Statistics 694: Data Science Bootcamp (2 units)

Summer 2020

Prof. Suess, Department of Statistics and Biostatistics, CSU East Bay

Lecture:

- MW 7 to 8:40pm, on-line by Zoom

Intructor: Prof. Eric A. Suess **Office:** NSc 319 **Phone:** 510-885-3879 **e-mail:** eric.suess@csueastbay.edu

Office Hours: MW 8:40 to 9:10pm, or by appointment

Class Website: cox.csueastbay.edu/~esuess/stat694

Required Text:

- Wickham, Grolemund R for Data Science
- Wickham, Advanced R
- Ismay, Kim, ModernDive
- Phillips, Yarrr
- Boehmke, Greenwell, Hands-On Machine Learning with R
- Kross, Unix Workbench
- Silge, Robinson, TidyText
- Wickham, Mastering Shiny
- Sievert, Interactive web-based data visualization with R, plotly, and shiny
- Janssens, Data Science at the Command Line
- Allen, Creating APIs in R with Plumber

Reference Texts:

- Baumer, Kaplan, Horton, Modern Data Science with R, CRC Press, 2017.

Further References:

- Bryan, Hester, Happy Git and GitHub for the useR
- VanderPlas, Python Data Science Handbook
- Klok, Nazarathy, Statistics with Julia
- Garrels, Bash Beginners Guide

Material To Be Covered:

This is the zero course in the sequence of Data Science courses offered by the Department of Statistics and Biostatistics for the MS Data Science Concentration. The Data Science courses are specifically for registered students in the MS Statistics program.

The sequence of courses are:

0. Stat. ??? Data Technologies or Data Science Bootcamp or Foundations in R for Data Science
1. Stat. 650 Advanced R for Data Science
2. Stat. 651 Data Visualization
3. Stat. 652 Statistical Learning
4. Stat. 653 Statistical Natural Language Processing
5. Stat. 654 Introduction to Applied Deep Learning

These courses are intended to be taken in order as they build upon each other, but you can discuss taking the courses out of order with instructor approval.

(Alternatively, students who have completed the 650 sequence, this course can be used to reinforce your R Notebook skills, your understanding of data structures in R, expand your experience with the dataset used in the 650 sequence, and to research topics beyond the classes.)

The topics in this course will be drawn from a variety of on-line book, blog posts, YouTube videos, and Github. For each class there will be other supporting refrences. The intension is to develop a collection of resources for the 650 sequence. And to have experience working with the main datasets that will be part of the 650 sequence, to try some and become aware of many modern data technolgies that are commonly used.

The main topics for this class:

- R Notebooks, the file system, your computer system, benchmarking, parallel processing, cluster computing, cloud computing, supper computers, data.frames, data.table, lists, Tidyverse, data, structured data, compressed data, zip and unzip.
- Structured vs unstructured data, AutoEDA, AutoML, h2O, Tensorflow/Keras, SQL, noSQL, SQLite, MongoDB, Small Data, Medium Data, Big Data, Spark.
- Unstructured data, images, music, video, APIs, Github, Slack, Loom.
- R, Python, Julia, Bash, SQL Jupyter Notebooks, Anaconda, Kaggle.
- Unstructured data, text, e-books, a-books, many small files into one dataset, break one Big file into smaller files, disk.frame.
- Other Topics: Shiny, R packages, using R to set up an API, Linux, BSD, Docker, Amazon AWS, Microsoft Azure, Google GCP, IBM Watson, Digital Ocean, Edge Computing, RasperberryPi, FitBit, iPhones, Android Phones.

The datasets for this class:

- iris, mtcars, diamonds, gapmider, nycflights, NHANES, MNIST, Boston Housing, cat-and-dogs
- bikeshare data, Lyft Bay Wheels, NYC Taxi data, Airline On-Time Statistics, Chicago Crime data, Fannie Mae
- Twiter, Gutenberg, NASA, Genius, Spotify, Arxiv, Internet Archive.

This course is the foundation course for the Data Science courses that have been developed for the MS Statistics Concentation in Data Science.

The sequence of courses are:

1. Stat. 650 Advanced R for Data Science
2. Stat. 651 Data Visualization
3. Stat. 652 Statistical Learning
4. Stat. 653 Statistical Natural Language Processing
5. Stat. 654 Introduction to Applied Deep Learning
6. Stat. 694 Research in Statistics and Biostatistics - Innovation

These courses are intended to be taken in order as they build upon each other, but you can discuss taking the courses out of order with instructor approval.

Homework: A list will also be on the website. Homework will be assigned weekly. Homework will be “due”" on Mondays, which means you should complete the homework and come to class prepared to ask questions. Homework will be “collected”" though Blackboard and needs to be submitted by Wednesday of the week the homework is due.

Project: Idea, Plan & Development, Data Product, Presentation/Video

Grading:

- Homework 50%
- Project 50%

Policy on Make-up Exams: You are expected to take the quizzes and exams at the scheduled times. In case of genuine emergency, illness or hardship, for which you can present written documentation I may agree to arrange for a make-up exam. Make-up exams must always be arranged BEFORE the regular exam is given and always take place AFTER the regular exam. Quizzes may not be made up!

Statistics 650 SLOs

Student Learning Outcomes (SLO's):

Students graduating with an M.S. in Statistics from Cal State East Bay will be able to:

1. Apply statistical methodologies, including a) descriptive statistics and graphical displays, b) probability models for uncertainty, stochastic processes, and distribution theory, c) hypothesis testing and confidence intervals, d) ANOVA and regression models (including linear, and multiple linear) and analysis of residuals from models and trends at the Master's level.
2. Derive basic theory underlying these methodologies.
3. Model practical problems for solutions using these methodologies.
4. Produce relevant computer output using standard statistical software and interpret the results appropriately.
5. Communicate statistical concepts and analytical results clearly and appropriately to others; and,
6. Employ theory, concepts, and terminology at a level that supports lifelong learning of related methodologies.