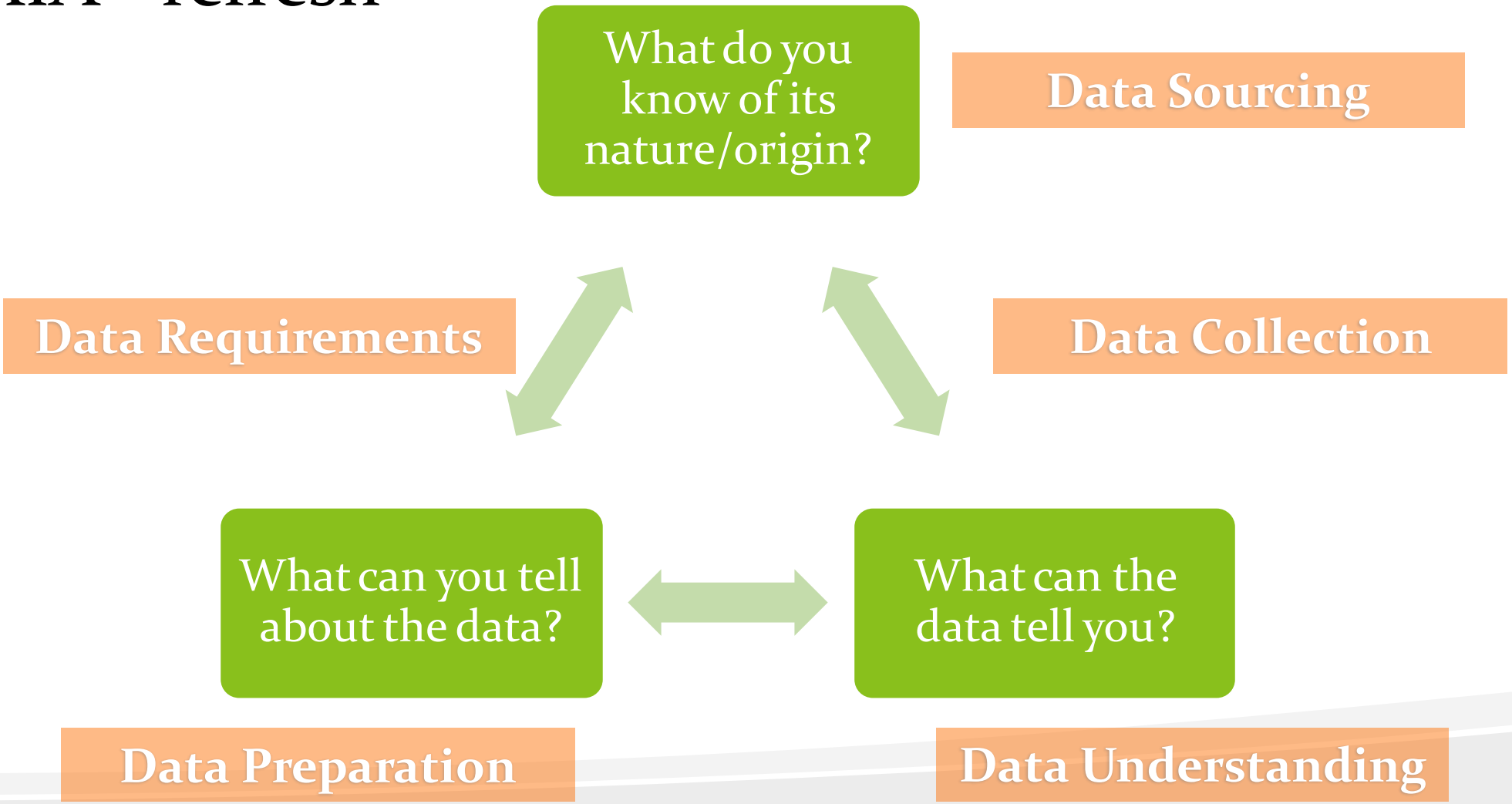
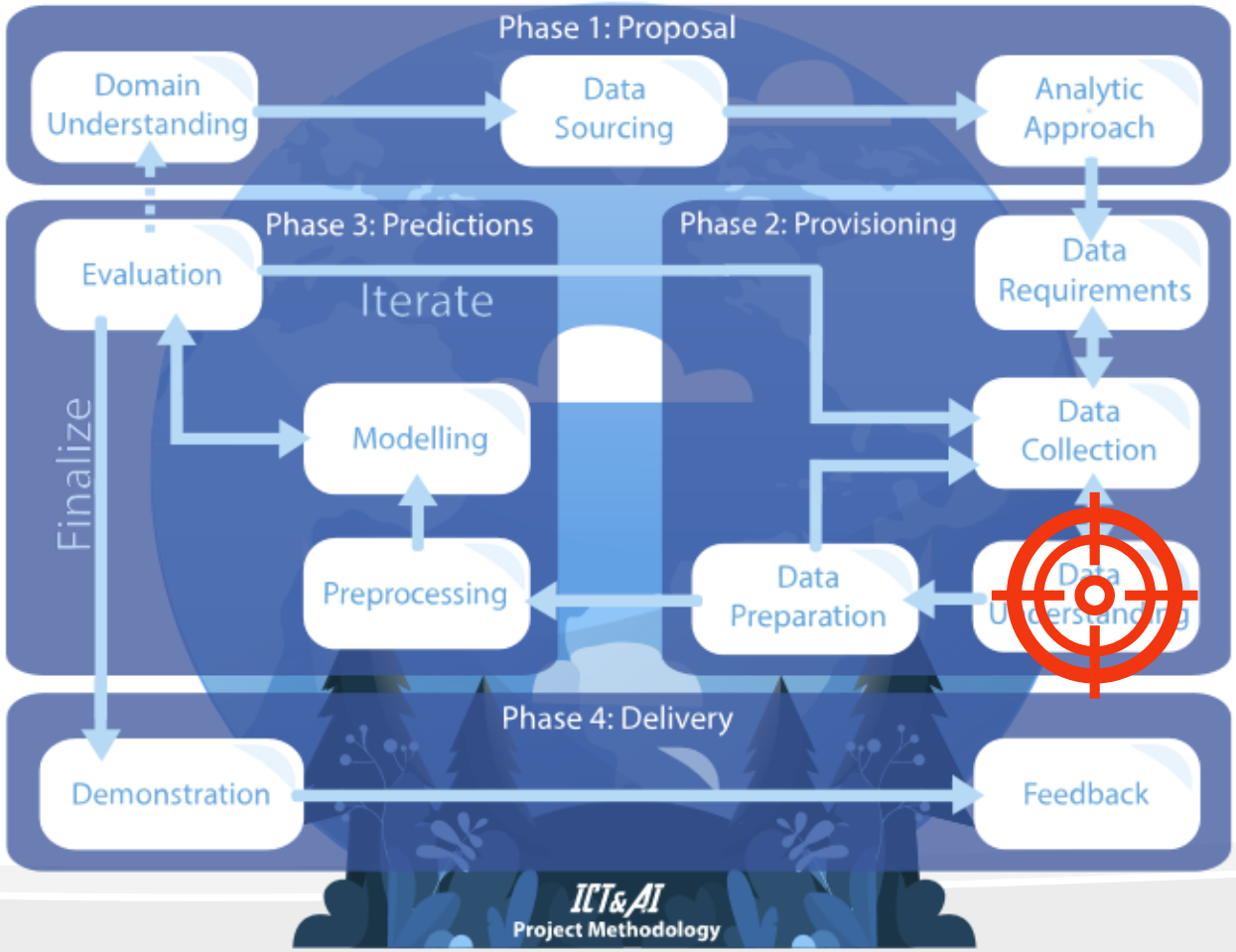
DAIA

Visualizations (basics) – week 2

DAIA - refresh



AI project methodology: your roadmap



Target variable, model and domain requirements, data source(s), ...
What do you want to achieve (predict)?

What data (quality) is required?
Define a data dictionary

How do you get (generate) and combine your data? Capture the process.

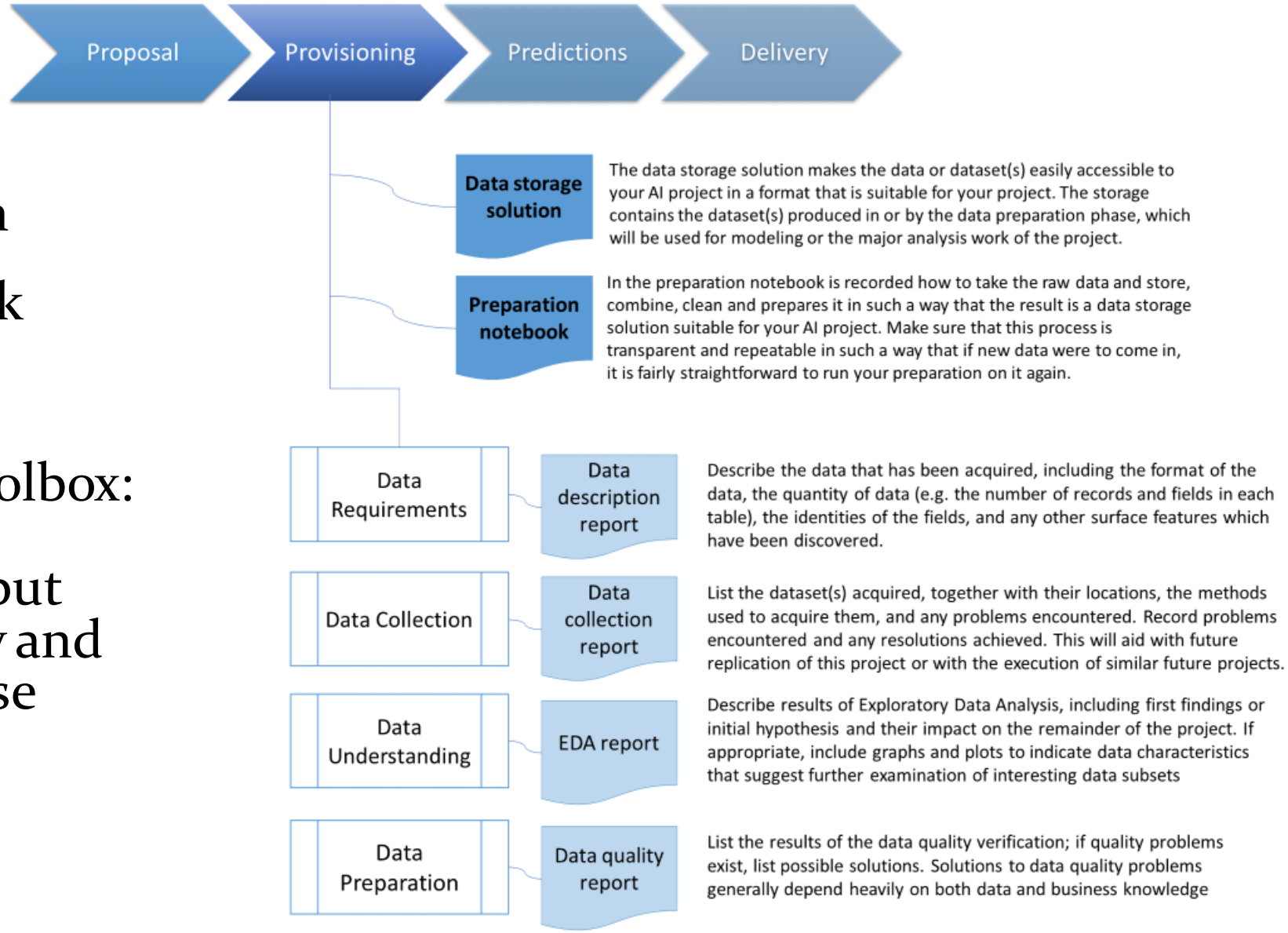
Explore your data (EDA and EDV)

Deliverables (beroepsproducten)

Prepared dataset:

- Data storage solution
- Preparation notebook

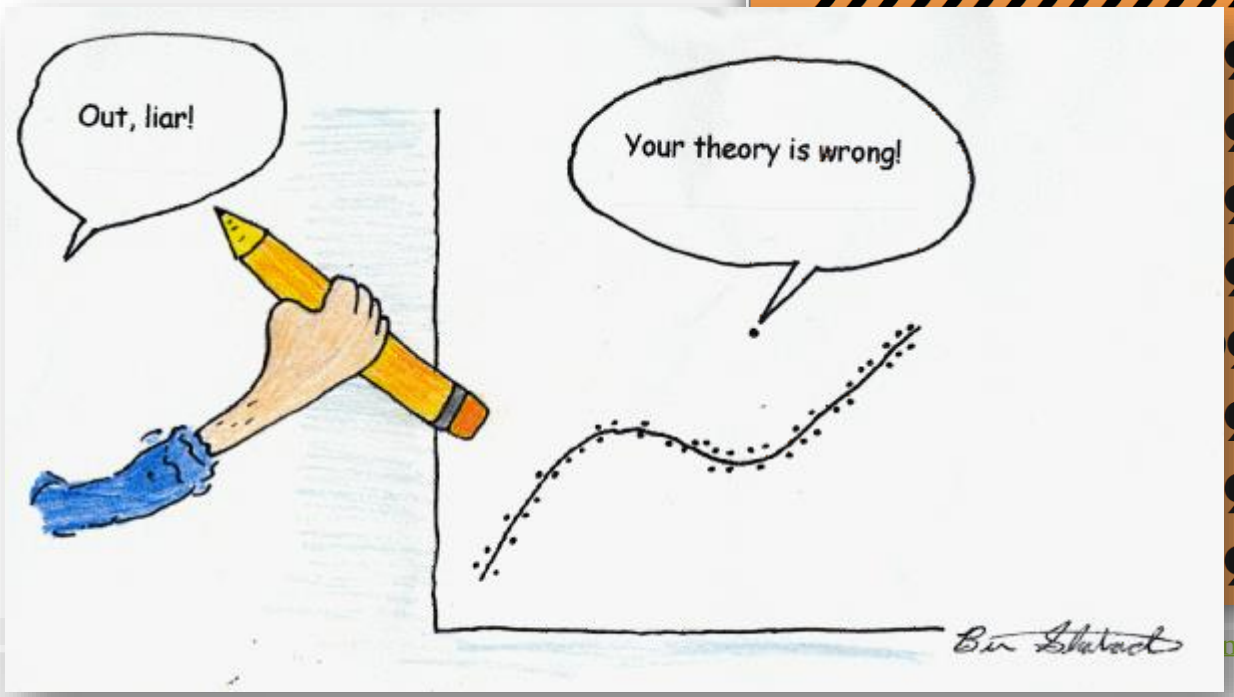
Provisioning is a big toolbox: you don't need to use/ demonstrate all tools, but select tools consciously and show how to apply those properly.

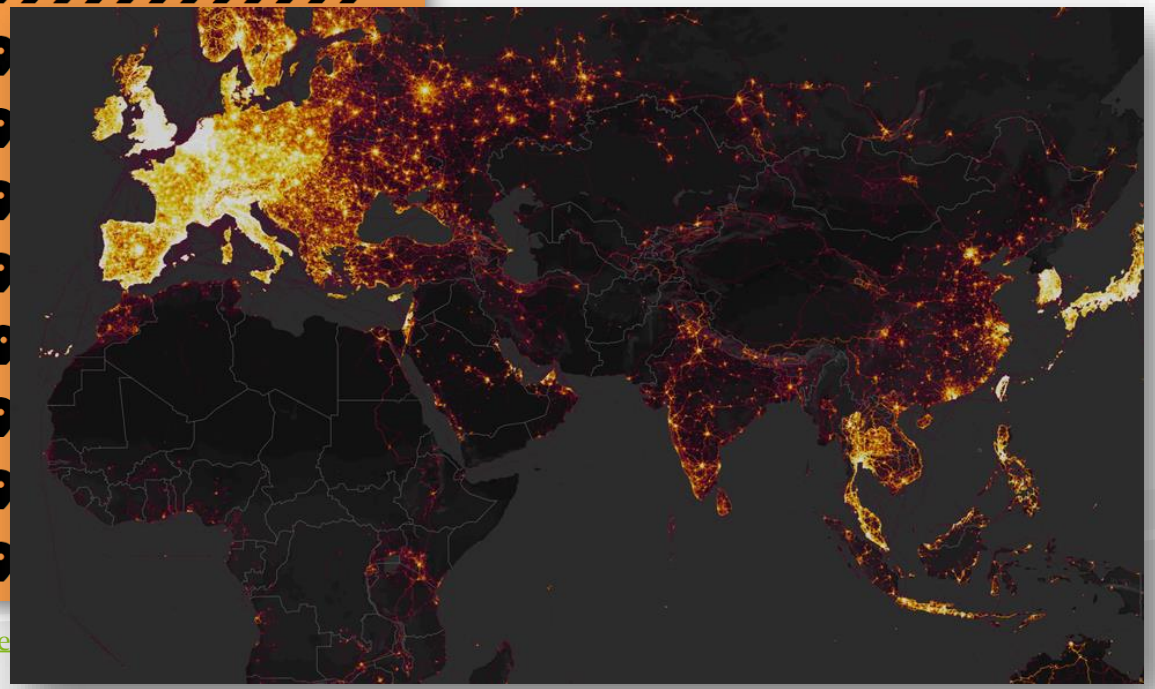


The exploration process

- Start with data
- Decide what your target audience is and what questions are interesting to investigate
- Explore the data to get insights, guided by the questions
- Start with looking at data distribution, relationships and timelines
- Enrich with new datasets, ask new questions
- Use descriptive statistics and advanced charts

Why exploring needs visualizations…

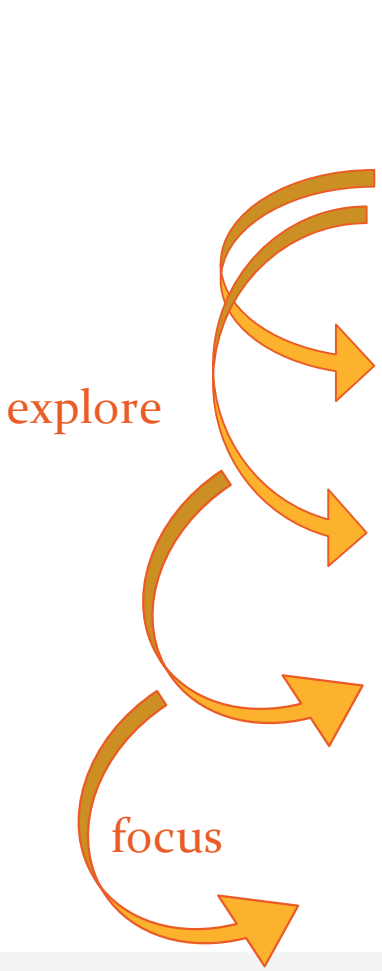




https://psychab.wordpress.com/2011/10/28/27/

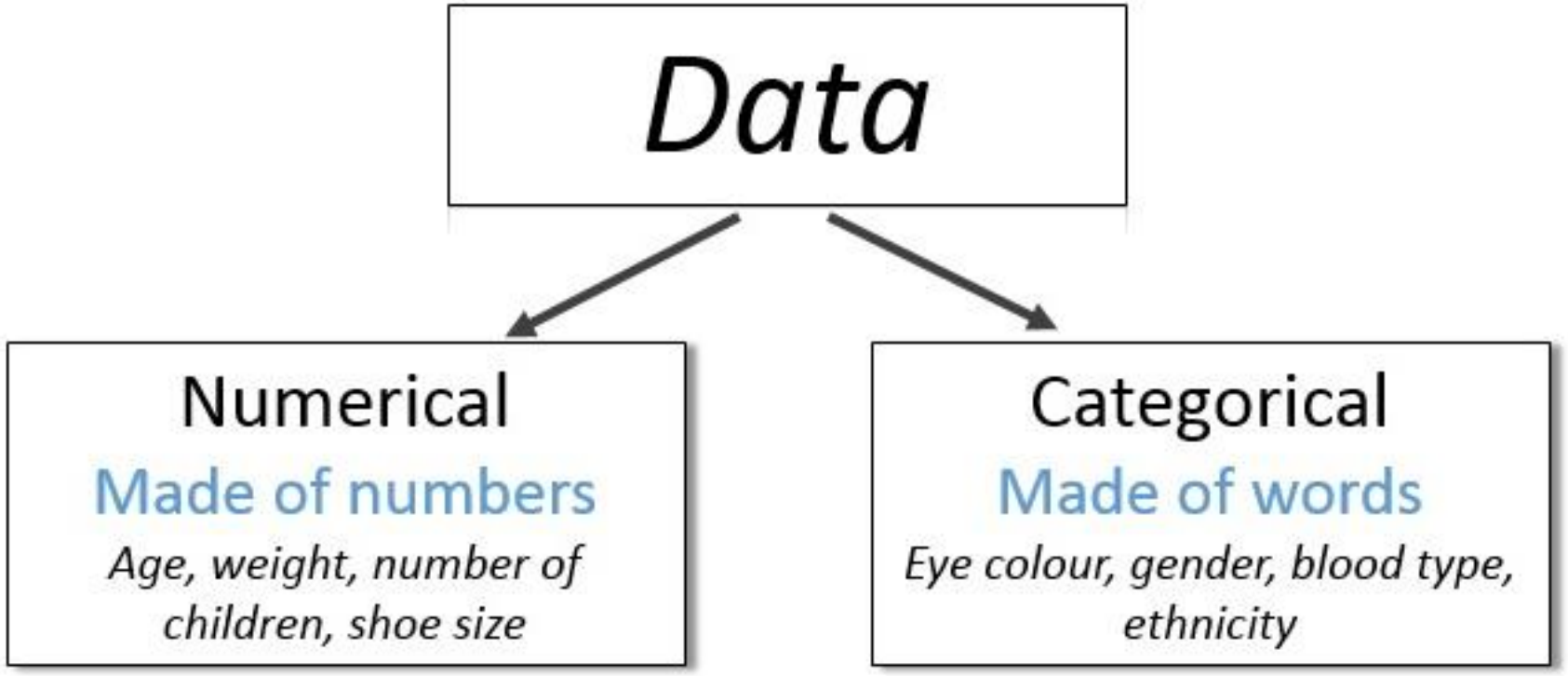
Source: https://www.popularmechanics.com/technology/apps/a15912407/strava-app-military-bases-fitbit-jogging/

The Exploration Process



Data	categorical and numerical *variables*, (not) clean, sufficiently diverse
Descriptive statistics	measures for *center* and *dispersion* of the dataset
Basic charts	Histogram, boxplot, barchart, scatterplot, linechart
Relevant insights	Outliers, Correlations, Clusters, Trends
Reporting	Dashboards, clear visuals, well-explained answers, advice

Types of Data



What about discrete and continuous numbers?

What about ordinal and nominal data?

Basic guidelines for data visualization

Descriptive Statistics

- central tendency
- Spread

Basic Charts

1. Histogram -> distribution, central tendency, spread
2. Boxplot -> distribution, central tendency, spread, comparisons
3. Bar Chart -> outliers, comparisons
4. Scatter Plot -> outliers, correlations
5. Line Chart -> trends

Descriptive statistics: Central Tendency

Measures to describe the **center** of the set

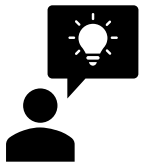
Mean = average

Median = middle value

Only for **numerical** data

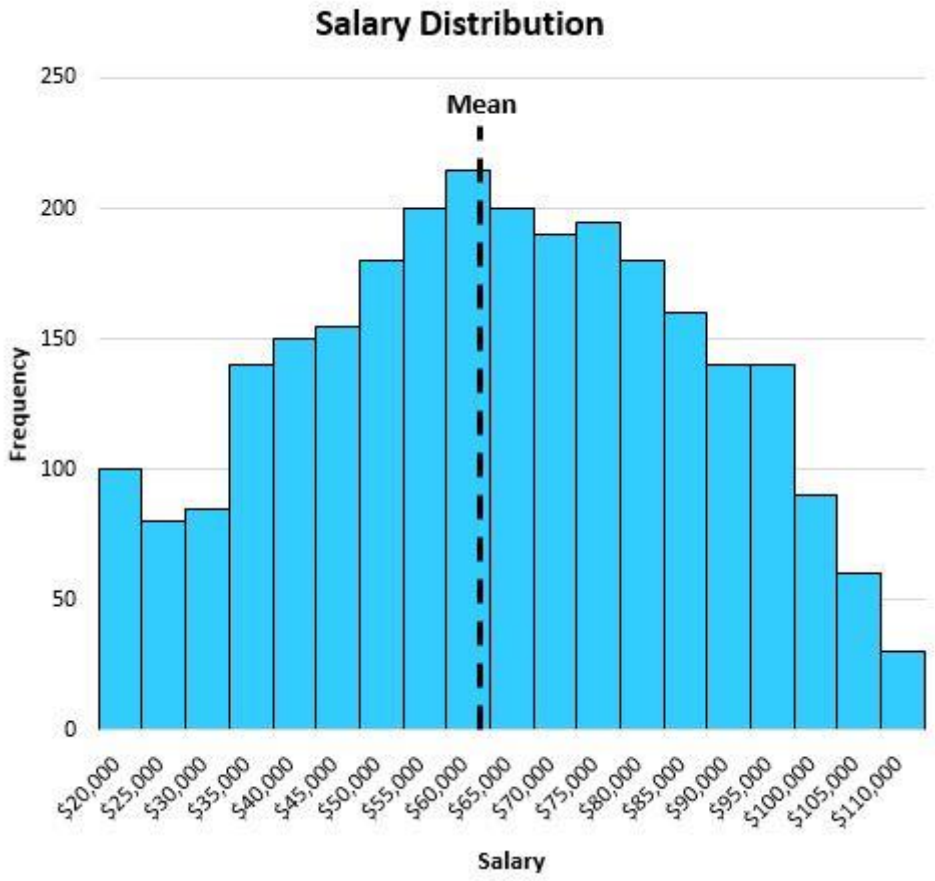
Mode = most frequent value

Typically for **categorical** data

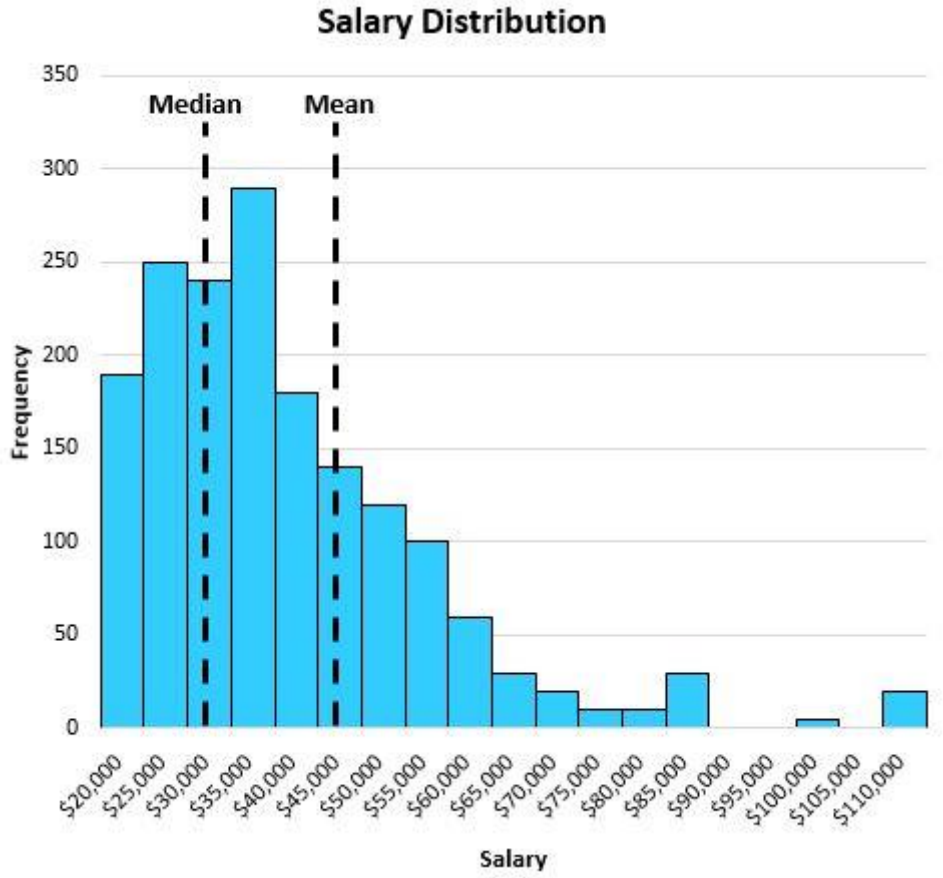


"Elon Musk walks into a bar, and everyone inside becomes a millionaire, on average."

Descriptive statistics: Central Tendency



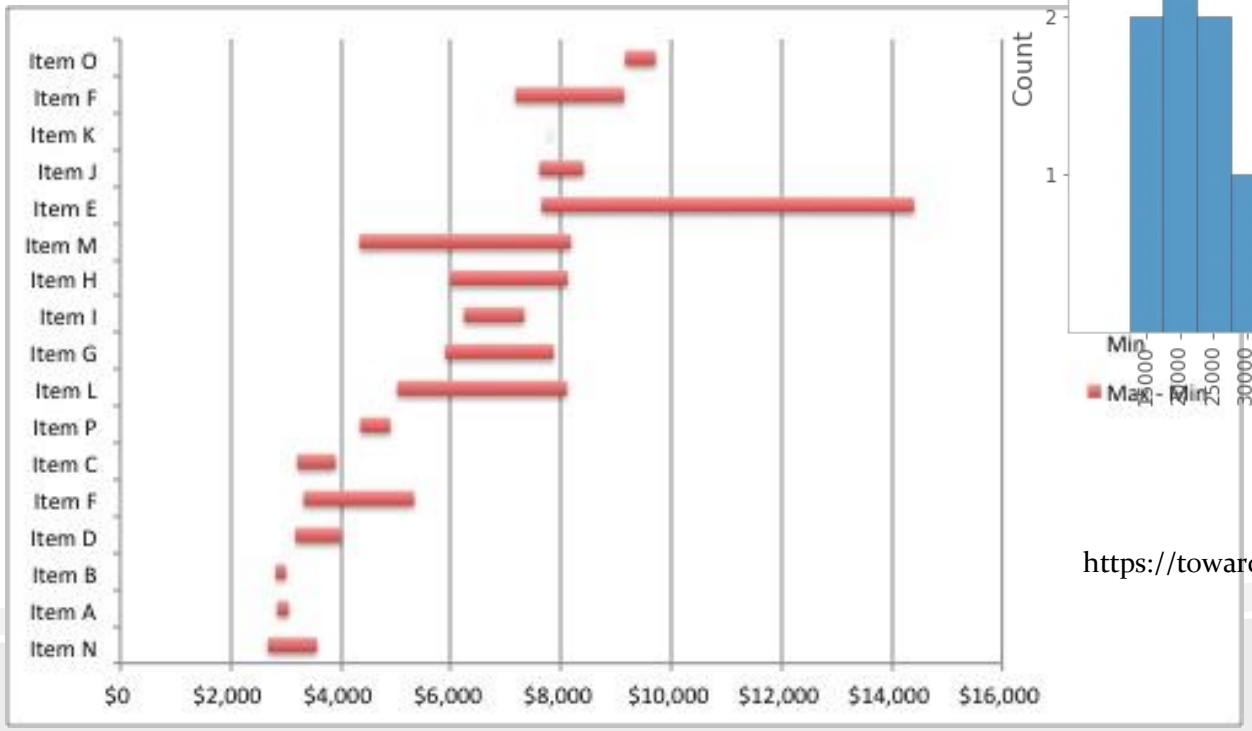
vs

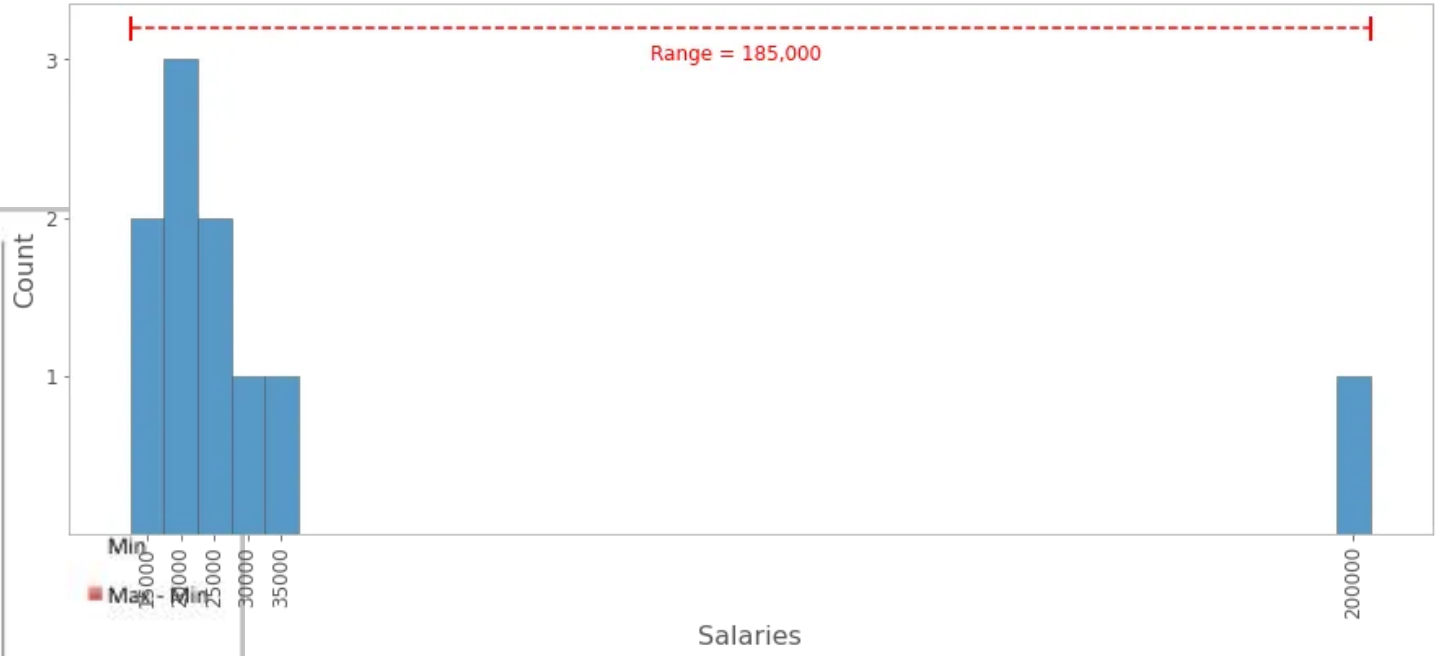


Descriptive statistics: Spread

Measures to describe the **spread/dispersion** of the set

- **Range** = max – min





https://towardsdatascience.com/statistics-02-measuring-and-visualizing-the-spread-of-data-2fc31d928830

https://www.statisticshowto.com/probability-and-statistics/normal-distributions/

Descriptive statistics: Spread

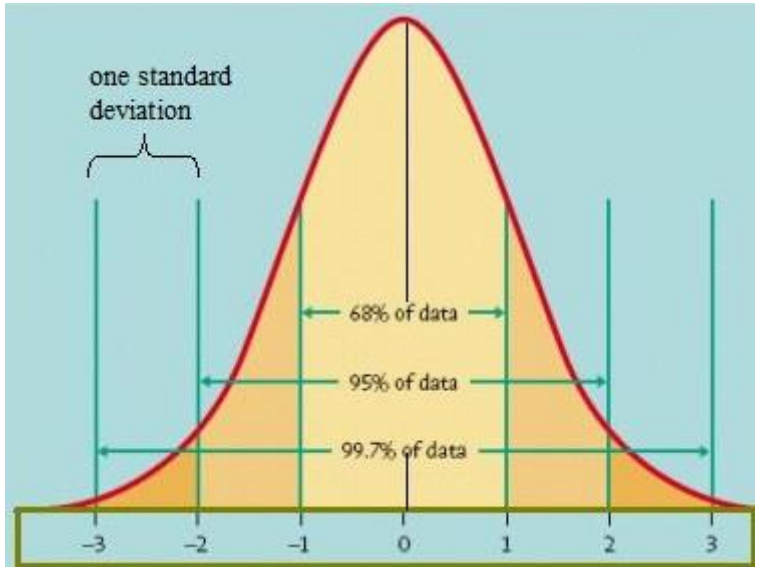
Measures to describe the **spread/dispersion** of the set

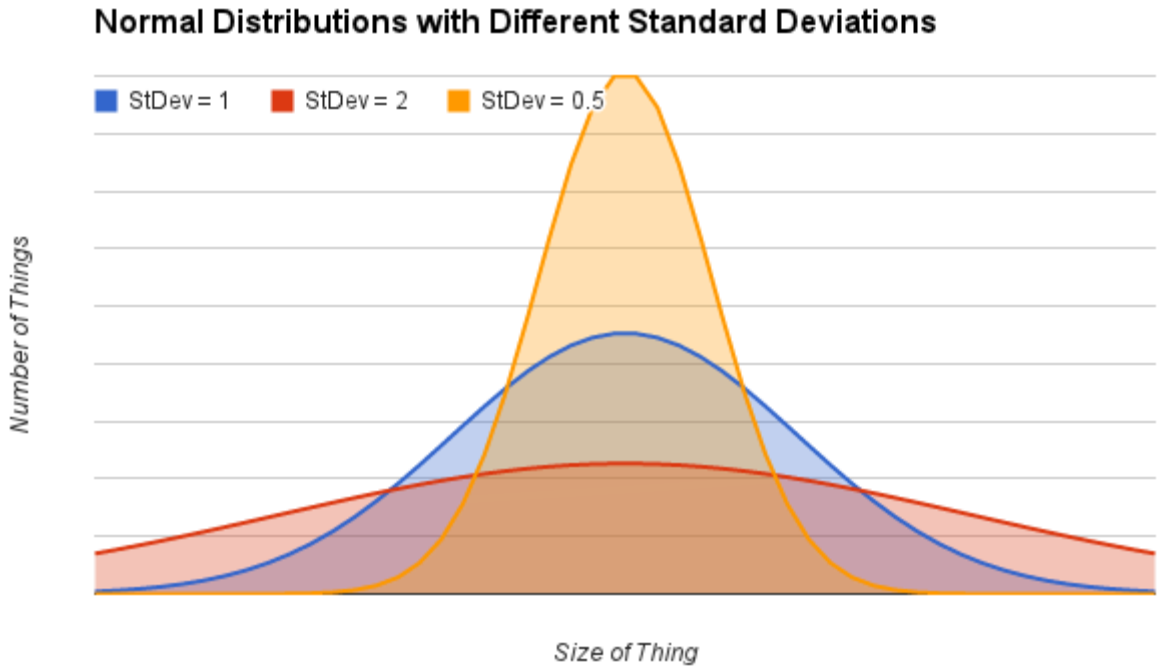
- **Standard deviation** σ
- variance = σ^2

Only for **numerical** data

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

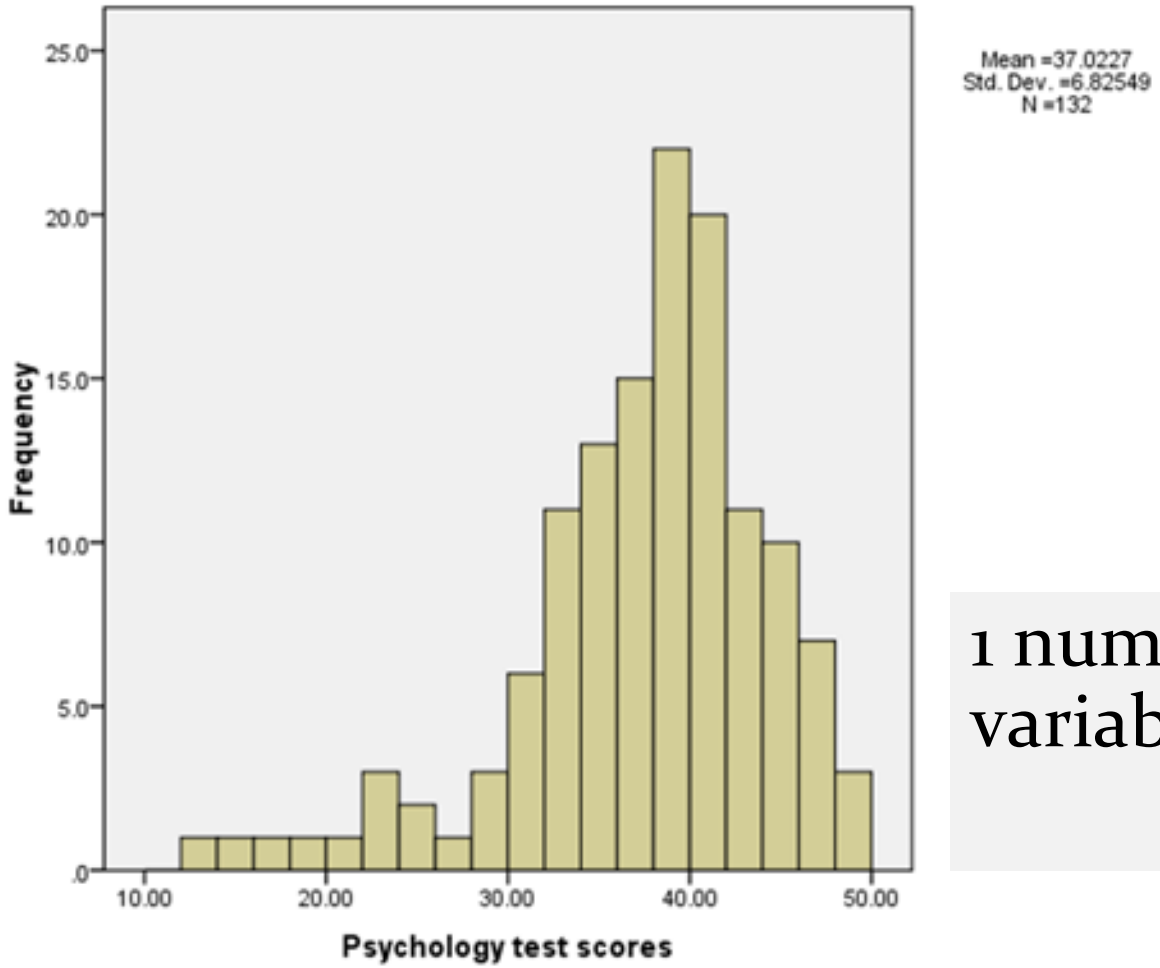
'distance' to the mean value



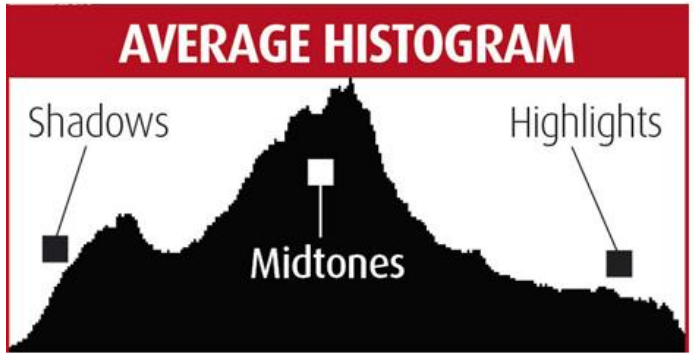


http://sam-koblenski.blogspot.com/2014/09/everyday-statistics-for-programmers_25.html

Basic Charts – 1. Histogram



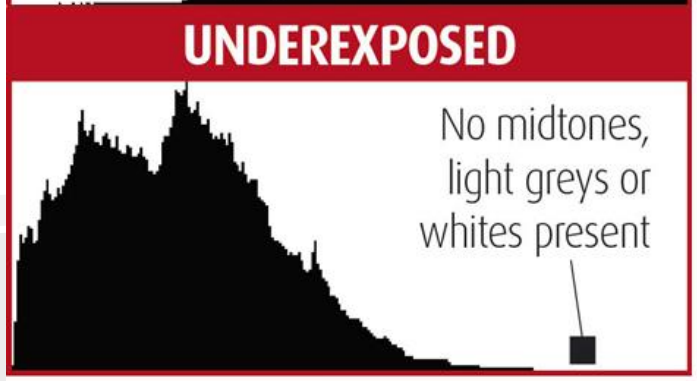
1 nummerical variable



correct exposure

=

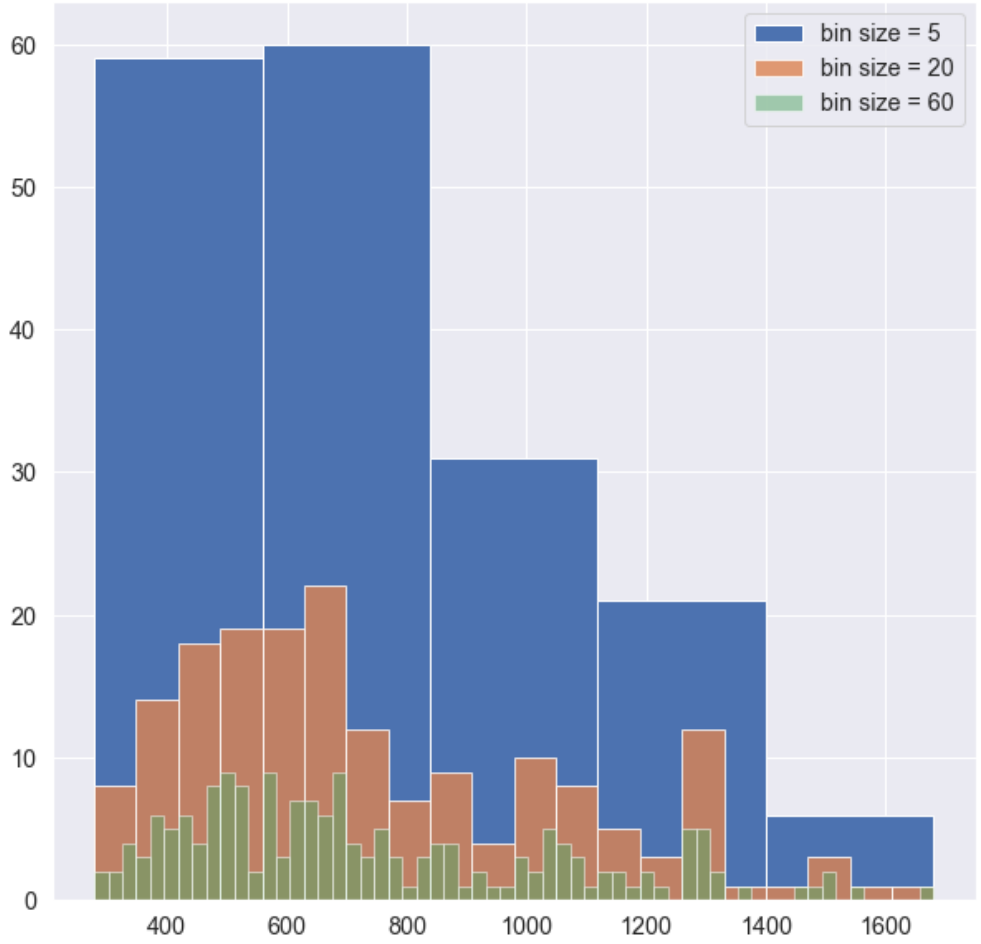
the peak of the black mountain sitting halfway between shadows and highlights



Basic Charts – 1. Histogram

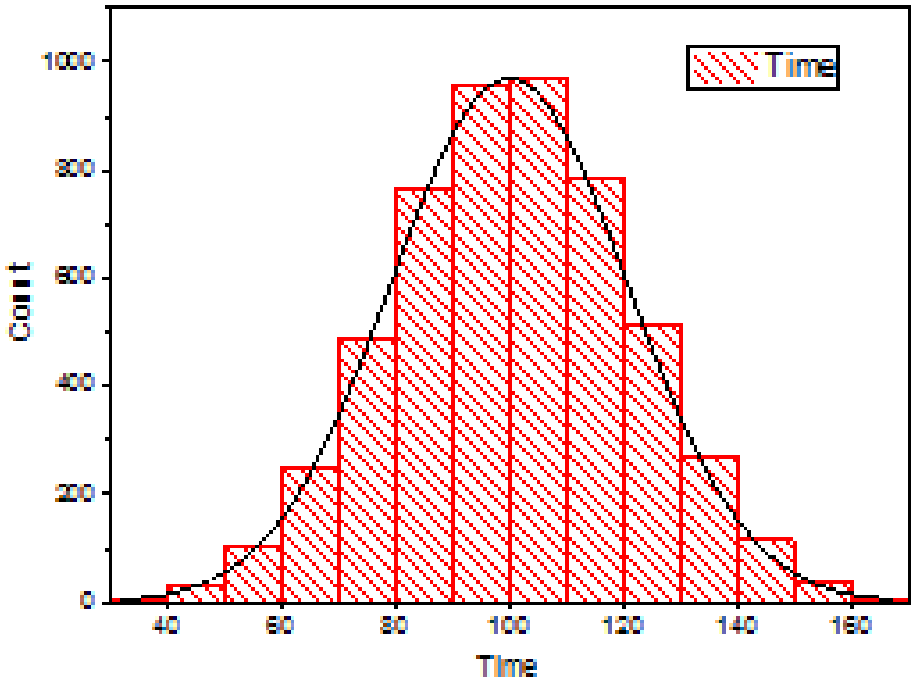
Bin size:

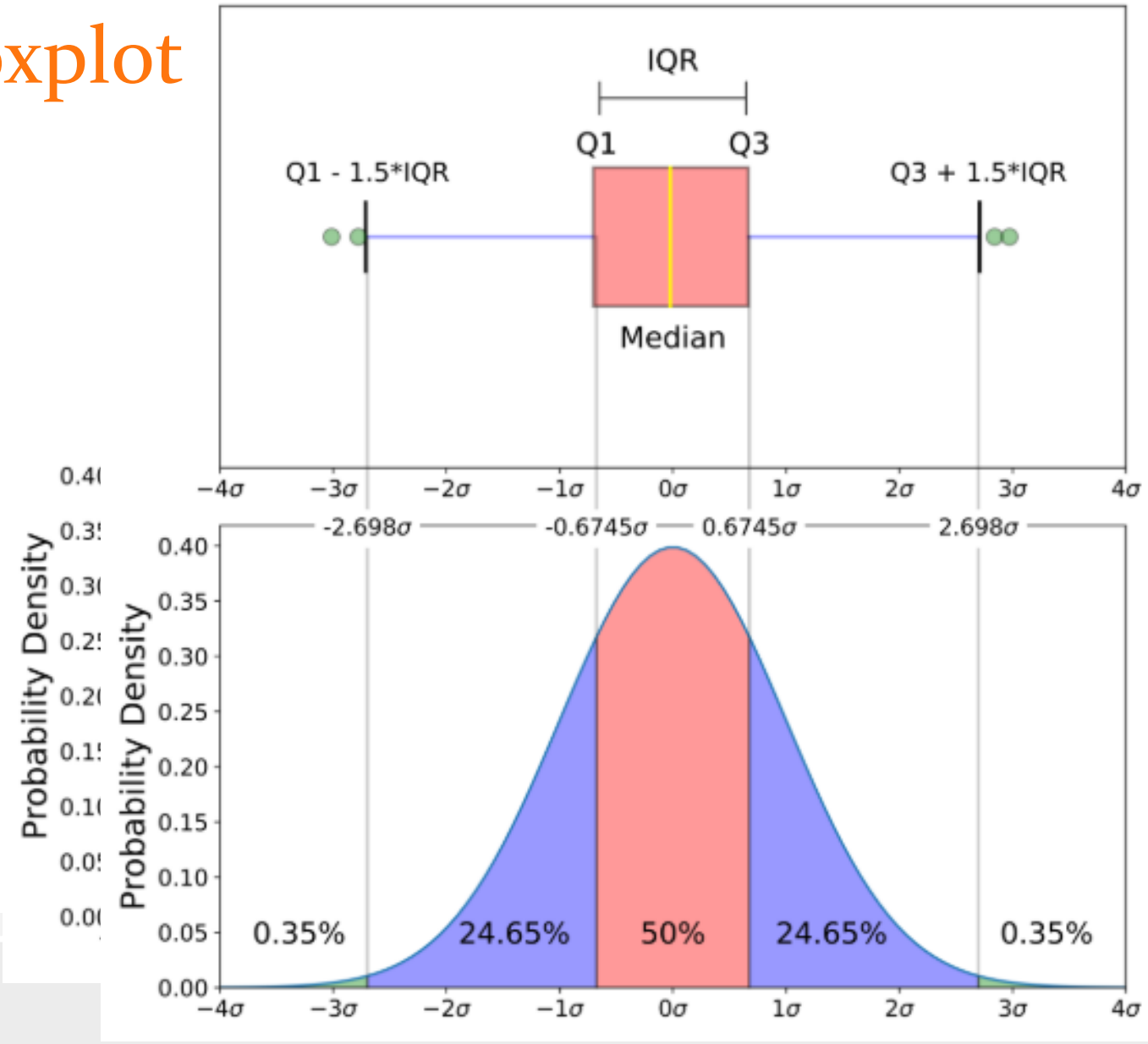
- Too small: no grouping/ aggregation effect
- Too big: no differentation




```
#(bin) size does matter!
plt.hist([winedata['proline']], bins= 5, alpha=1, label='bin size = 5')
plt.hist([winedata['proline']], bins= 20, alpha=0.8, label='bin size = 20')
plt.hist([winedata['proline']], bins= 60, alpha=0.5, label='bin size 60’)
plt.legend(loc='upper right')
plt.show()
```

Basic Charts – 2. Boxplot

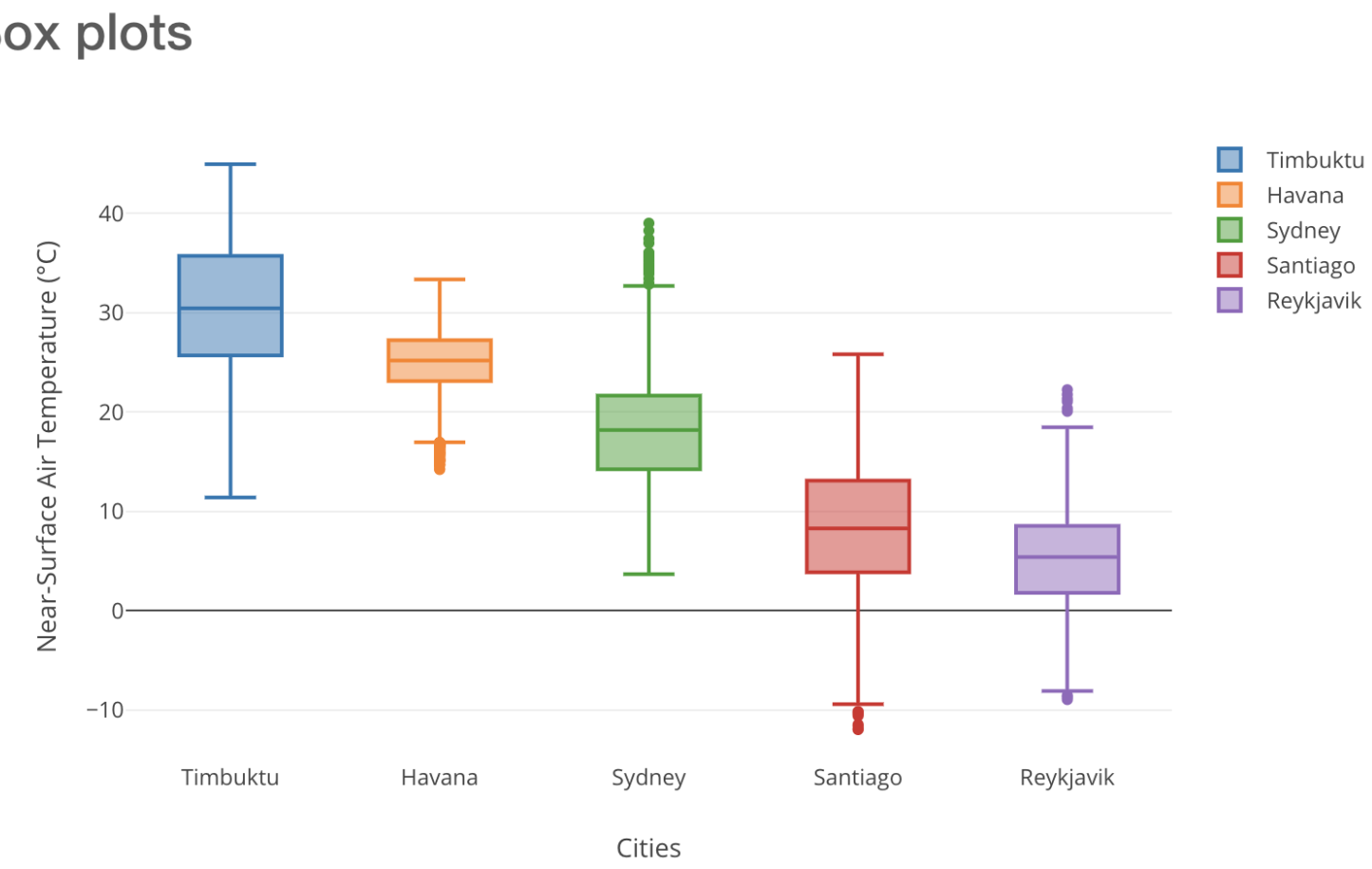




Basic Charts – 2. Boxplot(s)

Boxplots can show a lot of information in 1 graph

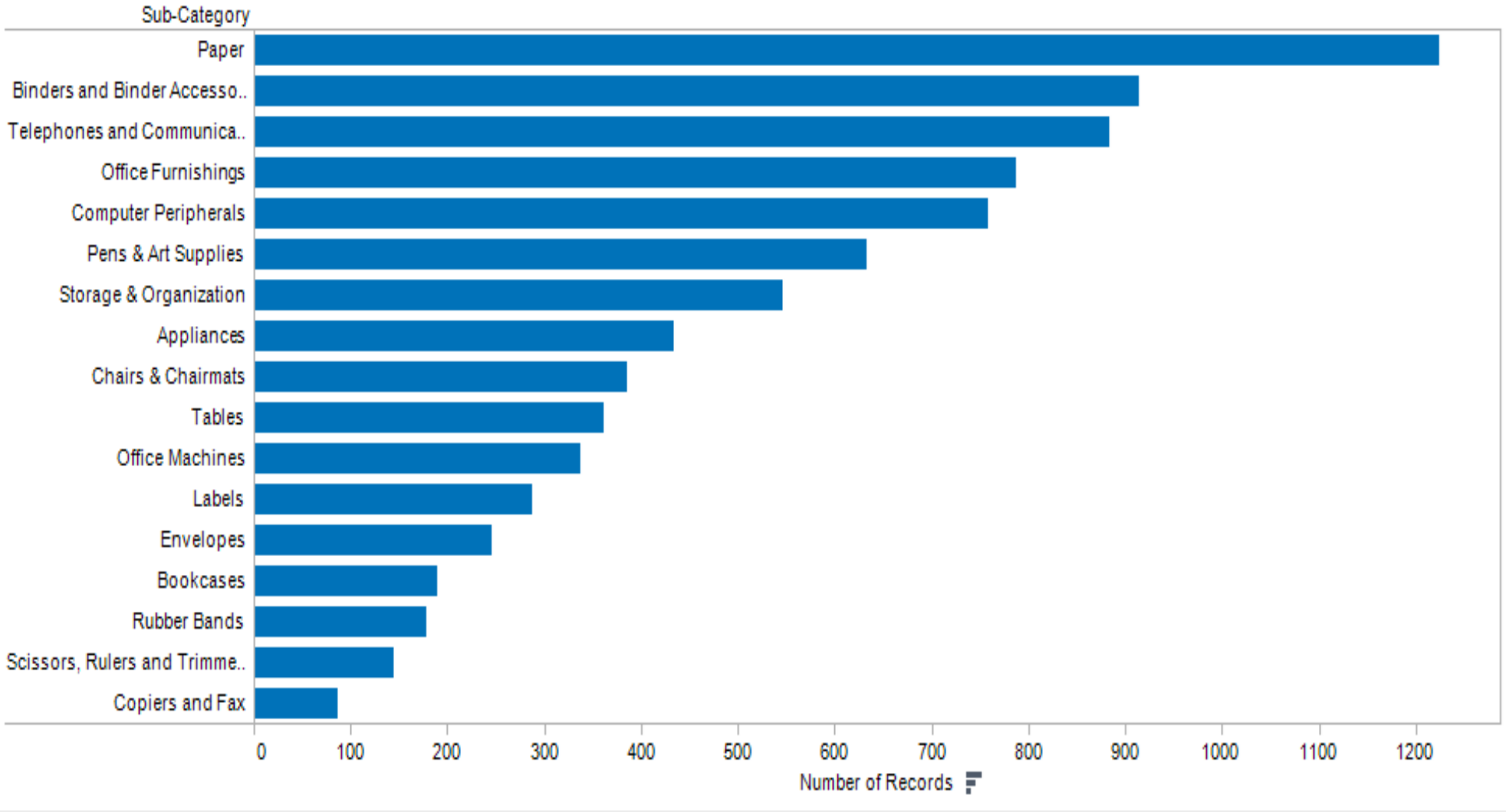
Make sure you can explain what you see (and what strikes you)



https://cds.climate.copernicus.eu/toolbox/doc/gallery/54_box_plots.html

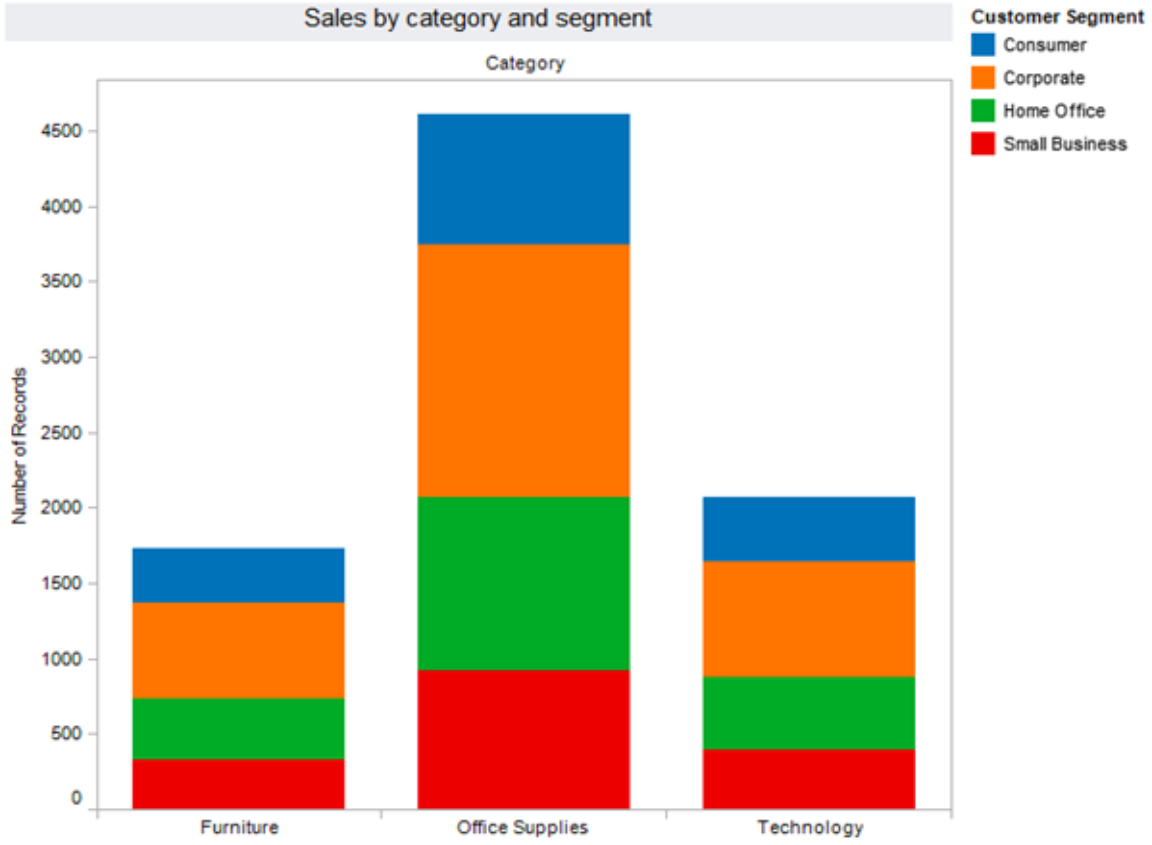
1 category, 1 nummerical

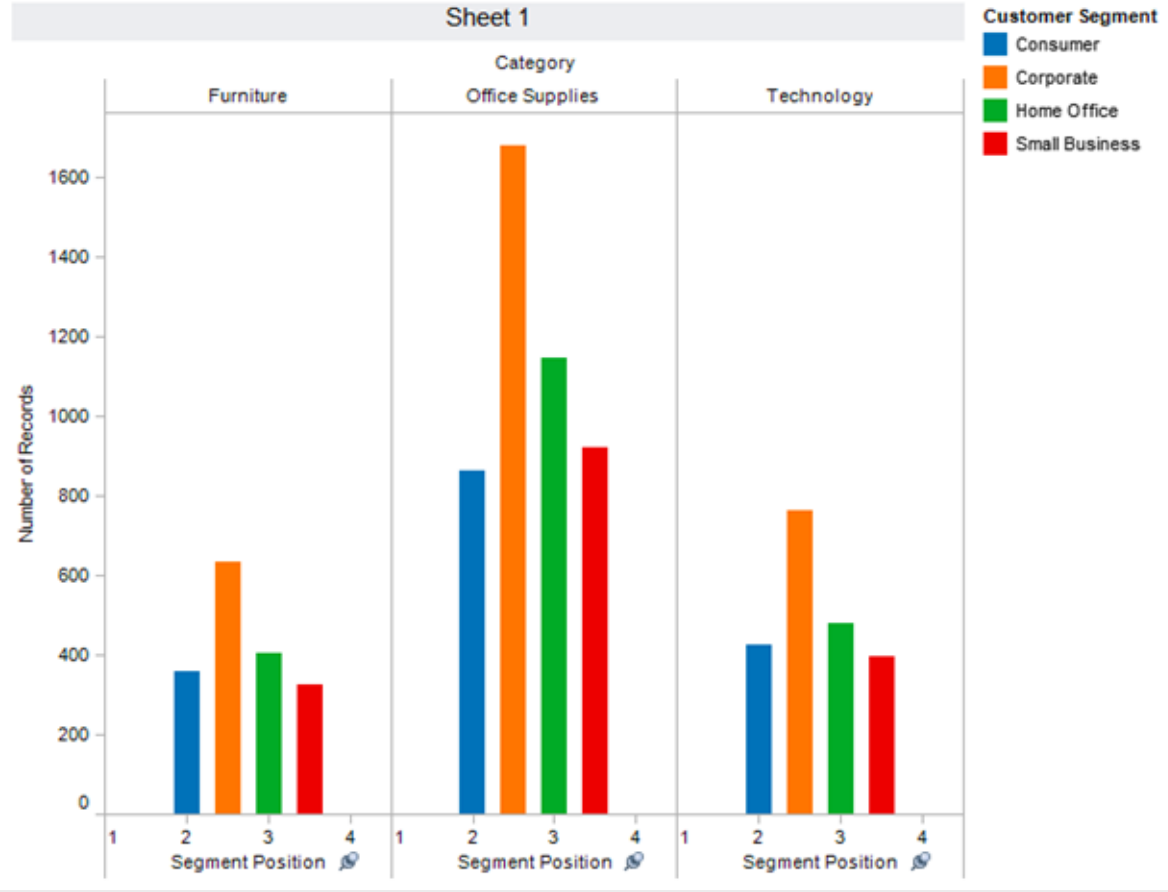
Basic Charts – 3. Bar Chart



… stacked and grouped

2 categorical keys, 1 quant. value



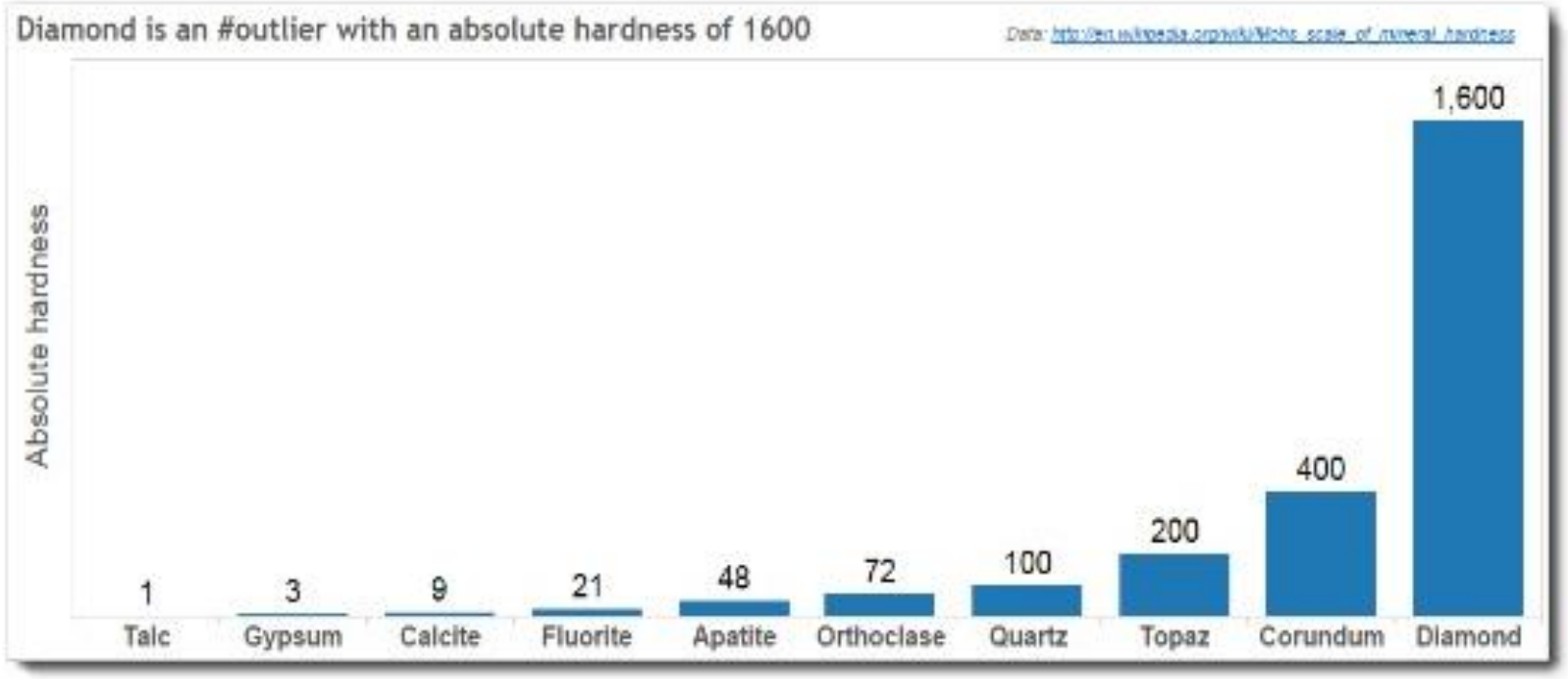


… insights through Outliers and Comparisons

Insights about how **special** something is

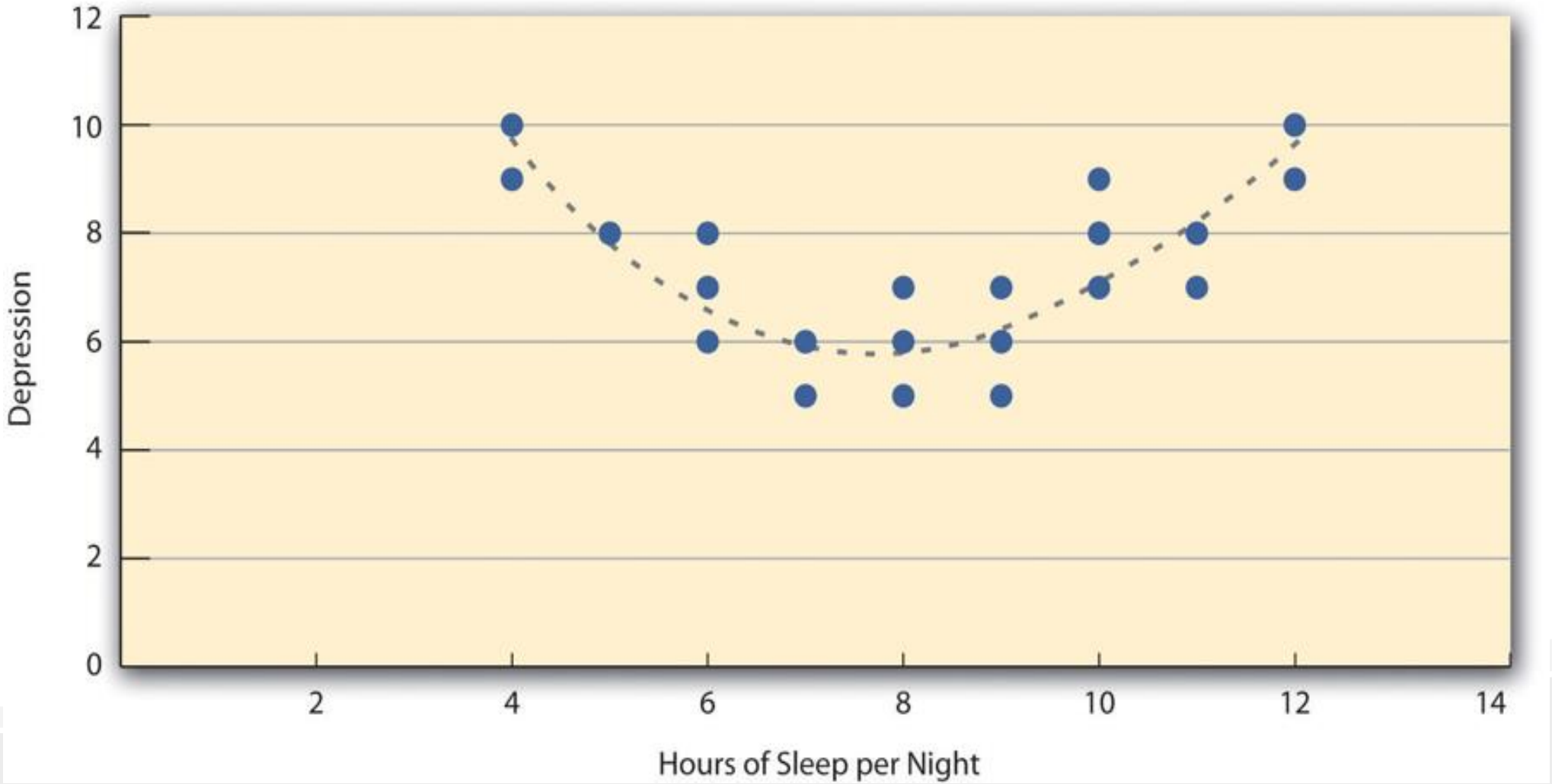
or **larger/smaller** than something else

Which mineral is an #outlier? Why, Diamond: 1,600 on Moh's Scale of Hardness.



2 numerical variables

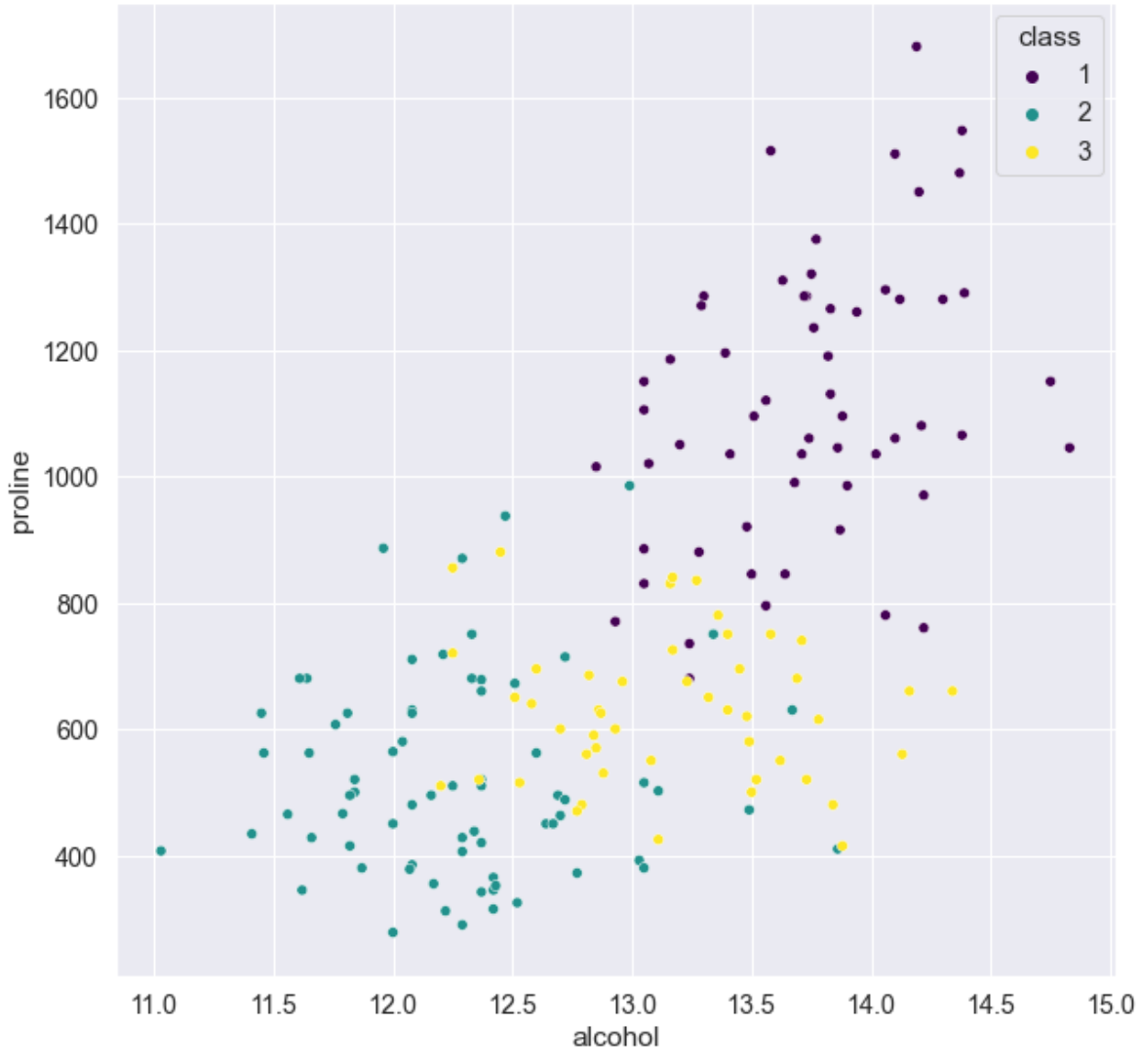
Basic Charts – 4. Scatter plot



Basic Charts – 4. Scatter plot

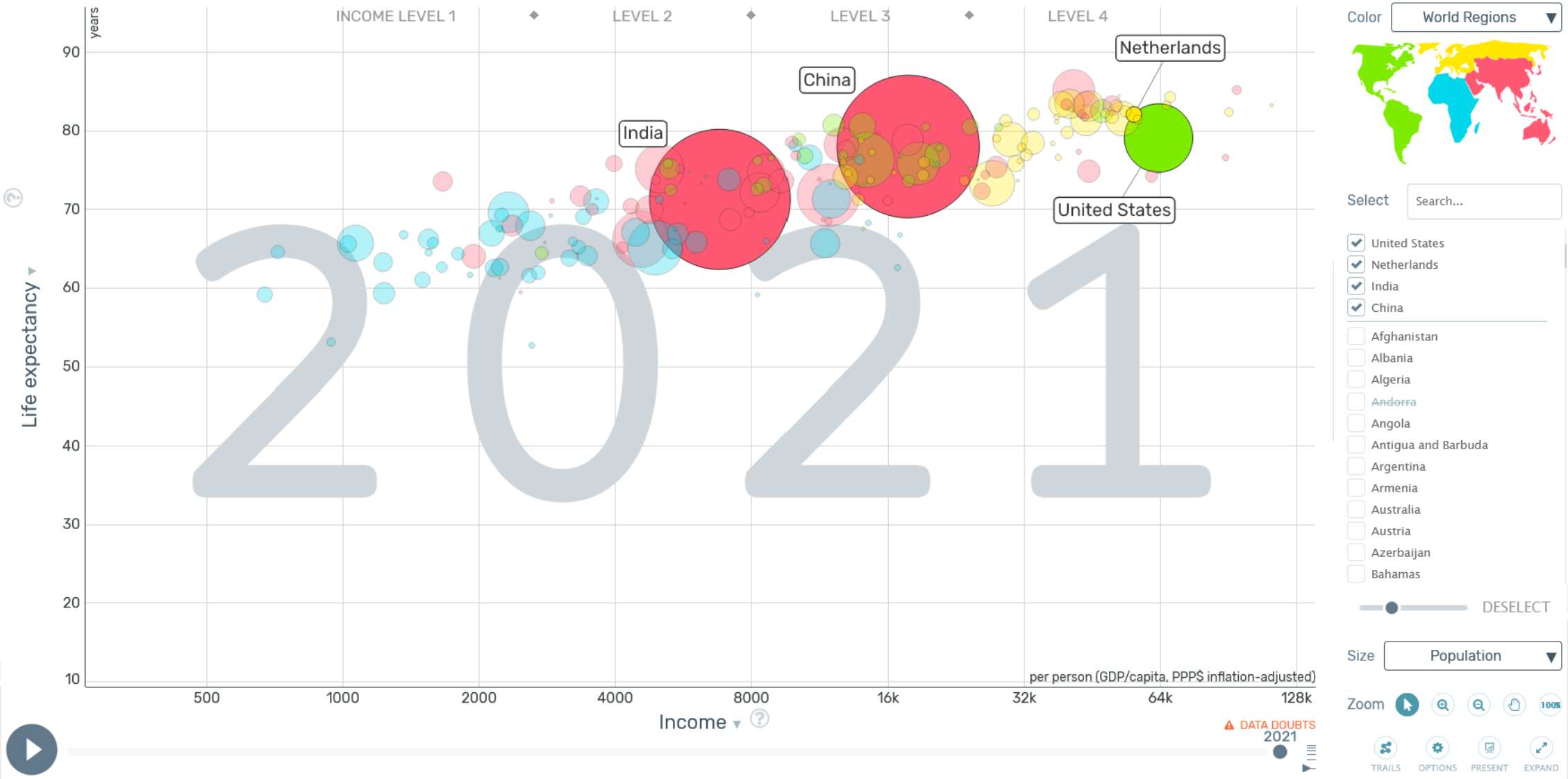
```
#plot 2 features as data points (a 3rd feature as hue)

sns.scatterplot(data=winedata, x='alcohol',
y='proline', hue='class', palette='viridis')
```

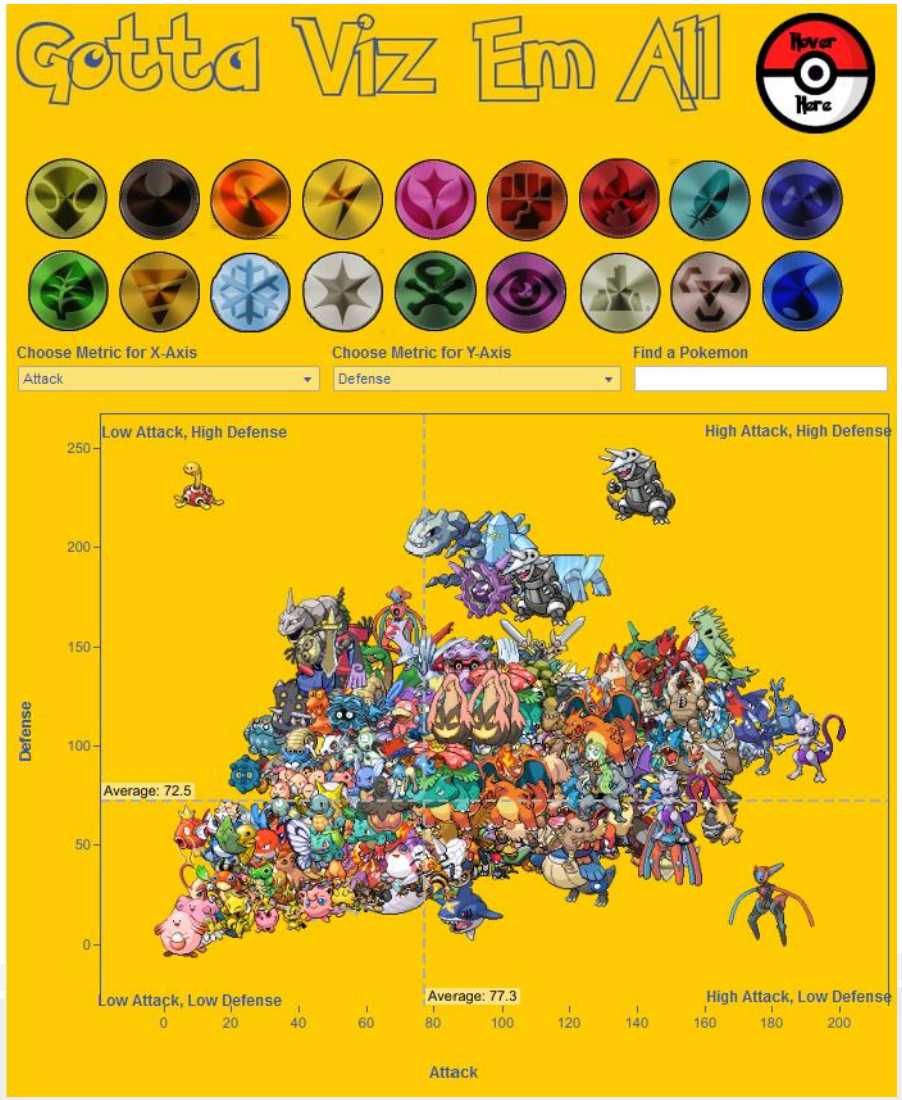


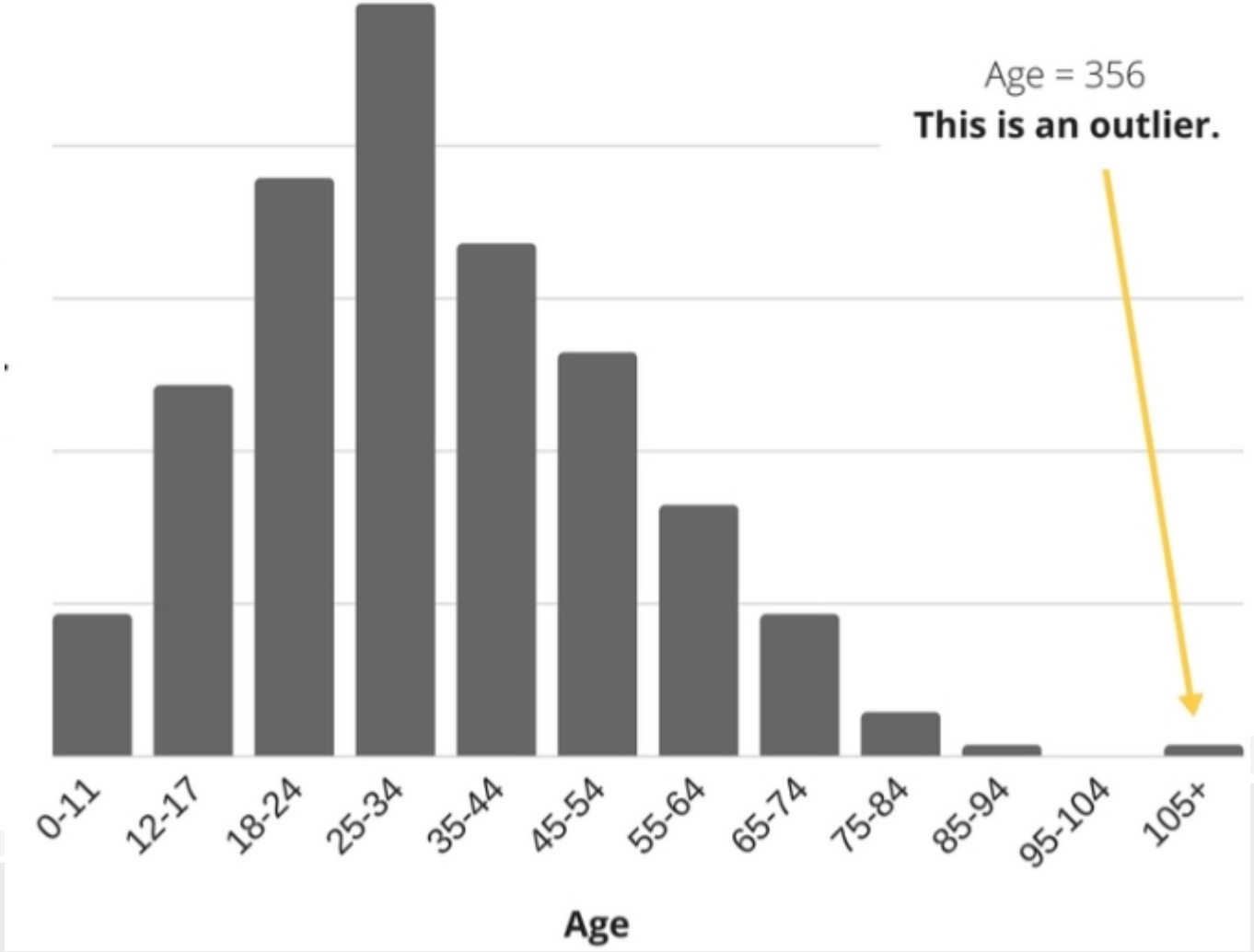
2 numerical variables, and more

Basic Charts – 4. Scatter plot



… insights through Outliers

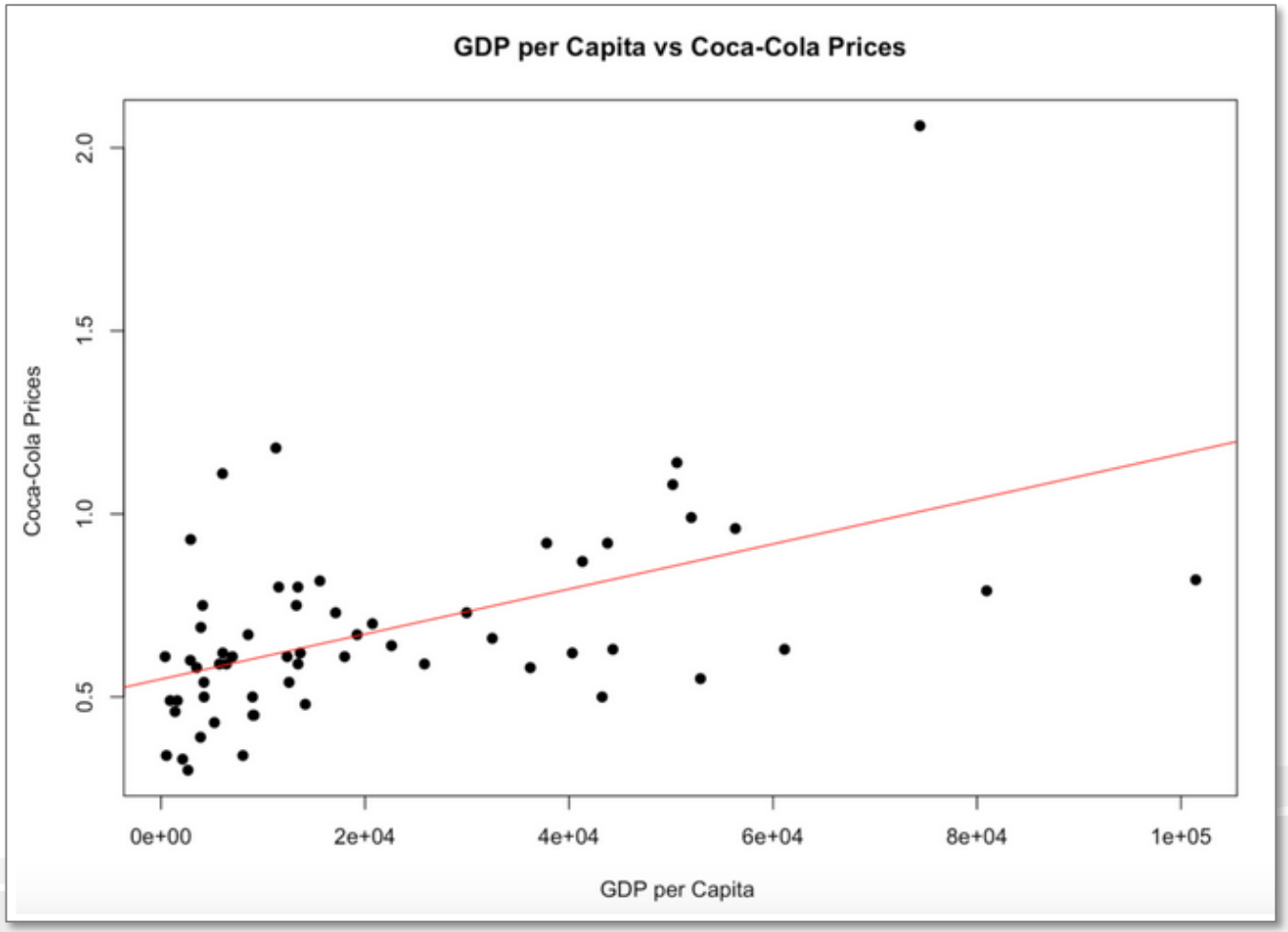




Source: https://medium.com/analytics-vidhya/its-all-about-outliers-cbe172aa1309

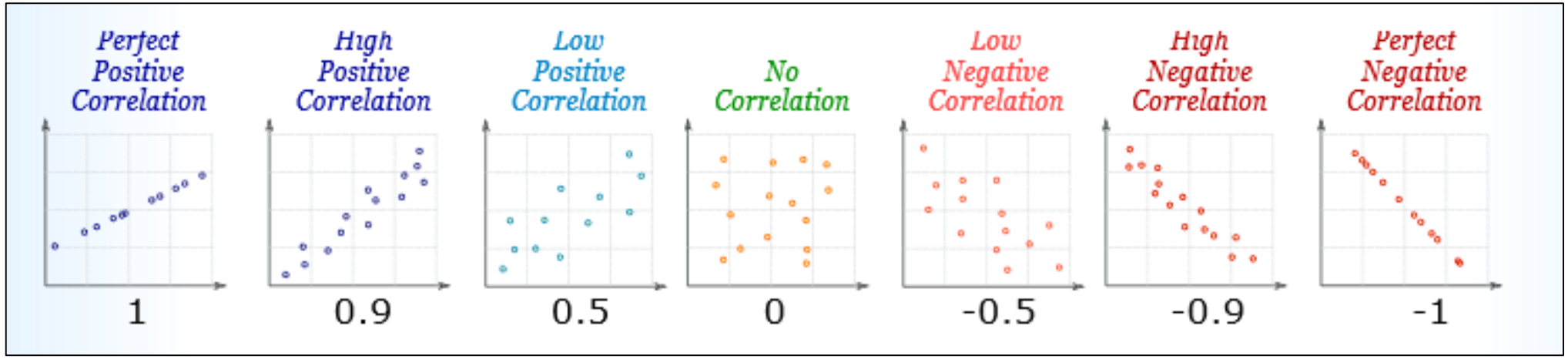
… insights through Correlations

Insights about how strongly numerical features are **linked** and therefore possibly influence each other.



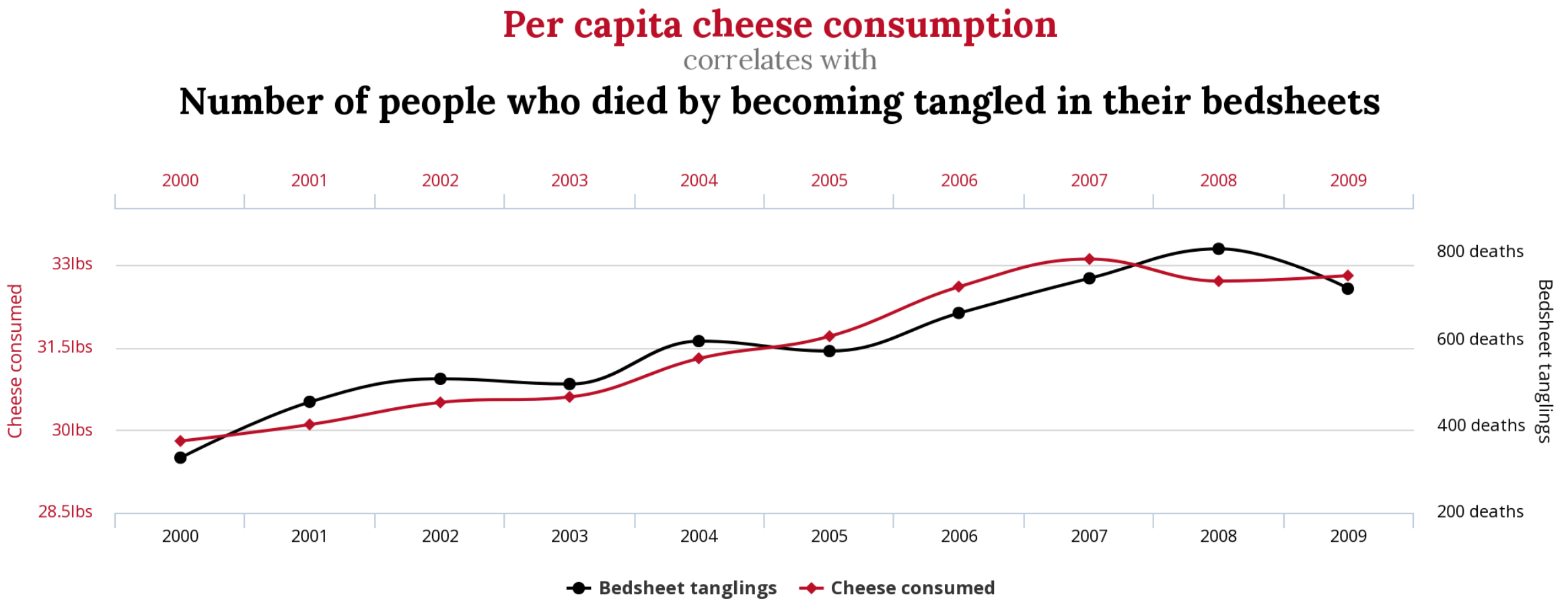
… insights through Correlations

also as a numerical value, most commonly used is the **correlation coefficient R**



Try this fun game: http://guessthecorrelation.com/

… insights through Correlations

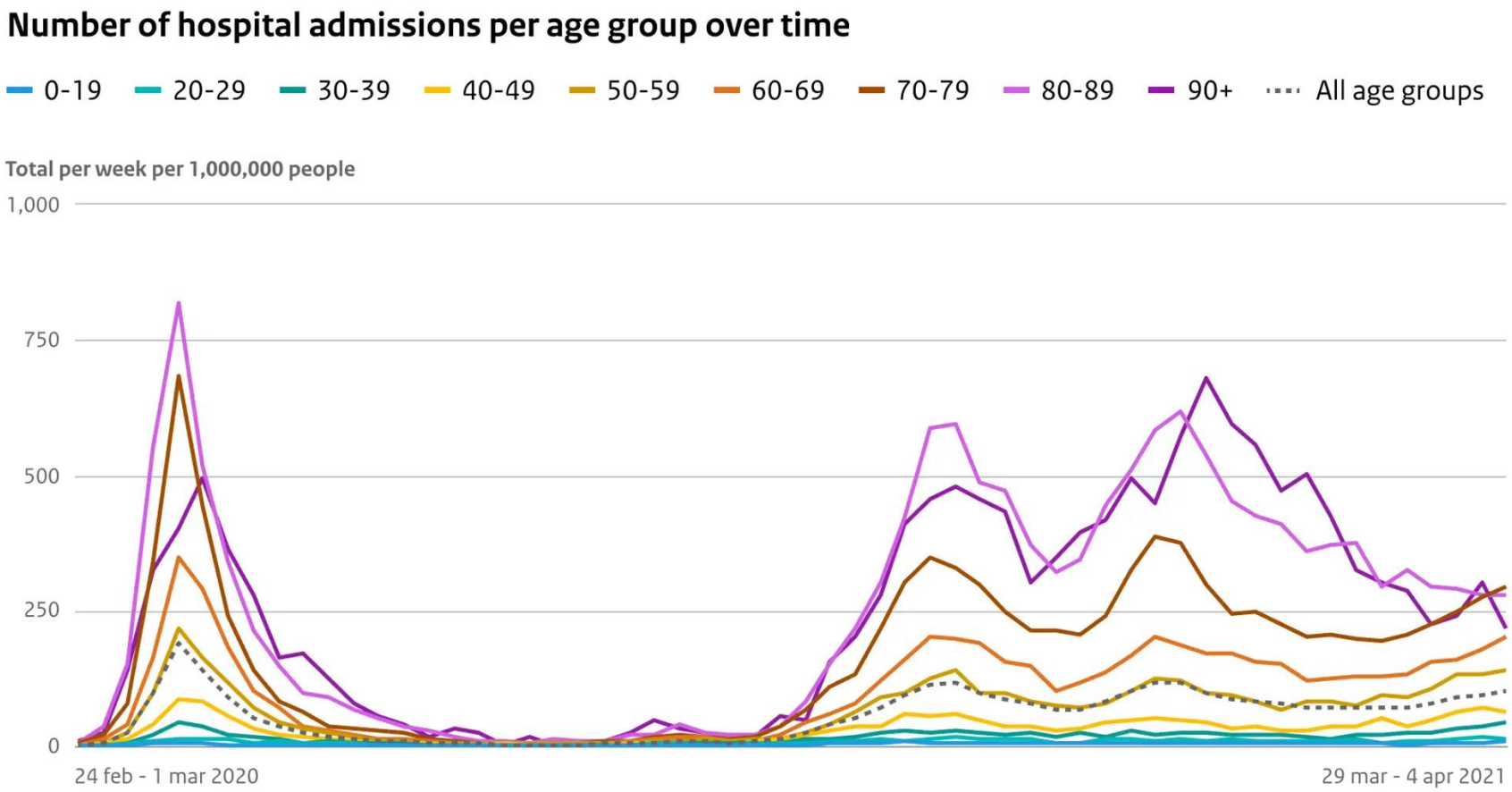


Correlation does not always mean a cause-effect relation!

See many more interesting examples here: http://www.tylervigen.com/spurious-correlations

2 numerical variables, usually x = time

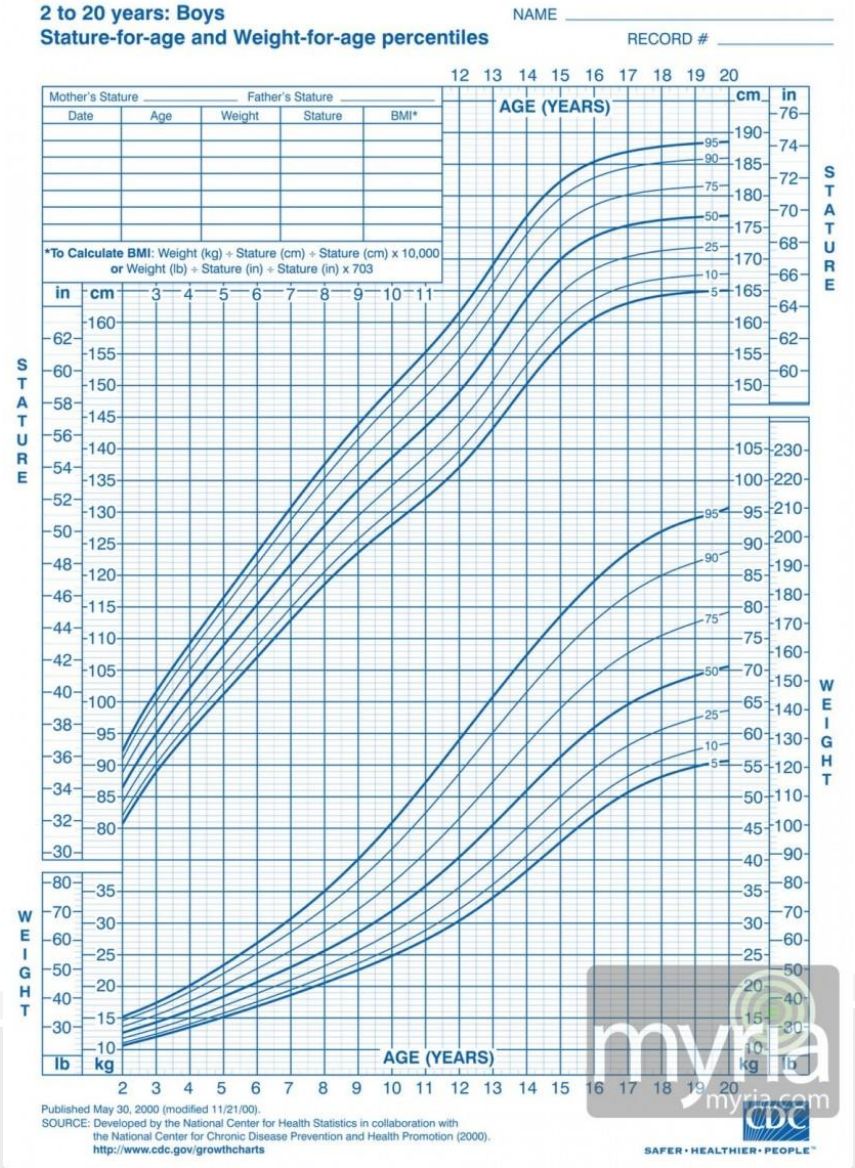
Basic Charts – 5. Line chart

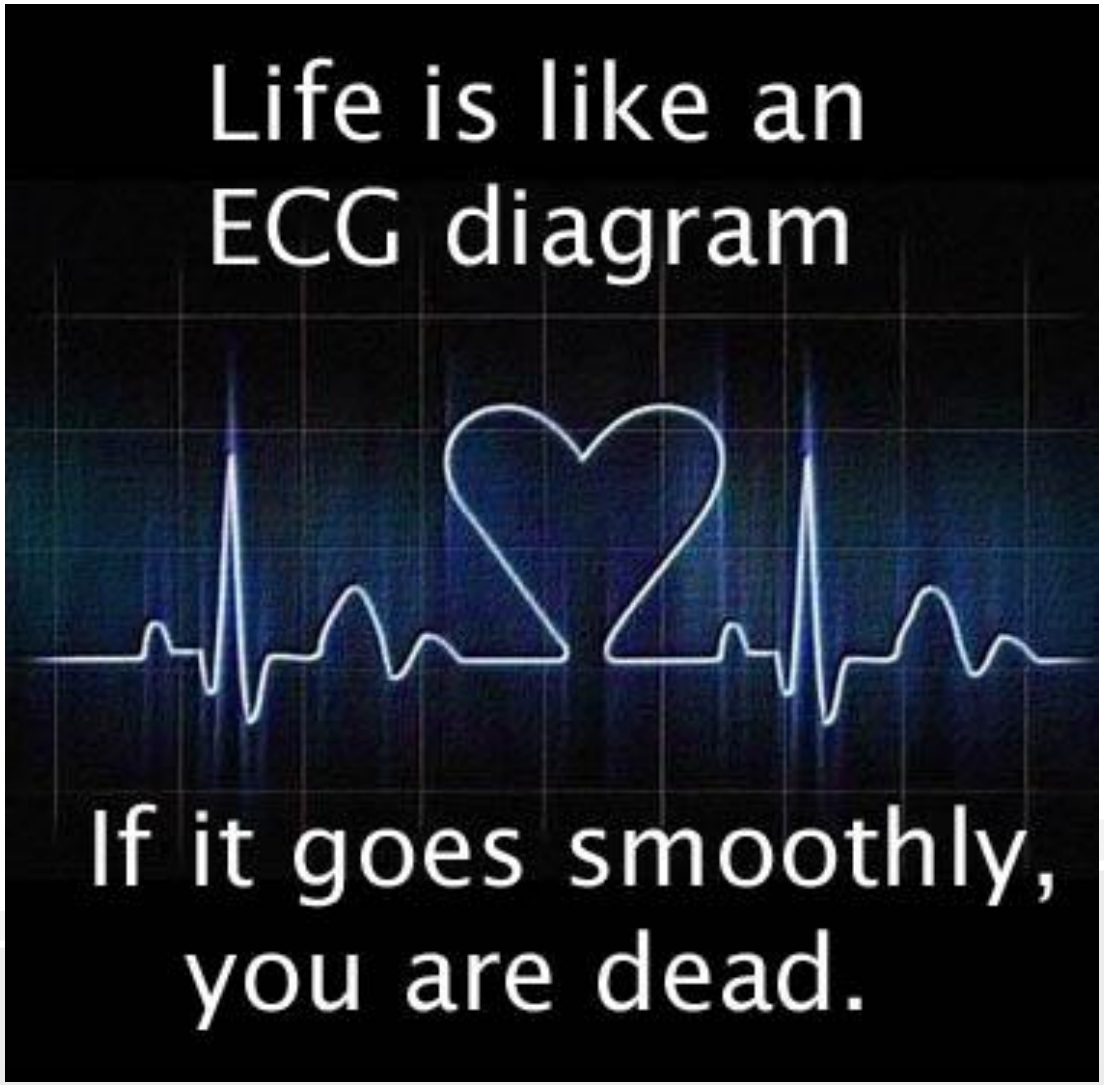


Source: https://coronadashboard.government.nl/artikelen/leeftijdsverdeling-van-covid-patienten-in-de-ziekenhuizen

Basic Charts – 5. Line chart

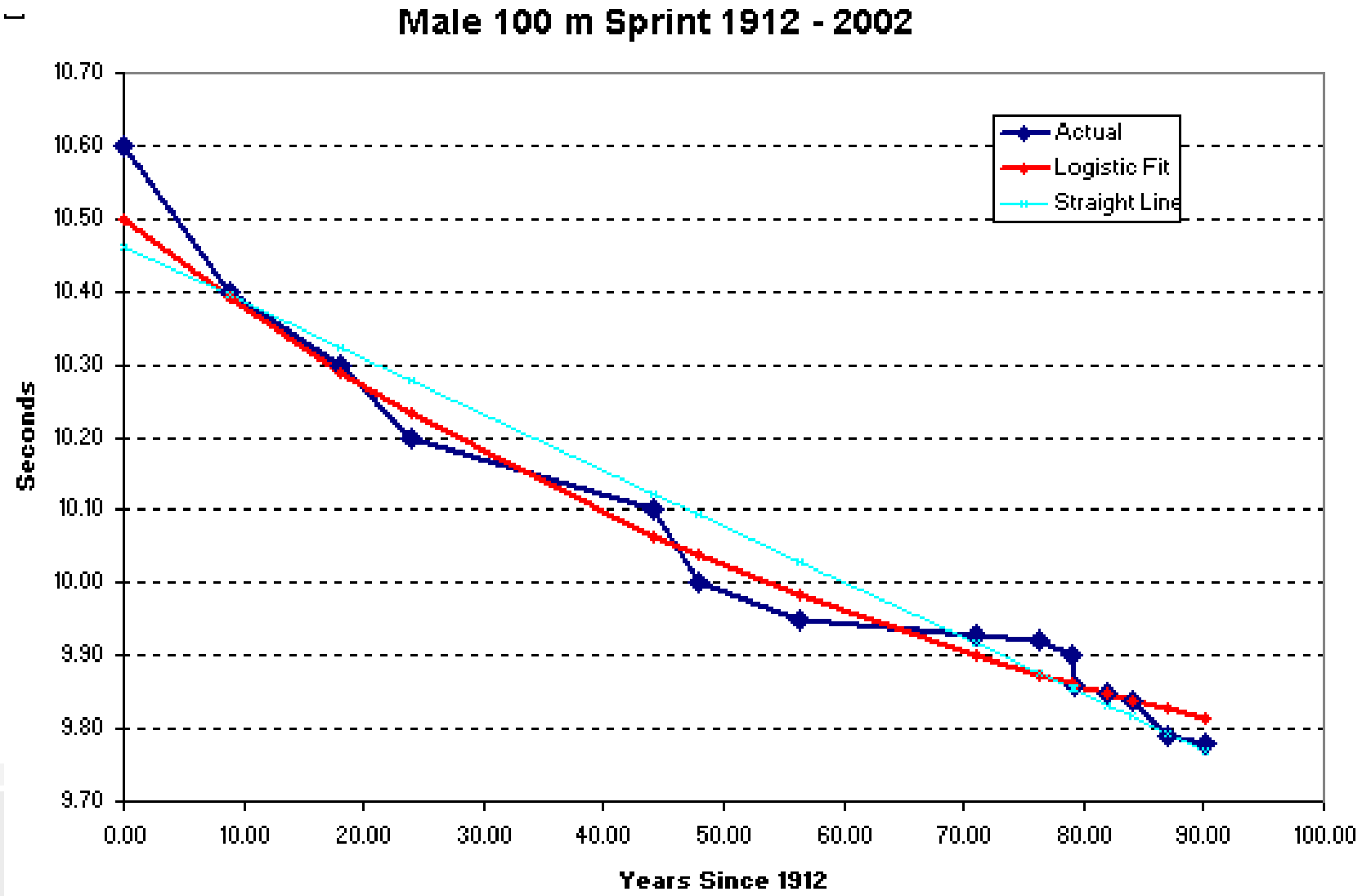
2 numerical variables, usually X = time





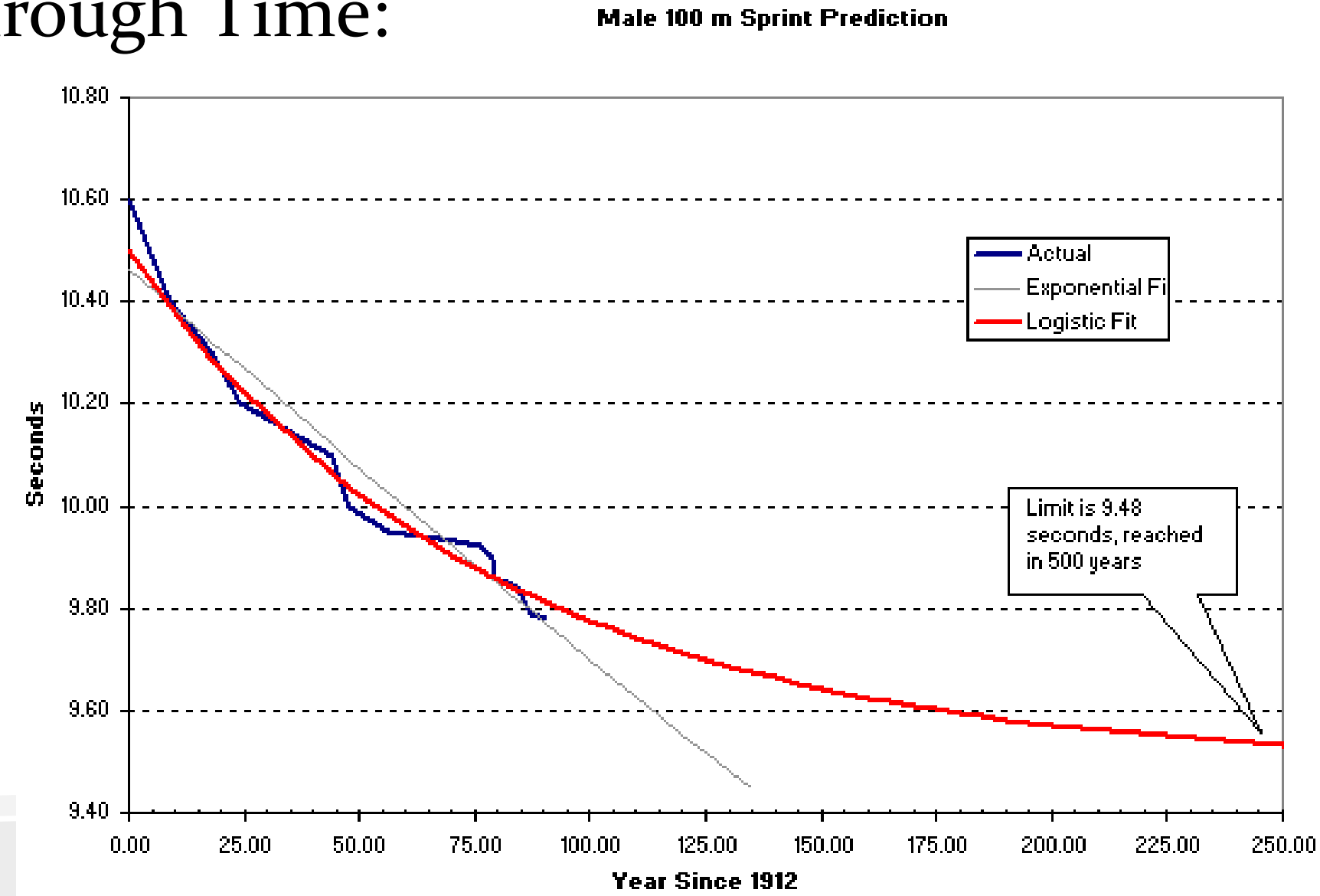
… insights through Time:

Trends

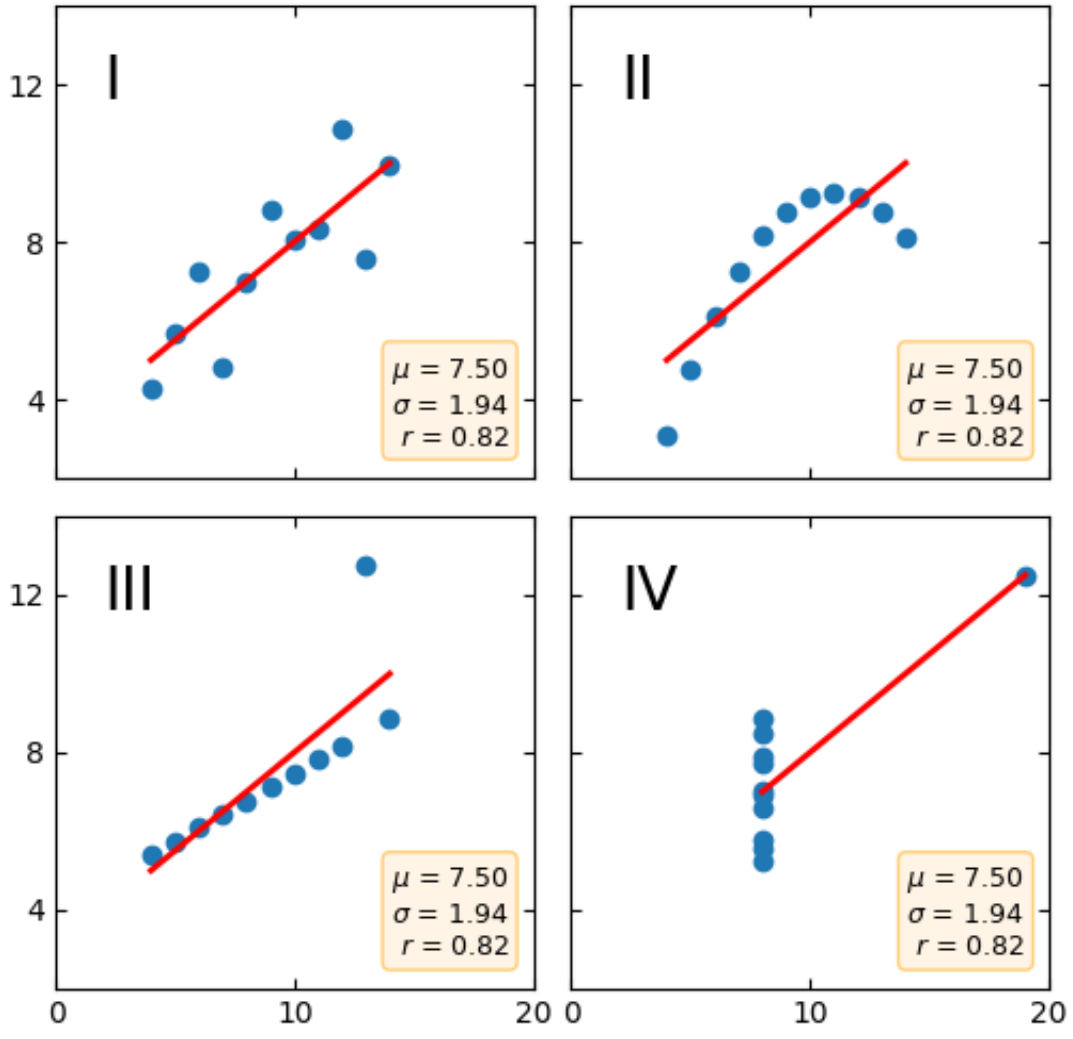


… insights through Time:

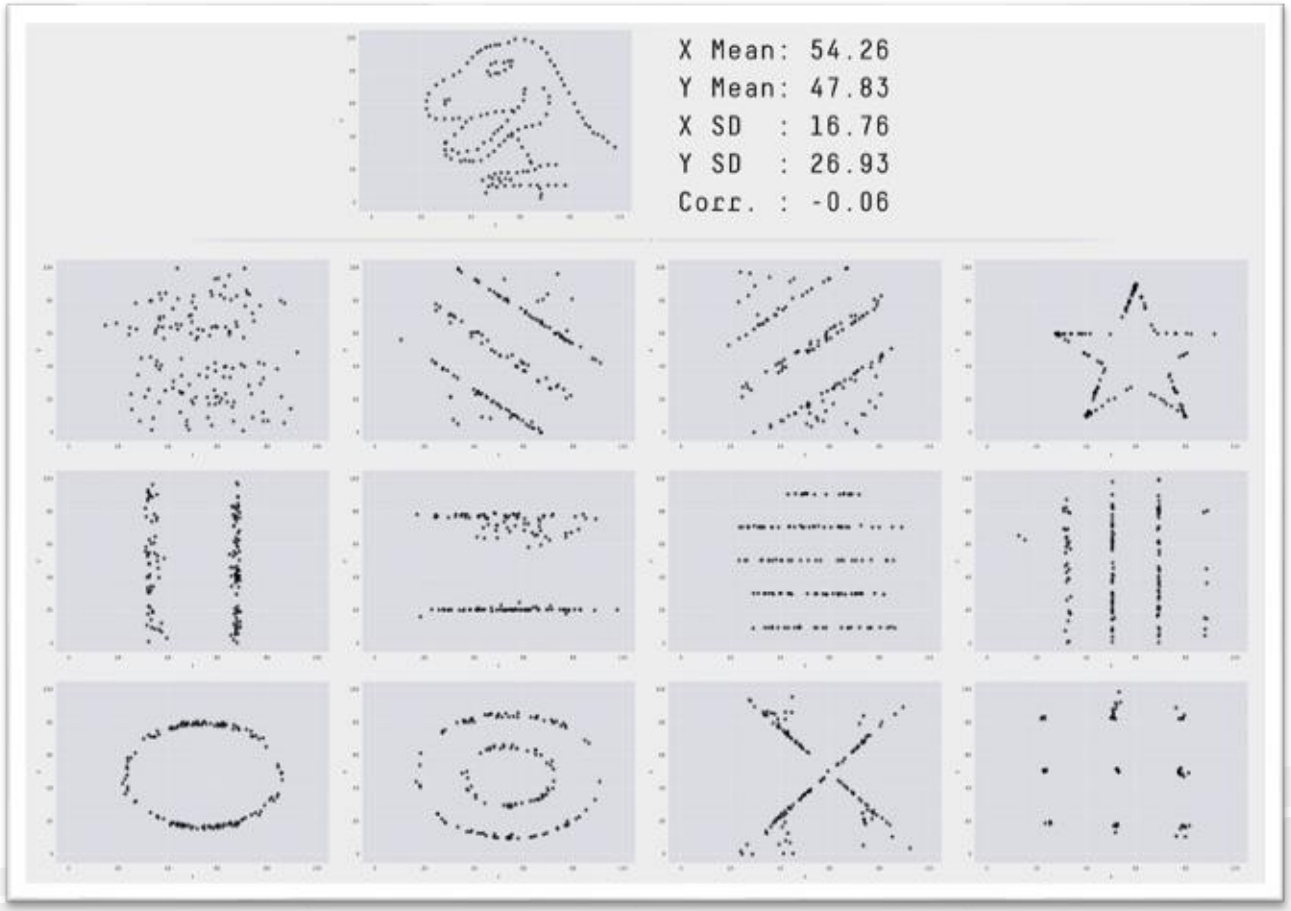
Trends



Why is data visualization important?

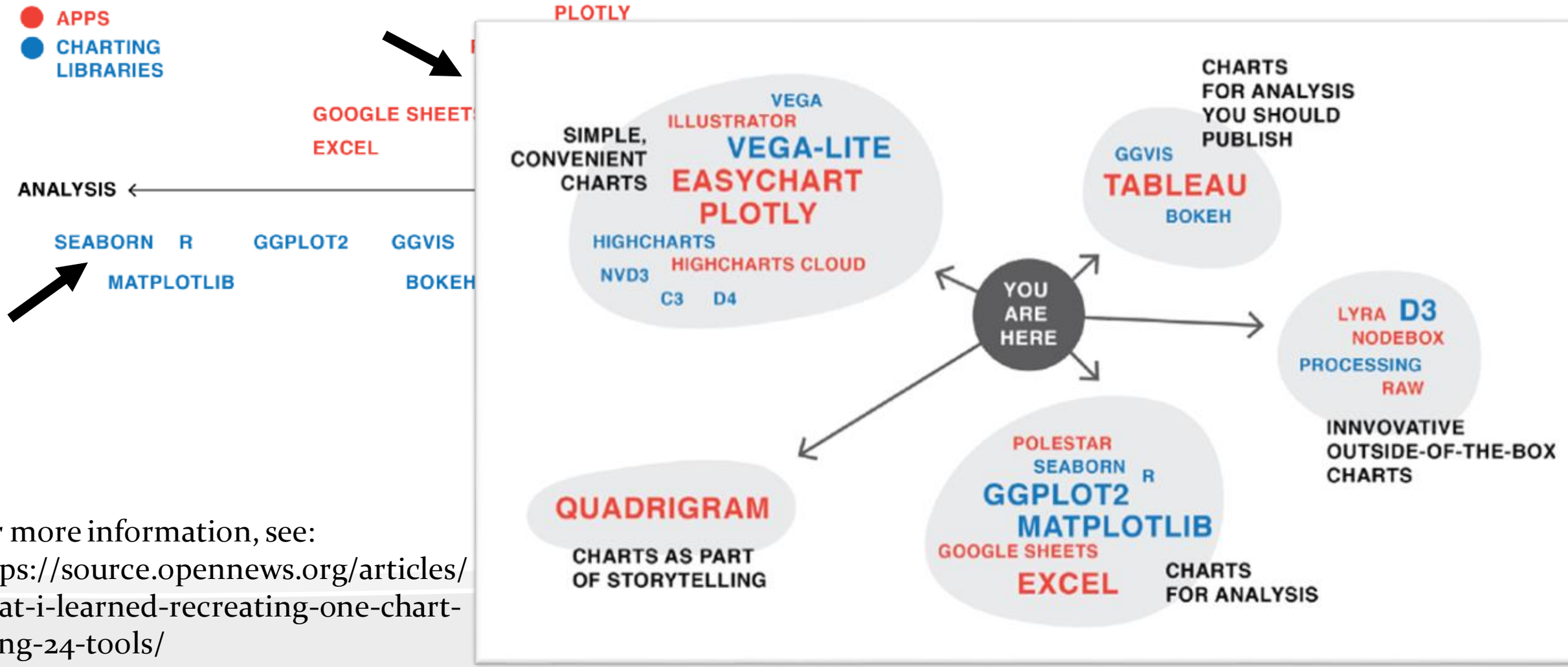


Source: https://matplotlib.org/3.2.1/gallery/specialty_plots/anscombe.html



Source: https://www.autodesk.com/research/publications/same-stats-different-graphs

Popular visualization tools



For more information, see:
https://source.opennews.org/articles/what-i-learned-recreating-one-chart-using-24-tools/

Tableau and/or Python

- Simple for the exploration phase.
- Allows also for publishing online interactive visualizations with storytelling elements.
- Allows for connecting with Big Data sources such as Spark and MongoDB.

- Complete control and reproducibility of datasets
- Active community
- Integration with complex libraries like machine learning
- Can serve web based visualizations (via, for instance, Dash by Plotly)

Tools for data exploration

- reproducibility of steps is key; that dismisses traditional Excel
- 2 classes of tools
  - end-user tools (GUI based), e.g. Tableau, KNIME, Power BI
  - scripting tools: SQL, Python, R, Julia, JavaScript, ...

Typical problems you can run into when using data prep tools

- data import decisions are usually based on the first n (n < 1000) rows
- dates, dates, dates
- discerning between missing values and empty values/strings
- adaptors for specific data sources / data storage solutions (HTML, XML, databases, filesystems, applications)
- level of documentability: Notebooks shine in this respect

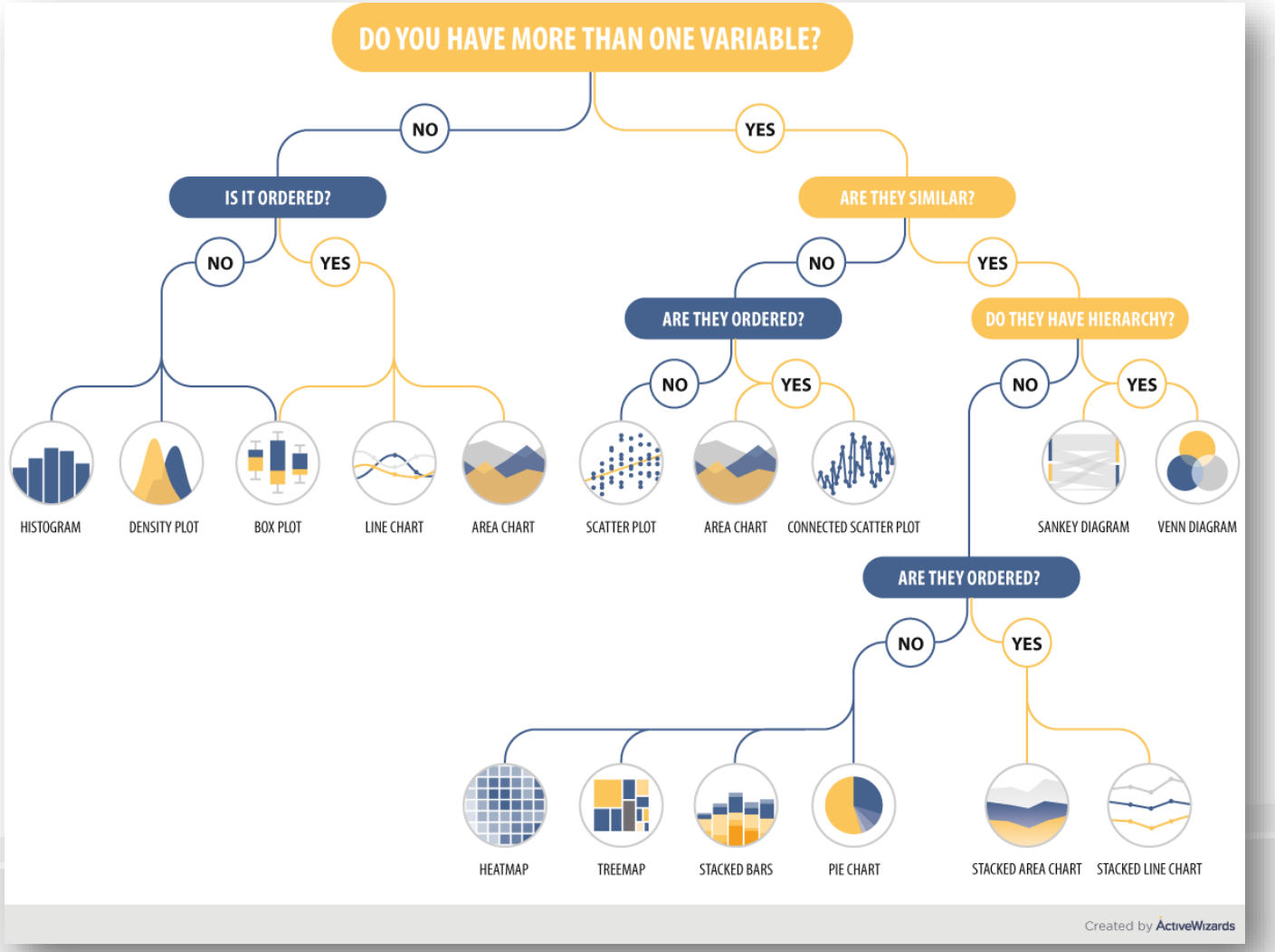
pandas - what should you know

- inspired by R data frames, built on NumPy (and Matplotlib)
- there's more than one way you can do it. That makes using it convenient and difficult at the same time
- no smooth learning curve, although you can get initial work done with a small subset
- it takes quite some time (and necessary mileage) to become a pandas expert
- Numpy is stable, pandas is improved and extended at a rapid rate
- pandas: almost anything regarding data preparation can be done with it
- DataFrame size is bound to available, physical memory:
  - no fancy storage solutions in core pandas
  - a lot of fancy stuff is recently released or in the making (parquet, Arrow)

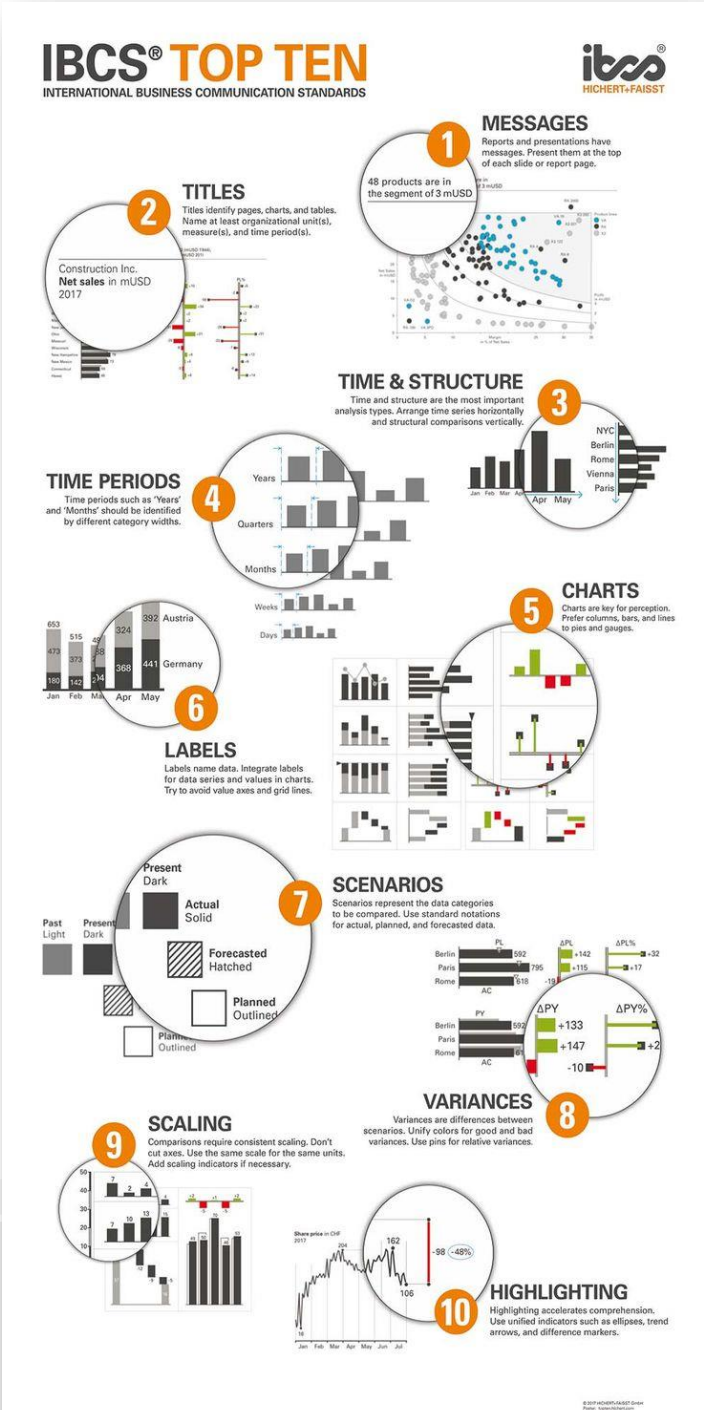
pandas - basic (necessary) subset

- importing data
  - read from an Excel or csv file
  - read from a Web API
- transforming data
  - handling missing values (NaN, None)
  - mutating: adding columns based on values in other columns
  - subsetting and filtering
  - grouping and aggregating
- presenting data
  - using pandas' built in plot functionality
  - using Seaborn
  - using MatPlotLib (only limited subset)

Useful standards, guidelines



https://activewizards.com/content/blog/How_to_Choose_the_Right_Chart_Type/chart-types-infographics04.png



https://www.ibcs-shop.com/product/ibcs-top-ten-pdf-of-poster/

Useful resources

- A comprehensive Chart Guide poster (+ large hi-res version)

- https://elitedatascience.com/python-seaborn-tutorial - short tutorial on the **Seaborn** library, a fast way of making beautiful charts in Jupyter Notebooks

- http://seaborn.pydata.org/examples/index.html - fancy chart examples made with Seaborn, with code samples

Go prepare your ingredients!

Check the exercise(s) on Canvas