Colin Horgan

Stream Data Analysis Project

***Data Visualization and Standardization***

After importing Stream\_data into RStudio, I first looked at the summary statistics for the raw data of the relevant variables (mayfly, caddisfly, stonefly, flies, and beetles). It was clear that the data was not standardized and so I attempted several transformations. I log­10-transformed the data (this provided better results than the use of a natural log) and performed a Tukey transformation on the data. The summary statistics are provided in Table 1 and the respective histograms in Figures 1, 2, and 3.

A close up of text on a white background

Description automatically generated

*Table 1: Summary statistics of relevant variables from Stream\_data.xslx. Raw data is shown as “Stream\_raw”, Log-transformed data is shown as “Stream\_log”, and Tukey-transformed data is shown as “Stream\_Tuk”.*

Clearly Stream\_log and Stream\_Tuk demonstrate significant improvements in standardization compared to Stream\_raw. However, both Stream\_log and Stream\_Tuk have their abnormalities. Stream\_log lacks very large irregularities in mean and standard deviation, however contains heavy skewness in positive and negative directions. Stream\_Tuk has very tight skewness (with the exception of stonefly), has closer standard deviations among groups compared to Stream\_log (with the exception of flies), but has higher ratios between means compared to Stream\_log. To get a better sense of which transformation to choose we will look at the data in histograms.

*Figure 1: Histogram plots of Stream\_raw.*

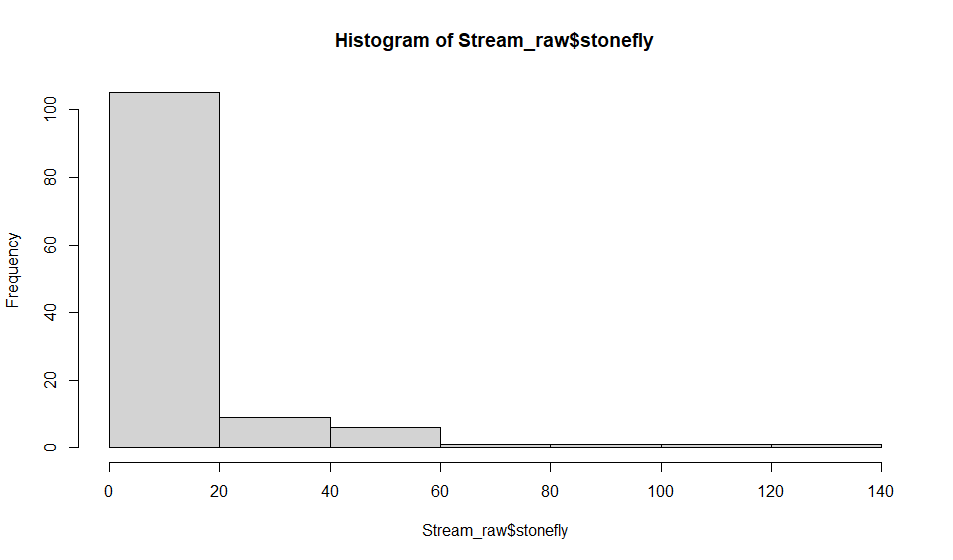
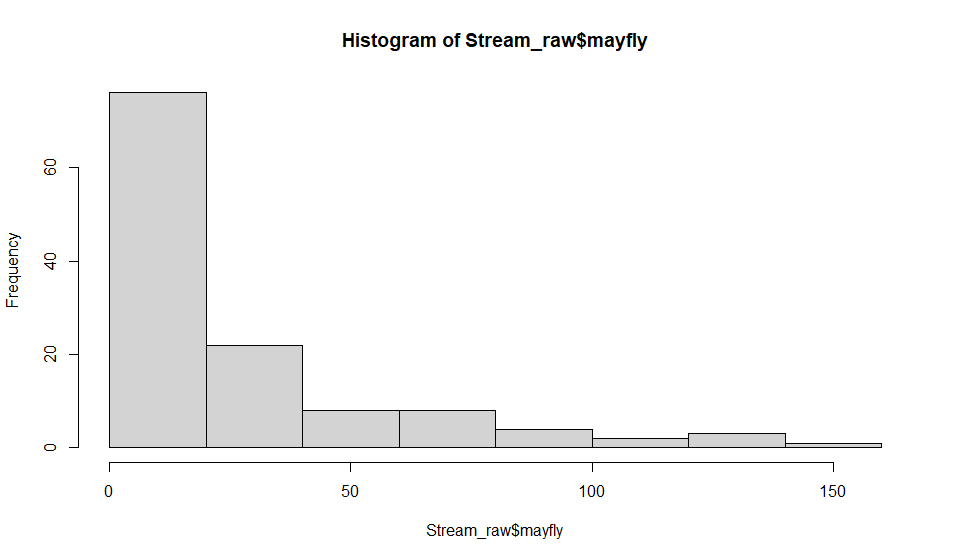
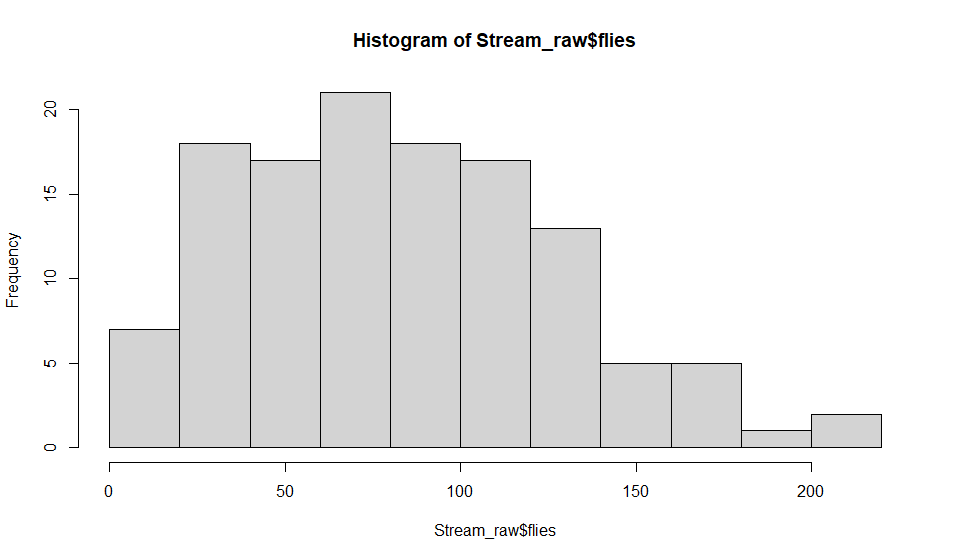
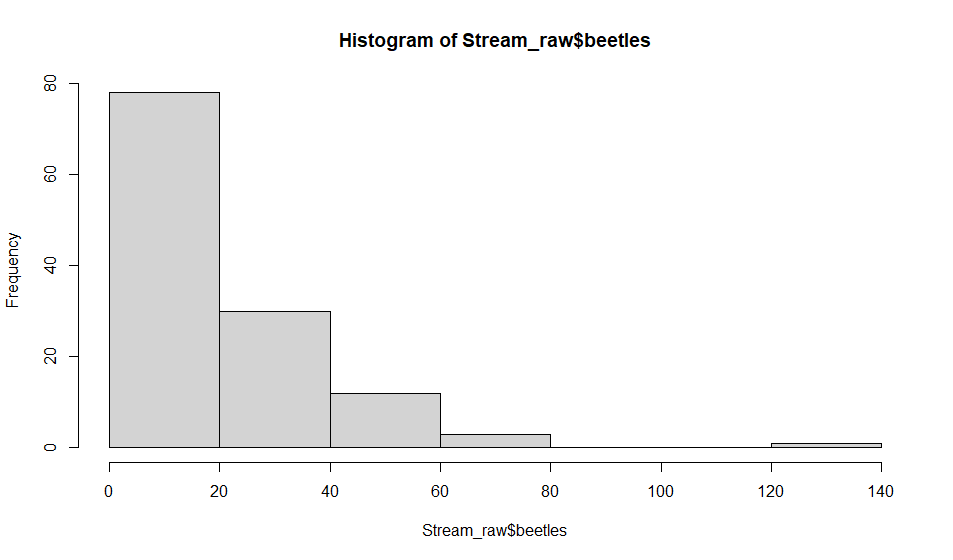
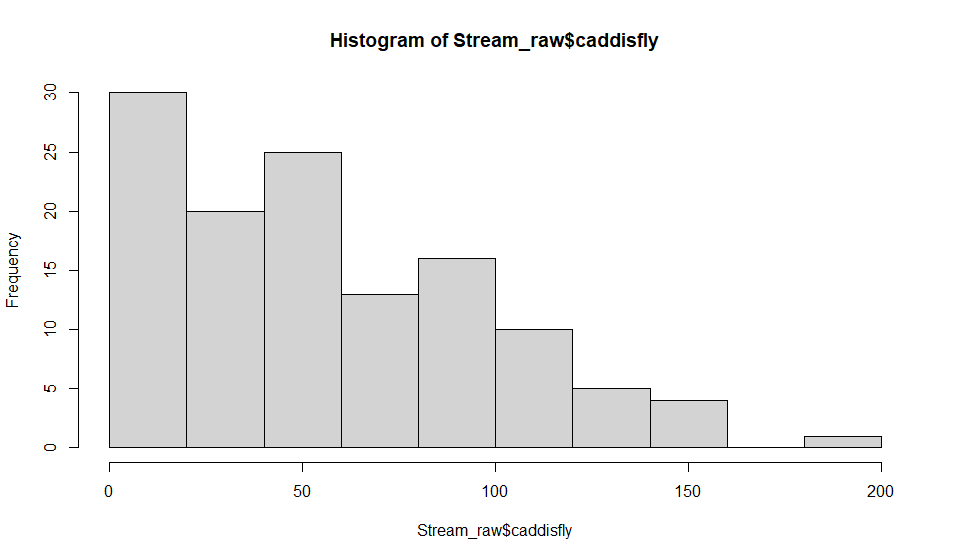
*A) mayfly,*

*B) caddisfly,*

*C) stonefly,*

*D) flies,*

*E) beetles.*



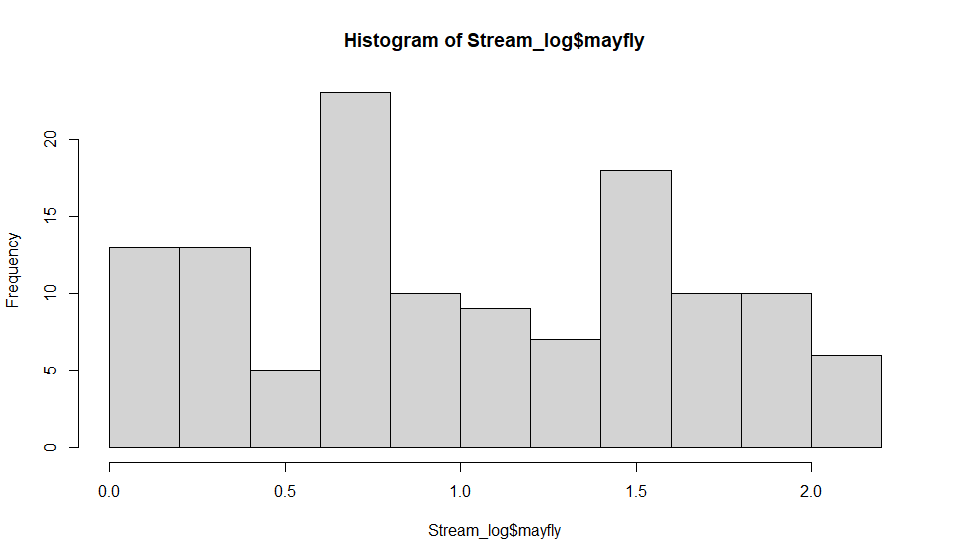
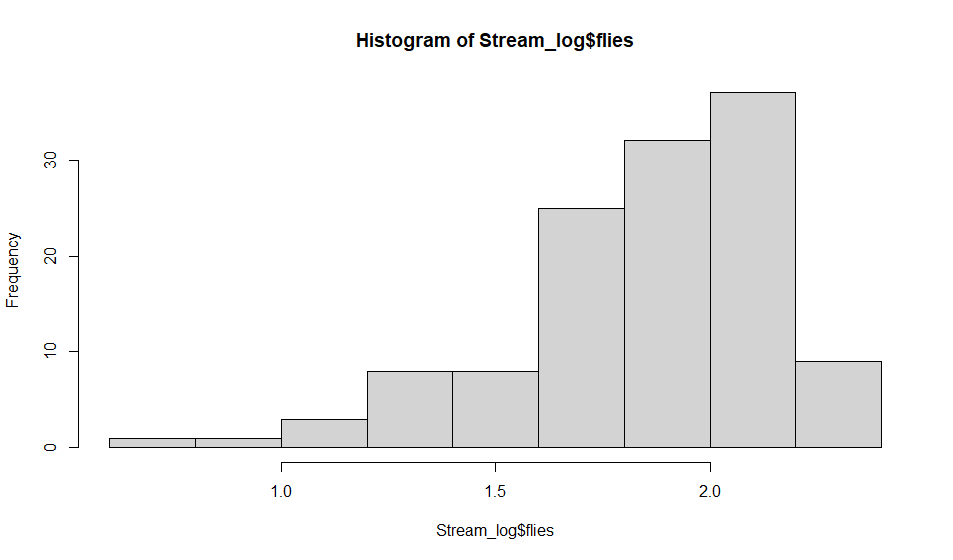
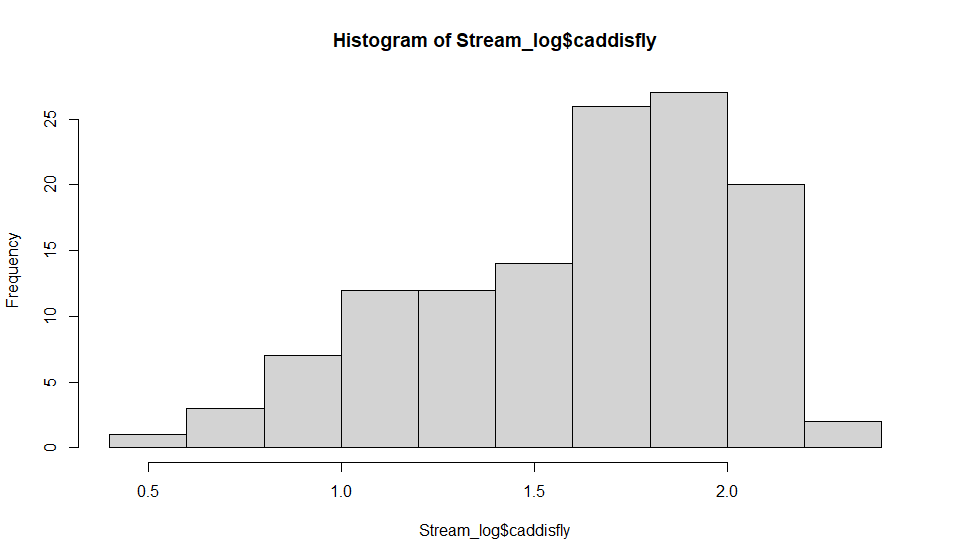
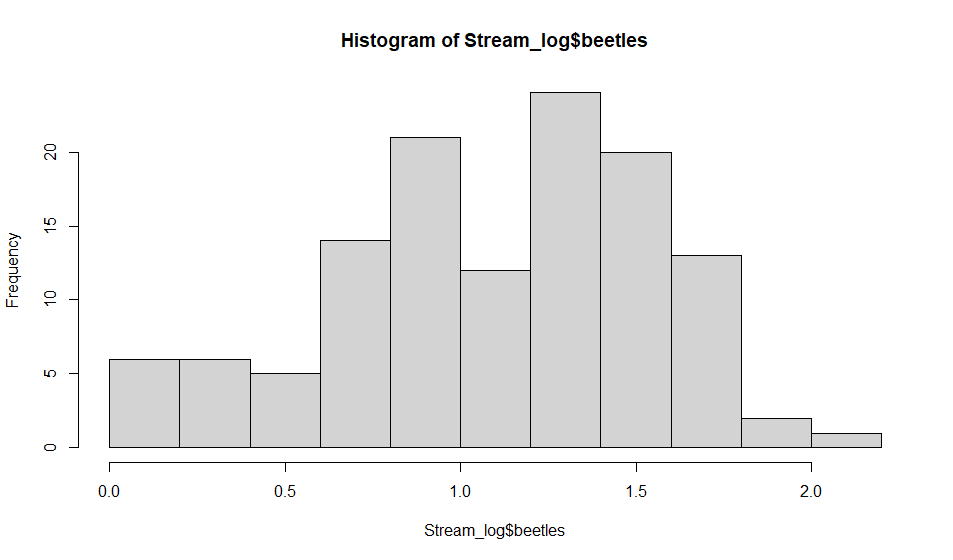
**A**

**B**

**C**

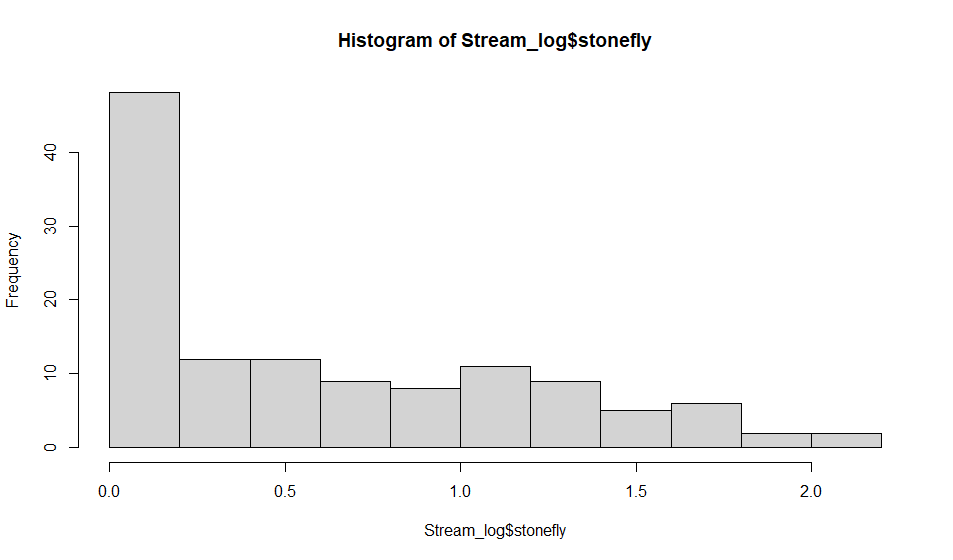
**D**

**E**



**A**

**B**



**C**

**D**

**E**

*Figure 2: Histogram plots of Stream\_log.*

*A) mayfly,*

*B) caddisfly,*

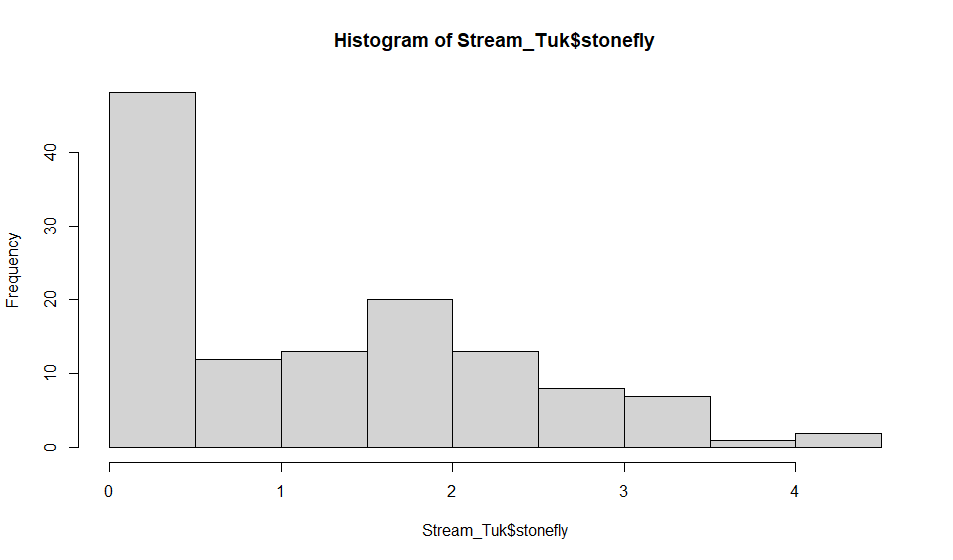
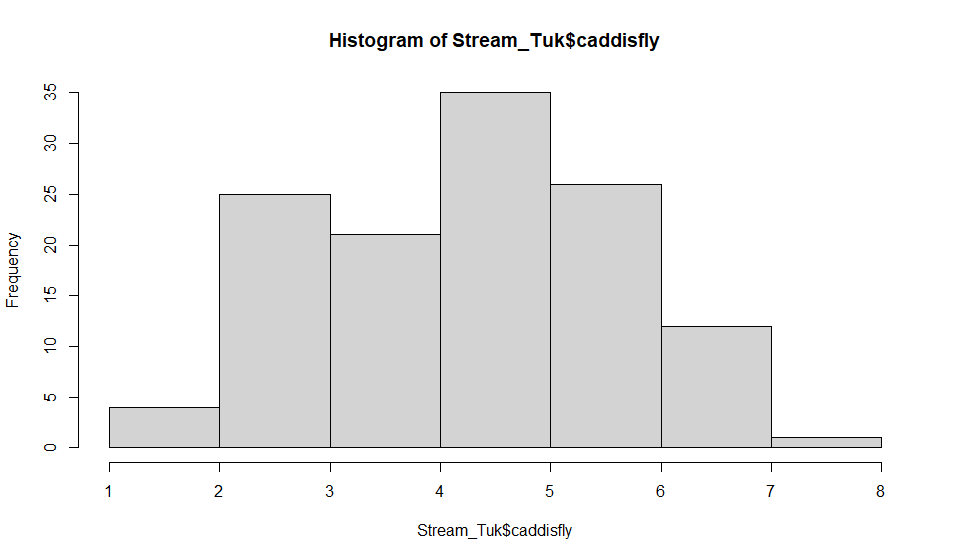
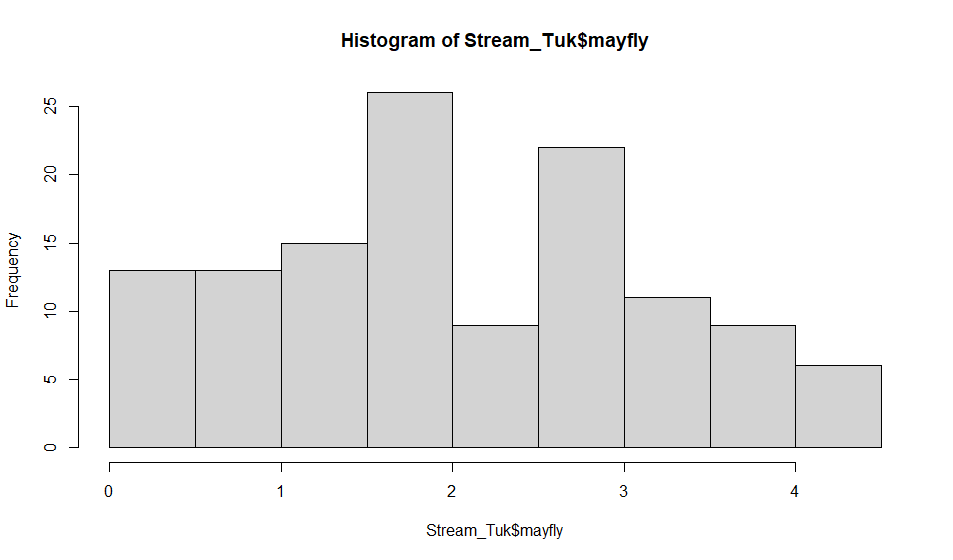
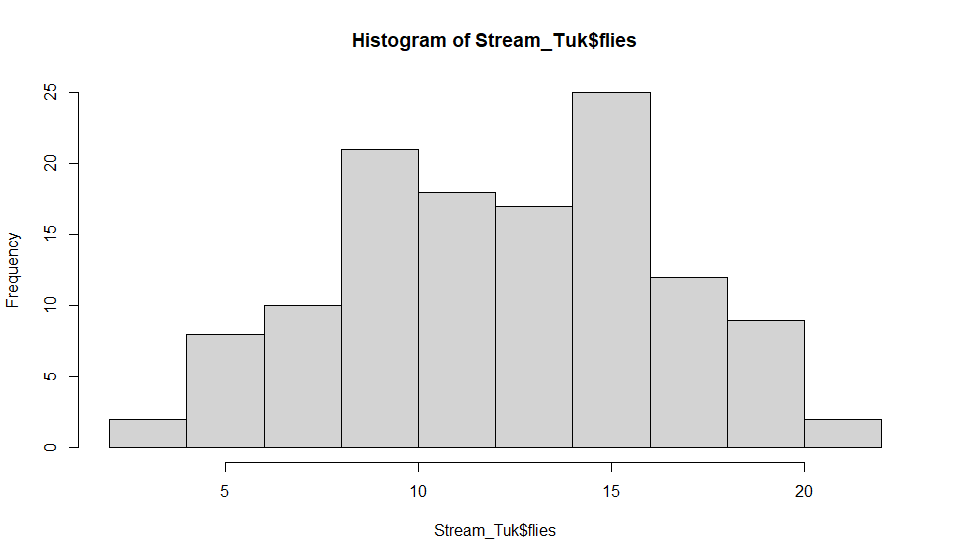
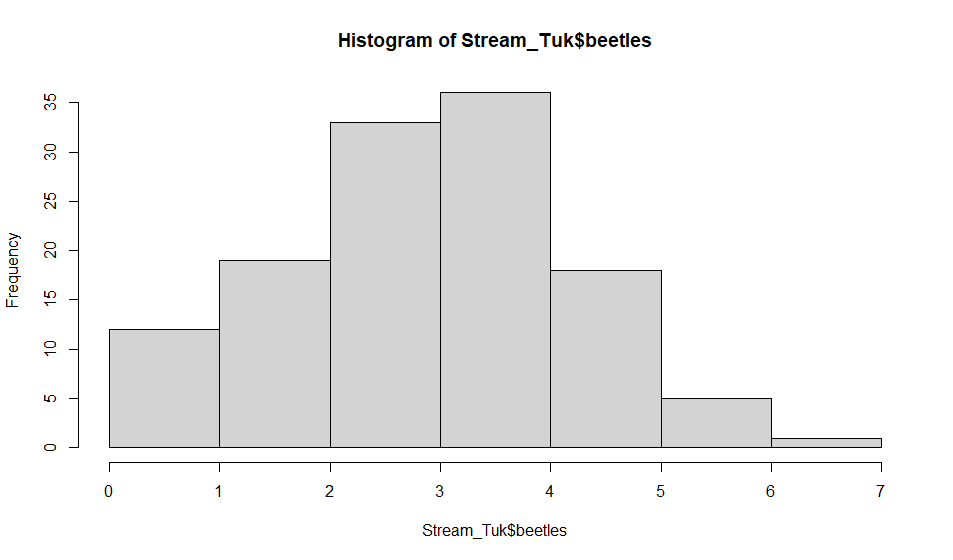
*C) stonefly,*

*D) flies,*

*E) beetles.*

*Figure 3: Histogram plots of Stream\_Tuk.*

*A) mayfly, B) caddisfly, C) stonefly, D) flies, E) beetles.*



**A**

**B**

**C**

**D**

**E**

Qualitatively speaking, the histograms from the Stream\_tuk dataset seem to be more normally distributed (Figure 3). Numerically, the summary data from the Stream\_log dataset seem to have more heterogeneity of variance but are more skewed. It’s clear both datasets present issues of normality, but moving forward I will use the Stream\_tuk dataset as it is qualitatively more normal, less skewed, and the standard deviations (with the exception of flies) are far tighter than that of Stream\_log.

**Analysis of the Impacts of Stream Quality and Diversity on Insect Species**

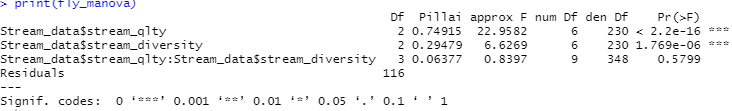
For Part B, we are concerned with the effect of stream quality and biodiversity on the number of mayflies, caddisflies, and stoneflies in the stream. Our Null hypothesis is that there is no effect.

Independent variables: Stream\_qlty, Stream\_diversity

Dependent variables: mayfly, caddisfly, stonefly

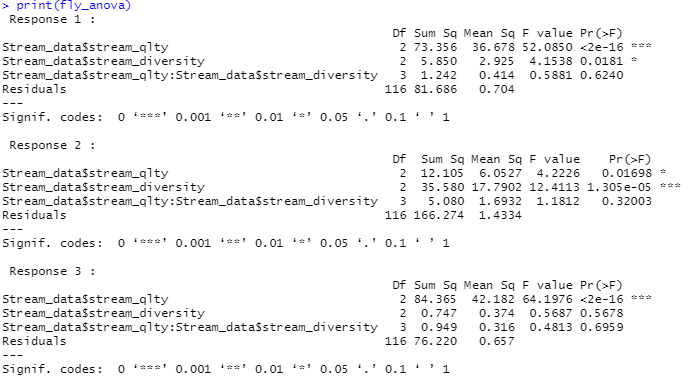
We can directly assess the interaction effects of our MANOVA by specifying a linear model where our dependent variables are dependent on the product of both Stream\_qlty and Stream\_diversity (notice that if there was an interaction between these two variables this would be true). The source code and results of this MANOVA are reported in Table 2. Given that the output of our MANOVA main effects are highly significant, we can reject the null hypothesis and proceed to multiple comparisons. Additionally, since our test returned a non-significant result for the interaction between Stream\_qlty and Stream\_diversity, we can conclude there is no interaction present.





*Table 2: Results of a 2 [Stream\_qlty, Stream\_diversity] x 3 [mayfly, caddisfly, stonefly] Manova returned a highly significant effect of stream\_qlty and stream\_diversity, but no qlty x diversity interaction.*

Next, we want to see which of our dependent variables are affected by the independent variables of Stream\_qlty and Stream\_diversity using ANOVA. The results of these tests are presented in Table 3. From the result of each of these tests, we can conclude that Stream\_qlty and Stream\_diversity effect the number of mayflies, caddisflies, and stoneflies. However, the number of mayflies and caddisflies are affected by both stream\_qlty and stream\_diversity whereas the number of stoneflies are only affected by stream\_qlty.



*Table 3: Results of ANOVA on each dependent variable. Response 1, 2, and 3 represent outputs from ANOVA on mayflies, caddisflies, and stoneflies – respectively.*