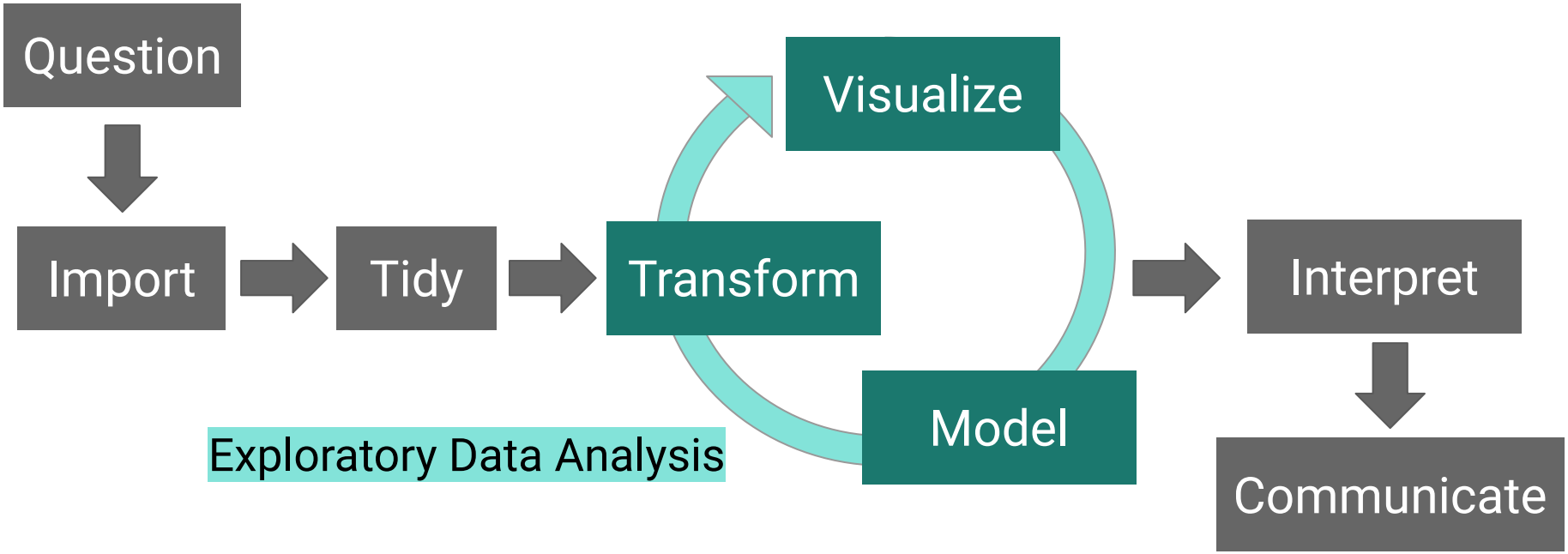
Exploratory Data Analysis

Lecture 7

Objectives

- to perform exploratory data analysis
- to visualize data using ggplot2

Motivation

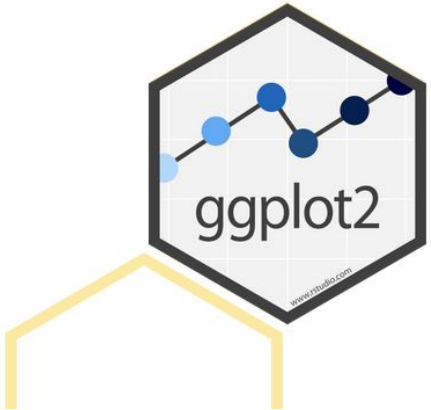


Modified from https://r4ds.had.co.nz/explore-intro.html#explore-intro

ggplot2 package

Tidyverse

Packages



ggplot2

ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. You provide the data, tell ggplot2 how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details. Go to docs...

https://www.tidyverse.org/packages/

EDA as a process

- Generate questions about your data
- Search for answers by visualising, transforming, and modelling your data
- Use what you learn to refine your questions and/or generate new questions

Questions to interrogate your data

- What type of variation occurs within my variables?
- What type of covariation occurs between my variables?

What is a variable?

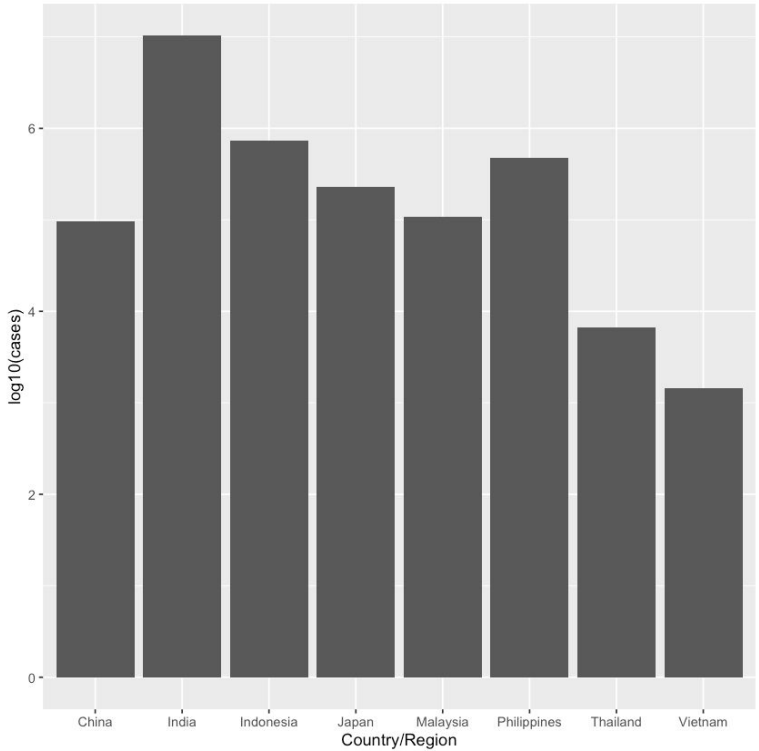
- A **variable** is a quantity, quality, or property that you can measure.
- A **value** is the state of a variable when you measure it.
- The value of a variable may change from measurement to measurement.

Variation

- **Variation** is the tendency of the values of a variable to change across observations.
- Each of your measurements will include a small amount of error that varies from measurement to measurement.
- Variations in data may represent important new findings or only errors in typing or coding which need to be corrected.

Visualizing distributions

- How you visualise the distribution of a variable will depend on whether the variable is categorical or continuous
- A variable is **categorical** if it can only take one of a small set of values.
- Categorical variables are usually saved as factors or character vectors.



Visualizing using **ggplot2**

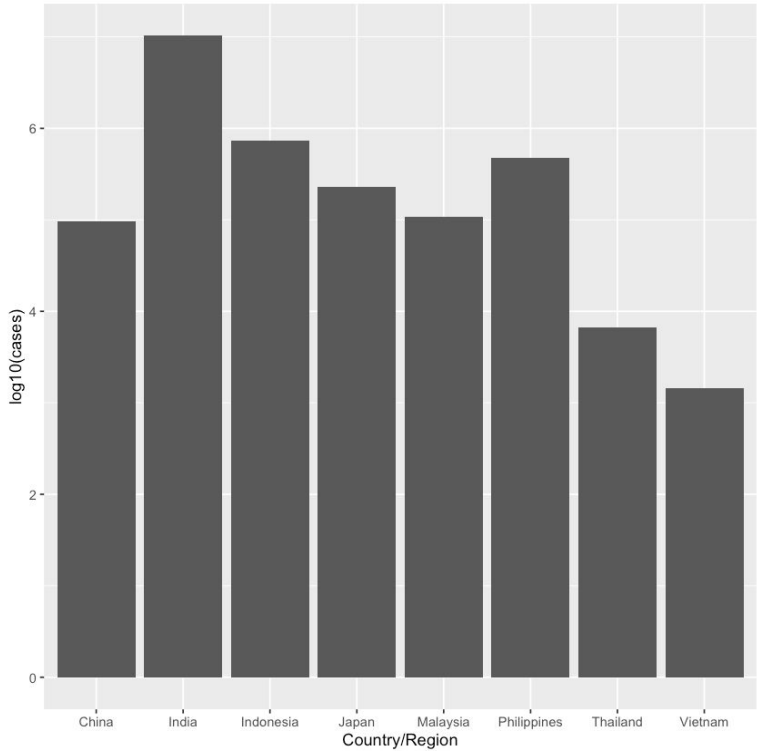
- **data** - in tidy format
- **geometry** - type of graph (scatter plot, bar graph, etc.)
- **aesthetic mapping** - map variables of the data to location of x- and y-axes, color, size

ggplot2 is based on the **grammar of graphics**, the idea that you can build every graph from the same components: a **data** set, a **coordinate system**, and geoms—visual marks that represent data points.

A tibble: 8 × 5

Country/Region	date	cases	population	rate
<chr>	<date>	<dbl>	<dbl>	<dbl>
China	2020-12-29	95797	1402667000	6.829632
India	2020-12-29	10244852	1380004000	742.378428
Indonesia	2020-12-29	727122	273524000	265.834808
Japan	2020-12-29	227415	125769000	180.819598
Malaysia	2020-12-29	108615	32366000	335.583637
Philippines	2020-12-29	471526	109581000	430.299048
Thailand	2020-12-29	6690	69800000	9.584527
Vietnam	2020-12-29	1454	97339000	1.493749

```
> ggplot(data = covid) +
      geom_col(mapping = aes(
            x = `Country/Region`,
            y = log10(cases)
                  )
            )
)
```

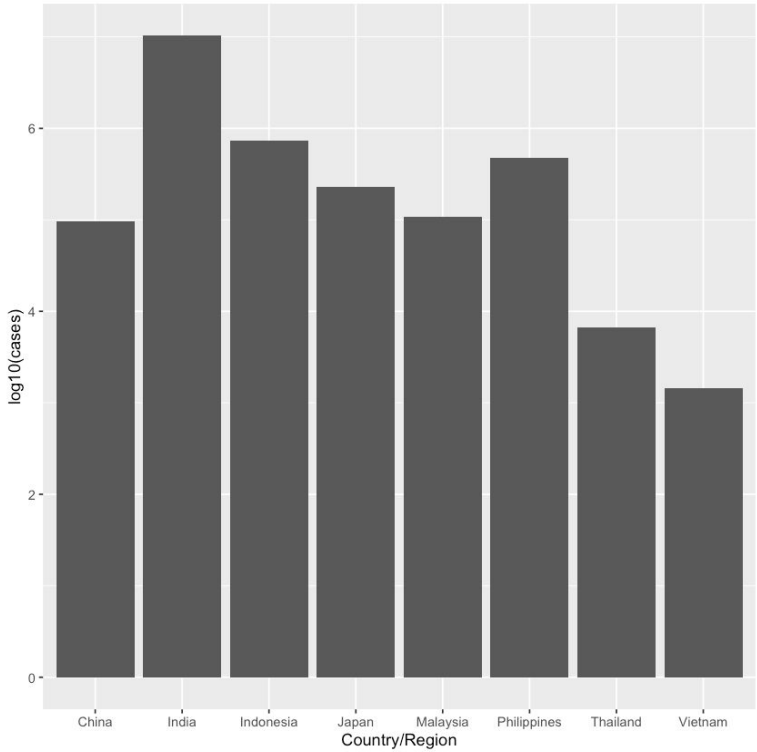


Visualizing distributions

- Visualizations at this stage are rough for personal consumption only
- Not necessarily publication ready

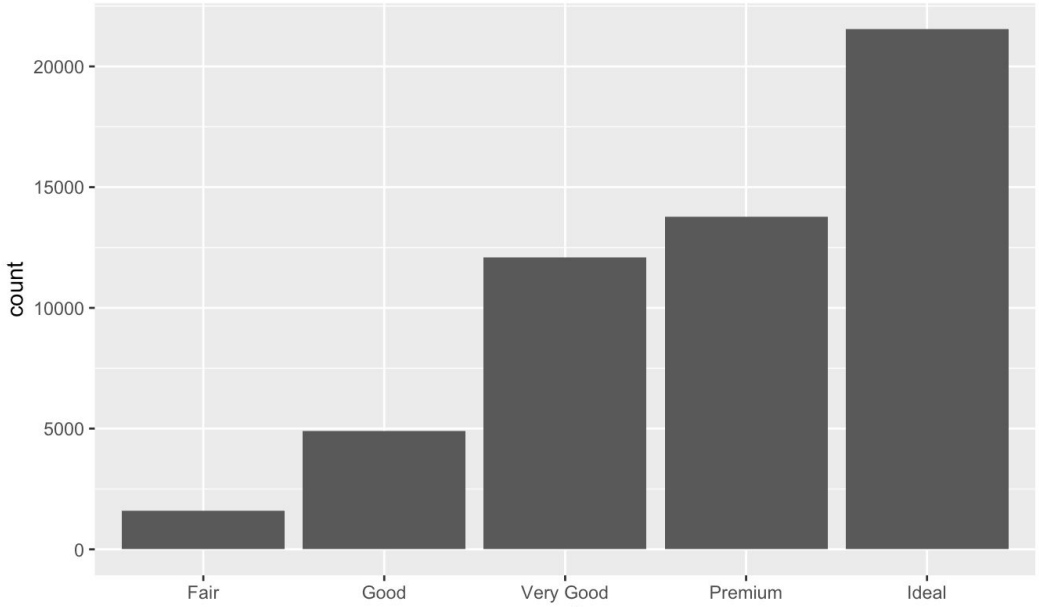
Visualizing categorical variables

- **Nominal** measurements have no intrinsic order and the difference between levels of the variable have no meaning
- Sex, race, exposure category (yes/no)



Visualizing categorical variables

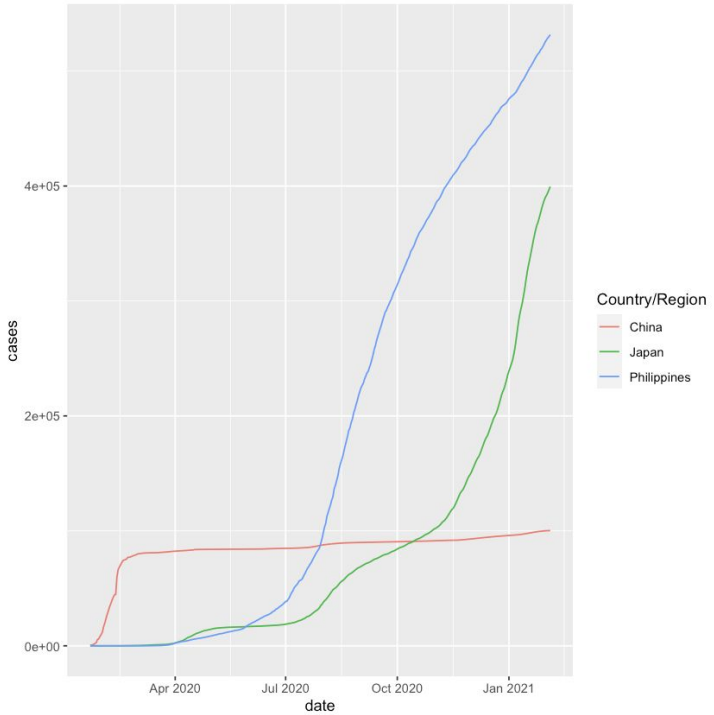
- **Ordinal** variables have an intrinsic order
- “low”, “medium”, “high”



Visualizing discrete variables

- A variable is **discrete** if values are integers (e.g. number of persons, counts)
- Arithmetic-scale line graph to show patterns of trends over some variable, e.g. time

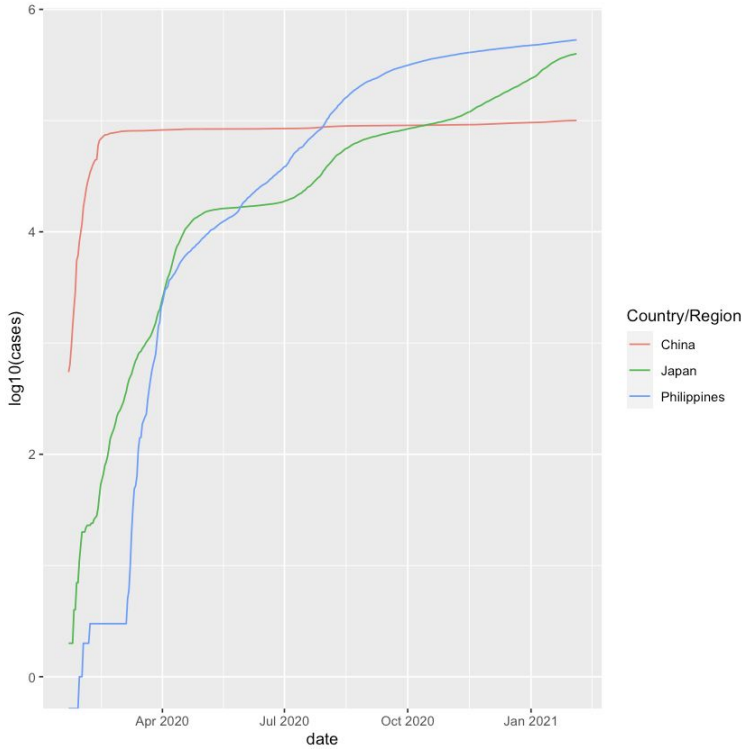
```
> confirmed_covid %>%
      ggplot(aes( x = date,
            y = cases,
            color = `Country/Region`)) +
      geom_line()
```



Visualizing discrete distributions

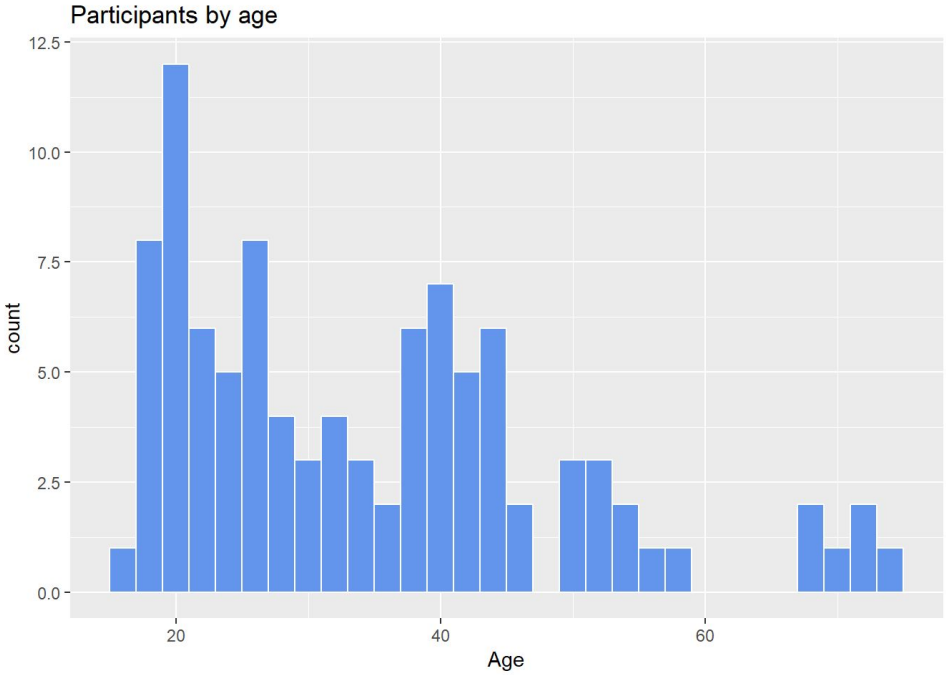
- A variable is **discrete** if values are integers (e.g. number of persons, counts)
- Semi-logarithmic-scale line graph to show patterns with large values

```
> confirmed_covid %>%
      ggplot(aes( x = date,
            y = log10(cases),
            color = `Country/Region`)) +
      geom_line()
```



Visualizing continuous distributions

- A variable is **continuous** if it can take any of an infinite set of ordered values.
- Numbers and date-times are two examples of continuous variables.
- A **histogram** is a graph of the frequency distribution of a continuous variable, based on class intervals.



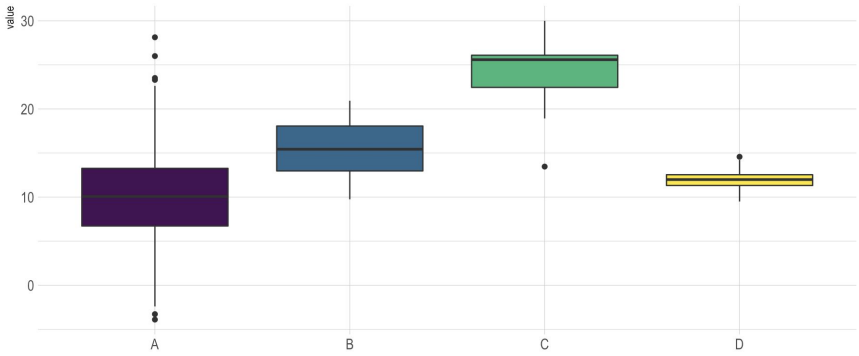
Typical values

- In both bar charts and histograms, tall bars show the common values of a variable, and shorter bars show less-common values
- Places that do not have bars reveal values that were not seen in your data.
- Look for anything unexpected:

- Which values are the most common? Why?
- Which values are rare? Why? Does that match your expectations?
- Can you see any unusual patterns? What might explain them?

Unusual values

- Outliers are observations that are unusual
- Data points that don’t seem to fit the pattern
- Sometimes outliers are data entry errors
- Other times outliers suggest importan new science



https://www.data-to-viz.com/caveat/boxplot.html

Missing values

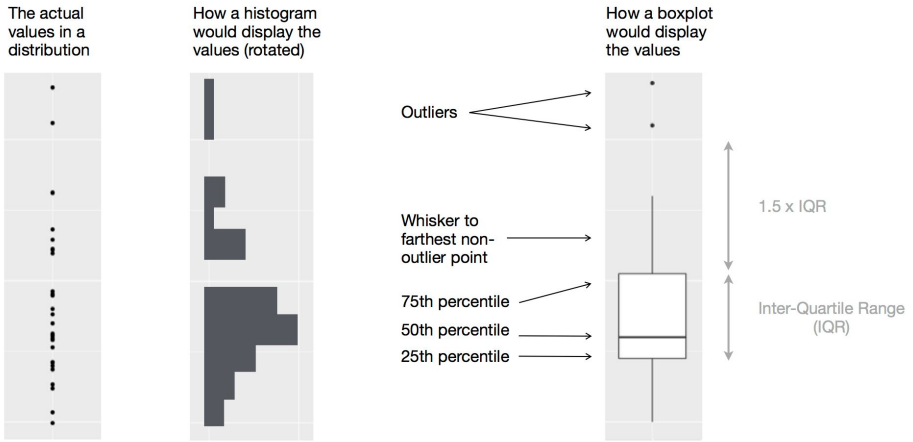
- Drop entire row with unusual values with caution.
- Replace the unusual values with missing values, NA.

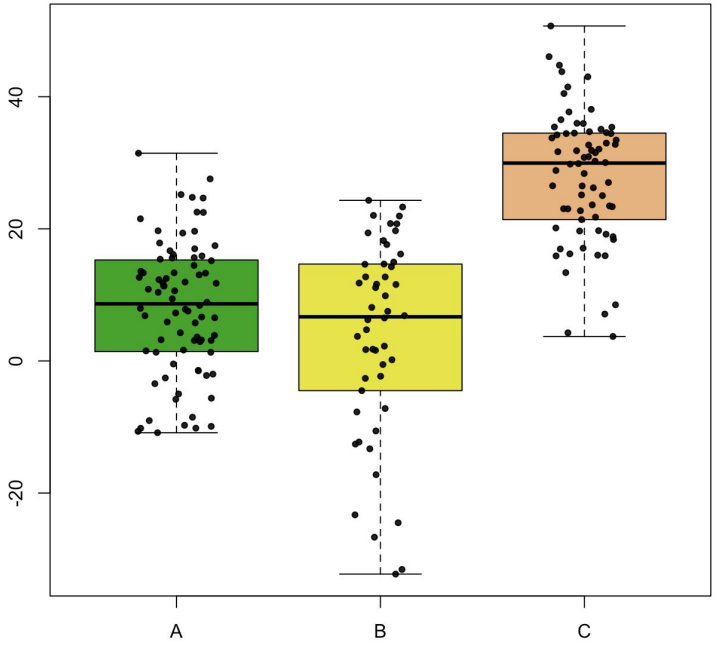
Covariation

- If variation describes the behavior within a variable, covariation describes the behavior between variables.
- Covariation is the tendency for the values of two or more variables to vary together in a related way.
- The best way to spot covariation is to visualise the relationship between two or more variables.

Covariation: a categorical and continuous variable

boxplot

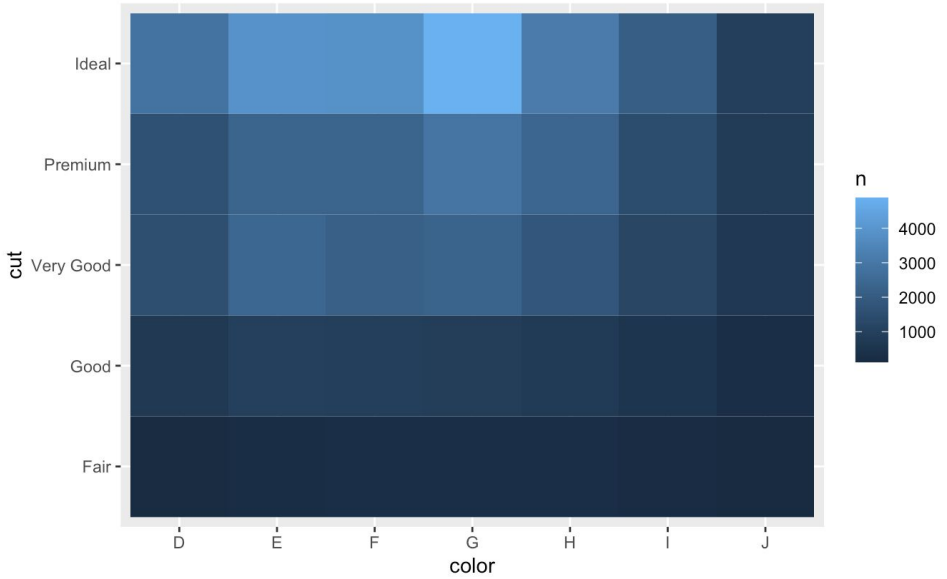




https://www.r-graph-gallery.com/96-boxplot-with-jitter.html

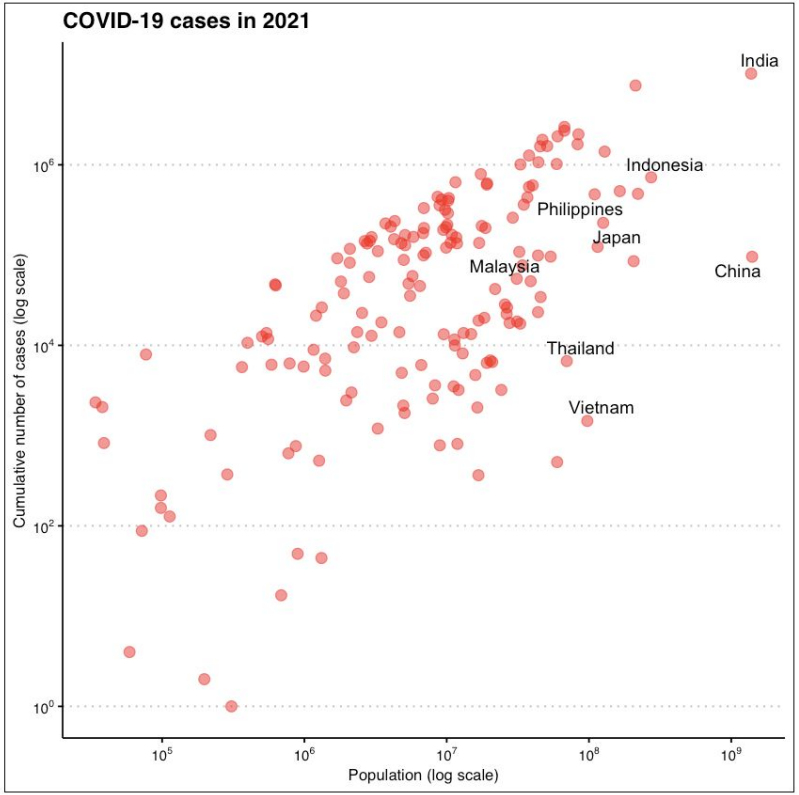
Covariation: two categorical variables

- To visualize the covariation between two categorical variables, you'll need to count the number of observations for each combination.



Covariation: two continuous variables

scatter plot



Pattern and models

- Patterns in your data provide clues about relationships.
- If a systematic relationship exists between two variables it will appear as a pattern in the data.

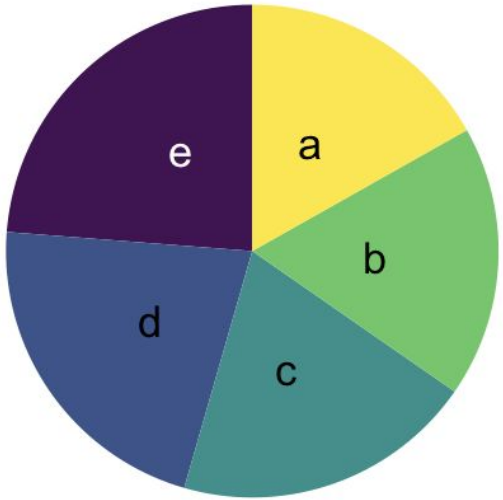
- Could this pattern be due to coincidence (i.e. random chance)?
- How can you describe the relationship implied by the pattern?
- How strong is the relationship implied by the pattern?
- What other variables might affect the relationship?
- Does the relationship change if you look at individual subgroups of the data?

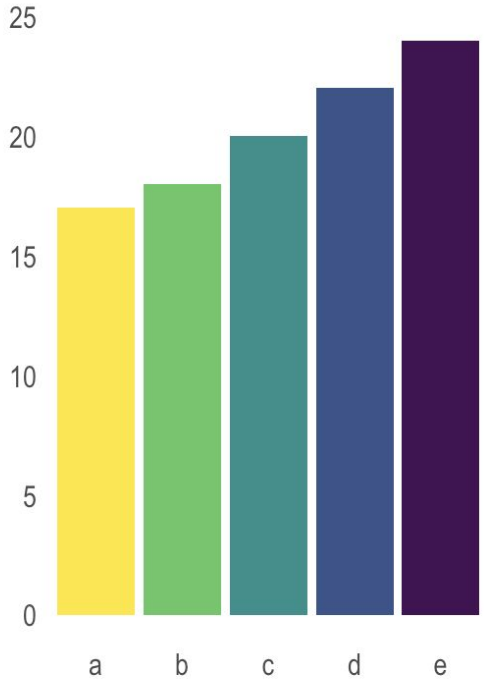
Guide to Selecting a Graph or Chart to Illustrate Epidemiologic Data

Type of Graph or Chart	When to Use
Arithmetic scale line graph	Show trends in numbers or rates over time
Semilogarithmic scale line graph	Display rate of change over time; appropriate for values ranging over more than 2 orders of magnitude
Histogram	Show frequency distribution of continuous variable; for example, number of cases during epidemic (epidemic curve) or over longer period of time
Frequency polygon	Show frequency distribution of continuous variable, especially to show components
Cumulative frequency	Display cumulative frequency for continuous variables
Scatter diagram	Plot association between two variables
Simple bar chart	Compare size or frequency of different categories of a single variable
Grouped bar chart	Compare size or frequency of different categories of 2 4 series of data
Stacked bar chart	Compare totals and illustrate component parts of the total among different groups
Deviation bar chart	Illustrate differences, both positive and negative, from baseline
100% component bar chart	Compare how components contribute to the whole in different groups
Pie chart	Show components of a whole
Spot map	Show location of cases or events
Area map	Display events or rates geographically
Box plot	Visualize statistical characteristics (median, range, asymmetry) of a variable's distribution

https://www.cdc.gov/csels/dsepd/ss1978/lesson4/summary.html

No appetite for pie charts?





https://www.data-to-viz.com/caveat/pie.html

Take-away message

- perform EDA first before any analysis
- graphs do not have to be pretty