

Exploratory Data Analysis (Visualization)

In groups of 2-3, you will be using the data visualization techniques discussed thus far in class to glean information from new datasets that you select yourself.

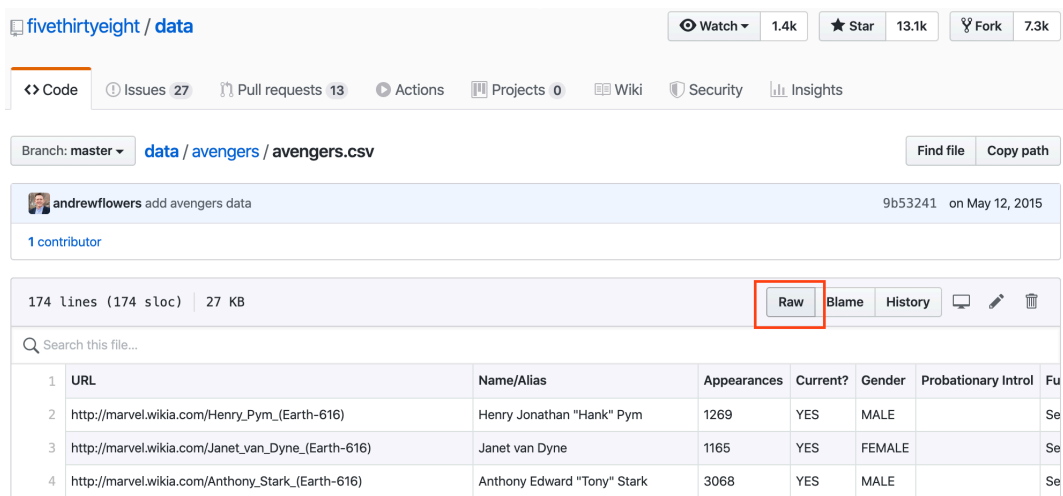
Pick one person to type up your group's code and one person who will present your results to the class.

There is an (incomprehensible) amount of data available on the internet. Some of that data has luckily already been curated for us, for example in the repository at:

<https://github.com/fivethirtyeight/data>

Each directory in this repository contains a description of the data, as well as a comma-separated values (.csv) file that contains the actual data.

1. As a group, spend ~10 minutes clicking on the different datasets on that page and reading their descriptions, noting ones that your group thinks might be interesting to explore.
2. Select a dataset to focus on – you will be focusing on this dataset for the next couple of classes, so make sure everyone likes the dataset. Once you've picked a dataset, let the Professor know so you can have it approved.
3. Once I've approved your dataset, download the .csv file to your CS1070 directory. To do this, you will need to click on the link to the file. This will open it up on a github page, but you'll then need to click the button for the "raw" file before you can download it:



The screenshot shows the GitHub repository page for `fivethirtyeight/data`. The file `data/avengers/avengers.csv` is selected. The page shows the file's history, with the latest commit by `andrewflowers` on May 12, 2015. The file size is 27 KB. The 'Raw' button is highlighted with a red box. Below the file information, a search bar is visible, and a table of data is shown.

1	URL	Name/Alias	Appearances	Current?	Gender	Probationary Introl	Fu
2	http://marvel.wikia.com/Henry_Pym_(Earth-616)	Henry Jonathan "Hank" Pym	1269	YES	MALE		Se
3	http://marvel.wikia.com/Janet_van_Dyne_(Earth-616)	Janet van Dyne	1165	YES	FEMALE		Se
4	http://marvel.wikia.com/Anthony_Stark_(Earth-616)	Anthony Edward "Tony" Stark	3068	YES	MALE		Se

This will open the plain .csv file in your browser. Click File – Save As (or equivalent) to save the .csv file to your CS1070 directory. (if you don't follow these steps, the file you download might be a webpage rather than a .csv file)

4. Run your jupyter notebook and create a new notebook named “exploratory_visualization.ipynb”
5. import pandas
6. Copy and paste the load_file function from our previous notebooks. Use this function to load your .csv file into a pandas DataFrame.
7. Pandas has the four different visualization tools that we discussed last class built in. Read the documentation for how to use these functions at:
 - <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.plot.line.html>
 - <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.plot.bar.html>
 - <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.plot.scatter.html>
 - <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.plot.hist.html>

In particular, focus on the example code that shows how you can generate each of the plots directly from a DataFrame.

8. Create a variety of plots that help visualize the dataset that you downloaded – pick different columns to plot over time, columns to compare against each other in a scatter plot, etc. The professor will walk around and help suggest plots to generate.

Each plot should only require ~1-3 lines of code to generate. Hint: you might want to review the code from last class’s exercise to help remember how to grab specific rows or columns of your DataFrame. Another hint: google is your friend – for example:

<https://www.google.com/search?q=get+row+from+pandas+dataframe>

9. Select the plots that you think are most interesting to present to the class and share in a brief presentation (~3 minutes each). Your presentation should:
 - Explain what data is contained in the dataset you selected (and perhaps why your group thought it was interesting)
 - Describe the plots you’ve selected – explaining what the axes of the plot are, and what you think the plots communicate about the data
 - Explain why the type of visualization you used was the appropriate one for the data you’re visualizing.