Data Visualization

ISTP Summer JC

Anthony Hung

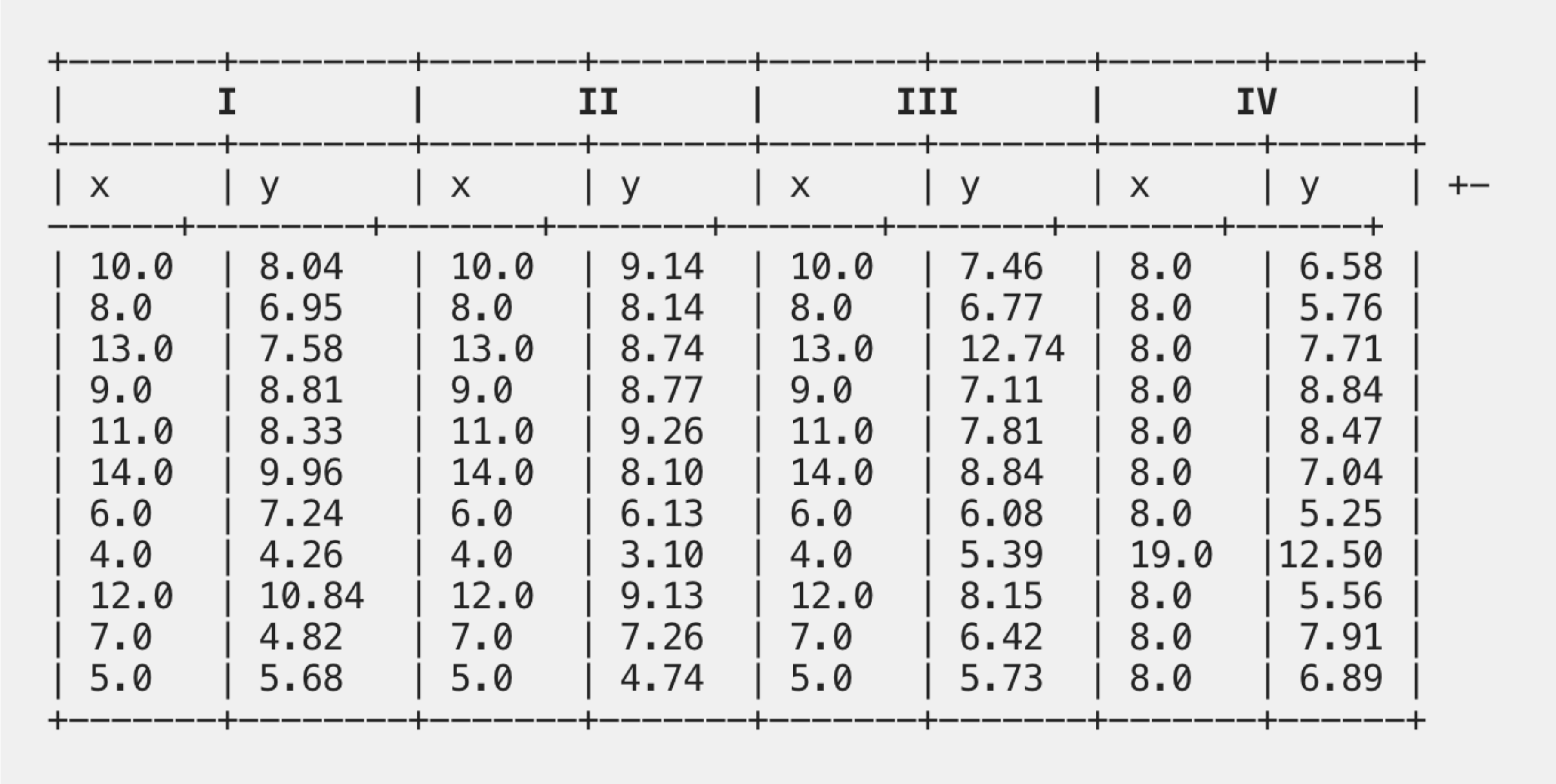
Outline

- Data visualization: Data exploration vs Data presentation
  - Tips/Tricks/Resources
- Grammar of Graphics
- Exercise in R: ggplot2

Data visualization is important not only to present data to others, but to better understand it yourself

- Data exploration
  - Find the message in the data
- Data presentation
  - Present the message in the data to others (who may or may not be familiar with your work)

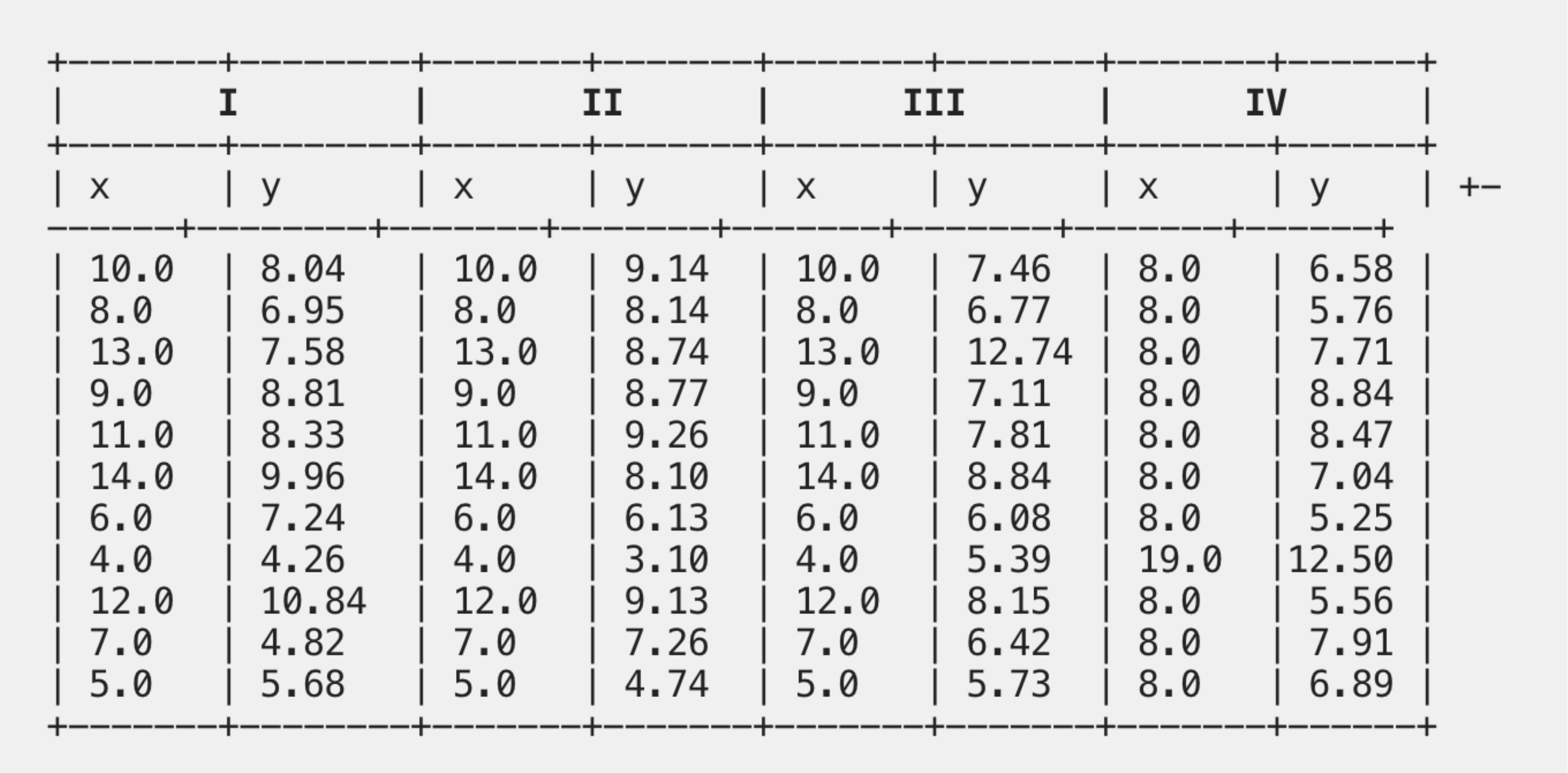
Data exploration



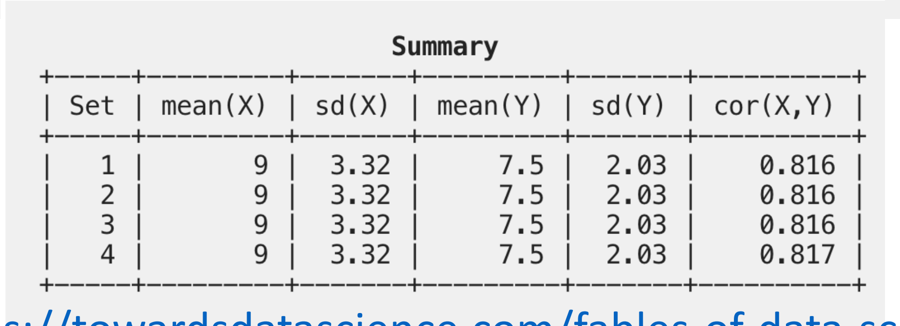
```
+-------+--------+-------+--------+-------+--------+-------+-------+
|       I        |       II       |       III      |       IV      |
+-------+--------+-------+--------+-------+--------+-------+-------+
| x     | y      | x     | y      | x     | y      | x     | y     | +-
-----+--------+-------+--------+-------+--------+-------+-------+
| 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  |
+-------+--------+-------+--------+-------+--------+-------+-------+
```

https://towardsdatascience.com/fables-of-data-science-anscombes-quartet-2c2e1a07fbe6

Data exploration



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

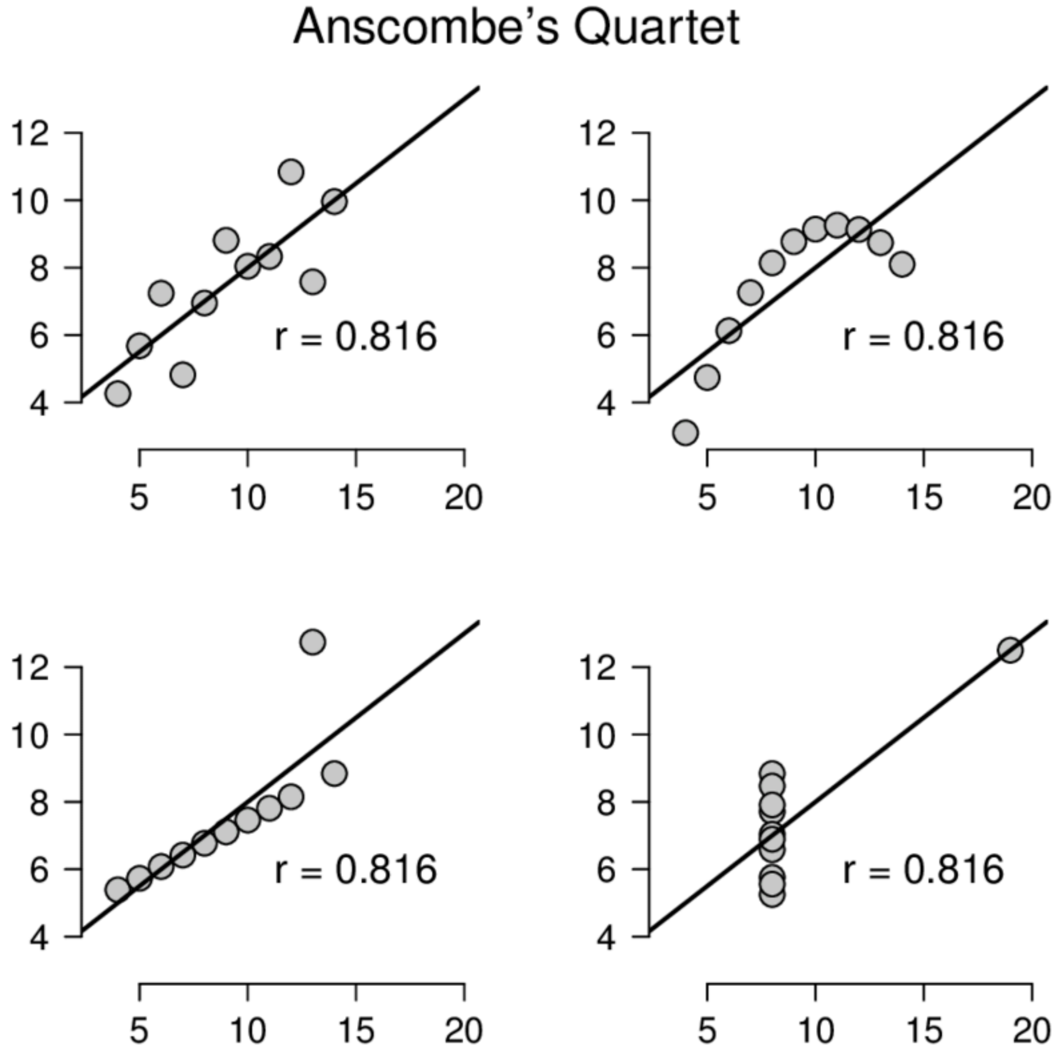


Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

https://towardsdatascience.com/fables-of-data-science-anscombes-quartet-2c2e1a07fbe6

Visualization reveals hidden relationships



Heathcote, Andrew & Brown, Scott & Wagenmakers, Eric-Jan. (2015). An Introduction to Good Practices in Cognitive Modeling. 10.1007/978-1-4939-2236-9_2.

Data presentation

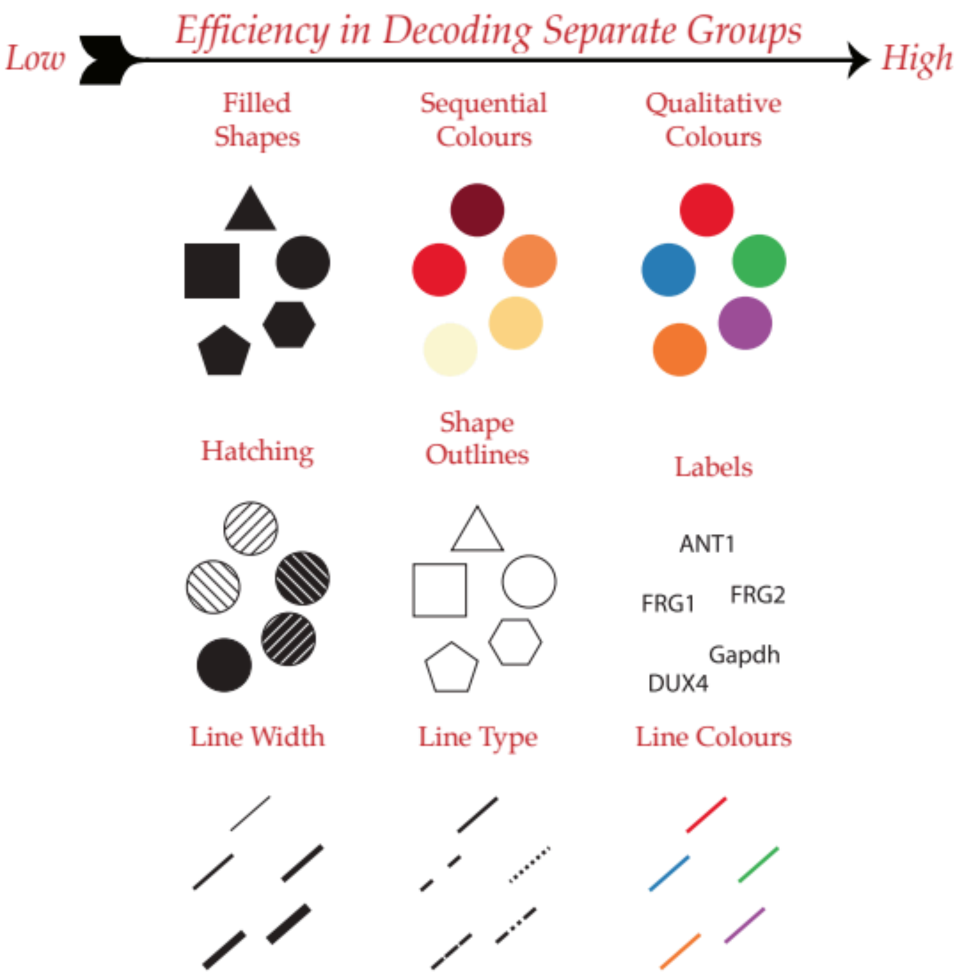
- Summarize information and present in a meaningful, non-confusing, non-manipulative way
- Consider what the ***message*** of each visual is, and use tools to best present that message
    - Best type of chart
    - Best way to discriminate datapoints

Choosing the right type of chart

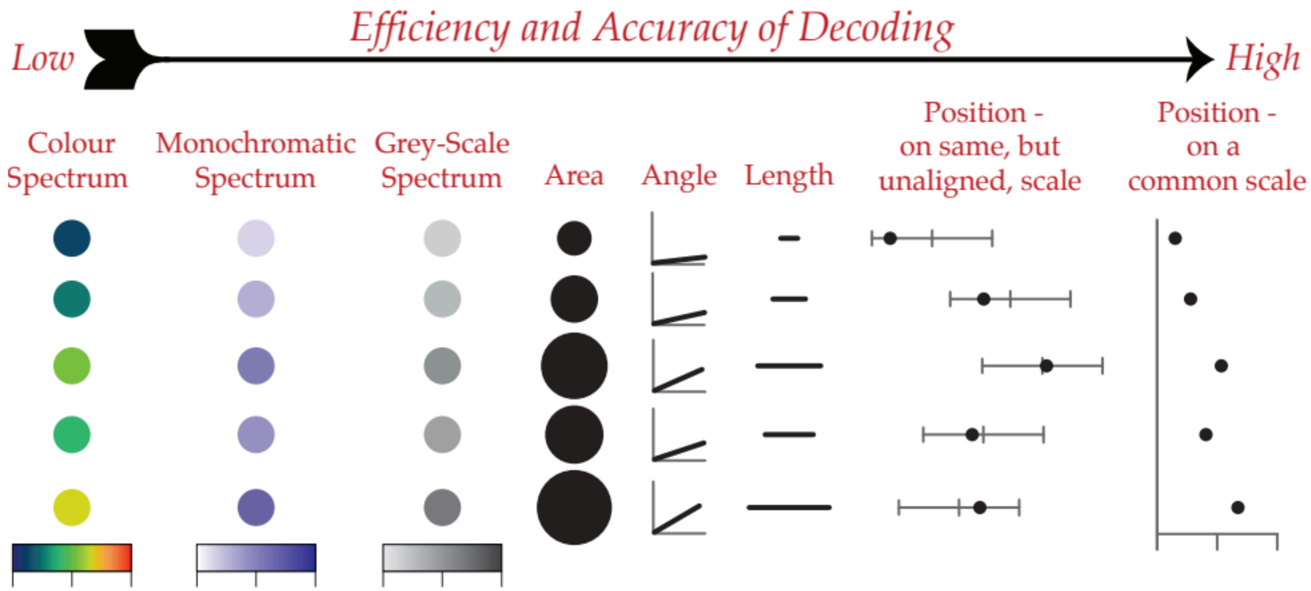
- https://datavizcatalogue.com/
- http://datavizproject.com/
- https://github.com/ft-interactive/chart-doctor/tree/master/visual-vocabulary
- http://chartmaker.visualisingdata.com/
- https://xeno.graphics/

Discriminating between datapoints

Categorical Variables



Continuous Variables


Low
Efficiency and Accuracy of Decoding
High
Colour Spectrum
Monochromatic Spectrum
Grey-Scale Spectrum
Area
Angle
Length
Position - on same, but unaligned, scale
Position - on a common scale

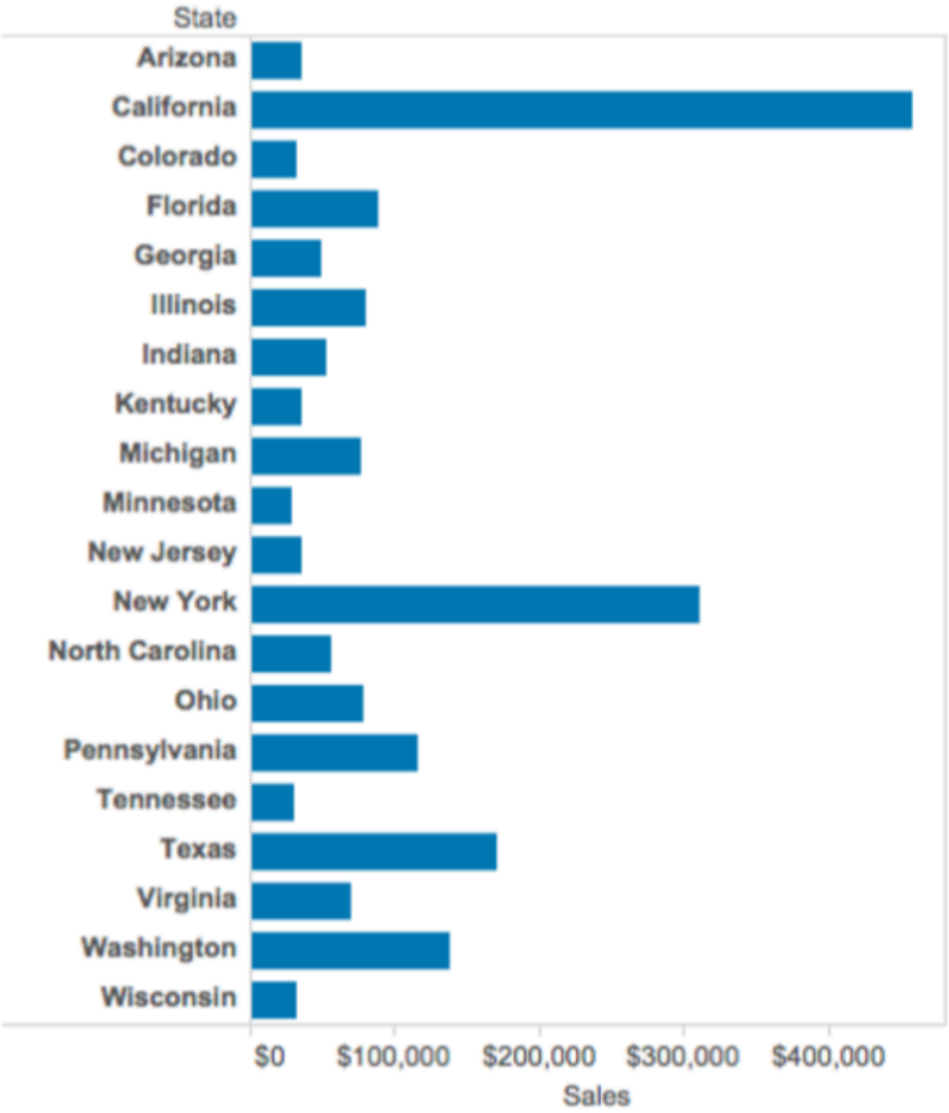
Choosing color palettes

- https://color.adobe.com/explore
- https://coolors.co/fcfafa-c8d3d5-a4b8c4-6e8387-0cca4a

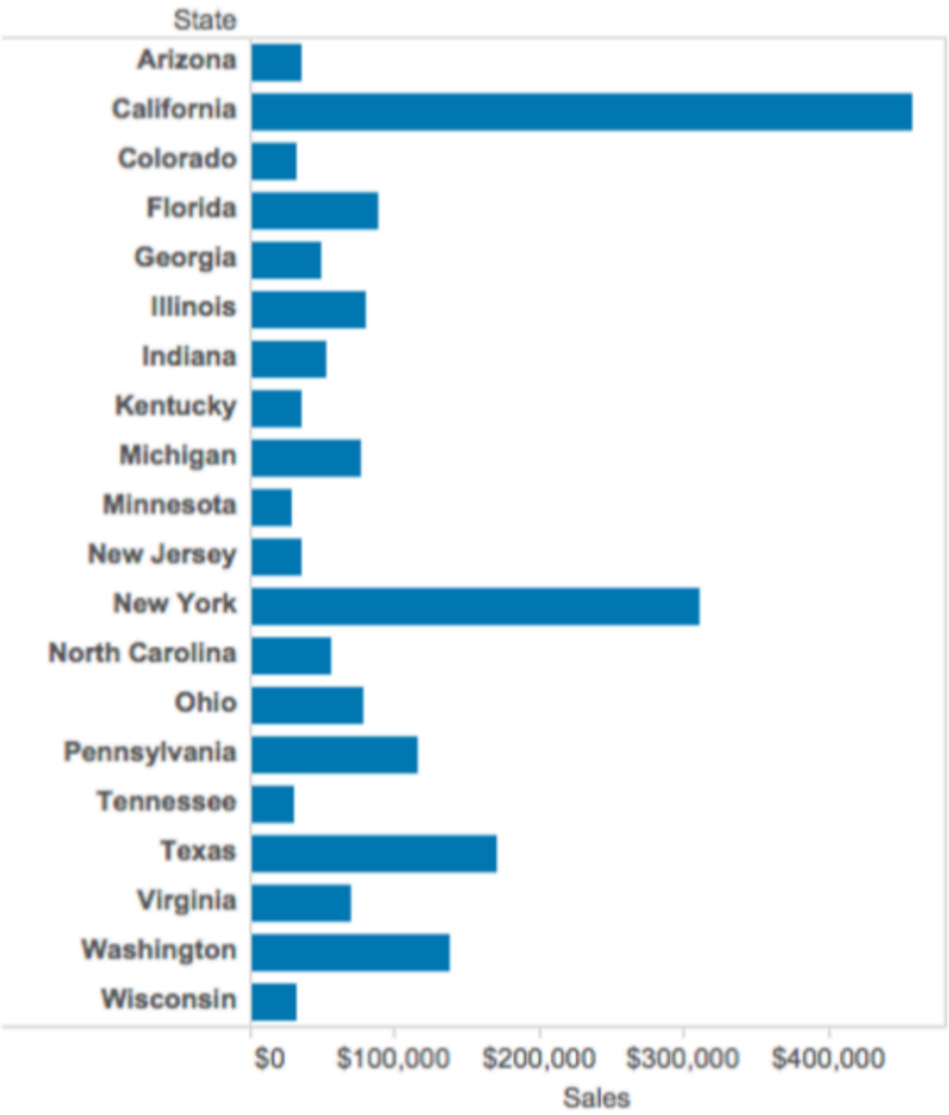
Missteps to avoid

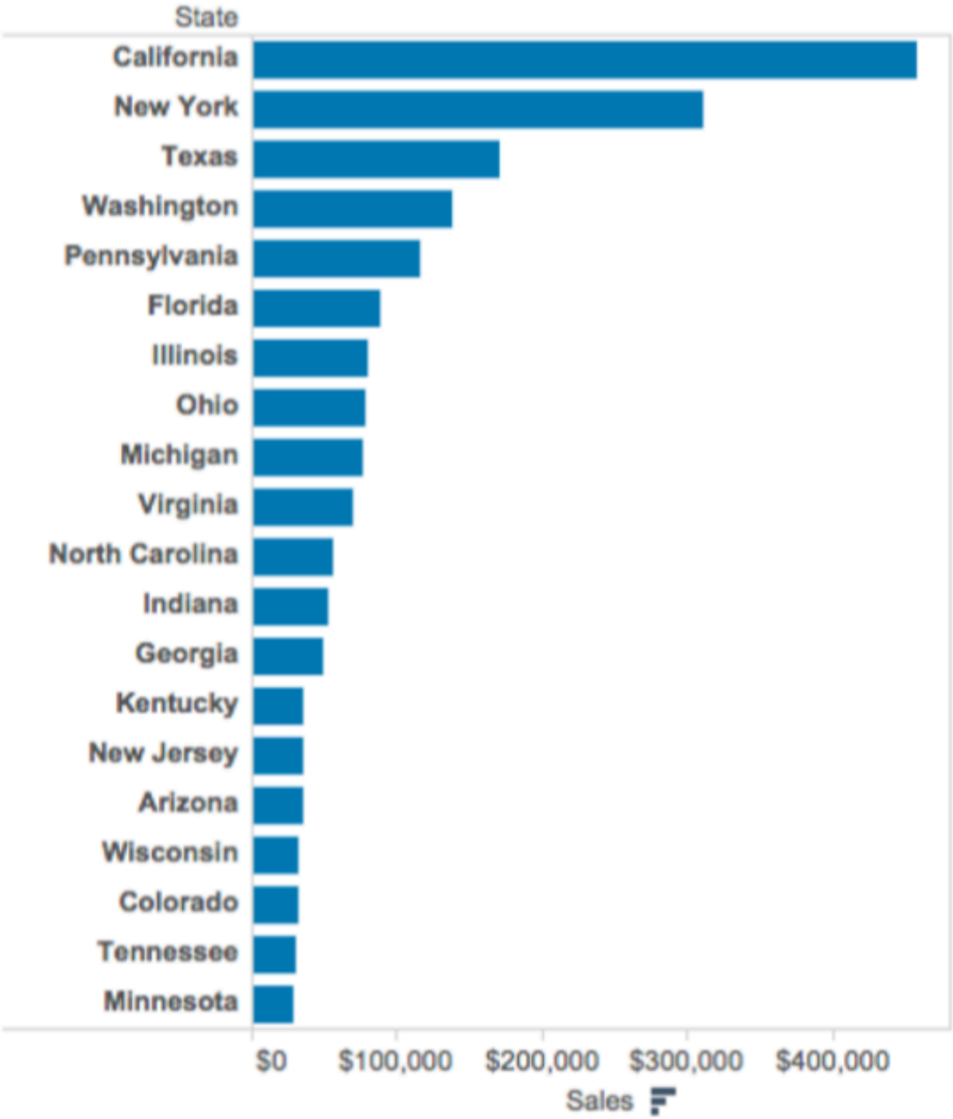
- Default ordering/Alphabetical ordering hides patterns in data
- Not all the data tells a story (emphasize the important data)

Default ordering/Alphabetical ordering hides patterns in data

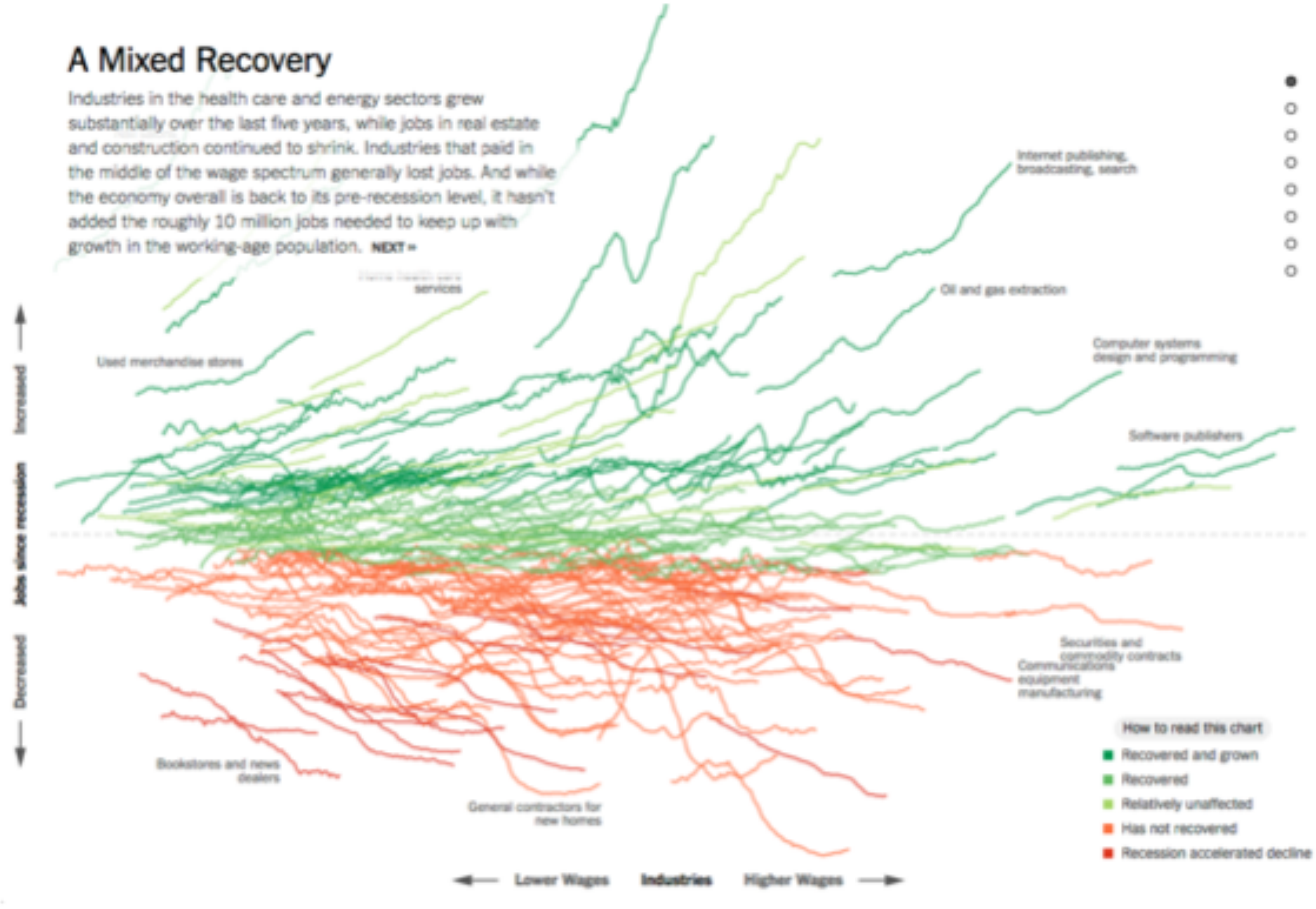


Default ordering/Alphabetical ordering hides patterns in data





Not all the data tells a story



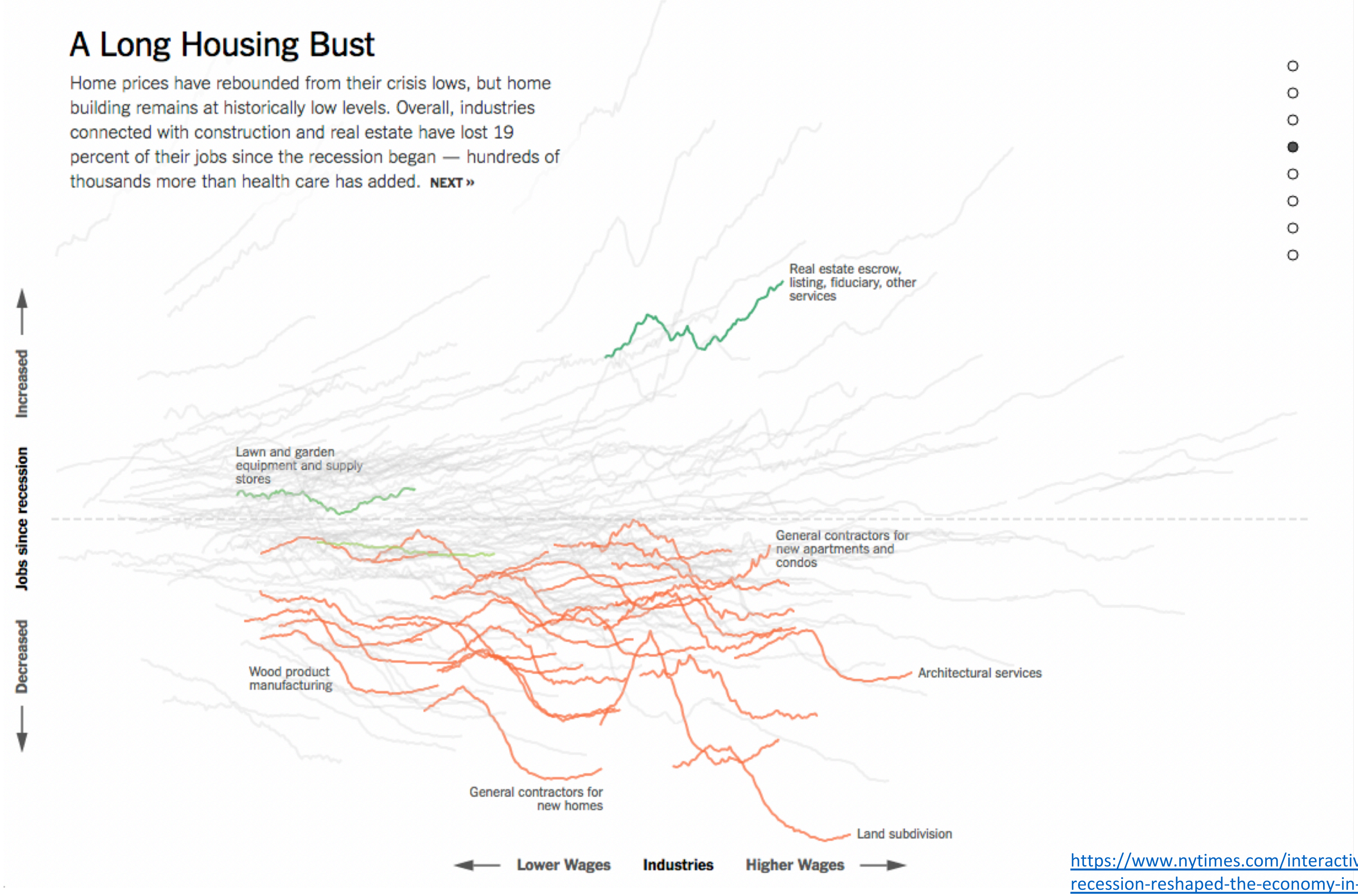
A Mixed Recovery

Industries in the health care and energy sectors grew substantially over the last five years, while jobs in real estate and construction continued to shrink. Industries that paid in the middle of the wage spectrum generally lost jobs. And while the economy overall is back to its pre-recession level, it hasn't added the roughly 10 million jobs needed to keep up with growth in the working-age population. NEXT »

Home health care services
Used merchandise stores
Internet publishing, broadcasting, search
Oil and gas extraction
Computer systems design and programming
Software publishers
Securities and commodity contracts
Communications equipment manufacturing
Bookstores and news dealers
General contractors for new homes
Increased
Jobs since recession
Decreased
Lower Wages
Industries
Higher Wages
How to read this chart
🟩 Recovered and grown
🟩 Recovered
🟩 Relatively unaffected
🟧 Has not recovered
🟥 Recession accelerated decline

https://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html

Not all the data tells a story

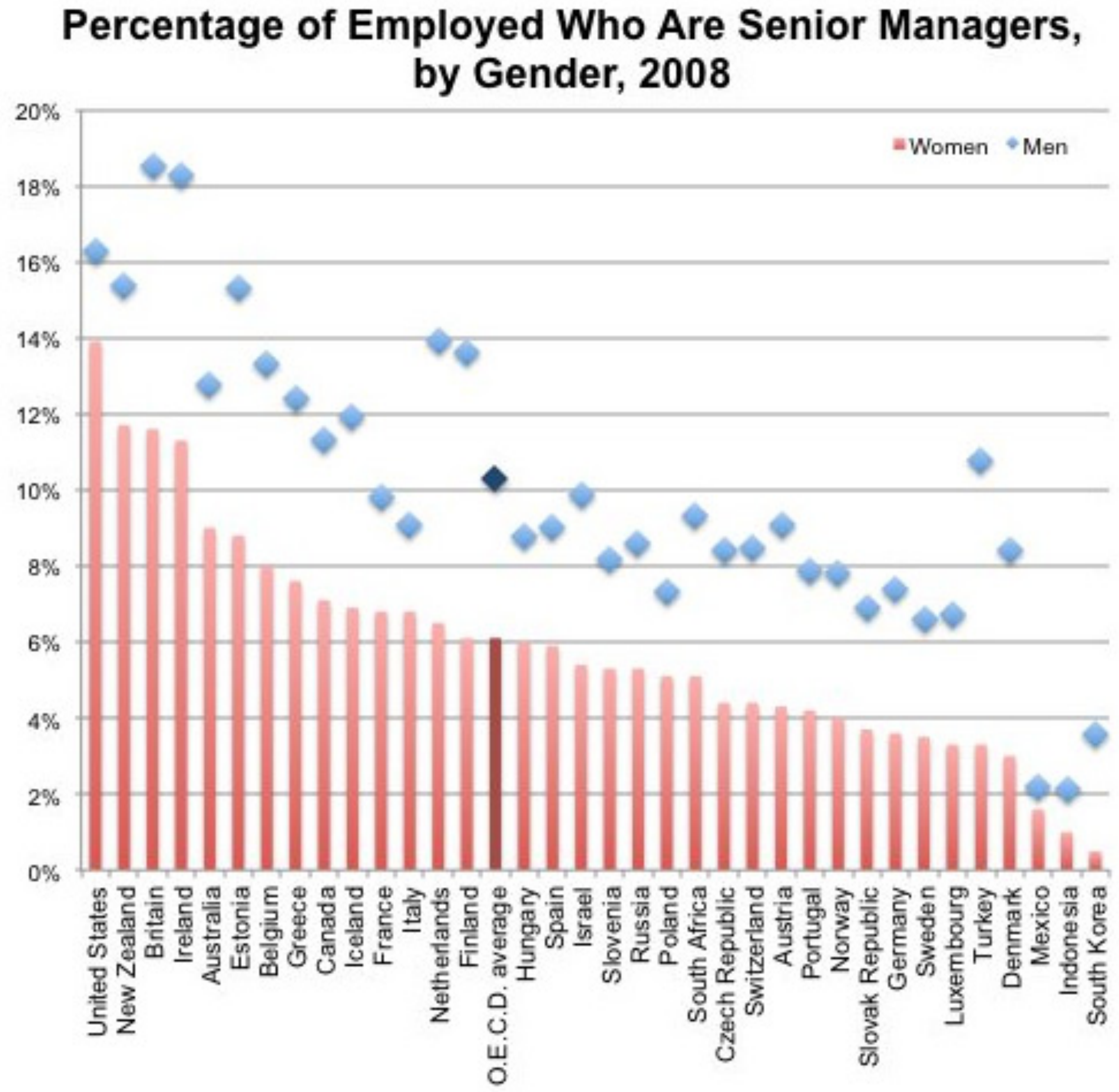


A Long Housing Bust

Home prices have rebounded from their crisis lows, but home building remains at historically low levels. Overall, industries connected with construction and real estate have lost 19 percent of their jobs since the recession began — hundreds of thousands more than health care has added. **NEXT »**

https://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html

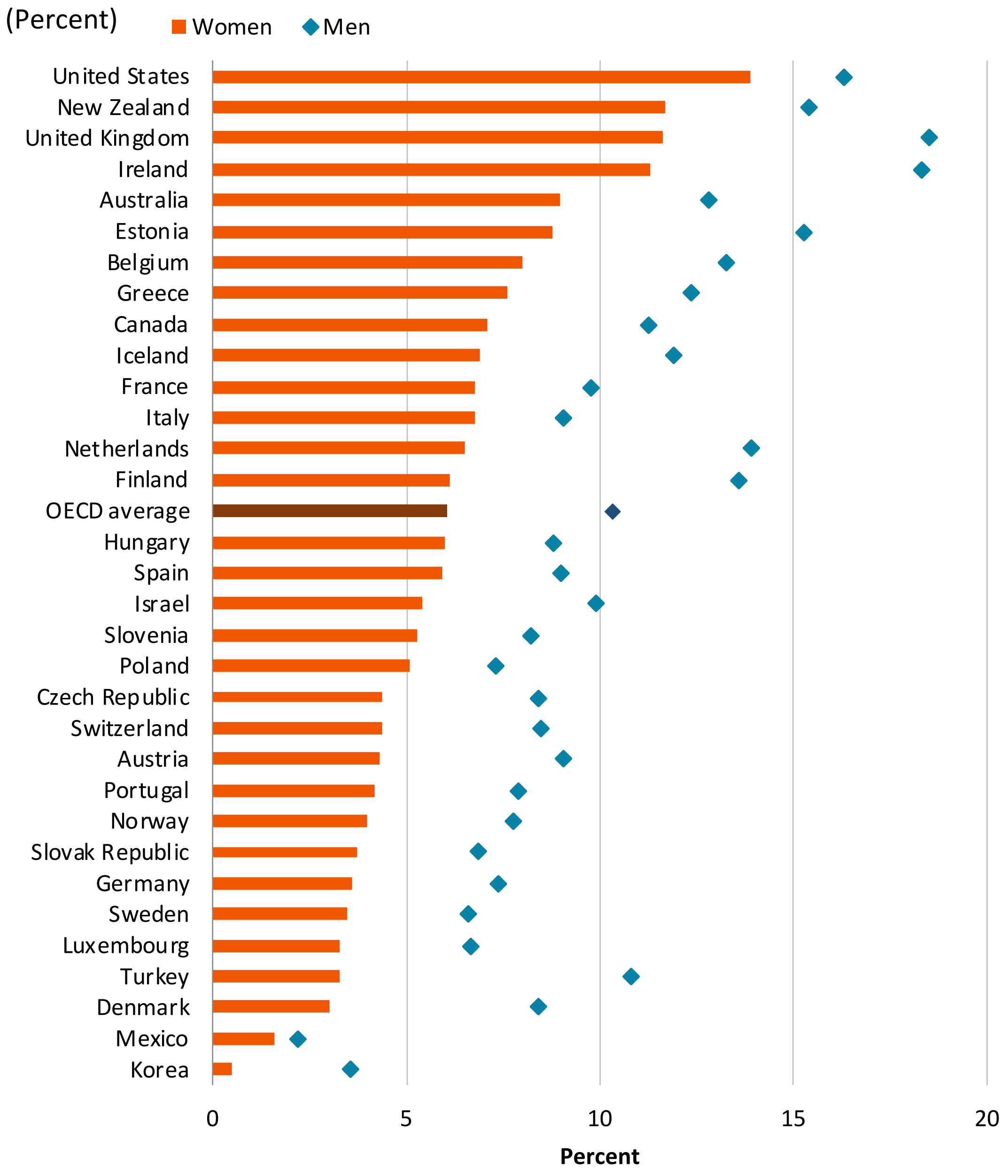
Critique/Fix some issues



http://economix.blogs.nytimes.com/2013/04/02/comparing-the-worlds-glass-ceilings/?smid=tw-share

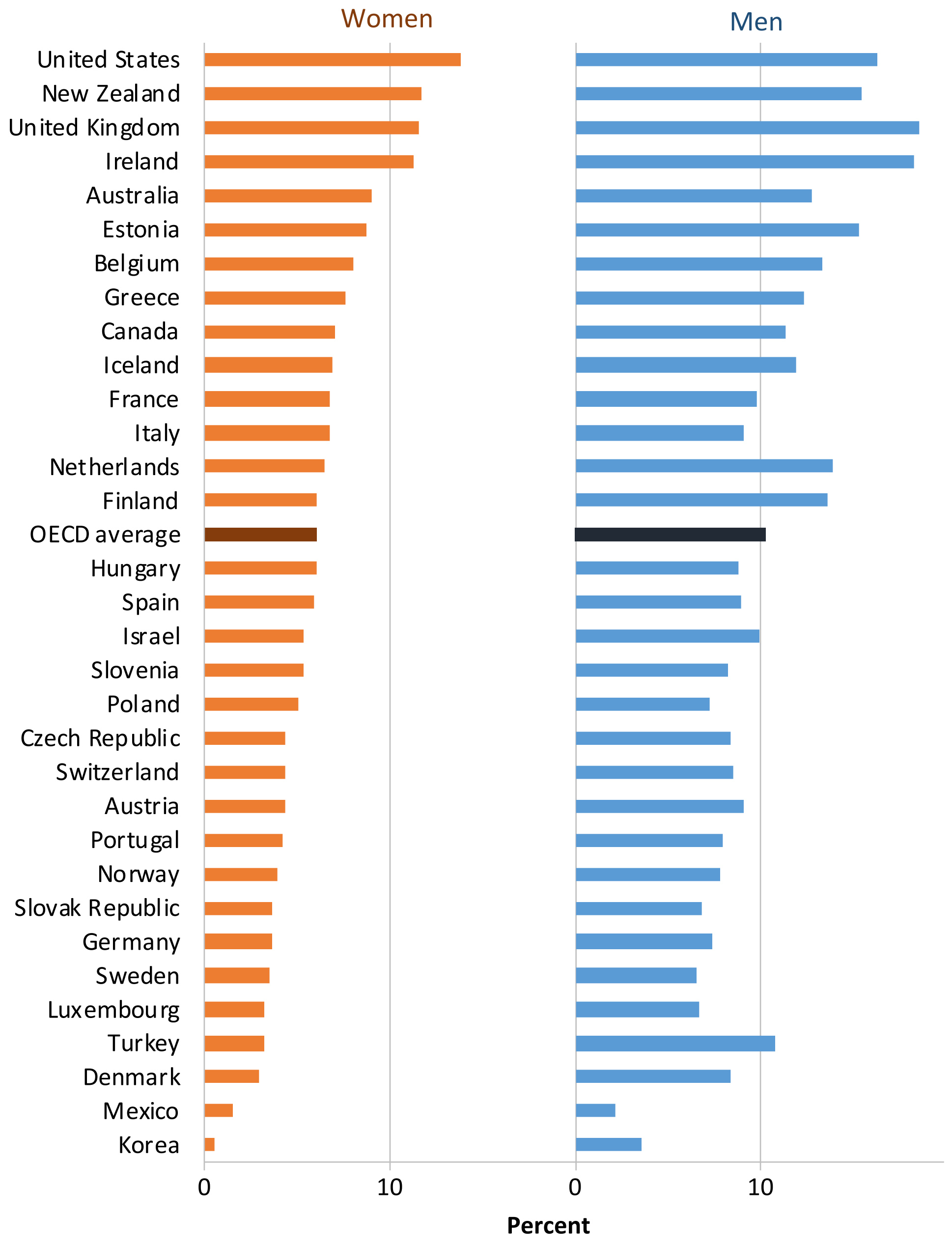
http://webcache.googleusercontent.com/search?q=cache:hl7F7jJRd6YJ:thewhyaxis.info/gap-remake/+&cd=1&hl=en&ct=clnk&gl=us

Percentage of Employed Who are Senior Managers, by Gender, 2008

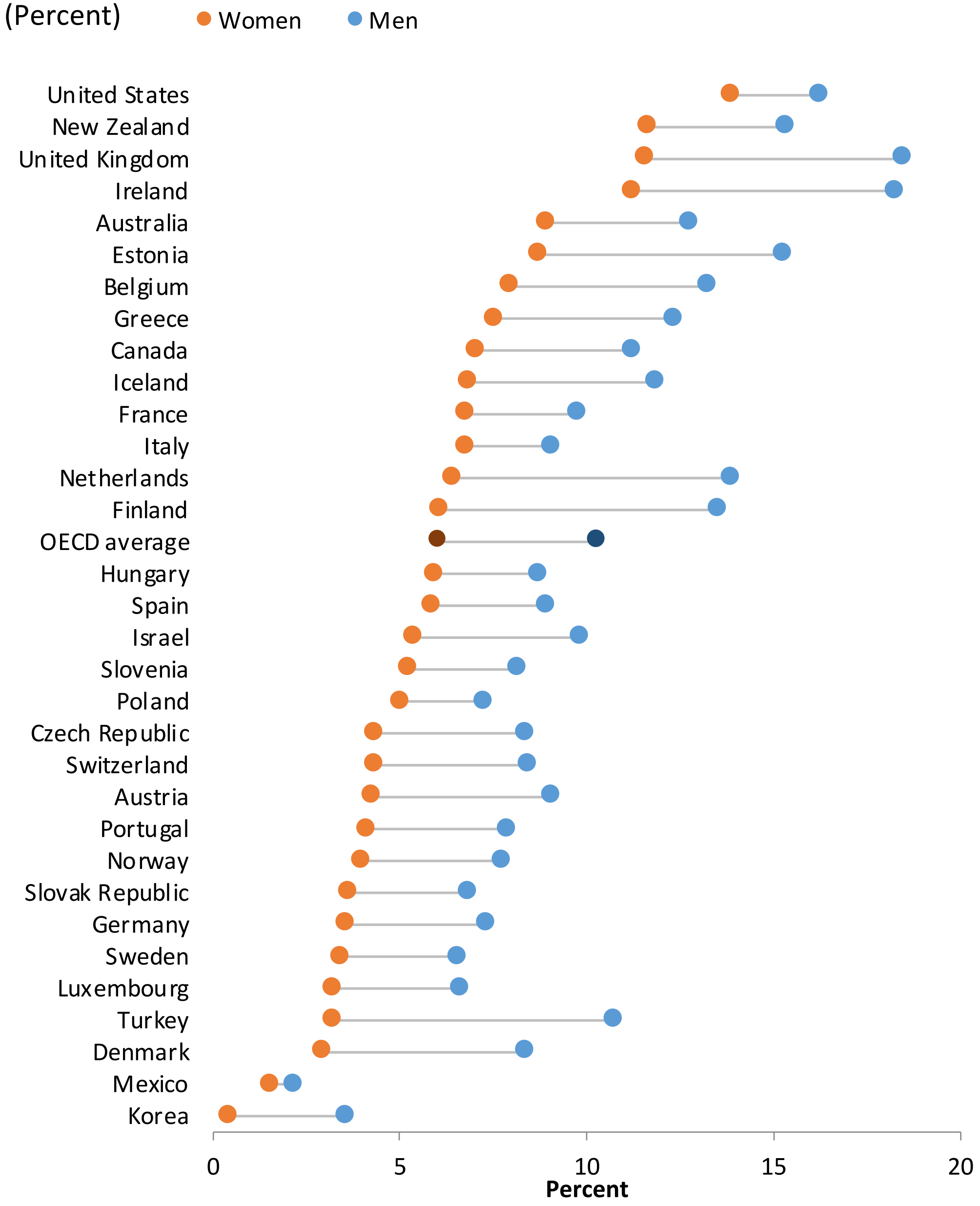


Percentage of Employed Who are Senior Managers, by Gender, 2008

(Percent)

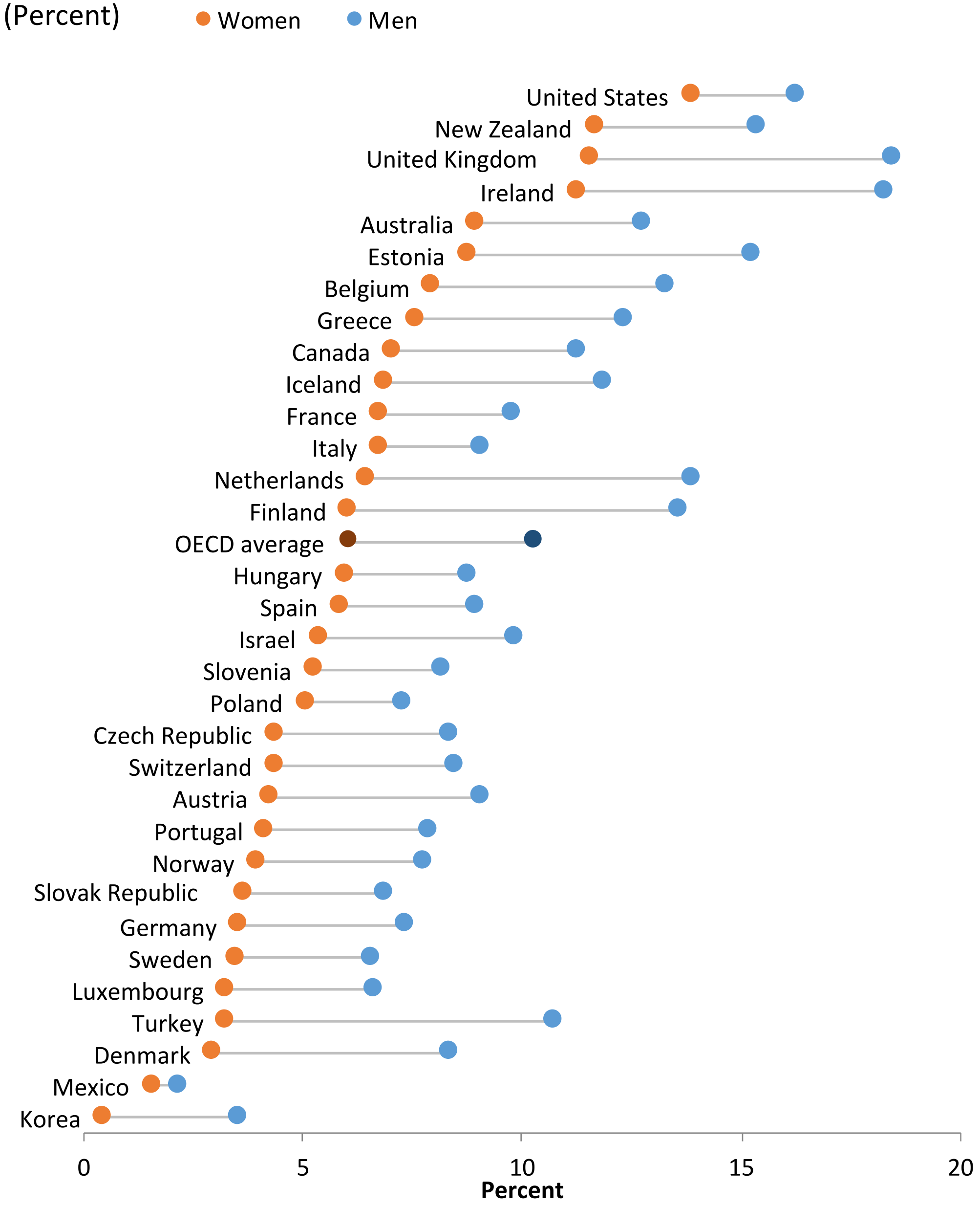


Percentage of Employed Who are Senior Managers, by Gender, 2008



Percentage of Employed Who are Senior Managers, by Gender, 2008

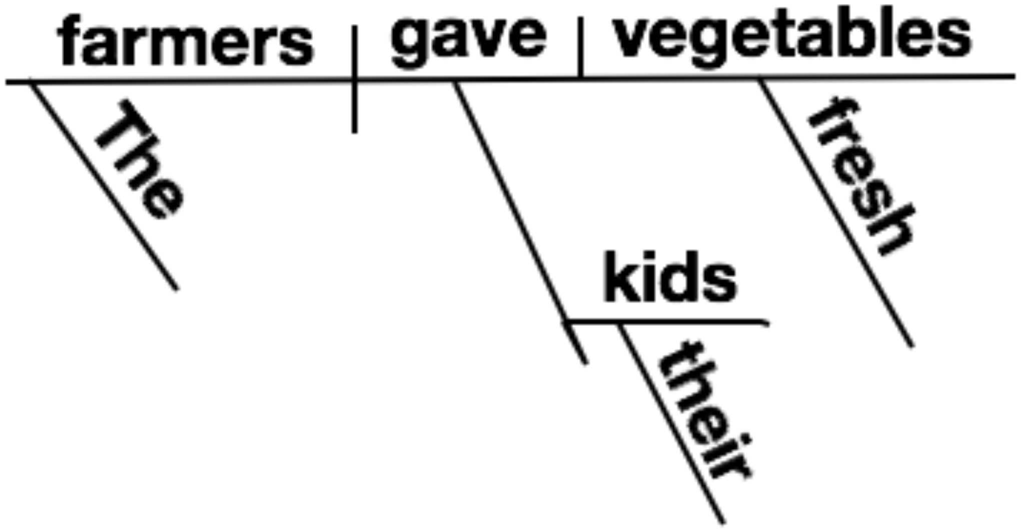
(Percent)

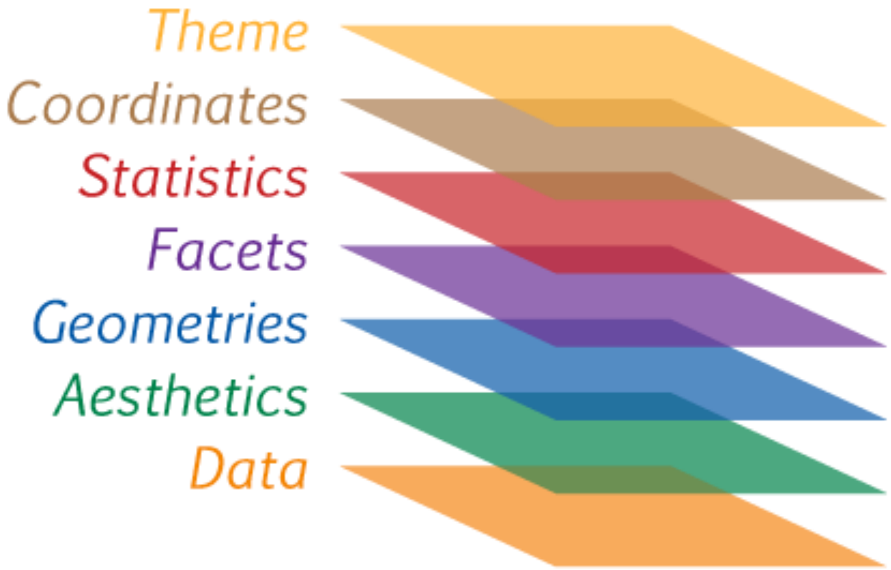


Grammar of Graphics

Grammar: framework that defines and establishes the structure of a language. If words are the lego blocks of language, grammar is the instruction booklet for how to put them together to make meaningful sentences.

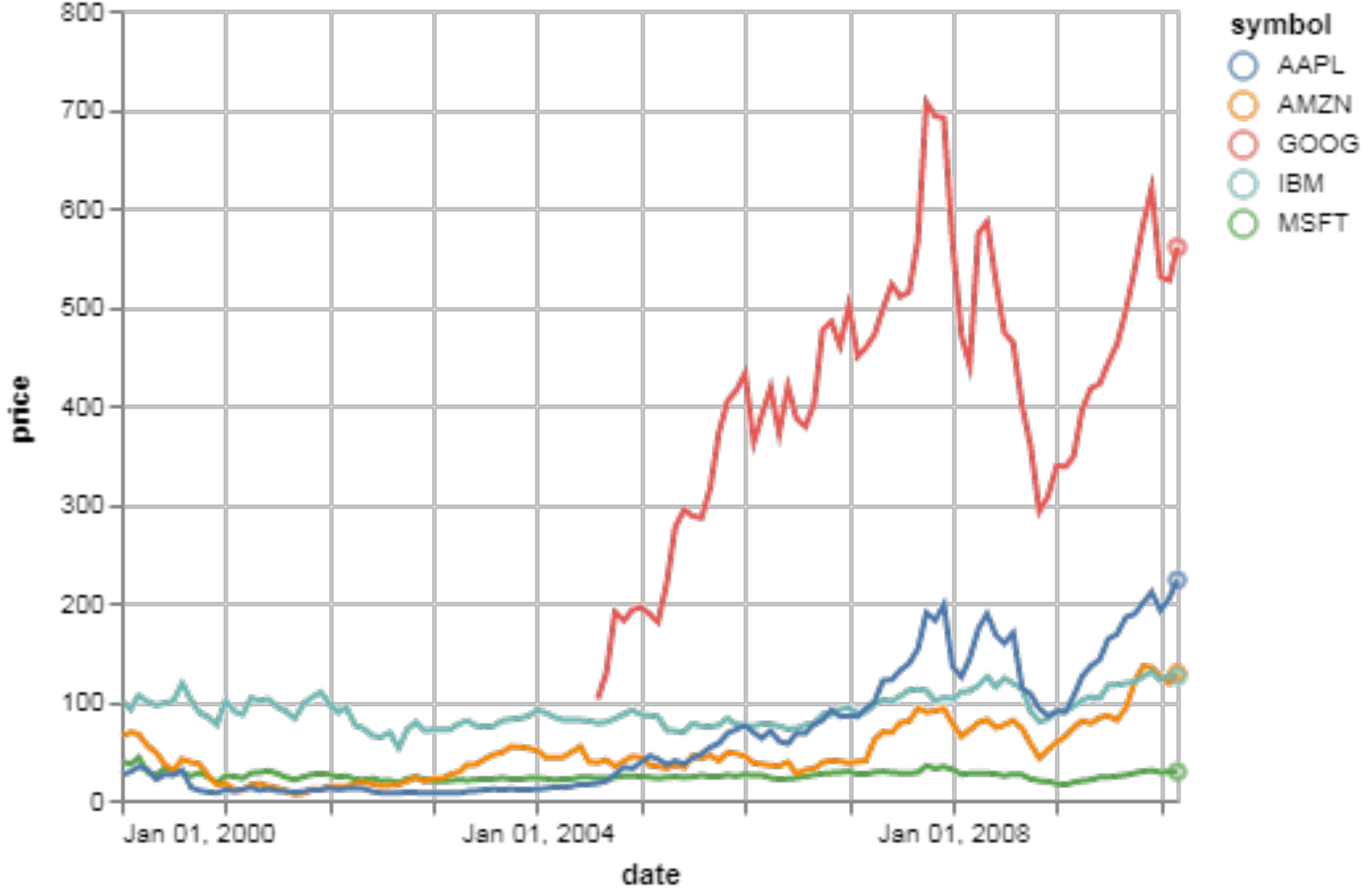
A grammar of graphics is a framework that breaks down any graphic into its components (layers) to allow us to concisely describe it.





Layers of any visualization

1. **Data**: Always start with the data, identify the dimensions you want to visualize.
2. **Aesthetics**: Confirm the axes based on the data dimensions, positions of various data points in the plot. Also check if any form of encoding is needed including size, shape, color and so on which are useful for plotting multiple data dimensions.
3. **Scale:** Do we need to scale the potential values, use a specific scale to represent multiple values or a range?
4. **Geometric objects:** These are popularly known as ‘geoms’. This would cover the way we would depict the data points on the visualization. Should it be points, bars, lines and so on?
5. **Statistics:** Do we need to show some statistical measures in the visualization like measures of central tendency, spread, confidence intervals?
6. **Facets:** Do we need to create subplots based on specific data dimensions?
7. **Coordinate system:** What kind of a coordinate system should the visualization be based on — should it be cartesian or polar?



https://github.com/altair-viz

Exercise in R

https://rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

References

- Everything Data course: “Effective Visualizations”, Duke University