

STAGE 2 FINAL PROJECT

Data Science Batch 51

FourSight Team

RAKAMIN



ANGGOTA KELOMPOK



Mufti Habibie Alayubi

Data Scientist



Syahdilla Fitri U

Project Manager



Johannes

Data Analyst



Ismawardani

Data engineer



MATERI PRESENTASI

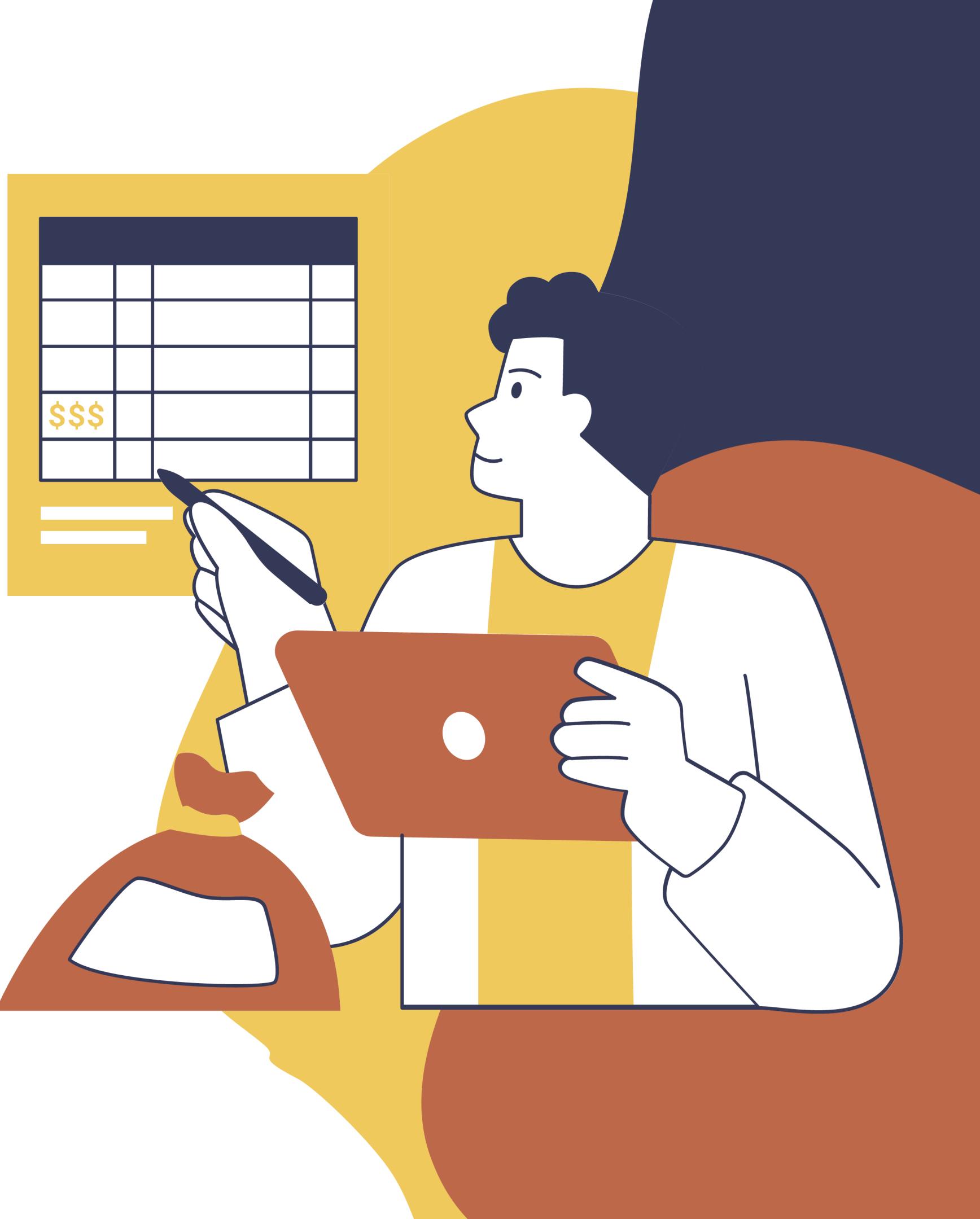


- 1 Model Selection
- 2 Model training
- 3 Profilisasi Cluster
- 4 Streamlit User Interface



1. MODEL SELECTION

Machine learning (ML) memainkan peran penting dalam industri perhotelan dan penyewaan properti seperti Airbnb. Tiga kategori utama dalam ML adalah Supervised Learning, Unsupervised Learning, dan Reinforcement Learning. Dari ketiganya, **Clustering**, yang termasuk dalam Unsupervised Learning, menjadi fokus utama dalam analisis Airbnb.



MENGAPA CLUSTERING?

Clustering memungkinkan pengelompokan data berdasarkan karakteristik tertentu, yang bermanfaat bagi Airbnb dalam:

- Mengidentifikasi segmen pelanggan berdasarkan preferensi dan perilaku.
- Menganalisis pola harga dan lokasi untuk strategi penetapan harga yang lebih optimal.
- Mengelompokkan properti berdasarkan fitur untuk pengalaman pengguna yang lebih baik.

Beberapa metode clustering yang sering digunakan dalam bisnis Airbnb meliputi **K-Means Clustering**, Hierarchical Clustering, dan DBSCAN.



JURNAL YANG MENDUKUNG ANALISIS CLUSTERING DALAM BISNIS AIRBNB

Airbnb Clustered Price Ranges

Daniel Forgosh, Brandon Ubiera, and Bemnut Nuru
Computer Science
Hood College
djf7@hood.edu, bu2@hood.edu, bn3@hood.edu

December 21, 2018

Dalam jurnal "Airbnb Clustered Price Ranges" oleh Daniel Forgosh, Brandon Ubiera, dan Bemnut Nuru, para peneliti memilih model clustering untuk menganalisis data geografis Airbnb di New York City. Mereka menyatakan bahwa "Clustering is a powerful method to data mine geographical data. It can group coordinate data points that consist of latitude and longitude."

Untuk membangun model mereka, peneliti menggunakan algoritma clustering untuk mengelompokkan data koordinat (lintang dan bujur) dari semua listing Airbnb di New York City. Tujuannya adalah menemukan distribusi frekuensi harga dalam setiap cluster yang terbentuk, sehingga dapat mengidentifikasi pola harga berdasarkan lokasi geografis.



Dynamic Pricing Analytic of Airbnb Amsterdam Using K-Means Clustering

Publisher: IEEE

Cite This

PDF

ABSTRAK

Abstrak: Penelitian ini menggunakan model K-Means Clustering untuk menganalisis harga dinamis Airbnb di Amsterdam, dengan akurasi penelitian mencapai 96,21%.

METODE

- Pengumpulan Data → Harga, lokasi, fasilitas Airbnb.
- Klastering (K-Means) → Menentukan jumlah klaster optimal.
- Analisis → Identifikasi pola harga & klaster terbaik.
- Penetapan Harga Dinamis → Menyesuaikan harga secara real-time.
- Hasil: Akurasi 96,21%, pendapatan meningkat.

2. MODEL

.....

TRAINING

Pelatihan model adalah proses melatih algoritma ML dengan data pelatihan yang memadai untuk menunjukkan korelasi antara hasil dan variabel yang memengaruhi. Dalam hal ini diperlukan data cleaning dan feature engineering. Model yang dipilih menggunakan K-Means karena dianggap mampu digunakan untuk data dengan skala besar.



2. MODEL TRAINING

HANDLING
MISSING VALUE

.....

HANDLING
INCONSISTENCY

```
listing_id      0
date            0
available       0
price           0
dtype: int64
```

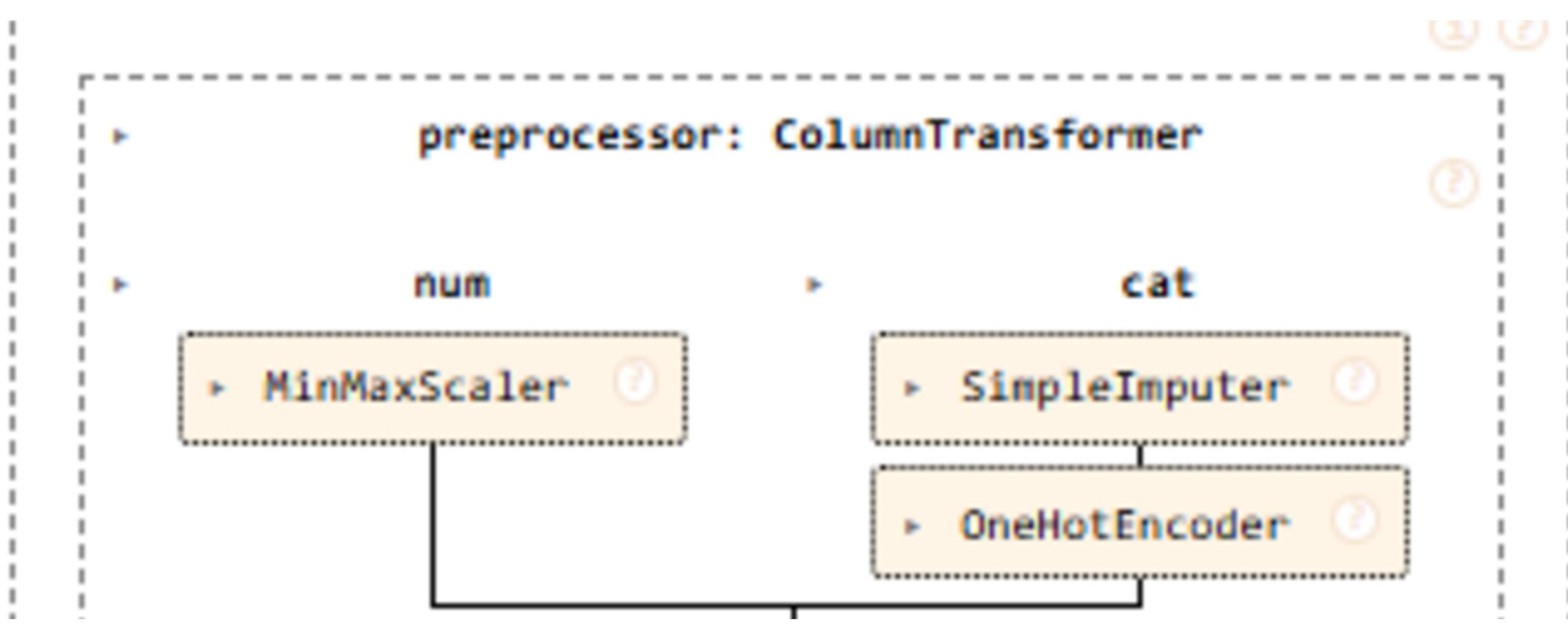
```
0    85.0
1    85.0
2    0.0
3    0.0
4    0.0
Name: price, dtype: float64
```

CLUSTERING LISTING DATA

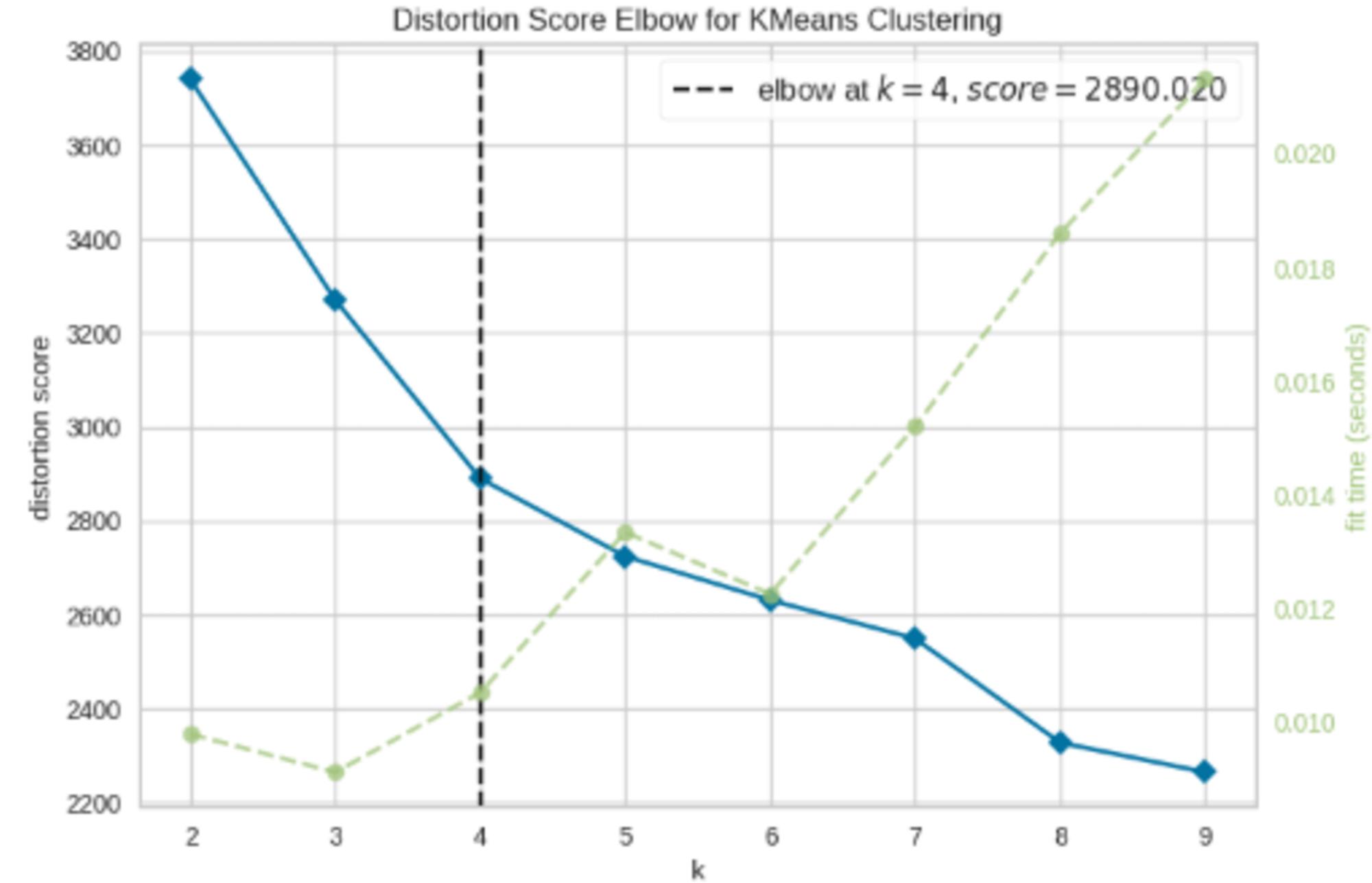
	property_type	room_type	price	review_scores_rating	accommodates	bedrooms	beds	bed_type	host_is_superhost	bathrooms	availability_365	number_of_reviews	host_response_time	city_normalized	distance_to_city
0	Apartment	Entire Home/Apt	85.0	95.0	4.0	1.0	1.0	Real Bed	F	1.0	346	62.0	Within A Few Hours	Seattle	4.439052
1	Apartment	Entire Home/Apt	150.0	96.0	4.0	1.0	1.0	Real Bed	T	1.0	291	43.0	Within An Hour	Seattle	4.441933
2	House	Entire Home/Apt	975.0	97.0	7.0	3.5	3.5	Real Bed	F	1.0	220	20.0	Within A Few Hours	Seattle	3.833194
3	Apartment	Entire Home/Apt	100.0	96.0	3.0	0.0	2.0	Real Bed	F	1.0	143	0.0	Within An Hour	Seattle	4.543407
4	House	Entire Home/Apt	450.0	92.0	6.0	3.0	3.0	Real Bed	F	1.0	365	38.0	Within An Hour	Seattle	4.240533
...
3813	House	Entire Home/Apt	359.0	88.0	6.0	3.0	3.0	Real Bed	F	1.0	32	1.0	Within A Few Hours	Seattle	6.770782
3814	Apartment	Entire Home/Apt	79.0	100.0	4.0	1.0	2.0	Real Bed	F	1.0	273	2.0	Within An Hour	Seattle	4.930030
3815	House	Entire Home/Apt	93.0	96.0	2.0	1.0	1.0	Real Bed	F	1.0	88	0.0	Within An Hour	Seattle	12.855801
3816	Condominium	Entire Home/Apt	99.0	96.0	2.0	0.0	1.0	Real Bed	F	1.0	179	0.0	Within An Hour	Seattle	5.140297
3817	Apartment	Entire Home/Apt	87.0	96.0	3.0	2.0	1.0	Real Bed	F	1.0	7	0.0	Within A Day	Seattle	3.961612

3818 rows × 15 columns

HANDLING MISSING VALUE, HANDLING INCONSISTENCY DAN HANDLING OUTLIER, MAKA KEMUDIAN DILAKUKAN CLUSTERING LISTING DATA DAN MENGHASILKAN DATA DI ATAS.



1. `MinMaxScaler` adalah fungsi untuk melakukan normalisasi data. Artinya skala data akan diubah sehingga hanya bernilai antara 0 dan 1. Nilai 0 untuk data terendah, dan nilai 1 untuk data tertinggi.
2. `Imputer` dari Scikit-Learn yang digunakan untuk mengganti data yang hilang dengan nilai tertentu. Ini dapat mempercepat proses data preprocessing.
3. `One hot encoder` digunakan untuk melakukan label encoding khusus pada label-label yang memiliki lebih dari dua kategori. Hal ini memungkinkan setiap kategori untuk dibuatkan satu kolom khusus, sehingga menjadi binary. Teknik ini digunakan untuk menghindari bias yang bersifat ordinal pada `LabelEncoding`.



Dari inisialisasi Kmeans diperoleh jumlah kluster yang optimal dari visualizer, dan menampilkan label yang diperoleh.

Berdasarkan uji Elbow, Cluster dengan perubahan signifikan terdapat di cluster 4

3. Profilisasi Cluster

cluster	price	review scores rating	accommodates	bedrooms	beds	host is superhost	availability 365	number of reviews	distance to city	host response time
1	137,5231092	97,46008403	3,386554622	1,1470588	1,63445	1	233,3046218	32,60504	4,6635	1,384453782
2	72,64636076	95,45490506	2,013449367	1	1,12579	0,19778481	267,7950949	18,79035	5,80292	1,564082278
3	122,375642	95,05796038	2,967718269	0,9545121	1,40352	0	230,0190756	13,88408	4,10887	1,638297872
4	230,1104895	95,15244755	5,882517483	2,5048951	3,02028	0,072727273	239,8321678	11,05035	5,3595	1,713286713

Merah : Tertinggi

Orange : Tertinggi Ke-2

Kuning : Tertinggi Ke-3

Hijau : Terendah

Cluster 1:

- Harga (price): Rp137 ribuan (sedang, kuning)
 - Review score rating: 97,46 (tertinggi, merah) → Sangat baik
 - Superhost: (tertinggi, merah) → Semua host di cluster adalah Superhost
 - Jarak ke pusat kota: (terdekat kedua)
 - Respon time host: 1,38 jam (tercepat, hijau)
- Kategori: Properti berkualitas tinggi, harga sedang, superhost, dan dekat kota.

Cluster 2 (Harga termurah, sedikit fasilitas, jarak terjauh)

- Harga: Rp72 ribuan (termurah, hijau)
 - Review score rating: 95,45 (terendah, hijau)
 - Superhost: 19,8% (tertinggi kedua, oranye)
 - Kapasitas: 2 orang, 1 kamar tidur → Paling minimalis
 - Jarak ke pusat kota: (terjauh, merah)
- Kategori: Budget-friendly, fasilitas minimal, jarak agak jauh dari pusat kota.

Cluster 3 (Fasilitas paling sedikit, superhost sedikit)

- Harga: Rp122 ribuan (sedang, kuning)
 - Superhost: 0% (terendah, hijau) → Tidak ada Superhost
 - Kapasitas: 3 orang, 1 kamar tidur → Mirip Cluster 2
 - Jarak ke pusat kota: (terdekat, hijau)
 - Respon time host: 1,63 jam (ke-2 tertinggi, oranye)
- Kategori: Properti dekat kota, tapi hostnya bukan Superhost.

Cluster 4 (Harga tertinggi, kapasitas besar, review bagus)

- Harga: Rp230 ribuan (tertinggi, merah)
 - Kapasitas: 5-6 orang, 2,5 kamar tidur (tertinggi, merah)
 - Jarak ke pusat kota: (ke-3 tertinggi, oranye)
 - Respon time host: 1,71 jam (tertinggi, merah)
- Kategori: Properti mewah, harga mahal, cocok untuk keluarga/banyak orang.

Profilisasi Cluster

Cluster	roomtype Entire Home/Apt	room type Private Room	room type Shared Room	bed type Airbed	bed type Couch	bed type Futon	bed type Pull-Out Sofa	bed type Real Bed
1	0,9811	0,0000	0,0189	0,0042	0,0021	0,0168	0,0063	0,9706
2	0,0000	0,9177	0,0823	0,0158	0,0079	0,0396	0,0245	0,9122
3	1,0000	0,0000	0,0000	0,0037	0,0015	0,0117	0,0081	0,9751
4	0,9944	0,0000	0,0056	0,0000	0,0000	0,0000	0,0028	0,9972

Merah : Tertinggi

Orange : Tertinggi Ke-2

Kuning : Tertinggi Ke-3

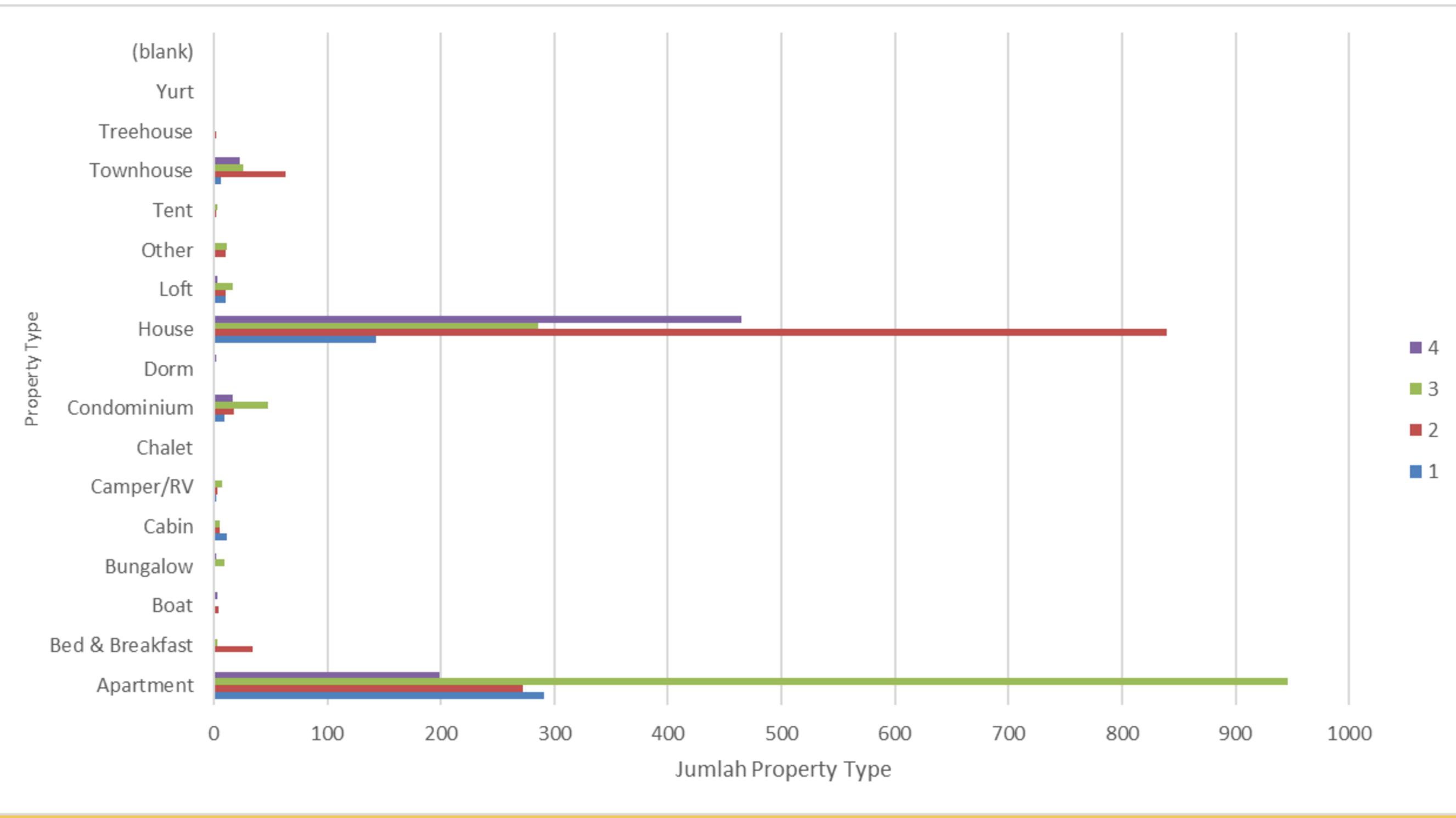
Hijau : Terendah

Cluster 1: Mayoritas terdiri dari Entire Home/Apt dengan harga menengah (137,52) dan jumlah ulasan tertinggi. Mayoritas menggunakan Real Bed, menunjukkan kenyamanan standar dengan akses penuh ke akomodasi.

Cluster 2: Didominasi oleh Private Room dengan harga paling rendah (72,65). Berbagai jenis tempat tidur digunakan, termasuk Futon yang cukup tinggi, menandakan opsi ekonomis dengan fasilitas terbatas.

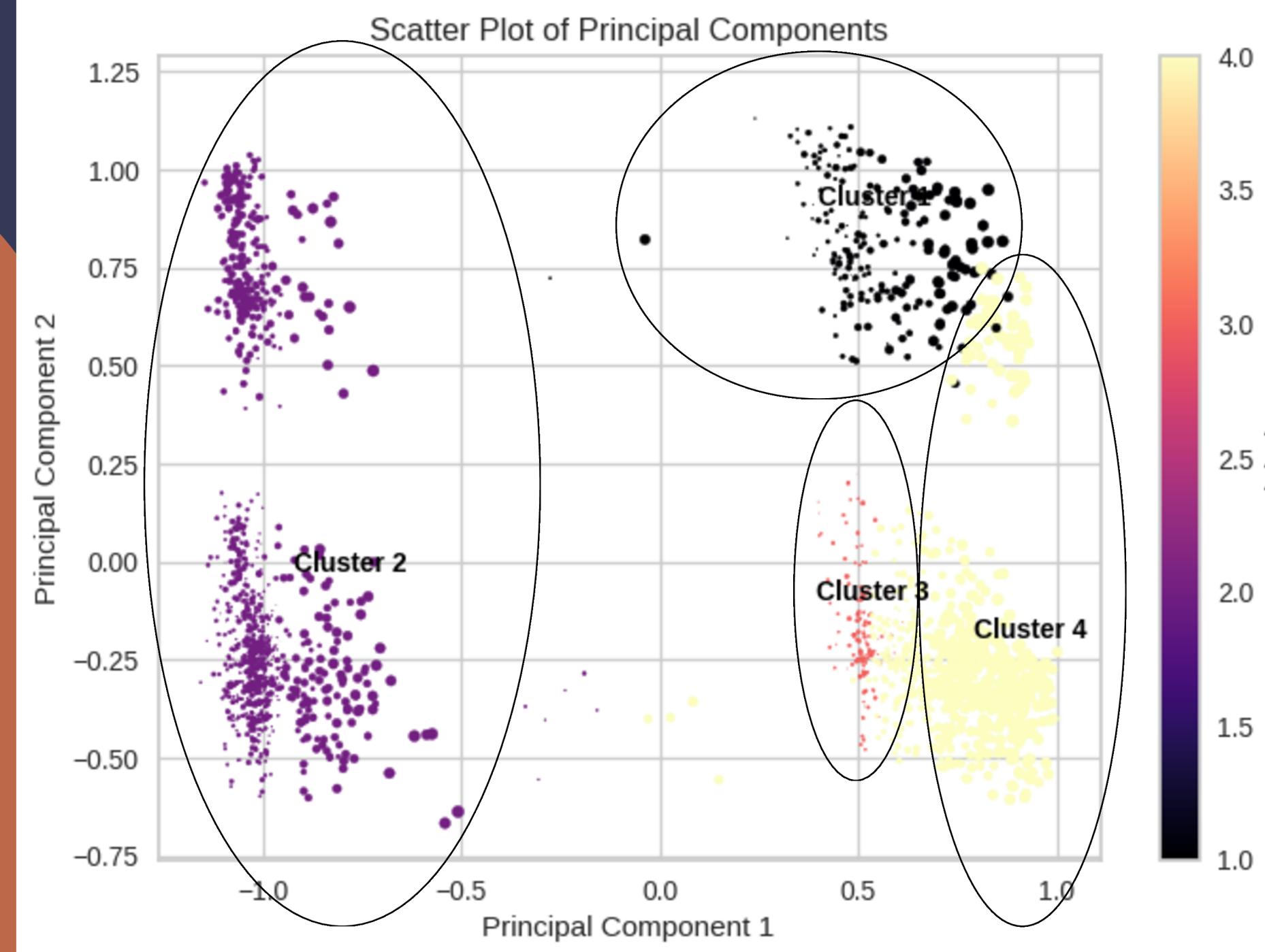
Cluster 3: Seluruh room type nya adalah Entire/Apartemen dengan harga menengah (122,37), tanpa Shared Room, dan didominasi Real Bed, mencerminkan kenyamanan lebih baik dibandingkan Cluster 2.

Cluster 4: Hampir semua Entire Home/Apt dengan harga tertinggi (230,11), sangat didominasi Real Bed, mencerminkan akomodasi premium dengan kenyamanan terbaik.



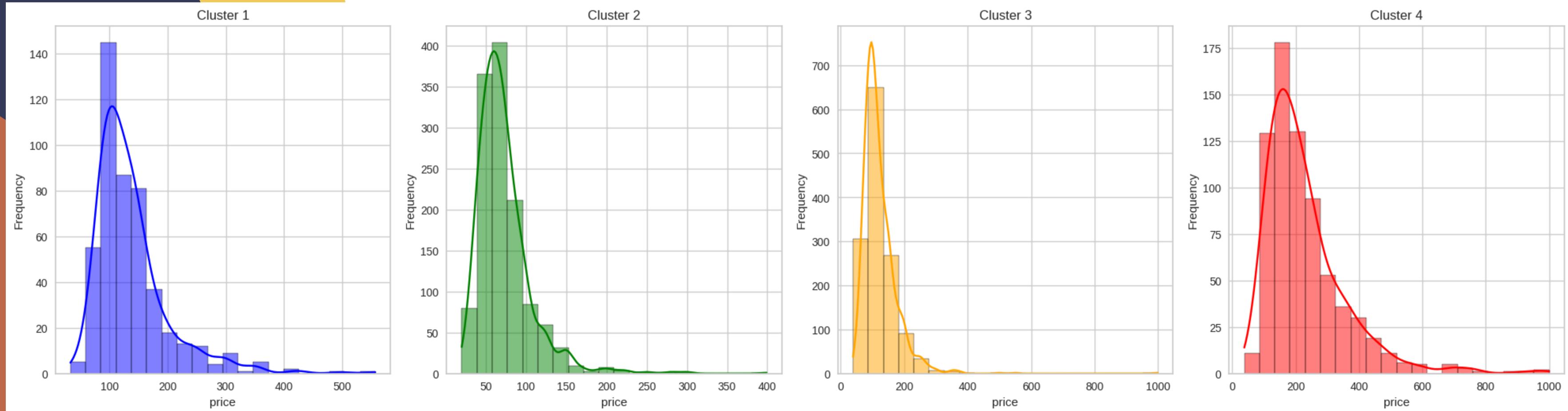
Cluster 2 (merah) didominasi oleh House, menunjukkan bahwa properti ini paling banyak digunakan dalam kategori harga rendah. Cluster 3 (hijau) lebih banyak memiliki Apartment, menandakan properti ini umumnya berada di kelas harga menengah. Cluster 4 (ungu) yang memiliki harga paling tinggi, juga didominasi oleh House, tetapi dalam jumlah lebih sedikit dibanding Cluster 2, menunjukkan properti dengan fasilitas lebih eksklusif. Cluster 1 (biru) memiliki distribusi yang lebih merata, tetapi lebih banyak pada Apartment dan beberapa tipe properti unik seperti Treehouse atau Loft.

Scatter Plot PCA



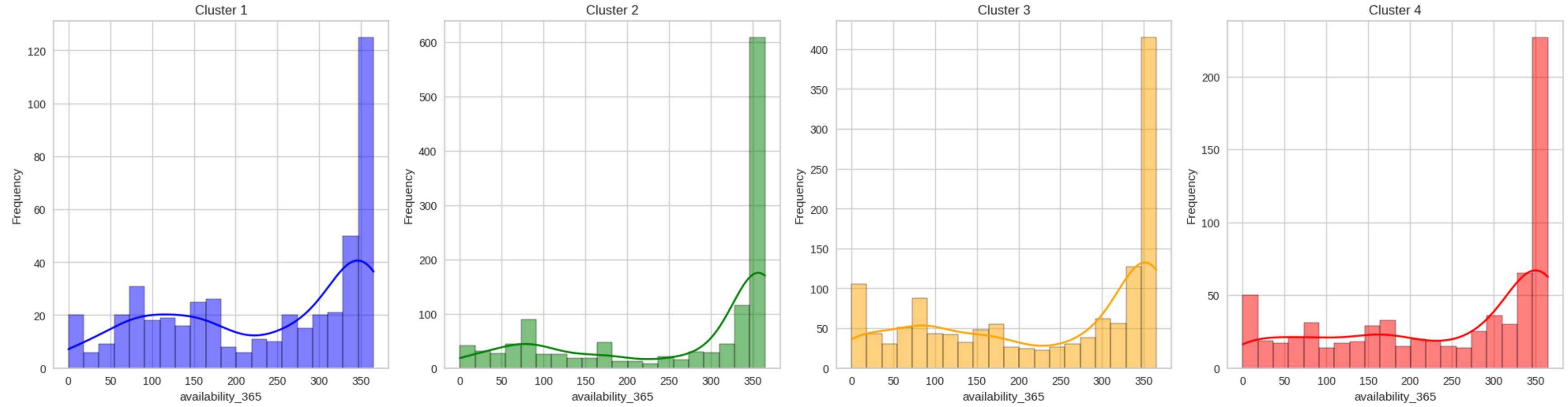
Terlihat bahwa antar cluster cukup terpisah dengan baik

Distribusi Harga per Cluster



- Cluster 2 (Hijau) memiliki harga paling rendah dan cenderung lebih homogen.
- Cluster 3 (Oranye) memiliki distribusi yang sangat sempit dan terkonsentrasi di harga rendah.
- Cluster 1 (Biru) dan Cluster 4 (Merah) memiliki harga yang lebih tinggi, dengan Cluster 4 memiliki sebaran yang lebih luas.
- Cluster 4 menunjukkan variasi harga terbesar, dengan beberapa properti yang jauh lebih mahal dibandingkan cluster lain.

Distribusi Ketersediaan 365 Hari per Cluster



- Cluster 2 dan 3 memiliki pola yang lebih jelas, dengan mayoritas properti tersedia sepanjang tahun (365 hari).
- Cluster 1 dan 4 lebih tersebar, menunjukkan kombinasi antara properti dengan ketersediaan tinggi dan rendah.
- Jika ingin mencari properti yang tersedia sepanjang tahun, Cluster 2 dan 3 lebih dominan dibandingkan lainnya.
- Untuk variasi ketersediaan yang lebih fleksibel, Cluster 1 dan 4 lebih sesuai.

KESIMPULAN CLUSTERING

Berdasarkan hasil clustering, Cluster 2 merupakan pilihan paling murah dengan fasilitas minimalis dan lokasi agak jauh dari pusat kota, cocok bagi wisatawan dengan anggaran terbatas. Cluster 3 juga memiliki harga relatif rendah dengan fasilitas terbatas, namun lebih dekat ke pusat kota meskipun tanpa Superhost. Cluster 1 menawarkan properti berkualitas tinggi dengan harga sedang, didominasi oleh Superhost, serta memiliki akses cepat ke pusat kota dan responsivitas host yang tinggi. Sementara itu, Cluster 4 adalah kategori paling mahal, menawarkan kapasitas besar dan kenyamanan lebih, menjadikannya pilihan ideal untuk keluarga atau kelompok besar yang mencari akomodasi eksklusif.

- Berdasarkan hasil clustering, Cluster 2 merupakan pilihan paling murah dengan fasilitas minimalis dan lokasi agak jauh dari pusat kota, cocok bagi wisatawan dengan anggaran terbatas. Cluster 3 juga memiliki harga relatif rendah dengan fasilitas terbatas, namun lebih dekat ke pusat kota meskipun tanpa Superhost. Cluster 1 menawarkan properti berkualitas tinggi dengan harga sedang, didominasi oleh Superhost, serta memiliki akses cepat ke pusat kota dan responsivitas host yang tinggi. Sementara itu, Cluster 4 adalah kategori paling mahal, menawarkan kapasitas besar dan kenyamanan lebih, menjadikannya pilihan ideal untuk keluarga atau kelompok besar yang mencari akomodasi eksklusif.
-
-
-
-
-

4. Streamlit User Interface

<https://recommendation-airbnb-foursight.streamlit.app>

Airbnb Recommendation System

Find your perfect Airbnb property based on your preferences!



Created by : Foursight

Show raw data

Select Cluster

Traveler's Choice

Price Range: 50.00 - 200.00

Review Scores: 80 - 100

Min Bedrooms: 1

Min Bathrooms: 1

Search

Lovely Queen Anne Cottage, 2 BR
\$109.00

Elegance in Historic Queen Anne
\$100.00

1914 Stunner View
\$100.00

Select Cluster

Comfort Living

Price Range: 20.00 - 1000.00

Review Scores: 20 - 100

Min Bedrooms: 1

Min Bathrooms: 1

Search

Property Name	Price
Tiny Garden cabin on Queen Anne	\$60.00
Queen Anne Private Bed and Bath	\$80.00
Cozy Queen Anne Finished Basement	\$99.00
Park Life in Lower Queen Anne	\$66.00
Private room in upper QA w/ view!	\$70.00
Hand crafted Gypsy Wagon with heart	\$69.00
Private unit in a 1920s mansion	\$120.00
Urban Charm Downtown Views	\$90.00
Ballard Artist's Studio	\$99.00
One bedroom with Lounge	\$150.00
Cozy room in Modern	\$69.00

TERIMA KASIH

