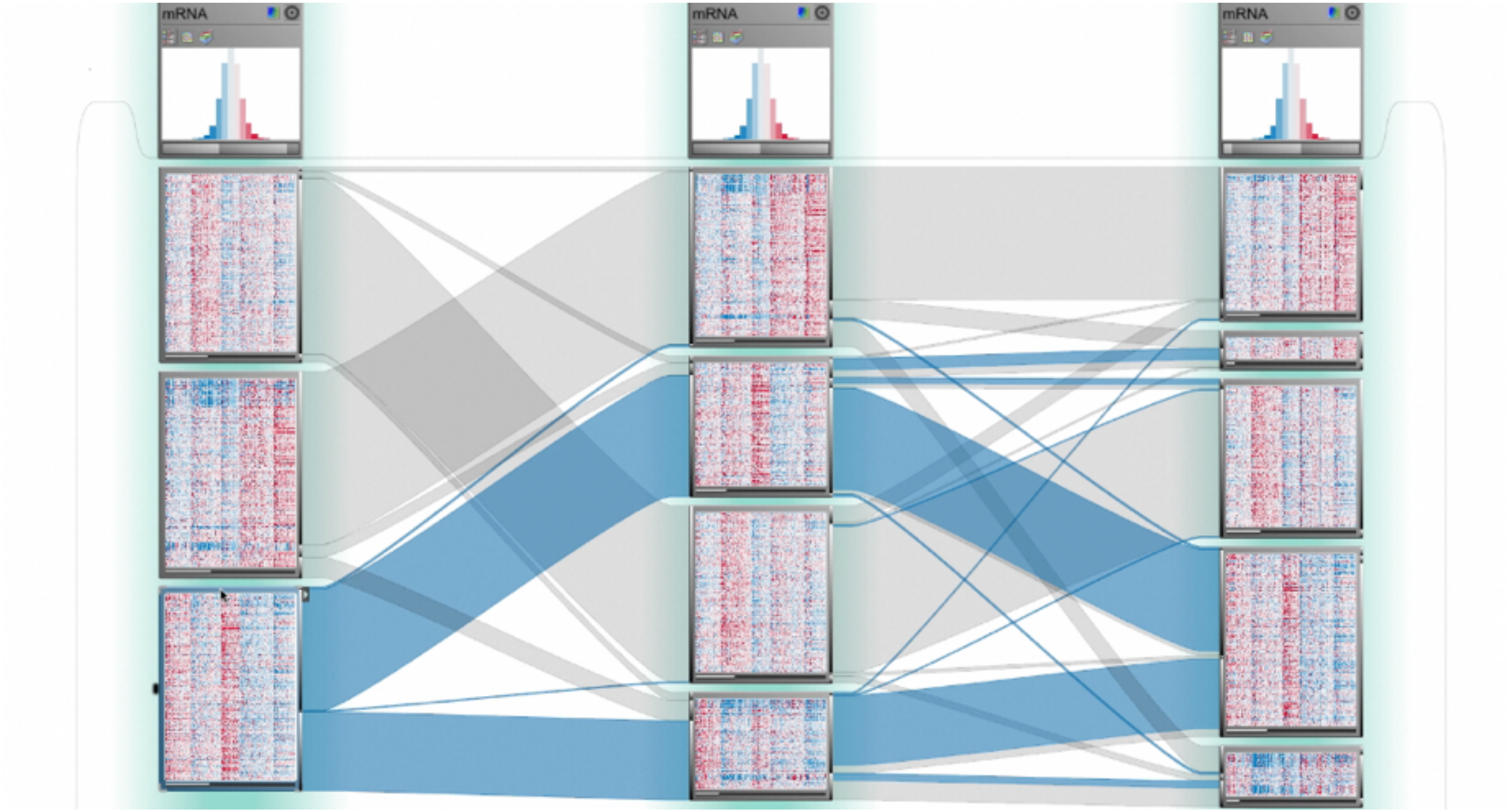
Combining information visualization theory and the grammar of graphics to do and teach modern data analysis

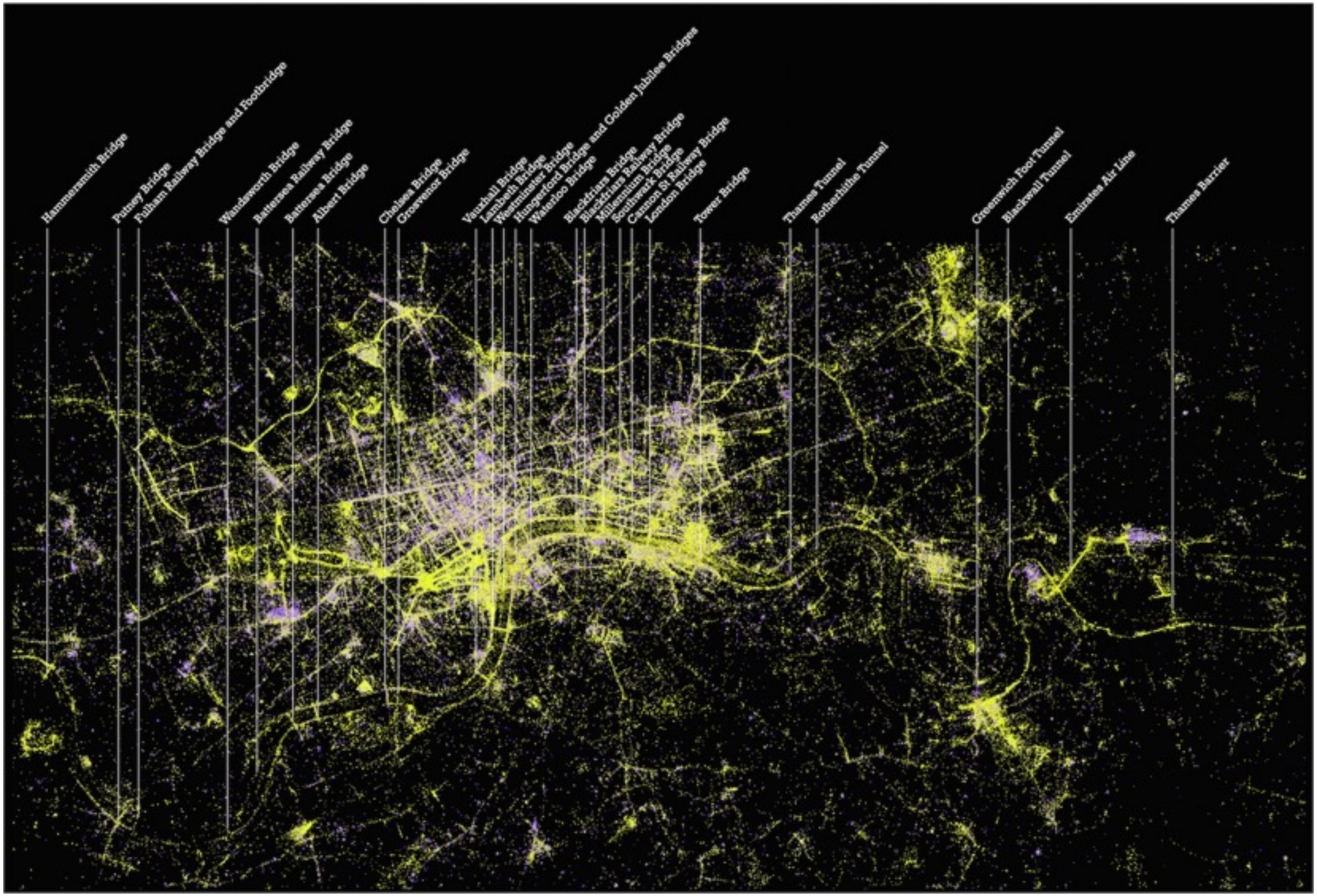
6th February 2018

Roger Beecham
www.roger-beecham.com

Justin Matejka, Tovi Grossman, George Mitzmaurice



Marc Streit, Alexander Lex, Samuel Gratzl, Hanspeter Pfister, Nils Gehlenbourg



Alex Kachkaev, Jo Wood

p5*

p5*

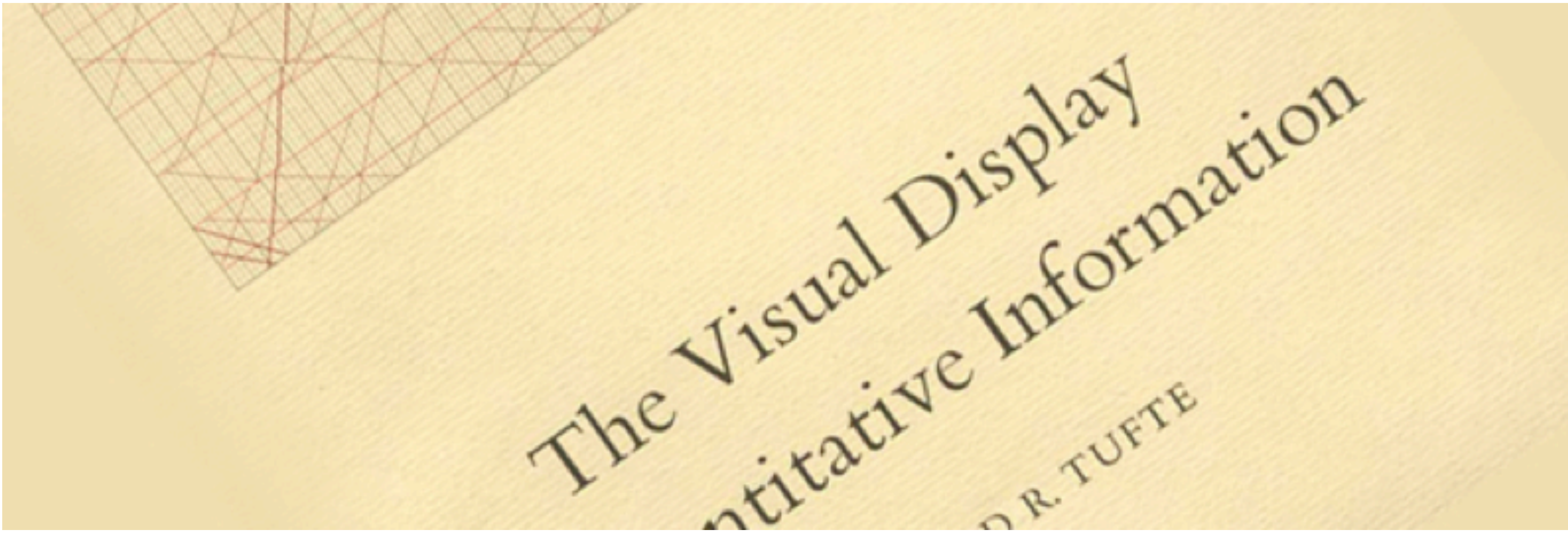
Vega-Lite

Vega

ggplot2

Vega

p5*



Data graphics visually display measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading, and color.

Tufte, 1983

Effective data graphics should

1. Show the data

2. Induce the viewer to think about the substance of the data rather than about graphic design

3. Avoid distorting what the data have to say

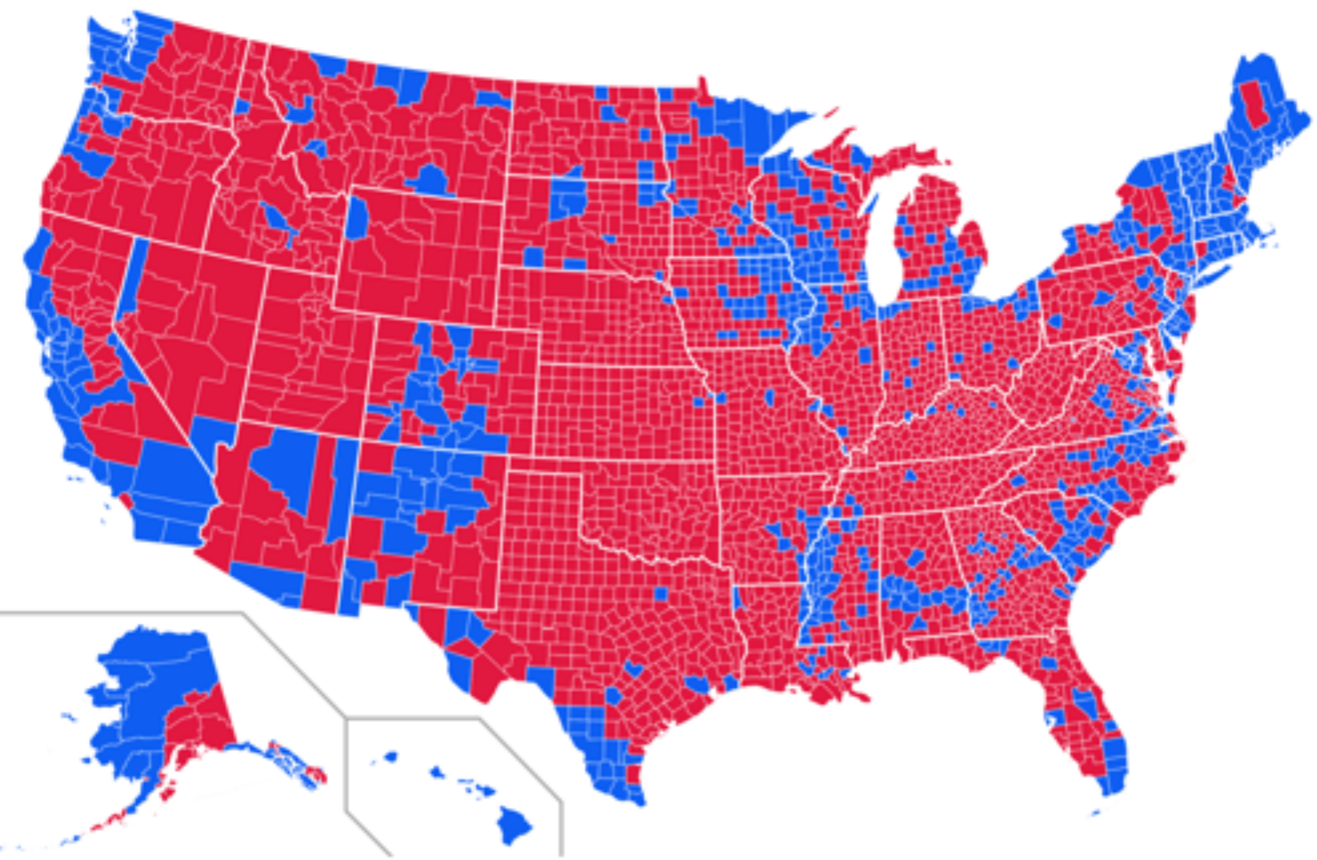
4. Present many numbers in a small space

5. Make large data sets coherent

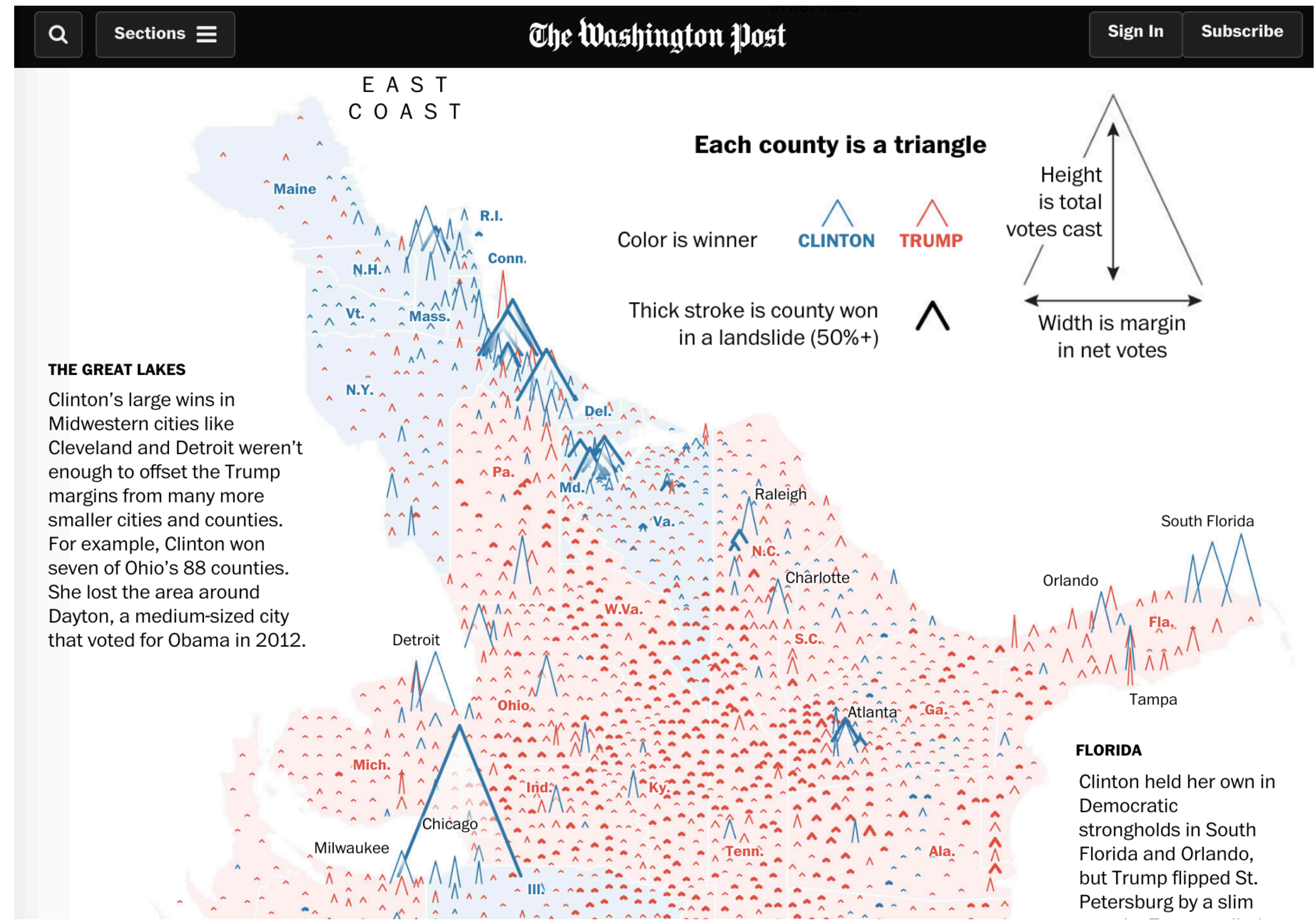
6. Encourage the eye to compare different pieces of data

7. Reveal the data at several levels of detail from a broad overview to a fine structure

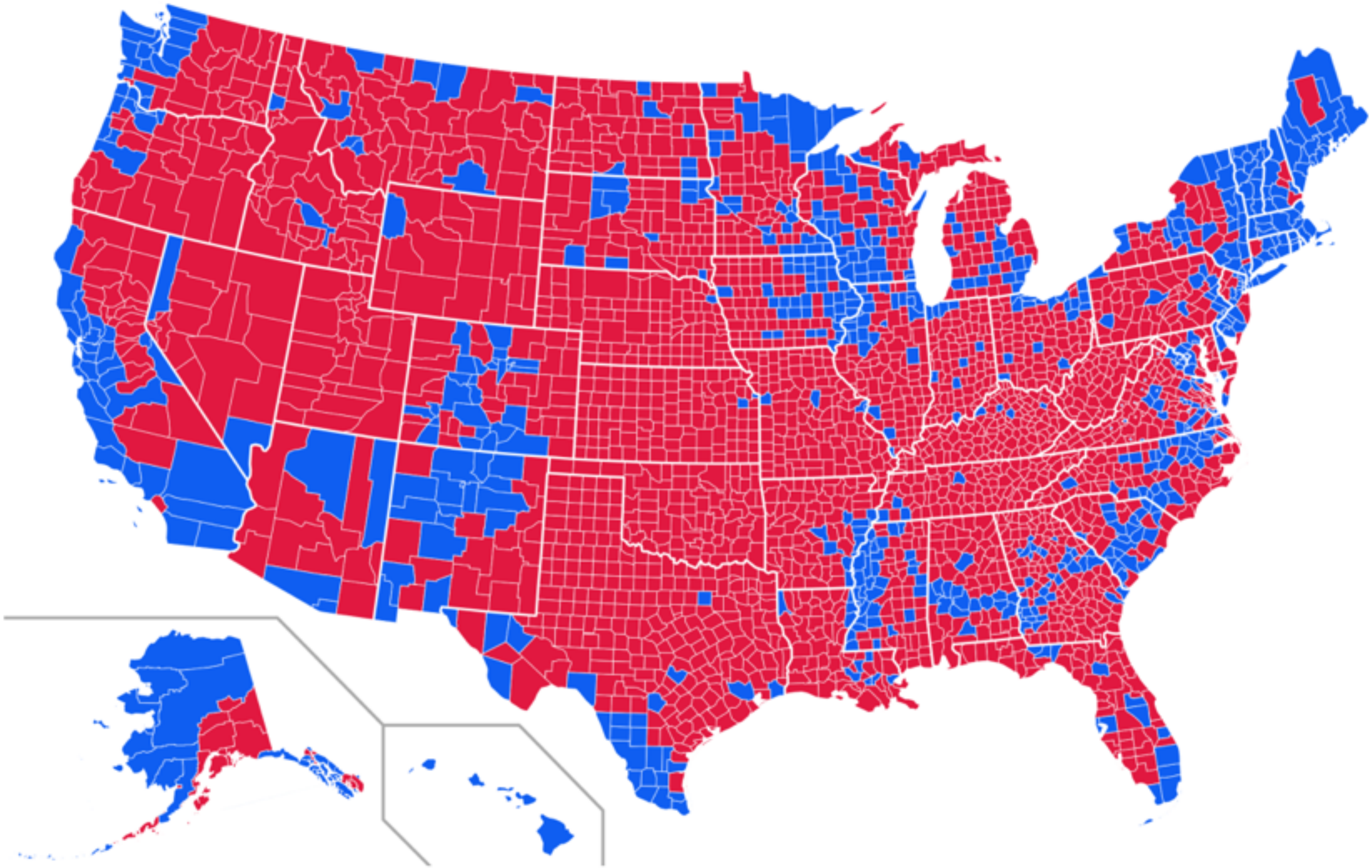
Tufte (1983: 13)



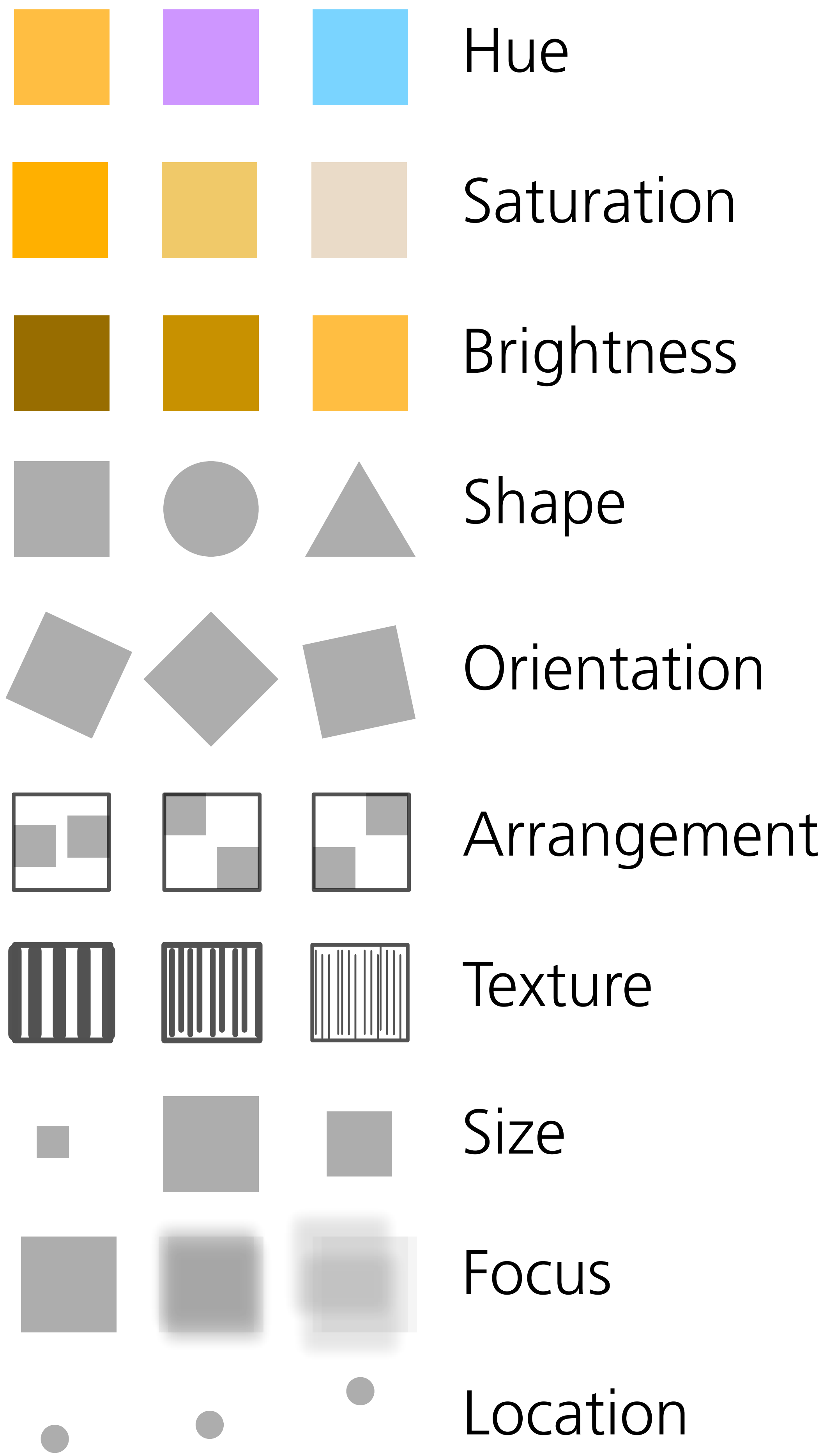
Natalie Schmidt, on Medium



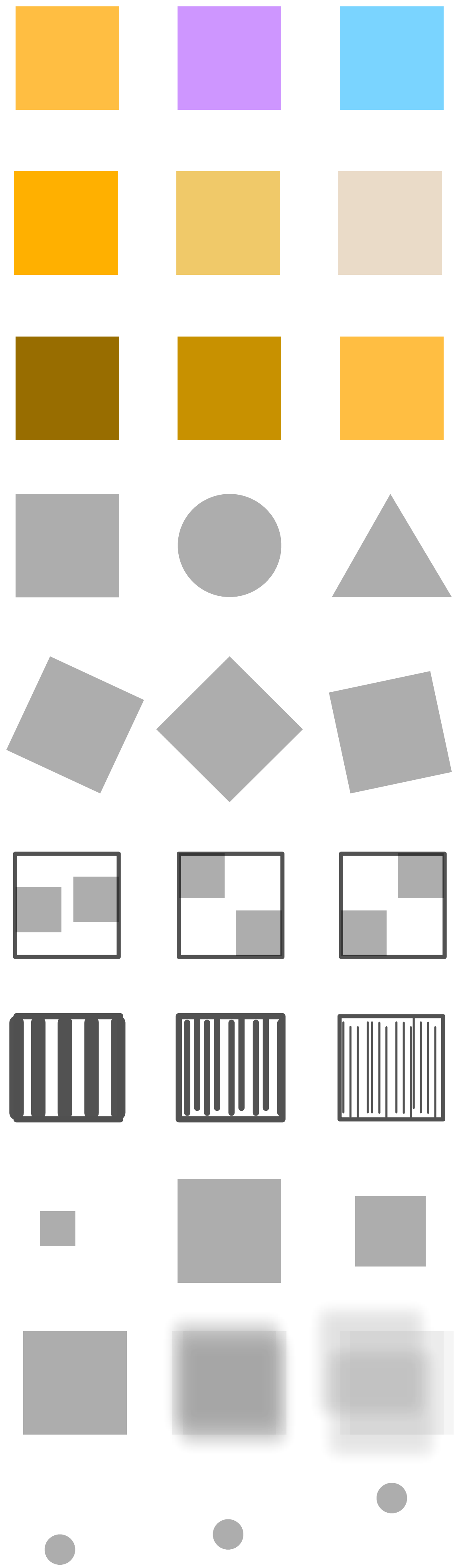
Lazaro Gamio and Dan Keating, Washington Post



Natalie Schmidt, on Medium


Hue
Saturation
Brightness
Shape
Orientation
Arrangement
Texture
Size
Focus
Location

Bertin, 1983

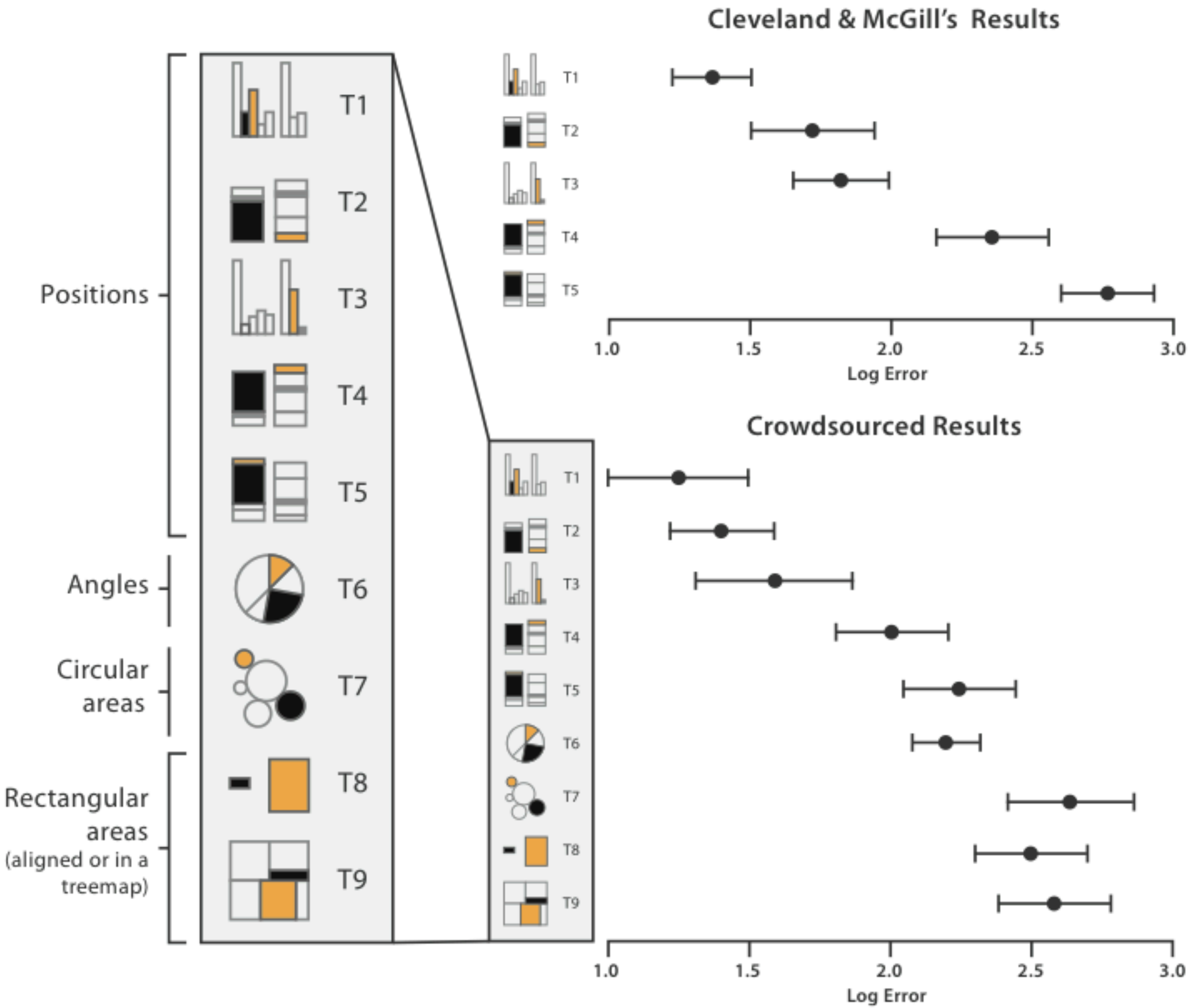


Selective: Change in this visual variable alone is enough to allow a symbol to be selected from a group.

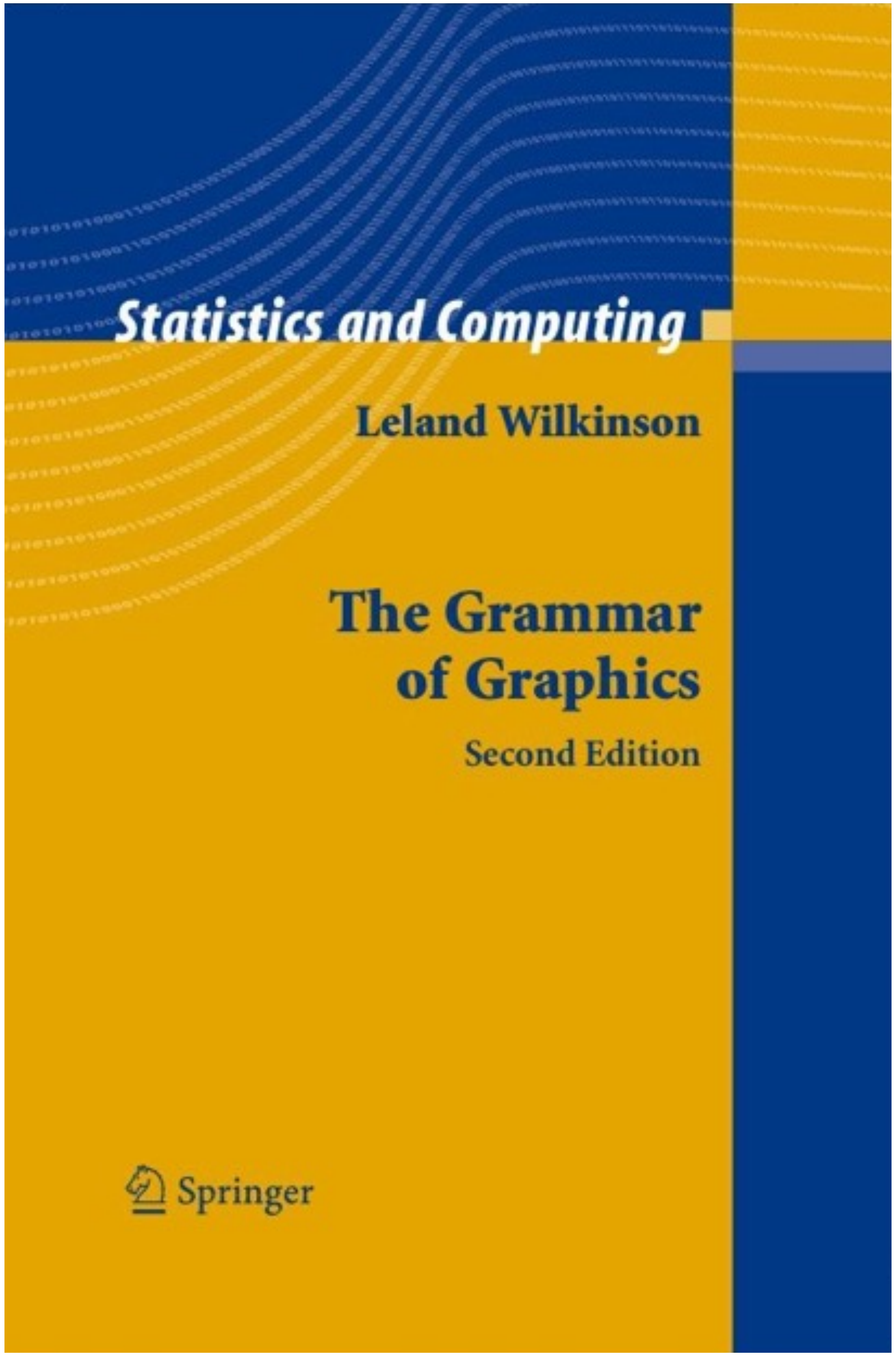
Associative: Symbols that are alike in all other ways can be grouped according to change in this visual variable.

Quantitative: A numerical reading is obtainable from changes in this visual variable.

Order: Changes in this variable perceived as ordered



Heer & Bostock 2010



Data

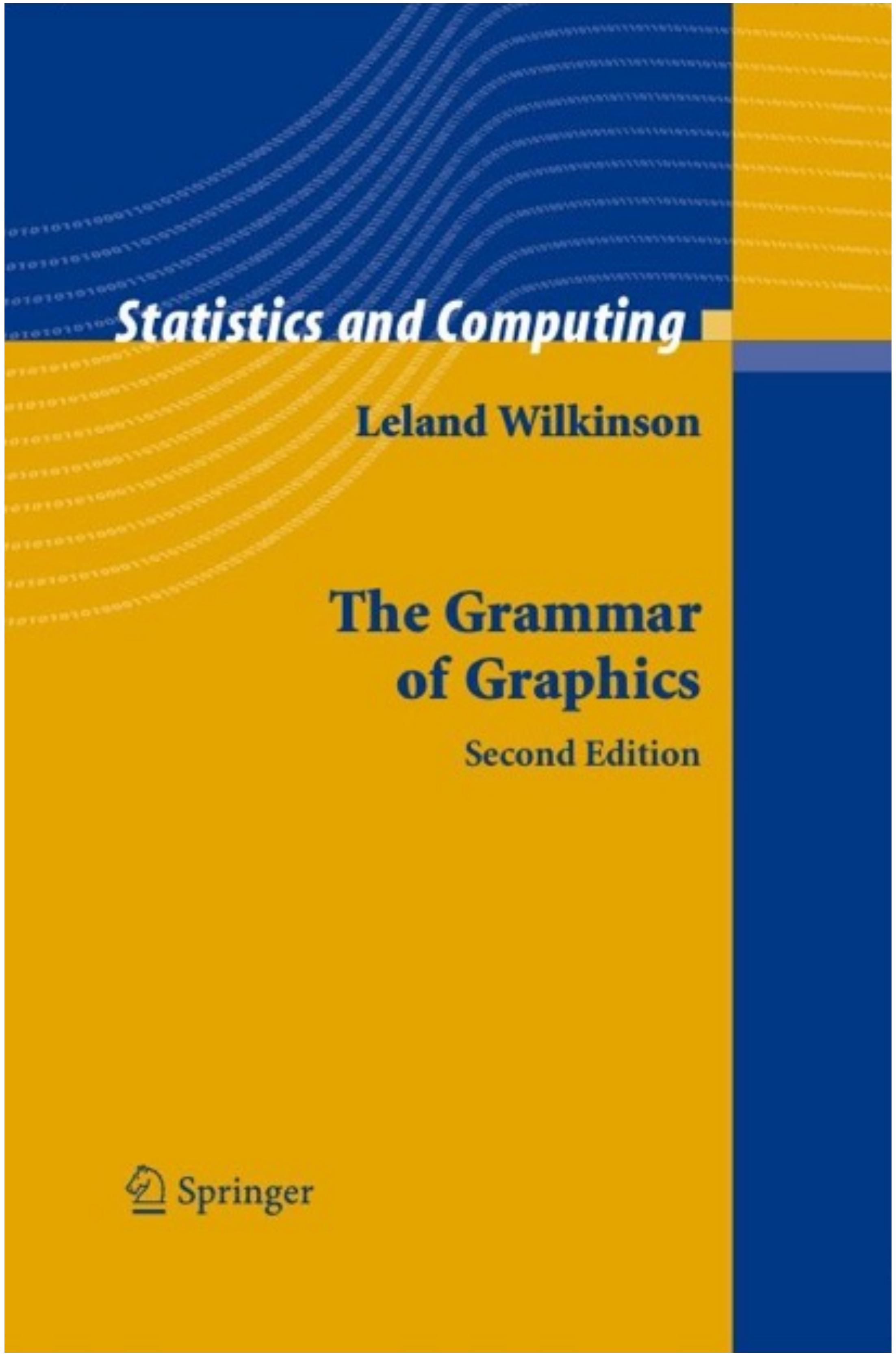
Transformation

Element

Scale

Guide

Coord

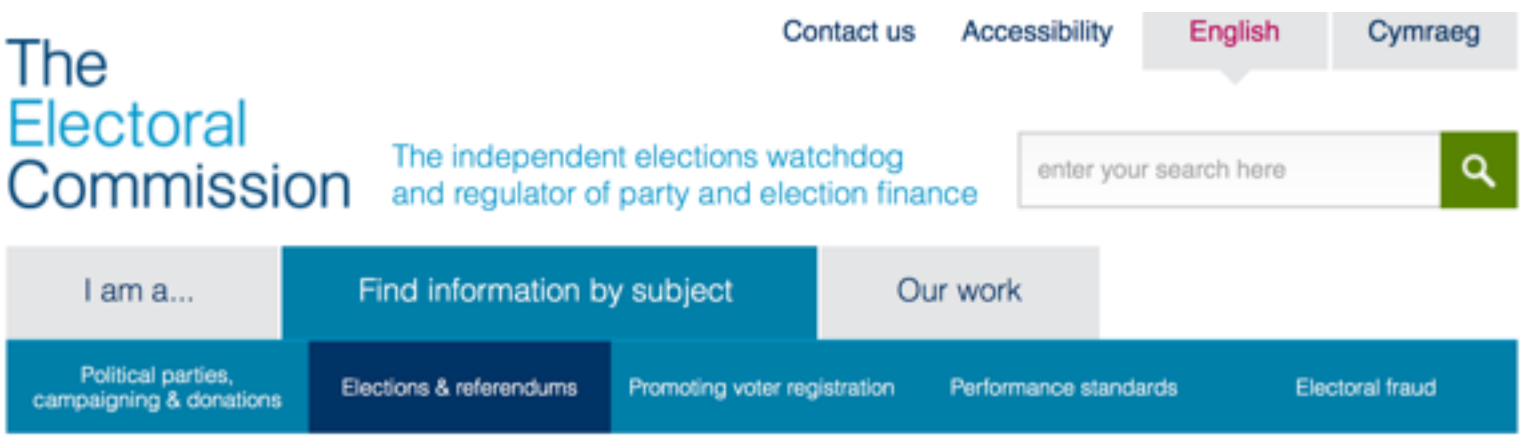


ggplot2

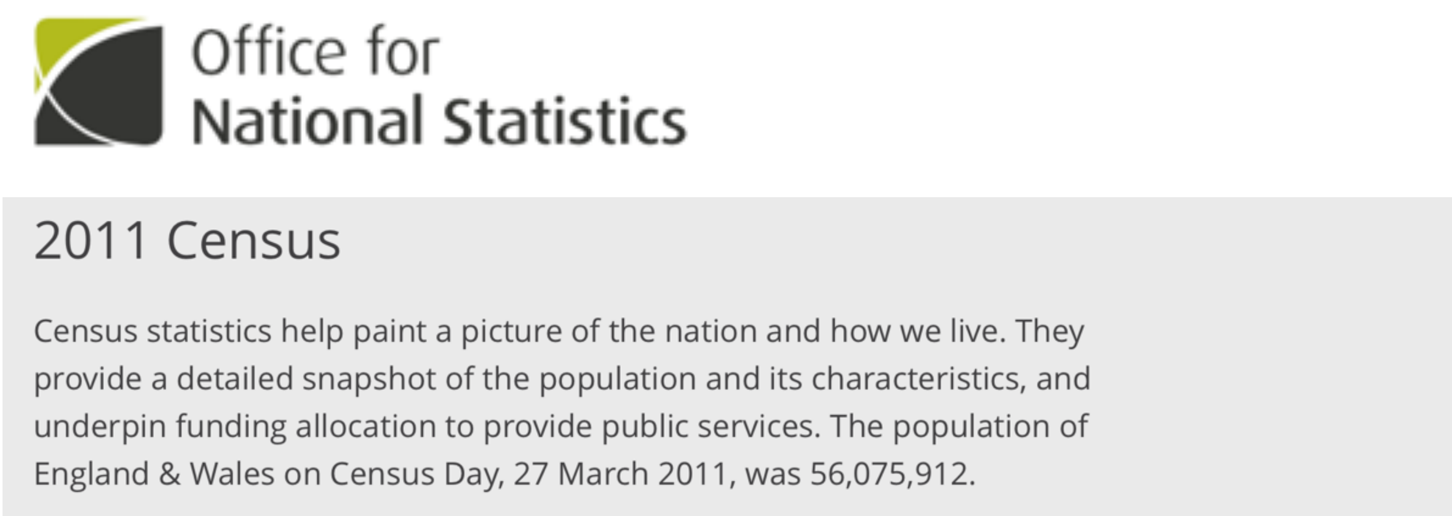
Vega-Lite

Data	variables you want to represent
Aesthetics	mapping of data to visual channels
Geom	shapes to represent data (point, line, bar)
Facets	split on a (nominal/ordinal) variable to generate small multiples
Statistics	aggregates using statistical models
Coordinates	plotting space you are using
Themes	non-data ink: design with a particular visual fonts, colours and other design elements.

+ informed defaults



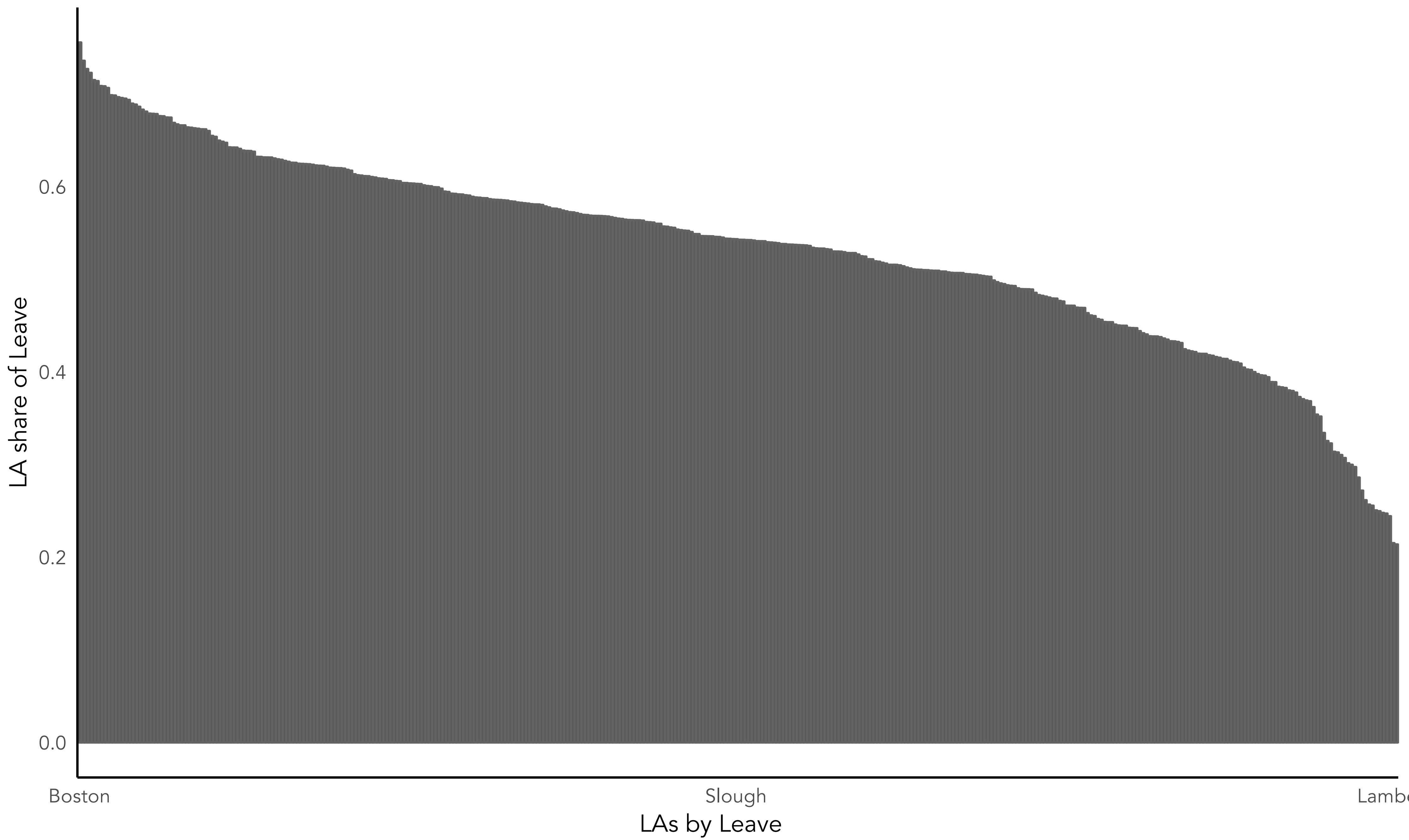
Brexit data: share of leave vote by Local Authority



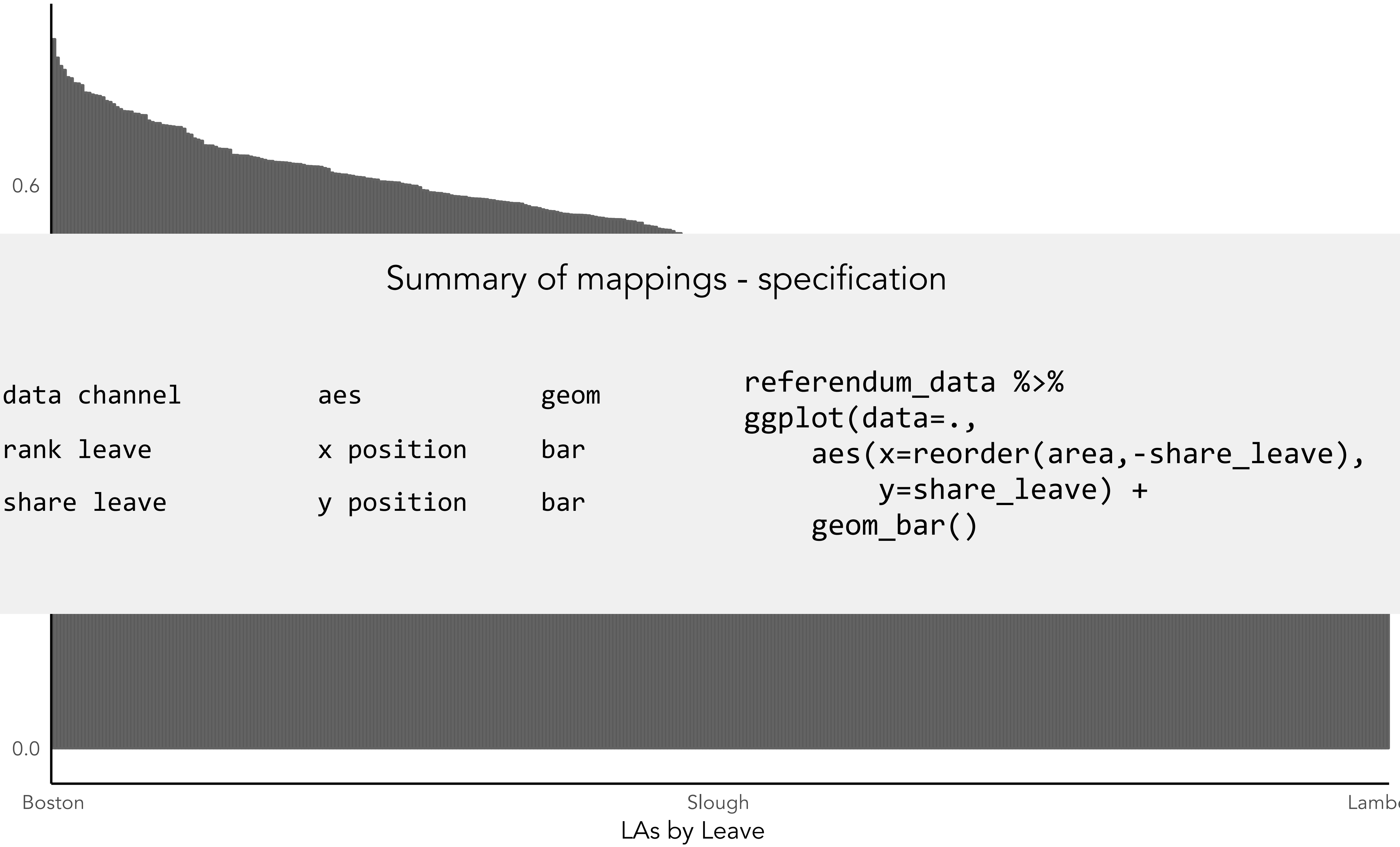
Demographics data: skills levels, occupation and diversity by Local Authority

LAs ordered by share of Leave

LAs ordered by share of Leave



LAs ordered by share of Leave

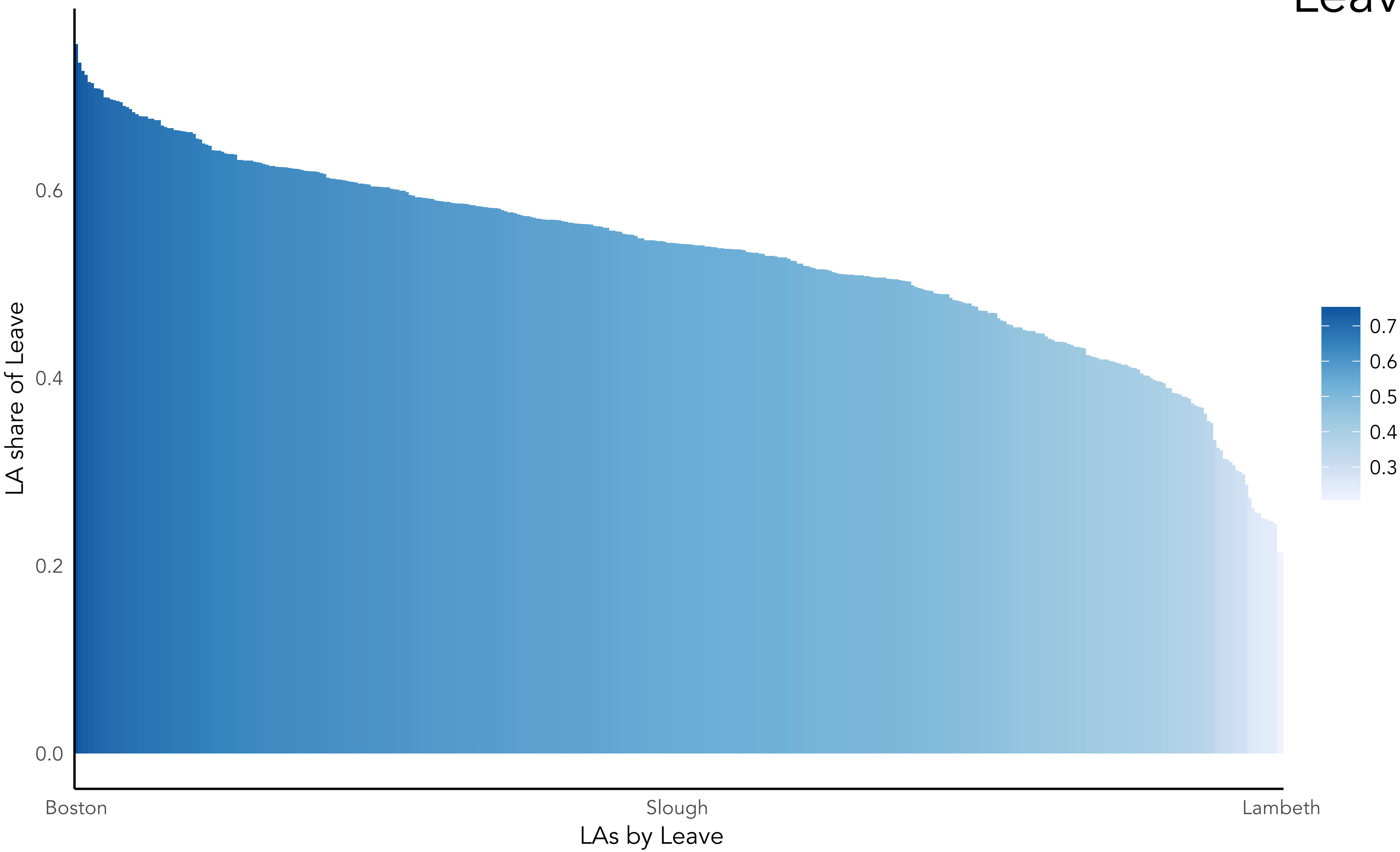


Summary of mappings - specification

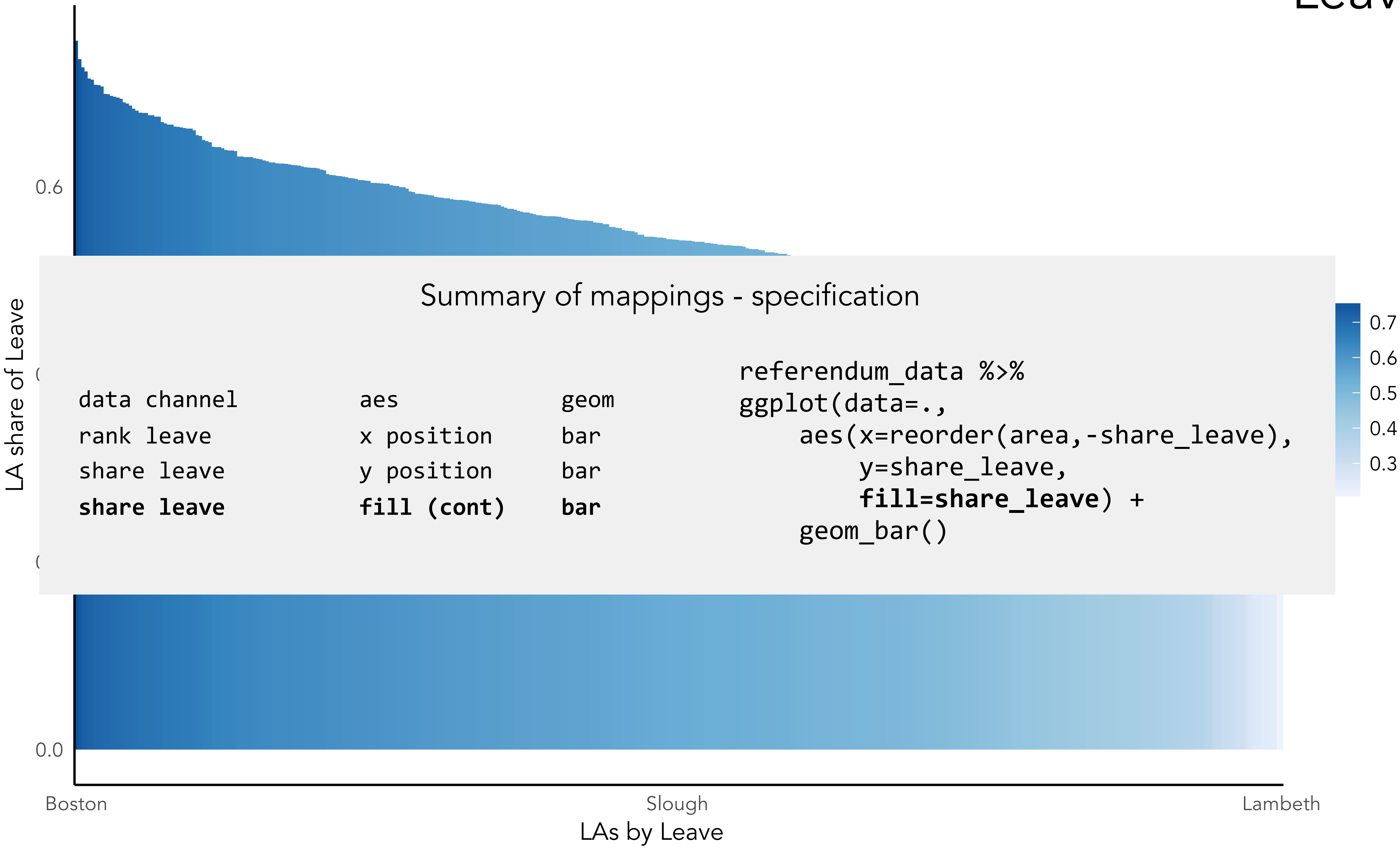
data channel	aes	geom
rank leave	x position	bar
share leave	y position	bar

```
referendum_data %>%
ggplot(data=.,
    aes(x=reorder(area,-share_leave),
        y=share_leave) +
    geom_bar()
```

LAs ordered by share of Leave



LAs ordered by share of Leave



Summary of mappings - specification

data channel	aes	geom
rank leave	x position	bar
share leave	y position	bar
share leave	**fill (cont)**	**bar**

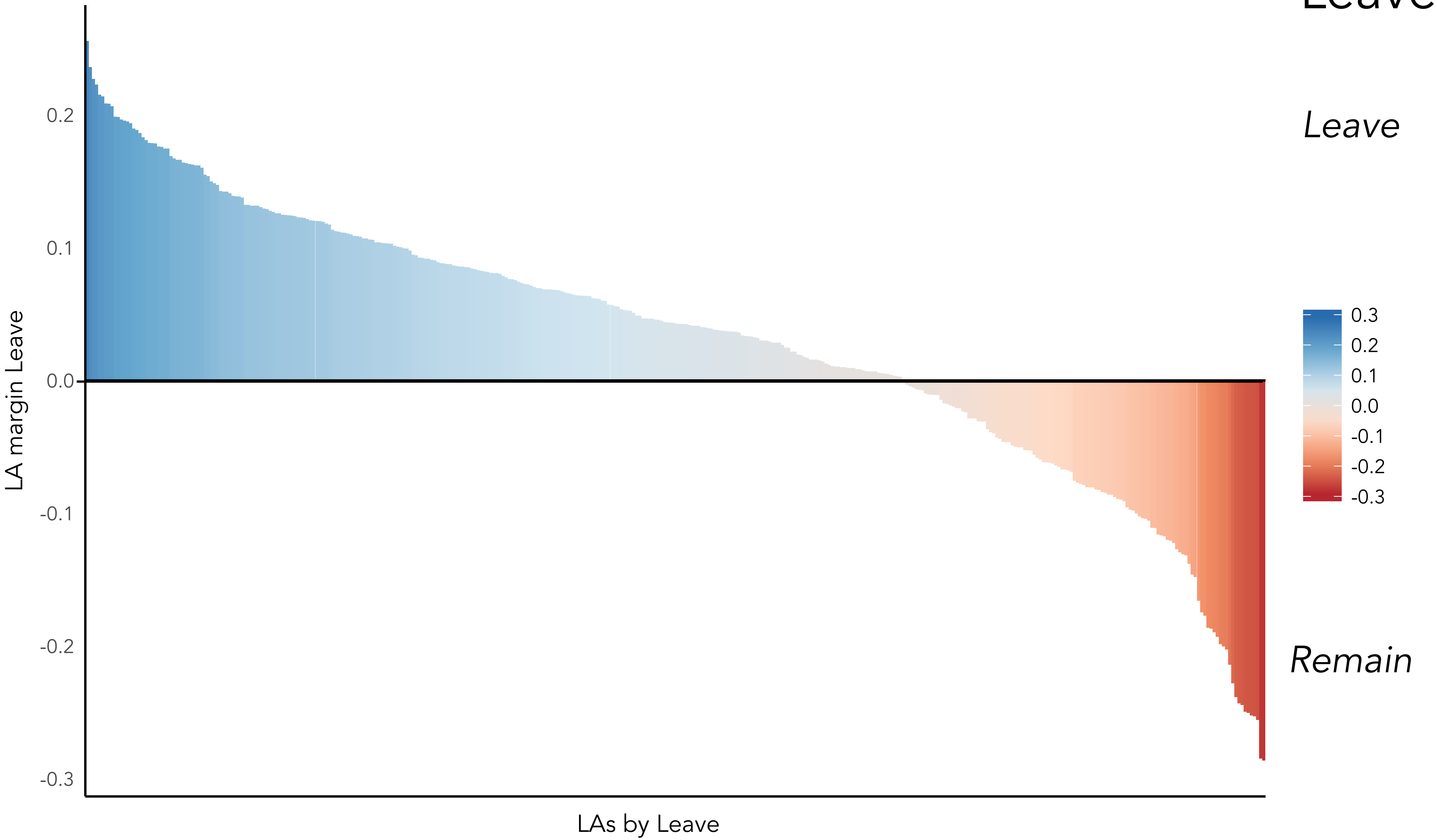
```
referendum_data %>%
ggplot(data=.,
    aes(x=reorder(area,-share_leave),
        y=share_leave,
        fill=share_leave) +
    geom_bar()
```

LAs ordered by share of Leave

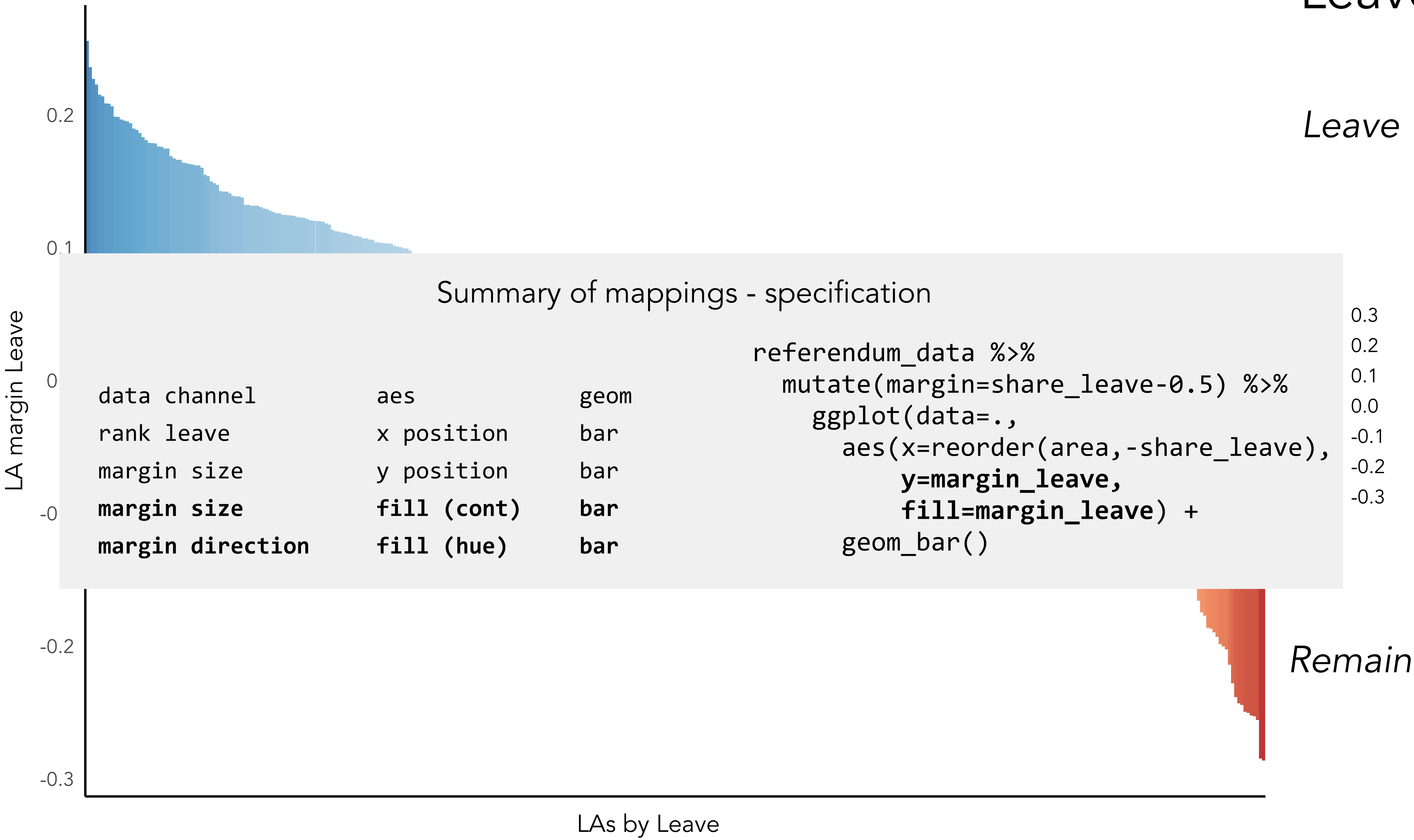
Leave

Remain

LAs ordered by share of Leave



LAs ordered by share of Leave

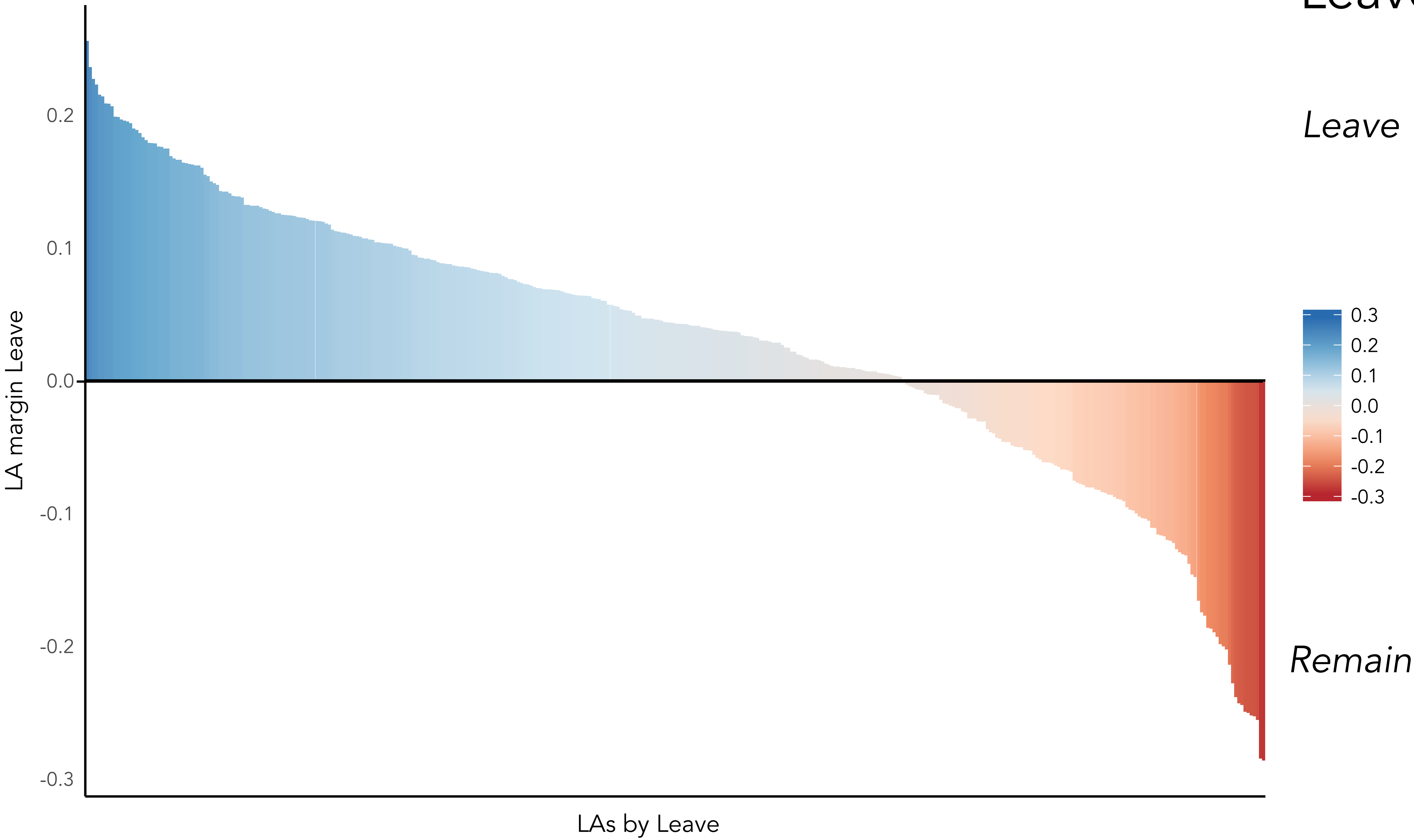


Summary of mappings - specification

data channel	aes	geom
rank leave	x position	bar
margin size	y position	bar
margin size	**fill (cont)**	**bar**
margin direction	**fill (hue)**	**bar**

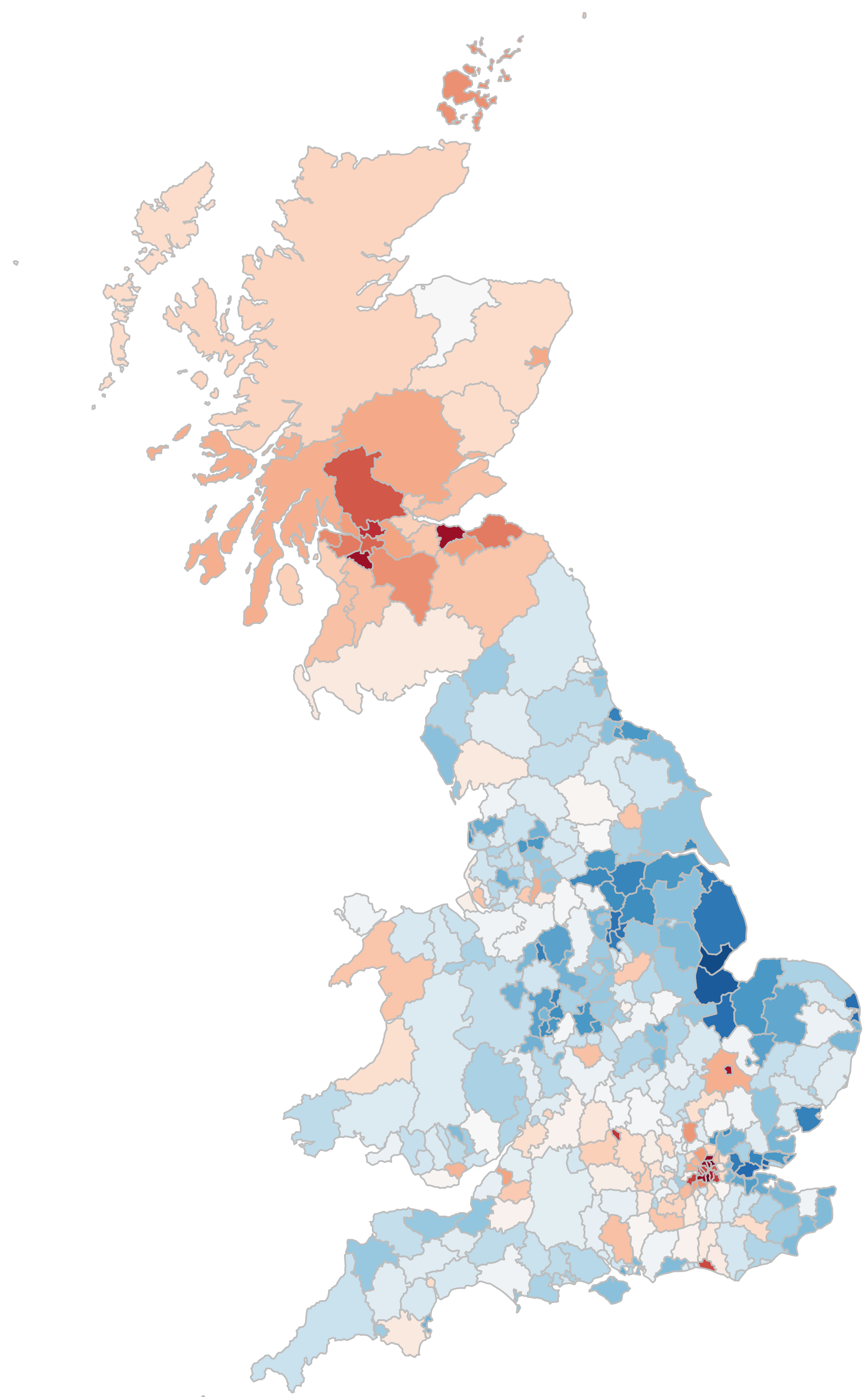
```
referendum_data %>%
  mutate(margin=share_leave-0.5) %>%
    ggplot(data=.,
      aes(x=reorder(area,-share_leave),
          y=margin_leave,
          fill=margin_leave) +
      geom_bar()
```

LAs ordered by share of Leave

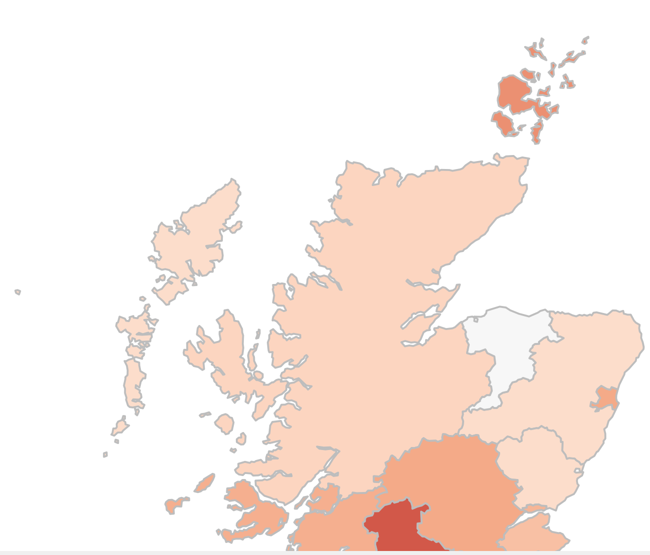


LAs ordered by geospatial position

LAs ordered by geospatial position



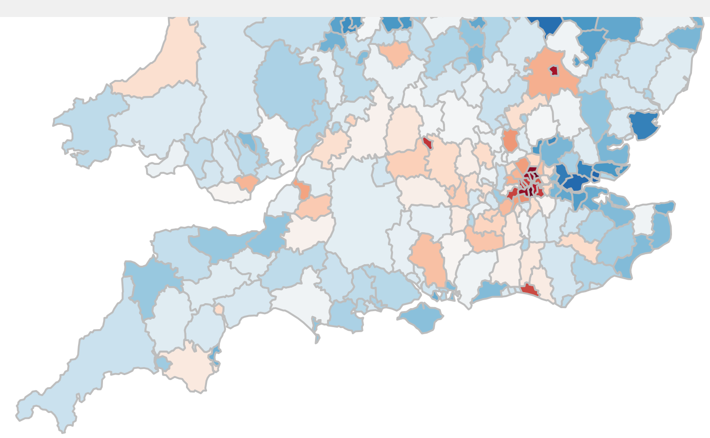
LAs ordered by geospatial position



Summary of mappings - specification

data channel	aes	geom
la position	**x,y position**	**poly**
margin size	fill (cont)	poly
margin direction	fill (hue)	poly

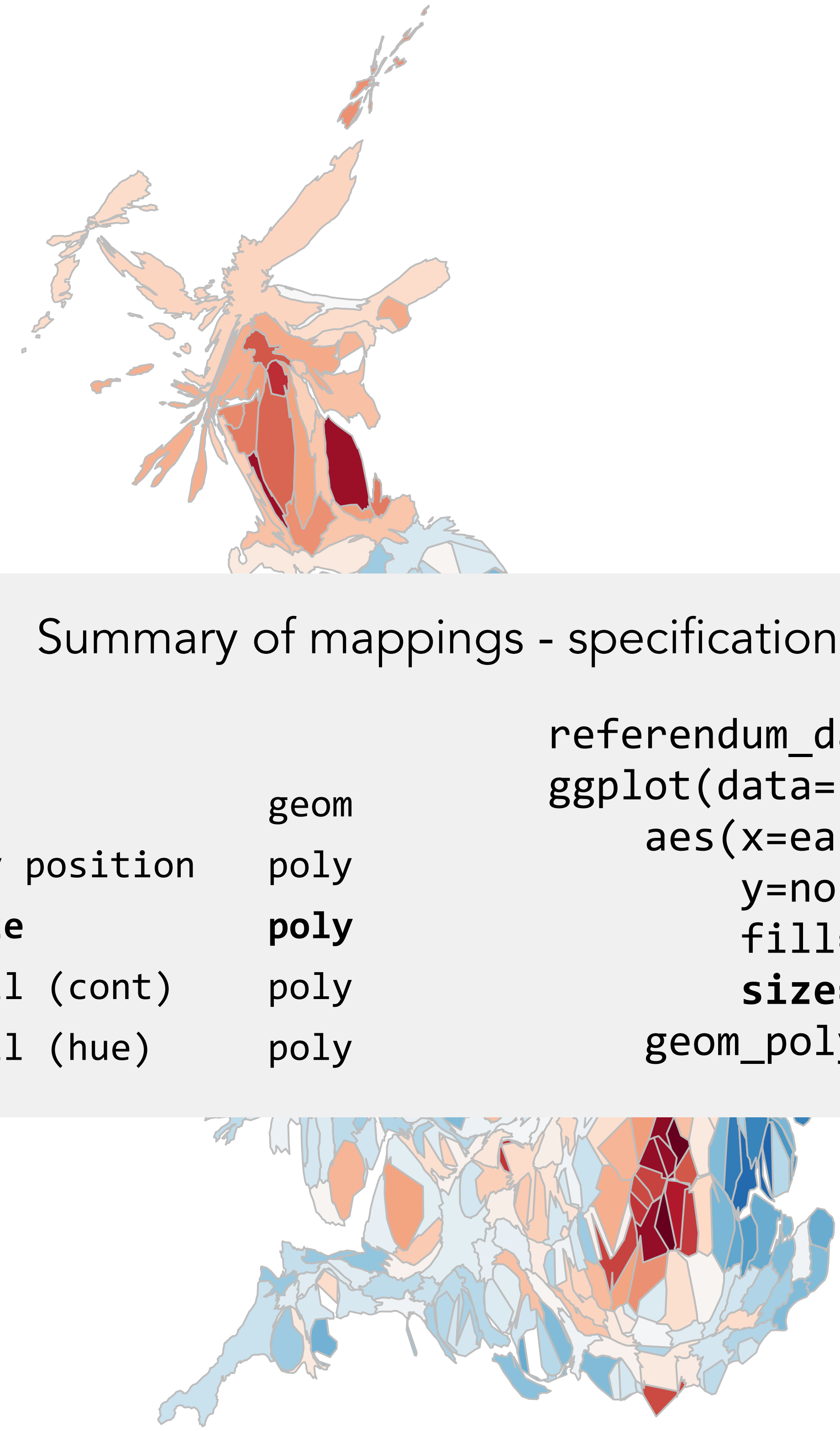
```
referendum_data %>%
ggplot(data=.,
    aes(x=easting,
        y=northing,
        fill=share_leave) +
    geom_polygon()
```



LAs ordered by geospatial position

LAs ordered by geospatial position

LAs ordered by geospatial position



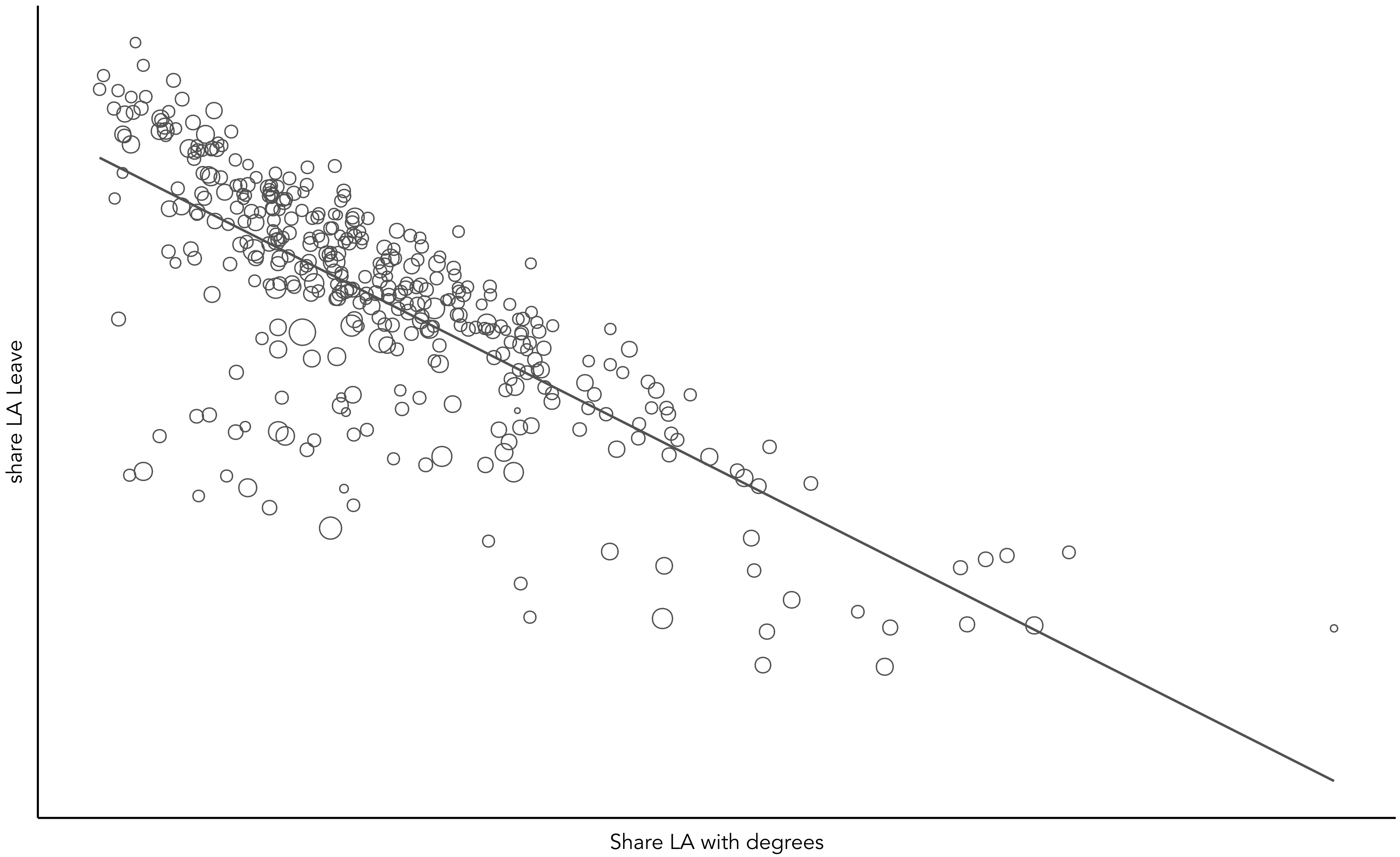
Summary of mappings - specification

data channel	aes	geom
la position	x,y position	poly
la area	**size**	**poly**
margin size	fill (cont)	poly
margin direction	fill (hue)	poly

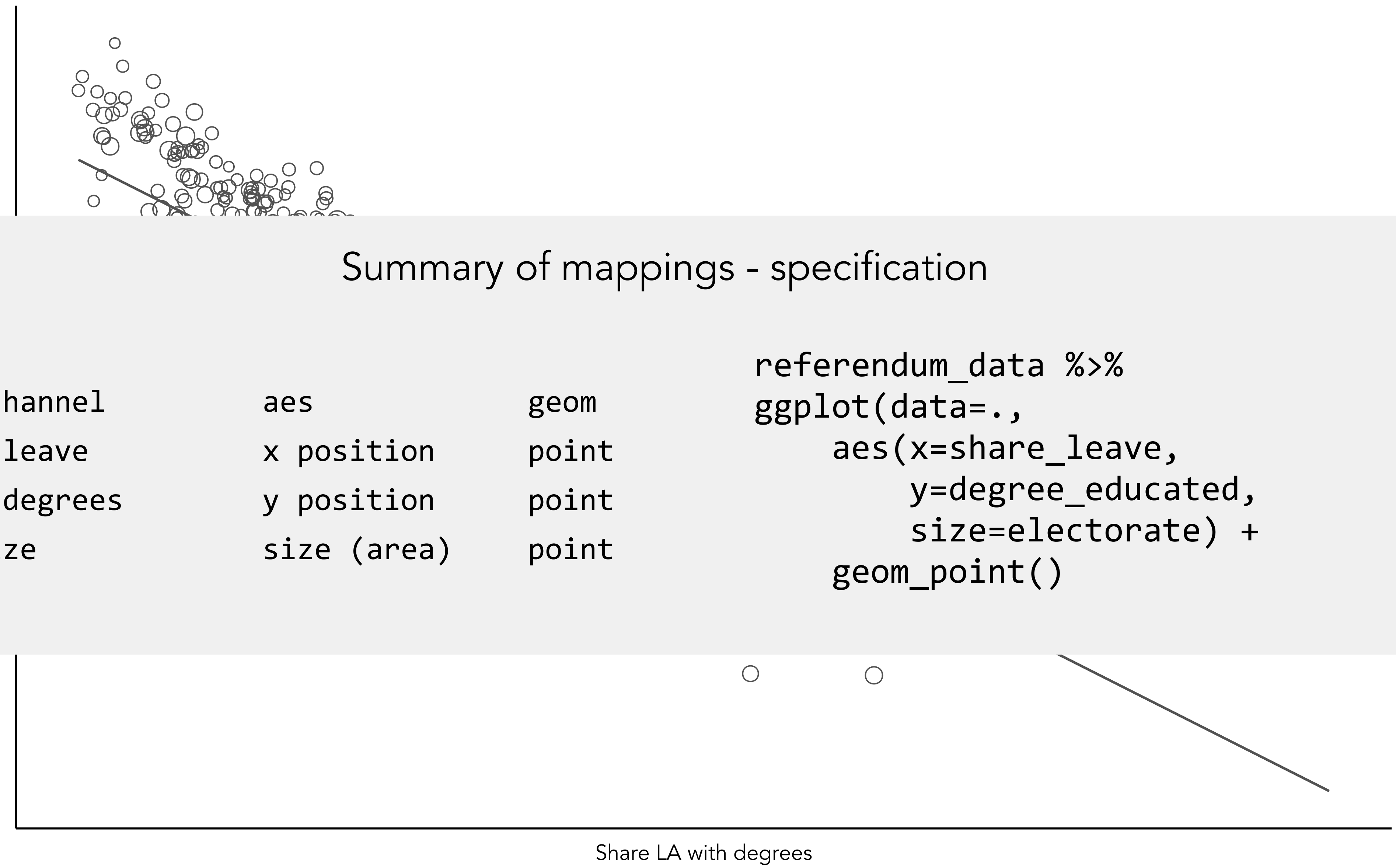
```
referendum_data %>%
ggplot(data=.,
    aes(x=easting,
        y=northing,
        fill=share_leave,
        size=area) +
    geom_polygon()
```

Leave vote by degree-level education

Leave vote by degree-level education



Leave vote by degree-level education

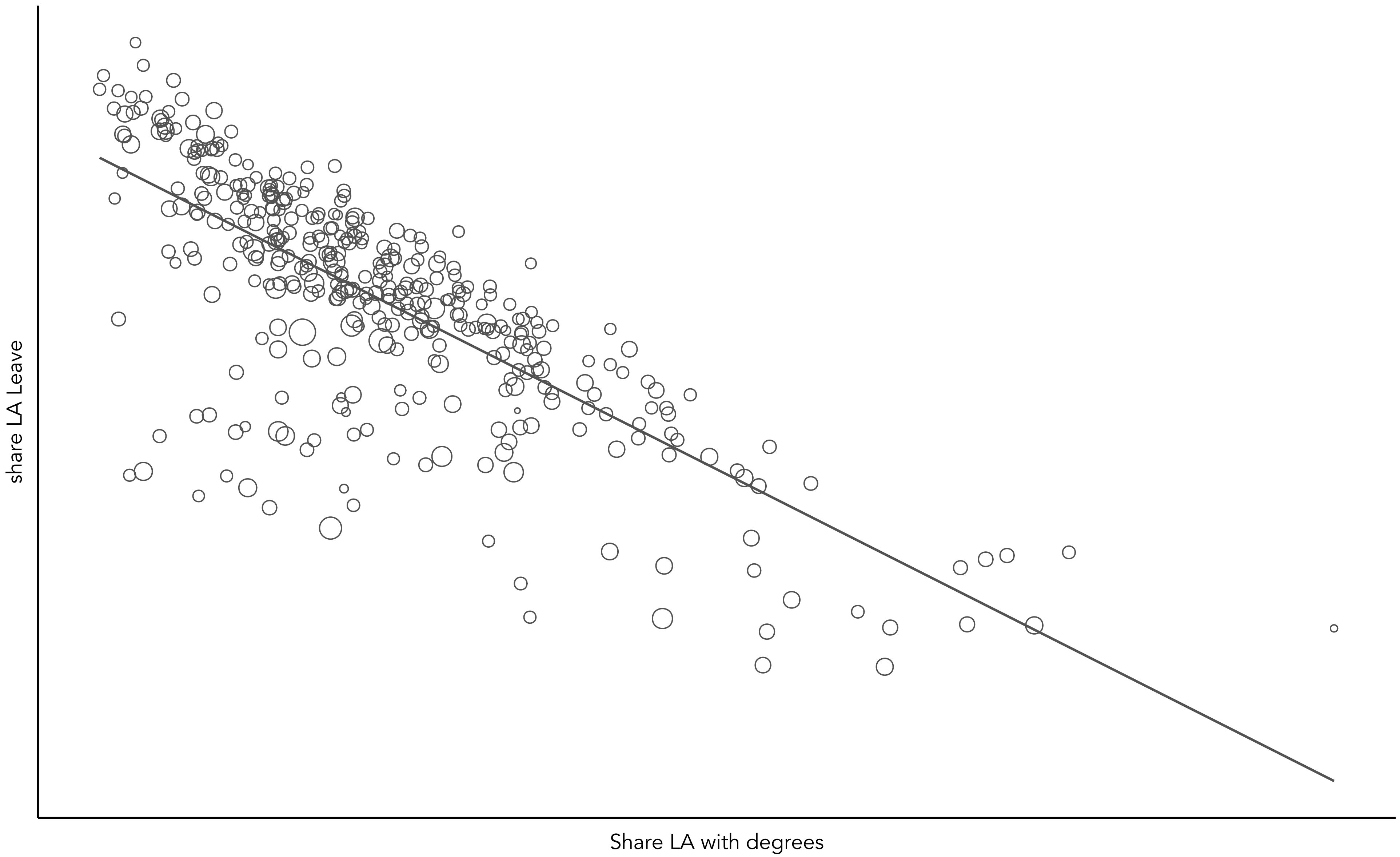


Summary of mappings - specification

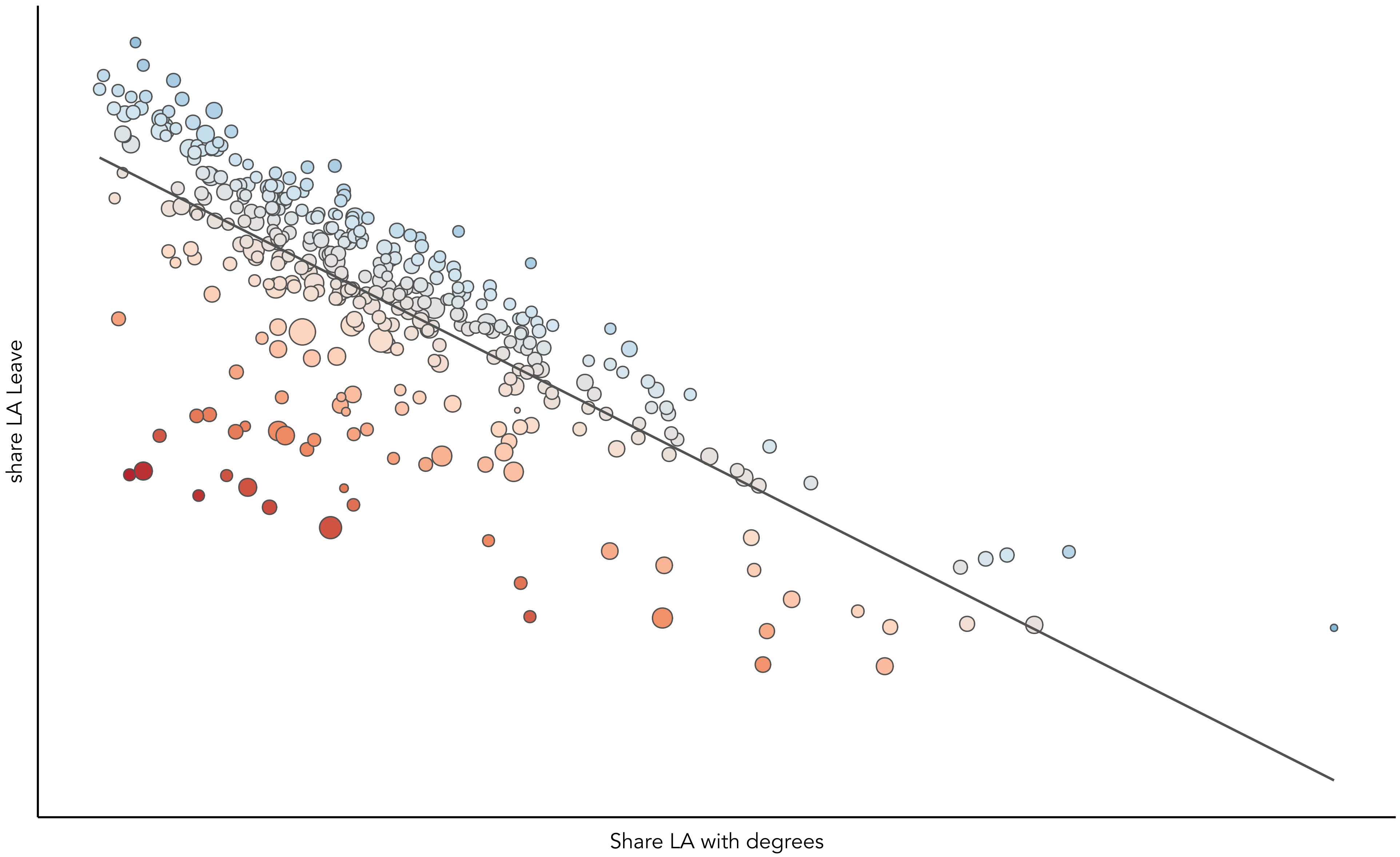
data channel	aes	geom
share leave	x position	point
share degrees	y position	point
pop size	size (area)	point

```
referendum_data %>%
ggplot(data=.,
    aes(x=share_leave,
        y=degree_educated,
        size=electorate) +
    geom_point()
```

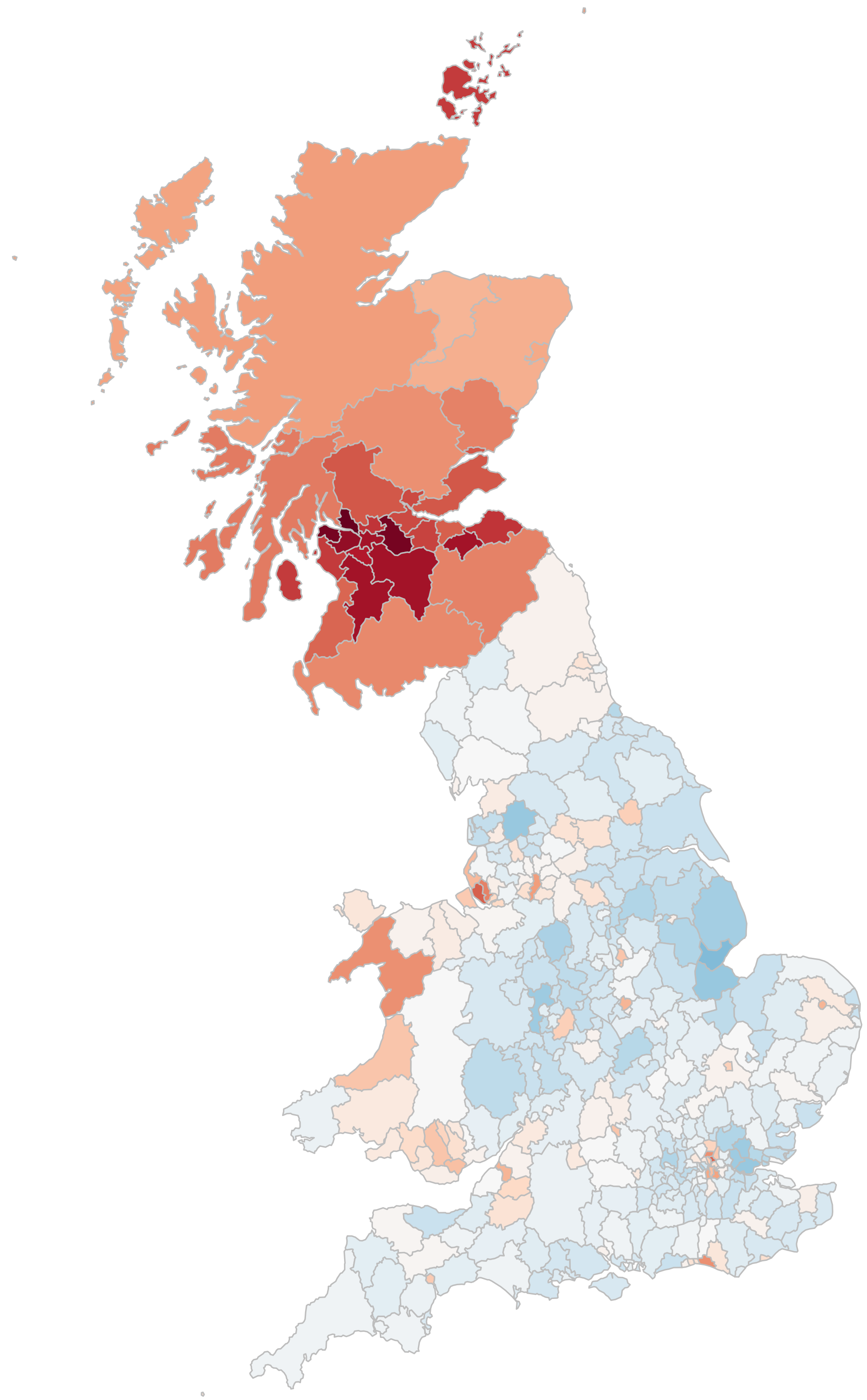
Leave vote by degree-level education



Leave vote by degree-level education

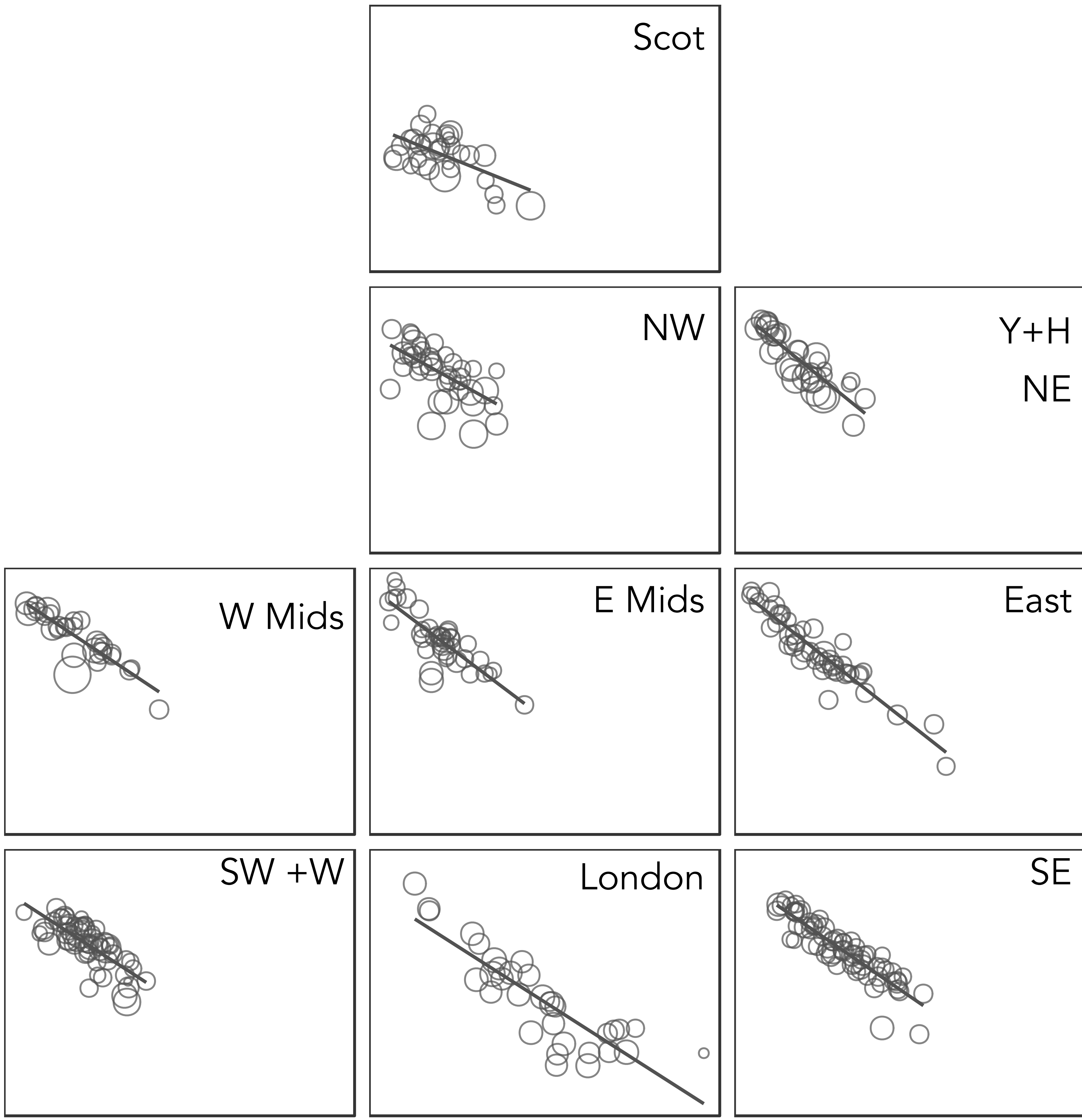


Leave vote by degree-level education

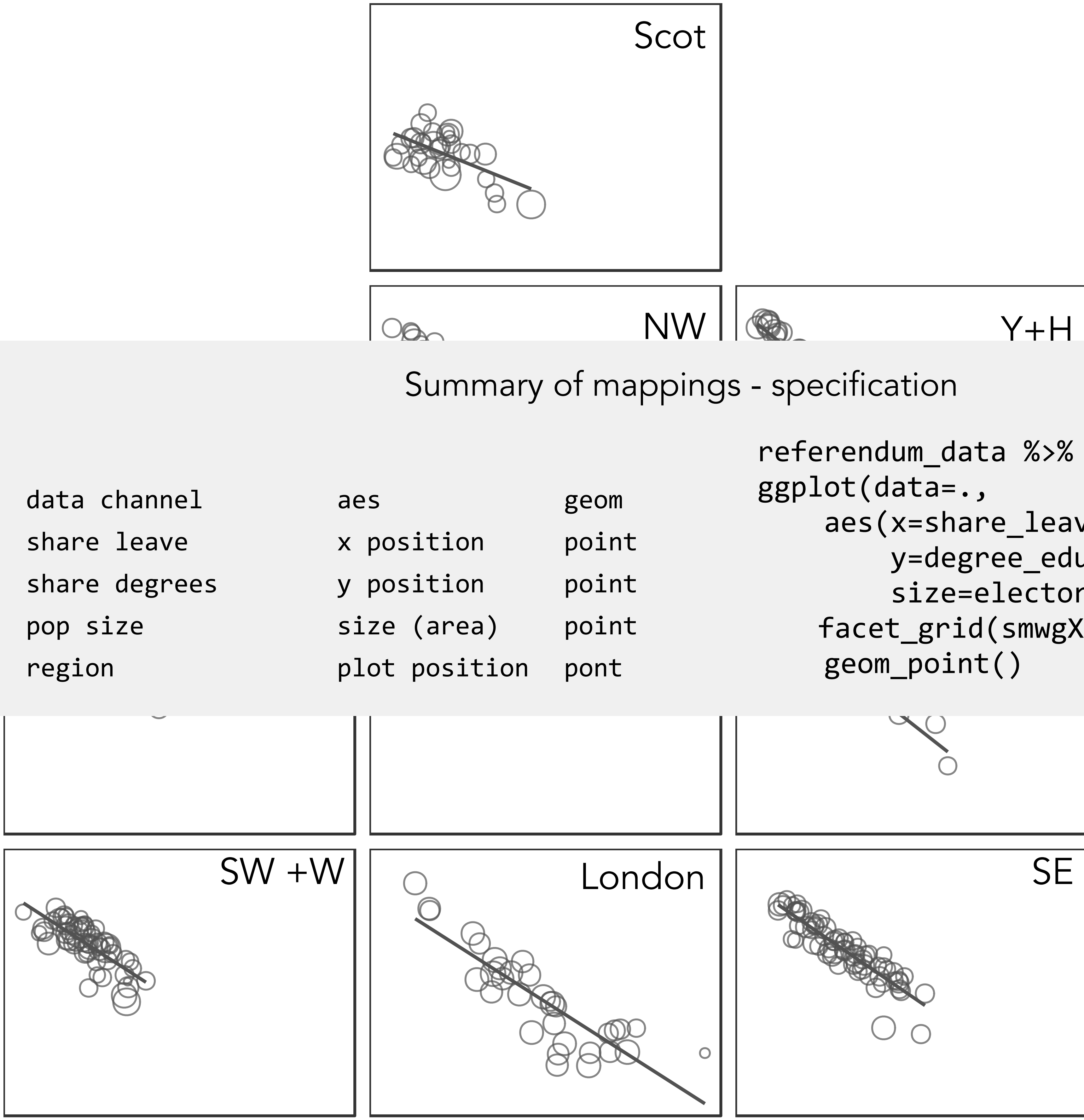


Leave vote by degree-level education faceted by region

Leave vote by degree-level education faceted by region



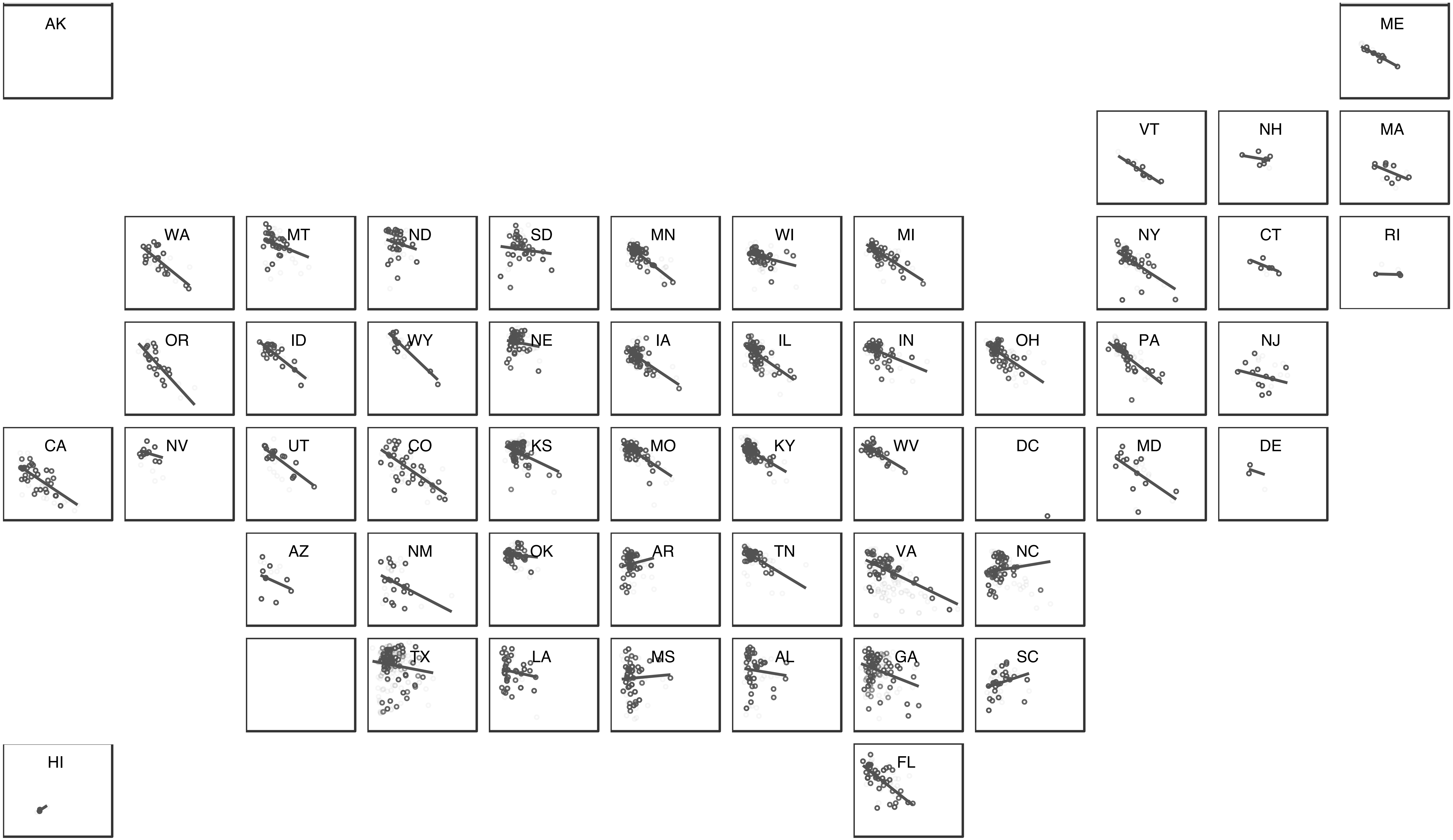
Leave vote by degree-level education faceted by region



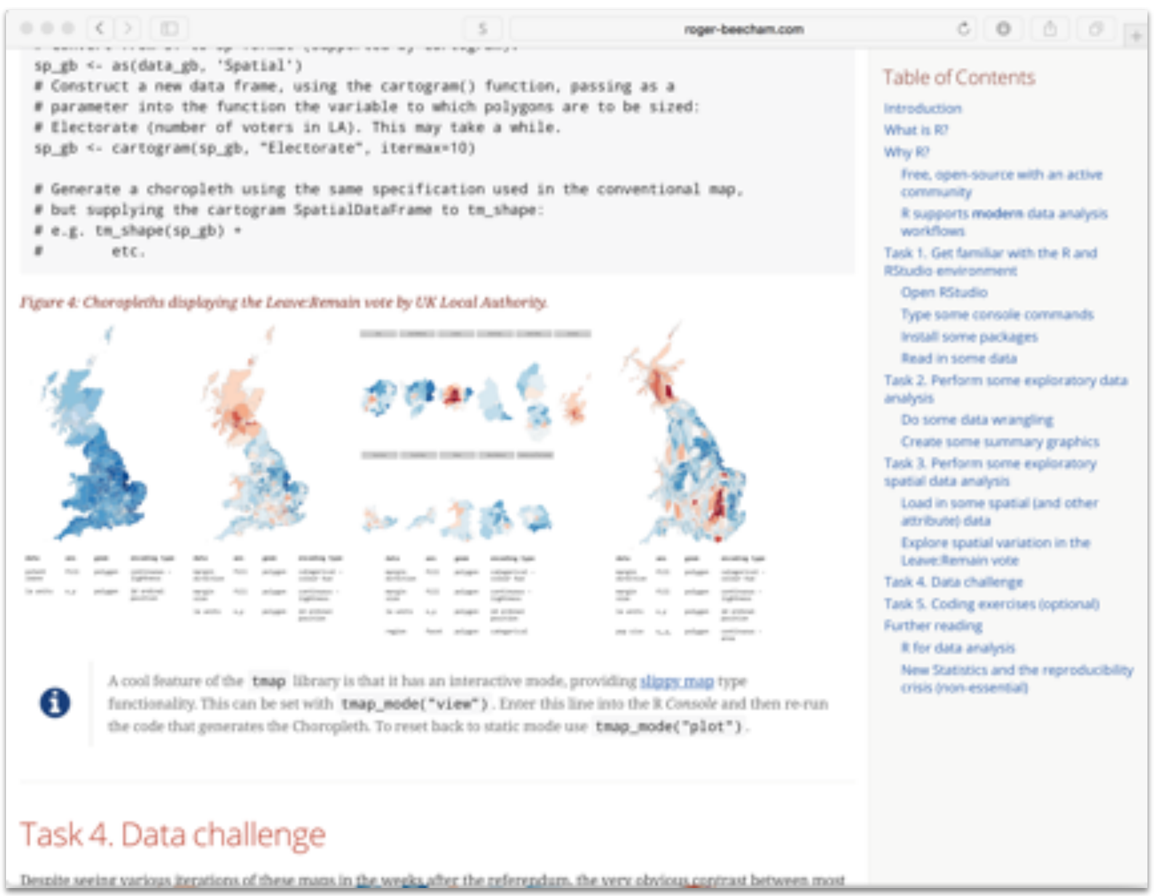
Summary of mappings - specification

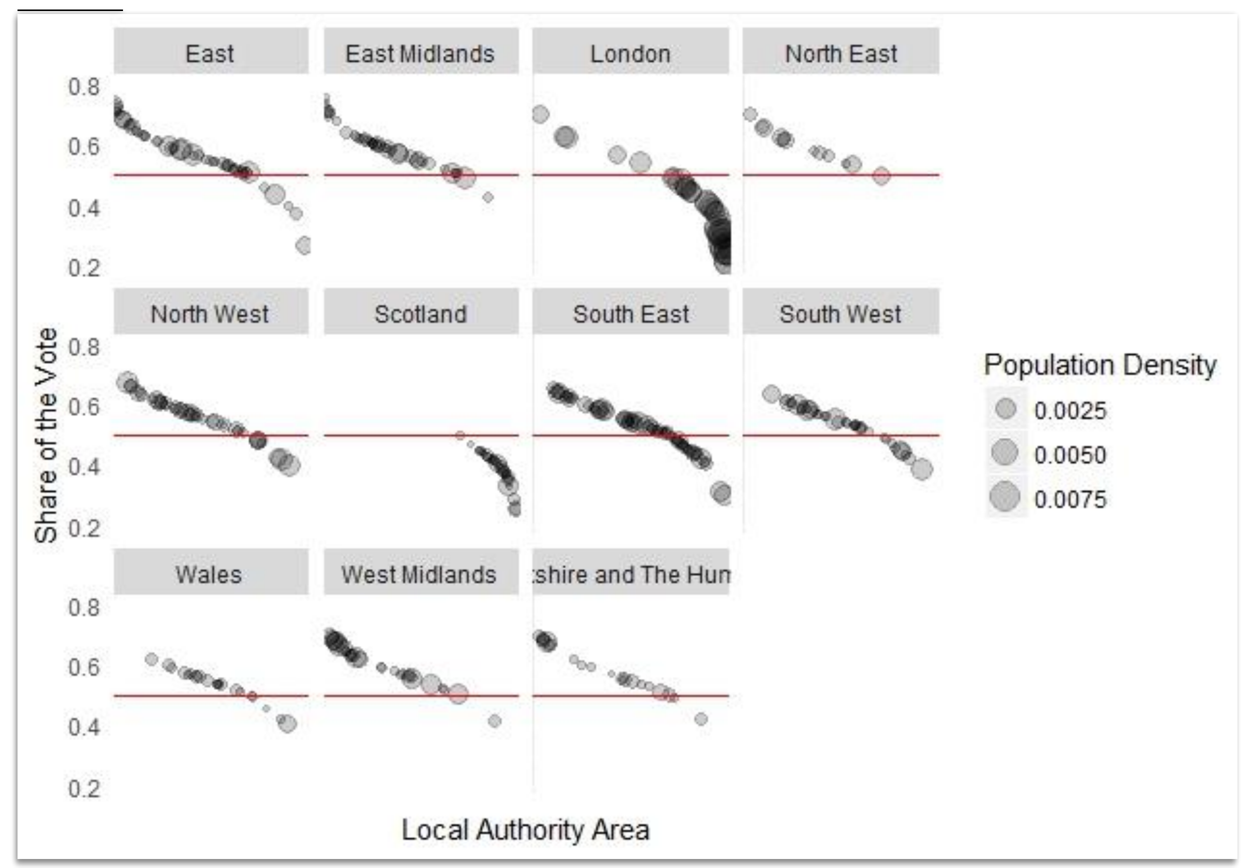
data channel	aes	geom
share leave	x position	point
share degrees	y position	point
pop size	size (area)	point
region	plot position	pont

```
referendum_data %>%
ggplot(data=.,
    aes(x=share_leave,
        y=degree_educated,
        size=electorate) +
   facet_grid(smwgX~smwgY) +
    geom_point()
```


AK
ME
VT
NH
MA
WA
MT
ND
SD
MN
WI
MI
NY
CT
RI
OR
ID
WY
NE
IA
IL
IN
OH
PA
NJ
CA
NV
UT
CO
KS
MO
KY
WV
DC
MD
DE
AZ
NM
OK
AR
TN
VA
NC
TX
LA
MS
AL
GA
SC
HI
FL

Combining information visualization theory and the grammar of graphics to do and **teach modern data analysis**





Term 1 GIS MSc students

Combining information visualization theory and the grammar of graphics to **do** and teach modern **data analysis**

JOURNAL OF SPATIAL INFORMATION SCIENCE
Number N (YYYY), pp. xx–yy
doi:10.5311/JOSIS.YYYY.II.NNN

RESEARCH ARTICLE

Locally-varying explanations behind the United Kingdom's vote to leave the European Union

Roger Beecham[1], Aidan Slingsby[2], and Chris Brunsdon[3]

[1]University of Leeds, UK
[2]City, University of London, UK
[3]Maynooth University, Republic of Ireland

Received: M, D, Y; returned: M, D, Y; revised: M, D, Y; accepted: M, D, Y.

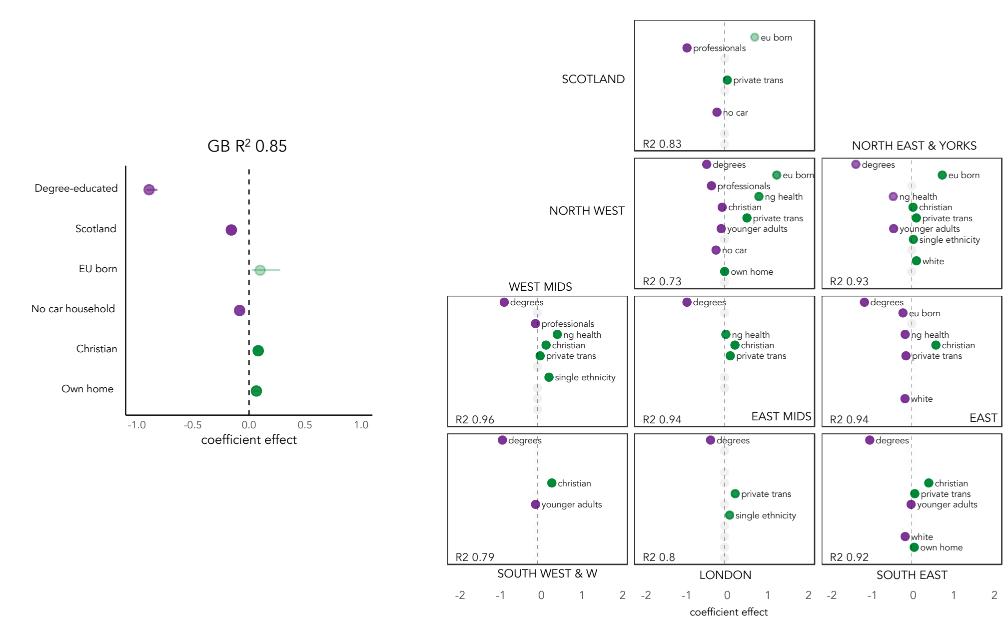
Abstract: Explanations behind area-based (Local Authority-level) voting preference in the 2016 referendum on membership of the European Union are explored using aggregate-level data. Developing local models, special attention is paid to whether variables explain the vote equally well across the country. Variables describing the post-industrial and economic 'successfulness' of Local Authorities most strongly discriminate variation in the vote. To a lesser extent this is the case for variables linked to 'metropolitan' and 'big city' contexts, which assist the Remain vote, those that distinguish more traditional and 'nativist' values, assisting Leave, and those loosely describing material outcomes, again reinforcing Leave. Whilst variables describing economic competitiveness co-vary with voting preference equally well across the country, the importance of secondary variables – those distinguishing metropolitan settings, values and outcomes – does vary by region. For certain variables and in certain areas, the direction of effect on voting preference reverses. For example, whilst levels of European Union migration mostly assist the Remain vote, in parts of the country the opposite effect is observed.

Keywords: European Union; referendum; multi-level modelling; geographically-weighted statistics; LASSO; area-based analysis.

variables distinguishing LAs that are within London and Scotland. The line through the regression coefficients in Figure 4 and their transparency is determined by 95% confidence intervals calculated via a bootstrap.

The model created under this LASSO procedure identified six variables. *Degree-educated* contributes the largest coefficient effect. Holding the other variables constant, a one percent point increase in the *degree-educated* population decreases the leave vote by 0.9 percent points. The fact that Scotland is selected by the LASSO procedure is instructive: there is something fundamentally different about Scotland, not accounted for completely by census variables, that lowers preference for Leave (by 16% points after controlling for demographics). The effect of the *EU-born* variable is counter to that expected. In Figure 1 the variable appears negatively correlated with Leave and we speculate might represent economic opportunity and relative diversity. After controlling for variation in other demographic characteristics, the model suggests an increase in the *EU-born* population in fact *increases* the Leave vote. Notice, however, the large confidence interval around this coefficient. Given the resampling procedure used to generate our bootstrap, this interval indicates that the effect of *EU-born* is likely to vary across LAs.

4.3 Region-specific explanations implied by local models



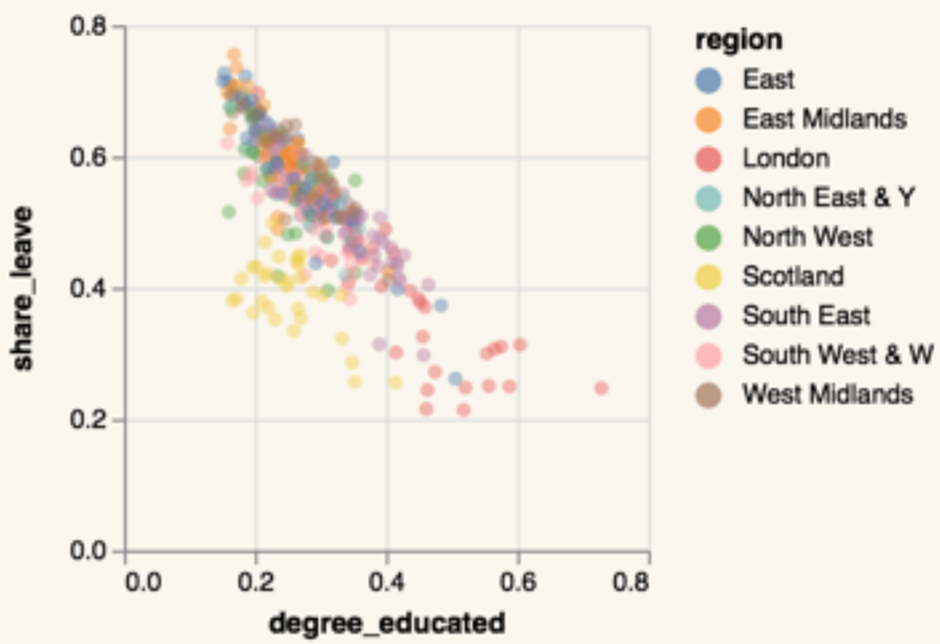
(a) (b)

Figure 4: Coefficients for multivariate models fit to data for GB (4a) and super-regions (4b) and annotated with adjusted R^2. Positive coefficients are green, negative purple and colour lightness varies according to a 95% Confidence Interval calculated via a bootstrap. Note that the GB model was specified with additional dummy variables for Scotland and London.

Vega-Lite

Data	static or data source
Transform	filter, aggregation, binning
Mark	point, line, bar, polygon
Encoding	mapping between data and mark properties
Scale	functions that map data values to visual values
Guides	axes and legends

Vega-Lite



Leave against degree-educated

```
"data": {"url": "data/data_gb.csv"},
"mark": {"type": "point", "filled": true},
"encoding": {
   "x": {"field": "degree_educated", "type": "quantitative"},
   "y": {"field": "share_leave", "type": "quantitative"},
   "color": {"field": "region", "type": "nominal"}
}
```

Plot grammar

Vega-Lite


```
"data": {"url": "data/data_gb.csv"},
"repeat" : {"column": ["degree_educated", "not_good_health",
"private_transport_to_work"]},
"spec": {
   "mark": {"type": "point", "filled": true},
   "encoding": {
     "x": {"field": {"repeat": "column"}, "type": "quantitative"},
     "y": {"field": "share_leave", "type": "quantitative"},
     "color": {"field": "region", "type": "nominal"}
   }
}
```

Vega-Lite

Grammar of Interaction

selections

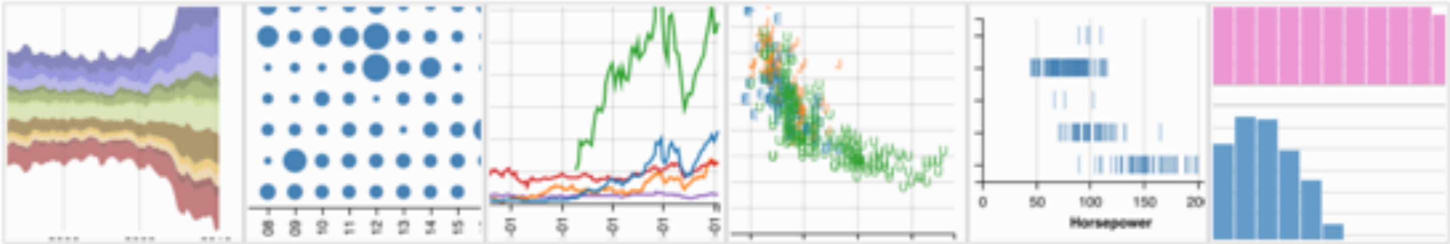
map user input (e.g. mouse moves)

into data queries

which drive conditional encodings, filter data points etc.

```
"data": {"url": "data/data_gb.csv"},
"mark": {"type": "point", "filled": true},
"selection" : {"picked": {"type": "single", "on":"mouseover"}},
"encoding": {
   "x": {"field": "degree_educated", "type": "quantitative"},
   "y": {"field": "share_leave", "type": "quantitative"},
   "color": {
     "condition":
        {"selection": "picked", "field": "region", "type": "nominal"}, "value": "grey"}
}
```

Declarative Visualization in Python

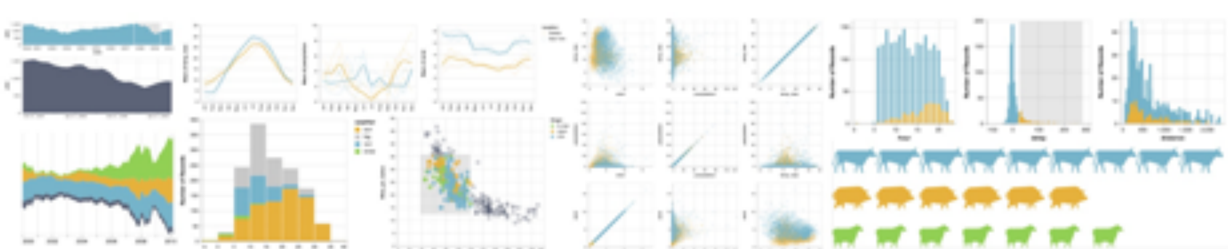


Altair is a declarative statistical visualization library for Python, based on Vega-Lite.

With Altair, you can spend more time understanding your data and its meaning. Altair's API is simple, friendly and consistent and built on top of the powerful Vega-Lite visualization grammar. This elegant simplicity produces beautiful and effective visualizations with a minimal amount of code.

github.com/altair-viz/altair

elm-vega



Declarative visualization for Elm

This library allows you to create Vega-Lite specifications in Elm providing a pure functional interfa
visualization construction.

The library does not generate graphical output directly, but instead it allows you to create a JSON
sent to the Vega-Lite runtime to create the output. This is therefore a 'pure' Elm package without
dependencies.

github.com/gicentre/elm-vega

Teaching materials	github.com/rogerbeecham/intro-visual-data-analysis/
Paper and code	github.com/rogerbeecham/brexit-analysis/

OBSERVABLE