HW2P4

September 21, 2016

0.0.1 Homework 2: More Exploratory Data Analysis

0.1 Gene Expression Data and Election Polls

Due: Thursday, September 29, 2016 11:59 PM

Submission Instructions To submit your homework, create a folder named **lastname_firstinitial_hw#** and place your IPython notebooks, data files, and any other files in this folder. Your IPython Notebooks should be completely executed with the results visible in the notebook. We should not have to run any code. Make sure to share the private repo with my github account (mdog) and submit the repo path through blackboard.

0.2 Introduction

John Tukey wrote in Exploratory Data Analysis, 1977: "The greatest value of a picture is when it forces us to notice what we never expected to see." In this assignment we will continue using our exploratory data analysis tools, but apply it to new sets of data: gene expression and polls from the 2012 Presidental Election and from the 2014 Senate Midterm Elections.

First: You will use exploratory data analysis and apply the singular value decomposition (SVD) to a gene expression data matrix to determine if the the date that the gene expression samples are processed has large effect on the variability seen in the data.

Second: You will use the polls from the 2012 Presidential Elections to determine (1) Is there a pollster bias in presidential election polls? and (2) Is the average of polls better than just one poll?

Finally: You will use the HuffPost Pollster API to extract the polls for the current 2014 Senate Midterm Elections and provide a preliminary prediction of the result of each state.

Data We will use the following data sets:

1. A gene expression data set called `exprs_GSE5859.csv` and sample annotation table called `sampleinfo_GSE5859.csv` which are both available on Github in the 2014_data repository: expression data set and sample annotation table.

2. Polls from the 2012 Presidential Election: Barack Obama vs Mitt Romney. The polls we will use are from the Huffington Post Pollster.

3. Polls from the 2014 Senate Midterm Elections from the HuffPost Pollster API.

0.3 Load Python modules

```
In [1]: # special IPython command to prepare the notebook for matplotlib
        %matplotlib inline

        import requests
        from io import StringIO
        import numpy as np
        import pandas as pd # pandas
        import matplotlib.pyplot as plt # module for plotting
        import datetime as dt # module for manipulating dates and times
        import numpy.linalg as lin # module for performing linear algebra operatio
```

0.4 Problem 4

In this last problem, we will use the polls from the 2014 Senate Midterm Elections from the HuffPost Pollster API to create a preliminary prediction of the result of each state.

The HuffPost Pollster API allows you to access the data as a CSV or a JSON response by tacking ".csv" or ".json" at the end of the URLs. For example the 2012 Presidential Election could be accessed as a .json instead of a .csv

Problem 4(a) Read in the polls for **all** of the 2014 Senate Elections using the HuffPost API. For example, we can consider the 2014 Senate race in Kentucky between Mitch McConnell and Alison Grimes.

To search for the 2014 Senate races, use the `topics` parameter in the API [read more about topics here].

```
In [36]: url_str = "http://elections.huffingtonpost.com/pollster/api/charts/?topic=
```

To list all the URLs related to the 2014 Senate races using the pollster API, we can use a list comprehension:

```
In [37]: election_urls = [election['url'] + '.csv' for election in requests.get(url
         election_urls

Out[37]: [u'http://elections.huffingtonpost.com/pollster/2014-kentucky-senate-mccon
          u'http://elections.huffingtonpost.com/pollster/2014-arkansas-senate-cotto
          u'http://elections.huffingtonpost.com/pollster/2014-michigan-senate-land-
          u'http://elections.huffingtonpost.com/pollster/2014-louisiana-senate-cass
          u'http://elections.huffingtonpost.com/pollster/2014-new-hampshire-senate-
          u'http://elections.huffingtonpost.com/pollster/2014-west-virginia-senate-
          u'http://elections.huffingtonpost.com/pollster/2014-new-hampshire-senate-
          u'http://elections.huffingtonpost.com/pollster/2014-north-carolina-senate
          u'http://elections.huffingtonpost.com/pollster/2014-virginia-senate-gille
          u'http://elections.huffingtonpost.com/pollster/2014-colorado-senate-gardn
          u'http://elections.huffingtonpost.com/pollster/2014-illinois-senate-oberw
          u'http://elections.huffingtonpost.com/pollster/2014-alaska-senate-sulliva
          u'http://elections.huffingtonpost.com/pollster/2014-iowa-senate-ernst-vs-
          u'http://elections.huffingtonpost.com/pollster/2014-mississippi-senate-co
          u'http://elections.huffingtonpost.com/pollster/2014-oregon-senate-wehby-v
          u'http://elections.huffingtonpost.com/pollster/2014-georgia-senate-perdue
          u'http://elections.huffingtonpost.com/pollster/2014-louisiana-senate-sass
          u'http://elections.huffingtonpost.com/pollster/2014-south-dakota-senate-r
          u'http://elections.huffingtonpost.com/pollster/2014-maine-senate-collins-
          u'http://elections.huffingtonpost.com/pollster/2014-minnesota-senate-mcfa
          u'http://elections.huffingtonpost.com/pollster/2014-texas-senate-cornyn-v
          u'http://elections.huffingtonpost.com/pollster/2014-south-carolina-senate
          u'http://elections.huffingtonpost.com/pollster/2014-south-carolina-senate
          u'http://elections.huffingtonpost.com/pollster/2014-oklahoma-senate-inhof
          u'http://elections.huffingtonpost.com/pollster/2014-new-mexico-senate-weh
          u'http://elections.huffingtonpost.com/pollster/2014-new-jersey-senate-bel
          u'http://elections.huffingtonpost.com/pollster/2014-idaho-senate-risch-vs
          u'http://elections.huffingtonpost.com/pollster/2014-tennessee-senate-alex
          u'http://elections.huffingtonpost.com/pollster/2014-wyoming-senate.csv',
          u'http://elections.huffingtonpost.com/pollster/2014-kansas-senate-roberts
          u'http://elections.huffingtonpost.com/pollster/2014-hawaii-senate-cavasso
          u'http://elections.huffingtonpost.com/pollster/2014-oklahoma-senate-lankf
          u'http://elections.huffingtonpost.com/pollster/2014-montana-senate-daines
          u'http://elections.huffingtonpost.com/pollster/2014-rhode-island-senate-z
          u'http://elections.huffingtonpost.com/pollster/2014-massachusetts-senate-
          u'http://elections.huffingtonpost.com/pollster/2014-delaware-senate-wade-
```

Because there so many Senate races, we can create a dictionary of pandas DataFrames that will be keyed by the name of the election (a string).

```
In [38]: def build_frame(url):
             """
             Returns a pandas DataFrame object containing
             the data returned from the given url
             """
             source = requests.get(url).text

             # Use StringIO because pd.DataFrame.from_csv requires .read() method
             s = StringIO(source)

             return pd.DataFrame.from_csv(s, index_col=None).convert_objects(
                     convert_dates="coerce", convert_numeric=True)
```

```
In [39]: # Makes a dictionary of pandas DataFrames keyed on election string.
         dfs = dict((election.split("/")[-1][:-4], build_frame(election)) for elect
```

Show the head of the DataFrame containing the polls for the 2014 Senate race in Kentucky between McConnell and Grimes.

```
In [40]: #your code here
```

Problem 4(b) For each 2014 Senate race, create a preliminary prediction of the result for that state.

```
In [42]: #your code here
```

1 Submission Instructions

To submit your homework, create a folder named **lastname_firstinitial_hw#** and place your IPython notebooks, data files, and any other files in this folder. Your IPython Notebooks should be completely executed with the results visible in the notebook. We should not have to run any code. Make sure to share the private repo with my github account (mdog) and submit the repo path through blackboard.

```
In [ ]:
```