Weekly Challenge #1 | Exploring Immigration Data

In this weekly challenge, we'll be exploring data on H2 visa applications, which are requests by companies to hire foreign workers for non-agricultural jobs within the United States. The U.S. Department of Labor releases this data as Excel (.XLS) files and BuzzFeed News has been tracking and hosting these datasets on Github. The dataset we'll be working with represents H2 applications from October 1, 2010 to March 31, 2016.

Since you'll be working on this challenge on your own computer, you'll need to have Python and Pandas installed. The workflow we recommend is writing and running your code in Jupyter Notebook so your work is stored in a notebook file. Then, you can upload the notebook to your Github and send the link to the notebook to members in the community to get help. Lastly, once you've finished the project, the notebook contains your work and thought process and you can add it to your portfolio.

Links

Install Python and Pandas:
https://www.dataquest.io/mission/118/project-python-and-pandas-installation/)

Installing and using Jupyter Notebook:
https://www.dataquest.io/mission/162/guided-project-using-jupyter-notebook/

Example rendered Jupyter Notebook on Github:
https://github.com/dataquestio/solutions/blob/master/Mission216Solutions.ipynb

Raw data from US Dept of Labor (XLS files):
https://github.com/BuzzFeedNews/H-2-certification-data/tree/master/data/raw

Cleaned up, aggregated CSV file (right click **Raw** and save the file to your computer):
https://github.com/BuzzFeedNews/H-2-certification-data/blob/master/data/processed/H-2-certification-decisions.csv

Read more about H2 visas and how they work:
https://www.uscis.gov/working-united-states/temporary-workers/h-2b-temporary-non-agricultural-workers

Data Exploration

1. Read about how BuzzFeed compiled the data and what each column represents in the Readme file - https://github.com/BuzzFeedNews/H-2-certification-data#standardized-data

2. Read the CSV file into a Pandas Dataframe and understand the different options and ranges of values for the most important columns:
    a. **visa_type**
        i. How many different visa types and how many cases for each?
    b. **last_event_date**
        i. How are date values represented?
    c. **case_status**
        i. Probably the most important column to understand the different values since this describes the status of the application.
        ii. This column contains many values, which can be simplified further for analysis. Create a new column with simplified values for easier analysis.
    d. **n_requested**
        i. Calculate summary statistics and look for any outliers!
    e. **N_certified**
        i. Calculate summary statistics and look for any outliers!
    f. **job_title**
        i. How many different jobs represented?
    g. **organization_flag**
        i. What are the different types of organizations and the distribution of values?
3. For fields with many missing values, determine which ones have enough useful values to keep and which ones you should avoid using in your analysis.
4. Use the **is_duplicate** field to remove duplicates from the dataset.
5. Since each row doesn't represent an application for a single worker, any data analyses done will be on a case level not an individual worker level. Brainstorm ideas to modify the data itself or your analysis when trying to do data analysis that's centered on workers themselves (over the cases / applications).
6. Based on your initial exploration, clean up the dataset accordingly.

Data Analysis

To answer these questions, use a combination of quantitative measures and data plots to answer these questions more thoroughly.

1. How has the number of approved applications changed over time?
2. How have the number of approved **workers** changed over time?
3. Which locations and types of businesses request the most visas and how has this changed over time?
4. Which positions are the most frequently requested for these visas?

Bonus

Look for new dataset(s) to deepen your analysis. Here are some ideas:

- What datasets can help you further segment the **job_title** column?
- What datasets, if any, can help you further segment the **employer_name** column?