For the "Applied machine learning" taught as part of the Data Science masters program at DSI in spring 2017, I'm looking for real-world dataset with associated machine learning tasks. I'm looking for heterogeneous tabular data that has some of the following characteristics:

- mixture of continuous and discrete variables
- high cardinality discrete variables ("many levels")
- missing values
- outliers
- other types of "unclean data"
- string columns
- date columns
- correlated groups of observations

The main focus of the class is i.i.d. data, but time-series data is also of interest. Ideally the data would be public or at least I would be able to publicly share the data. I'm interested in the following tasks:

- classification
- regression
- outlier detection and anomaly detection
- feature selection / determining relevant features

I'm mostly interested in tasks that allow a quantitative evaluation of a machine learning approach, so having ground truth for some of the data, or another way of evaluating a machine learning approach would be preferred.

Contact acm2248@columbia.edu if your data set fits their needs.