Assignment 3

Haviland Wright

September 26, 2018

Reading:
Tidy Data
An introduction to data cleaning with R
R for Data Science Chapters 7,8

The purpose of this assignment is to provide a counterpoint of practical application to the lessons in your R for Data Science textbook. In the textbook, the exercises explain and illustrate the use of tidyverse tools. Exercises deepen your understanding by posing questions with verifyable answers. The class assignments attempt to recreate the more ambiguous situations that tend to occurr in practice.

Assignment 3 is the first in a series of assignments in which you will use tidyverse tools for data explorations and for reporting what you have found in a way that allows others to follow your steps.

In class, I demonstrated two databases that provide query tools through which you can set parameters for data selection. In the Financial Crimes Enforcement Network, I showed the Suspicious Activities Report statistics tool. I also demonstrated the Quick Stats interface that provides access to a database from U.S. Department of Agriculture's National Agricultural Statistics Service.

Your assignment is to use tidy verse tools to explore these databases and report your findings. You don't need to be comprehensive. Define a topic to focus your exploration and report on the topic.

Working with the Suspicious Activity Report database

In the Suspicious Activities Report database, for example, getting an idea of the number of reports by year and state might be one way to start.

Table 1: Securities/Futures Suspicious Activity Reports by Year

Year	Count
2012	1111
2013	17091
2014	28295
2015	27631
2016	27387
2017	33657
2018	25060

Table 2: Securities/Futures Suspicious Activity Reports – Top 10 States

State	Count
California	24835
Massachusetts	18747
New York	16786
Rhode Island	15997
Utah	9702
Florida	7664
New Jersey	7434
Missouri	7214
Nebraska	6738
Virginia	6168

A next step might be to see the kinds us suspicious activities get reported.

Table 3: Top 20 Reported Susicious Activities

Suspicious.Activity	Count
Suspicious EFT/Wire Transfers	18327
Embezzlement/Theft/Disappearance of Funds	17577
Suspicion Concerning the Source of Funds	13866
Other Fraud (Type)	12906
Two or More Individuals Working Together	12039
Unauthorized Electronic Intrusion	10508
Elder Financial Exploitation	9938
Market Manipulation/Wash Trading	9117
Transaction(s) Below CTR Threshold	8886
Insider Trading	8752
Forgeries	8566
Suspicious Use of Multiple Accounts	7641
Credit/Debit Card	7549
Mail	6650
Provided Questionable or False Documentation	6016
Transaction Out of Pattern for Customer(s)	5283
Counterfeit Instrument	4537
Other Money Laundering	3779
Misuse of Position or Self-Dealing	3590
Transaction(s) Below BSA Recordkeeping Threshold	3122

Ag Data: how about some milk?

The First think you notice about USDA Quick Stats is the large amount of gratuitous data it delivers. In this case, the 21 variables delivered by the database are quickly reduced to 7 variables.

Once you have done a little initial cleaning, you can explore the data for patterns, problems, and releatioships.

Keep in mind that you cannot "discover" hypotheses and then test them with your data.

At this point, however, you're probably just trying to sort out basic information you need to access the data at all.

If you work with the code in this document, you should have enough information to make these simple plots for the year 2012:

Plots of Sales by State, Sales by Size of Inventory, Sales by Area of the operation.

Now ask youself: Are the sales relationships uniform across states? regions? NAICS classification?