**Required**:

1. “Principles of Data Integration.” Doan, Halevy, and Ives.
2. “Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.” 2nd Edition. McKinney.

**Recommended**:

1. “Practical Statistics for Data Scientists: 50 Essential Concepts.” Bruce, Bruce.
2. “Data Wrangling with Python: Tips and Tools to Make Your Life Easier.” Kazil, Jarmul.
3. “The Visual Display of Quantitative Information.” 2nd Edition. Tufte.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Weekly Schedule** | | | | |
| **Week** | **Topics** | **Learning Goal** | **Evidence Required to Demonstrate Goals Are Achieved.**  **Be precise!!** | **Learning Activities That Will Produce the Evidence (quizzes, coding homework, etc.)**  **Be precise!! If it is a quiz, give some sample questions. If it is a coding assignment, give the specifications for the inputs, outputs, and algorithms to be implemented.** |
| 1 | Course overview. Garbage in, garbage out: how dirty data can impact analysis. Errors vs. artifacts. Sources of errors in data and their telltale signs in data sets. | Understand that centering and scalling | Do character variable have valid values?  Are numeric variables within range?  Are there missing values?  Are there duplicated values?  Are values unique for some variables, for example, ID variables?  Are the dates valid? Do we need to combine multiple data files? | Explore two very popular open source tools: OpenRefine and Data Wrangler |
| 2 | Exploratory Data Analysis. Visualization tools. Regular expressions. [OpenRefine](http://openrefine.org/). |  |  | data transformation which usually means that we apply a mathematical function to each data value. Perhaps the most common data transformation is centering and scaling of a single variable. For those of you with statistics background, you can think of it as calculating z-score of each observed value. That is, each data value is first reduced by the mean and then divided by the standard deviation. |
| 3 | Univariate outlier detection. Robust Statistics and Estimators. |  |  |  |
| 4 | Multivariate outlier detection. Robust Multivariate Estimation. |  |  |  |
| 5 | Case study: outlier detection in financial data – opportunity, error, or artifact? Missing values. Imputing missing values. |  |  |  |
| 6 | Resampling Techniques. Frequency outliers. |  |  |  |
| 7 | Data Sampling:  Sample and selection bias.  Confidence intervals.  Distributions: normal, t, binomial, Poisson. |  |  |  |
| 8 | Deduplication. Case study: sales data, accounting changes, and the surprise jump in revenue |  |  |  |
| 9 | Data Transformation: Restructuring, Enriching. |  |  |  |
| 10 | Univariate entity resolution (a.k.a. string matching) |  |  |  |
| 11 | Schema Matching and Mapping |  |  |  |
| 12 | Multivariate entity resolution (a.k.a. data matching) |  |  |  |
| 13 | Review |  |  |  |