How To Think About Visualizing Your Data

Kieran Healy
Duke University

WAPOR / December 15th, 2023
The Problem
YOU SHOULD LOOK AT YOUR DATA
SEEING IS NOT AS SIMPLE AS IT LOOKS
These are two perfectly geometrical circles.
Why this stuff matters
Anscombe's Quartet
Cairo; Matejka & Fitzmaurice
Gun deaths in Florida

Number of murders committed using firearms

2005 Florida enacted its ‘Stand Your Ground’ law

Source: Florida Department of Law Enforcement

C. Chan 16/02/2014
Which of the following statements best describes the data in the graph below?

Pew Research

A. In recent years, the rate of cavities has increased in many countries
B. In some countries, people brush their teeth more frequently than in other countries
C. The more sugar people eat, the more likely they are to get cavities
D. In recent years, the consumption of sugar has increased in many countries
Which of the following statements best describes the data in the graph below?

![Graph showing correlation between average sugar consumption and average number of decayed teeth per person in different countries.]

Pew Research

A. In recent years, the rate of cavities has increased in many countries
B. In some countries, people brush their teeth more frequently than in other countries
C. The more sugar people eat, the more likely they are to get cavities
D. In recent years, the consumption of sugar has increased in many countries
Which of the following statements best describes the data in the graph below?

Pew Research

A. In recent years, the rate of cavities has increased in many countries

B. In some countries, people brush their teeth more frequently than in other countries

C. The more sugar people eat, the more likely they are to get cavities

D. In recent years, the consumption of sugar has increased in many countries
Some charts are better than others
BAD TASTE
BAD DATA
BAD PERCEPTION
1. Bad Taste
Simplify, simplify?

The Visual Display of Quantitative Information

Edward Tufte
“Chart junk”

Edward Tufte
...And its limits

Three kinds of boxplot
2. Bad Data
Percentage of people who say it is “essential” to live in a democracy


New York Times
Graph by Erik Voeten, based on WVS 5
3. Bad Perception
Contrasts and Edges

Mach Bands
Contrasts and Edges

Mach Bands
Contrasts and Edges

Four gray rectangles
Contrasts and Edges

With a gradient background
Contrast and Color

Colin Ware
Contrast and Color

Colin Ware
Adelson’s checkerboard

Edward Adelson
Visual Tasks for Decoding Graphs
What’s a graph, anyway?
Data

Encoded or mapped

Some sort of faithful visual representation

Encode ...
... and decode
Stacked columns
William Cleveland
William Cleveland
Rainbow gradients are not linear in the luminance channel

Perceptually Uniform Sequential colormaps

Luminance-balanced gradients
Luminance-balanced sequential palettes
Luminance-balanced diverging and qualitative/unordered palettes
A rough hierarchy of mappings for data
<table>
<thead>
<tr>
<th>Property</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position on a common scale</td>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
<tr>
<td>Position on unaligned scale</td>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
<tr>
<td>Length (1D as size)</td>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
<tr>
<td>Tilt or Angle</td>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
<tr>
<td>Area (2D as size)</td>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
<tr>
<td>Depth (3D as Position)</td>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
<tr>
<td>Color luminance [brightness]</td>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
<tr>
<td>Color saturation [intensity]</td>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
<tr>
<td>Curvature</td>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
<tr>
<td>Volume (3D as size)</td>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
</tbody>
</table>
Learning How
The R Project for Statistical Computing
The ggplot2 package
ggplot implements a grammar of graphics
The grammar is a set of rules for how to produce graphics from data, by *mapping* data to or *representing* it by geometric *objects* (like points and lines) that have aesthetic *attributes* (like position, color, size, and shape), together with further rules for transforming data if needed, for adjusting scales and their guides, and for projecting results onto some coordinate system.
Like other rules of syntax, the grammar limits what you can validly say but it doesn’t automatically make what you say sensible or meaningful.
Data represented by visual elements; like position, length, color, and size;
Each measured on some scale;
Each scale with a labeled guide;
With the plot itself also titled and labeled.
How does ggplot do this?
**ggplot2**’s flow of action

<table>
<thead>
<tr>
<th>gdp</th>
<th>lifexp</th>
<th>pop</th>
<th>continent</th>
</tr>
</thead>
<tbody>
<tr>
<td>340</td>
<td>65</td>
<td>31</td>
<td>Euro</td>
</tr>
<tr>
<td>227</td>
<td>51</td>
<td>200</td>
<td>Amer</td>
</tr>
<tr>
<td>909</td>
<td>81</td>
<td>80</td>
<td>Euro</td>
</tr>
<tr>
<td>126</td>
<td>40</td>
<td>20</td>
<td>Asia</td>
</tr>
</tbody>
</table>

What we start with
A Gapminder Plot

**Life Expectancy**

**log GDP**

**Continent**
- Asia
- Euro
- Amer

**Population (m)**
- 0-35
- 36-100
- >100

Where we're going
ggplot’s flow of action

1. Tidy Data

```
p <- ggplot(data = gapminder, ...)```

<table>
<thead>
<tr>
<th>gdp</th>
<th>lifexp</th>
<th>pop</th>
<th>continent</th>
</tr>
</thead>
<tbody>
<tr>
<td>340</td>
<td>65</td>
<td>31</td>
<td>Euro</td>
</tr>
<tr>
<td>227</td>
<td>51</td>
<td>200</td>
<td>Amer</td>
</tr>
<tr>
<td>909</td>
<td>81</td>
<td>80</td>
<td>Euro</td>
</tr>
<tr>
<td>126</td>
<td>40</td>
<td>20</td>
<td>Asia</td>
</tr>
</tbody>
</table>

2. Map the Aesthetics

```
p <- ggplot(data = gapminder,
        mapping = aes(x = gdp,
                       y = lifexp,
                       size = pop,
                       color = continent))
```

3. Pick a Geom

```
p + geom_point()
```

Core steps

1. Get the data in the right shape. This is usually long format.
2. Decide how your variables will be represented by things you can see.
3. Decide what kind of plot, or series of plots, you want to draw.
ggplot’s flow of action

4. Fix Scales and Co-Ordinate System

```r
p + coord_cartesian() + scale_x_log10()
```

4. Adjust scales and their markings. Not just x and y but also color, size, etc,

5. Add Labels and Adjust Guides

```r
p + labs(x = “log GDP”,
        y = “Life Expectancy”,
        title = “A Gapminder Plot”)
```

5. Label your plot and adjust how the guides are displayed.

6. Add or Adjust Themes

```r
p + theme_minimal()
```

6. Style or adjust plot elements that are not directly representing your data.

Optional steps
You can think of this as a series of transformations, carried out by functions.
Example: Gapminder
Start with the data

### gapminder

```
# A tibble: 1,704 × 6
  country continent year  lifeExp pop  gdpPercap
  <fct>    <fct>   <int>  <dbl>  <int>     <dbl>
1 Afghanistan Asia 1952 28.8 8425333   779.         
2 Afghanistan Asia 1957 30.3 9240934   821.         
3 Afghanistan Asia 1962 32.0 10267083  853.         
4 Afghanistan Asia 1967 34.0 11537966  836.         
5 Afghanistan Asia 1972 36.1 13079460  740.         
6 Afghanistan Asia 1977 38.4 14880372  786.         
7 Afghanistan Asia 1982 39.9 12881816  978.         
8 Afghanistan Asia 1987 40.8 13867957  852.         
9 Afghanistan Asia 1992 41.7 16317921  649.         
10 Afghanistan Asia 1997 41.8 22227415  635.         
# i 1,694 more rows
```

```
dim(gapminder)
```

```
[1] 1704  6
```
Create a plot object

Data is the `gapminder` table.

```r
gapminder
  ggplot()
```
Map variables to aesthetics

Tell `ggplot` which visual elements represent which columns

```r
gapminder %>%
ggplot(mapping = aes(x = gdpPerCap, 
y = lifeExp, 
size = pop, 
color = continent))
```
What sort of plot?

gapminder ▷

```r
ggplot(mapping = aes(x = gdpPerCap,
                      y = lifeExp,
                      size = pop,
                      color = continent))
```

This empty plot has no geoms.
Pick a geom

gapminder ▷
  ggplot(mapping = aes(x = gdpPercap,
                      y = lifeExp,
                      size = pop,
                      color = continent)) +

  geom_point()

A scatterplot of Life Expectancy vs GDP
```
gapminder
  ggplot(mapping = aes(x = gdpPerCap, y = lifeExp, size = pop, color = continent)) +
  geom_smooth()
```

A smoothed lineplot of Life Expectancy vs GDP
Build your plots layer by layer

gapminder ▷
  ggplot(mapping = aes(x = gdpPercap,
                      y = lifeExp,
                      size = pop,
                      color = continent)) +

  geom_point()
This process is additive

gapminder ▶
ggplot(mapping = aes(x = gdpPercap,
        y = lifeExp,
        size = pop,
        color = continent)) +
geom_point() +
geom_smooth()
Functions take arguments

```r
gapminder
  ggplot(mapping = aes(x = gdpPerCap,
                       y = lifeExp,
                       size = pop,
                       color = continent)) +

  geom_point() +
  geom_smooth(method = "lm")
```

An ill-advised linear fit
#### A tibble: 1,704 × 6

<table>
<thead>
<tr>
<th>country</th>
<th>continent</th>
<th>year</th>
<th>lifeExp</th>
<th>pop</th>
<th>gdpPerCap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1952</td>
<td>28.8</td>
<td>8425333</td>
<td>779.</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1957</td>
<td>30.3</td>
<td>9240934</td>
<td>821.</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1962</td>
<td>32.0</td>
<td>10267083</td>
<td>853.</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1967</td>
<td>34.0</td>
<td>11537966</td>
<td>856.</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1972</td>
<td>36.1</td>
<td>13079460</td>
<td>740.</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1977</td>
<td>38.4</td>
<td>14888372</td>
<td>786.</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1982</td>
<td>39.9</td>
<td>12881816</td>
<td>978.</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1987</td>
<td>40.8</td>
<td>13867957</td>
<td>852.</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1992</td>
<td>41.7</td>
<td>16317921</td>
<td>649.</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>Asia</td>
<td>1997</td>
<td>41.8</td>
<td>22227415</td>
<td>635.</td>
</tr>
</tbody>
</table>

# 1,694 more rows
```r
gapminder |> 
ggplot(mapping = aes(x = gdpPerCap, 
y = lifeExp, 
size = pop, 
color = continent))
```
```r
gapminder |> 
ggplot(mapping = aes(x = gdpPercap, 
                      y = lifeExp, 
                      size = pop, 
                      color = continent)) + 
geom_point()
```
ggpminder |> 
  ggplot(mapping = aes(x = gdpPerCap, 
                       y = lifeExp, 
                       size = pop, 
                       color = continent)) +
  geom_point() +
  geom_smooth(method = "lm")
ggminder $>$

ggplot(mapping = aes(x = gdpPercap,
                   y = lifeExp,
                   size = pop,
                   color = continent)) +

geom_point() +
geom_smooth(method = "lm") +
scale_x_log10()
```r
# Economic Growth and Life Expectancy

ggplot(mapping = aes(x = gdpPerCap, y = lifeExp, size = pop, color = continent)) +
  geom_point() +
  geom_smooth(method = "lm") +
  scale_x_log10() +
  labs(x = "GDP Per Capita (log scale)", y = "Life Expectancy in Years", size = "Population", color = "Continent",
       title = "Economic Growth and Life Expectancy", subtitle = "Data points are country-years", caption = "Source: Gapminder.")
```

Source: Gapminder.
Nearly there

Economic Growth and Life Expectancy
Data points are country-years

Source: Gapminder.
Simplify to one line

gapminder |> ggplot(mapping = aes(x = gdpPerCap, y = lifeExp)) + geom_point(mapping = aes(size = pop, color = continent)) + geom_smooth(method = "lm") + scale_x_log10() + labs(x = "GDP Per Capita (log scale)", y = "Life Expectancy in Years", size = "Population", color = "Continent", title = "Economic Growth and Life Expectancy", subtitle = "Data points are country-years", caption = "Source: Gapminder.")
Economic Growth and Life Expectancy
Data points are country-years

Life Expectancy in Years

Source: Gapminder.
Remove the lm SE band

ggplot(mapping = aes(x = gdpPerCap, y = lifeExp)) +
geom_point(mapping = aes(size = pop, color = continent)) +
geom_smooth(method = "lm", se = FALSE) +
scale_x_log10(labels = label_dollar()) +
scale_size(labels =
  label_number(scale_cut = cut_short_scale())) +
labs(x = "GDP Per Capita (log scale)", y = "Life Expectancy in Years", size = "Population", color = "Continent", title = "Economic Growth and Life Expectancy", subtitle = "Data points are country-years", caption = "Source: Gapminder.")

Economic Growth and Life Expectancy
Data points are country-years

- Africa
- Americas
- Asia
- Europe

- 250M
- 500M
- 750M
- 1,000

Source: Gapminder.
Use the **alpha** channel

```r
ggplot(mapping = aes(x = gdpPerCap, y = lifeExp)) +
geom_point(mapping = aes(size = pop, color = continent), alpha = 0.3) +
geom_smooth(method = "lm", se = FALSE) +
scale_x_log10(labels = label_dollar()) +
scale_size(labels = label_number(scale_cut = cut_short_scale())) +
guides(color = guide_legend(override.aes = list(alpha = 1))) +
labs(x = "GDP Per Capita (log scale)", y = "Life Expectancy in Years", size = "Population", color = "Continent", title = "Economic Growth and Life Expectancy", subtitle = "Data points are country-years", caption = "Source: Gapminder."
```

**Economic Growth and Life Expectancy**

Data points are country-years

- **Africa**
- **Americas**
- **Asia**
- **Europe**

![Graph showing economic growth and life expectancy](source)
Example: Faceting
# A tibble: 1,784 × 6

  country continent year  lifeExp  pop  gdpPerCap
  <fct>    <fct>   <int>    <dbl>  <int>     <dbl>
1 Afghanistan Asia    1952    28.8      842533    779.
2 Afghanistan Asia    1957    30.3     9240934    821.
3 Afghanistan Asia    1962    32.0    10267083    853.
4 Afghanistan Asia    1967    34.0    11557966    836.
5 Afghanistan Asia    1972    36.1    13079460    740.
6 Afghanistan Asia    1977    38.4    14880272    786.
7 Afghanistan Asia    1982    39.9    12881816    978.
8 Afghanistan Asia    1987    40.8    13867957    852.
9 Afghanistan Asia    1992    41.7    16317921    649.
10 Afghanistan Asia   1997    41.8    22227415    635.
# : 1,694 more rows
Facets are a powerful tool

gapminder >
filter(continent != "Oceania")

# A tibble: 1,680 × 6
  country         continent year  lifeExp     pop  gdpPercap
  <fct>            <fct>    <int>     <dbl>   <int>    <dbl>
1 Afghanistan     Asia       1952     28.8  8425333      779.  
2 Afghanistan     Asia       1957     30.3  9240934      821.  
3 Afghanistan     Asia       1962     32.0 10267083      853.  
4 Afghanistan     Asia       1967     34.0 11237966      856.  
5 Afghanistan     Asia       1972     36.1 13079460      740.  
6 Afghanistan     Asia       1977     38.4 14880272      786.  
7 Afghanistan     Asia       1982     39.9 12881816      978.  
8 Afghanistan     Asia       1987     40.8 13867957      852.  
9 Afghanistan     Asia       1992     41.7 16317921      649.  
10 Afghanistan    Asia       1997     41.8 22227415      635.  
# : 1,670 more rows
Facets are a powerful tool

gapminder >>
filter(continent != "Oceania") >>
ggplot(mapping = aes(x = year,
                    y = gdpPerCap))

[Chart showing GDP per capita over time, with years from 1950 to 2000 on the x-axis and GDP per capita on the y-axis.]
Facets are a powerful tool

gapminder |> filter(continent != "Oceania") |> ggplot(mapping = aes(x = year, y = gdpPercap)) + geom_line(mapping = aes(group = country), color = "gray70")
Facets are a powerful tool

gapminder |> filter(continent != "Oceania") |> ggplot(mapping = aes(x = year, y = gdpPercap)) + geom_line(mapping = aes(group = country), color = "gray70") + geom_smooth(mapping = aes(group = continent), method = "loess")
Facets are a powerful tool

```r
# Load data
# (Assuming gapminder data is loaded)

# Filter out Oceania continent
continent <- filter(gapminder, continent != "Oceania")

# Create a ggplot
ggplot(continent, aes(x = year, y = gdpPerCap)) +
  geom_line(aes(group = country), color = "gray70") +
  geom_smooth(aes(group = continent), method = "loess") +
  scale_y_log10(labels = label_dollar)
```
Facets are a powerful tool

gapminder >
  filter(continent != "Oceania") 
  ggplot(mapping = aes(x = year, 
                      y = gdpPerCap)) + 
  geom_line(mapping = aes(group = country), 
            color = "gray70") + 
  geom_smooth(mapping = aes(group = continent), 
              method = "loess") + 
  scale_y_log10(labels = label_dollar()) + 
  facet_wrap(~ continent)
Facets are a powerful tool

gapminder |> 
  filter(continent != "Oceania") |> 
  ggplot(mapping = aes(x = year, 
                      y = gdpPerCap)) + 
  geom_line(mapping = aes(group = country), 
            color = "gray70") + 
  geom_smooth(mapping = aes(group = continent), 
              method = "loess") + 
  scale_y_log10(labels = label_dollar()) + 
  facet_wrap(~ continent) + 
  labs(x = "Year", 
       y = "GDP per capita", 
       title = "Per Capita GDP over time, by Continent", 
       subtitle = "Excludes Oceania. Gray lines are countries", 
       caption = "Source: Gapminder.")
Facets are a powerful tool

ggplot(mapping = aes(x = year,
    y = gdpPerCap)) +
  geom_point() +
  geom_smooth(mapping = aes(group = country),
    color = "gray80") +
  scale_y_log10(labels = label_dollar()) +
  facet_wrap(~ continent) +
  labs(x = "Year",
    y = "GDP per capita",
    title = "Per Capita GDP over time, by Continent",
    subtitle = "Excludes Oceania. Gray lines are countries",
    caption = "Source: Gapminder.")

Source: Gapminder:
Per Capita GDP over time, by Continent
Excludes Oceania. Gray lines are countries

Source: Gapminder.
Facets are often better than Guides
Consider a basic crosstab

```
rel_by_region ▸ select(-n) ▸
pivot_wider(names_from = religion, values_from = pct) ▸
janitor::adorn_totals(where = "col")
```

<table>
<thead>
<tr>
<th>bigregion</th>
<th>Protestant</th>
<th>Catholic</th>
<th>Jewish</th>
<th>None</th>
<th>Other (Missing)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northeast</td>
<td>32</td>
<td>33</td>
<td>5.53</td>
<td>23</td>
<td>5.7</td>
<td>0.20</td>
</tr>
<tr>
<td>Midwest</td>
<td>47</td>
<td>25</td>
<td>0.43</td>
<td>23</td>
<td>4.7</td>
<td>0.72</td>
</tr>
<tr>
<td>South</td>
<td>62</td>
<td>15</td>
<td>1.05</td>
<td>16</td>
<td>4.8</td>
<td>1.05</td>
</tr>
<tr>
<td>West</td>
<td>38</td>
<td>25</td>
<td>1.58</td>
<td>28</td>
<td>7.6</td>
<td>0.16</td>
</tr>
</tbody>
</table>
We might write ...

```
rel_by_region <-
ggplot(mapping = aes(x = bigregion, y = pct, fill = religion)) +
geom_col(position = "dodge") + labs(x = "Region", y = "Percent", fill = "Religion")
```

Is this an effective graph? Not really!
Try **faceting instead**

```r
rel_by_region <- ggplot(mapping = aes(x = pct, y = reorder(religion, -pct), fill = religion)) + geom_col() + guides(fill = "none") + facet_wrap(~ bigregion, nrow = 1) + labs(x = "Percent", y = NULL)
```

Facets **reduce the number of guides and legends** the viewer needs to consult. Notice how this graph no longer requires the bars to be in color.

**Reorder your categories** by the value of what they summarize. This is true for all kinds of summary graphs involving categorical variables, not just facets.

**Try putting the categories on the y-axis.** This is a very useful trick. It makes graphs compact and table-like, and avoids x-axis labels being in the wrong orientation, or you having to figure out how to put them at an angle.
Graphing in practice
A scatterplot
Trendlines
Workhorses

Point-and-range across categories
Manhattan Building Heights by Year of Construction
Show Ponies

Age Distribution of Congressional Representatives, 1945-2019

Trend line is mean age; bands are 25th and 75th percentiles of the range.
Youngest and oldest percentiles are named instead of being shown by points.

Congressional comparison
Show Ponies

Several plots at once
Show Ponies

Vaccination Exemption Rates in California Kindergartens
Percent of Kindergarteners with a Personal Belief Exemption, by Type and Size of School.
Show Ponies

Dogs of New York

Based on data from New York City's Dog Licensing System, these maps show the relative prevalence of the twenty five most common breeds of dog, by zip code.

Faceted maps
Unicorns ...
Swedish mortality
... or monsters

Network chords
Summary
Your toolkit can’t make you honest, or correct
But you should still look at (and show) your data
Consider who your audience is, including when it’s yourself.
Present findings in substantive terms
Show degrees of confidence or uncertainty
Show the data when you can
But remember, these points apply just as well to presenting data in *any* format: tables, models, text, whatever. **Graphs are not special in this respect!**
Learn More

https://socviz.co
Full draft of *Data Visualization: A Practical Introduction* (Princeton 2019)

https://kieranhealy.org
Examples, R data packages (including for the GSS), code