Lecture 9: Tables and plots

Concise data summary

Dr Milan Valášek

23 November 2020

UNIVERSITY OF SUSSEX

Overview

Why use tables and plots?

Tables

- What makes a good table?
- Frequency tables
- Summary tables

Plotting basics

- What makes a good plot?
- Structure of a plot
- Frequency plots
- Summary plots

All plots and tables in the slides made in **R**

Data

- 4,468 films pulled from IMDB.com
- Inclusion criteria
    - Rated at least 5,000 times
    - Released in 2010 or later
- Data available here

Of the 4,468 titles, 52 (1.16%) were produced, or co-produced by at least one African country, 3,009 (67.35%) by at least one country in the Americas, 1,058 (23.68%) by a country in Asia, 1,604 (35.9%) in Europe, and 153 (3.42%) in Oceania. The sum of these numbers is necessarily higher than the total number of titles in the data set, as one title can have multiple production attributions and so can count towards several "Continent" categories.

The number of IMDB ratings ranged from 5,003 to 2.03×10^{6}, with a mean of 6.31×10^{4} and *SD* of 1.23×10^{5}. The average user rating for a given title spanned the range 1–9.2; M=6.37, SD=1.02.

Information on estimated budget was only available for 2,170 of the titles in the data set: 18–3.56×10^{8}, M=3.96×10^{7}, SD=5.09×10^{7}.

Why and when to use tables and plots?

Why

- Plots and tables allow us to convey a lot of information using relatively small amounts of space
- They structure the information we're communicating so that it's easier to understand than a wall of text
- Good tables and plots are simply #aesthetic

When

- Tables and plot are not just for reports
- They are a good way of exploring data before analysis in order for us to get to know them
- Not all plots and table we create should be put in reports/papers
- If we are including them in reports/papers, they should be used to convey important information that would be cumbersome to convey in body text

Tables

- Tidy way of presenting a lot of numbers
- A good table should be easy to read, well-organised, and clear
- Good for exploring and summarising data (this lecture), and presenting results (future modules)

Structural elements

- Number: all tables should be numbered and the number should be referenced in paper/report
- Title: should be descriptive
- Header: clearly indicates what the data in each column mean
- Body: logically organised into rows and columns
- Note: optional, provides additional information necessary to correctly interpret data in the table

Frequency tables

Table 1

Distribution of film titles by continent with absolute and relative frequencies

Continent	*N* *	%†
Africa	52	1.2
Americas	3,009	67.3
Asia	1,058	23.7
Europe	1,604	35.9
Oceania	153	3.4

Note:

Based on a sample of 4,468 full feature films released since 2010 with 5,000 or more ratings on IMDB.com.

* Column does not add up to total number of titles, as a single title can have multiple continent attributions.

† Percentage of total number of titles (does not add up to 100%).

Grouped frequency tables

Table 2

Distribution of film titles by genres for Europe and the Americas

	Americas		Europe	
Continent	*N* *	%†	*N* *	%†
Drama	1,704	56.6	1,075	67.0
Thriller	1,142	38.0	620	38.7
Comedy	1,069	35.5	439	27.4
Action	779	25.9	363	22.6
Romance	551	18.3	325	20.3
Crime	547	18.2	295	18.4
Adventure	541	18.0	267	16.6
Other	2,934	97.5	1,548	96.5

Note:

Based on a sample of 4,468 full feature films released since 2010 with 5,000 or more ratings on IMDB.com.

* Column does not add up to total number of titles, as a single title can have multiple genre attributions.

† Percentage of total number of titles within given continent (does not add up to 100%)

Summary tables

Table 3

Summary statistics of the measured continuous variables

	N	Min	Max	*M*	*SD*	*SEM*
Average IMDB rating	4,468	1	9.2	6.37	1.02	0.02
Number of IMDB ratings	4,468	5,003	2.03×10^{6}	6.31×10^{4}	1.23×10^{5}	1844.18
Estimated budget (USD)	2,170	18	3.56×10^{8}	3.96×10^{7}	5.09×10^{7}	1.09×10^{6}
Opening week grossing (USD)	2,842	63	3.57×10^{8}	1.14×10^{7}	2.52×10^{7}	4.73×10^{5}
Runtime in minutes	4,387	45	321	109.22	20.33	0.31

Plots

- Sometimes, a picture is worth a thousand words
- Great for communicating information about data that takes a lot of space to explain in writing
- Good graphics should be both clear and packed full of information

Structural elements

- Number: just like tables, all plots should be numbered and the number should be referenced
- Title: should be descriptive
- Axes: clearly labelled, with sensible ticks along them, and units of measurement
- Graphics: clear, well designed, good size
- Legend: if graphical elements are used to distinguish levels of variables, legend must be provided
- Note: optional, provides additional information necessary to correctly interpret the plot

Frequency plots

- Good for exploring distributions of data
- They are intended for you, the analyst, not for the readers of your paper/report
- They can be nice but take up too much space, use up too much ink, and convey too little information

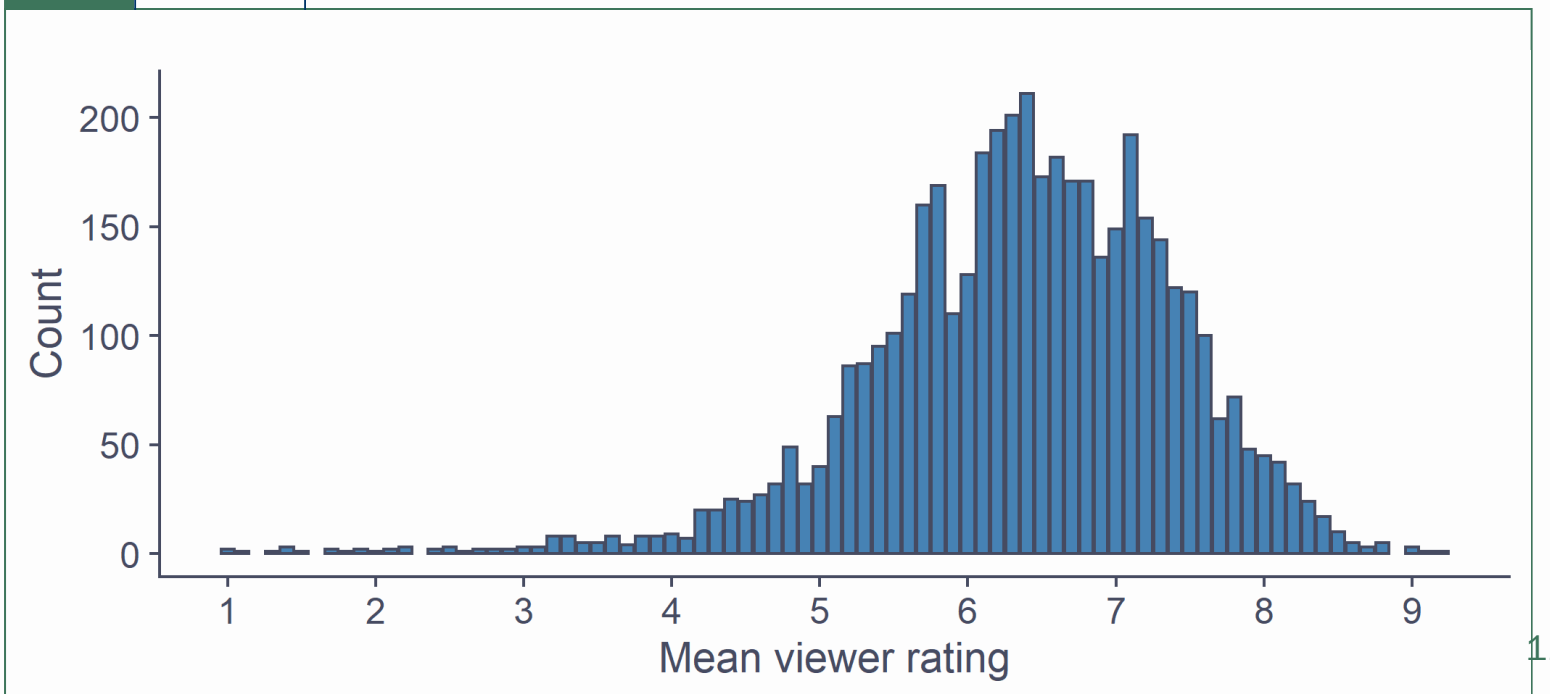
Histogram

- Useful for plotting distributions of continuous variables only (*interval* and *ratio*)
- Data need to be binned; width of individual bins is our decision, explicit or implicit

Figure 1

A histogram of the distribution of average viewer ratings in the sample

Plot | R code



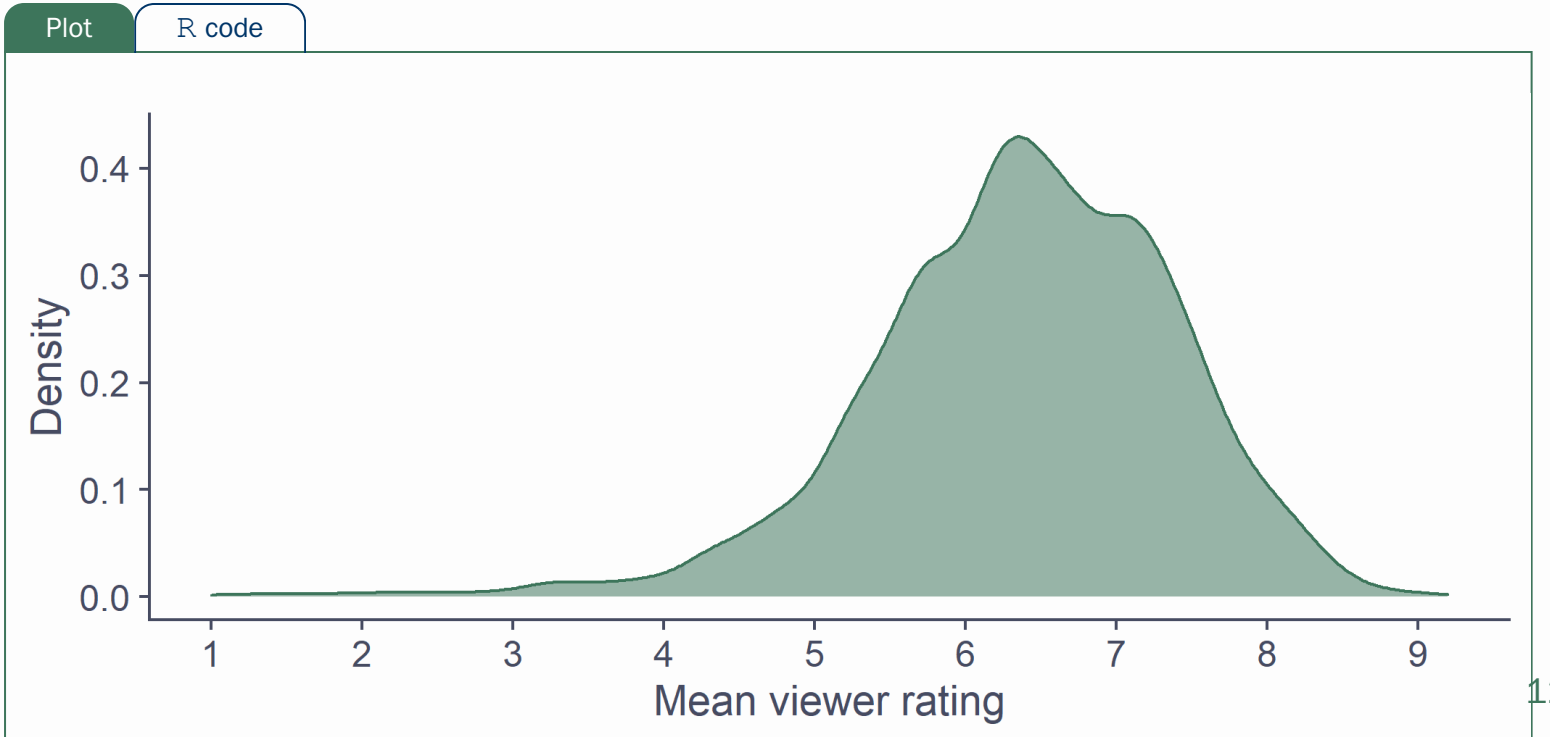
Density plots

- Also for continuous variables only
- Simulate what a histogram with infinitely narrow bins would look like

Figure 2

Density plots provide a smoother representation of the distribution of a variable

Plot | R code



Bar charts

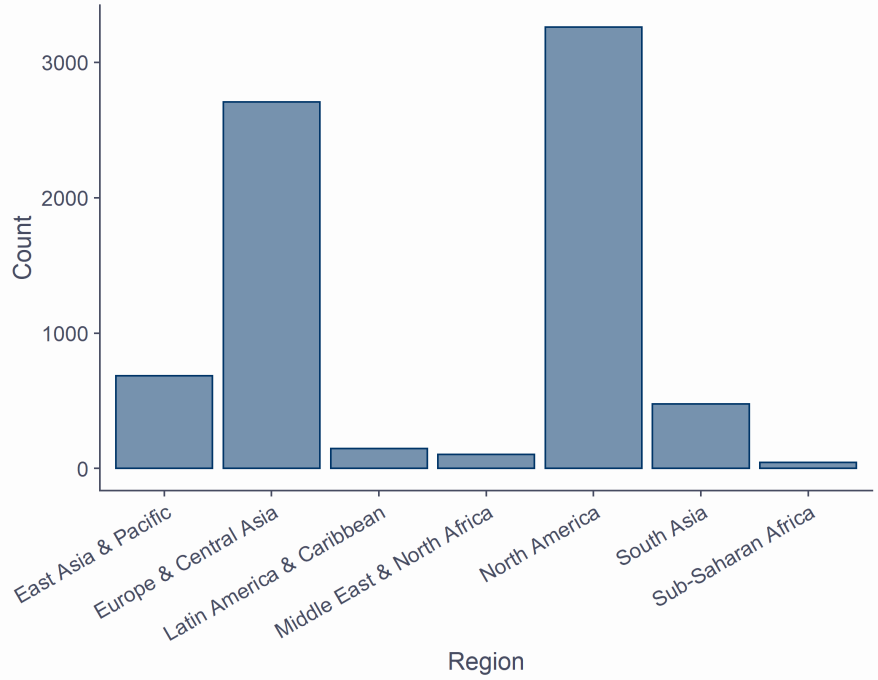
- Visualise distributions of categorical data (*nominal* and *ordinal*)
- They are still used for summarising data
- Even the APA website shows them in their list of sample figures
- You shouldn't do it though - it's a waste of space!

Simple bar chart

Figure 3

A bar chart showing the number of films produced within a region

Plot | R code



Note. Categorisation of countries into regions is based on World Bank Development Indicators.

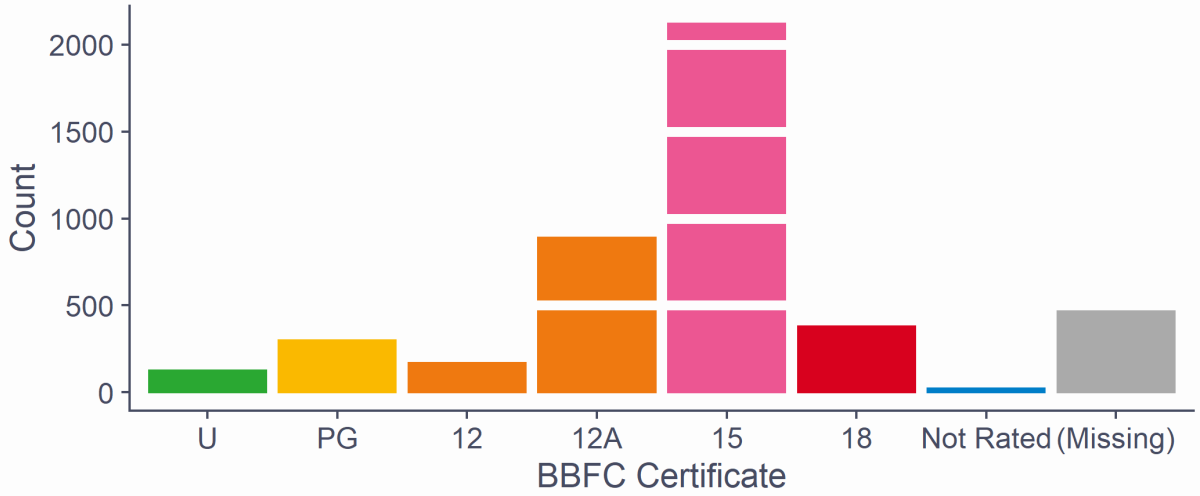
Bar charts

- Colour isn't necessary here but it's at least meaningful (British Board of Film Classification labels)

Figure 4

Bar charts can also show distributions of ordinal variables

Plot | R code



Note. U: universally suitable; PG: parental guidance recommended; 12: suitable for children from 12 years of age; 12A: suitable for children from 12 years of age accompanied by adult; 15: suitable for 15-year-olds and older; 18: only suitable for adults.

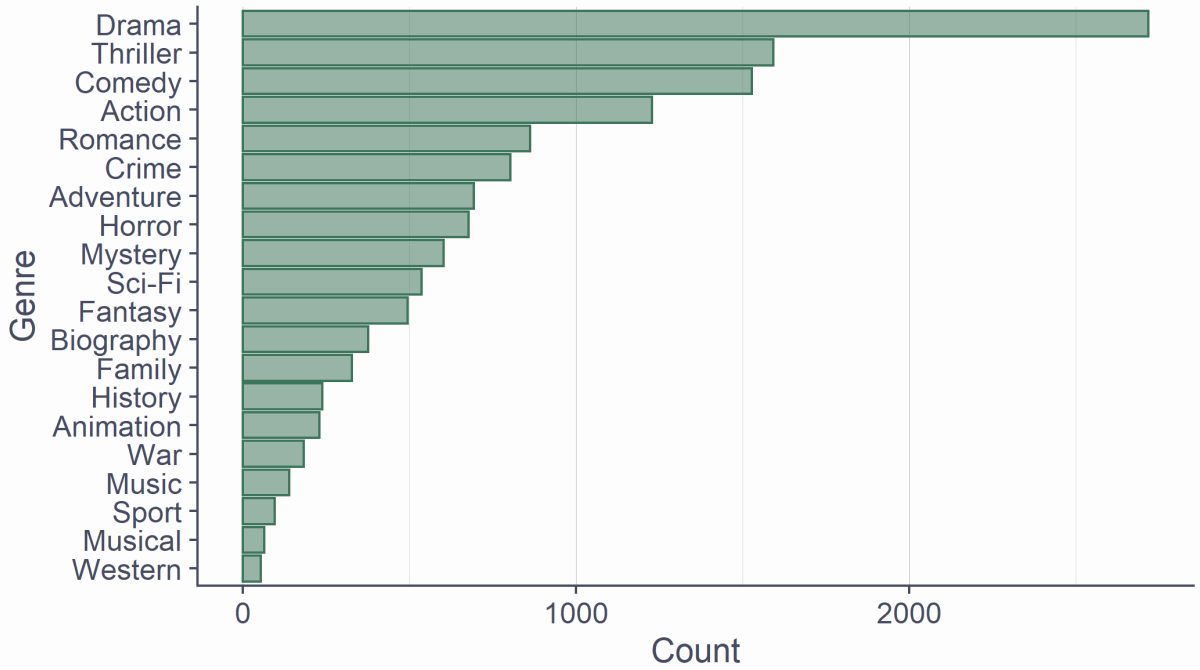
Bar charts

- Sometimes it makes more sense to flip them horizontally
- Grid lines can help comparing things in all kinds of plots

Figure 5

An example of a horizontal bar chart with grid lines

Plot | R code



Summary plots

- Unlike frequency plots, their primary aim is to summarise the data in term of key statistics
- They are often used to compare variables across groups
- *Some of them* can be used to gauge differences between groups and relationships between variables
- They are *not a substitute for data analysis!*

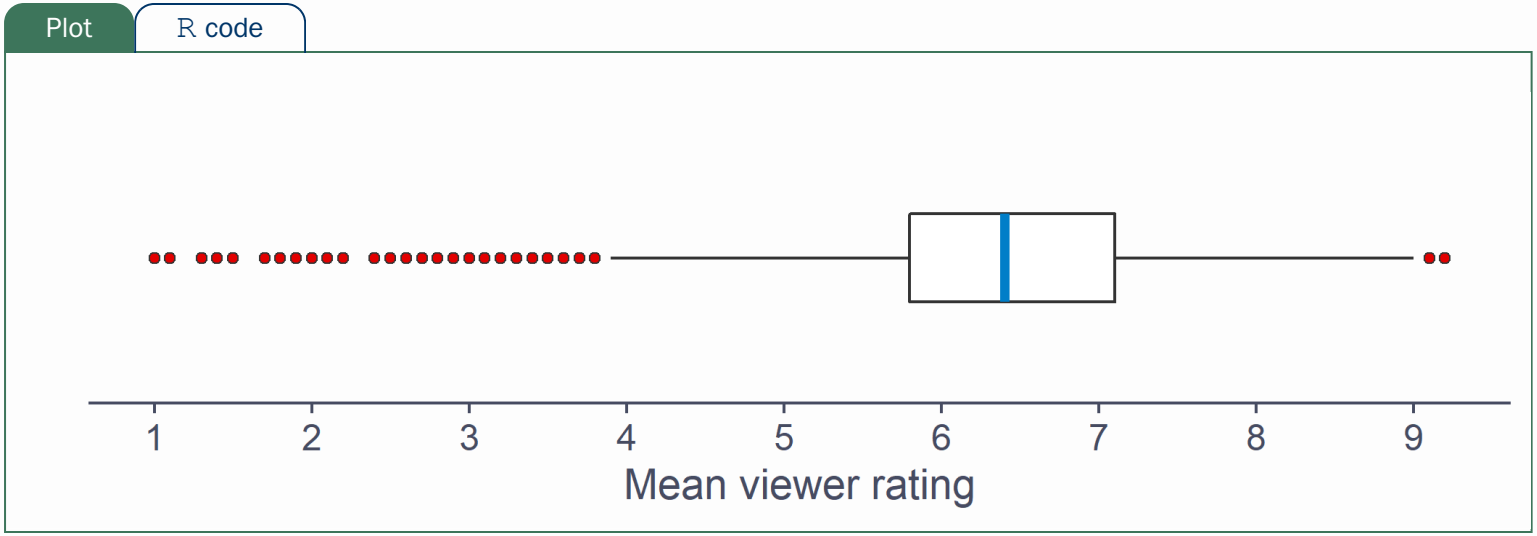
Box plot

- AKA box-and-whiskers plots

Figure 6

A box plot gives a more detailed info about a variable's distribution and basic descriptive statistics

Plot | R code



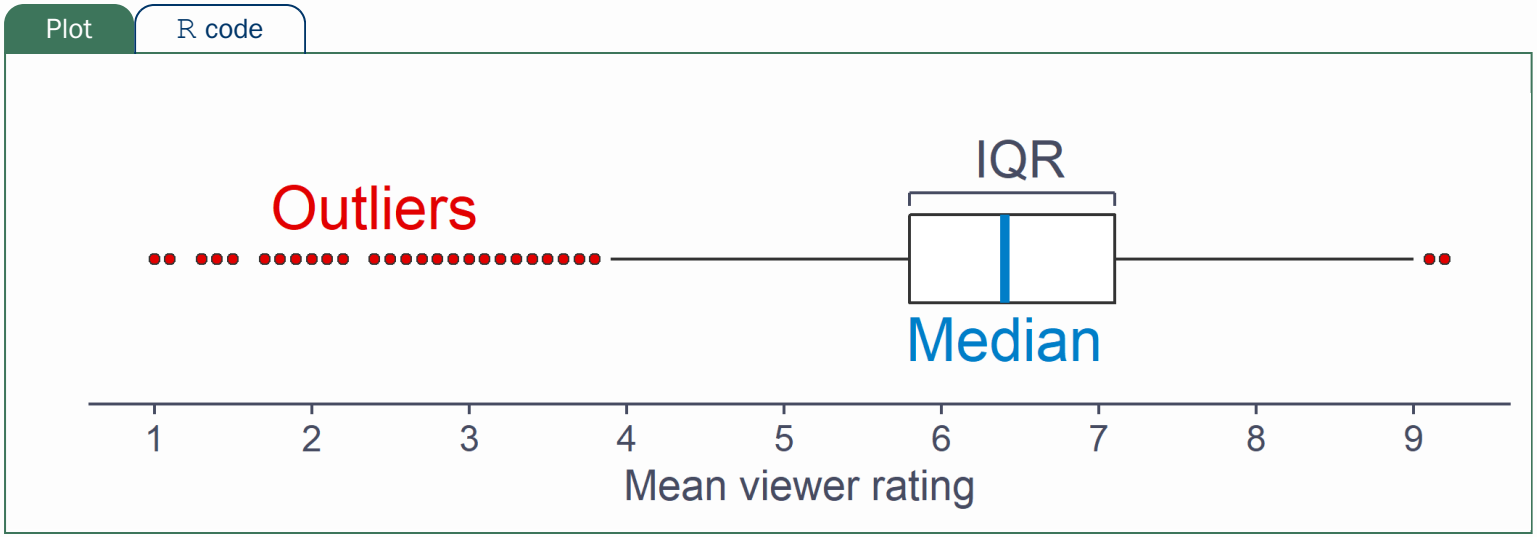
Box plot

- AKA box-and-whiskers plots

Figure 6

A box plot gives a more detailed info about a variable's distribution and basic descriptive statistics

Plot | R code

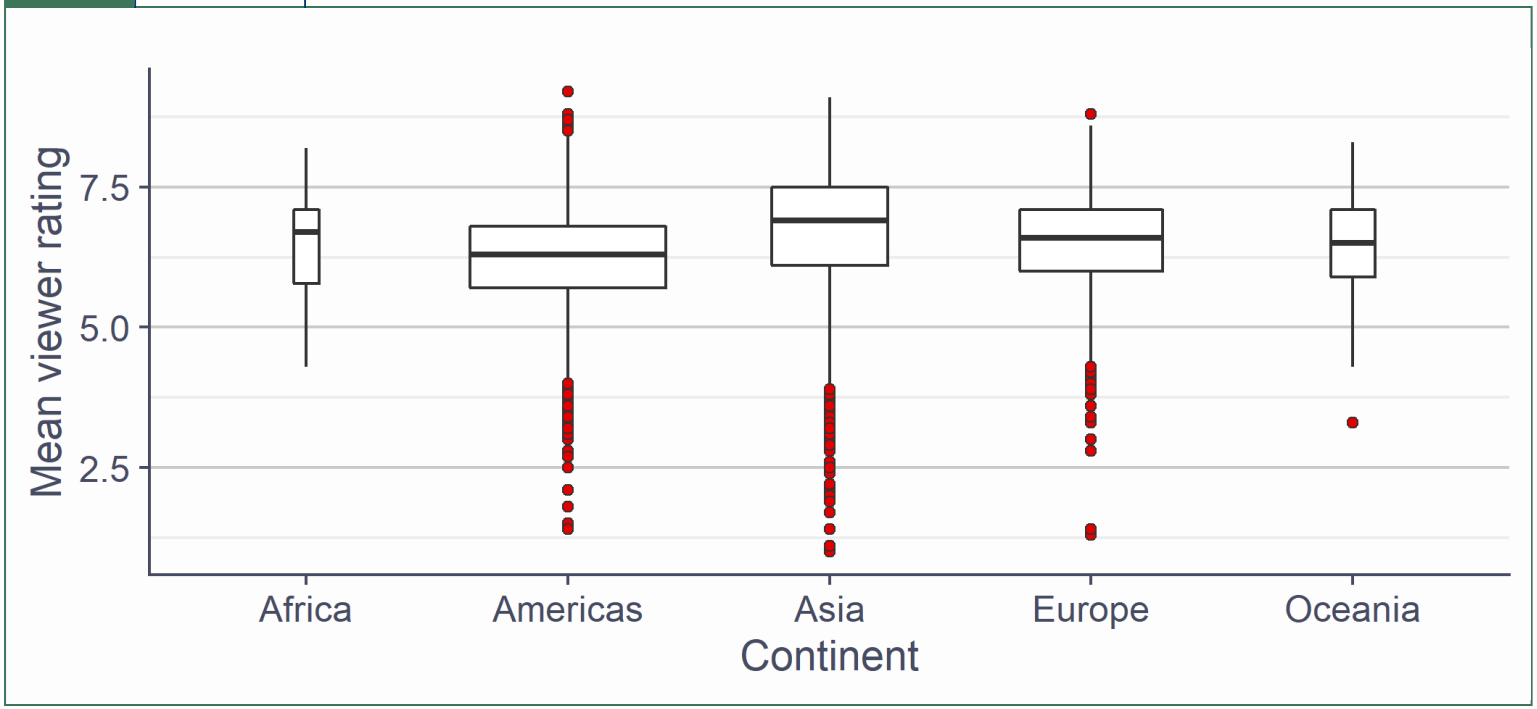


Grouped box plot

Figure 7

Boxplots are useful to compare distributions of a variable between groups

Plot | R code



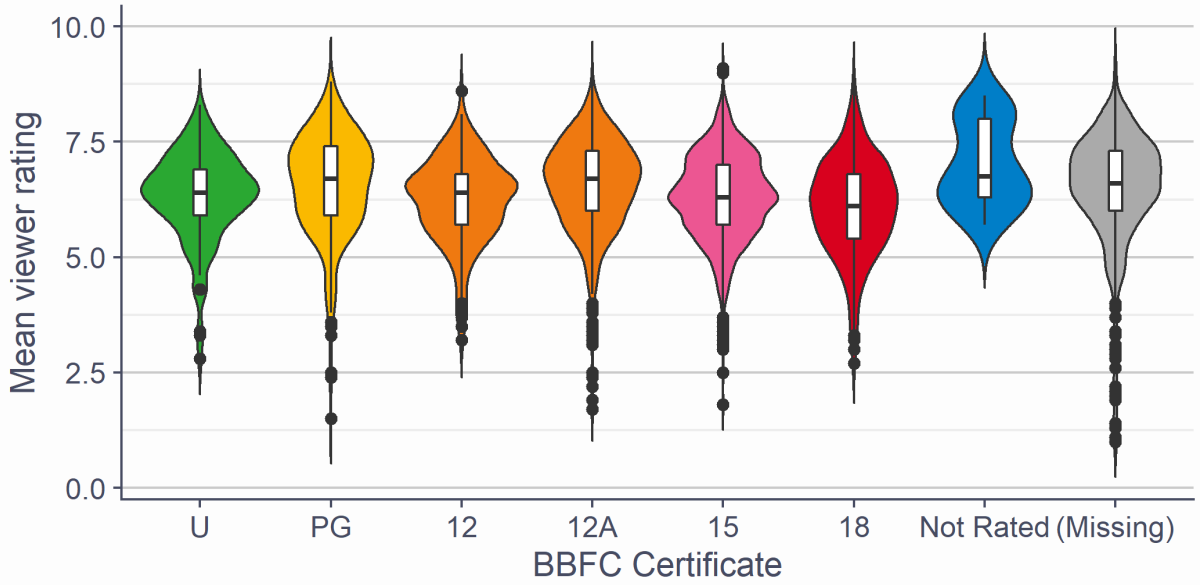
Note. Width of boxes proportional to number of observation per category of the Continent variable

Violin plot

Figure 8

Violin plots with inserted box plots showing mean viewer rating by different levels of BBFC certificate

Plot | R code



Note. U: universally suitable; PG: parental guidance recommended; 12: suitable for children from 12 years of age; 12A: suitable for children from 12 years of age accompanied by adult; 15: suitable for 15-year-olds and older; 18: only suitable for adults.

Errorbar plot

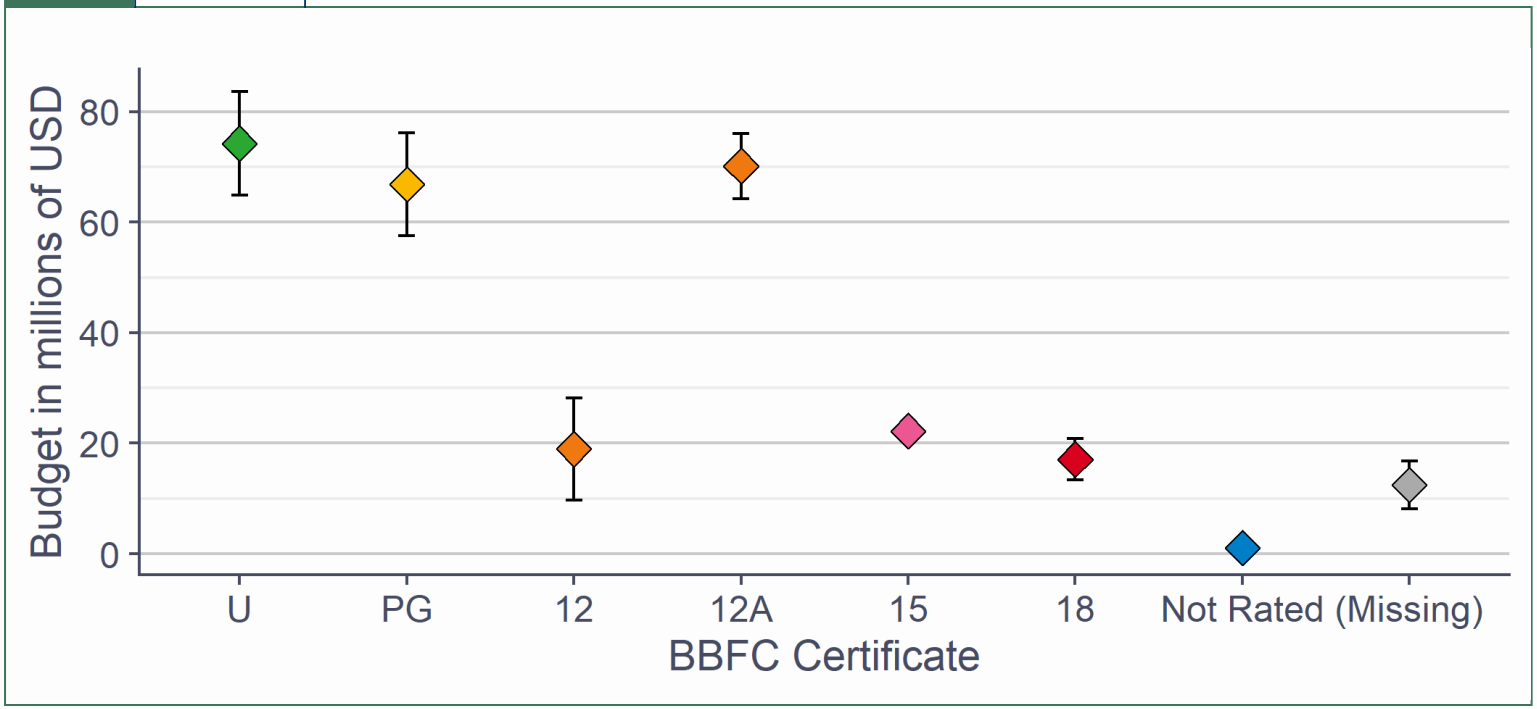
- Errorbar plots are great for showing means and spread/inferential statistics
- Some of them can be used to gauge *statistical* differences between groups
- Error bars can show several things (*e.g,* standard deviations, standard errors, their multiples)
- Plot should clearly indicate what they represent
- Pay attention to what the error bars mean
- Interpretation of plots changes based on what the bars show!

Errorbar plot

Figure 9

Errorbar plot showing mean estimated budget of movies by level of BBFC certificate

Plot | R code



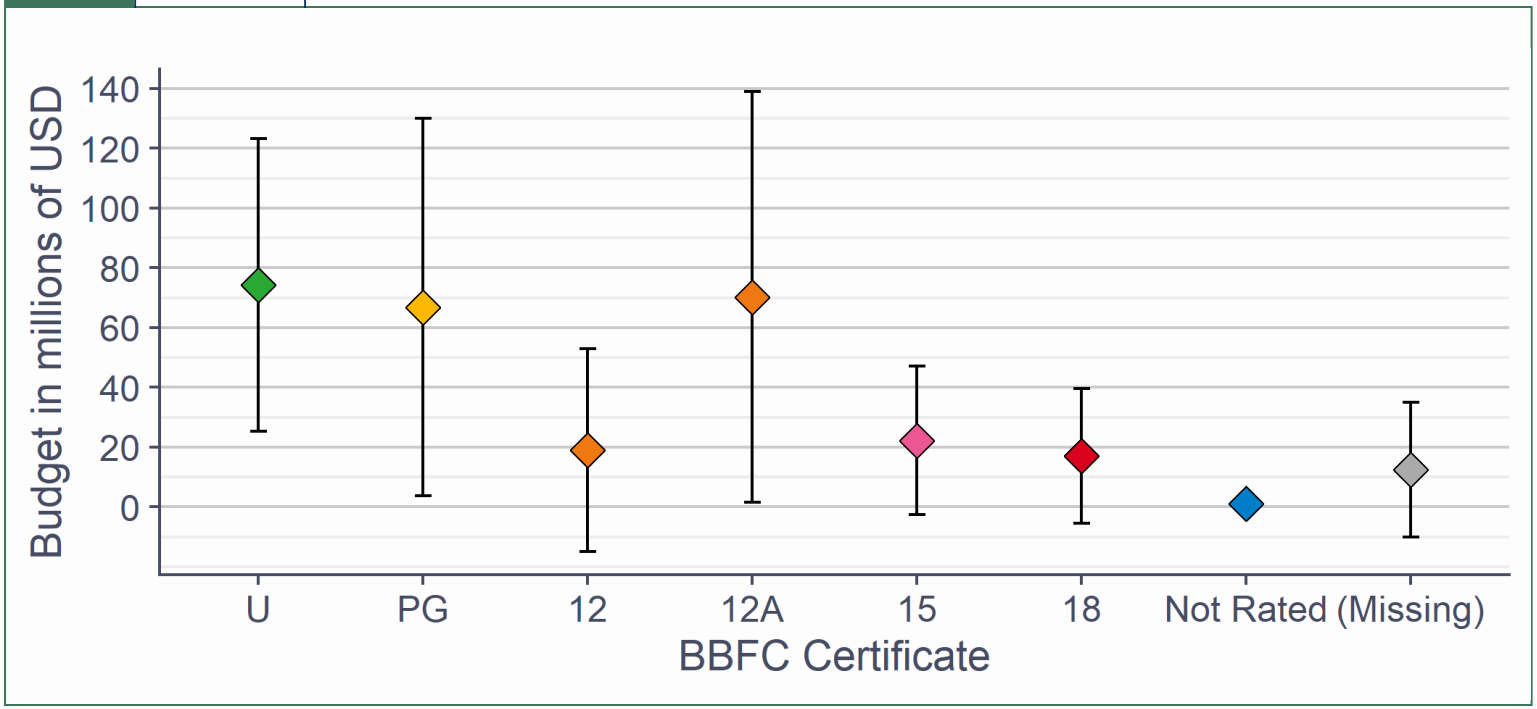
Note. Error bars represent ±2 *SE* from the mean.

Errorbar plot

Figure 10

Errorbar plot showing mean estimated budget of movies by level of BBFC certificate

Plot | R code



Note. Error bars represent ±1 *SD* from the mean.

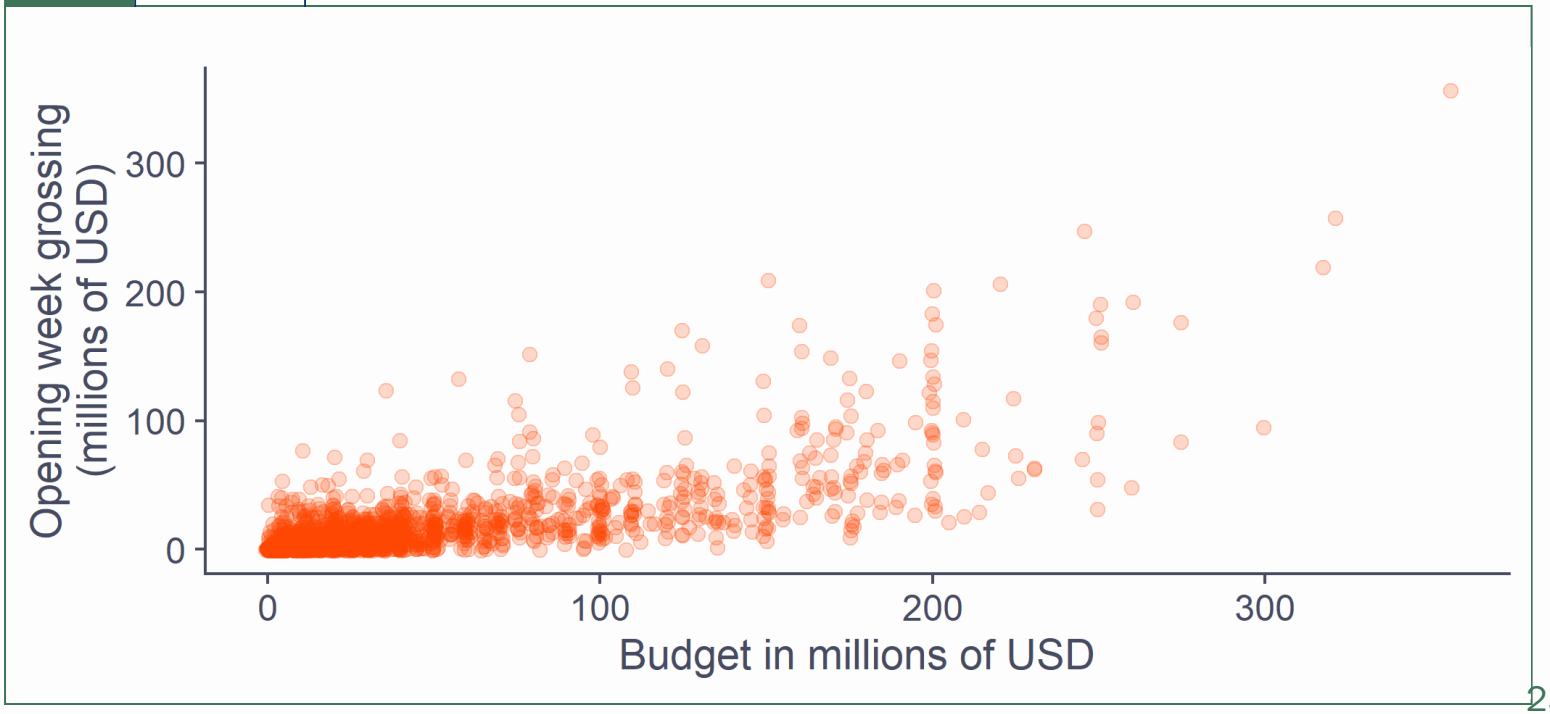
Scatter plot

- Best way to show relationships between two continuous variables.

Figure 11

The relationship between a film's budget and its opening week box office performance

Plot | R code



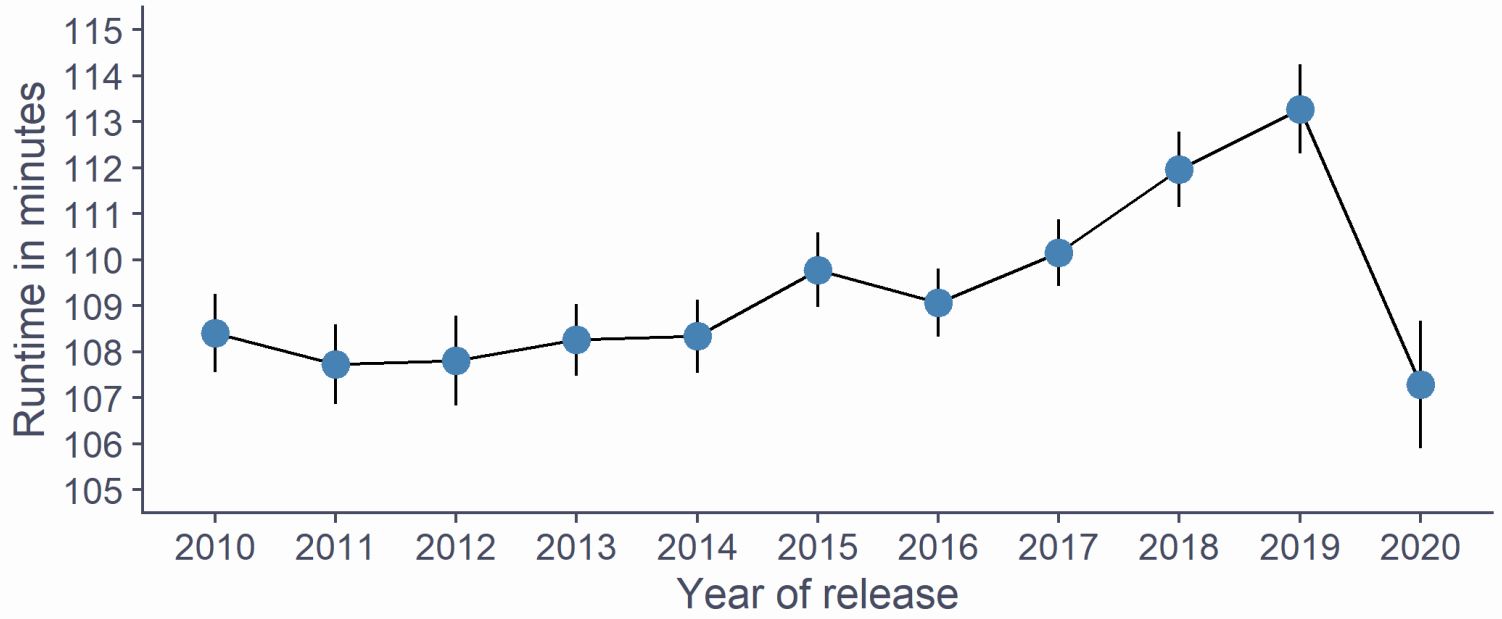
Line plot

- Great for highlighting repeated measures and within-subject structure

Figure 12

Average film runtime over time in the second decade of the 21st century

Plot | R code



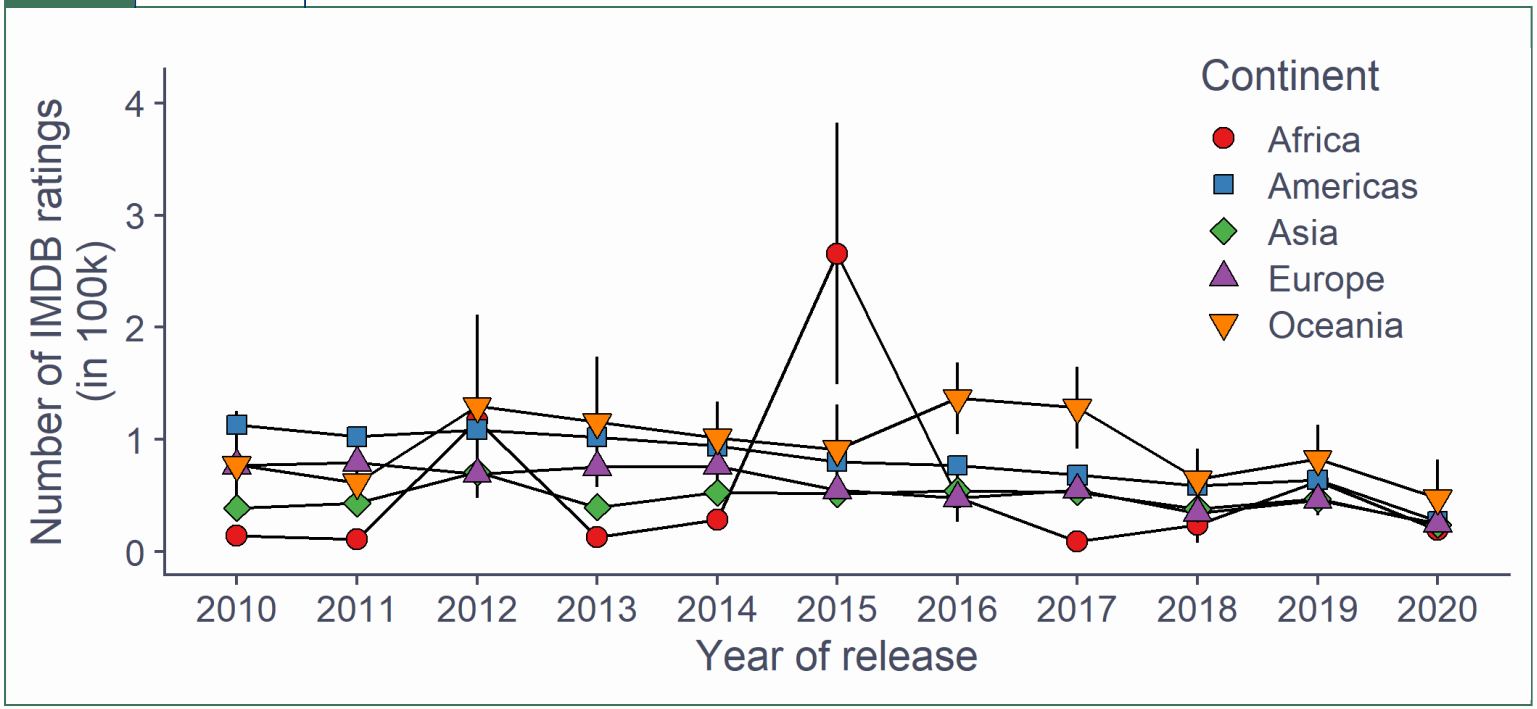
Note. Error bars represent +1 *SE* from the mean

Lot of info!

Figure 13

Mean number of IMDB viewer rating of films by continent over time

Plot | R code



Note. Error bars represent ±1 *SE* from the mean.

Sometimes less is more!

- Plots and tables should *supplement* body text, not repeat what's already there
- There's no need to show the same thing in a table AND in a plot at the same time!
- Always make sure font size and size of graphics are big enough to be easily legible
- For more detailed APA guidelines, see the APA website
    - They also provide sample tables and figures. (But remember, not bar charts for summary stats! Not ever!)


That's all Folks!