COM-3590: Data Cleaning and Transformation

Course Information	
Prerequisite	COM 3640 - Programming Languages, MAT 2461 - Probability Theory, MAT 2462 - Mathematical Statistics
Description	In real-world situations, data scientists must be able to use data from many dirty, autonomous, and heterogeneous data sources that are far from being ready to be analyzed. Preparing the data for analysis (often referred to as "data wrangling") involves four different tasks: cleaning, sampling, transformation, and integration. • **Cleaning** is the detection and removal of noise, i.e., dirty data, from a data set. Speaking very broadly, an instance is considered "dirty" if it is, in some way, inaccurate or duplicate. • **Sampling** is drawing a representative subset of the population of interest from the data set. Sampling may be used either to reduce the data set to a tractable size or to isolate the population of interest from the remainder of the data. • **Transformation** involves taking an existing data set and mapping it from its existing schema to the schema required for the desired analysis. May include restructuring the schema and/or enriching it with additional data from other sources. • **Integration** is the process of combining two or more sets of data into a consistent, unified view. The data to be integrated often is stored in multiple data sources which differ in their storage formats, query languages, schema/metadata languages, and provenance. Integration occurs at both the schema and instance levels, and includes entity resolution, which is the detection of when multiple data instances refer to the same real-world entity. For each of these tasks, interactive tools are useful both for preparing small data sets as well as for investigating the general quality or structure of a large data set. When dealing with large data sets measuring in many thousands or millions of rows, however, programmatic quantitative approaches are an absolute necessity to make data preparation a realistic task. This course covers both interactive tools and quantitative approaches to each of these tasks. Because data preparation is a focus of significant R&D and small advances may have major impacts on one's productivity, the course also introduces students to the communities of research and practice that continue to advance the state of the art enabling students to stay abreast of valuable advances in this area.
Course Outcomes	• Students will be able to apply descriptive statistics to explore a data set • Students will be able to use data visualization tools to understand and explain the characteristics of a data set • Students will be able to write programs to clean data sets • Students will be able to derive from existing data sets new data sets that are ready for analysis, via transformation, integration, and sampling
Major Topics Covered in Course	•**Data Cleaning** •**Data Sampling** •**Data Transformation** •**Data Integration** •**Use of Data Visualization in all the above**
Text Book(s)	**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: 3) "Practical Statistics for Data Scientists: 50 Essential Concepts." Bruce, Bruce. 4) "Data Wrangling with Python: Tips and Tools to Make Your Life Easier." Kazil, Jarmul. 5) "The Visual Display of Quantitative Information." 2nd Edition. Tufte.
Assignments	There will be weekly programming assignments to implement the methods studied that week and apply them to a data set provided by the professor. For the first eight weeks, each assignment will involve doing an initial analysis of a data set before cleaning it, and documenting the results & recommendations reached. Students will then apply one or more approaches to clean the data and then redo their analysis and document the new results & recommendations. Both the initial results & recommendations as well as those reached after data cleaning will be compared to the "ground truth", i.e., reality, of what is happening in the business or organization. For the last four weeks, which focus on data transformation, assignments will involve taking multiple related data sets, transforming them correctly to form an accurate global picture, and testing the transformed data set against some assignment-specific metrics that will indicate if the transformation was done correctly.
Assignment Grading	Mathematically correct implementation of the given method/approach: 40% Applying the method/approach and achieving the desired data cleaning or transformation results: 60% **Assignments submitted after the due date & time receive a grade of zero.**
Exams	There will be a final exam at the end of the semester
Components of Student's Grade	Assignments: 60% Final exam: 40%
Grading Scale:	A = 93-100% B+ = 87-89% C+ = 77-79% D+ = 67-69% A- = 90-92% B = 83-86% C = 73-76% D = 65-66% B- = 80-82% C- = 70-72% F = 64 and lower
Credits	3
Weekly Schedule	Two sessions a week, 75 minutes each
Attendance Policy	• Attendance of every session is mandatory. • Every unexcused absence incurs a penalty of one point off the final exam. • Every third unexcused absence additionally incurs the penalty of the student's final grade in the course being lowered by two letter "places" (e.g. from an A- to a B, from a B+ to a B-, etc.)
Y.C. C.S. Department Academic Integrity Policy	**If you need help with any aspect of any Y.C. C.S. course, please reach out to your professor and/or TA - we are there to help. Do not under any circumstances resort to cheating or plagiarizing in any way.** All academic integrity cases in Y.C. C.S. classes will be handled as follows: **1)** Every case will be referred to the dean's office for investigation and disciplinary measures - no exceptions **2)** The first time a student is caught cheating or plagiarizing on any part of any work item (e.g. a homework assignment, exam, etc.) for any Y.C. C.S. course, he will receive a zero on the work item on which he cheated or plagiarized and have his final grade in the course lowered by an entire letter (e.g. from B+ to C+.) Repeat offenders, whether they repeat in a single

	semester or across multiple semesters, will be dealt with more stringently. These penalties have been, and will be, applied even if it means a senior not graduating and/or a student having to take the course all over again. For more information, please see Yeshiva College's Academic Integrity Policy.

Weekly Schedule	
Week	**Topics**
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.
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

Students With Disabilities

The Office of Academic Support provides services and resources designed to help students on Wilf Campus develop more efficient and effective study skills and strategies. Individual support is available in areas such as time management and organization, active reading, note-taking, exam preparation and test-taking skills. The office is located in Furst Hall, Suite 412. To schedule an appointment, call 646-592-4285 or email academicsupport.wilf@yu.edu.