Week1, Assignment 1

Gabi Rivera

```
library(tidyverse)
library(GGally)
```

Problem 3.1 (30 points)

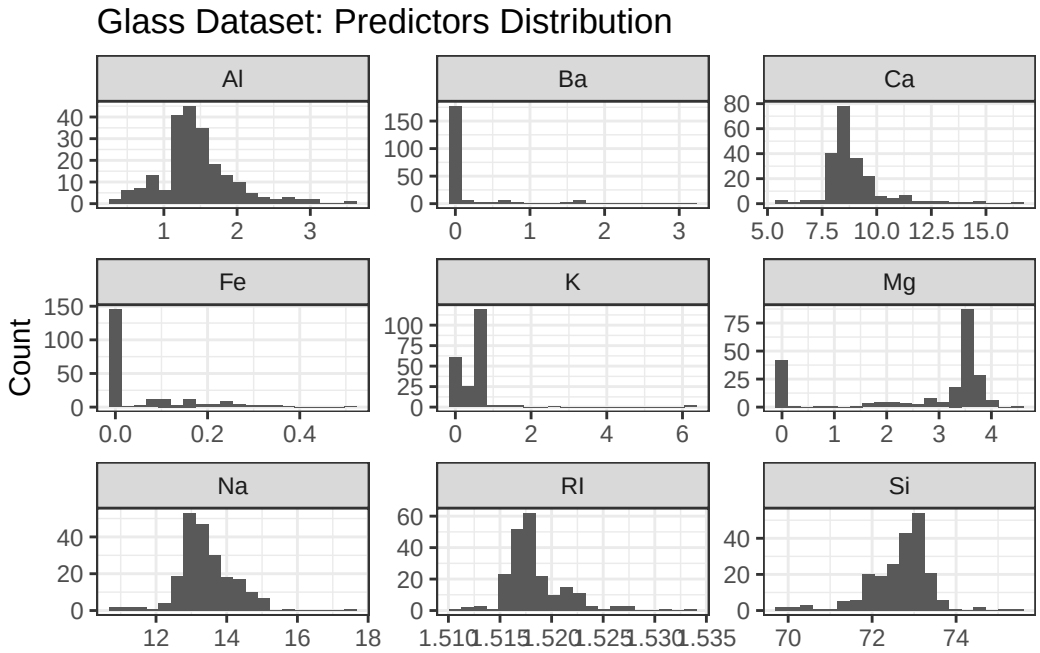
The UC Irvine Machine Learning Repository contains a Glass Identification Data Set. The data consist of 214 glass samples labeled as one of seven class categories. There are nine predictors, including the refractive index and percentages of eight elements: Na, Mg, Al, Si, K, Ca, Ba, and Fe. The data can be accessed via:

```
library(mlbench)

# Data upload and general information:
data(Glass)
```

3.1.a (10 points)

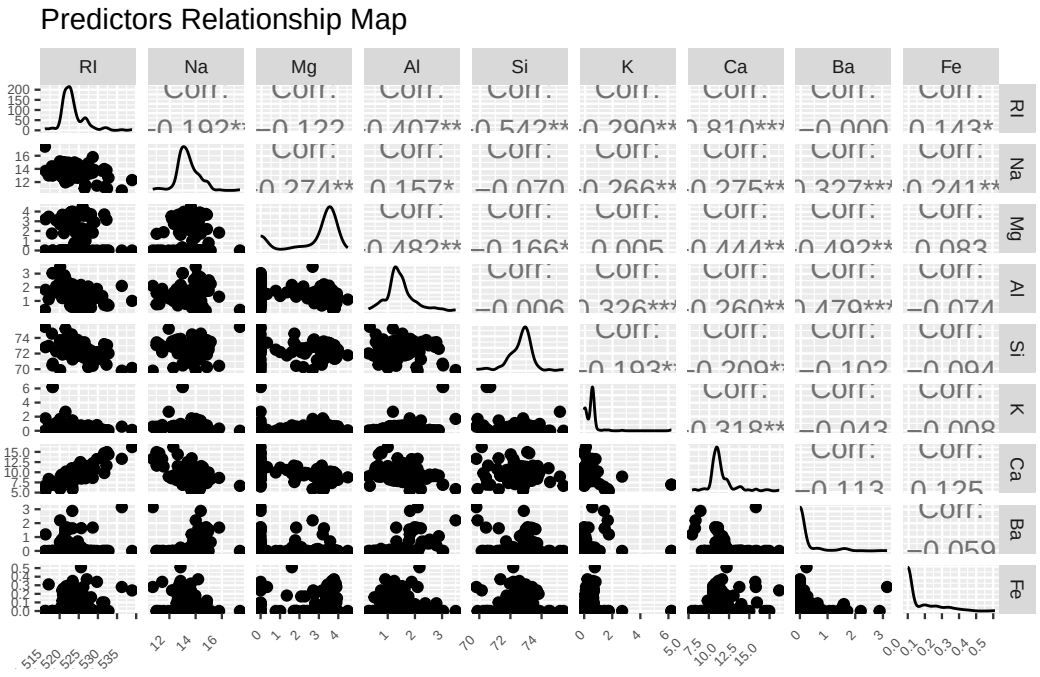
Using visualizations, explore the predictor variables to understand their distributions …

```
# Create histograms for each predictors:
Glass |>
  pivot_longer(-Type, names_to = 'Element', values_to = 'value')  |>
  ggplot(aes(x = value)) +
  geom_histogram(bins = 20) +
  facet_wrap(~Element, scales = 'free', ncol = 3) +
  theme_bw() +
  labs(title = 'Glass Dataset: Predictors Distribution', x = NULL,
       y = "Count")
```



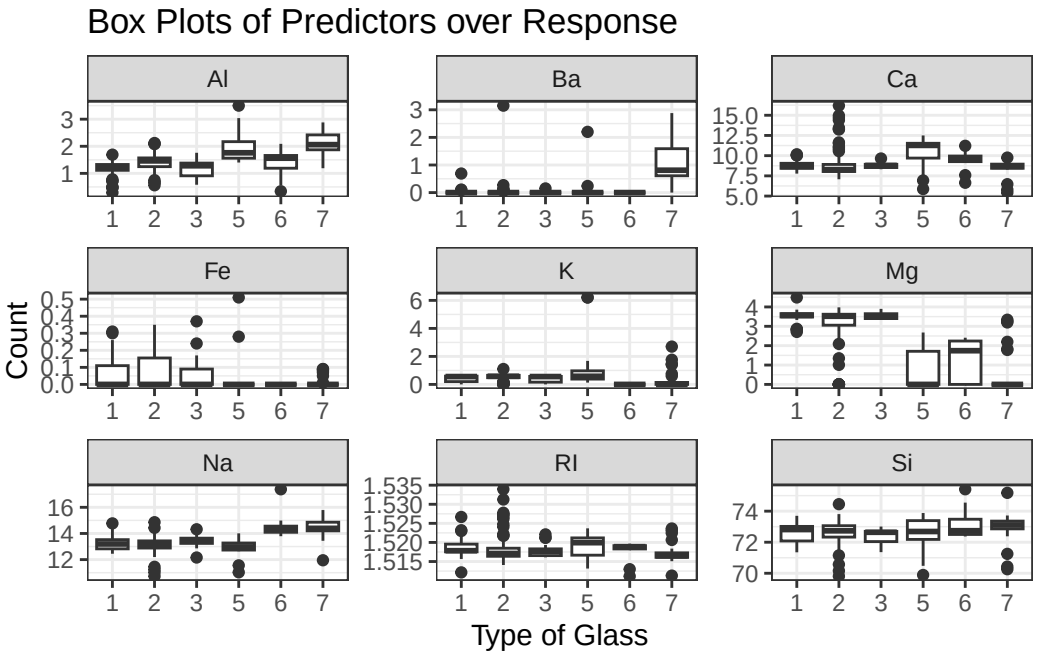
as well as the relationships between predictors.

```
# Create relationship map across predictors:
Glass |>
  select(-Type) |>
  ggpairs(title = 'Predictors Relationship Map', progress = TRUE,
          mapping = aes(text = list(size = 5))) +
  theme_grey(base_size = 9) +
  theme(axis.text.x = element_text(angle = 45, hjust = 2, size = 5),
        axis.text.y = element_text(size = 5))
```



Explore the relationship between predictors and response.

```
# Create box plots of response for each predictors:
Glass |>
  pivot_longer(-Type, names_to = 'Element', values_to = 'value')  |>
  ggplot(aes(Type, value)) +
  geom_boxplot() +
  facet_wrap(~Element, scales = 'free', ncol = 3) +
  theme_bw() +
  labs(title = 'Box Plots of Predictors over Response', x = "Type of Glass",
       y = "Count")
```



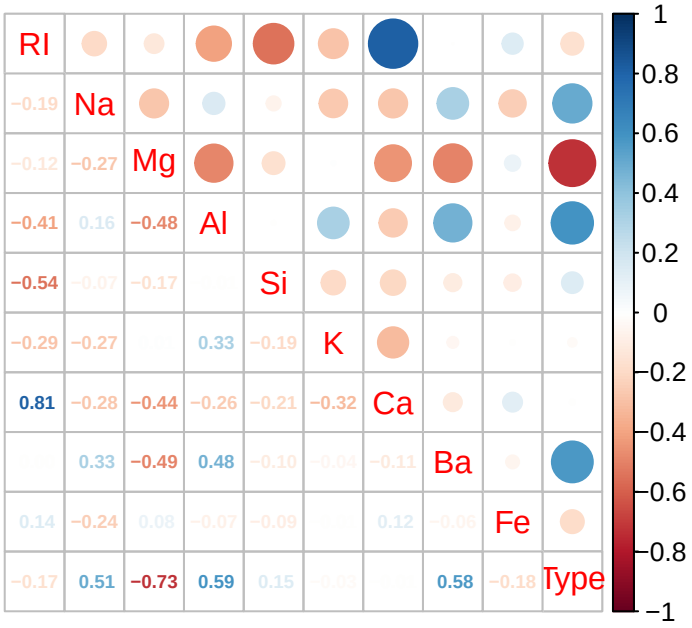
Which elements do you think will be good/poor predictors (based on the visualizations)?

Looks like AI is normally distributed compared to the rest. AI have slightly high correlation with RI, Ba, and Mg. Na, K, and Fe has the least correlation amongst the other predictors which may result to least chance of multicollinearity. K and Fe shows the least ideal distribution which are way skewed to the right compared to Na. They will need to be normalized for certain models. Box plot spread for Na looks like each type of glass are well spaced together with AI element.

Compute the correlations between the predictors and the `Type` variable.

```
suppressPackageStartupMessages(library(corrplot))
library(corrplot)

# Create correlation map including the Type variable:
Glass |>
  mutate(Type = as.numeric(Type)) |>
  cor() |>
  corrplot.mixed(title = "Correlation Plot Between Type and Predictors",
                 tl.cex = 1, number.cex = 0.6, mar = c(0, 0, 1, 0))
```

Correlation Plot Between Type and Predictors



Which elements do you think will be good/poor predictors (based on the correlation calculation)?

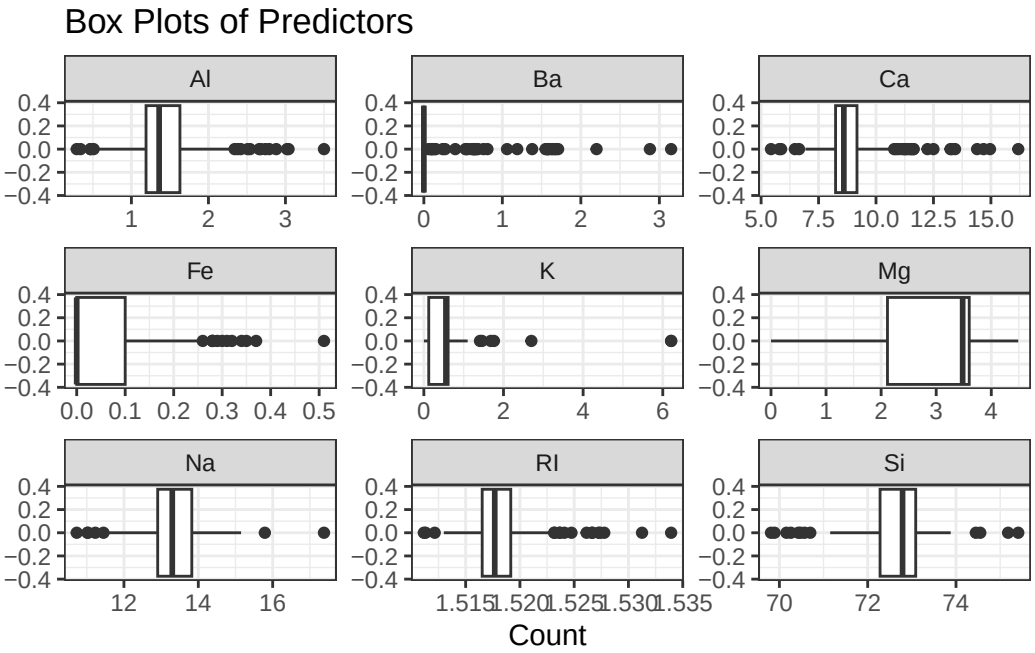
From the correlation matrix, the type variable have least correlation with Si, K, Ca, and Fe elements. Overall, Fe and K have the least correlation with the rest of the variables. Based of off less than +/- 0.5 coefficient score.

3.1.b (10 points)

Are there any outliers in the data?

Yes, even from previous visualization there are definitely outliers in the data looking at the distribution and the box plots shown in 3.1.a.

```
# Create box plots for each elements and RI to show outliers:
Glass |>
  pivot_longer(-Type, names_to = 'Element', values_to = 'value')  |>
  ggplot(aes(value)) +
  geom_boxplot() +
  facet_wrap(~Element, scales = 'free', ncol = 3) +
  theme_bw() +
  labs(title = 'Box Plots of Predictors', x = "Count",
       y = NULL)
```



If we look at each predictors, only Mg show no outliers from its box plot. The rest have outliers.

Are any predictors skewed?

Yes, all of the predictors has some level of skewness. Si and Mg are negatively skewed the rest are skewed to the right. Na is the closes to being normally distributed.

```
library(e1071)

# Skewness scores:
Glass |>
  select(-Type)  |>
  sapply(skewness) |>
  round(2)
```

```
   RI    Na    Mg    Al    Si     K    Ca    Ba    Fe
 1.60  0.45 -1.14  0.89 -0.72  6.46  2.02  3.37  1.73
```

3.1.c (10 points)

Are there any relevant transformations of one or more predictors that might improve the classification model? Assume the model requires the predictors to have approximately symmetric

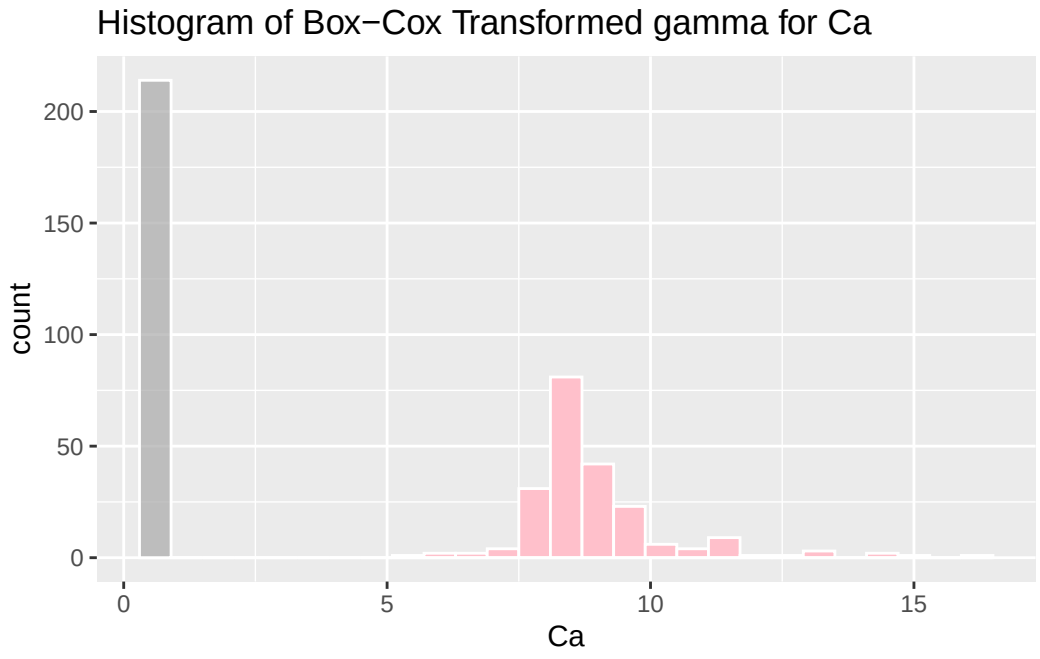
distribution. Apply relevant transformations to the predictors and observe the changes to the distributions of predictors.

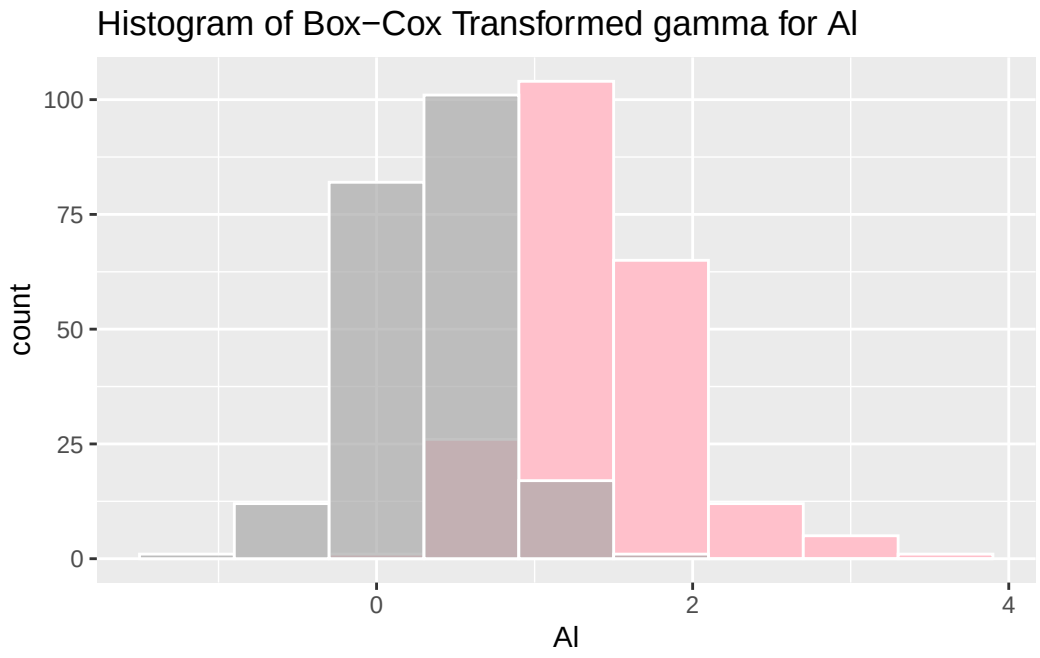
For right skewness, log and square-root transformation might improve the model. Box-Cox transformation deals with positive and negative skew through optimal power transformation.

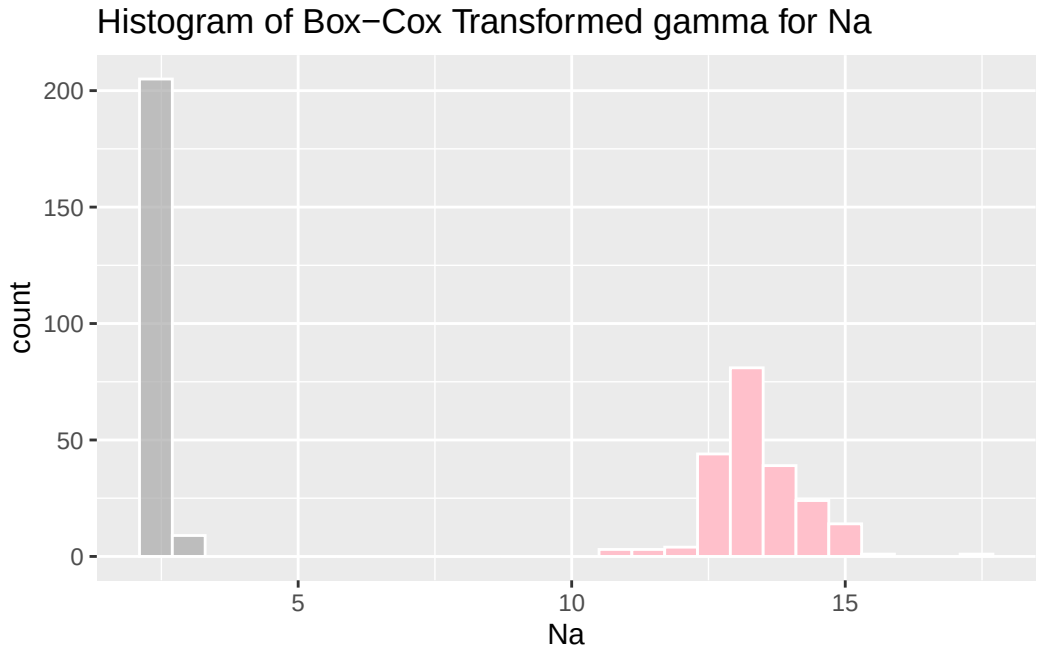
```
suppressPackageStartupMessages(library(caret))
library(caret)

# Calculate boxcoxtrans gamma transformation for each predictors:
for (col in names(Glass)[-10]) {
  bct <- BoxCoxTrans(Glass[[col]], na.action = na.pass)
  Glass[[paste0("gamma_", col)]] <- predict(bct, Glass[[col]])
}

# Skewness scores:
Skewness_scr <-
Glass |>
  select(-Type)  |>
  sapply(skewness) |>
  round(2)

skewness_score <- as.data.frame(t(Skewness_scr))

# Calculate % difference for each predictors skewness score:
calculate_diff <- function(col) {
  return (((skewness_score[[col]] - skewness_score[[paste0("gamma_", col)]]) /
          ((skewness_score[[col]] + skewness_score[[paste0("gamma_", col)]]) /
             2)) * 100)
}

gpredictors <- c("RI", "Na", "Mg", "Al", "Si", "K", "Ca", "Ba", "Fe")
results_list <- lapply(gpredictors, calculate_diff)
names(results_list) <- paste0(gpredictors, "_percent_diff")
skew_p_diff <- t(round((do.call(rbind, results_list)),2))
skew_p_diff
```

```
     RI_percent_diff Na_percent_diff Mg_percent_diff Al_percent_diff
[1,]            1.89             175               0          163.27
     Si_percent_diff K_percent_diff Ca_percent_diff Ba_percent_diff
[1,]           10.22              0          241.53               0
     Fe_percent_diff
[1,]               0
```

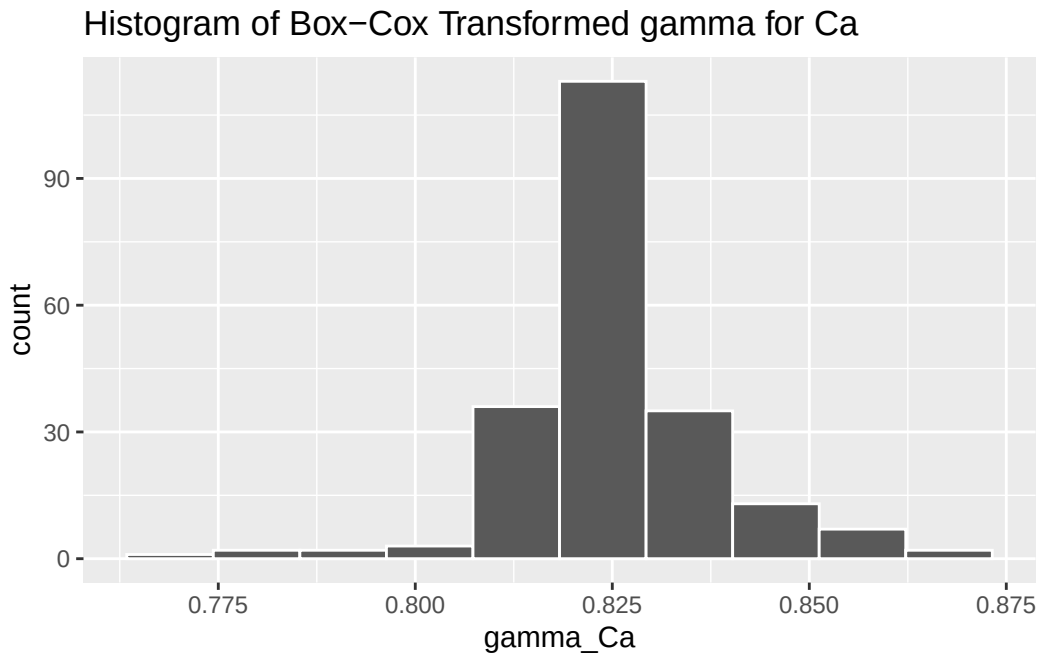
```r
# Create histogram loop:
plot_histogram <- function(Glass, predictor) {
  print(
  ggplot(Glass, aes(x = !!sym(predictor))) +
    geom_histogram(bins = 20, binwidth = .6, fill = "pink",
                   color = "white") +
    geom_histogram(bins = 20, aes(x = !!sym(
      paste("gamma_", predictor, sep = ""))),
                   binwidth = .6, fill = "darkgray",
      color = "white",  alpha = .7) +
    labs(title = paste("Histogram of Box-Cox Transformed gamma for",
                       predictor))
  )
}

# Run predictors through the loop:
# Only Ca and Al have more than 50% percent diff for skewness.
predictor_columns <- c("Ca","Al", "Na")
for (predictor in predictor_columns) {
  plot_histogram(Glass, predictor)
}
```



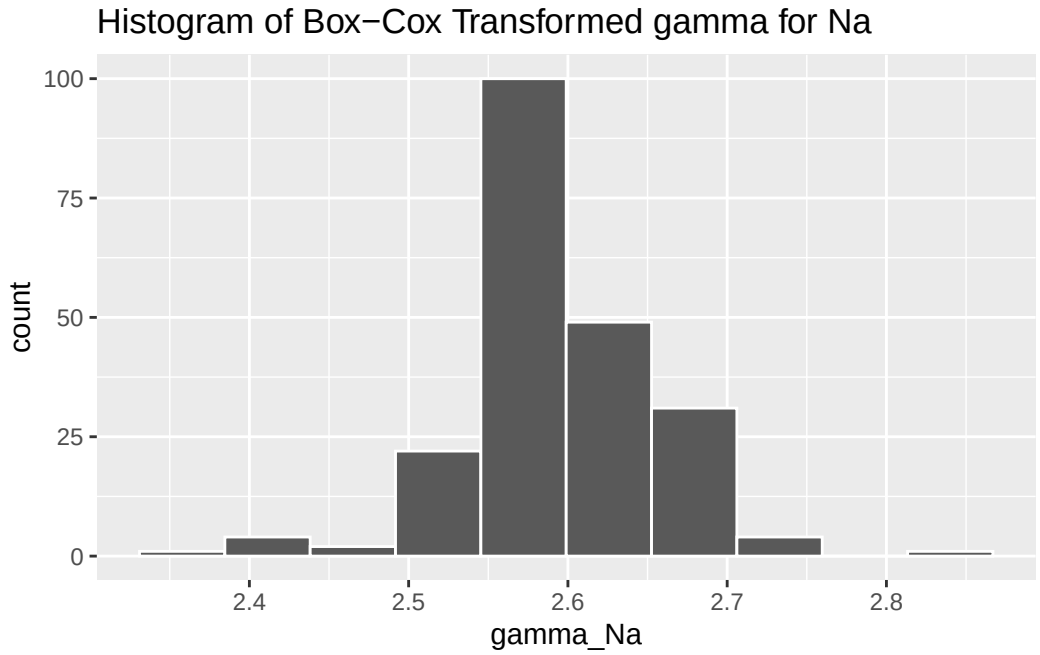





```
# Another look at gamam_Ca:
Glass |>
  ggplot(aes(x = gamma_Ca)) +
  geom_histogram(bins= 10, color = "white") +
  labs(title ="Histogram of Box-Cox Transformed gamma for Ca")
```




```
# Another look at gamam_Na:
Glass |>
  ggplot(aes(x = gamma_Na)) +
  geom_histogram(bins= 10, color = "white") +
  labs(title ="Histogram of Box-Cox Transformed gamma for Na")
```



Problem 3.2 (20 points)

The image below shows a scatter plot matrix of the continuous features of a dataset. Discuss the relationships between the features in the dataset that this scatter plot highlights. Make sure to discuss relationships between all pairs.

1. *LifeExpectancy has a strong correlation coefficient with InfantMortality at 1 cc. The scatter plot for LifeExpenctancy and InfantMortality is linear (negative).*
2. *LifeExpectancy has a strong positive correlation with Education at 0.477 cc.*
3. *LifeExpectancy has a strong positive correlation with Health at 0.572 cc.*
4. *LifeExpectancy has a strong positive correlation with HealthUSD at 0.776 cc.*
5. *InfantMortality has a strong negative correlation with Education at -0.504 cc.*
6. *InfantMortality has a strong negative correlation with Health at -0.500 cc.*
7. *InfantMortality has a strong negative correlation with HealthUSD at -0.736 cc.*
8. *Education has a strong positive correlation with Health at 0.689 cc.*
9. *Education has a strong positive correlation with HealthUSD at 0.527 cc.*
10. *Health has a strong positive correlation with HealthUSD at 0.760 cc.*

Problem 3.3 (10 points)

Discuss the relationships between the variables shown in below visualizations:

3.3.a (5 points)

The visualization below illustrates the relationship between Diastolic BP and Tachycardia, left most plot has data where Tachycardia = true and false (the full study population).

No-Tachycardia (False) is more normally distributed compared to the Diastolic BP control. Both have relatively the same peak location at ~80 BP. No-Trachycardia gained slightly higher density count compared to the control at > 0.04.

With-Tachycardia (True) have similar bimodal distribution as the Diastolic BP only. The first peak location is shifted more to the right ~90 BP compared to the control. With-Tachycardia lost about 0.01 density count compared to Diastolic BP at its highest peak. The smaller peak of Trachycardia (True) has peak ~120.

Overall, the normal distribution of Trachycardia (False) averages in a lower Diastolic BP compared to Tarchycardia (True) and control for mean/median. Modes of Tachycardia (False)

and control are relatively the same. Mode/mean/media of Tachycardia (True) are more shifted to the right compared to the other two plots. Tachycardia (True) is the contributor for the bimodal pattern seen in Diastolic BP control, especially the tailing.

3.3.b (5 points)

The visualization below illustrates the relationship between Height and Tachycardia, left most plot has data where Tachycardia = true and false.

No-Tachycardia (False) show relatively similar bimodal pattern for Height distribution as well as modes location with the control. No-Tachycardia has slightly lower density count compared to the control.

Same with Trachycardia (True), the bimodal pattern is relatively similar with the control. The highest peak mode of Trachycardia (True) is shifted to the right. The peaks has slightly gained to a higher density count compared to the control. Tachycardia (True) have a steeper dip on the tailed compared to the rest.

Overall, height has relatively similar pattern and distribution between control, Tachycardia (False), and Tachycardia (True). Ever slightly difference is observed.

Problem 3.4 (30 points)

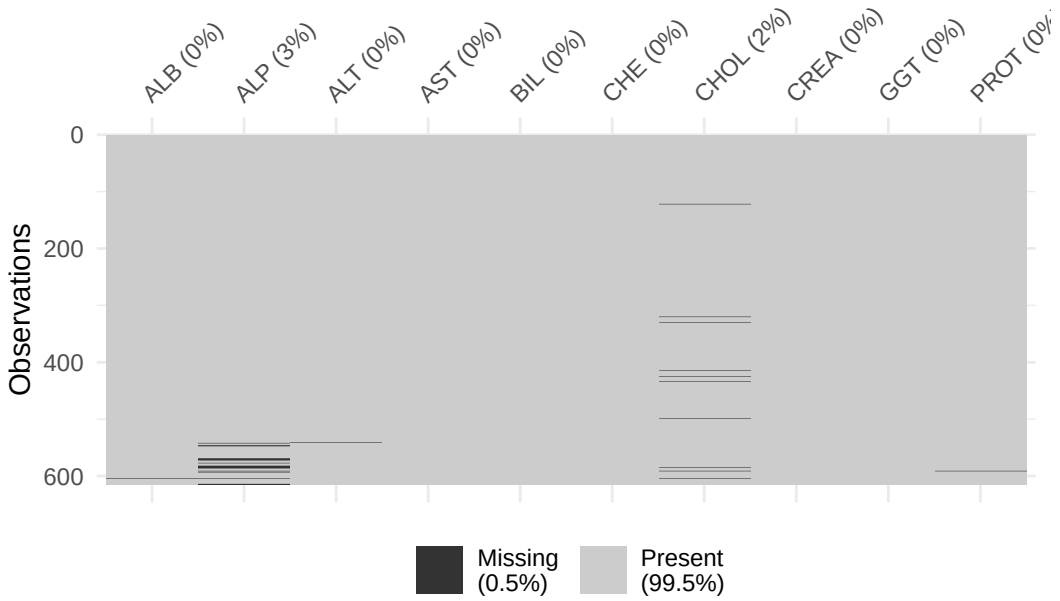
Use the HCV Data Set at the UCI Machine Learning Repository (or download the `hcvdat0.csv` file in Canvas) and pick the numeric predictors (you can do this by excluding columns "X" , "Category", "Age" and "Sex") to perform the following analysis in R:

```
csv <- list.files(here::here(), pattern = 'hcvdat0.csv',
                  recursive = TRUE) |> head(1)
hcv <- read_csv(csv, show_col_types = FALSE) |>
        select(-c(1, Category, Age, Sex))
```

```
New names:
* `` -> `...1`
```

3.4.a. Are there any missing data in the predictors? Identify all the predictors with missing values (5 points)

```
library(naniar)

# Identify missing values:
vis_miss(hcv)
```



Yes, the resource page from UCI identify the dataset as having missing values. ALB, ALP, ALT, CHOL, and PROT are the lab tests that show missing values from the missing value plot.

3.4.b. Summarize the missing data by each predictor. (5 points)

```
# Missing values for each predictor:
hcv[hcv == ""] <- NA
na_value <- sapply(hcv, function(x) sum(is.na(x)))
na_value
```

```
 ALB  ALP  ALT  AST  BIL  CHE CHOL CREA  GGT PROT
   1   18    1    0    0    0   10    0    0    1
```

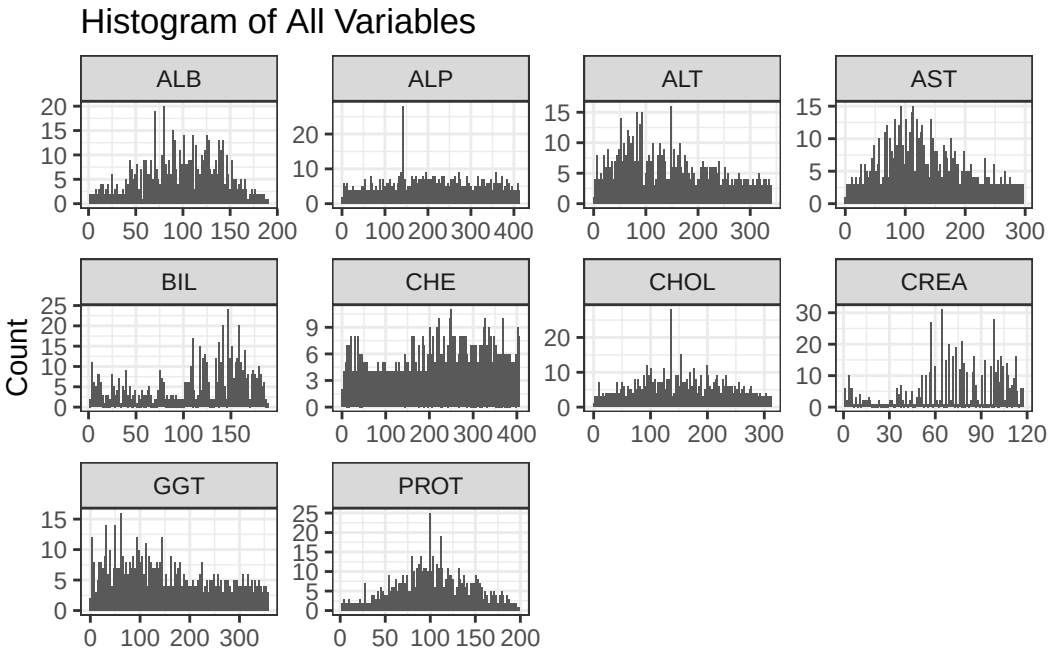
```
# Impute mode into missing values of each predictor:
mode_impute <- function(x) {
  mode_val <- names(sort(table(x), decreasing = TRUE))[1]
```

```
  x[is.na(x)] <- mode_val
  return(x)
}

# Apply mode imputation to each column
hcv_imputed <- as.data.frame(lapply(hcv, mode_impute))
hcv_imputed[] <- lapply(hcv_imputed, function(x)
  as.numeric(as.factor(x)))

# Check imputed missing values:
impt_na_value <- sapply(hcv_imputed, function(x)
  sum(is.na(x)))
impt_na_value
```

```
 ALB  ALP  ALT  AST  BIL  CHE CHOL CREA  GGT PROT
   0    0    0    0    0    0    0    0    0    0
```

3.4.c. Plot the histograms of predictors and visually identify predictors with skewed distributions. (5 points)

```
# Create historgram fro each predictors:
hcv_imputed |>
  pivot_longer(cols = everything(), names_to = "variable",
               values_to = "value") |>
  ggplot(aes(x = value)) +
  geom_histogram(bins = 100) +
  facet_wrap(~ variable, scales = 'free', ncol = 4) +
  theme_bw() +
  labs(title = 'Histogram of All Variables',
       x = NULL, y = "Count")
```



From the visualization, looks like CHOL and PROT have fairly normal distribution while the rest have some form of skewness to them.

3.4.d. Compute skewness using the skewness function from the e1071 package. Are the skewness values aligning with the visual interpretations from part c. (5 points)

```
# Skewness scores:
hcv_imputed |>
  sapply(skewness) |>
  round(2)
```

```
  ALB   ALP   ALT   AST   BIL   CHE  CHOL  CREA   GGT  PROT
-0.16  0.01  0.40  0.35 -0.61 -0.23  0.05 -0.69  0.37 -0.07
```

ALP, CHOL, and PROT has the lowest skewness score and for the most part the visualization aligns. Except for ALP, the distribution is flat and have a peak outlier that might be contributing to the skewness score. The rest have expected skewness and aligns with the scoring well.

3.4.e. Apply box-cox transformations to the data and then recompute the skewness metrics and report the differences; does box-cox transformation help mitigate skewness? (5 points)

```
# Calculate boxcoxtrans gamma transformation for each predictors:
for (col in names(hcv_imputed)) {
  bct_h <- BoxCoxTrans(hcv_imputed[[col]],
                       na.action = na.pass)
  hcv_imputed[[paste0("gamma_", col)]] <-
    predict(bct_h, hcv_imputed[[col]])
}

# Skewness scores:
skew <- hcv_imputed |>
          sapply(skewness) |>
          round(2)
skew <- as.data.frame(t(skew))


# Calculate % difference for each predictors skewness score:
calculate_diff0 <- function(col) {
  return (((skew[[col]] - skew[[paste0("gamma_", col)]]) /
          ((skew[[col]] + skew[[paste0("gamma_", col)]]) /
             2)) * 100)
}

gpredictors0 <- c("ALB", "ALP", "ALT", "AST", "BIL", "CHE",
                  "CHOL", "CREA", "GGT", "PROT")
results_list0 <- lapply(gpredictors0, calculate_diff0)
names(results_list0) <- paste0(gpredictors0, "_percent_diff")
skew_p_diff0 <- t(round((do.call(rbind, results_list0)),2))
skew_p_diff0
```

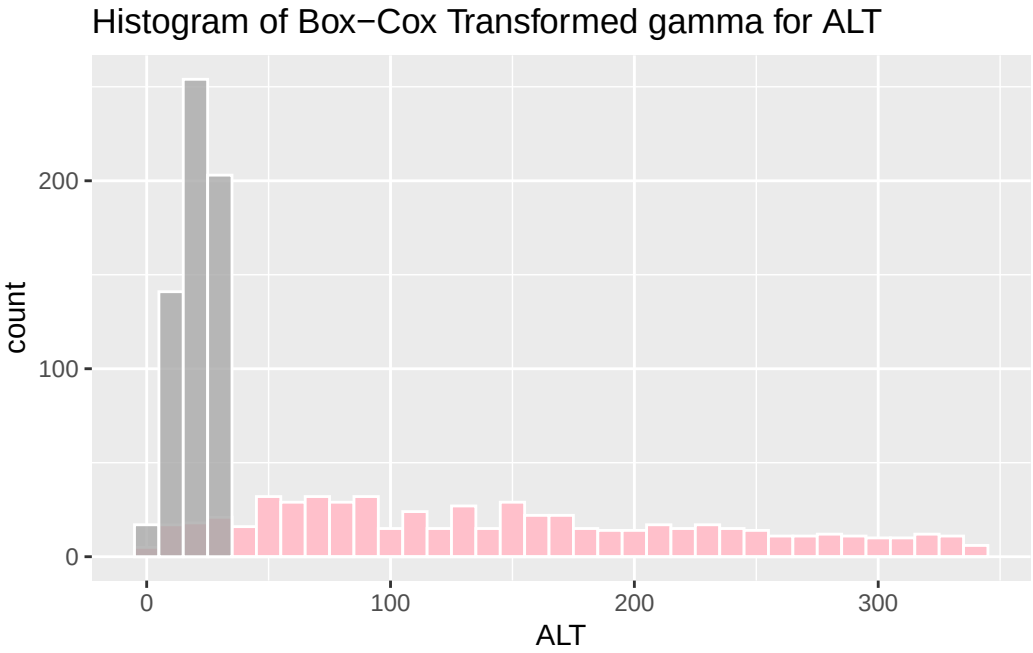
```
     ALB_percent_diff ALP_percent_diff ALT_percent_diff AST_percent_diff
[1,]                0                0              600           577.78
     BIL_percent_diff CHE_percent_diff CHOL_percent_diff CREA_percent_diff
[1,]                0                0                 0                40
     GGT_percent_diff PROT_percent_diff
[1,]           786.67                 0
```

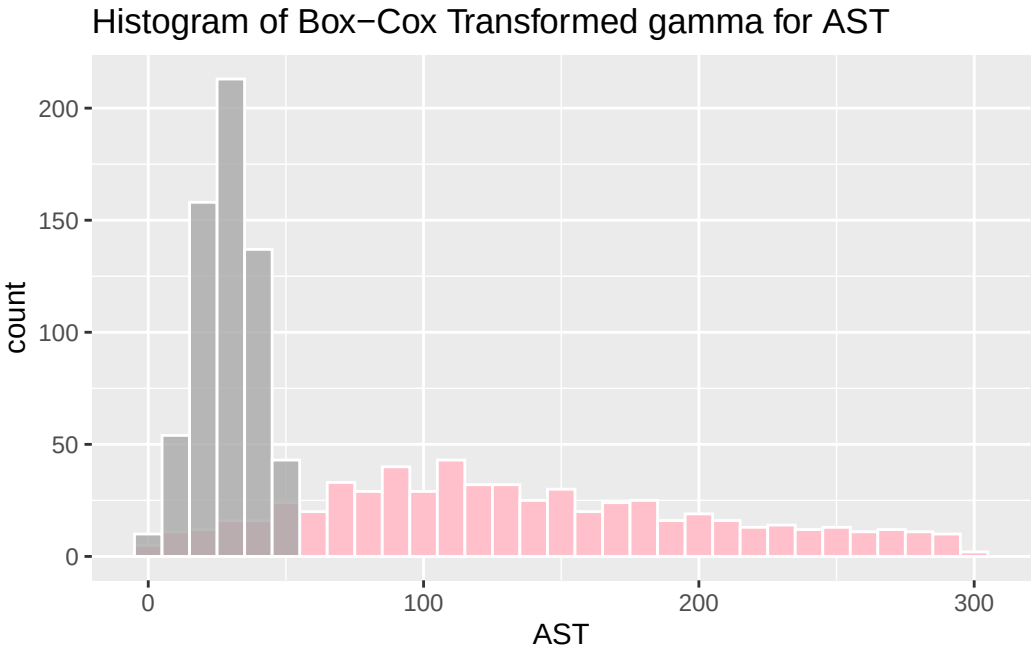
Lab tests that benefited the most through Box-Cox transformation are ALT, AST, ant GTT

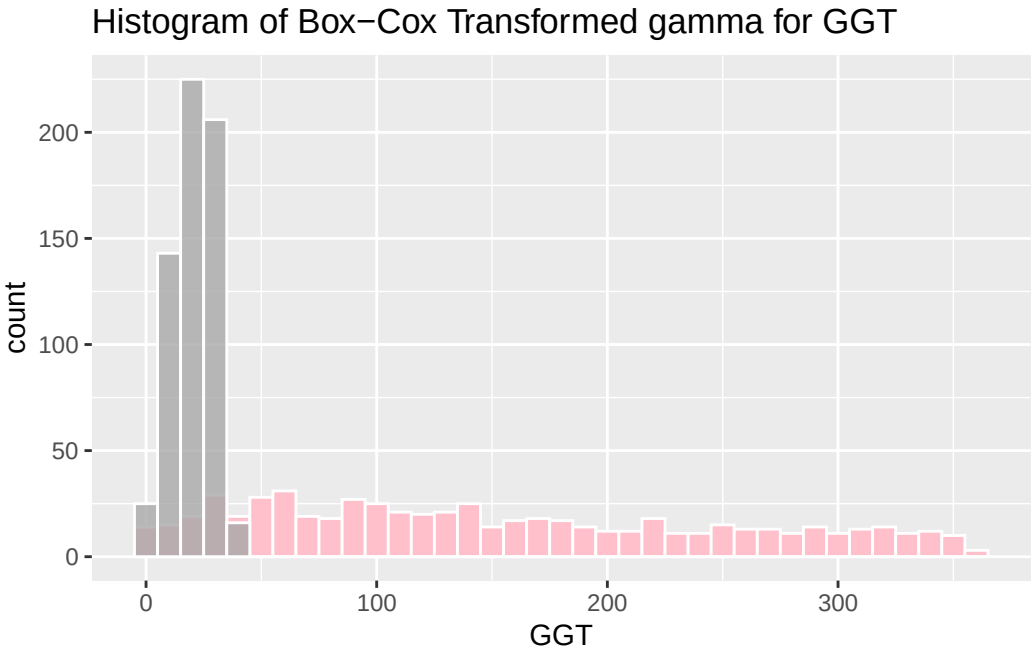
base off of greater than 50% diff criteria. It seems like Box-Cox transformation tremendously improved the three predictors based on their % diff gain.

3.4.f. Plot histograms of transformed predictors to observe changes to skewness visually. (5 points)

```
# Create histogram loop:
plot_histogram0 <- function(hcv_imputed, predictor0) {
  print(
  ggplot(hcv_imputed, aes(x = !!sym(predictor0))) +
    geom_histogram(bins = 10, binwidth = 10, fill = "pink",
                   color = "white") +
    geom_histogram(bins = 10, aes(x = !!sym(
      paste("gamma_", predictor0, sep = ""))),
                   binwidth = 10, fill = "darkgray",
      color = "white", alpha = .8) +
    labs(title = paste("Histogram of Box-Cox Transformed gamma for",
                       predictor0))
  )
}

# Run predictors through the loop:
# Only Ca and Al have more than 50% percent diff for skewness.
predictor_columns0 <- c("ALT","AST", "GGT")
for (predictor0 in predictor_columns0) {
  plot_histogram0(hcv_imputed, predictor0)
}
```


Histogram of Box−Cox Transformed gamma for ALT
count
200
100
0
0
100
200
300
ALT


Histogram of Box−Cox Transformed gamma for AST
count
200
150
100
50
0
0
100
200
300
AST



Looking at the visualizations, the Box-Cox transformation to ALT, AST, and GGT overall help mitigate the skewness of the each lab test distribution. The process condensed the distribution and somewhat formed a classic bell shape curve as a sign of being normally distributed.