Best Practices Around Data Visualisation

Dr. Niccole M. Pamphilis

10 June 2024

Overview

What will we talk about:

- What data visualisations are
- Why standards matter
- How and when to use tables
- How and when to use graphs
- Warnings around creating data visualisation

What we will not be talking about:

- Code for graphs
- How to generate data visualisations

Admin Notes

Contact details:

email: Niccole.Pamphilis@glasgow.ac.uk

Email me for a copy of the slides

Visualizing Data

Why do we have data visualisations?

Data sets can contain a lot of information

While raw data in small amounts can be visually taken in:

- How quickly can you locate a typical case?
- How quickly can you determine the variation/spread of the data?

Data visualisationss allow for quick, succinct presentation of information

What are Data Visualisations

What is your best guess at what a data visualisation is?

A visual representation of data

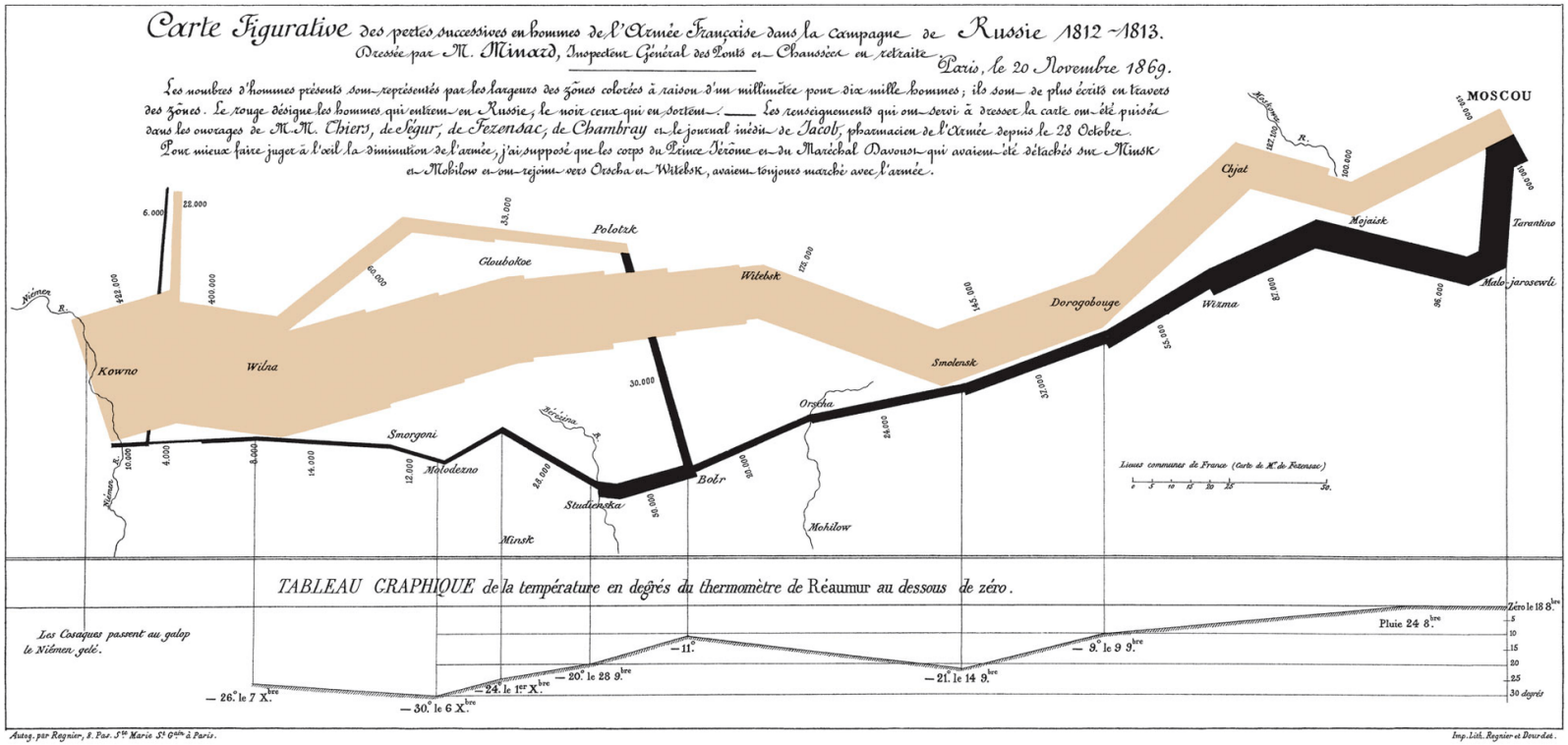
Does the amount of data matter?

Does the type of data matter?

What are some examples of data visualisations we use?

- Tables
- Graphs
- Infographics

Examples of Historical Data Visualisations



Source: Charles Minards Flow Map of Napoleons Russian Campaign of 1812

Examples of Historical Data Visualisations

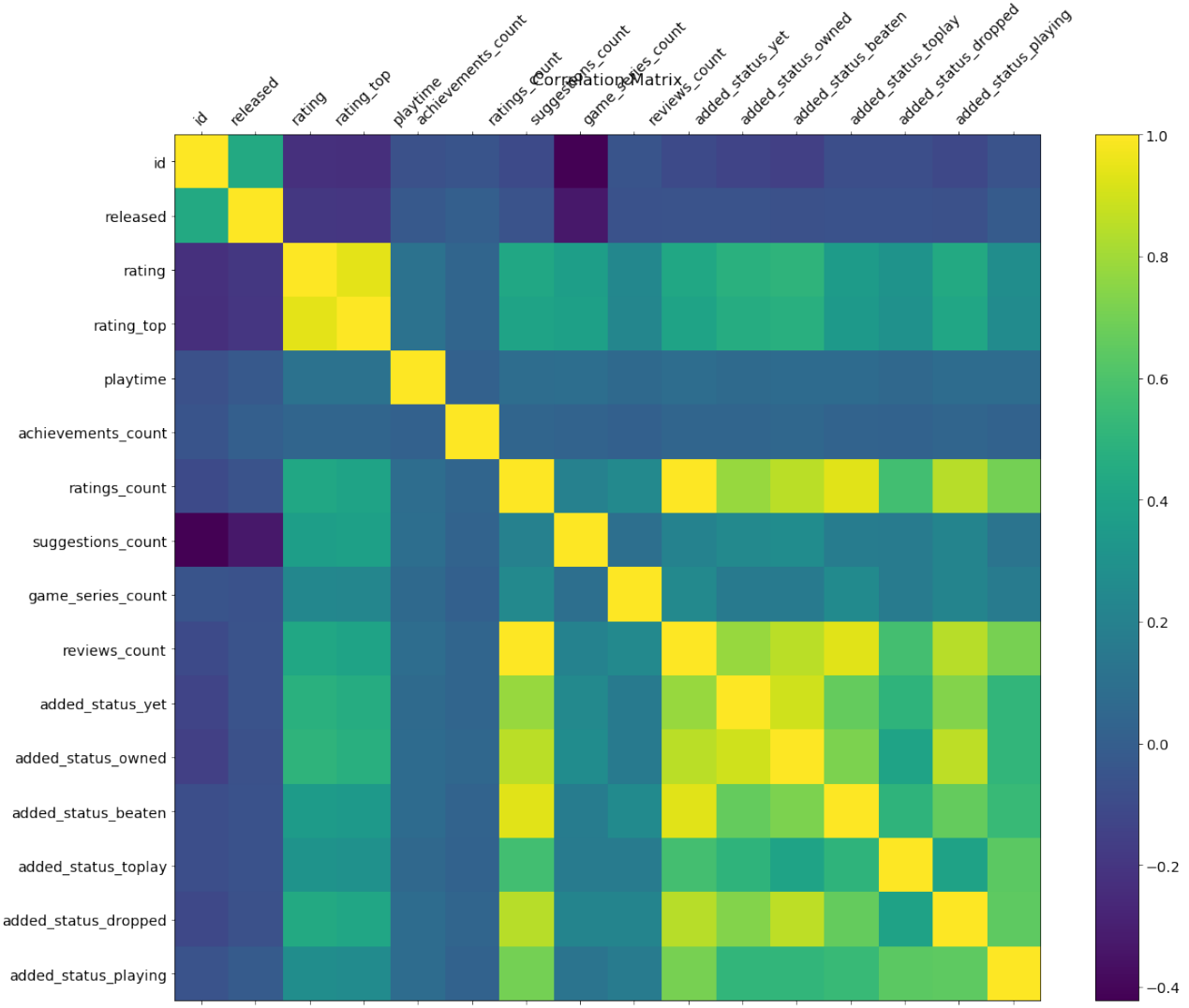
Source: Dr. John Snow: p. 97-120 of the "Report on the cholera outbreak in the Parish of St. James, Westminster, during the autumn of 1854"

Why Standards

Why do we need standards for data visualisations

- Eyes lie
- Over complications
- Wrong visualisation for data type
- Things may look cutting-edge, but the brain cannot always keep up

Is this Good?



Source:https://www.kaggle.com/code/joshuaisanan/predicting-the-number-of-reviews-of-a-video-game/notebook

Discussion

What did you like about the data visualisation?

What did you dislike about the data visualisation?

Was this a realistic example?

What is a Table

What are the features of a data table?

- Rows and columns
- Variables names
- Statistics
- Numbers

Anything else necessary?

Is it easy to make a good data table?

When are Tables Useful?

Small amounts of information

- Summary statistics
- Model results

Ideally, information that fits on one page

What does a Table Need?

- Clear main title
  - So reader knows what they are looking at
- Clear variable descriptors
  - Not variable names directly from data set
  - So readers know what is going on
- Clear statistical terms
  - Not everyone technical jargon
- Source of the data used
  - Transparency
  - Replication
- Note/legend if relevant
- Understandable outside of context

What to Avoid and Why

Not always easy or straight forward process data, particularly in visual presentations

Avoid small font size do not make the reader work hard

Avoid the use of lines

- Vertical
- Horizontal

Do not assume people understand symbols

- What does $\star$ mean

Do not assume people read the text around the table

Variable names used in data sets

- Do not require readers to have access to code books or guess what variables capture

Table Example

Government Spending on Social Protection

Country	Mean	Standard Deviation	Minimum	Maximum
Austria	61.13	0.86	59.84	62.65
Belgium	59.27	1.55	57.19	62.27
Canada	41.62	1.25	38.71	43.04
Czech Republic	52.31	0.98	50.00	54.05
Denmark	62.45	1.75	60.03	65.54
Estonia	49.05	1.79	47.48	55.09
Finland	63.72	2.11	61.25	68.73
France	61.89	0.69	60.90	63.33
Germany	65.72	0.92	64.22	67.23
Greece	63.57	3.99	59.27	71.53
Hungary	60.50	2.01	57.29	64.14
Iceland	35.75	2.76	32.62	40.86
Italy	50.92	3.69	45.21	55.74

Is this table fit for purpose?

What needs to be improved?

Table Example, cont.

Government Spending on Social Protection

Country	Mean	Standard Deviation	Minimum	Maximum
Austria	61.13	0.86	59.84	62.65
Belgium	59.27	1.55	57.19	62.27
Canada	41.62	1.25	38.71	43.04
Czech Republic	52.31	0.98	50.00	54.05
Denmark	62.45	1.75	60.03	65.54
Estonia	49.05	1.79	47.48	55.09
Finland	63.72	2.11	61.25	68.73
France	61.89	0.69	60.90	63.33
Germany	65.72	0.92	64.22	67.23
Greece	63.57	3.99	59.27	71.53
Hungary	60.50	2.01	57.29	64.14
Iceland	35.75	2.76	32.62	40.86
Italy	50.92	3.69	45.21	55.74

Improve the Table, Task

Proportion of Total Government Spending on Social Protection

Country	Mean	Standard Deviation	Minimum	Maximum
Austria	61.13	0.86	59.84	62.65
Belgium	59.27	1.55	57.19	62.27
Canada	41.62	1.25	38.71	43.04
Czech Republic	52.31	0.98	50.00	54.05
Denmark	62.45	1.75	60.03	65.54
Estonia	49.05	1.79	47.48	55.09
Finland	63.72	2.11	61.25	68.73
France	61.89	0.69	60.90	63.33
Germany	65.72	0.92	64.22	67.23
Greece	63.57	3.99	59.27	71.53
Hungary	60.50	2.01	57.29	64.14
Iceland	35.75	2.76	32.62	40.86
Italy	50.92	3.69	45.21	55.74

Note:Data obtained from the OECD for the years 1990-2011 for most countries.

Why do We Use Graphs

Large quantities of information

Trends and patterns across space or time

Highlight particular points

Types of Graphs

Variety of graphs depending on data and need

Discrete versus continuous data

One variable or more

Patterns/trends/relationships

Types of Graphs, cont.

Not an exhaustive list

- Line graphs
- Bar charts and histograms
- Boxplots
- Scatter plots
- Heat maps

Constant across Types

- Clear titles
- Clear labeled axes
- Clear visible plotting symbols
  - Default options not always the best
- Proportions of axes
- Avoid truncation
  - Obscures story of the data
  - Can make comparison harder
- Clear notes
- Understandable outside of context

Line Graphs, What

Line Chart

- A means of showing changes in values across time
- Across the x-axis we place the measure of time
- Across the y-axis we place the scale for frequency/proportions of variable of interest

Useful for looking at trends across time

Useful for visual comparison across observations/groups/categories across time

Line Graphs, When

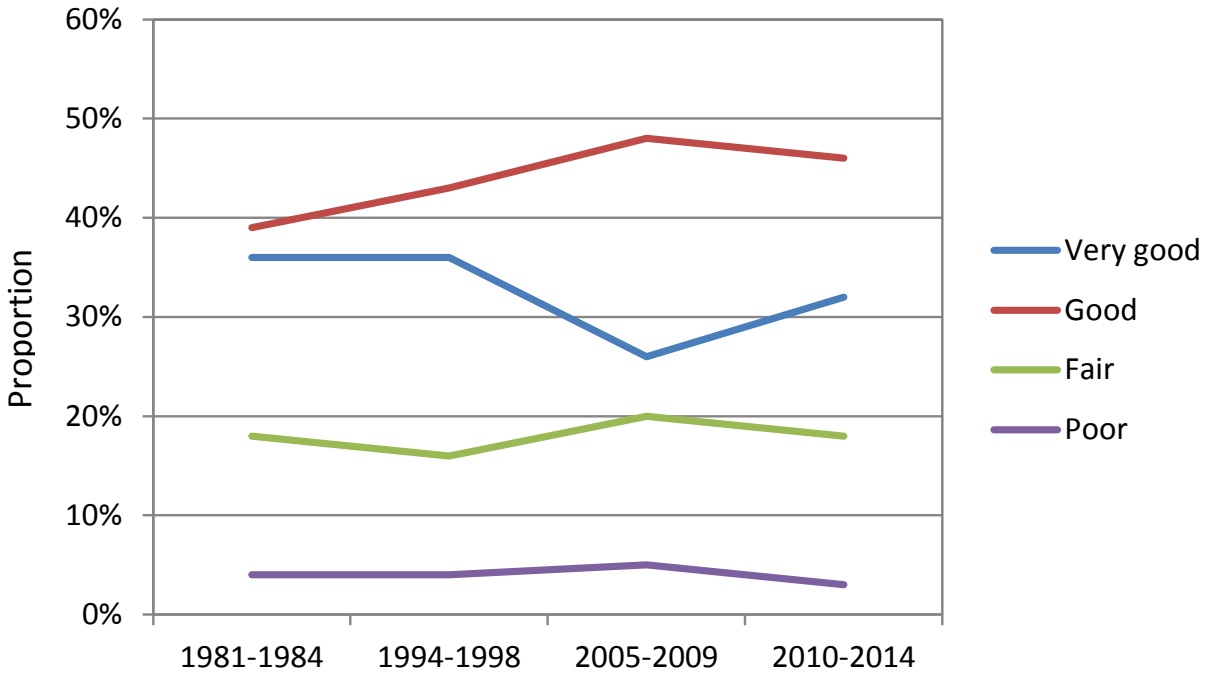
Ideally, multiple consistently spaced time points

Interest in observed variation of outcome

Line Chart

All in all, how would you describe your state of health these days? Would you say it is...

Australia



Source: World Values Survey

What are some issues with the above graph that you can see

Warning about Line Graphs

What do you normally do between the data points on a line chart?

What does connecting the dots imply to the reader?

Is this assumption frequently tested?

Bar Charts/Histograms, What

- Bar chart
  - Discrete Variables with +2 Categories
  - Denotes frequency at which observations fall into a category
  - Can show frequency or percentage of observations per category
  - Not useful for changes over time
- Histogram
  - Continuous Variables (Interval/Ratio Level)
  - Distribution of Data
    - Are observations clustered together?
    - Are observations spread out over the range of possible values?
    - Are the values normally distributed?

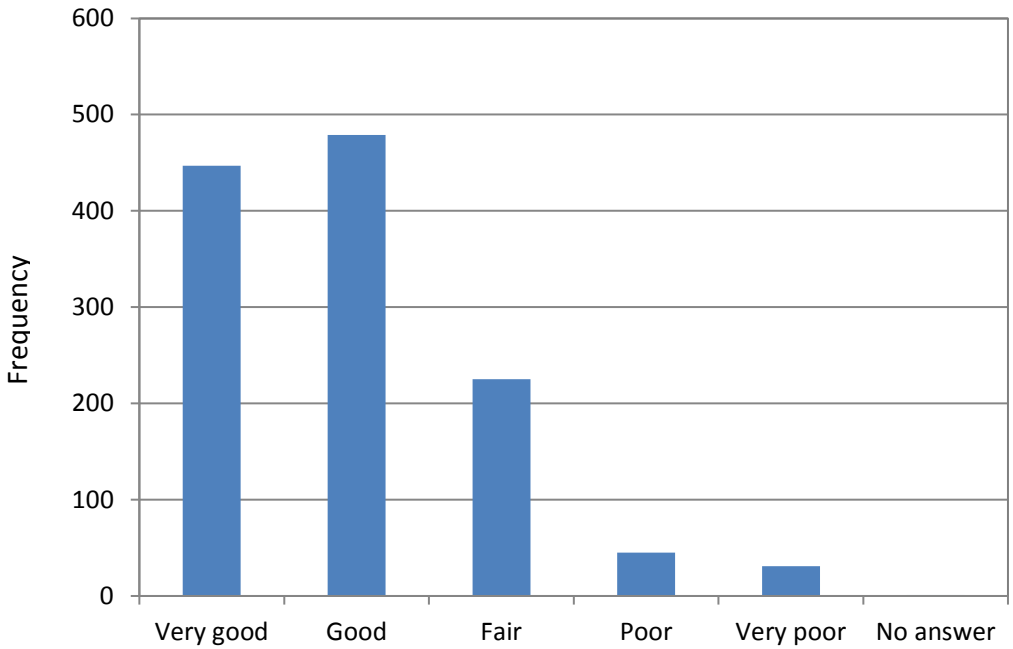
Bar Charts/Histograms, When

Interest is in the distribution of values across categories/ranges

Can include second variable to compare distributions across subsets

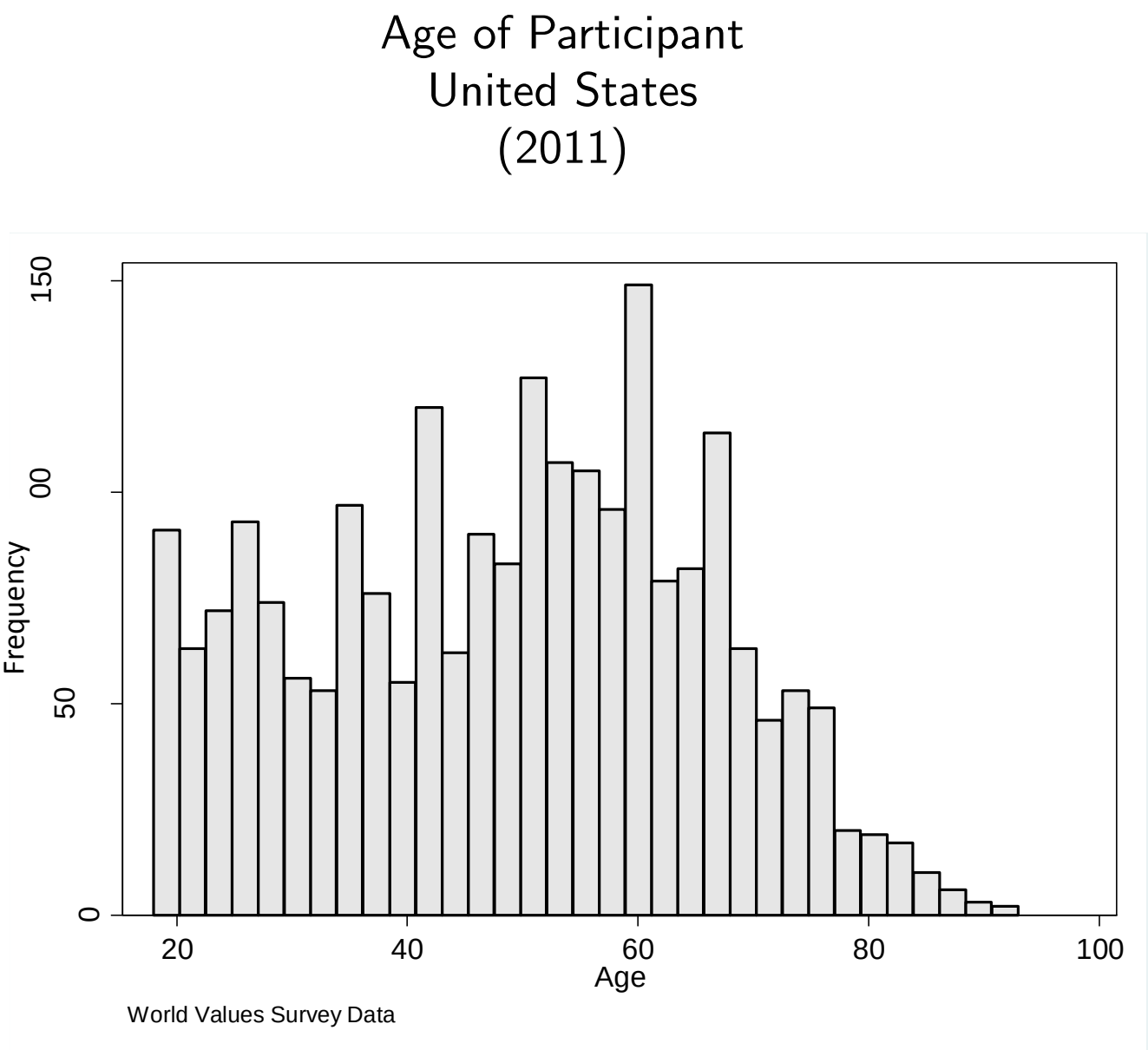
Bar Char, Example

All in all, how would you describe your state of health these days? Would you say it is...
Australia
(1981)



Source: World Values Survey

Histogram, Example



Boxplots,What

Continuous data

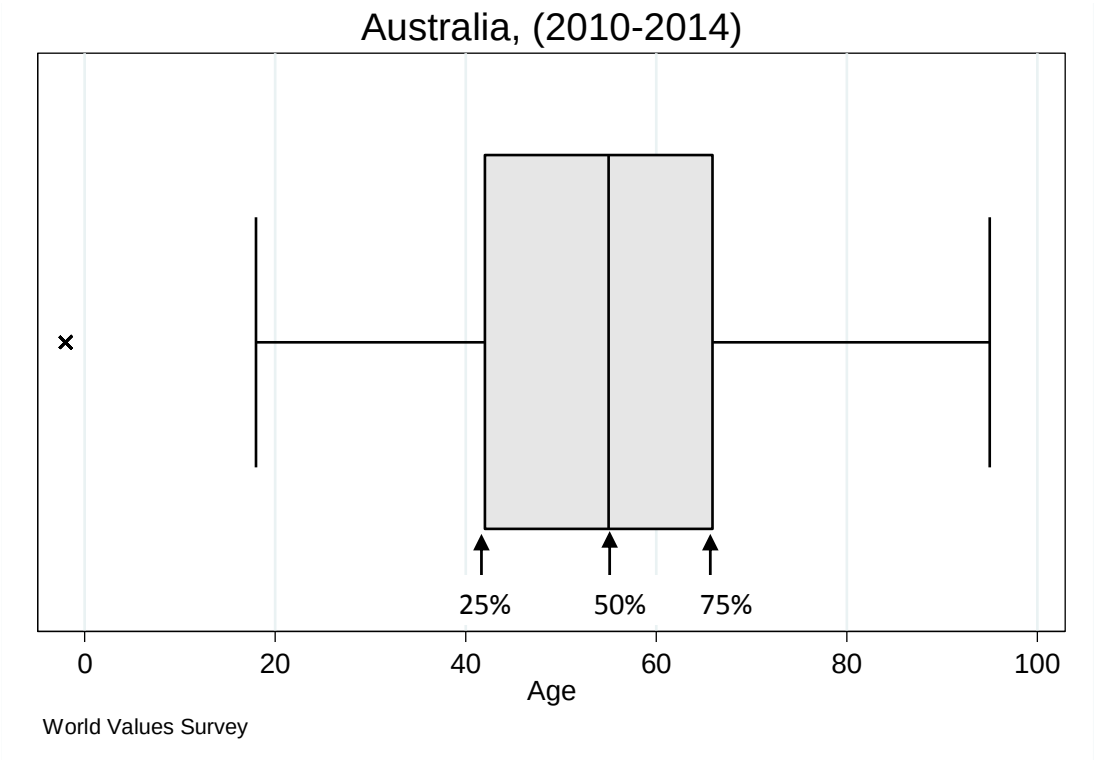
Interest is in distributions of data

Can be done across subsets of the data

Box Plots

Shows the Distribution of a Variable, using Quartiles

The values that denote the 25%, 50% (median) and 75% data values



Good for locating outliers

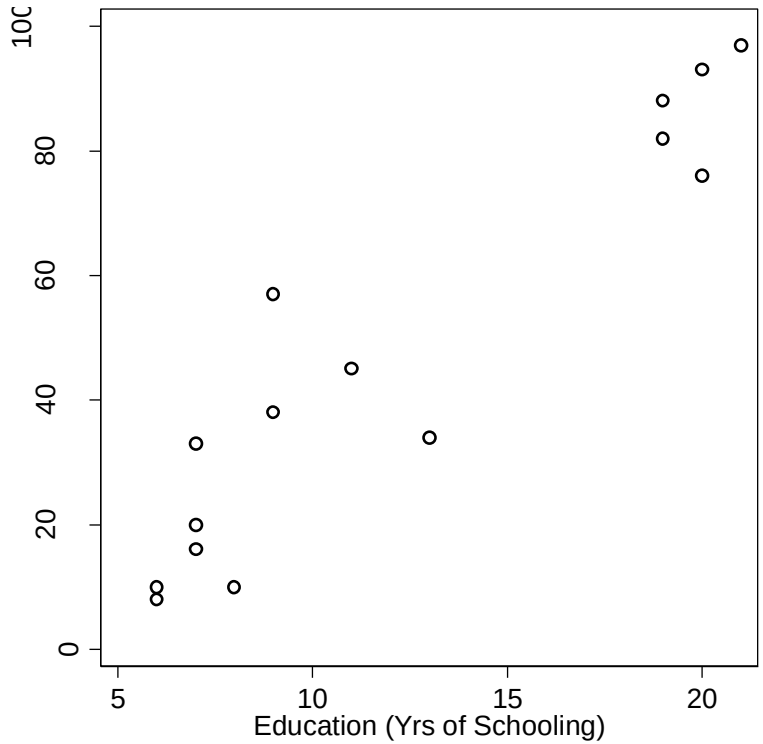
Scatterplots, What

Bi-variate relationships/patterns

Mapping model results onto data

Diagnostic/ initial data examination

Scatter Plot, Example



What's missing from the graph?

Other Data Visualisation Types

Many varieties of data visualisations exist

- Heat maps
- Correlation grids
- Violin plots

But be careful in selection and creation

Warning on Data Visualisations

There are some basic practices that everyone can engage with

Default graphic selection is meant to optimize image not comprehension

Check and adjust:

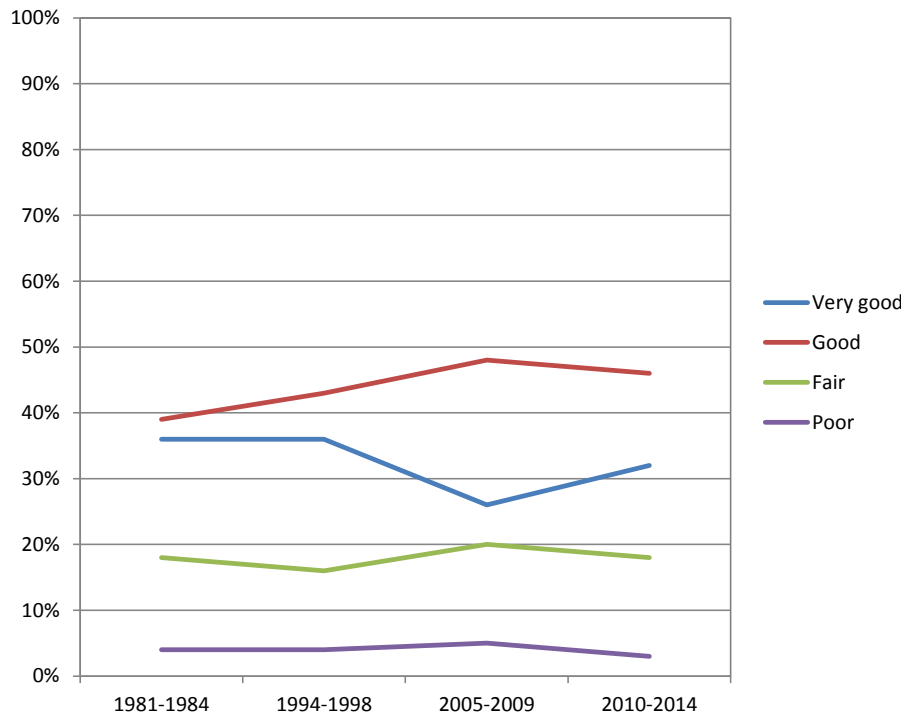
- Scale
- Truncation
- Labels and titles
- Plotting symbols
- Color

Scale

The size of the x-axis relative to the y-axis

Ideally, this should be 1 to 1

All in all, how would you describe your state of health these days? Would you say it is...
Australia

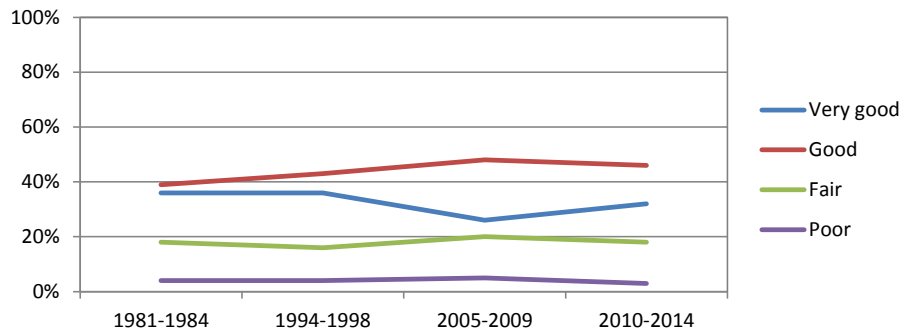


Source: World Values Survey

Scale, cont.

But, changing the ratio, can change the story

All in all, how would you describe your state of health these days? Would you say it is...
Australia

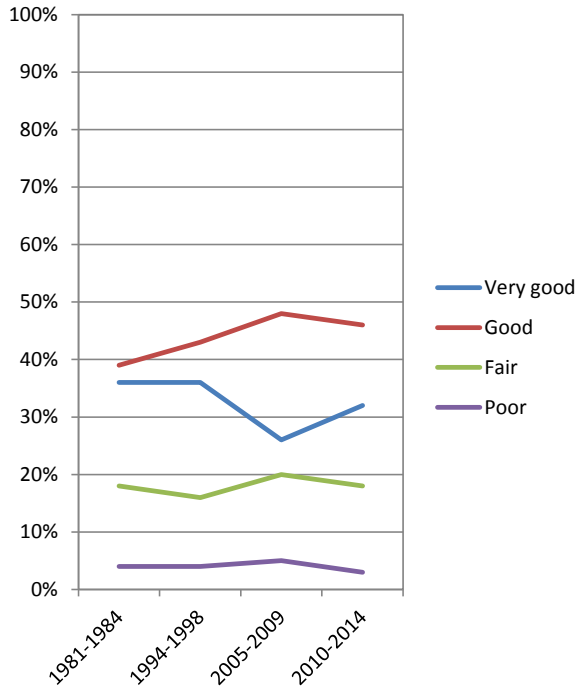


Source: World Values Survey

Scale, cont.

All in all, how would you describe your state of health these days? Would you say it is...

Australia



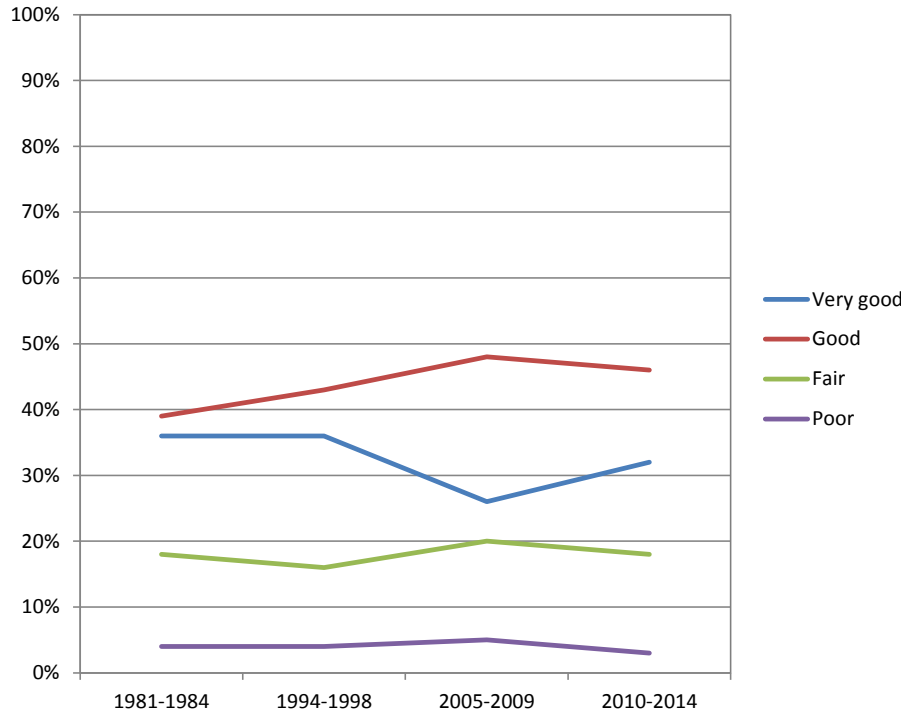
Source: World Values Survey

Truncated Axes

Omitting the full range of possible values can skew the story

All in all, how would you describe your state of health these days? Would you say it is...

Australia

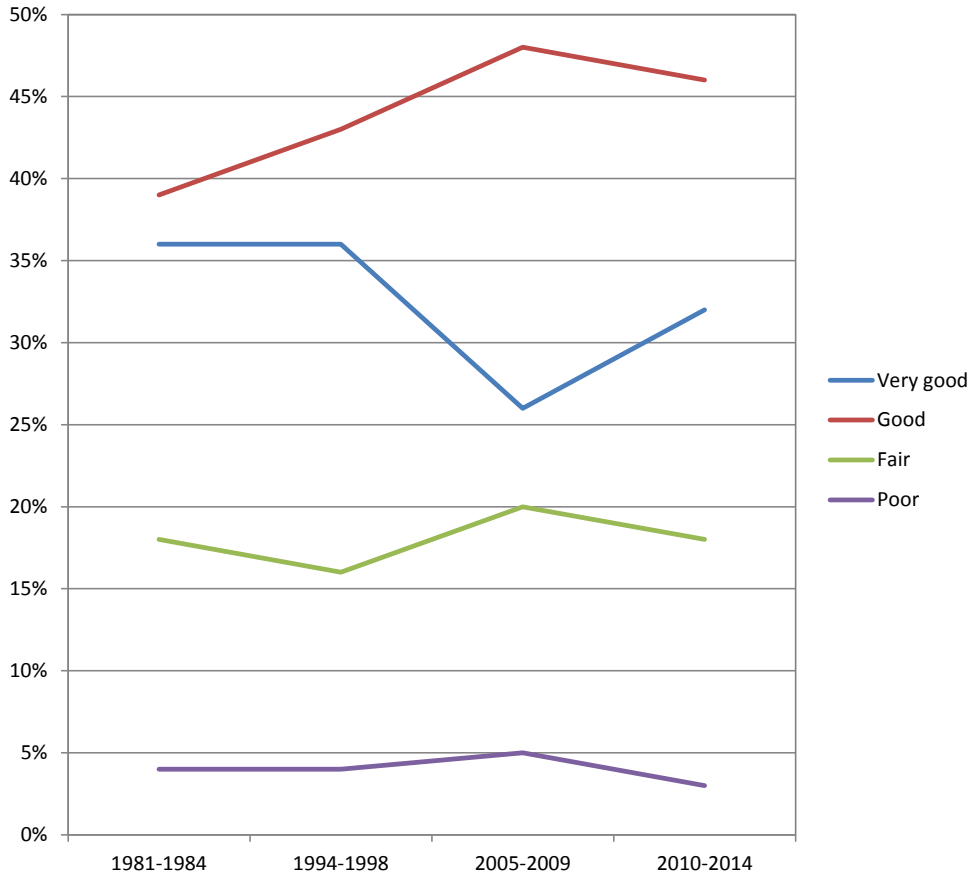


Source: World Values Survey

Truncation, cont.

All in all, how would you describe your state of health these days? Would you say it is...

Australia



Source: World Values Survey

Histograms, Bin Size

Bin size: The range of values within a bar

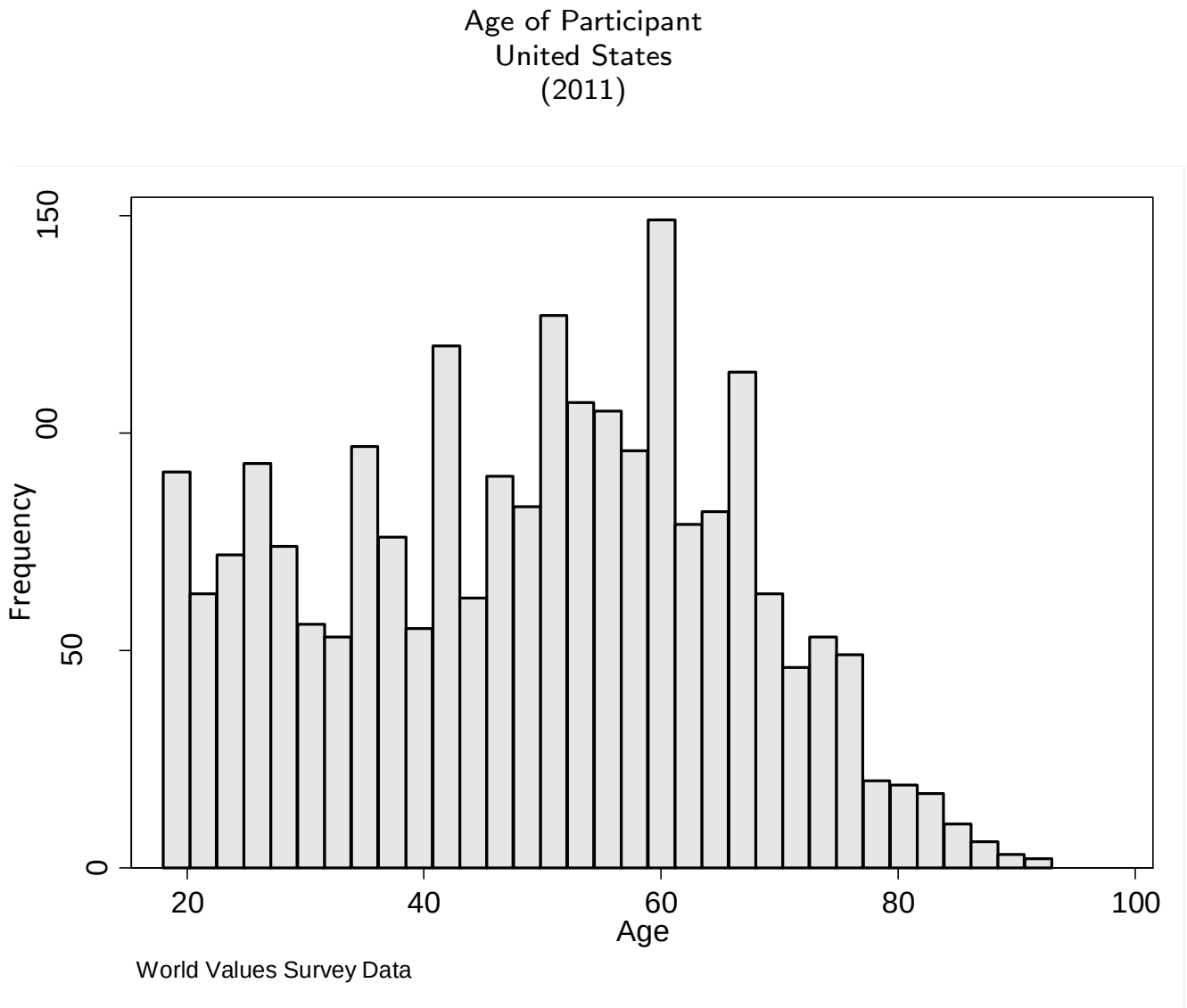
Changing the size will alter the smoothness of the graph

Changing the size may change the difficulty to see the spread of data values

Stats programs will automatically determine what values it calculates should be grouped together

- You are not bound by what the program does
- Play around with the bin sizes to see what the data look like

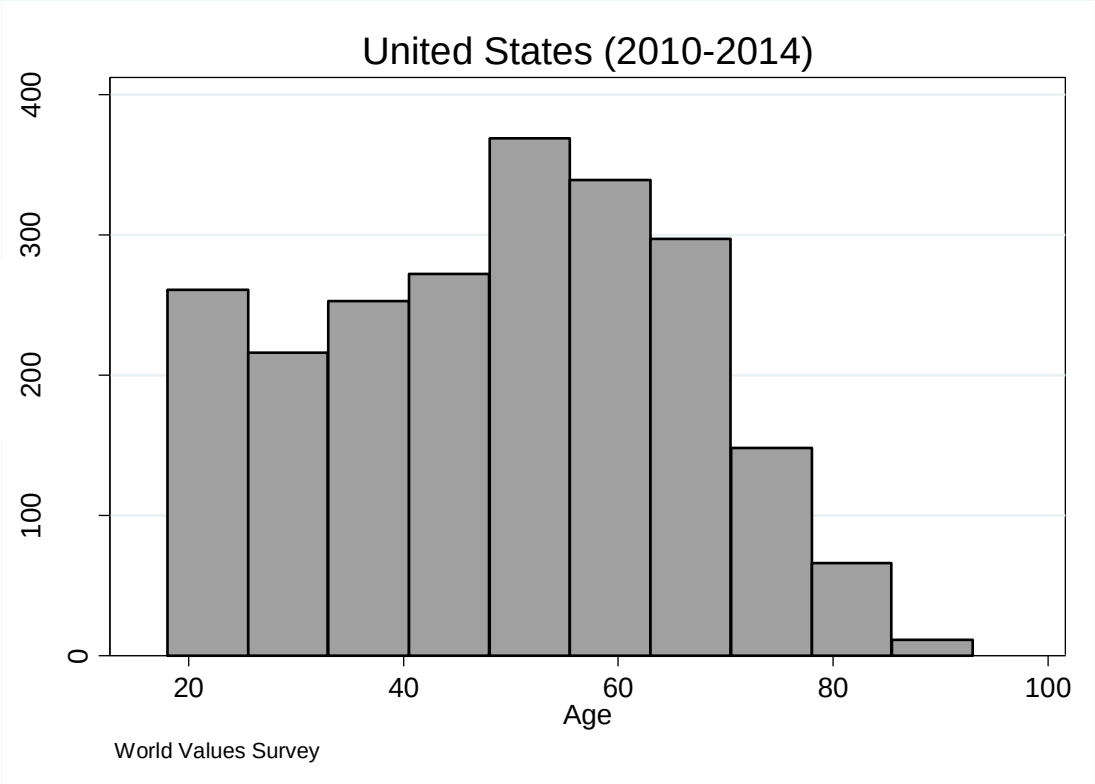
Bin Size, example



Bin Size, cont.

Fewer Bins

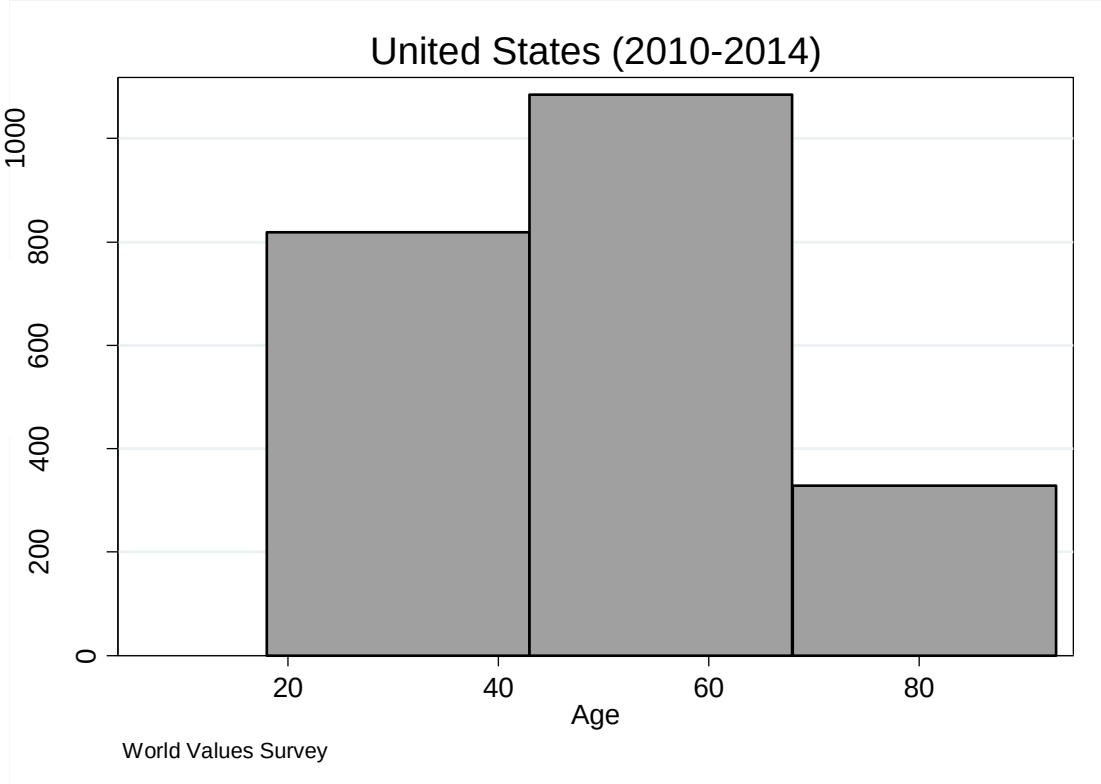
Age of Participant
United States
(2011)



Bin Size, cont.

Fewer Bins

Age of Participant
United States
(2011)



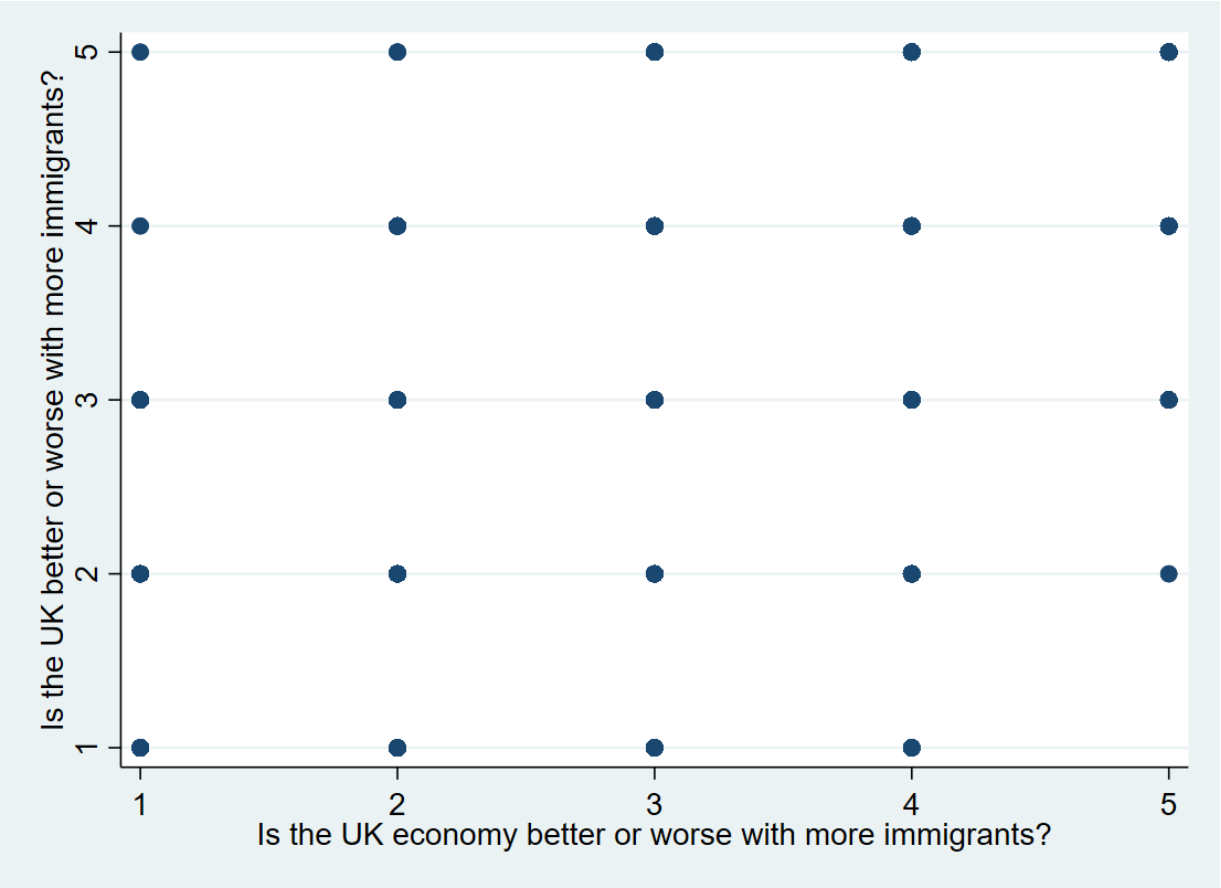
Hidden Observations

Be careful with plotting symbols used

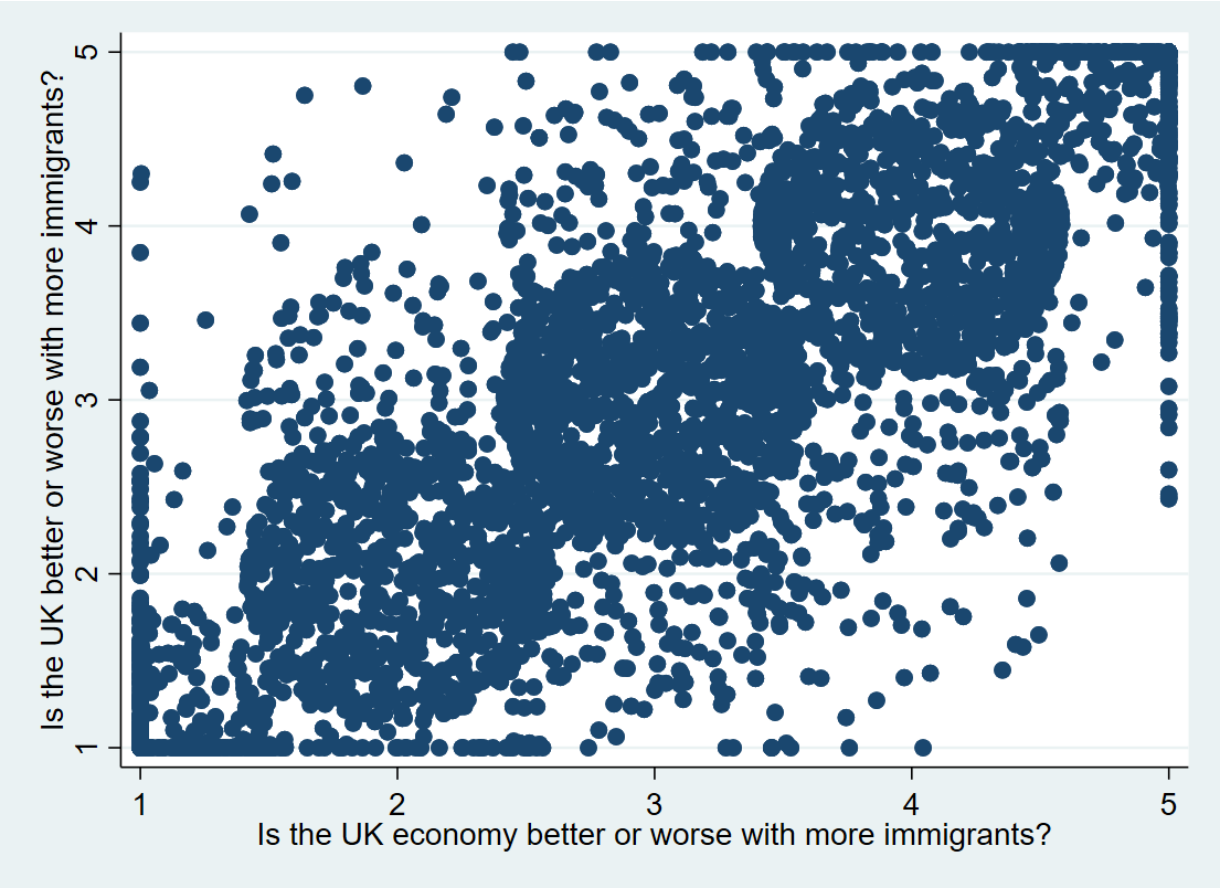
Consider using open/empty symbols versus solid plotting symbols

Consider "jittering" your data points

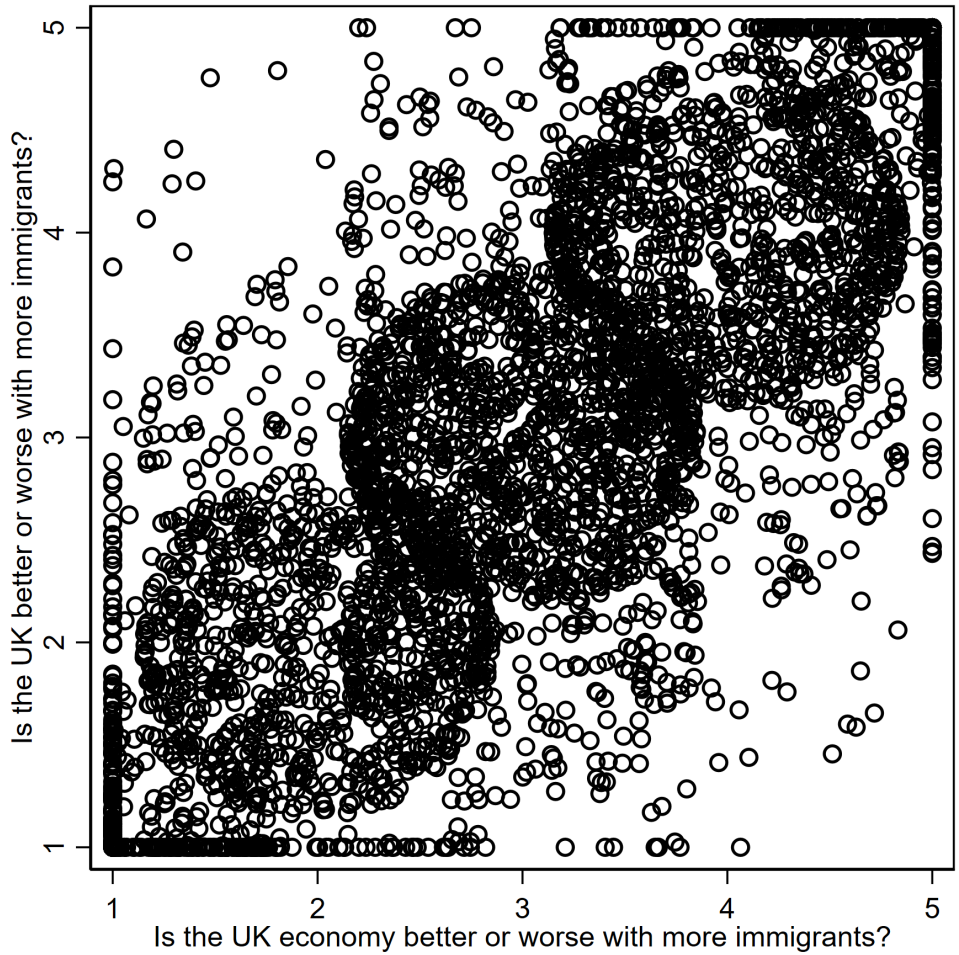
Original Graph



Graph+Jittering



Graph+Jitter+Hollow Symbols



How Short-Cuts can Fail Us

Eyes give cues of what is around

Brain puts cues together to make best guess based on prior experiences

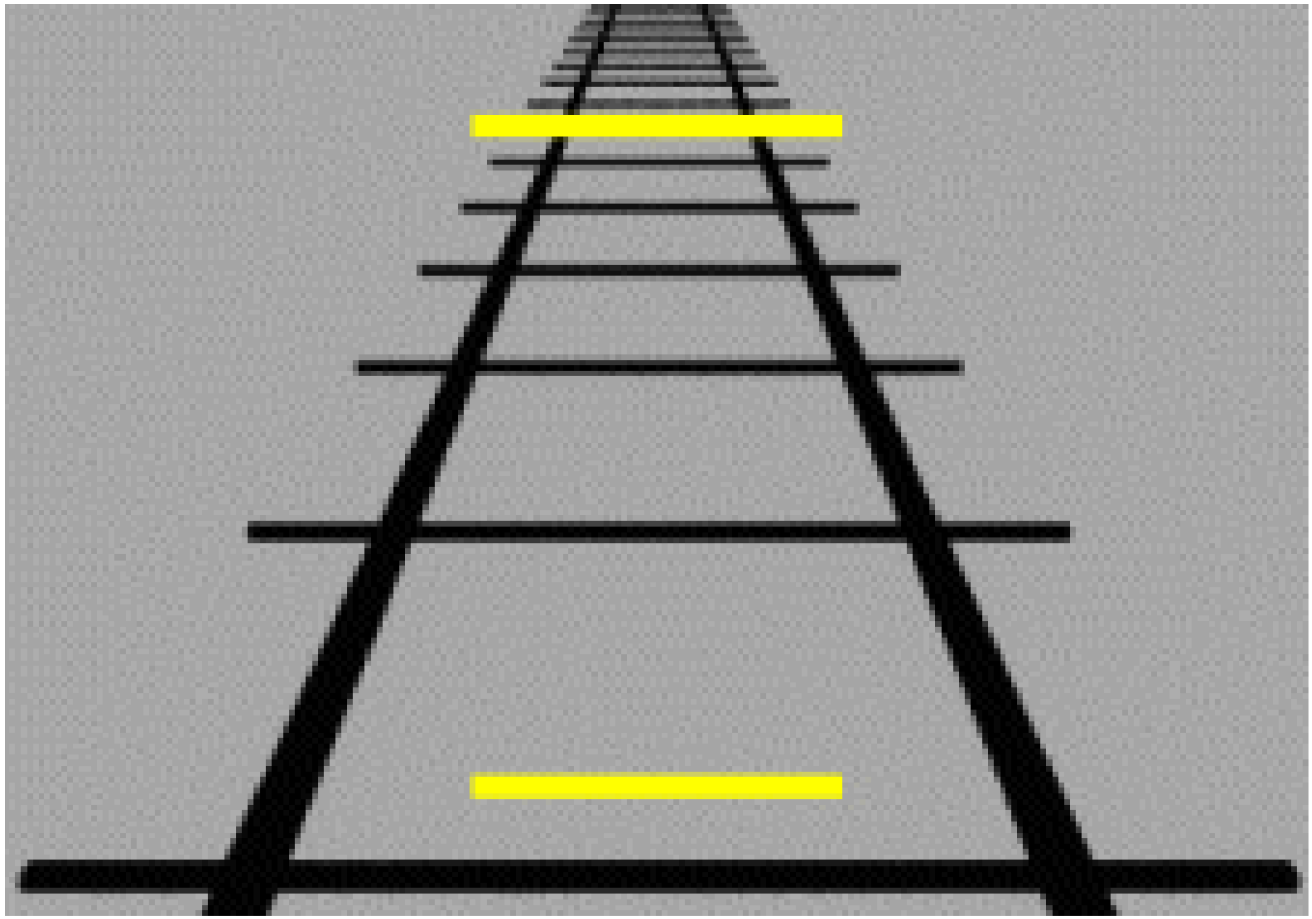
Think of driving and using your mirrors

- Not constant update of what is around you but estimates

Brain uses short-cuts when looking at data visualisations too

The Eyes Lie, Example 1

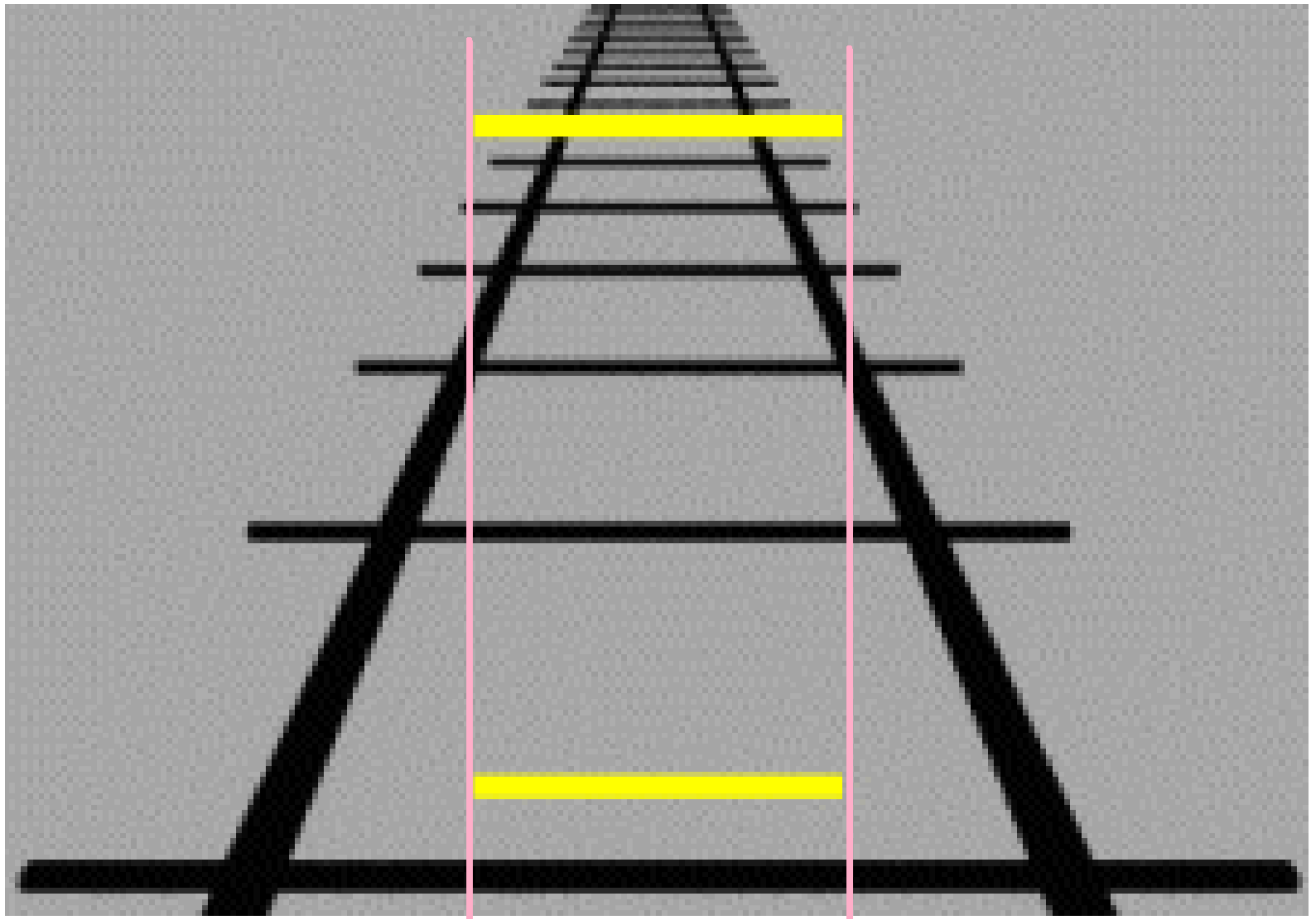
Which line is bigger?



By Tony Philips, National Aeronautics and Space Adm. - NASA - Summer Moon Illusion (image link), Public Domain, https://commons.wikimedia.org/w/index.php?curid=1211098

The Eyes Lie, Example 1

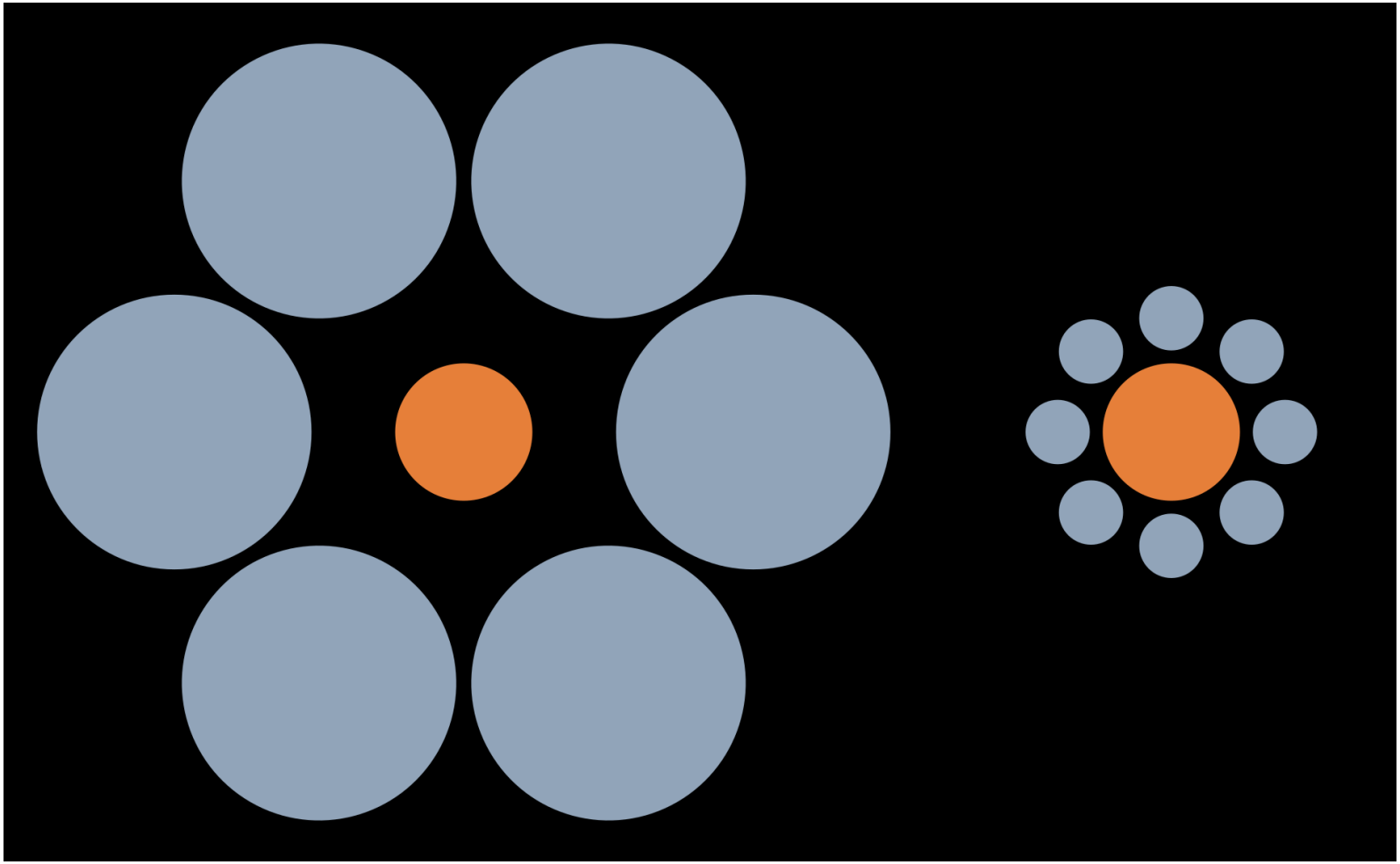
Which line is bigger?



By Tony Philips, National Aeronautics and Space Adm. - NASA - Summer Moon Illusion (image link), Public Domain, https://commons.wikimedia.org/w/index.php?curid=1211098

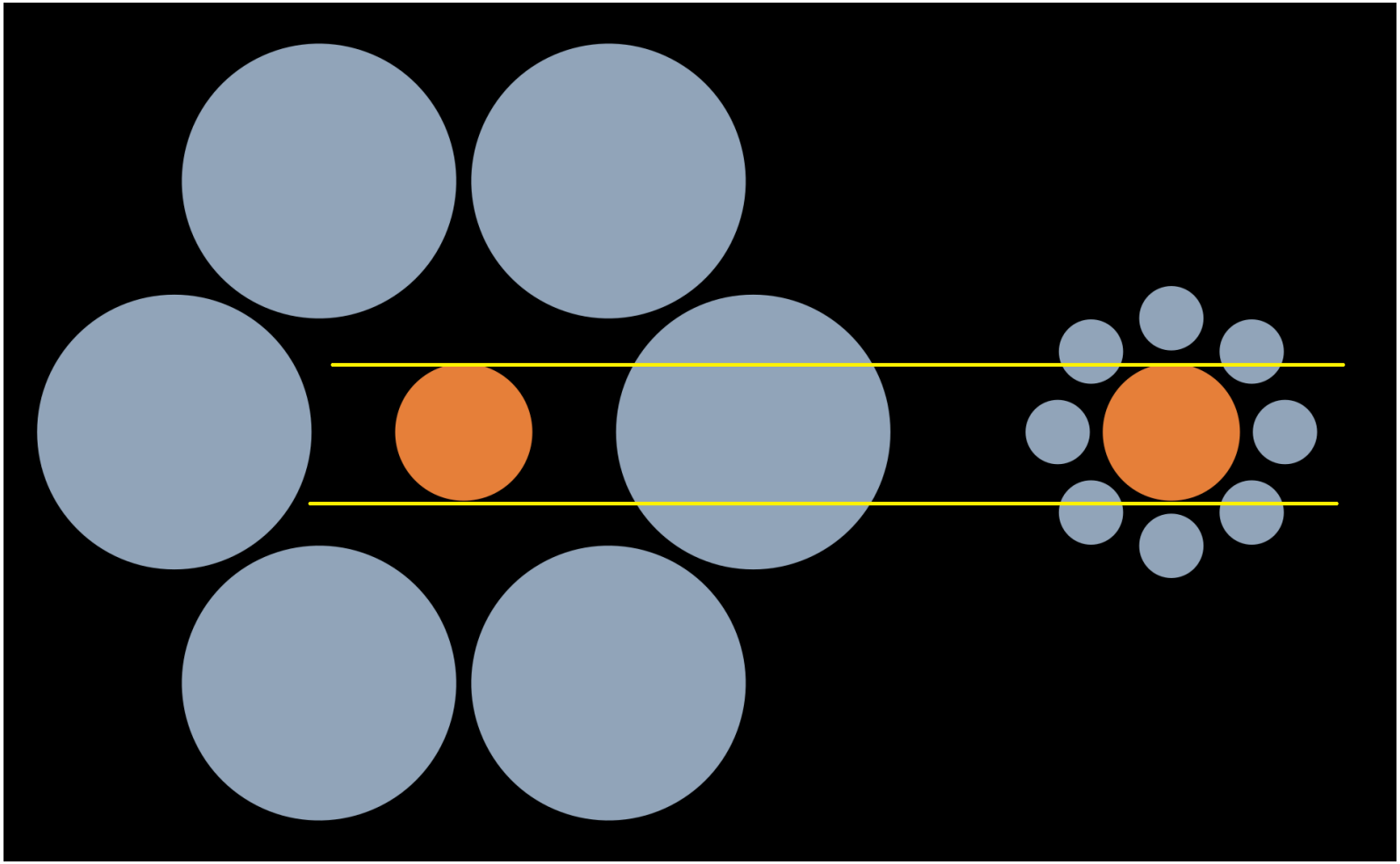
The Eyes Lie, Example 2

Which circle is smaller?



The Eyes Lie, Example 2

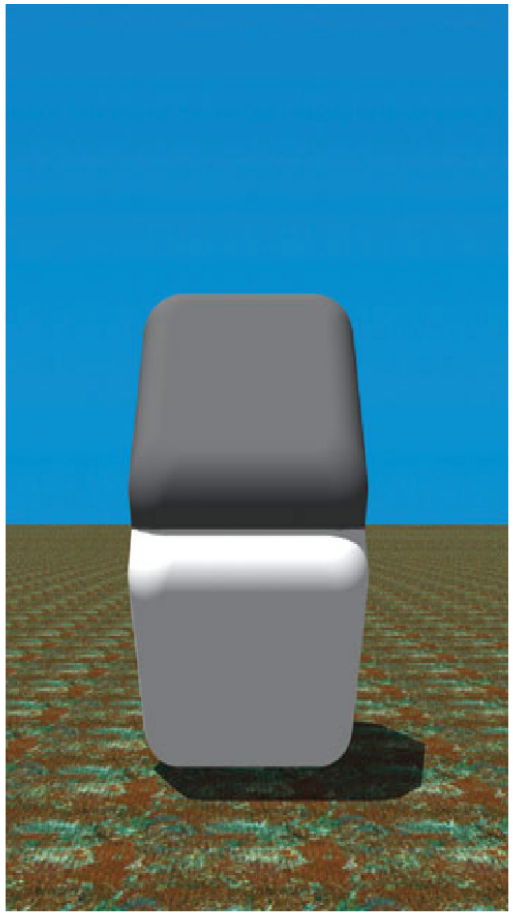
Which circle is smaller?



By No machine-readable author provided. Phrood~commonswiki assumed (based on copyright claims). - No machine-readable source provided. Own work assumed (based on copyright claims)., Public Domain, https://commons.wikimedia.org/w/index.php?curid=828098By No machine-readable author provided. Phrood~commonswiki assumed (based on copyright claims). - No machine-readable source provided. Own work assumed (based on copyright claims)., Public Domain, https://commons.wikimedia.org/w/index.php?curid=828098

The Eyes Lie, Example 3

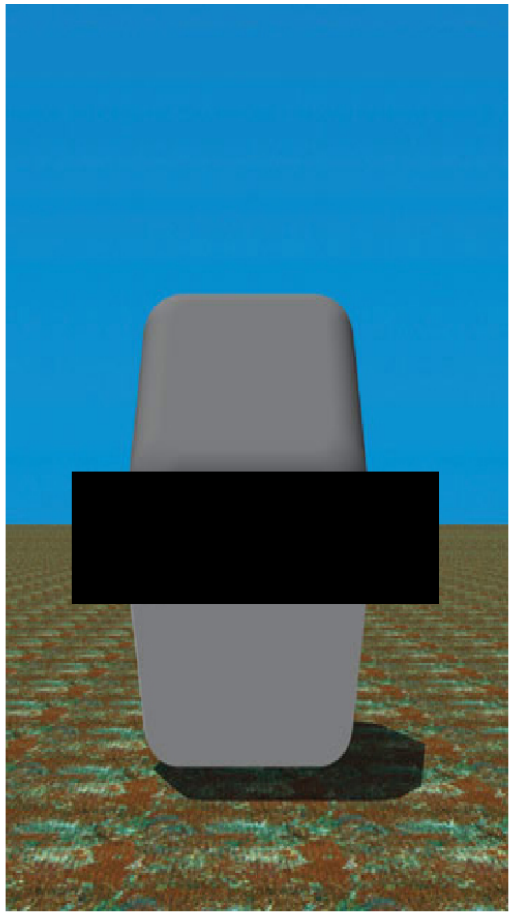
What color are the two sides of the box?



Dale Purves

The Eyes Lie, Example 3

What color are the two sides of the box?

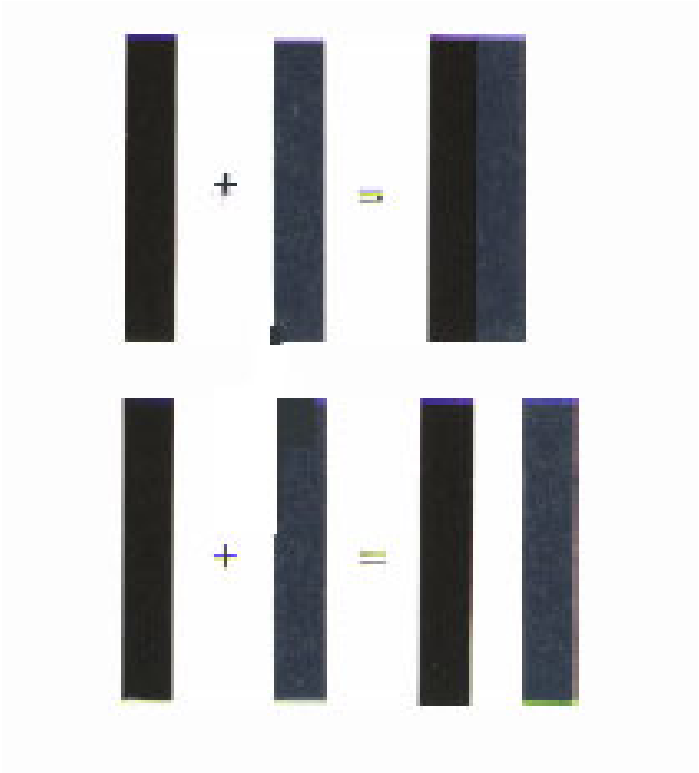


Dale Purves

1+1=3

Brain will interpret blank space as data at times

How many lines are there?



Levels of Processing

There are some things that are area easier for the brain to process than others

A hierarchy was proposed at one point based on research

Elements listed from easiest to process to more difficult

- Position on aligned scales
- Positions on unaligned scales
- Length
- Slopes/angles
- Area
- Volume
- Curvatures
- Shading
- Color saturation

Hierarchy of Basic Graphs

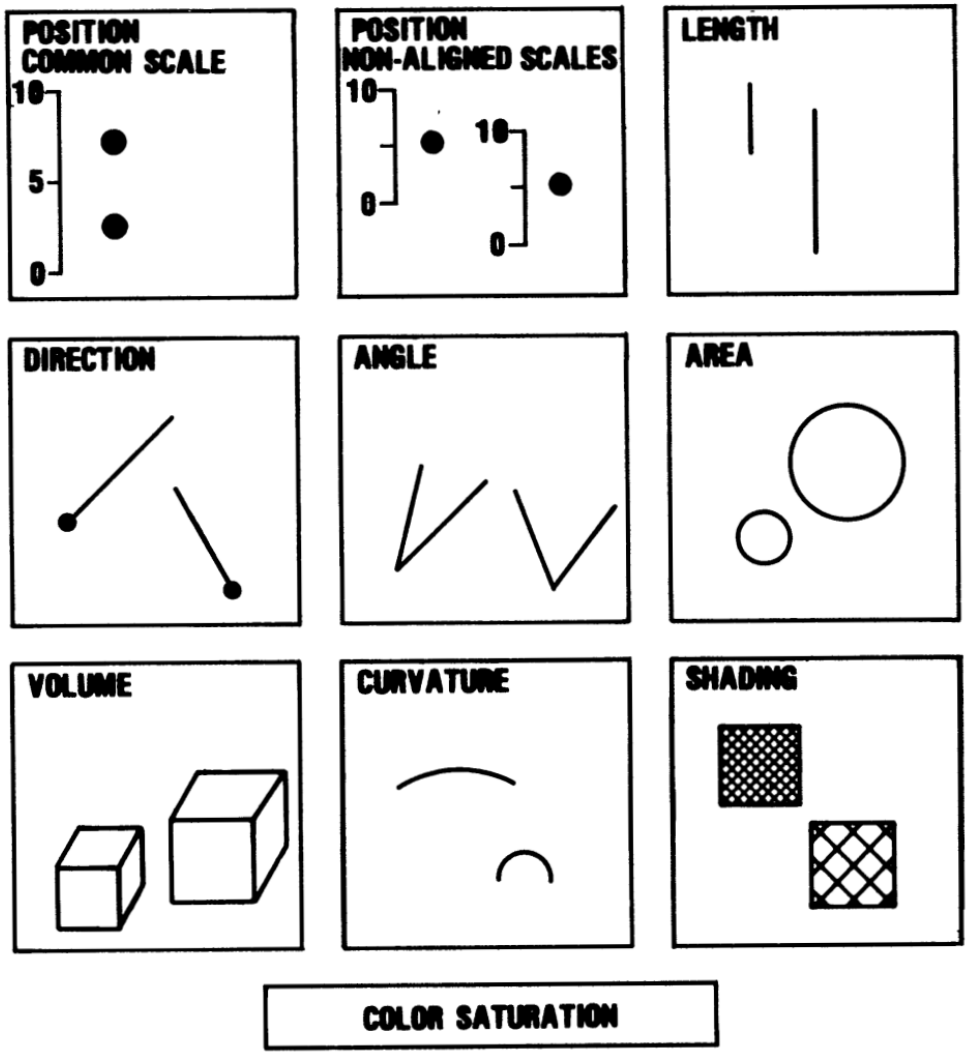
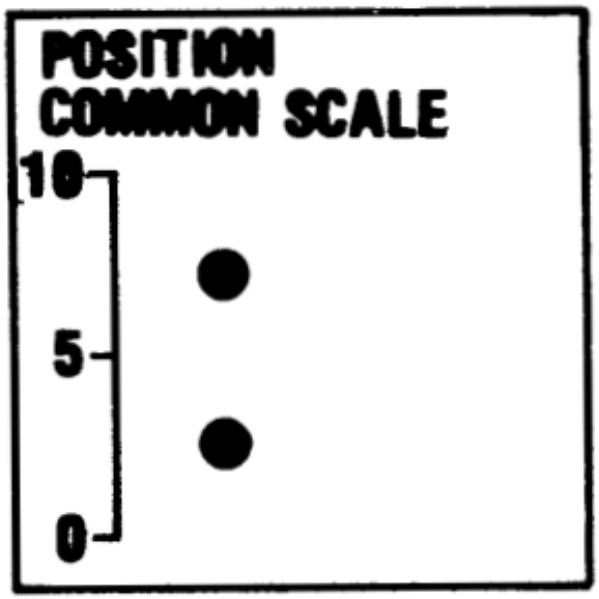
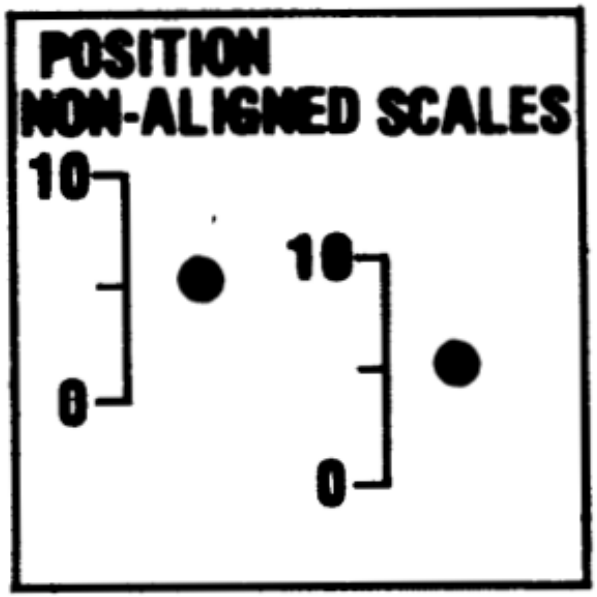


Figure 1. Elementary perceptual tasks.

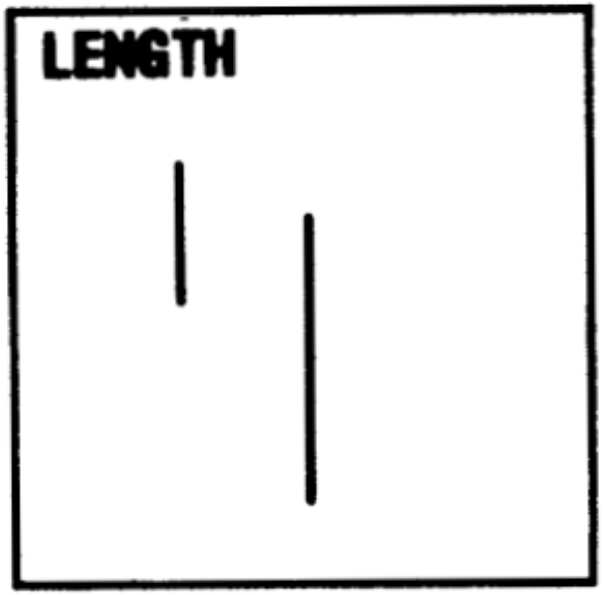
Hierarchy of Basic Graphs, cont.



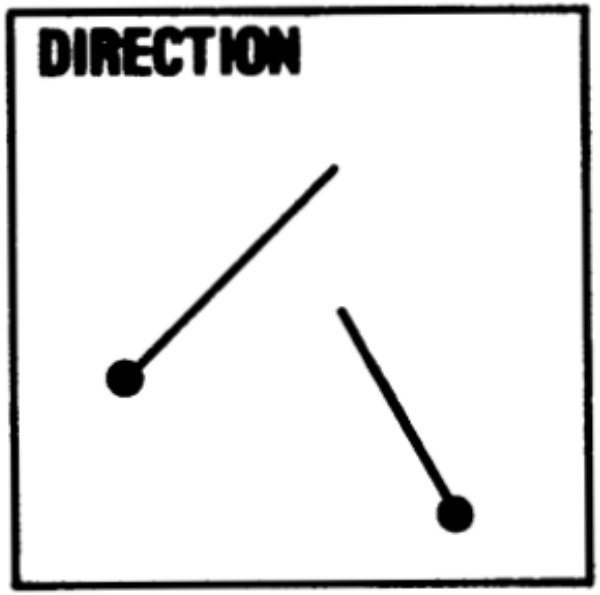
Hierarchy of Basic Graphs, cont.



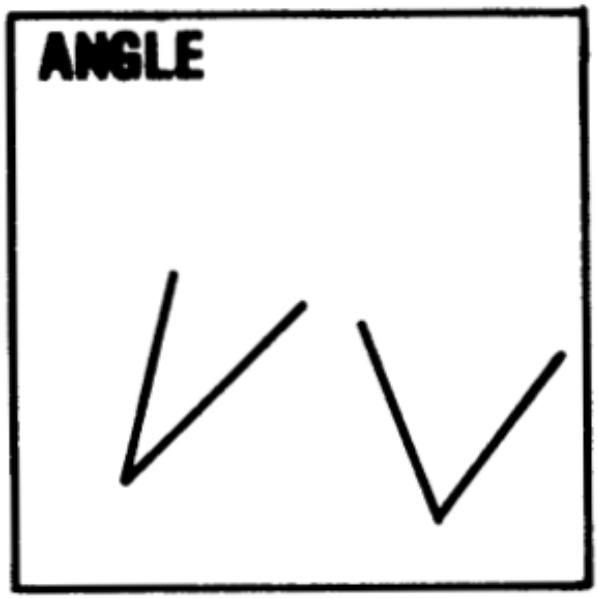
Hierarchy of Basic Graphs, cont.



Hierarchy of Basic Graphs, cont.



Hierarchy of Basic Graphs, cont.



Color

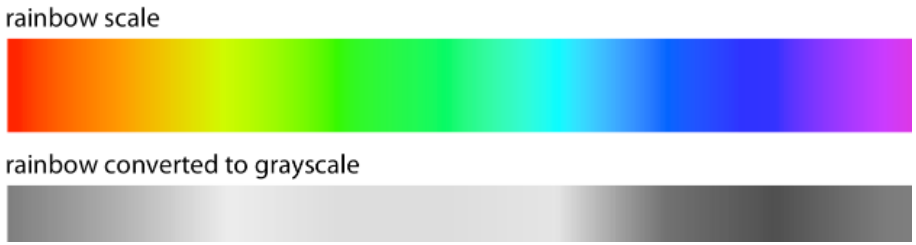
First, not everyone sees color the same way

Print usually in grey scale

Color itself is not necessarily intuitive to provide visual ordering

If a key is necessary to decode the color pattern, less useful

Look at gray scale options



Opt for variation in line patterns:

- Solid line
- Dashed line
- Dashed-dot

Missing Information

Omitted information can skew understanding

What is the scale of something

What do symbols mean

What is actually being measured

Missing information

source: Erikson Times

Labels

Labels are necessary for clarity

Labels include:

- Main Title
- Y and X axis titles
- Symbols
- Color

Labels should describe data and not be data set variable names

Challenge of 3 Dimensions

Technology advances and increases ease of making complex things

Does not mean our brain catches up to this

Keep in mind how visualisations will be viewed

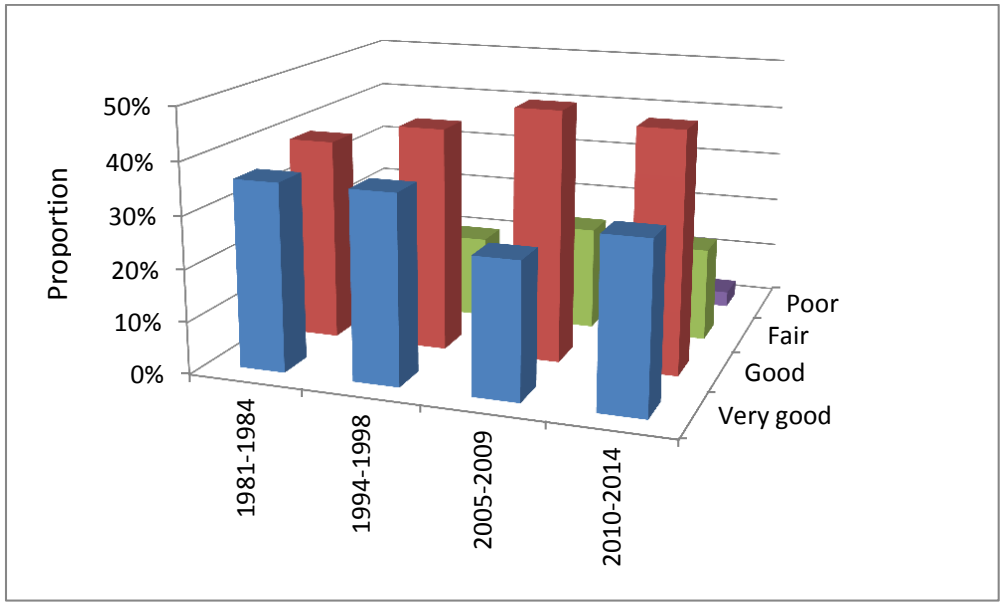
Can reader see all the data?

Have you introduced any bias into the image?

3 Dimensional Challenge

All in all, how would you describe your state of health these days? Would you say it is...

Australia



Source: World Values Survey

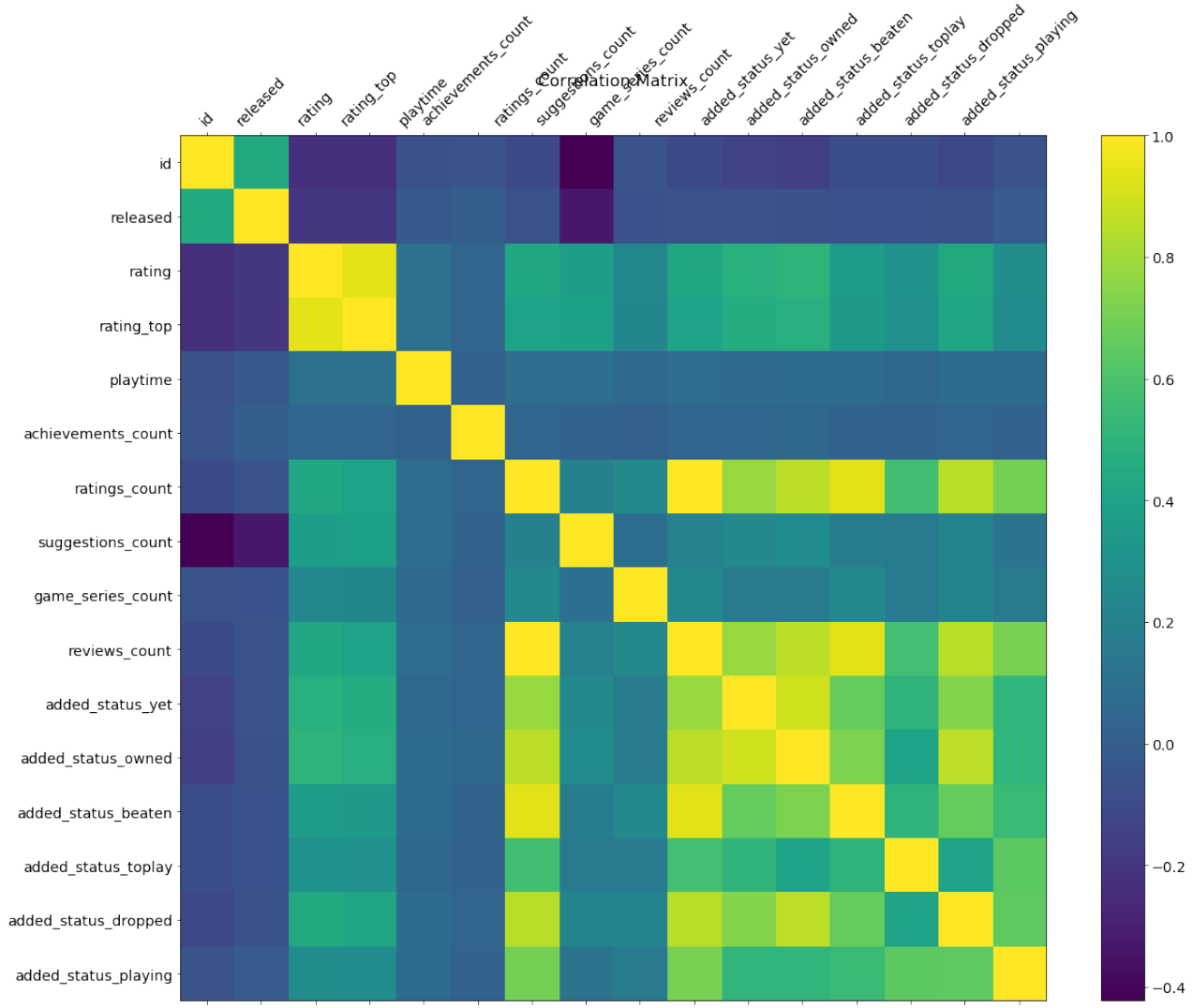
Any issues with this graph?

What are the values for fair and poor?

Perspective?

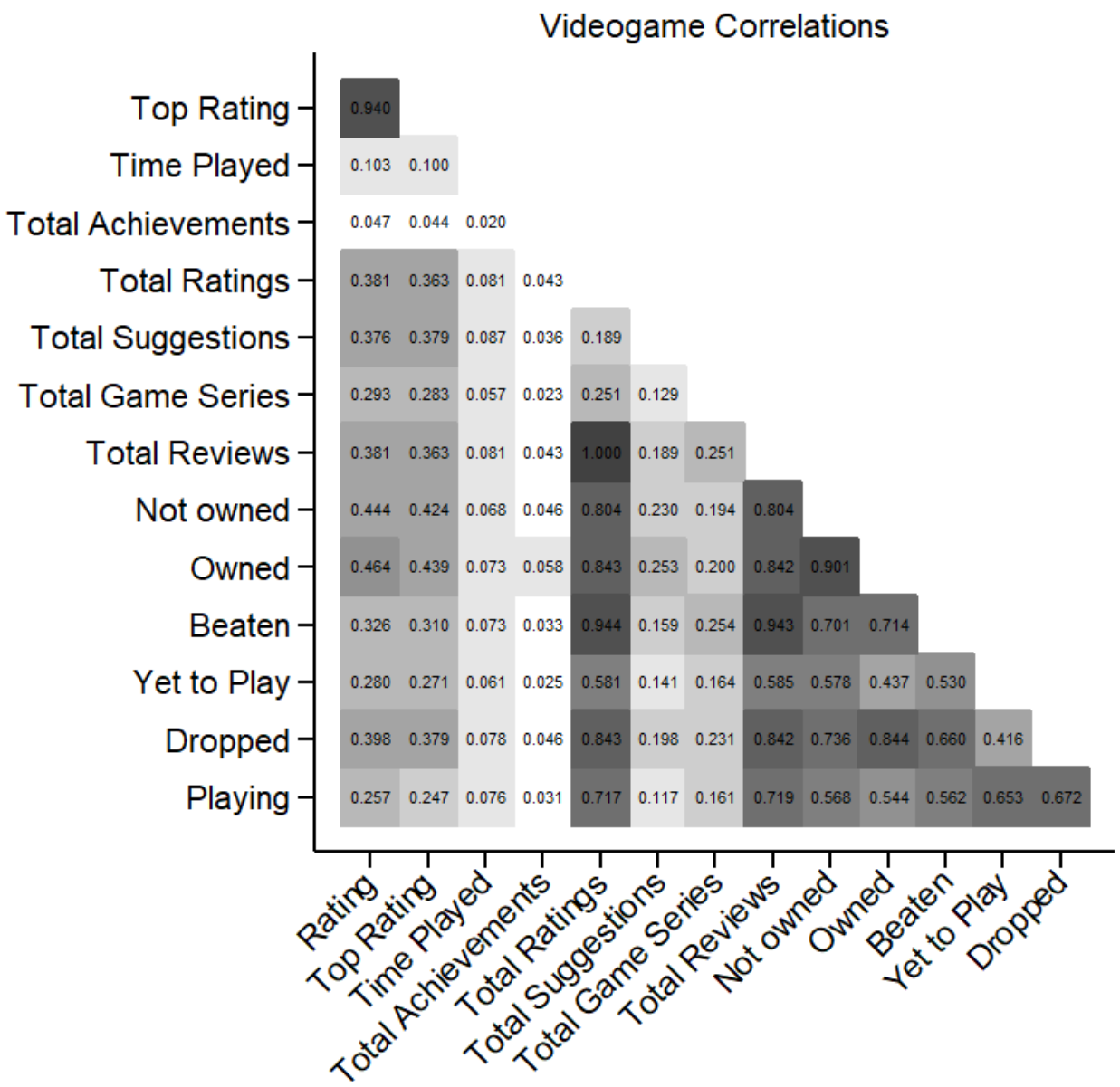
Graph Revisited

Let's look at the graph from the start



What would you change now?

Graph Revised



Erasing Principle

Erasing principle is the idea of removing unneeded pieces

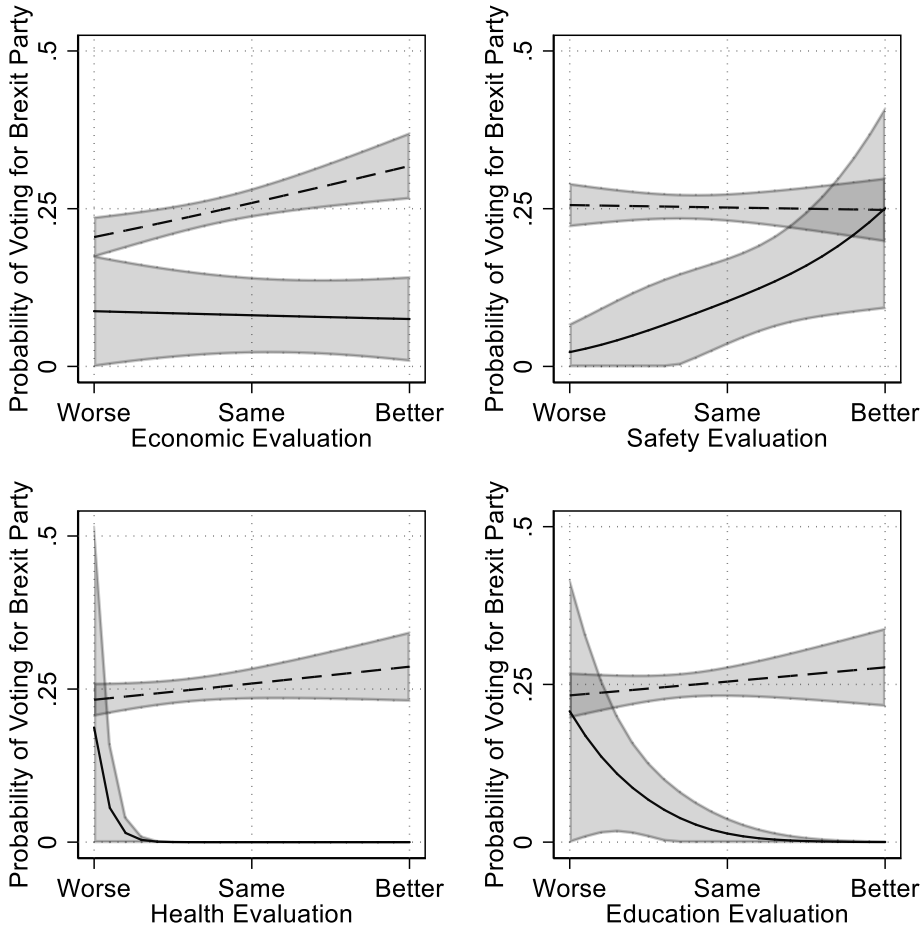
If you have three graphs in an image and they all use the same color scheme, do you need three legends?

If you have three graphs lined up that have the same y-axis do you need three sets of labels?

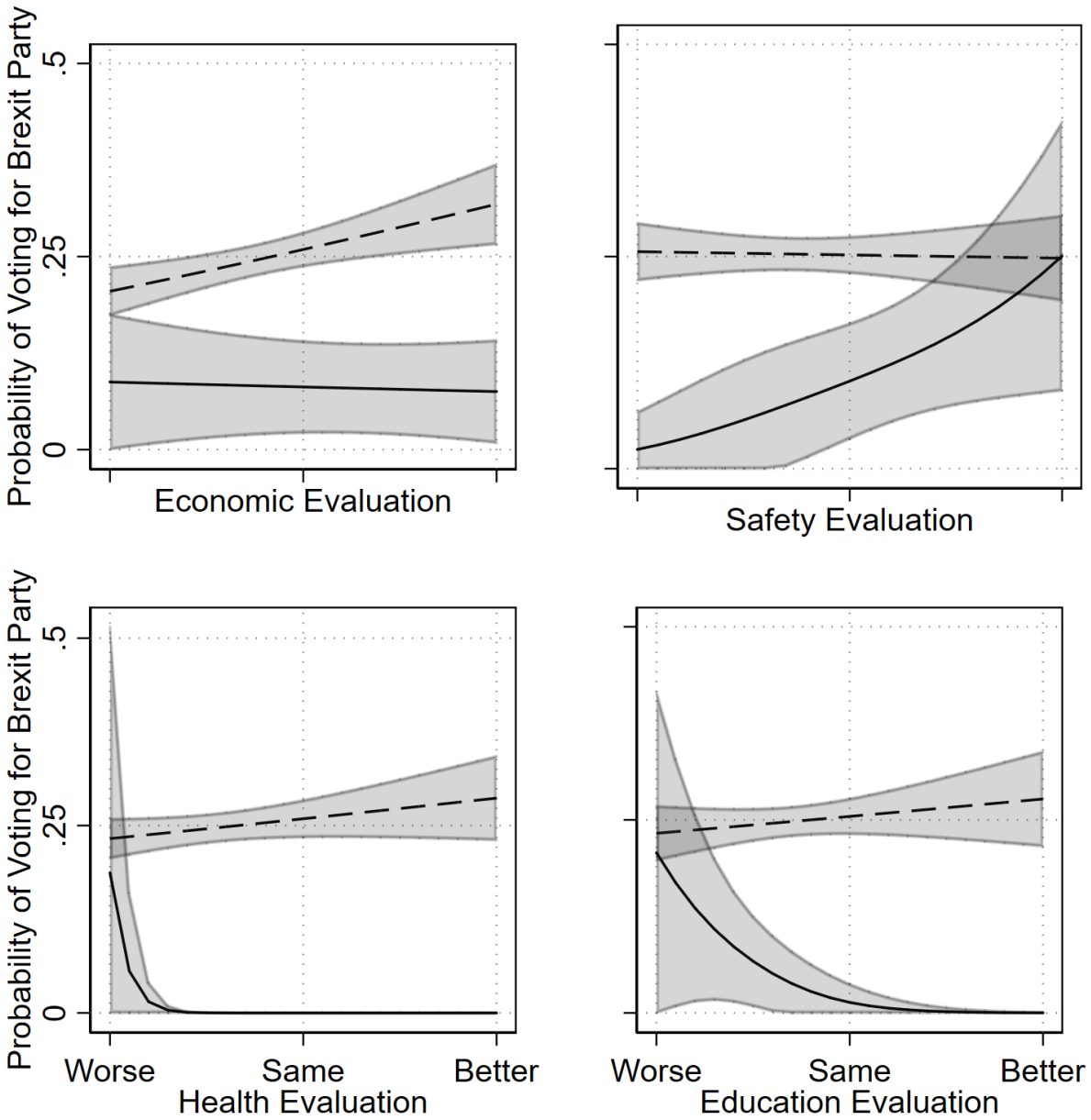
Do you need need the information to understand the image?

Erasing Example

Figure 2: **Predicted Probability Plots of Voting for the Brexit Party based on Attribution of Responsibility and Evaluations.**



Erasing Example, cont.



Remove Redundant Data

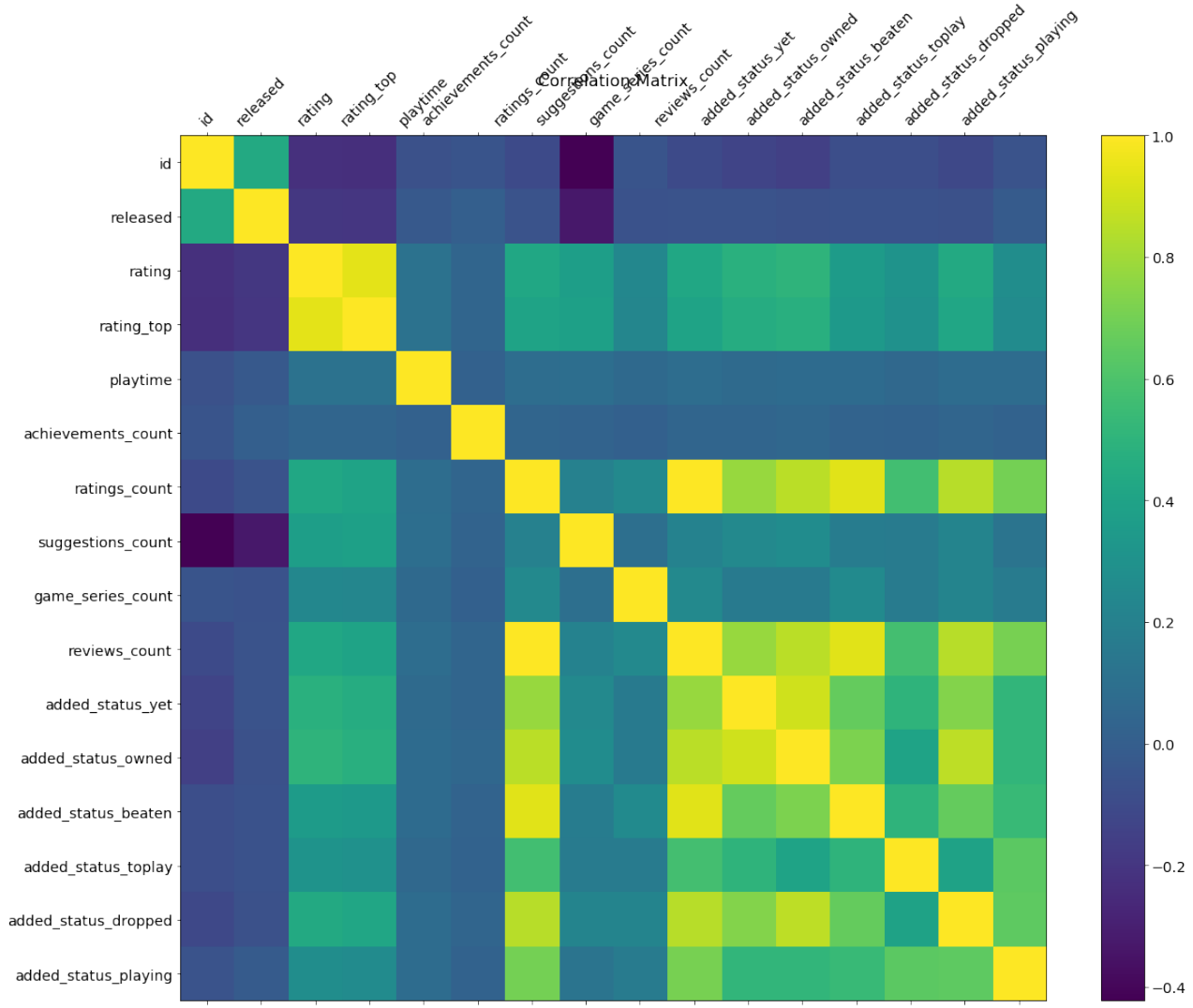
In addition to erasing unneeded details

Remove duplicate data/information

Let's look back at our starting example

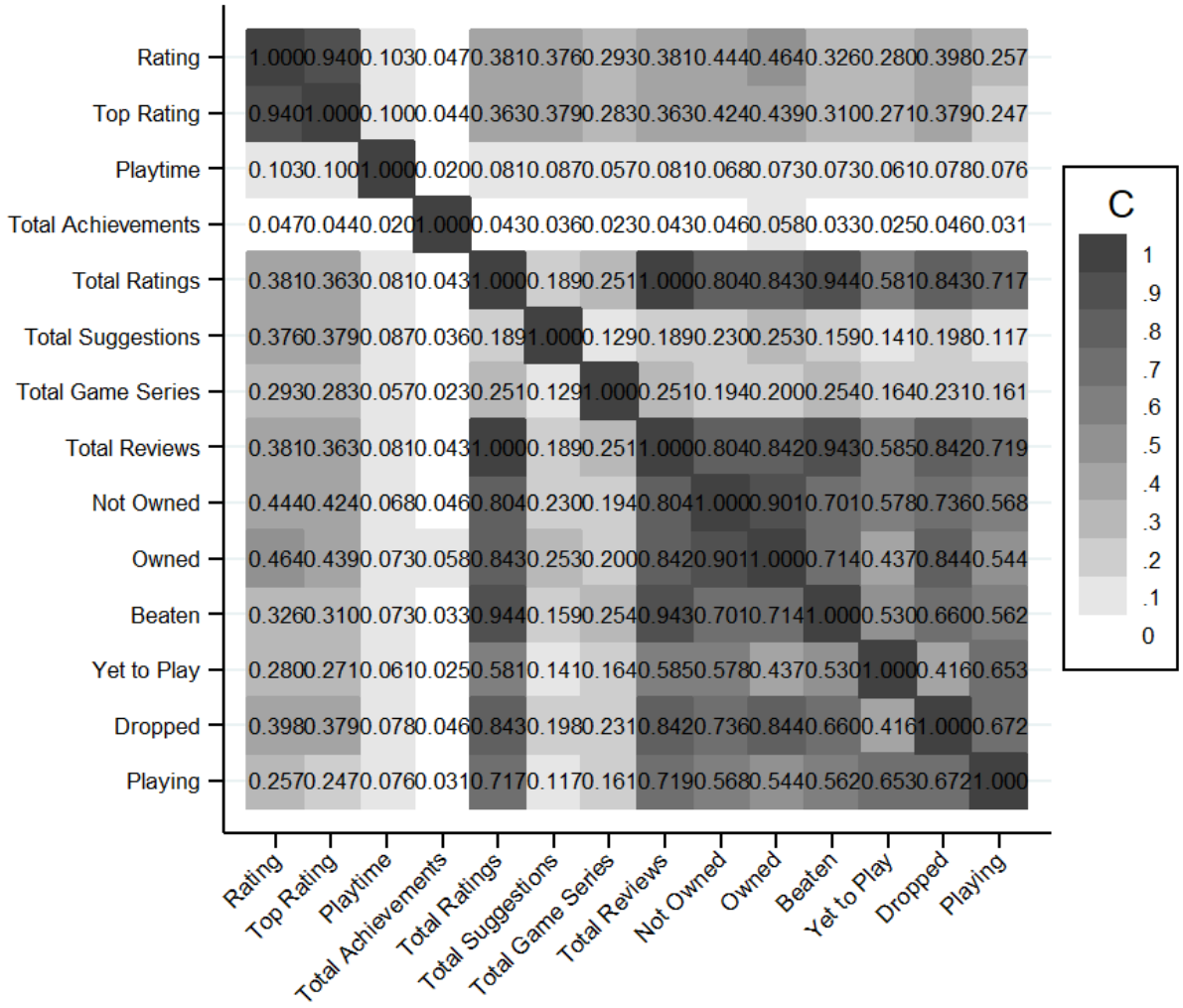
Graph Revisited

Let's look at the graph from the start



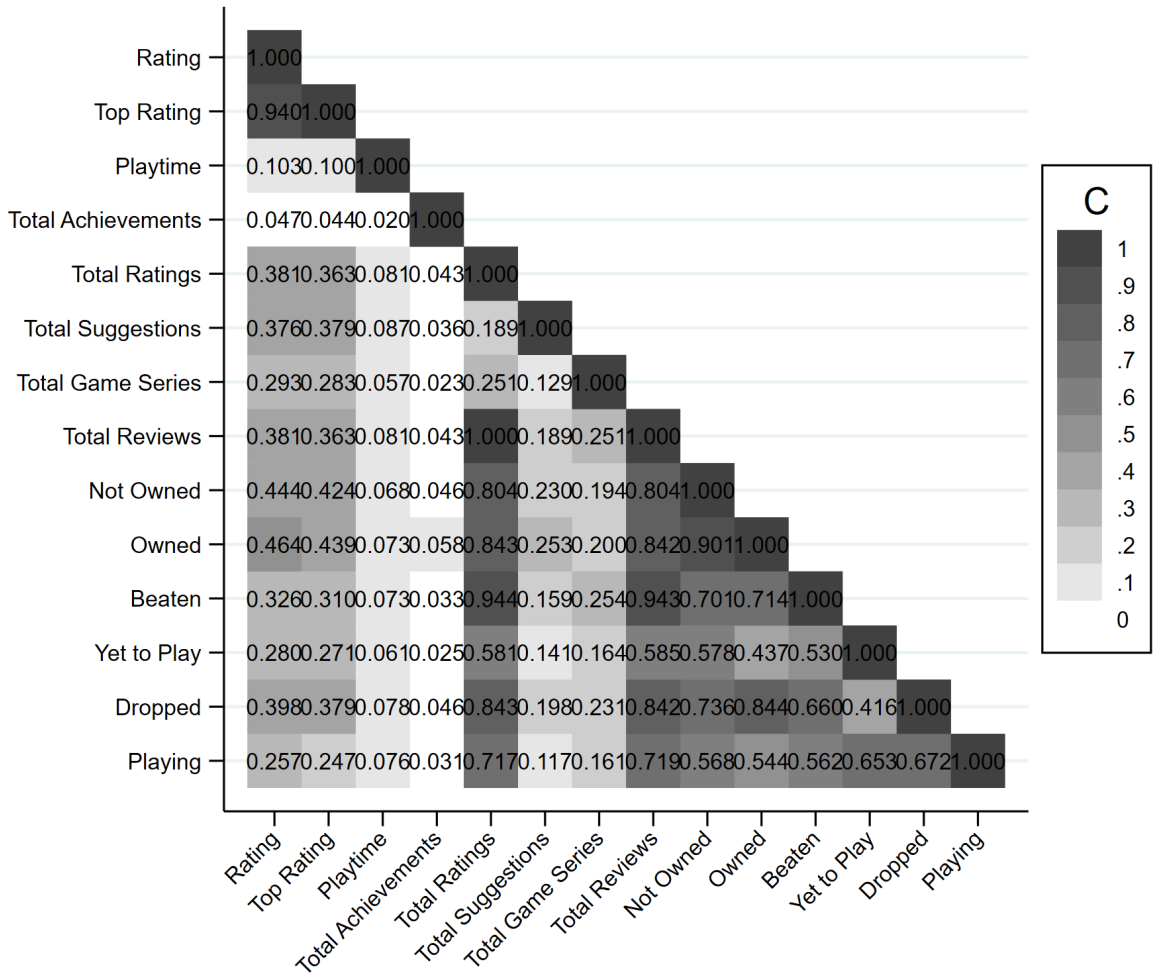
What data repeats itself?

Redundant Data



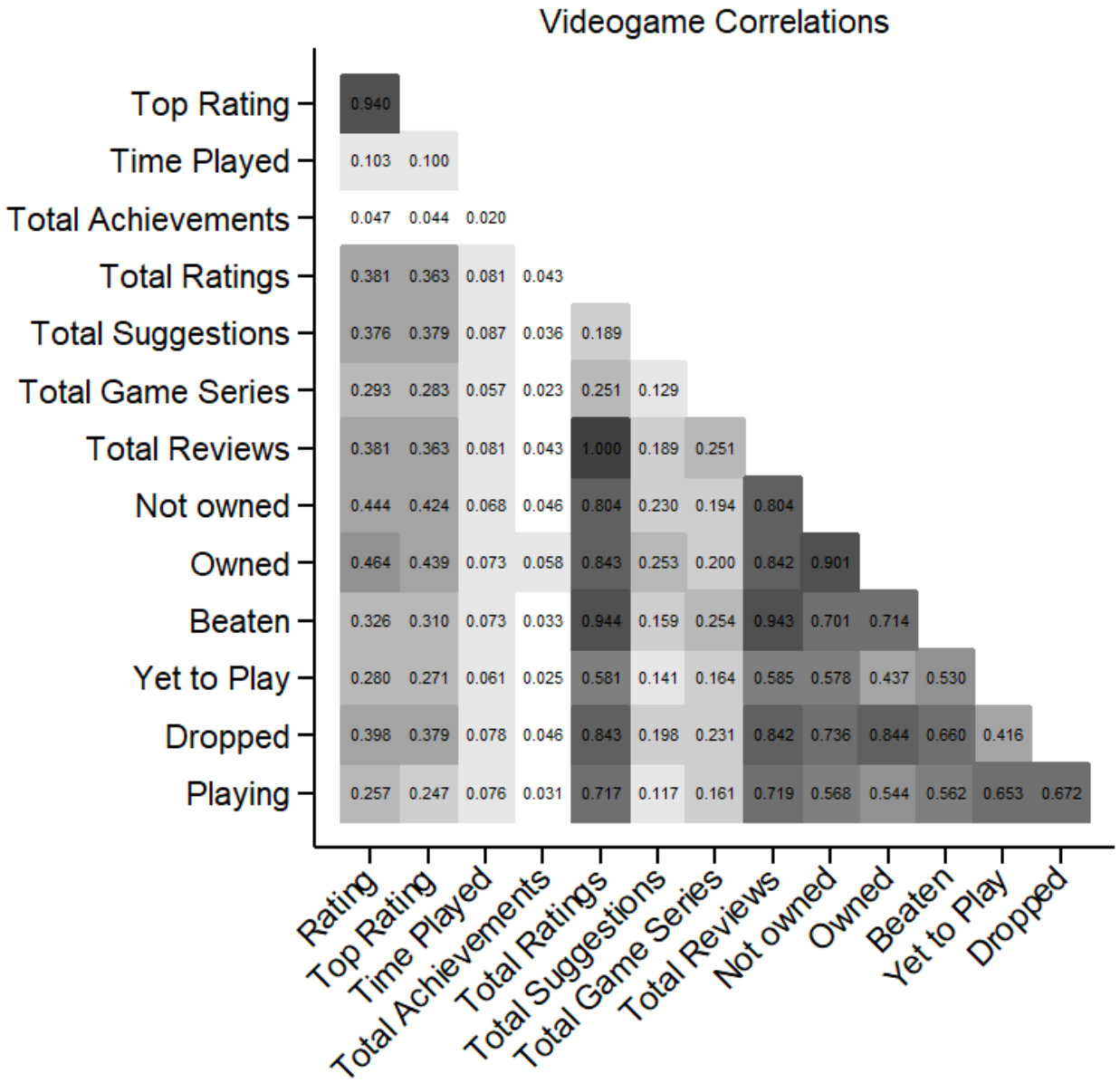
What is redundant now?

Redundant Data



What is redundant now?

Redundant Data



Revise and Edit

Preparing data visualisations in not a one-time process

Revise, examine, revise, examine, go back

Sometimes you go too far

- Think of the histogram and bins example

Key Points

Key take aways:

- More is not always better
- Fancy is not always clear
- All data visualisations should seek to make patterns and messages easier to see
- Default settings not always best
- Revising and editing takes time

Warning to Makers

Beware of unintended bias being introduced

Be aware of accessibility issues

Be careful not to mask/hide your data

Visuals should be accessible outside of the writing

Warning to Readers

Check the data sources

Check the scales and axes

Ask what information is missing

Conclusion

Questions?

Contact information: Niccole.Pamphilis@glasgow.ac.uk

Potential Reading

- Healy, K. Data Visualization: A Practical Introduction. 2018.
  `https://socviz.co/index.htmla`
- Tufte, E., Visual explanations: Images and quantities, evidence and narrative. Graphics Press, 1997.
- Cleveland, WS, and R. McGill.Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. 1984. Journal of the American Statistical Association. 79.387:531-54.
- Tukey, JW. Exploratory data analysis. Vol. 2. Reading, MA: Addison-wesley, 1977.