PERBANDINGAN METODE K-MEANS CLUSTERING ANALYSIS DENGAN PCA DAN TANPA PCA PADA SEGMENTASI DATA PENJUALAN BERLIAN

MULTIVARIATE ANALYSIS

Oleh:

1. **Federico Saputra - 23101910069**
2. **Kevin Adrian Halim - 23101910055**
3. **Nicolas Maria Andre Gozali - 23101910040**
4. **Tang Owenn Gimli - 23101910008**
5. **Wesley Mulia Salim - 23101910052**

Universitas Prasetiya Mulya
2022

DAFTAR ISI

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Dalam suatu proses analisis, penggunaan teknik analisis yang tepat menjadi kunci untuk memberikan rekomendasi ataupun keputusan yang baik kepada pemangku keputusan. Pada analisis terhadap data yang memiliki banyak variabel, salah satu teknik analisis yang dikenal adalah *Multivariate Analysis*. Menurut Dillon dan Goldstein dalam Waluyo (2007), *multivariate analysis* didefinisikan sebagai semua metode statistik yang menganalisis beberapa pengukuran (variabel-variabel) yang ada pada setiap objek dalam satu atau banyak sampel secara simultan. Berdasarkan definisi ini, dapat diketahui bahwa *multivariate analysis* adalah teknik analisis secara simultan terhadap data yang memiliki variabel lebih banyak dari dua.

Menurut Dillon dan Goldstein dalam Waluyo (2007), teknik analisis multivariat secara umum dapat dibagi menjadi dua kelompok besar, yaitu :

1. *Dependence Methods* : Teknik multivariat yang didalamnya terdapat variabel atau set variabel terikat (*dependent variable*) dan variabel lainnya sebagai variabel bebas (*independent variable*).
2. *Interdependence Methods* : Teknik multivariat di mana semua variable dianalisis secara simultan, tidak ada variabel yang didefinisikan bebas atau terikat.

Dalam analisis menggunakan teknik *multivariate*, terdapat beberapa jenis analisis yang dapat digunakan seperti *Factor Analysis, Principal Component Analysis, Cluster Analysis, Multidimensional Scaling, Canonical Correlation Analysis*, dan masih banyak lainnya.

Dalam penelitian kali ini, jenis analisis yang akan digunakan adalah *Cluster Analysis* dan *Principal Component Analysis.* Kedua jenis analisis ini dipilih karena data yang digunakan adalah data yang diasumsikan sebagai data penjualan berlian (*diamond*) pada sebuah toko fiktif, yang setiap unit *diamond* yang keluar memiliki ciri khas masing-masing, seperti carat, cut, color, length, width, depth, dan lainnya.

Dari data ini, peneliti ingin melakukan pengelompokan (dalam hal ini *Cluster Analysis*) untuk membantu pemilik toko berlian dalam proses sortir produk yang masuk berdasarkan *cluster*-nya masing-masing dan melacak penjualan *diamond* terbanyak setiap bulannya berdasarkan *cluster*-nya. Tujuan dilakukannya *clustering* ini adalah untuk mempermudah penjual dalam melihat keseluruhan penjualan setiap bulannya dalam kelompok tertentu. Untuk lebih lanjutnya, penelitian ini akan membandingkan data asli yang di-*clustering* dengan data yang telah dilakukan PCA lalu di-*clustering*. Hal ini dilakukan untuk melihat perbandingan hasil yang ditemukan, agar mampu memberikan hasil dan rekomendasi terbaik bagi pemilik toko dan pada akhirnya mampu membantu dalam pemilihan keputusan dalam rangka peningkatan penjualan.

1.2 Rumusan Masalah

Rumusan masalah yang akan dihadapi dalam analisis ini adalah :

1.2.1 Dalam suatu pengelompokan *clustering*, variabel apa saja yang dipakai dan berapa banyak *cluster* yang dihasilkan?

1.2.2 Bagaimana suatu produk dapat diidentifikasikan atau termasuk dalam suatu *cluster*?

1.3 Tujuan Analisis

Analisis ini dilakukan dengan tujuan :

1.3.1 Untuk melihat perbandingan dari hasil antara *cluster analysis* yang masing-masing menggunakan data yang telah dilakukan PCA dan tidak dilakukan PCA.

1.3.2 Untuk menentukan setiap ciri-ciri berlian tertentu agar dapat dimasukkan atau didaftarkan ke dalam suatu *cluster* dan mempermudah identifikasinya.

1.4 Manfaat Analisis

Manfaat dari analisis ini adalah penulis berharap agar hasil yang dikemukakan dapat menjadi rekomendasi dalam memberikan bantuan kepada pemegang kepentingan di toko berlian tersebut untuk mengambil keputusan melihat dari banyaknya penjualan terhadap suatu *cluster*. Hasil *clustering* ini diharapkan dapat membantu pemilik toko untuk melihat suatu berlian berdasarkan kelompoknya, agar memudahkan ketika ada barang baru yang masuk dan melacak penjualan produk setiap bulannya; agar setiap *cluster* dengan penjualan yang rendah dapat ditingkatkan ataupun tidak dilanjutkan penjualannya, atau *cluster* dengan penjualan yang tinggi dapat dipertahankan ataupun ditingkatkan lagi.

BAB 2
LANDASAN TEORI

2.1 Analisis Multivariat

Analisis multivariat adalah studi statistik data di mana beberapa pengukuran dilakukan pada setiap unit percobaan dan di mana hubungan antara pengukuran multivariat dan strukturnya penting. Kategorisasi analisis multivariat yang modern dan tumpang tindih adalah: model multivariat normal dan umum dan teori distribusi; studi dan pengukuran hubungan; perhitungan probabilitas daerah multidimensi; dan eksplorasi struktur dan pola data. Distribusi normal multivariat memainkan peran sentral dalam analisis multivariat dengan cara yang sama seperti distribusi normal univariat memainkan peran sentral. Distribusi multivariat yang lebih baru dapat memodelkan data ketika distribusi normal multivariat tidak memadai. Untuk data multidimensi, hubungan antar variabel sangat penting untuk dijelajahi. Teknik yang berguna untuk memahami dan mengukur ini termasuk analisis regresi multivariat dan berbagai pengertian korelasional seperti korelasi parsial dan korelasi kanonik. Pendekatan untuk menghitung probabilitas multidimensi yang rumit termasuk memperoleh batas bawah untuk probabilitas dan menggunakan teknik pendekatan numerik. Eksplorasi struktur dan pola untuk kumpulan data multivariat yang kompleks sangat penting untuk analisis data modern dan penambangan data. Alat multivariat, berguna dalam konteks ini, termasuk analisis komponen utama, analisis kanonik, analisis faktor, analisis jalur, metode persamaan struktural, pengelompokan, dan analisis diskriminan (I. Olkin, A.R. Sampson, 2001).

2.2 *Kaiser Meyer Olkin*

Kaiser Meyer Olkin (KMO) dan Measure of Sampling Adequacy (MSA). KMO adalah uji yang menunjukkan apakah metode sampling yang dipergunakan telah memenuhi syarat. Statistik yang menunjukkan proporsi varians dalam variabel yang mungkin disebabkan oleh faktor faktor yang mendasarinya.

Kaiser Meyer Olkin dipergunakan untuk meneliti ketepatan suatu analisis faktor dengan melakukan perbandingan koefisien korelasi sampel yang diamati dengan koefisien korelasi parsial. Ketepatan dari proses analisis yang dilakukan ditentukan dengan nilai KMO. Jika nilai KMO berada di antara 0.5 sampai 1. Kriteria nilai uji KMO dari matriks antara variabel diantaranya adalah:

1. Nilai uji $0,9 < KMO \leq 1,00$ = sangat memuaskan,
2. Nilai uji $0,8 < KMO \leq 0,9$ = sangat baik,
3. Nilai uji $0,7 < KMO \leq 0,8$ = baik,
4. Nilai uji $0,6 < KMO \leq 0,7$ = cukup baik,
5. Nilai uji $0,5 < KMO \leq 0,6$ = jelek. Nilai uji $\leq 0,5$ = ditolak.

2.3 *Clustering*

Clustering adalah salah satu teknik analisis data eksplorasi yang paling umum digunakan untuk mendapatkan intuisi tentang struktur data. Ini dapat didefinisikan sebagai tugas mengidentifikasi sub kelompok dalam data sedemikian rupa sehingga titik data dalam sub kelompok yang sama (cluster) sangat mirip sedangkan titik data dalam kelompok yang

berbeda sangat berbeda. Dengan kata lain, peneliti mencoba untuk menemukan sub kelompok yang homogen dalam data sedemikian rupa sehingga titik data di setiap klaster semirip mungkin menurut ukuran kesamaan seperti jarak berbasis euclidean atau jarak berbasis korelasi. Keputusan ukuran kesamaan mana yang akan digunakan adalah spesifik aplikasi (Imad Dabbura, 2018).

Analisis pengelompokan dapat dilakukan berdasarkan fitur dimana peneliti mencoba menemukan subkelompok sampel berdasarkan fitur atau berdasarkan sampel dimana peneliti mencoba menemukan subkelompok fitur berdasarkan sampel. Peneliti akan membahas disini pengelompokan berdasarkan fitur. Clustering digunakan dalam segmentasi pasar; dimana peneliti mencoba mencari pelanggan yang mirip satu sama lain baik dari segi perilaku atau atribut, segmentasi/kompresi citra; tempat peneliti mencoba mengelompokkan wilayah yang serupa, pengelompokan dokumen berdasarkan topik, dll.

Tidak seperti *supervised learning*, pengelompokan dianggap sebagai metode un*supervised learning* karena peneliti tidak memiliki kebenaran dasar untuk membandingkan output dari algoritma pengelompokan dengan label yang sebenarnya untuk mengevaluasi kinerjanya. Peneliti hanya ingin mencoba menyelidiki struktur data dengan mengelompokkan titik data ke dalam sub kelompok yang berbeda.

2.4 *Principal Component Analysis*

(PCA) merupakan teknik analisis multivariate yang dilakukan dengan mereduksi variabel ke dalam variabel baru dengan jumlah lebih sedikit dari variabel asal (Johnson, 2002). PCA merupakan sebuah metode yang dapat digunakan untuk memberi bobot masing-masing kriteria yang digunakan dalam pemilihan suatu keputusan. Prosedur PCA pada dasarnya bertujuan untuk menyederhanakan variabel yang diamati dengan cara menyusutkan (mereduksi) dimensinya. Prosedur tersebut dilakukan dengan cara menghilangkan korelasi antara variabel bebas melalui transformasi variabel bebas asal ke variabel baru yang tidak berkorelasi sama sekali atau yang biasa disebut dengan principal component (Sharma, 2006).

PCA adalah sebuah teknik untuk membangun variabel-variabel baru yang merupakan kombinasi linear dari variabel-variabel asli (Soemartini, 2008). Jumlah maximum dari variabel-variabel baru ini akan sama dengan jumlah dari variabel lama, dan variabel-variabel baru ini tidak saling berkorelasi satu sama lain. Prosedur PCA pada dasarnya adalah bertujuan untuk menyederhanakan variabel yang diamati dengan cara menyusutkan (mereduksi) dimensinya. Hal ini dilakukan dengan cara menghilangkan korelasi antara variabel bebas melalui transformasi variabel bebas asal ke variabel baru yang tidak berkorelasi sama sekali atau yang biasa disebut dengan Principal Component. Setelah beberapa komponen hasil PCA yang bebas multikolinearitas diperoleh, maka komponen-komponen tersebut menjadi variabel bebas baru yang akan diregresikan atau dianalisa pengaruhnya terhadap variabel tak bebas (Y) dengan menggunakan analisis regresi.

PCA adalah metode statistik yang bertujuan untuk mereduksi data serta mengidentifikasi komponen yang dapat diperhitungkan variabilitas keseluruhan dalam variabel yang menjadi pertimbangan. Komponen utama adalah kombinasi linear dari perhitungan variabel untuk variabilitas umum. Langkah-langkah yang terlibat dari penerapan PCA adalah (1) ekstraksi komponen awal, (2) penentuan komponen yang signifikan, dipertahankan dalam model, (3) rotasi matriks berdasarkan faktor pembebanan untuk

mendapatkan solusi, (4) interpretasi solusi,(5) penghitungan skor untuk masing-masing faktor dan skor umum, (6) sintesis hasil dalam tabel (Petrisoret al.,2012).

2.5 *K-means Clustering*

K-Means adalah salah satu algoritma pembelajaran tanpa pengawasan paling sederhana yang memecahkan masalah pengelompokan yang terkenal. Prosedur ini mengikuti cara yang sederhana dan mudah untuk mengklasifikasikan kumpulan data yang diberikan melalui sejumlah klaster tertentu (asumsikan k klaster) tetap secara apriori (Kodinariya, Trupti M., and Prashant R. Makwana, 2013).

Ide utamanya adalah mendefinisikan k centroid, satu untuk setiap cluster. Centroid ini harus ditempatkan dengan cara yang cerdik karena lokasi yang berbeda menyebabkan hasil yang berbeda. Jadi, pilihan yang lebih baik adalah menempatkan mereka sejauh mungkin dari satu sama lain. Langkah selanjutnya adalah mengambil setiap titik milik kumpulan data yang diberikan dan mengaitkannya ke centroid terdekat. Ketika tidak ada poin yang tertunda, langkah pertama selesai dan grup awal selesai. Pada titik ini, perlu dilakukan penghitungan ulang k centroid baru sebagai pusat cluster yang dihasilkan dari langkah sebelumnya. Setelah k centroid baru ini, pengikatan baru harus dilakukan antara titik data yang sama dan centroid baru terdekat. Sebuah loop telah dibuat. Sebagai hasil dari loop ini, mungkin terlihat bahwa k centroid mengubah lokasinya selangkah demi selangkah sampai tidak ada lagi perubahan yang dilakukan. Dengan kata lain centroid tidak bergerak lagi. Cara kerja algoritma *K-means Clustering* adalah sebagai berikut:

1. Menentukan jumlah *cluster* K.
2. Menentukan centroid awal dengan terlebih dahulu mengacak dataset dan kemudian secara acak memilih K titik data untuk centroid tanpa penggantian.
3. Terus iterasi sampai tidak ada perubahan pada centroid dimana penetapan titik data ke *cluster* tidak berubah.
4. Hitung jumlah kuadrat jarak antara titik data dan semua centroid.
5. Tetapkan setiap titik data ke *cluster* terdekat (centroid).
6. Hitung centroid untuk *cluster* dengan mengambil rata-rata dari semua titik data yang dimiliki setiap *cluster*.

Pendekatan yang kmeans ikuti untuk memecahkan masalah disebut *Expectation-Maximization*. *Step-E* digunakan untuk menetapkan titik data ke *cluster* terdekat. *M-step* digunakan untuk menghitung centroid dari setiap *cluster*.

Algoritma *K-means* sangat populer dan dapat diaplikasikan dalam berbagai hal seperti segmentasi pasar, pengelompokan dokumen, segmentasi gambar dan kompresi gambar, dll. Tujuan biasanya ketika kita menjalani analisis klaster adalah untuk mendapatkan intuisi yang bermakna tentang struktur data yang sedang kita tangani dan *cluster-then-predict* di mana model yang berbeda akan dibangun untuk sub kelompok yang berbeda jika kami percaya ada variasi yang luas dalam perilaku sub kelompok yang berbeda. Contohnya adalah mengelompokkan pasien ke dalam sub kelompok yang berbeda dan membangun model untuk setiap sub kelompok untuk memprediksi kemungkinan risiko serangan jantung (Imad Dabbura, 2018).

Berlawanan dengan *supervised learning* dimana kami memiliki dasar untuk mengevaluasi kinerja model, analisis pengelompokan tidak memiliki metrik evaluasi yang solid yang dapat kami gunakan untuk mengevaluasi hasil dari algoritma pengelompokan yang

berbeda. Selain itu, karena *K-means* membutuhkan *k* sebagai input dan tidak mempelajarinya dari data, tidak ada jawaban yang benar dalam hal jumlah cluster yang harus kita miliki dalam masalah apa pun. Terkadang pengetahuan domain dan intuisi dapat membantu tetapi biasanya tidak demikian. Dalam metodologi prediksi klaster, kita dapat mengevaluasi seberapa baik kinerja model berdasarkan klaster K yang berbeda karena klaster digunakan dalam pemodelan *downstream*. Dua metode metrik yang sering digunakan adalah *Elbow method* dan *Silhouette analysis.*

2.6 Dataset

Dataset adalah suatu database di dalam memori (in-memory). Dataset memiliki semua karakteristik, fitur dan fungsi dari database biasa. Dataset dapat memiliki banyak tabel, dan tabel tabel dapat memiliki hubungan (relationship). Tabel-tabel pada suatu dataset dapat memiliki foreign key dan integritas referensial. Dataset adalah objek yang merepresentasikan data dan relasinya di memory. Strukturnya mirip dengan data yang ada di database. Dataset berisi koleksi dari data tabel dan data. Jenis dataset ada dua macam yaitu : Private Dataset Private dataset yaitu dataset yang dapat diambil dari organisasi yang kita jadikan tempat atau objek penelitian. Adapun contoh-contohnya seperti instansi, rumah sakit, pabrik, perusahaan jasa, etc. Public Dataset Public dataset yaitu dataset yang dapat diambil dari repository public yang telah disepakati oleh para peneliti.

Dataset yang digunakan oleh peneliti bernama *Diamonds,* dimana dataset tersebut dapat digunakan untuk menganalisa berlian berdasarkan potongan, warna, kejernihan, harga dan atribut lainnya. Dataset ini berisi harga dan atribut lain dari hampir 54.000 berlian. Variabel-variabel yang terdapat pada dataset ini adalah sebagai berikut:

1. *Price* - Harga dalam dolar AS
2. *Carat* - Berat dari berlian
3. *Cut* - Kualitas dari potongan
4. *Color* - Warna dari berlian
5. *Clarity* - Pengukuran kejernihan dari suatu berlian
6. *X length* - Panjang dalam mm
7. *Y width* - Lebar dalam mm
8. *Z depth* - Kedalaman dalam mm
9. *Depth* - Total persentase kedalaman ($\frac{z}{mean(x,y)} = \frac{2*z}{(x+y)}$)
10. *Table* - Lebar dari atas berlian relatif terhadap titik terlebar

BAB 3
ANALISIS DAN HASIL

3.1 *Tools and Libraries*

R adalah bahasa dan lingkungan untuk komputasi statistik dan grafik. Ini adalah proyek GNU yang mirip dengan bahasa dan lingkungan S yang dikembangkan di Bell Laboratories (sebelumnya AT&T, sekarang Lucent Technologies) oleh John Chambers dan rekan-rekannya. R dapat dianggap sebagai implementasi yang berbeda dari S. Ada beberapa perbedaan penting, tetapi banyak kode yang ditulis untuk S berjalan tidak berubah di bawah R.

R menyediakan berbagai macam statistik (pemodelan linier dan nonlinier, uji statistik klasik, analisis deret waktu, klasifikasi, pengelompokan) dan teknik grafis, dan sangat dapat dikembangkan. Bahasa S sering menjadi kendaraan pilihan untuk penelitian dalam metodologi statistik, dan R menyediakan rute Open Source untuk berpartisipasi dalam aktivitas tersebut.

Dalam pengerjaan ini, peneliti menggunakan *libraries* yang tersedia dalam bahasa pemrograman R. Adapun *libraries* yang digunakan dalam penelitian ini adalah sebagai berikut :

1. Psych : Fungsi utamanya untuk analisis multivariat dan konstruksi skala menggunakan analisis faktor, analisis komponen utama, analisis klaster, dan analisis keandalan, meskipun yang lain menyediakan statistik deskriptif dasar
2. Corrplot : Corrplot menyediakan alat eksplorasi visual pada matriks korelasi yang mendukung penataan ulang variabel otomatis untuk membantu mendeteksi pola tersembunyi di antara variabel
3. Ggplot2 : Paket visualisasi data sumber terbuka untuk bahasa pemrograman statistik R
4. Tidyverse : Data manipulation
5. Cluster : Clustering algorithms
6. Factoextra : Clustering visualization

3.2 *Data Cleaning*

Data cleaning adalah suatu prosedur untuk memastikan kebenaran, konsistensi, dan kegunaan suatu data yang ada dalam dataset. Pada penelitian ini, peneliti menggunakan beberapa metode untuk membuat data semakin bagus dan konsisten. Beberapa metode data cleaning yang dipakai adalah :

1. Konversi data menjadi numerik :
   Pada tahap ini, terdapat beberapa *variabel* yang masih dalam bentuk kategori. Maka dari itu, peneliti menggunakan metode konversi variabel agar semua variabel menjadi numerical. Adapun kategori yang di *convert* adalah sebagai berikut :
   1. *Cut* :
      a. Ideal = 5
      b. Premium = 4
      c. Very Good = 3
      d. Good = 2
      e. Fair = 1
   2. *Color* :

a. J = 1
b. I = 2
c. H = 3
d. G = 4
e. F = 5
f. E = 6
g. D = 7

3. *Clarity :*
   a. I1 = 1
   b. SI2 = 2
   c. SI1 = 3
   d. VS2 = 4
   e. VS1 = 5
   f. VVS2 = 6
   g. VVS1 = 7
   h. IF1 = 8

```
diamonds2$cut2[diamonds2$cut == "Ideal"] <- 5
diamonds2$cut2[diamonds2$cut == "Premium"] <- 4
diamonds2$cut2[diamonds2$cut == "Very Good"] <- 3
diamonds2$cut2[diamonds2$cut == "Good"] <- 2
diamonds2$cut2[diamonds2$cut == "Fair"] <- 1

diamonds2$color2[diamonds2$color =="J" ] <-1
diamonds2$color2[diamonds2$color =="I" ] <-2
diamonds2$color2[diamonds2$color =="H" ] <-3
diamonds2$color2[diamonds2$color =="G" ] <-4
diamonds2$color2[diamonds2$color =="F" ] <-5
diamonds2$color2[diamonds2$color =="E" ] <-6
diamonds2$color2[diamonds2$color =="D" ] <-7

diamonds2$clarity2[diamonds2$clarity=="I1"] <-1
diamonds2$clarity2[diamonds2$clarity=="SI2"] <-2
diamonds2$clarity2[diamonds2$clarity=="SI1"] <-3
diamonds2$clarity2[diamonds2$clarity=="VS2"] <-4
diamonds2$clarity2[diamonds2$clarity=="VS1"] <-5
diamonds2$clarity2[diamonds2$clarity=="VVS2"] <-6
diamonds2$clarity2[diamonds2$clarity=="VVS1"] <-7
diamonds2$clarity2[diamonds2$clarity=="IF1"] <-8
```

Konversi kategorikal ke numerical

2. Mengambil nilai yang tidak memiliki nilai Null :
   Pada tahap ini, peneliti menyadari bahwa terdapat nilai-nilai yang memiliki nilai 0. Peneliti menyadari bahwa hal tersebut merupakan kesalahan dan karena peneliti memiliki cukup banyak data, peneliti tidak mengikutsertakan baris-baris yang memiliki nilai 0. Hasil akhir dari metode ini mereduksi baris dari 17.000 menjadi 16.416 baris.

```
numeric_predictors=c('carat',"cut2","color2","depth","table","price","x","y","z","clarity2")
df <- diamonds2[,numeric_predictors]
df <-  subset(df, x!=0 & y!=0 & z!=0 & carat!=0  & price!=0  & clarity2!=0)
dim(df)
```

```
[1] 16416    10
```

Hasil dimensi

3. Mengeliminasi *Outliers* menggunakan metode jangkauan interkuartil :
   Pada tahap ini, peneliti menggunakan metode jangkauan interkuartil untuk mengeliminasi *outliers* yang ada. Metode jangkauan interkuartil merupakan suatu metode yang cukup umum untuk mendeteksi outliers dan mengeliminasikannya. Adapun tahap-tahap yang dilakukan adalah menghitung IQR untuk setiap variabel, lalu menghitung *lower threshold,* dan juga menghitung *upper threshold*. Setelah itu, untuk setiap variabel akan dilakukan pemilihan sesuai dengan threshold yang ada. Hal tersebut dilakukan sebanyak 15 kali agar benar-benar tidak terdapat outliers dalam dataset tersebut. Hasil akhir dari pengeliminasian outliers ini mereduksi baris dari 16.416 menjadi 14.772 baris.

```
[1] 14308    10
[1] 13920    10
[1] 13858    10
[1] 13840    10
[1] 13773    10
[1] 13670    10
[1] 13670    10
[1] 13670    10
[1] 13670    10
[1] 13670    10
```

Hasil Dimensi saat melakukan eliminasi outliers menggunakan IQR

4. Memilih variabel berdasarkan *Kaiser-Meyer-Olkin* (KMO)
   Pada tahap ini, peneliti menggunakan Metode KMO untuk meneliti variabel mana saja yang layak dipilih untuk melakukan analisis faktor. Dari hasil kalkulasi dengan threshold > 0.5, dapat disimpulkan bahwa variabel-variabel yang tidak diikutsertakan adalah cut2 (kategori unacceptable), color2 (kategori unacceptable) , depth (kategori unacceptable),

dan table (kategori miserable). Hasil akhir dari metode KMO adalah peneliti mendapatkan 6 variabel penting yaitu carat, price, x, y ,z , dan clarity2.

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor(df))
Overall MSA =  0.73
MSA for each item =
   carat     cut2   color2    depth    table    price        x        y        z clarity2
    0.87     0.49     0.39     0.09     0.50     0.80     0.84     0.82     0.72     0.54
```

5. Scaling

   Pada tahap ini, peneliti melakukan scaling menggunakan salah satu function dari R programming yaitu scale(). Tahap ini dilakukan agar range dari setiap data itu memiliki range yang sama sehingga tidak menimbulkan bias ketika melakukan clustering.

```{r}
df2<-df
df <- na.omit(df)
df <- scale(df)
head(df)
```

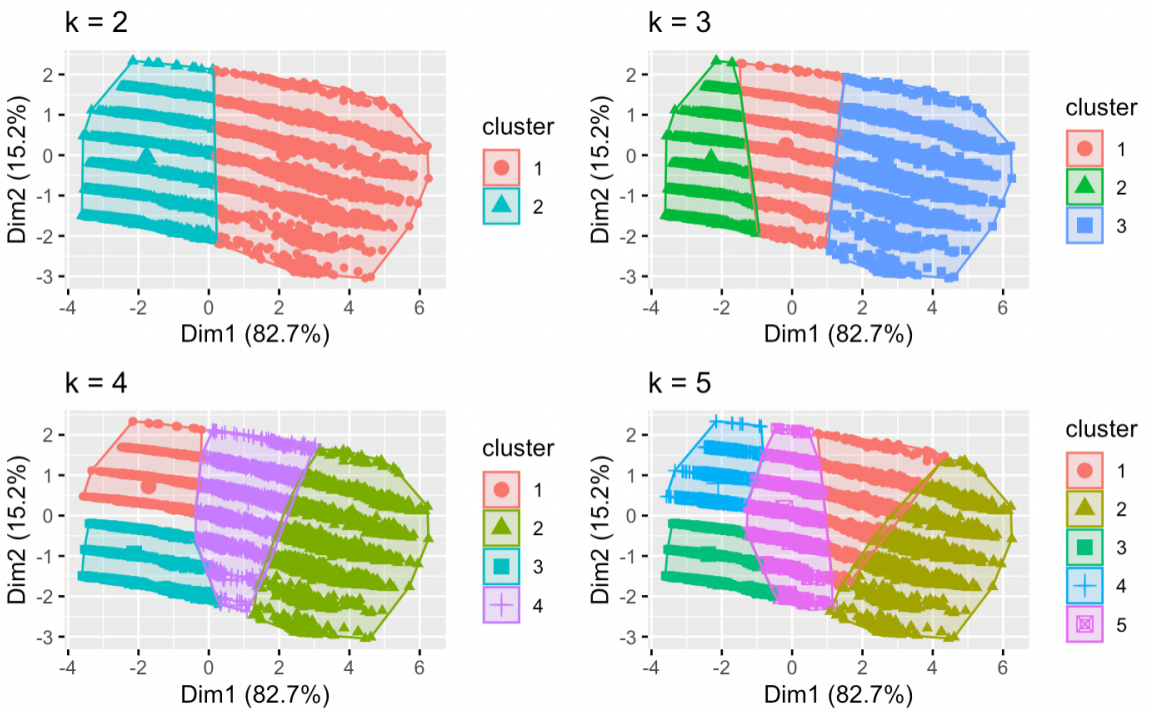
```
          carat      price          x          y          z   clarity2
[1,] -1.1444002 -1.0009130 -1.3452860 -1.3262388 -1.26872801  0.0491915
[2,] -1.0867787 -0.9436282 -1.1531721 -1.1326557 -1.21704041  0.7183047
[3,] -0.0784026 -0.4212241  0.2236445  0.2654443 -0.02822547  1.3874178
[4,] -0.8274820 -0.7491160 -0.7902902 -0.8100172 -0.71739355  2.0565310
[5,]  0.5842445  0.7321687  0.6185454  0.6633651  0.78154702  0.0491915
[6,] -1.1155894 -1.0398155 -1.1958641 -1.1756741 -1.30318642 -1.2890348
```

Hasil dari scaling

3.3 *Clustering*

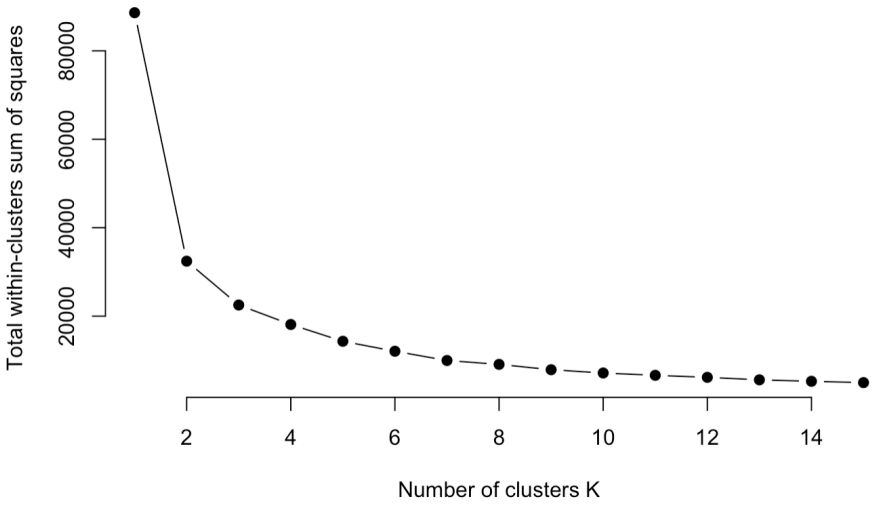
3.3.1 Data *Clustering* menggunakan *K-Means*

Pada tahap ini, peneliti memasukan data kepada algoritma K-Means untuk menghasilkan clustering terhadap data, yang berguna untuk membantu melakukan segmentasi pada data. Melalui gambar yang ada di bawah, secara kasar peneliti dapat menyimpulkan bahwa K-Means ini dapat menghasilkan *clustering* yang cukup baik. Hasil yang cukup baik ini dapat dilihat dari tidak ada tumpang tindihnya antara cluster. Dari Informasi di bawah juga dapat dilihat bahwa Dim1 memiliki nilai 82.7% dan Dim2 memiliki nilai sebesar 15.2%. Hal ini menunjukan bahwa walaupun direduksi ke dalam 2 dimensi, informasi yang dapat diberikan memiliki nilai sebesar 97.9% dari dimensi original, yaitu 6 dimensi.

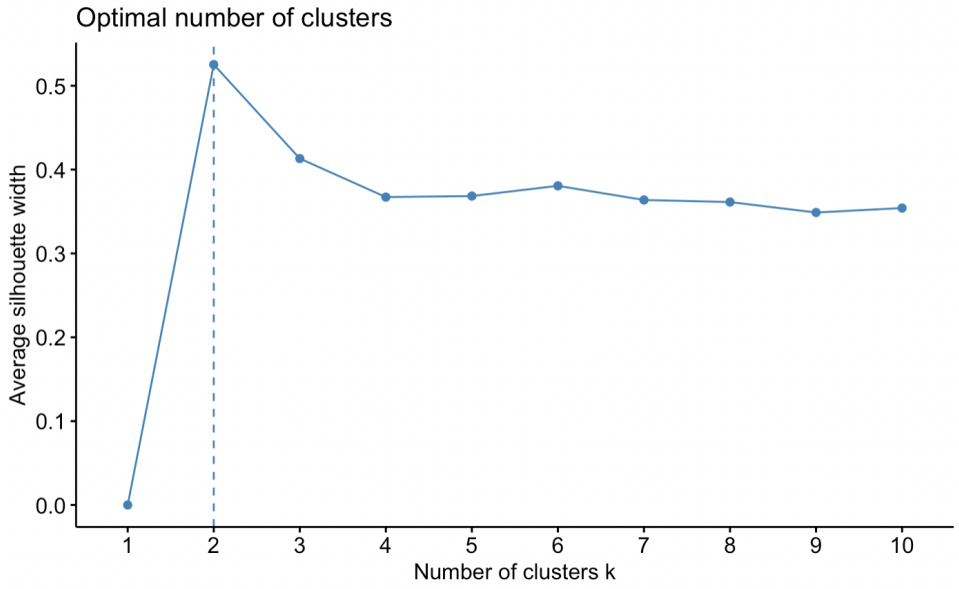


Hasil Clustering

Peneliti menggunakan metode *Elbow Method* untuk mencari berapa banyak *clusters* yang perlu dipakai sehingga dapat menghasilkan clustering yang maksimal dan juga tidak *over fitting*. Menggunakan metode siku, peneliti dapat memprediksi bahwa dengan menggunakan cluster sebesar 3, akan menghasilkan cluster yang cukup dan pas sehingga tidak terjadinya *over fitting*.



Total within-cluster sum of squares



Average Silhouette width

Berdasarkan hasil cluster yang diperoleh, dapat dilihat bahwa dengan menggunakan 3 cluster, kita memiliki 4853 data yang dikategorikan ke cluster 1, 5301 data yang dikategorikan cluster 2, 4154 data yang dikategorikan cluster 3. Dari informasi dibawah kita juga dapat melihat bahwa *Within cluster sum of squares by cluster (WCSS)* memiliki nilai sebesar 73.9%. Angka ini menunjukan kepadatan tiap cluster dan juga menunjukan seberapa mirip data-data point di tiap clusternya.

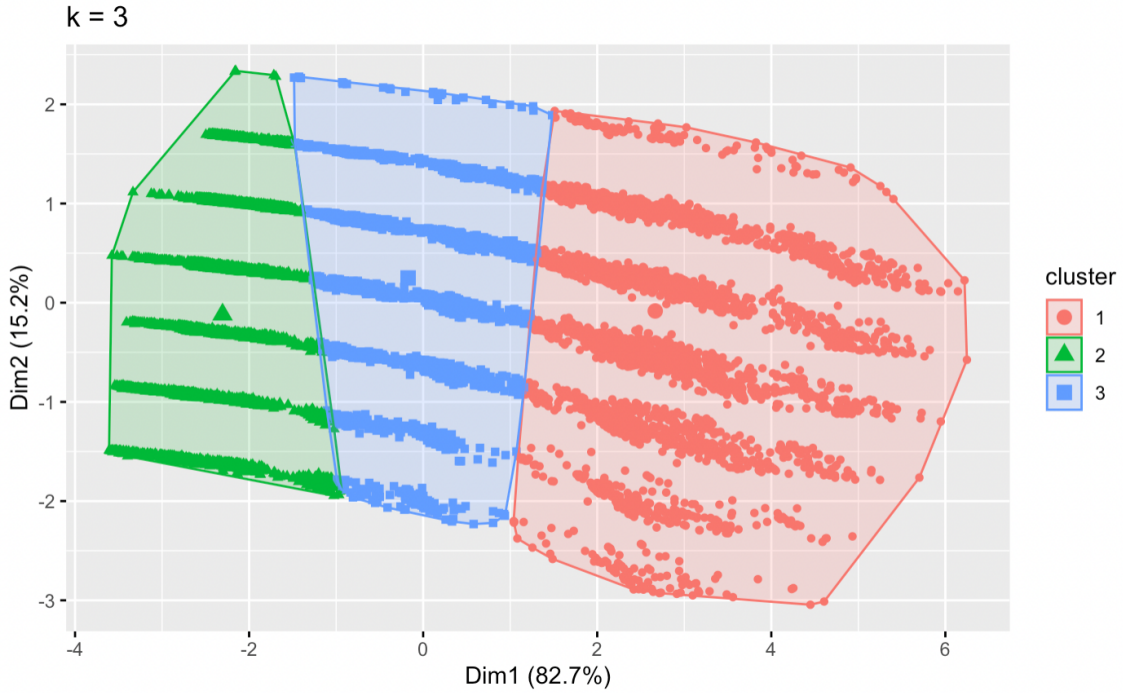
```
K-means clustering with 3 clusters of sizes 4853, 5301, 4154

Cluster means:
       carat      price            x            y            z   clarity2
1  1.2041498  1.1838429  1.170622337  1.169021185  1.168675206 -0.4079082
2 -0.9810666 -0.8482793 -1.065496972 -1.066499003 -1.064605898  0.5033508
3 -0.1548159 -0.3005442 -0.007903407 -0.004754116 -0.006765747 -0.1657882

Clustering vector:
   [1] 2 2 2 1 2 1 1 3 2 3 1 1 2 1 1 1 1 1 2 2 2 2 3 1 3 3 2 1 2 1 1 2 2 3 1 1 3 3 3 2 3 2 1 3 1 2 2 1 3 1 2
  [81] 1 1 2 3 1 3 3 2 1 2 1 1 2 3 3 1 2 2 3 1 1 2 1 2 1 2 3 2 2 2 2 3 1 2 1 2 3 2 2 1 2 2 3 1 1 3 1 1 1 1 1
 [161] 2 1 3 3 1 2 1 1 2 3 1 1 3 2 1 3 1 1 2 2 2 1 2 3 2 2 2 3 2 3 2 3 3 2 2 3 1 3 2 2 2 2 3 2 1 2 2 3 2 1 1
 [241] 1 2 2 1 1 1 2 1 1 3 3 2 1 2 3 3 2 2 2 2 1 2 1 1 2 1 2 3 3 2 2 3 1 2 1 3 1 1 2 3 3 3 3 1 1 2 2 3 2 1 2
 [321] 2 1 1 1 1 3 1 2 2 1 3 3 2 3 3 1 1 1 1 2 2 3 2 2 1 3 3 3 1 2 1 3 3 1 2 3 2 1 2 1 2 1 2 2 1 1 2 1 2 3 3
 [401] 1 1 1 2 3 2 2 3 2 1 3 3 3 2 1 1 2 2 1 3 3 2 2 2 1 2 3 1 2 1 3 1 2 2 3 3 3 1 1 1 3 3 1 2 1 1 2 3 2 3 3
 [481] 3 1 3 2 2 2 2 1 1 3 1 1 3 3 3 1 2 2 2 3 2 3 1 2 2 2 3 2 2 2 1 3 3 1 1 3 2 2 3 3 2 1 1 3 2 3 3 1 3 2 2
 [561] 3 2 2 2 3 2 2 2 1 2 1 1 2 1 2 2 2 3 2 1 1 2 2 3 2 1 2 1 2 2 3 1 2 3 3 2 3 3 3 1 3 2 1 2 1 3 3 1 2 3 1
 [641] 1 1 2 1 2 2 3 1 3 3 3 3 1 2 1 2 2 1 1 3 3 2 1 3 2 2 1 1 1 1 2 2 2 3 1 3 1 2 1 3 2 2 2 1 3 1 3 2 3 2 1
 [721] 3 2 3 1 2 1 1 3 2 1 2 1 1 1 1 2 3 2 2 1 3 3 2 2 3 3 2 3 2 1 2 2 1 1 1 2 3 1 3 3 1 3 3 1 2 2 1 1 1 2 2
 [801] 2 3 1 2 1 2 2 3 2 2 3 2 2 3 2 3 3 1 2 2 2 1 3 2 2 3 1 2 2 1 2 2 2 3 2 3 1 2 1 2 1 2 1 3 2 1 2 1 2 2
 [881] 2 1 1 1 1 3 3 3 1 1 3 2 2 1 2 1 1 1 3 1 1 2 1 3 2 2 3 3 2 3 3 1 1 3 2 2 1 2 3 1 2 1 3 3 1 2 2 3 3 1 2
 [961] 3 2 2 2 1 1 2 1 1 3 2 1 3 3 3 2 1 3 2 1 2 2 1 3 1 3 1 1 3 1 3 3 2 2 1 2 2 3 1 1
 [ reached getOption("max.print") -- omitted 13308 entries ]

Within cluster sum of squares by cluster:
[1] 10323.013  6772.443  5285.279
 (between_SS / total_SS =  73.9 %)
```

Hasil Summary dari 3 Cluster menggunakan K-Means



Visualisasi Cluster menggunakan k = 3

3.3.2 Data Clustering mengkombinasikan PCA dengan K-Means

Di tahap ini, peneliti melakukan proses *PCA* sebelum menghasilkan *K-Means Clustering*. Tujuan peneliti melakukan proses tersebut adalah untuk mendapatkan hasil clustering yang lebih akurat, karena *PCA* dipercaya dapat melakukan *noise reduction* pada data yang akan dipakai sebelum dilakukannya *clustering*. Dikutip dari penelitian yang dilakukan oleh Ding dan He (2004), bahwa mereka mendukung argumen mengenai peran *PCA* dalam melakukan *K-Means Clustering*.

Setelah dilakukan *PCA* terhadap data, *summary PCA* dapat dilihat dari gambar di bawah ini. Hasil tersebut menyatakan bahwa *principal component* yang pertama dapat merepresentasikan 83.09% total varians dari dataset. Kemudian 14.79% dari *principal component* kedua, 1.61% dari *principal component* ketiga, dan seterusnya. Oleh karena itu, peneliti mengambil 2 *principal component* pertama dengan *cumulative proportion* 97.88% untuk merepresentasikan seluruh dataset.

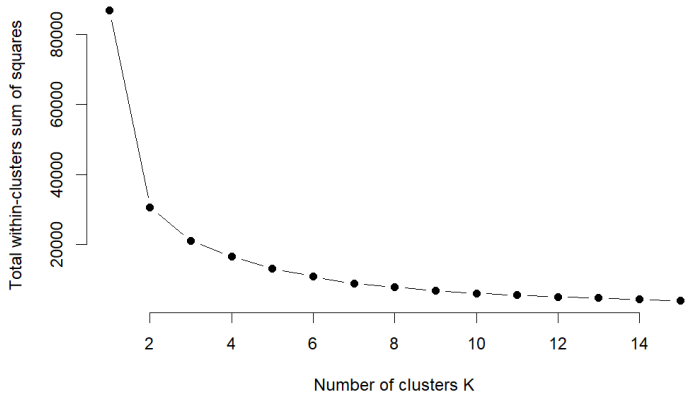
```
Importance of components:
                          Comp.1    Comp.2     Comp.3      Comp.4      Comp.5       Comp.6
Standard deviation     2.2339032 0.9449722 0.30647789 0.119280954 0.085056284 0.0362330092
Proportion of Variance 0.8317206 0.1488288 0.01565478 0.002371324 0.001205762 0.0002188052
Cumulative Proportion  0.8317206 0.9805493 0.99620411 0.998575433 0.999781195 1.0000000000

Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
carat     0.445                0.884  0.125
price     0.417  0.270 -0.850 -0.176
x         0.445         0.280 -0.269  0.367 -0.718
y         0.445         0.280 -0.316  0.372  0.696
z         0.445         0.276 -0.120 -0.843
clarity2 -0.186  0.960  0.207
```

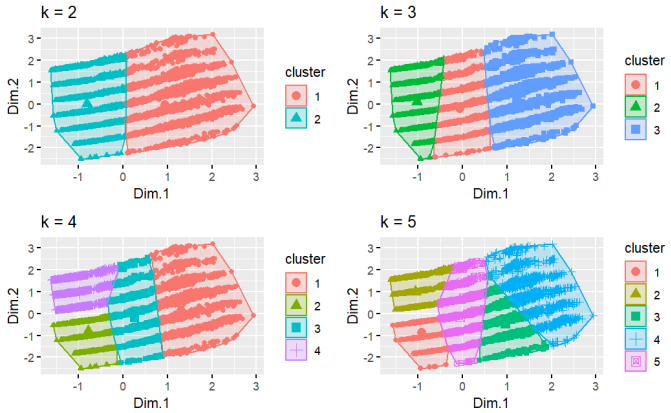
Dua *principal component* pertama ini (*Comp.1* dan *Comp.2*) akan peneliti gunakan untuk dilakukannya K-Means Clustering, dengan tujuan agar nilai *WCSS* dari clustering lebih baik dari sebelum dilakukannya *PCA.* Data yang dipakai untuk clustering diperoleh dari koordinat individual kedua *principal component* pertama seperti yang dapat dilihat pada gambar di bawah ini.

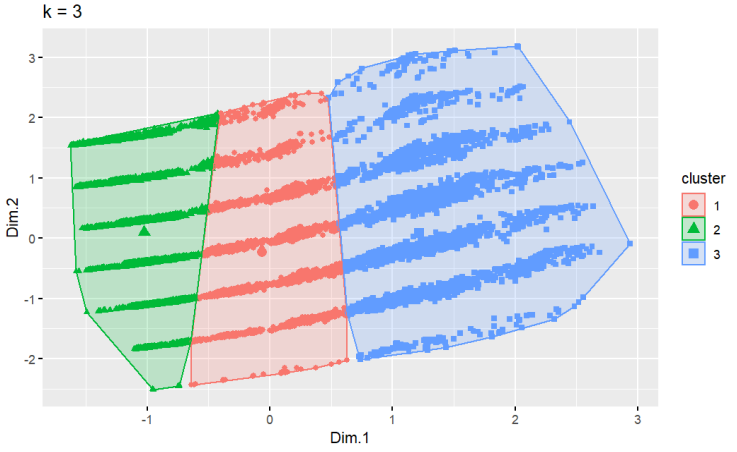
	Dim.1	Dim.2
1	-2.628056	-0.4310554
2	-2.509400	0.2470077
3	-2.037113	1.6400532
4	1.525317	0.3166233
5	-2.262836	-1.7198388
6	2.194902	0.4828619

Sebelum melakukan *K-Means Clustering*, peneliti menggunakan *Elbow Method* untuk menentukan jumlah *cluster* yang maksimal agar tidak terjadinya *overfitting*. Hasil dari *Elbow Method* dapat dilihat dari gambar di bawah ini. Dari gambar tersebut, peneliti dapat memprediksi bahwa dengan menggunakan 3 *clusters*, akan menghasilkan *clusters* yang cukup dan tidak menyebabkan *overfitting.*



Pada gambar di bawah ini, dapat dilihat *clustering* yang dilakukan bagus dan tidak bertumpang-tindih. Dan berdasarkan *Elbow Method* yang dilakukan sebelumnya, peneliti akan menggunakan *data clustering* dengan k = 3.





Dapat dilihat bahwa dengan menggunakan 3 *cluster*, kita memiliki 3947 data yang dikategorikan ke *cluster* 1, 4544 data yang dikategorikan *cluster* 2, 5179 data yang dikategorikan *cluster* 3. Dari informasi dibawah kita juga dapat melihat bahwa *Within cluster sum of squares by cluster (WCSS)* memiliki nilai sebesar 76%. Dan ini merupakan hasil yang lebih baik dibandingkan dengan *K-Means Cluster* sebelum dilakukannya *PCA*.

```
K-means clustering with 3 clusters of sizes 3947, 4544, 5179

Cluster means:
      Dim.1       Dim.2
1 -0.137297 -0.21497674
2  2.725353  0.07872291
3 -2.286559  0.09476661

Clustering vector:
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21   22   23   24   25   26
   3    3    3    2    3    2    1    3    1    2    2    3    2    2    2    2    3    3    3    3    1    2    1    1    3    3
  27   28   29   30   31   32   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48   49   50   51   52
   2    3    3    1    2    2    1    1    2    3    1    3    2    1    2    3    3    2    1    2    3    1    3    3    3    2
  53   54   55   56   57   58   59   60   61   62   63   64   65   66   67   68   69   70   71   72   73   74   75   76   77   78
   3    1    3    1    3    2    2    2    2    1    1    3    2    3    2    2    3    2    2    3    1    2    2    3    2    2
  79   80   81   82   83   84   85   86   87   88   89   90   91   92   93   94   95   96   97   98   99  100  101  102  103  104
   3    1    1    1    3    2    3    2    2    3    1    1    2    3    3    1    2    2    3    2    3    3    1    3    3    3
 105  106  107  108  109  110  111  112  113  114  115  116  117  118  119  120  121  122  123  124  125  126  127  128  129  130
   1    2    3    2    3    1    3    3    2    3    3    1    2    2    1    2    2    2    2    2    2    1    1    3    3    2
 131  132  133  134  135  136  137  138  139  140  141  142  143  144  145  146  147  148  149  150  151  152  153  154  155  156
   3    2    1    3    3    3    2    2    2    3    3    3    3    2    1    3    2    3    1    1    1    3    2    1    1    2
 157  158  159  160  161  162  163  164  165  166  167  168  169  170  171  172  173  174  175  176  177  178  179  180  181  182
   3    2    2    3    2    2    1    3    2    2    2    2    3    3    3    2    3    1    3    3    3    1    3    1    3    1
 183  184  185  186  187  188  189  190  191  192  193  194  195  196  197  198  199  200  201  202  203  204  205  206  207  208
 833  834  835  836  837  838  839  840  841  842  843  844  845  846  847  848  849  850  851  852  853  854  855  856  857  858
   2    2    3    3    3    3    3    3    2    1    2    3    1    3    1    2    3    2    2    2    1    2    1    3    3    3
 859  860  861  862  863  864  865  866  867  868  869  870  871  872  873  874  875  876  877  878  879  880  881  882  883  884
   2    2    1    2    3    2    1    3    3    1    1    3    1    1    2    2    1    3    3    2    3    1    2    3    2    1
 885  886  887  888  889  890  891  892  893  894  895  896  897  898  899  900  901  902  903  904  905  906  907  908  909  910
   1    3    3    1    1    2    3    1    2    2    1    1    2    1    2    2    2    3    1    2    1    1    1    3    2    2
 911  912  913  914  915  916  917  918  919  920  921  922  923  924  925  926  927  928  929  930  931  932  933  934  935  936
   3    3    3    3    3    3    3    3    3    1    3    3    3    2    2    3    2    2    1    3    2    1    1    1    3    2
 937  938  939  940  941  942  943  944  945  946  947  948  949  950  951  952  953  954  955  956  957  958  959  960  961  962
   1    3    2    3    3    2    1    2    1    2    2    2    2    1    1    3    3    2    3    3    1    2    2    3    2    3
 963  964  965  966  967  968  969  970  971  972  973  974  975  976  977  978  979  980  981  982  983  984  985  986  987  988
   2    1    3    2    2    3    1    1    1    1    3    2    2    1    1    2    1    1    2    1    2    1    3    3    2    3
 989  990  991  992  993  994  995  996  997  998  999 1000
   3    1    3    1    2    2    3    2    1    2    1    2
 [ reached getOption("max.print") -- omitted 12670 entries ]

Within cluster sum of squares by cluster:
[1] 4850.731 8056.542 6357.480
 (between_SS / total_SS =  76.0 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
```

3.3.3 Analisis Data Berdasarkan Hasil Cluster PCA

Untuk melakukan analisis berdasarkan *Cluster PCA*, peneliti menambahkan (*append*) kolom hasil *clustering* pada data sebelum *scaling*. Peneliti juga menambahkan kolom volume, yaitu hasil perkalian dari variabel *x, y, dan z*, untuk mempermudah analisis dan visualisasi dari data.

carat <dbl>	price <int>	x <dbl>	y <dbl>	z <dbl>	clarity2 <dbl>	cluster <int>	cluster_pca <int>	volume <dbl>
0.30	526	4.27	4.30	2.68	4	3	3	49.20748
0.32	660	4.45	4.48	2.71	5	3	3	54.02656
0.41	1115	4.79	4.78	3.00	7	3	3	68.68860
0.90	4580	6.11	6.15	3.87	4	2	2	145.42106
0.31	435	4.41	4.44	2.66	2	3	3	52.08386
1.01	5679	6.37	6.40	3.99	4	2	2	162.66432
0.52	1822	5.13	5.16	3.20	4	1	1	84.70656
0.33	1006	4.46	4.48	2.74	5	3	3	54.74739
0.64	2419	5.51	5.53	3.44	4	1	1	104.81783
1.21	4637	6.68	6.63	4.22	3	2	2	186.89705

Setelah itu, peneliti melakukan analisis pada data yang sudah dikelompokkan berdasarkan *cluster* dari hasil *PCA*: pada *cluster* 1, nilai min dari carat adalah 0.41, nilai max dari carat 0.92, dan nilai mean dari carat tersebut 0.64. Nilai min dari price adalah 452, nilai max dari price 5539, dan nilai mean dari price tersebut 2204.32. Nilai min dari clarity adalah 1, nilai max dari clarity 7, dan nilai mean dari clarity tersebut 3.79. Nilai min dari volume adalah 67.06, nilai max dari volume 149.6 dan nilai mean dari volume tersebut 104.62.

Untuk *cluster* 2, nilai min dari carat adalah 0.84, nilai max dari carat 1.94, dan nilai mean dari carat tersebut 1.12. Nilai min dari price adalah 1901, nilai max dari price 9712, dan nilai mean dari price tersebut 5733.21. Nilai min dari clarity adalah 1, nilai max dari clarity 7, dan nilai mean dari clarity tersebut 3.3. Nilai min dari volume adalah 137.16, nilai max dari volume 317.26 dan nilai mean dari volume tersebut 180.93.

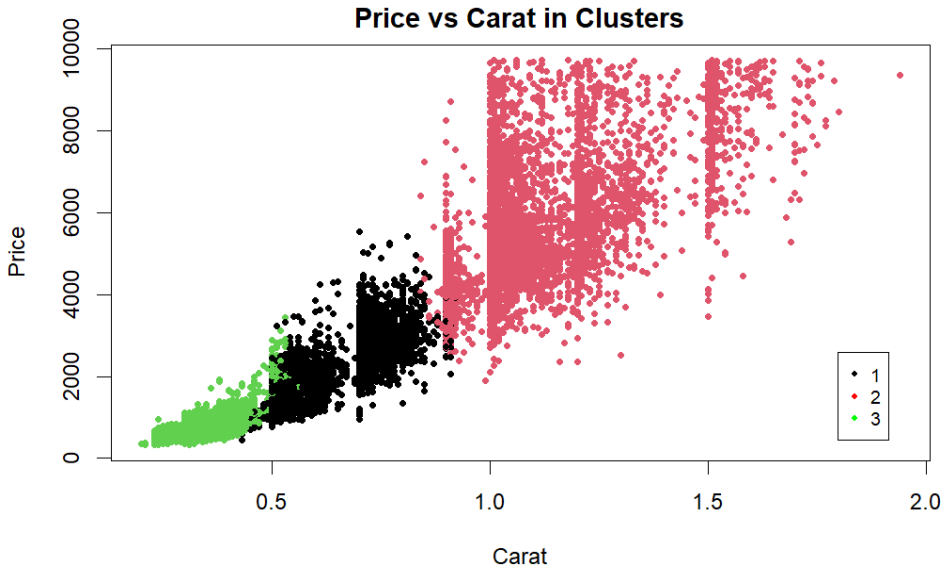
Dan untuk *cluster* 3, nilai min dari carat adalah 0.2, nilai max dari carat 0.56, dan nilai mean dari carat tersebut 0.35. Nilai min dari price adalah 326, nilai max dari price 3442, dan nilai mean dari price tersebut 815.93. Nilai min dari clarity adalah 1, nilai max dari clarity 7, dan nilai mean dari clarity tersebut 4.7. Nilai min dari volume adalah 33.41, nilai max dari volume 91.37 dan nilai mean dari volume tersebut 57.44.

cluster_pca <int>	mean_carat <dbl>	mean_price <dbl>	mean_clarity2 <dbl>	mean_volume <dbl>
1	0.6399501	2204.321	3.792010	104.62007
2	1.1154694	5733.218	3.303512	180.92613
3	0.3496030	815.936	4.700467	57.44045

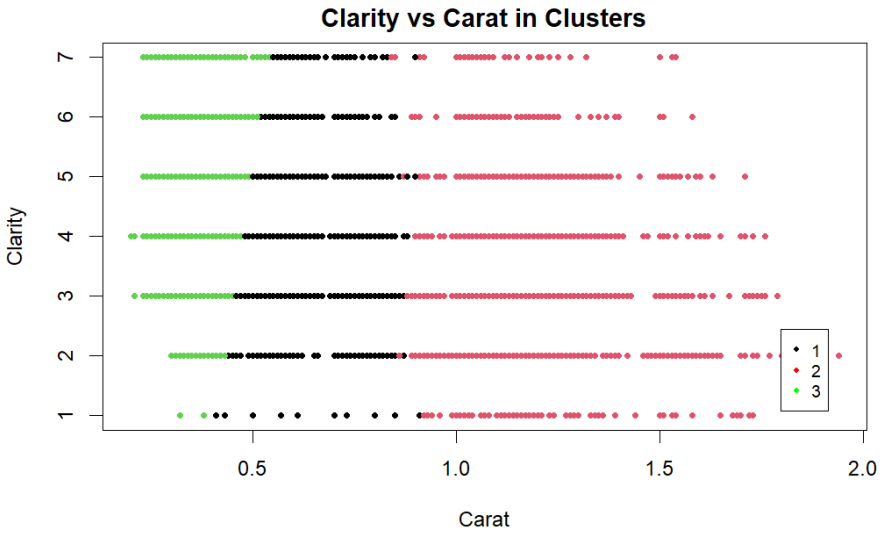
cluster_pca <int>	min_carat <dbl>	min_price <int>	min_clarity2 <dbl>	min_volume <dbl>
1	0.41	452	1	67.05678
2	0.84	1901	1	137.16480
3	0.20	326	1	33.41218

cluster_pca <int>	max_carat <dbl>	max_price <int>	max_clarity2 <dbl>	max_volume <dbl>
1	0.92	5539	7	149.60784
2	1.94	9712	7	317.25848
3	0.56	3442	7	91.37153

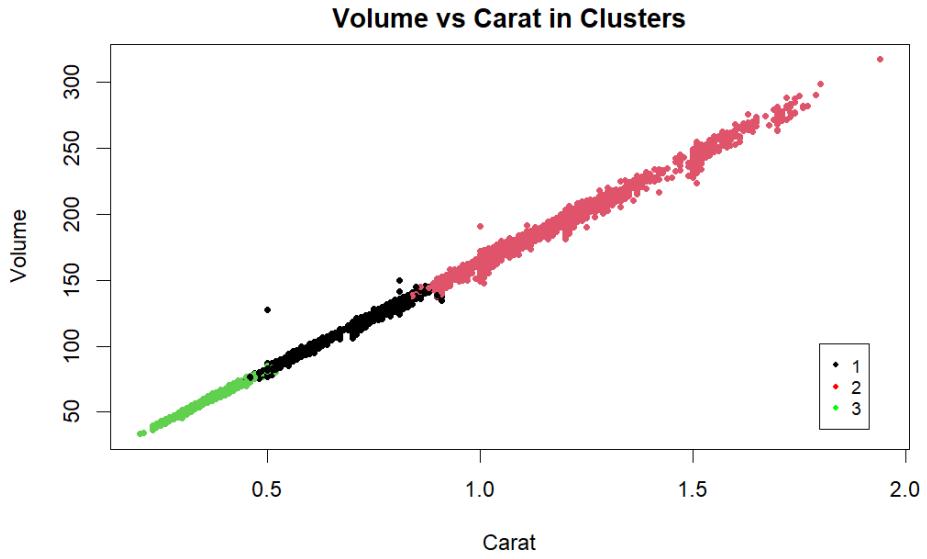
Pada gambar di bawah ini, peneliti menampilkan *clustering* dari data Price dan Carat. *Clustering* yang peneliti lakukan dapat membagi data dengan baik karena tidak banyak warna titik yang bertumpang-tindih antara *clusters*. Secara kasat mata, nilai 0 sampai 0.5 termasuk dalam *cluster* 3, nilai 0.5 sampai 0.75 termasuk dalam *cluster 1*, dan nilai 0.75 sampai 2 termasuk dalam *cluster* 2.



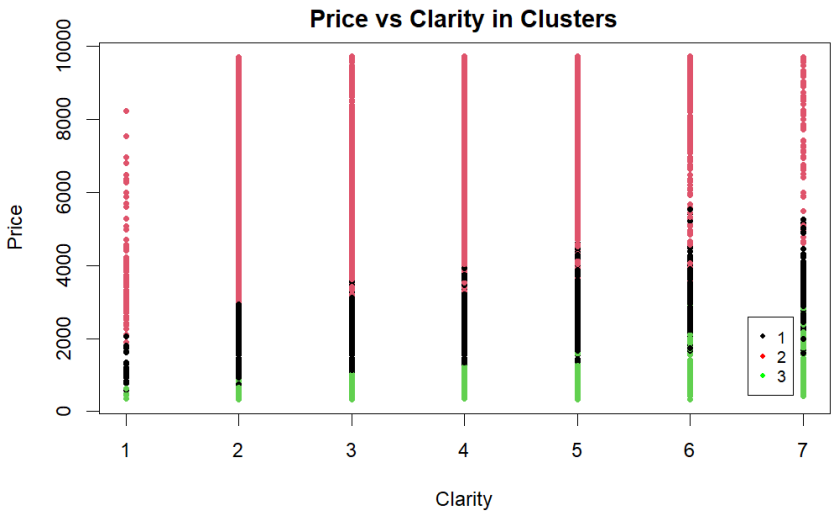
Pada gambar di bawah ini, peneliti menampilkan *clustering* dari data Clarity dan Carat. *Clustering* yang peneliti lakukan dapat membagi data dengan baik karena tidak banyak warna titik yang bertumpang-tindih antara *clusters*. Secara kasat mata, nilai 0 sampai 0.5 dari carat termasuk dalam *cluster* 3, nilai 0.5 sampai 0.75 dari carat termasuk dalam *cluster 1*, dan nilai 0.75 sampai 2 dari carat termasuk dalam *cluster* 2.



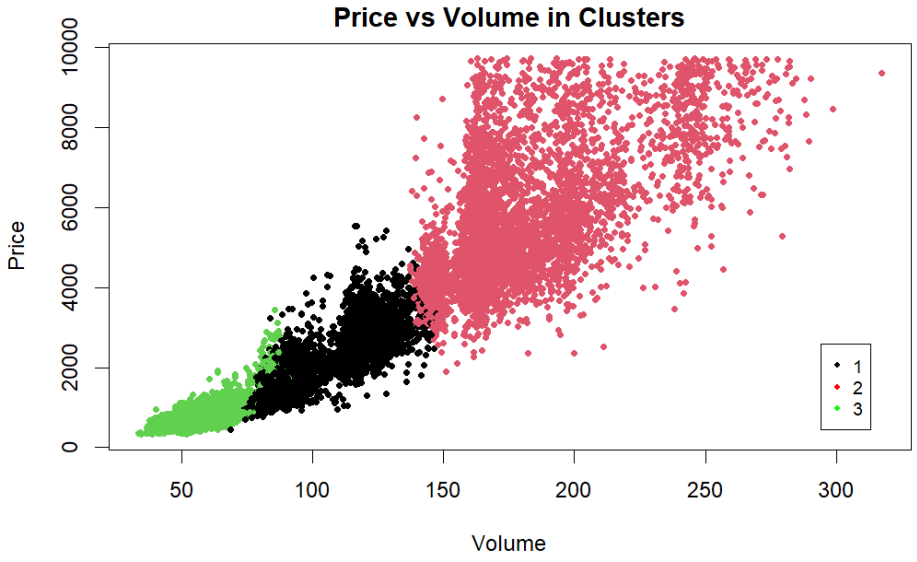
Pada gambar di bawah ini, peneliti menampilkan *clustering* dari data Volume dan Carat. *Clustering* yang peneliti lakukan dapat membagi data dengan baik karena tidak banyak warna titik yang bertumpang-tindih antara *clusters*. Secara kasat mata, nilai 0 sampai 0.5 dari carat termasuk dalam *cluster* 3, nilai 0.5 sampai 0.75 dari carat termasuk dalam *cluster 1*, dan nilai 0.75 sampai 2 dari carat termasuk dalam *cluster* 2.



Pada gambar di bawah ini, peneliti menampilkan *clustering* dari data Price dan Clarity. *Clustering* yang peneliti lakukan dapat membagi data dengan baik karena tidak banyak warna titik yang bertumpang-tindih antara *clusters*. Setiap nilai clarity memiliki segmentasi cluster nilai price yang berbeda-beda. Namun secara kasat mata, dapat diperkirakan bahwa nilai 0 sampai 2000 dari price termasuk dalam *cluster* 3 tergantung pada nilai claritynya, nilai 1500 sampai 5000 dari price termasuk dalam *cluster 1* tergantung pada nilai claritynya, dan nilai 2000 sampai 10000 dari price termasuk dalam *cluster* 2 tergantung pada nilai claritynya.



Pada gambar di bawah ini, peneliti menampilkan clustering dari data Price dan Volume. *Clustering* yang peneliti lakukan dapat membagi data dengan baik karena tidak banyak warna titik yang bertumpang-tindih antara *clusters*. Secara kasat mata, nilai 0 sampai 75 dari volume termasuk dalam *cluster* 3, nilai 75 sampai 150 dari volume termasuk dalam *cluster 1*, dan nilai 150 sampai 300 dari volume termasuk dalam *cluster* 2.



Pada gambar di bawah ini, peneliti menampilkan clustering dari data Volume dan Clarity. *Clustering* yang peneliti lakukan dapat membagi data dengan baik karena tidak banyak warna titik yang bertumpang-tindih antara *clusters*. Setiap nilai clarity memiliki segmentasi cluster nilai volume yang berbeda-beda. Namun secara kasat mata, dapat diperkirakan bahwa nilai 0 sampai 100 dari volume termasuk dalam *cluster* 3 tergantung pada nilai claritynya, nilai 75 sampai 150 dari volume termasuk dalam *cluster 1* tergantung pada nilai claritynya, dan nilai 150 sampai 300 dari volume termasuk dalam *cluster* 2.


Volume vs Clarity in Clusters
Volume
50
100
150
200
250
300
1
2
3
4
5
6
7
Clarity
1
2
3

BAB 4
PENUTUPAN

4.1 Kesimpulan

Metode yang digunakan dalam penelitian kali ini adalah *K-Means Clustering*. Akan tetapi dalam analisis yang dilakukan, metode yang akan dipakai ini akan digabungkan dengan *PCA.* Tujuan penggabungan kedua metode ini adalah untuk membandingkan hasil analisis terbaik kepada data yang dilakukan *PCA* terlebih dahulu dengan data yang tidak dilakukan *PCA*. Penggunaan metode *PCA* ini ditujukan untuk menutup salah satu kelemahan *K-Means Clustering*, yaitu kalkulasi yang buruk akibat tingginya dimensi/variabel. Walaupun *PCA* mampu untuk mereduksi dimensi yang menyebabkan kalkulasi yang lebih baik, pertukaran yang dialami akibat penggunaan *PCA* ini adalah data yang digunakan tidak dapat dijelaskan secara total (100%) akibat data digambarkan dalam 2 dimensi.

Akan tetapi, dalam penelitian yang telah dilakukan, hasil *PCA* yang didapatkan bernilai 98.05%. Hal ini berarti, dari *PCA* yang dilakukan dapat menjelaskan keseluruhan data sebanyak 98.05% dan sudah cukup baik. Selanjutnya, dengan menggunakan *K-Means Clustering*, didapati bahwa terdapat 3 *cluster* untuk data ini, yang memiliki kemiripan sebesar 76% untuk setiap data di dalam *cluster* tersebut. Untuk lebih jelasnya dapat dilihat pada tabel berikut, ciri-ciri suatu berlian (*diamond*) dapat masuk dalam suatu *cluster* :

cluster_pca <int>	mean_carat <dbl>	mean_price <dbl>	mean_clarity2 <dbl>	mean_volume <dbl>
1	0.6399501	2204.321	3.792010	104.62007
2	1.1154694	5733.218	3.303512	180.92613
3	0.3496030	815.936	4.700467	57.44045

cluster_pca <int>	min_carat <dbl>	min_price <int>	min_clarity2 <dbl>	min_volume <dbl>
1	0.41	452	1	67.05678
2	0.84	1901	1	137.16480
3	0.20	326	1	33.41218

cluster_pca <int>	max_carat <dbl>	max_price <int>	max_clarity2 <dbl>	max_volume <dbl>
1	0.92	5539	7	149.60784
2	1.94	9712	7	317.25848
3	0.56	3442	7	91.37153

Dengan demikian, penelitian ini mampu menjawab pertanyaan bahwa dengan data penjualan berlian ini, variabel yang dapat digunakan untuk mengelompokan suatu berlian ke dalam 3 *cluster* adalah *carat*, *price*, *clarity*, dan *volume*. Disini dapat dilihat juga bahwa kecenderungan pembeli berlian ada pada *cluster* 3, lalu disusul dengan *cluster* 2 dan 1. Artinya bahwa penjualan terbanyak berlian ada pada rata-rata harga yang paling rendah; dan orang akan langsung membeli berlian dengan harga yang paling rendah (*cluster* 3) atau paling tinggi (*cluster* 2), dan lebih sedikit yang membeli di harga pertengahan (*cluster* 1). Sehingga dengan hasil ini, diharapkan penjualan berlian dapat dilihat berdasarkan pengelompokan *cluster*-nya, dan mampu melacak dimana pertumbuhan penjualan berlian setiap interval waktunya (mingguan, bulanan, ataupun tahunan) dengan melihat persebaran di dalam *cluster*-nya.

DAFTAR PUSTAKA

Budi, A. (2018). *Penggunaan PCA (Principal Component Analysis)*. Dikutip pada 27 Agustus 2022, dari https://www.academia.edu/42215462/Penggunaan_PCA_Principal_Component_Analysis_

Dabbura, I. (2018, September 17). *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks*. Dikutip pada 27 Agustus 2022, dari https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Ding, C., & He, X. (n.d.). *K-means Clustering via Principal Component Analysis*. Dikutip pada 27 Agustus 2022, dari https://ranger.uta.edu/~chqding/papers/KmeansPCA1.pdf

Hidayat, A. (n.d.). *Penjelasan Tentang Analisis Multivariat Dan Jenisnya*. Dikutip pada 27 Agustus 2022, dari https://www.statistikian.com/2016/11/analisis-multivariat.html

Olkin, & Sampson. (2002, November 2). *Multivariate Analysis: Overview*. Dikutip pada 27 Agustus 2022, dari https://www.sciencedirect.com/science/article/pii/B0080430767004721

Waluyo, M. (2007, Juli). TEKNIK ANALISIS DATA MULTIVARIAT DENGAN STRUCTURAL EQUATION MODELLING (SEM). *TEKMAPRO Teknik Industri FTI UPNV Jatim*, *2*(2), 124-139. https://core.ac.uk/download/pdf/12218323.pdf

Yahya, & Mahpuz. (2019, Juli). *Penggunaan Algoritma K-Means Untuk Menganalisis Pelanggan Potensial Pada Dealer SPS Motor Honda Lombok Timur Nusa Tenggara Barat*. Dikutip pada 27 Agustus 2022, dari https://e-journal.hamzanwadi.ac.id/index.php/infotek/article/download/1447/pdf