You and your teammates form a **journalism department**. Your task is to provide interesting insights or analysis based on the file listed below.

Instead of worrying about "staying within the lines," be creative in how look for insights and in how you present them. In that sense, this is a very easy assignment. As long as you fulfill the requirements, you will get full marks for this assignment.

**However, your final notebook, containing your names, will be shared with all of your classmates** (who will be asked to not share it publicly).

I expect your final notebook to at least contain basic exploration of data. At least basic data cleaning. I expect you to show some interesting insights you were able to glean from the data set, supplemented with visual charts.

**Requirements:**

-Download file from https://www.foreignlaborcert.doleta.gov/performancedata.cfm#dis the latest version PERM "FY2020 Q3.xlsx"

-Data dictionary available under "PERM Record Layout.pdf" link

-Form groups of three and explore the data

-Do quick exploration of data

-Create histograms of variables

-Pay attention to which ones you like

-Investigate data exploration tools, some ideas:

-pandas profiling https://github.com/pandas-profiling/pandas-profiling

-dabl https://amueller.github.io/dabl/dev/index.html (from one of the core contributors of Scikit-Learn)

-think about joining other datasets to augment this file

-Come up with ideas about what you want to explore. Potential ideas:

-Which cases are more likely to succeed given a profession, city, lawyer, filing date?

-Use tf/idf or word vectors or whatever to summarize professions (NLP)

-What are the top newspapers for each city vs profession (blue collar vs white collar)

-Don't try to investigate everything! Pick a direction, let other teams follow their own paths

-Clean/normalize the data. Potential ideas:

-missing values?

-pay in terms of hourly vs annual?

-why are some descriptions all caps?

-Clearly communicate results.

-Use lessons from "Leadership" class (if you've taken it)

-But use jupyter, not power point (unless you want to use the RISE plugin for Jupyter or something similar)

-Create nice visualizations (optional)

-You must use GitHub to manage this project and coordinate among your teammates

-This should be very helpful: https://github.com/jupyterlab/jupyterlab-git

-This is an alternative, probably not as user friendly https://github.com/jupyter/nbdime

-Since this is a team project, you must come up with a good methodology for managing changes, here are some suggestions:

-http://drivendata.github.io/cookiecutter-data-science/

-https://github.com/quantumblacklabs/kedro

-Just give these links a quick read. You don't have to install any packages. Just pay attention to their recommendations.

-Pay attention to their directory structure and how they organize their notebooks, data assets, etc.

-Pay attention to how they name their files

-Pay attention to what other lessons you can learn from them, but trying to follow every suggestion will likely be a waste of time for such a short project

Ethics:

-This data contains LOTS of private information. Use it with care.

-Would you still be ok if your analysis became public, perhaps viral?

-Expect your anaysis to be challenged. Challenge the analysis of other teams

Deliverable:

-I expect one main python notebook containing your analysis

-You can upload a zipped folder which contains supporting notebooks which contain

-Your basic exploration

-Experiments you may have done

-A README file telling me where to find the main notebook

-Everything, except the main notebook is optional

General notes:

-You may use any package and learn from anyone (classmates, resources on the internet), but may not directly copy

-Feel free to reach out to your colleagues to see what they are doing. Set up zoom calls for walkthroughs of each other's work

-Investigate visualization packages, some ideas to get you started: matplotlib, seaborn, altair, bokeh, holoviews, bqplot

-You do not have to build prediction or clustering models, but you may build several if you like

-You can "hire" me to provide extra lectures on version control, visualization or any other topic. The cost is that you must convince at least 8 people to attend this extra lecture.

Here are some interesting notebooks I found online:

Notebook which is nicely done, although it is in French:

https://github.com/jhroy/theses/blob/master/theses.ipynb

World differences in infectious tuberculosis prevalence:

https://github.com/jadianes/data-journalism/tree/master/articles/tuberculosis-world-situation

Exploring Datasets with Python:

https://github.com/MartinSeeler/python-data-exploration/blob/master/Exploring%20Datasets.ipynb