pca_original_scale.R

```r
source("scripts/load_machines_subset.R")

compute_pca <- function(X = load_machines_subset(),
                                out_dir = "plots/plots_pca_original", std = FALSE) {

  pca <- prcomp(X, center = TRUE, scale. = std)
  pve  <- (pca$sdev^2) / sum(pca$sdev^2)
  cpve <- cumsum(pve)
  k <- which(cpve >= 0.95)[1]

  cat("\n==================== RESULTS ====================\n")
  cat("Number of variables (p):", ncol(X), "\n")
  cat("Number of observations (n):", nrow(X), "\n\n")

  cat("PVE (proportion variance explained) per PC:\n")
  print(pve)

  cat("\nCPVE (cumulative PVE):\n")
  print(cpve)

  cat("\nMinimum k for CPVE >= 0.95:\n")
  cat("k =", k, "\n")
  cat("CPVE[k] =", cpve[k], "\n")


  cat("Loadings (rotation) for retained PCs (1..k):\n")
  print(pca$rotation[, 1:k, drop = FALSE])

  abs_load <- abs(pca$rotation)
  top_pc1 <- sort(abs_load[, 1], decreasing = TRUE)
  cat("\nTop contributors to PC1 (absolute loading):\n")
  print(top_pc1)

  if (ncol(X) >= 2) {
    top_pc2 <- sort(abs_load[, 2], decreasing = TRUE)
    cat("\nTop contributors to PC2 (absolute loading):\n")
    print(top_pc2)
  }


  if (!dir.exists(out_dir))
    dir.create(out_dir)

  # -------------------------------
  # PLOT 1: Scree plot (PVE)
  # -------------------------------
  png(
    filename = file.path(out_dir, "01_scree_pve.png"),
    width = 1200,
    height = 800,
    res = 150
  )
  barplot(
    pve,
    names.arg = paste0("PC", seq_along(pve)),
    las = 2,
```

pca_original_scale.R

```r
  xlab = "Principal components",
  ylab = "Proportion of variance explained"
)
dev.off()

# ------------------------------
# PLOT 2: Cumulative variance + 95%
# ------------------------------
png(
  filename = file.path(out_dir, "02_cumulative_cpve.png"),
  width = 1200,
  height = 800,
  res = 150
)
plot(
  cpve,
  type = "b",
  pch = 19,
  ylim = c(0, 1),
  xlab = "Number of PCs",
  ylab = "Cumulative proportion of variance explained"
)
abline(h = 0.95, lty = 2)
abline(v = k, lty = 3)
text(
  k,
  cpve[k],
  labels = paste0("k = ", k, " (", round(100 * cpve[k], 2), "%)"),
  pos = 4,
  cex = 0.9
)
dev.off()

# ------------------------------
# PLOT 3: Scores plot (PC1 vs PC2)
# ------------------------------

library(ggplot2)
library(ggrepel)

scores <- pca$x
rownames(scores) <- rownames(X)

scores_df <- data.frame(
  PC1  = scores[, 1],
  PC2  = scores[, 2],
  name = rownames(scores)
)

med1 <- median(scores_df$PC1)
med2 <- median(scores_df$PC2)

scores_df$dist <- sqrt((scores_df$PC1 - med1)^2 + (scores_df$PC2 - med2)^2)

# top 25% have label
thr <- quantile(scores_df$dist, 0.75)
```

pca_original_scale.R

```
  scores_df$label <- ifelse(scores_df$dist > thr, scores_df$name, "")

  # jitter to split coincident points
  scores_df$PC1_j <- jitter(scores_df$PC1, amount = diff(range(scores_df$PC1)) * 0.01
)
  scores_df$PC2_j <- jitter(scores_df$PC2, amount = diff(range(scores_df$PC2)) * 0.01
)

  png(
    filename = file.path(out_dir, "03_scores_pc1_pc2.png"),
    width  = 800,
    height = 800,
    res    = 150
  )

  p <- ggplot(scores_df, aes(PC1_j, PC2_j)) +
    geom_point() +
    geom_vline(xintercept = 0, linetype = 2) +
    geom_hline(yintercept = 0, linetype = 2) +
    geom_text_repel(
      aes(label = label),
      size         = 3.5,
      max.overlaps = 50
    ) +
    labs(
      title = "Scores plot (PC1 vs PC2)",
      x     = "PC1 score",
      y     = "PC2 score"
    ) +
    coord_fixed() +
    theme_classic(base_size = 14) +
    theme(aspect.ratio = 1)

  print(p)
  dev.off()

  # -------------------------------
  # PLOT 4: Loadings barplots (PC1 and PC2)
  # -------------------------------
  png(
    filename = file.path(out_dir, "04_loadings_pc1_pc2.png"),
    width = 1400,
    height = 800,
    res = 150
  )
  par(mfrow = c(1, 2), mar = c(7, 4, 4, 1))

  barplot(pca$rotation[, 1],
          las = 2,
          main = "Loadings - PC1",
          ylab = "Loading")
  abline(h = 0, lty = 2)

  barplot(pca$rotation[, 2],
          las = 2,
          main = "Loadings - PC2",
          ylab = "Loading")
  abline(h = 0, lty = 2)

  par(mfrow = c(1, 1))
  dev.off()

  # -------------------------------
  # PLOT 5: Biplot (PC1 vs PC2)
  # -------------------------------
  png(
    filename = file.path(out_dir, "05_biplot_pc1_pc2.png"),
    width = 1200,
    height = 900,
    res = 150
  )
  biplot(pca,
        choices = c(1, 2),
        cex = 0.8,
        main = "")
  title("Biplot (PC1 vs PC2) - scores + loadings", line = 2)
  dev.off()

  # ======================================================================
  # Extreme analysis based on loadings + var values
  # ======================================================================

  extremes <- scores_df[order(-scores_df$dist), ]
  ext_names <- extremes$name[1:6]
  load <- pca$rotation

  # most contribution for pc1 and pc2
  vars_pc1 <- names(sort(abs(load[, 1]), decreasing = TRUE))[1:5]
  vars_pc2 <- names(sort(abs(load[, 2]), decreasing = TRUE))[1:5]
  vars_key <- unique(c(vars_pc1, vars_pc2))
  X_ext <- X[ext_names, vars_key, drop = FALSE]

  cat("\nScores of the machines farthest in the PC1–PC2 space:\n")
  print(head(extremos, 6))

  cat("\nMachines farthest in the PC1–PC2 space:\n")
  print(ext_names)

  cat("\nVariables with the highest loadings on PC1:\n")
  print(vars_pc1)

  cat("\nVariables with the highest loadings on PC2:\n")
  print(vars_pc2)

  cat("\nOriginal values of these variables for each extreme machine:\n")
  print(round(X_ext, 3))

  # ======================================================================


  cat("Saved plots in folder:", out_dir, "\n")
  cat("Files:\n")
```

pca_original_scale.R

```r
  cat("  01_scree_pve.png\n")
  cat("  02_cumulative_cpve.png\n")
  cat("  03_scores_pc1_pc2.png\n")
  cat("  04_loadings_pc1_pc2.png\n")
  cat("  05_biplot_pc1_pc2.png\n")
}

compute_pca()
```