

LAPORAN AKHIR

**Analisis *Clustering* Hotel di Pulau Bali Berdasarkan Harga, *Rating*,
Jumlah Ulasan, dan Fasilitas**



KELOMPOK: SD A1 - F

- | | |
|--------------------------|-----------|
| 1. Aqila Hana Winanggoro | 164221010 |
| 2. Herman Effendi | 164221056 |
| 3. Tika Dian Pangastuti | 164221061 |
| 4. Arkan Syafiq At'taqy | 164221062 |
| 5. Nisrina Khairunisa | 164221097 |

**TEAM-BASED PROJECT
MATA KULIAH DATA MINING I
PROGRAM STUDI TEKNOLOGI SAINS DATA
FAKULTAS TEKNOLOGI MAJU DAN MULTIDISIPLIN
UNIVERSITAS AIRLANGGA
2024**

DAFTAR ISI

DAFTAR ISI.....	1
DAFTAR TABEL.....	3
DAFTAR GAMBAR.....	4
BAB I.....	5
PENDAHULUAN.....	5
1.1 Latar Belakang.....	5
1.2 Rumusan Masalah.....	5
1.3 Tujuan.....	5
1.4 Manfaat.....	6
BAB II.....	7
TINJAUAN PUSTAKA.....	7
2.1 Hotel.....	7
2.2 Traveloka.....	7
2.3 Feature Scaling.....	7
2.4 Elbow Method.....	8
2.5 Silhouette Score.....	8
2.6 K-means Clustering.....	8
2.7 DBSCAN (Density-Based Spatial Clustering of Applications with Noise).....	9
BAB III.....	10
METODOLOGI.....	10
3.1 Sumber Data.....	10
3.2 Deskripsi Data.....	10
3.3 Metode Penelitian.....	10
BAB IV.....	14
HASIL DAN PEMBAHASAN.....	14
4.1 Pemahaman Data.....	14
4.2 Exploratory Data Analysis.....	16
4.3 Data Preprocessing.....	17
4.4 Data Clustering.....	20
BAB V.....	25
KESIMPULAN DAN SARAN.....	25
5.1 Kesimpulan.....	25
5.2 Saran.....	25
DAFTAR PUSTAKA.....	27
LAMPIRAN.....	28

DAFTAR TABEL

Tabel 4.1 Hasil Statistika Deskriptif.....	13
Tabel 4.2 Hasil Clustering K-Means.....	21
Tabel 4.3 Hasil Clustering DBSCAN.....	23

DAFTAR GAMBAR

Gambar 3.1 Dataset.....	9
Gambar 3.2 Dataset Setelah Preprocessing.....	10
Gambar 3.3 Hasil Penerapan PCA.....	10
Gambar 4.1 Data yang Terduplicat.....	13
Gambar 4.2 Outlier pada Data.....	14
Gambar 4.3 Nilai Hilang pada Data.....	14
Gambar 4.4 Heatmap Nilai Korelasi.....	15
Gambar 4.5 Grafik Distribusi.....	15
Gambar 4.6 Hasil Preprocessing Data Duplikat.....	16
Gambar 4.7 Dataset Hasil Preprocessing Kolom Teks.....	16
Gambar 4.8 Hasil Feature Engineering.....	17
Gambar 4.9 Hasil Feature Scaling.....	17
Gambar 4.10 Grafik Hasil Komponen Optimal.....	18
Gambar 4.11 Hasil Komponen Optimal.....	18
Gambar 4.12 Visualisasi Elbow Method.....	19
Gambar 4.13 Silhouette Score.....	20
Gambar 4.14 Hasil K-Means Clustering 2D.....	20
Gambar 4.15 Hasil DBSCAN Clustering 2D.....	22

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pariwisata telah menjadi primadona Negara Indonesia sejak dulu, terutama di wilayah Pulau Bali. Pulau Bali telah banyak menarik wisatawan dari penjuru dunia dan terus mengalami peningkatan dari waktu ke waktu. Dengan pesatnya pertumbuhan pariwisata ini, tentunya permintaan akomodasi yang berkualitas juga meningkat secara linear. Oleh karena itu, untuk menghadapi jumlah permintaan akomodasi, tentunya dibutuhkan pemahaman mengenai karakteristik hotel di Bali guna memenuhi kebutuhan dan preferensi wisatawan serta mendukung industri pariwisata secara

Meskipun Pulau Bali menawarkan berbagai pilihan akomodasi, informasi yang terperinci tentang perbedaan antara hotel-hotel tersebut masih terbatas. Oleh karena itu, analisis klaster hotel yang didasarkan pada harga, rating, jumlah review dan jumlah fasilitas menjadi penting untuk memberikan pemahaman yang lebih baik untuk pasar akomodasi di Bali. Dengan demikian, projek ini bertujuan untuk mengisi kesenjangan informasi dan menyediakan wawasan yang bermanfaat bagi wisatawan, penyedia hotel dan pemangku kepentingan pariwisata lainnya.

1.2 Rumusan Masalah

Dalam proyek ini, beberapa rumusan masalah yang relevan dapat diajukan, antara lain:

1. Bagaimana cara memetakan dan menganalisis klaster hotel di Bali berdasarkan harga, rating, review, dan fasilitas?
2. Bagaimana pengaruh variabel-variabel tersebut terhadap klaster hotel di Bali?
3. Apa saja karakteristik masing-masing klaster hotel yang dihasilkan dari analisis tersebut?

1.3 Tujuan

Tujuan dari proyek ini adalah:

1. Menyelidiki pola-pola klaster hotel di Bali berdasarkan variabel harga, rating, review, dan fasilitas.
2. Mengidentifikasi perbedaan karakteristik antara klaster hotel yang dihasilkan.
3. Membuat informasi yang penting kepada wisatawan terhadap karakteristik hotel yang ada.

1.4 Manfaat

Proyek ini diharapkan akan memberikan manfaat sebagai berikut:

1. Mendeskripsikan karakteristik tiap klaster hotel kepada wisatawan agar sesuai dengan preferensi wisatawan.
2. Membantu pengelola hotel dalam memahami preferensi dan kebutuhan pelanggan mereka.
3. Memberikan informasi kepada pemangku kepentingan pariwisata untuk perencanaan dan pengembangan sektor pariwisata di Bali.

BAB II

TINJAUAN PUSTAKA

2.1 Hotel

Berdasarkan data dari BPS tahun 2019, Provinsi Bali menempati posisi pertama dengan jumlah akomodasi terbanyak. Hal ini juga didukung dengan masyarakat lokal yang juga terlibat langsung dalam kegiatan pariwisata yang lambat laun akan meningkatkan kesejahteraan masyarakat. Sehingga Provinsi Bali sangat tergantung perekonomiannya pada sektor pariwisata.

Untuk mendukung pariwisata di Bali, dibangun beragam akomodasi hotel dengan fasilitas yang memadai. Fasilitas yang memadai ini akan meningkatkan kualitas pelayanan terhadap tamu dapat menggerakan roda bisnis perhotelan. Pengaruh kualitas pelayanan serta fasilitas hotel menjadi daya tarik yang kuat serta dapat mendorong tercapainya kepuasan tersendiri bagi banyak konsumen.

2.2 Traveloka

Traveloka merupakan sebuah perusahaan yang menyediakan berbagai kebutuhan perjalanan yang tersedia dalam satu platform. Traveloka dapat diakses melalui aplikasi mobile maupun melalui *website* traveloka.com. Perusahaan yang baru berdiri pada Bulan Oktober 2012 ini memiliki rata-rata pengunjung *website* sekitar 20 ribuan perhari yang menjadikannya sangat terkenal di Asia Tenggara, termasuk Indonesia.

Traveloka menyediakan pelayanan jasa, seperti pemesanan hotel *online*, pemesanan tiket pesawat, tiket kereta api, bus, rekreasi, serta hiburan lainnya. *Website* traveloka.com digunakan sebagai pendukung dalam kemajuan serta pengembangan traveloka menjadi sebuah media untuk mendapatkan informasi pemesanan transportasi maupun akomodasi secara cepat, tepat, efektif, dan efisien.

2.3 Feature Scaling

Feature scaling merupakan suatu proses dalam preprocessing data yang penting dalam algoritma *machine learning*. Proses ini bertujuan untuk menstandarkan *range* yang besar pada sebuah variabel data sebelum dimasukan ke dalam model. Proses ini menjadi penting karena algoritma *machine learning* bekerja lebih baik dan cepat ketika variabel berada pada skala yang mirip dan mendekati distribusi normal. Dengan adanya proses *feature scaling* juga dapat mengurangi kemungkinan suatu fitur mendominasi karena memiliki skala yang besar dibandingkan yang lain. Beberapa jenis feature scaling, yaitu;

- *Standardization*

Metode ini mengasumsikan data berdistribusi normal dengan menghitung nilai rata-rata dan deviasi standar dengan persamaan normalisasi *Z-score*. Scaling ini tidak kuat terhadap outlier karena akan berpengaruh pada saat menghitung rata-rata dan deviasi standar.

- *Min-Max Scaler*

Metode ini mentransformasi data menjadi memiliki range yang kecil, umumnya [0,1] atau [-1,1] jika terdapat nilai negatif pada data. *Scaling* ini dapat

optimal ketika data tidak berdistribusi Gaussian dan deviasi standarnya sangat kecil.

- *Robust Scaler*

Metode ini untuk menstandarkan skala data dengan menghitung median dan IQR. Median dan skala data dihilangkan sesuai dengan rentang kuartil sehingga dapat tahan terhadap outlier pada data.

2.4 Elbow Method

Elbow method atau metode siku adalah teknik yang umum digunakan dalam algoritma pengelompokan (*clustering*) untuk menentukan jumlah *cluster* yang optimal pada suatu dataset. Metode ini akan menghitung jumlah kuadrat dari kesalahan *intra-cluster* pada setiap klaster dan kemudian mengidentifikasi titik yang memiliki nilai menurun tajam. Titik yang nilainya menurun tersebut menunjukkan jumlah *cluster* yang optimal (Hadanny et al., 2022).

Teknik ini dapat diterapkan dalam berbagai disiplin ilmu dengan konteks yang berbeda-beda juga. Para peneliti menggunakan metode ini untuk menentukan jumlah klaster untuk mengetahui segmentasi pelanggan dalam analisis pasar (Alamsyah et al., 2022). dengan metode siku yang membantu dalam memilih jumlah *cluster optimal*, maka metode ini dapat meningkatkan kinerja algoritma pengelompokan seperti K-means.

2.5 Silhouette Score

Pada penentuan jumlah *cluster* yang optimal tidak ada satu metode yang dapat cocok dengan semua data, tergantung pada perhitungan kemiripan antar data. Mengingat pada hirarki *clustering*, penentuan nilai k dapat cenderung subjektif, diperlukan metode penentuan nilai k yang dapat dibuktikan, salah satunya yaitu *silhouette score*.

Nilai *silhouette* mendekati +1 membuktikan bahwa titik data berada pada klaster yang tepat. Pada nilai *silhouette* mendekati 0, membuktikan titik data mungkin saja berada pada klaster yang lain. Sedangkan pada nilai *silhouette* yang mendekati -1, maka membuktikan bahwa titik data berada klaster yang salah. Dapat dikatakan bahwa nilai *silhouette* yang tinggi mengindikasikan *clustering* yang baik sehingga dapat membantu dalam menentukan nilai k yang optimal.

2.6 K-means Clustering

Clustering merupakan sebuah metode mencari klaster atau kelompok dari sebuah dataset yang dikelompokan berdasarkan kemiripan yang paling dekat. Pada pendekatan non parametrik, metode *clustering* sering kali dihitung berdasarkan kemiripan antar data. Hal ini dapat terbagi menjadi dua metode, yaitu metode hirarki dan partisional dimana metode partisional lebih umum digunakan.

Salah satu algoritma pada metode *partitional* yang populer adalah *K-means clustering*. Secara umum, algoritma k-means menghitung jarak antar data untuk menentukan kesamaannya. Dimana yang memiliki kemiripan tinggi akan berada

dalam *cluster* yang sama, sedangkan yang tidak memiliki kesamaan akan berada pada *cluster* yang berbeda. Cara kerja dari algoritma ini yaitu mencari pusat *cluster* (*centroid*) secara acak yang kemudian dihitung jarak masing-masing titik data ke *centroid* tersebut dan titik-titik tersebut dikelompokkan berdasarkan jarak ke *centroid* terdekat. Kumpulan dari titik dari centroid yang sama itu membentuk *cluster*. *Centroid* akan terus diperbarui dan mencapai optimal ketika tidak ada perubahan lagi pada titik data dalam *cluster*.

Algoritma ini dikenal dengan keefisienan yang baik sehingga dapat diterapkan dalam berbagai bidang yang luas karena sifatnya yang sederhana dan linear. Dibandingkan dengan metode lainnya seperti spectral clustering, DBSCAN, dan mean-shift clustering, k-means dipilih karena kecepatan komputasinya yang lebih baik. Jika diterapkan dalam konteks segmentasi berupa gambar, algoritma ini terbukti memberikan hasil yang optimal. Dengan tambahan teknik median-cut untuk memperbaiki proses segmentasi lebih lanjut (Rosado et al., 2016).

2.7 DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*)

DBSCAN merupakan algoritma *clustering* berbasis kepadatan atau kerapatan. Kepadatan yang terkait satu sama lain ini dihitung jumlah titik dalam suatu wilayahnya dengan jari-jari yang ditentukan di sekitar titik tersebut. Titik-titik dengan kepadatan yang berada di atas ambang batas yang ditentukan akan dikelompokkan dalam satu *cluster*. Pada Algoritma DBSCAN, jumlah *cluster* yang terbentuk bergantung kepada dua parameter. Parameter yang pertama, yaitu epsilon (ϵ) yang akan menjadi nilai dari jari-jari pada sebuah *cluster*. Selain itu terdapat parameter *minimum points* atau minimal banyaknya data pada suatu *cluster*.

Algoritma ini menghasilkan tiga jenis status dari setiap data, di antara nya adalah noise sebagai titik data yang tidak terjangkau *cluster*, core sebagai titik pusat pada klaster, dan border sebagai titik batas luar dari sebuah *cluster*. DBSCAN cukup berbeda jika dibandingkan dengan clustering lainnya. Hal ini dikarenakan kemampuannya dalam melakukan analisis pada dataset yang besar dan tidak memerlukan penentuan jumlah *cluster* terlebih dahulu.

BAB III

METODOLOGI

3.1 Sumber Data

Sumber data : Data yang digunakan bersumber dari hasil scraping pada website Traveloka (<https://www.traveloka.com>), di mana kami mengambil informasi tentang hotel-hotel di Bali pada tanggal 25 Mei 2024 sampai 26 Mei 2024.

3.2 Deskripsi Data

Dataset berisi informasi tentang hotel-hotel di Bali. Berikut adalah penjelasan untuk setiap variabel:

- Nama: Nama hotel
- Harga: Harga per malam di hotel, ditampilkan dalam Rupiah.
- Reviews: Jumlah ulasan yang diterima oleh hotel
- Rating: Rating hotel
- Facilities: Fasilitas yang tersedia di hotel

	nama	harga	reviews	rating	facilities
0	ASTON Kuta Hotel & Residence	Rp 734.194	(2.4K reviews)	8.6 (2.4K reviews)	Hot tub\nKids club\nMassage\nFitness center\nP...
1	Brits Hotel Legian	Rp 422.004	(2.1K reviews)	8.6 (2.1K reviews)	Top picked by Family with children\nPay at Hot...
2	Harper Kuta by ASTON	Rp 650.826	(2.9K reviews)	8.6 (2.9K reviews)	Pay at Hotel\nMassage\nAirport transfer\nSauna...
3	Atanaya Kuta Bali	Rp 670.000	(10K reviews)	8.6 (10K reviews)	Babysitting\nPoolside bar\nWheelchair accessib...
4	Wyndham Garden Kuta Beach Bali	Rp 1.167.769	(1.4K reviews)	8.1 (1.4K reviews)	Pay at Hotel\nBabysitting\nBicycle rental\nCar...
...
259	Hotel Karthi	Rp 289.256	(693 reviews)	8.4 (693 reviews)	Massage
260	Bali Bungalo	Rp 590.909	(58 reviews)	8.1 (58 reviews)	Bicycle storage\nClothes dryer\nBicycle rental...
261	Cattleya Pool Suite - Seminyak by Marbella	Rp 910.301	(219 reviews)	8.2 (219 reviews)	Pay at Hotel\nBanana boat\nBeach volleyball\nB...
262	The Akasha Luxury Villas	Rp 3.910.050	(37 reviews)	8.7 (37 reviews)	Surfing\nHorse riding\nMassage\nKitchenette\nE...
263	Interconnection Hostel Kuta	Rp 460.542	(70 reviews)	8.8 (70 reviews)	Car rental

Gambar 3.1 Dataset

3.3 Metode Penelitian

Langkah-langkah yang dilakukan dalam pembuatan model clustering adalah sebagai berikut:

1. Data *Preprocessing*
 - Menangani Missing Value: Menggunakan metode “.isna()” dari Pandas untuk mengidentifikasi missing value dalam data.
 - Label Encoding: Mengubah nilai-nilai kategorik pada variabel “facilities” menjadi numerik menggunakan label encoding dan menghitung jumlah total fasilitas tersebut.
 - Penanganan Outlier: Menggunakan Robust Scaler untuk menangani outlier pada variabel “harga” dan “reviews”. Robust Scaler menggunakan median dan interquartile range (IQR) untuk menangani outlier, sehingga lebih robust terhadap nilai ekstrem.
 - Feature Scaling: Standar Scaler digunakan untuk standarisasi variabel “rating” dan “total_facilities”, sementara Min-Max Scaler digunakan untuk melakukan scaling pada variabel “total_facilities”. Hal ini diperlukan agar variabel dengan skala yang berbeda memiliki pengaruh

yang seimbang dalam proses clustering. Hasil dari penerapan feature scaling adalah sebagai berikut:

	harga	reviews	rating	total_facilities
0	0.061511	-0.213966	0.345248	1.0
1	-0.217362	-0.214804	0.345248	1.0
2	-0.012960	-0.212570	0.345248	0.8
3	0.004168	-0.192737	0.345248	0.4
4	0.448815	-0.216760	-1.078900	1.0

Gambar 3.2 Dataset Setelah Preprocessing

2. *Exploratory Data Analysis (EDA)*

- Korelasi Heatmap: Menghitung dan memvisualisasikan koefisien korelasi antara setiap pasangan variabel numerik dengan heatmap menggunakan metode pearson dengan library “pandas” dan “seaborn” serta “matplotlib.pyplot” untuk memplot heatmap.
- Boxplot: Menggunakan library “seaborn” dan “matplotlib.pyplot” serta menambahkan loop “for”. Hal ini untuk mendeteksi outlier dan memvisualisasikan distribusi data secara grafis.
- Distribusi plot: Menggunakan library “seaborn” dan “matplotlib.pyplot” serta menambahkan loop “for”. Hal ini untuk melihat distribusi frekuensi nilai-nilai dalam suatu variabel.
- Pair plot: Menggunakan library “seaborn” untuk semua pasangan variabel numerik, dengan pewarnaan (hue) berdasarkan jumlah total facilities. Hal ini membantu dalam melihat pola hubungan antara variabel secara keseluruhan.

3. *Principal Component Analysis (PCA)*

- Penerapan PCA: Reduksi dimensi data dan memilih komponen utama yang signifikan. PCA dilakukan dengan melatih model pada seluruh variabel dan memvisualisasikan plot varians pada setiap komponen untuk menentukan jumlah komponen optimal yang menjelaskan setidaknya 95% dari total variansi data.
- Transformasi data: Data akan ditransformasi dan digabung dengan komponen yang baru dan menghasilkan variabel PC1 dan PC2. Hasil dari penerapan PCA adalah sebagai berikut:

	PC1	PC2
harga	1.419780	0.098510
reviews	-0.374905	0.996036
rating	0.508536	0.454244
total_facilities	0.035955	0.071112

Gambar 3.3 Hasil Penerapan PCA

4. Implementasi *K-Means Clustering*

- Pemilihan parameter K: Menggunakan metode elbow dengan library scikit-learn, Yellowbrick, dan matplotlib untuk menentukan cluster optimal dengan memplot jumlah cluster versus sum of squared distance (inertia) dan mencari nilai elbow dimana penurunan inertia mulai melambat.
- Silhouette score: Digunakan untuk mengevaluasi seberapa baik cluster terbentuk dengan library scikit-learn dan matplotlib. Nilai silhouette untuk setiap sampel akan dihitung dan divisualisasikan dalam silhouette plot yang menampilkan seberapa baik setiap sampel cocok dengan cluster mereka sendiri dibandingkan dengan cluster lain.
- Proses clustering: Menggunakan metode K-Means dari library scikit-learn dengan memilih variabel "harga", "rating", dan "total_facilities" berdasarkan heatmap korelasi yang menunjukkan hubungan yang kuat. Jumlah cluster yang digunakan adalah 5, hasil dari analisis elbow yang telah dilakukan sebelumnya.
- Pelabelan cluster: Pelabelan cluster dan visualisasi plot 2D serta 3D menggunakan Matplotlib bertujuan untuk membantu memahami struktur dan distribusi cluster dalam data yang telah direduksi. Plot 2D menampilkan titik-titik data dalam dua dimensi dengan warna yang berbeda untuk setiap cluster, serta centroid cluster yang ditampilkan dengan bentuk bintang. Hasil visualisasi ini memberikan gambaran yang lebih jelas tentang struktur dan distribusi cluster dalam data, memudahkan dalam mendekripsi pola dan anomali dalam data yang telah direduksi.

5. Implementasi DBSCAN

- Penentuan nilai eps optimal: Menggunakan k-distance graph dengan algoritma Kneedle untuk menemukan elbow point pada grafik jarak k-terdekat yang dihitung dari data menunjukkan perubahan paling signifikan dalam grafik jarak dan digunakan sebagai nilai eps optimal.
- Penerapan algoritma DBSCAN: Menggunakan variabel 'harga', 'rating', dan 'total_facilities', DBSCAN mengelompokkan titik data berdasarkan kepadatan. Titik yang tidak cukup dekat dengan titik lain (berdasarkan eps) dan tidak memenuhi jumlah minimum sampel akan dianggap sebagai noise. Minimum sample adalah parameter yang menentukan jumlah titik data minimum yang harus ada dalam radius eps agar suatu titik dianggap sebagai core point. Jika suatu titik adalah core point, maka semua titik dalam jarak eps dari titik tersebut dianggap sebagai bagian dari cluster yang sama.
- Penentuan jumlah cluster: Jumlah cluster ditentukan dari jumlah label unik yang dihasilkan oleh DBSCAN, dikurangi satu jika terdapat label -1 yang menunjukkan noise. Jumlah noise adalah jumlah titik yang diberi label -1.

- Visualisasi hasil clustering: Menggunakan scatter plot untuk memplot titik-titik data dan Convex Hull untuk menggambar batas cluster untuk visualisasi 2D. Hasil visualisasi ini berfungsi untuk memahami struktur cluster yang terbentuk dan mendeteksi outlier dalam data, serta untuk menemukan cluster dengan bentuk yang tidak beraturan dan mendeteksi outlier secara efektif tanpa perlu menentukan jumlah cluster sebelumnya.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Pemahaman Data

Melakukan analisis terhadap kumpulan data yang diberikan dengan melakukan pengecekan statistika deskriptif, duplikat data, data yang hilang, dan penciran data.

4.1.1 Statistika Deskriptif

	Nama Hotel	Harga Hotel	Reviews	Rating Hotel	Fasilitas Hotel
Jumlah	264	264	264	264	264
Nilai unik	255	243	187	242	223
Terbanyak	Coast Boutique Hotel	Rp 247.934	(1.5K reviews)	8.5 (54 reviews)	Pay at Hotel
Frekuensi	2	5	7	2	8

Tabel 4.1 Hasil Statistika Deskriptif

Dari tabel statistika deskriptif diatas, diperoleh jumlah data nya sebanyak 264 pada variabel nama hotel, harga hotel, jumlah review, rating hotel, serta fasilitas hotel. Pada variabel harga, terdapat paling banyak 5 hotel yang memiliki harga yang sama, yaitu Rp247.934. Rating terbanyak pada data ini adalah 8,5 sedangkan jumlah orang yang me-review hotel paling banyak adalah 1500 review. Pada fasilitas hotel, paling banyak merupakan *pay at hotel* atau dapat melakukan pembayaran langsung ketika sampai di hotel tersebut.

4.1.2 Pengecekan Data Duplikat

```

Jumlah data duplikat: 9
          nama      harga      reviews  \
245      Coast Boutique Hotel  Rp 810.114  (151 reviews)
246      Away Bali Legian Camakila Resort  Rp 2.258.924  (54 reviews)
247      The Bandha Hotel & Suites  Rp 6.832.244  (68 reviews)
248      Destiny Villas and Residence  Rp 2.640.000  (45 reviews)
249      Ping Hotel  Rp 458.182  (244 reviews)
250      Sri Krisna II  Rp 351.240  (67 reviews)
251      Kelan-Tel Bali  Rp 304.132  (34 reviews)
252  Sun Island Boutique Villas & Spa Seminyak  Rp 2.541.322  (11 reviews)
253      Amadea Resort & Villas Seminyak Bali  Rp 1.444.724  (79 reviews)

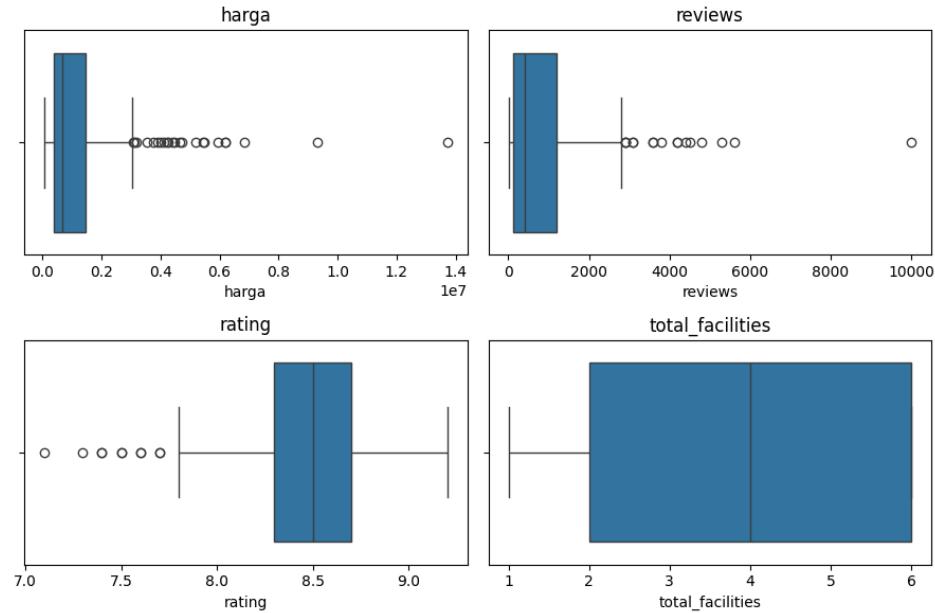
          rating           facilities
245  8.7 (151 reviews)  Pay at Hotel\nAirport transfer
246  8.5 (54 reviews)  Kids club\nMassage\nFitness center\nWheelchair...
247  8.9 (68 reviews)  Hot tub\nBeach towel\nMassage\nFitness center\n2+
248  8.6 (45 reviews)  Babysitting\nBicycle rental\nCar rental\nMassa...
249  8.4 (244 reviews)  Pay at Hotel\nCar rental
250  8.5 (67 reviews)  Pay at Hotel\nBraille or raised signage\nHot t...
251  7.4 (34 reviews)  Bicycle rental\nCar rental\nPool sun loungers\...
252  8.5 (11 reviews)  Pay at Hotel\nCar rental\nMassage\nKitchenette...
253  8.7 (79 reviews)  Massage\nSpa\nBar

```

Gambar 4.1 Data yang Terduplikat

Dari hasil kode diatas didapat bahwa terdapat 9 data duplikat yang terdapat pada baris ke 245-253, sehingga pada preprocessing data nantinya data duplikat ini akan dihapus.

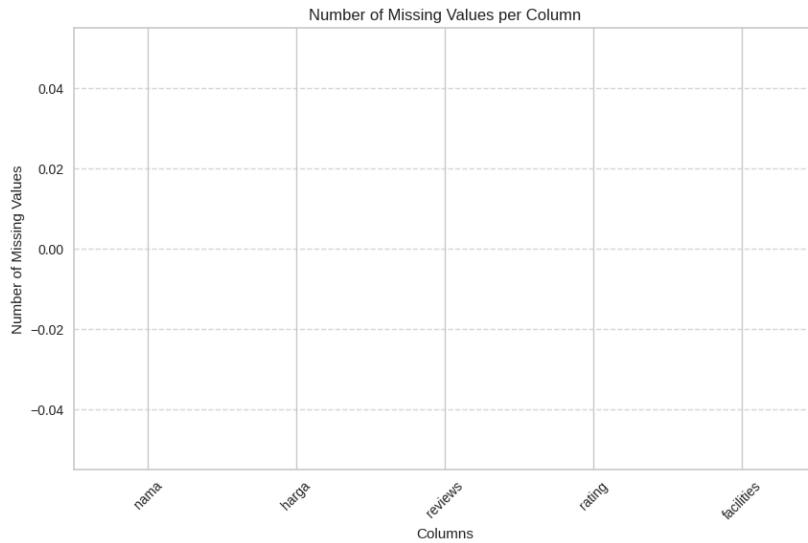
4.1.3 Pengecekan Outlier



Gambar 4.2 Outlier pada Data

Dari visualisasi diatas didapat bahwa pada variabel ‘harga’, ‘total_facilities’, dan ‘rating’ terdapat banyak outlier, sedangkan pada variabel ‘total_facilities’ tidak terdapat outlier.

4.1.4 Pengecekan Missing Values



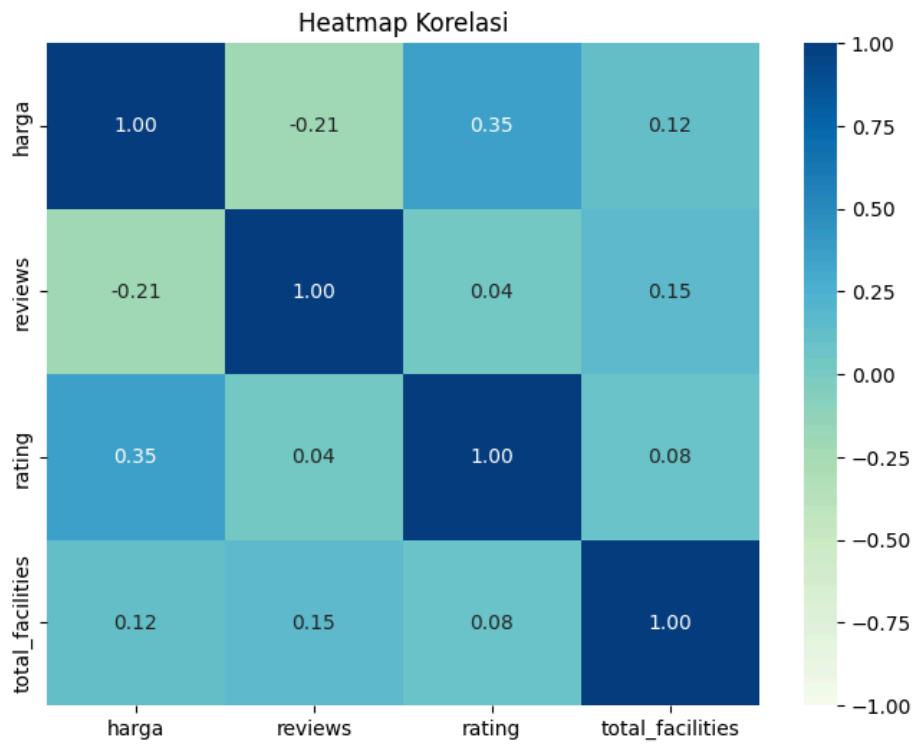
Gambar 4.3 Nilai Hilang pada Data

Pada visualisasi diatas, dapat dilihat bahwa tidak terdapat *missing values* pada tiap-tiap variabel yang terdapat pada data.

4.2 Exploratory Data Analysis

Melakukan visualisasi data menggunakan korelasi heatmap dan distribusi plot untuk memahami hubungan antar variabel

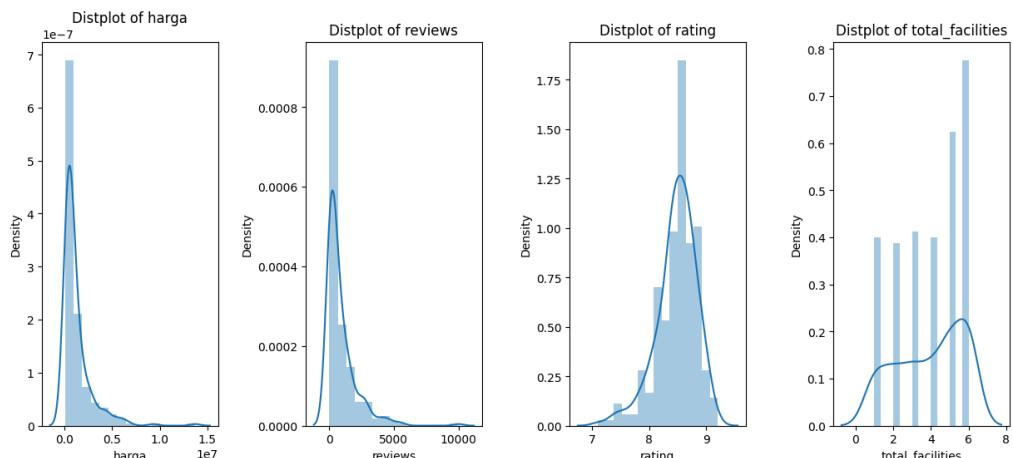
4.2.1 Korelasi Heatmap



Gambar 4.4 Heatmap Nilai Korelasi

Dari korelasi heatmap diatas didapat bahwa variabel ‘harga’ dan ‘rating’ memiliki korelasi tertinggi dengan nilai korelasi 0.35, dan variabel ‘rating’ dan ‘reviews’ memiliki korelasi terendah dengan nilai korelasi 0.04.

4.2.2 Distribution Plot



Gambar 4.5 Grafik Distribusi

Dari tabel diatas didapat bahwa variabel ‘harga’ dan ‘reviews’ cenderung skewed ke kanan, pada variabel ‘total_facilities’ data cenderung *skewed* ke kiri, dan pada variabel ‘rating’ data mendekati kurva lonceng atau berdistribusi normal.

4.3 Data *Preprocessing*

Mengatasi nilai duplikat pada data, menangani teks agar bisa dilakukan clustering, membuat variabel baru, melakukan scaling pada data, serta mereduksi dimensi data agar dapat dilakukan analisis clustering.

4.3.1 Mengatasi Data Duplikat

```
▶ # Menghapus data duplikat
df= df.drop_duplicates()

# Menampilkan jumlah baris setelah menghapus duplikat
print(f"Jumlah baris setelah menghapus duplikat: {df.shape[0]}")
```

➡ Jumlah baris setelah menghapus duplikat: 255

Gambar 4.6 Hasil Preprocessing Data Duplikat

Dari code diatas dilakukan ‘drop’ yaitu penghapusan baris pada data duplikat, sehingga data yang awalnya berjumlah 264 baris berkurang menjadi 255 baris.

4.3.2 Mengatasi Kolom Teks

	nama	harga	reviews	rating	facilities
0	ASTON Kuta Hotel & Residence	734194	2400	8.6	[22, 23, 31, 19, 38, 43]
1	Brits Hotel Legian	422004	2100	8.6	[48, 34]
2	Harper Kuta by ASTON	650826	2900	8.6	[34, 31, 1, 43, 3]
3	Atanaya Kuta Bali	670000	10000	8.6	[2, 38, 37]
4	Wyndham Garden Kuta Beach Bali	1167769	1400	8.1	[34, 2, 10, 13, 31]
...
259	Hotel Karthi	289256	693	8.4	[31]
260	Bali Bungalo	590909	58	8.1	[11, 15, 10, 13]
261	Cattleya Pool Suite - Seminyak by Marbella	910301	219	8.2	[34]
262	The Akasha Luxury Villas	3910050	37	8.7	[45, 31, 24, 17]
263	Interconnection Hostel Kuta	460542	70	8.8	[13]
255 rows × 6 columns					

Gambar 4.7 Dataset Hasil Preprocessing Kolom Teks

Dari tabel diatas dilakukan text preprocess dan mapping pada variabel ‘harga’, ‘rating’, dan ‘facilities’ dari yang sebelumnya memuat data berupa teks menjadi data numerik, sehingga bisa dilakukan analisis klastering.

4.3.3 Feature Engineering

	nama	harga	reviews	rating	facilities	total_facilities
0	ASTON Kuta Hotel & Residence	734194	2400	8.6	[22, 23, 31, 19, 38, 43]	6
1	Brits Hotel Legian	422004	2100	8.6	[48, 34]	6
2	Harper Kuta by ASTON	650826	2900	8.6	[34, 31, 1, 43, 3]	5
3	Atanaya Kuta Bali	670000	10000	8.6	[2, 38, 37]	3
4	Wyndham Garden Kuta Beach Bali	1167769	1400	8.1	[34, 2, 10, 13, 31]	6
...
259	Hotel Karthi	289256	693	8.4	[31]	1
260	Bali Bungalo	590909	58	8.1	[11, 15, 10, 13]	6
261	Cattleya Pool Suite - Seminyak by Marbella	910301	219	8.2	[34]	3
262	The Akasha Luxury Villas	3910050	37	8.7	[45, 31, 24, 17]	5
263	Interconnection Hostel Kuta	460542	70	8.8	[13]	1

255 rows × 6 columns

Gambar 4.8 Hasil Feature Engineering

Dari tabel diatas dilakukan feature engineering sehingga mendapatkan variabel baru yaitu ‘total_facilities’ yang memuat jumlah fasilitas pada sebuah hotel yang berguna untuk analisis *clustering*.

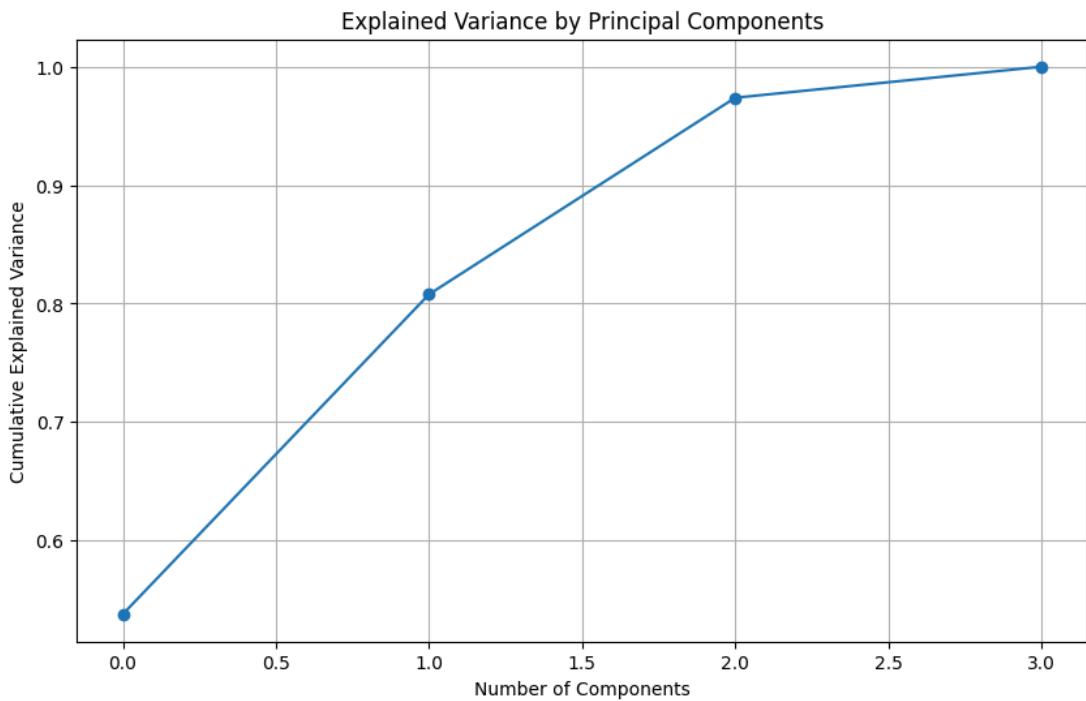
4.3.4 Feature Scaling

	harga	reviews	rating	total_facilities
0	0.065847	1.837263	0.345972	1.0
1	-0.224777	1.559871	0.345972	1.0
2	-0.011762	2.299584	0.345972	0.8
3	0.006087	8.864540	0.345972	0.4
4	0.469471	0.912621	-1.086218	1.0

Gambar 4.9 Hasil Feature Scaling

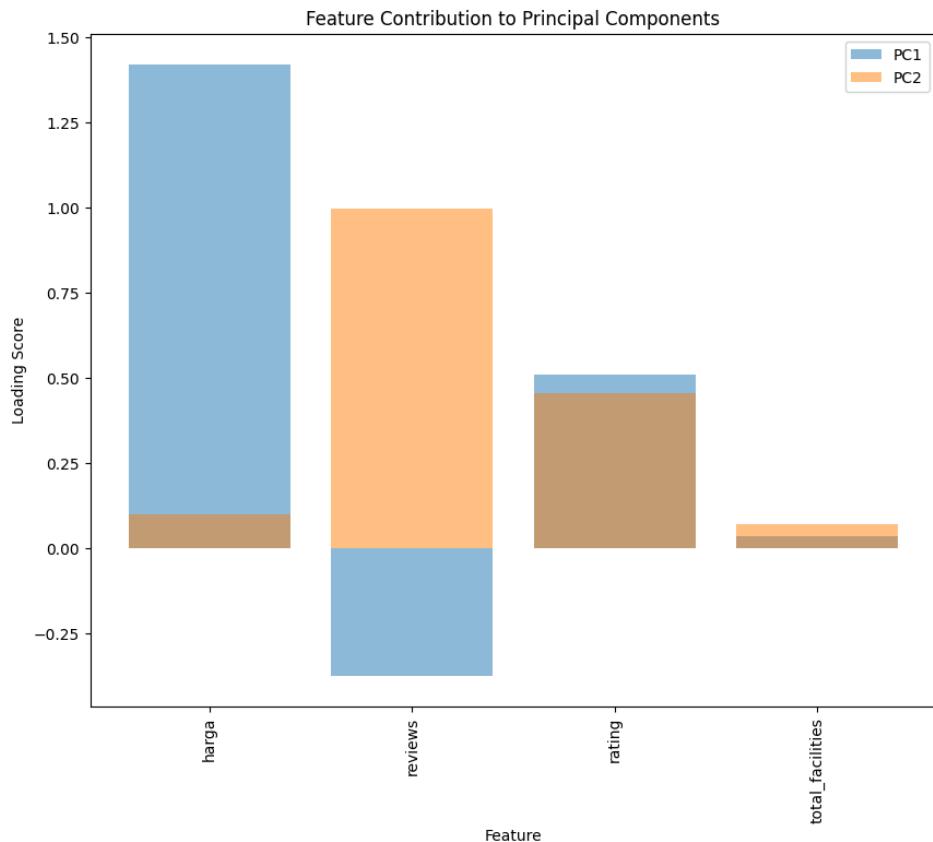
Dari tabel diatas dilakukan *robust scaling* pada variabel ‘harga’ dan ‘reviews’ dikarenakan terdapat outlier. Kemudian menggunakan *standard scaling* pada variabel ‘rating’ dikarenakan distribusi data mendekati distribusi normal. Dan terakhir menggunakan min-max *scaling* pada variabel ‘total_facilities’ agar didapatkan nilai pada rentang 0-1.

4.3.5 Principal Component Analysis (PCA)



Gambar 4.10 Grafik Hasil Komponen Optimal

Berdasarkan visualisasi diatas didapat bahwa $n = 2$ karena telah menjelaskan setidaknya 95% variansi.



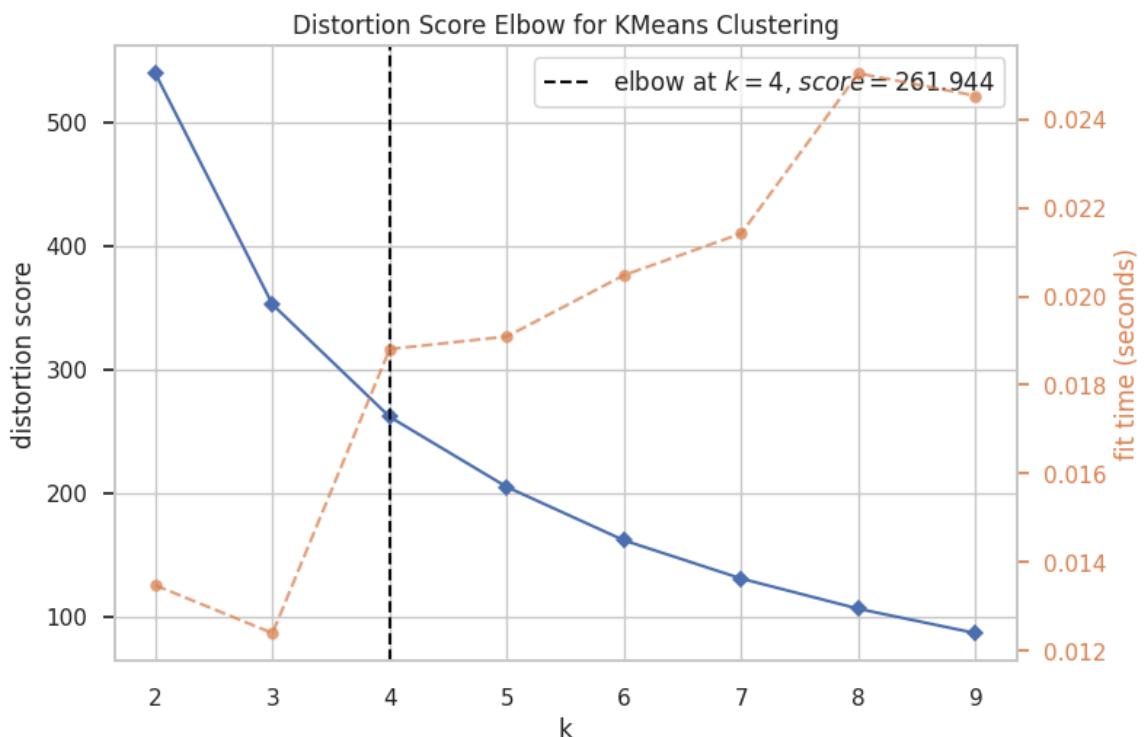
Gambar 4.11 Hasil Komponen Optimal

Berdasarkan visualisasi diatas PC1 merepresentasikan banyak kontribusi dari variabel ‘harga’, dengan sedikit kontribusi dari variabel ‘reviews’ dan ‘rating’. Sedangkan PC2 merepresentasikan banyak kontribusi dari variabel ‘reviews’, dengan sedikit kontribusi dari variabel ‘rating’, ‘harga’ dan ‘total_facilities’.

4.4 Data *Clustering*

Melakukan analisis *clustering* terhadap data dengan melakukan pengecekan *elbow method*, pengecekan *silhouette score*, kemudian analisis *clustering* menggunakan metode K-Means dan DBSCAN yang divisualisasikan secara 2 dimensi.

4.4.1 Elbow Method



Gambar 4.12 Visualisasi *Elbow Method*

Berdasarkan visualisasi metode elbow, jumlah cluster yang optimal adalah 4 dengan skor distorsi 261.944 yang ditandai dengan garis vertikal putus-putus. Penurunan skor distorsi yang mulai melambat, menunjukkan bahwa menambah lebih banyak cluster, tidak memberikan peningkatan yang signifikan dalam kualitas cluster.

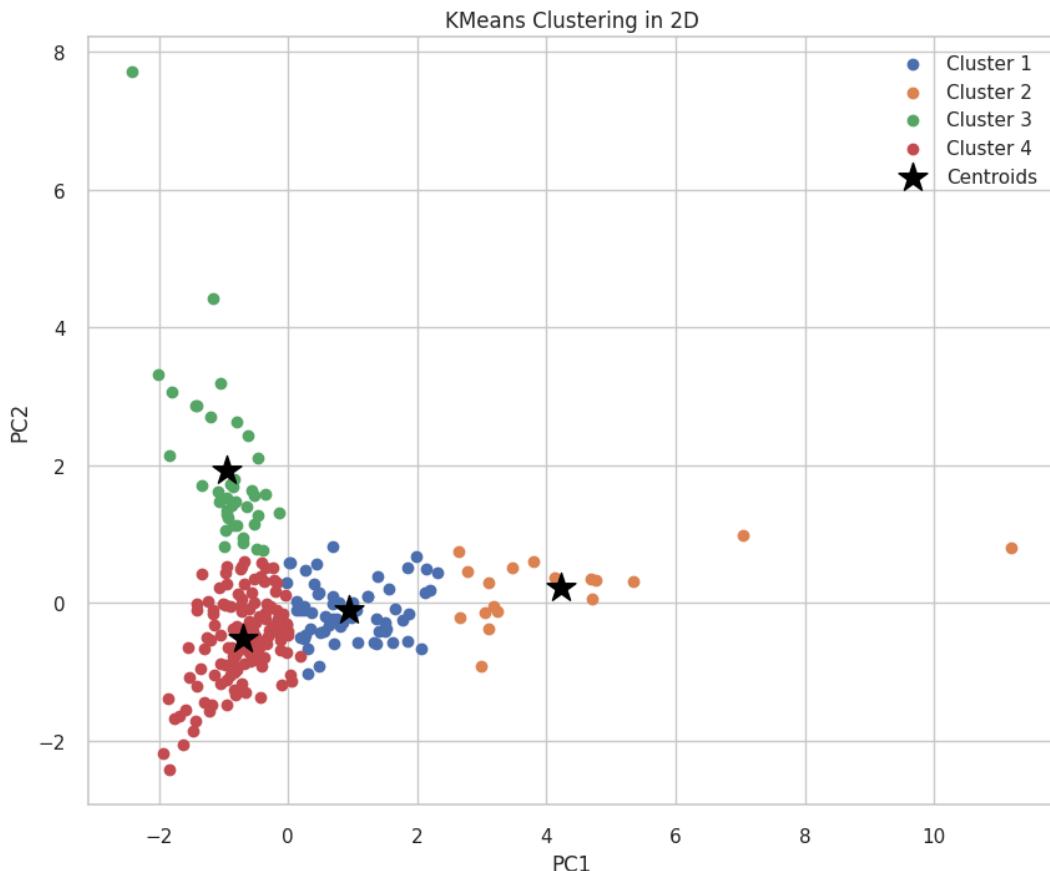
4.4.2 Silhouette Score

Number of Clusters	Silhouette Score
0	2
1	3
2	4

Gambar 4.13 Silhouette Score

Berdasarkan tabel, nilai rata-rata silhouette score untuk pembagian data menjadi 2 cluster adalah 0.541, yang menunjukkan tingkat pemisahan yang cukup baik dengan cluster yang terpisah secara signifikan. Ketika data dibagi menjadi 3 cluster, nilai rata-rata silhouette score adalah sekitar 0.511. Ini menunjukkan bahwa tingkat pemisahan menjadi lebih rendah dibandingkan dengan pembagian 2 cluster, dengan beberapa overlap antar cluster. Untuk pembagian data menjadi 4 cluster, nilai rata-rata silhouette score kembali menurun menjadi sekitar 0.44, yang menunjukkan bahwa tingkat pemisahan menjadi lebih rendah dibandingkan dengan pembagian 3 cluster, dengan beberapa overlap antar cluster.

4.4.3 K-Means Clustering



Gambar 4.14 Hasil K-Means Clustering 2D

Berdasarkan visualisasi K-Means Clustering 2D di atas antara PC1 dan PC2, didapatkan 4 klaster serta beberapa titik pusat (centroids). Klaster 1 (berwarna biru) memiliki nilai PC1 yang bervariasi dan terpusat di sekitar nilai negatif hingga positif kecil pada PC2. Hal ini mengindikasikan bahwa hotel-hotel dalam klaster ini cenderung memiliki harga yang menengah, rating yang cukup baik, jumlah reviewer yang moderat, dan fasilitas yang memadai.

Klaster 2 (berwarna oranye) terpusat pada nilai PC1 yang lebih tinggi dan nilai PC2 yang negatif hingga sedikit positif, menunjukkan bahwa hotel-hotel dalam klaster ini memiliki harga yang lebih tinggi, rating yang sangat baik, dan fasilitas yang lebih lengkap dibandingkan dengan klaster lainnya.

Klaster 3 (berwarna hijau) memiliki nilai PC1 yang lebih tinggi dengan nilai PC2 yang bervariasi dari negatif hingga positif tinggi, mengindikasikan bahwa hotel-hotel dalam klaster ini memiliki harga yang lebih terjangkau, rating yang lebih rendah, jumlah *reviewer* yang lebih sedikit, dan fasilitas yang lebih dasar.

Klaster 4 (berwarna merah) memiliki nilai PC1 negatif hingga sedikit positif dengan rentang nilai PC2 yang rendah hingga tinggi, mengindikasikan bahwa hotel-hotel dalam klaster ini memiliki kombinasi harga, rating, jumlah reviewer, dan fasilitas yang bervariasi namun tetap memiliki kesamaan tertentu yang membedakan mereka dari klaster lainnya.

Selain itu, titik-titik pusat (*centroid*) yang ditandai dengan bintang hitam menunjukkan pusat dari masing-masing klaster yang membantu dalam memahami karakteristik utama dari tiap klaster. Dengan demikian, analisis clustering ini memberikan wawasan yang jelas mengenai segmentasi hotel-hotel di Pulau Bali berdasarkan harga, rating, jumlah reviewer, dan fasilitas.

Klaster	Harga	Review	Rating	Total Fasilitas	Anggota
1	Rp1.976.745,26	369	8,69	4	61
2	Rp5.708.495,44	271	8,78	5	18
3	Rp614.669,00	3110	8,57	5	39
4	Rp513.285,24	559	8,32	4	137

Tabel 4.2 Hasil *Clustering K-Means*

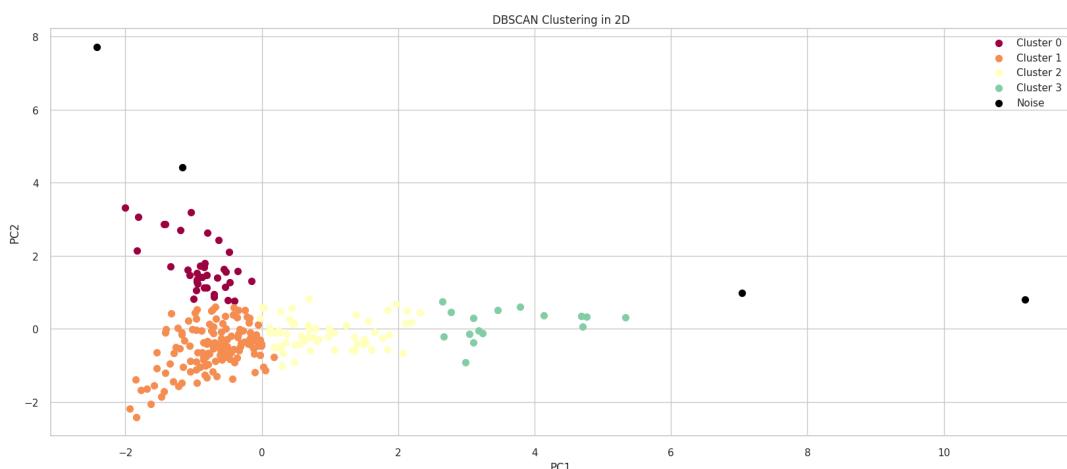
Berikut ini merupakan nilai rata-rata dari masing-masing variabel berdasarkan hasil *clustering* menggunakan metode K-Means. Dari statistika deskriptif nilai rata-rata tersebut, klaster 1 memiliki rata-rata harga dan rating cukup tinggi walaupun jumlah ulasan dan jumlah fasilitas bukan yang tertinggi.

Pada klaster 2 banyaknya ulasan memang paling sedikit di antara klaster lainnya (271 *reviews*), namun memiliki harga, *rating*, dan jumlah fasilitas tertinggi dibandingkan klaster lainnya.

Jumlah ulasan pada statistik deskriptif tersebut menandakan bahwa hotel tersebut populer di kalangan para turis. Sehingga dapat disimpulkan bahwa pada klaster 3 terdapat hotel-hotel yang populer dengan rata-rata harga sebesar Rp614,699, *rating* sebesar 8,75, dan total fasilitas sebanyak 5.

Klaster terakhir atau klaster 4 memiliki anggota yang paling banyak dimana rata-rata harga, *rating*, dan jumlah fasilitas hotel tersebut paling rendah dibandingkan klaster lain dengan jumlah ulasan yang cukup banyak.

4.4.4 DBSCAN Clustering



Gambar 4.15 Hasil DBSCAN Clustering 2D

Berdasarkan visualisasi DBSCAN Clustering 2D di atas antara PC1 dan PC2, didapatkan 4 klaster serta beberapa titik noise. Klaster 0 memiliki nilai PC1 dan PC2 yang bervariasi dengan konsentrasi utama pada nilai PC1 negatif dan nilai PC2 berkisar antara negatif hingga positif sedang. Hal ini mengindikasikan bahwa hotel-hotel dalam klaster ini cenderung memiliki harga yang terjangkau, rating yang beragam, jumlah reviewer yang sedang, dan fasilitas yang memadai.

Klaster 1 terpusat pada nilai PC1 rendah hingga positif dan nilai PC2 negatif hingga sedikit positif, menunjukkan bahwa hotel-hotel dalam klaster ini memiliki harga yang lebih tinggi, rating yang baik, dan fasilitas yang lebih lengkap dibandingkan dengan klaster lain.

Klaster 2 memiliki nilai PC1 positif dan nilai PC2 yang sangat rendah hingga sedikit positif, mengindikasikan bahwa hotel-hotel dalam klaster ini memiliki harga

yang lebih terjangkau, rating yang lebih rendah, jumlah reviewer yang lebih sedikit, dan fasilitas yang lebih dasar.

Klaster 3 memiliki nilai PC1 yang positif dengan rentang nilai PC2 yang rendah, mengindikasikan bahwa hotel-hotel dalam klaster ini cenderung memiliki kombinasi harga, rating, jumlah reviewer, dan fasilitas yang beragam namun tetap memiliki kesamaan tertentu yang membedakan mereka dari klaster lainnya.

Selain itu, terdapat beberapa titik noise yang tidak termasuk dalam klaster manapun, yang kemungkinan merupakan hotel-hotel dengan karakteristik yang sangat berbeda atau data yang tidak konsisten. Dengan demikian, analisis clustering ini memberikan wawasan yang jelas mengenai segmentasi hotel-hotel di Pulau Bali berdasarkan harga, rating, jumlah reviewer, dan fasilitas.

Klaster	Harga	Review	Rating	Total Fasilitas	Anggota
1	Rp612.447,62	2857	8,56	5	37
2	Rp513.285,24	559	8,32	4	137
3	Rp1.976.745,26	369	8,69	4	61
4	Rp4.983.063,62	244	8,78	5	16

Tabel 4.3 Hasil *Clustering* DBSCAN

Berikut merupakan nilai rata-rata dari masing-masing variabel yang ada berdasarkan hasil *clustering* menggunakan metode DBSCAN. Dapat dilihat bahwa hasil *clustering* tidak jauh berbeda dengan metode K-Means. Perbedaan yang paling terlihat adalah pada jumlah anggota masing-masing klaster. Hal ini dikarenakan pada metode DBSCAN terdapat nilai noise yang tidak masuk ke dalam cluster manapun. Pada hasil *clustering* ini terdapat 4 titik data noise. Data noise tersebut sedikit mempengaruhi pada statistik deskriptif walaupun persebarannya tetap sama dengan metode *K-Means*.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Pariwisata di Bali yang berkembang pesat meningkatkan permintaan akan akomodasi berkualitas. Studi ini menganalisis dan mengelompokkan hotel di Bali berdasarkan harga, rating, jumlah ulasan, dan fasilitas, memberikan wawasan berharga bagi wisatawan dan pemangku kepentingan pariwisata.

Berdasarkan hasil analisis dengan dua metode clustering, yaitu K-Means dan DBSCAN, dapat disimpulkan bahwa data hotel di Pulau Bali berdasarkan pada variabel harga, rating, jumlah viewer, dan fasilitas, diperoleh segmentasi hotel dengan karakteristik yang berbeda beda.

A. K-Means

1. Klaster 1: Hotel harga menengah dengan rating cukup baik, jumlah viewer moderat, dan fasilitas memadai.
2. Klaster 2: Hotel harga lebih tinggi dengan rating sangat baik, dan fasilitas lengkap.
3. Klaster 3: Hotel harga terjangkau dengan rating lebih rendah, jumlah viewer sedikit, dan fasilitas dasar.
4. Klaster 4: Hotel harga, rating, jumlah viewer, dan fasilitas yang beragam.

B. DBSCAN

1. Klaster 0: Hotel dengan harga terjangkau, rating beragam, jumlah reviewer sedang, dan fasilitas memadai.
2. Klaster 1: Hotel dengan harga lebih tinggi, rating baik, dan fasilitas lebih lengkap dibandingkan klaster lainnya.
3. Klaster 2: Hotel dengan harga terjangkau, rating lebih rendah, jumlah reviewer lebih sedikit, dan fasilitas dasar.
4. Klaster 3: Hotel dengan kombinasi harga, rating, jumlah reviewer, dan fasilitas yang beragam namun memiliki kesamaan tertentu.

DBSCAN dapat mengidentifikasi beberapa titik noise yang tidak termasuk dalam klaster manapun. Titik-titik ini merupakan hotel dengan karakteristik yang sangat berbeda atau data yang tidak konsisten. Hal ini dapat membantu dalam memahami pola distribusi hotel di Pulau Bali berdasarkan harga, rating, jumlah reviewer, dan fasilitas, memberikan wawasan yang lebih baik untuk segmentasi pasar dan strategi pemasaran.

5.2 Saran

1. Untuk Wisatawan: Gunakan informasi klaster ini untuk memilih hotel yang sesuai dengan preferensi dan anggaran. Misalnya, untuk pengalaman mewah, fokus pada hotel di Klaster 2 dari analisis K-Means atau Klaster 1 dari analisis DBSCAN. Sedangkan bagi wisatawan dengan anggaran terbatas, hotel di Klaster 3 dari kedua metode bisa menjadi pilihan yang baik.

2. Untuk Pemangku Kepentingan Pariwisata: Pemangku kepentingan dapat menggunakan hasil analisis ini untuk merencanakan strategi pemasaran yang lebih efektif, menargetkan segmen pasar spesifik sesuai dengan karakteristik tiap klaster. Informasi ini dapat membantu dalam pengembangan infrastruktur dan promosi wisata yang lebih terfokus, sesuai dengan kebutuhan dan preferensi wisatawan yang berbeda. Pengambil keputusan di sektor perhotelan dapat memanfaatkan pemetaan klaster ini untuk meningkatkan kualitas layanan dan fasilitas di hotel-hotel yang berada di klaster dengan rating dan fasilitas yang lebih rendah.

Dengan mengikuti saran ini, diharapkan wisatawan dapat membuat keputusan lebih baik, dan pemangku kepentingan dapat meningkatkan kualitas serta daya saing hotel di Bali, memberikan manfaat lebih besar bagi industri pariwisata secara keseluruhan.

DAFTAR PUSTAKA

- Birant, D. & Kut, A. 2007, 'ST-DBSCAN: An algorithm for clustering spatial-temporal data', *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 208-221.
- Dewi, C., Siam, E. P., Wijayanti, G. A., Putri, M., Aulia, N. & Nooraeni, R. 2021, 'Comparison of DBSCAN and K-Means clustering for grouping the village status in Central Java 2020', *Jurnal Matematika, Statistika dan Komputasi*, vol. 17, no. 3, pp. 394-404.
- Hadanny, A., Harland, T. A., Khazen, O., DiMarzio, M., Telkes, I. & Pilitsis, J. G. 2022, 'In Reply: Development of Machine Learning-Based Models to Predict Treatment Response to Spinal Cord Stimulation', *Neurosurgery*, vol. 91, no. 2, pp. e68-e70.
- Margaretha, F., Wirawan, S. E. & Wowor, W. 2022, 'The influence of service quality toward customer loyalty at five-star hotel in Bali', *International Journal of Social and Management Studies*, vol. 3, no. 2, pp. 175-186.
- Pandey, P. 2020, Feature Scaling: MinMax, Standard and Robust Scaler, *Machine Learning Geek*, viewed 2 June 2024, <https://machinelearninggeek.com/feature-scaling-minmax-standard-and-robust-scaler/>
- Rohman, S. & Pratama, R. N. 2022, 'Customer Segmentation Using the Integration of the Recency Frequency Monetary Model and the K-Means Cluster Algorithm', *Scientific Journal of Informatics*, vol. 9, no. 2, pp. 189.
- Rosado, L., Correia da Costa, J. M., Elias, D. & S Cardoso, J. 2016, 'A review of automatic malaria parasites detection and segmentation in microscopic images', *Anti-Infective Agents*, vol. 14, no. 1, pp. 11-22.
- Sendi, M., Salat, D., Miller, R. & Calhoun, V. 2022, 'Two-step clustering-based pipeline for big dynamic functional network connectivity data', *Frontiers in Neuroscience*, vol. 16, pp. 895637.
- Shahapure, K. R. & Nicholas, C. 2020, 'Cluster quality analysis using silhouette score', in 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pp. 747-748, IEEE.
- Yadi, Y. 2018, 'Analisa usability pada website Traveloka', *Jurnal Ilmiah BETRIK: Besemah Teknologi Informasi dan Komputer*, vol. 9, no. 03, pp. 172-180.
- Zhao, H. 2022, 'Design and implementation of an improved k-means clustering algorithm', *Mobile Information Systems*, vol. 2022, pp. 1-10.

LAMPIRAN

Code Analisis: [!\[\]\(7fd808d098fc71ab2be986223535f4b7_img.jpg\) DM1_SD-A1_010_056_061_062_097.ipynb](#)

Code Scraping: [Scraping](#)

Dataset: [Dataset](#)