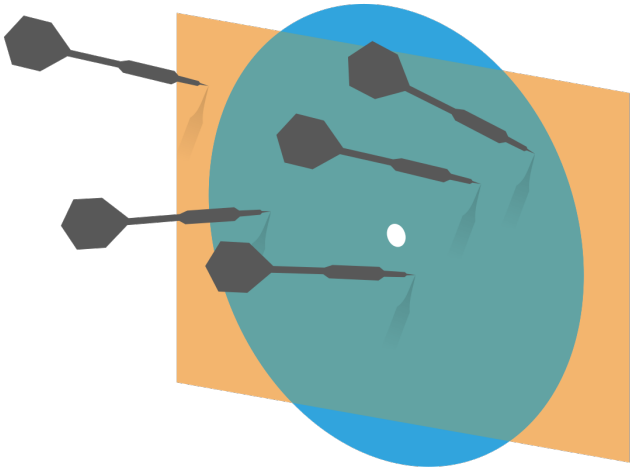
Probability and Statistics for Computer Science



Credit: wikipedia

"The eternal mystery of the world is its comprehensibility … The fact that it is comprehensible is a miracle."
– Albert Einstein

Hongye Liu, Teaching Assistant Prof, Course CS361, UIUC, 8.25.2020

* Upon entry, speakers of the students are muted for the quality of sound in Zoom room.

* Please "raise up hand" to speak, the audio will be unmuted for you.

* You can use "chat" to write private note to the instructor.

* You can use "chat" to ask questions and write comments.

* Don't share your screen during this lecture.

Test Poll1

- Have you read the syllabus on the course website?

  A. Yes. B. No.

Test Poll2

- Have you done the survey on the course Compass website?

  A. Yes. B. No.

Test Poll3

- Have you watched the welcome video in the Orientation module?

  A. Yes. B. No.

Objectives

- Welcome/Orientation
- Big picture of the contents
- Lecture 1 - Data Visualization & Summary (I)

Vision

- Passion for learning
- Compassion for each other

How to succeed in this course?

- Factors that will hinder you from success
- Factors that will help you succeed

Avoid these that could cause failure

- Academic integrity infraction – by all means!
- Missing homeworks or project
- Late/Poor homeworks or project
- Insufficient viewing of the contents
- Poor time management
- Too many challenging classes at the same time
- Not motivated/not interested in the topic

Factors that will help you succeed

- Try your best to be engaged/motivated, learn from the course and from each other
- Be **Active** in class participation
- Do as much practice as possible, not just the homeworks and project.
- Read the textbook and other recommended books.
- Clear your doubts/misconceptions **asap (every lecture/discussion is important)**

Interactions are important!

- Try to go to office hours as much as possible
- Try to meet or talk to the instructor as least once personally
- You are encouraged to join the team work
- Show compassion via community service

We will try to customize for students in international locations for team work

- Please answer this poll:

Are you in an international location that has more than 3hrs time difference from Central USA?

A. Yes

B. No

Graded Team work

How many ways to arrange "COVID"?

5×4×3×2×1 = 120
5!

Deduction	If
0	120
−1	60
−2	other

Your grading decision, reason

Write feedbacks to the "Student".

Team members: Leader ×××
. . . ——

Extra Points

Quizzes

Course materials

- Compass Course Site

  Find it through Compass for CS361 Fall 2020 AL1

- Public Website

  https://courses.engr.illinois.edu/cs361/fa2020/

Lecture videos and ClassTranscribe

- Lecture and discussion will be recorded and accessible at https://mediaspace.illinois.edu/
- ClassTranscribe provides transcripts for these videos

  https://classtranscribe.illinois.edu/home
- The specific links are all on Compass

Our Staff

Instructor: Hongye Liu

Teaching Assistants: Enyi Jiang (ADA)
Anay Pattanaik (ADB),
Nathan (ADC, ADD),
Aditya Karan (ADE),
Jinglin Chen (ADF),

Office hours are yet to be finalized.

Our Staff (II)

Course Assistants: Ajay Fewell, Brian Yang, Chenhui Zhang, Muhammed Imran, Vishesh Gupta, Yuxin Wang, and Zihan Xu.

What are the contents?

- Probability and Statistics in action

Randomness

- What does this course teach?

Textbook: Forsyth, D. A. "Probability and Statistics for Computer Science," Springer (2018)

- Why are there 4 sections? How are they related?

This field really started with gaming

- We are familiar with flipping a coin or throwing a dice, the result is uncertain!

Head
Or Tail?

Which side
is front?

Life is uncertain so aim for long-term average

- We repeat a lot of experiments and see if there is regularity

Head
Or Tail?

Which side
is front?

Throwing a lot of “coins” for many times in one touch

- Galton board, the Bead Machine

https://www.youtube.com/watch?v=Kq7e6cj2nDw

Simulation of random draw of a picture on computer

- It's the same as throwing a 4-sided die.

Probability and Statistics
Experiment in action

Break out !!

What does this course teach?

- Describing Datasets

  Summary & visualization

- Probability

- Inference – Statistical Inference

- Tools – Machine Learning tools

Chances in numbers

Look inside the box

Describing datasets (Summary & visualization)

Descriptive & Graphical

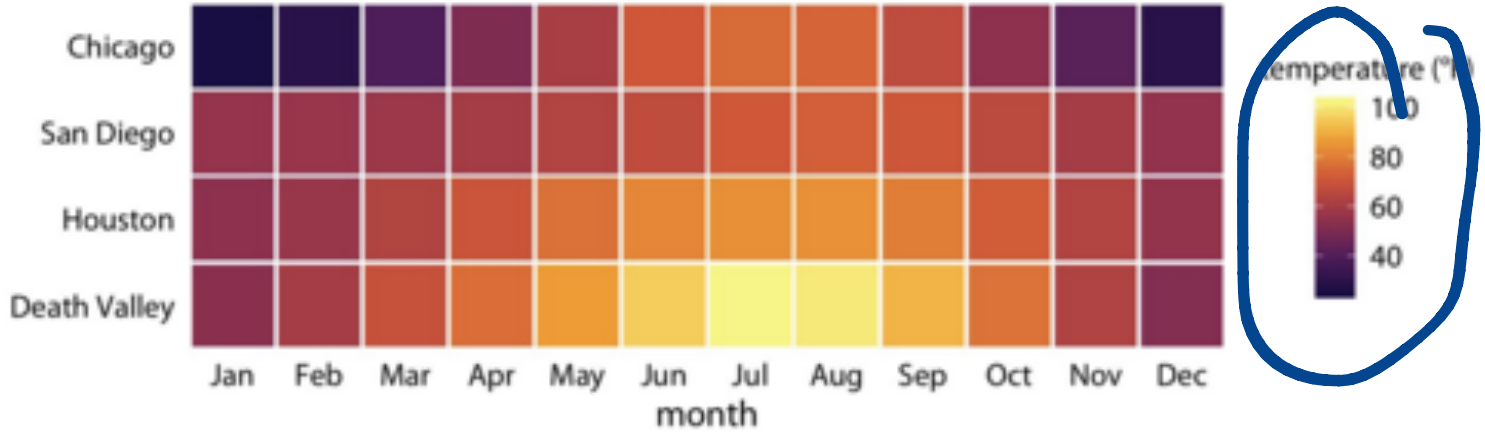


Figure 2-4. Monthly normal mean temperatures for four locations in the US. Data source: NOAA.

Summarization of 4 locations' annual mean temperature by month

Probability

- **Mathematical**

Romeo and Juliet have a date

Each arrives with a delay btw 0 and 1 hour. The first to arrive leaves after 1/4 hour. All pairs of delays are equally likely.

What's the probability that they will meet?

Probability

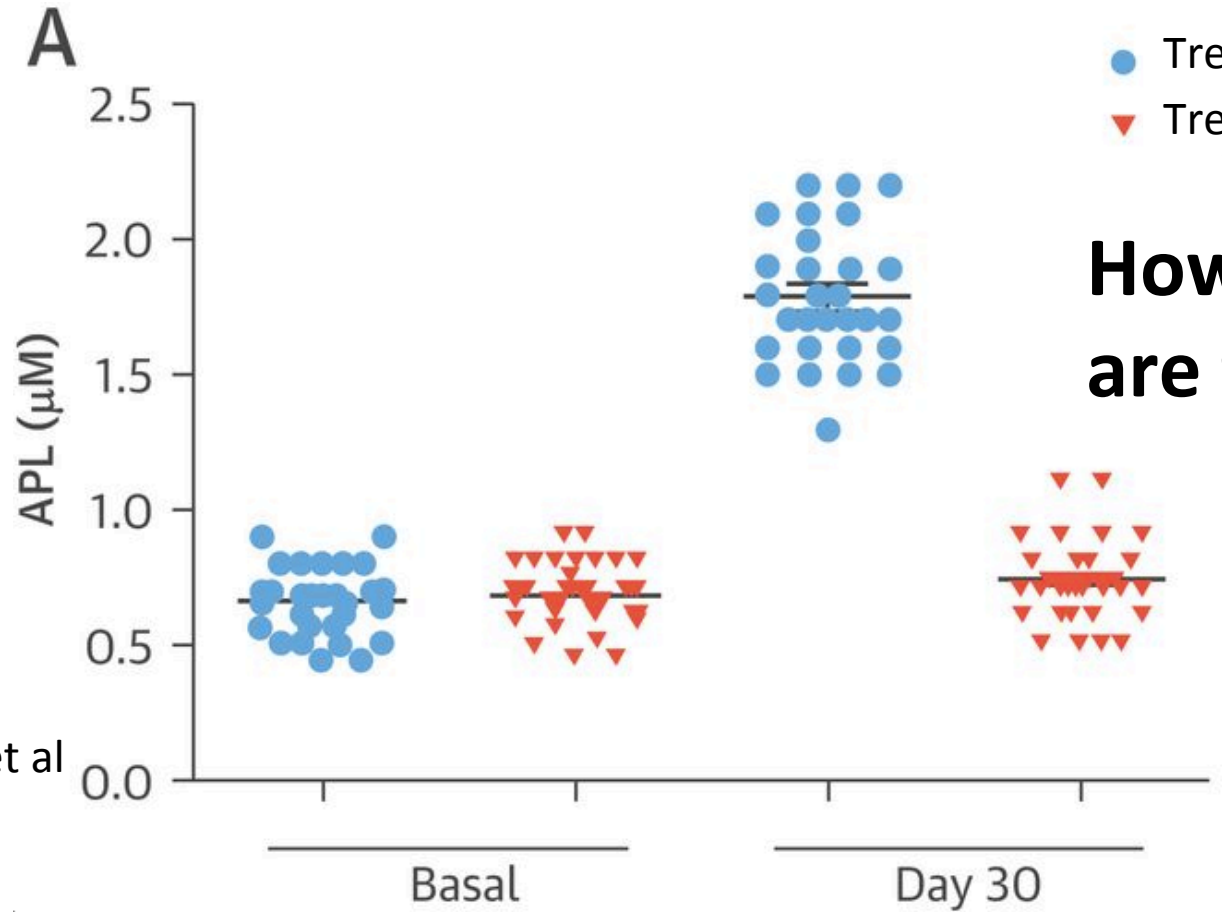
- **Mathematical**

How many slots are empty on average for a simple hashing?

(N items, K slots)

Inference

- **Analytical**



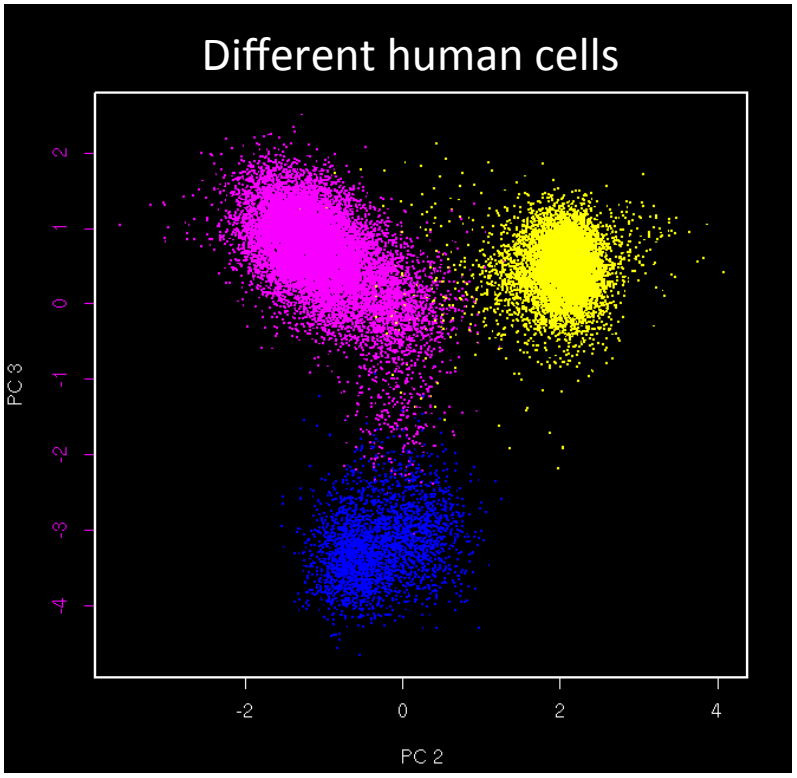
How different are they?

Significance is p-vals

J Fromonot et al
JACC 2016

Tools (Machine learning)

- **Algorithmical**



High-dimensional or complex shaped data sets need tools!
Humans are limited in 2-3D.
Machine learning is Highly desired!
Often depends on Statistics.

Why these 4 sections?

- Summary & visualization

  Graphical

- Probability

  Mathematical

- Inference – Statistical Inference

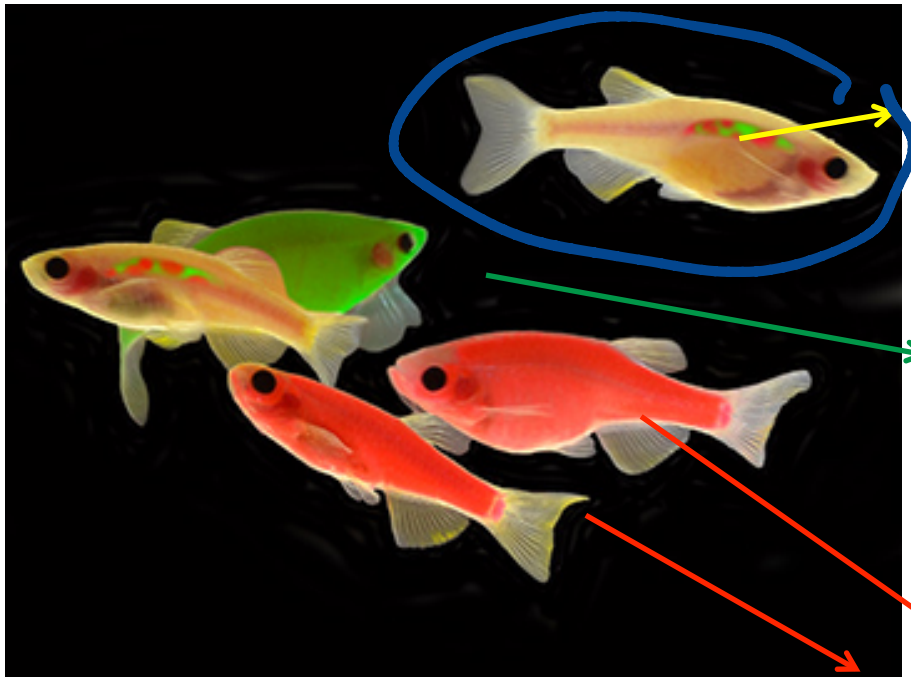
  Analytical

- Tools – Machine Learning tools

  Algorithmical

Why these 4 sections?

- The common thread is **Data.**
- We are doing computer science and so are like these yellow fish



What is special of Data? For Data?

Why these 4 sections?

- Real world data is often high dimensional and complex
- These 4 parts of knowledge or techniques are inseparably/ organically connected in many real world applications.

Not just big

Not in piece meal fashion

What do we emphasize?

- Mathematical principle
- Critical thinking
- Working with real world data

not only using tools

ask questions

LECTURE 1

Q. What do you feel about it when we speak of data visualization?

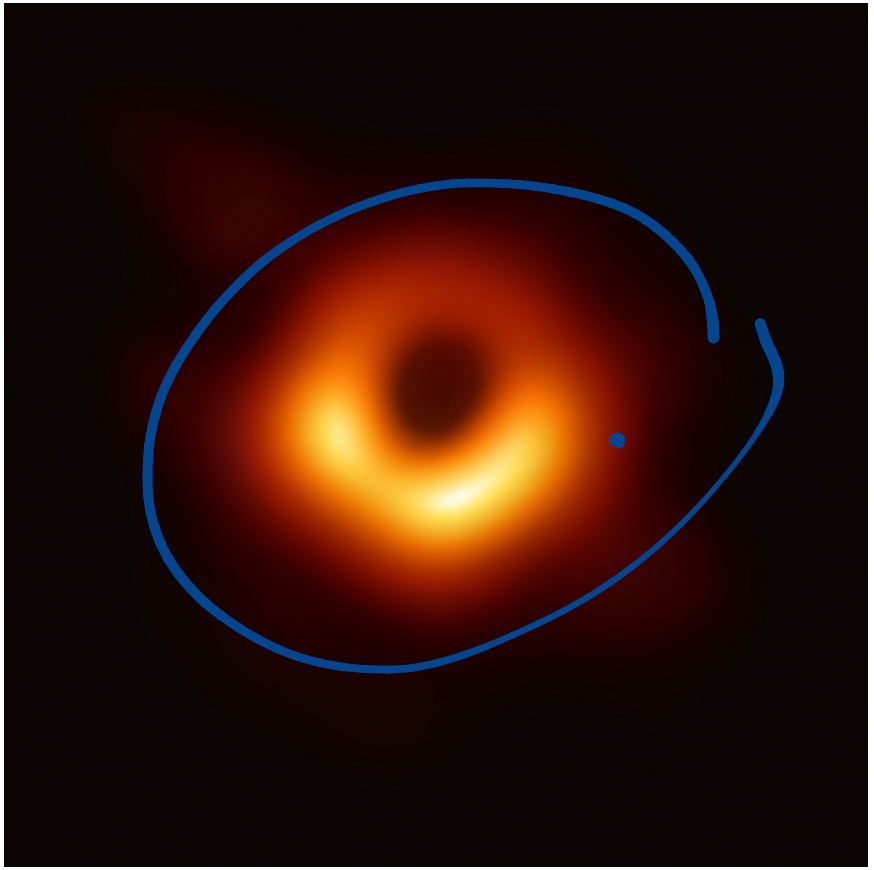
Example 1: Black hole

Constructed image using data collected from many different telescopes' view of the same object

This project received a 3million-dollar award

invisible → visible

more insights of data



Credit: NASA

Example 2: Four seasons by Vivaldi

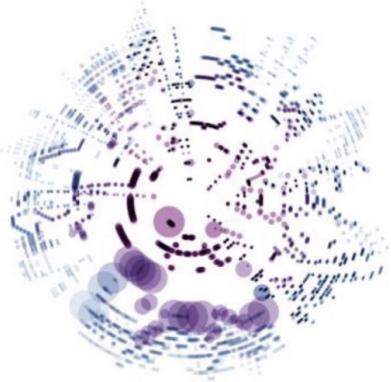
Pitch is shown by the distance from center;
Length of the note is the size of dot
Instrument is shown by the color

https://medium.com/future-today/off-the-staff-an-experiment-in-visualizing-notes-from-music-scores-58f6ee9f0cef

Spring

Summer

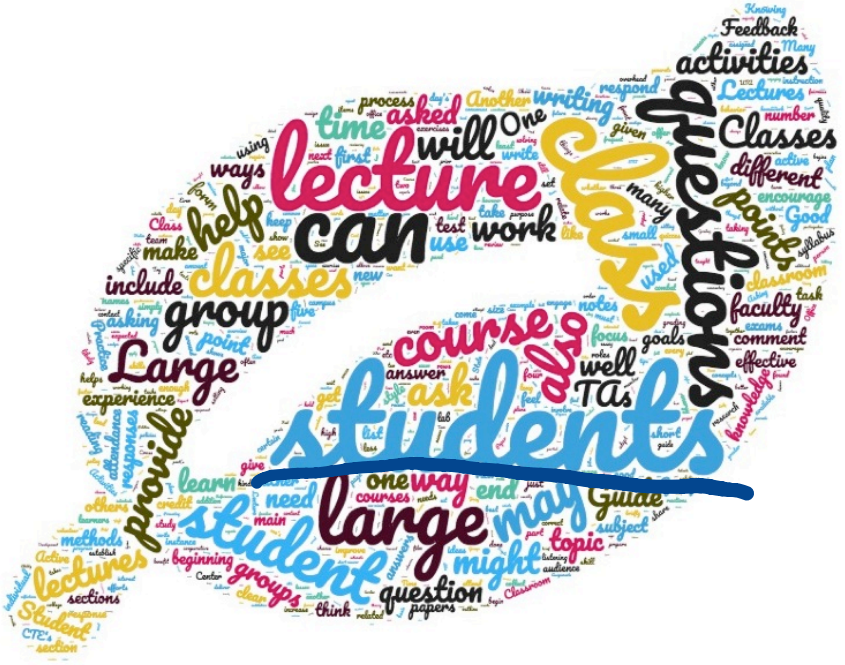
Autumn



Winter

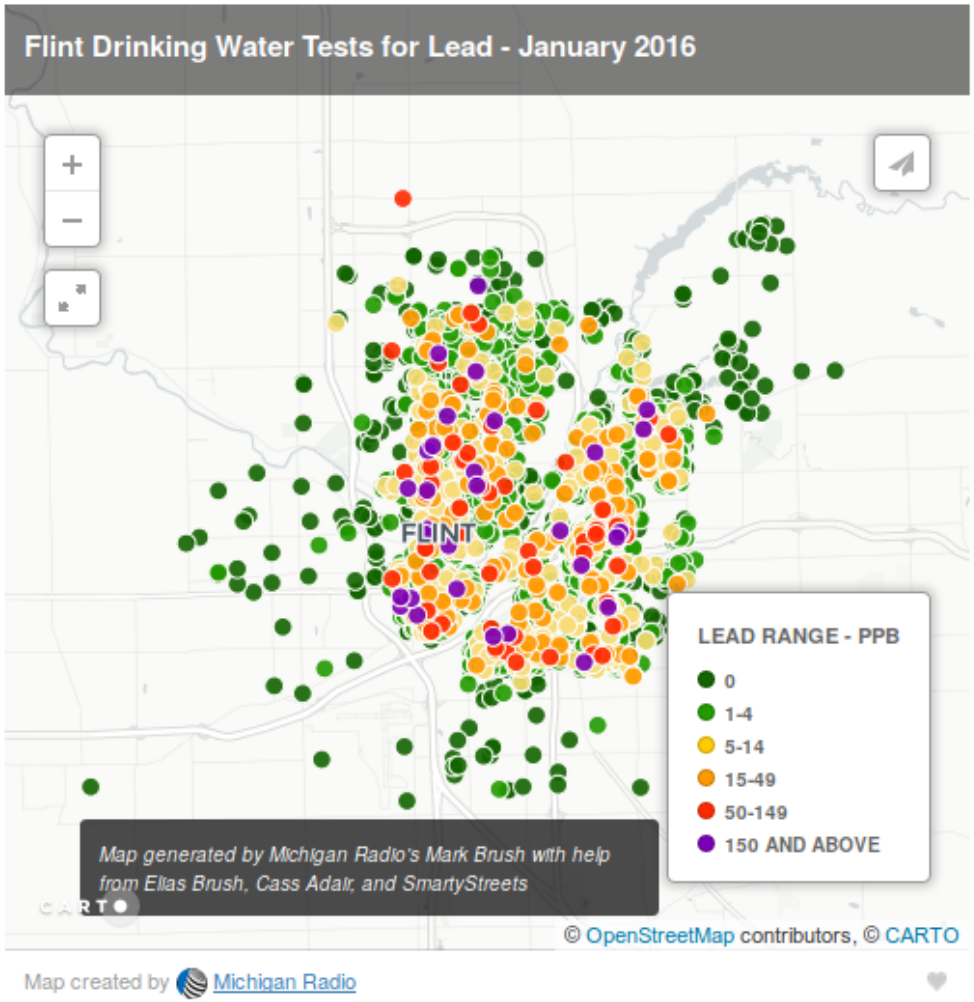
Example 3: Word cloud

Frequency of words of a document in novel visual presentation



Example 4: GIS map

Color scaled dots show the lead level in water in an area in Michigan



Lecture I: Data Visualization &Summary

- Datasets {x} – a set of N items x_i, i=1…N, each of which is a tuple

An example

Proteins →

Cells ↓

Cell ID	CD45	CD3e	CD19	CD11b	Ki67
1	7.10543765	1.99490875	2.13073358	7.82894178	2.57289058
2	6.5957055	4.65342077	1.62918585	0.88137359	0.88137359
3	6.81991147	1.76259579	4.63429706	2.74452653	0.88137359
4	6.90112651	1.41502227	4.54593607	0.88137359	0.88137359
5	6.75571436	2.87597714	2.18671075	6.72464322	0.91192661
6	7.39538689	2.55285118	4.55845203	1.57273629	0.88137359
7	6.50181654	0.9030504	0.88137359	6.55459538	1.61883699
8	6.60986569	2.1753298	1.52779681	6.44086205	1.5347653
9	6.97651408	2.38246511	1.90249637	3.41580053	1.85303806
10	7.14397512	3.36924119	9.23325502	4.79035059	0.88137359

Each row is a tuple

Lecture I: Data Visualization &Summary

- Convention: columns are the ***features***; the number of features is ***dimension.***

Proteins →

Cells ↓

Cell ID	CD45	CD3e	CD19	CD11b	Ki67
1	7.10543765	1.99490875	2.13073358	7.82894178	2.57289058
2	6.5957055	4.65342077	1.62918585	0.88137359	0.88137359
3	6.81991147	1.76259579	4.63429706	2.74452653	0.88137359
4	6.90112651	1.41502227	4.54593607	0.88137359	0.88137359
5	6.75571436	2.87597714	2.18671075	6.72464322	0.91192661
6	7.39538689	2.55285118	4.55845203	1.57273629	0.88137359
7	6.50181654	0.9030504	0.88137359	6.55459538	1.61883699
8	6.60986569	2.1753298	1.52779681	6.44086205	1.5347653
9	6.97651408	2.38246511	1.90249637	3.41580053	1.85303806
10	7.14397512	3.36924119	9.23325502	4.79035059	0.88137359

Each row is a tuple with dimension =5

Data types

- Categorical

positive negative

- Ordinal

Ratings

Grades

A B C D

- Continuous

temp. height

Data types

- Categorical

  Smoker or non-Smoker, Female or Male etc.

- Ordinal

  Not satisfied, satisfied, very satisfied

- Continuous (any real number within a range)

  Temperature

Q. Which of the following data is not categorical?

A. Number of enrolled students in a class
B. Weight of apples in a grocery store
C. Instruments played by an orchestra
D. Type of chemical reagents in a lab
E. A & B

Simple Visualization of Data

- General principles
- Bar chart
- Histogram
- Conditional histogram

Simple Visualization of Data

- General principles

  Must not mislead or distort;

  Aesthetically pleasing;

  Clear, Attractive, Convincing;

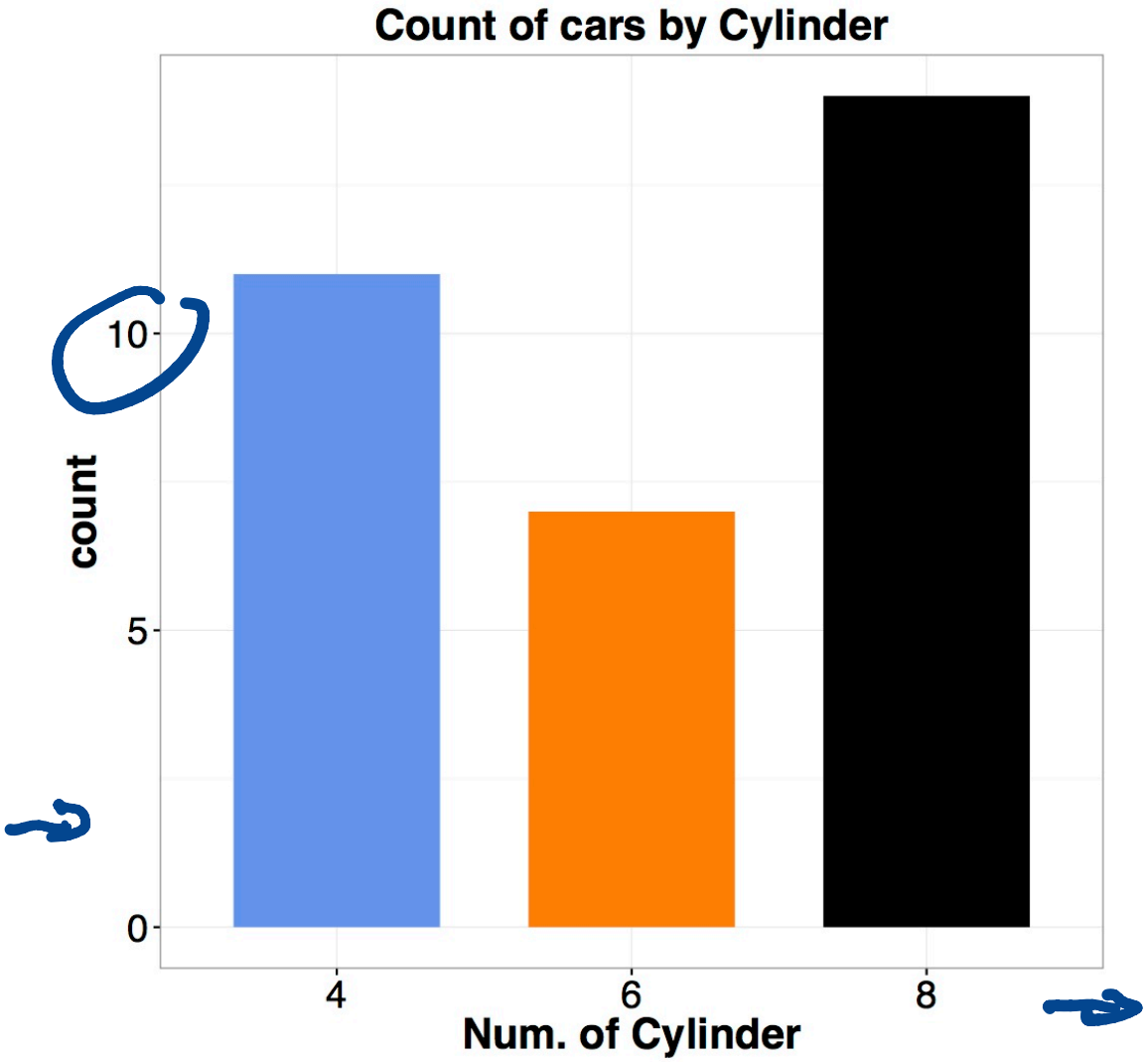
  Show message/significance.

Simple Visualization of Data

- Bar chart

A set of bars that are organized by categorical or ordinal feature

Data: “mtcars”



An example of good, ugly, bad, wrong

Dr. Wilke illustrated the difference between ***good***, ***ugly, bad and wrong visualization***

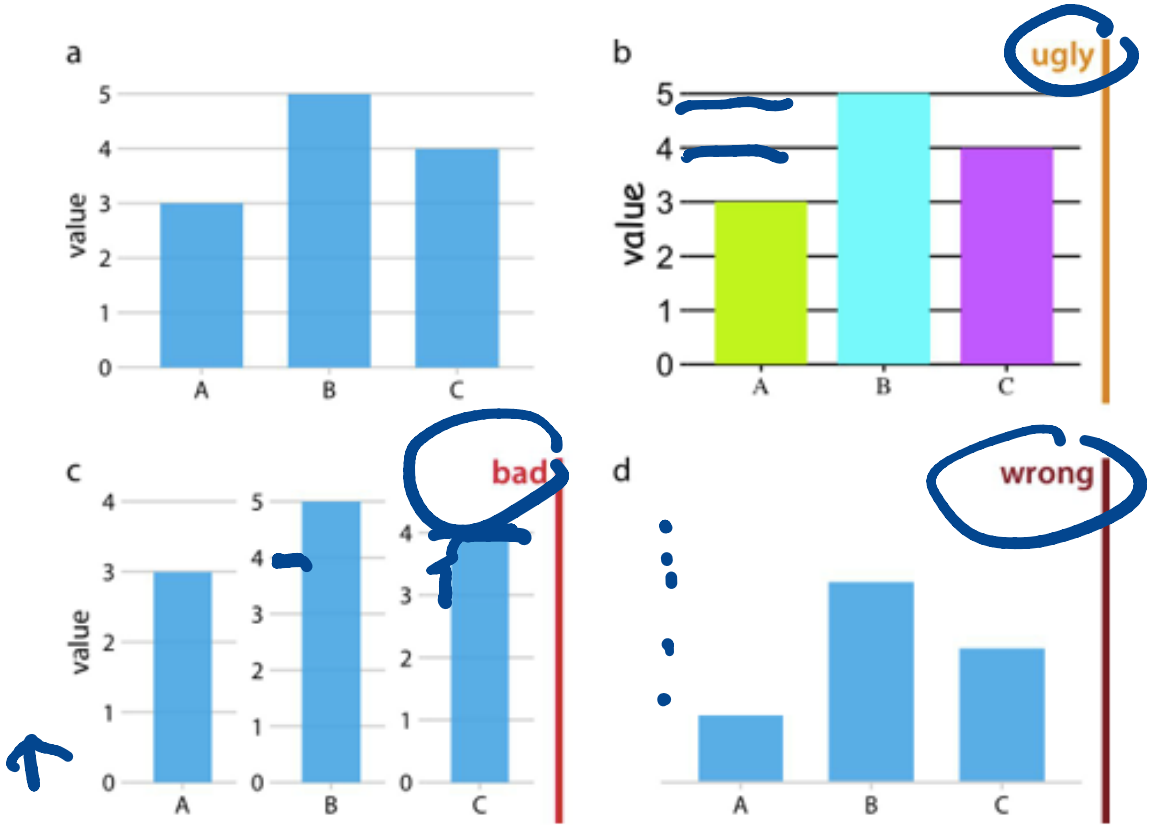
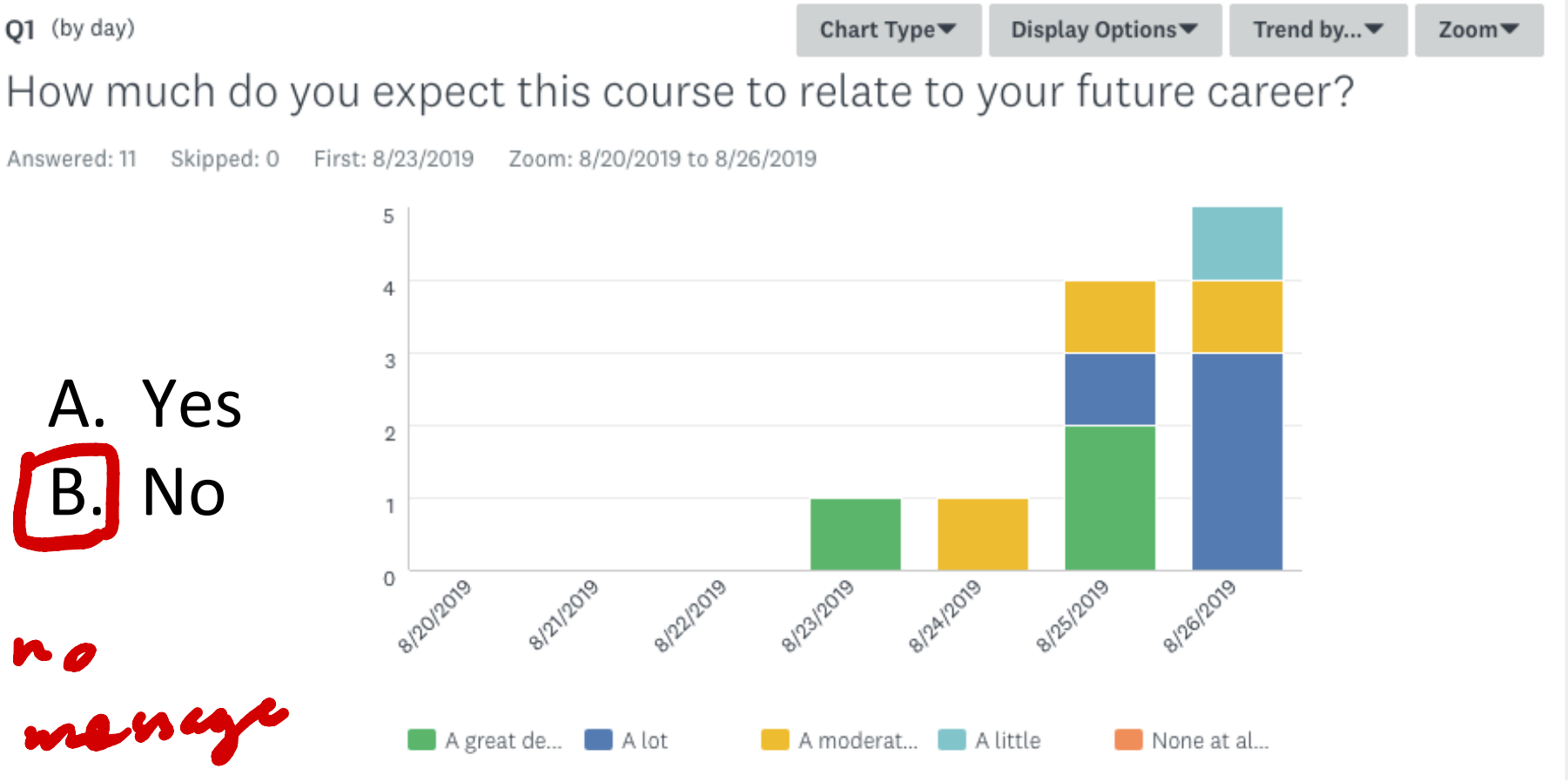


Figure 1-1. Examples of ugly, bad, and wrong

C. Wilke "Fundamentals of Data Visualization"

Q: Is this a good bar chart?



A. Yes

B. No

no message

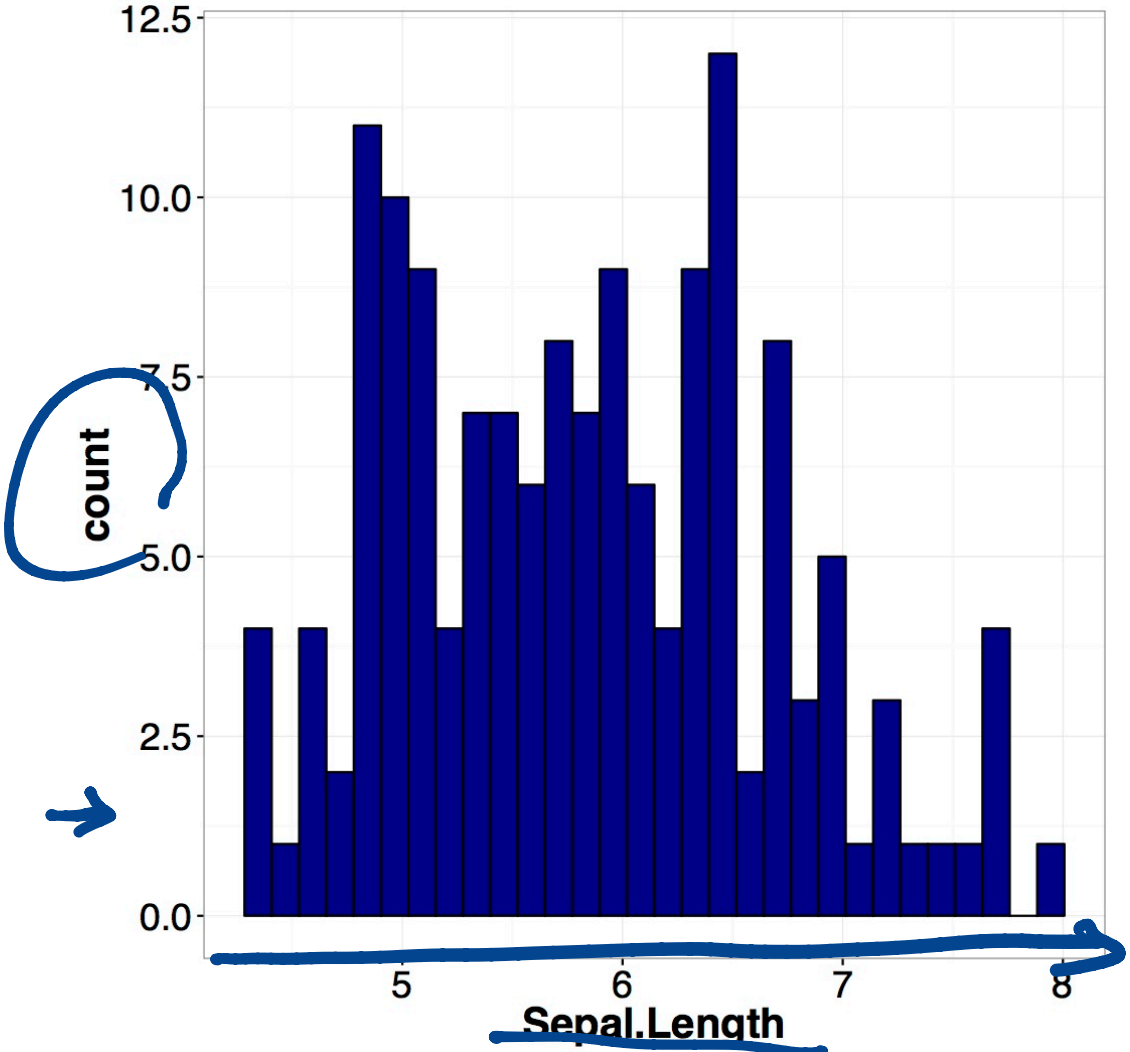
How about using a color scale

Visualizing Data with Histogram

- Histogram

A set of bars that are organized by bins that contain numerical data (discrete or continuous)

Data: “iris”

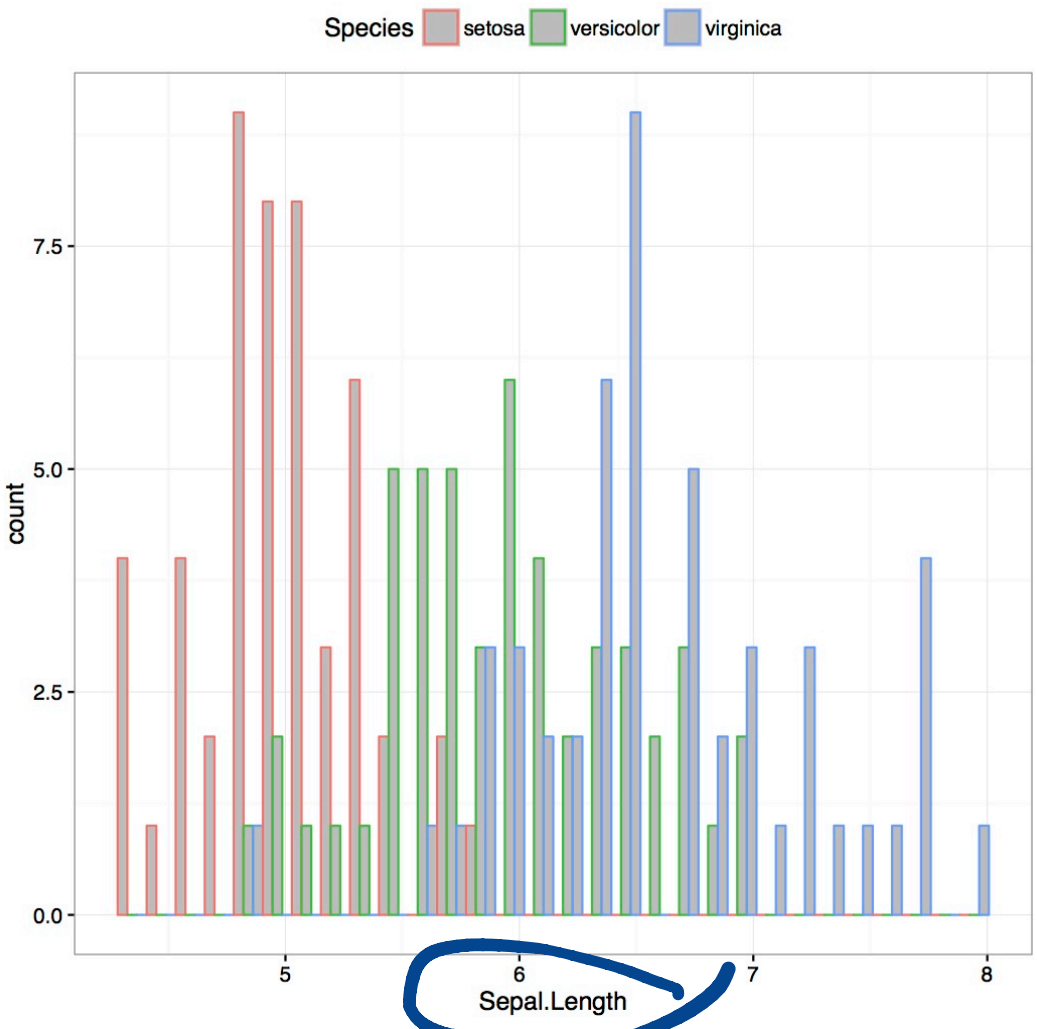


Visualizing Data with Histogram (II)

- Conditional histogram

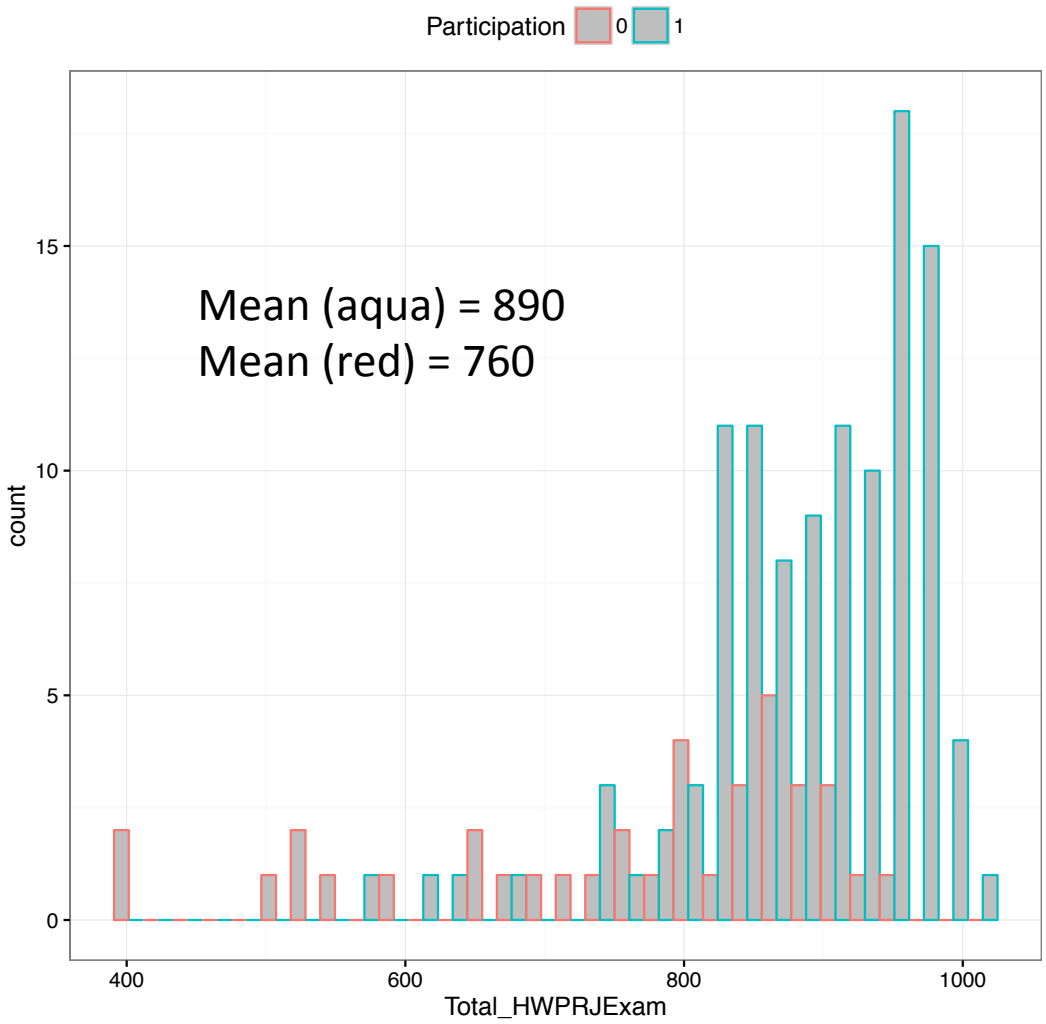
Histogram generated by subsets of the data

Data: “iris”



Visualizing Data with Histogram (III)

- Conditional histogram



Data: Combined Score (HWs, Prj and Exams) grouped by students with full participation or **not full** in CS361 fall 2019

Additional References

- Charles M. Grinstead and J. Laurie Snell "Introduction to Probability”
- Morris H. Degroot and Mark J. Schervish "Probability and Statistics”

See you next time

See you!