DSCI 210: Data Science

Instructor: Chris Malone

Office: Gildemeister 124C

Email: cmalone@winona.edu

Office Hours: [Link](http://course1.winona.edu/cmalone/schedule.htm)

**Text (Optional)**: e-Book: An Introduction to Data Science (Version 3) by Jeffrey Stanton, Syracuse University

**Learning Outcomes**

1. Students will be able to identify and describe the methods and techniques commonly used in data science
2. Students will be able to apply the data science process to an data science problem
3. Students will be able to extract and assemble data from a variety of sources
4. Students will be able to demonstrate the ability to clean and prepare data for an analysis
5. Students will be able to conduct basic analyses and provide visualizations of data using commonly used tools in data science
6. Students will conduct a data science project which will require the formulation of the problem, collection and assemble of appropriate data, a thorough analysis of the data, and a proposed solution that can be defended to both a technical, e.g. data scientists, and a non-technical, e.g. executive, audience.

**Assessment Method #1 -- Homework**:

Homework assignments will be given throughout the semester.  I will collect several of your homework assignments. I strongly encourage you to stay current in your homework assignments.  Late homework assignments will not be accepted after they are returned. Quizzes may be utilized as well to assess your learning. Quizzes may *or* may not be announced.

**Assessment Method #2 -- Exams**:

There will be one midterm exam and one final exam for this course. I will test your ability to make conclusions and/or extensions to current methods.  More than likely these exams will consist of an in-class portion and an out-of-class portion. If you know you are going to miss an exam, the exam *must* be taken early. Makeup exams will not be given.

**Assessment Method #3 -- Project**:

There will be one substantial required project in this course. This project will require you to implement the data science techniques learned in this class. Evaluation of this project will likely consist of a written component, oral presentation, and creation of a poster.

**Grades**:

Your grade will be determined by your performance on exams, project, homework & quizzes.  My “target” for the number of points is two exams = 250pts, project = 75 points, and homework/quizzes = 150pts.  I do no weighting, so a point is worth a point in this class.

Your final grade will be determined using the following percentages.

|  |  |
| --- | --- |
| Your Percentage | Grade |
| greater than 90%    | A |
| 80% - 90% | B |
| 70% - 80% | C |
| 60% - 70% | D |
| less than 60% | F |

**Computing**:

We will be using Excel and the R program extensively in this class. The R software is available for free from the R-project.org website

**Extras**:

* I encourage you to use a 3-ring binder for this class because class material will be a combination of note taking, handouts, and lots of computer output.
* Attendance in mandatory.  If you miss class, it is your responsibility to get the material and get yourself caught up.
* If necessary, I reserve the right to make policy changes for this course as the semester progresses.

**Course Outline**:

1. Introduction to Data
	1. Size and scope of data
	2. Data Storage and Warehousing
	3. Data Formats (e.g. time-series, shape files, etc.)
	4. Structured versus unstructured data
2. Introduction to the Data Science Process
	1. Formulating research questions
	2. Assembly of appropriate data
	3. Analysis and Reporting
	4. Repeating the process
3. Management and Preparation of Data
	1. Data integrity and quality
	2. Recoding / creating new variables
	3. Sub-setting, filtering, and collapsing data
4. Analysis and Visualization of Data
	1. Methods to summarize data
	2. Methods to visualization data
	3. Additional summary and visualization methods (e.g. aggregation, conditioning, etc.)
5. Introduction to modeling with data
	1. Introduction to regression problems
	2. Introduction to classification problems
6. Data Science Project