**COMP 5355 – DATA ANALYSIS**

**Course Syllabus**

(January 2019)

Semester 2019 Spring

Class: [**HOLD**] Room: 839

**[Denis Akhiyarov]**

Office: remote

Phone: 9182003994

Office Hours: by appointment

[[email] dakhiyarov@na.edu](mailto:[email]%20dakhiyarov@na.edu)

**Prerequisite(s)**

(None) Programming knowledge and statistical knowledge expected

**Course Description**

This course investigates data scraping/sampling/cleaning in order to get an informative, manageable data set; data storage and management in order to be able to access data - especially big data - quickly and reliably during - subsequent analysis; exploratory data analysis to generate hypotheses and intuition about the data; prediction based on statistical tools such as regression, classification, and clustering and communication of results through visualization, and interpretable summaries.

**Instructional Hours/Credits**

Lecture - 45 Clock Hours / 3 Semester Credits

**Textbooks, Instructional Materials and References**

[Python Data Science Handbook](http://shop.oreilly.com/product/0636920034919.do) by Jake VanderPlas:

<https://jakevdp.github.io/PythonDataScienceHandbook/>

Hands-On Machine Learning with Scikit-Learn and TensorFlow

<https://github.com/ageron/handson-ml>

Python Machine Learning (2nd edition)

<https://github.com/rasbt/python-machine-learning-book-2nd-edition>

**Relation to Program Learning Outcomes (PLOs)**

|  |  |
| --- | --- |
| **MS in CS - Program Learning Outcomes (PLOs)** | **Assessment** |
| 1. Demonstrate advanced skills of computing theory and algorithms. | **U** |
| 1. Develop sophisticated knowledge of operating systems and hardware. |  |
| 1. Apply advanced practices of software design and development cycle. | **U** |
| 1. Analyze data using advanced computing tools. | **U** |
| 1. Develop state-of-the-art skills of computer networks. | **U** |
| 1. Engage and apply secure practices in various digital environments. | **E** |
| **Key: I=Introduced; E=Emphasized; U=Utilized; A=Comprehensive Assessment** | |

**Learning Objectives/Outcomes**

Students will be able to:

* Use Python and other tools to scrape, clean, and process data **[PLO 1,3 & 4]**
* Use data management techniques to store data locally and on cloud **[PLO 4 & 5]**
* Use statistical methods and visualization to quickly explore data **[PLO 4]**
* Apply statistics and computational analysis to make predictions based on data **[PLO 1,4]**
* Apply basic computer science concepts such as modularity, abstraction, and encapsulation to data analysis problems **[PLO 3 & 4]**
* Implement data-intensive computations on cluster and cloud infrastructures using distributed computing **[PLO 4 & 6]**
* Effectively communicate the outcome of data analysis using descriptive statistics and visualizations **[PLO 4]**

**Course Outline**

* Data Types
* Visualization of Multi-Dimensional Data
* Distance.
* Clustering and Dimensionality Reduction
* Data Scraping
* Statistical Modeling
* Regression and Classification
* Machine Learning
* Decision trees, ensemble learning
* Support vector machine
* Clustering methods
* Distributed computing
* Graph Visualization

**Homework Expectation**

Students are expected to spend approximately six (6) hours a week, on average, completing homework assignments in order to achieve the learning objectives for this 15 week lecture course. This meets the Federal Government’s expectation of two hours of homework for each hour of lecture.

**Academic Honesty:**

Each student assumes the responsibilities of being a member of the NAC academic community. All acts of plagiarism are not tolerated including: cheating, claiming one’s work as their own, fabrication and helping one to commit any of these acts. Any violations of academic honesty will receive strict disciplinary action, which can include suspension and even expulsion from NAC.

**Instructional Methods**

**Lectures**

The class meets twice a week for lectures and joint class activities. The class activities are designed to help you master the relevant materials, to work on your homework in groups, and to get you started on your project. The weekly schedule of lectures is posted on the course web site.

**Project**

Towards the middle of the course you will start to work on a data science project. The goal of the project is to go through the complete data science process to answer questions you have about some topic of your own choosing. You will acquire the data, design your visualizations, run statistical analysis, and communicate the results. Part of this project will be assessed midway through the course and then the rest of it will be assessed at the end of the course.

You will work closely with other classmates in a 3-4 person project team. You can come up with your own teams. If you can’t find a partner we will team you up randomly. We recognize that individual schedules, different time zones, preferences, and other constraints might limit your ability to work in a team. If this the case, ask us for permission to work alone.

**Homework**

The homework is going to provide an opportunity to learn data science skills and to test your understanding of the material. See the homework as an opportunity to learn, and not to “earn points”. The homework will also be graded to reflect this objective.

**Reading Assignments**

The course schedule includes required readings. The goal of the reading assignments is to prepare for class, to familiarize yourself with new terminology and definitions, and to determine which part of the subject needs more attention. The homework assignments may contain questions about the mandatory readings. When answering those please be brief and to the point!

**Library**

Since this course assignments and project requires research you are expected to know how to use the library's resources including the available databases, periodicals, and journals. If you are not familiar with using the library, please ask for assistance from the library's personnel, take workshops provided by the library, or visit the library's website at<http://na.edu/academics/library.html>

**Accommodations:**

Students that require any accommodation (such are students with disabilities, religious conflicts, etc…) should notify the instructor as early as possible and accommodations will be made on an individual basis in adherence with the regulations outlined in the Student Handbook.[can be modified, optional]

**Assessment Criteria and Methods of Evaluating Students**

Quizzes 30% (include rubric for this one if course is online)

Project 40%

Homework 30%

|  |  |
| --- | --- |
| A | 96~100 |
| A- | 91~95 |
| B+ | 86~90 |
| B | 81~85 |
| B- | 76~80 |
| C+ | 71~75 |
| C | 66~70 |
| C- | 61~65 |
| D+ | 56~60 |
| D | 50~55 |
| F | Below 50 |