Clustering Daftar Harga Rumah di Jakarta Dengan Algoritma K-Means

Zyhan Faradilla Daldiri[-1], Muhammad Rafly[-2], Ionia Veritawati[-3]
Program Studi Teknik Informatika
Universitas Pancasila, Jakarta, Indonesia
4519210108@univpancasila.ac.id[1], 4519210096@univpancasila.ac.id[2]
ionia.veritawati@univpancasila.ac.id[3]

Abstrak—Harga rumah di setiap daerah berbeda-beda sesuai dengan daerah dan kategorinya masing-masing. Khususnya harga rumah yang berada di kota misalnya Jakarta. Di Jakarta sendiri memiliki harga rumah yang berbeda sesuai dengan kategorinya. Banyak masyarakat yang tidak mengetahui apakah harga rumah tersebut termasuk murah atau mahal. Maka diberikan solusi untuk mengkategorikan harga rumah yang ada di Jakarta dengan *clustering* menggunakan algoritma *K-Means*. Algoritma *K-Means* dapat membantu untuk mengkategorikan harga rumah di Jakarta dengan 8 atribut yang digunakan terdapat nomor data, nama rumah, harga dari rumah, jumlah luas bangunan, jumlah luas tanah, jumlah kamar tidur, jumlah kamar mandi, dan jumlah kapasitas mobil dalam garasi. Dengan dilakukan penelitian menggunakan algoritma *K-Means* pada k = 5 didapatkan *class* data harga termurah pada class 0 sampai yang termahal pada *class* 4. Dan hasil validitas dari *silhoutte score* yaitu 0,626.

Kata Kunci : Harga Rumah, Algoritma K-Means, Clustering.

I. PENDAHULUAN

Kebutuhan setiap manusia selain dari sandang dan pangan adalah kebutuhan untuk mempunyai tempat tinggal yang nyaman. Rumah adalah sebuah tempat tinggal yang dibutuhkan bagi setiap orang untuk berlindung dan beristirahat [1]. Selain hal tersebut rumah juga bermanfaat sebagai tempat sosialisasi dengan sanak saudara yang ada untuk berkumpul atau bersilaturahmi. Maka dari itu setiap orang pasti membutuhkan rumah sebagai tempat tinggal. Kenyamanan sebuah rumah memiliki beberapa aspek seperti lingkungan, luas bangunan rumah, dan faktor internal seperti keluarga yang membuat suasana rumah menjadi nyaman [2]. Dikarenakan hal tersebut yang membuat setiap orang selektif dalam memilih rumah. Selain hal tersebut juga dilihat juga harga dari rumah tersebut apakah termasuk harga yang relevan atau tidaknya.

Harga rumah pada setiap daerah berbeda-beda biasanya harga rumah di kota lebih mahal daripada di desa. Contohnya di Jakarta sendiri memiliki harga rumah yang cukup tinggi tergantung dari tipe rumah dan luas bangun rumah tersebut. Luas bangunan rumah menjadi salah satu faktor dalam menentukan sebuah harga rumah [3]. Dikarenakan perbedaan tersebut perlu dilakukannya sebuah teknik untuk mengelompokkan daftar harga rumah yang ada di Jakarta untuk membedakan harga pada setiap luas bangunan rumah yang ada mulai dari harga termurah sampai tertinggi. Pengelompokan ini dapat dilakukan dengan salah satu metode dalam data mining yaitu clustering menggunakan algoritma *K-Means.*

Clustering itu sendiri adalah sekumpulan objek dalam membedakan objek-objek yang mirip antara satu sama lainnya untuk dijadikan kedalam sebuah anggota kelompok setiap cluster yang ada. *Clustering* berfungsi untuk mempartisi data tanpa penamaan label pada setiap kelompok yang ada. Salah satu algoritma clustering sendiri adalah algoritma *K-Means* [4].

Algoritma *K-Means* adalah algoritma untuk mengelompokkan data yang banyak menjadi beberapa sekelompok data yang sama. Terdapat beberapa pendekatan untuk menghasilkan cluster-cluster yang ada, salah satunya yaitu membentuk aturan pada setiap anggota dalam kelompok yang sama berdasaran tingkat kesamaan diantara anggota lainnya. Selain hal tersebut lalu membuat sekumpulan data dengan mengukur beberapa faktor dalam pengelompokkan tersebut sebagai parameter clustering dan adanya centroid sebagai pusat cluster tersebut. Metode *K-Means* adalah metode algoritma yang hanya bekerja pada atribut numerik [5].

Dalam hal ini telah dilakukan penelitian seblumnya oleh Hendi Sukma yang berjudul Clustering Data Siswa SMPN 6 Palangkaraya Untuk Menentukan Kelayakan Bantuan Siswa Miskin dan Berprestasi pada tahun 2021 yang membahas mengenai pengelompokan siswa yang layak atau tidak layak untuk mendapatkan bantuan siswa miskin dan berprestasi menggunakan algoritma *K-Means* dengan hasil yang didapatkan adalah pengelompokkan data berdasarkan cluster yang dikategorikannya [6].

Sehingga penelitian tersebut dapat dijadikan sebagai referensi metode untuk mengelompokkan daftar harga rumah yang ada di Jakarta berdasarkan faktor numerik seperti luas bangunan, luas tanah, dan lainnya untuk mengetahui dari harga tertinggi sampai terendah.

II. TINJAUAN PUSTAKA

A. *Clustering*

Clustering merupakan salah satu proses data mining yang bertujuan untuk mengelompokkan beberapa data kedalam cluster. Cluster yaitu kelompok data yang sama obkjeknya dan tidak sama dengan cluster lainnya[7]. Pada proses *clustering* disebut juga sebagai tahap menentukan atau mendeskripsikan nilai numerik pada tingkat kemiripan atau ketidakmiripan data.

B. Algoritma *K-Means*

Algoritma *K-Means* adalah salah satu algoritma pembelajaran data *mining*. Pada dasarnya, kumpulan data dipartisi menjadi sekumpulan k, di mana k mewakili jumlah cluster. Cluster mengklasifikasikan pengamatan dalam banyak kelompok, sehingga pengamatan dalam kelompok yang sama atau mungkin serupa (yaitu, kesamaan kelas yang tinggi atau variasi dalam klaster), sedangkan pengamatan dari kelompok yang berbeda atau tidak sama (yaitu antar kelompok rendah dalam kesamaan kelas). Hal ini ditentukan dari poin rata-rata pada setiap cluster terhadap *centroid* [8]. Pada algoritma *K-Means* perlu menghitung jarak data ke *centroid* terdekat dengan rumus :

$$D(x,y) = \sqrt{(xi - yi)^2 + (xi - yi)^2} \qquad (1)$$

Dimana :
D(x,y) = Jarak data ke *centroid*
x= *Record / Data*
y= *Centroid /* Pusat cluster

Dalam menentukan k yang terbaik digunakan *elbow method*. *Elbow method* adalah metode yang berguna untuk menguji performa tingkat konsistensi jumlah cluster yang tepat. Ditandai dengan grafik yang memiliki lekukan dengan kriteria siku. Nilai pada lekukan kriteria siku inilah menjadi nilai k yang terbaik [9].

Untuk uji validitas *cluster* yang digunakan yaitu dengan validitas *Silhouette* dengan menghitung rata-rata nilai pada setiap himpunan data. Menghitung rata-rata nilai ini yaitu nilai *separation* dikurangi nilai *compactness* dibagi nilai maksimum keduanya. Hasil cluster yang terbaik diambil dari nilai *Silhoutte* yang mendekati nilai 1. Dapat dilihat rentang nilai dan interpretasi pada Tabel 1.1 [9].

Tabel 1. Rentang Nilai *Silhouette* dan Interpretasi

Rentang Nilai	Interpretasi
0.71 – 1.0	Tinggi
0.51 – 0.70	Beralasan
0.26 – 0.50	Rendah
<0.25	Tidak Ditemukan Struktur yang Substansial

C. Rumah

Rumah adalah bangunan yang memiliki fungsi sebagai tempat tinggal dan sarana pembinaan keluarga (Undang-Undang No.4 Tahun 1992). Pengertian rumah secara luas adalah tempat tinggal dengan memenuhi syarat-syarat kelayakan bangunan untuk kehidupan [10].

D. Faktor Pengaruh Harga Rumah

Harga rumah pasti dipengaruhi oleh beberapa faktor yang ada seperti luas tanah, luas bangunan, jumlah kamar mandi, jumlah kamar tidur, dan lingkungannya. Dari segi lingkungan pada akses kesehatan, tingkat kriminalitas, tingkat kebisingan, dan kualitas udara [11].

III. METODOLOGI PENELITIAN

Metodologi yang digunakan untuk penelitian ini adalah seperti pada Gambar 3.1 berikut.

Gambar 3.1 Metodologi Penelitian

Metodologi penelitian yang ada pada Gambar 3.1 dijelaskan sebagai berikut:

a. Identifikasi masalah : Mengidentifikasi masalah apa yang sedang terjadi dan mencari data yang dapat digunakan.

 E-ISSN : 2722-0346

b. Data harga rumah di Jakarta : Data ini adalah data yang dapat digunakan pada penelitian untuk dikelola yang diambil dari situs kaggle https://www.kaggle.com/datasets/wisnuanggara/daftar-harga-rumah yang terdapat 8 atribut yaitu NO : nomor data; NAMA RUMAH : title rumah; HARGA : harga dari rumah; LB : jumlah luas bangunan; LT : jumlah luas tanah; KT : jumlah kamar tidur; KM : jumlah kamar mandi; GRS : jumlah kapasitas mobil dalam garasi. Data tersebut berjumlah 1100 data.
c. Pembersihan data : Melakukan data *cleaning* untuk melihat data yang kosong atau bernilai *null* dan *duplicate* data.
d. Pengelolahan data dengan metode *clustering* menggunakan algoritma *K-Means* : Mengelola data harga rumah di Jakarta menggunakan algoritma *K-Means* dengan mengambil data uji, menentukan jumlah cluster (k), menghitung jarak data ke *centriod*, dan mengelompokkan data berdasarkan jarak *minimum* ke *centroid*.
e. Analisa hasil penelitian : Menganalisis hasil yang didapat dengan melihat *score* dari penelitian yang dilakukan.

IV. HASIL DAN DISKUSI

Penelitian ini menggunakan *tools Google Collaboratory* untuk mengimplementasikan algoritma *K-Means*. Pada permasalahan harga rumah yang berada di Jakarta diambil *dataset* dari *kaggle* yang memperlihatkan data awal saja yaitu seperti pada Gambar 4.1 berikut.

	NO	NAMA RUMAH	HARGA	LB	LT	KT	KM	GRS
0	1	Rumah Murah Hook Tebet Timur, Tebet, Jakarta S...	3800000000	220	220	3	3	0
1	2	Rumah Modern di Tebet dekat Stasiun, Tebet, Ja...	4600000000	180	137	4	3	2
2	3	Rumah Mewah 2 Lantai Hanya 3 Menit Ke Tebet, T...	3000000000	267	250	4	4	4
3	4	Rumah Baru Tebet, Tebet, Jakarta Selatan	430000000	40	25	2	2	0
4	5	Rumah Bagus Tebet komp Gudang Peluru lt 350m, ...	9000000000	400	355	6	5	3

Gambar 4.1 *Dataset* Daftar Harga Rumah

Selanjutnya memeriksa dataset apakah terdapat *null* atau tidak seperti pada Gambar 4.2.

```
[3]  data.isna().sum()

NO            0
NAMA RUMAH    0
HARGA         0
LB            0
LT            0
KT            0
KM            0
GRS           0
dtype: int64
```

Gambar 4.2 Data *Null*

Gambar 4.2 menunjukkan bahwa *dataset* tersebut tidak ada yang memiliki nilai *null*. Selanjutnya mengecek data yang *duplicate* seperti pada Gambar 4.3.

```
[4]  data.dropna(inplace=True)

[5]  data.duplicated().sum()

0
```

Gambar 4.3 *Duplicate* Data

Gambar 4.3 menunjukkan bahwa dataset tersebut tidak memiliki *duplicate* data. Selanjutnya mengecek info data yang digunakan seperti pada Gambar 4.4.

```
[6]  data.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1010 entries, 0 to 1009
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   NO          1010 non-null   int64
 1   NAMA RUMAH  1010 non-null   object
 2   HARGA       1010 non-null   int64
 3   LB          1010 non-null   int64
 4   LT          1010 non-null   int64
 5   KT          1010 non-null   int64
 6   KM          1010 non-null   int64
 7   GRS         1010 non-null   int64
dtypes: int64(7), object(1)
memory usage: 71.0+ KB

[7]  data.shape

(1010, 8)
```

Gambar 4.4 Data Info

Gambar 4.4 memperlihatkan info dari data yag digunakan terdapat 8 kolom yang terdiri dari no, nama rumah, harga, LB, LT, KT, KM, dan GRS juga memiliki 1010 data.

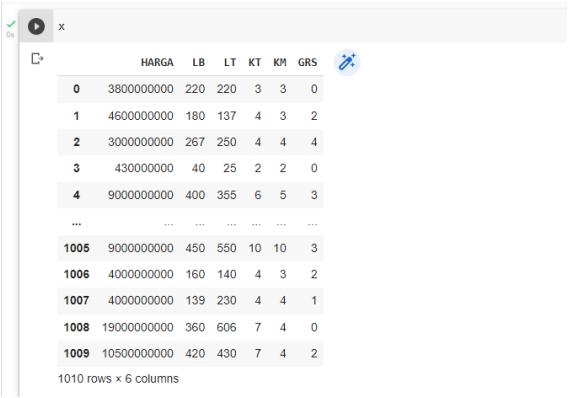
Setelah melakukan pembersihan data dan info data, selanjutnya dilakukan pegelolahan data dengan menggunaan algoritma *K-Means*. Pertama diambil data yang memiliki kategori numerik dan data numerik yang digunakan dibuat sebagai *array* pada sebuah variabel seperti pada Gambar 4.5.

```
[9]  numerik = ['HARGA', 'LB', 'LT', 'KT', 'KM', 'GRS']

[10] x = pd.concat([data[numerik]],axis=1)
```

Gambar 4.5 Mengambil Data Numerik

Dan datanya terlihat seperti pada Gambar 4.6 berikut.

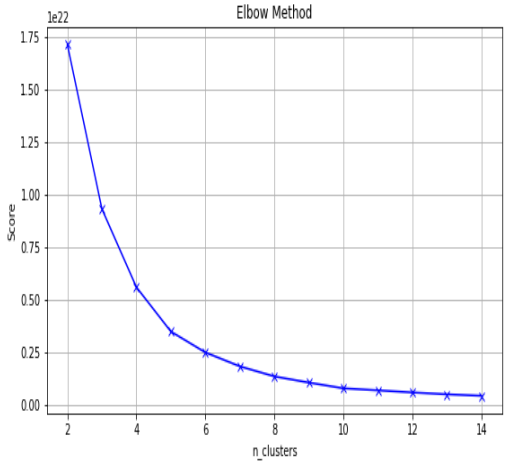


	HARGA	LB	LT	KT	KM	GRS
0	3800000000	220	220	3	3	0
1	4600000000	180	137	4	3	2
2	3000000000	267	250	4	4	4
3	430000000	40	25	2	2	0
4	9000000000	400	355	6	5	3
...	...	...	...	...	...	...
1005	9000000000	450	550	10	10	3
1006	4000000000	160	140	4	3	2
1007	4000000000	139	230	4	4	1
1008	19000000000	360	606	7	4	0
1009	10500000000	420	430	7	4	2

1010 rows × 6 columns

Gambar 4.6 Data Numerik

Setelah memilih data numerik untuk dilakukan uji coba, lalu menentukan nilai k yang akan digunakan untuk uji coba dengan elbow *method.*



Gambar 4.7 Grafik *Elbow Method*

Gambar 4.7 menunjukkan grafik dari *elbow method* dalam menentukan k yang digunakan diambil nilai k pada lekukan yang terlihat siku pada grafik tersebut. Lekukan yang terlihat siku terdapat pada nilai k = 5. Maka digunakan k = 5 sebagai nilai k terbaik untuk melakukan pengolahan data dengan algoritma *K-Means*.

Selanjutnya dilakukan perhitungan untuk melihat nilai jarak objek data ke *centroid* dengan *array* seperi pada Gambar 4.8.

```
array([[7.42440325e+09, 3.05986395e+02, 2.54187075e+02, 5.13605442e+00,
        3.97619048e+00, 2.06462585e+00],
       [2.47955882e+10, 6.12367647e+02, 6.22602941e+02, 5.85294118e+00,
        4.88235294e+00, 4.26470588e+00],
       [4.28461538e+10, 6.78846154e+02, 7.33230769e+02, 5.46153846e+00,
        5.15384615e+00, 3.38461538e+00],
       [1.42976759e+10, 4.36490741e+02, 4.01861111e+02, 5.33333333e+00,
        4.20370370e+00, 2.61111111e+00],
       [3.29270082e+09, 1.74075901e+02, 1.32459203e+02, 4.09867173e+00,
        3.07779886e+00, 1.36053131e+00]])
```

Gambar 4.8 Jarak Objek Data Ke *Centroid*

Lalu dilakukan perhitungan algoritma *K-Means* dalam mengelompokkan data berdasarkan *class* yang ditentukan dengan k = 5, maka menghasilkan seperti pada Gambar 4.9.

```
[23] kmeans = KMeans(n_clusters=5).fit(x)
     hasil_kmeans = kmeans.labels_
     hasil_kmeans

     array([0, 0, 0, ..., 0, 2, 4], dtype=int32)
```

Gambar 4.9 *K-Means* Label

Untuk melihat hasil *class K-Means* pada data maka dapat dilihat pada Gambar 4.10 berikut.

```
[25] data['Class'] = hasil_kmeans

[26] data
```

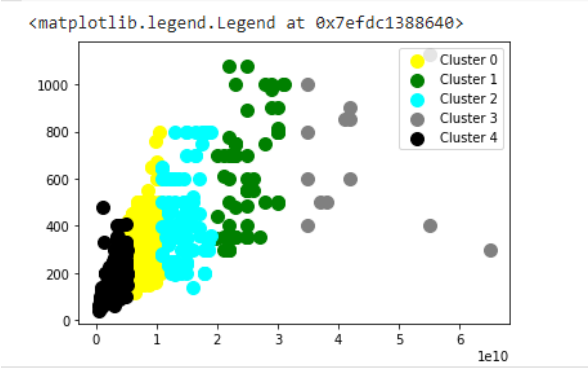
	NO	NAMA RUMAH	HARGA	LB	LT	KT	KM	GRS	Class
0	1	Rumah Murah Hook Tebet Timur, Tebet, Jakarta S...	3800000000	220	220	3	3	0	0
1	2	Rumah Modern di Tebet dekat Stasiun, Tebet, Ja...	4600000000	180	137	4	3	2	0
2	3	Rumah Mewah 2 Lantai Hanya 3 Menit Ke Tebet, T...	3000000000	267	250	4	4	4	0
3	4	Rumah Baru Tebet, Tebet, Jakarta Selatan	430000000	40	25	2	2	0	0
4	5	Rumah Bagus Tebet komp Gudang Peluru lt 350m, ...	9000000000	400	355	6	5	3	4
...	...	...	...	...	...	...	...	...	...
1005	1006	Rumah Strategis Akses Jalan 2mobil Di Menteng ...	9000000000	450	550	10	10	3	4
1006	1007	Tebet Rumah Siap Huni Jln 2 Mbl Nyaman	4000000000	160	140	4	3	2	0
1007	1008	Di Kebun Baru Rumah Terawat, Area Strategis	4000000000	139	230	4	4	1	0
1008	1009	Dijual Cepat Rumah Komp Depkeu Dr Soepomo Tebe...	19000000000	360	606	7	4	0	2
1009	1010	Dijual Rumah Kokoh Di Gudang Peluru	10500000000	420	430	7	4	2	4

1010 rows × 9 columns

Gambar 4.10 Data Hasil *K-Means*

Untuk menunjukkan pengelompokkan data berdasarkan *class* yang ditentukan yang diaplikasikan ke dalam bentuk *scatter plot* seperti pada Gambar 4.11.

 E-ISSN : 2722-0346



Gambar 4.11 Pengelompokkan Data

Untuk hasil setiap *class* yang ada terdapat berapa data maka dapat dilihat pada Gambar 4.12.

```
[27] data['Class'].value_counts()

0    527
4    294
2    108
3     68
1     13
Name: Class, dtype: int64
```

Gambar 4.12 Jumlah Data Setiap *Class*

Dari Gambar 4.12 dapat dilihat bahwa terdapat 5 *class* dimulai dari *class* 0 yang terdapat 527 data, *class* 1 yang terdapat 13 data, *class* 2 yang terdapat 108 data, *class* 3 yang terdapat 68 data, dan *class* 4 yang terdapat 294. Untuk melakukan pengujian validitas digunakan *silhouette score.*

```
[24] sscore_kmeans = silhouette_score(x,hasil_kmeans)
     sscore_kmeans

0.6269069705004773
```

Gambar 4.13 *Silhouette Score*

Gambar 4.13 menunjukkan hasil *silhouette score* yaitu 0,626 yang berarti hasil tersebut memiliki interpretasi hasil yang beralasan.

V. KESIMPULAN

Dari penelitian yang telah dilakukan maka diambil kesimpulan pada daftar harga rumah di Jakarta dengan menggunakan algoritma *K-Means* pada k = 5 didapatkan *class* data harga termurah pada class 0 sampai yang termahal pada *class* 4.

DAFTAR PUSTAKA

[1] Purba, Fernando. 2015. Faktor-Faktor Yang Mempengaruhi Permintaan Rumah Pada Perumahan Citra Garden Padang Bulan Medan. Fakultas Ekonomi. Universitas Negeri Medan. Jl. Willem Iskandar / Pasar V, Medan, Sumatera Utara,20221.

[2] Muttaqin, Najmul. 2017. Implementasi Metode Fuzzy C-Means (FCM) Clustering Dalam Sistem Pendukung Keputusan Untuk Menentukan Pembelian Rumah. Teknik Informatika. Universitas Islam Negeri Gunung Djati Bandung. Jalan A.H. Nasution No. 105, Cipadung, Cibiru 40614.

[3] Noveandri, Muhammad Faris. 2016. Analisis Pengaruh Harga, Fasilitas, Lokasi, dan Lingkungan Terhadap Keputusan Pembelian Rumah Di Perumahan Cluster (Studi pada Konsumen Penghuni Perumahan Cluster di Kota Cilacap). Fakultas Ekonomi dan Bisnis. Universitas Muhammadiyah Purwokerto. Jl. KH. Ahmad Dahlan, Dusun III, Dukuhwaluh, Kec. Kembaran, Kabupaten Banyumas, Jawa Tengah 53182.

[4] Mauliadi, Rafki. 2022. Data Mining Menggunakan Algoritma K-Means Clustering dalam Analisis Tingkat Potongan Harga Terhadap Harga Jual Sepeda Motor Honda. *Jurnal Informatika Ekonomi Bisnis*, 4(4), 124-129. ISSN : 2714-8491.

[5] Dhuhita, Windha Mega Pradnya. 2015. Clustering Menggunakan Metode K-Means Untuk Menentukan Status Gizi Balita. *Jurnal Informatika*, 15(2), 160-174.

[6] Sukma, Hendy. 2021. Clustering Data Siswa SMPN 6 Palangkaraya Untuk Menentukan Kelayaan Bantuan Siswa Miskin dan Berprestasi. Fakultas Tenik Informatika.Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Palangkaraya. Jl. G. Obos No.114, Menteng, Kec. Jekan Raya, Kota Palangka Raya, Kalimantan Tengah 74874.

[7] Rahmah, Evi., Haerani, Elin., dll. 2022. Penerapan Algoritma K-Medoids Clustering Untuk Menentukan Strategi Promosi Pada Data Mahasiswa (Studi Kasus : Stikes Perintis Padang). *Jurnal Nasional Komputasi dan Teknologi Informasi*, 5(3), 556-564. ISSN : 2621-3052.

[8] Kilinc, Betul Kan., Tug, Ilkay. 2019. The Examination of Real Estate Prices in Istanbul by Using Hybrid Hierarchical K-Means Clustering. *y-BIS*, 208-212.

[9] Aditya, Agil,. dkk. 2020. Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama Di Indonesia Tahun 2018/2019. *Jurnal Media Informatika Budidarma*, 4(1), 51-48. DOI : 10-30865.

[10] Ningrum, Tyas Puspita. 2018. Kajian Perubahan Fungsi Rumah Tinggal Menjadi Rumah Kos Di Sekitar Kampus Universitas Muhammadiyah Purwokerto. Pendidikan Geografi. Universitas Muhammadiyah Purwokerto. Jl. KH. Ahmad Dahlan, Dusun III, Dukuhwaluh, Kec. Kembaran, Kabupaten Banyumas, Jawa Tengah 53182.

[11] Kurniawan, Anindista Tursilo. 2017,. dkk. 2017. Analisis Faktor-Faktor yang Mempengaruhi Harga Jual Rumah Di Kabupaten Sukoharjo dan Karanganyar. Teknik Industri. Universitas Sebelas Maret. Jl. Ir. Sutami 36A Surakarta 57126.