Data visualization

APAM E4990
Modeling Social Data

Jake Hofman

Columbia University

February 15, 2019

Why visualize?

1. To explore and understand data
2. To communicate with your readers

Anscombe's quartet (1973)[1]

What's the difference between these four data sets?

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

[1] https://www.jstor.org/stable/2682899

Anscombe's quartet

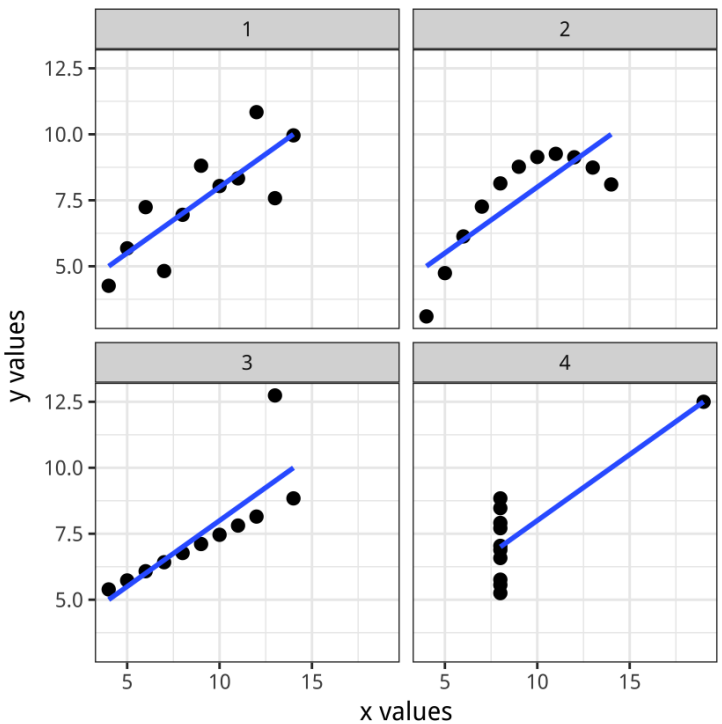
What's the difference between these four data sets?

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Property	Value
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression	0.67

Anscombe's quartet[2]

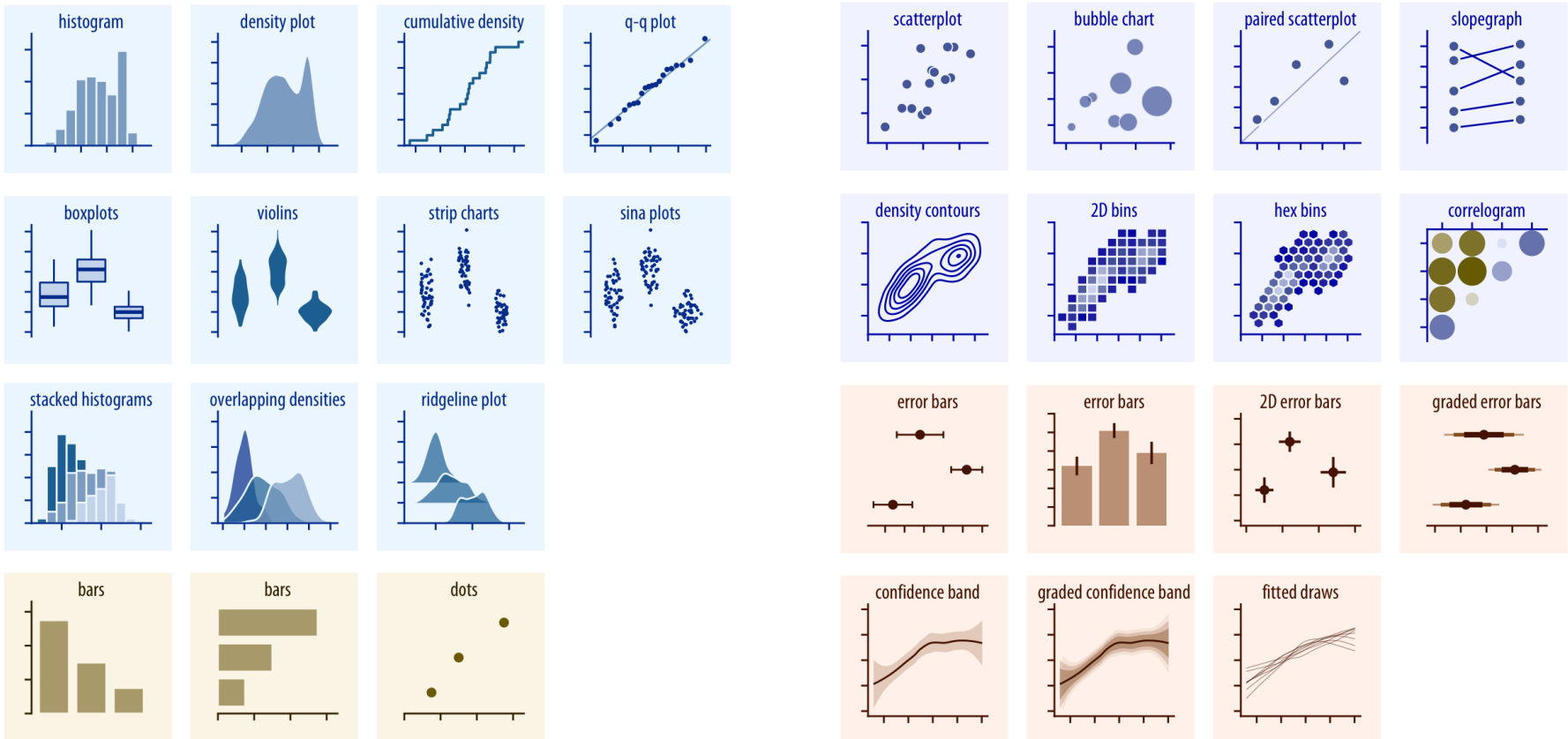
What's the difference between these four data sets?



Property	Value
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression	0.67

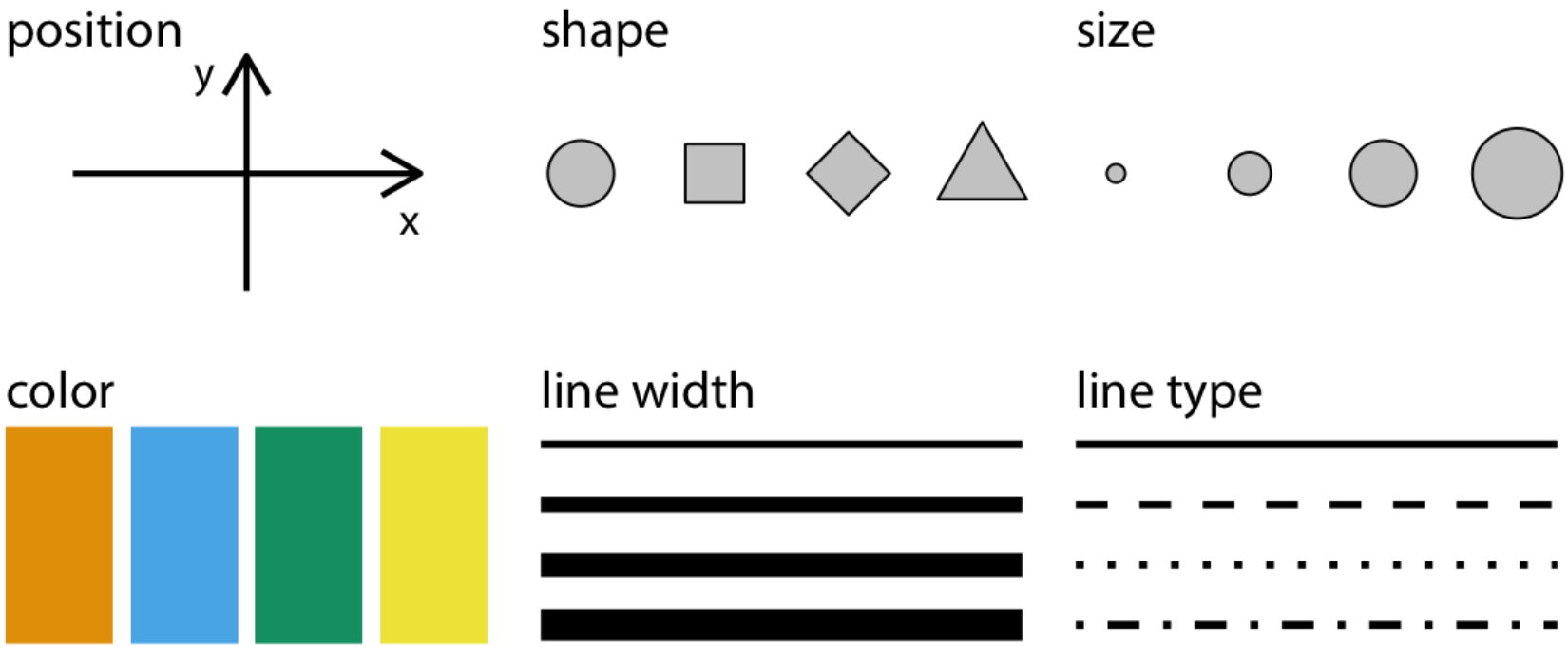
[2] http://vissoc.co/lookatdata.html

So. Many. Options.[3]



[3] https://serialmentor.com/dataviz/directory-of-visualizations.html

Even. More. Options.



Good plots (a la Mackinlay 1986)

Good plots should express the facts effectively as possible

- "Tell the truth and nothing but the truth"
- Use encodings that people can easily decode
- Make a clear and concise point
- Have a one sentence take-away

Good plots (a la Mackinlay 1986)

Automating the Design of Graphical Presentations of Relational Information

JOCK MACKINLAY
Stanford University

The goal of the research described in this paper is to develop an application-independent presentation tool that automatically designs effective graphical presentations (such as bar charts, scatter plots, and connected graphs) of relational information. Two problems are raised by this goal: The codification of graphic design criteria in a form that can be used by the presentation tool, and the generation of a wide variety of designs so that the presentation tool can accommodate a wide variety of information. The approach described in this paper is based on the view that graphical presentations are sentences of graphical languages. The graphic design issues are codified as expressiveness and effectiveness criteria for graphical languages. Expressiveness criteria determine whether a graphical language can express the desired information. Effectiveness criteria determine whether a graphical language exploits the capabilities of the output medium and the human visual system. A wide variety of designs can be systematically generated by using a composition algebra that composes a small set of primitive graphical languages. Artificial intelligence techniques are used to implement a prototype presentation tool called APT (A Presentation Tool), which is based on the composition algebra and the graphic design criteria.

Some visualizations are better[4] than others

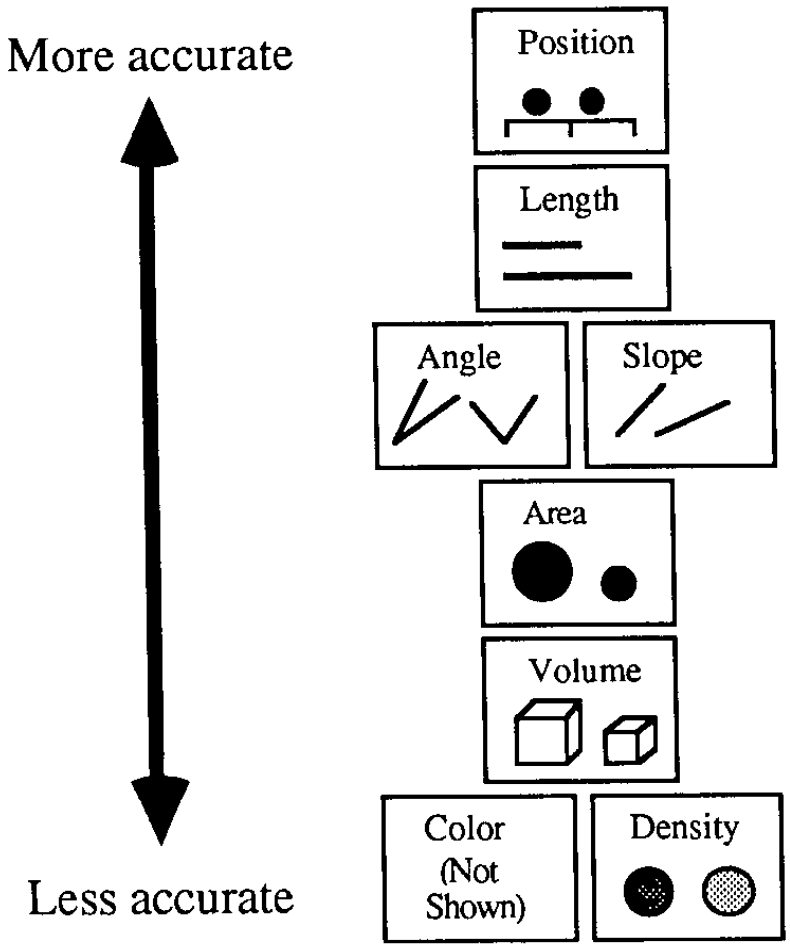


Fig. 14. Accuracy ranking of quantitative perceptual tasks. Higher tasks are accomplished more accurately than lower tasks. Cleveland and McGill empirically verified the basic properties of this ranking.

[4] Perceived more accurately

Clevland & McGill 1984

What percent is the smaller region of the larger region?

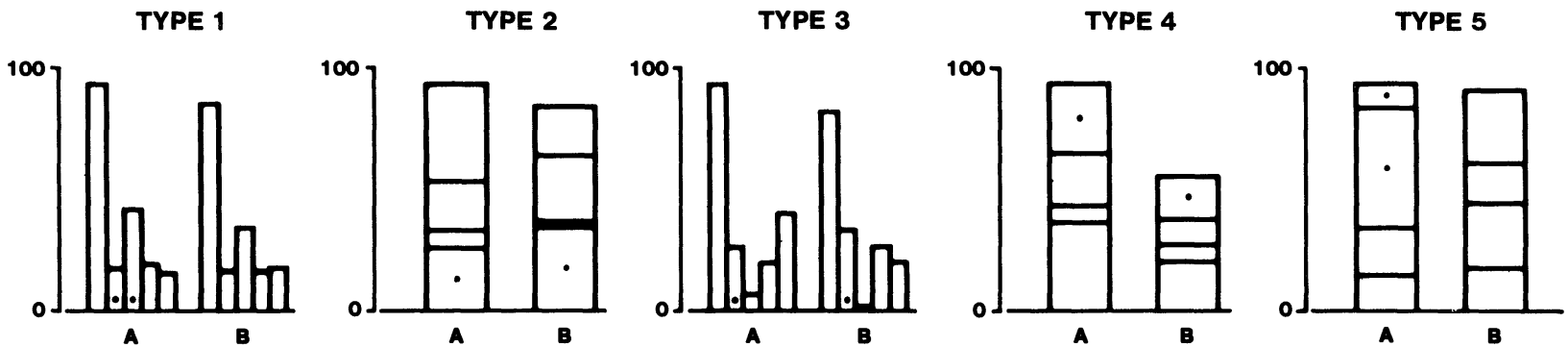
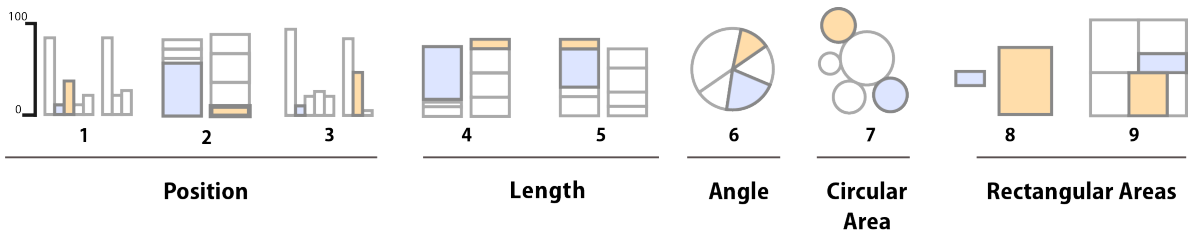
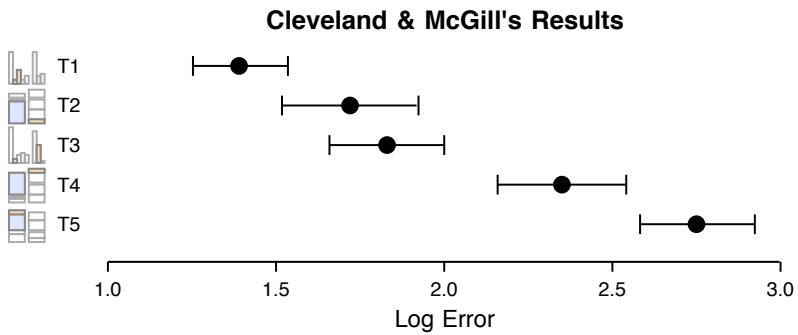
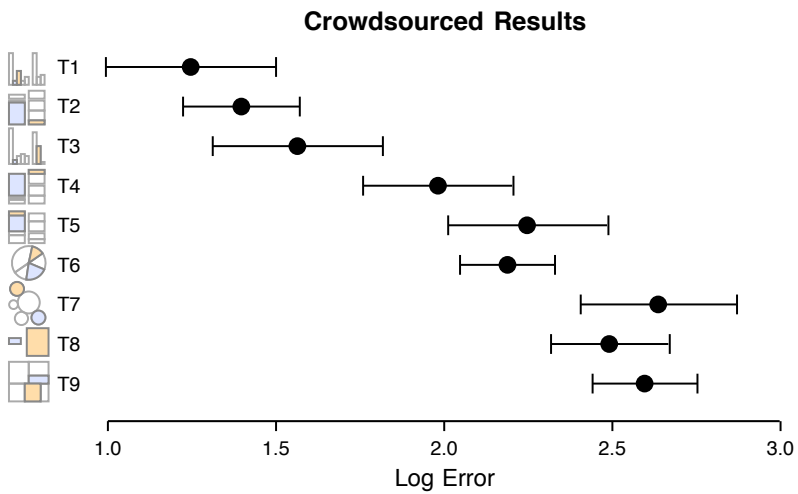


Figure 4. Graphs from position–length experiment.

Clevland & McGill 1984 / Heer & Bostock 2010







Diffrent strokes for different data types

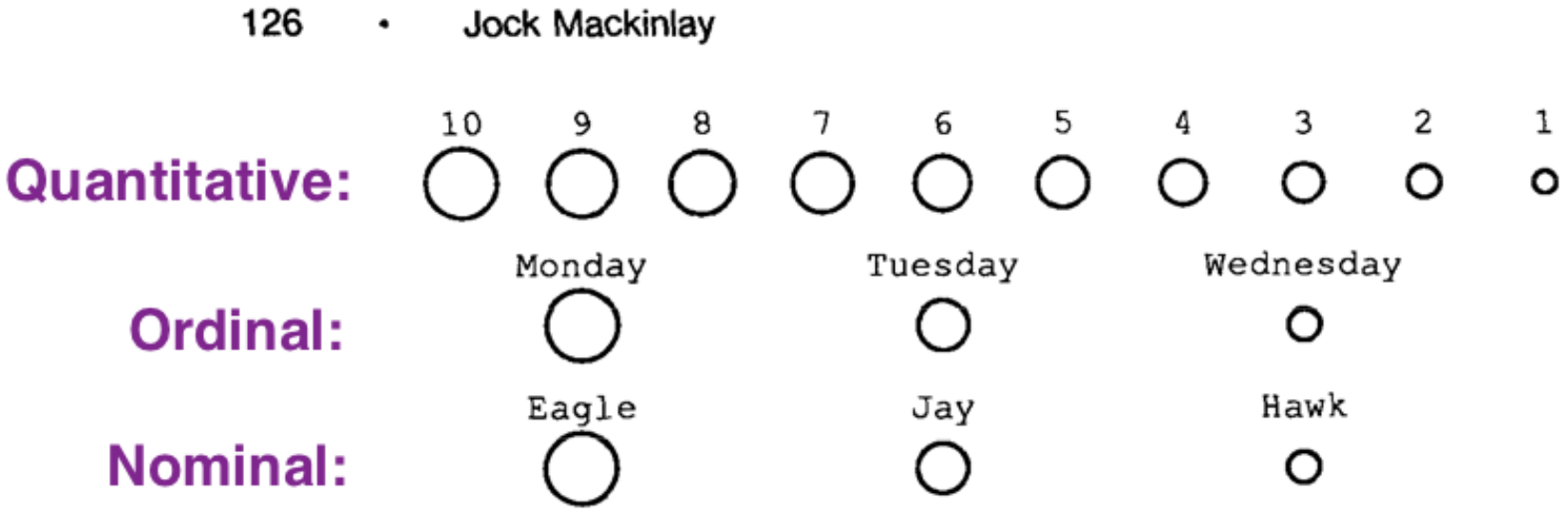


Fig. 16. Analysis of the area task. The top case shows that area is moderately effective for encoding quantitative information. The middle case shows that it is possible to encode ordinal information as long as the step size between areas is large enough so that the values are not confused. The bottom case shows that it is possible to encode nominal information, but people may perceive an ordinal encoding.

- Quantitative: numerical values in a range (e.g., height)
- Ordinal: categories with natural ordering (e.g., day of week)
- Nominal: categories with no natural ordering (e.g., gender)

Diffrent strokes for different data types

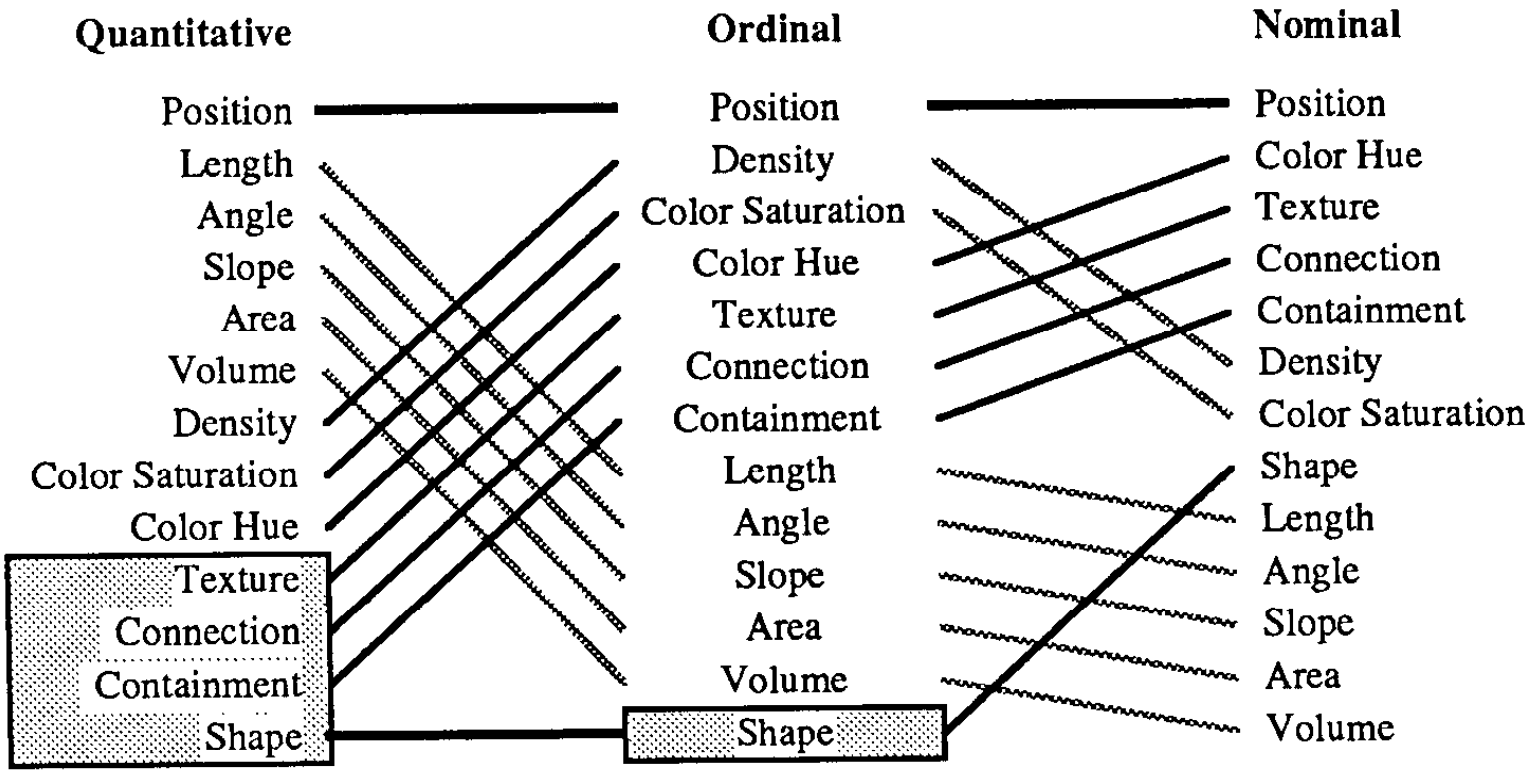
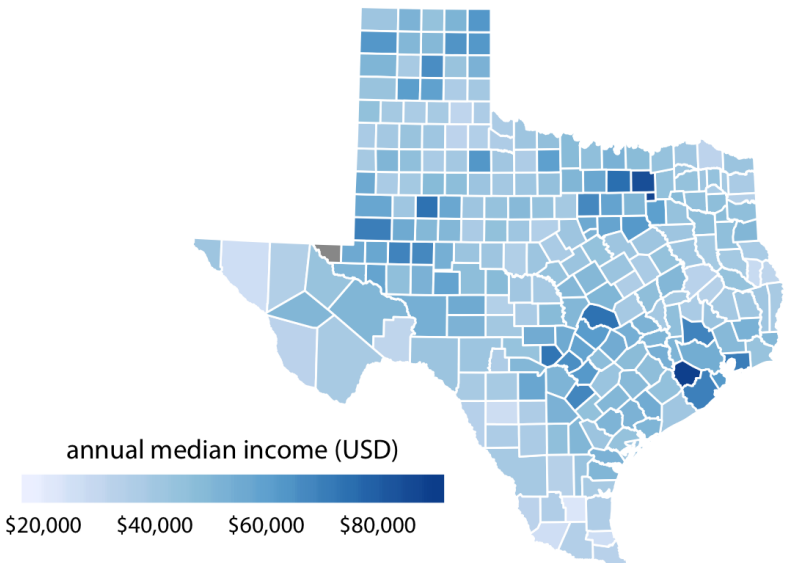


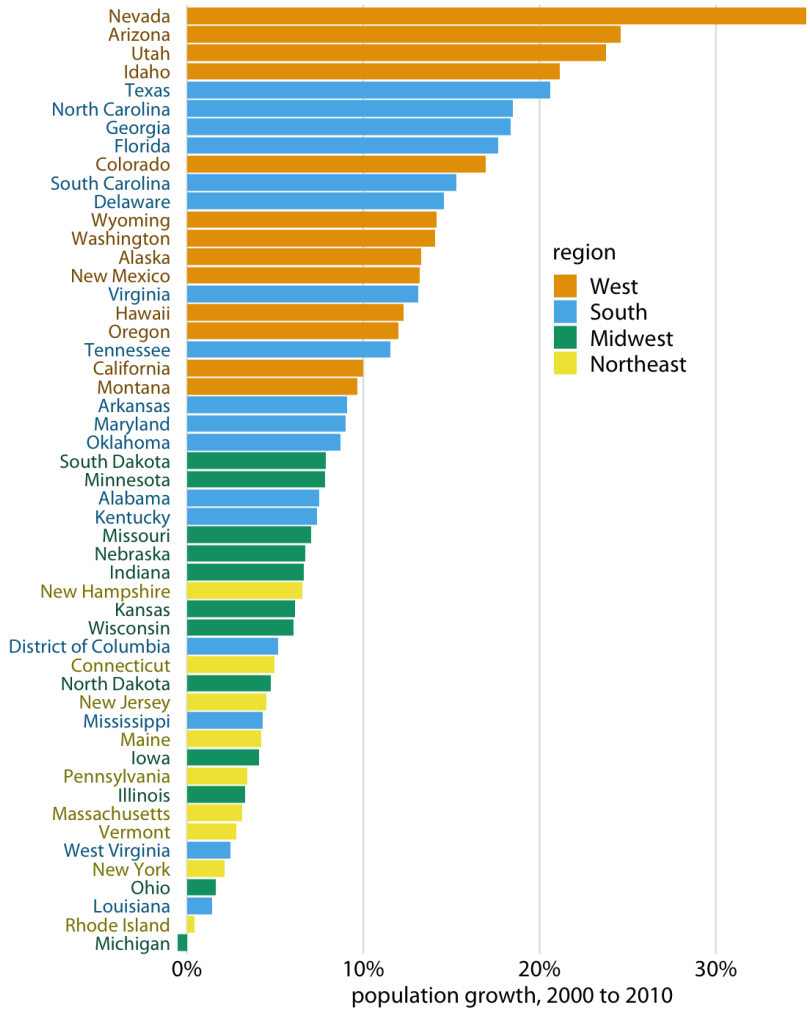
Fig. 15. Ranking of perceptual tasks. The tasks shown in the gray boxes are not relevant to these types of data.

Diffrent colors for different data types[5]

Quantitative



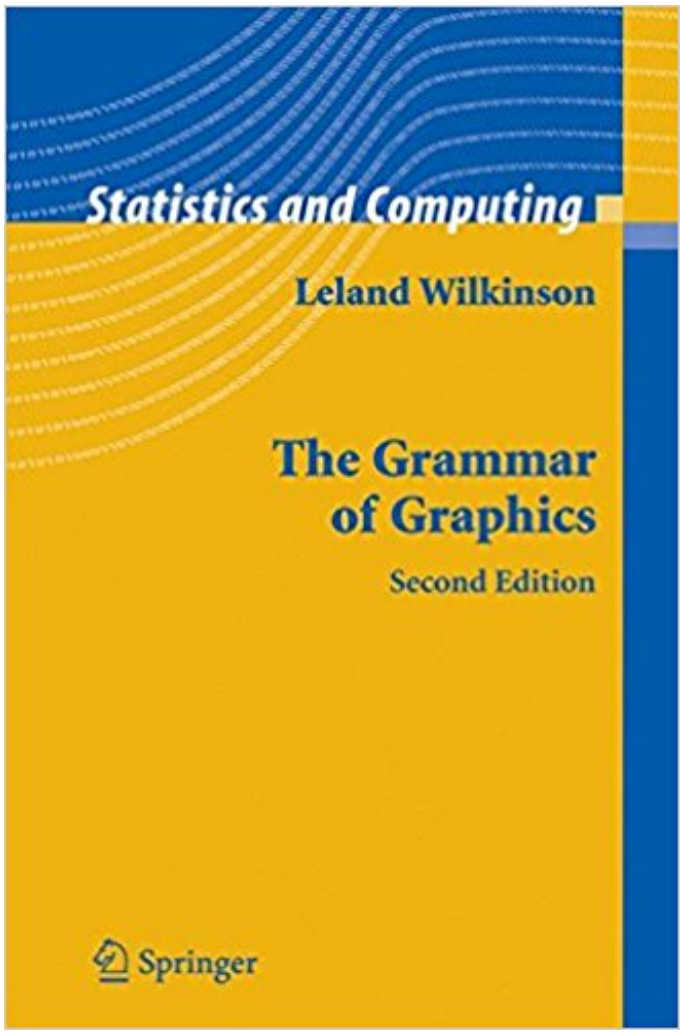
Nominal



[5] https://serialmentor.com/dataviz/color-basics.html

A grammar of graphics

A language to describe the components of a graphic



Grammar of graphics a la ggplot2[6]

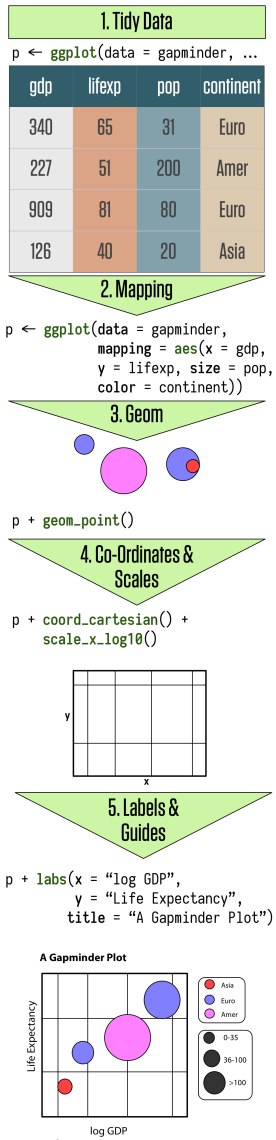
3.10 The layered grammar of graphics

In the previous sections, you learned much more than how to make scatterplots, bar charts, and boxplots. You learned a foundation that you can use to make *any* type of plot with ggplot2. To see this, let's add position adjustments, stats, coordinate systems, and faceting to our code template:

```
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(
     mapping = aes(<MAPPINGS>),
     stat = <STAT>,
     position = <POSITION>
  ) +
  <COORDINATE_FUNCTION> +
  <FACET_FUNCTION>
```

[6] http://r4ds.had.co.nz

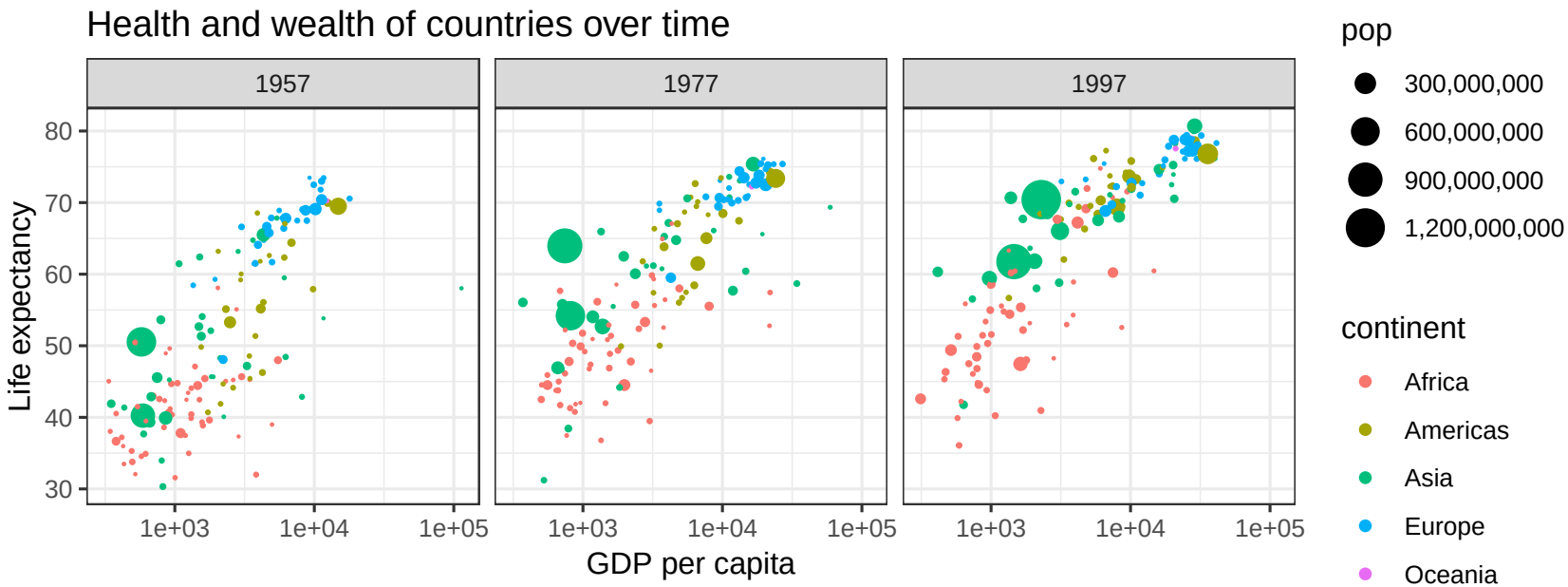
Grammar of graphics a la ggplot2 [7]



1. Get your data into the right format
2. Map variables to aesthetics
3. Choose a geometry for your plot
4. Set co-ordinate system and scales
5. Add annotations, legends, and labels

[7] http://vissoc.co/makeplot.html

Grammar of graphics a la ggplot2




```
ggplot(data = gapminder,
       aes(x = gdpPercap, y = lifeExp,
           size = pop, color = continent)) +
  geom_point() + scale_x_log10() +
  scale_size_area(label = comma) +
  labs(x = 'GDP per capita', y = 'Life expectancy',
       title = 'Health and wealth of countries over time')
  facet_wrap(~ year)
```

Benefits

- Lowers the barrier to asking questions of your data
- Lets you explore more, and faster
- Easily produces publication-ready plots
- Large and active user base for support

Acknowledgements

Slides are generously adapted from Çağatay Demiralp, whose slides are generously adapted from Jeff Heer's Data Visualization course