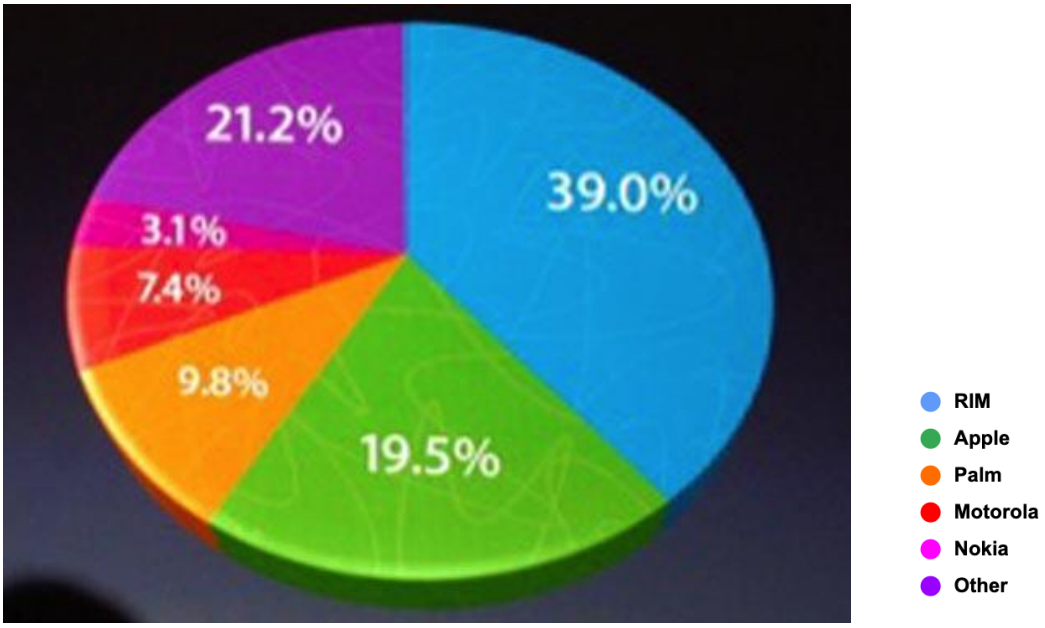
Lab 04: Functions, Visualizations

Data 8 Discussion Worksheet

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

1. The following graphic is 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.



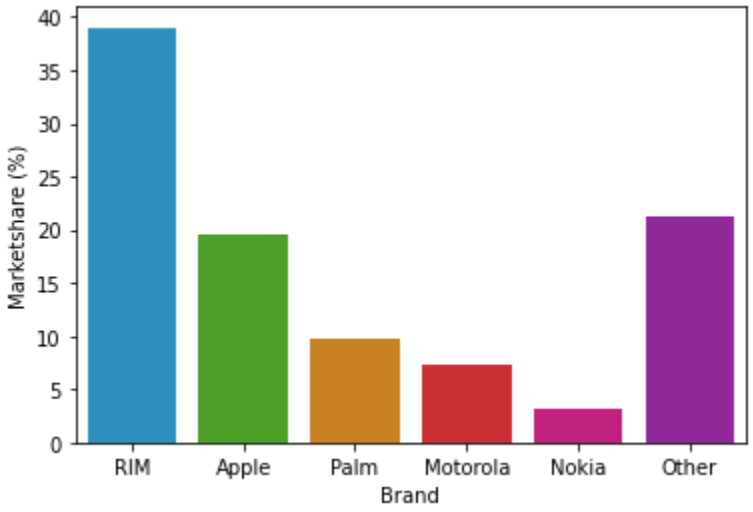
(Source: https://www.wired.com/2008/02/macworlds-iphon/)

a. What features could potentially make this visualization misleading?

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.

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

The best visualization for this data would be a simple bar chart with the correct labels. Each bar would have the same width. Here's an example (courtesy of Ellen Persson):

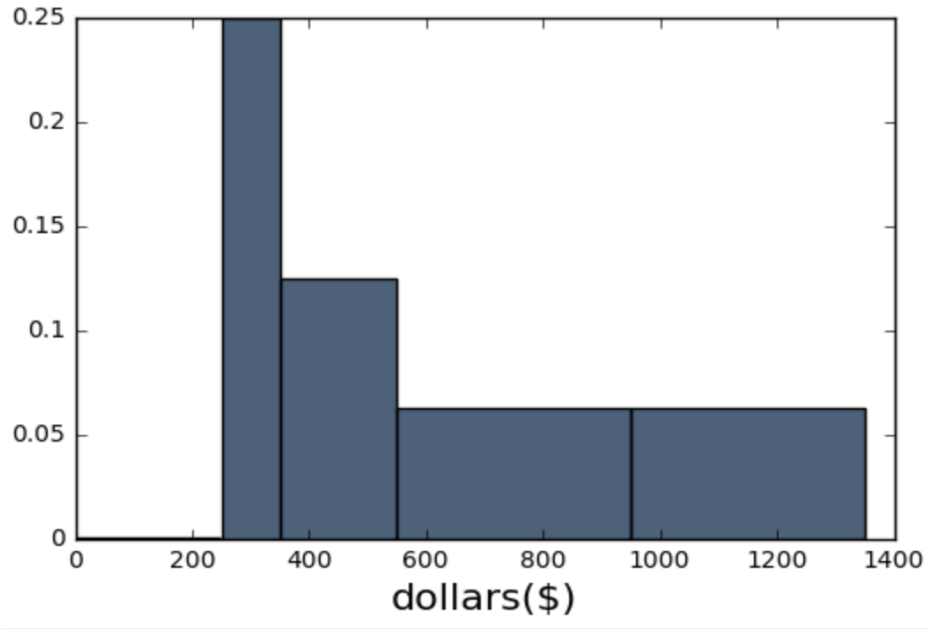


2. The table below shows the distribution of rents paid by students in Boston. The first column consists of ranges of monthly rent, in dollars. Ranges include the left endpoint but not the right. The second column shows the percentage of students who pay rent in each of the ranges.

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, in the correct units?
   A. 12.5% per student
   B. 0.125% per student
   C. 0.125% per dollar
   D. 12.5% per dollar

Solution: C

$$\text{height} = \frac{\text{area}}{\text{width}} = \frac{25\%}{\$550 - \$350} = 0.125\% \text{ per dollar}$$

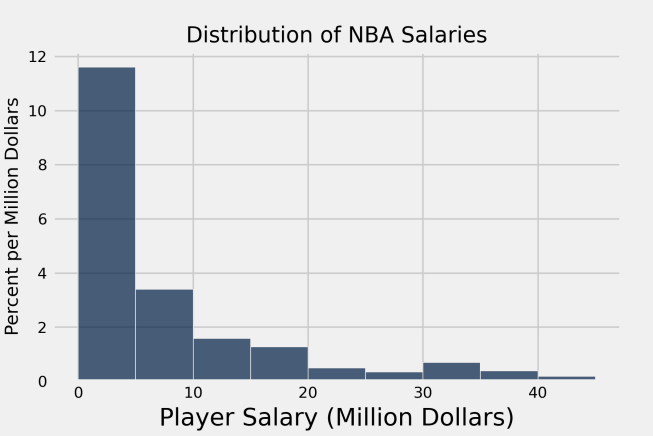
c. **True or False (*Explain*)**: If we combine the [250, 350) and [350, 550) bins together, the height of the new bin would be **greater than** the heights of both of the old bins.

False: When we combine bins together, the height of the new bin is the weighted average of the old bin heights. Thus, the new bin height will be greater than the [350, 550) bin, but less than the [250, 350) bin. If we calculate the new height, it will be:

Area / width = 50% / ((350 - 250) + (550 - 350)) = 0.167% per dollar

3. The table `nba` has a column labeled "`Player Salary`" containing the 2021-22 salaries of 538 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)	Height (Percent per Million Dollars)
[0, 5)	11.61
[5, 10)	3.4
[10, 15)	1.59
[15, 20)	1.28
[20, 25)	0.5
[25, 30)	0.35
[30, 35)	0.7
[35, 40)	0.39
[40, 45)	0.19

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

Which range contains more players: [0,5) or [5,20)? What percentage of players are in this range? Explain your choice. Feel free to use a calculator for your arithmetic calculations.

Solution:

Through calculation, we find that `[0,5)` has more players, because the area of the bars represents the percent of players, and there is a greater percent of players in the range `[0,5)`.

- Area of the `[0,5)` range = 5 * 11.61 = 58.05% = 312.3 players

- Area of the $[5, 20)$ range = 5 * 3.4 + 5 * 1.59 + 5 * 1.28 = 17 + 7.95 + 6.5 = 31.35% = 168.7 players

313 players > 169 players