Unsupervised K-Means Analysis on the Personality Synthetic Dataset

Jimmy Wu

2025-11-27

Introduction

In this part of the project, we apply unsupervised K-means clustering to the *Personality Synthetic Dataset* to investigate whether natural groupings of personality traits emerge without using any labels. Our goals are:

- Identify how many personality clusters exist using common model-selection diagnostics (elbow plot and silhouette analysis).
- Interpret the discovered clusters by summarizing their personality trait profiles.
- Visualize the structure of these clusters using PCA to examine their separability.
- Assess whether the clusters are meaningful, stable, and interpretable in a psychological sense.

Namely, 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)
```

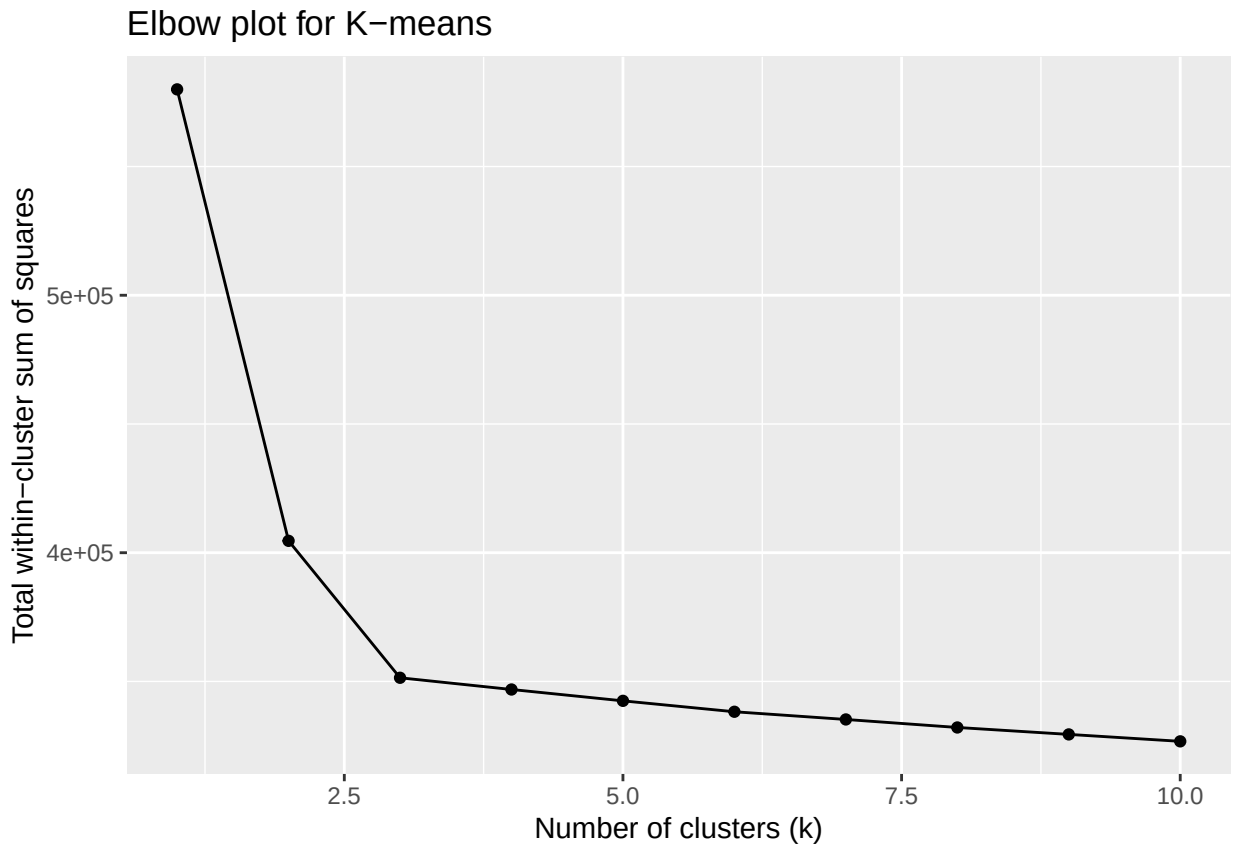
Select the value of k

```
# 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")
```



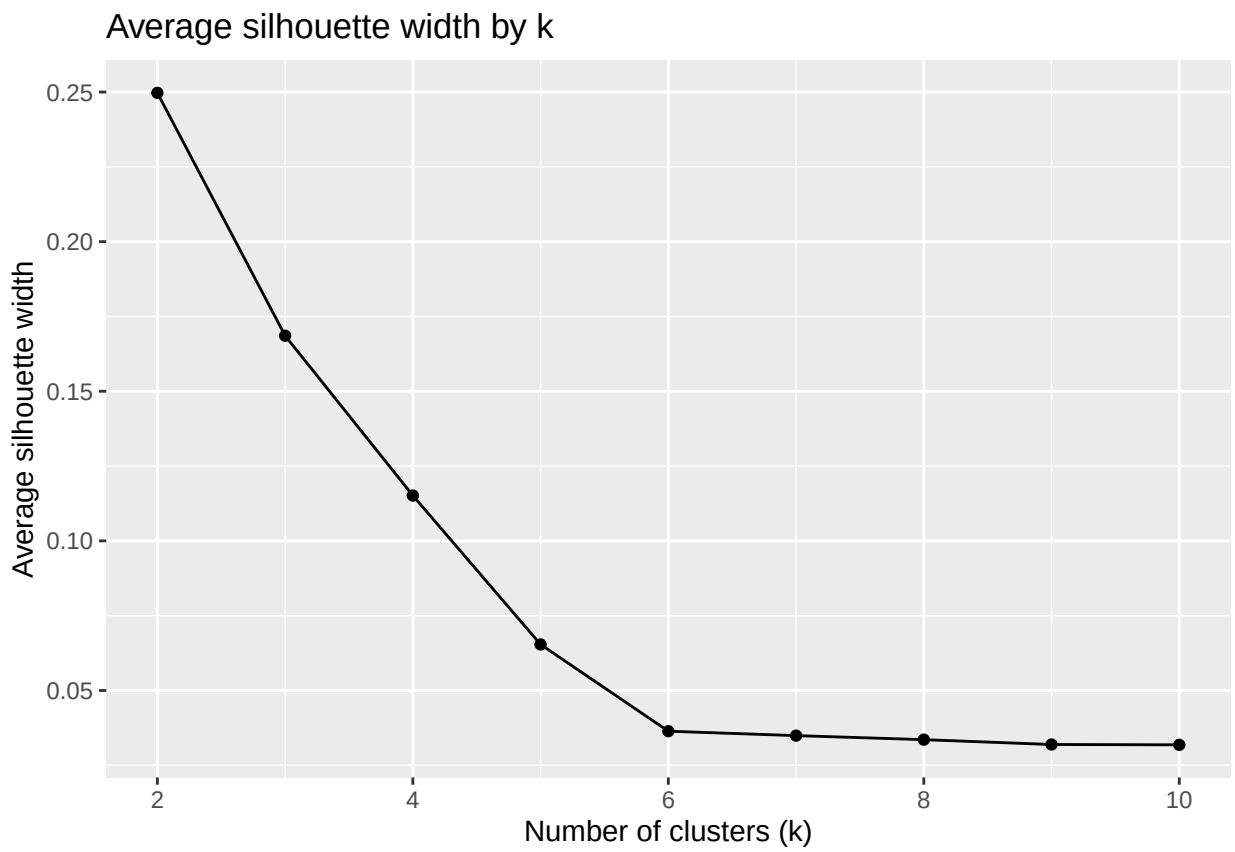
The elbow plot shows the total within-cluster sum of squares decreasing sharply from k = 1 to k = 2, with diminishing improvements for k $\geq$ 3. This indicates that going beyond two clusters adds little additional explanatory power.

```
# 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")
```



The silhouette analysis shows the highest average silhouette width at k = 2, with values decreasing steadily at higher k.

Based on the results from both methods, k = 2 is the optimal number of groups, both statistically and structurally. Therefore, we will specify k = 2 for the K-means model fit.

Fit K-means model

```
set.seed(20251127)

k_best <- 2
km_final <- kmeans(dat_scaled, centers = k_best, nstart = 50)

dat_clusters <- dat_raw %>%
  filter(complete.cases(dat_num)) %>%
```

```
  mutate(cluster = factor(km_final$cluster))

# Cluster sizes
cluster_sizes <- dat_clusters %>%
  count(cluster, name = "n")

kable(cluster_sizes)
```

cluster	n
1	8161
2	11839

```
cluster_summary <- dat_clusters %>%
  group_by(cluster) %>%
  summarise(across(where(is.numeric), mean, .names = "mean_{.col}"),
            .groups = "drop")

print(cluster_summary)
```

```
## # A tibble: 2 x 30
##   cluster mean_social_energy mean_alone_time_preference mean_talkativeness
##   <fct>                <dbl>                      <dbl>              <dbl>
## 1 1                     7.58                       3.33               7.64
## 2 2                     4.08                       6.92               4.07
## # i 26 more variables: mean_deep_reflection <dbl>, mean_group_comfort <dbl>,
## #   mean_party_liking <dbl>, mean_listening_skill <dbl>, mean_empathy <dbl>,
## #   mean_creativity <dbl>, mean_organization <dbl>, mean_leadership <dbl>,
## #   mean_risk_taking <dbl>, mean_public_speaking_comfort <dbl>,
## #   mean_curiosity <dbl>, mean_routine_preference <dbl>,
## #   mean_excitement_seeking <dbl>, mean_friendliness <dbl>,
## #   mean_emotional_stability <dbl>, mean_planning <dbl>, ...
```

K-means produced two large and stable clusters consisting of 8,161 individuals and 11,839 individuals correspondingly, giving us confidence that the personality dataset contains two major profiles. This distribution is not perfectly balanced, but both clusters remain large and well-represented.

The cluster summary table reveals two clear personality types. Cluster 1, on the one hand, is featured with high social energy, talkativeness, party liking, friendliness, and excitement seeking, with low alone-time preference and routine preference. With these features observed, we conclude that this cluster resembles extroverted personality.

Cluster 2, on the other hand, is featured with high alone-time preference, deep reflection, and routine preference, with lower scores on social and energetic traits. With these characteristics displayed, we conclude that this cluster resembles introverted personality.

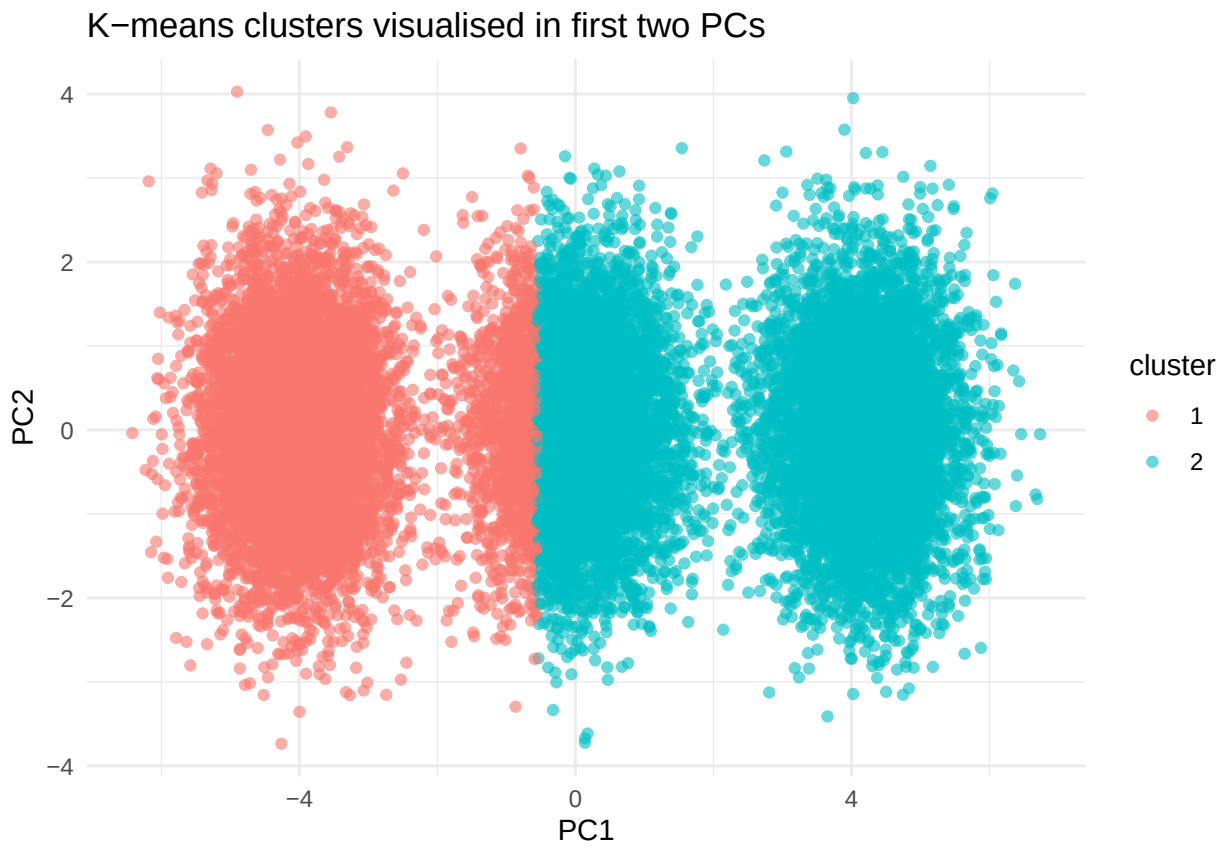
The dataset naturally splits into two psychologically meaningful clusters, corresponding roughly to extroversion vs. introversion. This suggests that K-means has captured real structure that aligns with expectation from general knowledge.

```
pca <- prcomp(dat_scaled, center = TRUE, scale. = TRUE)

pca_df <- as_tibble(pca$x[, 1:2]) %>%
```

```
  rename(PC1 = 1, PC2 = 2) %>%
  mutate(cluster = dat_clusters$cluster)

ggplot(pca_df, aes(x = PC1, y = PC2, colour = cluster)) +
  geom_point(alpha = 0.6) +
  labs(title = "K-means clusters visualised in first two PCs") +
  theme_minimal()
```



The PCA plot shows that Cluster 1 and Cluster 2 form two distinct clouds of points. Separation is strongest along PC1, suggesting this component captures the extroversion–introversion axis. This visualization confirms that the clusters are structurally distinct and interpretable within psychological constructs.

Overall, K-means clustering reveals that the *Personality Synthetic Dataset* contains two dominant personality groups. Model-selection diagnostics, trait summaries, and PCA visualization all support this two-cluster structure. These groups correspond closely to introversion and extroversion, one of the most widely studied axes in personality psychology.