Homework

# Statistics and Functions Homework – Cleaning/Munging Dataframes

Often when you get a dataset, it is not in the exact format you want/need. So, you have to refine the dataset into something more useful – this is often called “data munging.”

In this lab, you need to read in a dataset and work on that (in a dataframe). Then, we will explore the distribution within the dataset.

**Step 1: Create a function (named readStates) to read a CSV file into R**

1. You need to read a URL, not a local file to your computer.

2. The file is a dataset on state populations (within the United States).

The URL is:

[www2.census.gov/programs-surveys/popest/tables/2010-2011/state/totals/nst-est2011-01.csv](http://www2.census.gov/programs-surveys/popest/tables/2010-2011/state/totals/nst-est2011-01.csv)

(Note that you might need to use https:// rather than http://)

*Hint: Google “read.csv” and “url” with respect to R commands.*

**Step 2: Clean the dataframe**

3. Note the issues that need to be fixed (removing columns, removing rows, changing column names).

4. Within your function, make sure there are 51 rows (one per state + the district of Columbia). Make sure there are only 5 columns with the columns having the following names (stateName, Census, Estimates, Pop2010, Pop2011).

5. Make sure the last four columns are numbers (i.e. not strings).

**Step 3: Store and explore the dataset**

6. Store the dataset into a dataframe, called dfStates.

7. Test your dataframe by calculating the mean for the 2011 data, by doing:

mean(dfStates$Pop2011)

🡪 you should get an answer of 6,109,645

**Step 4: Find the state with the highest population**

8. Based on the 2011 data, what is the population of the state with the highest population? What is the name of that state?

9. Sort the data, in increasing order, based on the 2011 data.

**Step 5: Explore the distribution of the states**

10. Write a function that takes two parameters. The first is a vector and the second is a number.

11. The function will return the percentage of elements within the vector that is less than the same (i.e. cumulative distribution below the value provided).

12. For example, if the vector had 5 elements (1,2,3,4,5), with 2 being the number passed into the function, the function would return 0.2 (since 20% of the numbers were below 2).

13. Test the function with the vector “dfStates$Pop2011”, and the mean of “dfStates$Pop2011.”

There are many ways to write this function (described in point 10) – so please try to write multiple versions of this function – which do you think is best?