Analisis *Cluster* Pada Data Ulasan Destinasi Pariwisata Asia Timur di Situs TripAdvisor

Dina Lestari[1, a)], Dwi Ariyanti Nur Latifah[2, b)], Rahayu Isnaini[3, c)], Richwani Neysa Febriana[4, d)], Tara Dwipa Ardhatu Timur[5, e)], Tarisa Putri Cahyani[5, f)]

[1,2,3,4,5] *Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Gadjah Mada, Indonesia*

[a)] Corresponding author: dinalestari057@mail.ugm.ac.id
[b]dwiariyanti02@mail.ugm.ac.id
[c)]rahayu.isnaini@mail.ugm.ac.id
[d)]richwani.neysa@mail.ugm.ac.id
[e)]taradwipa02@mail.ugm.ac.id
[f)]tarisa.putri.cahyani@mail.ugm.ac.id

Abstrak: *Review* merupakan suatu pendapat langsung dari seseorang. *Review* juga menjadi salah satu faktor yang menentukan keputusan seseorang, dengan menunjukkan bahwa orang dapat mengambil jumlah *Review* sebagai indikator popularitas tempat atau nilai dari suatu tempat dengan mempengaruhi kemauan untuk mendatangi suatu tempat pariwisata. Penelitian ini membahas tentang Penerapan metode K-Means untuk *Review* dataset travel. Dataset ini diambil langsung melalui halaman situs web UCI Machine Learning Store dengan jumlah information sebanyak 980 record, yang terdiri dari 10 variabel atau atribut yaitu *Art Galleries*, *Dance Clubs*, *Juice Bars*, *Restaurants*, *Museums*, *Resorts*, *Parks/Picnic Spots*, *Beaches*, *Theaters*, *Religious Institutions*. Proses cluster dibagi kedalam 2 (dua) cluster dan akan ditinjau metode manakah yang lebih baik digunakan. Berdasarkan hasil analisis telah didapatkan metode clustering terbaik yaitu metode K-Means cluster. Hasil clustering ini dapat digunakan sebagai rekomendasi untuk menentukan destinasi yang cocok dikunjungi oleh masing masing user sehingga user dapat mempersiapkan perjalanan agar lebih efektif.

LATAR BELAKANG

Sektor Pariwisata berperan penting dalam meningkatkan perekonomian suatu negara, khususnya dalam mengurangi jumlah pengangguran dan meningkatkan produktivitas suatu negara. Sektor Pariwisata merupakan salah satu sektor strategis yang harus dimanfaatkan untuk pembangunan kepariwisataan sebagai bagian dari pembangunan suatu negara. Pembangunan kepariwisataan mempunyai tujuan akhir meningkatkan pendapatan masyarakat yang pada akhirnya meningkatkan kesejahteraan masyarakat dan pertumbuhan ekonomi.

Untuk menjawab tantangan industri pariwisata berbagai situs online dikembangan untuk memberikan pelayanan berupa booking online, baik tiket pesawat, hotel, hingga akomodasi, dan memberikan informasi wisata di berbagai daerah. Salah satu situs online pariwisata yang banyak digunakan masyarakat global saat ini adalah TripAdvisor. Pengguna jasa TripAdvisor memiliki fitur dimana pengunjung memungkinkan memberi online review berupa rating bagi tempat

wisata yang pernah dikunjungi. Tripadvisor mendukung orang dalam fase pre-travel juga fase post-travel, untuk berbagi pengalaman, me-review hotel-hotel dan destinasi wisata, dan memposting foto dan video dari perjalanan mereka. Rating tersebut sebagai penilaian yang dapat digunakan oleh berbagai pihak dalam menentukan destinasi unggulan, meningkatkan kualitas pelayanan, dan memberikan informasi terkait destinasi unggulan tersebut.

Dengan penerapan data mining banyaknya data yang masuk dapat dikelola dengan relatif cepat sehingga mudah dalam menemukan pengetahuan atau informasi.Salah satu metode pada data mining, yaitu *Clustering*. *Clustering* adalah pengelompokkan dari sejumlah data menjadi suatu kelompok-kelompok data tertentu. Untuk meringkas data atau sejumlah variabel menjadi lebih sedikit, analisis *cluster* ini menjadi sangat berguna. *Clustering* dilakukan dengan mengelompokan objek–objek berdasarkan kesamaan karakteristik tertentu diantara objek–objek yang akan diteliti. Selain itu *clustering* digunakan untuk mengetahui struktur data yang bisa diterapkan lebih lanjut pada berbagai aplikasi secara luas. Pada studi kasus ini digunakan metode K-Means dan K-Medoids karena akan dilakukan perbandingan kedua metode tersebut dalam menganalisa pola pengelompokkan dan menentukkan metode mana yang terbaik dalam pengolahan dataset tersebut.

Dari latar belakang yang telah dijelaskan diatas, maka akan dilakukan *clustering* user berdasarkan kategori-kategori destinasi pada dataset ulasan destinasi pariwisata Asia Timur dalam situs TripAdvisor dengan menggunakan metode K-Means dan K-Medoids.

TUJUAN DAN MANFAAT

Tujuan dari analisis ini yaitu untuk mengelompokkan user berdasarkan kategori-kategori destinasi dengan menggunakan metode K-Means dan K-Medoids serta membandingkan hasil clustering dari kedua metode tersebut. Hasil analisis ini diharapkan dapat menambah wawasan dan pengetahuan terkait destinasi pariwisata yang akan dikunjungi serta dapat menjadi bahan pertimbangan bagi wisatawan maupun pengelola wisata.

DASAR TEORI

Analisis klaster merupakan salah satu jenis permasalahan dalam data mining. Data mining sendiri menurut David Hand, Heikki Mannila, dan Padhraic Smyth dari MIT dalam Larose (2006) adalah analisis terhadap data (biasanya data yang berukuran besar) untuk menemukan hubungan yang jelas serta menyimpulkannya yang belum diketahui sebelumnya dengan cara terkini dipahami dan berguna bagi pemilik data tersebut. Sedangkan analisis klaster dalam data mining (dikenal juga dengan istilah clustering) adalah metode yang digunakan untuk membagi rangkaian data menjadi beberapa grup berdasarkan kesamaan-kesamaan yang telah ditentukan (Gorunescu, 2011).

Non hierarchical methods merupakan pengelompokan berbasis sekatan atau partitioning yang menghasilkan partisi dari data sehingga objek dalam cluster lebih mirip satu sama lain daripada objek yang ada dalam cluster lain (Triyanto, 2015). Berbeda dengan hirarki, prosedur pengelompokan ini tidak dilakukan secara bertahap, dan jumlah cluster ditentukan terlebih dahulu (Machfudhoh, 2013). Biasanya metode non hirarki diterapkan pada data yang memiliki jumlah sangat banyak.

Pusat cluster atau biasa disebut centroid yang dipilih pada metode ini merupakan pusat cluster sementara dengan terus memperbaharui pusat cluster untuk tiap iterasi sampai kriteria pemberhentian tercapai, sehingga dimungkinkan bahwa objek yang telah berada pada suatu kelompok atau cluster tertentu dapat pindah ke cluster yang lain. (Safe'i, 2018)

Salah satu metode pengelompokan non hirarki adalah K-Means dan K-Medoids.

➢ K-Means Clustering

K-Means merupakan salah satu metode pengelompokan data non-hirarki yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih cluster. Metode ini mempartisi data ke dalam cluster sehingga data berkarakteristik sama dimasukkan ke dalam satu cluster yang sama dan data yang berkarakteristik berbeda dikelompokkan ke dalam cluster yang lain. Adapun tujuan pengelompokan data ini adalah untuk meminimalkan fungsi objektif yang diatur dalam proses pengelompokan, yang pada umumnya berusaha meminimalkan variasi di dalam suatu cluster dan memaksimalkan variasi antar cluster. (Supranto, 2004)

K-Means menggunakan "mean" yaitu rata-rata sebagai pusat cluster-nya, sehingga K-Means tidak robust atau tidak cukup baik terhadap data yang memiliki outlier. Sehingga, untuk mengatasi hal tersebut dapat digunakan metode K-Medoids untuk mengelompokkan data yang memiliki outlier. (Larasati, 2017)

➢ K-Medoids Clustering

K-Medoids atau Partitioning Around Method (PAM) adalah metode cluster non hirarki yang merupakan varian dari metode K-Means. K-Medoids hadir untuk mengatasi kelemahan K-Means yang sensitif terhadap outlier karena suatu objek dengan suatu nilai yang besar mungkin secara substansial menyimpang dari distribusi data (Jiawei & Kamber, 2006). Hal ini didasarkan pada penggunaan medoids bukan dari pengamatan mean yang dimiliki oleh setiap cluster, dengan tujuan mengurangi sensitivitas dari partisi sehubungan dengan nilai ekstrim yang ada dalam dataset (Vercellis, 2009)

K-Medoids menggunakan metode pengelompokan partisi untuk mengelompokan sekumpulan n objek menjadi sejumlah k cluster. Algoritma ini menggunakan objek pada kumpulan objek yang mewakili sebuah cluster. Objek yang mewakili sebuah cluster disebut dengan medoids.Medoid merupakan objek yang letaknya terpusat di dalam suatu cluster sehingga robust terhadap outlier. Cluster dibangun dengan menghitung kedekatan yang dimiliki antara medoids dengan objek non medoids (Setyawati, 2017). Sama halnya dengan K-Means, metode K-Medoid lebih menguntungkan diterapkan pada data yang sangat besar.

METODE ANALISIS

Berdasarkan dataset ulasan destinasi pariwisata Asia Timur dalam situs TripAdvisor dan dengan tujuan dari analisis yang dilakukan yaitu untuk mengelompokkan User berdasarkan ulasannya di berbagai kategori destinasi, maka dilakukan analisis *clustering* dengan menggunakan metode K-Means dan K-Medoids karena terdapat cukup banyak *outlier* pada data sehingga digunakan kedua metode tersebut untuk membandingkan hasil *clustering* manakah yang lebih baik. Lebih lanjut, pembahasan ini akan dijelaskan pada bagian Hasil Analisis dan Pembahasan

TENTANG DATA

Kumpulan data berasal dari TripAdvisor.com berisi ulasan tentang destinasi dalam 10 kategori yang disebutkan di seluruh Asia Timur. Setiap peringkat wisatawan dipetakan sebagai Sangat Baik (4), Sangat Baik (3), Rata-Rata (2), Buruk (1), dan Sangat Buruk (0) dan peringkat rata-rata digunakan untuk setiap kategori per User.

Data yang digunakan terdiri dari 980 observasi dan 11 variabel. Variabel-variabel yang digunakan adalah sebagai berikut:

a. **User_ID** : User ID
b. **Category_1** : Rata-rata *feedback* User pada *Art Galleries*
c. **Category_2** : Rata-rata *feedback* User pada *Dance Clubs*
d. **Category_3** : Rata-rata *feedback* User pada *Juice Bars*
e. **Category_4** : Rata-rata *feedback* User pada *Restaurants*
f. **Category_5** : Rata-rata *feedback* User pada *Museums*
g. **Category_6** : Rata-rata *feedback* User pada *Resorts*
h. **Category_7** : Rata-rata *feedback* User pada *Parks/Picnic Spots*
i. **Category_8** : Rata-rata *feedback* User pada *Beaches*
j. **Category_9** : Rata-rata *feedback* User pada *Theaters*
k. **Category_10** : Rata-rata *feedback* User pada *Religious Institutions*

Variabel-variabel diatas digunakan sebagai indikator dalam mengelompokkan User berdasarkan ulasannya di berbagai destinasi/kategori.

HASIL ANALISIS DAN PEMBAHASAN

Input Data

	User ID	Art Galleries	Dance Clubs	Juice Bars	Restaurants	Museums	Resorts	Parks/picnic spots	Beaches	Theaters	Religious institutions
0	10001	0.93	1.8	2.29	0.62	0.80	2.42	3.19	2.79	1.82	2.42
1	10002	1.02	2.2	2.66	0.64	1.42	3.18	3.21	2.63	1.86	2.32
2	10003	1.22	0.8	0.54	0.53	0.24	1.54	3.18	2.80	1.31	2.50
3	10004	0.45	1.8	0.29	0.57	0.46	1.52	3.18	2.96	1.57	2.86
4	10005	0.51	1.2	1.18	0.57	1.54	2.02	3.18	2.78	1.18	2.54

Gambar 1. Input data tripadvisor_review.csv

Sebelum dilakukan analisis data, dilakukan input dataset 'tripadvisor_review.csv' serta memanggil 5 data pertama dari dataset 'tripadvisor_review.csv' untuk mengecek kesesuaian data yang terinput dengan dataset yang dimiliki.

Data Unik

```
0      False
1      False
2      False
3      False
4      False
       ...
975    False
976    False
977    False
978    False
979    False
Length: 980, dtype: bool
```

	User ID	Art Galleries	Dance Clubs	Juice Bars	Restaurants	Museums	Resorts	Parks/picnic spots	Beaches	Theaters	Religious institutions
0	10001	0.93	1.80	2.29	0.62	0.80	2.42	3.19	2.79	1.82	2.42
1	10002	1.02	2.20	2.66	0.64	1.42	3.18	3.21	2.63	1.86	2.32
2	10003	1.22	0.80	0.54	0.53	0.24	1.54	3.18	2.80	1.31	2.50
3	10004	0.45	1.60	0.29	0.57	0.46	1.52	3.18	2.96	1.57	2.86
4	10005	0.51	1.20	1.18	0.57	1.54	2.02	3.18	2.78	1.18	2.54
...	...	...	...	...	...	...	...	...	...	...	...
975	10976	0.74	1.12	0.30	0.53	0.88	1.38	3.17	2.78	0.99	3.20
976	10977	1.25	0.92	1.12	0.38	0.78	1.68	3.18	2.79	1.34	2.80
977	10978	0.61	1.32	0.67	0.43	1.30	1.78	3.17	2.81	1.34	3.02
978	10979	0.93	0.20	0.13	0.43	0.30	0.40	3.18	2.98	1.12	2.46
979	10980	0.93	0.56	1.13	0.51	1.34	2.36	3.18	2.87	1.34	2.40

980 rows × 11 columns

Gambar 3. Cek adanya data duplikat

Selanjutnya akan dicek apakah pada dataset mengandung data duplikat. Berdasarkan output yang didapatkan terlihat bahwa pada pemanggilan data pertama keterangan jumlah data yaitu sebanyak 980 data, pada pemanggilan data kedua setelah dilakukan penghapusan data duplikat agar menjadi data yang unik keterangan jumlah data yaitu sebesar 980. Karena pemanggilan data sebelum dan setelah penghapusan data sama yaitu 980 maka data yang dimiliki tidak terduplikat atau data yang dimiliki sudah unik.

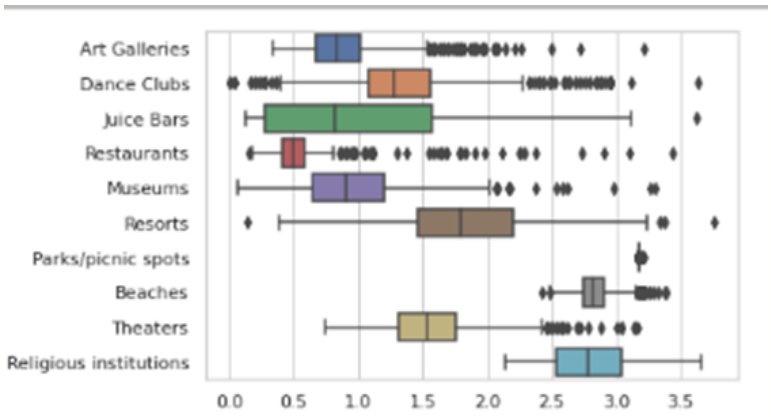
Missing Value

```
User ID                 0
Art Galleries           0
Dance Clubs             0
Juice Bars              0
Restaurants             0
Museums                 0
Resorts                 0
Parks/picnic spots      0
Beaches                 0
Theaters                0
Religious institutions  0
dtype: int64
```

Gambar 4. Cek adanya data Missing Value

Pada pengujian Missing Value secara keseluruhan didapatkan output pada masing masing kategori yaitu 0 yang berarti bahwa tidak ditemukan adanya masing value pada masing masing kategori.

Mengidentifikasi Outlier



Identifikasi Outlier pada dataset 'tripadvisor_review.csv' dilakukan dengan membuat Boxplot pada masing masing kategori. Output yang dihasilkan terlihat bahwa pada data kategori Religion Institutions tidak ditemukan adanya titik titik diluar boxplot yang berarti bahwa tidak ditemukan adanya outlier pada data kategori Religion Institutions, namun pada data kategori selain kategori Religion Institutions menunjukkan adanya titik titik diluar kotak boxplot sehingga teridentifikasi adanya outlier pada data.

Data_Clustering

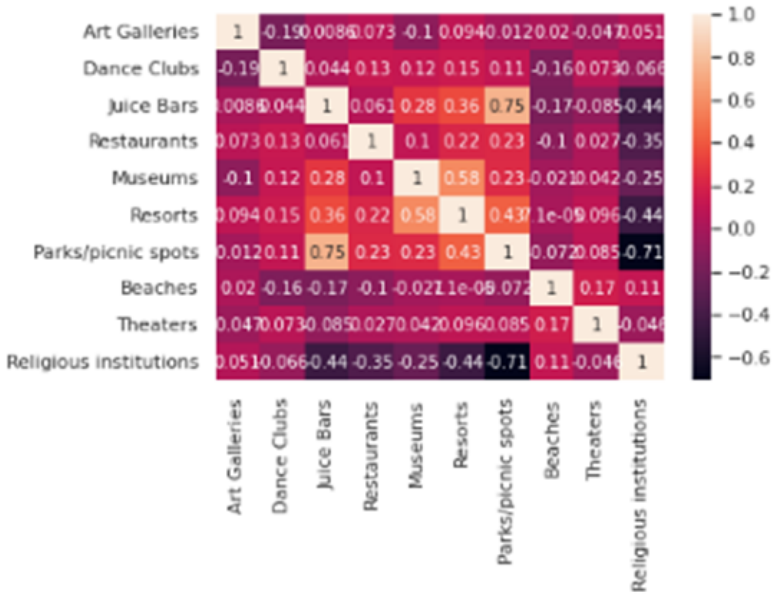
User ID	Art Galleries	Dance Clubs	Juice Bars	Restaurants	Museums	Resorts	Parks/picnic spots	Beaches	Theaters	Religious institutions
10001	0.93	1.80	2.29	0.62	0.80	2.42	3.19	2.79	1.82	2.42
10002	1.02	2.20	2.66	0.64	1.42	3.18	3.21	2.63	1.86	2.32
10003	1.22	0.80	0.54	0.53	0.24	1.54	3.18	2.80	1.31	2.50
10004	0.45	1.80	0.29	0.57	0.46	1.52	3.18	2.96	1.57	2.86
10005	0.51	1.20	1.18	0.57	1.54	2.02	3.18	2.78	1.18	2.54
...	...	...	...	...	...	...	...	...	...	...
10976	0.74	1.12	0.30	0.53	0.88	1.38	3.17	2.78	0.99	3.20
10977	1.25	0.92	1.12	0.38	0.78	1.68	3.18	2.79	1.34	2.80
10978	0.61	1.32	0.67	0.43	1.30	1.78	3.17	2.81	1.34	3.02
10979	0.93	0.20	0.13	0.43	0.30	0.40	3.18	2.98	1.12	2.46
10980	0.93	0.56	1.13	0.51	1.34	2.36	3.18	2.87	1.34	2.40

980 rows × 10 columns

Gambar 5. Tampilan data_clustering

Didefinisikan data_clustering sebagai dataframe dari dataset yang memiliki total data sebesar 980 data.

Uji Asumsi Multikolinearitas



Gambar 6. Heatmap data_clustering

Pada analisis clustering diperlukan adanya uji asumsi multikolinearitas. Pada output yang ditampilkan pada data train didapatkan bahwa mayoritas korelasi antar data < 0,7, namun terdapat korelasi antara kategori juice bars dan kategori park/picnic spots serta kategori juice bars dan kategori Religious institutions yang memiliki nilai korelasi ≥ 0,7 yang masing masing secara berurutan bernilai 0,75 serta 0,71. Karena terdapat 2 korelasi antar data kategori yang memiliki nilai korelasi ≥ 0,7 sehingga dapat dikatakan bahwa asumsi no multikolinearitas tidak terpenuhi

Menghitung VIF

	fatures	VIF
0	Art Galleries	9.510295
1	Dance Clubs	10.099982
2	Juice Bars	3.752489
3	Restaurants	5.579521
4	Museums	8.933049
5	Resorts	24.526804
6	Parks/picnic spots	646.854951
7	Beaches	475.220916
8	Theaters	20.812945
9	Religious institutions	120.639574

Gambar 7. Nilai VIF masing masing variabel

Untuk mendeteksi adanya multikolinearitas pada data dilakukan pengujian nilai VIF. Berdasarkan output didapatkan nilai VIF pada masing masing kategori yang meliputi kategori

Art Galleries (9,510295), kategori Dance Clubs (10,099982), Kategori Juice Bars (3,752489), kategori Restaurant (5,579521), kategori Museums (8,933049), kategori Resort (24,526004), kategori Park/Picnic Spots (646,854951), kategori Beaches (475,220916), kategori Theaters (20,812945), kategori Religious Institutions (120,639574). Pada pengujian selanjutnya akan digunakan data kategori yang memiliki nilai ≤ 10. Sehingga yang akan digunakan untuk pengujian selanjutnya yaitu data kategori Art Galleries, Dance Clubs, Juice Bars, Restaurant, Museum

	VIF	Features
1	6.282038	Dance Clubs
4	5.512570	Museums
0	5.104403	Art Galleries
3	4.617672	Restaurants
2	2.859720	Juice Bars

Gambar 8. Nilai VIF kelima variabel yang signifikan

Setelah dilakukan drop data kategori yang memiliki VIF > 10 yaitu kategori 6 sampai 10, dilakukan pengujian nilai VIF kembali, didapatkan nilai VIF yang baru adalah kategori Dance Clubs (6,282038), kategori Museums (5,512570), kategori Art Galleries (5,104403), kategori Restaurant (4,617672), kategori Juice Bars (2,859720). Karena nilai VIF pada semua kategori < 10 maka kelima kategori dapat digunakan untuk analisis cluster

	User ID	Art Galleries	Dance Clubs	Juice Bars	Restaurants	Museums
0	10001	0.93	1.80	2.29	0.62	0.80
1	10002	1.02	2.20	2.66	0.64	1.42
2	10003	1.22	0.80	0.54	0.53	0.24
3	10004	0.45	1.80	0.29	0.57	0.46
4	10005	0.51	1.20	1.18	0.57	1.54
...	...	...	...	...	...	...
975	10976	0.74	1.12	0.30	0.53	0.88
976	10977	1.25	0.92	1.12	0.38	0.78
977	10978	0.61	1.32	0.67	0.43	1.30
978	10979	0.93	0.20	0.13	0.43	0.30
979	10980	0.93	0.56	1.13	0.51	1.34

980 rows × 6 columns

Gambar 9. Tampilan data_clustering2

Didefinisikan data_clustering2 sebagai data frame dari dataset yang memiliki total data sebesar 980 data.

```
10001    1
10645    1
10647    1
10648    1
10649    1
        ..
10331    1
10332    1
10333    1
10334    1
10980    1
Name: User ID, Length: 980, dtype: int64
```

Gambar 10. Variabel counts masing masing variabel

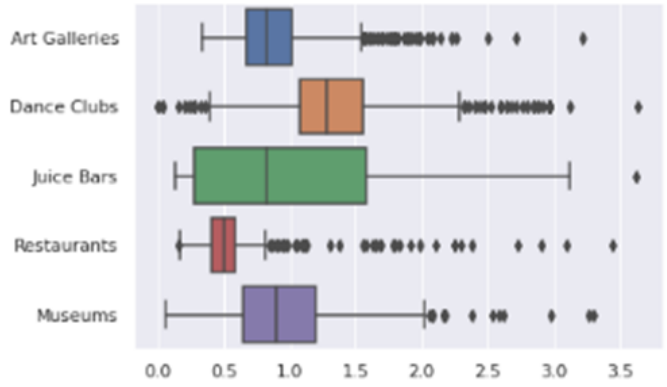
Selanjutnya akan dicek value counts data yang dimiliki tiap user ID, berdasarkan output diketahui tiap user ID hanya memiliki value counts sebesar 1

User ID	Art Galleries	Dance Clubs	Juice Bars	Restaurants	Museums
10001	0.93	1.80	2.29	0.62	0.80
10002	1.02	2.20	2.66	0.64	1.42
10003	1.22	0.80	0.54	0.53	0.24
10004	0.45	1.80	0.29	0.57	0.46
10005	0.51	1.20	1.18	0.57	1.54
...	...	...	...	...	...
10976	0.74	1.12	0.30	0.53	0.88
10977	1.25	0.92	1.12	0.38	0.78
10978	0.61	1.32	0.67	0.43	1.30
10979	0.93	0.20	0.13	0.43	0.30
10980	0.93	0.56	1.13	0.51	1.34

980 rows × 5 columns

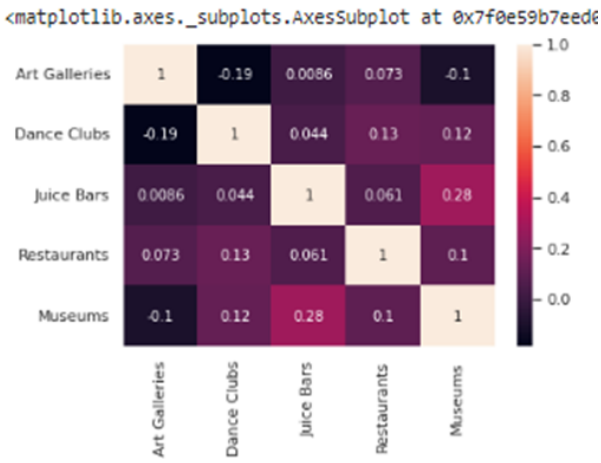
Gambar 11. Tampilan data_clustering3

Setelah dilakukan pengecekan value counts, dilakukan pengelompokan data user ID dengan menggunakan group by yang selanjutnya akan disimpan di dalam variabel data_clustering3



Gambar 12. Pengecekan data outlier dengan Boxplot

Dilakukan pengecekan data outlier kembali menggunakan boxplot, pada kelima data kategori menunjukkan adanya titik titik diluar kotak boxplot sehingga teridentifikasi adanya outlier pada data.



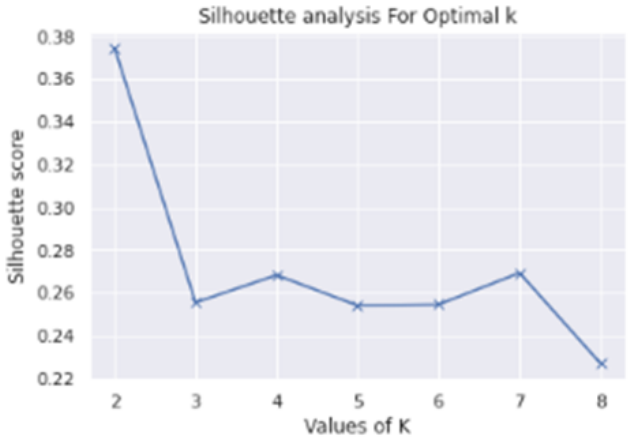
Gambar 13. Heatmap data_clustering3

Selanjutnya akan dicek kembali multikolinearitas pada data dengan menggunakan data_clstering3. Berdasarkan output yang diperoleh didapatkan keseluruhan korelasi variabel memiliki nilai < 0,7 sehingga dapat disimpulkan bahwa asumsi no multikolinearitas pada data_clstering3 terpenuhi

K-Means Clustering

```
2
Silhouetter Score: 0.374
3
Silhouetter Score: 0.255
4
Silhouetter Score: 0.268
5
Silhouetter Score: 0.254
6
Silhouetter Score: 0.255
7
Silhouetter Score: 0.269
8
Silhouetter Score: 0.227
```

Gambar 14. Nilai Silhouette



Gambar 15. Tampilan grafik Silhouette

Selanjutnya dilakukan clustering data dengan metode K-Means cluster, cluster yang akan dibuat sebanyak 7 cluster yang meliputi cluster 2 sampai 8 serta akan dicari nilai silhouette masing masing cluster. Berdasarkan output yang diperoleh didapatkan nilai silhouette pada 2 cluster (0,374), 3 cluster (0,255), 4 cluster (0,268), 5 cluster (0,254), 6 cluster (0,255), 7 cluster (0,269), 8 cluster (0,227). Didapatkan nilai silhouette terbaik yaitu pada 2 cluster dengan nilai silhouette sebesar 0,374

User ID	Art Galleries	Dance Clubs	Juice Bars	Restaurants	Museums	cluster
10001	0.93	1.80	2.29	0.62	0.80	0
10002	1.02	2.20	2.66	0.64	1.42	0
10003	1.22	0.80	0.54	0.53	0.24	1
10004	0.45	1.80	0.29	0.57	0.46	1
10005	0.51	1.20	1.18	0.57	1.54	0
...	...	...	...	...	...	...
10976	0.74	1.12	0.30	0.53	0.88	1
10977	1.25	0.92	1.12	0.38	0.78	1
10978	0.61	1.32	0.67	0.43	1.30	1
10979	0.93	0.20	0.13	0.43	0.30	1
10980	0.93	0.56	1.13	0.51	1.34	1

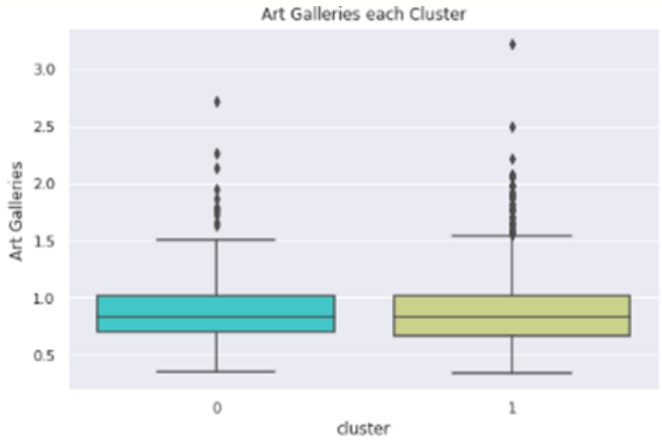
980 rows × 6 columns

Gambar 16. Tampilan data hasil clustering

```
1    609
0    371
Name: cluster, dtype: int64
```

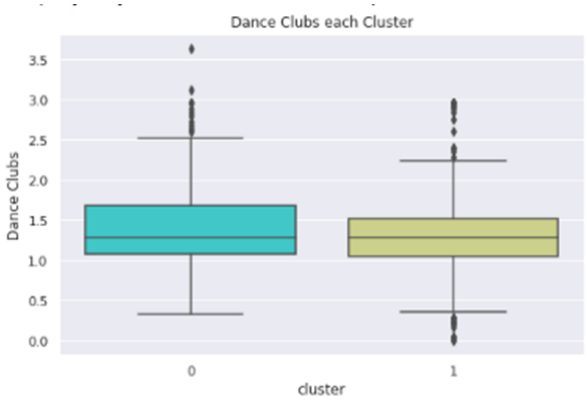
Gambar 17. Jumlah data per cluster

Setelah dilakukan pengelompokan cluster dengan menggunakan metode K-Means didapatkan 2 cluster yaitu cluster 1 yang memiliki jumlah data sebanyak 609 dan cluster 0 yang memiliki data sebanyak 371 data



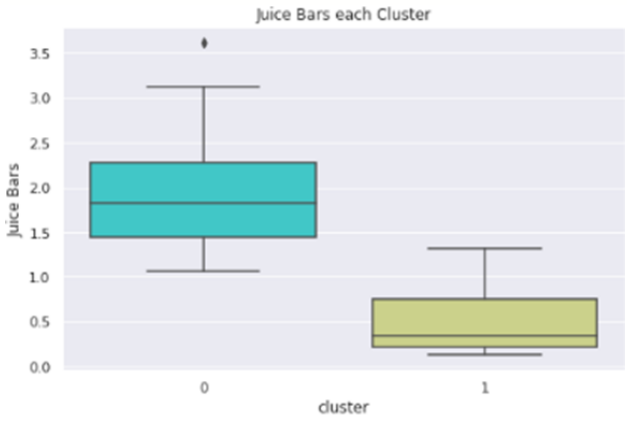
Gambar 18. Boxplot kategori Art Galleries

Selanjutnya akan dibuat boxplot pada masing masing kategori untuk melihat adanya perbedaan diantara kedua cluster. Berdasarkan output yang didapatkan pada kategori Art and Galeries untuk kedua cluster relatif sama serta pada masing masing cluster memiliki data outlier.



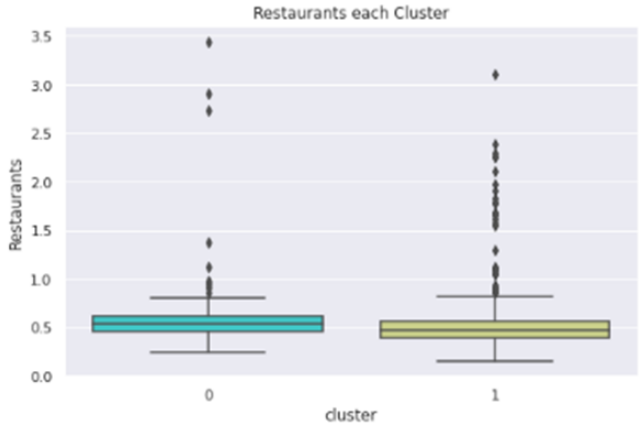
Gambar 19. Boxplot kategori Dance Clubs

Berdasarkan output yang didapatkan pada kategori Dance Clubs untuk kedua cluster relatif sama, namun pada cluster 0 memiliki lebar boxplot yang lebih besar dari cluster 1 serta pada masing masing cluster memiliki data outlier.



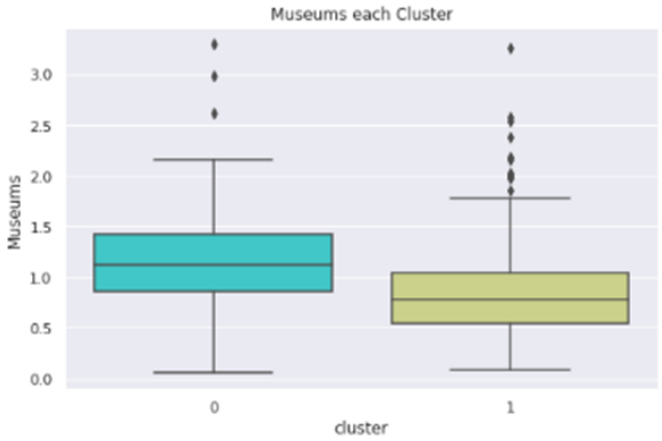
Gambar 20. Boxplot kategori Juice Bars

Berdasarkan output yang didapatkan pada kategori Juice Bars pada cluster 0 memiliki nilai Pusat, Q1, dan Q3 yang lebih tinggi dari cluster 1 serta terdapat 1 outlier pada data cluster 0



Gambar 21. Boxplot kategori Restaurants

Berdasarkan output yang didapatkan pada kategori Restaurant untuk kedua cluster relatif sama serta pada masing masing cluster memiliki data outlier.



Gambar 22. Boxplot kategori Museums

Berdasarkan output yang didapatkan pada kategori Museums terlihat pada cluster 0 memiliki nilai Pusat, Q1, dan Q3 yang lebih tinggi dari cluster 1 serta terdapat beberapa outlier pada data cluster 0 maupun cluster 1

K-Medoids

```
2
Silhouetter Score: 0.371
3
Silhouetter Score: 0.253
4
Silhouetter Score: 0.257
5
Silhouetter Score: 0.222
6
Silhouetter Score: 0.208
7
Silhouetter Score: 0.180
8
Silhouetter Score: 0.182
```

Gambar 23. Nilai silhouette

Gambar 24. Grafik nilai silhouette

Selanjutnya dilakukan clustering data dengan metode K-Medoids cluster, cluster yang akan dibuat sebanyak 7 cluster yang meliputi 2 sampai 8 cluster serta akan dicari nilai silhouette masing masing cluster. Berdasarkan output yang diperoleh didapatkan nilai silhouette pada 2 cluster (0,371), 3 cluster (0,253), 4 cluster (0,257), 5 cluster (0,222), 6 cluster (0,208), 7 cluster (0,180), 8 cluster (0,182). Didapatkan nilai silhouette terbaik yaitu pada 2 cluster dengan nilai silhouette sebesar 0,371

User ID	Art Galleries	Dance Clubs	Juice Bars	Restaurants	Museums	cluster
10001	0.93	1.80	2.29	0.62	0.80	0
10002	1.02	2.20	2.66	0.64	1.42	0
10003	1.22	0.80	0.54	0.53	0.24	1
10004	0.45	1.80	0.29	0.57	0.46	1
10005	0.51	1.20	1.18	0.57	1.54	0
...	...	...	...	...	...	...
10976	0.74	1.12	0.30	0.53	0.88	1
10977	1.25	0.92	1.12	0.38	0.78	1
10978	0.61	1.32	0.67	0.43	1.30	1
10979	0.93	0.20	0.13	0.43	0.30	1
10980	0.93	0.56	1.13	0.51	1.34	0

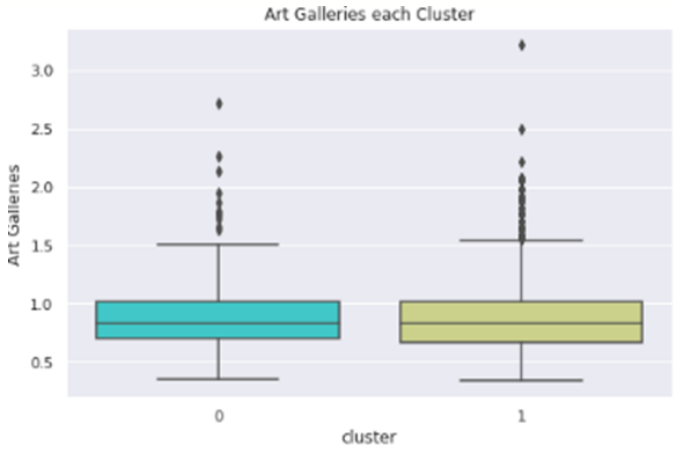
980 rows × 6 columns

Gambar 24. Data hasil clustering

```
1    596
0    384
Name: cluster, dtype: int64
```

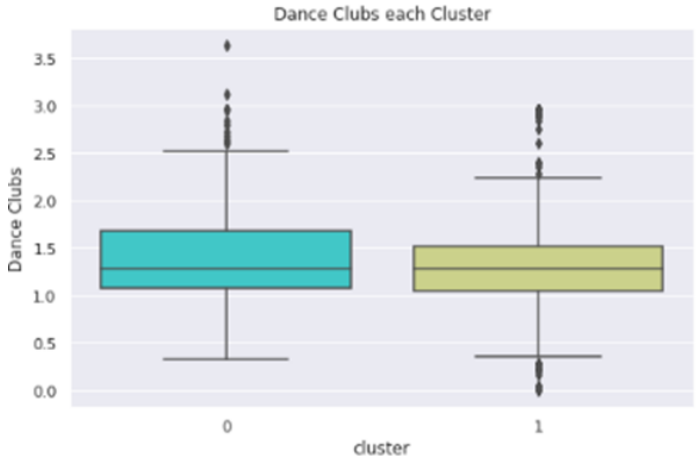
Gambar 25. Jumlah data per cluster

Setelah dilakukan pengelompokan cluster dengan menggunakan metode K-Medoids didapatkan 2 cluster yaitu cluster 1 yang memiliki jumlah data sebanyak 596 dan cluster 0 yang memiliki data sebanyak 384 data



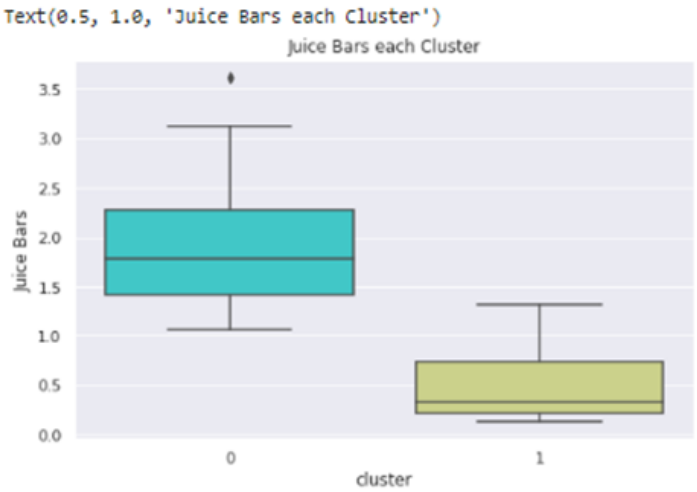
Gambar 26. Boxplot kategori Art galleries

Selanjutnya akan dibuat boxplot pada masing masing kategori untuk melihat adanya perbedaan diantara kedua cluster. Berdasarkan output yang didapatkan pada kategori Art Galleries untuk kedua cluster relatif sama serta pada masing masing cluster memiliki data outlier



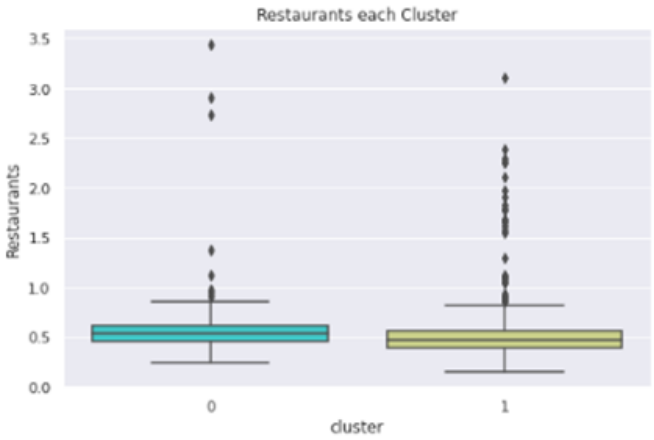
Gambar 27. Boxplot kategori Dance Clubs

Berdasarkan output yang didapatkan pada kategori Dance Clubs untuk kedua cluster relatif sama, namun pada cluster 0 memiliki lebar boxplot yang lebih besar dari cluster 1 serta pada masing masing cluster memiliki data outlier.



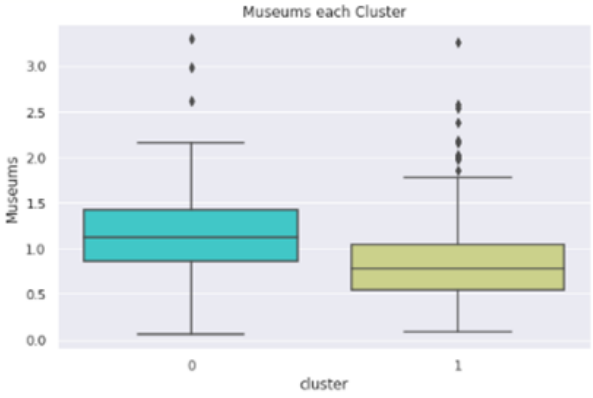
Gambar 28. Boxplot kategori Juice Bars

Berdasarkan output yang didapatkan pada kategori Juice Bars pada cluster 0 memiliki nilai Pusat, Q1, dan Q3 yang lebih tinggi dari cluster 1 serta terdapat 1 data outlier pada data cluster 0. Selain itu mean data pada cluster 1 tampak lebih dekat dengan Q1 yang menandakan bahwa distribusi tidak normal



Gambar 28. Boxplot kategori Restaurant

Berdasarkan output yang didapatkan pada kategori Restaurant untuk kedua cluster relatif sama serta pada masing masing cluster memiliki data outlier. Nilai outlier dari data cluster 1 cenderung lebih banyak daripada outlier pada cluster 0



Gambar 29. Boxplot kategori Museums

Berdasarkan output yang didapatkan pada kategori Museums pada cluster 0 memiliki nilai Pusat, Q1, dan Q3 yang lebih tinggi dari cluster 1 serta terdapat beberapa data outlier pada data cluster 0 maupun cluster 1

KESIMPULAN DAN SARAN

Berdasarkan analisis clustering yang telah dilakukan didapatkan pada clustering K-Means nilai silhouette yang terbaik adalah 2 cluster dengan nilai 0,374, pada clustering K-Medoids didapatkan nilai silhouette terbaik yaitu pada 2 cluster dengan nilai silhouette sebesar 0,371. Oleh karena itu berdasarkan nilai silhouette yang didapatkan pada kedua metode, metode clustering yang paling baik digunakan untuk dataset ini adalah metode K-Means Cluster karena memiliki nilai silhouette lebih tinggi dari metode K-Medoids dengan nilai silhouette sebesar

0,374 serta akan dibentuk 2 cluster dengan cluster 1 sebanyak 609 data dan cluster 0 sebanyak 371 data

Berdasarkan hasil analisis telah didapatkan metode clustering terbaik yaitu metode K-Means cluster. Saran yang dapat kami berikan yaitu karena user telah di cluster sesuai dengan rating yang telah diberikan pada beberapa destinasi, hasil clustering ini dapat digunakan sebagai rekomendasi untuk menentukan destinasi yang cocok dikunjungi oleh masing masing user sehingga user dapat mempersiapkan perjalanan agar lebih efektif.

DAFTAR PUSTAKA

Vandanu Aditya Putra Sangga. (2018). *Perbandingan Algoritma K-Means dan Algoritma K-Medoids Dalam Pengelompokkan Komoditas Peternakan di Provinsi Jawa Tengah Tahun 2015*. In *Tugas Akhir*. Yogyakarta: Universitas Islam Indonesia.

Jiawei, H., & Kamber, M. (2006). *Data Mining: Concept and Techniques Second Edition*. Morgan Kaufmann.

Nurjannah, Maya. (2021). *Penerapan Algoritma K-Means Untuk Analisis Data Ulasan Di Situs Tripadvisor*. Vol 3 No 1 (2021): Jurnal Responsif : Riset Sains dan Informatika.

Santoso, S. (2012). *Aplikasi SPSS pada Statistik Parametrik*. Jakarta: PT. Elex Media Komputindo.

Safe'i, A. I. (2018). *Aplikasi K-Means untuk Pengelompokan Kabupaten dan Kota Berdasarkan Produktivitas Tanaman Pangan di Provinsi Jawa Tengah tahun 2015*. In *Tugas Akhir*. Yogyakarta: Universitas Islam Indonesia.