DATA 601: Introduction to Data Science

Murat Guner

Fall 21

E-mail: mguner@umbc.edu
Office Hours: **With Appointment**
Office: BSE IV 3121

Web: **Course's Github Page**
Class Hours: T/Th 3-4:15pm
Class Room: ...

- I return messages within 48 hours, excluding breaks and weekends.

- **<u>Covid-19:</u>** Due to the global pandemic, our classes will be held remotely this semester. Please see **this Google doc** this Google doc for UMBC Policies and Resources during COVID-19.

- If your question is something about your code or something we covered in class then please don't send me email instead schedule an appointment.

- I am available after class for questions and help. Outside of that timeframe, please schedule an appointment.

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, web scraping; life cycle of data science projects and different roles in a data science team, and some ethical issues in data science.

Required Materials

Textbook

- Course notes available on Github.

- Weekly reading materials will be shared via Blackboard system.

- **R for data science**

- **Python Data Science Handbook**

Hardware Requirements

- Students should have a working computer with working camera and microphone in case needed.

Prerequisites/Core requirements

Core requirements: Basic familiarity with Statistics, Basic familiarity with Python programming.

Course Objectives

Successful students will be able to:

1. Describe the key activities in a data science project.

2. Use commonly used Python packages for data visualization.

3. Use Pandas library for data transformations.

4. Evaluate the quality of a dataset for a given data science project.

5. Create functions and programs that are cleaning and transforming raw data sets.

6. Load, read and write from different data sources.

7. Apply basic statistical knowledge in a data science project to test and verify hypotheses.

8. Create new data sets by combining data from different sources.

9. Understand the role of modeling in a data science project.

10. Apply best practices of communication for reporting when they finish a data science project.

Course Structure

Class Structure

1. Students will complete assigned homework, readings, quizzes.

2. This course incorporates a variety of hands-on labs and practical exercises to engage students and prepare them for challenges they may encounter in the workplace.

3. Students will occasionally present their solutions to homework assignments in class.

Expectations From Students

.... Here write expectations from students

Assignments

.... Here Assignment policies

Grading Policy

.... Here Grading Policy

Course Policies

During Class

I understand that the electronic recording of notes will be important for class and so computers will be allowed in class. Please refrain from using computers for anything but activities related to the class. Phones are prohibited as they are rarely useful for anything in the course. Eating and drinking are allowed in class but please refrain from it affecting the course. Try not to eat your lunch in class as the classes are typically active.

Attendance Policy

Attendance Policy

Policies on Incomplete Grades and Late Assignments

Incomplete grade policy
Late Assignment Policies

Academic Integrity and Honesty

UMBC Policies for Academic Integrity

Equity and Inclusion

UMBC Policies for Accommodation
UMBC Policies against Discrimination

Schedule and weekly learning goals

The schedule is tentative and subject to change.

Week 01, 09/13 - 09/17: 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.
- Coding Basics: Naming conventions, objects and their methods, built-in functions etc.

Week 02, 09/20 - 09/24: Basics of visualization with Python

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

Week 03, 09/27 - 10/01: Data transformations

- Installing and importing 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 04, 10/04 - 10/08: Data Transformations - Part-II

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

Week 05, 10/11 - 10/15: Exploratory data analysis

- Variation
- Typical values
- Unusual values
- Heatmaps
- Patterns and models

Week 06, 10/18 - 10/22: Loading and wrangling data

- Introductions to different data formats
- Reading csv, excel, html files
- Handling numbers
- Handling strings
- Working with dates and times
- Tidying data

Week 07, 10/25 - 10/29: Relational data

- Merging, concatenation, joining datasets.

Week 08, 11/01 - 11/05: Working with strings

- String operations
- Exact matches
- fuzzy matches

Week 09, 11/08 - 11/12: Working with date and time

- Timezone handling
- Moving Averages
- Resampling methods for time series
- Time series visualizations

Week 10, 11/15 - 11/19: Basics of object oriented programming

- Introduction to classes and objects
- Global variables
- Inheritance

Week 11, 11/22 - 11/26: Web Scraping and API

- Basic HTML
- Best Practices of Web Scraping
- Use and Benefits of APIs

Week 12, 11/29 - 12/03: Modeling

- Place of modeling in a data science project
- Linear Regression
- Visualizing Models

Week 13, 12/06 - 12/10: Report and Commminication

- Structure of a Data Analysis Write-up
- Project Structure and Deliverables
- Reproducibility
- Project Structure and Deliverables

Week 14, 12/13 - 12/17: Misc

- ...