Course Syllabus

SI 618 – Data Manipulation and Analysis

School of Information, University of Michigan, Fall 2019

Class sessions: Tuesdays 4-7pm, Room North Quad 2255

Instructor: Ceren Budak (cbudak@umich.edu)
- Office: NQ 3336, office hours: 1-3pm Wednesdays

GSI: Scott Henry (schenry@umich.edu)
- Office hours: 11am-1pm on Mondays NQ1270

Instructional Aides: TBD
- Office hours: TBD

Note: Some syllabus details are subject to change.

Course Overview

This course aims to help students get started with their own data harvesting, processing, aggregation, and analysis. Data analysis is crucial to evaluating and designing solutions and applications, as well as understanding user’s information needs and use. In many cases the data we need to access is distributed online among many webpages, stored in a database, or available in a large text file. Often these data (e.g. web server logs) are too large to obtain and/or process manually. Instead, we need an automated way of gathering the data, parsing it, and summarizing it, before we can do more advanced analysis. Therefore, students will learn to use Python and its modules to accomplish these tasks in a 'quick and easy' yet useful and repeatable way. Next, students will learn techniques of exploratory data analysis, using scripting, text parsing, structured query language, regular expressions, graphing, and clustering methods to explore data. R modules will be used to accomplish these tasks. Students will be able to make sense of and see patterns in otherwise intractable quantities of data.

The skills students will learn include the following:

- Introduction to Pandas and Jupyter notebooks
- Refreshers on: Fetching and parsing web content and querying SQL databases
- Big data processing
- Converting messy data into a form that can be analyzed using R
- Connect a database to R to simplify repeated analysis of changing data.
• Compute and visualize summary statistics of datasets.
• Master the specification of graphical displays using the 'grammar of graphics' via ggplot2.
• Combine the use of graphics with data manipulation to visualize relationships between variables.
• Use subscripting to select subsets of data to analyze.
• Use factors to analyze categorical data.
• Produce polished information graphics for publication.

Course Evaluation

• Homework assignments: 50%
• Labs: 20%
• Course Project: 30%

Text

Required:


Recommended:


Other related works:


(There will be multiple other sources used throughout the course, but I will note them in the slides)

Classroom Policy

Students are asked to attend class on time and remain through the entire class. Students will need to bring their laptops for the in-class labs. I don’t take attendance in this class, but we do have in-class labs that must be completed for credit. So it will be in your best interest to attend.
Original Work

Unless otherwise specified in an assignment all submitted work must be your own, original work. Any excerpts, statements, or phrases from the work of others must be clearly identified as a quotation, and a proper citation provided. Any violation of the School’s policy on Academic and Professional Integrity (stated in the Master’s and Doctoral Student Handbooks) will result in serious penalties, which might range from failing an assignment, to failing a course, to being expelled from the program. Violations of academic and professional integrity will be reported to UMSI Student Affairs. The faculty instructor determines the consequences impacting assignment or course grades; the Assistant Dean for Academic and Student Affairs may impose additional sanctions.

Accommodations for Students with Disabilities

If you think you need an accommodation for a disability, please let me know at your earliest convenience. Some aspects of this course, the assignments, the in-class activities, and the way we teach may be modified to facilitate your participation and progress. As soon as you make me aware of your needs, we can work with the Office of Services for Students with Disabilities (SSD) to help us determine appropriate accommodations. SSD (734-763-3000; http://www.umich.edu/sswd/) typically recommends accommodations through a Verified Individualized Services and Accommodations (VISA) form. I will treat any information you provide as private and confidential.

Student Mental Health and Wellbeing

The University of Michigan is committed to advancing the mental health and wellbeing of its students. If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. For help, contact Counseling and Psychological Services (CAPS) at (734) 764- 8312 and https://caps.umich.edu/ during and after hours, on weekends and holidays, or through its counselors physically located in schools on both North and Central Campus. You may also consult University Health Service (UHS) at (734) 764-8320 and https://www.uhs.umich.edu/mentalhealthsvcs, or for alcohol or drug concerns, see www.uhs.umich.edu/aodresources.

For a listing of other mental health resources available on and off campus, visit: http://umich.edu/~mhealth/

Course Requirements

You are required to bring a laptop to class, and at your first opportunity, install Python for the in-class lab assignments.

Schedule (Tentative)

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>What is due before class</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 3</td>
<td>Course Introduction</td>
<td>Install Python</td>
</tr>
<tr>
<td></td>
<td>Introduction to Pandas + Jupyter notebooks</td>
<td></td>
</tr>
<tr>
<td>Date</td>
<td>Topic</td>
<td>Lab/HW</td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------------------------------------------------</td>
<td>------------</td>
</tr>
<tr>
<td>September 10</td>
<td>Refresher on data collection and processing: (i) Fetching and Parsing Web content: HTML, XML, JSON. Web APIs. (ii) Querying data in a SQL Database</td>
<td>Lab 1, HW 1</td>
</tr>
<tr>
<td>September 17</td>
<td>Big data processing on Hadoop: MapReduce, Hive and Spark</td>
<td>Lab 2, HW 2</td>
</tr>
<tr>
<td>September 24</td>
<td>Big data processing on Hadoop: MapReduce, Hive and Spark (2)</td>
<td>Lab 3, HW 3</td>
</tr>
<tr>
<td>October 1</td>
<td>Big data processing on Hadoop: MapReduce, Hive and Spark (3)</td>
<td>Lab 4, HW 4, Data gathering and processing project proposals</td>
</tr>
<tr>
<td>October 8</td>
<td>Advanced Topics in Data Collection: Crowdsourcing + Cross Project Group Feedback session</td>
<td>Lab 5, HW 5</td>
</tr>
<tr>
<td>October 22</td>
<td>Introduction to R, RStudio and R Markdown + Mid-point short presentations</td>
<td>Install R. Data gathering and processing project reports</td>
</tr>
<tr>
<td>October 29</td>
<td>How to manipulate data frames. How to use qplot plyr. Basic statistics</td>
<td>Lab 6, HW 6</td>
</tr>
<tr>
<td>November 5</td>
<td>Smoothing and Trend-finding. Building ggplot Layer by Layer</td>
<td>Lab 7, HW 7, Exploratory data analysis project proposals</td>
</tr>
<tr>
<td>Date</td>
<td>Details</td>
<td>Details</td>
</tr>
<tr>
<td>--------------------</td>
<td>--------------------------------------------------------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>November 12</td>
<td>ggplot2 Toolbox</td>
<td>Lab 8, HW 8</td>
</tr>
<tr>
<td></td>
<td>Finding relationships between variables Time series</td>
<td></td>
</tr>
<tr>
<td>November 19</td>
<td>Exploratory Cluster Analysis</td>
<td>Lab 9, HW 9</td>
</tr>
<tr>
<td>November 26</td>
<td>Principal Component Analysis and Exploratory Factor Analysis</td>
<td>Lab 10, HW 10</td>
</tr>
<tr>
<td>December 3</td>
<td>Outlier Detection</td>
<td>Work on your projects</td>
</tr>
<tr>
<td>December 10</td>
<td>Project presentations</td>
<td>Exploratory data analysis project Reports</td>
</tr>
</tbody>
</table>