Data Science

**Course:** CSC 405/605 – Data Science

**Time:** Tuesday and Thursday, 8 am – 9:10 am

**Instructor:** Dr. Qianqian Tong

**Location:** Petty

**Office Hours:** Tuesday and Thursday 9:10 am – 10:00 am @ Petty 152

**Email:** [q\_tong@uncg.edu](mailto:q_tong@uncg.edu)

Course Description

In a world with ever increasing data generated both by humans and machines alike, the field of

computer science has seen a transition from computation-intensive solutions to data-intensive

ones. Often in such a scenario, solutions to real-world problems can be derived/learned by

analyzing disparate, complex, and messy datasets using Data Science methods and approaches.

This course is highly interactive, and will explore the theories, techniques, and the tools

necessary to gain insights from such datasets. Using a problem-based learning philosophy,

students are expected to make use of such technologies to design data solutions that can process and analyze real-world datasets for a variety of scientific, social, and environmental challenges.

The core topics addressed by the course will be:

• Programming with Data

• Data Mining, Munging, Wrangling

• Statistics, Analytics, Representation, Visualization

• Introduction to Applied Machine-Learning

Prerequisites

A grade of B+ or better in [CSC 330](https://catalog.uncg.edu/search/?P=CSC%20330) and ([STA 271](https://catalog.uncg.edu/search/?P=STA%20271) or [STA 290](https://catalog.uncg.edu/search/?P=STA%20290)), or permission of instructor (prior programming and statistics experience is required).

Textbooks

There is no required text for the course. Class slides will be available for download. Suggested

textbooks are: 1) Building Machine Learning Systems with Python (Richert and Coelho), 2) Data

Science from Scratch (Joel Grus)

Course Overview

This course is highly interactive and based on the problem-based learning philosophy;

Students are expected to make use of said technologies to design highly scalable systems

that can process and analyze real-world datasets for a variety of scientific, social, and environmental challenges.

Course Topics and Schedule (Tentative)

1. **Introduction to Data Science: (Week 1-3)**

o Data Science Introduction

o Class Project discussion

o Programming prepare

1). Re/Introduction to Python

2). IPython, IPython-Notebook

o Data Science Reproducibility

1). Setting up your Repository – Data, Code, and Documentation

2). Using Version Control with Git

o Final Project Discussions - Goals and Requirements

2. **Data Munging, Wrangling, Cleaning (Week 4-5)**

o Data Structures

o Data Manipulation

1). Selection - Indexing

2). Handling Missing Data

3). Aggregation

4). Descriptive Statistics

5). Merging / Join

6). Working with Date-Time

o Assignment 1 submission

o Project Review - Stage I

3. **Data and Statistics (Week 6-9)**

o Distributions

o Estimates

o Statistical Hypothesis Testing

o Correlation

o Distribution Estimators: MoM, MLE, KDE

o Project Review - Stage II

4. **Introduction to Applied Data Modeling: (Weeks 10-12)**

o Applied Machine Learning

o Mathematical optimization (if time allowed)

o Stochastic thinking (if time allowed)

o Regression and Feature Selection

o Bias versus Variance

o Clustering and Dimensionality Reduction

o Validation and Model Performance

o Assignment 2 submission

o Project Review - Stage III

5. **Data Visualization (Week 13-14)**

o Graph Generation

1). Types of Graphs

2). Customizing Plots

3). Visualizing Errors

4). Interactive / Dynamic Graphs

o Visualization Best Practices

o Project Review - Stage IV

6**. Project Presentations: (Week 15–16)**

o Assignment 3 submission

o Project Review - Stage V

o Graduate Students report submission

Grading Policy

Grade Max% to Min%

A 100% to 94%

A- < 94% to 90%

B+ < 90% to 87%

B < 87% to 84%

B- < 84% to 80%

C+ < 80% to 77%

C < 77% to 74%

C- < 74% to 70%

D+ < 70% to 67%

D < 67% to 64%

D- < 64% to 60%

F < 60% to 55%

1. Class Participation: 10%

Attendance is required for all the class. Students must let the instructor know the reason for missing in person class, else will lose credit for absence from class. Email should be sent one-day before class begin.

2. Assignments (3): 30%

Three programming-based assignments will be given covering the utilization of the tools learned in class. Each assignment accounts for 10 points. Absolutely no collaboration on assignments. Students must upload (Notebooks) individual assignments to canvas before deadline. Later submission (within one week) will have a 20% deduction.

3. Project: Each stage project 50% + Final Presentation 10%

The final project of the class will focus on the end-to-end development of an analytical

model. The project will be split into the following stages:

o Stage I. Data/Project Understanding,

o Stage II. Data Modeling,

o Stage III. Distributions and Hypothesis Testing,

o Stage IV. Basic Machine Learning,

o Stage V. Visualization and Dashboard.

This will be a team-based effort, where in the first week of the course the students split into teams of 3-5 students. After completing each stage, the teams will have to give a short presentation (3-5 mins) and a report (1 page) of their progress with the project. The projects will be open-source and the teams will have to use GitHub as their code repository. Upon completion of the project the teams will present their software along with the results in form of a presentation (15 minutes).

Each Stage of the Final Project has 100 points. They will be equally weighted for the project final score. Each stage consists of: 1). Report; 2). Code Jupyter/IPython Notebooks; 3). Presentation. To get the full points in each stage you need to finish all of the deliverables.

**Graduate Students Only**: Stage V has 50 points for your project and 50 points for project report (IEEE format). Minimum 5 pages for single author, 8 for 2 authors, and 12 for 3 authors (figures and references included). The due date should before the final week.

Academic Honesty Policy

The instructor will deal strictly with any violations of academic honesty and integrity in this course. ***Absolutely no discussion, collaboration, copying, and sharing on assignments. This includes coping from the internet. Any student who violates this policy will receive “F” directly in the course. The instructor will report the case to the university.***

Special Needs and/or Disabilities

Students with disabilities should have documentation from the Office of Accessibility Resources & Services. This documentation should be provided to the instructor for review. In the case of major provisions such as separate testing environment or test-readers, the student must make arrangements with Office of Accessibility Resources & Services so that suitable accommodations can be provided.

COVID Statement

As we return for spring 2023, please uphold UNCG’s culture of care to limit the spread of covid-19 and other airborne illnesses. These actions include, but are not limited to:

* Engaging in proper hand-washing hygiene
* Self-monitoring for symptoms of covid-19
* Staying home when ill
* Complying with directions from health care providers or public health officials to isolate if ill
* Completing a [self-report](https://veoci.com/v/p/132667/workflow/fs2x25pzqnd5) when experiencing covid-19 symptoms or testing positive for covid-19
* Following the CDC's [exposure guidelines](https://www.cdc.gov/coronavirus/2019-ncov/your-health/if-you-were-exposed.html) when exposed to someone who has tested positive for covid-19
* Staying informed about the University's policies and announcements via the [covid-19](https://covid.uncg.edu/) website