**Part A Cleaning and EDA**

* Data cleaning
  + Are there missing values?
  + Are there inappropraite values?
  + Remove or impute any bad data.
* Answer the following questions for the data in each column:
  + How is the data distributed?
  + What are the summary statistics?
  + Are there anomalies/outliers?
* Plot each colmun as appropriate for the data type:
  + Write a summary of what the plot tells you.
* Are any of the columns correlated?

**Part B Writing a web scraper**

Find a public website. You must post your website domain (e.g. amazon.com) on the class piazza and get an OK before using it as every student needs to use a *different* website.

* Website
  + Collect all of the external links (there must be some on the page of your )
  + Associate the link with a textual description of it from the website.
  + Write a function to check whether the link is valid.
  + Save the external links(urls), textual description, a boolean for valid, and the last vaild datetime check to an excel file.

**List of datasets for machine learning research**

* [List of datasets for machine learning research](https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research)
* [UC Irvine Machine Learning Repository](https://archive.ics.uci.edu/ml/)
* [Public Data Sets : Amazon Web Services](https://aws.amazon.com/datasets/)
* [freebase](https://developers.google.com/freebase/)
* [Google Public Data Explorer](https://www.google.com/publicdata/directory)
* [datahub](http://datahub.io/)
* [data.gov](https://www.data.gov/)

**Data cleaning checklist**

* Save original data
* Identify missing data
* Identify placeholder data (e.g. 0's for NA's)
* Identify outliers
* Check for overall plausibility and errors (e.g., typos, unreasonable ranges)
* Identify highly correlated variables
* Identify variables with (nearly) no variance
* Identify variables with strange names or values
* Check variable classes (eg. Characters vs factors)
* Remove/transform some variables (maybe your model does not like categorial variables)
* Rename some variables or values (if not all data is useful)
* Check some overall pattern (statistical/ numerical summaries)
* Possibly center/scale variables

**Exploratory Data Analysis checklist**

* Suggest hypotheses about the causes of observed phenomena
  + Assess assumptions on which statistical inference will be based
  + Support the selection of appropriate statistical tools and techniques
  + Provide a basis for further data collection through surveys or experiments

*Five methods that are must have*:

* Five number summaries (mean/median, min, max, q1, q3)
* Histograms
* Line charts
* Box and whisker plots
* Pairwise scatterplots (scatterplot matrices)
* What values do you see?
* What distributions do you see?
* What relationships do you see?
* What relationships do you think might benefit the prediction problem?
* Answer the following questions for the data in each column:
  + How is the data distributed?
  + Test distribution assumptions (e.G. Normal distributions or skewed?)
  + What are the summary statistics?
  + Are there anomalies/outliers?
* Identify useful raw data & transforms (e.g. log(x))
* Identify data quality problems
* Identify outliers
* Identify subsets of interest
* Suggest functional relationships

Last update September 5, 2017