MTAT.03.319

# Business Data Analytics

## Lecture 2
## Descriptive analysis and visualization

Rajesh Sharma
https://css.cs.ut.ee/

# Are you sure about data?


1973 admissions' data

| | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| **Total** | 8442 | **44%** | 4321 | 35% |

Source: https://en.wikipedia.org/wiki/Simpson%27s_paradox

# Are you sure about data?

Investigation


| | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| **Total** | 8442 | **44%** | 4321 | 35% |

Source: https://en.wikipedia.org/wiki/Simpson%27s_paradox

# Are you sure about data?

| | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| **Total** | 8442 | **44%** | 4321 | 35% |

| Department | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| **A** | ***825*** | 62% | 108 | **82%** |
| **B** | ***560*** | 63% | 25 | **68%** |
| **C** | 325 | **37%** | ***593*** | 34% |
| **D** | 417 | 33% | 375 | **35%** |
| **E** | 191 | **28%** | ***393*** | 24% |
| **F** | 373 | 6% | 341 | **7%** |

Source: https://en.wikipedia.org/wiki/Simpson%27s_paradox

# Are you sure about data?

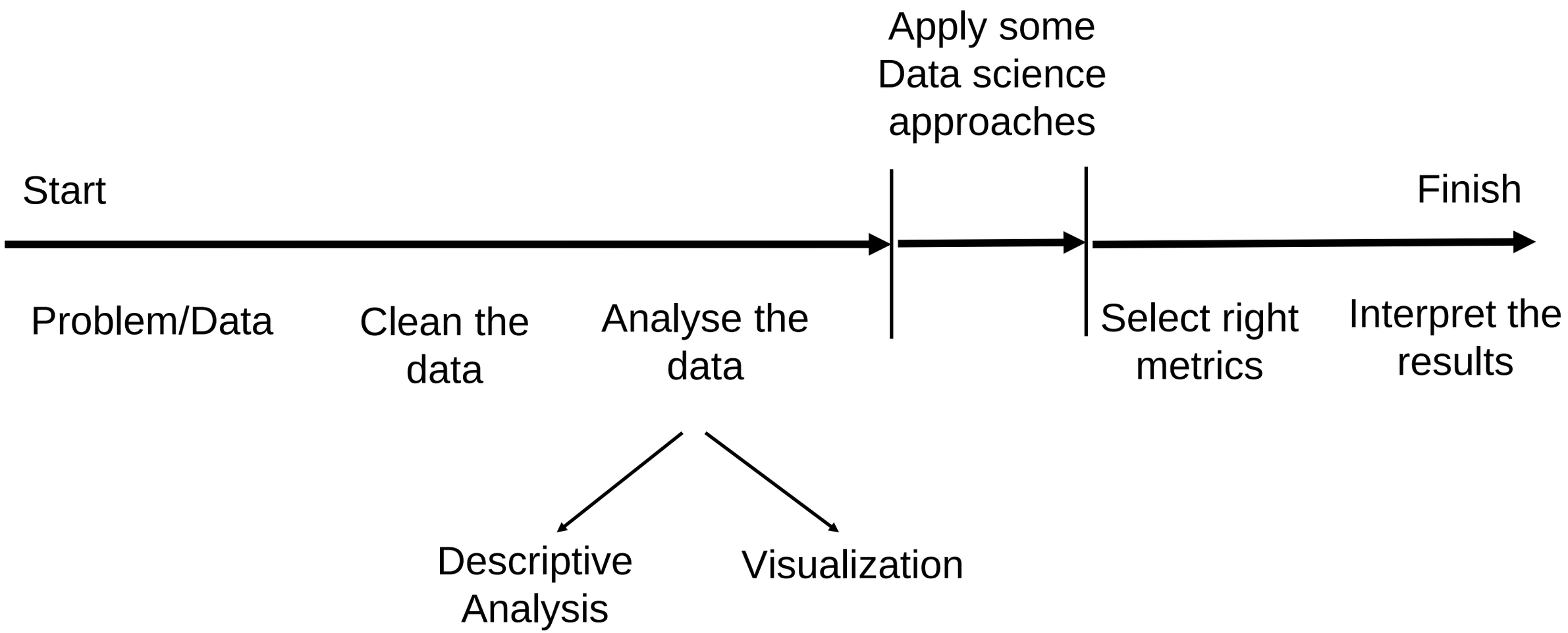
**Simpson's paradox**

| | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| **Total** | 8442 | **44%** | 4321 | 35% |

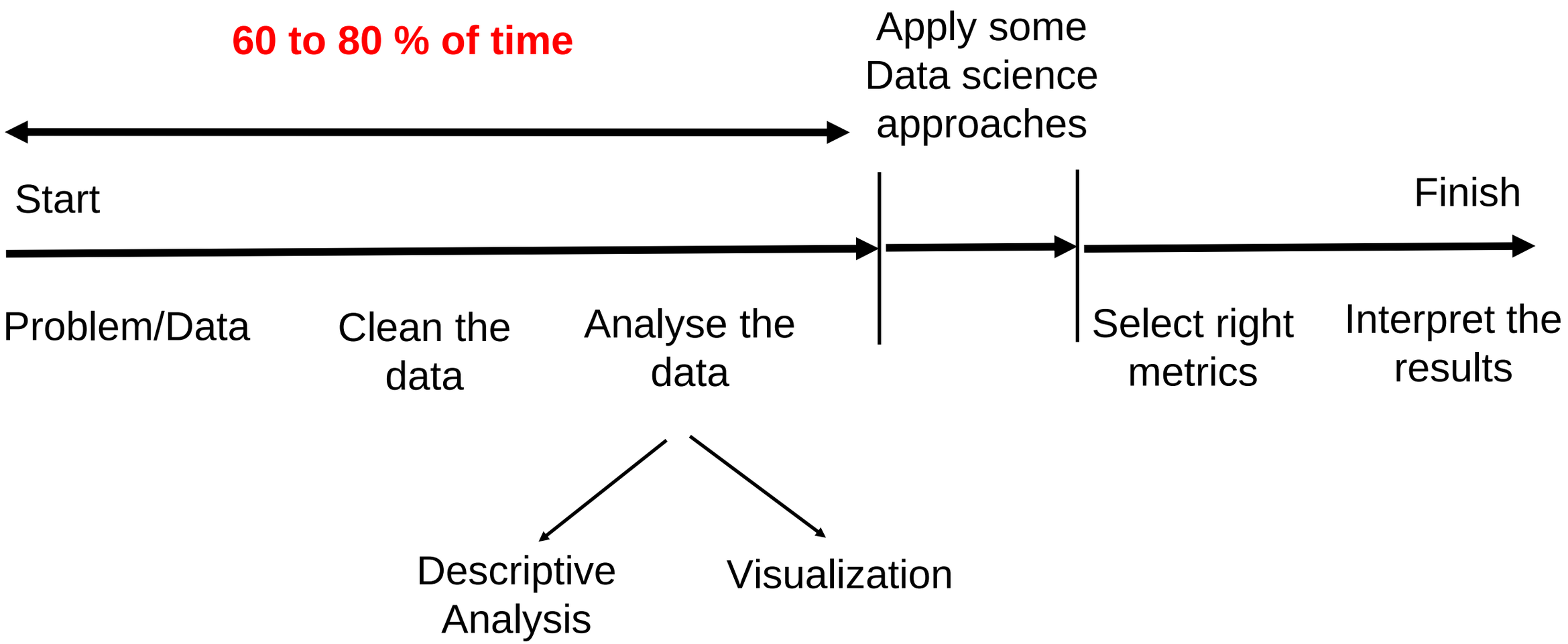
| Department | Men | | Women | |
|---|---|---|---|---|
| | Applicants | Admitted | Applicants | Admitted |
| **A** | ***825*** | 62% | 108 | **82%** |
| **B** | ***560*** | 63% | 25 | **68%** |
| **C** | 325 | **37%** | ***593*** | 34% |
| **D** | 417 | 33% | 375 | **35%** |
| **E** | 191 | **28%** | ***393*** | 24% |
| **F** | 373 | 6% | 341 | **7%** |

Source: https://en.wikipedia.org/wiki/Simpson%27s_paradox

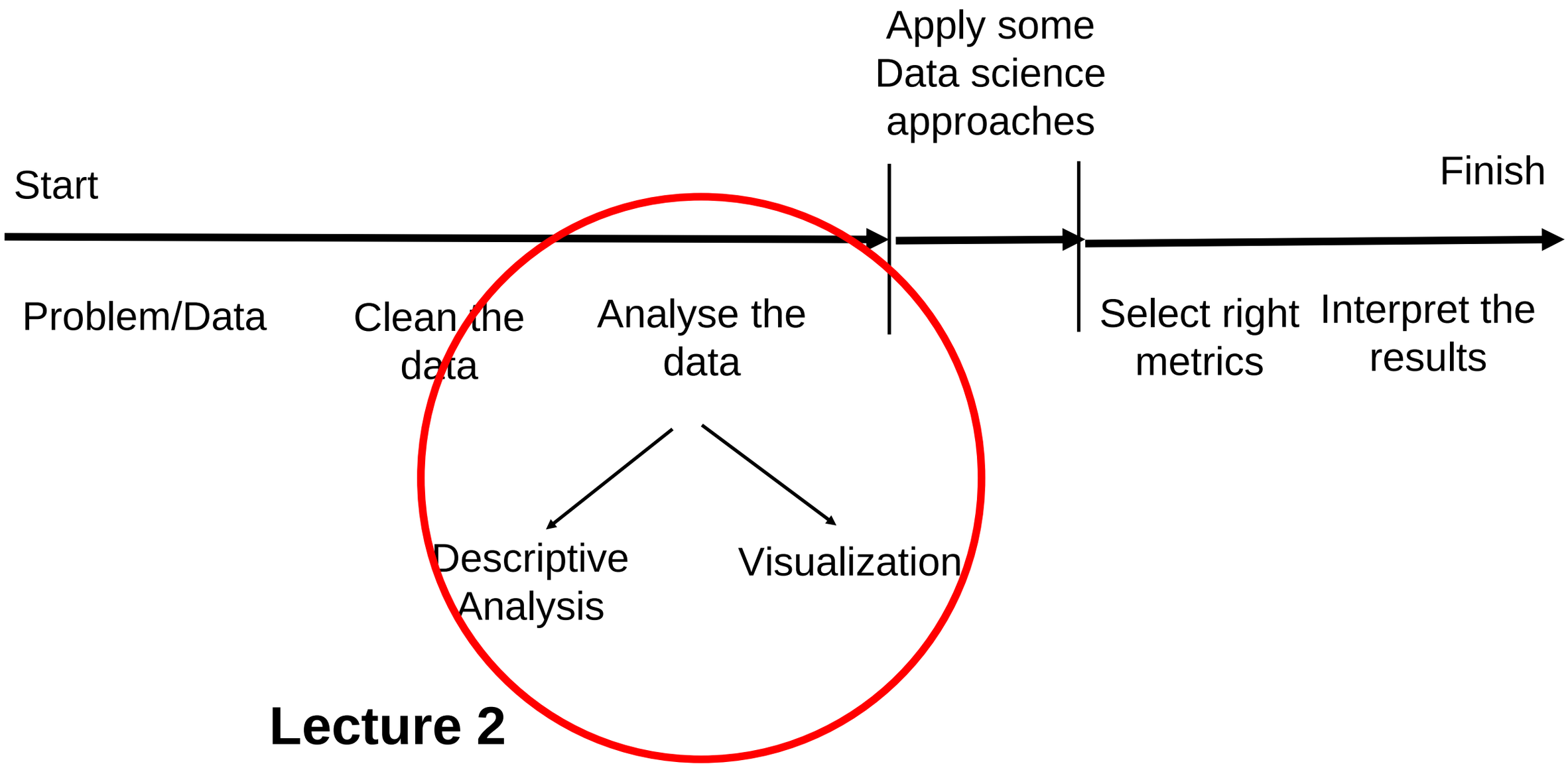
# A Generic data science approach



# A Generic data science approach



# A Generic data science approach




**Lecture 2**

# Last Lecture and this ..

Lecture 1: BDA basics

1. Business sponsor: Increase sales in Bank
2. Objective is defined: Let us do cross/up selling
3. Domain expert: Finance expert
4. Data steward: Prepare data (find which tables have information of customers, combine with sales etc.)
5. Data Analyst: Descriptive analysis and visualization

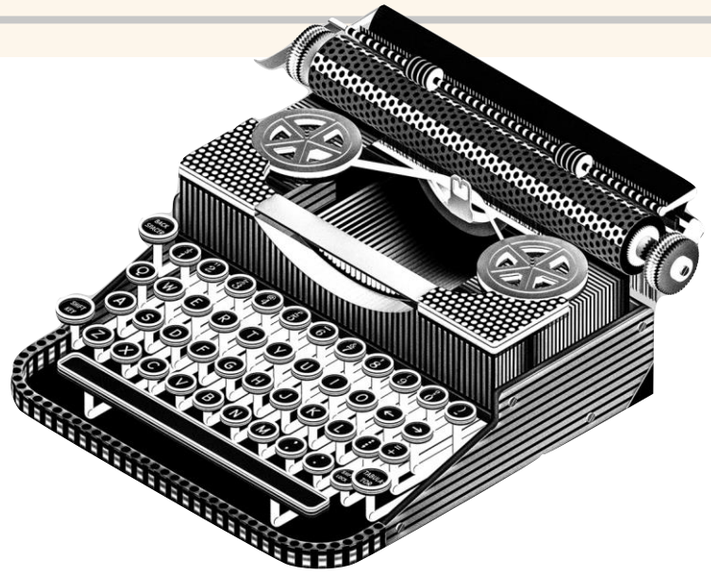
Lecture 2

# Lecture 2:
# Descriptive analysis
# and visualization

# Outline

- Data
  - Intuition Vs. Data based decisions
  - Characteristics of Data
  - Common Risks
- Descriptive Analysis
  - Understanding data with statistical measures
- Visualizations
  - Understand data by plotting the data

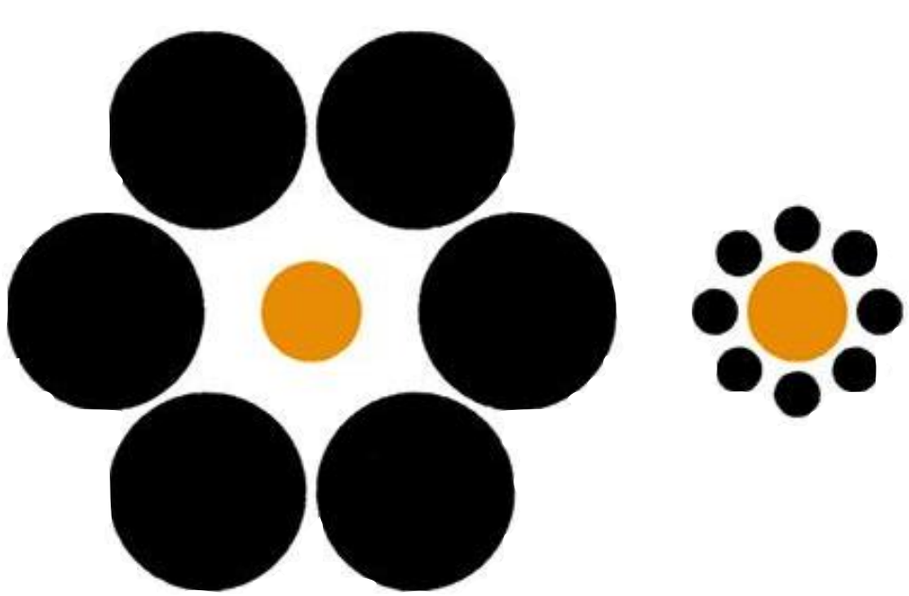
"about 1/5 of business decision-makers don't really understand what big data is or still believe that big data is a lot of hype"



source: Forrester's Global Business Technographics Data And Analytics Survey, 2015

# Informative business decisions vs. intuition

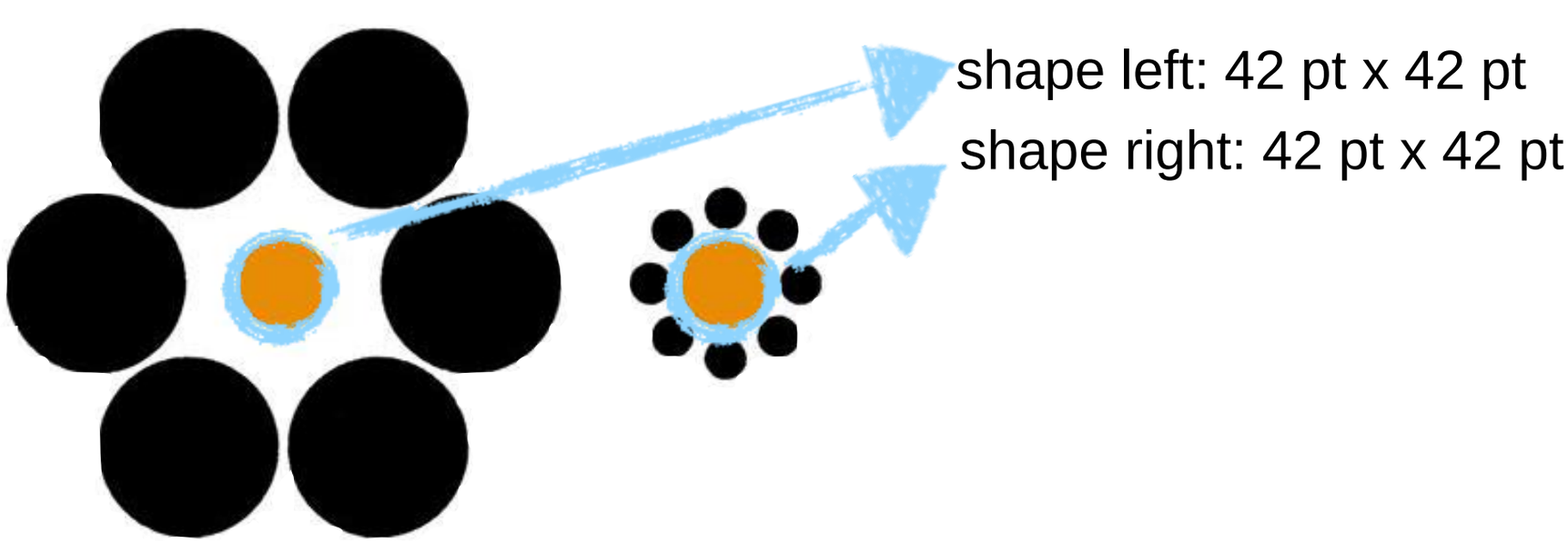
- many business decisions remain based on intuitive hunches, not facts



source: Forrester’s Global Business Technographics Data And Analytics Survey, 2015

# Informative business decisions vs. intuition

many business decisions remain based on
intuitive hunches, not facts




analytics helps to reduce the gap between intuition
and factual decision-making

source: Forrester’s Global Business Technographics Data And Analytics Survey, 2015

# Informative business decisions vs. intuition

- many business decisions remain based on intuitive hunches, not facts

- analytics helps to reduce the gap between intuition and factual decision-making

- sophisticated data usage brings competitive gains


image source: making of mass effect andromeda on behance

# Informative business decisions vs. intuition

- many business decisions remain based on intuitive hunches, not facts

- analytics helps to reduce the gap between intuition and factual decision-making

- sophisticated data usage brings competitive gains

- data does not speak for itself. it should be analyzed to take full advantage of its potential

source: Forrester’s Global Business Technographics Data And Analytics Survey, 2015

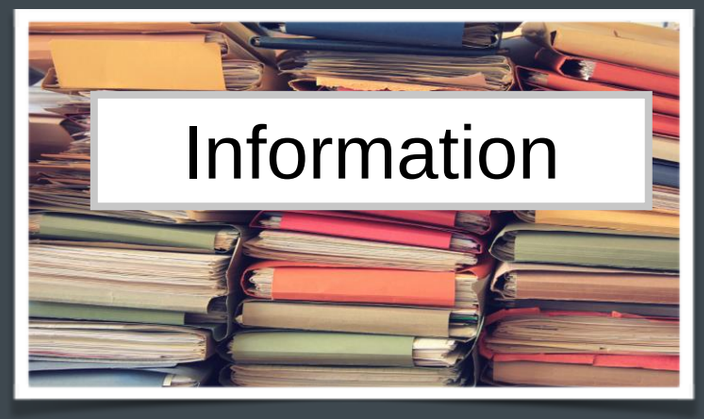
# Data is not yet knowledge

collect


Example: The price of crude oil is $80 per barrel

clean organize

summarize filter




Example: The price of crude oil has risen from $70 to $80 per barrel

analyze


Example: Should the company buy petrol or not ?

action

decision making


Example: When crude oil prices go up by $10 per barrel, it's likely that petrol prices will rise by 2p per litre

# Usable data is

clean

consistent

comprehensive

current

# Most common risks

- Organization will not have the expertise to use the tools
- Organization will not have the expertise of concepts and techniques
- Business people will not understand how to obtain business values out of BA

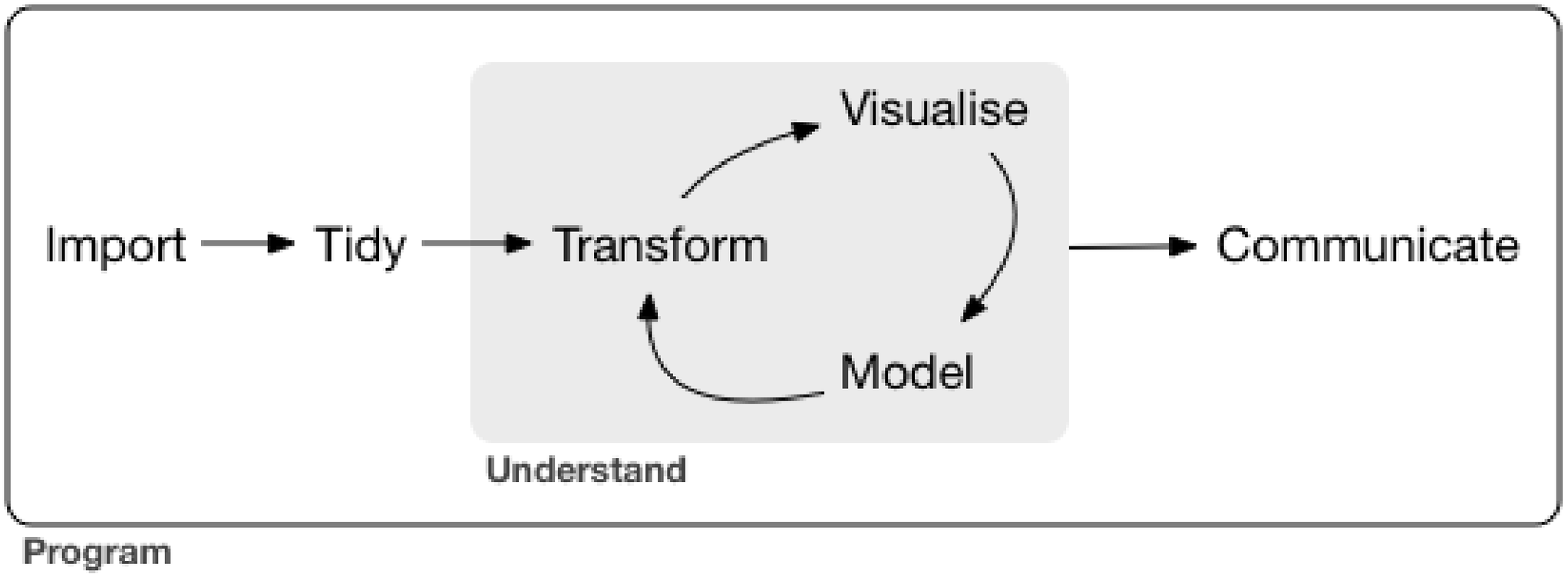
# Most common risks

- Organization will not have the expertise to use the tools
- Organization will not have the expertise of concepts and techniques
- Business people will not understand how to obtain business values out of BA

Examples are easy and clean.
Real data is messy.

# Analysis: principles

# Steps of data analysis




source: R for Data Science
http://r4ds.had.co.nz/introduction.html

# What is tidy data

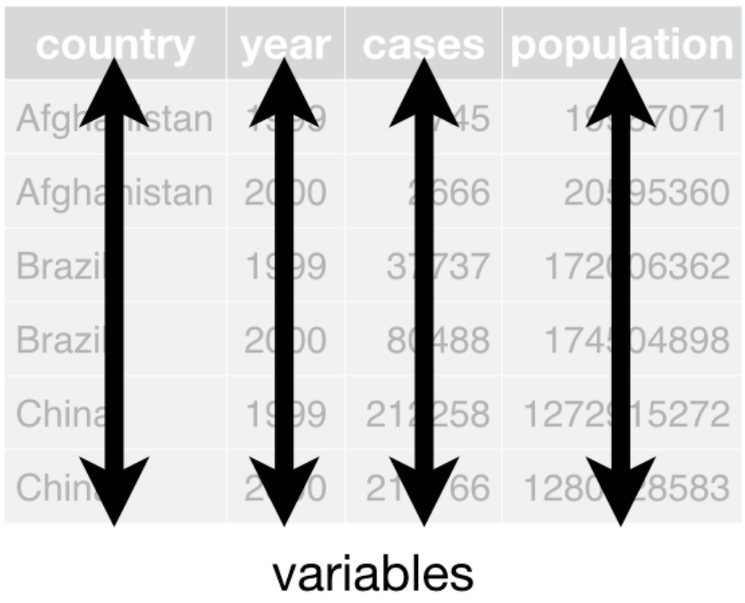
Tidy format: common name for common statistical form called a model matrix or data matrix.

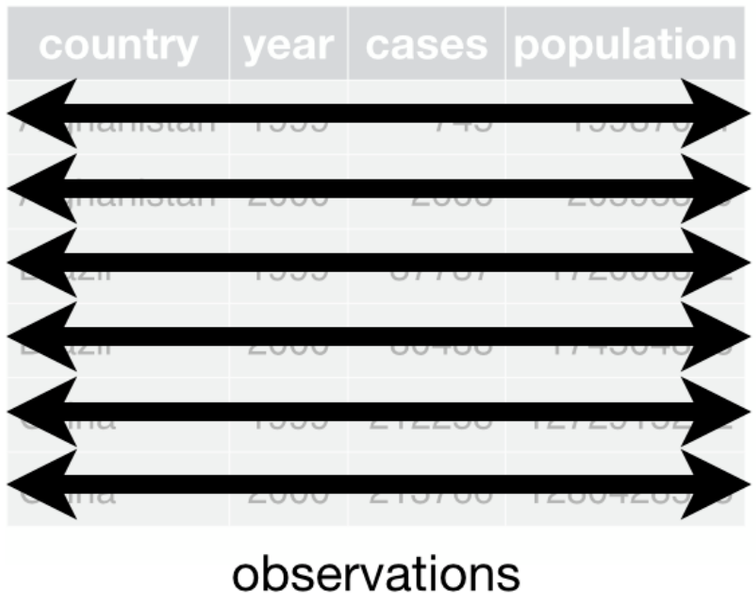
3 principles:

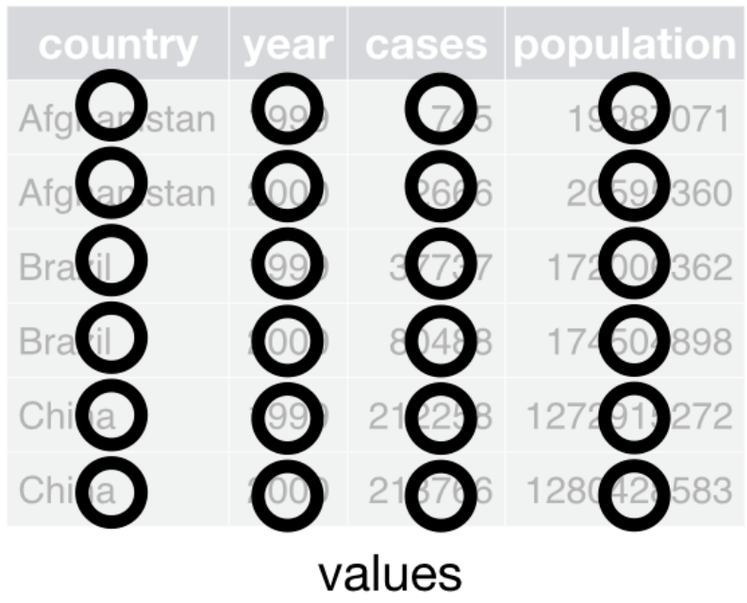
each variable forms a column

each observations forms a row

each type of observational unit forms a table




source: R for Data Science
http://r4ds.had.co.nz/introduction.html

# 2 Kinds of data formats

wide format

| | treatmenta | treatmentb |
|---|---|---|
| John Smith | — | 2 |
| Jane Doe | 16 | 11 |
| Mary Johnson | 3 | 1 |

wide format

| | John Smith | Jane Doe | Mary Johnson |
|---|---|---|---|
| treatmenta | — | 16 | 3 |
| treatmentb | 2 | 11 | 1 |

long format

| person | treatment | result |
|---|---|---|
| John Smith | a | — |
| Jane Doe | a | 16 |
| Mary Johnson | a | 3 |
| John Smith | b | 2 |
| Jane Doe | b | 11 |
| Mary Johnson | b | 1 |

# 2 Kinds of data formats

wide format

| | treatmenta | treatmentb |
|---|---|---|
| John Smith | — | 2 |
| Jane Doe | 16 | 11 |
| Mary Johnson | 3 | 1 |

wide format

| | John Smith | Jane Doe | Mary Johnson |
|---|---|---|---|
| treatmenta | — | 16 | 3 |
| treatmentb | 2 | 11 | 1 |

long format

| person | treatment | result |
|---|---|---|
| John Smith | a | — |
| Jane Doe | a | 16 |
| Mary Johnson | a | 3 |
| John Smith | b | 2 |
| Jane Doe | b | 11 |
| Mary Johnson | b | 1 |

Missing data (simple solution: remove it)

# Imputation Techniques

## Univariate Imputation

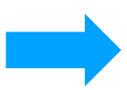
Fills the values in a feature by using the non-missing values from that feature.

Replacement of missing values either by:

- Constant value
- Mean, Median or other statistical value of the feature
- Most or least frequent etc

Can work on Categorical or Numeric data

| ID | Name | Age |
|---|---|---|
| 1 | ABC | 20 |
| 2 | | 10 |
| 3 | ABC | |
| 4 | XYZ | 20 |



| ID | Name | Age |
|---|---|---|
| 1 | ABC | 20 |
| 2 | ABC | 10 |
| 3 | ABC | 20 |
| 4 | XYZ | 20 |

## Multivariate Imputation

Uses the entire feature set to estimate the missing values.

| ID | F1 | F2 |
|---|---|---|
| 1 | 1 | 0.5 |
| 2 | 4 | 2 |
| 3 | 10 | |
| 4 | | 20 |

| ID | F1 | F2 |
|---|---|---|
| 1 | 1 | 0.5 |
| 2 | 4 | 2 |
| 3 | 10 | 5 |
| 4 | 40 | 20 |

Source: https://medium.com/fintechexplained/what-are-imputers-in-data-science-b72f8308322b

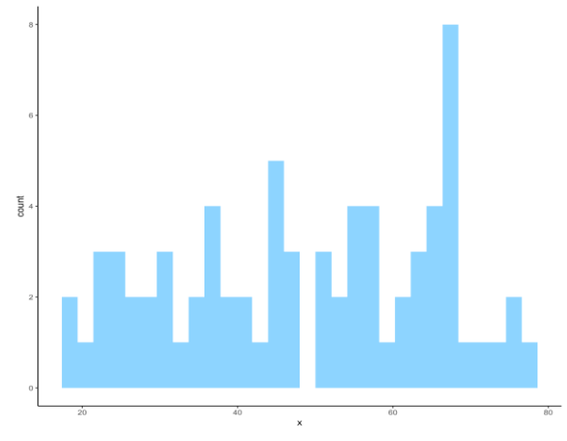
# Exploratory phase

Explore via

## Descriptive statistics

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 18.72 | 35.13 | 51.54 | 48.60 | 63.33 | 77.95 |

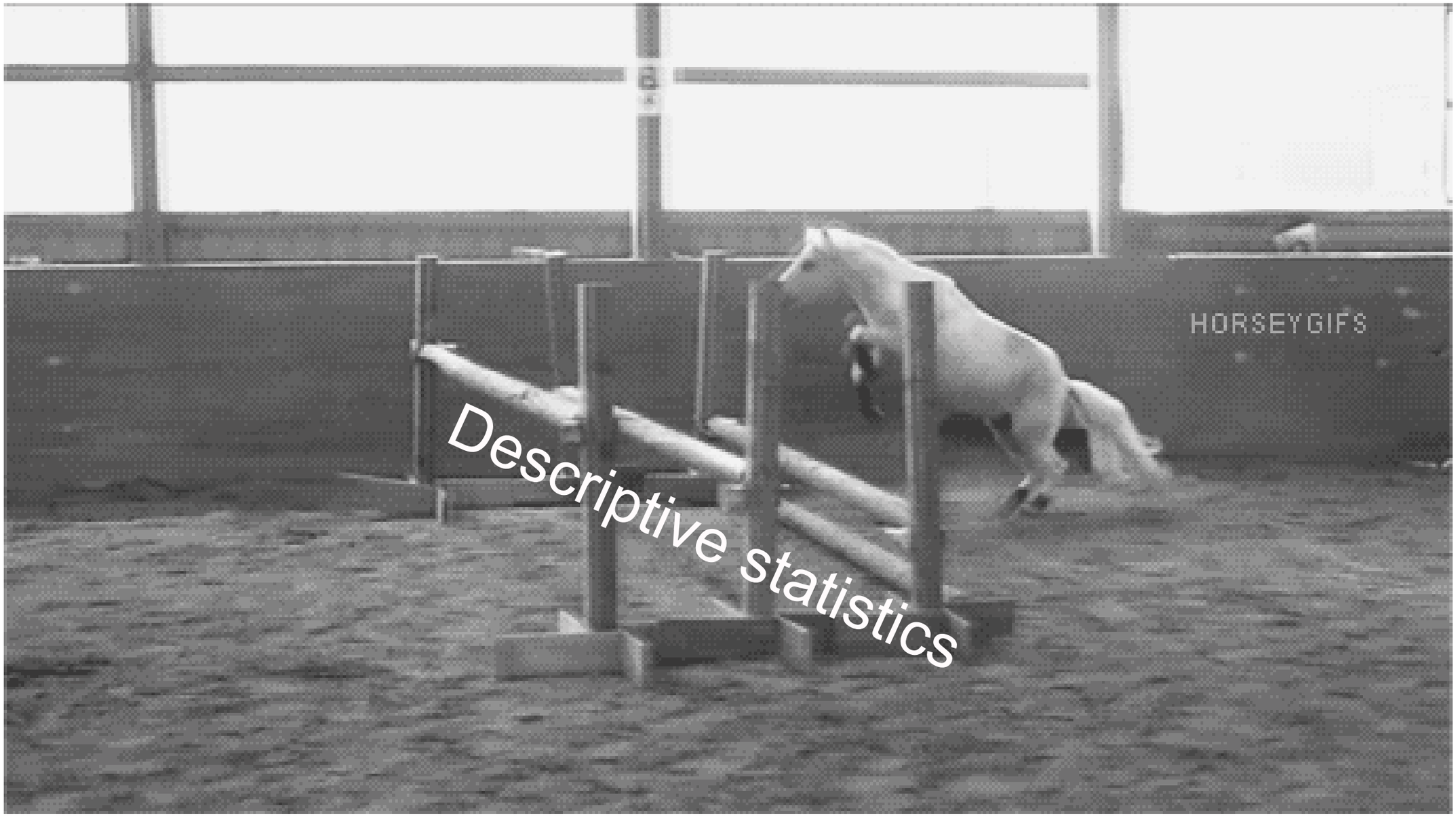
## Descriptive Plots



# Data types

- categorical
  - binary
  - nominal
  - ordinal
- numerical
  - discrete
  - continuous





HORSEYGIFS
Descriptive statistics

- Central tendency measures: Mean, Mode, Median.

  computed to provide a ‘center’ around which observations are distributed

- Variation measures: Variance, Standard deviation.

  describe ‘data spread’ or the distance from the center.

- Relative measures: Percentiles

  description of relative positions of observations

# The Mode, the Median, and the Mean

x <- c(4,5,2,5,0,0,4,0,9,3)

mode: 0

sort(x): 0 0 0 2 3 4 4 5 5 9

median: 3.5

sum(x)/length(x)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

mean: 3.2

# Variance and standard deviation

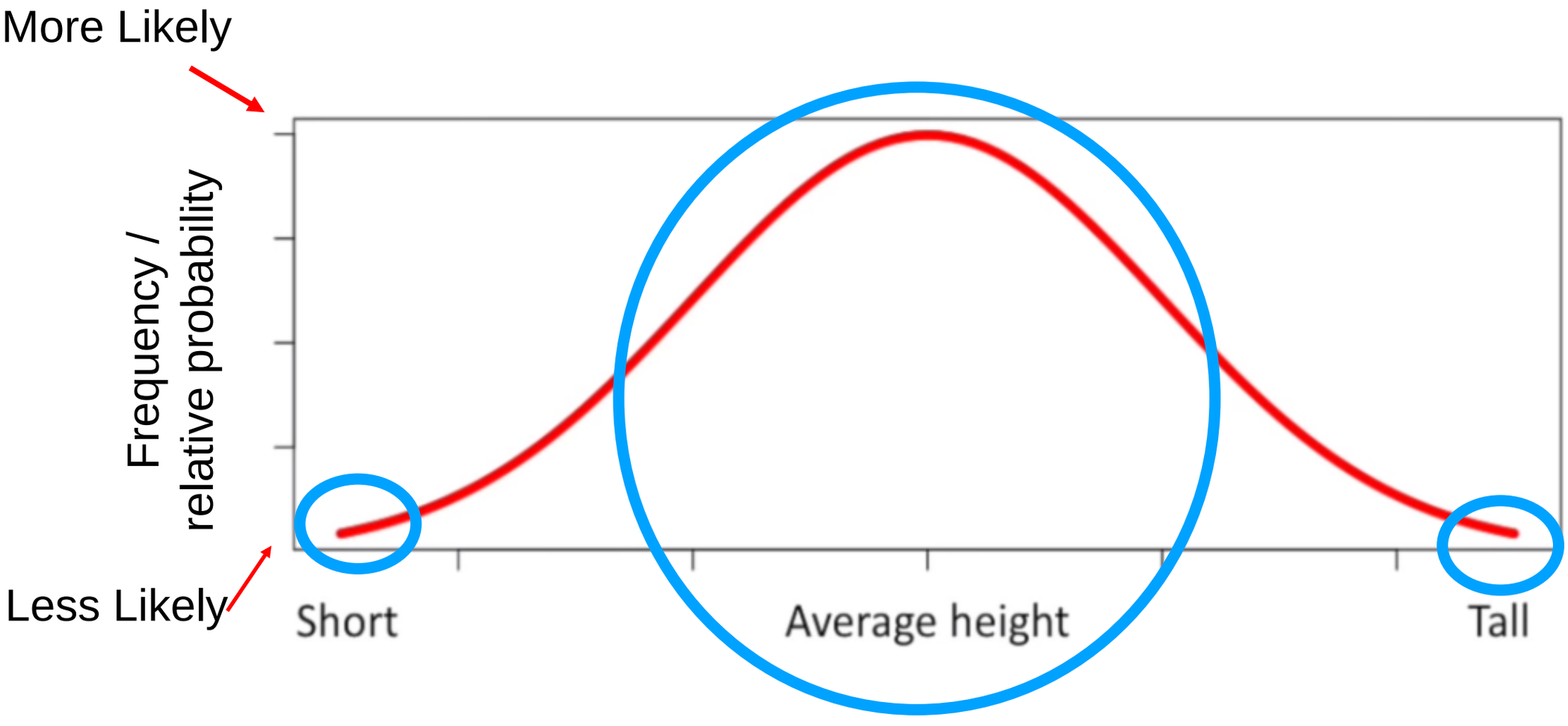
$$\text{variance} = \frac{\sum(\text{each observation} - \text{mean})^2}{\text{number of observations}}$$

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

$$\text{standard deviation} = \sqrt{\text{variance}}$$

# Data Distribution

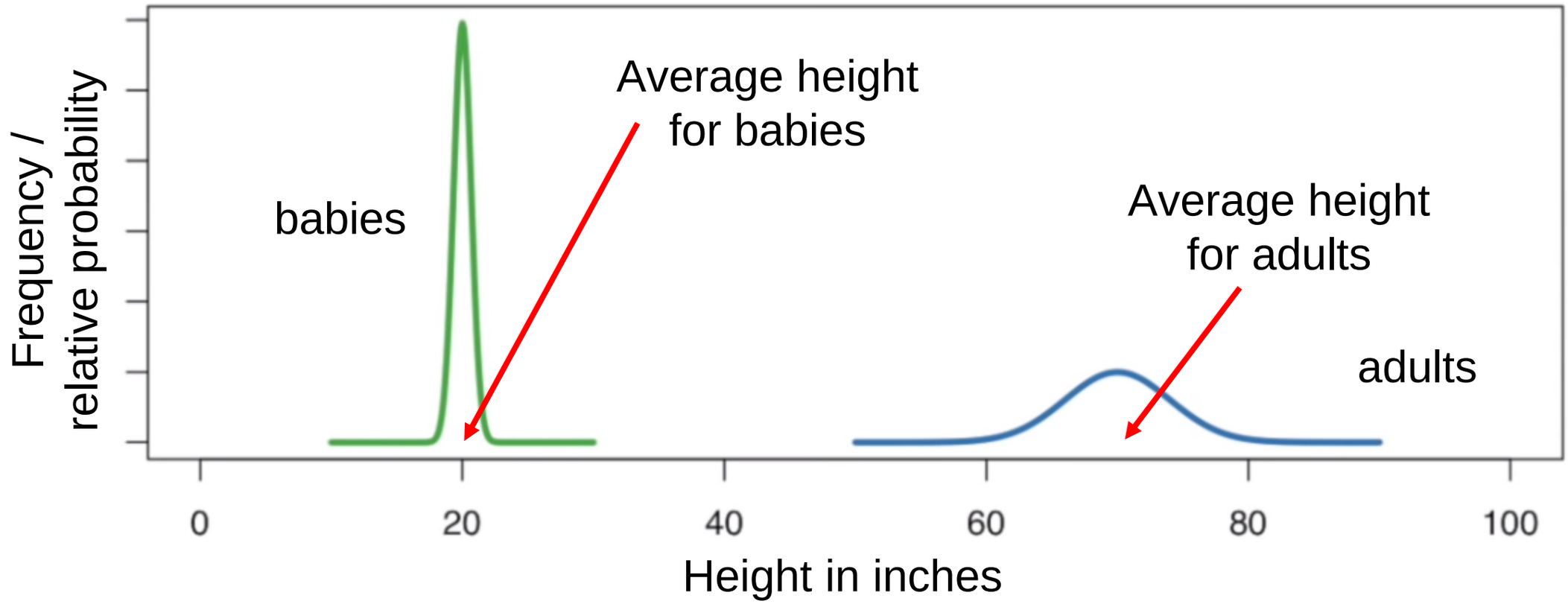
## Normal / Bell / Gaussian




Human height measurements

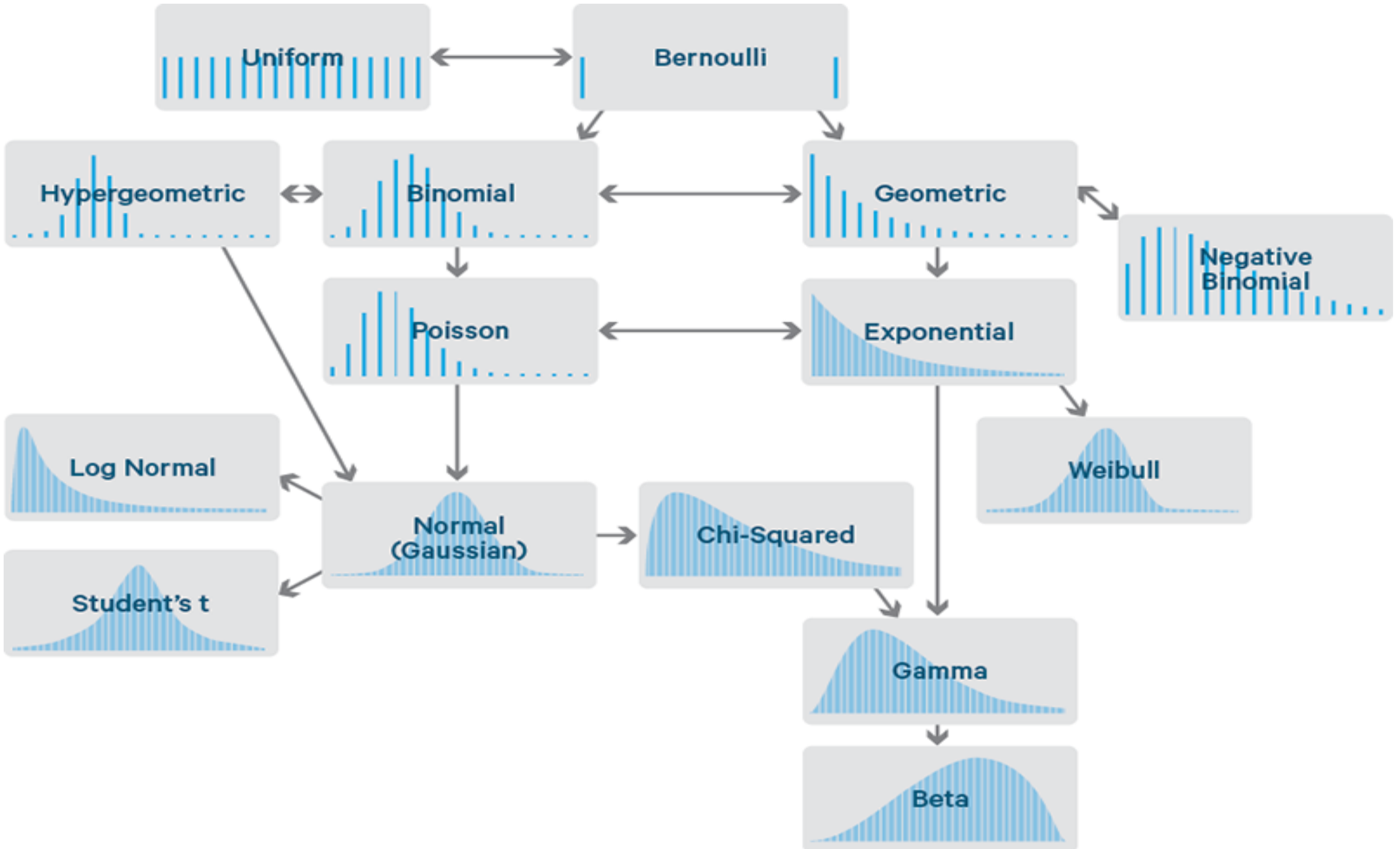
# Data Distribution

## Ideal case: Normal / Bell / Gaussian



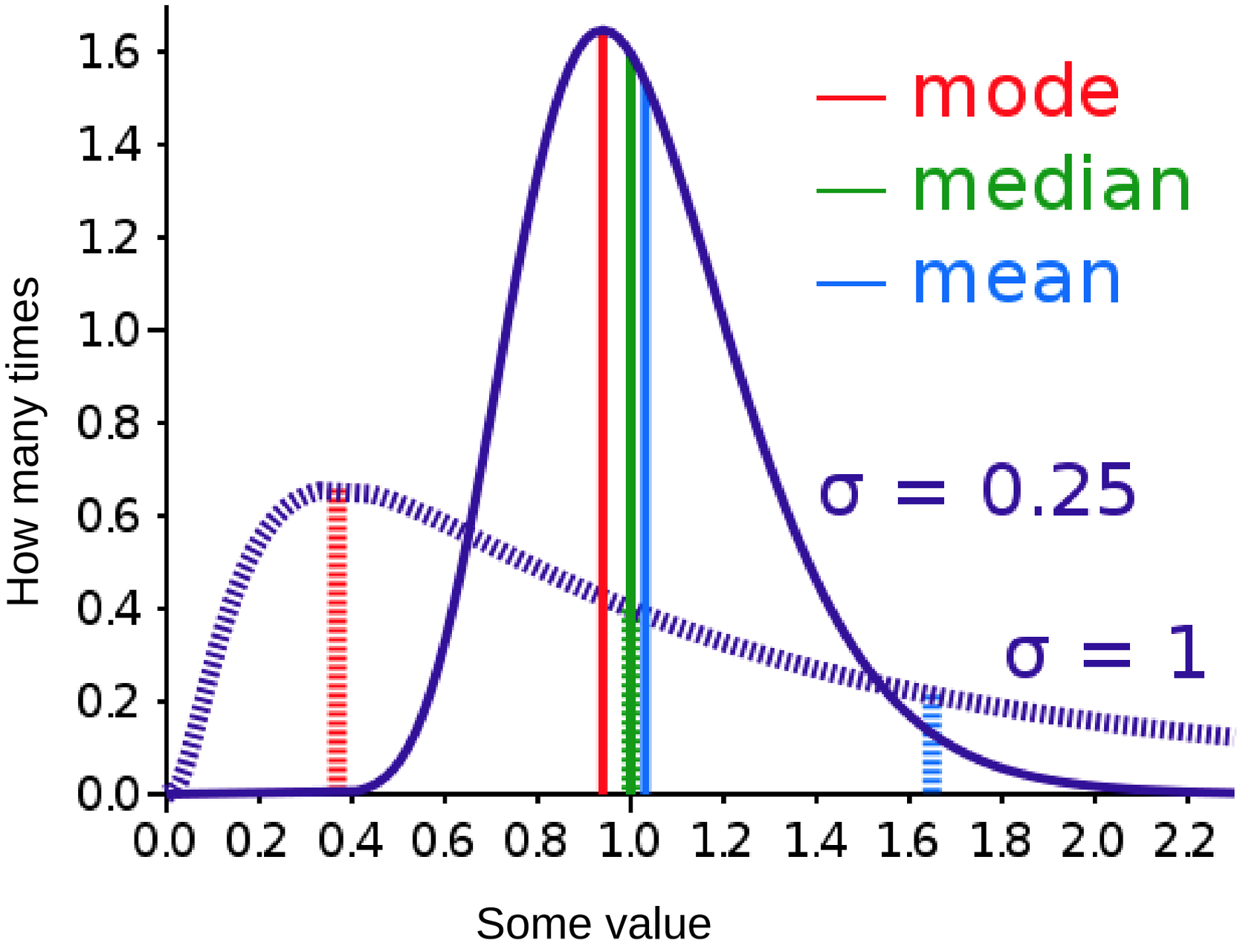

- Normal distribution always centered on average value
- Width of the curve is defined by the std deviation (4 for adults, and 0.6 for babies)
- 95% of the measurements fall between +/-2 std. deviations around the mean
- To draw a normal distribution:
    1) Avg measurement: Center of the curve.
    2) Std. Deviation: How wide the curve should be

# Distributions




Source: http://blog.cloudera.com/blog/2015/12/common-probability-distributions-the-data-scientists-crib-sheet/

# The Mode, the Median, and the Mean



- Central tendency measures: Mean, Mode, Median.

  computed to provide a ‘center’ around which observations are distributed

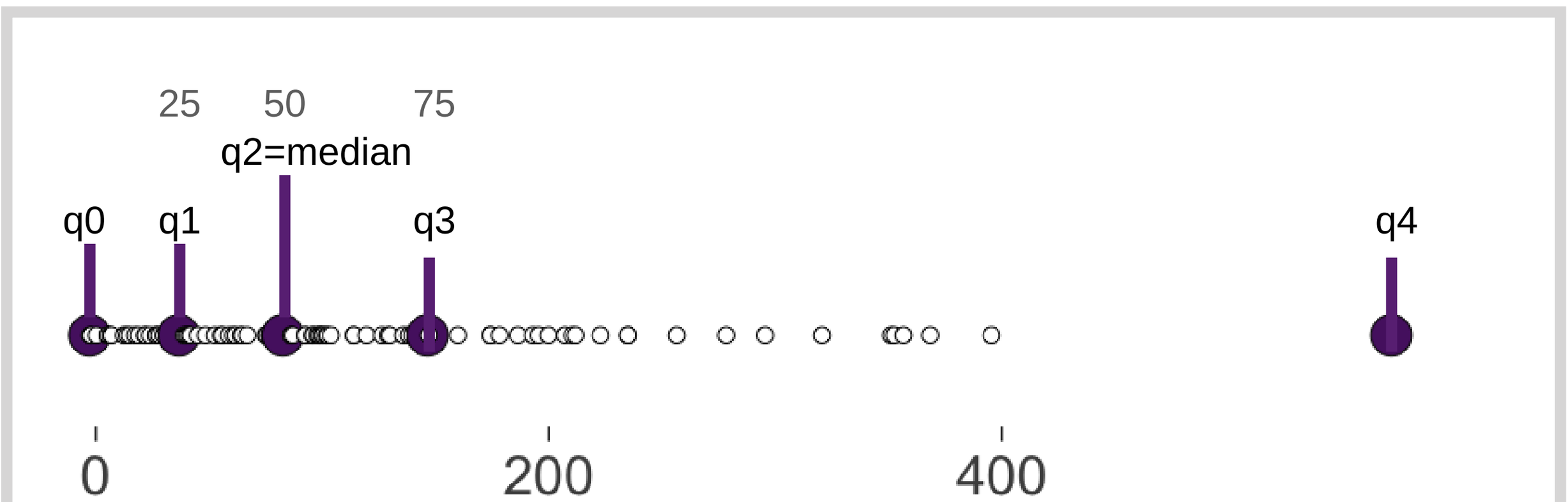
- Variation measures: Variance, Standard deviation.

  describe ‘data spread’ or the distance from the center.

- Relative measures: Percentiles

  description of relative positions of observations

# Quartiles, percentiles and IQR




the first quartile, $Q_1$, is the value for which 25% of the observations are smaller and 75% are larger

only 25% of the observations > Q3

# Percentiles, Quartiles and IQR

the nth percentile is a value such that n% of the observations fall at or below it

3, 12, 15, 16, 16, 17, 19, 34

What is the percentile ranking of 17?
= (# values below x *100)/n = (5*100)/8 = 62.5 %

What value exists at the percentile ranking of 25% ?
Value # = (Percentile * (n+1))/100  = (25 * (8+1))/100
Value# = 2.25 (Take average of 2$^{nd}$ and 3$^{rd}$ values)
Value# = (12+15)/2 = 13.5

# Outlier Detection

- Identifying the data points which are behaving very differently from the most of the data points.

- **Q:** Should we remove these Outliers ? **Please note:** they are not necessarily bad data points.
  **A:** Depends on your objective.

- We will see different ways to identify outliers.
  **Visualization techniques**: Box plots, Scatter plot, Histogram, line plots, Bar plots, Distribution plot
  **Mathematical functions:** IQR, Z- score

# Outlier using 1.5*IQR rule

IQR = Q3 – Q1 Q1 = ¼(n+1) Q3 = ¾(n+1)

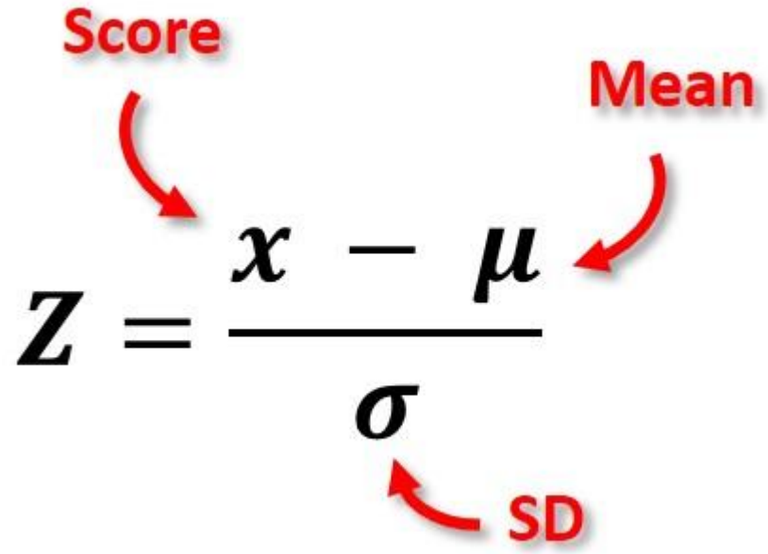
- 3, 12, 15, 16, 16, 17, 19, 34
- Min = 3, Q1 = 13.5, Med = 16, Q3 = 18, Max = 34
- 1.5 (IQR) Rule
  = 1.5( Q3 – Q1)
  = 1.5 (18 – 13.5)
  = 6.75
- Outliers
  - Lower Outliers = Q1 – 6.75 = 13.5 – 6.75 = 6.75
  - Upper Outliers = Q3 + 6.75 = 18 + 6.75 = 24.75

One of the mathematical ways to identify outliers.

# Z-Score

- Reassign the values based on mean and standard deviation

- The set of new values are either 0, or more or less than 0 (mean)




- Values more than +3 and less -3 can be considered as outliers

One of the mathematic ways to identify outliers.

# Normalization and Standardization

Normalization is between certain ranges It happens generally when you involve more than 1 feature . You bring them in certain ranges.

Xi_new= (Xi-Xmin)/(Xmax-Xmin)

| ID | Salary | Age |
|---|---|---|
| 1 | 10,000 | 20 |
| 2 | 20.000 | 10 |
| 3 | 30,000 | 30 |
| 4 | 50,000 | 20 |

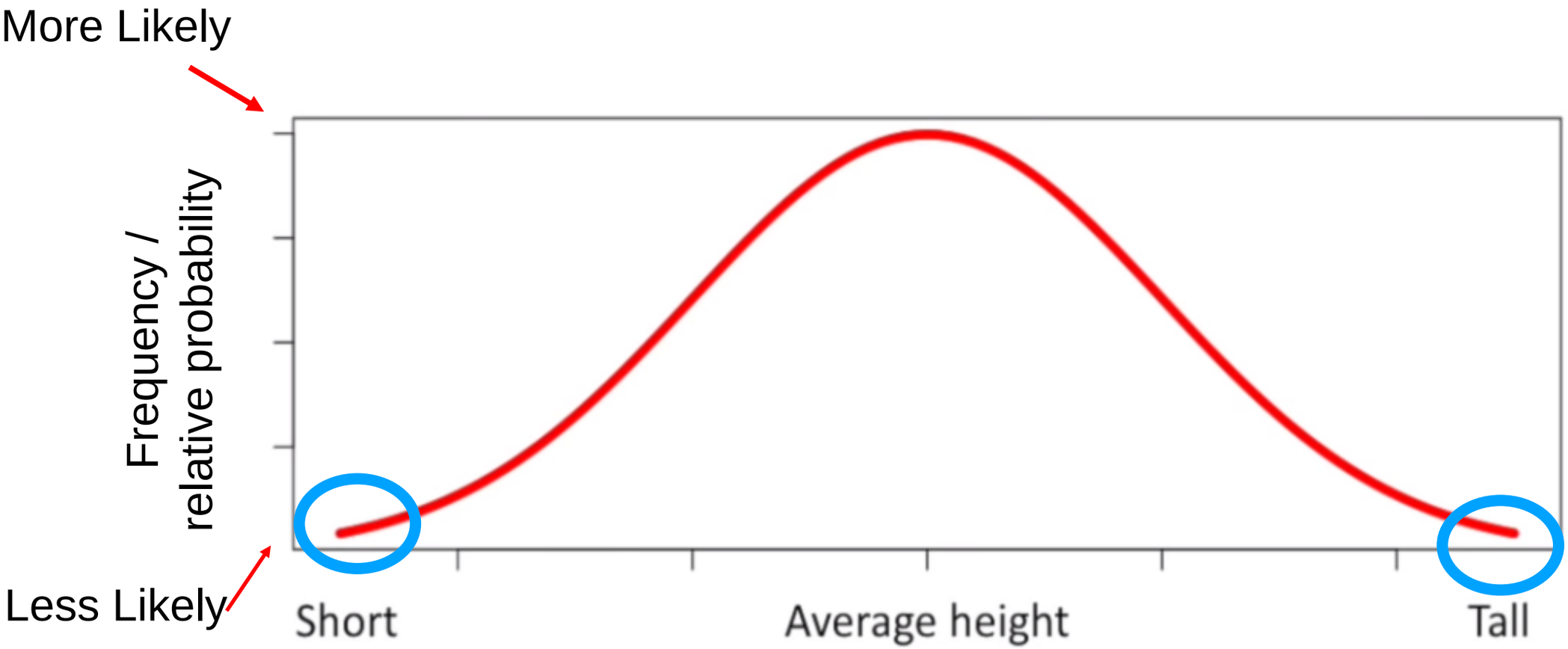
Standardization: You bring the data around a mean value.

Z-score

| ID | Salary | Age |
|---|---|---|
| 1 | 10,000 | 20 |
| 2 | 20.000 | 10 |
| 3 | 30,000 | 30 |
| 4 | 100,000 | 20 |

# Data Distribution

## Normal / Bell / Gaussian



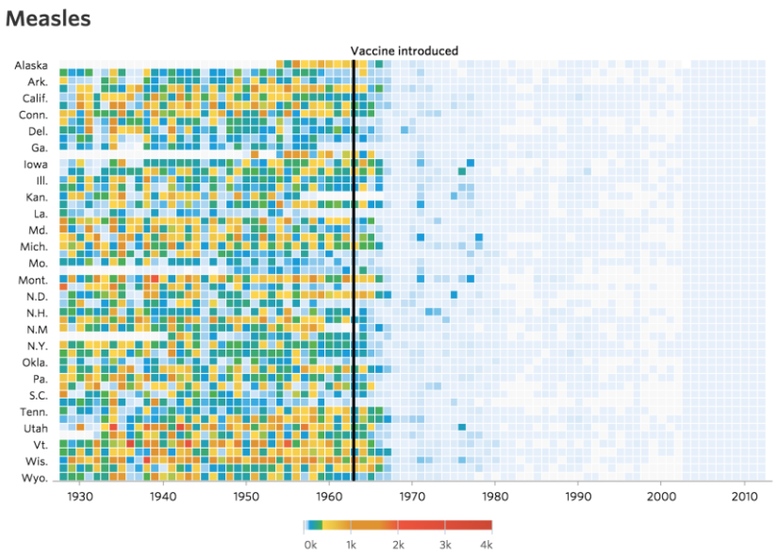

Human height measurements

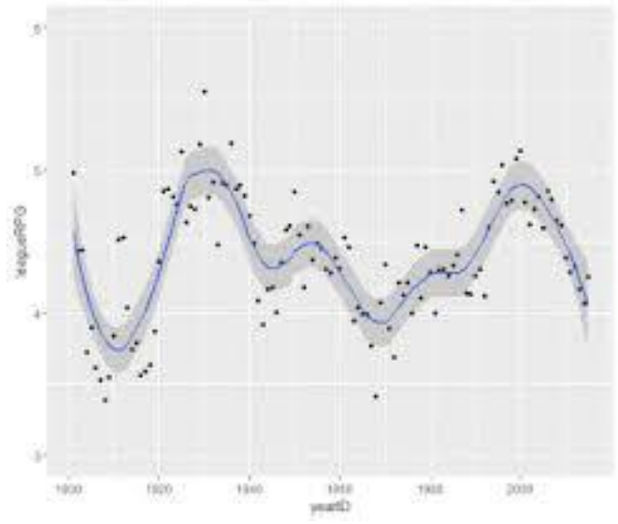
Data points possibly after the blue circle are outliers

# Visualization

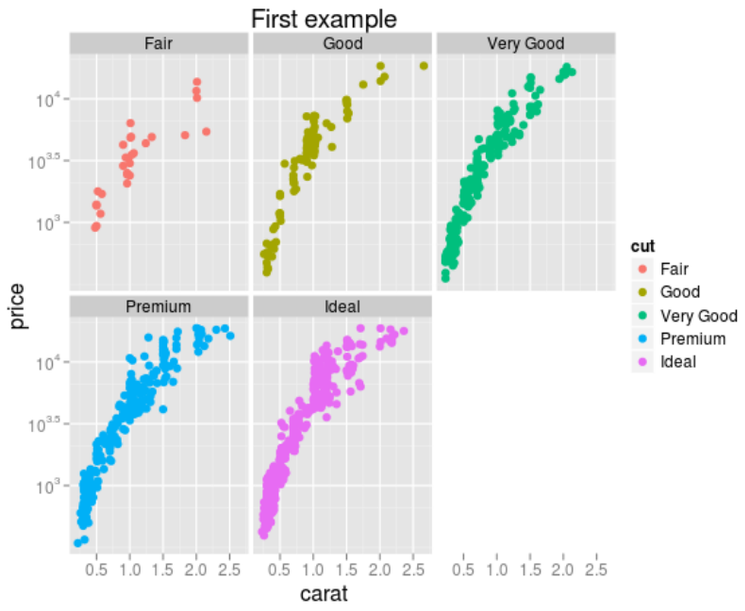
A visualization is a graphical representation designed to enable exploration, analysis, or communication

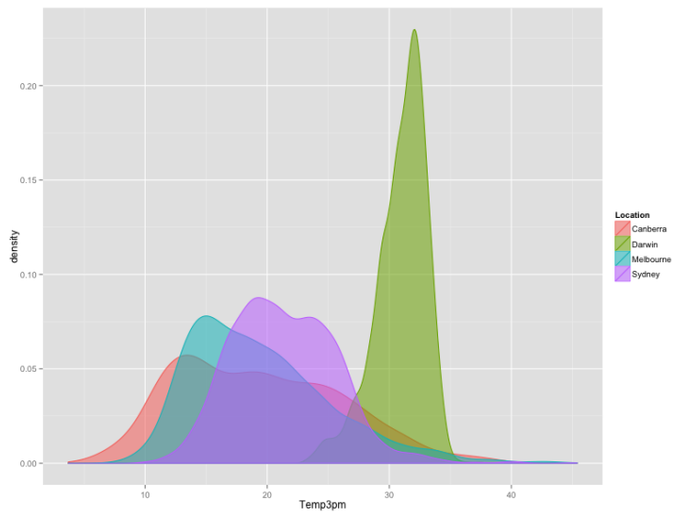
# The goal of the visualization



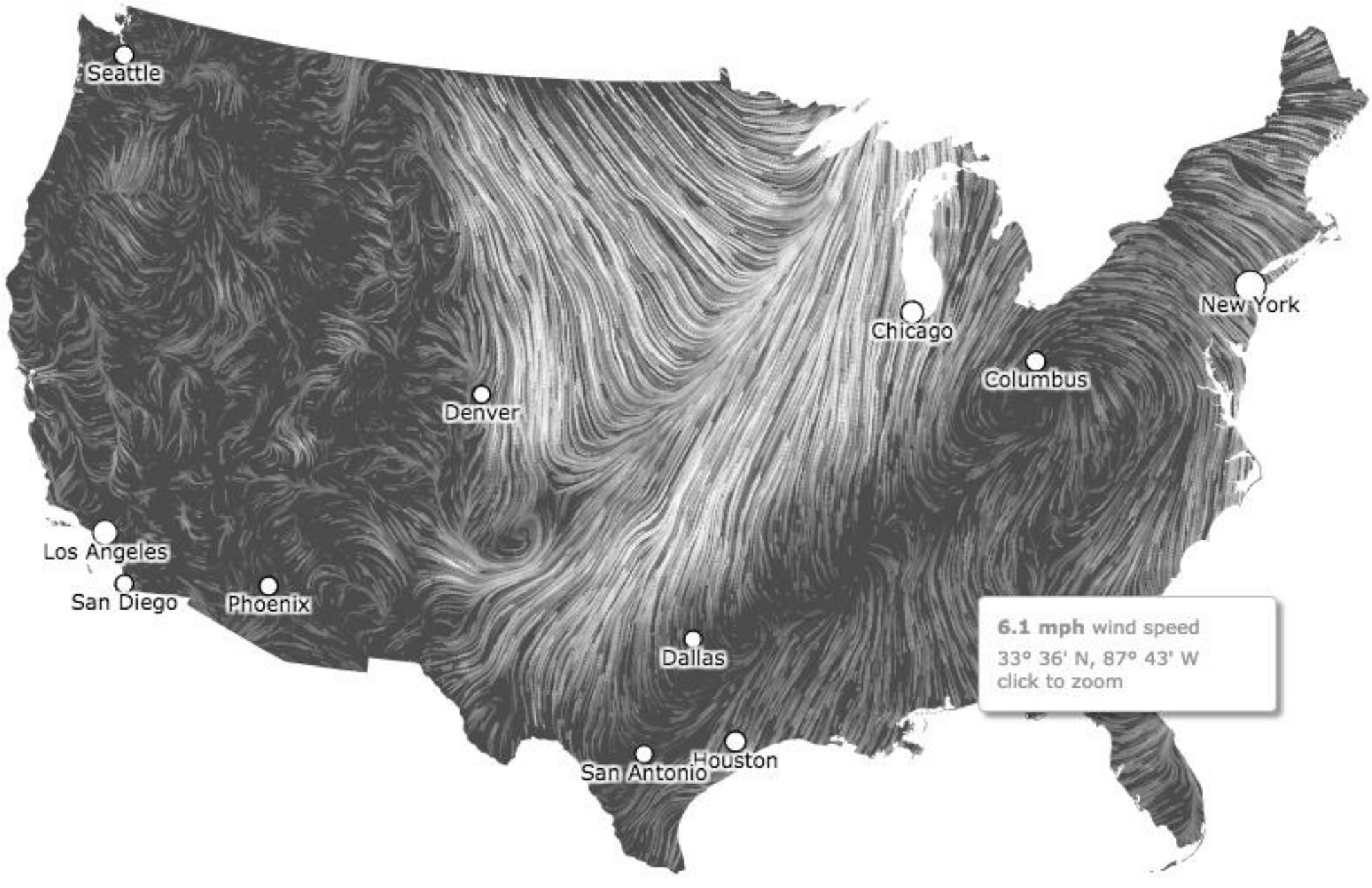
# Data exploration



# The goal of the visualization

## Conveying the message



CUT THROUGH B2B INFORMATION OVERLOAD

# *with* EFFECTIVE STORYTELLING

Infographics, whitepapers, 24-hour news reports — everyone in B2B is overloaded with information. Here's how **TELLING A STORY** can cut through the noise and prioritize what you want your B2B audience to care about

EVERYONE IS OVERLOADED WITH INFORMATION

HOW TO HELP YOUR CONTENT RISE ABOVE THE NOISE

## EVERYONE WANTS MULTIMEDIA

"SHOW, DON'T TELL"

Press releases that contain multimedia get...

**77%**

more responses compared to text-only releases

Blog posts with videos are linked to

**3X** more than text-only posts

IT'S CLASSIC STORYTELLING ADVICE AND TRUER THAN EVER IN THE AGE OF INFORMATION OVERLOAD

## NO ONE READS

MAKE SURE YOUR MESSAGE DOESN'T GET SKIMMED

**79%**

**scan the web** rather than reading word by word

**100,500**

The number of **digital words consumed** each day by an average US citizen

FRAME IT AS A STORY WITH A STRONG TITLE AND INTERESTING HOOK TO GET YOUR AUDIENCE THE INFORMATION YOU WANT THEM TO HAVE

BROUGHT TO YOU BY LOOKBOOKHQ AND Beutler Ink

## EVERYONE IS BUSY

Professionals spend...

**51%**

of their time **managing information** instead of acting on it

**90%**

of surveyed professionals admit to having thrown away important information **without reading it**

MANAGE INFORMATION FOR YOUR AUDIENCE

DON'T RELY ON THEM TO PIECE YOUR STORY TOGETHER FOR THEMSELVES—TELL THEM THE STORY YOU WANT THEM TO HEAR

## MARKETING IS NOISY

DON'T LET YOUR BRAND GET LOST IN MARKETING OVERLOAD

The **number of ads** served up on the Internet in 2012

**5.3 TRILLION**

It's estimated that people see:

**5,000**

marketing messages per day

MAKE IT MEMORABLE THE SAME WAY WRITERS MAKE THEIR CHARACTERS MEMORABLE—TELL A STORY ABOUT WHO YOU ARE

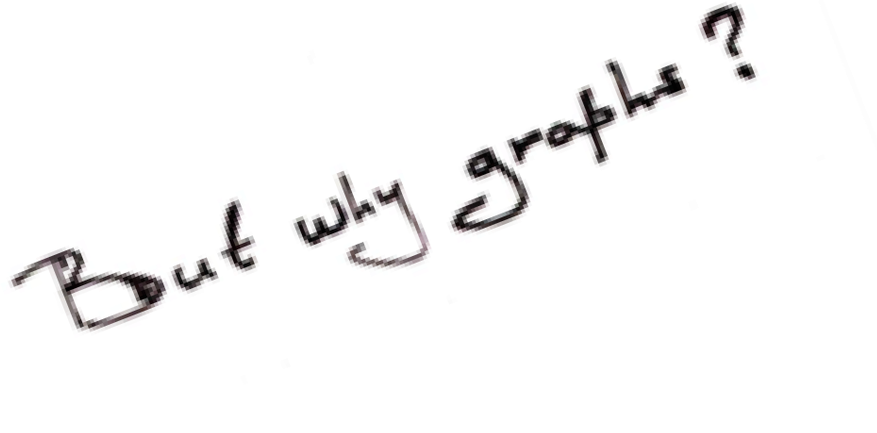
THE PROBLEMS OF INFORMATION OVERLOAD IN MAKING YOUR VOICE HEARD ARE MANY,

BUT THEIR SOLUTION IS SIMPLE - YOUR STORY, TOLD BY YOU

| | x | y |
|---|---|---|
| 1: | 23.0769 | 70.3125 |
| 2: | 24.3590 | 81.0817 |
| 3: | 26.9231 | 90.3125 |
| 4: | 29.7436 | 86.8510 |
| 5: | 31.5385 | 82.2356 |
| 6: | 34.3590 | 76.8510 |
| 7: | 38.9744 | 77.6202 |
| 8: | 42.8205 | 79.5433 |
| 9: | 22.3077 | 63.3894 |
| 10: | 22.0513 | 53.3894 |
| 11: | 24.6154 | 47.2356 |
| 12: | 28.7179 | 41.4663 |

<truncated>

| | | |
|---|---|---|
| 68: | 21.2821 | 46.4663 |
| 69: | 27.1795 | 48.7740 |
| 70: | 31.0256 | 49.1587 |
| 71: | 35.1282 | 49.5433 |
| 72: | 40.2564 | 51.4663 |
| 73: | 45.8974 | 53.0048 |



```
          x       y
 1: 23.0769 70.3125
 2: 24.3590 81.0817
 3: 26.9231 90.3125
 4: 29.7436 86.8510
 5: 31.5385 82.2356
 6: 34.3590 76.8510
 7: 38.9744 77.6202
 8: 42.8205 79.5433
 9: 22.3077 63.3894
10: 22.0513 53.3894
11: 24.6154 47.2356
12: 28.7179 41.4663
    <truncated>

68: 21.2821 46.4663
69: 27.1795 48.7740
70: 31.0256 49.1587
71: 35.1282 49.5433
72: 40.2564 51.4663
73: 45.8974 53.0048
```

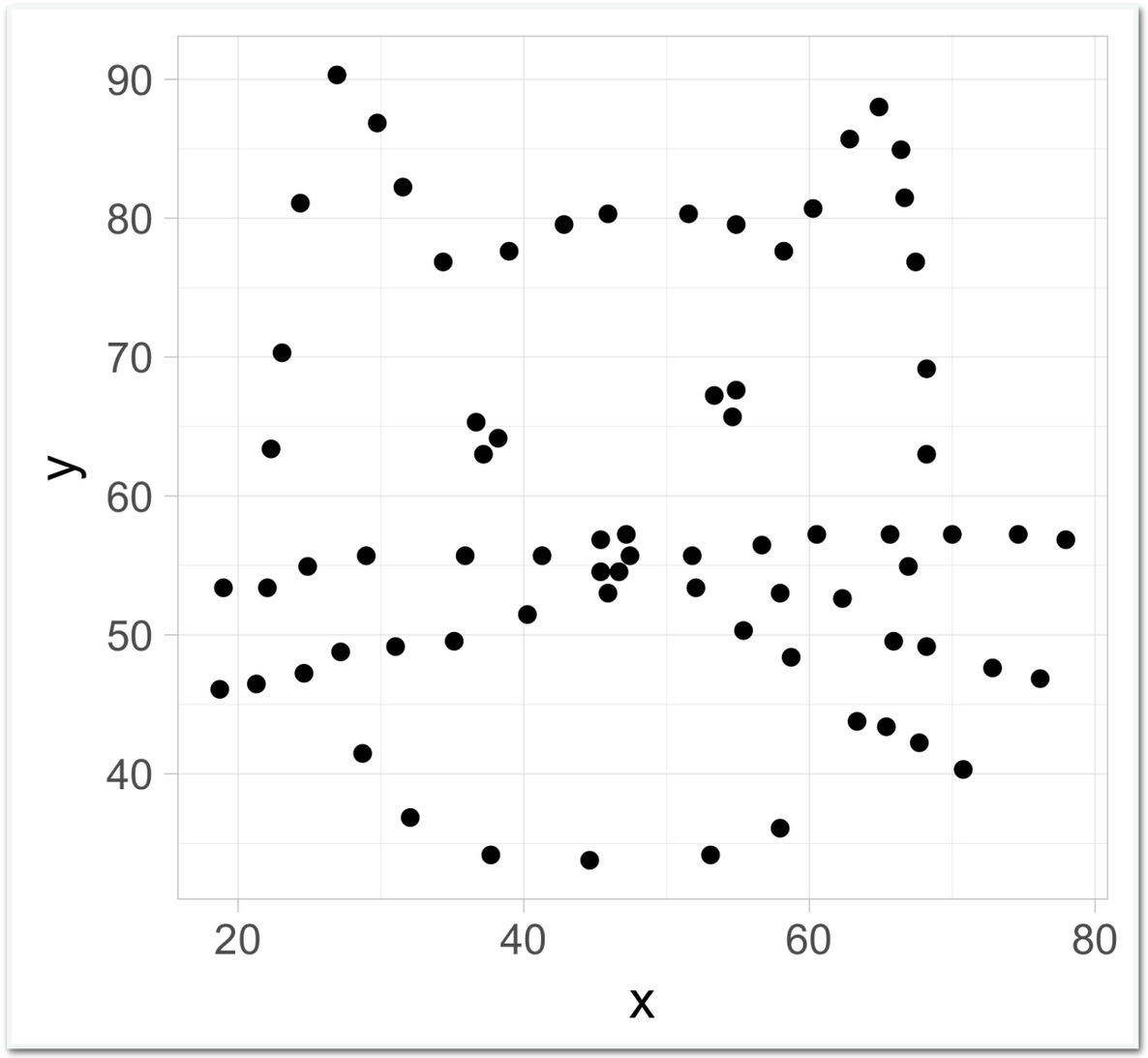
```
       x               y        
 Min.   :18.72   Min.   :33.77  
 1st Qu.:35.13   1st Qu.:49.16  
 Median :51.54   Median :55.70  
 Mean   :48.60   Mean   :59.43  
 3rd Qu.:63.33   3rd Qu.:69.16  
 Max.   :77.95   Max.   :90.31  
```

```
> cor(dt$x, dt$y)
[1] -0.005949079
```

| | x | y |
|---|---|---|
| 1: | 23.0769 | 70.3125 |
| 2: | 24.3590 | 81.0817 |
| 3: | 26.9231 | 90.3125 |
| 4: | 29.7436 | 86.8510 |
| 5: | 31.5385 | 82.2356 |
| 6: | 34.3590 | 76.8510 |
| 7: | 38.9744 | 77.6202 |
| 8: | 42.8205 | 79.5433 |
| 9: | 22.3077 | 63.3894 |
| 10: | 22.0513 | 53.3894 |
| 11: | 24.6154 | 47.2356 |
| 12: | 28.7179 | 41.4663 |

<truncated>

| | | |
|---|---|---|
| 68: | 21.2821 | 46.4663 |
| 69: | 27.1795 | 48.7740 |
| 70: | 31.0256 | 49.1587 |
| 71: | 35.1282 | 49.5433 |
| 72: | 40.2564 | 51.4663 |
| 73: | 45.8974 | 53.0048 |




http://robertgrantstats.co.uk/drawmydata.html

# Visualization ABCs

bar chart

multi-set bar chart

histogram

density plot

boxplot

line chart

scatter plot

network/graph

http://www.datavizcatalogue.com/

# Dataset

A transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

The company mainly sells unique all-occasion gifts.

Many customers of the company are wholesalers.

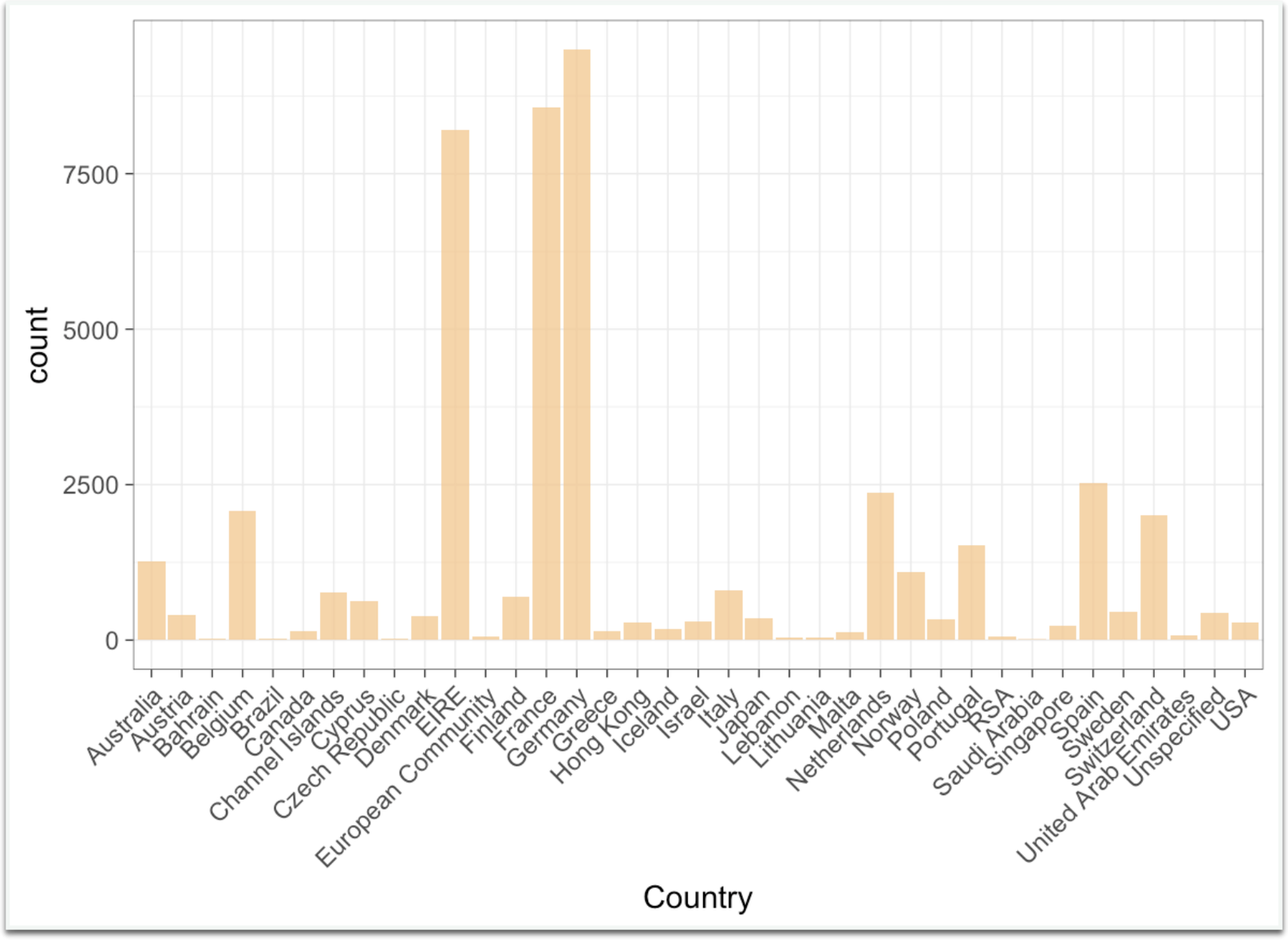
# Dataset

Head command

| | InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Country |
|---|---|---|---|---|---|---|---|---|
| 1: | 536365 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 01/12/10 08:26 | 2.55 | 17850 | United Kingdom |
| 2: | 536365 | 71053 | WHITE METAL LANTERN | 6 | 01/12/10 08:26 | 3.39 | 17850 | United Kingdom |
| 3: | 536365 | 84406B | CREAM CUPID HEARTS COAT HANGER | 8 | 01/12/10 08:26 | 2.75 | 17850 | United Kingdom |
| 4: | 536365 | 84029G | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 01/12/10 08:26 | 3.39 | 17850 | United Kingdom |
| 5: | 536365 | 84029E | RED WOOLLY HOTTIE WHITE HEART. | 6 | 01/12/10 08:26 | 3.39 | 17850 | United Kingdom |
| 6: | 536365 | 22752 | SET 7 BABUSHKA NESTING BOXES | 2 | 01/12/10 08:26 | 7.65 | 17850 | United Kingdom |

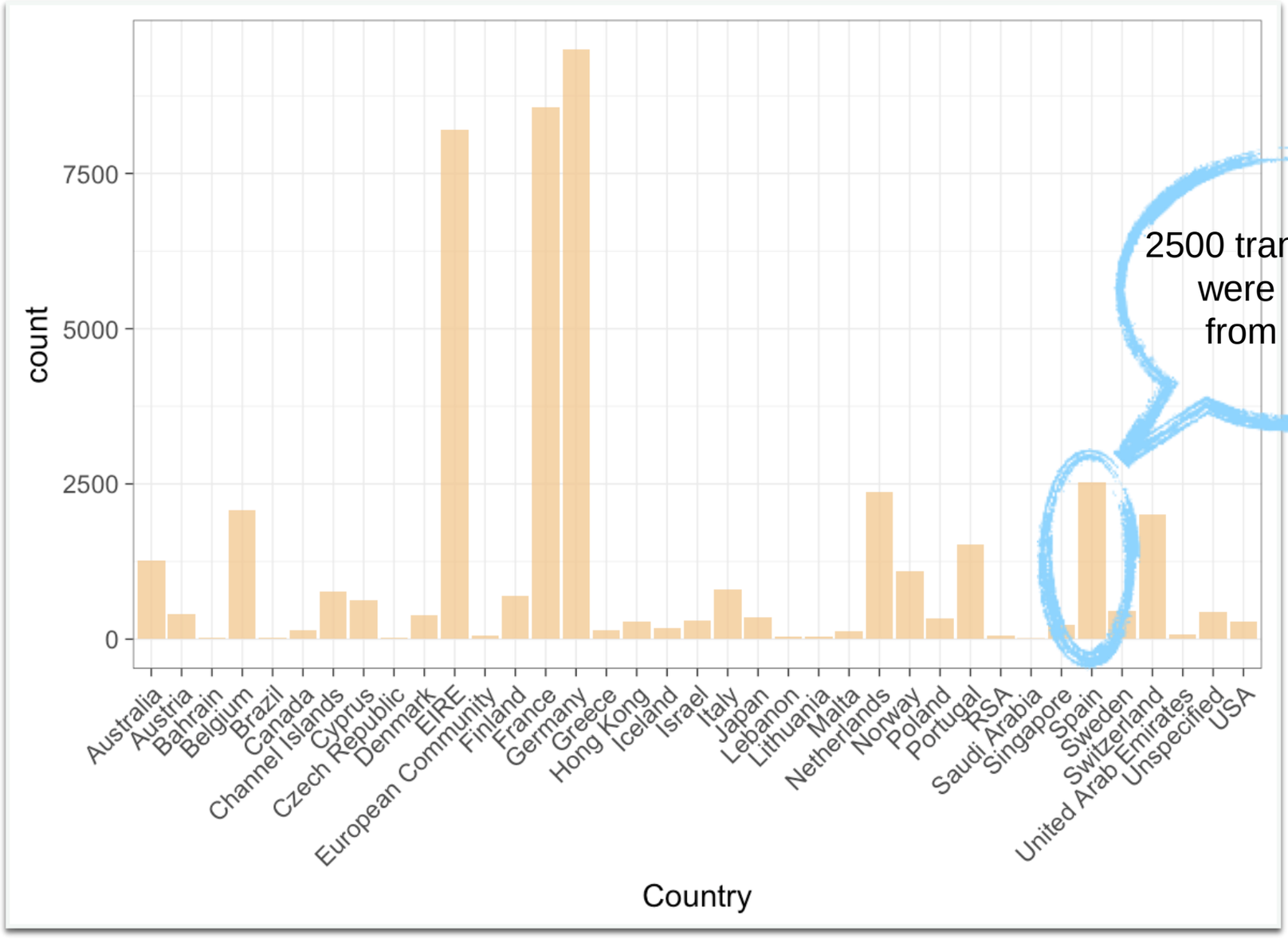
# Bar chart

Description of one discrete feature that displays counts



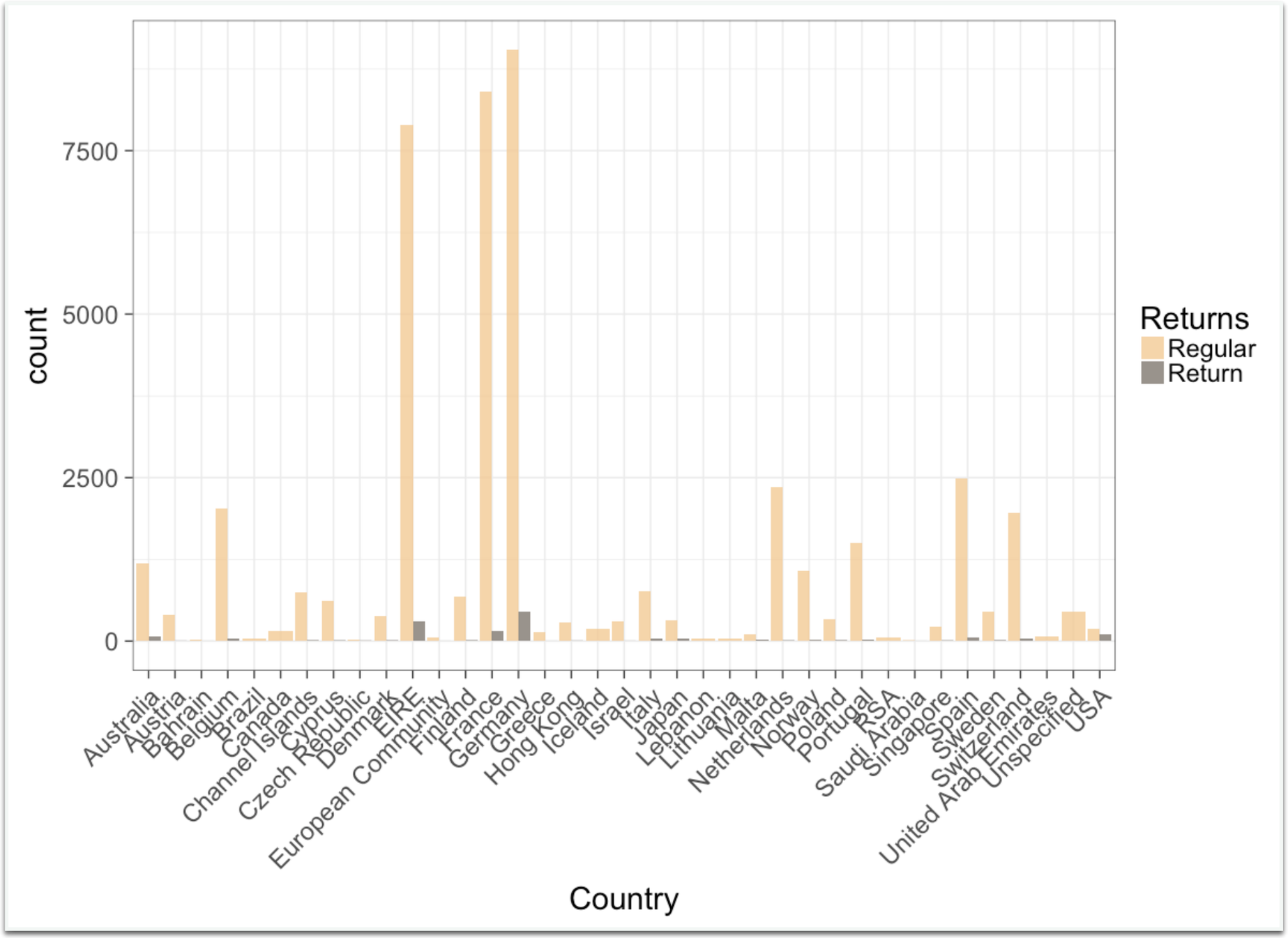
# Bar chart

Description of one discrete feature that displays counts



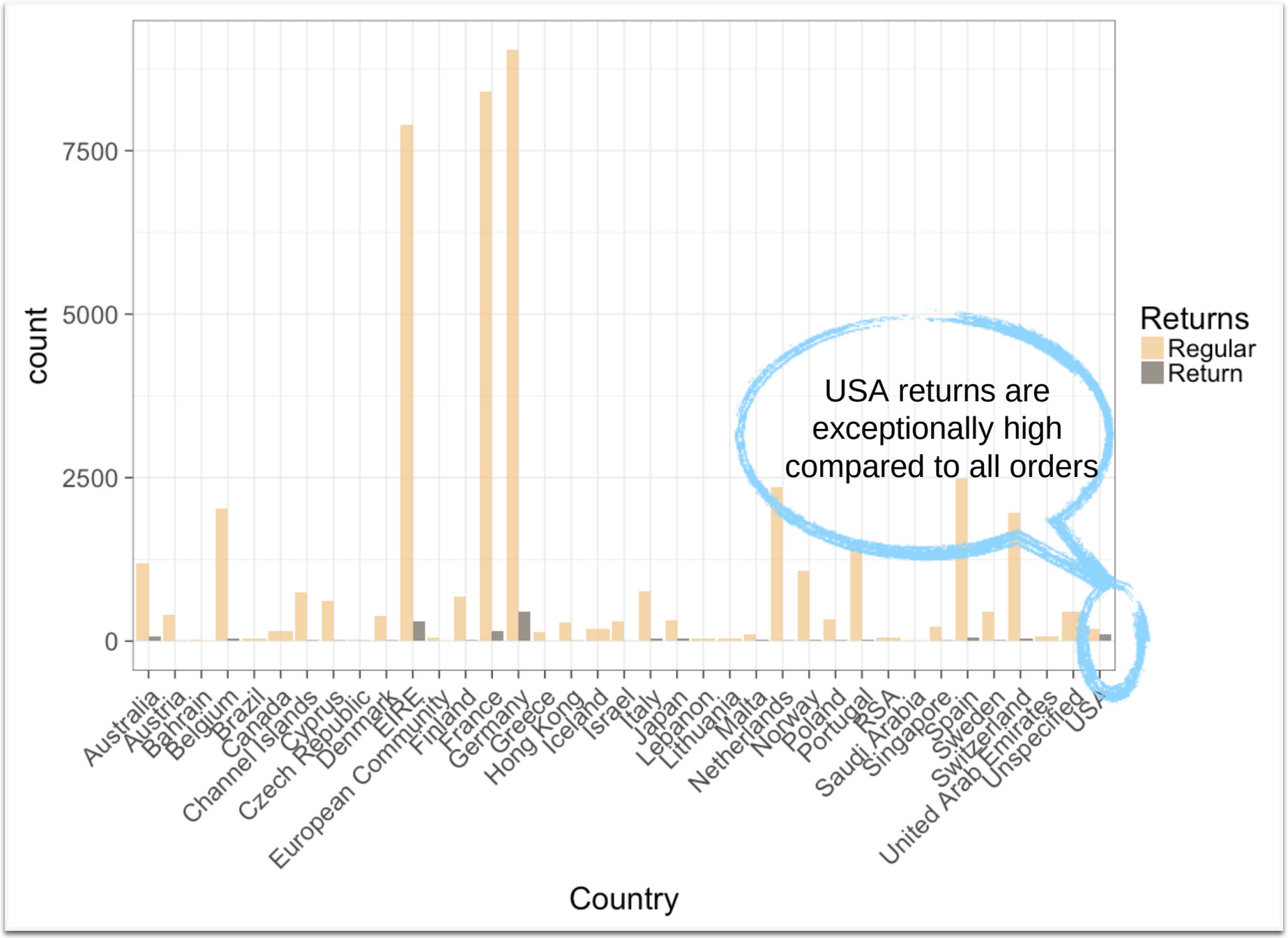
# Multi-set bar chart

Description of two discrete features that displays counts



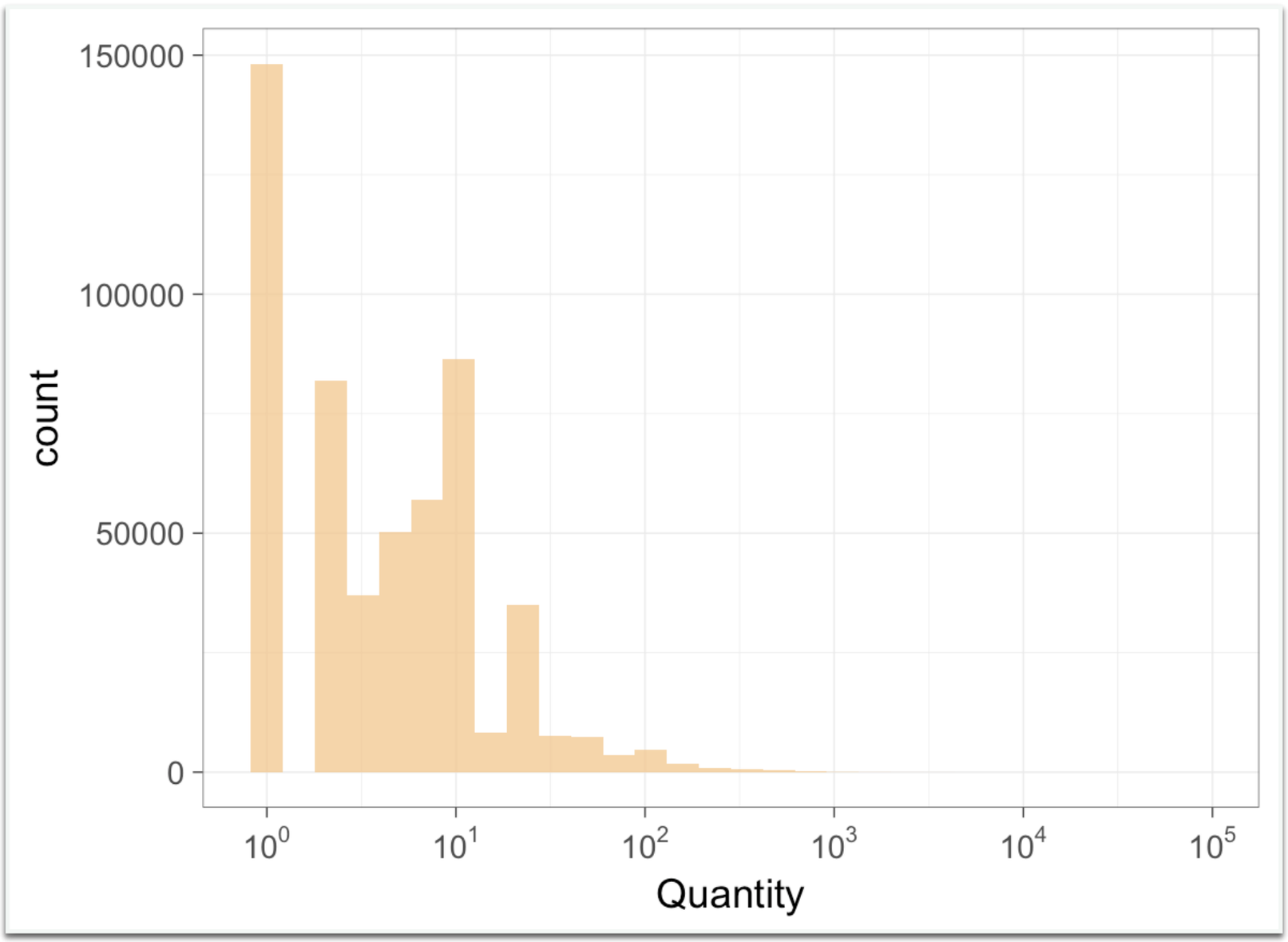
# Multi-set bar chart

Description of two discrete features that displays counts



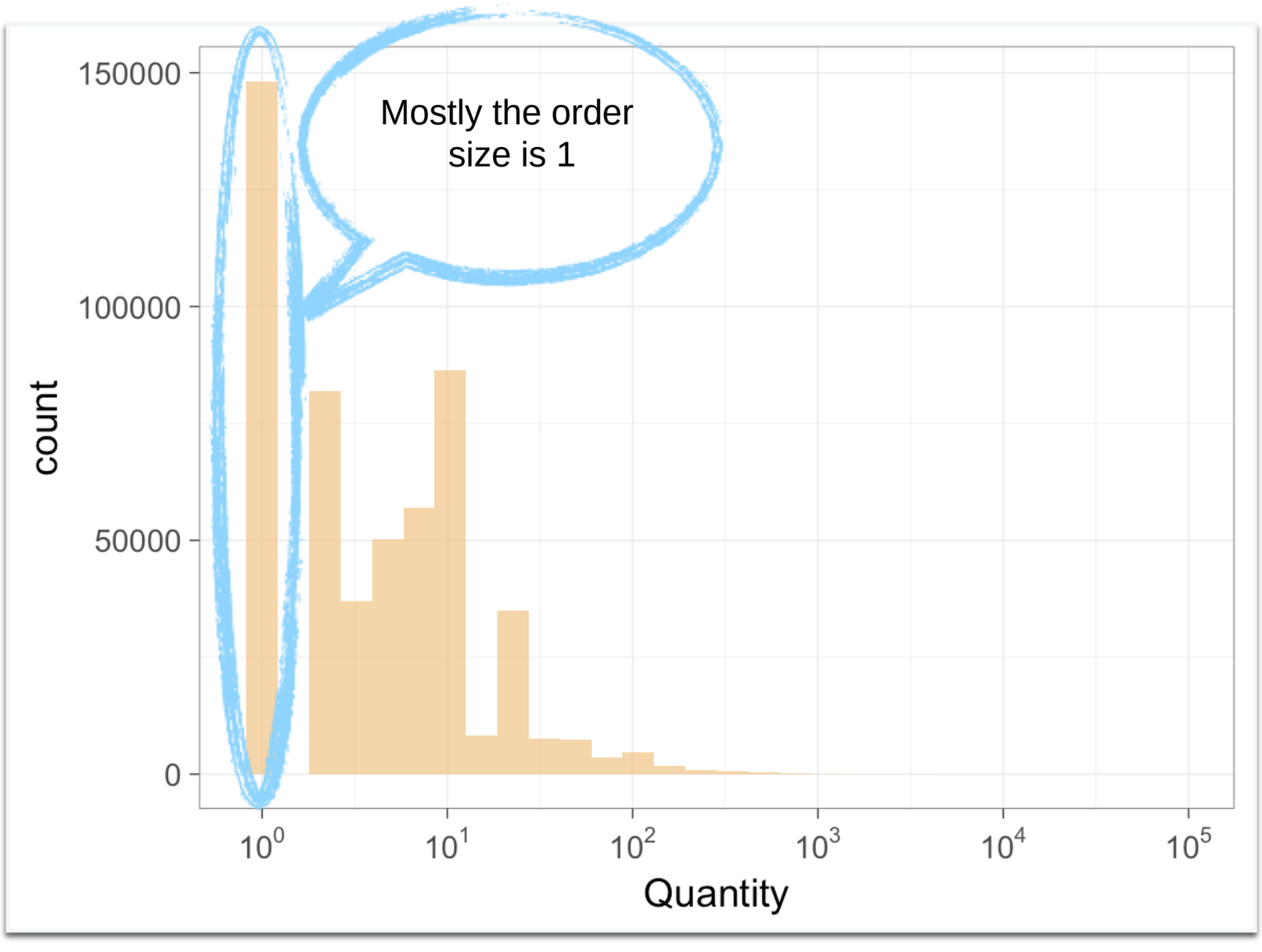
# Histogram

Description of one continuous feature. Displays general distribution



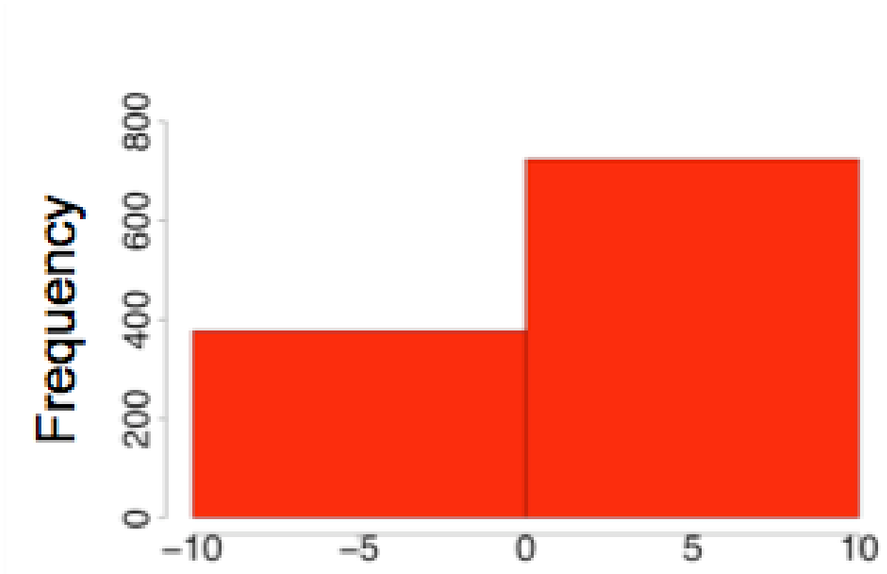
# Histogram

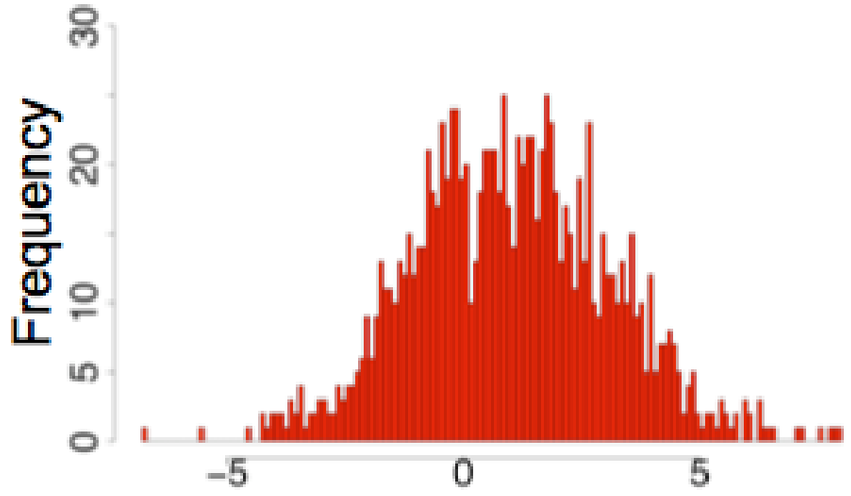
Description of one continuous feature. Displays general distribution

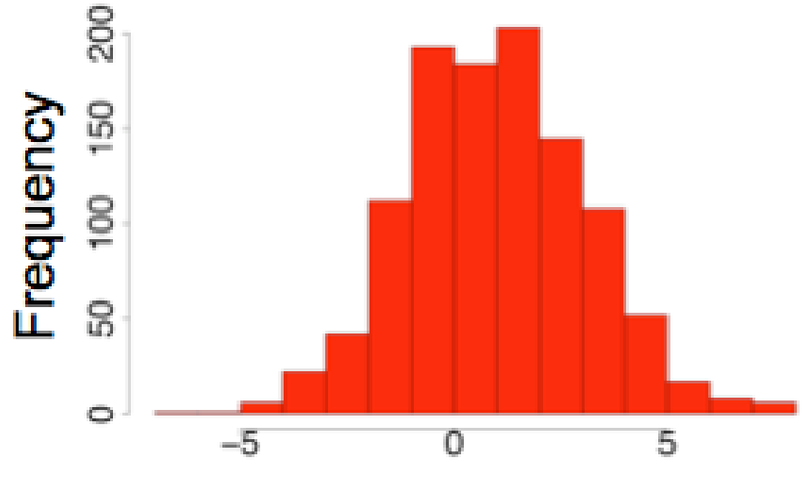


# Histogram

Effect of a bin size on histogram

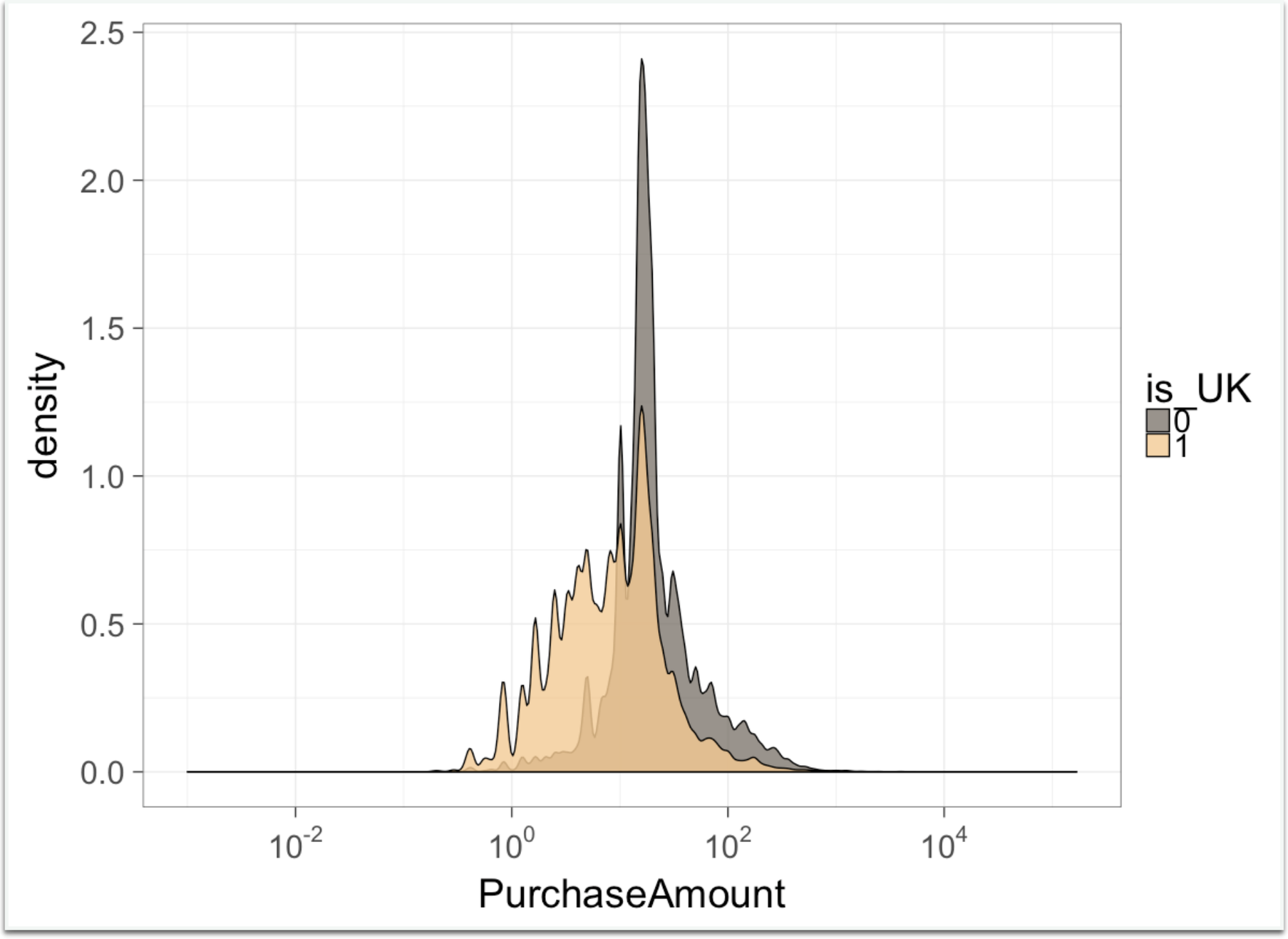


# Density

Description of one continuous feature. Displays smoothed general distribution



# Box plot example (different dataset)

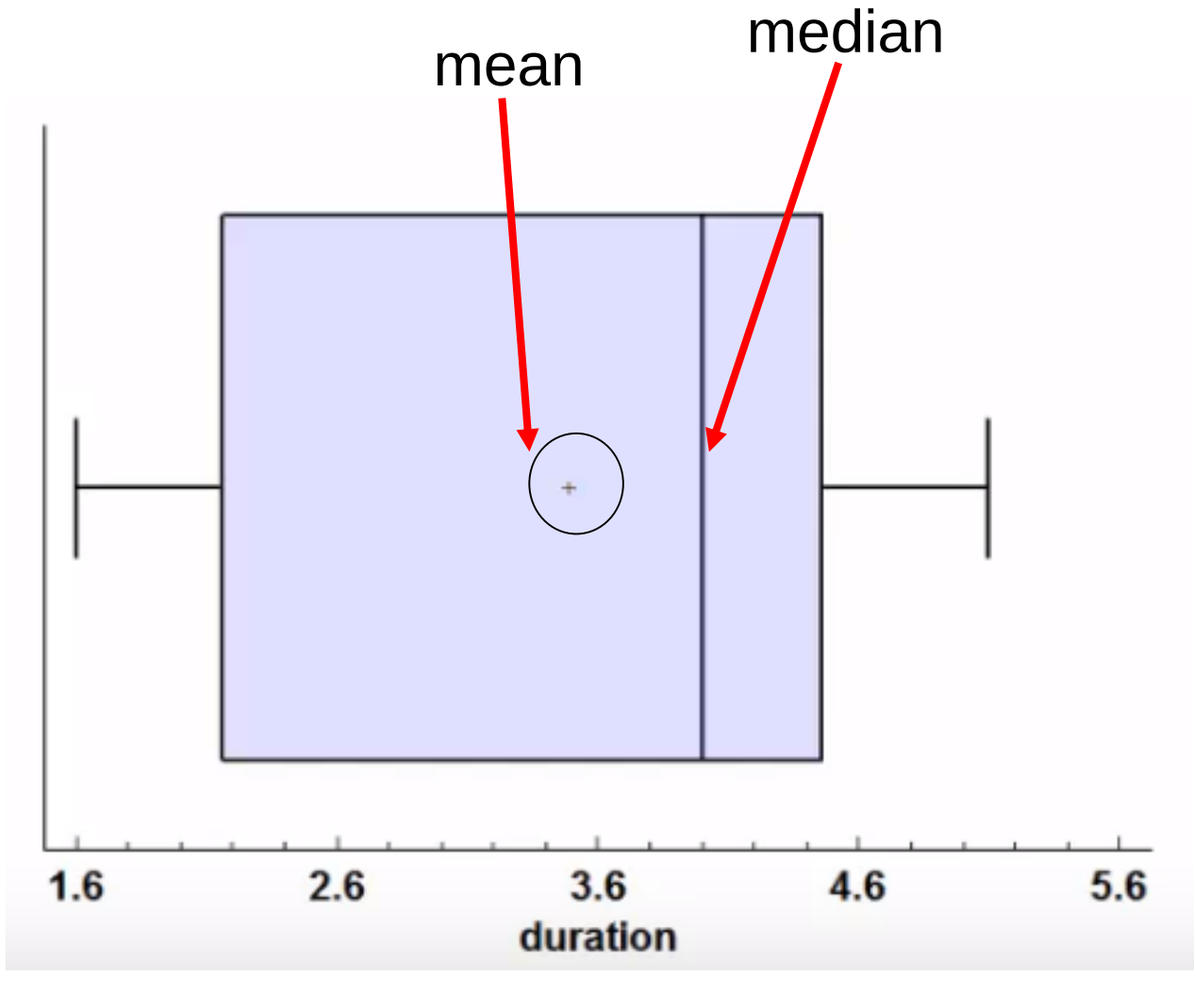
data.head()

the length of the waiting period until the next one (in mins)

| | eruptions | waiting |
|---|---|---|
| 1 | 3.600 | 79 |
| 2 | 1.800 | 54 |
| 3 | 3.333 | 74 |
| 4 | 2.283 | 62 |
| 5 | 4.533 | 85 |
| 6 | 2.883 | 55 |

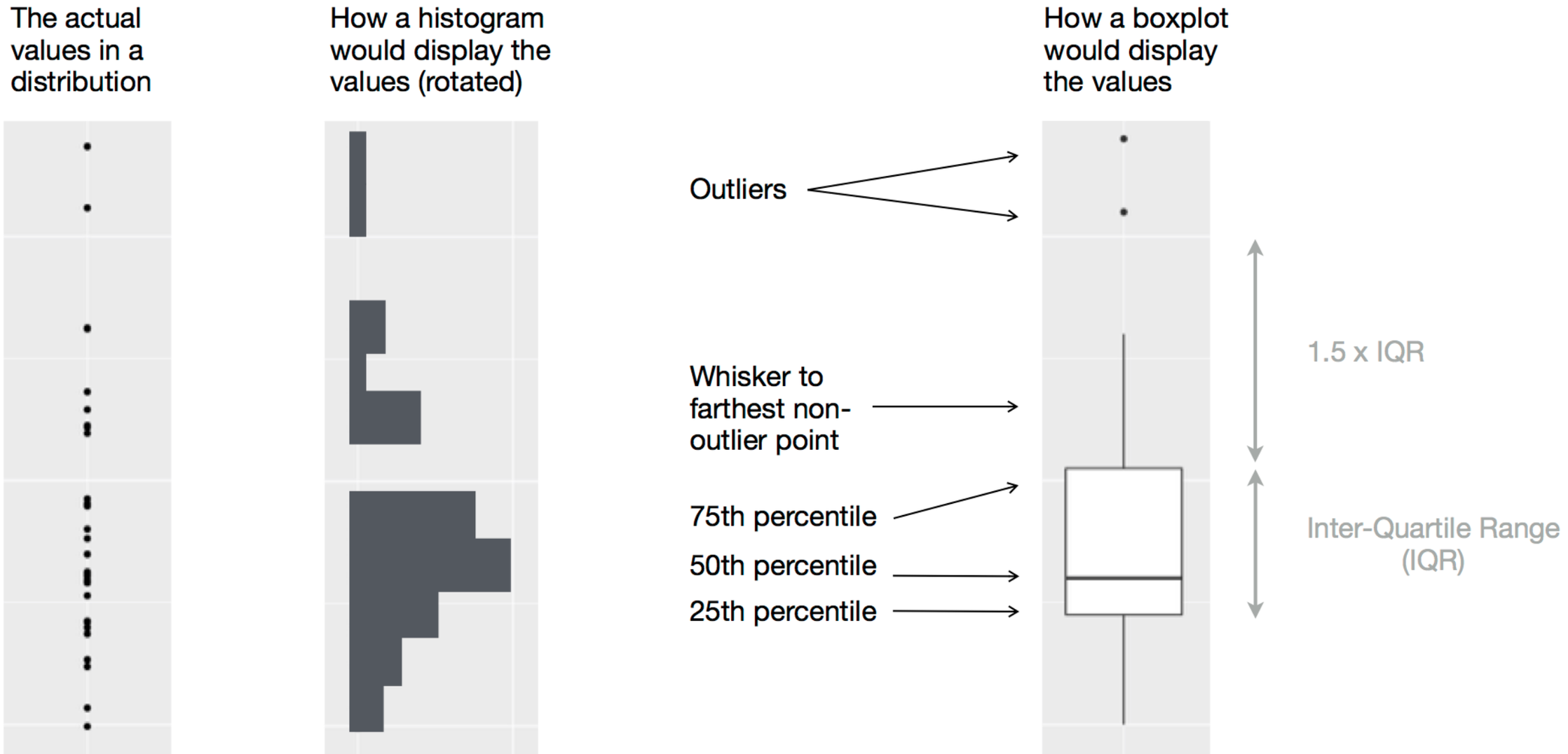
the duration of the geyser eruptions (in mins)

data.shape()
[272 2



# Boxplot

## Also knows as box and whisker plot




One of the graphical way to identify outliers.

source: Alberto Cairo

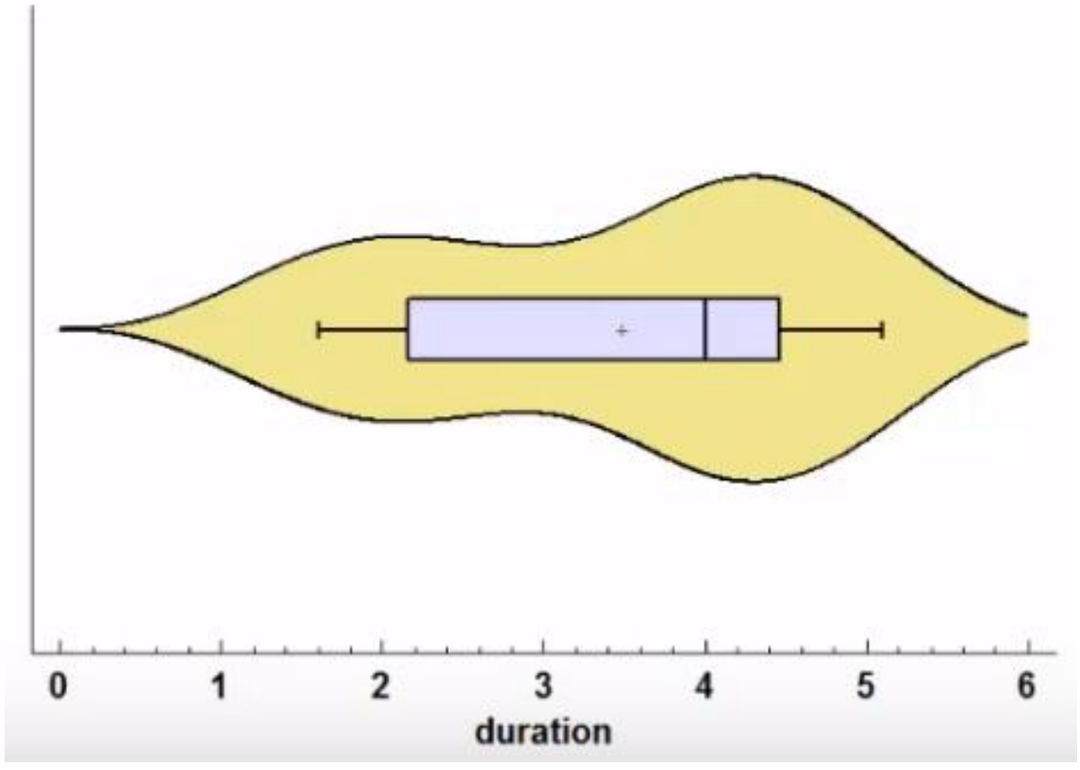
# Violin Plots

**Violin plots:  Combines Boxplot with a nonparametric density parameter (**probability density of the data at different values  -- in the simplest case this could be a histogram).

Above and below the box plot:
Estimates of the density function of the observations.

They are Created using non parametric density estimator (smooths out the probabilities using an interval of particular bandwidth)

If you change the bandwidth, it change how much smoothing is applied to the density function.



# Violin Plots

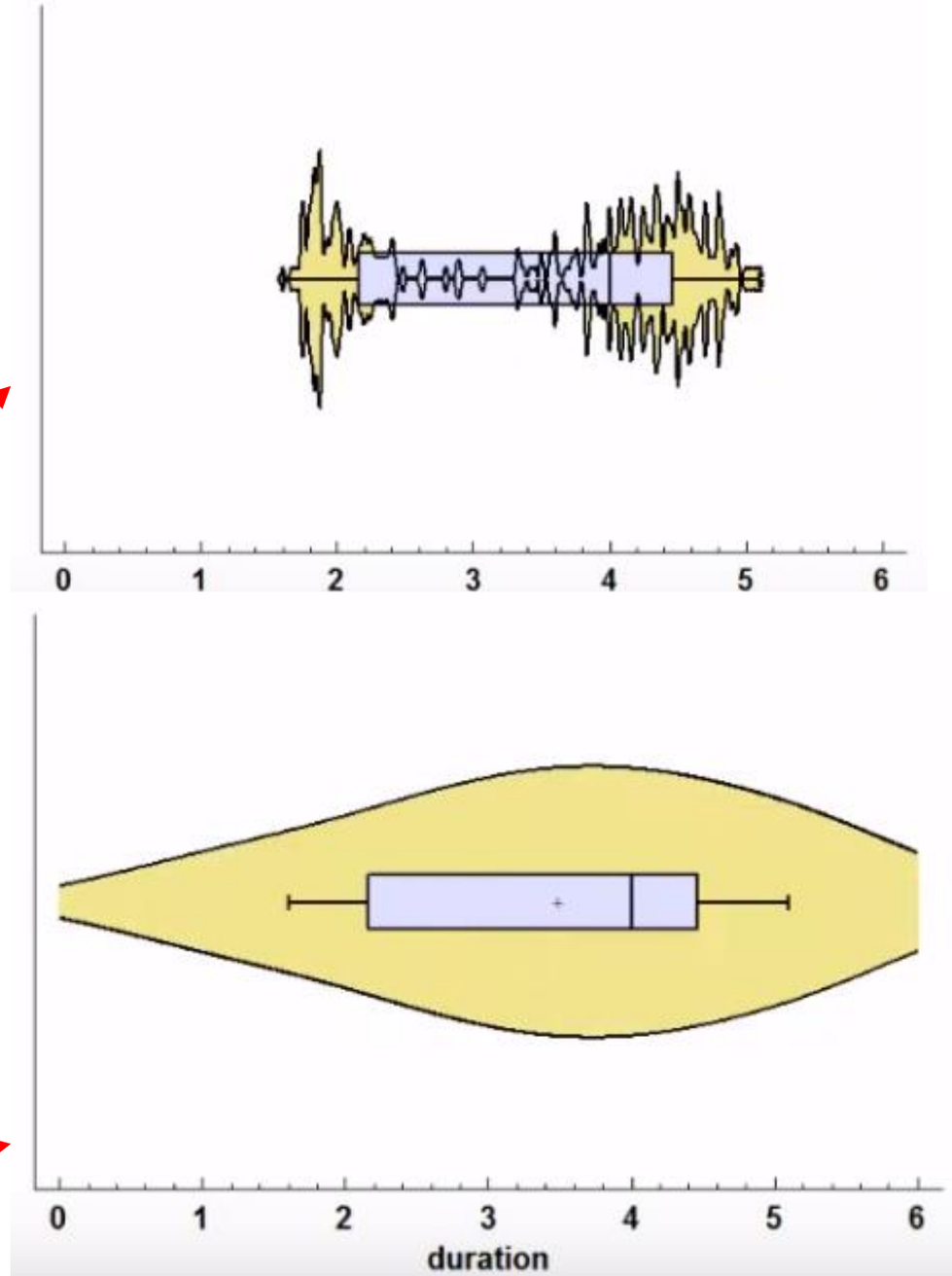
**Violin plots: Combines Boxplot with a nonparametric density parameter (**probability density of the data at different values -- in the simplest case this could be a histogram).

Above and below the box plot:
Estimates of the density function of the observations.

They are Created using non parametric density estimator (smooths out the probabilities using an interval of particular bandwidth)

If you change the bandwidth, it change how much smoothing is applied to the density function.

Small bandwidth: too much detail, nothing useful
Large bandwidth: might miss interesting aspects of the data.



# Violin Plots

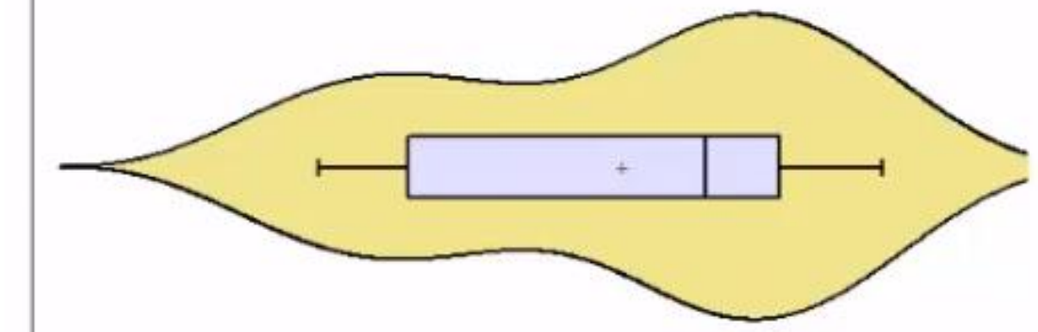
**Violin plots: Combines Boxplot with a nonparametric density parameter (**probability density of the data at different values -- in the simplest case this could be a histogram).

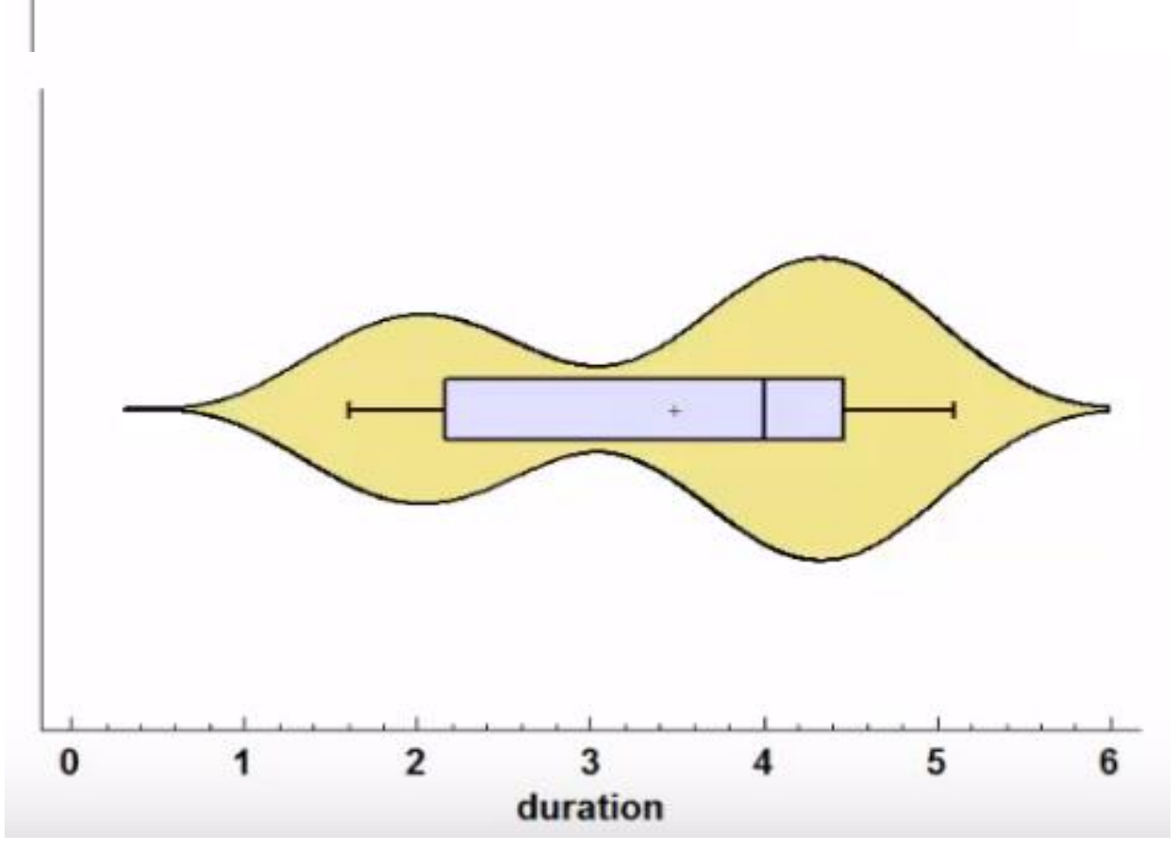
Above and below the box plot:
Estimates of the density function of the observations.

They are Created using non parametric density estimator (smooths out the probabilities using an interval of particular bandwidth)

If you change the bandwidth, it change how much smoothing is applied to the density function.

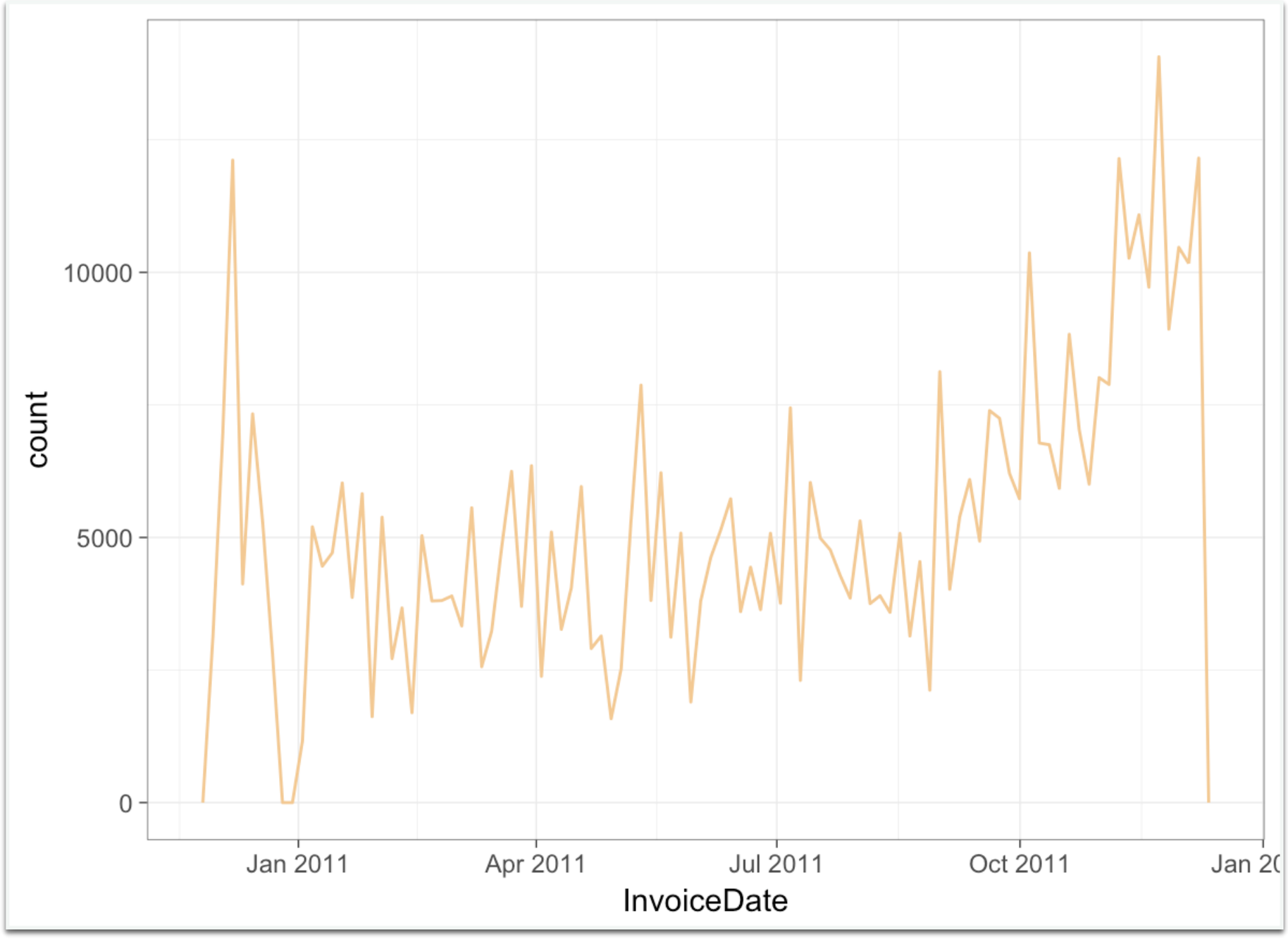
Select a bandwidth: shows most important features of data sample.
Eruptions: Bimodal.





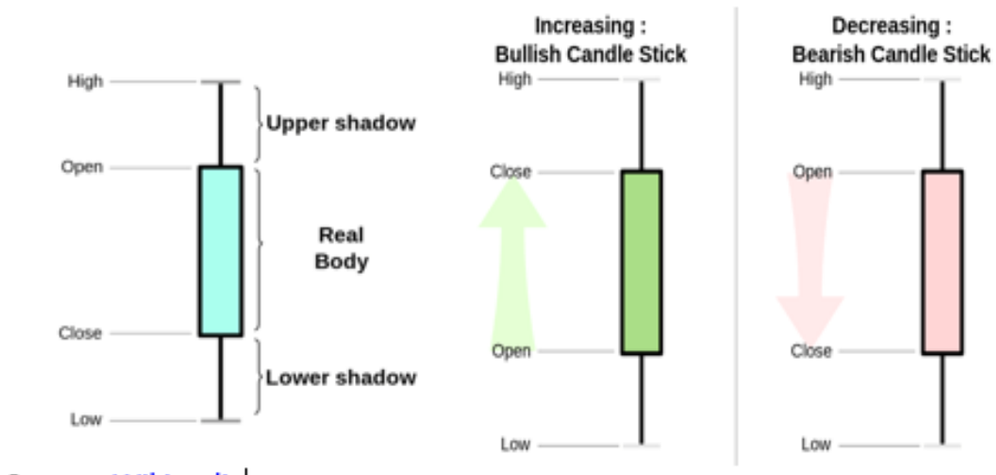
# Line chart

2 continuous features. Usually time series



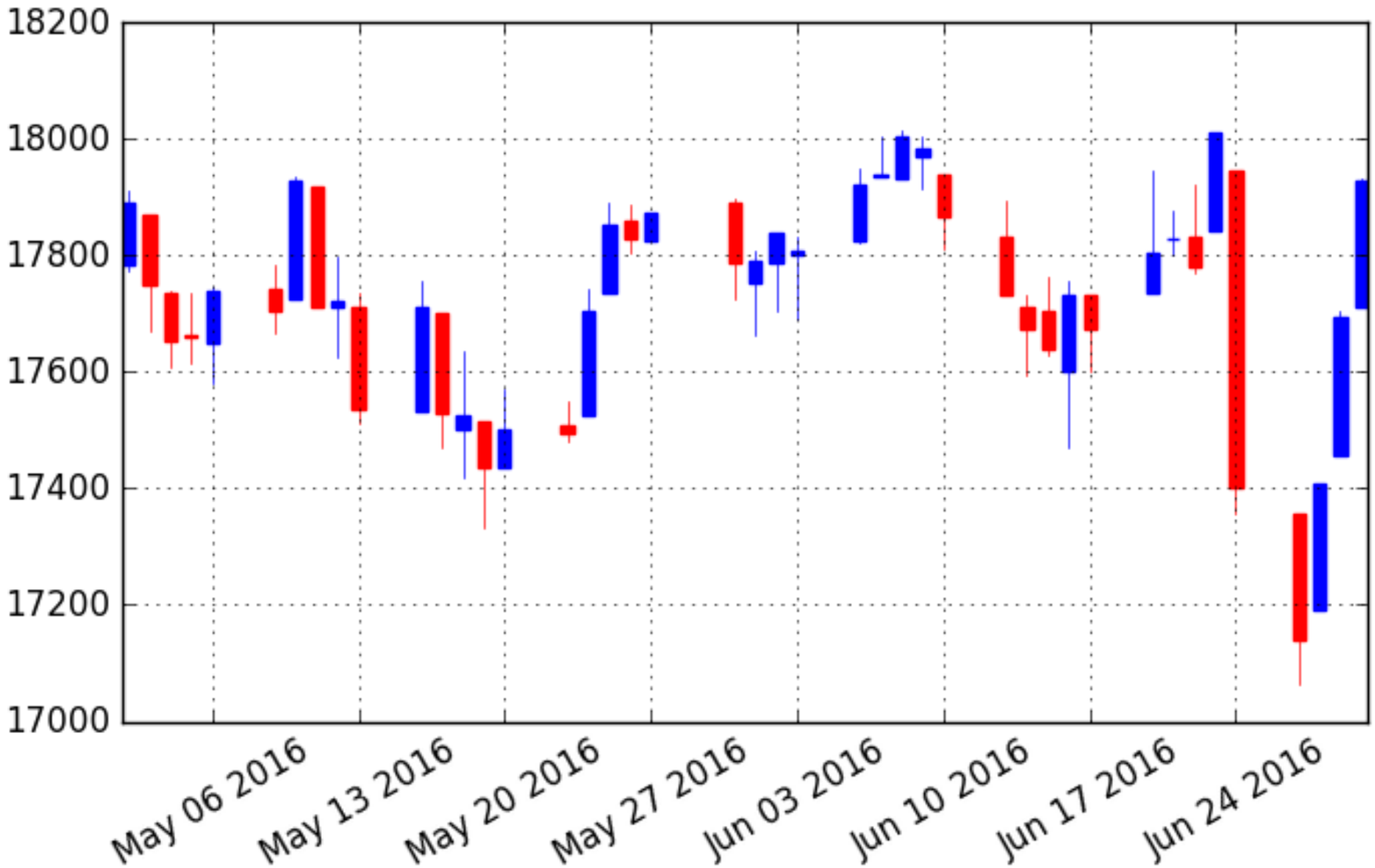
# Candlestick plots

## (Entries, Exits, Risk Management)



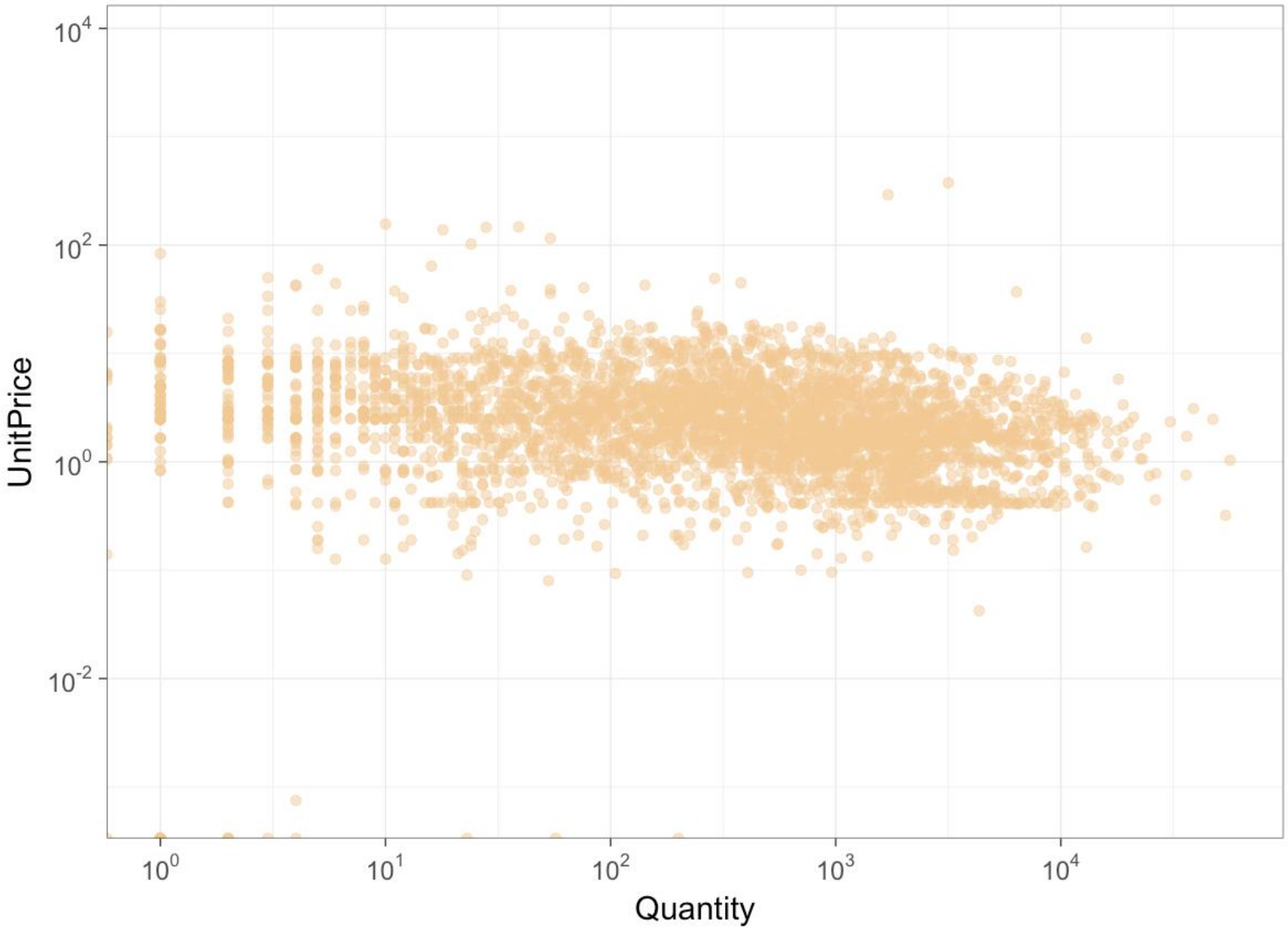

Source: Wikipedia

# Stock price data using candle stick plot



# Scatter plot

Relationship (Correlation) of 2 continuous features.

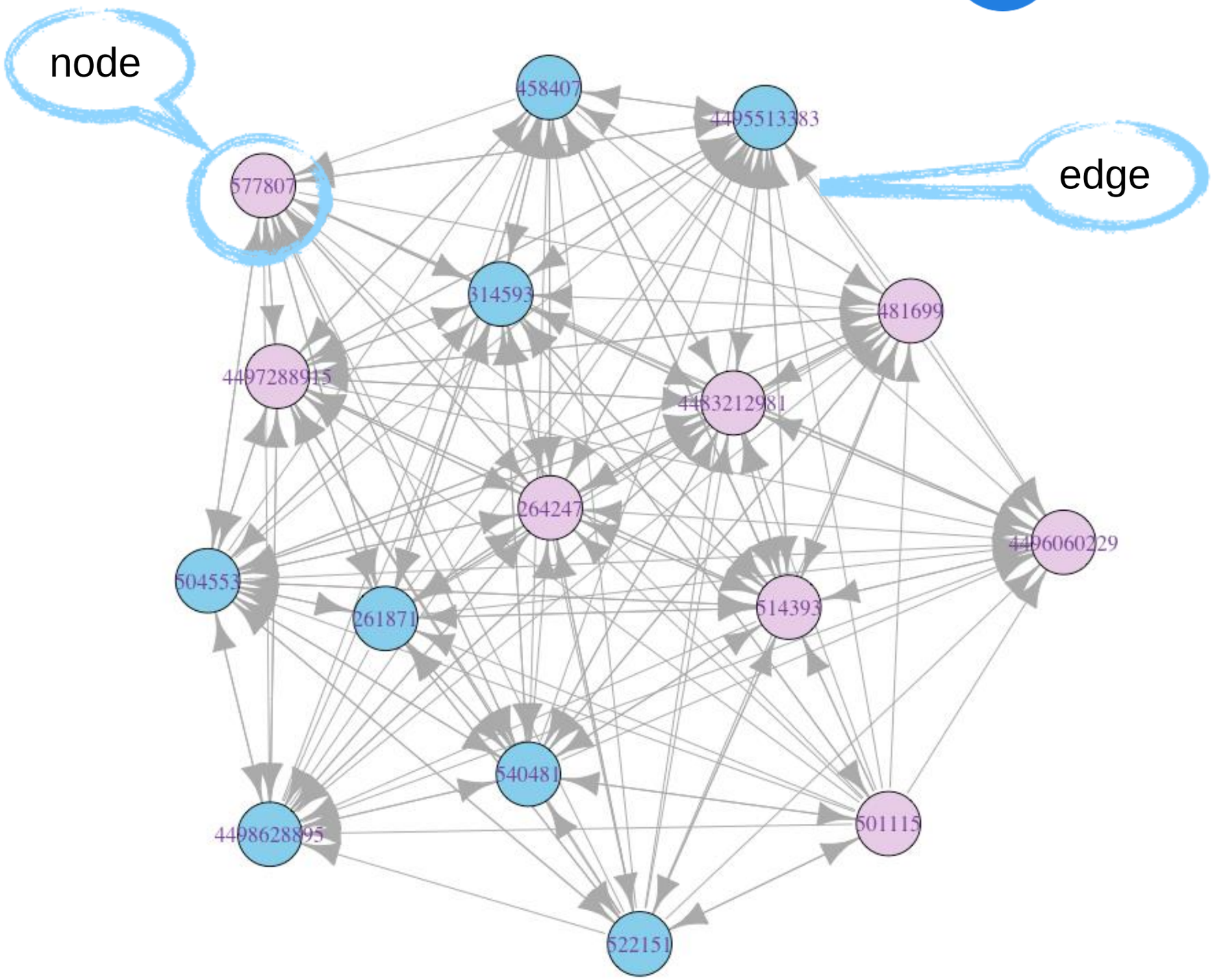



One of the graphical way to identify outliers.

# Scatter plot if used for correlation

Correlation tells you how two variables are correlated ?
Positive and Negative correlation

# Network

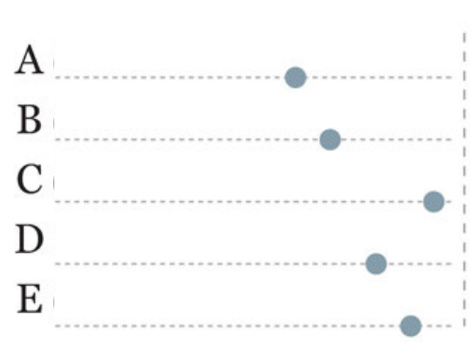



New course: Network Science, Jan 2020

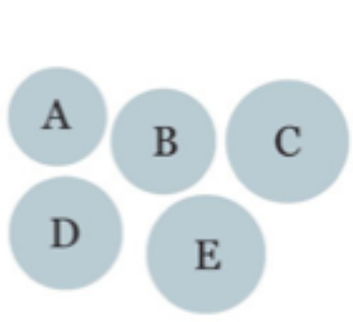
# Tip1: What is the best way to understand correctly the differences  without reading the numbers?

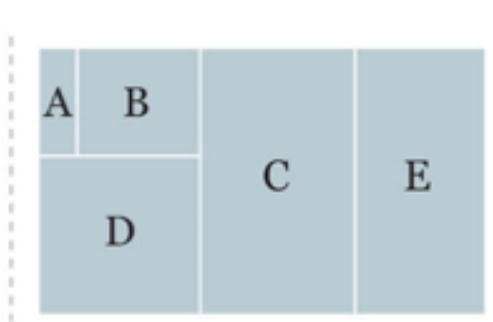
Length or height


Position

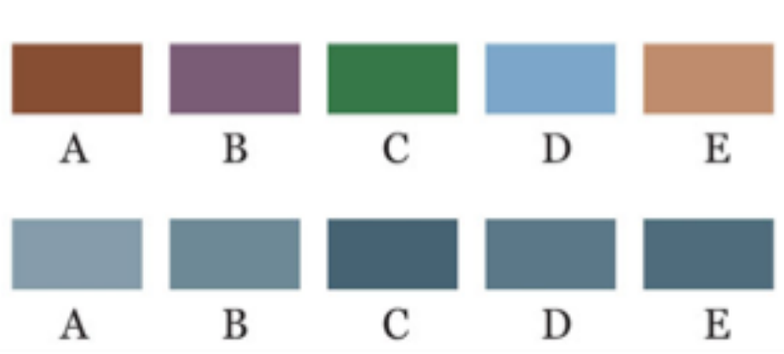



Area




Hue and shade



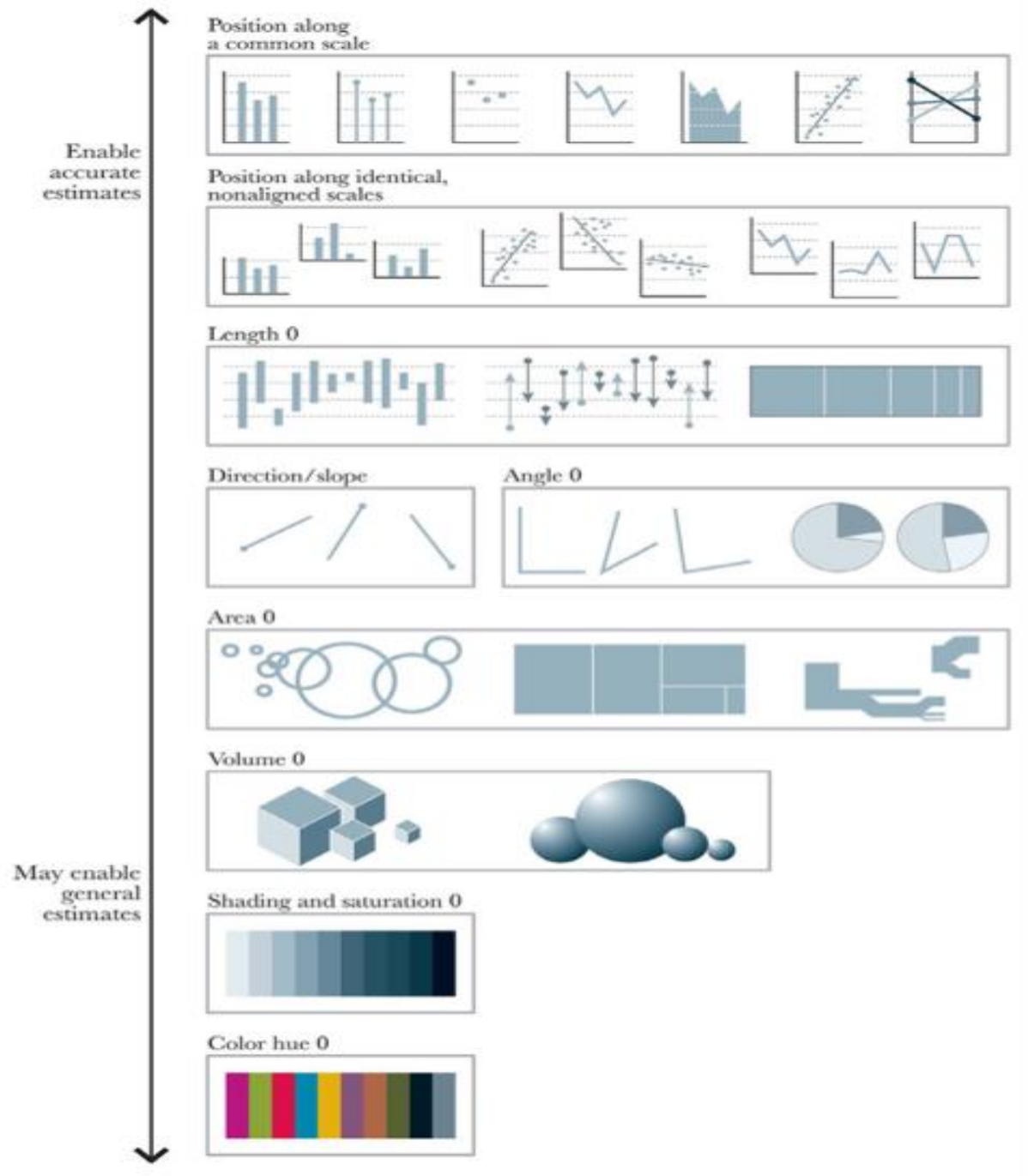

Angle/area


Line weight


Figures represented
in all these graphics:
22%, 25%, 34%, 29%, 32%

source: Alberto Cairo

# Tip2: Mode detailed or overall idea




source: Alberto Cairo

# Chart Suggestions—A Thought-Starter

What would you like to show?

## *Comparison*

### *Relationship*

- Two Variables
  - Scatter Chart
  - Bubble Chart
- Three Variables

### *Distribution*

- Single Variable
  - Few Data Points → Column Histogram
  - Many Data Points → Line Histogram
- Two Variables → Scatter Chart
- Three Variables → 3D Area Chart

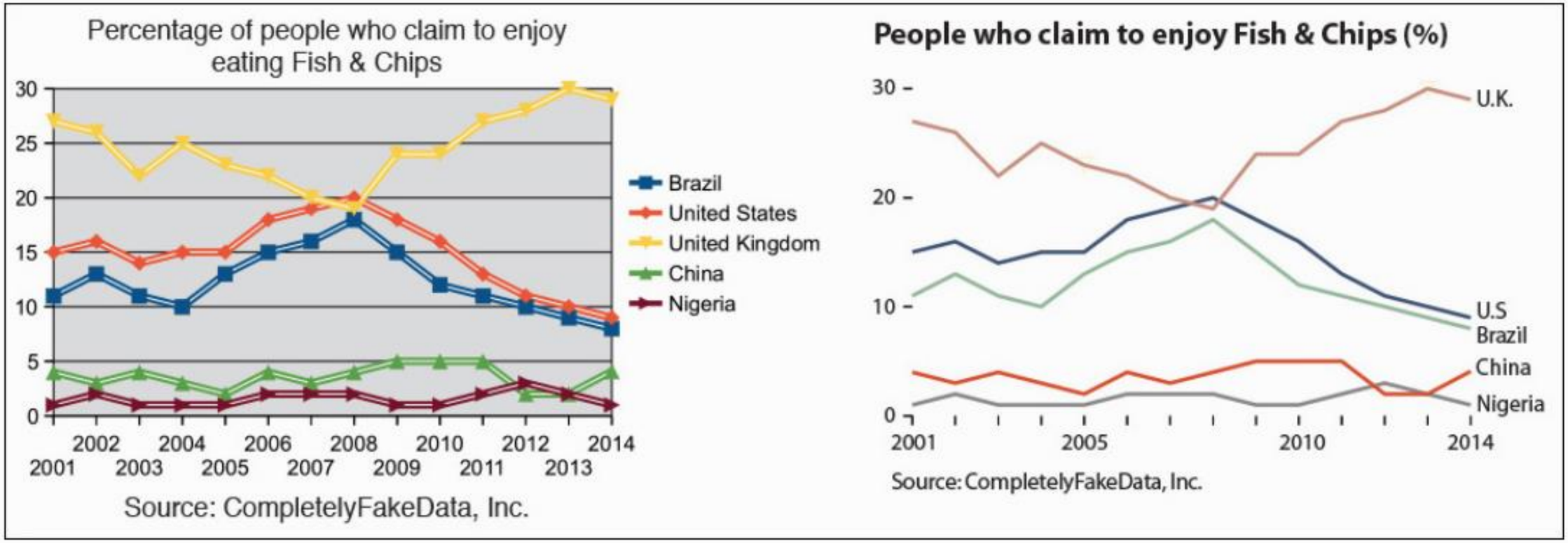
## *Composition*

- Changing Over Time
  - Few Periods
    - Only Relative Differences Matter → Stacked 100% Column Chart
    - Relative and Absolute Differences Matter → Stacked Column Chart
  - Many Periods
    - Only Relative Differences Matter → Stacked 100% Area Chart
    - Relative and Absolute Differences Matter → Stacked Area Chart
- Static
  - Simple Share of Total → Pie Chart
  - Accumulation or Subtraction to Total → Waterfall Chart
  - Components of Components → Stacked 100% Column Chart with Subcomponents

Over Time

- Among Items
  - Two Variables per Item
    - Variable Width Column Chart
    - Table or Table with Embedded Charts
  - Many Categories
    - Many Items → Bar Chart
    - Few Items → Column Chart
  - Few Categories
- One Variable per Item
  - Cyclical Data → Circular Area Chart
  - Non-Cyclical Data → Line Chart
  - Many Periods
- Single or Few Categories → Column Chart
- Many Categories → Line Chart
- Few Periods

# Which one is better?




source: Alberto Cairo

Carefully select a plot to make a point and show your data

# Better data beats fancier algorithms.

- Identifying relevant data and removing irrelevant data
- Duplicate entries
- Irrelevant observations
- Fix Irregular cardinality and structural errors
- Outliers
- Missing data treatment

Source: https://medium.com/limitedio/data-cleaning-the-secret-ingredient-to-the-success-of-any-data-science-project-e42d5e0df05a

# Summary

Previous Lecture: Introduction to BDA

This Lecture:

- Understanding data through measurements & plots
- Measurements: mean, standard deviations, percentiles, quartiles, data distributions. ….
- Plots: Which plot can represents this data best ?

Next Lecture: Customer Segmentation

- First Model: Techniques to segment customers’ data