Pandas Assignment

Urban Institute

Due: Before April 13th

Module & Data Import

Start by opening a new Jupyter notebook and importing Python's `numpy`, `pandas`, and `matplotlib` modules. Then use the pandas `read_csv` method to read in the GSS2016.csv data set. This is the newly released General Social Survey for 2016. The data set is produced by NROC at the University of Chicago, and you can see the documentation here (LINK).

Basic Data Exploration

Answer the following questions using Pandas:

- How many columns are there in the dataframe?
- Use the `value_counts()` (LINK) method to create a frequency table of the `kidsinhh` column (a dummy in which 2 indicates yes and 1 indicates no). Make sure to set the `dropna` argument is to `False` so NANs are included.
- How many rows of data are there in which the `genegen` variable is less than or equal to three (indicating respondents think genetically modified crops are harmful to the environment)? What percentage of total respondents (those who are not NAN) does this group represent?
- The `cohort` variable is the respondent's year of birth. Calculate the average, median, and standard deviation for this column. Use the `quantile()` method to calculate the 10th and 90th percentile - the documentation is available here (LINK).
- Subset the original dataframe to create two new, smaller dataframes of just respondents who:
  - are White (`race` variable is equal to 1) **AND** lived in a rural area when young (`res16` is less than or equal to 3);
  - are Black (`race` variable is equal to 2) **OR** they lived in either New England or the Pacific region when young (`reg16` variable is either 1 or 9).

Using Excel

Reading from xls or xlsx files works very similarly to csv files, provided they aren't heavily formatted in a way that makes it hard to read the data in the form of an array. Use the attached file, `some_cities.xlsx`, and do the following:

- Use the Pandas read_excel method (LINK) to load the two sheets (month1 and month2) in as separate dataframes.

Merging & Concatenating

Continuing with the Excel data, read the documentation to accomplish two basic joining tasks:

- Use the concat function (LINK to "stack" the months together in a "long" format. You'll need to create a column in each dataframe that tells the month first, since it was denoted in the source data by which Excel sheet it was on, and not by the data itself.

- Use the merge method (LINK) to join the data together in a "wide" format, using the city and state columns as the merge keys. This will require you to spend a bit more time understanding the arguments you can use with Pandas "merge". In particular, make use of the "suffixes" argument to address the overlapping names, value1 and value2. Also try it with the different join types, "inner", "outer", "left" and "right". What does that change?

Assignment Submission

Send your completed upyter notebook to jlevy@urban.org and aharris@urban.org.

More Resources:

- Pandas - Essential Basic Functionality (LINK)
- 10 Minutes to Pandas (LINK)
- Summarising and Aggregation of Grouped Data in Pandas (LINK)