STAT 29000 Project 6

Topics: Analysis across 100 AirBnB cities

Motivation: We spent considerable time, in the previous project, building a large data frame, based on the AirBnB data. In this project, we begin to explore this data frame. We basically do some exploratory data analysis, to begin to better understand this data.

Context: Data wrangling is the first step in analysis. A colleague in industry told me this week that she estimates that 80 percent of the work in data analysis comes down to simple data wrangling!

Scope: We develop some experience in this project with parsing and manipulating strings. We begin to develop an appreciation for the richness of this data set. The project ends with an open-ended question about the nature of pricing the properties in this data set.

Throughout this project, we use `bigDF` to refer to the big data frame built in STAT 29000 Project 5. The `bigDF` data frame should have 1340094 rows and 16 columns.

Question 1: Parsing strings to extract information

The `rownames` of `bigDF` are so long that it makes it hard to see the data when we type, for instance, `head(bigDF)`. The `head` is not very readable, with these long `rownames`. Nonetheless, the `rownames` contain crucial information. We can think about the `rownames` as character strings, with 10 occurrences of "/" and (therefore) 11 substrings. (Hint: We might do a `strsplit` with "/" as the splitting character.)

The first 5 such substrings of the `rownames` come from `/class/datamine/data/airbnb` so we do not need those. The 6th through 9th substrings of the `rownames` contain the information about (respectively) the country, the state/region, the city, and the date that the information was downloaded from the internet.

1a. Make a new columns of `bigDF` called (respectively) `country`, `region`, `city`, and `downloadeddate`, which give this information.

1b. Now we do not need the `rownames` anymore. Change the `rownames` so that the first of the `rownames` is 1, the second of the `rownames` is 2, the third of the `rownames` is 3, etc. There are 1340094 `rownames` altogether.

Question 2: Introduction to string manipulation

2a. Some elements of the `name` column of `bigDF` are written using capital letters (along with, perhaps, some numbers, special characters, etc.). How many entries of the `name` column are written in this dramatic way, i.e., every letter that appears is a capital letter?

2b. Import the 2-column file called `review.csv` for Los Angeles. This file has 1427153 review dates. Extract the years from these dates (there are multiple ways to do this; you might choose to use `strsplit` again). For each year from 2009 to 2019, find the number of reviews for properties in Los Angeles.

Question 3: A Pricing Mystery

3a. Which city in the United States has the most expensive average AirBnB listing prices?

3b. Why do you think that is the case? Justify your answer using the data.

Project Submission:

Submit your solutions for the project at this URL: https://classroom.github.com/a/q2XKoW2A using the instructions found in the GitHub Classroom instructions folder on Blackboard.