

LAPORAN AKHIR

“Clustering Kos Dari Harga dan Fasilitas Untuk Menemukan Opsi Paling Ekonomis Sesuai Kebutuhan”



KELOMPOK: SD-A1 - KELOMPOK E

- | | |
|---------------------------------------|-------------|
| 1. Ilham Dicky Darmawan | (164221023) |
| 2. Erwina Yolavania | (164221037) |
| 3. Raissa Dinda Maya Sabilla | (164221069) |
| 4. Patricia Dewinta Wahyu Krisnayanti | (164221079) |
| 5. Fadli Muhammad | (164221081) |

TEAM-BASES PROJECT

MATA KULIAH DATA MINING I

PROGRAM STUDI TEKNOLOGI SAINS DATA

FAKULTAS TEKNOLOGI MAJU DAN MULTIDISIPLIN

UNIVERSITAS AIRLANGGA

2024

DAFTAR ISI

BAB I.....	1
PENDAHULUAN.....	1
1.1. Latar Belakang.....	1
1.2. Rumusan Masalah.....	2
1.3. Manfaat Penelitian.....	2
1.4. Tujuan Penelitian.....	2
BAB II.....	3
TINJAUAN PUSTAKA.....	3
2.1. Data Mining.....	3
2.2. Clustering.....	3
2.3. K-Means.....	3
BAB III.....	5
METODOLOGI PENELITIAN.....	5
3.1 Diagram Alur Proses Penelitian.....	5
3.2. Pengumpulan Data.....	5
3.3. Variabel Penelitian.....	5
3.4. Metode Data Preprocessing.....	6
3.5. Metode Reduksi Dimensi.....	7
3.5.1 Principal Component Analysis.....	7
3.5.2 Loadings (Variabel Kontribusi).....	7
3.6. Metode Analisis Data.....	7
3.6.1. K-Means Clustering.....	8
3.6.2. Density-Based Spatial Clustering of Application with Noise.....	8
BAB IV.....	9
HASIL ANALISIS DAN PEMBAHASAN.....	9
4.1. Deskripsi Data.....	9
4.2. Data Pre-Processing.....	9
4.3. Modelling.....	13
4.3.1. K-Means Clustering.....	13
4.3.2. DBSCAN Clustering.....	18
4.4. Evaluasi Metode.....	23
BAB V.....	24
KESIMPULAN DAN SARAN.....	24
5.1 Kesimpulan.....	24
5.2 Saran.....	24
DAFTAR PUSTAKA.....	25
LAMPIRAN.....	26

DAFTAR GAMBAR

Gambar 1. Diagram Alur Penelitian.....	5
Gambar 2. Code Check Duplicate Data.....	9
Gambar 3. Code Check Duplicate Data.....	9
Gambar 4. Code Check Missing Value.....	10
Gambar 5. Code Checking Outlier.....	10
Gambar 6. Code Explore Data Kecamatan.....	11
Gambar 7. Code Explore Data Gender.....	11
Gambar 8. Code Scalling Data Numerik.....	11
Gambar 9. Code Dummy Variabel Kategorik.....	12
Gambar 10. Code PCA.....	12
Gambar 11. Visualisasi PCA.....	13
Gambar 12. Visualisasi Elbow Method.....	14
Gambar 13. Code K-Means Clustering.....	14
Gambar 14. Code K-Means Clustering.....	14
Gambar 15. Visualisasi K-Means Clustering.....	15
Gambar 16. Visualisasi Boxplot (K-Means Clustering).....	17
Gambar 17. Visualisasi Histogram (K-Means Clustering).....	18
Gambar 18. Code K-Distance Graph.....	18
Gambar 19. Code DBSCAN Clustering.....	19
Gambar 20. Visualisasi DBSCAN Clustering.....	19
Gambar 21. Visualisasi Boxplot (DBSCAN Clustering).....	22
Gambar 22. Visualisasi Histogram (DBSCAN Clustering).....	22
Gambar 23. Code Silhouette Score (K-means Clustering).....	23
Gambar 24. Code Silhouette Score (DBSCAN Clustering).....	23

DAFTAR TABEL

Tabel 1. Variabel Penelitian.....	6
Tabel 2. Statistika Deskriptif (K- Means Clustering).....	16
Tabel 3. Statistika Deskriptif (DBSCAN Clustering).....	21

BAB I

PENDAHULUAN

1.1. Latar Belakang

Kos merupakan salah satu pilihan akomodasi yang populer bagi mahasiswa yang mencari tempat tinggal yang terjangkau dan nyaman. Berdasarkan data dari Kementerian Riset, Teknologi, dan Pendidikan Tinggi (Kemendiknas), jumlah mahasiswa di Indonesia pada tahun 2022 mencapai lebih dari 9 juta orang. Sebagian besar dari mereka membutuhkan tempat tinggal yang ekonomis dekat kampus untuk menghemat. Hal ini menciptakan permintaan yang tinggi terhadap kos, terutama di kota-kota besar yang salah satunya adalah Surabaya.

Kos biasanya menawarkan berbagai fasilitas yang dapat mempengaruhi harga, seperti kamar mandi dalam, AC, WiFi, dan akses 24 jam. Faktor keamanan, kebersihan, dan lingkungan sekitar juga menjadi pertimbangan penting dalam pemilihan kos.

Dengan semakin banyaknya pilihan kos yang tersedia, terutama di kota-kota besar, calon penghuni seringkali menghadapi kesulitan dalam menemukan kos yang paling sesuai dengan kebutuhan dan anggaran mereka. Oleh karena itu, metode clustering dapat digunakan untuk mengelompokkan kos berdasarkan harga dan fasilitas, sehingga memudahkan pencarian kos yang ideal.

Web Mamikos, sebagai salah satu platform pencarian kos terkemuka di Indonesia, menyediakan database yang luas dengan lebih dari 2 juta kamar kos yang tersebar di banyak kota di seluruh Indonesia. Platform ini memanfaatkan teknologi untuk memberikan informasi yang akurat tentang ketersediaan kamar, detail fasilitas dan harga.

Penelitian ini bertujuan untuk menerapkan algoritma clustering, seperti K-means, untuk mengelompokkan data kos dari Mamikos berdasarkan harga dan fasilitas. Dengan demikian, penelitian ini diharapkan dapat memberikan wawasan baru dalam pencarian kos yang lebih efisien dan ekonomis.

Dengan menggunakan metode clustering, kita dapat mengidentifikasi pola dan tren dalam data kos, yang pada akhirnya dapat membantu calon penghuni menemukan kos yang paling sesuai dengan preferensi dan kebutuhan mereka, sambil tetap mempertimbangkan anggaran yang mereka miliki.

1.2. Rumusan Masalah

1. Bagaimana mengidentifikasi dan mengelompokkan (clustering) kos berdasarkan harga dan fasilitas yang mempengaruhi?
2. Bagaimana menerapkan algoritma K-Means dan DBSCAN untuk mencari biaya kos yang paling ekonomis dan sesuai dengan kebutuhan fasilitas?
3. Bagaimana hasil dari pola-pola visualisasi yang akan dihasilkan dari metode clustering serta informasi apa yang dihasilkan?

1.3. Manfaat Penelitian

1. Meningkatkan efisiensi pengguna dalam pencarian kos.
2. Membantu pengguna dalam menemukan opsi kos yang ekonomis.
3. Mempermudah pengelola kos dalam menyusun strategi pasar.
4. Menyediakan basis data yang berguna untuk penelitian selanjutnya.

1.4. Tujuan Penelitian

1. Mengidentifikasi dan menampilkan kelompok kos berdasarkan harga dan fasilitas yang mempengaruhi.
2. Mengetahui penerapan metode K-Means dan DBSCAN clustering untuk mencari biaya kos paling ekonomis dan sesuai dengan kebutuhan fasilitas.
3. Menghasilkan informasi mengenai dan pola visualisasi kos terbaik yang disesuaikan berdasarkan hasil clustering.

BAB II

TINJAUAN PUSTAKA

2.1. Data Mining

Data mining adalah proses mengekstraksi dan mengidentifikasi informasi yang bermanfaat serta mengumpulkan pengetahuan dari kumpulan data besar dengan menggunakan pendekatan statistik, matematika, kecerdasan buatan, dan pembelajaran mesin (Ginting et al., 2019). Ada berbagai subbidang dalam data mining, seperti penemuan pengetahuan (knowledge discovery) dan pengenalan pola (pattern recognition) (Bastian et al., n.d.).

2.2. Clustering

Clustering merupakan salah satu teknik utama dalam penambangan data (data mining). Algoritma clustering berfungsi dengan mengelompokkan data ke dalam kumpulan data yang relevan (Studi & Informatika, 2022). Proses ini melibatkan pemisahan satu set objek data ke dalam kelompok-kelompok terpisah (Apulino Iman Seno Aji et al., 2021).

2.3. K-Means

K-Means adalah teknik data mining yang digunakan untuk mengelompokkan data ke dalam satu atau lebih cluster. Data dengan karakteristik serupa ditempatkan dalam cluster yang sama, sedangkan data dengan karakteristik berbeda ditempatkan dalam cluster yang berbeda namun serupa (Dinata et al., 2020). Algoritma K-Means melibatkan dua langkah utama: mengidentifikasi lokasi pusat setiap cluster dan menentukan anggota dari setiap cluster tersebut (Sagala, 2021).

2.4. PCA (*Principal Component Analysis*)

PCA adalah teknik analisis statistik yang digunakan untuk mengurangi dimensi data dengan mempertahankan sebanyak mungkin variasi dalam data. Dengan mereduksi jumlah dimensi, PCA membantu dalam visualisasi dan pemrosesan data yang lebih efisien (Jolliffe, 2002). Langkah-langkah untuk menerapkan PCA sebelum menggunakan DBSCAN adalah sebagai berikut:

1. Standardisasi data.
2. Menerapkan PCA untuk mengurangi dimensi data.

3. Menggunakan komponen utama hasil PCA untuk mengelompokkan data dengan DBSCAN

2.5. DBSCAN

DBSCAN adalah teknik data mining yang digunakan untuk mengelompokkan data berdasarkan kepadatan. Data yang terletak di area dengan kepadatan tinggi dikelompokkan dalam cluster yang sama, sementara data yang terletak di area dengan kepadatan rendah dianggap sebagai noise atau outlier (GeeksforGeeks, 2023). Algoritma DBSCAN melibatkan dua parameter utama: *eps*, yang menentukan radius lingkungan sekitar setiap titik data, dan *MinPts*, jumlah minimum titik data yang diperlukan untuk membentuk sebuah cluster (scikit-learn, 2024).

2.6 Kost

Kost adalah layanan penyediaan akomodasi yang menawarkan kamar atau ruang hunian dengan biaya sewa yang ditetapkan untuk periode tertentu, biasanya dibayar setiap bulan. Layanan kost seringkali mencakup fasilitas tambahan seperti listrik, air, dan internet, serta dapat menyediakan pilihan kamar dengan berbagai ukuran dan fasilitas untuk memenuhi kebutuhan beragam penyewa (Suryadi, 2022). Konsep kost memungkinkan individu, terutama pelajar dan pekerja, untuk menemukan tempat tinggal yang terjangkau dan nyaman di lokasi yang strategis (Hartanto, 2023).

2.7 Pengaruh Harga dan Fasilitas pada Kost

Harga dan fasilitas merupakan faktor penting yang mempengaruhi keputusan individu dalam memilih kost. Harga yang kompetitif sering kali menjadi pertimbangan utama bagi penyewa, terutama bagi pelajar dan pekerja dengan anggaran terbatas (Suryadi, 2022). Fasilitas yang ditawarkan, seperti listrik, air, dan internet, serta fasilitas tambahan seperti keamanan, area parkir, dan dapur bersama, juga berperan penting dalam menarik penyewa yang mencari kenyamanan dan kemudahan (Hartanto, 2023).

BAB III

METODOLOGI PENELITIAN

3.1 Diagram Alur Proses Penelitian

Diagram alur ini menggambarkan langkah-langkah metodologi yang digunakan dalam penelitian ini, mulai dari pengumpulan data hingga analisis akhir.



Gambar 1. Diagram Alur Penelitian

3.2. Pengumpulan Data

Pengumpulan data adalah rangkaian langkah yang sistematis dan terstruktur yang dilakukan untuk mengumpulkan informasi atau data. Data yang digunakan dalam penelitian ini diperoleh melalui teknik web scraping yang dapat diakses di <https://mamikos.com/>. *Web scraping* adalah metode untuk mengekstrak dan mengumpulkan data dari situs web. Kami mengumpulkan informasi terbaru dan akurat tentang harga, fasilitas, dan lokasi kos yang tersedia di Mamikos.

3.3. Variabel Penelitian

Untuk menentukan clustering kos berdasarkan harga, fasilitas dan jenis kos, kita perlu memahami variabel-variabel yang akan digunakan, yaitu:

Nama Variabel	Tipe Data	Deskripsi Variabel
<i>Name</i>	Kategorik	Nama dari kos yang berada di sekitar Universitas Airlangga
<i>Price</i>	Numerik	Harga kos per bulan
Kecamatan	Kategorik	Lokasi kos berdasarkan kecamatan
<i>Gender</i>	Kategorik	Tipe kos berdasarkan jenis kelamin

		(khusus laki-laki, khusus perempuan, atau campur (laki-laki dan perempuan))
Akses 24 Jam	Biner	Ketersediaan fasilitas akses 24 jam pada kos
Kasur	Biner	Ketersediaan fasilitas kasur pada kos
AC	Biner	Ketersediaan fasilitas AC pada kos
Kamar Mandi Dalam	Biner	Ketersediaan fasilitas kamar mandi dalam pada kos
Kloset Duduk	Biner	Ketersediaan fasilitas kloset duduk pada kos
WiFi	Biner	Ketersediaan fasilitas WiFi pada kos

Tabel 1. Variabel Penelitian

3.4. Metode Data *Preprocessing*

Kumpulan data ini akan diproses memerlukan pekerjaan awal dikarenakan memiliki masalah seperti missing value dan lainnya sehingga diperlukannya preprocessing data meliputi :

1. Data Cleaning

Dilakukan untuk menghilangkan data noise (data yang tidak relevan) dan data yang tidak konsisten

2. Transformasi Data

Dalam proses ini, kualitas dari data mining dapat ditentukan dari transformasi data. Pada metode clustering misalnya, pada variabel 'price' perlu dilakukan scaling data agar berada dalam rentang 0 hingga 1 atau perubahan data kategorik menjadi dummy variabel untuk mewakili kategori atau kelompok dalam data

3. Principal Component Analysis

Teknik ini digunakan untuk mereduksi dimensi dari dataset untuk mengidentifikasi variabel-variabel yang penting, yang paling menjelaskan sebagian besar varians dalam dataset

3.5. Metode Reduksi Dimensi

3.5.1 Principal Component Analysis

Analisis komponen utama (PCA) adalah metode statistik multivariat yang menggabungkan informasi dari beberapa variabel yang diamati pada subjek yang sama menjadi variabel yang lebih sedikit, disebut komponen utama (PCs). Langkah - Langkah dari PCA yaitu :

1. Standarisasi variabel
2. Reduksi dimensi
3. Scaling dan Interpretasi dari Biplot
4. Opsi untuk Mengurangi Penekanan pada Kasus atau Variabel dalam Biplot
5. Opsional Penambahan Variabel Tambahan ke dalam Biplot

3.5.2 Loadings (Variabel Kontribusi)

Loadings adalah koefisien dari kombinasi linear yang digunakan untuk membuat komponen utama. Secara matematis, loadings adalah elemen-elemen dari matriks eigenvector V (arah dari variasi maksimum data) yang diskalakan dengan akar dari eigenvalues L (nilai yang mengukur seberapa besar variasi dalam data sepanjang arah eigenvector). Ini berarti bahwa loadings menunjukkan seberapa besar kontribusi setiap variabel asli terhadap komponen utama.

$$Loadings = Eigenvector \times \sqrt{Eigenvalue}$$

Dalam konteks PCA, Loadings dari *eigenvectors* dan *eigenvalues* memungkinkan kita untuk memahami fitur mana yang paling mempengaruhi komponen utama. Ini membantu dalam menginterpretasikan data dan menemukan pola yang signifikan.

3.6. Metode Analisis Data

Menurut Han dkk. (2011: 444), analisis cluster adalah proses membagi sekumpulan objek data menjadi beberapa subset. Setiap subset adalah sebuah cluster di mana objek-objek di dalamnya memiliki kesamaan satu sama lain, tetapi berbeda dari objek-objek di cluster lainnya. Dalam hal ini, analisis klaster digunakan untuk mengelompokkan harga kos berdasarkan fasilitas yang tersedia. Dengan tujuan untuk memberikan informasi mengenai kelompok kos, serta membantu pengguna menemukan kos yang ekonomis dan sesuai dengan kebutuhan fasilitasnya.

3.6.1. K-Means Clustering

K-Means merupakan metode yang sederhana untuk analisis clustering. Algoritma K-Means menggunakan centroid untuk membentuk cluster. Centroid merupakan titik tengah dari sebuah cluster dan berupa nilai. Centroid ini digunakan untuk menghitung jarak antara objek data dan centroid tersebut. Sebuah objek data dimasukkan ke dalam sebuah cluster jika jaraknya paling dekat dengan centroid cluster tersebut. Langkah-langkah k-means clustering adalah menentukan jumlah cluster yang ingin dibentuk. Objek atau elemen awal dalam cluster dapat dipilih secara acak sebagai titik tengah (centroid) cluster. Algoritma K-Means kemudian mengulangi langkah-langkah berikut hingga tercapai kestabilan (tidak ada objek yang dipindahkan lagi):

1. Menentukan koordinat titik tengah setiap cluster.
2. Menghitung jarak setiap objek ke koordinat titik tengah.
3. Mengelompokkan objek-objek berdasarkan jarak terdekatnya ke centroid.

3.6.2. Density-Based Spatial Clustering of Application with Noise

DBSCAN adalah algoritma pengelompokan yang mengidentifikasi wilayah-wilayah berkepadatan tinggi dan membaginya menjadi kelompok dengan memanfaatkan ϵ (epsilon) sebagai radius ketegangan dan MinObj sebagai batas minimum kepadatan objek. Algoritma ini dapat mendeteksi outlier atau noise, menghasilkan cluster lebih akurat untuk data dalam jumlah besar, dan tidak memerlukan penentuan jumlah cluster di awal. Langkah-langkah DBSCAN dengan menggunakan dataset hasil PCA adalah:

1. Menggunakan dataset yang telah direduksi dimensinya dengan PCA untuk meningkatkan efisiensi komputasi.
2. Menentukan nilai epsilon (ϵ) dan MinObj yang optimal untuk menentukan parameter DBSCAN.
3. Menerapkan DBSCAN dengan mengelompokkan objek-objek berdasarkan kepadatan objek yang memenuhi kriteria epsilon (ϵ) dan MinObj.
4. Objek yang tidak termasuk dalam cluster dianggap sebagai noise atau outlier.

BAB IV

HASIL ANALISIS DAN PEMBAHASAN

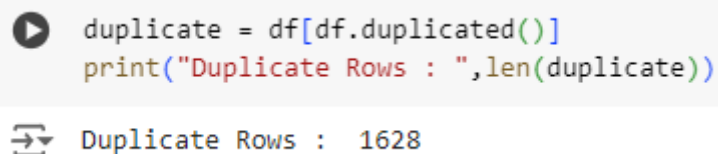
4.1. Deskripsi Data

Data yang digunakan terdiri dari 240 kos dan terdapat 10 variabel yang tersedia untuk kos. Variabel tersebut terdiri dari *name*, *price*, kecamatan, *gender*, akses 24 jam, kasur, AC, kamar mandi dalam, kloset duduk dan Wi-Fi. Variabel-variabel tersebut menunjukkan berbagai aspek yang dapat digunakan untuk menganalisis pengelompokan harga kos. Tipe data untuk variabel harga adalah data numerik, variabel *name*, kecamatan dan *gender* adalah data kategori sedangkan untuk variabel akses 24 jam, kasur, AC, kamar mandi dalam, kloset duduk dan Wi-Fi.

4.2. Data Pre-Processing

Sebelum dilakukan metode klaster, perlu dilakukan data preprocessing untuk mengatasi masalah pada data seperti inkonsistensi data, data tidak lengkap dan redundansi data yang terdapat saat *scrapping* pada tahap awal. Data preprocessing ini juga melakukan perubahan status barang yang semula *true* / *false* menjadi 1 / 0. Tahap preprocessing yang kita lakukan sebagai berikut :

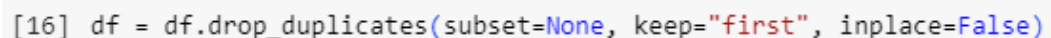
1. Check Duplicate Data



```
duplicate = df[df.duplicated()]
print("Duplicate Rows : ",len(duplicate))
```

➡ Duplicate Rows : 1628

Gambar 2. Code Check Duplicate Data



```
[16] df = df.drop_duplicates(subset=None, keep="first", inplace=False)
```

Gambar 3. Code Check Duplicate Data

Ditemukan 1628 duplikasi data dalam dataset. Setelah dilakukan penghapusan pada duplikasi data, masih tetap ditemukan beberapa nilai yang hilang (*null*) pada kolom 'kos_andalan', 'gender', dan 'rating'. Oleh karena itu perlu dilakukan penghapusan pada kolom tersebut.

2. Checking Missing Value

```
missing_values = df.isnull().sum()
missing_percentage = (df.isnull().sum() / len(df)) * 100

print(f'Missing Values:\n{missing_values}\nPersentase:\n{missing_percentage}')
```

Gambar 4. Code Check Missing Value

Terdapat missing value pada variabel 'rating', 'gender' dan 'kos_andalan'. Dikarenakan missing value yang ditemukan pada variabel 'rating' dan 'kos_andalan' cukup banyak maka diperlukan penghapusan variabel pada data ini.

3. Checking Outlier

```
[20] def cek_outlier(df, column):
      Q1 = df[column].quantile(0.25)
      Q3 = df[column].quantile(0.75)
      IQR = Q3 - Q1
      lower_bound = Q1 - 1.5 * IQR
      upper_bound = Q3 + 1.5 * IQR
      outliers = df[(df[column] < lower_bound) | (df[column] > upper_bound)]
      return outliers

      cek_outlier(df, 'price')
```

Gambar 5. Code Checking Outlier

Dilakukan pengecekan pencilan data pada variabel price menggunakan interquartile range (IQR) method guna mengetahui rentangan nilai dari masing masing kolom dan apakah terdapat outlier atau tidak.

4. Explore Data Kecamatan & Gender

```
[23] konsistensi_kecamatan = {  
      'Surabaya Gubeng': 'Kecamatan Gubeng',  
      'Surabaya Tambaksari': 'Kecamatan Tambaksari',  
      'Surabaya Tegalsari': 'Kecamatan Tegalsari',  
      'Surabaya Genteng': 'Kecamatan Genteng',  
      'Surabaya Wonokromo': 'Kecamatan Wonokromo',  
      'Surabaya Mulyorejo': 'Kecamatan Mulyorejo',  
      'Surabaya Sukolilo': 'Kecamatan Sukolilo'  
    }  
  
df['kecamatan'] = df['kecamatan'].replace(konsistensi_kecamatan)
```

Gambar 6. Code Explore Data Kecamatan

Karena masih terdapat inkonsistensi data pada variabel kecamatan seperti 'Kecamata Mulyorejo' dan 'Surabaya Mulyorejo' yang seharusnya satu kesatuannya sama.

```
[ ] df['gender'].unique()  
  
array(['Campur', 'Putri', 'Putra'], dtype=object)
```

Gambar 7. Code Explore Data Gender

Menampilkan value dari variabel 'gender'

5. Scalling Data Numerik

```
[27] scaler = StandardScaler()  
  
df_cleaned['price'] = scaler.fit_transform(df_cleaned['price'].values.reshape(-1, 1))
```

Gambar 8. Code Scalling Data Numerik

Menskala variabel 'price' antara 0 dan 1 untuk membantu peningkatan kinerja algoritma supaya bekerja lebih baik. Ketidakstabilan dapat terjadi ketika variabel memiliki value yang jauh lebih besar atau lebih kecil dibandingkan yang lain.

6. Dummy Variabel Kategorik

```
df_cleaned = pd.get_dummies(df_cleaned, columns=['kecamatan', 'gender'])

# Ganti nilai True/False menjadi 1/0 hanya untuk kolom 'kecamatan' dan 'gender'
kecamatan_columns = [col for col in df_cleaned.columns if col.startswith('kecamatan_')]
gender_columns = [col for col in df_cleaned.columns if col.startswith('gender_')]

# Ubah nilai True/False menjadi 1/0 menggunakan looping
for col in kecamatan_columns + gender_columns:
    df_cleaned[col] = df_cleaned[col].astype(int)
```

Gambar 9. Code Dummy Variabel Kategorik

Tujuan dari dummy ini agar setiap kategori dapat diwakili oleh satu dummy variabel, sehingga model dapat mempelajari efek unik dari setiap kategori.

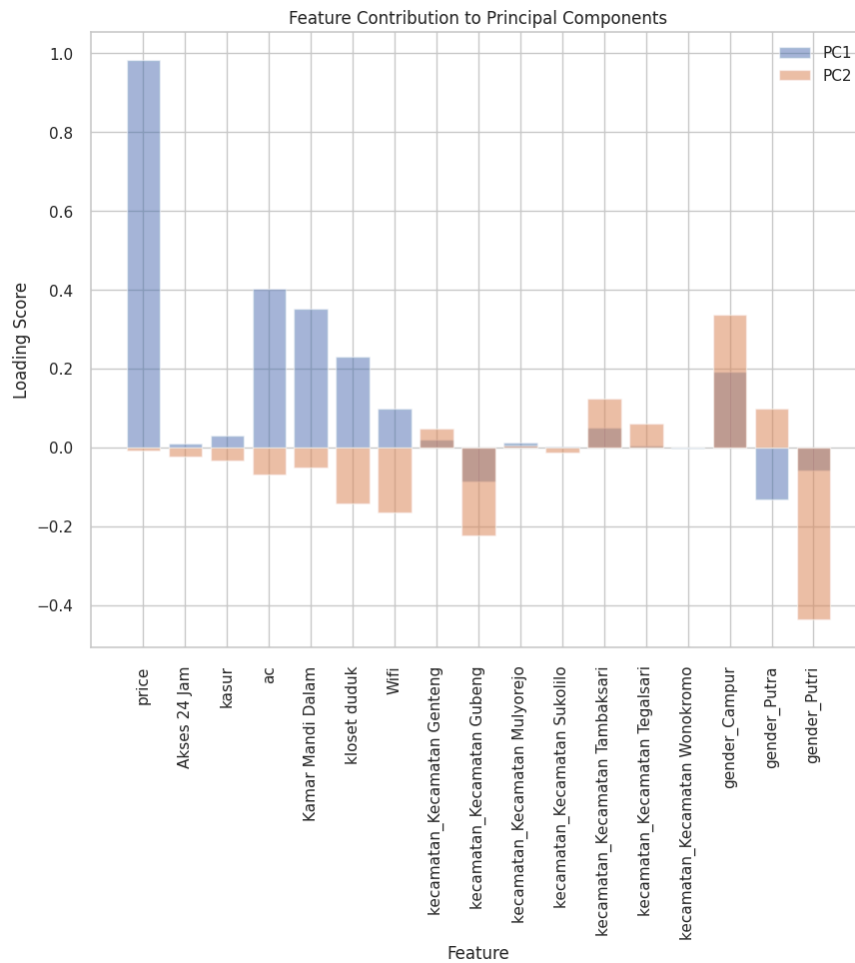
7. Principal Component Analysis

```
# PCA
pca = PCA(n_components=2)
pca_result = pca.fit_transform(df_pca)

pca_df = pd.DataFrame(data=pca_result, columns=['PC1', 'PC2'])
```

Gambar 10. Code PCA

Tujuan dari PCA sendiri adalah untuk mengurangi dimensi data dengan mempertahankan sebagian besar varians asli. Dikarenakan K-Means dan DBSCAN dapat terganggu kinerjanya jika data memiliki dimensi yang tinggi.



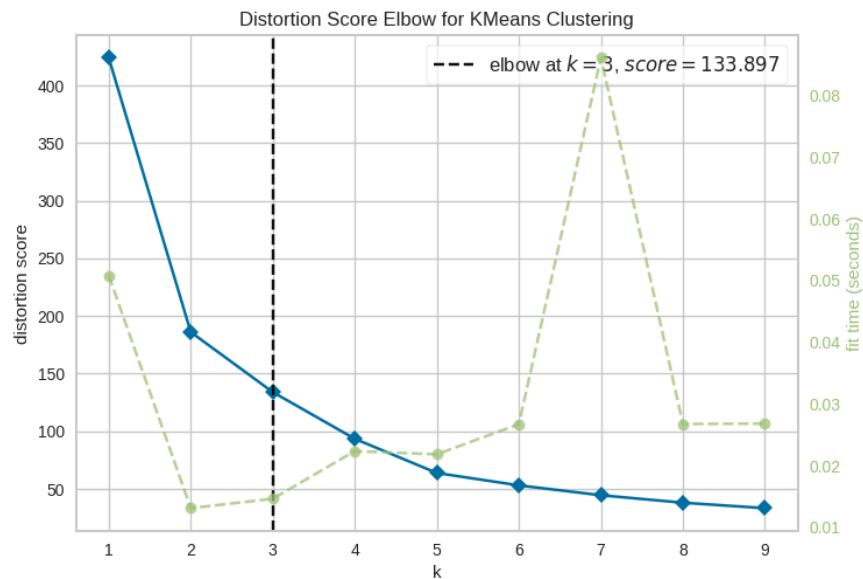
Gambar 11. Visualisasi PCA

Visualisasi dari hasil PCA dengan nilai eigen menunjukkan bagaimana setiap fitur yang ada dalam dataset berkontribusi terhadap komponen utama dalam PCA. Variabel 'price', 'Akses 24 jam', 'kasur', 'ac', dan 'kamar mandi dalam' berkontribusi terhadap PC1. Variabel yang berkontribusi paling besar pada PC1 adalah variabel 'price'. Sedangkan variabel 'kecamatan_Kecamatan Genteng', 'kecamatan_Kecamatan Tambaksari', 'kecamatan_Kecamatan Tegalsari', 'gender_Campur' dan 'gender_Putra' berkontribusi terhadap PC2. Variabel yang berkontribusi positif paling besar pada PC2 adalah 'gender_Campur'.

4.3. Modelling

4.3.1. K-Means Clustering

1. Elbow Method



Gambar 12. Visualisasi Elbow Method

Sebelum melakukan clustering, perlu menentukan jumlah kluster optimal yang bisa dibentuk menggunakan metode elbow. Didapatkan bahwa jumlah kluster optimal yang bisa dibentuk adalah sebanyak 3.

2. Modelling

```
kmeans = KMeans(n_clusters=3, random_state=42)
cluster_labels = kmeans.fit_predict(pca_df)
print(f'Hasil Clustering:\n{cluster_labels}')
```

Gambar 13. Code K-Means Clustering

```
pd.Series(kmeans.labels_).value_counts()
```

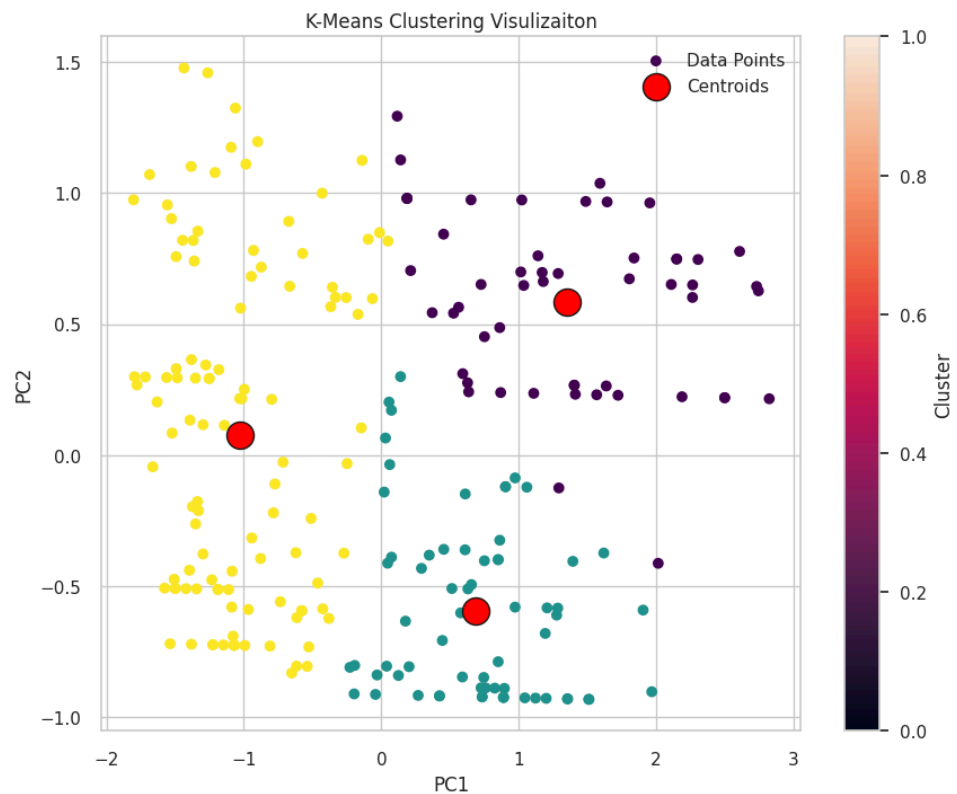
2	114
1	66
0	53

Name: count, dtype: int64

Gambar 14. Code K-Means Clustering

Hasil dari K-Means clustering menunjukkan 3 kluster yang terbentuk, jumlah data dari kluster 0 adalah 53, kluster 1 adalah 66 dan kluster 2 adalah 114.

3. Visualisasi Clustering



Gambar 15. Visualisasi K-Means Clustering

Plot menunjukkan visualisasi dari hasil clustering menggunakan algoritma K-Means dengan 3 cluster berdasarkan fitur PC1 dan PC2. Masing-masing cluster digambarkan dengan warna yang berbeda (ungu, hijau dan kuning). Titik-titik merah adalah centroid atau pusat dari masing-masing cluster.

4. Statistika Deskriptif Clustering

	Cluster 0	Cluster 1	Cluster 2
Jumlah Data	53	66	114
Harga Rata - Rata	1.809.693	1.517.927	765.029
Harga Minimum	1.200.000	900.000	400.000
Harga Maksimum	2.708.750	2.300.000	1.300.000
Rata-rata PC1	0.29908622019 095765	0.10877432461 154608	-0.19555838808 675524

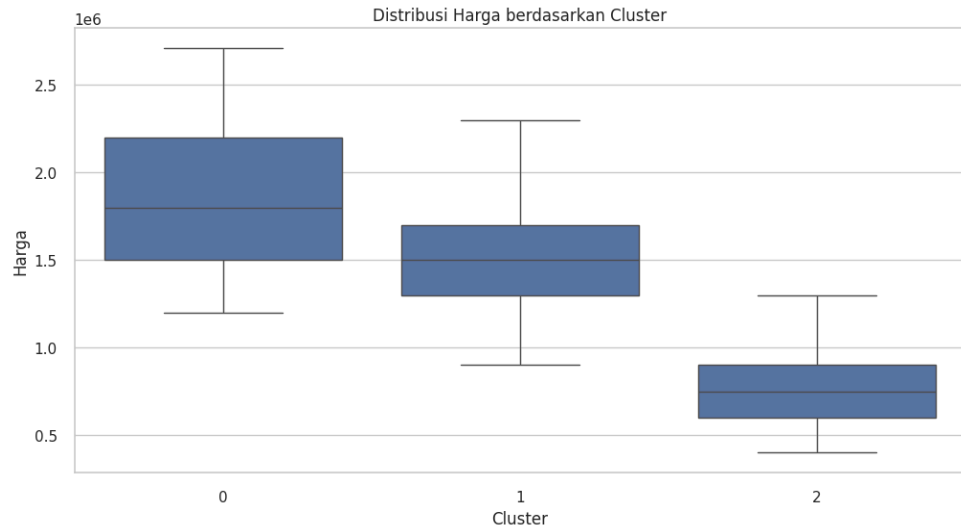
Rata-rata PC2	0.07716619931 161514	-0.04839445596 450456	-0.01168299666 5063611
Akses 24 Jam (1)	42	50	85
Akses 24 Jam (0)	11	16	29
Kasur (1)	51	66	105
Kasur (0)	2	0	9
AC (1)	51	61	15
AC (0)	2	5	99
Kamar Mandi Dalam (1)	44	52	24
Kamar Mandi Dalam (0)	9	14	90
Kloset Duduk (1)	45	63	50
Kloset Duduk (0)	8	3	64
Wi-Fi (1)	40	61	79
Wi-Fi (0)	13	5	35

Tabel 2. Statistika Deskriptif (K- Means Clustering)

- Cluster 0 terdiri dari 53 kos dengan harga rata rata per bulannya adalah Rp. 1.809.639, harga minimumnya Rp. 1.220.000 dan harga maksimumnya Rp. 2.780.000. Sebagian kos memiliki fasilitas akses 24 jam, kasur, AC, kamar mandi dalam, kloset duduk, dan WiFi.
- Cluster 1 terdiri dari 66 kos dengan harga rata rata per bulannya adalah Rp. 1.517.927, harga minimumnya Rp. 970.000 dan harga maksimumnya Rp. 2.708.750. Sebagian besar kos memiliki fasilitas akses 24 jam, kasur, AC, kamar mandi dalam, kloset duduk, dan WiFi.
- Cluster 2 terdiri dari 114 kos dengan harga rata rata sebesar Rp. 765.029, harga minimumnya Rp. 325.000 dan harga maksimumnya

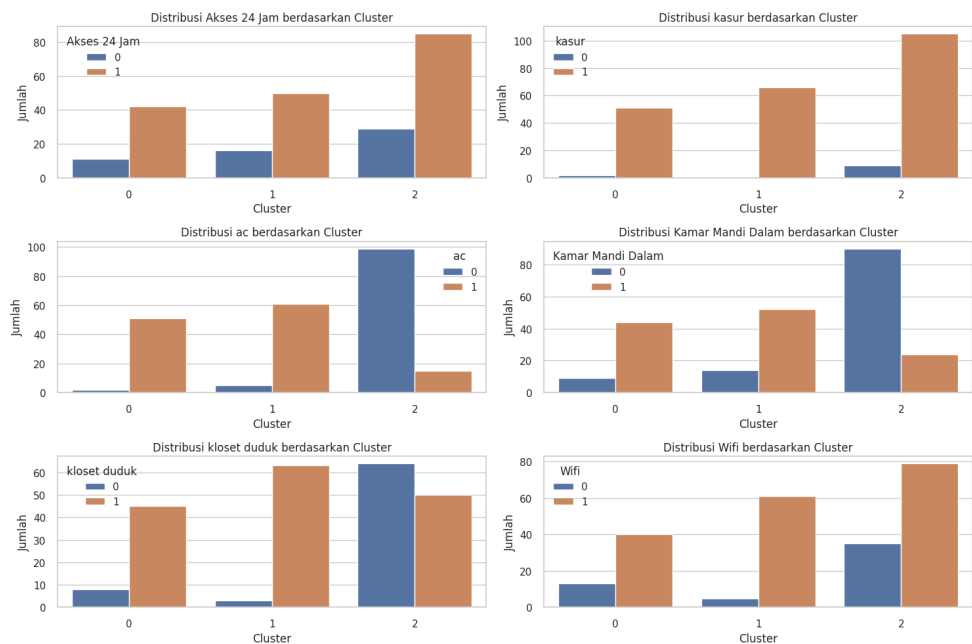
Rp. 1.100.000. Sebagian besar kos memiliki fasilitas akses 24 jam, kasur, AC, kamar mandi dalam, kloset duduk, dan WiFi, meskipun persentasenya mungkin lebih rendah dibandingkan cluster lainnya.

5. Visualisasi Pola Clustering



Gambar 16. Visualisasi Boxplot (K-Means Clustering)

Cluster 0 memiliki harga yang paling tinggi dengan rentang yang paling luas, sementara cluster 2 memiliki harga yang paling rendah dengan rentang yang paling sempit. Cluster 1 berada di tengah dengan harga dan rentang yang relatif sedang.



Gambar 17. Visualisasi Histogram (K-Means Clustering)

- Fasilitas akses 24 jam banyak dimiliki oleh cluster 0, cluster 1 dan cluster 2.
- Fasilitas kasur banyak dimiliki oleh cluster 0, cluster 1 dan 2.
- Fasilitas ac banyak dimiliki oleh cluster 0 dan 1, sedangkan pada cluster 2 banyak yang tidak memiliki fasilitas ac.
- Fasilitas kamar mandi dalam banyak dimiliki oleh cluster 0 dan 1, sedangkan pada cluster 2 banyak yang tidak memiliki kamar mandi dalam.
- Fasilitas kloset duduk banyak dimiliki oleh cluster 0 dan cluster 1, sedangkan pada cluster 2 banyak yang tidak memiliki kamar mandi dalam
- Fasilitas Wi-Fi banyak dimiliki oleh cluster 0, cluster 1, dan cluster 2

4.3.2. DBSCAN Clustering

1. K-Distance Graph

```
[ ] # Menentukan eps menggunakan k-distance graph
min_samples = 3

neighbors = NearestNeighbors(n_neighbors=min_samples)
neighbors_fit = neighbors.fit(df_dbscan)
distances, indices = neighbors_fit.kneighbors(df_dbscan)

# Sort the distances (of the k-th nearest neighbor)
distances = np.sort(distances[:, min_samples - 1])

# Menggunakan algoritma Kneedle untuk menemukan eps optimal
kneedle = Kneelocator(range(len(distances)), distances, S=1.0, curve='convex', direction='increasing')
optimal_eps = distances[kneedle.elbow]

print(f'Epsilon optimal: {optimal_eps}')
print(f'MinPts optimal: {min_samples}')
```

Epsilon optimal: 0.3104176356110685
MinPts optimal: 3

Gambar 18. Code K-Distance Graph

Dengan menggunakan k-distance graph dan algoritma *kneedle* didapatkan nilai epsilon optimal sebesar 0.3104176356110685 dan jumlah minimum titik dalam radius epsilon yaitu 3.

2. Modelling

```

# Terapkan DBSCAN dengan nilai eps optimal
db = DBSCAN(eps=optimal_eps, min_samples=min_samples).fit(df_dbscan)
labels = db.labels_

# Jumlah cluster dan noise
n_clusters = len(set(labels)) - (1 if -1 in labels else 0)
n_noise = list(labels).count(-1)

print('Estimated number of clusters: %d' % n_clusters)
print('Estimated number of noise points: %d' % n_noise)

```

```

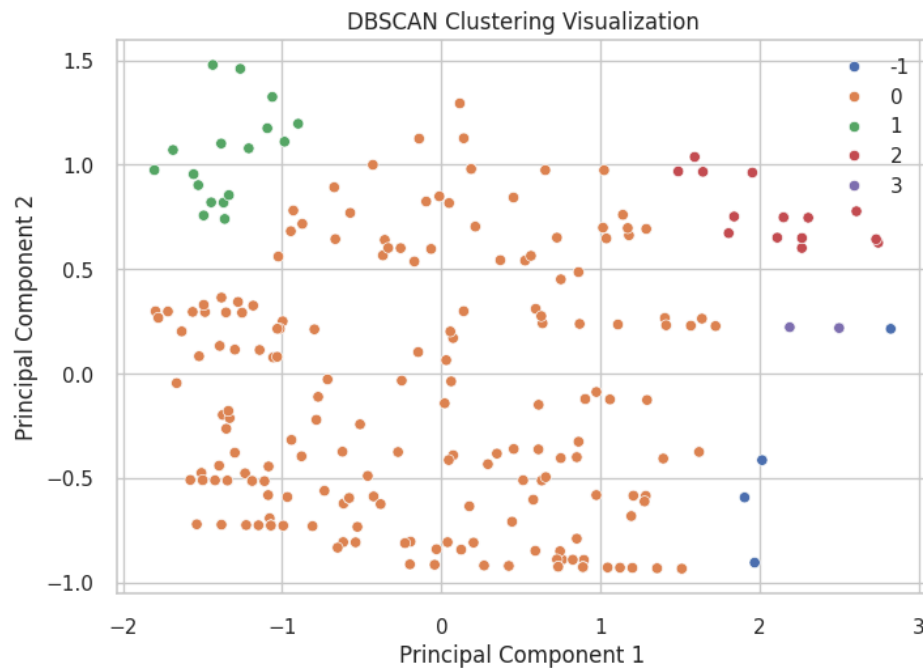
Estimated number of clusters: 4
Estimated number of noise points: 4

```

Gambar 19. Code DBSCAN Clustering

DBSCAN clustering menunjukkan bahwa jumlah klaster yang bisa dibentuk adalah sebanyak 4 klaster dan terdapat 4 titik yang merupakan noise. Noise adalah titik yang tidak termasuk dalam cluster manapun, karena tidak memenuhi kriteria jarak atau jumlah titik minimum.

3. Visualisasi Clustering



Gambar 20. Visualisasi DBSCAN Clustering

Hasil dari visualisasi DBSCAN dengan 4 cluster, yaitu cluster 0, 1, 2, dan 3 berdasarkan fitur PC1 dan PC2. Masing-masing cluster dibedakan dengan warna yang berbeda (hijau, merah, oranye dan ungu). Serta terdapat 4 noise yang ditandai dengan label -1.

4. Statistika Deskriptif

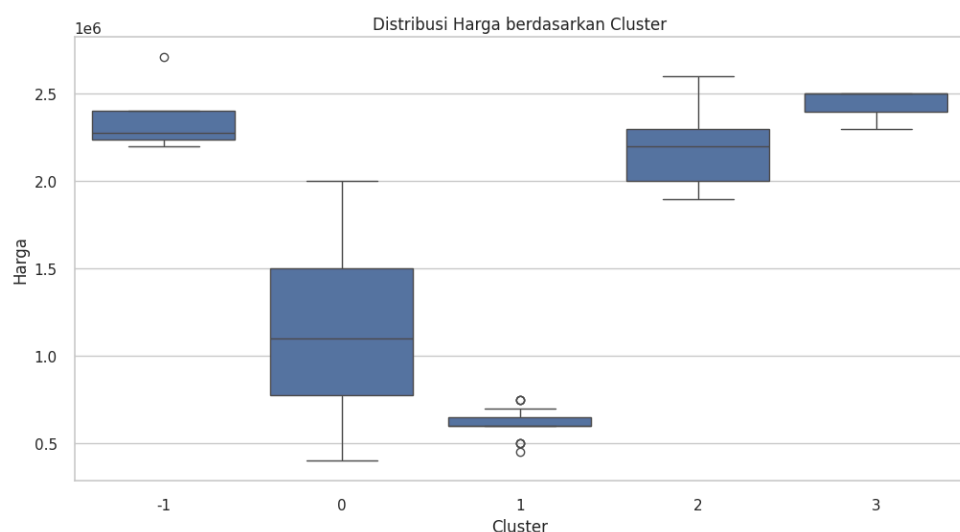
	Label -1 (Noise)	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Jumlah Data	4	194	17	15	3
Harga rata - rata	2.364.687,5	1.149.235	611.764	2.213.333	2.433.333
Harga Minimum	2.200.000	400.000	450.000	1.900.000	2.300.000
Harga Maksimum	2.708.750	2.000.000	750.000	2.600.000	2.500.000
Rata - rata PC1	11405141917126900	-0.08128749305330829	-0.11026873575667073	0.7383851635190702	0.827521225088872
Rata - rata PC2	-0.17754847870820234	-0.007551799619089829	0.09603327013756853	0.0642072973696724	-0.3832931582198336
Akses 24 jam (1)	3	148	12	11	3
Akses 24 jam (0)	1	46	5	4	0
Kasur (1)	4	185	15	15	3
Kasur (0)	0	9	2	0	0
AC (1)	4	105	0	15	3
AC (0)	0	89	17	0	0
K. Mandi Dalam (1)	4	98	0	15	3
K. Mandi Dalam (0)	0	96	17	0	0
Kloset Duduk (1)	4	136	3	12	3
Kloset Duduk (0)	0	58	14	3	0
Wi-Fi (1)	4	153	6	14	3

Wi-Fi (0)	0	41	11	1	0
-----------	---	----	----	---	---

Tabel 3. Statistika Deskriptif (DBSCAN Clustering)

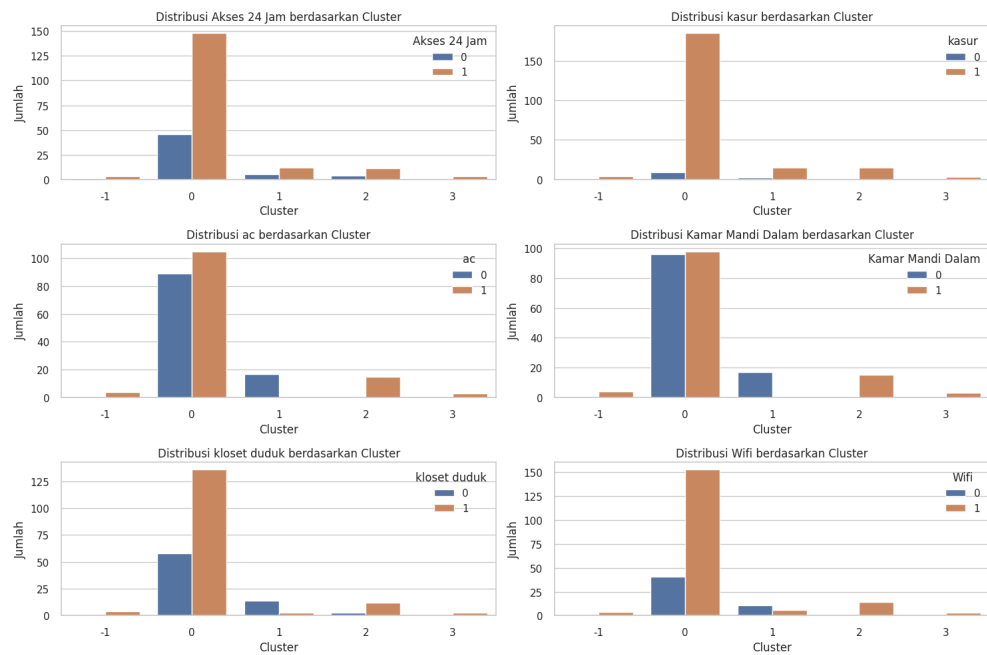
- Label -1 adalah noise, terdiri dari 4 kos dengan harga rata rata per bulannya adalah Rp. 2.364.687,5, harga minimumnya Rp. 2.200.000 dan harga maksimumnya Rp. 2.708.750. Sebagian besar kos memiliki seluruh fasilitas yang ada.
- Cluster 0 terdiri dari 194 kos dengan harga rata rata per bulannya adalah Rp. 1.149.235, harga minimumnya Rp. 400.000 dan harga maksimumnya Rp. 2.000.000. Sebagian besar kos memiliki seluruh fasilitas yang ada.
- Cluster 1 terdiri dari 17 kos dengan harga rata rata per bulannya adalah Rp. 611.764, harga minimumnya Rp. 450.000 dan harga maksimumnya Rp. 750.000. Sebagian besar kos tidak memiliki fasilitas AC, kloset duduk, kamar mandi dalam dan Wi-Fi.
- Cluster 2 terdiri dari 15 kos dengan harga rata rata per bulannya adalah Rp. 2.213.333, harga minimumnya Rp. 1.900.000, harga maksimumnya Rp. 2.600.000. Sebagian besar kos memiliki seluruh fasilitas yang ada.
- Cluster 3 terdiri dari 3 kos dengan harga rata rata per bulannya Rp. 2.433.333, harga minimumnya Rp. 2.300.000, harga maksimumnya Rp. 2.500.000. Semua kos memiliki seluruh fasilitas yang ada.

5. Visualisasi Pola Clustering



Gambar 21. Visualisasi Boxplot (DBSCAN Clustering)

Titik-titik di cluster -1 adalah noise, namun menunjukkan harga yang cenderung tinggi. Cluster 3 memiliki harga yang paling tinggi dengan rentang yang sempit, sementara cluster 1 memiliki rentang yang paling sempit dan harga yang paling rendah. Cluster 0 memiliki rentang harga yang paling luas. Cluster 2 memiliki harga yang cenderung tinggi dengan rentang yang cukup lebar, tetapi tidak se beragam cluster 0.



Gambar 22. Visualisasi Histogram (DBSCAN Clustering)

- Fasilitas akses 24 jam dimiliki oleh seluruh cluster.
- Fasilitas kasur dimiliki oleh seluruh cluster.
- Fasilitas AC dimiliki oleh seluruh cluster, kecuali pada cluster 1 tidak ada yang memiliki fasilitas AC.
- Fasilitas kamar mandi dalam banyak dimiliki oleh seluruh cluster, kecuali pada cluster 1 tidak ada yang memiliki fasilitas kamar mandi dalam.
- Fasilitas kloset duduk banyak dimiliki oleh seluruh cluster, kecuali pada cluster 1 banyak yang tidak memiliki fasilitas kloset duduk.
- Fasilitas Wi-Fi banyak dimiliki oleh semua cluster, kecuali pada cluster 1 banyak yang tidak memiliki fasilitas Wi-Fi.

4.4. Evaluasi Metode

1. Silhouette score metode K-means

```
[ ] kmeans_silhouette = silhouette_score(df_kmeans, cluster_labels)
    print(f"K-means Silhouette Score: {kmeans_silhouette}")
```

⇒ K-means Silhouette Score: 0.5297213082395246

Gambar 23. Code Silhouette Score (K-means Clustering)

2. Silhouette score pada metode DBSCAN

```
[ ] dbscan_silhouette = silhouette_score(df_dbscan, labels)
    print(f"DBSCAN Silhouette Score: {dbscan_silhouette}")
```

⇒ DBSCAN Silhouette Score: 0.2642574074353147

Gambar 24. Code Silhouette Score (DBSCAN Clustering)

Berdasarkan nilai Silhouette dari K-means Clustering dan DBSCAN Clustering, didapatkan bahwa metode K-means memiliki Silhouette Score yang lebih tinggi (0.529) daripada DBSCAN (0.264). Hal ini menunjukkan bahwa hasil clustering dari K-means lebih baik dalam memisahkan kluster dan memiliki titik-titik data yang lebih dekat dalam kluster yang sama dibandingkan dengan DBSCAN. Dalam hal ini, K-means lebih baik dibandingkan dengan DBSCAN.

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan penelitian yang telah dilakukan mengenai pengelompokan kos berdasarkan harga dan fasilitas menggunakan algoritma K-Means dan DBSCAN, dapat disimpulkan beberapa hal penting. Pertama, algoritma K-Means berhasil mengelompokkan kos menjadi tiga cluster utama: Cluster 0 yang mencakup kos dengan harga tertinggi dan fasilitas lengkap, Cluster 1 dengan harga sedang dan fasilitas lengkap, serta Cluster 2 dengan harga terendah dan fasilitas yang lebih terbatas. Sedangkan algoritma DBSCAN, menghasilkan empat cluster dan satu cluster noise, cluster 3 memiliki harga tertinggi dan memiliki semua fasilitas, sementara Cluster 1 memiliki harga terendah dan memiliki fasilitas yang terbatas. Kedua algoritma menghasilkan informasi yang memudahkan pengguna untuk memilih kos yang paling ekonomis dan sesuai dengan kebutuhan fasilitas. Informasi tentang outlier yang dihasilkan oleh DBSCAN juga membantu dalam mengidentifikasi kos yang mungkin menawarkan nilai lebih atau kurang dari yang diharapkan.

5.2 Saran

Berdasarkan kesimpulan yang telah diuraikan, beberapa saran dapat diberikan yaitu, jika tujuan melakukan clustering adalah untuk mengidentifikasi kelompok kos dengan pola harga dan fasilitas yang jelas, maka K-Means adalah metode yang tepat karena memberikan struktur cluster yang terorganisir. Namun, jika bertujuan untuk mendeteksi outlier dan mendapatkan gambaran yang lebih detail mengenai kos yang tidak sesuai dengan pola umum, maka DBSCAN adalah pilihan yang lebih baik. Kemudian, hasil clustering ini dapat diterapkan dalam aplikasi pencarian kos untuk memberikan rekomendasi yang lebih relevan kepada pengguna. Penyedia kos juga dapat memanfaatkan hasil clustering untuk menyesuaikan strategi pemasaran mereka agar lebih kompetitif di pasar. Untuk penelitian lanjutan dapat mempertimbangkan faktor seperti lokasi, keamanan, dan kebersihan untuk membuat clustering yang lebih komprehensif. Dengan menerapkan saran-saran ini, diharapkan pencarian kos menjadi lebih efisien dan efektif, serta memberikan kemudahan bagi pengguna dalam menemukan kos yang paling ekonomis dan sesuai dengan kebutuhan fasilitas mereka.

DAFTAR PUSTAKA

- Artana, P.N.P., Mandyartha, E.P., & Prami S, M.H. (2023). *Penerapan Data Mining pada Algoritma Hierarchical Clustering tentang Pengelolaan Mitra Perjalanan Wisatawan Bali Backpacker*. Jurnal Mahasiswa Teknik Informatika (JATI). Surabaya, Jawa Timur: Universitas Pembangunan Nasional "Veteran" Jawa Timur.
- Ayadi, A., Kusriani, & Pramono, E. (2020). *Perbandingan Tingkat Performa Metode K-Means dan Hierarchical Clustering pada Sistem Rekomendasi Pemilihan Kost*. Magister Teknik Informatika Universitas Amikom Yogyakarta. Yogyakarta, Daerah Istimewa Yogyakarta, Indonesia.
- Daldiri, Z.F., Rafly, M., & Veritawati, I. (2022). *Clustering Daftar Harga Rumah di Jakarta dengan Algoritma K-Means*. Journal of Informatics and Advanced Computing (JIAC). Jakarta, Indonesia: Universitas Pancasila.
- Budiman, S.A.D., Safitri, D., & Ispriyanti, D. (2016). Perbandingan Metode K-Means dan Metode DBSCAN Pada Pengelompokan Rumah Kost Mahasiswa di Kelurahan Tembalang Semarang. JURNAL GAUSSIAN, 5(4), 757-762.
- Chouinard, J. (n.d.). PCA loadings: Interpretation & examples. Jean-Christophe Chouinard. Diakses pada 4 Juni 2024, dari <https://www.jcchouinard.com/pca-loadings/>
- Everitt, B. (1980). Cluster analysis (2nd ed.). London: Social Science Research Council.
- Hawkins, D. M. (1980). Identification of Outliers. Chapman and Hall. (2008).
- Jain, A. K. (2010). Data Clustering: 50 years beyond k-means. Pattern Recognition Letters, 31(8), 651–666.
- Qolbi, A.A. (2016). Penerapan Metode Clustering K-Means Terhadap Dosen Berdasarkan Publikasi Jurnal Nasional dan Internasional. Skripsi. Semarang: Jurusan Ilmu Komputer FMIPA Unnes.
- Rizki, M. F., & Farikhin, F. (2021). Sistem Rekomendasi Kost Menggunakan Metode K-Means Clustering di Kota Malang. Jurnal Sisfokom (Sistem Informasi Dan Komputer), 10(1), 36-41.
- Pramono, A. H., & Fitriana, A. R. (2020). Analisis Kelayakan Kost Mahasiswa dengan Metode K-Means Clustering. Jurnal Ilmiah Teknologi Informasi Terapan (JITTER), 5(2), 84-92.
- Wijaya, A., & Hadi, S. (2019). Pengelompokan Data Kost Mahasiswa Menggunakan Metode K-Means Clustering. Jurnal Mantik Penusa, 3(2), 94-101.


Studi & Informatika. (2022). Pengenalan Clustering dalam Data Mining. Yogyakarta: Penerbit Andi.

LAMPIRAN

Kode Analisis:

 UAS_DM1_023_037_069_079_081.ipynb

Kode Scrapping:

 UAS(Scrapping)_DM1_023_037_069_079_081.ipynb

Dataset:

 data_mamikos_scrapping_1-juni