HW 7: Data with Pandas

Python DeCal- Spring 2024

Due on 4/3/2024 at 11:59 pm

In this homework assignment, we'll explore how to use Pandas with some interesting real-life datasets. Using the basic understanding we developed in class, and your knowledge of looking through the pandas documentation, you'll be producing some pretty cool data analysis and visualizations.
If there are any questions, please feel free to come in to our office hours, reach out on Ed, or send any of us an email.

1 Problem 1: Tweet Analysis

This problem will take a look at some of AOC's tweets. AOC's tweets are stored in a file called `AOC_recent_tweets.txt` which you are provided with.

- Our first step is going to be loading in this dataset. Load in her tweets and index it by the `id` column. This should be saved as a Pandas DataFrame in your Jupyter notebook, please display this in your submission.

- In the DataFrame above, we can see that there is a `created_at` column that tells us what time AOC tweeted out this specific tweet. But, this data is currently stored in a little bit of an odd format. In this question, write a function called `time_in_hours` that will take in a column like the `created_at` column which is in the datetime format, and converts it into hours.

  This is the conversion that you will use in the function:

  $$dec.hour = hour + \frac{minute}{60} + \frac{second}{60^2}$$

  Hint: Look up how to use datetime accessors to do this. DateTime is new Python datatype!

- Now that you have a function to change the datetimes into decimal hours, Take the data in the `created_at` column and convert it to decimalized hour to create a new `hours` column.

- Lastly, save a DataFrame with just the columns: `created_at,hours` and `full_text` as a CSV file to your system. Please submit this CSV file alongside your Jupyter Notebook submission for this assignment.

2 Problem 2: Planets Planets Planets!

In this question, we're going to explore a planets dataset that is one of the many example datasets in Seaborn (some of these are super cool to explore, so if you want to know more, check out this link).This dataset contains information about various exoplanets based on NASAs exoplanet catalog. The columns provide information of when these exoplanets were found, how they were found, and some of their basic properties (mass,period,distance,etc...).
Now, we want to recreate two famous plots in the exoplanet community. Though you can use MatPlotLib to create these visualizations, we recommend that you use Seaborn (especially for the second plot).

To access the dataset, use `sns.load_dataset('planets')` which should give you a Pandas DataFrame. Make sure to store this dataframe as you will use it for the visualizations described below.

- The first should be a scatter plot of orbital period on the x-axis and mass on the y- axis of all of the exoplanets in the catalog, with individual points colored by discovery method. Please ensure to include clear titles, labels and a legend for your plot and make sure that the plot is displayed in your submission.

  Hint: Perhaps try a log scale to make your data display well.

- The second should be a bar chart of how many exoplanets were discovered by year, and also should be categorized with different colors according to the different discovery methods.Once again, please ensure to include clear titles, labels and a legend for your plot and make sure the plot is displayed in your submission.

  Hint: It would be a good idea to remove NaNs from this dataset.
  Another Hint: This will be a stacked barplot. What parameters can you use in Seaborn to enable this?