PH142 Review Session: Week 1

PH142 GSI team
July 6th, 2023

Announcement

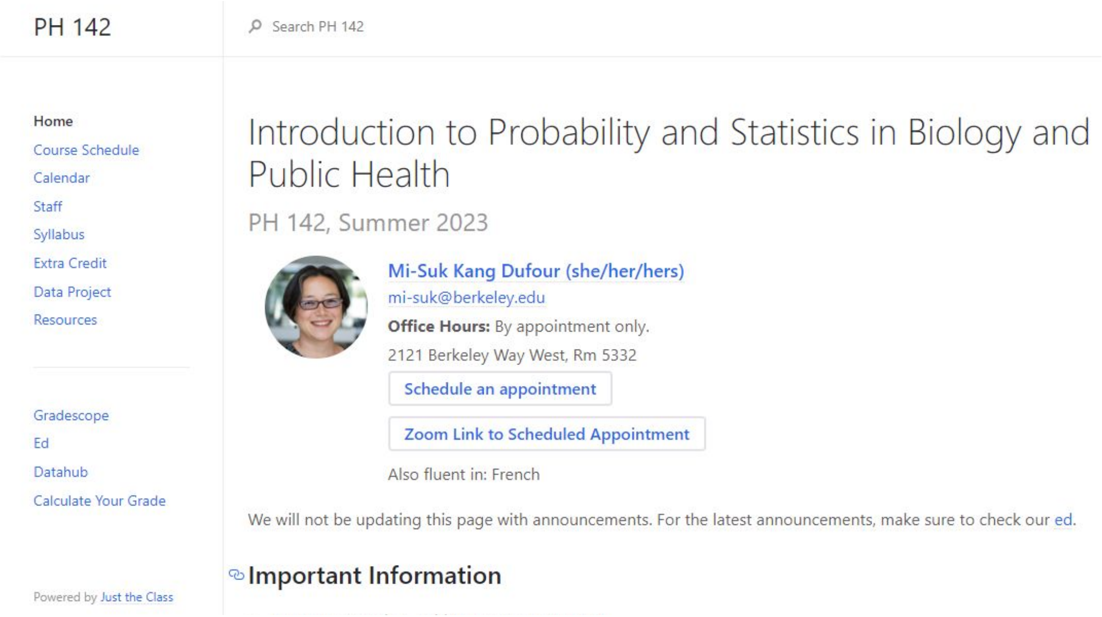
- Deadlines
  - Lecture quizzes on Gradescope
  - Lab 1: due 07/07 on Datahub
  - Lab 2: due 07/10 on Datahub
  - Homework 1&2: not turned in - for you to use as practice
- Midterm I: released 07/14 - available until 5pm 7/15
- Data Project
  - Check out instructions on https://ph142-ucb.github.io/su23/data-proj/.
  - Part I due on 07/17, 10pm PST

Objectives

- Summarize key course technologies, resources, and policies.
- Review materials from lectures 1-2.
    - PPDAC Approach
    - Categorical Data Visualization
    - Intro to R

Key Technologies & Resources

Course website: https://ph142-ucb.github.io/su23/



Introduction to Probability and Statistics in Biology and Public Health

PH 142, Summer 2023

Mi-Suk Kang Dufour (she/her/hers)
mi-suk@berkeley.edu
Office Hours: By appointment only.
2121 Berkeley Way West, Rm 5332

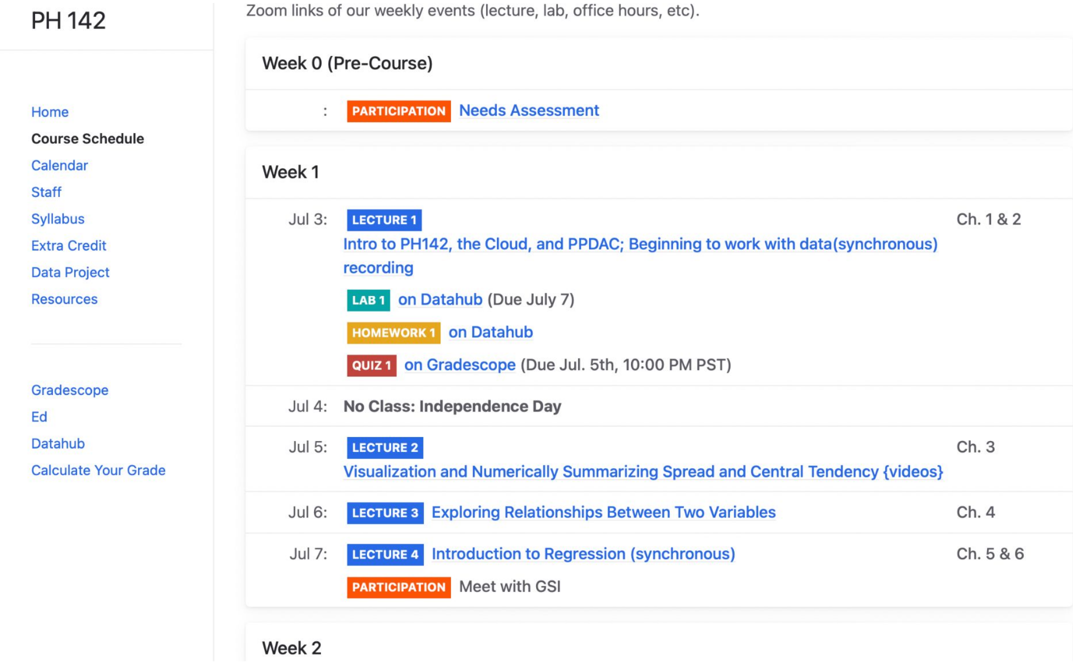
Also fluent in: French

We will not be updating this page with announcements. For the latest announcements, make sure to check our ed.

Important Information

Key Technologies & Resources

Accessing slides and recordings:



PH 142

- Home
- **Course Schedule**
- Calendar
- Staff
- Syllabus
- Extra Credit
- Data Project
- Resources

- Gradescope
- Ed
- Datahub
- Calculate Your Grade

Zoom links of our weekly events (lecture, lab, office hours, etc).

Week 0 (Pre-Course)

: PARTICIPATION Needs Assessment

Week 1

Jul 3: LECTURE 1 Intro to PH142, the Cloud, and PPDAC; Beginning to work with data(synchronous) recording — Ch. 1 & 2

LAB 1 on Datahub (Due July 7)

HOMEWORK 1 on Datahub

QUIZ 1 on Gradescope (Due Jul. 5th, 10:00 PM PST)

Jul 4: **No Class: Independence Day**

Jul 5: LECTURE 2 Visualization and Numerically Summarizing Spread and Central Tendency {videos} — Ch. 3

Jul 6: LECTURE 3 Exploring Relationships Between Two Variables — Ch. 4

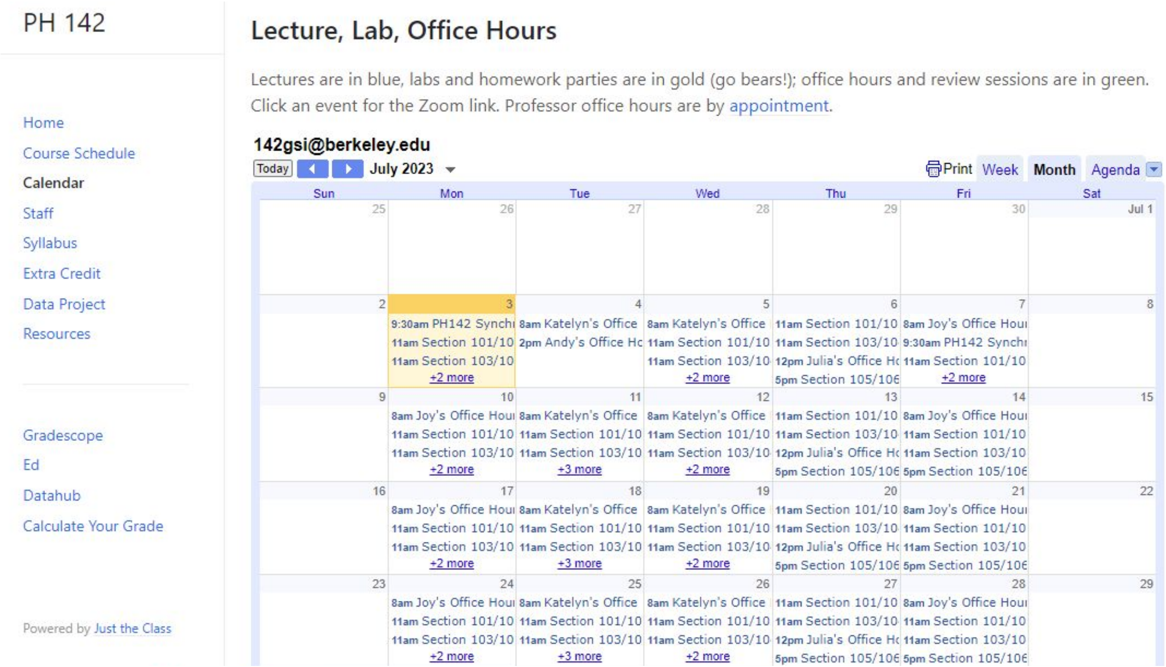
Jul 7: LECTURE 4 Introduction to Regression (synchronous) — Ch. 5 & 6

PARTICIPATION Meet with GSI

Week 2

Key Technologies & Resources

Course Calendar:



PH 142

- Home
- Course Schedule
- Calendar
- Staff
- Syllabus
- Extra Credit
- Data Project
- Resources

- Gradescope
- Ed
- Datahub
- Calculate Your Grade

Powered by Just the Class

Lecture, Lab, Office Hours

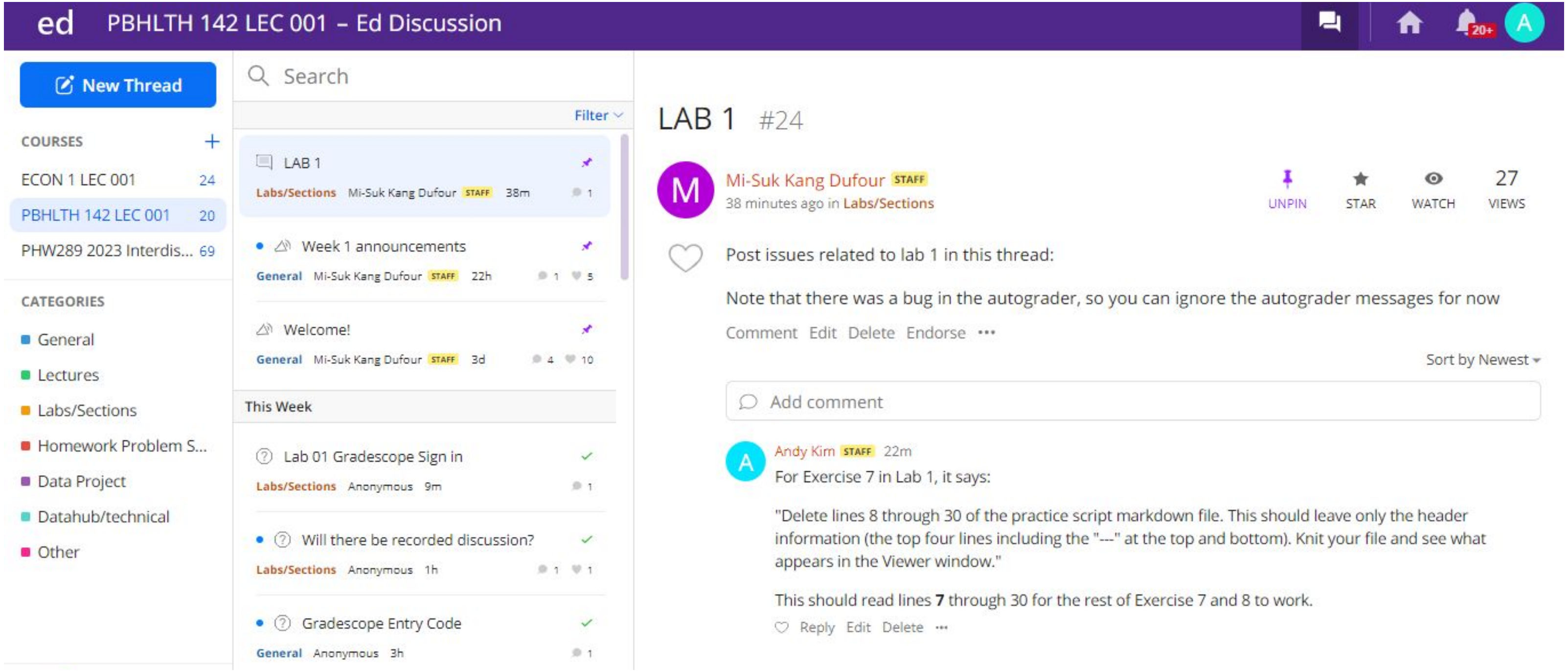
Lectures are in blue, labs and homework parties are in gold (go bears!); office hours and review sessions are in green. Click an event for the Zoom link. Professor office hours are by appointment.

142gsi@berkeley.edu

Sun	Mon	Tue	Wed	Thu	Fri	Sat
25	26	27	28	29	30	Jul 1
2	3 9:30am PH142 Synchı 11am Section 101/10 11am Section 103/10 +2 more	4 8am Katelyn's Office 2pm Andy's Office Hc	5 8am Katelyn's Office 11am Section 101/10 11am Section 103/10 +2 more	6 11am Section 101/10 11am Section 103/10 12pm Julia's Office Hc 5pm Section 105/106	7 8am Joy's Office Houı 9:30am PH142 Synchı 11am Section 101/10 +2 more	8
9	10 8am Joy's Office Houı 11am Section 101/10 11am Section 103/10 +2 more	11 8am Katelyn's Office 11am Section 101/10 11am Section 103/10 +3 more	12 8am Katelyn's Office 11am Section 101/10 11am Section 103/10 +2 more	13 11am Section 101/10 11am Section 103/10 12pm Julia's Office Hc 5pm Section 105/106	14 8am Joy's Office Houı 11am Section 101/10 11am Section 103/10 5pm Section 105/106	15
16	17 8am Joy's Office Houı 11am Section 101/10 11am Section 103/10 +2 more	18 8am Katelyn's Office 11am Section 101/10 11am Section 103/10 +3 more	19 8am Katelyn's Office 11am Section 101/10 11am Section 103/10 +2 more	20 11am Section 101/10 11am Section 103/10 12pm Julia's Office Hc 5pm Section 105/106	21 8am Joy's Office Houı 11am Section 101/10 11am Section 103/10 5pm Section 105/106	22
23	24 8am Joy's Office Houı 11am Section 101/10 11am Section 103/10 +2 more	25 8am Katelyn's Office 11am Section 101/10 11am Section 103/10 +3 more	26 8am Katelyn's Office 11am Section 101/10 11am Section 103/10 +2 more	27 11am Section 101/10 11am Section 103/10 12pm Julia's Office Hc 5pm Section 105/106	28 8am Joy's Office Houı 11am Section 101/10 11am Section 103/10 5pm Section 105/106	29

Key Technologies & Resources

Ed discussion: edstem.org



ed
PBHLTH 142 LEC 001 – Ed Discussion
20+
A
New Thread
COURSES
+
ECON 1 LEC 001 24
PBHLTH 142 LEC 001 20
PHW289 2023 Interdis... 69
CATEGORIES
General
Lectures
Labs/Sections
Homework Problem S...
Data Project
Datahub/technical
Other
Search
Filter
LAB 1
Labs/Sections Mi-Suk Kang Dufour STAFF 38m 1
Week 1 announcements
General Mi-Suk Kang Dufour STAFF 22h 1 5
Welcome!
General Mi-Suk Kang Dufour STAFF 3d 4 10
This Week
Lab 01 Gradescope Sign in
Labs/Sections Anonymous 9m 1
Will there be recorded discussion?
Labs/Sections Anonymous 1h 1 1
Gradescope Entry Code
General Anonymous 3h 1
UNPIN
STAR
WATCH
27
VIEWS
Comment Edit Delete Endorse •••
Sort by Newest
Add comment
Reply Edit Delete •••

LAB 1 #24

M Mi-Suk Kang Dufour STAFF
38 minutes ago in Labs/Sections

Post issues related to lab 1 in this thread:

Note that there was a bug in the autograder, so you can ignore the autograder messages for now

A Andy Kim STAFF 22m

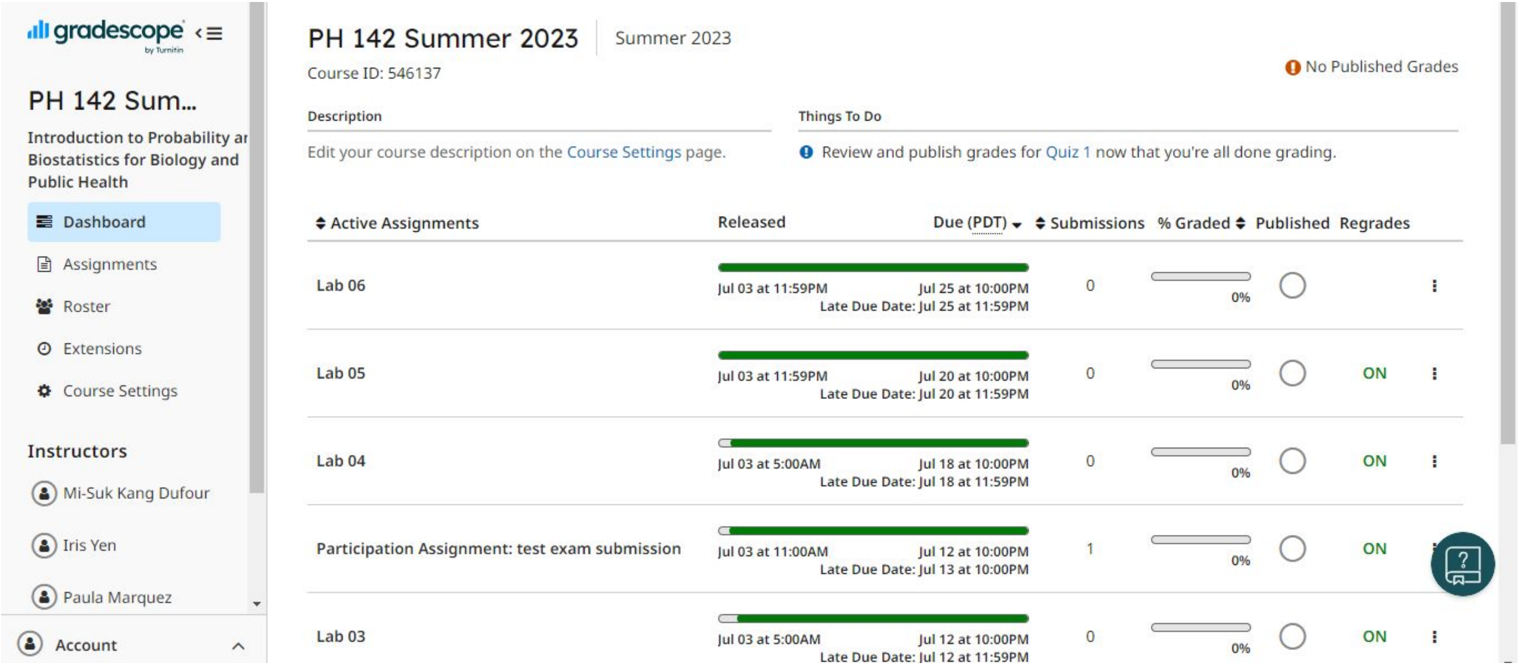
For Exercise 7 in Lab 1, it says:

"Delete lines 8 through 30 of the practice script markdown file. This should leave only the header information (the top four lines including the "---" at the top and bottom). Knit your file and see what appears in the Viewer window."

This should read lines **7** through 30 for the rest of Exercise 7 and 8 to work.

Key Technologies & Resources

Gradescope: gradescope.com



PH 142 Sum...

Introduction to Probability ar
Biostatistics for Biology and
Public Health

Instructors

- Mi-Suk Kang Dufour
- Iris Yen
- Paula Marquez

PH 142 Summer 2023 | Summer 2023

Course ID: 546137

No Published Grades

Description

Edit your course description on the Course Settings page.

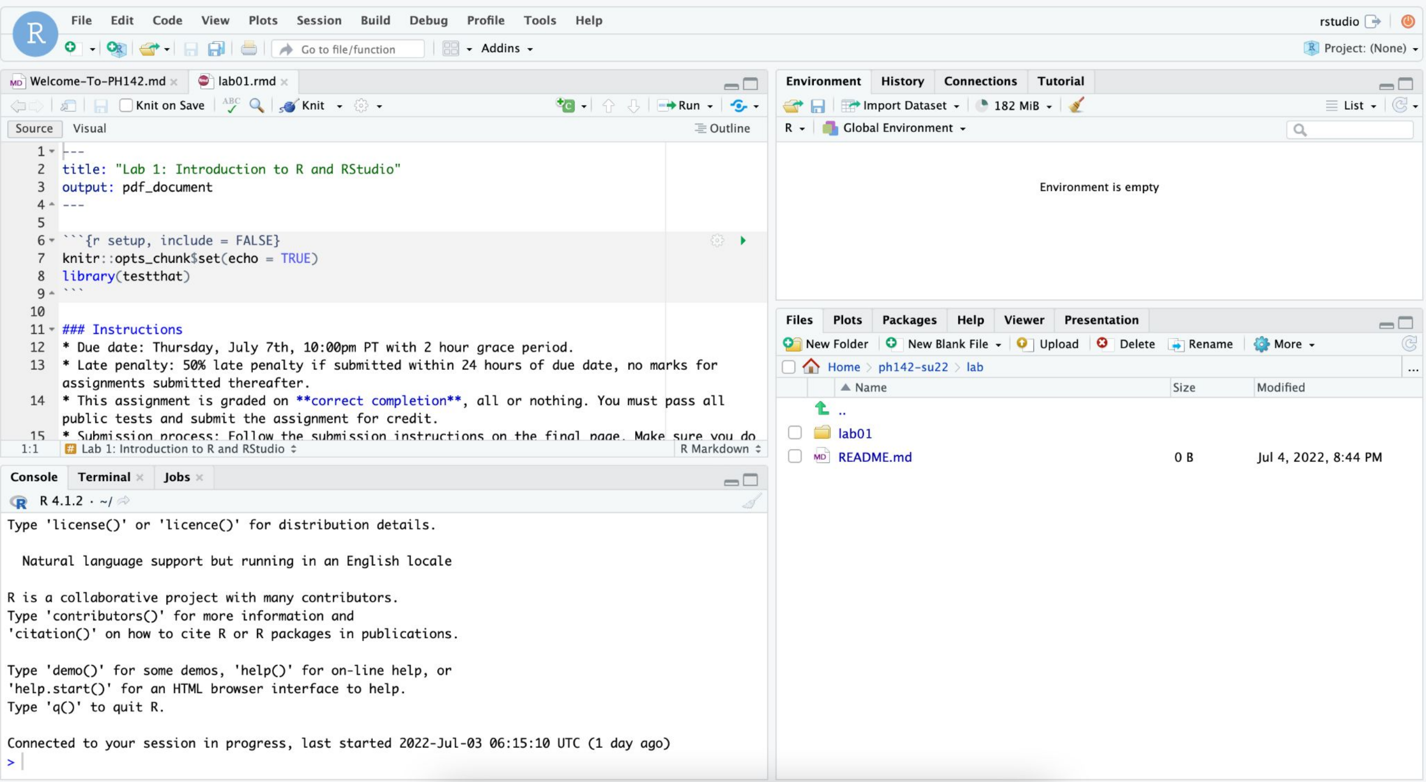
Things To Do

Review and publish grades for Quiz 1 now that you're all done grading.

Active Assignments	Released	Due (PDT)	Submissions	% Graded	Published	Regrades
Lab 06	Jul 03 at 11:59PM	Jul 25 at 10:00PM Late Due Date: Jul 25 at 11:59PM	0	0%	○	
Lab 05	Jul 03 at 11:59PM	Jul 20 at 10:00PM Late Due Date: Jul 20 at 11:59PM	0	0%	○	ON
Lab 04	Jul 03 at 5:00AM	Jul 18 at 10:00PM Late Due Date: Jul 18 at 11:59PM	0	0%	○	ON
Participation Assignment: test exam submission	Jul 03 at 11:00AM	Jul 12 at 10:00PM Late Due Date: Jul 13 at 10:00PM	1	0%	○	ON
Lab 03	Jul 03 at 5:00AM	Jul 12 at 10:00PM Late Due Date: Jul 12 at 11:59PM	0	0%	○	ON

Key Technologies & Resources

Datahub:



~~~
---
title: "Lab 1: Introduction to R and RStudio"
output: pdf_document
---

```{r setup, include = FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(testthat)
```

### Instructions
* Due date: Thursday, July 7th, 10:00pm PT with 2 hour grace period.
* Late penalty: 50% late penalty if submitted within 24 hours of due date, no marks for
assignments submitted thereafter.
* This assignment is graded on **correct completion**, all or nothing. You must pass all
public tests and submit the assignment for credit.
* Submission process: Follow the submission instructions on the final page. Make sure you do
~~~

```
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Connected to your session in progress, last started 2022-Jul-03 06:15:10 UTC (1 day ago)
>
```

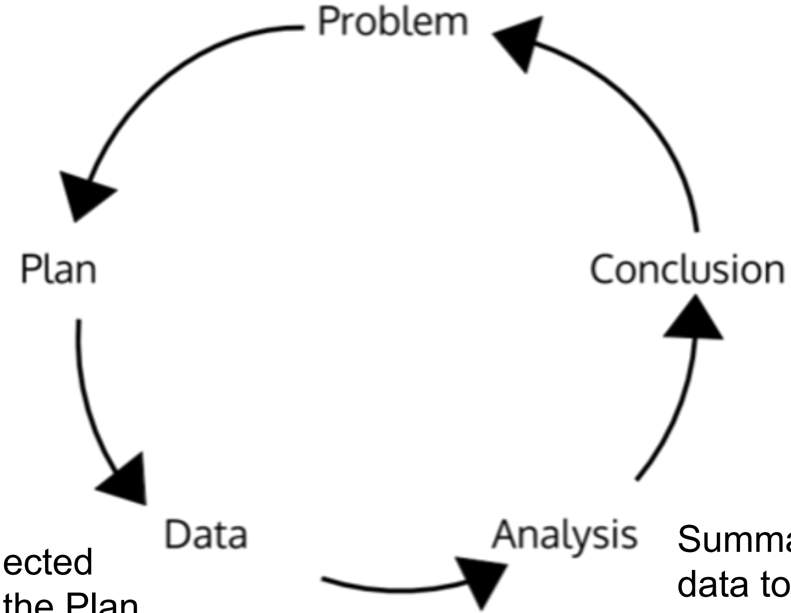
Environment is empty

Name	Size	Modified
..		
lab01		
README.md	0 B	Jul 4, 2022, 8:44 PM

PPDAC Approach

A 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 collected according to the Plan

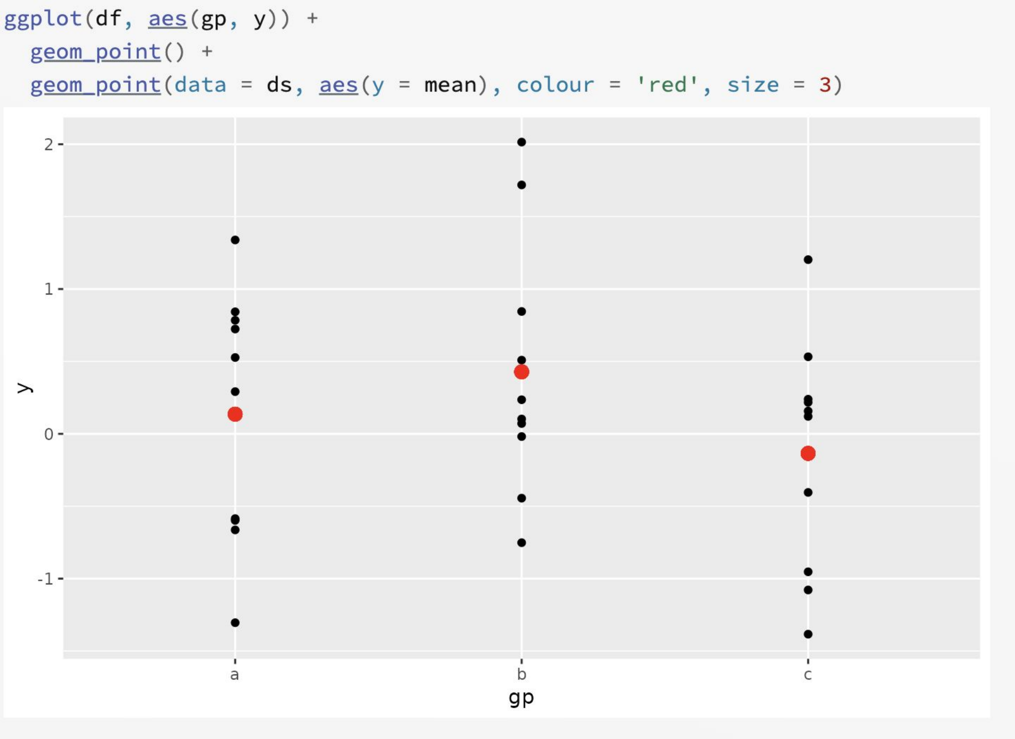
Summarization and analysis of the data to answer the questions posed by the Problem

Visualization of Categorical Data: “ggplot2”

1. Install and load the “ggplot2” package
   a. install.packages(“ggplot2”)
   b. library(ggplot2)
2. Specify your data, and what to have on the x and y axes
3. Create a plot: geom_ functions (many options)
   a. geom_bar, geom_histogram, geom_point, geom_line
   b. Tip: picture how you want to visualize your data in your head first, then pick the function that helps you achieve your goal :)
4. Change the style of your plot
   a. labs() function: update your main title, axis names, caption
   b. theme() function: change the size of your title, font, and position

Visualization of Categorical Data: “ggplot2”

e.g. https://ggplot2.tidyverse.org/reference/ggplot.html



```
ggplot(df, aes(gp, y)) +
  geom_point() +
  geom_point(data = ds, aes(y = mean), colour = 'red', size = 3)
```

Types of Variables

- Categorical: a variable that has grouping levels
  - Nominal: no underlying order or rank, e.g. blood type, zip code
  - Ordinal: with an underlying order or rank, e.g. blood pressure level (low, normal, high)
- Quantitative: a numeric variable which you can perform mathematical operations on
  - Discrete: can be counted, e.g. the number of cookies in the bag you got from a bakery
  - Continuous: can be measured precisely, with a rule or scale, e.g. 5.34 grams of cornstarch

Intro to R

- Library
  - A library is a package of functions, and you can load this package of functions by running library(ggplot2), library(dplyr)
  - Make sure you have them first, otherwise you need to do install.packages() first
- Read your data: e.g. how to read a .csv file
  - `library(readr)`
  - `mydata <- read_csv(‘my_data.csv’)`
- Some functions to get a quick look of your data
  - head(mydata): shows the first 6 rows of the dataset
  - dim(mydata): shows the total number of rows by the total number of columns
  - names(mydata) or colnames(mydata): shows all the variable names (column names) of the dataset
  - str(mydata): summarizes the information above and more

Data manipulation: functions in library(dplyr)

First, do `library(dplyr)` to have the package in your environment.

- rename() → renames variables (columns)
  - `new_dataset <- old_dataset %>% rename(new_name = old_name)`
  - or: `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)`

Data manipulation: functions in library(dplyr)

- 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

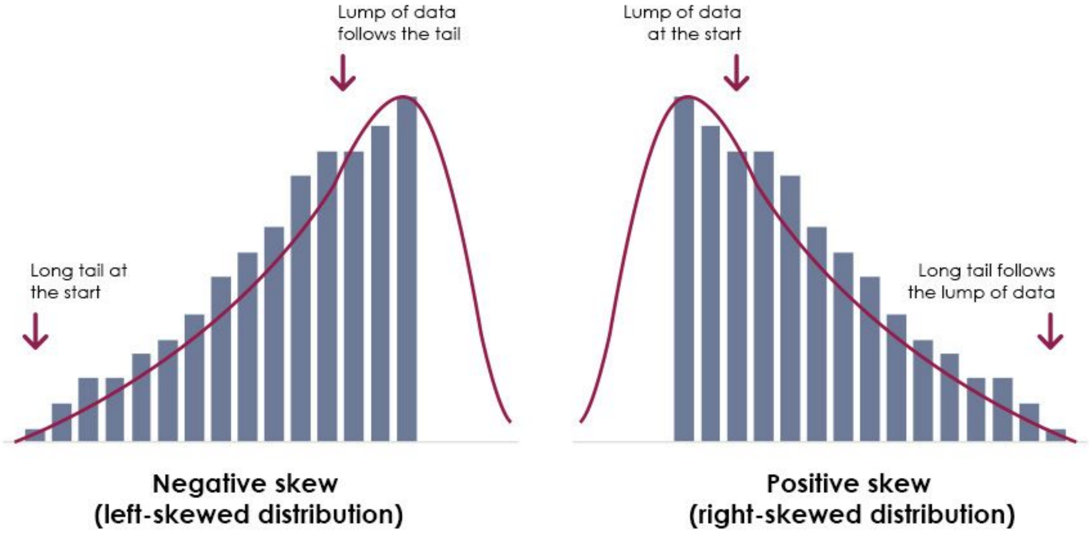
  -
    ```
    lake_data %>% summarize(mean_ph = mean(ph), sd_ph = sd(ph))
    ```

Measure of Central Tendency

- Mean and median are **approximately equal** when…
  - Distribution is symmetric
  - Data has one peak
  - There are no outliers

- Outliers: large effect on the mean

- Skewed data: mean ≠ median
  - Skewed right:
    mean > median
  - Skewed left:
    mean < median



Picture source: https://www.cambridgemaths.org/blogs/skewed-usage-skewed-distribution/

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))
```

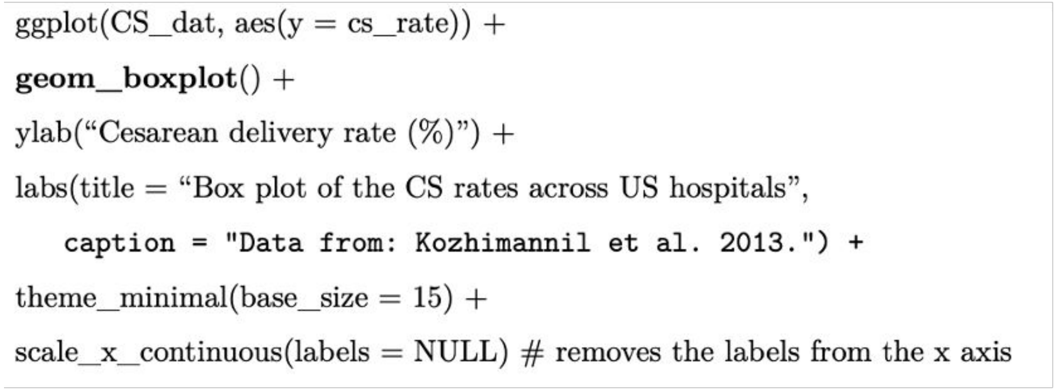
- Sample variance (s^2)
- Sample standard deviation (s)

```
CS_dat %>% summarize(cs_sd = sd(cs_rate), cs_var = var(cs_rate))
```

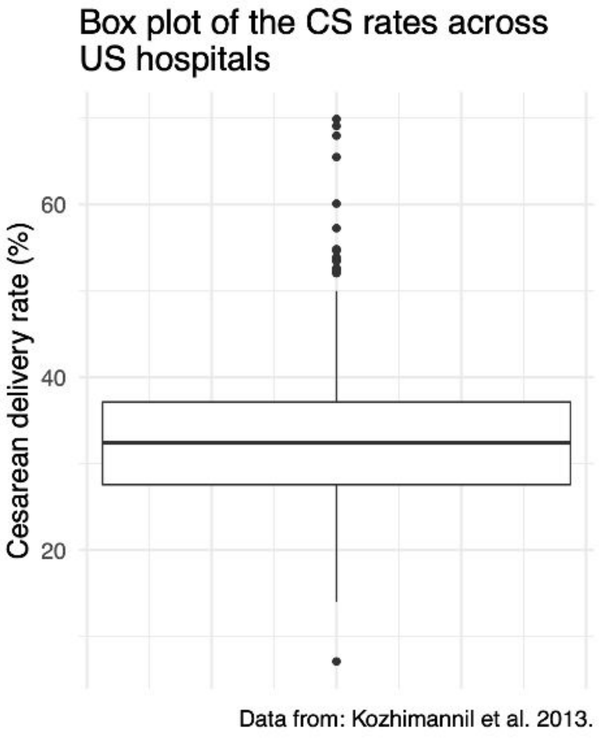
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Box plot

- 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
```



Box plot of the CS rates across US hospitals

Data from: Kozhimannil et al. 2013.

Common Errors

- Two different code chunks are named the same thing.
- The same variable names that are listed in the instructions are used in your work.
- If your data isn’t running, try reloading your past code chunks first.
- 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?