Responsi IF-F

2024-11-20

Responsi

Instruksi

1. Baca soal dengan seksama dan jawab dengan sesuai
2. Kerjakan secara mandiri
3. Waktu mengerjakan adalah 2 jam dan pengumpulan diberi waktu tambahan 10 menit (13.00 - 15.10)
4. Kumpulkan dalam bentuk pdf dengan format penamaan NIM_Nama_Responsi.pdf

Import Library (5 poin)

Import library yang dibutuhkan secara berkala.

```r
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 1.2.0 --

## v broom        1.0.6     v recipes      1.1.0
## v dials        1.3.0     v rsample      1.2.1
## v dplyr        1.1.4     v tibble       3.2.1
## v ggplot2      3.5.1     v tidyr        1.3.1
## v infer        1.0.7     v tune         1.2.1
## v modeldata    1.4.0     v workflows    1.1.4
## v parsnip      1.2.1     v workflowsets 1.1.0
## v purrr        1.0.2     v yardstick    1.3.1

## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x purrr::discard() masks scales::discard()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x recipes::step()  masks stats::step()
## * Learn how to get started at https://www.tidymodels.org/start/
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v lubridate 1.9.3     v stringr   1.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x readr::col_factor() masks scales::col_factor()
## x purrr::discard()    masks scales::discard()
## x dplyr::filter()     masks stats::filter()
## x stringr::fixed()    masks recipes::fixed()
## x dplyr::lag()        masks stats::lag()
## x readr::spec()       masks yardstick::spec()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(here)
```

```
## here() starts at D:/KULIAH IF/SEMESTER 5/PRAK DATA SCIENCE/Praktikum (Practice)/Responsi/responsi_if
```

```r
library(ggplot2)
```

Import Dataset (5 poin)

Import dataset housing1.csv dan housing2.csv yang terlampir di SPADA.

```r
# housing1.csv
path = here('data-raw', 'housing1.csv')
housing1 = read.csv(path)
head(housing1,5)
```

```
##   id SquareFeet Bedrooms Bathrooms
## 1  1       2126        4         1
## 2  2       2459        3         2
## 3  3       1860        2         1
## 4  4       2294        2         1
## 5  5       2130        5         2
```

```r
# housing2.csv
path = here('data-raw', 'housing2.csv')
housing2 = read.csv(path)
head(housing2,5)
```

```
##   id Neighborhood YearBuilt    Price
## 1  1        Rural      1969 215355.3
## 2  2        Rural      1980 195014.2
## 3  3       Suburb      1970 306891.0
## 4  4        Urban      1996 206786.8
## 5  5       Suburb      2001 272436.2
```

Preprocessing Data

Join Table (10 poin)

Gabungkan kedua dataset yang sudah di-import berdasarkan kolom yang sama.

```r
housing_combined = left_join(housing1, housing2, by = 'id')
head(housing_combined, 5)
```

```
##   id SquareFeet Bedrooms Bathrooms Neighborhood YearBuilt    Price
## 1  1       2126        4         1        Rural      1969 215355.3
## 2  2       2459        3         2        Rural      1980 195014.2
## 3  3       1860        2         1       Suburb      1970 306891.0
## 4  4       2294        2         1        Urban      1996 206786.8
## 5  5       2130        5         2       Suburb      2001 272436.2
```

```r
# View(housing_combined)
```

Encoding Data (15 poin)

Pada kolom Neighborhood, tipe datanya masih berupa character. Ubah menjadi factor, lalu tampilkan apa saja levelnya.

```r
housing_combined$Neighborhood = factor(housing_combined$Neighborhood)
class(housing_combined$Neighborhood)
```

```
## [1] "factor"
```

```r
levels(housing_combined$Neighborhood)
```

```
## [1] "Rural"  "Suburb" "Urban"
```

Ubah tiap level menjadi numerik agar bisa dilakukan clustering.

```r
housing_combined$NeighborhoodLevel = as.numeric(housing_combined$Neighborhood)
```

Data Filtering (7 poin)

Karena data rumah terlalu banyak, gunakan data rumah yang dibangun pada tahun 1995-2005 saja.

```r
new_housing = housing_combined %>%
  filter(YearBuilt >= 1995 & YearBuilt <= 2005)
# View(new_housing)
```

Scaling Data (8 poin)

Tiap kolom masih memiliki range yang beragam. Seragamkan range dari tiap kolom yang bertipe numerik (kecuali id).

```r
housing_scaled = new_housing %>%
  select(-id,-Neighborhood) %>%
  scale()

head(housing_scaled,10)
```

```
##        SquareFeet   Bedrooms     Bathrooms     YearBuilt      Price
##   [1,]  0.5053398 -1.3472543 -1.2291536164 -1.257828e+00 -0.2208389
##   [2,]  0.2209671  1.3356189  0.0003199255  3.144059e-01  0.6460503
##   [3,] -1.5286185  1.3356189  0.0003199255  1.257746e+00 -1.6842611
##   [4,]  1.1833259  0.4413279  1.2297934673 -3.144877e-01  2.4034478
##   [5,]  0.7532989 -1.3472543 -1.2291536164  1.257746e+00  0.6952597
##   [6,] -0.3425763  1.3356189 -1.2291536164  6.288527e-01  0.2703595
##   [7,] -1.0708478  0.4413279  1.2297934673 -1.257828e+00 -1.7889936
##   [8,]  1.0081940 -1.3472543  1.2297934673  3.144059e-01  2.5398133
##   [9,] -0.8541003 -0.4529632  0.0003199255  6.288527e-01 -0.8792149
##  [10,] -0.2974928 -1.3472543  0.0003199255 -4.091163e-05 -0.4417277
##        NeighborhoodLevel
##   [1,]       1.226386695
##   [2,]       0.003976095
##   [3,]       1.226386695
##   [4,]       1.226386695
##   [5,]      -1.218434505
##   [6,]      -1.218434505
##   [7,]      -1.218434505
##   [8,]       0.003976095
##   [9,]       0.003976095
##  [10,]      -1.218434505
```

Data Modelling

Tentukan Nilai k (25 poin)

Sebelum membuat model, tentukan jumlah cluster atau nilai k yang paling optimal dengan menggunakan Elbow Method. Jangan lupa beri keterangan pada grafik.

```r
set.seed(123)
wcss = sapply(1:10, function(k) {
  kmeans(housing_scaled, centers = k, nstart = 25)$tot.withinss
})
```

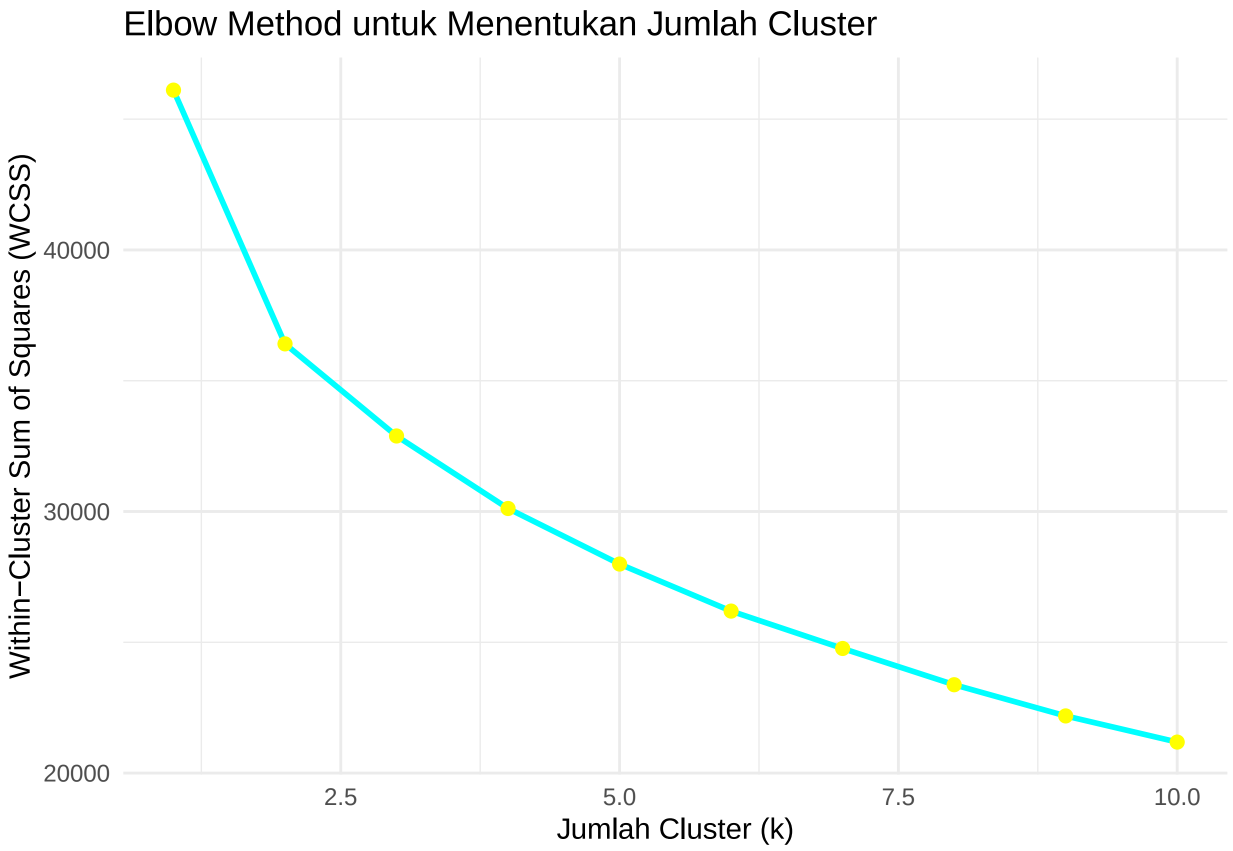
```
## Warning: did not converge in 10 iterations
```

```
## Warning: Quick-TRANSfer stage steps exceeded maximum (= 384300)
```

```
## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
```

```r
elbow_plot = data.frame(Clusters = 1:10, WCSS = wcss)

ggplot(
  elbow_plot,
  aes(
    x = Clusters,
    y = WCSS
    )
  ) +
  geom_line(
    color = "cyan",
    size = 1
  ) +
  geom_point(
    color = "yellow",
    size = 2
    ) +
  labs(
    title = "Elbow Method untuk Menentukan Jumlah Cluster",
    x = "Jumlah Cluster (k)",
    y = "Within-Cluster Sum of Squares (WCSS)"
  ) +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



Buat Cluster (12 poin)

Karena sudah menemukan jumlah cluster yang ideal, buat cluster menggunakan metode k-means. Hasil cluster dimasukkan sebagai kolom baru pada dataset housing sebelum scaling. Ubah tipe data kolom cluster menjadi factor.

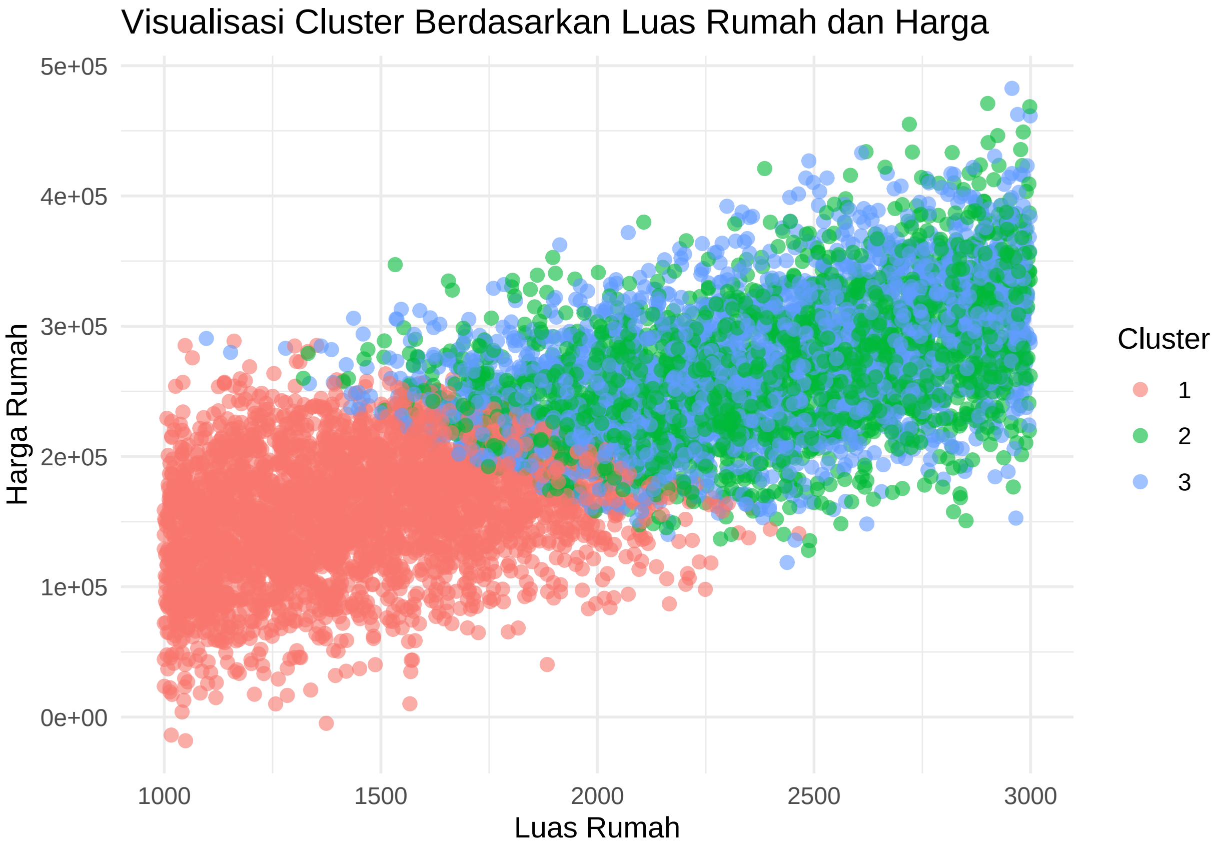
```
set.seed(123)
optimal_k = 3
kmeans_result = kmeans(housing_scaled, centers = optimal_k, nstart = 25)

new_housing$Cluster = as.factor(kmeans_result$cluster)

head(new_housing)
```

```
##   id SquareFeet Bedrooms Bathrooms Neighborhood YearBuilt     Price
## 1  4       2294        2         1        Urban      1996 206786.79
## 2  5       2130        5         2       Suburb      2001 272436.24
## 3 10       1121        5         2        Urban      2004  95961.93
## 4 21       2685        4         3        Urban      1999 405523.83
## 5 26       2437        2         1        Rural      2004 276162.86
## 6 27       1805        5         1        Rural      2002 243985.21
##   NeighborhoodLevel Cluster
## 1                 3       2
## 2                 2       3
## 3                 3       1
## 4                 3       3
## 5                 1       2
## 6                 1       3
```

Visualisasi Data (13 poin)

Visualisasikan cluster dengan menggunakan ggplot2. Buat grafik luas rumah dengan harga, lalu beri warna sesuai cluster. Berikan keterangan pada grafik.

```
ggplot(
  new_housing,
  aes(
    x = SquareFeet,
    y = Price,
    color = Cluster
    )
  ) +
  geom_point(
    alpha = 0.6,
    size = 2
  ) +
  labs(
    title = "Visualisasi Cluster Berdasarkan Luas Rumah dan Harga",
    x = "Luas Rumah",
    y = "Harga Rumah",
    color = "Cluster"
  ) +
  theme_minimal()
```


Visualisasi Cluster Berdasarkan Luas Rumah dan Harga
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
Harga Rumah
1000
1500
2000
2500
3000
Luas Rumah
Cluster
1
2
3