# 17.5: Data Manipulation in R

R is a very flexible tool for manipulating data into various subsets and forms. There are many useful packages and functions for doing this, including the dplyr package, tidyr package, and more. R and its packages will allow users to transform their data from long to wide formats, remove NA values, recode variables, etc. In order to make the downloaded data more manageable for the book, we are going to do two things. First, we want to restrict our data to one wave. The data we downloaded represent many waves of a quarterly survey that is sent to a panel of Oklahoma residents on weather, climate and policy preferences. This book will not venture into panel data analysis or time series analysis, as it is an introductory text, and therefore we simply want one cross section of data for our analysis. This can be done with one line of code:

```
# ds<-subset(ds, ds$wave_id == "Wave 12 (Fall 2016)")
```

What this line of code is doing is creating an object, that we have again named ds in order to overwrite our old object, that has only the 12th wave of data from the survey. In effect, this is removing all rows in which waveid, the variable that indicates the survey wave, does not equal twelve. Across these many waves, many different questions are asked and various variables are collected. We now want to remove all columns or variables that were not collected in wave twelve. This can also be done with one line of code:

```
# ds<-ds[, !apply(is.na(ds), 2, all)]
```

This line of code is a bit more complicated, but what it is essentially doing is first searching all of ds for NA values using the is.na function. It is then returning a logical value of TRUE or FALSE—if a cell does have an NA then the value returned is TRUE and vice versa. It is then searching by column, which is represented by the number 2 (rows are represented by the number 1), to see if all of the values are TRUE or FALSE. This then returns a logical value for the column, either TRUE if all of the rows/cells are NAs or FALSE if at least one row/cell in the column is not an NA. The ! is then reversing the TRUE and FALSE meanings. Now TRUE means a column that is not all NA and therefore one we want to keep. Finally, the brackets are another way to subset our data set. This allows us to keep all columns where the returned value is TRUE, or not all values were NA. Because we are concerned with columns, we write the function after the comma. If we wanted to do a similar thing but with rows we would put the function before the comma. Finally, we want to save this dataset to our working directory which will be explained in the following sectio