

# Introduction to #DataVisualization

**Prerequisite:** You have **basic understanding of Python programming** and/or you have successfully **completed the module 1: Python Fundamentals**.

**Course environment:**

- Zoom for the live classes
- Jupyter Notebook for coding
- Google Classroom for the HW/quiz submission
- GitHub for the course materials

## Course objectives:

1. Matplotlib, a powerful Python data visualization library. Matplotlib provides the building blocks to create rich visualizations of many different kinds of datasets. You will learn how to create visualizations for different kinds of data and how to customize, automate, and share these visualizations.
2. Seaborn, for Statistical analysis and to create informative and attractive visualizations. You'll also learn about some of Seaborn's advantages as a statistical visualization tool.
3. Use of Numpy and Pandas for data analysis
4. Become a data storyteller

## **Course outline:**

Lecture 1: Importance of Data visualization, example of Bad data visualization

Lecture 2: Pie Chart, Bar plot

Lecture 3: Time-series plot, Line plot, Scatter plot, Regression plot

Lecture 4: Statistical analysis- Box plot, Histogram, Cumulative distribution function

Lecture 5: Advanced data visualization- Pair plot, Heatmap

Lecture 6: Case study 1

Lecture 7: Case study 2

Lecture 8: Data Visualization in Natural Language Processing (NLP) and Image Processing

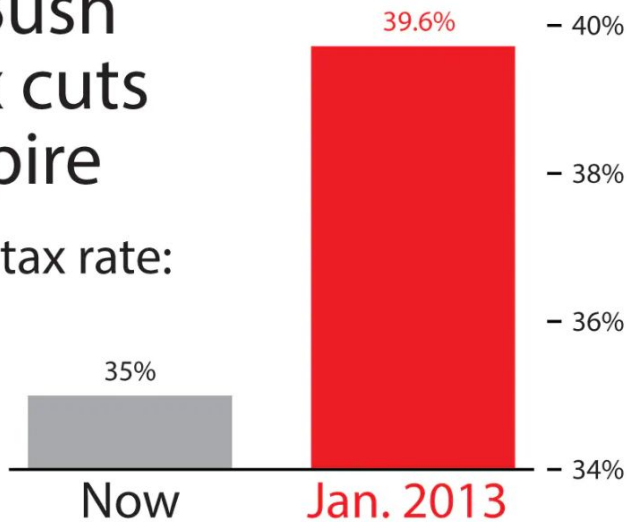
Contents may change based on the student's requirement or time constraint.

## Example of Bad Data Visualization

### Misleading

If Bush  
tax cuts  
expire

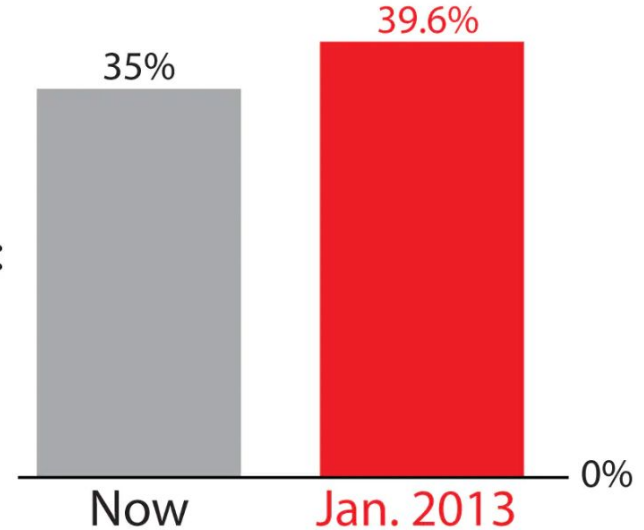
Top tax rate:



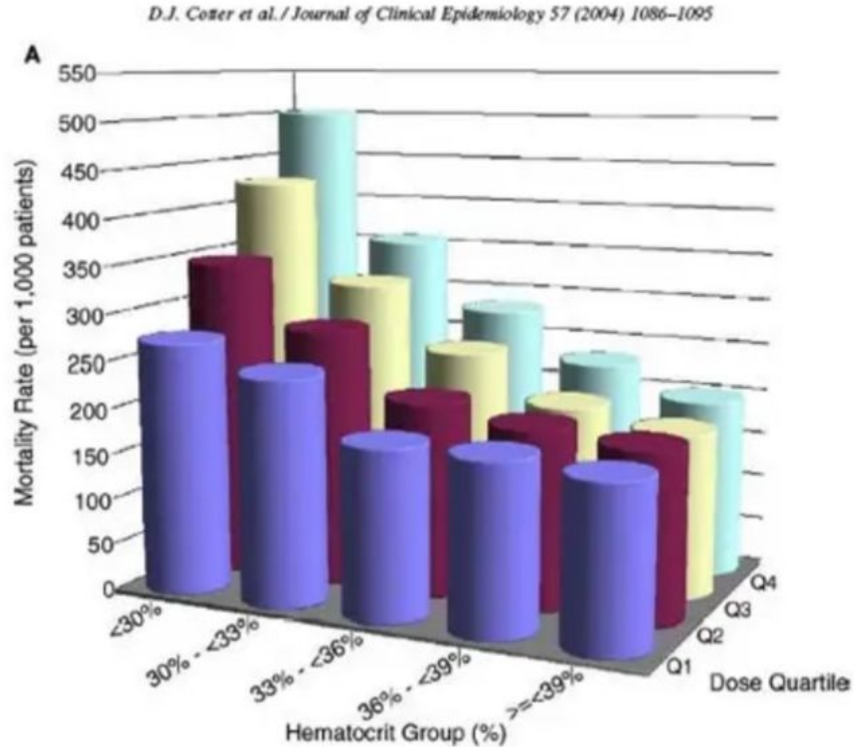
### More accurate

If Bush  
tax cuts  
expire

Top tax rate:



# Example of Bad Data Visualization

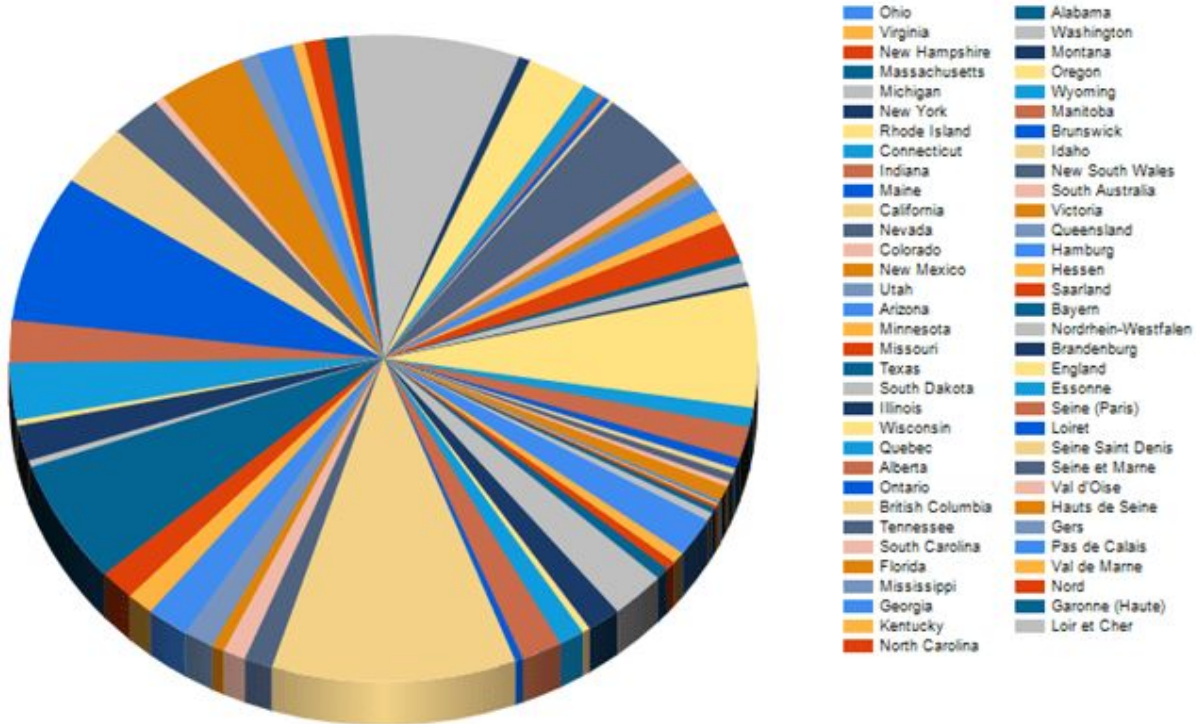


## Example of Bad Data Visualization



# Example of Bad Data Visualization

Chart Title

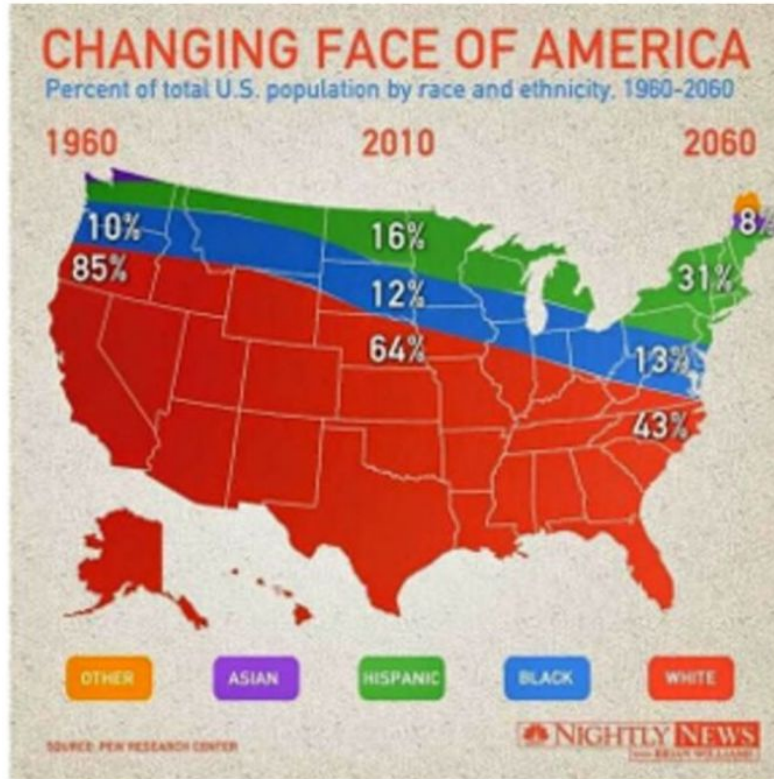




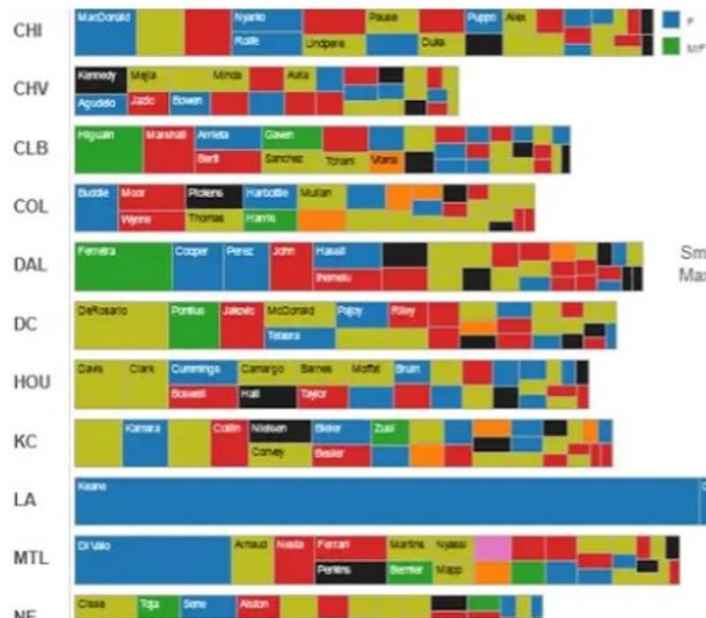
# Example of Bad Data Visualization



# Example of Bad Data Visualization



# Example of Bad Data Visualization



## MLS Salaries per May 1, 2013 MLS Players Union Release

Main graph driven by Base Salary

Smaller graph is driven by estimated salary cap cost of top 20 highest-paid, non-Generation-Adidas players on each club. Maximum of \$365,750 with adjustments for lower-paid DPs, and Youth DPs. **This is still only an estimate of cap usage.**

It has been noted by MLS clubs that this data (both individual salaries & positions) can be a bit inaccurate.

If you like these visualizations, please follow me on Twitter [@SoccerStatHunt](#)

created using Tableau Public

## Example of Bad Data Visualization



# Example of Bad Data Visualization

## Cities with best Batting averages

Sydney comes first with 5416 runs at an average of 45.37. It has produced some of the most prolific batsmen like David Warner, Steven Smith, Michael Clarke, Steve Waugh, and Mark Waugh. Next is Launceston with 2657 runs at an average of 44.80. This is mainly due to players like Ricky Ponting, David Boon and George Bailey.



## Cities with best Bowling averages

Sydney is on top with 156 wickets at an average of 34.5, thanks to bowlers like Mitchell Starc, Jason Gillespie, Pat Cummins, and Nathan Bracken. Christchurch is second with 179 wickets at an average of 34.6. The city has produced some very fine bowlers like Shane Bond, Richard Hadlee, Matt Henry and Chris Harris. It is also the birthplace of England's hero from the final, Ben Stokes.



# Example of Bad Data Visualization

## MOST WICKETS IN DEATH OVERTS IN ODIS

SINCE THE START OF JANUARY 2017

■ WKTS ■ AVE



NUMBERS UPDATED TILL MAY 14, 2019

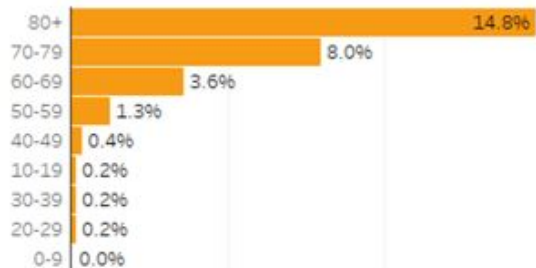


# Example of Bad Data Visualization

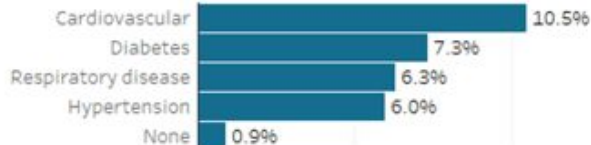
## Death rate varies by age, health and sex

Case fatality ratio

Age



Health Condition



Sex



Source: Chinese Centre for Disease Control and Prevention

NOTE: this is a redesign of a BBC graphic by Andy Cotgreave, exploring axis lengths. The redesign is in response to a tweet from Alice Casey (@cased)

## What percent of people who contract coronavirus die (estimated)?

Case fatality ratio

Age



Health Condition



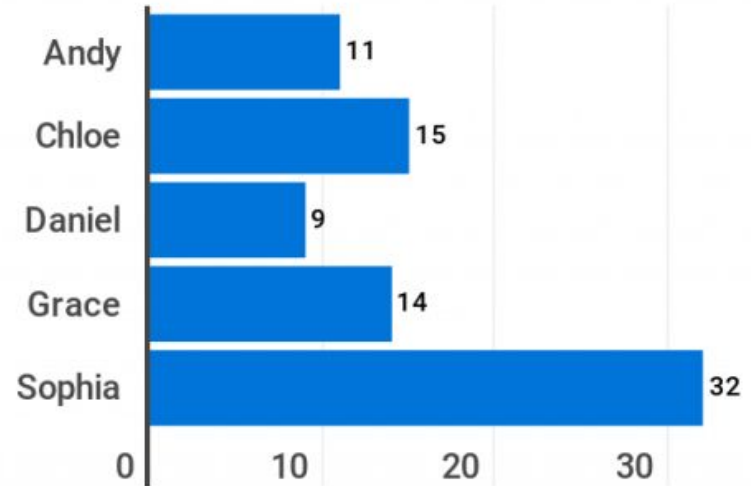
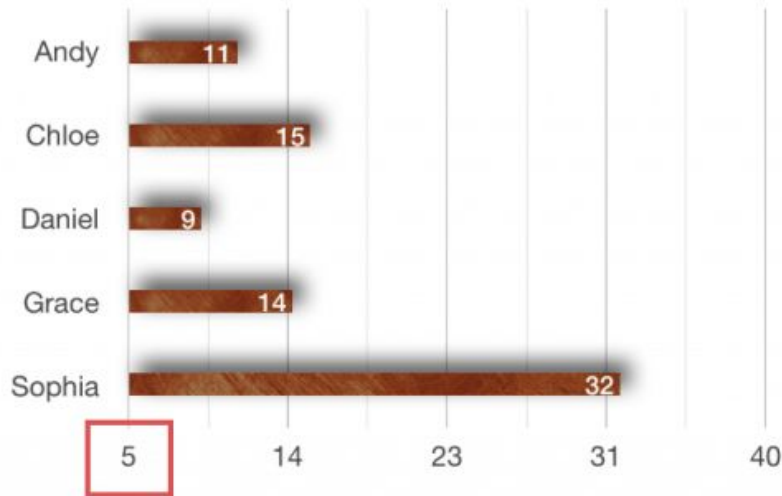
Sex



Source: Chinese Centre for Disease Control and Prevention

NOTE: this is a redesign of a BBC graphic by Andy Cotgreave, exploring axis lengths. The redesign is in response to a tweet from Alice Casey (@cased)

## Example of Bad Data Visualization



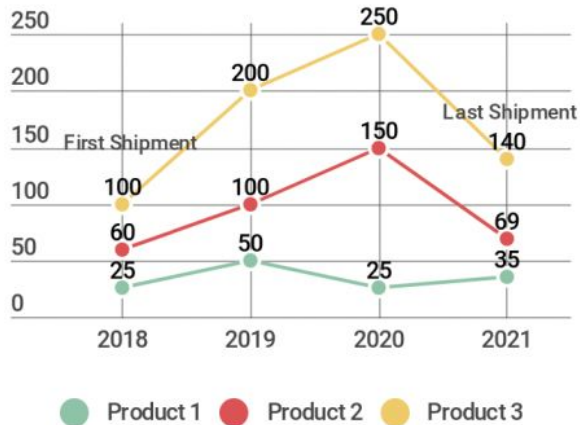


# Example of Bad Data Visualization

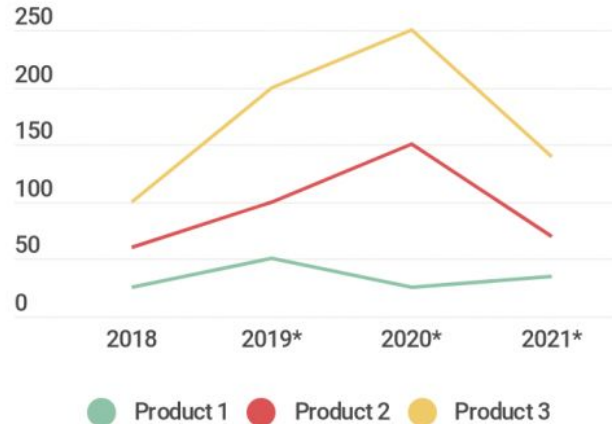


## Bad, Vague Title

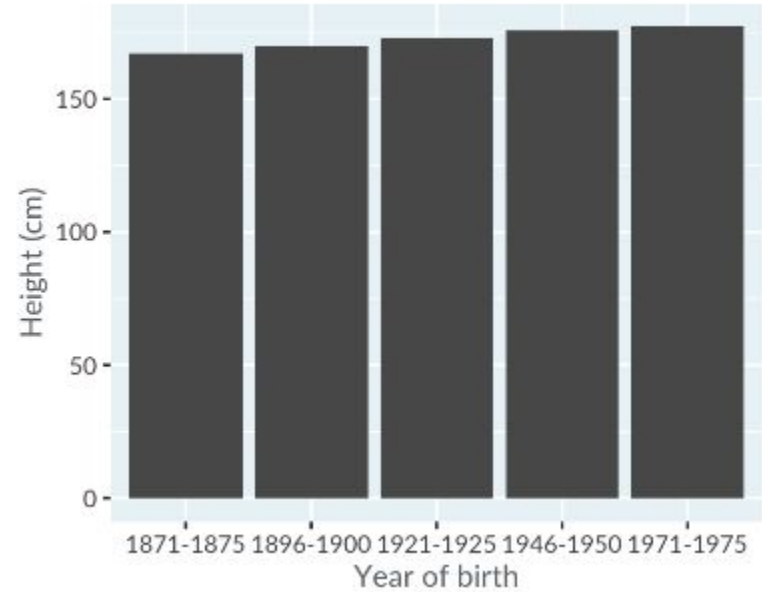
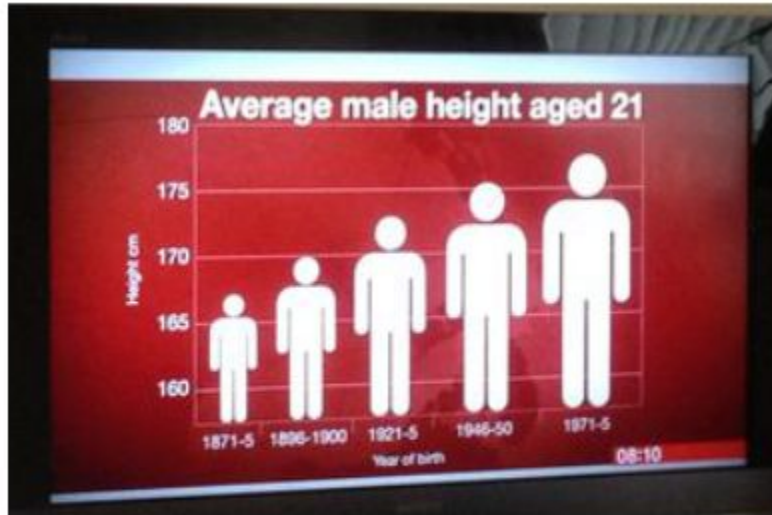
Text explaining the data stated below.



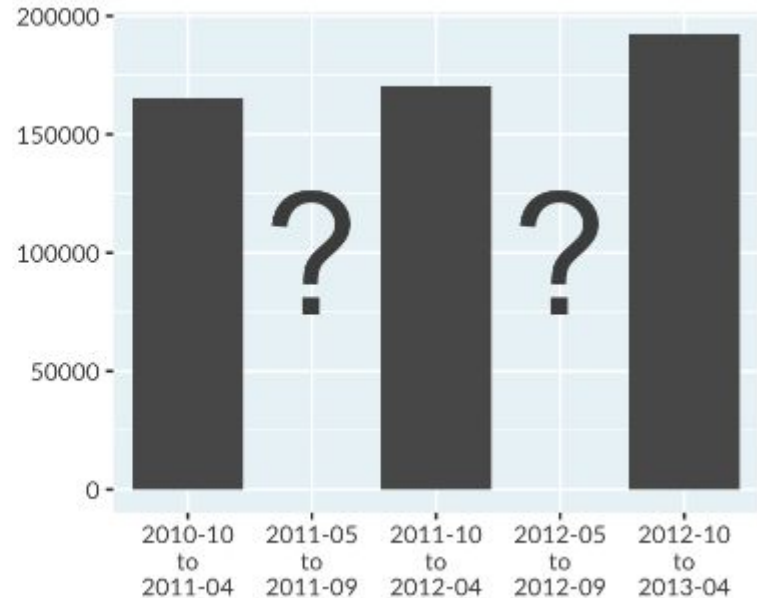
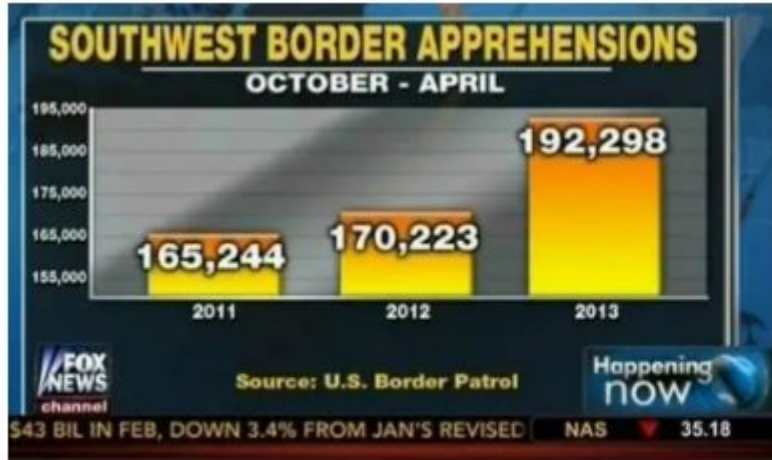
## Good, Memorable Title



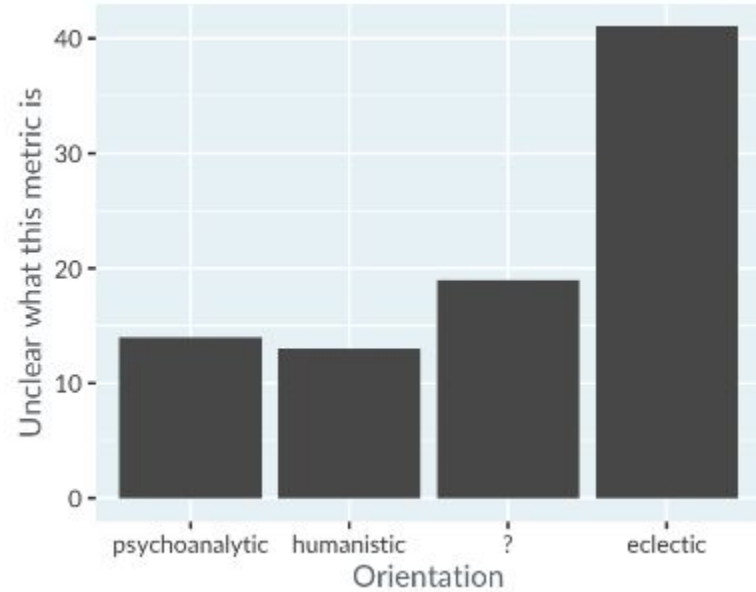
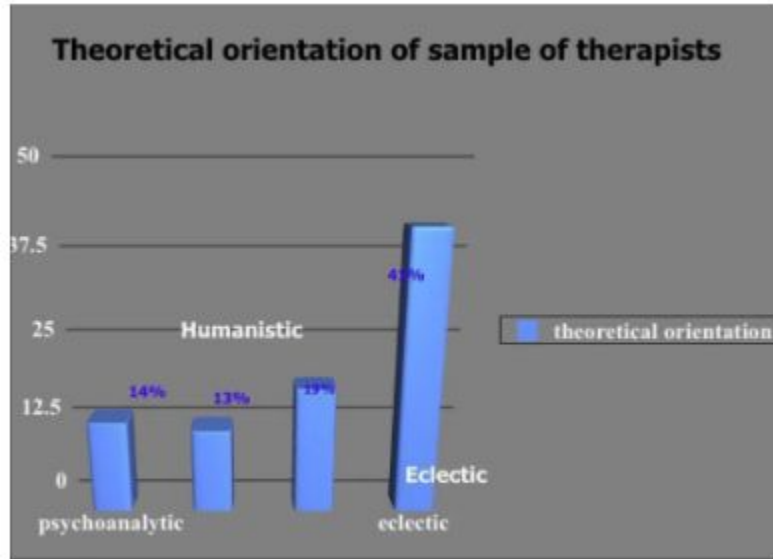
## Example of Bad Data Visualization



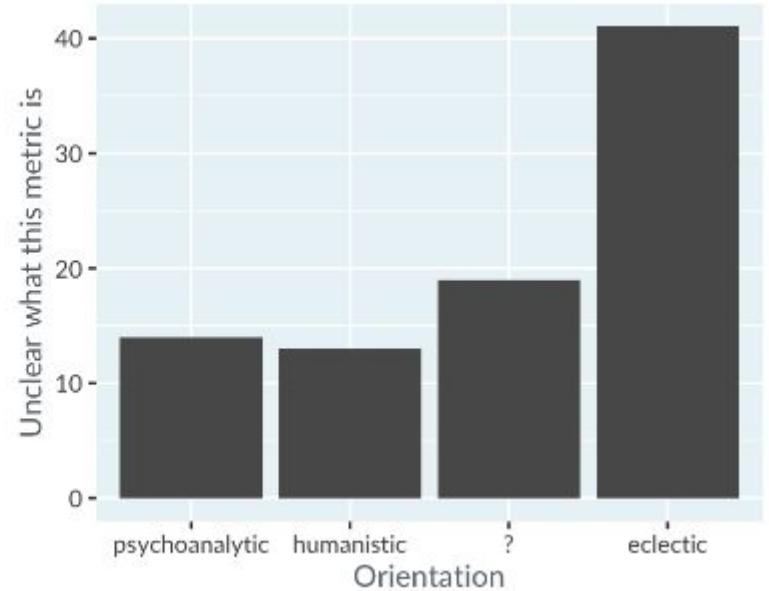
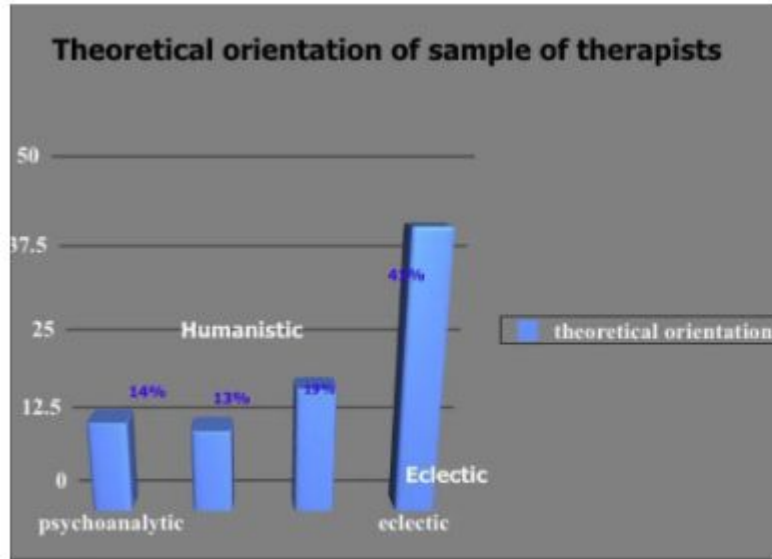
# Example of Bad Data Visualization



# Example of Bad Data Visualization



# Example of Bad Data Visualization



# References

- [1] <https://www.jotform.com/blog/bad-data-visualization/>
- [2] <https://medium.com/nightingale/bad-data-visualization-in-the-time-of-covid-19-5a9f8198ce3e>
- [3] <https://www.espncricinfo.com/story/which-top-cricket-city-would-win-the-world-cup-1196522>
- [4] <https://www.tableau.com/about/blog/2020/3/covid-19-resources-data-viz-best-practices>
- [5] <https://infogram.com/blog/do-this-not-that-data-visualization-before-and-after-examples/>
- [6] <https://www.washingtonpost.com/business/2019/10/14/youve-been-reading-charts-wrong-heres-how-pro-does-it/>
- [7] <https://junkcharts.typepad.com/>
- [8] [https://www.reddit.com/r/dataisugly/comments/dh6yra/an\\_actual\\_graph\\_presented\\_in\\_my\\_psychology/](https://www.reddit.com/r/dataisugly/comments/dh6yra/an_actual_graph_presented_in_my_psychology/)