

Outline
Data types
Three most common data types
Proportions data
Comparing Proportions for many categories
Point Data
Stacking principle and bar plot
Point charts
Distributional data
Comparing distributions for many categories
Summary

Data Visualisation Using R

Lecture-1

Suman Rakshit

School of EECMS, Curtin University

February 8, 2024



Curtin University

Main objectives

- ▶ Identify the major issues and directions in contemporary data visualisation.
- ▶ Selecting appropriate visualisation techniques for the real world problem.
- ▶ Produce high quality figures for effective communication and scientific interpretation.

Outline
Data types
Three most common data types
Proportions data
Comparing Proportions for many categories
Point Data
Stacking principle and bar plot
Point charts
Distributional data
Comparing distributions for many categories
Summary

Main objectives

- ▶ Identify the major issues and directions in contemporary data visualisation.
- ▶ Selecting appropriate visualisation techniques for the real world problem.
- ▶ Produce high quality figures for effective communication and scientific interpretation.

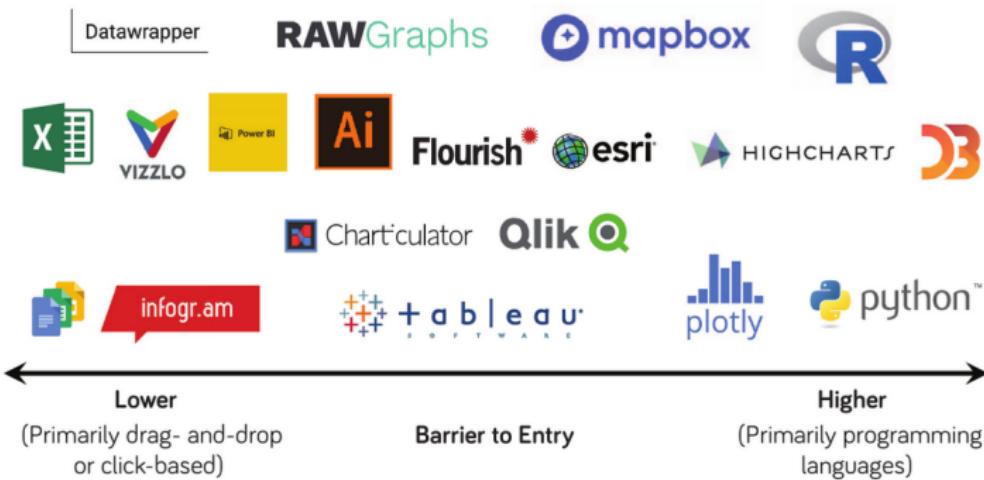
Outline
Data types
Three most common data types
Proportions data
Comparing Proportions for many categories
Point Data
Stacking principle and bar plot
Point charts
Distributional data
Comparing distributions for many categories
Summary

Main objectives

- ▶ Identify the major issues and directions in contemporary data visualisation.
- ▶ Selecting appropriate visualisation techniques for the real world problem.
- ▶ Produce high quality figures for effective communication and scientific interpretation.

Outline
Data types
Three most common data types
Proportions data
Comparing Proportions for many categories
Point Data
Stacking principle and bar plot
Point charts
Distributional data
Comparing distributions for many categories
Summary

Data Visualisation Platforms



Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

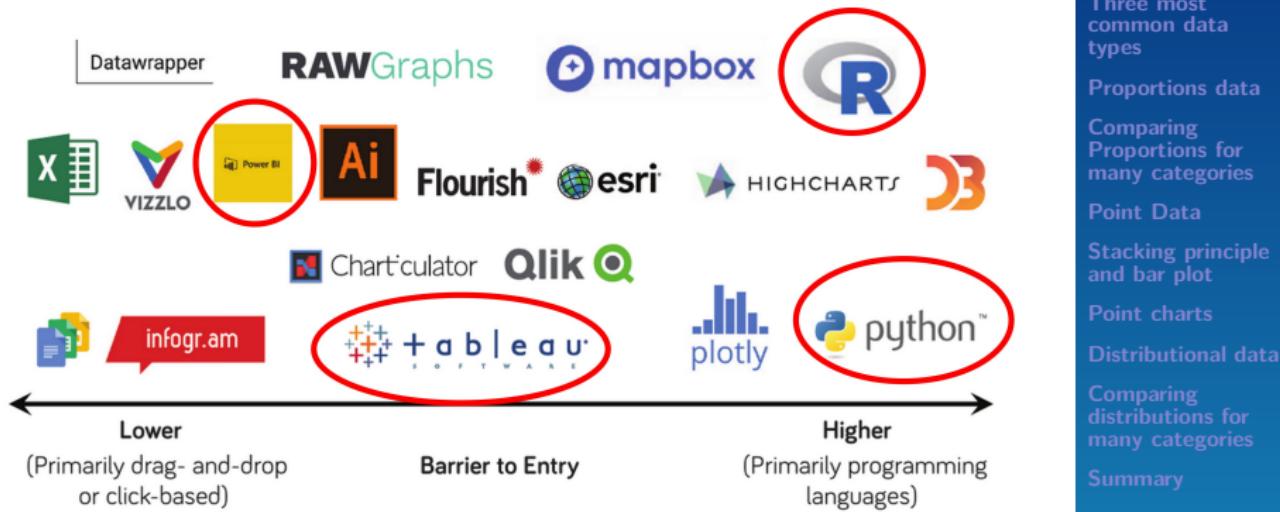
Point charts

Distributional data

Comparing distributions for many categories

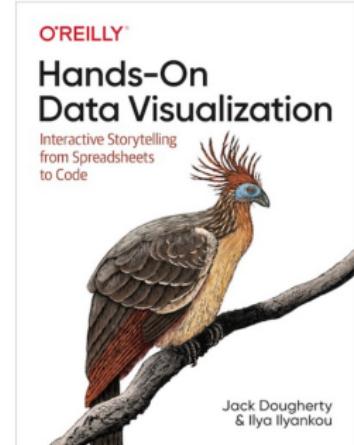
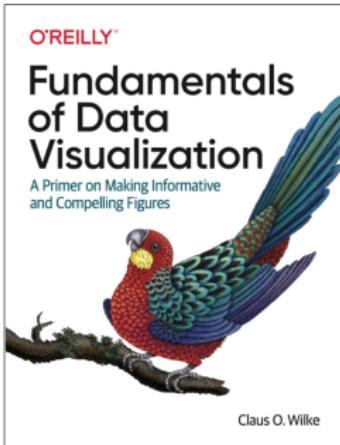
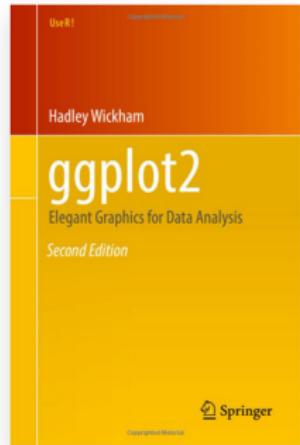
Summary

Visualisation Platforms taught in this Class



- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Books on Data Visualisation



- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Outline

1. Different Visualisations for different data types
2. Three most common data types: (i) proportions, (ii) point data, and (iii) distributions.
3. Proportions data examples and charts
4. Comparing Proportions for many categories
5. Point data examples and charts
6. Bar chart for point data (stacking principle)
7. Point chart for point data
8. Distributional data
9. Comparing distributional data of many categories
10. Summary

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary

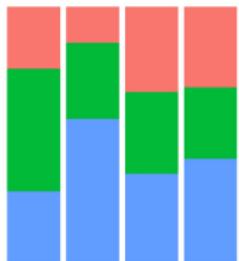


Different visualisations for different data

Proportion Data

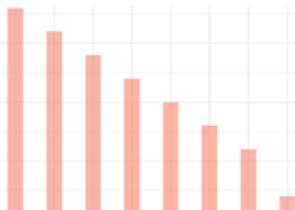


Pie chart



Stacked Bar chart

Point Data

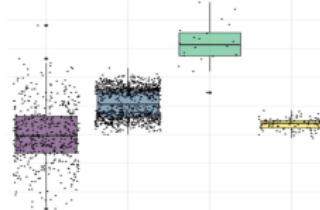


Bar chart



Point chart

Univariate Distribution



Boxplot



Kernel Density

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary

Figure 1: Standard charts for three common data types.

Examples of three data types

- ▶ **Proportions** (Percentage data): Data representing proportion (or percentage) of the whole population — all proportions (percentages) should add upto one (hundred). Examples: (i) *proportions of total revenue generated by different departments*, (ii) *percentages of total gun deaths for all states*, or (iii) *proportions of reported disease cases by the world health organisation (WHO) for different disease types*.
- ▶ **Point data:** Data representing a single summary. Examples: (i) *Total number of disease cases in a state*, (ii) *Risk of gun murder for each state in a year*, or (iii) *Median house prices in different suburbs*.
- ▶ **Distributional data:** Data representing many samples from a population. Examples: (i) *Blood sugar levels of patients after taking a new drug treatment*, (ii) *Samples from agricultural field trials to measure the risk of a fungal disease*, or (iii) *Exam marks of students*.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary

Examples of three data types

- ▶ **Proportions** (Percentage data): Data representing proportion (or percentage) of the whole population — all proportions (percentages) should add upto one (hundred). Examples: (i) *proportions of total revenue generated by different departments*, (ii) *percentages of total gun deaths for all states*, or (iii) *proportions of reported disease cases by the world health organisation (WHO) for different disease types*.
- ▶ **Point data:** Data representing a single summary. Examples: (i) *Total number of disease cases in a state*, (ii) *Risk of gun murder for each state in a year*, or (iii) *Median house prices in different suburbs*.
- ▶ **Distributional data:** Data representing many samples from a population. Examples: (i) *Blood sugar levels of patients after taking a new drug treatment*, (ii) *Samples from agricultural field trials to measure the risk of a fungal disease*, or (iii) *Exam marks of students*.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary



Examples of three data types

- ▶ **Proportions** (Percentage data): Data representing proportion (or percentage) of the whole population — all proportions (percentages) should add upto one (hundred). Examples: (i) *proportions of total revenue generated by different departments*, (ii) *percentages of total gun deaths for all states*, or (iii) *proportions of reported disease cases by the world health organisation (WHO) for different disease types*.
- ▶ **Point data:** Data representing a single summary. Examples: (i) *Total number of disease cases in a state*, (ii) *Risk of gun murder for each state in a year*, or (iii) *Median house prices in different suburbs*.
- ▶ **Distributional data:** Data representing many samples from a population. Examples: (i) *Blood sugar levels of patients after taking a new drug treatment*, (ii) *Samples from agricultural field trials to measure the risk of a fungal disease*, or (iii) *Exam marks of students*.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary



USA state-wise gun murder data for 2010

```
> library(dslabs)
> data("murders")
> help(murders)
```

US gun murders by state for 2010

Description

Gun murder data from FBI reports. Also contains the population of each state.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary

Details

- state. US state
- abb. Abbreviation of US state
- region. Geographical US region
- population. State population (2010)
- total. Number of gun murders in state (2010)

Figure 2: Description of *murders* dataset.

Summary of murders data

```
> summary(murders)
   state          abb          region
Length:51    Length:51    Northeast : 9
Class :character Class :character South      :17
Mode  :character Mode  :character North Central:12
                                         West      :13

population      total
Min.   : 563626  Min.   : 2.0
1st Qu.: 1696962 1st Qu.: 24.5
Median  : 4339367 Median  : 97.0
Mean    : 6075769 Mean   :184.4
3rd Qu.: 6636084 3rd Qu.:268.0
Max.   :37253956 Max.   :1257.0
```

Figure 3: Summary of gun violence deaths (*total*) for 51 states in 4 regions.

Outline
Data types
Three most common data types
Proportions data
Comparing Proportions for many categories
Point Data
Stacking principle and bar plot
Point charts
Distributional data
Comparing distributions for many categories
Summary

Get a peek at **murders** dataset

state	abb	region	population	total
Alabama	AL	South	4779736	135
Alaska	AK	West	710231	19
Arizona	AZ	West	6392017	232
Arkansas	AR	South	2915918	93
california	CA	West	37253956	1257
Colorado	CO	West	5029196	65
Vermont	VT	Northeast	625741	2
Virginia	VA	South	8001024	250
Washington	WA	West	6724540	93
West Virginia	WV	South	1852994	27
Wisconsin	WI	North Central	5686986	97
Wyoming	WY	West	563626	5

Figure 4: First and last six observations of the **murders** data.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary

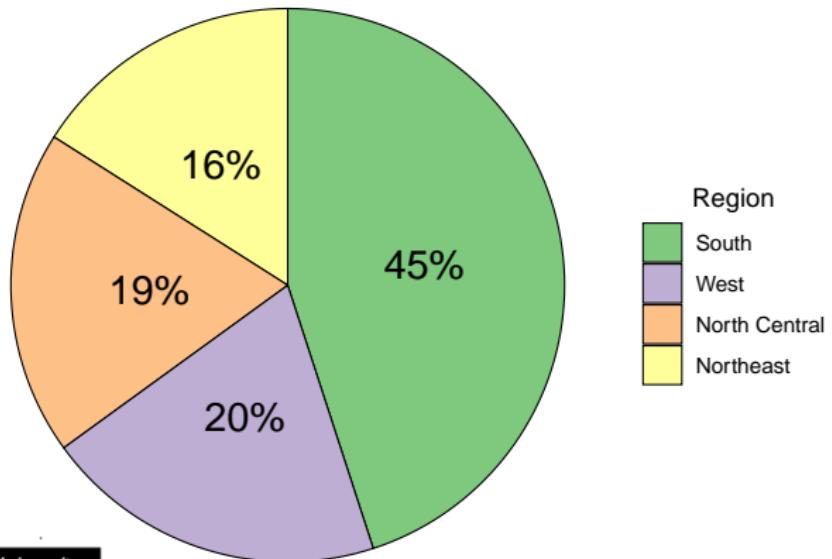


Proportions data example

Suppose we are asked to show the **percentages** of total gun deaths in 2010 for four regions.

Pie chart is a standard tool to display this information.

Regional breakdown of 2010 gun murders



Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary



Pros and Cons of Pie chart

Pie charts have a really bad rap in the academic circle, but industry people love their pie charts.

1. Shortcomings

- 1.1 Pie slices can be imprecise, as they are made using angles.
- 1.2 Pie slices with similar percentages can be hard to distinguish.
- 1.3 Human eyes are better suited to distinguish objects based on length or size than angles.
- 1.4 Pie charts can be easily misused: (i) percentages not adding up to 100, (ii) too many slices of pie, (iii) creation of 3D pies.

2. Advantages

- 2.1 Extremely intuitive to any audience if the number of classes used are small.
- 2.2 Conveys the main message straightaway.

Outline
Data types
Three most common data types
Proportions data
Comparing Proportions for many categories
Point Data
Stacking principle and bar plot
Point charts
Distributional data
Comparing distributions for many categories
Summary

Cousins of Pie — Waffle and Donut charts



Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary



Waffle chart of murders data

- ▶ Use if one needs more precision than Pie chart.
- ▶ Encodes proportions in area, not in angles.
- ▶ Squares can be counted to compute the percentages — each waffle piece represents 1% of the total.

Outline
Data types
Three most common data types
Proportions data
Comparing Proportions for many categories
Point Data
Stacking principle and bar plot
Point charts
Distributional data
Comparing distributions for many categories
Summary

Waffle chart: Regional breakdown of 2010 gun murders

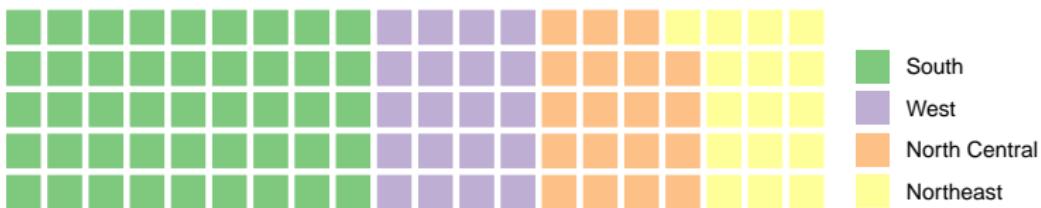
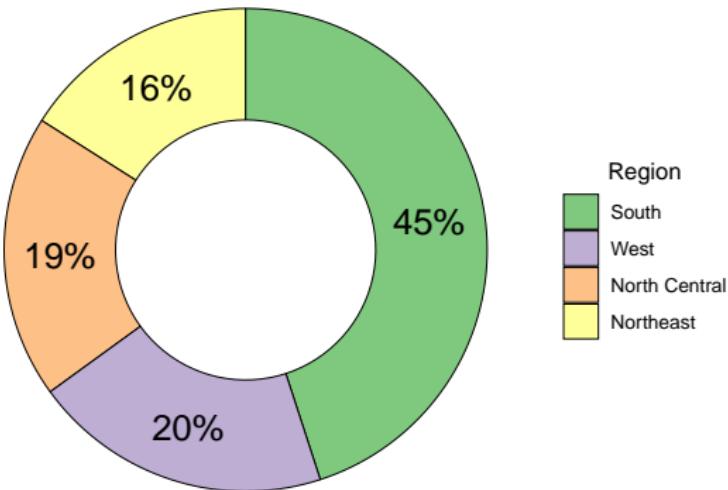


Figure 5: South: 45% ; West: 20% ; North Central: 19% ; Northeast: 16%.

Donut chart of murders data

Donut chart: Regional breakdown of 2010 gun murders



- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Figure 6: Donut chart is considered more aesthetically pleasing to eyes than its older cousin Pie chart.

Disease data from WHO

region	countryCode	country	disease	year	cases
EMR	AFG	Afghanistan	measles	2016	638
EUR	ALB	Albania	measles	2016	17
AFR	DZA	Algeria	measles	2016	41
EUR	AND	Andorra	measles	2016	0
AFR	AGO	Angola	measles	2016	53
AMR	ATG	Antigua and Barbuda	measles	2016	0
WPR	VUT	Vanuatu	yfever	1980	0
AMR	VEN	Venezuela (Bolivarian Republic of)	yfever	1980	4
WPR	VNM	Viet Nam	yfever	1980	0
EMR	YEM	Yemen	yfever	1980	0
AFR	ZMB	Zambia	yfever	1980	0
AFR	ZWE	zimbabwe	yfever	1980	0

Figure 7: First and last six rows of the disease data published by World Health Organisation (WHO).

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary



Summary of WHO's Disease data

region	countryCode	country	
AFR :10481	AFG : 223	Afghanistan	: 223
AMR : 7805	AGO : 223	Albania	: 223
EMR : 4683	ALB : 223	Algeria	: 223
EUR :11819	AND : 223	Andorra	: 223
SEAR: 2453	ARE : 223	Angola	: 223
WPR : 6021	ARG : 223	Antigua and Barbuda:	223
	(Other):41924	(other)	:41924

disease	year	cases
diphtheria:7178	Min. :1980	Min. : 0
measles :7178	1st Qu.:1991	1st Qu.: 0
mumps :3686	Median :2001	Median : 0
pertussis :7178	Mean :2000	Mean : 1816
polio :7178	3rd Qu.:2009	3rd Qu.: 35
rubella :3686	Max. :2016	Max. :1122285
yfever :7178		

Figure 8: Incidences of seven diseases across six regions over the years between 1980 and 2016.

- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Disease distributions across Regions

Our aim is to compare the distributions of different diseases across the seven regions of the world for the year 2016.

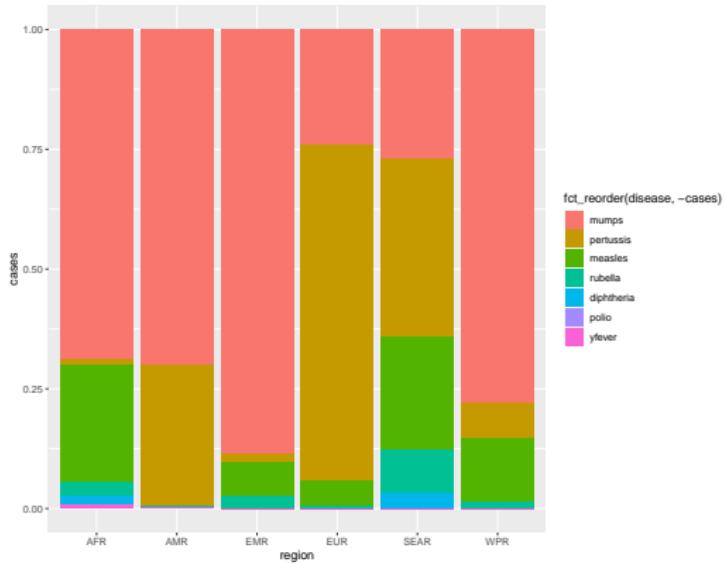
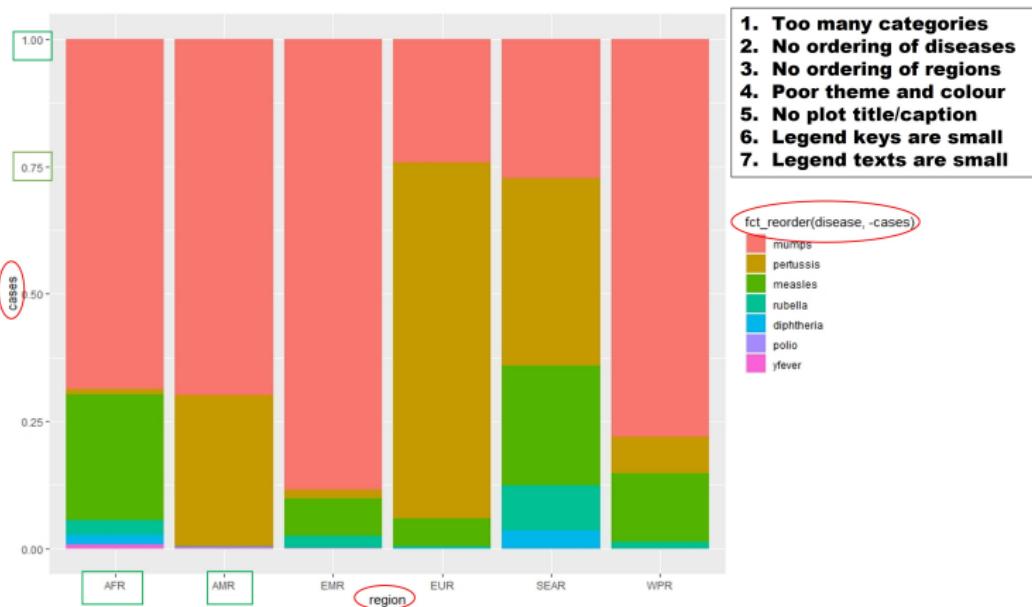


Figure 9: A barebone stacked barplot comparing diseases across regions.

- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

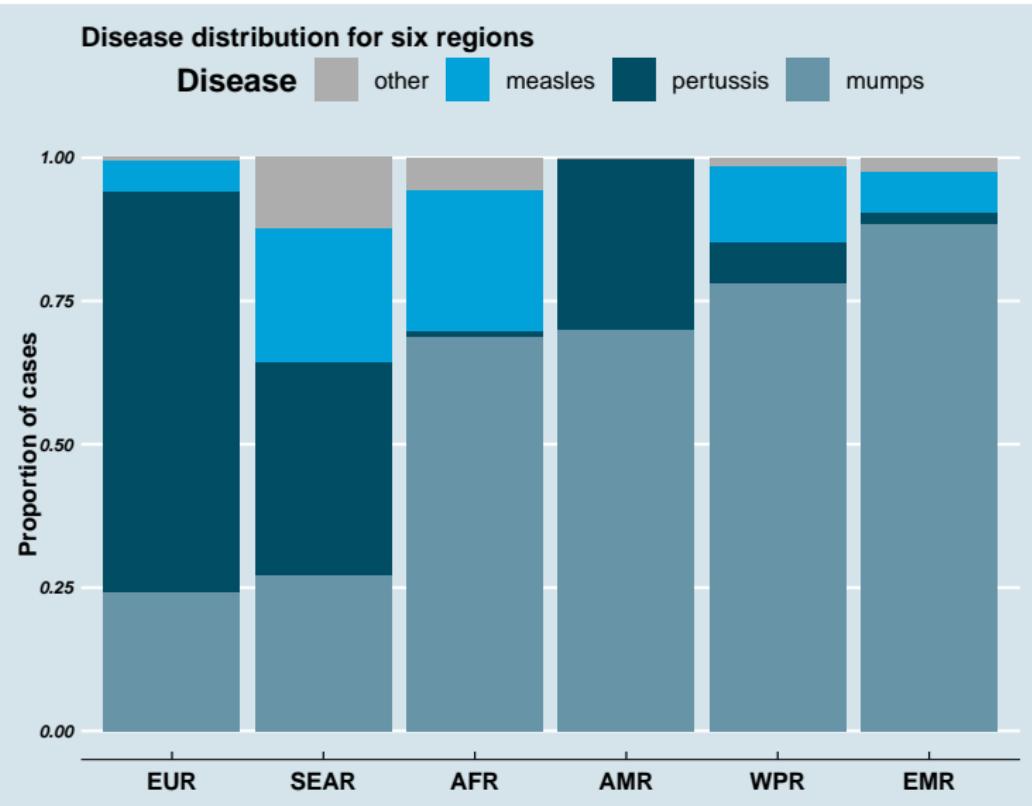
Problems with our stacked barplot

There are several visualisation issues with our plot.



- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Enhanced stacked barplot

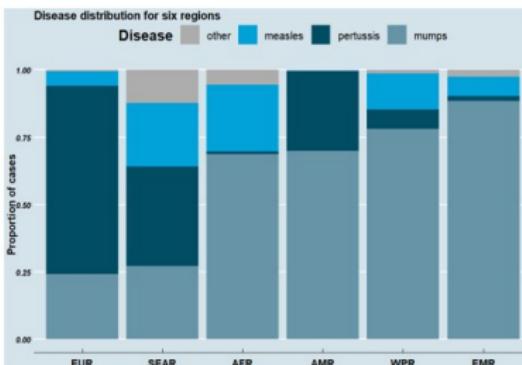


- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary



Changes made to enhance the plot

1. Reduced number of groups, and focused on most prevalent diseases.
2. Ordered diseases from most (mumps) to least prevalent (measles).
3. 'Other' category is of least importance --- put at the top of the bar.
4. The regions are ordered based on most proportion of mumps cases to least proportion mumps cases.
5. Selected a theme used in the famous Economist journal.
6. Increased the size of legend keys, legend texts, and axis texts.
7. Added an informative title to the plot.



Outline
Data types
Three most common data types
Proportions data
Comparing Proportions for many categories
Point Data
Stacking principle and bar plot
Point charts
Distributional data
Comparing distributions for many categories
Summary

Inspecting association between factors

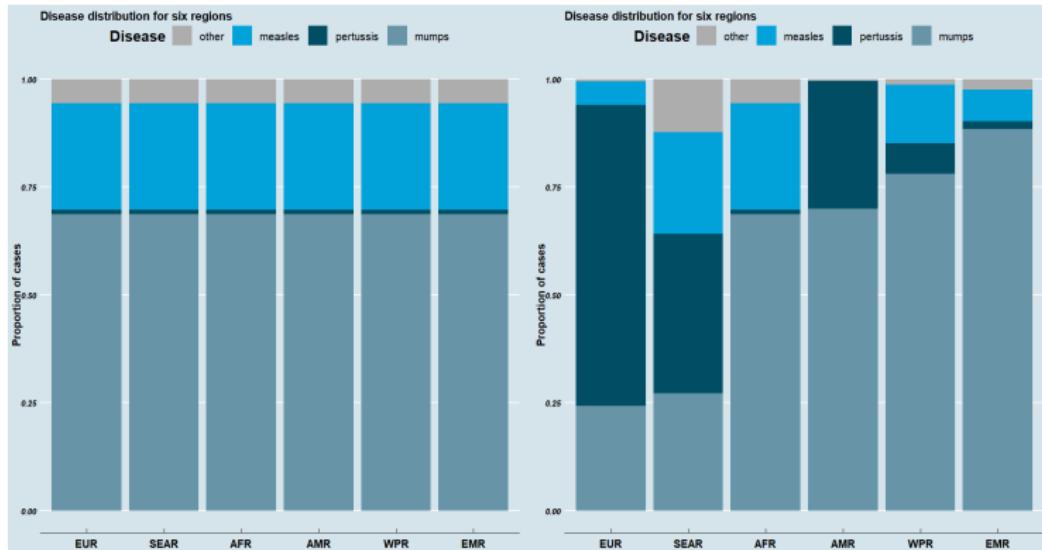


Figure 10: Left: The factors Disease and Region are not associated; Right: There is some relationship between factors Disease and Region — Disease prevalence does differ based on Region.

- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Point Data

- ▶ Point data represents **single observation** for each category. For example, (i) number of disease cases for each country, (ii) gun deaths per 1000 people in each state, or (iii) logarithm of GDP value for each country.
- ▶ Two charts are used typically to display such data — (i) **Bar chart** and (ii) **Point chart**.
- ▶ **Bar charts** are appropriate for variables that have some notion of stacking or accumulation. For example (i) the number of incidents of a particular disease or (ii) project expenditures in dollar amount.
- ▶ **Bar charts** are not appropriate for variables that do not have this property of addition or accumulation. For example, odds ratios, percentiles, or any non-linear transformations should be displayed using Point charts.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary



Point Data

- ▶ Point data represents **single observation** for each category.
For example, (i) number of disease cases for each country,
(ii) gun deaths per 1000 people in each state, or (iii)
logarithm of GDP value for each country.
- ▶ Two charts are used typically to display such data — (i) **Bar chart** and (ii) **Point chart**.
- ▶ **Bar charts** are appropriate for variables that have some notion of stacking or accumulation. For example (i) the number of incidents of a particular disease or (ii) project expenditures in dollar amount.
- ▶ **Bar charts** are not appropriate for variables that do not have this property of addition or accumulation. For example, odds ratios, percentiles, or any non-linear transformations should be displayed using Point charts.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary



Point Data

- ▶ Point data represents **single observation** for each category. For example, (i) number of disease cases for each country, (ii) gun deaths per 1000 people in each state, or (iii) logarithm of GDP value for each country.
- ▶ Two charts are used typically to display such data — (i) **Bar chart** and (ii) **Point chart**.
- ▶ **Bar charts** are appropriate for variables that have some notion of stacking or accumulation. For example (i) the number of incidents of a particular disease or (ii) project expenditures in dollar amount.
- ▶ **Bar charts** are not appropriate for variables that do not have this property of addition or accumulation. For example, odds ratios, percentiles, or any non-linear transformations should be displayed using Point charts.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary



Point Data

- ▶ Point data represents **single observation** for each category. For example, (i) number of disease cases for each country, (ii) gun deaths per 1000 people in each state, or (iii) logarithm of GDP value for each country.
- ▶ Two charts are used typically to display such data — (i) **Bar chart** and (ii) **Point chart**.
- ▶ **Bar charts** are appropriate for variables that have some notion of stacking or accumulation. For example (i) the number of incidents of a particular disease or (ii) project expenditures in dollar amount.
- ▶ **Bar charts** are not appropriate for variables that do not have this property of addition or accumulation. For example, odds ratios, percentiles, or any non-linear transformations should be displayed using Point charts.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary



Stacking Principle

Bar charts should be used to represent data that have some sort of accumulating property to them. You should ask yourself *if you could stack the units of the measure on top of each other*. For example, revenue (money) generated by different projects can be thought of as a *stack of dollar bills*.



- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Point Data Example from **murders** data

Question: Identify top 10 states based on USA gun murder deaths recorded in the **murders** dataset, and create an appropriate chart to present these data.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary



Horizontal Bar Plots are excellent

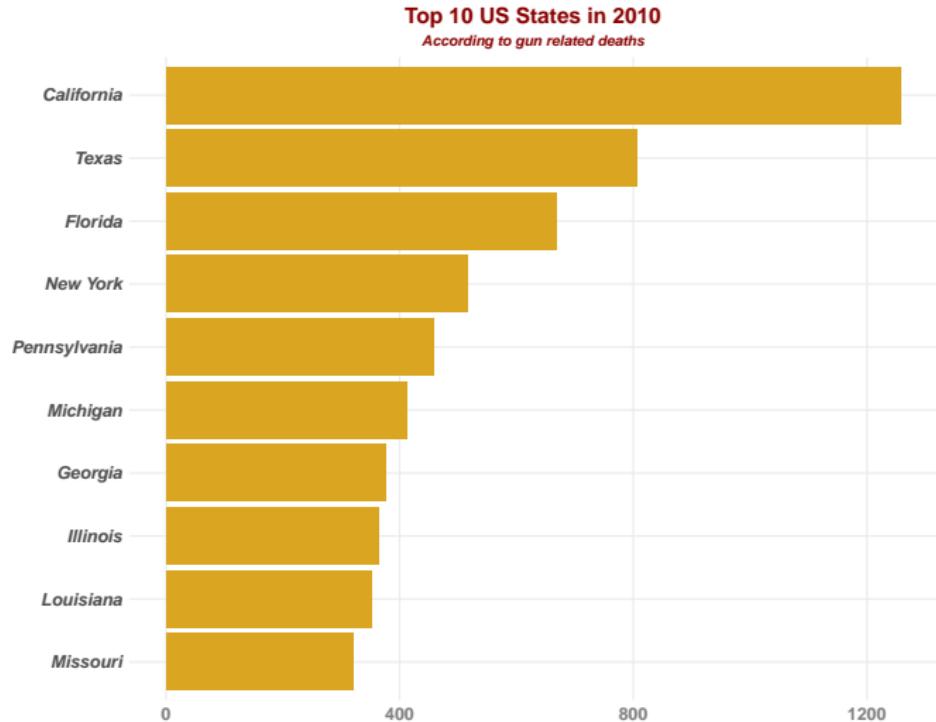


Figure 11: Axis of bar charts should start from zero.

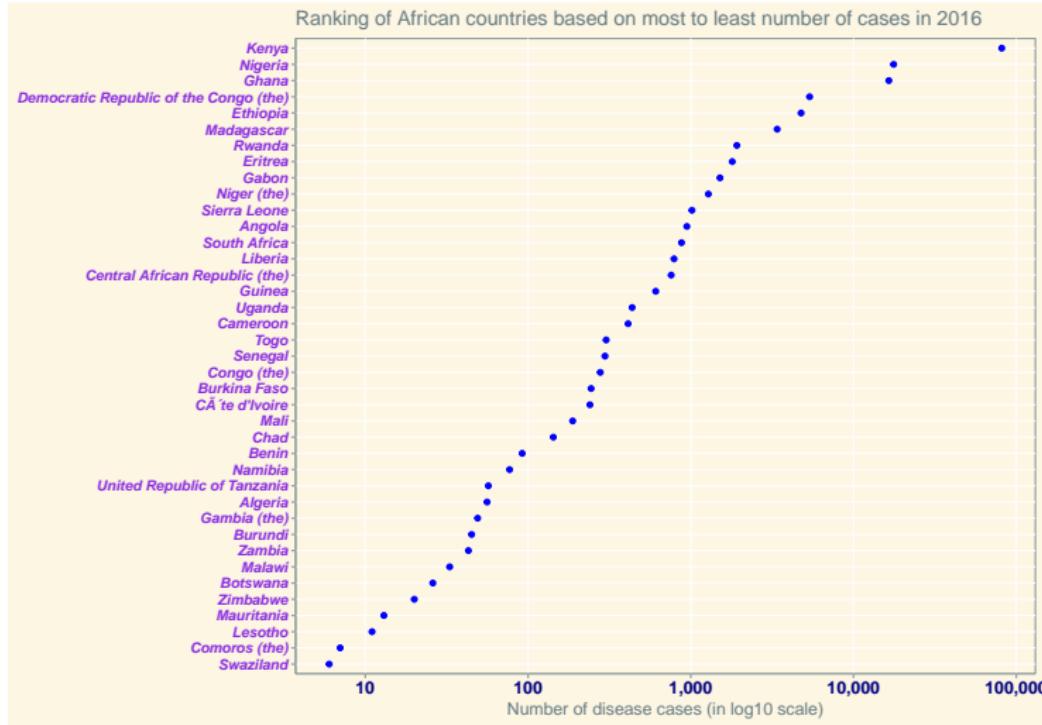
- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Point Charts

- ▶ When we have point data that do not satisfy the stacking principle, we should use **Point charts** to illustrate these data.
- ▶ Many point data are not stackable, for example ratios, percentiles or different sensor measurements such as NDVI, soil moisture content, or temperature.
- ▶ Non-linear transformations such as logarithm, square-root or exponentially transformed data should be plotted using point charts.
- ▶ Easy to construct — simply remove the bar and replace the top of the bar with a point.

Outline
Data types
Three most common data types
Proportions data
Comparing Proportions for many categories
Point Data
Stacking principle and bar plot
Point charts
Distributional data
Comparing distributions for many categories
Summary

Point chart example (\log_{10} (cases))



Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary

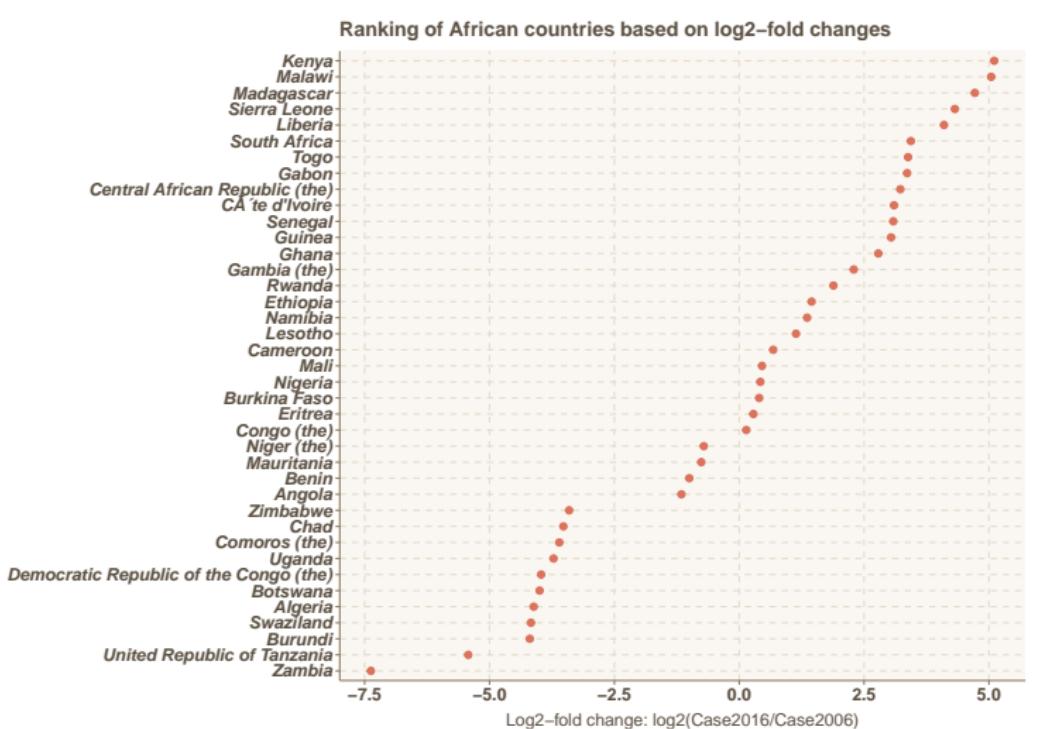
Point Data Example from disease data

Question: Rank countries from the AFR region based on the \log_2 -fold change in the number of cases from 2006 to 2016. The \log_2 -fold change is given by:

$$\log_2 \left(\frac{\text{Case2016}}{\text{Case2006}} \right).$$

- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Log2-fold change of African nations



- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Properties of log2-fold changes

- ▶ Log-fold change is **symmetric around zero**.
- ▶ Value of log2-fold change of 1 means **two-times larger**, while the value of -1 means **two-times smaller**.
- ▶ The value of 0 is the **focal/turning point**, where the decreasing number of cases switches to increasing number of cases.
- ▶ **Adding a line at the 0 value** would provide a visual aid to distinguish countries with declining number of disease cases from those with increasing number of disease cases.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

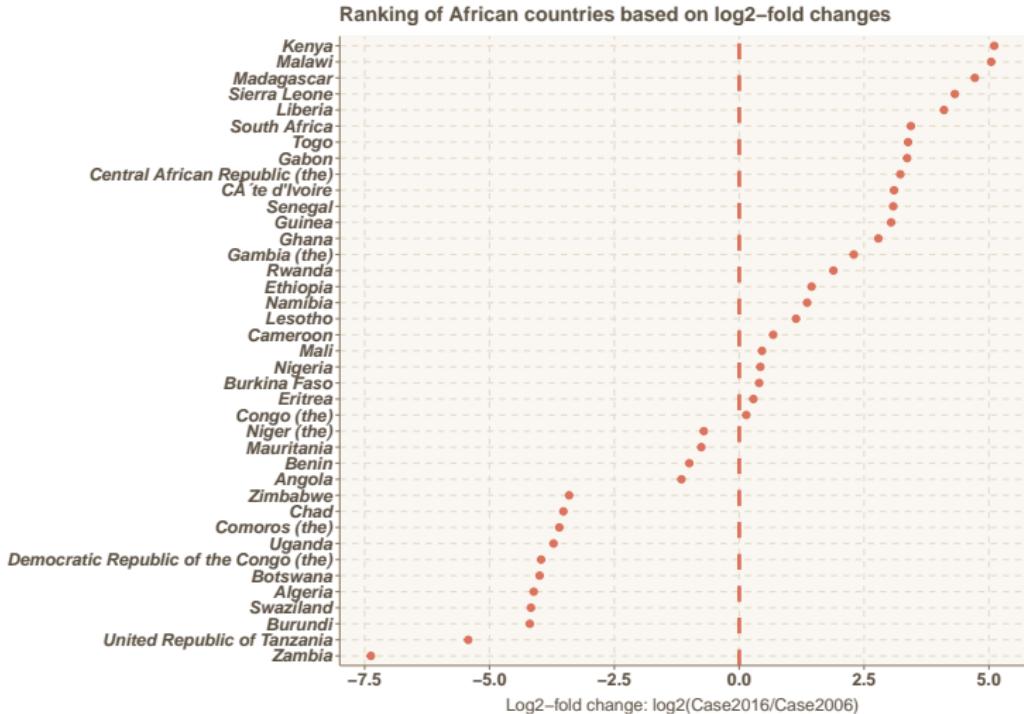
Point charts

Distributional data

Comparing distributions for many categories

Summary

Include an anchor at the focal point



- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Data from a single distribution

The distributional data are observed when several **samples** are gathered from a population. The **shape** and **percentiles** of the distribution can often reveal very important and interesting facts about the population. For example (i) *age distribution of the supporters of a political party may help to design a targeted campaigning strategy*, (ii) *income distribution of a state may help to design various socio-economic programs by the state government*, or (iii) *age distribution of infected population may help to design an appropriate vaccination plan*.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary



Distributional data example: iris data

```
> head(iris)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

```
> summary(iris)
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
Median :5.800	Median :3.000	Median :4.350	Median :1.300	virginica :50
Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199	
3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800	
Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500	

Figure 12: First six rows and summary of the *iris* dataset.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

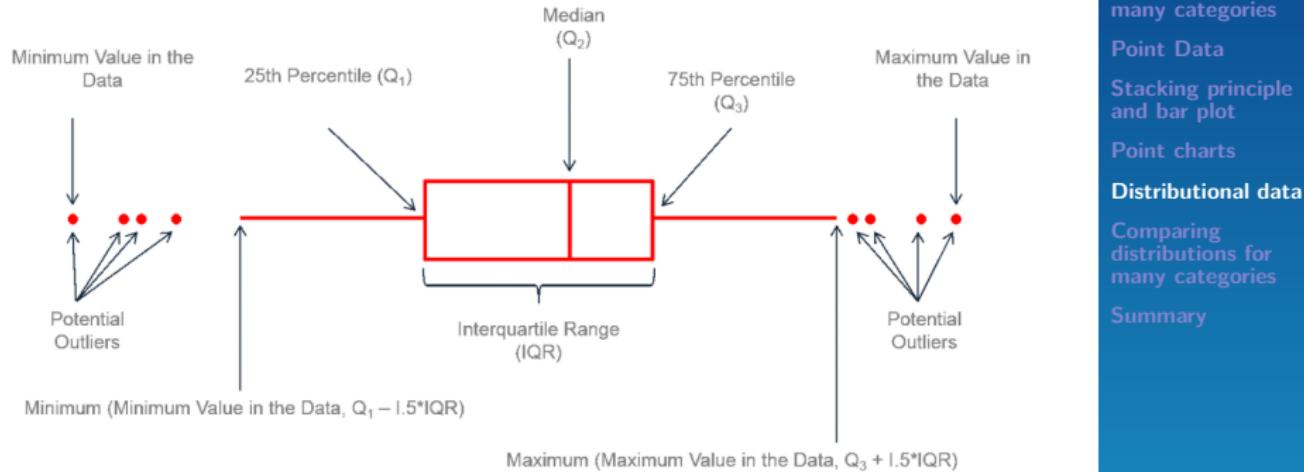
Distributional data

Comparing distributions for many categories

Summary

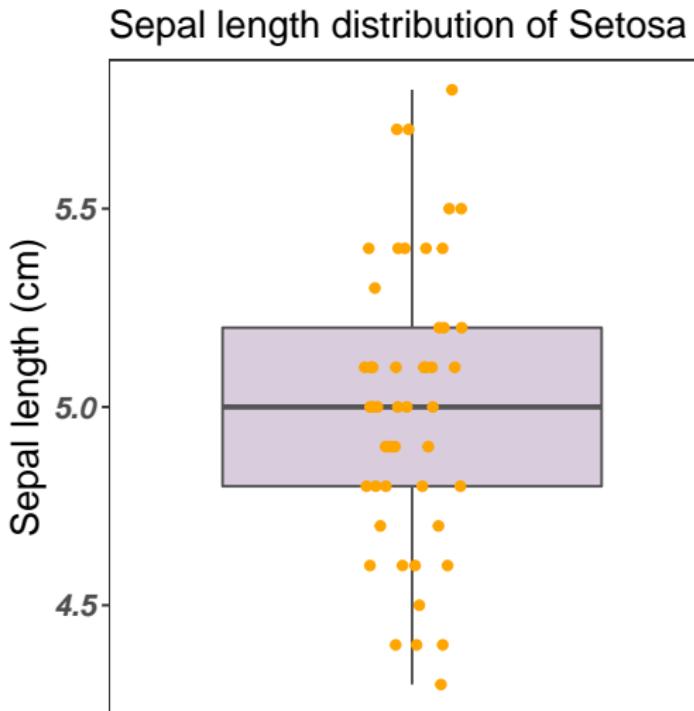
Boxplot for a quick distributional summary

Boxplot provides a **concise visualisation** of the distribution of the data. It shows the **central tendency** (median) and the (inter-quartile) **range** of the data.



- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Boxplot of Sepal.Length data



- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Beyond Boxplot

- ▶ Boxplot summarises the distribution using only five numbers: (i) median, (ii) first quartile, (iii) third quartile, (iv) ‘minimum’, and (v) ‘maximum’.
- ▶ Boxplot does not indicate much about the peaks in the distribution — is the distribution bimodal?
- ▶ To obtain more information about the distribution, we should use the histogram or the kernel density plot. However, boxplots are excellent tool for comparing distributions corresponding to many categories (we will come back to this point a bit later).

Outline
Data types
Three most common data types
Proportions data
Comparing Proportions for many categories
Point Data
Stacking principle and bar plot
Point charts
Distributional data
Comparing distributions for many categories
Summary

Histogram

Histograms are created by first binning the data and then counting the number of observations in each bin. As a data scientist, you should experiment with different binwidth values (or equivalently, with the number of bins).

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

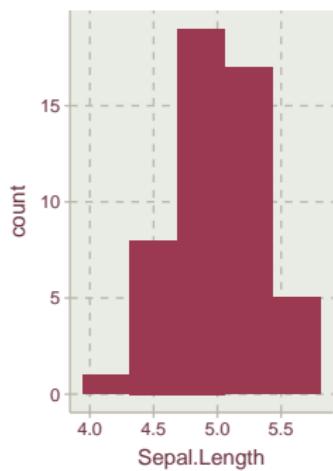
Summary



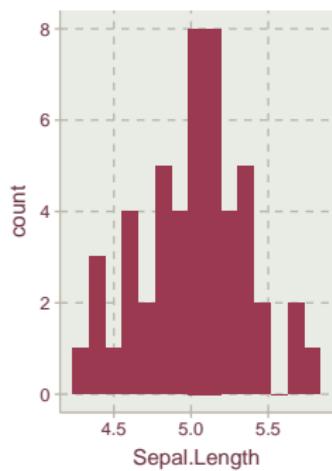
Histograms with different binwidths

The distribution of Sepal length is fairly symmetric, with most values falling around 5.0 cm. in the middle of the distribution. The frequency declines very similarly as we move away from the centre toward the two ends.

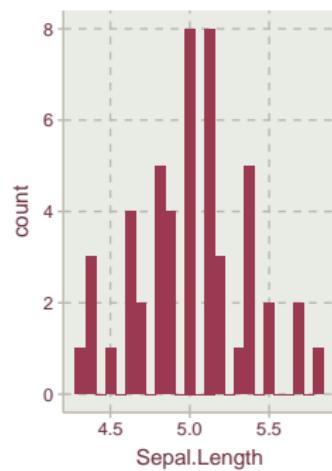
Number of bins = 5



Number of bins = 15



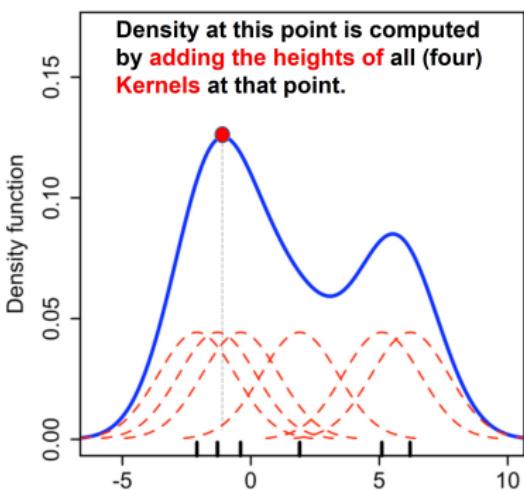
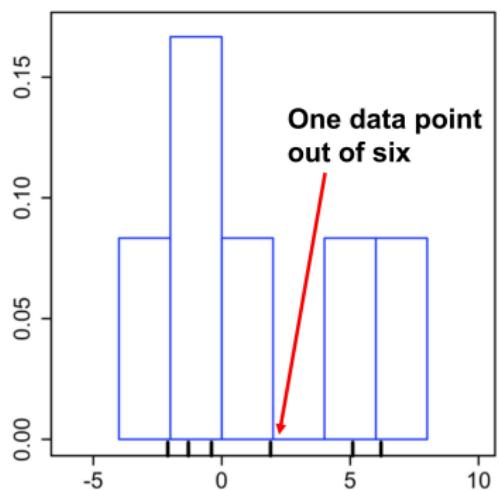
Number of bins = 25



- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Kernel Density Estimator (KDE)

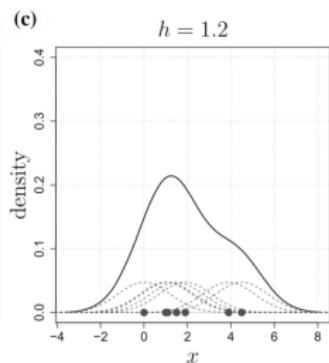
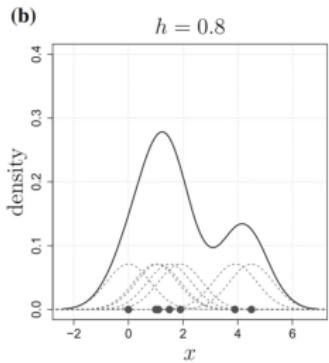
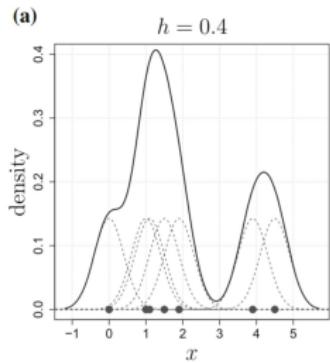
Kernel density estimator produces a smooth curve (**a smoother version of density histogram**) to represent the underlying density.



- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

KDE with different bandwidths

- ▶ Bandwidth is crucial to determine the shape of the estimated density.
- ▶ Too smaller bandwidth will display many peaks (which may be due to sampling bias).
- ▶ Too larger bandwidth will smooth out many potential real peaks.



Outline
Data types
Three most common data types
Proportions data
Comparing Proportions for many categories
Point Data
Stacking principle and bar plot
Point charts
Distributional data
Comparing distributions for many categories
Summary

KDE of Petal.Length data

Kernel Density Estimates of the Petal.Length for three different bandwidth values.

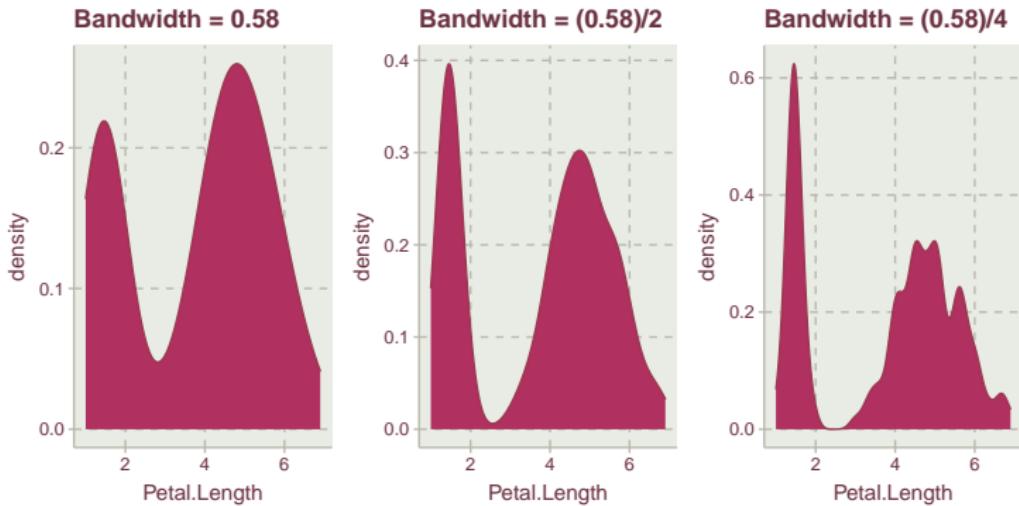


Figure 13: Default bandwidth of 0.58 is chosen in `geom_density()` and calculated using the function `bw.nrd0()`.

Outline
Data types
Three most common data types
Proportions data
Comparing Proportions for many categories
Point Data
Stacking principle and bar plot
Point charts
Distributional data
Comparing distributions for many categories
Summary

Petal Length Density

- ▶ Density of Petal length has two clear modes — one around 1.5 cm and the other around 5 cm.
- ▶ There could be a possible third peak around 5.5 cm, but it could be an artefact due to small bandwidth selection.
- ▶ Distinct peaks in the density may point to a missing factor.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary



Density peaks and missing factor

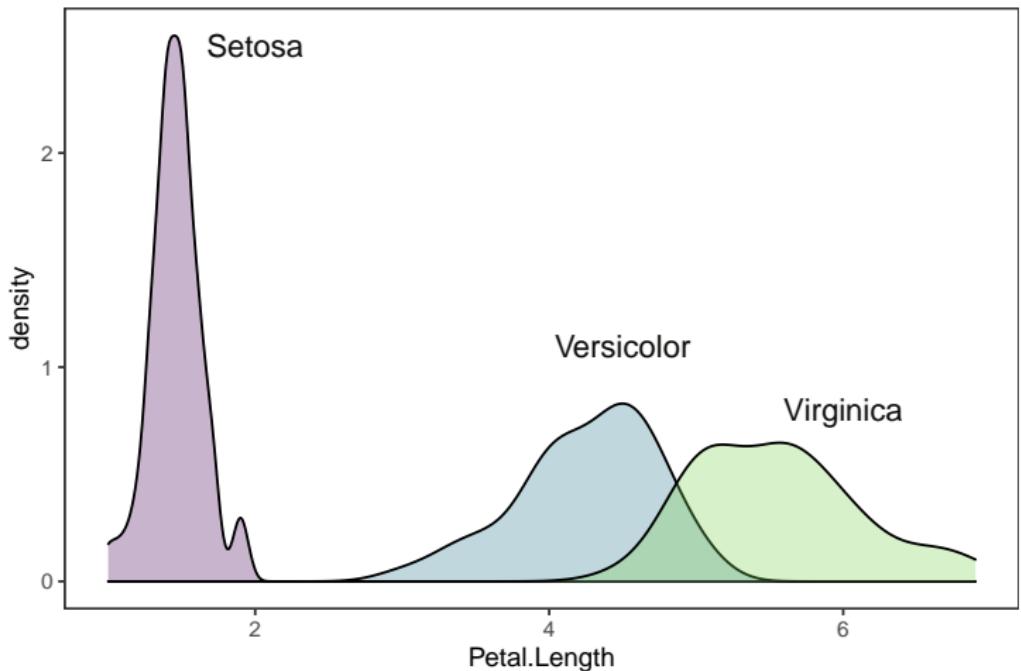
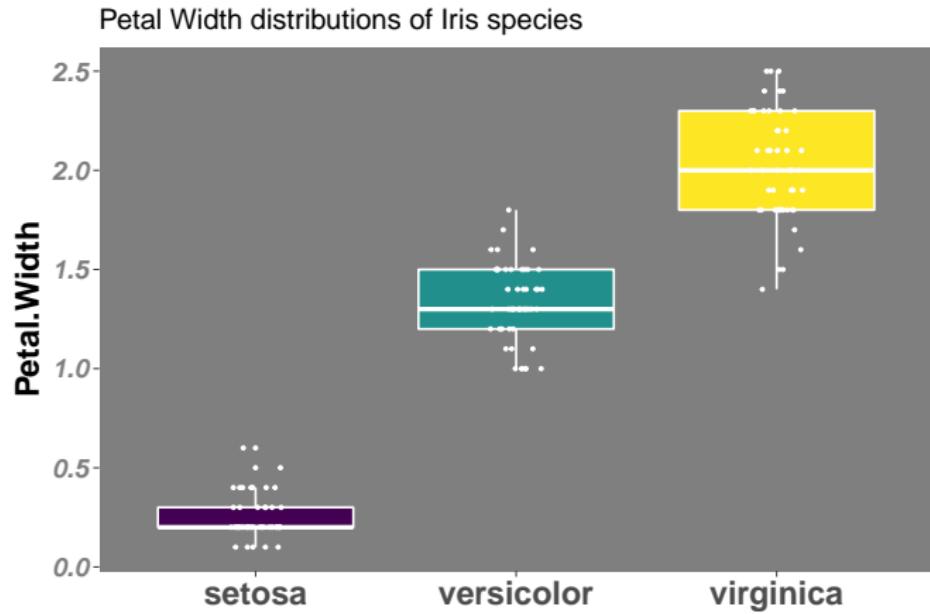


Figure 14: The missing factor is **Species**; however, it is hard to distinguish **Versicolor** and **Virginica** based on Petal Lengths.

- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary

Comparing distributions of many categories

Compare the Petal Width distributions of three iris species
— Boxplots are the most concise way to compare the three distributions.

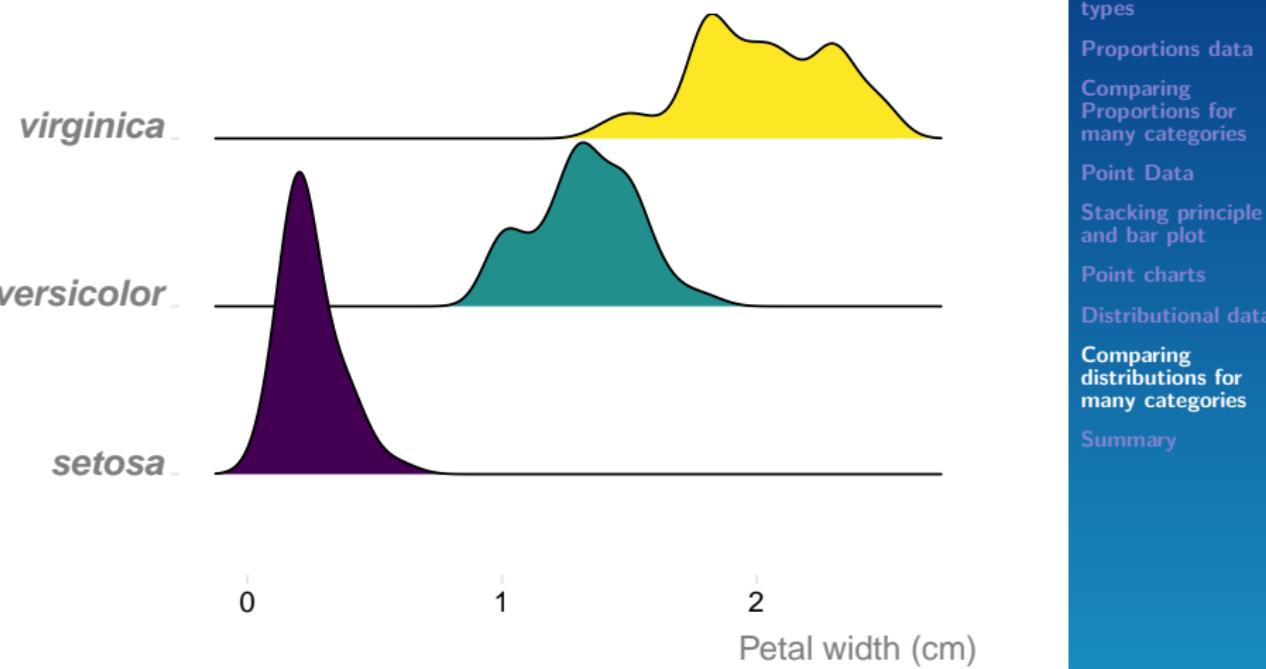


- Outline
- Data types
- Three most common data types
- Proportions data
- Comparing Proportions for many categories
- Point Data
- Stacking principle and bar plot
- Point charts
- Distributional data
- Comparing distributions for many categories
- Summary



KDE for Petal.Width distributions

We can also use **KDE** to compare the Petal Width distributions of three iris species.



Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary



Summary

- ▶ Use pie, waffle or donut charts for proportions data from a single population, but use Stacked Bar plot for comparing proportions data of many populations/categories.
- ▶ For point data satisfying stacking principle, use Bar charts. For other point data, use point charts instead.
- ▶ For distributional data, you can use boxplots, histograms or kernel density plots. You can also use a fancier plot, called the Violin plot, which works similarly as the kernel density plot.

Outline

Data types

Three most common data types

Proportions data

Comparing Proportions for many categories

Point Data

Stacking principle and bar plot

Point charts

Distributional data

Comparing distributions for many categories

Summary

