SCIENCE AND INFORMATION TECHNOLOGY
SINTECH
JOURNAL

SINTECH JOURNAL | ISSN 2598-7305 | E-ISSN 2598-9642
Vol. 6 No 3 – Desember 2023 | https://s.id/sintechjournal
DOI : https://doi.org/10.31598
Publishing : Prahasta Publisher

Recognizing Hotel Visitors Preferences Based on Service Consumption Level Using K-Means Method

Ni Wayan Sumartini Saraswati[1], I Kadek Agus Bisena[2], I Dewa Made Krishna Muku[3], Gede Gana Eka Krisna[4]

[1,2,3,4] Department of Informatic Engineering, Faculty of Technology and Informatics, Institut Bisnis dan Teknologi Indonesia
Jl. Tukad Pakerisan No.97, Panjer, Denpasar Selatan, Denpasar, Bali, Indonesia

e-mail: sumartini.saraswati@gmail.com[1], agus.bisena@instiki.ac.id[2], dewamuku@instiki.ac.id[3], ganae93@gmail.com[4]

Received : November, 2023	Accepted : December, 2023	Published : December, 2023

Abstract

Consumer segmentation remains a pertinent and valuable area of study today, offering substantial benefits such as enhanced marketing cost efficiency and improved effectiveness in customer retention efforts. Based on the level of hotel service consumption, this research identifies consumer clusters based on hotel consumer preferences using Adiwana Unagi Suites visitor data and the K-Means Clustering method to determine customer segments based on the level and type of service consumption. The clusters formed will be identified based on their characteristics. In this way, hotel management can target certain types of service promotions better and more precisely. The ideal cluster obtained is 4 clusters based on silhouette analysis, namely a silhouette score with an average of 0.339 and a silhouette coefficient plot that is the most balanced compared to other cluster plots. The results of the four clusters can describe customer preferences, namely the Silver cluster (cluster 0) with the characteristics of high room consumption, while the other clusters have low consumption, the Gold cluster (cluster 1) with the characteristics of low room consumption while the other clusters have high consumption, the Bronze cluster (cluster 2) with the lowest consumption characteristics, and the Diamond cluster (cluster 3) with the highest consumption characteristics.

Keywords: *hotel, k-means, clustering, silhouette score*

Abstrak

Segmentasi konsumen tetap menjadi bidang studi yang relevan dan berharga saat ini, karena menawarkan manfaat besar seperti peningkatan efisiensi biaya pemasaran dan peningkatan efektivitas dalam upaya retensi pelanggan. Berdasarkan tingkat konsumsi layanan hotel, penelitian ini mengidentifikasi cluster konsumen berdasarkan preferensi konsumen hotel dengan menggunakan data pengunjung Adiwana Unagi Suites dan metode K-Means Clustering untuk menentukan segmen pelanggan berdasarkan tingkat dan jenis konsumsi layanan. Cluster yang terbentuk akan diidentifikasi berdasarkan karakteristiknya. Dengan cara ini, manajemen hotel dapat menargetkan jenis promosi layanan tertentu dengan lebih baik dan tepat. Cluster ideal yang didapat adalah 4 cluster berdasarkan analisis silhouette, yaitu silhouette score dengan rata-rata 0,339 dan plot silhouette coefficient paling seimbang dibandingkan plot cluster lainnya. Hasil dari semua cluster dapat menggambarkan preferensi pelanggan

yaitu cluster Silver (cluster 0) dengan karakteristik konsumsi kamar tinggi, sedangkan cluster lainnya memiliki konsumsi rendah, cluster Gold (cluster 1) dengan karakteristik konsumsi kamar rendah sedangkan klaster lainnya memiliki konsumsi tinggi, klaster Bronze (klaster 2) dengan karakteristik konsumsi terendah, dan klaster Diamond (klaster 3) dengan karakteristik konsumsi tertinggi.

Kata Kunci: *hotel, k-means, clustering, silhouette score*

1. INTRODUCTION

Bali is a very popular tourist destination in the world, thus enabling the Indonesian tourism industry to grow rapidly. The advancement of the tourism industry is inseparable from the hospitality business progress. In line with the development of hotels in Bali, many new competitors are starting to appear in the hotel business, so existing hotels will compete to acquire consumers according to their respective classes. Hotel business actors need competitive intelligence to be able to continue to improve services in their businesses so as to win the competition.

Adiwana Unagi Suites is one of the 4-star hotels located in the Ubud area, Gianyar Regency, which applies the concept of traditional and modern combination. Tourists, who are the primary target market for this hotel, are foreign tourists with different service interests. Tourists staying at Adiwana Unagi Suites consume several types of services, such as room consumption, F&B consumption, spa consumption, and other consumption.

Utilizing analysis of hotel consumer data will assist hotel management in determining the right steps to take in making business decisions, especially in conducting promotions. If the hotel obtains information on the characteristics of tourists staying overnight based on the consumption level carried out. In that case, it will become the basis for more effective and efficient customer retention maintenance activities. It can be done by grouping consumers based on consumption levels using the clustering method. K-Means is a good clustering method, as shown by the following studies.

This research also proposed the Silhouette coefficients method to determine the right number of clusters as an advanced analysis when the elbow method has not found the right number of clusters from the existing data.

Research [1] used the K-Means method to cluster tourist data visiting ASEAN to find Indonesia's position in grouping tourist visits. The resulting clustering was K = 3, consisting of C1 (a high visitor cluster),C2 (moderate/medium visitor cluster), and C3 (low visitor cluster). The research results showed that Indonesia was included in cluster C2. This clustering could be an effective way to create a more mature strategy for the government to increase tourism in Indonesia. In addition, [2] also used K-Means to identify potential tourists and provide benefits for Bendesa Hotel. The resulting accuracy reached 84.4% with four clusters where the grouping used attributes, such as gender, profession, age, and continent that could describe the character of tourists. Research conducted by [3] used K-Means for Mount Emei tourism clustering using public opinion that could be an influence on potential tourists and produces five clusters. Furthermore, [4] analyzed tourist profiles in Turkey using K-Means. This research divided it into several clusters based on their characteristics. There were three attributes used in describing clusters: frequency of tourist vacations, time of booking and holidays, and age. The results showed that these three attributes were very important parameter characteristics in analyzing tourist profiles. Research conducted by [5] used the K-means method to analyze the influence of dominant visitors based on the characteristics of the variables that become the mapping of entrance tickets to increase tourism potential in Madura. This was research divided into three clusters, namely high cluster (C1), medium cluster (C2), and low cluster (C3), based on several characteristics, such as gender, age, occupation, education, and marital status. The result could be that the marital status characteristics have an accuracy of up to 0.81. Another research conducted by [6] made a recommendation system for three points: rating predictions on reviews, feedback models, and knowledge-based recommendation systems. It was done as a recommendation and promotion system that

could automatically analyze reviews in the USA. K-means was used to improve the quality of the recommendation system and was included in topic modeling, which aimed to identify a series of topics and terms that underlie these topics. [7] implemented K-means in looking for patterns in the causes of tourist satisfaction at Madani hotels in terms of service, facilities, comfort, and price. This research aimed to increase the number of hotel visitors. The results showed a satisfaction score of 11 for comfort, 14 for service, 13 for facilities, and 12 for price. The research conducted by [8] used the Naïve Bayes method for classification and K-Means for sentiment clustering in hotel reviews. K-means was used to group text and get information related to topics from content reviews in each group. This research resulted in three clusters, where each cluster had a word frequency often mentioned and could take the main discussion orientation. The implementation of K-Means in other tourism groupings was carried out by [9] using ten tourism objects in Pagar Alam City and producing four clusters, namely C1 (high cluster), C2 (moderate cluster), C3 (low cluster), and C4 (very low clusters). This research could provide information for a decision to increase the number of visitors.

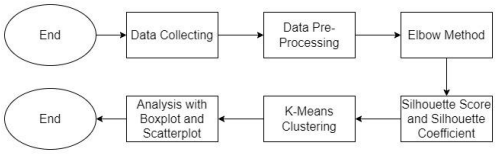
Research conducted by [10] compared K-Means and DBSCAN in grouping tourism in Madura, which showed that the K-Means method was superior with a Silhouette Index (SI) value of 0.6902. Research [11] also used K-Means for sentiment and descriptive analysis to find the right patterns and methods. The analysis used was Vader sentiment and SentiWordNet for hotel reviews. This research resulted in SSE reaching a 4.611% error. Another study conducted by [12] used the K-Means-based Ordered Weighted Average (OWA) method in clustering customers staying at hotels. This method performed clustering on multi-criteria data. The results showed that this method improved performance by 21.6% and reduced the number of convergent iterations by 48.46%.

K-Means implementation combined with PSO and ANN in data analysis for customer segmentation in Saudi Arabia showed that the resulting clusters reached 6, and the MSE value reached 0.09847. In research conducted by [13], K-Means gave higher results with the SVR-SMO method to carry out big data analysis on tourist decision-making about hotels in Mecca with segmentation expressing tourist satisfaction. Research conducted by [14] showed that K-Means produced an accuracy of up to 96.21% on dynamic price analysis, which could help Airbnb choose optimal property prices based on the selected season and aimed to provide an overview of strategies to adjust prices as optimally as possible so that it could be a strategy to compete with competitors.

Based on this, previous studies show that K-Means provides the best results. However, there is no research that can find the number of clusters in hotels that can describe consumer preferences based on service consumption such as room consumption, F&B consumption, spa consumption, and other consumption, so this research uses K-Means to find the ideal number of clusters in Adiwana reservation data Unagi Suites to reflect consumer preferences. The K-Means method will also be assisted using silhouette analysis and the elbow method in determining the most optimal number of clusters.

2. RESEARCH METHOD

This research has several stages which are shown in the flow diagram below.



Picture 1. Research Flowchart

This research was started by acquiring data obtained from Adiwana Unagi Suites. The data used was the daily data of Adiwana Unagi Suites hotel visitors in Excel form, with a total of 154 data and 13 attributes, including Guest Name, Country of Origin, Travel Type, Booking Agent, Check-In Date, Check-Out Date, Length of Stay, Number of guests per booking, Room Type, Room Consumption, F&B Consumption, Spa Consumption, and Other Consumption.

The second stage is the pre-processing process. Data pre-processing is carried out to process data according to the format required by the clustering model. In other words, data that is not needed will be discarded, or data that does not match the format will be changed for the

clustering process, such as handling missing values and data normalization.

When there is a case where one attribute has a value range that is very far from the value range of another attribute, this will affect the analysis results. Therefore, pre-processing is needed to normalize the values for these attributes. This research normalizes these attributes using standardScaler from Scikit-Learn. StandardScaler is formulated as follows [15].

$$StandardScaler=\frac{x_i\text{-}mean(x)}{std(x)} \quad (1)$$

After the pre-processing stage was complete, the data was run using the elbow method to determine the optimal number of clusters. The elbow method is used to find optimal cluster values for K-Means or other unsupervised learning methods. This method, which is also known as "knee of a curve", can also produce a smooth curve, so it will be difficult to determine the K value because the elbow points are difficult to distinguish [16]. Therefore, this research also proposes silhouette analysis. The cluster recommendation from the Silhouette score or silhouette coefficient method was used as a support for the elbow method to produce the optimal number of clusters. This silhouette analysis can be described in the following formula.

$$S_i=\frac{b_i\text{-}a_i}{max(b_i,a_i)} \quad (2)$$

Where a is the average distance between point one and all points in the same cluster, while b is the average distance between point one and all points in the nearest cluster [17].

In this research, the K-Means clustering method was applied as a method for grouping hotel visitors based on the service consumption level. K-Means is a partition-based unsupervised learning method. K-Means is based on distance or centroid (cluster center) [18]. K-Means minimizes the Sum Squared of Error (SSE) value between data objects and K centroids. There are two ways to calculate distance in K-Means, namely using Euclidean Distance which is shown in formula 3 and City Block (Manhantan Distance) which is shown in formula 4 [19].

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i\text{-}y_i)^2} \quad (3)$$

$$d(x,y)=\sum_{i=1}^{n}|x_i\text{-}y_1| \quad (4)$$

The results of the optimal number of clusters from the elbow method and silhouette analysis will be continued to the K-Means model process and produce clusters for each data. Because the cluster results are in array form. Then, the results will be converted into a form of data visualization. Visualizing this data will make analysis easier, so boxplots and scatterplots are used to understand the characteristics of the clusters formed.

3. RESULT AND DISCUSSION

After carrying out data acquisition, the data obtained entered data pre-processing, i.e., the data was processed from the raw data obtained to produce data that can be analyzed. The raw data that will go through data pre-processing is shown in Figure 2.

	Nama Tamu	Asal Negara	Travel Type	Booking Agent	Tanggal Check In	Tanggal Checkout	Lama Menginap	Jumlah tamu per booking	Tipe Kamar	Konsumsi atas Kamar	Konsumsi F&B	Konsumsi Spa	Konsumsi Lainnya
0	KEONTAY MARQUELLE ADAMS	United States	Leisure	Expedia Pay at Hotel	22-Aug-22	27-Aug-22	5.0	2.0	UNAGI SUITES	10756795.0	8404600.0	3600000.0	3600000.0
1	Agnes Lee	Australia	Leisure	WEBSITE	30-Aug-22	2-Sep-22	3.0	2.0	ONE BEDROOM POOL VILLA	14715000.0	6690090.0	1350000.0	1450000.0
2	Craig Shankar	Canada	Leisure	WEBSITE	24-Aug-22	31-Aug-22	7.0	2.0	ONE BEDROOM POOL VILLA	20071800.0	9680090.0	3500000.0	4500000.0
3	Carlos Corral	Spain	Leisure	WEBSITE	23-Aug-22	29-Aug-22	6.0	2.0	ONE BEDROOM POOL VILLA	15244999.0	9432900.0	1300000.0	1550000.0
4	Makaoui Stephane	France	Leisure	WEBSITE	21-Aug-22	26-Aug-22	5.0	2.0	UNAGI SUITES	10887500.0	4560300.0	4000000.0	3500000.0

Picture 2. Raw Data

In this research, the data pre-processing stage was carried out by dividing the data into cluster variables: room consumption, F&B consumption, spa consumption, and other consumption.

```
X = df.iloc[:,9:13].values
```

Picture 3. Data Pre-Processing Code

Using the line of code above, the selected variable from the “df” data frame, namely columns 9 to 12, will be stored in a new data frame, namely “X”. After dividing the data, the new data still contained non-numeric data, so it could not carry out the clustering process. Furthermore, data cleaning was carried out by changing the empty value “NaN” to 0 so that the clustering process could be carried out to produce data like the following, as shown in Table 1.

Table 1: Example of Top 5 Rows of Data Pre-Processing Results

	Room	F&B	Spa	Others
0	10756795.0	840450.0	350000.0	350000.0
1	14715000.0	659009.0	135000.0	145000.0
2	20071800.0	568009.0	350000.0	450000.0
3	15244999.0	543290.0	130000.0	155000.0
4	10887500.0	456030.0	400080.0	350000.0

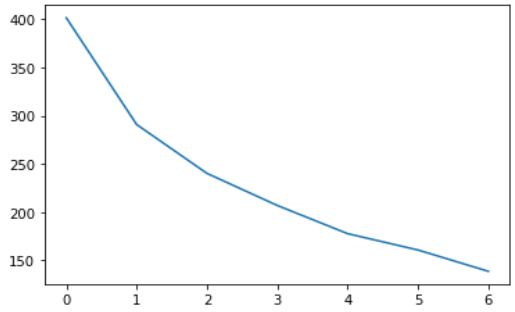
Standard Scaler is a class of sklearn that normalizes data so that the data used does not have large deviations. Standard Scaler is a preprocessing method that standardizes features by removing averages and scaling unit variances. Standard Scaler follows the Standard Normal Distribution (SND). Therefore, it makes mean = 0 and scales the data to unit variance. The Standard Scaler results from the data in this research are shown in Picture 4.

```
array([[ 1.12872023,  4.69459965,  3.16533487,  4.7482751 ],
       [ 2.04884515,  3.5186754 ,  0.8508793 ,  1.63090567],
       [ 3.29408767,  2.92890182,  3.16533487,  6.26894311],
       [ 2.1720488 ,  2.76869729,  0.79705475,  1.78297247],
       [ 1.15910394,  2.20316276,  3.70444155,  4.7482751 ],
       [ 0.754521  ,  2.20299425,  0.31263382, -0.57406295],
       [ 1.13474676,  2.11249318,  3.27298397,  0.33833786],
       [ 0.59480693,  2.11245429,  0.09733563,  0.26230446],
       [ 3.37317061,  2.06691858,  0.79705475, -0.57406295],
       [-0.27223478,  1.98293872,  2.85315249,  1.25073867],
       [ 3.2959706 ,  1.85103553,  2.84238758,  1.25073867],
       [ 0.42593765,  1.84284935, -0.6023835 , -0.57406295],
       [-0.91850863,  1.80240384,  0.79705475, -0.57406295],
       [ 0.60596502,  1.75512928,  0.79705475,  0.71850486],
       [-0.33291618,  1.74293198,  0.25911069,  0.03420426],
       [ 1.51534911,  1.63268969,  4.24182584,  1.25073867],
       [ 0.97139237,  1.61324661,  0.79705475,  0.11023766],
       [ 1.19106721,  1.54196178, -0.6023835 ,  0.33833786],
       [-0.15978382,  1.51610895,  1.71207207, -0.57406295],
```

Picture 4. Standard Scaler Normalization Result

In general, the research could use the elbow method graph to find the ideal number of clusters. In cluster analysis, the elbow method was a heuristic used to determine the number of clusters in a data set. The method consisted of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.

The elbow method graph was built from the data in this research, as shown in Figure 3. This research used the number of clusters where the graph changed drastically from horizontal to vertical from the previous one, which is called the center of the elbow. From Figure 3, this research has difficulty determining the center point of the elbow, so it used further analysis using the silhouette score.



Picture 5. Elbow Method Chart

Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in neighboring clusters and thus provides a way to assess parameters like the number of clusters visually. This measure has a range of [-1, 1].

This research used silhouette score analysis to determine the right number of clusters for the data. The research calculated the average value of the silhouette score for the number of clusters from 2 to 6. The results are as follows.

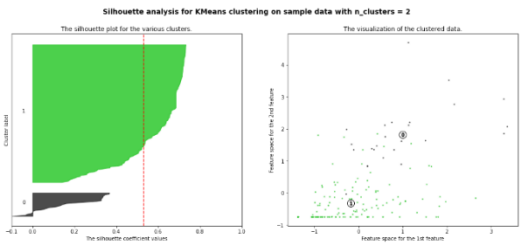
a. For n_clusters = 2 The average silhouette_score is : 0.5302394463526667
b. For n_clusters = 3 The average silhouette_score is : 0.36414785226070606
c. For n_clusters = 4 The average silhouette_score is : 0.3390753942363806
d. For n_clusters = 5 The average silhouette_score is: 0.33717866859937307
e. For n_clusters = 6 The average silhouette_score is : 0.34038627373953373

Thus far, this research obtained the number of clusters = 2, which gave the highest average silhouette value. Therefore, the number of clusters may be the right number of clusters. But furthermore, this research will analyze the respective silhouette coefficients shown in Picture 6 to 10 to provide a further picture of the best clusters.

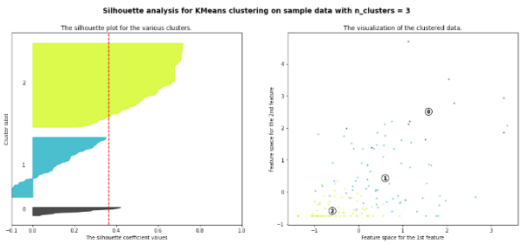
Silhouette coefficients are a cluster evaluation method that combines cohesion, and separation methods. Cohesion is measured by counting all objects contained in a cluster and separation is measured by calculating the average distance of each object in a cluster to the nearest cluster. The distance between data was calculated using the Euclidean distance formula [20].

Silhouette coefficients (as these values are referred to) near +1 indicated that the sample was far from the neighboring clusters. A value of 0 indicated that the sample was on or very close to the decision boundary between two neighboring clusters, and negative values indicated that those samples might have been assigned to the wrong cluster.

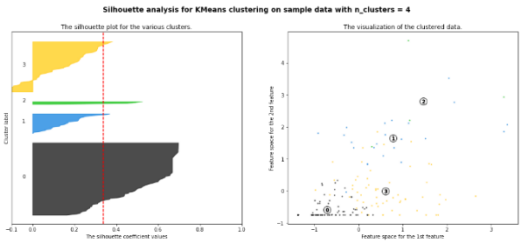
From the analysis of these figures, all graphs showed that there were samples that might be grouped in the wrong cluster, characterized by negative silhouette scores. The number of clusters = 4 illustrated the best choice for the given data due to the presence of all clusters with above-average silhouette scores, and also, from the thickness of the silhouette plot, the cluster size could be visualized as relatively more balanced.



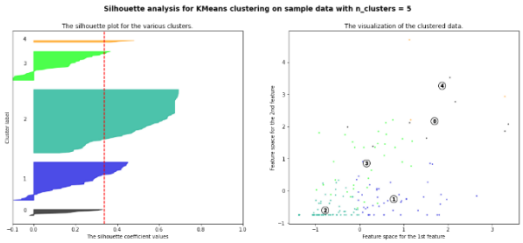
Picture 6. Silhouette Analysis with Two Cluster



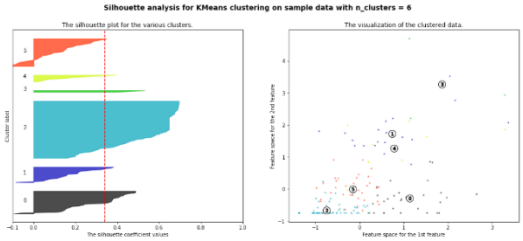
Picture 7. Silhouette Analysis with Three Cluster



Picture 8. Silhouette Analysis with Four Cluster



Picture 9. Silhouette Analysis with Five Cluster



Picture 10. Silhouette Analysis with Six Cluster

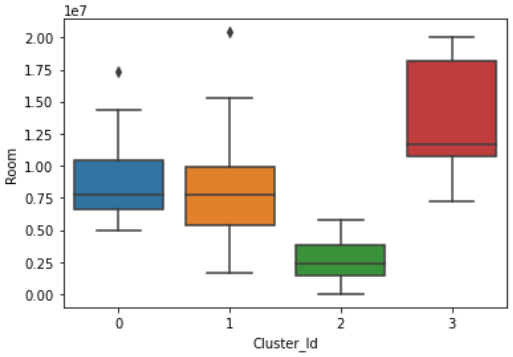
3.1 Clustering Result with Four Cluster

The research visualized the clustering results in a boxplot from Picture 11 to Picture 14 to illustrate each cluster's character analysis based on each service's consumption level. The summary of the analysis with our ranking is compiled in Table 2. From table 2, customer preferences can be described. In this research, cluster 3 is said to be the Diamond cluster or the cluster with the highest rank with the highest service consumption for rooms, FNB, spa and others compared to other clusters. Customer preference for cluster 1 is called gold even though the room consumption is smaller than cluster 0, but other consumption such as fnb, spa, and other have higher consumption compared to clusters 0 and 2. Meanwhile in cluster 0 the customer preference is said to be silver even though the room consumption is higher compared to cluster 1, but the consumption of the other three services is smaller. For cluster 3, which is called the bronze cluster, all service consumption is the lowest compared to other clusters. So, from this cluster it can be found how customer preferences can help hotels to increase sales from consumers.

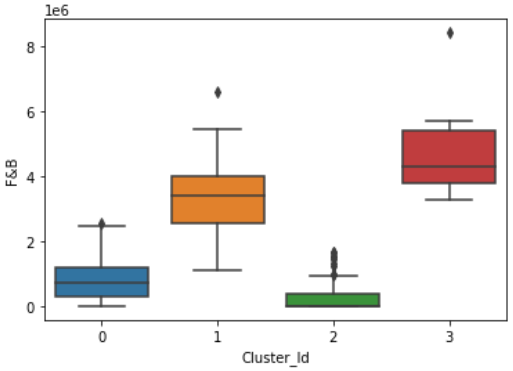
Table 2: Summary of Cluster Analysis of Boxplots

Cluster	Rooms Consumption Rank	F&B Consumption Rank	Spa Consumption Rank	Others Consumption Rank	Cluster Name
0	2	3	3	3	Silver

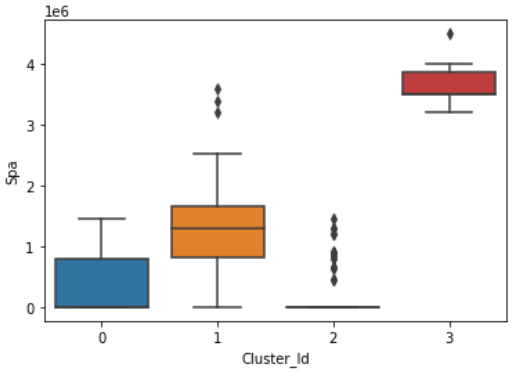
1	3	2	2	2	Gold
2	4	4	4	4	Bronze
3	1	1	1	1	Diamond



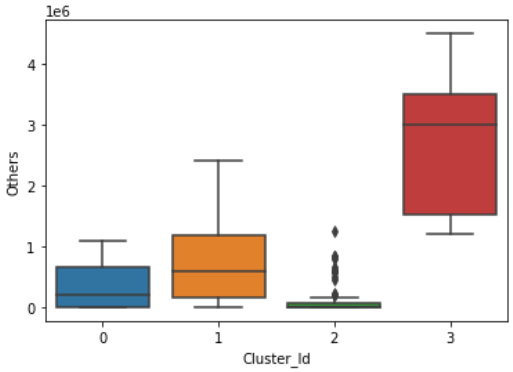
Picture 11. Boxplot for Variable 'Room'



Picture 12. Boxplot for Variable 'F&B'



Picture 13. Boxplot for Variable 'Spa'



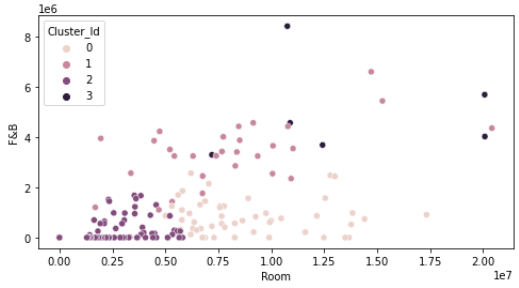
Picture 14. Boxplot for Variable 'Others'

To further analyze the consumer consumption characteristics of each cluster, this research used paired scatterplots for the combination of room, F&B, and Spa services shown in Figure 15 to Figure 17.

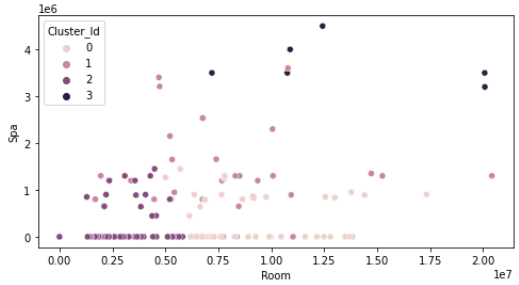
This research analyzed the graph and obtained the summary, as shown in Table 3. The unique thing was that from the division of 4 clusters, which obtained that cluster 0 had the characteristics of guests whose room consumption was good but other services was low. From the results shown, the boxplot and scatterplot are able to describe customer preferences at the Adiwana Unagi Suites hotel based on consumption of these services.

Table 3: Summary of Cluster Analysis of Scatterplot

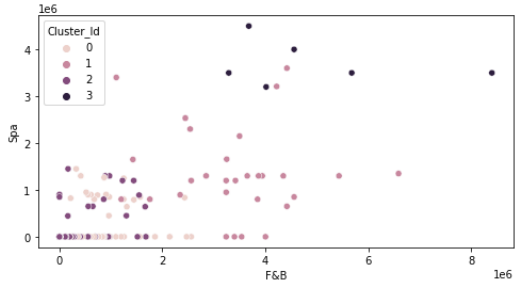
Cluster	Rooms and F&B	Rooms and Spa	F&B and Spa
0	F&B Low, Room Medium	Spa Low, Room Medium	Both Medium to Low
1	Both Medium	Both Medium	Both Medium
2	Both Low	Both Low	Both Low
3	Both High	Both High	Both High



Picture 15. Scatterplot for "Room" and "F&B" Variables



Picture 16. Scatterplot for "Room" and "Spa" Variables



Picture 17. Scatterplot for "F&B" and "Spa" Variables

4. CONCLUSION

Based on the analysis of the silhouette score and silhouette coefficients, it was found that the right number of clusters for our data was 4 clusters. Through the analysis of boxplot and scatterplot, it was defined that the cluster with the highest service consumption in all types was cluster 3, which was named the diamond cluster. The medium consumption in all services was cluster 1, named the gold cluster. While the cluster 0 had consumption for medium rooms, but consumption for other services was low, which was named the silver cluster. Lastly, consumption for all types of services was low in cluster 3, named the bronze cluster. Therefore, in this research, K-Means can help provide the optimal number of clusters and an overview of consumer preferences supported by the elbow method and silhouette analysis.

STATEMENT OF APPRECIATION

Our appreciation goes to DRPM INSTIKI for its support for this research, both material and non-material support. Hopefully this research will be useful for further research.

REFERENCES

[1] G. M. Kario and E. Amalia, "K-Means Algorithm Implementation for Clustering of Foreign Tourists Visiting," *Int. J. Open Inf. Technol.*, vol. 9, no. 6, pp. 20–27, May 2021, Accessed: Aug. 04, 2023. [Online]. Available: http://injoit.org/index.php/j1/article/view/1101

[2] I. Gede, K. K. Putra, W. Surya Dharma, G. Karang, and K. Putra, "Application of The K-Means Clustering Method To Search For Potential Tourists of Bendesa Hotel," *TIERS Inf. Technol. J.*, vol. 4, no. 1, pp. 8–15, Jun. 2023, doi: 10.38043/TIERS.V4I1.4297.

[3] Z. Li, L. Jia, and B. Su, "Improved K-Means Algorithm for Finding Public Opinion of Mount Emei Tourism," *Proc. - 2019 15th Int. Conf. Comput. Intell. Secur. CIS 2019*, pp. 192–196, Dec. 2019, doi: 10.1109/CIS.2019.00048.

[4] M. E. Yildirim, M. Kaya, and I. Furkanince, "A Case Study: Unsupervised Approach for Tourist Profile Analysis by K-means Clustering in Turkey," *J. Internet Comput. Serv.*, vol. 23, no. 1, pp. 11–17, 2022, doi: 10.7472/JKSII.2022.23.1.11.

[5] A. Jauhari, D. R. Anamisa, and F. A. Mufarroha, "Analysis of Clusters Number Effect Based on K-Means Method for Tourist Attractions Segmentation," *J. Phys. Conf. Ser.*, vol. 2406, no. 1, p. 012024, Dec. 2022, doi: 10.1088/1742-6596/2406/1/012024.

[6] R. K. Mishra, J. A. A. Jothi, S. Urolagin, and K. Irani, "Knowledge based topic retrieval for recommendations and tourism promotions," *Int. J. Inf. Manag. Data Insights*, vol. 3, no. 1, p. 100145, Apr. 2023, doi: 10.1016/J.JJIMEI.2022.100145.

[7] D. Masri, A. Apriyandi, and B. Harahap, "Implementation of K-Means for Analysis of Factors Causing Consumer Satisfaction at Madani Hotel Medan City," *Bull. Comput. Sci. Electr. Eng.*, vol.

3, no. 2, pp. 66–72, Dec. 2022, doi: 10.25008/BCSEE.V3I2.1162.

[8] R. N. Juliadi and Y. Puspitarani, "Supervised Model for Sentiment Analysis Based on Hotel Review Clusters using RapidMiner," *Sink. J. dan Penelit. Tek. Inform.*, vol. 7, no. 3, pp. 1059–1066, Aug. 2022, doi: 10.33395/SINKRON.V7I3.11564.

[9] D. Puspita and S. Sasmita, "APPLICATION OF K-MEANS ALGORITHM IN GROUPING OF CITY TOURISM CITY PAGAR ALAM," *Sink. J. dan Penelit. Tek. Inform.*, vol. 7, no. 1, pp. 28–32, Jan. 2022, doi: 10.33395/SINKRON.V7I1.11220.

[10] A. Jauhari, D. R. Anamisa, F. A. Mufarroha, and I. O. Suzanti, "Grouping Madura Tourism Objects with Comparison of Clustering Methods," *Proceeding - IEEE 8th Inf. Technol. Int. Semin. ITIS 2022*, pp. 119–123, 2022, doi: 10.1109/ITIS57155.2022.10009968.

[11] M. Beny Pangestu, A. Ridho Barakbah, and T. Hadiah Muliawati, "Data analytics for hotel reviews in multi-language based on factor aggregation of sentiment polarization," *IES 2020 - Int. Electron. Symp. Role Auton. Intell. Syst. Hum. Life Comf.*, pp. 324–331, Sep. 2020, doi: 10.1109/IES50839.2020.9231625.

[12] P. J. Pons-Vives, M. Morro-Ribot, C. Mulet-Forteza, and O. Valero, "An Application of Ordered Weighted Averaging Operators to Customer Classification in Hotels," *Math. 2022, Vol. 10, Page 1987*, vol. 10, no. 12, p. 1987, Jun. 2022, doi: 10.3390/MATH10121987.

[13] A. Alsayat, "Customer decision-making analysis based on big social data using machine learning: a case study of hotels in Mecca," *Neural Comput. Appl.*, vol. 35, no. 6, pp. 4701–4722, Feb. 2023, doi: 10.1007/S00521-022-07992-X/TABLES/5.

[14] Fitrianingsih, D. A. Rahayu, and F. R. Zazila, "Dynamic Pricing Analytic of Airbnb Amsterdam Using K-Means Clustering," *2022 7th Int. Conf. Informatics Comput. ICIC 2022*, 2022, doi: 10.1109/ICIC56845.2022.10006966.

[15] A. B. Adetunji, O. N. Akande, F. A. Ajala, O. Oyewo, Y. F. Akande, and G. Oluwadara, "House Price Prediction using Random Forest Machine Learning Technique," *Procedia Comput. Sci.*, vol. 199, pp. 806–813, Jan. 2022, doi: 10.1016/J.PROCS.2022.01.100.

[16] A. J. Onumanyi, D. N. Molokomme, S. J. Isaac, and A. M. Abu-Mahfouz, "AutoElbow: An Automatic Elbow Detection Method for Estimating the Number of Clusters in a Dataset," *Appl. Sci. 2022, Vol. 12, Page 7515*, vol. 12, no. 15, p. 7515, Jul. 2022, doi: 10.3390/APP12157515.

[17] L. George and P. Sumathy, "An integrated clustering and BERT framework for improved topic modeling," *Int. J. Inf. Technol.*, vol. 15, no. 4, pp. 2187–2195, Apr. 2023, doi: 10.1007/S41870-023-01268-W/FIGURES/6.

[18] P. Sharma, "The Ultimate Guide to K-Means Clustering Definition, Methods and Applications," *Analytics Vidhya*, 2019. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/#What_Is_K-Means_Clustering? (accessed Mar. 03, 2023).

[19] H. Bonthu, "Understanding KMeans Clustering. Introduction," *Analytics Vidhya*, 2021. https://www.analyticsvidhya.com/blog/2021/08/kmeans-clustering/

[20] S. Paembonan and H. Abduh, "Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat," *PENA Tek. J. Ilm. Ilmu-Ilmu Tek.*, vol. 6, no. 2, p. 48, 2021, doi: 10.51557/pt_jiit.v6i2.659.