INTRODUCTION

To Data Science

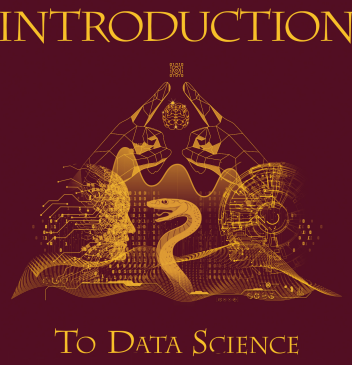
Introduction to Data Science

Collecting, Exploring and Cleaning Data - Class 3

Giora Simchoni

gsimchoni@gmail.com and add #intro2ds in subject

Stat. and OR Department, TAU



Common Data Formats in Data Science

INTRODUCTION

To Data Science

CSV: Comma Separated Values

```
id,drugName,condition,review,rating
206461,Valsartan,Left Ventricular Dysfunction,"""It has n
95260,Guanfacine,ADHD,"""My son is halfway through his fo
We have tried many different medications and so far this
92703,Lybrel,Birth Control,"""I used to take another oral
The positive side is that I didn&#039;t have any other si
138000,Ortho Evra,Birth Control,"""This is my first time
35696,Buprenorphine / naloxone,Opiate Dependence,"""Subox
```

```
import pandas as pd

df = pd.read_csv('../datasets/drugs.csv')

print(df.head())
```

```
      id                 drugName                      condition  \
0  206461                Valsartan  Left Ventricular Dysfunction
1   95260                Guanfacine                          ADHD
2   92703                    Lybrel               Birth Control
3  138000               Ortho Evra               Birth Control
4   35696  Buprenorphine / naloxone           Opiate Dependence

                                              review  rating
0  "It has no side effect, I take it in combinati...       9
1  "My son is halfway through his fourth week of ...       8
2  "I used to take another oral contraceptive, wh...       5
3  "This is my first time using any form of birth...       8
4  "Suboxone has completely turned my life around...       9
```

JSON: JavaScript Object Notation

```
{"fisrtname": "John",
 "lastname": "Smith",
 "age": "27",
"address": {"streetAddress": "21 2nd Street",
  "city": "New York",
  "state": "NY",
  "postalCode": "10021-3100"},
 "children": [],
 "spouse": null}
```

```
import json

data = dict()

with open('../datasets/test.json') as f:
    data=json.load(f)

data
```

```
{'fisrtname': 'John',
 'lastname': 'Smith',
 'age': '27',
 'address': {'streetAddress': '21 2nd Street',
  'city': 'New York',
  'state': 'NY',
  'postalCode': '10021-3100'},
 'children': [],
 'spouse': None}
```

Plain Text

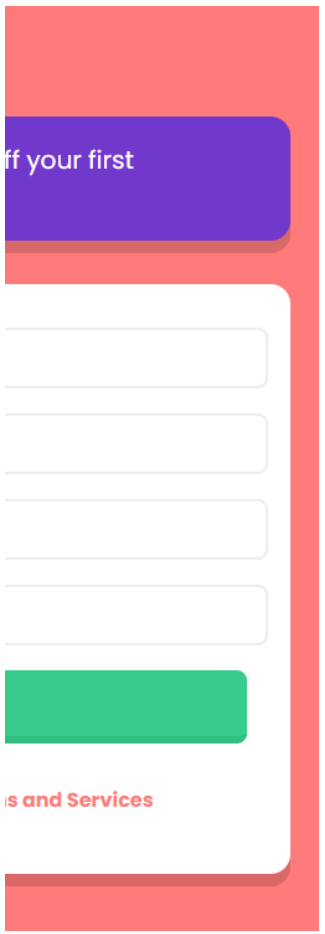
```
If you should go skating
On the thin ice of modern life
Dragging behind you the silent reproach
Of a million tear-stained eyes
Don't be surprised when a crack in the ice
Appears under your feet.
```

```
with open('../datasets/test.txt') as f:
    lines = f.readlines(1000)

lines
```

```
['If you should go skating,\n',
 'On the thin ice of modern life,\n',
 'Dragging behind you the silent reproach,\n',
 'Of a million tear-stained eyes,\n',
 "Don't be surprised when a crack in the ice,\n",
 'Appears under your feet.']
```

HTML

```
<body>
  <main>
    <!-- intro section -->
    <section class="intro">
      <h1 class="title">Black Friday Deals</h1>
      <p>Get up to 50% off on all our products and services. Hurry up, the offer ends in 24 hours.</p>
    </section>

    <!-- sign-up section -->
    <section class="sign-up">
      <p class="sign-up-para">Sign up for our newsletter and get 10% off your first purchase</p>
      <!-- the form itself -->
      <form class="sign-up-form">
        <div class="form-input">
          <input type="text" name="first-name" id="first-name" placeholder="First Name" required>
          <span>!</span>
          <p class="warning">First name cannot be empty</p>
        </div>

        <div class="form-input">
          <input type="text" name="last-name" id="last-name" placeholder="Last Name" required>
          <span>!</span>
          <p class="warning">Last name cannot be empty</p>
        </div>

        <div class="form-input">
          <input type="email" name="email" id="email" placeholder="Email Address" required>
          <span>!</span>
          <p class="warning">Looks like this is not an email</p>
        </div>

        <div class="form-input">
          <input type="Password" name="Password" id="Password" placeholder="Password" required>
          <span>!</span>
          <p class="warning">Password cannot be empty</p>
        </div>

        <input class="submit-btn" type="submit" value="Claim your offer">
        <p class="form-term">By clicking the button, you are agreeing to our <span>Terms and Services</span> </p>
      </form>
```



HTML

```
from bs4 import BeautifulSoup

with open('../datasets/test.html') as f:
    soup = BeautifulSoup(f, 'html.parser')

print(soup.prettify())
```

```
<!DOCTYPE html>
<html lang="en">
 <head>
  <meta charset="utf-8"/>
  <meta content="IE=edge" http-equiv="X-UA-Compatible"/>
  <meta content="width=device-width, initial-scale=1.0" name="viewport"/>
  <title>
   Sample HTML Code - NewsLetter Form
  </title>
  <style>
   @import url("https://fonts.googleapis.com/css2?family=Poppins:wght@400;500;600;700&display=swap");

    body {
      display: flex;
      justify-content: center;
      padding: 3rem 0;
      font-family: "Poppins", sans-serif;
      font-size: 1rem;
      color: white;
      background-color: #ff7a7a;
    }
```

Collecting Data

INTRODUCTION

To Data Science

Where do(es) data come from?

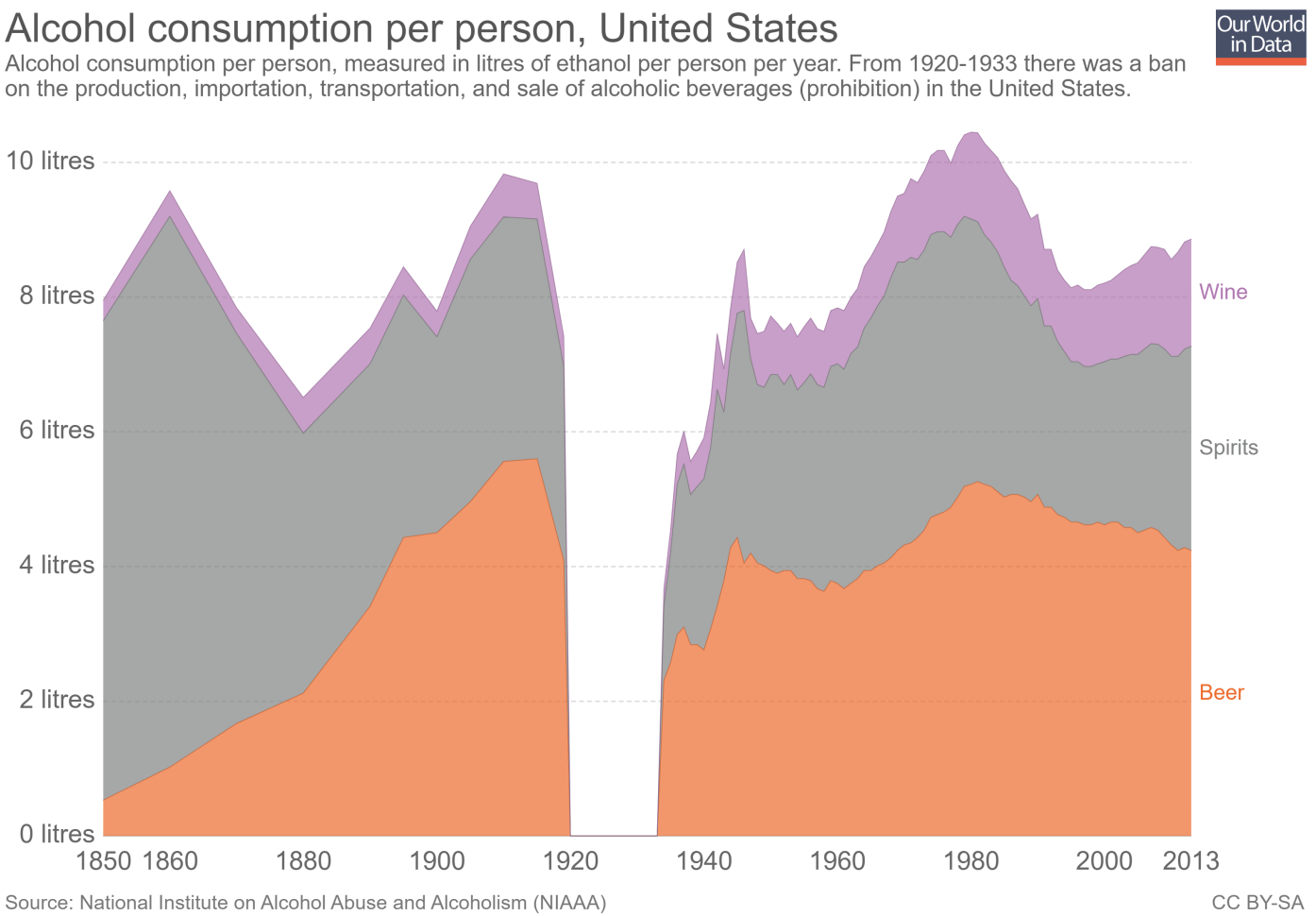
- Then: Manual Curation
- Now: Automatic Curation

Then: Manual Curation (I)

If your parents have not taken note, anywhere, of how tall you were at the age of 1 - we may never be able to extract this information.

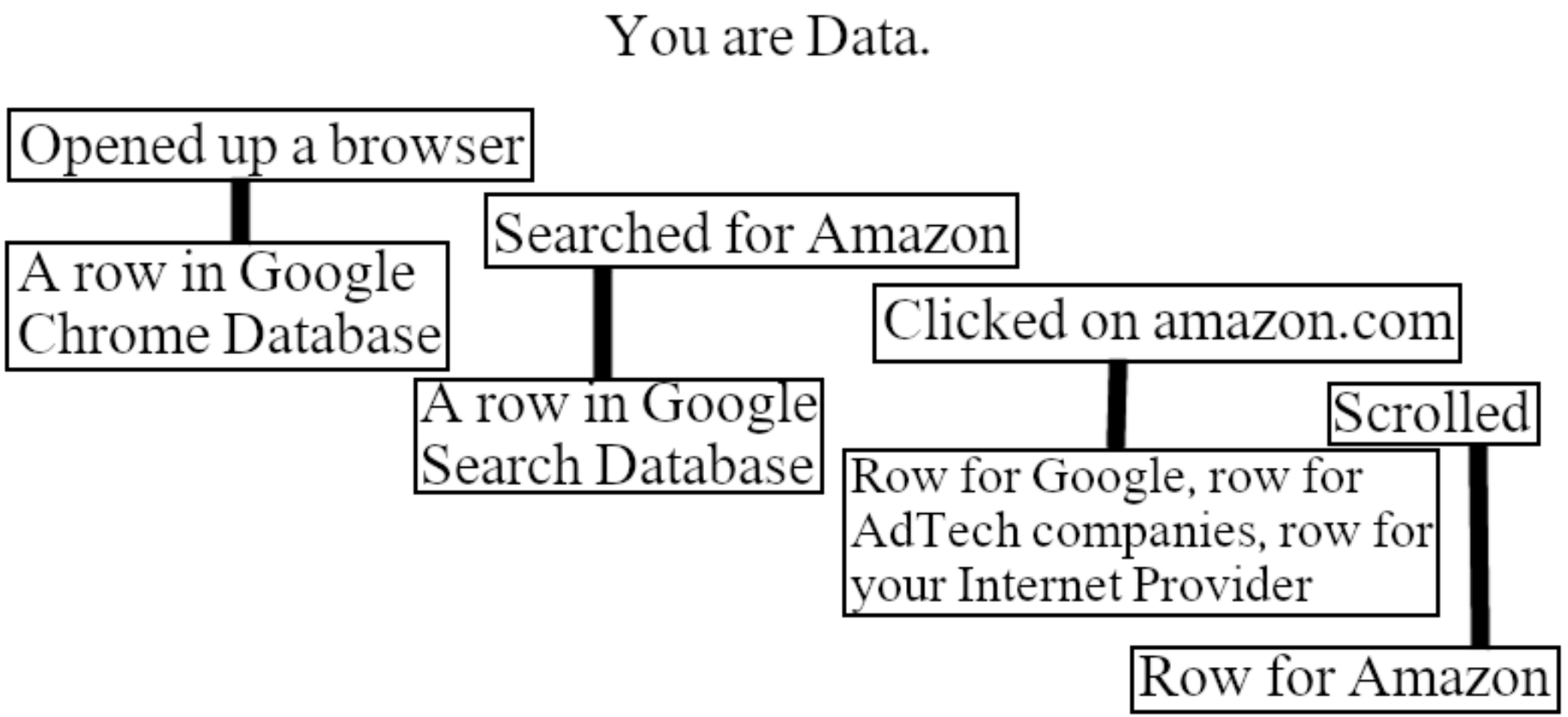
Then: Manual Curation (II)

If the US government had not seen fit to estimate and record the level of alcohol consumption of its citizens, we would never have known.



Now: Automatic Curation (I)

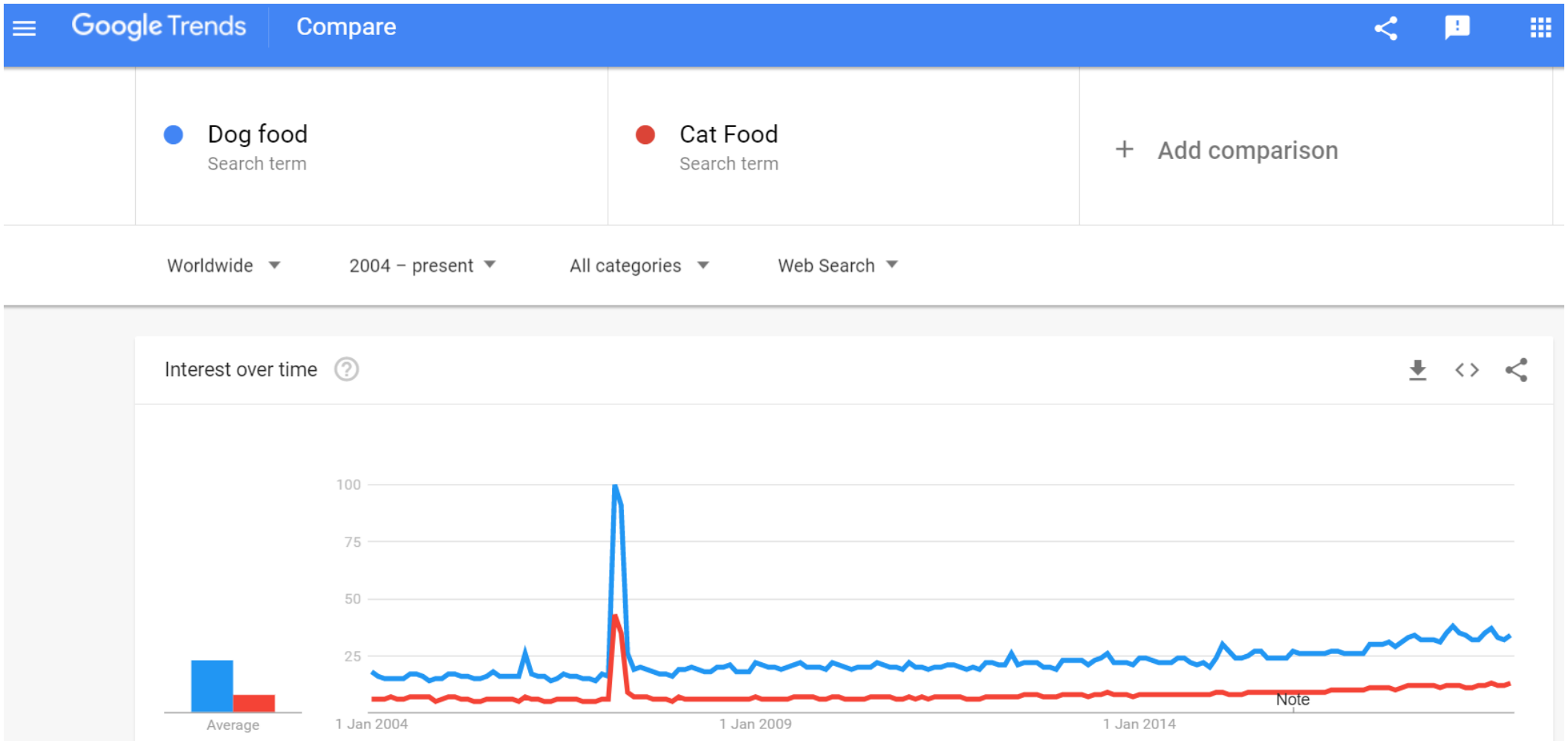
Have you ever opened up an Internet browser, searched for “Amazon”, clicked on amazon.com and scrolled around to check the price of a T-shirt? You don’t have to be logged in. You don’t have to buy. You are data.



Web Scraping

- Public APIs
- Beautiful Soup

Public APIs (I)



Beautiful Soup

You may not even need an API. Would you look to have all the Beatles records in a single table on your machine?

The following code scrapes the Wikipedia page for the and creates a table, out of "thin air". See more advanced examples in recitation.

```
1 from bs4 import BeautifulSoup
2 import requests
3 import re
4 import pandas as pd
5
6 url = 'https://en.wikipedia.org/wiki/The_Beatles_discography'
7 r = requests.get(url)
8 soup = BeautifulSoup(r.content, 'html.parser')
```

```
albums = dict()
id = 0
albums[id] = dict()
tables = soup.find_all('table')
for table in tables:
    caption = table.find('caption')
    if caption is not None:
        header = caption.get_text()
        if re.match(re.compile('^List of(.+?)albums'), header):
            rows = table.find_all('tr')
            for row in rows:
                title_col = row.find('th')
                if title_col is not None and 'scope' in title_col.attrs and\
                title_col.attrs['scope'] == 'row':
                    title_cell = title_col.find('a')
                    if title_cell is not None and title_cell.attrs is not None and\
                    'title' in title_cell.attrs:
                        albums[id]['name'] = title_cell.attrs['title']
                        release_col = row.find('td')
                        release_date, release_label = get_release_details(release_c
```

```
albums_df = pd.DataFrame.from_dict(albums, orient ='index')
albums_df.head(5)
```

	name	release_date	release_label
0	Please Please Me	22 March 1963	Parlophone
1	With the Beatles	22 November 1963	Parlophone (UK), Capitol (Canada), Odeon (France)
2	A Hard Day's Night (album)	10 July 1964	Parlophone
3	Beatles for Sale	4 December 1964	Parlophone
4	Help!	6 August 1965	Parlophone

Small Data, Big Data

INTRODUCTION

To Data Science

What's in a name?

These definitions are constantly changing.

- "Everything processed in Excel is small data." (Rufus Pollock, The Guardian)
- "[Big Data] is data so large it does not fit in main memory" (Leskovec et al., Mining of Massive Datasets)

Or maybe we should define the size of our data according how easy it is to process and understand it?

- "[Small Data is] data that has small enough size for human comprehension." (jWork.ORG)
- "data sets that are too large or complex for traditional data-processing application software to adequately deal with" (Wikipedia)

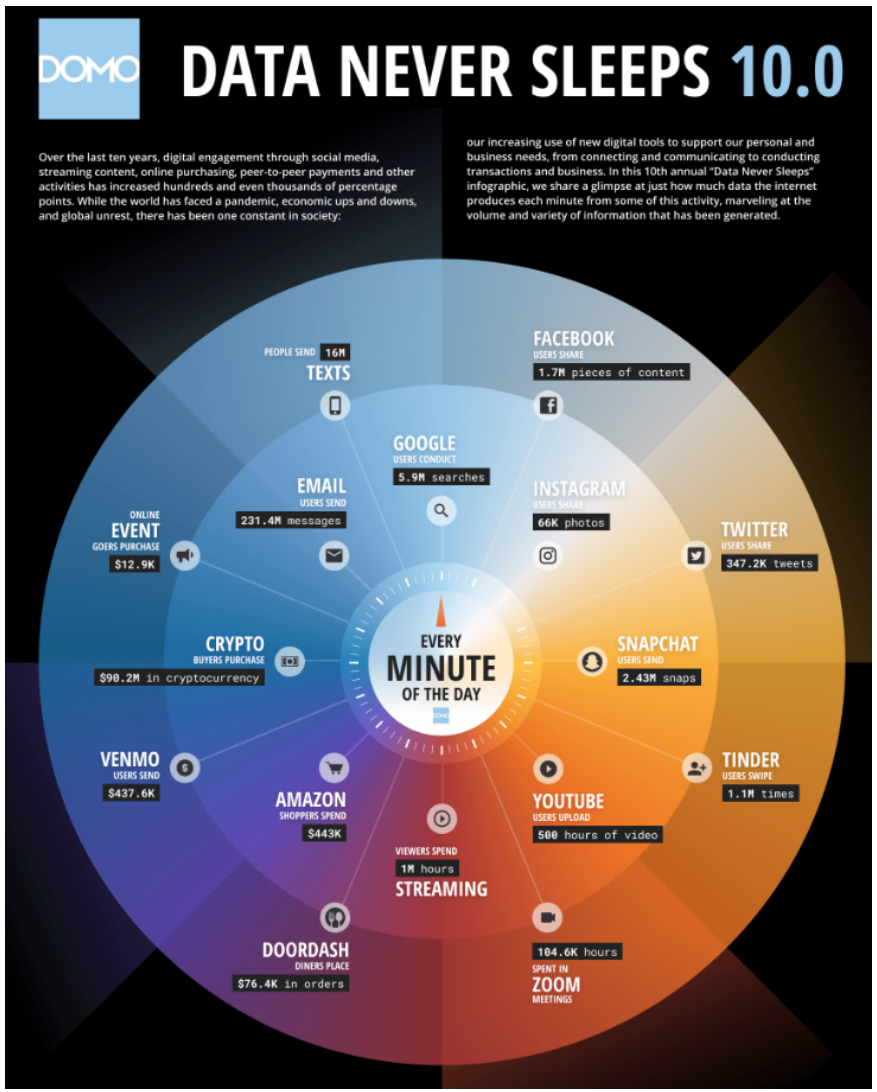
What's in a name? (II)

The actual definition should probably merge both of the above.

- Excel can fit 1M rows, 16K columns of double numbers. Try loading a matrix such as this into Matlab, Python or R, multiply and invert it - you can't. So isn't that Big?
- Facebook generates 4 Petabytes of data, daily. That's 4K Terabytes or 4M Gigabytes. (Brandwatch.com) But a Facebook Data Scientist in daily life typically needs only a copy of some of these data, which fits in her PC. Isn't that small?

Web data is Big Data

We can all agree *this* is big: (Domo.com)



Exploring Data: Basic Plots

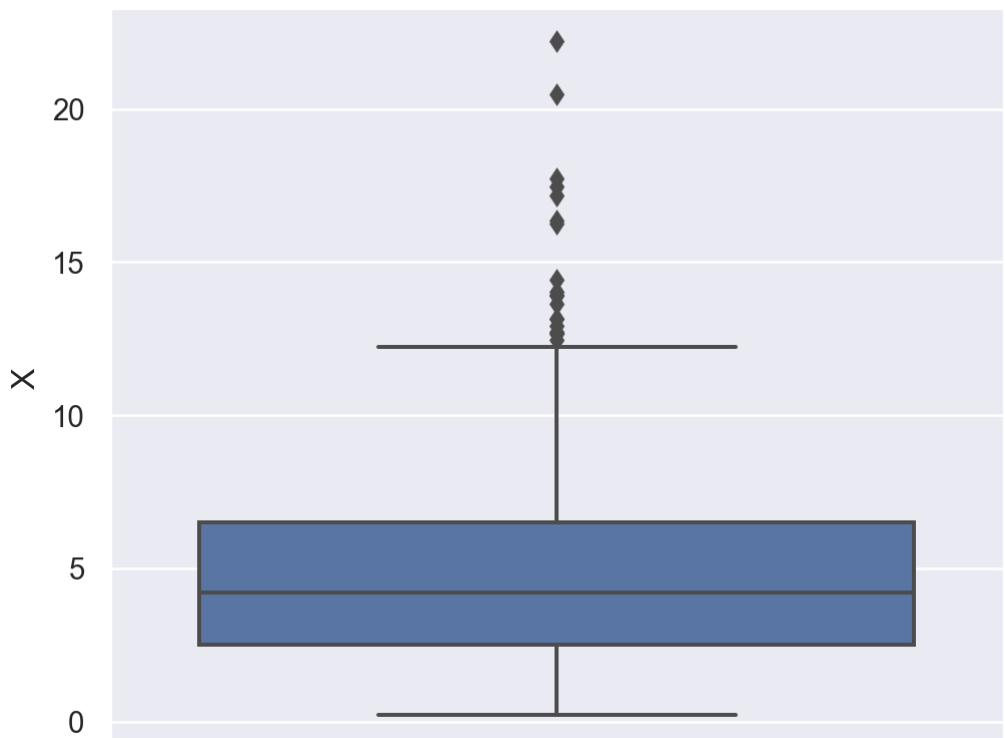
INTRODUCTION

TO DATA SCIENCE

Boxplot

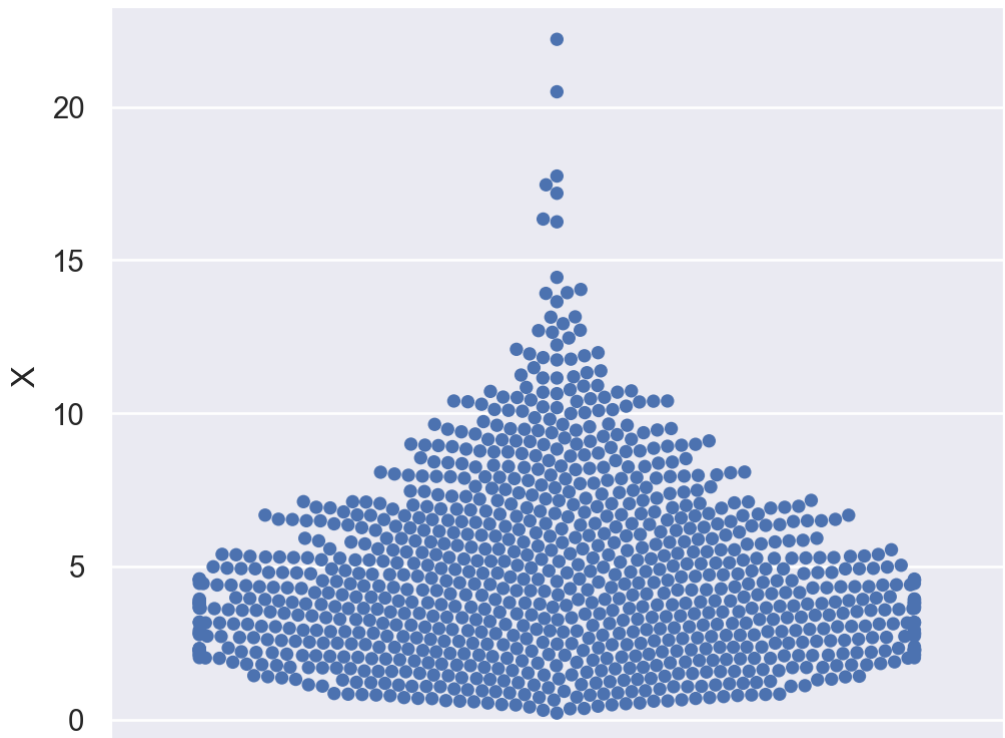
```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

```
sns.boxplot(y = X)
plt.ylabel('X')
plt.show()
```



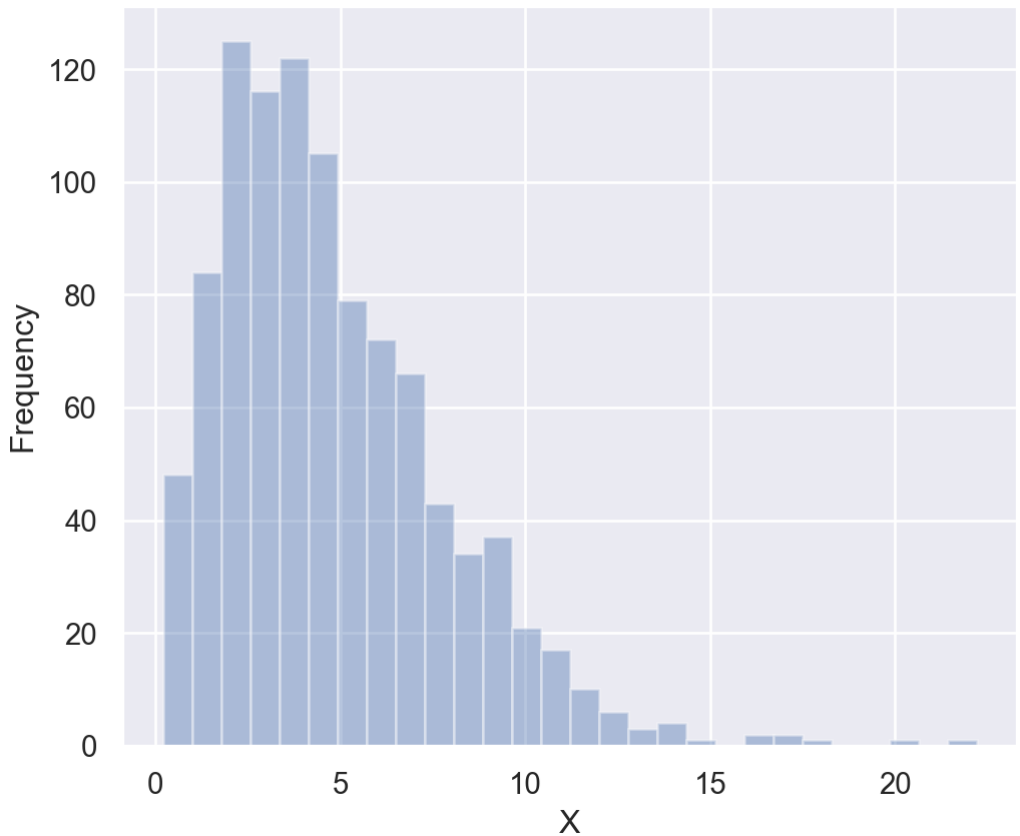
Swarmplot

```
sns.swarmplot(y = X)
plt.ylabel('X')
plt.show()
```



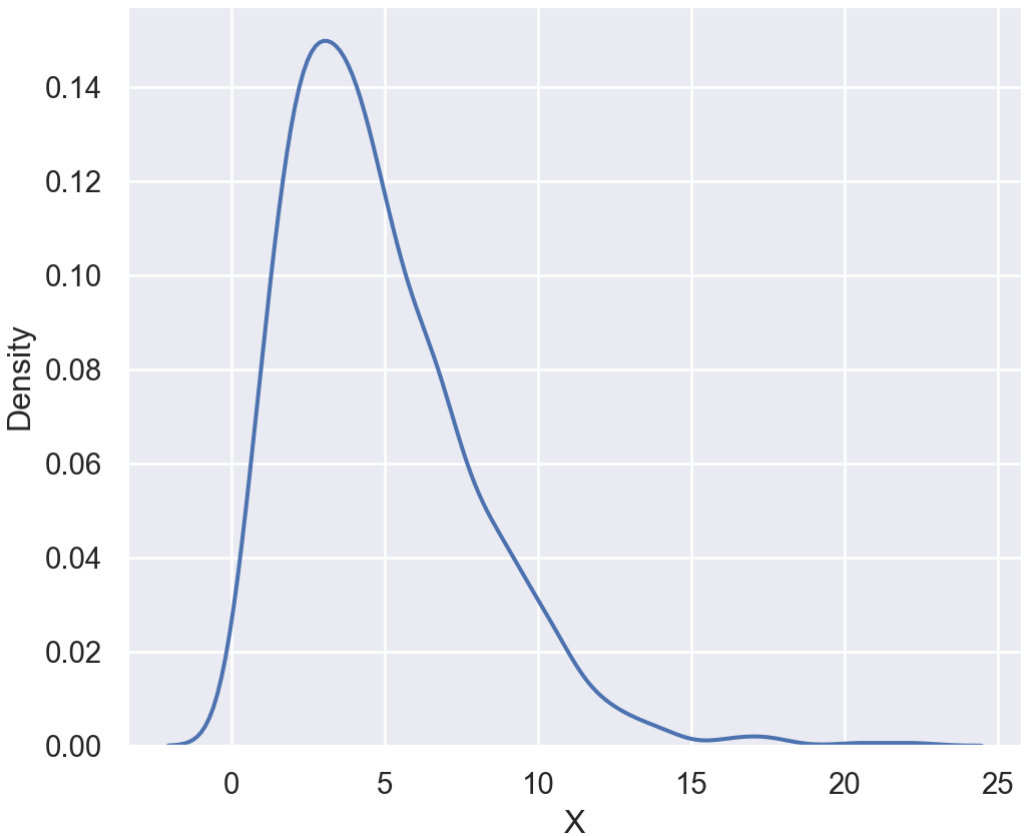
Histogram

```
sns.distplot(X, kde = False)
plt.ylabel('Frequency')
plt.xlabel('X')
plt.show()
```



Density Plot

```
sns.distplot(X, hist = False)
plt.ylabel('Density')
plt.xlabel('X')
plt.show()
```



Density plot: kernel density estimation / convolution

The way to get a smooth estimate of the distribution in density plot is by defining a kernel which “smoothes” the data. Mathematically we define a kernel weight function $w : \mathbb{R} \to \mathbb{R}^+$ as:

1. Non-negative and symmetric: $w(x) = w(-x)$
2. Integrates to 1: $\int_{\mathbb{R}} w(x)dx = 1$

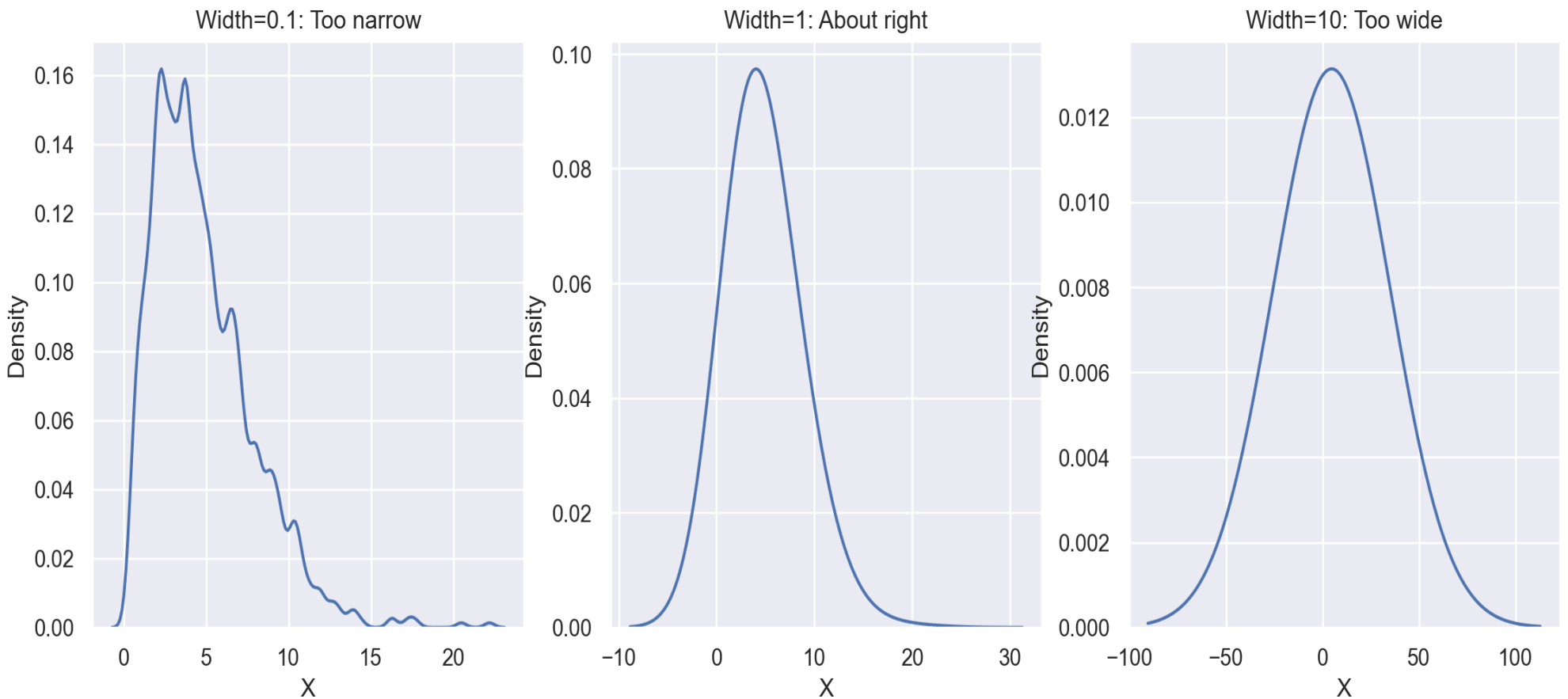
And then the density kernel estimate is: $J(x) = \frac{1}{n} \cdot \sum_{i=1}^{n} w(x_i - x)$.

Nice property: $\int_{\mathbb{R}} J(x)dx = 1$

- Wide w: smooth estimate, but it may not reflect the real data
- Narrow w: very non-smooth description

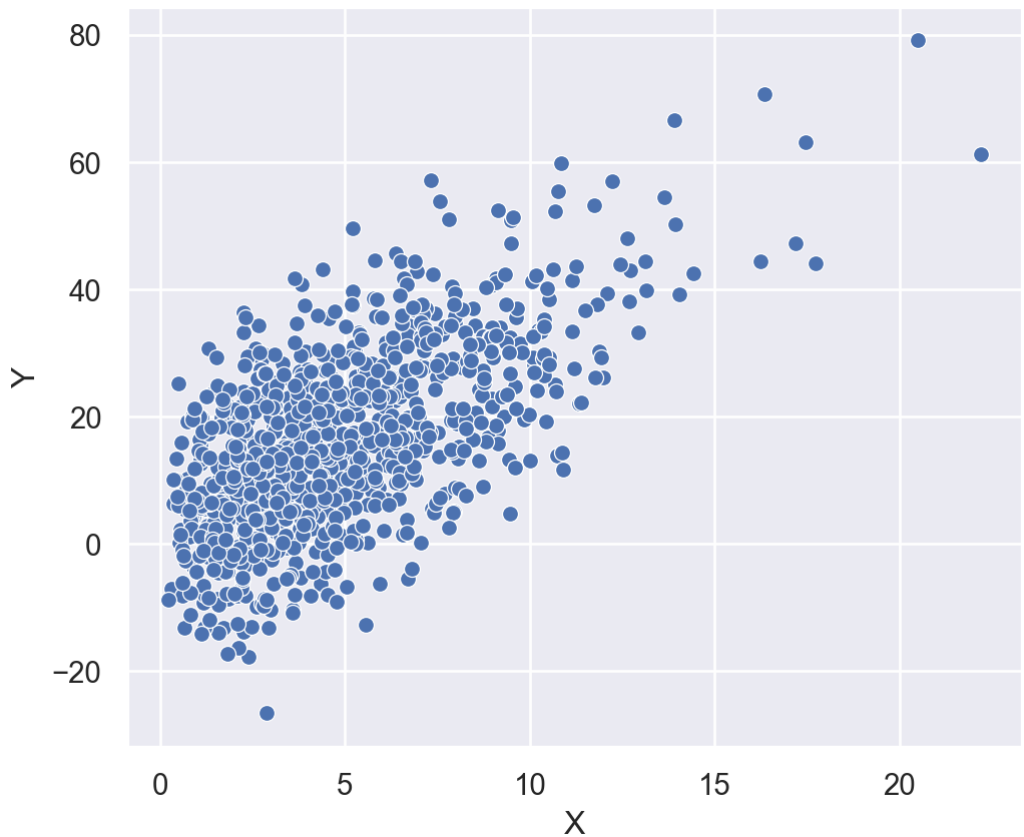
For example smoothing this same dataset with too narrow or too wide window:

▶ Code



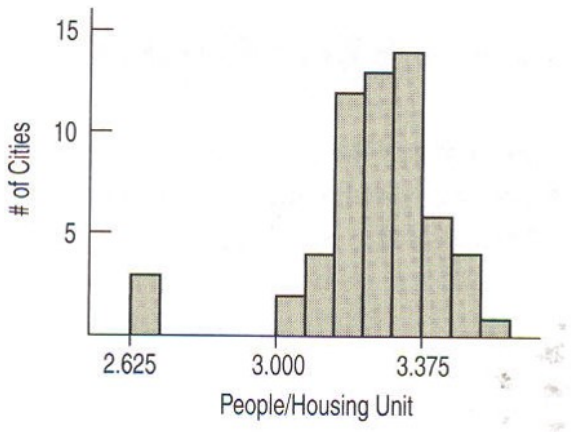
Scatterplot

```
Y = X * 3 + 2 + np.random.normal(0, 10, 1000)
sns.scatterplot(X, Y)
plt.ylabel('Y')
plt.xlabel('X')
plt.show()
```

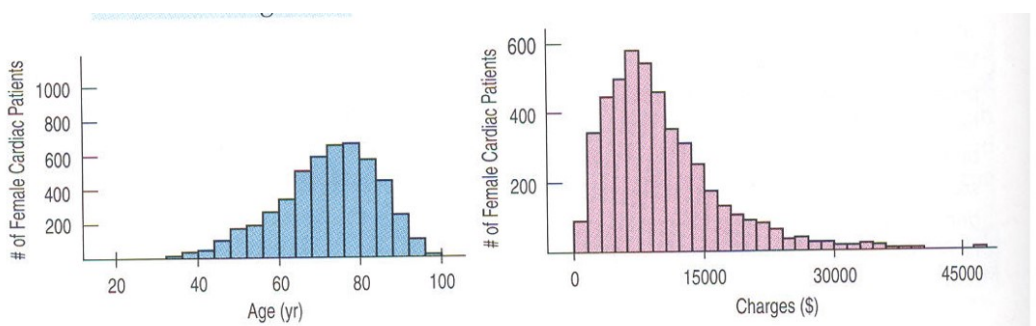


What can we learn from simple plots?

Look at outliers:



See the shape and tail direction:



Age at heart attack (left) and cost of hospitalization (right)

Exploring Data: Summary Statistics

INTRODUCTION

TO DATA SCIENCE

Location

"Where is this X located? Where is the central mass?"

- Mean of empirical distribution (=average):

$$Mean(X) = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

- Median:

$$Med(X) = m \; s.t. \; P(X \leq m) = P(X \geq m) = 0.5$$

- Mode:

$$Mode(X) = Most \; frequent \; value \; in \; a \; dataset$$

```
mean = np.mean(X)
median = np.median(X)
hist, _ = np.histogram(X, bins=range(20))
mode = list(range(20))[hist.argsort()[::-1][0]]
plt.figure(figsize= (4,4))
sns.distplot(X, bins = range(20), kde = False)
plt.plot([mean, mean], [0, 160], linewidth=2, color='r')
plt.plot([median, median], [0, 160], linewidth=2, color='g')
plt.plot([mode, mode], [0, 160], linewidth=2, color='b')
plt.legend({'Mean':mean,'Median':median,'Mode':mode})
plt.xlabel('X')
plt.ylabel('Frequency')
plt.show()
```

Dispersion

“Is X widely spread out? Does it concentrate narrowly around the mean?”

- Quantiles/Percentiles:

$$Q(X, q) = v \; s.t. \; P(X \leq v) = 1 - P(X \geq v) = q$$

- Range:

$$Range(X) = Max(X) - Min(X)$$

- Inter-Quartile-Range:

$$IQR(X) = Q(X, 0.75) - Q(X, 0.25)$$

Dispersion

- (Empirical) Variance:

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(X_i - Mean(X))^2$$

- Standard Deviation:

$$STD(X) = \sqrt{Var(X)}$$

```
print(f'90th percentile: {np.percentile(X, 90) :.2f}')
print(f'Range: {np.max(X) - np.min(X) :.2f}')
print(f'IQR: {np.percentile(X, 75) - np.percentile(X, 25) :.2f}')
print(f'Variance: {np.var(X) :.2f}')
print(f'Standard Deviation: {np.std(X) :.2f}')
```

```
90th percentile: 8.95
Range: 22.00
IQR: 3.97
Variance: 9.12
Standard Deviation: 3.02
```

Shape

"Is X symmetric or not? How 'tailed' is it?"

- Skewness:

$$Skew(X) = \frac{1}{N} \frac{\sum_{i=1}^{N} (X_i - Mean(X))^3}{STD(X)^3}$$

```
from scipy import stats

print(f'Skewness: {stats.skew(X) :.2f}')
```

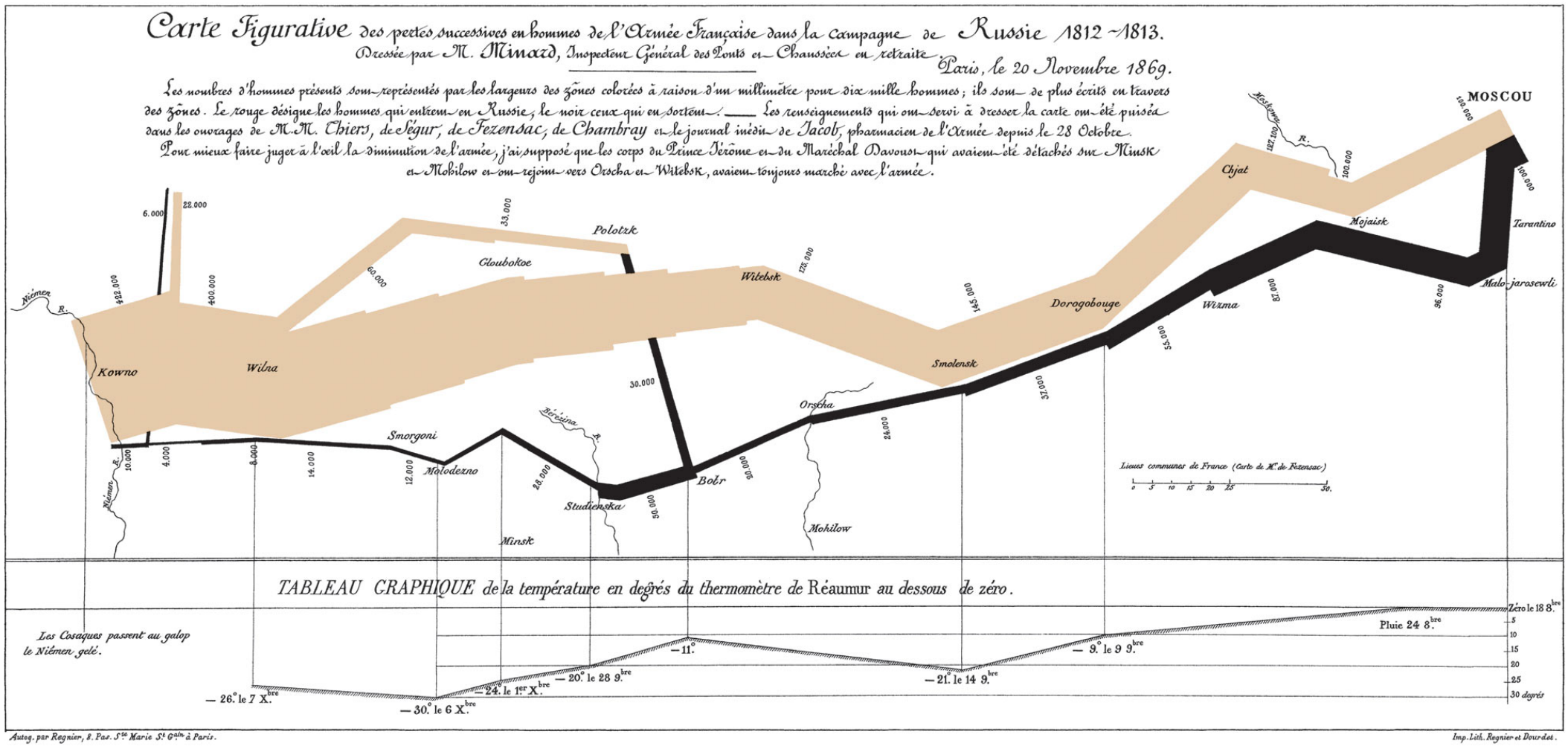
```
Skewness: 1.22
```

Advanced Visualization

INTRODUCTION

TO DATA SCIENCE

Minard's Napoleon March

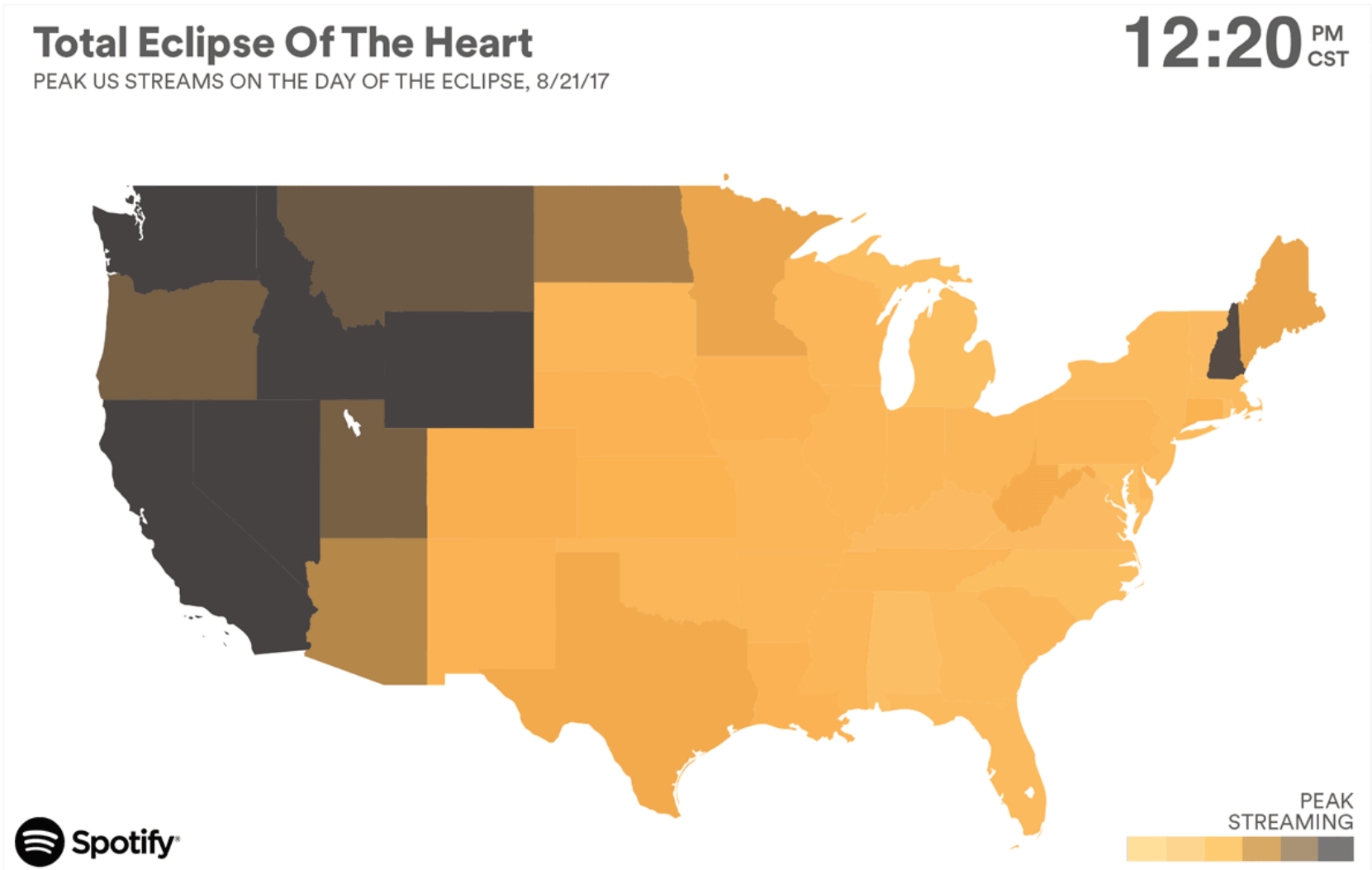


Heatmaps

source

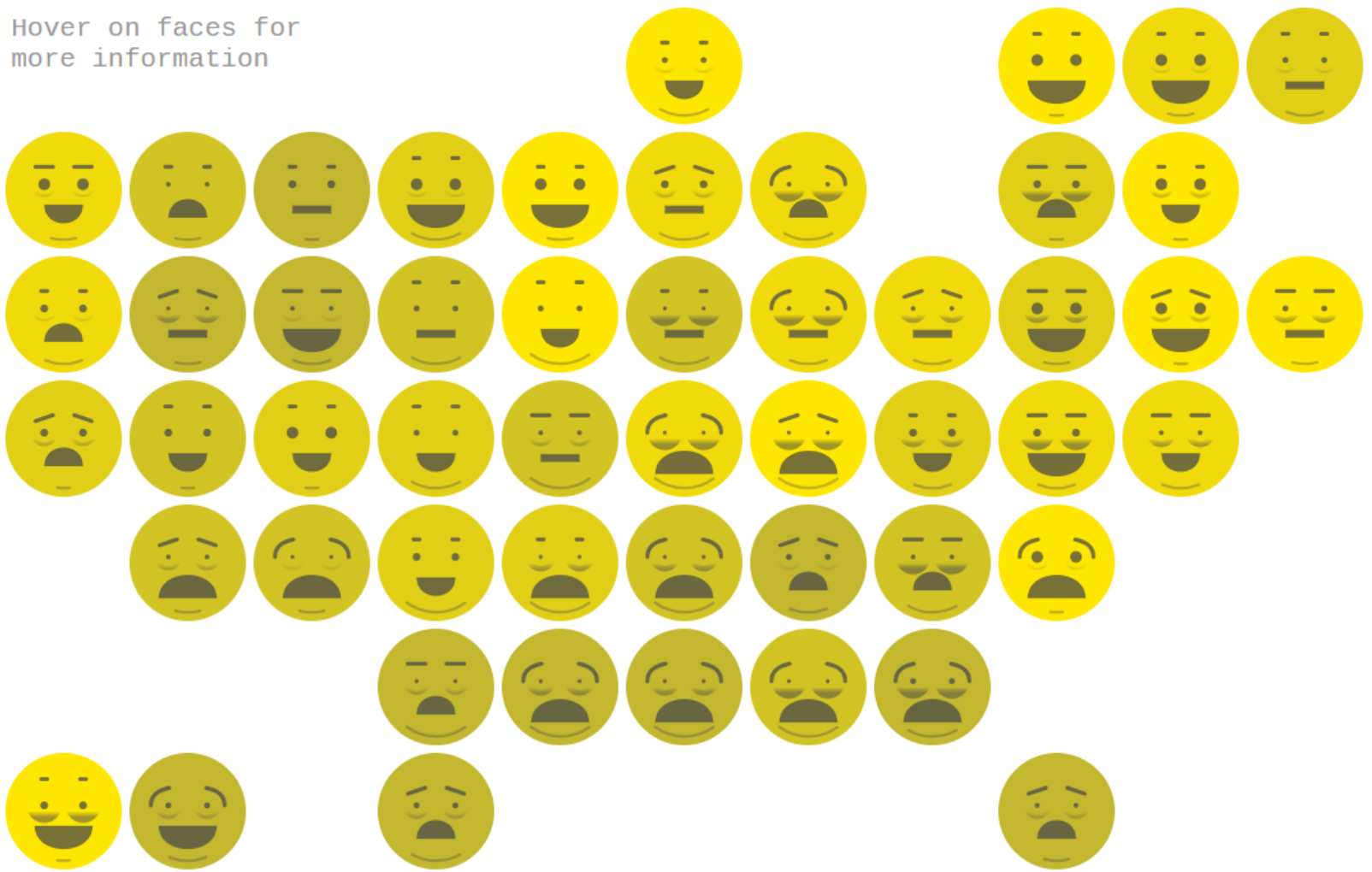
Spotify: Total Eclipse of the Heart

source

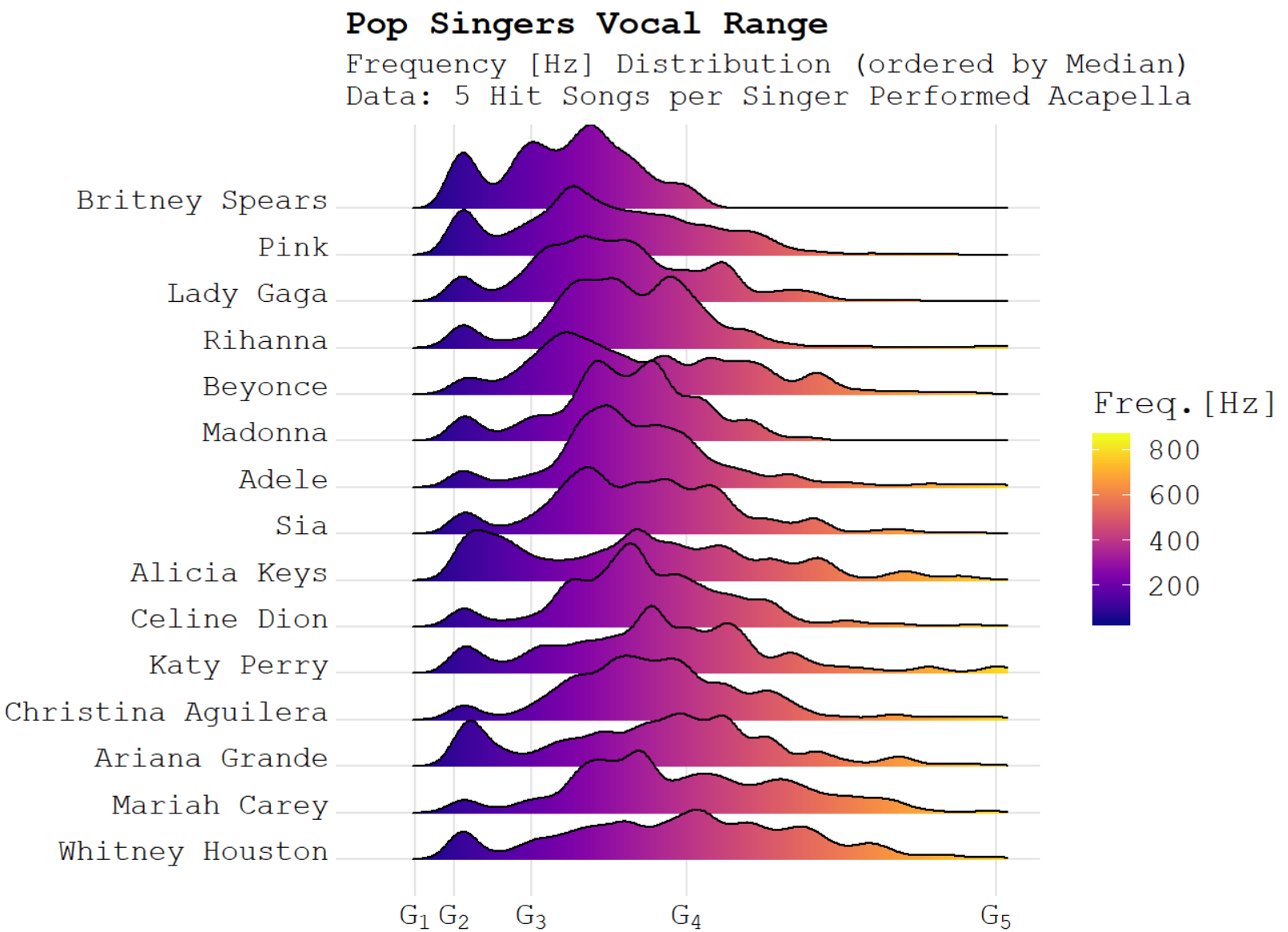


Chernoff Faces

source



Ridge plot (a.k.a Joy plot)

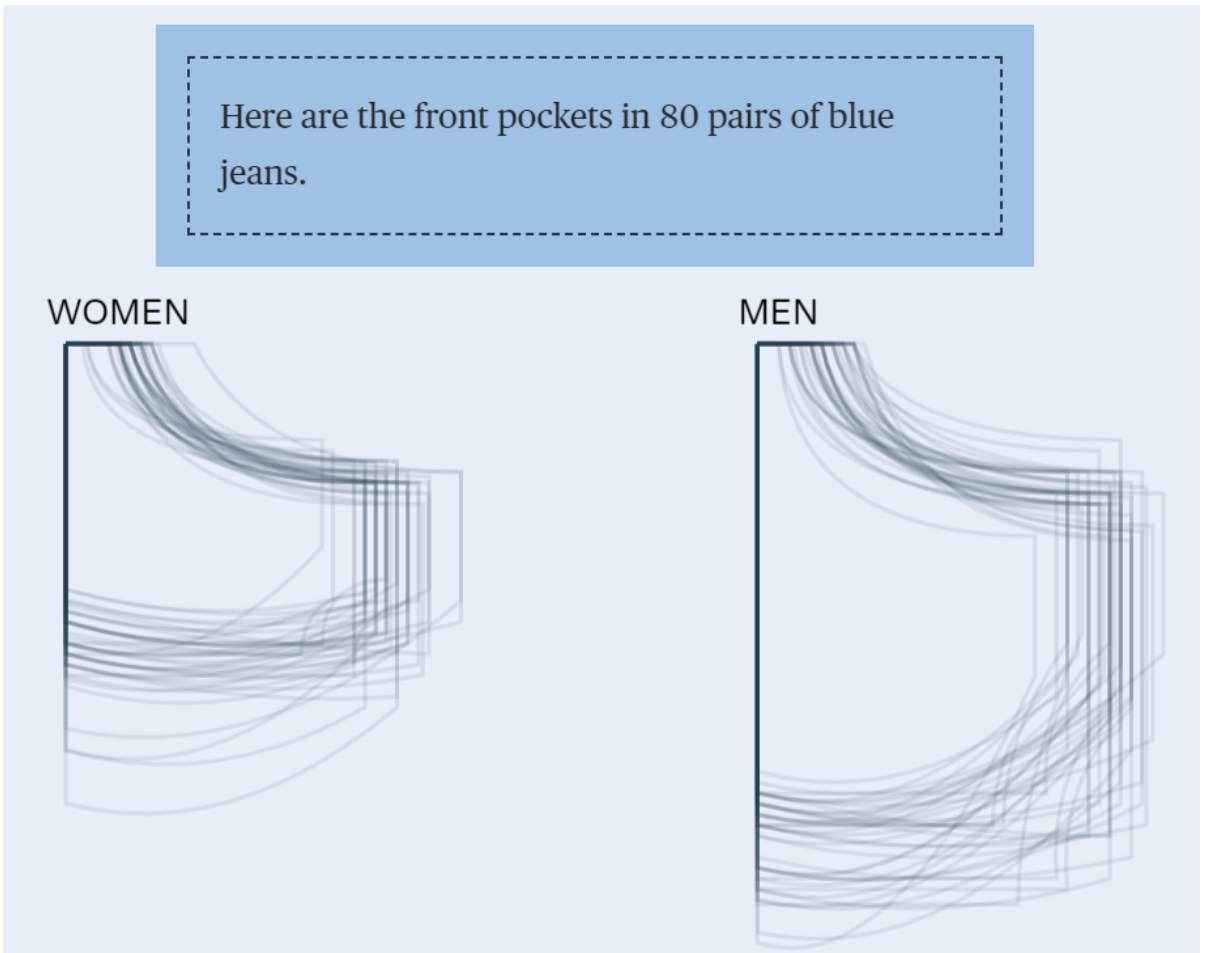


Flowing Data: A Day in the Life of Americans

source

Pudding: Women's Pockets (and every single post on their site!)

source



The Gapminder story

The history of the world encapsulated in a simple visualization:

source

Dangers of Dirty Data

INTRODUCTION

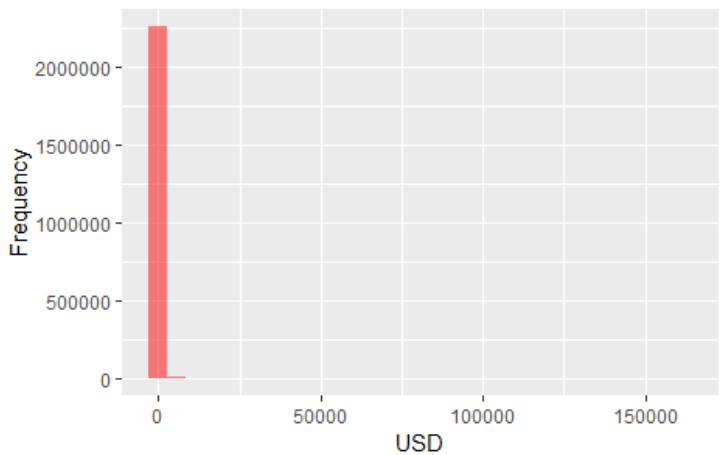
TO DATA SCIENCE

What could be dirty about data?

- The data itself
- The data's structure

The data itself: Outliers

- Numerical Outliers: This is a histogram of random ~2.3 million transactions on ebay US website in over a few weeks in 2013 (source):



- Textual Outliers: The Blog Authorship Corpus consists of 19K posts by bloggers from blogger.com in 2004. These are actual words used in the 10-20 age group:

> aaaaaaaaaaaaaaaaaaaaargh, lolzi, jfjgfjhgjhfjgfjf, roflmfao, duuuuuuh, walang, dunno

The data itself: Missing data

source

כל כותרות היום

קמיל פוקס על כישלון המדגמים: "כשסיפרו לי על התוצאות, חשבתי שאני מת"

בכיר באחד הערוצים: "טעינו בגדול. גילינו פער של 5%-6% בתוצאות בין המדגם לבין הקלפי - אנשים נכנסו לקלפי המדגמית ושיקרו" ■ ד"ר מינה צמח: "סגרנו את קלפיות המדגם בשעה שמונה וחצי בערב אבל ראינו שככל שהשעה יותר מאוחרת, כך הליכוד מתחזק"

שמור 176 16

עדי דברת-מזריץ | התראות במייל
14:52 18.03.2015

קריאת זן

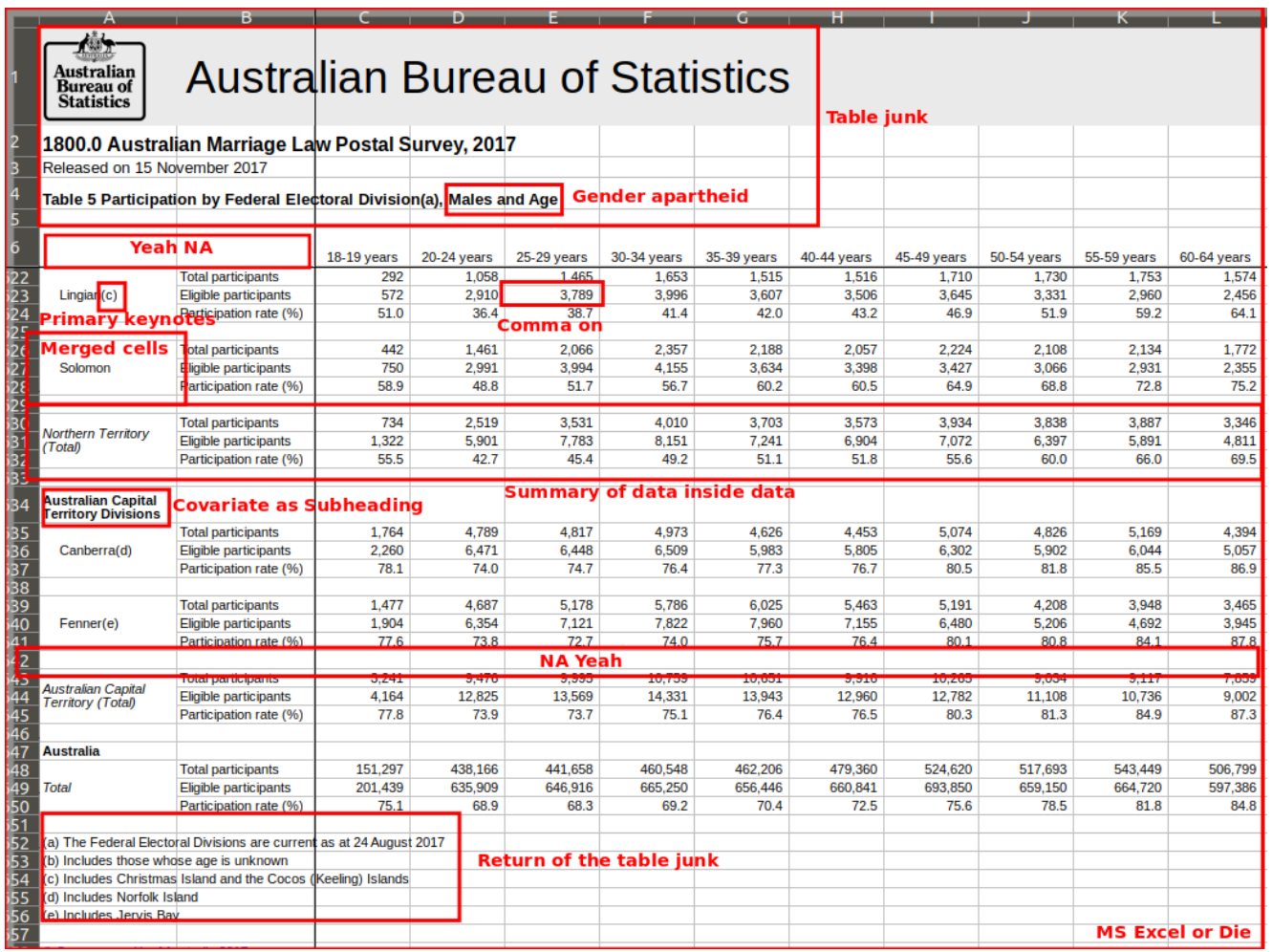
פרופסור קמיל פוקס שמע לראשונה על כך שתוצאות המדגם שהציג אתמול בערוץ 10 שגויות רק כשהתעורר היום בבוקר משיחת טלפון של מפיק ערוץ 10, שהזמין אותו לאולפן כדי לדון בתוצאות האמת של הבחירות. "זוגתי אמרה לי 'שמעת מה קרה?' וסיפרה לי על תוצאות האמת, ואני חשבתי שאני מת".

פוקס סיפר כי "זו לא הפעם הראשונה שזה קורה לי - בשנת 96' אמרתי לחיים יבין שתוצאות המדגם מלמדות על תיקו עם יתרון קל לפרס ואז הלכתי לישון עם פרס והתעוררתי עם ביבי (נתניהו). אבל הפעם התחושה הרבה יותר גרועה כי הפער בין תוצאות המדגם לתוצאות האמת יותר גדולות ובשנת 96' הפער היה רק של 16 אלף קולות", אומר פוקס ל-TheMarker ביום שאחרי מה שייזכר

The data's structure

In a word: Excel.

source



Australian Bureau of Statistics

Table junk

1800.0 Australian Marriage Law Postal Survey, 2017

Released on 15 November 2017

Table 5 Participation by Federal Electoral Division(a), Males and Age — Gender apartheid

Yeah NA

		18-19 years	20-24 years	25-29 years	30-34 years	35-39 years	40-44 years	45-49 years	50-54 years	55-59 years	60-64 years
Lingiari(c)	Total participants	292	1,058	1,465	1,653	1,515	1,516	1,710	1,730	1,753	1,574
	Eligible participants	572	2,910	3,789	3,996	3,607	3,506	3,645	3,331	2,960	2,456
	Participation rate (%)	51.0	36.4	38.7	41.4	42.0	43.2	46.9	51.9	59.2	64.1
Solomon	Total participants	442	1,461	2,066	2,357	2,188	2,057	2,224	2,108	2,134	1,772
	Eligible participants	750	2,991	3,994	4,155	3,634	3,398	3,427	3,066	2,931	2,355
	Participation rate (%)	58.9	48.8	51.7	56.7	60.2	60.5	64.9	68.8	72.8	75.2
Northern Territory (Total)	Total participants	734	2,519	3,531	4,010	3,703	3,573	3,934	3,838	3,887	3,346
	Eligible participants	1,322	5,901	7,783	8,151	7,241	6,904	7,072	6,397	5,891	4,811
	Participation rate (%)	55.5	42.7	45.4	49.2	51.1	51.8	55.6	60.0	66.0	69.5
Australian Capital Territory Divisions											
Canberra(d)	Total participants	1,764	4,789	4,817	4,973	4,626	4,453	5,074	4,826	5,169	4,394
	Eligible participants	2,260	6,471	6,448	6,509	5,983	5,805	6,302	5,902	6,044	5,057
	Participation rate (%)	78.1	74.0	74.7	76.4	77.3	76.7	80.5	81.8	85.5	86.9
Fenner(e)	Total participants	1,477	4,687	5,178	5,786	6,025	5,463	5,191	4,208	3,948	3,465
	Eligible participants	1,904	6,354	7,121	7,822	7,960	7,155	6,480	5,206	4,692	3,945
	Participation rate (%)	77.6	73.8	72.7	74.0	75.7	76.4	80.1	80.8	84.1	87.8
Australian Capital Territory (Total)	Total participants	3,241	9,476	9,995	10,759	10,651	9,916	10,265	9,034	9,117	7,859
	Eligible participants	4,164	12,825	13,569	14,331	13,943	12,960	12,782	11,108	10,736	9,002
	Participation rate (%)	77.8	73.9	73.7	75.1	76.4	76.5	80.3	81.3	84.9	87.3
Australia											
Total	Total participants	151,297	438,166	441,658	460,548	462,206	479,360	524,620	517,693	543,449	506,799
	Eligible participants	201,439	635,909	646,916	665,250	656,446	660,841	693,850	659,150	664,720	597,386
	Participation rate (%)	75.1	68.9	68.3	69.2	70.4	72.5	75.6	78.5	81.8	84.8

Primary keynotes

Merged cells

Comma on

Summary of data inside data

Covariate as Subheading

NA Yeah

(a) The Federal Electoral Divisions are current as at 24 August 2017
(b) Includes those whose age is unknown
(c) Includes Christmas Island and the Cocos (Keeling) Islands
(d) Includes Norfolk Island
(e) Includes Jervis Bay

Return of the table junk

MS Excel or Die

Spreadsheet Blunder

source

BBC Sign in Home News Sport Reel Worklife Travel

NEWS

Home | War in Ukraine | Climate | Video | World | UK | Business | Tech | Science | Entertainment & Arts

Tech

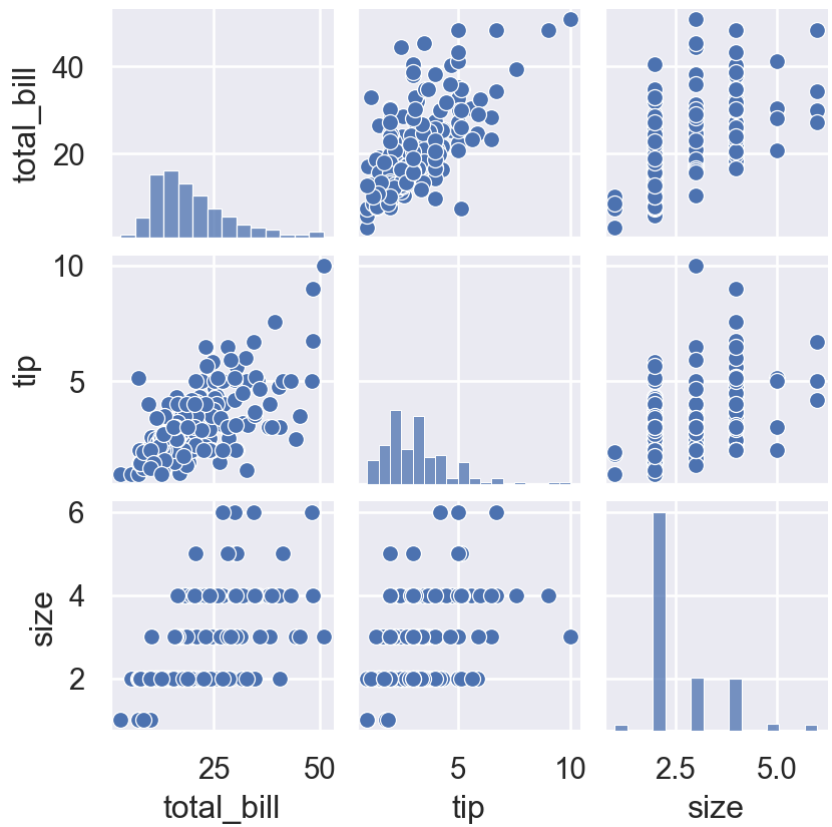
Excel: Why using Microsoft's tool caused Covid-19 results to be lost

5 October 2020

Some advice on cleaning data

Plot first:

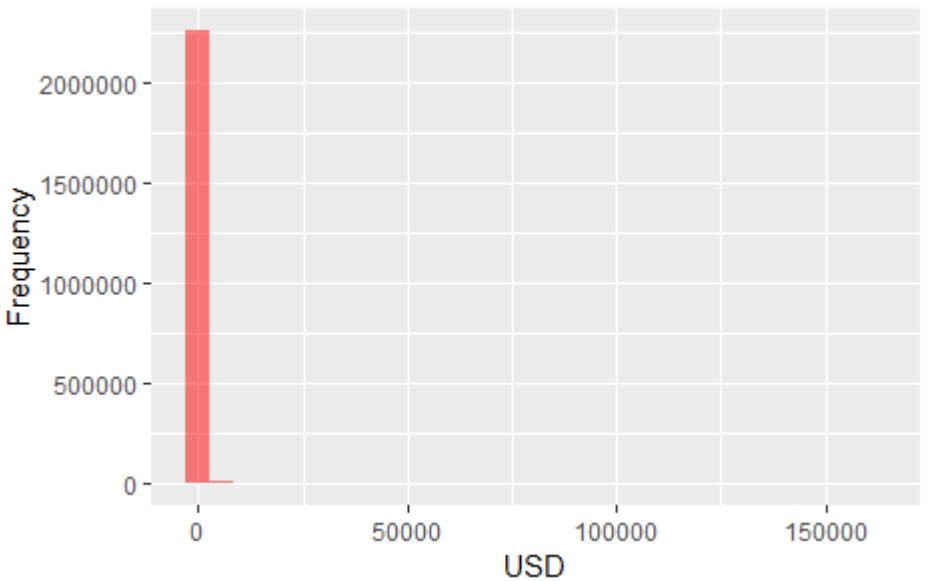
```
tips = sns.load_dataset('tips')
sns.pairplot(tips, height=1.5)
plt.show()
```

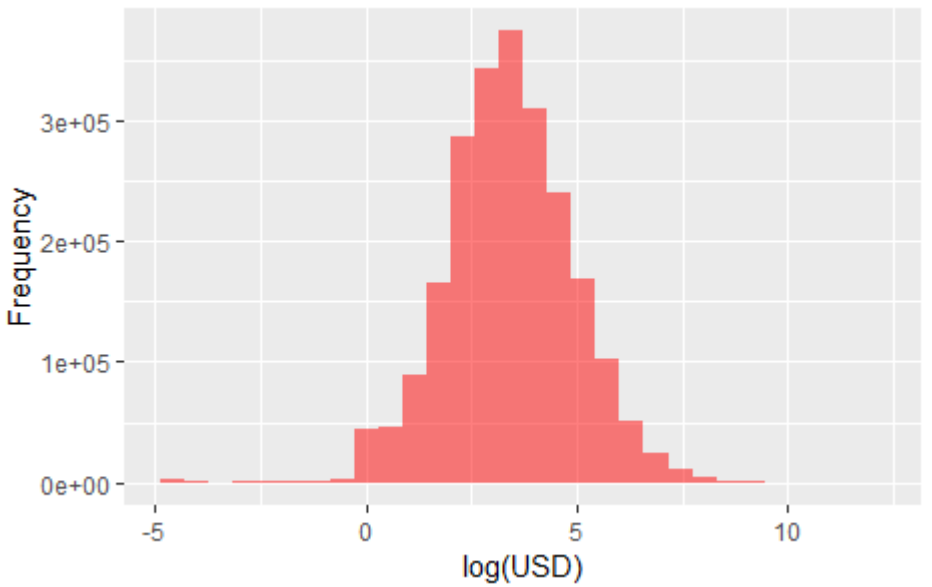


Some advice on cleaning data

Apply common transformations:

Here's how ebay's 2.3 million transactions look with a log transformation:

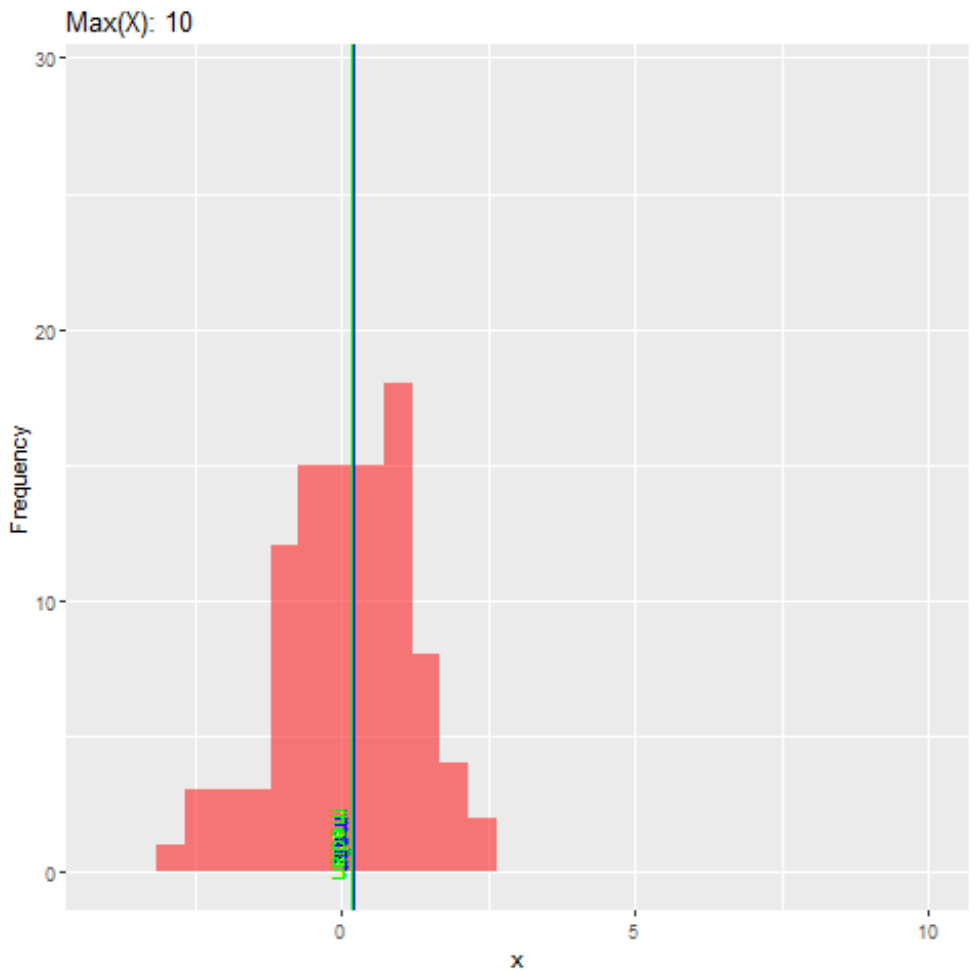




Some advice on cleaning data

Use robust statistics (an entire field in Statistics):

For example the Median is much more robust to extreme values than the mean:



Some advice on cleaning data

Tidy your data:

> Each variable is a column, each observation is a row, and each type of observational unit is a table. (Hadley Wickham)

Untidy (wide):

	SubjectName	Weight_Age20	Weight_Age30
1	John Doe	80	90
2	Jane Doe	60	50

Tidy (long):

	SubjectName	Age	Weight
1	John Doe	20	80
2	John Doe	30	90
3	Jane Doe	20	60
4	Jane Doe	30	50

Intro to Data Science