DATA 601: Introduction to Data Science

Dr. Chalachew Jemberie

Fall,2023

**E-mail:** [cjember1@umbc.edu](mailto:cjember1@umbc.edu)

**Consultation:** Thursday 6.30-7.00PM **Class Hours:** M 7:10-9:40pm

**Office:** TBD

**Classroom: Information Technology 237**



* For course consultation about any Data Science project or problem, please email me 8 hours in advance.
* If you have question about debugging your code, homework or project work, please schedule an appointment with our Technical Assistant, Mani Teja Gunda( [mgunda1@umbc.edu](mailto:mgunda1@umbc.edu))

# Course Description

This class prepares students for more advanced topics in data science and introduces some of the tools and notions which are frequently used in the industry. Topics include: a review of Python programming and most fundamental modules; acquisition, handling, and working with different data; exploratory data analysis with statistics; data visualization and web scraping; life cycle of data science projects and different roles in a data science team, and ethical issues in data science.

# Student Learning Outcomes

# By the end of this course, successful students will be able to:

1. Describe the key activities in a data science project and understand the role of modeling in a data science project.
2. Use popular Python packages for exploratory data analyses, data visualization, and transformations.
3. Create functions and programs that clean, merge, and transform raw data sets and evaluate their quality for a given data science project.
4. Apply basic statistical knowledge in a data science project to test and verify hypotheses.
5. Apply best practices of communication for reporting upon completing a data science project.

# Format and Procedures

This will be an in-person course and depending on the pandemic situation there might be online (synchronous) components. Computers with internet connection and working cameras are required for the online lectures. We will also be using a software called [Collaboratory Notebooks](https://research.google.com/colaboratory/faq.html) as the main medium of delivery for the lecture materials. For more information and to get familiarity with it, [please check this tutorial.](https://colab.research.google.com/notebooks/intro.ipynb) In addition to these:

1. Students will complete assigned homework, readings, quizzes.
2. Students will engage with hands-on labs and practical exercises to prepare them for challenges they may encounter in the workplace.
3. Students will occasionally present their solutions to homework assignments in class.
4. Students who are participating in the class online should be able to share both their video and audio.

# Course Requirements

## Textbook

We will not be following one single textbook in this course. Weekly reading materials and relevant course materials will be shared beforehand via Blackboard and/or the course’s Github repository. In addition to these, I will be following the logical structures of the following textbook.

* + [**Python Data Science Handbook**](https://jakevdp.github.io/PythonDataScienceHandbook/): (2016): We will follow this book in the programming with Python parts of the course. The sections related to Panda’s library are especially well written.
    - There is also a Python Data Science Handbook by Jake VanderPlas (2022)
  + [Python for Data Analysis](https://github.com/wesm/pydata-book) by Wes McKinney, O’Reilly, 2nd edition, 2017. Please also review the latest edition (<https://wesmckinney.com/book>)
  + [Python Data Analytics](https://github.com/Apress/python-data-analytics-2e) With Pandas, NumPy, and Matplotlib, by Fabio Nelli, Apress, 2nd edition, 2018

## Hardware Requirements

* + Web browser capable of running technical notebooks
  + A computer with sufficient internet speed for online lectures. Make sure that your computer has video and microphone access.

## Quizzes

There will be some quizzes in the beginning of the lectures to assess students’ understanding of the reading assignments. Also, at the end of some lectures, students might be given quizzes to assess their understanding of the covered material.

## Homework

There will be homework assigned to students roughly every two weeks. Depending on the scale of the homework, students will be given one or two weeks to submit their homework and other than exceptional circumstances this homework will be graded within a week.

Please return your solution notebooks in Blackboard with the following filename convention: Lastname\_HWXY.ipynb (e.g. Jemberie\_HW02.ipynb). If there is someone with the same last-name in the class, please add the initial of your first name between your last name and HWXY.

## Attendance

In every lecture, I will take attendance at some point during the lecture. Students who miss this part of the lecture will be considered as absent and except for medical situations (or some other formal/written excuse) no excuse will be accepted for missing a class. Attendance will contribute to 5 percent of the final grade.

# Grading

In the final grade, the assignments will have the following weights:

|  |  |  |
| --- | --- | --- |
| Attendance (14 Lectures) | — | 5% |
| Quizzes (4 quizzes total) | — | 9% |
| Homework (7 HW total) | — | 56% |
| Projects (2 Projects total) | — | 30% |

## Grading Distribution

Final letter grades will be assigned as follows: (Grades will be rounded upwards.)

|  |  |  |
| --- | --- | --- |
| 94-100 | — | A |
| 93 - 90 | — | A- |
| 87 - 89 | — | B+ |
| 83 - 86 | — | B |
| 80 - 82 | — | B- |
| 77 - 79 | — | C+ |
| 73 - 76 | — | C |
| 70 - 72 | — | C- |
| 67 - 69 | — | D+ |
| 0 - 59 | — | F |

# Schedule and weekly learning goals

The schedule is tentative and subject to change.

Perquisite Overview -Refreshment

* Introduction to Probability (Conditional Probability, Law of Total Probability, Random Variables, Probability Distributions)
* Linear Algebra-Fundamental of Matrix and Eigen Values
* Introduction of Statistics (Properties of Random Variables, Law of Large Numbers, Central Limit Theorem, Hypothesis Testing, Sampling Distribution and Statistics, Random Sampling, Normal Distribution, Binomial Distribution)
* Introduction to Jupyter notebook and Markdown
* Python Data Types-List &Tuples, Dictionaries and Sets

**Week 01:** Introduction to DATA601 and data science

* + Course logistics, resources, and policies.
  + Introduction and review of basic tools.
  + Overview of the course: what you will learn and what we will not cover.
  + Overview of the Python and Anaconda Data Science Environment
  + Creating and Executing Code in a Jupyter Notebook
  + Introduction to Markdown
  + Introduction to Python Programming
    - Coding Basics: Naming conventions, objects and their methods, built-in functions etc.
    - Extra-1: Introduction to Python
    - Extra-2: Introduction to Variables in Python

**Week2:** Python Programming

* + Conditionals
  + Loops
  + Input/output
  + Functions
  + Intro to Numpy

**Week 03:** Basics of object-oriented programming

* + Introduction to classes and objects
  + Global variables
  + Inheritance

**Week 04:** Automation and Version Control (Tentative)

* + make
  + git

**Week 05:** Basics of visualization with Python

* + Visualization with matplotlib
  + Figure, axis, titles, labels etc.
  + Adding subplots to a figure
  + Bar charts and box plots
  + Advanced plotting

**Week 06:** Pandas

* + Dataframes vs series
  + Choosing an observation
  + Choosing a variable
  + Renaming, dropping a column, sorting values in a column
  + Boolean mask and filtering
  + Detecting missing values
  + Describe, info, etc.

**Week 07:** Data Transformations - Part-II

* + Operations on series: +/-/\*, log, ranking, cutting etc.
  + Grouping and aggregation
  + Vectorized string operations
  + Pivot tables

**Fall Break**

**Week 09:** Exploratory data analysis

* + Variation
  + Typical values
  + Unusual values
  + Heatmaps
  + Patterns and models

**Week 10:** Loading and wrangling data

* + Introductions to different data formats
  + Reading csv, excel, html files
  + Handling numbers
  + Handling strings
  + Tidying data

**Week 11:** Working with date, time, and REX

* + Timezone handling
  + Moving Averages
  + Resampling methods for time series
  + Time series visualizations
  + REX

**Week 12:** Relational data (Exercise will be both in Panadas or SQL)

* + Merging, concatenation, joining datasets.

**Week 13:** Introduction to Regression (1 of 3)

* + Distributions
  + Statistics with python
  + Introduction to Linear Regression

**Week 14:** Introduction to Regression (2 of 3)

* + Linear Regression (Cont.)
  + Introduction to Logistic Regression

**Week 15:** Introduction to Regression (3 of 3)

* + Feature Selection

**Week 16:** Review and Ethics

* + Ethics, Privacy, and Data Science
  + Review

# Course Policies

## Policies on Incomplete Grades and Late Assignments

Late/incomplete assignments will be accepted if an extension has been agreed to in advance. Emergency situations will be handled on a case-by-case basis with appropriate justification or documentation.

Incomplete grades are granted only for extenuating circumstances and your request is made before the last week of class.

# Institutional Policies

## Covid-19 Policies

Please see [this Google doc](https://docs.google.com/document/d/1xWWGAR8qEzKYr7qaVHoEhvO6lyXIyn6M3M7EFZPJQgA/edit?usp=sharing) for UMBC Policies and Resources during COVID-19.

## Academic Integrity and Honesty

By enrolling in this course, each student assumes the responsibilities of an active participant in UMBC’s scholarly community in which everyone’s academic work and behavior are held to the highest standards of honesty. Cheating, fabrication, plagiarism, and helping other to commit these acts are all forms of academic dishonesty, and they are wrong. Academic misconduct could result in disciplinary action that may include, but is not limited to failure, suspension or dismissal.

Refer to the UMBC policy on Academic Integrity: [UMBC Academic Integrity Policy](https://catalog.umbc.edu/content.php?catoid=17&navoid=879&academic-integrity)

## Diversity Statement

It is my intent that students from all diverse backgrounds and perspectives be well-served by this course, that students’ learning needs be addressed both in and out of class, and that the diversity that the students bring to this class be viewed as a resource, strength and benefit. It is my intent to present materials and activities that are respectful of diversity: gender identity, sexuality, disability, age, socioeconomic status, ethnicity, race, nationality, religion, and culture. Your suggestions are encouraged and appreciated. Please let me know ways to improve the effectiveness of the course for you personally, or for other students or student groups.

## Student Disability Services

UMBC is committed to eliminating discriminatory obstacles that may disadvantage students based on disability. Services for students with disabilities are provided for all students qualified under [the Americans with Disabilities](https://en.wikipedia.org/wiki/Americans_with_Disabilities_Act_of_1990) [Act (ADA) of 1990](https://en.wikipedia.org/wiki/Americans_with_Disabilities_Act_of_1990), [the ADAAA of 2009](https://en.wikipedia.org/wiki/ADA_Amendments_Act_of_2008), and [Section 504 of the Rehabilitation Act](https://en.wikipedia.org/wiki/Section_504_of_the_Rehabilitation_Act) who request and are eligible for accommodations. The Office of Student Disability Services (SDS) is the UMBC department designated to coordinate reasonable accommodations that would allow students to have equal access and inclusion in all courses, programs, and activities of the University. If you have a documented disability and need to request academic accommodations, please register with the Office of Student Disability Services (SDS) as soon as possible. To begin the registration process please visit the SDS website and review the registration information, including disability documentation

guidelines and how to submit the SDS registration form online using the confidential data management software called [Accommodate](https://sds.umbc.edu/accommodations/registering-with-sds/).

Once accommodations have been approved, you and your instructors will be notified via an emailed accommodation letter from the SDS office. Both the SDS office and Shady Grove’s [Center for Academic Success(CAS)](https://shadygrove.umd.edu/student-services/center-for-academic-success) will work with you to ensure you receive the approved accommodations. If you have any questions or concerns, please contact the [Office of Student Disability Services SDS](https://sds.umbc.edu/) via [disAbility@umbc.edu](mailto:disAbility@umbc.edu) or phone at 410-455-2459. Please note that accommodations are not retroactive and begin once SDS sends an approved accommodation letter. For more information on the services CAS provides, please contact Mary Gallagher (maryg@umd.edu) or visit [Student-Services/Center](https://shadygrove.umd.edu/student-services/center-for-academic-success) [for Academic Success](https://shadygrove.umd.edu/student-services/center-for-academic-success)

## Title IX Statement

Any student who has experienced sexual harassment or assault, relationship violence, and/or stalking is encouraged to seek support and resources. There are a number of resources available to you. Please see [The Office of Equity and](https://oei.umbc.edu/sample-title-ix-responsible-employee-syllabus-language/) [Inclusion Website](https://oei.umbc.edu/sample-title-ix-responsible-employee-syllabus-language/) for recently updated UMBC Policies and Resources during COVID-19.

With that said, as an instructor, I am considered a Responsible Employee, as per [UMBC’s Interim Policy on](https://oei.umbc.edu/umbcs-policy-on-prohibited-sexual-misconduct-interpersonal-violence-and-other-related-misconduct/) [Prohibited Sexual Misconduct, Interpersonal Violence, and Other Related Misconduct](https://oei.umbc.edu/umbcs-policy-on-prohibited-sexual-misconduct-interpersonal-violence-and-other-related-misconduct/). This means that while I am here to listen and support you, I am required to report disclosures of sexual assault, domestic violence, relationship violence, stalking, and/or gender-based harassment to the University’s Title IX Coordinator. The purpose of these requirements is for the University to inform you of options, supports, and resources.

You can utilize support and resources even if you do not want to take any further action. You will not be forced to file a police report, but please be aware, depending on the nature of the offense, the University may take action.

* + [The Counseling Center](https://counseling.umbc.edu/). phone: 410-455-27742 (M-F 8:30am - 5pm)
  + [University Health Services](https://uhs.umbc.edu/): 410-455-2542 (M-F 8:30am - 5pm)
  + For after-hours emergency consultation, call [the police](https://police.umbc.edu/) at 410-455-5555 Other on-campus supports and resources:
  + [The Women’s Center](https://womenscenter.umbc.edu/)(available to students of all genders): 410-455-2714 (M-Th 9:30am -6pm, F 9:30am - 4pm)
  + [Title IX Coordinator](https://oei.umbc.edu/sexual-misconduct-policy-and-procedures/): 410-455-1606 (M-F 9am - 5pm)