Week 1 Review Session

Public Health 142 • July 8, 2021

GSI: Chandler Beon Slides: Saher Daredia

Announcements

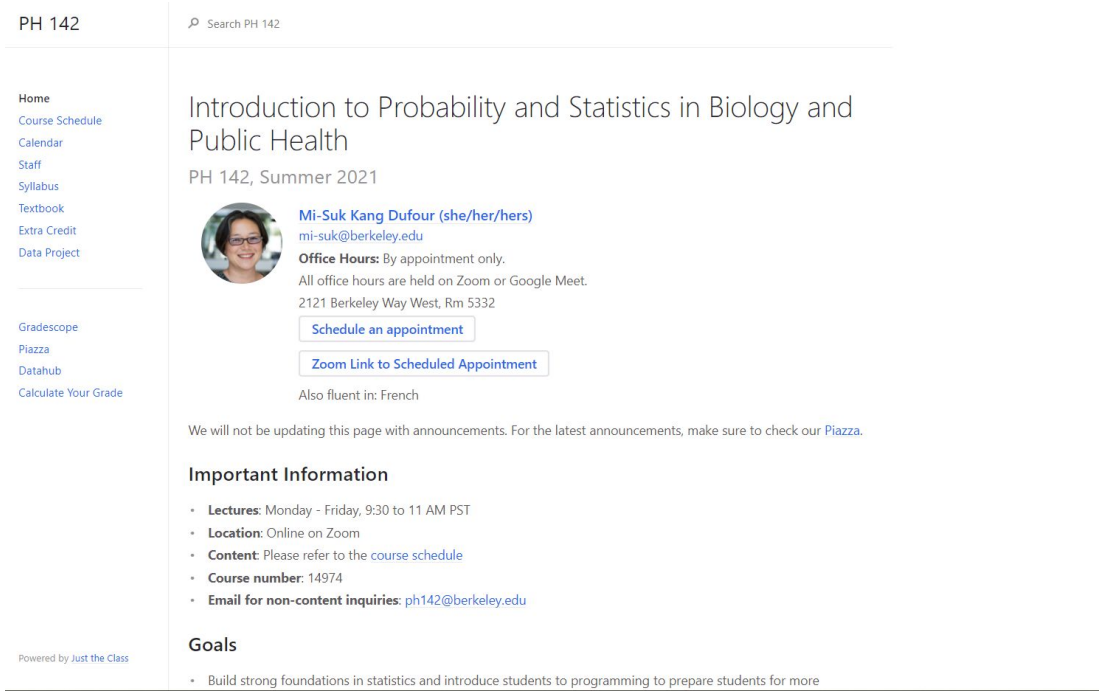
- **Homework 1/Lab 1**
  - Lab 1 due Friday, July 9th @ 10 PM PST
  - Homework 1 Solutions released Friday, July 9th
  - Need help? Piazza, Office Hours, and Homework Parties!
- **Homework 2/Lab 2**
  - Released tomorrow!
  - Due Tuesday, July 13th @ 10 PM PST
- **Midterm 1**
  - Coming soon — Friday July 16th!
- **Data Project**
  - Project specs available on https://ph142-ucb.github.io/su21/data-proj/
  - Project groups / GSI Assignments have been released!

Objectives

1. Summarize key course technologies and policies
2. Review material from lectures #1-2
    - PPDAC Approach
    - Visualizing Categorical Data
    - R Basics

Key Course Technologies

Course Website: https://ph142-ucb.github.io/su21/



Key Course Technologies

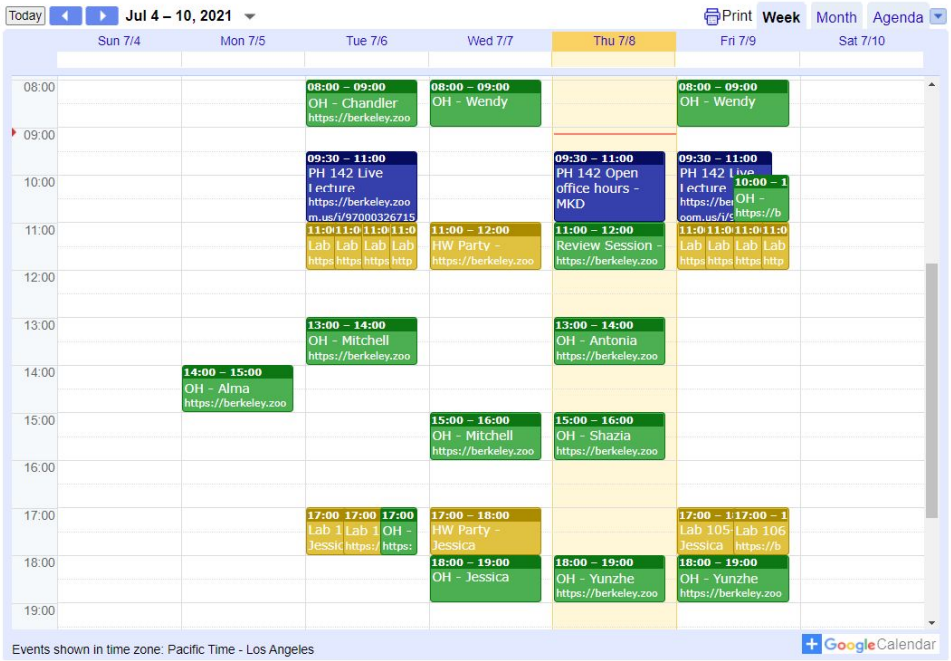
Accessing Slides and Recordings

Week 1

Jul 5:	**No Class: Independence Day**	
Jul 6:	LECTURE 1 LIVE Intro to PH142, Datahub and PPDAC; Visualizing Distributions for One Variable (recording) PARTICIPATION 'Homework' 0 (PDF) LAB 1 on Datahub (Recording) (Due July 9) HOMEWORK 1 on Datahub (Sol. released July 9) QUIZ 1 on Gradescope	Ch. 1 & 2
Jul 7:	LECTURE 2 Working with Data and Numerically Summarizing Spread and Central Tendency (Handout) (Recordings)	Ch. 3
Jul 8:	LECTURE 3 Exploring Relationships Between Two Variables (recording) REVIEW Week 1 Review QUIZ 2	Ch. 4
Jul 9:	LECTURE 4 LIVE Introduction to Regression LAB 2 (Recording)(Due July 13) HOMEWORK 2 (Sol. released July 13)	Ch. 5 & 6

Key Course Technologies

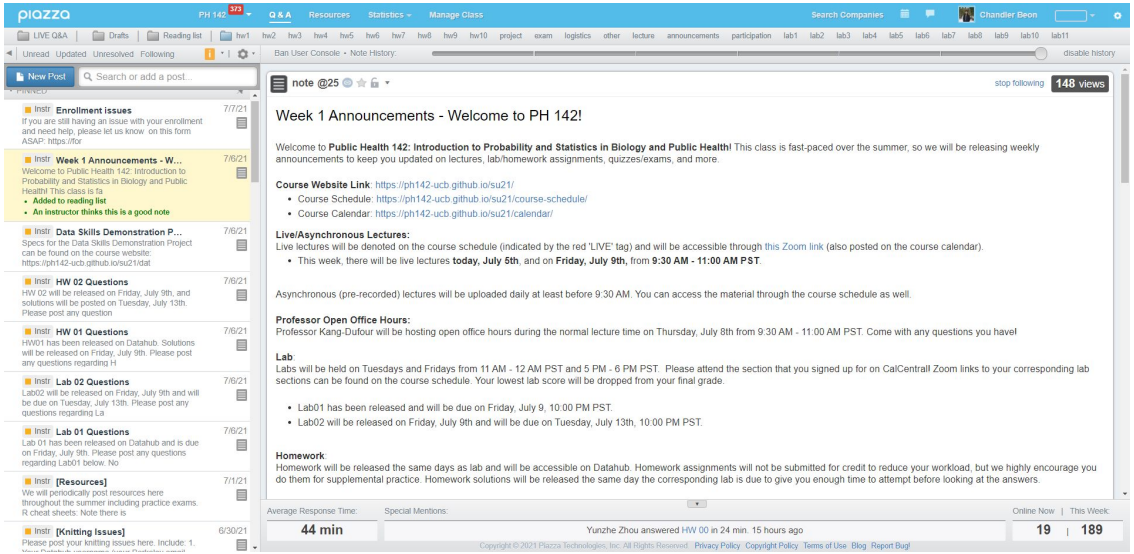
Course Calendar



Key Course Technologies

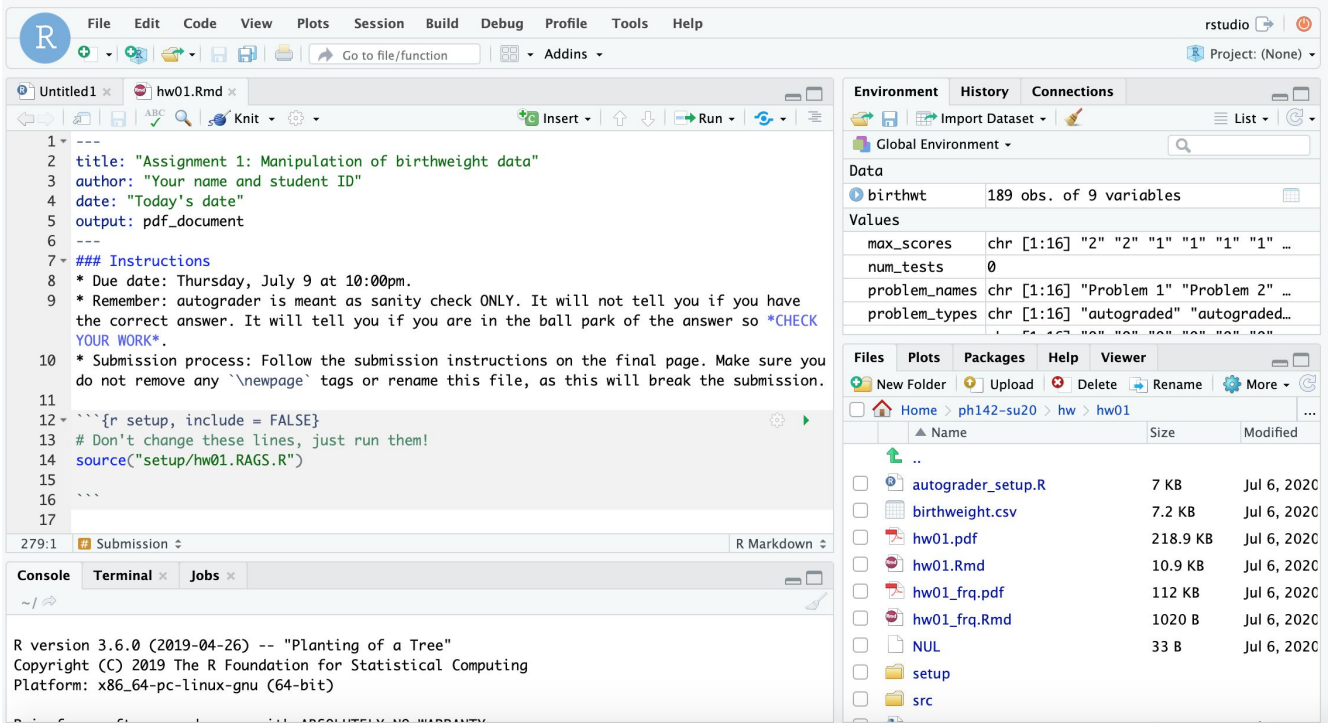
Gradescope: Make sure to enroll using the entry code **P5352G**

Piazza



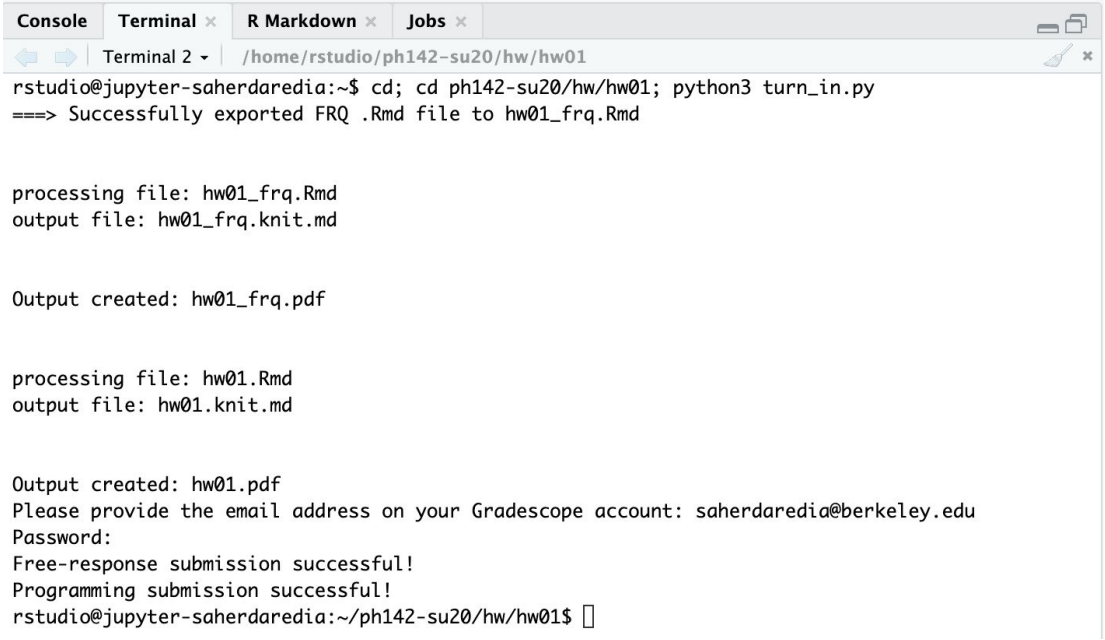
Key Course Technologies

Datahub



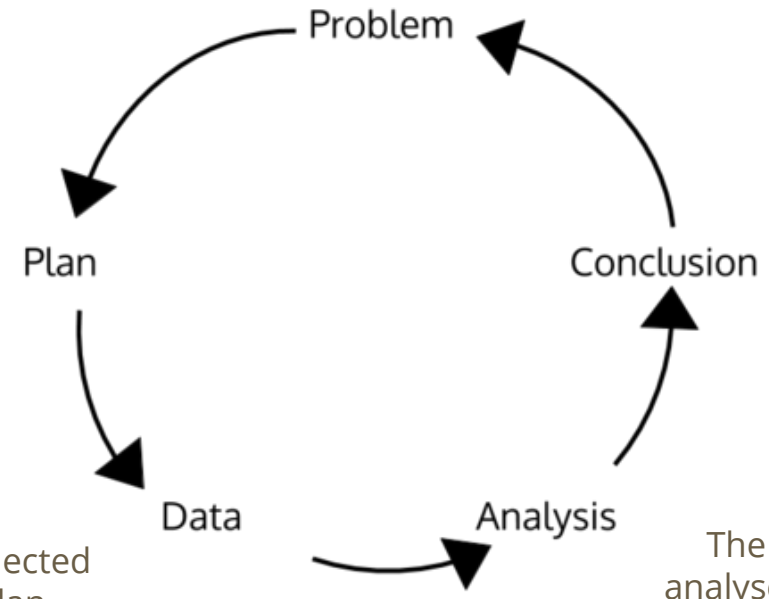
Key Course Technologies

Datahub: Knit BEFORE typing code in “Terminal” to submit labs/HW!



PPDAC Approach

Clear statement of what we are trying to achieve



The procedures we use to carry out the study

Conclusions are drawn about what has been learned about answering the Problem

The data which is collected according to the Plan

The data is summarized and analysed to answer the questions posed by the Problem

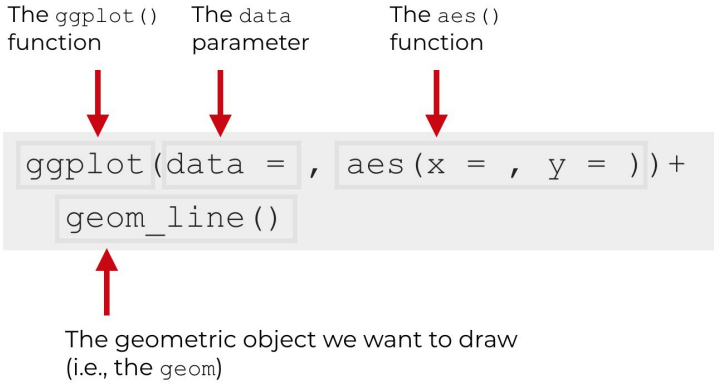
Visualization of Categorical Data: ggplot2

1. **Install and load the ggplot2 package**
   - `install.packages(ggplot2)`
   - `library(ggplot2)`

2. **Specify the data set and what goes on the x and y axes**

3. **Use a 'geom_' function to create our chart**
   - Many different types including: `geom_bar`, `geom_histogram`, `geom_point`, `geom_line`

4. **Change the style of your plot**
   - `labs()` function = specify the main title, axis titles, caption
   - `theme()` function = change title sizes, fonts, and positions



Types of Variables

Categorical

A variable that has grouping levels

Nominal

No underlying order or rank

Eye color

Ordinal

Can be ordered or ranked

Committee position (President, VP, Secretary etc.)

Quantitative

A numeric variable that you can perform mathematical operations on

Discrete

Can be counted

Number of reviews a restaurant has on Yelp

Continuous

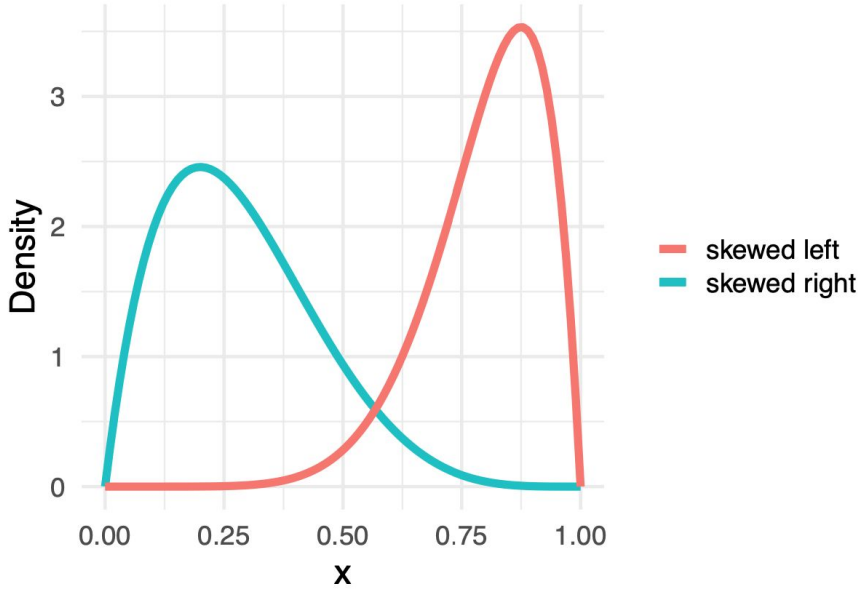
Can be measured precisely, with a ruler or scale

Insurance charges $

R Basics

- **Library** → package of R functions (e.g. dplyr)
  - Remember: you must load the required libraries every time you start a new session/file
  - Ex) `library(dplyr)`

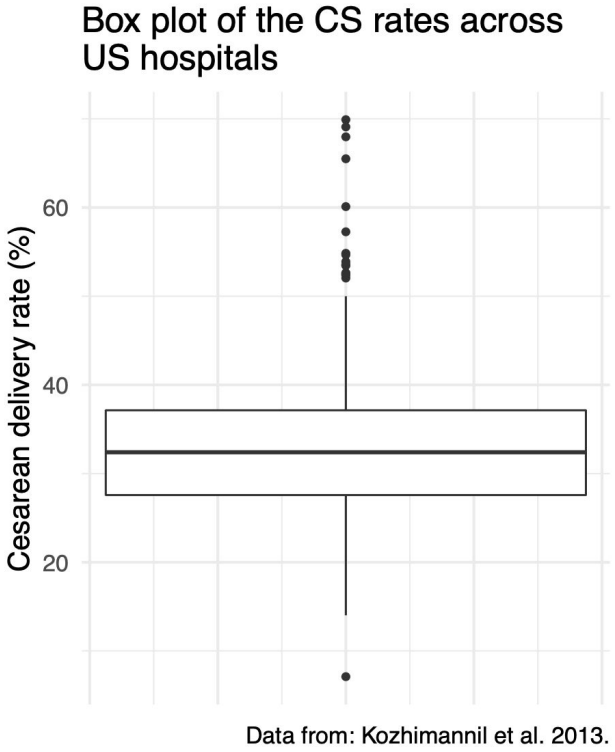
- **Reading in a CSV file**
  - ```
    library(readr)
    your_data <- read_csv("name_of_dataset.csv")
    ```

- **Four useful functions to describe your dataset**
  - `head(your_data)`: Shows the first six rows of the supplied dataset
  - `dim(your_data)` **:** Provides the number of rows by the number of columns
  - `names(your_data)`: Lists the variable names of the columns in the dataset
  - `str(your_data)`: Summarizes the above information and more
    ```

dplyr functions for data manipulation

You must load the dplyr library to access these functions

- rename() → renames variables (columns)
  - `new_dataset <- old_dataset %>% rename(new_name = old_name)`
  - Alternatively: `new_dataset <- rename(old_dataset, new_name = old_name)`
- select() → subsets variables (columns)
  - `smaller_data <- old_data %>% select(variable1, variable2, variable3)`
  - `smaller_data <- select(old_data, variable1, variable2, variable3)`
  - `smaller_data <- select(old_data, variable1:variable3)`
  - To keep all variables other than variable1: `smaller_data <- old_data %>% select(- variable1)`
- arrange() → orders observations (rows) by a certain variable (column) or variables (columns)
  - `lake_data %>% arrange(ph)`
  - `lake_data %>% arrange(age_data, ph)`

dplyr functions (cont.)

- filter() → selects a subset of rows by certain conditions
  - If we want condition A AND condition B to be satisfied, use **,** or **&**
  - If we want condition A OR condition B to be satisfied, use **|** or **%in%**
  - `lake_data %>% filter(age_data == "recent")`
  - `lake_data %>% filter(lakes %in% c("Alligator", "Blue Cypress"))`
  - `lake_data %>% filter(ph > 6 | chlorophyll > 30)`
- mutate() → creates new variables
  - `lake_data_new <- lake_data %>% mutate(actual_fish_sample = number_fish_sampled * 100)`
- group_by() → groups the data by a categorical variable
  - `lake_data %>% group_by(age_data)%>% summarize(mean_ph = mean(ph))`
- summarize() → applies summary functions to calculate statistics
  - `lake_data %>% summarize(mean_ph = mean(ph), sd_ph = sd(ph))`

Measures of Central Tendency

- Mean and median are approximately equal when...
  1. Distribution is symmetric
  2. Data has one peak
  3. There are no outliers
- Outliers → large effect on the mean
- Skewed data: mean ≠ median
  - Skewed right: mean > median
  - Skewed left: mean < median



Measures of Spread

- Range = max - min
- IQR = Q3 - Q1
  - Five number summary in R!
    - ```
      CS_dat %>% summarize(min = min(cs_rate),
      Q1 = quantile(cs_rate, 0.25), median = median(cs_rate),
      Q3 = quantile(cs_rate, 0.75), max = max(cs_rate))
      ```
- Sample variance (s^2)
- Sample standard deviation (s)
  - `CS_dat %>% summarize(cs_sd = sd(cs_rate), cs_var = var(cs_rate))`

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
      ```

Box Plots

- Center line → median
- Top of box → Q3
- Bottom of box → Q1
- Top of top whisker → max value or highest point that is below Q3 + 1.5*IQR
- Bottom of bottom whisker → min value or lowest point that is above Q1 - 1.5*IQR
- Data points above and below whiskers → outliers

```
ggplot(CS_dat, aes(y = cs_rate)) +
geom_boxplot() +
ylab("Cesarean delivery rate (%)") +
labs(title = "Box plot of the CS rates across US hospitals",
    caption = "Data from: Kozhimannil et al. 2013.") +
theme_minimal(base_size = 15) +
scale_x_continuous(labels = NULL) # removes the labels from the x axis
```



Common errors

1. Do not name two code chunks the same thing
2. Use the variable names that are listed in the instructions
3. If your data isn't running, try reloading your past code chunks first
4. If you want to see the output of your data, just retype the name of your variable in a new line within the same code chunk and run again

Questions?