

CSC 405/605/705: Data Science

Dr. Xiaochen Li

About Me

- Assistant Professor, Department of Computer Science, University of North Carolina at Greensboro
- Former Visiting Ph.D. Student and Postdoctoral Researcher, University of Virginia
- Ph.D. in Cyberspace Security, Zhejiang University (China)
- Research focus: **Data Privacy in Cyberspace Security**, with applications to **statistical analysis** and **large language models (LLMs)**



Email: x_li12@uncg.edu

Office: Petty 153

Office Hour: Monday Wednesday 10:00 am – 11:00 am



What is the course about?

- Programming your way into Data Science
- It is not a Statistics or an AI or a Visualization course
- The course contains parts of everything
- Learn about lot of tools and how to use them in innovative ways
 - We added a topic that learn how to use modern AI tools to efficiently build and communicate data science projects!
- We will work with real-world data
- Hopefully develop some cool projects

Course Information

You can find all course-related information on this course website, including lecture slides, assignments, and other resources. You can also find this link on Canvas.

<https://xiaochenli-w.github.io/CSC-405-605-705-Data-Science-Fall-2026/>

Course Information

- **Prerequisites**

A grade of C or better in CSC 330 and (STA271 or STA290), or permission of instructor (prior programming and statistics experience is required).

- **Experience in:**

- Programming skills Python

We will go through Introduction to Python

You would have to work hard in the early weeks to get comfortable with Python

- Terminal, Command-Line, Installing packages and configuring a development environment, Basic debugging skills

Course Information

- **Textbooks**

There is no required textbook 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 Topics

1. Introduction to Data Science (2 lectures)

- What is Data Science? The Data Science workflow
- Introduction to essential development tools:
- Jupyter Notebook
- Git and GitHub

2. Python Programming for Data Science (7 lectures)

- Python programming fundamentals
- Data manipulation with Pandas
 - Data loading and inspection
 - Data selection and indexing
 - Descriptive statistics
 - Data cleaning:
 - Handling missing values
 - Removing duplicates
 - Detecting and treating outliers
 - Correcting inconsistent data formats
- Numerical computing with NumPy

Course Topics

3. Statistics for Data Science (6 lectures)

- Descriptive statistical measures
- Probability distributions
- Distribution estimation methods
 - Method of Moments (MoM)
 - Maximum Likelihood Estimation (MLE)
 - Kernel Density Estimation (KDE)
- Correlation analysis
- Statistical hypothesis testing

4. Data Visualization (3 lectures)

- Principles of effective data visualization
- Creating visualizations with Python
 - Common chart types
 - Plot customization
 - Interactive and dynamic visualizations

Course Topics

5. Introduction to Applied Data Modeling (2 lectures)

- Decision Trees
- Regression models
- Model validation and performance evaluation

6. AI-Assisted Data Analysis (3 lectures)

- Introduction to AI-assisted Data Science
- Building interactive data applications with Streamlit
- Integrating LLMs using LangChain
- Building an end-to-end data science project with LLMs

7. Final Project Presentations (2 lectures)

- Student project presentations

Course Information – Grading Policy

Grading Policy

Grade	Score Max %	Score Min %
A	100%	92%
A-	< 92%	89%
B+	< 89%	86%
B	< 86%	83%
B-	< 83%	80%
C+	< 80%	77%
C	< 77%	74%
C-	< 74%	70%
D+	< 70 %	67%
D	< 67%	64%
D-	< 64 %	60 %
F	< 60%	0 %



Course Information – Grading Policy

For undergraduates:

Category	
Assignments (2)	26% (Each accounts for 13%)
In-class quizzes (3)	24% (Each accounts for 8%)
Project	50%

For graduates:

Category	
Assignments (2)	26% (Each accounts for 13%)
In-class quizzes (3)	24% (Each accounts for 8%)
Project	40%
Research Report	10%

Grading Policy

- [In class quizzes (3): 24%]

Three in-class quizzes will be conducted during the semester, each contributing 8% to the final grade. **Quizzes must be completed in class as scheduled; failure to attend or complete a quiz during class will result in no credit.**

Quiz dates are provided at the beginning of the semester; therefore, personal travel or other foreseeable commitments are not acceptable excuses. Make-up quizzes will be granted only in documented cases of serious emergencies or other exceptional circumstances, and only with prior approval whenever possible.

Course Information

- **[Assignments (2): 26%]**

Two programming-based assignments will be given covering the utilization of the tools learned in class. Each assignment accounts for 13 points. **Absolutely no collaboration on assignments.** Students must upload individual assignments to canvas before deadline.

Assignments and the course project are due by 11:59 PM on the stated due date. Submissions received within 7 calendar days after the deadline will incur a 25% late penalty. Work submitted more than 7 calendar days late will not be accepted.

Course Information

- **[Project: 50% (40% for Graduate Students)]**

Each student will complete an individual AI-assisted data science project that demonstrates the complete data science workflow using the concepts and techniques learned throughout the semester. Students are expected to independently design, implement, and present a complete data analysis project using Streamlit as an interactive reporting platform.

Individual Project

➤ **Each project should include the following stages:**

- Data acquisition
- Data cleaning and preprocessing
- Exploratory Data Analysis (EDA)
- Statistical analysis
- Machine learning modeling
- Model evaluation
- Data visualization
- AI-assisted interpretation and discussion
- Final interactive report

Individual Project

➤ Each student is responsible for:

- Selecting one real-world dataset from a publicly available repository (e.g., Kaggle, UCI Machine Learning Repository, Data.gov, etc.).
- Defining at least one data-driven research question or business problem.
- Choosing appropriate analytical methods and machine learning models.
- Submit a project proposal during the semester.
- Developing a Streamlit application that presents the complete analysis.
- Maintaining an individual GitHub repository containing all source code.
- Submitting a PDF version of the final Streamlit report together with the GitHub repository link.

Publicly Available Repository

Academic Datasets

- UC Irvine Machine Learning Repository (<http://archive.ics.uci.edu/ml/>)
- Stanford Large Network Dataset Collection (<http://snap.stanford.edu/data/>)
- Inter-university Consortium for Political and Social Research (<http://www.icpsr.umich.edu/>)
- Pittsburgh Science of Learning Center's DataShop (<https://pslcdatashop.web.cmu.edu/>)
- Academic Torrents (<http://academictorrents.com/>)

Private Companies

- Data.World (<https://data.world/>)
- Quandl Financial Data (<https://www.quandl.com/>)
- Amazon Web Services Public Data Sets (<http://aws.amazon.com/datasets/>)
- Kaggle (<http://www.kaggle.com/>)
- Nytimes (<http://developer.nytimes.com/docs>)

Publicly Available Repository

Gov. and NGO's

- Data.gov (<https://www.data.gov/>)
- NYC Open Data (<https://nycopendata.socrata.com/>)
- DC Open Data Catalog (<http://data.dc.gov/>)
- OpenDataDC (<http://www.opendatadc.org/>)
- DataLA (<https://data.lacity.org/>)
- Project Open Data Dashboard (<http://data.civicagency.org/>)
- data.gov.uk (<http://data.gov.uk/>)
- US Census Bureau (<http://www.census.gov/>)
- World Bank Open Data (<http://data.worldbank.org/>)
- Humanitarian Data Exchange (<http://docs.hdx.rwlabs.org/>)
- Sunlight Foundation (<http://sunlightfoundation.com/api/>)
- ProPublica Data Store (<https://projects.propublica.org/data-store/>)

Publicly Available Repository

Other resources

- 20 Big Data Sources (<http://www.smartdatacollective.com/bernardmarr/235366/big-data-20-free-big-data-sources-everyone-should-know>)
- Center for Data Innovation (<http://www.datainnovation.org/category/publications/dataset-blog/>)
- Data Science Central (<http://www.datasciencecentral.com/>)
- Python API's (<http://www.pythonforbeginners.com/api/listof-python-apis>)
- PyCoders Weekly (<http://pycoders.com/>)

Individual Project

➤ Important Project Dates:

Date	Milestone
October 30	Project Proposal Due
November 9	Instructor Feedback on Project Proposal Returned
November 30	Final Project Report Submission Due
November 30	Student Project Presentations (Session I)
December 2	Student Project Presentations (Session II)

Individual Project

➤ Project Grading Policy:

Task	100 points total (worth 50% for undergraduate and 40% for graduate in final scores)
Project Proposal	20 points
GitHub repository	30 points (Source code 20 points, Version control 10 points)
Final Project Report	50 points
Presentation	Bonus 0~10 points

Individual Project

➤ Project Presentation (Bonus Opportunity):

During the final two class meetings (Nov 30 & Dec 2), 10 students will present their projects (about 5 presentations per class, 15-20 minutes each).

Presentation opportunities are intended **as a bonus rather than a course requirement**.

- Students may volunteer to present.
- If more than 10 students volunteer, presenters will be selected by random drawing.
- If fewer than 10 students volunteer, the remaining presentation slots will be filled by randomly selecting additional students from the class.
- Students who present may earn up to 10 bonus points, based on the quality of the presentation.

Individual Project

➤ Project Proposal (20 Points):

- The project proposal is intended to help students receive early feedback and guidance before beginning the full project.
- The proposal should demonstrate that the project is feasible, well planned, and aligned with the course objectives.
- Students should download and complete the Project Proposal Template provided and submit it by the posted deadline.
- There is no page requirement; however, the proposal should address all sections of the template with sufficient detail.

Individual Project

- **Reproducible AI-Assisted Workflow (30 points):**
 - An individual GitHub repository (must be publicly accessible during grading)
 - The README must include: Project title, Dataset source (URL), Project objective, Instructions for running the code
 - Version control (GitHub commit history) will be considered as part of the evaluation.

Individual Project

➤ Streamlit Report Components (30 points):

- The report must present the complete data science workflow.
- **Dataset overview**
- **Exploratory analysis**
You **must** demonstrate:
 - Identification and handling of missing values
 - Detection of outliers or anomalies
 - Use of:
 - Group-by / aggregation
 - Descriptive statistics
- Simply dropping data without justification is not acceptable.
- **Statistical Analysis & Hypothesis Testing**
You must perform at least one formal statistical analysis (e.g., t-test, ANOVA, Chi-square test, correlation analysis)
For each test: State the null hypothesis; Explain why the test is appropriate; report test statistic and p-value; interpret the result in context.
- **Machine learning results**
You must implement:
 - One baseline machine learning model (e.g., Logistic Regression, Decision Tree, Linear Regression)
 - One more advanced model (e.g., Random Forest, XGBoost, LightGBM, Neural Networks)

Individual Project

➤ Streamlit Report Components (30 points):

- The report must present the complete data science workflow.

- **Visualizations**

You must present multiple informative figures:

- Appropriate chart selection
- Clear labels and captions
- Interpretation of every figure

Visualizations should support the project's conclusions rather than simply displaying data.

- **Final conclusions**

The Project Requirements Are Your Roadmap

Follow them step by step to build the entire project.

Individual Project

➤ AI-Assisted Analysis (20 points):

- The Streamlit application must integrate an LLM through LangChain/LangGraph and an LLM API (e.g., OpenAI/Gemini API). **Using web-based AI tools (e.g., ChatGPT's web interface) does not satisfy this requirement.**
- Students should design effective prompts to obtain meaningful insights and assistance from the AI assistant.
- For each major data preparation or analysis decision, students should:
 - formulate an effective prompt,
 - critically evaluate the AI-generated response,
 - explain whether the recommendation was accepted, modified, or rejected, and justify the decision.

Individual Project

➤ AI-Assisted Analysis (20 points):

- AI may be used to assist with tasks such as code generation, visualization, statistical interpretation, and result explanation, but **all AI interactions must be performed through the integrated LangChain + API workflow.**
- Representative examples of AI interactions should be documented in the final report to demonstrate how AI contributed to the project's development.

Simply accepting AI-generated suggestions without critical evaluation or explanation will receive limited credit.

Course Information

- **[Research Report: 10%] (Graduate Students Only)**

Graduate students are required to submit a project report using the **IEEE Conference double-column format**. Please use the official IEEE template (<https://www.overleaf.com/latex/templates/ieee-conference-template/grfzhhnscsfqn>).

Students should choose a project topic relevant to the course and conduct a brief literature review by reading **2–4 related research papers**, including **at least one paper published after January 2025**. The report should clearly reference and discuss these papers.

The report should be **at least 3 pages for a single author** and **5 pages for two or more authors**, inclusive of figures and references. The final report must be submitted before the final week of the course.

Research Report - Grading Policy

- **Research Report: 10% (Graduate Students Only)**

Failure to satisfy **any** of the following requirements will result in a maximum possible score of **80%** on the report:

- The report does not focus on at least one clearly defined data science technique or methodology.
- The required IEEE Conference template is not used.
- The report does not meet the minimum page requirement.
- The report lacks original analysis, discussion, or conclusions.
- The report does not include at least one figure or table.
- The report contains substantial AI-generated content without meaningful original analysis.

The final report must be submitted before the final class (Dec 2) of the course.



You are graduate students. I expect graduate-level thinking, not just more pages

Course Information – Attendance Policy

➤ Attendance Policy:

Regular attendance is expected and strongly encouraged. Routine attendance will not normally be recorded. Students are responsible for managing their own learning and ensuring they remain current with all course materials, announcements, and in-class activities.

If overall class attendance or participation becomes unsatisfactory, the instructor may introduce attendance tracking and/or participation incentives during the semester. Any such changes will be announced in advance.

Course Information

- **Midterm Grades:**

A midterm grade (<https://spartancentral.uncg.edu/student-records/grades/>) will be assigned to all undergraduate students and made available in UNCGenie. This grade reflects your current performance based on completed coursework and is intended **for feedback purposes only**; it will **not** be used in calculating the final course grade. The midterm grade is meant to help you assess whether you are on track or need to make adjustments.

Students receiving a **D or F** at midterm are encouraged to contact the instructor to discuss possible strategies and options for continuing in the course. Final grades will be determined based on all required work completed throughout the semester.

Course Information

- **Suggestions**

- Please pay close attention to the course website (<https://xiaochenli-w.github.io/CSC-405-605-705-Data-Science/>). As the course progresses, lecture slides and assignments will be continuously posted there. In particular, **make sure to check the Schedule module**, where I will post reminders about when assignments are released and due, the timing of upcoming quizzes, and progress updates for different project stages.
- Participation and interaction are strongly encouraged in class.
- Start working on the project early. Emails with questions about assignments or projects sent within two days of the submission deadline may not receive a response.

Course Information

The course is going to be tough, especially for students with limited programming experience

- Work hard, be rewarded with a good data science experience;
- Will talk about the benefits later in course intro;

Do not cheat in the course – Result will be an ‘F’ grade.

- Assignment solutions are unique, differs from student to student.
- **All assignments and course work in this class are individual work. Collaboration is not allowed unless I explicitly tell you otherwise.;**

Course Information

Use of AI Tools

The use of AI tools is **allowed** when used responsibly for **learning purposes** such as understanding concepts, debugging code or understanding error messages. However, all submitted work must reflect the **students' own understanding**. Students are expected to write and modify code by hand, make independent analytical decisions, and fully understand the methods and results they present. Students may be asked to explain or clarify any part of their submitted report, code, or analysis. Inability to adequately explain submitted work may result in grade penalties, including partial credit loss or individual score adjustments.

I'm Looking Forward to a Great Semester Together!

I hope you enjoy the course, challenge yourself, and get something valuable out of it. Let's learn, build, and grow together this semester.

