# STAT 004 - Elements of Data Science

**Instructor Information**


- Instructor: Yingzhuo (Joyce) Fu. **joyce.fu@ucr.edu**
- TA: Namhwa Lee nlee098@ucr.edu
  Yudi Mu ymu015@ucr.edu


**Meeting Information**

Lecture: Mon-Thur 10:00 – 11:20 am Zoom Meeting Registration Link
Discussion: TBD
Joyce's Office Hours: Mon-Thur 9:00am -10:00am or by Appt Zoom Link: same as lecture
TA Office Hours: TBD
**Prerequisite**: None

**Course Description**

(4 credits) This course is an elementary introduction to data science. It is designed for students from all majors who are interested in learning about data science. Topics include summarizing and visualizing data using R or Python; causality, randomness, decision making in the presence of uncertainty, prediction and classification.

**Textbook**: No specific textbook is required
Recommended Readings:

- **[AD-DataScience]** Adhikari, Ani and DeNero, John. **"** *The Foundations of Data Science".* The electronic version is free at *https://www.inferentialthinking.com/chapters/intro*
- **[OS-DataScience]** O'Neil, Schutt. *"Doing Data Science: Straight Talk from the Frontline". (2013) 1st Edition*
- **[DC-OpenIntroStat]** Diez, Cetinkaya-Rundel, Barr. *"OpenIntro Statistics"* (2019) 4th Edition. The electronic version is free at https://leanpub.com/openintro-statistics
- **[GS-CartoonStat]** Gonick, Smith. *"The Cartoon Guide to Statistics"* (2005) 1st Edition
- **[BF-Data]** Bueno de Mesquita, Ethan, Anthoney Flower. "Thinking Clearly with Data" (2021)

**Grading System**

| | | |
|---|---|---|
| Problem Sets | 20% | Weekly (10 HWs in total) |
| Application Report | 10% | |
| Discussion Participation | 5% | |
| Quizzes | 15% | Monday Weeks 2, 3, 5 |
| Midterm Exam | 20% | Monday Week 4 |
| Final Exam | 30% | Thursday Week 5 |
| Bonus | | |

**Course Structure**

| Activity | Platform | Delivery | Details |
|---|---|---|---|
| Lecture | Zoom | Synchronous | Recordings will be available on Canvas by 5pm of the day for students that are unable to attend or prefer to attend asynchronously. |
| Discussion | Zoom | Synchronous | Recordings will be available on Canvas by 5pm of the day for students that are unable to attend or prefer to attend asynchronously. |
| Participation | Canvas | Asynchronous | Participation will be assessed through contributions to the weekly discussion boards in Canvas. |
| Problem Sets | Canvas | Asynchronous | Available in Canvas in the weekly modules. |
| Quizzes | Canvas | Synchronous | Open notes. In class. Available in Canvas in the weekly modules. 20-minute time limit. |
| Exams | Canvas and Yuja | Synchronous | Closed book. One page (two-sided) of cheatsheet is allowed. Students may use a calculator. Any necessary probability tables or Python output will be provided. |
| Technical Support | Canvas/YuJa | Asynchronous | Short tutorial videos for Python will be available on Canvas in the weekly modules. |

**Discussion**

Discussion time will be used as additional time for computing demonstration in Python and getting help on homework questions. You will have the opportunity to play with data. Basic data processing knowledge within Python will be introduced. Topics such as data types, data frames, iteration, and conditional statements will be covered. However, this course emphasizes less on syntax and more on logic and main workflow. Programming code examples will be provided.

**Software installation**

- Download Python at https://www.python.org/downloads/. The website will automatically detect your system type and you only need to choose a version that is 3.5 or up.
- Python can also be used online or when the computer you are using doesn't have Python installed.
    - For python, go to https://cocalc.com/doc/jupyter-notebook.html.

  NOTE: This site require creating an account to keep projects saved in the cloud.

**Learning Community**

Student participation in the course will be measured using contributions to the discussion boards in Canvas. Posts will be graded credit/no credit based on appropriateness (e.g., yes/no, I agree/disagree, or similar responses will not receive credit). Proper conduct and etiquette are expected – be respectful of your classmates and their opinions/posts. Each appropirate post can earn you 0.5% of bonus point. Total bonus is capped at 2%.

Examples of appropriate posts are:

1. Here is an interesting video about misleading statistics. https://www.youtube.com/watch?v=sxYrzzy3cq8
2. I have a question about understanding the conditional probability. What's the difference between a conditoinal probability and the usual probability?
3. Re: post#2. My understanding is that conditional probability has an extra piece of information, which helps to narrow down the possible outcomes. For example, if you roll a dice and it's given the number is an even number, then the probability of getting a 4 is 1/3, not 1/6.

**Problem sets**

Weekly problem sets will consist of roughly eight to ten questions. The questions will provide an opportunity for students to strengthen their understanding of the topics covered in lecture and also help students to prepare for the exams. The assignments will be posted and submitted through Canvas by the posted due date. Grading will be based on completeness as well as accuracy. The **lowest** homework score will be **dropped**.

**Application Experiment Report**

To get a complete and in-depth understanding of the key concepts covered in this class, multiple application experiments such as a poker game, Monty Hall problem, or to choose your side (KNN) will be carried out by students. The requirements for experiment reports will be posted in Canvas. It is strongly suggested to use excel or google sheet for data recording purposes. It would be helpful to work in a word processor of your choice (MS work, Google Doc, etc) and then submit your summary and conclusions in PDF form so we can see the results in the way that it is supposed to be. Reports will be submitted in Canvas.

**Quizzes and Exams**

There will be three quizzes, one midterm, and one final exam. Quizzes will be available in Canvas during class time and have a 20-minute time limit once started. There are **no** make-ups for quizzes. Exams will be completed using Yuja Proctoring and Gradescope. If there are situations preventing you from taking a midterm or final exam, you must contact the instructor at least 24 hours in advance (or within 24 hours in case of an emergency). Proper documentation must be provided in order for a make-up of midterm or final exam to be approved.

**Workload**

You are expected to put in approximately 4-6 hours of work per week outside of class. The time will vary from student to student, but it mostly includes the time to preview, review course materials, and finish assignments.

**Tips for Success**

1. Interest is the best motivation. Find a problem or a situation that you are puzzled upon and could be related to data science.
2. Understanding the basic concept/logic is much more important than just carrying out the calculation process correctly. Do the application experiments carefully and try to find more examples related to your daily interests.
3. Ask questions to me, your fellow students, or your TA. There are no dumb questions except the questions that are not asked.
4. Find a study group. You are encouraged to discuss any in-class questions and/or homework questions with your classmates. Questions may also be posted in the discussion boards in Canvas which will be monitored by the Instructor and TAs. Just make sure that you turn in your own work.

**Technology Requirements**

Because this course takes place online, you will need the following:

**Hardware**

- Access to a current Mac or PC (with a fast processor and speakers)
- Webcam and microphone (to participate in any video components, e.g., live sessions, remote proctoring, video presentations)

**Software**

- Windows 8.1 or OS 10.14 or above is recommended.
- Please Note: Some software applications (such as Respondus LockDown Browser) are not supported by Chromebooks.

**Internet Connection**

You will need a reliable internet connection. Test your internet speed using speedtest.net.
Recommended internet speeds:

- Ping response is less than 100ms
- Download speed is at least 10Mbps
- Upload speed is at least 2Mbps

**Browsers**

- Chrome or Firefox work best (avoid Safari)
- Make sure you are using the most up-to-date browser

If you do not have or cannot acquire these items, you can apply to the Loan2Learn program to get a loaner device. Please keep me aware of any difficulty or delay in acquiring these devices.

For technology support, contact Bearhelp@ucr.edu.

**Course Modules**

| Module | Topics | Recommended Reading |
|---|---|---|
| 1 | **Introduction to Data Science**<br><br>How data facilitates big decision making in business, science, politics.<br><br>Data visualization. Explore statistics with two examples. | [OS-DataScience]<br>CH 1 What is data science?<br><br>[AD-DataScience]<br>CH 1.3 Plotting the Classics |
| 2 | **Data Basics and Introduction to Python**<br><br>Data types, data storage and processing using "table" in Python.<br><br>Iterations, conditional statement for simple programming<br><br>Examine center and variation<br><br>**Quiz** | [AD-DataScience]<br>CH 4, 6, 7 Data Types, Tables, visualization<br><br>[DC-OpenIntroStat]<br>CH 2.1.2 - 2.1.7 Examining Numerical Data |
| 3 | **Causality and Experiment**<br><br>Case study of John Snow and the Broad Street pump<br><br>Establishing causality<br><br>Randomization | [AD-DataScience]<br>CH 2 Causality and Experiment<br><br>[DC-OpenIntroStat]<br>CH 1.3 Sampling Principle |
| 4 | **Randomness**<br><br>Monty Hall problem to introduce randomness<br><br>Simulations to show Law of Large Numbers<br><br>Conditional probability and Bayes Rule<br><br>Random variables, variability with simulation<br><br>**Quiz** | [GS-CartoonStat]<br>CH 3 Probability<br><br>[DC-OpenIntroStat]<br>CH 3 Probability<br><br>[GS-CartoonStat]<br>CH 4 Random Variables |
| 5 | **Distributions**<br><br>Binomial Distribution, Bernoulli random variable, binomial probabilities<br><br>Normal Distribution, standardizing, tail area<br>Recommended Readings | [GS-CartoonStat]<br>CH 5 A Tale of Two Distributions<br><br>[DC-OpenIntroStat]<br>CH 4.1, 4.3 Distributions of Random Variables |

| | | |
|---|---|---|
| 6 | **Population and Sample**<br><br>Sampling Strategies, source of bias<br><br>Point estimate, variability of the estimates<br><br>Simulations to demonstrate variability of the estimates<br>**Midterm** | [GS-CartoonStat]<br>CH 6 Sampling<br><br>[DC-OpenIntroStat]<br>CH 5.1 Point Estimates and Sampling Variability |
| 7 - 8 | **Decision Making in the Presence of Uncertainty**<br><br>Simulation to visualize the variability of sample mean<br><br>Confidence Interval for mean/proportion<br><br>Hypothesis testing, A/B testing<br><br>**Quiz** | [AD-DataScience]<br>CH 11 Testing Hypothesis,<br>CH 12.1 A/B testing<br><br>[GS-CartoonStat]<br>CH 7 Confidence Interval<br>CH 8 Hypothesis Testing<br><br>[DC-OpenIntroStat]<br>CH 5.2, 5.3 Confidence Interval, Hypothesis Testing |
| 9 | **Predictions**<br><br>Correlations<br><br>Regression line for two variables<br><br>Diagnostic of linear regression | [AD-DataScience]<br>CH 16 Inference for Regression<br><br>[GS-CartoonStat]<br>CH 11 Regression<br><br>[BF-Data]<br>CH5 Regression for Describing and Forecasting |
| 10 | **Classification and Clustering**<br><br>Prediction and classification with k-nearest neighbours (KNN)<br><br>Visualizing k-means clustering<br><br>Training set and validation set, overfitting | [AD-DataScience]<br>CH 17 Classification |
| Finals | **Exam** Augst 25 Thursday in class | |