

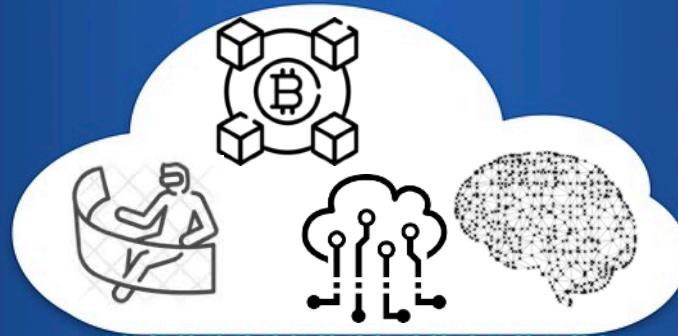
# **DATA HACK 2023 MAKER WORKSHOP:**

## **DATA SCRAPING AND DATA CLEANING**

Bernard Suen  
Center for Entrepreneurship  
Chinese University of Hong Kong

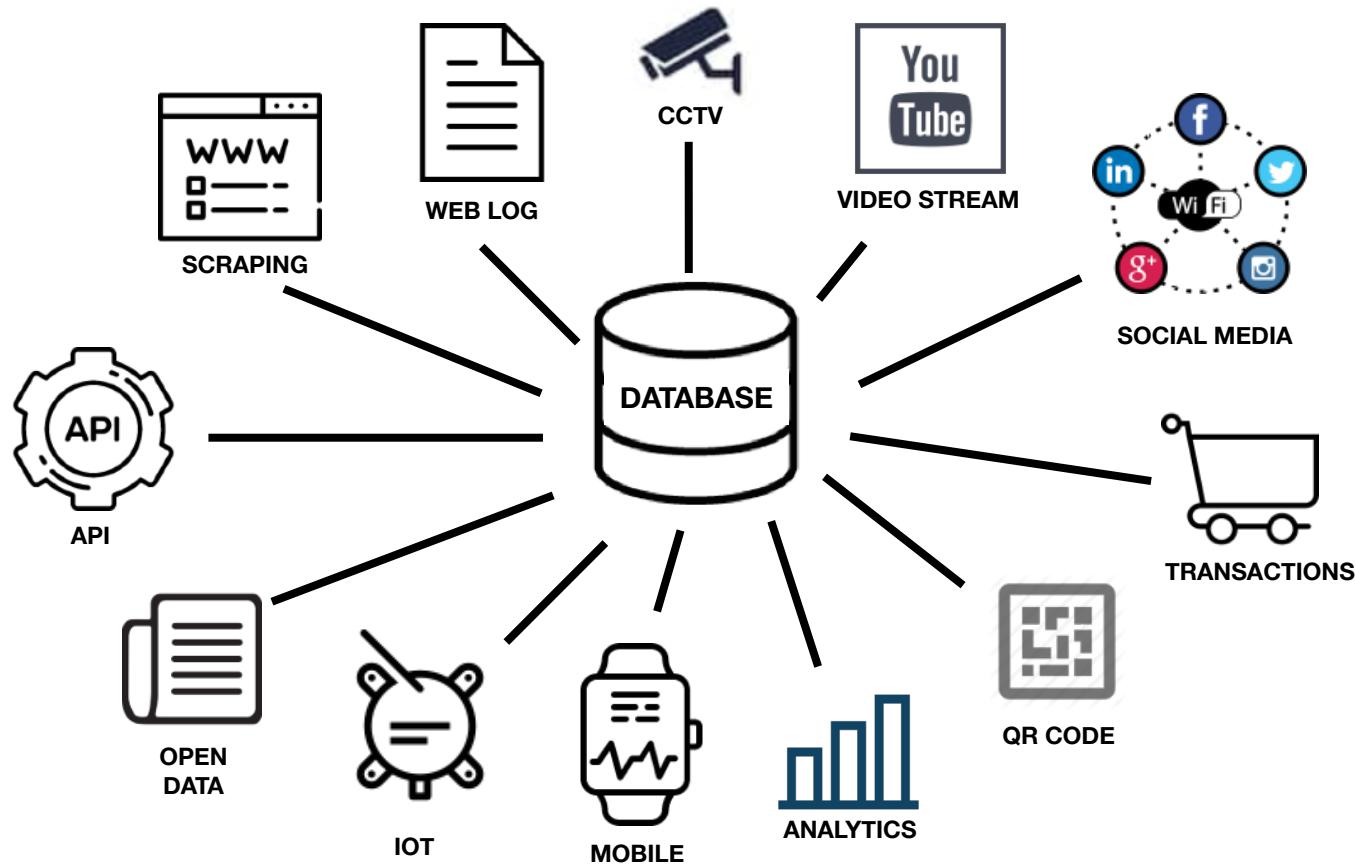
**Today's agenda.**

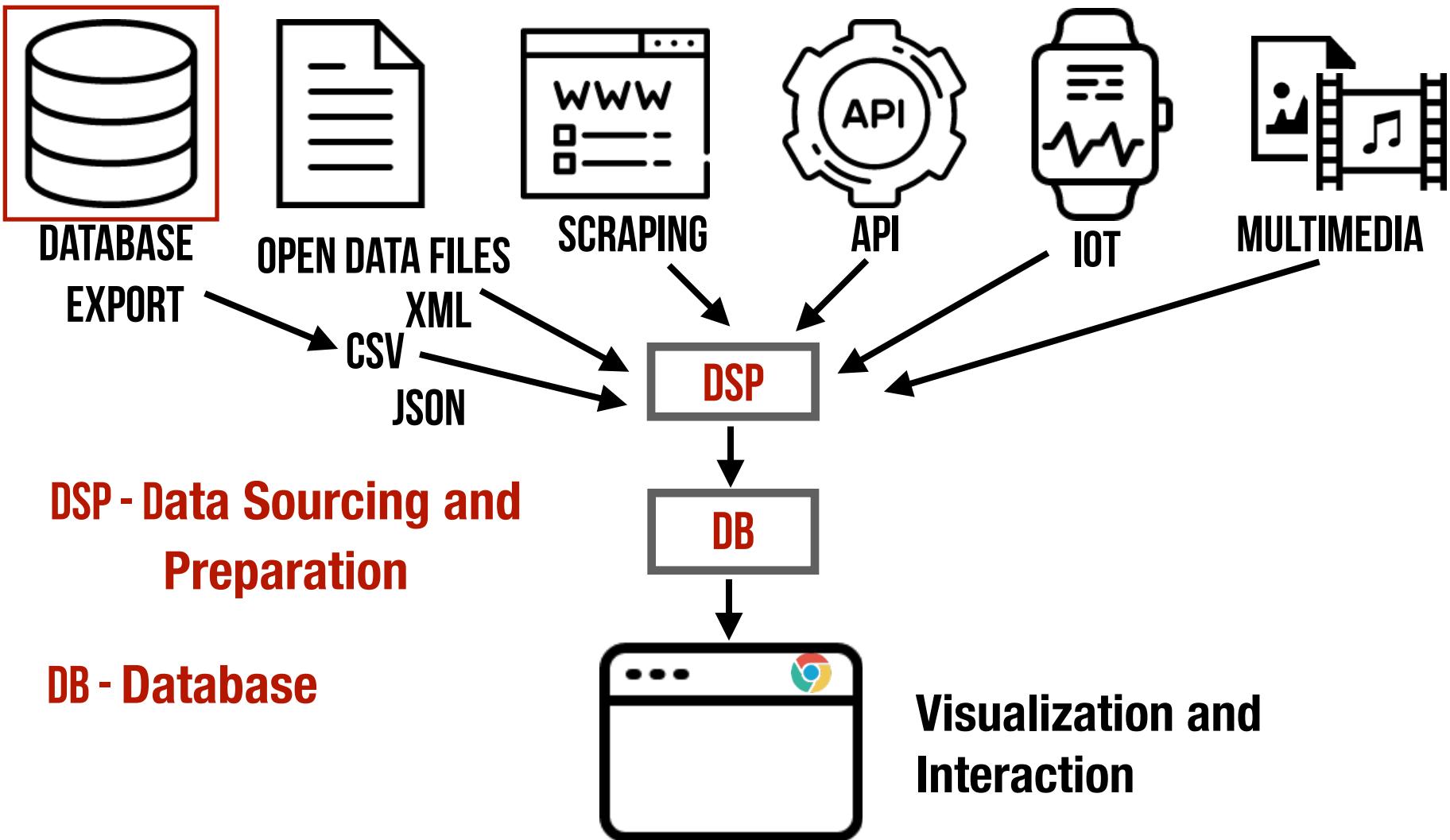
1. Most common **sources** of data
2. Data **scraping** with ParseHub
3. Data **cleaning** with OpenRefine



```
1 01 01 00 10 10 01 00 10 01 10 00 10 11 00 11 00 0  
1 01 11 11 10 11 00 11 10 10 01 10 10 11 01 10 11 10 0  
0 10 10 10 10 10 00 00 11 00 00 00 00 00 01 01 11 00 10 0  
0 11 01 11 00 00 01 00 11 11 00 11 11 01 10 00 10 00 1  
0 00 11 00 01 11 01 10 10 01 00 11 11 11 01 01 01 11 0  
1 11 10 01 01 01 00 01 01 00 01 00 11 00 00 01 10 0  
1 10 01 01 01 11 01 00 10 00 00 10 00 01 10 11 00 10 1  
1 00 10 00 01 00 10 11 01 00 00 10 10 10 01 10 01 10 0  
1 10 11 00 01 11 00 10 01 11 11 11 11 10 11 01 01 1  
0 00 11 00 11 10 10 00 00 11 01 11 10 11 01 10 01 01 0  
1 01 00 01 01 00 01 10 01 00 00 11 10 00 11 00 10 01 0  
1 10 10 00 10 10 10 11 10 01 10 01 11 00 01 11 00 01 1  
0 00 01 10 01 00 10 00 11 00 10 00 10 00 11 00 00 01 0  
1 01 11 01 01 10 11 00 00 11 10 01 10 00 00 01 10 01 1  
0 11 00 11 01 00 01 11 11 10 10 10 11 11 11 00 00 10 0  
1 11 01 10 11 11 01 00 01 00 01 11 00 00 01 11 00 00 0  
1 11 01 00 10 11 01 11 11 10 11 00 11 10 01 01 11 10 0  
0 01 01 00 10 11 11 00 11 01 00 10 10 00 00 00 11 11 1  
0 10 11 01 00 01 11 10 11 10 01 01 11 10 00 01 00 00 0  
0 01 11 01 10 10 01 11 00 00 11 00 00 01 10 01 00 01 10 1  
1 00 01 00 11 11 10 10 10 11 11 01 11 10 11 10 11 10 11 1  
1 10 11 11 10 00 10 01 01 00 11 11 00 11 01 10 10 00 0  
1 11 10 11 11 10 00 10 10 00 10 01 11 01 11 10 00 11 0  
1 10 01 00 10 01 11 00 10 00 10 01 01 01 01 11 10 10 00 1  
0 10 00 11 10 10 01 11 00 00 10 10 11 00 10 01 11 11 01 1  
0 00 01 11 00 00 01 11 00 10 00 10 00 00 01 01 11 01 11 00 0  
1 11 11 01 00 10 00 11 10 10 00 11 00 00 10 11 10 10 0  
1 01 11 11 11 11 01 00 11 01 11 00 11 01 10 00 00 01 1  
0 11 00 00 11 01 01 11 01 11 01 00 11 10 11 01 10 10 1  
1 10 10 11 00 00 00 10 00 10 10 00 11 00 11 10 11 11 10 1  
1 10 01 10 11 00 01 01 01 10 10 10 11 11 11 10 01 01 0  
1 11 10 01 01 00 11 01 01 11 00 10 11 00 11 01 10 00 1  
0 01 10 00 01 10 11 10 01 00 00 00 11 00 11 01 00 01 0  
1 11 10 11 10 00 00 11 00 00 11 11 00 11 01 11 01 11 0  
0 00 00 01 10 01 01 10 11 11 10 10 01 01 00 11 11 00 0  
0 11 00 00 00 11 10 01 00 00 00 10 11 11 11 01 01 0  
0 11 11 10 11 11 10 00 10 00 01 11 10 11 00 01 00 11 0  
0 11 01 00 01 00 11 10 10 10 10 00 10 11 11 00 00 01 11 1
```

# Structured and Un-structured Data Sources





- Dashboard
- Posts
- Media
- Post Grid Combo
- Pages
- Comments
- Restaurants
- Appearance
- Plugins
- Users
- Tools
- Settings
- Pods Admin
- All Export
- New Export
- Manage Exports
- Settings
- WP Data Access
- Ultimate CSV Importer Free
- FakerPress
- Collapse menu

## WP ALL EXPORT New Export

[Support](#) | [Documentation](#)

First, choose what to export.

 Specific Post Type

 WP\_Query Results

Choose a post type...

Created by  soflyy

- Dashboard
- Posts
- Media
- Post Grid Combo
- Pages
- Comments
- Restaurants
- Appearance
- Plugins
- Users
- Tools
- Settings
- Pods Admin

### All Export

- New Export
- Manage Exports
- Settings

### WP Data Access

- Ultimate CSV Importer Free

### FakerPress

Collapse menu

Thank you for creating with [WordPress](#).

Version 6.1.1

Choose a post type...

Choose a post type...

Posts

Pages

Taxonomies

Comments

Users

Global Styles

Layouts

Lottie Animations

Navigation Menus

Pod Fields

Pod Groups

Pod Templates

Pods

Post Grid

Restaurants

Saved Templates

Template Parts

Templates

W Datahack 2023 0 + New Howdy, admin

WP ALL EXPORT Drag & Drop

Support | Documentation

302 Restaurants will be exported

Drag & drop data to include in the export file.

ID Title Content

Add Field Add All Clear All Preview

Available Data

Standard Media Custom Fields Other ACF

All Export

New Export Manage Exports Settings

WP Data Access

Ultimate CSV Importer Free

FakerPress

Collapse menu

Advanced Options

Export Type

Save settings as a template Load Template... ▾

< Back Continue >

W Datahack 2023 0 + New

WP ALL EXPORT

# Confirm & Run

Support | Documentation

Export Complete

Time Elapsed 00:00:04

100%

Exported 302

What's next?

Download Scheduling External Apps Export, Edit, Import

Click to Download

CSV Bundle

The bundle contains your exported data and a settings file for WP All Import. Upload the Bundle to WP All Import on another site to quickly import this data.

Public URL

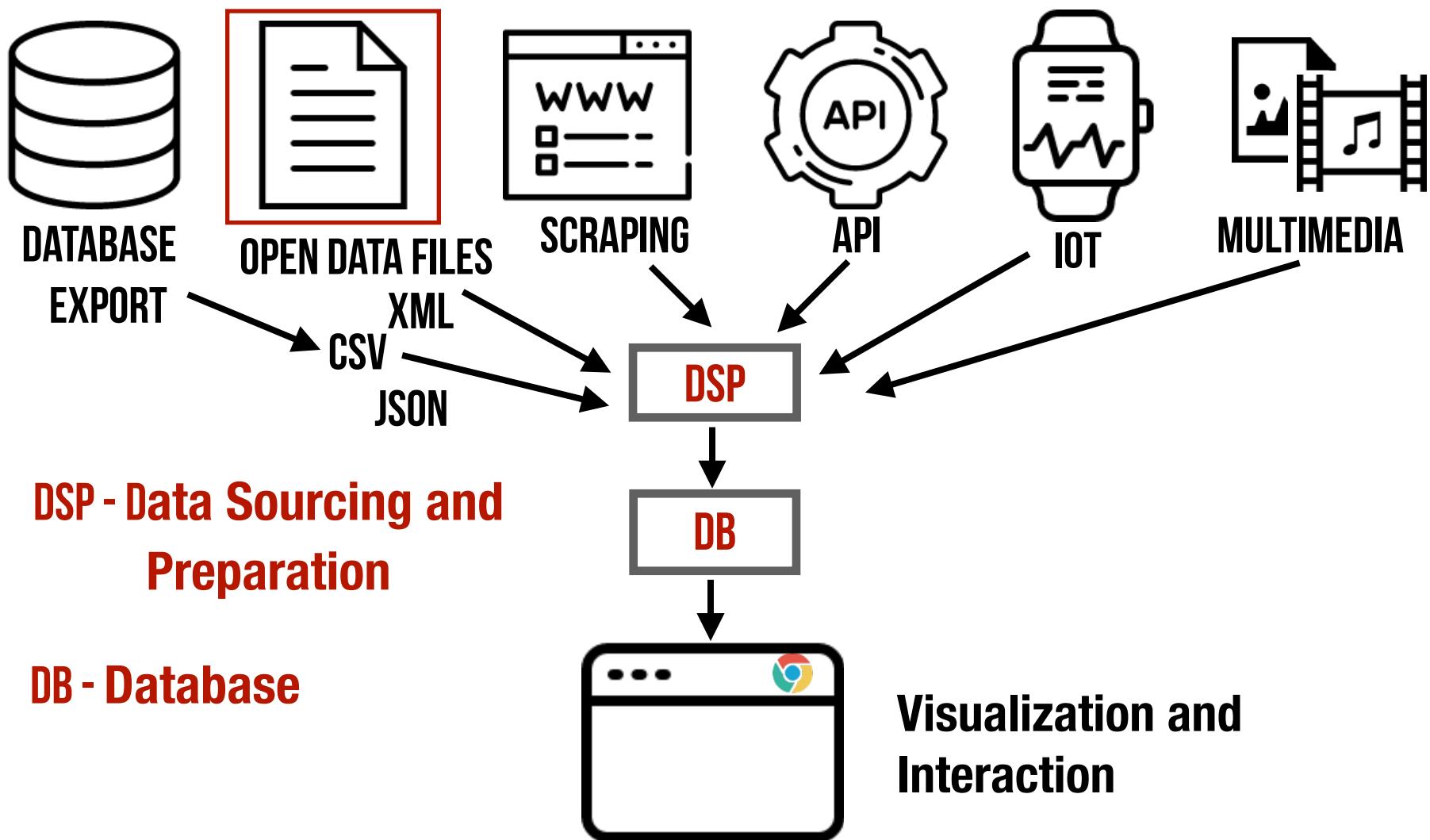
[https://dev-hackathon-2023.pantheonsite.io/wp-load.php?  
security\\_token=c6650df68d693461&export\\_id=1&action=get\\_data](https://dev-hackathon-2023.pantheonsite.io/wp-load.php?security_token=c6650df68d693461&export_id=1&action=get_data)

This URL will always provide the export file from this export, even if the file name changes.

Created by  soflyy

Thank you for creating with [WordPress](#).

Version 6.1.1



## 2020-2022 Annual Open Data Plans Are Now Published



New to The Site? [START HERE](#)

Search Data e.g. population 

BROWSE DATASETS

<https://data.gov.hk/en/>

DATASETS

GEOSPATIAL DATA



# CITY DASHBOARD

City dynamics at a glance

New to The Site? **START HERE**

Search Data e.g. population

**BROWSE DATASETS**<https://data.gov.hk/en-data/dataset/hk-lcsd-facility-facility-bkbc>



HOME DATA LEARN COMMUNITY



ENG ▾

Home > Datasets > Basketball Courts (Free Outdoor Pitches/Courts)

## Basketball Courts (Free Outdoor Pitches/Courts)

LEISURE AND CULTURAL SERVICES DEPARTMENT |

Recreation and Culture



UPDATE FREQUENCY: AS AND WHEN NEW FACILITY IS ADDED OR AMENDMENT IS MADE

Location of Basketball Courts (Free Outdoor Pitches/Courts)

---

Data Dictionary : [https://www.lcsd.gov.hk/datagovhk/facility/facility-bkbc\\_data\\_dictionary.pdf](https://www.lcsd.gov.hk/datagovhk/facility/facility-bkbc_data_dictionary.pdf)

---

1 JSON file(s)



Add All to Queue

[{"District\_en": "Kowloon City", "District\_cn": "九龍城區", "Name\_en": "Carpenter Road Park", "Name\_cn": "賈炳達道公園", "Address\_en": "Carpenter Road, Kowloon City, Kowloon.", "Address\_cn": "九龍城賈炳達道", "GIHS": "9G5i7NFpXL", "Court\_no\_en": "4", "Court\_no\_cn": "4", "Ancillary\_facilities\_en": "<li>Toilets are provided in the Park.<br><li>Other facilities include a jogging track, 7 fitness stations, a cycling track, 2 sets of children's playground, an elderly fitness station, a 7-a-side hard-surface soccer pitch and 2 volleyball courts.<br><li>Barrier Free Facilities: Accessible Toilet, Tactile Guide Path, Braille Directory Map\\Floor Plan<br>", "Ancillary\_facilities\_cn": "<li>公園內設有洗手間<br><li>其他設施包括1條緩跑徑、7個健身站、1條單車徑、2個兒童遊樂場、1個長者健身站、1個7人硬地足球場和2個排球場<br><li>無障礙設施: 暢通易達洗手間、觸覺引路帶、觸覺點字及觸覺平面圖。<br>", "Opening\_hours\_en": "7 am to 11 pm daily", "Opening\_hours\_cn": "每日上午7時至晚上11時", "Phone": "2716 9962", "Remarks\_en": "", "Remarks\_cn": "", "Longitude": "114-11-27", "Latitude": "22-19-52"}, {"District\_en": "Kowloon City", "District\_cn": "九龍城區", "Name\_en": "Ho Man Tin Park", "Name\_cn": "何文田公園", "Address\_en": "No.1 Chung Yee Street, Ho Man Tin, Kowloon.", "Address\_cn": "九龍何文田忠義街一號", "GIHS": "MSKgwBPmtd", "Court\_no\_en": "2", "Court\_no\_cn": "2", "Ancillary\_facilities\_en": "<li>Men's and ladies' changing rooms and toilets<br><li>A fee-charging car park (including 1 designated disabled parking space)<br><li>Other facilities include a hard-surface 7-a-side soccer pitch cum handball court, a children's playground and a jogging track with 6 fitness stations.<br><li>Barrier Free Facilities: Accessible Toilet, Tactile Guide Path, Braille Directory Map\\Floor Plan", "Ancillary\_facilities\_cn": "<li>男、女更衣室及洗手間<br><li>1個收費停車場 (設有1個殘疾人士專用車位)<br><li>其他設施包括1個硬地7人足球場兼手球場、1個兒童遊樂場、1條緩跑徑和6個健身站<br><li>無障礙設施: 暢通易達洗手間、觸覺引路帶、觸覺點字及觸覺平面圖", "Opening\_hours\_en": "7 am to 11 pm daily", "Opening\_hours\_cn": "每日上午7時至晚上11時", "Phone": "2762 7837", "Remarks\_en": "", "Remarks\_cn": "", "Longitude": "114-10-50", "Latitude": "22-18-44"}, {"District\_en": "Kowloon City", "District\_cn": "九龍城區", "Name\_en": "Hoi Sham Park", "Name\_cn": "海心公園", "Address\_en": "Yuk Yat Street, Tokwawan, Kowloon.", "Address\_cn": "九龍土瓜灣旭日街", "GIHS": "nZ2deVMDpF", "Court\_no\_en": "1", "Court\_no\_cn": "1", "Ancillary\_facilities\_en": "<li>Men's and ladies' toilets<br><li>Other facilities include 2 hard-surface 5-a-side soccer, a children's playground and elderly fitness equipment.<br><li>Barrier Free Facilities: Accessible Toilet, Tactile Guide Path, Braille Directory Map\\Floor Plan", "Ancillary\_facilities\_cn": "<li>男、女洗手間<br><li>其他設施包括2個硬地5人足球場、兒童遊樂場及長者健體設施<br><li>無障礙設施: 暢通易達洗手間、觸覺引路帶、觸覺點字及觸覺平面圖", "Opening\_hours\_en": "7 am to 11 pm daily", "Opening\_hours\_cn": "每日上午7時至晚上11時", "Phone": "2334 3576 \ 2762 2083", "Remarks\_en": "", "Remarks\_cn": "", "Longitude": "114-11-30", "Latitude": "22-18-54"}, {"District\_en": "Kowloon City", "District\_cn": "九龍城區", "Name\_en": "Junction Road Park", "Name\_cn": "聯合道公園", "Address\_en": "Junction Road, Kowloon City, Kowloon.", "Address\_cn": "九龍聯合道", "GIHS": "9G5i7NFpXL", "Court\_no\_en": "4", "Court\_no\_cn": "4", "Ancillary\_facilities\_en": "<li>Toilets are provided in the Park.<br><li>Other facilities include a jogging track, 7 fitness stations, a cycling track, 2 sets of children's playground, an elderly fitness station, a 7-a-side hard-surface soccer pitch and 2 volleyball courts.<br><li>Barrier Free Facilities: Accessible Toilet, Tactile Guide Path, Braille Directory Map\\Floor Plan<br>", "Ancillary\_facilities\_cn": "<li>公園內設有洗手間<br><li>其他設施包括1條緩跑徑、7個健身站、1條單車徑、2個兒童遊樂場、1個長者健身站、1個7人硬地足球場和2個排球場<br><li>無障礙設施: 暢通易達洗手間、觸覺引路帶、觸覺點字及觸覺平面圖。<br>", "Opening\_hours\_en": "7 am to 11 pm daily", "Opening\_hours\_cn": "每日上午7時至晚上11時", "Phone": "2716 9962", "Remarks\_en": "", "Remarks\_cn": "", "Longitude": "114-11-27", "Latitude": "22-19-52"}]

- Convert JSON to CSV (<https://codebeautify.org/jsonviewer>)
- Convert XML to CSV (<https://www.convertcsv.com/xml-to-csv.htm>)
- Import CSV into Excel or Numbers

JSON Viewer★

Save & Share

Sample

```
1 [{"District_en": "Kowloon City", "District_cn": "九龍城區",  
  "Name_en": "Carpenter Road Park", "Name_cn": "賈炳達道公園",  
  "Address_en": "Carpenter Road, Kowloon City, Kowloon.",  
  "Address_cn": "九龍城賈炳達道", "GIHS": "9G5i7NFpXL",  
  "Court_no_en": "4", "Court_no_cn": "4",  
  "Ancillary_facilities_en": "<li>Toilets are provided in the Park.<br><li>Other facilities include a jogging track, 7 fitness stations, a cycling track, 2 sets of children's playground, an elderly fitness station, a 7-a-side hard-surface soccer pitch and 2 volleyball courts.<br><li>Barrier Free Facilities: Accessible Toilet, Tactile Guide Path, Braille Directory Map</Floor Plan<br>"  
  "Ancillary_facilities_cn": "<li>公園內設有洗手間<br><li>其他設施包括1條緩跑徑、7個健身站、1條單車徑、2個兒童遊樂場、1個長者健身站、1個7人硬地足球場和2個排球場<br><li>無障礙設施：暢通易達洗手間、觸覺引路帶、觸覺點字及觸覺平面圖。<br>", "Opening_hours_en": "7 am to 11 pm daily",  
  "Opening_hours_cn": "每日上午7時至晚上11時", "Phone": "2716 9962", "Remarks_en": "", "Remarks_cn": "", "Longitude": "114-11-27", "Latitude": "22-19-52"}, {"District_en": "Kowloon City", "District_cn": "九龍城區", "Name_en": "Ho Man Tin Park", "Name_cn": "何文田公園", "Address_en": "No.1 Chung Yee Street, Ho Man Tin, Kowloon.", "Address_cn": "九龍何文田忠義街一號", "GIHS": "MSKgWBPMtd", "Court_no_en": "2", "Court_no_cn": "2", "Ancillary_facilities_en": "<li>Men's and ladies' changing rooms and toilets<br><li>A fee-charging car park (including 1 designated disabled parking space)<br><li>Other facilities include a hard-surface 7-a-side soccer pitch cum handball court, a children's playground and a jogging track with 6 fitness stations.<br><li>Barrier Free Facilities: Accessible Toilet, Tactile Guide Path, Braille Directory Map</Floor Plan", "Opening_hours_en": "7 am to 11 pm daily", "Opening_hours_cn": "每日上午7時至晚上11時", "Phone": "2716 9962", "Remarks_en": "", "Remarks_cn": "", "Longitude": "114-11-27", "Latitude": "22-19-52"}]
```

**Result mode:**

**view** ▾

---

**Load Url**

---

**Browse**

---

**Tree Viewer**

---

2 Tab Space ▾

---

**Beautify**

---

(i) X

# Website Designer Near You

Stallions SEO LLC

We want to make you  
businesses phone ring!

---

**OPEN**

---

**Minify**

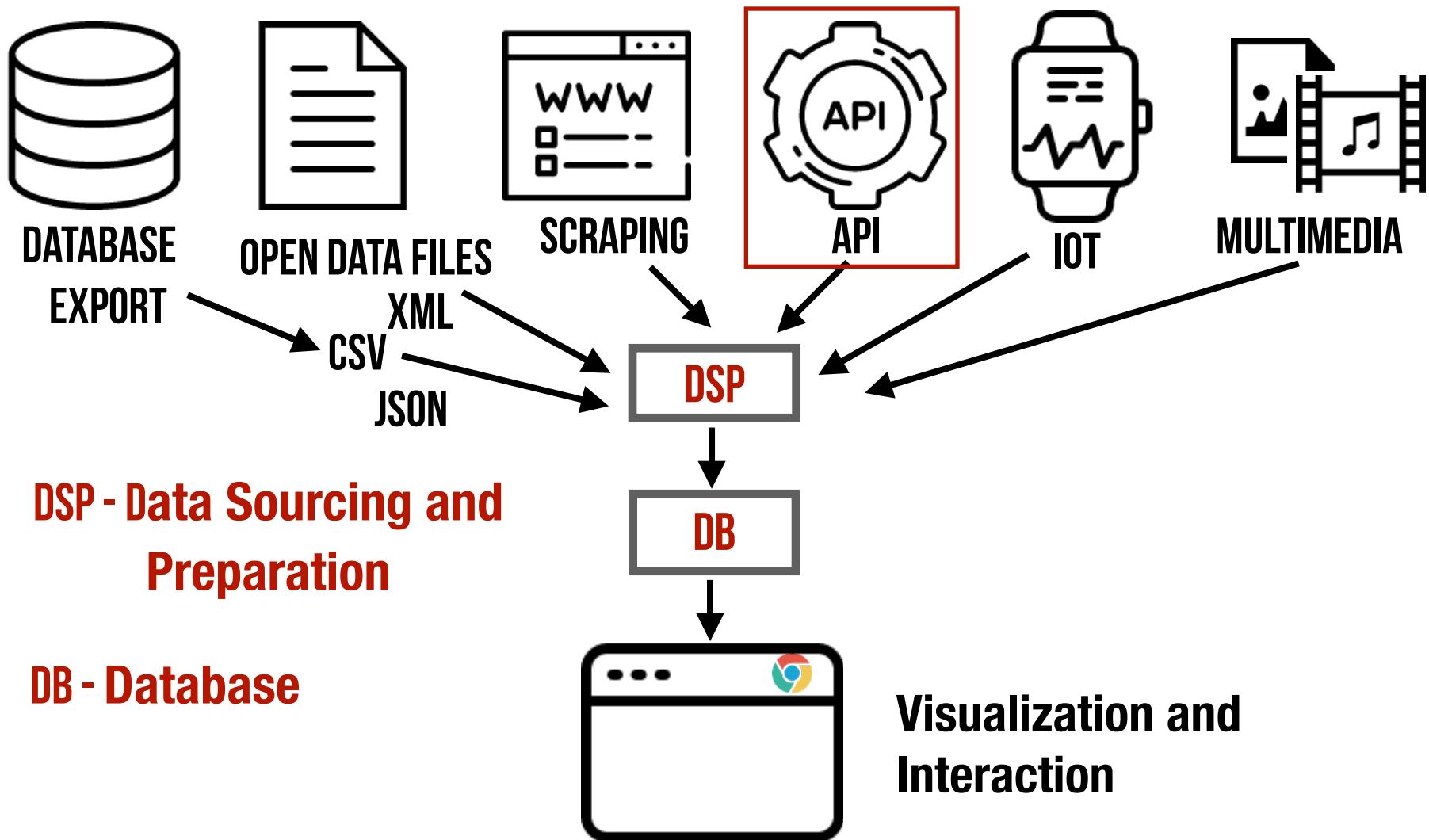
---

**Validate**

```
array > 0 >
▼ array [228]
  ▼ 0 {18}
    District_en : Kowloon City
    District_cn : 九龍城區
    Name_en : Carpenter Road Park
    Name_cn : 賈炳達道公園
    Address_en : Carpenter Road, Kowloon City, Kowloon.
    Address_cn : 九龍城賈炳達道
    GIHS : 9G5i7NFpXL
    Court_no_en : 4
    Court_no_cn : 4
    Ancillary_facilities_en : <li>Toilets are provided in the Park.<br><li>Other facilities include a jogging track, 7 fitness stations, a cycling track, 2 sets of children's playground, an elderly fitness station, a 7-a-side hard-surface soccer pitch and 2 volleyball courts.<br><li>Barrier Free Facilities: Accessible Toilet, Tactile Guide Path, Braille Directory Map/Floor Plan<br>
    Ancillary_facilities_cn : <li>公園內設有洗手間<br><li>其他設施包括1條緩跑徑、7個健身站、1條單車徑、2個兒童遊樂場、1個長者健身站、1個7人硬地足球場和2個排球場<br><li>無障礙設施 暢通易達洗手間、觸覺引路帶、輪椅通道
```

<https://codebeautify.org/jsonviewer>

District_en	District_cn	Name_en	Name_cn	Address_en
Kowloon City	九龍城區	Carpenter Road Park	賈炳達道公園	Carpenter Road, Kowloon City, Kowloon.
Kowloon City	九龍城區	Ho Man Tin Park	何文田公園	No.1 Chung Yee Street, Ho Man Tin, Kowloon.
Kowloon City	九龍城區	Hoi Sham Park	海心公園	Yuk Yat Street, Tokwawan, Kowloon.
Kowloon City	九龍城區	Junction Road Park	聯合道公園	195 Junction Road, Kowloon City.
Kowloon City	九龍城區	Kam Shing Road Recreation Ground	金城道遊樂場	Kam Shing Road, Kowloon.
Kowloon City	九龍城區	Kau Pui Lung Road Playground	靠背壘道遊樂場	Kau Pui Lung Road, Kowloon.
Kowloon City	九龍城區	Kent Road Garden	根德道花園	Kent Road, Kowloon.
Kowloon City	九龍城區	King Wan Street Playground	景雲街遊樂場	King Wan Street, To Kwa Wan, Kowloon.
Kowloon City	九龍城區	King's Park High Level Service Reservoir Playground	京士柏上配水庫遊樂場	Chung Hau Street, Ho Man Tin, Kowloon.
Kowloon City	九龍城區	Kowloon Tsai Park	九龍仔公園	13 Inverness Road, Kowloon City, Kowloon.
Kowloon City	九龍城區	Lung Cheung Road Playground	龍翔道遊樂場	Beacon Hill Road, Kowloon.
Kowloon City	九龍城區	Ma Tau Wai Road Playground	馬頭圍道遊樂場	Ma Tau Wai Road, Tokwawan, Kowloon.
Kowloon City	九龍城區	Oxford Road Playground	牛津道遊樂場	Oxford Road, Kowloon.
Kowloon City	九龍城區	Peace Avenue Playground	太平道遊樂場	Peace Avenue, Ho Man Tin, Kowloon.
Kowloon City	九龍城區	Perth Street Sports Ground	巴富街運動場	Shek Ku Street , Ho Man Tin, Kowloon.
Kowloon City	九龍城區	Pui Ching Road Playground	培正道遊樂場	Pui Ching Road, Ho Man Tin, Kowloon.
Kowloon City	九龍城區	Rutland Quadrant Children's Playground	律倫街兒童遊樂場	Rutland Quadrant, Kowloon.
Kowloon City	九龍城區	Sung Wong Toi Playground	宋王臺遊樂場	Sung Wong Toi Road, Kowloon.
Kowloon City	九龍城區	Tai Wan Road Playground	大環道遊樂場	Tai Wan Road, Hung Hom, Kowloon.
Kowloon City	九龍城區	Tai Wan Shan Park	大環山公園	Wan Hoi Street, Hung Hom, Kowloon.
Kowloon City	九龍城區	To Kwa Wan Recreation Ground	土瓜灣遊樂場	66 Ha Ha Ling Road, To Kwa Wan, Kowloon





百度地图  
开放平台

首页

功能与服务

解决方案

开发文档

合作咨询

控制台

登录

# 百度地图发布全球API试用版

拉开海外地图开发市场序幕 · 全球POI超过1.5亿 · 覆盖209个国家和地区



Android定位SDK v7.4 **NEW!**  
优化离线定位内部策略



2018年春运出行仪表盘 **NEW!**  
春运综合出行信息播报



iOS地图SDK v3.4.4  
适配 iOS 11 定位



iOS定位SDK v1.1  
适配 iOS 11 永久定位的设置



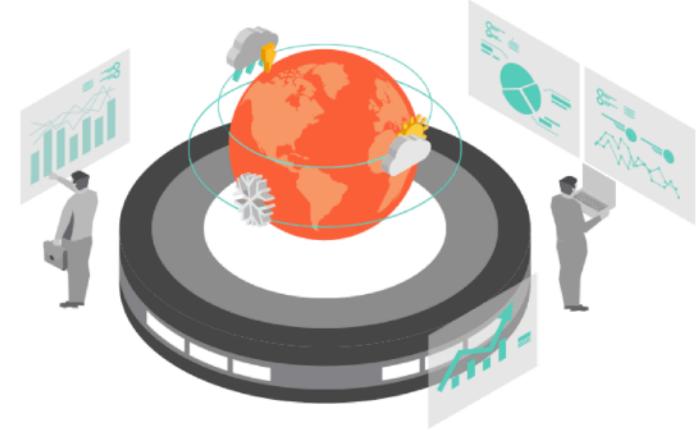
# **API Demonstration (in JSON format)**

# OpenWeather global services

Weather forecasts, nowcasts and history in fast and elegant way

2 Billion Forecasts Per Day  
2,500 new subscribers a day

2,600,000 customers  
20+ weather APIs



Search city

Search



Different Weather?

Metric: °C, m/s

Imperial: °F, mph

3:39pm, Sep 12

London, GB



<https://openweathermap.org/>



Weather in your city

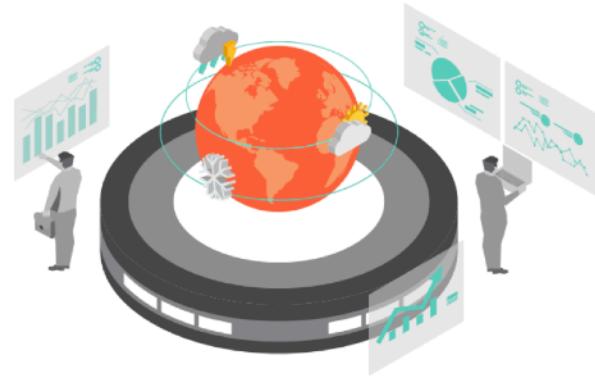
Get Started API Pricing Maps FAQ Partners Blog Marketplace Sign in Support

# OpenWeather global services

Weather forecasts, nowcasts and history in fast and elegant way

2 Billion Forecasts Per Day  
2,500 new subscribers a day

2,600,000 customers  
20+ weather APIs



Search city

Search



Different Weather?

Metric: °C, m/s Imperial: °F, mph

3:39pm, Sep 12

London, GB

21°C





Weather in your city

Guide API Dashboard Marketplace Pricing Maps Our Initiatives Partners Blog For Business support ▾

format.

## Call current weather data

### How to make an API call

#### API call

```
https://api.openweathermap.org/data/2.5/weather?lat={lat}&lon={lon}&appid={API key}
```



#### Parameters

**lat, lon** required Geographical coordinates (latitude, longitude). If you need the geocoder to automatically convert city names and zip-codes to geo coordinates and the other way around, please use

[Geocoding API](#)

<https://openweathermap.org/current>

Call current weather data

[How to make an API call](#)

Bulk downloading

Weather fields in API response

JSON

XML

List of condition codes

Min/max temperature in current weather API and forecast API

Other features

Geocoding API

Built-in geocoding

Built-in API request by city name

Built-in API request by city ID

Built-in API request by ZIP code

Format

Units of measurement

Multilingual support

Call back function for JavaScript code



Weather in your city

Get Started API Pricing Maps FAQ Partners Blog Marketplace support ▾ Support

## By city name

Description:

You can call by city name or city name, state code and country code. API responds with a list of weather parameters that match a search request.

API call:

`api.openweathermap.org/data/2.5/weather?q={city name}&appid={your api key}`

`api.openweathermap.org/data/2.5/weather?q={city name},{state code}&appid={your api key}`

`api.openweathermap.org/data/2.5/weather?q={city name},{state code},{country code}&appid={your api key}`

Parameters:

**q** city name, state code and country code divided by comma, use ISO 3166 country codes. You can specify the parameter not only in English. In this case, the API response should be returned in the same language as the language of requested location name if the location is in our predefined list of more than 200,000 locations.

Examples of API calls:

`api.openweathermap.org/data/2.5/weather?q=London`

`api.openweathermap.org/data/2.5/weather?q=London,uk`

Searching by states available only for the USA locations.

There is a possibility to receive a central district of the city/town with its own parameters

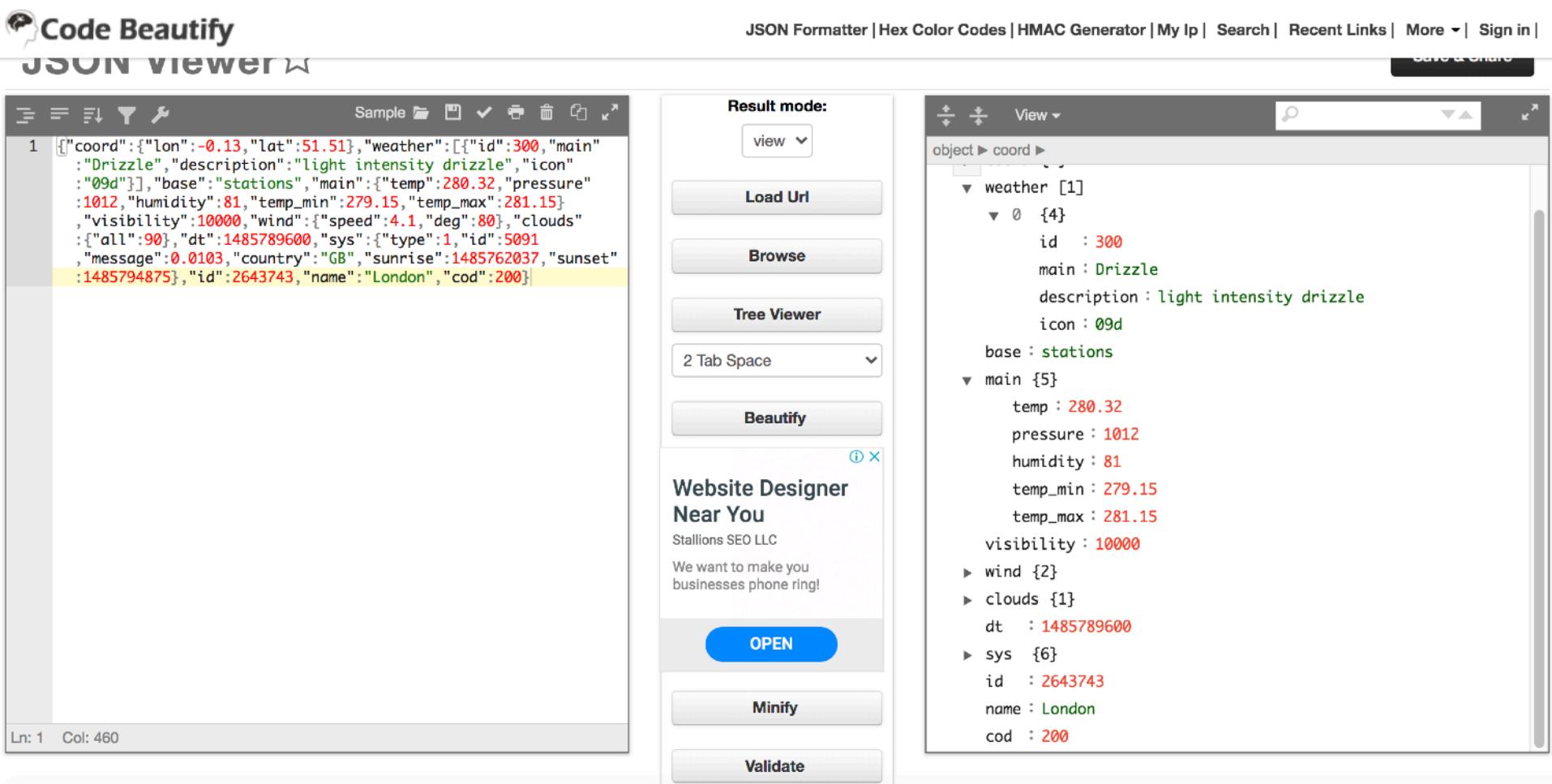
- List of condition codes
- Min/max temperature in current weather API and forecast API
- Other features
  - Format
  - Units format
  - Multilingual support
  - Call back function for JavaScript code

\*\*\*

← → ⌂ https://api.openweathermap.org/data/2.5/weather?q=London&appid=xxxxxxxxxxxxxx

```
{"coord":{"lon":-0.1257,"lat":51.5085}, "weather":[{"id":804,"main":"Clouds","description":"overcast clouds","icon":"04n"}], "base":"stations", "main": {"temp":282.23,"feels_like":280.5,"temp_min":280.8,"temp_max":283.76,"pressure":1034,"humidity":87}, "visibility":10000, "wind": {"speed":3.09,"deg":270}, "clouds": {"all":100}, "dt":1675460291, "sys": {"type":2,"id":2075535,"country":"GB","sunrise":1675409784,"sunset":1675443126}, "timezone":0, "id":2643743, "name":"London", "cod":200}
```

```
{"coord":{"lon":-0.1257,"lat":51.5085}, "weather":  
[{"id":804,"main":"Clouds","description":"overcast  
clouds","icon":"04n"}], "base":"stations", "main":  
{"temp":282.23,"feels_like":280.5,"temp_min":280.8,"temp_max":283.76,"press  
ure":1034,"humidity":87}, "visibility":10000, "wind":  
{"speed":3.09,"deg":270}, "clouds":{"all":100}, "dt":1675460291, "sys":  
{"type":2,"id":2075535,"country":"GB","sunrise":1675409784,"sunset":1675443  
126}, "timezone":0, "id":2643743, "name": "London", "cod":200}
```



<https://codebeautify.org/jsonviewer>

## Weather API

City:

Paris

Submit

overcast clouds

17°

Paris

## Weather API

City:

Shanghai

Submit

light rain

25°

Shanghai

## Weather API

City:

Hong Kong

Submit

light rain

30°

Hong Kong



## Meituan API Endpoints Example

<https://developer.meituan.com/ka/docs/wmapi-1>

到餐API文档

> 团购API

> 茶饮版 API

智能版 API

门店-获取营业日起...

获取授权资源ID列表

门店-获取物品（分...

总部-获取物品（分...

门店-获取物品类别...

总部-获取物品类别...

门店-获取供应商（...

总部-获取供应商（...

门店-获取物品单位...

总部-获取物品单位 ...

### HTTP请求示例

```
(提示: 实际请求需要进行 urlencode 编码)
POST /rms/base/v1/auth/resources/get HTTP/1.1
Host: api-open-cater.meituan.com
Content-Type: application/x-www-form-urlencoded; charset=utf-8

appAuthToken=qwe&
timestamp=123&
sign=sdsadfsfd&
charset=utf-8&
developerId=100&
version=1&
biz={}
```

### REQUEST

### 响应示例

#### Json示例

```
{
  "code": "OP_SUCCESS",
  "msg": "成功",
  "traceId": 123,
  "data": {
    "resources": {
      "orgType": "1 总部, 5 门店",
      "rootOrgId": 123,
      "orgId": 123
    }
  }
}
```

### RESPONSE

评价管理

外卖门店在延迟发配送名单内才能设置延迟时间

活动管理

订单管理

订单推送

取消订单推送

退款信息推送

图片上传

门店装修

安心卡

在线联系IM

接口code码集合

&gt; 到餐API文档

&gt; 茶饮版 API

API接口	接口描述
poi/save	创建或更新门店信息
poi/getids	获取门店ID
poi/mget	批量获取门店详细信息
poi/close	门店设置为休息状态
poi/online	门店设置为上线状态
poi/updatepromoteinfo	更改门店公告信息
poiTag/list	获取门店品类列表
shippingtime/update	更新门店营业时间
poi/logistics/isDelayPush	查询门店是否延迟发配送
poi/logistics/setDelayPush	设置门店延迟发配送时间



## Developer Resources

Search REST API



### CHAPTERS

[REST API Handbook](#)[Key Concepts](#)[Frequently Asked Questions](#)[Using the REST API](#)[Extending the REST API](#)[Endpoint Reference](#)[Glossary](#)[Changelog](#)

Browse: Home / REST API Handbook

## REST API Handbook



The WordPress REST API provides an interface for applications to interact with your WordPress site by sending and receiving data as [JSON](#) (JavaScript Object Notation) objects. It is the foundation of the [WordPress Block Editor](#), and can likewise enable your theme, plugin or custom application to present new, powerful interfaces for managing and publishing your site content.

Using the WordPress REST API you can create a plugin to provide an entirely new admin experience for WordPress, build a brand new interactive front-end experience, or bring your WordPress content into completely separate applications.

The REST API is a developer-oriented feature of WordPress. It provides data access to the content of your site, and implements the same authentication restrictions — content that is public on your site is generally publicly accessible via the REST API, while private content, password-protected content, internal users, custom post types, and metadata is only available with authentication or if you specifically set it to be so. If you are not a developer, the most important thing to understand about the API is that it enables the block editor and modern plugin interfaces without compromising the security or privacy of your site.

### TOPICS

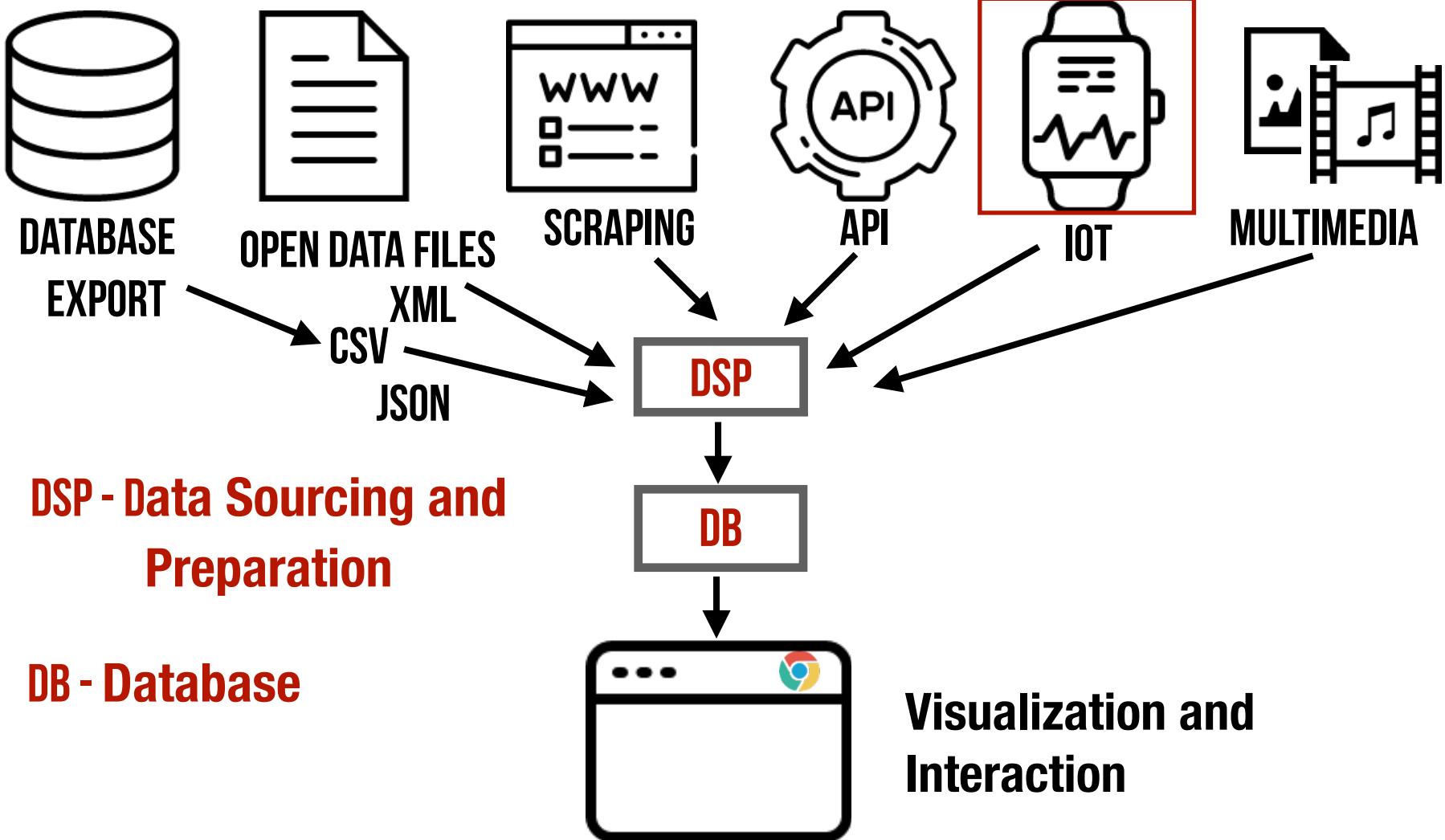
[What Is A REST API?](#)[Using the WordPress REST API](#)[Next Steps](#)

## 🔗 What Is A REST API?

An API is an Application Programming Interface. REST, standing for "REpresentational State Transfer," is a set of concepts for modeling and accessing your application's data as interrelated objects and collections. The WordPress REST API provides REST endpoints (URLs) representing the posts, pages, taxonomies, and other built-in WordPress data types. Your application can send and receive JSON data to these endpoints to query, modify and create content on your site. JSON is an open standard data format that is lightweight and human-

<https://dev-hackathon-2023.pantheonsite.io/wp-json/wp/v2/restaurant>

<https://codebeautify.org/jsonviewer>



# Popular IoT Prototyping Boards



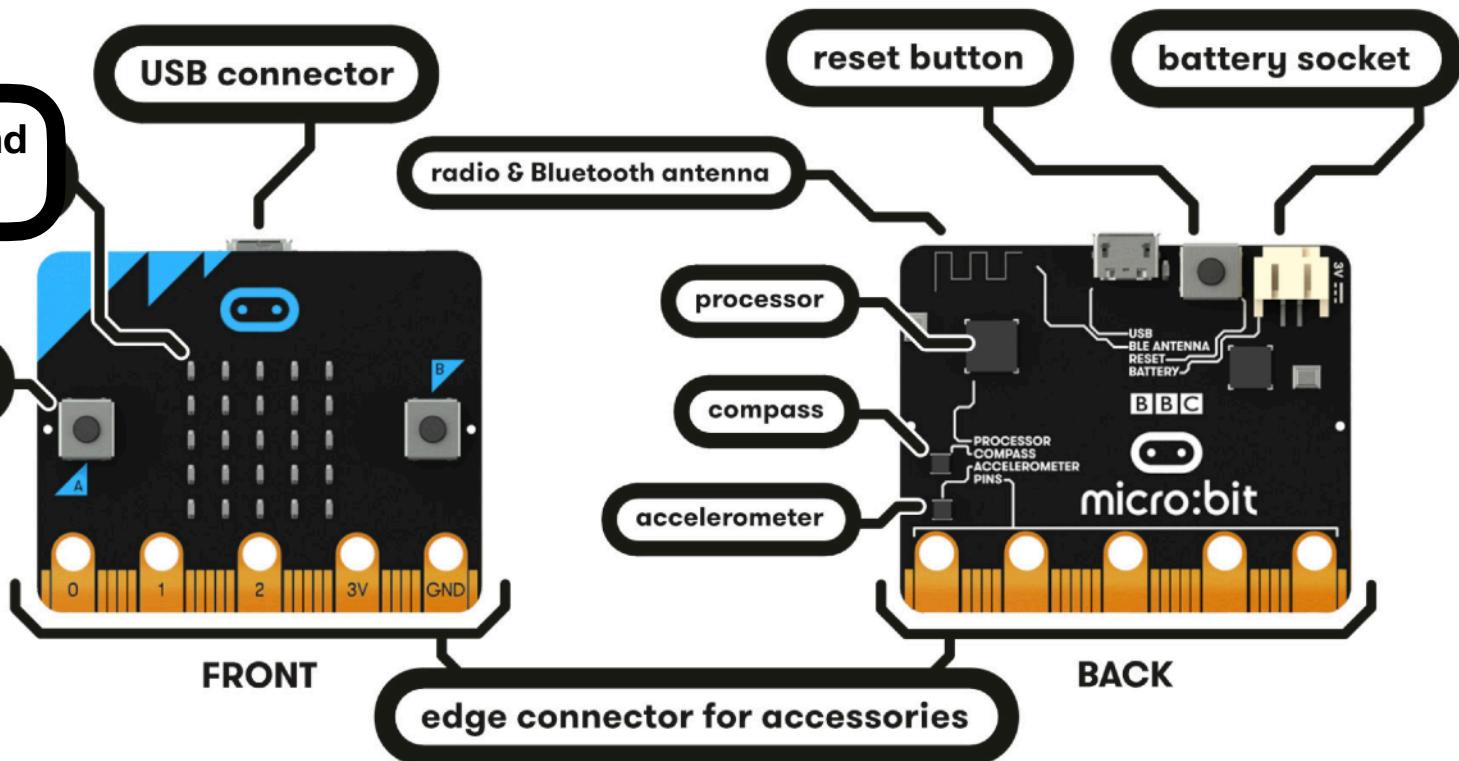
Adruino



Raspberry Pi



Micro:bit



## Basic Structure & Components of Micro:bit

source: <https://microbit.org/guide/features>

Server: localhost:3306 » Database: abc » Table: logs

Browse Structure SQL Search Insert Export Import Privileges Operations Triggers

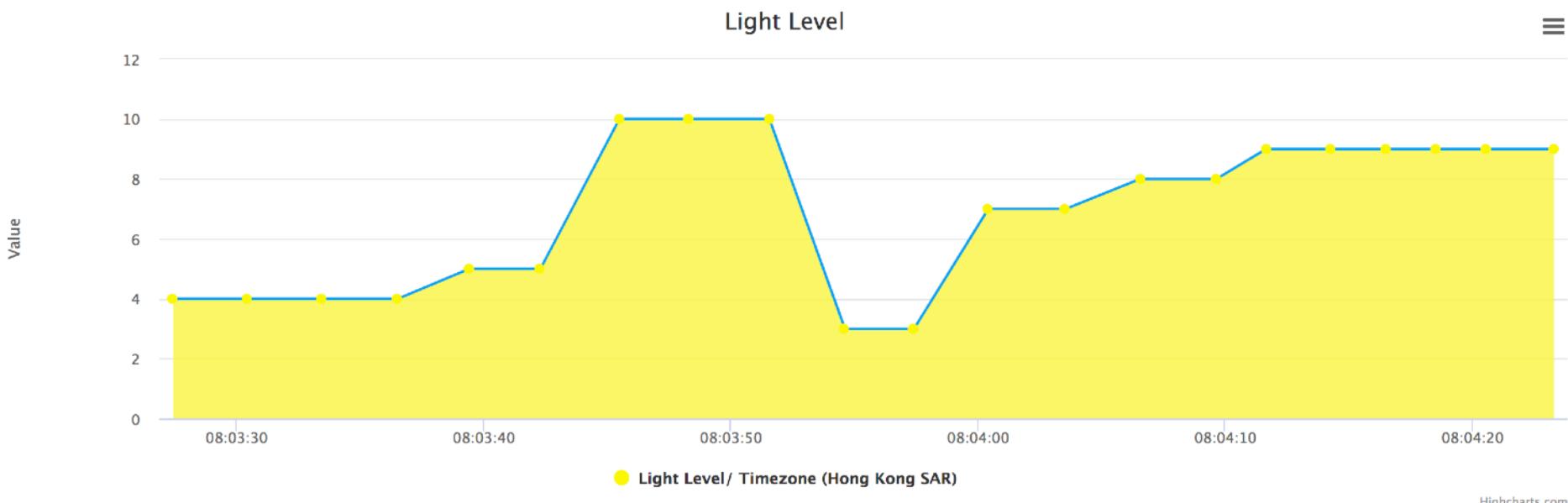
Number of rows: 25 Filter rows: Search this table Sort by key: None

+ Options

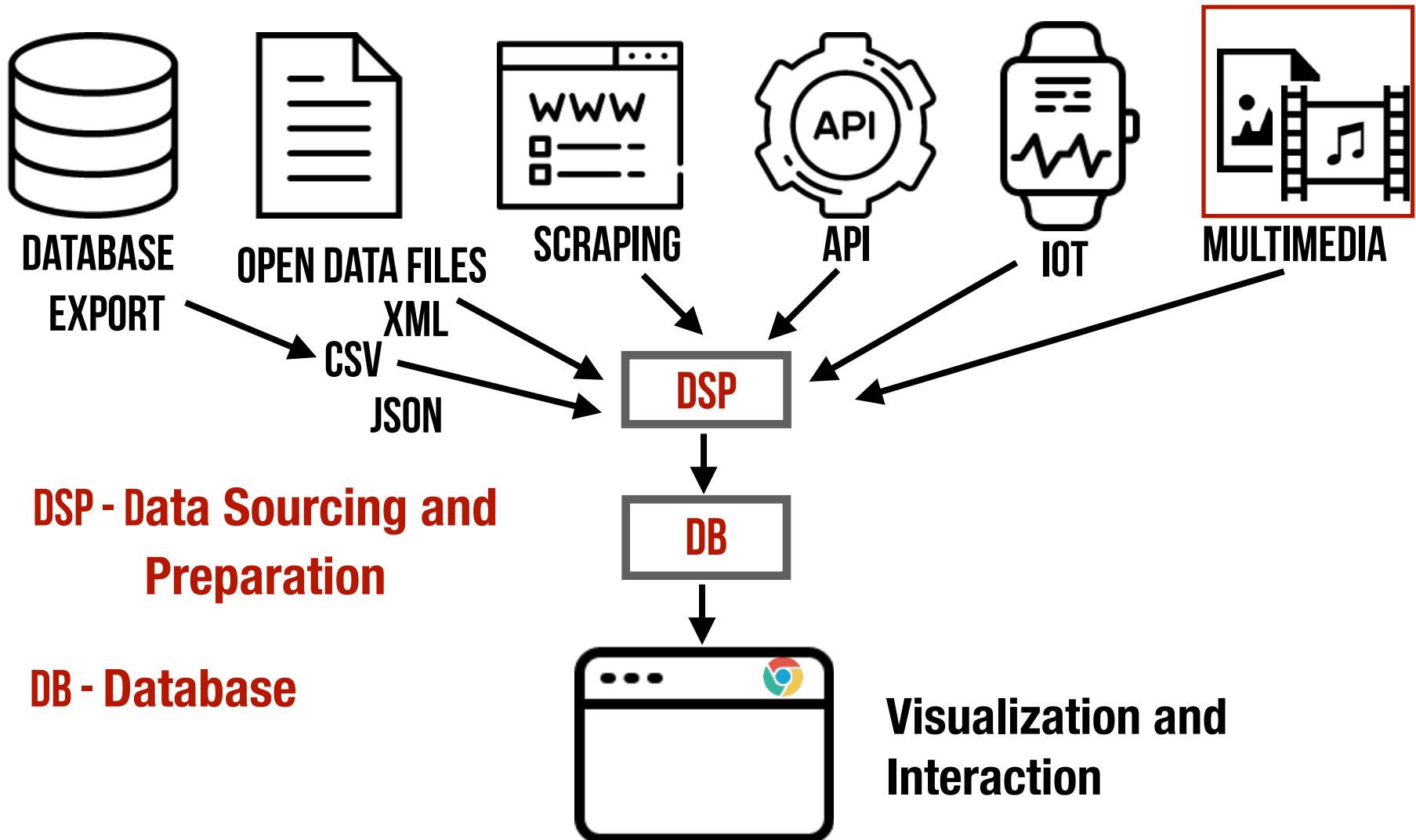
				logtID	lightLevel	timeStamp
				logtID	lightLevel	timeStamp
<input type="checkbox"/>		<a href="#">Edit</a>		15181	4	2020-09-11 22:55:24
<input type="checkbox"/>		<a href="#">Edit</a>		15182	5	2020-09-11 22:55:30
<input type="checkbox"/>		<a href="#">Edit</a>		15183	4	2020-09-11 22:55:36
<input type="checkbox"/>		<a href="#">Edit</a>		15184	7	2020-09-11 22:55:42
<input type="checkbox"/>		<a href="#">Edit</a>		15185	5	2020-09-11 22:55:48
<input type="checkbox"/>		<a href="#">Edit</a>		15186	7	2020-09-11 22:55:54
<input type="checkbox"/>		<a href="#">Edit</a>		15187	7	2020-09-11 22:56:00
<input type="checkbox"/>		<a href="#">Edit</a>		15188	5	2020-09-11 22:56:06
<input type="checkbox"/>		<a href="#">Edit</a>		15189	9	2020-09-11 22:56:12
<input type="checkbox"/>		<a href="#">Edit</a>		15190	4	2020-09-12 19:50:56
<input type="checkbox"/>		<a href="#">Edit</a>		15191	9	2020-09-12 19:51:02
<input type="checkbox"/>		<a href="#">Edit</a>		15192	10	2020-09-12 19:51:08
<input type="checkbox"/>		<a href="#">Edit</a>		15193	3	2020-09-12 19:51:17
<input type="checkbox"/>		<a href="#">Edit</a>		15194	9	2020-09-12 19:51:23
<input type="checkbox"/>		<a href="#">Edit</a>		15195	9	2020-09-12 19:51:29

Check all With selected: Edit Copy Delete Export

# Light Level : 9

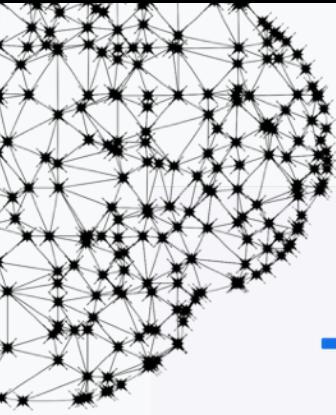


Thanks to [Soumil Nitin Shah](#) and [Peter Kazarinoff](#) for inspiring me to create this Python Flask data logger.



# **Image and Voice Recognition and Modelling**

## **Using Machine/Deep Learning**



# Teachable Machine

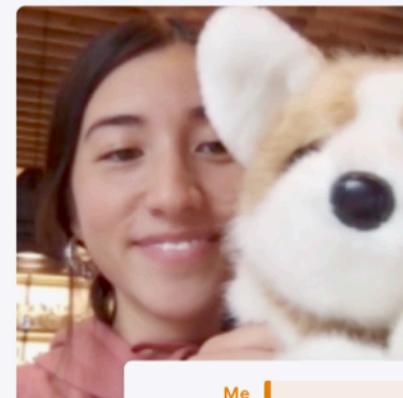
**Train a computer to recognize your own images, sounds, & poses.**

A fast, easy way to create machine learning models for your sites, apps, and more – no expertise or coding required.

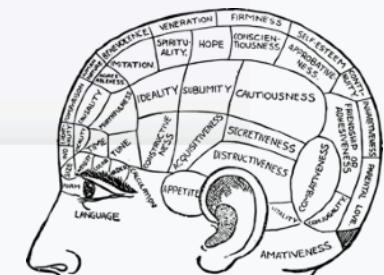
[Get Started](#)



<https://teachablemachine.withgoogle.com/>



Me  
Me + Dog <3  
98%



# ≡ Teachable Machine

## Mask

Add Image Samples:



## No-Mask

Add Image Samples:



⋮

## Training

Train Model

Advanced 

⋮

## Preview

 Export Model

You must train a model on the left before you can preview it here.

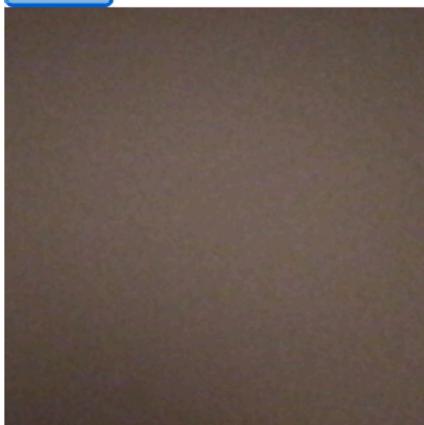
 Add a class

 English (U.S.) 

release-2-2-2 - 2.2.2#062770 - 83

Teachable Machine Image Model

Start



Mask: 0.36

No Mask: 0.64

# Teachable Machine

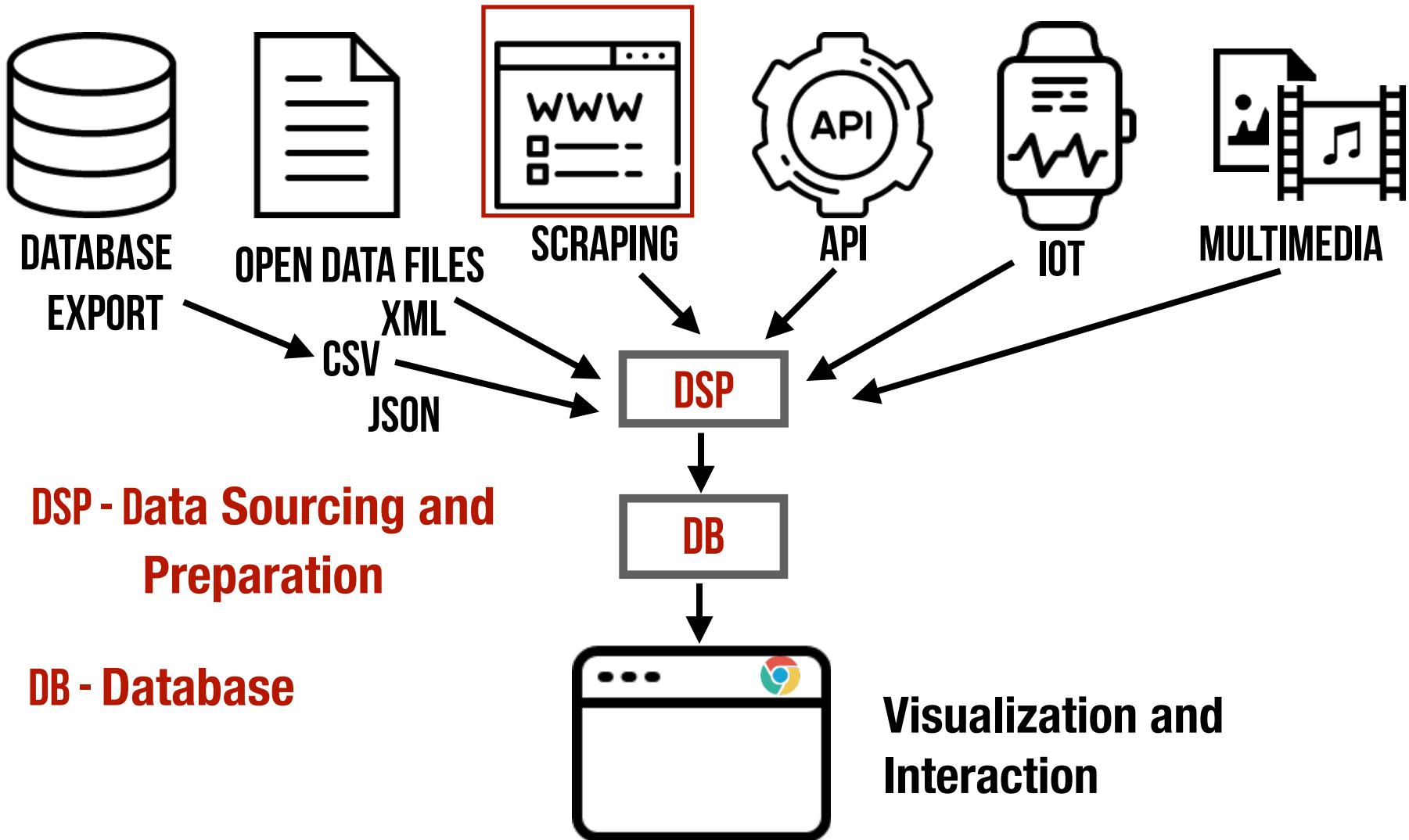
Teachable Machine is an online platform for ML beginners to experience Google's powerful TensorFlow deep learning tech. You can use it to recognize:

- Images
- Audios
- Poses

The simple user interface allows you to train the ML model to recognize images, audios, and poses with ease.

The result can be downloaded for storage or stored online.

You can easily embed the trained model into web pages, with just a few lines of JavaScript codes. The result can be integrated with HTML/CSS codes for more engaging user experience.



# **What is web scraping?**

**Web scraping is a process of fetching web pages from a website and extracting specific data from it.**

- ✓ Scrap stock prices
- ✓ Scrap product information
- ✓ Scrap race scores
- ✓ Scrap market trend data
- ✓ Scrap demographics and census data
- ✓ Social media and web articles, etc.



# **How does it work?**

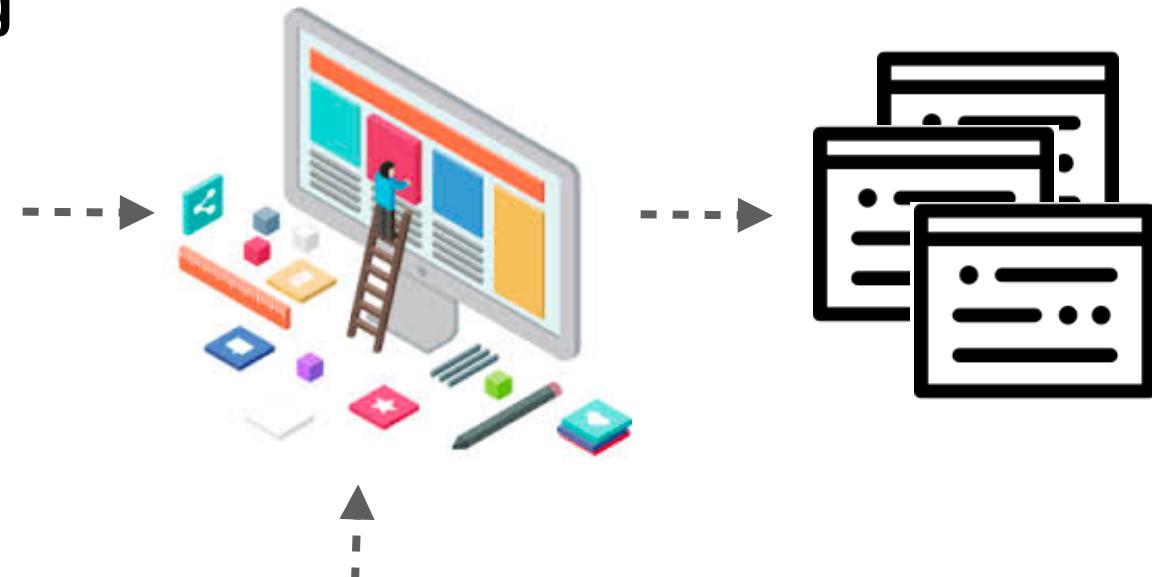
# Data

## Dynamic Web Building

ID	PHONE	POPULARNAME	PREFERREDNAME	LATITUDE	LONGITUDE
1194620	00994614	popular_name_00994614	preferred_name_00994614	23.789675	88.897865
1194621	00994615	popular_name_00994615	preferred_name_00994615	23.789675	88.897865
1194622	00994616	popular_name_00994616	preferred_name_00994616	23.789675	88.897865
1194623	00994617	popular_name_00994617	preferred_name_00994617	23.789675	88.897865
1194624	00994618	popular_name_00994618	preferred_name_00994618	23.789675	88.897865
1194625	00994619	popular_name_00994619	preferred_name_00994619	23.789675	88.897865
1194626	00994620	popular_name_00994620	preferred_name_00994620	23.789675	88.897865
1194627	00994621	popular_name_00994621	preferred_name_00994621	23.789675	88.897865
1194628	00994622	popular_name_00994622	preferred_name_00994622	23.789675	88.897865
1194629	00994623	popular_name_00994623	preferred_name_00994623	23.789675	88.897865
1194630	00994624	popular_name_00994624	preferred_name_00994624	23.789675	88.897865
1194631	00994625	popular_name_00994625	preferred_name_00994625	23.789675	88.897865

# Web Publishing

# Web

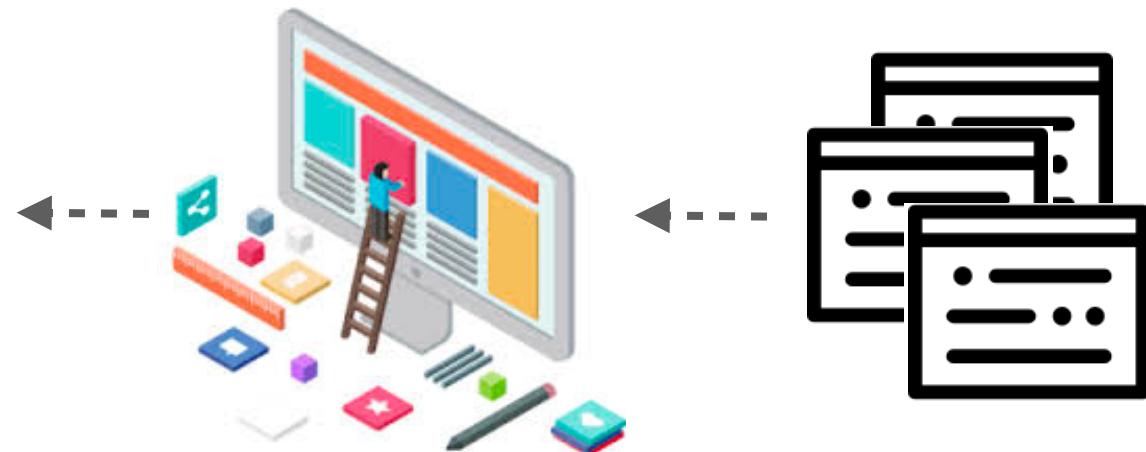


## Static Web Building

# Data

# Web Scraping

# Web



# **Scraping in Practice**

# Why scraping?

1. In the absence of open data, API and other means of sourcing data, scraping is a viable alternative.
2. Useful for collecting market data to conduct competitive analysis (e.g. price comparison)
3. As part of a data understanding exercise, gather data from multiple sources, including scraping, can help to see the big picture.
4. In exploratory data investigation, the more data the better for discovering patterns
5. Machine learning requires enormous amount of training and testing data for building predictive models.

# **Download Data Tools for Today**



## A web scraping tool that is easy to use

ParseHub is a free web scraping tool. With our advanced web scraper, extracting data is as easy as clicking the data you need.

Download our free app



<https://www.parsehub.com/>



Open a website

Download our [desktop app](#). Choose a site to scrape data from.



Click to select data

Get data from multiple pages. Interact with AJAX, forms, dropdowns, etc.



Download results

Access data via JSON, Excel and [API](#). Data is collected by our servers.



Projects

Runs

My Account

Integrations

Plans & Billing

Tutorials

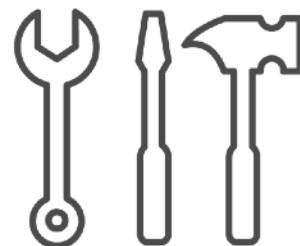
Documentation

API

Contact

Log Out

## Recent projects



+ New Project

covid19 hkej.com Project



travelchinaguide.com Project



startupbeat.hkej.com Project



basketball-reference.com P...



+ See more

## Recent runs

## Interactive Tutorials

✓ Learn the Basics (8 min)

✓ Select and Download Data (3 min)

✓ Group data with Relative Select (3 min)

✓ Click and Navigate to Links (4 min)

## Written Tutorials

1 Parsehub 101

2 Pagination ('next' page buttons)

3 Scrape product details

4 Scrape leads from directories



WIKIPEDIA  
The Free Encyclopedia

Article Talk

Not logged in Talk Contributions Create account Log in

Read

Edit

View history

Search Wikipedia

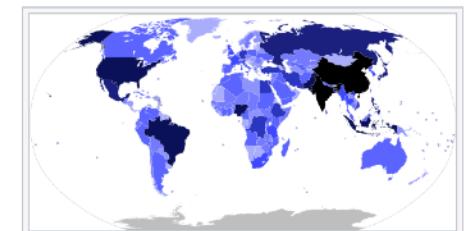


# List of countries and dependencies by population

From Wikipedia, the free encyclopedia

This is a **list of countries and dependencies by population**. It includes **sovereign states**, inhabited **dependent territories** and, in some cases, **constituent countries** of sovereign states, with inclusion within the list being primarily based on the ISO standard **ISO 3166-1**. For instance, the **United Kingdom** is considered as a single entity, while the constituent countries of the **Kingdom of the Netherlands** are considered separately. In addition, this list includes certain **states with limited recognition** not found in ISO 3166-1.

Also given in percent is each country's population compared with the **world population**, which the **United Nations** estimates at 7.82 billion as of today.



Map of countries and territories by population in 2019



A cartogram of the world population in 2018

## Contents [hide]

- 1 Method
- 2 Sovereign states and dependencies by population
- 3 See also
  - 3.1 Lists of countries by population
    - 3.1.1 Continental

[https://en.wikipedia.org/wiki/List\\_of\\_countries\\_and\\_dependencies\\_by\\_population](https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population)

## Sovereign states and dependencies by population [\[ edit \]](#)

Note: A number rank is assigned to the 193 member states of the United Nations and the two observer states to the United Nations General Assembly. Dependent territories and constituent countries that are parts of sovereign states are shown in *italics* and not assigned a numbered rank. In addition, sovereign states with limited recognition are not assigned a number rank.

Rank	Country (or dependent territory)	Population	% of world	Date	Source (official or UN)
1	China <sup>[b]</sup>	1,405,239,080	18.0%	9 Nov 2020	National population clock <sup>[3]</sup>
2	India <sup>[c]</sup>	1,369,434,247	17.5%	9 Nov 2020	National population clock <sup>[4]</sup>
3	United States <sup>[d]</sup>	330,617,844	4.22%	9 Nov 2020	National population clock <sup>[5]</sup>
4	Indonesia	269,603,400	3.45%	1 Jul 2020	National annual projection <sup>[6]</sup>
5	Pakistan <sup>[e]</sup>	220,892,331	2.82%	1 Jul 2020	UN Projection <sup>[2]</sup>
6	Brazil	212,311,986	2.71%	9 Nov 2020	National population clock <sup>[7]</sup>
7	Nigeria	206,139,587	2.63%	1 Jul 2020	UN Projection <sup>[2]</sup>
8	Bangladesh	160,615,066	2.17%	9 Nov 2020	National population

◀ bernard@intechni...

Projects

Runs

My Account

Integrations

Plans & Billing

Tutorials

Documentation

API

Contact

Log Out

## My Projects

+ New Project

Import Project

Export All

Click 'New Project' to start.

catalog.hkstore.com demo 1



basketball-reference.com Proj...



OpenRice



startupbeat.hkj.com Project



## Untitled Project

Enter a website you'd like to extract data from

**Start project on this URL**

Enter URL and click  
'Start project on this  
URL'.

The screenshot shows the ParseHub landing page. At the top, there's a search bar with the URL "countries\_and\_dependencies\_by\_population" and a green button below it labeled "Start project on this URL". A red box highlights this area. The main content includes a "Welcome!" message, "Beginner Tutorials" (with 5 numbered steps), and "Advanced Tutorials" (with 5 numbered steps). Below these are tabs for "CSV/Excel", "JSON", and "CSV/Excel Wide (beta)". A large box indicates where data previews will appear, with the text "A preview of your data will appear here". At the bottom, there are checkboxes for "Show more data" and "Visuals enabled (advanced)", and a blue circular icon with a white message icon.

Welcome!

Beginner Tutorials

1. Start Here: Create your first project
2. Extract text from a web page
3. Extract data from many pages (pagination)
4. Run project & download Excel & JSON data
5. Use the REST API

Advanced Tutorials

1. Collect data on a schedule
2. Enter text into a search box
3. Get data from behind a log-in
4. Infinite scrolling pages
5. Enter URLs for ParseHub to crawl

CSV/Excel    JSON    CSV/Excel Wide (beta)

A preview of your data will appear here

Show more data ?     Visuals enabled (advanced) ?

API

Tutorials

Contact

Select page

Select selection1 (202)

Extract name

Extract url

Selection Node: 

All td s  
> All a s

Wait up to 60 seconds for elements to appear.

W List of countries and dependencies

https://en.wikipedia.org/wiki/List\_of\_countries\_and\_dependencies\_by\_population

	Rank	Country (or dependent territory)	Population	% of world	Date	Source (official or UN)
1	China[b]	 China[b]			2020	National population clock[3]
2	India[c]	 India[c]			2020	National population clock[4]
3	United States[d]	 United States[d]	330,544,466	4.23%	27 Oct 2020	National population clock[5]
4	Indonesia	 Indonesia	269,603,400	3.45%	1 Jul 2020	National annual

Click first target as anchor column.

CSV/Excel JSON CSV/Excel Wide (beta)

selection1_name	selection1_url
China	<a href="https://en.wikipedia.org/wiki/Demographics_of_China">https://en.wikipedia.org/wiki/Demographics_of_China</a>
India	<a href="https://en.wikipedia.org/wiki/Demographics_of_India">https://en.wikipedia.org/wiki/Demographics_of_India</a>
United States	<a href="https://en.wikipedia.org/wiki/Demographics_of_United_States">https://en.wikipedia.org/wiki/Demographics_of_United_States</a>
Indonesia	<a href="https://en.wikipedia.org/wiki/Demographics_of_Indonesia">https://en.wikipedia.org/wiki/Demographics_of_Indonesia</a>
Pakistan	<a href="https://en.wikipedia.org/wiki/Demographics_of_Pakistan">https://en.wikipedia.org/wiki/Demographics_of_Pakistan</a>
Brazil	<a href="https://en.wikipedia.org/wiki/Demographics_of_Brazil">https://en.wikipedia.org/wiki/Demographics_of_Brazil</a>
Nigeria	<a href="https://en.wikipedia.org/wiki/Demographics_of_Nigeria">https://en.wikipedia.org/wiki/Demographics_of_Nigeria</a>
Bangladesh	<a href="https://en.wikipedia.org/wiki/Demographics_of_Bangladesh">https://en.wikipedia.org/wiki/Demographics_of_Bangladesh</a>

This is a live preview. When you are ready to run your project, click Get Data.



en.wikip...

BROWSE

Select page

Select country (202) +

Extract name

Extract url

Get Data

main\_template

Enter 'Select' name and 'Extract' column name as you wish.

Wait up to 60 seconds for elements to appear.

W List of countries and depende X +

https://en.wikipedia.org/wiki/List\_of\_countries\_and\_dependencies\_by\_population

	Rank	Country (or dependent territory)	Population	% of world	Date	Source (official or UN)
1	China <sup>b</sup>	1,405,050,320	18.0%	27 Oct 2020	National population clock <sup>[3]</sup>	
2	India <sup>c</sup>	1,368,910,880	17.5%	27 Oct 2020	National population clock <sup>[4]</sup>	
3	United States <sup>d</sup>	330,544,466	4.23%	27 Oct 2020	National population clock <sup>[5]</sup>	
4	Indonesia	269,603,400	3.45%	1 Jul 2020	National annual	

CSV/Excel JSON CSV/Excel Wide (beta)

country_name	country_url
China	https://en.wikipedia.org/wiki/Demographics_of_China
India	https://en.wikipedia.org/wiki/Demographics_of_India
United States	https://en.wikipedia.org/wiki/Demographics_of_United_States
Indonesia	https://en.wikipedia.org/wiki/Demographics_of_Indonesia
Pakistan	https://en.wikipedia.org/wiki/Demographics_of_Pakistan
Brazil	https://en.wikipedia.org/wiki/Demographics_of_Brazil
Nigeria	https://en.wikipedia.org/wiki/Demographics_of_Nigeria
Bangladesh	https://en.wikipedia.org/wiki/Demographics_of_Bangladesh

This a live preview. When you are ready to run your project, click Get Data.

Show more data ?  Visuals enabled (advanced) ?

API Tutorials Contact

main\_template

en.wikip...

BROWSE

Attach already-selected elements to related elements (auto-extracts data if possible). Good for grouping different fields together in your data.

Relative Select

Click

Advanced

Select page

Select country (202)

Extract name

Extract url

Get Data

Use 'Relative Select' to select remaining columns with 1st column as the anchor.

W List of countries and depend...

Relative Select

country\_name

China

India

United States

Indonesia

Pakistan

Brazil

Nigeria

Bangladesh

Source (official or UN)

Date

of world

18.0%

27 Oct 2020

National population clock<sup>[3]</sup>

17.5%

27 Oct 2020

National population clock<sup>[4]</sup>

4.23%

27 Oct 2020

National population clock<sup>[5]</sup>

3.45%

1. Jul 2020

National annual

This a live preview. When you are ready to run your project, click Get Data.

Show more data

Visuals enabled (advanced)

API Tutorials Contact

main\_template

country_name	Source	Date	Value
China	National population clock <sup>[3]</sup>	27 Oct 2020	18.0%
India	National population clock <sup>[4]</sup>	27 Oct 2020	17.5%
United States	National population clock <sup>[5]</sup>	27 Oct 2020	4.23%
Indonesia	National annual	1. Jul 2020	3.45%
Pakistan			
Brazil			
Nigeria			
Bangladesh			

en.wikip...

BROWSE

Select page +

Select country +

Extract name

Extract url

Relative selection1 (1) +

> End to

Wait up to 60 seconds for elements to appear.

main\_template

API Tutorials Contact

List of countries and dependencies

https://en.wikipedia.org/wiki/List\_of\_countries\_and\_dependencies\_by\_population

	Rank	Country (or dependent territory)	Population	% of world	Date	Source (official or UN)
	1	China <sup>b</sup>	1,405,050,320	18.0%	27 Oct 2020	National population clock <sup>[3]</sup>
	2	India <sup>c</sup>	1,368,910,880	17.5%	27 Oct 2020	National population clock <sup>[4]</sup>
	3	United States <sup>d</sup>	330,544,466	4.23%	27 Oct 2020	National population clock <sup>[5]</sup>
	4	Indonesia	269,603,400	3.45%	1 Jul 2020	National annual

CSV/Excel JSON CSV/Excel Wide (beta)

country_name	country_url	country_selection1
China	https://en.wikipedia.org/wiki/Demographics_of_China	1,405,050,320
India	https://en.wikipedia.org/wiki/Demographics_of_India	
United States	https://en.wikipedia.org/wiki/Demographics_of_United_States	
Indonesia	https://en.wikipedia.org/wiki/Demographics_of_Indonesia	
Pakistan	https://en.wikipedia.org/wiki/Demographics_of_Pakistan	

This is a live preview. When you are ready to run your project, click Get Data.

Show more data  Visuals enabled (advanced)

en.wikip...

BROWSE

main\_template

Select page +  
Select country +  
Extract name  
Extract url  
Relative population (202) ↕ - +  
Get Data

W List of countries and dependencies

https://en.wikipedia.org/wiki/List\_of\_countries\_and\_dependencies\_by\_population

	Country (or dependent territory)	Rank	Population	% of world	Date	Source Select Mode (official or UN)
1	China <sup>b</sup>	1	1,405,050,320	18.0%	27 Oct 2020	National population clock <sup>[3]</sup>
2	India <sup>c</sup>	2	1,368,910,880	17.5%	27 Oct 2020	National population clock <sup>[4]</sup>
3	United States <sup>d</sup>	3	330,544,466	4.23%	27 Oct 2020	National population clock <sup>[5]</sup>
4	Indonesia	4	269,603,400	3.45%	1 Jul 2020	National annual

Rename the column name as you wish.  
Make sure no blank space and odd symbols.

CSV/Excel JSON CSV/Excel Wide (beta)

country_name	country_url	country_population
China	https://en.wikipedia.org/wiki/Demographics_of_China	1,405,050,320
India	https://en.wikipedia.org/wiki/Demographics_of_India	1,368,910,880
United States	https://en.wikipedia.org/wiki/Demographics_of_United_States	330,544,466
Indonesia	https://en.wikipedia.org/wiki/Demographics_of_Indonesia	269,603,400
Pakistan	https://en.wikipedia.org/wiki/Demographics_of_Pakistan	220,892,331

This is a live preview. When you are ready to run your project, click Get Data.



en.wikip...

BROWSE

Select page

Select country

Extract name

Extract url

Relative population

Relative percent\_of\_world

Relative date (202)

Get Data

When done defining the columns, click 'Get Data'.

main\_template

List of countries and dependencies by population

Country

Rank (or dependent territory) Population % of world Date

Source: Select Mode (official or UN)

	Rank (or dependent territory)	Population	% of world	Date	Source: Select Mode (official or UN)
1	China <sup>[b]</sup>	1,405,050,320	18.0%	27 Oct 2020	National population clock <sup>[3]</sup>
2	India <sup>[c]</sup>	1,368,910,880	17.5%	27 Oct 2020	National population clock <sup>[4]</sup>
3	United States <sup>[d]</sup>	330,544,466	4.23%	27 Oct 2020	National population clock <sup>[5]</sup>
4	Indonesia	269,603,400	3.45%	1 Jul 2020	National annual

CSV/Excel JSON CSV/Excel Wide (beta)

country_name	country_url	country_population	country_percent_of...	country_date
China	https://en.wikipedia.org/wiki/Demographics_of_China	1,405,050,320	18.0%	27 Oct 2020
India	https://en.wikipedia.org/wiki/Demographics_of_India	1,368,910,880	17.5%	27 Oct 2020
United States	https://en.wikipedia.org/wiki/Demographics_of_the_United_States	330,544,466	4.23%	27 Oct 2020
Indonesia	https://en.wikipedia.org/wiki/Demographics_of_Indonesia	269,603,400	3.45%	1 Jul 2020
Pakistan	https://en.wikipedia.org/wiki/Demographics_of_Pakistan	220,892,331	2.82%	1 Jul 2020

This is a live preview. When you are ready to run your project, click Get Data.

Show more data ?  Visuals enabled (advanced) ?



← ⌂ en.wikipedia.org Project

Run your project on ParseHub's  
servers, once

Test Run



Run



Schedule

Edit project

Previous Runs

Click 'Run' to start scraping.

Your results will appear here after you've run your project.



 Edit project

Data is being collected. Please wait.

Starting up...

Refreshing status in **2** second(s). [Refresh now](#)

Download Data

CSV/Excel



JSON

API

 Cancel Run

All dates and times are in UTC +0000.

Empty file with no results? [Click here](#) to fix.

CSV file too big? Save the JSON file and [click here to convert to CSV](#).

Run Details

Settings

Started job



 Edit project

When done scraping, download the data as CSV file and import into a spreadsheet or SQL database.

Your data is ready! Click on the green buttons to download.

#### Download Data

CSV/Excel

JSON

API

Template Name  
main\_template

Pages Scrapped  
1 

All dates and times are in UTC +0000.

Empty file with no results? [Click here](#) to fix.

CSV file too big? Save the JSON file and [click here](#) to convert to CSV.

#### Run Details

Status	complete
Pages	1 collected
Initialized	2020-10-28T14:43:31

#### Settings

URL

[https://en.wikipedia.org  
/wiki/List\\_of\\_countries\\_and\\_dependencies\\_by\\_population](https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population)



country_name	country_url	country_population	country_percent_of_world	country
China	<a href="https://en.wikipedia.org/wiki/Demographics_of_China">https://en.wikipedia.org/wiki/Demographics_of_China</a>	1,405,050,320	18.0%	27 Oct 2
India	<a href="https://en.wikipedia.org/wiki/Demographics_of_India">https://en.wikipedia.org/wiki/Demographics_of_India</a>	1,368,910,880	17.5%	27 Oct 2
United States	<a href="https://en.wikipedia.org/wiki/Demographics_of_United_States">https://en.wikipedia.org/wiki/Demographics_of_United_States</a>	330,544,466	4.23%	27 Oct 2
Indonesia	<a href="https://en.wikipedia.org/wiki/Demographics_of_Indonesia">https://en.wikipedia.org/wiki/Demographics_of_Indonesia</a>	269,603,400	3.45%	1 Jul 202
Pakistan	<a href="https://en.wikipedia.org/wiki/Demographics_of_Pakistan">https://en.wikipedia.org/wiki/Demographics_of_Pakistan</a>	220,892,331	2.82%	1 Jul 202
Brazil	<a href="https://en.wikipedia.org/wiki/Demographics_of_Brazil">https://en.wikipedia.org/wiki/Demographics_of_Brazil</a>	212,254,617	2.71%	27 Oct 2
Nigeria	<a href="https://en.wikipedia.org/wiki/Demographics_of_Nigeria">https://en.wikipedia.org/wiki/Demographics_of_Nigeria</a>	206,139,587	2.64%	1 Jul 202
Bangladesh	<a href="https://en.wikipedia.org/wiki/Demographics_of_Bangladesh">https://en.wikipedia.org/wiki/Demographics_of_Bangladesh</a>	169,542,128	2.17%	27 Oct 2
Russia	<a href="https://en.wikipedia.org/wiki/Demographics_of_Russia">https://en.wikipedia.org/wiki/Demographics_of_Russia</a>	146,748,590	1.88%	1 Jan 202
Mexico	<a href="https://en.wikipedia.org/wiki/Demographics_of_Mexico">https://en.wikipedia.org/wiki/Demographics_of_Mexico</a>	127,792,286	1.63%	1 Jul 202
Japan	<a href="https://en.wikipedia.org/wiki/Demographics_of_Japan">https://en.wikipedia.org/wiki/Demographics_of_Japan</a>	125,880,000	1.61%	1 Oct 20
Philippines	<a href="https://en.wikipedia.org/wiki/Demographics_of_Philippines">https://en.wikipedia.org/wiki/Demographics_of_Philippines</a>	109,349,790	1.40%	27 Oct 2
DR Congo	<a href="https://en.wikipedia.org/wiki/Demographics_of_Democratic_Republic_of_the_Congo">https://en.wikipedia.org/wiki/Demographics_of_Democratic_Republic_of_the_Congo</a>	101,935,800	1.30%	1 Jul 202
Egypt	<a href="https://en.wikipedia.org/wiki/Demographics_of_Egypt">https://en.wikipedia.org/wiki/Demographics_of_Egypt</a>	101,100,567	1.29%	27 Oct 2
Ethiopia	<a href="https://en.wikipedia.org/wiki/Demographics_of_Ethiopia">https://en.wikipedia.org/wiki/Demographics_of_Ethiopia</a>	100,829,000	1.29%	1 Jul 202
Vietnam	<a href="https://en.wikipedia.org/wiki/Demographics_of_Vietnam">https://en.wikipedia.org/wiki/Demographics_of_Vietnam</a>	96,483,981	1.23%	1 Jul 202
Iran	<a href="https://en.wikipedia.org/wiki/Demographics_of_Iran">https://en.wikipedia.org/wiki/Demographics_of_Iran</a>	83,898,529	1.07%	27 Oct 2
Turkey	<a href="https://en.wikipedia.org/wiki/Demographics_of_Turkey">https://en.wikipedia.org/wiki/Demographics_of_Turkey</a>	83,154,997	1.06%	31 Dec 2
Germany	<a href="https://en.wikipedia.org/wiki/Demographics_of_Germany">https://en.wikipedia.org/wiki/Demographics_of_Germany</a>	83,122,889	1.06%	30 Jun 2
France	<a href="https://en.wikipedia.org/wiki/Demographics_of_France">https://en.wikipedia.org/wiki/Demographics_of_France</a>	67,132,000	0.858%	1 Sep 20

**Scraping multiple pages.**

◀ bernard@intechni...

Projects

Runs

My Account

Integrations

Plans & Billing

Tutorials

Documentation

API

Contact

Log Out

## My Projects

+ New Project

Import Project

Export All

Search Projects

catalog.hkstore.com demo 1



basketball-reference.com Proj...



OpenRice



startupbeat.hkj.com Project



## Untitled Project

Enter a website you'd like to extract data from

[https://catalog.hkstore.com/catalog\\_tc\\_hk/m](https://catalog.hkstore.com/catalog_tc_hk/m)

[Start project on this URL](#)

Enter URL and click  
'Start project on this  
URL'.

The screenshot shows a web browser window with the URL [https://catalog.hkstore.com/catalog\\_tc\\_hk/men](https://catalog.hkstore.com/catalog_tc_hk/men) in the address bar. The page title is '男裝'. The main content area displays a grid of men's apparel products, including shirts and jackets. On the left, there is a sidebar with categories like '分類', '熱賣系列', '商品分類', '鞋款', and '服裝類別', each with a '+' sign to expand. At the bottom of the sidebar, there are buttons for 'CSV/Excel', 'JSON', and 'CSV/Excel Wide (beta)'. A teal banner at the bottom says 'This a live preview. When you are ready to run your project, click Get Data.' Below the banner is a large empty box for data visualization. At the very bottom, there are checkboxes for 'Show more data' and 'Visuals enabled (advanced)', and a link to the URL again: [https://catalog.hkstore.com/catalog\\_tc\\_hk/men](https://catalog.hkstore.com/catalog_tc_hk/men).

catalog.... BROWSE

Select page +

Select & Extract selection1 (1) +

Get Data

Selection Node: Edit

All elements with class product-item-brand  
> All elements

Wait up to 60 seconds for elements to appear.

男裝 https://catalog.hkstorecom/catalog\_tc\_hk/men

註冊成為新會員即送\$100網店優惠券。立即註冊 Select Mode

CATALOG MARATHON SPORTS 馬拉松 搜索

最新商品 品牌 聯乘系列 男裝 女裝 童裝 電子禮品卡 SALE

配件類別 + 袋款 + 品牌 + 性別 + 活動類型 + 功能 + 顏色 + 價格 +

SPAN Adidas

Stan Smith 中性球鞋 \$899

加入購物車 加入喜愛清單

Adidas Puffylette 中性無鞋帶 \$599

Click first target as anchor column.

CSV/Excel JSON CSV/Excel Wide (beta)

selection1

Adidas

This is a live preview. When you are ready to run your project, click Get Data.

Show more data Visuals enabled (advanced)

catalog....

BROWSE

Select page +

Select selection1 (24) - +

Extract name

Get Data

Selection Node: [Edit](#)

All elements with class `product-item-brand`  
> All elements s

Wait up to 60 seconds for elements to appear.

男裝

https://catalog.hkstore.com/catalog\_to\_hk/men

訂單淨值滿\$400即享免費門市取貨

Select Mode

CATALOG MARATHON SPORTS 馬拉松

最新商品 品牌 聯乘系列 男裝 女裝 童裝 電子禮品卡 SALE

搜索

鞋款 +

服裝類別 +

配件類別 +

袋款 +

品牌 +

性別 +

Adidas Stan Smith

Retropy E5

Puffylette

活動類型 + 中性球鞋

男裝球鞋

功能 +

顏色 + DIV

Make sure the first columns are properly defined with all the targets bounded by the green borders.

CSV/Excel JSON CSV/Excel Wide (beta)

selection1\_name

Adidas

Adidas

Adidas

PUMA

This a live preview. When you are ready to run your project, click Get Data.

Show more data ?  Visuals enabled (advanced) ?

catalog....

男裝

https://catalog.hkstore

Select Mode

Select page

Select selection1(24)

Extract name

Get Data

Relative Select

Click

Advanced

... main\_template

鞋款

Always use the first column as the anchor to ‘Relative Select’ each new column.

功能

顏色

CSV/Excel JSON CSV/Excel Wide (beta)

selection1\_name

Adidas

Adidas

Adidas

PUMA

This a live preview. When you are ready to run your project, click Get Data.

Show more data Visuals enabled (advanced)

The screenshot shows a web browser window with a scraping tool overlay. The tool has a sidebar on the left with sections for 'main\_template', 'catalog....', 'Select page', 'Select selection1(24)', 'Extract name', 'Get Data', and a tooltip for 'Relative Select'. The main area displays a shoe catalog from 'hkstore'. A tooltip for 'Relative Select' is overlaid on the interface. The catalog shows several shoe models like 'Retropy E5' and 'Puffylette' with their prices (\$899 and \$599 respectively). The bottom of the tool has a 'CSV/Excel' section with a preview of the extracted data for 'selection1\_name' (listing 'Adidas' and 'PUMA' multiple times) and a note about being a live preview.

訂單淨值滿\$400即享免費門市取貨

Select Mode

CATALOG MARATHON SPORTS 馬拉松

最新商品 品牌 聯乘系列 男裝 女裝 童裝 電子禮品卡 SALE

搜索

會員登入 | 註冊

Always use the first column as the anchor to 'Relative Select' each new column.

選項	Adidas Stan Smith	Adidas Retropy E5	Adidas Puffylette
鞋款	Stan Smith	Retropy E5	Puffylette
服裝類別	中性球鞋	男裝球鞋	中性無鞋帶
	\$899	\$999	\$599

https://catalog.hkstore.com/catalog\_tc\_hk/men/adidas-108gx-8

CSV/Excel JSON CSV/Excel Wide (beta)

selection1_name	selection1_series	selection1_series_url
Adidas	Stan Smith	https://catalog.hkstore.com/catalog_tc_hk/men/adidas-108gx-8
Adidas	Retropy E5	https://catalog.hkstore.com/catalog_tc_hk/men/adidas-108hq6460
Adidas	Puffylette	https://catalog.hkstore.com/catalog_tc_hk/men/adidas-108hr1481

This is a live preview. When you are ready to run your project, click Get Data.

Show more data ?    Visuals enabled (advanced) ?

Select page



Select selection1



Extract name

Relative series (24)



Extract series

Delete Command

Extract series\_url



Get Data

Remove URL  
link if you don't  
need it.

男裝 +

[https://catalog.hkstore.com/catalog\\_tc\\_hk/men](https://catalog.hkstore.com/catalog_tc_hk/men)

註冊成為新會員即送\$100網店優惠券。 [立即註冊](#)

Select Mode

CATALOG MARATHON SPORTS

最新商品 品牌 聯乘系列 男裝 女裝 童裝 電子禮品卡 SALE



PUMA Scuff	Hoka Bondi 8	Keen Uneek SNK Sneaker
中性涼鞋	男裝登山鞋	男裝球鞋
\$259	\$1,380	\$1,090

chrome://phapp/content/views/index.html

CSV/Excel JSON CSV/Excel Wide (beta)

selection1_name	selection1_series	selection1_series_url
Adidas	Stan Smith	<a href="https://catalog.hkstore.com/catalog_tc_hk/men/adidas-108gx-8">https://catalog.hkstore.com/catalog_tc_hk/men/adidas-108gx-8</a>
Adidas	Retropy E5	<a href="https://catalog.hkstore.com/catalog_tc_hk/men/adidas-108hq6460">https://catalog.hkstore.com/catalog_tc_hk/men/adidas-108hq6460</a>
Adidas	Puffylette	<a href="https://catalog.hkstore.com/catalog_tc_hk/men/adidas-108hr1481">https://catalog.hkstore.com/catalog_tc_hk/men/adidas-108hr1481</a>

This a live preview. When you are ready to run your project, click Get Data.

 Show more data 
 Visuals enabled (advanced) 


catalog... BROWSE

Select page (1)

Select selection1

Extract name

Relative series

Extract series

Relative style

Extract style

Relative price

Extract price

New Select Command

Get Data

Targets one or more elements for a command (auto-extracts data if possible).

Select

Relative Select

Click

Advanced

男裝

訂單淨值滿\$400即享免費門市取貨

Select Mode

會員登入 | 註冊

分類

熱賣系列

商品分類

鞋款

服裝類別

男裝 女裝 童裝 電子禮品卡 SALE

Adidas Stan Smith

Adidas Retropy E5

Adidas Puffylette

Click the 'Select page' + button to create the next page.

CSV/Excel JSON CSV/Excel Wide (beta)

selection1_name	selection1_series	selection1_style	selection1_price
Adidas	Stan Smith	中性球鞋	\$899
Adidas	Retropy E5	男裝球鞋	\$999
Adidas	Puffylette	中性無鞋帶	\$599
PUMA	Scuff	中性涼鞋	\$259

This a live preview. When you are ready to run your project, click Get Data.

Show more data ?  Visuals enabled (advanced) ?

API Tutorials Contact

catalog... BROWSE

Select Mode

... main\_template

Select page +  
Select selection1 +  
Extract name  
Relative series  
Extract series  
Relative style  
Extract style  
Relative price  
Extract price  
Select & Extract selection2(1) - +  
New Click Command  
Get Data

**Relative Select**  
Click Advanced ▾

Click on an already-selected element. Good for dropdowns, popups or navigating to new pages.  
常见问题

關於我們  
關於Catalog  
關於送貨  
關於退貨  
聯絡我們

追蹤我們  
免費訂閱Catalog電郵通訊，收取有關Catalog最新優惠及產品資訊。  
輸入電郵地址 提交

马拉松 Sports App

Marathon Sports App

男裝

https://catalog.hkstore.com/catalog\_tc\_hk/men

訂單淨值滿\$400即享免費門市取貨

CATALOG MARATHON SPORTS 馬拉松

最新商品 品牌 聯乘系列 男裝 女裝 童裝 電子禮品卡 SALE

1 2 3 4 5 >

Click the 'Click' button to start defining the next page.

selection1_name	selection1_series	selection1_style	selection1_price	selection2	selection2_url
Adidas	Stan Smith	中性球鞋	\$899	頁 下一頁	https://catalog.hkstore.com/catalog_tc_hk/men?p=2
Adidas	Retropy E5	男裝球鞋	\$999	頁 下一頁	https://catalog.hkstore.com/catalog_tc_hk/men?p=2
Adidas	Puffylette	中性無鞋帶	\$599	頁 下一頁	https://catalog.hkstore.com/catalog_tc_hk/men?p=2

This a live preview. When you are ready to run your project, click Get Data.

Show more data ⓘ  Visuals enabled (advanced) ⓘ

catalog... BROWSE

Select page  
Select selection1  
Extract name  
Relative series  
Extract series  
Relative style  
Extract style  
Relative price  
Extract price  
Select selection2  
Click each selection... (1) and go to main\_template

Get Data

Loads a new page    Uses AJAX

Go to Existing Template main\_template

Go to Another Project

Wait up to 5 seconds for page to load

Repeat the Current Template 0 more time(s). (0 = ∞)

Click setup

Is a next page button?

**Click 'Yes' to confirm.**

Examples of next page buttons

Examples of non next page buttons

Director: David Leitch  
Writers: Rhett Reese, Paul Wernick [more credit >](#)

UNLOCKED Apple iPhone 6 16GB/64GB/128GB with Warranty

iPhone 6 16GB 64GB 128GB

Buy It Now

Free Shipping

Free Returns

30-day Refund

13 new & refurbished from \$199.99

追蹤我們

免費訂閱Catalog電郵通訊，收取有關Catalog最新優惠及產品資訊。

輸入電子郵件地址 提交

Marathon Sports App

Marathon Sports App

selection1_name	selection1_series	selection1_style	selection1_price
Adidas	Stan Smith	中性球鞋	\$899
Adidas	Retropy E5	男裝球鞋	\$999
Adidas	Puffylette	中性無鞋帶	\$599
PUMA	Scuff	中性涼鞋	\$259

This is a live preview. When you are ready to run your project, click Get Data.

Show more data [?](#)  Visuals enabled (advanced) [?](#)



catalog... BROWSE

Select page  
Select selection1  
Extract name  
Relative series  
Extract series  
Relative style  
Extract style  
Relative price  
Extract price  
Select selection2  
Click each selection... (1) and go to main\_template  
Get Data

Loads a new page    Uses AJAX

Go to Existing Template main\_template  
Go to Another Project

Wait up to 5 seconds for page to load  
Repeat the Current Template 0 more time(s). (0 = ∞)

API Tutorials Contact

Click each s... CATALOG

Click setup

Is [ ] a next page button?

Yes No

This click takes you to the next page of results. It will repeat the current template (pagination help).

Repeat the Current Template 0 more time(s). (0 = ∞)

Click 'Repeat Current Template' to confirm.

Repeat Current Template

男裝 童裝 電子禮品卡 SALE

立即註冊

會員登入 | 註冊

1 2 3 4 5 >

購物指南 VIP會員及積分計劃 電子禮品卡 員工註冊

聯絡我們 門市地址 工作機會

追蹤我們 免費訂閱Catalog電郵通訊，收取有關Catalog最新優惠及產品資訊。

輸入電子郵件 提交

Marathon Sports App Marathon Sports App

CSV/Excel JSON CSV/Excel Wide (beta)

selection1_name	selection1_series	selection1_style	selection1_price
Adidas	Stan Smith	中性球鞋	\$899
Adidas	Retropy E5	男裝球鞋	\$999
Adidas	Puffylette	中性無鞋帶	\$599
PUMA	Scuff	中性涼鞋	\$259

This a live preview. When you are ready to run your project, click Get Data.

Show more data Visuals enabled (advanced)

The screenshot shows a web scraping project in progress. On the left, the 'main\_template' sidebar contains various selection and extraction rules. A 'Click setup' dialog is open in the center, asking if the highlighted element is a 'next page button'. It includes options to 'Yes' or 'No', a note about pagination, and a 'Repeat the Current Template' button which is highlighted with a red box. Below the dialog, a large red text overlay says 'Click 'Repeat Current Template' to confirm.' To the right, a live preview table displays data for four shoe products from Adidas and Puma, including names, series, styles, and prices. At the bottom, there are checkboxes for 'Show more data' and 'Visuals enabled (advanced)'.

main\_template

catalog.... 

...  
Select page +  
Select selection1 +  
Extract name  
Relative series  
Extract series  
Relative style  
Extract style  
Relative price  
Extract price  
Select selection2  
**Click each selection... (1) **  
**and go to main\_template **

**Get Data**

**Next page defined.**

Go to Existing Template **main\_template** ▾  
 Go to Another Project

Wait up to **5** seconds for page to load

Repeat the Current Template **0** more time(s). (0 = ∞)

[API](#) [Tutorials](#) [Contact](#)

男裝  男裝  男裝 

https://catalog.hkstore.com/catalog\_tc\_hk/men?p=2

Click each s... Select Mode

訂單淨值滿\$400即享免費門市取貨

CATALOG MARATHON SPORTS 馬拉松

最新商品 品牌 聯乘系列 男裝 女裝 童裝 電子禮品卡 SALE

主頁 > 男裝

男裝 25-48 / 2507件 排序 最新上架

過濾

分類 +  
熱賣系列 +  
商品分類 +  
鞋款 +  
服裝類別 +

CSV/Excel JSON CSV/Excel Wide (beta)

selection1_name	selection1_series	selection1_style	selection1_price
New Balance	2002R	中性球鞋	\$1,199
New Balance	2002R	中性球鞋	\$1,199
New Balance	2002R	中性球鞋	\$1,199
PLASTIC THING X CATALOG系列	Plastic Thing 為食妹刺繡圖案外套	中性外套	\$699

This a live preview. When you are ready to run your project, click Get Data.

Show more data   
 Visuals enabled (advanced) 

## catalog...

Settings

Get Data

Select

Save

Sele

Undo

E Redo

F Export

Delete Project

Exit

Logout

Extract price

Select selection2

Click each selection... (1)



and go to main\_template

Get Data

Click the  
'Settings'  
option to  
define max #  
of pages to  
be scraped.

API

Tutorials

Contact

catalog...

男裝 男裝 +

https://catalog.hkstore.com/catalog\_to\_hk/men?p=2

Click each s... 訂單淨值滿\$400即享免費門市取貨 Select Mode

CATALOG MARATHON SPORTS HK

最新商品 品牌 聯乘系列 男裝 女裝 童裝 電子禮品卡 SALE

分類 +

熱賣系列 +

商品分類 +

鞋款 +

服裝類別 +

配件類別 +

袋款 +

品牌 +

性別 +

New Balance 2002R 中性球鞋

New Balance 2002R 中性球鞋

New Balance 2002R 中性球鞋

CSV/Excel JSON CSV/Excel Wide (beta)

selection1_name	selection1_series	selection1_style	selection1_price
New Balance	2002R	中性球鞋	\$1,199
New Balance	2002R	中性球鞋	\$1,199
New Balance	2002R	中性球鞋	\$1,199
PLASTIC THING X CATALOG系列	Plastic Thing 為食妹刺繡圖案外套	中性外套	\$699

This a live preview. When you are ready to run your project, click Get Data.

Show more data  Visuals enabled (advanced)

**catalog... BROWSE**

**Back to Commands**

**Project title**: catalog.hkstore.com Project

**Starting Site**: https://catalog.hkstore.com/cat

**Starting Template**: main\_template

**Web Hook**:

**Project Token**: You must save your project first  
tkOX156BQ5W2

**Max Workers**: 0

**Max Pages**:  2

**Enable Email Notifications**:

**Load JavaScript & Images**:

**Rotate IP Addresses**:

**Starting Value**:

**Max # of pages to be scraped.**

男裝 男裝

https://catalog.hkstore.com/catalog\_to\_hk/men?p=2

Click each s... 標題

訂單淨值滿\$400即享免費門市取貨 Select Mode

CATALOG MARATHON SPORTS

最新商品 品牌 聯乘系列 男裝 女裝 童裝 電子禮品卡 SALE

分類 +

熱賣系列 +

商品分類 +

鞋款 +

服裝類別 +

配件類別 +

袋款 +

品牌 +

性別 +

New Balance 2002R 中性球鞋

New Balance 2002R 中性球鞋

New Balance 2002R 中性球鞋

CSV/Excel JSON CSV/Excel Wide (beta)

selection1_name	selection1_series	selection1_style	selection1_price
New Balance	2002R	中性球鞋	\$1,199
New Balance	2002R	中性球鞋	\$1,199
New Balance	2002R	中性球鞋	\$1,199
PLASTIC THING X CATALOG系列	Plastic Thing 為食妹刺繡圖案外套	中性外套	\$699

This a live preview. When you are ready to run your project, click Get Data.

Show more data ?  Visuals enabled (advanced) ?

catalog.hkstore.com Project

Edit project

Your data is ready! Click on the green buttons to download.

Download Data

CSV/Excel

JSON

API

Template Name  
main\_template

Pages Scrapped  
2

# of pages scraped.

All dates and times are in UTC +0000.

Empty file with no results? [Click here](#) to fix.

CSV file too big? Save the JSON file and [click here](#) to convert to CSV.

## Details

Status	complete
Pages	2 collected
Initialized	2022-10-03T02:55:52
Start Time	2022-10-03T02:55:53
Finished	2022-10-03T02:56:29
API Key	tkOX156BQ5W2
Project Token	t2hdrZsNoUGS
Run Token	tuSJ0HKZxiyo

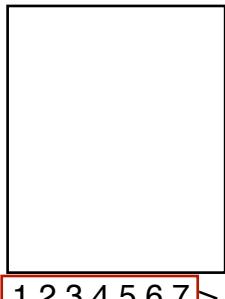
## Settings

URL	<input type="text" value="https://catalog.hkstore.com/catalog_tc_hk/men"/>
Starting Template	main_template
Starting Value	<input type="text" value="0"/>
Load Javascript	true
Rotate IPs	false

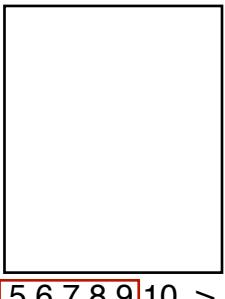


**Periodic scraping in smaller batches.**

**Monday  
morning**

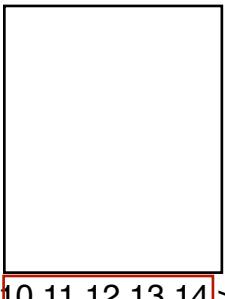


1 2 3 4 5 6 7 >

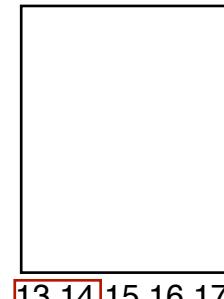


5 6 7 8 9 10 >

**Monday  
afternoon**

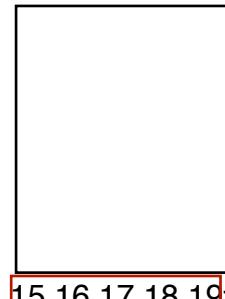


10 11 12 13 14 >

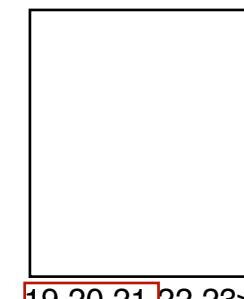


13 14 15 16 17 >

**Tuesday**



15 16 17 18 19 >



19 20 21 22 23 >

**More Tutorials Can be Found Below**

<https://www.youtube.com/channel/UCRfbcJQu9HZvc7uks2wmGzw/playlists>

# **Legal and Ethical Considerations.**

scraping and how they play out.

Share this



## Is Web Scraping Legal?

Your first thought might be to look at the legal side of things.

The truth is that the legality of web scraping is still relatively up in the air.

Meaning that there are currently no specific laws that refer to the legality of web scraping. So it is neither legal or illegal.

<https://www.parsehub.com/blog/web-scraping-ethical/>

# Scraping Publicly Available Information

Another factor to keep in mind is the type of data you'd be scraping. In our case, we always refer to publicly available data.

This is data that has been made public by the owner of said data. Private and leaked information is not considered publicly available information.

# Ethics in Web Scraping



James Densmore Jul 23, 2017 · 3 min read



We all scrape web data. Well, those of us who work with data do. Data scientists, marketers, data journalists, and the data curious alike. Lately, I've been thinking more about the ethics of the practice and have been dissatisfied by the lack of consensus on the topic.

Let me be clear that I'm talking **ethics** not the law. The law in regards to scraping web data is complex, fuzzy and ripe for reform, but that's another matter. It's not that no one is thinking, or writing, about the ethics in scraping but rather that both those scraping and those being scraped can't

<https://towardsdatascience.com/ethics-in-web-scraping-b96b18136f01>

## The Ethical Scraper

I, the web scraper will live by the following principles:

WRITTEN BY

James Densmore

Data Science and  
Data Engineering  
Consultant at Data  
Liftoff  
<https://www.dataliftoff.com>

Follow

---

397

5

0

- If you have a public API that provides the data I'm looking for, I'll use it and avoid scraping all together.
- I will always provide a User Agent string that makes my intentions clear and provides a way for you to contact me with questions or concerns.
- I will request data at a reasonable rate. I will strive to never be confused for a DDoS attack.
- I will only save the data I absolutely need from your page. If all I need is OpenGraph meta-data, that's all I'll keep.
- I will respect any content I do keep. I'll never pass it off as my own.
- I will look for ways to return value to you. Maybe I can drive some (real) traffic to your site or credit you in an article or post.
- I will respond in a timely fashion to your outreach and work with you towards a resolution.
- I will scrape for the **purpose of creating new value from the data**, not to duplicate it.

# Clean up data in **OpenRefine**



A free, open source,  
powerful tool for working  
with messy data



[Home](#)  
[Community](#)  
[Documentation](#)  
**[Download](#)**  
[Data Privacy](#)  
[Contact Us](#)  
[Blog](#)

## Download

On this page you will find a list of OpenRefine distributions and extensions available for download. Are we missing something? Want to fix a typo? You can [submit changes](#).

### Official Distribution

Read the [installation instructions](#).

You can also download all official releases and source from our [GitHub releases page](#)

#### OpenRefine 3.5.2

The latest stable release of OpenRefine 3.5, released on January 26, 2021. Please [backup your workspace directory](#) before installing and report any problems that you encounter. A change log is provided on [the release page](#).

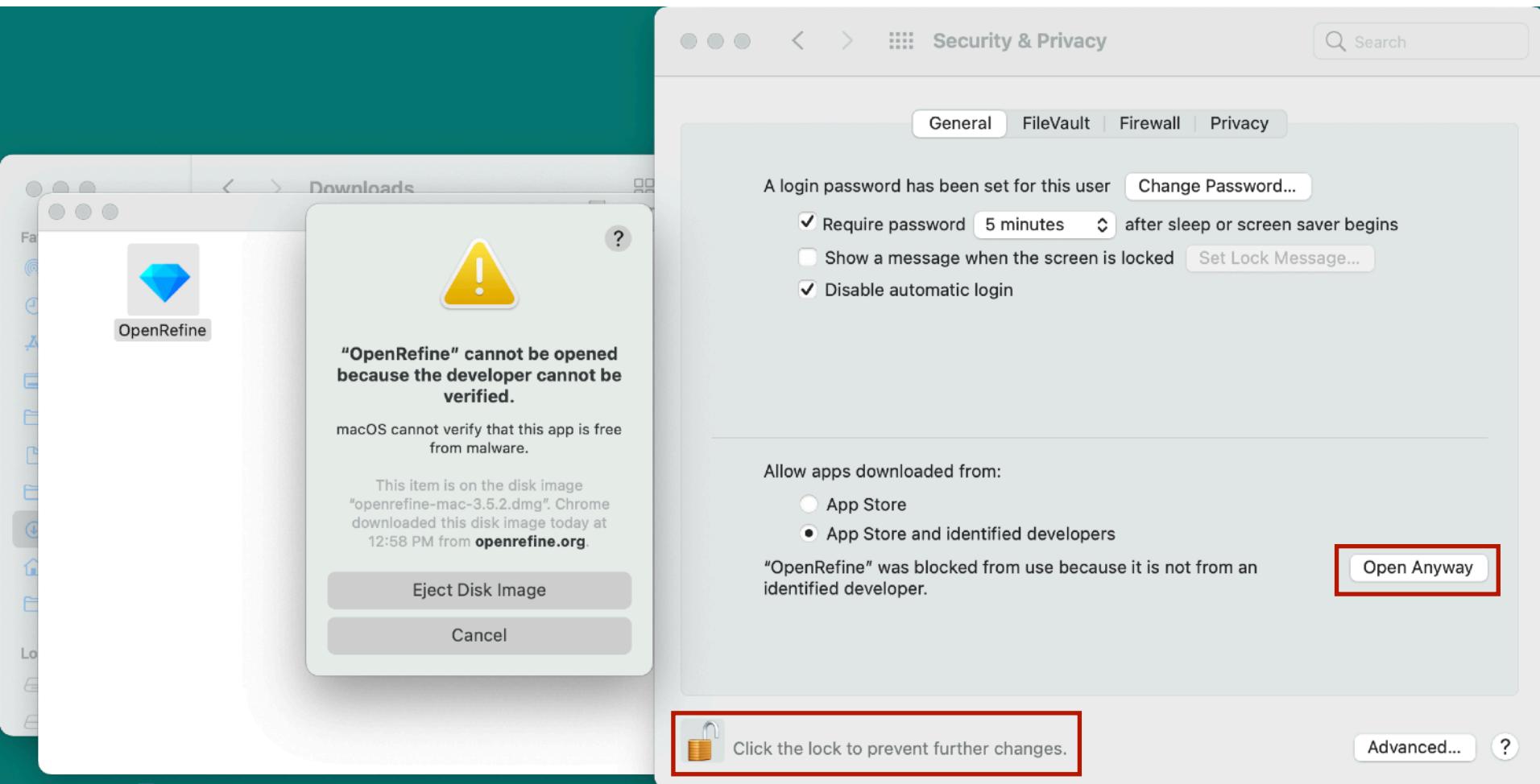
- **Windows kit**, This requires Java to be installed on your computer. Download, unzip, and double-click on `openrefine.exe` or `refine.bat` if the former does not work.
- **Windows kit with embedded Java**, includes [OpenJDK Java](#), available under the [GPLv2+CE](#) license. Download, unzip, and double-click on `openrefine.exe` or `refine.bat` if the former does not work.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it. You do not need to install Java separately.
- **Linux kit**, Download, extract, then type `./refine` to start. This requires Java to be installed on your computer.

#### OpenRefine 3.4.1

The previous stable release of OpenRefine, released on September 24, 2020. Please [backup your workspace directory](#) before installing and report any problems that you encounter. A change log is provided on [the release page](#).

- **Windows kit**, This requires Java to be installed on your computer. Download,

<https://openrefine.org/download.html>



## **Most Basic Data Cleaning Tasks Part 1**

- 1. Resolve inconsistent entries.**
- 2. Rename, combine and split columns for making further data presentation and processing easier.**
- 3. Transform incorrect code (e.g. wrong data types such as text instead of numeric and calculations)**

# **Tutorial**

<http://d3-media.blogspot.hk/2013/11/how-to-refine-your-data.html>

# **Practice Dataset**

<https://raw.githubusercontent.com/suentze2020/datahack/main/demo-refine-Sheet.csv>

## OpenRefine refine Sheet1 csv Permalink

Facet / Filter Undo / Redo 0 / 0

### Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?  
[Watch these screencasts](#)

## Cleaning column 1: company names

Click on the drop down menu of header of column 1, company, and choose text facet.

**25 rows**

Show as: rows records Show: 5 10 25 50 rows

All	company	Product code / n	address	city	country	name
1.	Facet		Text facet	arnhem	the netherlands	dhr p. jansen
2.		Text filter	Numeric facet	arnhem	the netherlands	dhr p. hansen
3.		Edit cells	Timeline facet	arnhem	the netherlands	dhr j. Gansen
4.		Edit column	Scatterplot facet	arnhem	the netherlands	dhr p. mansen
5.		Transpose	Custom text facet...	arnhem	the netherlands	dhr p. franssen
6.		Sort...	Custom Numeric Facet...	arnhem	the netherlands	dhr p. bansen
7.		View	Customized facets	arnhem	the netherlands	dhr p. vansen
8.			Leeuwardenweg 180	arnhem	the netherlands	dhr p. bransen
9.			Leeuwardenweg 181	arnhem	the netherlands	dhr p. janssen
10.		Reconcile	Leeuwardenweg 182	arnhem	the netherlands	mevr l. rokken
11.	ak zo	q-5	Leeuwardenweg 183	arnhem	the netherlands	mevr l. lokken
12.	akzo	q-9	Leeuwardenweg 184	arnhem	the netherlands	mevr l. mokken
13.	akzo	x-8	Delfzijlstraat 54	arnhem	the netherlands	mevr l. mokken
14.	phillips	p-56	Delfzijlstraat 55	arnhem	the netherlands	mevr l. mokken
15.	fillips	v-67	Delfzijlstraat 56	arnhem	the netherlands	mevr l. mokken
16.	philips	v-21	Delfzijlstraat 57	arnhem	the netherlands	mevr l. sokken
17.	Van Houten	x-45	Delfzijlstraat 58	arnhem	the netherlands	mevr l. wokken
18.	van Houten	v-56	Delfzijlstraat 59	arnhem	the netherlands	mevr l. kokken
19.	van houten	v-65	Delfzijlstraat 60	arnhem	the netherlands	mevr l. Bokken
20.	van houten	x-21	Delfzijlstraat 61	arnhem	the netherlands	mevr l. dokken
21.	Van Houten	p-23	Jourestraat 23	arnhem	the netherlands	mevr l. gokken
22.	unilver	x-3	Jourestraat 24	arnhem	the netherlands	mevr l. stokken
23.	unilever	q-4	Jourestraat 25	arnhem	the netherlands	mevr l. rokken
24.	Unilever	q-6	Jourestraat 26	arnhem	the netherlands	mevr l. rokken
25.	unilever	q-8				

**OpenRefine refine Sheet1 csv** Permalink

Facet / Filter Undo / Redo 0 / 0

Refresh Reset All Remove All

**company**

19 choices Sort by: name count

**Cluster**

**25 rows**

Show as: rows records Show: 5 10 25 50 rows

	company	Product code / n	address	city	country	name
1.	Philips	p-5	Groningen singel 147	armhem	the netherlands	dhr p
2.	phillips	p-43	Groningen singel 148	armhem	the netherlands	dhr p
3.	philips	x-3	Groningen singel 149	armhem	the netherlands	dhr p
4.	philips	x-34	Groningen singel 150	armhem	the netherlands	dhr p
5.	philips	x-12	Groningen singel 151	armhem	the netherlands	dhr p
6.	philipS	p-23	Groningen singel 152	armhem	the netherlands	dhr p
7.	akzo	v-43	Leeuwardenweg 178	armhem	the netherlands	dhr p
8.	Akzo	v-12	Leeuwardenweg 179	armhem	the netherlands	dhr p
9.	AKZO	x-5	Leeuwardenweg 180	armhem	the netherlands	dhr p
10.	akzo	p-34	Leeuwardenweg 181	armhem	the netherlands	dhr p
11.	ak zo	q-5	Leeuwardenweg 182	armhem	the netherlands	mevr
12.	akzo	q-9	Leeuwardenweg 183	armhem	the netherlands	mevr
13.	akzo	x-8	Leeuwardenweg 184	armhem	the netherlands	mevr
14.	philips	p-56	Delfzijlstraat 54	armhem	the netherlands	mevr
15.	fillips	v-67	Delfzijlstraat 55	armhem	the netherlands	mevr
16.	philips	v-21	Delfzijlstraat 56	armhem	the netherlands	mevr
17.	Van Houten	x-45	Delfzijlstraat 57	armhem	the netherlands	mevr
18.	van Houten	v-56	Delfzijlstraat 58	armhem	the netherlands	mevr
19.	van houten	v-65	Delfzijlstraat 59	armhem	the netherlands	mevr
20.	van houten	x-21	Delfzijlstraat 60	armhem	the netherlands	mevr
21.	Van Houten	p-23	Delfzijlstraat 61	armhem	the netherlands	mevr
22.	uniliver	x-3	Jourestraat 23	armhem	the netherlands	mevr
23.	unilever	q-4	Jourestraat 24	armhem	the netherlands	mevr
24.	Unilever	q-6	Jourestraat 25	armhem	the netherlands	mevr
25.	unilever	q-8	Jourestraat 26	armhem	the netherlands	mevr

**OpenRefine refine Sheet1 csv** Permalink

Facet / Filter Undo / Redo 0 / 0

Refresh Reset All Remove All

**company**

19 choices Sort by: name count

**Cluster**

**25 rows**

Show as: rows records Show: 5 10 25 50 rows

**Cluster & Edit column "company"**

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, "Unilever" and "Unilever" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person.

Method: key collision Keying Function: fingerprint

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	5	• akzo (3 rows) • AKZO (1 rows) • Akzo (1 rows)	<input type="checkbox"/>	akzo
3	5	• Van Houten (2 rows) • van houten (2 rows) • van Houten (1 rows)	<input type="checkbox"/>	Van Houten
3	4	• philips (2 rows) • Phillips (1 rows) • philipS (1 rows)	<input type="checkbox"/>	philips
2	3	• unilever (2 rows) • Unilever (1 rows)	<input type="checkbox"/>	unilever
1	1	• philips (1 rows)	<input type="checkbox"/>	philips
1	1	• ak zo (1 rows)	<input type="checkbox"/>	ak zo
1	1	• philips (1 rows)	<input type="checkbox"/>	philips

Select All Unselect All Export Clusters Merge Selected & Re-Cluster

On the left you see the companies listed. Click on cluster and choose 'key collision' and 'finger print' for clustering; that is correcting the names. Choose 'merge selected & re cluster'.

### Cluster & Edit column "company"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method key collision

Keying Function **fingerprint**

8 clusters found

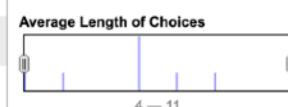
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
1	6	• AKZO7 (6 rows)	<input type="checkbox"/>	AKZO7
1	3	• Unilever4 (3 rows)	<input type="checkbox"/>	Unilever4
1	7	• philips9 (7 rows)	<input type="checkbox"/>	philips9
1	1	• fillips (1 rows)	<input type="checkbox"/>	fillips
1	1	• philips (1 rows)	<input type="checkbox"/>	philips
1	5	• Van Houten5 (5 rows)	<input type="checkbox"/>	Van Houten5
1	1	• akz0 (1 rows)	<input type="checkbox"/>	akz0
1	1	• univer (1 rows)	<input type="checkbox"/>	univer

Select All

Unselect All

Export Clusters

Merge Selected & Re-Clu



### Cluster & Edit column "company"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method key collision

Keying Function **ngram-fingerprint**

Ngram Size 2

4 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
1	9	• philips9 (9 rows)	<input type="checkbox"/>	philips9
1	5	• Van Houten5 (5 rows)	<input type="checkbox"/>	Van Houten5
1	4	• Unilever4 (4 rows)	<input type="checkbox"/>	Unilever4
1	7	• AKZO7 (7 rows)	<input type="checkbox"/>	AKZO7

Select All

Unselect All

Export Clusters

Merge Selected & Re-Cluster

Merge Selected & Close Close

# Rows in Cluster

4 — 9

Average Length of Choices

5 — 11

Continue clustering with a new algorithm: ngram-fingerprint. Continue till you have the following names:

AKZO7  
philips9  
Unilever4  
Van Houten5

Facet / Filter   Undo / Redo 3 / 3

Refresh   Reset All   Remove All

company change

4 choices Sort by: name count Cluster

AKZO7 7  
philips9 9  
Unilever4 4  
Van Houten5 5  
Facet by choice counts

25 rows

Show as: rows records   Show: 5 10 25 50 rows

All	company	Product code / n	address	city	country	name	
1.	Facet		Groningsingel 147	arnhem	the netherlands	dhr p. jansen	
2.	Text filter		Groningsingel 148	arnhem	the netherlands	dhr p. hansen	
3.			Groningsingel 149	arnhem	the netherlands	dhr j. Gansen	
4.	Edit cells		Transform...				
5.	Edit column		Common transforms				
6.	Transpose						
7.	Sort...		Trim leading and trailing whitespace				
8.	View		Collapse consecutive whitespace				
9.			Unescape HTML entities				
10.	Reconcile		Replace Smart quotes with ascii				
11.	AKZO7	q-5	To titlecase				
12.	AKZO7	q-9	To uppercase				
13.	AKZO7	x-8	To lowercase				
14.	philips9	p-56	To number				
15.	philips9	v-67	To date				
16.	philips9	v-21	To text				
17.	Van Houten5	x-45	To null				
18.	Van Houten5	v-56	To empty string				
19.	Van Houten5	v-65					
20.	Van Houten5	x-21	Delfzijlstraat 54	arnhem	the netherlands	mevr l. Bokken	
21.	Van Houten5	p-23	Delfzijlstraat 55	arnhem	the netherlands	mevr l. dokken	
22.	Unilever4	x-3	Delfzijlstraat 56	arnhem	the netherlands	mevr l. gokken	
23.	Unilever4	q-4	Delfzijlstraat 57	arnhem	the netherlands	mevr l. stokken	
24.	Unilever4	q-6	Jourestraat 23	arnhem	the netherlands	mevr l. rokken	
25.	Unilever4	q-8	Jourestraat 24	arnhem	the netherlands	mevr l. rokken	
			Jourestraat 25	arnhem	the netherlands	mevr l. rokken	
			Jourestraat 26	arnhem	the netherlands	mevr l. rokken	

Now we have the correct spelling of the company names; let's turn them into uppercase. Choose from column company in the drop down menu edit cells, common transformations, into uppercase.

OpenRefine refine Sheet1 csv Permalink

Facet / Filter Undo / Redo 4 / 4

Text transform on 18 cells in column company:  
value.toUpperCase() Undo

Open... Export... Help

Extensions: Wikidata ▾

25 rows

Show as: rows records Show: 5 10 25 50 rows

« first < previous 1 - 25 next > last »

	company	product code / n	address	city	country	name
1.	PHILIPS9	p-5	Groningen singel 147	arnhem	the netherlands	dhr p. jansen
2.	PHILIPS9	p-43	Groningen singel 148	arnhem	the netherlands	dhr p. hansen
3.	PHILIPS9	x-3	Groningen singel 149	arnhem	the netherlands	dhr j. Gansen
4.	PHILIPS9	x-34	Groningen singel 150	arnhem	the netherlands	dhr p. mansen
5.	PHILIPS9	x-12	Groningen singel 151	arnhem	the netherlands	dhr p. fransen
6.	PHILIPS9	p-23	Groningen singel 152	arnhem	the netherlands	dhr p. franssen
7.	AKZO7	v-43	Leeuwardenweg 178	arnhem	the netherlands	dhr p. bansen
8.	AKZO7	v-12	Leeuwardenweg 179	arnhem	the netherlands	dhr p. vansen
9.	AKZO7	x-5	Leeuwardenweg 180	arnhem	the netherlands	dhr p. bransen
10.	AKZO7	p-34	Leeuwardenweg 181	arnhem	the netherlands	dhr p. janssen
11.	AKZO7	q-5	Leeuwardenweg 182	arnhem	the netherlands	mevr l. rokken
12.	AKZO7	q-9	Leeuwardenweg 183	arnhem	the netherlands	mevr l. lokken
13.	AKZO7	x-8	Leeuwardenweg 184	arnhem	the netherlands	mevr l. mokken
14.	PHILIPS9	p-56	Delfzijlstraat 54	arnhem	the netherlands	mevr l. mokken
15.	PHILIPS9	v-67	Delfzijlstraat 55	arnhem	the netherlands	mevr l. mokken
16.	PHILIPS9	v-21	Delfzijlstraat 56	arnhem	the netherlands	mevr l. mokken
17.	VAN HOUTEN5	x-45	Delfzijlstraat 57	arnhem	the netherlands	mevr l. sokken
18.	VAN HOUTEN5	v-56	Delfzijlstraat 58	arnhem	the netherlands	mevr l. wokken
19.	VAN HOUTEN5	v-65	Delfzijlstraat 59	arnhem	the netherlands	mevr l. kokken
20.	VAN HOUTEN5	x-21	Delfzijlstraat 60	arnhem	the netherlands	mevr l. Bokken
21.	VAN HOUTEN5	p-23	Delfzijlstraat 61	arnhem	the netherlands	mevr l. dokken
22.	UNILEVER4	x-3	Jourestraat 23	arnhem	the netherlands	mevr l. gokken
23.	UNILEVER4	q-4	Jourestraat 24	arnhem	the netherlands	mevr l. stokken
24.	UNILEVER4	q-6	Jourestraat 25	arnhem	the netherlands	mevr l. rokken
25.	UNILEVER4	q-8	Jourestraat 26	arnhem	the netherlands	mevr l. rokken

All company names are now capitalised.

Facet / Filter Undo / Redo 4 / 5 Extract... Apply...

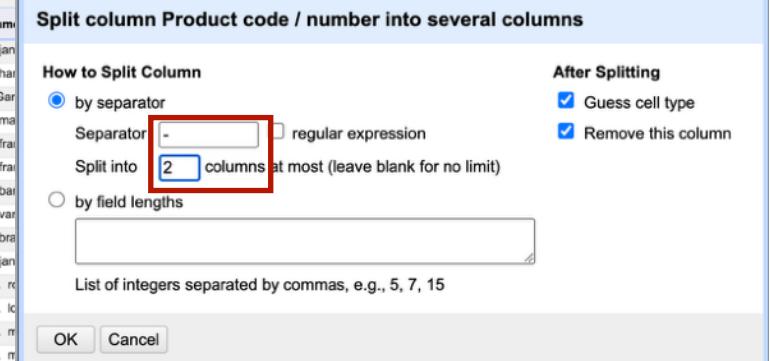
Facet / Filter Undo / Redo 7 / 7 Refresh Reset All Remove All

**company** change  
4 choices Sort by: name count Cluster

AZKO7 7  
PHILIPS9 9  
UNILEVER4 4  
VAN HOUTENS 5  
Facet by choice counts

25 rows Show as: rows records Show: 5 10 25 50 rows

		company	Product code / n	address	city	country	name
1.	PHILIPS9	Facet	Groningsingel 147	amhem	the netherlands	dhr p. jan	
2.	PHILIPS9	Text filter	Groningsingel 148	amhem	the netherlands	dhr p. har	
3.	PHILIPS9	Edit cells	Groningsingel 149	amhem	the netherlands	dhr j. Gar	
4.	PHILIPS9	Edit column	Groningsingel 150	amhem	the netherlands	dhr p. ma	
5.	PHILIPS9	Transpose				r p. fra	
6.	PHILIPS9	Sort...				r p. fra	
7.	AKZ07	View	Add column based on this column...			r p. bar	
8.	AKZ07		Add column by fetching URLs...			r p. var	
9.	AKZ07		Add columns from reconciled values...			r p. bra	
10.	AKZ07		Rename this column			r p. jan	
11.	AKZ07	q-5	Remove this column			wvr l. ro	
12.	AKZ07	Le	Move column to beginning			wvr l. ic	
13.	AKZ07	q-9	Move column to end			wvr l. m	
14.	PHILIPS9	x-8	Move column left			wvr l. m	
15.	PHILIPS9	p-56	Move column right			wvr l. mokken	
16.	PHILIPS9	v-67				r p. malink	
17.	VAN HOUTENS	v-21					
18.	VAN HOUTENS	x-45	Delfzijlstraat 57	amhem	the netherlands	mevr l. st	
19.	VAN HOUTENS	v-56	Delfzijlstraat 58	amhem	the netherlands	mevr l. w	
20.	VAN HOUTENS	v-65	Delfzijlstraat 59	amhem	the netherlands	mevr l. lk	
21.	VAN HOUTENS	x-21	Delfzijlstraat 60	amhem	the netherlands	mevr l. B	
22.	UNILEVER4	p-23	Delfzijlstraat 61	amhem	the netherlands	mevr l. di	
23.	UNILEVER4	x-3	Jourestraat 23	amhem	the netherlands	mevr l. gr	
24.	UNILEVER4	q-4	Jourestraat 24	amhem	the netherlands	mevr l. st	
25.	UNILEVER4	q-6	Jourestraat 25	amhem	the netherlands	mevr l. rc	
		q-8	Jourestraat 26	amhem	the netherlands	mevr l. rc	



25 rows Show as: rows records Show: 5 10 25 50 rows

		company	Product code / n	address	city	country	name
1.	PHILIPS9	Facet	Groningsingel 147	amhem	the netherlands	dhr j. Gansen	
2.	PHILIPS9	Text filter	Groningsingel 148	amhem	the netherlands	dhr p. mansen	
3.	PHILIPS9	Edit cells	Groningsingel 149	amhem	the netherlands	dhr p. franssen	
4.	PHILIPS9	Edit column	Groningsingel 150	amhem	the netherlands	dhr p. bansen	
5.	PHILIPS9	Transpose				dhr p. vansen	
6.	PHILIPS9	Sort...				dhr p. bransen	
7.	AKZ07	View	Add column based on this column...			dhr p. janssen	
8.	AKZ07		Add column by fetching URLs...			mevr l. rokken	
9.	AKZ07		Add columns from reconciled values...			mevr l. lokken	
10.	AKZ07		Rename this column			mevr l. mokken	
11.	AKZ07	q	Remove this column			mevr l. mokken	
12.	AKZ07	q	Move column to beginning			mevr l. mokken	
13.	AKZ07	x	Move column to end			mevr l. mokken	
14.	PHILIPS9	p	Move column left			mevr l. mokken	
15.	PHILIPS9	v	Move column right			mevr l. mokken	
16.	PHILIPS9	v					

**Split column 2 product name and number. Go to edit column in the drop down menu and choose split into several columns. Use – as the separator and choose 2 columns. Give column 3 a new name: choose edit column and rename: number. Change the name of column 2 in product, by choosing in the drop down menu: edit column and rename column.**

Facet / Filter   Undo / Redo 7 / 7

Refresh   Reset All   Remove All

company change

4 choices Sort by: name count Cluster

AKZO7 7  
PHILIPS9 9  
UNILEVER4 4  
VAN HOUTEN5 5  
Facet by choice counts

<input checked="" type="checkbox"/>	All	<input checked="" type="checkbox"/>	company	<input checked="" type="checkbox"/>	Product	<input checked="" type="checkbox"/>	Number	<input checked="" type="checkbox"/>	address	<input checked="" type="checkbox"/>	city	<input checked="" type="checkbox"/>	country	<input checked="" type="checkbox"/>	name
1.	PHILIPS9	p		5	Groningen singel 147	arnhem	the netherlands	dhr p. jansen							
2.	PHILIPS9	p		43	Groningen singel 148	arnhem	the netherlands	dhr p. hansen							
3.	PHILIPS9	x		3	Groningen singel 149	arnhem	the netherlands	dhr j. Gansen							
4.	PHILIPS9	x		34	Groningen singel 150	arnhem	the netherlands	dhr p. mansen							
5.	PHILIPS9	x		12	Groningen singel 151	arnhem	the netherlands	dhr p. franssen							
6.	PHILIPS9	x		23	Groningen singel 152	arnhem	the netherlands	dhr p. franssen							
7.	AKZO7	v		43	Leeuwardenweg 178	arnhem	the netherlands	dhr p. bansen							
8.	AKZO7	v		12	Leeuwardenweg 179	arnhem	the netherlands	dhr p. vansen							
9.	AKZO7	x		5	Leeuwardenweg 180	arnhem	the netherlands	dhr p. bransen							
10.	AKZO7	p		34	Leeuwardenweg 181	arnhem	the netherlands	dhr p. janssen							
11.	AKZO7	q		5	Leeuwardenweg 182	arnhem	the netherlands	mevr l. rokken							
12.	AKZO7	q		9	Leeuwardenweg 183	arnhem	the netherlands	mevr l. lokken							
13.	AKZO7	x		8	Leeuwardenweg 184	arnhem	the netherlands	mevr l. mokken							
14.	PHILIPS9	p		56	Delfzijlstraat 54	arnhem	the netherlands	mevr l. mokken							
15.	PHILIPS9	v		67	Delfzijlstraat 55	arnhem	the netherlands	mevr l. mokken							
16.	PHILIPS9	v		21	Delfzijlstraat 56	arnhem	the netherlands	mevr l. mokken							
17.	VAN HOUTEN5	x		45	Delfzijlstraat 57	arnhem	the netherlands	mevr l. sokken							
18.	VAN HOUTEN5	v		56	Delfzijlstraat 58	arnhem	the netherlands	mevr l. wokken							
19.	VAN HOUTEN5	v		65	Delfzijlstraat 59	arnhem	the netherlands	mevr l. kokken							
20.	VAN HOUTEN5	x		21	Delfzijlstraat 60	arnhem	the netherlands	mevr l. Bokken							
21.	VAN HOUTEN5	p		23	Delfzijlstraat 61	arnhem	the netherlands	mevr l. dokken							
22.	UNILEVER4	x		3	Jourestraat 23	arnhem	the netherlands	mevr l. gokken							
23.	UNILEVER4	q		4	Jourestraat 24	arnhem	the netherlands	mevr l. stokken							
24.	UNILEVER4	q		6	Jourestraat 25	arnhem	the netherlands	mevr l. rokken							
25.	UNILEVER4	q		8	Jourestraat 26	arnhem	the netherlands	mevr l. rokken							

Product code and number are now separated.

In column 2 we have a code for the product but not the name of the **product**. Here is the conversion:

p=radio

v=tv

x=computer

q=tablet

Go in column 2 to drop down menu, choose edit cells and transform. Now we have to write some code in order to the transformation:

Use the following expression:

value.replace("p", "radio")

The left column show the original and the right the result of the transformation.

Do this for all different products.

company	Product	Number	address	city	country	name
AKZO7	1. PHILIPS9	5	Groningen singel 147	arnhem	the netherlands	dhr p. jansen
	2. PHILIPS9	43	Groningen singel 148	arnhem	the netherlands	dhr p. hansen
	3. PHILIPS9				the netherlands	dhr j. Gansen
VAN HOUTENS	4. PHILIPS9				the netherlands	dhr p. mansen
	5. PHILIPS9				the netherlands	dhr p. franssen
	6. PHILIPS9				the netherlands	dhr p. franssen
	7. AKZO7				the netherlands	dhr p. banzen
	8. AKZO7				the netherlands	dhr p. vansen
	9. AKZO7				the netherlands	dhr p. branson
	10. AKZO7				the netherlands	dhr p. janssen
	11. AKZO7				the netherlands	mevr l. lokken
	12. AKZO7				the netherlands	mevr l. mokken
	13. AKZO7				the netherlands	mevr l. mokken
	14. PHILIPS9	56	Delfzijlstraat 54	arnhem	the netherlands	mevr l. mokken
	15. PHILIPS9	67	Delfzijlstraat 55	arnhem	the netherlands	mevr l. mokken
	16. PHILIPS9	21	Delfzijlstraat 56	arnhem	the netherlands	mevr l. mokken

Text transform on 5 cells in column Product:  
grel:value.replace("q","tablet") [Undo](#)

Facet / Filter [Undo / Redo](#) 11 / 11

Refresh [Reset All](#) [Remove All](#)

company [change](#)

4 choices Sort by: name count [Cluster](#)

AKZO7 7  
PHILIPS9 9  
UNILEVER4 4  
VAN HOUTEN5 5  
Facet by choice counts

**25 rows**

Show as: [rows](#) [records](#) Show: [5](#) [10](#) [25](#) [50](#) rows

<input checked="" type="checkbox"/> All	<input checked="" type="checkbox"/> company	<input checked="" type="checkbox"/> Product	<input checked="" type="checkbox"/> Number	<input checked="" type="checkbox"/> address	<input checked="" type="checkbox"/> city	<input checked="" type="checkbox"/> country	<input checked="" type="checkbox"/> name
1.	PHILIPS9	radio	5	Groningen singel 147	arnhem	the netherlands	dhr p. jansen
2.	PHILIPS9	radio	43	Groningen singel 148	arnhem	the netherlands	dhr p. hansen
3.	PHILIPS9	computer	3	Groningen singel 149	arnhem	the netherlands	dhr j. gansen
4.	PHILIPS9	computer	34	Groningen singel 150	arnhem	the netherlands	dhr p. mansen
5.	PHILIPS9	computer	12	Groningen singel 151	arnhem	the netherlands	dhr p. fransen
6.	PHILIPS9	radio	23	Groningen singel 152	arnhem	the netherlands	dhr p. franssen
7.	AKZO7	tv	43	Leeuwardenweg 178	arnhem	the netherlands	dhr p. bansen
8.	AKZO7	tv	12	Leeuwardenweg 179	arnhem	the netherlands	dhr p. vansen
9.	AKZO7	computer	5	Leeuwardenweg 180	arnhem	the netherlands	dhr p. bransen
10.	AKZO7	radio	34	Leeuwardenweg 181	arnhem	the netherlands	dhr p. janssen
11.	AKZO7	tablet	5	Leeuwardenweg 182	arnhem	the netherlands	mevr l. rokken
12.	AKZO7	tablet	9	Leeuwardenweg 183	arnhem	the netherlands	mevr l. lokken
13.	AKZO7	computer	8	Leeuwardenweg 184	arnhem	the netherlands	mevr l. mokken
14.	PHILIPS9	radio	56	Delfzijlstraat 54	arnhem	the netherlands	mevr l. mokken
15.	PHILIPS9	tv	67	Delfzijlstraat 55	arnhem	the netherlands	mevr l. mokken
16.	PHILIPS9	tv	21	Delfzijlstraat 56	arnhem	the netherlands	mevr l. mokken
17.	VAN HOUTEN5	computer	45	Delfzijlstraat 57	arnhem	the netherlands	mevr l. sokken
18.	VAN HOUTEN5	tv	56	Delfzijlstraat 58	arnhem	the netherlands	mevr l. wokken
19.	VAN HOUTEN5	tv	65	Delfzijlstraat 59	arnhem	the netherlands	mevr l. kokken
20.	VAN HOUTEN5	computer	21	Delfzijlstraat 60	arnhem	the netherlands	mevr l. Bokken
21.	VAN HOUTEN5	radio	23	Delfzijlstraat 61	arnhem	the netherlands	mevr l. dokken
22.	UNILEVER4	computer	3	Jourestraat 23	arnhem	the netherlands	mevr l. gokken
23.	UNILEVER4	tablet	4	Jourestraat 24	arnhem	the netherlands	mevr l. stokken
24.	UNILEVER4	tablet	6	Jourestraat 25	arnhem	the netherlands	mevr l. rokken
25.	UNILEVER4	tablet	8	Jourestraat 26	arnhem	the netherlands	mevr l. rokken

Product renamed to more meaningful representation.

Facet / Filter Undo / Redo 11 / 11

Refresh Reset All Remove All

company change  
4 choices Sort by: name count Cluster

AKZO7 7  
PHILIPS9 9  
UNILEVER4 4  
VAN HOUTEN5 5  
Facet by choice counts

25 rows Show as: rows records Show: 5 10 25 50 rows

	company	Product	Number	address	city	country	name
1.	PHILIPS9	radio	5	Facet	nhem	the netherlands	dhr p. jansen
2.	PHILIPS9	radio	43	Text filter	nhem	the netherlands	dhr p. hansen
3.	PHILIPS9	computer	3	Edit cells	amsterdam	the netherlands	dhr p. jansen
4.	PHILIPS9	computer	34	Transform...	amsterdam	the netherlands	dhr p. jansen
5.	PHILIPS9	computer	12	Edit column	amsterdam	the netherlands	dhr p. jansen
6.	PHILIPS9	radio	23	Transpose	amsterdam	the netherlands	dhr p. jansen
7.	AKZO7	tv	43	Sort...	Blank down	the netherlands	dhr p. jansen
8.	AKZO7	tv	12	View	Leeuwardenweg 182	the netherlands	dhr p. jansen
9.	AKZO7	computer	5	Reconcile	Leeuwardenweg 183	the netherlands	dhr p. jansen
10.	AKZO7	radio	34	Cluster and edit...	Leeuwardenweg 184	the netherlands	dhr p. jansen
11.	AKZO7	tablet	5	Replace	Delfzijlstraat 54	the netherlands	mehr l. mokken
12.	AKZO7	tablet	9		Delfzijlstraat 54	the netherlands	mehr l. mokken
13.	AKZO7	computer	8		Delfzijlstraat 54	the netherlands	mehr l. mokken
14.	PHILIPS9	radio	56		Delfzijlstraat 54	the netherlands	mehr l. mokken
15.	PHILIPS9	tv	67		Delfzijlstraat 54	the netherlands	mehr l. mokken
16.	PHILIPS9	tv	21		Delfzijlstraat 54	the netherlands	mehr l. mokken
17.	VAN HOUTEN5	computer	45		Delfzijlstraat 54	the netherlands	mehr l. mokken
18.	VAN HOUTEN5	tv	56		Delfzijlstraat 54	the netherlands	mehr l. mokken
19.	VAN HOUTEN5	tv	65		Delfzijlstraat 54	the netherlands	mehr l. mokken
20.	VAN HOUTEN5	computer	21		Delfzijlstraat 54	the netherlands	mehr l. mokken
21.	VAN HOUTEN5	radio	23		Delfzijlstraat 54	the netherlands	mehr l. mokken
22.	UNILEVER4	computer	3		Delfzijlstraat 54	the netherlands	mehr l. mokken
23.	UNILEVER4	tablet	4		Delfzijlstraat 54	the netherlands	mehr l. mokken
24.	UNILEVER4	tablet	6		Delfzijlstraat 54	the netherlands	mehr l. mokken
25.	UNILEVER4	tablet	8		Delfzijlstraat 54	the netherlands	mehr l. mokken

Custom text transform on column address

Expression: cells["address"].value + ", " + cells["city"].value + ", " + cells["country"].value

Language: General Refine Expression Language (GREL)

Preview History Starred Help

row value	cells["address"].value + ", ..."
1. Groningsingel 147	Groningsingel 147, amhem, the netherlands
2. Groningsingel 148	Groningsingel 148, amhem, the netherlands
3. Groningsingel 149	Groningsingel 149, amhem, the netherlands
4. Groningsingel 150	Groningsingel 150, amhem, the netherlands
5. Groningsingel 151	Groningsingel 151, amhem, the netherlands
6. Groningsingel 152	Groningsingel 152, amhem, the netherlands

On error:  keep original  set to blank  store error  Re-transform up to 10 times until no change

OK Cancel

In order to geocode we have to bring all address information into one column. First address and city. Go to address choose edit column, add column based on this column: use the following transformation:

```
cells["address"].value + ", " + cells["city"].value + ", " + cells["country"].value
cells['address'].value + ', ' + cells['city'].value + ', ' + cells['country'].value
```

25 rows

Show as: rows records Show: 5 10 25 50 rows

All	company	Product	Number	address	city	country	name
1.	PHILIPS9	radio	5	Groningsingel 147, arnhem, the netherlands	arnhem	the netherlands	dhr p. jansen
2.	PHILIPS9	radio	43	Groningsingel 148, arnhem, the netherlands	arnhem	the netherlands	dhr p. hansen
3.	PHILIPS9	computer	3	Groningsingel 149, arnhem, the netherlands	arnhem	the netherlands	dhr j. Gansen
4.	PHILIPS9	computer	34	Groningsingel 150, arnhem, the netherlands	arnhem	the netherlands	dhr p. mansen
5.	PHILIPS9	computer	12	Groningsingel 151, arnhem, the netherlands	arnhem	the netherlands	dhr p. franssen
6.	PHILIPS9	radio	23	Groningsingel 152, arnhem, the netherlands	arnhem	the netherlands	dhr p. franssen
7.	AKZO7	tv	43	Leeuwardenweg 178, arnhem, the netherlands	arnhem	the netherlands	dhr p. jansen
8.	AKZO7	tv	12	Leeuwardenweg 179, arnhem, the netherlands	arnhem	the netherlands	dhr p. hansen
9.	AKZO7	computer	5	Leeuwardenweg 180, arnhem, the netherlands	arnhem	the netherlands	dhr j. Gansen
10.	AKZO7	radio	34	Leeuwardenweg 181, arnhem, the netherlands	arnhem	the netherlands	dhr p. mansen
11.	AKZO7	tablet	5	Leeuwardenweg 182, arnhem, the netherlands	arnhem	the netherlands	dhr p. franssen
12.	AKZO7	tablet	9	Leeuwardenweg 183, arnhem, the netherlands	arnhem	the netherlands	dhr p. jansen
13.	AKZO7	computer	8	Leeuwardenweg 184, arnhem, the netherlands	arnhem	the netherlands	dhr p. hansen
14.	PHILIPS9	radio	56	Delfzijlstraat 54, arnhem, the netherlands	arnhem	the netherlands	dhr j. Gansen
15.	PHILIPS9	tv	67	Delfzijlstraat 55, arnhem, the netherlands	arnhem	the netherlands	dhr p. mansen
16.	PHILIPS9	tv	21	Delfzijlstraat 56, arnhem, the netherlands	arnhem	the netherlands	dhr p. franssen
17.	VAN HOUTEN5	computer	45	Delfzijlstraat 57, arnhem, the netherlands	arnhem	the netherlands	dhr p. jansen
18.	VAN HOUTEN5	tv	56	Delfzijlstraat 58, arnhem, the netherlands	arnhem	the netherlands	dhr p. hansen
19.	VAN HOUTEN5	tv	65	Delfzijlstraat 59, arnhem, the netherlands	arnhem	the netherlands	dhr j. Gansen
20.	VAN HOUTEN5	computer	21	Delfzijlstraat 60, arnhem, the netherlands	arnhem	the netherlands	dhr p. mansen
21.	VAN HOUTEN5	radio	23	Delfzijlstraat 61, arnhem, the netherlands	arnhem	the netherlands	dhr p. franssen
22.	UNILEVER4	computer	3	Jourestraat 23, arnhem, the netherlands	arnhem	the netherlands	dhr p. jansen
23.	UNILEVER4	tablet	4	Jourestraat 24, arnhem, the netherlands	arnhem	the netherlands	dhr p. hansen
24.	UNILEVER4	tablet	6	Jourestraat 25, arnhem, the netherlands	arnhem	the netherlands	dhr j. Gansen
25.	UNILEVER4	tablet	8	Jourestraat 26, arnhem, the netherlands	arnhem	the netherlands	dhr p. mansen

After combining the address, city, and country columns into one column, we can now remove the city and country columns.

25 rows  
Show as: rows records Show: 5 10 25 50 rows

All	company	Product	Number	address	city	country	name
1.	PHILIPS9	radio	5	Groningsingel 147, arnhem, the netherlands	arnhem	the netherlands	dhr p. jansen
2.	PHILIPS9	radio	43	Groningsingel 148, arnhem, the netherlands	arnhem	the netherlands	dhr p. hansen
3.	PHILIPS9	computer	3	Groningsingel 149, arnhem, the netherlands	arnhem	the netherlands	dhr j. Gansen
4.	PHILIPS9	computer	34	Groningsingel 150, arnhem, the netherlands	arnhem	the netherlands	dhr p. mansen
5.	PHILIPS9	computer	12	Groningsingel 151, arnhem, the netherlands	arnhem	the netherlands	dhr p. franssen
6.	PHILIPS9	radio	23	Groningsingel 152, arnhem, the netherlands	arnhem	the netherlands	dhr p. franssen
7.	AKZO7	tv	43	Leeuwardenweg 178, arnhem, the netherlands	arnhem	the netherlands	dhr p. jansen
8.	AKZO7	tv	12	Leeuwardenweg 179, arnhem, the netherlands	arnhem	the netherlands	dhr p. hansen
9.	AKZO7	computer	5	Leeuwardenweg 180, arnhem, the netherlands	arnhem	the netherlands	dhr j. Gansen
10.	AKZO7	radio	34	Leeuwardenweg 181, arnhem, the netherlands	arnhem	the netherlands	dhr p. mansen
11.	AKZO7	tablet	5	Leeuwardenweg 182, arnhem, the netherlands	arnhem	the netherlands	dhr p. franssen
12.	AKZO7	tablet	9	Leeuwardenweg 183, arnhem, the netherlands	arnhem	the netherlands	dhr p. jansen
13.	AKZO7	computer	8	Leeuwardenweg 184, arnhem, the netherlands	arnhem	the netherlands	dhr p. hansen
14.	PHILIPS9	radio	56	Delfzijlstraat 54, arnhem, the netherlands	arnhem	the netherlands	dhr j. Gansen
15.	PHILIPS9	tv	67	Delfzijlstraat 55, arnhem, the netherlands	arnhem	the netherlands	dhr p. mansen
16.	PHILIPS9	tv	21	Delfzijlstraat 56, arnhem, the netherlands	arnhem	the netherlands	dhr p. franssen
17.	VAN HOUTEN5	computer	45	Delfzijlstraat 57, arnhem, the netherlands	arnhem	the netherlands	dhr p. jansen
18.	VAN HOUTEN5	tv	56	Delfzijlstraat 58, arnhem, the netherlands	arnhem	the netherlands	dhr p. hansen
19.	VAN HOUTEN5	tv	65	Delfzijlstraat 59, arnhem, the netherlands	arnhem	the netherlands	dhr j. Gansen
20.	VAN HOUTEN5	computer	21	Delfzijlstraat 60, arnhem, the netherlands	arnhem	the netherlands	dhr p. mansen
21.	VAN HOUTEN5	radio	23	Delfzijlstraat 61, arnhem, the netherlands	arnhem	the netherlands	dhr p. franssen
22.	UNILEVER4	computer	3	Jourestraat 23, arnhem, the netherlands	arnhem	the netherlands	dhr p. jansen
23.	UNILEVER4	tablet	4	Jourestraat 24, arnhem, the netherlands	arnhem	the netherlands	dhr p. hansen
24.	UNILEVER4	tablet	6	Jourestraat 25, arnhem, the netherlands	arnhem	the netherlands	dhr j. Gansen
25.	UNILEVER4	tablet	8	Jourestraat 26, arnhem, the netherlands	arnhem	the netherlands	dhr p. mansen

Facet	► dhr p. jansen
Text filter	► dhr p. hansen
Edit cells	► dhr j. Gansen
<b>Edit column</b>	► dhr p. mansen
Transpose	► Split into several columns...
Sort...	► Add column based on this column...
View	► Add column by fetching URLs...
Reconcile	► Rename this column
Remove this column	► Move column to beginning
	► Move column to end
	► Move column left
	► Move column right

## **Most Basic Data Cleaning Part 2**

- 4. Remove duplicates.**
- 5. Remove blanks.**

Select: NBA\_Page()



Select Row



Extract Rank

Relative Pk

Relative Player

Relative College

Get Data

Selection Node:

p 1st body

2014 NBA Draft | Basketball-Reference.com

[https://www.basketball-reference.com/draft/NBA\\_2014.html](https://www.basketball-reference.com/draft/NBA_2014.html)

Sports Reference | Baseball | Football (college) | Basketball (college) | Hockey | Football | Blog | Stathead | Widgets

Login | Logout | Chat

Select Mode

Enter Person, Team, Section, etc

Search

BASKETBALL REFERENCE

Players Teams Seasons Leaders Scores (2) WNBA Draft Stathead Newsletter Full Site Menu Below ▾

## 2014 NBA Draft

« 2013 NBA Draft | 2015 NBA Draft »

Date: Thursday, June 26, 2014  
 Location: New York, New York  
 Number of Picks: 60 (53 played in NBA)  
 First Overall Pick: [Andrew Wiggins](#) (23.5 Win Shares)  
 Most Win Shares: [N. Jokić](#) (74.7), [C. Capela](#) (49.7) and [J. Embiid](#) (39.6)  
 All-Stars: 4 ([J. Embiid](#), [N. Jokić](#), [Z. LaVine](#) and [J. Randle](#))

 via Sports Logos.net About logos

HURRY, ENDS SOON 

Draft History Draft Years ▾

CSV/Excel	JSON	CSV/Excel Wide (beta)				
Row_Rank	Row_Pk	Row_Pk_url	Row_Player	Row_Player_url	Row_College	Row_College_url
1	1	<a href="https://stathead.com/basketball/draft_finder.cgi?request...year_min=&amp;year_max=&amp;round_min=&amp;round_max=&amp;order_by=rank&amp;sort_order=asc&amp;format=json&amp;rows=1000">https://stathead.com/basketball/draft_finder.cgi?request...year_min=&amp;year_max=&amp;round_min=&amp;round_max=&amp;order_by=rank&amp;sort_order=asc&amp;format=json&amp;rows=1000</a>	Andrew Wiggins	<a href="https://www.basketball-reference.com/players/w/wiggian01.html">https://www.basketball-reference.com/players/w/wiggian01.html</a>	Kansas	<a href="https://www.basketball-reference.com/friv/draft.fcgi?college=kansas&amp;year=2014&amp;round=1&amp;order_by=rank&amp;sort_order=asc&amp;format=json&amp;rows=1000">https://www.basketball-reference.com/friv/draft.fcgi?college=kansas&amp;year=2014&amp;round=1&amp;order_by=rank&amp;sort_order=asc&amp;format=json&amp;rows=1000</a>
2	2	<a href="https://stathead.com/basketball/draft_finder.cgi?request...year_min=&amp;year_max=&amp;round_min=&amp;round_max=&amp;order_by=rank&amp;sort_order=asc&amp;format=json&amp;rows=1000">https://stathead.com/basketball/draft_finder.cgi?request...year_min=&amp;year_max=&amp;round_min=&amp;round_max=&amp;order_by=rank&amp;sort_order=asc&amp;format=json&amp;rows=1000</a>	...	<a href="https://www.basketball-reference.com/players/w/wiggian01.html">https://www.basketball-reference.com/players/w/wiggian01.html</a>	...	<a href="https://www.basketball-reference.com/friv/draft.fcgi?college=kansas&amp;year=2014&amp;round=1&amp;order_by=rank&amp;sort_order=asc&amp;format=json&amp;rows=1000">https://www.basketball-reference.com/friv/draft.fcgi?college=kansas&amp;year=2014&amp;round=1&amp;order_by=rank&amp;sort_order=asc&amp;format=json&amp;rows=1000</a>

This is a live preview. When you are ready to run your project, click Get Data.

Show more data   Visuals enabled (advanced) 



Facet / Filter

Undo / Redo 0 / 0

**Using facets and filters**

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?

[Watch these screencasts](#)

**63 rows**

Extensions: Wikidata

Show as: **rows** records

Show: 5 10 25 50

**rows**

« first &lt; previous 1 - 10 next &gt; last »

All	Column	Rk	Pk	Tm	Player	College	Yrs	G	MP	PTS	TRB	AST	FG%	FT%
1. 0	1	1	1	CLE	Andrew Wiggins	Kansas	8	536	18929	10454	2371	1248	.446	.341
2. 1	2	2	2	MIL	Jabari Parker	Duke	8	302	8464	4351	1682	605	.495	.324
3. 2	3	3	3	PHI	Joel Embiid	Kansas	6	269	8221	6649	3022	832	.485	.331
4. 3	4	4	4	ORL	Aaron Gordon	Arizona	8	464	13232	5889	2939	1145	.451	.320
5. 4	5	5	5	UTA	Dante Exum		6	245	4545	1389	441	515	.407	.305
6. 5	6	6	6	BOS	Marcus Smart	Oklahoma State	8	459	13540	4667	1631	1943	.375	.319
7. 6	7	7	7	LAL	Julius Randle	Kentucky	8	458	14032	8007	4224	1534	.483	.344
8. 7	8	8	8	SAC	Nik Stauskas	Michigan	5	335	6662	2272	688	513	.389	.353
9. 7	8	8	8	SAC	Nik Stauskas	Michigan	5	335	6662	2272	688	513	.389	.353

1. Run the sort function on the column which contains duplicates. For instructions on sorting, see the [Sorting Data](#) section.
2. After you have sorted the column, choose "Sort" and then "reorder rows permanently."
3. Go to the column with duplicates and click on the arrow button in the column header.
4. Choose "Edit cells" and then select "Blank down."
  - a. "Blank down" will detect if two rows following each other have the same content. If they do, the second row will be "blanked out" and the cell values removed.

Source: library.illinois.edu

## Facet / Filter

Undo / Redo 0 / 0

## Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?

## Watch these screencasts

rows											« first < previous 1 - 10 next > last »				
Show as: rows records		Show: 5 10 25 50													
All	Column	Row	PK	TM	Player	College	Yrs	G	MP	PTS	TRB	AST	FG%		
1.	Facet		CLE		Andrew Wiggins	Kansas	8	536	18929	10454	2371	1248	.446	.341	
2.	Text filter		MIL		Jabari Parker	Duke	8	302	8464	4351	1682	605	.495	.324	
3.	Edit cells		PHI		Joel Embiid	Kansas	6	269	8221	6649	3022	832	.485	.331	
4.	Edit column		ORL		Aaron Gordon	Arizona	8	464	13232	5889	2939	1145	.451	.320	
5.	Transpose		UTA		Dante Exum		6	245	4545	1389	441	515	.407	.305	
6.	Sort...		BOS		Marcus Smart	Oklahoma State	8	459	13540	4667	1631	1943	.375	.319	
7.	View		LAL		Julius Randle	Kentucky	8	458	14032	8007	4224	1534	.483	.344	
8.	7	8	8	SAC	Nik Stauskas	Michigan	5	335	6662	2272	688	513	.389	.353	
9.	7	8	8	SAC	Nik Stauskas	Michigan	5	335	6662	2272	688	513	.389	.353	

1. Run the **sort** function on the column which contains duplicates. For instructions on sorting see the [Sorting Data](#) section.
  2. After you have **sorted the column**, choose "Sort" and then "reorder rows permanently."
  3. Go to the column with duplicates and click on the arrow button in the column header.
  4. Choose "Edit cells" and then select "Blank down."
    - a. "Blank down" will detect if two rows following each other have the same content. If they do, the second row will be "blanked out" and the cell values removed.

Source: library.illinois.edu

OpenRefine nba csv Permalink

Facet / Filter Undo / Redo 0 / 0

63 rows

Show as: rows records Show: 5 10 25 50 rows

Extensions: Wikidata ▾

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?  
[Watch these screencasts](#)

**Sort by Column**

Sort cell values as

text  case-sensitive  
 numbers  
 dates  
 booleans

Position blanks and errors

Valid values  
 Errors  
 Blanks

Drag and drop to re-order

smallest first  largest first

	G	MP	PTS	TRB	AST	FG%
536	18929	10454	2371	1248	.446	.341
302	8464	4351	1682	605	.495	.324
269	8221	6649	3022	832	.485	.331
464	13232	5889	2939	1145	.451	.320
245	4545	1389	441	515	.407	.305
459	13540	4667	1631	1943	.375	.319
458	14032	8007	4224	1534	.483	.344
335	6662	2272	688	513	.389	.353
335	6662	2272	688	513	.389	.353
					.459	.308

1. Run the sort function on the column which contains duplicates. For instructions on sorting, see the [Sorting Data](#) section.
2. After you have sorted the column, choose "Sort" and then "reorder rows permanently."
3. Go to the column with duplicates and click on the arrow button in the column header.
4. Choose "Edit cells" and then select "Blank down."
  - a. "Blank down" will detect if two rows following each other have the same content. If they do, the second row will be "blanked out" and the cell values removed.

Source: library.illinois.edu

Facet / Filter

Undo / Redo 0 / 0

Extensions: [Wikidata ▾](#)

### Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?

[Watch these screencasts](#)

**63 rows**

Show as: **rows** records

Show: **5 10 25 50** rows

Sort ▾

« first < previous **1 - 10** next > last »

All	Column	Rk	Pk	Tm	Player	Remove sort	Reorder rows permanently	TRB	AST	FG%	...
1.	0	1	1	CLE	Andrew Wiggins			2371	1248	.446	.341
2.	1	2	2	MIL	Jabari Parker			1682	605	.495	.324
3.	2	3	3	PHI	Joel Embiid	Kansas	6	269	8221	6649	3022
4.	3	4	4	ORL	Aaron Gordon	Arizona	8	464	13232	5889	2939
5.	4	5	5	UTA	Dante Exum		6	245	4545	1389	441
6.	5	6	6	BOS	Marcus Smart	Oklahoma State	8	459	13540	4667	1631
7.	6	7	7	LAL	Julius Randle	Kentucky	8	458	14032	8007	4224
8.	7	8	8	SAC	Nik Stauskas	Michigan	5	335	6662	2272	688
9.	7	8	8	SAC	Nik Stauskas	Michigan	5	335	6662	2272	688
											513
											.389
											.353
											.459
											.308

1. Run the sort function on the column which contains duplicates. For instructions on sorting, see the [Sorting Data](#) section
2. After you have sorted the column choose 'Sort' and then 'reorder rows permanently.'
3. Go to the column with duplicates and click on the arrow button in the column header.
4. Choose "Edit cells" and then select "Blank down."
  - a. "Blank down" will detect if two rows following each other have the same content. If they do, the second row will be "blanked out" and the cell values removed.

Source: library.illinois.edu

## 63 rows

Show as: rows records

Show: 5 10 25 50 rows

&lt;&lt; first &lt; previous 1 - 10 next &gt; last &gt;&gt;

Using facets and filters 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?

[Watch these screencasts](#)

All	Column	Rk	Pk	Tm	Player	College	Yrs	G	MP	PTS	TRB	AST	FG%	
1.	Facet			CLE	Andrew Wiggins	Kansas	8	536	18929	10454	2371	1248	.446	.341
2.	Text filter			MIL	Jabari Parker	Duke	8	302	8464	4351	1682	605	.495	.324
3.	Edit cells					Transform...								
4.	Edit column					Common transforms								
5.	Transpose					Fill down								
6.	Sort...					Blank down								
7.	View					Split multi-valued cells...								
8.	Reconcile					Join multi-valued cells...								
9.						Cluster and edit...								
						Replace								

1. Run the sort function on the column which contains duplicates. For instructions on sorting, see the [Sorting Data](#) section.
2. After you have sorted the column, choose "Sort" and then "reorder rows permanently."
3. Go to the column with duplicates and click on the arrow button in the column header.
4. Choose "Edit cells" and then select "Blank down."
  - a. "Blank down" will detect if two rows following each other have the same content. If they do, the second row will be "blanked out" and the cell values removed.

Facet / Filter Undo / Redo 2 / 2

**63 rows**

Show as: **rows** records Show: 5 10 25 50 rows

« first < previous 1 - 10 next > last »

All	Column	Rk	Pk	Tm	Player	College	Yrs	G	MP	PTS	TRB	AST	FG%	
1.	0	1	1	CLE	Andrew Wiggins	Kansas	8	536	18929	10454	2371	1248	.446	.341
2.	1	2	2	MIL	Jabari Parker	Duke	8	302	8464	4351	1682	605	.495	.324
3.	2	3	3	PHI	Joel Embiid	Kansas	6	269	8221	6649	3022	832	.485	.331
4.	3	4	4	ORL	Aaron Gordon	Arizona	8	464	13232	5889	2939	1145	.451	.320
5.	4	5	5	UTA	Dante Exum		6	245	4545	1389	441	515	.407	.305
6.	5	6	6	BOS	Marcus Smart	Oklahoma State	8	459	13540	4667	1631	1943	.375	.319
7.	6	7	7	LAL	Julius Randle	Kentucky	8	458	14032	8007	4224	1534	.483	.344
8.	7	8	8	SAC	Nik Stauskas	Michigan	5	335	6662	2272	688	513	.389	.353
9.		8	8	SAC	Nik Stauskas	Michigan	5	335	6662	2272	688	513	.389	.353

1. Run the sort function on the column which contains duplicates. For instructions on sorting, see the [Sorting Data](#) section.
2. After you have sorted the column, choose "Sort" and then "reorder rows permanently."
3. Go to the column with duplicates and click on the arrow button in the column header.
4. Choose "Edit cells" and then select "Blank down."
  - a. "Blank down" will detect if two rows following each other have the same content. If they do, the second row will be "blanked out" and the cell values removed.

Source: library.illinois.edu

Facet / Filter

Undo / Redo 2 / 2

### Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?

[Watch these screencasts](#)

5. After you have used the “Blank down” function, use the “[Facet by blank](#)” to identify rows with blank cell values for that column.
6. From the facet window, select the “true” option.
7. Go to the column labeled “All” and click on the arrow button, then select “Edit rows” and choose “Remove all matching rows.”
8. All rows with the identified duplicates will be removed. To restore the full data view, simply reset the facets

**63 rows**

Extensions: Wikidata

Show as: **rows** records

Show: 5 10 25 50 rows

< first < previous **1 - 10** next > last >

All	Column	Rk	Pk	Tm	Player	College	Yrs	G	MP	PTS	TRB	AST	FG%	...	
1.	Facet					Text facet	sas	8	536	18929	10454	2371	1248	.446	.341
2.						Numeric facet	e	8	302	8464	4351	1682	605	.495	.324
3.						Timeline facet	sas	6	269	8221	6649	3022	832	.485	.331
4.						Scatterplot facet	ona	8	464	13232	5889	2939	1145	.451	.320
5.						Custom text facet...		6	245	4545	1389	441	515	.407	.305
						Custom Numeric Facet...									
						Customized facets									
						Word facet	631	1943	.375	.319					
						Duplicates facet	224	1534	.483	.344					
						Numeric log facet	88	513	.389	.353					
						1-bounded numeric log facet	88	513	.389	.353					
						Text length facet	742	257	.459	.308					
						Log of text length facet									
						Unicode char-code facet									
						Facet by error									
						Facet by null									
						Facet by empty string									
						Facet by blank (null or empty string)									

Source: library.illinois.edu

Facet / Filter Undo / Redo 2 / 2

Refresh

Reset All Remove All

 Column change invert reset

2 choices Sort by: name count

false 62

true 1

exclude

Facet by choice counts

## 1 matching rows (63 total)

Extensions: Wikidata

Show as: rows records Show: 5 10 25 50 rows

« first &lt; previous 1 - 1 next &gt; last »

All	Column	Rk	Pk	Tm	Player	College	Yrs	G	MP	PTS	TRB	AST	FG%	3P%
9.		8	8	SAC	Nik Stauskas	Michigan	5	335	6662	2272	688	513	.389	.353

Facet / Filter Undo / Redo 2 / 2

Refresh Reset All Remove All

**62 matching rows (63 total)**

Show as: **rows** records Show: 5 10 25 50 rows « first < previous **1 - 10** next > last »

All  Column  Rk  Pk  Tm  Player  College  Yrs  G  MP  PTS  TRB  AST  FG%  :

	Rk	Pk	Tm	Player	College	Yrs	G	MP	PTS	TRB	AST	FG%
1.	0	1	1	CLE	Andrew Wiggins	Kansas	8	536	18929	10454	2371	.446 .341
2.	1	2	2	MIL	Jabari Parker	Duke	8	302	8464	4351	1682	.495 .324
3.	2	3	3	PHI	Joel Embiid	Kansas	6	269	8221	6649	3022	.485 .331
4.	3	4	4	ORL	Aaron Gordon	Arizona	8	464	13232	5889	2939	.451 .320
5.	4	5	5	UTA	Dante Exum		6	245	4545	1389	441	.407 .305
6.	5	6	6	BOS	Marcus Smart	Oklahoma State	8	459	13540	4667	1631	.375 .319
7.	6	7	7	LAL	Julius Randle	Kentucky	8	458	14032	8007	4224	.483 .344
8.	7	8	8	SAC	Nik Stauskas	Michigan	5	335	6662	2272	688	.389 .353
10.	8	9	9	CHH	Noah Vonleh	Indiana	7	339	5683	1660	1742	.459 .308
11.	9	10	10	PHI	Elfrid Payton	Louisiana	8	455	12909	4920	1936	.450 .288

Source: library.illinois.edu

**OpenRefine nba dataset csv Permalink**

Facet / Filter Undo / Redo 0 / 5 Refresh Reset All Remove All

62 matching rows (63 total)

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

Extensions: Wikidata ▾

Rk Column Player College Yrs G MP PTS TRB AST FG% 3P% FT% MP2 PTS2 TRB2 AST2 WS WS/48 BPM

1. 0 1 Facet Text facet Numeric facet

2. 1 2 Edit cells Timeline facet

3. 2 3 Edit column Scatterplot facet

4. 3 4 Transpose Custom text facet...

5. 4 5 Sort... Customized facets

6. 5 6 View Word facet

7. 6 7 Reconcile Duplicates facet

8. 7 8 auskas Michigan 5 Numeric log facet

9. 7 8 10 10 CHI Noah Vonleh Indiana 5 1-bounded numeric log facet

10. 8 9 11 11 PHI Elfrid Payton Louisiana 5 Text length facet

11. 9 10 12 12 DEN Doug McDermott Creighton 5 Log of text length facet

12. 10 11 13 13 ORL Dario Šarić UCLA 5 Unicode char-code facet

13. 11 12 14 14 MIN Zach LaVine T.J. Warren NC State 5 Facet by error

14. 12 13 15 15 PHO Adreian Payne Michigan State 4 Facet by null

15. 13 14 16 16 ATL James Young Kentucky 4 Facet by empty string

16. 14 15 17 17 CHI Jusuf Nurkić Syracuse 4 Facet by blank (null or empty string)

17. 15 16 18 18 BOS Tyler Ennis Michigan State 5

18. 16 17 19 19 CHI Gary Harris Michigan State 5

19. 17 18 20 20 TOR Bruno Caboclo Michigan State 5

20. 18 19 21 21 OKC Mitch McGary Michigan State 5

21. 19 20 22 22 MEM Jordan Adams UCLA Michigan State 5

22. 20 21 23 23 UTA Rodney Hood Duke Michigan State 5

23. 21 22 24 24 CHH Shabazz Napier UConn Michigan State 5

24. 22 23 25 25 HOU Clint Capela Michigan State 5

25. 23 24 26 26 MIA P.J. Hairston UNC Michigan State 5

26. 24 25 27 27 PHO Bogdan Bogdanović Michigan State 5

27. 25 26 28 28 LAC C.J. Wilcox Washington Michigan State 5

28. 26 27 29 29 OKC Josh Huestis Stanford Michigan State 5

29. 27 28 30 30 SAS Kyle Anderson UCLA Michigan State 5

30. 28 29 31 31 Rk Damien Inglis Michigan State 5

31. 29 30 32 32 MIL Damien Inglis Michigan State 5

1. To remove blank rows, just use facet by blank.

# **Most Basic Data Cleaning and Wrangling Tasks**

- 1. Remove blanks**
- 2. Remove duplicates**
- 3. Resolve inconsistent entries**
- 4. Transform incorrect data (e.g. wrong data types such as text instead of numeric and calculations)**
- 5. Rename, combine and split columns for making further data presentation and processing easier**
- 6. Impute missing entries (need expertise to judge usage instead of relying on statistics alone)**

# **Analysing Cleaned Data Using SQL**

**OpenRefine** nba run results csv [Permalink](#)

Facet / Filter Undo / Redo 0 / 0 [Extensions: Wikidata](#)

### 60 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows [first](#) [previous](#) [1 of 1 page](#) [next](#) [last](#)

**Using facets and filters** 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? [Watch these screencasts](#)

	Col_Rank	Col_Player	Col_College	Col_Yrs	Col_Game	Col_Minutes_Played	Col_Points	Col_Total_Rebounds	Col_Assists	Col_Field_Goal_Percentage	Col_Three_Pt_Percentage	Col_L
1.	1	Andrew Wiggins	Kansas	8	598	20935	11519	2649	1393	0.448	0.350	0.723
2.	2	Jabari Parker	Duke	8	310	8535	4380	1703	610	0.494	0.326	0.743
3.	3	Joel Embiid	Kansas	6	328	10240	8535	3732	1080	0.490	0.338	0.810
4.	4	Aaron Gordon	Arizona	8	528	15277	6887	3310	1313	0.460	0.323	0.702
5.	5	Dante Exum	[REDACTED]		245	4545	1389	441	515	0.407	0.305	0.764
6.	6	Marcus Smart	Oklahoma State	8	520	15484	5438	1867	2318	0.382	0.321	0.780
7.	7	Julius Randle	Kentucky	8	518	16145	9191	4814	1842	0.472	0.332	0.743
8.	8	Nik Stauskas	Michigan	6	343	6701	2290	691	515	0.389	0.354	0.812
9.	9	Noah Vonleh	Indiana	7	339	5683	1660	1742	257	0.459	0.308	0.691
10.	10	Elfrid Payton	Louisiana	8	500	13383	5036	2007	2868	0.447	0.287	0.623
11.	11	Doug McDermott	Creighton	8	527	11012	4846	1236	496	0.476	0.409	0.820
12.	12	Dario Šarić	[REDACTED]		356	8965	4180	2086	699	0.441	0.357	0.838
13.	13	Zach LaVine	UCLA	8	478	15219	9466	1876	1860	0.461	0.386	0.830
14.	14	T.J. Warren	NC State	7	332	9553	5142	1372	397	0.507	0.357	0.780
15.	15	Adreian Payne	Michigan State	4	107	1403	429	315	66	0.406	0.254	0.680
16.	16	Jusuf Nurkić	[REDACTED]		411	9815	5015	3513	903	0.500	0.232	0.672
17.	17	James Young	Kentucky	4	95	812	219	96	28	0.367	0.277	0.563
18.	18	Tyler Ennis	Syracuse	4	186	2336	779	250	359	0.419	0.317	0.768
19.	19	Gary Harris	Michigan State	8	468	13473	5526	1162	965	0.448	0.363	0.812
20.	20	Bruno Caboclo	[REDACTED]		105	1293	442	268	72	0.403	0.308	0.836
21.	21	Mitch McGary	Michigan	2	52	557	227	183	17	0.527	0.000	0.580
22.	22	Jordan Adams	UCLA	2	32	263	101	30	19	0.402	0.385	0.607
23.	23	Rodney Hood	Duke	8	448	10977	4656	1160	736	0.420	0.366	0.841
24.	24	Shabazz Napier	UConn	6	345	5986	2433	657	849	0.397	0.345	0.815
25.	25	Clint Capela	[REDACTED]		471	12614	5851	5023	465	0.624	0.000	0.527
26.	26	P.J. Hairston	UNC	2	111	2000	664	266	59	0.343	0.295	0.810
27.	27	Bogdan Bogdanović	[REDACTED]		316	9041	4501	1083	1073	0.439	0.384	0.822
28.	28	C.J. Wilcox	Washington	3	66	376	132	31	30	0.370	0.333	0.813
29.	29	Josh Huestis	Stanford	3	76	1068	187	180	23	0.346	0.312	0.240
30.	30	Kyle Anderson	UCLA	8	505	10583	3366	2220	1167	0.476	0.334	0.711
31.	31	Damien Ingles	[REDACTED]		20	156	36	31	10	0.351	0.231	0.875
32.	32	K.J. McDaniels	Clemson	3	148	2092	782	324	95	0.412	0.290	0.776
33.	33	Joe Harris	Virginia	8	414	10601	4584	1344	679	0.483	0.439	0.780
34.	34	Cleanthony Early	Wichita State	2	56	801	241	123	42	0.346	0.263	0.750
35.	35	Jarnell Stokes	Tennessee	3	28	151	67	40	7	0.581		0.531

[OpenRefine](#) nba run results csv Permalink

Facet / Filter Undo / Redo 0 / 0 < 60 rows Show as: rows records Show: 5 10 25 50 100 500 1000 rows < first < previous 1 of 1 page next > last >

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? Watch these screencasts

Extensions: Wikidata ▾

	All	Col_Rank	Col_Player	Col_College	Col_Yrs	Col_Game	Col_Minutes_Played	Col_Points	Col_Total_Rebounds	Col_Assists	Col_Field_Goal_Percentage	Col_Three_Pt_Percentage	Col_F
1.	1	Andrew Wiggins	Facet	Text facet				11519	2649	1393	0.448	0.350	0.723
2.	2	Jabari Parker	Text filter	Numeric facet				4380	1703	610	0.494	0.326	0.743
3.	3	Joel Embiid	Edit cells	Timeline facet				8535	3732	1080	0.490	0.338	0.810
4.	4	Aaron Gordon	Edit column	Scatterplot facet				6887	3310	1313	0.460	0.323	0.702
5.	5	Dante Exum	Transpose	Custom text facet...				1389	441	515	0.407	0.305	0.764
6.	6	Marcus Smart	Sort...	Custom Numeric Facet...				5438	1867	2318	0.382	0.321	0.780
7.	7	Julius Randle	View	Customized facets				9191	4814	1842	0.472	0.332	0.743
8.	8	Nik Stauskas	Reconcile	Word facet				515	389			0.354	0.812
9.	9	Noah Vonleh		Duplicates facet				257	459			0.308	0.691
10.	10	Elfrid Payton		Numeric log facet				2868	447			0.287	0.623
11.	11	Doug McDermott	Creighton	1-bounded numeric log facet				496	476			0.409	0.820
12.	12	Dario Šarić		Text length facet				699	441			0.357	0.838
13.	13	Zach LaVine	UCLA	Log of text length facet				1860	461			0.386	0.830
14.	14	T.J. Warren	NC State	Unicode char-code facet				397	507			0.357	0.780
15.	15	Adreian Payne	Michigan State	Facet by error				66	406			0.254	0.680
16.	16	Jusuf Nurkić		Facet by null				903	500			0.232	0.672
17.	17	James Young	Kentucky	Facet by empty string				28	367			0.277	0.563
18.	18	Tyler Ennis	Syracuse	Facet by blank (null or empty string)				359	419			0.317	0.768
19.	19	Gary Harris	Michigan State					965	448			0.363	0.812
20.	20	Bruno Caboclo						72	403			0.308	0.836
21.	21	Mitch McGary	Michigan					17	527			0.000	0.580
22.	22	Jordan Adams	UCLA					19	402			0.385	0.607
23.	23	Rodney Hood	Duke					736	420			0.366	0.841
24.	24	Shabazz Napier	UConn					849	397			0.345	0.815
25.	25	Clint Capela						4656	1160			0.000	0.527
26.	26	P.J. Hairston	UNC					2433	657			0.295	0.810
27.	27	Bogdan Bogdanović						4501	1083			0.384	0.822
28.	28	C.J. Wilcox	Washington					5851	5023			0.624	0.813
29.	29	Josh Huestis	Stanford					132	31			0.333	0.240
30.	30	Kyle Anderson	UCLA					187	180			0.312	0.711
31.	31	Damien Inglis						3366	2220			0.476	0.711
32.	32	K.J. McDaniels	Clemson					1167	1083			0.334	0.875
33.	33	Joe Harris	Virginia					782	324			0.412	0.776
34.	34	Cleanthony Early	Wichita State					4584	1344			0.483	0.780
35.	35	Jarnell Stokes	Tennessee					241	123			0.346	0.750
								67	40			0.581	0.531

javascript:{}

**OpenRefine** nba run results csv Permalink

Facet / Filter Undo / Redo 0 / 0 Refresh Reset All Remove All

Extensions: Wikidata ▾ Open... Export ▾ Help

**14 matching rows** (60 total)

Show as: rows records Show: 5 10 25 50 100 500 1000 rows « first < previous 1 of 1 page next » last »

	Col_Rank	Col_Player	Col_College	Col_Yrs	Col_Game	Col_Minutes_Played	Col_Points	Col_Total_Rebounds	Col_Assists	Col_Field_Goal_Percentage	Col_Three_Pt_Percentage	Col_F
5. 5	Dante Exum			6	245	4545	1389	441	515	0.407	0.305	0.764
12. 12	Dario Šarić			5	356	8965	4180	2086	699	0.441	0.357	0.838
16. 16	Jusuf Nurkić			8	411	9815	5015	3513	903	0.500	0.232	0.672
20. 20	Bruno Caboclo			7	105	1293	442	268	72	0.403	0.308	0.836
25. 25	Clint Capela			8	471	12614	5851	5023	465	0.624	0.000	0.527
27. 27	Bogdan Bogdanović			5	316	9041	4501	1083	1073	0.439	0.384	0.822
31. 31	Damien Inglis			1	20	156	36	31	10	0.351	0.231	0.875
41. 41	Nikola Jokić			7	527	16018	10364	5456	3281	0.542	0.345	0.830
43. 43	Edy Tavares			2	13	101	33	32	4	0.625		0.273
51. 51	Thanasis Antetokounmpo			4	127	1159	398	250	83	0.517		0.545
52. 52	Vasilije Micić											
53. 53	Alessandro Gentile											
54. 54	Nemanja Dangubić											
57. 57	Louis Labeyrie											

javascript:{}

**OpenRefine nba run results csv** [Permalink](#)

Facet / Filter Undo / Redo 0 / 0 < Extensions: Wikidata ▾

Refresh Reset All Remove All

**46 matching rows (60 total)**

Show as: rows records Show: 5 10 25 50 100 500 1000 rows « first < previous 1 of 1 page next » last »

**Col\_College** change invert reset

2 choices Sort by: name count

false 46 exclude

true 14

Facet by choice counts

All Col\_Rank Col\_Player Col\_College Col\_Yrs Col\_Game Col\_Minutes\_Played Col\_Points Col\_Total\_Rebounds Col\_Assists Col\_Field\_Goal\_Percentage Col\_Three\_Pt\_Percentage Col\_F

			Kansas	8	598	20935	11519	2649	1393	0.448	0.350	0.723
			Duke	8	310	8535	4380	1703	610	0.494	0.326	0.743
			Kansas	6	328	10240	8535	3732	1080	0.490	0.338	0.810
			Arizona	8	528	15277	6887	3310	1313	0.460	0.323	0.702
			Oklahoma State	8	520	15484	5438	1867	2318	0.382	0.321	0.780
			Kentucky	8	518	16145	9191	4814	1842	0.472	0.332	0.743
			Michigan	6	343	6701	2290	691	515	0.389	0.354	0.812
			Indiana	7	339	5683	1660	1742	257	0.459	0.308	0.691
			Louisiana	8	500	13383	5036	2007	2868	0.447	0.287	0.623
			Creighton	8	527	11012	4846	1236	496	0.476	0.409	0.820
			UCLA	8	478	15219	9466	1876	1860	0.461	0.386	0.830
			NC State	7	332	9553	5142	1372	397	0.507	0.357	0.780
			Michigan State	4	107	1403	429	315	66	0.406	0.254	0.680
			Kentucky	4	95	812	219	96	28	0.367	0.277	0.563
			Syracuse	4	186	2336	779	250	359	0.419	0.317	0.768
			Michigan State	8	468	13473	5526	1162	965	0.448	0.363	0.812
			Michigan	2	52	557	227	183	17	0.527	0.000	0.580
			UCLA	2	32	263	101	30	19	0.402	0.385	0.607
			Duke	8	448	10977	4656	1160	736	0.420	0.366	0.841
			UConn	6	345	5986	2433	657	849	0.397	0.345	0.815
			UNC	2	111	2000	664	266	59	0.343	0.295	0.810
			Washington	3	66	376	132	31	30	0.370	0.333	0.813
			Stanford	3	76	1068	187	180	23	0.346	0.312	0.240
			UCLA	8	505	10583	3366	2220	1167	0.476	0.334	0.711
			Clemson	3	148	2092	782	324	95	0.412	0.290	0.776
			Virginia	8	414	10601	4584	1344	679	0.483	0.439	0.780
			Wichita State	2	56	801	241	123	42	0.346	0.263	0.750
			Tennessee	3	28	151	67	40	7	0.581		0.531
			LSU	4	147	1684	509	352	69	0.402	0.360	0.663
			UConn									
			Colorado	8	387	10206	5042	1167	1933	0.411	0.322	0.791
			Syracuse	8	555	14454	6329	2189	747	0.452	0.349	0.723
			Michigan	7	304	5299	1804	776	238	0.457	0.373	0.779
			Nick Johnson	1	28	262	74	39	11	0.347	0.238	0.680

Filter Undo / Redo 0 / 0 Reset All Remove All

**46 matching rows (60 total)**

Show as: rows records Show: 5 10 25 50 100 500 1000 rows « first

_College	Col_Rank	Col_Player	Col_College	Col_Yrs	Col_Game	Col_Minutes_Played	Col_Points	Col_Total_Rebounds	Col_Assists	Col_Field_Go	Col_F
	1. 1	Andrew Wiggins	Kansas	8	598	20935	11519	2649	1393	0.448	
	2. 2	Jabari Parker	Duke	8	310	8535	4380	1703	610	0.494	
	3. 3	Joel Embiid	Kansas	6	328	10240	8535	3732	1080	0.490	
	4. 4	Aaron Gordon	Arizona	8	528	15277	6887	3310	1313	0.460	
	6. 6	Marcus Smart	Oklahoma State	8	520	15484	5438	1867	2318	0.382	
	7. 7	Julius Randle	Kentucky	8	518	16145	9191	4814	1842	0.472	
	8. 8	Nik Stauskas	Michigan	6	343	6701	2290	691	515	0.389	
	9. 9	Noah Vonleh	Indiana	7	339	5683	1680	1742	257	0.459	
	10. 10	Elfrid Payton	Louisiana	8	500	13383	5036	2007	2868	0.447	
	11. 11	Doug McDermott	Creighton	8	527	11012	4846	1236	496	0.476	
	13. 13	Zach LaVine	UCLA	8	478	15219	9466	1876	1860	0.461	
	14. 14	T.J. Warren	NC State	7	332	9553	5142	1372	397	0.507	
	15. 15	Adreian Payne	Michigan State	4	107	1403	429	315	66	0.406	
	17. 17	James Young	Kentucky	4	95	812	219	96	28	0.367	
	18. 18	Tyler Ennis	Syracuse	4	186	2336	779	250	359	0.419	0.317 0.768
	19. 19	Gary Harris	Michigan State	8	468	13473	5526	1162	965	0.448	0.363 0.812
	21. 21	Mitch McGary	Michigan	2	52	557	227	183	17	0.527	0.000 0.580
	22. 22	Jordan Adams	UCLA	2	32	263	101	30	19	0.402	0.385 0.607
	23. 23	Rodney Hood	Duke	8	448	10977	4656	1160	736	0.420	0.366 0.841
	24. 24	Shabazz Napier	UConn	6	345	5986	2433	657	849	0.397	0.345 0.815
	26. 26	P.J. Hairston	UNC	2	111	2000	664	266	59	0.343	0.295 0.810
	28. 28	C.J. Wilcox	Washington	3	66	376	132	31	30	0.370	0.333 0.813
	29. 29	Josh Huestis	Stanford	3	76	1068	187	180	23	0.346	0.312 0.240
	30. 30	Kyle Anderson	UCLA	8	505	10583	3366	2220	1167	0.476	0.334 0.711
	32. 32	K.J. McDaniels	Clemson	3	148	2092	782	324	95	0.412	0.290 0.776
	33. 33	Joe Harris	Virginia	8	414	10601	4584	1344	679	0.483	0.439 0.780
	34. 34	Cleanthony Early	Wichita State	2	56	801	241	123	42	0.346	0.263 0.750
	35. 35	Jarnell Stokes	Tennessee	3	28	151	67	40	7	0.581	
	36. 36	Johnny O'Bryant	LSU	4	147	1684	509	352	69	0.402	0.360 0.663
	37. 37	DeAndre Daniels	UConn								
	38. 38	Spencer Dinwiddie	Colorado	8	387	10206	5042	1167	1933	0.411	0.322 0.791
	39. 39	Jerami Grant	Syracuse	8	555	14454	6329	2189	747	0.452	0.349 0.723
	40. 40	Glenn Robinson III	Michigan	7	304	5299	1804	776	238	0.457	0.373 0.779
	42. 42	Nickeil Alexander-Walker	Arizona	1	26	260	74	26	11	0.247	0.226 0.600

```
In [3]: 1 import pandas as pd
2 import sqlite3
3 con = sqlite3.connect('nba_db.db')
4 sql = "SELECT * FROM nba run results WHERE Col Game > 400"
5 df = pd.read_sql_query(sql,con)
6 con.close()
7 df
```

Out[3]:

Col_Rank	Col_Player	Col_College	Col_Yrs	Col_Game	Col_Minutes_Played	Col_Points	Col_Total_Rebounds	Col_Assists	Col_Field_Goal_Percentage	Col_T
0	1	Andrew Wiggins	Kansas	8	598	20935	11519	2649	1393	0.448
1	4	Aaron Gordon	Arizona	8	528	15277	6887	3310	1313	0.460
2	6	Marcus Smart	Oklahoma State	8	520	15484	5438	1867	2318	0.382
3	7	Julius Randle	Kentucky	8	518	16145	9191	4814	1842	0.472
4	10	Elfrid Payton	Louisiana	8	500	13383	5036	2007	2868	0.447
5	11	Doug McDermott	Creighton	8	527	11012	4846	1236	496	0.476
6	13	Zach LaVine	UCLA	8	478	15219	9466	1876	1860	0.461
7	19	Gary Harris	Michigan State	8	468	13473	5526	1162	965	0.448
8	23	Rodney Hood	Duke	8	448	10977	4656	1160	736	0.420
9	30	Kyle Anderson	UCLA	8	505	10583	3366	2220	1167	0.476
10	33	Joe Harris	Virginia	8	414	10601	4584	1344	679	0.483
11	39	Jerami Grant	Syracuse	8	555	14454	6329	2189	747	0.452
12	45	Dwight Powell	Stanford	8	511	9720	3926	2349	524	0.585
13	46	Jordan Clarkson	Missouri	8	600	16191	9216	1971	1526	0.439

```
1 import pandas as pd
2 import sqlite3
3 con = sqlite3.connect('nba_db.db')
4 sql = SELECT Col_College, sum(Col_Game) AS [Total # of Games] FROM nba_run_results GROUP BY Col_College ORDER BY |
5 df = pd.read_sql_query(sql,con)
6 con.close()
7 df
```

	Col_College	Total # of Games
0	UCLA	1015.0
1	Kansas	926.0
2	Duke	758.0
3	Syracuse	741.0
4	Michigan	699.0
5	Oklahoma State	633.0
6	Kentucky	613.0
7	Missouri	600.0
8	Stanford	587.0
9	Michigan State	575.0
10	Arizona	556.0
11	Creighton	527.0
12	Louisiana	500.0
13	Virginia	414.0
14	Colorado	387.0
15	UConn	345.0
16	Indiana	339.0
17	NC State	332.0
18	Tennessee	151.0
19	Clemson	148.0
20	LSU	147.0

```

1 # CAST(ROUND(Column_Name, 2) AS DECIMAL(10,2)
2 import pandas as pd
3 import sqlite3
4 con = sqlite3.connect('nba_db.db')
5 # sql = "SELECT Col_Player, Col_College, Col_Game, Col_Yrs, ROUND((Col_Game*1.0/Col_Yrs),2) AS [Games/Years], ROUND(CAST(Col_Game AS FLOAT)/Col_Yrs,2) AS [Avg_Points]"
6 sql = "SELECT Col_Player, Col_College, Col_Game, Col_Yrs, ROUND(CAST(Col_Game AS FLOAT)/Col_Yrs,2) AS [Games/Years]"
7 df = pd.read_sql_query(sql,con)
8 con.close()
9 df

```

	Col_Player	Col_College	Col_Game	Col_Yrs	Games/Years	Avg Points
0	Zach LaVine	UCLA	478.0	8.0	59.75	158.43
1	Joel Embiid	Kansas	328.0	6.0	54.67	156.13
2	Andrew Wiggins	Kansas	598.0	8.0	74.75	154.10
3	Julius Randle	Kentucky	518.0	8.0	64.75	141.95
4	Jordan Clarkson	Missouri	600.0	8.0	75.00	122.88
5	Jabari Parker	Duke	310.0	8.0	38.75	113.03
6	T.J. Warren	NC State	332.0	7.0	47.43	108.42
7	Aaron Gordon	Arizona	528.0	8.0	66.00	104.35
8	Spencer Dinwiddie	Colorado	387.0	8.0	48.38	104.23
9	Gary Harris	Michigan State	468.0	8.0	58.50	94.46
10	Jerami Grant	Syracuse	555.0	8.0	69.38	91.23
11	Joe Harris	Virginia	414.0	8.0	51.75	88.58
12	Marcus Smart	Oklahoma State	520.0	8.0	65.00	83.66
13	Rodney Hood	Duke	448.0	8.0	56.00	83.14
14	Elfrid Payton	Louisiana	500.0	8.0	62.50	80.58
15	Doug McDermott	Creighton	527.0	8.0	65.88	73.56
16	Dwight Powell	Stanford	511.0	8.0	63.88	61.46
17	Kyle Anderson	UCLA	505.0	8.0	63.13	53.32
18	Shabazz Napier	UConn	345.0	6.0	57.50	42.31
19	Glenn Robinson III	Michigan	304.0	7.0	43.43	41.54
20	Nik Stauskas	Michigan	343.0	6.0	57.17	40.06

**Thank you for your time!**