Project 3

TidyTuesday project's 2024-08-06/olympics.csv is the dataset we will use for this project.

```
# Read csv

download.file("https://raw.githubusercontent.com/rfordatascience/tidytuesday/
main/data/2024/2024-08-06/olympics.csv", destfile = "olympics.csv")
olympics <- readr::read_csv("olympics.csv")
```

```
Rows: 271116 Columns: 15
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (10): name, sex, team, noc, games, season, city, sport, event, medal
dbl  (5): id, age, height, weight, year

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
olympics
```

```
# A tibble: 271,116 × 15
      id name     sex     age height weight team  noc   games  year season city
   <dbl> <chr>    <chr> <dbl>  <dbl>  <dbl> <chr> <chr> <chr> <dbl> <chr>  <chr>
 1     1 A Dijia… M        24    180     80 China CHN   1992…  1992 Summer Barc…
 2     2 A Lamusi M        23    170     60 China CHN   2012…  2012 Summer Lond…
 3     3 Gunnar … M        24     NA     NA Denm… DEN   1920…  1920 Summer Antw…
 4     4 Edgar L… M        34     NA     NA Denm… DEN   1900…  1900 Summer Paris
 5     5 Christi… F        21    185     82 Neth… NED   1988…  1988 Winter Calg…
 6     5 Christi… F        21    185     82 Neth… NED   1988…  1988 Winter Calg…
 7     5 Christi… F        25    185     82 Neth… NED   1992…  1992 Winter Albe…
 8     5 Christi… F        25    185     82 Neth… NED   1992…  1992 Winter Albe…
 9     5 Christi… F        27    185     82 Neth… NED   1994…  1994 Winter Lill…
10     5 Christi… F        27    185     82 Neth… NED   1994…  1994 Winter Lill…
# ℹ 271,106 more rows
# ℹ 3 more variables: sport <chr>, event <chr>, medal <chr>
```

Introduction:

This project uses the TidyTuesday "olympics" dataset (2024-08-06), which you can find more information here: https://github.com/rfordatascience/tidytuesday/tree/main/data/2024/2024-08-06

The full CSV contains athlete records across all sports and Olympic Games, including these key columns:

id, name (athlete identifiers) sex, age, height, weight (demographics and anthropometrics) team, noc, games, year, season, city (nationality and edition) sport, event, medal (competition details)

For our analysis, we subset to Taekwondo competitors. We also created a binary medalist flag from the medal column. We will focus on the variables age, height, weight, sex, year, and medalist to uncover underlying athlete "archetypes" and see how those profiles relate to medal success and gender composition across different Olympic years.

Question:

What combinations of factors define athlete profiles most likely to win a medal in Olympic Taekwondo?

Approach:

We will begin by subsetting the full Olympics dataset to only Taekwondo competitors.

Next, we will standardize (scale) age, height, and weight variables. Then we will run a principal components analysis to capture the main axes of variation in athlete body size and age. Reducing to the first two PCs allows us to summarize each athlete in two dimensions while preserving as much variance as possible. We will then apply k-means clustering on the PC scores to let the data itself reveal natural "archetypes" of athletes. Around k value of 5 will be good enough to divide the archetypes. By grouping similar athletes together, we can interpret which combinations of age, height, and weight tend to cluster—and then see which clusters are most associated with medal success.

Finally, we will produce two complementary visualizations: A PCA scatter (PC1 vs. PC2), colored by archetype and faceted by Olympic year, to show how these profiles distribute over time; a compound bar chart—side by side—showing each archetype's medal-winning rate and its gender composition, to directly compare which profiles are most successful and how they break down by sex. This workflow (combining PCA for dimension reduction, clustering for profile discovery, and targeted plots for interpretation) will reveal the combinations of factors that define the athlete profiles most likely to win medals in Olympic Taekwondo.

Analysis:

```
# Data filtering

taekwondo <- olympics |>
  filter(!is.na(height)) |>              # only keep athletes with known height
  filter(!is.na(weight)) |>              # only keep athletes with known weight
  filter(sport == "Taekwondo") |>        # keep only Taekwondo sport
  mutate(
    medalist = case_when(                # add column to track medalist vs not
      is.na(medal) ~ "non-medalist",
     !is.na(medal) ~ "medalist"        # any medals (Gold, Silver, Bronze) count
```

```
    )
  )

taekwondo
```

```
# A tibble: 596 × 16
      id name     sex     age height weight team  noc   games  year season city
   <dbl> <chr>    <chr> <dbl>  <dbl>  <dbl> <chr> <chr> <chr> <dbl> <chr>  <chr>
 1    53 Talaat … M        24    172     58 Egypt EGY   2000…  2000 Summer Sydn…
 2    65 Patimat… F        21    165     49 Azer… AZE   2016…  2016 Summer Rio …
 3   165 Nia Nic… F        20    175     56 Unit… USA   2004…  2004 Summer Athi…
 4   353 Rasul A… M        19    183     74 Kyrg… KGZ   2008…  2008 Summer Beij…
 5   353 Rasul A… M        23    183     74 Kyrg… KGZ   2012…  2012 Summer Lond…
 6   608 Ahmad A… M        20    178     68 Jord… JOR   2016…  2016 Summer Rio …
 7   612 Mohamma… M        28    183     68 Jord… JOR   2012…  2012 Summer Lond…
 8   658 Jaouad … M        23    175     64 Belg… BEL   2016…  2016 Summer Rio …
 9   666 Aziz Ac… M        28    180     68 Germ… GER   2000…  2000 Summer Sydn…
10  1084 Shimaa … F        19    163     54 Egypt EGY   2000…  2000 Summer Sydn…
# ℹ 586 more rows
# ℹ 4 more variables: sport <chr>, event <chr>, medal <chr>, medalist <chr>
```

```
# Perform PCA on age, height, weight

features_scaled <- taekwondo |>
  select(age, height, weight) |>
  scale(center = TRUE, scale = TRUE)

pca_res <- prcomp(features_scaled)

# Extract PC scores and bind metadata

scores <- as_tibble(pca_res$x) |>
  bind_cols(taekwondo |> select(medalist, sex, year))

# Cluster athletes in PC space (k-means with k = 5)

set.seed(123)
k <- 5
km <- kmeans(scores |> select(PC1, PC2), centers = k)
scores <- scores |>
  mutate(archetype = factor(km$cluster))
```

```
# Check profiles of athletes by archetype

profiles <- taekwondo |>
```

```
  bind_cols(archetype = scores$archetype) |>
  group_by(archetype) |>
  summarise(
    n            = n(),
    mean_age     = mean(age),
    sd_age       = sd(age),
    mean_height  = mean(height),
    sd_height    = sd(height),
    mean_weight  = mean(weight),
    sd_weight    = sd(weight)
  )

print(profiles)
```

```
# A tibble: 5 × 8
  archetype     n mean_age sd_age mean_height sd_height mean_weight sd_weight
  <fct>     <int>    <dbl>  <dbl>       <dbl>     <dbl>       <dbl>     <dbl>
1 1           107     27.5   2.66        169.      4.86        58.8      6.42
2 2           117     20.9   2.05        165.      5.67        53.7      4.43
3 3           108     28.2   2.82        182.      4.38        75.3      7.42
4 4           181     21.1   2.21        178.      4.70        68.3      7.01
5 5            83     25.4   3.15        193.      5.07        90.5      9.02
```

```
# Update archetype and labels with profiles data

scores <- scores |>
  mutate(
    archetype = factor(archetype,
      levels = c(1, 2, 3, 4, 5),
      labels = c(
        "Veteran lightweights",
        "Rookie lightweights",
        "Veteran heavyweights",
        "Rookie heavyweights",
        "Super heavy elites"
      )
    ),
    sex = recode(sex,
      F = "Female",
      M = "Male"
    )
  )
```

```
# Prepare visualizations
```

```r
# Visualization 1 — PCA scatter colored by archetype, faceted by year

p_scatter <- ggplot(scores, aes(x = PC1, y = PC2, color = archetype)) +
  geom_point(alpha = 0.7) +
  facet_wrap(~ year) +

  # Set a custom palette for archetypes

  scale_color_manual(values = c(
    "Veteran lightweights"   = "#A23C42",
    "Rookie lightweights"    = "#3B8EA5",
    "Veteran heavyweights"   = "#F6AD55",
    "Rookie heavyweights"    = "#68D391",
    "Super heavy elites"     = "#805AD5"
  )) +

  labs(
    title = "Athlete Archetypes in PCA Space",
    x     = "PC1",
    y     = "PC2",
    color = "Archetype"
  ) +

  theme_minimal() +
  theme(

    # set background colors

    plot.background   = element_rect(fill = "#FEF8F0", color = NA),
    panel.background  = element_rect(fill = "#FEF8F0", color = NA),
    legend.background = element_rect(fill = "#FEF8F0", color = NA),

    # position legend inside at bottom-right

    legend.position     = c(0.95, 0.05),
    legend.justification = c("right", "bottom")
  )
```

```
Warning: A numeric `legend.position` argument in `theme()` was deprecated in
ggplot2
3.5.0.
ℹ Please use the `legend.position.inside` argument of `theme()` instead.
```

```r
# Visualization 2 - Compute summary statistics by archetype

medal_rate <- scores |>
```

```r
  group_by(archetype) |>
  summarise(rate = mean(medalist == "medalist"))

sex_comp <- scores |>
  count(archetype, sex) |>
  group_by(archetype) |>
  mutate(prop = n / sum(n))

# Order archetype levels by ascending medal rate

ordered_levels <- medal_rate |>
  arrange(rate) |>
  pull(archetype)

medal_rate <- medal_rate |>
  mutate(archetype = factor(archetype, levels = ordered_levels))

sex_comp <- sex_comp |>
  mutate(archetype = factor(archetype, levels = ordered_levels))

# Visualization 2a - Bar chart of medal-winning rate by archetype

p_medal <- ggplot(medal_rate, aes(x = archetype, y = rate)) +
  geom_col(fill = "#805AD5") +
  scale_y_continuous(labels = percent_format()) +
  labs(
    title = "Medal Winning Rate by Archetype",
    x     = "Archetype",
    y     = "Medal Rate"
  ) +
  coord_flip() +
  theme_minimal() +
  theme(
    plot.background   = element_rect(fill = "#FEF8F0", color = NA),
    panel.background  = element_rect(fill = "#FEF8F0", color = NA),
    legend.background = element_rect(fill = "#FEF8F0", color = NA)
  )

# Visualization 2b - Stacked bar chart of gender composition by archetype

p_sex <- ggplot(sex_comp, aes(x = archetype, y = prop, fill = sex)) +
  geom_col() +
  scale_fill_manual(values = c(
    "Female" = "#A23C42",
    "Male"   = "#3B8EA5"
  )) +
  labs(
    title = "Gender Composition by Archetype",
```

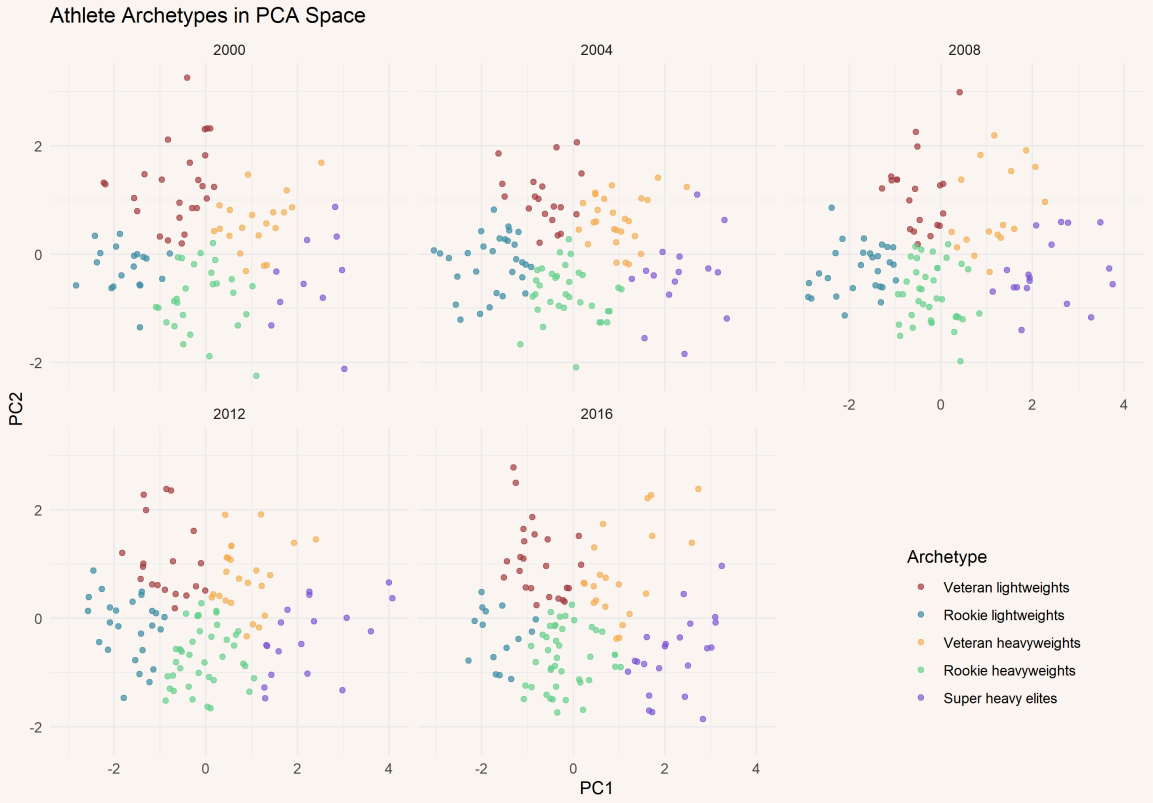
```
    x     = NULL,
    y     = "Proportion",
    fill  = "Sex"
  ) +
  coord_flip() +
  theme_minimal() +
  theme(
    plot.background   = element_rect(fill = "#FEF8F0", color = NA),
    panel.background  = element_rect(fill = "#FEF8F0", color = NA),
    legend.background = element_rect(fill = "#FEF8F0", color = NA),
    axis.text.y       = element_blank(),
    axis.ticks.y      = element_blank()
  )

# Visualization 2 Compound plot — medal rates + gender composition side by side

p_composite <- p_medal + p_sex + plot_layout(ncol = 2)
```

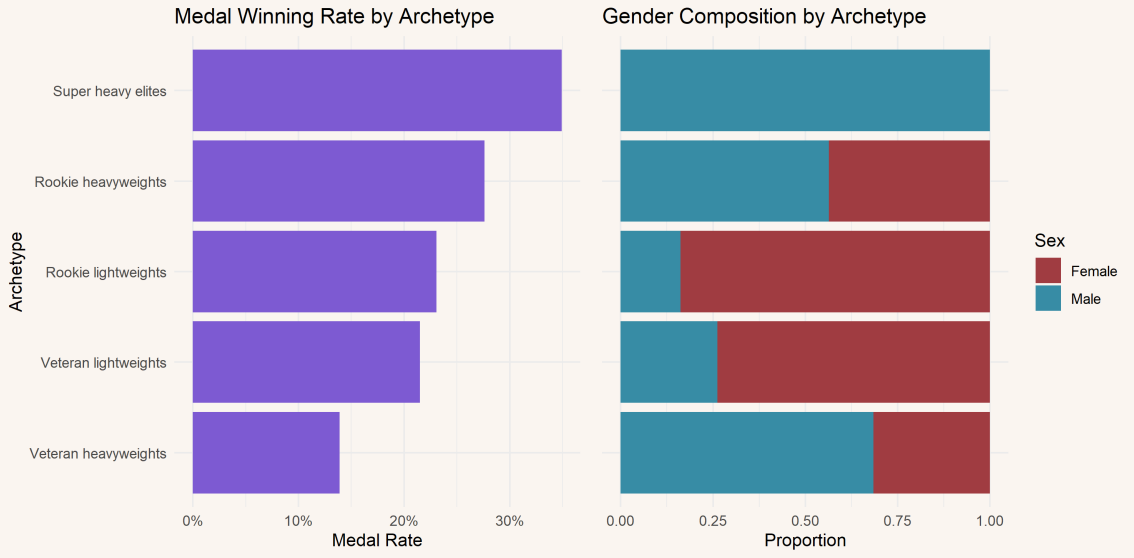
```
# Render visualization 1

print(p_scatter)
```



```
# Render visualization 2

print(p_composite)
```



Discussion:

The analysis reveals that the tallest, heaviest athletes in their mid-20s—the "Super-heavy elites", enjoy the highest medal rate (around 35%). This is followed by the younger, tall "Rookie heavyweights" (about 28%), while both lightweight clusters medal at roughly 20–25% and the "Veteran heavyweights" lag furthest behind (~13%). Heavy-division archetypes are predominantly male (70–80%), whereas lightweight clusters skew female (up to 70% women in the rookie lightweights). Taken together, this suggests that in Olympic Taekwondo a profile combining mid-20s age, extreme height, and high weight, particularly among male competitors, is most likely to win a medal.