SoccerWorksheet

# Load packages  
library(readr)

Warning: package 'readr' was built under R version 4.4.3

library(dplyr)

Warning: package 'dplyr' was built under R version 4.4.3

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':  
  
 filter, lag

The following objects are masked from 'package:base':  
  
 intersect, setdiff, setequal, union

library(ggplot2)

Warning: package 'ggplot2' was built under R version 4.4.3

library(GGally)

Warning: package 'GGally' was built under R version 4.4.3

Registered S3 method overwritten by 'GGally':  
 method from   
 +.gg ggplot2

library(tidyr)

Warning: package 'tidyr' was built under R version 4.4.3

library(cluster)  
library(factoextra)

Warning: package 'factoextra' was built under R version 4.4.3

Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(fmsb)

Warning: package 'fmsb' was built under R version 4.4.3

# Load data  
fifa <- read\_csv("../Data/players\_22.csv", show\_col\_types = FALSE)  
  
# Clean and select relevant columns  
fifa\_clean <- fifa |>  
 select(  
 short\_name, age, nationality\_name, club\_name, overall, potential,  
 pace, shooting, passing, dribbling, defending, physic  
 ) |>  
 filter(if\_all(c(pace, shooting, passing, dribbling, defending, physic), ~ !is.na(.))) |>  
 distinct()  
  
# Save for reuse  
write\_csv(fifa\_clean, "fifa\_clean.csv")

# —- PART 1: EXPLORATORY ANALYSIS —-

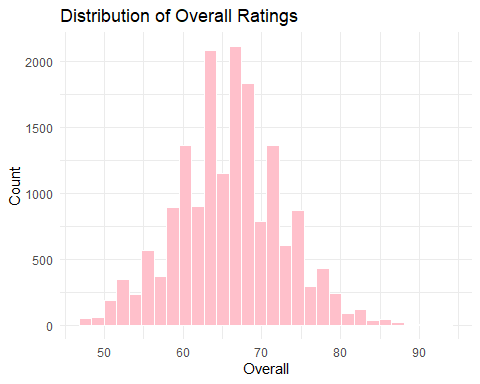
# Q1: What are the average values for age, overall rating, and key skills?

summary\_stats <- fifa\_clean |>  
 summarise(  
 avg\_age = mean(age),  
 avg\_overall = mean(overall),  
 avg\_pace = mean(pace),  
 avg\_shooting = mean(shooting),  
 avg\_passing = mean(passing),  
 avg\_dribbling = mean(dribbling),  
 avg\_defending = mean(defending),  
 avg\_physic = mean(physic)  
 )  
print(summary\_stats)

# A tibble: 1 × 8  
 avg\_age avg\_overall avg\_pace avg\_shooting avg\_passing avg\_dribbling  
 <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 25.1 65.9 68.2 52.3 57.3 62.6  
# ℹ 2 more variables: avg\_defending <dbl>, avg\_physic <dbl>

# Q2: What does the distribution of overall player ratings look like?

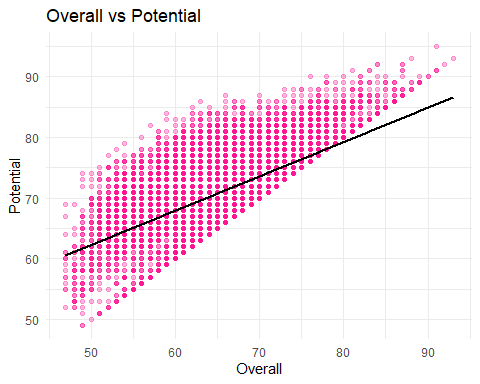
ggplot(fifa\_clean, aes(x = overall)) +  
 geom\_histogram(bins = 30, fill = "pink", color = "white") +  
 labs(title = "Distribution of Overall Ratings", x = "Overall", y = "Count") +  
 theme\_minimal()



# Q3: How does overall rating relate to potential?

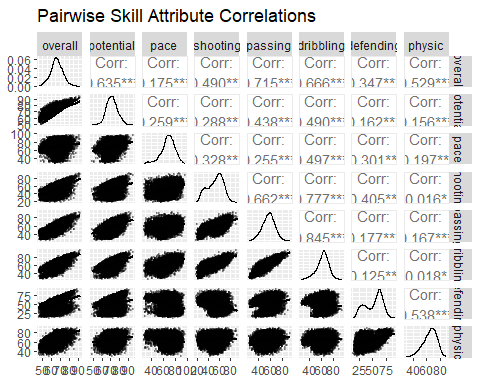
ggplot(fifa\_clean, aes(x = overall, y = potential)) +  
 geom\_point(alpha = 0.3, color = "deeppink") +  
 geom\_smooth(method = "lm", se = FALSE, color = "black") +  
 labs(title = "Overall vs Potential", x = "Overall", y = "Potential") +  
 theme\_minimal()

`geom\_smooth()` using formula = 'y ~ x'



# Q4: Create a pairwise correlation plot of all skill attributes.

skill\_subset <- fifa\_clean |>  
 select(overall, potential, pace, shooting, passing, dribbling, defending, physic)  
  
ggpairs(skill\_subset, lower = list(continuous = wrap("smooth", alpha = 0.3, size = 0.1)),  
 title = "Pairwise Skill Attribute Correlations")



# —- PART 2: K-MEANS CLUSTERING —-

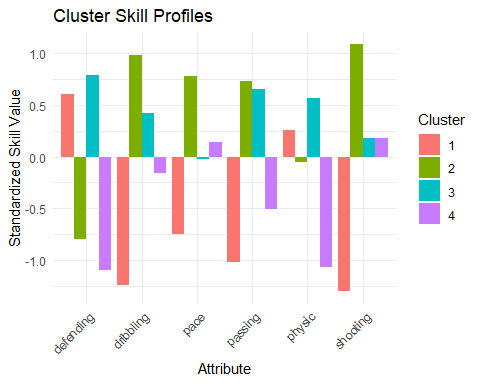
# Prepare numeric skills and scale  
fifa\_cluster <- fifa\_clean |>  
 select(pace, shooting, passing, dribbling, defending, physic)  
fifa\_scaled <- scale(fifa\_cluster)  
  
# Q5: Use K-means to group players into 4 clusters  
set.seed(123)  
k4 <- kmeans(fifa\_scaled, centers = 4, nstart = 20)

Warning: Quick-TRANSfer stage steps exceeded maximum (= 855350)

fifa\_clean <- fifa\_clean |>  
 mutate(Cluster = as.factor(k4$cluster))

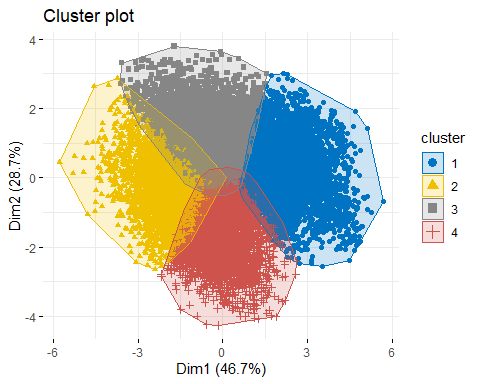
# Q6: Plot the standardized skill profiles for each cluster

centers <- as\_tibble(k4$centers) |> mutate(Cluster = as.factor(1:4)) |>   
 pivot\_longer(-Cluster, names\_to = "Attribute", values\_to = "Value")  
  
ggplot(centers, aes(x = Attribute, y = Value, fill = Cluster)) +  
 geom\_col(position = "dodge") +  
 labs(title = "Cluster Skill Profiles", y = "Standardized Skill Value") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



# Q7: Visualize clusters using PCA

fviz\_cluster(k4, data = fifa\_scaled, geom = "point", palette = "jco", ggtheme = theme\_minimal())



# —- PART 3: CLUSTER INTERPRETATION —-

# Q8: How do average raw skill ratings vary across clusters?

cluster\_summary <- fifa\_clean |>  
 group\_by(Cluster)|>  
 summarise(across(c(pace, shooting, passing, dribbling, defending, physic), mean))  
  
print(cluster\_summary)

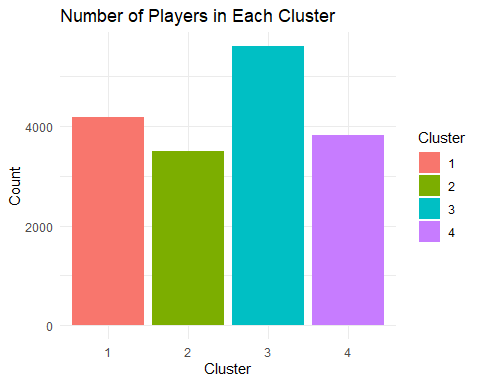
# A tibble: 4 × 7  
 Cluster pace shooting passing dribbling defending physic  
 <fct> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
1 1 60.1 34.1 47.1 50.6 61.4 67.4  
2 2 76.7 67.5 64.7 72.0 38.8 64.3  
3 3 67.9 54.8 63.8 66.6 64.5 70.3  
4 4 69.8 54.8 52.2 61.0 34.1 54.4

# Q9: Which cluster has the most players?

cluster\_sizes <- fifa\_clean |> count(Cluster, sort = TRUE)  
print(cluster\_sizes)

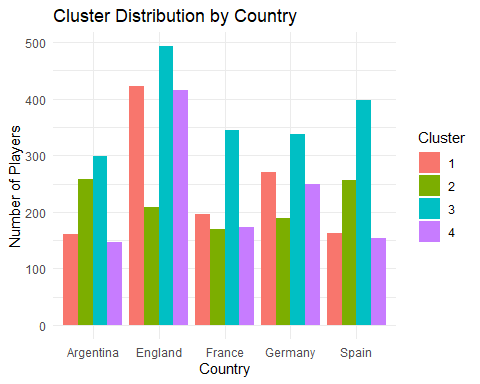
# A tibble: 4 × 2  
 Cluster n  
 <fct> <int>  
1 3 5611  
2 1 4183  
3 4 3818  
4 2 3495

ggplot(cluster\_sizes, aes(x = Cluster, y = n, fill = Cluster)) +  
 geom\_col() +  
 labs(title = "Number of Players in Each Cluster", x = "Cluster", y = "Count") +  
 theme\_minimal()



# Q10: How are clusters distributed among top 5 nationalities?

top\_countries <- fifa\_clean |>  
 count(nationality\_name, sort = TRUE) |>  
 slice\_max(n, n = 5) |>  
 pull(nationality\_name)  
  
fifa\_top <- fifa\_clean |>  
 filter(nationality\_name %in% top\_countries)  
  
ggplot(fifa\_top, aes(x = nationality\_name, fill = Cluster)) +  
 geom\_bar(position = "dodge") +  
 labs(title = "Cluster Distribution by Country", x = "Country", y = "Number of Players") +  
 theme\_minimal()



# —- PART 4: PLAYER ARCHETYPES —-

# Q11: Create a radar chart of skill profiles by cluster

# Create radar chart of player archetypes by cluster  
  
# Step 1: Compute means  
radar\_data <- fifa\_clean |>  
 group\_by(Cluster) |>  
 summarise(  
 Overall = mean(overall),  
 Pace = mean(pace),  
 Shooting = mean(shooting),  
 Passing = mean(passing),  
 Dribbling = mean(dribbling),  
 Defending = mean(defending),  
 Physic = mean(physic)  
 ) |>   
 as.data.frame()  
  
# Step 2: Convert to proper format for radar chart  
rownames(radar\_data) <- radar\_data$Cluster  
radar\_data <- radar\_data[, -1]  
  
# Step 3: Add max and min rows  
radar\_plot\_data <- rbind(  
 rep(100, 7), # max  
 rep(0, 7), # min  
 radar\_data # actual cluster data  
)  
  
# Step 4: Draw radar chart  
radarchart(radar\_plot\_data, axistype = 1,  
 pcol = c("pink", "hotpink", "blanchedalmond", "orchid"),  
 pfcol = c("#FF69B480", "#FF149380", "#FFDAB980", "#DA70D680"),  
 plwd = 2, plty = 1,  
 cglcol = "grey", cglty = 1,  
 axislabcol = "grey", vlcex = 0.8,  
 title = "Radar Profiles of Player Clusters")  
  
legend("topright", legend = paste("Cluster", 1:4),  
 col = c("pink", "hotpink", "blanchedalmond", "orchid"),  
 lty = 1, lwd = 2, bty = "n")

