Computation, Problem Set #2, Visualization and Pandas

OSM Lab, Jan Ertl

Due Tuesday, July 3 at 6:00pm

Do the following Exercises from the Brigham Young University Applied Mathematics and Computational Emphasis (ACME) Python labs Humpherys and Jarvis (2017) and from Richard Evans.

1. **Exercises from ACME: Intro to Matplotlib lab.** Do problems 1 through 6 from Intro to Matplotlib lab. You will need to download the FARS.npy file, which is saved in the course repository.

2. **Exercises from ACME: Data Visualization lab.** Do problems 1 through 6 from Data Visualization lab. You will need to download the anscombe.npy, MLB.npy, earthquakes.npy, and countries.npy files, which are saved in the course repository.

3. **Exercises from ACME: Pandas 1 lab.** Do problems 1 through 6 from Pandas 1 lab. You will need to download the crime_data.txt and titanic.csv files, which are saved in the course repository.

4. **Exercises from ACME: Pandas 2 lab.** Do problem 1 from Pandas 2 lab. You will need to download the titanic.csv file, which is saved in the course repository.

5. **Exercises from ACME: Pandas 3 lab.** Do problems 1 and 2 from Pandas 3 lab. You will need to import the `iris`, `poisons`, and `diamonds` datasets from the `pydataset` module using command `import pydataset as data`. You will also need to download the titanic.csv file, which is saved in the course repository.

6. **Exercises from ACME: Pandas 4 lab.** Do problems 1 through 6 from Pandas 4 lab. You will need to download the DJIA.csv, paychecks.csv, finances.csv, and website_traffic.csv files, which are saved in the course repository.

7. **A lifetime of temperatures.** Assume the following lifetime of events for an individual named Ricardo.

Date	Location	Description
Jan-22-1975	Indianapolis, Indiana	Born
Aug-01-1980	Pittsburgh, Pennsylvania	moved to Pittsburgh
Jul-14-1988	Pittsburgh, Pennsylvania	little league all-star team wins regional championship
Aug-17-1993	Miami, Florida	began university
May-05-1998	Washington, DC	takes first full-time job
May-15-2003	Washington, DC	gets married
Jun-03-2006	Chicago, Illinois	takes second full-time job
Oct-01-2010	Chicago, Illinois	first child born
Oct-31-2016	Chicago, Illinois	died of happiness while writing Python code

Go to the National Centers for Environmental Information Climate Data Online Search page to get the temperature data for the five cities above in the correct time periods. Use U.S. historical temperature data to make a scatterplot of temperature for each day of Ricardo's life. Just get the maximum and minimum temperatures for each day.

(a) On the x-axis plot day of the one complete year. Let day 1 be September 21 (the first day of Fall), and let day 366 be September 20 (the last day of summer). This means that you have to include the extra day in leap year (every four years) of February 29. For example, if an individual lived for 20 years, then each day of the year on the plot would have twenty days of temperature data.

(b) On the y-axis plot the high temperature and low temperature (in degrees Fahrenheit) for each day of the individual's life.

(c) Choose a marker size that is small enough to see the data fairly well. Let the general markers be circles with black marker fill color.

(d) Use maroon marker fill color (University of Chicago color) to highlight the temperatures from during Ricardo's Illinois period.

(e) Use yellow marker fill color with bold black marker outline for the life events of "born", "little league all-star team wins regional championship". Place descriptive text with pointer lines near these markers.

(f) Put some form of descriptive label on the x-axis and y-axis. An example of a D3 version of a similar plot for another person is here.

8. **3D histogram.** Read in the data file `lipids.csv`. These data are cholesterol and trigliceride measurements from the blood of 371 individuals, 51 of which had no history of current evidence of heart disease and 320 of which have evidence of heart disease. Make the following histograms and answer the following questions using only the observations for the 320 diseased individuals (`diseased==1`).

(a) Make a 2D (one-dimensional) frequency (not count) histogram of the cholesterol variable with 25 equally spaced bins. Save this histogram to your `images/` directory. What is the midpoint of the bin with the highest frequency?

(b) Make a 3D (two-dimensional) frequency histogram with 25 equally spaced bins in both the cholesterol and trigliceride dimensions. Save this histogram to your `images/` directory. What is the key new characteristic that emerges from the data?

(c) If you had to interpret the findings from the 3D histogram, what group or groups might you say have the highest risk for heart disease?

9. **Comparing segments of time series.** Read in the U.S. total nonagricultural jobs time series data from the file `payems.csv`. This file contains the month of the jobs total as well as the total number of jobs in thousands (you take the number in the data and multiply it by 1000 to get the real total). This tells you the historical time series of total number of workers employed in a given month in the United States. The data are annual from 7/1/1929 to 7/1/1938 and are monthly from 1/1/1939 to the 5/1/2018.

   The NBER business cycle dates webpage gives the beginning and ending dates of every recession back to 1857. The Great Depression is the peak-to-trough from August 1929 to March 1933. The U.S. has had 14 recessions from the Great Depression to the present period. The most recent U.S. recession, termed the Great Recession, was the longest since the Great Depression (18 months). You will create a plot that compares job growth in each or these most recent 14 recessions. Save your plot in the `images/` folder. An example of this type of plot for stock prices is here.

   (a) Create 14 segments of the job growth data series that start one year (12 months) before the peak date of the recession and end 10 years and 5 months after the peak at the beginning of the recession. Note that you will not have one year of prior data for the Great Depression.

   (b) Normalize each of the 14 series such that the jobs level at the peak date equals 1. That is, divide each element of a given series by the jobs level at the peak of that series. This transforms each series to a percent change from the peak jobs level.

   (c) Plot each of the 14 series as a line plot (with no markers) on the same axes with the peak jobs level of 1 on top of each other for each series.

   (d) Let each line plot be a different combination of color and linestyle.

   (e) Make a legend outside of the axes on the right that gives the beginning date of each recession as its label.

   (f) Label the y axis "`Jobs/peak`" and label the x axis "`Time from peak`".

   (g) Label the x-axis with the following 9 labels: `-1yr, peak, +1yr, +2yr, +3yr, +4yr, +5yr, +6yr, +7yr, +8yr, +9yr, +10yr`.

(h) Make a dashed grey horizontal line at `Jobs/peak=1` and a dashed grey vertical line at `peak`.

(i) Make the line plot for the Great Depression black and solid, and make the line plot for the Great Recession red and sold. Make both line plots thicker than the other 12 plots.

(j) Are there any U.S. recessions beside the Great Depression that have been worse than the Great Recession in terms of jobs?

(k) Are there any ways in which the Great Recession has been worse than the Great Depression in the United States?

References

Humpherys, Jeffrey and Tyler Jarvis, "Computational Labs for Foundations of Applied Mathematics, Volumes I and II," 2017.