Teaching Statistical Programming and Data Analysis with “Data Hospital”

Earvin Balderama
California State University, Fresno
Department of Mathematics

earvin@csufresno.edu

Statistics courses

- MATH 11: Elementary Statistics
- MATH 101: Statistical Methods
- ***MATH 105*: Statistical Programming and Data Analysis**
- MATH 106. Applied Linear Statistical Models
- MATH 107: Mathematical Statistics
- MATH 108: Advanced Mathematical Statistics
- MATH 109: Applied Probability
- MATH 137: Exploring Statistics
- MATH 191T: Research Seminar in Statistics (1 unit)

Required course materials

Software

Textbooks

- Hands-On Programming with R. *Garrett Grolemund (2014).*
- Introduction to Data Science. *Rafael Irizarry (2020).*
- R for Data Science. *Garrett Grolemund & Hadley Wickham (2017).*

Why  ?

- Created specifically for **statistical computing, graphics and data analysis**
- Strong and widespread community of users and developers
- Latest methods contributed by top researchers in the field
- Used by many other disciplines outside of statistics
- Relatively easy language to learn
- Easy to find help online
- Free and open source

Why  Studio®?

- User-friendly **integrated development environment.**
- Workspace organized in a single window.
- Run other coding languages: Python, C++, SQL, D3, etc.
- Create dynamic documents: Rmarkdown, knitr, Sweave
  - LaTeX, Word, html
- Integration with Github.
- Desktop or Server formats.
- Also free (but have commercial versions as well).

Statistical Programming and Data Analysis

Fall 2019: Face-to-face

- 14 students
- Grading
  - 20% --- Attendance/Participation
  - 15% --- Challenges
  - **25% --- Mini Projects (Data Hospital)**
  - 40% --- Final Project
- Students worked in the same team for all Mini Projects.

Fall 2020: Virtual synchronous

- 11 students
- Grading
  - 20% --- Discussion Board
  - 20% --- Challenges
  - **20% --- Mini Projects (Data Hospital)**
  - 40% --- Final Project
- Students randomly placed in teams for each Mini Project.

Data Hospital

Team Mini-Projects,
consisting of R coding problems and/or data analysis

Inspiration

In an academic hospital, e.g., Stanford Hospital, UCLA Medical Center, UCSF, etc., (or think *House*, or *Scrubs*, or *Grey's Anatomy*), **attending physicians** teach and train **teams of residents** ("doctors-in-training") as they make their "**rounds**" from patient to patient to evaluate, discuss, and treat their health issues.

Inspiration

In an academic hospital, e.g., Stanford Hospital, UCLA Medical Center, UCSF, etc., (or think *House*, or *Scrubs*, or *Grey's Anatomy*), **attending physicians** teach and train **teams of residents** ("doctors-in-training") as they make their "**rounds**" from patient to patient to evaluate, discuss, and treat their health issues.

- Residents gather and synthesize patient information, and present diagnosis and treatment plan.
- The Attending acts as a facilitator, offers feedback and guidance.
- Active learning experience: Residents **learn by doing**.
    - Practice **communication** skills through live presentations.
    - Receive **feedback** from attending and fellow residents.
    - Work as a **team**.

Data Hospital

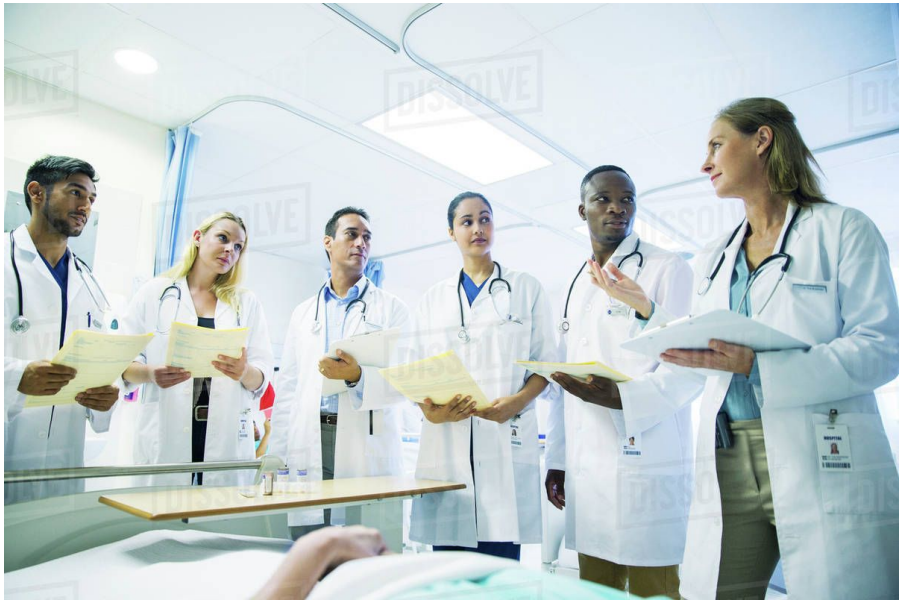
I will be the “**Attending**” and students will take the role of “**resident**,” training to become a data analyst while treating the data as the “**patient**.”

- The Attending (me)
- Resident teams (you)
  - Chief Resident
  - Interns

Two sets of *Rounds*

Presenting your data to the class

1. Assessment and Planning
2. Results and Follow-up



The data will be your "patient"!

Team mini-projects

1. Your team will be assigned one or more data sets
2. Determine how you want to proceed with the analysis of such data
3. **Present the data to the class during the “Assessment & Planning” rounds**
4. Consider feedback from your fellow *residents* and *Attending*
5. Use R to “treat” the data
6. **Present your results and conclusions at the “Results & Follow-Up” rounds.**
7. Post results in discussion thread, and provide feedback to other teams.

Typical week

Monday: Teams form and get assigned data.

Wednesday: **Initial Assessment and Planning** Rounds.

- Learn history/background information of your data.
- Identify research question(s) and purpose of study.
- Identify type of data structure; observational units and variable types.
- Any interesting features of the data set?
- Convey plans for exploratory data analysis.

Friday: **Results and Follow-up** Rounds.

- Brief summary of data.
- Plots and graphics.
- Methods; functions and packages used.
- Results and conclusions.
- Further research questions?
- Any changes in diagnosis?
- Unexpected outcomes?

Patients

Mini Project #1

- `datasets::airquality`
- `datasets::iris`
- `datasets::mtcars`
- `datasets::USArrests`
- `datasets::toothgrowth`

Mini Project #2

- `dslabs::greenhouse_gases`
- `dslabs::historic_co2`
- `dslabs::temp_carbon`
- `dslabs::stars`
- `dslabs::gapminder`

Mini Project #3

- SMS Spam Collection
- US Election 2020 Tweets

Mini Project #4

- County Health Rankings

Mini Project #5

- 2019 Data Science Bowl

Basic R programming, functions, data structures, data visualization, data transformation

Mini Project #1

- `datasets::airquality`
- `datasets::iris`
- `datasets::mtcars`
- `datasets::USArrests`
- `datasets::toothgrowth`

Mini Project #2

- `dslabs::greenhouse_gases`
- `dslabs::historic_co2`
- `dslabs::temp_carbon`
- `dslabs::stars`
- `dslabs::gapminder`

Mini Project #3

- SMS Spam Collection
- US Election 2020 Tweets

Mini Project #4

- County Health Rankings

Mini Project #5

- 2019 Data Science Bowl

Basic R programming, functions, data structures, data visualization, data transformation

Mini Project #1

- `datasets::airquality`
- `datasets::iris`
- `datasets::mtcars`
- `datasets::USArrests`
- `datasets::toothgrowth`

Data Wrangling, tidying data, relational data

Mini Project #2

- `dslabs::greenhouse_gases`
- `dslabs::historic_co2`
- `dslabs::temp_carbon`
- `dslabs::stars`
- `dslabs::gapminder`

Mini Project #3

- SMS Spam Collection
- US Election 2020 Tweets

Mini Project #4

- County Health Rankings

Mini Project #5

- 2019 Data Science Bowl

Basic R programming, functions, data structures, data visualization, data transformation

Mini Project #1

- `datasets::airquality`
- `datasets::iris`
- `datasets::mtcars`
- `datasets::USArrests`
- `datasets::toothgrowth`

Data Wrangling, tidying data, relational data

Mini Project #2

- `dslabs::greenhouse_gases`
- `dslabs::historic_co2`
- `dslabs::temp_carbon`
- `dslabs::stars`
- `dslabs::gapminder`

String processing, regular expressions

Mini Project #3

- SMS Spam Collection
- US Election 2020 Tweets

Mini Project #4

- County Health Rankings

Mini Project #5

- 2019 Data Science Bowl

Basic R programming, functions, data structures, data visualization, data transformation

Mini Project #1

- `datasets::airquality`
- `datasets::iris`
- `datasets::mtcars`
- `datasets::USArrests`
- `datasets::toothgrowth`

Data Wrangling, tidying data, relational data

Mini Project #2

- `dslabs::greenhouse_gases`
- `dslabs::historic_co2`
- `dslabs::temp_carbon`
- `dslabs::stars`
- `dslabs::gapminder`

String processing, regular expressions

Mini Project #3

- SMS Spam Collection
- US Election 2020 Tweets

Statistical modeling, regression, linear models, disease mapping

Mini Project #4

- County Health Rankings

Mini Project #5

- 2019 Data Science Bowl

Observations

- Students went beyond the assignment and learned some background information about the dataset prior to analyzing the data
  - Results and conclusions became more meaningful
- Lessons learned carried over to final projects
- Increased class participation

What's the ***best*** thing about this course?

- *"I really liked the group work because it allowed us to bounce ideas off each other."*
- *"...the mini-hospitals we did as a class were fun and provided a great way to interact with students I might have never had the chance to meet."*
- *"I really enjoyed that homework was project based. It made me want to learn while not feeling like the work was overwhelming."*
- *"It was more about us learning than being perfect."*

What's the ***worst*** thing about this course?

- *"The worst thing about this MATH 105 course is that it is taught online."*
- *"Really wish it was in person"*
- *"The worse thing was the constant random groups. It would of been nice to have a group and stick with them for half the semester then maybe switch again."*
- *"The worst things about this class were that I sometimes felt rushed to answer breakout room questions and I did 3 out of 4 group projects either entirely or almost entirely by myself."*

Other student comments

- *“The day we analyzed candidate tweets was the best because it was really **relevant** since the campaign for presidency was happening at that time. So it would be cool to do more relevant topics in the future.”*

- *“This class overall has been one of my favorite classes in all of undergrad.”*

- *“... my only career choice ... has been a sports statistician, but thanks to this course, I have now **broadened** that to a data analyst as well.”*

Thank you!

Teaching Statistical Programming and Data Analysis with “Data Hospital”

Earvin Balderama
California State University, Fresno
Department of Mathematics

earvin@csufresno.edu