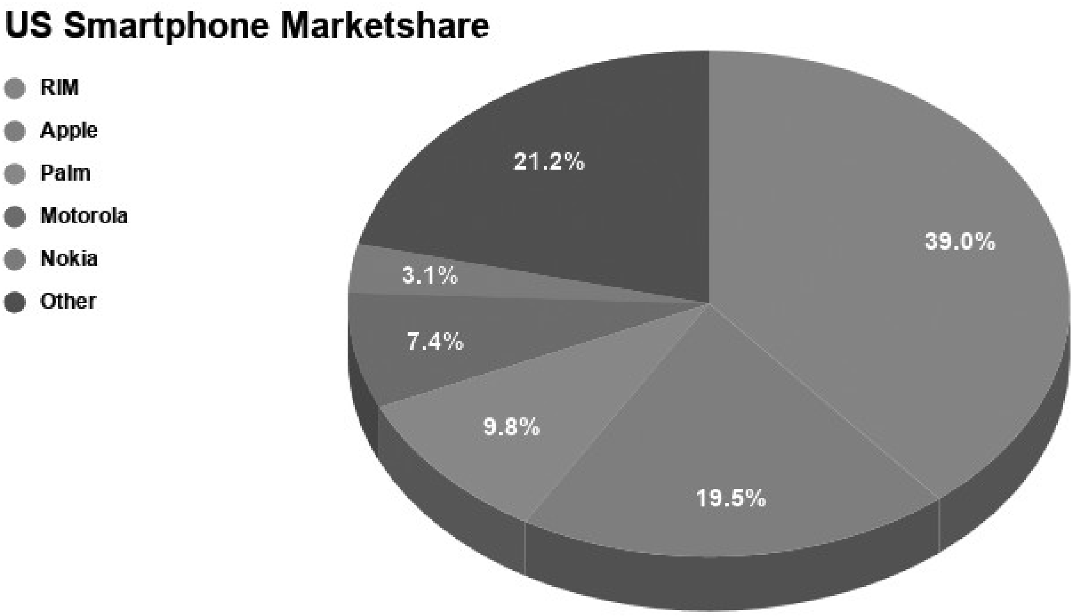
CS 2065 2022
Discussion: Visualizations and Histograms (Lab 04)

An extremely important aspect of data science is *visualizing* the data in a precise, consistent manner. This week, we will first examine some instances of bad visualizations, and think about how we can improve them. Then, we will transition to focus on *histograms*, which are powerful visualizations used to display the distribution of values for numerical data.

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/)



a) What features could potentially make this visualization misleading?

b) Suppose the underlying data was accessible to you. How would you choose to visualize the data?

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!

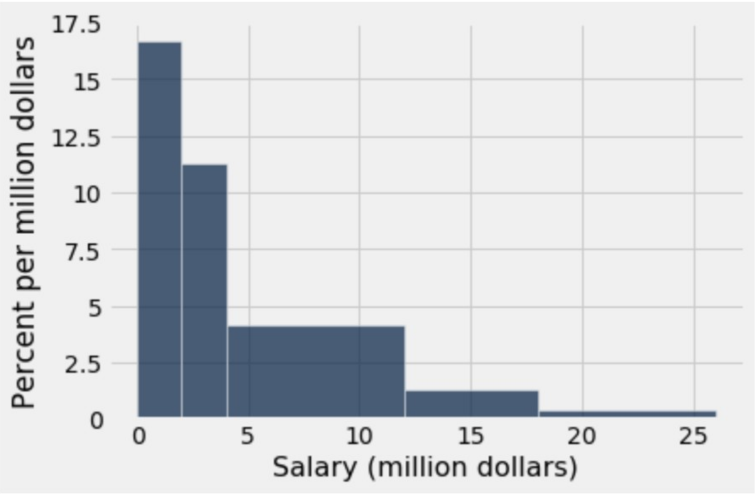
b) What is the height of the bar over the bin 350-550 on the density scale, in the correct units?

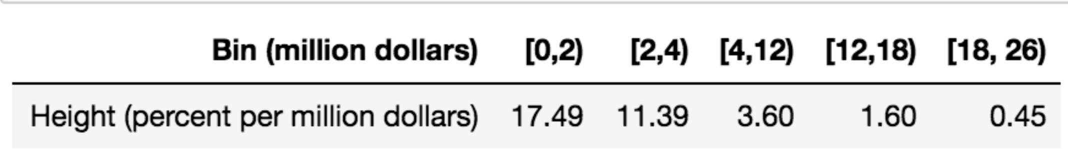
A. 12.5% per student
B. 0.125% per student
C. 0.125% per dollar
D. 12.5% per dollar

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

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

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.

a) Which range contains more players: `[0,4)` or `[4,18)`? How many players are in this range? Explain your choice.