MSIS 2802: Data Science Analysis with Python

Instructor ~ Michael Schermann, mschermann@scu.edu, +1 408 554 6832, Lucas Hall 316A
Teaching Assistant ~ Meiyuan Li, mli@scu.edu

Contents

Motivation

The objective of this course is to teach the **analytical mindset & programming skills** relevant to data science. You will learn how to explore and analyze data using data science libaries and the Python programming language. More specifically, you will learn how to work with Jupyter notebooks, NumPy, Pandas, Seaborn, and Scikit-learn.

Students will learn skills that cover the various phases of **exploratory data analysis**:

- analytical thinking,
- importing data,
- cleaning and transforming data,

- grouping and aggregation,
- using time series and statistical modeling &
- visualization and communication of results.

Upon successful completion of this course, you have acquired the fundamental skills for a data-centric career in the Silicon Valley.

Learning objectives

- Understand the fundamentals of Python,
- Prepare data sources for analysis,
- Develop solutions to data science problems,
- Create a persuasive data analysis.

Course Logistics

Online Python Tutorial

The goal is to spend as much time in the classroom on establishing and honing your analytical mindset. Thus, as part of the class, you will complete the online Python tutorial on codecademy.com. This is an excellent, self-guided and self-paced tutorial on the Python programming language. You are required to complete the tutorial by **January 26, 2018**. This will allow us to spend the time in class on covering the basics of data science.

Textbooks

- https://jakevdp.github.io/PythonDataScienceHandbook - A compendium on Python in a data-intense environment. **You should buy this book because it will be an invaluable companion for the next couple of years.**
- https://jakevdp.github.io/WhirlwindTourOfPython - An excellent introduction to Python for the data scientist.
- https://wiki.python.org/moin/BeginnersGuide/Programmers - An extensive list of how to program in Python.

Technology

The hands-on elements in this course use the **Jupyter** environment. Jupyter is a web applications for literate programming. Literate programming integrates the explanation of the code with the actual code, which is particularly important in data science environments.

PLEASE NOTE: This course involves extensive programming and computer work, both in and out of class. You are required to bring a laptop to class each time and have Jupyter installed on your laptop. If you run into any issues, please let me or the TA know.

Communication

I am committed to your learning success. Please feel free to contact me with any questions regarding this course. If I am not able to help you myself, I will forward your request to someone who can.

1. If you have general questions about course material, assignments, etc. please write them into this FAQ document (accessible only with SCU ID).
2. Before you write an email, please read and comment in the FAQ document (accessible only with SCU ID).
3. If you send me an email that contains questions of interest to the whole class, I will answer them in the FAQ document (accessible only with SCU ID).
4. My office hours are Tuesdays and Thursdays after class from 9:00pm to 10:00pm. Please make an appointment here.
5. Please make an appointment whether you want to meet during office hours or outside of my office hours. A meeting request must have a specific agenda. I am available via phone, zoom, or face-to-face.
6. I post all course material, course information, announcements, and updates on Camino. Please make sure that your correct email address is listed in Camino so that you do not miss important information.

Class Meetings

Class meetings are **Tuesdays and Thursdays**, **7:35 PM to 8:50 PM** in **Vari Hall 134**.

Each class will focus on **one specific conceptual issue** and discusses solutions to this issue. During the class meeting, you have to work on **data science problems**. At the end of each meeting, **I may ask you to present and explain your results**.

Assignments

> *"What it boils down to is one per cent inspiration and ninety-nine per cent perspiration." (Thomas Edison)*

Your mastery of the learning objectives will be examined through the online Python tutorial, a midterm exam, weekly homework, and a team project. **There will be no final exam.**

The following table links the learning objectives of this class with the assignments and shows the maximum number of points that you can achieve of each assignment in the final grade.

Learning Objective	Assignment	Max. Points
Understand the fundamentals of Python.	Online tutorial	10
Prepare data sources for analysis.	Mid-term exam	30
Develop solutions to data science problems.	Weekly homework	30
Create a persuasive data analysis.	Team Project	30
Total		**100**

The final grade distribution is as follows.

Points	Letter Grade
100-94	A
>94-90	A-
>90-87	B+
>87-84	B
>84-80	B-

Points	Letter Grade
>80-77	C+
>77-74	C
>74-70	C-
>70-0	F

My grading criteria are as follows:

- **A grades** (4.0) reflect work that meets all assignment objectives at the highest possible level and sometimes goes beyond that. The submitted work is of superior quality and could be used in productive environments with no or minimal revisions. Typically, no more than 40% of participants in a course receive an A grade.
- **B grades** (3.0) reflect work that meets all assignment objectives at a level that is above average but not exceptional. The submitted work shows high levels of competency and could be used in productive environments with some editing.
- **C grades** (2.0) reflect work that meets all course objectives at an average level but is not exceeding expected standards. The submitted work lacks a clear in-depth understanding of the subject and could be used in productive environments only with extensive editing. Typically, at least 5% of participants in a course receive a C grade.
- **F grades** (0.0) reflect work that does not meet course objectives and is below minimum standards. Submissions are late without prior consultation with the instructor, miss the assignment objectives, or show a clear lack of learning progress. Also, repeated violations of the academic integrity standards result in an overall F grade.

I reserve the right to change the grading to accommodate special circumstances and opportunities. Any changes, however, will be discussed and announced in class and on Camino.

Online tutorial

You will submit a confirmation of completing the online tutorial no later than **January 26, 2018 (11:59 PM)**. I will **not** accept late submission without prior notice or without a doctor's note. I am aware that sometimes life goes crazy but please notify me in advance and we will work it out.

Mid-term exam

In the mid-term exam you will demonstrate your ability to prepare a dataset for analysis. Commonly, data scientists spend about 60-80 percent of their time with data preparation (cleaning, exploring, transforming, etc.).

The mid-term exam consists of a in-class component and a take-home component:

- The **in-class component (20 points)** of the mid-term exam will be a 60 minutes long multiple-choice quiz with 20 questions on **February, 8 2018** during class session.
- The **take-home component (10 points)** is due on **February, 10 2018 (11:59 PM)**.

Both components will be administered through Camino. I will **not** accept late submission for the take-home component without prior notice or without a doctor's note. I am aware that sometimes life goes crazy but please notify me in advance and we will work it out.

PLEASE NOTE: Both components of the mid-term exam are "open book", which means that you may use the internet, class notes, books, etc. during the exam time.

PLEASE NOTE: The take-home component is individual work. You **may discuss** the take-home component but you **must not share** intermediate or final solutions with fellow students. Please review the academic integrity rules below.

I will evaluate the take-home submissions based on the following criteria:

Criteria	Metrics	Max. Points
Data description	Understandability (1), Completeness (1)	2
Data preparation & use	Clarity (1), Explanations (1), Class Concepts (1)	3
Finding	Persuasiveness (1), Evidence (2)	3
Style	Professionality (1), Originality (1)	2
Total		**10**

Persuasiveness metrics assess the structure and the strength of your findings from a managerial perspective (Can you answer the "So what"-question? Do you provide actionable insights?). **Evidence** metrics assess the functionality, efficiency, and reproducibility of your findings from a data science perspective (Did you produce codes that other data scientists can build upon?).

Homework

Eight (8) weekly homework assignments help you to apply the class material to real-world data science problems. You will face increasingly difficult problems as we progress in the quarter. Homework assignments will be published each week on Tuesdays and are due on Fridays (11:59 PM) of the same week. No homework during the week of the mid-term exam and the final week. I will **not** accept late submission without prior notice or without a doctor's note. I am aware that sometimes life goes crazy but please notify me in advance and we will work it out.

PLEASE NOTE: Homework is individual work. You **may discuss** homework assignments but you **must not share** intermediate or final solutions with fellow students. Please review the academic integrity rules below.

Homework	Due	Max. Points
1	January, 12 2018 (11:59 PM)	3
2	January, 19 2018 (11:59 PM)	3
3 1	January, 26 2018 (11:59 PM)	3
4	February, 2 2018 (11:59 PM)	3
5	February, 16 2018 (11:59 PM)	4
6	February, 23 2018 (11:59 PM)	4
7	March, 2 2018 (11:59 PM)	5
8	March, 9 2018 (11:59 PM)	5
Total		**30**

You have to submit **one** Jupyter notebook as a **github** link (I will not accept any other forms of submission.). Make sure that the notebook renders correctly on github.

I will evaluate homework submissions based on the following criteria:

Criteria	Question	Max. Points
Completion	All objective completed?	1
Explanation	Do you explain how you arrived at the solution?	1-2
Efficiency	Can you show that your solution is efficient?	0-1
Style	Professionality (1)	1
Total		**3-5**

Team project

The objective of this project is to demonstrate mastery of the class material. You will work in teams of **two students** on a dataset of your own choice. Your objective is to prepare the data and detect three **interesting**, **non-trivial**, and **somewhat unexpected** findings. It is your responsibility to explain and present **why** your findings are interesting, non-trivial, and somewhat unexpected.

You have to notify me of your choice of dataset by **February 6, 2018 (11:59 PM)**.

You are **highly** encouraged to structure the project according to the following schedule:

Phase	Should be done by
Identify Dataset	Week 5
Data preparation	Week 6
Finding 1	Week 7
Finding 2	Week 8
Finding 3	Week 9
Presentation	Week 10
Polishing	Finals Week

The project is due on **March, 20, 2018, 11:59 PM**. You have to submit **one** Jupyter notebook as a **github** link (I will not accept any other forms of submission.). Make sure that the notebook renders correctly on github. I will **not** accept late submission without prior notice or without a doctor's note. I am aware that sometimes life goes crazy but please notify me in advance and we will work it out.

The Jupyter notebook should includes the following components:

1. Project statement (Motivation, specific goals, data).
2. "Making-of" documentation (Details of your development process, data wrangling steps, your reasoning, detours, literature, etc.).
3. Description of your results.
4. Roadmap for future analysis/enhancements of your analysis

I will evaluate project submissions based on the following criteria:

Criteria	Metrics	Max. Points
Data description	Understandability (1), Completeness (1)	2
Data preparation & use	Clarity (2), Explanations (2), Class Concepts (2), Efficiency (2)	8
Finding 1	Persuasiveness (3), Evidence (3)	6

Criteria	Metrics	Max. Points
Finding 2	Persuasiveness (3), Evidence (3)	6
Finding 3	Persuasiveness (3), Evidence (3)	6
Style	Professionality (1), Originality (1)	2
Total		**30**

Persuasiveness metrics assess the structure and the strength of your findings from a managerial perspective (Can you answer the "So what"-question? Do you provide actionable insights?). **Evidence** metrics assess the functionality, efficiency, and reproducibility of your findings from a data science perspective (Did you produce codes that other data scientists can build upon?).

How to get an A in this course

I firmly believe that mastery of data science requires constant practice. You will ace this course if you:

- Adhere to the academic integrity standards outlined below.
- Be ready for class meetings, which means you have done the homework and read the textbook.
- Participate in the class discussions, ask questions, and share experiences.
- Support your teammates (If you can explain it to a fellow student, you know that you have understood it yourself).
- Start early on the assignments, **seek feedback from me, the TA, and other sources**.
- Continuously think about **why** you are doing something in your assignments. This is far more important than **what** you are doing.
- Answer the **'boss question'** before submitting **any** deliverable: Would you send your submission **as is** to your boss or to a recruiter? If not, please do not submit it.

Course Schedule

Week	Class Meeting	Topic
1	January, 9	Introduction
1	January, 11	Jupyter in a whirlwind
2	January, 16	Series
2	January, 18	Indexing & selection
3	January, 23	Dataframes
3	January, 25	Grouping & aggregation
4	January, 30	String operations
4	February, 1	Apply
5	February, 6	Review
5	February, 8	Midterm
6	February, 13	Matplotlib & seaborn
6	February, 15	Take-home review
7	February, 20	Exploratory classification
7	February, 22	Predictive classification
8	February, 27	Clustering
8	March, 1	Regression

Week	Class Meeting	Topic
9	March, 6	Team project review
9	March, 8	Putting it all together
10	March, 13	Outlook
10	March, 15	Team presentations

I reserve the right to change the schedule to accommodate special circumstances and opportunities. Any changes, however, will be discussed and announced in class and on Camino.

Academic Integrity

The Academic Integrity pledge is an expression of the University's commitment to fostering an understanding of and commitment to a culture of integrity at Santa Clara University. The Academic Integrity pledge, which applies to all students, states:

> *"I am committed to being a person of integrity. I pledge, as a member of the Santa Clara University community, to abide by and uphold the standards of academic integrity contained in the Student Conduct Code."*

You are expected to uphold the principles of this pledge for all work in this class. For more information about Santa Clara University's academic integrity pledge and resources about ensuring academic integrity in your work, see www.scu.edu/academic-integrity.

In particular, I expect that you give credit to any material (including but not limited to journal articles, web article, blog posts, images, data sets, libraries, APIs, and any media) that you have used for completing any assignment in this class. Being able to give credit by referencing sources consistently and correctly is evidence of mastery of a topic. It shows that you are able to construct original arguments that are backed with verifiable evidence. Failing to give credit is a sign of an inadequate learning progress. It shows that you have not understood the topic well enough to formulate your own arguments in relation to already existing ideas.

During your work in this class, you will use, modify, or extend digital content that you have found online. You will also use libraries, APIs, code snippets, and data sets that have been created by others. In every piece of work (presentations, assignments, etc.), you must acknowledge work, source code, data sets, and any other content that was not produced by you. Acknowledgements must be easily identifiable, inseparable from your content, and must not violate licenses.

Failure to provide appropriate acknowledgements will result in an F grade for that assignment. Repeated failure to provide appropriate acknowledgements will result in an F grade for the entire course.

During the first class, we will discuss this digital content policy. After this class, I will strictly enforce this policy. If you have doubts, contact me.

Course Conduct

My responsibility

I will support you in your learning in this class and beyond to the best of my abilities. If I am not able to help you myself, I will identify someone who can. I will evaluate your contribution solely based on the standards set by this syllabus. Changes to the syllabus will be highlighted, discussed during class sessions, and will be published on Camino.

Your responsibility

By enrolling in this class, you agree to the requirements stated in this syllabus. You will operate with integrity in your dealings with me and your fellow students. You will engage the learning materials with appropriate attention and dedication and maintain their engagement when challenged by difficult learning activities. You will contribute to the learning of others and you will perform to standards set by this syllabus.

Mutual respect is the foundation of this course. No one will be criticized for being wrong. Appropriate conduct includes honesty, self-respect, respect for others, and compliance with university policies and standards. Computers in the classroom should be used only for completing course-related work and for taking notes; cell phones must be turned off or muted.

Attendance Policy

Please let me know via email during the first two weeks of the course if you have any conflicts between a course element (class meeting, assignment) and another vital commitment (another course, work, university-related extracurricular activities, religious commitments). At my discretion, I will you provide with alternative means to complete the course element.

I am aware that many of you have multiple commitments. You should attend at least 80 percent of all scheduled class meetings. If you miss more than 20 percent of scheduled classes, you will receive reduction by one letter grade.

University Policies

Disability Resources

If you have a disability for which accommodations may be required in this class, please contact Disabilities Resources (Benson Hall 216, 408-554-4109) as soon as possible to discuss your needs and register for accommodations with the University. If you have medical needs related to pregnancy, you may also be eligible for accommodations. If you have already arranged accommodations through Disabilities Resources, please discuss them with me during my office hours as soon as possible.

While I am happy to assist you, I am unable to provide accommodations until I have received verification from Disabilities Resources. If you are in doubt of whether you are eligible for accommodations, I encourage you to contact Disabilities Resources (Benson Hall 216, 408-554-4109). The Disabilities Resources office would be grateful for advance notice of at least two weeks.

Accommodations for Pregnancy and Parenting

In alignment with Title IX of the Education Amendments of 1972, and with the California Education Code, Section 66281.7, Santa Clara University provides reasonable accommodations to students who are pregnant, have recently experienced childbirth, and/or have medically needs related to childbirth. Pregnant and parenting students can often arrange accommodations by working directly with their instructors, supervisors, or departments. Alternatively, a pregnant or parenting student experiencing related medical conditions may request accommodations through Disabilities Resources (Benson Hall 216, 408-554-4109).

Discrimination and Sexual Misconduct (Title IX)

Santa Clara University upholds a zero-tolerance policy for discrimination, harassment and sexual misconduct. If you (or someone you know) have experienced discrimination or harassment, including sexual assault,

domestic/dating violence, or stalking, I encourage you to tell someone promptly. For more information, please consult the University's Gender-Based Discrimination and Sexual Misconduct Policy at http://bit.ly/2ce1hBb or contact the University's EEO and Title IX Coordinator, Belinda Guthrie, at 408-554-3043, bguthrie@scu.edu. Reports may be submitted online through https://www.scu.edu/osl/report/ or anonymously through Ethicspoint https://www.scu.edu/hr/quick-links/ethicspoint/

Acknowledgement

This syllabus was inspired by Aleszu Bajak's syllabus. The class was developed by Michele Samorani at Santa Clara University.