Course description: An introduction to extracting knowledge from data. Students are introduced to working with and analyzing big data. Topics include scraping and cleaning data, data wrangling (manipulation of large data sets), data visualization, and creating reproducible results.

Prerequisites: Successful completion of a statistics course (BA/ECO 220, BIO 206, MTH 135, MTH 335, PSY 201, PS 301, SSC 252 or AP credit for Statistics) OR successful completion of CS 152.

Required materials:


- R - free software environment for statistical computing and graphics. Available for download here [http://www.r-project.org/](http://www.r-project.org/)


Overview: Data Science is a field emerging this century. As the Internet has become pervasive, massive data sets are collected and, often, readily available. Traditional statistical methods for analysis of data are incomplete. The data scientist, like the statistician, often starts with visualization of patterns. Visualization of patterns allows the analyst to consider the data from a point of view which can complement the statistician’s point of view. Interactive graphs supplement the traditional graphs of statistics allowing new questions to be asked and answered as patterns emerge. Additionally, the data are not in nice, neat data tables. The structure of the raw data must be considered. Large data sets always require data cleaning, reformatting, combining with other data sets with similar information and/or other manipulation techniques. In this class, we will work with large data sets learning techniques that can be used with massive data sets.

Goals:

- Learn to ask questions which can be answered with data
- Understand the fundamental ideas of data science
- Obtain and prepare data for the analysis of a problem
- Create appropriate visualizations of data
- Use techniques that facilitate reproducible results
Objectives: At the end of this course, students will be able to

- Demonstrate that he or she can use R with minimal guidance.
- Use open data sites to obtain large data sets.
- Create clean and useful data tables using data manipulation techniques including, but not limited to, join, merge, subset and filter.
- Create data visualizations that both answer questions and allow new questions to be asked.
- Use and explain techniques that allow the end user to reproduce results using a new data set.

Specific learning objectives and how they are assessed can be found on the D2L site.

Expectations:

**In class:** Most class meetings will consist of both lecture and activities and examples exploring data science ideas and techniques. Please come to class prepared and willing (eager?!) to work during class time and to collaborate with your peers and to ask questions of me. This preparation will not only help you to learn the material and perform well in the course, but it will also produce a much more enjoyable learning environment for all of us. Class attendance is required, as the in-class activities should prove to be valuable learning experiences. Needless to say, you are responsible for everything presented in class, the sections assigned for reading, and assignments used for formative evaluations.

**Outside-of-class:** Extensive work for this course, typically involving 8-12 hours per week will be divided among:

- working on assignments and writing code and supporting text for your assignments and projects
- struggling with the computer software
- reviewing, completing and correcting your class notes

**Integrity:** Central to the purpose and pursuit of any academic community is academic integrity. All members of the Westminster community, including students, faculty, staff, and administrators, are expected to maintain the highest standards of honesty and integrity, in keeping with the philosophy and mission of the College. Academic dishonesty is a profound violation of this code of behavior.

The paragraph above is taken from the Westminster College Undergraduate Catalog. It is imperative that you never submit the work of others as though it is your own work nor should you ever allow anyone else to use your work without giving credit to you. The penalty for academic dishonesty in this class is minimally the grade of 0 on the assignment and, except for unusual circumstances, a grade of F for the course. Any event of academic dishonesty is reported to the Dean of the College. Other details of violations and consequences are given in the Catalog.

**Group work is expected.** Within the group work context, it is possible to misunderstand exactly what it means to be responsible for “doing your own work.” As such, I wish to define specifically
what I expect. Group study is a proper and effective way to study if all of the participants have done their full share of the work. Unless otherwise specified, you may and should discuss problems together and reach conclusions together. But it is a form of dishonesty for a student who has not attended class, read the assignment, or thought about the problem on his or her own to try to use the ideas developed by a group or claim credit for work to which he or she has not contributed. It is also a form of dishonesty to encourage or allow such practices on the part of others.

Aside from integrity issues, writing solutions on your own allows you to determine how much of the material you understand versus how much you can “do.” At some point you are responsible for understanding each and every piece of the problems and will need to be able to write the complete solutions in your own words.

Sometimes, you will be able to “find” and copy code that someone else has written. Other times, you will be expected to create your own. Unless otherwise specified, you may use code that you find on the internet or other sources (including your classmates) as long as you cite your source. If you have used someone else’s code/work and your source is not cited, I will assume you have created the code on your own. I will consider this plagiarism. If the project or assignment specifies that you are not to use outside resources, I expect you to work on your own or with your designated partner. For these assignments, you are not allowed to give or receive help from animate or inanimate objects.

Assessment: The following three methods will be used to assess the objectives of this course.

- **Class assignments:** There will short in-class computer assignments which will be started in class and completed by the next class period. (In other classes these assignments would be considered homework.) Usually, these assignments will be done in pairs. Students who are absent during the class when the assignment is started will receive a zero on that assignment and will not be allowed to turn in the assignment even if it is finished outside of class. The class assignments are graded complete and at least 90% correct, (3%), complete but between 70% and 90% (2%) or not complete (0%). Students are allowed to have one incomplete assignment over the course of the semester. Each assignment is worth 3% of the final grade. I expect to assign 9 of these assignments. Homework received after the due date and time will receive, at most, 1% and may receive less if the assignment is not complete at a 70% level.

- **Quizzes:** Weekly quizzes, given on Tuesdays, will evaluate your understandings of the required readings, lectures and your individual ability to use the software. These quizzes will typically last 15 minutes, be short answer and/or involve writing or evaluating brief code. The quizzes will be averaged together at the end of the semester. The lowest quiz will be dropped. NO makeups will be given.

- **Projects:** Four projects will be assigned over the semester. Details of each project’s requirements will be given at least two weeks prior to the due date. Generally, each project will be done with one colleague of your choosing, demonstrate understanding of specific techniques and objectives and be presented to the class. One project will be presented at URAC.
Grade Calculations:

- Class Assignments 24%
- Quiz average 16%
- Project 1 10%
- Project 2 10%
- Project 3 15%
- Project 4 25%

- For every unexcused absence beyond the first one, you will lose 5% from your final percentage. For example, if you have a 82% with 3 unexcused absences your final grade will be recorded as $82 - (3 - 1) \times 5 = 72\%$

Grade cutoffs will be no higher than A: 93, A-: 90, B+: 87, B: 83, B-: 80, C+: 77, C: 73, C-: 70, D: 60.

See you in class!