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

# *STA217: Introduction to Data Science*

**Lecturer:** Yifang MA (马一方)
**Office:** 405-2, Innovation Park 8# （创园 8 栋 405-2）
**Email:** mayf@sustech.edu.cn (MA)

**Lecture Hours and Venue:** Monday (Every week, 每周 3-4 节), 一教 401
Wednesday (Only even week, 双周 3-4 节), 一教 401

**Other useful information:**
Lecture slides will be uploaded to the Blackboard.
Class communications will be done through QQ Group: 839454354

**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
- Scraping data from web (Optional)

Part 2 Mathematics, Programming and Data Science
- Basic probabilities
- Single variable analysis
- Normal distributions
- Data relationships
- Numerical computing using numpy & scipy
- Analyzing tabular data using pandas
- Skewed data
- Basic graph/network data structure
- Time series, text analysis

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

Interactive visualization

Part 4 Practice
Regression
Selected machine learning practice
Code optimization
Network analysis and visualization

**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) (50%)
3) Project (20%)
   - Content and substance: 70 percent
   - Presentation and format: 20 percent
   - English and writing: 10 percent
4) Final Presentation (20%)
   - Submission of project title, research objectives and study plan (1 mark)
   - Group discussion with lecturer (1 mark)
   - Class Presentation (18 marks)
     - 9 points for group presentation
       - ✓ Content (4 marks): Key ideas; Research purpose; Methods, etc.
       - ✓ Structure (1 marks): Logically organized
       - ✓ Visualization (2 marks)
       - ✓ Delivery, Timing, Teamwork (2 marks)
     - 9 points for peer reviews
       - ✓ Reasonable evaluation with comments (1 marks)
       - ✓ Peer reviews from other groups (8 marks)
   - 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 3.7+). 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 3.7+ version of Anaconda (See our installation instruction). I assume you have a personal computer to work from.

# Course Agenda (Tentative)

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