DSC 201: Data Analysis & Visualization

Introduction

Dr. David Koop
About Me

• 3rd Year at UMass Dartmouth
• Teaching: Data Science, Data Management, and Visualization
• Research: Visualization, Provenance, Geographical Data Analysis
• Contacting me:
  - Office: Dion 308A
  - Email: dkoop
  - Phone: x6692
About You

• Year?

• Programming Background:
  - DSC 101? CIS 180/181?
  - Python? Scripting languages?
  - Data Analysis?
  - Visualization?

• Why Data Science?
Course Structure

- Balance lectures with concepts and examples
- Try things out, bring a laptop if you have one
- Assignments will focus on real (or at least real-ish) data
- Exams will test both concepts and problem solving
- Concepts include data structures (CIS 360 background)
http://www.cis.umassd.edu/~dkoop/dsc201
Course Material

• Course Website
  - http://www.cis.umassd.edu/~dkoop/dsc201
  - All material will be posted there
  - myCourses for turning in assignments

• Textbook: *Python for Data Analysis* by Wes McKinney, 2012
  - Good reference for data science topics in Python
  - McKinney created the Pandas package
Course Material

• Software:
  - Anaconda Python Distribution (https://www.continuum.io/downloads): makes installing python and python packages easier
  - Jupyter Notebook: Web-based interface for interactively writing and executing Python code
  - Pandas: Python library for analyzing datasets
  - Matplotlib, Bokeh, Tableau: Visualization tools
Grading

- Assignments: 50%
- Exams: 45%
- Class Participation: 5%
- Late Policy
Exams

- Midterm: October 25 in class
- Final: December 16, 11:30am-2:30pm
- Exams may not be rescheduled. Exams can only be made up in case of a documented emergency.
Accommodation Policy

• Please contact me at the beginning of the semester and provide the appropriate paperwork from the Center for Access and Success.

• Please update me if anything changes during the semester.

• Center for Access and Success: Pine Dale Hall Room 7136, x8711, access_success@umassd.edu
Office Hours

• Monday: 3-5pm
• Tuesday: 10:30-11:30am
• Thursday: 10:30-11:30am
• and by appointment (email me)
Academic Honesty

• Do not cheat!

• You will receive a zero for any assignment/exam/etc. where cheating has occurred. Repeat offenders will fail the course.

• You may discuss problems and approaches with other students

• You may not copy or transcribe code from another source
What do you do when faced with a problem?
Suppose your car won't start
Types of Problem Solvers

• The Excuse-maker: makes excuses about things being too difficult
• The Critic: points out flaws—doesn't look for a solution
• The Dreamer: envisions goals but doesn't try to implement a solution
• The Go-Getter: fast and never gives up but does a lot of extra work
• Goal: Mix some of the attributes above but try to structure the solution process

[Problem Solving 101, K. Watanabe]
Problem Solving

• Gather information
• Identify problems (questions)
• Consider various methods and solutions
• Decide on an approach and execute
• [Loop, Mix]
Data Science (aka Modern Problem Solving)

- Information often involves (large) datasets
- Methods and solutions often involve computers
Python

• Don't worry if you don't have any experience with it!
• Programming language to get things done
• Lots of libraries so you don't have to reinvent the wheel
• Syntax is fairly readable
• Indentation (spaces) are very important
Chicago Food Inspections

- Data: Information about food facility inspections in Chicago
- Data Source: https://data.cityofchicago.org/Health-Human-Services/Food-Inspections/4ijn-s7e5/data
- Fields: Name, Facility Type, Risk, Violations, Location
- What questions are interesting about this dataset?
Chicago Food Inspections Exploration

• Based on David Beazley's PyData Chicago talk
• YouTube video: https://www.youtube.com/watch?v=j6VSAsKAj98
Homework

- Install Anaconda (Python 3)
- Try "Hello World" in python
- Chapter 1 of *Python for Data Analysis*