# DATS 6103 – Introduction to Data Mining
# CRN 56030
# Tue 7:10 PM – 9:40 PM
# January 17, 2017 – May 1, 2017

## INSTRUCTOR

Name: Nima Zahadat, Ph.D.

Term: Spring 2017

Campus Address: GWU Media and Public Affairs (MPA) 305

E-mail: nzahadat@gwu.edu

Office hours: By appointment

## RESOURCES

A Programmer's Guide to Data Mining, The Ancient Art of the Numerati.

Han and Kamber , Data Mining: Concepts and Techniques, Third Edition.

Charu C. Aggarwal, Data Mining: The Textbook,
http://www.charuaggarwal.net/Data-Mining.htm

The Elements of Statistical Learning, 2nd edition by Hastie, Tibshirani, and Friedman (http://www-stat.stanford.edu/~tibs/ElemStatLearn)

## COURSE DESCRIPTION

This course is a survey of concepts, principles, and techniques in data mining, including classification, association, and cluster analyses. Students learn to apply data mining methods to real-world problems with minimal rigorous mathematical understanding of the underpinnings of the methods. The course helps build a good foundation for taking advanced courses in the data science curriculum and for applying the basic techniques to practical problems. Data based examples and exercises using R, Python, and other tools are integrated into class activities.

## OBJECTIVES

Be able to

1. Develop code in Python

2. Explain and use the mining process for descriptive and predictive analytics
3. Explore data using various mining and visualization techniques
4. Understand and apply the core data mining methods of classification, association and analysis

## TOPICS

**Data Mining and Analysis Overview:** the nature of data, Python, and other tools; graphical display of data; classification and estimation; parametric vs. nonparametric methods; generalization, over and under fitting, and cross-validation.

**Basic Concepts of Pattern Recognition:** basic statistical descriptions of data; data visualization; measuring similarity and dissimilarity; parametric density estimation; maximum likelihood; generative classification; regression analysis; optimization; nonparametric estimation; histogram; kernel density estimation; bias-variance tradeoff.

**Data Processing**: data quality, cleaning, noise; data reduction, integration; basics of data warehousing, and online analytical processing; types of data (structured, non-vector, compressed); outliers and robustness; corrupted, noisy, expensive, and heterogeneous data.

**Current and Emerging Application Areas:** Digital Humanities and data mining; Information Extraction from text corpora

## COURSE PREREQUISITES

You are expected to have a basic knowledge of statistics (e.g., at the level of STAT 2118 Regression Methods that covers analysis of research data through simple and multiple regression and correlation). The course prerequisite is DATS 6101 or permission of the instructor.

## ASSESSMENTS

There will be two individual projects.

There will be weekly assignments.

## GRADING

Grade scale is as follows:

| | |
|---|---|
| 97 – 100% | A+ |

| | |
|---|---|
| 93 – 96% | A |
| 90 – 92% | A- |
| 87 – 89% | B+ |
| 83 – 86% | B |
| 80 – 82% | B- |
| 77 – 79% | C+ |
| 73 – 76% | C |
| 70 – 72% | C- |
| 60 – 69% | D |
| < 60% | F |

Your projects are graded based on the following rubrics:

-Followed directions (5)
-Ease of use and ease of understanding of the programs and dataset (5)
-Presentation/Video design, quality, professionalism (5)
-Originality of topic, concepts, ideas (5)
-Creativity in design, organization, presentation (10)
-Working programs and tools (10)
-Research level, validity of research sources (5)
-Technical knowledge (5)
-Files properly packaged and named (5)
-Professionalism in organization and presentation (5)
-Usefulness of the dataset and associated tools and graphics (5)
-Clarity of concepts, analyses, video (5)
-Focus; is the project and its analyses focused and easy to follow/understand? (5)
-Difficulty level of the project in terms of research, technology, preparation (5)
-Organization of presentation, material, concepts, analyses, conclusions (5)
-Error free (5)
-Presented within time (5)
-Analysis including observed patterns, code analysis, conclusions, predictions (5)

**Note that a project that doesn't work or is fairly incomplete will not receive partial credit and will be given a grade of 0. The rubrics apply only to working projects.**

## INDIVIDUAL PROJECTS

The individual projects will constitute the following:

1. The first project will be a presentation (PowerPoint and hands-on) on the military spending of the top 10 or more countries in the world

2. The second project will be a working data mining project on a topic of your choice that is approved by the instructor.

## ASSIGNMENT SUBMISSION

-Assignments are due on time; see due dates on Blackboard
-Be sure to put all your files into a **folder** and name the folder like this

"DATS 6103 - Individual Project # - First Last"
**as in**
"DATS 6103 - Individual Project 1 - Bugs Bunny"
**and for the Final Project**
"DATS 6103 - Final Project - Bugs Bunny"
if your name is Bugs Bunny; otherwise use your own name

-Pay attention to the spacing and capitalization; do not add your own formatting
-Zip the folder and name it using the same formatting
-Unless your name is Bugs Bunny, use your own name
-Upload your file to Blackboard

## READING ASSIGNMENTS

You are required to read roughly one chapter in your book every week.

## EMAIL ETIQUETTE

In the age of technology, when most forms of communication are electronic, it is important to adopt a proper etiquette to communicate with one another. It is asked that students use salutation when sending emails to their instructors and also make sure to SIGN their name and include their class/section at the end of the email. The instructor reserves the right NOT to reply to emails that are not properly addressed or do not have a signature. Students should also use their GWU email for any correspondence with the instructors. Students are required to check their emails daily and especially the morning before class.

## ACADEMIC INTEGRITY

Students are responsible for understanding the George Washington University’s Honor Code's provisions. In the spirit of the code, a student's word is a declaration of good faith acceptable as truth in all academic matters. Cheating and attempted cheating, plagiarism, lying, and stealing of academic work and related materials constitute Honor Code violations. These will not be tolerated. The code states: “Academic dishonesty is defined as cheating of any kind, including misrepresenting one's own work, taking credit for the work of others without crediting them and without appropriate authorization, and the fabrication of information.” For the remainder of the code, see:
http://www.gwu.edu/~ntegrity/code.html

## SUPPORT FOR STUDENTS OUTSIDE THE CLASSROOM

*DISABILITY SUPPORT SERVICES (DSS)*
Any student who may need an accommodation based on the potential impact of a disability should contact the Disability Support Services office at 202-994-8250 in the Marvin Center, Suite 242, to establish eligibility and to coordinate reasonable accommodations. For additional information please refer to: http://gwired.gwu.edu/dss/

*UNIVERSITY COUNSELING CENTER (UCC)* ***202-994-5300***
The University Counseling Center (UCC) offers 24/7 assistance and referral to address students' personal, social, career, and study skills problems. Services for students include:

- crisis and emergency mental health consultations
- confidential assessment, counseling services (individual and small group), and referrals
  **http://gwired.gwu.edu/counsel/CounselingServices/AcademicSupportServices**

## PROPOSED SCHEDULE (SUBJECT TO CHANGE)

| | |
|---|---|
| Week 01: | Going over the syllabus |
| 01/17 – 01/20 | Class expectations |
| | Getting environments ready: Anaconda setup |
| | Learning Python |
| Week 02: | Learning Python |
| 01/23 – 01/27 | Procedural aspects |
| Week 03: | Learning Python |
| 01/29 – 02/03 | OOP aspects |
| Week 04: | Working with datasets in Python |
| 02/06 – 02/10 | Data analysis with Python and Pandas |
| | Pandas basics |
| | IO basics |
| Week 05: | Building datasets |
| 02/13 – 02/17 | Concatenating and appending data-frames |
| | Joining and merging data-frames |
| Week 06: | Percent change and correlation tables |
| 02/21 – 02/24 | Handling missing data |
| | Matplotlib intro and line |
| | Legends, titles, and labels |
| Week 07: | Matplotlib bar charts and histograms |
| 02/27 – 03/03 | Scatter plots |

| | |
|---|---|
| | Stack plots |
| | Pie charts |
| | Loading data from files |
| Week 08: | Baby names dataset |
| 03/06 – 03/10 | |
| Week 09: | No class (spring break) |
| 03/13 – 03/17 | |
| Week 10: | First project due |
| 03/20 – 03/24 | Work on project |
| Week 11: | Getting data online |
| 03/27 – 03/31 | Converting data |
| | Basic customization |
| | More customization of colors and fills |
| Week 12: | FBI homicide dataset |
| 04/03 – 04/07 | |
| Week 13: | Reading website data |
| 04/10 – 04/14 | Creating datasets from websites |
| Week 14: | Facts of life presentation |
| 04/17 – 04/21 | Work on second project |
| Week 15: | Second project due |
| 04/24 – 04/28 | |

## SECURITY

In the case of an emergency, if at all possible, the class should shelter in place. If the building that the class is in is affected, follow the evacuation procedures for the building. After evacuation, seek shelter at a predetermined rendezvous location.