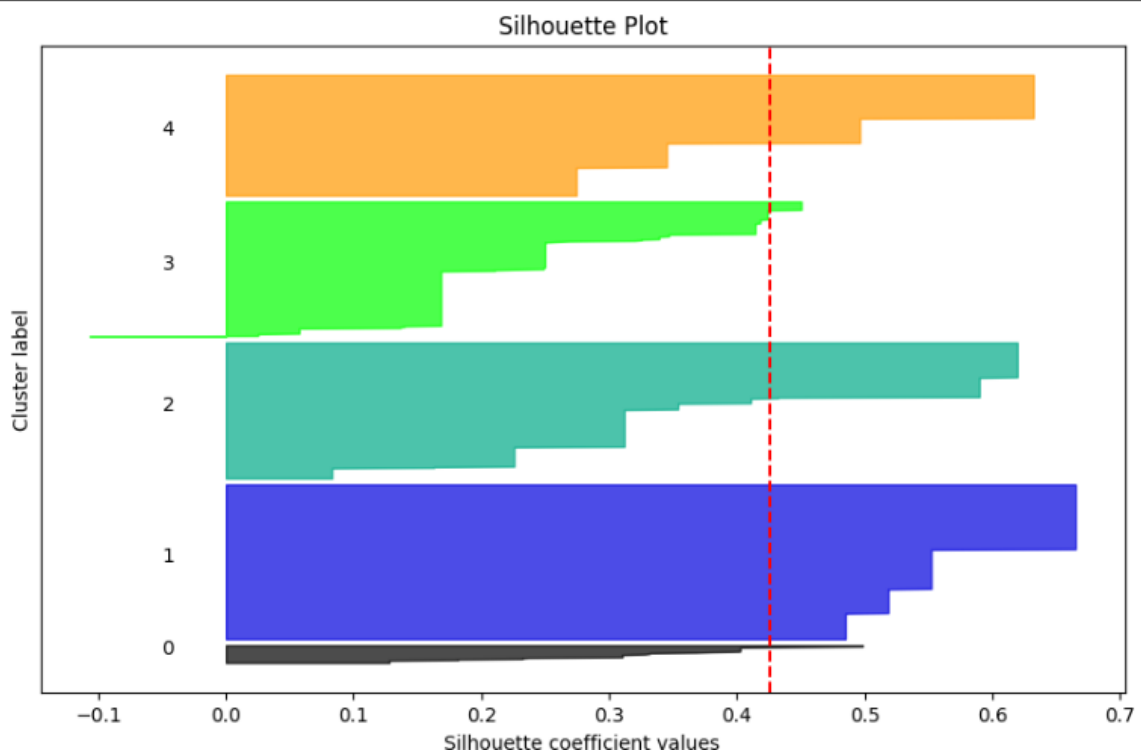


FIGURE 6

Silhouette plot of K Means clustering

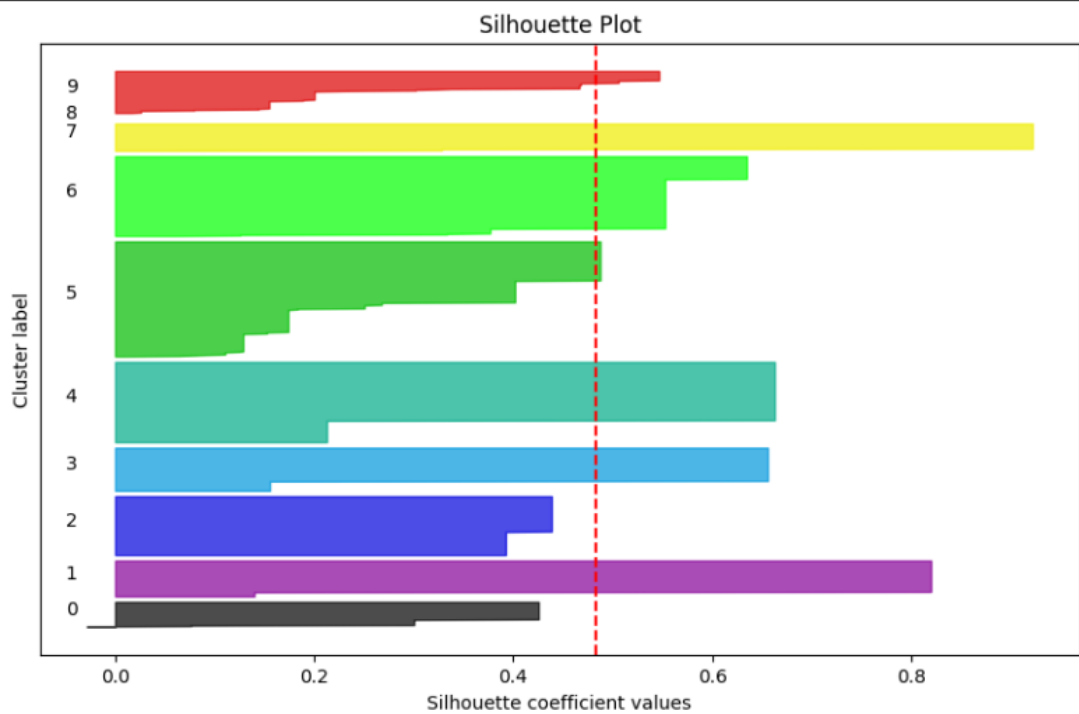


FIGURE 7

Silhouette plot of RM K Means clustering

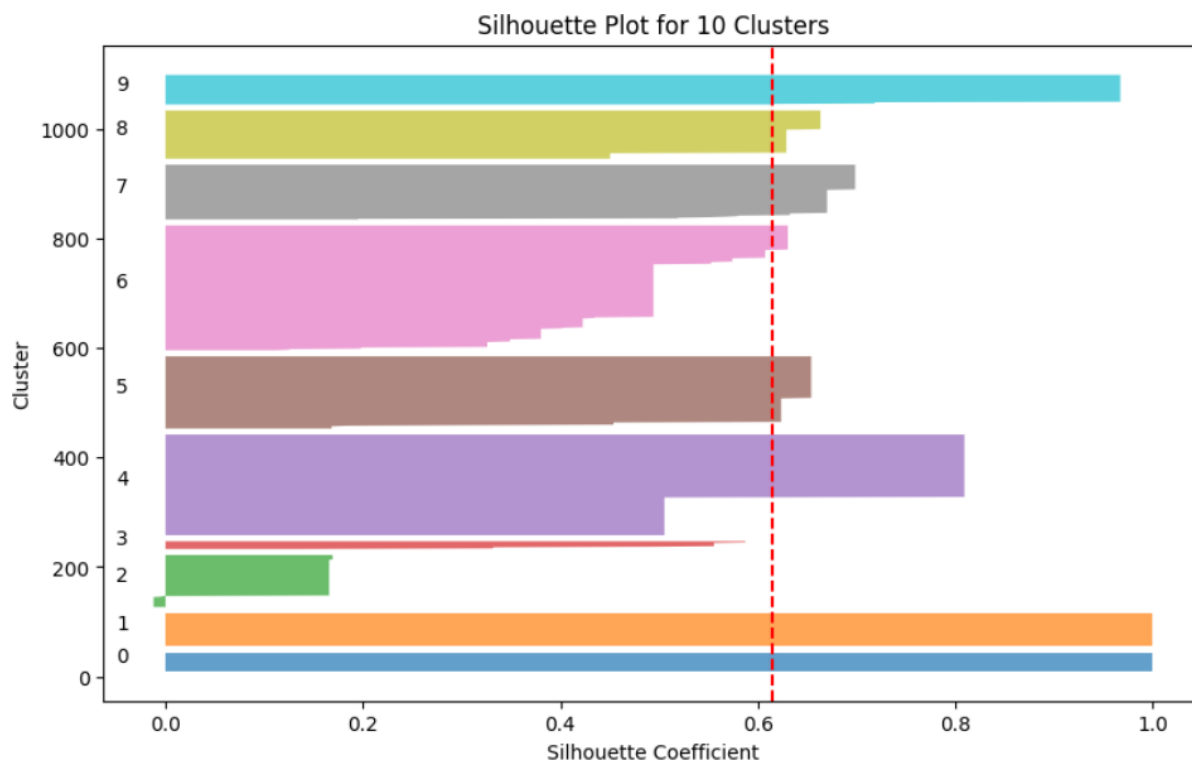


FIGURE 8

Silhouette plot of Box-Cox transformed RM K Means clustering

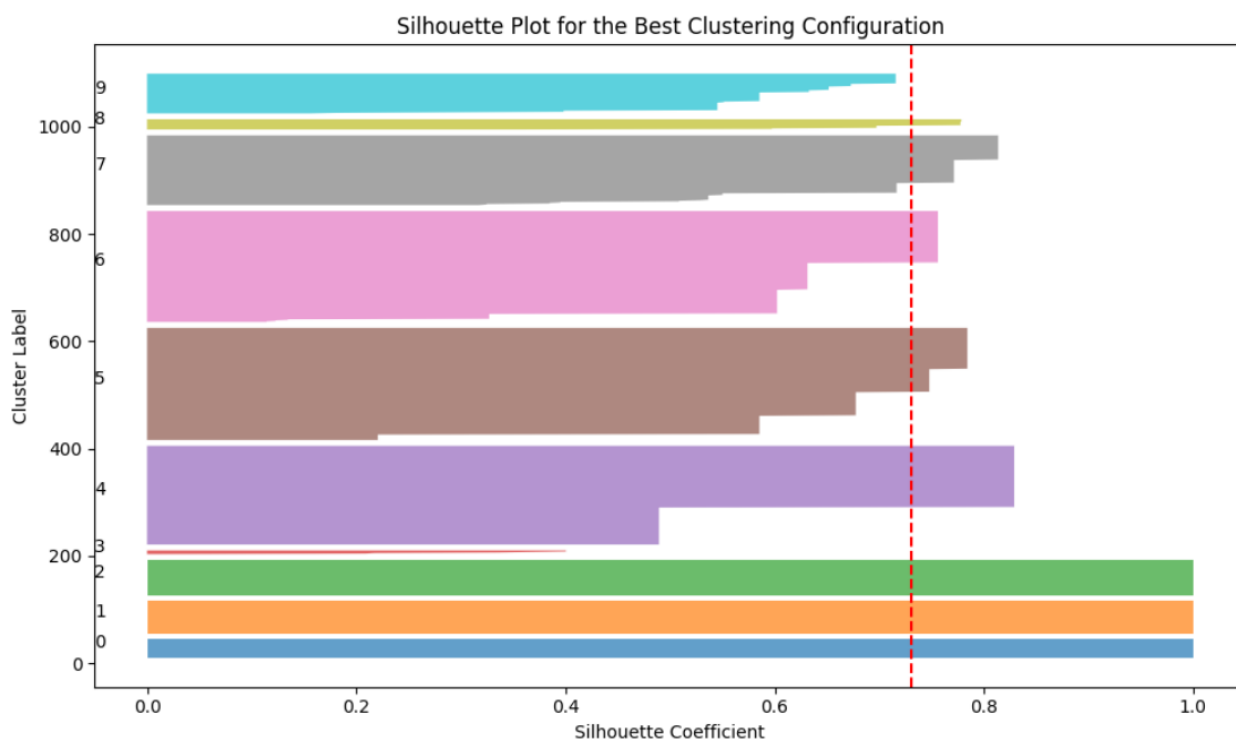


FIGURE 9

Silhouette plot of data tuned RM K Means clustering

The Box-Cox transformed RM K means algorithm gives a silhouette index of 0.61 signifying that acknowledging the skewness of the dataset and simply transforming all three attributes using the Box-Cox transformation can significantly boost the silhouette score of the clustering algorithm. However, the data tuned RM K-Means gives a staggering result of 0.71 as the silhouette index implying that for best results, one must address the needs of each attribute individually and perform the most apt transformation.

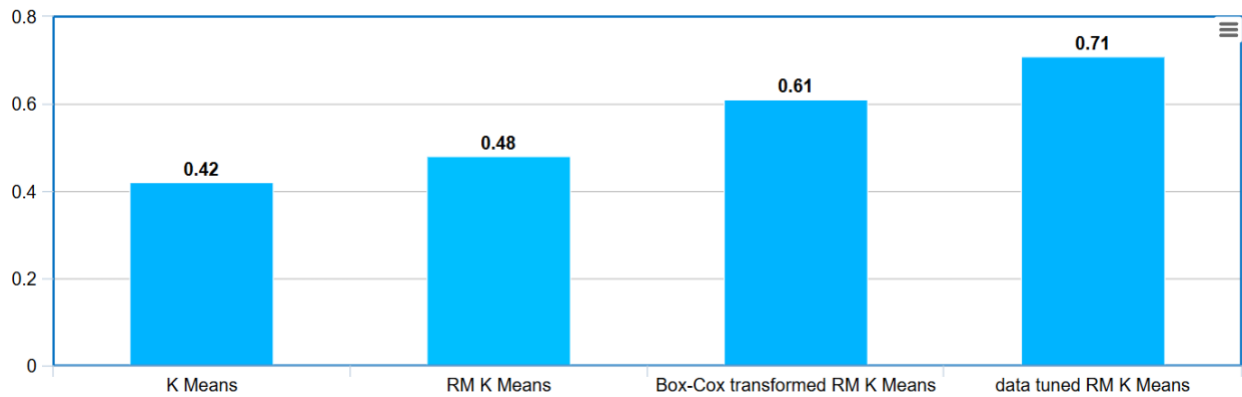


FIGURE 10

Average Silhouette width

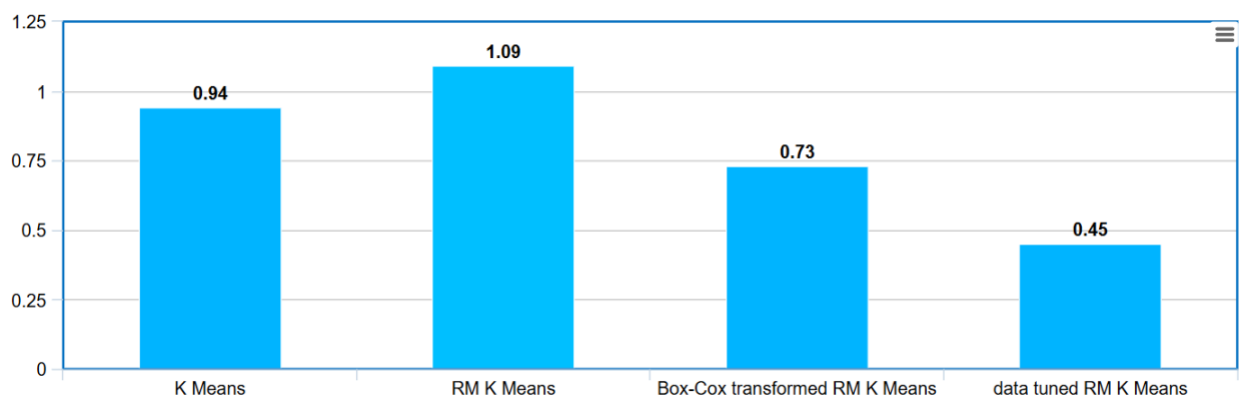


FIGURE 11

Average Davies Bouldin width

Moreover, the regular K-means gives a Davies Bouldin Index (DBI) of 0.94, the RM K means gives a DBI of 1.09, Box-Cox transformed RM K-means algorithms gives a DBI of 0.73.

The best DBI was also given by the data tuned RM K means algorithm. It gives 0.455 which denotes a credible advancement in the clustering of data as a lower Davies Bouldin score indicates better clustering.

Overall, we can conclude that the Data Tuned RM K means clustering algorithm performs the best clustering in comparison to K Means, RM K means and Box-Cox transformed RM K means clustering algorithm.

References

- [1] A. Joy Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa. RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences*, 33(10):1251–1257, 2021. DOI:<https://doi.org/10.1016/j.jksuci.2018.09.004>
- [2] Tushar Kansal, Suraj Bahuguna, Vishal Singh, and Tanupriya Choudhury. Customer Segmentation using K-means Clustering. In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pages 135–139, 2018. DOI:10.1109/CTEMS.2018.8769171
- [3] Morteza Namvar, Mohammad R. Gholamian, and Sahand KhakAbi. A Two Phase Clustering Method for Intelligent Customer Segmentation. In *2010 International Conference on Intelligent Systems, Modelling and Simulation*, pages 215–219, 2010. DOI:10.1109/ISMS. 2010.48
- [4] Mediana Aryuni, Evaristus Didik Madyatmadja, and Eka Miranda. Customer Segmentation in XYZ Bank Using K-Means and K-Medoids Clustering. In *2018 International Conference on Information Management and Technology (ICIMTech)*, pages 412–416, 2018. DOI:10. 1109/ICIMTech.2018.8528086
- [5] Amit Kumar. *Customer Segmentation of Shopping Mall Users Using K-Means Clustering*. 12 2022. DOI:10.4018/978-1-6684-5727-6.ch013
- [6] Sarahm, Customer Segmentation Using RFM Analysis, Version 1, 2023. <https://www.kaggle.com/code/sarahm/customer-segmentation-using-rfm-analysis/input>
- [7] Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy*, 23(6), 2021. DOI:10.3390/e23060759
- [8] Meysam Abbasi, Farahnaz Ahang, Hassan Ghaffari, Mohamad Mehdi, and Abdolmajid Imani. Customer Segmentation to Identify Key Customers Based on RFM Model by Using Data Mining Techniques. 11:62–76, 01 2022. DOI:10.22105/riej.2021.291738.122
- [9] Paul von Hippel. Skewness. In Miodrag Lovric, editor, *International Encyclopedia of Statistical Science*, pages 1340–1342, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. DOI:10.1007/978-3-642-04898-2_525
- [10] Andrew Leung. Chapter Twenty-One - Maximum likelihood estimation, Actuarial

Principles, 2022. DOI:<https://doi.org/10.1016/B978-0-32-390172-7.00025-7>

Vaibhav VINIL

design proj 40.docx

 APPLICATIONS OF TURBOCHARGERS IN THE MARINE INDUSTRY

 PS I - Summer term 2023-24

 Birla Institute of Technology & Science Pilani

Document Details

Submission ID**trn:oid::1:3121040807****Submission Date****Dec 22, 2024, 10:43 PM GMT+4****Download Date****Dec 22, 2024, 10:44 PM GMT+4****File Name****design_proj_40.docx****File Size****544.1 KB****22 Pages****4,468 Words****24,532 Characters**





14% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

-  **47 Not Cited or Quoted 13%**
Matches with neither in-text citation nor quotation marks
-  **4 Missing Quotations 2%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 8%  Internet sources
- 10%  Publications
- 8%  Submitted works (Student Papers)