DATA 8
Spring 2021

Discussion 4

Visualizations and Histograms

Materials: tinyurl.com/d8-disc04
or access through kevin-miao.com
under teaching

Today

- Check-In
- Announcements
- Review: Visualizations
- Worksheet
  - Challenge Question (last question on worksheet) is optional

Check-In

- How do you feel today?
  - How were last week's homework/lab?
  - How do you feel about the material in the class?
  - Are you getting enough support?
- Always feel free to email, private message me or stay after class to discuss these matters

If you ever have any extenuating circumstances, please reach out to and we will figure out how we can accommodate you!

Announcements

- **Midterm** will be held on March 12, 7-9PM PT
  - Are you in a different time zone and is this time inconvenient? Please complete the alternate exam form on Piazza.
- **The vitamin** will be **due today**
- **Homework 3** is **due tomorrow** (early submission tonight)
- **Project 1** will be released on **Friday**
  - Projects are basically longer homeworks with checkpoints
  - Project Partner
    - You will be able to pair up with another person in your designated lab section
    - Matching Forms will be sent out in due course

Visualizations

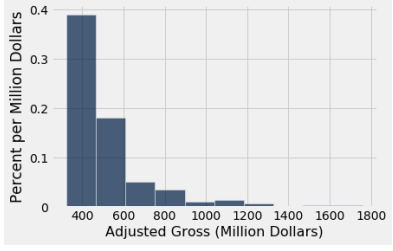
- **Line Graphs**
  - Sequential (Time/Distance) data
    - *Number of movies* over the *years*
- **Scatter Plot**
  - X and Y are both numerical (X and Y axis are interchangeable)
  - Testing for any **association**
    - Relationship between parent's height and child's height
- **Distributions**
  - How are data points spread out over certain categories or bins?
  - **Categorical**: Bar Charts, Pie Charts
    - # Students and their ZIP Codes
  - **Numerical**: Histograms
    - # Students and their grade percentage in Data 8

Histograms

- What is the difference between **histograms** and **bar charts?**

on the y-axis for histograms we have densities
on bar chart we have percentages and/or number!

- **Histograms**
  - **Areas** as **percentages**
  - **Height** as **densities**
  - The complete area under a histogram is always 1
  - Bins (can be arbitrary)
  - Formulas:



$$height = \frac{\%\ in\ a\ bin}{width\ of\ the\ bin}$$

$$area = \% = width\ of\ bin * height\ of\ bar$$

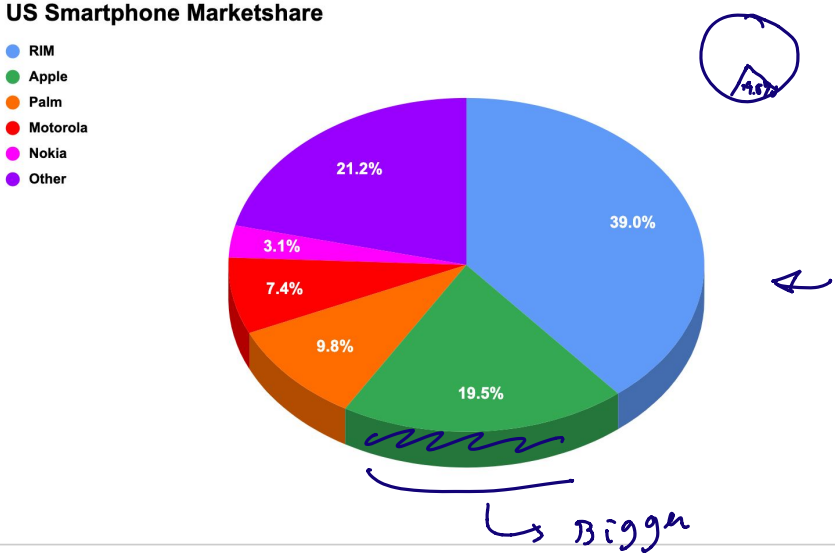
width height

To the worksheet! ✍️

tinyurl.com/d8-disc04

Question 1

Question 1. The following graphic is a recreation of a graphic presented by Steve Jobs in a keynote at Macworld in 2008. Discuss the graph below with your neighbors, then answer the questions below. (Source: https://www.wired.com/2008/02/macworlds-iphon/)



1a) What's so misleading about this graph?

- Graph is tilted making some slices appear bigger
- Because humans are bad at angles, pie charts might not be the best choice.

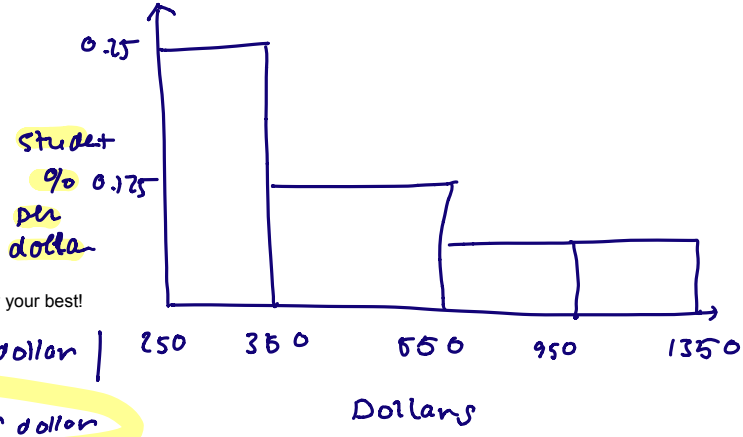
1b) How would you visualize it instead?

Bar Chart

Question 2a

Question 2. The table below shows the distribution of rents paid by students in a college town. The first column consists of ranges of monthly rent, in dollars. Ranges include the left endpoint but not the right. The second column shows the percentage of students who pay rent in each of the ranges.

Dollars	Student (%)
250-350	25
350-550	25
550-950	25
950-1350	25



a) Draw a histogram of the data. You do not have to be precise with your drawing, but try your best! Make sure you label your axes!

Height (250-350): $\frac{25\%}{100\$}$ = 0.25 % per dollar

(350-550): $\frac{25\%}{200\$}$ = 0.125 % per dollar

(550-950): $\frac{25\%}{400\$}$ = 0.0675 % per dollar

(950-1350): $\frac{25\%}{400\%}$ = 0.0675 % per dollar

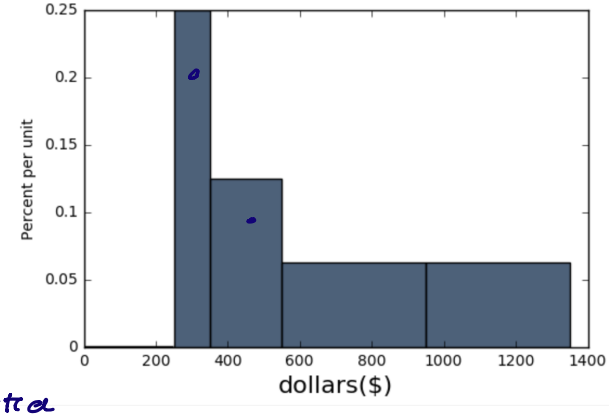
Question 2bcd

b) What is the height of the bar over the bin 350-550 on the density scale, in the correct units?

A. 12.5% per student
B. 0.125% per student
C. 0.125% per dollar
D. 12.5% per dollar

→ 0.125

Heights are all over the place



c) True or false (explain): The data show that the rents are evenly distributed over the interval 250-1350.

False, we look at the density scale and the heights are not the same.

d) True or False (explain): The data show that the rents are evenly distributed over the interval 550-950.

False, we don't know how it's distributed in the bin

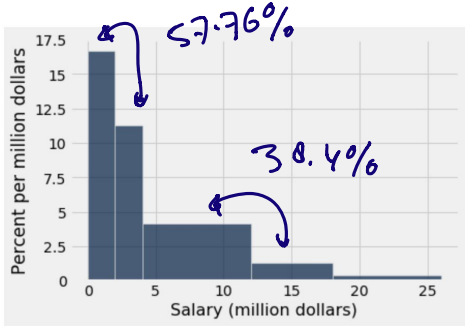
550$ - 950$ → 25%

→ $550 → 25%

→ $949 → 25%

Question 3

Question 3. The table `nba` has a column labeled `salary` containing the 2015-2016 salaries of NBA players. The following histogram was generated by calling `nba.hist(...)`. Also included below is a table with the bins and their corresponding heights.



Bin (million dollars)	[0,2)	[2,4)	[4,12)	[12,18)	[18, 26)
Height (percent per million dollars)	17.49	11.39	3.60	1.60	0.45

The interval `[a,b)` contains all values that are greater than or equal to a and less than b.

Which range contains more players: [0, 4) or [4, 18)? How many players are in that range? Explain.

players [0, 4) ← more players

2 * 17.49 + 2 * 11.39 = 57.76

[4, 18)

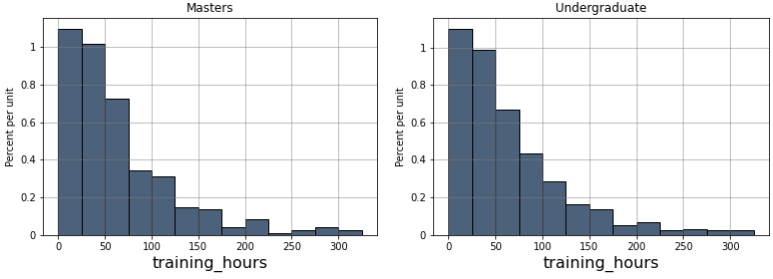
3.6 * 8 + 1.6 * 6 = 38.4

Can we calculate how many players are in the bin [18, 20)?

- We don't have the total # of players
- We can't look inside the bin!

Question 4

Question 4. Below are two histograms visualizing the number of training hours for a data scientist's job. The first histogram is that of a data scientist whose highest level of education is a master's degree and the second histogram is that of a data scientist whose highest level of education is an undergraduate degree. The sample size of masters students is 400, and the sample size of the undergraduate students is 200.

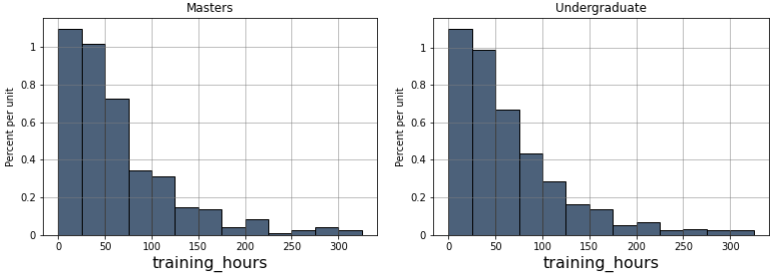


4A) (I) The number of masters data scientists with training hours between 75 (inclusive) and 100 (exclusive) hours (the height of the bin is 0.35) vs. (II) The number of undergraduate data scientists with training hours between 75 (inclusive) and 100 (exclusive) hours (the height of the bin is 0.42)

A. (I) is smaller
B. (II) is smaller
C. (I) and (II) are roughly equal
D. There is not enough information to compare (I) and (II)

Question 4

Question 4. Below are two histograms visualizing the number of training hours for a data scientist's job. The first histogram is that of a data scientist whose highest level of education is a master's degree and the second histogram is that of a data scientist whose highest level of education is an undergraduate degree. The sample size of masters students is 400, and the sample size of the undergraduate students is 200.



4B) What sample size for the undergraduate students could have made it so that (I) and (II) are roughly equal?

A. 167
B. 333
C. No change to the sample size would have made a difference
D. (I) and (II) are already roughly equal

Question 5 (Challenge)

Question 5 (Challenge). Below are two tables that represent the data of the same distribution and table, but with different histograms and bin widths. Fill out Histogram 3 with the appropriate y-axis values given the new bin values so that Histogram 3 also represents the data of the same distribution and table as Histograms 1 and 2.

Histogram 1	
X-axis (unit)	Y-axis (% per unit)
[0, 5)	10.5
[5, 10)	5.5
[10, 15)	2.1
[15, 20)	1
[20, 25)	0.5
[25, 30)	0.4

% for [5, 10)

5 * 5.5

Histogram 2	
X-axis (unit)	Y-axis (% per unit)
[0, 2.5)	8
[2.5, 10)	8
[10, 15)	2.1
[15, 17.5)	0.6
[17.5, 25)	0.8
[25, 30)	0.4

% for [2.5, 10)

7.5 * 8

Histogram 3		
	X-axis (unit)	Y-axis (% per unit)
a	[0, 2.5)	8
b	[2.5, 5)	$\frac{7.5 * 8 - 5 * 5.5}{2.5}$
c	[5, 17.5)	
d	[17.5, 20)	
e	[20, 30)	

End of Section
How did I do?

https://tinyurl.com/kevind8feedback