
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/>

PH 142

Home
Course Schedule
Calendar
Staff
Syllabus
Textbook
Extra Credit
Data Project


Gradescope
Piazza
Datahub
Calculate Your Grade

Powered by Just the Class

Search PH 142

Introduction to Probability and Statistics in Biology and Public Health

PH 142, Summer 2021



Mi-Suk Kang Dufour (she/her/hers)
mi-suk@berkeley.edu
Office Hours: By appointment only.
All office hours are held on Zoom or Google Meet.
2121 Berkeley Way West, Rm 5332
[Schedule an appointment](#)
[Zoom Link to Scheduled Appointment](#)

Also fluent in: French

We will not be updating this page with announcements. For the latest announcements, make sure to check our [Piazza](#).

Important Information

- **Lectures:** Monday - Friday, 9:30 to 11 AM PST
- **Location:** Online on Zoom
- **Content:** Please refer to the [course schedule](#)
- **Course number:** 14974
- **Email for non-content inquiries:** ph142@berkeley.edu

Goals

- Build strong foundations in statistics and introduce students to programming to prepare students for more

Key Course Technologies

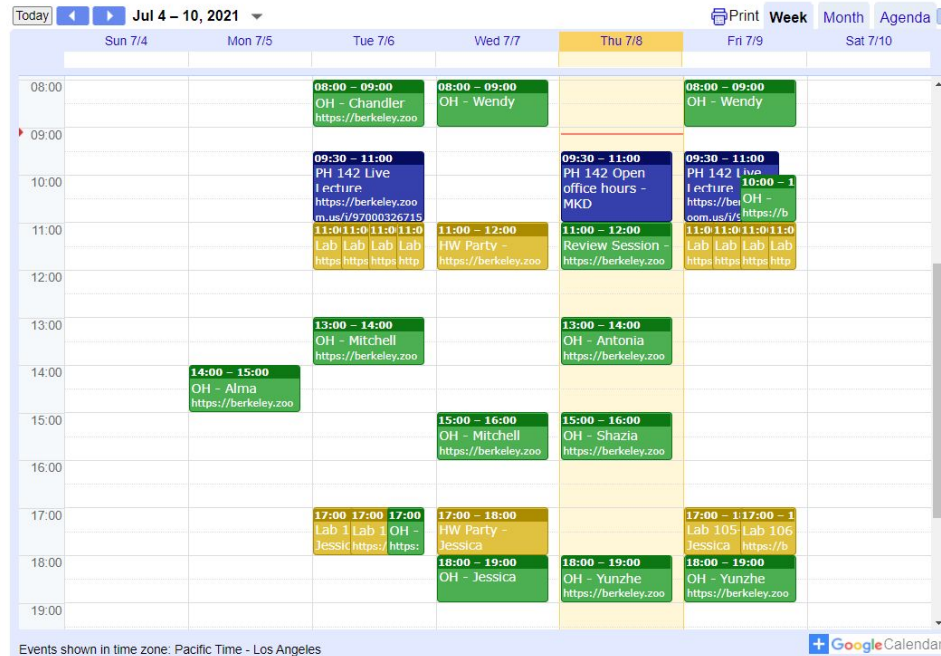
Accessing Slides and Recordings

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

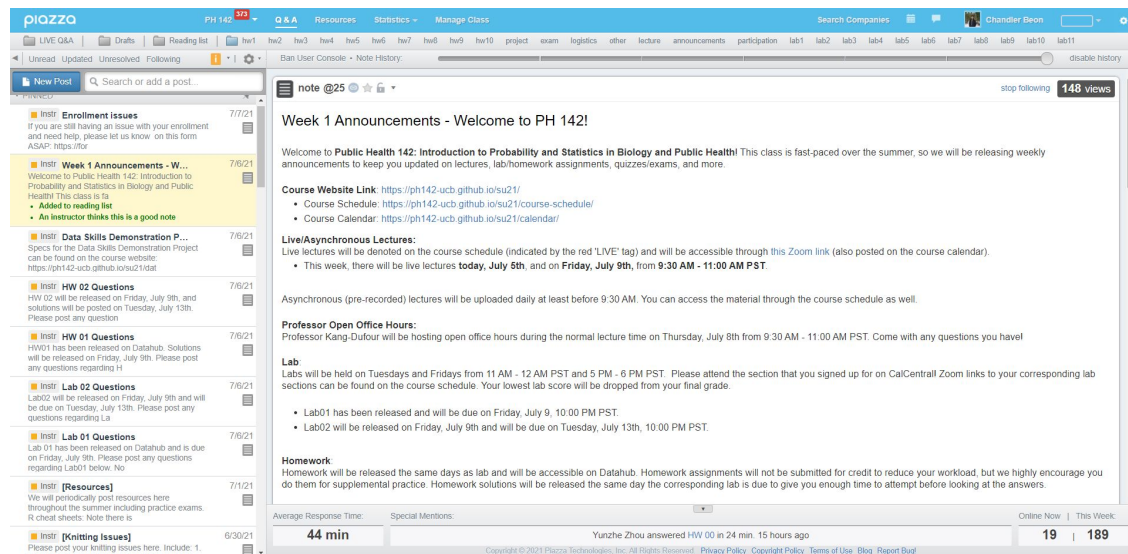
Course Calendar



Key Course Technologies

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

Piazza



The screenshot shows the Piazza forum interface for a course. The top navigation bar includes links for LIVE Q&A, Drafts, Reading list, and a search bar. The main content area displays a list of posts on the left and a detailed view of a selected post on the right. The selected post is titled "Week 1 Announcements - Welcome to PH 142!" and contains information about the course, including a welcome message, course website link, course schedule, course calendar, live/asynchronous lectures, professor open office hours, lab sections, and homework assignments. The post is marked as "note" and has 148 views. The bottom of the interface shows a response time of 44 minutes and a list of recent activity.

Week 1 Announcements - Welcome to PH 142!

Welcome to **Public Health 142: Introduction to Probability and Statistics in Biology and Public Health!** This class is fast-paced over the summer, so we will be releasing weekly announcements to keep you updated on lectures, lab/homework assignments, quizzes/exams, and more.

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

- Course Schedule: <https://ph142-ucb.github.io/su21/course-schedule/>
- Course Calendar: <https://ph142-ucb.github.io/su21/calendar/>

Live/Asynchronous Lectures:
Live lectures will be denoted on the course schedule (indicated by the red 'LIVE' tag) and will be accessible through [this Zoom link](#) (also posted on the course calendar).

- This week, there will be live lectures **today, July 5th**, and on **Friday, July 9th**, from **9:30 AM - 11:00 AM PST**.

Asynchronous (pre-recorded) lectures will be uploaded daily at least before 9:30 AM. You can access the material through the course schedule as well.

Professor Open Office Hours:
Professor Kang-Dukour will be hosting open office hours during the normal lecture time on Thursday, July 8th from 9:30 AM - 11:00 AM PST. Come with any questions you have!

Lab
Labs will be held on Tuesdays and Fridays from 11 AM - 12 AM PST and 5 PM - 6 PM PST. Please attend the section that you signed up for on CalCentral Zoom links to your corresponding lab sections can be found on the course schedule. Your lowest lab score will be dropped from your final grade.

- Lab01 has been released and will be due on Friday, July 9, 10:00 PM PST.
- Lab02 will be released on Friday, July 9th and will be due on Tuesday, July 13th, 10:00 PM PST.

Homework:
Homework will be released the same days as labs and will be accessible on Datahub. Homework assignments will not be submitted for credit to reduce your workload, but we highly encourage you do them for supplemental practice. Homework solutions will be released the same day the corresponding lab is due to give you enough time to attempt before looking at the answers.

Average Response Time: **44 min**

Special Mentions: Yunzhe Zhou answered HW 00 in 24 min 15 hours ago

Online Now: **19** | This Week: **189**

Copyright © 2021 Piazza Technologies, Inc. All rights reserved. [Privacy Policy](#) [Copyright Policy](#) [Terms of Use](#) [Blog](#) [Report Bug](#)

Key Course Technologies

Datahub

The screenshot displays the RStudio IDE interface. The main editor window shows a script titled 'hw01.Rmd' with the following content:

```
1 ---
2 title: "Assignment 1: Manipulation of birthweight data"
3 author: "Your name and student ID"
4 date: "Today's date"
5 output: pdf_document
6 ---
7 ### Instructions
8 * Due date: Thursday, July 9 at 10:00pm.
9 * Remember: autograder is meant as sanity check ONLY. It will not tell you if you have
  the correct answer. It will tell you if you are in the ball park of the answer so *CHECK
  YOUR WORK*.
10 * Submission process: Follow the submission instructions on the final page. Make sure you
  do not remove any ``\newpage`` tags or rename this file, as this will break the submission.
11
12 ```{r setup, include = FALSE}
13 # Don't change these lines, just run them!
14 source("setup/hw01.RAGS.R")
15 ```
16
17
```

The Environment pane on the right shows the Global Environment with the following data:

Variable	Value
birthwt	189 obs. of 9 variables
max_scores	chr [1:16] "2" "2" "1" "1" "1" "1" ...
num_tests	0
problem_names	chr [1:16] "Problem 1" "Problem 2" ...
problem_types	chr [1:16] "autograded" "autograded..."

The Files pane on the right shows the directory structure:

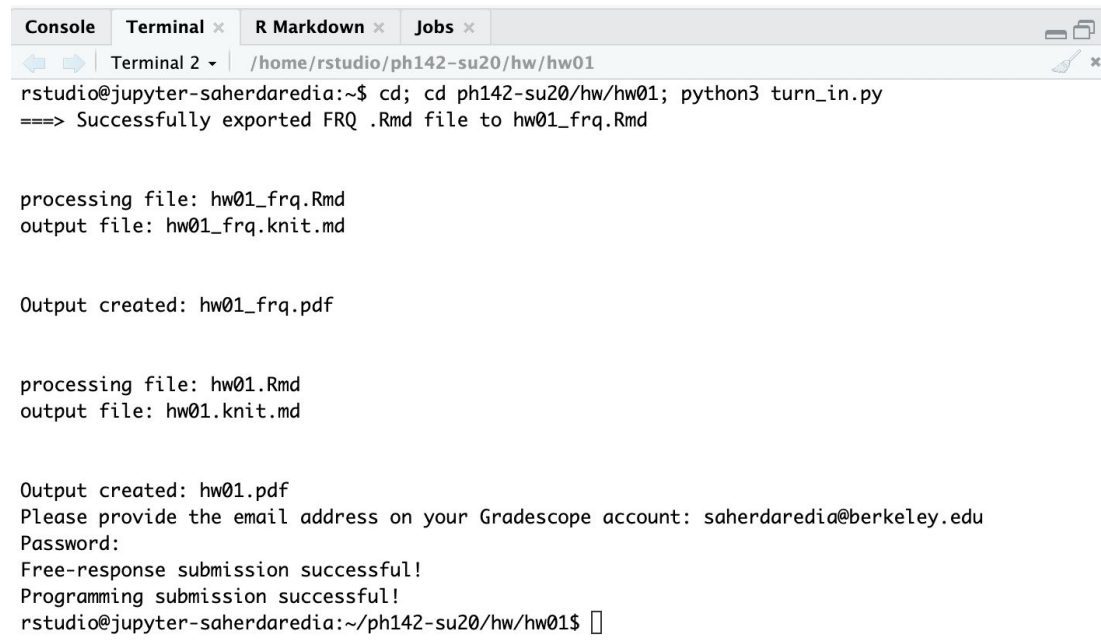
- Home > ph142-su20 > hw > hw01
- Files: .., autograder_setup.R (7 KB, Jul 6, 2020), birthweight.csv (7.2 KB, Jul 6, 2020), hw01.pdf (218.9 KB, Jul 6, 2020), hw01.Rmd (10.9 KB, Jul 6, 2020), hw01_frq.pdf (112 KB, Jul 6, 2020), hw01_frq.Rmd (1020 B, Jul 6, 2020), NUL (33 B, Jul 6, 2020), setup, src.

The Console pane at the bottom shows the R version and platform information:

```
R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)
```


Key Course Technologies

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

A screenshot of an RStudio terminal window. The window has tabs for 'Console', 'Terminal', 'R Markdown', and 'Jobs'. The 'Terminal' tab is active, showing a command prompt at the path '/home/rstudio/ph142-su20/hw/hw01'. The user has run a command to export an Rmd file and then knit it. The output shows the files created and a successful submission message to Gradescope.

```
rstudio@jupyter-saherdaredia:~$ cd; cd ph142-su20/hw/hw01; python3 turn_in.py
==> Successfully exported FRQ .Rmd file to hw01_frq.Rmd

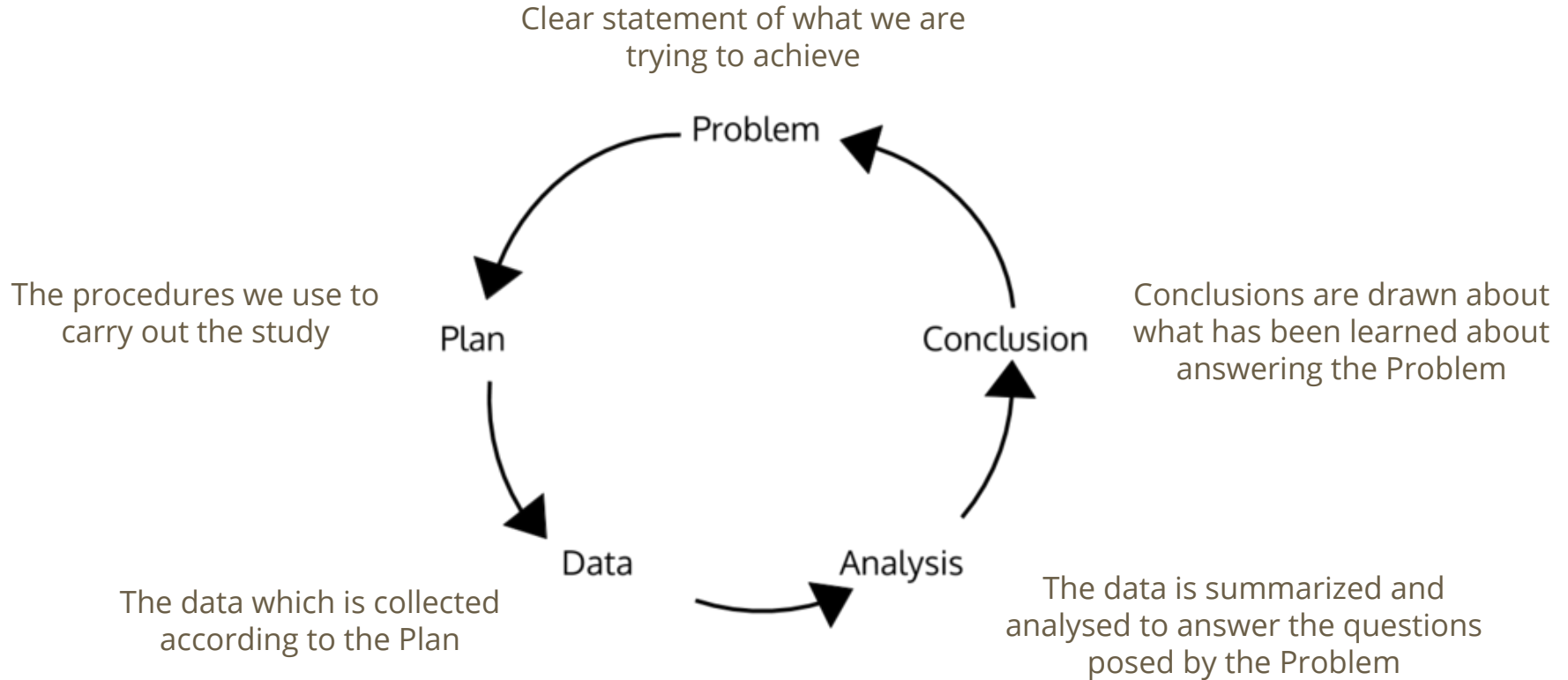
processing file: hw01_frq.Rmd
output file: hw01_frq.knit.md

Output created: hw01_frq.pdf

processing file: hw01.Rmd
output file: hw01.knit.md

Output created: hw01.pdf
Please provide the email address on your Gradescope account: saherdaredia@berkeley.edu
Password:
Free-response submission successful!
Programming submission successful!
rstudio@jupyter-saherdaredia:~/ph142-su20/hw/hw01$
```

PPDAC Approach



Visualization of Categorical Data: ggplot2

1. Install and load the ggplot2 package

- `install.packages(ggplot2)`
- `library(ggplot2)`

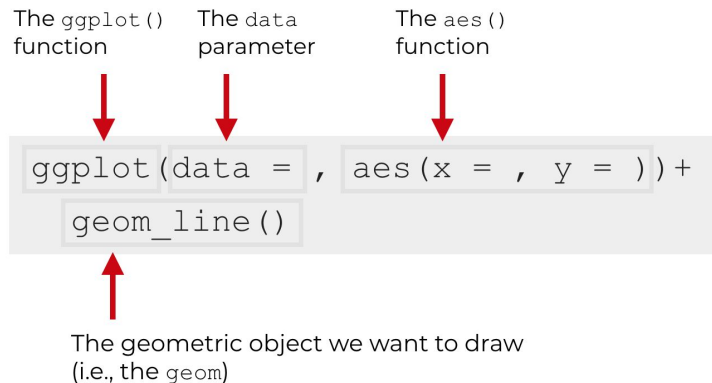
2. Specify the data set and what goes on the x and y axes

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

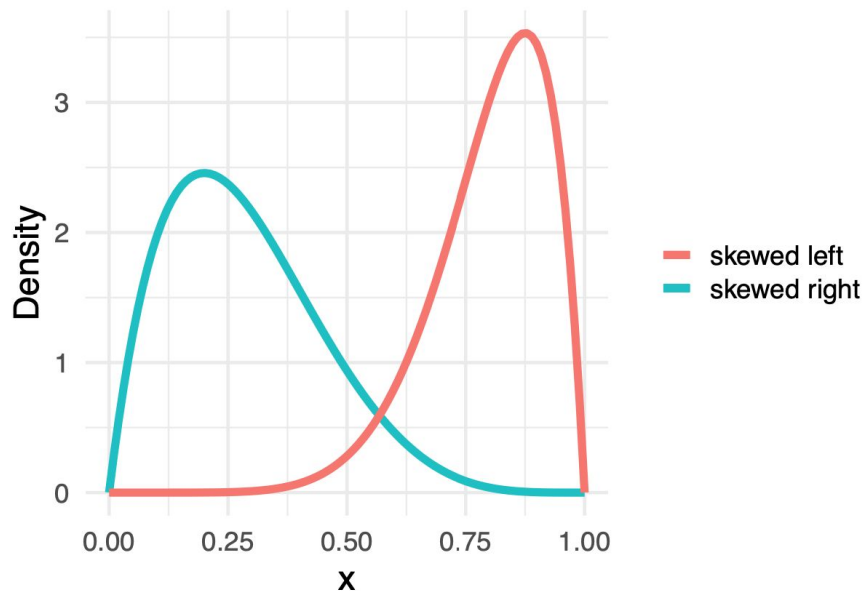
- `rename()` → renames variables (columns)
 - `new_dataset <- old_dataset %>% rename(new_name = old_name)`
 - Alternatively: `new_dataset <- rename(old_dataset, new_name = old_name)`
- `select()` → subsets variables (columns)
 - `smaller_data <- old_data %>% select(variable1, variable2, variable3)`
 - `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 \neq median
 - Skewed right: mean $>$ median
 - Skewed left: mean $<$ median



Measures of Spread

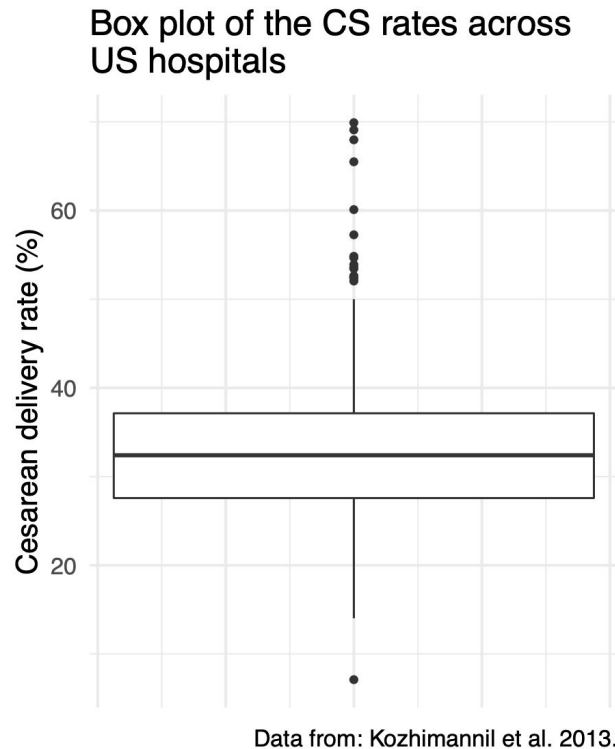
- Range = max - min
- IQR = Q3 - Q1
 - Five number summary in R!
 - ```
CS_dat %>% summarize(min = min(cs_rate),
 Q1 = quantile(cs_rate, 0.25), median = median(cs_rate),
 Q3 = quantile(cs_rate, 0.75), max = max(cs_rate))
```
- Sample variance ( $s^2$ )
- Sample standard deviation ( $s$ )
  - ```
CS_dat %>% 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 $Q3 + 1.5 \cdot IQR$
- Bottom of bottom whisker → min value or lowest point that is above $Q1 - 1.5 \cdot IQR$
- 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?