Pandas Assignment

Urban Institute

Due: Wednesday, April 26

Module & Data Import

To start, we'll be using the same dataset as in the last homework, the General Social Survey for 2016. Start by opening a new Jupyter notebook and importing Python's `numpy` and `pandas` modules. Then use the pandas `read_csv` method to read in the GSS2016.csv data set; see the data documentation here (LINK).

Grouping & Aggregations

- Create a new variable `age_rounded` which rounds `age` to the next multiple of 10. For example, the `age_rounded` would be 30 for ages 25, 27, or 30; it would be 40 for ages 31, 39, or 40.
- Group the dataset by the `age_rounded` variable. For each `age_rounded` group, find the total number of facebook users (with variable `facebook`).
- Using the original data set, find the average number of children (note that the `childs` variable gives the number of children in each family) for users of Facebook.
- For each `age_rounded` group, again sum up the total number of facebook users, but this time weight each facebook user by the number of children that user has. For example, a facebook user with 4 children would add 4 to the total, a facebook user with 2 children would add 2 children to the total, and a non-facebook user with any number of children adds 0 to the total. *Hint*: This can be done in one line of code (though that is not required).
- Save the result from the last exercise to a new variable. What is the type of this new variable? *Hint*: Use the python built-in function `type()`.
- This new variable has a Pandas type that we briefly mentioned in class, and which you'll see referenced in documentation and online as you get more advanced. Read the relevant Pandas documentation here (LINK).
- Return to the original dataset. How many records are in the dataset?
- Randomly generate a list of dates where each date references a different day (see the example from class) and add that as a column in the data set named `date_of_survey`.
- Using `date_of_survey`, what is the total number of facebook users (unweighted) who were surveyed in each month? In each year?

Datatypes and `read_csv`

Now load the file co_validate.csv using read_csv. Notice the warning it gives; if you Google it, one of the first links is to this link (LINK).

Read over the accepted answer. Now explore the dataframe you created, and see what dtypes it assigned. When your data is small relative to your computer's memory, letting it automatically determine dtypes is just fine, but it quickly becomes a problem with larger data. Try the following:

- Look at the data in Excel; notice the top two lines aren't data. Tell read_csv to skip those.
- Specify correct dtypes so that the warning goes away when you load the dataframe.

- EINs have meaningful leading zeros that get chopped off when converted to integers; fix it so they remain after loading.
- Try loading it while setting a column as the index from the start.
- Load only a subset of the columns.

Assignment Submission

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

More Resources:

- Pandas - Essential Basic Functionality (LINK)
- 10 Minutes to Pandas (LINK)
- Summarizing and Aggregation of Grouped Data in Pandas (LINK)
- `read_csv` Pandas Documentation (LINK)
- Data Science in Pandas Cheat Sheet (LINK)