

# WS #20 - k-means clustering

Monday, November 24, 2025

Your Name: \_\_\_\_\_

Names of people you worked with: \_\_\_\_\_

Name one tradition around this week's time off that you grew up with. Will you do it this year?

## Task:

Consider the following observations:

```
# A tibble: 5 x 3
  ID   Age Income
<dbl> <dbl> <dbl>
1     1    25  80000
2     2    30 100000
3     3    40  90000
4     4    30  50000
5     5    40 110000
```

- Find the (Euclidean) distance between person 1 and person 2.
- Find the (Euclidean) distance between person 1 and person 3.

```
kmeans_data |>
  mutate(scale(Age), scale(Income))
```

```
# A tibble: 5 x 5
  ID   Age Income `scale(Age)`[,1] `scale(Income)`[,1]
<dbl> <dbl> <dbl>         <dbl>         <dbl>
1     1    25  80000        -1.19         -0.261
2     2    30 100000        -0.447         0.608
3     3    40  90000         1.04         0.174
4     4    30  50000        -0.447        -1.56
5     5    40 110000         1.04         1.04
```

- Using the scaled data, find the (Euclidean) distance between person 1 and person 2.
- Using the scaled data, find the (Euclidean) distance between person 1 and person 3.
- Who do you believe is closer to person 1 (person 2 or person 3)?

**Solution:**

- a.  $\sqrt{5^2 + 20000^2} = 20,000.00062$
- b.  $\sqrt{15^2 + 10000^2} = 10,000.011$
- c.  $\sqrt{(-1.192 + 0.447)^2 + (-0.261 - 0.608)^2} = 1.145$
- d.  $\sqrt{(-1.192 - 1.044)^2 + (-0.261 - 0.174)^2} = 2.278$
- e. Seemingly, person 1 is farther from person 3 because although their income differences are large in magnitude, they are not particularly different relative to the variability of incomes across the five individuals. Indeed, it is the ages that distinguish person 1 and 3 more strongly.