EDA

Paul K. Yu

2024-09-23

Load libraries

```r
library(ggplot2)
library(tidyverse)
```

We want to do an exploratory data analysis on the damage data.

```r
head(metadata_prep)
```

```
##                                   X.SampleID  Date_100  Date_200  Date_300
## 1 16925_PostMedieval_ChelseaOldChurch_OCU00 1800_1900 1800_2000 1600_1900
## 2       16969_Medieval_BermondseyAbbey_BA84      <NA>      <NA>      <NA>
## 3         16937_Medieval_StMaryGraces_MIN86 1400_1500 1400_1600 1300_1600
## 4         16939_Medieval_StMaryGraces_MIN86 1400_1500 1400_1600 1300_1600
## 5         16948_Medieval_StMaryGraces_MIN86 1300_1400 1200_1400 1300_1600
## 6       16900_PostMedieval_CrossBones_REW92      <NA>      <NA> 1600_1900
##   BlackDeath_PrePost EarlyDate LateDate MedievalPostMedieval          Cemetry
## 1               Post      1836     <NA>         PostMedieval ChelseaOldChurch
## 2             Across      1066     1540             Medieval  BermondseyAbbey
## 3               Post      1400     1538             Medieval     StMaryGraces
## 4               Post      1400     1538             Medieval     StMaryGraces
## 5               Post      1350     1400             Medieval     StMaryGraces
## 6               Post      1598     1853         PostMedieval       CrossBones
##   MaxillaMandible BuccalLingual SubSupragingival   Tooth Tooth_Simplified
## 1        Mandible Interproximal            Supra  Canine           Canine
## 2        Mandible       Lingual            Supra   Molar            Molar
## 3        Mandible       Lingual            Supra   Molar            Molar
## 4        Mandible Interproximal            Supra   Molar            Molar
## 5        Mandible       Lingual            Supra   Molar            Molar
## 6        Mandible Interproximal            Supra Incisor          Incisor
##   BlackDeath_1346_1353 DeltaD_mean DeltaD_mean_methano DeltaD_mean_por
## 1                 Post  0.03159436         0.033232946     0.007384087
## 2               Across  0.04112947         0.056094304     0.061636551
## 3                 Post  0.03960263         0.054139003     0.053491125
## 4                 Post  0.04259754         0.056497442     0.067990389
## 5               Across  0.04002394         0.009694423     0.052791102
## 6                 Post  0.01938508         0.016877111     0.002777638
##   DeltaD_mean_strep
## 1       0.006438698
## 2       0.002875072
## 3       0.038610533
## 4       0.001558243
## 5       0.003304606
## 6       0.003928007
```

Structure of the data

```r
str(metadata_prep)
```

```
## 'data.frame':    126 obs. of  19 variables:
##  $ X.SampleID          : chr  "16925_PostMedieval_ChelseaOldChurch_OCU00" "16969_Medieval_BermondseyA
##  $ Date_100            : chr  "1800_1900" NA "1400_1500" "1400_1500" ...
##  $ Date_200            : chr  "1800_2000" NA "1400_1600" "1400_1600" ...
##  $ Date_300            : chr  "1600_1900" NA "1300_1600" "1300_1600" ...
##  $ BlackDeath_PrePost  : chr  "Post" "Across" "Post" "Post" ...
##  $ EarlyDate           : chr  "1836" "1066" "1400" "1400" ...
##  $ LateDate            : chr  NA "1540" "1538" "1538" ...
##  $ MedievalPostMedieval: chr  "PostMedieval" "Medieval" "Medieval" "Medieval" ...
##  $ Cemetry             : chr  "ChelseaOldChurch" "BermondseyAbbey" "StMaryGraces" "StMaryGraces" ...
##  $ MaxillaMandible     : chr  "Mandible" "Mandible" "Mandible" "Mandible" ...
##  $ BuccalLingual       : chr  "Interproximal" "Lingual" "Lingual" "Interproximal" ...
##  $ SubSupragingival    : chr  "Supra" "Supra" "Supra" "Supra" ...
##  $ Tooth               : chr  "Canine" "Molar" "Molar" "Molar" ...
##  $ Tooth_Simplified    : chr  "Canine" "Molar" "Molar" "Molar" ...
##  $ BlackDeath_1346_1353: chr  "Post" "Across" "Post" "Post" ...
##  $ DeltaD_mean         : num  0.0316 0.0411 0.0396 0.0426 0.04 ...
##  $ DeltaD_mean_methano : num  0.03323 0.05609 0.05414 0.0565 0.00969 ...
##  $ DeltaD_mean_por     : num  0.00738 0.06164 0.05349 0.06799 0.05279 ...
##  $ DeltaD_mean_strep   : num  0.00644 0.00288 0.03861 0.00156 0.0033 ...
```

Summary of the data

```r
summary(metadata_prep)
```

```
##   X.SampleID          Date_100           Date_200           Date_300
##  Length:126         Length:126         Length:126         Length:126
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  BlackDeath_PrePost  EarlyDate           LateDate         MedievalPostMedieval
##  Length:126         Length:126         Length:126         Length:126
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    Cemetry          MaxillaMandible    BuccalLingual      SubSupragingival
##  Length:126         Length:126         Length:126         Length:126
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
```

```
##
##
##     Tooth           Tooth_Simplified   BlackDeath_1346_1353  DeltaD_mean
##  Length:126         Length:126         Length:126           Min.   :0.001263
##  Class :character   Class :character   Class :character     1st Qu.:0.025439
##  Mode  :character   Mode  :character   Mode  :character     Median :0.032291
##                                                             Mean   :0.032399
##                                                             3rd Qu.:0.039823
##                                                             Max.   :0.051045
##                                                             NA's   :5
##  DeltaD_mean_methano DeltaD_mean_por    DeltaD_mean_strep
##  Min.   :0.002018    Min.   :0.000574   Min.   :0.001347
##  1st Qu.:0.015111    1st Qu.:0.012232   1st Qu.:0.006117
##  Median :0.036106    Median :0.054815   Median :0.039205
##  Mean   :0.034821    Mean   :0.045910   Mean   :0.034599
##  3rd Qu.:0.048835    3rd Qu.:0.067990   3rd Qu.:0.054557
##  Max.   :0.213169    Max.   :0.095906   Max.   :0.076097
##  NA's   :11          NA's   :5          NA's   :6
```

From the mapDamage website

DeltaD, the cytosine deamination probability in double strand context.

DeltaS, the cytosine deamination probability in single strand context.

```r
DeltaD_mean <- metadata_prep$DeltaD_mean
DeltaD_mean_methano <- metadata_prep$DeltaD_mean_methano
DeltaD_mean_por <- metadata_prep$DeltaD_mean_por
DeltaD_mean_strep <- metadata_prep$DeltaD_mean_strep

# For DeltaD_mean:
# Get NA indices
na_indices <- which(is.na(DeltaD_mean))

# Calculate mean
mean_value <- mean(DeltaD_mean, na.rm = TRUE)

# Impute NA values with the calculated mean
DeltaD_mean[na_indices] <- mean_value

# Recheck NA (should be zero)
sum(is.na(DeltaD_mean))
```

```
## [1] 0
```

```r
# For DeltaD_mean_methano:
# Get NA indices
na_indices <- which(is.na(DeltaD_mean_methano))

# Calculate mean
mean_value <- mean(DeltaD_mean_methano, na.rm = TRUE)

# Impute NA values with the calculated mean
DeltaD_mean_methano[na_indices] <- mean_value
```

```
# Recheck NA (should be zero)
sum(is.na(DeltaD_mean_methano))
```

```
## [1] 0
```

```
# For DeltaD_mean_por:
# Get NA indices
na_indices <- which(is.na(DeltaD_mean_por))

# Calculate mean
mean_value <- mean(DeltaD_mean_por, na.rm = TRUE)

# Impute NA values with the calculated mean
DeltaD_mean_por[na_indices] <- mean_value

# Recheck NA (should be zero)
sum(is.na(DeltaD_mean_por))
```

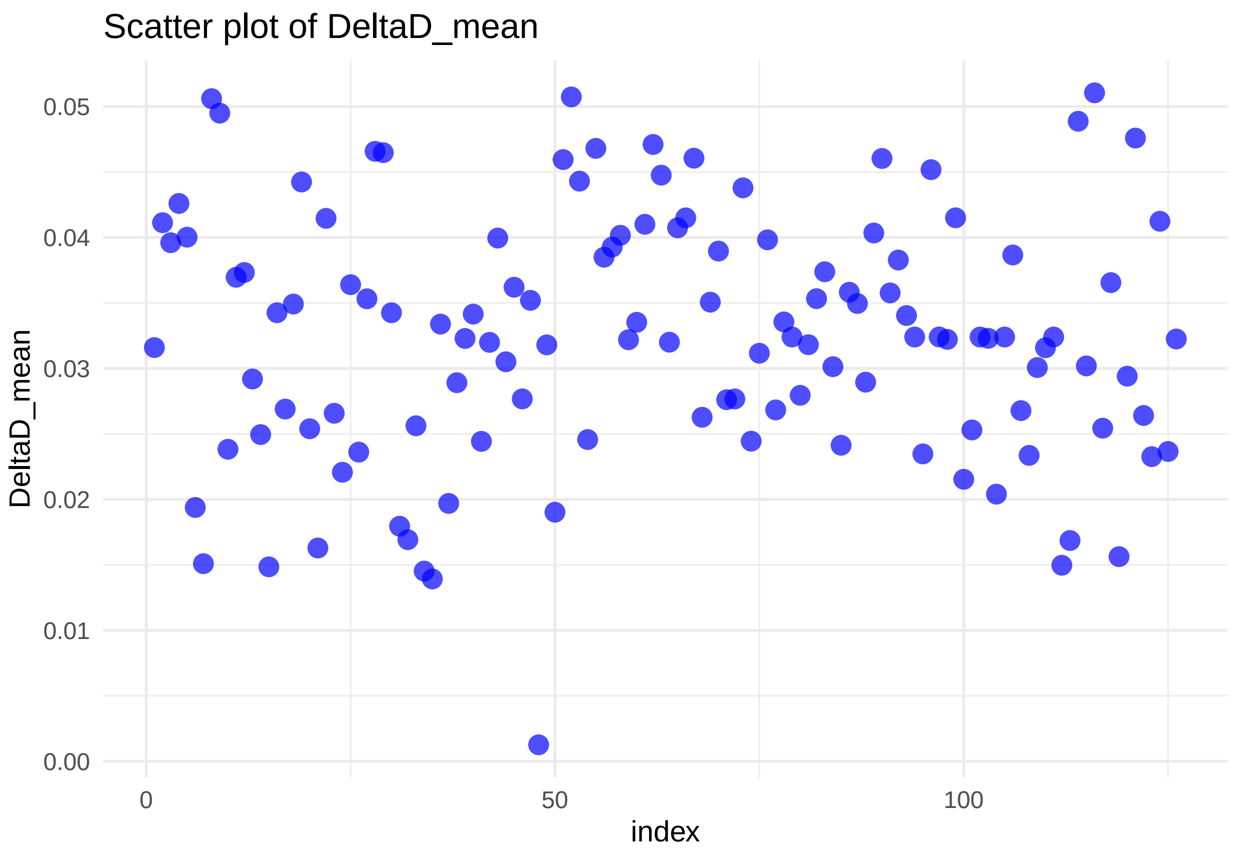
```
## [1] 0
```

```
# For DeltaD_mean_strep:
# Get NA indices
na_indices <- which(is.na(DeltaD_mean_strep))

# Calculate mean
mean_value <- mean(DeltaD_mean_strep, na.rm = TRUE)

# Impute NA values with the calculated mean
DeltaD_mean_strep[na_indices] <- mean_value

# Recheck NA (should be zero)
sum(is.na(DeltaD_mean_strep))
```

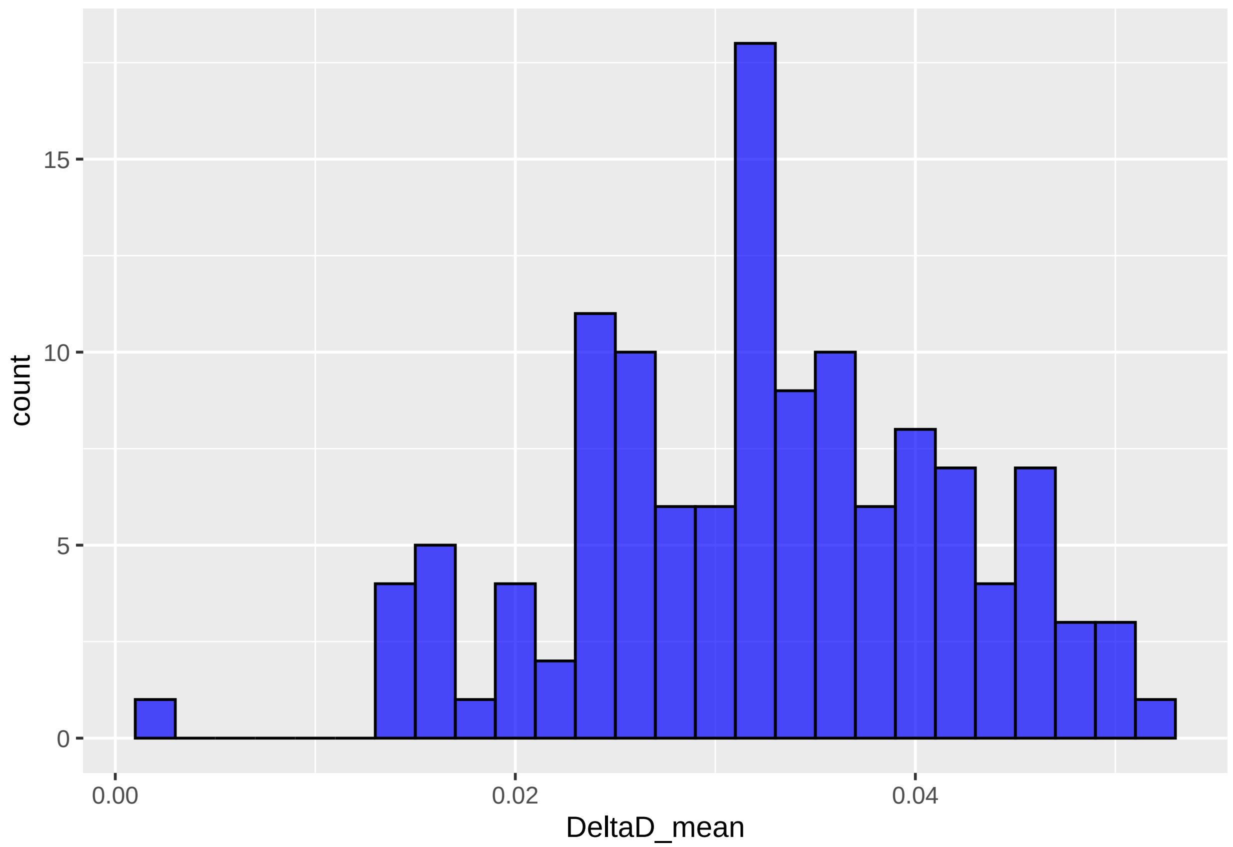
```
## [1] 0
```

Plots for DeltaD_mean

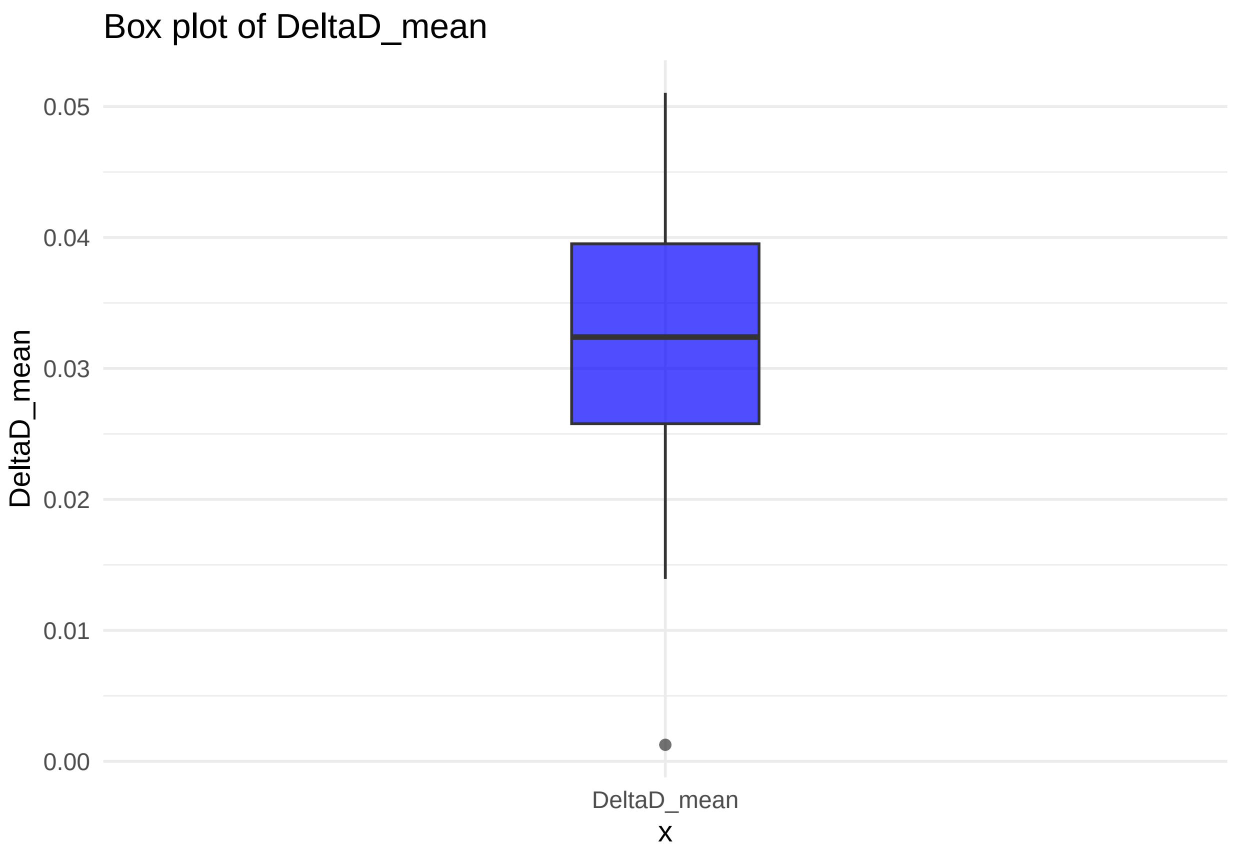
```
index <- which(!is.na(DeltaD_mean))

ggplot(data.frame(DeltaD_mean), aes(x = index, y = DeltaD_mean)) +
    geom_point(color = "blue", size = 3, alpha = 0.7) +
    theme_minimal() +
    ggtitle(paste("Scatter plot of DeltaD_mean"))
```



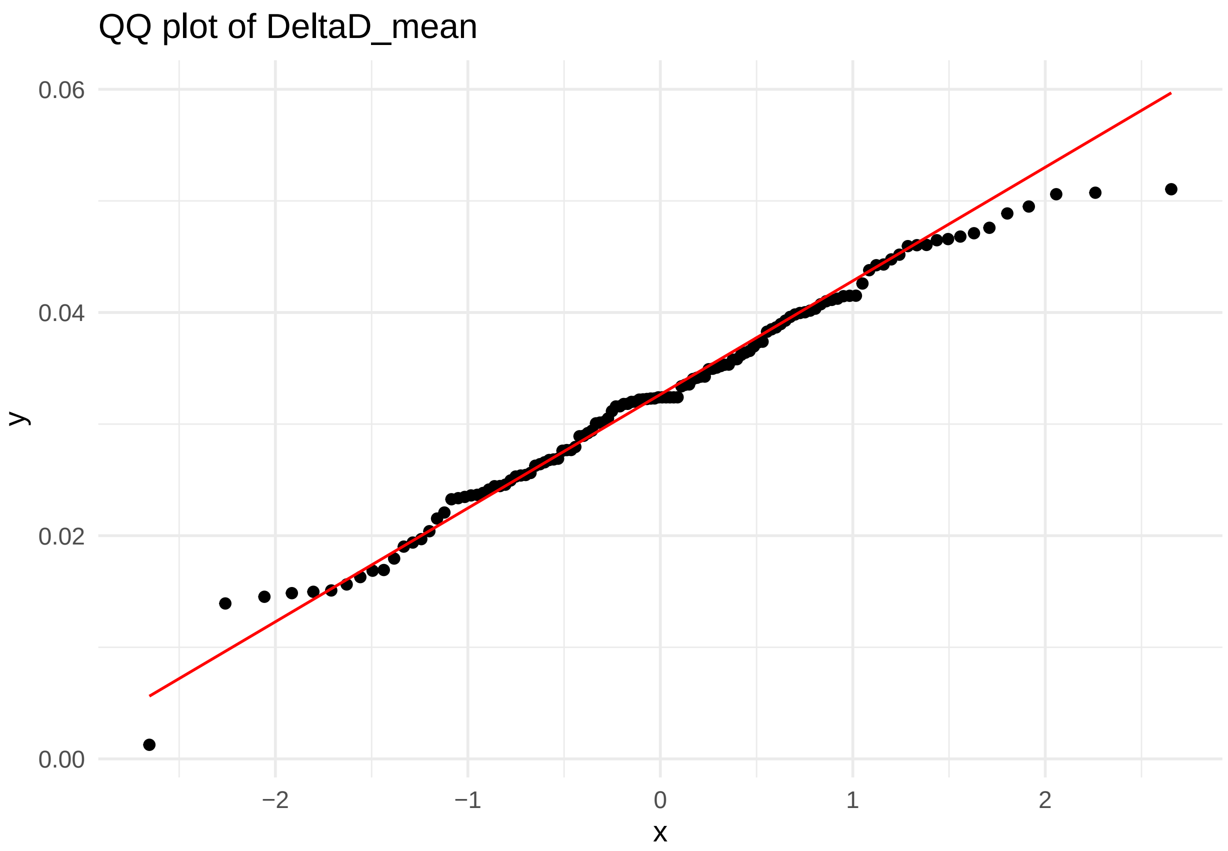
```
ggplot(data.frame(DeltaD_mean), aes(x = DeltaD_mean)) +
    geom_histogram(binwidth = 0.002, fill = "blue", color = "black", alpha = 0.7)
```



```
    theme_minimal() +
    ggtitle(paste("Histogram of DeltaD_mean"))

ggplot(data.frame(DeltaD_mean), aes(x = "DeltaD_mean", y = DeltaD_mean)) +
    geom_boxplot(fill = "blue", alpha = 0.7, width = 0.2) +
    theme_minimal() +
    ggtitle(paste("Box plot of DeltaD_mean"))
```



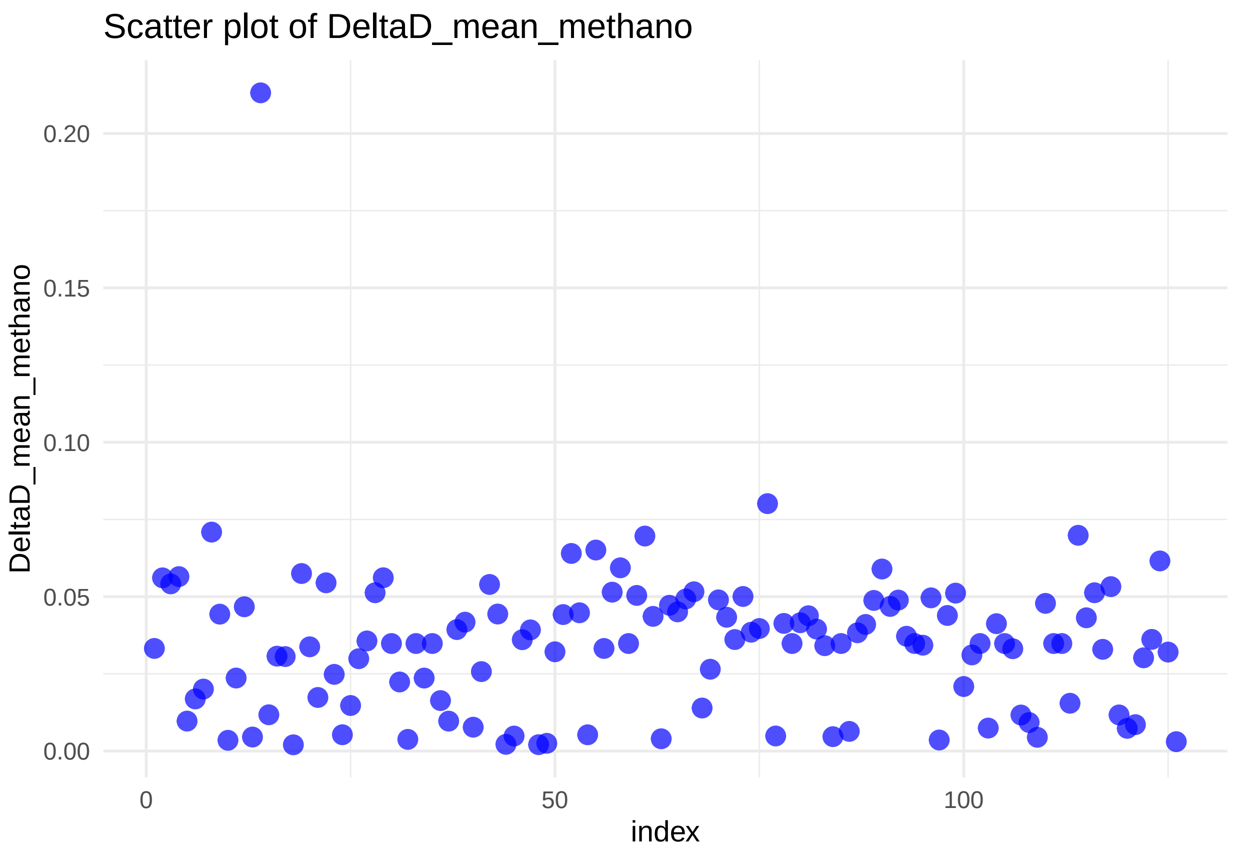
```r
ggplot(data.frame(DeltaD_mean), aes(sample = DeltaD_mean)) +
    geom_qq() +
    geom_qq_line(color = "red") +
    theme_minimal() +
    ggtitle(paste("QQ plot of DeltaD_mean"))
```



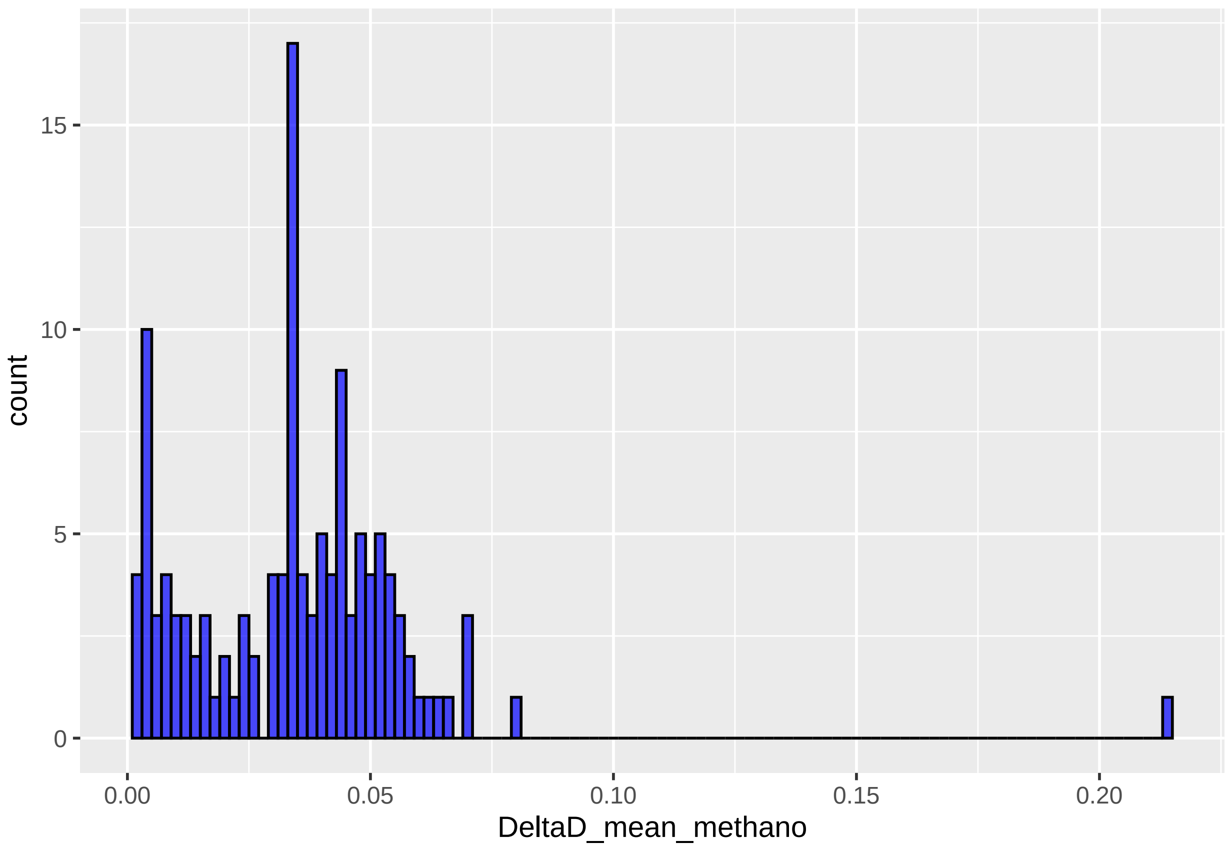
```r
shapiro.test(DeltaD_mean)
```

Plots for DeltaD_mean_methano

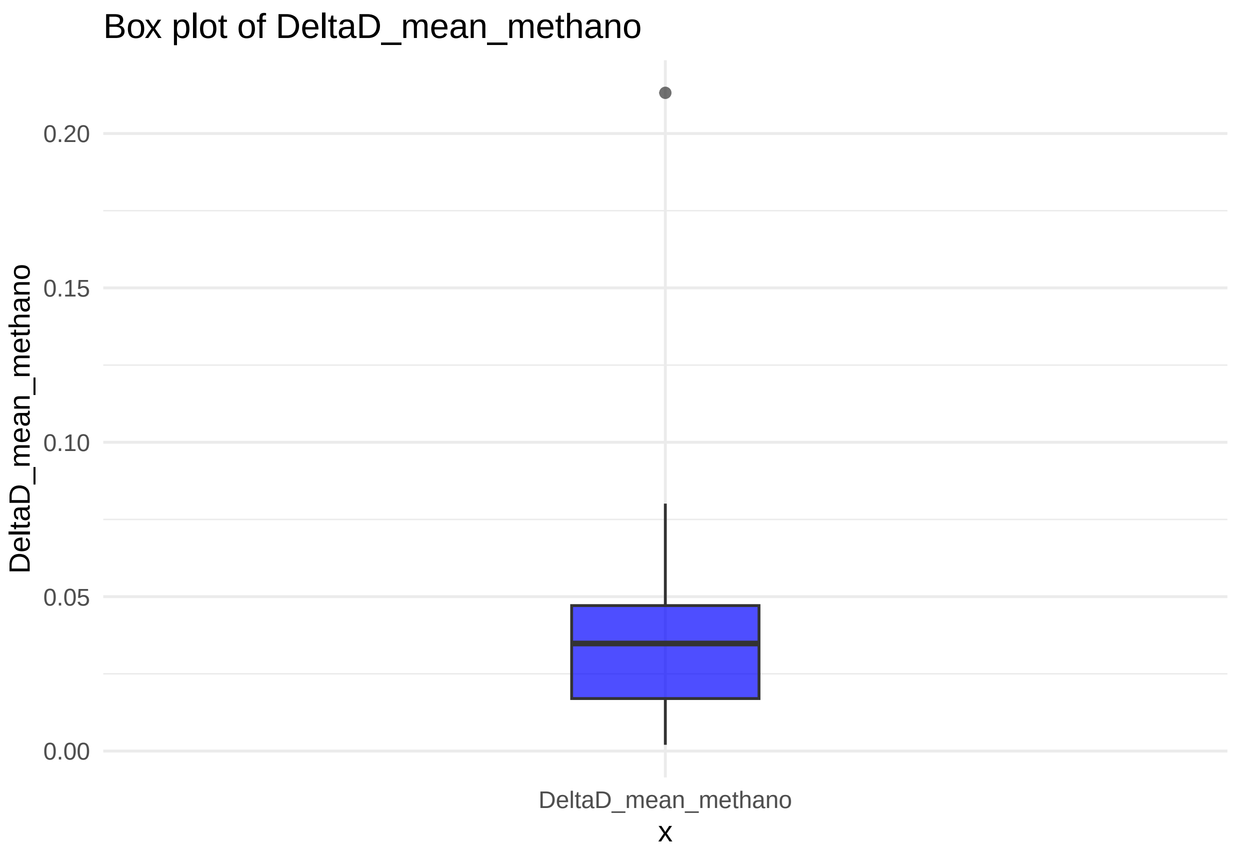
```r
index <- which(!is.na(DeltaD_mean_methano))

ggplot(data.frame(DeltaD_mean_methano), aes(x = index, y = DeltaD_mean_methano)) +
    geom_point(color = "blue", size = 3, alpha = 0.7) +
    theme_minimal() +
    ggtitle(paste("Scatter plot of DeltaD_mean_methano"))
```



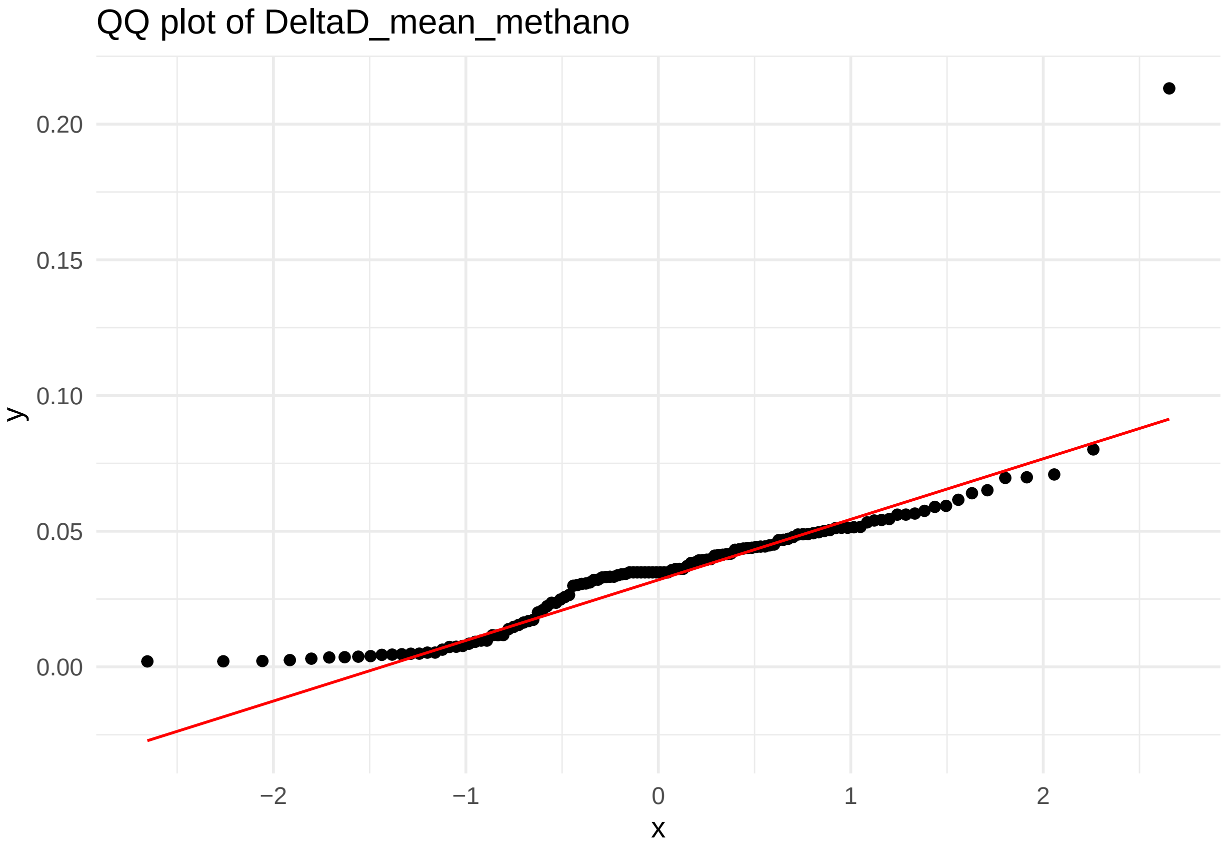
```
ggplot(data.frame(DeltaD_mean_methano), aes(x = DeltaD_mean_methano)) +
    geom_histogram(binwidth = 0.002, fill = "blue", color = "black", alpha = 0.7)
```



```
    theme_minimal() +
    ggtitle(paste("Histogram of DeltaD_mean_methano"))

ggplot(data.frame(DeltaD_mean), aes(x = "DeltaD_mean_methano", y = DeltaD_mean_methano)) +
    geom_boxplot(fill = "blue", alpha = 0.7, width = 0.2) +
    theme_minimal() +
    ggtitle(paste("Box plot of DeltaD_mean_methano"))
```



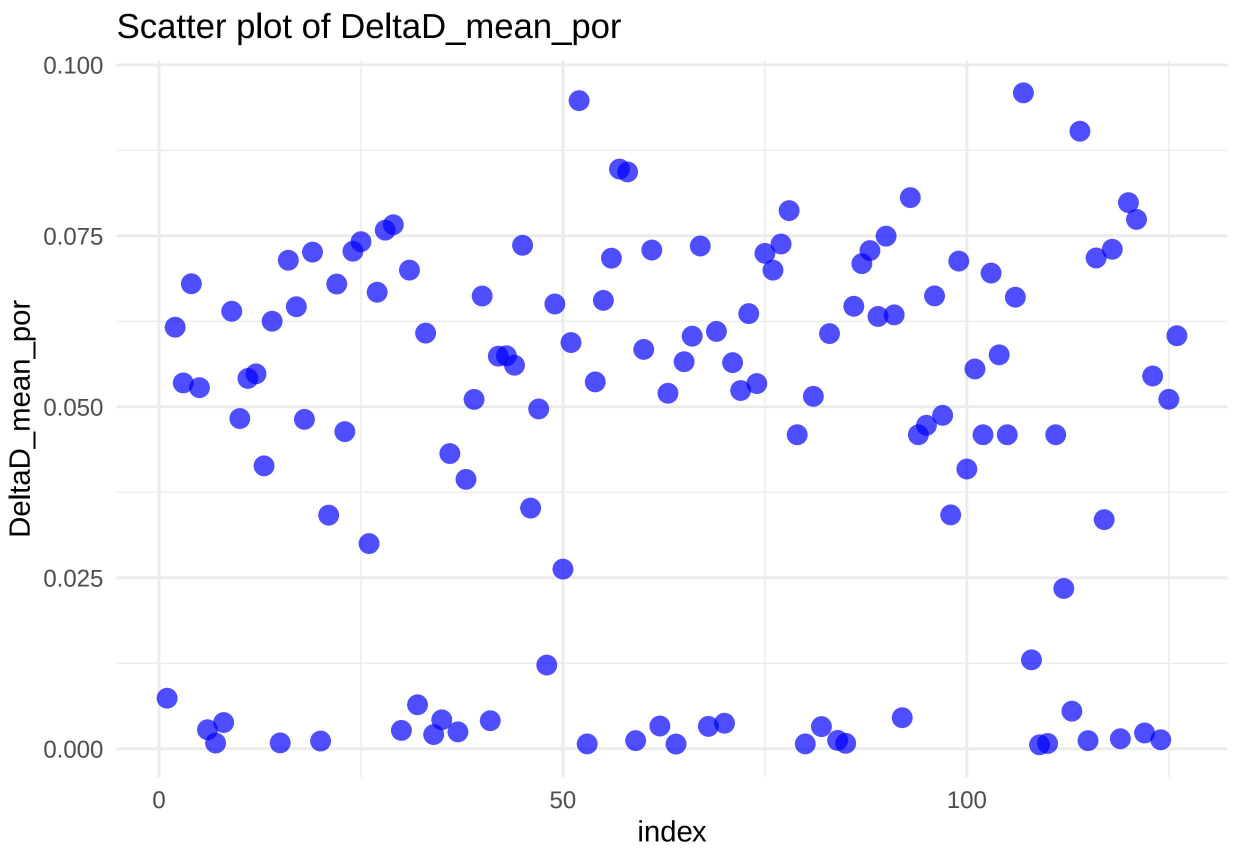
```r
ggplot(data.frame(DeltaD_mean_methano), aes(sample = DeltaD_mean_methano)) +
    geom_qq() +
    geom_qq_line(color = "red") +
    theme_minimal() +
    ggtitle(paste("QQ plot of DeltaD_mean_methano"))
```



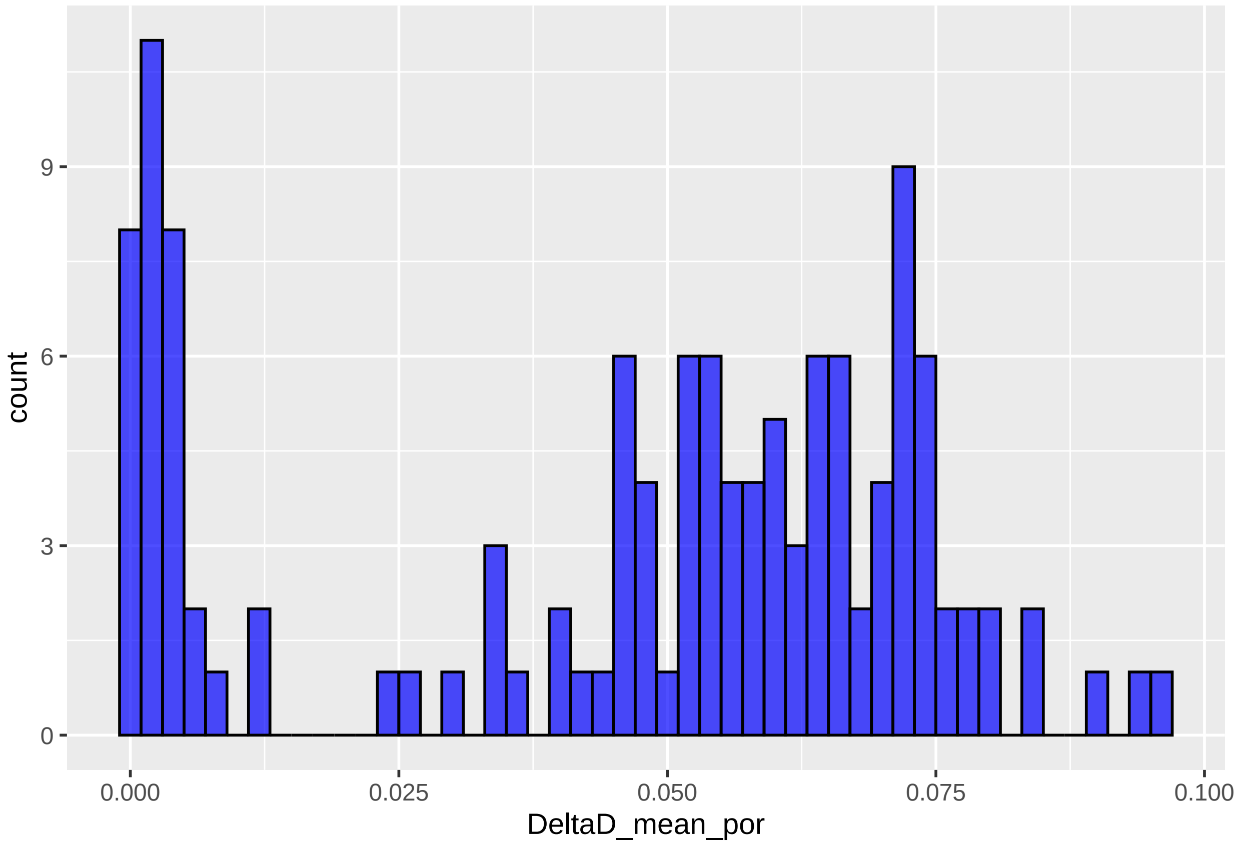
```
shapiro.test(DeltaD_mean_methano)
```

Plots for DeltaD_mean_por

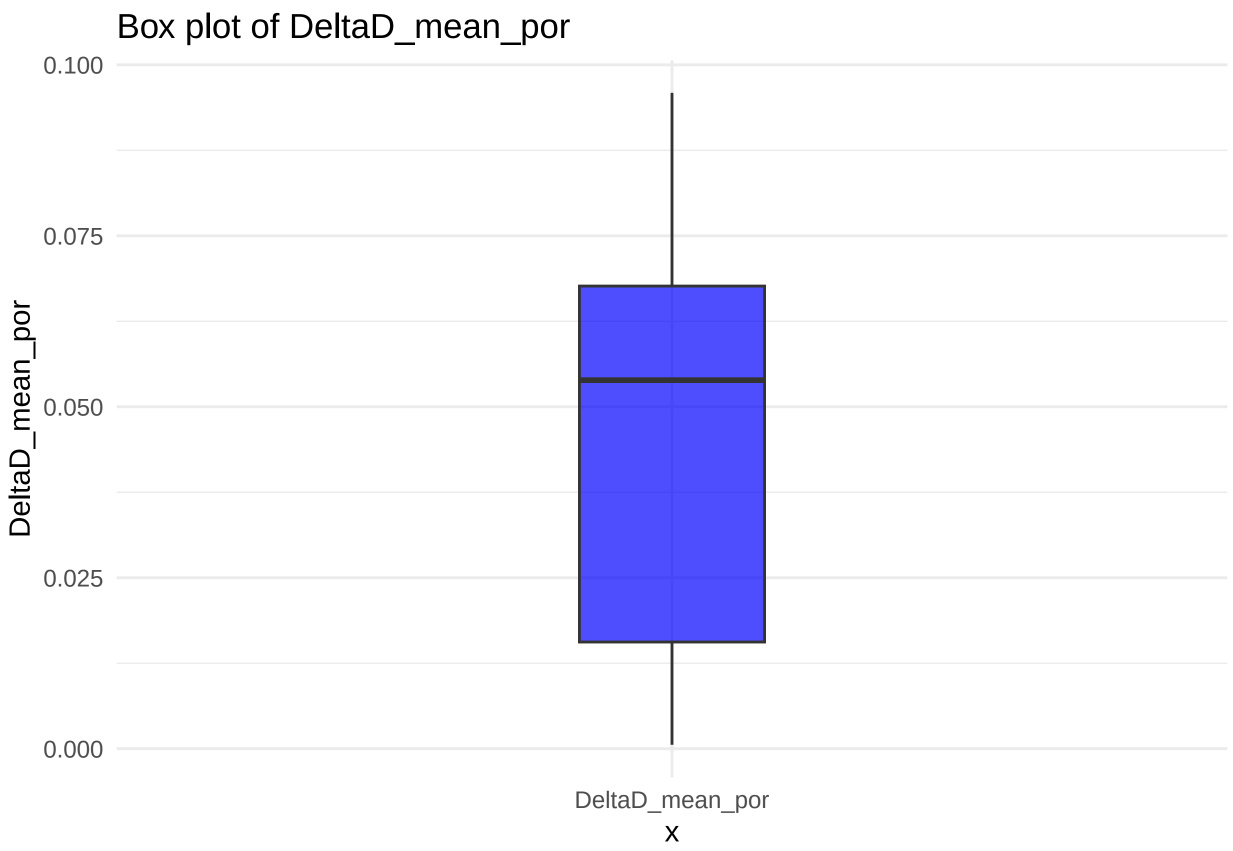
```
index <- which(!is.na(DeltaD_mean_por))

ggplot(data.frame(DeltaD_mean_por), aes(x = index, y = DeltaD_mean_por)) +
    geom_point(color = "blue", size = 3, alpha = 0.7) +
    theme_minimal() +
    ggtitle(paste("Scatter plot of DeltaD_mean_por"))
```



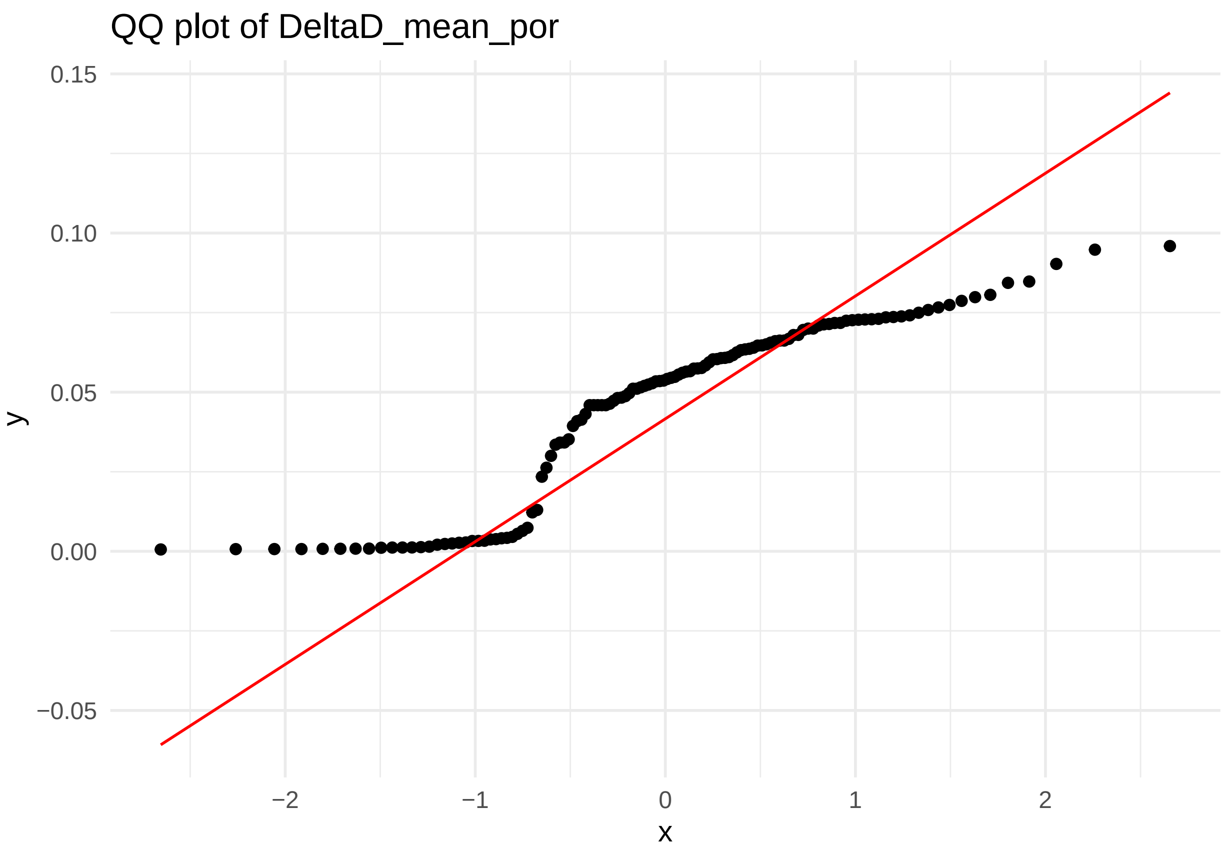
```r
ggplot(data.frame(DeltaD_mean_por), aes(x = DeltaD_mean_por)) +
    geom_histogram(binwidth = 0.002, fill = "blue", color = "black", alpha = 0.7)
```



```
    theme_minimal() +
    ggtitle(paste("Histogram of DeltaD_mean_por"))

ggplot(data.frame(DeltaD_mean), aes(x = "DeltaD_mean_por", y = DeltaD_mean_por)) +
    geom_boxplot(fill = "blue", alpha = 0.7, width = 0.2) +
    theme_minimal() +
    ggtitle(paste("Box plot of DeltaD_mean_por"))
```



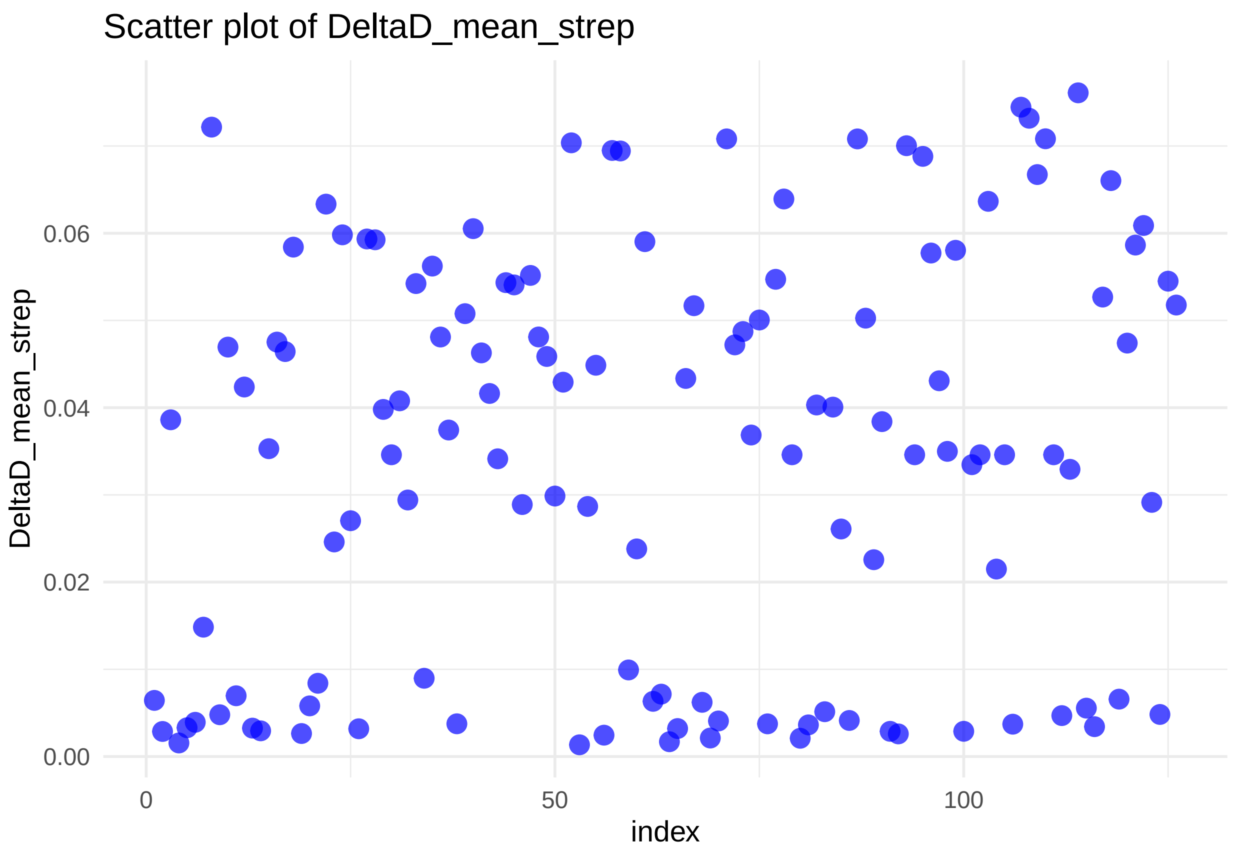
```r
ggplot(data.frame(DeltaD_mean_por), aes(sample = DeltaD_mean_por)) +
    geom_qq() +
    geom_qq_line(color = "red") +
    theme_minimal() +
    ggtitle(paste("QQ plot of DeltaD_mean_por"))
```



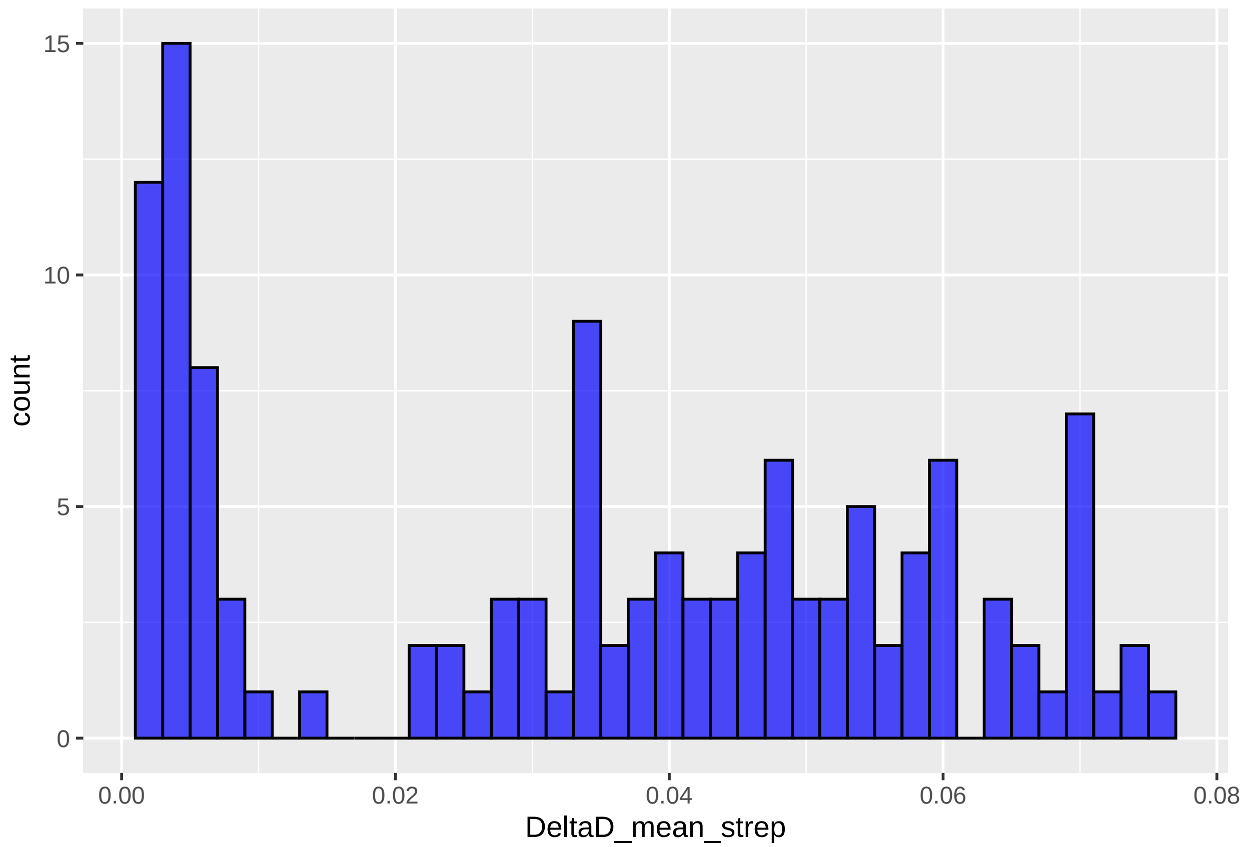
```
shapiro.test(DeltaD_mean_por)
```

Plots for DeltaD_mean_strep

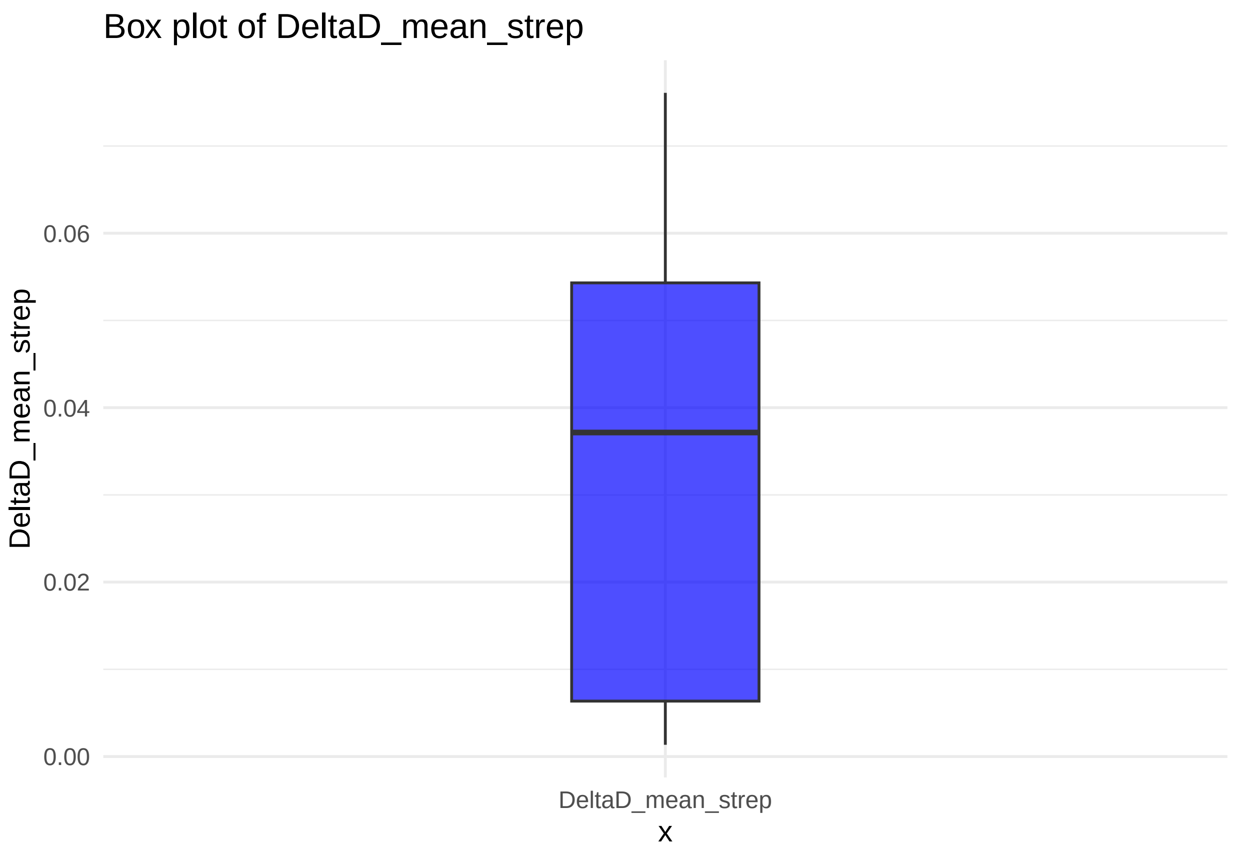
```
index <- which(!is.na(DeltaD_mean_strep))

ggplot(data.frame(DeltaD_mean_strep), aes(x = index, y = DeltaD_mean_strep)) +
    geom_point(color = "blue", size = 3, alpha = 0.7) +
    theme_minimal() +
    ggtitle(paste("Scatter plot of DeltaD_mean_strep"))
```



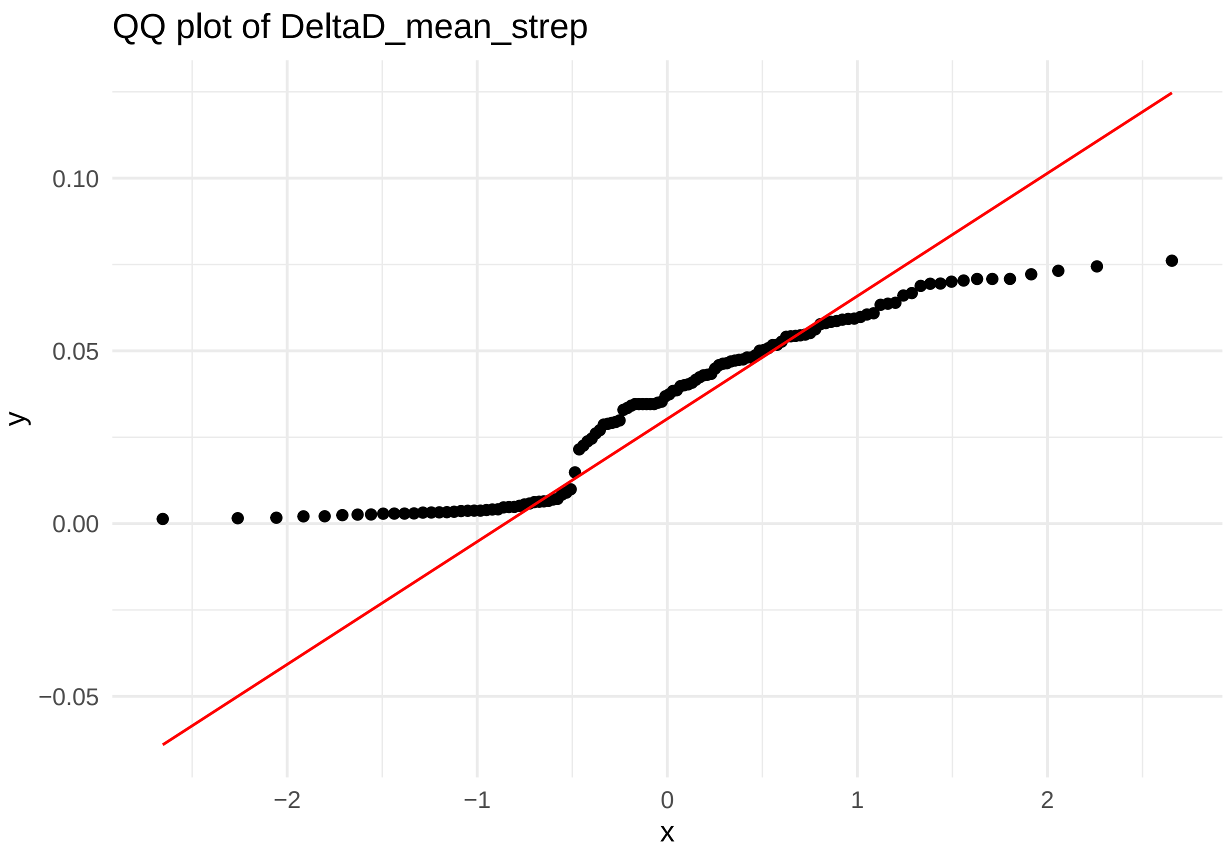
```
ggplot(data.frame(DeltaD_mean_strep), aes(x = DeltaD_mean_strep)) +
    geom_histogram(binwidth = 0.002, fill = "blue", color = "black", alpha = 0.7)
```



```
    theme_minimal() +
    ggtitle(paste("Histogram of DeltaD_mean_strep"))

ggplot(data.frame(DeltaD_mean), aes(x = "DeltaD_mean_strep", y = DeltaD_mean_strep)) +
    geom_boxplot(fill = "blue", alpha = 0.7, width = 0.2) +
    theme_minimal() +
    ggtitle(paste("Box plot of DeltaD_mean_strep"))
```



```r
ggplot(data.frame(DeltaD_mean_strep), aes(sample = DeltaD_mean_strep)) +
    geom_qq() +
    geom_qq_line(color = "red") +
    theme_minimal() +
    ggtitle(paste("QQ plot of DeltaD_mean_strep"))
```



```
shapiro.test(DeltaD_mean_strep)
```