Week 5 Quiz

The answers to the following questions should be placed in a single document file or text file.

Week 5 quiz is due end of day on Friday September 26th. Solutions to all quiz exercises will be

posted on Saturday September 27th. Your grade will be based on two randomly selected exercises.

Reading Quiz: “Tidy Data”

1. Data \_\_\_\_ is what Hadley Wickham calls structuring datasets to facilitate analysis.

* Data Tidying

2. In tidy data, there are three properties. One of these properties states that “each type of

observational unit forms a table.” What are the other two properties of tidy data.

* Each variable forms a column
* Each observation forms a row

3. According to Hadley Wickham, what are the five most common problems with messy datasets?

* Column headers are values, not variable names
* Multiple variables are stored in one column
* Variables are stored in boh rows and columns
* Multiple types of observational units are stored in the same table
* A single observational unit is stored in multiple tables

4. Provide a one sentence definition for each of the four fundamental verbs of data manipulation:

a. Filter

* Filtering is subsetting or removing observations based on some condition.

b. Transform

* Transforming data is the act of adding or modifying variables, including modifications done on single variables (e.g. log-transformation), or multiple variables (e.g. computing density from weight and volume).

c. Aggregate

* Aggregating involves collapsing multiple values into a single value (by summing, taking means, high, low, etc.)

d. Sort

* Sorting is changing the order of observations, usually sorting by one of the columns to gain more information about that attribute.

5. What is the function in base R used for filtering?

* According to Hadley Wickam, the subset() function is used for filtering. I actually rely a bit more on using logical indexes in dataframes. For example, selecting all rows where a certain criteria is met can be done like this:
  + Df[Df$columnname == “value”,]

##########################################################################

Reading Quiz: “Introduction to dplyr”

1. What are the advantages and disadvantages of using the dplyr package versus the plyr package?

* The advantages of dplyr compared to plyr are the ability to connect to external databases (SQL, etc,) it’s faster, and it allows you to use the “grammer of data manipulation.” The functions are more easily spelled out, and you’re able to string them together in a logical way. The only disadvantage I see is dplyr works exclusively with dataframes. But, for what you’re doing, this is probably the sort of data being used.

2. What are the advantages and disadvantage of using plyr versus the base R apply family of

functions?

* There are similar advantages here: grammer, speed, and the connection to databases. The base R apply family of functions are useful, but don’t even really approach the usefulness of base plyr for split/apply/combine methods (I always think of apply as a replacement for a for loop rather than really useful for split/apply/combine methodology.) dplyr is also more consistent than base R, and built to work with data frames while other base R functions are built to work around vectors.

##########################################################################

Reading Quiz: “tidyr”

1. What are the two verbs (functions) in the tidyr package that correspond to the spreadsheet

functions pivot and unpivot?

* Gather() is the verb for “pivot” and spread() is the verb for “unpivot.”