Week 4, Day 1

Sarah Moore

2022-10-11

Agenda for today

- Debrief of your first data visualization blog.
- Moving away from abstract ideas of data viz and just the technical emphasis to some concrete ideas of real design *principles*.
- Thus far, we have focused on some basic choices in terms of the tools that you have at your disposal. Today we'll instead focus on thinking of the thought process that precedes declaring design.

Moving on to principles

- Claus O. Wilke's Fundamentals of Data Visualization
- Assigned for this week and next to read around in chapters 2-25, but the whole book is really readable.
- I am using this book to frame today's lecture and I think much of this information underlies the things we have already talked about.

Our data today

Martinez i Coma, Ferran; Lee Morgenbesser, 2020, "Replication Data for Election turnout in authoritarian regimes", https://doi.org/10.7910/DVN/3KSEUO, Harvard Dataverse, V2.

Martinez i Coma, Ferran; Morgenbesser, Lee. 2020. "Election turnout in authoritarian regimes", Electoral Studies, Volume 68. https://doi.org/10.1016/j.electstud.2020.102222.

Coming up with a basic "directory"[1]

Type of Information	**Suggested Visualization**
Amounts	bars, dots, heatmap
Distributions	histogram, density plot, qq-plot, boxplot, violin plot, strip chart
Proportions	bars, density plot, mosaiac plot, treemap, parallel set
Relations	scatterplot, bubblechart, slope-graph, contour plot, bins, correl-ogram, line graph

- Each of these suggestions can be further modified to be inclusive of more information, by stacking, grouping, faceting, adding information on uncertainty, or changing coordinate systems.

[1] Adapted from Wilke ch. 5

Choosing an axis

- Cartesian coordinate system, standard 2-dimensional x-y, with units spaced uniformly across the grid, *usually the default and implicit choice.*
- Sometimes we will need to modify the standard (0,0) starting point. . . but the scale itself is still on the same linear basis of the Cartesian system.
- A nonlinear axis, such as a root scale or a log scale, entails an axis where the difference between each unit is not uniform. Data plotted on a non-linear scale would entail the same interpretation as transformed data on a linear scale. However, in most instances, the change to a nonlinear scale may not be advisable for lay audiences.

Curved Axes

- A polar coordinate system entails a curved axis (think pie graph!). Rather than specifying (x,y) values to occupy a coordinate space, a polar system specifies an angle and radial distance from the origin.
- Geospatial data also requires a curved axis via plotting by latitude and longitude (or other standardized measures that ultimately rely on the geospatial coordinate system). Luckily, for geospatial data there are many ways to offload burden of knowing exactly *how* to plot on a specific mapping projection. Instead, the question is *what tool to use*.

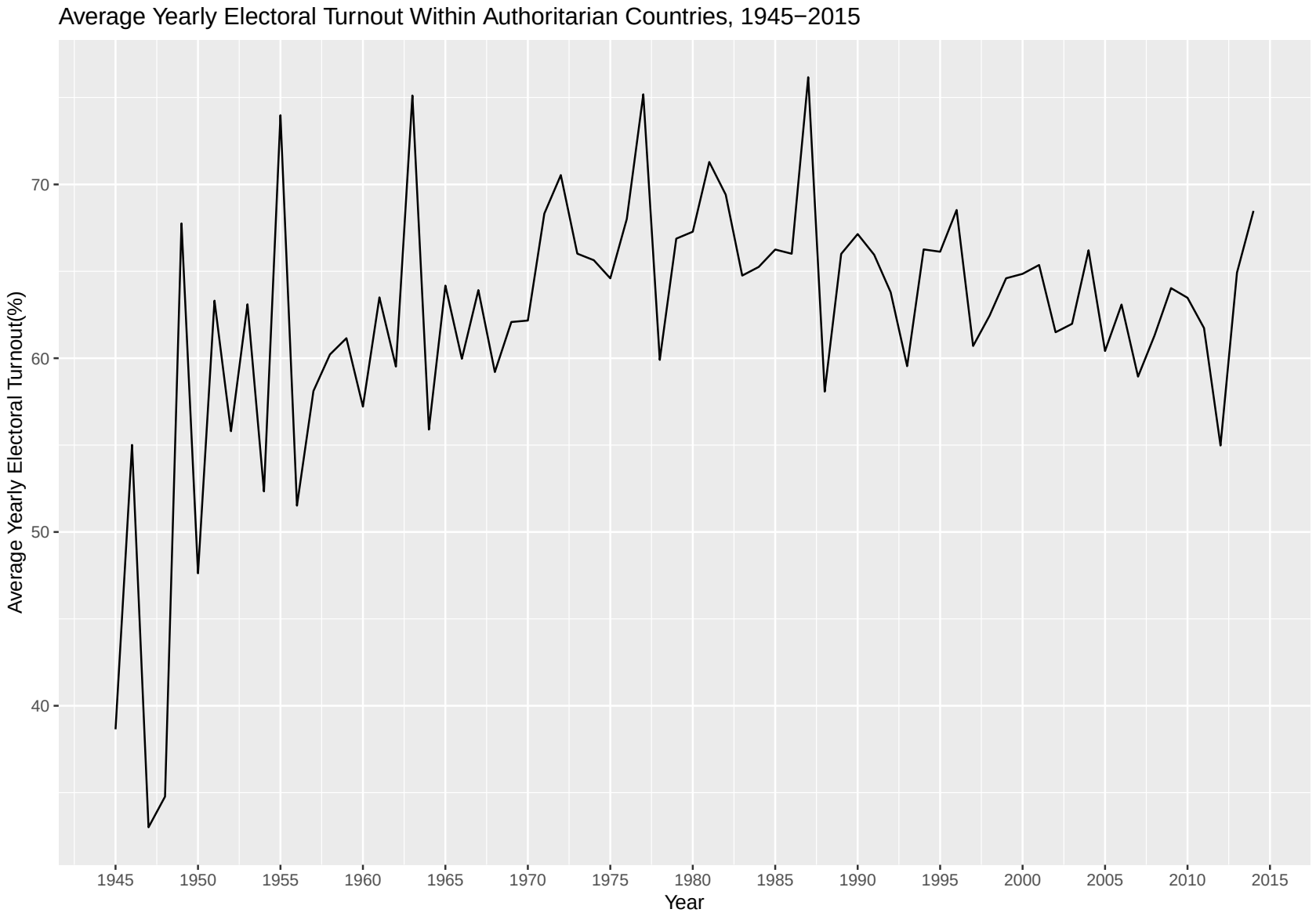
Beyond the two-dimensional space

- There are ways to map and visualize multi-dimensional relationships between data. For example, with contour densities between variables x, y, z. But maybe not the best case if we can confine our visualizations to the 2-d space?

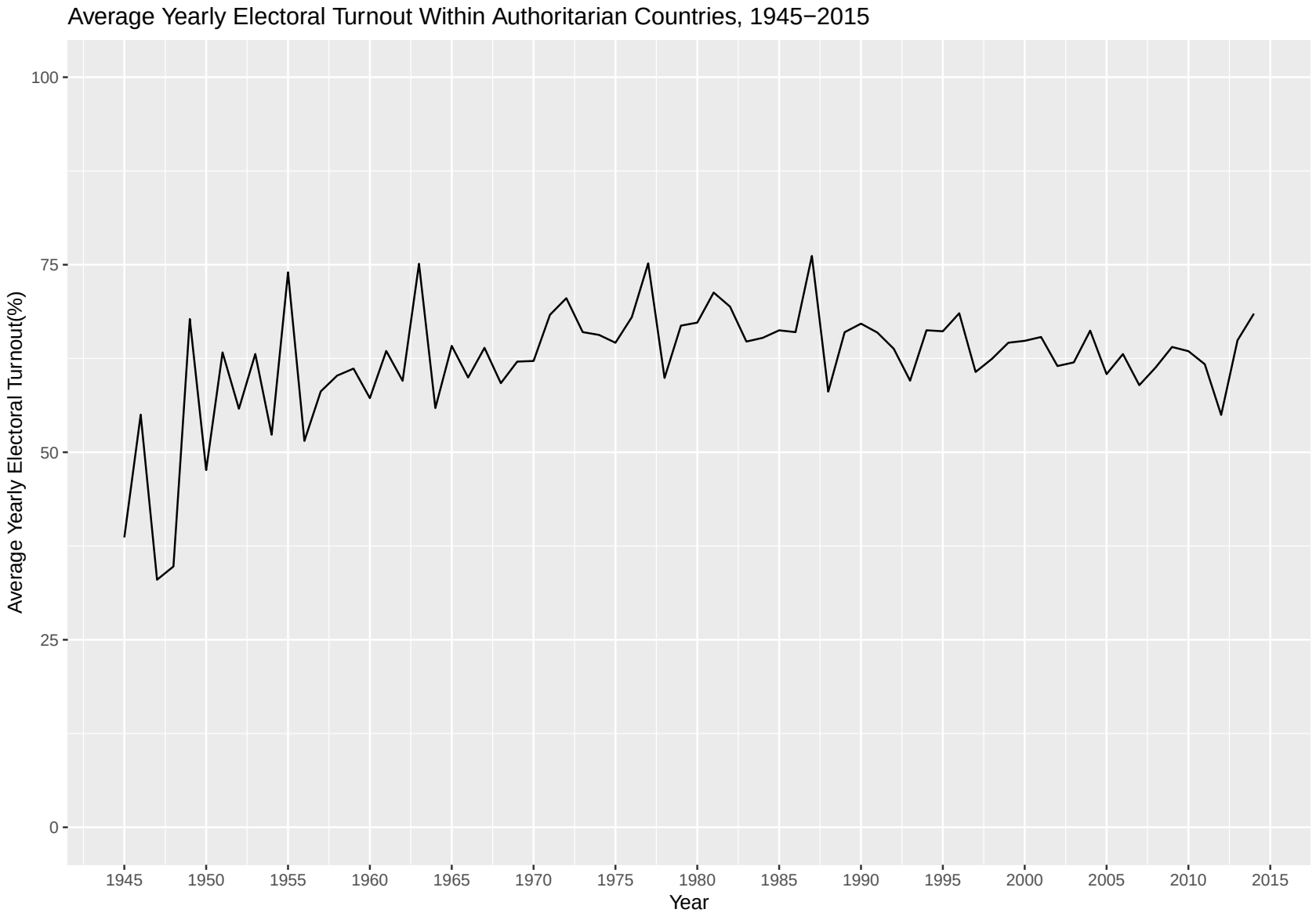
Once we have a choice of axis + geometric function...

- What do we do? What are our limits?
- Implement the *principle of proportional ink*: **the size of the ink representing data points need to be *proportional* to the data values they represent**.
  - Where ink refers to whatever graphical element is being used to represent the data point mapped.
- Ultimately, this refers to coherence and consistency between what we are visualizing relative to what we are representing.

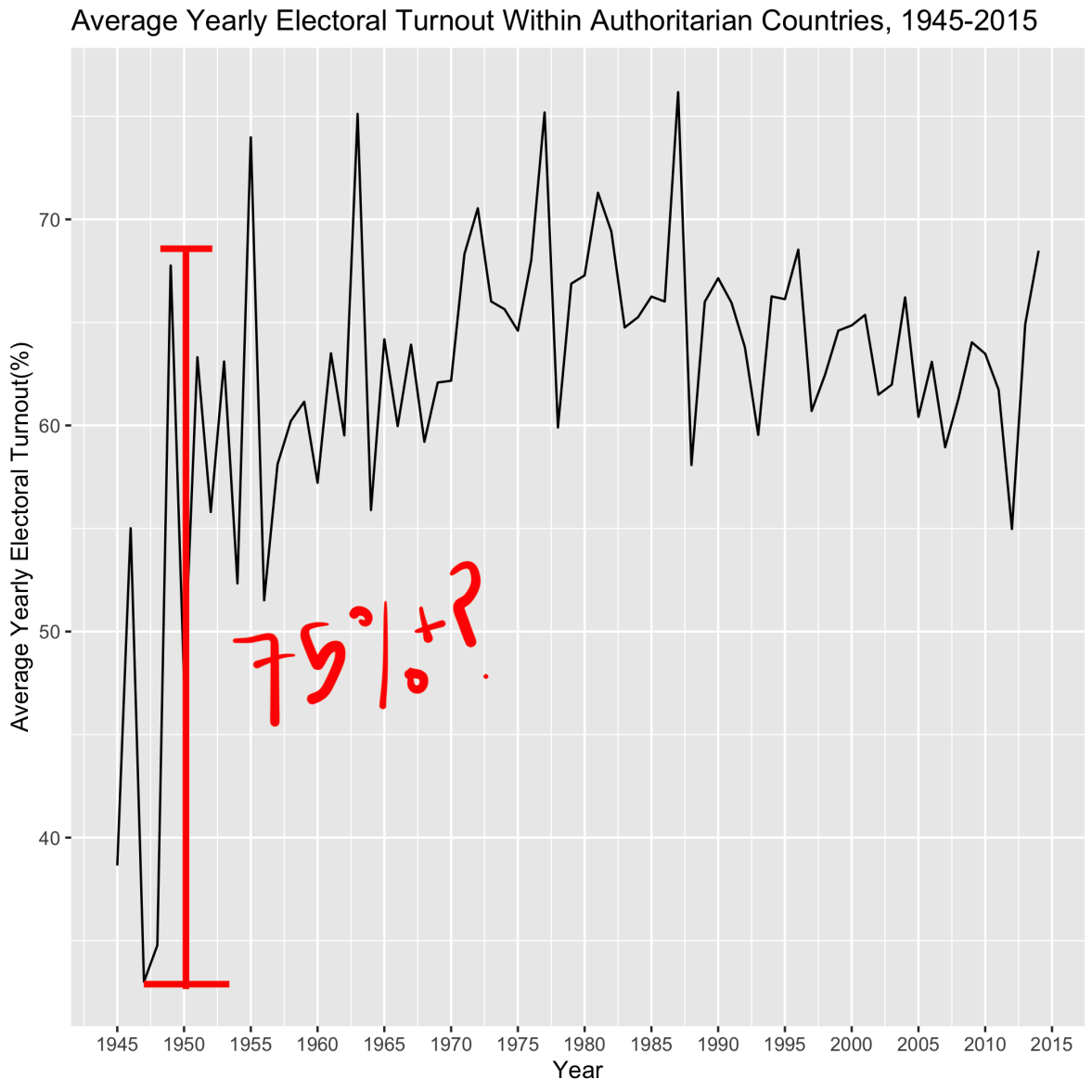
Consider the y-axis between this:



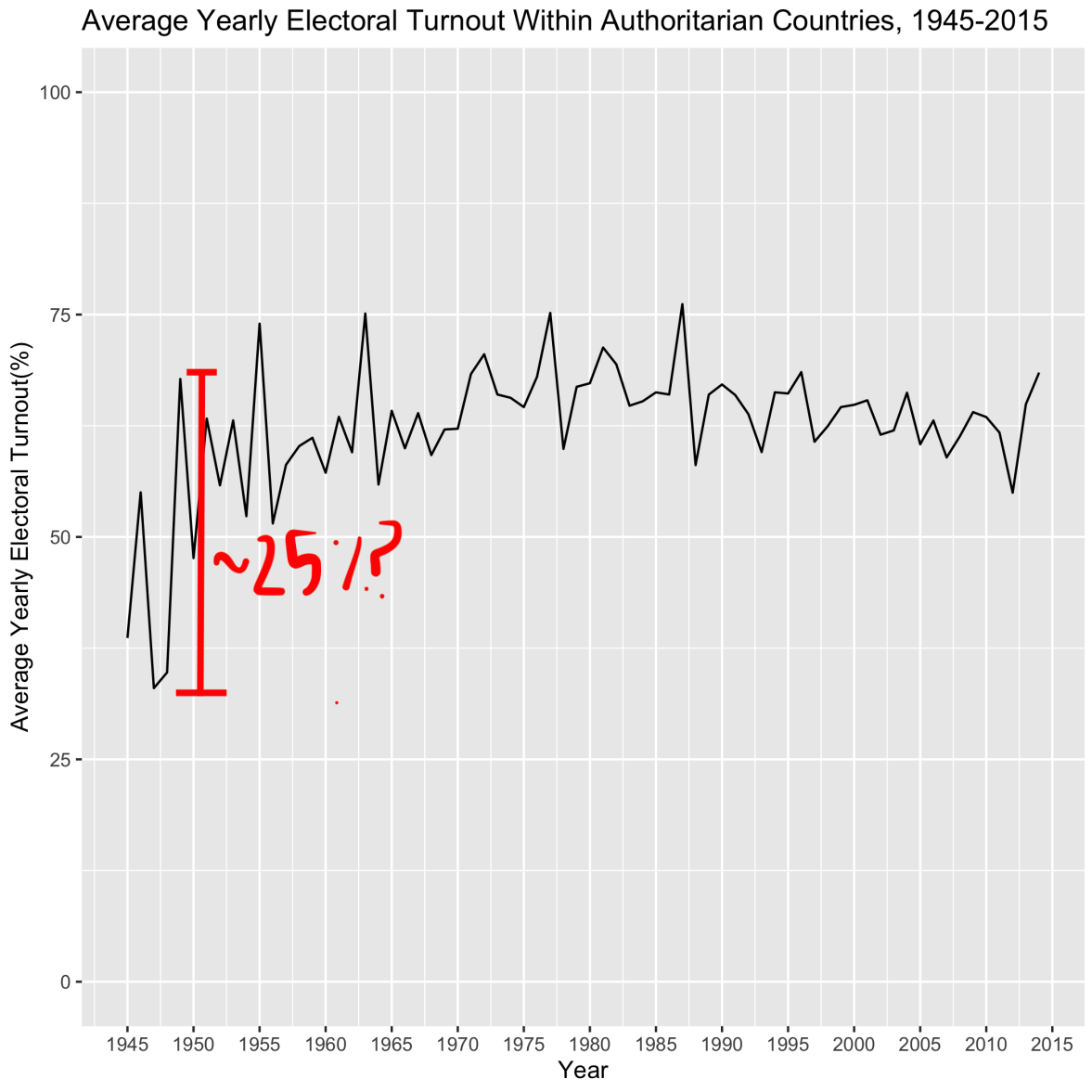
And this:


Average Yearly Electoral Turnout Within Authoritarian Countries, 1945−2015
Average Yearly Electoral Turnout(%)
100
75
50
25
0
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
Year

How does this violate this principle?



How does this violate this principle?



Some general rules:

- So if we expect our data to start at 0, then the scale should reflect that 0. *BUT* a lot of social data does not have a *natural zero*. For example, something on a scale from 1 to 6, ACT scores, other sorts of standardized measures.
- Treemaps, mosaic plots, and pie graphs don't have a typical axes. Instead the surface area of these types of plots accounts for the "amount" of data that they represent. People are actually not so great at guessing angles. So, these plots are often not *recommended*. but if you find them particularly necessary, you can always add text labels to clarify the amounts that you want to represent.

Similarly. . .

- We want to find ways to be transparent of overlapping data and the extent of coverage of our data.

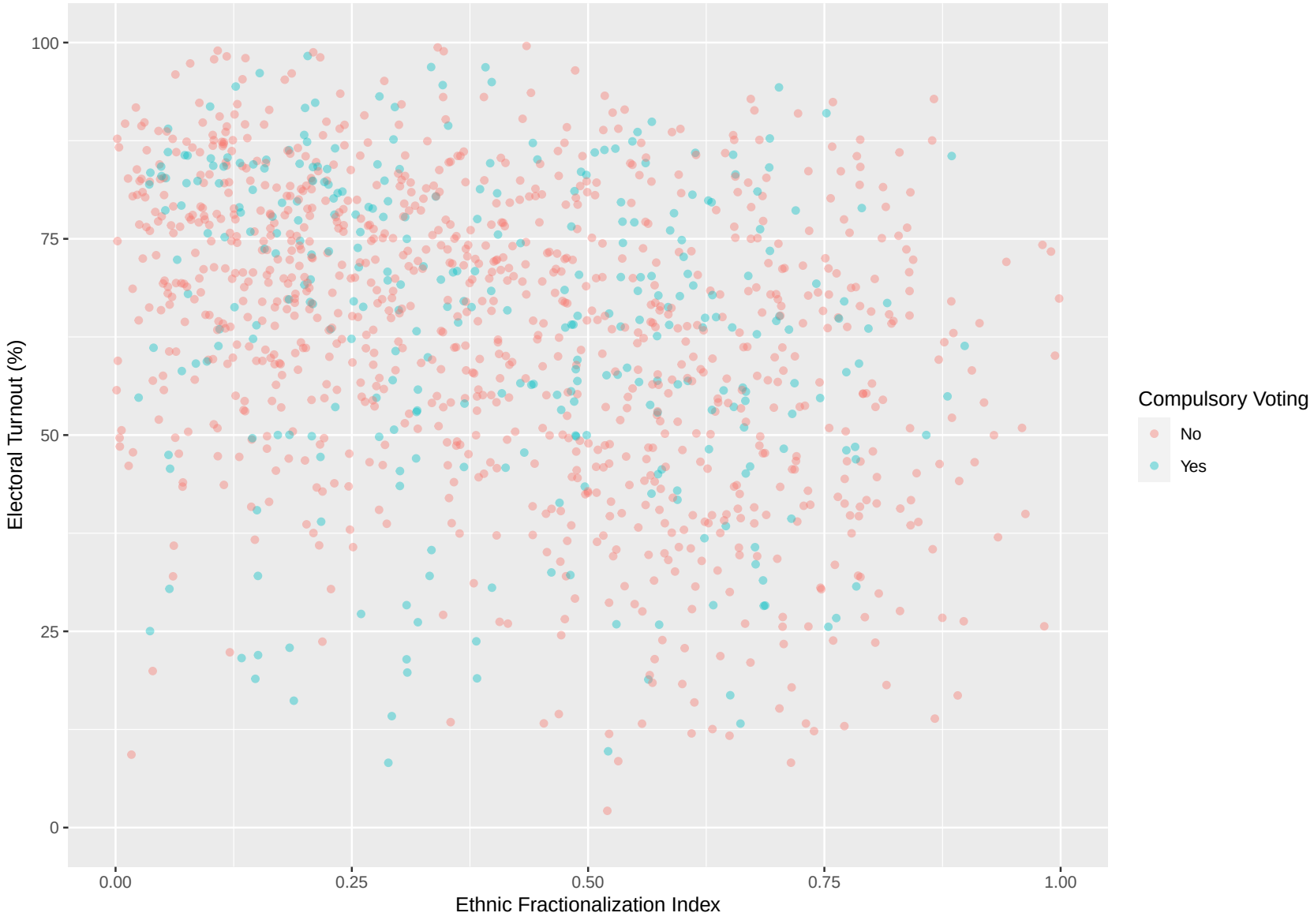
Take this plot for example

What's the problem with this?

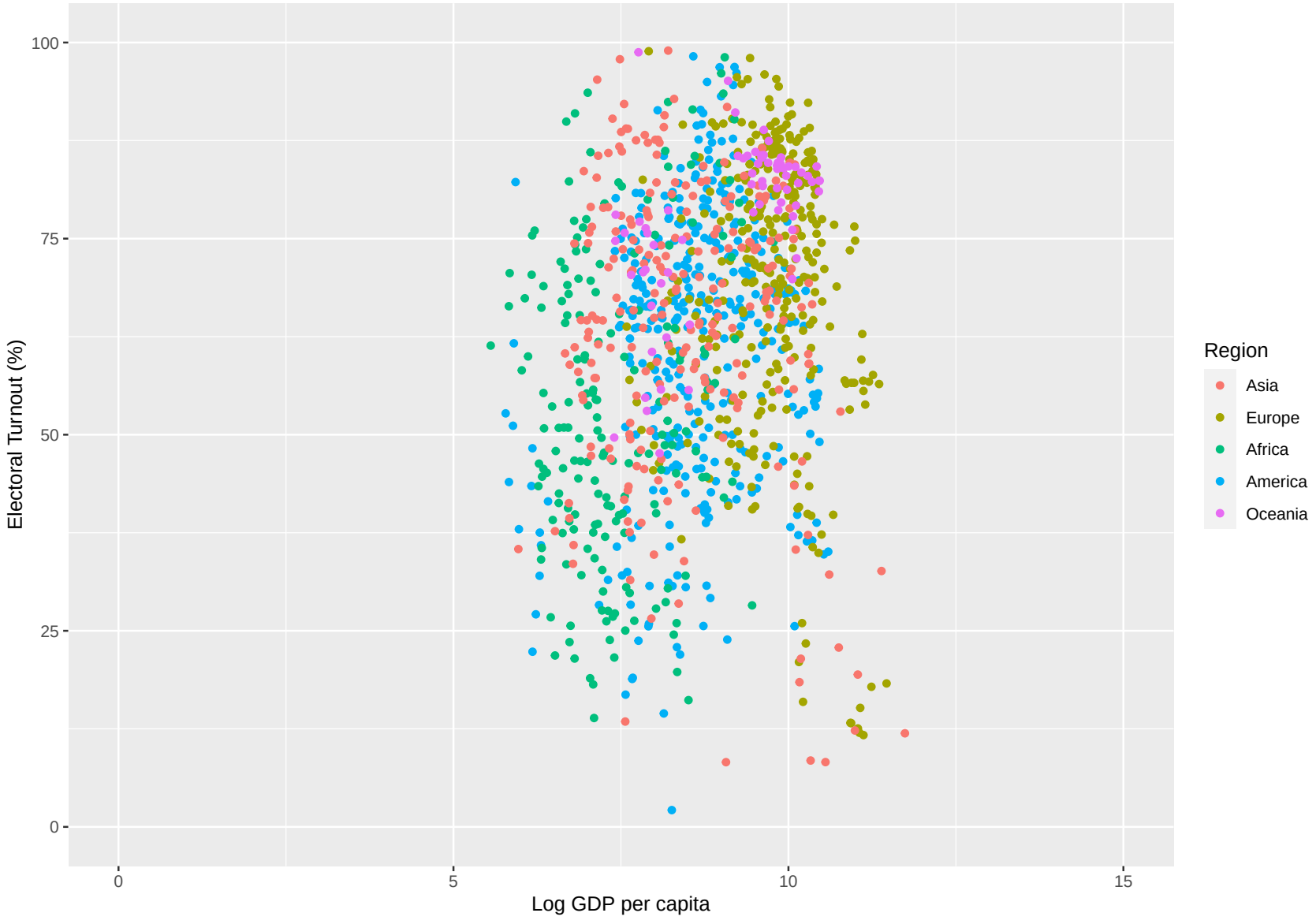
What's the problem with this?

We can see there are a bunch of dots, but can't *truly* discern the density where there are clusters. This happens a lot with categorical and rounded data.

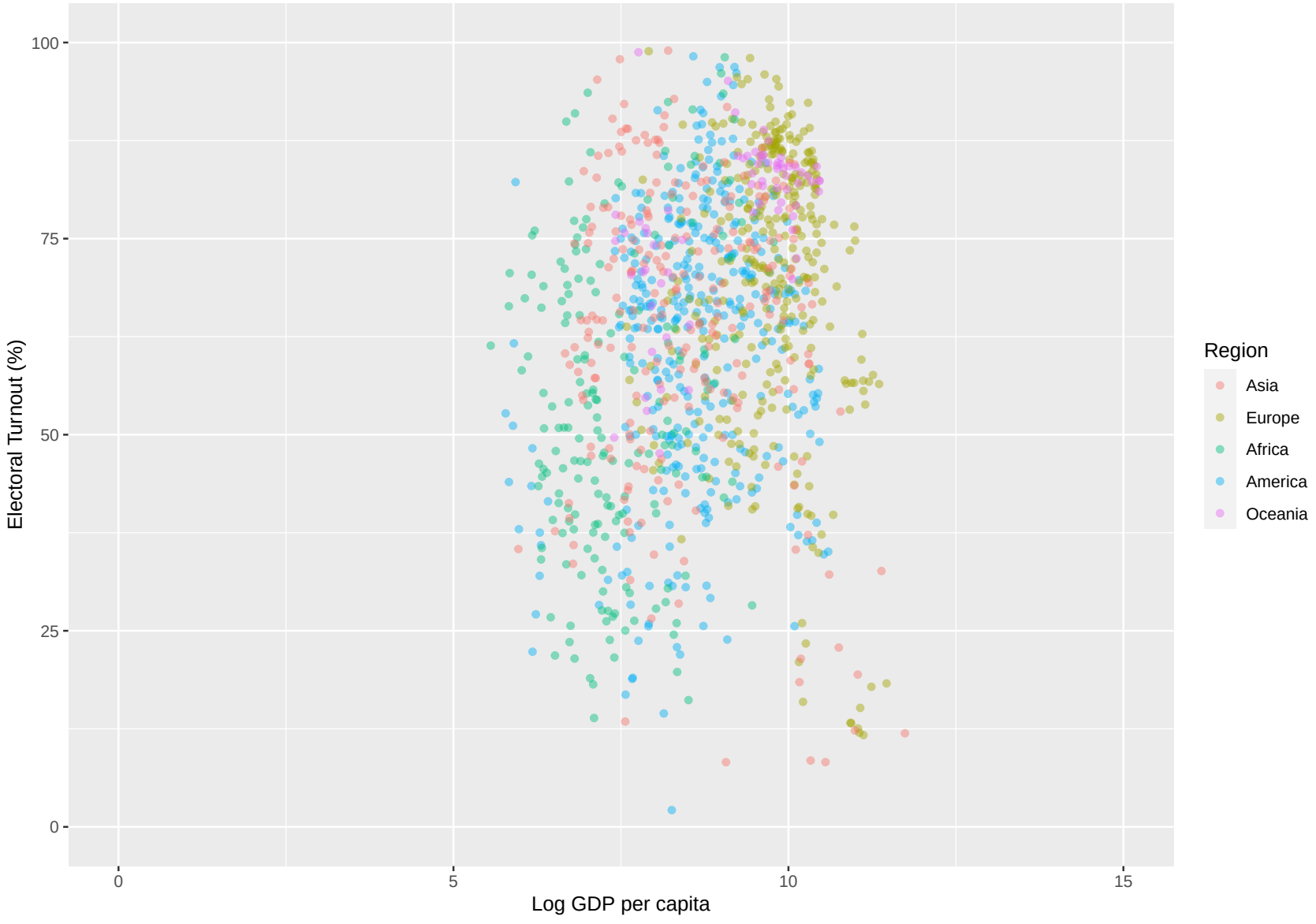
Add some jitter and change the transparency of the points



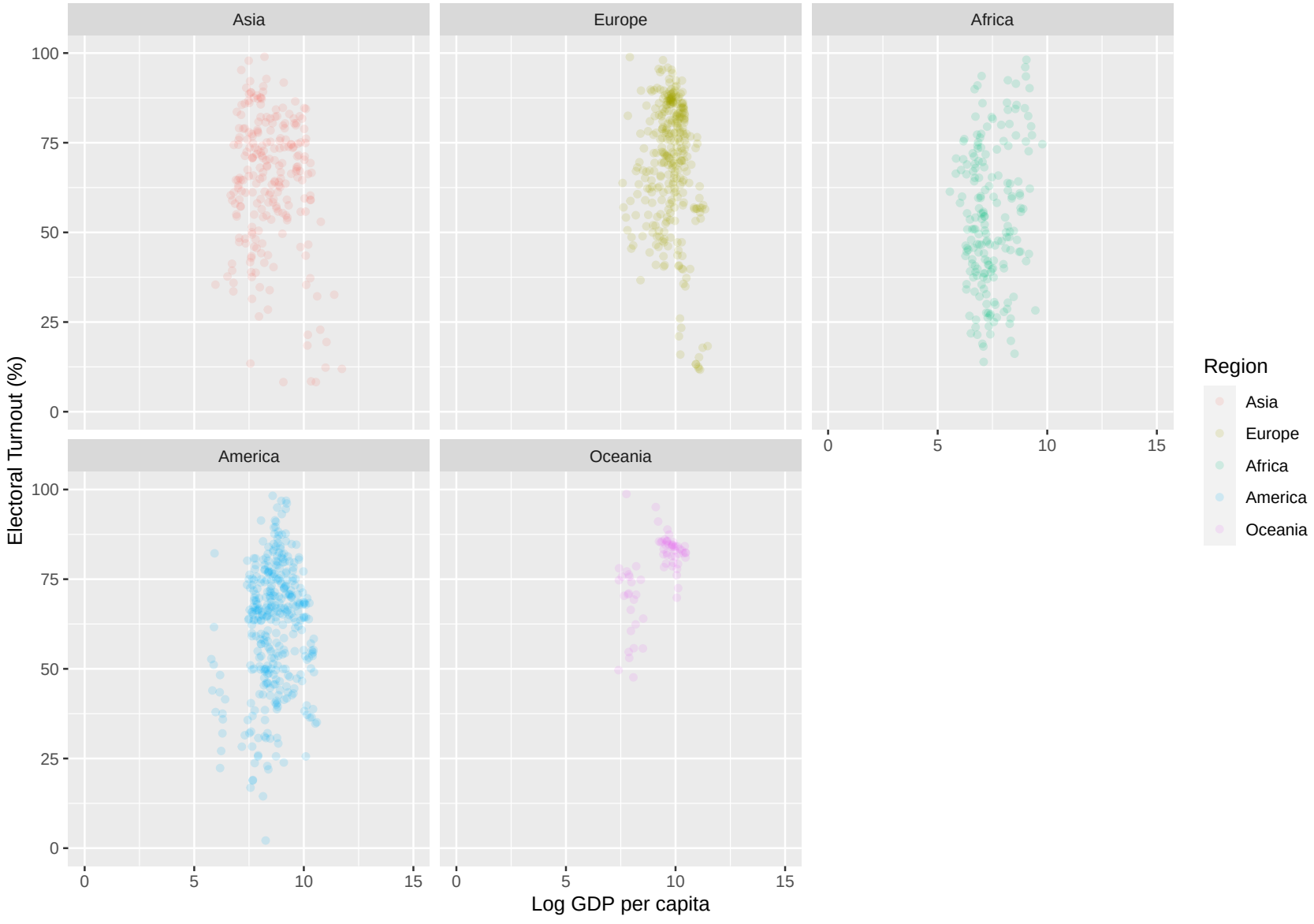
Take another example



One solution



Another solution



When to use color?

1) Distinguishing groups
2) Represent data values
3) Highlighting information

Color for distinguishing discrete categories

- Qualitative scales such as ethnicity, region, type of government, etc. can easily be differentiated by color. This means that we can make a qualitative scale an auxiliary mapping variable in addition to the typical x and y variables.
- We have talked about how discrete variables have no inherent order. This means that the colors that correspond to a categorical, discrete variable should not appear to reflect some order.
- Most color palettes will contain differentiated colors for discrete values, such that a sequential order to the data is not implied. However, if we make a manual palette, this will be a point to keep in mind.

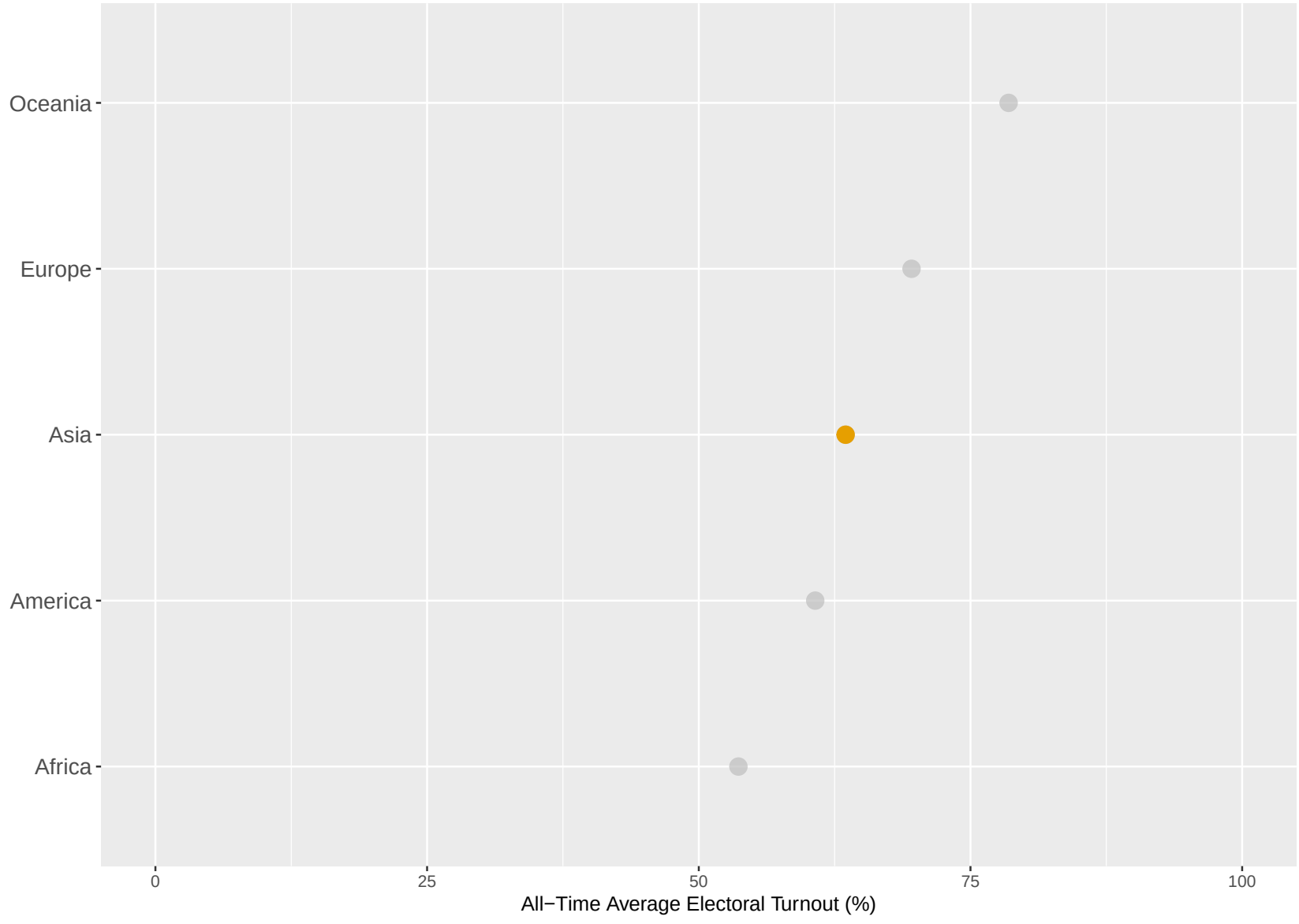
Color for representing data values

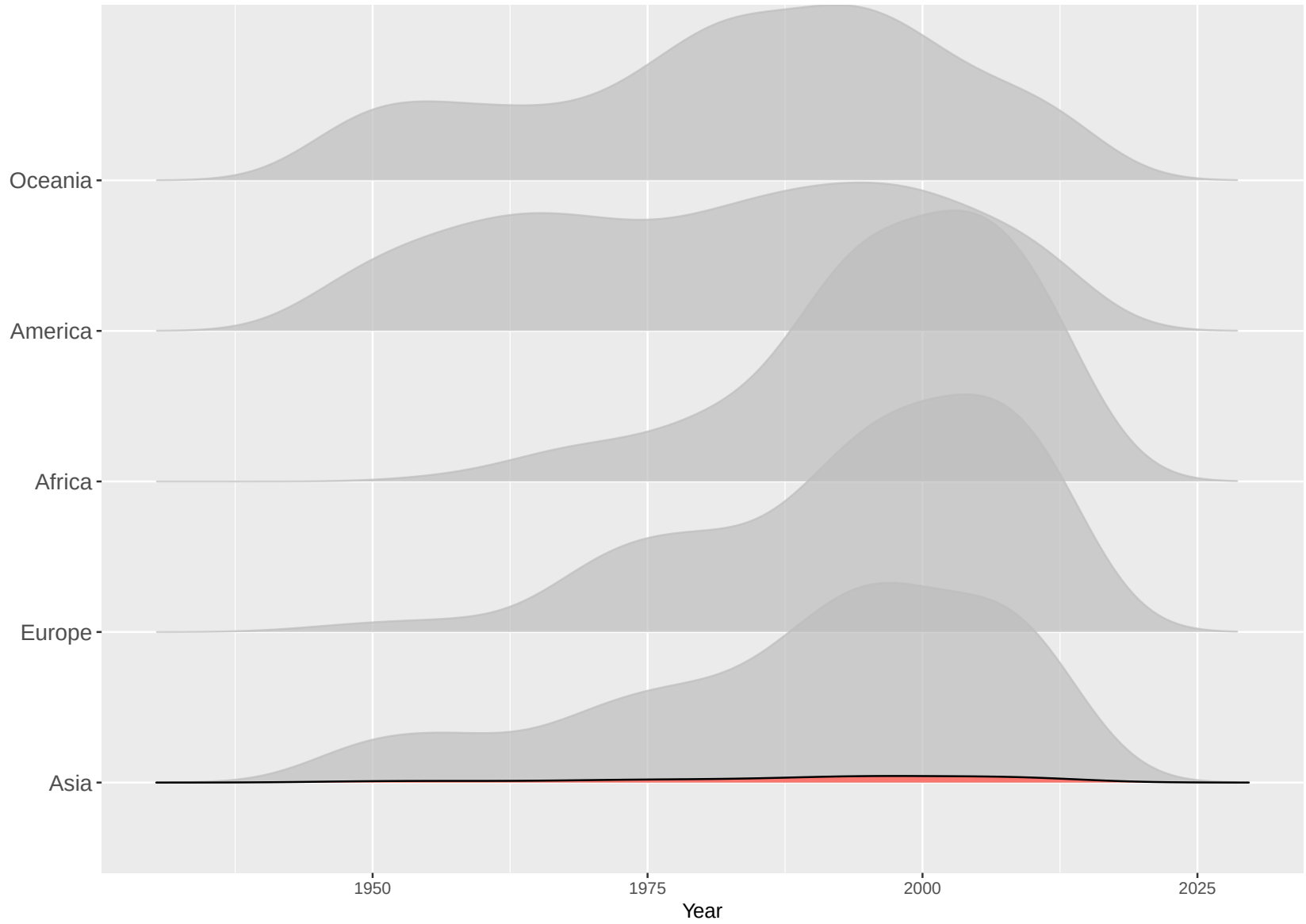
- Continuous values can also be represented by color.
- As opposed to the discrete application of color, we can enact some ordered scale or gradient with a sequential color scale for continuous values.
- Sequential color scales indicate the values in the data that are smaller or larger relative to the full data, which further indicates the magnitude of difference between two data points.

Color to highlight

- Sometimes the questions that we address in social science and data visualization are about key groups within a category.
- Finding a way to highlight those key categories amid the rest of the data can be important to show viewers the trends among that key category within the full data. In this case, we want to **accent** our key category, without implanting color for all categories in the data.

Example of highlighting




Oceania
America
Africa
Europe
Asia
1950
1975
2000
2025
Year

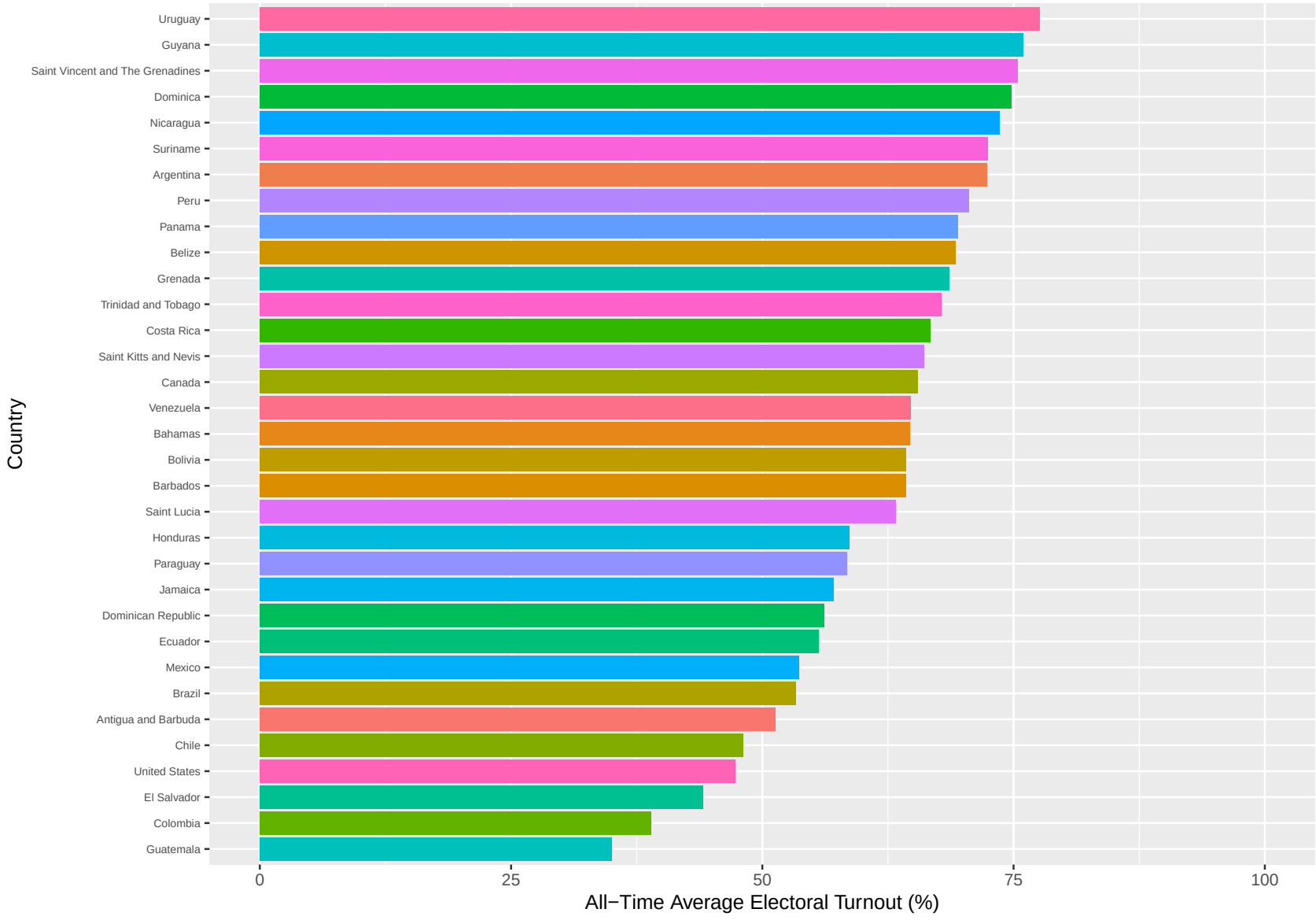
When is color overenhancing?

- So far color and fill options have looked like attractive options for including one more piece of information – either discrete or continuous.
- However, color has limits. Some of you might remember the line graph from last week that included a color line for each of the departments in Colombia... not great!
- Remember that we are asking our audience to engage in a visualization tasks that helps to reduce the complexity of trends in data. To that point, it's on us to find a minimally-intensive task.
- **RULE OF THUMB**: Qualitative color scales best for 3-5 categories. Beyond that, we can use some direct labels where appropriate.

Color for the sake of color?

- Coloring just because is also not a great idea.
- Adding color "just because" overcomplicates the visualization task given that there is nothing to be discerned from the color variation.
- Furthermore, we don't want to choose color options that over-saturate the colors visualized. If color becomes so saturated that there is no ability to compare between some values, then it is not a useful inflection point in the visualization.

Too much color!!!



Color is a boon and a bane. . .

- Color-vision deficiencies are really quite widespread. Therefore, our decisions need to based in part with these different abilities in mind. Our best bet is to usually choose a color palette that we *know* avoids potential issues regarding *ability* to interpret.
- The readings to skim cover accessibility in this regard a little further, so you should take a look.
- Furthermore, **cvd tools** help you to visualize possible issues that your visualizations might have w.r.t. color-related visual impairments.

Example:
https://www.color-blindness.com/coblis-color-blindness-simulator/

What about shapes?

- The guidelines regarding the use of shapes is pretty much the same.
- However, I would suggest using this option sparingly. Especially if the choice is between color, faceting, or shape– choosing shape is our worst case scenario.

Use of shape doesn't really help us all that much in this case