COMP-4447: Data Science Tools 1

Course Overview

The objective of Data Science Tools 1 is to learn various tools to perform data collection, cleanup, summarization, and visualization (exploratory data analysis [EDA]).

Objectives

At the end of the course, students should be able to:

- Understand and create reproducible data science workflow.
- Perform Git tools workflow.
- Perform data science at the command prompt. Linux command line, bash, basic awk and sed.
- Perform data collection from various web resources like webpage or RESTAPI call.
- Perform data cleanup and imputation using Pandas (groupby, apply, aggregate, pivot table, etc.).
- Perform data summarization and visualization using Pandas, Seaborn, and Matplotlib.
- Perform time series data and feature engineering.
- Should be able to use scientific Python library (NumPy, etc.) and data science libraries.

Textbooks and Materials

There are no required textbooks for this course.

Grading

Assignment/Assessment	Points	Weight of Final Grade
Week 1 Programming Assignment	3	Programming Assignments 1 - 8 are worth 60% homework weight. Percentage is distributed according to homework points.
Week 2 Programming Assignment	5	
Week 3 Programming Assignment	6	
Week 4 Programming Assignment	8	
Week 5 Programming Assignment	9	
Week 6 Programming Assignment	12	
Week 7 Programming Assignment	6	
Week 8 Programming Assignment	8	
Week 5 Midterm	12	25%
Week 9: no homework. Work on the final project and keep updating github.		
Week 10 Final Project notebook due. See the Final Project rubric for grading details.		15%

Note that Week 5 has homework and a midterm. Please plan accordingly.

The midterm will be available immediately following the Week 5 live session and will be due 24 hours prior to the Week 6 live session. You will have 1 hour and 45 minutes to complete the midterm.

Grading Scale

grade range [('A', >=93), ('A_minus', >=89), ('B_plus', >=85), ('B', >=81), ('B_minus', >=77), ('C_plus', >=73), ('C', >=69), ('C_minus', >=65), ('D_plus', >61), ('D', >=57), ('D_minus', >=53), ('F', < 53)]

Assignment and Assessment Information

Detailed instructions for the weekly programming assignments are available in the Online Campus. All assignments are due 48 hours after the live session.

Weekly Schedule

Readings should be completed before the live session of the week in which they are assigned. Ideally, readings should be completed before other asynchronous content is started.

The notebooks for async portion are located at

https://git.cs.du.edu/psnegi/online_tools1_notebooks

You may need a login name/password to get access to the folder/repository.

Attendance Policy

Attendance at all live session meetings is mandatory.

Datasets for Final Project.

We are looking around to find noisy datasets for practice.

- Datasets for data cleaning practice by Rachael Tatman
- Datasets for Data Mining and Data Science
- The EU Open Data Portal
- World Bank Open Data
- The home of the U.S. government's open data
- Awesome public datasets on github

- Web scraping, web API (for natural language processing, one can use the *New York Times*, Twitter, etc.). We will cover these topics in the course.

Final Project Details and Rubric

In Tools 1, we are concerned with data cleaning and exploratory analysis. Please select a project that has enough scope for the following activities.

For the final project, you will create a Github repository for your project and tag it with the label TOOL1_FINAL_PROJECT by the due date. The github repository must have a .ipynb notebook file with output and associated code. Having output in the notebook cell is very important if your dataset is big or we won't be able to run the notebook in a reasonable time. Also, we should be able to run your project with a Binder link. The binder link should be in the README.md file. Please check https://mybinder.org/ to see how to create a Binder link. If this service is down, this step is not required.

Your final report should read like a data-driven story/scientific study (data science). This is really important if you want to publish your story as a blog on the web or share with stakeholders. Scientific publications have their own style and content requirements.

Use code cells and markdown cells to carry out your analysis. Please write the report using the following section format guidelines. You can create more sections if it is more natural to do so, depending on your project. Please write each section like a report and address the points mentioned in the following rubric. Try to make your report more enjoyable to read.

- Proper tagging of Github repository for final report as per deadlines (0.5 = 0.25 + 0.25 points)

- Dataset and motivation slide (1 points)

  How/why the dataset was collected and a description of the metadata of your dataset.

- Actual task definition/research question (2 points)

What real-world problem are you trying to solve? What are the input and output of your analysis?

- Literature review (2 points)

What other work has been done in this area, and how is your work novel compared to others?

- Quality of cleaning (6 points, 2 points each)

- Data cleaning and type conversion activity. Please share anything unusual you faced during this activity.

- What did you do about missing values and why? Handling missing values properly is very important.

- New feature/attribute creation and data summary statistics and interpretation.

- Visualization (8 points, 2 points each)

- Data visualization activity (box plot, bar plot, violin plot, and pairplot to see relationships and distribution, etc.).

- Describe anything you find in the data after each visualization.

- What data visualization helped you understand about data distribution.

- What you did about possible outlier as per data distribution visualization. (Did you confirm with your client whether it is actually an outlier or put a disclosure statement in your notebook if you decided to remove it?)

Program Mission

Our MS in data science provides students with a broad course of study in programming, algorithms, statistics, and data management, as well as a depth of understanding in specific fields such as data mining, machine learning, and parallel systems. Graduates of the data science program go on to work in a wide variety of career settings, including business, government, education, and the natural sciences.

Honor Code and Academic Integrity

All students are expected to abide by the University of Denver Honor Code. These expectations include the application of academic integrity and honesty in your class participation and assignments. Violations of these policies include but are not limited to:

- Plagiarism, including any representation of another’s work or ideas as one’s own in academic and educational submissions.
- Cheating, including any actual or attempted use of resources not authorized by the instructor(s) for academic submissions.
- Fabrication, including any falsification or creation of data, research, or resources to support academic submissions.

Violations of the Honor Code may have serious consequences including, but not limited to, a zero for an assignment or exam, a failing grade in the course, and reporting of violations to the Office of Student Conduct.

Diversity, Inclusiveness, Respect

DU has a core commitment to fostering a diverse learning community that is inclusive and respectful. Our diversity is reflected by differences in race, culture, age, religion, sexual orientation, socioeconomic background, and myriad other social identities and life experiences. The goal of inclusiveness in a diverse community encourages and appreciates expressions of different ideas, opinions, and beliefs so that conversations and interactions that could potentially be divisive turn instead into opportunities for intellectual and personal enrichment.

A dedication to inclusiveness requires respecting what others say, their right to say it, and the thoughtful consideration of others' communication. Both speaking up AND listening are valuable tools for furthering thoughtful, enlightening dialogue. Respecting one another's individual differences is critical in transforming a collection of diverse individuals into an inclusive, collaborative, and excellent learning community. Our core commitment shapes our core expectation for behavior inside and outside of the classroom.