Week 5 quiz

“Tidy Data”

1. “Tidying” is what Hadley Wickham calls structuring the data for analysis.
2. The other two parts are that each variable forms a column, and each observation forms a row.
3. The five most common problems with messy data are:

Column headers as values and not variable names

Multiple variables stored in a column

Variables are stored in both rows and columns

Multiple types of observations are stored in a singled data table

A single observation unit is stored in multiple tables

1. Once sentence about the four verbs:
   1. A filter creates a subset of the data based on some condition.
   2. A transform modifies an existing variable or adds a new variable to the data set, and can be a function of other variables.
   3. Aggregate data is a summary of values, like a mean or summation of the values by group.
   4. A sort changes the order the observations in the data.
2. The function in base R used for filtering is the subset function.

Intro to dplyr

1. The dplyr package is faster than plyr, and provides a better joins, but it is designed only to work with data frames.
2. The plyr apply functions are designed to work on a dataframe, where the base apply functions are designed for vector application. This is can be an advantage or disadvantage depending on what your data looks like.

Tidyr

1. The corresponding to the spreadsheet functions pivot is ‘gather’, ‘spread’ corresponds to unpivot.