Week 2

Pedro Henrique Brant

2025-07-17

```
library(tidyverse)
library(here)
library(janitor)
library(patchwork)
library(rlang)
library(e1071)
```

```
df <- readRDS(here("Output", "merged_dataset_fregni_plus_students.rds")) %>%
  clean_names()
```

```
explore_continuous_var <- function(data, var) {
  var_sym <- ensym(var)
  var_name <- as_string(var_sym)

  # Extract variable and compute basic quantities
  x <- data %>% pull({{ var }})
  x_non_na <- na.omit(x)
  n_total <- length(x)
  n_missing <- sum(is.na(x))
  missing_pct <- round(n_missing / n_total * 100, 2)

  # Compute stats
  stats_list <- list(
    variable = var_name,
    n = n_total,
    n_missing = n_missing,
    missing_pct = missing_pct,
    min = min(x, na.rm = TRUE),
    `25%` = quantile(x, 0.25, na.rm = TRUE),
    median = median(x, na.rm = TRUE),
    `75%` = quantile(x, 0.75, na.rm = TRUE),
    max = max(x, na.rm = TRUE),
    IQR = IQR(x, na.rm = TRUE),
    mean = mean(x, na.rm = TRUE),
    SD = sd(x, na.rm = TRUE),
    skewness = skewness(x_non_na, na.rm = TRUE),
    kurtosis = kurtosis(x_non_na, na.rm = TRUE)
  )

  # Convert to tibble with two columns
  summary_tbl <- tibble(
```

```
    statistic = names(stats_list),
    value = unlist(stats_list)
  )

  # Plotting tibble
  df_plot <- tibble(x = x)

  # Plot 1: Density
  density_plot <- ggplot(df_plot, aes(x = x)) +
    geom_density(fill = "lightblue", alpha = 0.5, na.rm = TRUE) +
    geom_vline(xintercept = median(x_non_na), color = "blue", linetype = "dashed") +
    geom_vline(xintercept = mean(x_non_na), color = "red", linetype = "dashed") +
    annotate("text", x = median(x_non_na), y = Inf,
             label = paste0("Median: ", round(median(x_non_na), 2)),
             vjust = -0.5, hjust = 0, color = "blue", size = 3) +
    annotate("text", x = mean(x_non_na), y = Inf,
             label = paste0("Mean: ", round(mean(x_non_na), 2)),
             vjust = -1.5, hjust = 0, color = "red", size = 3) +
    labs(title = "Density Plot", x = var_name, y = "Density") +
    theme_minimal()

  # Plot 2: Histogram
  hist_plot <- ggplot(df_plot, aes(x = x)) +
    geom_histogram(bins = 30, fill = "gray80", color = "black", na.rm = TRUE) +
    labs(title = "Histogram", x = var_name, y = "Count") +
    theme_minimal()

  # Plot 3: Q-Q
  qq_plot <- ggplot(tibble(x = x_non_na), aes(sample = x)) +
    stat_qq() +
    stat_qq_line(color = "red") +
    labs(
      title = "Q-Q Plot",
      subtitle = paste("n =", length(x_non_na)),
      x = "Theoretical Quantiles",
      y = "Sample Quantiles"
    ) +
    theme_minimal()

  # Plot 4: Boxplot
  box_plot <- ggplot(df_plot, aes(y = x)) +
    geom_boxplot(fill = "lightgreen", na.rm = TRUE) +
    coord_flip() +
    labs(title = "Boxplot", y = var_name) +
    theme_minimal()

  # Combine all 4 plots
  combined_plot <- (density_plot | hist_plot) / (qq_plot | box_plot)

  return(list(summary = summary_tbl, plot = combined_plot))
}
```

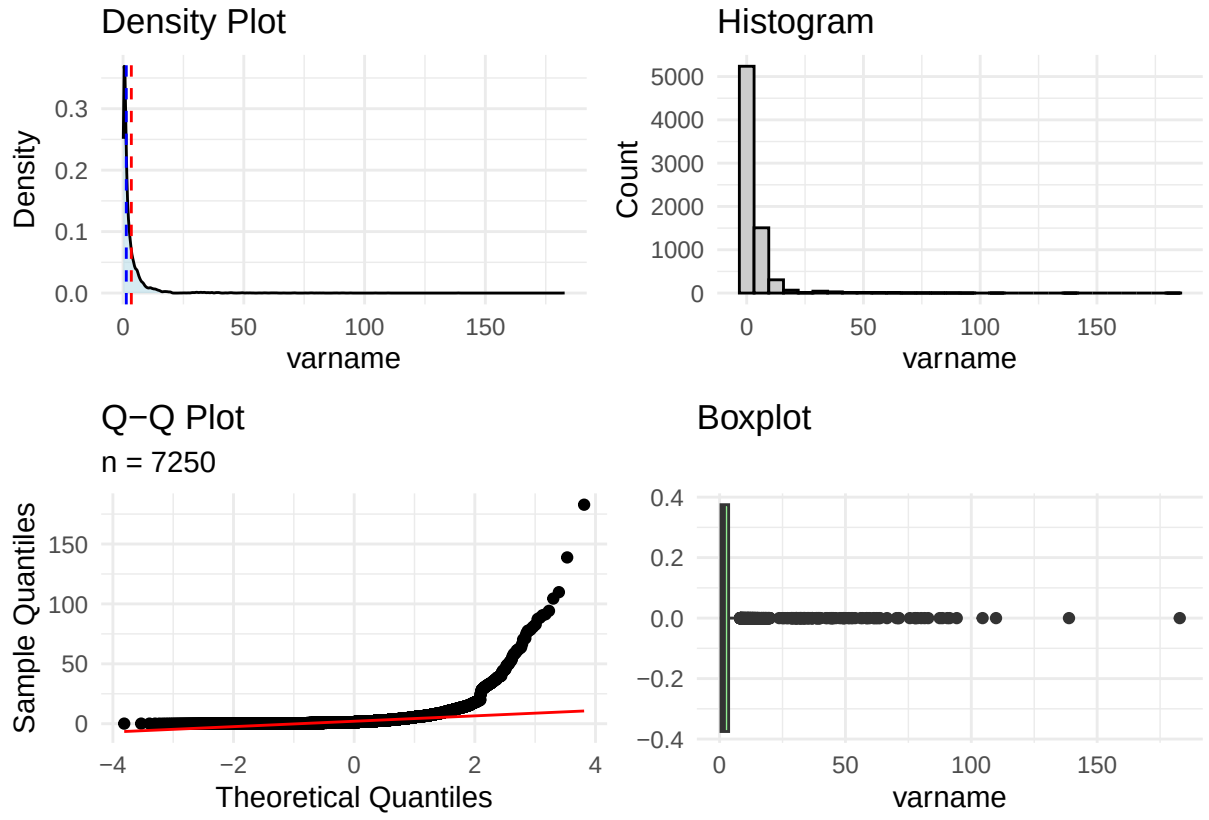
```r
biomarker_vars <- c(
  hs_crp    = "lbxhscrp",
  insulin   = "lbxin",
  glucose   = "lbxglu",
  ferritin  = "lbxfer",
  tg        = "lbxtr",
  hdl       = "lbdhdd",
  ldl       = "lbdldl",
  total_chol = "lbxtc",
  hba1c     = "lbxgh",
  neutrophil_count = "lbdneno",
  lymphocyte_count = "lbdlymno"
)


results <- imap(biomarker_vars, function(varname, label) {
  cat("Exploring", label, "(", varname, ")...\n")
  result <- explore_continuous_var(df, varname)
  list(name = label, varname = varname, result = result)
})
```

```
## Exploring hs_crp ( lbxhscrp )...
## Exploring insulin ( lbxin )...
## Exploring glucose ( lbxglu )...
## Exploring ferritin ( lbxfer )...
## Exploring tg ( lbxtr )...
## Exploring hdl ( lbdhdd )...
## Exploring ldl ( lbdldl )...
## Exploring total_chol ( lbxtc )...
## Exploring hba1c ( lbxgh )...
## Exploring neutrophil_count ( lbdneno )...
## Exploring lymphocyte_count ( lbdlymno )...
```

```r
results
```

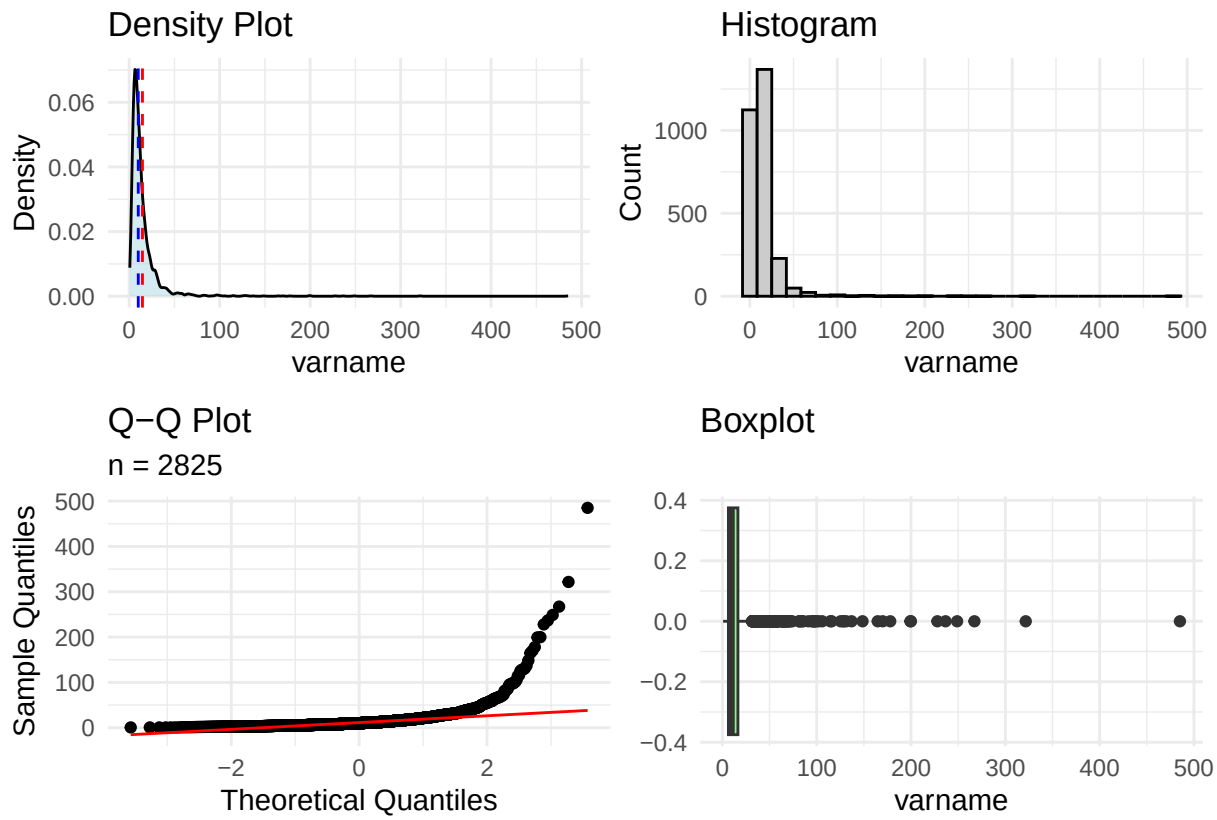
```
## $hs_crp
## $hs_crp$name
## [1] "hs_crp"
##
## $hs_crp$varname
## [1] "lbxhscrp"
##
## $hs_crp$result
## $hs_crp$result$summary
## # A tibble: 14 x 2
##    statistic   value
##    <chr>       <chr>
##  1 variable    varname
##  2 n           9254
##  3 n_missing   2004
##  4 missing_pct 21.66
##  5 min         0.11
##  6 25%         0.56
```

```
##  7 median      1.355
##  8 75%         3.59
##  9 max         182.82
## 10 IQR         3.03
## 11 mean        3.43972
## 12 SD          7.41174192797466
## 13 skewness    8.3309143833476
## 14 kurtosis    110.413930230871
##
## $hs_crp$result$plot
```



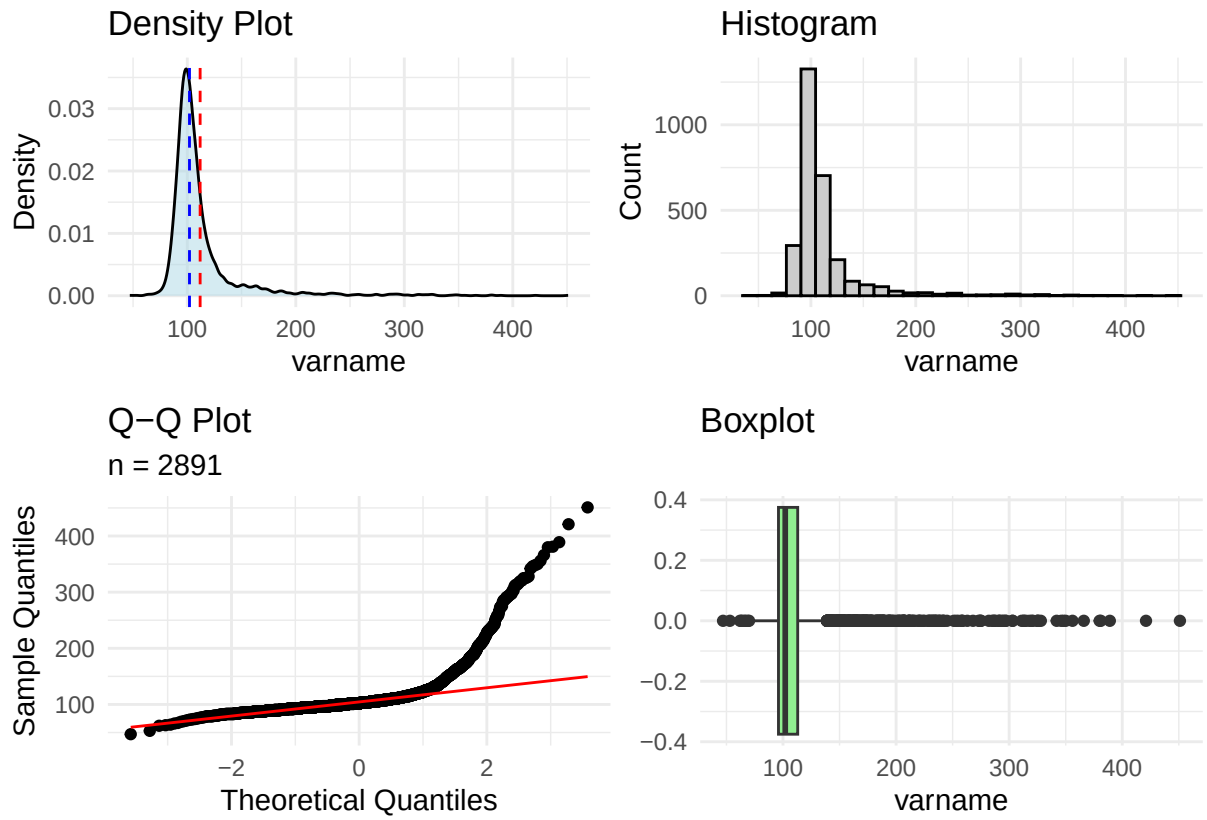

```
##
##
##
## $insulin
## $insulin$name
## [1] "insulin"
##
## $insulin$varname
## [1] "lbxin"
##
## $insulin$result
## $insulin$result$summary
## # A tibble: 14 x 2
##    statistic   value
```

```
##    <chr>       <chr>
##  1 variable    varname
##  2 n           9254
##  3 n_missing   6429
##  4 missing_pct 69.47
##  5 min         0.71
##  6 25%         6.38
##  7 median      10.04
##  8 75%         16.47
##  9 max         485.1
## 10 IQR         10.09
## 11 mean        14.6706619469027
## 12 SD          20.3753710567905
## 13 skewness    9.68211487824672
## 14 kurtosis    152.698462294906
##
## $insulin$result$plot
```



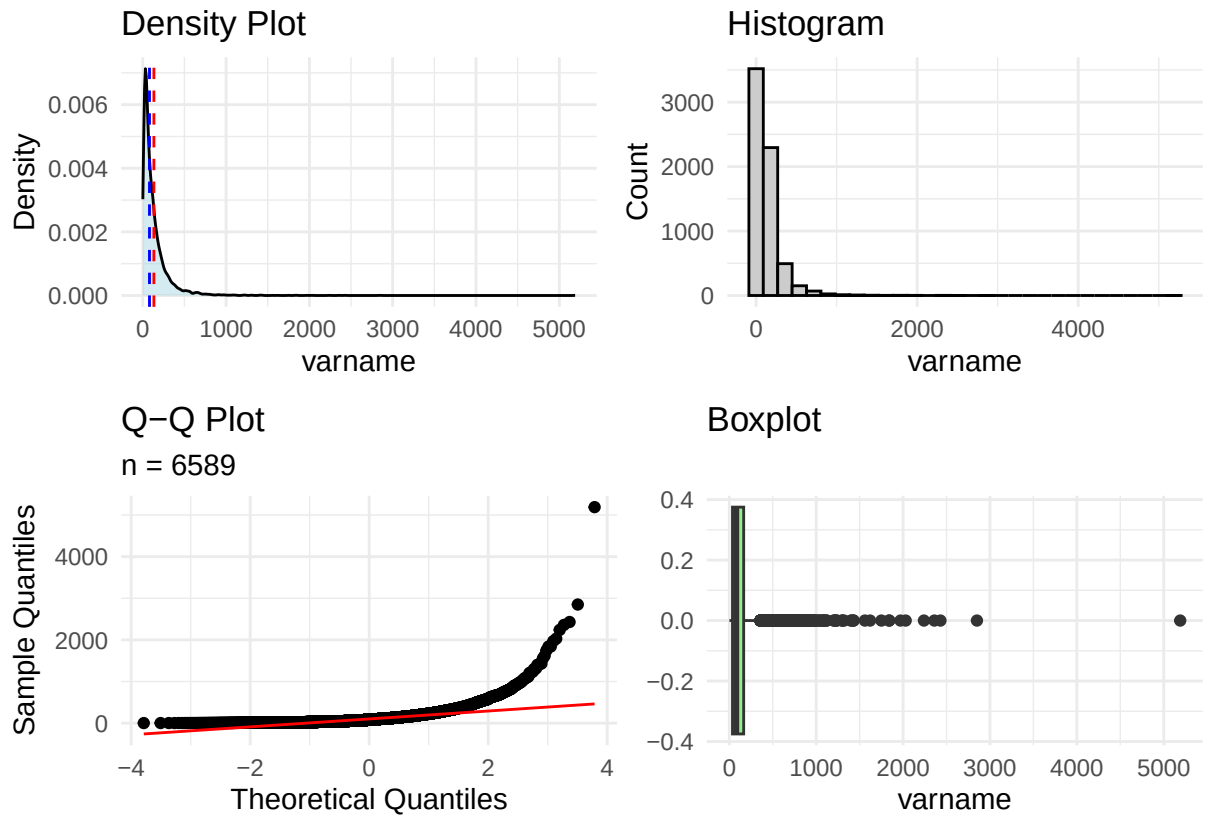

```
##
##
##
## $glucose
## $glucose$name
## [1] "glucose"
##
```

```
## $glucose$varname
## [1] "lbxglu"
##
## $glucose$result
## $glucose$result$summary
## # A tibble: 14 x 2
##    statistic   value
##    <chr>       <chr>
##  1 variable    varname
##  2 n           9254
##  3 n_missing   6363
##  4 missing_pct 68.76
##  5 min         47
##  6 25%         96
##  7 median      102
##  8 75%         113
##  9 max         451
## 10 IQR         17
## 11 mean        111.803182289865
## 12 SD          35.5331465251848
## 13 skewness    4.00089794296835
## 14 kurtosis    20.8627939578962
##
## $glucose$result$plot
```



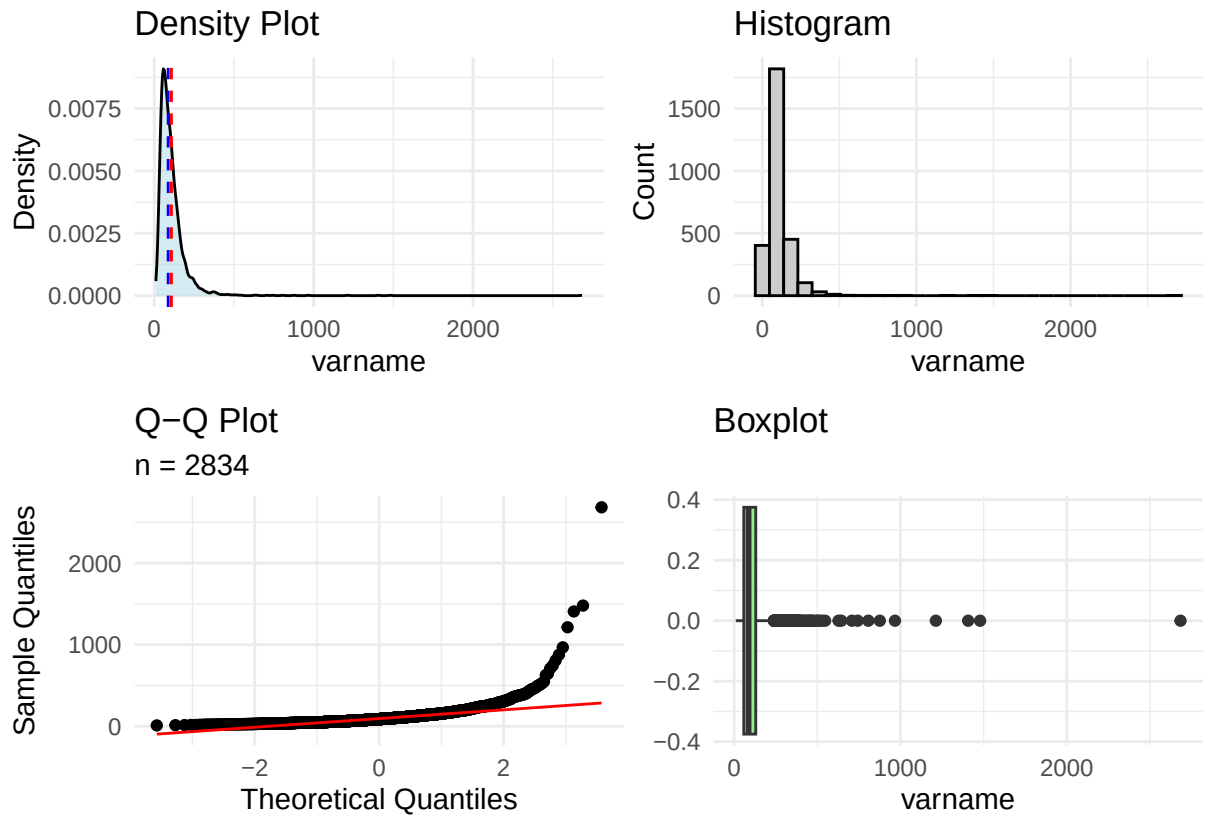

```
##
```

```
##
##
## $ferritin
## $ferritin$name
## [1] "ferritin"
##
## $ferritin$varname
## [1] "lbxfer"
##
## $ferritin$result
## $ferritin$result$summary
## # A tibble: 14 x 2
##    statistic   value
##    <chr>       <chr>
##  1 variable    varname
##  2 n           9254
##  3 n_missing   2665
##  4 missing_pct 28.8
##  5 min         1.04
##  6 25%         36.6
##  7 median      80.7
##  8 75%         165
##  9 max         5190
## 10 IQR         128.4
## 11 mean        133.394927910153
## 12 SD          180.307802965244
## 13 skewness    7.25384245546476
## 14 kurtosis    124.979711257697
##
## $ferritin$result$plot
```



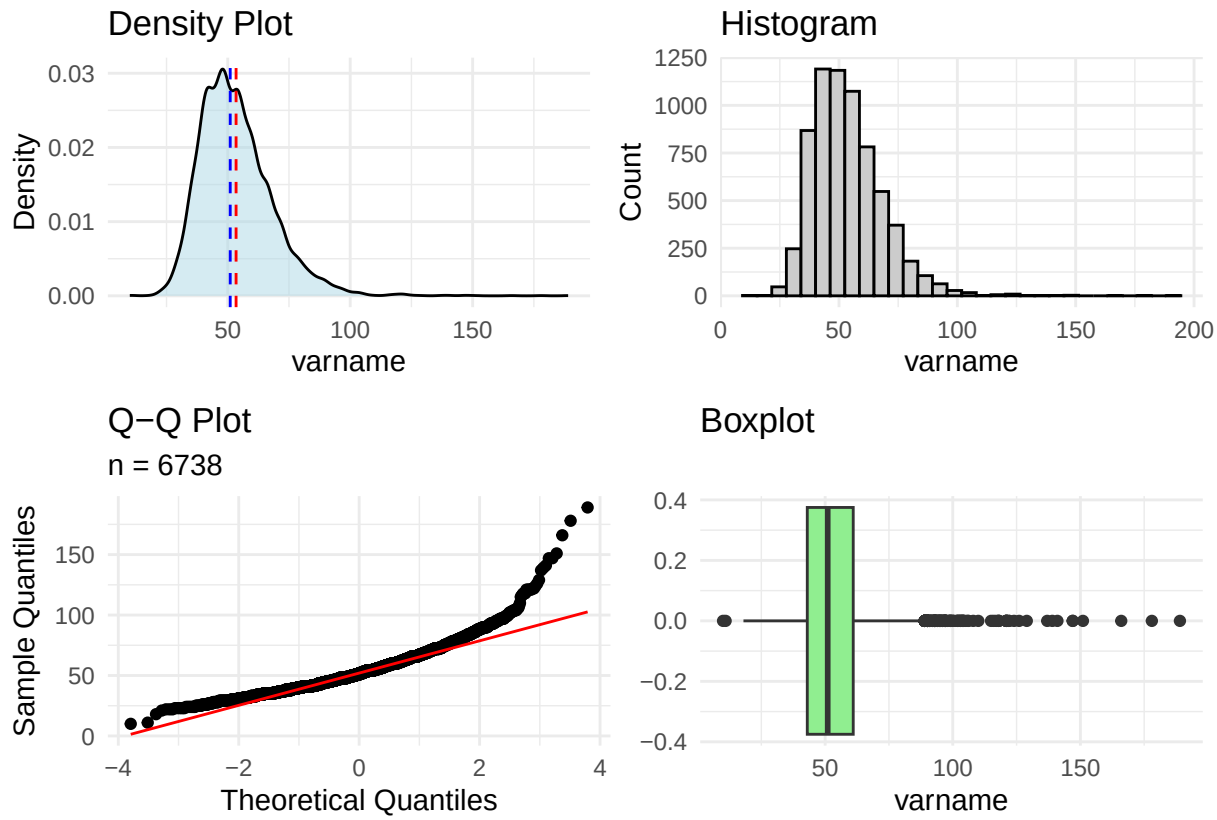

```
##
##
##
## $tg
## $tg$name
## [1] "tg"
##
## $tg$varname
## [1] "lbxtr"
##
## $tg$result
## $tg$result$summary
## # A tibble: 14 x 2
##    statistic   value
##    <chr>       <chr>
##  1 variable    varname
##  2 n           9254
##  3 n_missing   6420
##  4 missing_pct 69.38
##  5 min         10
##  6 25%         58
##  7 median      87
##  8 75%         130
##  9 max         2684
## 10 IQR         72
## 11 mean        107.344389555399
```

```
## 12 SD          98.2648543532061
## 13 skewness    10.0532032949495
## 14 kurtosis    201.579399375063
##
## $tg$result$plot
```



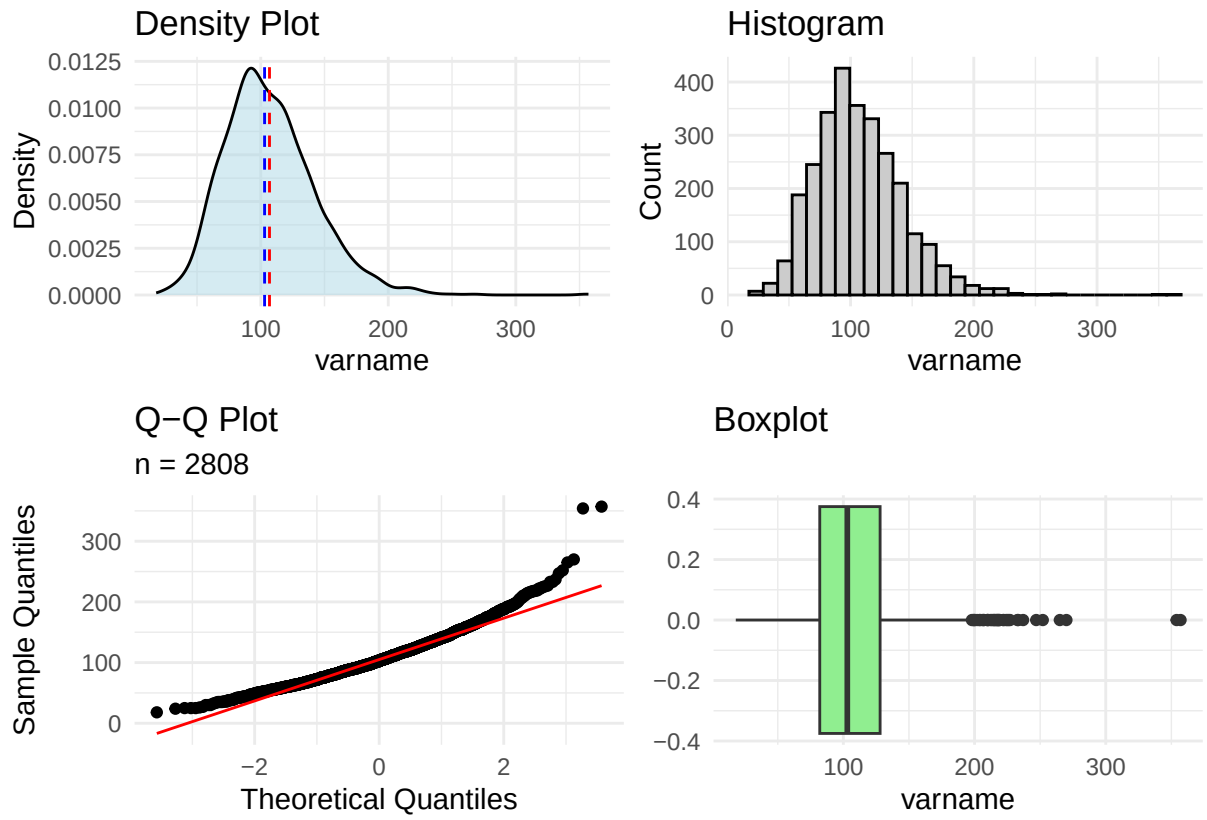

```
##
##
##
## $hdl
## $hdl$name
## [1] "hdl"
##
## $hdl$varname
## [1] "lbdhdd"
##
## $hdl$result
## $hdl$result$summary
## # A tibble: 14 x 2
##    statistic   value
##    <chr>       <chr>
##  1 variable    varname
##  2 n           9254
##  3 n_missing   2516
##  4 missing_pct 27.19
```

```
##  5 min          10
##  6 25%          43
##  7 median       51
##  8 75%          61
##  9 max          189
## 10 IQR          18
## 11 mean         53.3925497180172
## 12 SD           14.7458439839249
## 13 skewness     1.22660698968594
## 14 kurtosis     4.34634439458705
##
## $hdl$result$plot
```



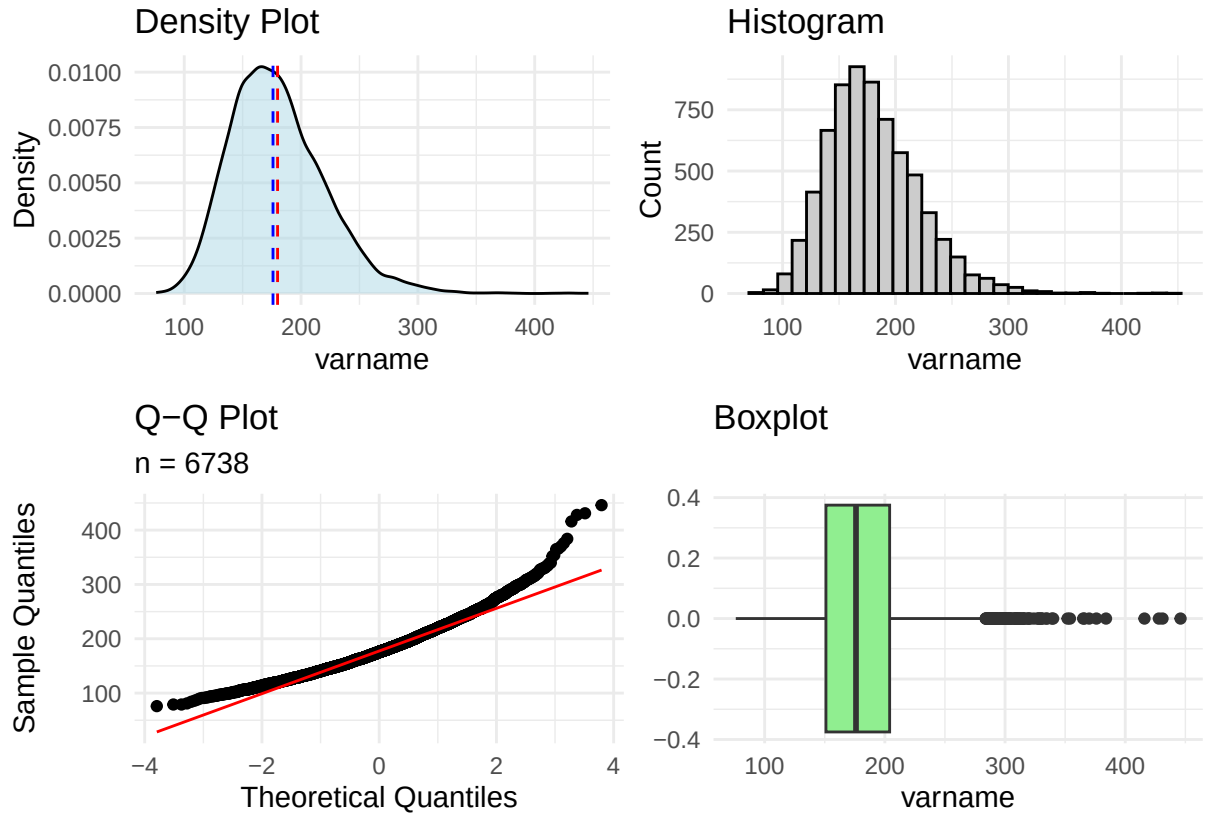

```
##
##
##
## $ldl
## $ldl$name
## [1] "ldl"
##
## $ldl$varname
## [1] "lbdldl"
##
## $ldl$result
## $ldl$result$summary
```

```
## # A tibble: 14 x 2
##    statistic   value
##    <chr>       <chr>
##  1 variable    varname
##  2 n           9254
##  3 n_missing   6446
##  4 missing_pct 69.66
##  5 min         18
##  6 25%         82
##  7 median      103
##  8 75%         128
##  9 max         357
## 10 IQR         46
## 11 mean        106.85292022792
## 12 SD          35.5860410625961
## 13 skewness    0.833083043467597
## 14 kurtosis    2.1242030510364
##
## $ldl$result$plot
```

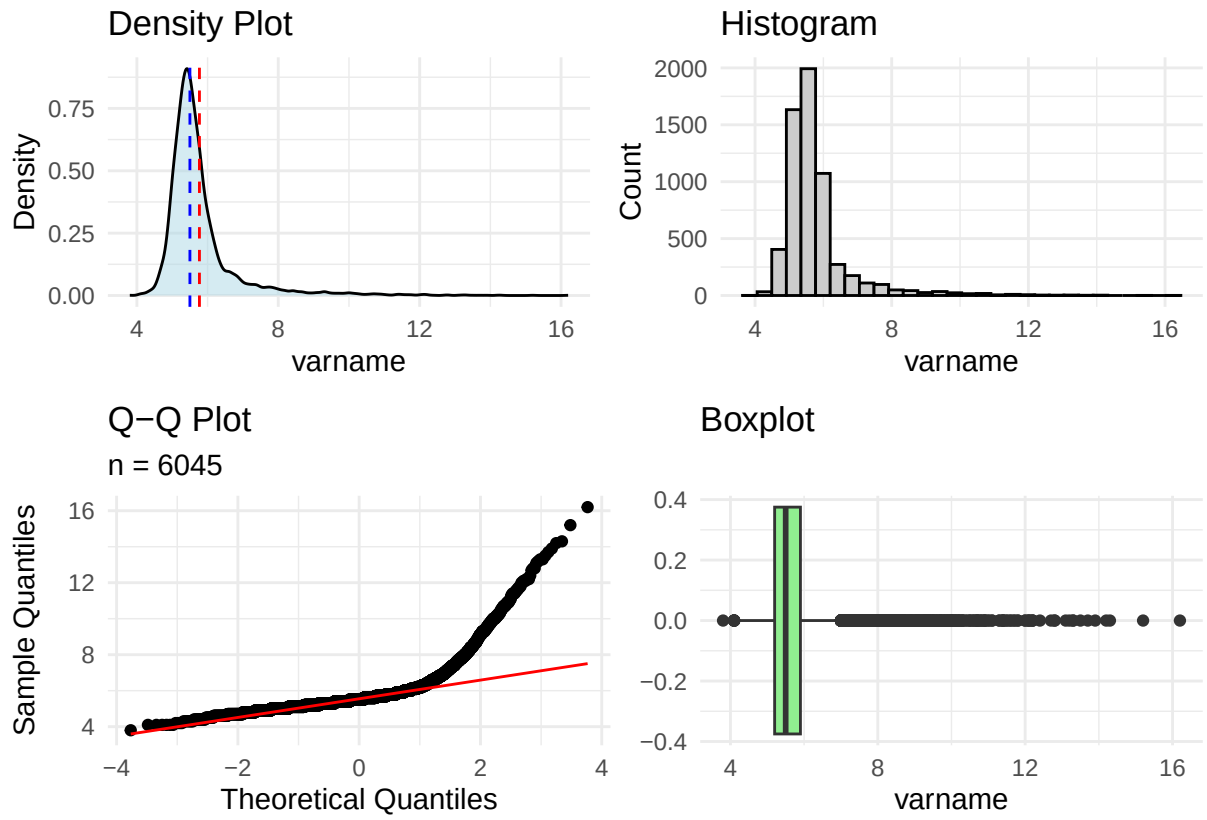



```
##
##
##
## $total_chol
## $total_chol$name
```

```
## [1] "total_chol"
##
## $total_chol$varname
## [1] "lbxtc"
##
## $total_chol$result
## $total_chol$result$summary
## # A tibble: 14 x 2
##    statistic   value
##    <chr>       <chr>
##  1 variable    varname
##  2 n           9254
##  3 n_missing   2516
##  4 missing_pct 27.19
##  5 min         76
##  6 25%         151
##  7 median      176
##  8 75%         204
##  9 max         446
## 10 IQR         53
## 11 mean        179.894627485901
## 12 SD          40.6022481382962
## 13 skewness    0.778016325190274
## 14 kurtosis    1.43276278678026
##
## $total_chol$result$plot
```

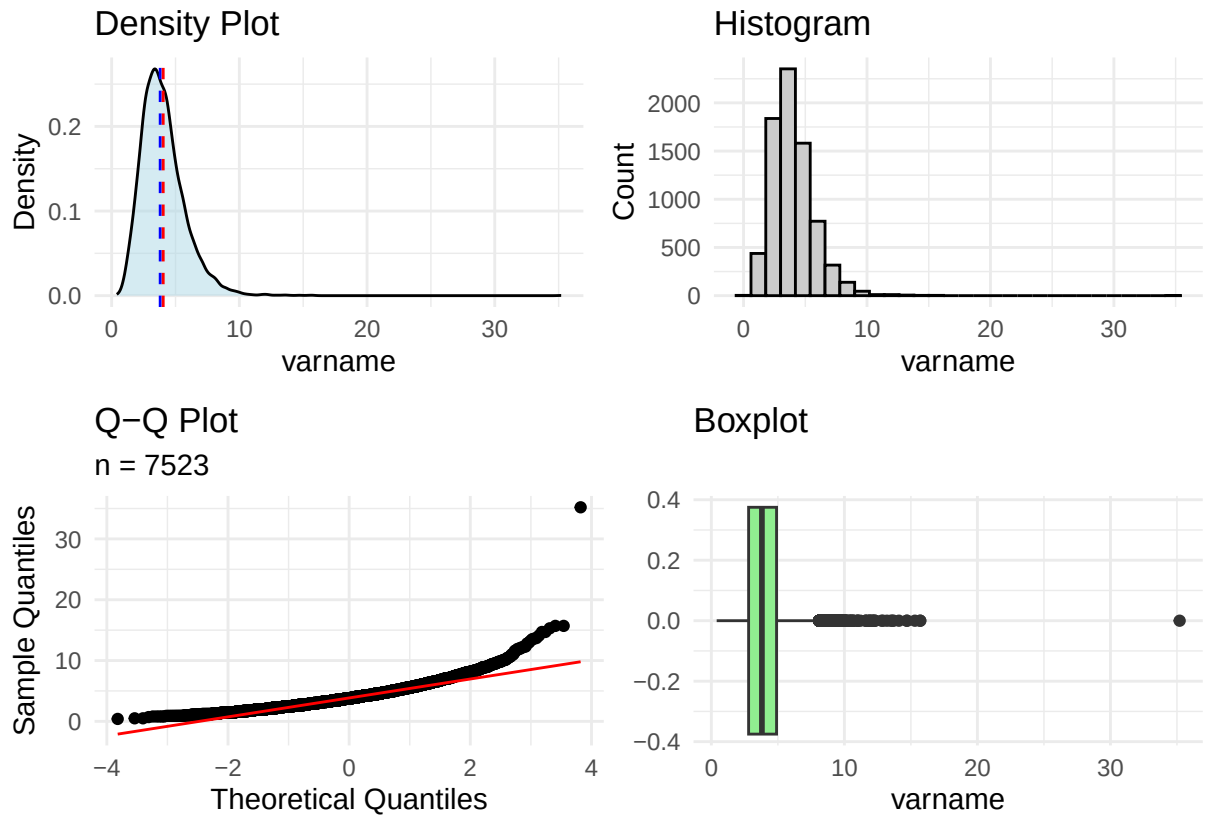


```
##
##
##
## $hba1c
## $hba1c$name
## [1] "hba1c"
##
## $hba1c$varname
## [1] "lbxgh"
##
## $hba1c$result
## $hba1c$result$summary
## # A tibble: 14 x 2
##    statistic   value
##    <chr>       <chr>
##  1 variable    varname
##  2 n           9254
##  3 n_missing   3209
##  4 missing_pct 34.68
##  5 min         3.8
##  6 25%         5.2
##  7 median      5.5
##  8 75%         5.9
##  9 max         16.2
## 10 IQR         0.7
## 11 mean        5.76956162117452
## 12 SD          1.03783802498172
## 13 skewness    3.36637346414582
## 14 kurtosis    16.2147829767829
##
## $hba1c$result$plot
```




```
##
##
##
## $neutrophil_count
## $neutrophil_count$name
## [1] "neutrophil_count"
##
## $neutrophil_count$varname
## [1] "lbdneno"
##
## $neutrophil_count$result
## $neutrophil_count$result$summary
## # A tibble: 14 x 2
##    statistic   value
##    <chr>       <chr>
##  1 variable    varname
##  2 n           9254
##  3 n_missing   1731
##  4 missing_pct 18.71
##  5 min         0.4
##  6 25%         2.8
##  7 median      3.8
##  8 75%         4.9
##  9 max         35.2
## 10 IQR         2.1
## 11 mean        4.03464043599628
```

```
## 12 SD          1.73222705478062
## 13 skewness    1.87383806381105
## 14 kurtosis    16.4624777861025
##
## $neutrophil_count$result$plot
```




```
##
##
##
## $lymphocyte_count
## $lymphocyte_count$name
## [1] "lymphocyte_count"
##
## $lymphocyte_count$varname
## [1] "lbdlymno"
##
## $lymphocyte_count$result
## $lymphocyte_count$result$summary
## # A tibble: 14 x 2
##    statistic   value
##    <chr>       <chr>
##  1 variable    varname
##  2 n           9254
##  3 n_missing   1731
##  4 missing_pct 18.71
```

```
##  5 min          0.4
##  6 25%          1.8
##  7 median       2.3
##  8 75%          2.9
##  9 max          358.8
## 10 IQR          1.1
## 11 mean         2.50159510833444
## 12 SD           4.3164997104012
## 13 skewness     75.3088413408158
## 14 kurtosis     6174.81974438113
##
## $lymphocyte_count$result$plot
```