STATISTICS 167, Spring 2019
Introduction to Data Science

Course Description and Objectives
Data science is an emerging field about processing and managing big data, extracting and discovering structures and knowledge from the data, and making predictions and decisions. Data science is an interdisciplinary field at the frontier of statistics and computer science.

In this course, we will explore real data sets, practice scalable data management and analytical skills, explore various statistical learning models, and determine what conclusions we can draw from the data. During the class, you will perform big data analyses with existing software packages, and write simple programs to implement and extend key techniques. You will also need to write reports about your analyses.

Upon completing this course, you will be able to tackle real-world data-driven problems, by:
(1) efficiently managing massive data sets;
(2) exploring various visual data analytics to understand the data sets
(3) selecting the appropriate statistical and machine learning models (both supervised and unsupervised methods) and justifying your choices;
(4) implementing these methods programmatically (with R);
(5) critically and thoroughly evaluating the results;
(6) presenting and communicating your results to non-statisticians.

Prerequisites: STAT 100B or STAT 155 (with a grade of C- or better) or equivalents; STAT 147 (with a grade of C- or better); CS010 (with a grade of C- or better)

Teaching team
Instructor: Dr. Wenxiu Ma  Email: wenxiu.ma@ucr.edu
Office: Olmsted Hall 1423  Office Hours: Monday 3:10 pm - 5:00 pm

TA: Luke Klein  Email: lklei001@ucr.edu
Office: Olmsted Hall 1108  Office Hours: Wednesday 11:00 am - 12:00 pm

Class Hours
Lecture: MWF 02:10 pm - 03:00 pm  Olmsted Hall 1431
Discussion: F 01:10 pm - 02:00 pm  Olmsted Hall 1411

Textbook
  Electronic copy of the textbook is freely available at http://r4ds.had.co.nz/

Optional readings:
  Electronic copy (Restricted to UC campus) http://dx.doi.org/10.1007/978-3-319-24277-4
  Electronic copy is freely available online at http://www-bcf.usc.edu/~gareth/ISL/

Course Websites
• iLearn website: https://ilearn.ucr.edu/[announcements; lecture notes; course material]
• Piazza website: https://piazza.com/ucr/spring2019/stat167 [discussion board]
• Poll Everywhere website https://pollev.com/stat167 [in-class participation]

These course websites are integral parts of this class. Please be sure to check the course websites and your email account on a regular basis for class announcements, homework assignments, and miscellaneous information.

Grading system
Lecture Participation: 10%  Midterm: 15%
Discussion: 15%  Final Project: 30%
Homework: 30%

<table>
<thead>
<tr>
<th>Grade</th>
<th>A+</th>
<th>A</th>
<th>A-</th>
<th>B+</th>
<th>B</th>
<th>B-</th>
<th>C+</th>
<th>C</th>
<th>C-</th>
<th>D+</th>
<th>D</th>
<th>D-</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>TBD</td>
<td>&gt;92</td>
<td>90&lt;92</td>
<td>87&lt;90</td>
<td>82&lt;87</td>
<td>80&lt;82</td>
<td>77&lt;80</td>
<td>72&lt;77</td>
<td>70&lt;72</td>
<td>67&lt;70</td>
<td>62&lt;67</td>
<td>60&lt;62</td>
<td>&lt;60</td>
<td></td>
</tr>
</tbody>
</table>

Grades are not curved or adjusted. This is because the grade is based on individual effort and not on comparative effort. The grade of A+ will be given to the top students in the course.

R and RStudio
R is a free, open-source programming language for statistical computing. Almost all of our work in this class will be done using R.
RStudio is a free, open-source R programming environment. It contains a built-in code editor, many features to make working with R easier, and works the same way across different operating systems. Use of RStudio is strongly recommended for this course.

Lecture Participation
This course will consist of lectures and discussion sessions. You are expected to attend all lectures. The lectures will cover the material necessary to prepare you for the discussion sessions, the homework assignments.
Lecture participation is tracked by Poll Everywhere, a web-based classroom polling system. During lectures, your instructor will display Poll Everywhere activities on-screen and you can respond via web-enabled devices (phones, tablets, or laptops). It is recommended to install the Poll Everywhere app on your iOS or Android device. Poll Everywhere is integrated with iLearn, so there’s no need to create your own account. Remember to log in with your UCR email address and your CAS credentials to receive credit for your responses.
Discussion
The discussion section will meet weekly in the OLMH 1411 computer lab. During the discussion sessions, data analysis concepts will be reviewed and discussed, then you will practice programming and data analytical skills on example datasets. You will submit your completed exercises at the end of each discussion session. **There are no make-up discussion sessions.** To account for emergencies that prevent discussion attendance, your lowest discussion score will be dropped.

Homework
Homework will be assigned on a regular basis, approximately one per week. Homework assignments will consist of both data analytical and programming exercises, providing opportunities to practice the statistical and computational concepts discussed in the class. The assignments will be posted on the course website.

All assignments must be turned in electronically. You must submit your assignment using R Markdown. Exceptions may be made, with prior permission, for those who want to use Sweave/knitr. (If you don’t know what it is, plan to use R Markdown.) Work submitted in any other format will receive an automatic grade of 0, without exceptions.

Every file you submit should have a file name which includes your first and last name, your Student ID, and clearly indicates the assignment type (homework, lab, etc.) and its number. For example, FirstnameLastname-SID-HW1.rmd and FirstnameLastname-SID-HW1.html.

**Late Policy:**
No assignment will be accepted after 2 days past its initial deadline.
- Up to 24 hours late, the assignment will incur a 10 percent reduction;
- Up to 48 hours late, the assignment will incur a 30 percent reduction.

**Each student will have a total of two “free” late days (a late day is 24 hours of lateness). There are no partial days.** Start early and make sure your environment is working correctly and you are able to produce a working R Markdown document. If you submit multiple attempts, the final attempt will be graded. If your first submission is on time, but your final submission is late, your will incur the late submission penalty.

**Collaboration, Copying and Plagiarism:**
You are encouraged to discuss homework assignments with each other, but the submitted work must be your original work (both written work and computer code). **Explicitly sharing your written work or code with someone else is not allowed.**

Midterm
There will be one take-home midterm exam. The midterm exam will be scheduled approximately in the end of the sixth week. The midterm exam is open book/internet access, but absolutely no communicating with other humans.

Final Project
There is no written final exam for this course. Instead, you will work on a final project in small groups (4-5 persons per group). For the final project, you are expected to perform a complete data analysis workflow on a real data application. You will conduct exploratory data analysis, choose appropriate supervised or unsupervised learning models, then
implement necessary analyses with your own R code, and determine what conclusions to draw from the data (if any). Each group will have regular meetings with the instructor and/or the TA to discuss the progress of your final project. More details of the final project are to be announced.

At the end of the quarter, you will present your final project and findings in class. Each group will also submit a final project written report (using R Markdown).

**Computer Lab Open Hours**

**OLMH 1316 computer lab**
- Tuesday and Thursday, 4:10pm - 8:00pm: open to campus (subject to change according to Student Computing scheduling)
- Friday, 9:10am - 3:00pm (beginning at Week 2): for Statistics course-related work. You will need a special permission slip (available from your TA). You must sign in and out and your permission slip must be kept visible while you are in the lab. If you are found doing something else, you will be asked to leave the lab. The lab will be monitored by Statistics 48, 100A, and 100B TAs, so you will NOT be able to ask them for help!

**OLMH 1318 computer lab**
- Monday to Friday, 8:10am - 4:30pm. You must get a permission slip (available from your TA). Please be sure you have your ID with you and say that you are in STAT 167.

*^^^ WELCOME TO THE COURSE! *^^^*

**Tentative Course Plan**

*Please understand that the following is a tentative schedule. Announcements of skipped or added material will be made in class and on the course websites.*

- **Week 1: Introduction**
  - Introduction to data science
  - Introduction to R, RStudio, and R Markdown
- **Week 2-3: Exploratory data analysis and visualization**
  - Visualization with ggplot2
  - Data transformation with dplyr
- **Week 4-5: Supervised learning methods**
  - Linear regression: univariate and multivariate linear regression
  - Classification: nearest neighbor classification, logistic regression
  - Model selection and validation: the validation set approach, cross-validation
- **Week 6: Data management and analytics at large scale**
  - Databases and the relational algebra: dplyr and SQL
  - Data import with readr
  - Data cleaning with tidyr
- **Week 7: Unsupervised learning methods**
  - Clustering: k-means and hierarchical clustering, principle component analysis
- **Week 8: Special topics and final project presentations**