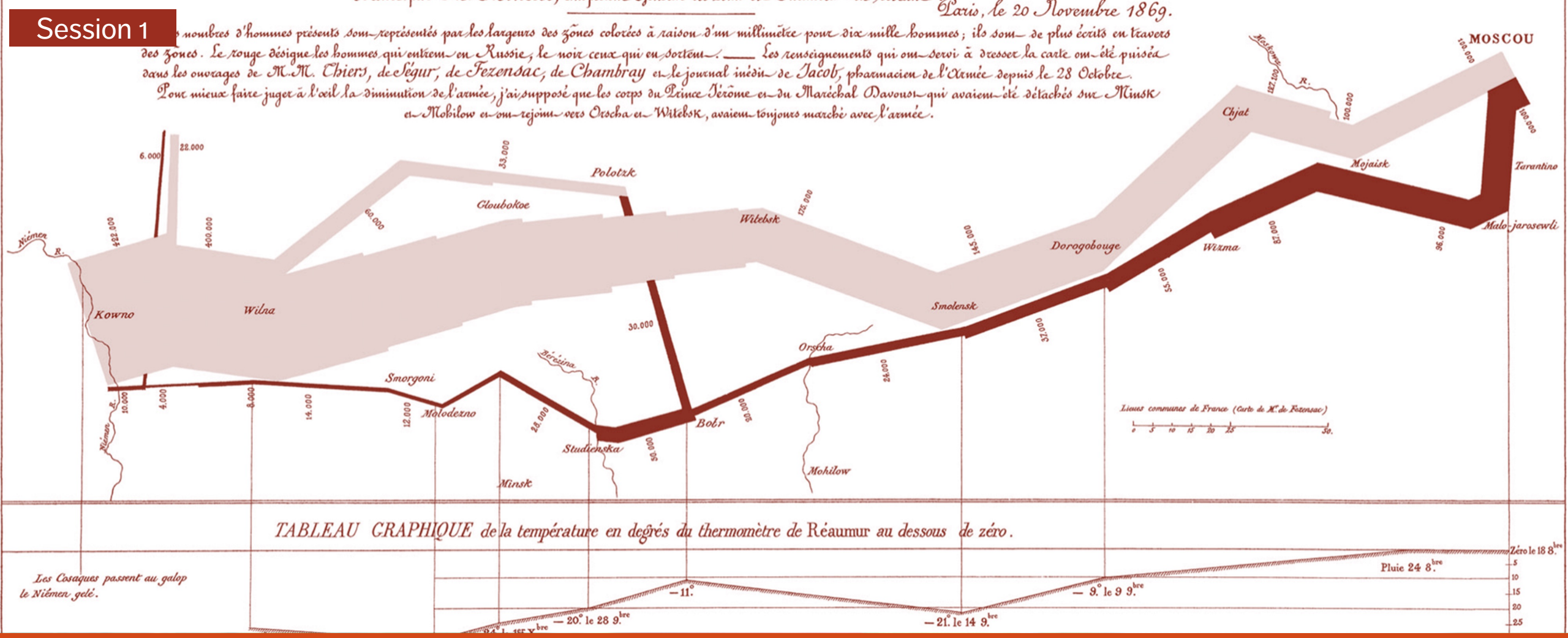
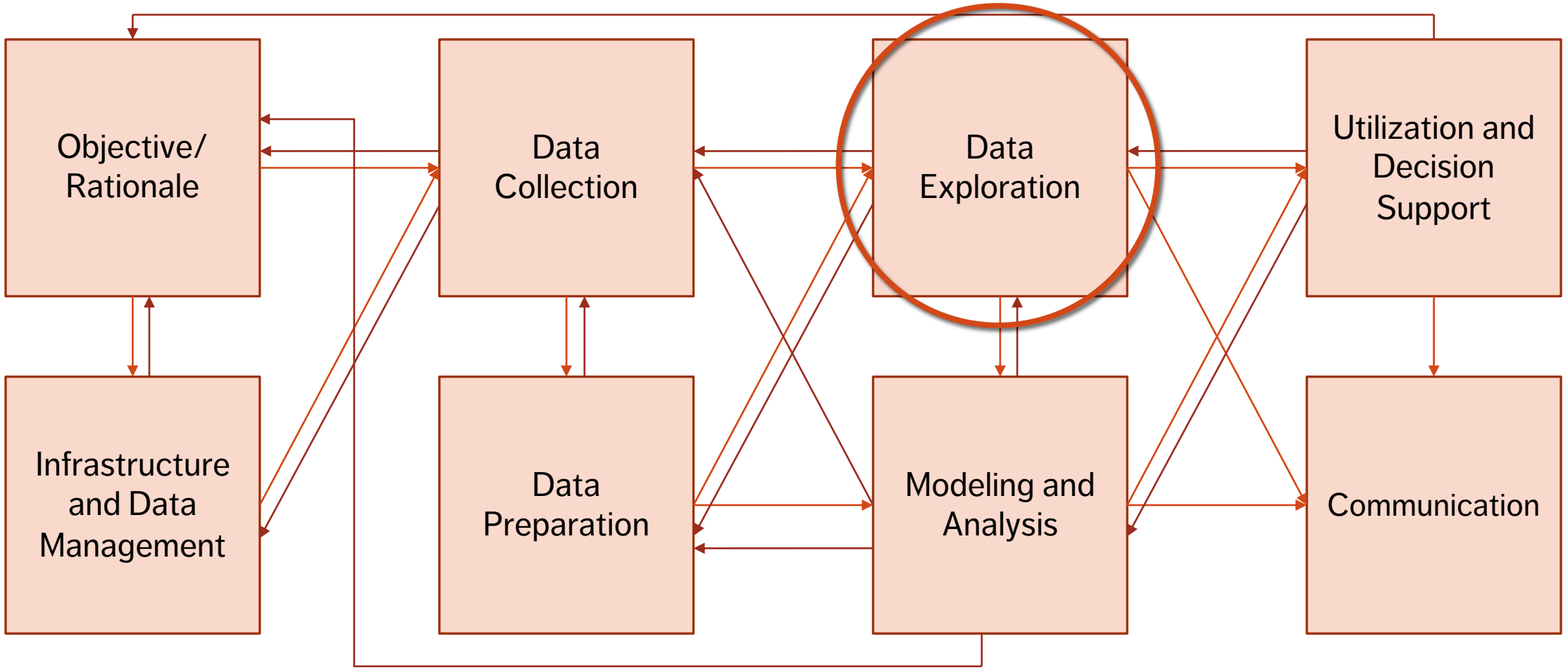


1. Exploratory Data Analysis

The (Messy) Analysis Process



Some Basic Questions

What **system** does your data represent – **objects**, **attributes**, **relationships**?

How does it represent this system – i.e., the **data model**?

Who made this dataset? When? For what purpose?

Assuming a flat file – what do the rows and columns represent?

Do you have enough information (e.g., **metadata**) to answer these questions? Where can you find more information?

Non-Visualization Summaries

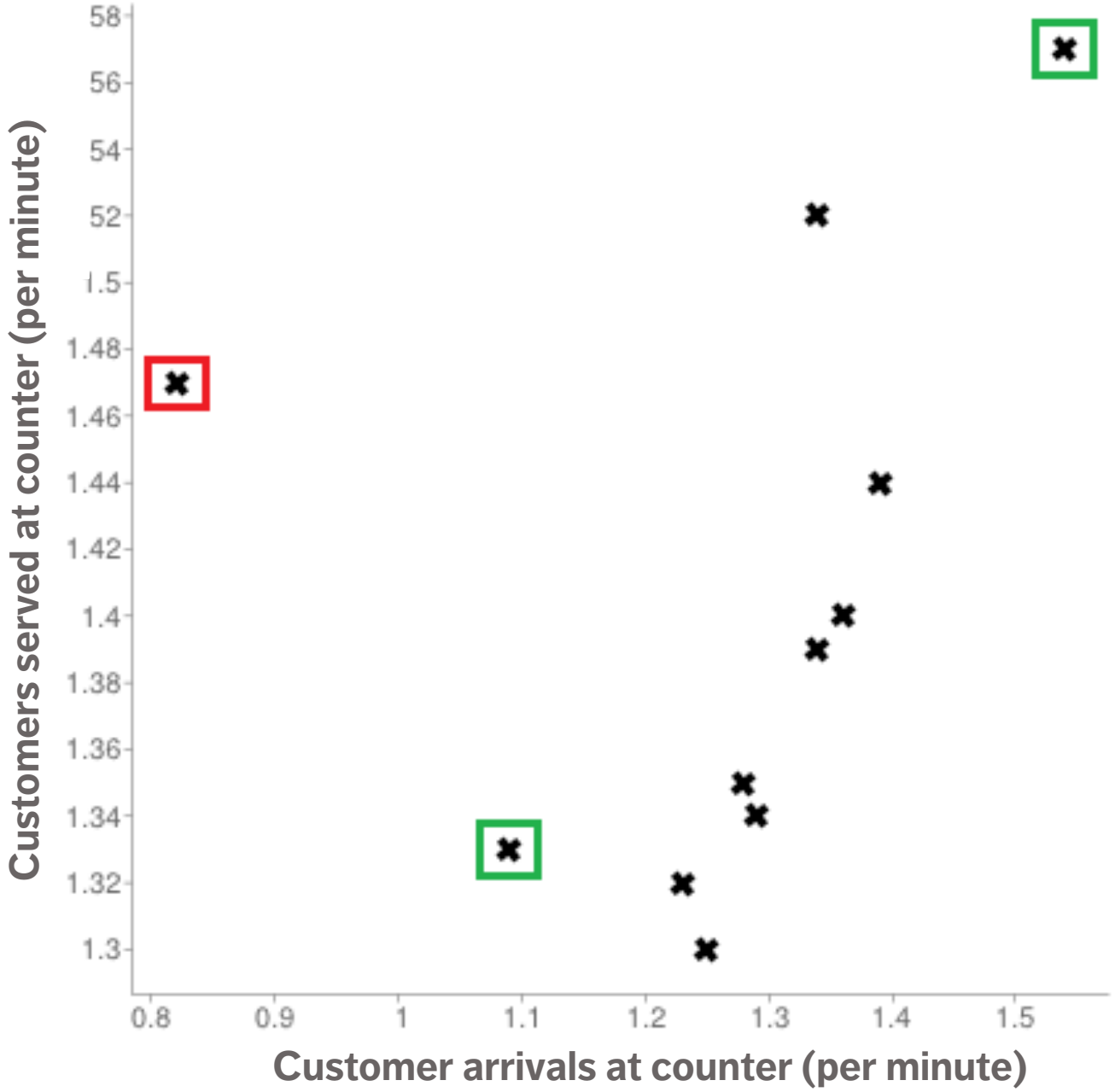
```
       Cl                NO3                NH4
Min.   :  0.222   Min.   : 0.000   Min.   :    5.00
1st Qu.: 10.994   1st Qu.: 1.147   1st Qu.:   37.86
Median : 32.470   Median : 2.356   Median :  107.36
Mean   : 42.517   Mean   : 3.121   Mean   :  471.73
3rd Qu.: 57.750   3rd Qu.: 4.147   3rd Qu.:  244.90
Max.   :391.500   Max.   :45.650   Max.   :24064.00
NA's   :16        NA's   :2        NA's   :2

   season
Length:340
Class :character      autumn spring summer winter
Mode  :character          80     84     86     90
```

Pre-Analysis Use

Data visualization can be used to set the stage for analysis:

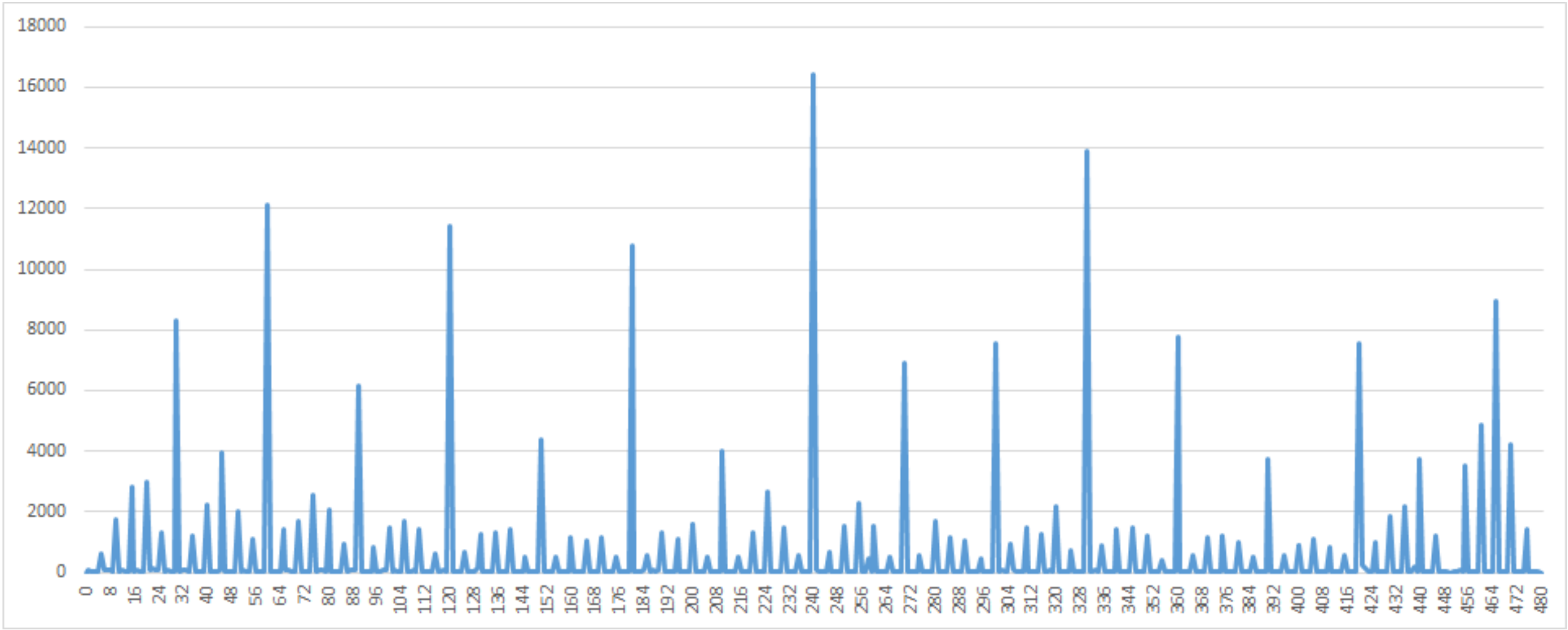
- **detecting anomalous entries**
  invalid entries, missing values, outliers
- **shaping the data transformations**
  binning, standardization, Box-Cox transformations, PCA-like transformations
- **getting a sense for the data**
  data analysis as an art form, exploratory analysis
- **identifying hidden data structure**
  clustering, associations, patterns informing the next stage of analysis

[Personal dataset]



Session 1

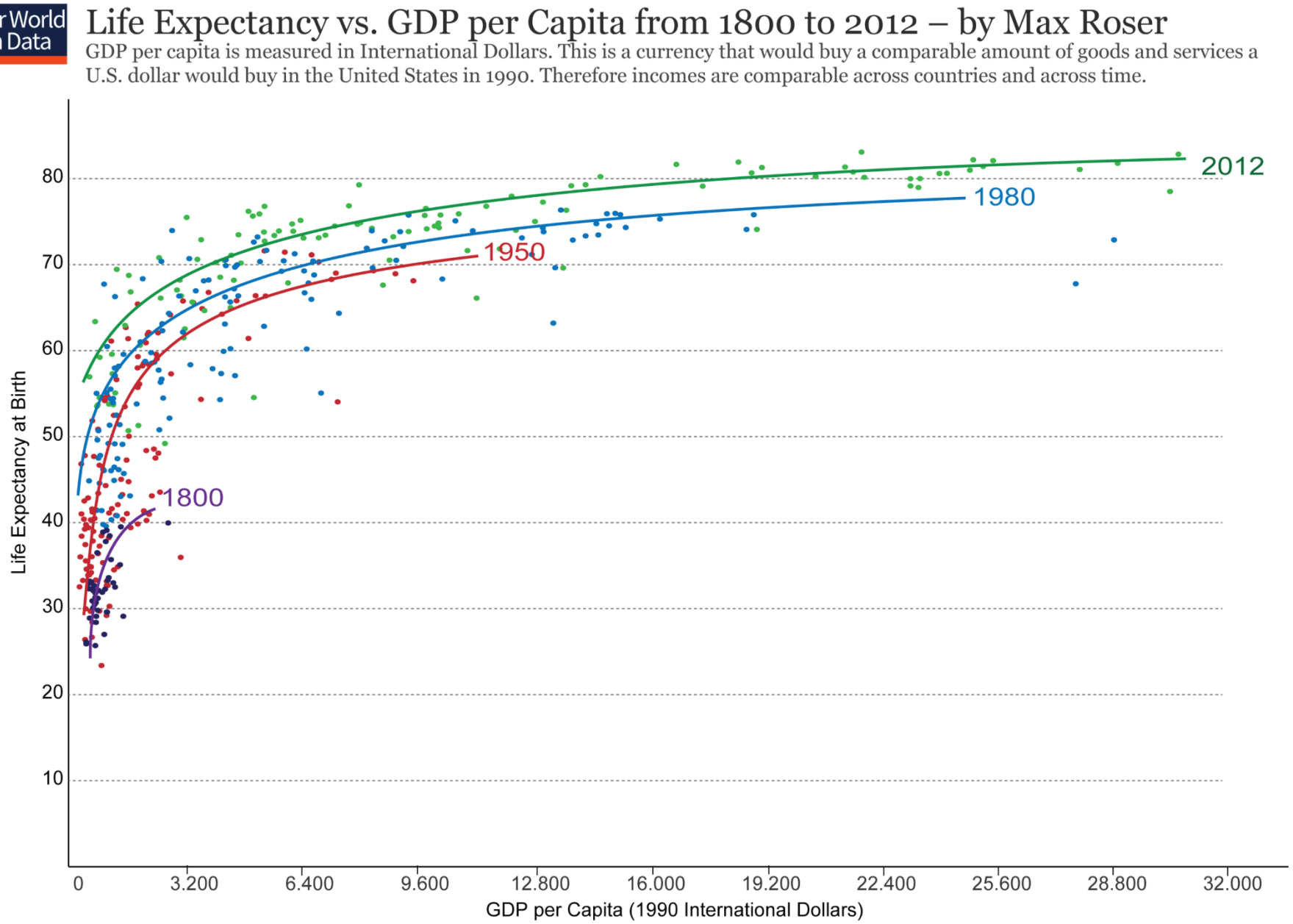
[Personal dataset]



Self-reported work hours (mins)

Session 1

[https://ourworldindata.org/life-expectancy]



This graph displays the correlation between life expectancy and GDP per capita.

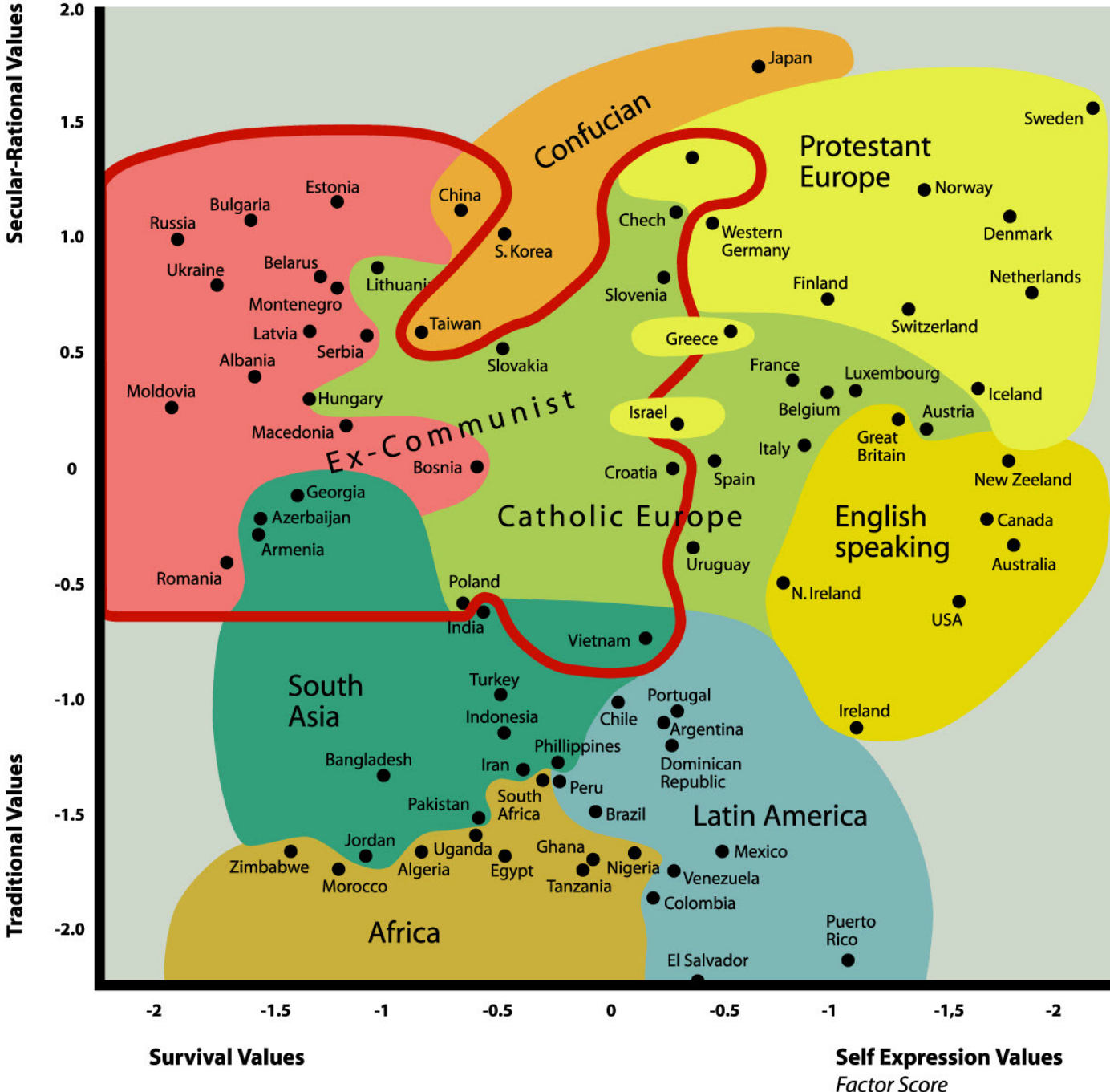
Countries with higher GDP have a higher life expectancy, in general.

The relationship seems to follow a logarithmic trend: the unit increase in life expectancy per unit increase in GDP decreases as GDP per capita increases.

Session 1

Inglehart-Welzel's Global Cultural Map (2010-2014)

[https://en.wikipedia.org/wiki/World_Values_Survey]



Traditional values
importance of religion, parent-child ties, deference to authority and traditional family values.

Secular-rational values
less emphasis on religion, traditional family values and authority.

Survival values
emphasis on economic and physical security.

Self-expression values
high priority to environmental protection, growing tolerance of foreigners, gays and lesbians and gender equality

Workhorse Data Exploration Charts

Text and Tables

Rug Charts/Number Lines

Histograms/Bar Charts

Boxplots

Line Graphs

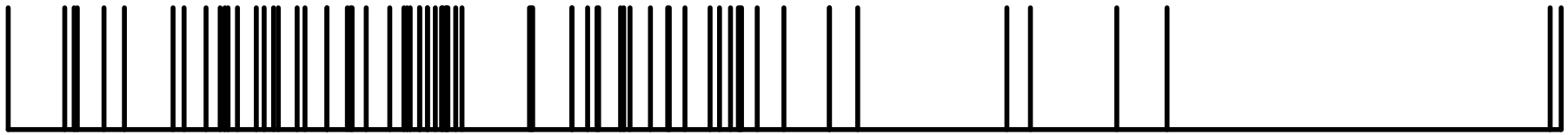
Scatterplots

Line Chart/Rug Chart

Gaps in the number line: **absence** of those numeric values in the data.

Remember: this is (possibly) different from the order that values appear in the dataset – since it is a number line, it shows where the values fall numerically.

If some values are identical, they lie on top of each other (use **jitter**?).

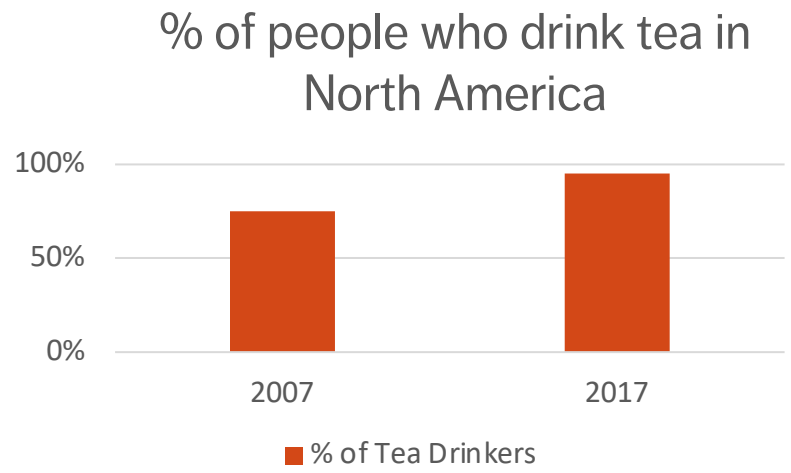


Simple Text

One or two numbers to focus on.

Good at “setting the scene”.

Draws focus to an area of the report.



95% of the population drinks tea today compared to **75%** in 2007

Tables

Tables interact with our **verbal** system, which means we **read** them:

- used to compare values
- audiences will look for their rows

Name	Last Year	This Year
Bob	20	30
Fred	30	40
George	10	15

Table design needs to **blend** into background

- the data should stand out, not the borders
- dense table/data: use alternating row colour

Name	Last Year	This Year
Bob	20	30
Fred	30	40
George	10	15

Table Heatmaps

Leverage colour to convey magnitude

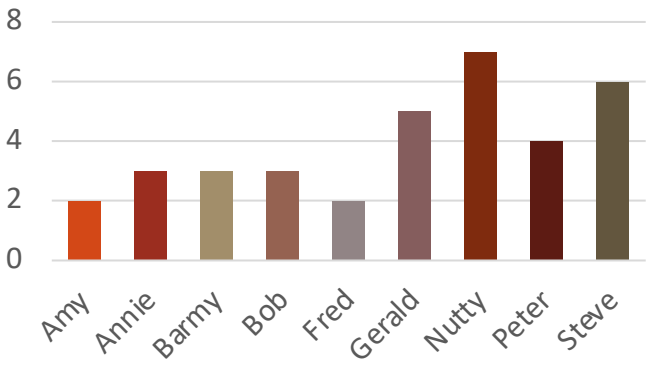
- use **single colour saturation** rather than differentiation (different colours)
- with a legend (white = low, blue = high), numbers can be removed without altering the message

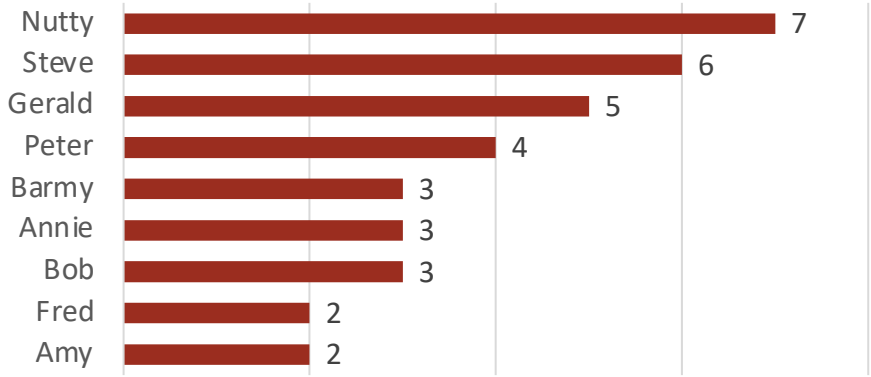
	Last Year	This Year	Next Year	Optimum
George	20	20	20	20
Peter	40	35	30	25
John	10	10	5	5
Sandra	25	30	35	40

	Last Year	This Year	Next Year	Optimum
George	20	20	20	20
Peter	40	35	30	25
John	10	10	5	5
Sandra	25	30	35	40

	Last Year	This Year	Next Year	Optimum
George				
Peter				
John				
Sandra				

Bar Charts





Very versatile and useful.

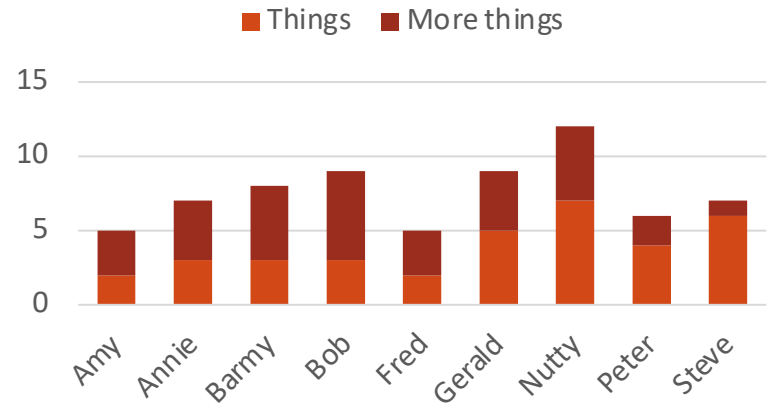
ALWAYS (?) have a zero baseline.

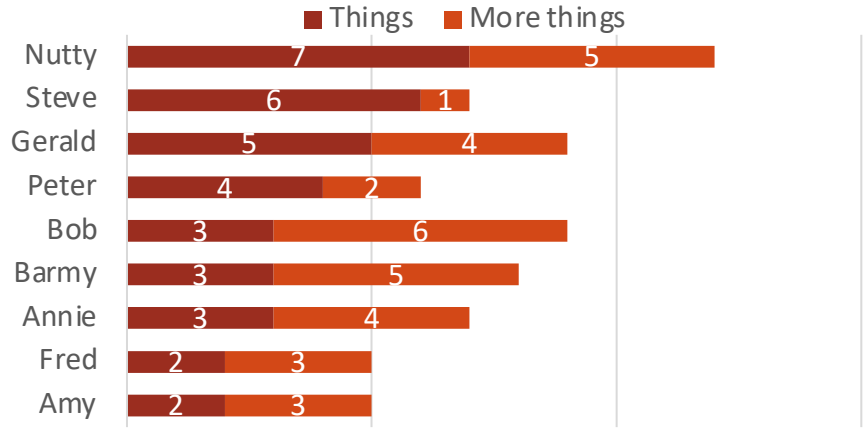
Use graph axis OR data labels. Axis for broad statements, data labels for more detail.

Horizontal charts are apparently **easier to read** (according to many studies).

Think about the ordering of categories.

Stacked Bar Charts



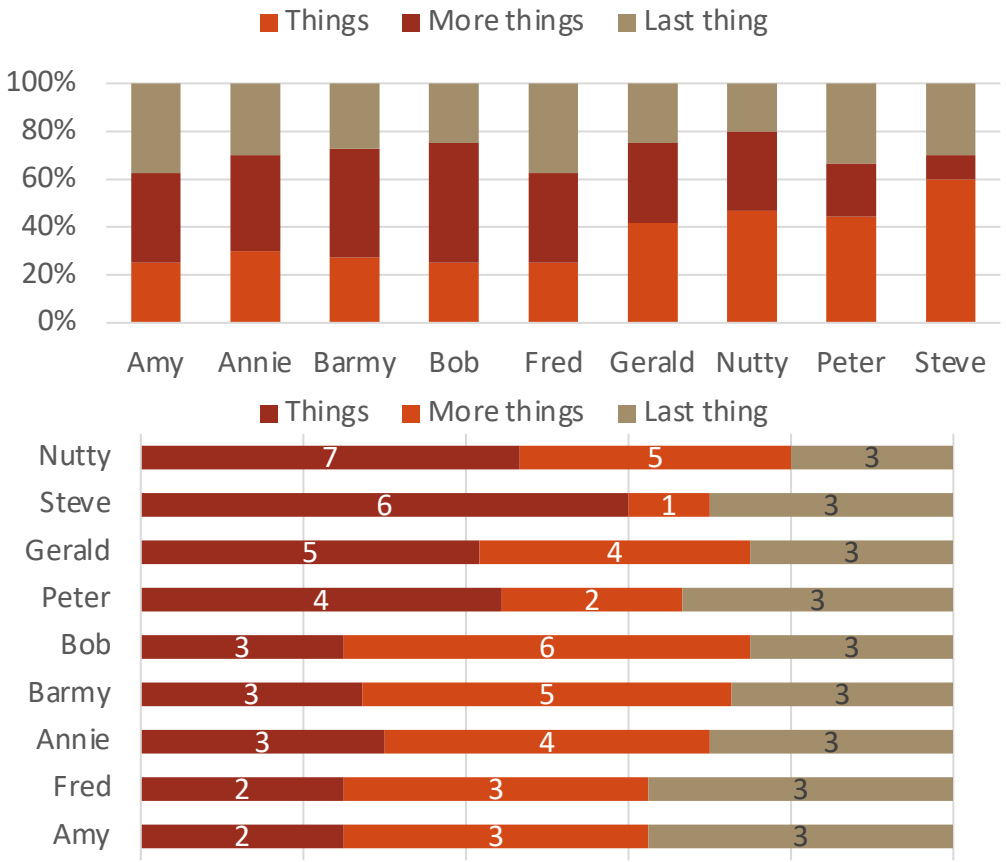


Designed for **comparing totals**, but can quickly become **overwhelming**.

Hard to sort / order.

Filtering is complicated in Power BI (what do you click on & how the chart responds when filter is clicked on?)

100% Bar Charts



Work well for visualizing **proportions** of a whole on a scale from negative to positive.

Consistent baseline on far left and right.

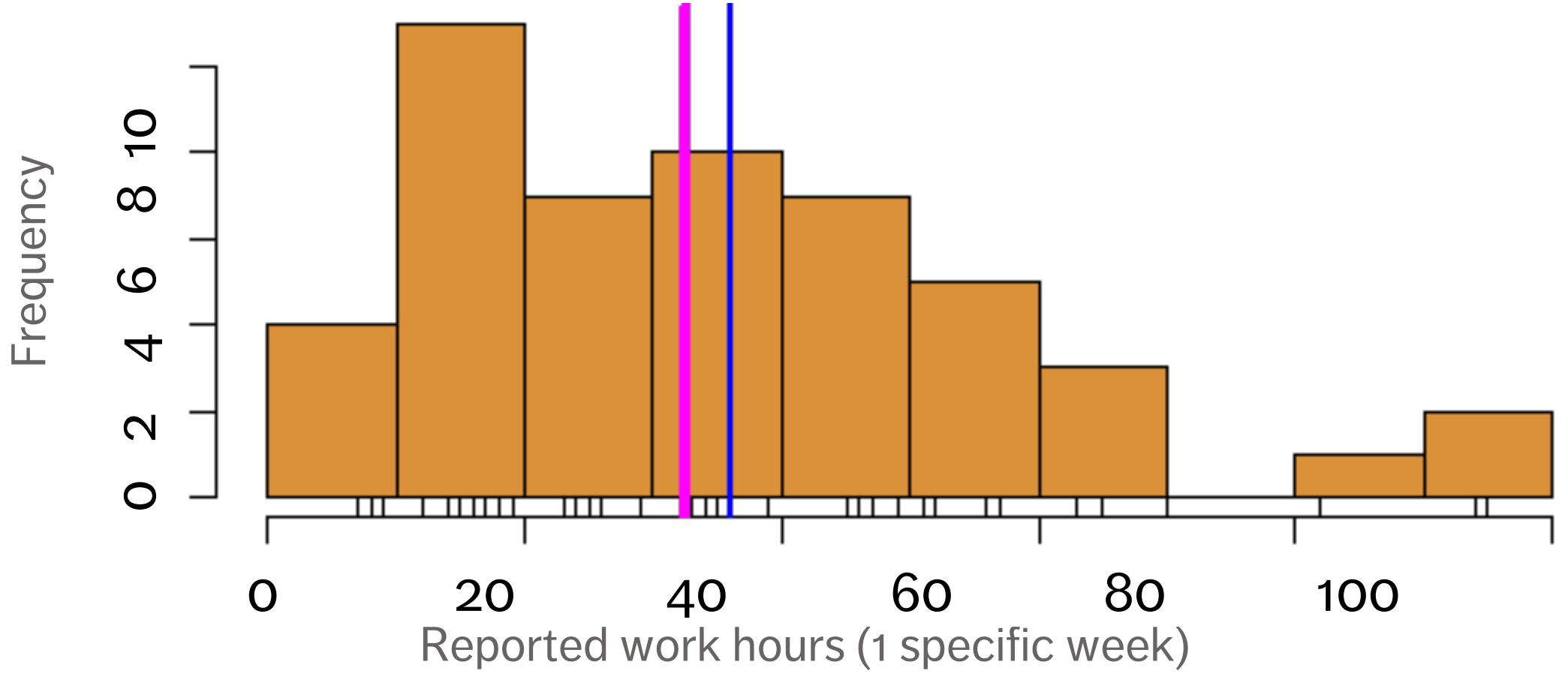
Easy to compare.

No relative measure to **magnitude** of data.

Research shows that horizonal is easier to process than vertical.

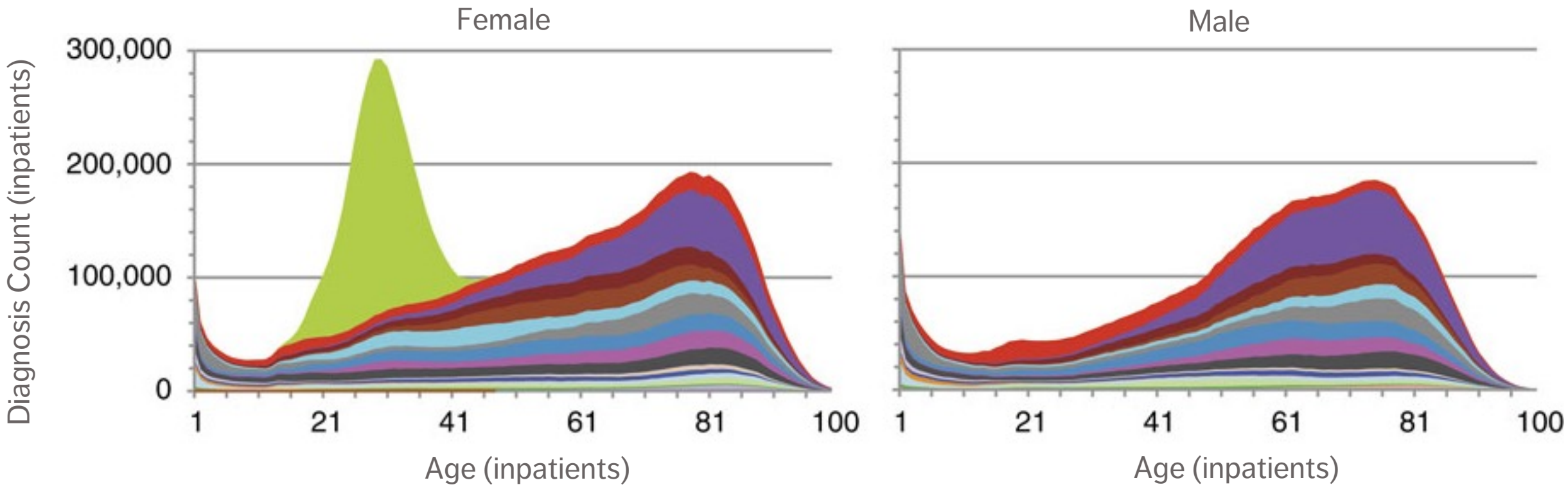
[Personal file]

Histogram



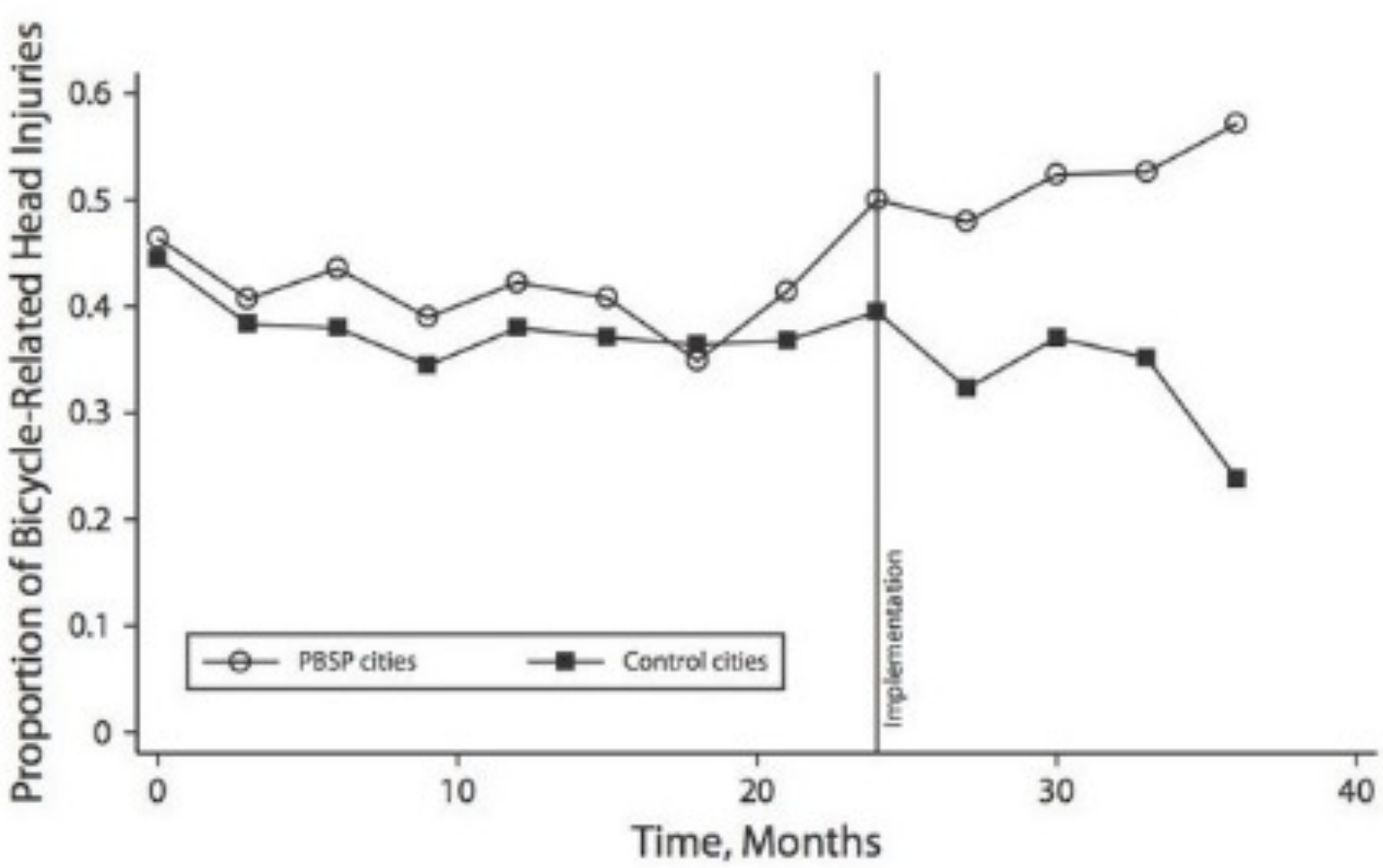
[A.B. Jensen et al., *Danish Medical Study*, Nature Communications]

Stacked Histograms



[Graves et al., Am.J.Phys.Health, 2014]

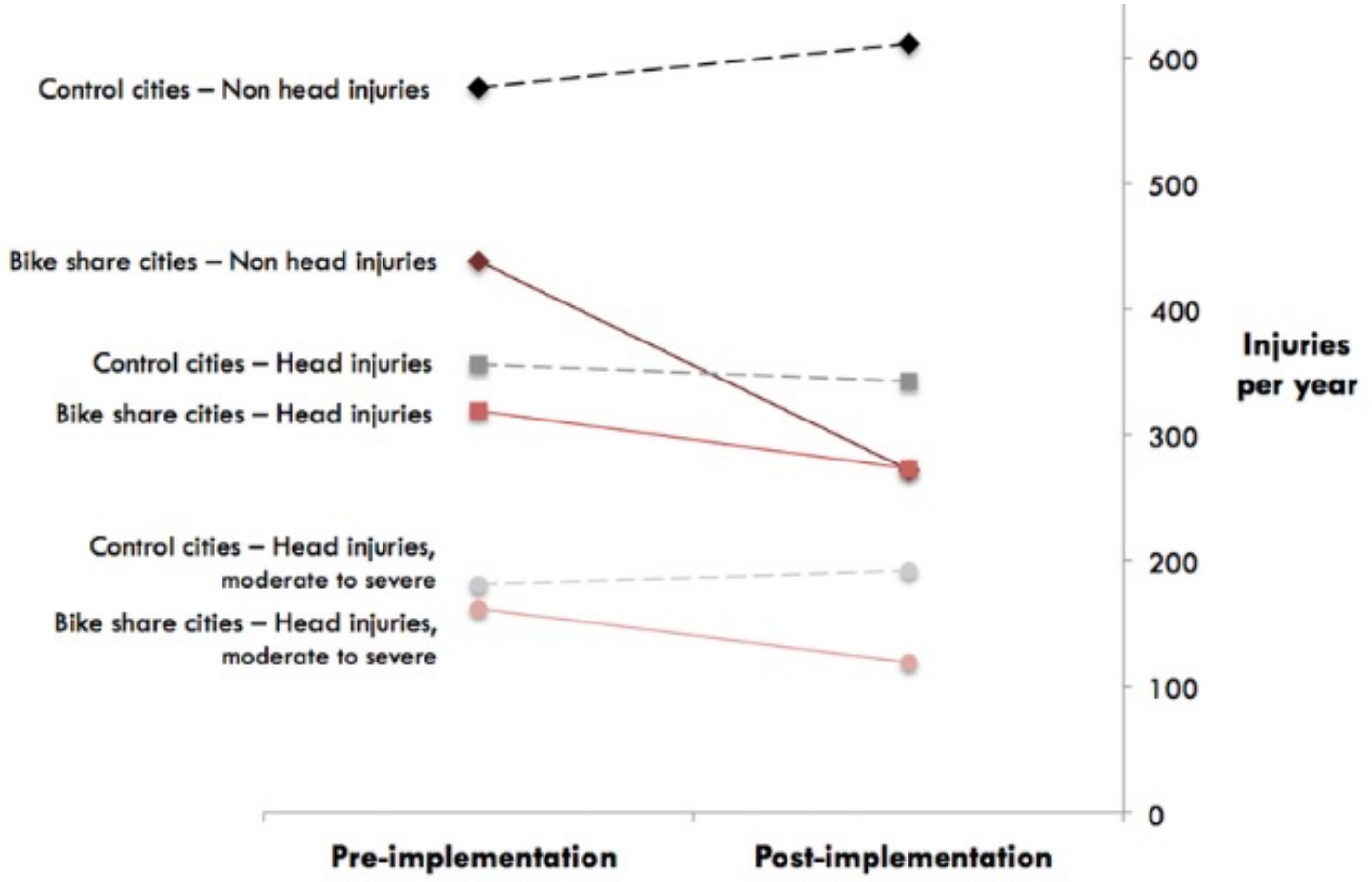
Line Graphs



Proportion of all bicycle-related injuries that were classified as head injuries among cities with public bike share programs and control cities, centered on intervention date (vertical line); North America.

[Teschke's rebuttal]

Line Graphs

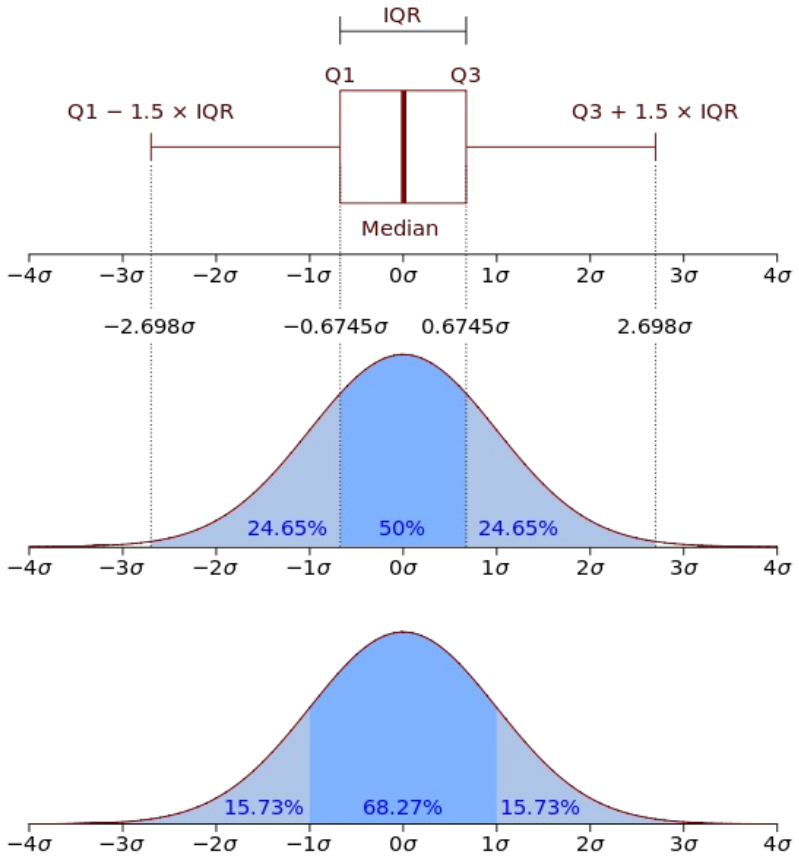


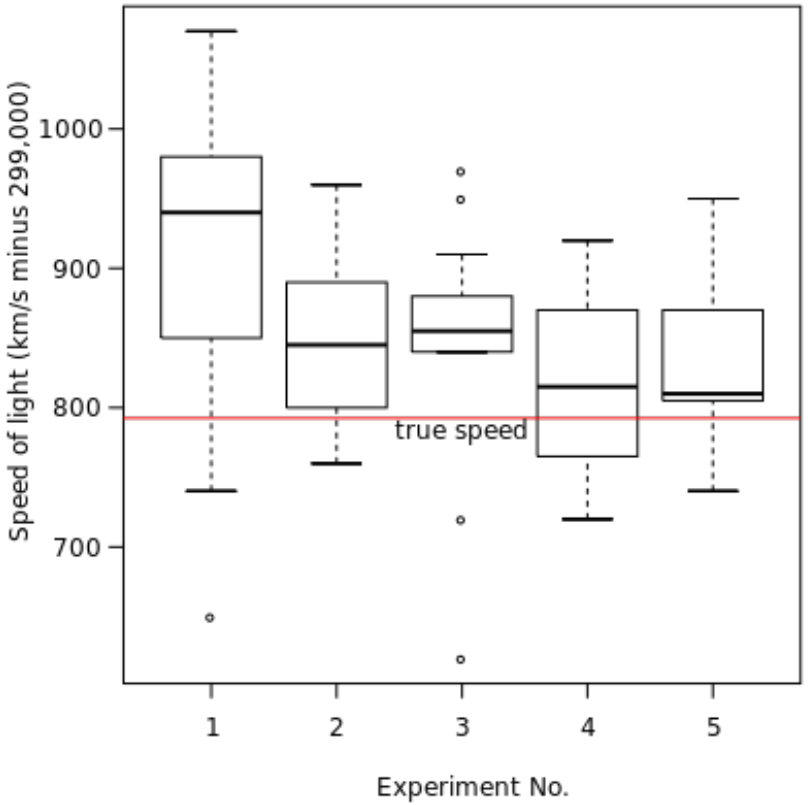
Data from new study show declines in all injuries, including head injuries after bike share system implemented.

Because head injuries decline less than other injuries, they are now a larger proportion of all injuries.

[Files from Wikipedia]

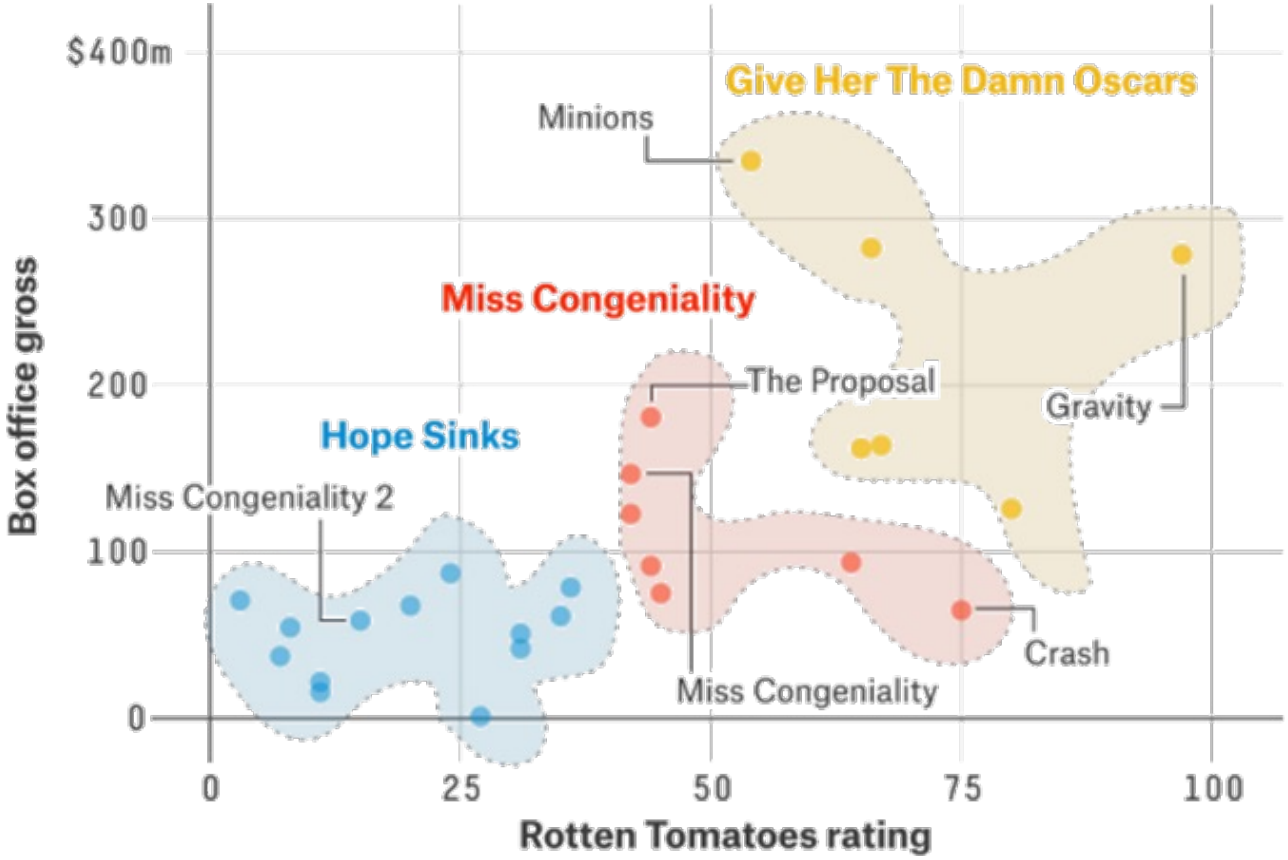
Boxplots

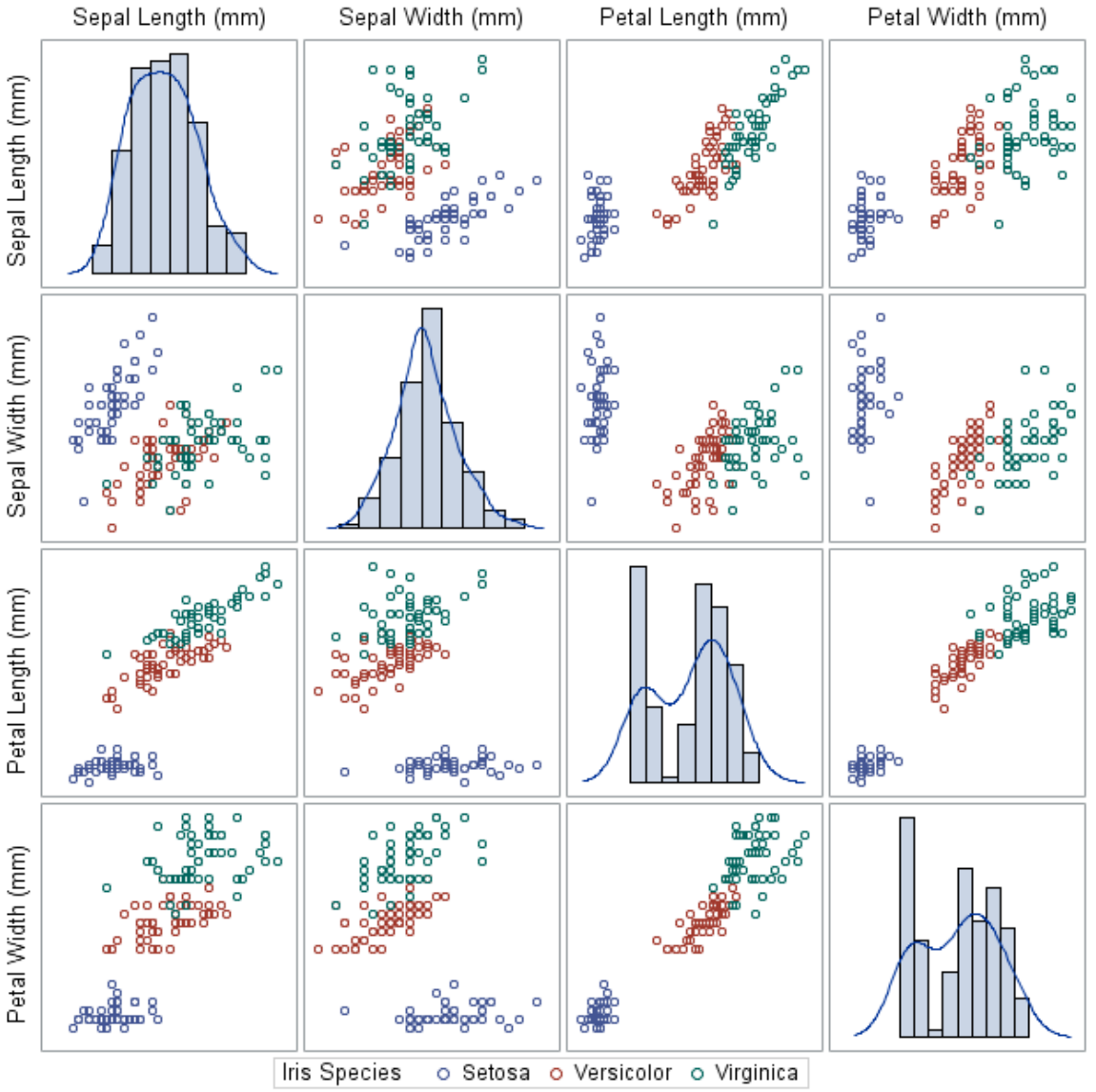




Michelson-Morley Experiment to Determine the Speed of Light

Scatterplots





Suggested Reading

Exploratory Data Analysis

Data Understanding, Data Analysis, Data Science
Data Visualization and Data Exploration

Data and Charts

- Pre-Analysis Uses

The Practice of Data Visualization
Basics of Data Visualization

Data Exploration

Workhorse Data Visualizations

Exercises

Exploratory Data Analysis

1. Find examples of data presentations that you consider to be particularly insightful and/or powerful. Discuss their strengths/weaknesses.
2. Find examples of data presentations that you consider to be particularly misleading and/or useless. Discuss their strengths/weaknesses.
3. How do you think new technologies (e.g. virtual or augmented reality, 3D-printing, wearable computing) will influence data presentations?