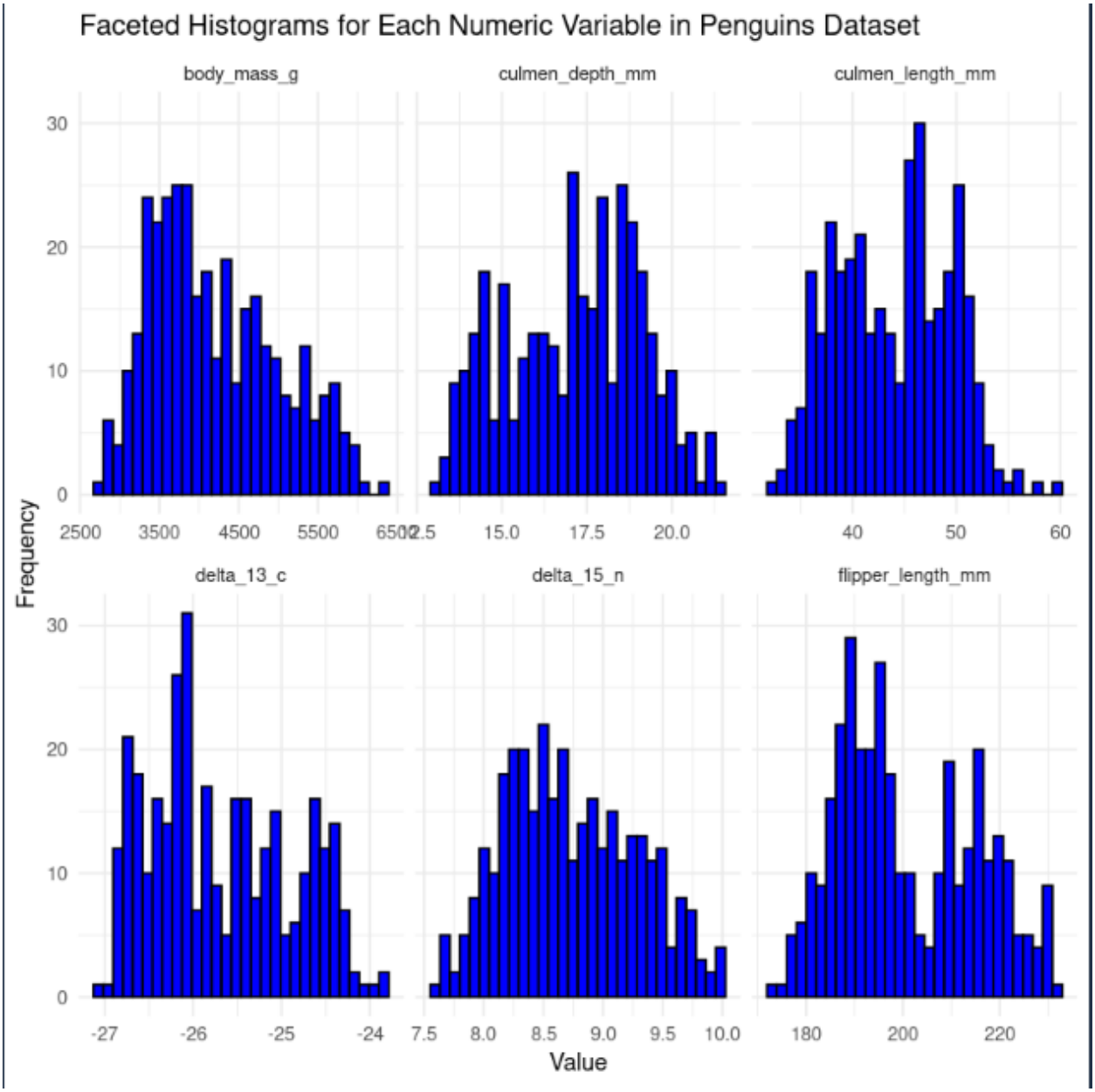
Statistical Evaluation: `palmerpenguins`

First, we examined the distributions of all of the numeric variables in the dataset: `body_mass_g`, `culmen_depth_mm`, `culmen_length_mm`, `delta_13_c`, `delta_15_n`, and `flipper_length_m`. Applying the Empirical Rule to any of these variables requires that the variable be approximately symmetrical in its distribution. None of these variables appear to meet this criterion.



Body Mass (g) Histogram

The histogram for body mass appears to be roughly unimodal and slightly right-skewed, indicating that most penguins have a body mass around the center of the distribution with fewer penguins having higher body masses. The empirical rule, which states that approximately 68%, 95%, and 99.7% of the data fall within one, two, and three standard deviations from the mean, respectively, assumes a normal distribution. Given the skewness observed, the empirical rule may not strictly apply here.

Culmen Depth (mm) Histogram

The culmen depth shows a bimodal distribution with two distinct peaks. This suggests there are possibly two different subgroups within the penguin population with different average culmen depths. Since the distribution is not normal and has multiple peaks, the empirical rule would not be appropriate for this variable.

Culmen Length (mm) Histogram

Similar to culmen depth, the culmen length is also bimodal, which suggests the presence of two subgroups with different culmen lengths. The bimodal nature again indicates that the empirical rule would not be appropriate here.

Delta 13 C (‰) Histogram

The distribution of delta 13 C appears multimodal with multiple peaks, which might indicate the presence of different dietary or environmental subgroups within the penguins. Multimodal distributions are not suitable for the application of the empirical rule, as it does not follow a normal distribution pattern.

Delta 15 N (‰) Histogram

This histogram is also multimodal, similar to delta 13 C, which suggests complexity in the data that might be related to different feeding habits or habitats. The empirical rule cannot be applied due to the lack of a normal distribution.

Flipper Length (mm) Histogram

The histogram for flipper length appears unimodal with a slight skew. There is one prominent peak, and the data tails off as flipper length increases. While there is a single peak, the skewness might still pose an issue for the empirical rule, although this variable might be the best candidate for its application compared to the others if one considers applying some skewness correction.

In summary, the empirical rule is most useful for data that are symmetrically distributed and bell-shaped (normal distribution). In these histograms, many of the variables show signs of bimodal or multimodal distributions, which suggests that

different subpopulations exist within the data. The empirical rule does not apply well to such distributions.

Chebyshev’s Rule Applied to Culmen Length (mm)

Chebyshev's Rule, also known as Chebyshev's Inequality, is a statistical rule that provides a lower bound on the probability that a random variable lies within a certain number of standard deviations from the mean. Specifically, for any real number $K>1$, at least $(1-1/k^2$)f the values lie within (k) standard deviations of the mean. This rule applies to any probability distribution, regardless of its shape. *By Chebyshev's Rule, at least 75 % of the penguins' culmen lengths are between 33 mm and 54.8 mm.*

Outliers

The R script evaluated Z-scores for “flipper_length_mm”. “culmen_depth_mm”, and “body_mass_g.” None of these variables had any outliers +/- 3 standard deviations from the mean.

Coefficient of Variation Comparison

The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. It is calculated as the ratio of the standard deviation to the mean, and it is often expressed as a percentage. The CV is useful because it allows for comparison of the variability of different datasets with different units or means. We will calculate the CV for two or more numerical variables from the `penguins_numeric` dataset and compare their variability.

The coefficients of variation (CV) for the variables `culmen_length_mm`, `flipper_length_mm`, and `body_mass_g` from the `penguins_numeric` dataset are 12.43%, 6.999%, and 19.086% respectively. Here's what these values indicate about the variability of each measurement:

- Culmen Length (mm): With a CV of 12.43%, this suggests that the culmen length measurements have a moderate level of variability relative to the mean culmen length. This means that while there is some variation in culmen length among the penguins, it is not excessively high.
- Flipper Length (mm): The CV of 6.99% for flipper length indicates that this variable has relatively low variability compared to its mean. This implies that the flipper lengths of the penguins are quite consistent, with less variation from the average flipper length.
- Body Mass (g): The body mass has a CV of 19.086%, which is higher than the other two variables. This higher CV suggests that there is a greater level of

variability in the body mass of the penguins. In other words, the weights of the penguins are more spread out from the average weight, indicating a wider range of body mass values within the dataset.

In summary, the CV provides a standardized measure of variability that is independent of the unit of measurement. By comparing the CVs, we can conclude that the body mass of the penguins is the most variable trait among the three, while flipper length is the least variable.