**Biostatistics**

**Lab 8**

*Tasks*

More dplyr (Mutate and Data Transformations)

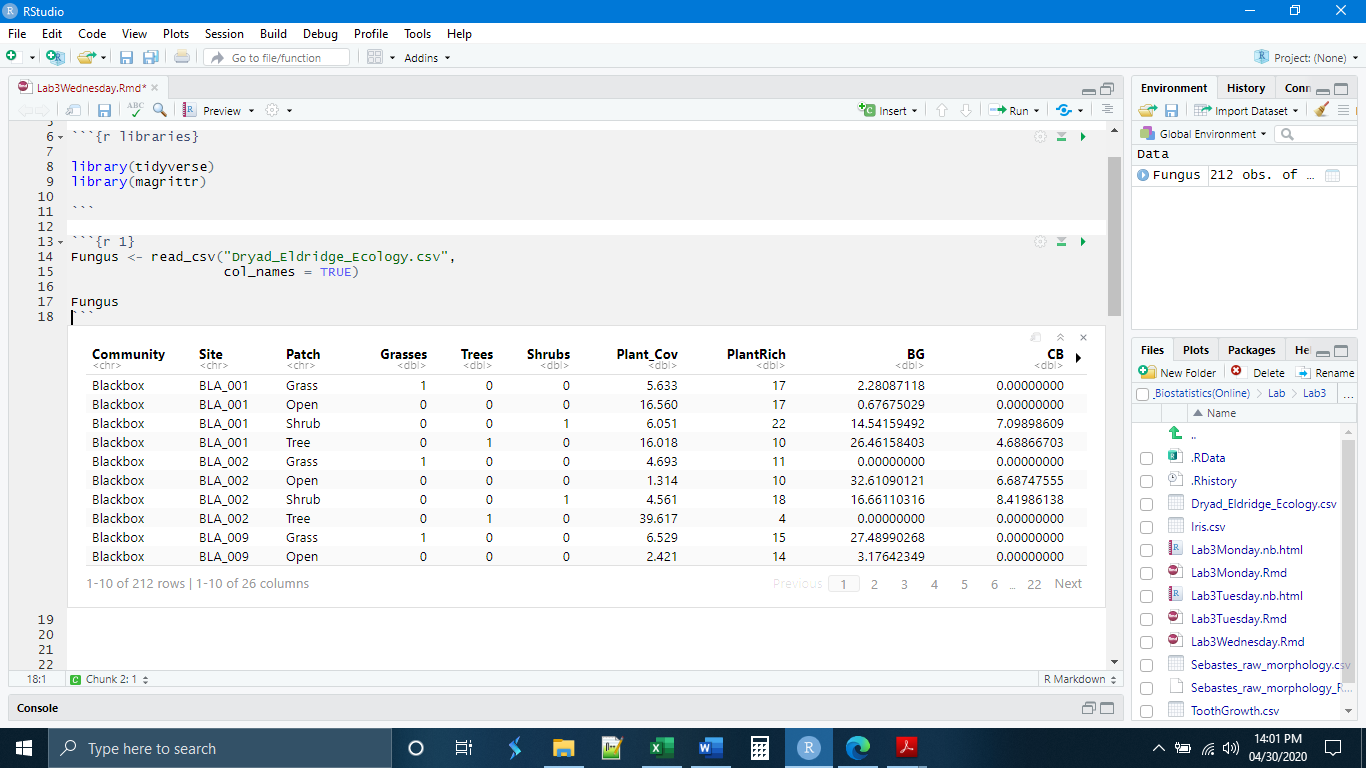
*Introduction*

Yesterday we went over t-Testing in R. We also went over ways to assess if the assumptions necessary for a proper implementation of the t-Test were met. Today will be a short lab in which we learn one technique for transforming our data in order to satisfy the assumption of normality. A data transformation is the application of a mathematical operation to all of the elements of a variable. A transformation can change the collective properties of a group of observations without actually changing their relative measurement. If we alter the collective properties of the observations, it may make certain mathematical analyses, such as t-Tests, possible. Today, we will learn how to do a logarithmic transformation.

*Set-Up and Data*

If you have not done so already, download from Blackboard all of the data files necessary for the lab this week and save them in the folder for today. The file that will be most important for today is the Dryad\_Eldridge\_Ecology.csv data.

Start R Studio and open a new R Notebook file. Save this file in today’s folder where the data files are. Delete the text that comes in the notebook template, and use the first chunk to load libraries. It is good practice to also present your session info, make a last chunk to do this. Also, read in the Dryad\_Eldridge\_Ecology.csv data.

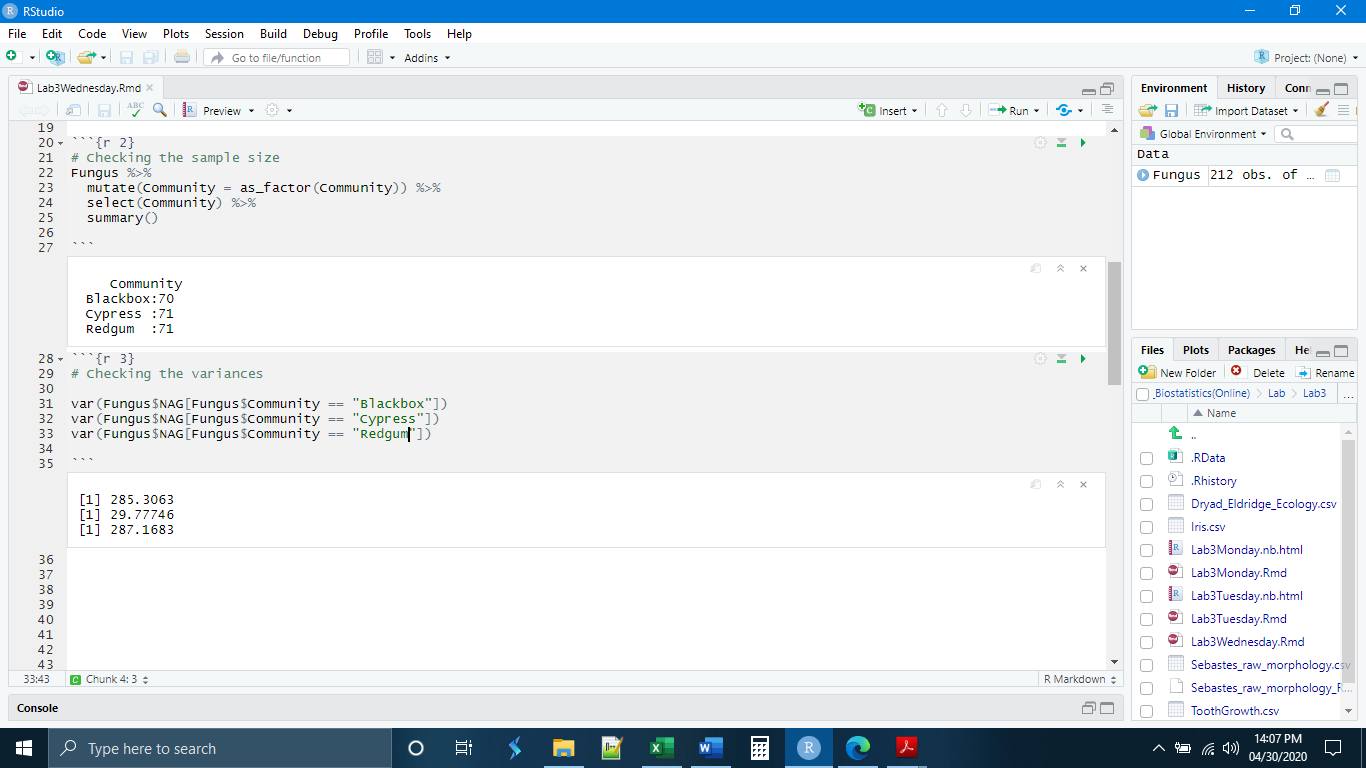


*Assessing Assumptions*

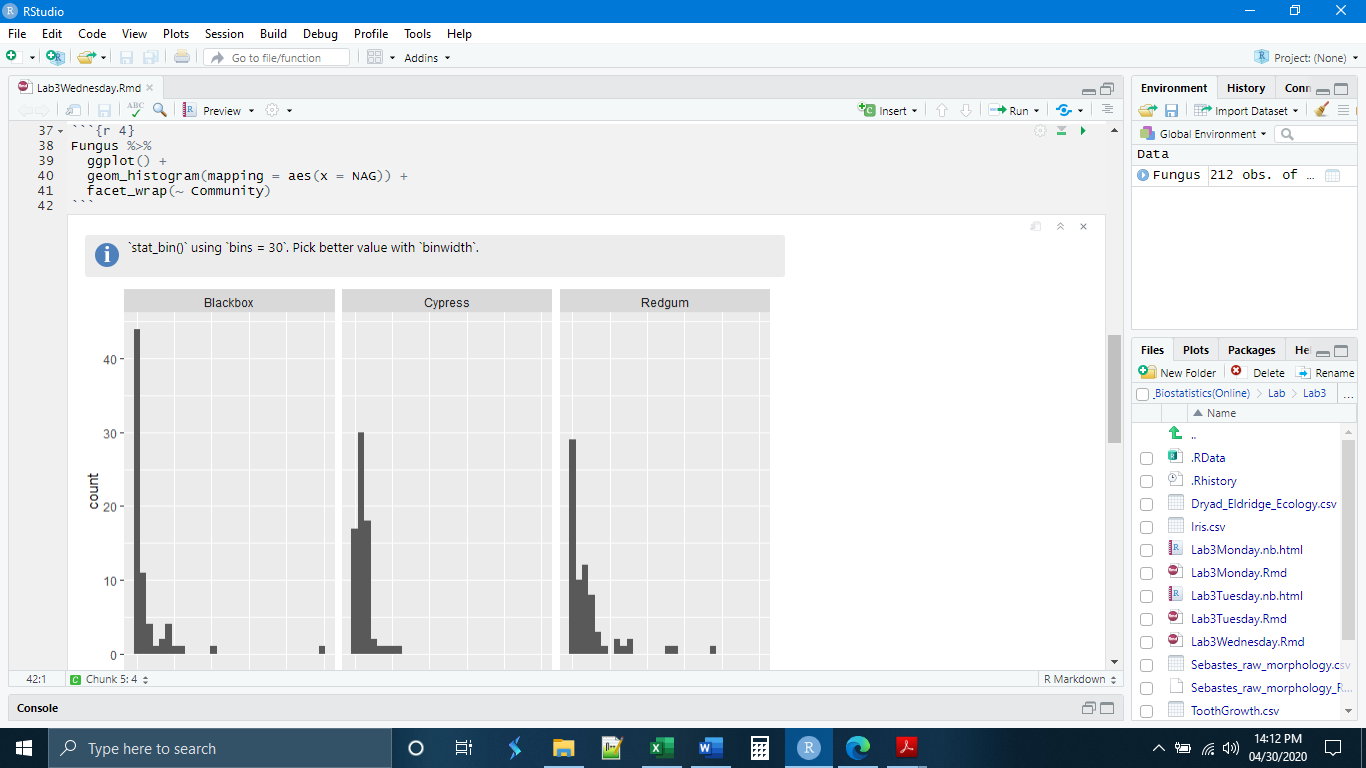
These data are derived from a study that was interested in changes in the soil microbial communities under different animal grazing conditions. Let us assume that we were interested in the NAG (N-acetyl-beta-glucosaminidase) concentration in the soil. This is an enzyme that serves as a proxy (indicator) of nitrogen degradation activity in the soil. We decide that we want to test to see if the NAG concentrations are different in the three different ecological community types that were surveyed in the study.

Before we bother to run our test, we want to make sure that we have met the assumptions required for the test. Let us assume that the conditions associated with data collection, such as sample independence have been met. We want to check for large sample size, even variances, and normality.

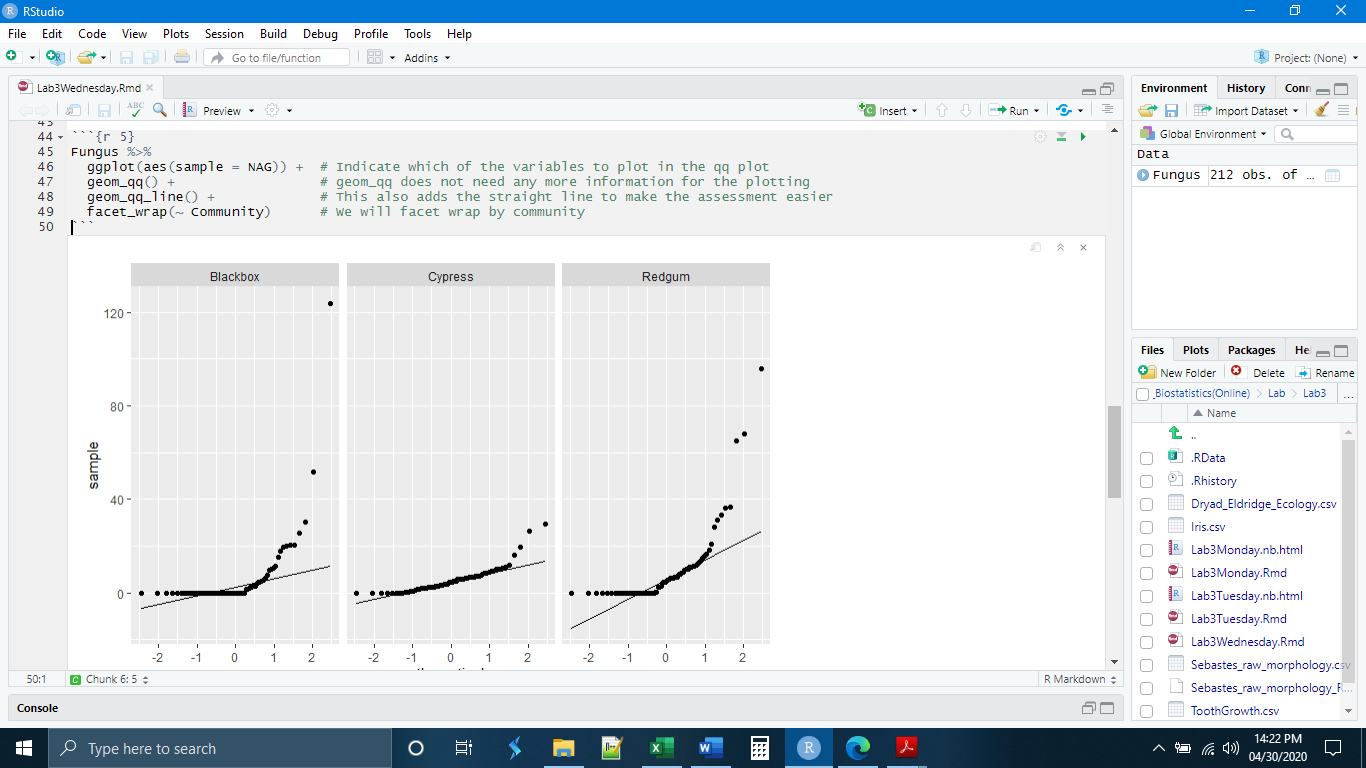
Checking first for sample size, we can see that all three communities have samples larger than 30, which is good. The variances are uneven. It looks like it would be appropriate to compare the blackbox woodland community and the redgum woodland community.



Now let’s check for normality. One way to do this is by making a histogram. Ignoring the cypress community, we see that both blackbox and redgum are heavily skewed right.

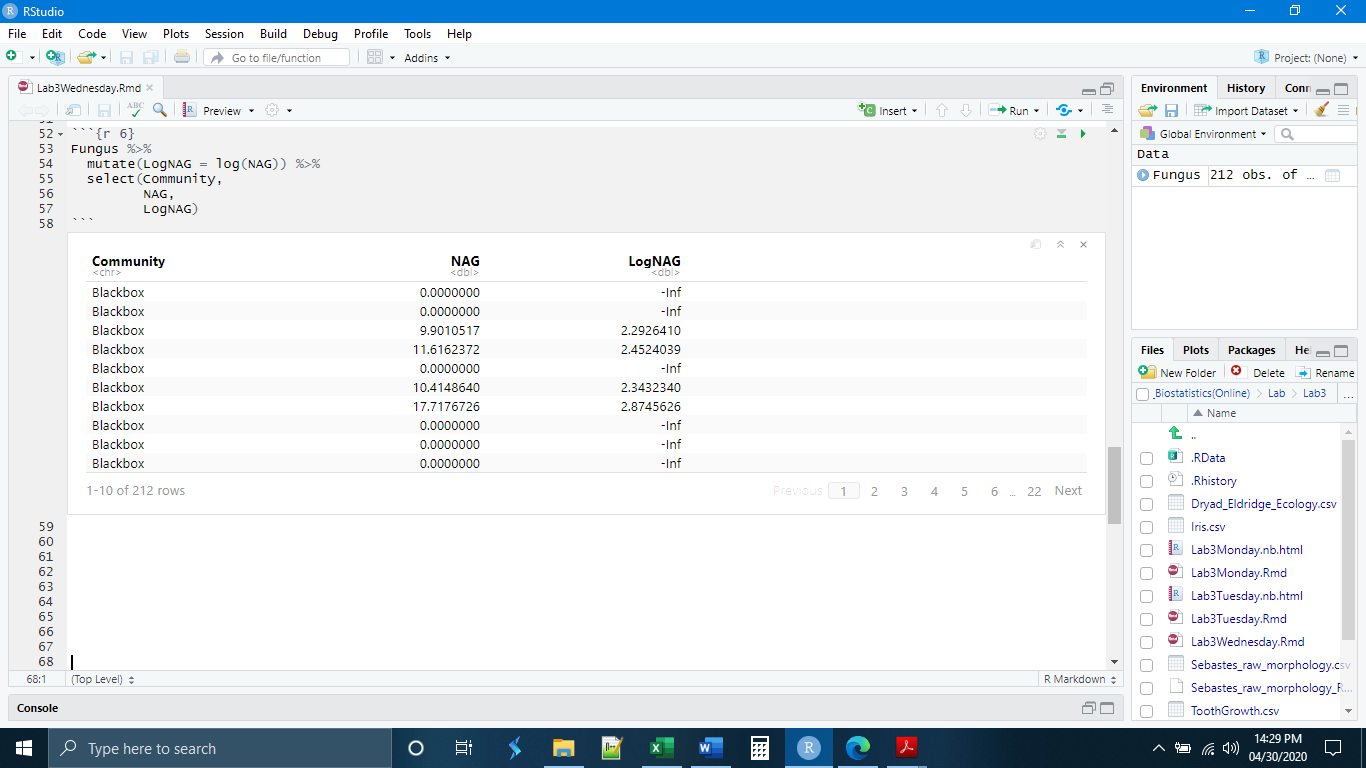


A second way to assess the shape of the distribution of a set of observations is the q-q plot, which plots the expected quantiles of the data points against the observed quantiles of the data points. If the data are normally distributed, the points will fall on a straight line. The syntax in ggplot is a little different from what we are used to, the data being indicated upfront:

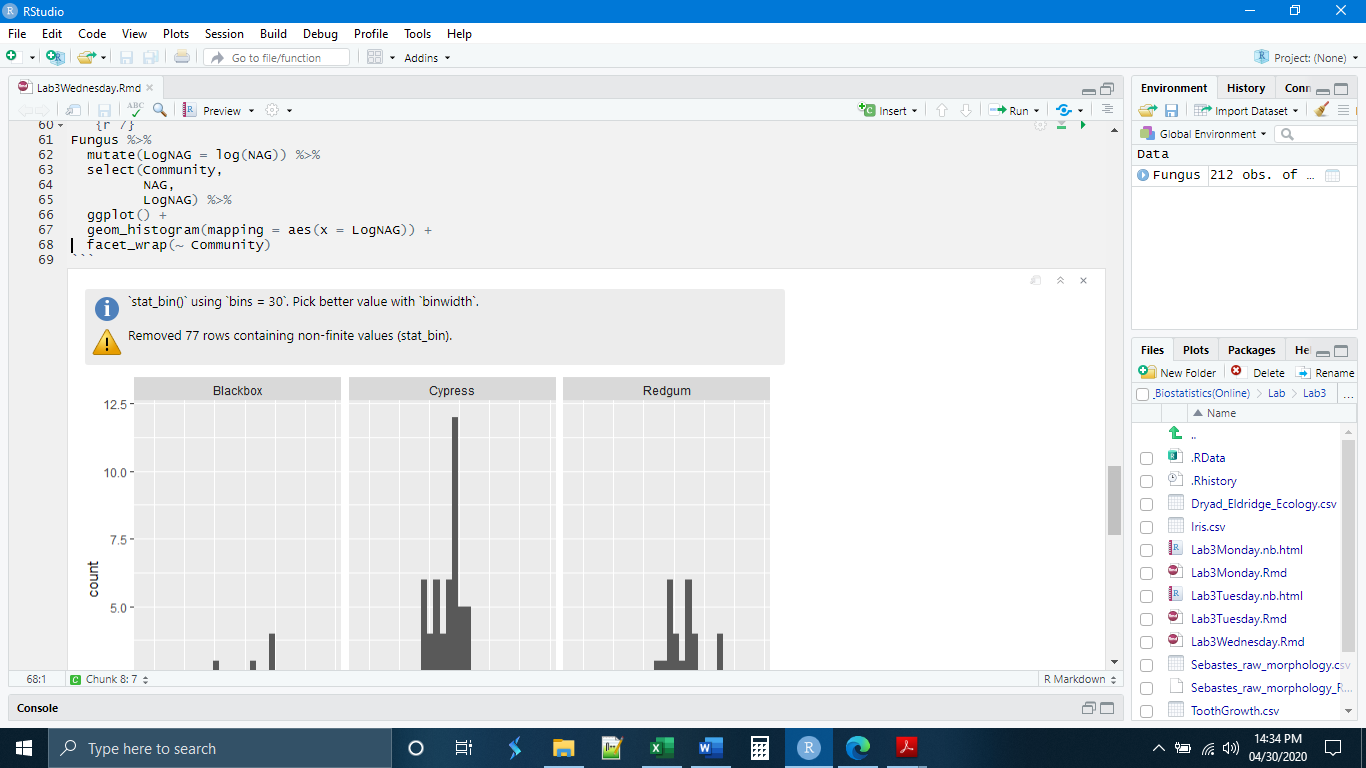


*Using Mutate to Transform (log-normal)*

Both the q-q plot and the histogram indicated that our data were not normally distributed. Moreover, both indicated that our data seem to be distributed lognormally. A lognormal (log-normal) distribution is one that becomes normal if the logarithm is taken of every value. There are some disadvantages to this and it should only be done if necessary. For example, we lose some of our data because log(0) is in undefined value.



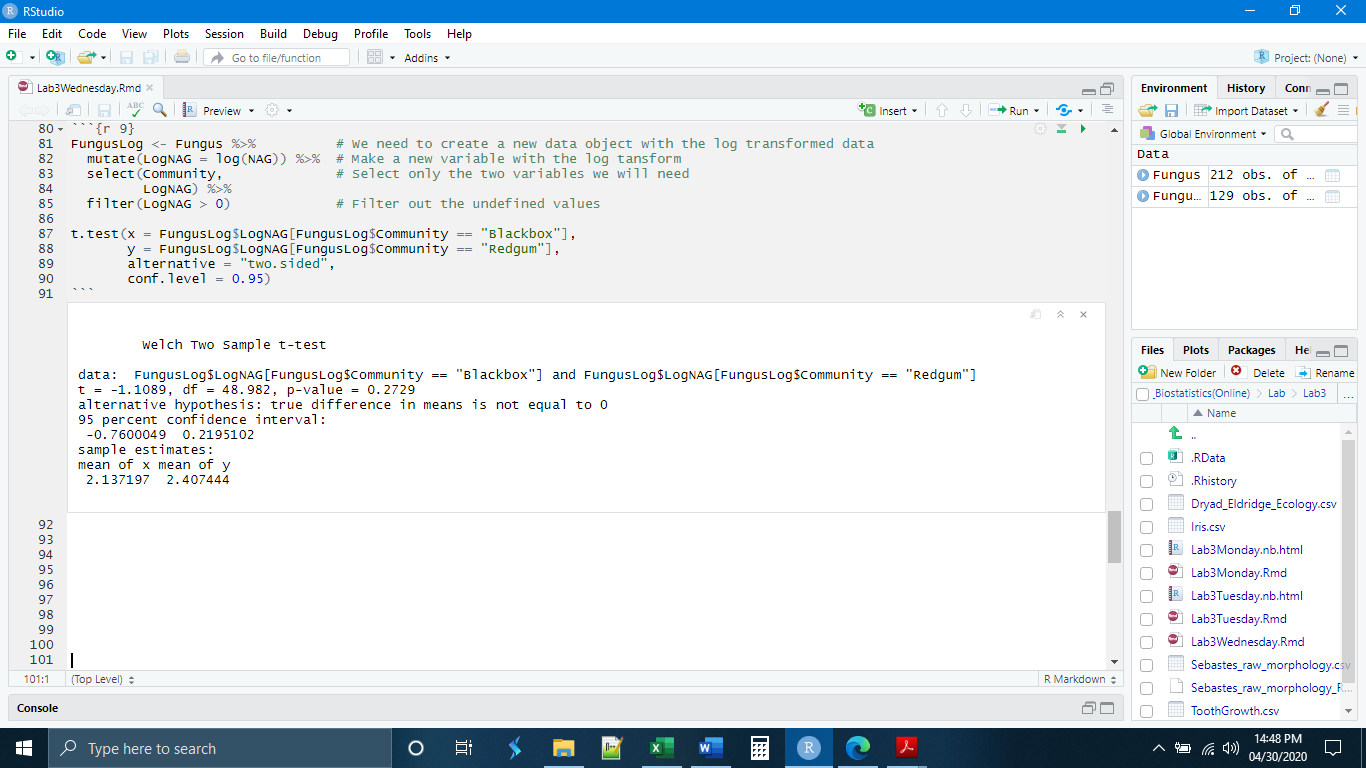
If we check our histograms and q-q plots using the log values, we will see that our distribution has improved.





*The t-Test on Transformed Data*

With these improved data, we can now perform our t-Test on two means. We first have to save the log transformed data as a new object and remove the undefined values.



Here it looks like, at least in the log transformed data, the two communities were not different with respect to NAG concentrations. Any attempt to interpret the results of transformed data reveals a second disadvantage, the interpretation is not straightforward. We never tested the difference in mean NAG concentrations. What we tested was the difference in mean log(concentrations). We would be required to report any findings in those terms. As a result, data transformations can be done, but only if the data are quite badly skewed or if the sample size is small. In general, larger sample sizes can compensate for violations of normality. If you have more than 50 observations per group and the data look generally normal, a transformation is not advised.

*Exercises*

**1.** Make a new chunk called “chunk 5” and write code to

- Read in the Fungus data (from Blackboard) and save it as a new data object

- Generate histograms in ggplot of the BASI2 observations (basidiomycete abundances), facet wrapping by community.

(Hint: The histogram code and the read\_csv() code cannot be a single pipe because you need to save the data as its own object)

**2.** Make a new chunk called “chunk 6” and write code to determine the variances of the BASI2 variable for each community. Which two communities have comparable variances? Record your answer as a comment in the code.

**3.** Make a new chunk called “chunk 7” and write code to

- Perform a log transformation of the BASI2 variable, creating a new data object to save the resulting data.

- Generate histograms in ggplot of the new log transformed data, facet wrapping by community. Did the log transform help to normalize the data? Record your answer as a comment in the code.

- Perform a two-sample t-Test of the log transformed basidiomycete concentrations to test if there are differences in the basidiomycete abundances in the blackbox and cypress communities.