UCLA Extension Data Science Intensive

Instructor: William Yu

Project 3

10/8/2018

- Submit your results (including R script and any output files you got) through Canvas.

A. Visualization of Data

- In the class, we plot the U.S. county map with color for CHCI in 2016 (period: '12-'16; midyear: 2014). Now apply it to plot the similar chart for CHCI in 2009 (period: '05-'09; midyear: 2007). Export the chart as a pdf file. (Above the chart, click Export → Save as image, choose the directory).
  Note: the data is from chci.csv in Project 2 folder. The R script is D03b_map.
- In the class, we plot the correlation between life expectancies and fertility rates among countries in 2016 with bubble size for population. Now apply it to plot the similar chart for the year in 1960.
  Note: The R script is D03a_ggplot.

B. Explore Zillow Prize Project

- Go to https://www.kaggle.com/c/zillow-prize-1 to read the details of Zillow Prize competition.
- Although it is too late to enter the competition, we still can use the Zillow data to analyze. Basically, Zillow wants to use this competition to improve its Zestimate prediction of real-world home sales price.
- In project 3 folder, I have downloaded all the Zillow prize data for you to practice. Put the data into your computer (It is big so it could take a while to load into R). I have prepared a P03_zillow R script for you. Go run though the script one by one. It provides a lot of interesting and useful tools for data exploration and visualization.
- In the data, the main interest is its forecast error:

  logerror = log(Zestimate)-log(SalePrice)

- The goal is to figure out what variables among 58 variables could predict (or correlate) the logerror. If we find those, Zillow could simply add those variables into their model and consequentially reduce the logerror.
- After went through all those interesting codes in P03_zillow.R, run two simple linear regressions (y1 is abs_logerror and y2 is logerror) and explain the results.

```
Fit1 = lm (abs_logerror ~ tax_property + area_live_finished + num_garage  + num_bedroom + num_bathroom, data=cor_tmp)
```

  Note that to run this regression, the data to use is cor_tmp.

- We will come back to analyze the dataset later on after we learn more machine learning tools.

C. Bonus project (No due date)

Do you recall in the first week I showed you this cool dynamic correlation between life expectancy and fertility rate among countries from 1960 to 2016? In fact, in the class (2016) and this project (1960) we have done this similar task. If you can figure out a way in R to present this dynamics interaction, you will get a bonus.

Link:
https://www.google.com/publicdata/explore?ds=d5bncppjof8f9_&ctype=b&strail=false&nselm=s&met_x=sp_dyn_le00_in&scale_x=lin&ind_x=false&met_y=sp_dyn_tfrt_in&scale_y=lin&ind_y=false&met_s=sp_pop_totl&scale_s=lin&ind_s=false&dimp_c=country:region&ifdim=country&iconSize=0.5&uniSize=0.035