Analisis Data *House Price Prediction* di Kota Seattle, Amerika Serikat dengan Metode *K-Means Clustering* dan *K-Medoids Clustering*

Dina Aprilia Aktarani[1, a)], Diva Kemala[1, b)], Fitriana Murtafiah[2, c)], Julius Satya Ratnandi [1, d)], Maristy Widya Pangestika[1, e)], Tiara Yosianti Solekhah [1, f)]

Afiliasi Penulis

[1] *Mahasiswa S1 Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Gadjah Mada, Sekip Utara BLS 21 Yogyakarta, Indonesia*

[2] *Mahasiswa S1 Program Studi Ilmu Aktuaria, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Gadjah Mada, Sekip Utara BLS 21 Yogyakarta, Indonesia*

Email Penulis

[a)]Penulis yang bersesuaian: *dinaaktarani12@mail.ugm.ac.id*

[b)] *divakemala@mail.ugm.ac.id*

[c)] *fitriana.murtafiah@mail.ugm.ac.id*

[d)] *julius.satya.r@mail.ugm.ac.id*

[e)] *maristy.widya@mail.ugm.ac.id*

[f)] *tiara.y@mail.ugm.ac.id*

Abstract. Rumah adalah bangunan yang berfungsi sebagai tempat tinggal atau hunian dan sarana pembinaan keluarga. Terdapat berbagai jenis dan spesifikasi rumah yang dipasarkan, terlebih pada kota besar yang jumlah rumahnya terbilang sangat banyak. Untuk mengetahui rumah mana yang memiliki kriteria sesuai dengan keinginan calon pembeli, diperlukan adanya pengelompokan rumah berdasarkan faktor-faktor tertentu. Oleh karena itu, penelitian ini bertujuan untuk mengelompokkan rumah pada salah satu kota besar di Amerika Serikat yaitu kota Seattle menggunakan data *House Price Prediction* dari situs *Kaggle*. Pengelompokan rumah di Seattle didasarkan pada harga, jumlah kamar tidur, luas bangunan rumah, luas lahan, dan jumlah lantai (tingkat) dalam rumah. Dengan menggunakan metode *k-means clustering* dan *k-medoids clustering* diperoleh masing-masing dua kluster rumah untuk pengelompokan rumah di kota Seattle, Amerika Serikat.

Kata Kunci : kluster, rumah, k-means, k-medoids

LATAR BELAKANG

Keberlangsungan hidup manusia disertai dengan berbagai kebutuhan. Kebutuhan muncul karena naluri dari manusia untuk bertahan hidup. Kebutuhan dapat dikategorikan berdasarkan beberapa faktor, salah satunya adalah tingkat kepentingan. Berdasarkan kepentingannya, kebutuhan dapat dibedakan menjadi 3 jenis, yakni kebutuhan primer, sekunder, dan tersier. Kebutuhan primer adalah kebutuhan yang paling mendasar bagi manusia untuk dapat bertahan hidup. Salah satu contoh kebutuhan primer adalah papan.

Papan adalah tempat tinggal atau tempat beristirahat yang aman bagi manusia, seperti rumah. Berdasarkan Pasal 1 Bab 1 Undang-Undang Republik Indonesia Nomor 4 Tahun 1992 tentang Perumahan dan Permukiman, definisi rumah adalah bangunan yang berfungsi sebagai tempat tinggal atau hunian dan sarana pembinaan keluarga. Berdasarkan definisi tersebut, fungsi rumah tidak hanya sebagai tempat tinggal manusia, namun juga menjadi tempat bagi manusia untuk tumbuh, berkembang dan hidup sebagai makhluk sosial di lingkungan keluarga.

Pada kota-kota besar di dunia, jumlah rumah bisa mencapai puluhan bahkan ratusan ribu unit. Salah satunya di kota Seattle, Amerika Serikat. Kota yang berada di negara bagian Washington ini adalah salah satu kota besar dari segi populasi di Amerika Serikat. Berdasarkan data *American Community Survey* tahun 2014, setidaknya terdapat 324.490 unit rumah yang ada di wilayah kota Seattle, dengan 304.564 unit yang dihuni. Artinya, terdapat 19.926 unit rumah yang masih belum dihuni dan dipasarkan sepanjang tahun 2014.

Berbagai jenis dan spesifikasi rumah dipasarkan di kota Seattle. Umumnya, calon pembeli rumah akan mempertimbangkan faktor kualitatif maupun kuantitatif. Contoh faktor kuantitatif adalah harga, luas bangunan, jumlah kamar, dan lain sebagainya. Faktor ini disesuaikan dengan kebutuhan, keinginan, dan kemampuan dari calon pembeli. Ketiga kriteria tersebut bisa dikatakan saling bertolak belakang, di mana orang tentu mencari rumah yang sesuai dengan kebutuhan dan keinginannya namun sebisa mungkin mengeluarkan uang yang tidak banyak. Agar dapat mengetahui rumah mana saja yang memiliki kriteria tertentu yang sesuai dengan keinginan calon pembeli, maka dapat dilakukan pengelompokkan rumah-rumah berdasarkan faktor-faktor kuantitatif.

Analisis *cluster* adalah metode multivariat yang bertujuan untuk mengelompokkan sampel objek berdasarkan beberapa variabel terukur ke dalam sejumlah *cluster* (kelompok) yang berbeda, sehingga objek yang serupa ditempatkan pada *cluster* yang sama. Analisis *cluster* terdiri dua metode, yaitu metode hierarki dan metode non-hierarki. Metode hierarki digunakan apabila belum diketahui banyaknya *cluster* yang akan dibentuk. Sementara metode non-hierarki digunakan apabila sudah diketahui banyaknya *cluster* yang ingin dibentuk (Sihombing dkk, 2020). Salah satu jenis metode non-hierarki adalah metode partisi atau *partitioning methods*. Metode partisi merupakan metode yang melakukan optimasi pada penempatan objek, di mana objek dalam satu *cluster* akan memiliki karakteristik yang serupa dan berbeda dengan objek pada *cluster* lainnya. Pada metode partisi terdapat beberapa metode, antara lain metode K-Means dan metode K-Medoids. Metode K-Means adalah metode yang melakukan penentuan jumlah *cluster* terlebih dahulu, sedangkan metode K-Medoids menjadikan beberapa objek sebagai *medoids* atau pusat dari *cluster*, sehingga jumlah *cluster* akan menyesuaikan dengan banyaknya *medoids* (Musfiani, 2019).

Berdasarkan uraian di atas, kami akan melakukan pengelompokkan rumah yang dijual di kota Seattle dengan menggunakan analisis *cluster*. Pengelompokkan dilakukan dengan menggunakan data yang diperoleh dari situs *kaggle*, yakni data harga rumah di negara bagian Washington pada tanggal 2 Mei 2014 hingga 10 Juli 2014. Metode yang akan kami gunakan adalah metode partisi dengan metode K-Means dan K-Medoids. Pengelompokkan rumah didasarkan pada harga, luas bangunan, luas lahan, jumlah kamar tidur, dan jumlah lantai (tingkat) bangunan.

TUJUAN DAN MANFAAT

- Mengimplementasikan analisis *cluster* dalam persoalan dunia nyata, yakni pengelompokkan rumah di kota Seattle
- Mengetahui pengelompokan rumah di kota Seattle berdasarkan variabel harga, luas bangunan, luas lahan, jumlah kamar tidur, dan jumlah lantai (tingkat) bangunan, sehingga memudahkan calon pembeli rumah di kota Seattle dalam menentukan karakteristik rumah seperti apa yang diinginkan

DASAR TEORI

Analisis *cluster* merupakan proses mempartisi sekumpulan objek data (atau pengamatan) ke dalam subset. Setiap subset adalah sebuah *cluster* sehingga objek-objek dalam sebuah *cluster* mirip satu sama lain, namun berbeda dengan objek di *cluster* lain (Han, 2012).

Tujuan utama dari metode *clustering* adalah pengelompokan sejumlah data/objek ke dalam *cluster* (group) sehingga dalam setiap *cluster* akan berisi data yang semirip mungkin (Santosa, 2007). *Clustering* melakukan pengelompokan data yang didasarkan pada kesamaan antar objek dalam setiap *cluster* sehingga akan menghasilkan kelompok-kelompok yang memiliki homogenitas tinggi antar objek di dalam suatu *cluster* dan heterogenitas yang tinggi antar *cluster*-nya.

Prosedur dalam algoritma *cluster* yang umum digunakan yaitu *partitioning methods* dan *hierarchical methods*. Perbedaan mendasar dari kedua jenis *clustering* tersebut adalah sebagai berikut.

a. *Partitioning methods*

Metode partisi membagi observasi dari suatu himpunan menjadi beberapa kelompok *cluster* eksklusif dimana setiap observasi harus bergabung dengan tepat satu kelompok saja. Dalam metode partisi, *cluster* dibentuk dengan mengoptimalkan kriteria partisi objektif, seperti fungsi ketidaksamaan berdasarkan jarak, sehingga karakteristik observasi dalam *cluster* serupa satu sama lain dan berbeda dengan observasi di *cluster* lain dalam hal atribut kumpulan data. Metode partisi yang umum digunakan adalah K-Means dan K-Medoids.

Algoritma K-Means didasarkan pada penentuan *centroid* dari sebuah *cluster* sebagai nilai rata-rata dari titik-titik di dalam *cluster*. Langkah-langkah algoritma K-Means adalah sebagai berikut:

1. Tentukan nilai k sebagai jumlah *cluster* yang ingin dibentuk.
2. Tentukan k *centroid* (titik pusat *cluster*) awal secara random.
3. Hitung jarak setiap data ke masing-masing *centroid*.
4. Kelompokkan setiap data berdasarkan jarak terdekat antara data dengan *centroid*-nya.
5. Tentukan posisi *centroid* baru dengan cara menghitung nilai rata-rata dari data-data yang ada pada *centroid* yang sama.
6. Kembali ke langkah 3 jika posisi *centroid* baru dengan *centroid* lama tidak sama.

Penentuan jumlah *cluster* k dalam K-means *clustering* dapat dilakukan dengan memberikan kisaran perkiraan nilai k kemudian menggunakan teknik analisis untuk menentukan k terbaik dengan membandingkan hasil pengelompokan yang diperoleh untuk nilai k yang berbeda. K-means *clustering* sensitif terhadap *noise* dan pencilan karena sejumlah kecil data semacam itu dapat memengaruhi nilai rata-rata secara substansial.

Berbeda dengan K-Means *clustering* yang menggunakan rata-rata sebagai *centroid*, K-Medoids *clustering* menggunakan medoid sebagai *centroid* sehingga tidak sensitif terhadap pencilan. Salah satu algoritma K-Medoids yang sering digunakan adalah Partitioning around Medoids (PAM). Meskipun PAM lebih *robust* terhadap pencilan n dibandingkan K-Means *clustering* dan bekerja dengan sangat memuaskan untuk dataset berukuran kecil, namun algoritma PAM tidak efisien dalam menangani kumpulan data menengah dan besar. Algoritma PAM adalah sebagai berikut:

1. Menentukan terlebih dahulu banyaknya cluster yang akan dibentuk (k).
2. Menentukan k titik awal medoid secara acak.
3. Melakukan perhitungan jarak masing-masing observasi ke masing-masing medoid.

4. Memasukkan observasi ke masing-masing cluster dengan mengambil jarak yang paling dekat dari masing-masing observasi, kemudian hitung total jaraknya.
5. Memilih secara acak observasi masing-masing cluster sebagai kandidat medoid baru.
6. Menghitung jarak setiap observasi non-medoid dengan medoid baru dan menempatkan tiap observasi non-medoid tersebut ke kandidat medoid terdekat, kemudian hitung total jaraknya.
7. Menghitung total simpangan (S) dengan menghitung nilai total jarak baru – total jarak lama. Apabila total simpangan bernilai negatif, maka kita menggunakan medoids baru sebagai medoids.
8. Dilakukan perulangan terhadap langkah (5) sampai langkah (7) hingga tidak terjadi perubahan medoid sehingga didapatkan cluster beserta anggota cluster masing-masing.

b. *Hierarchical methods*

Metode berhierarki digunakan untuk mengelompokkan pengamatan secara terstruktur berdasarkan kemiripan sifatnya dan kelompok yang diinginkan belum diketahui banyaknya. Metode ini dilakukan dengan mengelompokkan objek dalam diagram pohon. Ada dua cara untuk mendapatkan kelompok dengan metode pengelompokan hierarki yaitu dengan cara penggabungan (*agglomerative*) dan pemisahan kelompok (*divisive*) (Mattjik & Sumertajaya, 2011).

Pada awal teknik penggabungan, setiap objek dianggap sebagai suatu *cluster* tersendiri. Kemudian, objek-objek tersebut dikelompokkan mulai dari yang ukuran ketakmiripan yang terkecil atau jaraknya terdekat. Bila suatu cluster sudah terisi dengan dua atau beberapa objek, ukuran ketakmiripan yang digunakan adalah antar cluster. Misalkan ukuran ketakmiripan antara cluster ke-i dan cluster ke-j adalah $d_{i,j}$ serta ukuran ketakmiripan antara cluster ke-j dengan cluster (i,j) adalah $d_{k(i,j),}$ terdapat beberapa ukuran ketakmiripan antar cluster antara lain:

1. Pautan Tunggal

   Metode ini didasarkan pada jarak minimum. Ukuran ketakmiripan yang digunakan adalah:

   $$d_{(k(i,j)} = minimum(d_{ki,}d_{kj})$$

2. Pautan Lengkap

   Metode ini didasarkan pada jarak maksimum. Ukuran ketakmiripan yang digunakan adalah:

   $$d_{(k(i,j)} = maksimum(d_{ki,}d_{kj})$$

3. Rataan

   Metode ini didasarkan pada jarak rata-rata antara pengamatan. Ukuran ketakmiripan yang digunakan adalah:

   $$d_{(k(i,j)} = \frac{n_i}{(n_i + n_j)}d_{ki} + \frac{n_j}{(n_i + n_j)d_{kj}}$$

4. Sentroid (Centroid Method)

   Jarak antara dua cluster adalah jarak antara pusat cluster. Ukuran ketakmiripan yang digunakan adalah:

   $$d_{(k(i,j)} = \frac{n_i}{(n_i + n_j)}d_{ki} + \frac{n_j}{(n_i + n_j)d_{kj}} + \frac{(n_i n_j)}{(n_i + n_j)^2}d_{ij}$$

5. Ward's Error Sum of Square Method

   Metode ini tidak menghitung jarak antar cluster-nya, tetapi membentuk cluster dengan memaksimalkan ukuran kehomogenannya. Cluster yang memiliki kombinasi jumlahan kuadrat error terkecil adalah yang terbaik. Jumlahan kuadrat yang diminimalkan sering disebut Error Sums of Square (ESS).

Pada teknik pembagian, diawali dengan suatu cluster yang berisikan seluruh objek yang ada. Kemudian, cluster ini dibagi menjadi 2, masing-masingnya dibagi menjadi 2 lagi, dan seterusnya. Objek dengan ketakmiripan terbesar atau jarak terjauh dipisahkan sehingga membentuk cluster yang lebih kecil (splinter). Pemisahan ini dilakukan hingga mencapai jumlah cluster yang diinginkan. Untuk metode pembagiannya, didasarkan pada perhitungan rata-rata jarak masing-masing objek dengan objek pada kelompok splinter dan rata-rata jarak objek tersebut dengan objek lain pada kelompoknya. Hal ini sering disebut Splinter Average Distance Method.

METODE

Pada data *House Price Prediction* akan dibentuk *cluster* dari dataset yang ada untuk memudahkan calon pembeli dalam menentukan kriteria rumah pilihannya. Metode yang digunakan dalam studi ini adalah *clustering* dengan metode partisi yang berupa K-Means dan K-Medoids. Pada analisis cluster ini akan digunakan variabel *price, bedrooms, sqft_lot, sqft_living,* dan *floors* pada kota Seattle. Clustering ini dilakukan untuk mengelompokkan rumah pada kota Seattle. Adapun tahapan teknik analisis yang akan dilakukan yaitu:

1. Melakukan pendeteksian missing value
2. Melakukan uji asumsi
3. Menentukan jumlah cluster (K-Means)
4. Menentukan jumlah cluster (K-Medoids)
5. Memilih metode clustering yang lebih baik digunakan

DATA YANG DIGUNAKAN

Data yang digunakan pada penelitian ini adalah data sekunder yakni data yang diperoleh secara tidak langsung atau diperoleh dari sumber yang sudah ada. Digunakan data *House Price Prediction* dari situs *Kaggle* yang berisikan kumpulan data rumah di negara Washington pada tanggal 2 Mei 2014 hingga 10 Juli 2014, terdapat sebanyak 4600 observasi dan 18 variabel sebagai berikut

- *Price* : Harga dari rumah
- *Bedrooms* : Jumlah kamar tidur yang ada dalam rumah
- *Bathrooms* : Jumlah kamar mandi yang ada dalam rumah
- *Sqft_living* : Luas bangunan rumah dalam satuan square feet
- *Sqft_lot* : Luas lahan dalam satuan square feet
- *Floors* : Jumlah lantai (tingkat) dalam rumah
- *Waterfront* : Apakah rumah menghadap ke tepi laut atau tidak (tidak = 0, iya = 1)
- *View* : Seberapa bagus pemandangan tepi laut dari rumah (0-4)
- *Condition* : Seberapa baik kondisi rumah secara keseluruhan (1-5)
- *Sqft_above* : Luas rumah yang berada di atas tanah dalam satuan square feet
- *Sqft_basement* : Luas rumah yang berada di bawah tanah dalam satuan square feet

- *Yr_built* : Tahun ketika rumah dibangun
- *Yr_renovated* : Tahun ketika rumah terakhir direnovasi
- *Street* : Nama jalan lokasi rumah
- *City* : Nama kota lokasi rumah
- *Statezip* : Kode pos
- *Country* : Nama negara lokasi rumah

Dalam studi kasus ini hanya akan dilakukan analisis kluster perumahan pada kota Seattle yang memiliki data penjualan rumah paling banyak yaitu sebanyak 1573 observasi. Pengelompokan akan dilakukan berdasarkan variabel *street* dan hanya akan menggunakan 5 variabel karakteristik yaitu *price, bedrooms, sqft_living, sqft_lot,* dan *floors*.

ANALISIS

Melakukan Import Data

Penelitian dilakukan dengan menggunakan *google colab*. Data yang digunakan adalah data House Price Prediction dari platform Kaggle yang terdiri dari kumpulan data penjualan properti berupa rumah di US sebanyak 4600 observasi dan memiliki 18 variabel. Pertama-tama melakukan import data terlebih dahulu ke *google colab*.

```
1 from google.colab import drive
```

```
1 drive.mount('/content/drive',force_remount=True)
```

```
Mounted at /content/drive
```

```
1 data = pd.read_csv("/content/drive/MyDrive/Kelompok 19 Prak Mining/Data/data.csv")
```

```
1 data
```

	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basement	yr_built	yr_renovated	
0	2014-05-02 00:00:00	3.130000e+05	3.0	1.50	1340	7912	1.5	0	0	3	1340	0	1955	2005	De
1	2014-05-02 00:00:00	2.384000e+06	5.0	2.50	3650	9050	2.0	0	4	5	3370	280	1921	0	Bl
2	2014-05-02 00:00:00	3.420000e+05	3.0	2.00	1930	11947	1.0	0	0	4	1930	0	1966	0	14
3	2014-05-02 00:00:00	4.200000e+05	3.0	2.25	2000	8030	1.0	0	0	4	1000	1000	1963	0	85
4	2014-05-02	5.500000e+05	4.0	2.50	1940	10500	1.0	0	0	4	1140	800	1976	1992	17
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
4595	2014-07-09 00:00:00	3.081667e+05	3.0	1.75	1510	6360	1.0	0	0	4	1510	0	1954	1979	1
4596	2014-07-09 00:00:00	5.343333e+05	3.0	2.50	1460	7573	2.0	0	0	3	1460	0	1983	2009	14
4597	2014-07-09 00:00:00	4.169042e+05	3.0	2.50	3010	7014	2.0	0	0	3	3010	0	2009	0	In
4598	2014-07-10 00:00:00	2.034000e+05	4.0	2.00	2090	6630	1.0	0	0	3	1070	1020	1974	0	(
4599	2014-07-10 00:00:00	2.206000e+05	3.0	2.50	1490	8102	2.0	0	0	4	1490	0	1990	0	18 2

4600 rows × 18 columns

Data telah berhasil diimport ke *google colab*.

Melakukan Pengecekan terhadap Missing Value

```
data.isna().sum()
```

```
date             0
price            0
bedrooms         0
bathrooms        0
sqft_living      0
sqft_lot         0
floors           0
waterfront       0
view             0
condition        0
sqft_above       0
sqft_basement    0
yr_built         0
yr_renovated     0
street           0
city             0
statezip         0
country          0
dtype: int64
```

Berdasarkan output, diketahui bahwa pada seluruh variabel tidak terdapat missing value ditandai dengan angka 0 pada setiap variabel.

Melakukan clustering data

Clustering dilakukan berdasarkan variabel street dan hanya akan menggunakan 5 variabel karakteristik yaitu price, bedrooms, sqft_living, sqft_lot, dan floors.

```
house = pd.DataFrame(data,columns=['street','price','sqft_lot','sqft_living','floors','bedrooms','city'])
```

```
house
```

	street	price	sqft_lot	sqft_living	bedrooms	floors	city
0	18810 Densmore Ave N	3.130000e+05	7912	1340	3.0	1.5	Shoreline
1	709 W Blaine St	2.384000e+06	9050	3650	5.0	2.0	Seattle
2	26206-26214 143rd Ave SE	3.420000e+05	11947	1930	3.0	1.0	Kent
3	857 170th Pl NE	4.200000e+05	8030	2000	3.0	1.0	Bellevue
4	9105 170th Ave NE	5.500000e+05	10500	1940	4.0	1.0	Redmond
...	...	...	...	...	...	...	...
4595	501 N 143rd St	3.081667e+05	6360	1510	3.0	1.0	Seattle
4596	14855 SE 10th Pl	5.343333e+05	7573	1460	3.0	2.0	Bellevue
4597	759 Ilwaco Pl NE	4.169042e+05	7014	3010	3.0	2.0	Renton
4598	5148 S Creston St	2.034000e+05	6630	2090	4.0	1.0	Seattle
4599	18717 SE 258th St	2.206000e+05	8102	1490	3.0	2.0	Covington

4600 rows × 7 columns

Selanjutnya data tiap variabel akan diurutkan dari data terbesar hingga data terkecil untuk melihat nilai dari tiap data berdasarkan pada variabel city.

```
house["city"].value_counts()
```

```
Seattle                1573
Renton                  293
Bellevue                286
Redmond                 235
Issaquah                187
Kirkland                187
Kent                    185
Auburn                  176
Sammamish               175
Federal Way             148
Shoreline               123
Woodinville             115
Maple Valley             96
Mercer Island            86
Burien                   74
Snoqualmie               71
Kenmore                  66
Des Moines               58
North Bend               50
Covington                43
Duvall                   42
Lake Forest Park         36
```

```
Bothell                33
Newcastle              33
SeaTac                 29
Tukwila                29
Vashon                 29
Enumclaw               28
Carnation              22
Normandy Park          18
Clyde Hill             11
Medina                 11
Fall City              11
Black Diamond           9
Ravensdale              7
Pacific                 6
Algona                  5
Yarrow Point            4
Skykomish               3
Preston                 2
Milton                  2
Inglewood-Finn Hill     1
Snoqualmie Pass         1
Beaux Arts Village      1
Name: city, dtype: int64
```

Berdasarkan output dapat dilihat bahwa berdasarkan pada variabel city yang paling banyak adalah pada Seattle yakni sejumlah 1573. Kemudian untuk analisis selanjutnya akan digunakan data rumah pada Seattle.

```
house_seattle=house[house['city']=='Seattle']
```

```
house_seattle
```

	street	price	sqft_lot	sqft_living	floors	bedrooms	city
1	709 W Blaine St	2.384000e+06	9050	3650	2.0	5.0	Seattle
5	522 NE 88th St	4.900000e+05	6380	880	1.0	2.0	Seattle
9	6811 55th Ave NE	6.400000e+05	6200	1520	1.5	4.0	Seattle
11	3838-4098 44th Ave NE	1.400000e+06	4000	2920	1.5	4.0	Seattle
13	2504 SW Portland Ct	3.650000e+05	6435	1090	1.0	3.0	Seattle
...	...	...	...	...	...	...	...
4582	312 NE 81st St	4.060625e+05	4650	1290	1.0	2.0	Seattle
4585	4324 Dayton Ave N	4.868950e+05	3330	1890	1.5	3.0	Seattle
4591	3529 SW Webster St	3.961667e+05	5752	1880	1.0	3.0	Seattle
4595	501 N 143rd St	3.081667e+05	6360	1510	1.0	3.0	Seattle
4598	5148 S Creston St	2.034000e+05	6630	2090	1.0	4.0	Seattle

1573 rows × 7 columns

Setelah memanggil data rumah pada Seattle, kemudian akan dilakukan pengecekan apakah terdapat data duplikat dengan berdasarkan pada variabel street karena variabel ini digunakan sebagai variabel yang bersifat unik.

```
house_seattle['street'].duplicated().sum()
```

```
30
```

Berdasarkan pada output, terdapat 30 data duplikat berdasarkan pada variabel street.

```
tes=np.fromfunction(lambda i, j: i+j+1, (1573,1))
str(tes)
```

```
'[[1.000e+00]\n [2.000e+00]\n [3.000e+00]\n ...\n [1.571e+03]\n [1.572e+03]\n [1.573e+03]]'
```

```
house_seattle.insert(1, 'ID', tes)
house_seattle
```

	street	ID	price	sqft_lot	sqft_living	floors	bedrooms	city
1	709 W Blaine St	1.0	2.384000e+06	9050	3650	2.0	5.0	Seattle
5	522 NE 88th St	2.0	4.900000e+05	6380	880	1.0	2.0	Seattle
9	6811 55th Ave NE	3.0	6.400000e+05	6200	1520	1.5	4.0	Seattle
11	3838-4098 44th Ave NE	4.0	1.400000e+06	4000	2920	1.5	4.0	Seattle
13	2504 SW Portland Ct	5.0	3.650000e+05	6435	1090	1.0	3.0	Seattle
...	...	...	...	...	...	...	...	...
4582	312 NE 81st St	1569.0	4.060625e+05	4650	1290	1.0	2.0	Seattle
4585	4324 Dayton Ave N	1570.0	4.868950e+05	3330	1890	1.5	3.0	Seattle
4591	3529 SW Webster St	1571.0	3.961667e+05	5752	1880	1.0	3.0	Seattle
4595	501 N 143rd St	1572.0	3.081667e+05	6360	1510	1.0	3.0	Seattle
4598	5148 S Creston St	1573.0	2.034000e+05	6630	2090	1.0	4.0	Seattle

Berdasarkan output di atas, telah ditambahkan kolom baru untuk ‘ID’. Berikut syntax untuk menggabungkan kolom ‘street’ dengan ‘ID’.

```
1 house_seattle.loc[1:4598,'street+id'] = house_seattle['street'].str.cat(house_seattle['ID'].astype('str'),sep='; ID : ')
```

```
house_seattle
```

	street	ID	price	sqft_lot	sqft_living	floors	bedrooms	city	street+id
1	709 W Blaine St	1.0	2.384000e+06	9050	3650	2.0	5.0	Seattle	709 W Blaine St; ID : 1.0
5	522 NE 88th St	2.0	4.900000e+05	6380	880	1.0	2.0	Seattle	522 NE 88th St; ID : 2.0
9	6811 55th Ave NE	3.0	6.400000e+05	6200	1520	1.5	4.0	Seattle	6811 55th Ave NE; ID : 3.0
11	3838-4098 44th Ave NE	4.0	1.400000e+06	4000	2920	1.5	4.0	Seattle	3838-4098 44th Ave NE; ID : 4.0
13	2504 SW Portland Ct	5.0	3.650000e+05	6435	1090	1.0	3.0	Seattle	2504 SW Portland Ct; ID : 5.0
...	...	...	...	...	...	...	...	...	...
4582	312 NE 81st St	1569.0	4.060625e+05	4650	1290	1.0	2.0	Seattle	312 NE 81st St; ID : 1569.0
4585	4324 Dayton Ave N	1570.0	4.868950e+05	3330	1890	1.5	3.0	Seattle	4324 Dayton Ave N; ID : 1570.0
4591	3529 SW Webster St	1571.0	3.961667e+05	5752	1880	1.0	3.0	Seattle	3529 SW Webster St; ID : 1571.0
4595	501 N 143rd St	1572.0	3.081667e+05	6360	1510	1.0	3.0	Seattle	501 N 143rd St; ID : 1572.0
4598	5148 S Creston St	1573.0	2.034000e+05	6630	2090	1.0	4.0	Seattle	5148 S Creston St; ID : 1573.0

1573 rows × 9 columns

```
house_seattle = house_seattle.drop(house_seattle.columns[[0,1,7]], axis=1)
house_seattle
```

	price	sqft_lot	sqft_living	floors	bedrooms	street+id
1	2.384000e+06	9050	3650	2.0	5.0	709 W Blaine St; ID : 1.0
5	4.900000e+05	6380	880	1.0	2.0	522 NE 88th St; ID : 2.0
9	6.400000e+05	6200	1520	1.5	4.0	6811 55th Ave NE; ID : 3.0
11	1.400000e+06	4000	2920	1.5	4.0	3838-4098 44th Ave NE; ID : 4.0
13	3.650000e+05	6435	1090	1.0	3.0	2504 SW Portland Ct; ID : 5.0
...	...	...	...	...	...	...
4582	4.060625e+05	4650	1290	1.0	2.0	312 NE 81st St; ID : 1569.0
4585	4.868950e+05	3330	1890	1.5	3.0	4324 Dayton Ave N; ID : 1570.0
4591	3.961667e+05	5752	1880	1.0	3.0	3529 SW Webster St; ID : 1571.0
4595	3.081667e+05	6360	1510	1.0	3.0	501 N 143rd St; ID : 1572.0
4598	2.034000e+05	6630	2090	1.0	4.0	5148 S Creston St; ID : 1573.0

1573 rows × 6 columns

Berdasarkan output data di atas, terdapat kolom baru pada bagian kanan yakni street+id. Kolom ‘street’, ‘ID’, dan ‘city’ dihapus agar tidak menjadi rancu. Kemudian dilanjutkan dengan melakukan pengecekan terhadap variabel street+id untuk mengecek apakah masih terdapat data duplikat atau tidak.

```
house_seattle['street+id'].duplicated().sum()
```

```
0
```

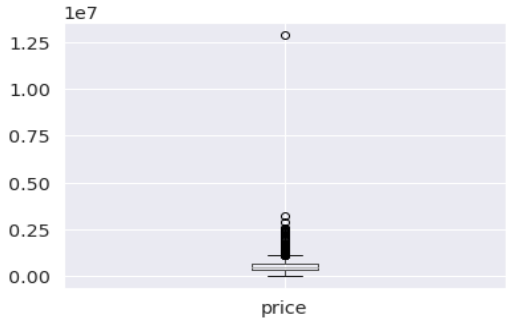
Berdasarkan pada output, tidak terdapat data duplikat berdasarkan pada variabel street+id.

```
data_clustering = house_seattle.groupby('street+id').sum()
data_clustering
```

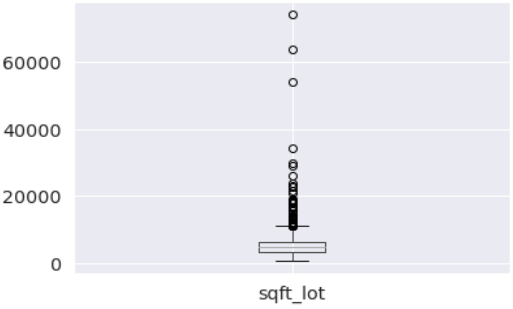
street+id	price	sqft_lot	sqft_living	floors	bedrooms
10 W Etruria St; ID : 1342.0	625000.0	1641	1820	3.0	3.0
100 20th Ave E; ID : 1308.0	600000.0	2002	910	1.5	2.0
100 24th Ave E; ID : 64.0	460000.0	929	1230	2.0	2.0
10000-10026 S 100th St; ID : 1443.0	284000.0	8800	1880	1.0	4.0
10005 16th Ave S; ID : 1086.0	265000.0	9450	1620	1.5	3.0
...	...	...	...	...	...
9853 Arrowsmith Ave S; ID : 1175.0	450000.0	5650	2310	1.0	4.0
9854 25th Ave SW; ID : 821.0	148000.0	8261	620	1.0	1.0
9957 Rainier Ave S; ID : 909.0	418000.0	6250	2360	1.0	4.0
Burke-Gilman Trail; ID : 1434.0	1675000.0	8343	3490	2.0	3.0
Schmitz Park to Alki Trail; ID : 630.0	544000.0	8203	1790	1.5	3.0

1573 rows × 5 columns

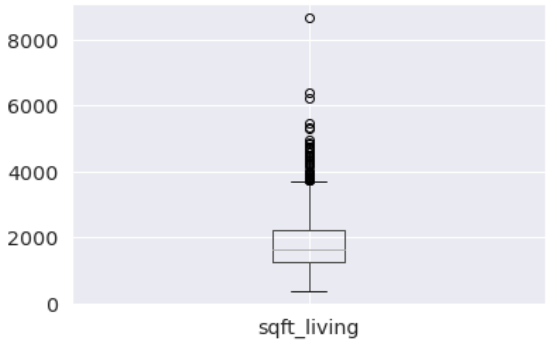
Berdasarkan output di atas, data telah bersifat unik berdasarkan pada variabel karakteristiknya. Setelah itu, dilanjutkan dengan membuat boxplot pada variabel karakteristik price.



Untuk variabel Price terlihat bahwa data mengumpul pada harga kisaran 1000000 hingga 3000000 dan terdapat 1 data outlier yang bernilai diatas 12500000 atau jika melihat pada data yang dimiliki data outlier tersebut adalah data rumah yang terletak pada 5426 40th Ave W dengan harga yang dimiliki yaitu senilai 12899000.

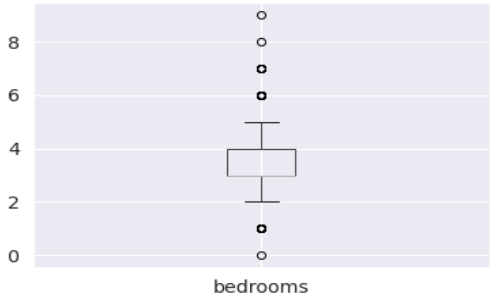


Untuk variabel sqft_lot terlihat bahwa data mengumpul pada nilai 1000 hingga 30000 dan memiliki 3 outlier dengan nilai kisaran lebih dari 50000.



Untuk variabel sqft_living terlihat bahwa data mengumpul pada nilai 500 hingga 6000 dan memiliki 1 outlier dengan nilai kisaran lebih dari 8000.

Untuk variabel floors terlihat bahwa data banyaknya lantai rumah mengumpul pada 1 hingga 3,5 lantai dan tidak memiliki outlier.



Untuk variabel bedrooms terlihat bahwa data banyaknya jumlah kamar dalam rumah mengumpul pada 3 hingga 4 kamar dan terdapat beberapa rumah yang memiliki 0-1 kamar dan ada pula yang memiliki lebih dari 6 kamar.

Melakukan Uji Asumsi

Terdapat 3 asumsi yang harus dipenuhi dalam melakukan analisis cluster, yaitu:

1. **Asumsi Kecukupan Sampel**

Sampel yang dimiliki harus bisa merepresentasikan populasi. Untuk menunjukkan kecukupan sampel, dapat dilakukan analisis dengan uji KMO dan Bartlett. Apabila data yang dimiliki adalah merupakan populasi itu sendiri, maka asumsi ini secara otomatis terpenuhi.

KMO Test

```
[37]  1 !pip install factor_analyzer
```

```
from factor_analyzer.factor_analyzer import calculate_kmo
kmo_all,kmo_model=calculate_kmo(data_clustering)
kmo_model
```

```
0.5675784415412319
```

Bartlett Test

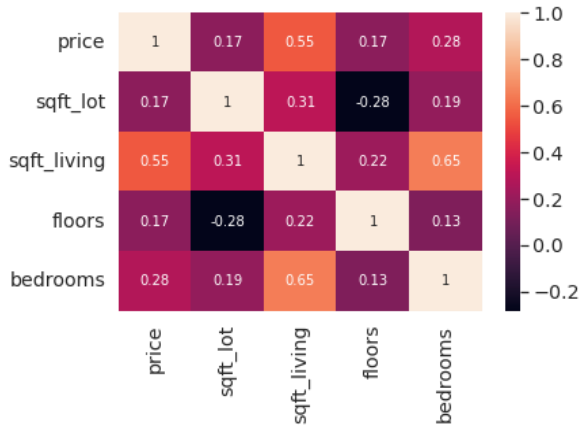
+ Code + Text

```
[ ] stats.bartlett(house_seattle.price,house_seattle.sqft_lot,house_seattle.sqft_living,house_seattle.floors,house_seattle.bedrooms)
```

```
BartlettResult(statistic=105656.27851765024, pvalue=0.0)
```

Hasil KMO test untuk data_clustering mencapai >0.5 yaitu sebesar 0.56758. Hal ini menunjukan bahwa asumsi kecukupan data telah terpenuhi. Dari Bartlett's Test diperoleh p-value sebesar 0.000. Karena p-value < 0.05, maka kriteria sudah terpenuhi. Dengan demikian dapat dikatakan bahwa variabel dan sampel yang memenuhi asumsi kecukupan sampel dan dapat digunakan untuk analisis lanjutan.

2. Asumsi No Multikolinearitas

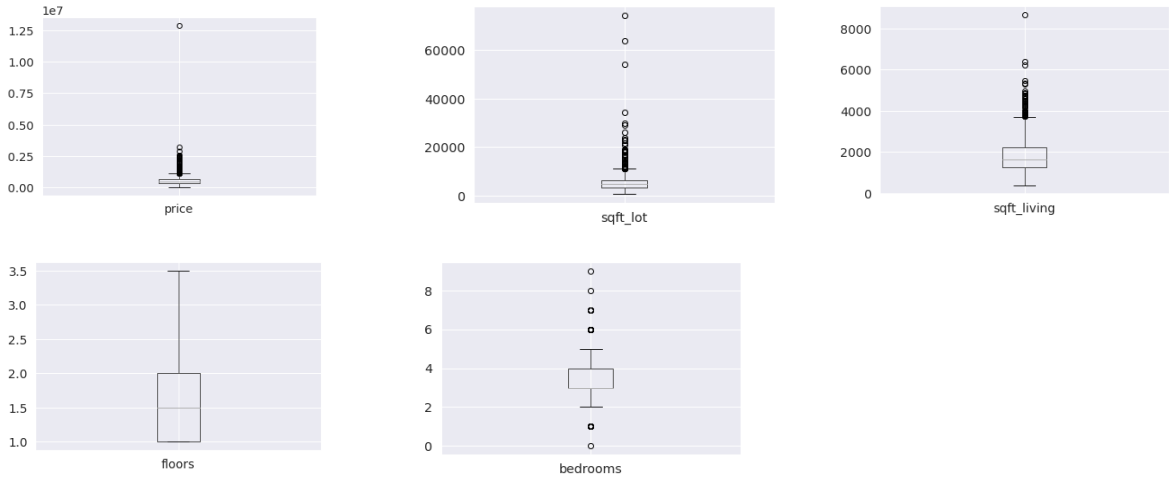


Dari output yang didapatkan bahwa semua variabel menunjukkan nilai korelasi lebih kecil dari 0.7 dan lebih besar dari -0.7 yang berarti bahwa semua variabel tidak menunjukkan adanya korelasi yang kuat. Dikarenakan korelasi yang kuat menandakan adanya hubungan linier positif yang kuat, maka dapat disimpulkan dari output bahwa tidak terdapat adanya multikolinearitas atau asumsi no multikolinearitas sudah terpenuhi. Dengan keterangan lebih lanjut sebagai berikut :

- Didapatkan hasil korelasi antara variabel price dengan sqft_lot adalah sebesar 0,17. Karena nilai 0,17 tersebut kurang dari 0.7 dan lebih dari -0.7, maka variabel tidak menunjukkan adanya korelasi yang kuat atau dapat dikatakan bahwa tidak terdapat adanya multikolinearitas.
- Didapatkan hasil korelasi antara variabel price dengan sqft_living adalah sebesar 0,55. Karena nilai 0,55 tersebut kurang dari 0.7 dan lebih dari -0.7, maka variabel tidak menunjukkan adanya korelasi yang kuat atau dapat dikatakan bahwa tidak terdapat adanya multikolinearitas.
- Didapatkan hasil korelasi antara variabel price dengan floors adalah sebesar 0,17. Karena nilai 0,17 tersebut kurang dari 0.7 dan lebih dari -0.7, maka variabel tidak menunjukkan adanya korelasi yang kuat atau dapat dikatakan bahwa tidak terdapat adanya multikolinearitas.

- Didapatkan hasil korelasi antara variabel price dengan bedrooms adalah sebesar 0,28. Karena nilai 0,28 tersebut kurang dari 0.7 dan lebih dari -0.7, maka variabel tidak menunjukkan adanya korelasi yang kuat atau dapat dikatakan bahwa tidak terdapat adanya multikolinearitas.
- Didapatkan hasil korelasi antara variabel sqft_lot dengan sqft_living adalah sebesar 0,31. Karena nilai 0,31 tersebut kurang dari 0.7 dan lebih dari -0.7, maka variabel tidak menunjukkan adanya korelasi yang kuat atau dapat dikatakan bahwa tidak terdapat adanya multikolinearitas.
- Didapatkan hasil korelasi antara variabel sqft_lot dengan floors adalah sebesar -0,28. Karena nilai -0,28 tersebut kurang dari 0.7 dan lebih dari -0.7, maka variabel tidak menunjukkan adanya korelasi yang kuat atau dapat dikatakan bahwa tidak terdapat adanya multikolinearitas.
- Didapatkan hasil korelasi antara variabel sqft_lot dengan bedrooms adalah sebesar 0,19. Karena nilai 0,19 tersebut kurang dari 0.7 dan lebih dari -0.7, maka variabel tidak menunjukkan adanya korelasi yang kuat atau dapat dikatakan bahwa tidak terdapat adanya multikolinearitas.
- Didapatkan hasil korelasi antara variabel sqft_living dengan floors adalah sebesar 0,22. Karena nilai 0,22 tersebut kurang dari 0.7 dan lebih dari -0.7, maka variabel tidak menunjukkan adanya korelasi yang kuat atau dapat dikatakan bahwa tidak terdapat adanya multikolinearitas.
- Didapatkan hasil korelasi antara variabel sqft_living dengan bedrooms adalah sebesar 0,65. Karena nilai 0,65 tersebut kurang dari 0.7 dan lebih dari -0.7, maka variabel tidak menunjukkan adanya korelasi yang kuat.
- Didapatkan hasil korelasi antara variabel floors dengan bedrooms adalah sebesar 0,13. Karena nilai 0,13 tersebut kurang dari 0.7 dan lebih dari -0.7, maka variabel tidak menunjukkan adanya korelasi yang kuat.

3. **Asumsi Bebas Pencilan**
   Untuk metode K-Means, diperlukan asumsi tambahan yakni data bebas pencilan. Dari hasil pengecekan menggunakan boxplot diperoleh hasil berikut.

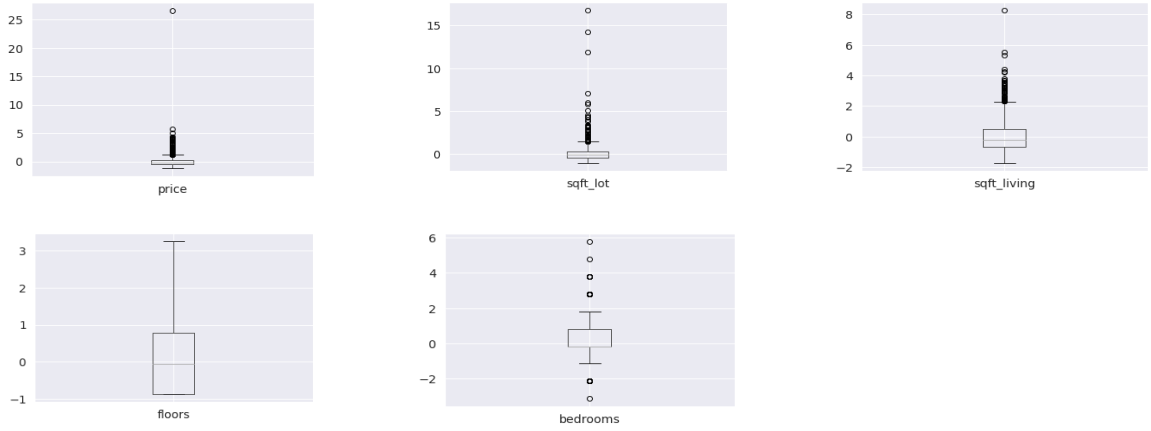


Terlihat semua variabel kecuali ‘floors’ memiliki pencilan. Maka akan dilakukan standarisasi pada semua variabel agar tidak terdapat perbedaan satuan antar variabel.

```
aaa=['price','sqft_lot','sqft_living','floors','bedrooms']
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
data_clustering[aaa] = sc.fit_transform(data_clustering[aaa])
data_clustering
```

street+id	price	sqft_lot	sqft_living	floors	bedrooms
10 W Etruria St; ID : 1342.0	0.097629	-0.895304	-0.010459	2.427328	-0.168847
100 20th Ave E; ID : 1308.0	0.043586	-0.807611	-1.114130	-0.050941	-1.163586
100 24th Ave E; ID : 64.0	-0.259057	-1.068262	-0.726026	0.775148	-1.163586
10000-10026 S 100th St; ID : 1443.0	-0.639523	0.843744	0.062311	-0.877031	0.825893
10005 16th Ave S; ID : 1086.0	-0.680596	1.001641	-0.253024	-0.050941	-0.168847
...	...	...	...	...	...
9853 Arrowsmith Ave S; ID : 1175.0	-0.280674	0.078553	0.583825	-0.877031	0.825893
9854 25th Ave SW; ID : 821.0	-0.933519	0.712812	-1.465849	-0.877031	-2.158326
9957 Rainier Ave S; ID : 909.0	-0.349850	0.224304	0.644467	-0.877031	0.825893
Burke-Gilman Trail; ID : 1434.0	2.367452	0.732731	2.014959	0.775148	-0.168847
Schmitz Park to Alki Trail; ID : 630.0	-0.077471	0.698723	-0.046844	-0.050941	-0.168847

1573 rows × 5 columns

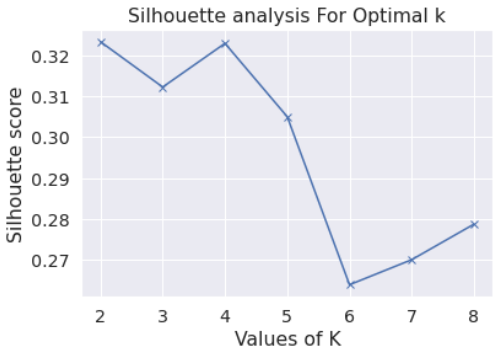


Dapat dilihat standarisasi tidak terlalu berpengaruh terhadap data, namun data hasil standarisasi tetap lebih baik daripada data asli, sehingga akan digunakan data hasil standarisasi untuk analisis cluster.

Dikarenakan semua asumsi sudah terpenuhi yaitu asumsi kecukupan sampel, asumsi no multikolinearitas, dan asumsi bebas pencilan, maka uji dapat dilanjutkan ke uji selanjutnya yaitu clustering data menggunakan metode K-Means Clustering dan K-Medoids (PAM).

K-MEANS

Membuat Plot Silhouette



Berdasarkan output di atas, terlihat bahwa cluster yang paling bagus nilainya terletak di 2. Jadi, dapat disimpulkan bahwa akan digunakan 2 cluster.

```
data_result_kmeans = data_clustering.copy()

data_result_kmeans['cluster'] = bestclusters

data_result_kmeans
```

Selanjutnya akan dibuat variabel baru bernama ‘cluster’ yang berisi cluster-cluster yang terbentuk dalam bestclusters. Cluster yang didapatkan di K=2.

street+id	price	sqft_lot	sqft_living	floors	bedrooms	cluster
10 W Etruria St; ID : 1342.0	0.097629	-0.895304	-0.010459	2.427328	-0.168847	1
100 20th Ave E; ID : 1308.0	0.043586	-0.807611	-1.114130	-0.050941	-1.163586	1
100 24th Ave E; ID : 64.0	-0.259057	-1.068262	-0.726026	0.775148	-1.163586	1
10000-10026 S 100th St; ID : 1443.0	-0.639523	0.843744	0.062311	-0.877031	0.825893	1
10005 16th Ave S; ID : 1086.0	-0.680596	1.001641	-0.253024	-0.050941	-0.168847	1
...	...	...	...	...	...	...
9853 Arrowsmith Ave S; ID : 1175.0	-0.280674	0.078553	0.583825	-0.877031	0.825893	0
9854 25th Ave SW; ID : 821.0	-0.933519	0.712812	-1.465849	-0.877031	-2.158326	1
9957 Rainier Ave S; ID : 909.0	-0.349850	0.224304	0.644467	-0.877031	0.825893	0
Burke-Gilman Trail; ID : 1434.0	2.367452	0.732731	2.014959	0.775148	-0.168847	0
Schmitz Park to Alki Trail; ID : 630.0	-0.077471	0.698723	-0.046844	-0.050941	-0.168847	1

1573 rows × 6 columns

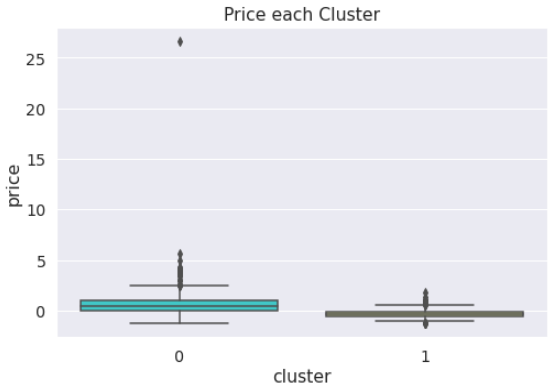
Berdasarkan output di atas, terdapat kolom baru bernama cluster dimana kolom ini menunjukkan pembagian cluster tiap data rumah. Dikarenakan sebelumnya sudah didapatkan bahwa akan digunakan 2 kluster, maka tiap data terbagi menjadi 2 kluster. Kluster pertama adalah 1 dan kluster kedua adalah 0.

```
1 data_result_kmeans['cluster'].value_counts()
```

```
1    1099
0     474
Name: cluster, dtype: int64
```

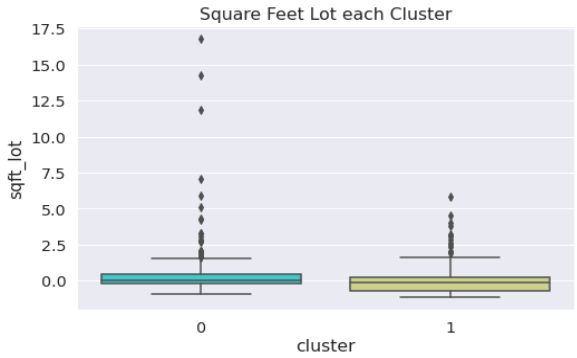
Berdasarkan output di atas, ditunjukkan jumlah data rumah di tiap klusternya. Untuk kluster pertama yaitu 1 didapatkan bahwa ada sebanyak 1099 rumah yang masuk ke kluster pertama. Selanjutnya, untuk kluster kedua yaitu 0 didapatkan bahwa ada sebanyak 474 rumah yang masuk ke kluster kedua. Dapat diketahui bahwa jumlah kluster pertama lebih banyak daripada kluster kedua.

Boxplot variabel Price



Berdasarkan output pada boxplot, apabila harga rumah berkisar antara -0,5-2,5 maka akan masuk ke dalam cluster 0 sedangkan apabila harga rumah berkisar antara -0,3-0,1 maka akan masuk ke dalam cluster 1. Pada cluster 0 terdapat adanya outlier sedangkan pada cluster 1 tidak terdapat adanya outlier.

Boxplot variabel square feet lot

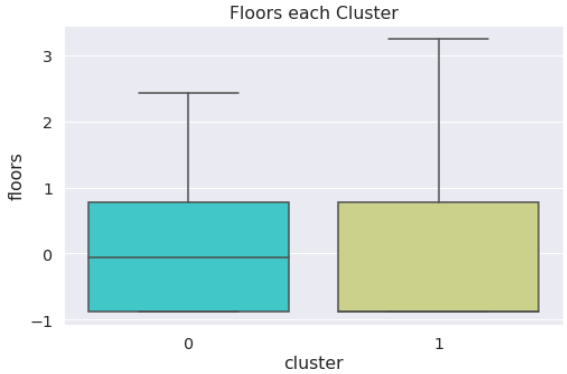


Berdasarkan output pada boxplot, pada cluster 0 sqft lot berkisar antara -0,5-2 dan terdapat adanya outlier, sedangkan pada cluster 1 sqft lot rumah berkisar antara -0,6-2 dan terdapat adanya outlier.

Boxplot variabel square feet living

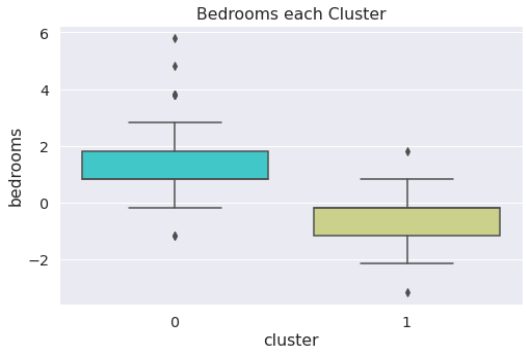
Berdasarkan output pada boxplot sqft living, pada cluster 0 sqft living berkisar antara -0,8-3 dan terdapat adanya outlier, sedangkan pada cluster 1 sqft living rumah berkisar antara -1,8-1 dan terdapat adanya outlier.

Boxplot variabel Floors



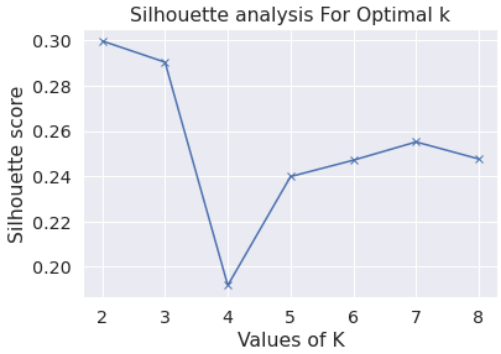
Berdasarkan output pada boxplot variabel floors, pada cluster 0 berkisar antara -0.8-2.5 dan tidak terdapat adanya outlier, sedangkan pada cluster 1 berkisar antara -0.8-3,25 dan tidak terdapat adanya outlier.

Boxplot variabel Bedrooms



Berdasarkan output pada boxplot variabel bedrooms, pada cluster 0 bedrooms berkisar di 0-3 dan terdapat adanya outlier, sedangkan pada cluster 1 bedrooms rumah berkisar antara -2,1-1 dan terdapat adanya outlier.

K-MEDOIDS



Berdasarkan output di atas, terlihat bahwa cluster yang paling bagus nilainya terletak di 2. Jadi, dapat disimpulkan bahwa akan digunakan 2 cluster. Selanjutnya akan dibuat variabel baru bernama ‘cluster’ yang berisi cluster-cluster yang terbentuk dalam bestclusters. Cluster yang didapatkan di K=2.

```
1 data_result_kmedoids = data_clustering.copy()
```

```
1 data_result_kmedoids['cluster'] = bestclusters
```

```
1 data_result_kmedoids
```

street+id	price	sqft_lot	sqft_living	floors	bedrooms	cluster
10 W Etruria St; ID : 1342.0	0.097629	-0.895304	-0.010459	2.427328	-0.168847	1
100 20th Ave E; ID : 1308.0	0.043586	-0.807611	-1.114130	-0.050941	-1.163586	1
100 24th Ave E; ID : 64.0	-0.259057	-1.068262	-0.726026	0.775148	-1.163586	1
10000-10026 S 100th St; ID : 1443.0	-0.639523	0.843744	0.062311	-0.877031	0.825893	0
10005 16th Ave S; ID : 1086.0	-0.680596	1.001641	-0.253024	-0.050941	-0.168847	1
...	...	...	...	...	...	...
9853 Arrowsmith Ave S; ID : 1175.0	-0.280674	0.078553	0.583825	-0.877031	0.825893	0
9854 25th Ave SW; ID : 821.0	-0.933519	0.712812	-1.465849	-0.877031	-2.158326	1
9957 Rainier Ave S; ID : 909.0	-0.349850	0.224304	0.644467	-0.877031	0.825893	0
Burke-Gilman Trail; ID : 1434.0	2.367452	0.732731	2.014959	0.775148	-0.168847	0
Schmitz Park to Alki Trail; ID : 630.0	-0.077471	0.698723	-0.046844	-0.050941	-0.168847	1

1573 rows × 6 columns

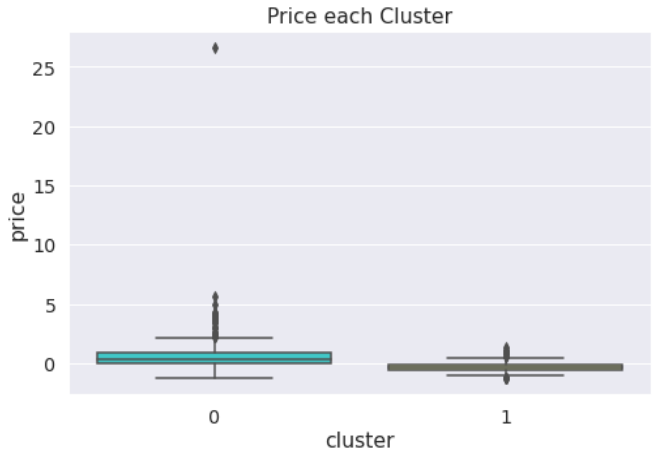
Berdasarkan output di atas, terdapat kolom baru bernama cluster dimana kolom ini menunjukkan pembagian cluster tiap data rumah. Dikarenakan sebelumnya sudah didapatkan bahwa akan digunakan 2 kluster, maka tiap data terbagi menjadi 2 kluster. Kluster pertama adalah 0 dan kluster kedua adalah 1.

```
data_result_kmedoids['cluster'].value_counts()
```

```
1    1017
0     556
Name: cluster, dtype: int64
```

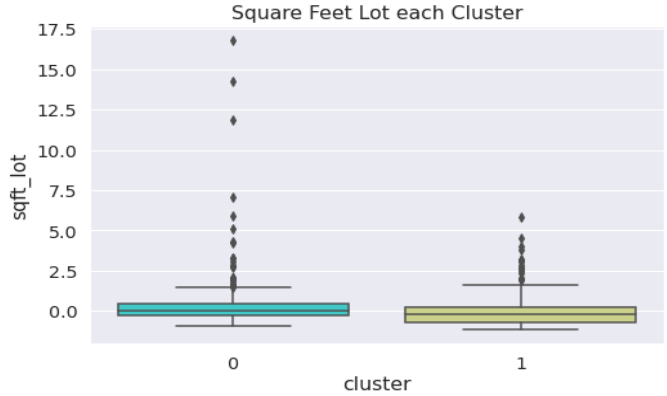
Berdasarkan output di atas, ditunjukkan jumlah data rumah di tiap klusternya. Untuk kluster pertama yaitu 0 didapatkan bahwa ada sebanyak 556 rumah yang masuk ke kluster pertama. Selanjutnya, untuk kluster kedua yaitu 1 didapatkan bahwa ada sebanyak 1017 rumah yang masuk ke kluster kedua. Dapat diketahui bahwa jumlah kluster kedua lebih banyak dibandingkan kluster yang lain.

Boxplot variabel Price



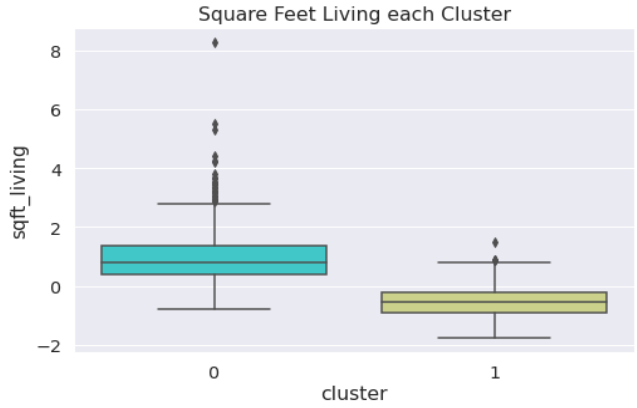
Berdasarkan output pada boxplot, cluster 0 berkisar antara -0,5-2,5, cluster 1 cenderung berkumpul di sekitar angka -0,1-2.

Boxplot variabel Square Feet Lot



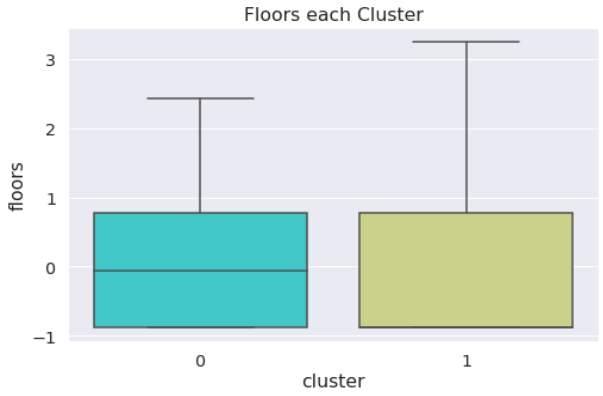
Berdasarkan output pada boxplot, pada cluster 0 sqft lot berkisar antara -0,5-2,2 dan terdapat adanya outlier, cluster 1 sqft lot rumah berkisar antara -0,6-2,3 dan terdapat adanya outlier.

Boxplot variabel Square Feet Living



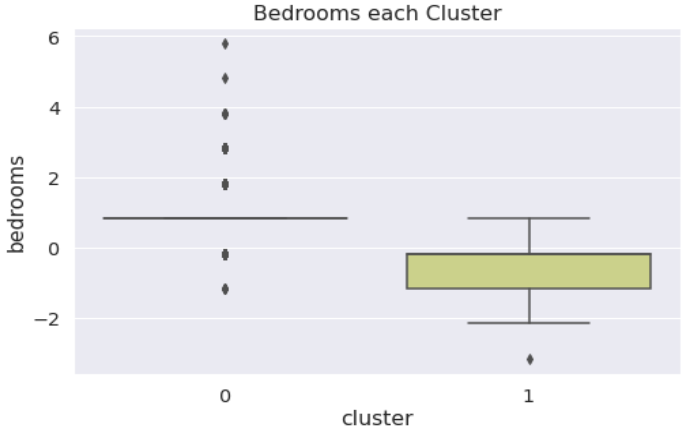
Berdasarkan output pada boxplot sqft living, pada cluster 0 sqft living berkisar antara -0,8-3 dan terdapat adanya outlier, cluster 1 sqft living rumah berkisar antara -1,8-1 dan terdapat adanya outlier.

Boxplot variabel Floors



Berdasarkan output pada boxplot variabel floors, pada cluster 0 berkisar antara -0.8-2.5 dan tidak terdapat adanya outlier, cluster 1 berkisar antara -0.8-3,25 dan tidak terdapat adanya outlier.

Boxplot variabel Bedrooms



Berdasarkan output pada boxplot variabel bedrooms, pada cluster 0 bedrooms cenderung berkumpul di angka 1 terdapat adanya outlier, sedangkan pada cluster 1 bedrooms rumah berkisar antara -2,1-3 dan terdapat adanya outlier.

KESIMPULAN DAN SARAN

Kesimpulan

Dilakukan analisis clustering terhadap data *house price prediction* dari situs *Kaggle*. Berdasarkan hasil analisis K-Means Clustering, didapat nilai silhouette terbaik adalah pada 2 cluster dengan nilai 0.32329. Artinya terbentuk 2 cluster, yaitu cluster 0 yang terdiri dari 1099 data dan cluster 1 sebanyak 474 data. Pada analisis K-Medoids Clustering juga terbentuk 2 cluster dengan nilai silhouette terbaik yaitu 0,300. Artinya terbentuk 2 cluster, yaitu cluster 0 yang terdiri dari 1017 data dan cluster 1 sebanyak 556 data. Dari kedua nilai silhouette masing-masing metode tersebut, dapat dilihat bahwa metode clustering yang paling baik untuk digunakan pada dataset ini adalah metode K-Means Clustering karena memiliki nilai silhouette yang lebih tinggi.

Kemudian jika ditinjau berdasarkan nilai mean untuk setiap variabel pada masing-masing metode cluster didapat kesimpulan sebagai berikut:

Metode K-Means Clustering:

- Cluster 0 memiliki rata-rata nilai price (harga rumah) yang lebih tinggi dari cluster 1
- Cluster 0 memiliki rata-rata nilai square feet lot (luas lahan) yang lebih tinggi dari cluster 1
- Cluster 0 memiliki rata-rata nilai square feet living (luas bangunan) yang lebih tinggi dari cluster 1
- Cluster 0 memiliki rata-rata floor (jumlah lantai) yang sama dengan cluster 1
- Cluster 0 memiliki rata-rata bedrooms (jumlah kamar) yang lebih dari cluster 1

Metode K-Medoids Clustering:

- Cluster 0 memiliki rata-rata nilai price (harga rumah) yang lebih tinggi dari cluster 1
- Cluster 0 memiliki rata-rata nilai square feet lot (luas lahan) yang lebih tinggi dari cluster 1
- Cluster 0 memiliki rata-rata nilai square feet living (luas bangunan) yang lebih tinggi dari cluster 1
- Cluster 0 memiliki rata-rata floor (jumlah lantai) yang sama dengan cluster 1
- Cluster 0 memiliki rata-rata bedrooms (jumlah kamar) yang lebih dari cluster 1

Saran

Berdasarkan hasil analisis yang diperoleh, dapat diketahui bahwa metode K-Means Clustering lebih baik dalam melakukan pengelompokkan. Adapun saran yang dapat kami berikan kepada calon pembeli adalah untuk dapat menggunakan hasil clustering ini sebagai rekomendasi dalam menentukan rumah yang akan dibeli. Jika calon pembeli memiliki modal lebih banyak, kami juga menyarankan untuk membeli rumah yang terdapat pada cluster 0.

REFERENSI

Jiawei Han, M. K. (2012). *Data Mining : Concepts and Techniques*. Waltham,: Elsevier Inc.

Mattjik, A. A., & Sumertajaya, I. M. (2011). Sidik Peubah Ganda (G. N. A. Wibawa & A. F. Hadi, Eds.). Bandung: IPB Press.

Santosa, B. (2007). *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.

Shree. (2017). *House price prediction*, Kaggle, diakses pada 21 Maret 2022, https://www.kaggle.com/datasets/shree1992/housedata?select=data.csv

W. Song and S. C. Park, "A Novel Document Clustering Model Based on Latent Semantic Analysis," pp. 539–542, 2007

Musfiani. (2019). Analisis Cluster dengan Menggunakan Metode Partisi pada Pengguna Alat Kontrasepsi di Kalimantan Barat. Buletin Ilmiah Math. Stat. Dan Terapannya (Bimaster), 893–896.

Sihombing, R., Rachmatin, D., Dahlan, J., & Kunci, K. (2020). Program Aplikasi Bahasa R untuk Pengelompokan Objek Menggunakan Metode K-Medoids Clustering. 7, 58–79.

United States Census Bureau. (2014). *Selected Housing Characteristics*. Diakses pada 25 Maret 2022, dari https://data.census.gov/cedsci/table?q=seattle%20city%20DP04&g=1600000US5363000&tid=ACSDP1Y2014.DP04

Presiden Republik Indonesia. 1984. Undang-Undang Republik Indonesia Nomor 4 Tahun 1992 tentang Perumahan Dan Permukiman. Lembaran Negara Republik Indonesia Tahun 1992, No. 23. Sekretariat Negara. Jakarta.