Mid-Term Project:

During the course, you will be working on a mid-term project that takes you through a data cleansing project. You will select a dataset, perform various cleansing methods using Python to the dataset and submit your newly formatted readable dataset.

The first step is selecting a dataset. Your dataset must have a minimum of 1000 records and 15-20 variables.

The authors of our book *Data Wrangling with Python* have some suggestions where datasets can be found on Pages 130-140 of your book.

Some other helpful places to find datasets include:

- https://community.tableau.com/docs/DOC-10635
- https://www.kaggle.com/datasets
- http://www.data.gov
- http://www.science.gov
- http://data.gov.uk
- http://gss.norc.org/
- http://www.europeansocialsurvey.org

There are no restrictions on what dataset you use, other than you cannot use the specific datasets used in the book, and your dataset for the mid-term and final project must be different.

You will turn in your mid-term at the end of Week 6.

The following is due submitted to the assignment link or submit a link to your GitHub repository to the assignment link:

- Your dataset with the following transformations
    - Replace headers (*Data Wrangling with Python* pg. 154 – 163)
    - Format Data to a Readable Format (*Data Wrangling with Python* pg. 164 – 168)
    - Identify outliers and bad data (*Data Wrangling with Python* pg. 169 – 174)
    - Find Duplicates (*Data Wrangling with Python* pg. 175 – 178)
    - Conduct Fuzzy Matching (if you don't have an obvious example to do this with in your data, create categories and use Fuzzy Matching to lump data together) (*Data Wrangling with Python* pg. 179 – 188)
- Using Python, submit your results via your notebook or export your code and submit via the assignment link. You must show your code and work for full credit.
- A 250-word paper summarizing your steps and any challenges you ran into during the project. You should also outline any decisions you had to make while transforming the data (for example, if you decided to remove duplicates, how did you arrive at that decision?).

Remember – your GitHub repository can act as a portfolio for potential employers! I would highly suggest using this to submit your work, so you can fill it with good content that demonstrates the projects you are working on!