**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. |  |  |  |
| 2 | Exploratory Data Analysis. Visualization tools. Regular expressions. [OpenRefine](http://openrefine.org/). |  |  |  |
| 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 |  |  |  |