Unit-2

Visualizing Distributions

Visualizing Amounts:-

1. Bar plots
2. Grouped and stacked Bars
3. Dot plots and Heat Maps:-

Bar plots:-

A Bar plot is a plot that presents categorical data with rectangular bars with length proportional to the values that they represent.

A Bar graph is a tool used in visualization to compare data among categories using bars.

A Bargraph can be presented horizontally or vertically. A Bar plot relation is direct that is the longer the bar, the greater its value.

A Bargraphs consists of two axes.

→ the horizontal axis (x axis) shows the data categories (Vertical graph)

→ the vertical axis (y axis) is the scale (Vertical graph).

Features of bar plot

1. A bar chart helps to compare a group of data between different groups.
2. A graph represents the relationship between the two axes (x & y axis)
3. It represent categories on one axis and a discrete value on the other.
4. Bar charts shows big changes in data over time.

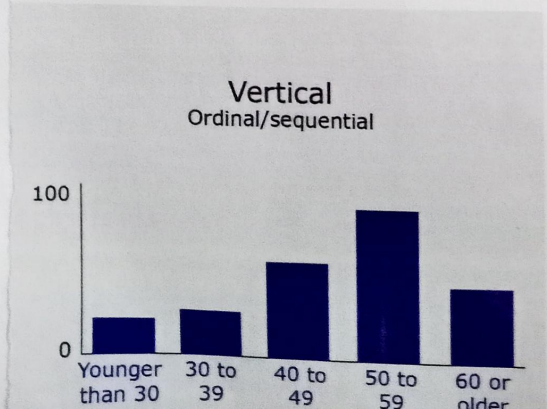
Types of Bar Graphs :-

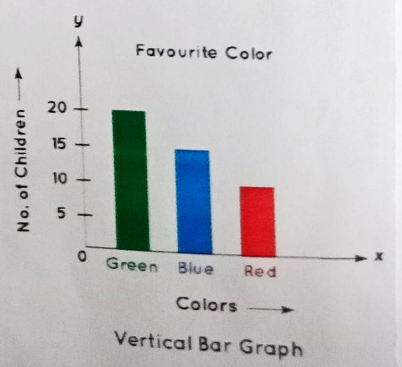
Bar Graphs are mainly classified into two types.

1. Vertical Bar Graph.
2. Horizontal Bar Graph.

1. Vertical Bar Graph :-

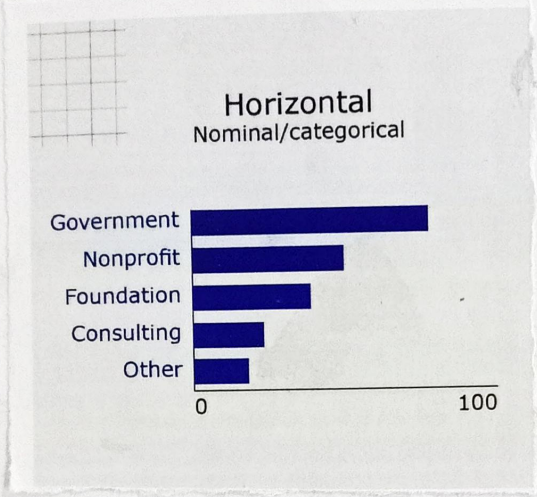
The rectangular bars are vertically drawn on the x-axis and the y-axis shows the value of the height of the rectangular bars which represents the quantity of the variables written on the x-axis.

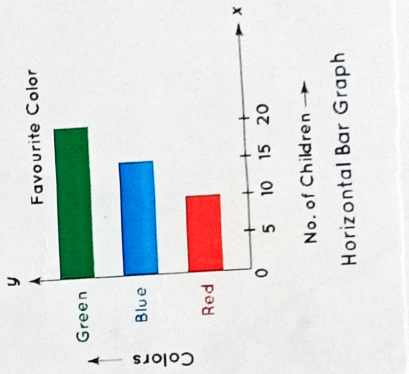




2. Horizontal Bar Graphs :-

The rectangular bars are horizontally drawn on the y-axis and the x-axis shows the length of the bars equal to the values of different variables present in the data.



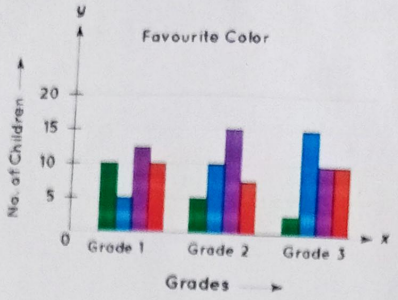


these are two more types of bar graphs. They are.

1. Grouped Bar Graph
2. Stacked Bar Graph.

Grouped Bar Graph :-

the grouped bar graph is also referred the clustered bar graph. It is used to show the discrete value for two or more categorical data. In this, rectangular bars are grouped by position for levels of one categorical variable, with the same colors showing the secondary category level within each group. It can be shown both vertically and horizontally.

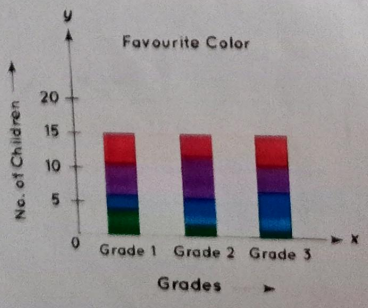


Grouped Bar Graph

Stacked Bar Graph

The stacked bar graph is also ~~refered~~ to as the composite bar graph. It divides the whole bar into different ~~categories~~ parts. In this each part of a bar is represented using different colors to easily identify the different categories. It requires specific labeling to indicate the different parts of the bar.

In the stacked bar graph every rectangular bar represents the whole, and each segment in the ~~legut~~ rectangular bar shows the different parts of the whole. It can be shown vertically or horizontally



Stacked Bar Graph

Dot plot :-

A Dot plot, also known as a strip plot or dot chart, it is a simple form of data visualization that consists of data points plotted as dots on a graph with an x and y axis. These types of charts are used to graphically depict certain data trends or groupings.

Dot plots typically contain the following elements.

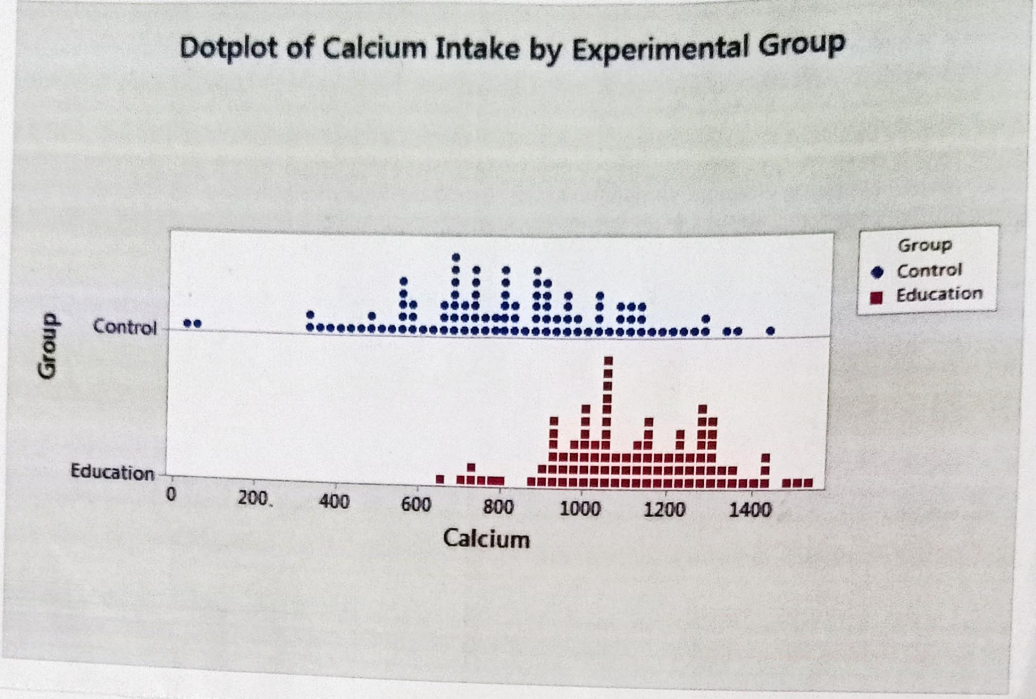
1. X-axis divided into ranges of values (bins) for the variable.
2. Dot height representing the frequency of observed values falling within each bin
3. Dots representing observations
4. optionally, Dot plots can display multiple distributions, allowing you to compare them.

The Dot plots has are two types

1. cleveland Dot plot
2. Wilkinson Dot plot.

→ the cleveland Dot plot is useful when using multiple variables as it does not require the axis to start at zero, allowing for the use of a log axis.

→ the wilkinson dot plot lays out data much like a histogram. It has individual data points, whereas a histogram lays out the data in bins.

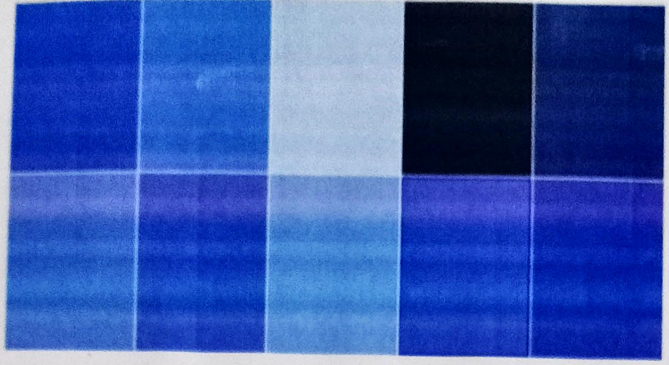


Heat Map:-

Heat map visualisation is to visualize the relationship between columns, represented in a matrix type view. The heat map uses color and intensity of the color to show the relationship between two columns.

→ Heatmaps are also alot more visual than standard analytics reports, which can make them easier to analyze at a glance.

→ This make them more accessible, particularly to people who are not accustomed to analyzing large amounts of data.



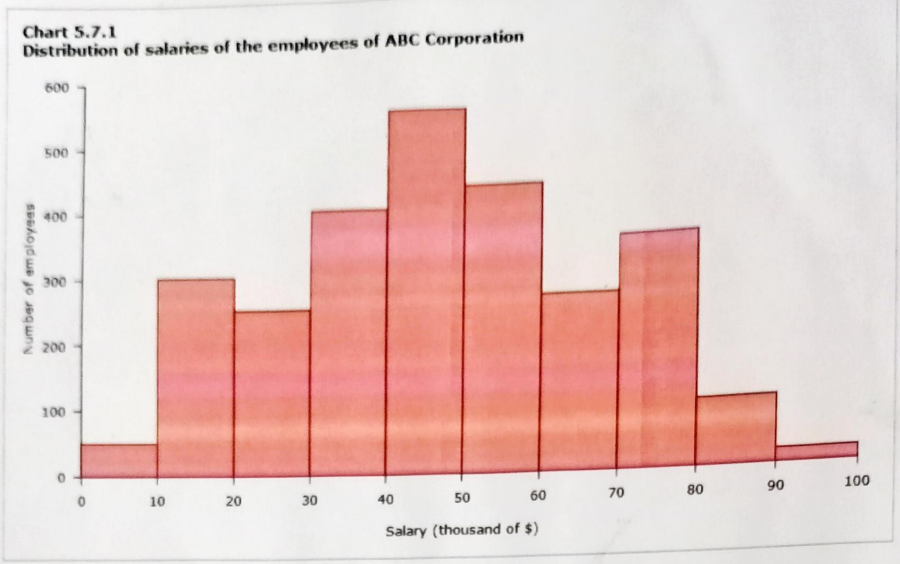
Visualizing Distributions: Histograms & Density plots :-

Histogram :-

It is a popular graphing tool. It is used to summarize discrete or continuous data that are measured on an interval scale.

A Histogram divides up the range of possible values in a dataset into classes or groups. For each group, a rectangle is constructed with a base length equal to the range of values in that specific group and a length equal to the number of observations falling into that group. A histogram has an appearance similar to a vertical barchart, but there are no gaps between them

EX1-

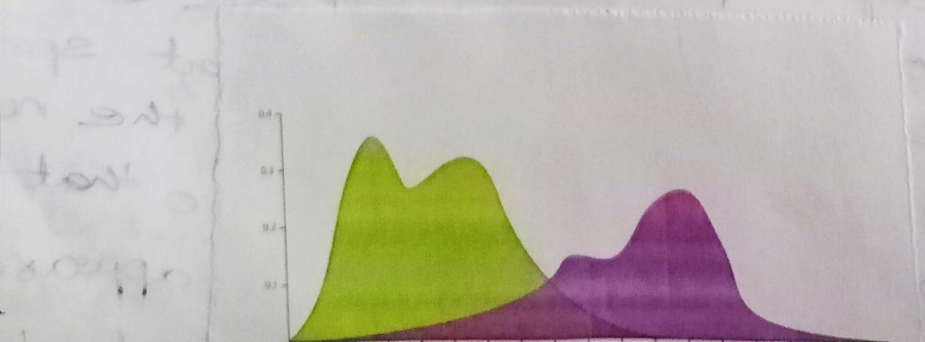


Data table for Chart 5.7.1

The following table presents the differences between a histogram and vertical bar graph.

Density plot

A density plot can be seen as an extension of the histogram. The density plot can smooth out the distribution of values and reduce the noise. It visualizes the distribution of data over a given period and the peaks show where values are concentrated.



Visualizing a single distribution:

We can obtain a sense of the age distribution among the passengers by grouping all passengers into bins with comparable ages and then counting the number of passengers in each bin.

The number of passenger with known age on the titanic is.

Agerange	count
0-5	36
6-10	19
11-15	18
16-20	99
21-25	139
26-30	121
31-35	76
36-40	74
41-45	54
46-50	50
51-55	26
56-60	22
61-65	16
66-70	3
71-75	3

We can visualize this table by drawing filled rectangles whose heights correspond to the counts and whose widths correspond to the width of the age bins such a visualization

is called a histogram. Histograms are generated by binning the data, their exact visual appearance depends on the choice of the bin width.

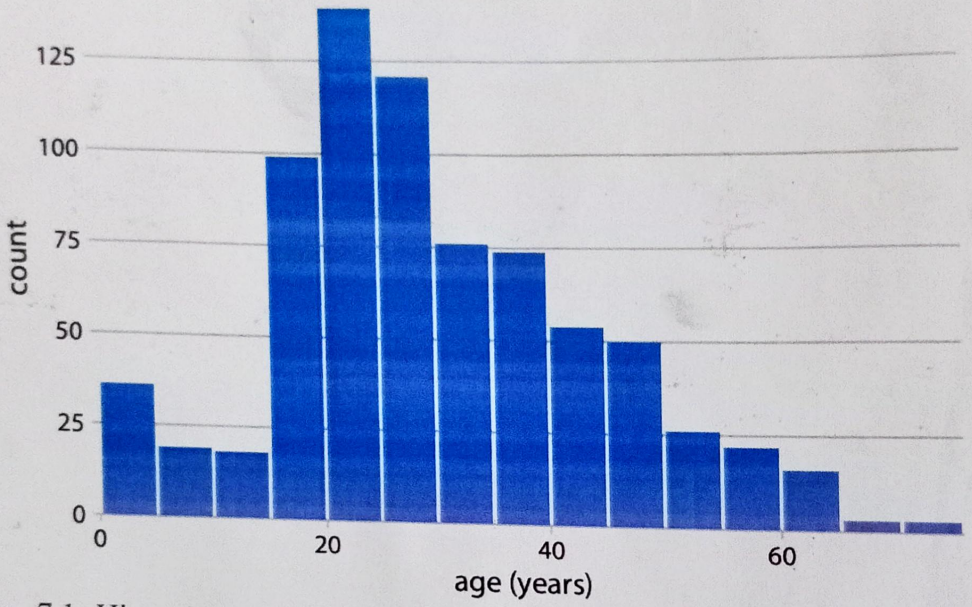


Figure 7.1: Histogram of the ages of Titanic passengers.

For the age distribution of Titanic passengers, we can see that a bin width of one year is too small and a bin width of fifteen years is too large, whereas bin width between three to five years work fine.

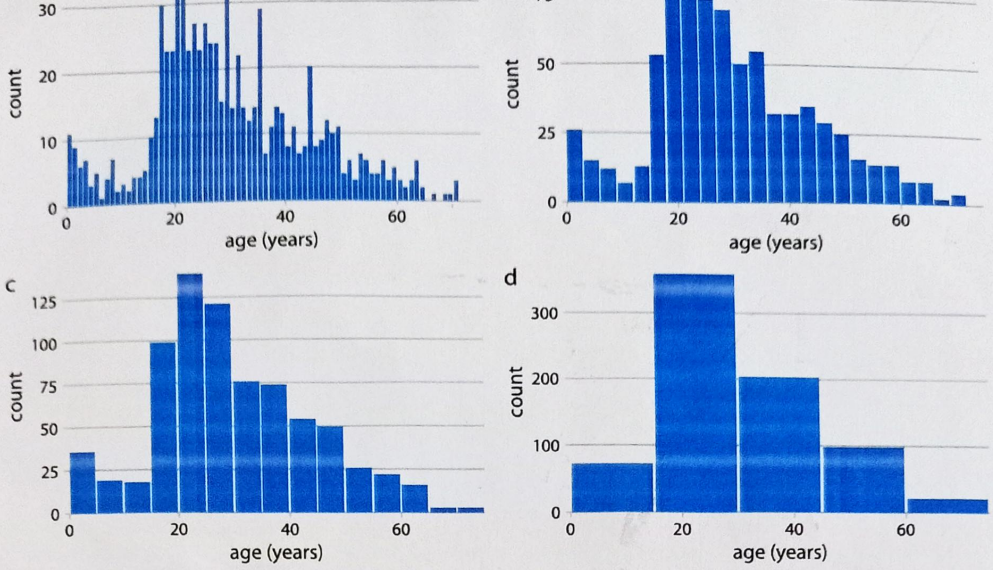


Figure 7.2: Histograms depend on the chosen bin width. Here, the same age distribution of Titanic passengers is shown with four different bin widths: (a) one year; (b) three years; (c) five years; (d) fifteen years.

This curve needs to be estimated from the data and the most commonly used method for this estimation. procedure is called kernal density estimation.

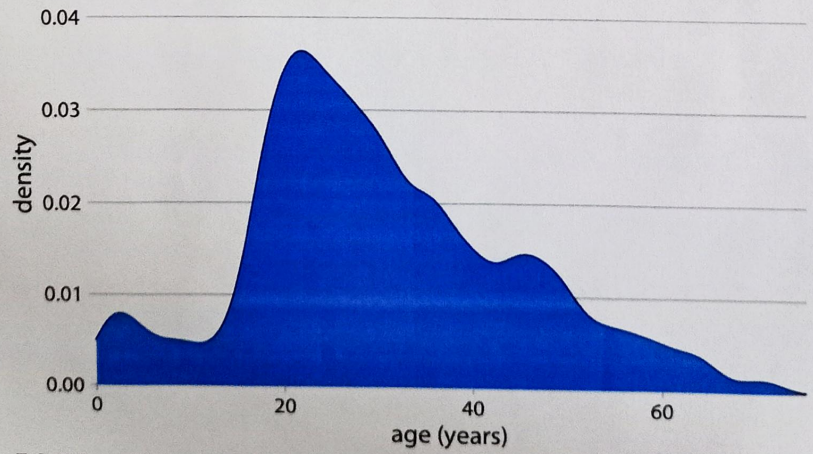


Figure 7.3: Kernel density estimate of the age distribution of passengers on the Titanic. The height of the curve is scaled such that the area under the curve equals one. The density estimate was performed with a Gaussian kernel and a bandwidth of 2

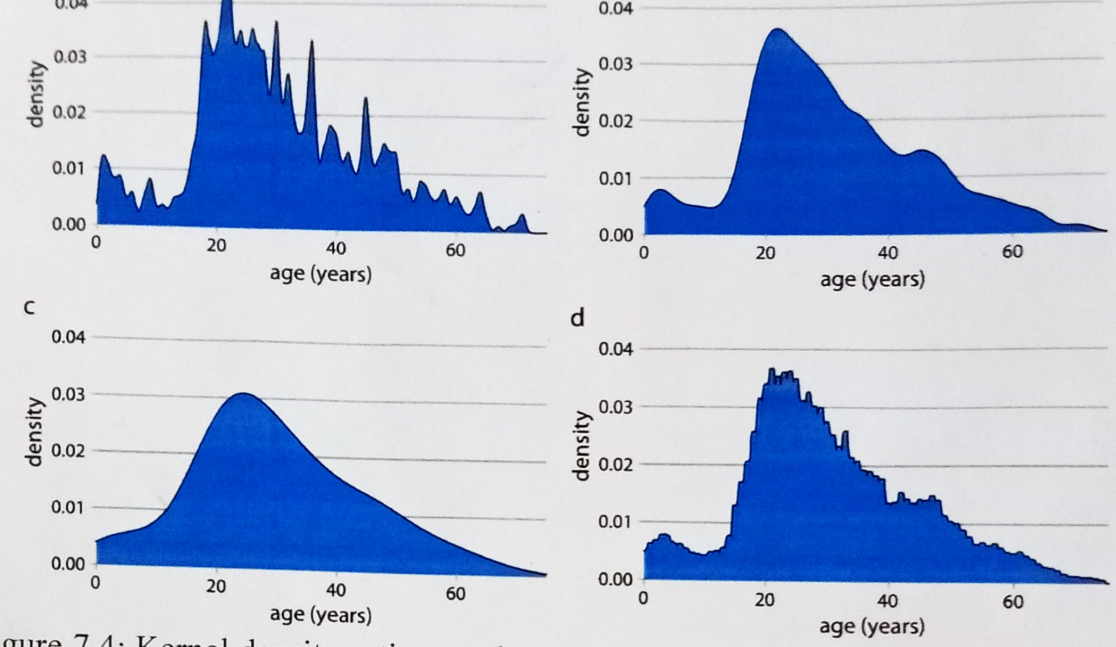


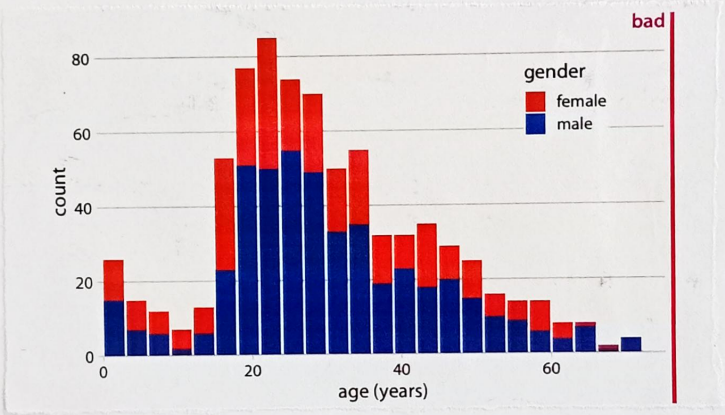
Figure 7.4: Kernel density estimates depend on the chosen kernel and bandwidth. Here, the same age distribution of Titanic passengers is shown for four different combinations of these parameters: (a) Gaussian kernel, bandwidth = 0.5; (b) Gaussian kernel, bandwidth = 2; (c) Gaussian kernel, bandwidth = 5; (d) Rectangular kernel, bandwidth = 2.

For example, in case of the age distribution, the data range on the x-axis goes from 0 to approximately 75. We expect the mean height of the density curve to be $1/75 = 0.013$.

We see that the y values range from 0 to approximately 0.04, with an average of somewhere close to 0.01

Visualizing multiple distribution at the same time:

In many scenarios, we have multiple distributions we would like to visualize simultaneously. Here we like to see how the ages of Titanic passengers are distributed between men and women.



This type of visualization should be avoided. There are two key problems here,
First, from just looking at the figure,
→ it is never entirely clear where exactly the bars begin.
→ the bar heights for the female counts cannot be directly compared to each other, because the bars all start at a different height.

The Density estimates of the ages of male and female Titanic passengers. To highlight that there were more male than female passengers, the density curves were scaled such the area under each curve corresponds to the total no of male and female passengers with known age.

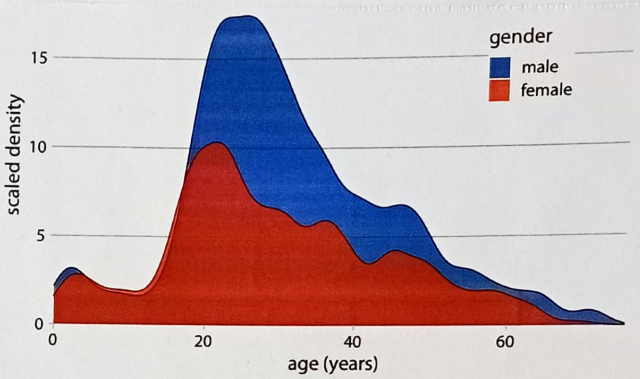


Figure 7.8: Density estimates of the ages of male and female Titanic passengers. To highlight that there were more male than female passengers, the density curves were scaled such that the area under each curve corresponds to the total number of male and female passengers with known age (468 and 288, respectively).

A solution that works well for this dataset is to show the age distributions of male and female

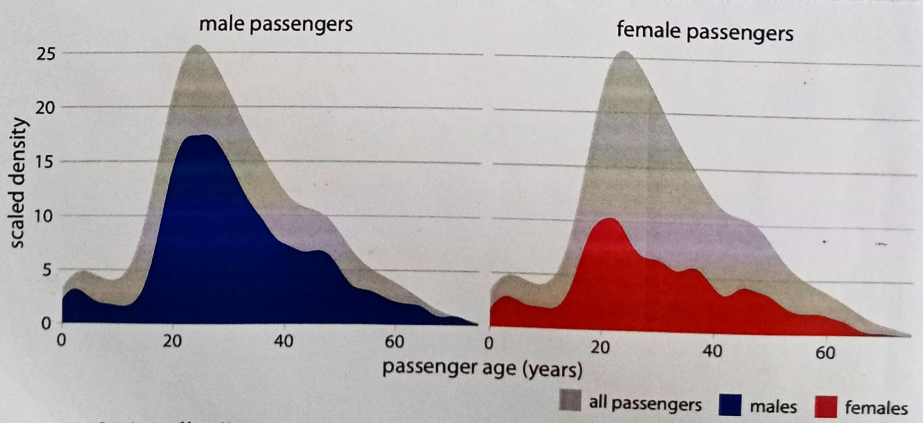


Figure 7.9: Age distributions of male and female Titanic passengers, shown as proportion of the passenger total. The colored areas show the density estimates of the ages of male and female passengers, respectively, and the gray areas show the overall passenger age distribution.

passengers separately.

→ Visualizing distributions: Empirical cumulative distribution functions and q-q plots:-

These types of visualizations require no arbitrary parameter choices and they show all the data at once.

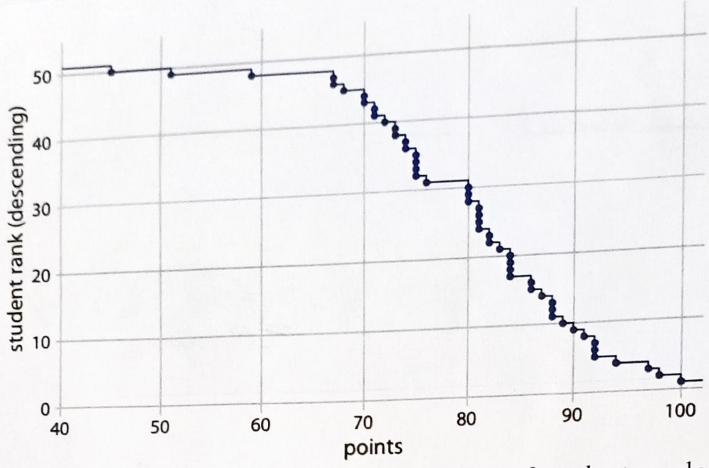
→ Empirical cumulative distribution functions

Assume our hypothetical class has 50 students and the students just completed an exam on which they could score between 0 and 100 points.

We can plot the total number of students that have received at most a certain number of points versus all possible point scores. This plot will be an ascending function starting at 0 for 0 points and ending at 50 for 100 points.

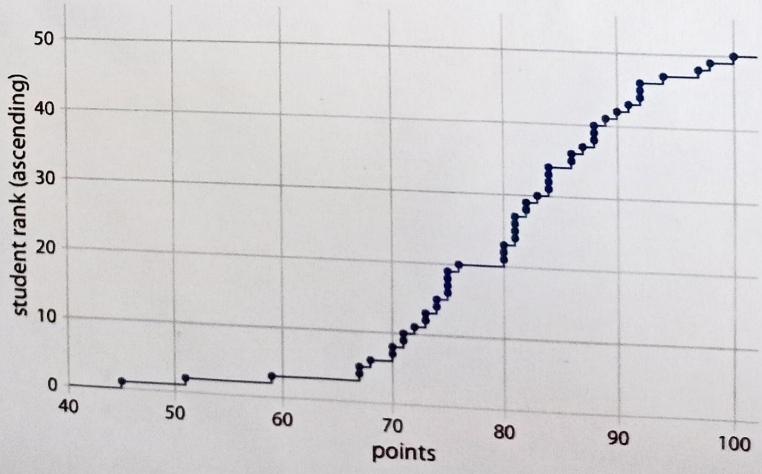
We can rank all students by the number of points they obtained in ascending order so that the student with few points receives the lowest rank and the student with the most points the highest rank then plot the rank versus the actual point obtained. The result is an empirical cummulative distribution function (ecdf) or simply cumulative distribution.

Each dot represent one student and the line visualize the highest student rank observed for any possible



Empirical cumulative distribution function of student grades for a hypothetical class of 50 students

If we rank the students the other way in descending order. This ranks simply flips the function on its head The result is still an ECDF.

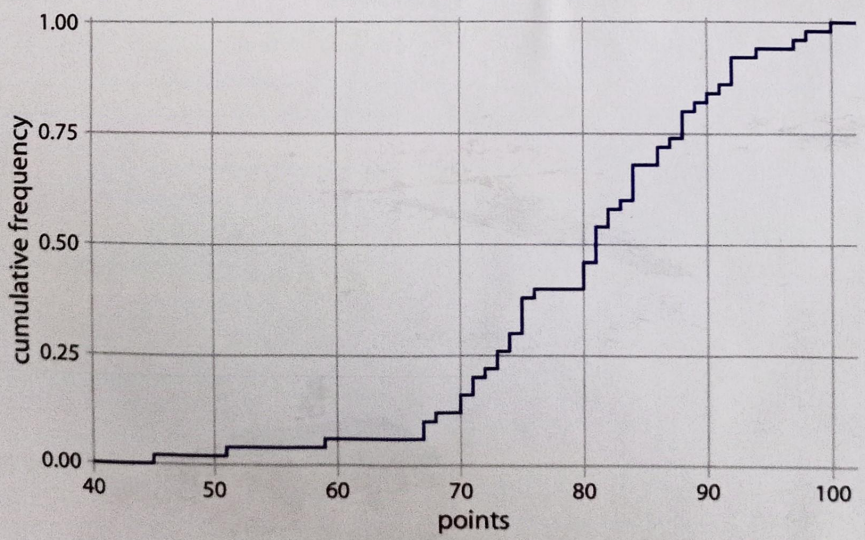


Distribution of student grades plotted as a descending

Ascending cumulative distribution functions are more widely known and commonly used that descending ECDF. But both have important applications.

In some applications, it is to draw the ecdf without highlighting the individual points and to normalize the rank by the maximum rank, so that the y axis represents the cumulative frequency.

For example 25% of the students received less than 75 points. The median point value corresponding to a cumulative frequency of 0.5 is 81.



Ecdf of student grades. The student ranks have been normalized to the total number of students, such that the **y** values plotted correspond to the fraction of students in the class with at most that many points

Highly skewness means a distribution curve has a shorter tail on one end of a distribution curve and a long tail on the other. If the skewness is between -1 and -0.5 or between

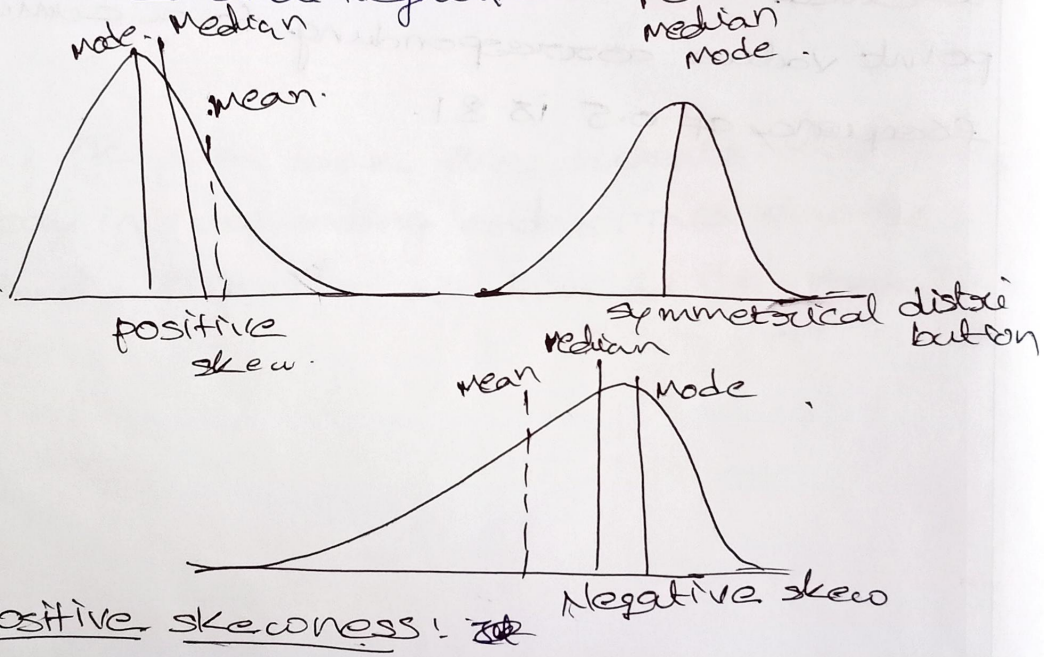
mode > median > mean

-1 and -0.5 → Negatively skewed.

0.5 and 1 → positively skewed

mean > median > mode.

These are two types of skewness positive and Negative.



Positive skewness:

It means when the tail on the right side of the distribution is longer. The mean and median will be greater than the mode.

Negative skewness

It means when the tail on the left side of the distribution is longer the mean and median will be less than the mode.

$$\text{mean} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

median = Middle element

mode = frequently occured elements.

Let us take an example dataset to calculate the mean, median mode of an grouped data.

Dataset:

Grade	f
40-49	3
50-59	5
60-69	6
70-79	9
80-89	8
90-100	7

Age	f
11-20	3
21-30	4
31-40	7
41-50	8
51-60	5
61-70	2
71-80	1

Mean :-

$$\text{Mean}(\bar{x}) = \frac{\Sigma fx}{\Sigma f} = \frac{1245}{30} = 41.5$$

Age	f	x	fx
11-20	3	$\frac{11+20}{2} = 15.5$	46.5
21-30	4	$\frac{21+30}{2} = 25.5$	102.0
31-40	7	$\frac{31+40}{2} = 35.5$	248.5
41-50	8	$\frac{41+50}{2} = 45.5$	364.0
51-60	5	$\frac{51+60}{2} = 55.5$	277.5
61-70	2	$\frac{61+70}{2} = 65.5$	131.0
71-80	1	$\frac{71+80}{2} = 75.5$	75.5

Median:

$$\text{Median} = lcb_m + \left(\frac{\frac{\Sigma f}{2} - cf_{bmc}}{f_m}\right) w$$

lcb_m → Lower class boundary.

$\frac{\Sigma f}{2}$ → $\frac{\text{sequency (Total)}}{2}$

cf_{bmc} → cumulative freq before Median class

f_m → freq of median class.

w → Width of the class

Age	f	lcb	cf
11-20	3	$\frac{10+11}{2} = \frac{21}{2} = 10.5$	3
21-30	4	$\frac{20+21}{2} = 20.5$	7
31-40	7	$\frac{30+31}{2} = 30.5$	14
41-50	8	$\frac{40+41}{2} = 40.5$	22
51-60	5	$\frac{50+51}{2} = 50.5$	27
61-70	2	$\frac{60+61}{2} = 60.5$	29
71-80	1	$\frac{70+71}{2} = 70.5$	30
	30		

Now find the median class.

$$\frac{\text{Total no of freq}}{2} = \frac{30}{2} = 15$$

Now check 15 will exist in the cf., No 15, so check the value near to 15 ie 22.

$$\therefore \text{median} = 40.5 + \left(\frac{15-14}{8}\right) 10$$

$$= 40.5 + \left(\frac{10}{8}\right)$$

$$= 40.5 + 1.25 = 41.75$$

Mode :- we calculate ~~Median~~ Mode based on the Median.

$$\text{Mode} = lcbm + \left(\frac{A_1 + A_2}{A_1 + A_2}\right) w$$

$A_1 \rightarrow$ Difference between freq of Median class and above class.

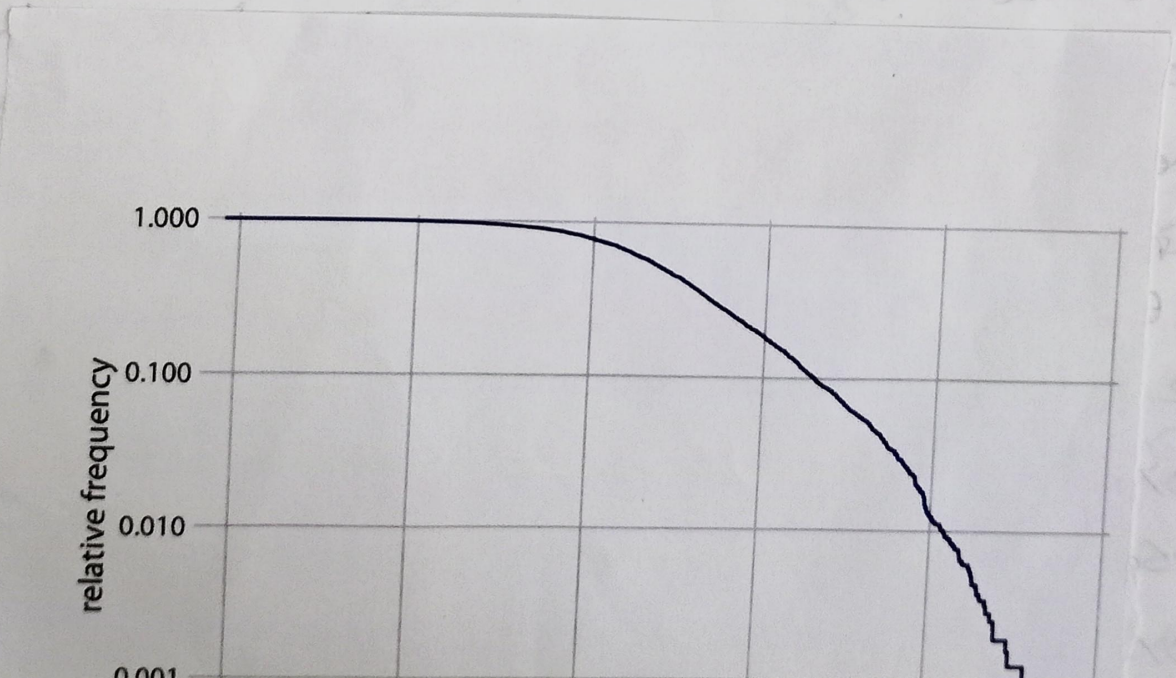
$A_1 = 8 - 7 = 1$ $\quad A_2 = 8 - 5 = 3$

$A_2 =$ Difference between freq of Median class and below class.

$$\text{Mode} = 40.5 + \left(\frac{1}{1+3}\right) \times 10$$

$$= 40.5 + \left(\frac{1}{4}\right) \times 10$$

$$= 40.5 + 2.5 = 42.5$$



q-q plot :- (quantile-quantile plot)

The q-q plot is a graphical technique for determining if two data sets come from populations with a common distribution.

The q-q plot is formed by.

1) Vertical axis : Estimate quantiles from data set 1

2) Horizontal axis : Estimate quantiles from data set 2.

3) The q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

4) A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately among the reference line.

5) By a quantile, we mean the fraction of points below the given value. That is, a quantile is the point at which x% percent of the data fall below and (100-x)% fall above that value.

6) Quantiles are values that split up a dataset into four equal parts.

Formula for calculate the quartiles for grouped data

$$Q_i = L + (i/4) * (iN/4 - M)$$

L → lower bound of the interval that contains the i^{th} quartile

c → class width.

F → the frequency of the interval that contains the i^{th} quartile

N → the total frequency

M → the cumulative frequency leading up to the interval that contains the i^{th} quartile.

suppose we have the following frequency distribution.

class	frequency	cumulative freq
1-5	6	6
6-10	19	25
11-15	13	38
16-20	20	58
21-25	12	70
26-30	11	81
31-35	6	87
36-40	5	92

suppose we like to calculate the Q_3 of this distribution.

the value of the third quartile will be located at position $iN/4$ in the distribution $iN/4 = \frac{3 \times 92}{4} = 69$

The interval that contains the third quartile will be the 21-25 interval since 69 is between the cumulative freq of 58 and 70 ie the class 21-25 with cf of 70.

→ Visualizing many distributions at once.

1) Visualizing distributions along the vertical axis

$\therefore L = 21$

$c = 25 - 21 = 4$.

$f = 12$.

$N = 92$

$M = 58$

∴ the cf leading upto 58 for 21-25 days

$$Q_3 = L + (c/f) * \left(\frac{3N}{4} - M\right)$$

$$= 21 + \left(\frac{4}{12}\right) * \left(\frac{3 \times 92}{4} - 58\right)$$

$$= 24 - 67.$$

Similar approach to calculate the values for the first and second quartiles.

When we perform this procedure for the student grades distribution from the beginning of this chapter, we obtain Figure 8.8.

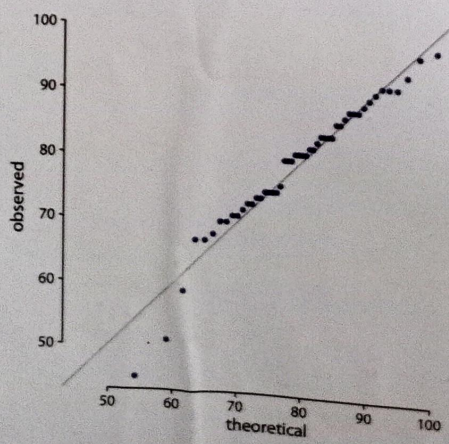


Figure 8.8: q-q plot of student grades.

Visualizing many distributions at once.

There are many scenarios in which we want to visualize multiple distributions at the same time.

there are two approaches

1. visualizing distribution along the vertical axis
2. visualizing distribution along the horizontal axis.

1) visualizing distribution along the vertical axis

The simplest approach to showing many distributions at once is to show their mean or median as points, with some indication of the variation around the mean or median shown by error bars.

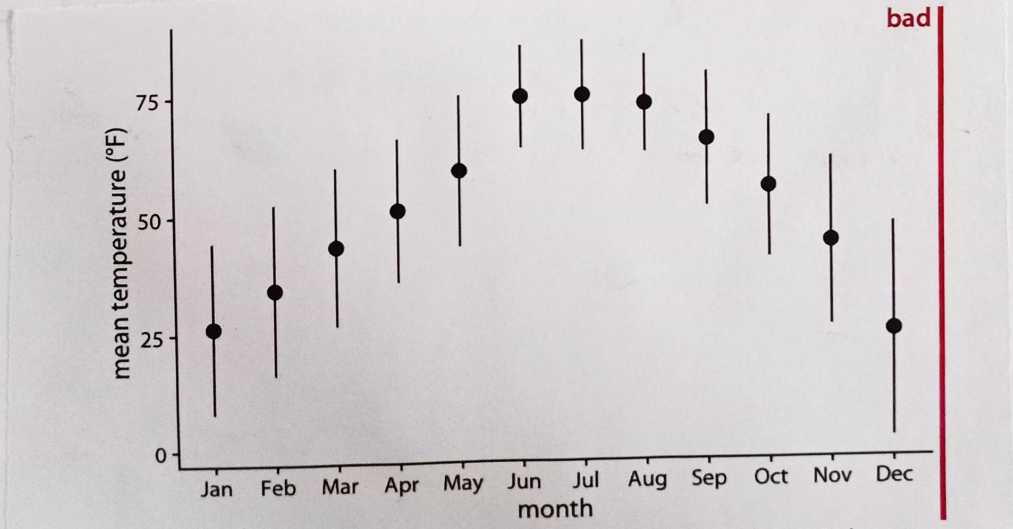
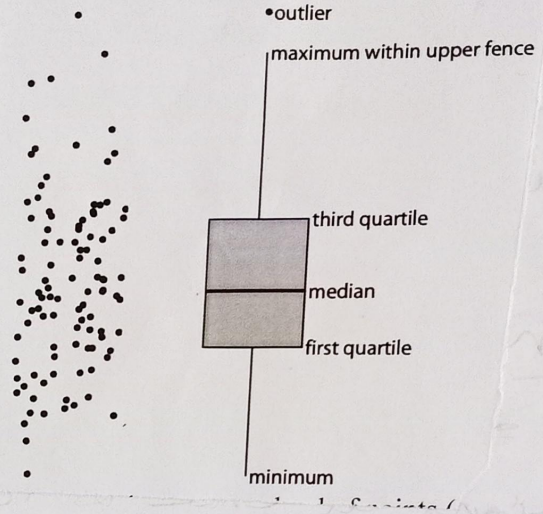
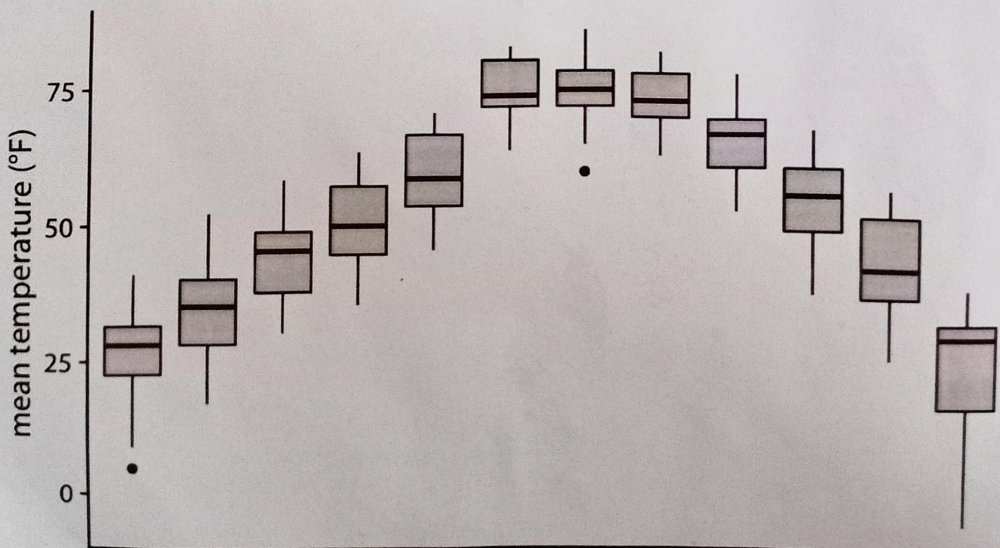


Figure 9.1: Mean daily temperatures in Lincoln, Nebraska in 2016. Points represent the average daily mean temperatures for each month, averaged over all days of the month, and error bars represent twice the standard deviation of the daily mean temperatures within each month. This figure has been labeled as "bad" because because error bars are conventionally used to visualize the uncertainty of an estimate, not the variability in a population. Data source: Weather Underground

A boxplot divides the data into quartiles and visualizes them in a standardized manner.



Boxplots are simple yet informative and they work well when plotted next to each other to visualize many distributions at once. In the following figure we can see that temperature is highly skewed in december and not very skewed at all in some other months.



violins can be used whenever one would otherwise use a boxplot and they provide a much more nuanced picture of the data. violin plots will accurately represent bimodal data whereas a boxplot will not

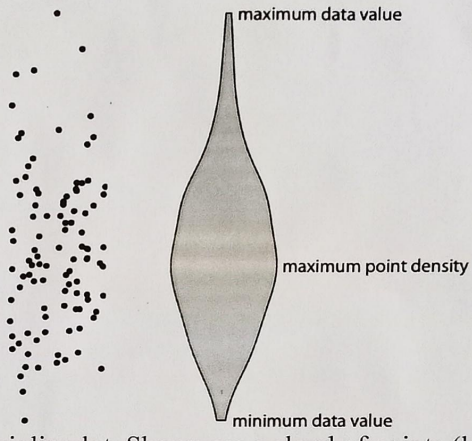
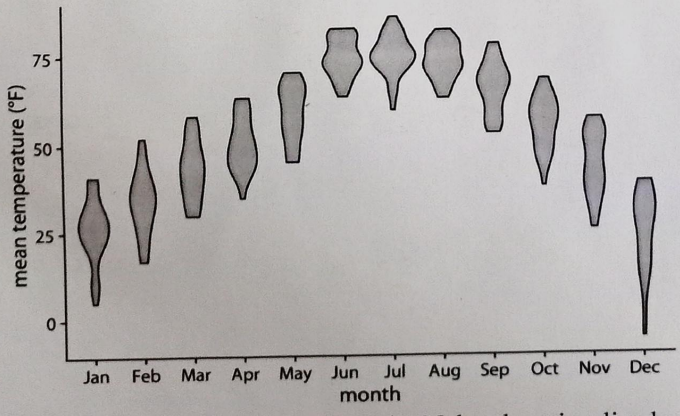


Figure 9.4: Anatomy of a violin plot. Shown are a cloud of points (left) and the corresponding violin plot (right). Only the y values of the points are visualized in the violin plot. The width of the violin at a given y value represents the point density at that y value. Technically, a violin plot is a density estimate rotated by 90 degrees and then mirrored. Violins are therefore symmetric. Violins begin and end at the minimum and maximum data values, respectively. The thickest part of the violin corresponds to the highest point density in the dataset.



: Mean daily temperatures in Lincoln, Nebraska, visualized as violin

strip charts are fine in principle as long as we make sure that we don't plot too many points on top to each other. A simple solution to overplotting is to spread out the points somewhat along the x-axis by adding some random noise in x dimension. this technique is called jittering

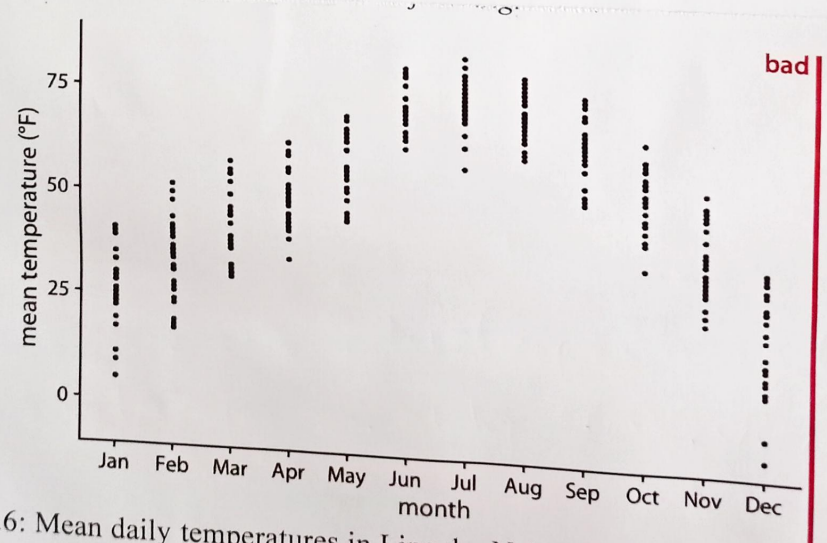


Figure 9.6: Mean daily temperatures in Lincoln, Nebraska, visualized as strip chart. Each point represents the mean temperature for one day. This figure is labeled as "bad" because so many points are plotted on top of each other that it is not possible to ascertain which temperatures were the most common in each month.

Another metho is sina plot, it is a hybrid between a violin plot and jittered points, and it shows each individual point while also visualizing the distributions

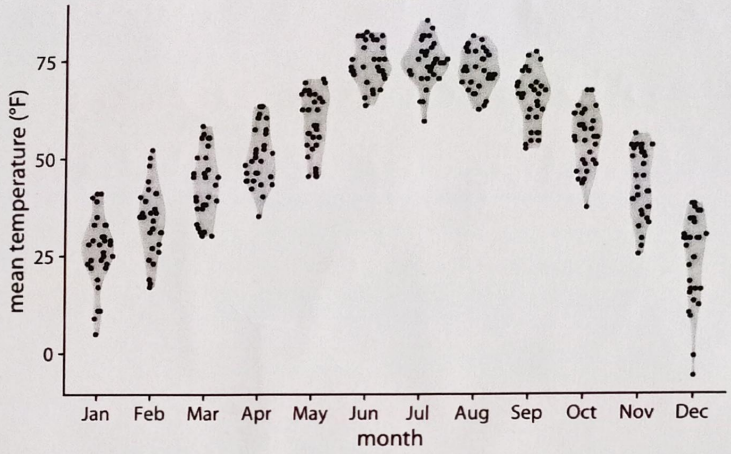
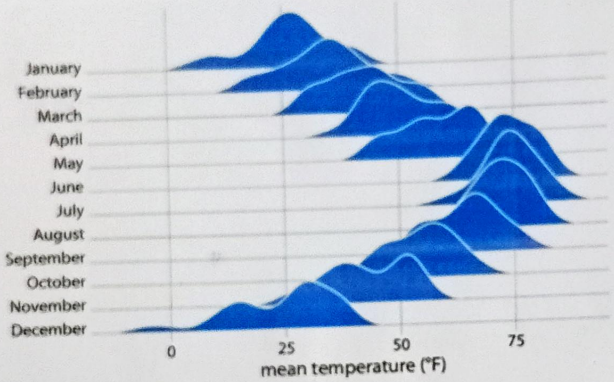


Figure 9.8: Mean daily temperatures in Lincoln, Nebraska, visualized as a sina plot (combination of individual points and violins). The points have been jittered along the *x* axis in proportion to the point density at the respective temperature. The name *sina plot* is meant to honor Sina Hadi Sohi, a student at the University of Copenhagen, Denmark, who wrote the first version of the code that researchers at the university used to make such plots (Frederik O. Bagger, personal communication).

Visualizing distribution along the horizontal axis

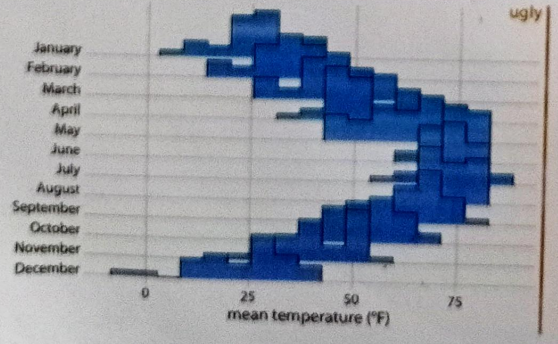
We visualized distributions along the horizontal axis using histograms and density plots. We have another visualization called ridgeline plot, because these plots look like mountain ridgelines. Ridgeline plots tends to work particularly well if we want to show trends in distributions over time.

Ridgeline plots tend to work particularly well if want to show trends in distribution over time. The standard ridgeline plot uses density estimates. It is quite closely related to the violin plot, but frequently evokes a more intuitive understanding of the data



Temperatures in Lincoln, Nebraska, in 2016, visualized as a ridgeline plot. For each month, we show the distribution of daily mean temperatures measured in Fahrenheit. Original figure concept: Wehrwein (2017).

We can use histogram instead of density plots in a ridgeline visualization.



Temperatures in Lincoln, Nebraska, in 2016, visualized as a ridgeline plot of histograms. The individual histograms don't separate well visually, and the overall

Ridgeline plots work well if we want to compare two trends over time. This is a scenario that arises commonly if we want to analyze the voting patterns of the members of two different parties.

We can make this comparision by staggering the distribution vertically by time and drawing two differently colored distribution at each time point representing the two parties.

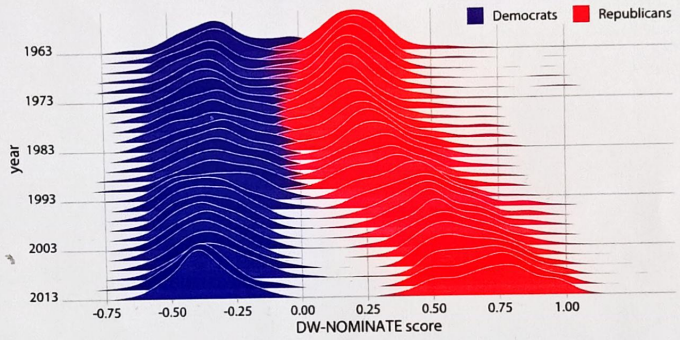


Figure 9.12: Voting patterns in the U.S. House of Representatives have become increasingly polarized. DW-NOMINATE scores are frequently used to compare voting patterns of representatives between parties and over time. Here, score distributions are shown for each Congress from 1963 to 2013 separately for Democrats and Republicans. Each Congress is represented by its first year. Original figure concept: McDonald (2017).

Assignment

1) Explain highly skewed distributions and ECDF with an example.
2) How to visualize multiple distribution at the same time.
3) Explain visualizing distributions in vertical and horizontal axes
4) Explain Briefly about visualizing