Structural Profiles in Powerlifting: A Statistical Analysis

Stefano Librizzi

2025-10-26

Contents

Introduction

Powerlifting is a strength sport that consists of lifting the maximum possible weight in three exercises: **Squat, Bench Press, and Deadlift**. The objective of this analysis is to explore the vast OpenIPF dataset, which

collects all competitions held by federations affiliated with the International Powerlifting Federation (IPF), to identify athlete "profiles" based on their specialization in one of the three lifts.

We will use **Cluster Analysis** techniques to group athletes with similar lifting structures and subsequently analyze how these groups differ by nation of origin and performance (measured by Goodlift Points).

The data is updated as of October 25, 2025.

---

1. Data Preparation: From Raw File to Analysis Dataset

Before starting the analysis, a robust data cleaning and transformation process is necessary. This section documents all the steps taken to convert the raw dataset downloadable on the site Open IPF, here named `openipf_25OTT25.csv`, into the clean file `DC.RData`, making the entire project transparent and reproducible.

1.1 Loading Libraries and Data

First, we load all the necessary libraries and the raw dataset.

```
# Libraries for data manipulation
library(dplyr)
library(tidyr)
library(data.table)
library(lubridate)

# Libraries for visualization
library(ggplot2)
library(patchwork)
library(ggtern)
library(plotly)
library(scales)

# Libraries for statistical analysis
library(cluster)
library(factoextra)
library(FactoMineR)

# Loading the raw dataset using fread for efficiency
openipf <- fread("openipf_25OTT25.csv")

# Converting the Date column to the correct format
openipf$Date <- as.Date(openipf$Date)

# Filtering to include only the most recent competitions (from 2021 onwards)
# to get a current and homogeneous picture of the sport.
Dati <- openipf %>%
  filter(year(Date) > 2020)

head(Dati)
```

```
##              Name    Sex  Event Equipment   Age AgeClass BirthYearClass
```

```
##             <char> <char> <char>    <char> <num>   <char>         <char>
## 1:  Kate MacDonald      F    SBD       Raw    41    40-44          40-49
## 2:       Uyen Pham      F    SBD       Raw    36    35-39          24-39
## 3:        Lily Ngo      F    SBD       Raw    24    24-34          24-39
## 4:     Kat Trinder      F    SBD       Raw    39    35-39          24-39
## 5: Aeva Petranovic      F    SBD       Raw    25    24-34          24-39
## 6:  Melissa Debono      F    SBD       Raw    37    35-39          24-39
##     Division BodyweightKg WeightClassKg Squat1Kg Squat2Kg Squat3Kg Squat4Kg
##       <char>        <num>         <char>    <num>    <num>    <num>    <num>
## 1: Masters 1        107.5            84+    105.0    115.0    125.0       NA
## 2:      Open         56.3             57    117.5    125.0   -130.0       NA
## 3:      Open         67.5             69    120.0    130.0    137.5       NA
## 4:      Open         74.8             76    120.0    130.0    135.0       NA
## 5:      Open         82.3             84    137.5    147.5    155.0       NA
## 6:      Open         80.7             84     85.0     95.0    110.0       NA
##    Best3SquatKg Bench1Kg Bench2Kg Bench3Kg Bench4Kg Best3BenchKg Deadlift1Kg
##           <num>    <num>    <num>    <num>    <num>        <num>       <num>
## 1:        125.0     65.0    -67.5     67.5       NA         67.5       145.0
## 2:        125.0     67.5     70.0     72.5       NA         72.5       125.0
## 3:        137.5     90.0     95.0     97.5       NA         97.5       140.0
## 4:        135.0     65.0    -70.0     70.0       NA         70.0       117.5
## 5:        155.0     80.0     85.0     87.5       NA         87.5       170.0
## 6:        110.0     65.0     70.0    -72.5       NA         70.0       105.0
##    Deadlift2Kg Deadlift3Kg Deadlift4Kg Best3DeadliftKg TotalKg  Place   Dots
##          <num>       <num>       <num>           <num>   <num> <char>  <num>
## 1:       157.5      -162.5          NA           157.5   350.0      1 290.79
## 2:       135.0       142.5          NA           142.5   340.0      1 392.71
## 3:       152.5       162.5          NA           162.5   397.5      1 410.38
## 4:       127.5       132.5          NA           132.5   337.5      1 329.18
## 5:       185.0      -192.5          NA           185.0   427.5      1 397.24
## 6:       122.5       132.5          NA           132.5   312.5      2 293.18
##     Wilks Glossbrenner Goodlift Tested   Country  State Federation
##     <num>        <num>    <num> <char>    <char> <char>     <char>
## 1: 286.01       243.65    61.31    Yes Australia    QLD       APLA
## 2: 398.37       352.34    80.50    Yes Australia    NSW       APLA
## 3: 405.70       357.88    83.38    Yes Australia    NSW       APLA
## 4: 321.38       282.73    67.05    Yes Australia    QLD       APLA
## 5: 385.22       337.25    81.39    Yes Australia    NSW       APLA
## 6: 284.57       249.44    59.98    Yes Australia    QLD       APLA
##    ParentFederation       Date MeetCountry MeetState MeetTown        MeetName
##              <char>     <Date>      <char>    <char>   <char>          <char>
## 1:              IPF 2025-08-17   Australia       NSW          East Coast Open
## 2:              IPF 2025-08-17   Australia       NSW          East Coast Open
## 3:              IPF 2025-08-17   Australia       NSW          East Coast Open
## 4:              IPF 2025-08-17   Australia       NSW          East Coast Open
## 5:              IPF 2025-08-17   Australia       NSW          East Coast Open
## 6:              IPF 2025-08-17   Australia       NSW          East Coast Open
##    Sanctioned
##        <char>
## 1:        Yes
## 2:        Yes
## 3:        Yes
## 4:        Yes
## 5:        Yes
```

```
## 6:        Yes
```

N.B.: *All observations prior to 2021 have been removed from the dataset. It was deemed more interesting to specifically study the situation in recent years, avoiding the analysis being influenced by previous periods that might have different dynamics and athlete distributions.*

1.2 Extracting the Best Performance per Athlete

The dataset contains multiple records for the same athlete if they have participated in multiple competitions. To avoid analyzing redundant data and to focus on peak performance, we extract **a single observation for each athlete**: their best absolute performance, measured by the Goodlift score.

We also filter the competitions, excluding major international events (like the IPF Worlds or EPF Europeans) to concentrate on data from **national competitions**, which better represent the athlete base of a country.

```
Dati_best <- Dati %>%
  # Excluding international/continental federations to isolate national meets
  filter(!(Federation %in% c("IPF", "EPF", "NAPF", "APF", "NPF", "SAPF",
                             "FESUPO", "OPF", "APL", "CPF"))) %>%
  # Grouping by athlete, weight class, and age class
  group_by(Name, WeightClassKg, AgeClass) %>%
  # For each group, we keep only the row with the highest Goodlift score
  slice_max(Goodlift, n = 1, with_ties = FALSE) %>%
  ungroup()

head(Dati_best)
```

```
## # A tibble: 6 x 42
##   Name            Sex   Event Equipment    Age AgeClass BirthYearClass Division
##   <chr>           <chr> <chr> <chr>      <dbl> <chr>    <chr>          <chr>
## 1 A Cudd          M     B     Single-ply  46   45-49    40-49          M-O
## 2 A Dhanush       M     SBD   Single-ply  20   20-23    19-23          Juniors
## 3 A Hemanth Kumar M     SBD   Raw         22.5 20-23    19-23          Juniors
## 4 A Jyoshna       F     SBD   Single-ply  17   16-17    14-18          Open
## 5 A Madhuri       F     SBD   Raw         14.5 13-15    14-18          Sub-Juni~
## 6 A Madhuri       F     SBD   Raw         16.5 16-17    14-18          Sub-Juni~
## # i 34 more variables: BodyweightKg <dbl>, WeightClassKg <chr>, Squat1Kg <dbl>,
## #   Squat2Kg <dbl>, Squat3Kg <dbl>, Squat4Kg <dbl>, Best3SquatKg <dbl>,
## #   Bench1Kg <dbl>, Bench2Kg <dbl>, Bench3Kg <dbl>, Bench4Kg <dbl>,
## #   Best3BenchKg <dbl>, Deadlift1Kg <dbl>, Deadlift2Kg <dbl>,
## #   Deadlift3Kg <dbl>, Deadlift4Kg <dbl>, Best3DeadliftKg <dbl>, TotalKg <dbl>,
## #   Place <chr>, Dots <dbl>, Wilks <dbl>, Glossbrenner <dbl>, Goodlift <dbl>,
## #   Tested <chr>, Country <chr>, State <chr>, Federation <chr>, ...
```

1.3 Final Cleaning and Sample Selection

We proceed with further cleaning to ensure the quality and consistency of our analysis sample: - Remove disqualified athletes ('DQ'). - Select only the 'SBD' event (Squat, Bench, Deadlift). - Include only the 'Raw' category to ensure comparability. - Exclude, for interpretability of results, the few observations with Sex not valued within (M,F). - Remove rows with missing data for the main lifts or Goodlift score. - Finally, select only the columns relevant to our analysis.

```r
D <- Dati_best %>%
  filter(
    Place != 'DQ',
    Event == 'SBD',
    Equipment == 'Raw',
    Sex != 'Mx',
    !is.na(Goodlift),
    !is.na(Best3SquatKg),
    !is.na(Best3BenchKg),
    !is.na(Best3DeadliftKg)
  ) %>%
  select(Name, Sex, Age, WeightClassKg, BodyweightKg, MeetCountry,
         Best3SquatKg, Best3BenchKg, Best3DeadliftKg, TotalKg, Goodlift)

head(D)
```

```
## # A tibble: 6 x 11
##   Name           Sex     Age WeightClassKg BodyweightKg MeetCountry Best3SquatKg
##   <chr>          <chr> <dbl> <chr>                <dbl> <chr>              <dbl>
## 1 A Hemanth Kum~ M      22.5 66                    63.8 India                175
## 2 A Madhuri      F      14.5 76                    74.5 India                112.
## 3 A Madhuri      F      16.5 76                    76   India                115
## 4 A Rajasekhar   M      53   93                    88.5 India                165
## 5 A V Anjana     F      23   84+                  111.  UAE                  160
## 6 A V V Satyana~ M      16.5 105                   96.3 India                150
## # i 4 more variables: Best3BenchKg <dbl>, Best3DeadliftKg <dbl>, TotalKg <dbl>,
## #   Goodlift <dbl>
```

1.4 Creating Percentage Variables

The goal of our analysis is to identify profiles based on an athlete's "structure," not their absolute strength. For this, we create three new variables that express each lift as a **percentage of the total lifted**. This allows us to fairly compare the specialization of a lighter athlete with that of a heavier one.

```r
DC <- D %>%
  mutate(
    PSquat = (Best3SquatKg / TotalKg) * 100,
    PBench = (Best3BenchKg / TotalKg) * 100,
    PDeads = (Best3DeadliftKg / TotalKg) * 100
  )
# Saving the clean and ready-to-use dataset, so we can load it directly
# in the future without repeating all the steps.
save(DC, file = 'DC.RData')

# Display the first few rows of the final dataset
head(DC)
```

```
## # A tibble: 6 x 14
##   Name           Sex     Age WeightClassKg BodyweightKg MeetCountry Best3SquatKg
##   <chr>          <chr> <dbl> <chr>                <dbl> <chr>              <dbl>
## 1 A Hemanth Kum~ M      22.5 66                    63.8 India                175
## 2 A Madhuri      F      14.5 76                    74.5 India                112.
```

```
## 3 A Madhuri      F      16.5 76                   76   India              115
## 4 A Rajasekhar   M      53   93                   88.5 India              165
## 5 A V Anjana     F      23   84+                 111.  UAE                160
## 6 A V V Satyana~ M      16.5 105                  96.3 India              150
## # i 7 more variables: Best3BenchKg <dbl>, Best3DeadliftKg <dbl>, TotalKg <dbl>,
## #   Goodlift <dbl>, PSquat <dbl>, PBench <dbl>, PDeads <dbl>
```

The DC dataset is now ready. It contains essential demographic and competition information and, most importantly, the three percentage variables (PSquat, PBench, PDeads) that will be the core of our Cluster Analysis.

---

2. Descriptive Analysis: The Composition of the Total

To compare athletes fairly, we use the percentage of each lift relative to the total, as calculated in the previous section. We analyze the distribution of these three variables to get an initial idea of the average powerlifter's structure.

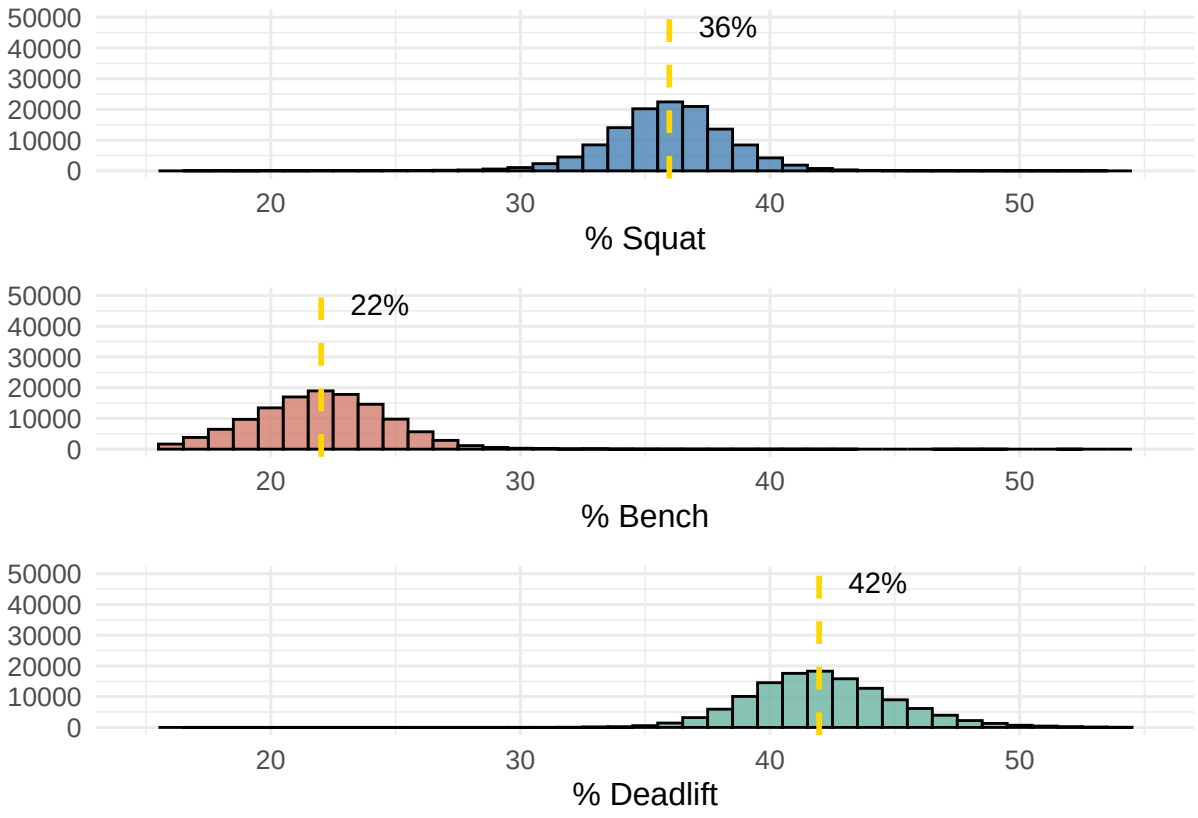
```
med_squat <- median(DC$PSquat, na.rm = TRUE)
med_bench <- median(DC$PBench, na.rm = TRUE)
med_deads <- median(DC$PDeads, na.rm = TRUE)

p1 <- ggplot(DC, aes(x = PSquat)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black", alpha = 0.8) +
  geom_vline(xintercept = med_squat, linetype = "dashed", color = "gold", linewidth = 1) +
  annotate("text", x = med_squat, y = Inf, label = paste0(round(med_squat, 1), '%'),
           vjust = 1.3, hjust = -0.5, color = "black", size = 4) +
  labs(x = "% Squat", y = NULL) +
  xlim(15, 55) + ylim(0, 50000) + theme_minimal()

p2 <- ggplot(DC, aes(x = PBench)) +
  geom_histogram(binwidth = 1, fill = "#d17d6d", color = "black", alpha = 0.8) +
  geom_vline(xintercept = med_bench, linetype = "dashed", color = "gold", linewidth = 1) +
  annotate("text", x = med_bench, y = Inf, label = paste0(round(med_bench, 1), '%'),
           vjust = 1.3, hjust = -0.5, color = "black", size = 4) +
  labs(x = "% Bench", y = NULL) +
  xlim(15, 55) + ylim(0, 50000) + theme_minimal()

p3 <- ggplot(DC, aes(x = PDeads)) +
  geom_histogram(binwidth = 1, fill = "#69b3a2", color = "black", alpha = 0.8) +
  geom_vline(xintercept = med_deads, linetype = "dashed", color = "gold", linewidth = 1) +
  annotate("text", x = med_deads, y = Inf, label = paste0(round(med_deads, 1), '%'),
           vjust = 1.3, hjust = -0.5, color = "black", size = 4) +
  labs(x = "% Deadlift", y = NULL) +
  xlim(15, 55) + ylim(0, 50000) + theme_minimal()

p1 / p2 / p3
```



This plot not suprisingly shows that, on average, the deadlift constitutes the largest portion of the total (median ~42%), followed by the squat (~35.6%) and the bench press (~22.3%).

---

3. Cluster Analysis: Identifying Athlete Profiles

3.1. Choosing the Number of Clusters (Combined Approach)

To determine the optimal number of clusters (k), the first step is to use a quantitative method like the "Elbow Method." This method calculates the within-cluster sum of squares (SSE) for different values of K. The "elbow" in the plot, where the SSE stops decreasing rapidly, indicates a good trade-off between the number of clusters and internal compactness.

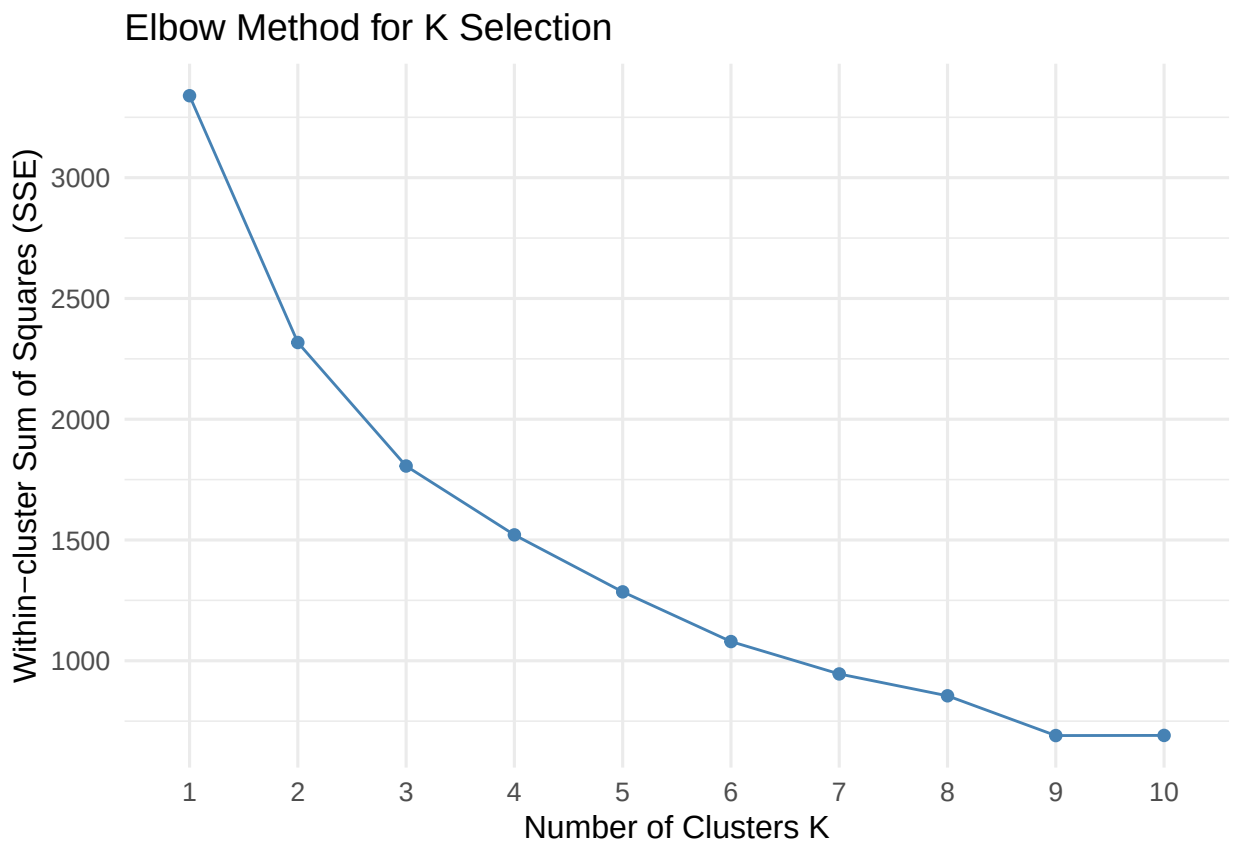
To make the calculation efficient on such a large dataset, we apply the method to a random sample of 1000 athletes.

```
DC_pct <- DC %>% select(PSquat, PBench, PDeads)
DC_scaled <- scale(DC_pct)

set.seed(123)
DC_sample_scaled <- DC_scaled %>% as.data.frame() %>% sample_n(1000)

fviz_nbclust(DC_sample_scaled, kmeans, method = "wss", nstart = 10) +
```

```
  labs(title = "Elbow Method for K Selection", x = 'Number of Clusters K', y = 'Within-cluster Sum of Sq
  theme_minimal()
```



As can be seen from the graph, there is no sharp, unambiguous "elbow." The most drastic decrease in SSE occurs when moving from K=1 to K=2. The transition from K=2 to K=3 also shows a considerable reduction, after which the curve's slope flattens. This suggests that both K=2 and K=3 could be reasonable choices from a purely statistical standpoint. In situations like this, **domain knowledge** is crucial to resolve ambiguity. Since powerlifting is defined by **three distinct disciplines (Squat, Bench Press, Deadlift)**, it makes a great deal of interpretive sense to try to identify three athlete profiles, each potentially specializing in one of these lifts. Therefore, backed by a quantitative indication that does not rule out K=3 and a solid theoretical justification, we choose to proceed with K=3 clusters.

3.2. Applying K-Means and Renaming Clusters

We apply the K-Means algorithm with k=3 to the entire dataset and then analyze the cluster means to give them intuitive names.

```
set.seed(123)
km_full <- kmeans(DC_scaled, centers = 3, nstart = 10)
DC$Cluster <- factor(km_full$cluster)

cluster_summary <- DC %>%
  group_by(Cluster) %>%
  summarise(
    Squat_Mean = mean(PSquat),
```

```
    Bench_Mean = mean(PBench),
    Deadlift_Mean = mean(PDeads),
    Num_Athletes = n()
  )
print(cluster_summary)
```

```
## # A tibble: 3 x 5
##   Cluster Squat_Mean Bench_Mean Deadlift_Mean Num_Athletes
##   <fct>        <dbl>      <dbl>         <dbl>        <int>
## 1 1             35.6       24.5          39.9        49030
## 2 2             38.0       20.1          41.9        41972
## 3 3             33.7       20.8          45.5        34311
```

```
# We rename the clusters based on the dominant lift to make them more readable
# Note: Cluster numbers might change on each run, so we adapt the names.
# Check the table above to confirm the correct mapping!
DC <- DC %>%
  mutate(Cluster = recode_factor(Cluster,
                                 `1` = "Bp",  # Bench Press dominant
                                 `2` = "Dl",  # Deadlift dominant
                                 `3` = "Sq")) # Squat dominant

# We reorder the levels for consistency in plots
DC$Cluster <- factor(DC$Cluster, levels = c('Sq', 'Bp', 'Dl'))
```

The table of means confirms the existence of three distinct profiles: a group with a higher-than-average Squat percentage (which we'll call 'Sq'), one with a dominant Bench Press percentage ('Bp'), and one specialized in the Deadlift ('Dl').

3.3. Cluster Visualization

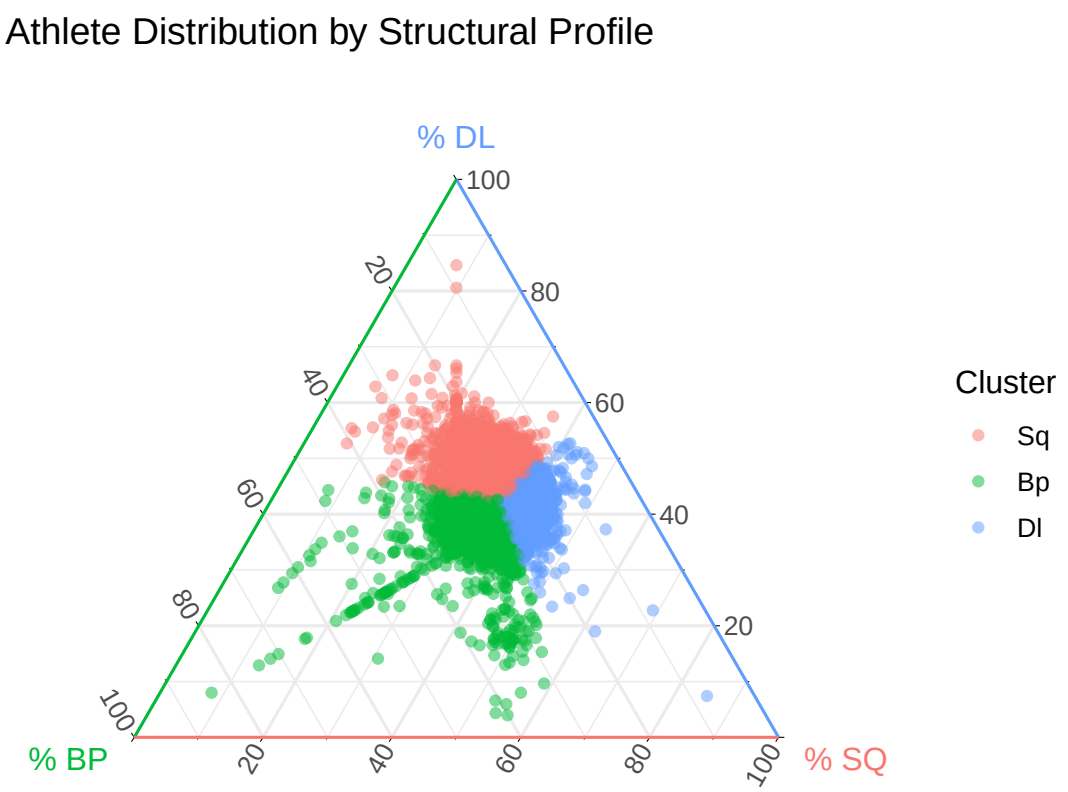
A ternary plot is perfect for visualizing compositional data like our three percentages. Each point represents an athlete, and its position depends on their lifting structure.

```
cluster_levels <- levels(as.factor(DC$Cluster))
auto_colors <- scales::hue_pal()(length(cluster_levels))
names(auto_colors) <- cluster_levels

ggtern(data = DC, aes(x = PBench, y = PDeads, z = PSquat, color = Cluster)) +
  geom_point(alpha = 0.5, size = 1.5) +
  labs(
    title = "Athlete Distribution by Structural Profile",
    x = "% BP",
    y = "% DL",
    z = "% SQ",
    color = "Cluster"
  ) +
  scale_color_manual(values = auto_colors) +
  theme_minimal() +
  theme(tern.axis.line.T = element_line(color = auto_colors["Dl"]),
    tern.axis.title.T = element_text(color = auto_colors["Dl"]),
```

```
  tern.axis.line.L = element_line(color = auto_colors["Bp"]),
  tern.axis.title.L = element_text(color = auto_colors["Bp"]),

  tern.axis.line.R = element_line(color = auto_colors["Sq"]),
  tern.axis.title.R = element_text(color = auto_colors["Sq"])
)
```



This plot clearly shows the three "islands" of density corresponding to the identified clusters, visually confirming the validity of our segmentation. Each corner of the triangle represents an athlete who totals 100% in that specific lift.

---

4. Association between Clusters and Nation

Is there a national "specialization"? Do some countries tend to produce more athletes of a certain profile? We use **Correspondence Analysis** to investigate the association between cluster membership and the country of competition.

4.1. Country Selection and Association Test

To avoid an overly crowded plot, we filter the dataset to include only the countries with the most athletes, covering about 90% of the observations (37 countries). We then perform a Chi-squared test to check if the association is statistically significant.

```r
# Select countries covering 90% of observations
country_counts <- as.data.frame(table(DC$MeetCountry)) %>%
  arrange(desc(Freq)) %>%
  mutate(Cumulative_Perc = cumsum(Freq) / sum(Freq))

selected_countries <- country_counts %>%
  filter(Cumulative_Perc <= 0.90) %>%
  pull(Var1)

# Filter the dataset
DC_filtered_countries <- DC %>% filter(MeetCountry %in% selected_countries)

# Create contingency table and test the association
contingency_table <- table(DC_filtered_countries$Cluster, DC_filtered_countries$MeetCountry)
chisq_result <- chisq.test(contingency_table)

print(chisq_result)
```

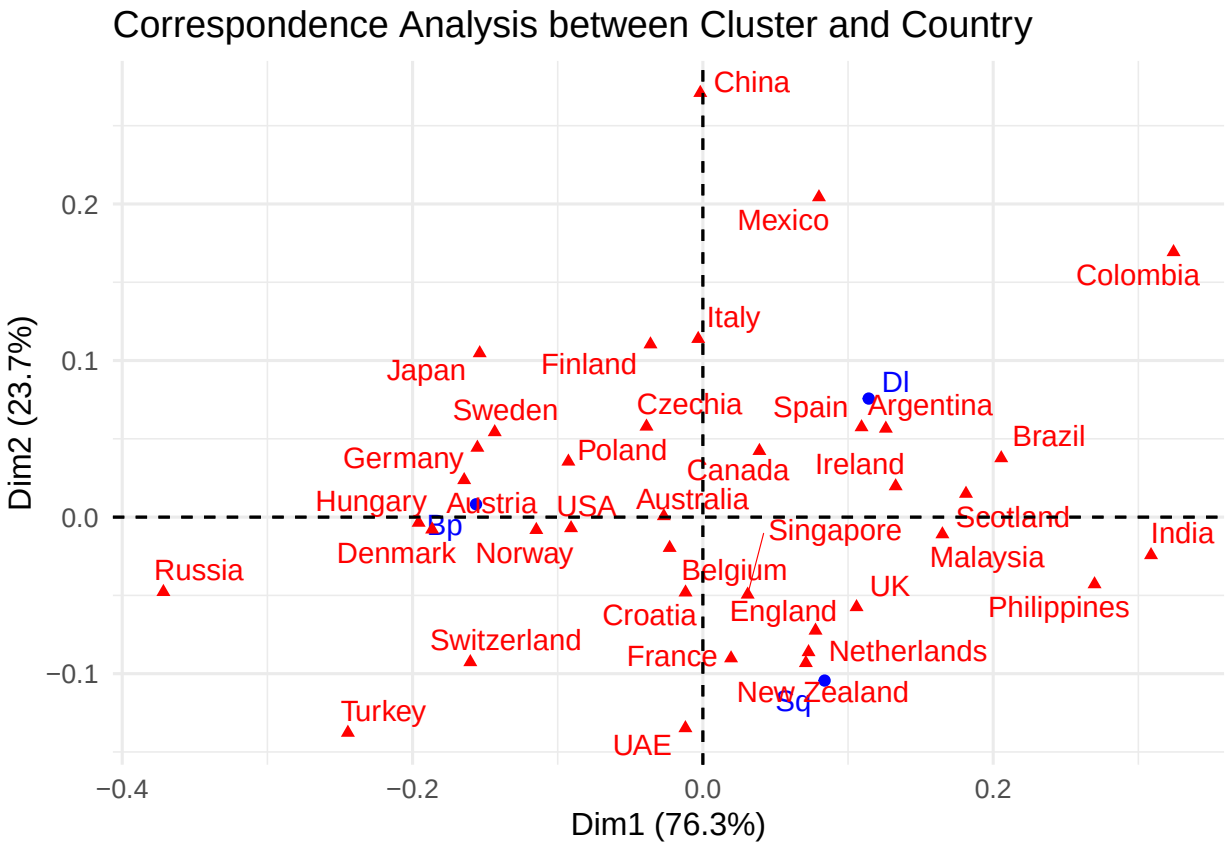
```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 2342.6, df = 72, p-value < 2.2e-16
```

The p-value of the Chi-squared test is extremely low (< 2.2e-16), which allows us to reject the hypothesis of independence. There is a statistically significant association between an athlete's profile and the country in which they compete.

4.2. Correspondence Analysis (CA Biplot)

The Correspondence Analysis biplot graphically displays this relationship. Proximity between a cluster point (blue) and a country point (red) indicates a stronger-than-average association. It's important to note that the two dimensions of the plot explain 100% of the total variance (Dim1: 76.3%, Dim2: 23.7%), so this biplot is a complete and faithful representation.

```r
res.ca <- CA(contingency_table, graph = FALSE)
fviz_ca_biplot(res.ca, repel = TRUE, title = "Correspondence Analysis between Cluster and Country") +
  theme_minimal()
```



*The biplot reveals very clear geographical patterns. The horizontal axis (Dim1) starkly contrasts the **Deadlifters (Dl)** on the left with the **Benchers (Bp)** on the right. The vertical axis (Dim2) isolates the **Squatters (Sq)** in the upper portion. Key associations emerge: * **'Sq' Cluster (Squatters):** Associated with nations like **New Zealand, Netherlands, UK, and Philippines.** * **'Bp' Cluster (Benchers):** Shows a strong association with a group of Central and Northern European countries, particularly **Germany, Austria, Poland, and Sweden**. **Russia** appears as an extreme case, heavily skewed towards this profile. * **'Dl' Cluster (Deadlifters):** Clearly associated with Latin American countries like **Colombia, Mexico, Brazil, and Argentina.** * **Average Profile:** Nations like the **USA, Australia, Canada, and Belgium** are very close to the center, indicating their athlete distribution across the three clusters is very similar to the global average.*

---

5. Comparing Performance Across Clusters

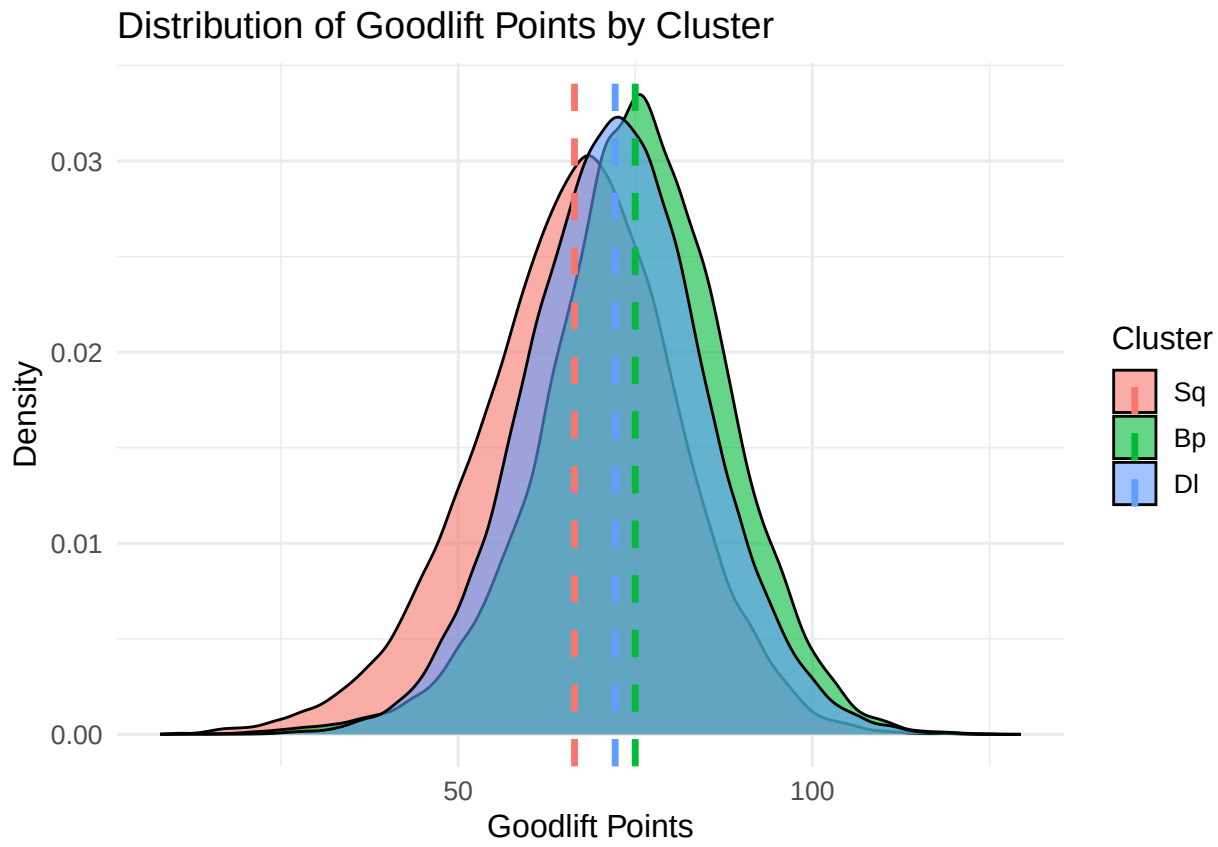
Which athlete profile is, on average, more successful? We compare **Goodlift Points** (a standardized score accounting for body weight and sex) across the three clusters.

5.1. Distribution of GL Points and ANOVA Test

We visualize the distributions of GL Points for each cluster and use an ANOVA test to verify if the differences in means are significant.

```r
# Calculate means for the vertical lines in the plot
cluster_means <- DC %>%
  group_by(Cluster) %>%
  summarise(mean_goodlift = mean(Goodlift, na.rm = TRUE))

# Density plot
ggplot(DC, aes(x = Goodlift, fill = Cluster)) +
  geom_density(alpha = 0.6) +
  geom_vline(data = cluster_means, aes(xintercept = mean_goodlift, color = Cluster),
             linetype = "dashed", size = 1.2) +
  labs(
    title = "Distribution of Goodlift Points by Cluster",
    x = "Goodlift Points",
    y = "Density"
  ) +
  theme_minimal()
```




```r
# ANOVA Test
summary(aov(Goodlift ~ Cluster, data = DC))
```

```
##                 Df   Sum Sq Mean Sq F value Pr(>F)
## Cluster          2  1496492  748246    4276 <2e-16 ***
## Residuals   125310 21925849     175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*Visually, the **Benchers (Bp)** cluster appears to have the highest average GL Points. The ANOVA test confirms that the differences between the groups are statistically highly significant (p-value < 2e-16). This result is counterintuitive, given that the bench press contributes the least to the total.*

5.2. Gender Analysis: A Hidden Factor?

The high performance of "Benchers" could be explained by a confounding variable. We investigate the gender composition of the clusters.

```
# Contingency table for gender and cluster
table(DC$Cluster, DC$Sex)
```

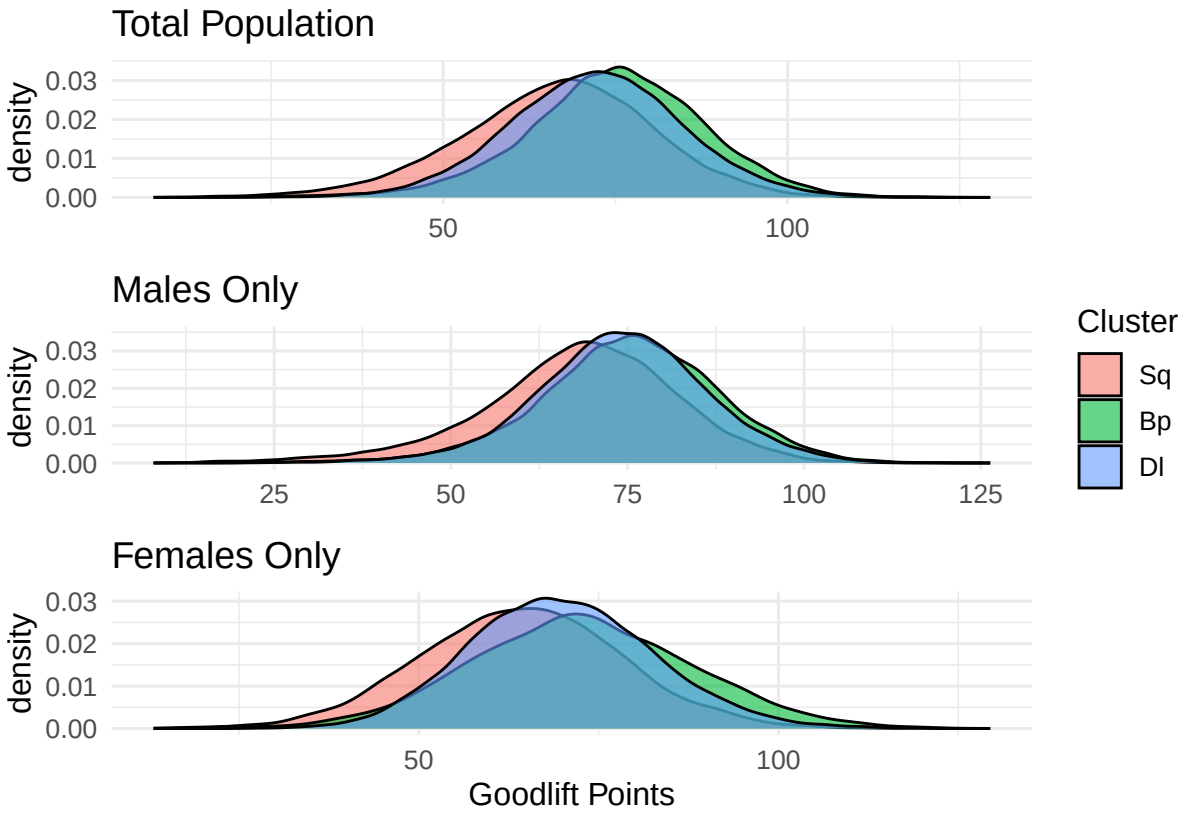
```
##
##          F     M
##   Sq 15142 19169
##   Bp  4303 44727
##   Dl 20253 21719
```

```
# Separate density plots for Males and Females
p_all <- ggplot(DC, aes(x = Goodlift, fill = Cluster)) +
  geom_density(alpha = 0.6) + labs(title = "Total Population", x = NULL) + theme_minimal()

p_male <- DC %>% filter(Sex == 'M') %>% ggplot(aes(x = Goodlift, fill = Cluster)) +
  geom_density(alpha = 0.6) + labs(title = "Males Only", x = NULL) + theme_minimal()

p_female <- DC %>% filter(Sex == 'F') %>% ggplot(aes(x = Goodlift, fill = Cluster)) +
  geom_density(alpha = 0.6) + labs(title = "Females Only", x = "Goodlift Points") + theme_minimal()

p_all / p_male / p_female + plot_layout(guides = "collect")
```



The cause is revealed. The table shows that the ***Benchers (Bp) cluster is over 90% male****, and males tend to have higher average GL Points. When we analyze the sexes separately, we see that* ***controlling for gender, the performance differences between clusters drastically decrease or even reverse****. The Deadlifters (Dl) cluster, for instance, has the lowest GL Points in both sexes.*

---

6. Conclusions

This analysis has allowed us to draw several key conclusions:

1. **There are three distinct athlete profiles** in powerlifting, characterized by the percentage dominance of one of the three lifts: Squat, Bench Press, or Deadlift.
2. **Cluster membership is significantly associated with the nation** where the athlete competes, suggesting possible "schools" or specializations at a geographical level.
3. The **"Benchers" cluster shows the highest GL Points scores**, but this effect is almost entirely explained by a **strong male prevalence** within it (gender gap).
4. **Controlling for gender**, the **"Deadlifter"** profile proves to be the one with the lowest average performance, while "Squatters" and "Benchers" are equivalent.

This notebook demonstrates how, starting from raw data, it is possible to use clustering and statistical analysis techniques to uncover hidden patterns and tell a complex and multi-faceted story.