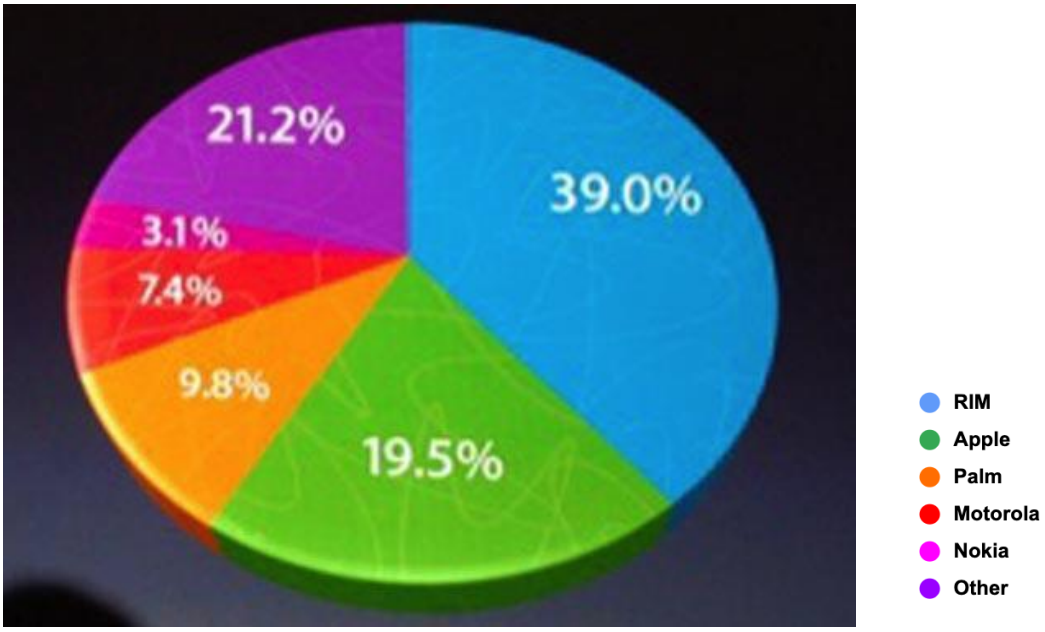
Lab 04: Functions, Visualizations

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.



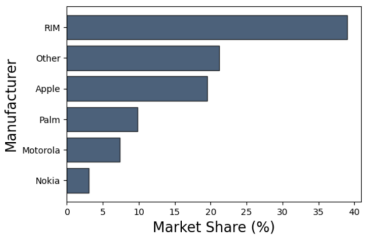
(Source: https://www.wired.com/2008/02/macworlds-iphon/)

a. What features could potentially make this visualization misleading?

Normally, pie charts are organized from most proportionality to least, which is not the case here seeing as 2nd and 3rd place are on either side of 1st place (RIM, light blue). Additionally, one needs the key to read this pie chart, whereas many pie charts are structured such that one would not need a legend just to read them. Finally, the image is blurry, and humans are bad at spatially estimating non-rectangular proportions where height/width is not the only determining factor of area. The green looks much larger than the purple.

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

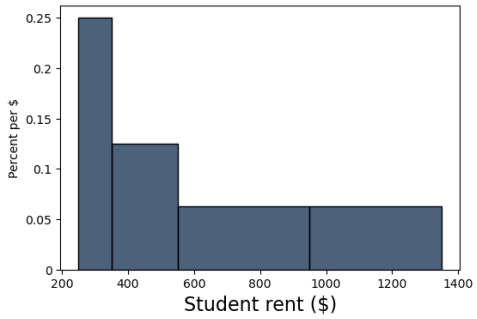
I would choose to visualize this data in a bar graph that has axes/individuals labeled so that area is much easier to see, and it would be sorted in descending order so that it is easy to see whose is greater.



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!

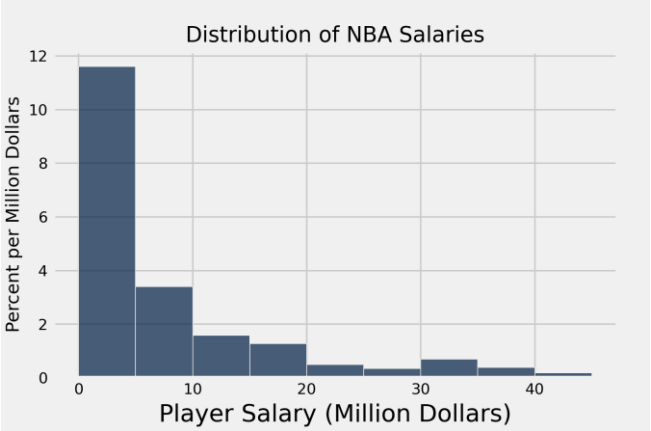


b. What is the height of the bar over the bin 350-550, in the correct units?
   A. 12.5% per student
   B. 0.125% per student
   C. 0.125% per dollar
   D. 12.5% per dollar

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: the height of the combined bin would be about 0.17 % per $, which is less than the current height of the [250, 350) bin and greater than the current height of the [350, 550) bin. This makes sense because the first bin is relatively crowded while the second bin is less crowded. It makes sense that taking up the same area would require the resulting rectangle to be shorter than the tall bin and taller than the short bin. This is not always the case, but giving more width is a way to reduce density and therefore height from a crowded bin, meaning that the tall bin would not be taller by adding more width unless being combined with a more dense (taller) bin.

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.

The [0, 5) bin has 11.61 percent per million dollars, but multiplying that by the millions of dollars in the bin (roughly 5-0=5), one is left with the percent of players in that bin which is 58.05% of players on this table whose salaries fall in the [0, 5) range. Meanwhile, the percentage of players in the [5, 20) bins can be calculated by 5*(3.4+1.59+1.28), which is 31.35% of players in this table. Therefore, because the [0, 5) bin has a higher percentage of players on the table encompassed (58.05% > 31.35%), more players have salaries in the [0, 5) bin than in the [5, 20) range. These calculations are made by the fact that the y-axis represents percent of players in the table per million dollars, while the x-axis represents millions of dollars of player salary. Therefore, multiplying percent of players in that bin per million dollars by millions of dollars in each bin simply results in the percentage of players from the graph in that bin.