Week 4 Overview

Your book *R for Everyone* throws out a statistic in Chapter 11 that you have already heard in this program, multiple times – 80% of a data scientists effort is spent cleaning data. And the reason we keep repeating it, is because it really might be the most important step in the data science process. If overlooked, your analysis is likely full of errors! It is so important; we have an entire course dedicated to learning how to transform data. While we will only be spending 2 weeks on the topic in this course, it is given two weeks because there are quite a few functions in R that you should know to clean, wrangle, munge, transform, etc. your data (there are so many terms for this step). This week and next, we will spend some time cleaning data in R, specifically a dataset that will be used for later analysis in future weeks.

Contents of the Week

Overview

Readings, Assignments and Tasks

4.1 Discussion/Participation

4.2 Exercise

Objectives

After completing this week, you should be able to:

Identify when data transformations are needed and be cognizant of the effects.

Packages available in R for data transformations

Implications and potential ethical risks when performing data transformations

Know when and how to clean data

Weekly Resources

Comprehensive R Archive Network

R Studio

Sage Publications. (2021). Discovering Statistics Using R

RStudio, PBC. (2021). RStudio Documentation

RStudio, PBC. (2021). R Studio Cheatsheets

Week 4 Readings, Assignments, and Tasks

Here are your tasks for this week:

Read the following:

- *R for Everyone:* Chapters 8 -11
- *Discovering Statistics Using R*: Chapter 5
- data-transformation.pdf

Complete the following:

- 4.1 Discussion/Participation
- 4.2 Exercise

4.1 Discussion/Participation

Here are optional topics for discussion via Teams this week.

Remember, these topics aren't required, but if you are struggling to know what to post about, these can be used to initiate discussion!

1. How do you pass arguments to a function and why would you want to do this?
2. What is the … argument in R?
3. What is the return command used for?
4. What are control statements? Provide some examples
5. What are compound tests?
6. What are loops used for?
7. What is the apply function used for? What are some variations to apply?
8. What is the aggregate function used for?
9. What is the plyr function?
10. What is meant by statistical assumptions?
11. What are parametric tests? How do you know if your data are parametric?
12. What is normally distributed data?
13. How do we deal with assumptions about normality?
14. How do you test if data is distributed normally? What is the Shapiro-Wilk test?
15. What is homogeneity of variance? What is Levene's test?
16. How do we deal with outliers?
17. What are some different types of data transformations outlined in your text?
18. What can go wrong with transforming data?

4.2 Exercise

Complete the following exercises by creating an R Script for each.

1. Test Scores
   a. A professor has recently taught two sections of the same course with only one difference between the sections. In one section, he used only examples taken from sports applications, and in the other section, he used examples taken from a variety of application areas. The sports themed section was advertised as such; so students knew which type of section they were enrolling in. The professor has asked you to compare student performance in the two sections using course grades and total points earned in the course. You will need to import the Scores.csv dataset that has been provided for you.
      i. Use the appropriate R functions to answer the following questions:
         1. What are the observational units in this study?
         2. Identify the variables mentioned in the narrative paragraph and determine which are categorical and quantitative?
         3. Create one variable to hold a subset of your data set that contains only the Regular Section and one variable for the Sports Section.
         4. Use the Plot function to plot each Sections scores and the number of students achieving that score. Use additional Plot Arguments to label the graph and give each axis an appropriate label. Once you have produced your Plots answer the following questions:
            a. Comparing and contrasting the point distributions between the two section, looking at both tendency and consistency: Can you say that one section tended to score more points than the other? Justify and explain your answer.
            b. Did every student in one section score more points than every student in the other section? If not, explain what a statistical tendency means in this context.
            c. What could be one additional variable that was not mentioned in the narrative that could be influencing the point distributions between the two sections?

2. We interact with a few datasets in this course, one you are already familiar with, the 2014 American Community Survey and the second is a Housing dataset, that provides real estate transactions recorded from 1964 to 2016. For this exercise, you need to start practicing some data transformation steps – which will carry into next week, as you learn some additional methods. For this week, using either dataset (or one of your own – although I will let you know ahead of time that the Housing dataset is used for a later assignment, so not a bad idea for you to get more comfortable with now!), perform the following data transformations:
    a. Use the apply function on a variable in your dataset
    b. Use the aggregate function on a variable in your dataset
    c. Use the plyr function on a variable in your dataset – more specifically, I want to see you split some data, perform a modification to the data, and then bring it back together
    d. Check distributions of the data
    e. Identify if there are any outliers
    f. Create at least 2 new variables