Welcome to DATA 151

I'm so glad you're here!

DATA 151: CLASS 3B
INTRODUCTION TO DATA SCIENCE (WITH R)

WORKING WITH DATA IN R

NOTES PREPARED BY PROF. KITADA SMALLEY (FALL 2022)

ANNOUNCEMENTS

RELEVANT READING

Introduction to Data Science:

- Tuesday:
  - Ch 1: Getting Started with R and R Studio
  - Ch 2: R Basics
- Thursday:
  - Ch 3: Programming basics

HOMEWORK REMINDER

Due next week:

- *HW #3: DC Introduction to Programming in R*
  - ***No submission on WISE necessary, do on DataCamp***
- *Project Milestone #1: Project Proposal*
  - ***Due on WISE 9/22***
  - *One submission per group*

RStudio®

LEARNING ABOUT DATA IN R

LEARNING ABOUT THE DATA

Details

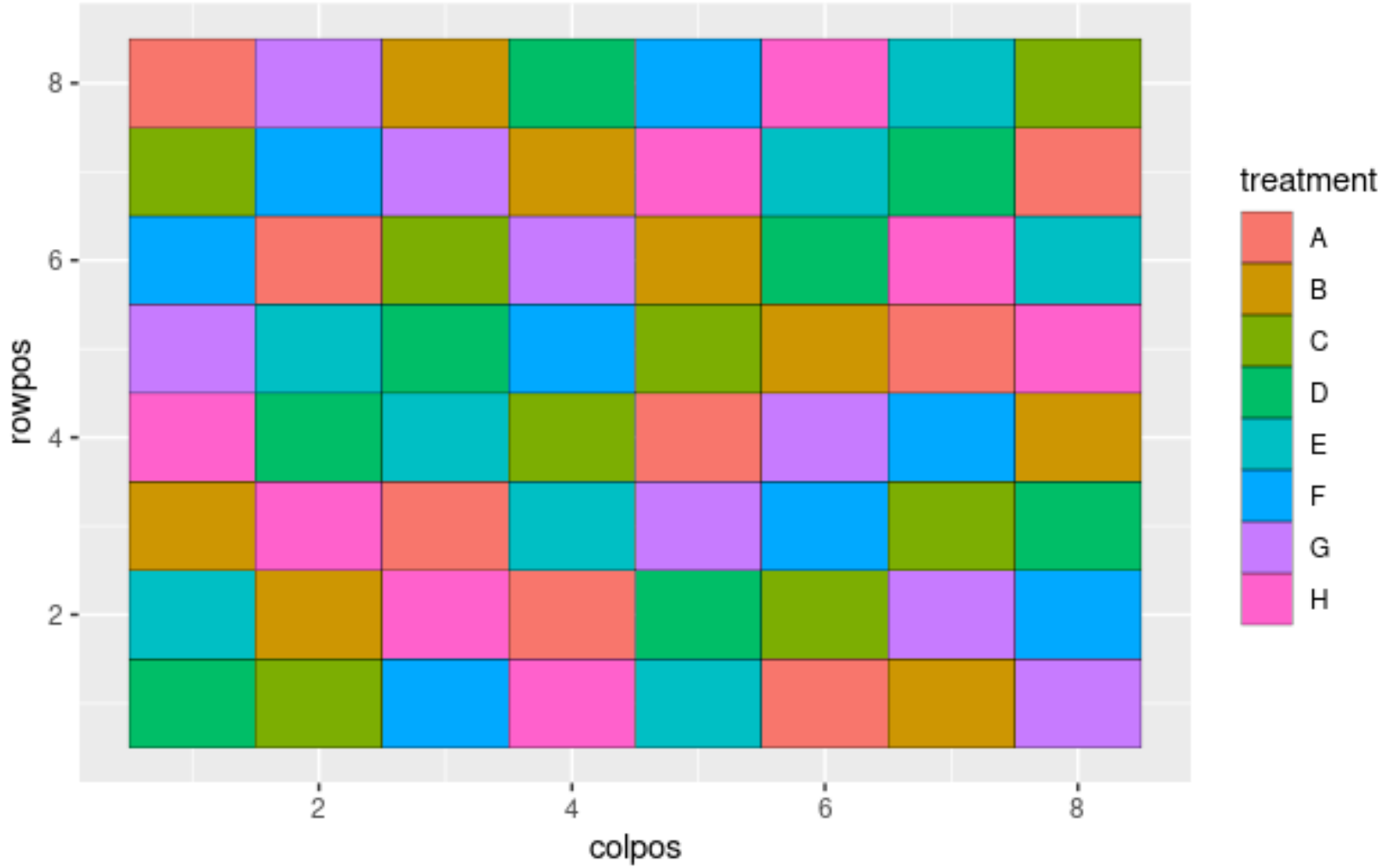
Individual cells of dry comb were filled with measured amounts of lime sulphur emulsion in sucrose solution. Seven different concentrations of lime sulphur ranging from a concentration of 1/100 to 1/1,562,500 in successive factors of 1/5 were used as well as a solution containing no lime sulphur.

The responses for the different solutions were obtained by releasing 100 bees into the chamber for two hours, and then measuring the decrease in volume of the solutions in the various cells.

An 8×8 Latin square design was used and the treatments were coded as follows:

A highest level of lime sulphur
B next highest level of lime sulphur
.
.
.
G lowest level of lime sulphur
H no lime sulphur

WHAT DOES THE EXPERIMENT LOOK LIKE?



WORKSHEET TIME!

Part I: Experimental Design:

1) What are the response and explanatory variables in this study?

2) Are the four principles of a randomized experiment present? Verify each and explain.

1.
2.
3.
4.

If the four principles are met, do you feel comfortable making a cause and effect conclusion?

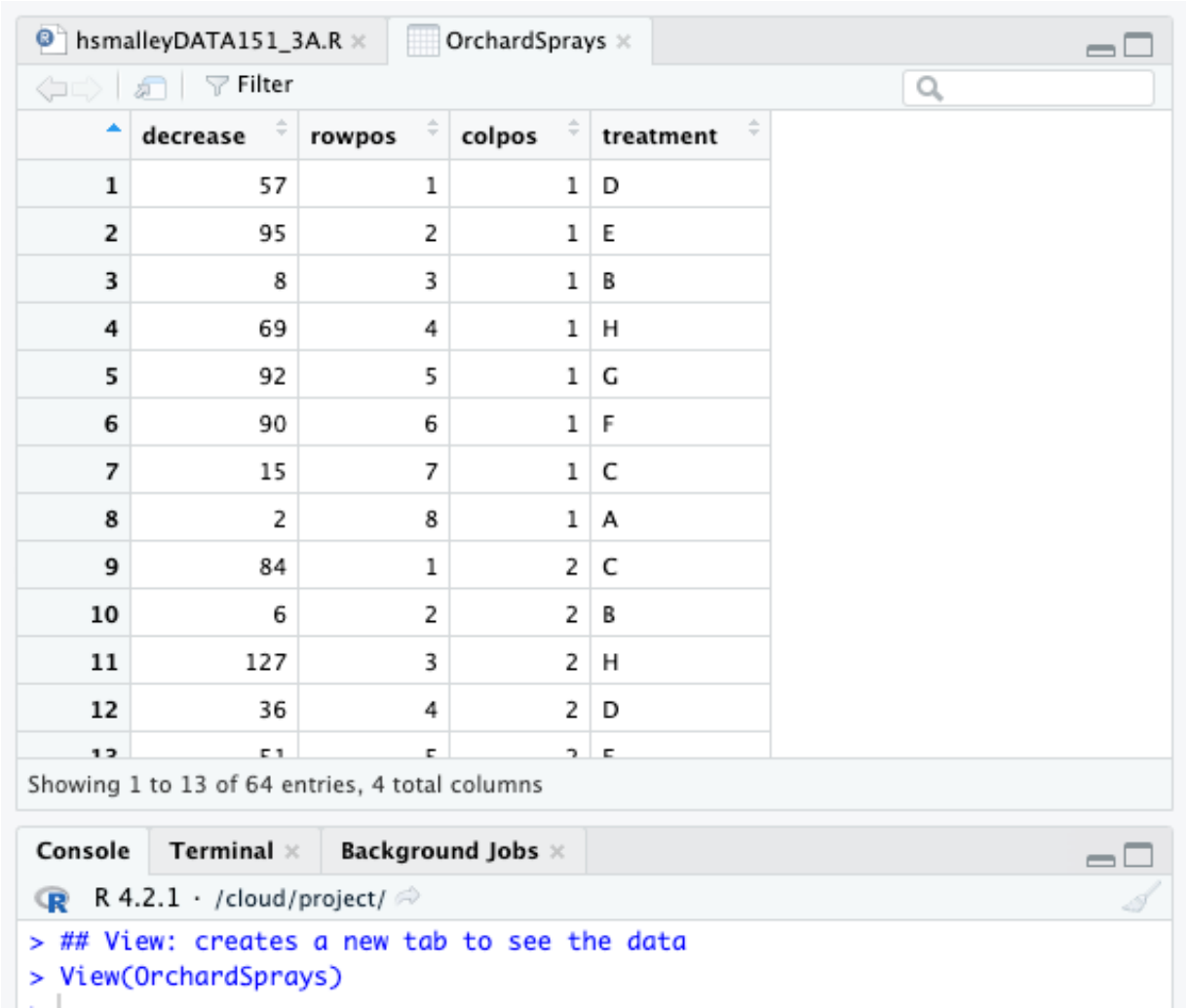
LOOKING AT THE DATA: HEAD

```
> ## STEP 5: Looking at the data
> ## head: first six rows
> head(OrchardSprays)
   decrease rowpos colpos treatment
1        57      1      1         D
2        95      2      1         E
3         8      3      1         B
4        69      4      1         H
5        92      5      1         G
6        90      6      1         F
```

LOOKING AT THE DATA: TAIL

```
> ## tail: last six rows
> tail(OrchardSprays)
   decrease rowpos colpos treatment
59       39      3      8         D
60       14      4      8         B
61       86      5      8         H
62       55      6      8         E
63        3      7      8         A
64       19      8      8         C
```

LOOKING AT THE DATA:VIEW



DATA STRUCTURE

```
> ## STEP 6: Looking at the data structure
> str(OrchardSprays)
'data.frame':   64 obs. of  4 variables:
 $ decrease : num  57 95 8 69 92 90 15 2 84 6 ...
 $ rowpos   : num  1 2 3 4 5 6 7 8 1 2 ...
 $ colpos   : num  1 1 1 1 1 1 1 1 2 2 ...
 $ treatment: Factor w/ 8 levels "A","B","C","D",..: 4 5 2 8 7 6 3 1 3 2 ...
```

This is a data.frame object!

WORKSHEET TIME!

Part II: Data Structure:

1) What do the rows of this dataset represent?

2) What do the columns of this dataset represent? Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

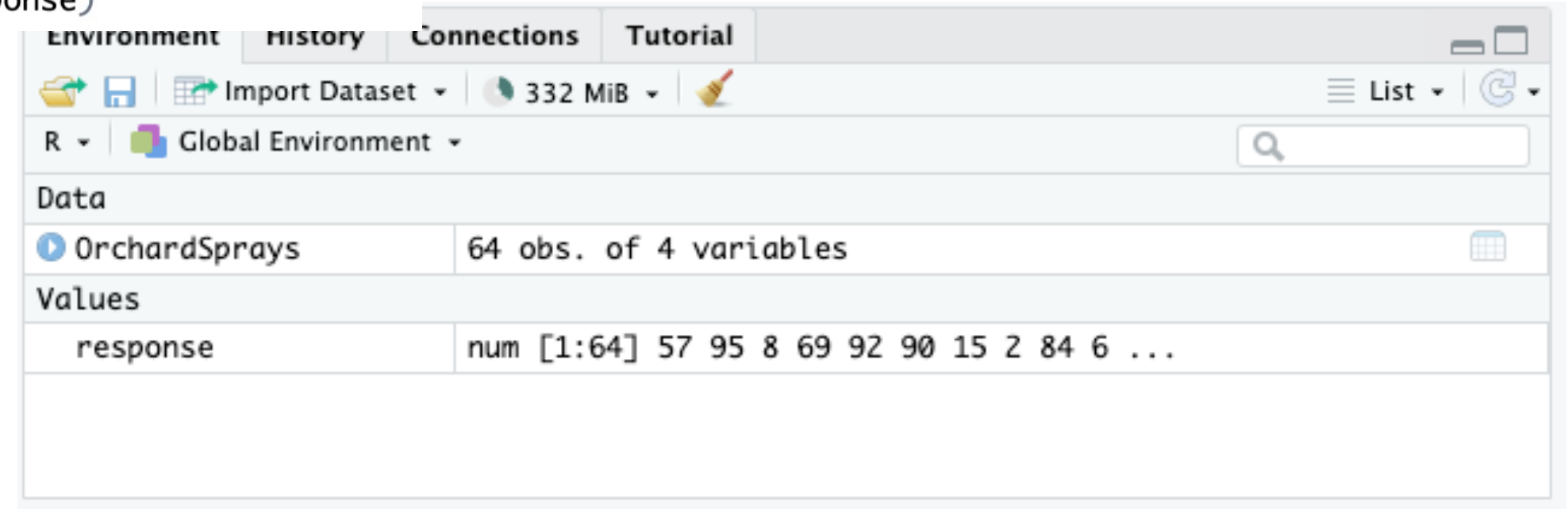
Part III: Hypothesis:

5) Based on the description of the study, what do you think the researcher's hypothesis was? Are higher or lower values of the response preferable? Explain.

VARIABLE ASSIGNMENT AND $ OPERATOR

```
response<-OrchardSprays$decrease

## what kind of class is this?
class(response)
```



Syntax for $ operator: `dataset$column`

TYPES OF VARIABLES

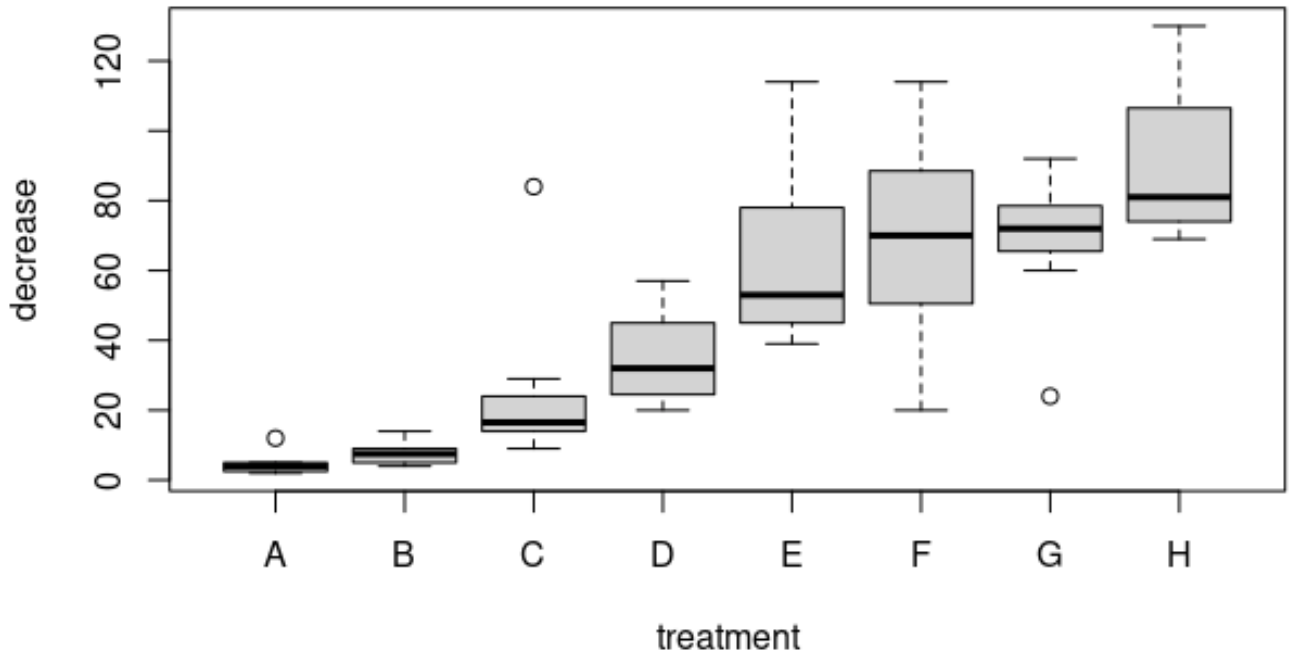
```
## there are 4 types of classes
# 1) factors
explanatory<-OrchardSprays$treatment
class(explanatory)
# 2) character strings
my_name<-"heather"
class(my_name)
# 3)
my_boolean<-TRUE
class(my_boolean)
# 4)
my_pie<-pi
class(my_pie)
# 5)
my_int<-13L
class(my_int)
```

BASIC GRAPHICS IN R

```
boxplot(decrease~treatment, data = OrchardSprays)
```

Numeric Response

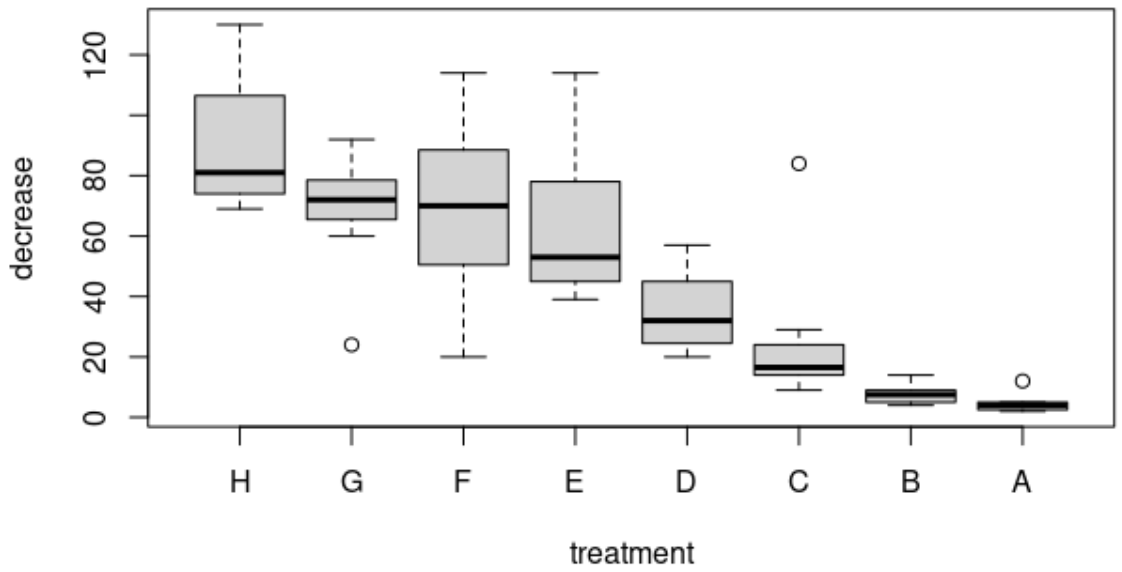
Categorical Variable



BONUS: REORDER FACTOR VARIABLES

```r
## BONUS: Reorder factors
### Is this the order we want?
OrchardSprays$treatment <- factor(OrchardSprays$treatment,
                                  levels=c('H', 'G', 'F', 'E',
                                           'D', 'C', 'B', 'A'))

## Plot again
boxplot(decrease~treatment, data = OrchardSprays)
```



WORKSHEET TIME!

Part IV: Basic Graphics in R:

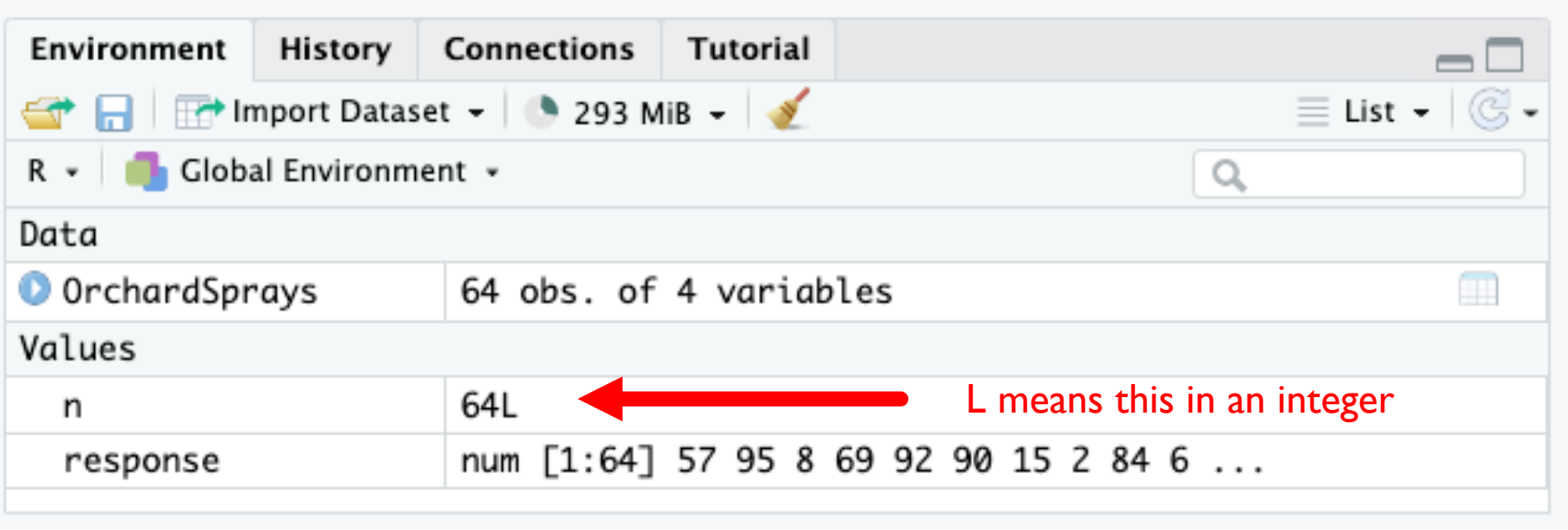
6) Describe any possible trends in these data. Explain in the context of this study.

Part V: Preliminary Conclusions:

7) Based on all the parts above, can you help the researcher find evidence to support or refute their hypothesis? Explain.

VECTORS

```
## STEP 9: Vectors
### vectors are one dimensional arrays
n<-length(response)
n
```



COMMON FUNCTIONS BUILT INTO R

```
## STEP 10: Common functions
## how much solution was consumed in the experiment?
sum(response)

## what is the average amount of solution consumed?
mean(response)
```

COMMON FUNCTIONS BUILT INTO R

```
## STEP 11: Using variables in operations
sum(response)/n
```

```
> ## STEP 11: Using variables in operations
> sum(response)/n
[1] 45.42188
>
> # verify
> mean(response)
[1] 45.42188
```