Advanced data visualizations with Stata

#StataViz ++

Asjad Naqvi

International Institute for Applied Systems Analysis (IIASA)
Wirtschaftsuniversität Wien (WU)

Stata UK Conference
10 September, 2021

Background

- Ph.D. Economics (2007-12), New School for Social Research, New York, USA.
- Started using Stata during my M.Sc. Economics degree (2003-04)
- Worked/still working on a ton of research projects mostly using Stata
  - Overseeing several large data projects
  - Going through more code than I like
- Why do all the dataviz stuff?
  - Online COVID-19 dataviz explosion + work-from-home $\rightarrow$ curiosity
  - Huge positive response from the online community

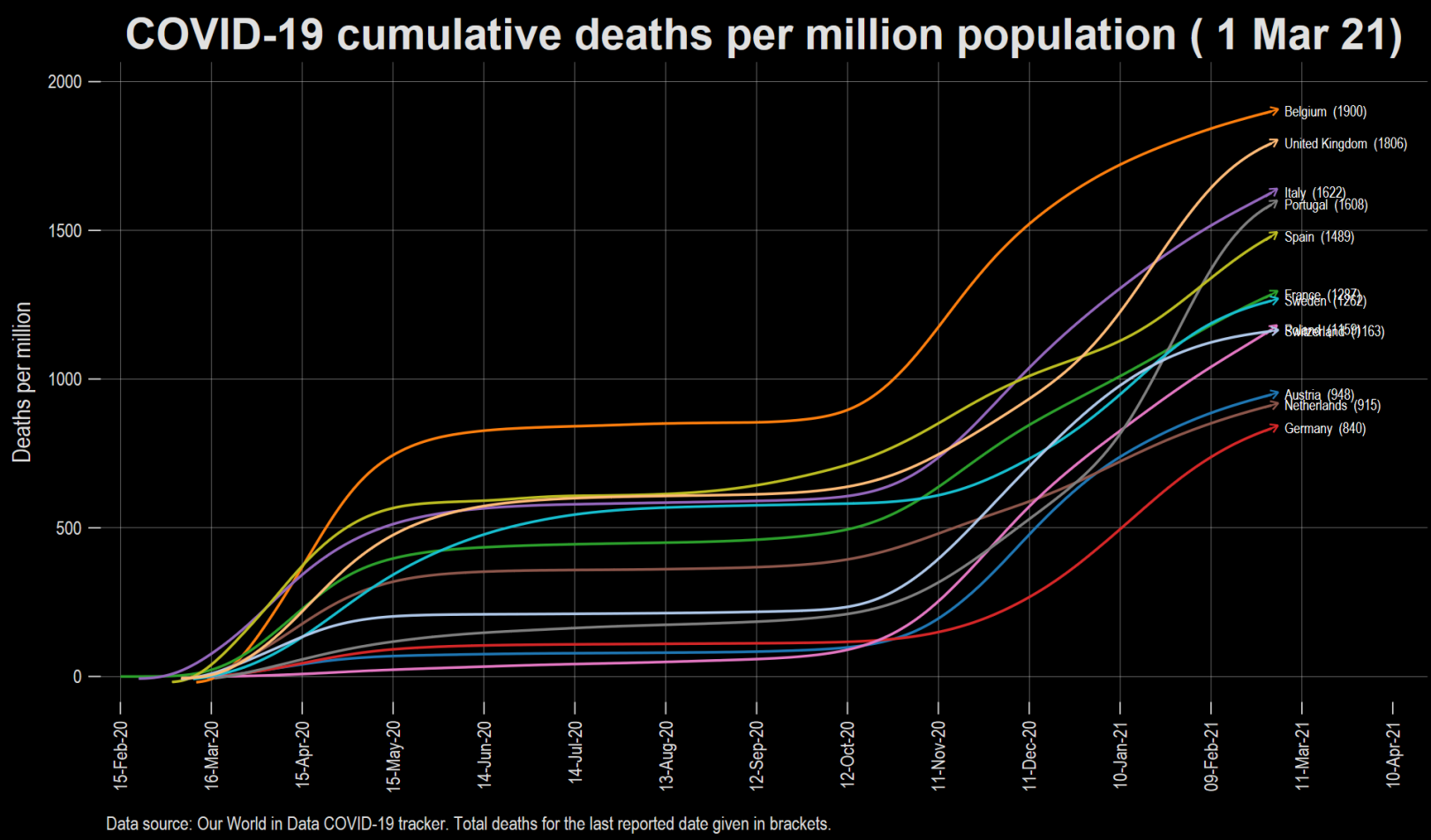
Why we need more #StataViz?

- Few official releases for graphs
- Still lots of options available in the default Stata structure
- Lots of great development in terms of dataviz packages:
  - `colorpalette`, `colrspace`, `heatplot` (Ben Jann), `spmap` (Maurizio Pisati), various packages by Nick Cox, `nwcommands` (Thomas Grund)

Why we need more #StataViz?

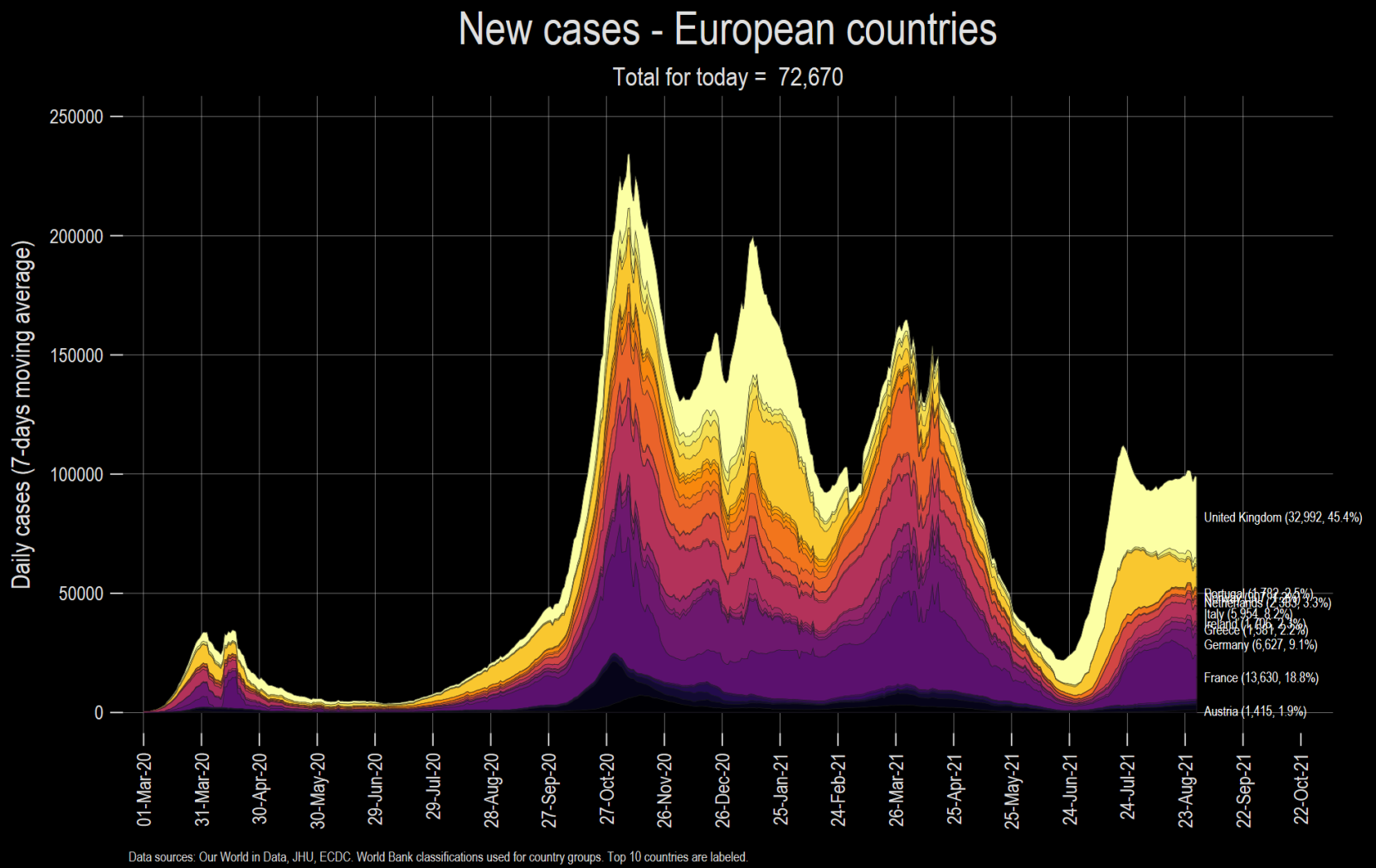
- Few official releases for graphs
- Still lots of options available in the default Stata structure
- Lots of great development in terms of dataviz packages:
  - `colorpalette`, `colrspace`, `heatplot` (Ben Jann), `spmap` (Maurizio Pisati), various packages by Nick Cox, `nwcommands` (Thomas Grund)
- What I will discuss here:
  - What has been developed
  - What is still possible
  - What core elements Stata should still add

colorpalette



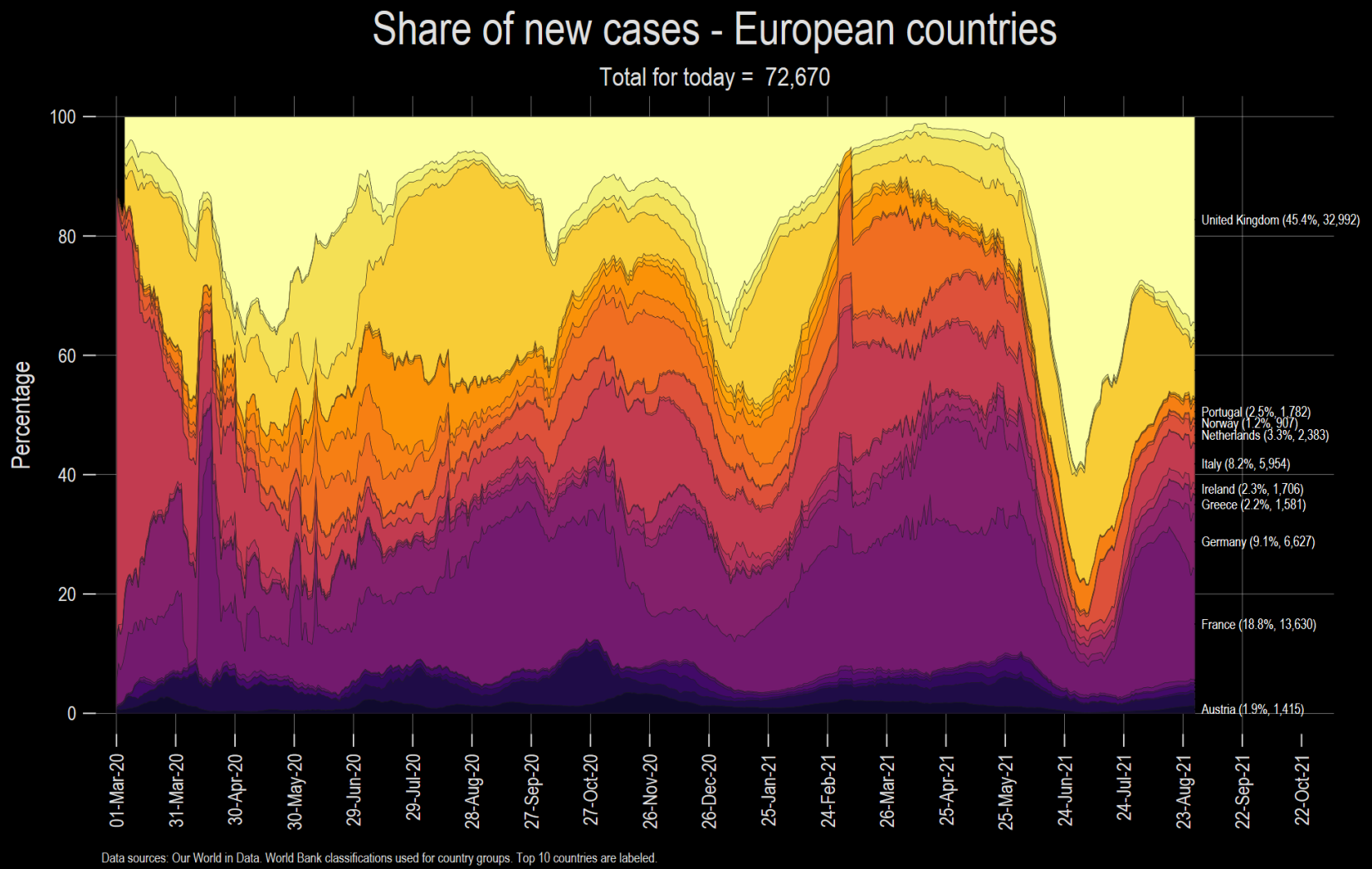
Guide: Arrows

colorpalette + stacked-area graphs



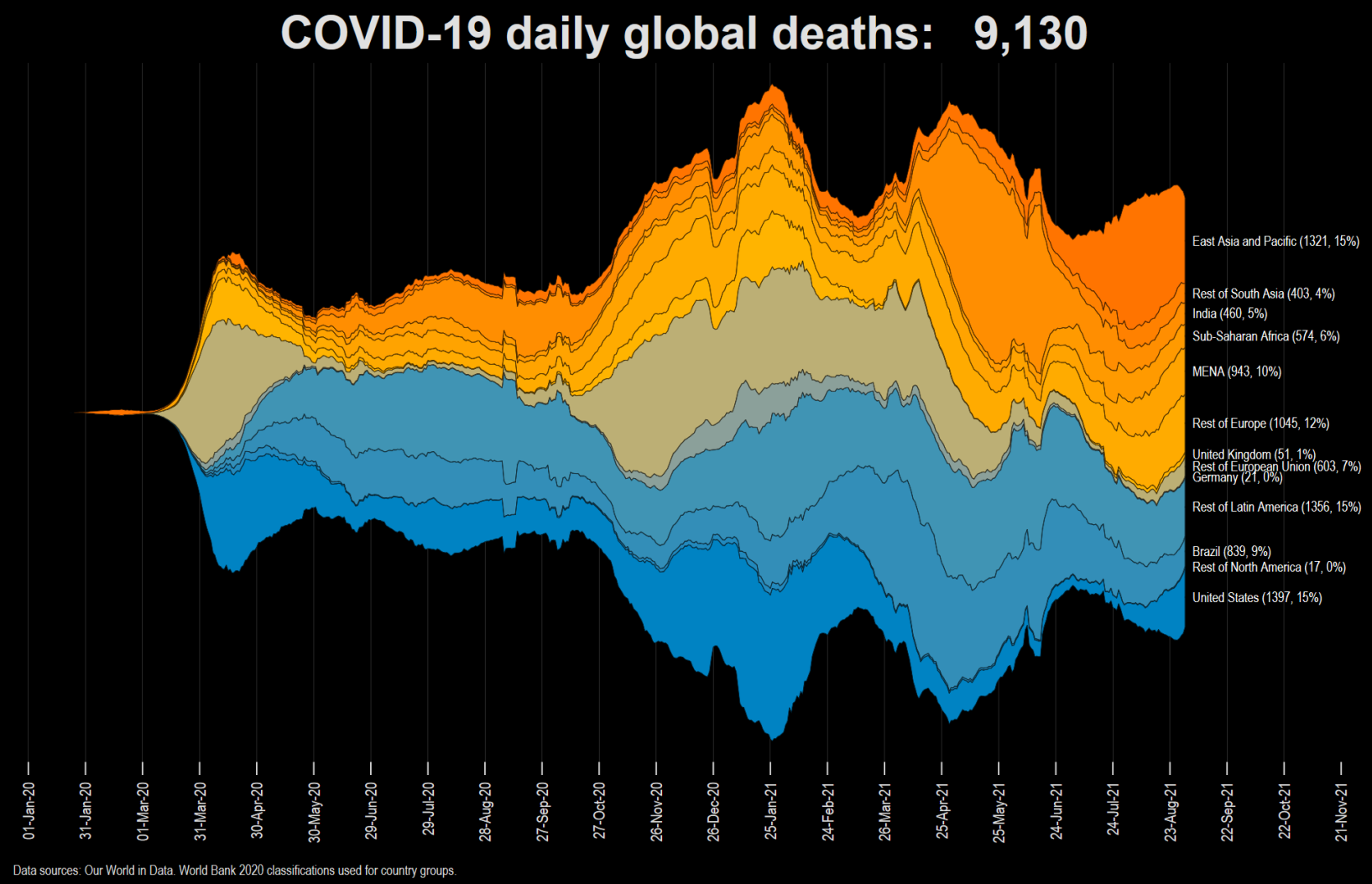
Guide: Stacked area graphs

colorpalette + stacked-area graphs



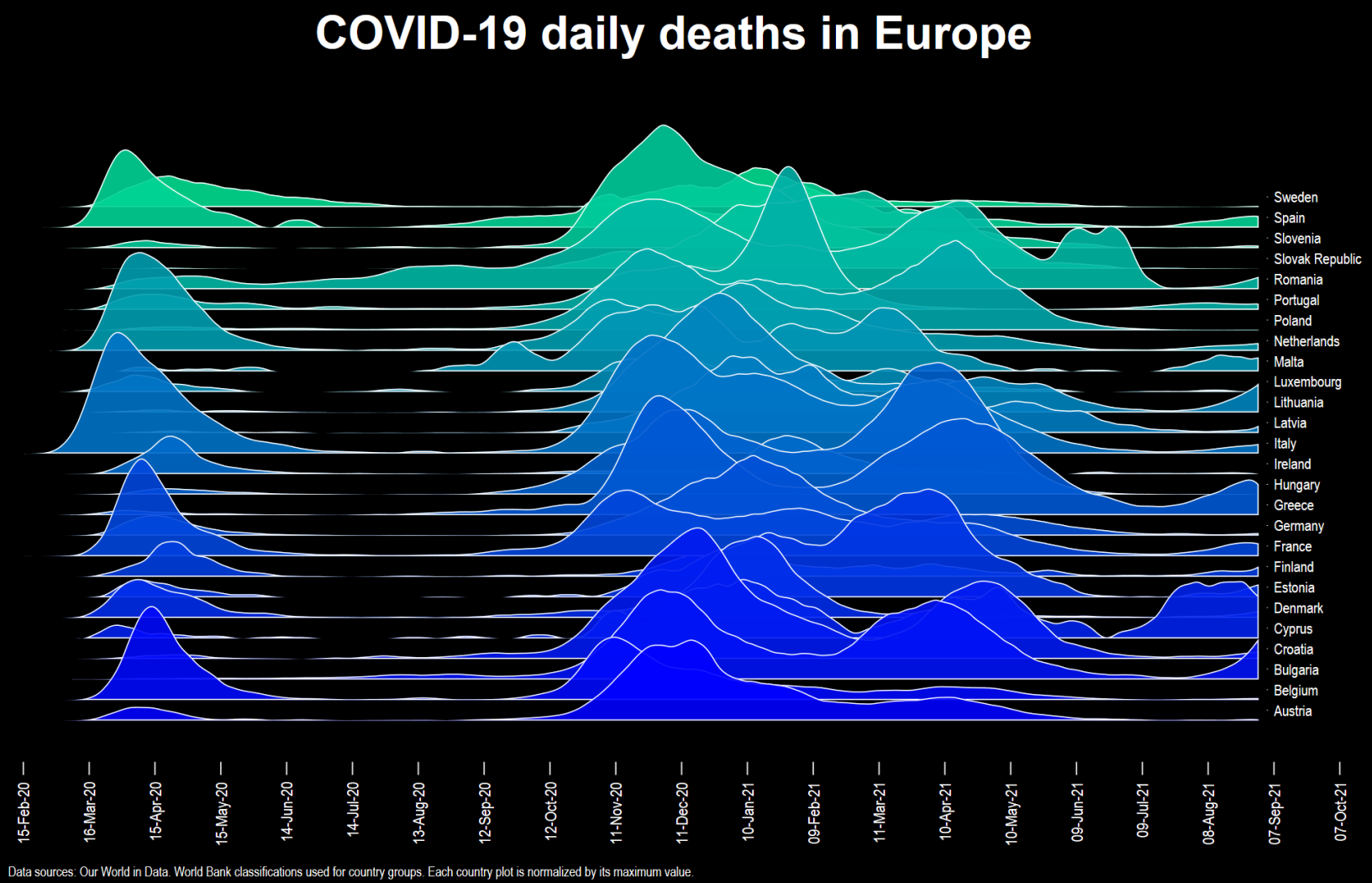
Guide: Stacked area graphs

colorpalette + stream plot

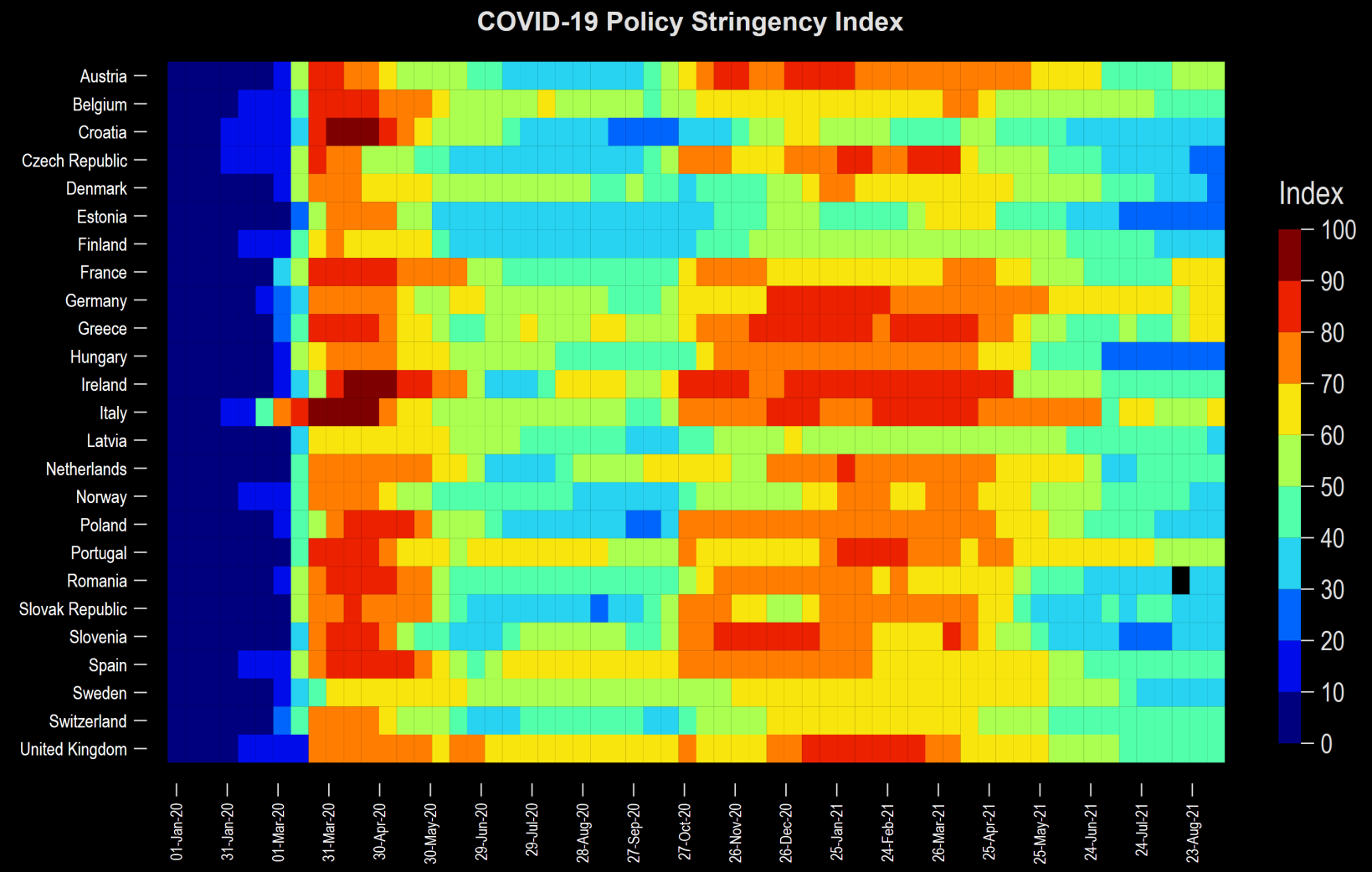


Guide: Stream plots

Ridgeline (Joy) plots

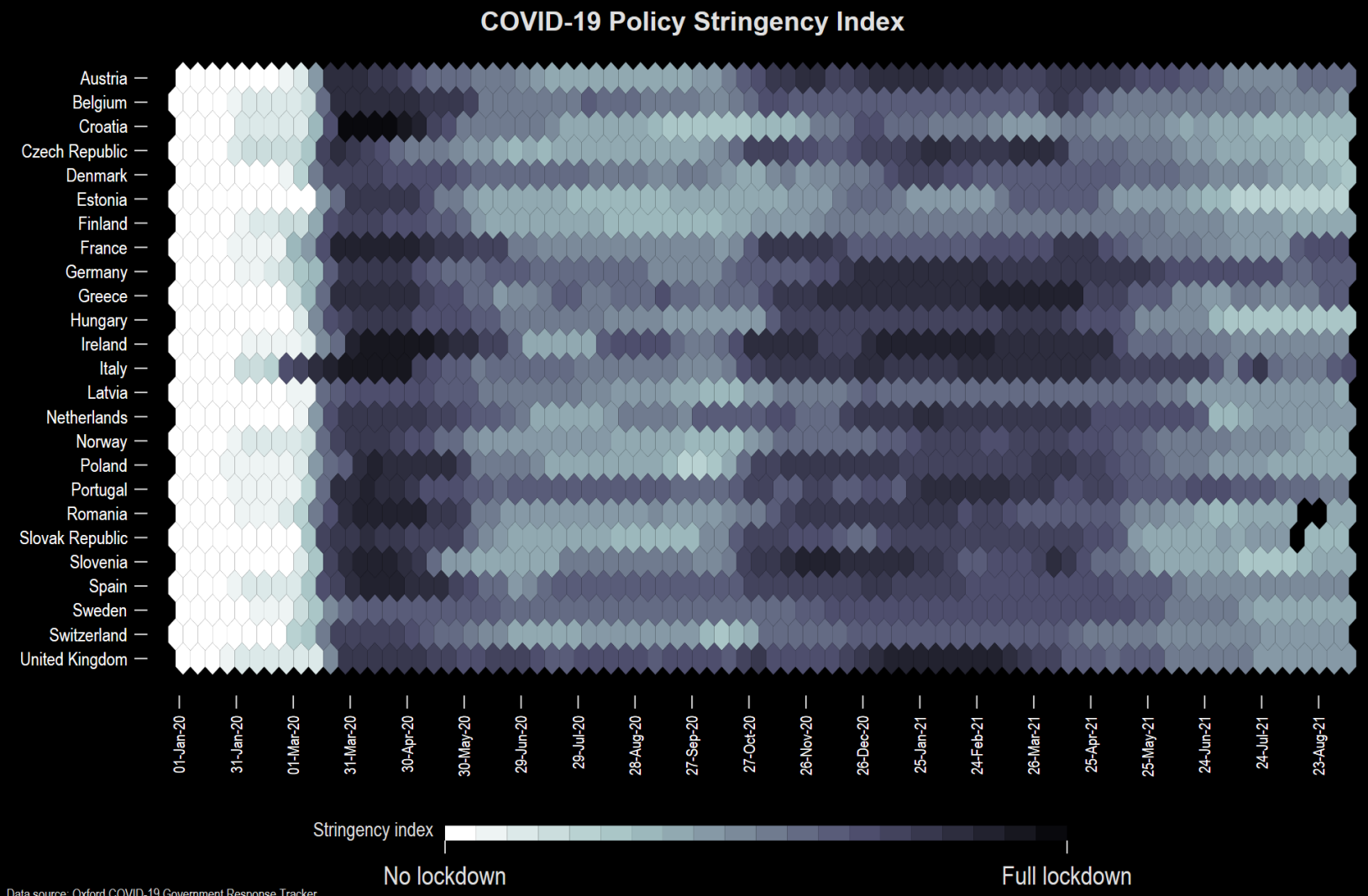


colorpalette + heatplot



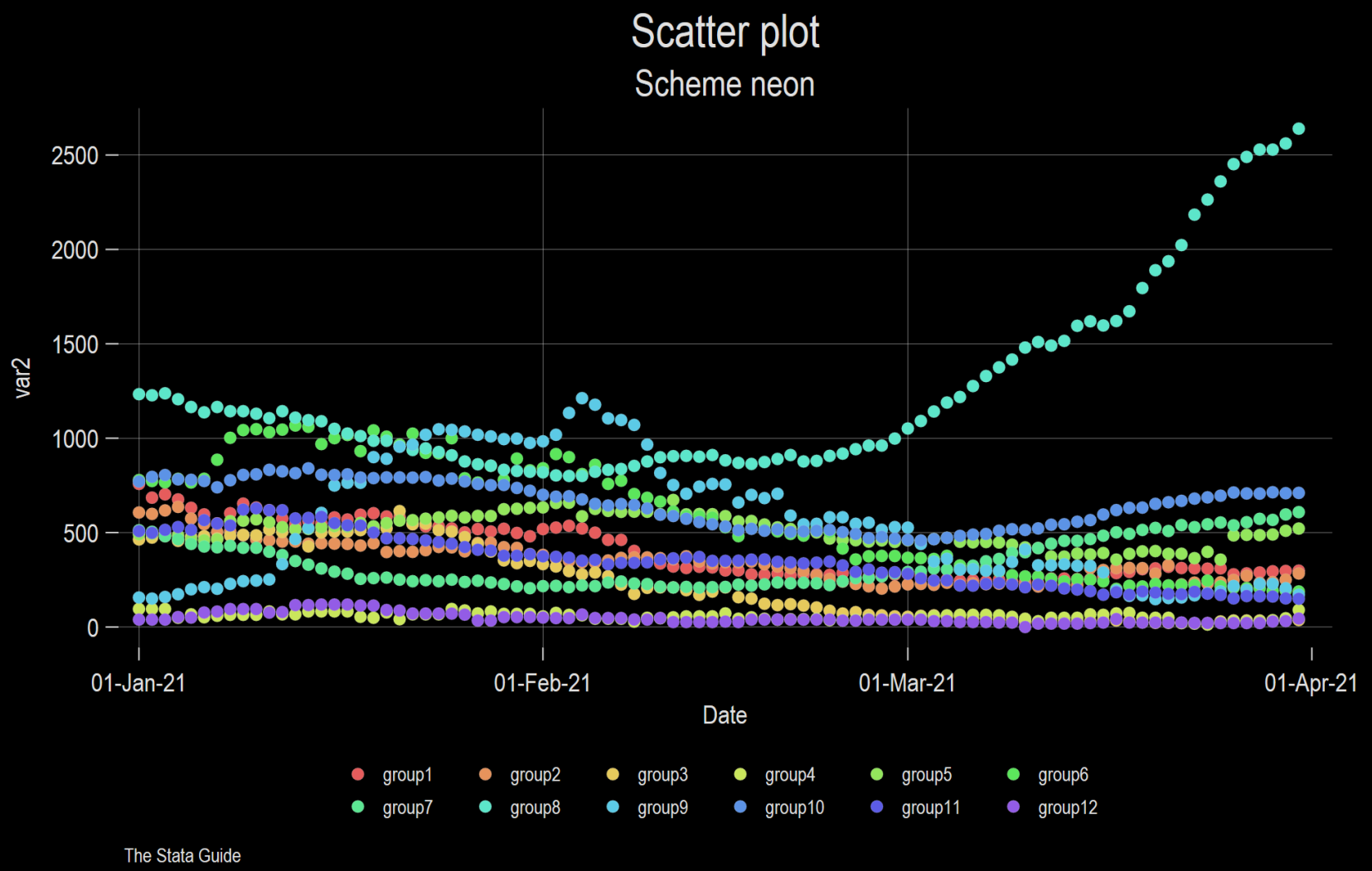
Guide: Heat plots

colorpalette + heatplot

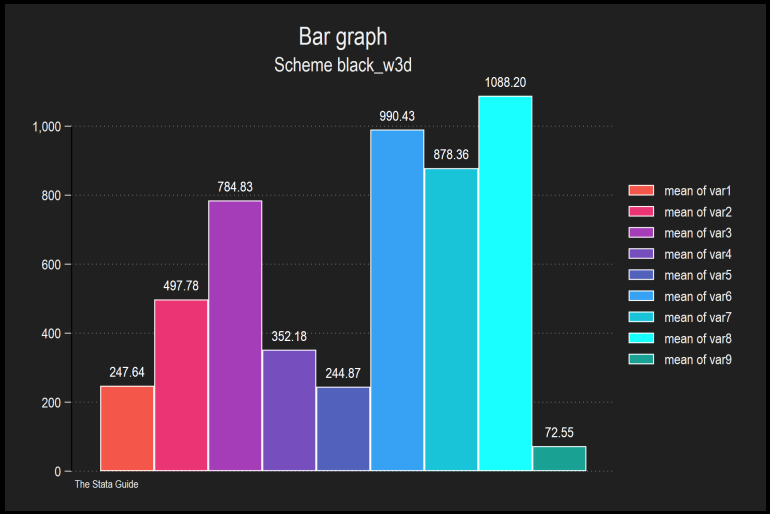


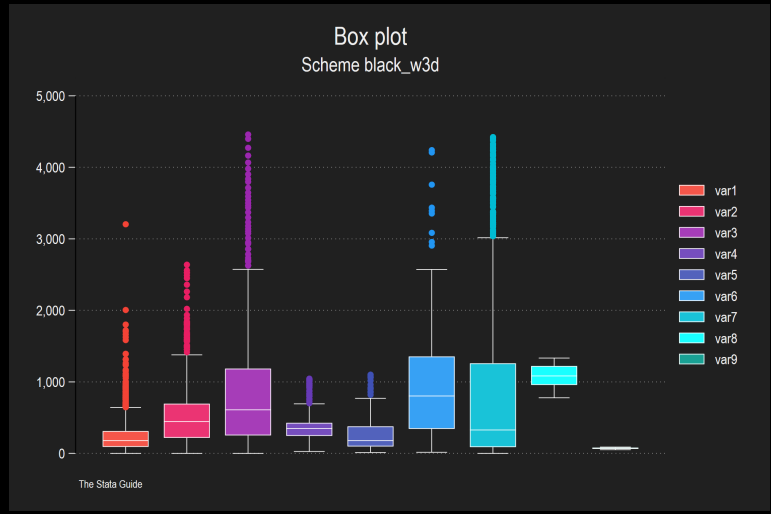
Guide: Heat plots

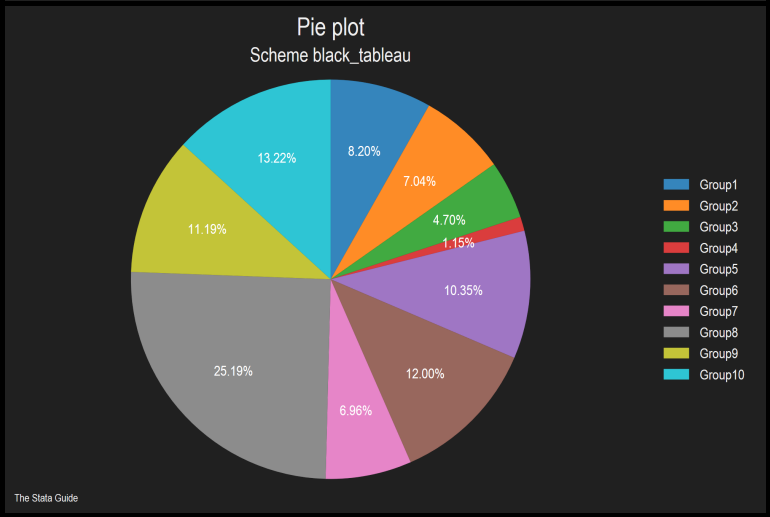
Schemes + colorpalette

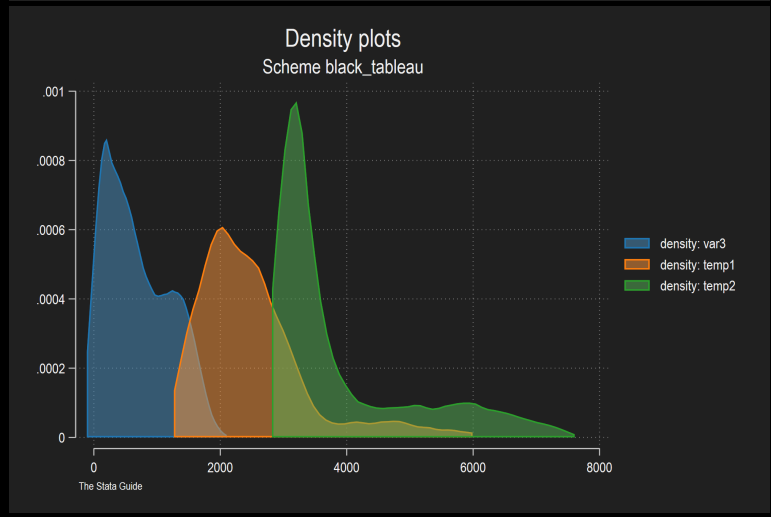


Schemes + colorpalette



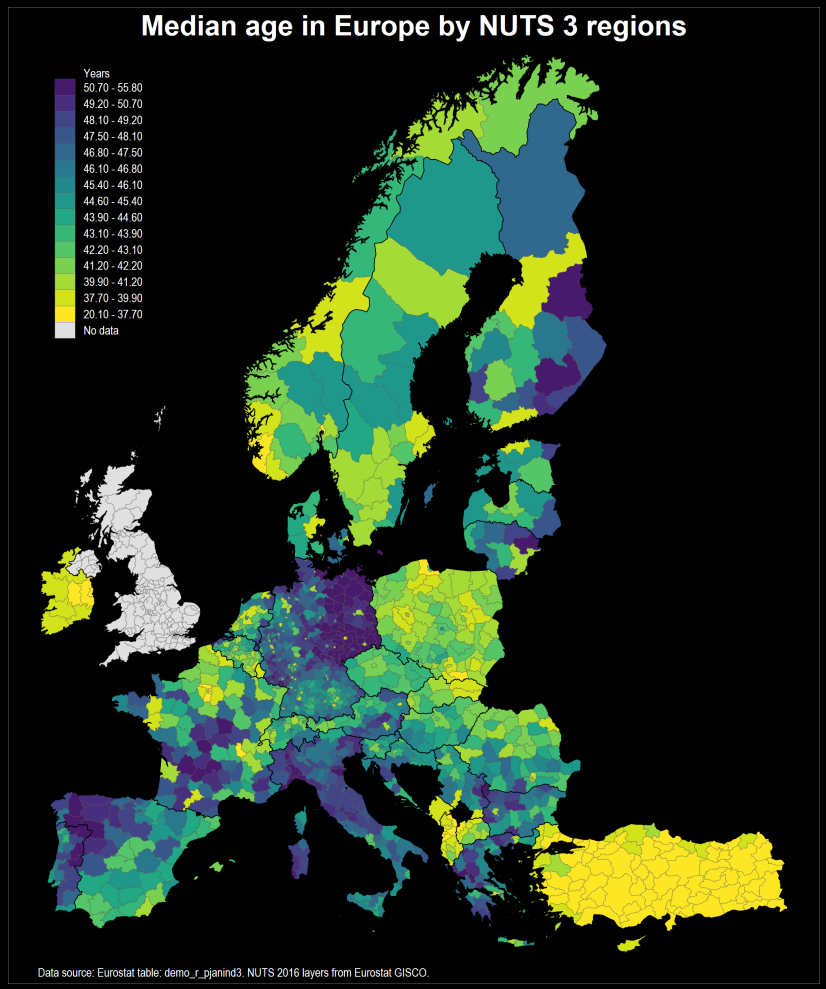


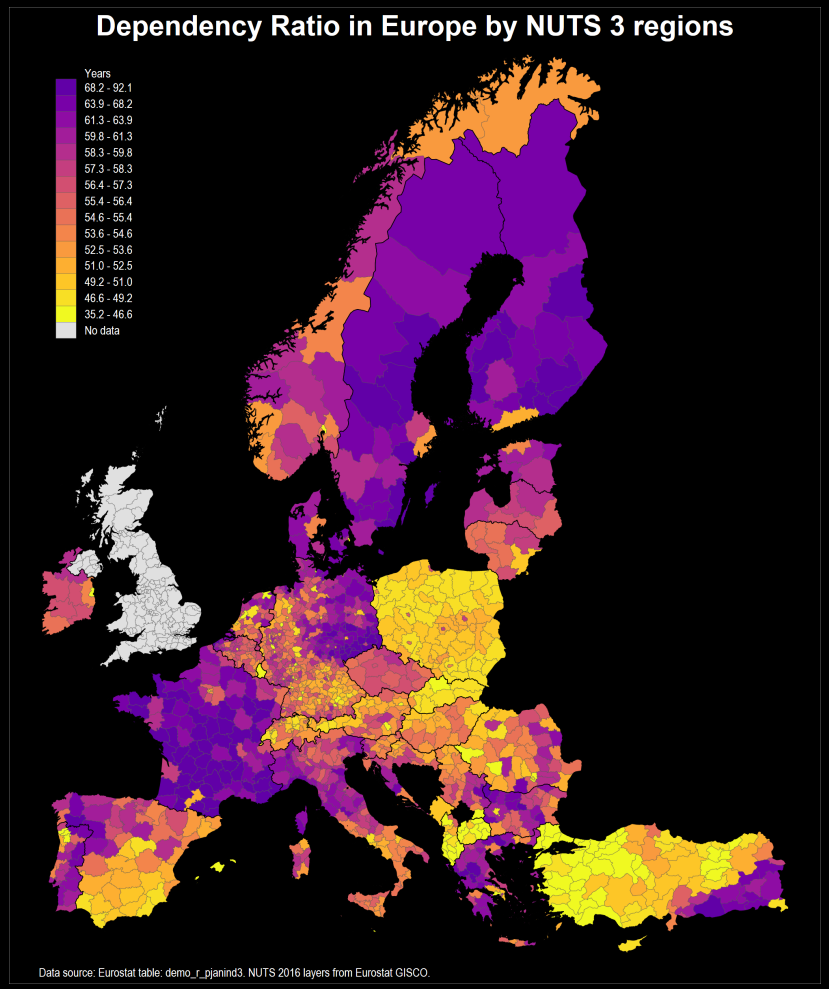




TSG Schemes

spmap + colorpalette





Guide: Maps II

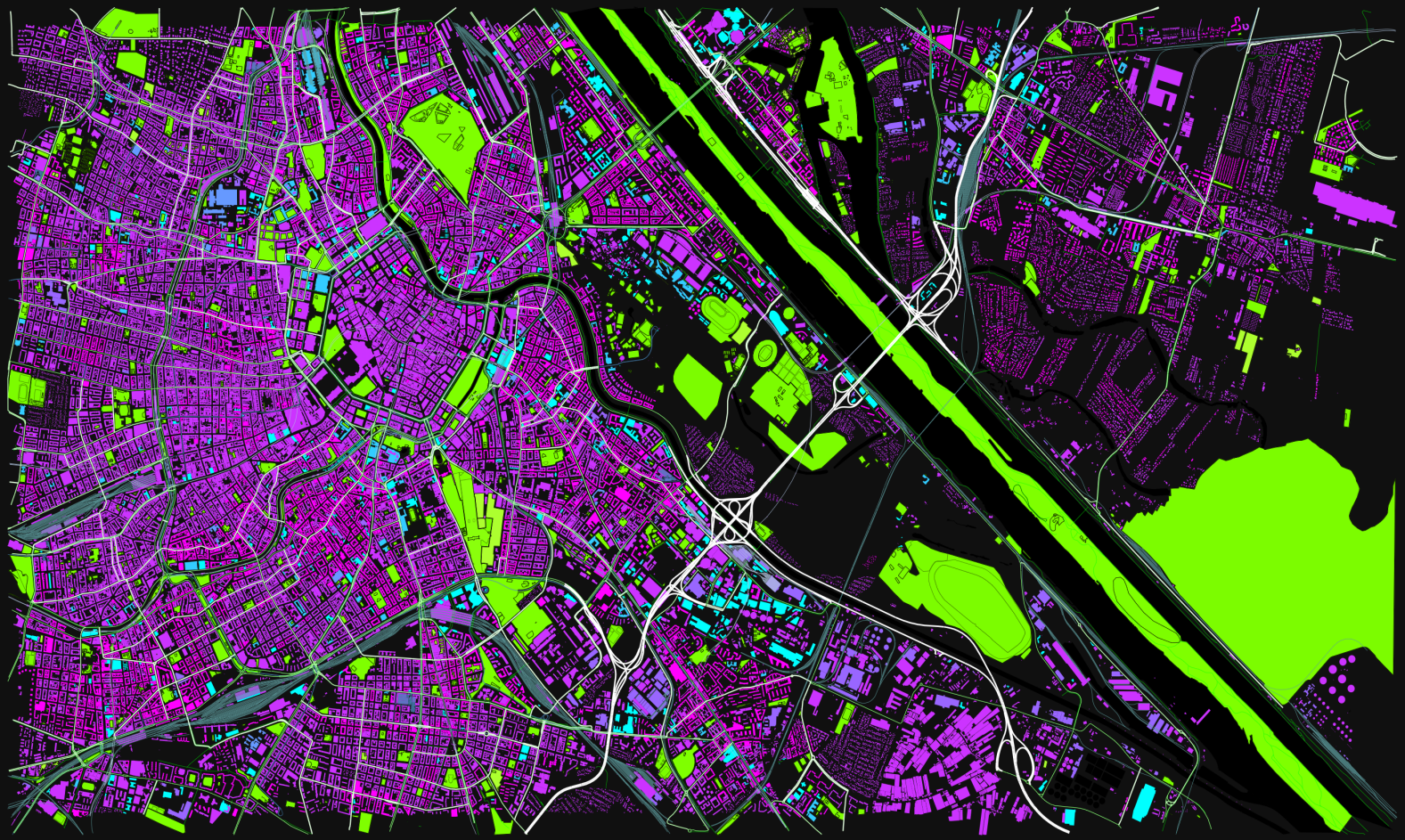
OSM + QGIS to Stata

- GIS data from OpenStreetMaps (OSM)
- QGIS can be used to extract multiple layers (points, lines, polygons)
- Why do this?
  - Fully customized maps
  - Spatial layer clipping + control over bounding box
  - Allow for very rich visualizations

Vienna

Guide: Advanced Mapping with Stata: OpenStreetMap (OSM) and QGIS

Vienna



Guide: Advanced Mapping with Stata: OpenStreetMap (OSM) and QGIS

New York

Guide: Advanced Mapping with Stata: OpenStreetMap (OSM) and QGIS

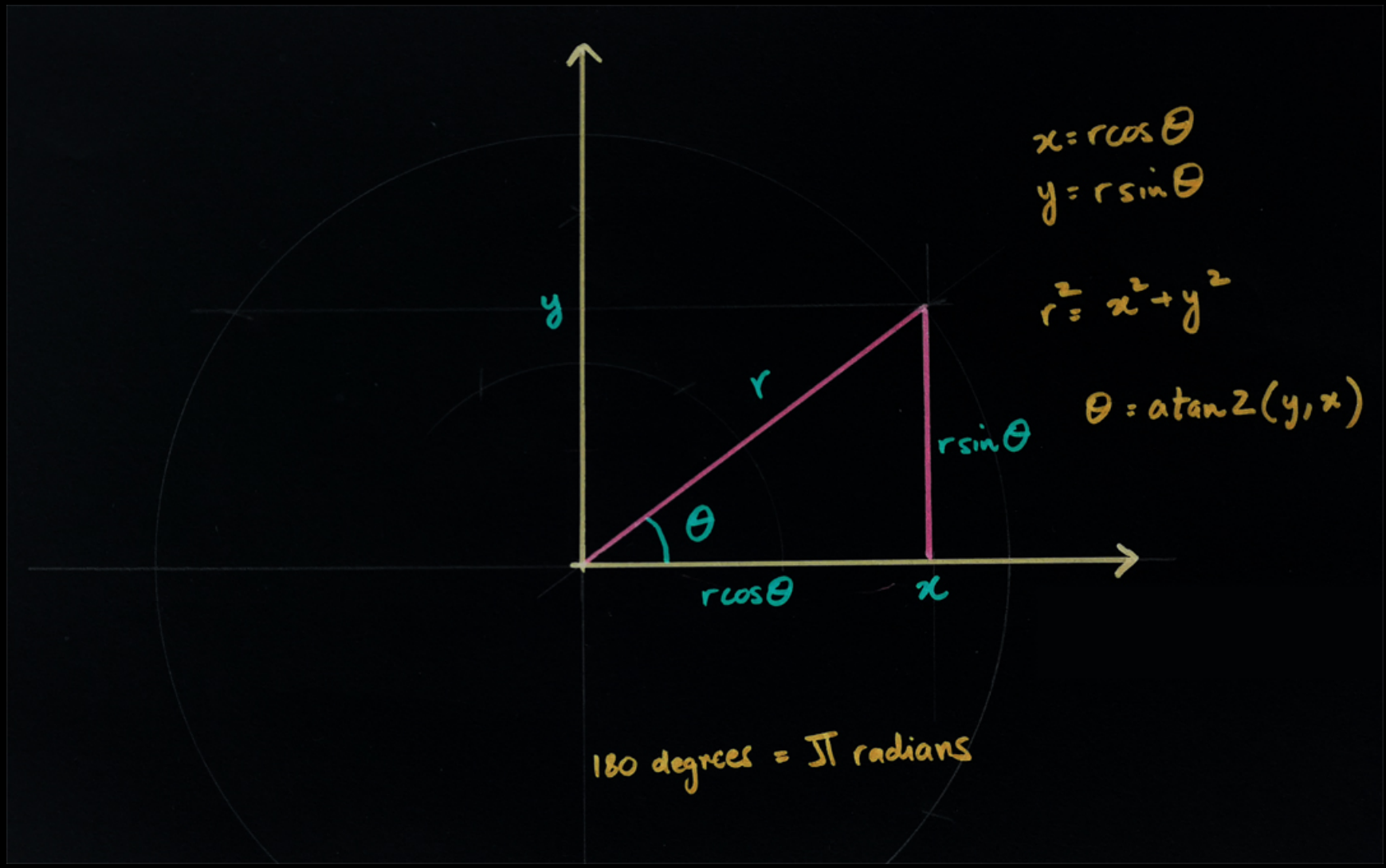
Maps + Networks

Social Connectivity Index - Austria (70+ pctile)

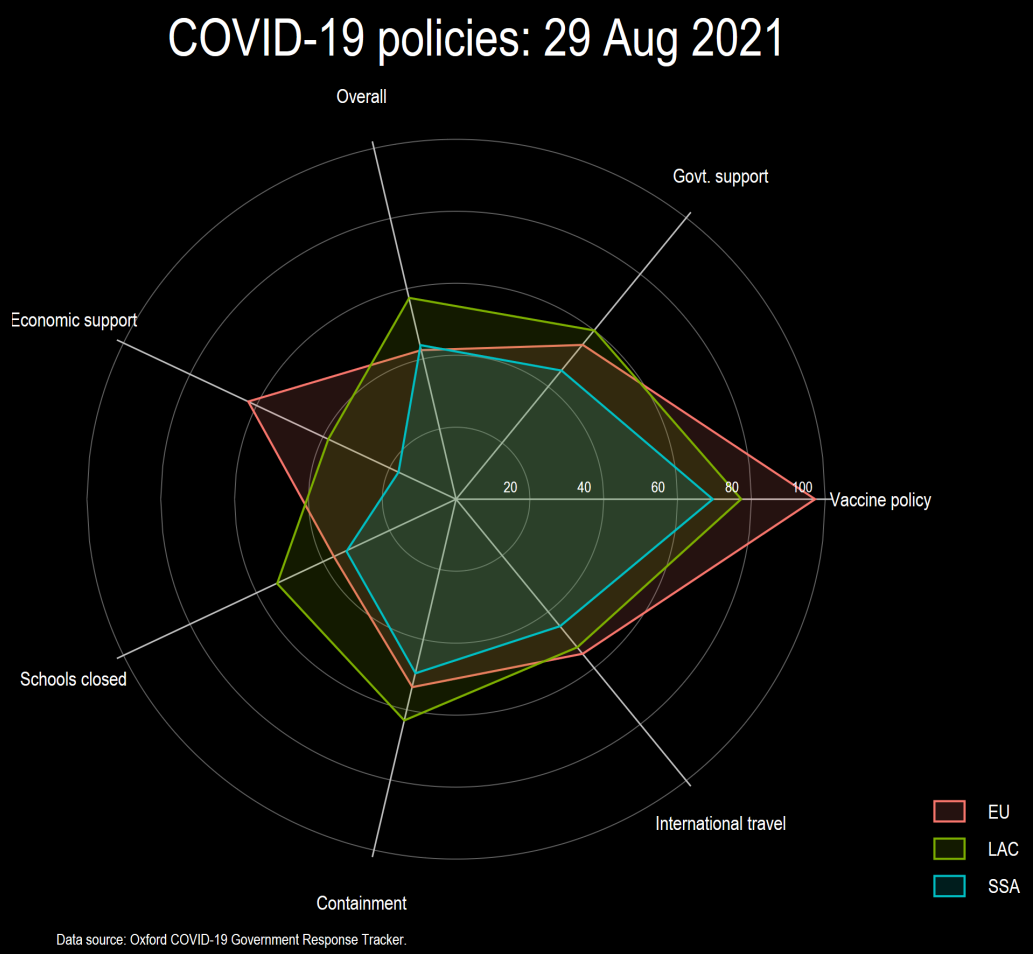
Map layers: GISCO Eurostat, SCI: Facebook. Within country tries not included. Darker shade is higher SCI.

Social Connectivity Index - UK (70+ pctile)

Map layers: GISCO Eurostat, SCI: Facebook. Within country tries not included. Darker shade is higher SCI.

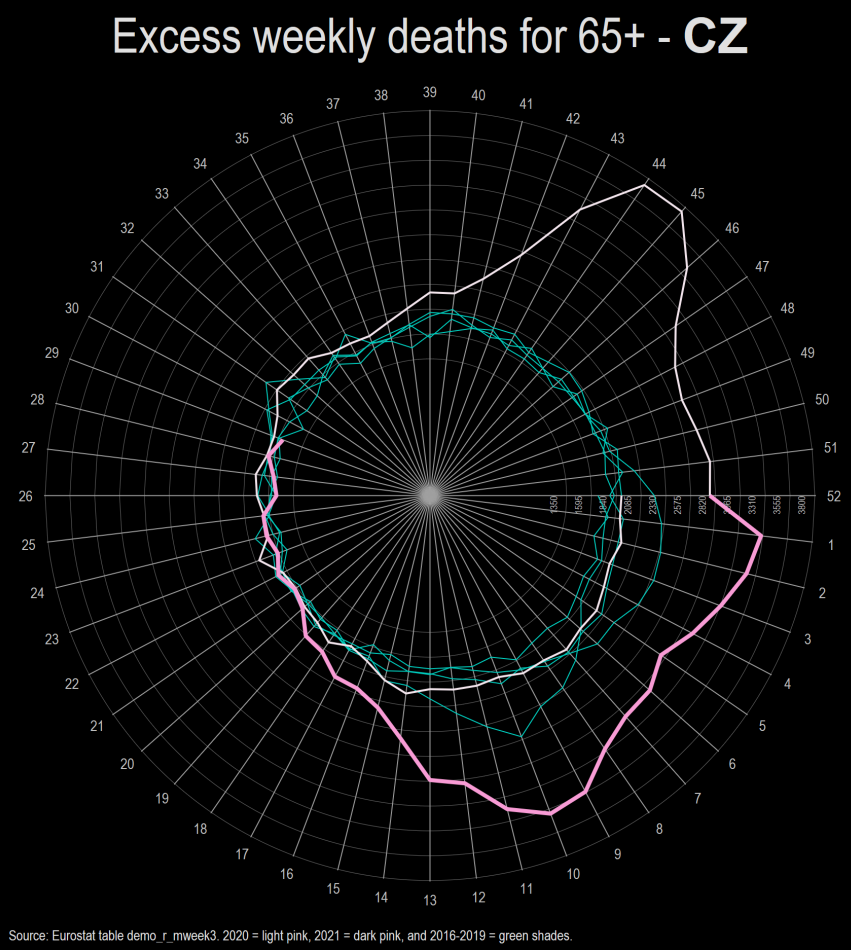

y
r
θ
r sin θ
r cos θ
x
x = r cos θ
y = r sin θ
r² = x² + y²
θ = atan2(y, x)
180 degrees = π radians

Spider plots



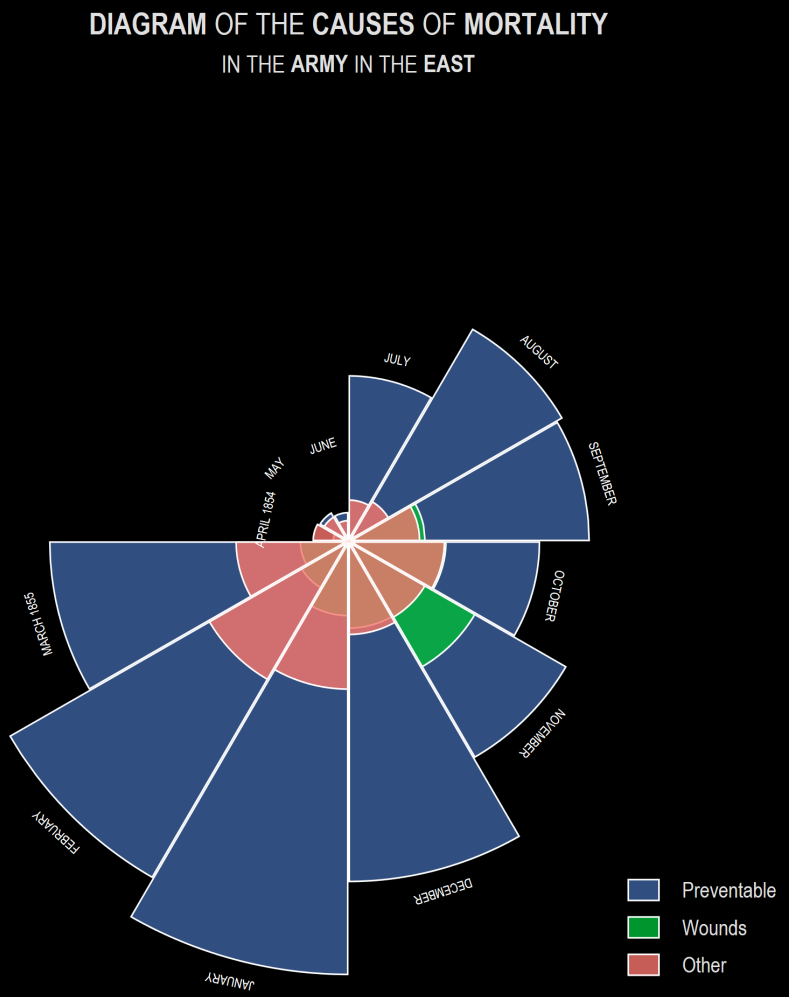
Guide: Stata graphs: Spider plots

Polar plots



Guide: Stata graphs: Polar plots

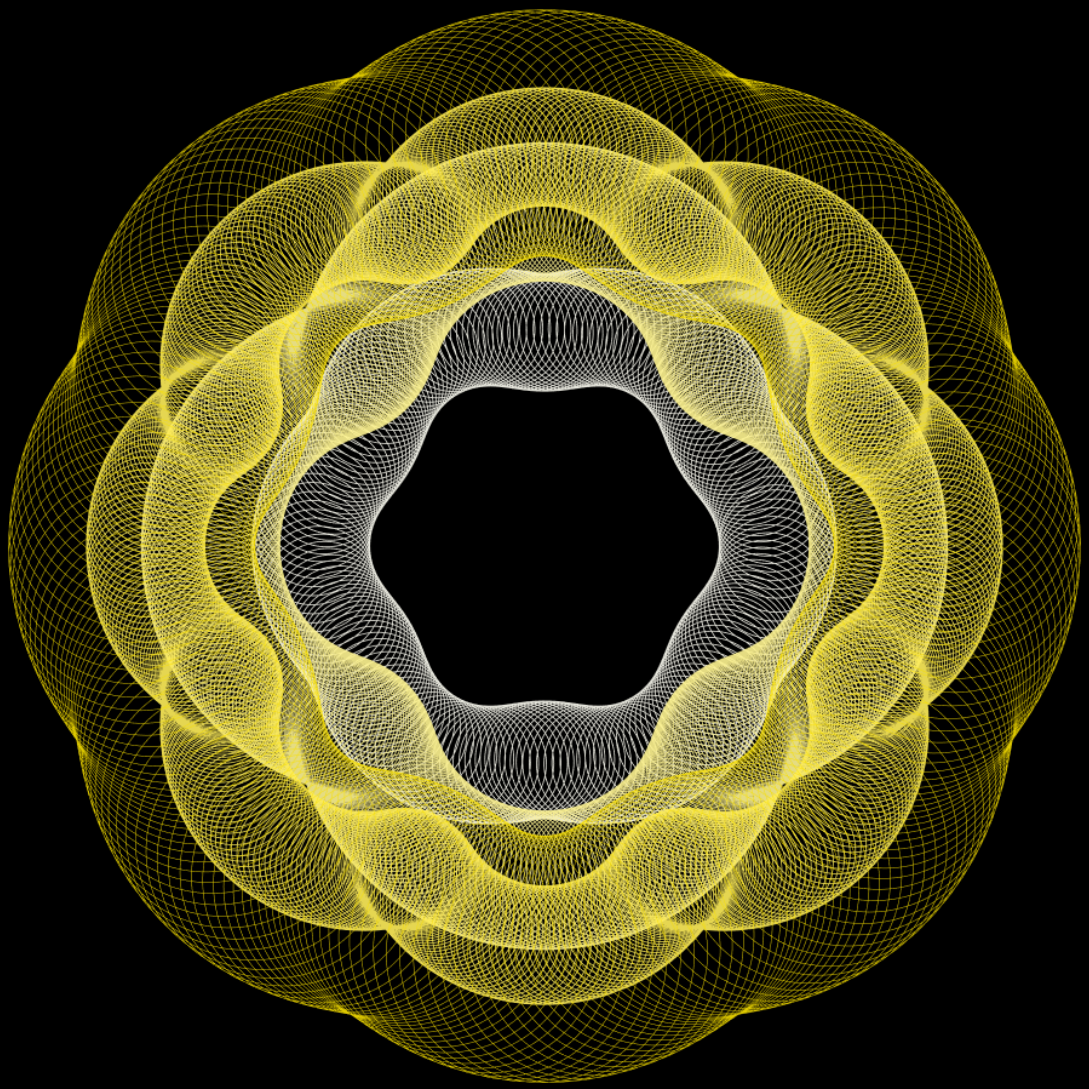
Coxcomb plots (by Florence Nightingale)



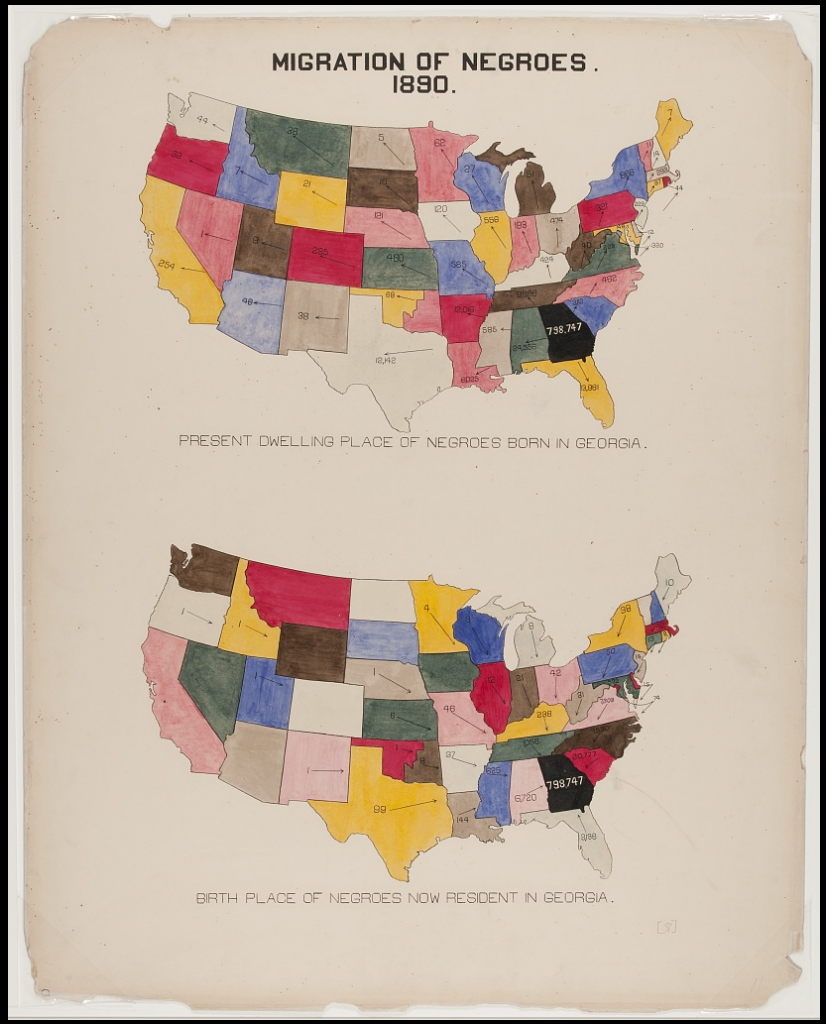
Guide: Stata graphs: Rose (Coxcomb) plots

Guide: Spiral text wrapping in Stata

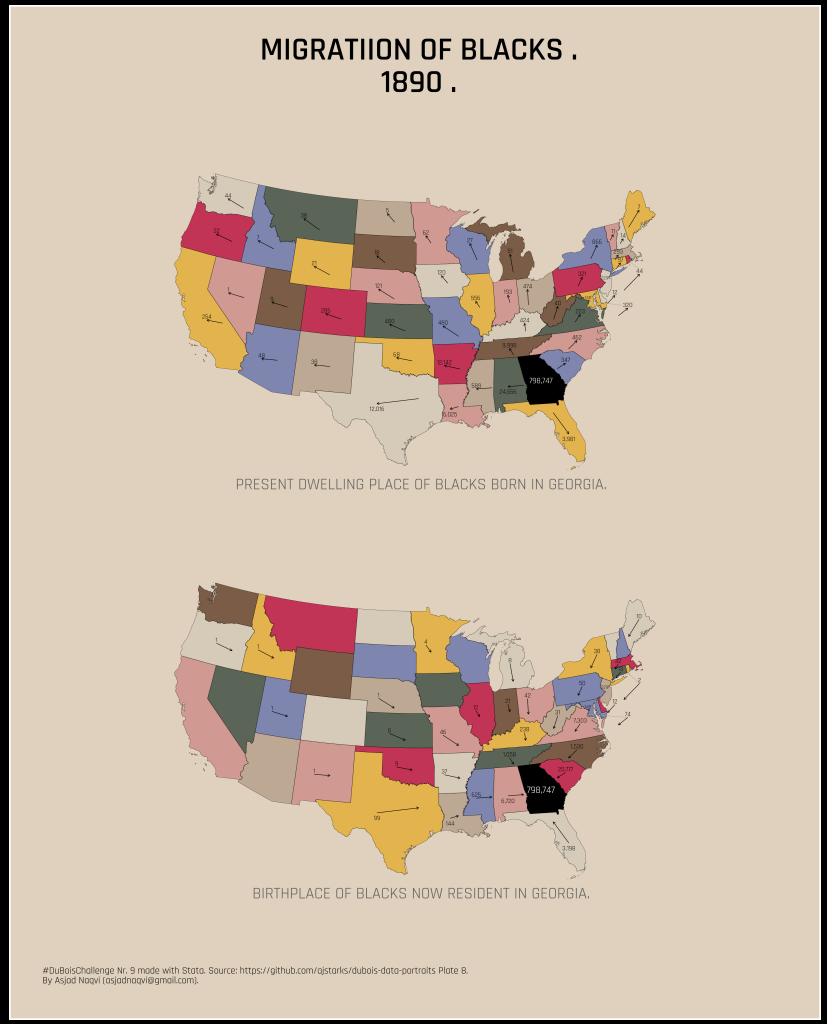
Hypotrochoids (spiralographs), Guilloché patterns



The Du Bois Challenge



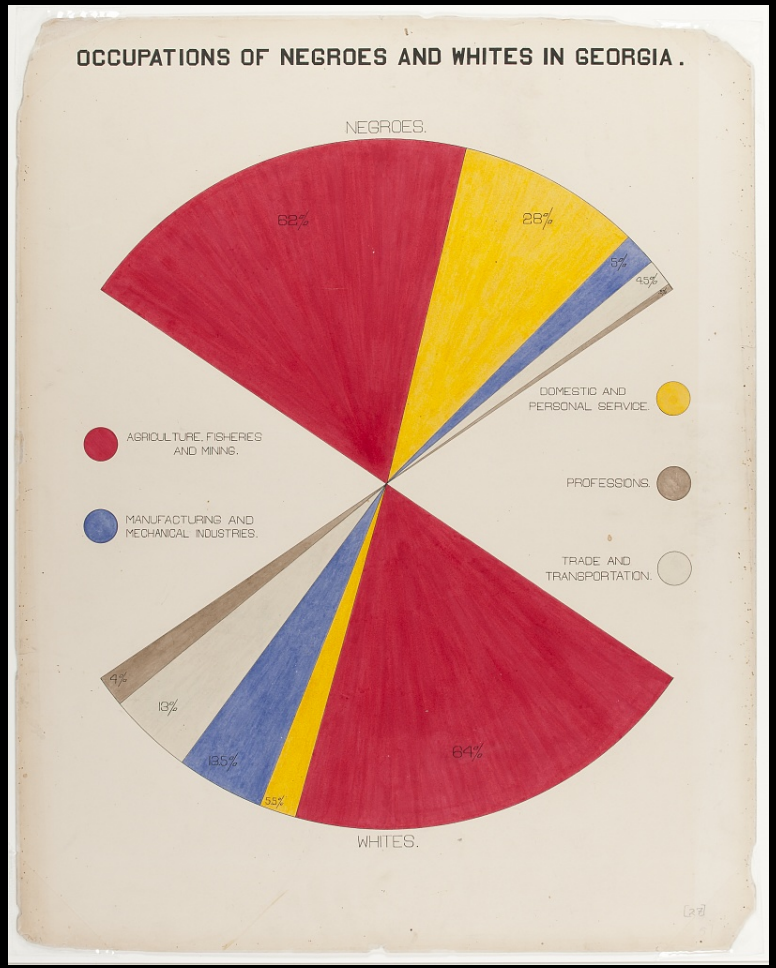
Original



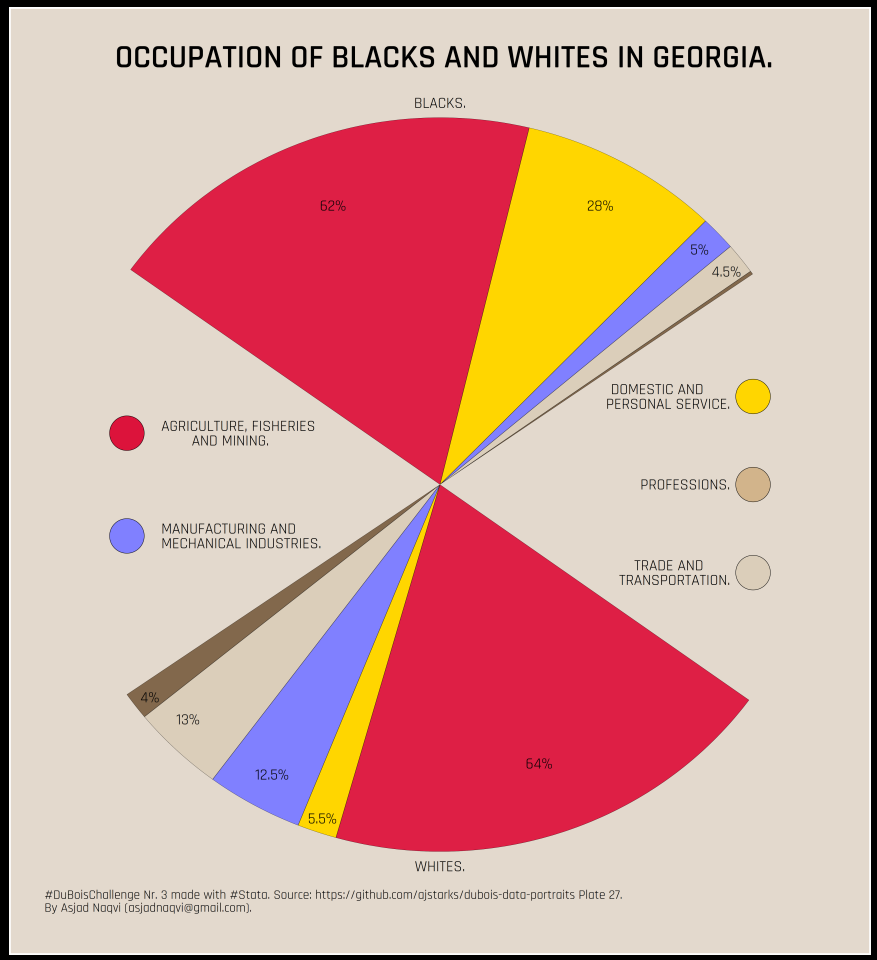
Replication

Data and code: https://github.com/asjadnaqvi/DuBoisChallenge-Stata. Explainer: The Do Bois Challenge in Stata.

The Du Bois Challenge



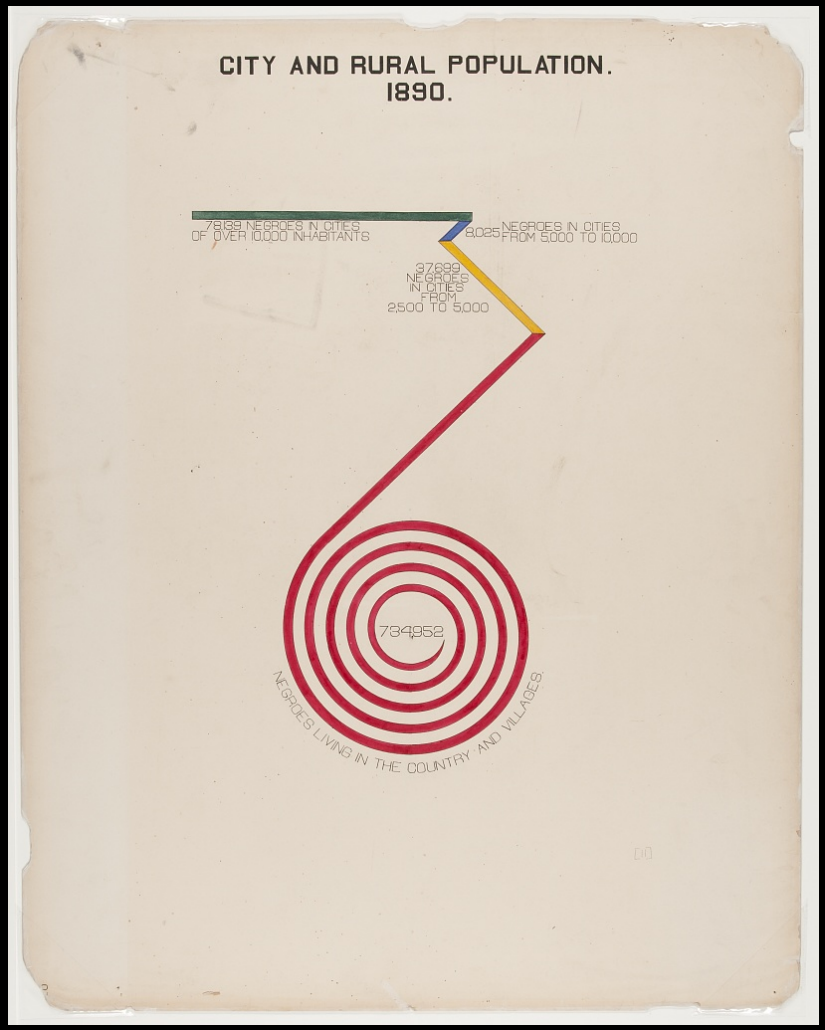
Original



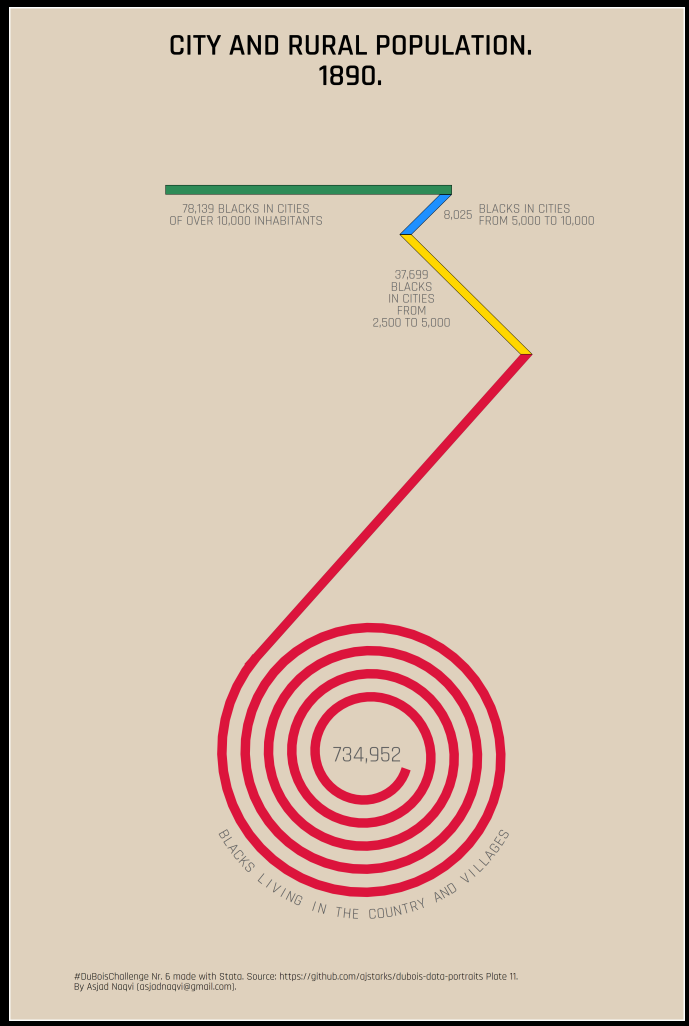
Replication

Data and code: https://github.com/asjadnaqvi/DuBoisChallenge-Stata. Explainer: The Do Bois Challenge in Stata.

The Du Bois Challenge



Original



Replication

Data and code: https://github.com/asjadnaqvi/DuBoisChallenge-Stata. Explainer: The Do Bois Challenge in Stata.

Relational and hierarchical datasets

- Both are in a network structure
- *Relational*: flows from one node to another for all layers
- *Hierarchical*: Flows are partitioned and ordered by layers

Dummy relational data

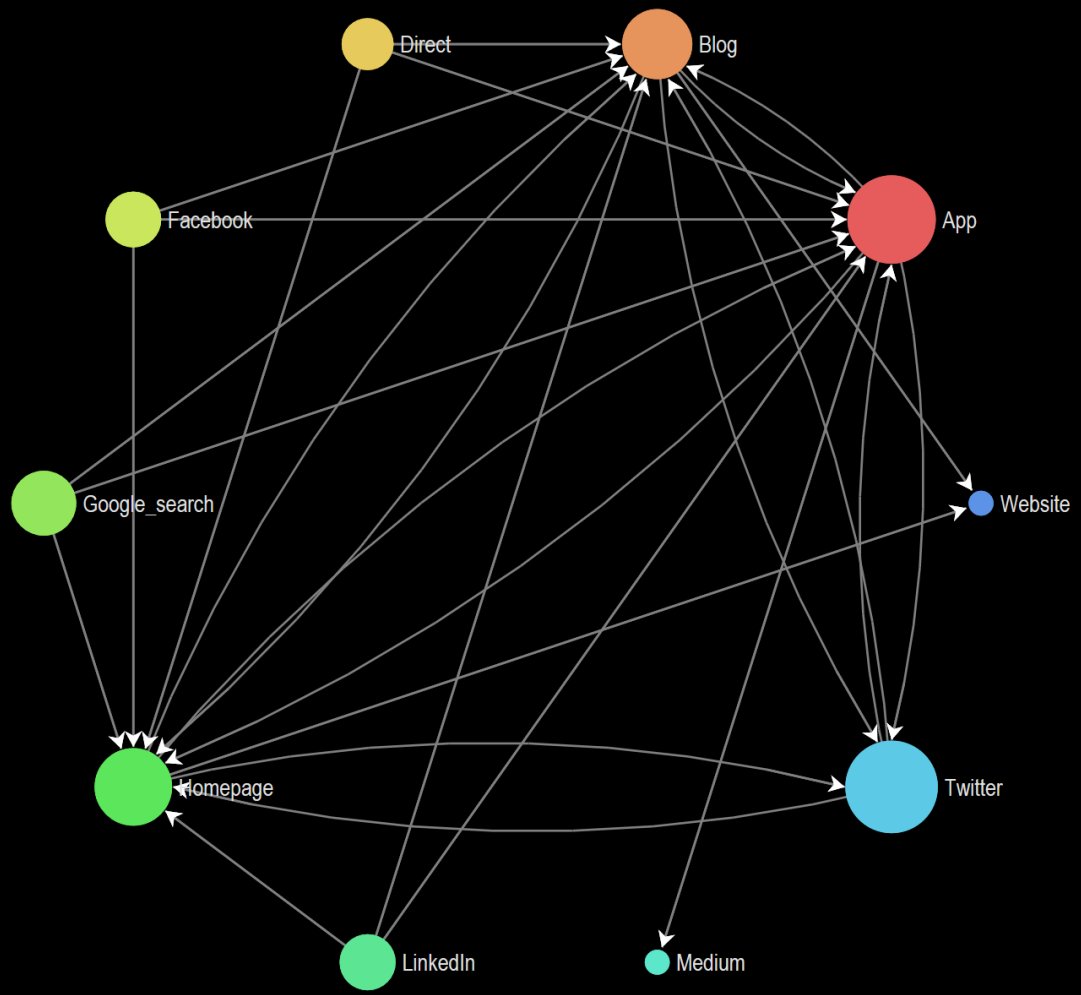
From	To	Value
App	Blog	325
App	Homepage	255
App	Medium	220
App	Twitter	180
Blog	App	144
Blog	Homepage	184
Blog	Twitter	320
Blog	Website	50
Direct	App	124
Direct	Blog	175
Direct	Homepage	119
Facebook	App	124
Facebook	Blog	167
Facebook	Homepage	185

From	To	Value
Google search	App	233
Google search	Blog	252
Google search	Homepage	131
Homepage	App	375
Homepage	Blog	287
Homepage	Twitter	50
Homepage	Website	100
LinkedIn	App	185
LinkedIn	Blog	173
LinkedIn	Homepage	123
Twitter	App	485
Twitter	Blog	341
Twitter	Homepage	219

A matrix representation

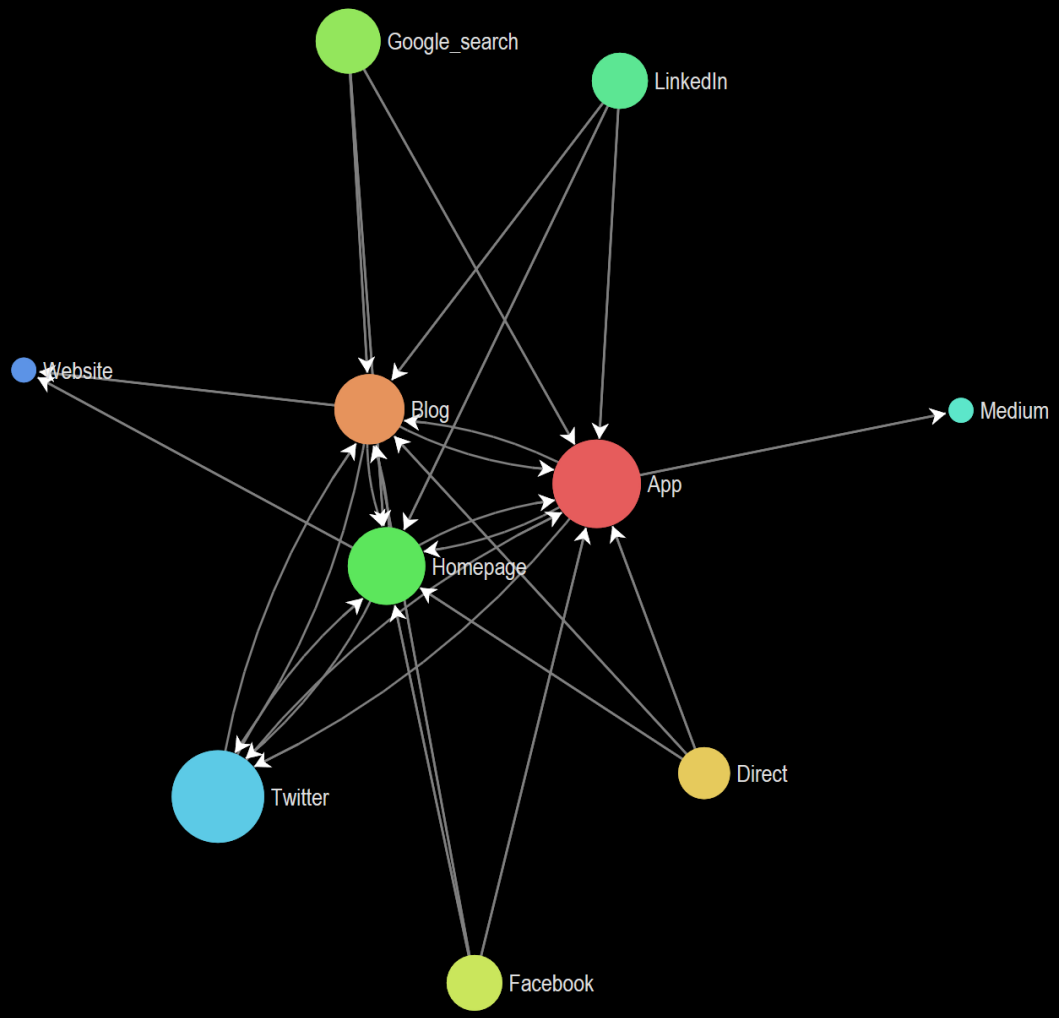
Package: `heatplot`

A network representation (circle layout)



Package: `nwcommands`

A network representation (MDS layout)



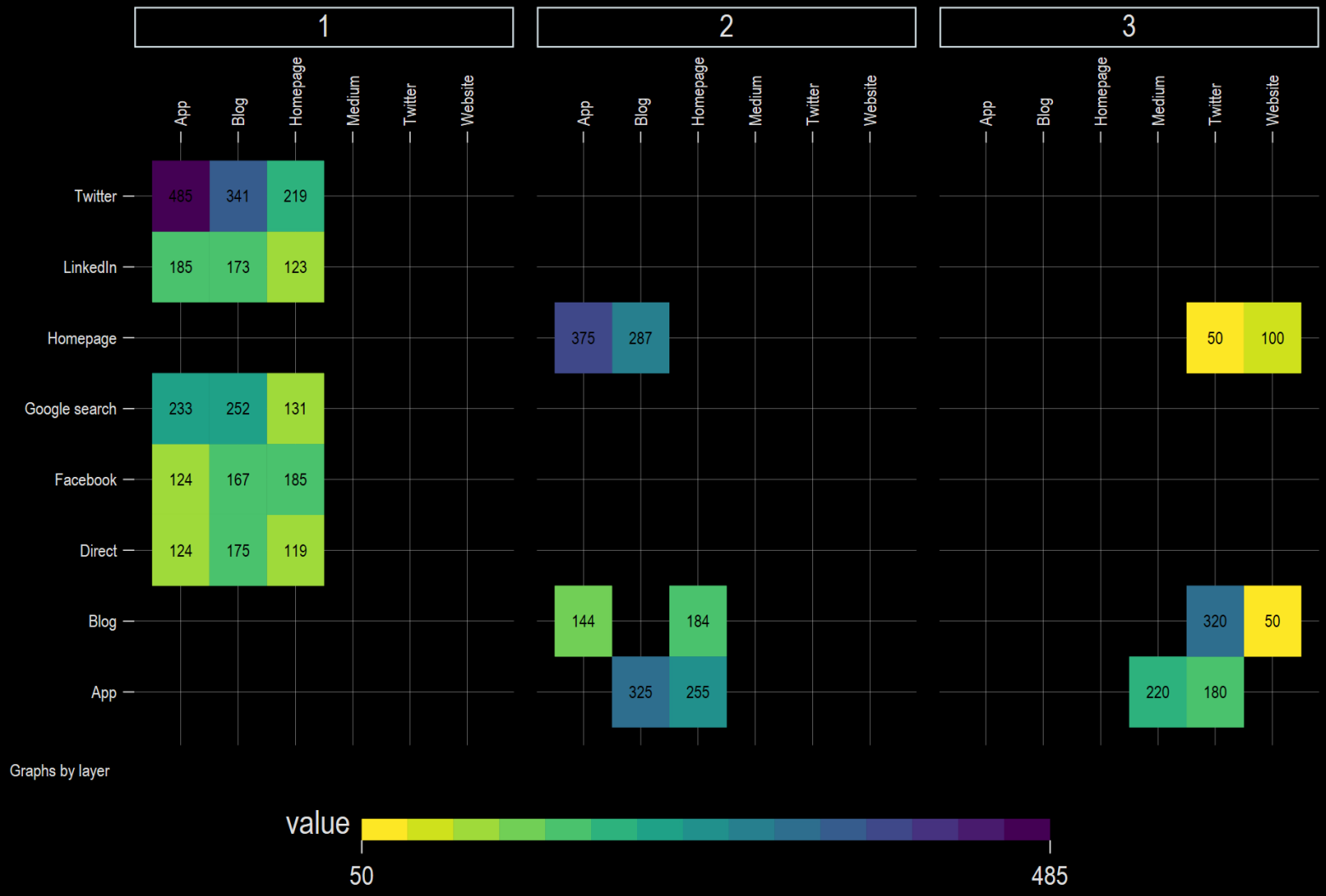
Package: `nwcommands`. MDS = Multi-Dimensional Scaling

Sample data - Hierarchical

Layer	From	To	Value
1	Twitter	Blog	341
1	Twitter	Homepage	219
1	Twitter	App	485
1	Google search	Blog	252
1	Google search	Homepage	131
1	Google search	App	233
1	Facebook	Blog	167
1	Facebook	Homepage	185
1	Facebook	App	124
1	LinkedIn	Blog	173
1	LinkedIn	Homepage	123
1	LinkedIn	App	185
1	Direct	Blog	175
1	Direct	Homepage	119
1	Direct	App	124

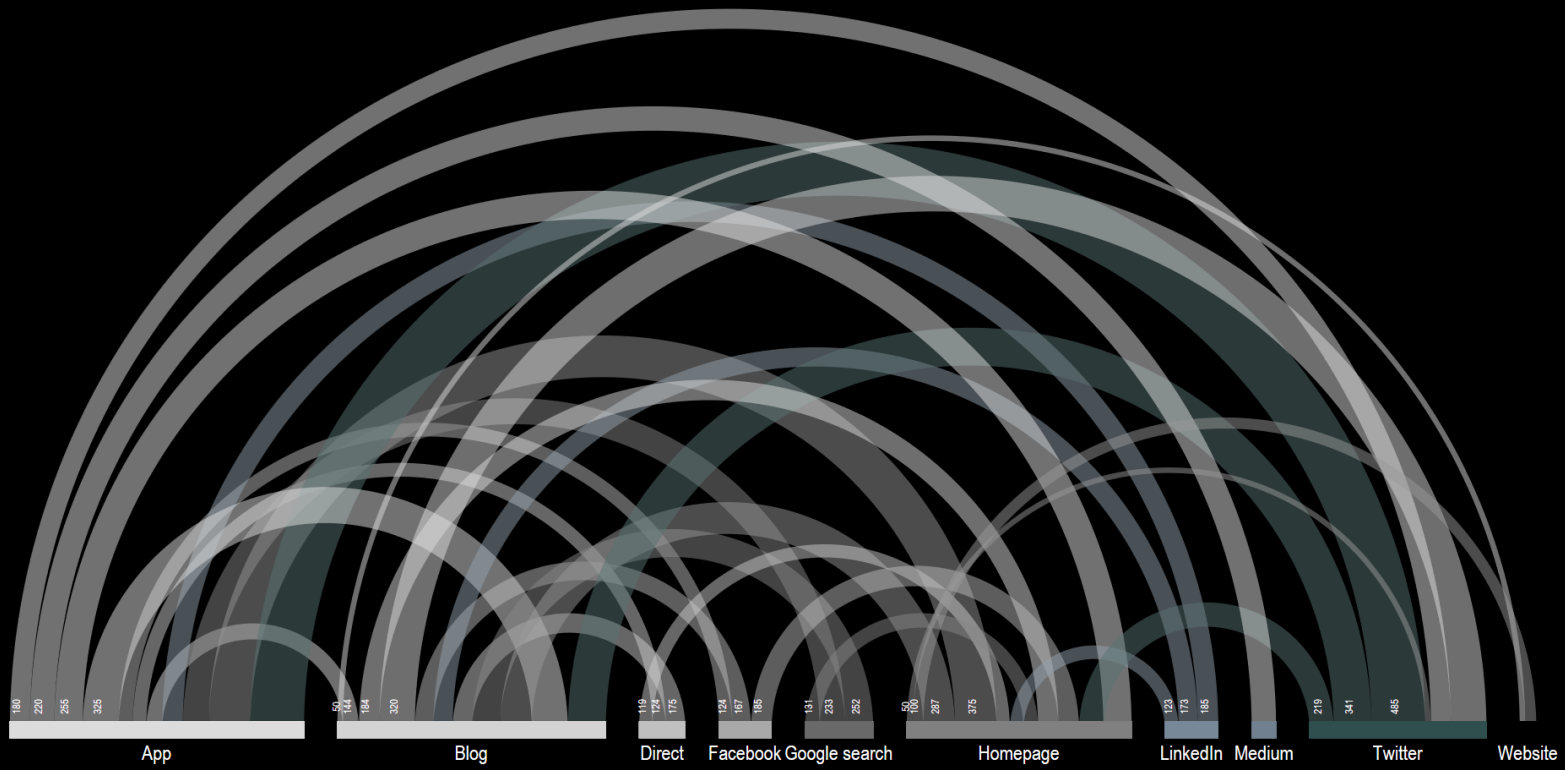
Layer	From	To	Value
2	Blog	Homepage	184
2	Blog	App	144
2	Homepage	Blog	287
2	Homepage	App	375
2	App	Homepage	255
2	App	Blog	325
3	Homepage	Website	100
3	Homepage	Twitter	50
3	App	Twitter	180
3	App	Medium	220
3	Blog	Twitter	320
3	Blog	Website	50

A matrix representation

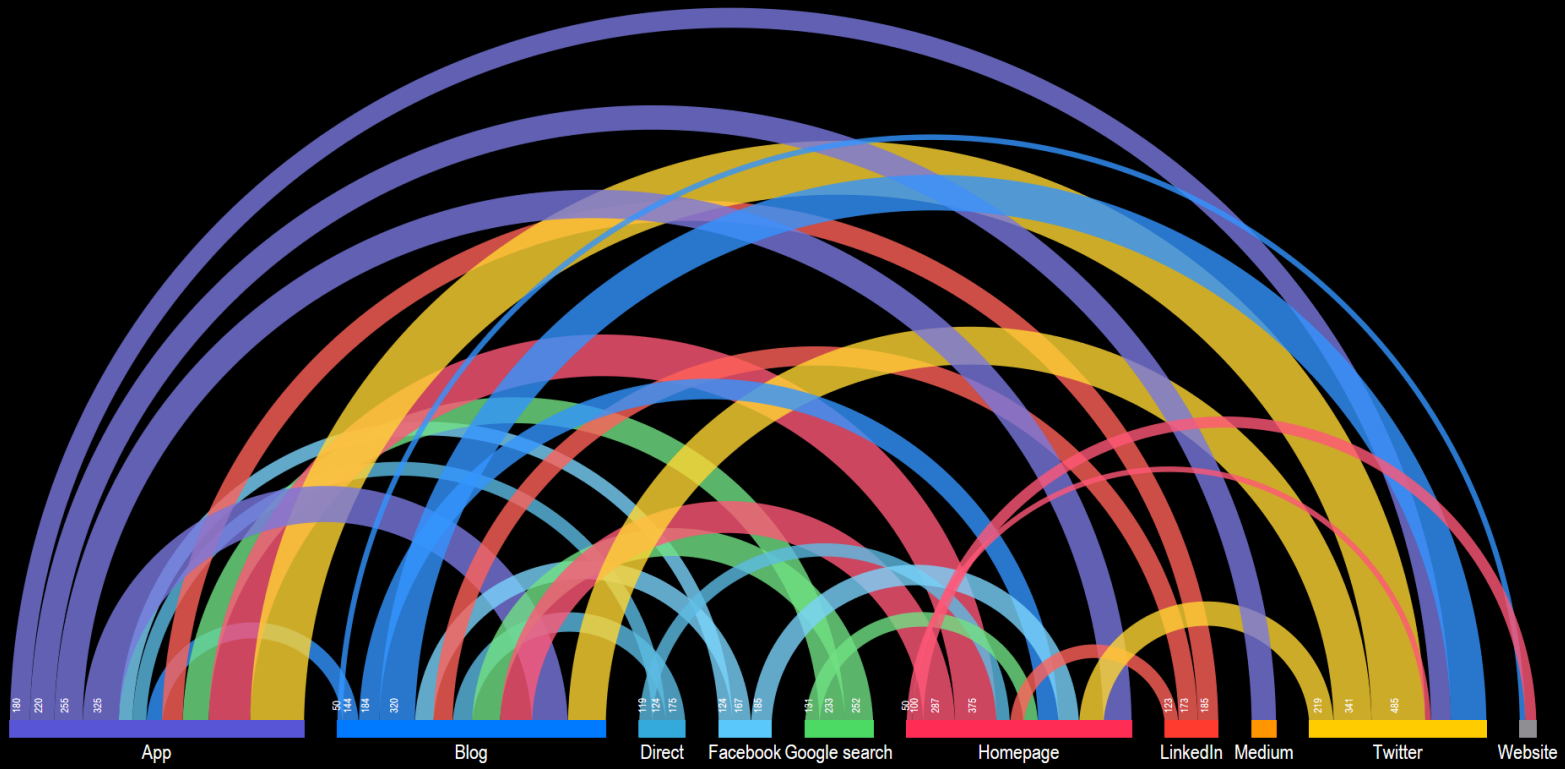


Package: `heatplot`

Arc


180
220
255
325
50
144
184
320
119
124
175
124
167
185
131
233
252
50
180
287
375
123
173
185
219
341
485
App
Blog
Direct
Facebook
Google search
Homepage
LinkedIn
Medium
Twitter
Website

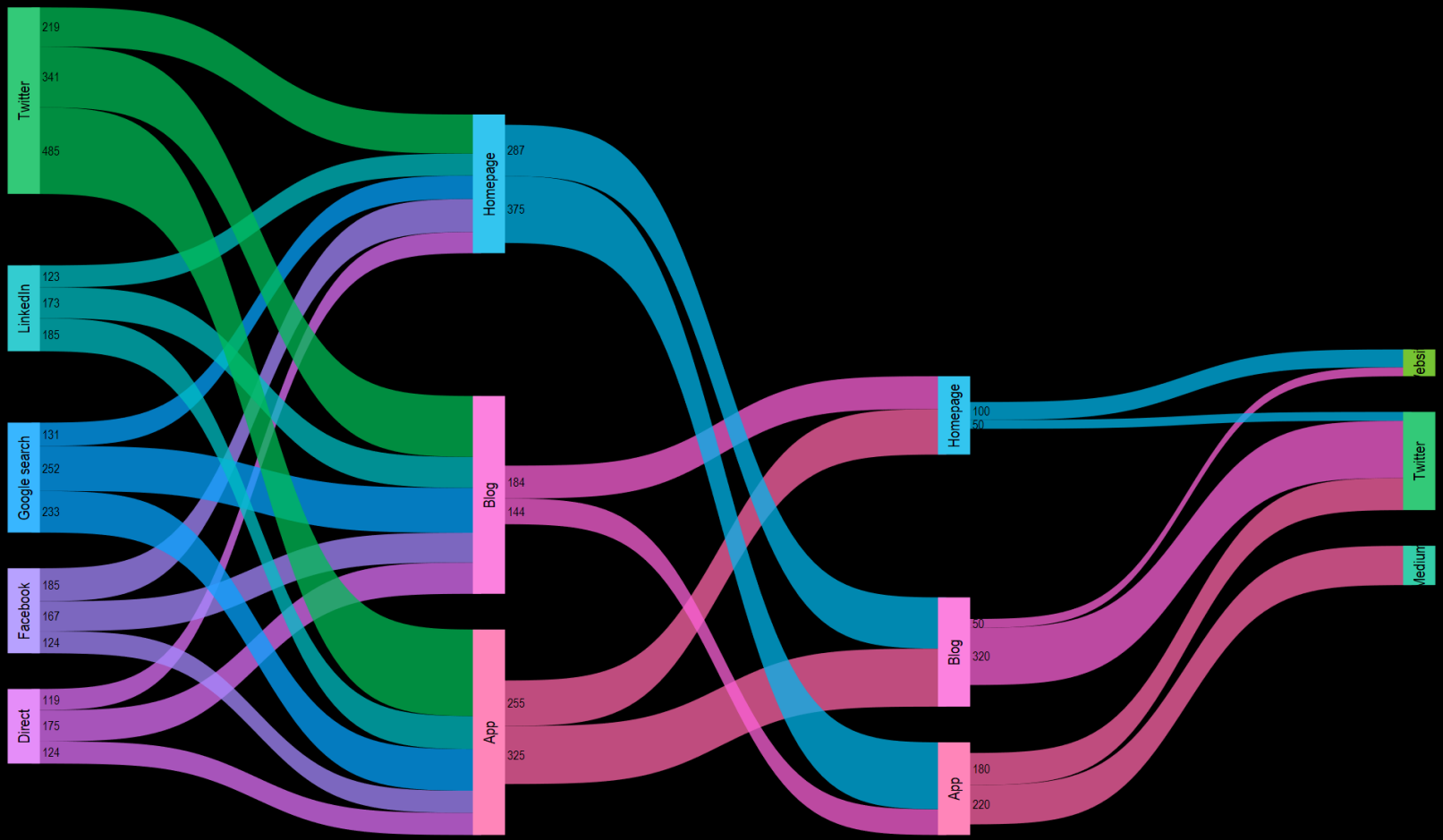
Arc


180
220
255
325
50
144
184
320
119
124
175
124
167
185
141
233
252
50
150
287
375
123
173
185
219
341
485
App
Blog
Direct
Facebook
Google search
Homepage
LinkedIn
Medium
Twitter
Website

Chord


Blog
App
Direct
Facebook
Website
Google search
Twitter
Homepage
Medium
Linkedin

Sankey


Twitter
219
341
485
LinkedIn
123
173
185
Google search
131
252
233
Facebook
185
167
124
Direct
119
175
124
Homepage
287
375
Blog
184
144
App
255
325
Homepage
100
50
Blog
50
320
App
180
220
ebsi
Twitter
Medium

Why do all of this in Stata?

- One can pass the DataViz stuff to more established languages:
  - R using `rcall` (Haghish 2019) or Python using PyStata
  - But these should be viewed more as stop-gap options
  - Develop and internalize the dataviz option within Stata
  - Why?

Why do all of this in Stata?

- One can pass the DataViz stuff to more established languages:
  - R using `rcall` (Haghish 2019) or Python using PyStata
  - But these should be viewed more as stop-gap options
  - Develop and internalize the dataviz option within Stata
  - Why?
- Easy to integrate with the Stata code workflow
- Reduce the language burden on end-users
- Time investment? One-time (high) sunk costs *or*
- Some of the `#StataViz` shown above can be easily packaged for end-users

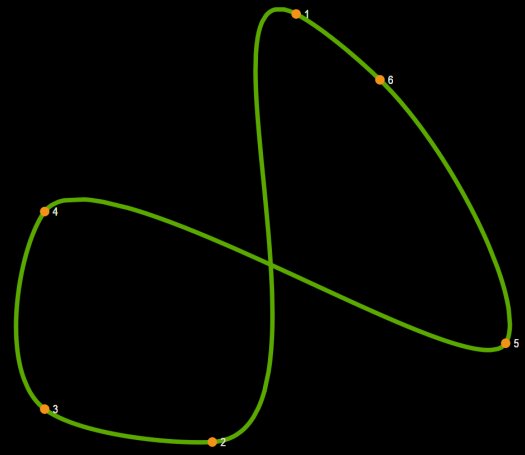
Core features that can take Stata to the next level (wish list)

- Ability to control marker size scaling
- Line weights (like marker weights)
- Allow angles/sizes/colors to be read from variables (like mlabels)
- Color/alpha scaling for markers/lines
- Color gradients for lines/areas
- Ability to add custom markers (or increase marker pool)
- Ability to added colored text in graphs
- Ability to add images in figures
- Ability to read images (pixel data/colors)
- Ability to draw/add/subtract areas between two or more functions
- Line curvature

Thank you!

For more #Stataviz:

- The Stata Guide on Medium
- GitHub: asjadnaqvi
- Twitter: @AsjadNaqvi



Catmull-Rom spline

A lot more is in development for next year