Department of Statistics and Data Science
Southern University of Science and Technology

STA217: Introduction to Data Science

Lecturer: Yifang MA (马一方)
Office: 353, College of Business (商学院 353)
Email: mayf@sustech.edu.cn (MA)

Lecture Hours: Monday (Every week)
Wednesday (Only even week)

Other useful information:
Lecture slides will be uploaded to the Blackboard.
Class communications will be done through WeChat (Group No. in Blackboard).

1. Course Objectives
This course uses a combination of theory and practice to help students fully understand the basic tools, theories, and methods in data science, including mathematical theories and common methods in data science; analysis and visualization of different data types; complex data cleaning, analyzing, and modeling, etc.

2. Pre-requisites
Before you take the class, you should know: MA102a Mathematical Analysis II / MA102B Calculus II A

3. Course Contents
Part 0 Course Overview
- Introduction to data science

Part 1 Python Programming
- Basics of python: data types; flow control; IO; function & modularity
- Python standard library; built-in functions
- Data structures and data wrangling
- Exploratory data analysis
- ~~Scraping data from web~~

Part 2 Foundational Elements for Programming and Data Science
- Single variable analysis
- Normal distributions
- Data relationships
- Numerical computing using numpy & scipy
- Analyzing tabular data using pandas
- Skewed data analysis
- Basic graph/network data structure

Part 3 Data Analysis and Visualization
- Exploratory data analysis and effective visualization: pandas/matplotlib/seaborn
- Trends, Category, Distribution visualization

Network analysis and visualization
Introduction to interactive visualization

Part 4 Practice
Modeling
Simulation
Code optimization

4. Learning Objectives and Outcomes

On successful completion of the course, students should be able to:

- Use Python and other tools to collect, clean, and process data.
- Use statistical methods to quickly explore, visualize, and describe complex data structures.
- Use data science theory to analyze, model, and predict real data.

5. Evaluation and Grading:

We will assign final grades based on four weighted components:

1) Class Attendance (10%)
2) Homework (about 7-10 Times) (30%)
3) Midterm exam (30%)
4) Project (15%)
   - Content and substance: 70 percent
   - Organization and format: 20 percent
   - English and writing: 10 percent
5) Final Presentation (15%)
   - Submission of project title, research objectives and study plan (1 mark)
   - Group discussion with lecturer (1 mark)
   - Class Presentation (10 marks from student peer reviews and 3 marks from the lecturer and TAs)
     - ✓ Content: Key ideas; Research purpose; Methods, etc.
     - ✓ Structure: Logically organized
     - ✓ Visualization
     - ✓ Delivery, Timing, Teamwork
   - In case of intragroup conflicts, the group members should report to the lecturer in time, the lecturer have the final judgement in mark assignment for each member.

6. Software and Programming

This is a programming-intensive course taught using **Python**, and homework and projects will use Python (version 3.10+). Python is very popular in industry and is free, easy to learn, and has many useful third-party packages. To support Windows, Mac, and Linux, please use:

- **Anaconda**. A free, scientifically focused "bundle" of Python and important Python libraries. It provides a text editor (**Spyder**), enhanced interactive prompt called IPython, and a graphical package manager.

You should download and install the Python 3.10+ version of Anaconda (See our installation instruction). I assume you have a personal computer to work from.

7. Course Material Sources and Attribution

The course materials were adapted from a number of sources. All materials were used for educational, non-commercial reasons only:

- Python data science handbook by Jake VanderPlas
- Introduction to Data Science and Programming by Michael Szell
- Scientific Python course by Roberta Sinatra
- A Whirlwind Tour of Python by Jake VanderPlas and James Bagrow (CC0)
- Data 8 and Data 100 from UC Berkeley Data Science Education Program (DSEP), et. al.
- Data Science Study Notes [数据科学札记] from https://www.cnblogs.com/feffery/
- 6.859 Interactive Data Visualization from mit
- Elements of Computational Communication by Chengjun Wang
- A network science class by Sean Cornelius, Emma Thompson and Albert Laszlo Barabasi
- Data4Sci by Bruno Gonçalves: https://github.com/DataForScience/
- INFO VIZ class by Lingfei Wu

The class draws heavily on materials and examples found online, and we try our best to give credit by attributing to the original source here. Please contact me if you find materials where credit is missing.

Course Agenda (Tentative)

	Monday	Wednesday	Assignments	Comment
第1周 秋季学期 11 廿七 12 廿八 13 廿九 14 三十 15 初一 16 初二 17 初三	Introduction			
第2周 秋季学期 18 初四 19 初五 20 初六 21 初七 22 初八 23 秋分 24 初十	Part 1	Part 1	Ass. 1 Released	
第3周 秋季学期 25 十一 26 十二 27 十三 28 十四 29 中秋节 30 十六 1 国庆节	Part 1			
国庆周 2 十八 3 十九 4 二十 5 廿一 6 廿二 7 廿三 8 寒露	假期			
第4周 秋季学期 9 廿五 10 廿六 11 廿七 12 廿八 13 廿九 14 三十 15 初一	Part 1	Part 1		
第5周 秋季学期 16 初二 17 初三 18 初四 19 初五 20 初六 21 初七 22 初八	Part 2			Final Project Assignment
第6周 秋季学期 23 重阳节 24 霜降 25 十一 26 十二 27 十三 28 十四 29 十五	Part 2	Part 2	Ass. 3 Released	
第7周 秋季学期 30 十六 31 十七 1 十八 2 十九 3 二十 4 廿一 5 廿二	Part 2			Final Project Meeting
第8周 期中考试周 6 廿三 7 廿四 8 立冬 9 廿六 10 廿七 11 廿八 12 廿九	Part 2	Part 2	Ass. 4 Released	Final Project Meeting
第9周 期中考试周 13 初一 14 初二 15 初三 16 初四 17 初五 18 初六 19 初七	Part 2			
第10周 秋季学期 20 儿童日 21 初九 22 小雪 23 十一 24 十二 25 十三 26 十四	Part 3	Part 3	Ass. 5 Released	
第11周 秋季学期 27 十五 28 十六 29 十七 30 十八 1 艾滋病日 2 二十 3 廿一	Part 3			Midterm
第12周 秋季学期 4 廿二 5 廿三 6 廿四 7 大雪 8 廿六 9 廿七 10 廿八	Part 3	Part 3	Ass. 6 Released	
第13周 秋季学期 11 廿九 12 三十 13 初一 14 初二 15 初三 16 初四 17 初五	Part 4			
第14周 秋季学期 18 初六 19 初七 20 初八 21 初九 22 冬至 23 十一 24 十二	Part 4	Part 4	Ass. 7 Released	
第15周 秋季学期 25 十三 26 十四 27 十五 28 十六 29 十七 30 十八 31 十九	Group Presentation			
第 16 周	Group Presentation	Summary and Conclusion		