Unsupervised_K_Means Analysis on the Personality Synthetic Dataset

Jimmy Wu

2025-11-27

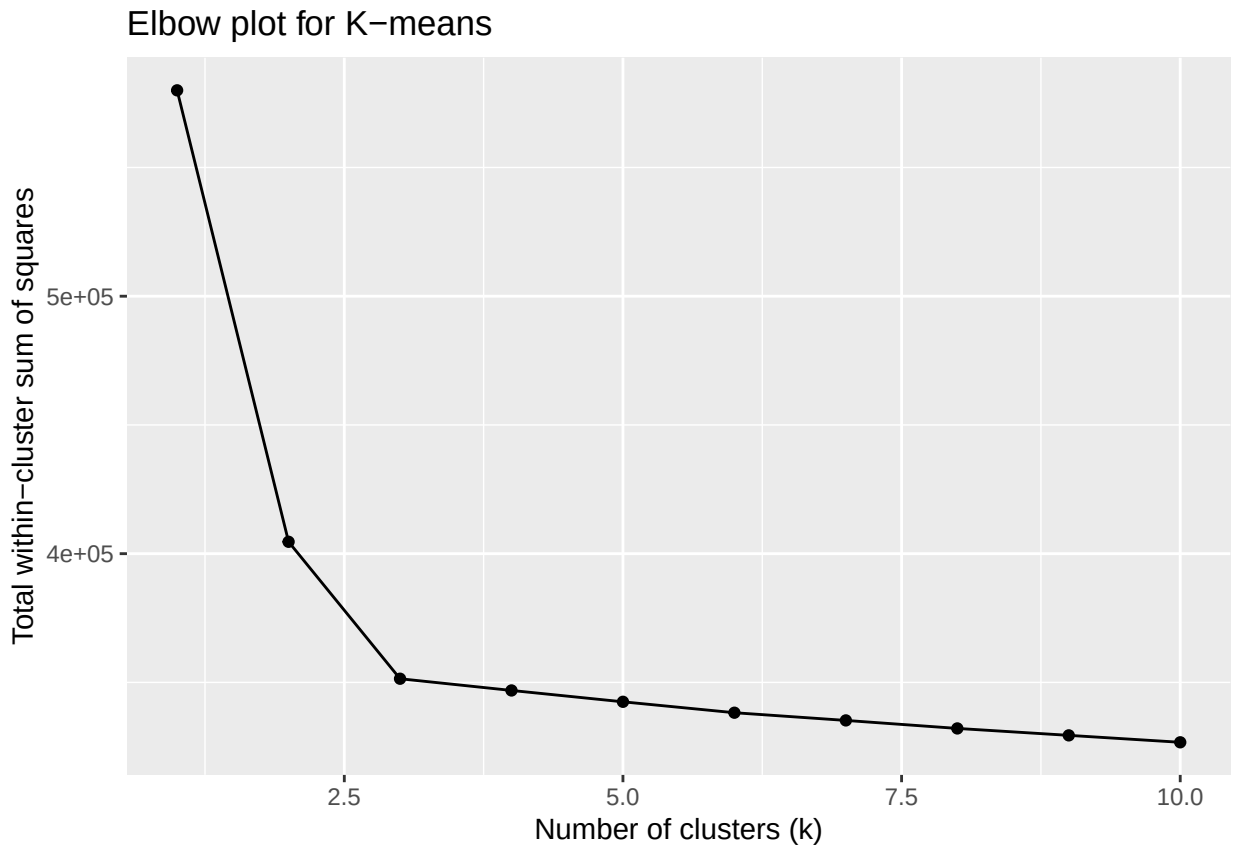
This part of the project answer the following questions: What natural personality clusters exist in the dataset, and what do they represent? Are they meaningful? How stable are they?

Load and Prepare the Dataset

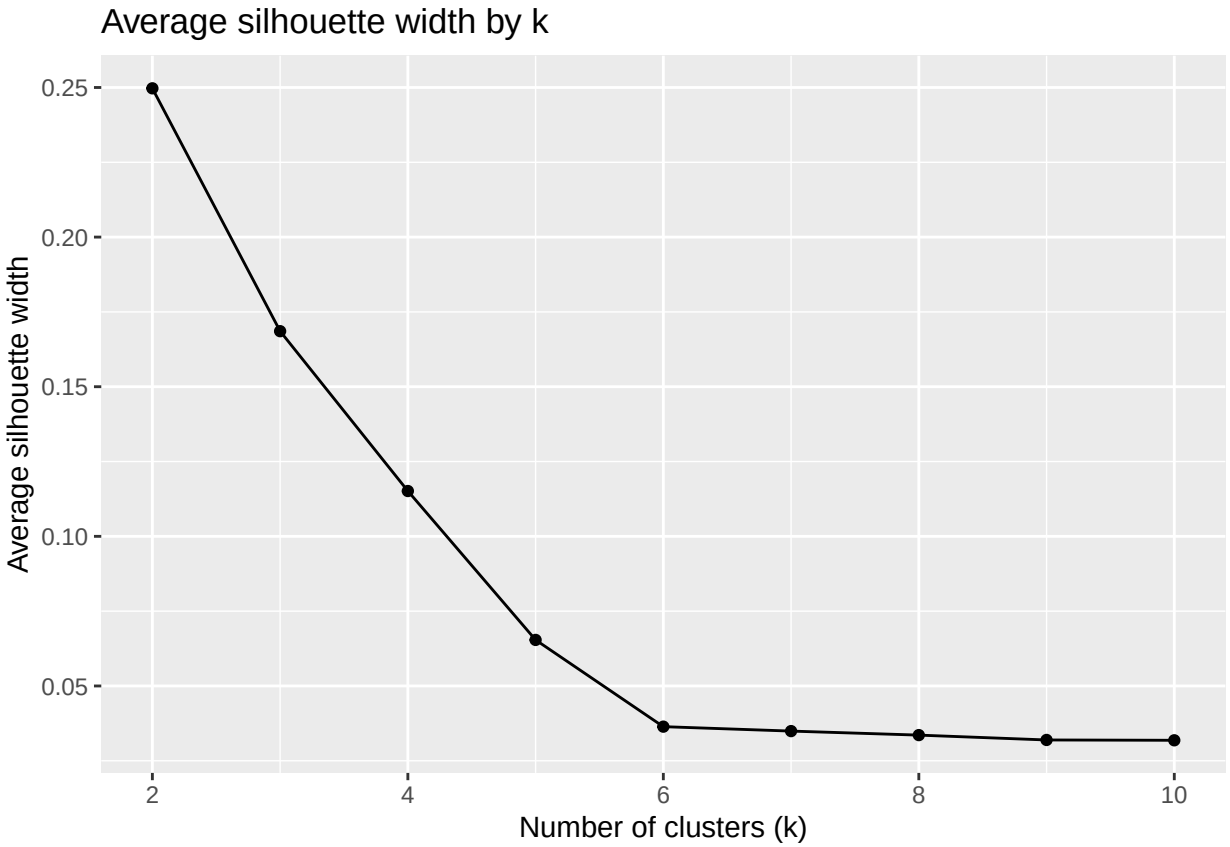
```
dat_raw <- read_csv("../data/raw/personality_synthetic_dataset.csv")

# Keep only numeric columns and remove missing numeric values
dat_num <- dat_raw %>%
  select(where(is.numeric)) %>%
  drop_na()

dat_scaled <- scale(dat_num)
```

Select the value of k

```
# Choose the value of k using elbow method
set.seed(20251126)
max_k <- 10

wss_df <- tibble(
  k = 1:max_k,
  tot_withinss = map_dbl(k, ~kmeans(dat_scaled, centers = .x, nstart = 20)$tot.withinss)
)

# Elbow plot
ggplot(wss_df, aes(x = k, y = tot_withinss)) +
  geom_line() +
  geom_point() +
  labs(title = "Elbow plot for K-means",
       x = "Number of clusters (k)",
       y = "Total within-cluster sum of squares")
```




```r
# Choose the value of k using average silhouette width method
set.seed(20251126)

sil_df <- tibble(
  k = 2:max_k,
  sil_width = map_dbl(k, ~{
    km <- kmeans(dat_scaled, centers = .x, nstart = 20)
    ss <- silhouette(km$cluster, dist(dat_scaled))
    mean(ss[, "sil_width"])
  })
)

ggplot(sil_df, aes(x = k, y = sil_width)) +
  geom_line() +
  geom_point() +
  labs(title = "Average silhouette width by k",
       x = "Number of clusters (k)",
       y = "Average silhouette width")
```


Average silhouette width by k
Average silhouette width
0.25
0.20
0.15
0.10
0.05
2
4
6
8
10
Number of clusters (k)