Web Scrapping for Grouping Star Hotels in Yogyakarta Using Hierarchical Cluster Analysis

Devina Gilar Fitri A.S
Department of Statistics
Universitas Islam Indonesia
Yogyakarta, Indonesia
devinagilar56@gmail.com

R.B Fajriya Hakim
Department of Statistics
Universitas Islam Indonesia
Yogyakarta, Indonesia
hakimf@uii.ac.id

***Abstract*—Yogyakarta Special Region is the second smallest province after DKI Jakarta Province which is located in the middle of Java Island, tourism is one of its flagship sectors, this is evidenced by DIY as the second largest tourist destination in Indonesia after Bali. in 2017 there were 5,097,000 domestic and foreign tourists visiting DIY, of which the number of tourists who used hotel accommodation was 1,326,031 with a total of 512 hotels, where from year to year, hotel occupancy rates decreased. Hotels play an important role in the tourism industry because they provide lodging, food and beverage facilities and services as well as other services for the public who temporarily stay and are managed commercially. Cluster analysis is a statistical technique that is useful for grouping objects or variables into certain groups where each object has properties and characteristics that are close together. To help overcome the problem of the low number of tourists staying at star hotels in Yogyakarta, grouping of star hotels was carried out based on similar characteristics such as facilities, prices, customer ratings, and the number of reviews in order to obtain descriptive results of star hotel data in Yogyakarta and grouping hotels based on their similar characteristics. Based on the results of the analysis that has been carried out, it is obtained the results that the scrapping data of Yogyakarta star hotels are very good, there is no assumption deviation, and there is no blank data. Then for the grouping results obtained are as many as 3 clusters, with each cluster having members, 6 for the first cluster, 2 for the second cluster, and 38 for the third cluster. Where from the results of this study hope can be additional information for people who want to visit Yogyakarta, as well as for certain parties in an effort to weigh business.**

Hierarchy Cluster, Average Linkage Method, Cluster with R, Web Scrapping, Hotel Yogyakarta, Yogyakarta Tourists, Tourism

I. INTRODUCTION

Yogyakarta Special Region is the second smallest province after DKI Jakarta Province which is located in the middle of Java Island. The total area of DIY Province is 3,185.80 km2 (DPPKA, 2017) with the system of the Sultanate government, ethnic and cultural diversity, friendliness and harmony of the people, and the many amazing natural attractions make Yogyakarta a special area.

Tourism is one of the mainstay sectors, this is evidenced by DIY as the second largest tourist destination in Indonesia after Bali (Kemendagri, 2016). According to the head of the DIY tourism service in (tribunjogja, 2018), it is said that in 2017 there were 4,700,000 domestic tourists and 397,000 foreign tourists visiting DIY. Based on the number of tourists, according to the website Statisticshotel.visitingjogja.com, it is known that in 2017 the number of foreign tourists using hotel accommodation was 136,641 and domestic tourists were 1,189,390 with a total of 512 hotels, from which the largest percentage was 82.22 % is the choice of tourists to stay in no-star hotels. From year to year, the occupancy rate of hotels decreased, the Head of IHGMA Chapter Yogyakarta, Herryadi Baiin acknowledged this. The hotel business is currently a bit heavy because there are a lot of new players appearing (Sindonews, 2017).

Hotels play an important role in the tourism industry because they provide facilities and services, lodging, food and beverages, and other services for the public who temporarily live and are managed commercially (Sihite, 2000). Promotional strategies with new concepts have always been a breakthrough in getting more guests. They asked the attention of the local government to pay attention to the competitive climate of the hospitality business (sindonews, 2017).

Cluster analysis is a statistical technique that is useful for grouping objects or variables into certain groups where each object has properties and characteristics that are close together. Hierarchical clusters are usually used for relatively few data samples of less than 100. Hierarchical clusters attempt to classify objects based on the similarity of the variables they have because hierarchically the process is to compare each pair of cases for a small number of cases. (Sitepu, et al., 2011).

So, from the existing problems, to help overcome the problem of the low number of tourists staying at star hotels in Yogyakarta, the grouping of star hotels is based on similar characteristics such as facilities, prices, customer ratings, and the number of reviews. The title raised in this study is "Grouping Star Hotels in Yogyakarta Based on Ratings, Facilities, Prices, and Number of Reviews Using Hierarchical Cluster Analysis" with the aim of obtaining descriptive data from star hotels in Yogyakarta and grouping hotels based on similar characteristics, which are expected this research can be additional information for people who want to visit Yogyakarta, as well as for certain parties in an effort to weigh business.

ISBN 978-602-53524-0-9

II. PURPOSE

The formulation of the problem raised from the background above is:

1) To find the general description of star hotel data in Yogyakarta according to tripadvisor.com
2) To find out the group results on star hotels based on rating, facilities, prices, and number of reviews from the tripadvisor.com website using hierarchical cluster analysis

III. METHODOLOGY

In this study, the data used were secondary data, which was obtained from the results of scrapping the tripadvisor.com web for the star hotel category in Yogyakarta. The variable data obtained include hotel name, rating, facilities, price, and the number of reviews. The data that has been obtained is then analyzed using hierarchical clusters, with 5 agglomerative methods namely: Single Linkage, Average Linkage, Complete Linkage, Centroid, and Ward which of the five methods will be compared to choose the method that produces the best cluster.

The research flow used has been adjusted back from research (Puspitasari & Susanti, 2016) and (Novyantika, 2018) as in the following figure,

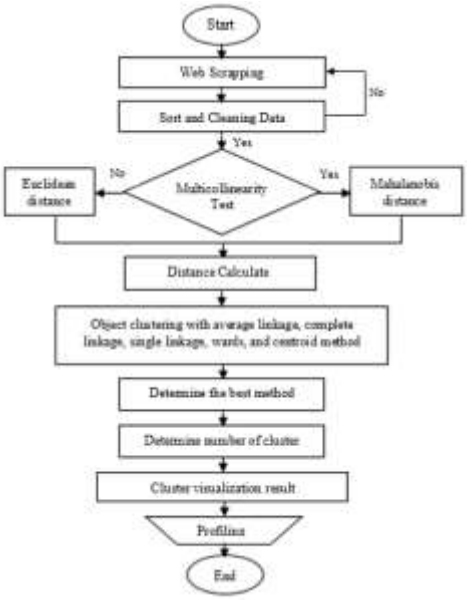

Figure 1. Research Flowchart

IV. RESULT AND DISCUSSION

A. *Descriptive Data of Star Hotels in Yogyakarta According to Tripadvisor.com*

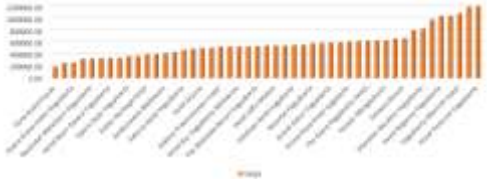

Figure 2. Hotel price charts from lowest to highest according to tripadvisor.com

Based on the graph above it is known that star hotels with the lowest prices are Duta Guest House and the most expensive Hotel Tentrem Yogyakarta. Then for hotels with the most reviews can be seen in the following picture,

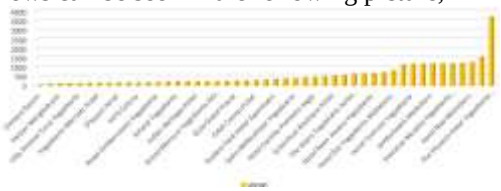

Figure 3. Number of hotel reviews by TripAdvisor

Based on the graph above, it is known that the most reviewed hotel reviews according to tripadvisor.com are The Phoenix Hotel Yogyakarta, and the least is Omkara Resort.

Then look at the percentage of rating on each hotel, the following graph,

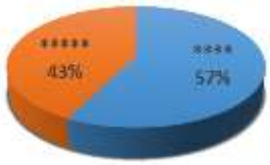

Figure 4. Percentage of rating according to tripadvisor.com

In the graph above, the percentage of star rating in Yogyakarta is presented according to tripadvisor.com, where it is known that there are only 2 rating options, namely 4 and 4.5 with a ratio of 43% and 57%,

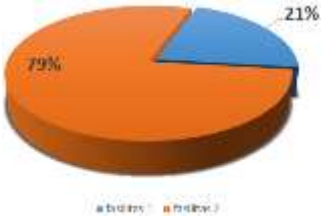

Figure 5. Percentage of facilities owned by hotels according to tripadvisor.com

B. *Cluster Assumption Test*

In cluster analysis, according to Hair in (Ningrat, et al., 2016) the assumption of multicollinearity between independent variables is very concerned because it can affect the validity of the results. Therefore, a multicollinearity test is performed on all data variables, and if there are multicollinearity variables, they will be reconsidered. The following are the results of the multicollinearity test obtained,

TABLE I. MULTICOLLINEARITY TEST RESULT

	rating	harga	fasilitas	ulasan
rating	1.0000000	0.0862871	-0.1834493	0.1416063
harga	0.0862871	1.0000000	-0.2712995	0.3520727
fasilitas	-0.1834493	-0.2712995	1.0000000	-0.2549298
ulasan	0.1416063	0.3520727	-0.2549298	1.0000000

According to (Yamin & Kurniawan, 2014), the data is said to have multicollinearity if the correlation value between

ISBN 978-602-53524-0-9

independent variables is> 0.70. From the results of the multicollinearity test, which is obtained, while the results obtained in table 4.1 above there is no value> 0.7 so that the assumption of no multicollinearity is fulfilled.

C. The Process of Getting the Best Cluster

After the cluster assumptions are met, then in this section is an important process in cluster analysis, which determines the best method of various sizes of distance contained in the hierarchy analysis than from the best method used to get the best cluster results. For the analysis phase and the discussion is as follows:

a) Distance Measuring Size

In distance measurement, there are many methods that can be used, but in this study, the distance measurement between objects is done using the usual method, Euclidean distance method where is the root of the sum of squares difference/deviation in the value for each variable

b) Selection of the Best Method

After knowing the distance, then the best method is determined. To determine the best cluster method there is still no theory or calculation in particular so that decisions depend on the researcher and the research context by not ignoring the substance, theory, and concepts that apply (Widiastuti & G, 2012). In this study referring to international journals by Sinan et al, the determination of the method is seen from the results of its cophenetic correlation, this cophenetic correlation is a measure of how faithfully a dendrogram maintains pairing distance between original unmodified data points. By using the formula (3.7) and the assistance of the R Program, which can be seen in appendix 2, the results of the cophenetic correlation values on the 5 methods are shown in the following table,

TABLE II. MULTICOLLINEARITY TEST RESULT

Method	Correlation *Cophenetic*
Single linkage	0.854
Average linkage	0.861
Complete linkage	0.850
Ward's	0.831
Centroid	0.860

Researchers decided to use the average linkage method to obtain the best cluster results in this study.

c) Cluster Formation

In this study the best grouping results are used, namely, the average linkage method that produces dendrogram as shown below,

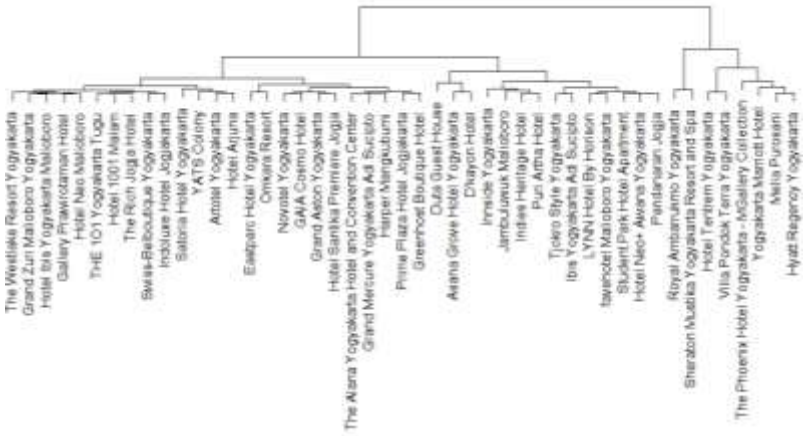

Figure 6. Number of hotel reviews by TripAdvisor

From the output dendrogram using the average linkage method it is known that there are 47 objects that have been grouped according to the hierarchy with the distance between objects that are owned are in the range 0 to 4e + 05, adjacent objects are shown by the location of the object on the dendrogram branch. Grouping starts from the observation pair with the distance closest to the average distance.

d) Determination of Number of Clusters

After knowing the results of the dendrogram in the following figure, in determining the number of clusters that will be formed in this hierarchy cluster, the researcher determines itself for the results of consideration, paying attention to the substance and conceptual aspects. In this study the researcher chose to use 3 clusters because the researchers wanted to classify the cluster results into 3 levels, namely low, medium, and high which were measured based on the average value of each cluster, as follows:

TABLE III. CLUSTER PROVINCIAL GROUP

Hotel	*Cluster*	Hotel	*Cluster*
Hyatt Regency	1	Fave Hotel	3
Meila Purosani	1	LYNN Hotel	3
Yogya Marriot	1	Ibis Hotel	3
The Phoenix Hotel	1	Tjokro Style	3
Villa Pondok Terra	1	Puri Artha Hotel	3
Tentrem Hotel	1	Indies Heritage	3
Sheraton Mustika	2	Jambu Luwuk Hotel	3
Royal Ambarukmo	2	Innside	3
Pandanaran Jogja	3	D'Kayon Hotel	3
Hotel Neo	3	Asana Grove	3
Student Park Hotel	3	Prime Plaza	3
Harper Mangkubumi	3	Santika Hotel	3

ISBN 978-602-53524-0-9

Grand Mercure	3	Grand Aston	3
The Alana	3	GAIA Cosmo	3
Novotel	3	YATS colony	3
Omkara resort	3	Satoria Hotel	3
Eastparc Hotel	3	Indoluxe	3
Arjuna Hotel	3	Swiss-Bel	3
Artotel	3	The Rich	3
1001 Malam	3	Neo Malioboro	3
The 101	3	Gallery Prawirotaman	3
Ibis Malioboro	3	Grand Zuri Malioboro	3
		The Westlake Resort	3

From the results of grouping above it is known that there are 3 clusters with the number of objects in cluster 1 as many as 6 hotels, objects in cluster 2 as many as 2 hotels, and objects in cluster 3 as many as 38 hotels.

e) Cluster Result Profiling

At this stage, the characteristics of each cluster will be described to explain that the clusters can differ in the relevant dimensions. The emphasis is on characteristics that are significantly different between clusters and predict members in a particular cluster (Widiastuti & G, 2012). To facilitate it, a cluster profiling table is presented as in the following table,

TABLE IV. CLUSTER PROFILING

Cluster	Number of members	Average of variables	Characteristic
1	6	272049.63	Average high prices, and lots of reviews
2	2	201837.90	The average price is medium, the facilities provided are only parking and free Wi-Fi
3	38	119210.70	Average low prices, and few reviews

Based on the results of the grouping characteristics obtained, it can be known that the profiling of each province in Indonesia has a different number of handling cases. The following are the centroid values of objects in each variable as presented in the following table,

TABLE V. THE AVERAGE VALUE OF OBJECTS IN A CLUSTER ON EACH VARIABLE

	rating	**price**	**facility**	**review**
Cluster 1	42.5	1087021.7	1.3	1133.0
Cluster 2	42.5	806636.5	2	670.5
Cluster 3	42	476347.4	1.8	451.64

CONCLUSION

From the results of the analysis that has been carried out until the final stage, there are some conclusions that can answer the formulation of the problem in this study, the conclusions are, among others,

a) Star hotel data in Yogyakarta according to TripAdvisor has 5 data variables, with a total of 47 objects, and has fulfilled the cluster assumptions

b) The best grouping results obtained using hierarchical clusters are the average linkage method in which there are 3 clusters with descriptions like the following:

i) The first cluster consists of 6 provinces, dominated by average high prices, and many customer reviews,

ii) Furthermore, cluster 2 consists of 2 provinces, which are dominated by standard hotel prices, and all facilities provided are only parking and free Wi-Fi,

iii) Then cluster 3 was 38 provinces, dominated by low hotel prices, and few reviews from customers.

ACKNOWLEDGMENT

Thank for Dr.R.B.Fajriya Hakim as a lecturer who has guided and provided his knowledge so that I can complete this paper, and thank you for my parent who has provided moral support and funds.

REFERENCES

[1] DPPKA, 2017. [Online] Available at: https://visitingjogja.com/downloads/Buku%20Statistik%20Kepariwisataan%20DIY%202016.pdf

[2] Ghozali, I., 2009. Analisis *Cluster*. Dalam: *Aplikasi Analisis Multivariate dengan Program SPSS.* Semarang: Badan Penerbit Universitas Diponegoro, p. 312.

[3] Hakam, M., Sudarno & Hoyyi, A., 2015. Analisis Jalur Terhadap Faktor-Faktor Yang Mempengaruhi Indeks Prestasi Kumulatif (IPK) Mahasiswa Statistika UNDIP. *Jurnal Gaussian*, 4(1), pp. 61-70.

[4] Hidayat, A., 2014. *Penjelasan Lengkap Tentang Analisis Cluster.* [*Online*]
Available at: https://www.statistikian.com/2014/03/analisis-*cluster*.html [Diakses 12 maret 2017].

[5] Johnson & Wichern, 2007. Applied Multivariate Statistical Analysis 6nd Edition. New Jersey: Pretice-Hall.

[6] Kemendagri, 2017. [*Online*] Available at: http://investasi.jogjakota.go.id/id/more/page/111/Profil-Kota-Yogyakarta

[7] Ningrat, D., Marruddani, D. & Wuryandari, T., 2016. Analisis *Cluster* dengan Algoritma K-Means dan Fuzzy C-Means *Cluster*ing Untuk Pengelompokan Data Obligasi Korporasi. *Jurnal Gaussian*, 5(4), pp. 641-650.

[8] Nugraha, I. & Dharmayanti, D., 2017. Penerapan Outlier Analysis Sebagai Salah Satu Rekomendasi Kelompok Belajar Terhadap Siswa Kelas 6 Di Sdn Pagelaran Ii. *Jurnal ilmiah komputer dan informatika (KOMPUTA).*

[9] Rachmatin, D., 2014. Aplikasi Metode-metode Agglomerative Dalam Analisis Klaster Pada Data Tingkat Polusi Udara. *Jurnal Ilmiah Program Studi Matematika STKIP Siliwangi Bandung*, 3(2), pp. 133-149.

[10] Sindonews, 2017. [*Online*] Available at: https://ekbis.sindonews.com/read/1199814/179/kuartal-i-2017-bisnis-perhotelan-di- yogyakarta-melesu-1493044383

[11] Saracli, S., Dogan, N. & Dogan, I., 2013. Comparison of Hierarchical *Cluster* Analysis Methods by Cophenetic Correlation. *Journal of Inequalities and Applications*, pp. 2-4.

[12] Sirojuddin, A., 2016. Analisis Cluster Pada Kabupaten/Kota di Provinsi Jawa Timur Berdasarkan Indikator Indeks Pembangunan Manusia. [Online]
Available at: http://etheses.uin-malang.ac.id/5762/1/12610033.pdf

ISBN 978-602-53524-0-9

[13] Sitepu, R., Irmeilyana & Gultom, B., 2011. Analisis *Cluster* Terhadap Tingkat Pencemaran Udara Pada Sektor Industri di Sumatera Selatan. *Jurnal Penelitian Sains*, 14(3(A)), pp. 11-17.

[14] Tribunjogja, 2013. [*Online*] Available at: http://jogja.tribunnews.com/2013/10/31/phri-catat-ada-1160-hotel-di-yogyakarta/

[15] Widiastuti, M. & G, E. Y., 2012. Pemetaan Kemiskinan Kabupaten/Kota di Provinsi Jawa Tengah Tahun 2002 dan 2010 Menggunakan Analisis Klaster. *Diponegoro Journal Of Economics*, 1(1), pp. 1-14.

[16] Yamin, S. & Kurniawan, H., 2009. SPSS Complete Teknik Analisis Statistik Terlengkap SPSS Seri 1. 1 penyunt. Jakarta: Salemba.

[17] Yamin, S. & Kurniawan, H., 2014. SPSS Complete: Teknik Analisis Statistik Terlengkap dengan Software SPSS. Jakarta: Salemba Infotek.

[18] Yamin, S., Rachmah, L. & Kurniawan, H., 2011. Regresi Korelasi dalam Genggaman Anda: Aplikasi dengan Software SPSS, Eviews, MINITAB, dan STATGRAPHICS. Jakarta: Salemba Empat.

AUTHORS PROFILE

Devina Gilar Fitri Ayu Sumardi is a bachelor student of department of statistics in the Islamic University of Indonesia, she is from North Kalimantan, and is currently 20 years old. active in the world of an organization from high school until now. since semester 4 she has begun to work in scientific writing to apply and utilize the knowledge she has in lectures.

Dr.Raden Bagus Fajriya Hakim, S.Si., M.Si is a lecturer at the Islamic University of Indonesia focusing on the field of big data or data science and live in Yogyakarta. He is an S1 and S2 alumnus of Gajah Mada University. He is an active lecturer, making a lot of learning content shared in several social media accounts, one of which is through his following youtube account https://www.youtube.com/channel/UCQzmv6yi_Ga8nW73L2kqoGQ

ISBN 978-602-53524-0-9