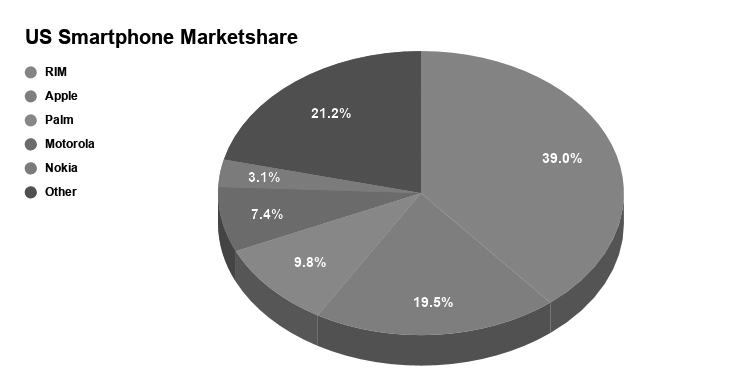
**Data 8 Spring 2020**

**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/>)

****

1. What features could potentially make this visualization misleading?

There are several features that could make this visualization misleading. Firstly, pie charts use angles to represent proportions, and people are generally bad at reading angles. Secondly, the chart is tilted backwards, which makes it seem like the slices in front have larger areas than those in back, making it seem like the Apple slice of the pie is much larger than it really is. In the original graphic, the size of the fonts in each slice are changed so that the slice for the other category has a larger font, making it appear smaller, and the slice for Apple has a smaller font, making it appear larger.

1. Suppose the underlying data was accessible to you. How would you choose to visualize that data?

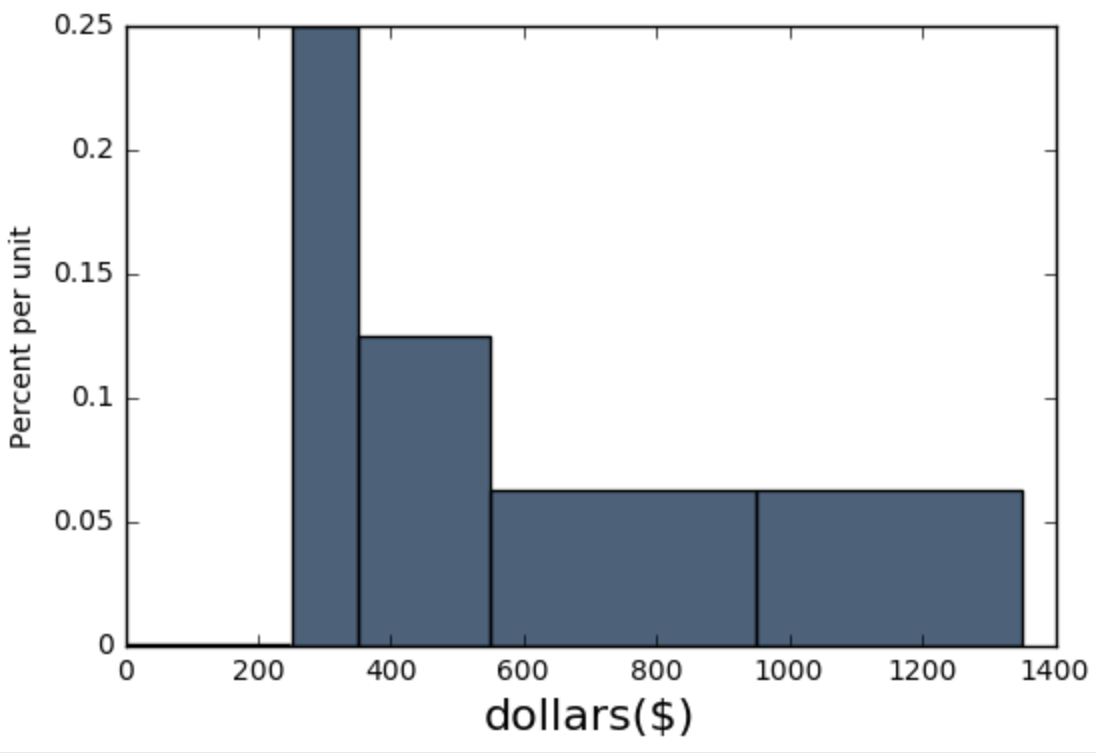
The best visualization for this data would be a simple bar chart with the correct labels. Each bar would have the same width.

**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!

**Solution:**



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

1. 12.5% per student
2. 0.125% per student
3. 0.125% per dollar
4. 12.5% per dollar

**Solution: C**



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

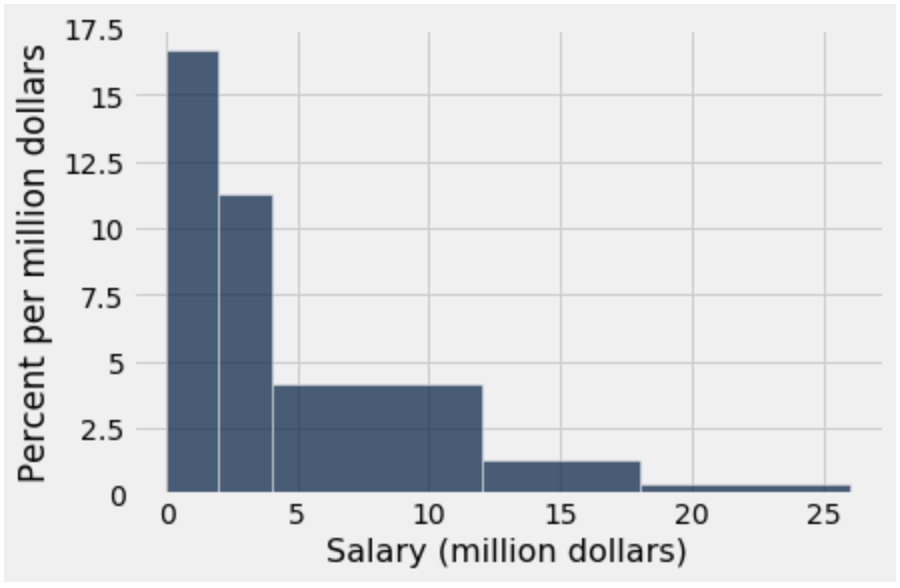
**Solution:**

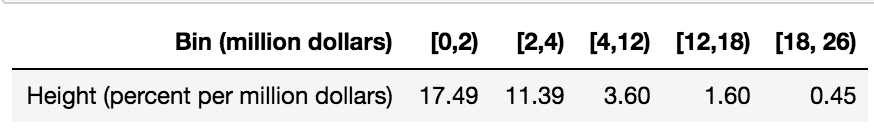
False, each bin contains 25% of the rents, but the bins don’t all have the same height. Moreover, we don’t know if the data is spread evenly within the bins, so we can’t say that the rents are evenly distributed.

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

**Solution:** False, we don’t know the distribution of data within a bin.

**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.

****

****

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 this range? Explain your choice.

**Solution:**

Through calculation, we find that [0,4)has more players, because the area of the bars represent the percent of players, and there is a greater percent of players in the range [0,4).

* Area of the [0,4) range = 2\* 17.49 + 2 \* 11.39 = 57.76
* Area of the [4,18) range = 8 \* 3.60 + 6 \* 1.60 = 38.4

57.76 > 38.4

We cannot know how many players are in the range because we do not know the total number of players.