Oct 14, 2024

Data Visualization

Week 4. Visualizing distributions

Reminder

- visualizing amounts
  - barplot and its variants
  - dotplot
  - heatmap
- problems in barplots
  - coordination flips
  - ordering bars

Introduction

What does distribution mean?

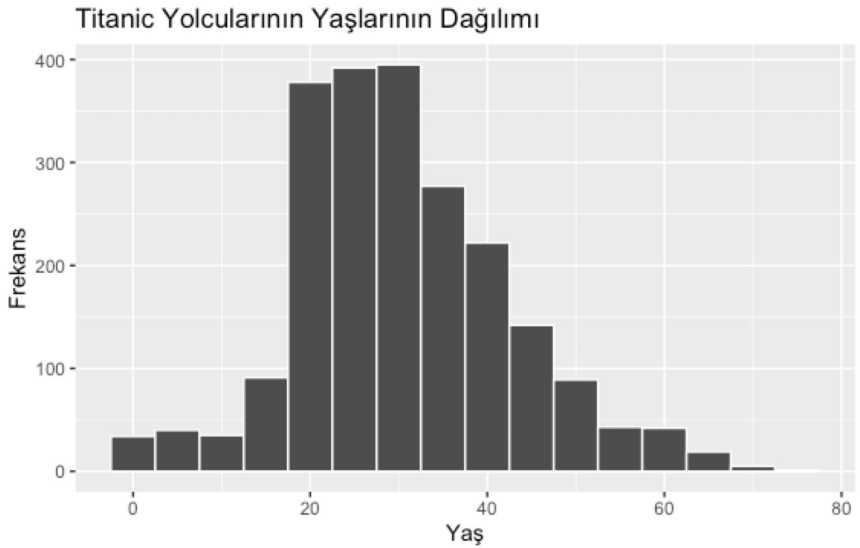
Visualizing Distributions

For visualizing the distribution of a variable:

- Histogram
- Kernel density estimation

1.Histogram

Histogram is created by visualizing the grouping of observed values based on fixed bin widths.



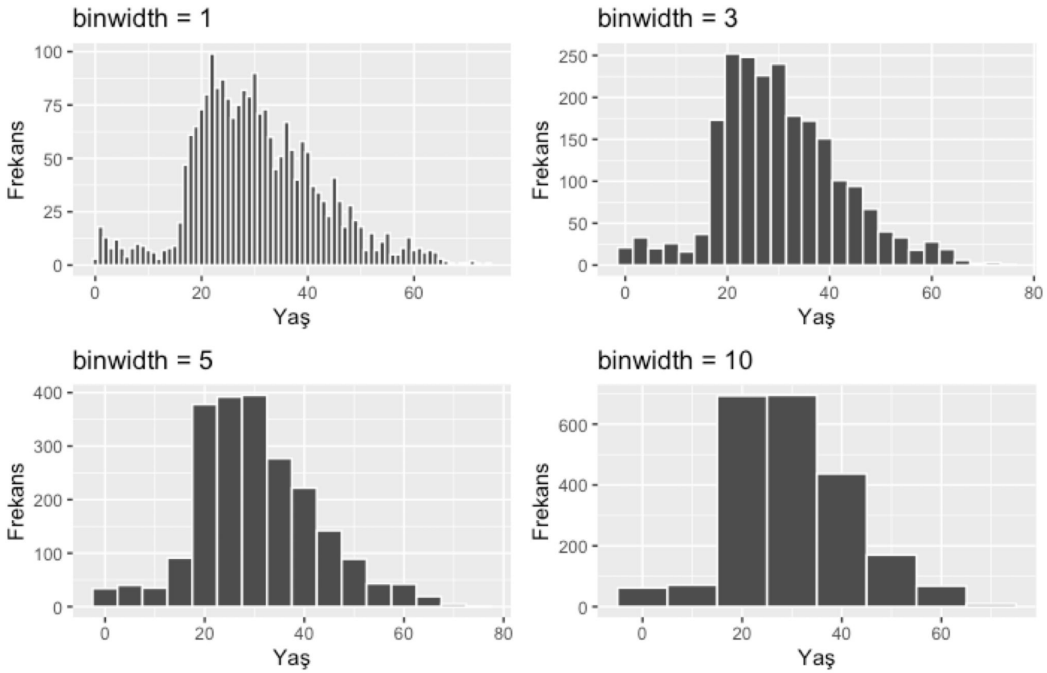
1.Histogram

The most significant issue in creating a histogram is that its appearance depends on the chosen bin width.

- If the bin width is set too small, excessive peaks may appear in the histogram, making interpretation difficult.
- If it is set too large, important variations within smaller intervals may be lost, making them undetectable.

Finding the appropriate bin width is achieved by experimenting with different bin widths and selecting the most suitable one.

1.Histogram


binwidth = 1
Frekans
Yaş
binwidth = 3
Frekans
Yaş
binwidth = 5
Frekans
Yaş
binwidth = 10
Frekans
Yaş

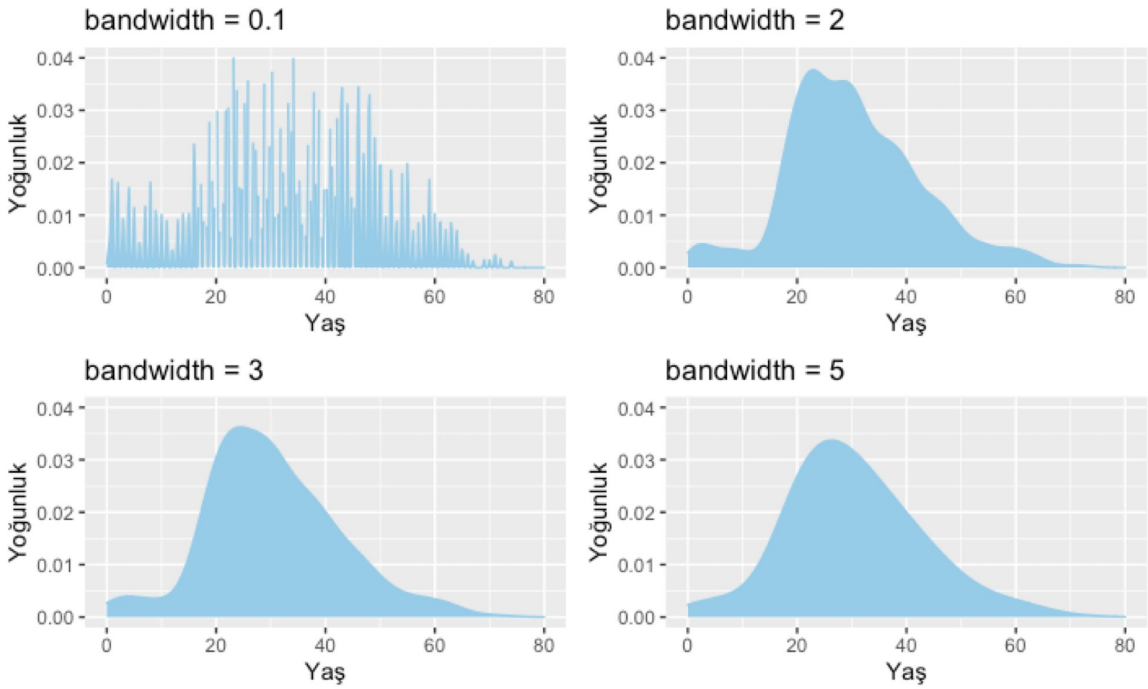
2.Kernel Density

Although histograms are more commonly used in practice, the use of kernel density estimation has increased in recent years.

2.Kernel Density

The most significant issue with kernel density estimations is that they can create the impression of observations in areas where no data points exist. For example, when visualizing a variable like age, which cannot take negative values, a negative age might appear in the plot. Careful attention is required to avoid such situations.

2.Kernel Density



Visualizing the distribution of two variables

We often encounter situations where it is necessary to visualize the distributions of two variables / levels.

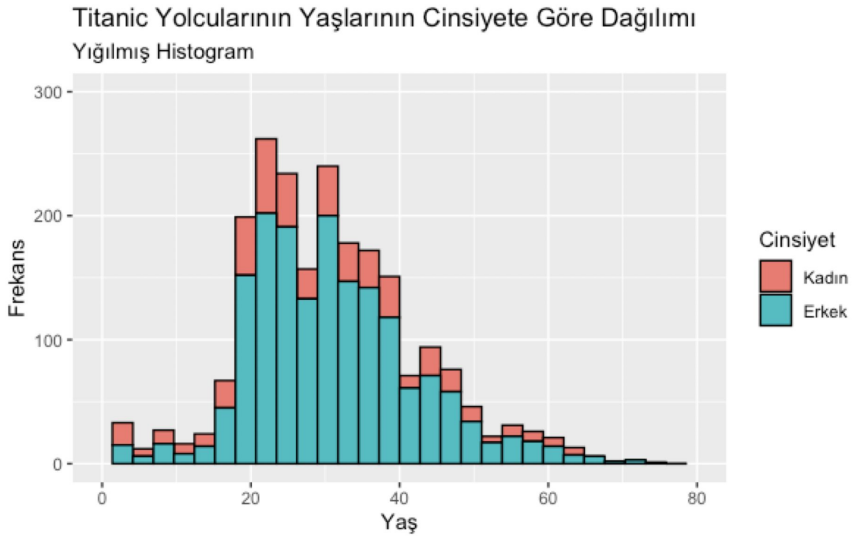
For example, when examining the age distributions of Titanic passengers by gender, we may need to answer the following questions:

- Are the average ages of male and female passengers similar?
- Is there a difference in passenger ages between genders?

In such cases, separate histograms can be created for each gender group, or a stacked histogram can be used.

1.Stacked Histogram

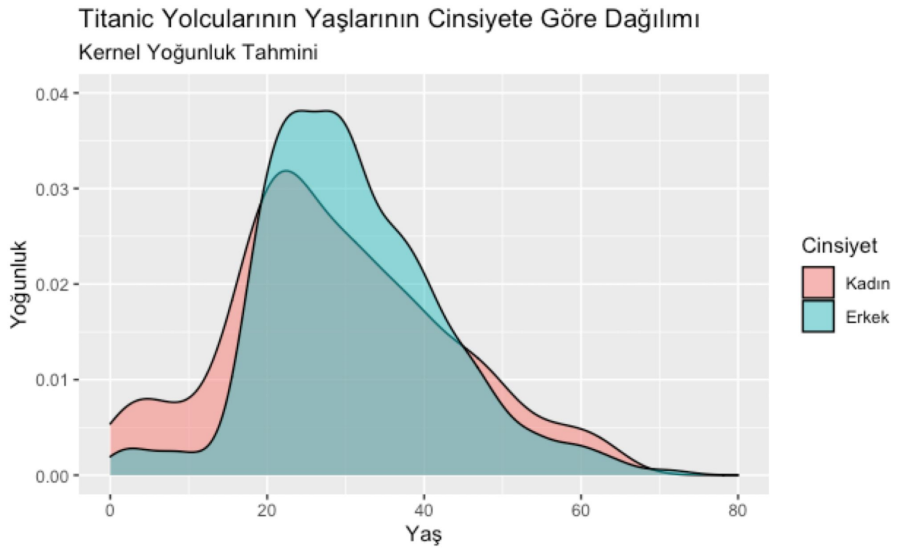
A stacked histogram involves drawing bars representing different groups on top of each other in different colors.



What are the two key issues in the plot?

2.Kernel Density

Due to the limitations of stacked histograms, using kernel density estimation for multiple groups is a better solution.



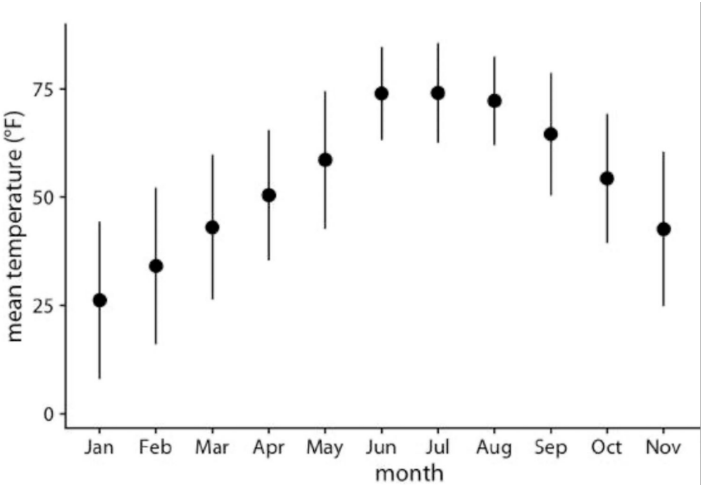
Visualizing the distribution of many variables

Situations may arise where the distributions of multiple variables need to be visualized simultaneously:

- Distribution of monthly temperatures
- Distribution of per capita income among countries
- ...

1.Error Bars

The simplest way to visualize multiple distributions simultaneously is by using error bars. Error bars can be created in various ways. One method is to represent the median as a point and show one standard deviation from the median with bars.



1.Error Bars

However, error bars have some limitations:

- They **contain limited information**, as they only show the median and standard deviation.
- **An explanation is needed**, as not everyone may know what the points and lines represent.
- They **do not indicate the symmetry or asymmetry** of the data distribution.

2.Boxplot

A boxplot visualizes the data by summarizing it with five points: minimum, first quartile, median, third quartile, and maximum.

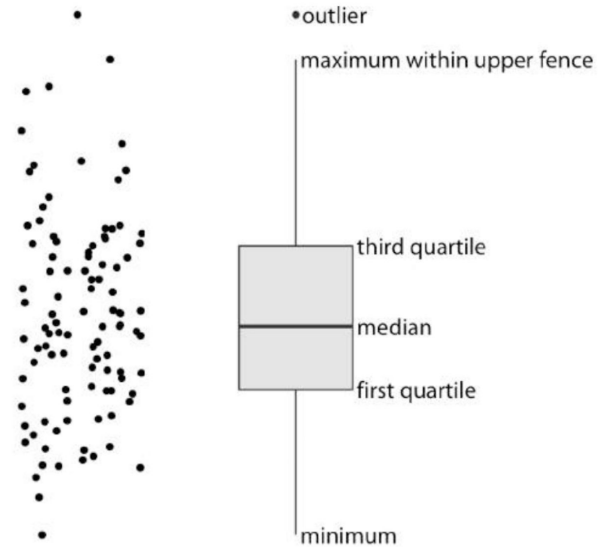


Figure 9-2. Anatomy of a boxplot. Shown are a cloud of points (left) and the corresponding boxplot (right).

2.Boxplot

Box plots can be used side by side to visualize multiple distributions.

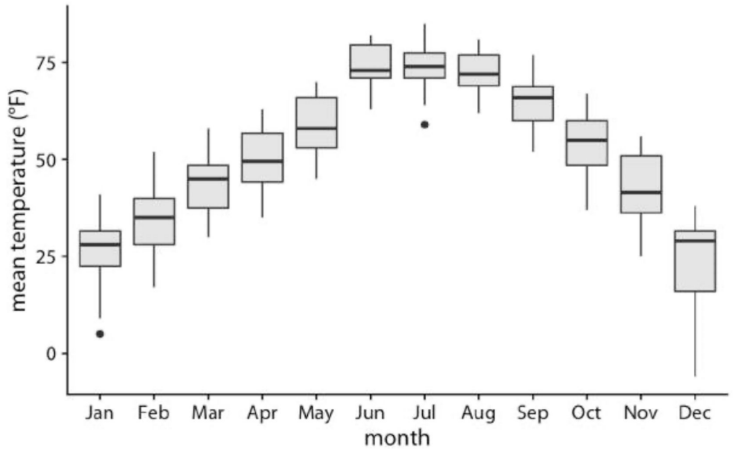


Figure 9-3. Mean daily temperatures in Lincoln, NE, visualized as boxplots. Data source: Weather Underground.

3.Violin plot

The most significant limitation of box plots is their inability to visualize bimodal distributions. In such cases, a violin plot is a good alternative.

- The width of the violin represents the density of observations at that point.
- Observations start at the minimum point and end at the maximum point.
- Before using a violin plot, it is essential to ensure that there are a sufficient number of observations.

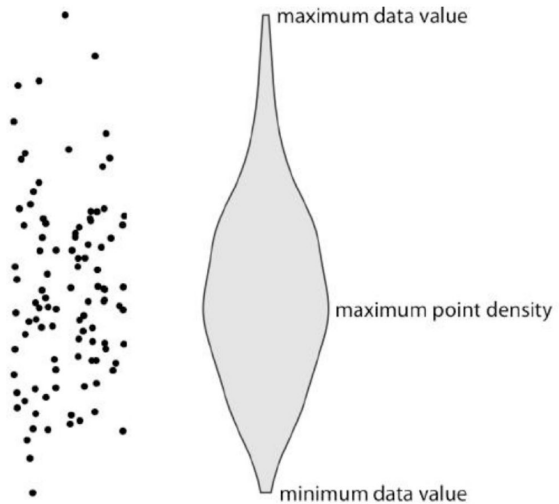


Figure 9-4. Anatomy of a violin plot. Shown are a cloud of points (left) and the corresponding violin plot (right).

3.Violin plot

Violin plots have some limitations because they are derived from density estimations:

- They can create the appearance of observations in areas where there are none.
- In areas with very few observations, they may suggest a misleadingly high density of observations.

To mitigate these limitations, they can be used in conjunction with strip plots.

3.Violin plot

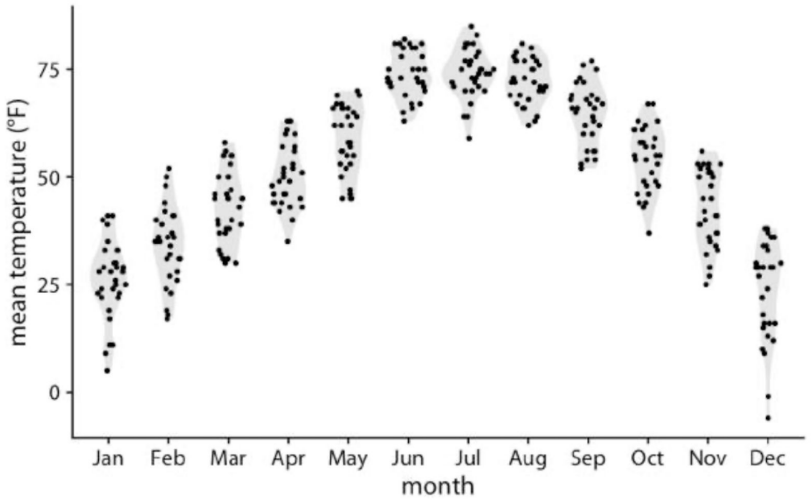


Figure 9-8. Mean daily temperatures in Lincoln, NE, visualized as sina plots (a combination of individual points and violins). The points have been jittered along the x *axis in proportion to the point density at the respective temperature. Here, the sina plots are shown superimposed on violin plots. Data source: Weather Underground.*

4.Ridgeline plot

Ridgeline plot is created by placing a continuous variable on the horizontal axis and a multilevel categorical variable (usually time) on the vertical axis. The general purpose of this type of plot is to observe changes in the distribution of the relevant variable over time.

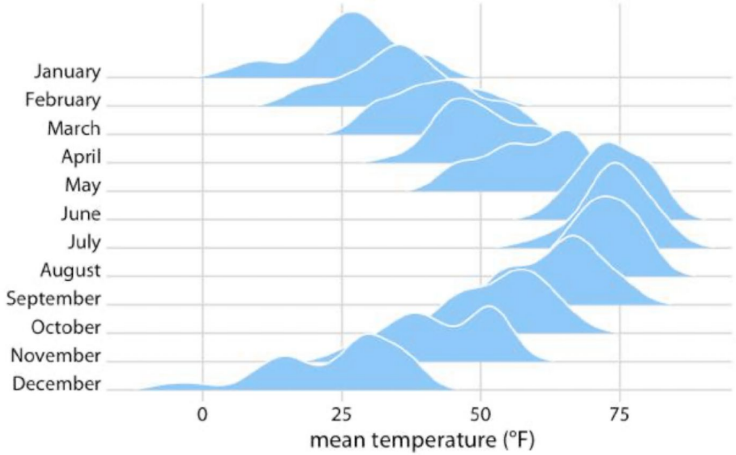


Figure 9-9. Temperatures in Lincoln, NE, in 2016, visualized as a ridgeline plot. For each month, we show the distribution of daily mean temperatures measured in Fahrenheit. Original figure concept: [Wehrwein 2017]. Data source: Weather Underground.

4.Ridgeline plot

It can be used to visualize the changes of a variable over many years.

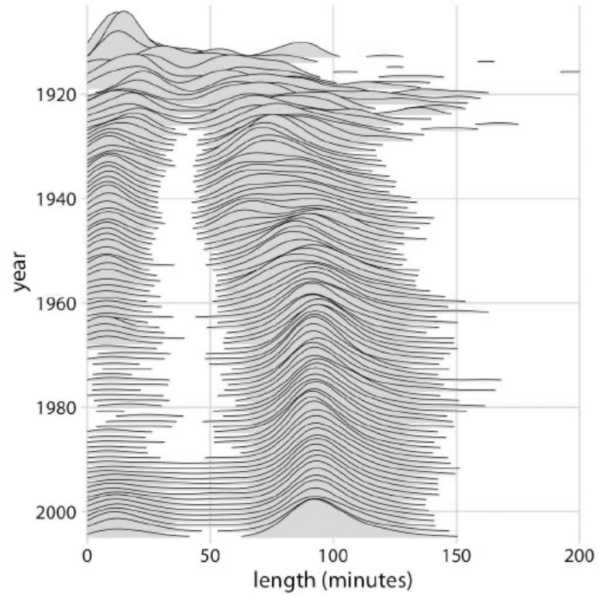
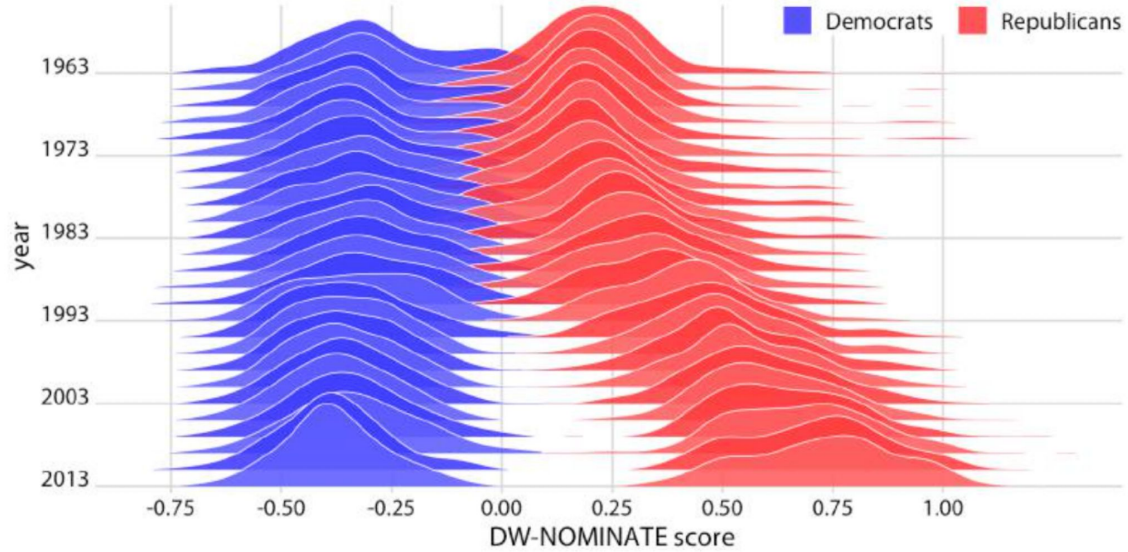


Figure 9-11. Evolution of movie lengths over time. Since the 1960s, the majority of all movies have been approximately 90 minutes long. Data source: Internet Movie Database (IMDB).

4.Ridgeline plot

It can be used to compare two trends over time.



Reference

The notes and plots in the presentation are compiled from Claus O. Wilke's book, *Fundamentals of Data Visualization*.