Week1, Assignment 1

Your Name

```
library(tidyverse)
library(GGally)
library(e1071)
library(caret)
```

Problem 3.1 (30 points)

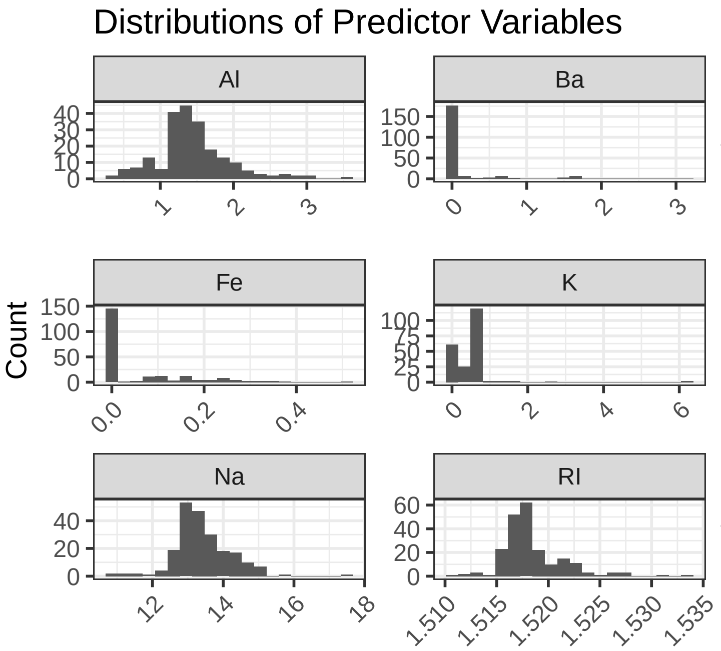
The UC Irvine Machine Learning Repository contains a Glass Identification Data Set. The data consist of 214 glass samples labeled as one of seven class categories. There are nine predictors, including the refractive index and percentages of eight elements: Na, Mg, Al, Si, K, Ca, Ba, and Fe. The data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

```
library(mlbench)
data(Glass)
```

3.1.a (10 points)

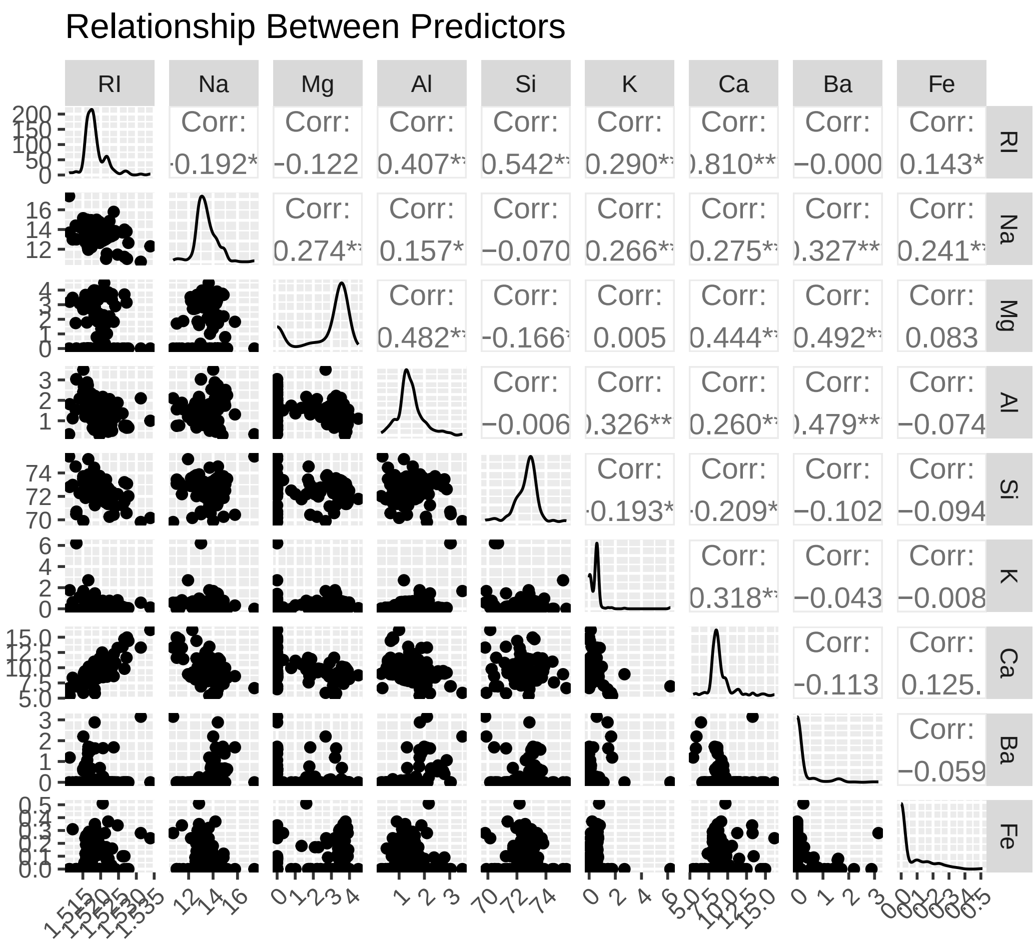
Using visualizations, explore the predictor variables to understand their distributions …

```
Glass |>
    pivot_longer(-Type, names_to = 'Element', values_to = 'value') |>
    ggplot(aes(x=value)) +
    facet_wrap(~Element, scales = "free", ncol = 3) +
    geom_histogram(bins = 20) +
    theme_bw() +
    labs(title = "Distributions of Predictor Variables", x = NULL, y = "Count") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



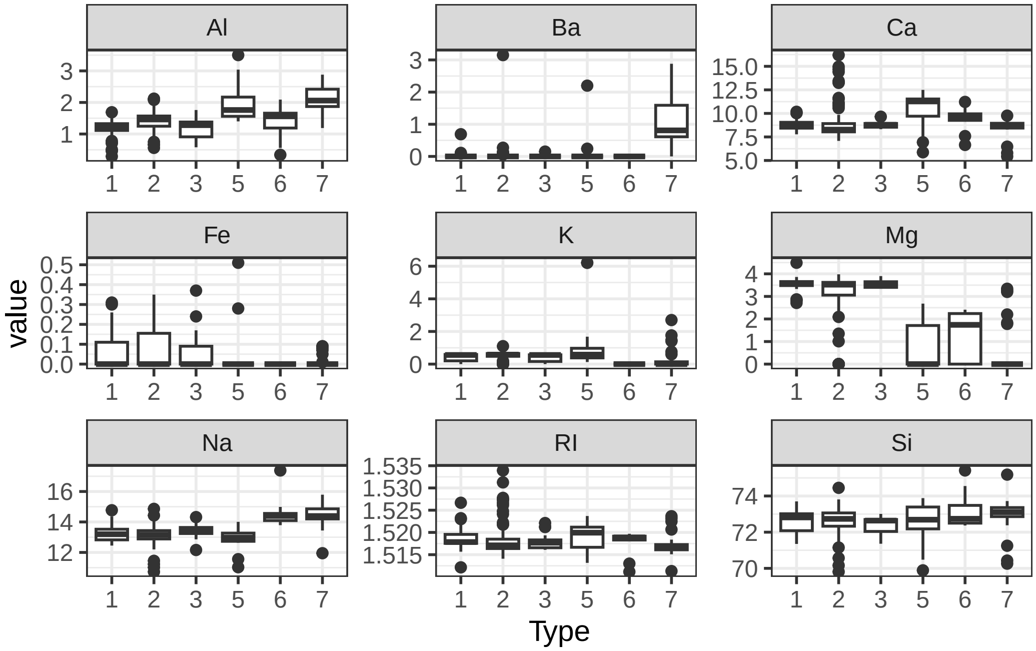
as well as the relationships between predictors.

```
Glass |>
    select(-Type) |>
    ggpairs(title = "Relationship Between Predictors", progress = FALSE)+
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Explore the relationship between predictors and response.

```
Glass |>
    pivot_longer(-Type, names_to = 'Element', values_to = 'value') |>
    ggplot(aes(Type, value)) +
    geom_boxplot() +
    facet_wrap(~Element, ncol = 3, scale = 'free') +
    theme_bw()
```



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

Ba, Mg, and Fe may be good predictors because the box plots show significant separation between the categories of the response variable. It is hard to say definitively from the plots how the predictors will inform a multivariate model.

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

```r
cor(Glass[1:9], as.numeric(Glass$Type))
```

```
           [,1]
RI -0.168739357
Na  0.506424080
Mg -0.728159518
Al  0.591197598
Si  0.149690687
K  -0.025834560
Ca -0.008997841
Ba  0.577676375
Fe -0.183206747
```

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

High correlations (> |0.7|) will likely be good predictors

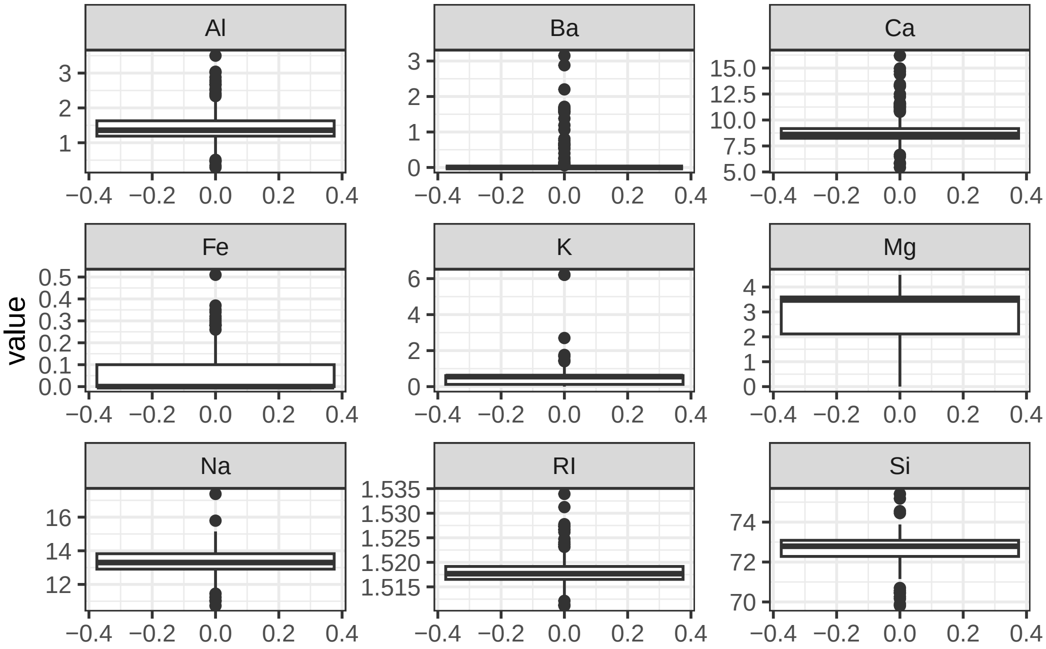
High: Mg (Negative)
Medium: Na (Positive), Al (+), Ba (+)
Low: RI (-), K (-), Ca (-), Fe (-)

3.1.b (10 points)

Are there any outliers in the data?

Based on the boxplots above all of the predictors show some outliers for at least one response category.
We can also boxplot the predictors as a complete set. `Mg` *is the only predictor that doesn't exhibit outliers.*

```
Glass |>
    pivot_longer(-Type, names_to = 'Element', values_to = 'value') |>
    ggplot(aes(y = value)) +
    geom_boxplot() +
    facet_wrap(~ Element, scale = 'free') +
    theme_bw()
```



Are any predictors skewed?

Based on the histograms above, all of the predictors show some skew. Most are right skewed except for `Mg` and `Si`, which are left skewed. `Na` exhibits the least skew while `K` is the most skewed. We can also calculate the skew for each predictor.

```
skew_glass <- Glass |>
    select(-Type) |>
    map_dbl(skewness)
skew_glass |> round(3)
```

```
    RI     Na     Mg     Al     Si      K     Ca     Ba     Fe 
 1.603  0.448 -1.136  0.895 -0.720  6.460  2.018  3.369  1.730 
```

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.

Hints:

- Skew values less than ±0.5 should be considered 'normal enough'
- Use 'caret::BoxCoxTrans()' with appropriate adjustments for non-positive values.
- Use transformations that improve skew by >50%

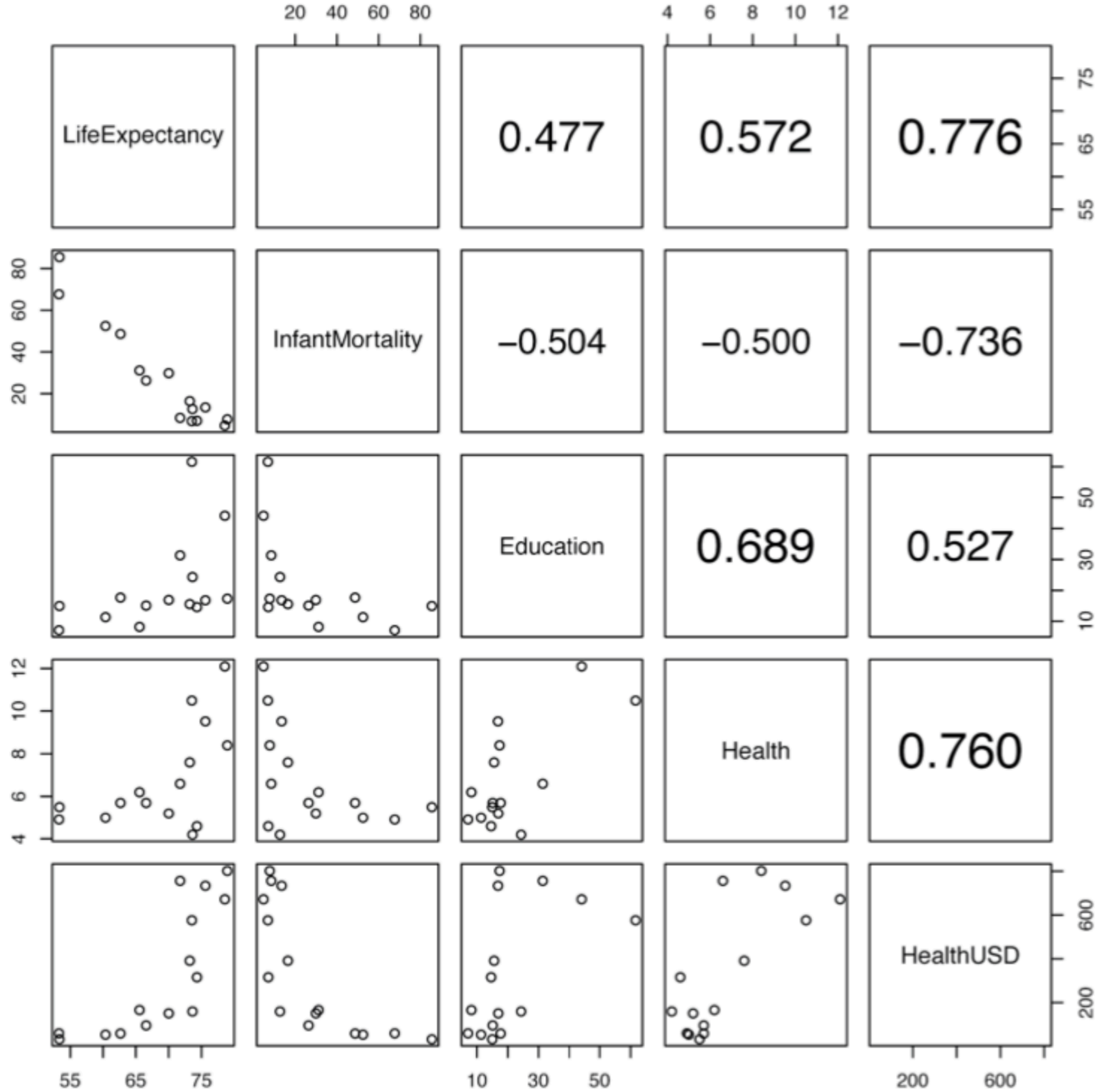
```
# write a function to apply a BoxCox transformation and compare skewness
bct_test <- function(x, property = 'skew') {
    stopifnot(property %in% c('skew', 'lambda'))
    x <- x[which(! is.na(x))] #hack to remove NA (so we can re-use in 3.4)
    x2 <- x + ifelse(any(x == 0), 0.0001, 0) #handle 0's
    bct <- BoxCoxTrans(x2)
    x_trans <- predict(bct, x2 )
    if (property == 'skew') return(e1071::skewness(x_trans))
    return(bct$lambda)
}

skew_glass_bct <- Glass |>
    select(-Type) |>
    map_dbl(bct_test)

bct_analysis <- tibble(
    Property = names(skew_glass),
    `Original Skew` = skew_glass,
    `Skew after BoxCox`= skew_glass_bct,
    Lambda = Glass |> select(-Type) |> map_dbl( ~ bct_test(.x, 'lambda'))
)

bct_keep <- bct_analysis |>
    filter(abs(`Original Skew`) > 0.5,
           abs(`Skew after BoxCox`) < 0.5)

bct_keep |>
    gt::gt()
```

Property	Original Skew	Skew after BoxCox	Lambda
Al	0.8946104	0.091058992	0.5
K	6.4600889	0.009536743	0.4
Ca	2.0184463	-0.193955732	-1.1

Grading Notes: The original key use high reductions in skewness as the criteria, so also included Ba and Fe. These should be accepted as well if they are justified by the student.

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.



Hint: There should be 10 combinations [$n(n-1)/2$] . The plot is missing a correlation coefficient for one – estimate what it is.

Example: *Var1 has a strongpositive correlation with Var2 of 0.XXX*

1. LifeExpectancy and InfantMortality have a strong negative correlation (~ -1)
2. LifeExpectancy and Education have a moderate positive correlation of 0.48
3. LifeExpectancy and Health have a moderate positive correlation of 0.57
4. LifeExpectancy and HealthUSD have a strong positive correlation of 0.78

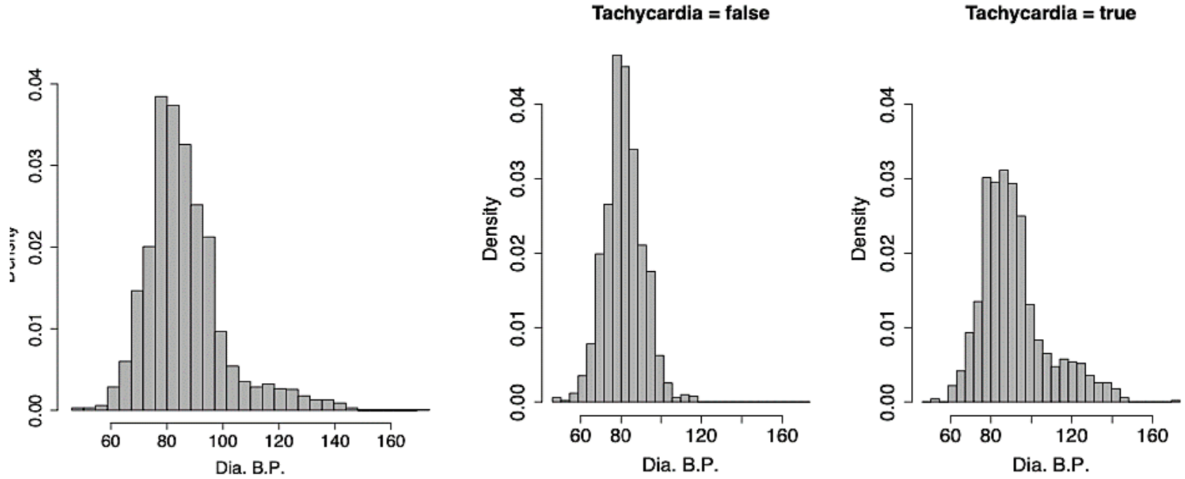
5. InfantMortality and Education have a moderate negative correlation of -0.50
6. InfantMortality and Health have a moderate negative correlation of -0.50
7. InfantMortality and HealthUSD have a strong negative correlation of -0.74
8. Education and Health have a strong positive correlation of 0.69
9. Education and HealthUSD have a strong positive correlation of 0.53
10. Health and HealthUSD have a strong positive correlation of 0.76

Problem 3.3 (10 points)

Discuss the relationships between the variables shown in below visualizations:

3.3.a (5 points)

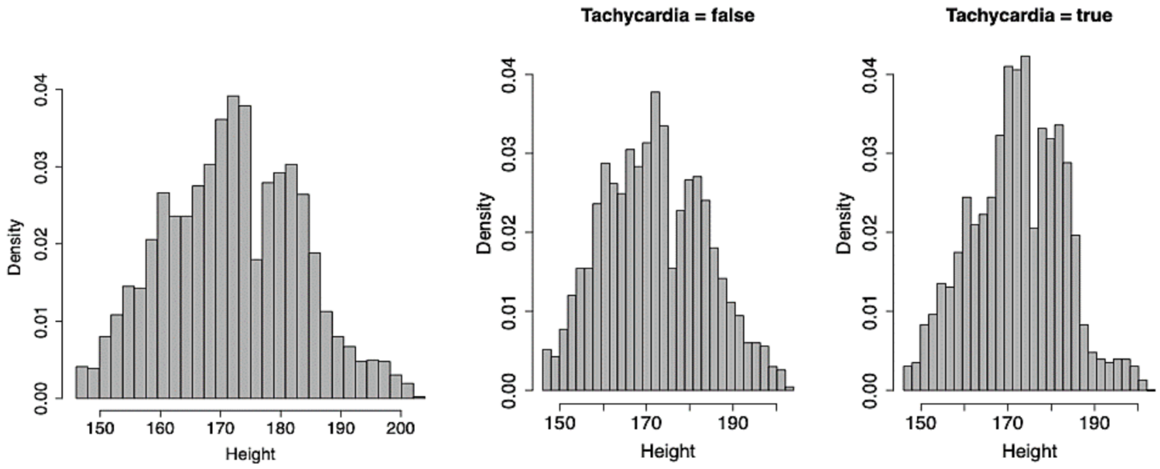
The visualization below illustrates the relationship between Diastolic BP and Tachycardia, left most plot has data where Tachycardia = true and false (the full study population).



Since the distributions of Diastolic BP with (range from 50 to 120, median of 75) and without (range from 50 to 150, median of 90) Tachycardia are quite different, we can infer that Tachycardia and Diastolic BP are related and patients with higher Diastolic BP have higher chance of having Tachycardia than patients with lower Diastolic BP.

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.



The distributions of height are very similar between patients with Tachycardia and patients without Tachy- cardia, they have similar range and median. We can infer that Tachycardia and height are not related based on these distributions.

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:

```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)

```r
missing_cnt <- map_int(hcv, ~ sum(is.na(.x)))
missing_names <- names(missing_cnt)[which(missing_cnt > 0)]
```

Yes, the following predictors have missing values: ALB, ALP, ALT, CHOL, PROT

3.4.b. Summarize the missing data by each predictor. (5 points) Hint: Use `purrr::map_int`, or `lapply`

```
missing_cnt
```

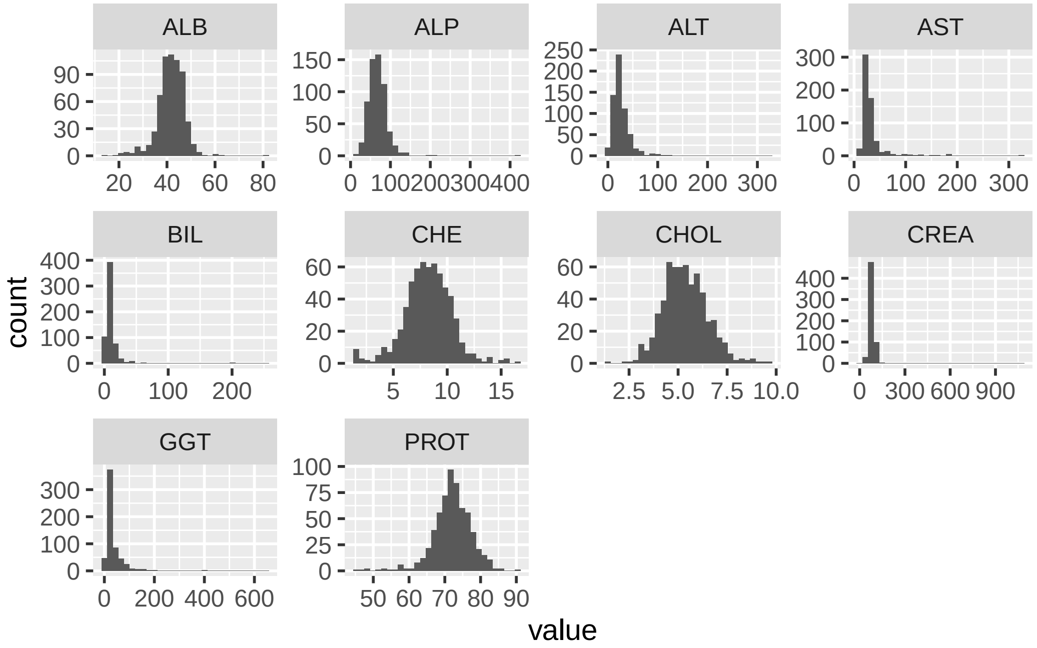
```
 ALB  ALP  ALT  AST  BIL  CHE CHOL CREA  GGT PROT 
   1   18    1    0    0    0   10    0    0    1 
```

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

```
hcv |>
    pivot_longer(cols = everything(), values_to = 'value', names_to = 'predictor') |>
    ggplot(aes(value)) +
    geom_histogram() +
    facet_wrap(~ predictor, scales = 'free')
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 31 rows containing non-finite values (`stat_bin()`).
```



Visually, ALP, ALT, AST, BIL, CREA, and GGT look right-skewed. PROT is slightly left-skewed.

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)

```
hcv |>
    map_dbl(~ e1071::skewness(.x, na.rm = TRUE))
```

```
       ALB        ALP        ALT        AST        BIL        CHE       CHOL
-0.1759048  4.6315552  5.4792399  4.9162540  8.3445765 -0.1096956  0.3739660
      CREA        GGT       PROT
15.0953748  5.6052871 -0.9589839
```

Yes. The skewness values align with the visual observations from part c.

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)

```
#reusing function from above
bct_test
```

```
function(x, property = 'skew') {
    stopifnot(property %in% c('skew', 'lambda'))
    x <- x[which(! is.na(x))] #hack to remove NA (so we can re-use in 3.4)
    x2 <- x + ifelse(any(x == 0), 0.0001, 0) #handle 0's
    bct <- BoxCoxTrans(x2)
    x_trans <- predict(bct, x2 )
    if (property == 'skew') return(e1071::skewness(x_trans))
    return(bct$lambda)
}
<bytecode: 0x7fa4b4aa5120>
```

```
hcv |>
    map_dbl(~ bct_test(.x))
```

```
        ALB         ALP         ALT         AST         BIL         CHE 
 0.32090260 -0.21816588 -0.42694320  0.05826619  0.05365506  0.17614896 
       CHOL        CREA         GGT        PROT 
 0.04167873  0.64684828  0.07373092 -0.45312501 
```

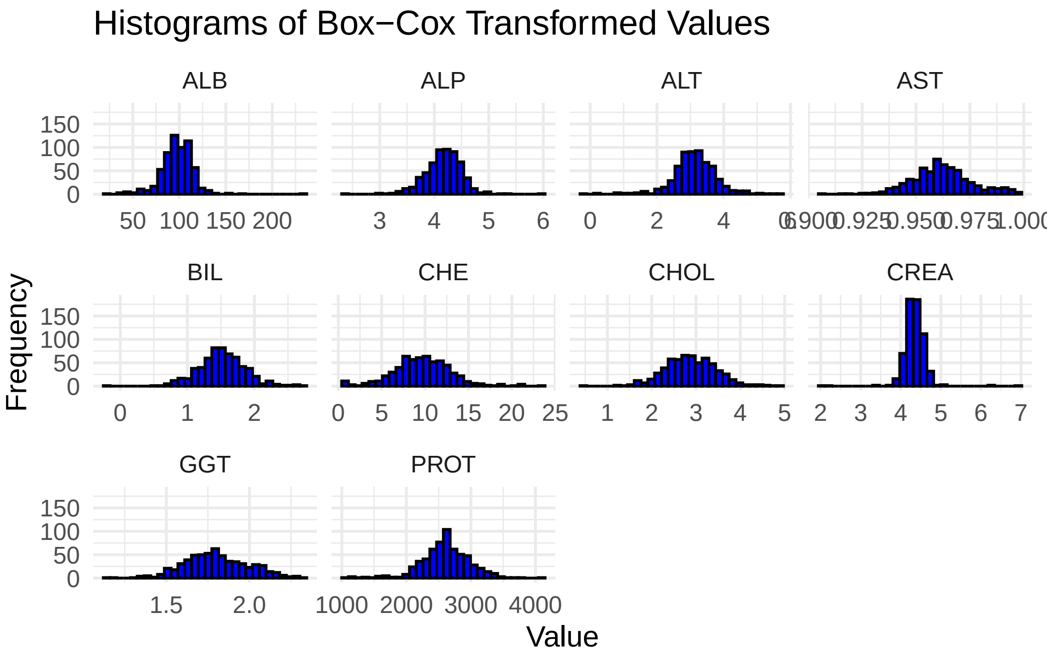
Based on above results BoxCox transformation helped reduce skewness of heavily skewed distributions.

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

```
bc_trans <- function(x) {
    if (any(x == 0, na.rm = TRUE)) x <- x + 0.0001
    bct <- BoxCoxTrans(x, na.rm = TRUE)
    predict(bct, x )
}
```

```
hcv_bct <- hcv |>
    map(bc_trans) |>
    enframe(name = "VectorID", value = "Values") %>%
    unnest(Values)
hcv_bct |>
    ggplot(aes(Values)) +
    geom_histogram(bins = 30, fill = "blue", color = "black") +
    facet_wrap(~ VectorID, scales = "free_x") +
    theme_minimal() +
    labs(x = "Value", y = "Frequency", title = "Histograms of Box-Cox Transformed Values")
```

```
Warning: Removed 31 rows containing non-finite values (`stat_bin()`).
```



It is clear visually that all the distributions look approximately symmetric and that BoxCox transformation helped reduce skewness.