Homework 1: CSc 599.69 Visualization, Spring 2017

Homework : CSc 599.69 Visualization, Spring 2017, Due Friday Feb. 10th

Instructor: Michael Grossberg

Submission through blackboard. Ipython notebooks in a repo for some problems, for others (problem 1,2) look for submission points in blackboard. They may not be up now but they will later.

Problem 1: Complete Python for Data Science

Complete the "Intro to Python for Data Science" and submit through blackboard proof of your completion.

Problem 2: Design Critique

Create a thread on blackboard to answer this question. Here is a source of good visualizations:

http://flowingdata.com/

and a source of less good ones:

http://wtfviz.net/

Please pick and critique one good (from flowing data) and one less good (from wtfviz) visualization.

Label and answer each of the following questions. Looking for at least 2 sentences per question. Be specific. Vague answers that lack depth will lose points.

1. Who is the audience? (expert? non-expert?)
2. What questions does this visualization answer?
3. What design principles best describe why it is good / bad?
4. Why do you like / dislike this visualization?
5. Can you suggest any improvements?

Make a separate thread for each visualization (total 2 threads) and link to the visualization. Comment on two other students critiques (agree/disagree). In your comment you must make a significant points. Don't just say "cool", or "dude that sucks".

Problem 3: Make line plots in Python

For problems 3-5 you will need to make and account on github. You will share the repo with me. My github username is "mdogy". The repo should be called CSc59969S17HW1, and put each problem in a separate folder. Put this one in folder called prob2.

We will make some simple line plots in matplotlib. We will use the U.S. Food Commodity Availability by Food Source, 1994-2008

https://www.ers.usda.gov/publications/pub-details/?pubid=81817

You will use the Appendix B: Consumption share tables with the link

https://www.ers.usda.gov/webdocs/publications/err221/appendix%20b%20(shares).xls?v=42725

You should open this file in excel or libre office to start, in order to make sense of the file. It uses worksheets and the columns are kind of messy because rows are grouped into sections. You should save it with an easy to use name like "USFoodCommodity.xls"

First of all there are 4 time periods considered: 1994-1998, 2003-2004, 2005-2006, and 2007-2008. For each time period there are two worksheets, a Food At Home Consumption (FAH), and a Food Away From Home (FAFH). Lets focus only on Food At Home.

Each worksheet is grouped into groups of rows the first being "LAFA away-from-home consumption amounts: 2007-08 means and confidence intervals for U.S., children, and adults", the second being "Appendix Table C-8 (cont'd). LAFA away-from-home consumption amounts: 2007-08 means and confidence intervals for boys, girls, men, and women," and so forth. Don't reformat with excel or open office. Use the "pandas.read_excel" function to load the data. Note that you will need to use the sheetname properly to select the sheets we need (FAH for each time). You will extract the data you need from the loaded dataframes and create new data frames with just the information you need, and the proper headers.

Lets concentrate on Men and Women and the at home consumption of a few kinds of fruit and dairy products to see if there are any trends. Make 4 line plots using python:

1. Fruit types over time (Men)
2. Fruit types over time (Women)
3. Dairy products over time (Men)
4. Dairy products over time (Women)

The fruit types you should consider will be "Apples as fruit", "Bananas", "Berries", "Grapes", "Melons", "Oranges, Total", "Other citrus fruit", "Stone fruit", "Tropical fruit". For dairy products you should look at "Fluid milk total", "Butter", "Cheese", "Yogurt", "Dairy, Other". Each product should have a line with its own color, line style. The graphs should have each axis labeled with the variable measured and units. There should be a legend to understand which product is which line. The data source and source url should appear below the graph in the image. This should be part of the same ipython notebook with the code extracing the data. We are going to just use mean pounds here, which, for Men, appears in the 7th column, after the 77th row. You will figure the details out through inspecting the sheets in a spreadsheet program and checking that you get the same numbers in the ipython notebook.

Below each plot image, use a markdown cell to write a caption. The caption should quickly summarize what is in the fig (2 sentences). Those summaries should fill in

some details in the attributes not obvious from reading the titles, legends and axis. You might get this information from the report summary.

Each caption should say something interesting thing we conclude from the figure. Like "in the figure we see an increase in the consumption of rabbit eyeballs", or "the overall consumption of bird droppings remained flat but rained higher than red grapes." Just to be clear, do not use these quotes as captions. The analysis should be your own.

Problem 4: Make a bar graphs in Python

For problems 3-5 you will need to make and account on github. You will share the repo with me. My github username is "mdogy". The repo should be called CSc59969S17HW1, and put each problem in a separate folder. Put this one in folder called prob4.

Go back to the data from problem 3. Compute the percent increase or decrease from 1994-1998 to 2007-2008 for each product. Again make 2 charts: one for men and one for women. Here group the bars reprenting fruit products, and dairy products separately with a space between them. Within each category of fruit or dairy, sort from largest decrease to largest increasing. Rather than legend, here use x axis to label by product type and use a 45 degree slant so that the labels all fit.

Problem 5: Data Set Types/Ipython Slides

For problems 3-5 you will need to make and account on github. You will share the repo with me. My github username is "mdogy". The repo should be called CSc59969S17HW1, and put each problem in a separate folder. Put this one in folder called prob4.

In this problem I want you to think about different data set types. Read the textbook to understand different data types. Find an example of a visualization (image) of each of these data types

(1) 1D Time Series

(2) 2D Scaler Field

(3) 3D Scaler Field

(4) 2D Vector Field

(5) 3D Shape

(6) Graph (not tree)

First collect (download) the images you find. Just to make this a double challenge, you will make this as an ipython slideshow. You will use the command

```
from IPython.display import display, Image
display(Image(filename='misc-you-dont-say-1.png'))
```

to display in images within your ipython notebook as "results". Next watch this: http://conference.scipy.org/scipy2014/schedule/presentation/1718/

and figure out how to make slides with your ipython notbook and export them to html/js slideshow in your repo. With each of the 6 examples you should have a short explanation of how this is a visualization of the data (what is the data and how it is this data type).