Chapter 3: Displaying Categorical and Numerical Data

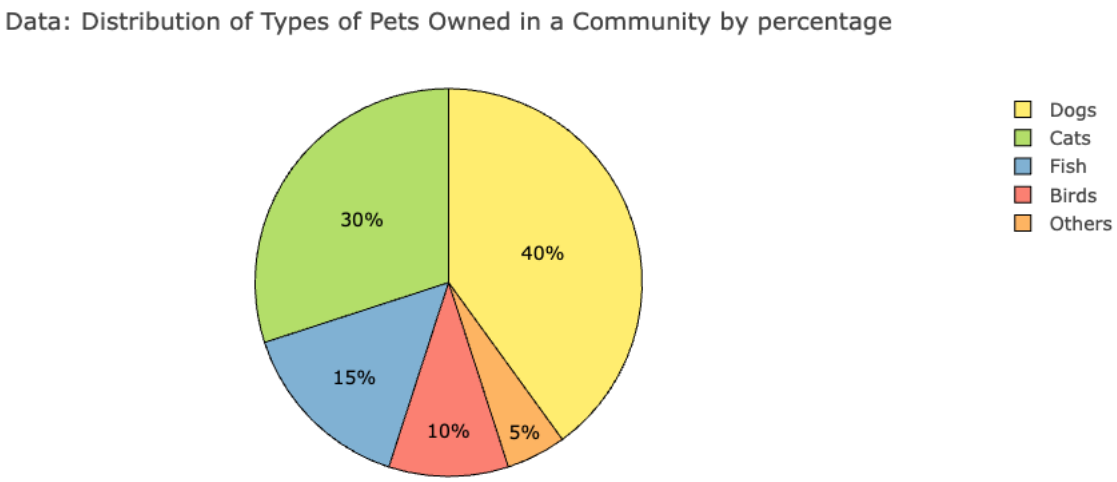
1 Organizing Categorical Data

Once we have a data set, we next need to organize and display the data in a way that helps us see patterns. With categorical variables, we are usually concerned with knowing how often a particular category occurs in our sample. We then (usually) want to compare how often a category occurs for one group with how often it occurs for another (liberal/conservative, man/woman). The two most common ways to display a categorical data are pie-chart and barchart.

1.1 Pie-chart:

A pie chart is the most common technique to represent categorical data. Each slice of a pie represents a category's contribution to the whole, with the size proportional to the quantity it represents. Pie charts work best with a small number of categories and are most useful for simple, straightforward data visualizations. However, they are not suitable for complex data or showing trends over time and can be difficult to interpret when slices are similar in size. Overusing pie charts or using them inappropriately can lead to misinterpretation.

For example, let's visualize the distribution of pet types in a community by percentage, where the data shows 40% dogs, 30% cats, 15% fishes, 10% birds, and 5% others.



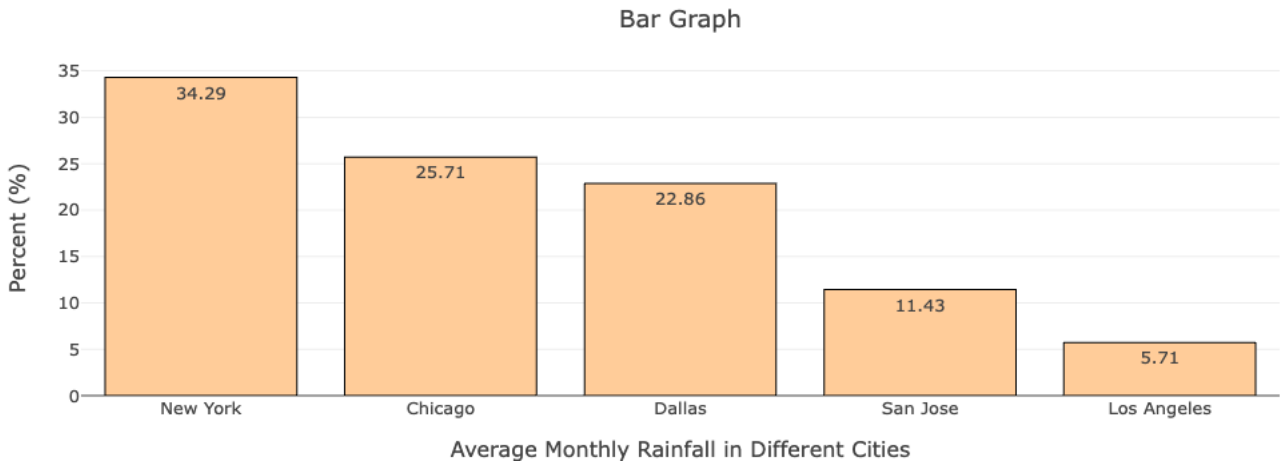
1.2 Bar charts

A bar chart is a graphical representation of data where the length or height of each bar is proportional to the value it represents. Bar charts are commonly used to compare quantities across different categories, making them ideal for visualizing discrete data. To create a bar chart, data is first collected and organized into categories, with each category assigned a corresponding value. These values are then represented by bars, either horizontally or vertically, with the axis providing a scale for measurement. Bar charts should be used when comparing different groups or tracking changes over time, especially when changes are larger. However, it is important to avoid common mistakes such as using too many categories, which can make the chart cluttered and hard to read, or not starting the y-axis at zero, which can distort the visual representation of the data. Additionally, bars should be of equal width and spaced evenly to ensure clarity and consistency.

For example, a bar chart could be used to display the average monthly rainfall in different cities for the data below.

Average Monthly Rainfall in Selected Cities

City	Average Monthly Rainfall (mm)
New York	120
Los Angeles	20
Chicago	90
San Jose	40
Dallas	80

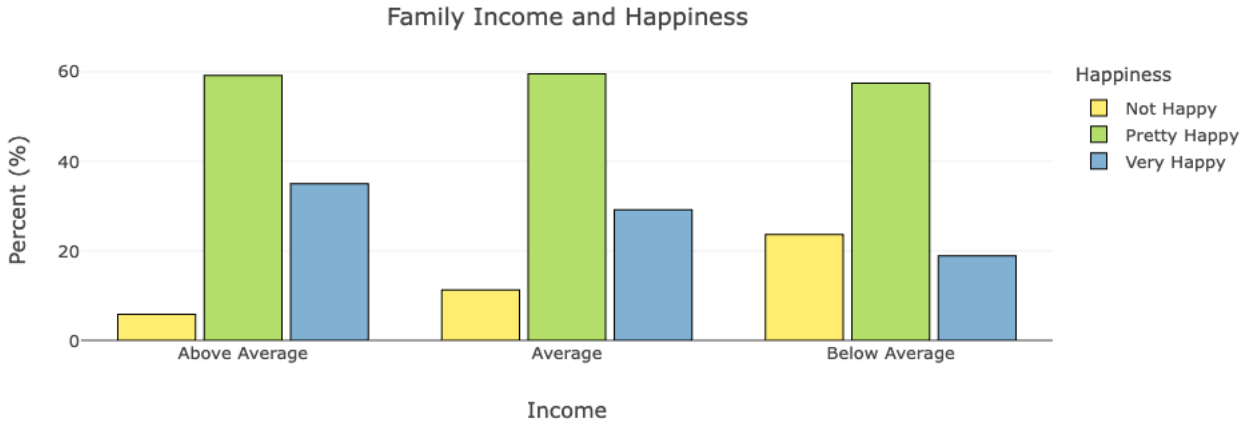


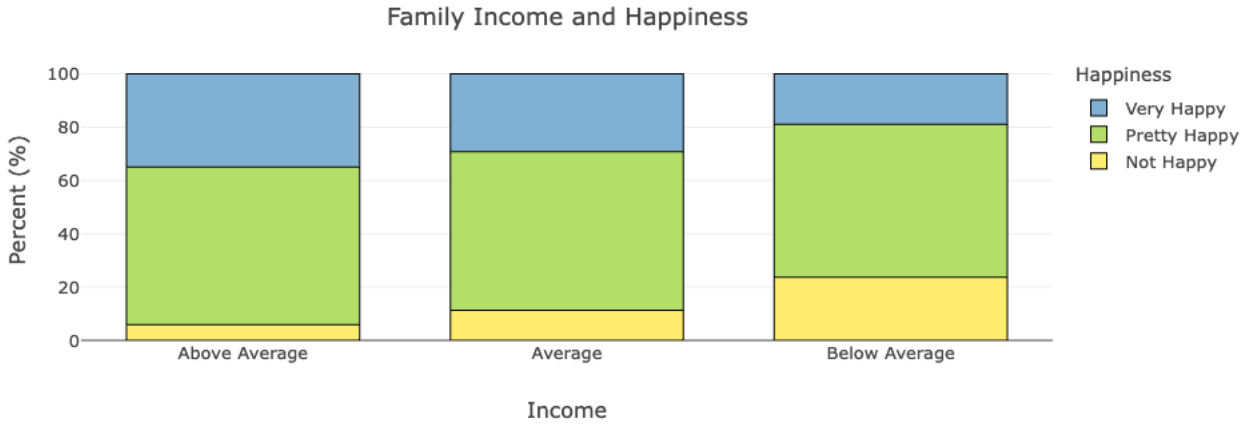
1.3 Side-by-side bar graphs and stacked bar graphs

A contingency table is a table that displays the results of two categorical variables simultaneously. It is also called a two-way table. Side-by-side bar graphs present data for two categorical variables from more than one group by creating two bars on the chart for each group - one bar for each variable. In the stacked bar graph, each bar represents the responses of one group. The height of each color within that bar represents a percentage of a particular response, and the combination of all colors represents the total (100%) of all responses within that group.

Example: The following contingency table displays how happiness varies with income levels. Let us represent this data in both side-by-side bar chart and a stacked bar chart.

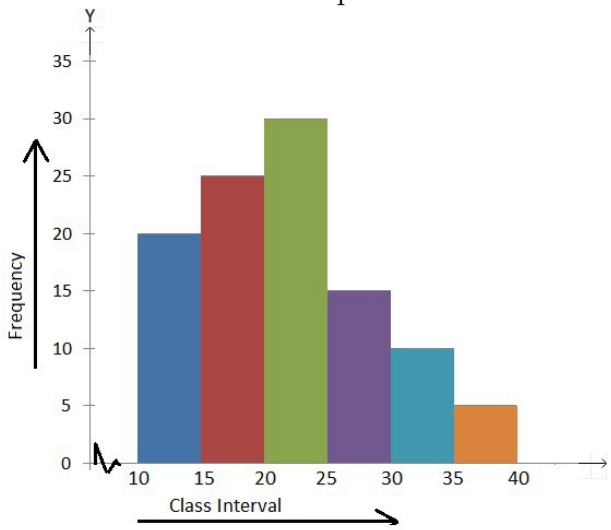
Income Level	**Happiness**		
	Not Happy	**Pretty Happy**	**Very Happy**
Above Average	21	213	126
Average	96	506	248
Below Average	143	347	114





2 Histograms

Histograms display your data by putting observations into intervals called bins of equal width. The height of the bin is the frequency. Frequency is the number of times a value was recorded when data was collected. The x axis be your independent numerical variable. And the y axis will have the various frequencies at which those variables occur.



3 Understanding quantitative data

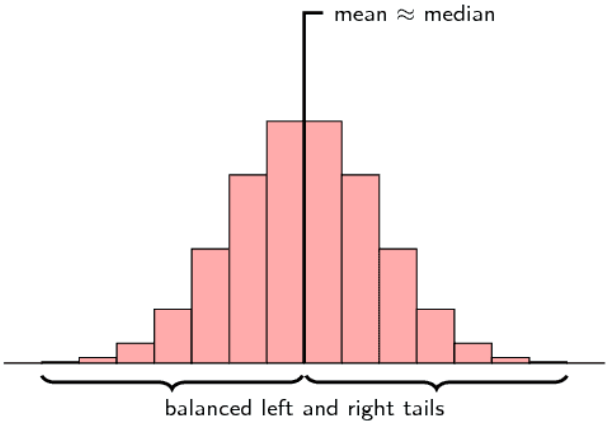
Four most common questions statisticians may ask themselves about their data are: **What is the shape of my data-set? Do I notice anything unusual? What is the center of my data-set? What is the spread of my data-set?** These questions can be answered when we study the distribution of our sample of data, and one of the most common ways to do it is to create a histogram.

4 Shape

Let's first begin with the shape. When you want to analyze the **shape** of a distribution, you want to ask yourself the following questions:

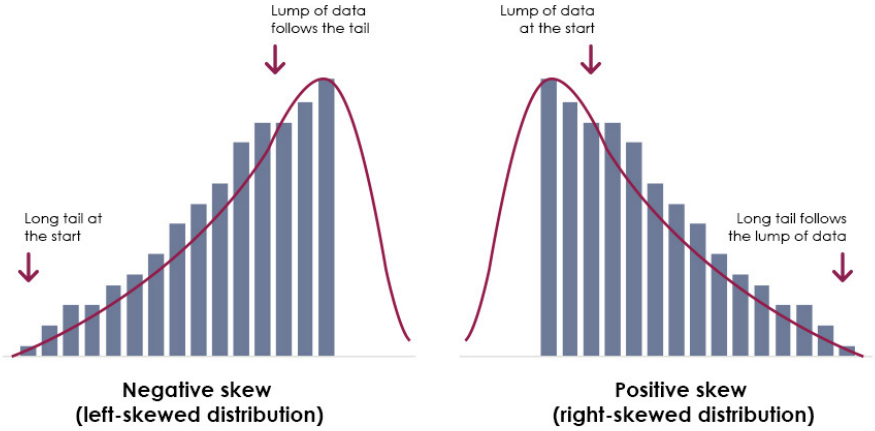
1. Is the distribution symmetric or skewed?
2. How many major mounds appear? None? One? Two? Several?

To answer these questions we are going to need to see a lot of examples. Let's first take a look at a **symmetrical histogram**. An important feature about the symmetrical histogram below is that the typical value falls around the center and the data is almost symmetrically distributed on both sides, hence the name.

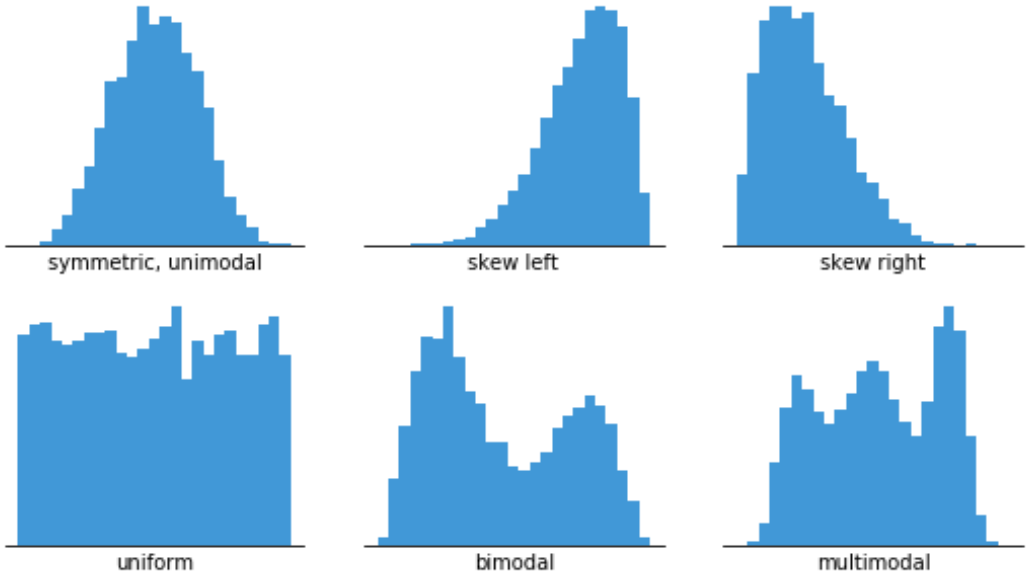


Now let's observed a skewed histogram. There are two types of skewed histograms, **right skewed** and **left skewed**.

- You will see in a **left skewed** distribution, lots of data is towards the right side of the histogram, which leaves the tail end to the left.
- In a **right skewed** distribution, lots of data is towards the left side of the histogram, which leaves the tail end to the right.



Now to answer question 2, you can also describe a histogram by how many peaks it has. A distribution may be uniform, have one major peak, two, or several. Those are shown below as well as the other shapes summarized.



Examples: What shape would you expect to see in a histogram of the following data sets?

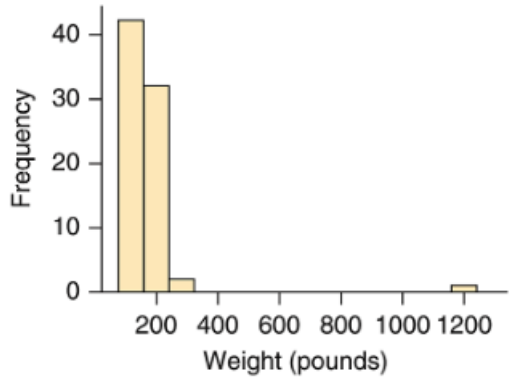
- GPA of college students - *Left Skewed*
- SAT scores - *Symmetric*
- Last digit of Social Security numbers for a random sample of students - - *Uniform*
- Income of USA residents - *Right Skewed*

What about modality in these data-sets?

- Morning and evening sales at a restaurant - *Bimodal*
- Men and women's heights - *Bimodal*

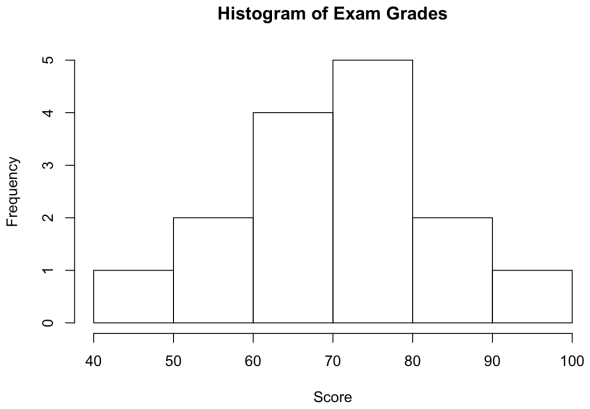
5 Do I notice anything unusual?

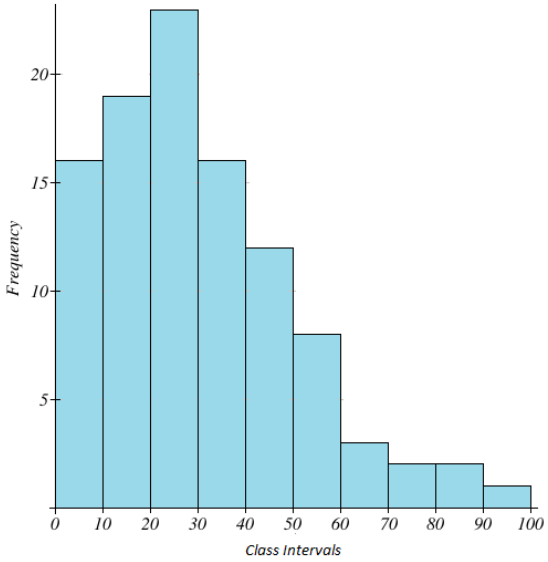
Some extremely large or small data values that don't fit the pattern of the rest of the data are called outliers. Outliers can be errors (such as typos) or be genuine. Genuine outliers are unusually interesting data values!
The histogram below shows one outlier on the right.



6 Center

The **center** of a distribution is like the typical amount. It is usually where you find most of your data. Based on the histograms below, what is the **shape** and what is the **center** or typical value?

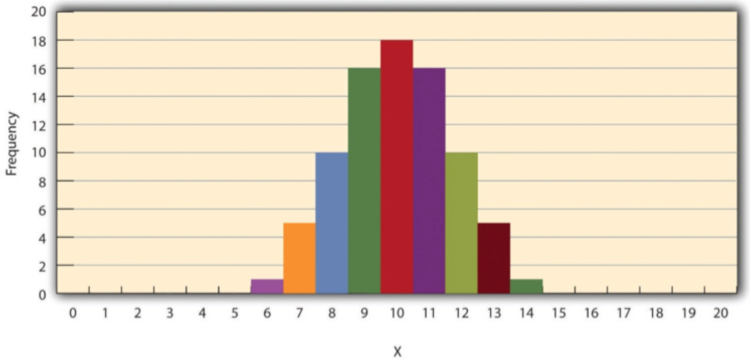


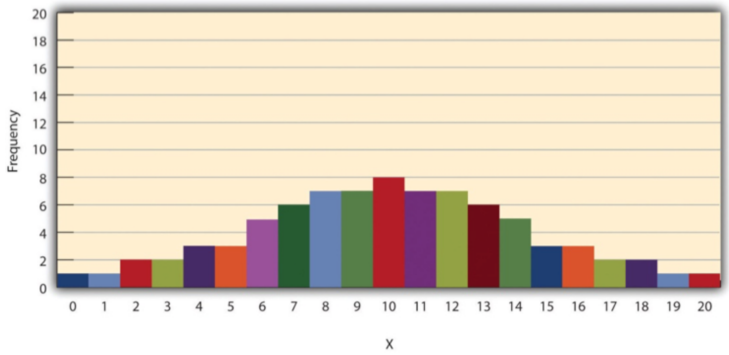


The histogram on the left has it's center between 70-75 and the histogram on the right has center somewhere between 28-33.

7 Variability/Spread

We describe the amount of variation in our data by looking at how much **spread** there is along the x axis. Based on the histograms below, which has more variability? Why?





Since the histogram on the bottom has more horizontal spread, it has more spread/variability in the data.