

Unsupervised Learning

December 9, 2024

Introduction

Machine learning bertujuan untuk membuat mesin yang belajar berdasarkan data. Machine learning terbagi dua:

1. Supervised Learning:

- Memiliki target variable.
- Tujuan: membuat model prediksi dengan training data.
- Model di-evaluasi dengan testing data (karena memiliki target variable)

2. Unsupervised Learning:

- Tidak memiliki target variable.
- Tujuan: mencari pola dalam data, yang berguna untuk menghasilkan informasi. Digunakan pada tahap pre-processing maupun Exploratory Data Analysis (EDA).
- Tidak ada target variable sehingga tidak ada evaluasi model

Dalam studi kasus ini, kita akan menggunakan metode **unsupervised learning** lebih tepatnya Clustering dengan KMeans, di mana variabel target tidak diketahui atau ditentukan.

Clustering

Teknik yang digunakan untuk mengelompokkan data berdasarkan kemiripan atau karakteristik tertentu.

Clustering bertujuan untuk menghasilkan cluster dimana:

- Observasi di satu cluster yang sama yang memiliki karakteristik yang mirip
- Observasi dari cluster yang berbeda memiliki karakteristik yang berbeda

K-Means

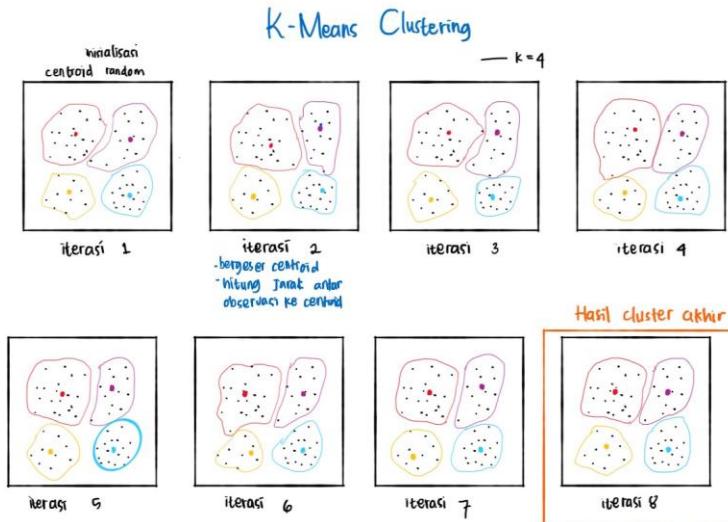
K-means adalah *centroid-based* clustering algorithms, artinya tiap cluster memiliki satu titik pusat (centroid) yang mewakili cluster tersebut.

K-means merupakan proses yang berulang dari:

1. **Random initialization:** meletakkan k centroid secara random

2. **Cluster assignment:** assign masing-masing observasi ke cluster terdekat, berdasarkan perhitungan jarak
3. **Centroid update:** menggeser centroid ke rata-rata (means) dari cluster yang terbentuk
4. Ulangi langkah 2 dan 3 sampai tidak ada observasi yang clusternya berubah lagi

Note: Banyaknya cluster k ditentukan oleh user.



K-means Workflow

Business Question: Whiskey Recommendation

Kita sebagai seorang data scientist sebuah toko whisky diminta untuk membuat product recommendation untuk whisky berdasarkan preferensi rasa masing-masing customer!

Tujuan: membentuk kelompok whisky yang memiliki karakteristik rasa khas pada tiap clusternya

Data yang digunakan berupa data penyulingan Malt Whisky dari 86 pabrik penyulingan, diperoleh dari penelitian Dr. Wisehart (Universitas St. Andrews). Setiap whisky diberi skor 0-4 dari 12 kategori cita rasa berdasarkan uji organoleptik:

Sumber: <https://github.com/sweis/whiskies/blob/master/whiskies.txt>

- Body: tingkat kekuatan rasa (light/heavy)
- Sweetness: tingkat rasa manis
- Smoky: tingkat rasa asap
- Medicinal: tingkat rasa pahit (obat)
- Tobacco: tingkat rasa tembakau
- Honey: tingkat rasa madu
- Spicy: tingkat rasa pedas
- Winey: tingkat rasa anggur

- Nutty: tingkat rasa kacang
- Malty: tingkat rasa gandum
- Fruity: tingkat rasa buah
- Floral: tingkat rasa bunga

Read Data

```
whisky <- read.csv("whiskies.txt")
glimpse(whisky)

#> Rows: 86
#> Columns: 17
#> $ RowID      <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,
#> 17, ...
#> $ Distillery <chr> "Aberfeldy", "Aberlour", "AnCnoc", "Ardbeg", "Ardmore",
#> "Ar...
#> $ Body        <int> 2, 3, 1, 4, 2, 2, 0, 2, 2, 2, 4, 3, 4, 2, 3, 2, 1, 2,
#> 2, 1, ...
#> $ Sweetness   <int> 2, 3, 3, 1, 2, 3, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2,
#> 2, 1, ...
#> $ Smoky       <int> 2, 1, 2, 4, 2, 1, 0, 1, 1, 2, 2, 1, 2, 2, 1, 2,
#> 3, 2, ...
#> $ Medicinal   <int> 0, 0, 0, 4, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
#> 1, 2, ...
#> $ Tobacco     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
#> 0, 0, ...
#> $ Honey        <int> 2, 4, 2, 0, 1, 1, 1, 2, 1, 0, 2, 3, 2, 2, 3, 2, 0, 1,
#> 2, 2, ...
#> $ Spicy        <int> 1, 3, 0, 2, 1, 1, 1, 0, 2, 1, 2, 2, 2, 1, 2, 1, 2,
#> 2, 2, ...
#> $ Winey       <int> 2, 2, 0, 0, 1, 1, 0, 2, 0, 0, 3, 1, 0, 0, 1, 1, 1, 2,
#> 1, 1, ...
#> $ Nutty       <int> 2, 2, 2, 1, 2, 0, 2, 2, 2, 3, 0, 2, 0, 2, 2, 0, 2,
#> 1, 2, ...
#> $ Malty       <int> 2, 3, 2, 2, 3, 1, 2, 2, 2, 1, 0, 2, 2, 2, 3, 2, 2, 2,
#> 1, 2, ...
#> $ Fruity      <int> 2, 3, 3, 1, 1, 1, 3, 2, 2, 2, 1, 2, 2, 3, 2, 2, 2, 2,
#> 1, 2, ...
#> $ Floral      <int> 2, 2, 2, 0, 1, 2, 3, 1, 2, 1, 2, 2, 2, 2, 2, 3, 2,
#> 2, 2, ...
#> $ Postcode    <chr> "\tPH15 2EB", "\tAB38 9PJ", "\tAB5 5LI", "\tPA42
#> 7EB", ...
#> $ Latitude    <int> 286580, 326340, 352960, 141560, 355350, 194050, 247670,
#> 340...
#> $ Longitude   <dbl> 749680, 842570, 839320, 646220, 829140, 649950, 672610,
#> 848...
```

note: k-means hanya bisa dilakukan pada data dengan tipe data numerik

Cleansing data

```

# meng-assign nilai dari kolom Distillery menjadi rownames
rownames(whisky) <- whisky$Distillery

# membuang kolom yang tidak digunakan
whisky_clean <- whisky %>%
  select(-c(RowID, Distillery, Postcode, Latitude, Longitude))
whisky_clean

#>           Body Sweetness Smoky Medicinal Tobacco Honey Spicy
Winey
#> Aberfeldy      2        2     2       0      0     2    1
2
#> Aberlour       3        3     1       0      0     4    3
2
#> AnCnoc         1        3     2       0      0     2    0
0
#> Ardbeg         4        1     4       4      0     0    2
0
#> Ardmore        2        2     2       0      0     1    1
1
#> ArranIsleOf    2        3     1       1      0     1    1
1
#> Auchentoshan   0        2     0       0      0     1    1
0
#> Auchroisk       2        3     1       0      0     2    1
2
#> Aultmore        2        2     1       0      0     1    0
0
#> Balblair        2        3     2       1      0     0    2
0
#> Balmenach       4        3     2       0      0     2    1
3
#> Belvenie        3        2     1       0      0     3    2
1
#> BenNevis        4        2     2       0      0     2    2
0
#> Benriach        2        2     1       0      0     2    2
0
#> Benrinnes       3        2     2       0      0     3    1
1
#> Benromach       2        2     2       0      0     2    2
1
#> Bladnoch         1        2     1       0      0     0    1
1
#> BlairAthol      2        2     2       0      0     1    2
2
#> Bowmore          2        2     3       1      0     2    2
1
#> Bruichladdich   1        1     2       2      0     2    2
1

```

#> Bunnahabhain	1	2	1	1	0	1	1
1							
#> Caol Ila	3	1	4	2	1	0	2
0							
#> Cardhu	1	3	1	0	0	1	1
0							
#> Clynelish	3	2	3	3	1	0	2
0							
#> Craigallechie	2	2	2	0	1	2	2
1							
#> Craigganmore	2	3	2	1	0	0	1
0							
#> Dailuaine	4	2	2	0	0	1	2
2							
#> Dalmore	3	2	2	1	0	1	2
2							
#> Dalwhinnie	2	2	2	0	0	2	1
0							
#> Deanston	2	2	1	0	0	2	1
1							
#> Dufftown	2	3	1	1	0	0	0
0							
#> Edradour	2	3	1	0	0	2	1
1							
#> GlenDeveronMacduff	2	3	1	1	1	1	1
2							
#> GlenElgin	2	3	1	0	0	2	1
1							
#> GlenGarioch	2	1	3	0	0	0	3
1							
#> GlenGrant	1	2	0	0	0	1	0
1							
#> GlenKeith	2	3	1	0	0	1	2
1							
#> GlenMoray	1	2	1	0	0	1	2
1							
#> GlenOrd	3	2	1	0	0	1	2
1							
#> GlenScotia	2	2	2	2	0	1	0
1							
#> GlenSpey	1	3	1	0	0	0	1
1							
#> Glenallachie	1	3	1	0	0	1	1
0							
#> Glendronach	4	2	2	0	0	2	1
4							
#> Glendullan	3	2	1	0	0	2	1
2							
#> Glenfarclas	2	4	1	0	0	1	2
3							

		3	2	2	0	0	2	2
#> RoyalLochnagar	2							
#> Scapa	1	2	2	1	1	0	2	1
#> Speyburn	0	2	4	1	0	0	2	1
#> Speyside	1	2	2	1	0	0	1	0
#> Springbank	1	2	2	2	2	0	2	2
#> Strathisla	2	2	2	1	0	0	2	2
#> Strathmill	0	2	3	1	0	0	0	2
#> Talisker	0	4	2	3	3	0	1	3
#> Tamdhu	1	1	2	1	0	0	2	0
#> Tamnavulin	0	1	3	2	0	0	0	2
#> Teaninich	0	2	2	2	1	0	0	2
#> Tobermory	0	1	1	1	0	0	1	0
#> Tomatin	1	2	3	2	0	0	2	2
#> Tomintoul	1	0	3	1	0	0	2	2
#> Tormore	1	2	2	1	0	0	1	0
#> Tullibardine	1	2	3	0	0	1	0	2
#>		Nutty	Malty	Fruity	Floral			
#> Aberfeldy		2	2	2	2			
#> Aberlour		2	3	3	2			
#> AnCnoc		2	2	3	2			
#> Ardbeg		1	2	1	0			
#> Ardmore		2	3	1	1			
#> ArranIsleOf	0	1	1	1	2			
#> Auchentoshan		2	2	3	3			
#> Auchroisk		2	2	2	1			
#> Aultmore		2	2	2	2			
#> Balblair		2	1	2	1			
#> Balmenach		3	0	1	2			
#> Belvenie	0	2	2	2	2			
#> BenNevis		2	2	2	2			
#> Benriach	0	2	3	2	2			
#> Benrinnes		2	3	2	2			
#> Benromach		2	2	2	2			
#> Bladnoch		0	2	2	3			

#> BlairAthol	2	2	2	2
#> Bowmore	1	1	1	2
#> Bruichladdich	2	2	2	2
#> Bunnahabhain	1	2	2	3
#> Caol Ila	2	1	1	1
#> Cardhu	2	2	2	2
#> Clynelish	1	1	2	0
#> Craigallechie	2	2	1	4
#> Craigganmore	2	2	2	2
#> Dailuaine	2	2	2	1
#> Dalmore	1	2	3	1
#> Dalwhinnie	1	2	2	2
#> Deanston	1	3	2	1
#> Dufftown	1	2	2	2
#> Edradour	4	2	2	2
#> GlenDeveronMacduff	0	2	0	1
#> GlenElgin	1	1	2	3
#> GlenGarioch	0	2	2	2
#> GlenGrant	2	1	2	1
#> GlenKeith	2	1	2	1
#> GlenMoray	2	2	2	4
#> GlenOrd	1	2	2	2
#> GlenScotia	2	2	1	1
#> GlenSpey	1	2	0	2
#> Glenallachie	1	2	2	2
#> Glendronach	2	2	2	0
#> Glendullan	1	2	3	2
#> Glenfarclas	2	3	2	2
#> Glenfiddich	0	2	2	2
#> Glengoyne	2	2	3	2
#> Glenkinchie	0	2	2	2
#> Glenlivet	1	2	2	3
#> Glenlossie	1	2	2	2
#> Glenmorangie	2	1	2	2
#> Glenrothes	1	2	2	0
#> Glenturret	2	2	1	2
#> Highland Park	1	2	1	1
#> Inchgower	1	2	1	2
#> Isle of Jura	2	1	1	1
#> Knochando	2	1	2	2
#> Lagavulin	1	1	1	0
#> Laphroig	1	1	0	0
#> Linkwood	0	1	3	2
#> Loch Lomond	1	2	1	2
#> Longmorn	3	3	2	3
#> Macallan	2	2	3	1
#> Mannochmore	2	1	2	2
#> Miltonduff	2	1	1	2
#> Mortlach	2	1	2	2
#> Oban	2	2	2	0

```

#> OldFettercairn      2     3     1     1
#> OldPulteney        2     2     2     2
#> RoyalBrackla       0     2     3     2
#> RoyalLochnagar     2     2     3     1
#> Scapa               2     2     2     2
#> Speyburn            0     2     1     2
#> Speyside            2     2     2     2
#> Springbank          2     1     0     1
#> Strathisla          3     3     3     2
#> Strathmill          2     1     3     2
#> Talisker            1     2     2     0
#> Tamdhu              1     2     2     2
#> Tamnavulin          2     1     2     3
#> Teaninich           0     0     2     2
#> Tobermory           1     2     2     2
#> Tomatin              1     2     0     1
#> Tomintoul            1     2     1     2
#> Tormore              2     1     0     0
#> Tullibardine         1     2     2     1

```

EDA & Scaling

Cek missing values

```

# cek NA
anyNA(whisky_clean)

#> [1] FALSE

```

Cek skala antar variabel

```

summary(whisky_clean)

#>      Body        Sweetness        Smoky        Medicinal
#> Min.   :0.00   Min.   :1.000   Min.   :0.000   Min.   :0.0000
#> 1st Qu.:2.00  1st Qu.:2.000  1st Qu.:1.000  1st Qu.:0.0000
#> Median :2.00  Median :2.000  Median :1.000  Median :0.0000
#> Mean   :2.07  Mean   :2.291  Mean   :1.535  Mean   :0.5465
#> 3rd Qu.:2.00  3rd Qu.:3.000  3rd Qu.:2.000  3rd Qu.:1.0000
#> Max.   :4.00  Max.   :4.000  Max.   :4.000  Max.   :4.0000
#>      Tobacco       Honey        Spicy        Winey
#> Min.   :0.0000  Min.   :0.000  Min.   :0.000  Min.   :0.0000
#> 1st Qu.:0.0000  1st Qu.:1.000  1st Qu.:1.000  1st Qu.:0.0000
#> Median :0.0000  Median :1.000  Median :1.000  Median :1.0000
#> Mean   :0.1163  Mean   :1.244  Mean   :1.384  Mean   :0.9767
#> 3rd Qu.:0.0000  3rd Qu.:2.000  3rd Qu.:2.000  3rd Qu.:1.0000
#> Max.   :1.0000  Max.   :4.000  Max.   :3.000  Max.   :4.0000
#>      Nutty        Malty        Fruity        Floral
#> Min.   :0.000  Min.   :0.000  Min.   :0.000  Min.   :0.000
#> 1st Qu.:1.000  1st Qu.:1.000  1st Qu.:1.000  1st Qu.:1.000
#> Median :2.000  Median :2.000  Median :2.000  Median :2.000

```

```
#> Mean      :1.465    Mean      :1.802    Mean      :1.802    Mean      :1.698
#> 3rd Qu.:2.000    3rd Qu.:2.000    3rd Qu.:2.000    3rd Qu.:2.000
#> Max.     :4.000    Max.     :3.000    Max.     :3.000    Max.     :4.000
```

Note: jika bertemu data dengan skala yang berbeda tetap harus discale

Pada data whisky, apakah skala nilai antar variable berbeda? Apakah perlu dilakukan scaling? Tidak. Karena data whisky sudah punya skala yang sama di tiap kolomnya yaitu skor 0-4

K-means

```
formula = kmeans(x, centers)
```

Parameter:

- `x`: dataset
- `centers`: banyaknya centroid k (banyaknya kelompok yang diinginkan)

Note: perlu dilakukan `set.seed()` karena terdapat random initialization pada tahap awal k-means

```
RNGkind(sample.kind = "Rounding")
set.seed(100)

whiz_cluster <- kmeans(whisky_clean, centers = 3) #kita mau coba buat 3
cluster
```

Hasil k-means:

- 1 Banyaknya observasi pada tiap cluster

```
# jumlah data tiap cluster
whiz_cluster$size

#> [1] 34 11 41
```

- 2 Letak pusat cluster/centroid, biasa digunakan untuk profiling cluster

```
# Letak pusat cluster atau centroid
whiz_cluster$centers

#>           Body Sweetness   Smoky Medicinal   Tobacco   Honey   Spicy
Winey      2.500000  2.323529  1.588235  0.1764706  0.05882353  1.8823529  1.647059
1.6764706
#> 2 2.909091  1.545455  2.909091  2.7272727  0.45454545  0.4545455  1.454545
0.5454545
#> 3 1.487805  2.463415  1.121951  0.2682927  0.07317073  0.9268293  1.146341
0.5121951
#>           Nutty   Malty   Fruity   Floral
#> 1 1.823529  2.088235  1.911765  1.7058824
```

```
#> 2 1.545455 1.454545 1.181818 0.5454545  
#> 3 1.146341 1.658537 1.878049 2.0000000
```

- 3 Label cluster untuk tiap observasi

```
# hasil clustering (Label cluster untuk tiap observasi)  
as.data.frame(whiz_cluster$cluster)
```

```
#>                               whiz_cluster$cluster  
#> Aberfeldy                      1  
#> Aberlour                        1  
#> AnCnoc                           3  
#> Ardbeg                            2  
#> Ardmore                          1  
#> ArranIsleOf                     3  
#> Auchentoshan                    3  
#> Auchroisk                        1  
#> Aultmore                         3  
#> Balblair                         3  
#> Balmenach                        1  
#> Belvenie                         1  
#> BenNevis                         1  
#> Benriach                         3  
#> Benrinnnes                       1  
#> Benromach                        1  
#> Bladnoch                          3  
#> BlairAthol                      1  
#> Bowmore                           1  
#> Bruichladdich                   1  
#> Bunnahabhain                    3  
#> Caol Ila                          2  
#> Cardhu                            3  
#> Clynelish                         2  
#> Craigallechie                    1  
#> Craigganmore                     3  
#> Dailuaine                        1  
#> Dalmore                           1  
#> Dalwhinnie                       3  
#> Deanston                          1  
#> Dufftown                          3  
#> Edradour                          1  
#> GlenDeveronMacduff              3  
#> GlenElgin                         3  
#> GlenGarioch                       3  
#> GlenGrant                         3  
#> GlenKeith                         3  
#> GlenMoray                         3  
#> GlenOrd                            1  
#> GlenScotia                        2  
#> GlenSpey                          3  
#> Glenallachie                      3
```

```

#> Glendronach           1
#> Glendullan            1
#> Glenfarclas           1
#> Glenfiddich          3
#> Glengoyne             3
#> Glenkinchie           3
#> Glenlivet              1
#> Glenlossie             3
#> Glenmorangie          3
#> Glenrothes             1
#> Glenturret              1
#> Highland Park           1
#> Inchgower              3
#> Isle of Jura            2
#> Knochando              1
#> Lagavulin              2
#> Laphroig                2
#> Linkwood                3
#> Loch Lomond              3
#> Longmorn                1
#> Macallan                1
#> Mannochmore             3
#> Miltonduff              3
#> Mortlach                1
#> Oban                     2
#> OldFettercairn          1
#> OldPulteney              2
#> RoyalBrackla            3
#> RoyalLochnagar           1
#> Scapa                     1
#> Speyburn                 3
#> Speyside                  3
#> Springbank                2
#> Strathisla                1
#> Strathmill                3
#> Talisker                   2
#> Tamdhu                     3
#> Tamnavulin                3
#> Teaninich                  3
#> Tobermory                  3
#> Tomatin                     1
#> Tomintoul                  3
#> Tormore                     3
#> Tullibardine               3

```

- ⚡ Banyaknya pengulangan (iterasi) algoritma k-means sampai dihasilkan cluster yang stabil

```

# berapa kali pengulangan sampai menghasilkan kelompok yang stabil
whiz_cluster$iter

```

```
#> [1] 2
```

Goodness of Fit

Kebaikan hasil clustering dapat dilihat dari 3 nilai:

1. Within Sum of Squares (\$withinss): jumlah jarak kuadrat dari tiap observasi ke centroid tiap cluster.
2. Between Sum of Squares (\$betweenss): jumlah jarak kuadrat terbobot dari tiap centroid ke rata-rata global. Dibobotkan berdasarkan banyaknya observasi pada cluster.
3. Total Sum of Squares (\$totss): jumlah jarak kuadrat dari tiap observasi ke rata-rata global.

```
# cek nilai wss
whiz_cluster$withinss

#> [1] 179.20588 65.45455 202.53659

whiz_cluster$tot.withinss

#> [1] 447.197

# cek nilai bss
whiz_cluster$betweenss

#> [1] 218.6402

# cek nilai tss
whiz_cluster$totss

#> [1] 665.8372

# cek rasio bss/tss
whiz_cluster$betweenss / whiz_cluster$totss

#> [1] 0.3283688
```

Insight: cluster yang dihasilkan ternyata kurang bagus, karena rasio bss/tss jauh dari 1

Clustering yang baik:

- WSS semakin **rendah**: jarak observasi di 1 kelompok yang sama semakin rendah, artinya tiap cluster memiliki karakteristik yang semakin mirip

$$\frac{BSS}{TSS} \approx 1$$

karena kelompok hasil clustering semakin mewakili persebaran data yang sesungguhnya

Sekarang, kita coba modeling k-means dengan k sebesar mungkin, misalnya 80. Kemudian cek WSS dan BSS/TSSnya, kita akan lihat apakah clustering yang terbentuk dapat dikatakan ideal atau tidak

```

# buat model kmeans
RNGkind(sample.kind = "Rounding")
set.seed(100)

whiz_cluster80 <- kmeans(whisky_clean, centers = 80)

# hitung rasio bss/totss
whiz_cluster80$betweenss/whiz_cluster80$totss

#> [1] 0.9902379

```

💡 Insight: Ternyata semakin besar nilai k (jumlah kelompok) yang dihasilkan, maka bss/tss nya semakin mendekati 1

```

# cek nilai wss
whiz_cluster80$withinss

#> [1] 1.5 0.0 0.0 2.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
#> [20] 1.5 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0
#> [39] 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.5 0.5 0.0
#> [58] 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.5
#> [77] 0.0 0.0 0.0 0.0

```

💡 Insight: WSS nya jadi banyak yang nol, karena banyak cluster yang anggotanya cuma 1 (centroid itu sendiri). Namun, kita tidak mau cluster yang seperti ini karena tujuan clustering hilang

Berdasarkan percobaan diatas, maka:

- Pemilihan k (banyaknya cluster) sangat mempengaruhi performa clustering
- Pemilihan banyak k akan membuat bss/tss bagus, namun tidak representatif karena bisa jadi ada satu cluster yang beranggotakan satu observasi sendiri (tujuan clustering tidak tercapai)

Pemilihan K Optimum

Semakin tinggi k :

- WSS semakin mendekati 0
- BSS semakin mendekati TSS (atau BSS/TSS mendekati 1)

Kalau begitu apakah kita selalu memilih $k = \text{banyak observasi}$? Bagaimana menentukan k optimum?

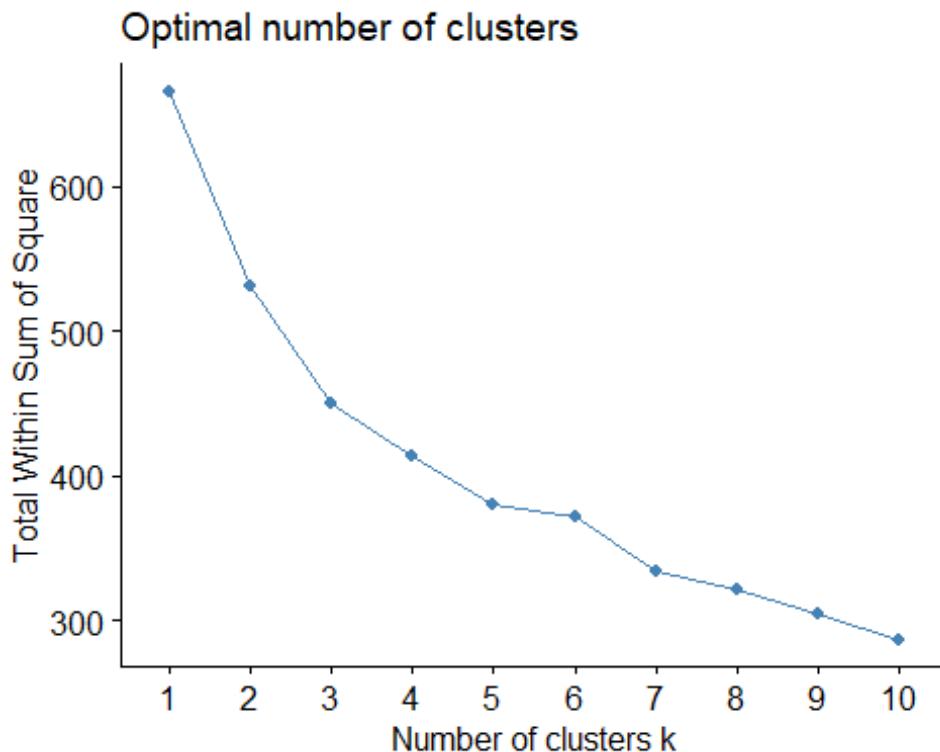
- Kebutuhan dari segi bisnis, dibutuhkan menjadi berapa kelompok
- Secara objektif: Elbow method, visualisasi dengan fviz_nbclust()

Elbow Method > Elbow Plot merupakan plot antara banyak klaster dengan total dari simpangan/variasi per kluster (total WSS).

Note: Banyak klaster yang dipilih adalah bagian “siku” atau titik dimana terdapat penurunan yang tajam sebelum titik tersebut dan disusul penurunan yang tidak tajam setelah titik tersebut. Hal ini karena penambahan jumlah klaster tidak membawa pengaruh banyak atas variasi yang ada di dalam klaster tersebut.

```
library(factoextra)

fviz_nbclust(
  x = whisky_clean, #data untuk clustering
  FUNcluster = kmeans, #algoritma kmeans
  method = "wss" #berdasarkan wss
)
```



Pilih nilai k dimana ketika k ditambah, penurunan total within sum of squares tidak terlalu drastis (atau dapat dikatakan sudah melandai)

- 💡 K optimum dari data whisky adalah 5

Kontruksi kembali k-means clustering menggunakan k optimum

```
RNGkind(sample.kind = "Rounding")
set.seed(100)

# k-means dengan k optimum
```

```

whisky_cluster_op <- kmeans(whisky_clean, centers = 5)

# WSS
whisky_cluster_op$tot.withinss #5 cluster

#> [1] 379.0445

whiz_cluster$tot.withinss #3 cluster

#> [1] 447.197

# BSS/TSS
whisky_cluster_op$betweenss / whisky_cluster_op$totss

#> [1] 0.4307249

# Cek berapa besar tiap cluster
whisky_cluster_op$size

#> [1] 21 6 22 14 23

```

Interpretation: Cluster Profiling

Membuat kolom baru yang berisikan informasi label dari cluster yang terbentuk menggunakan k optimum

```

whisky_clean$cluster <- as.factor(whisky_cluster_op$cluster)

# Cek head data
head(whisky_clean)

#>          Body Sweetness Smoky Medicinal Tobacco Honey Spicy Winey Nutty
#> Aberfeldy    2        2     2         0      0     2     1     2     2
#> Aberlour     3        3     1         0      0     4     3     2     2
#> AnCnoc       1        3     2         0      0     2     0     0     2
#> Ardbeg        4        1     4         4      0     0     2     0     1
#> Ardmore       2        2     2         0      0     1     1     1     2
#> ArranIsleOf   2        3     1         1      0     1     1     1     0
#>             Malty Fruity Floral cluster
#> Aberfeldy    2        2     2         1
#> Aberlour     3        3     2         1
#> AnCnoc       2        3     2         3
#> Ardbeg        2        1     0         2
#> Ardmore       3        1     1         4
#> ArranIsleOf   1        1     2         5

```

Grouping data based on cluster label

Melakukan grouping berdasarkan cluster yang terbentuk, untuk mengetahui karakteristik dari masing-masing cluster

```
as.data.frame(whisky_cluster_op$centers)
```

```

#>      Body Sweetness   Smoky Medicinal Tobacco Honey Spicy
#> 1 2.809524 2.428571 1.523810 0.04761905 0.0000000 1.8571429 1.6190476
#> 2 3.666667 1.500000 3.666667 3.33333333 0.6666667 0.1666667 1.6666667
#> 3 1.272727 2.363636 1.000000 0.18181818 0.0000000 0.7727273 0.7272727
#> 4 1.857143 2.000000 2.000000 1.14285714 0.2142857 1.2142857 1.3571429
#> 5 1.869565 2.478261 1.217391 0.26086957 0.1304348 1.4347826 1.7391304
#>      Winey Nutty Malty Fruity Floral
#> 1 2.0476190 2.095238 2.095238 2.190476 1.6190476
#> 2 0.5000000 1.166667 1.333333 1.166667 0.1666667
#> 3 0.4090909 1.500000 1.772727 2.000000 2.2272727
#> 4 0.8571429 1.642857 1.785714 1.000000 1.0714286
#> 5 0.7391304 0.826087 1.695652 1.913043 2.0434783

# melakukan profiling cluster dari data asli (supaya nantinya jika ketemu dengan perlu data yg perlu discaling sebelum kmeans nya , interpretasinya tidak salah)
whisky_centroid <- whisky_clean %>%
  group_by(cluster) %>%
  summarise_all(mean)
whisky_centroid

#> # A tibble: 5 × 13
#>   cluster Body Sweetness Smoky Medicinal Tobacco Honey Spicy Winey Nutty
#>   <dbl>   <dbl>     <dbl>   <dbl>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
#> 1 1       2.81     2.43    1.52    0.0476    0     1.86    1.62    2.05    2.10
#> 2 2       3.67     1.5     3.67    3.33     0.667   0.167   1.67    0.5     1.17
#> 3 3       1.27     2.36    1       0.182     0     0.773   0.727   0.409   1.5
#> 4 4       1.86     2       2       1.14     0.214   1.21    1.36    0.857   1.64
#> 5 5       1.87     2.48    1.22    0.261     0.130   1.43    1.74    0.739   0.826
#> # 2 more variables: Fruity <dbl>, Floral <dbl>

```

⚙️ Mempermudah profiling: tabel yang menampilkan cluster dengan nilai terendah dan tertinggi untuk masing-masing karakteristik whisky

```

library(tidyr)

whisky_centroid %>%
  pivot_longer(-cluster) %>%
  group_by(name) %>%
  summarize(
    kelompok_min = which.min(value),
    kelompok_max = which.max(value))

```

```
#> # A tibble: 12 × 3
#>   name      kelompok_min kelompok_max
#>   <chr>          <int>        <int>
#> 1 Body              3            2
#> 2 Floral             2            3
#> 3 Fruity             4            1
#> 4 Honey              2            1
#> 5 Malty              2            1
#> 6 Medicinal          1            2
#> 7 Nutty              5            1
#> 8 Smoky              3            2
#> 9 Spicy              3            5
#> 10 Sweetness          2            5
#> 11 Tobacco             1            2
#> 12 Winey              3            1
```

 Profiling tiap cluster

Cluster 1 :

- Paling tinggi di cita rasa : Fruity, Honey, Nutty, winey
- Paling rendah di cita rasa : Medicinal, Tobacco
- Label : Fruity whisky

Cluster 2 :

- Paling tinggi di cita rasa : Body, Medicinal, Smoky, Tobacco
- Paling rendah di cita rasa : Floral, Malty, Sweetness, Honey
- Label : Bitter whisky

Cluster 3 :

- Paling tinggi di cita rasa : Floral
- Paling rendah di cita rasa : Body
- Label : Floral whisky

Cluster 4 :

- Paling tinggi di cita rasa : -
- Paling rendah di cita rasa : Fruity
- Label : Mediocre whisky

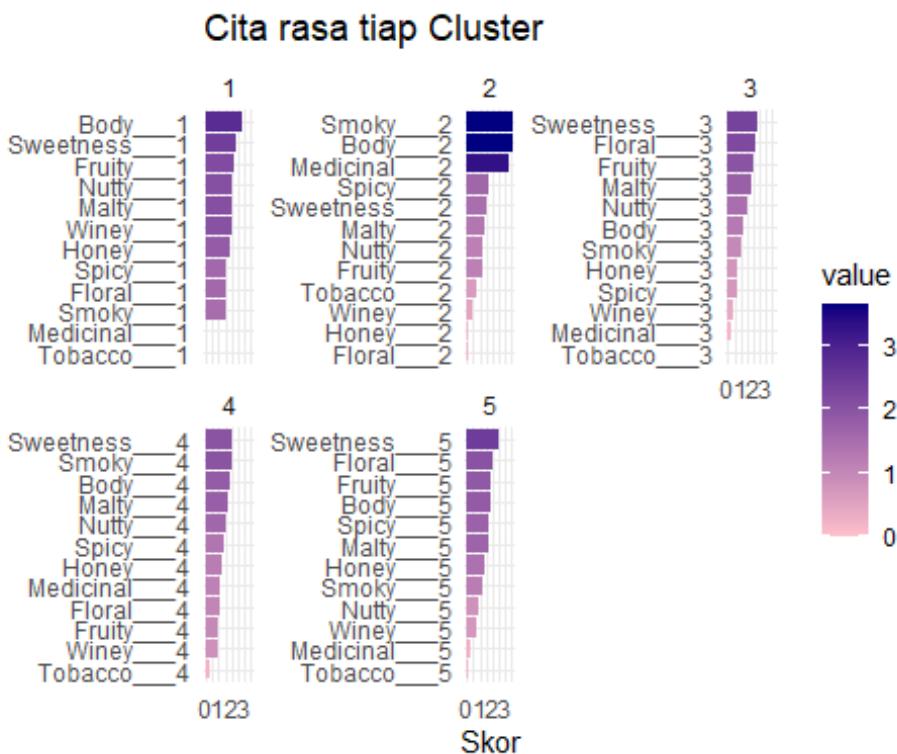
Cluster 5 :

- Paling tinggi di cita rasa : Spicy, Sweetness

- Paling rendah di cita rasa :-
- Label : Spicy & Sweetness whisky

❖ Untuk mempermudah profiling kita akan membentuk grafik yang mengurutkan cita rasa pada setiap cluster

```
whisky_centroid %>%
  pivot_longer(-cluster) %>%
  ggplot(aes(x = value, y = tidytext::reorder_within(name, value, cluster),
  fill = value)) +
  geom_col() +
  scale_fill_gradient(low = "pink", high = "navy") +
  facet_wrap(~cluster, scales = "free_y") +
  theme_minimal() +
  labs(title = "Cita rasa tiap Cluster",
       y = "",
       x = "Skor")
```



Case: Product Recommendation

Misal ada seorang pelanggan pecinta whisky “Laphroig” datang ke toko kita, namun stok whisky tersebut sedang kosong. Kira-kira whisky apa yang akan kita rekomendasikan?

- Identifikasi whisky “Laphroig” terdapat di cluster berapa

```
# your code here
whisky_clean["Laphroig",]
```

```

#>           Body Sweetness Smoky Medicinal Tobacco Honey Spicy Winey Nutty
Malty
#> Laphroig    4          2      4          4      1      0      0      1      1
1
#>           Fruity Floral cluster
#> Laphroig    0          0      2

```

- Memanggil whisky yang satu cluster dengan "Laphroig"

```

whisky_clean[whisky_clean$cluster==2,]

#>           Body Sweetness Smoky Medicinal Tobacco Honey Spicy Winey Nutty
Malty
#> Ardbeg     4          1      4          4      0      0      2      0      1
2
#> Caol Ila   3          1      4          2      1      0      2      0      2
1
#> Clynelish   3          2      3          3      1      0      2      0      1
1
#> Lagavulin  4          1      4          4      1      0      1      2      1
1
#> Laphroig   4          2      4          4      1      0      0      1      1
1
#> Talisker   4          2      3          3      0      1      3      0      1
2
#>           Fruity Floral cluster
#> Ardbeg     1          0      2
#> Caol Ila   1          1      2
#> Clynelish   2          0      2
#> Lagavulin  1          0      2
#> Laphroig   0          0      2
#> Talisker   2          0      2

```

 Solusi: Silakan rekomendasikan merk whisky lain yang ada di cluster 2