STAT 29000 Project 4

Topics: Utilizing information across data frames in R

Motivation: It is worthwhile to get familiar with methods for analyzing data, according to information available in a different data frame. This is a very practical skill. We often find the our data is spread across several tables. (Data comes like this because it is worthwhile to save data in an efficient way.)

Context: The concept in this project is something that you have already touched on, when working with databases last year. We can also pull data from more than one data frame in R. We can use data from one data frame to help us with the analysis of another data frame.

Scope: This basically boils down to becoming more and more savvy with the ways that data is labelled in R. As in many of our projects, we can do powerful things in R, with only a little bit of coding, if we make a simple (but careful) plan.

Question 1: Analyze flights according to their State of Origin

Before you start question 1, please consider the examples given here:

`/class/datamine/data/examples/stat29000project4examples.R`

These will help you to better understand the techniques that we will use in this project.

Read the data about the 2019 flights into a data frame called myDF.

Read the data about the airports from

http://stat-computing.org/dataexpo/2009/airports.csv

into a data frame called myairports.

1a. Using the information from the airports.csv file (which you stored in the data frame myairports), add a new column (to be called "myoriginstateabbr") to the data frame myDF, which lists the two-letter abbrevation of the State of the origin airport.

1b. Compare your new column (myoriginstateabbr, which you added to myDF), to the given data about state abbreviations in myDF$ORIGIN_STATE_ABR. These columns will mostly be the same. You will find, however, that two airports in this data set have a different two-letter abbreviation. What are the names of the two airports, for which the two-letter abbreviation is different?

Question 2: Analyze yellow taxi cab rides according to the Borough of their pickup location

Read the data about the June 2019 taxi cab trips into a data frame called myDF.

Read the data about the taxi zones from

https://s3.amazonaws.com/nyc-tlc/misc/taxi+_zone_lookup.csv

into a data frame called myzones.

2a. Add a new column (to be called "PUBorough") to the data frame myDF, which lists the Borough of the pickup location.

2b. Use the PUBorough column of myDF to discover how many yellow taxi cab rides originated from each Borough, in June 2019.

Question 3: Analyze the number of donations to each election campaign

Read the data about the 2020 election campaign contributions into a data frame called myDF.

Read the data about the election committees from

https://www.fec.gov/files/bulk-downloads/2020/committee_summary_2020.csv

into a data frame called mycommittees.

3a. Add a new column (to be called "CMTE_NM") to the data frame myDF, which lists the committee name to which each donation was made.

3b. Use the CMTE_NM column of myDF to discover the name of the committee that received the greatest number of contributions during the 2020 election (so far).

3c. Use the CMTE_NM column of myDF to discover the name of the committee that received the greatest total amount altogether (in dollars) of contributions during the 2020 election (so far).

Project Submission:

Submit your solutions for the project at this URL: https://classroom.github.com/a/DN4cf0DI using the instructions found in the GitHub Classroom instructions folder on Blackboard.