Data Bootcamp: Code Practice #3

Revised: October 10, 2018

Answer each of the questions below in a Jupyter Notebook.

1. Enter and run this code in a Jupyter cell to produce the dataframe `weo`:

```
import pandas as pd
data = {'BRA': [13.37, 13.30, 14.34, 15.07, 15.46, 15.98, 16.10],
        'JPN': [33.43, 31.83, 33.71, 34.29, 35.60, 36.79, 37.39],
        'USA': [48.30, 46.91, 48.31, 49.72, 51.41, 52.94, 54.60],
        'Year': [2008, 2009, 2010, 2011, 2012, 2013, 2014]}
weo  = pd.DataFrame(data)
```

   The numbers are GDP per person in thousands of US dollars, 2008 to 2014, variable PPPPC in the IMF's *World Economic Outlook* database.

   a. Explain the `import` statement.

   b. What type of object is `data`?

   c. Why does the last line have `pd` prior to the `DataFrame` function?

   d. What type of object is `weo`?

   e. How many rows does it have? Columns?

   f. What `dtypes` are the variables/columns? What does this mean?

   g. *Challenging.* What are each of these expressions? What type?
   ```
   weo['Year']
   weo[['Year']]
   weo[[3]]
   ```

   h. *Challenging.* Find and apply a method to convert `weo['Year']` to type `float`. *Hint:* The method begins with the letter `a`.

   i. Describe the result of the statement `t = weo.tail(3)`. What kind of object is `t`? What does it look like?

   j. How would you create a new dataframe that consists of the first 4 rows of `weo`?

   k. What type of object is `weo['BRA']`?

   l. Create a new variable equal to the ratio of Brazil's GDP per capita to Japan's and add it to the ataFrame.

   m. *Challenging.* Use the `drop()` method to eliminate this (new) variable from the dataframe.

   n. What are `weo`'s row and column labels?

   o. Set the index equal to the `Year` variable.

   p. Change the names of the other variables to Brazil, Japan, and United States.

q. Export the dataframe to an Excel spreadsheet.

r. What method would you use to compute the mean for each country? What are the means?

s. *Challenging.* How would you compute means across countries for each year?

t. Plot the data by applying a `plot` method to `weo`.

u. *Challenging.* Change the colors of the lines to green (Brazil), red (Japan), and blue (US).

v. *Challenging.* Do the same plot with a log scale. *Hint:* Read the documentation for the plot method.

w. Plot Brazil on its own.

2. Use `read_csv()` to read the responses of a previous semesters entry poll from

https://raw.githubusercontent.com/NYUDataBootcamp/Materials/master/Data/entry_poll_fall16.csv

a. Read the file and assign it to the variable `ep`.

b. Describe its contents. What are the variables? The responses?

c. What data types are the variables?

d. Change the variable names to something shorter.

e. *Challenging.* Describe what this code does:

```
ep[list(ep)[1]].value_counts()
```

Suggestion: Break it into two or more statements and explain them one at a time.

3. Consider the 538 college majors data at `url`:

```
url1 = 'https://raw.githubusercontent.com/fivethirtyeight/data/master/'
url2 = 'college-majors/recent-grads.csv'
url = url1 + url2
```

The variables are described at

https://github.com/fivethirtyeight/data/tree/master/college-majors

a. Create a dataframe `df538` from the csv file at `url` using `read_csv()`. What are its dimensions?

b. What argument/parameter would you use to read only the first ten lines of the file?

c. Extract the variables numbered `[2, 4, 15, 16, 17]`. What are the names of these variables? What do they represent?

d. Set the index equal to `Major`.

e. Use the `sort_values()` method to sort the data by `Total`.

f. What code would you use to extract the ten majors with the greatest number of people?

g. *Challenging.* Construct horizontal bar charts of the top ten majors sorted, first, by median salary and, second, by the salary of the 25th percentile. In each case plot just the variable you sorted on.