Traffic FInal Code

Erin Xu

2025-11-14

Setup

```
library(openxlsx)
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.5.2
```

```
## Warning: package 'readr' was built under R version 4.5.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.6
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.1     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
```

```
library(fastICA)
library(ggplot2)
library(ggvenn)
library(tidyr)
library(cluster)
file <- "traffic.xlsx"
sheet_names <- getSheetNames(file)
num_sheets <- length(sheet_names)
print(sheet_names)
```

```
##  [1] "Loc1"  "Loc2"  "Loc3"  "Loc4"  "Loc5"  "Loc6"  "Loc7"  "Loc8"  "Loc9"
## [10] "Loc10" "Loc11" "Loc12" "Loc13" "Loc14" "Loc15" "Loc16" "Loc17" "Loc18"
## [19] "Loc19" "Loc20" "Loc21" "Loc22" "Loc23" "Loc24" "Loc25" "Loc26"
```

```
print(num_sheets)
```

```
## [1] 26
```

```
set.seed(67)
df <- lapply(sheet_names, function(sheet) {
  as.matrix(read.xlsx(file, sheet = sheet, colNames = TRUE))
})

names(df) <- sheet_names

# Check one location
str(df[[1]])
```

```
##  num [1:288, 1:384] 138 129 121 115 109 104 100 97 95 93 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:288] "1" "2" "3" "4" ...
##   ..$ : chr [1:384] "WkDay-1" "WkDay-2" "WkDay-3" "WkDay-4" ...
```

Helpers for anomalies

```
choose_k_pca <- function(pca, threshold = 0.90) {
  var_expl <- pca$sdev^2
  var_expl <- var_expl / sum(var_expl)
  cumvar   <- cumsum(var_expl)
  k <- which(cumvar >= threshold)[1]
  return(k)
}

find_anomalies_from_scores <- function(score_mat, multiplier = 1.5) {
  # score_mat: rows = days, cols = components
  n_days <- nrow(score_mat)
  is_outlier <- rep(FALSE, n_days)

  for (j in seq_len(ncol(score_mat))) {
    x <- score_mat[, j]
    stats <- boxplot.stats(x, coef = multiplier)
```

```
    out_idx <- which(x %in% stats$out)
    is_outlier[out_idx] <- TRUE
  }

  which(is_outlier)  # returns row indices of anomalous days
}
```

```
analyze_location <- function(loc_name,
                             X,
                             var_expl_threshold = 0.90,
                             max_factors        = 3,
                             anomaly_coef       = 1.5,
                             do_ica             = FALSE,
                             n_ica_comp         = 3) {

  message("\nProcessing location: ", loc_name)

  # ---- 1. Transpose matrix so rows = days, columns = timepoints ----
  X_t <- t(X)
  X_centered <- scale(X_t, center = TRUE, scale = FALSE)

  # ---- 2. PCA ----
  pca <- prcomp(X_centered, center = FALSE, scale. = FALSE)
  k_pca <- choose_k_pca(pca, threshold = var_expl_threshold)

  pca_scores   <- pca$x[, 1:k_pca, drop = FALSE]
  pca_loadings <- pca$rotation[, 1:k_pca, drop = FALSE]

  # ---- 3. PCA-based anomaly detection ----
  pca_anom_idx  <- find_anomalies_from_scores(pca_scores, multiplier = anomaly_coef)
  pca_anom_days <- rownames(pca_scores)[pca_anom_idx]

  # ---- 4. Factor Analysis (on PCA scores for numerical stability) ----
  fa_model     <- NULL
  fa_scores    <- NULL
  fa_loadings  <- NULL
  fa_anom_days <- character(0)

  # Use fewer factors for interpretability (typically 2-5)
  n_factors <- min(max_factors, k_pca, 5)

  if (n_factors >= 1) {
    # Use FA on PCA scores to avoid singularity issues
    n_pcs_for_fa <- min(50, k_pca)
```

```
    n_factors_fa <- min(n_factors, n_pcs_for_fa - 1)

    if (n_factors_fa >= 1 && n_pcs_for_fa >= 3) {
      fa_model <- tryCatch(
        factanal(pca_scores[, 1:n_pcs_for_fa, drop = FALSE],
                 factors = n_factors_fa,
                 scores = "regression",
                 rotation = "varimax"),
        error = function(e) {
          message("  FA failed for ", loc_name, ": ", e$message)
          return(NULL)
        }
      )

      if (!is.null(fa_model)) {
        fa_scores <- fa_model$scores
        # Map loadings back to original time-of-day space
        fa_loadings <- pca_loadings[, 1:n_pcs_for_fa] %*% fa_model$loadings[, , drop = F
        fa_anom_idx <- find_anomalies_from_scores(fa_scores, multiplier = anomaly_coef)
        fa_anom_days <- rownames(fa_scores)[fa_anom_idx]
        message("  FA succeeded on PCA scores")
      }
    } else {
      message("  FA skipped for ", loc_name, ": insufficient PCs")
    }
  } else {
    message("  FA skipped for ", loc_name, ": insufficient factors")
  }

  # ---ica on PCA scores---

  ica_result <- NULL
  ica_anom_days <- character(0)
  ica_scores    <- NULL
  ica_loadings  <- NULL

if (do_ica && k_pca >= 2) {
  n_ica_to_use <- min(n_ica_comp, k_pca)

  ica_result <- tryCatch(
    fastICA(pca_scores, n.comp = n_ica_to_use, method = "C"),
    error = function(e) {
      message("  ICA failed for ", loc_name, ": ", e$message)
      return(NULL)
```

```
    }
  )

  if (!is.null(ica_result)) {
    # ICA scores (strength of each source per day)
    ica_scores <- ica_result$S

    # Map ICA loadings back to original time-of-day space
    # ica_result$A maps PCA space -> ICA sources
    # pca_loadings maps original space -> PCA space
    ica_loadings <- pca_loadings[, 1:n_ica_to_use, drop = FALSE] %*% ica_result$A

    if (!is.null(ica_scores) && nrow(ica_scores) > 0) {
      message("  ICA succeeded with ", ncol(ica_scores), " sources.")

      # Ensure day names - use pca_scores rownames which are correct!
      if (is.null(rownames(ica_scores))) {
        rownames(ica_scores) <- rownames(pca_scores)
      }

      # Calculate ICA anomalies
      ica_anom_idx <- find_anomalies_from_scores(ica_scores, multiplier = anomaly_coef)
      ica_anom_days <- rownames(ica_scores)[ica_anom_idx]

    } else {
      message("  ICA failed to produce valid scores for ", loc_name)
      ica_result <- NULL
      ica_scores <- NULL
      ica_loadings <- NULL
    }
  }
}

  # ---- 6. Return results (Now includes all ICA components) ----
  list(
    location       = loc_name,
    pca            = pca,
    k_pca          = k_pca,
    pca_scores     = pca_scores,
    pca_loadings   = pca_loadings,
    pca_anom_days  = pca_anom_days,
    fa_model       = fa_model,
    fa_scores      = fa_scores,
    fa_loadings    = fa_loadings,
```

```
    fa_anom_days   = fa_anom_days,
    ica            = ica_result,
    # ADDED: Explicitly return ICA components
    ica_scores     = ica_scores,
    ica_loadings   = ica_loadings,
    ica_anom_days  = ica_anom_days
  )
}

# ---- Run for all locations ----
location_results <- lapply(names(df), function(loc_name) {
  analyze_location(
    loc_name = loc_name,
    X        = df[[loc_name]],
    var_expl_threshold = 0.90,
    max_factors        = 3,
    anomaly_coef       = 1.5,
    do_ica             = TRUE
  )
})
```

```
##
## Processing location: Loc1

##   FA failed for Loc1: 3 factors are too many for 5 variables

##   ICA succeeded with 3 sources.

##
## Processing location: Loc2

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc3

##   FA failed for Loc3: 3 factors are too many for 5 variables

##   ICA succeeded with 3 sources.
```

```
##
## Processing location: Loc4

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc5

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc6

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc7

##   FA skipped for Loc7: insufficient PCs

##   ICA succeeded with 2 sources.

##
## Processing location: Loc8

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc9

##   FA failed for Loc9: 3 factors are too many for 4 variables

##   ICA succeeded with 3 sources.
```

```
##
## Processing location: Loc10

##   FA failed for Loc10: 3 factors are too many for 5 variables

##   ICA succeeded with 3 sources.

##
## Processing location: Loc11

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc12

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc13

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc14

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc15

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.
```

```
##
## Processing location: Loc16

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc17

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc18

##   FA failed for Loc18: 3 factors are too many for 5 variables

##   ICA succeeded with 3 sources.

##
## Processing location: Loc19

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc20

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc21

##   FA failed for Loc21: 3 factors are too many for 5 variables

##   ICA succeeded with 3 sources.
```

```
##
## Processing location: Loc22

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc23

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc24

##   FA failed for Loc24: 2 factors are too many for 3 variables

##   ICA succeeded with 3 sources.

##
## Processing location: Loc25

##   FA succeeded on PCA scores

##   ICA succeeded with 3 sources.

##
## Processing location: Loc26

##   FA skipped for Loc26: insufficient PCs

##   ICA succeeded with 2 sources.
```

```r
names(location_results) <- names(df)
```

List of anomalies

```r
# Print anomalies detected by each method for all locations
for (loc_name in names(location_results)) {
  res <- location_results[[loc_name]]
  cat("\n", rep("=", 60), "\n", sep = "")
  cat("Location:", loc_name, "\n")
  cat(rep("=", 60), "\n")
  cat("PCA: retained", res$k_pca, "components\n")
  cat("PCA anomalies (", length(res$pca_anom_days), "):",
      paste(res$pca_anom_days, collapse = ", "), "\n")
  cat("FA anomalies (", length(res$fa_anom_days), "):",
      paste(res$fa_anom_days, collapse = ", "), "\n")
  cat("ICA anomalies (", length(res$ica_anom_days), "):",
  paste(res$ica_anom_days, collapse = ", "), "\n")
  cat("\n")
}
```

```
##
## ============================================================
## Location: Loc1
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 5 components
## PCA anomalies ( 30 ): WkDay-1, WkDay-43, WkDay-46, WkDay-56, WkDay-77, WkDay-89, WkDa
## FA anomalies ( 0 ):
## ICA anomalies ( 17 ): WkDay-11, WkDay-38, WkDay-150, WkDay-189, WkDay-212, WkDay-213,
##
##
## ============================================================
## Location: Loc2
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 7 components
## PCA anomalies ( 20 ): WkDay-1, WkDay-37, WkDay-61, WkDay-79, WkDay-80, WkDay-118, WkD
## FA anomalies ( 10 ): WkDay-27, WkDay-37, WkDay-61, WkDay-91, WkDay-124, WkDay-147, Wk
## ICA anomalies ( 14 ): WkDay-1, WkDay-78, WkDay-79, WkDay-80, WkDay-120, WkDay-177, Wk
##
##
## ============================================================
## Location: Loc3
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 5 components
## PCA anomalies ( 24 ): WkDay-5, WkDay-23, WkDay-26, WkDay-28, WkDay-51, WkDay-72, WkDa
## FA anomalies ( 0 ):
## ICA anomalies ( 17 ): WkDay-17, WkDay-23, WkDay-28, WkDay-67, WkDay-76, WkDay-99, WkD
##
##
```

```
## ============================================================
## Location: Loc4
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 7 components
## PCA anomalies ( 20 ): WkDay-5, WkDay-27, WkDay-29, WkDay-34, WkDay-73, WkDay-91, WkDa
## FA anomalies ( 11 ): WkDay-5, WkDay-27, WkDay-34, WkDay-73, WkDay-95, WkDay-117, WkDa
## ICA anomalies ( 10 ): WkDay-27, WkDay-36, WkDay-159, WkDay-214, WkDay-241, WkDay-261,
##
##
## ============================================================
## Location: Loc5
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 7 components
## PCA anomalies ( 18 ): WkDay-1, WkDay-21, WkDay-117, WkDay-138, WkDay-201, WkDay-203,
## FA anomalies ( 5 ): WkDay-21, WkDay-201, WkDay-223, WkDay-228, WkDay-332
## ICA anomalies ( 8 ): WkDay-5, WkDay-6, WkDay-21, WkDay-203, WkDay-222, WkDay-223, WkD
##
##
## ============================================================
## Location: Loc6
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 6 components
## PCA anomalies ( 6 ): WkDay-111, WkDay-232, WkDay-267, WkDay-309, WkDay-311, WkDay-336
## FA anomalies ( 4 ): WkDay-232, WkDay-267, WkDay-309, WkDay-336
## ICA anomalies ( 3 ): WkDay-267, WkDay-309, WkDay-336
##
##
## ============================================================
## Location: Loc7
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 2 components
## PCA anomalies ( 4 ): WkDay-20, WkDay-220, WkDay-256, WkDay-259
## FA anomalies ( 0 ):
## ICA anomalies ( 15 ): WkDay-7, WkDay-8, WkDay-10, WkDay-20, WkDay-219, WkDay-220, WkD
##
##
## ============================================================
## Location: Loc8
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 8 components
## PCA anomalies ( 23 ): WkDay-1, WkDay-7, WkDay-48, WkDay-84, WkDay-86, WkDay-98, WkDay
## FA anomalies ( 10 ): WkDay-48, WkDay-86, WkDay-109, WkDay-113, WkDay-128, WkDay-132,
## ICA anomalies ( 6 ): WkDay-1, WkDay-7, WkDay-230, WkDay-232, WkDay-327, WkDay-382
##
##
```

```
## ============================================================
## Location: Loc9
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 4 components
## PCA anomalies ( 14 ): WkDay-24, WkDay-40, WkDay-50, WkDay-60, WkDay-127, WkDay-248, W
## FA anomalies ( 0 ):
## ICA anomalies ( 12 ): WkDay-24, WkDay-37, WkDay-50, WkDay-206, WkDay-248, WkDay-289,
##
##
## ============================================================
## Location: Loc10
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 5 components
## PCA anomalies ( 13 ): WkDay-46, WkDay-62, WkDay-86, WkDay-100, WkDay-169, WkDay-195,
## FA anomalies ( 0 ):
## ICA anomalies ( 7 ): WkDay-9, WkDay-10, WkDay-51, WkDay-62, WkDay-169, WkDay-171, WkD
##
##
## ============================================================
## Location: Loc11
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 6 components
## PCA anomalies ( 24 ): WkDay-1, WkDay-4, WkDay-7, WkDay-13, WkDay-15, WkDay-26, WkDay-
## FA anomalies ( 10 ): WkDay-1, WkDay-4, WkDay-26, WkDay-54, WkDay-66, WkDay-76, WkDay-
## ICA anomalies ( 9 ): WkDay-1, WkDay-2, WkDay-13, WkDay-14, WkDay-54, WkDay-222, WkDay
##
##
## ============================================================
## Location: Loc12
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 8 components
## PCA anomalies ( 23 ): WkDay-1, WkDay-4, WkDay-7, WkDay-19, WkDay-55, WkDay-131, WkDay
## FA anomalies ( 14 ): WkDay-7, WkDay-19, WkDay-55, WkDay-70, WkDay-77, WkDay-152, WkDa
## ICA anomalies ( 19 ): WkDay-1, WkDay-7, WkDay-19, WkDay-70, WkDay-77, WkDay-121, WkDa
##
##
## ============================================================
## Location: Loc13
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 9 components
## PCA anomalies ( 20 ): WkDay-1, WkDay-25, WkDay-31, WkDay-33, WkDay-34, WkDay-98, WkDa
## FA anomalies ( 9 ): WkDay-1, WkDay-50, WkDay-98, WkDay-139, WkDay-269, WkDay-271, WkD
## ICA anomalies ( 17 ): WkDay-1, WkDay-13, WkDay-16, WkDay-27, WkDay-37, WkDay-98, WkDa
##
##
```

```
## ============================================================
## Location: Loc14
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 8 components
## PCA anomalies ( 21 ): WkDay-1, WkDay-4, WkDay-24, WkDay-64, WkDay-106, WkDay-163, WkD
## FA anomalies ( 8 ): WkDay-24, WkDay-163, WkDay-205, WkDay-217, WkDay-274, WkDay-330,
## ICA anomalies ( 6 ): WkDay-81, WkDay-106, WkDay-175, WkDay-230, WkDay-232, WkDay-364
##
##
## ============================================================
## Location: Loc15
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 8 components
## PCA anomalies ( 20 ): WkDay-1, WkDay-5, WkDay-33, WkDay-66, WkDay-80, WkDay-83, WkDay
## FA anomalies ( 10 ): WkDay-1, WkDay-5, WkDay-33, WkDay-161, WkDay-232, WkDay-308, WkD
## ICA anomalies ( 8 ): WkDay-1, WkDay-161, WkDay-184, WkDay-214, WkDay-316, WkDay-317,
##
##
## ============================================================
## Location: Loc16
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 9 components
## PCA anomalies ( 29 ): WkDay-1, WkDay-9, WkDay-36, WkDay-45, WkDay-62, WkDay-77, WkDay
## FA anomalies ( 10 ): WkDay-1, WkDay-178, WkDay-230, WkDay-232, WkDay-260, WkDay-317,
## ICA anomalies ( 11 ): WkDay-9, WkDay-78, WkDay-191, WkDay-213, WkDay-222, WkDay-229,
##
##
## ============================================================
## Location: Loc17
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 8 components
## PCA anomalies ( 31 ): WkDay-1, WkDay-2, WkDay-6, WkDay-13, WkDay-46, WkDay-54, WkDay-
## FA anomalies ( 12 ): WkDay-1, WkDay-6, WkDay-13, WkDay-66, WkDay-80, WkDay-106, WkDay
## ICA anomalies ( 8 ): WkDay-2, WkDay-24, WkDay-54, WkDay-109, WkDay-212, WkDay-232, Wk
##
##
## ============================================================
## Location: Loc18
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 5 components
## PCA anomalies ( 15 ): WkDay-1, WkDay-26, WkDay-51, WkDay-93, WkDay-184, WkDay-199, Wk
## FA anomalies ( 0 ):
## ICA anomalies ( 5 ): WkDay-1, WkDay-239, WkDay-258, WkDay-259, WkDay-325
##
##
```

```
## ============================================================
## Location: Loc19
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 9 components
## PCA anomalies ( 21 ): WkDay-1, WkDay-14, WkDay-74, WkDay-76, WkDay-80, WkDay-114, WkD
## FA anomalies ( 10 ): WkDay-14, WkDay-80, WkDay-104, WkDay-114, WkDay-117, WkDay-176,
## ICA anomalies ( 13 ): WkDay-1, WkDay-24, WkDay-28, WkDay-74, WkDay-82, WkDay-137, WkD
##
##
## ============================================================
## Location: Loc20
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 6 components
## PCA anomalies ( 19 ): WkDay-1, WkDay-27, WkDay-29, WkDay-31, WkDay-49, WkDay-80, WkDa
## FA anomalies ( 6 ): WkDay-1, WkDay-29, WkDay-31, WkDay-49, WkDay-354, WkDay-377
## ICA anomalies ( 4 ): WkDay-1, WkDay-205, WkDay-206, WkDay-212
##
##
## ============================================================
## Location: Loc21
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 5 components
## PCA anomalies ( 10 ): WkDay-83, WkDay-133, WkDay-188, WkDay-215, WkDay-223, WkDay-224
## FA anomalies ( 0 ):
## ICA anomalies ( 7 ): WkDay-1, WkDay-52, WkDay-232, WkDay-264, WkDay-327, WkDay-328, W
##
##
## ============================================================
## Location: Loc22
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 6 components
## PCA anomalies ( 13 ): WkDay-1, WkDay-11, WkDay-69, WkDay-70, WkDay-127, WkDay-140, Wk
## FA anomalies ( 10 ): WkDay-11, WkDay-69, WkDay-70, WkDay-140, WkDay-159, WkDay-221, W
## ICA anomalies ( 13 ): WkDay-1, WkDay-2, WkDay-69, WkDay-71, WkDay-127, WkDay-222, WkD
##
##
## ============================================================
## Location: Loc23
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 6 components
## PCA anomalies ( 20 ): WkDay-1, WkDay-33, WkDay-35, WkDay-47, WkDay-137, WkDay-156, Wk
## FA anomalies ( 16 ): WkDay-1, WkDay-33, WkDay-37, WkDay-38, WkDay-47, WkDay-137, WkDa
## ICA anomalies ( 19 ): WkDay-1, WkDay-33, WkDay-37, WkDay-38, WkDay-137, WkDay-176, Wk
##
##
```

```
## ============================================================
## Location: Loc24
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 3 components
## PCA anomalies ( 5 ): WkDay-97, WkDay-162, WkDay-251, WkDay-278, WkDay-340
## FA anomalies ( 0 ):
## ICA anomalies ( 12 ): WkDay-57, WkDay-63, WkDay-100, WkDay-103, WkDay-162, WkDay-239,
##
##
## ============================================================
## Location: Loc25
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 7 components
## PCA anomalies ( 22 ): WkDay-1, WkDay-31, WkDay-33, WkDay-35, WkDay-36, WkDay-37, WkDa
## FA anomalies ( 6 ): WkDay-95, WkDay-102, WkDay-251, WkDay-257, WkDay-295, WkDay-384
## ICA anomalies ( 19 ): WkDay-35, WkDay-36, WkDay-37, WkDay-47, WkDay-52, WkDay-63, WkD
##
##
## ============================================================
## Location: Loc26
## = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = = =
## PCA: retained 2 components
## PCA anomalies ( 6 ): WkDay-1, WkDay-33, WkDay-226, WkDay-227, WkDay-235, WkDay-239
## FA anomalies ( 0 ):
## ICA anomalies ( 7 ): WkDay-18, WkDay-221, WkDay-222, WkDay-226, WkDay-231, WkDay-234,
```

Plot PCA Loadings

```r
plot_pca_loadings <- function(res, loc_name, n_comp = 3) {
  loadings <- res$pca_loadings[, 1:min(n_comp, ncol(res$pca_loadings)), drop = FALSE]

  # Convert time intervals to hours (288 intervals = 24 hours)
  time_hours <- seq(0, 24, length.out = nrow(loadings))

  df <- data.frame(Time = time_hours, loadings)
  colnames(df) <- c("Time", paste0("PC", 1:ncol(loadings)))

  df_long <- tidyr::pivot_longer(df, cols = -Time,
                                 names_to = "Component",
                                 values_to = "Loading")

  ggplot(df_long, aes(x = Time, y = Loading, color = Component)) +
    geom_line(size = 1) +
```

```
    facet_wrap(~Component, ncol = 1, scales = "free_y") +
    labs(title = paste("PCA Loadings -", loc_name),
         x = "Hour of Day",
         y = "Loading") +
    theme_minimal() +
    scale_x_continuous(breaks = seq(0, 24, 4))
}

for (loc_name in names(location_results)) {
  print(plot_pca_loadings(location_results[[loc_name]], loc_name))
}
```

View Anomalies with Boxplot

```
plot_pca_score_boxplot <- function(res, loc_name, include_fa = FALSE) {
  scores <- res$pca_scores
  days <- rownames(scores)

  df_long <- data.frame(
    Day = rep(days, ncol(scores)),
    PC = rep(colnames(scores), each = nrow(scores)),
    Score = c(scores)
  )

  # Identify anomalies
  df_long$IsPCAAnomaly <- ifelse(df_long$Day %in% res$pca_anom_days, "PCA Anomaly", "Nor

  # OPTIONALLY include FA anomalies
  if (include_fa && !is.null(res$fa_scores)) {
    df_long$IsFAAnomaly <- ifelse(df_long$Day %in% res$fa_anom_days, "FA Anomaly", "Norm
    df_long$AnomalyType <- ifelse(df_long$IsPCAAnomaly == "PCA Anomaly",
                                  "PCA Anomaly",
                                  ifelse(df_long$IsFAAnomaly == "FA Anomaly", "FA Anomal
  } else {
    df_long$AnomalyType <- df_long$IsPCAAnomaly
  }

  ggplot(df_long, aes(x = PC, y = Score)) +
    geom_boxplot(outlier.shape = NA) +   # Remove default black dots
    geom_jitter(aes(color = AnomalyType), alpha = 0.6, width = 0.15, size = 2) +
    scale_color_manual(values = c("Normal" = "gray",
                                  "PCA Anomaly" = "red",
```

```
                                    "FA Anomaly" = "blue")) +
    labs(title = paste("PC Score Boxplots with Anomalies -", loc_name),
         x = "Principal Component",
         y = "Score",
         color = "Detection Method") +
    theme_minimal() +
    theme(legend.position = "bottom")
}

for (loc_name in names(location_results)) {
  print(plot_pca_score_boxplot(location_results[[loc_name]], loc_name, include_fa=TRUE))
}
```

Anomalies overlay boxplot with labels

```
overlay <- function(res, loc_name, include_fa = FALSE) {
  scores <- res$pca_scores
  days <- rownames(scores)

  df_long <- data.frame(
    Day = rep(days, ncol(scores)),
    PC = rep(colnames(scores), each = nrow(scores)),
    Score = c(scores)
  )

  # Correct anomaly identification
  if (include_fa && !is.null(res$fa_scores)) {
    df_long$AnomalyType <- case_when(
      df_long$Day %in% intersect(res$pca_anom_days, res$fa_anom_days) ~ "Both PCA & FA",
      df_long$Day %in% setdiff(res$pca_anom_days, res$fa_anom_days) ~ "PCA Anomaly",
      df_long$Day %in% setdiff(res$fa_anom_days, res$pca_anom_days) ~ "FA Anomaly",
      TRUE ~ "Normal"
    )
  } else {
    df_long$AnomalyType <- ifelse(df_long$Day %in% res$pca_anom_days, "PCA Anomaly", "No
  }

ggplot(df_long, aes(x = PC, y = Score)) +
  geom_boxplot(outlier.shape = NA) +

  # Normal points first (light and small)
  geom_jitter(data = subset(df_long, AnomalyType == "Normal"),
```

```
              color = "gray70", alpha = 0.3, width = 0.15, size = 1) +

  # PCA-only anomalies
  geom_point(data = subset(df_long, AnomalyType == "PCA Anomaly"),
             aes(color = "PCA Anomaly"), size = 1.0, alpha = 0.9) +

  # FA-only anomalies
  geom_point(data = subset(df_long, AnomalyType == "FA Anomaly"),
             aes(color = "FA Anomaly"), size = 1.0, alpha = 0.9) +

  # Both PCA & FA
  geom_point(data = subset(df_long, AnomalyType == "Both PCA & FA"),
             aes(color = "Both PCA & FA"), size = 1.3, alpha = 1, shape = 8) +

  scale_color_manual(values = c(
    "PCA Anomaly" = "red",
    "FA Anomaly" = "blue",
    "Both PCA & FA" = "purple"
  )) +
  labs(title = paste("Anomalies Highlighted Clearly -", loc_name),
       x = "Principal Component", y = "Score", color = "Anomaly Source") +
  theme_minimal() +
  theme(legend.position = "bottom")
}

  for (loc_name in names(location_results)) {
  print(overlay(location_results[[loc_name]], loc_name, include_fa=TRUE))
}
```

visualize ica loadings

```
## ICA Source Loadings Plot
plot_ica_loadings <- function(res, loc_name, max_plot_comp = 4) {
  if (is.null(res$ica) || is.null(res$ica$A) || ncol(res$ica$A) == 0) {
    message("ICA Loadings (A matrix) not available for ", loc_name)
    return(NULL)
  }

  # ICA Loadings are stored in the 'A' matrix
  loadings <- res$ica$A
  n_comp <- ncol(loadings)
```

```
  # Limit components being plotted for stability
  n_comp_to_plot <- min(n_comp, max_plot_comp)
  loadings <- loadings[, 1:n_comp_to_plot, drop = FALSE]

  # Convert time intervals to hours
  time_hours <- seq(0, 24, length.out = nrow(loadings))

  df <- data.frame(Time = time_hours, loadings)
  colnames(df) <- c("Time", paste0("Source", 1:n_comp_to_plot))

  df_long <- tidyr::pivot_longer(df, cols = -Time,
                                 names_to = "Source",
                                 values_to = "Loading")

  ggplot(df_long, aes(x = Time, y = Loading, color = Source)) +
    geom_line(size = 1) +
    facet_wrap(~Source, ncol = 1, scales = "free_y") +
    labs(title = paste("ICA Source Loadings (A Matrix) -", loc_name),
         x = "Hour of Day",
         y = "Loading/Mixing Weight") +
    theme_minimal() +
    scale_x_continuous(breaks = seq(0, 24, 4))
}

# Loop over locations to generate the plots
for (loc_name in names(location_results)) {
  # Wrap in tryCatch for final safety
  tryCatch({
    print(plot_ica_loadings(location_results[[loc_name]], loc_name))
  }, error = function(e) {
    warning(paste("Failed to plot ICA Loadings for", loc_name, ":", e$message))
  })
}
```

Better: Plot Anomalous Days vs Normal Days

```
anomaly_timeline <- function(res, loc_name) {
  all_days <- rownames(res$pca_scores)
  df_timeline <- data.frame(DayName = all_days, DayIndex = 1:length(all_days))

  df_timeline <- df_timeline %>%
    mutate(AnomalyType = case_when(
```

```
    DayName %in% intersect(intersect(res$pca_anom_days, res$fa_anom_days), res$ica_ano
    DayName %in% setdiff(intersect(res$pca_anom_days, res$fa_anom_days), res$ica_anom_
    DayName %in% setdiff(intersect(res$pca_anom_days, res$ica_anom_days), res$fa_anom_
    DayName %in% setdiff(intersect(res$fa_anom_days, res$ica_anom_days), res$pca_anom_
    DayName %in% setdiff(setdiff(res$pca_anom_days, res$fa_anom_days), res$ica_anom_da
    DayName %in% setdiff(setdiff(res$fa_anom_days, res$pca_anom_days), res$ica_anom_da
    DayName %in% setdiff(setdiff(res$ica_anom_days, res$pca_anom_days), res$fa_anom_da
    TRUE ~ "Normal"
  ))

df_anomalies <- df_timeline %>% filter(AnomalyType != "Normal")

if (nrow(df_anomalies) == 0) {
  message("No anomalies to plot for ", loc_name)
  return(NULL)
}

ggplot(df_anomalies, aes(x = DayIndex, y = 1)) +

  # Clean horizontal baseline
  geom_hline(yintercept = 1, color = "gray85", linewidth = 0.3) +

  # nomaly points only
  geom_point(aes(color = AnomalyType, shape = AnomalyType), size = 3) +

  scale_color_manual(values = c(
    "PCA Only" = "red",
    "FA Only" = "blue",
    "ICA Only" = "green",
    "PCA & FA" = "purple",
    "PCA & ICA" = "#FF7F50",
    "FA & ICA" = "#00CED1",
    "All Three" = "black"
  )) +
  scale_shape_manual(values = c(
    "PCA Only" = 19,
    "FA Only" = 17,
    "ICA Only" = 15,
    "PCA & FA" = 8,
    "PCA & ICA" = 10,
    "FA & ICA" = 13,
    "All Three" = 4
  )) +
  labs(title = paste("Anomaly Temporal Sequence -", loc_name),
```

```
         x = "Day Index (Time Sequence)",
         y = "",
         color = "Detection Source",
         shape = "Detection Source") +
    theme_minimal() +
    theme(axis.text.y = element_blank(),
          axis.ticks.y = element_blank(),
          panel.grid.major.y = element_blank(),
          panel.grid.minor.y = element_blank(),
          legend.position = "bottom")
}

# Plot for all locations
for (loc_name in names(location_results)) {
  print(anomaly_timeline(location_results[[loc_name]], loc_name))
}
```

View Summary of All Locations

```
# Create summary table
summary_df <- data.frame(
  Location = names(location_results),

  # Total unique anomalies across all methods
  Total_Anomalies = sapply(location_results, \(x)
    length(union(union(x$pca_anom_days, x$fa_anom_days), x$ica_anom_days))),

  # All three methods detected
  All_Three = sapply(location_results, \(x)
    length(intersect(intersect(x$pca_anom_days, x$fa_anom_days), x$ica_anom_days))),

  # Two methods detected
  PCA_FA = sapply(location_results, \(x)
    length(setdiff(intersect(x$pca_anom_days, x$fa_anom_days), x$ica_anom_days))),

  PCA_ICA = sapply(location_results, \(x)
    length(setdiff(intersect(x$pca_anom_days, x$ica_anom_days), x$fa_anom_days))),

  FA_ICA = sapply(location_results, \(x)
    length(setdiff(intersect(x$fa_anom_days, x$ica_anom_days), x$pca_anom_days))),

  # Single method only
```

```
  PCA_Only = sapply(location_results, \(x)
    length(setdiff(setdiff(x$pca_anom_days, x$fa_anom_days), x$ica_anom_days))),

  FA_Only = sapply(location_results, \(x)
    length(setdiff(setdiff(x$fa_anom_days, x$pca_anom_days), x$ica_anom_days))),

  ICA_Only = sapply(location_results, \(x)
    length(setdiff(setdiff(x$ica_anom_days, x$pca_anom_days), x$fa_anom_days))),

  FA_Success = sapply(location_results, \(x) !is.null(x$fa_model)),
  ICA_Success = sapply(location_results, \(x) !is.null(x$ica_scores))
)

df_long_summary <- summary_df %>%
  pivot_longer(cols = -Location,
               names_to = "AnomalyType",
               values_to = "Count") %>%
  filter(Count > 0)



print(summary_df)
```

```
##       Location Total_Anomalies All_Three PCA_FA PCA_ICA FA_ICA PCA_Only FA_Only
## Loc1      Loc1              39         0      0       8      0       22       0
## Loc2      Loc2              27         0      8       9      0        3       2
## Loc3      Loc3              31         0      0      10      0       14       0
## Loc4      Loc4              27         3      7       1      0        9       1
## Loc5      Loc5              23         3      1       1      0       13       1
## Loc6      Loc6               6         3      1       0      0        2       0
## Loc7      Loc7              15         0      0       4      0        0       0
## Loc8      Loc8              25         0      9       5      0        9       1
## Loc9      Loc9              17         0      0       9      0        5       0
## Loc10    Loc10              17         0      0       3      0       10       0
## Loc11    Loc11              28         3      7       2      0       12       0
## Loc12    Loc12              29         8      4       5      2        6       0
## Loc13    Loc13              32         5      1       3      0       11       3
## Loc14    Loc14              26         0      7       2      0       12       1
## Loc15    Loc15              24         4      4       2      0       10       2
## Loc16    Loc16              39         0      6       5      0       18       4
## Loc17    Loc17              36         0     11       4      0       16       1
## Loc18    Loc18              17         0      0       3      0       12       0
## Loc19    Loc19              30         1      8       4      0        8       1
## Loc20    Loc20              21         1      4       2      0       12       1
```

```
## Loc21    Loc21              14         0      0       3      0         7       0
## Loc22    Loc22              22         1      8       4      0         0       1
## Loc23    Loc23              26         9      2       5      4         4       1
## Loc24    Loc24              15         0      0       2      0         3       0
## Loc25    Loc25              34         4      2       3      0        13       0
## Loc26    Loc26              12         0      0       1      0         5       0
##       ICA_Only FA_Success ICA_Success
## Loc1         9      FALSE        TRUE
## Loc2         5       TRUE        TRUE
## Loc3         7      FALSE        TRUE
## Loc4         6       TRUE        TRUE
## Loc5         4       TRUE        TRUE
## Loc6         0       TRUE        TRUE
## Loc7        11      FALSE        TRUE
## Loc8         1       TRUE        TRUE
## Loc9         3      FALSE        TRUE
## Loc10        4      FALSE        TRUE
## Loc11        4       TRUE        TRUE
## Loc12        4       TRUE        TRUE
## Loc13        9       TRUE        TRUE
## Loc14        4       TRUE        TRUE
## Loc15        2       TRUE        TRUE
## Loc16        6       TRUE        TRUE
## Loc17        4       TRUE        TRUE
## Loc18        2      FALSE        TRUE
## Loc19        8       TRUE        TRUE
## Loc20        1       TRUE        TRUE
## Loc21        4      FALSE        TRUE
## Loc22        8       TRUE        TRUE
## Loc23        1       TRUE        TRUE
## Loc24       10      FALSE        TRUE
## Loc25       12       TRUE        TRUE
## Loc26        6      FALSE        TRUE
```

Stacked bar chart of anomaly counts

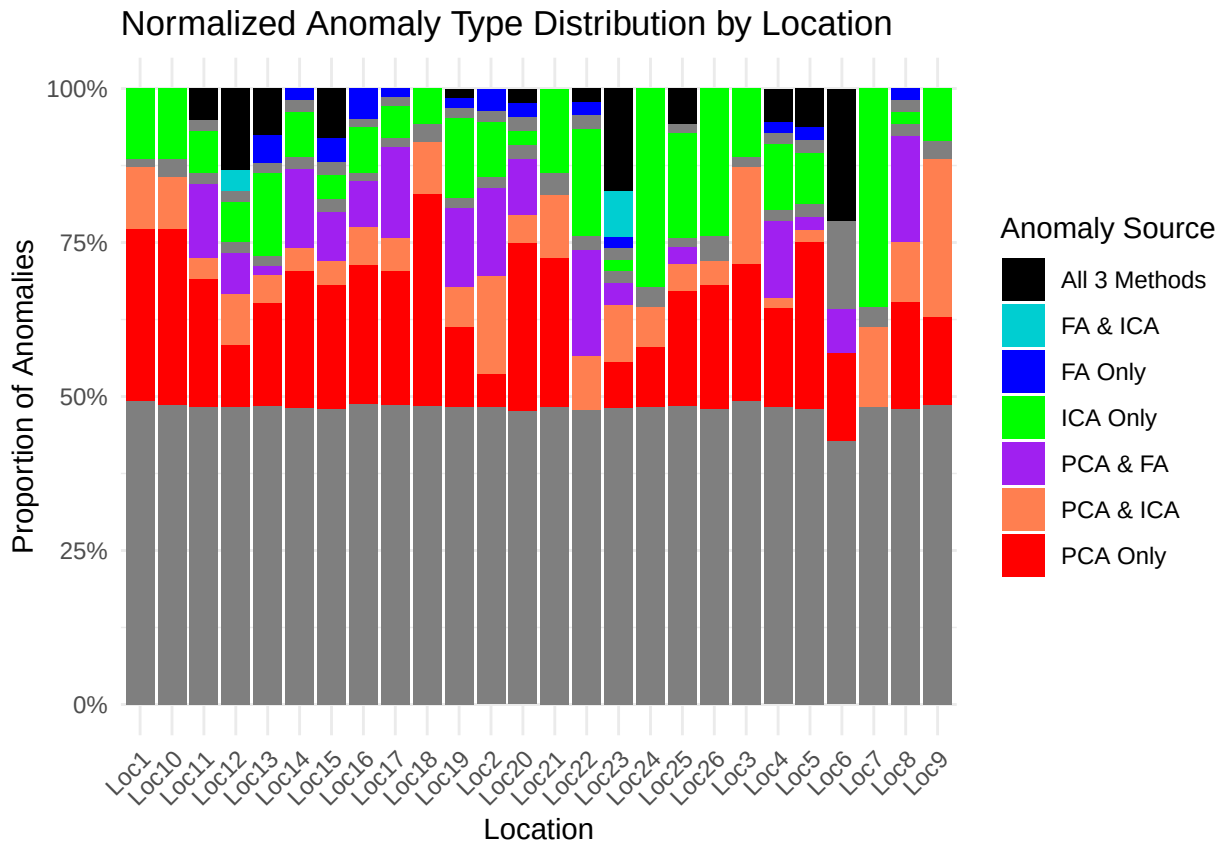
```
df_long_summary <- summary_df %>%
  pivot_longer(cols = c(PCA_Only, FA_Only, ICA_Only,
                        PCA_FA, PCA_ICA, FA_ICA, All_Three),
               names_to = "AnomalyType",
               values_to = "Count") %>%
  filter(Count > 0)
```

```r
fill_colors <- c(
  "PCA_Only" = "red",
  "FA_Only" = "blue",
  "ICA_Only" = "green",
  "PCA_FA" = "purple",
  "PCA_ICA" = "#FF7F50",
  "FA_ICA" = "#00CED1",
  "All_Three" = "black"
)

ggplot(df_long_summary, aes(x = Location, y = Count, fill = AnomalyType)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = fill_colors,
                    labels = c(
                      "PCA_Only" = "PCA Only",
                      "FA_Only" = "FA Only",
                      "ICA_Only" = "ICA Only",
                      "PCA_FA" = "PCA & FA",
                      "PCA_ICA" = "PCA & ICA",
                      "FA_ICA" = "FA & ICA",
                      "All_Three" = "All 3 Methods"
                    )) +
  labs(title = "Breakdown of Anomaly Types by Location",
       x = "Location",
       y = "Number of Anomalous Days",
       fill = "Anomaly Source") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

normalized

```r
# Compute proportion (normalized values)
df_long_summary <- df_long_summary %>%
  group_by(Location) %>%
  mutate(Proportion = Count / sum(Count)) %>%
  ungroup()

# Define colors for each category
fill_colors <- c(
  "PCA_Only" = "red",
  "FA_Only" = "blue",
  "ICA_Only" = "green",
```

```
  "PCA_FA" = "purple",
  "PCA_ICA" = "#FF7F50",
  "FA_ICA" = "#00CED1",
  "All_Three" = "black"
)

# Plot normalized stacked bar chart
ggplot(df_long_summary, aes(x = Location, y = Proportion, fill = AnomalyType)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = fill_colors,
                    labels = c(
                      "PCA_Only" = "PCA Only",
                      "FA_Only" = "FA Only",
                      "ICA_Only" = "ICA Only",
                      "PCA_FA" = "PCA & FA",
                      "PCA_ICA" = "PCA & ICA",
                      "FA_ICA" = "FA & ICA",
                      "All_Three" = "All 3 Methods"
                    )) +
  theme_minimal() +
  labs(title = "Normalized Anomaly Type Distribution by Location",
       x = "Location",
       y = "Proportion of Anomalies",
       fill = "Anomaly Source") +
  scale_y_continuous(labels = scales::percent_format()) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



component loadings heatmap

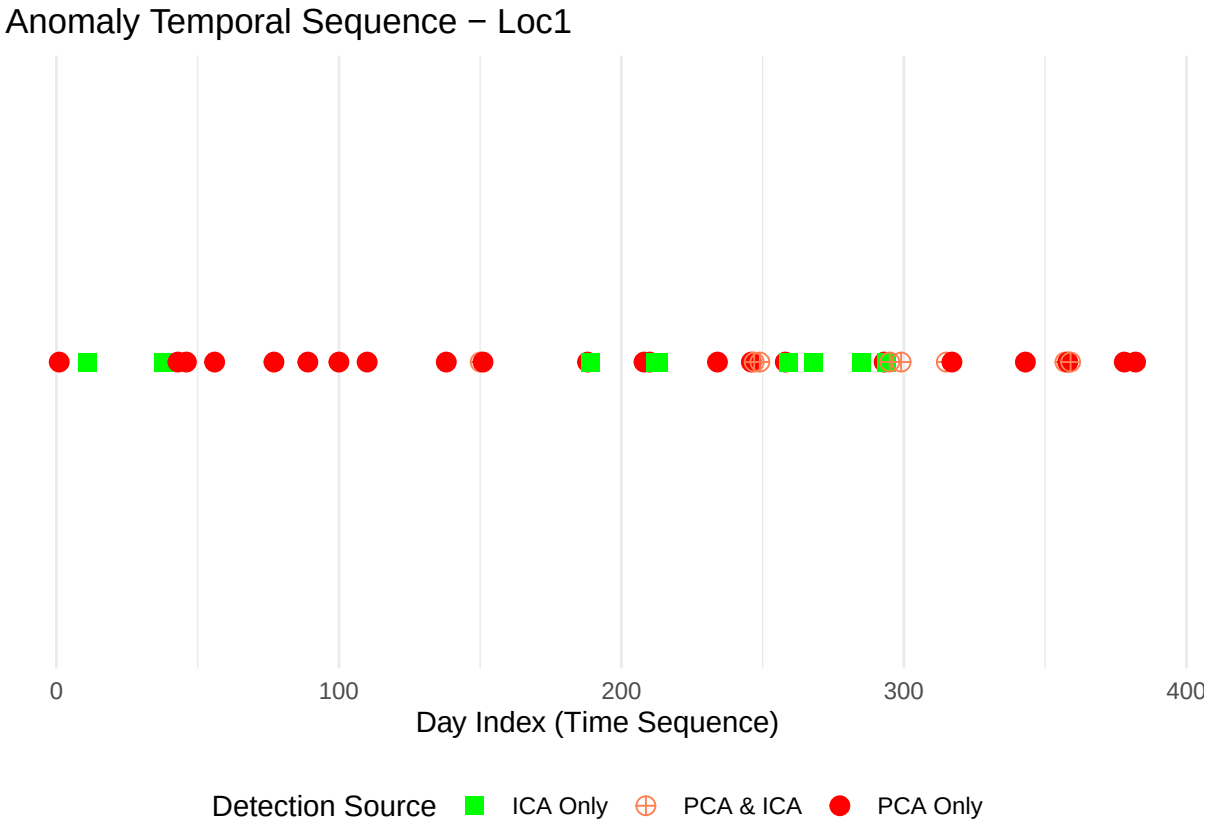
```r
# R Code for Component Loadings Heatmap (Example: Loc1):

loc_name <- names(location_results)[1]
res <- location_results[[loc_name]]
n_comp_to_plot <- min(3, ncol(res$pca_loadings))
loadings <- res$pca_loadings[, 1:n_comp_to_plot, drop = FALSE]

# Convert time intervals to index (1 to 288) and then approximate Hour
time_intervals <- 1:nrow(loadings)

df_loadings <- data.frame(TimeIndex = time_intervals, loadings)
colnames(df_loadings) <- c("TimeIndex", paste0("PC", 1:n_comp_to_plot))

df_long_loadings <- df_loadings %>%
  pivot_longer(cols = -TimeIndex,
               names_to = "Component",
               values_to = "Loading")

# Map TimeIndex to approximate Hour for axis readability (288 intervals = 24 hours)
df_long_loadings$Hour <- (df_long_loadings$TimeIndex - 1) * (24 / 288)
```

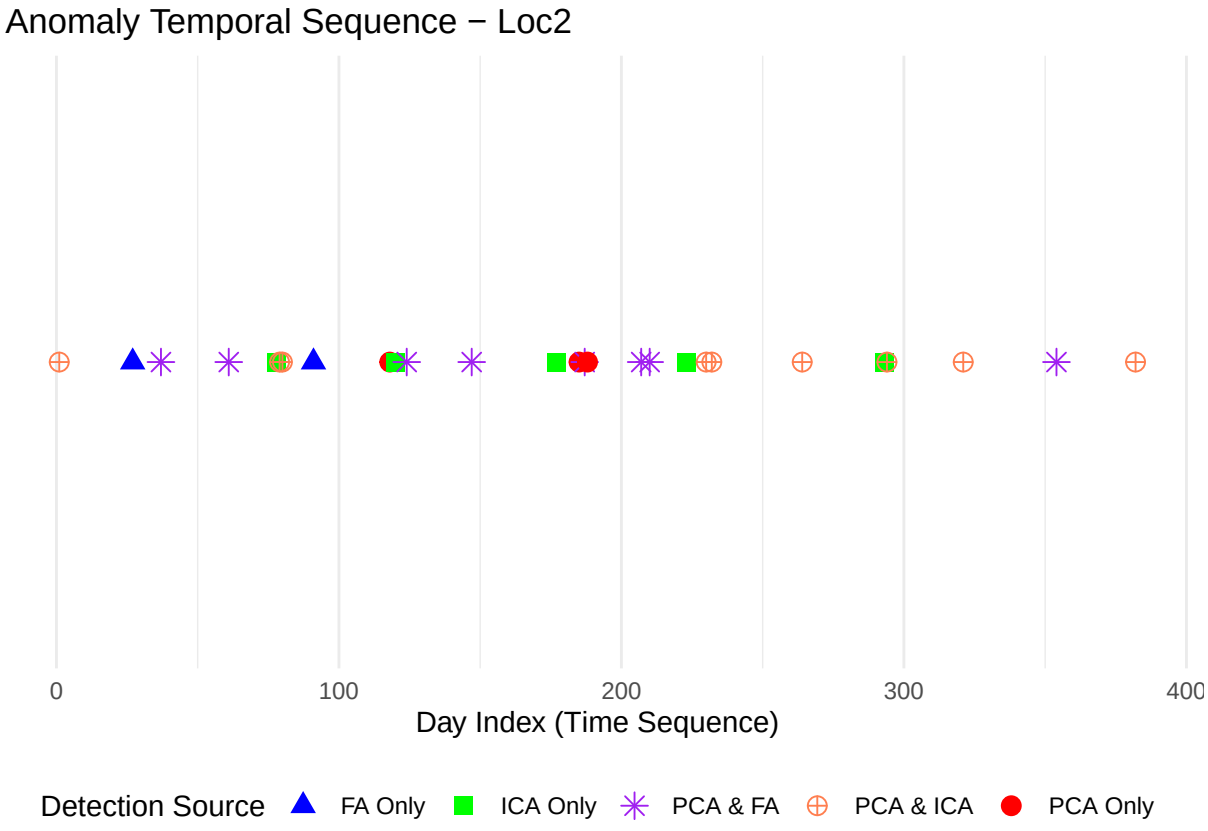
```
ggplot(df_long_loadings, aes(x = Hour, y = Component, fill = Loading)) +
  geom_tile(color = "white", linewidth = 0.5) +
  scale_fill_gradient2(low = "blue", mid = "white", high = "red", midpoint = 0,
                       name = "Loading Value") +
  labs(title = paste("PCA Component Loadings Heatmap -", loc_name),
       x = "Hour of Day",
       y = "Principal Component") +
  scale_x_continuous(breaks = seq(0, 24, 4)) +
  theme_minimal() +
  theme(legend.position = "bottom",
        axis.text.y = element_text(face = "bold"))
```

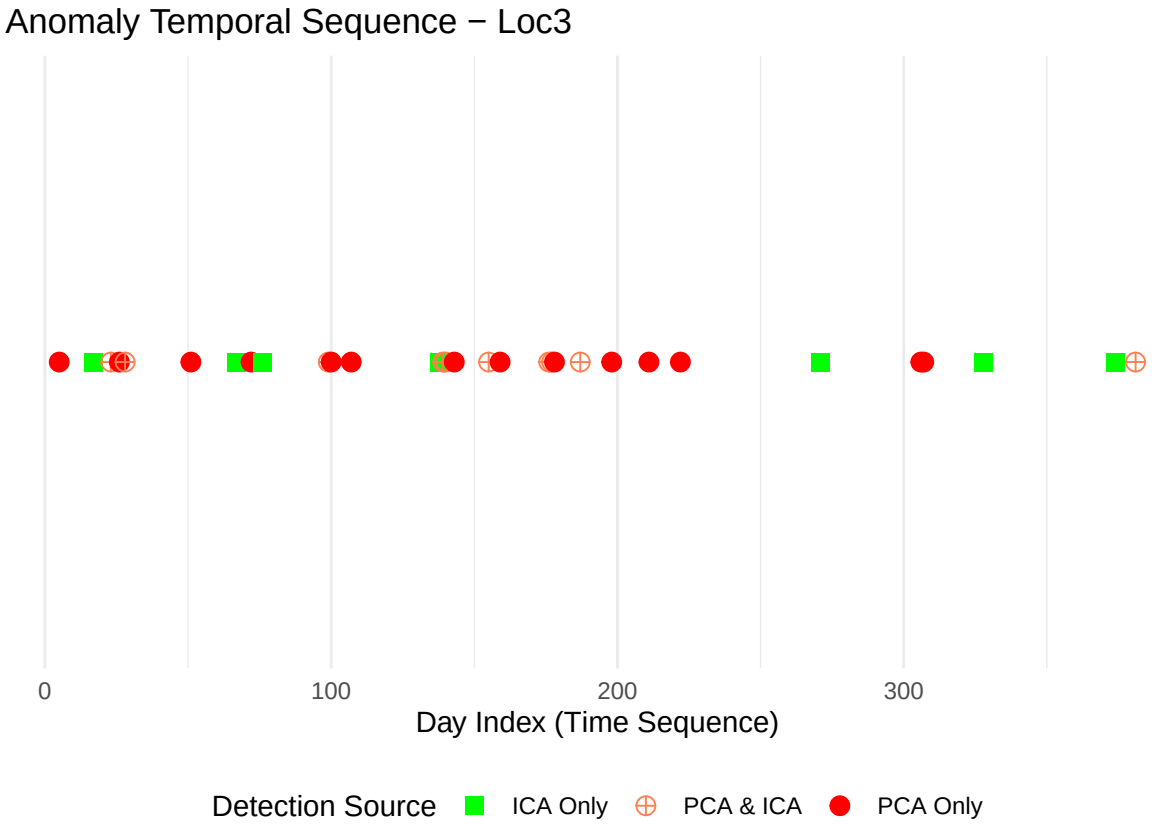
Anomaly timeline

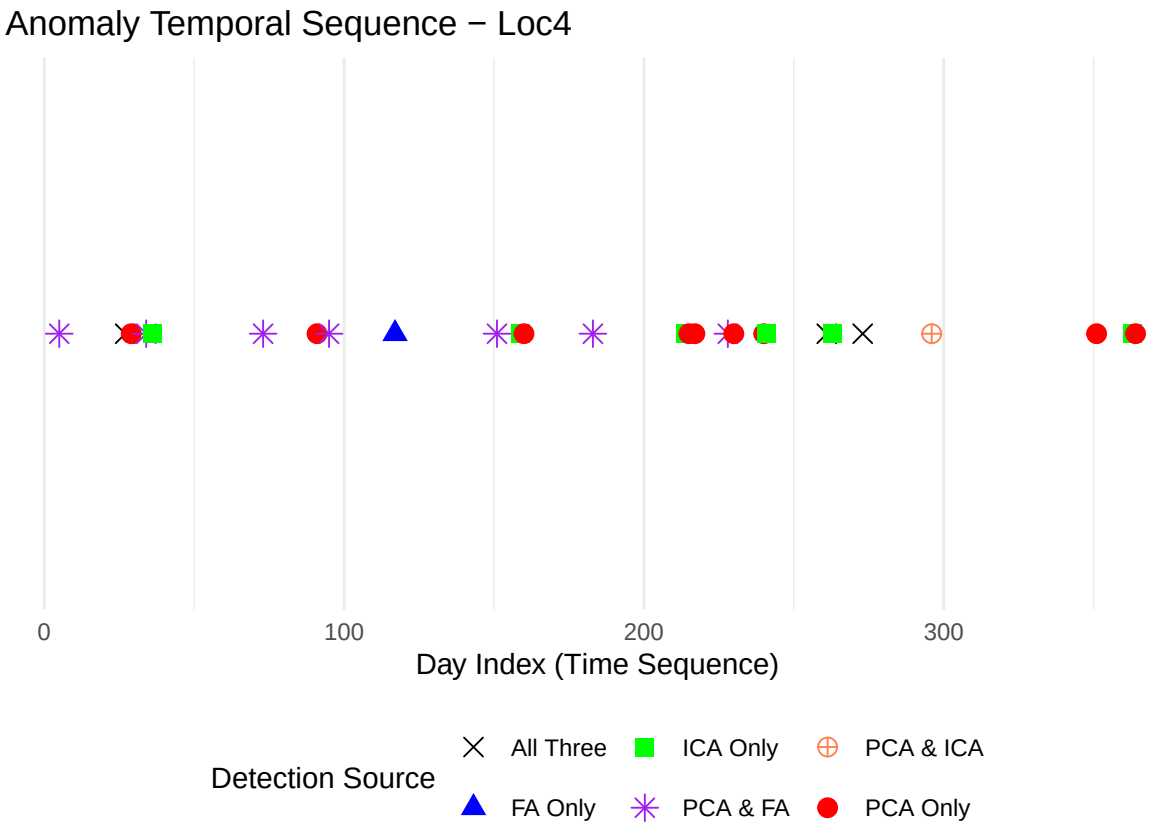
```
anomaly_timeline <- function(res, loc_name) {
  all_days <- rownames(res$pca_scores)
  df_timeline <- data.frame(DayName = all_days, DayIndex = 1:length(all_days))

  # Multi-method anomaly labeling
  df_timeline <- df_timeline %>%
    mutate(AnomalyType = case_when(
      DayName %in% intersect(intersect(res$pca_anom_days, res$fa_anom_days), res$ica_ano
      DayName %in% setdiff(intersect(res$pca_anom_days, res$fa_anom_days), res$ica_anom_
      DayName %in% setdiff(intersect(res$pca_anom_days, res$ica_anom_days), res$fa_anom_
      DayName %in% setdiff(intersect(res$fa_anom_days, res$ica_anom_days), res$pca_anom_
      DayName %in% setdiff(setdiff(res$pca_anom_days, res$fa_anom_days), res$ica_anom_da
      DayName %in% setdiff(setdiff(res$fa_anom_days, res$pca_anom_days), res$ica_anom_da
      DayName %in% setdiff(setdiff(res$ica_anom_days, res$pca_anom_days), res$fa_anom_da
      TRUE ~ "Normal"
    ))

  df_anomalies <- df_timeline %>% filter(AnomalyType != "Normal")

  if (nrow(df_anomalies) == 0) {
    message("No anomalies to plot for ", loc_name)
    return(NULL)
  }

  ggplot(df_timeline, aes(x = DayIndex, y = 1)) +
    # Normal timeline background
    # geom_segment(data = subset(df_timeline, AnomalyType == "Normal"),
    #              aes(x = DayIndex, xend = DayIndex, y = 0.95, yend = 1.05),
```

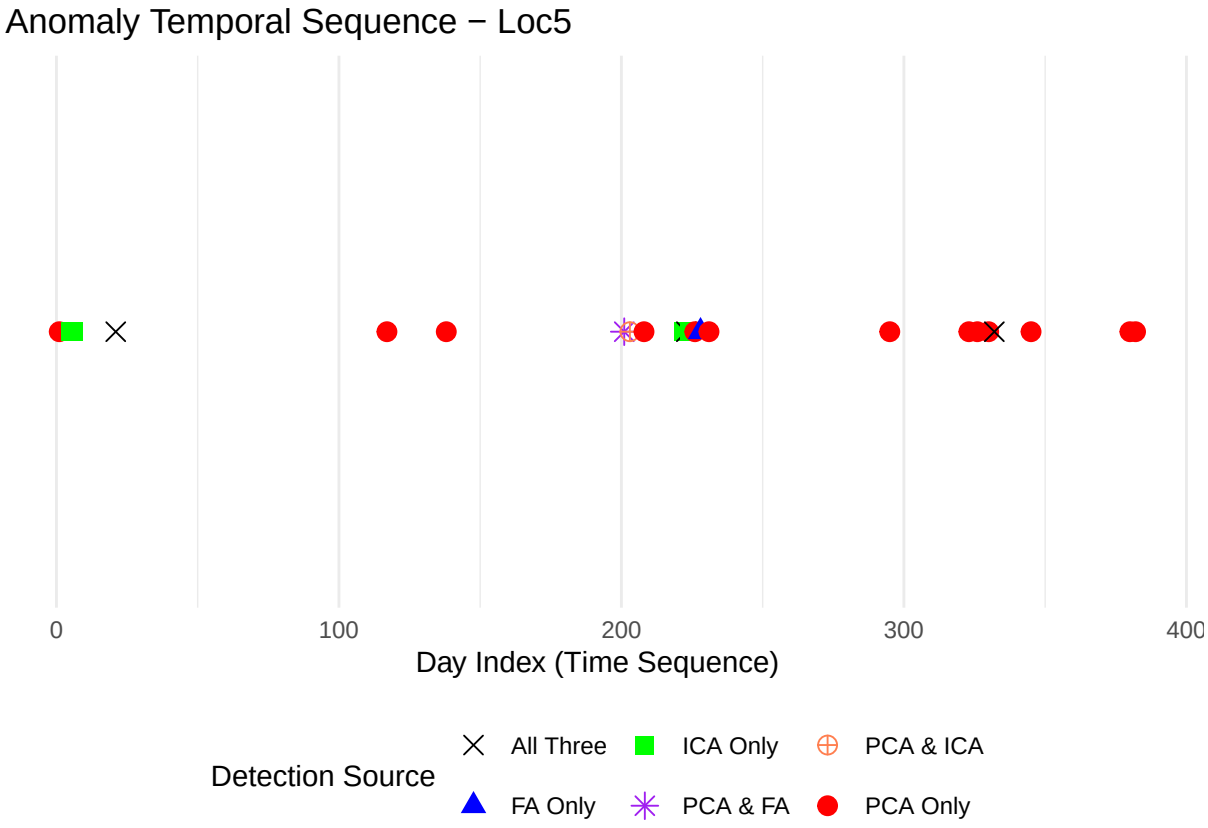
```
    #               color = "gray80", linewidth = 0.15) +

    # Highlight anomalies
    geom_point(data = df_anomalies,
               aes(color = AnomalyType, shape = AnomalyType),
               size = 3) +

    scale_color_manual(values = c(
      "PCA Only" = "red",
      "FA Only" = "blue",
      "ICA Only" = "green",
      "PCA & FA" = "purple",
      "PCA & ICA" = "#FF7F50",
      "FA & ICA" = "#00CED1",
      "All Three" = "black"
    )) +
    scale_shape_manual(values = c(
      "PCA Only" = 19,      # circle
      "FA Only" = 17,       # triangle
      "ICA Only" = 15,      # square
      "PCA & FA" = 8,       # star
      "PCA & ICA" = 10,     # diamond
      "FA & ICA" = 13,      # diamond plus
      "All Three" = 4       # X
    )) +

    labs(title = paste("Anomaly Temporal Sequence -", loc_name),
         x = "Day Index (Time Sequence)",
         y = "",
         color = "Detection Source",
         shape = "Detection Source") +

    theme_minimal() +
    theme(axis.text.y = element_blank(),
          axis.ticks.y = element_blank(),
          panel.grid.major.y = element_blank(),
          panel.grid.minor.y = element_blank(),
          legend.position = "bottom")
}

# Loop over locations to generate the plots
for (loc_name in names(location_results)) {
  print(anomaly_timeline(location_results[[loc_name]], loc_name))
}
```

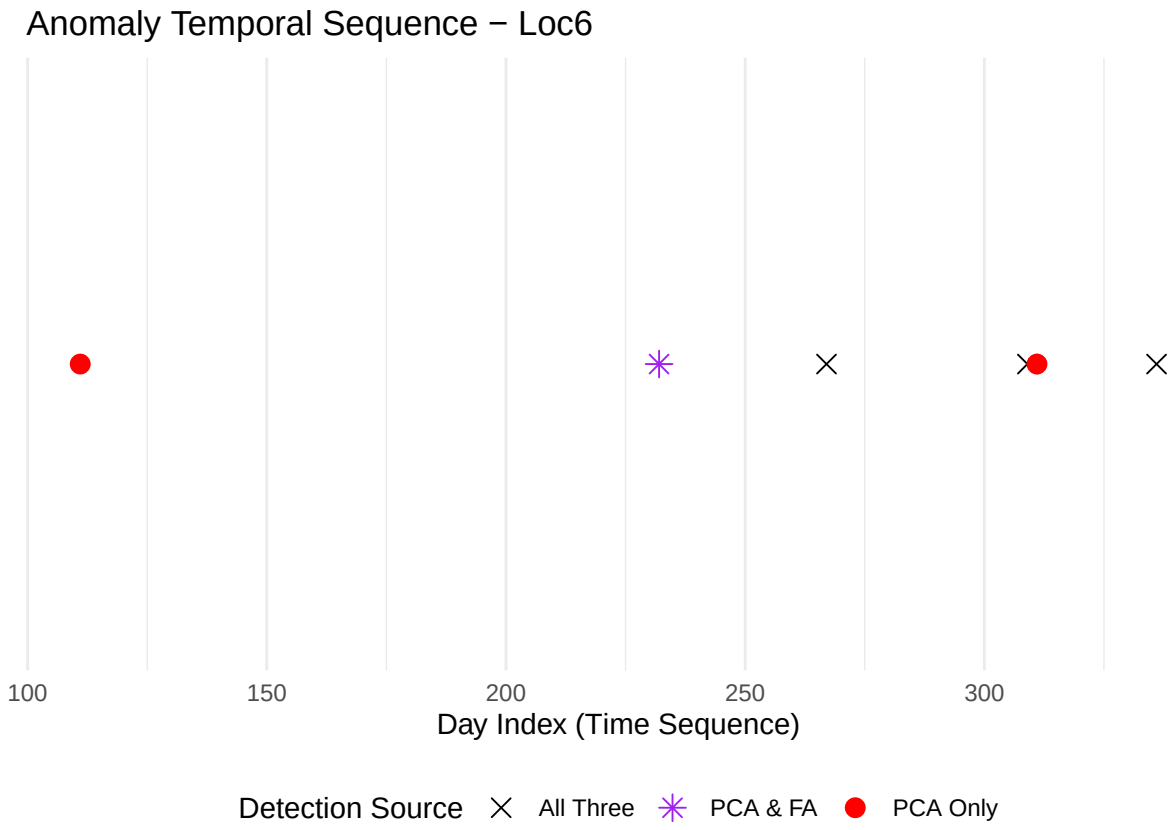

Anomaly Temporal Sequence − Loc1
0
100
200
300
400
Day Index (Time Sequence)
Detection Source
ICA Only
PCA & ICA
PCA Only

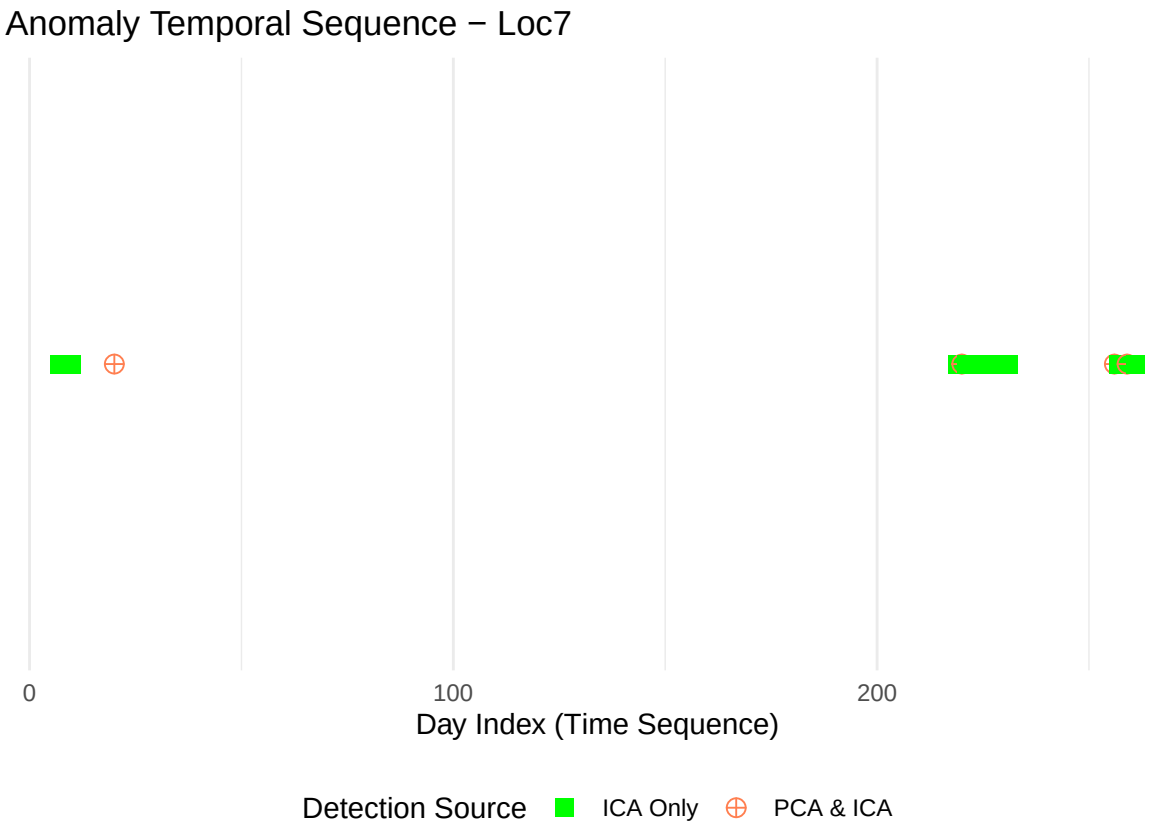

Anomaly Temporal Sequence − Loc2
0
100
200
300
400
Day Index (Time Sequence)
Detection Source
FA Only
ICA Only
PCA & FA
PCA & ICA
PCA Only

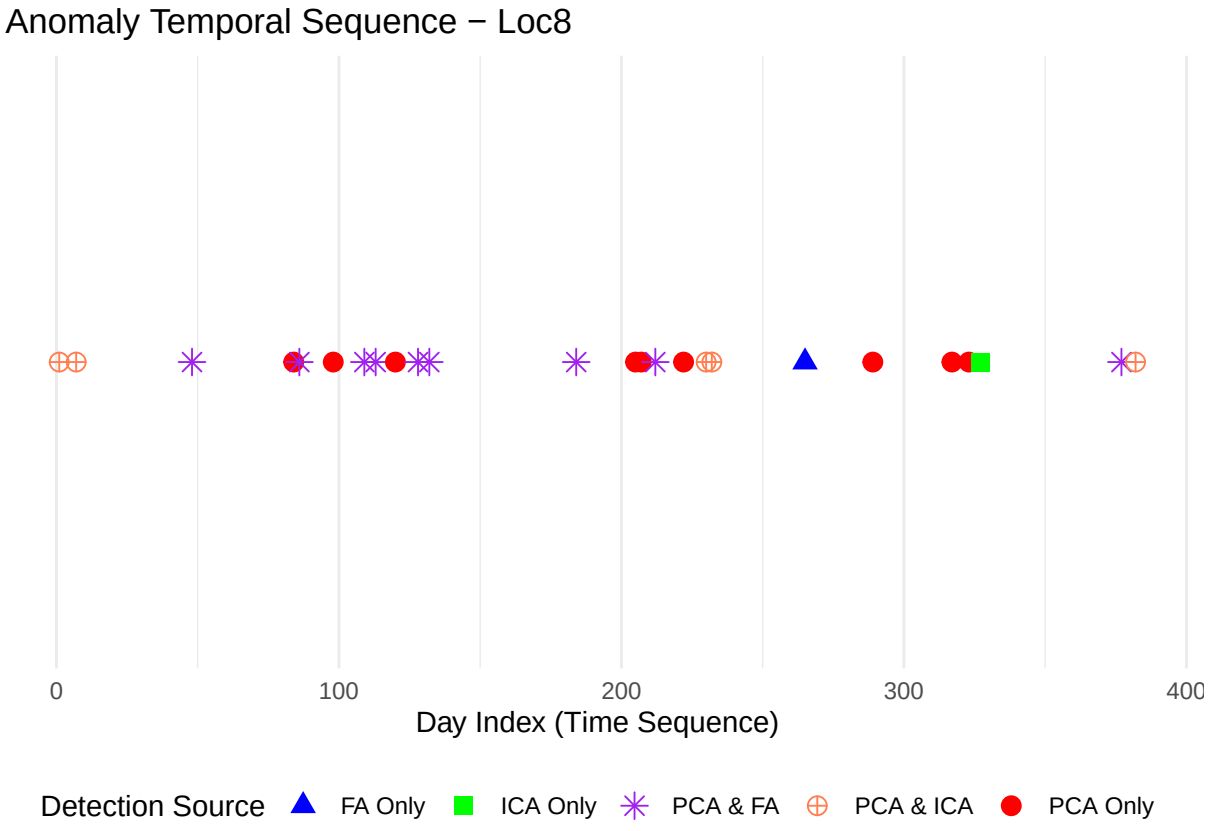

Anomaly Temporal Sequence − Loc3
0
100
200
300
Day Index (Time Sequence)
Detection Source
ICA Only
PCA & ICA
PCA Only

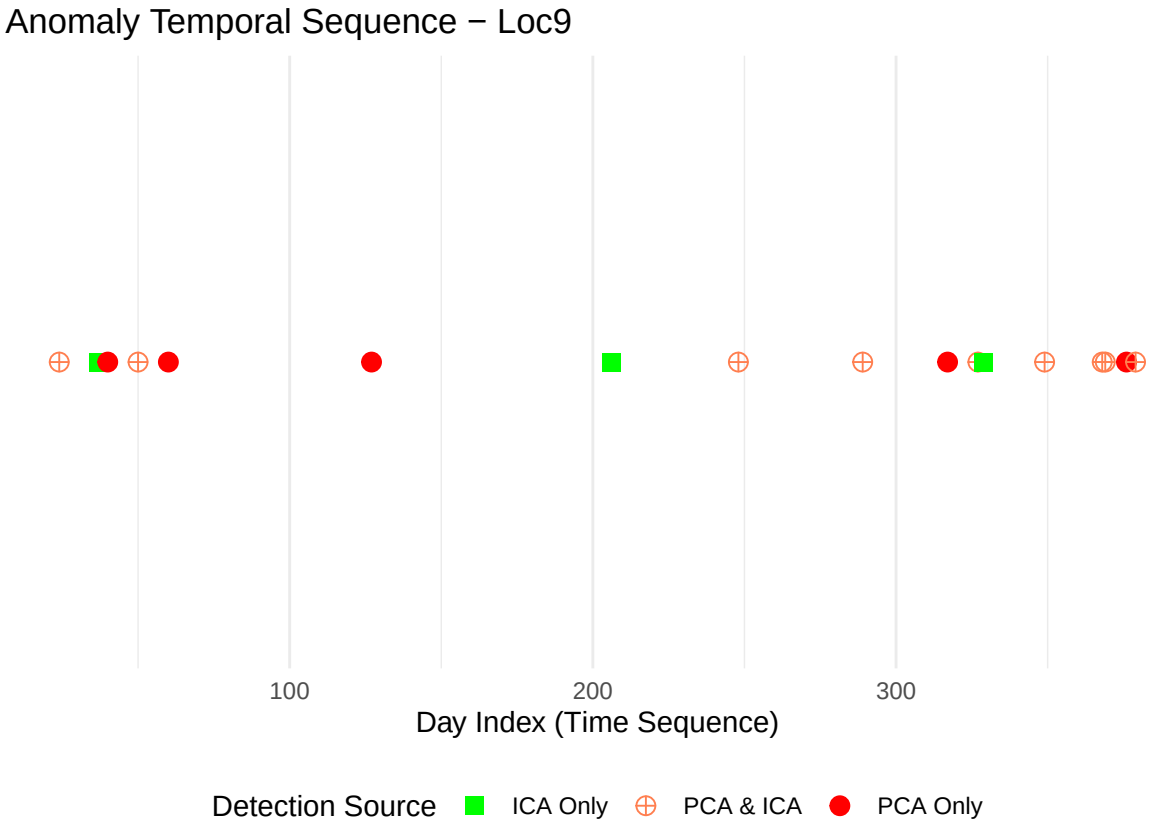

Anomaly Temporal Sequence − Loc4
0
100
200
300
Day Index (Time Sequence)
Detection Source
All Three
ICA Only
PCA & ICA
FA Only
PCA & FA
PCA Only

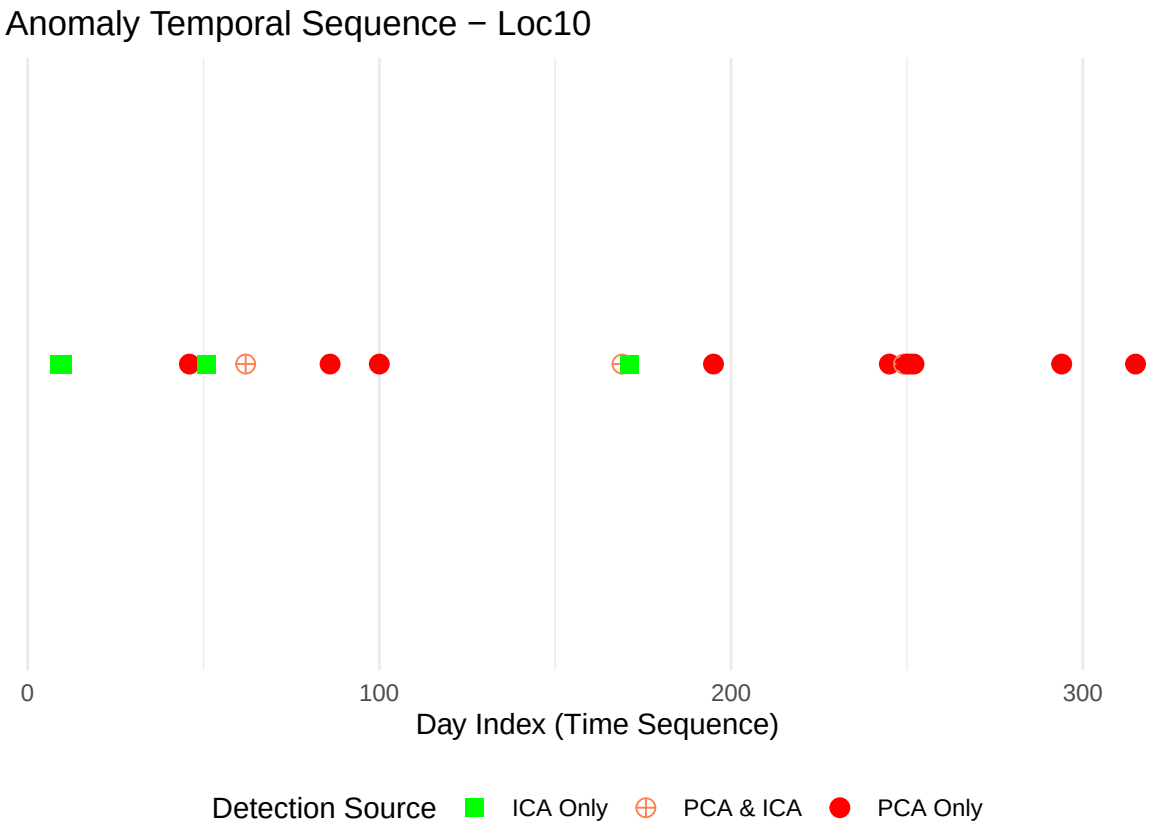

Anomaly Temporal Sequence − Loc5
0
100
200
300
400
Day Index (Time Sequence)
Detection Source
All Three
ICA Only
PCA & ICA
FA Only
PCA & FA
PCA Only

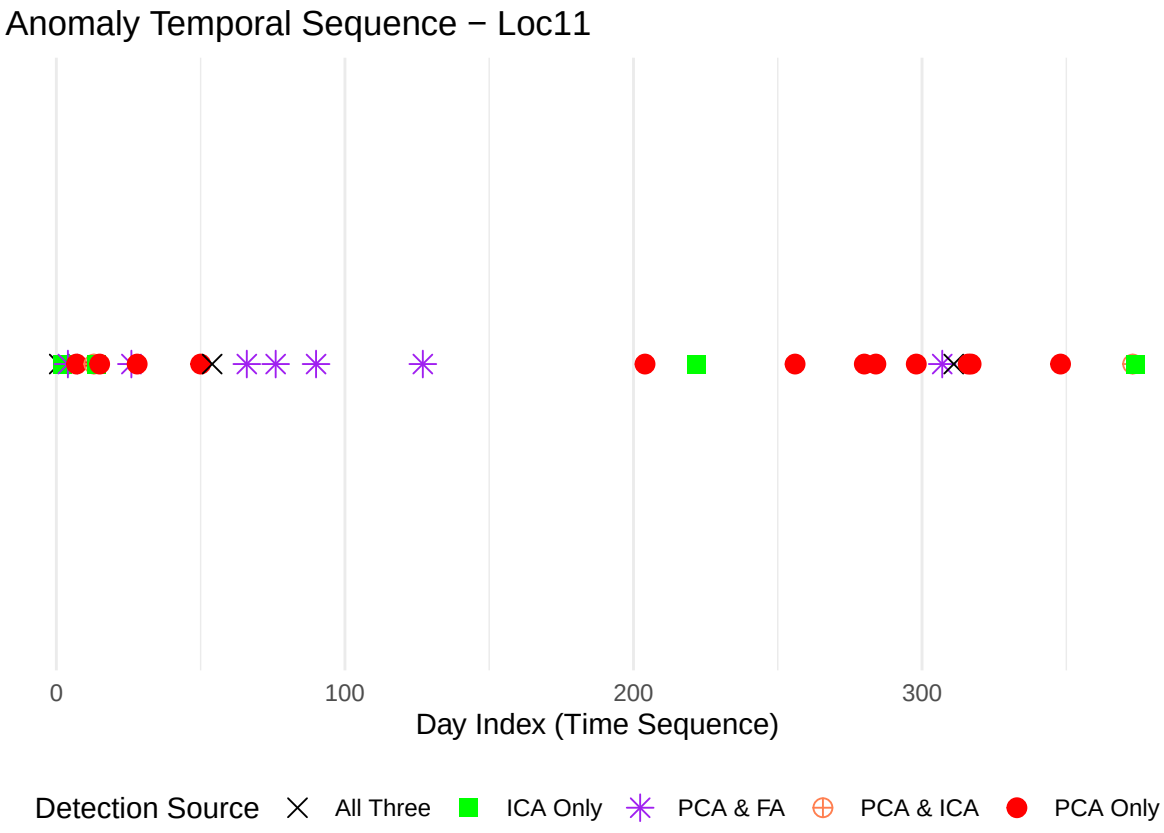

Anomaly Temporal Sequence − Loc6
100
150
200
250
300
Day Index (Time Sequence)
Detection Source
All Three
PCA & FA
PCA Only

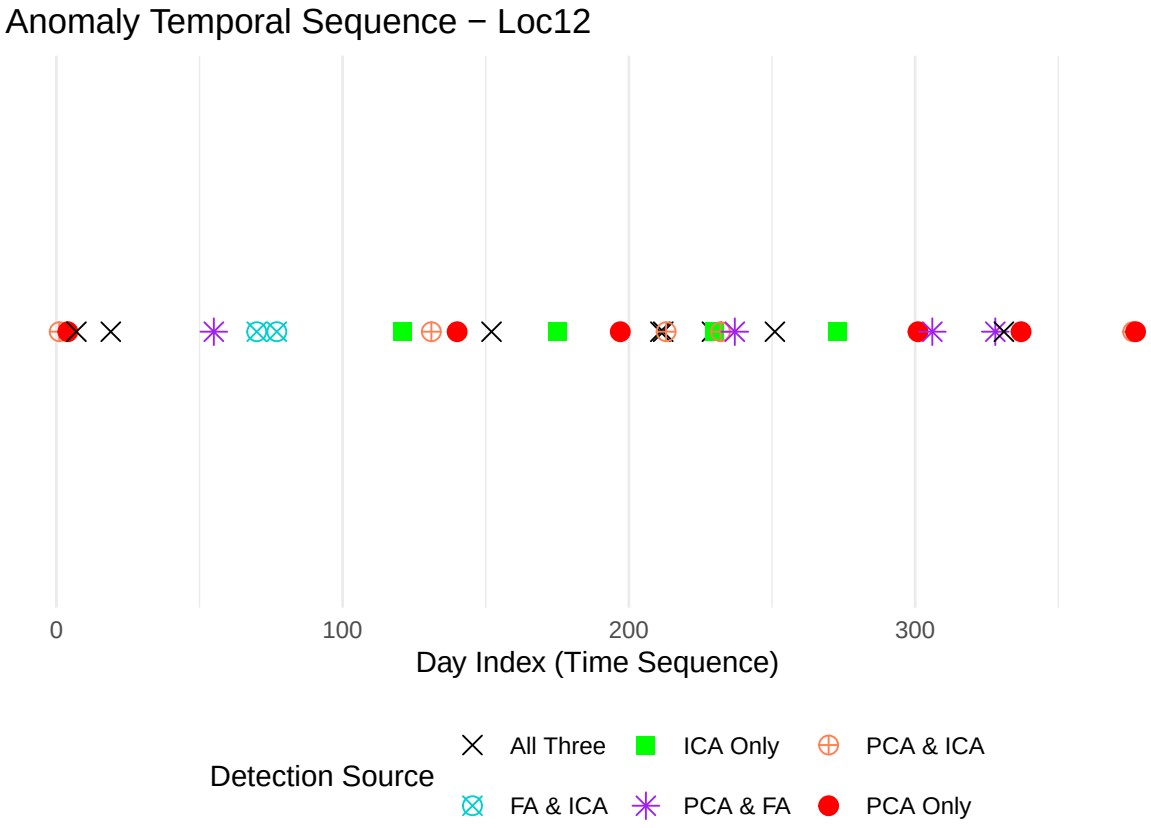

Anomaly Temporal Sequence − Loc7
0
100
200
Day Index (Time Sequence)
Detection Source
ICA Only
PCA & ICA

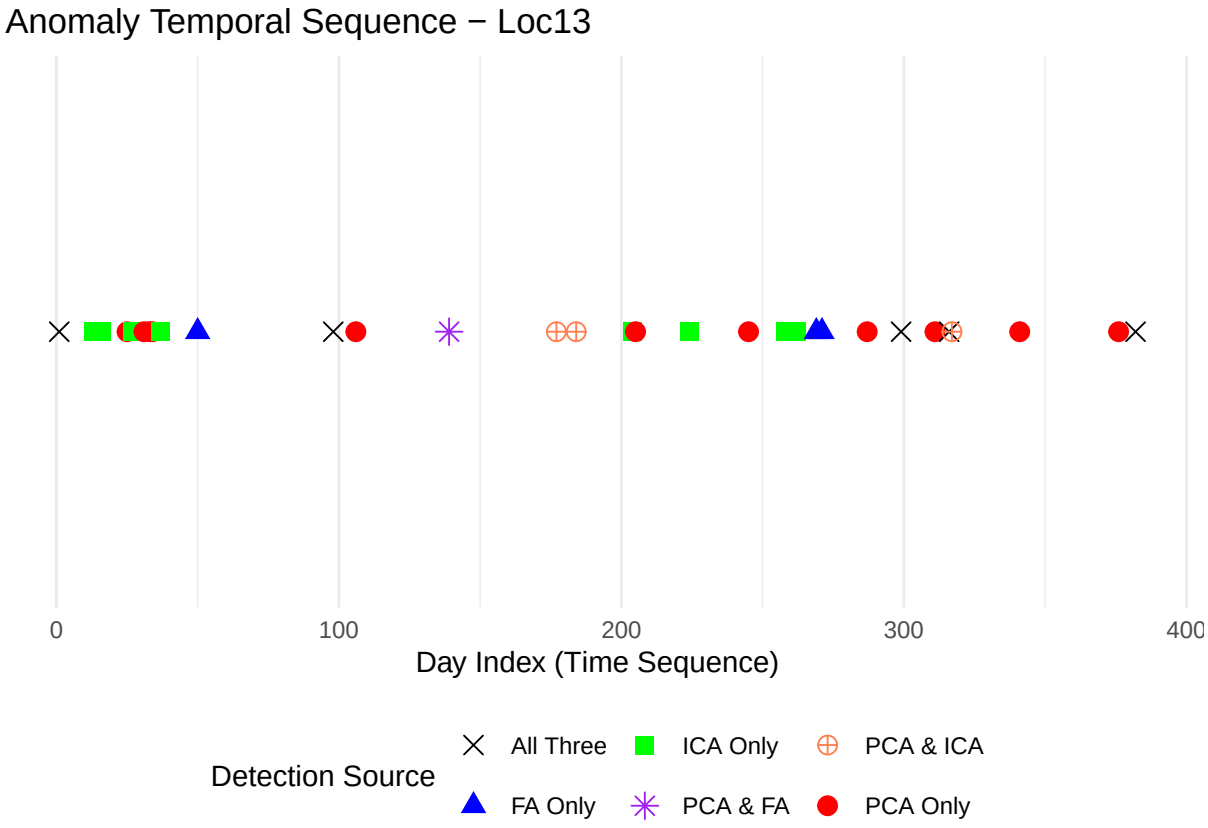

Anomaly Temporal Sequence − Loc8
0
100
200
300
400
Day Index (Time Sequence)
Detection Source
FA Only
ICA Only
PCA & FA
PCA & ICA
PCA Only

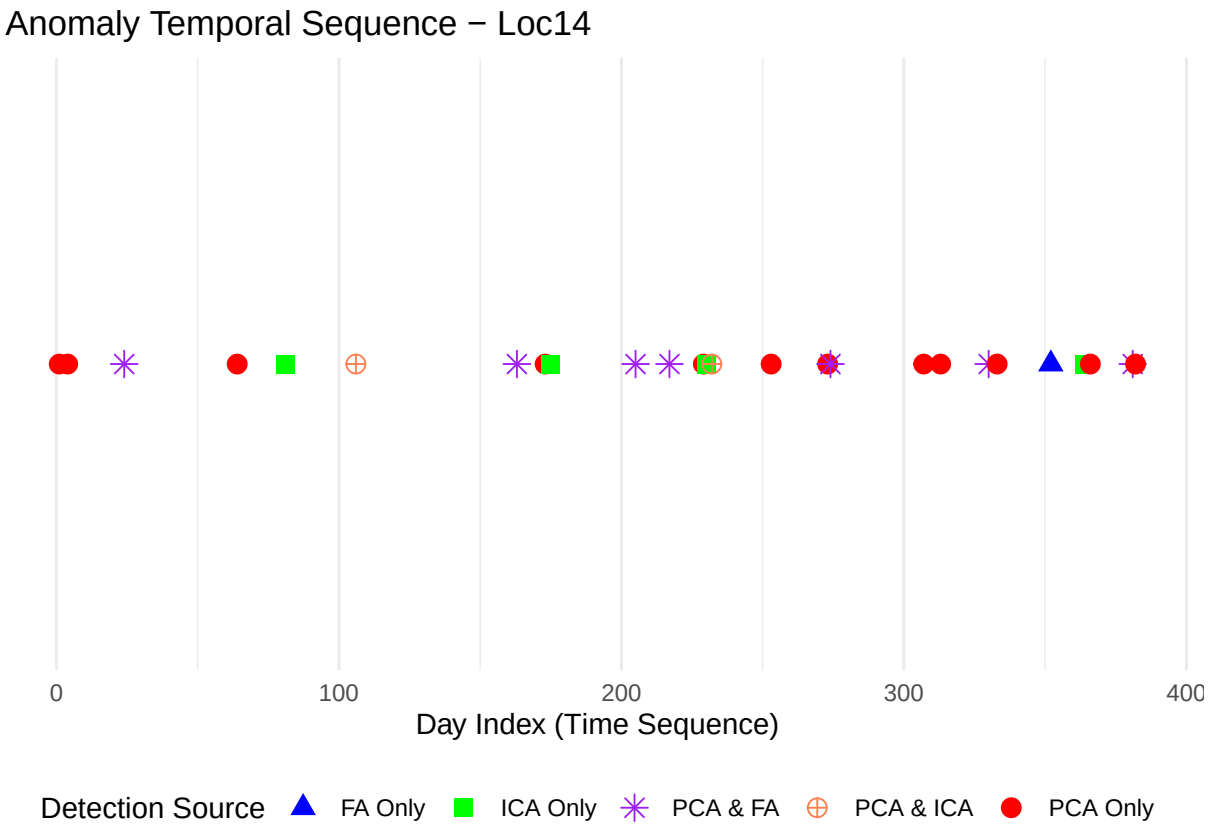

Anomaly Temporal Sequence − Loc9
100
200
300
Day Index (Time Sequence)
Detection Source
ICA Only
PCA & ICA
PCA Only

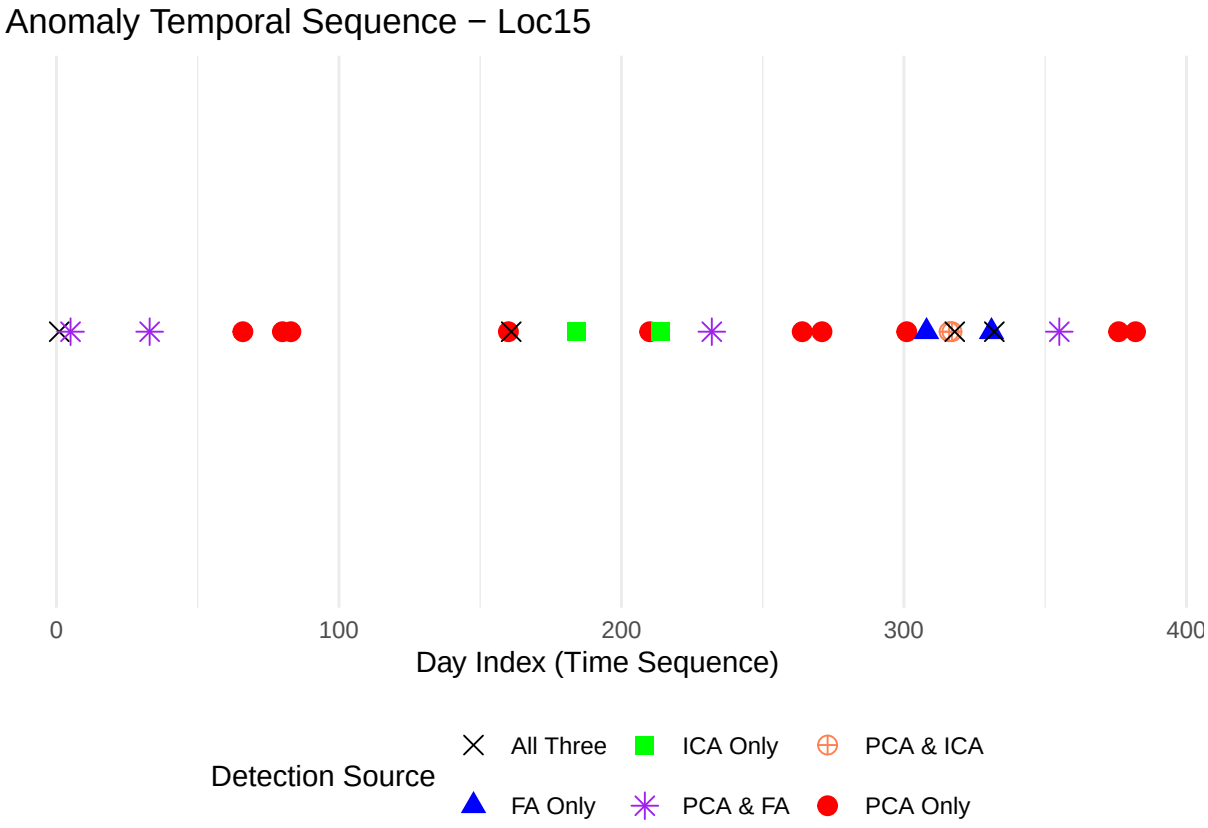

Anomaly Temporal Sequence − Loc10
0
100
200
300
Day Index (Time Sequence)
Detection Source
ICA Only
PCA & ICA
PCA Only

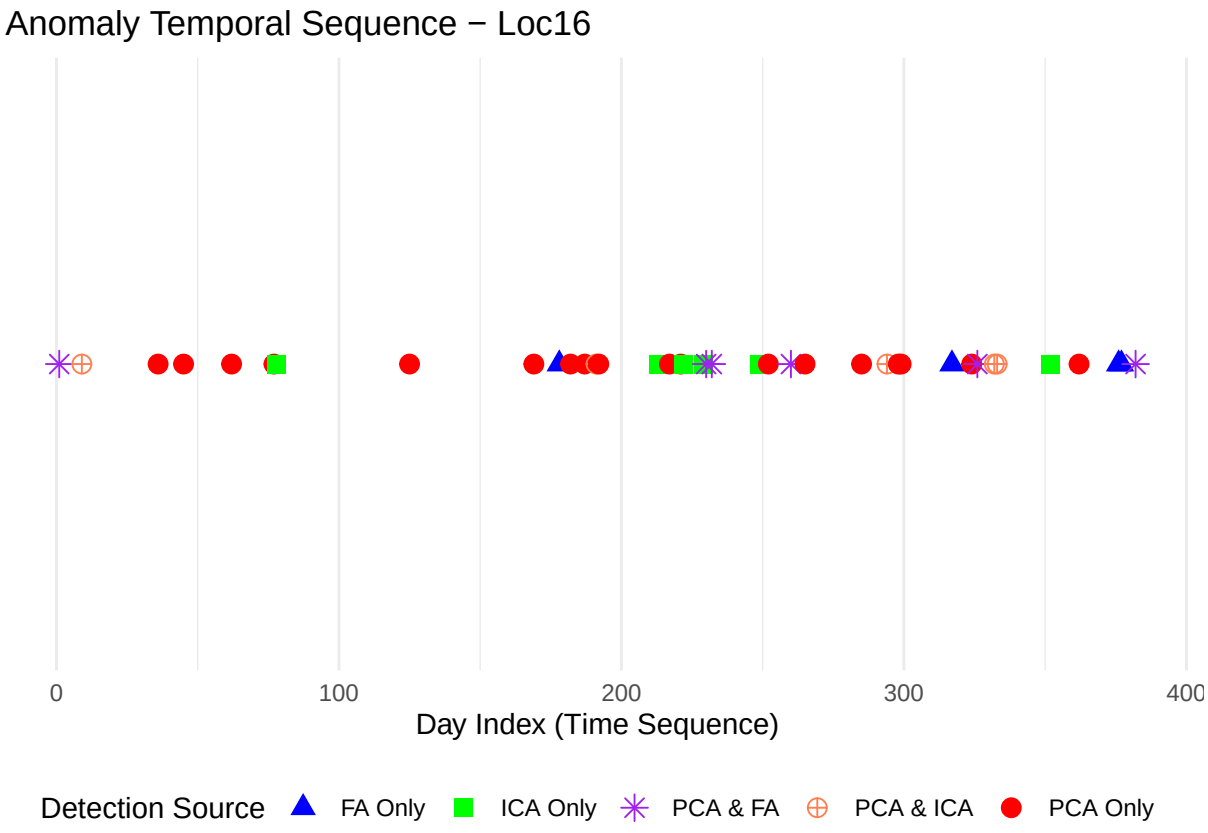

Anomaly Temporal Sequence − Loc11
0
100
200
300
Day Index (Time Sequence)
Detection Source
All Three
ICA Only
PCA & FA
PCA & ICA
PCA Only

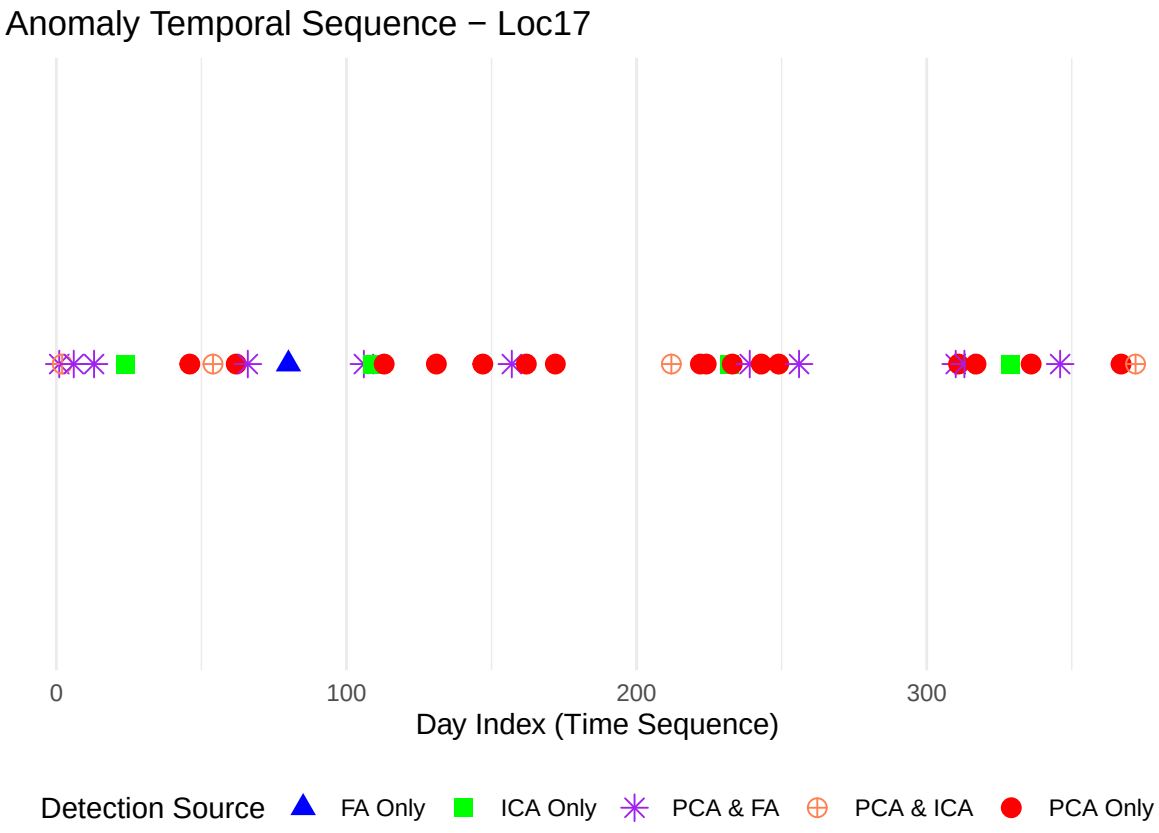

Anomaly Temporal Sequence − Loc12
0
100
200
300
Day Index (Time Sequence)
Detection Source
All Three
ICA Only
PCA & ICA
FA & ICA
PCA & FA
PCA Only

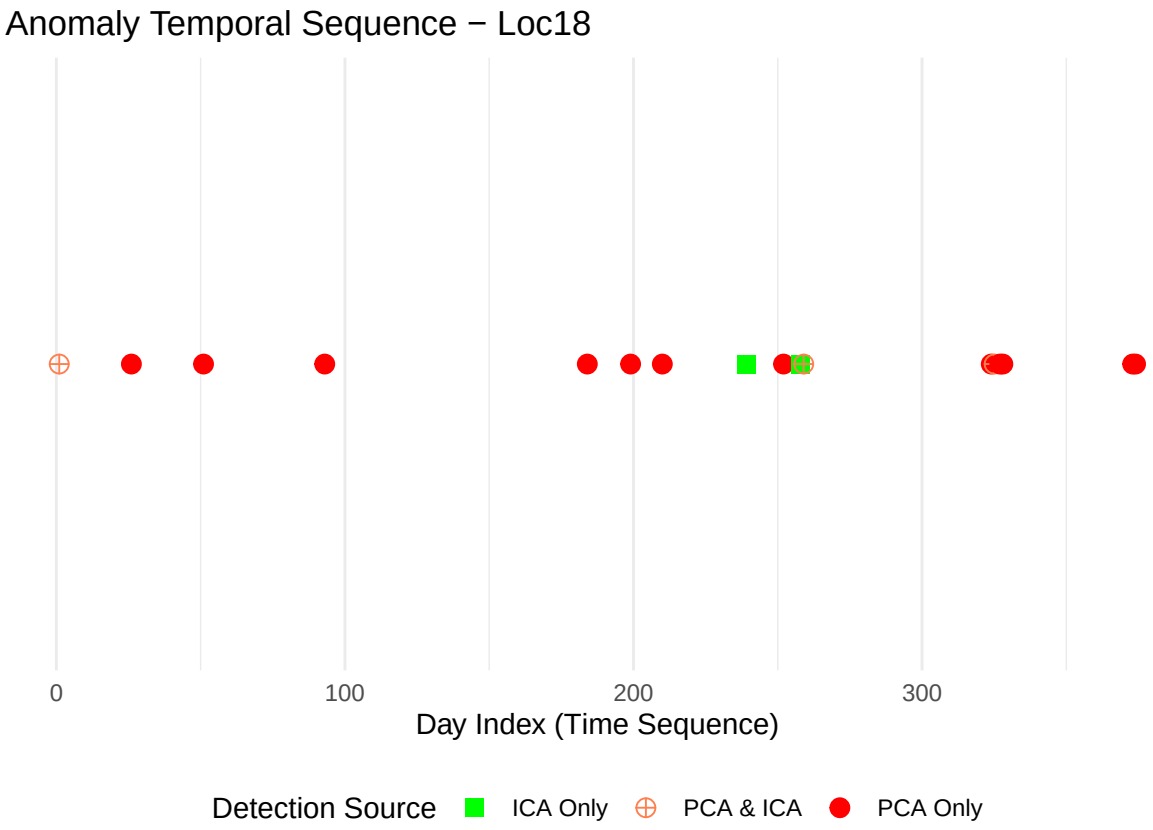

Anomaly Temporal Sequence − Loc13
0
100
200
300
400
Day Index (Time Sequence)
Detection Source
All Three
ICA Only
PCA & ICA
FA Only
PCA & FA
PCA Only

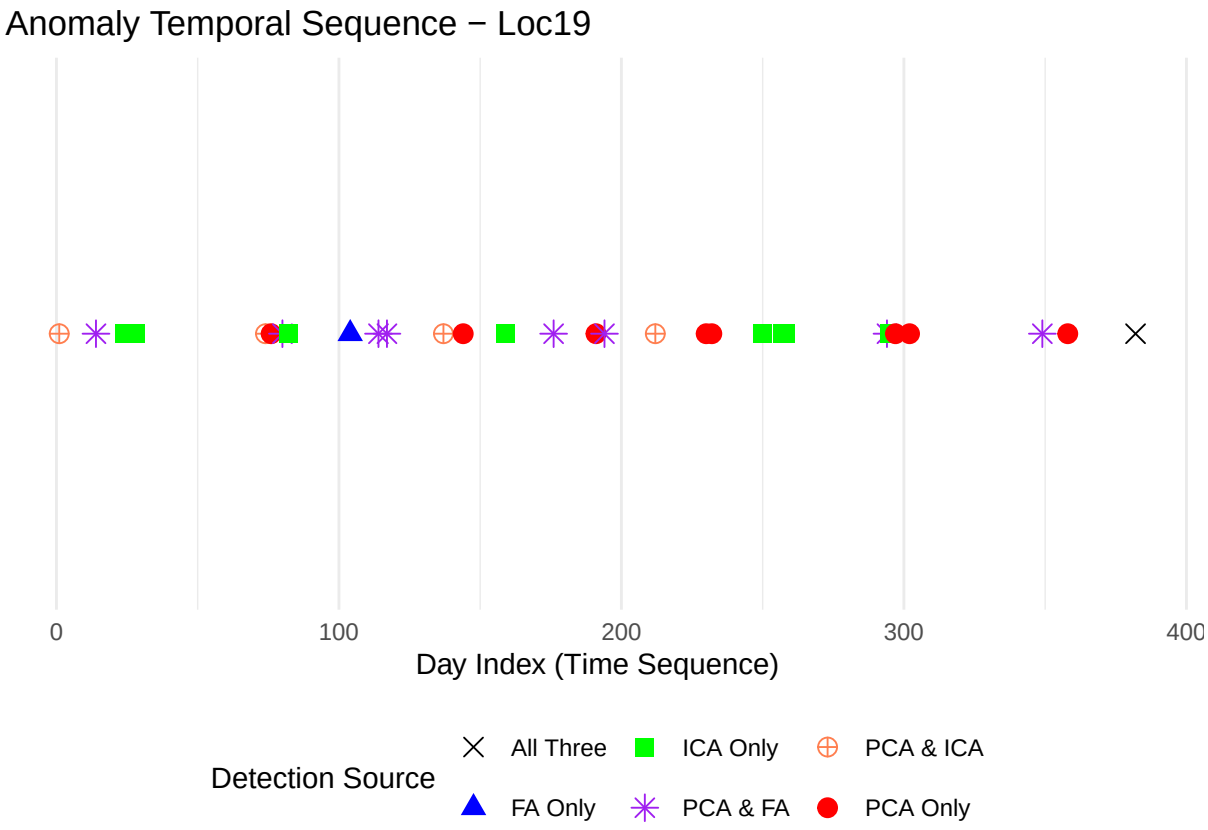

Anomaly Temporal Sequence − Loc14
0
100
200
300
400
Day Index (Time Sequence)
Detection Source
FA Only
ICA Only
PCA & FA
PCA & ICA
PCA Only

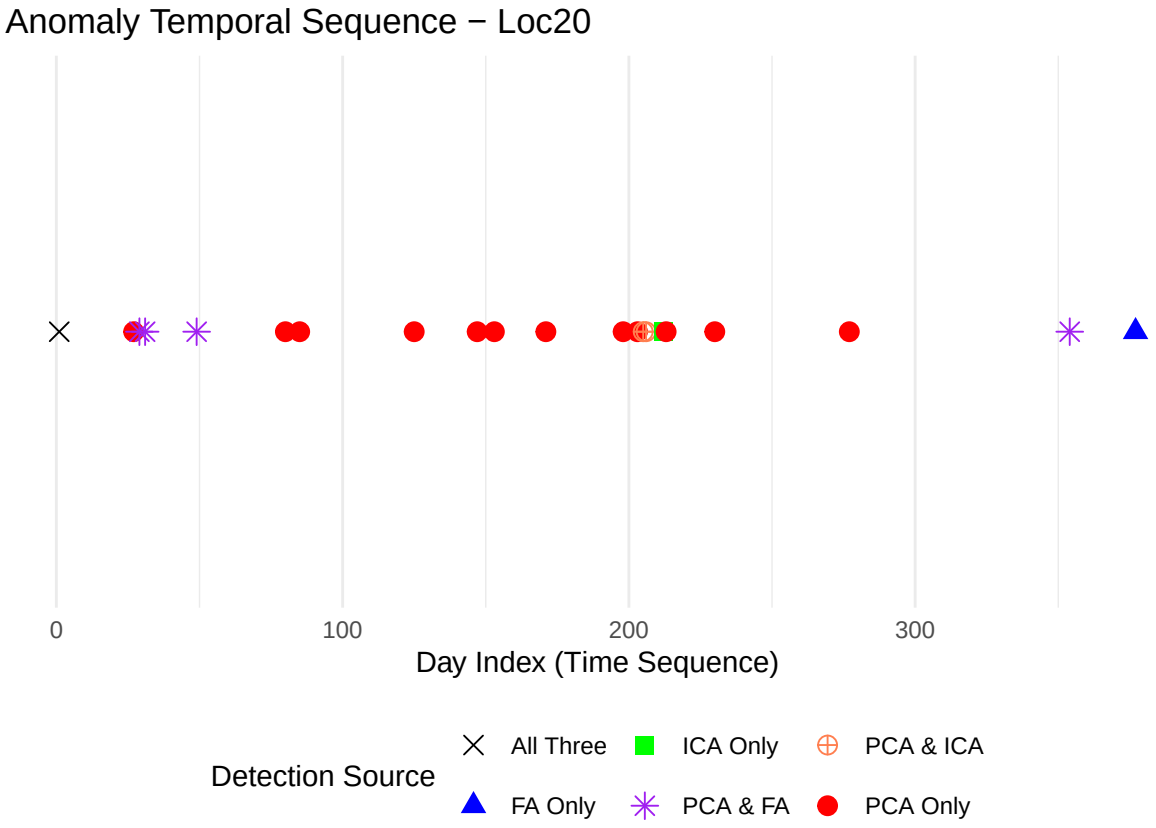

Anomaly Temporal Sequence − Loc15
0
100
200
300
400
Day Index (Time Sequence)
Detection Source
All Three
ICA Only
PCA & ICA
FA Only
PCA & FA
PCA Only

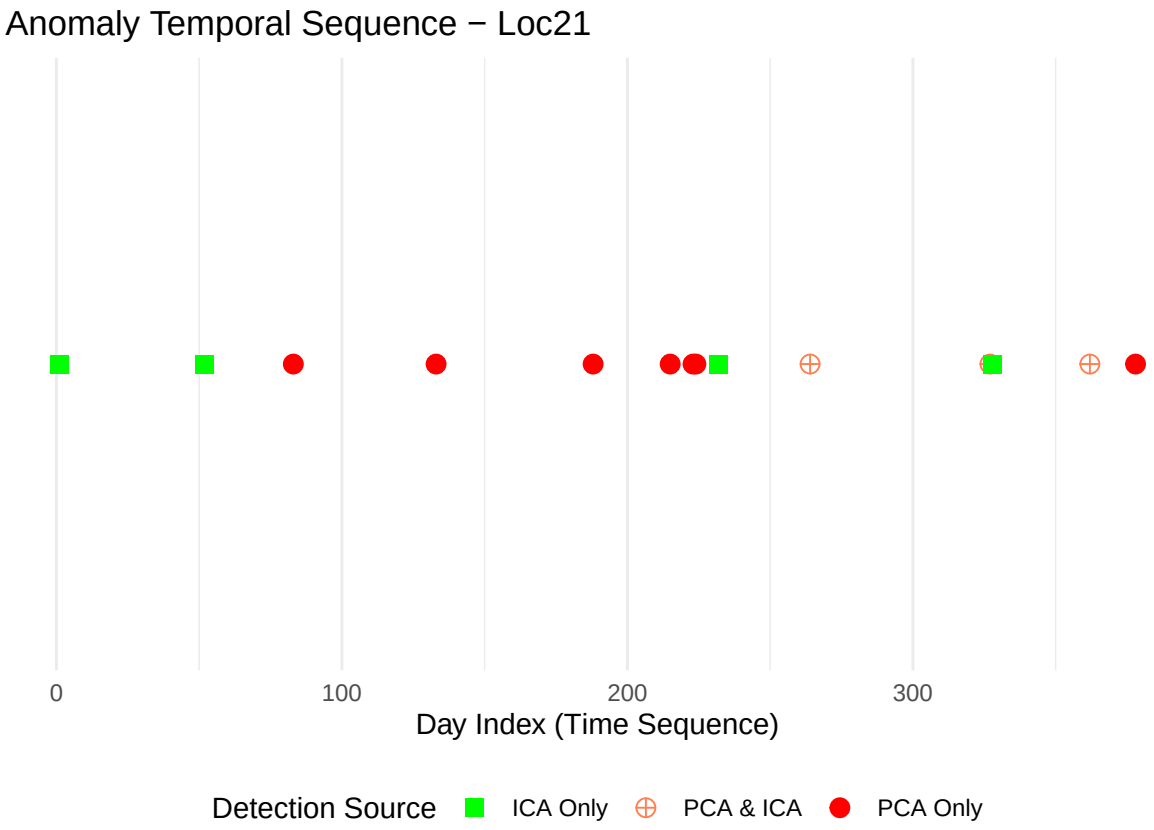

Anomaly Temporal Sequence − Loc16
0
100
200
300
400
Day Index (Time Sequence)
Detection Source
FA Only
ICA Only
PCA & FA
PCA & ICA
PCA Only

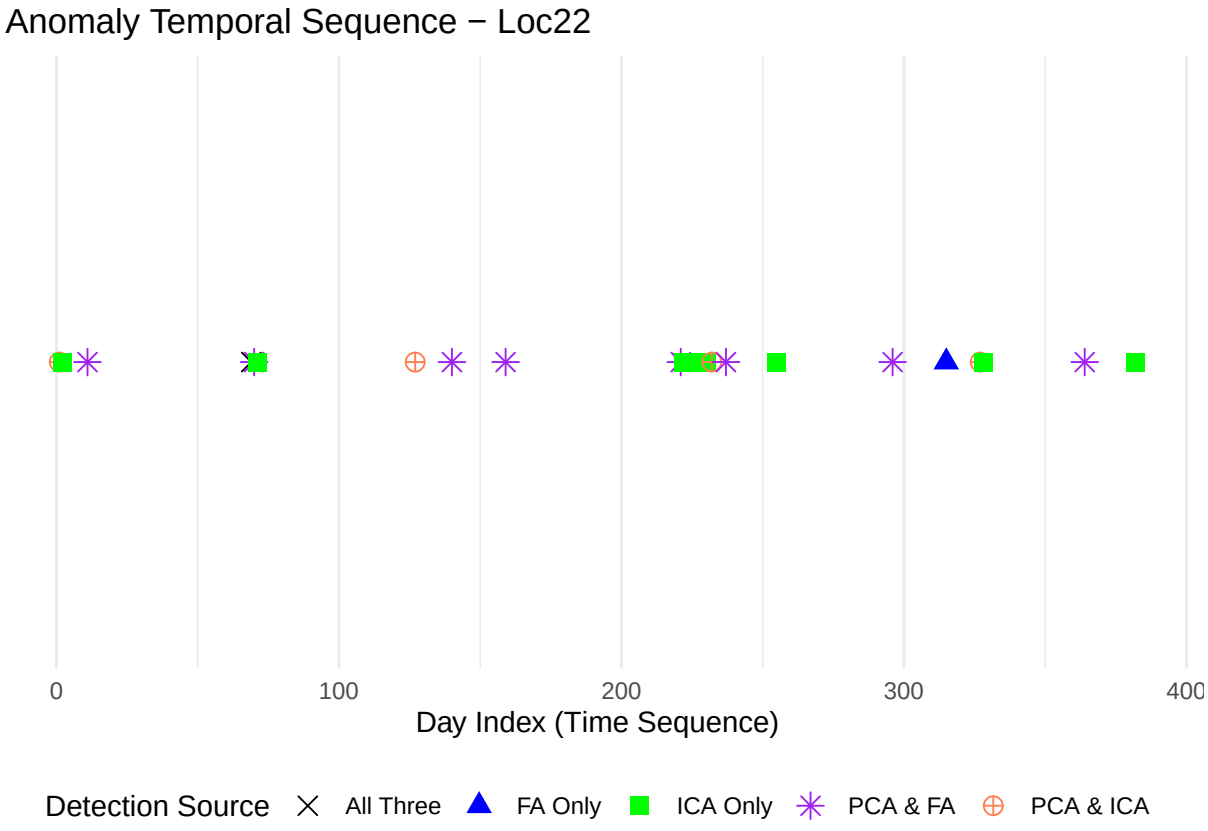

Anomaly Temporal Sequence − Loc17
0
100
200
300
Day Index (Time Sequence)
Detection Source
FA Only
ICA Only
PCA & FA
PCA & ICA
PCA Only

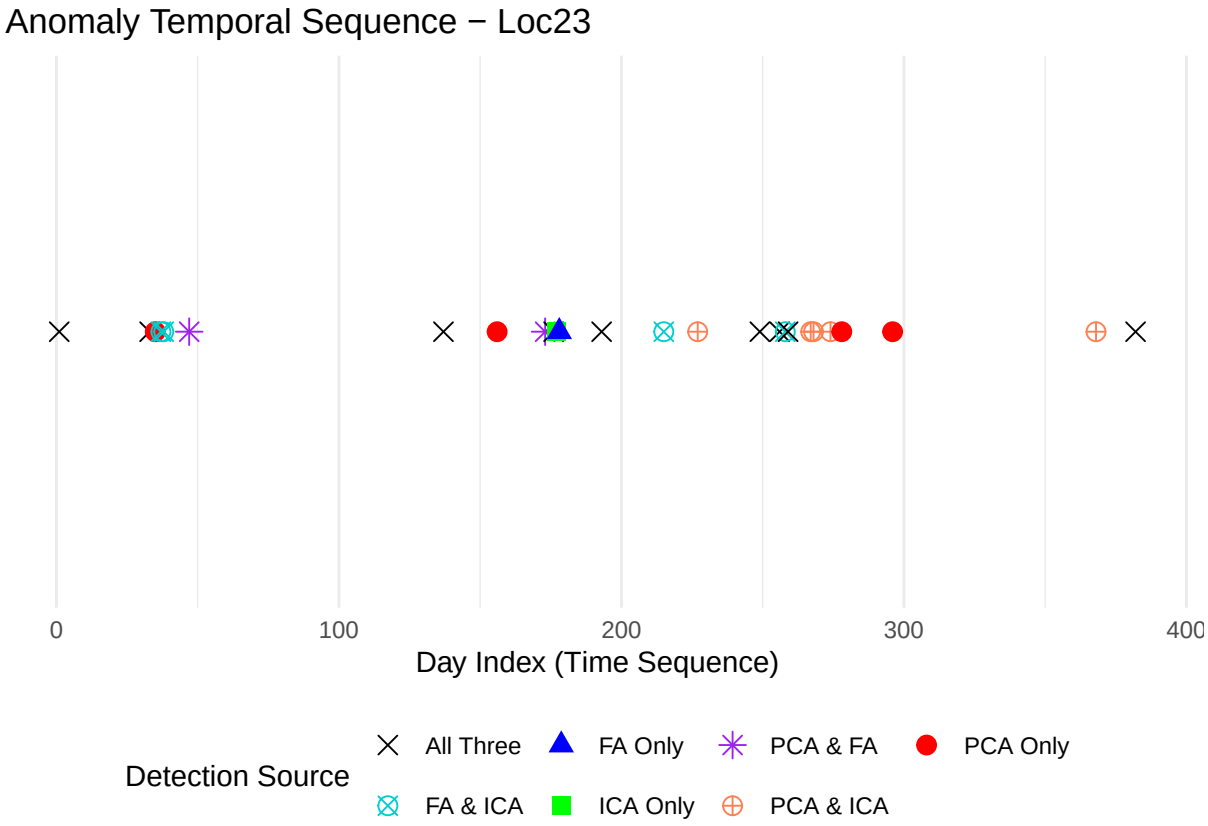

Anomaly Temporal Sequence − Loc18
0
100
200
300
Day Index (Time Sequence)
Detection Source
ICA Only
PCA & ICA
PCA Only

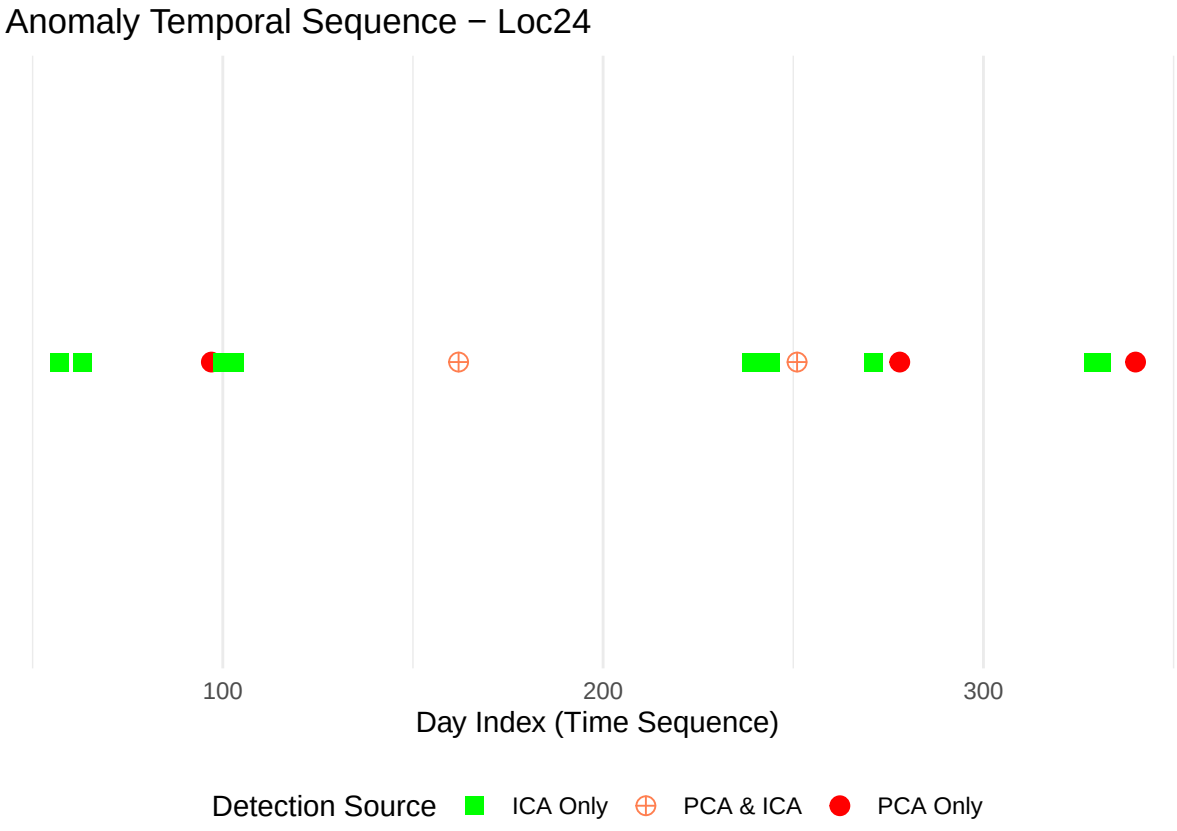

Anomaly Temporal Sequence − Loc19
0
100
200
300
400
Day Index (Time Sequence)
Detection Source
All Three
ICA Only
PCA & ICA
FA Only
PCA & FA
PCA Only

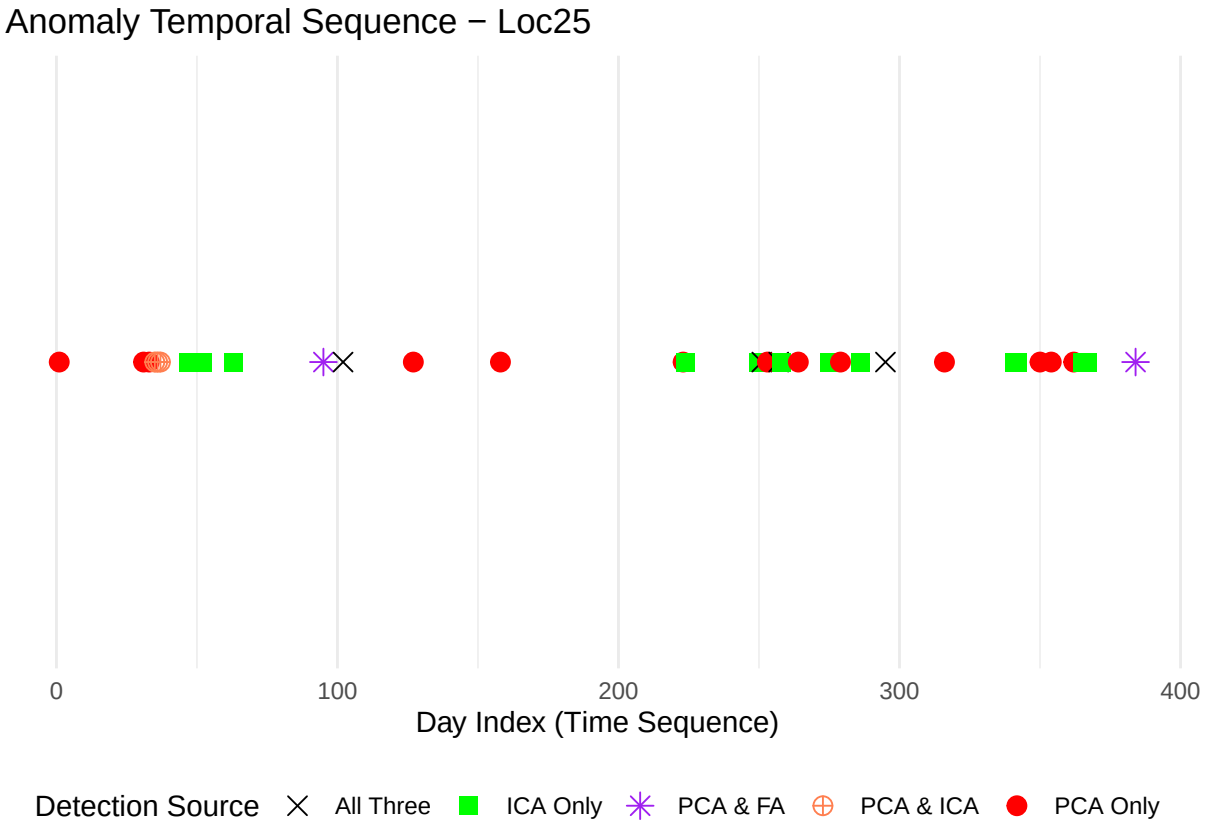

Anomaly Temporal Sequence − Loc20
0
100
200
300
Day Index (Time Sequence)
Detection Source
All Three
ICA Only
PCA & ICA
FA Only
PCA & FA
PCA Only

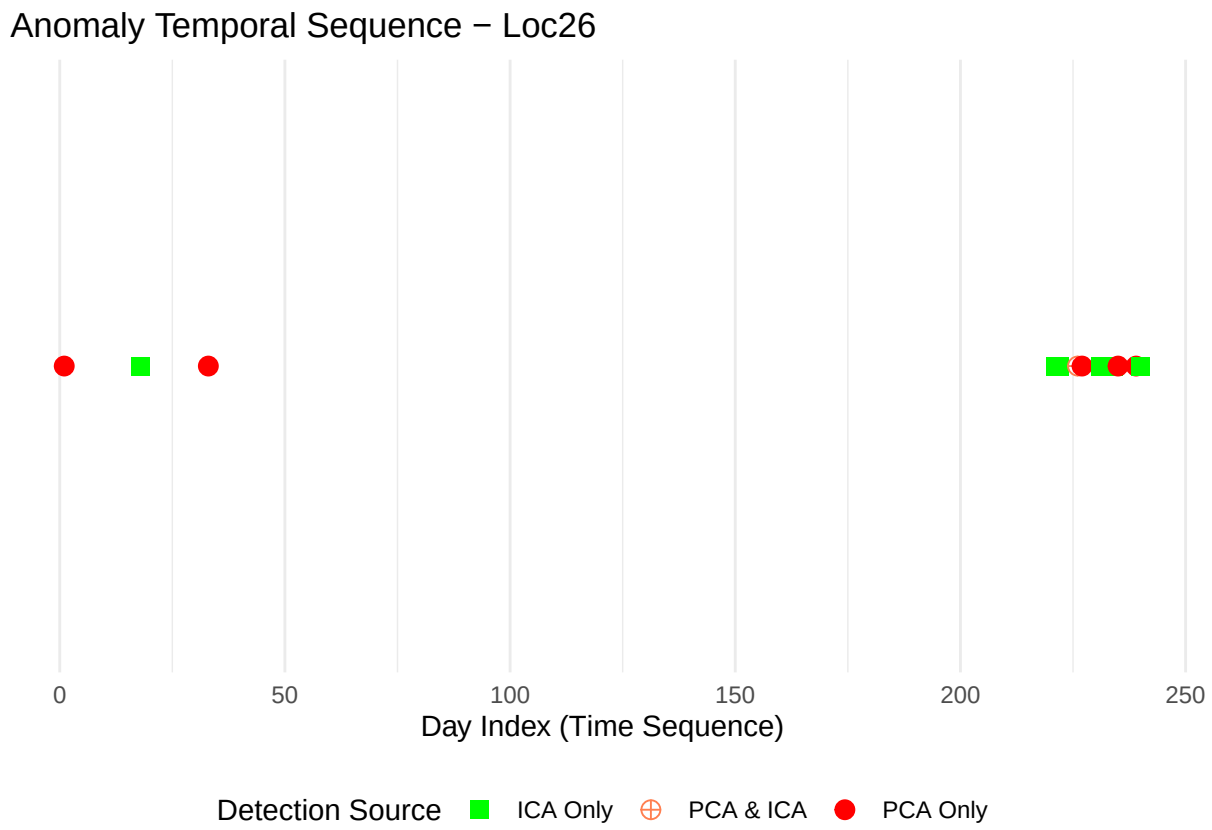

Anomaly Temporal Sequence − Loc21
0
100
200
300
Day Index (Time Sequence)
Detection Source
ICA Only
PCA & ICA
PCA Only


Anomaly Temporal Sequence − Loc22
0
100
200
300
400
Day Index (Time Sequence)
Detection Source
All Three
FA Only
ICA Only
PCA & FA
PCA & ICA


Anomaly Temporal Sequence − Loc23
0
100
200
300
400
Day Index (Time Sequence)
Detection Source
All Three
FA Only
PCA & FA
PCA Only
FA & ICA
ICA Only
PCA & ICA


Anomaly Temporal Sequence − Loc24
100
200
300
Day Index (Time Sequence)
Detection Source
ICA Only
PCA & ICA
PCA Only


Anomaly Temporal Sequence − Loc25
0
100
200
300
400
Day Index (Time Sequence)
Detection Source
All Three
ICA Only
PCA & FA
PCA & ICA
PCA Only



ICA boxplots

```r
## ICA Source Score Boxplot with Anomalies
plot_ica_score_boxplot <- function(res, loc_name, anomaly_coef = 1.5) {
  # Use res$ica_scores NOT res$ica$S!
  if (is.null(res$ica_scores) || nrow(res$ica_scores) == 0) {
    message("ICA scores not available or empty for ", loc_name)
    return(NULL)
  }

  scores <- res$ica_scores  # ← CHANGED FROM res$ica$S
  days <- rownames(scores)

  # Verify rownames exist
  if (is.null(days)) {
    stop("ICA scores missing row names for ", loc_name)
  }

  # ICA anomalies (already calculated)
```

```r
  ica_anom_days <- res$ica_anom_days

  # Debug
  cat("Location:", loc_name, "- ICA Anomalies detected:", length(ica_anom_days), "\n")
  if (length(ica_anom_days) > 0) {
    cat("  Anomaly days:", paste(head(ica_anom_days), collapse = ", "), "\n")
  }

  # Create the long data frame
  df_long <- data.frame(
    Day = rep(days, ncol(scores)),
    Source = rep(paste0("Source", 1:ncol(scores)), each = nrow(scores)),
    Score = c(scores),
    stringsAsFactors = FALSE
  )

  # Identify and label anomalies
  df_long$AnomalyType <- ifelse(df_long$Day %in% ica_anom_days, "ICA Anomaly", "Normal")

  # Debug
  cat("  Days marked as anomalies:", sum(df_long$AnomalyType == "ICA Anomaly"), "\n\n")

  # Create the plot - SEPARATE LAYERS
  ggplot(df_long, aes(x = Source, y = Score)) +
    geom_boxplot(outlier.shape = NA) +

    # Normal points first (light and small)
    geom_jitter(data = subset(df_long, AnomalyType == "Normal"),
                color = "gray70", alpha = 0.3, width = 0.15, size = 1) +

    # ICA anomalies on top (larger and bright)
    geom_point(data = subset(df_long, AnomalyType == "ICA Anomaly"),
               aes(color = "ICA Anomaly"), size = 1.0, alpha = 0.9) +

    scale_color_manual(values = c("ICA Anomaly" = "purple")) +
    labs(title = paste("ICA Source Score Boxplots with Anomalies -", loc_name),
         subtitle = paste("Total anomalies detected:", length(ica_anom_days)),
         x = "Independent Component Source",
         y = "Source Score",
         color = "Detection Method") +
    theme_minimal() +
    theme(legend.position = "bottom")
}
```

```r
# Loop over locations to generate the plots
for (loc_name in names(location_results)) {
  tryCatch({
    print(plot_ica_score_boxplot(location_results[[loc_name]], loc_name))
  }, error = function(e) {
    warning(paste("Failed to plot ICA Score Boxplot for", loc_name, ":", e$message))
  })
}
```

FA loadings

```r
## FA Factor Loadings Plot
plot_fa_loadings <- function(res, loc_name) {
  if (is.null(res$fa_loadings)) {
    message("FA results or loadings not available for ", loc_name)
    return(NULL)
  }

  loadings <- res$fa_loadings
  n_factors <- ncol(loadings)

  # Convert time intervals to hours (288 intervals = 24 hours)
  time_hours <- seq(0, 24, length.out = nrow(loadings))

  df <- data.frame(Time = time_hours, loadings)
  colnames(df) <- c("Time", paste0("Factor", 1:n_factors))

  df_long <- tidyr::pivot_longer(df, cols = -Time,
                                 names_to = "Factor",
                                 values_to = "Loading")

  ggplot(df_long, aes(x = Time, y = Loading, color = Factor)) +
    geom_line(size = 1) +
    facet_wrap(~Factor, ncol = 1, scales = "free_y") +
    labs(title = paste("FA Factor Loadings (Time Profile) -", loc_name),
         x = "Hour of Day",
         y = "Loading/Factor Weight") +
    theme_minimal() +
    scale_x_continuous(breaks = seq(0, 24, 4))
}

# Loop over locations to generate the plots
```

```r
for (loc_name in names(location_results)) {
  print(plot_fa_loadings(location_results[[loc_name]], loc_name))
}
```

Venn diagrams we are not using this

```r
plot_anomaly_venn<- function(res, loc_name) {
  anomaly_sets <- list(
    PCA = res$pca_anom_days,
    FA  = res$fa_anom_days,
    ICA = res$ica_anom_days
  )

  # Remove empty sets
  anomaly_sets <- anomaly_sets[sapply(anomaly_sets, length) > 0]

  if(length(anomaly_sets) < 2) {
    message("Not enough methods for Venn at ", loc_name)
    return(NULL)
  }

  ggvenn(
    anomaly_sets,
    fill_color = c("#FF6B6B", "#4D96FF", "#9B5DE5"),
    fill_alpha = 0.55,
    stroke_size = 0.8,
    text_size = 6,           # Much larger text
    show_percentage = FALSE # Turn off % labels
  ) +
  labs(title = paste("Anomaly Overlap -", loc_name)) +
  theme(plot.title = element_text(hjust = 0.5, size = 14, face = "bold"))
}


# Loop over locations
for (loc_name in names(location_results)) {
  tryCatch({
    print(plot_anomaly_venn(location_results[[loc_name]], loc_name))
  }, error = function(e) {
    warning(paste("Failed to plot Venn diagram for", loc_name, ":", e$message))
  })
}
```

Cluster

```r
# Step 1: Get the minimum number of retained PCs across all locations
min_k <- min(sapply(location_results, function(res) res$k_pca))
cat("Minimum number of available PCs across locations:", min_k, "\n")
```

```
## Minimum number of available PCs across locations: 2
```

```r
# Step 2: Extract ONLY the first min_k PCs from each location
pca_all_scores <- do.call(rbind, lapply(location_results, function(res) {
  if (ncol(res$pca_scores) >= min_k) {
    res$pca_scores[, 1:min_k, drop = FALSE]
  } else {
    NULL  # In case some failed completely
  }
}))

# Optional: remove duplicate row names
pca_all_scores <- pca_all_scores[!duplicated(rownames(pca_all_scores)), ]


# Step 3: Compute silhouette scores for different K

silhouette_scores <- data.frame(K = integer(), Silhouette = double())

for (k in 2:6) {
  km <- kmeans(scale(pca_all_scores), centers = k, nstart = 25)
  sil <- silhouette(km$cluster, dist(scale(pca_all_scores)))
  silhouette_scores <- rbind(silhouette_scores,
                             data.frame(K = k, Silhouette = mean(sil[, 3])))
}

print(silhouette_scores)
```

```
##   K Silhouette
## 1 2  0.3098061
## 2 3  0.3424861
## 3 4  0.3226218
## 4 5  0.3302555
## 5 6  0.3611448
```

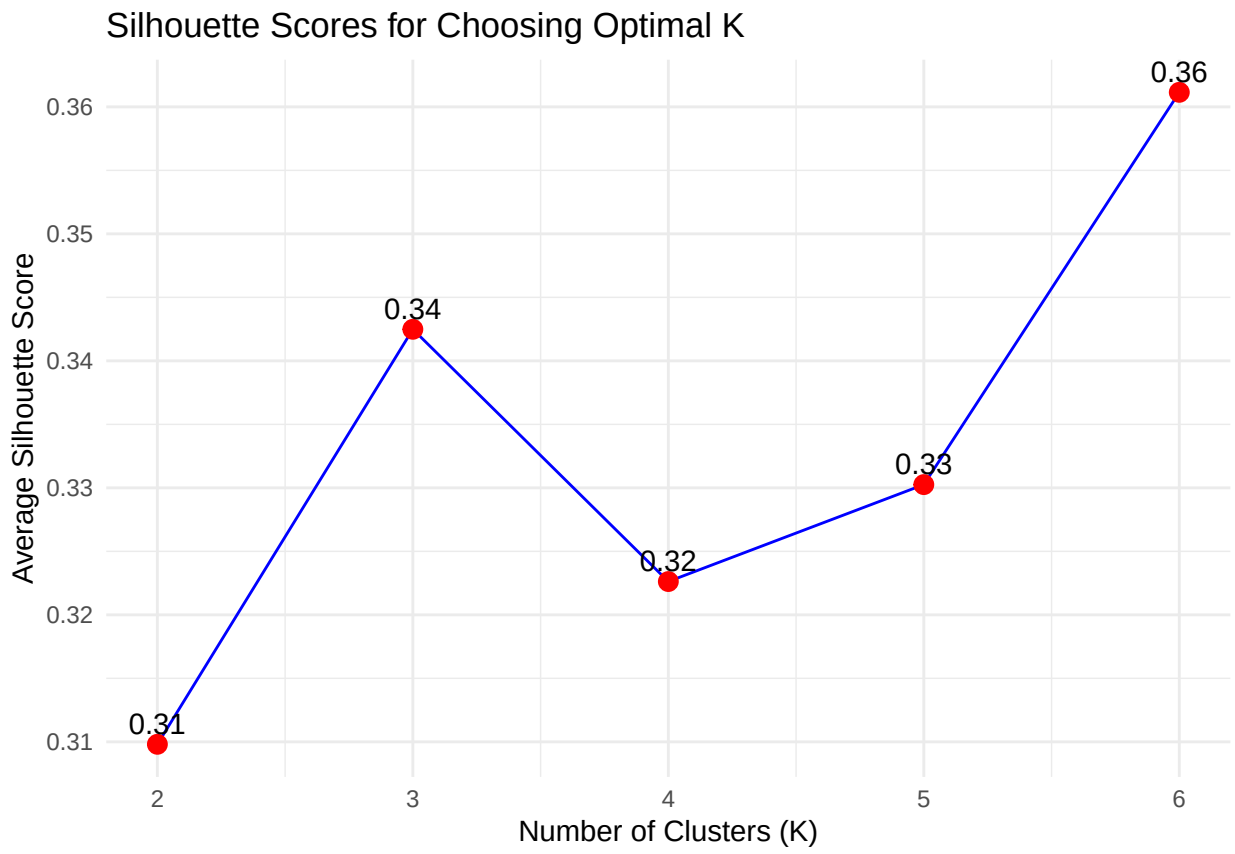
```r
best_k <- silhouette_scores$K[which.max(silhouette_scores$Silhouette)]
cat("Best number of clusters based on silhouette:", best_k, "\n")
```

```
## Best number of clusters based on silhouette: 6
```

```r
km_res <- kmeans(scale(pca_all_scores), centers = best_k, nstart = 25)

cluster_labels <- data.frame(
  Day = rownames(pca_all_scores),
  Cluster = as.factor(km_res$cluster)
)

head(cluster_labels)
```

```
##              Day Cluster
## WkDay-1 WkDay-1       5
## WkDay-2 WkDay-2       5
## WkDay-3 WkDay-3       5
## WkDay-4 WkDay-4       6
## WkDay-5 WkDay-5       3
## WkDay-6 WkDay-6       2
```

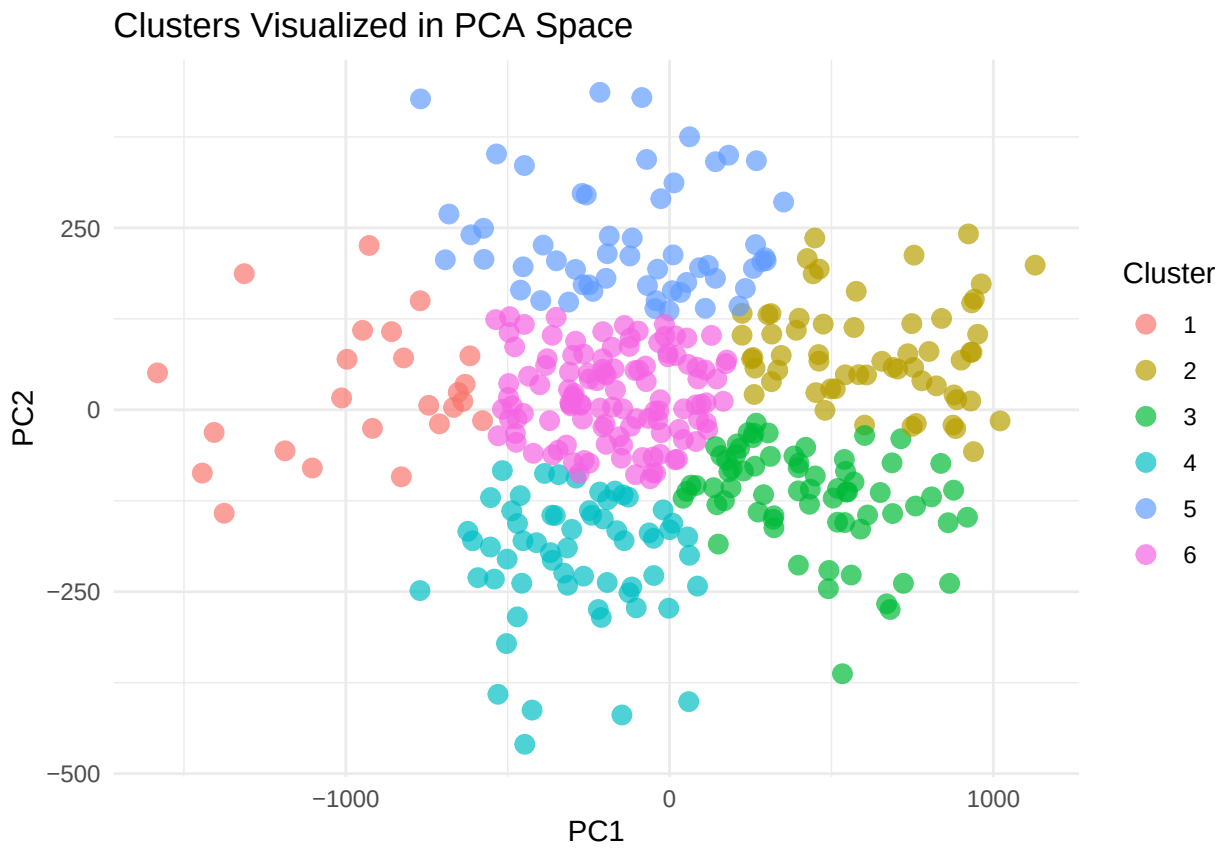
```r
ggplot(silhouette_scores, aes(x = K, y = Silhouette)) +
  geom_line(color = "blue") +
  geom_point(size = 3, color = "red") +
  geom_text(aes(label = round(Silhouette, 2)), vjust = -0.5) +
  labs(title = "Silhouette Scores for Choosing Optimal K",
       x = "Number of Clusters (K)",
       y = "Average Silhouette Score") +
  theme_minimal()
```



Plot PCA Score Scatterplot Colored by Cluster

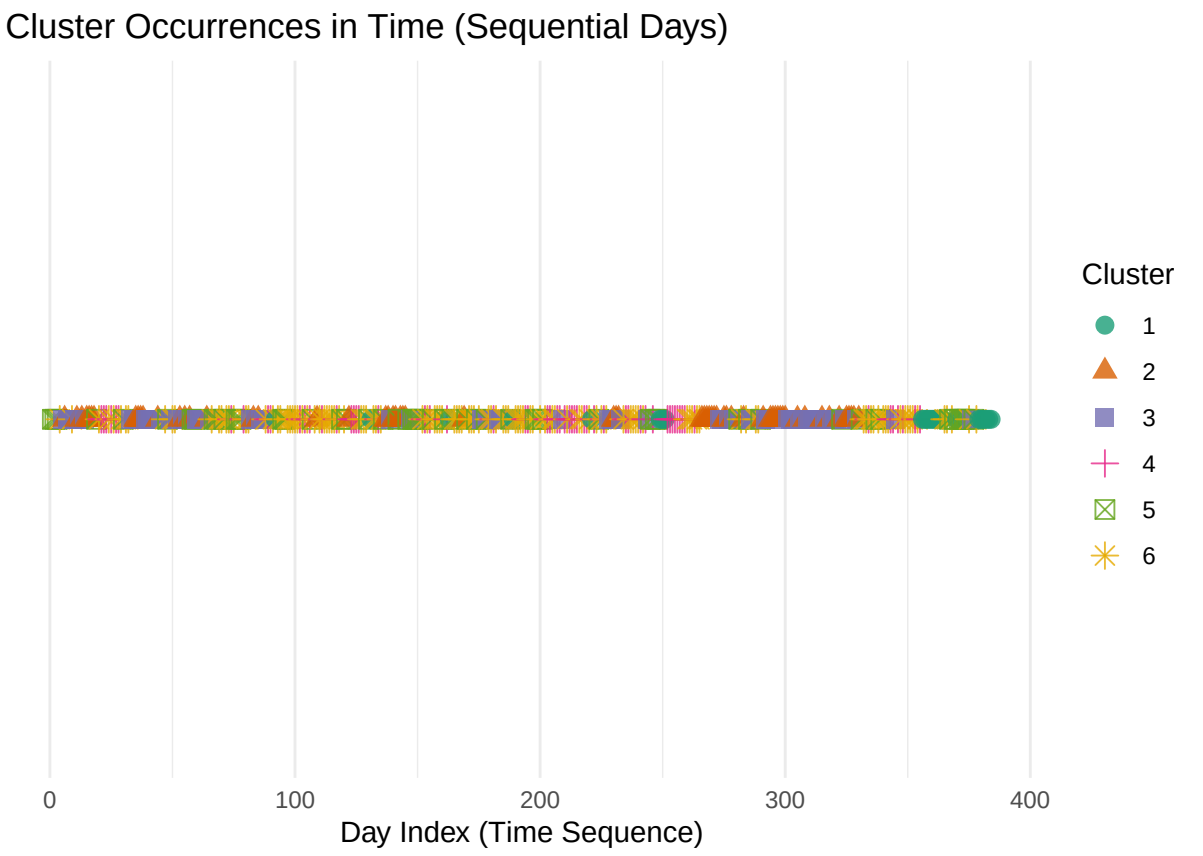
This shows how clusters separate in feature space, not along time.

```
pca_cluster_plot <- function(pca_all_scores, cluster_labels) {
  df <- data.frame(pca_all_scores[, 1:2], Cluster = cluster_labels$Cluster)
  colnames(df)[1:2] <- c("PC1", "PC2")

  ggplot(df, aes(x = PC1, y = PC2, color = Cluster)) +
    geom_point(size = 3, alpha = 0.7) +
    labs(title = "Clusters Visualized in PCA Space",
         x = "PC1",
         y = "PC2",
         color = "Cluster") +
    theme_minimal()
}

pca_cluster_plot(pca_all_scores, cluster_labels)
```




```r
plot_cluster_timeline <- function(cluster_labels) {
  cluster_labels$DayIndex <- 1:nrow(cluster_labels)  # Order by time

  ggplot(cluster_labels, aes(x = DayIndex, y = 1, color = Cluster, shape = Cluster)) +
    geom_point(size = 3, alpha = 0.8) +
    scale_color_brewer(palette = "Dark2") +
    labs(
      title = "Cluster Occurrences in Time (Sequential Days)",
      x = "Day Index (Time Sequence)",
      y = "",
      color = "Cluster",
      shape = "Cluster"
    ) +
    theme_minimal() +
    theme(
      axis.text.y = element_blank(),
      axis.ticks.y = element_blank(),
      panel.grid.major.y = element_blank(),
      panel.grid.minor.y = element_blank()
    )
}
```

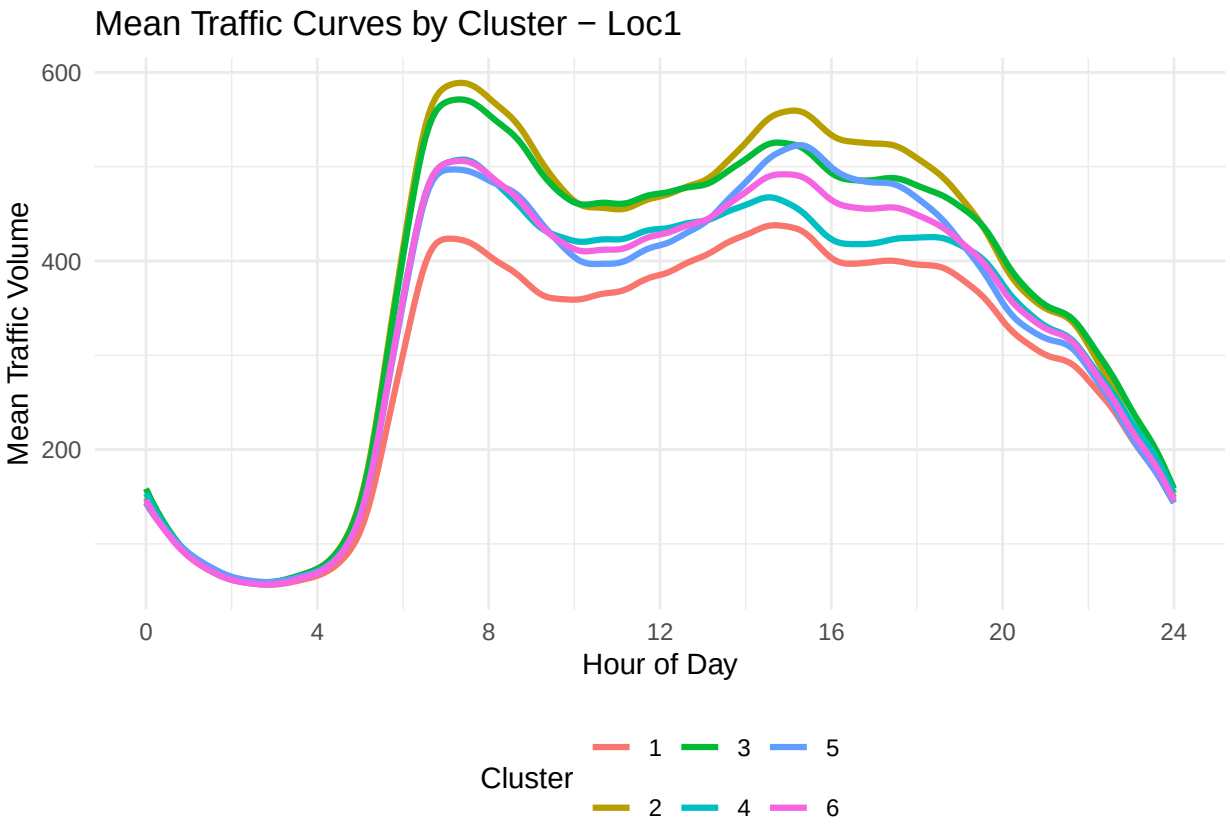
```r
# Example call:
plot_cluster_timeline(cluster_labels)
```

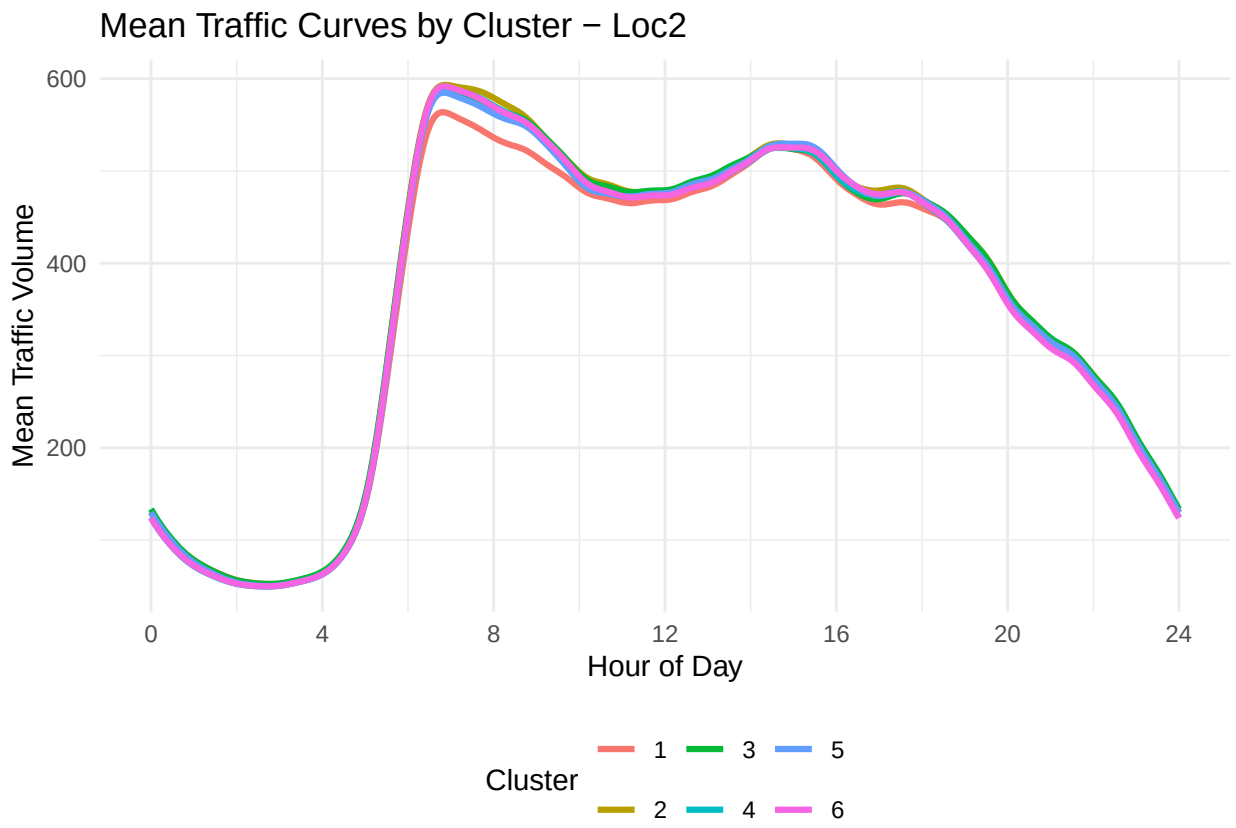


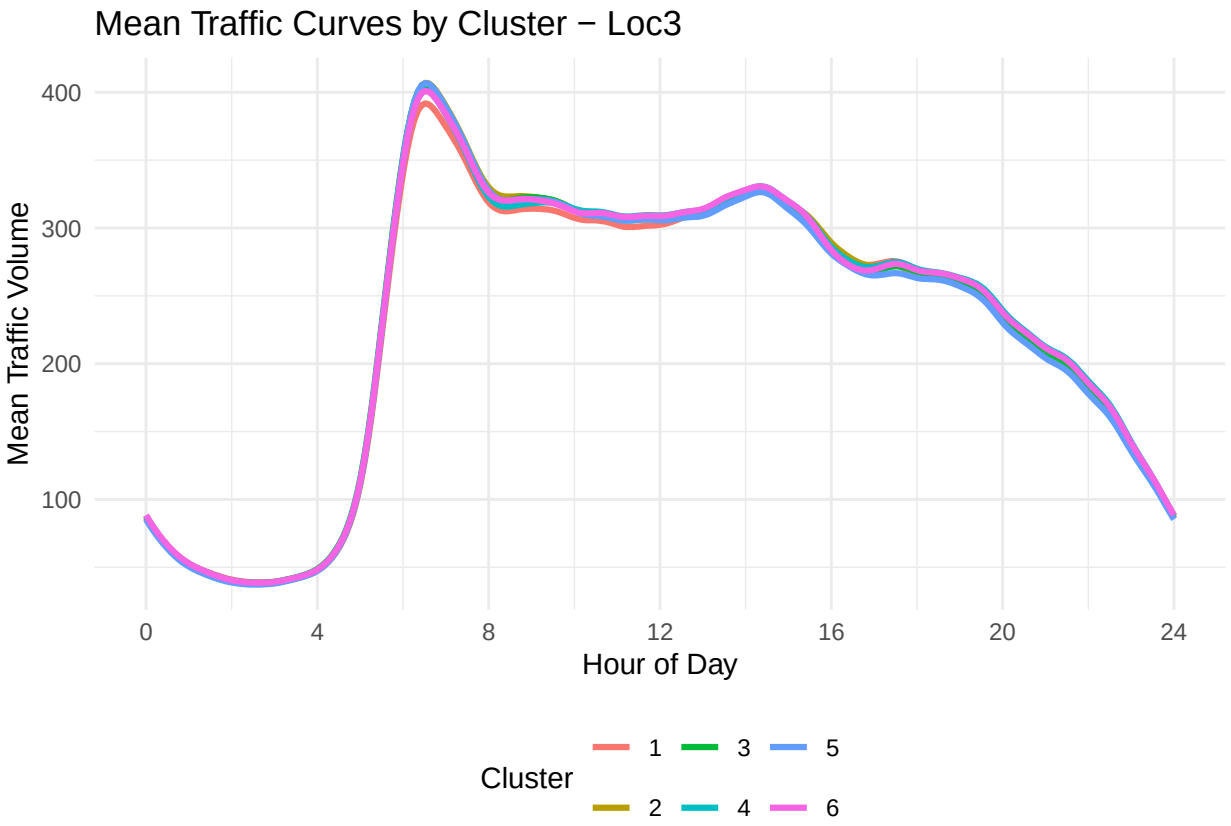

```r
plot_cluster_mean_shapes <- function(res, X, cluster_df, loc_name) {
  X_t <- t(X)

  df_long <- data.frame(
    Day = rownames(X_t),
    X_t,
    Cluster = cluster_df$Cluster[match(rownames(X_t), cluster_df$Day)]
  ) %>%
    filter(!is.na(Cluster)) %>%
    pivot_longer(cols = -c(Day, Cluster),
                 names_to = "TimePoint",
                 values_to = "Traffic") %>%
    mutate(TimeHour = (as.numeric(gsub("X", "", TimePoint)) - 1) * (24 / 287))

  df_mean <- df_long %>%
    group_by(Cluster, TimeHour) %>%
    summarise(MeanTraffic = mean(Traffic, na.rm = TRUE), .groups = "drop")
```

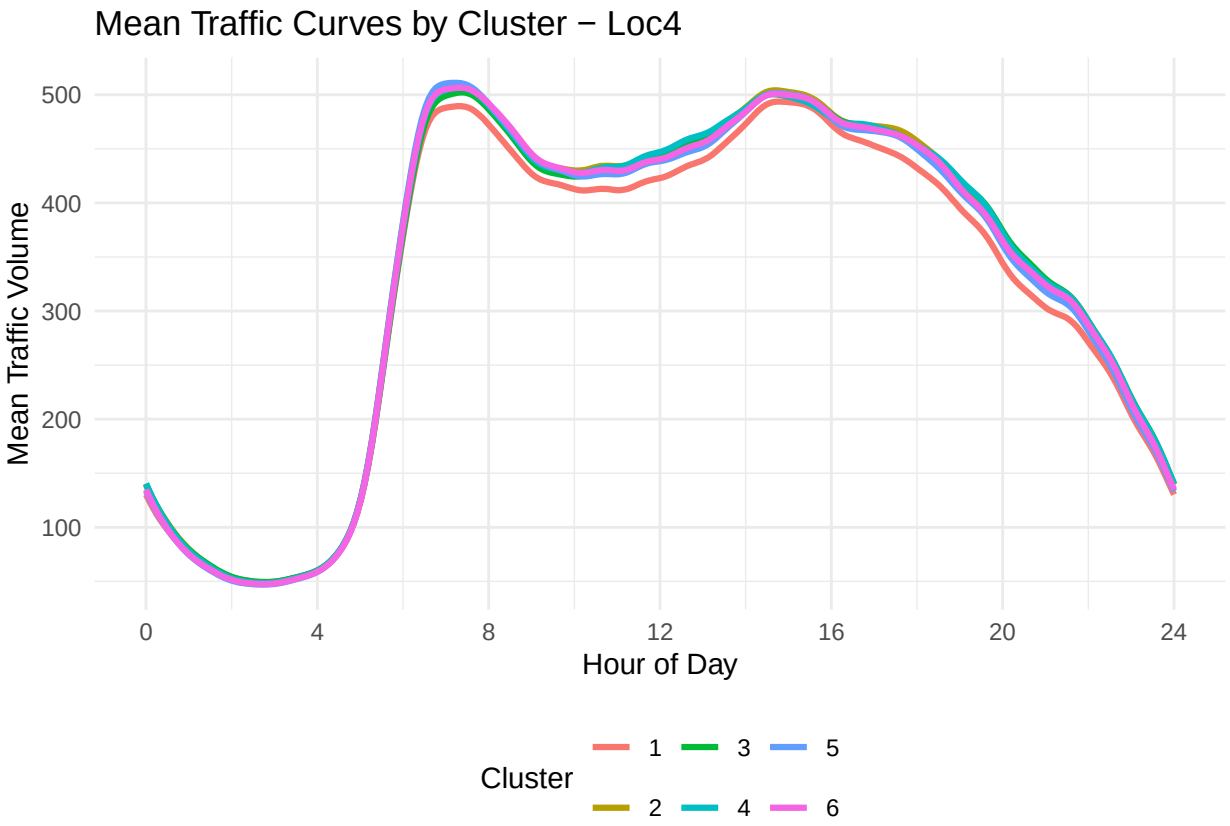
```
  ggplot(df_mean, aes(x = TimeHour, y = MeanTraffic, color = Cluster)) +
    geom_line(size = 1.1) +
    labs(
      title = paste("Mean Traffic Curves by Cluster -", loc_name),
      x = "Hour of Day",
      y = "Mean Traffic Volume",
      color = "Cluster"
    ) +
    scale_x_continuous(breaks = seq(0, 24, 4)) +
    theme_minimal() +
    theme(legend.position = "bottom") +
    guides(color = guide_legend(position = "bottom"))
}

for (loc_name in names(location_results)) {
  print(
    plot_cluster_mean_shapes(
      res = location_results[[loc_name]],  # model results for this location
      X = df[[loc_name]],                  # original traffic data
      cluster_df = cluster_labels,         # full anomaly cluster labels
      loc_name = loc_name                  # title
    )
  )
}
```

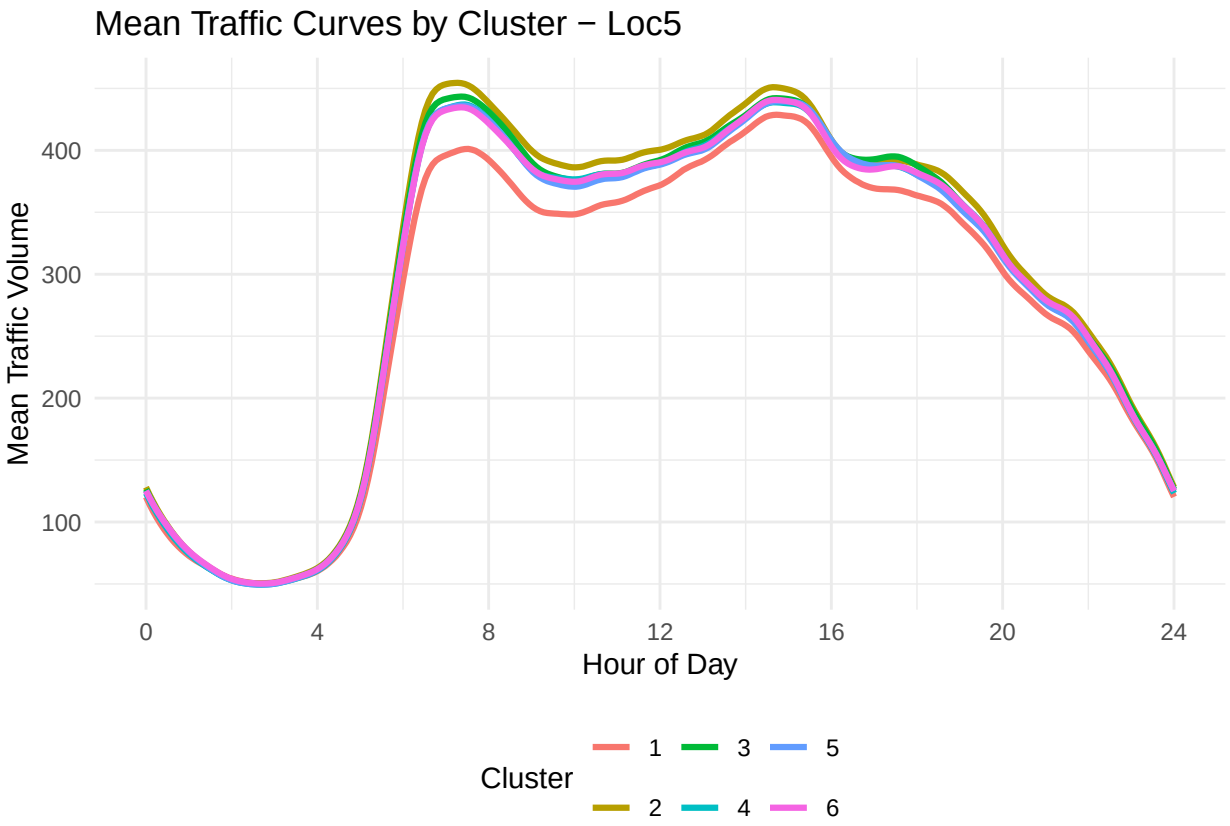
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

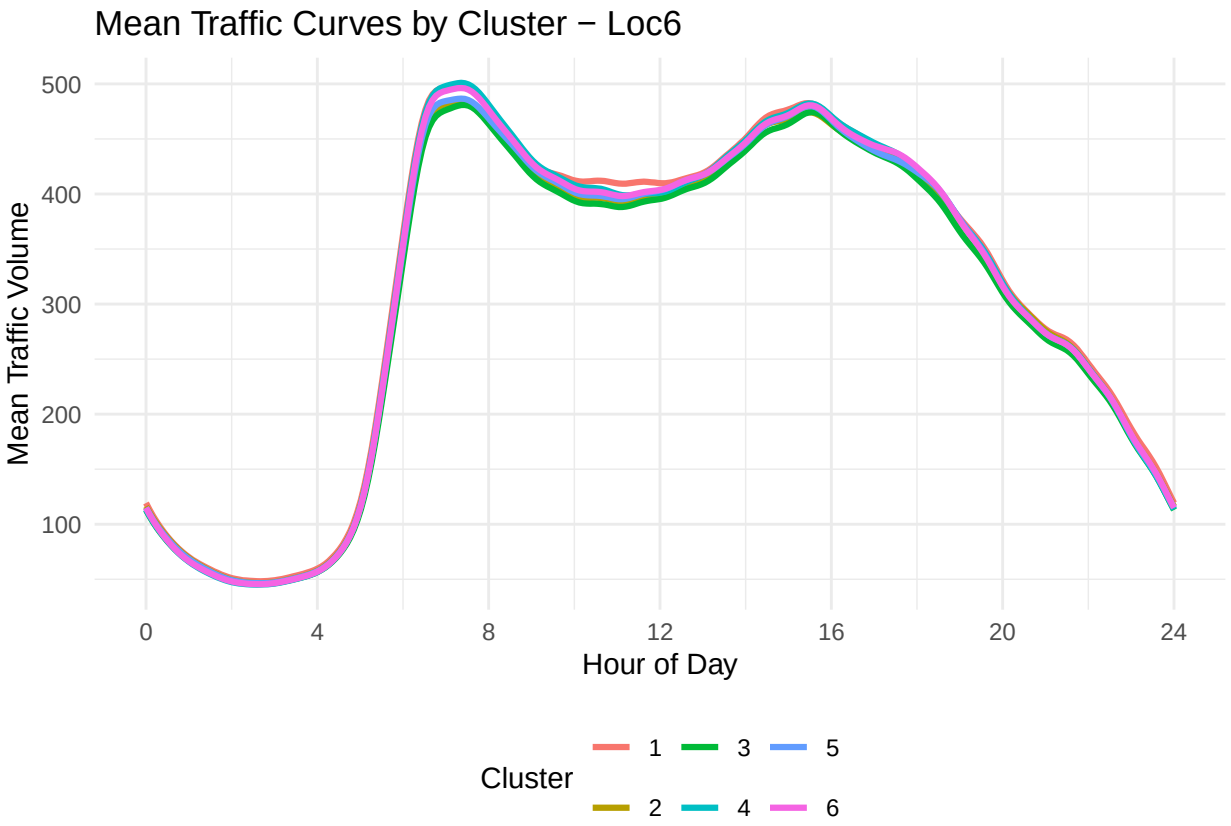

Mean Traffic Curves by Cluster − Loc1
Mean Traffic Volume
600
400
200
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

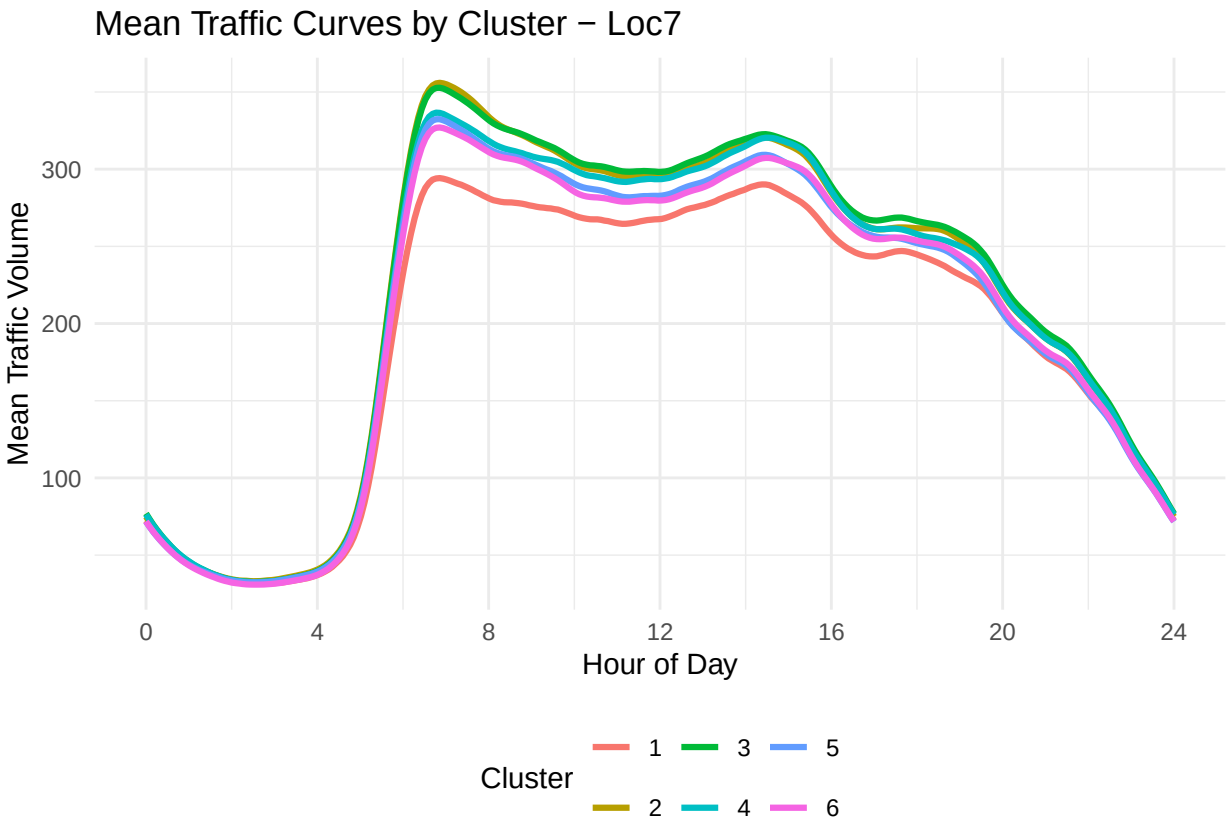

Mean Traffic Curves by Cluster − Loc2
600
400
200
Mean Traffic Volume
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

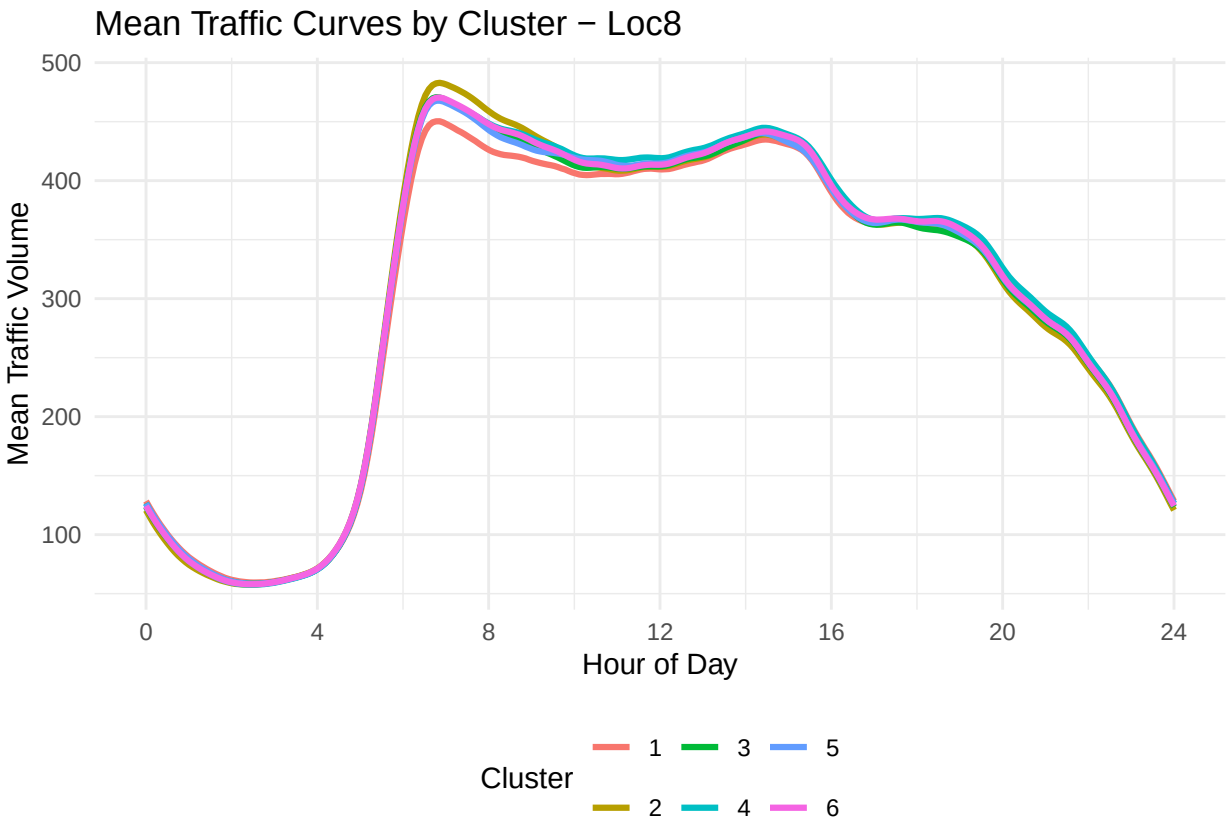

Mean Traffic Curves by Cluster − Loc3
Mean Traffic Volume
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

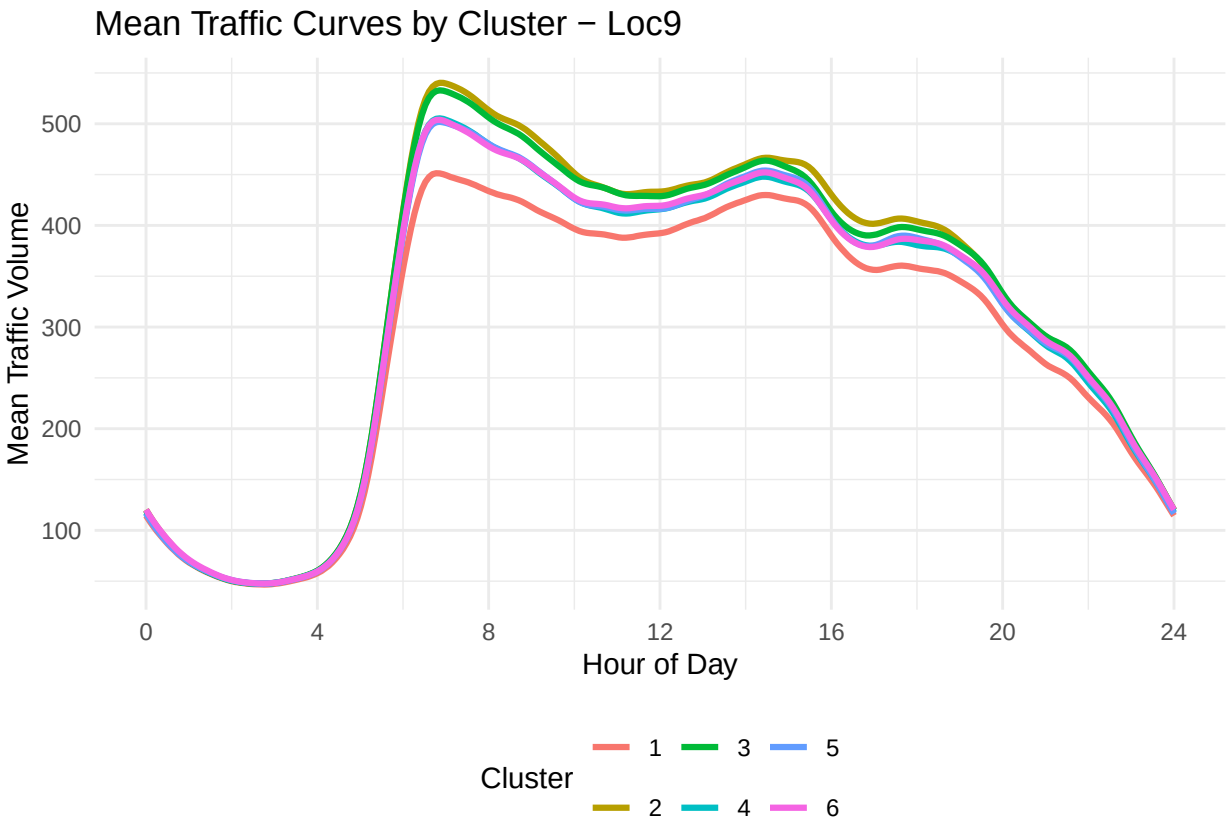

Mean Traffic Curves by Cluster − Loc4
Mean Traffic Volume
500
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

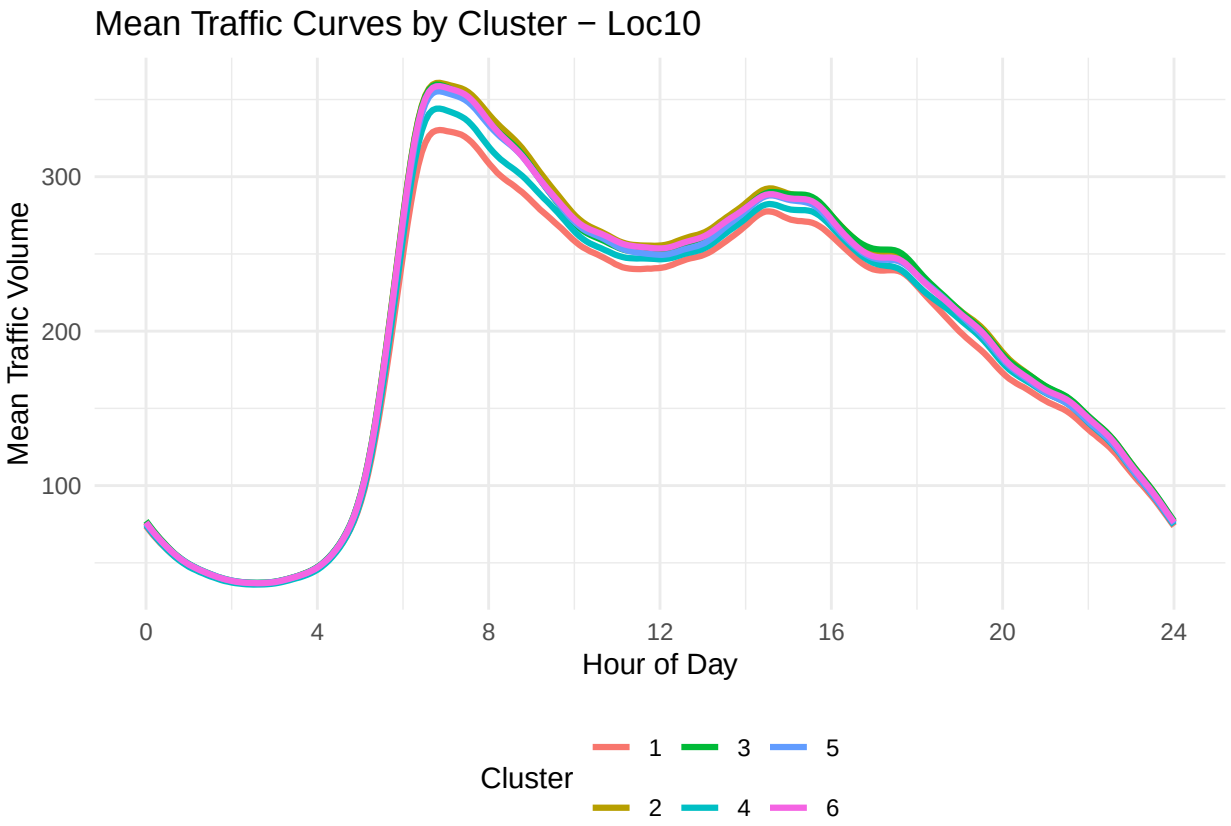

Mean Traffic Curves by Cluster − Loc5
Mean Traffic Volume
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

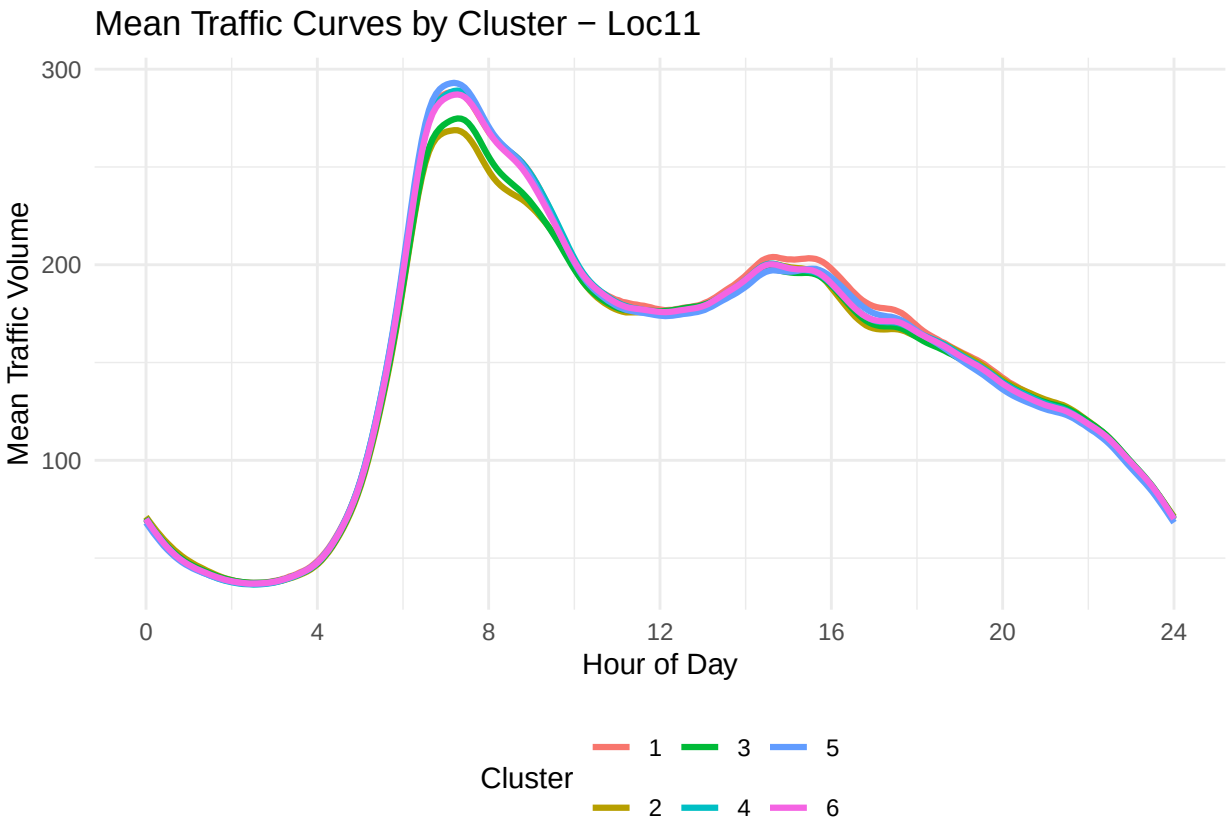

Mean Traffic Curves by Cluster − Loc6
Mean Traffic Volume
500
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

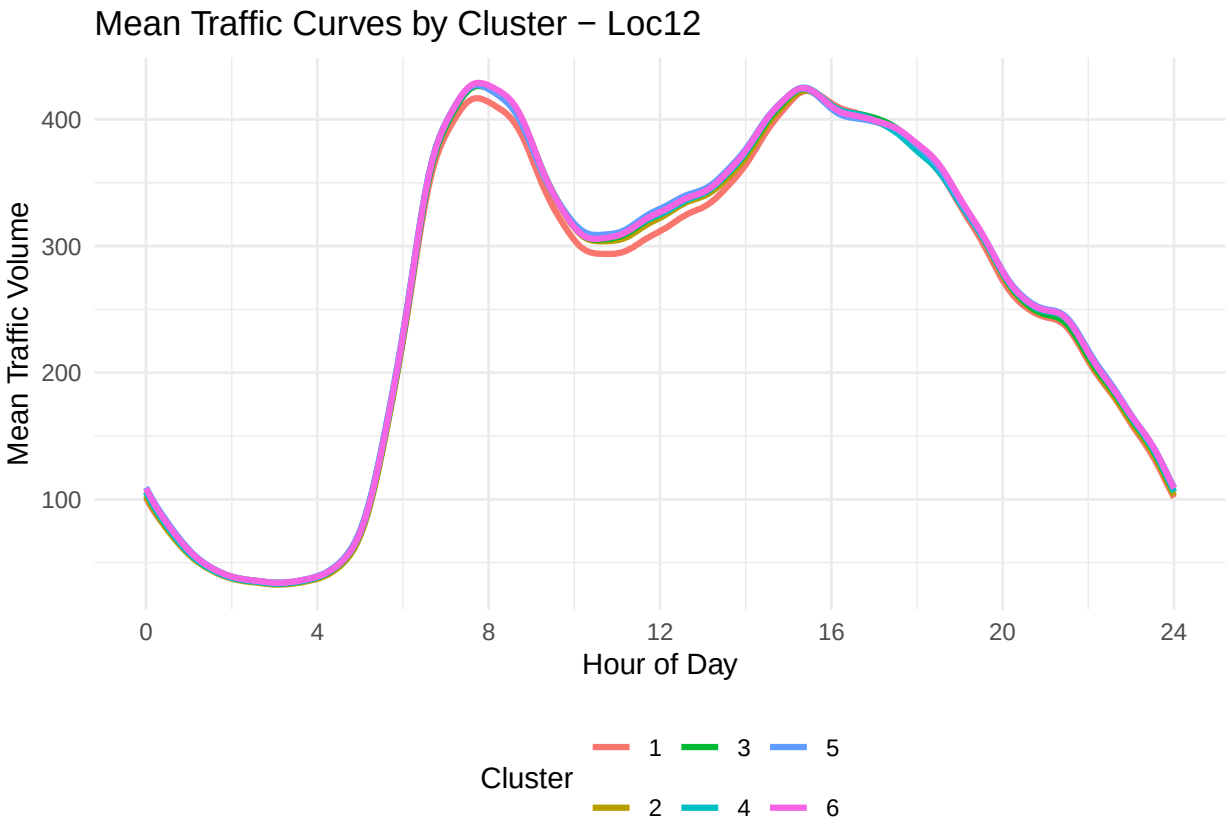

Mean Traffic Curves by Cluster − Loc7
Mean Traffic Volume
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

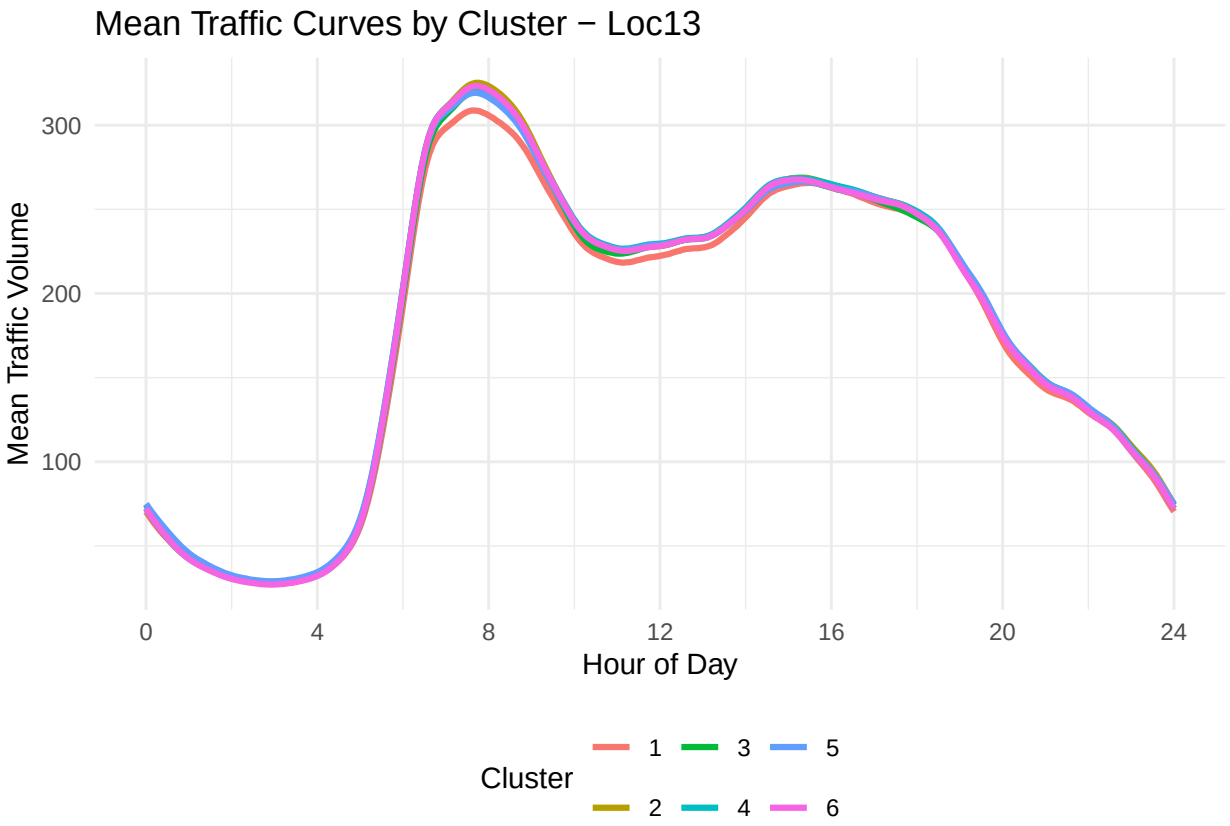

Mean Traffic Curves by Cluster − Loc8
Mean Traffic Volume
500
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

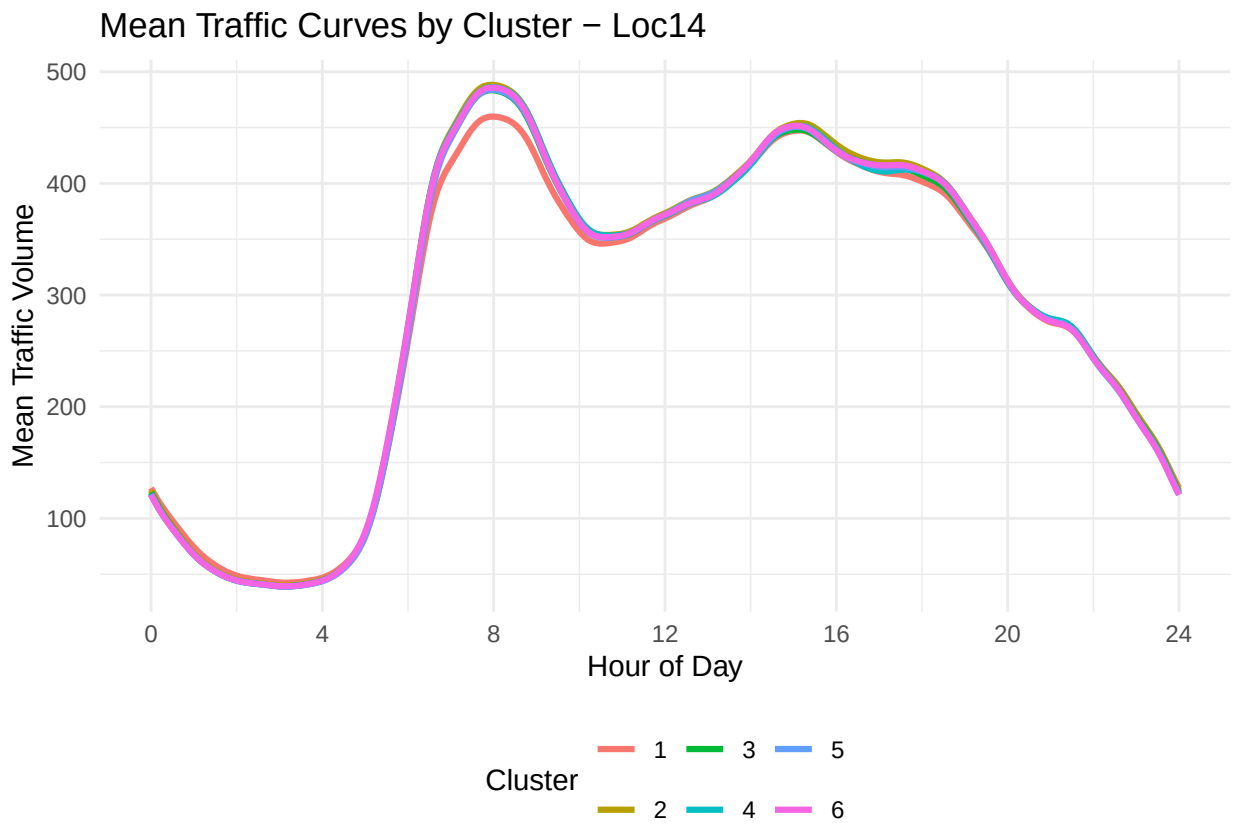

Mean Traffic Curves by Cluster − Loc9
Mean Traffic Volume
500
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

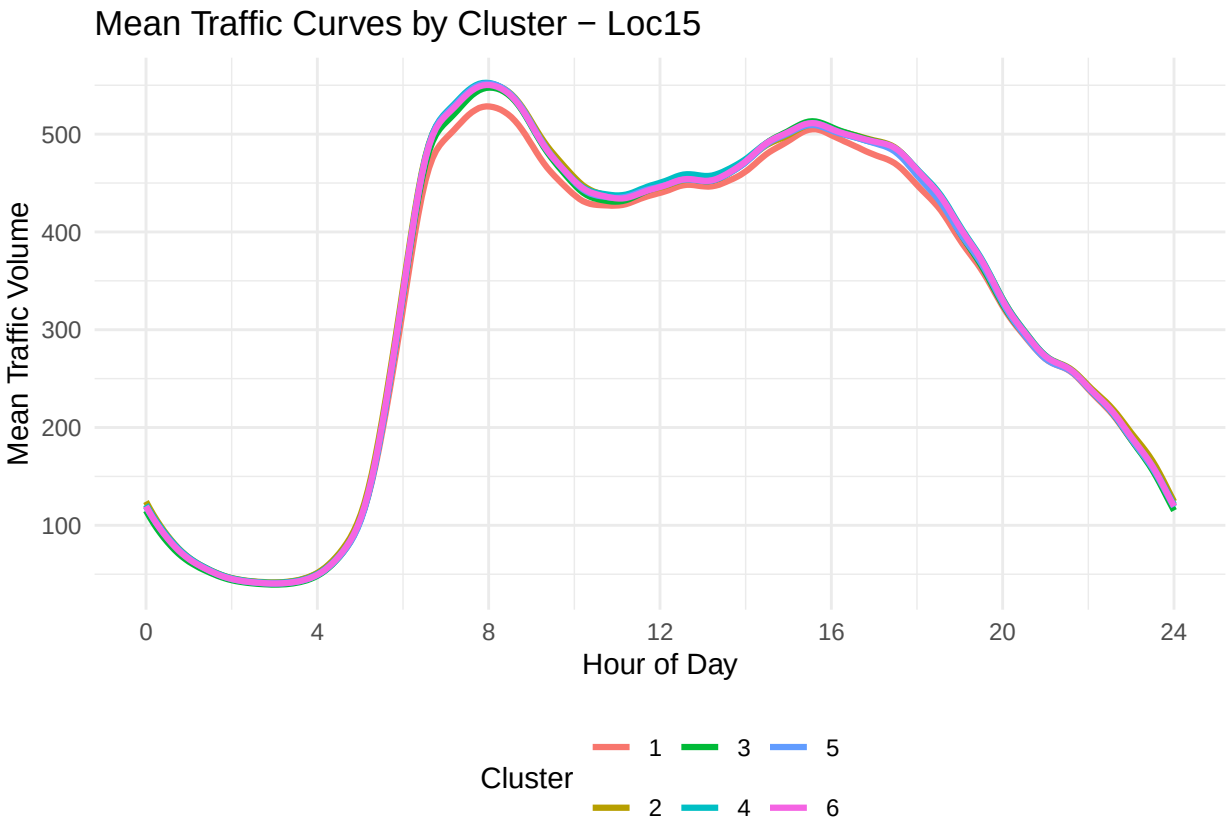

Mean Traffic Curves by Cluster − Loc10
Mean Traffic Volume
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

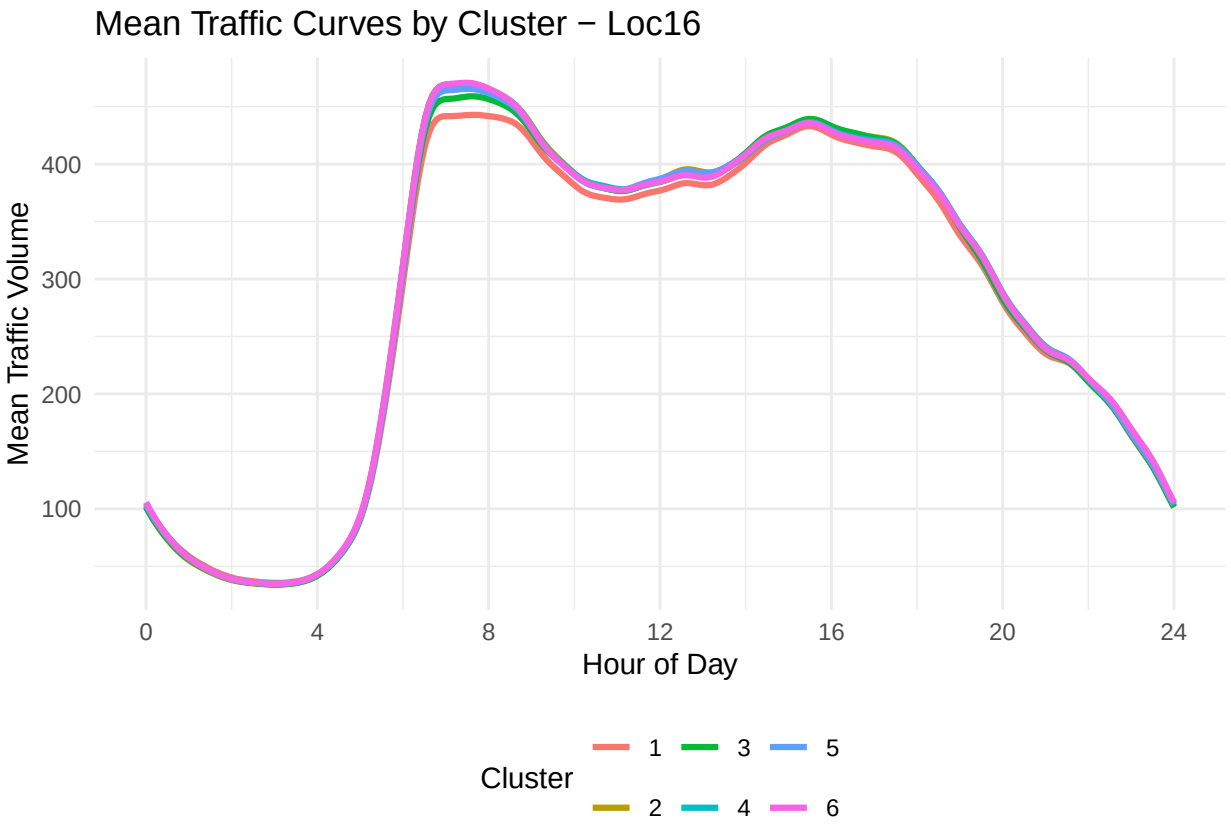

Mean Traffic Curves by Cluster − Loc11
Mean Traffic Volume
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

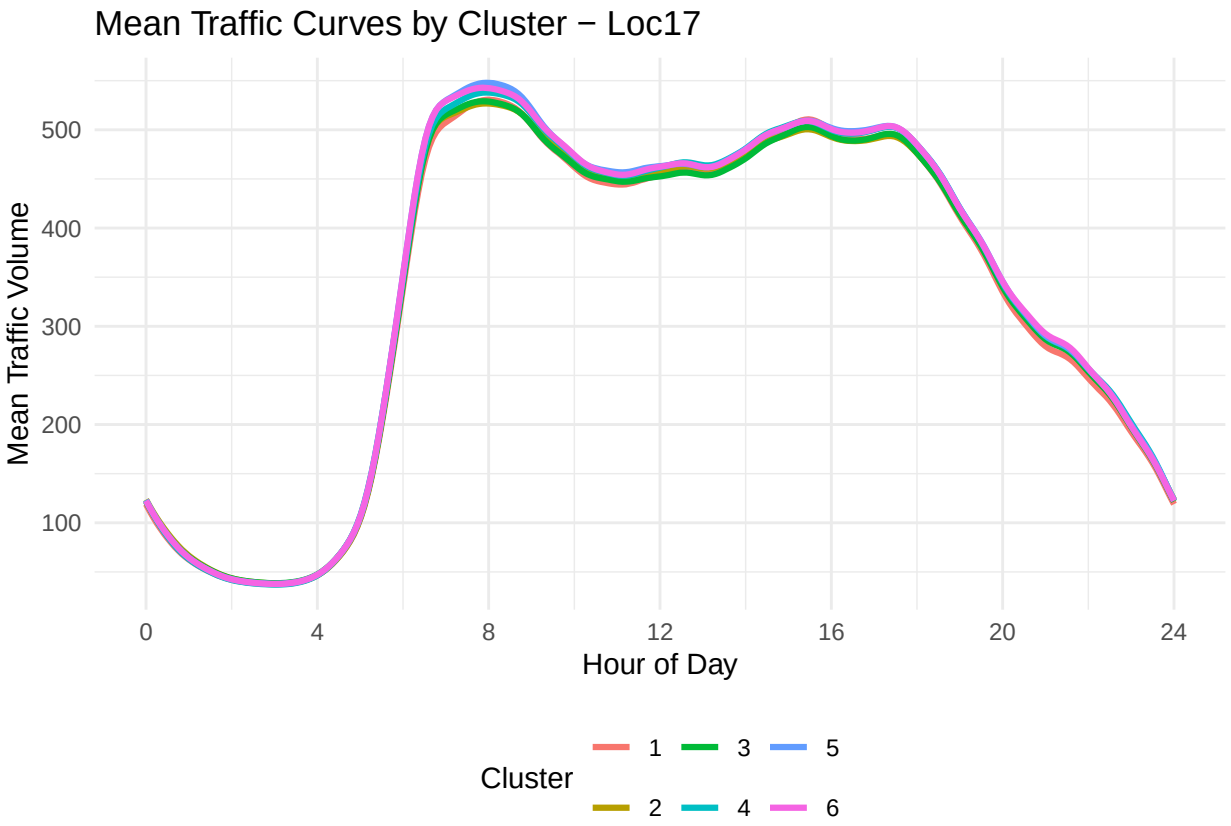

Mean Traffic Curves by Cluster − Loc12
Mean Traffic Volume
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

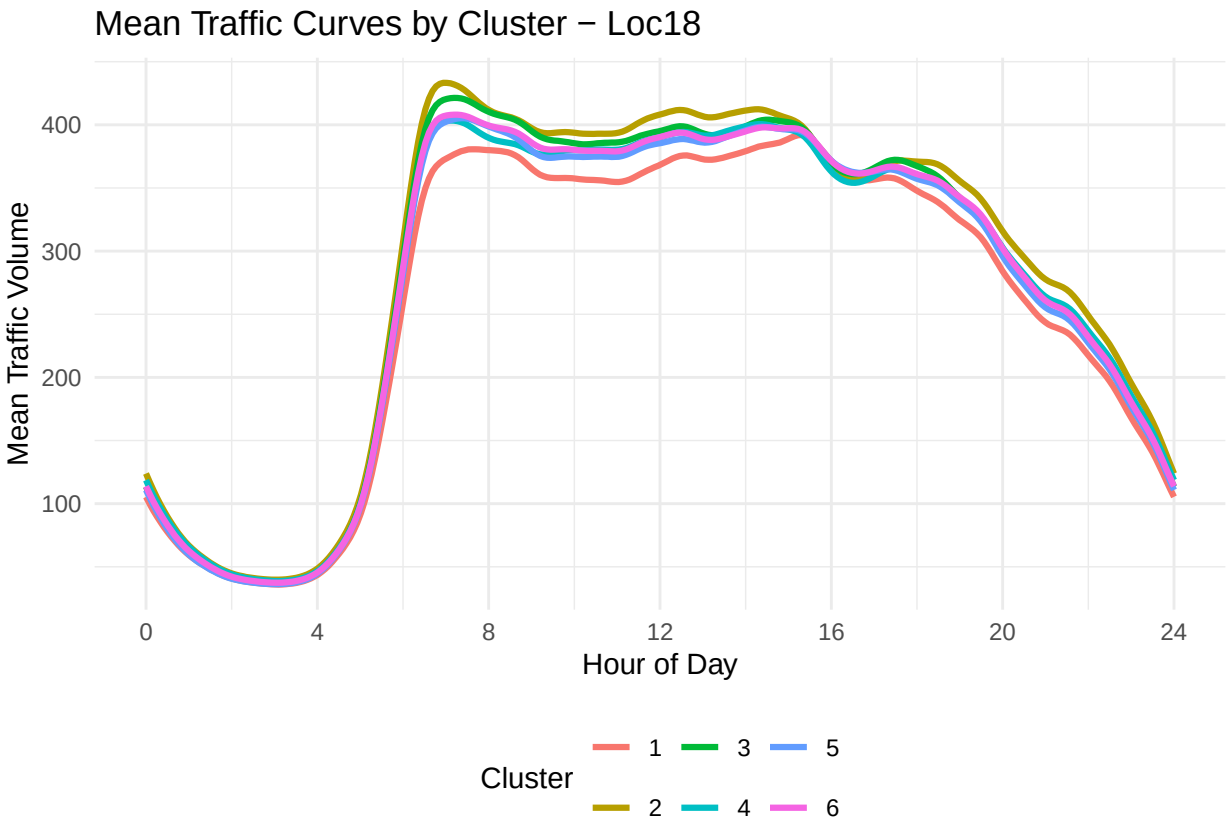

Mean Traffic Curves by Cluster − Loc13
Mean Traffic Volume
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

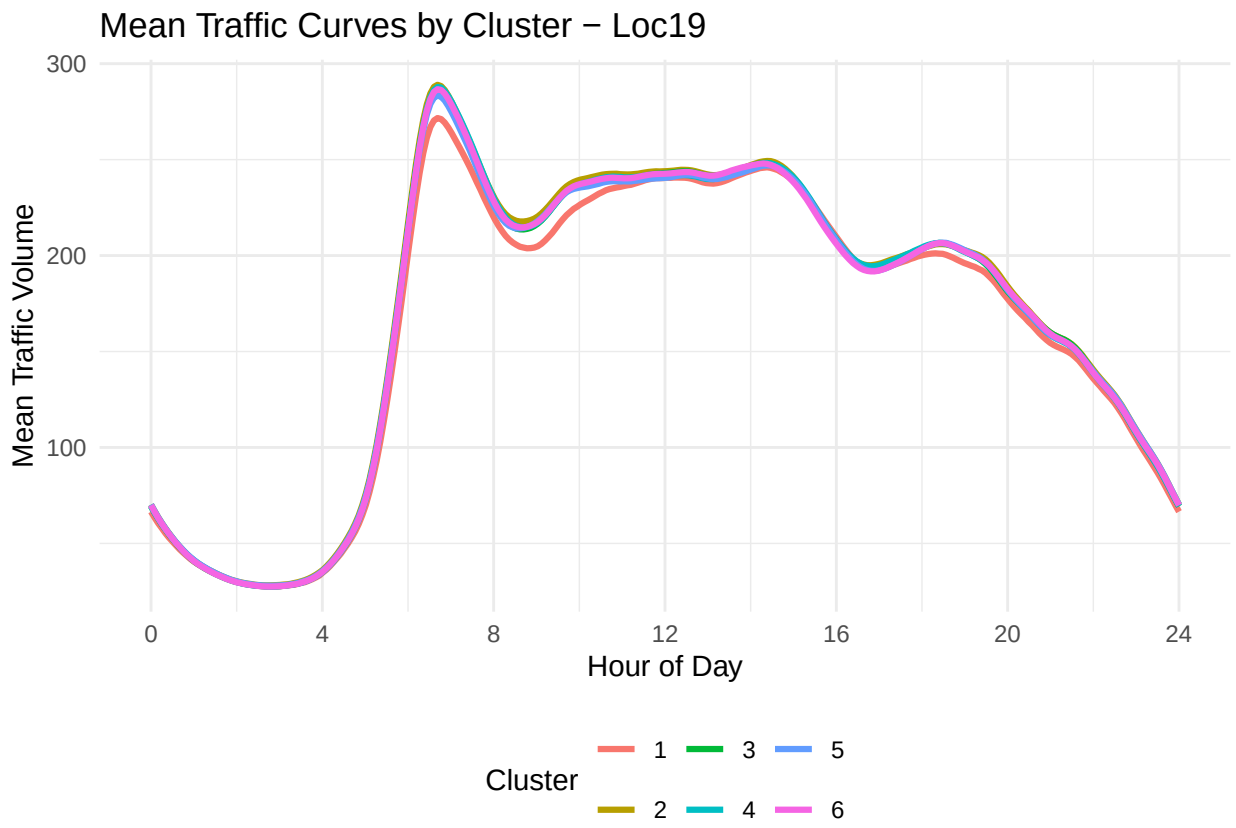

Mean Traffic Curves by Cluster − Loc14
Mean Traffic Volume
500
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

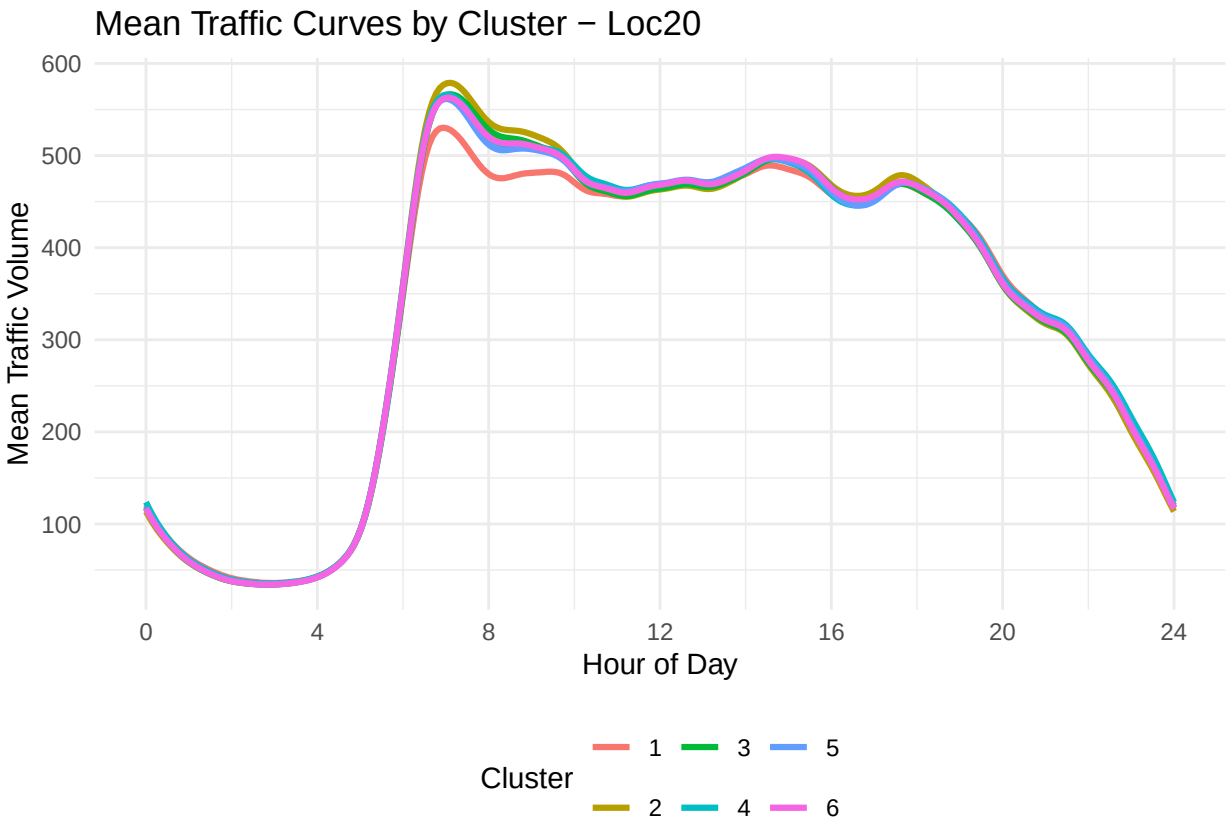

Mean Traffic Curves by Cluster − Loc15
Mean Traffic Volume
500
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

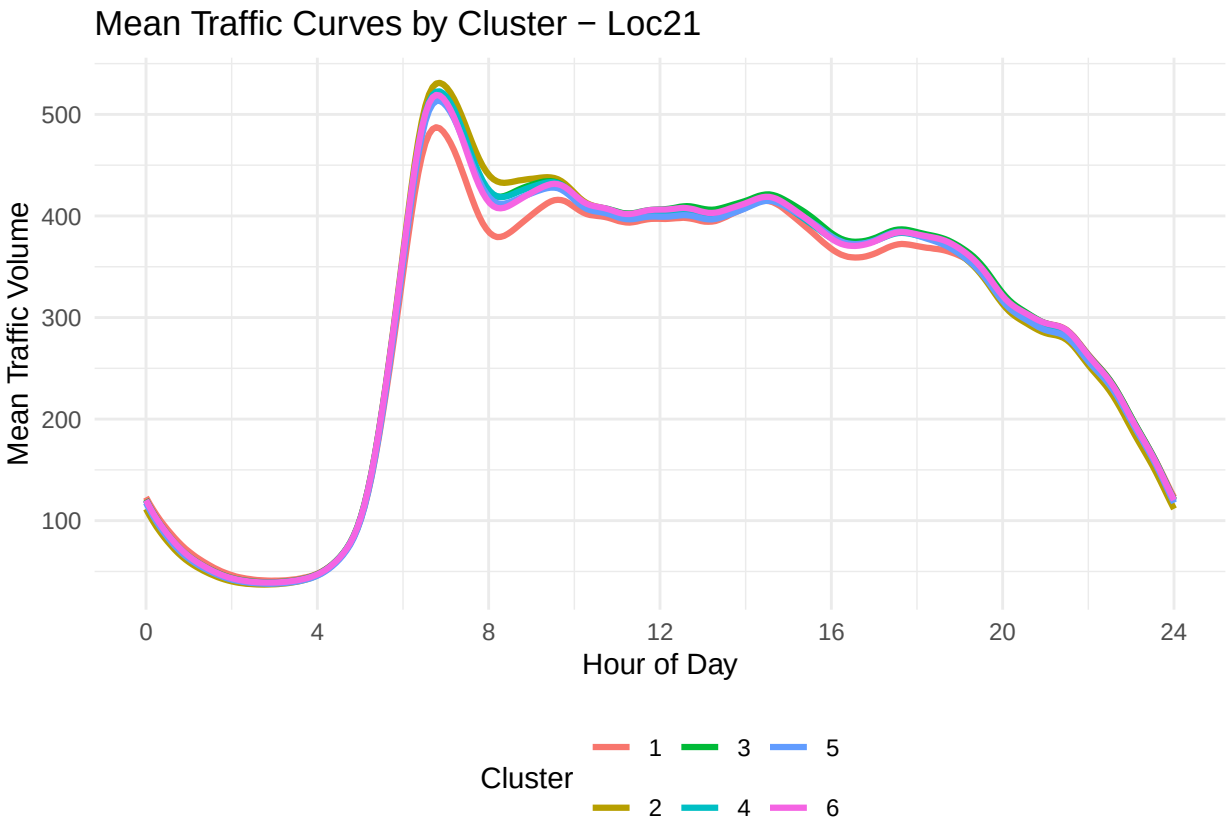

Mean Traffic Curves by Cluster − Loc16
Mean Traffic Volume
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

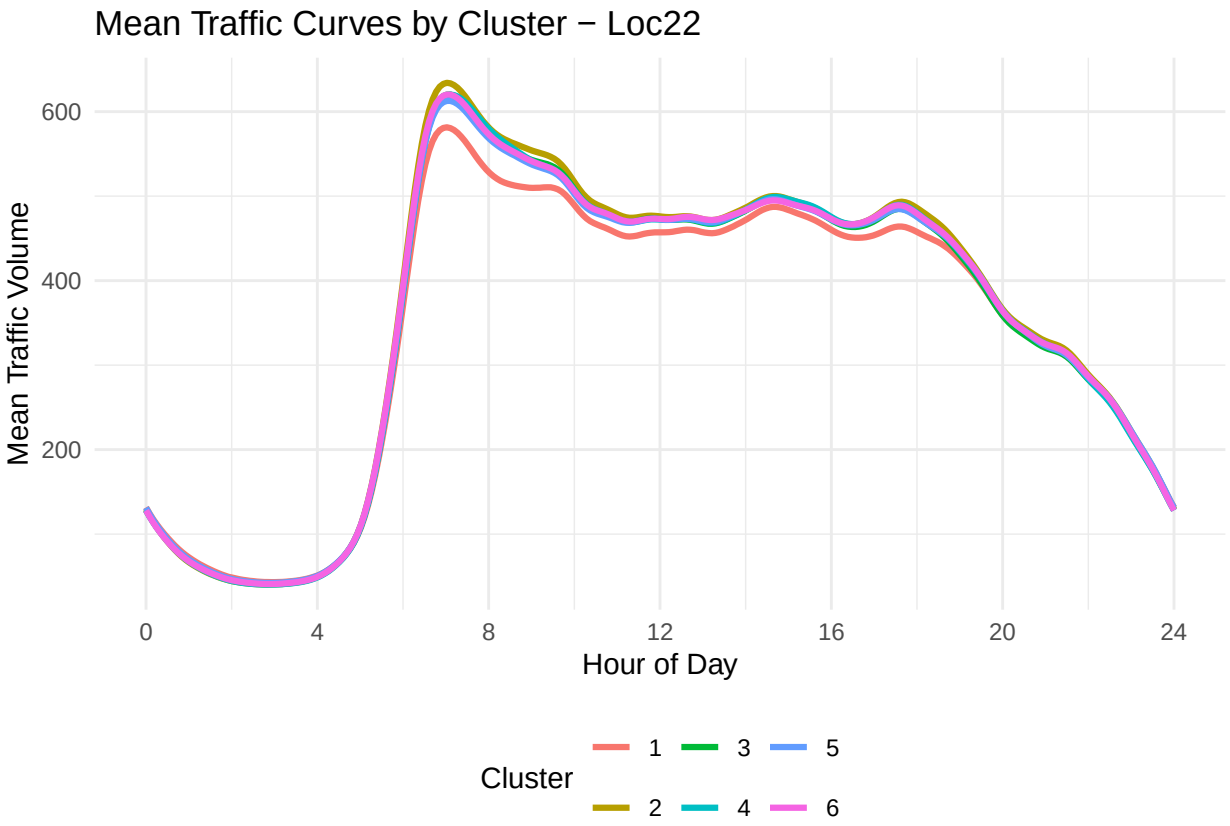

Mean Traffic Curves by Cluster − Loc17
Mean Traffic Volume
500
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

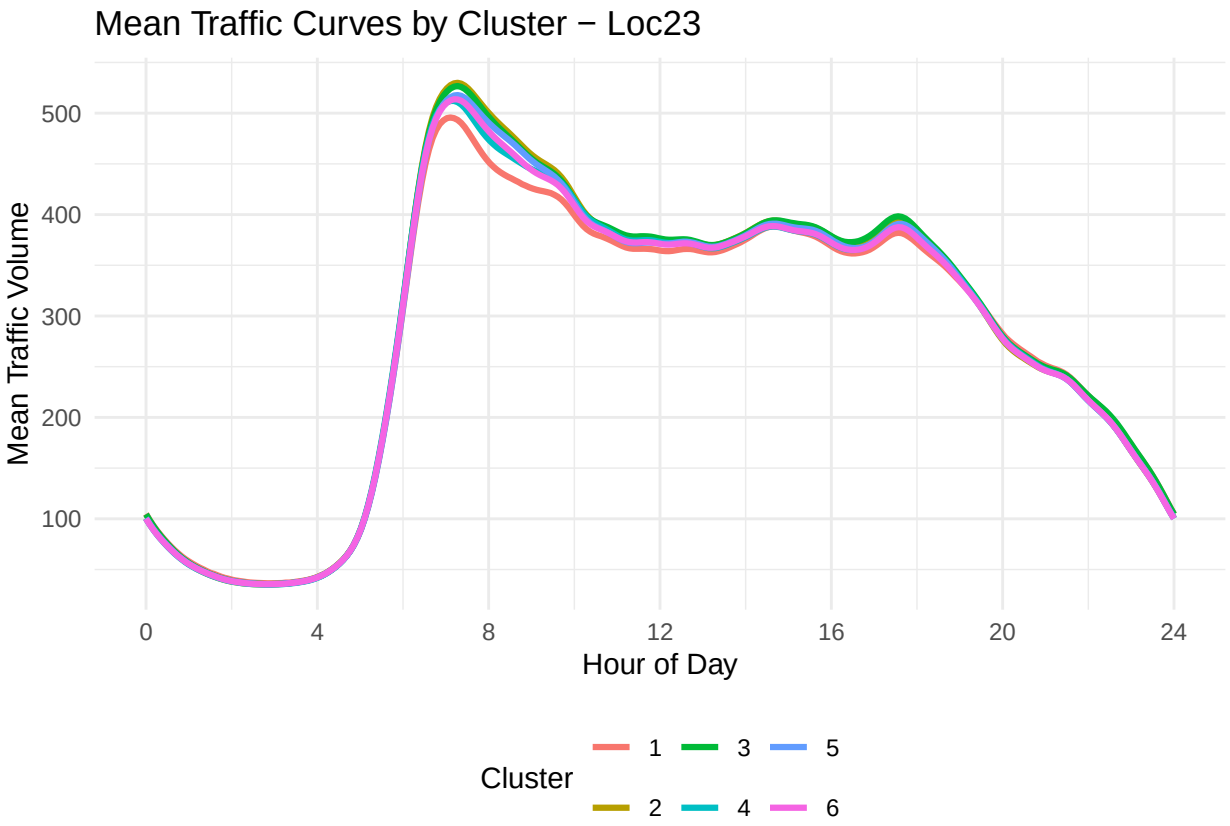

Mean Traffic Curves by Cluster − Loc18
Mean Traffic Volume
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

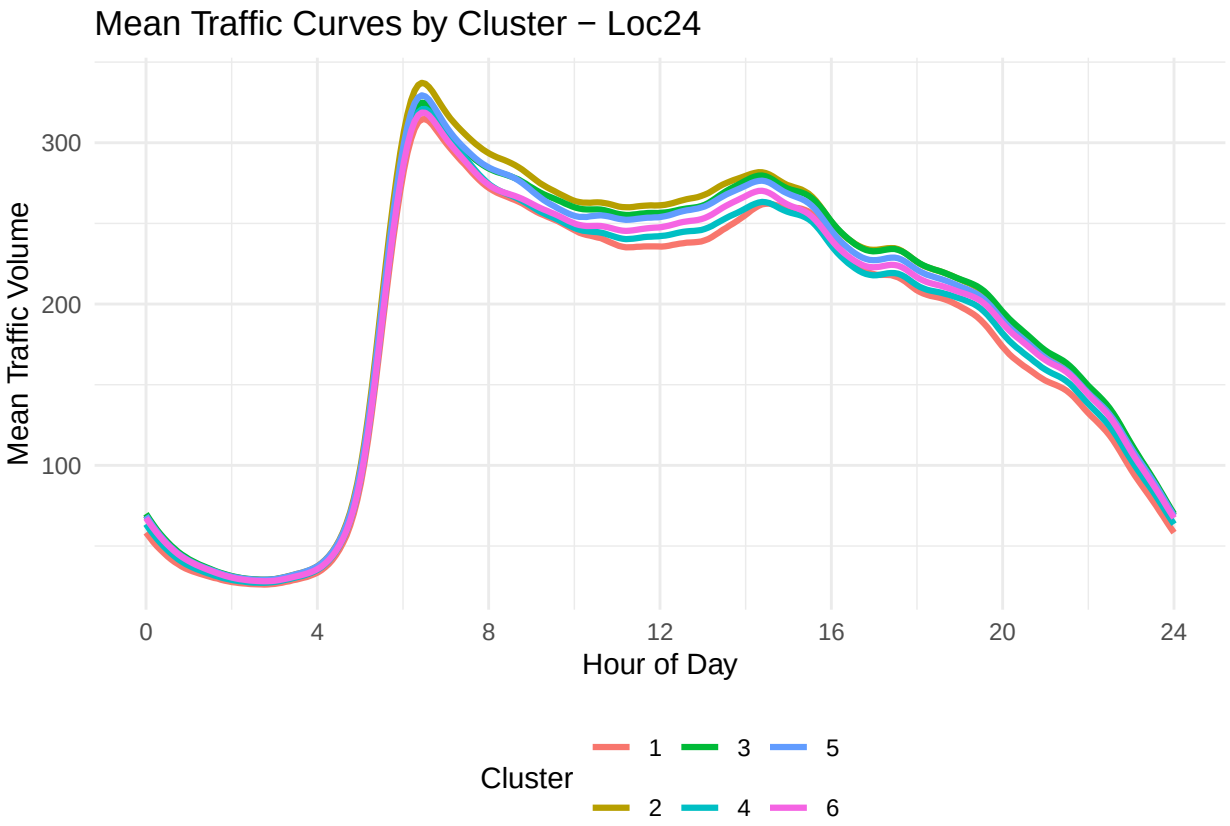

Mean Traffic Curves by Cluster − Loc19
Mean Traffic Volume
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

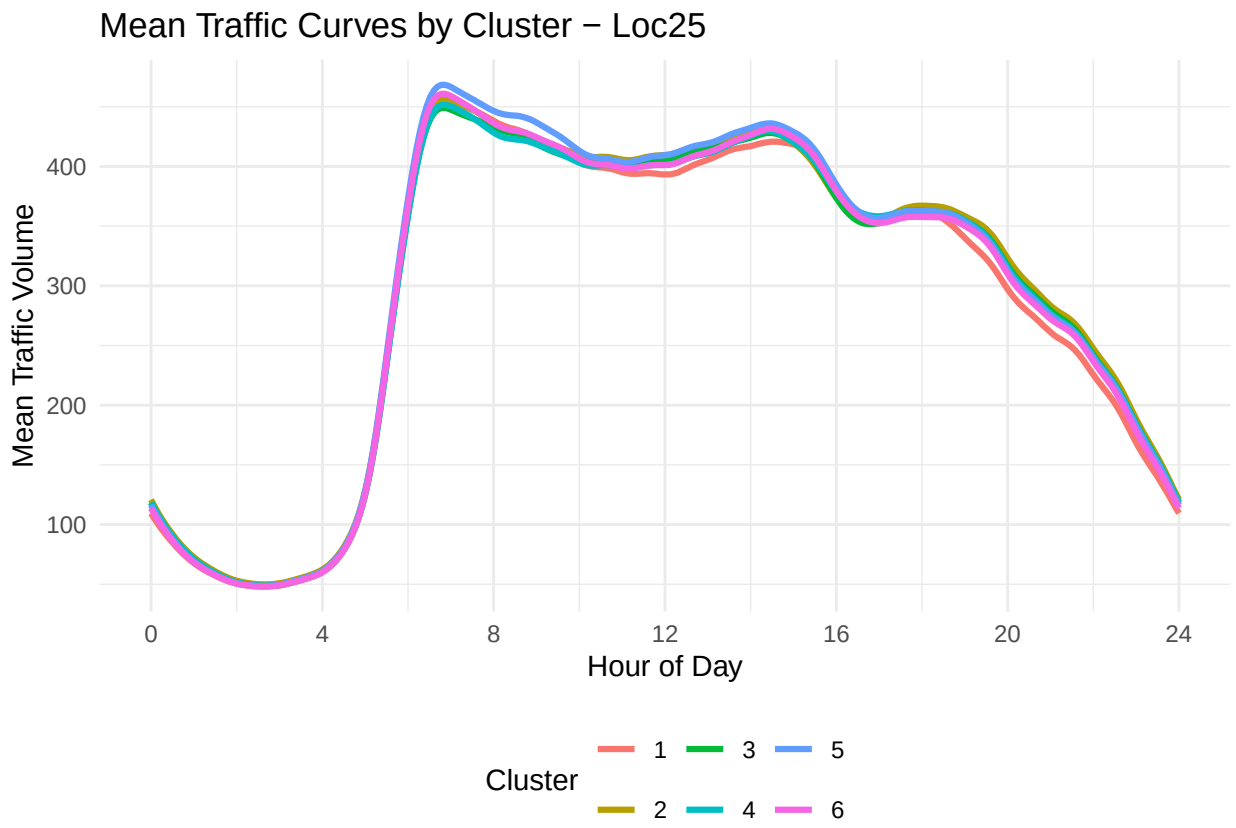

Mean Traffic Curves by Cluster − Loc20
Mean Traffic Volume
600
500
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6

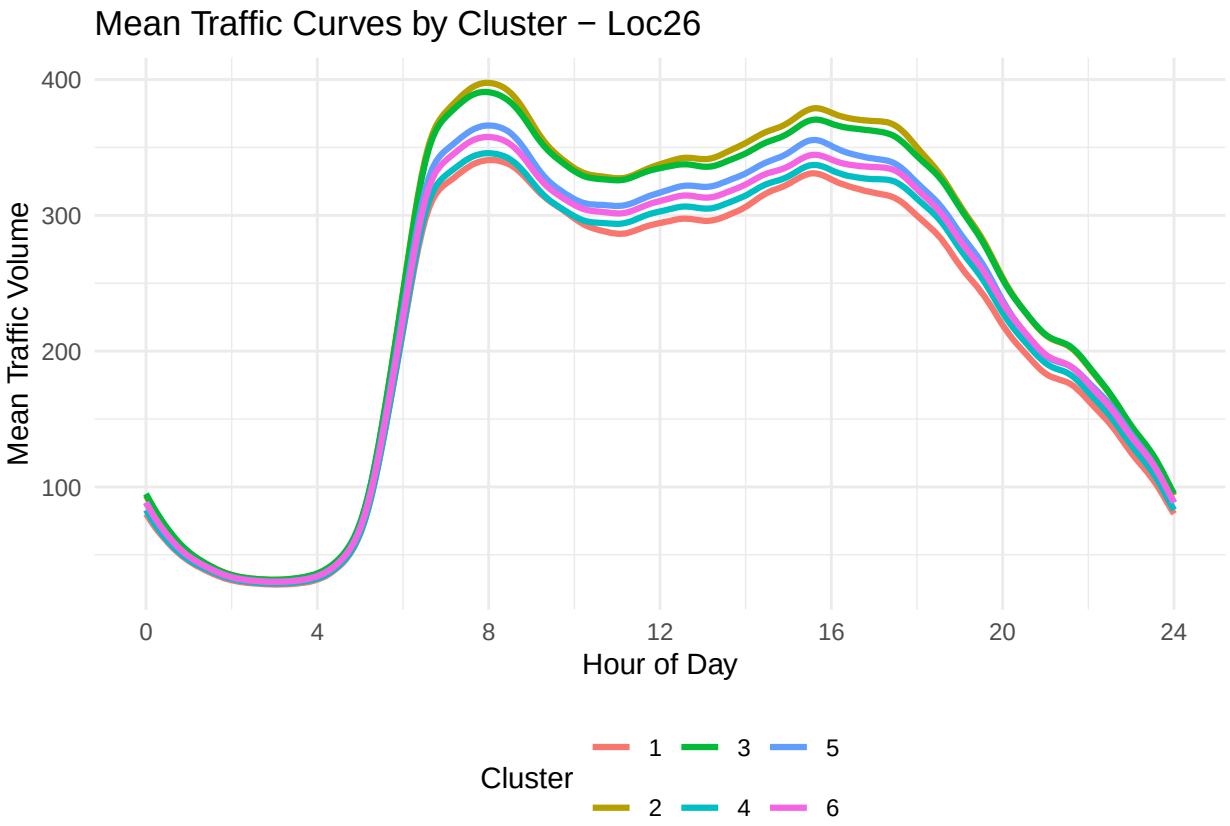

Mean Traffic Curves by Cluster − Loc21
Mean Traffic Volume
500
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6


Mean Traffic Curves by Cluster − Loc22
Mean Traffic Volume
600
400
200
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6


Mean Traffic Curves by Cluster − Loc23
Mean Traffic Volume
500
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6


Mean Traffic Curves by Cluster − Loc24
Mean Traffic Volume
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6


Mean Traffic Curves by Cluster − Loc25
Mean Traffic Volume
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6


Mean Traffic Curves by Cluster − Loc26
Mean Traffic Volume
400
300
200
100
0
4
8
12
16
20
24
Hour of Day
Cluster
1
3
5
2
4
6