# Laporan Tugas 4

# Analisis Cluster

Analisis Multivariat Kelas B

Kelompok L


| | |
|---|---|
| Evan Haryowidyatna | 2006485011 |
| Muhammad Jauhar Hakim | 2006463982 |
| Siskawati Simandalahi | 2006572970 |

Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Indonesia

Depok


Maret 2022

**Anggota kelompok:**

| No | Nama | NPM | Kontribusi | Tingkat kontribusi |
|---|---|---|---|---|
| 1 | Evan Haryowidyatna | 2006485011 | Mencari dan mengolah data lalu membuat laporan. | 100% |
| 2 | Muhammad Jauhar Hakim | 2006463982 | Mencari dan mengolah data lalu membuat laporan. | 100% |
| 3 | Siskawati Simandalahi | 2006572970 | Mencari dan mengolah data lalu membuat laporan. | 100% |

# I. Penjelasan Data

## A. Permasalahan

Data yang kami gunakan merupakan data risiko perilaku kanker serviks. Sumber data yang kami gunakan yaitu: https://archive.ics.uci.edu/ml/datasets/Cervical+Cancer+Behavior+Risk Data risiko perilaku kanker serviks terdiri dari 19 atribut data numerik yaitu :

1) behavior_eating (perilaku makan)

2) behavior_personalHygine (perilaku kebersihan pribadi)

3) intention_aggregation (niat agregasi)

4) intention_commitment (niat komitmen)

5) attitude_consistency (sikap konsisten)

6) attitude_spontaneity (sikap spontanitas)

7) norm_significantPerson (norma kepada orang penting)

8) norm_fulfillment (norma pemenuhan)

9) perception_vulnerability (persepsi kerentanan)

10) perception_severity (persepsi keparahan)

11) motivation_strength (motivasi kekuatan)

12) motivation_willingness (motivasi kemauan)

13) socialSupport_emotionality (dukungan sosial emosionalitas)

14) socialSupport_appreciation (dukungan sosial apresiasi)

15) socialSupport_instrumental (dukungan sosial intrumental)

16) empowerment_knowledge (pemberdayaan pengetahuan)

17) empowerment_abilities (pemberdayaan kemampuan)

18) empowerment_desires (pemberdayaan keinginan)

19) ca_cervix (merupakan atribut kategorik, 1 = memiliki kanker serviks, 0 = tidak memiliki kanker serviks)

## B. Tujuan

Analisis kluster atau analisis kelompok merupakan teknik analisa data yang bertujuan untuk mengelompokkan individu atau objek ke dalam beberapa kelompok yang memiliki sifat berbeda antar kelompok, sehingga individu atau objek yang terletak di dalam satu kelompok akan mempunyai sifat relatif homogen.

*Clustering* merupakan metode segmentasi data yang sangat berguna dalam prediksi dan analisa masalah bisnis tertentu. Misalnya Segmentasi pasar, marketing dan pemetaan zonasi wilayah. Selain itu dapat digunakan untuk identifikasi obyek dalam bidang berbagai bidang seperti computer vision dan image processing.

## C. Metode Analisis

Metode yang akan digunakan dalam Analisis Klaster ini adalah metode Hierarki dan Non-hierarki. Metode Hierarki adalah metode yang memulai pengelompokan dengan dua atau lebih objek yang mempunyai kesamaan paling dekat. Kemudian proses diteruskan ke objek lain yang mempunyai kedekatan kedua. Demikian seterusnya sehingga kluster akan membentuk semacam pohon di mana ada hierarki (tingkatan yang jelas) antara objek. Berbeda dengan metode hierarki, metode Non-hierarki justru di mulai dengan menentukan terlebih dahulu jumlah kluster yang diinginkan. Setelah jumlah kluster diketahui, baru proses kluster dilakukan tanpa mengikuti proses hierarki. Metode ini juga disebut K-Means cluster.

## D. Teori Ringkas

**Hierarchical Clustering**, pengelompokan data dilakukan dengan membuat suatu bagan hirarki (**dendrogram**) dengan tujuan menunjukkan kemiripan antar data. Setiap data yang mirip akan memiliki hubungan hirarki yang dekat dan menbentuk cluster data. Bagan hirarki akan terus terbentuk hingga seluruh data terhubung dalam bagan hirarki tersebut. Cluster dapat dihasilkan dengan memotong bagan hirarki pada level tertentu. Beberapa metode dalam hierarchical clustering yaitu *single linkage*, *complete linkage*, *average linkage*, dan *ward's minimum variance*.

**Non-Hierarchical Clustering** umumnya bertujuan untuk mengelompokkan data menjadi beberapa cluster yang lebih kecil. Pada prosesnya, setiap cluster akan memiliki titik pusat cluster (*centroid*) dan mencoba menghitung setiap data yang paling dekat dengan centroid tersebut. Terdapat beberapa metode pada non-hierarchical clustering yaitu seperti partitioning, campuran distribusi dan estimasi densitas. Dari ketiga metode tersebut yang paling umum merupakan partitioning. Metode dalam partitional clustering diantaranya *k-means*, *fuzzy k-means*, dan *mixture modeling.*

## II. Klustering Menggunakan Metode Hierarchical

A. Langkah Kerja

1. Import data dan standarisasi data

2. Menghitung matriks jarak (*dissimilarity*) menggunakan metode *euclidean distance* atau *manhattan distance*

3. klustering hierarkis menggunakan metode *single linkage*

4. klustering hierarkis menggunakan metode *complete linkage*

5. klustering hierarkis menggunakan metode *average linkage*

6. klustering hierarkis menggunakan metode *centroid linkage*

7. klustering hierarkis menggunakan metode *ward*

8. Analisis dendrogam serta penentuan berapa kluster yang terbentuk dan banyaknya anggota di dalamnya

9. Visualiasi *scatterplot* dalam 2 dimensi dengan PCA

B. Proses Komputasi

1. Import data dan standarisasi data

```
data <- read.table("sobar72.csv", header=TRUE, sep=",")
data <- as.matrix(data)
data

datanew <- scale(data)
datanew
```

2. Menghitung matriks jarak (*dissimilarity*) menggunakan metode *euclidean distance* atau *manhattan distance*

```
dist_data_euc <- dist(datanew, method="euclidean")
dist(datanew, method="euclidean")
dist(datanew, method="manhattan")
```

3. klustering hierarkis menggunakan metode *single linkage*

```
sin_hc <- hclust(dist_data_euc, method="single")

options(repr.plot.width = 13, repr.plot.height = 7, repr.plot.res = 100)
plot(hclust(dist_data_euc, method="single"))

options(repr.plot.width = 15, repr.plot.height = 7, repr.plot.res = 100)
fviz_dend(sin_hc, k = 4, rect = T, main = "Single Linkage Cluster")

single_clust <- cutree(sin_hc, k = 4)
table(single_clust)
```

4. klustering hierarkis menggunakan metode *complete linkage*

```
com_hc <- hclust(dist_data_euc, method="complete")


options(repr.plot.width = 13, repr.plot.height = 7, repr.plot.res = 100)
plot(hclust(dist_data_euc, method="complete"))

options(repr.plot.width = 15, repr.plot.height = 7, repr.plot.res = 100)
fviz_dend(com_hc, k = 4, rect = T, main = "Complete Linkage Cluster")

complete_clust <- cutree(com_hc, k = 4)
table(complete_clust)
```

5. klustering hierarkis menggunakan metode *average linkage*

```
ave_hc <- hclust(dist_data_euc, method="average")

options(repr.plot.width = 13, repr.plot.height = 7, repr.plot.res = 100)
plot(hclust(dist_data_euc, method="average"))\

options(repr.plot.width = 15, repr.plot.height = 6, repr.plot.res = 100)
fviz_dend(ave_hc, k = 4, rect = T, main = "Average Linkage Cluster")

avg_clust <- cutree(ave_hc, k = 4)
table(avg_clust)
```

6. klustering hierarkis menggunakan metode *centroid linkage*

```
cen_hc <- hclust(dist_data_euc, method="centroid")

options(repr.plot.width = 13, repr.plot.height = 7, repr.plot.res = 100)
plot(hclust(dist_data_euc, method="centroid"))

options(repr.plot.width = 15, repr.plot.height = 8, repr.plot.res = 100)
fviz_dend(cen_hc, k = 4, rect = T, main = "Centroid Linkage Cluster")

cen_clust <- cutree(cen_hc, k = 4)
table(cen_clust)
```

7. klustering hierarkis menggunakan metode *ward*

```
ward_hc <- hclust(dist_data_euc, method="ward.D2")

options(repr.plot.width = 13, repr.plot.height = 7, repr.plot.res = 100)
plot(hclust(dist_data_euc, method="ward.D2"))

options(repr.plot.width = 15, repr.plot.height = 8, repr.plot.res = 100)
```

```
fviz_dend(ward_hc, k = 4, rect = T, main = "Ward Min Variance Cluster")

ward_clust <- cutree(ward_hc, k = 4)
table(ward_clust)
```

8. Analisis dendrogam serta penentuan berapa kluster yang terbentuk dan banyaknya anggota di dalamnya

```
data.frame(complete = cor(complete_coph, dist_data_euc),
           single = cor(single_coph, dist_data_euc),
           average = cor(avg_coph, dist_data_euc),
           centroid = cor(centroid_coph, dist_data_euc),
           ward = cor(ward_coph, dist_data_euc)) %>%
  tidyr::pivot_longer(cols = colnames(.),names_to = "method", values_to = "co
rrelation")

#didapatkan jumlah anggota tiap kluster dengan metode ward
ward_clust <- cutree(ward_hc, k = 4)
table(ward_clust)
```

9. Visualiasi *scatterplot* dalam 2 dimensi dengan PCA

```
hc.cut <- hcut(datanew, k = 4, hc_method = "ward.D2")
hc.cut
fviz_cluster(hc.cut, ellipse.type = "convex")
```

C. Hasil Komputasi

1. Import data dan standarisasi data

- Import data

| behavior_sexualRisk | behavior_eating | behavior_personalHygine | intention_aggregation | intention_commitment | attitude_consistency | attitude_spontar |
|---|---|---|---|---|---|---|
| 10 | 13 | 12 | 4 | 7 | 9 | 10 |
| 10 | 11 | 11 | 10 | 14 | 7 | 7 |
| 10 | 15 | 3 | 2 | 14 | 8 | 10 |
| 10 | 11 | 10 | 10 | 15 | 7 | 7 |
| 8 | 11 | 7 | 8 | 10 | 7 | 8 |
| 10 | 14 | 8 | 6 | 15 | 8 | 10 |
| 10 | 15 | 4 | 6 | 14 | 6 | 10 |
| 8 | 12 | 9 | 10 | 10 | 5 | 10 |
| 10 | 15 | 7 | 2 | 15 | 6 | 10 |
| 7 | 15 | 7 | 6 | 11 | 8 | 8 |
| 7 | 15 | 7 | 10 | 14 | 7 | 9 |
| 10 | 15 | 8 | 9 | 15 | 7 | 10 |
| 10 | 15 | 12 | 10 | 15 | 6 | 10 |
| 9 | 12 | 14 | 9 | 15 | 10 | 9 |
| 2 | 15 | 15 | 6 | 13 | 8 | 9 |
| 10 | 15 | 7 | 6 | 14 | 8 | 8 |
| 10 | 15 | 9 | 7 | 6 | 8 | 8 |
| 10 | 12 | 7 | 5 | 10 | 8 | 8 |
| 10 | 11 | 12 | 2 | 10 | 8 | 8 |
| 10 | 12 | 12 | 8 | 10 | 8 | 6 |

- Standarisasi data

| behavior_sexualRisk | behavior_eating | behavior_personalHygine | intention_aggregation | intention_commitment | attitude_consistency | attitude_spontar |
|---|---|---|---|---|---|---|
| 0.2808717 | 0.0882285 | 0.30214667 | -1.42533462 | -2.6730651 | 1.1947676 | 0.9163363 |
| 0.2808717 | -0.7587651 | -0.02746788 | 0.76592714 | 0.2749104 | -0.1185647 | -1.0629501 |
| 0.2808717 | 0.9352221 | -2.66438430 | -2.15575520 | 0.2749104 | 0.5381014 | 0.9163363 |
| 0.2808717 | -0.7587651 | -0.35708243 | 0.76592714 | 0.6960498 | -0.1185647 | -1.0629501 |
| -1.4043583 | -0.7587651 | -1.34592609 | 0.03550656 | -1.4096470 | -0.1185647 | -0.4031880 |
| 0.2808717 | 0.5117253 | -1.01631154 | -0.69491403 | 0.6960498 | 0.5381014 | 0.9163363 |
| 0.2808717 | 0.9352221 | -2.33476975 | -0.69491403 | 0.2749104 | -0.7752309 | 0.9163363 |
| -1.4043583 | -0.3352683 | -0.68669699 | 0.76592714 | -1.4096470 | -1.4318971 | 0.9163363 |
| 0.2808717 | 0.9352221 | -1.34592609 | -2.15575520 | 0.6960498 | -0.7752309 | 0.9163363 |
| -2.2469733 | 0.9352221 | -1.34592609 | -0.69491403 | -0.9885077 | 0.5381014 | -0.4031880 |
| -2.2469733 | 0.9352221 | -1.34592609 | 0.76592714 | 0.2749104 | -0.1185647 | 0.2565742 |
| 0.2808717 | 0.9352221 | -1.01631154 | 0.40071685 | 0.6960498 | -0.1185647 | 0.9163363 |
| 0.2808717 | 0.9352221 | 0.30214667 | 0.76592714 | 0.6960498 | -0.7752309 | 0.9163363 |
| -0.5617433 | -0.3352683 | 0.96137578 | 0.40071685 | 0.6960498 | 1.8514338 | 0.2565742 |
| -6.4600482 | 0.9352221 | 1.29099033 | -0.69491403 | -0.1462289 | 0.5381014 | 0.2565742 |
| 0.2808717 | 0.9352221 | -1.34592609 | -0.69491403 | 0.2749104 | 0.5381014 | -0.4031880 |
| 0.2808717 | 0.9352221 | -0.68669699 | -0.32970374 | -3.0942045 | 0.5381014 | -0.4031880 |
| 0.2808717 | -0.3352683 | -1.34592609 | -1.06012432 | -1.4096470 | 0.5381014 | -0.4031880 |
| 0.2808717 | -0.7587651 | 0.30214667 | -2.15575520 | -1.4096470 | 0.5381014 | -0.4031880 |
| 0.2808717 | -0.3352683 | 0.30214667 | 0.03550656 | -1.4096470 | 0.5381014 | -1.7227123 |

2. Menghitung matriks jarak (*dissimilarity*) menggunakan metode *euclidean distance* atau *manhattan distance*

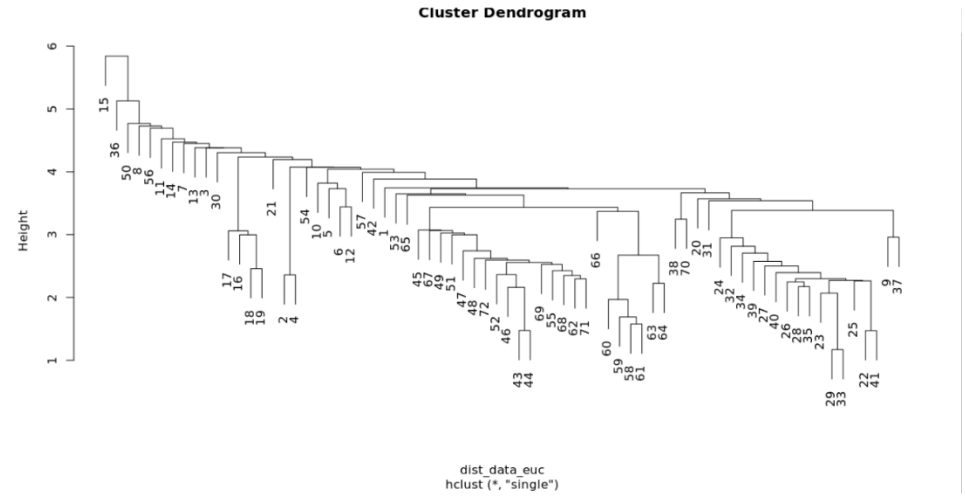
- *Euclidean distance*

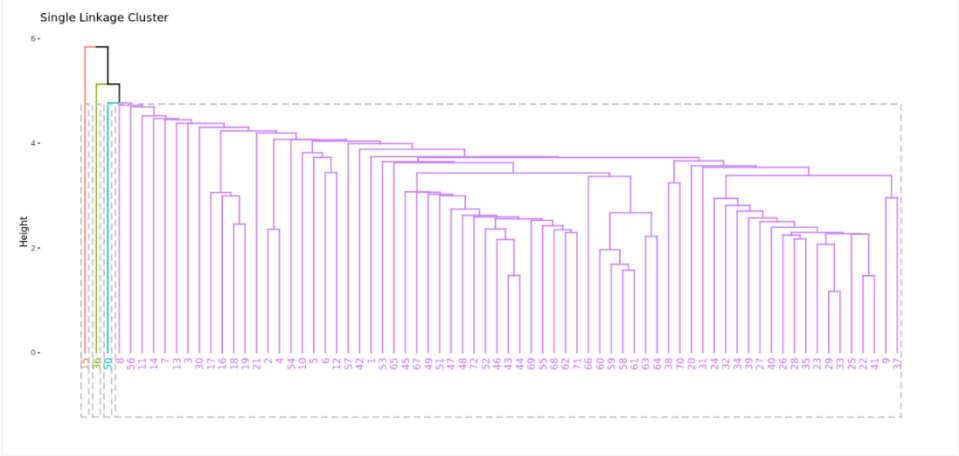
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 2 | 6.025901 | | | | | | |
| 3 | 6.074059 | 7.009447 | | | | | |
| 4 | 6.094742 | 2.359082 | 6.633271 | | | | |
| 5 | 4.603539 | 4.566632 | 5.304614 | 4.281029 | | | |
| 6 | 4.568646 | 4.773170 | 5.192080 | 4.134763 | 4.638993 | | |
| 7 | 6.617594 | 5.403798 | 4.634175 | 5.716588 | 5.473894 | 5.199341 | |
| 8 | 5.818105 | 4.928845 | 6.473883 | 5.448031 | 4.803268 | 5.117842 | 5.200128 |
| 9 | 5.861819 | 7.203653 | 5.876030 | 7.117532 | 6.947512 | 5.352856 | 6.218941 |
| 10 | 5.952587 | 5.005300 | 5.947272 | 5.396928 | 3.821007 | 5.005974 | 5.411228 |
| 11 | 6.568337 | 6.510186 | 4.876541 | 6.347303 | 4.864182 | 5.788372 | 5.418083 |
| 12 | 5.036607 | 5.012707 | 4.381560 | 4.621810 | 3.731245 | 3.441444 | 4.845460 |
| 13 | 6.288964 | 6.650751 | 5.509122 | 6.346141 | 5.873665 | 4.806355 | 5.407676 |
| 14 | 5.299484 | 4.674464 | 6.938215 | 5.159958 | 5.225355 | 5.360183 | 6.462662 |
| 15 | 7.977205 | 8.163665 | 9.054561 | 8.029057 | 6.548038 | 7.649938 | 9.288363 |
| 16 | 6.556076 | 5.745216 | 4.734097 | 5.783532 | 5.901797 | 5.187181 | 4.450679 |
| 17 | 4.822429 | 6.407682 | 5.994417 | 6.197884 | 5.210460 | 5.735144 | 6.224779 |
| 18 | 5.400254 | 5.752023 | 4.505248 | 5.262052 | 4.670079 | 5.089478 | 5.312857 |
| 19 | 5.263519 | 5.800218 | 5.691939 | 5.628564 | 5.451073 | 5.592236 | 5.993676 |
| 20 | 3.744674 | 4.074851 | 6.450992 | 4.357719 | 4.229611 | 4.828098 | 6.187296 |
| 21 | 5.503742 | 5.105765 | 5.853435 | 5.547238 | 5.838090 | 4.194969 | 5.332263 |
| 22 | 5.542596 | 4.499678 | 6.766868 | 4.766876 | 5.352744 | 4.450521 | 6.197372 |
| 23 | 5.928107 | 6.181350 | 7.860700 | 6.144388 | 6.445479 | 5.516186 | 7.776822 |
| 24 | 7.332413 | 7.249429 | 8.601046 | 7.195416 | 7.666805 | 6.567520 | 7.856486 |
| 25 | 5.413501 | 5.093870 | 6.500952 | 5.105651 | 5.525506 | 4.514099 | 6.310534 |
| 26 | 5.686659 | 5.878844 | 7.197468 | 5.752339 | 6.246628 | 4.707137 | 6.899239 |
| 27 | 4.965037 | 5.072415 | 7.280880 | 5.524702 | 6.062501 | 5.243277 | 7.024769 |
| 28 | 5.830499 | 6.500578 | 7.418156 | 6.488689 | 6.654553 | 5.281726 | 7.454224 |
| 29 | 5.951591 | 5.464491 | 7.540872 | 5.399999 | 5.694987 | 5.393315 | 7.266864 |
| 30 | 7.009094 | 5.051689 | 6.374592 | 4.866472 | 5.528309 | 5.864826 | 6.164242 |
| 31 | 5.995873 | 6.268674 | 7.461777 | 6.066401 | 6.220518 | 5.229039 | 7.565695 |
| 32 | 4.538414 | 5.881425 | 7.530595 | 6.609503 | 6.317492 | 5.983973 | 7.049740 |

- *Manhattan distance*

```
          1         2         3         4         5         6         7
2  21.512462
3  19.234737 23.360526
4  21.247012  4.068897 21.977842
5  16.732910 13.650023 17.381179 12.725343
6  13.021583 16.192098 17.234739 13.044545 16.509566
7  23.195776 16.916708 12.088935 18.487777 19.574484 16.827447
8  21.279559 17.416937 23.707868 19.446011 16.560580 17.535509 18.940098
9  19.962043 26.182668 17.076889 24.415710 25.499075 16.486161 18.776897
10 23.107083 16.232522 19.994577 17.803591 12.373134 15.427409 17.227806
11 22.455963 21.151332 14.382549 20.226653 14.684409 18.798536 17.565813
12 15.360211 15.754574 11.176708 13.987616 10.701537 11.112900 14.359972
13 19.940329 22.252090 15.904450 19.763767 19.439400 13.707458 15.669215
14 18.177304 16.151110 23.547575 17.210555 16.963654 19.236848 21.408116
15 22.113055 21.291130 23.878919 21.025680 18.345549 17.220248 27.517387
16 23.383101 17.294246 14.872189 18.239733 18.684306 16.045442 13.596538
17 17.370942 21.223483 19.019280 18.918209 17.312126 16.931347 20.510176
18 18.841789 19.812292 14.567574 16.585674 14.136275 17.038768 17.273104
19 19.069014 19.052384 18.076786 17.568582 17.543072 20.633330 20.921563
20 11.274955 14.883412 22.586224 15.701549 14.902304 16.111179 22.996396
21 18.249149 17.502030 15.860754 19.348054 21.685913 12.717413 14.554825
22 17.811100 14.496653 24.550678 15.556097 18.641802 15.689301 23.495004
23 19.758092 20.445421 26.849421 19.337692 22.406944 19.120345 27.487922
24 22.005910 24.971002 27.403252 23.863273 28.385167 20.988930 27.052451
25 17.634743 17.412364 22.968028 16.728992 19.305831 15.728272 22.498332
26 17.386115 19.259851 22.375924 18.152122 22.674015 15.086308 22.544270
27 17.462949 18.434447 25.648228 20.336171 21.670918 18.976606 25.463427
28 19.691945 21.752023 24.022525 20.644294 24.698427 16.265148 25.427515
29 19.920526 17.890772 27.425606 17.625322 19.868748 19.178643 27.614169
30 22.400015 17.644232 20.486490 16.960861 20.069238 18.557638 22.008799
31 20.220895 20.731766 25.071784 18.964808 22.359711 16.467414 26.476775
32 16.463589 22.774844 26.610407 25.760154 24.125599 20.946948 24.781118
```

3. klustering hierarkis menggunakan metode *single linkage*

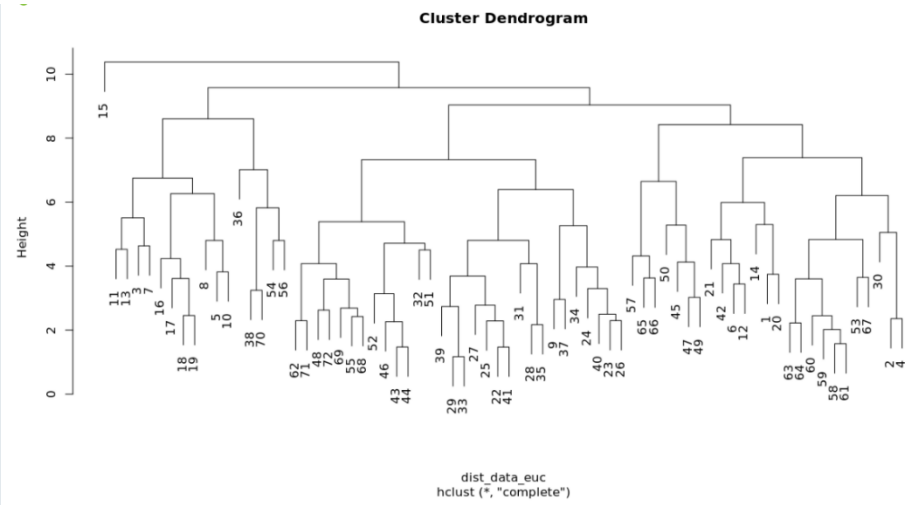


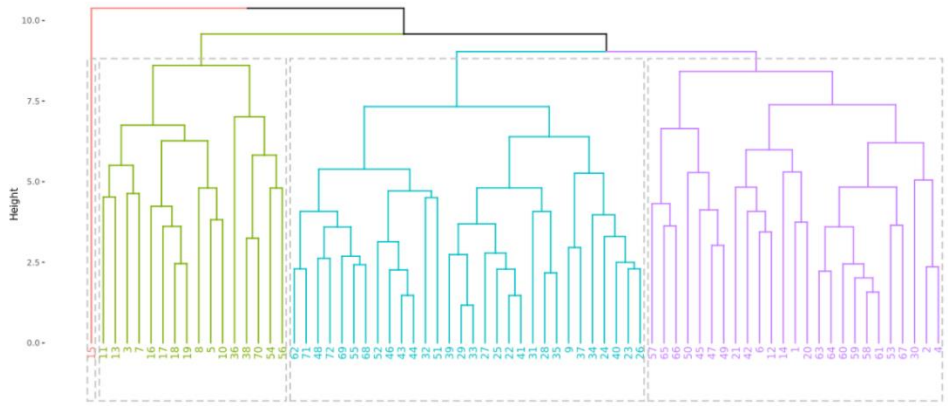



```
0.656760310363611
```

```
single_clust
 1  2  3  4
69  1  1  1
```

4. klustering hierarkis menggunakan metode *complete linkage*

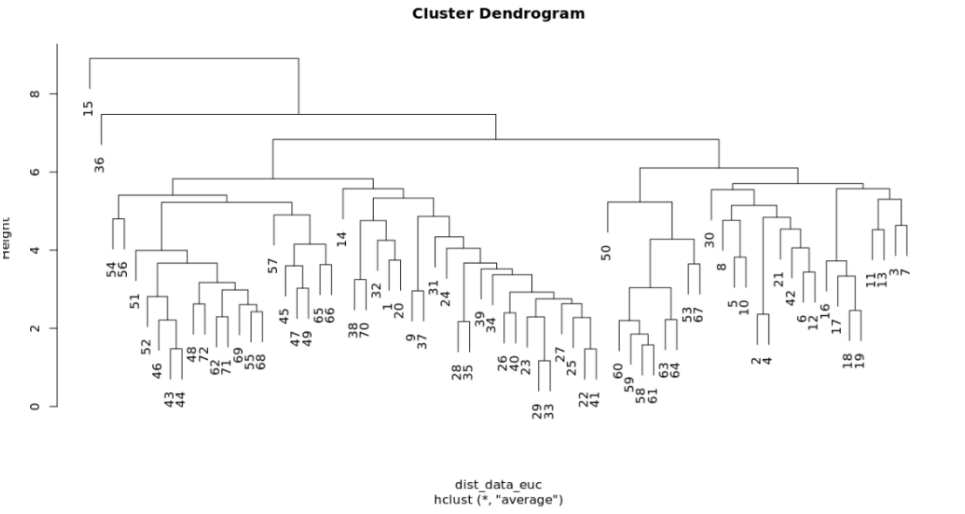


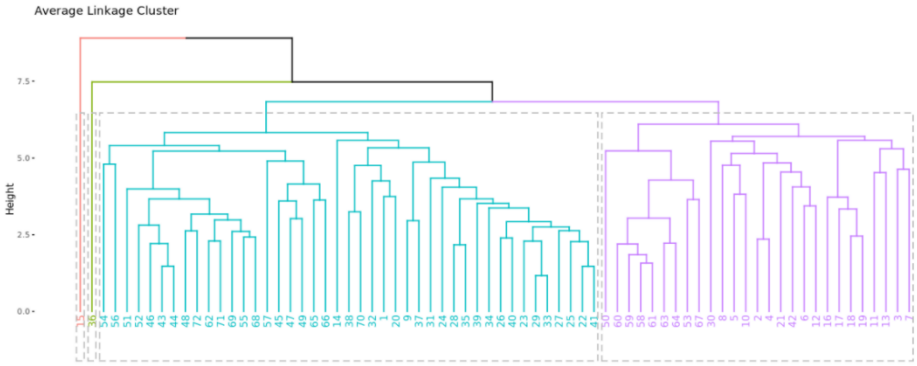



```
0.668176242147213
```

```
complete_clust
 1  2  3  4
25 16 30  1
```

5. klustering hierarkis menggunakan metode *average linkage*

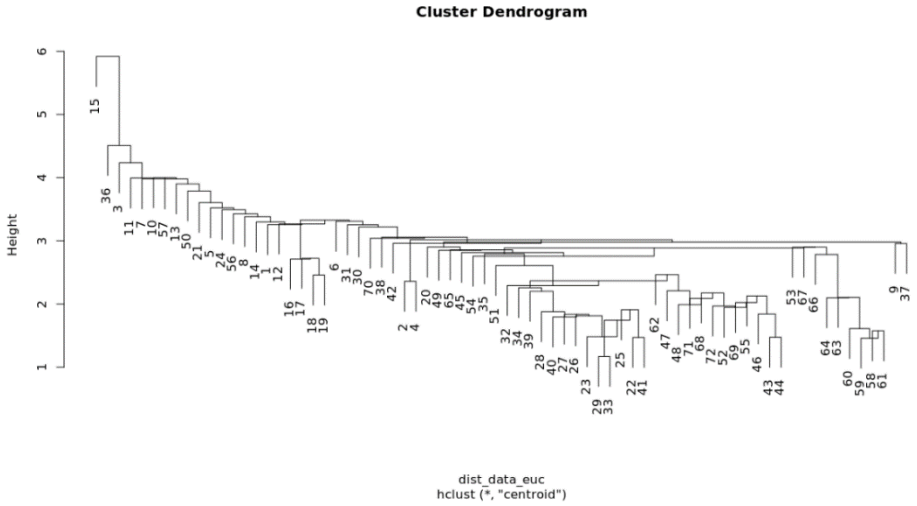


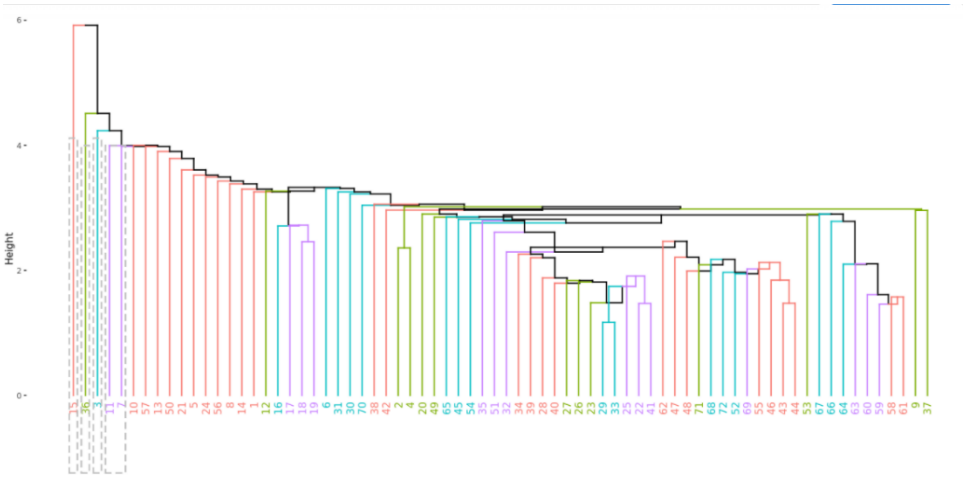



```
0.759360075781317
```

```
avg_clust
 1  2  3  4
43 27  1  1
```

6. klustering hierarkis menggunakan metode *centroid linkage*

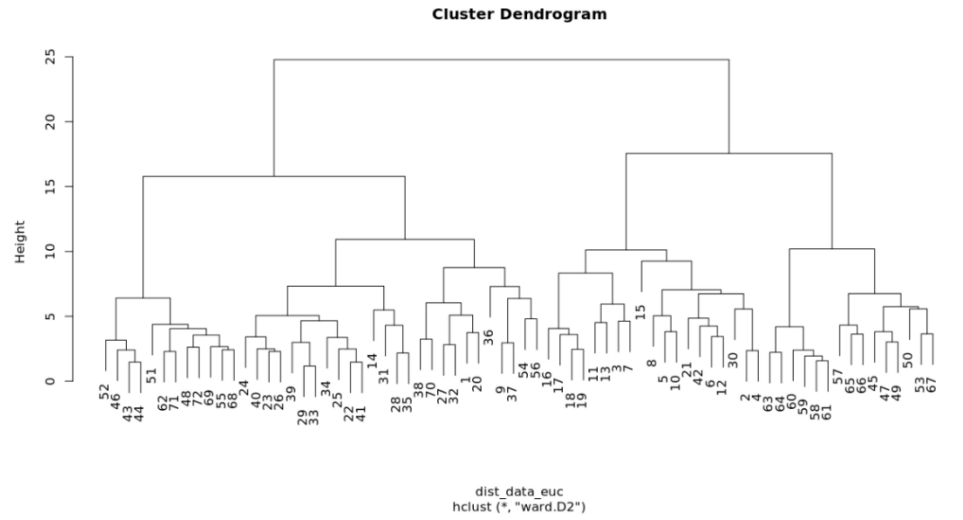


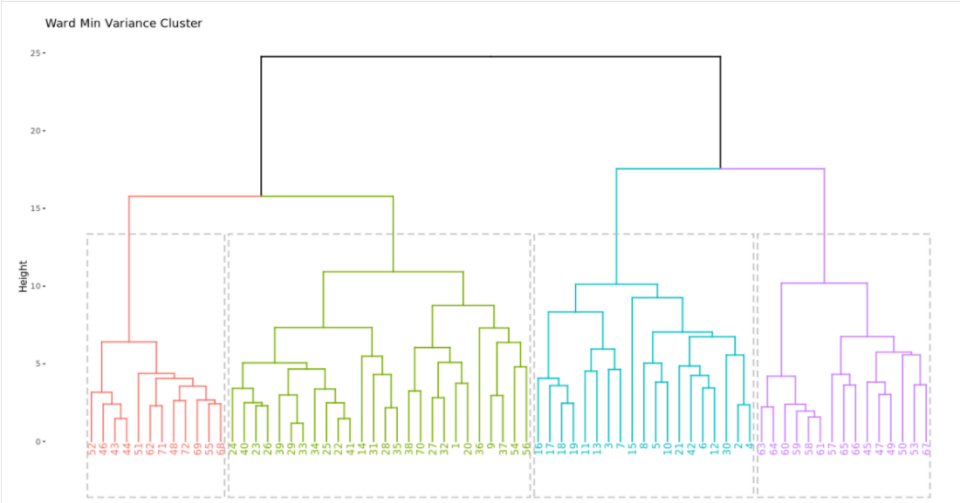



```
0.633357165447986
```

```
cen_clust
 1  2
71  1
```

7. klustering hierarkis menggunakan metode *ward*






```
0.603151420781048
```

```
ward_clust
 1  2  3  4
26 19 12 15
```

8. Analisis dendrogam serta penentuan berapa kluster yang terbentuk dan banyaknya anggota di dalamnya

| method | correlation |
|---|---|
| <chr> | <dbl> |
| complete | 0.6681762 |
| single | 0.6567603 |
| average | 0.7593601 |
| centroid | 0.6333572 |
| ward | 0.6031514 |

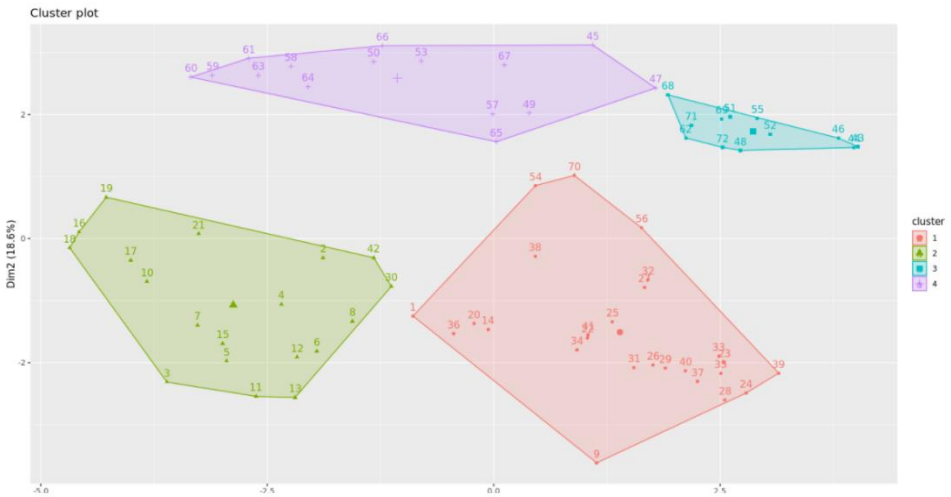
Dapat terlihat bahwa nilai korelasi cophenetic distance yang dihasilkan average linkage paling tinggi dibanding metode lainnya, disusul oleh complete linkage, dan single linkage.

Mempertimbangkan pula dendrogram yang dihasilkan, dendrogram dari average linkage menghasilkan cluster outlier yang hanya berisi cea_cervix, dan ditakutkan membuat hasil clustering tidak optimal.

Oleh karena itu, kami memilih metode ward minimum variance dengan pertimbangan hasil correlation yang lumayan tinggi dan menghasilkan dendrogram yang cukup mudah diinterpretasikan.

```
ward_clust
 1  2  3  4
26 19 12 15
```

9. Visualiasi *scatterplot* dalam 2 dimensi dengan PCA




D. Pembahasan

Kami melakukan klustering hirarkis dengan banyak metode di antaranya yaitu : single linkage, complete linkage, average linkage, centroid linkage, dan ward. Klustering kami mulai dengan menstandarisasi data terlebih dahulu agar dapat mencegah kesenjangan perbedaan dari satuan masing-masing variabel.

Lalu dilanjutkan dengan menghitung matriks dissimilarity dengan metode jarak euclidean dan jarak manhattan. Setelah itu kami membuat visualisasi dendogram untuk tiap metode. Dari analisis mengenai dendrogam tiap metode, kami memilih membaginya menjadi 4 kluster dan dengan mempertimbangkan nilai korelasi cophenetic distance dan visualisasi dari dendogram kami memilih menggunakan metode ward, karena pembagiannya merata dan juga tidak terdapat kluster outlier yang hanya berisi 1 observasi seperti yang ada pada metode average linkage.

Sehingga didapatkan anggota-anggota dari tiap kluster sebagai berikut :

Kluster 1 berisi 12 anggota yaitu anggotanya adalah observasi ke : 52, 46, 43, 44, 51, 62, 71, 48, 72, 69, 55, dan 68.

Kluster 2 berisi 26 anggota yaitu anggotanya adalah observasi ke : 24, 40, 23, 26, 39, 29, 33, 34, 25, 22, 41, 14, 31, 28, 35, 38, 70, 27, 32, 1, 20, 36, 9, 37, 54, dan 56.

Kluster 3 berisi 19 anggota yaitu anggotanya adalah observasi ke : 16, 17, 18, 19, 11, 13, 3, 7, 15, 8, 5, 10, 21, 42, 6, 12, 30, 2, dan 4.

Kluster 4 berisi 15 anggota yaitu anggotanya adalah observasi ke : 63, 64, 60, 59, 58, 61, 57, 65, 66, 45, 47, 49, 50, 53, dan 67.

## III. Klustering Menggunakan Metode Non-Hierarchical

A. Langkah Kerja

1. Penentuan jumlah kluster k yang diinginkan

2. Menentukan nilai *centroid cluster*. Nilai *centroid* merupakan rata-rata objek dalam kluster tersebut.

3. Menentukan jarak terdekat setiap objek terhadap tiap *centroid cluster* dengan menggunakan ukuran jarak *euclidean*

4. Hitung kembali *centroid* dari anggota kluster yang baru terbentuk

5. Jika *centroid cluster* tidak berubah maka langkah terhenti, namun jika nilai *centroid* berubah maka lakukan langkah ke 3 dan 4 sampai nilai *centroid* tidak berubah lagi.

B. Proses Komputasi

1. Import data dan standarisasi data

```
data <- read.table("sobar72.csv", header=TRUE, sep=",")
data <- as.matrix(data)
data

datanew <- scale(data)
datanew
```

2. Penentuan jumlah kluster k yang diinginkan

Dari metode klustering hirarki kami akan menggunakan $k = 4$

3. klustering non hirarkis dengan metode k-means

```
km <- kmeans(datanew, centers = 4)
str(km)

km

km$centers
```

4. Menentukan banyak anggota tiap kluster

```
km$size
```

5. Scatterplot dari 2 dimensi dengan PCA

```
fviz_cluster(km, data = datanew)
```

C. Hasil Komputasi

1. Import data dan standarisasi data

- Import data

| behavior_sexualRisk | behavior_eating | behavior_personalHygine | intention_aggregation | intention_commitment | attitude_consistency | attitude_spontar |
|---|---|---|---|---|---|---|
| 10 | 13 | 12 | 4 | 7 | 9 | 10 |
| 10 | 11 | 11 | 10 | 14 | 7 | 7 |
| 10 | 15 | 3 | 2 | 14 | 8 | 10 |
| 10 | 11 | 10 | 10 | 15 | 7 | 7 |
| 8 | 11 | 7 | 8 | 10 | 7 | 8 |
| 10 | 14 | 8 | 6 | 15 | 8 | 10 |
| 10 | 15 | 4 | 6 | 14 | 6 | 10 |
| 8 | 12 | 9 | 10 | 10 | 5 | 10 |
| 10 | 15 | 7 | 2 | 15 | 6 | 10 |
| 7 | 15 | 7 | 6 | 11 | 8 | 8 |
| 7 | 15 | 7 | 10 | 14 | 7 | 9 |
| 10 | 15 | 8 | 9 | 15 | 7 | 10 |
| 10 | 15 | 12 | 10 | 15 | 6 | 10 |
| 9 | 12 | 14 | 9 | 15 | 10 | 9 |
| 2 | 15 | 15 | 6 | 13 | 8 | 9 |
| 10 | 15 | 7 | 6 | 14 | 8 | 8 |
| 10 | 15 | 9 | 7 | 6 | 8 | 8 |
| 10 | 12 | 7 | 5 | 10 | 8 | 8 |
| 10 | 11 | 12 | 2 | 10 | 8 | 8 |
| 10 | 12 | 12 | 8 | 10 | 8 | 6 |

- Standarisasi data

| behavior_sexualRisk | behavior_eating | behavior_personalHygine | intention_aggregation | intention_commitment | attitude_consistency | attitude_spontar |
|---|---|---|---|---|---|---|
| 0.2808717 | 0.0882285 | 0.30214667 | -1.42533462 | -2.6730651 | 1.1947676 | 0.9163363 |
| 0.2808717 | -0.7587651 | -0.02746788 | 0.76592714 | 0.2749104 | -0.1185647 | -1.0629501 |
| 0.2808717 | 0.9352221 | -2.66438430 | -2.15575520 | 0.2749104 | 0.5381014 | 0.9163363 |
| 0.2808717 | -0.7587651 | -0.35708243 | 0.76592714 | 0.6960498 | -0.1185647 | -1.0629501 |
| -1.4043583 | -0.7587651 | -1.34592609 | 0.03550656 | -1.4096470 | -0.1185647 | -0.4031880 |
| 0.2808717 | 0.5117253 | -1.01631154 | -0.69491403 | 0.6960498 | 0.5381014 | 0.9163363 |
| 0.2808717 | 0.9352221 | -2.33476975 | -0.69491403 | 0.2749104 | -0.7752309 | 0.9163363 |
| -1.4043583 | -0.3352683 | -0.68669699 | 0.76592714 | -1.4096470 | -1.4318971 | 0.9163363 |
| 0.2808717 | 0.9352221 | -1.34592609 | -2.15575520 | 0.6960498 | -0.7752309 | 0.9163363 |
| -2.2469733 | 0.9352221 | -1.34592609 | -0.69491403 | -0.9885077 | 0.5381014 | -0.4031880 |
| -2.2469733 | 0.9352221 | -1.34592609 | 0.76592714 | 0.2749104 | -0.1185647 | 0.2565742 |
| 0.2808717 | 0.9352221 | -1.01631154 | 0.40071685 | 0.6960498 | -0.1185647 | 0.9163363 |
| 0.2808717 | 0.9352221 | 0.30214667 | 0.76592714 | 0.6960498 | -0.7752309 | 0.9163363 |
| -0.5617433 | -0.3352683 | 0.96137578 | 0.40071685 | 0.6960498 | 1.8514338 | 0.2565742 |
| -6.4600482 | 0.9352221 | 1.29099033 | -0.69491403 | -0.1462289 | 0.5381014 | 0.2565742 |
| 0.2808717 | 0.9352221 | -1.34592609 | -0.69491403 | 0.2749104 | 0.5381014 | -0.4031880 |
| 0.2808717 | 0.9352221 | -0.68669699 | -0.32970374 | -3.0942045 | 0.5381014 | -0.4031880 |
| 0.2808717 | -0.3352683 | -1.34592609 | -1.06012432 | -1.4096470 | 0.5381014 | -0.4031880 |
| 0.2808717 | -0.7587651 | 0.30214667 | -2.15575520 | -1.4096470 | 0.5381014 | -0.4031880 |
| 0.2808717 | -0.3352683 | 0.30214667 | 0.03550656 | -1.4096470 | 0.5381014 | -1.7227123 |

2. Penentuan jumlah kluster k yang diinginkan

K = 4

3. klustering non hirarkis dengan metode k-means

```
List of 9
 $ cluster     : int [1:72] 4 4 4 4 4 4 4 4 3 4 ...
 $ centers     : num [1:4, 1:20] 0.2247 0.0702 0.2426 -0.5174 0.06 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:4] "1" "2" "3" "4"
  .. ..$ : chr [1:20] "behavior_sexualRisk" "behavior_eating" "behavior_personalHygine" "intention_aggregation" ...
 $ totss       : num 1420
 $ withinss    : num [1:4] 119 179 227 300
 $ tot.withinss: num 825
 $ betweenss   : num 595
 $ size        : int [1:4] 15 16 22 19
 $ iter        : int 3
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
```

```
K-means clustering with 4 clusters of sizes 15, 16, 22, 19

Cluster means:
  behavior_sexualRisk behavior_eating behavior_personalHygine
1          0.22469733      0.05999538               0.7416327
2          0.07021791     -0.09705135              -0.1098715
3          0.24257098     -0.27751874               0.1373394
4         -0.51739516      0.35570017              -0.6520007
  intention_aggregation intention_commitment attitude_consistency
1            0.27898009          -0.03392512           0.14410174
2            0.03550656           0.35387404           0.25081000
3            0.11850890           0.08348343          -0.29765550
4           -0.38736852          -0.36788124           0.01968078
  attitude_spontaneity norm_significantPerson norm_fulfillment
1          -0.44717213              0.7991452        1.0556239
2           0.29780931              0.7449659        1.0216628
3           0.04664985             -0.7080254       -0.8863330
4           0.04822823             -0.4384249       -0.6674546
  perception_vulnerability perception_severity motivation_strength
1                1.1427665           1.1990900           0.2953416
2                0.7276285           0.8045665           0.3810859
3               -0.7899375          -0.7559286           0.2216440
4               -0.6002593          -0.7488938          -0.8107193
  motivation_willingness socialSupport_emotionality socialSupport_appreciation
1              0.8810004                  0.8726440                  0.7708317
2             -0.5312903                 -0.9066857                 -0.9419679
3              0.5802008                  0.6948182                  0.7739695
4             -0.9199357                 -0.7299311                 -0.7114904
  socialSupport_instrumental empowerment_knowledge empowerment_abilities
1                  0.7162464             0.7156323             0.8801212
2                 -1.1438705            -0.4246085            -0.5845333
3                  0.6923773             0.6150100             0.6301245
```

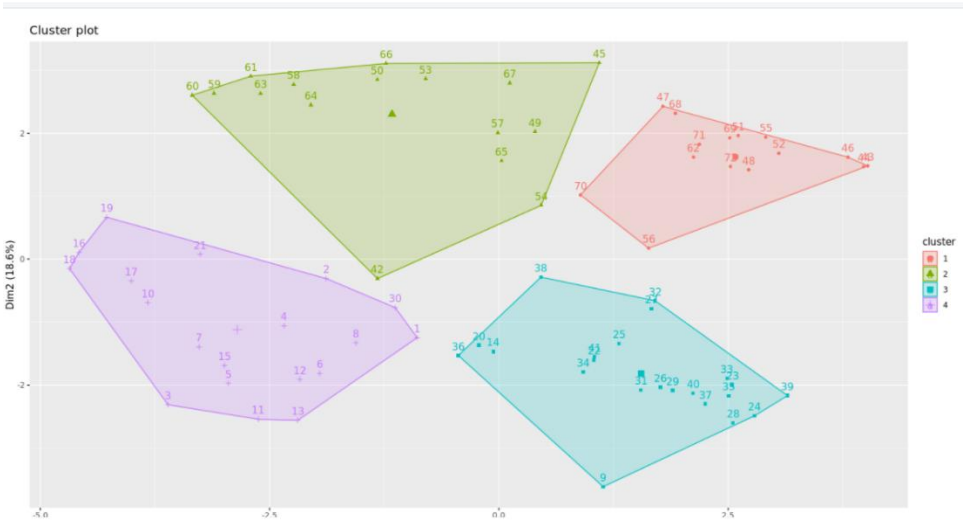
| | behavior_sexualRisk | behavior_eating | behavior_personalHygine | intention_aggregation | intention_commitment | attitude_consistency | attitude_sp |
|---|---|---|---|---|---|---|---|
| 1 | 0.22469733 | 0.05999538 | 0.7416327 | 0.27898009 | -0.03392512 | 0.14410174 | -0.4471721 |
| 2 | 0.07021791 | -0.09705135 | -0.1098715 | 0.03550656 | 0.35387404 | 0.25081000 | 0.29780931 |
| 3 | 0.24257098 | -0.27751874 | 0.1373394 | 0.11850890 | 0.08348343 | -0.29765550 | 0.04664985 |
| 4 | -0.51739516 | 0.35570017 | -0.6520007 | -0.38736852 | -0.36788124 | 0.01968078 | 0.04822823 |

A matrix: 4 × 20 of type dbl

4. Menentukan banyak anggota tiap kluster

15 · 16 · 22 · 19

5. Scatterplot dari 2 dimensi dengan PCA




D. Pembahasan

Pada klustering non-hirarkis kami menggunakan metode k-means. Dan dari hasil klustering hirarkis kami memutuskan menggunakan jumlah klusternya sebesar 4. Sama seperti pada klustering hirarkis, sebelum memulainya kami menstandarisasi datanya terlebih dahulu. Lalu kami melakukan klustering non-hirarkis metode k-means dengan bantuan program R.

Sehingga didapatkan anggota-anggota dari tiap kluster sebagai berikut :

Kluster 1 berisi 15 anggota yaitu anggotanya adalah observasi ke : 47, 68, 62, 71, 51, 55, 48, 52, 55, 46, 43, 44, 56, 70, dan 72.

Kluster 2 berisi 16 anggota yaitu anggotanya adalah observasi ke : 42, 54, 45, 49, 65, 57, 67, 53, 50, 66, 64, 58, 63, 61, 59, dan 60.

Kluster 3 berisi 19 anggota yaitu anggotanya adalah observasi ke : 16, 17, 18, 19, 10, 21, 2, 4, 7, 8, 3, 5, 12, 15, 6, 11, 13, 1, dan 30.

Kluster 4 berisi 22 anggota yaitu anggotanya adalah observasi ke : 9, 14, 20, 36, 38, 27, 32, 39, 24, 28, 23, 33, 35, 37, 40, 29, 26, dan 31.

## IV. Menentukan Metode Kluster Terbaik

Keunggulan dari algoritma *hierarchical* adalah kemudahan dalam memahami prosedur kerja dan penerapan pada data. Selain itu, penentuan *cluster* cukup jelas yaitu dengan melihat dendogram dan tidak memerlukan informasi tambahan. Namun, kekurangan yang jelas terlihat adalah ketika menghadapi data dengan jumlah observasi yang banyak. Kekurangan ini sangatlah berpengaruh pada lama proses pengerjaan/komputasi.

Jika dibandingkan dengan algoritma *hierarchical*, terlihat bahwa *k means* lebih cepat dalam mendapatkan solusinya pada data contoh diatas. Tentunya ini karena pemilihan pusat awal yang "baik". Dalam penentuan titik pusat di langkah pertama ini, anda harus berhati-hati dengan keadaan data. Hal ini karena jika anda memilih titik awal yang "buruk" hasil yang anda dapatkan juga akan buruk dalam artian tidak mencapai solusi optimal dalam masalah *clustering*. Bahkan, lebih buruk lagi adalah ada kemungkinan ada *cluster* kosong saat langkah pertama dengan kata lain solusi pun tidak bisa anda dapatkan.

Dalam hierarchical clustering, menentukan banyaknya cluster optimal yang harus dibentuk terbilang cukup sulit. Terdapat beberapa metode yang populer untuk menentukan banyaknya cluster yang optimal, di antaranya yaitu :

1. Elbow Method

   Tujuan awal dari clustering merupakan dibentuknya kelompok data di mana dalam sub-kelompok sehomogen mungkin dan antar sub-kelompok seheterogen mungkin, sehingga dengan menggunakan metode siku (elbow method) ini akan dicari k cluster yang optimal untuk meminimalkan WSS (Total Within-Cluster Sum of Square).

2. Average Silhoutte Method

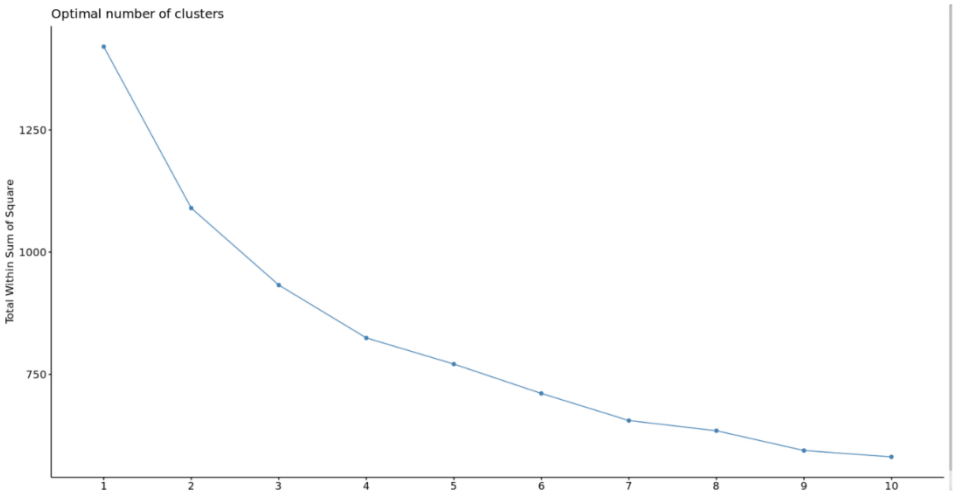
   Metode siluet rata-rata (average silhoutte) mengukur kualitas pengelompokan. Artinya, metode ini menentukan seberapa baik setiap objek terletak di dalam clusternya. Lebar siluet rata-rata yang tinggi menunjukkan pengelompokan yang baik. Metode siluet rata-rata menghitung siluet rata-rata pengamatan untuk nilai k kluster yang berbeda. Jumlah optimal kluster k adalah yang memaksimalkan rata-rata siluet pada rentang nilai yang mungkin untuk k

3. Gap Statistic

   Metode Gap Statistic telah diterbitkan oleh R. Tibshirani, G. Walther, dan T. Hastie (Standford University, 2001). Pendekatan ini dapat diterapkan pada metode pengelompokan apa pun (K-means clustering atau hierarchy clustering). Metode Gap Statistic membandingkan total variasi di dalam cluster untuk nilai k yang berbeda dengan nilai yang diharapkan di bawah distribusi referensi nol dari data (yaitu distribusi tanpa pengelompokan yang jelas). Dataset referensi dihasilkan menggunakan simulasi Monte Carlo dari proses pengambilan sampel. Artinya, untuk setiap variabel ($x_i$) dalam kumpulan data, metode ini akan menghitung rentangnya (min ($x_i$),max ($x_i$)) dan menghasilkan nilai untuk n titik secara seragam dari interval minimal hingga maksimal.

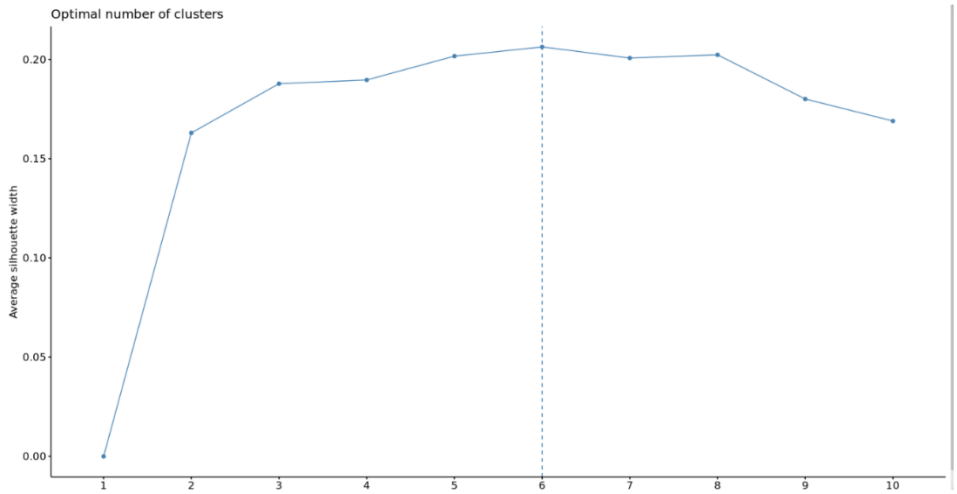
A. Metode Siku (Elbow Method)

```
#Elbow Method
fviz_nbclust(datanew, kmeans, method = "wss")
```



## B. Metode Siluet Rata-Rata (Average Silhoutte)

```
#Average Silhoutte Method
fviz_nbclust(datanew, kmeans, method = "silhouette")
```

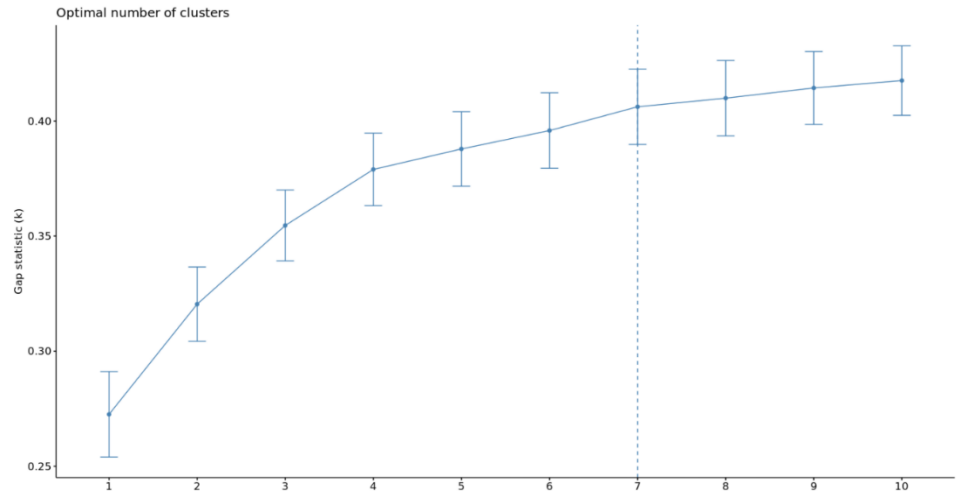



## C. Metode Gap Statistic

```
# compute gap statistic


set.seed(123)
gap_stat <- clusGap(datanew, FUN = kmeans, nstart = 25,
                    K.max = 10, B = 50)
# Print the result
print(gap_stat, method = "firstmax")

fviz_gap_stat(gap_stat)
```



D. Penentuan jumlah kluster optimal

Dapat dilihat dari ketiga metode di atas, jumlah kluster k optimal sebesar 6 ataupun 7 sebagaimana dapat dilihat pada metode siluet rata-rata titik tertinggi berada pada k=6 sedangkan pada gap statistics, dapat dilihat bahwa yang terbaik berada di 7. Oleh karena itu kami memilih k=6 untuk menentukan jumlah cluster terbaik.

E. Kluster akhir

Akan dilakukan kluster non hirarkis metode k-means dengan jumlah kluster $k = 6$.

```
kmfin <- kmeans(datanew, centers = 6)
str(kmfin)
```

```
List of 9
 $ cluster     : int [1:72] 3 3 5 3 3 5 5 3 1 6 ...
 $ centers     : num [1:6, 1:20] 0.2809 0.2207 -0.1021 0.0702 0.0511 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:6] "1" "2" "3" "4" ...
  .. ..$ : chr [1:20] "behavior_sexualRisk" "behavior_eating" "behavior_personalHygine" "intention_aggregation" ...
 $ totss       : num 1420
 $ withinss    : num [1:6] 132 97.9 159.4 178.5 131.8 ...
 $ tot.withinss: num 717
 $ betweenss   : num 703
 $ size        : int [1:6] 18 14 11 16 11 2
 $ iter        : int 3
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
```

```
kmfin
```

```
K-means clustering with 6 clusters of sizes 18, 14, 11, 16, 11, 2

Cluster means:
  behavior_sexualRisk behavior_eating behavior_personalHygine
1          0.28087166     -0.00588190              0.19227516
2          0.22068488      0.20922759              0.79656850
3         -0.10213515     -0.91276394             -0.26718755
4          0.07021791     -0.09705135             -0.10987152
5          0.05106757      0.62722443             -0.89645170
6         -4.35351072      0.93522211             -0.02746788
  intention_aggregation intention_commitment attitude_consistency
1            0.29926955           0.39189357           -0.4104164
2            0.34854395           0.06434074            0.2097684
3           -0.19689999          -0.95022227           -0.2379586
4            0.03550656           0.35387404            0.2508100
5           -0.66171309          -0.18451434            0.1799199
6           -0.69491403          -0.56736830            0.5381014
  attitude_spontaneity norm_significantPerson norm_fulfillment
1           0.18326727            -0.79011533       -0.9367601
2          -0.26181038             0.86106446        1.0653271
3          -0.76305825            -0.31399388       -0.4880620
4           0.29780931             0.74496588        1.0216628
5           0.37653093            -0.56026359       -0.6177317
6          -0.07330691            -0.06772417       -1.1178859
  perception_vulnerability perception_severity motivation_strength
1               -0.8867972          -0.8004824          0.29880602
2                1.2162718           1.2298957          0.28643310
3               -0.5666879          -0.6222674          0.07991781
4                0.7276285           0.8045665          0.38108593
5               -0.3540692          -0.7291964         -1.62079627
6               -1.2895915          -0.4084094          0.73185821
  motivation_willingness socialSupport_emotionality socialSupport_appreciation
```

```
kmfin$centers
```

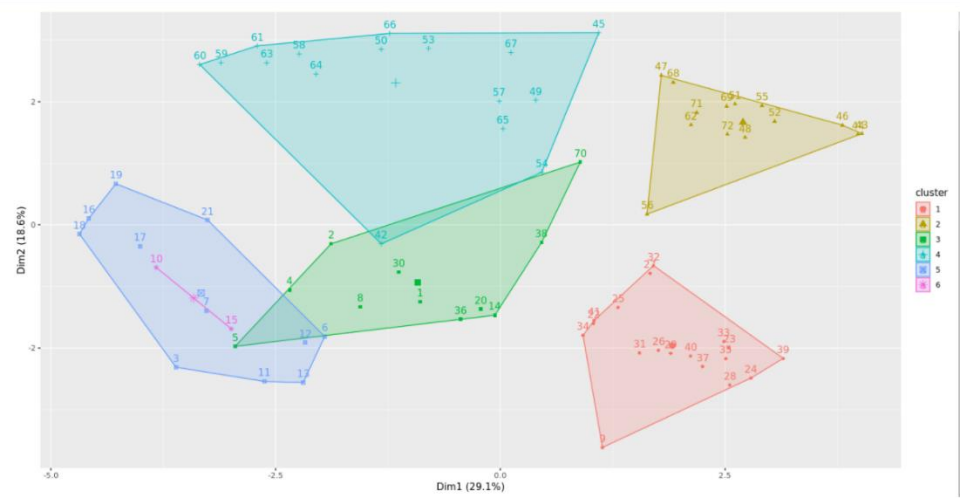
| | behavior_sexualRisk | behavior_eating | behavior_personalHygine | intention_aggregation | intention_commitment | attitude_consistency | attitude_sp |
|---|---|---|---|---|---|---|---|
| 1 | 0.28087166 | -0.00588190 | 0.19227516 | 0.29926955 | 0.39189357 | -0.4104164 | 0.18326727 |
| 2 | 0.22068488 | 0.20922759 | 0.79656850 | 0.34854395 | 0.06434074 | 0.2097684 | -0.26181038 |
| 3 | -0.10213515 | -0.91276394 | -0.26718755 | -0.19689999 | -0.95022227 | -0.2379586 | -0.76305825 |
| 4 | 0.07021791 | -0.09705135 | -0.10987152 | 0.03550656 | 0.35387404 | 0.2508100 | 0.29780931 |
| 5 | 0.05106757 | 0.62722443 | -0.89645170 | -0.66171309 | -0.18451434 | 0.1799199 | 0.37653093 |
| 6 | -4.35351072 | 0.93522211 | -0.02746788 | -0.69491403 | -0.56736830 | 0.5381014 | -0.07330691 |

A matrix: 6 × 20 of type dbl

```
kmfin$size
```

18 · 14 · 11 · 16 · 11 · 2

```
#Perhatikan ini hanya 2 dimensi dengan PCA, perlu diketahui ada > 15 variabel
fviz_cluster(kmfin, data = datanew)
```




Kami melakukan klustering non hirarkis sama seperti pada sebelumnya namun yang hanya membedakan sekarang kami menggunakan $k = 6$ yang menandakan jumlah kluster yang ingin dihasilkan sebesar 6 kluster.

Sehingga didapatkan anggota-anggota dari tiap kluster sebagai berikut :

Kluster 1 berisi 2 anggota yaitu anggotanya adalah observasi ke : 15 dan 17

Kluster 2 berisi 11 anggota yaitu anggotanya adalah observasi ke : 6, 7, 11, 12, 13, 3, 16, 17, 18, 19, dan 21.

Kluster 3 berisi 11 anggota yaitu anggotanya adalah observasi ke : 1, 2, 4, 5, 8, 14, 20, 30, 36, 38, dan 70.

Kluster 4 berisi 18 anggota yaitu anggotanya adalah observasi ke : 9, 22, 23, 24, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 37, 39, 40, dan 41.

Kluster 5 berisi 14 anggota yaitu anggotanya adalah observasi ke : 43, 44, 46, 47, 48, 51, 52, 55, 56, 62, 68, 69, 71, dan 72.

Kluster 6 berisi 16 anggota yaitu anggotanya adalah observasi ke : 42, 45, 49, 50, 53, 54, 57, 58, 59, 60, 61, 63, 64, 66, dan 67.