Section 001
CDS 101 *Introduction to Computational and Data Sciences*

Class 6: Data visualization II

February 8, 2018

General

Annoucements

- Visualization mini-assignment due on February 9 at 5:00pm
- Reading for next class: *R for Data Science* - chapter 3, section 3.7 through to the end of section 3.10
- I will be uploading Homework 1 on visualizing a dataset by Friday, will be due on February 23rd.
- I am being lenient on late submissions of mini-assignments during these first three weeks of class, anything due after today will be subject to the standard penalties stated in the syllabus

How to describe visualizations

A taxonomy for data graphics

- We can break visualizations down into four basic elements:
  - Visual cues
  - Coordinate system
  - Scale
  - Context

Visual cues

- These are the building blocks of any given visualization.
- Identify 9 separate visual cues.

Cues 1–9

1. **Position** (numerical) where in relation to other things?
2. **Length** (numerical) how big (in one dimension)?
3. **Angle** (numerical) how wide? parallel to something else?
4. **Direction** (numerical) at what slope? In a time series, going up or down?
5. **Shape** (categorical) belonging to which group?
6. **Area** (numerical) how big (in two dimensions)?
7. **Volume** (numerical) how big (in three dimensions)?
8. **Shade** (either) to what extent? how severly?
9. **Color** (either) to what extent? how severly? Beware of red/green color blindness.

Coordinate systems

1. **Cartesian** This is the familiar (x, y)-rectangular coordinate system with two perpendicular axes
2. **Polar**: The radial analog of the Cartesian system with points identified by their radius ρ and angle θ
3. **Geographic**: Locations on the curved surface of the Earth, but represented in a flat two-dimensional plane

Scale

1. **Numeric**: A numeric quantity is most commonly set on a *linear*, *logarithmic*, or *percentage* scale.

2. **Categorical**: A categorical variable may have no ordering or it may be *ordinal* (position in a series).

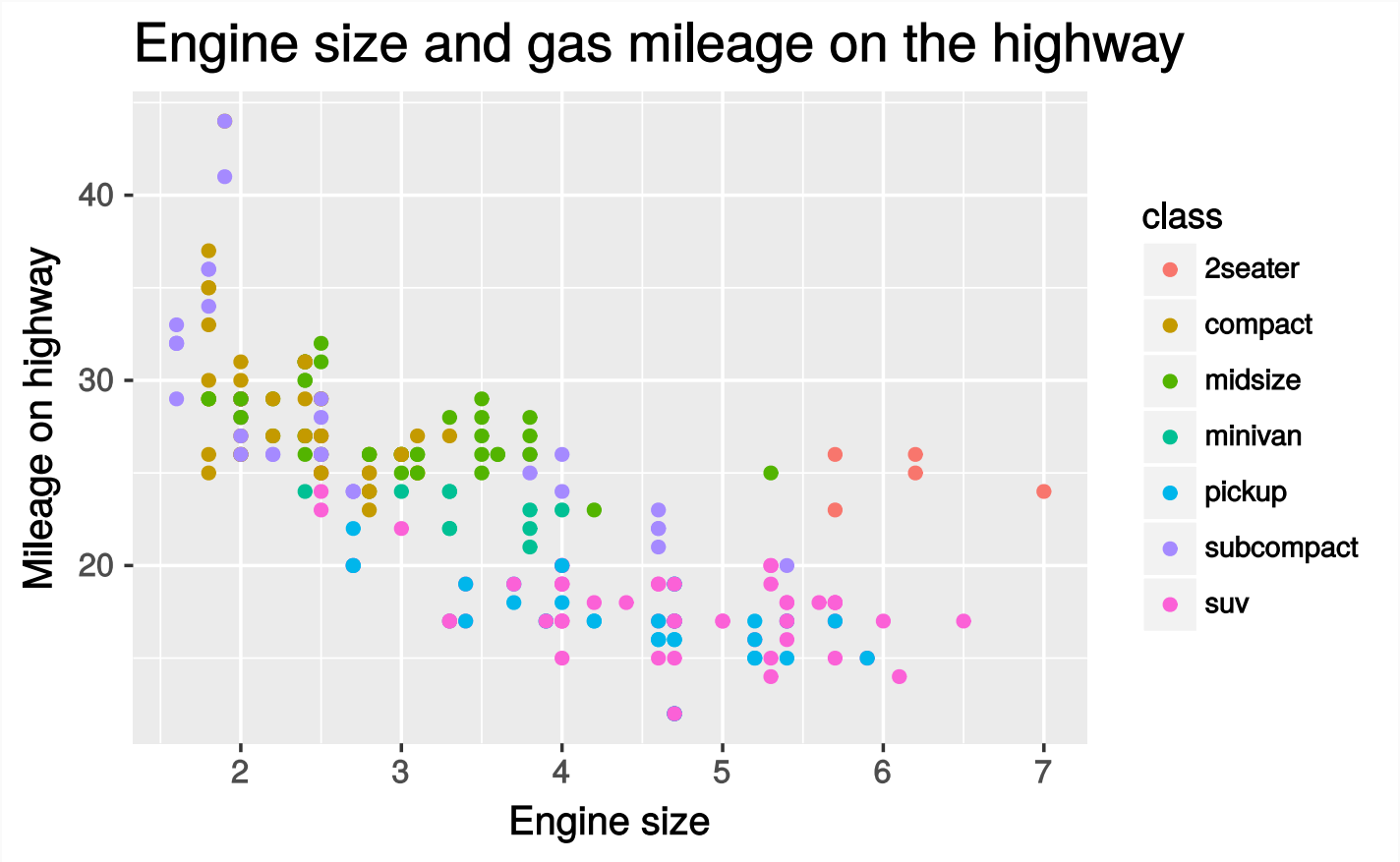
3. **Time**: A numeric quantity with special properties. Because of the calendar, it can be specified using a series of units (year, month, day). It can also be considered cyclically (years reset back to January, a spring oscillating around a central position).

Context

- Annotations and labels that draw attention to specific parts of a visualization.
  - Titles, subtitles
  - Axes labels that depict scale (tick mark labels) and indiciate the variable
  - Reference points or lines
  - Other markups such as arrows, textboxes, and so on (it's possible to overdo these)

Example plot

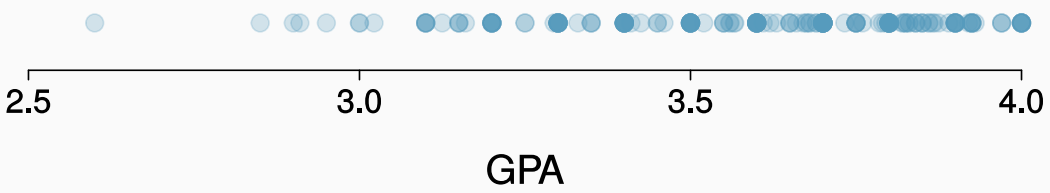
How many of the previous elements can you identify in this plot?



Examining numerical data

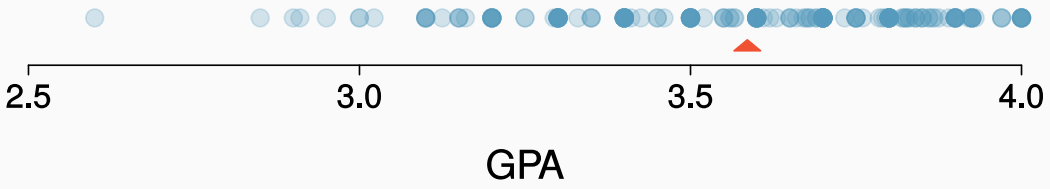
Dot plots

Useful for visualizing one numerical variable. Darker colors represent areas where there are more observations.



How would you describe the distribution of GPAs in this data set? Make sure to say something about the center, shape, and spread of the distribution.

Dot plots & mean



- The **mean**, also called the **average** (marked with a triangle in the above plot), is one way to measure the center of a **distribution** of data.
- The mean GPA is 3.59.

Mean

- The **sample mean**, denoted as $\bar{x}$, can be calculated as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

  where $x_1, x_2, \cdots, x_n$ represent the **n** observed values.

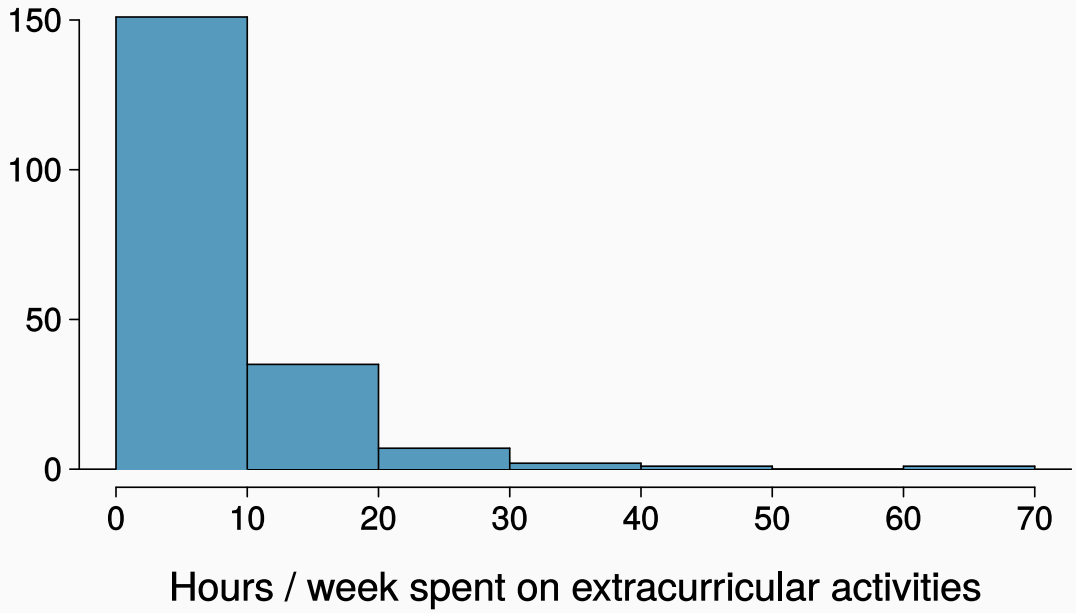
- The **population mean** is also computed the same way but is denoted as μ. It is often not possible to calculate μ since population data are rarely available.

- The sample mean is a **sample statistic**, and serves as a **point estimate** of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population), it is usually a pretty good estimate.

Histograms and shape

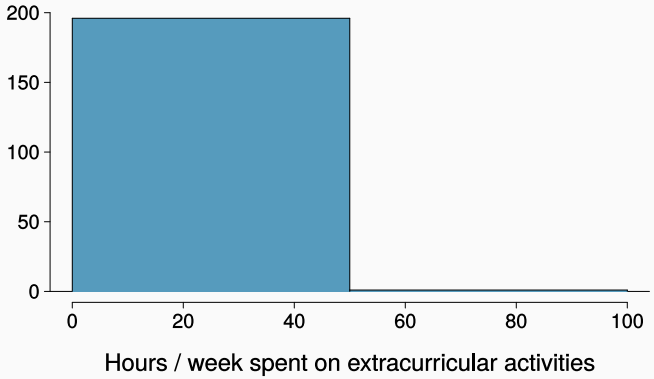
Histograms — Extracurricular hours

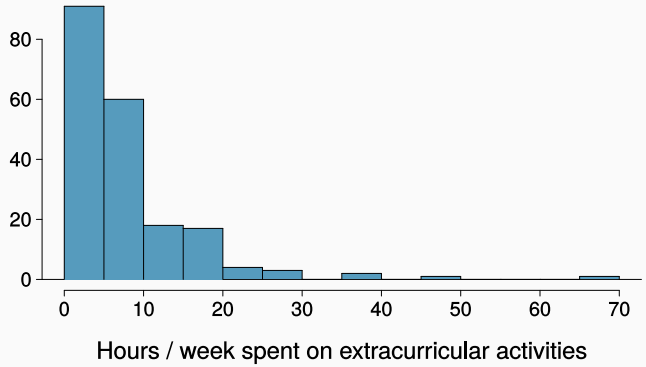
- Histograms provide a view of the **data density**. Higher bars represent where the data are relatively more common.
- Histograms are especially convenient for describing the **shape** of the data distribution.
- The chosen **bin width** can alter the story the histogram is telling.

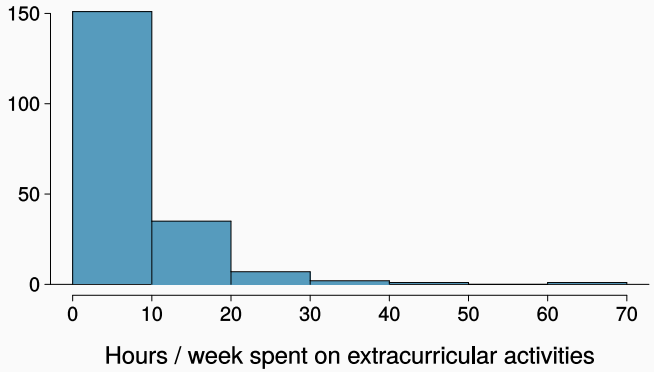


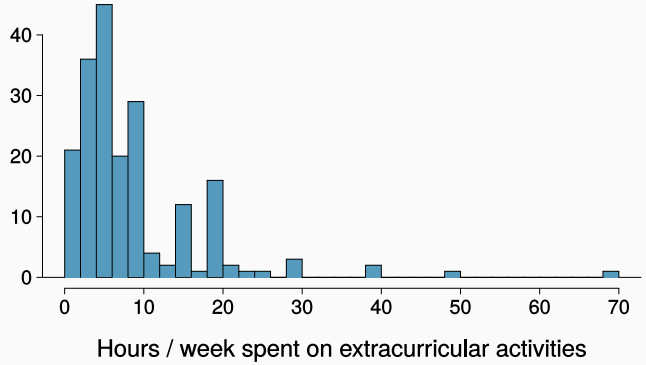
Bin width

Which one(s) of these histograms are useful? Which reveal too much about the data? Which hide too much?



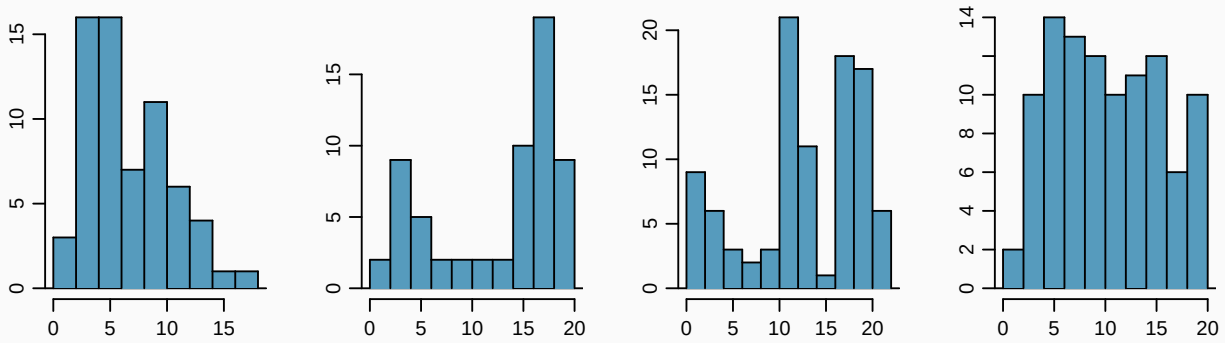






Shape of a distribution: modality

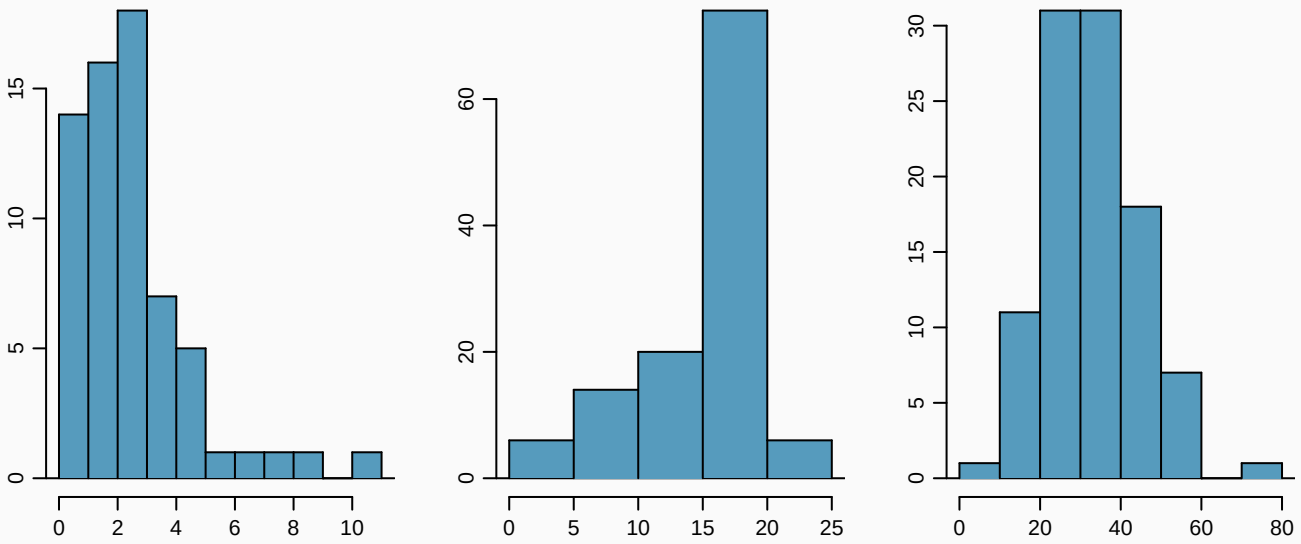
Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?



Note: In order to determine modality, step back and imagine a smooth curve over the histogram – imagine that the bars are wooden blocks and you drop a limp spaghetti over them, the shape the spaghetti would take could be viewed as a smooth curve.

Shape of a distribution: skewness

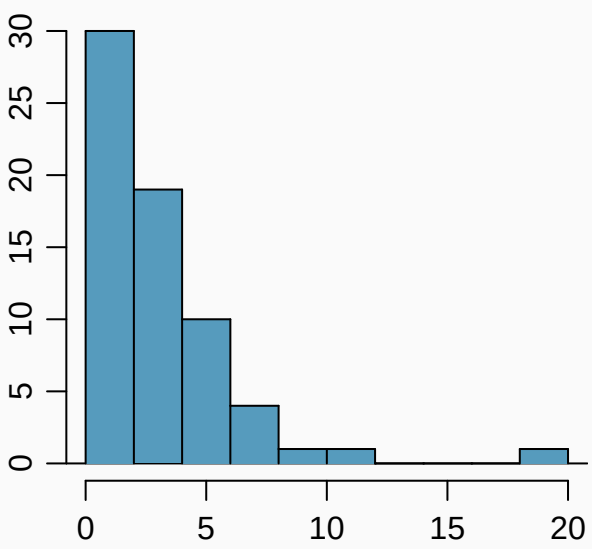
Is the histogram *right skewed*, *left skewed*, or *symmetric*?

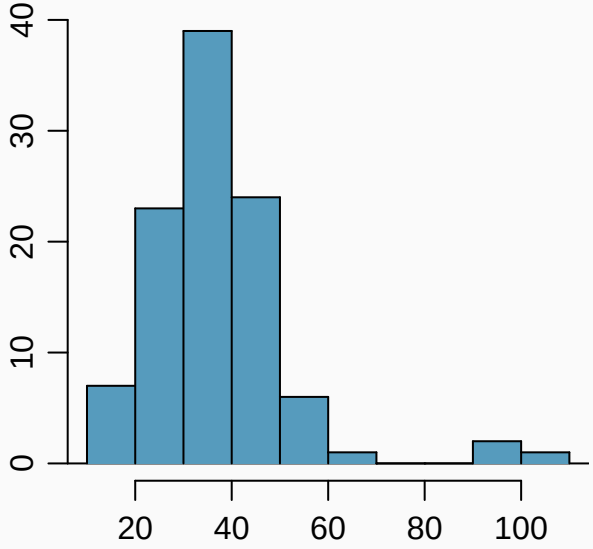


Note: Histograms are said to be skewed to the side of the long tail.

Shape of a distribution: unusual observations

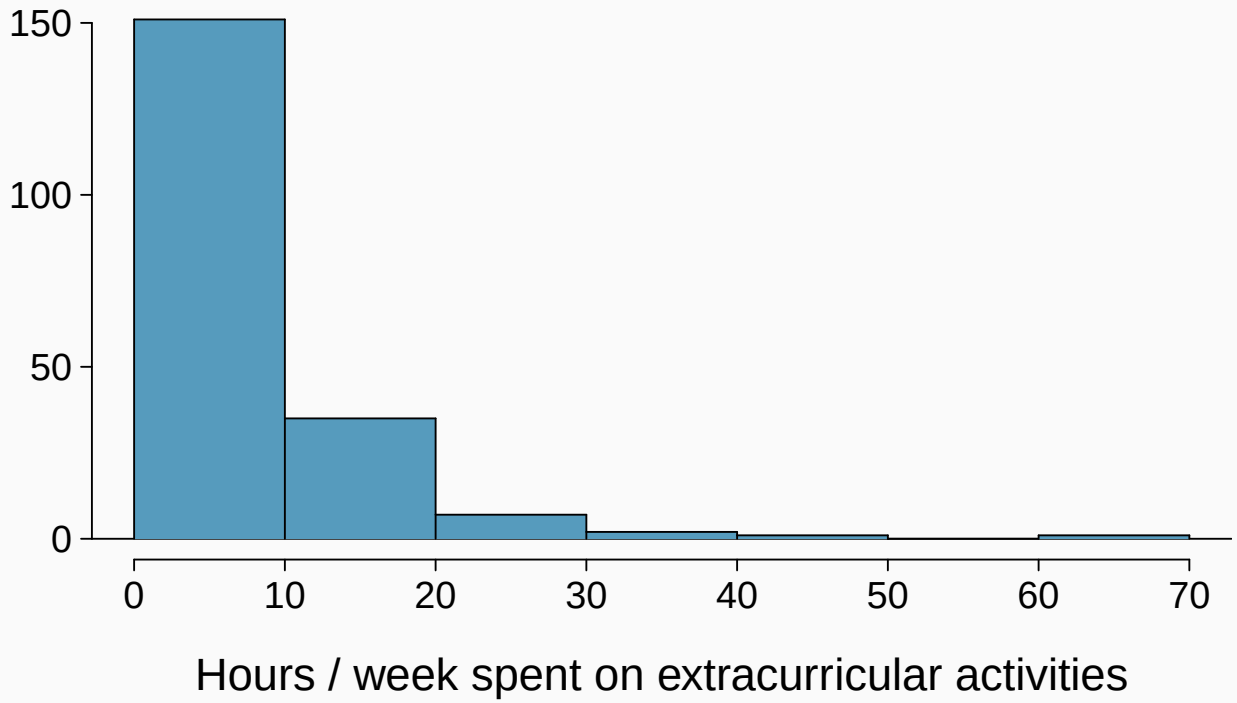
Are there any unusual observations or potential *outliers*?





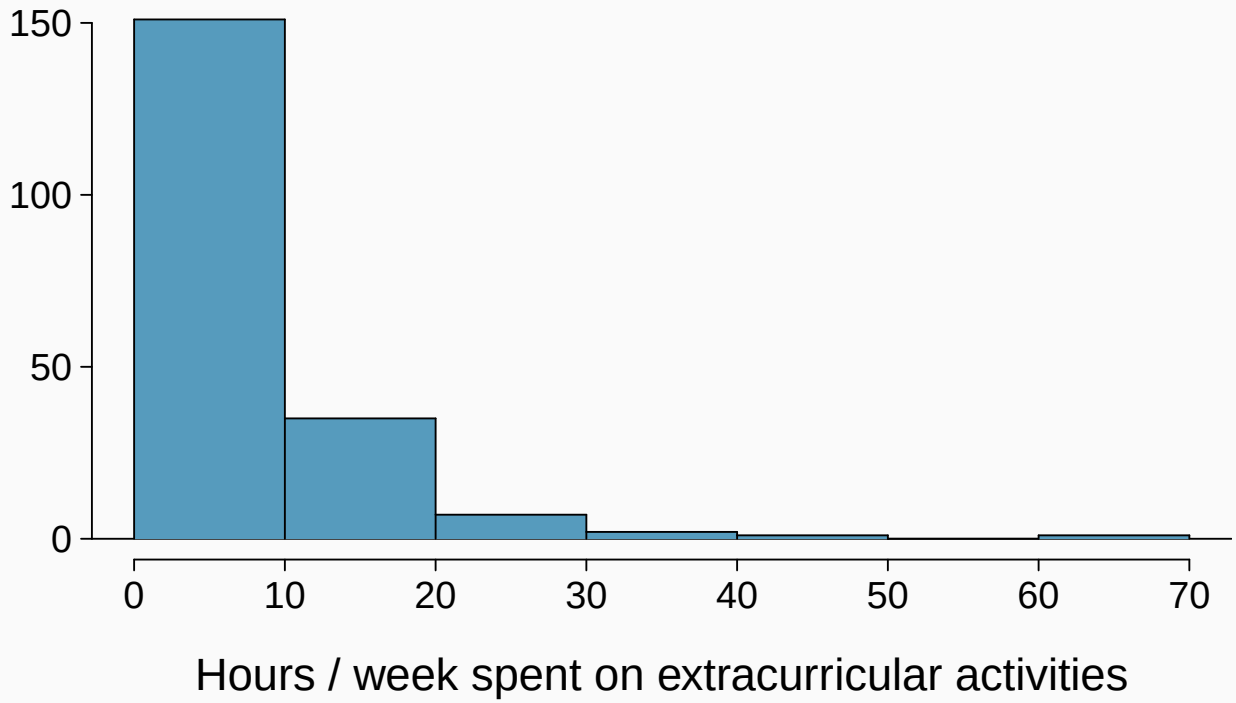
Extracurricular activities

How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?



Extracurricular activities

How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?



Unimodal and right skewed, with a potentially unusual observation at 60 hours/week.

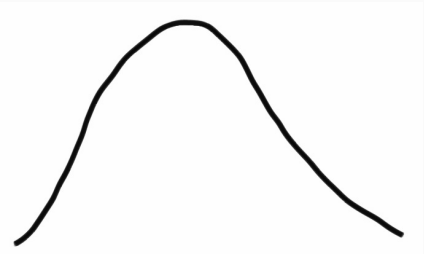
Commonly observed shapes of distributions

- modality

Commonly observed shapes of distributions

- modality

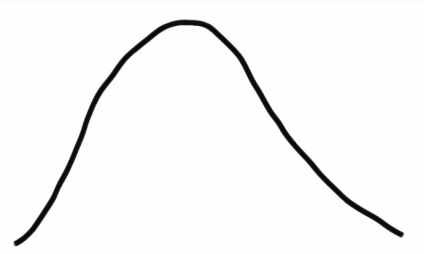
unimodal



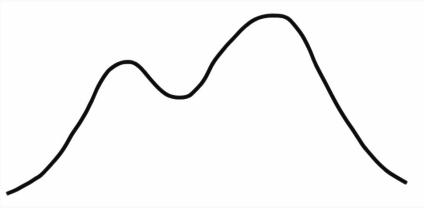
Commonly observed shapes of distributions

- modality

unimodal



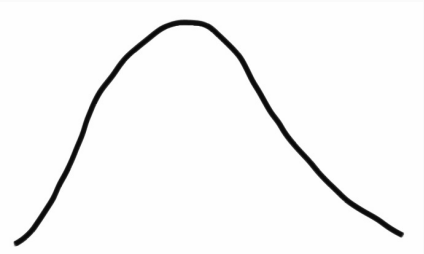
bimodal



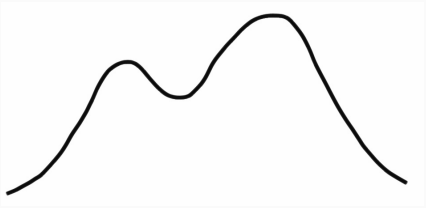
Commonly observed shapes of distributions

- modality

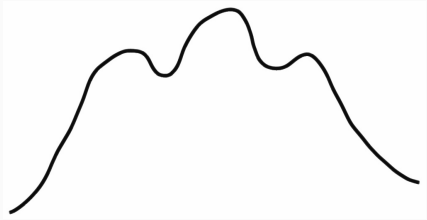
unimodal



bimodal



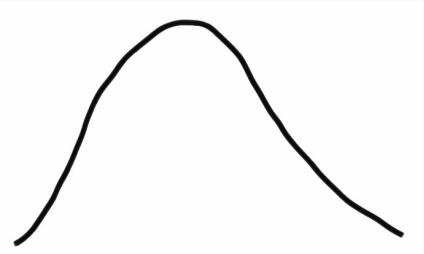
multimodal



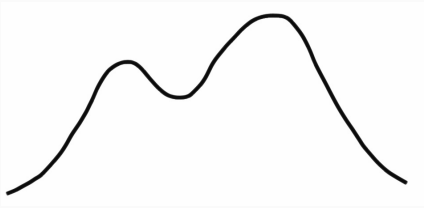
Commonly observed shapes of distributions

- modality

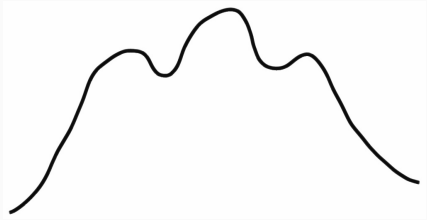
unimodal



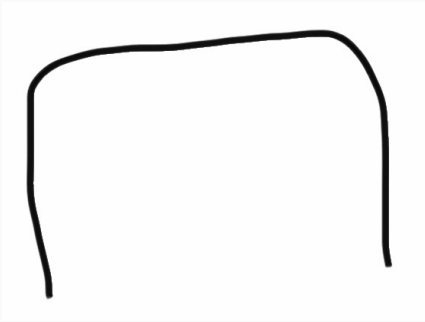
bimodal



multimodal



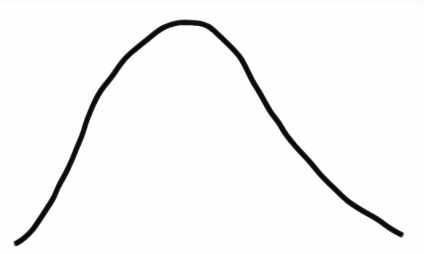
uniform



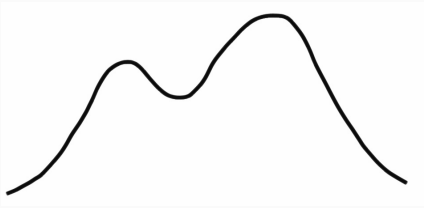
Commonly observed shapes of distributions

- modality

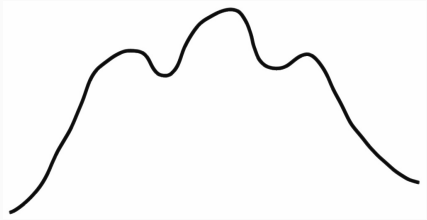
unimodal



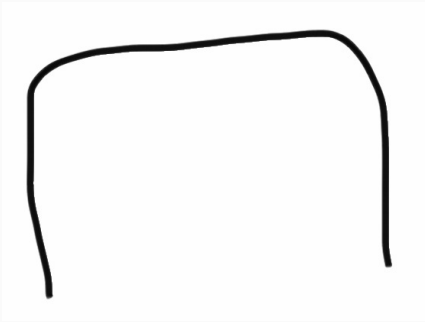
bimodal



multimodal



uniform

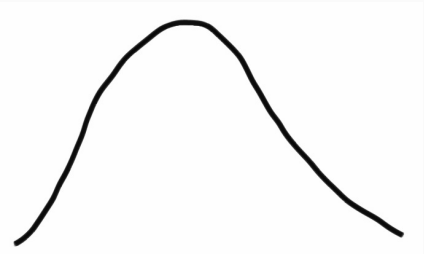


- skewness

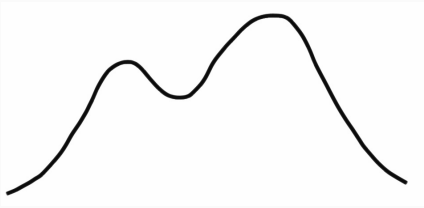
Commonly observed shapes of distributions

- modality

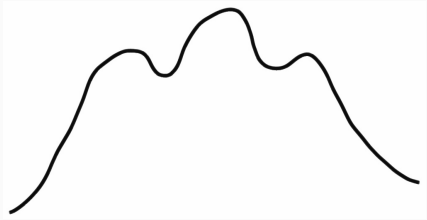
unimodal



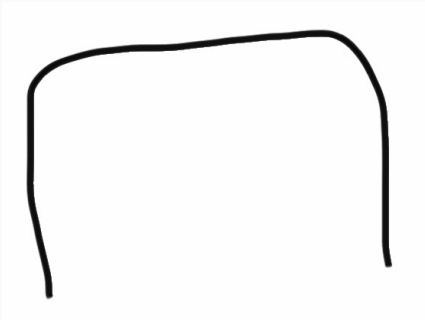
bimodal



multimodal

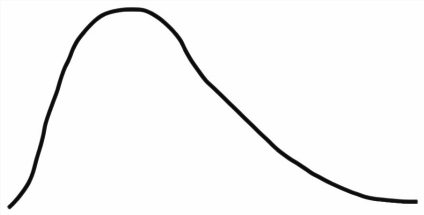


uniform



- skewness

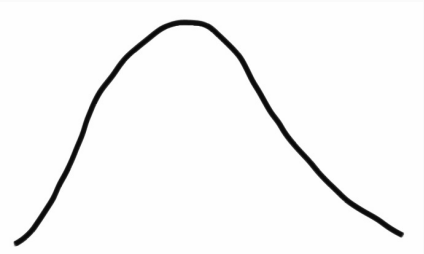
right skew



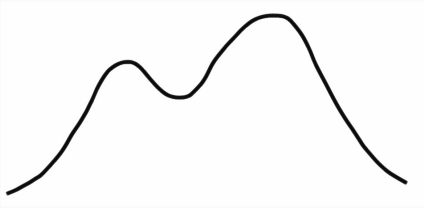
Commonly observed shapes of distributions

- modality

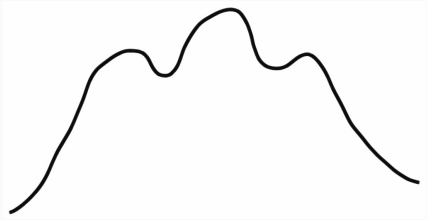
unimodal



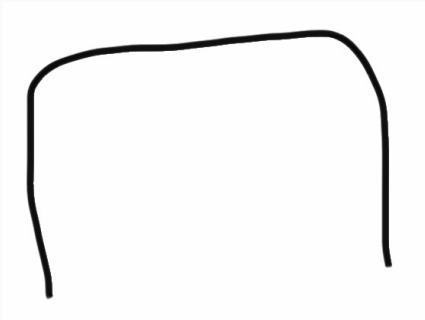
bimodal



multimodal

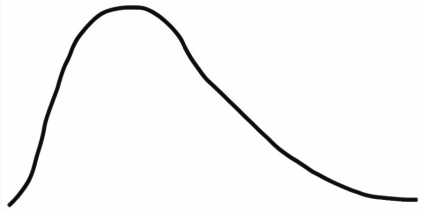


uniform

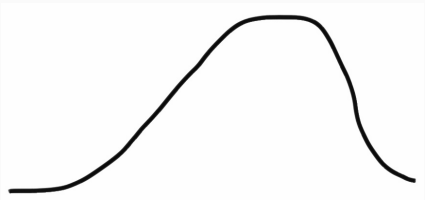


- skewness

right skew



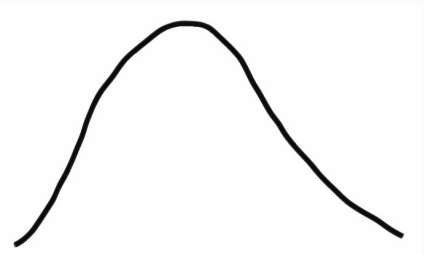
left skew



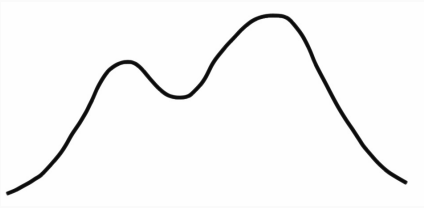
Commonly observed shapes of distributions

- modality

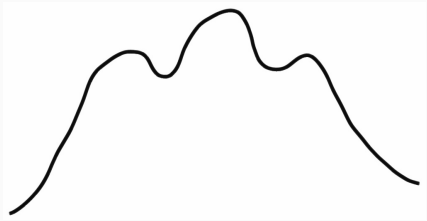
unimodal



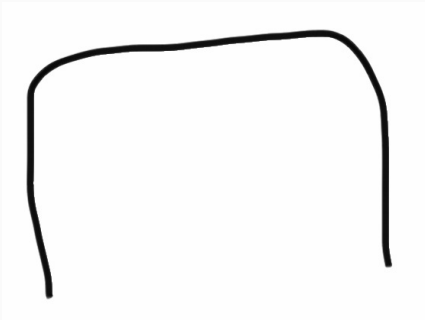
bimodal



multimodal

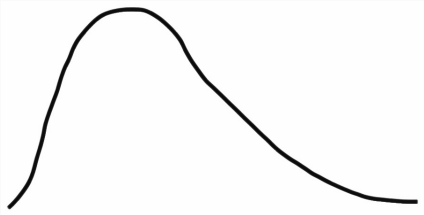


uniform

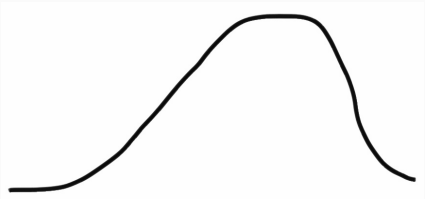


- skewness

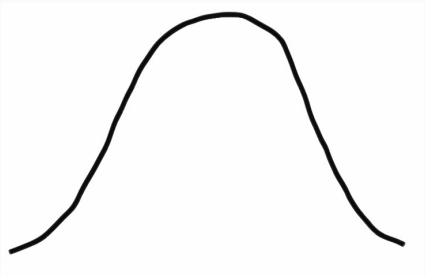
right skew



left skew



symmetric



Practice

Which of these variables do you expect to be uniformly distributed?

(a) weights of adult females

(b) salaries of a random sample of people from North Carolina

(c) house prices

(d) birthdays of classmates (day of the month)

Practice

Which of these variables do you expect to be uniformly distributed?

(a) weights of adult females

(b) salaries of a random sample of people from North Carolina

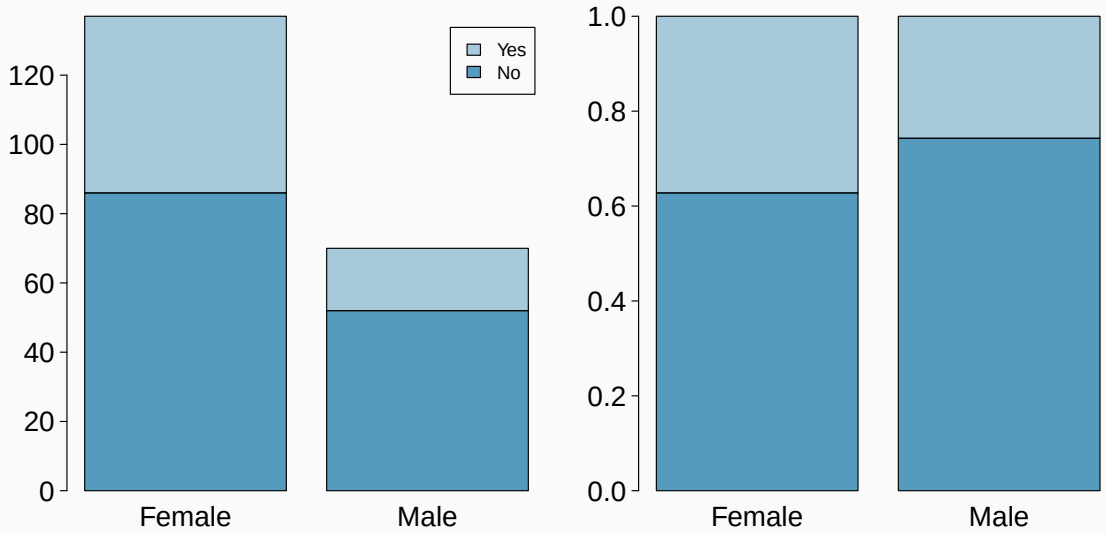
(c) house prices

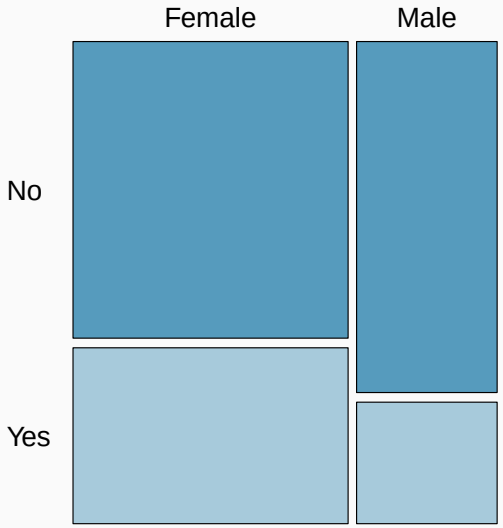
(d) *birthdays of classmates (day of the month)*

Considering categorical data

Segmented bar and mosaic plots

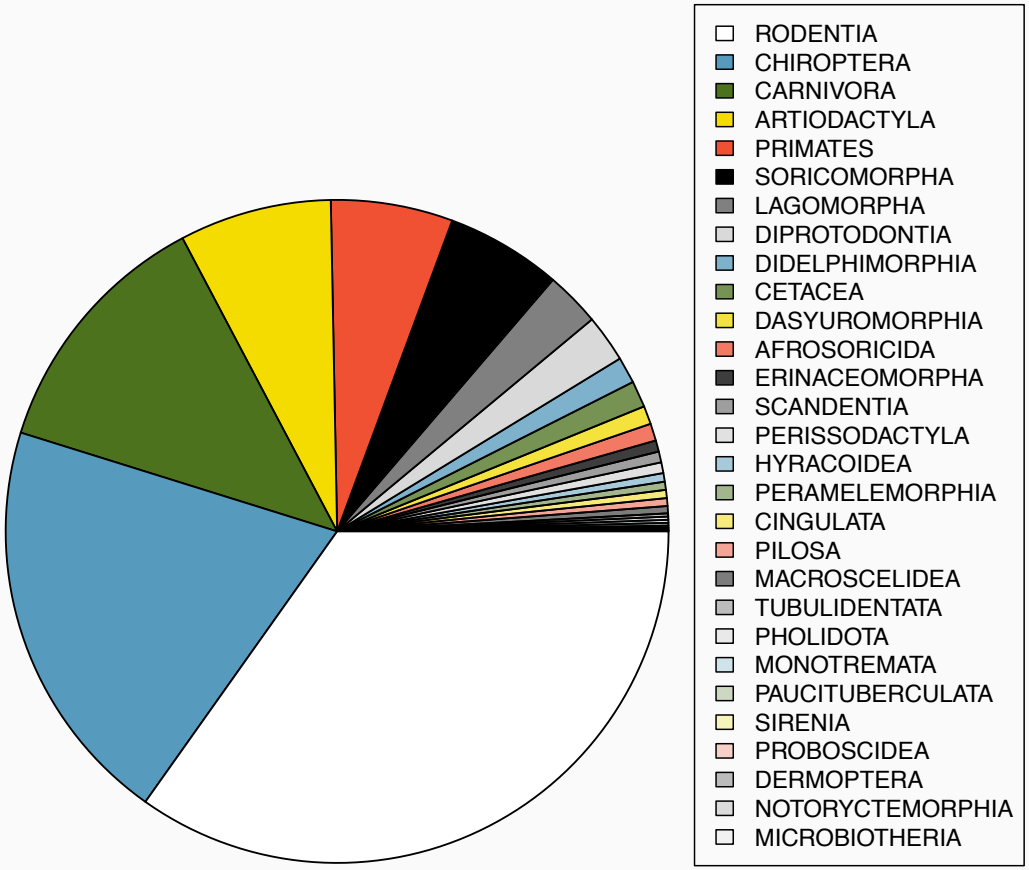
What are the differences between the three visualizations shown below?





Pie charts

Can you tell which order encompasses the lowest percentage of mammal species?



Data from *http://www.bucknell.edu/msw3*.

Credits

The material in How to describe visualizations was adapted from *Modern Data Science with R* by Benjamin Baumer, Daniel Kaplan, and Nicholas Horton, chapter 2.

Content in the sections **Examining Numerical Data** and **Histograms and shape**, as well as the slides with blue headers adapted from the chapter 1 OpenIntro Statistics slides developed by Mine Çetinkaya-Rundel and made available under the CC BY-SA 3.0 license.