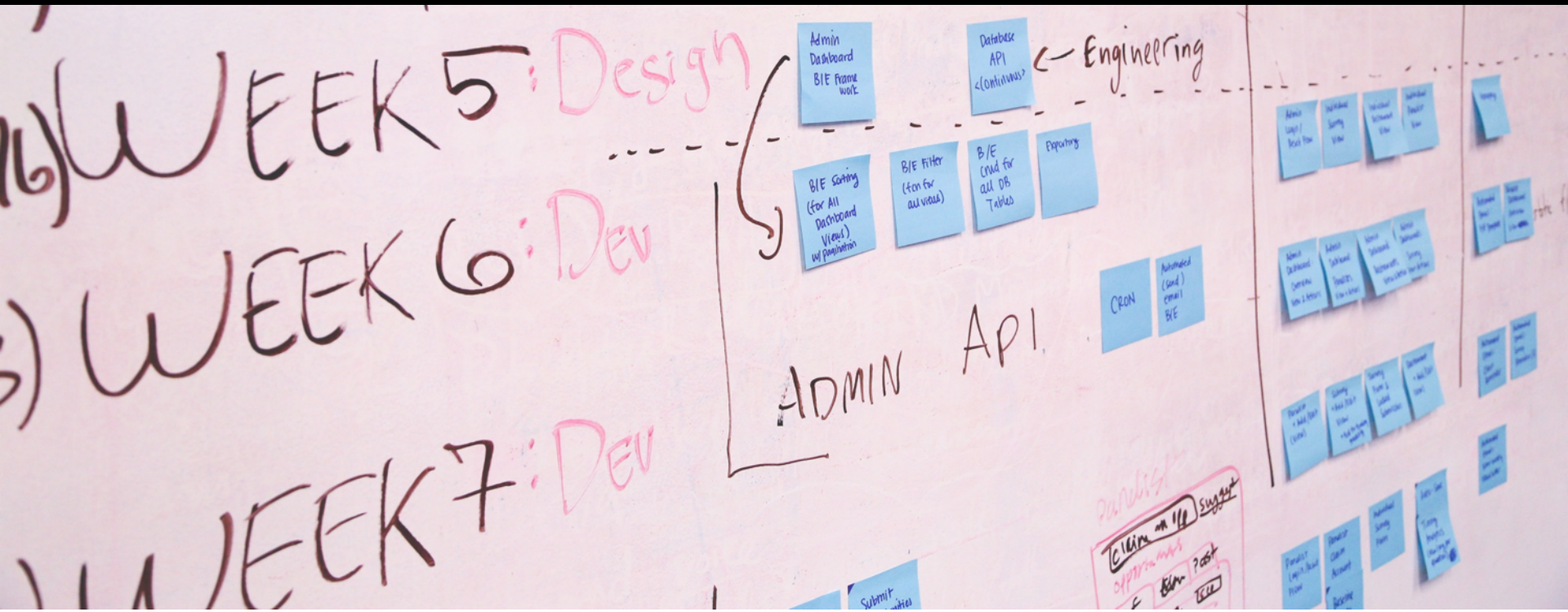
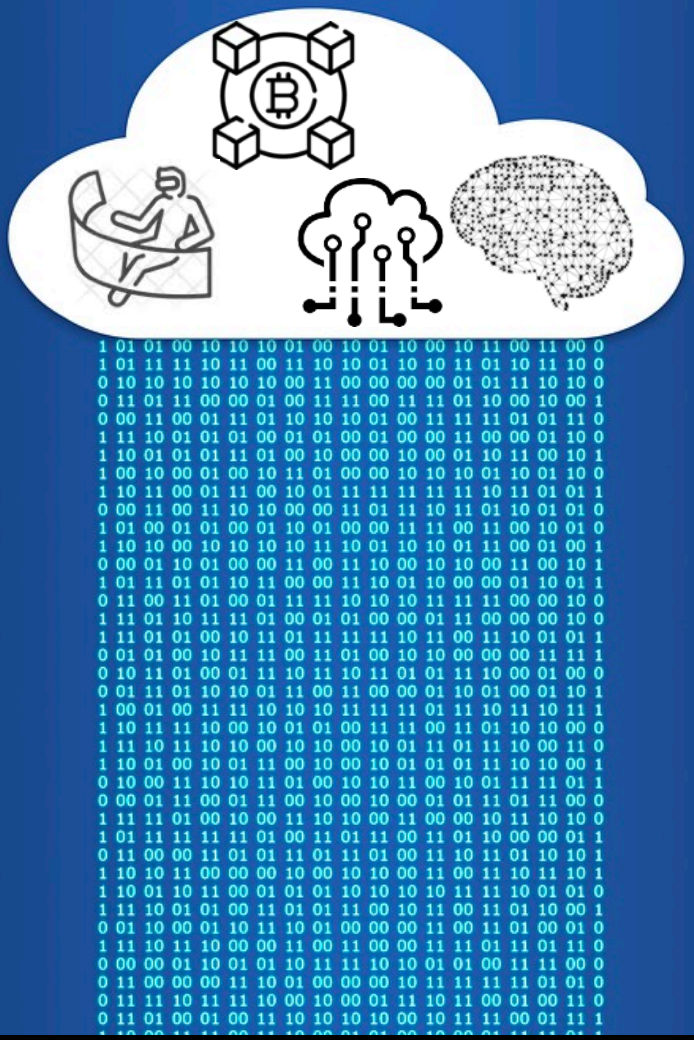


DATA HACK 2023 MAKER WORKSHOP:
DATA SCRAPING AND DATA CLEANING

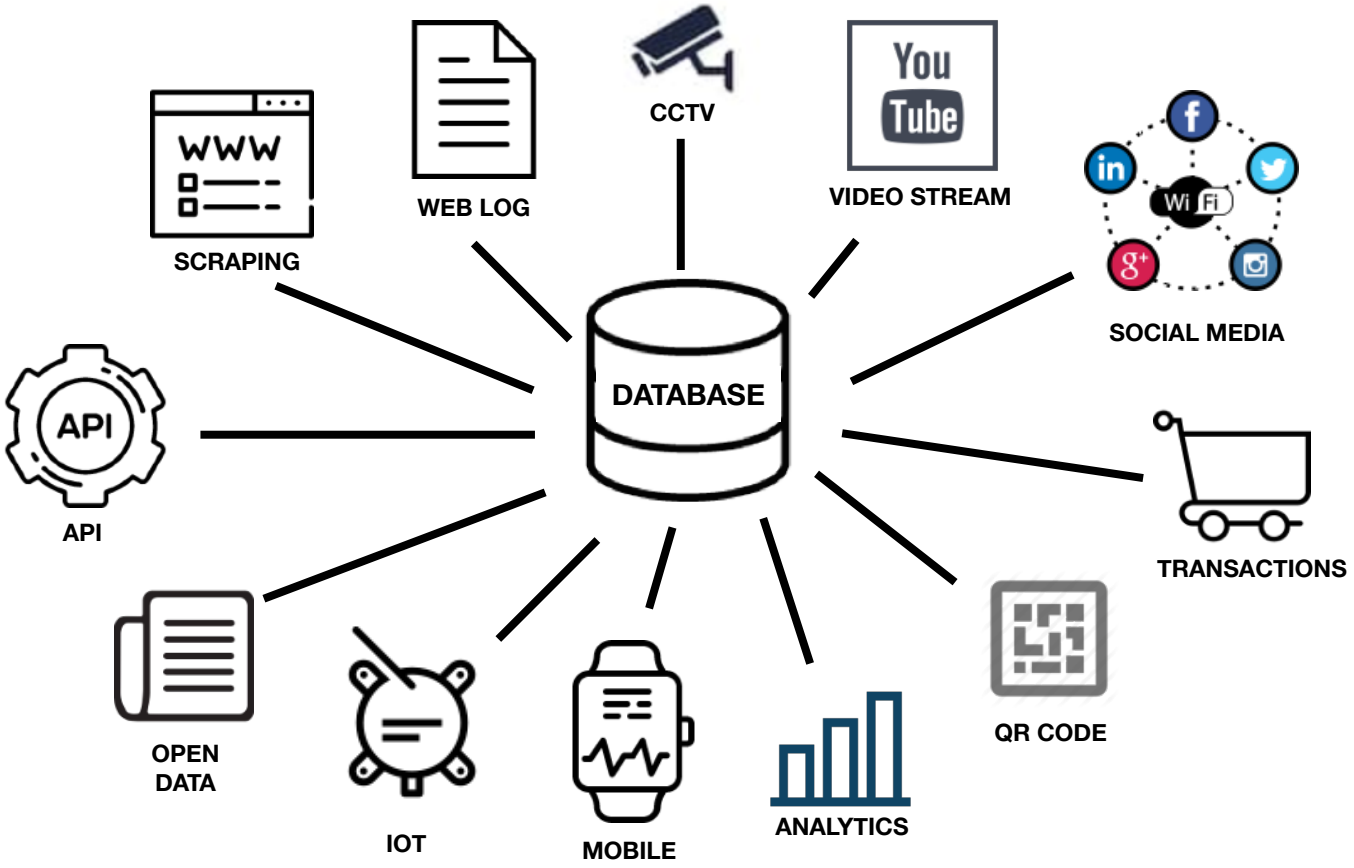
Bernard Suen
Center for Entrepreneurship
Chinese University of Hong Kong

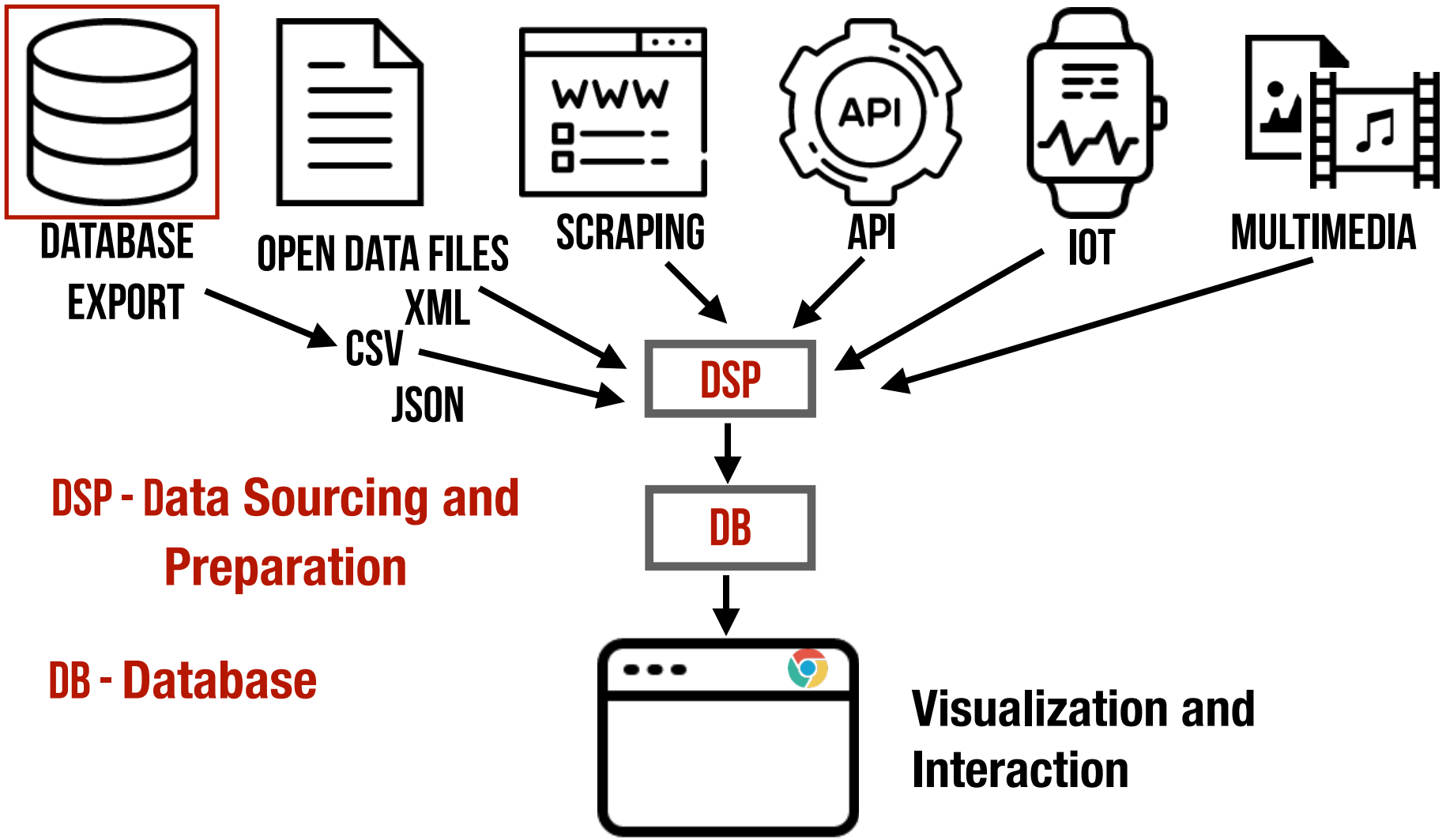
Today's agenda.

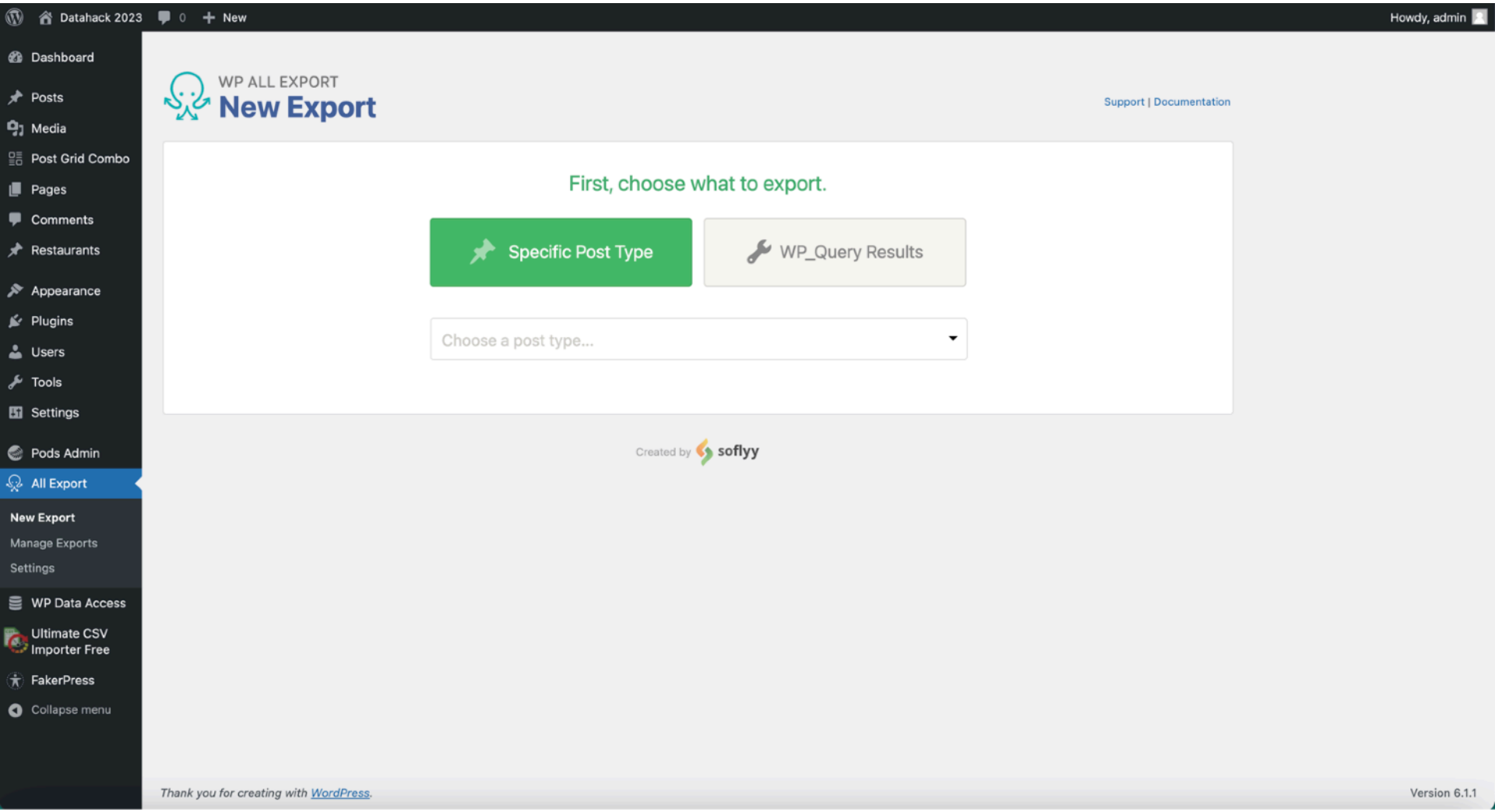
1. Most common sources of data
2. Data scraping with ParseHub
3. Data cleaning with OpenRefine

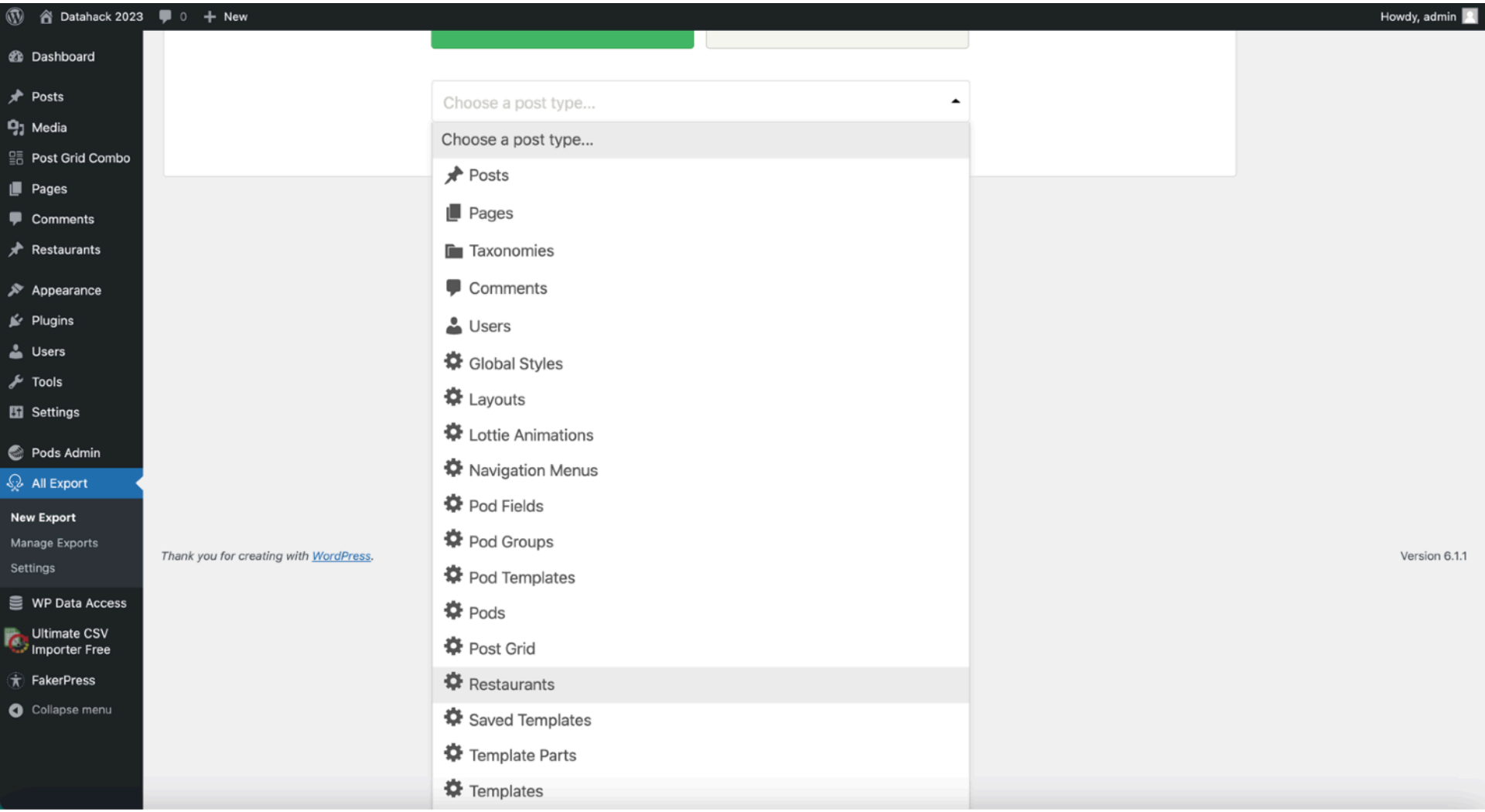


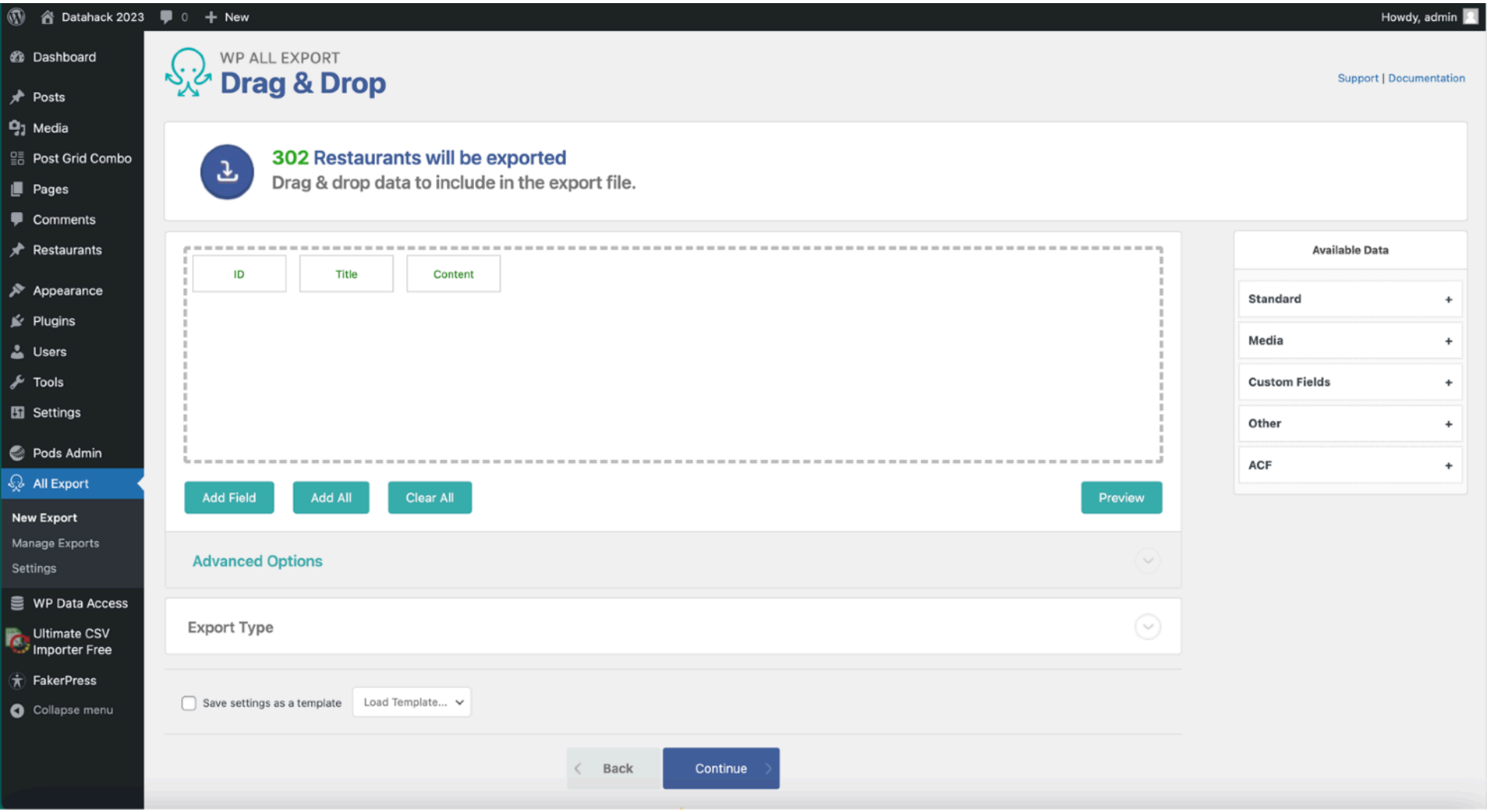
Structured and Un-structured Data Sources

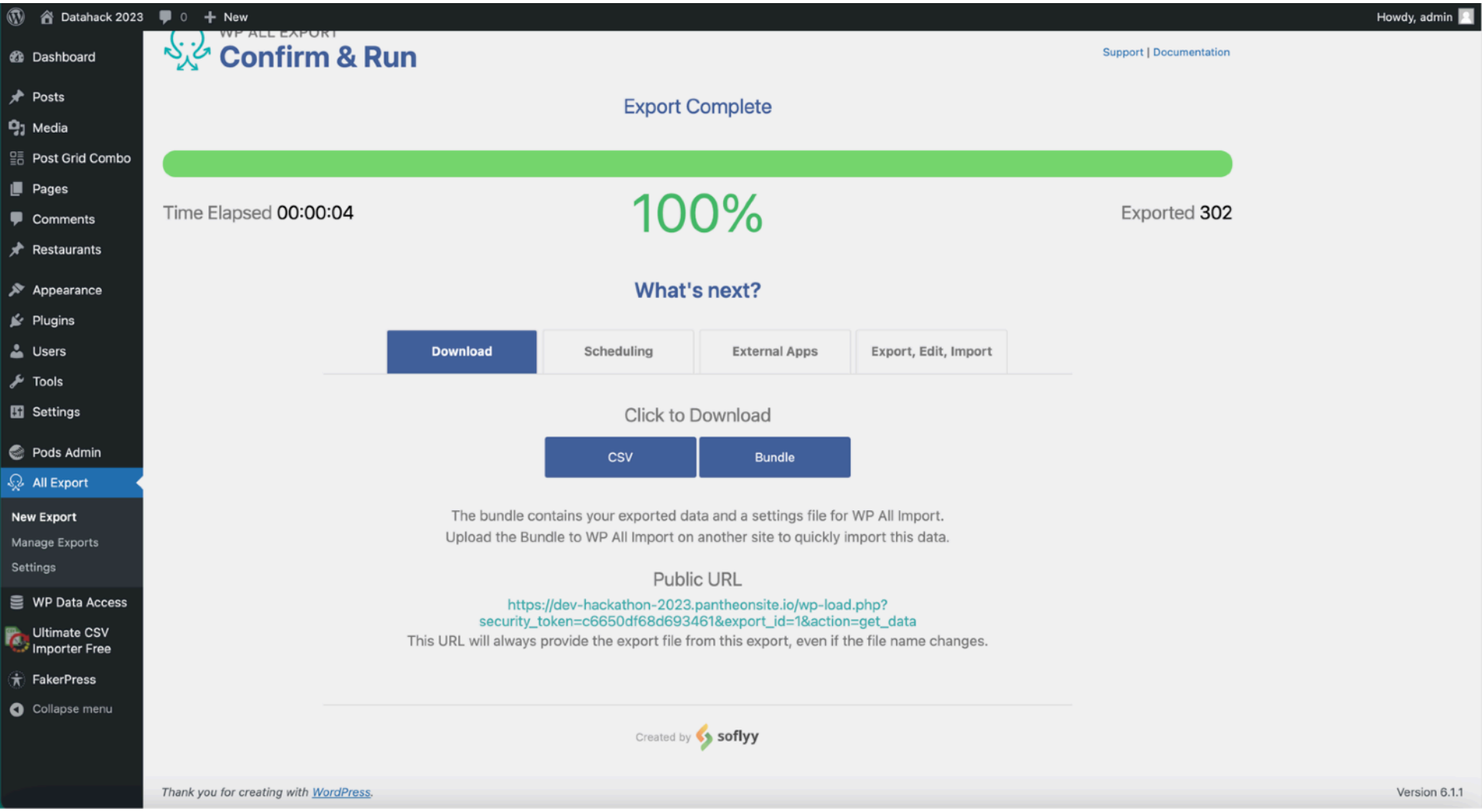


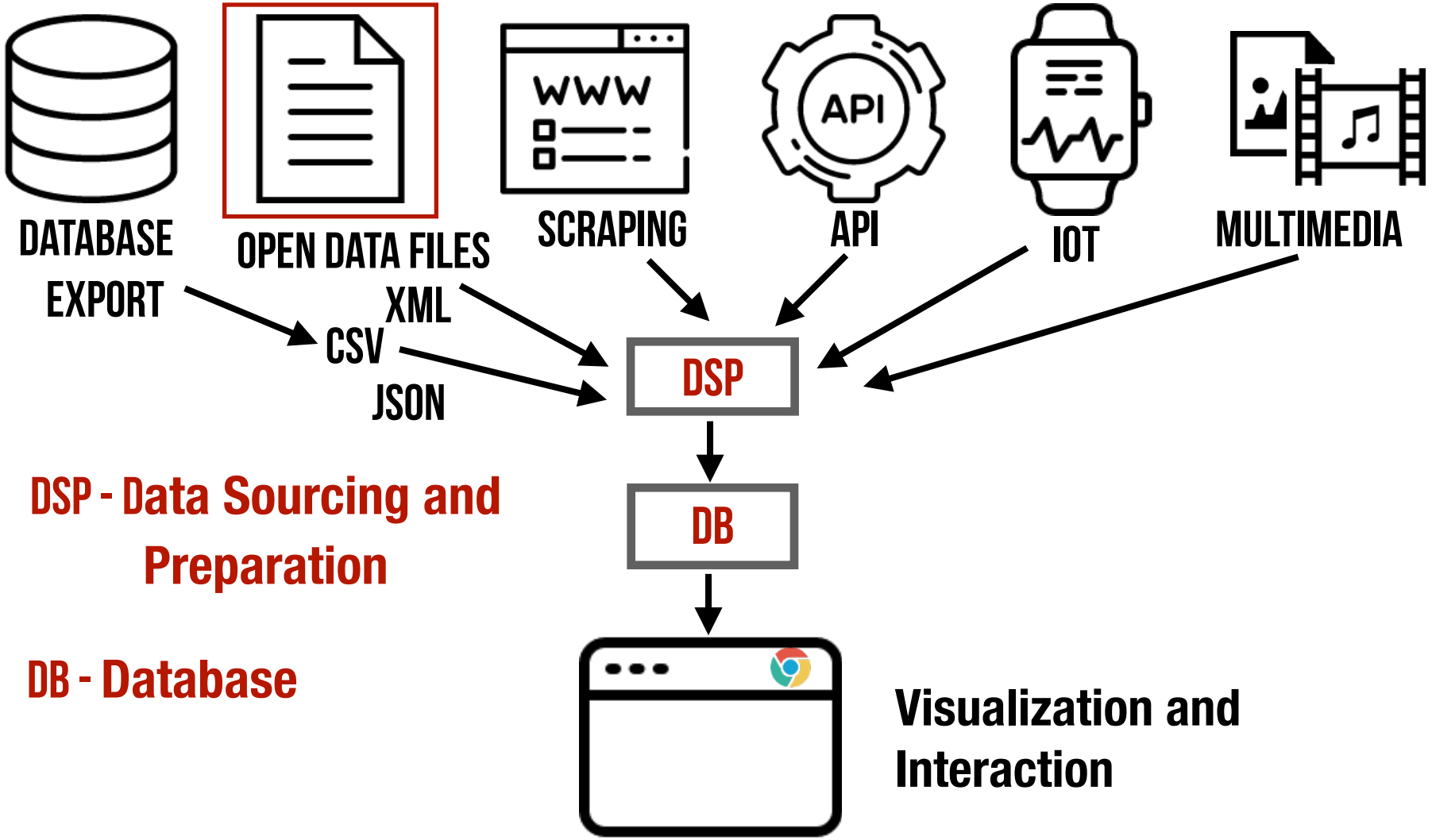

DATABASE
EXPORT
OPEN DATA FILES
XML
CSV
JSON
SCRAPING
API
IOT
MULTIMEDIA
DSP
DB
DSP - Data Sourcing and Preparation
DB - Database
Visualization and Interaction

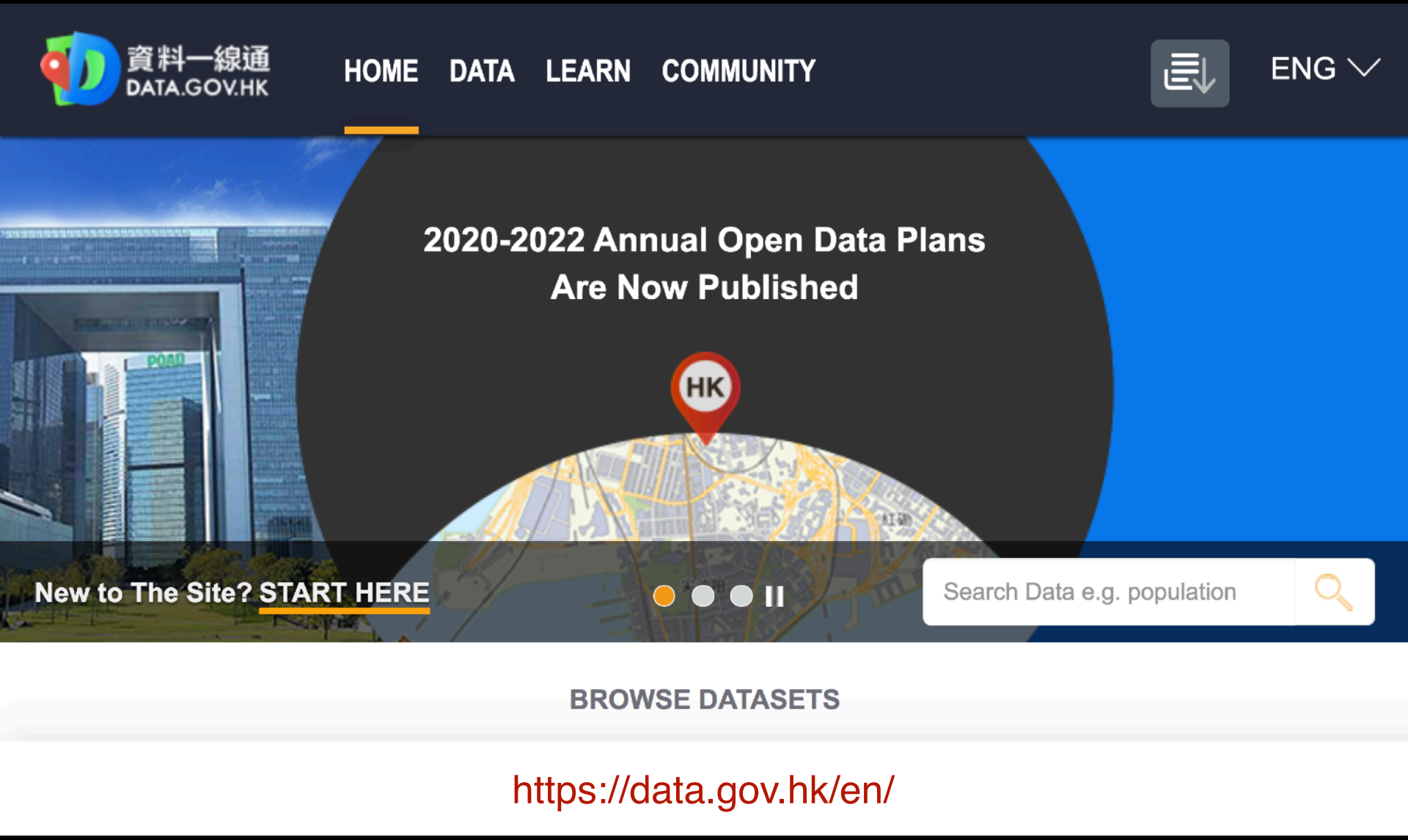

Datahack 2023
0
New
Howdy, admin
Dashboard
Posts
Media
Post Grid Combo
Pages
Comments
Restaurants
Appearance
Plugins
Users
Tools
Settings
Pods Admin
All Export
New Export
Manage Exports
Settings
WP Data Access
Ultimate CSV Importer Free
FakerPress
Collapse menu
WP ALL EXPORT
New Export
Support | Documentation
First, choose what to export.
Specific Post Type
WP_Query Results
Choose a post type...
Created by soflyy
Thank you for creating with WordPress.
Version 6.1.1


Datahack 2023
0
New
Howdy, admin
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
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
Thank you for creating with WordPress.
Version 6.1.1

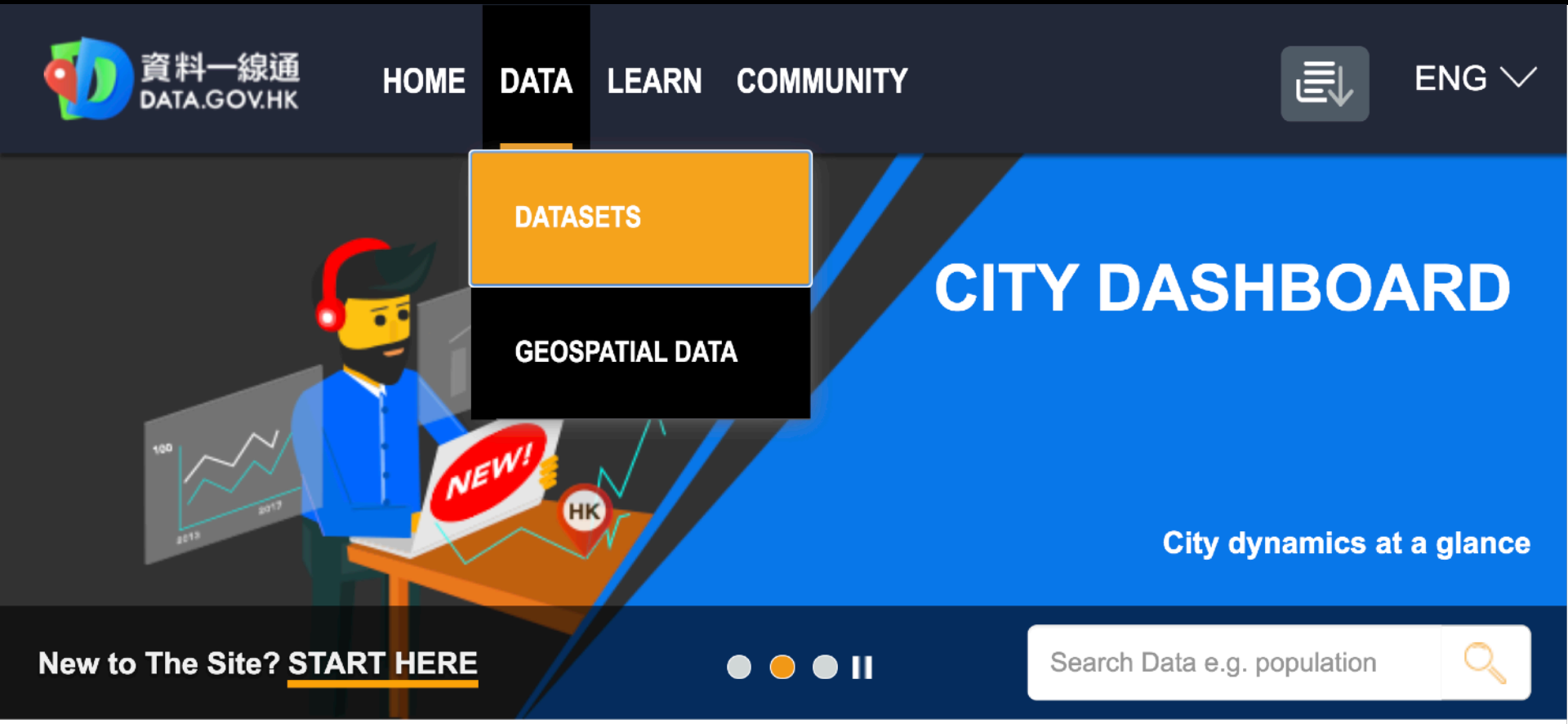

Datahack 2023
0
New
Howdy, admin
Dashboard
Posts
Media
Post Grid Combo
Pages
Comments
Restaurants
Appearance
Plugins
Users
Tools
Settings
Pods Admin
All Export
New Export
Manage Exports
Settings
WP Data Access
Ultimate CSV Importer Free
FakerPress
Collapse menu
WP ALL EXPORT
Drag & Drop
Support | Documentation
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
Advanced Options
Export Type
Save settings as a template
Load Template...
Back
Continue


Datahack 2023
0
New
Howdy, admin
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
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
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


DATABASE
EXPORT
OPEN DATA FILES
XML
CSV
JSON
SCRAPING
API
API
IOT
MULTIMEDIA
DSP
DB
DSP - Data Sourcing and Preparation
DB - Database
Visualization and Interaction



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



https://data.gov.hk/en-data/dataset/hk-lcsd-facility-facility-bkbc

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 Dictionar(ies) : https://www.lcsd.gov.hk/datagovhk/facility/facility-bkbc_data_dictionary.pdf

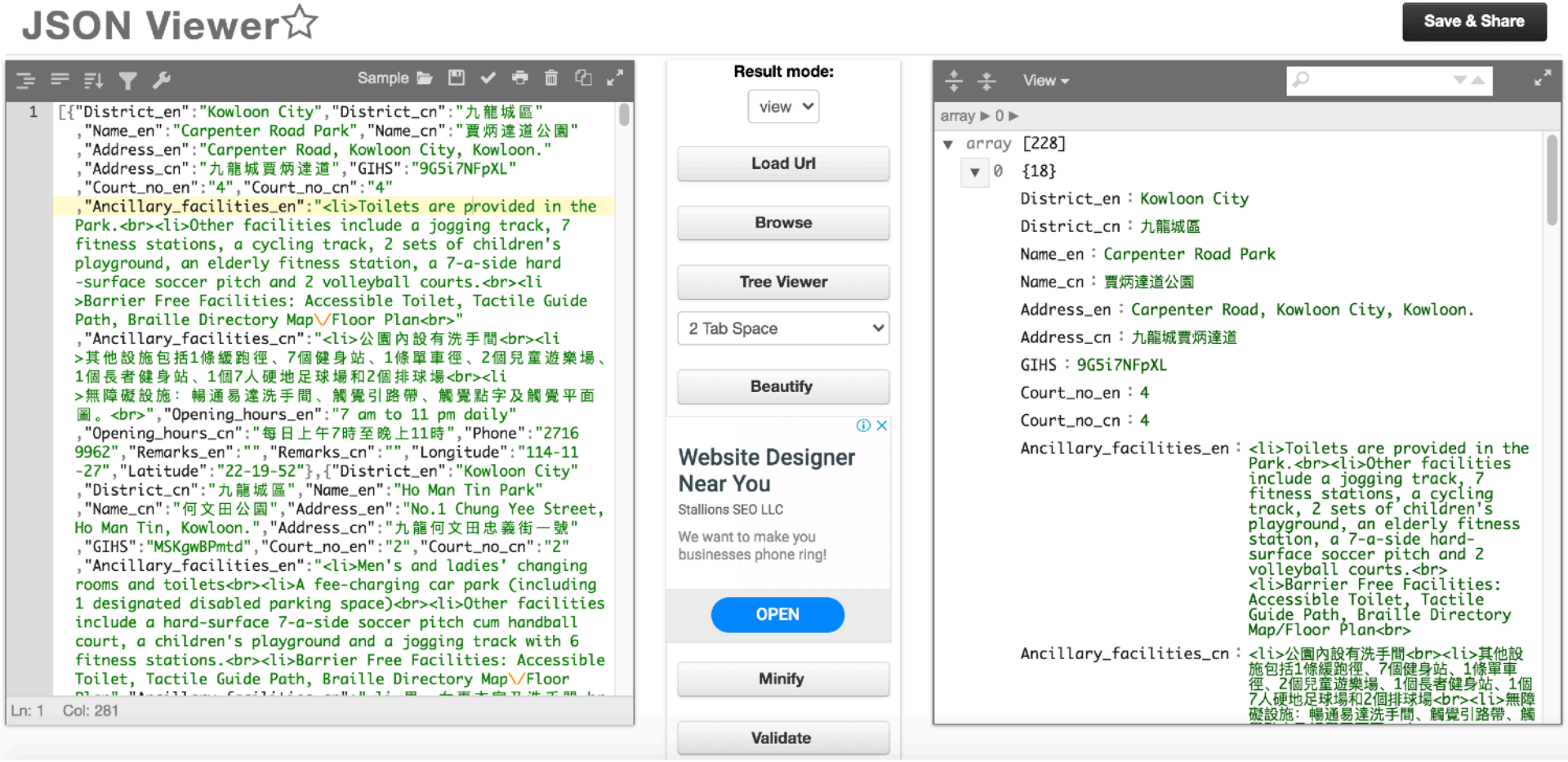
---

1 JSON file(s)

Add All to Queue

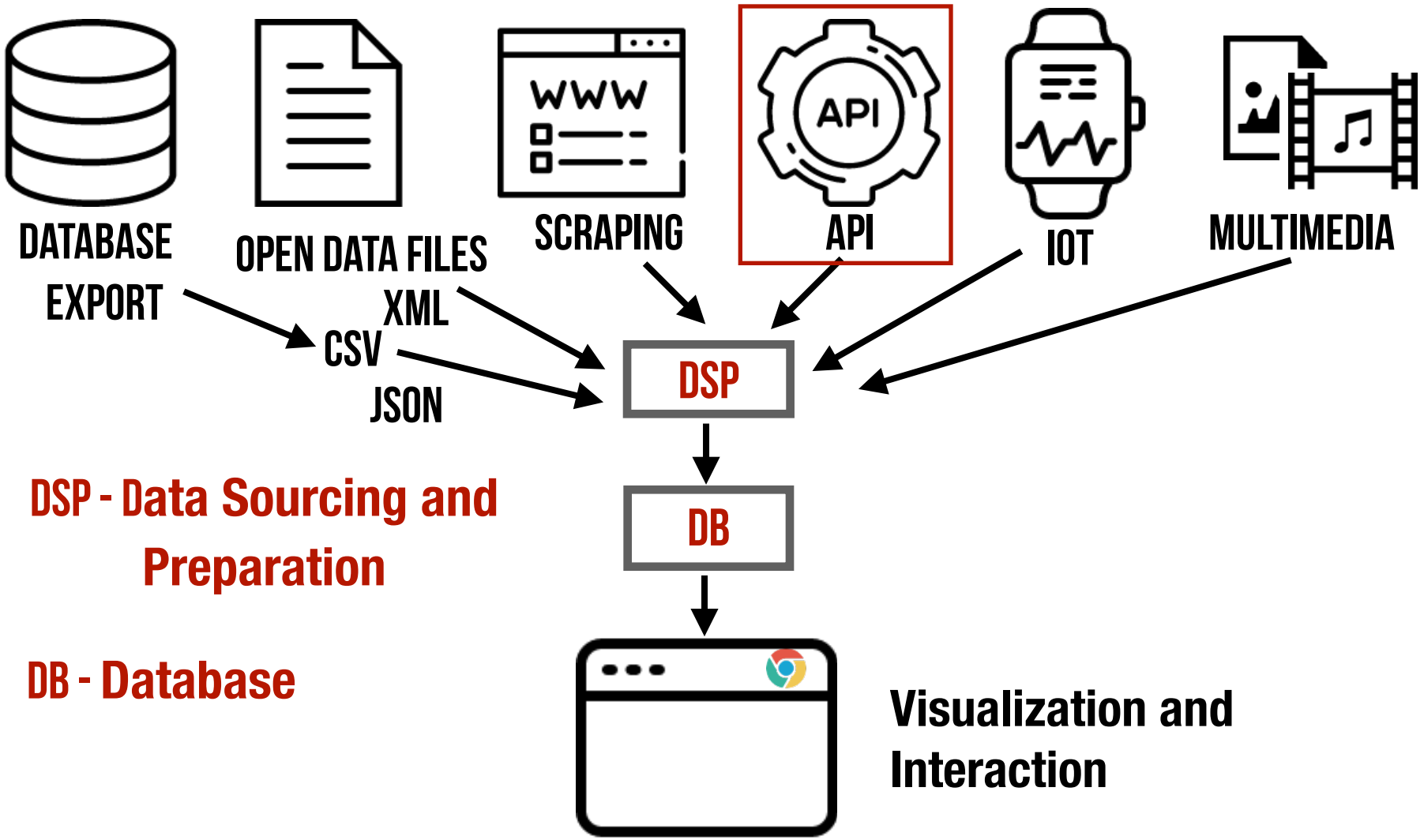
[{"District_en":"Kowloon City","District_cn":"九龍城區","Name_en":"Carpenter Road Park","Name_cn":"賈炳達道公園","Address_en":"Carpenter Road, Kowloon City, Kowloon.","Address_cn":"九龍城賈炳達道","GIHS":"9G5i7NFpXL","Court_no_en":"4","Court_no_cn":"4","Ancillary_facilities_en":"<li>Toilets are provided in the Park.
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.
Barrier Free Facilities: Accessible Toilet, Tactile Guide Path, Braille Directory Map\/Floor Plan
","Ancillary_facilities_cn":"<li>公園內設有洗手間
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
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":"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
A fee-charging car park (including 1 designated disabled parking space)
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.
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
1個收費停車場 (設有1個殘疾人士專用車位)
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
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
Other facilities include 2 hard-surface 5-a-side soccer, a children's playground and elderly fitness equipment.
Barrier Free Facilities: Accessible Toilet, Tactile Guide Path, Braille Directory Map\/Floor Plan","Ancillary_facilities_cn":"<li>男、女洗手間
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
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

- 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



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	靠背壟道遊樂場	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 Heung Road, To Kwa Wan, Kowloon


DATABASE
EXPORT
OPEN DATA FILES
XML
CSV
JSON
SCRAPING
API
API
IOT
MULTIMEDIA
DSP
DB
DSP - Data Sourcing and Preparation
DB - Database
Visualization and Interaction


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

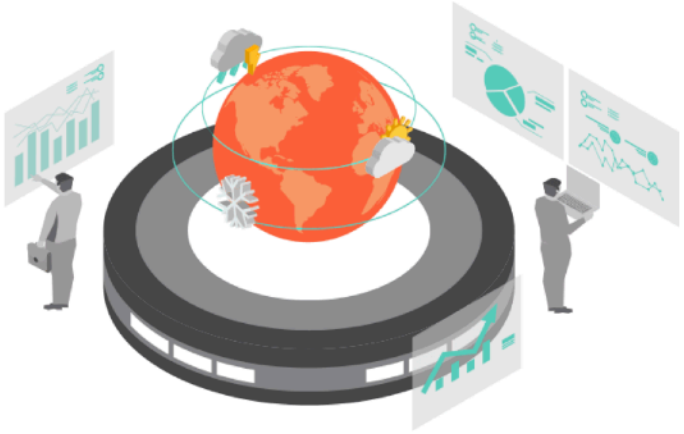
API Demonstration (in JSON format)

OpenWeather

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/

OpenWeather

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

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 automatic convert city names and zip-codes to geo coordinates and the other way around, please use

https://openweathermap.org/current

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

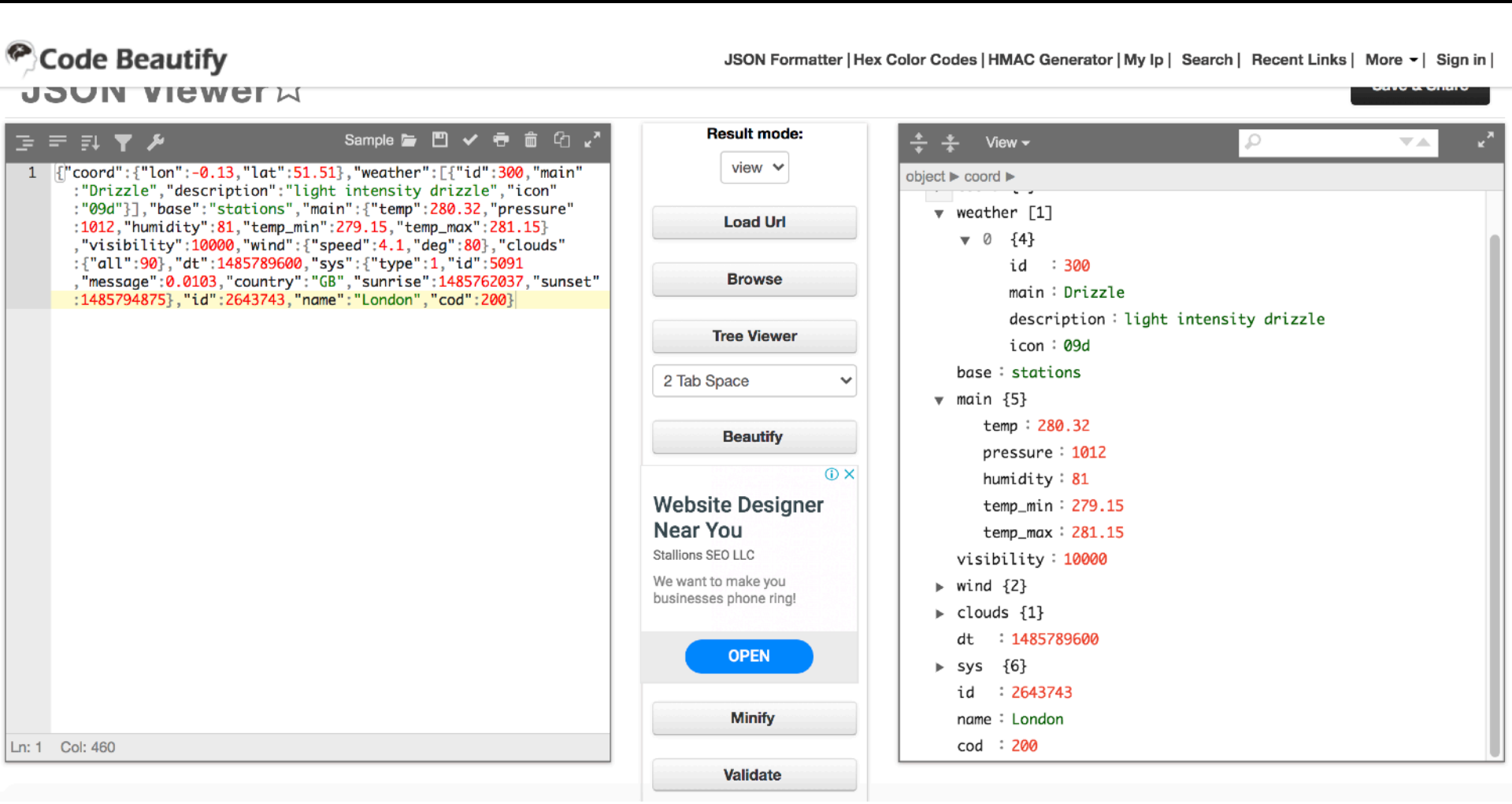
> Searching by states available only for the USA locations.
>
> There is a possibility to receive a central district of the city/town with its own parameters

***

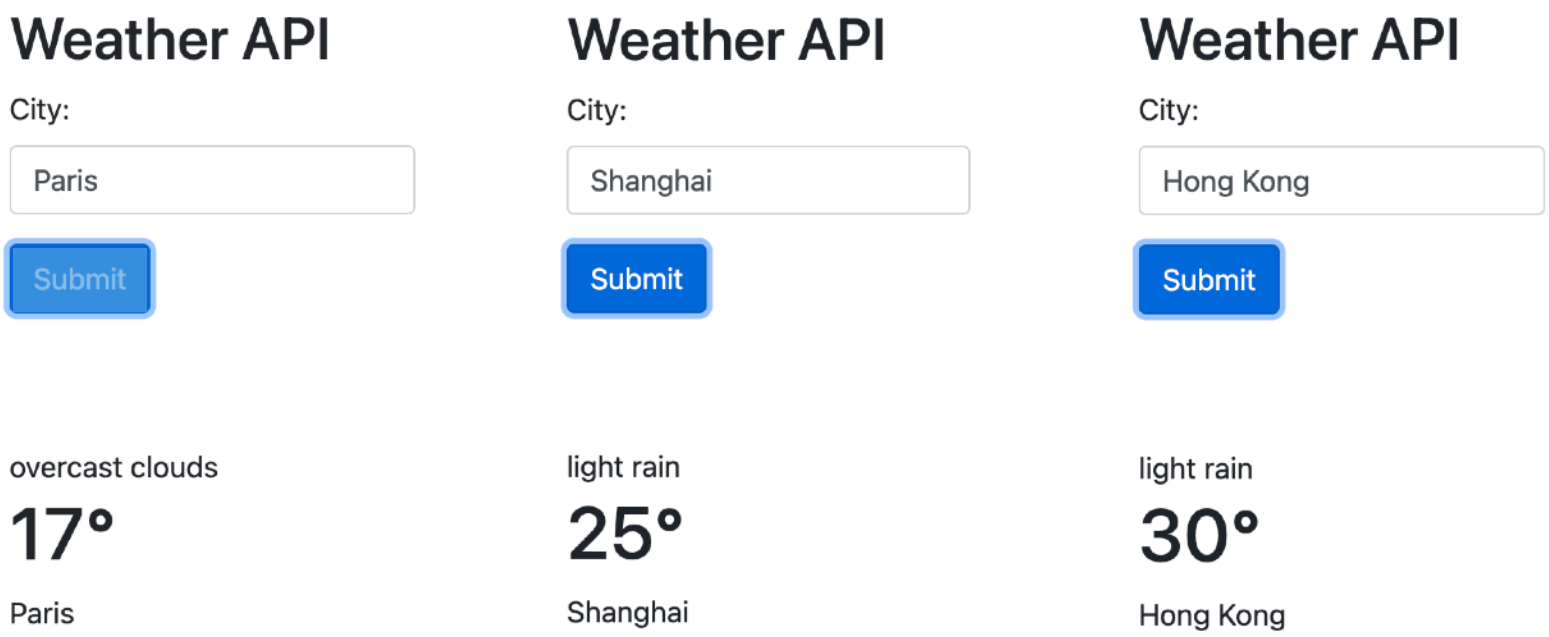
https://api.openweathermap.org/data/2.5/weather?q=london&appid=xxxxxxxxxxxx

{"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

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

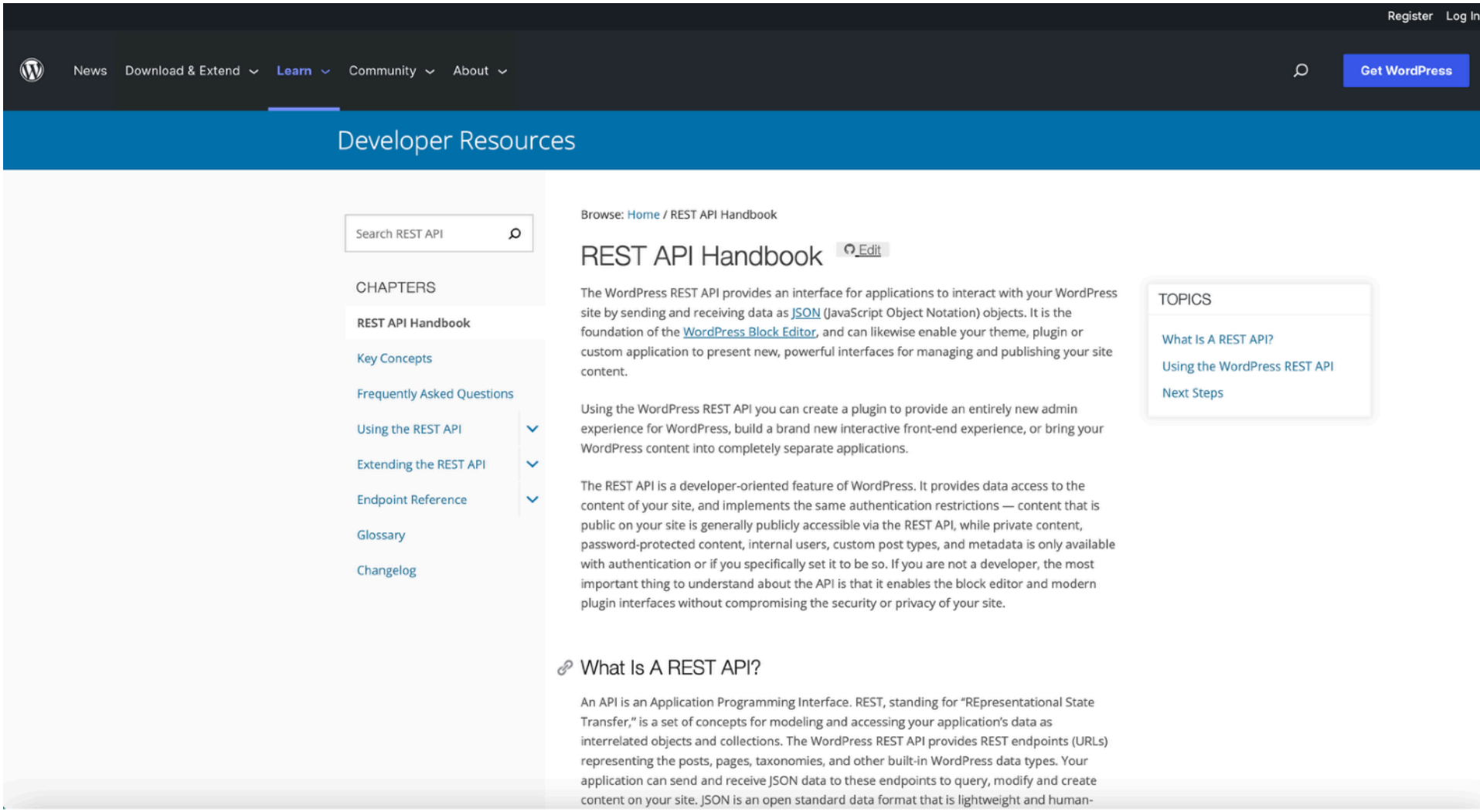
```
{
"code":"OP_SUCCESS",
"msg":"成功",
"traceId":123,
"data":{
  "resources": {
    "orgType": "1 总部, 5 门店",
    "rootOrgId": 123,
    "orgId": 123
  }
```

RESPONSE

评价管理
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
> 到餐API文档
> 茶饮版 API

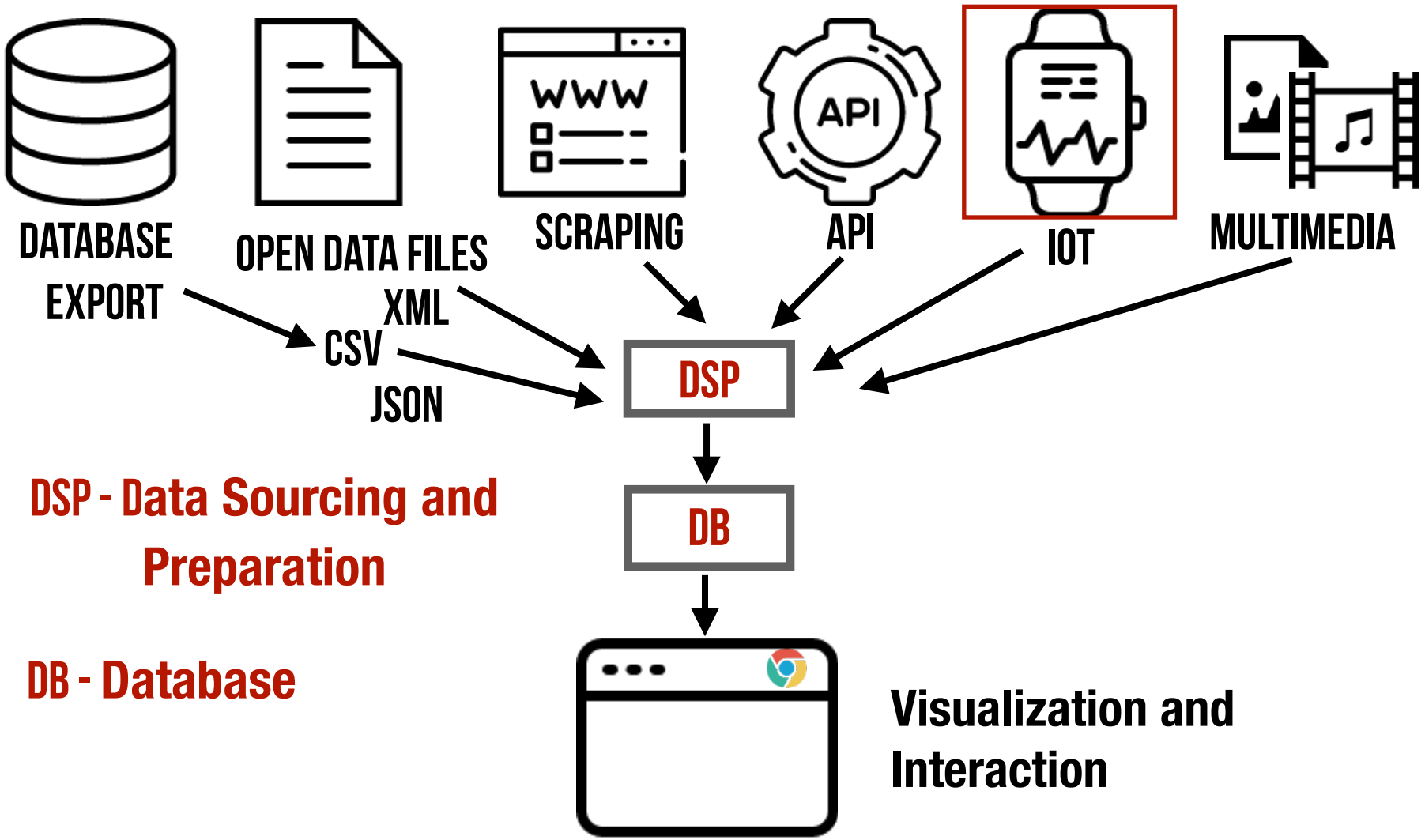
外卖门店在延迟发配送名单内才能设置延迟时间

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

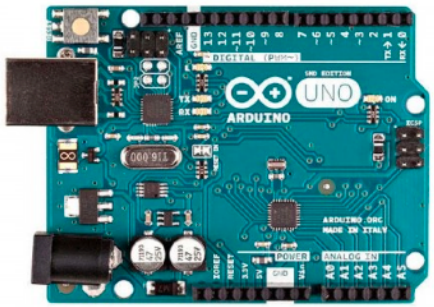

Register
Log In
News
Download & Extend
Learn
Community
About
Get WordPress
Developer Resources
Search REST API
CHAPTERS
REST API Handbook
Key Concepts
Frequently Asked Questions
Using the REST API
Extending the REST API
Endpoint Reference
Glossary
Changelog
Browse: Home / REST API Handbook
REST API Handbook
Edit
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

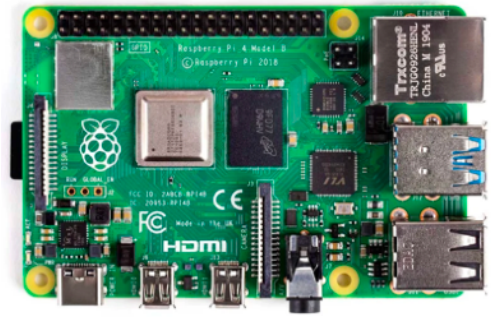
https://codebeautify.org/jsonviewer


DATABASE EXPORT
OPEN DATA FILES
XML
CSV
JSON
SCRAPING
API
IOT
MULTIMEDIA
DSP
DB
DSP - Data Sourcing and Preparation
DB - Database
Visualization and Interaction

Popular IoT Prototyping Boards

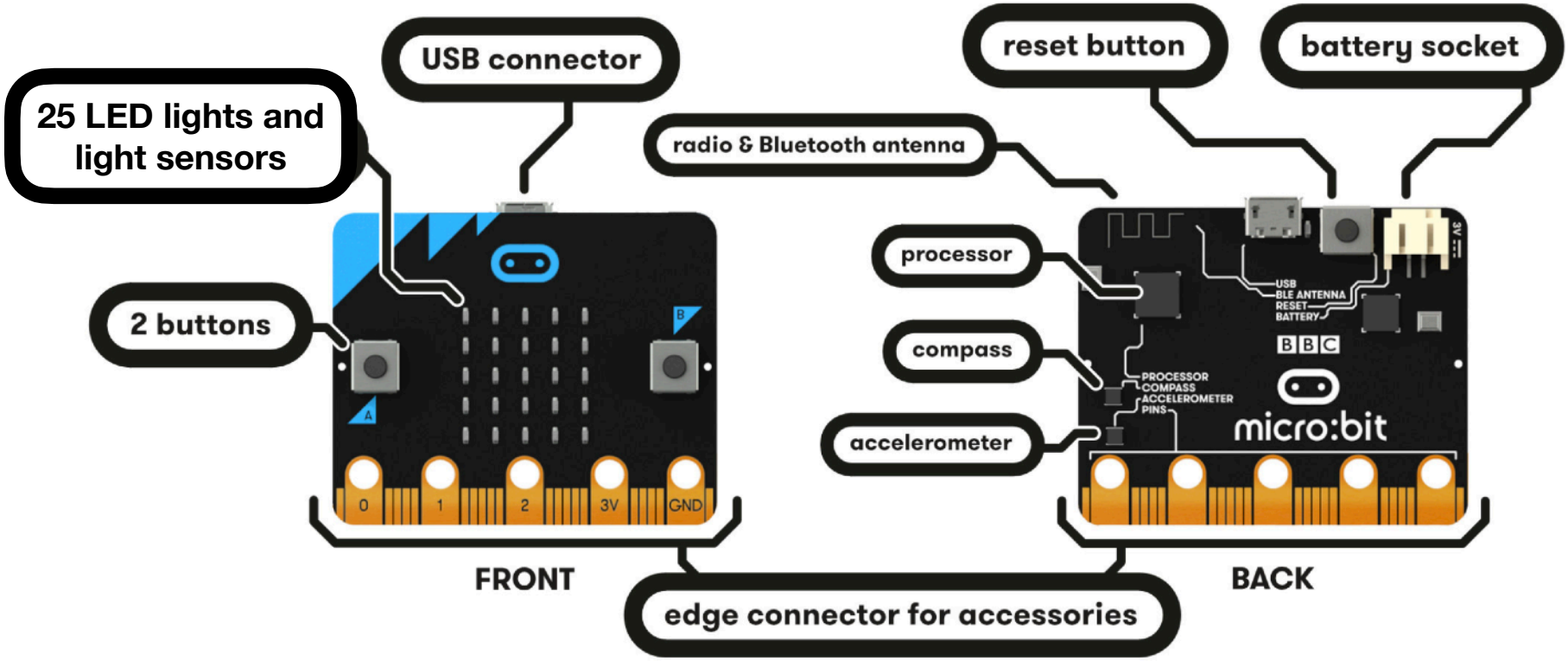


Adruino



Raspberry Pi

Micro:bit



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

Basic Structure & Components of Micro:bit

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

Browse | Structure | SQL | Search | Insert | Export | Import | Privileges | Operations | Triggers

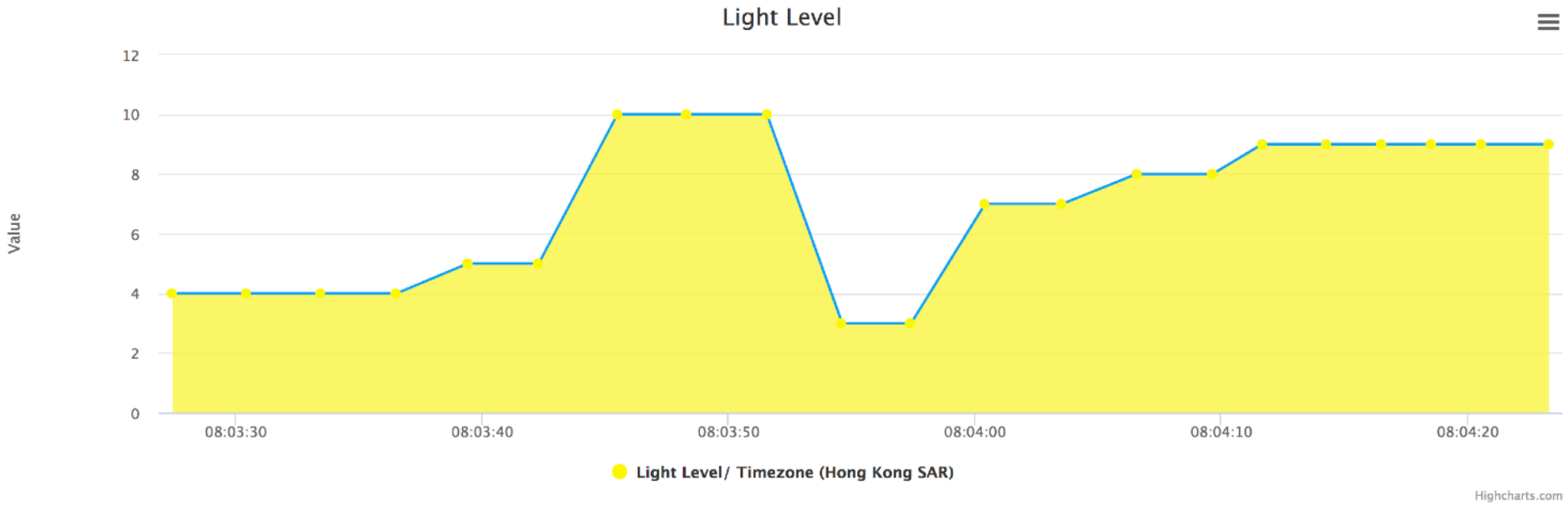
<< < 507 | Number of rows: 25 Filter rows: Search this table Sort by key: None

+ Options

				logtID	lightLevel	timeStamp
☐	Edit	Copy	Delete	15181	4	2020-09-11 22:55:24
☐	Edit	Copy	Delete	15182	5	2020-09-11 22:55:30
☐	Edit	Copy	Delete	15183	4	2020-09-11 22:55:36
☐	Edit	Copy	Delete	15184	7	2020-09-11 22:55:42
☐	Edit	Copy	Delete	15185	5	2020-09-11 22:55:48
☐	Edit	Copy	Delete	15186	7	2020-09-11 22:55:54
☐	Edit	Copy	Delete	15187	7	2020-09-11 22:56:00
☐	Edit	Copy	Delete	15188	5	2020-09-11 22:56:06
☐	Edit	Copy	Delete	15189	9	2020-09-11 22:56:12
☐	Edit	Copy	Delete	15190	4	2020-09-12 19:50:56
☐	Edit	Copy	Delete	15191	9	2020-09-12 19:51:02
☐	Edit	Copy	Delete	15192	10	2020-09-12 19:51:08
☐	Edit	Copy	Delete	15193	3	2020-09-12 19:51:17
☐	Edit	Copy	Delete	15194	9	2020-09-12 19:51:23
☐	Edit	Copy	Delete	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.

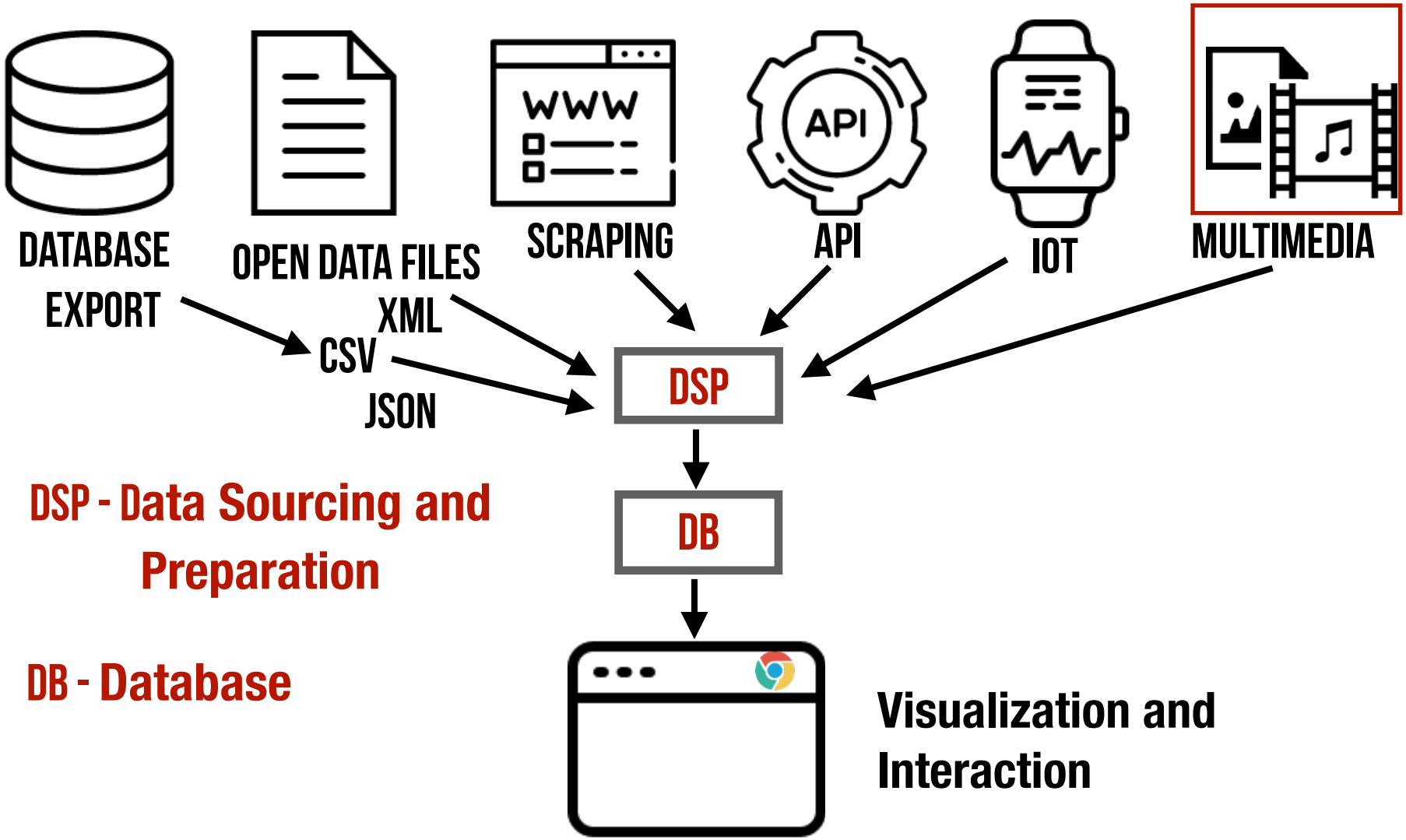

DATABASE EXPORT
OPEN DATA FILES
XML
CSV
JSON
SCRAPING
API
IOT
MULTIMEDIA
DSP
DB
DSP - Data Sourcing and Preparation
DB - Database
Visualization and Interaction

Image and Voice Recognition and Modelling Using Machine/Deep Learning

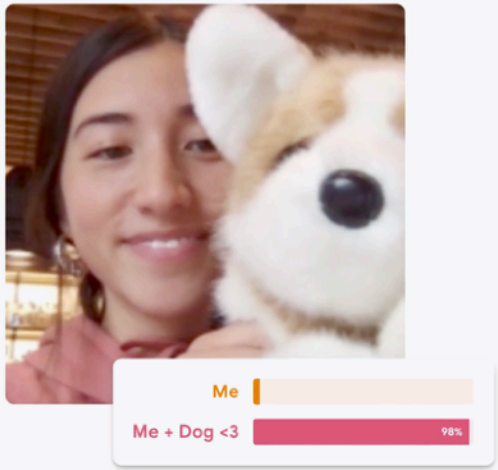
About FAQ Get Started

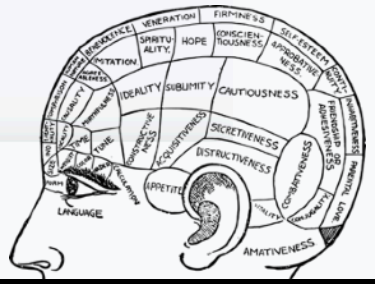
Teachable Machine

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

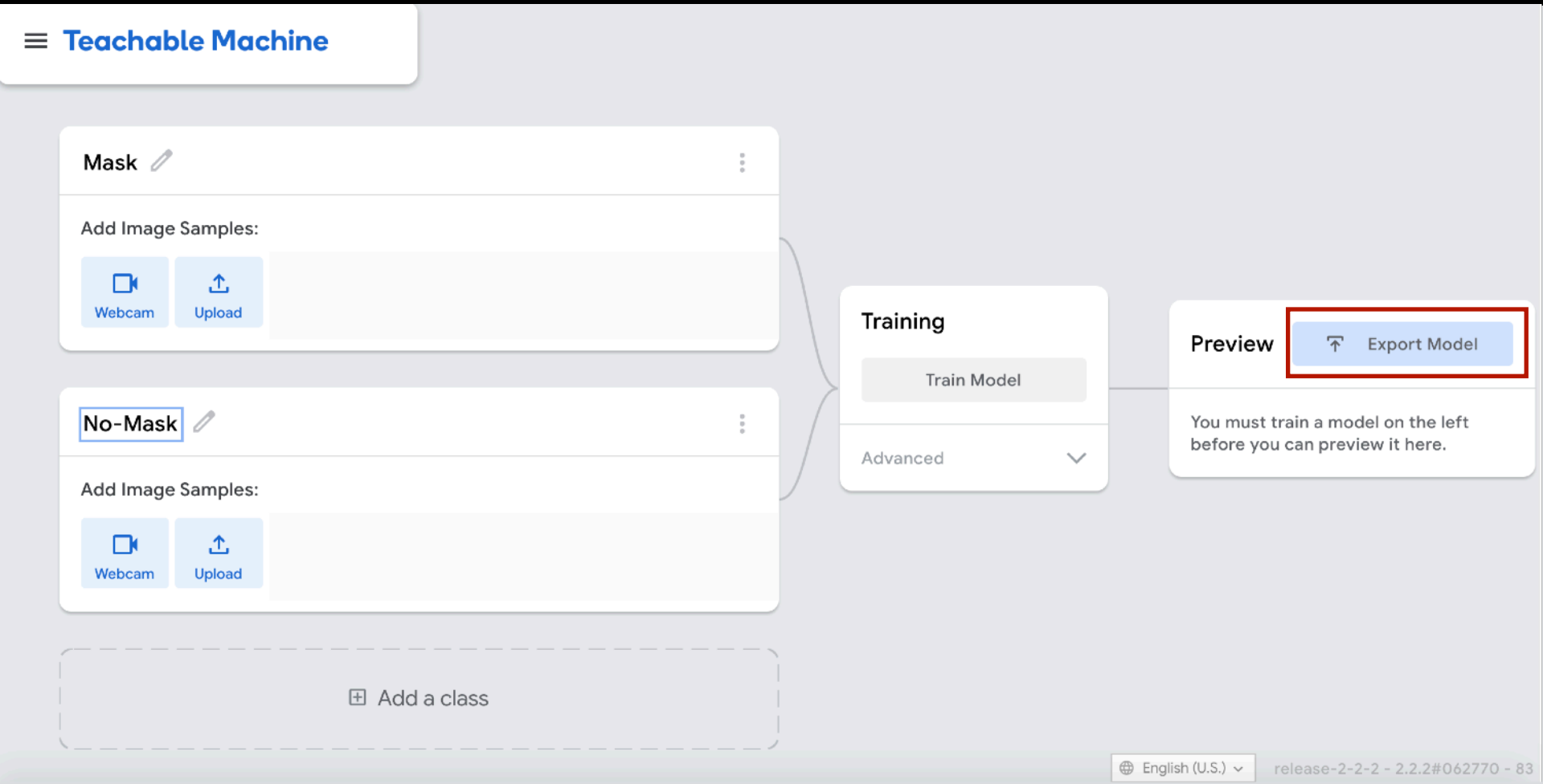
A fast, easy way to create machine learning models for your sites, apps, and more – no expertise or coding required.

Get Started



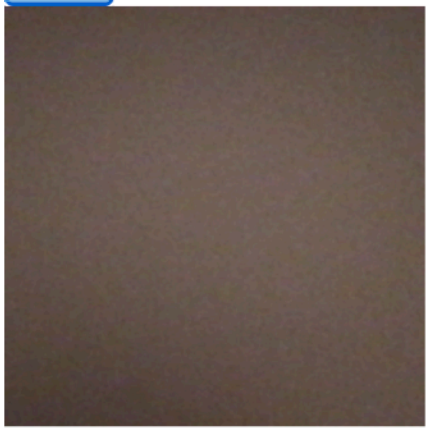


https://teachablemachine.withgoogle.com/


Teachable Machine
Mask
Add Image Samples:
Webcam
Upload
No-Mask
Add Image Samples:
Webcam
Upload
Add a class
Training
Train Model
Advanced
Preview
Export Model
You must train a model on the left before you can preview it here.
English (U.S.)
release-2-2-2 - 2.2.2#062770 - 83

Teachable Machine Image Model

Start



Mask: 0.36

No Mask: 0.64

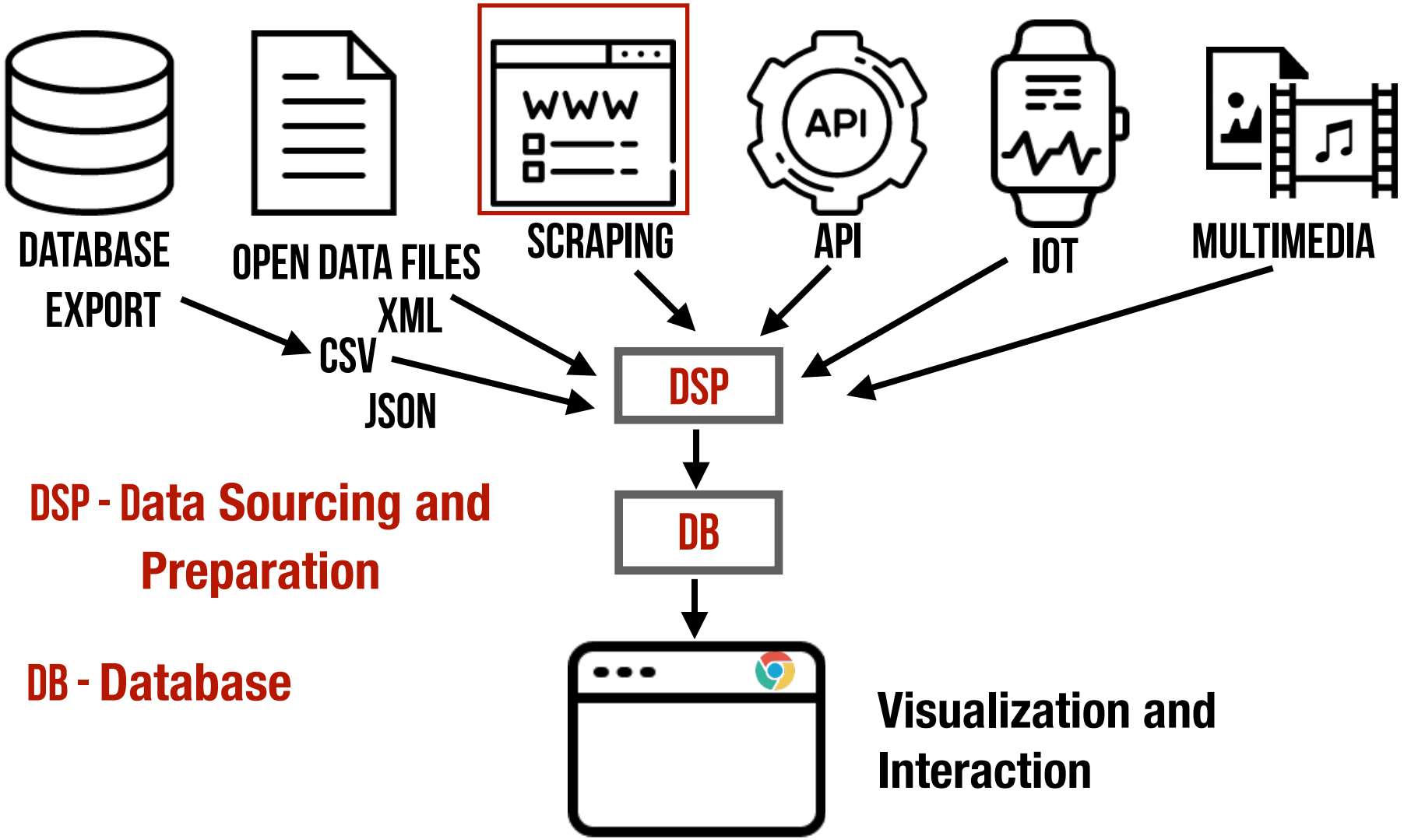
Teachable Machine

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.


DATABASE EXPORT
OPEN DATA FILES
XML
CSV
JSON
SCRAPING
API
IOT
MULTIMEDIA
DSP
DB
DSP - Data Sourcing and Preparation
DB - Database
Visualization and Interaction

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.

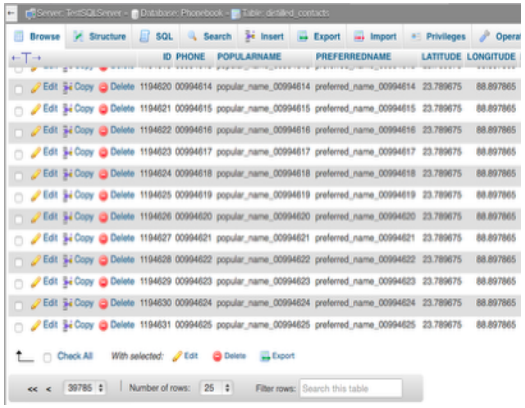
How does it work?

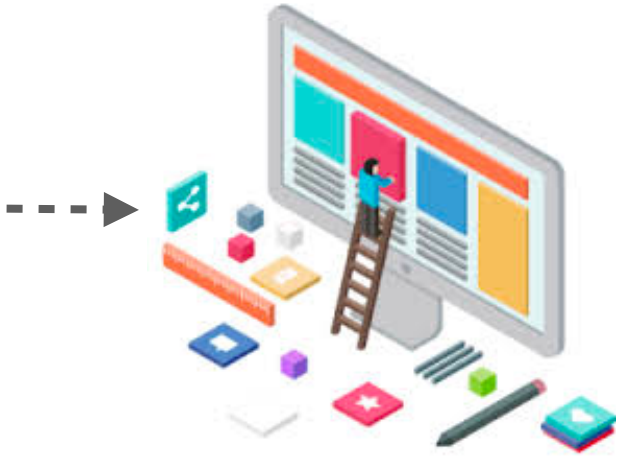
Data

Web Publishing

Web

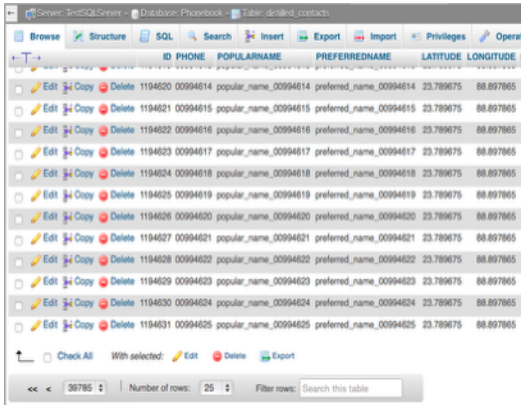
Dynamic Web Building



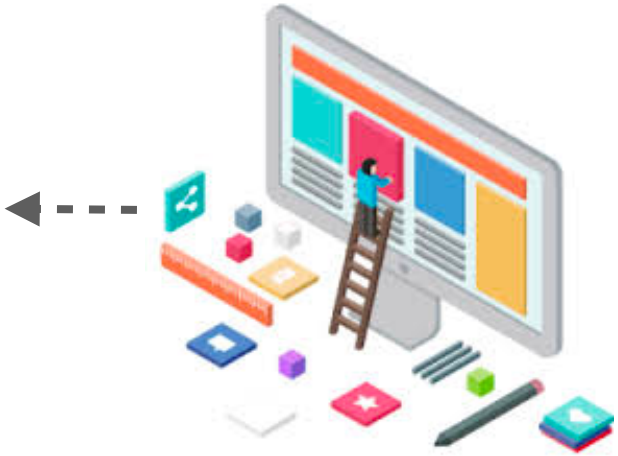


Static Web Building

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

https://www.parsehub.com

Apps | Cyberport Bootca... | Blockchain | NLP and Chatbot | FinTech | E-Commerce | Classes | Development Tools | Data Science | Machine Learning | Music | Drupal | Cloud Providers | Other Bookma

parsehub

Features Download Pricing Help Blog Log in Sign up

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

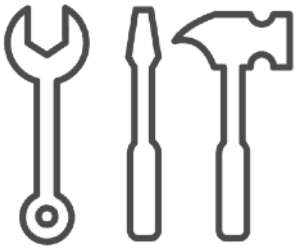
Access data via JSON, Excel and API. Data is collected by our servers.

iting for nexus-websocket-a.intercom.io...

bernard@intechni...

- 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

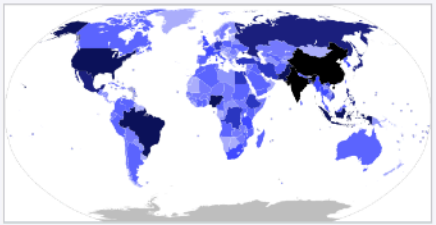
1. Parsehub 101
2. Pagination ('next' page buttons)
3. Scrape product details
4. Scrape leads from directories

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

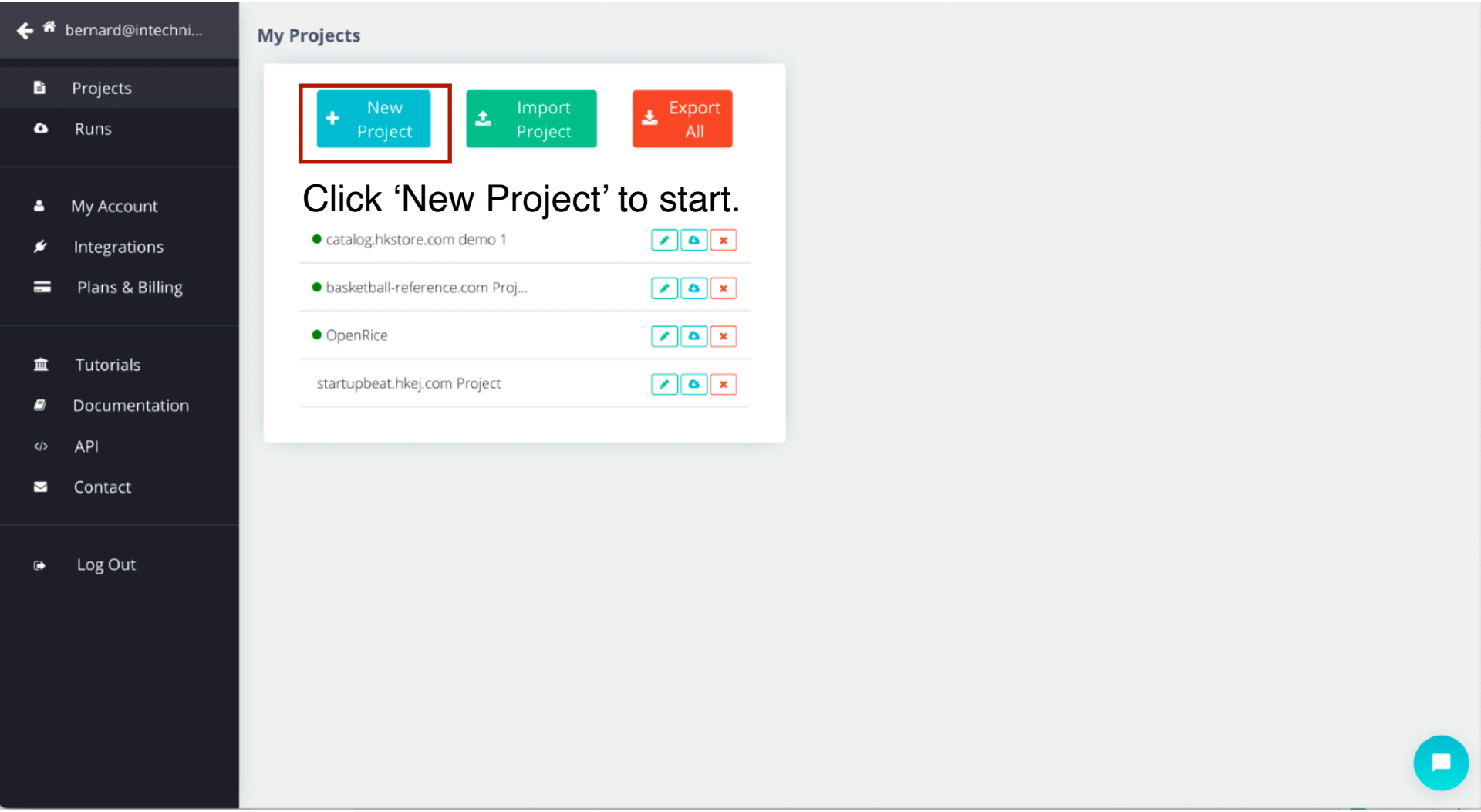
A cartogram of the world population in 2018

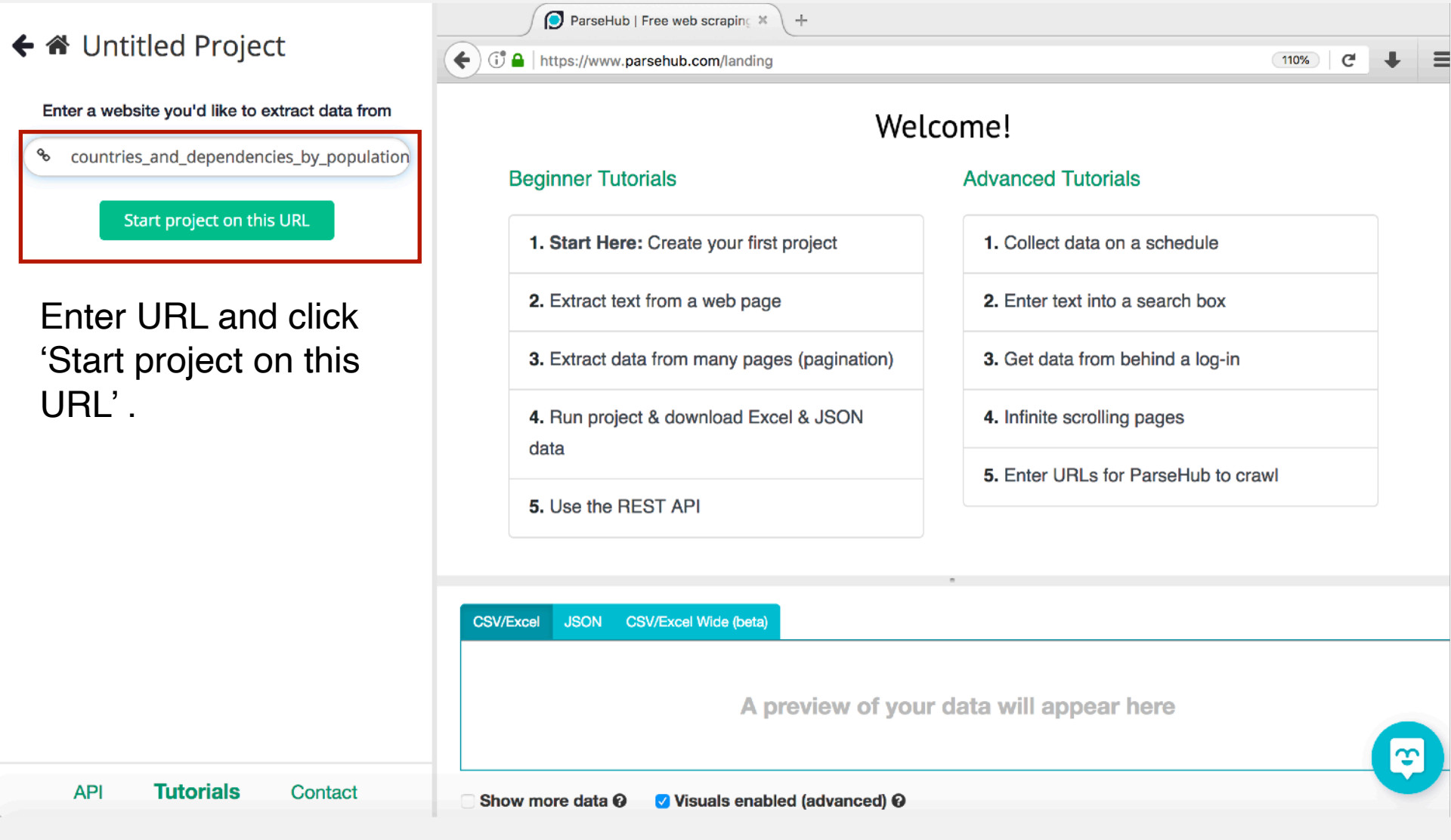
https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population

Sovereign states and dependencies by population [edit]

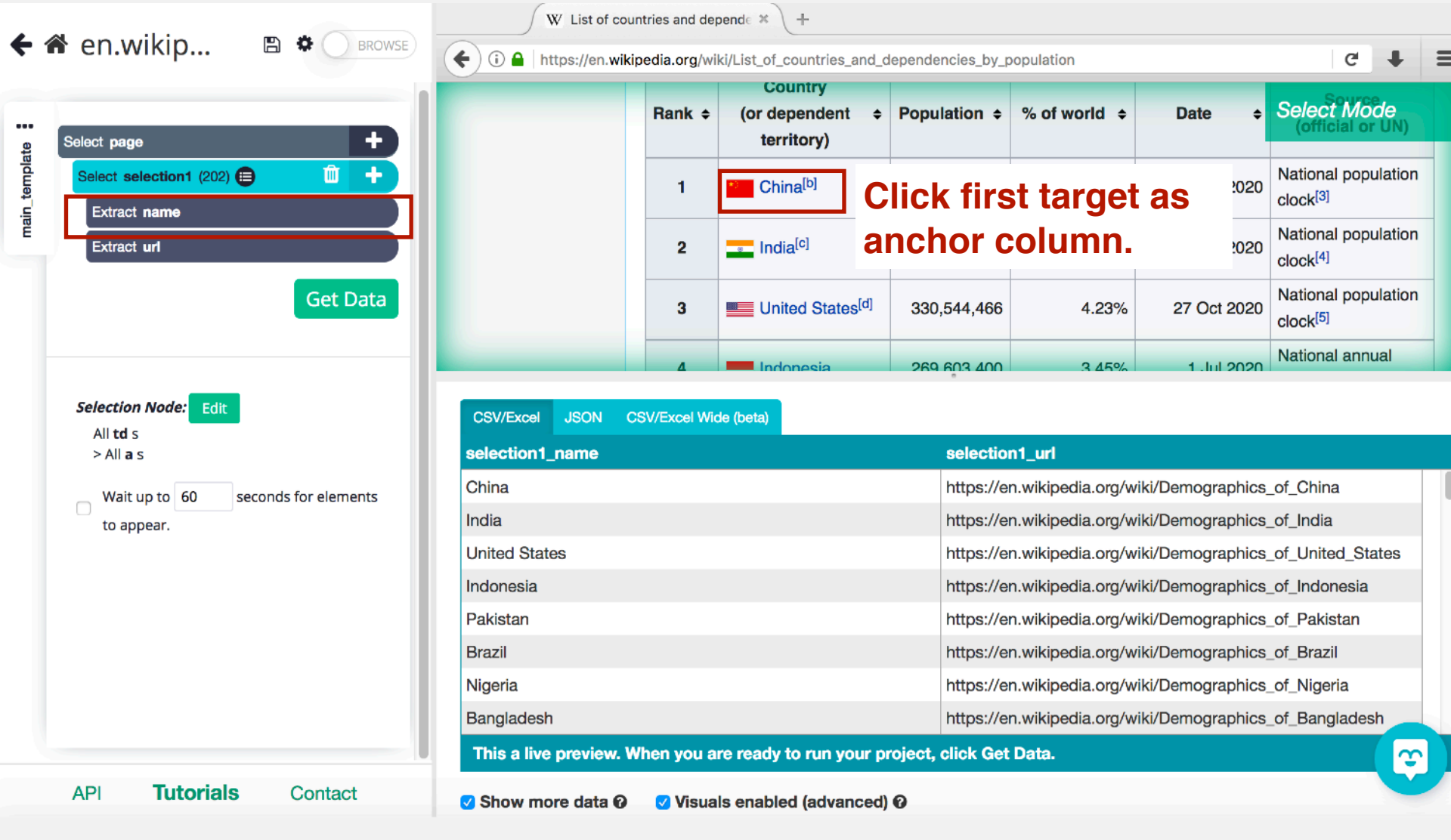
Note: A number rank is assigned to the 193 member states of the United Nations and the two observer states to the United Nations General Assembly. Dependent territories and constituent countries that are parts of sovereign states are shown in *italics* and not assigned a numbered rank. In addition, sovereign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

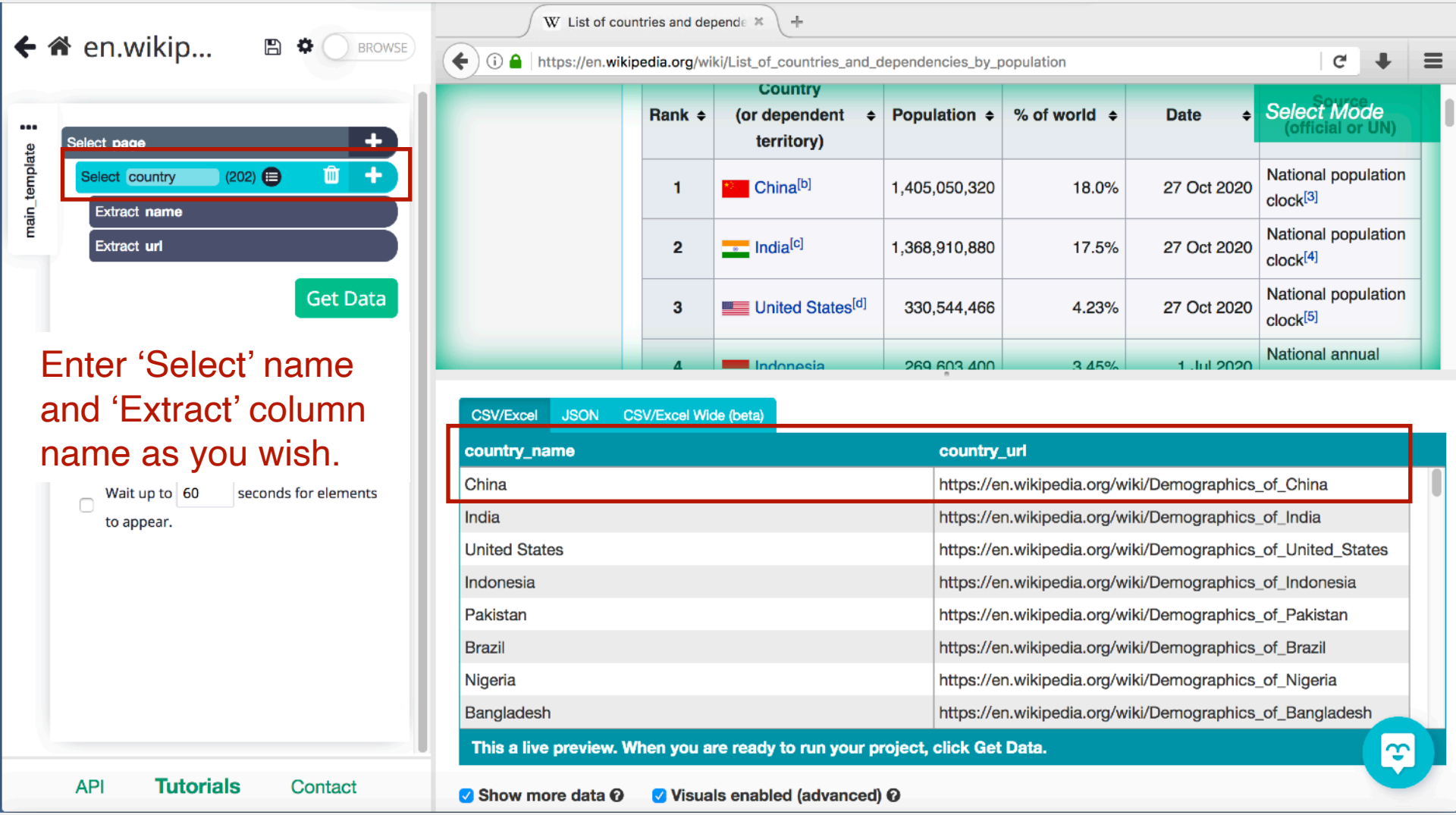
Rank	Country (or dependent territory)	Population	% of world	Date	Source (official or UN)
1	China[b]	1,405,239,080	18.0%	9 Nov 2020	National population clock[3]
2	India[c]	1,369,434,247	17.5%	9 Nov 2020	National population clock[4]
3	United States[d]	330,617,844	4.22%	9 Nov 2020	National population clock[5]
4	Indonesia	269,603,400	3.45%	1 Jul 2020	National annual projection[6]
5	Pakistan[e]	220,892,331	2.82%	1 Jul 2020	UN Projection[2]
6	Brazil	212,311,986	2.71%	9 Nov 2020	National population clock[7]
7	Nigeria	206,139,587	2.63%	1 Jul 2020	UN Projection[2]
8	Bangladesh	160,615,266	2.17%	9 Nov 2020	National population

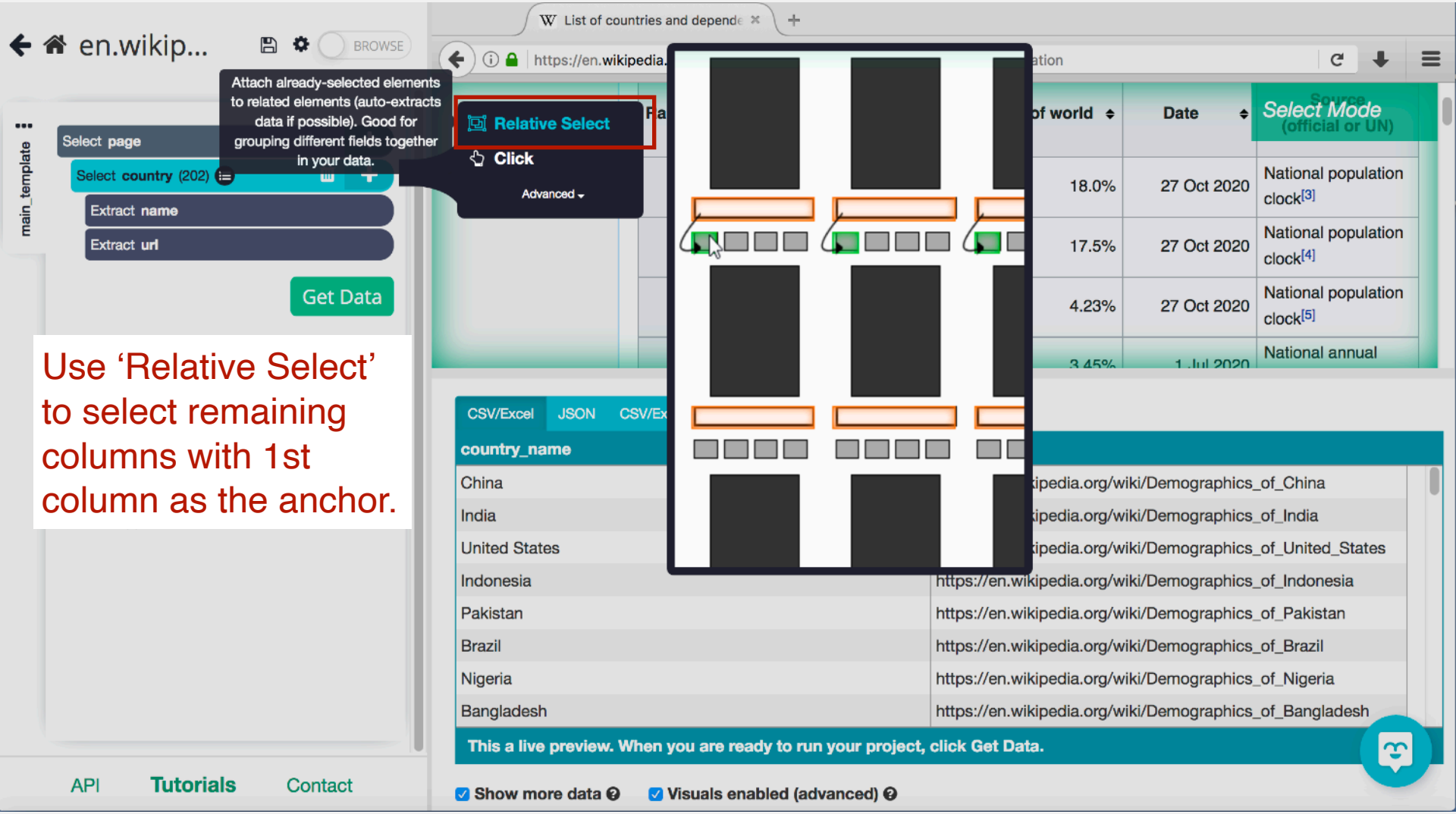

bernard@intechni...
Projects
Runs
My Account
Integrations
Plans & Billing
Tutorials
Documentation
API
Contact
Log Out
My Projects
New Project
Import Project
Export All
Click ‘New Project’ to start.
catalog.hkstore.com demo 1
basketball-reference.com Proj...
OpenRice
startupbeat.hkej.com Project



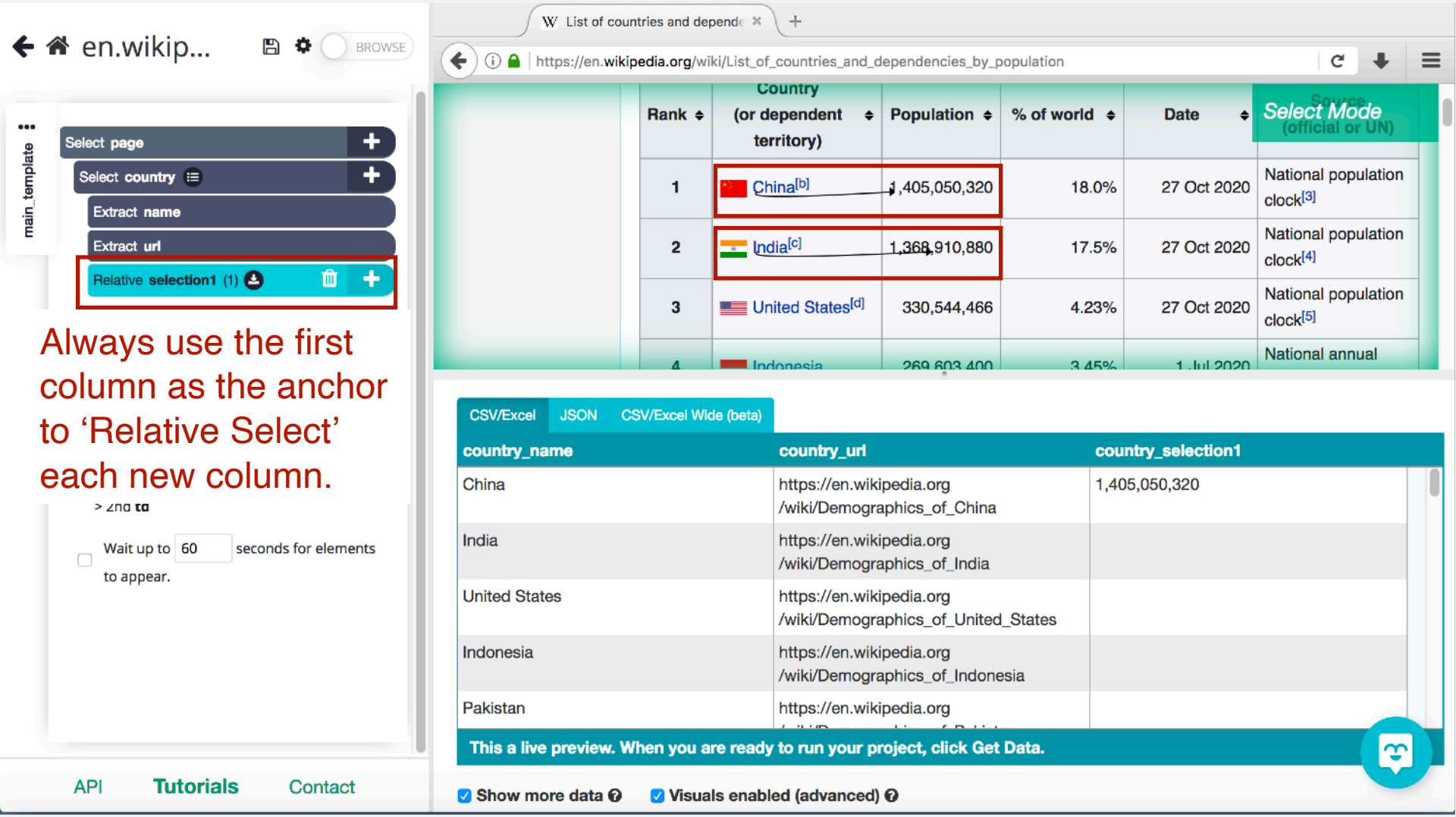
Enter URL and click ‘Start project on this URL’ .

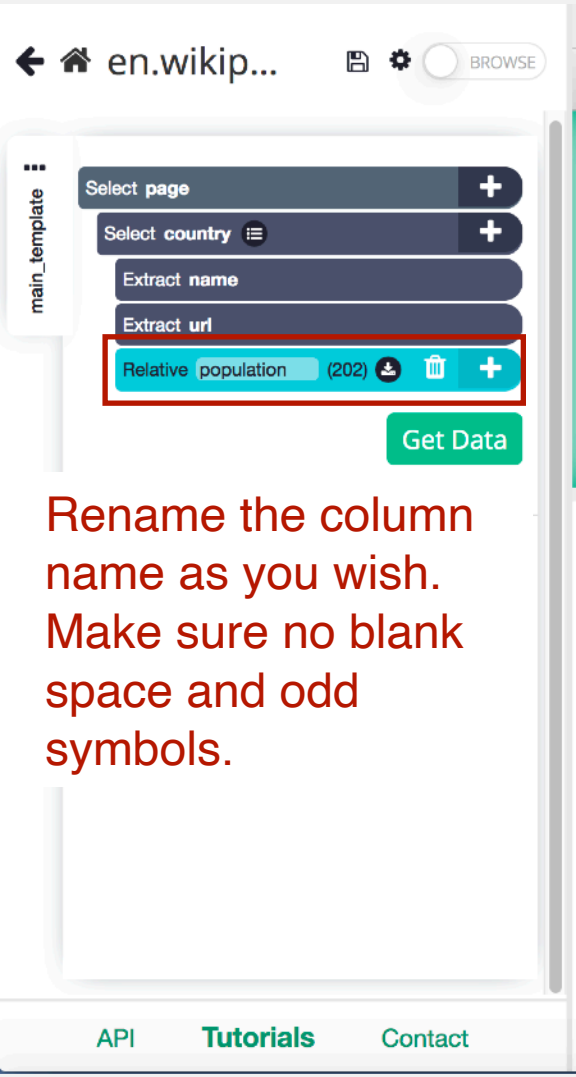

en.wikip...
BROWSE
main_template
Select page
Select selection1 (202)
Extract name
Extract url
Get Data
Selection Node:
Edit
All td s
> All a s
Wait up to 60 seconds for elements to appear.
API
Tutorials
Contact
List of countries and depend
https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population
Rank
Country (or dependent territory)
Population
% of world
Date
Source (official or UN)
Select Mode
1
China[b]
National population clock[3]
2
India[c]
National population clock[4]
3
United States[d]
330,544,466
4.23%
27 Oct 2020
National population clock[5]
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
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)


en.wikip...
BROWSE
main_template
Select page
Select country (202)
Extract name
Extract url
Get Data
Enter 'Select' name and 'Extract' column name as you wish.
Wait up to 60 seconds for elements to appear.
API
Tutorials
Contact
List of countries and depende
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)
Select Mode
1 China[b] 1,405,050,320 18.0% 27 Oct 2020 National population clock[3]
2 India[c] 1,368,910,880 17.5% 27 Oct 2020 National population clock[4]
3 United States[d] 330,544,466 4.23% 27 Oct 2020 National population clock[5]
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)

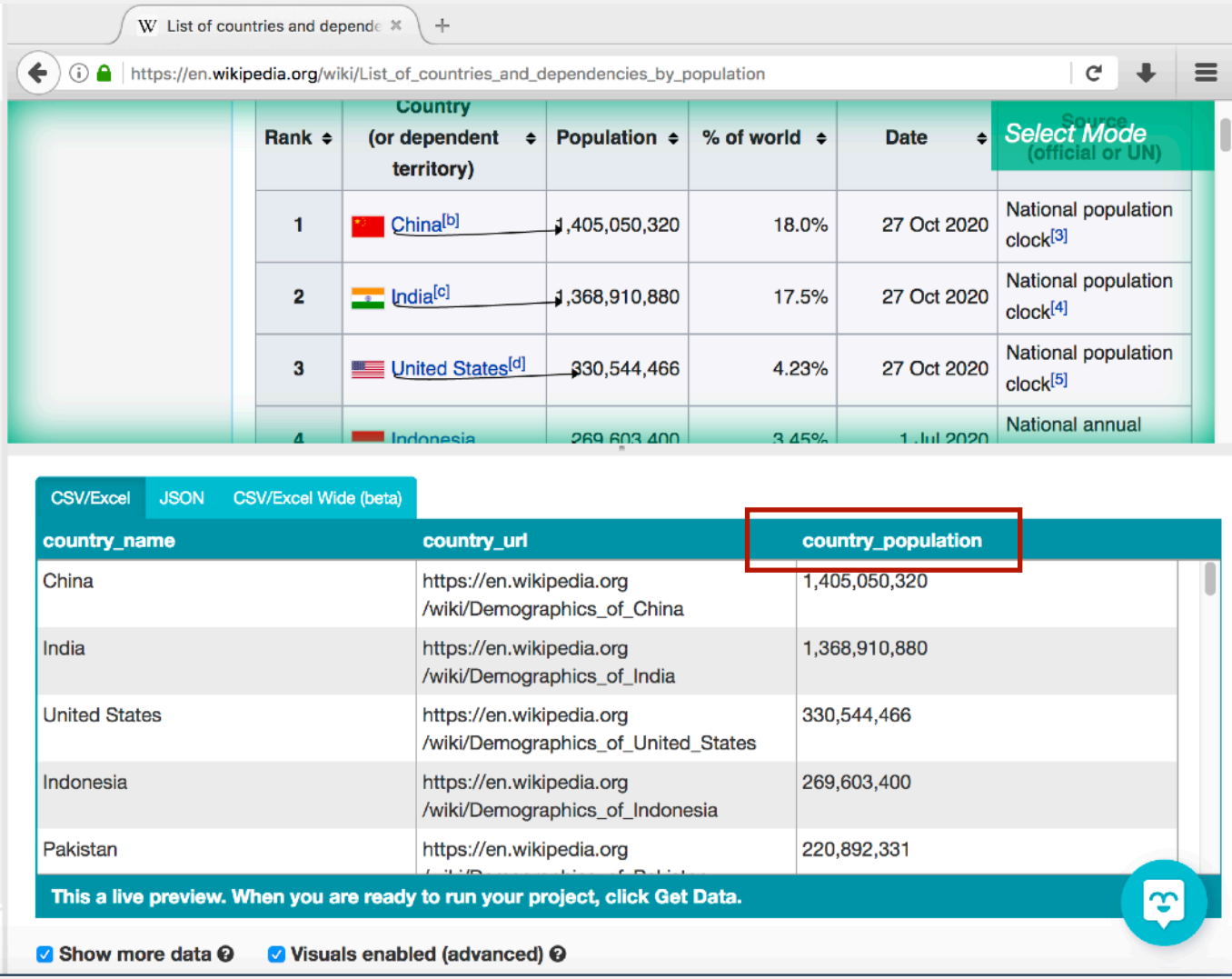


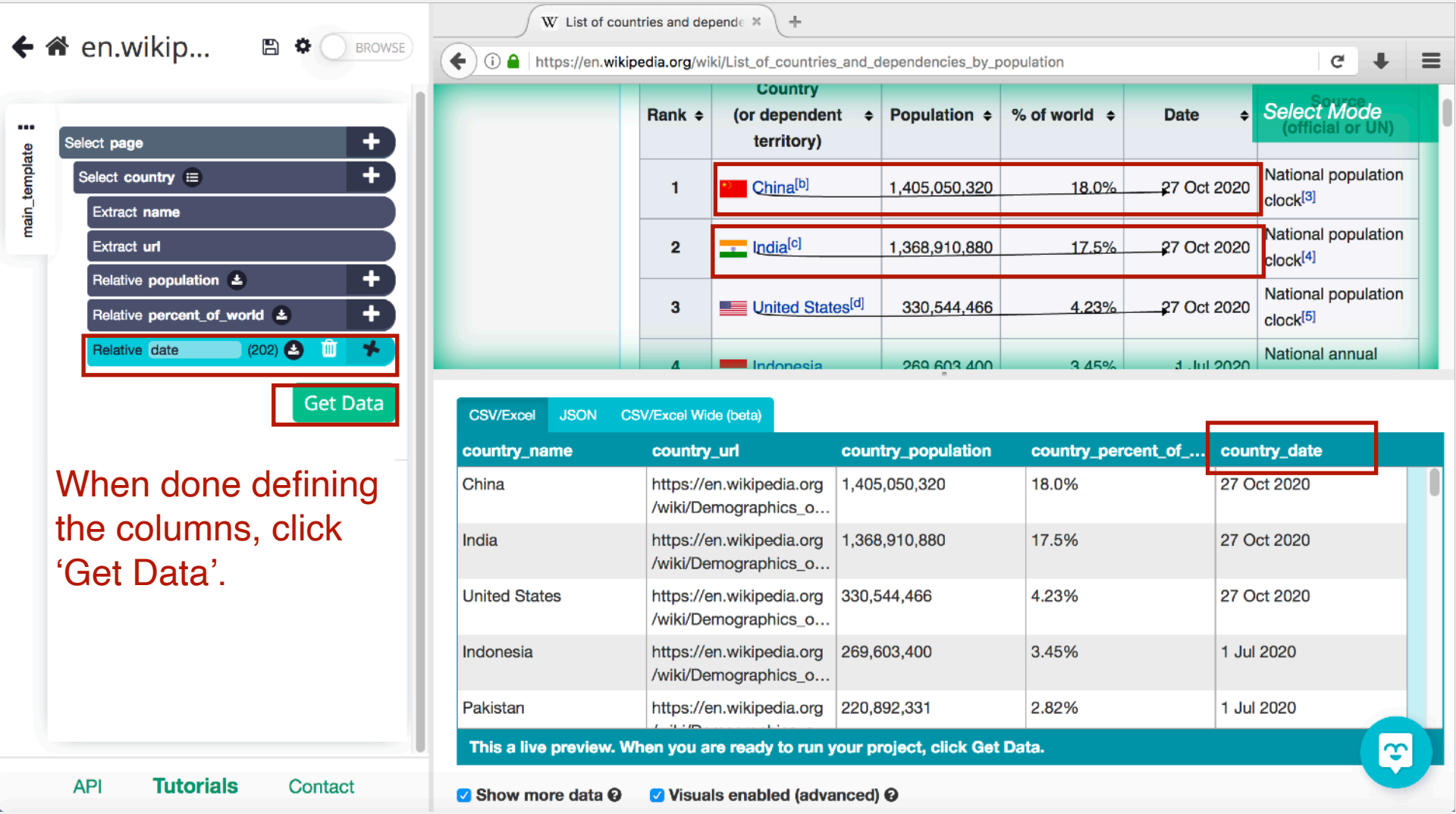
Use ‘Relative Select’ to select remaining columns with 1st column as the anchor.

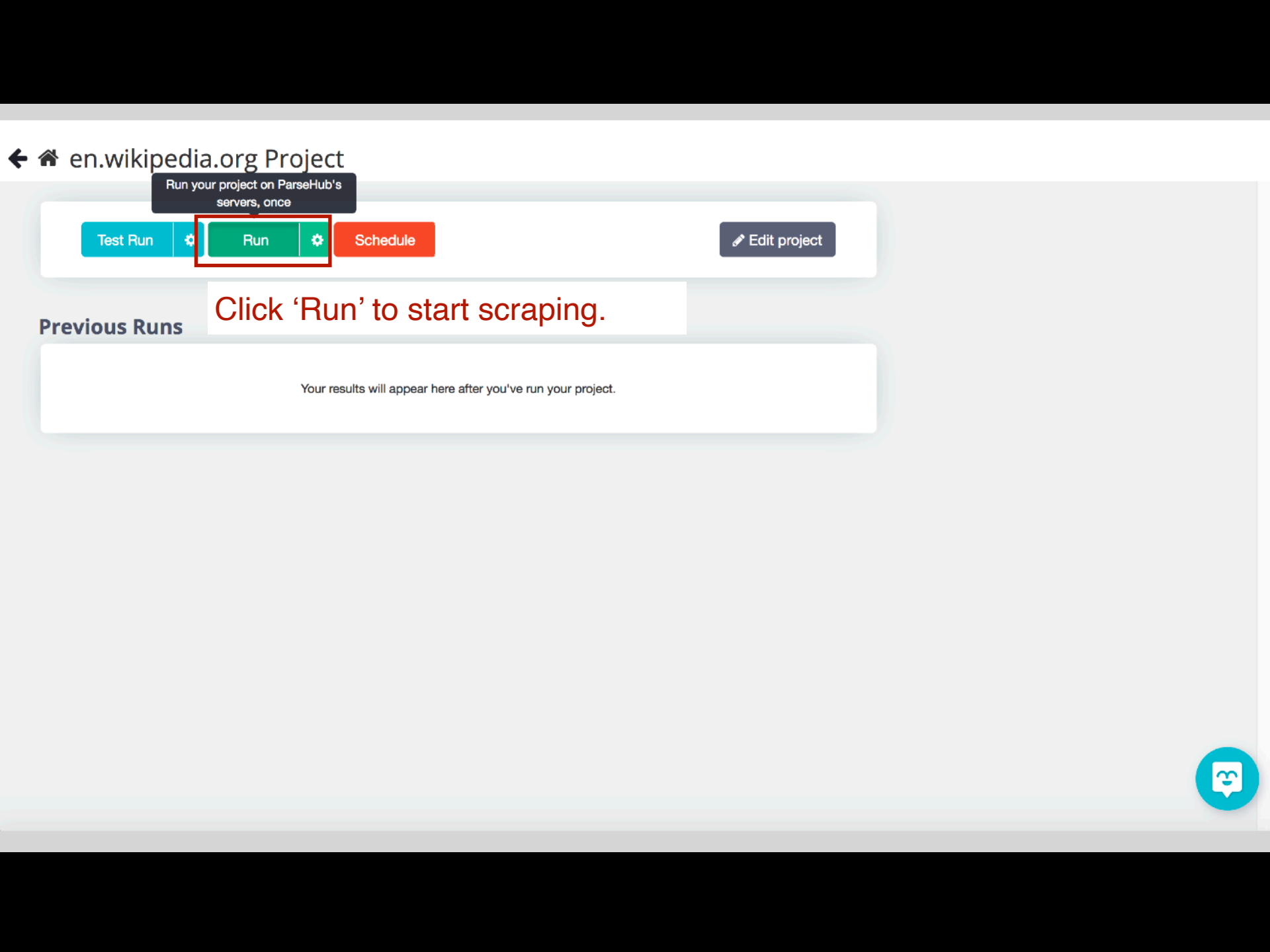

en.wikip...
BROWSE
main_template
Select page
Select country
Extract name
Extract url
Relative selection1 (1)
Always use the first column as the anchor to 'Relative Select' each new column.
Wait up to 60 seconds for elements to appear.
API
Tutorials
Contact
List of countries and depend
https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population
Rank
Country (or dependent territory)
Population
% of world
Date
Select Mode
(official or UN)
1
China[b]
1,405,050,320
18.0%
27 Oct 2020
National population clock[3]
2
India[c]
1,368,910,880
17.5%
27 Oct 2020
National population clock[4]
3
United States[d]
330,544,466
4.23%
27 Oct 2020
National population clock[5]
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
This a live preview. When you are ready to run your project, click Get Data.
Show more data
Visuals enabled (advanced)



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




en.wikip...
BROWSE
main_template
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'.
API
Tutorials
Contact
List of countries and depende
https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population
Rank
Country (or dependent territory)
Population
% of world
Date
Select Mode
(official or UN)
1 China[b] 1,405,050,320 18.0% 27 Oct 2020 National population clock[3]
2 India[c] 1,368,910,880 17.5% 27 Oct 2020 National population clock[4]
3 United States[d] 330,544,466 4.23% 27 Oct 2020 National population clock[5]
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_o... 1,405,050,320 18.0% 27 Oct 2020
India https://en.wikipedia.org/wiki/Demographics_o... 1,368,910,880 17.5% 27 Oct 2020
United States https://en.wikipedia.org/wiki/Demographics_o... 330,544,466 4.23% 27 Oct 2020
Indonesia https://en.wikipedia.org/wiki/Demographics_o... 269,603,400 3.45% 1 Jul 2020
Pakistan https://en.wikipedia.org 220,892,331 2.82% 1 Jul 2020
This 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
Click ‘Run’ to start scraping.
Previous Runs
Your results will appear here after you've run your project.

en.wikipedia.org 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.

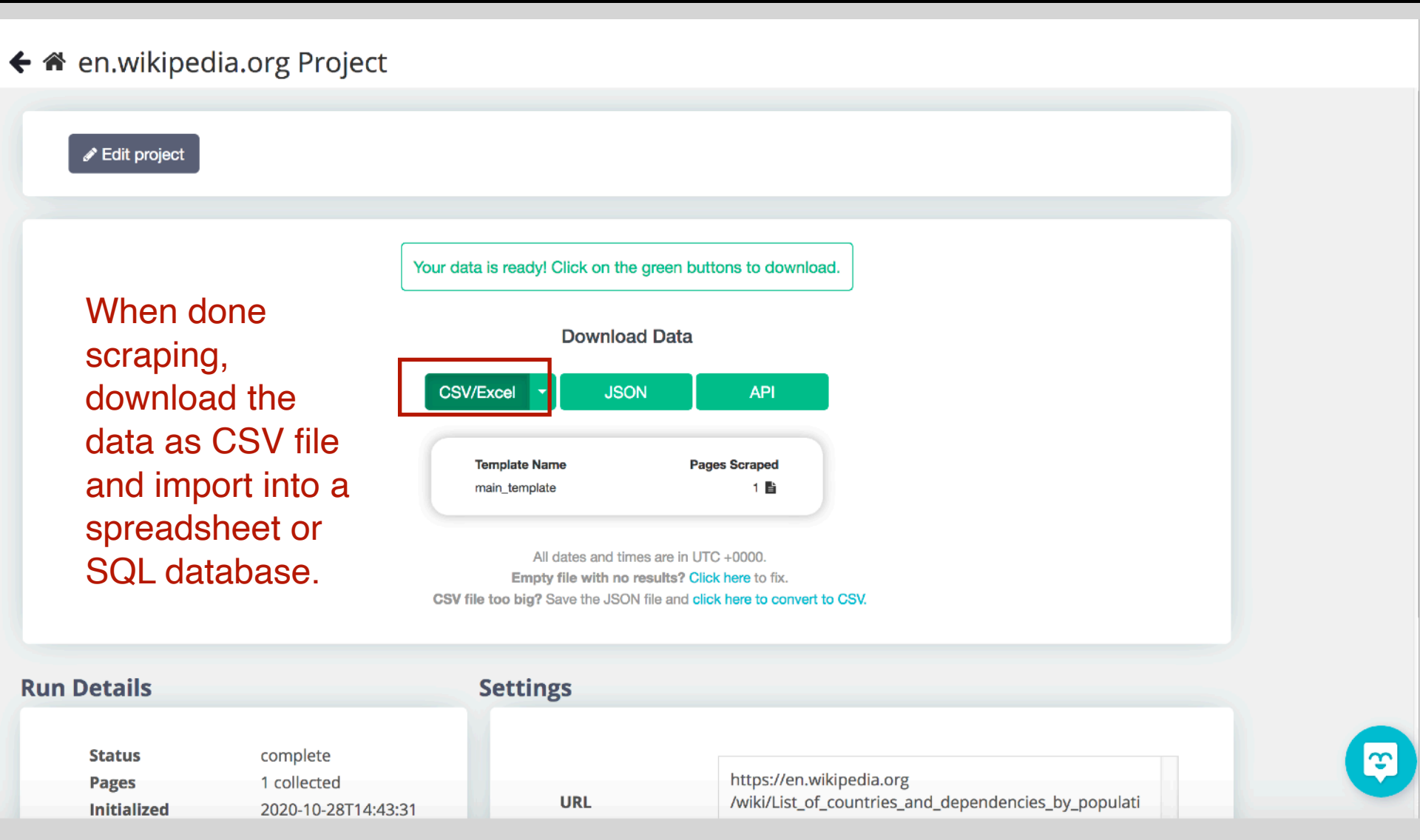
Empty file with no results? Click here to fix.

CSV file too big? Save the JSON file and click here to convert to CSV.

Run Details

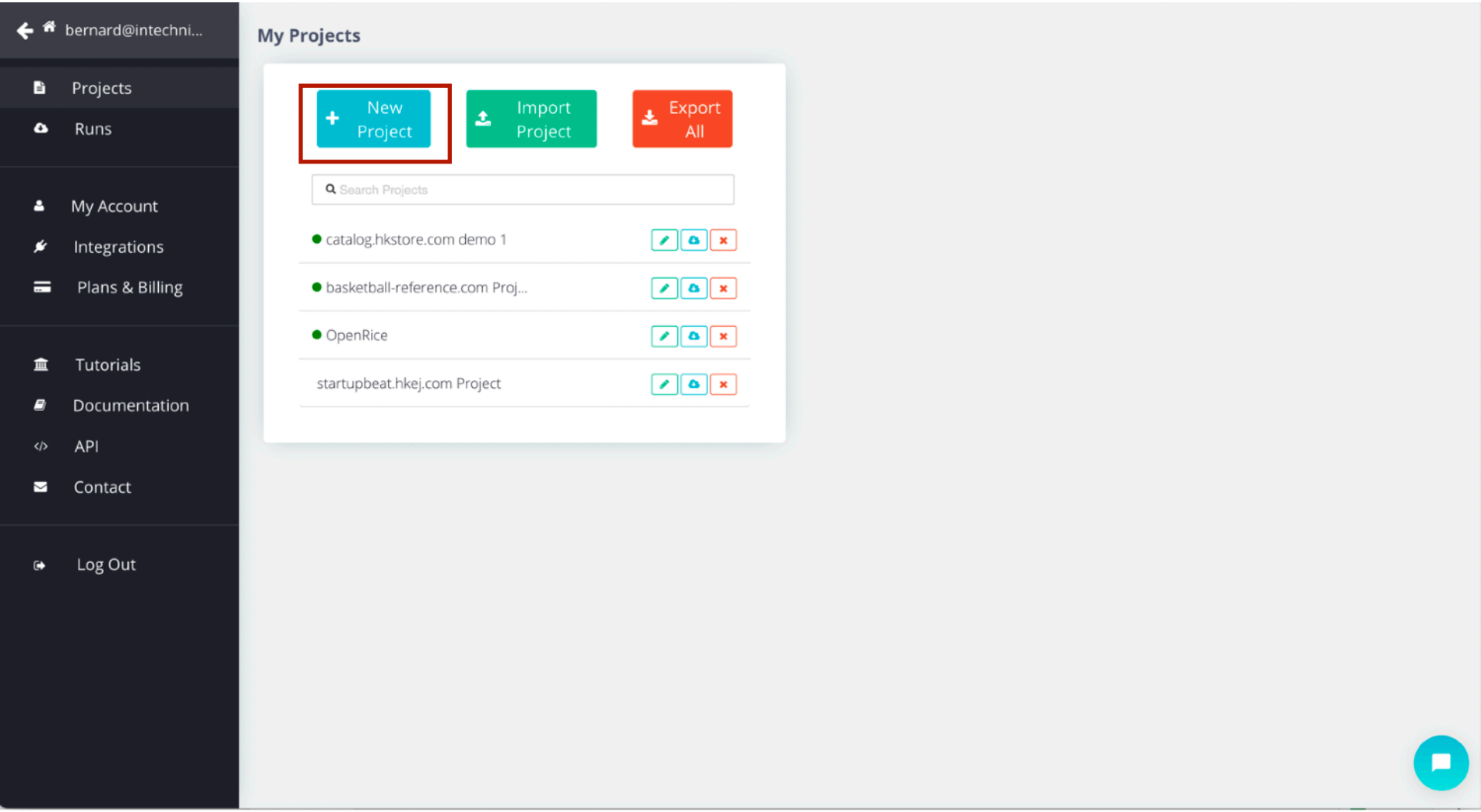
Settings

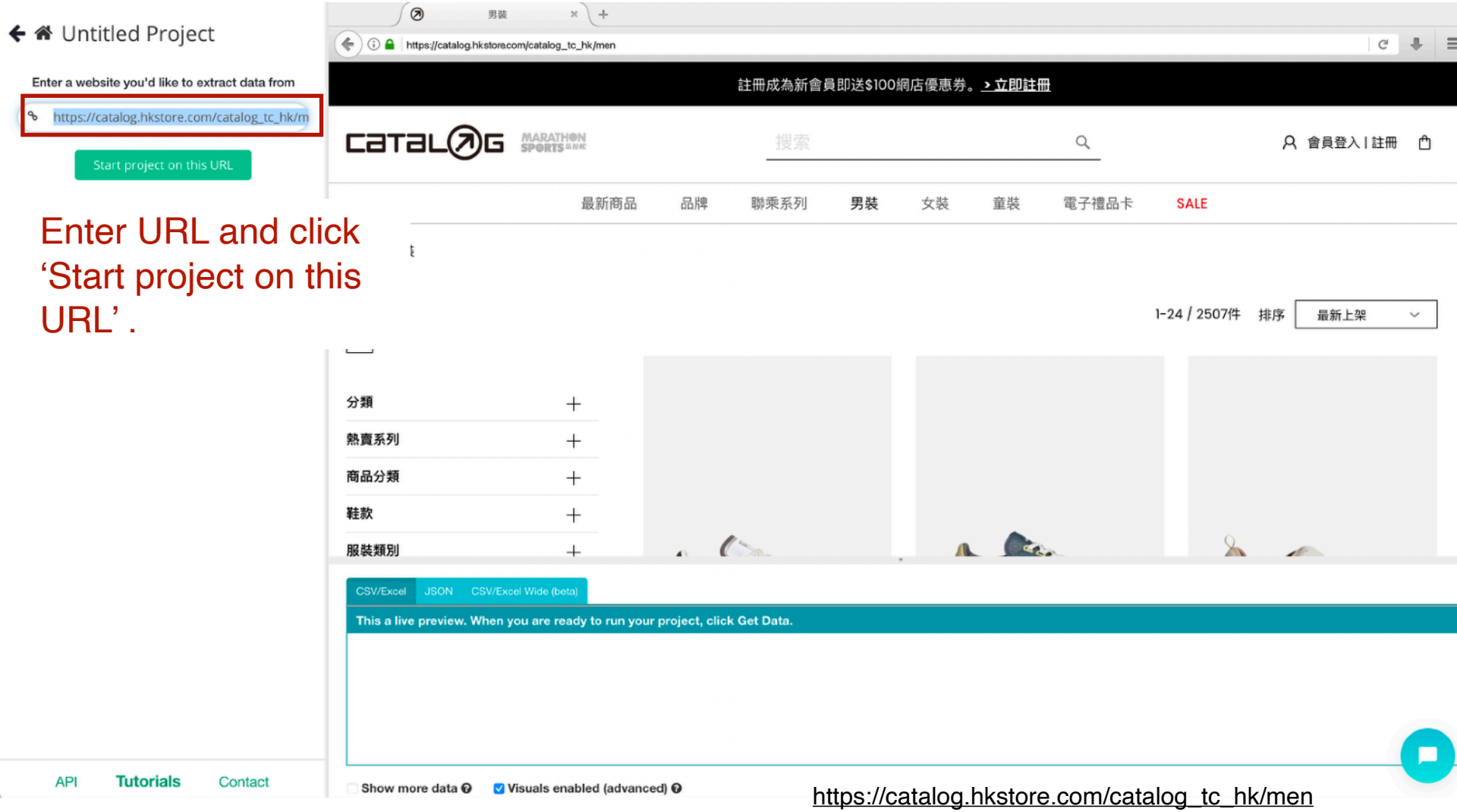
Started job


en.wikipedia.org Project
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
Pages Scraped
main_template
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
Settings
Status
complete
Pages
1 collected
Initialized
2020-10-28T14:43:31
URL
https://en.wikipedia.org
/wiki/List_of_countries_and_dependencies_by_populati

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

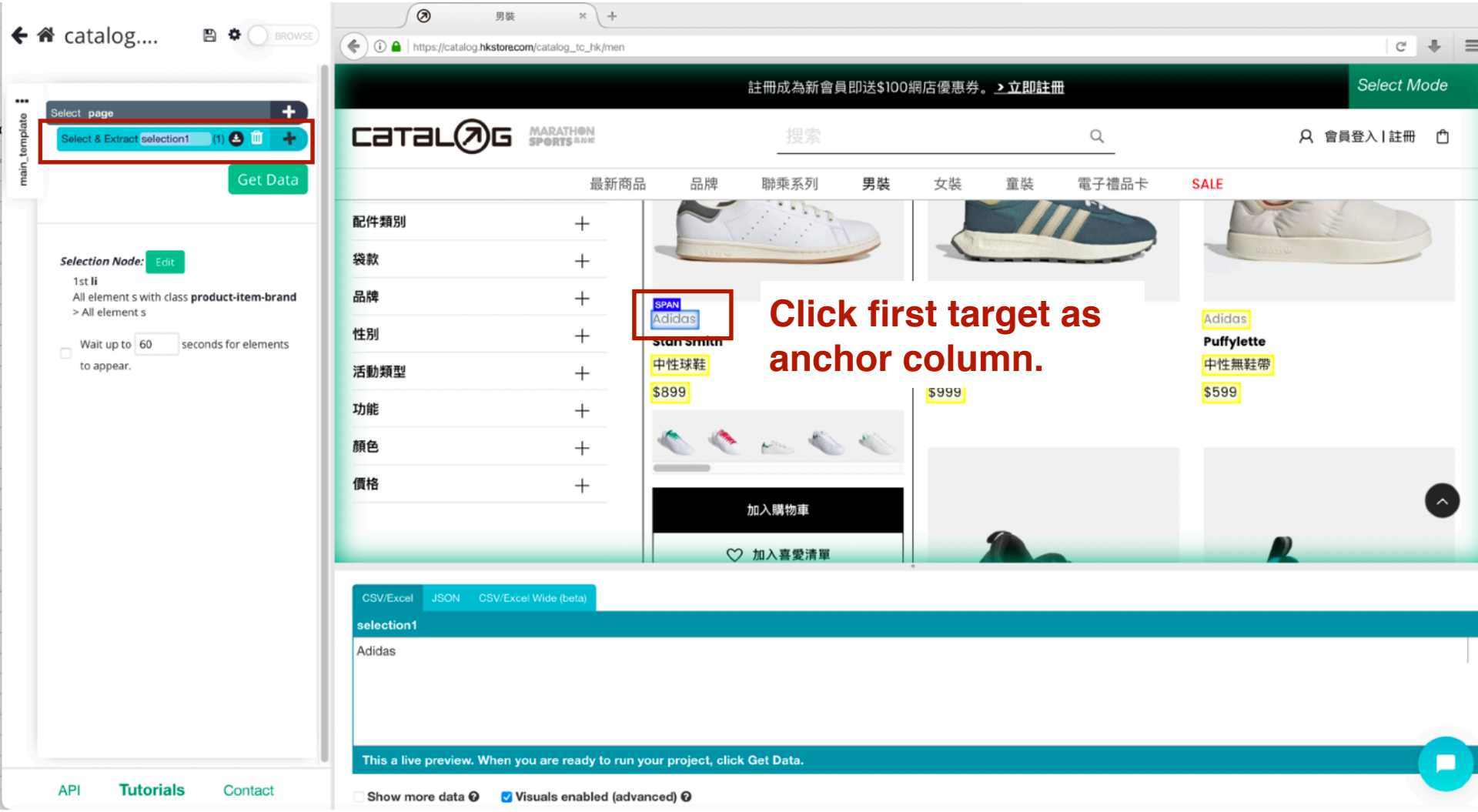
Scraping multiple pages.

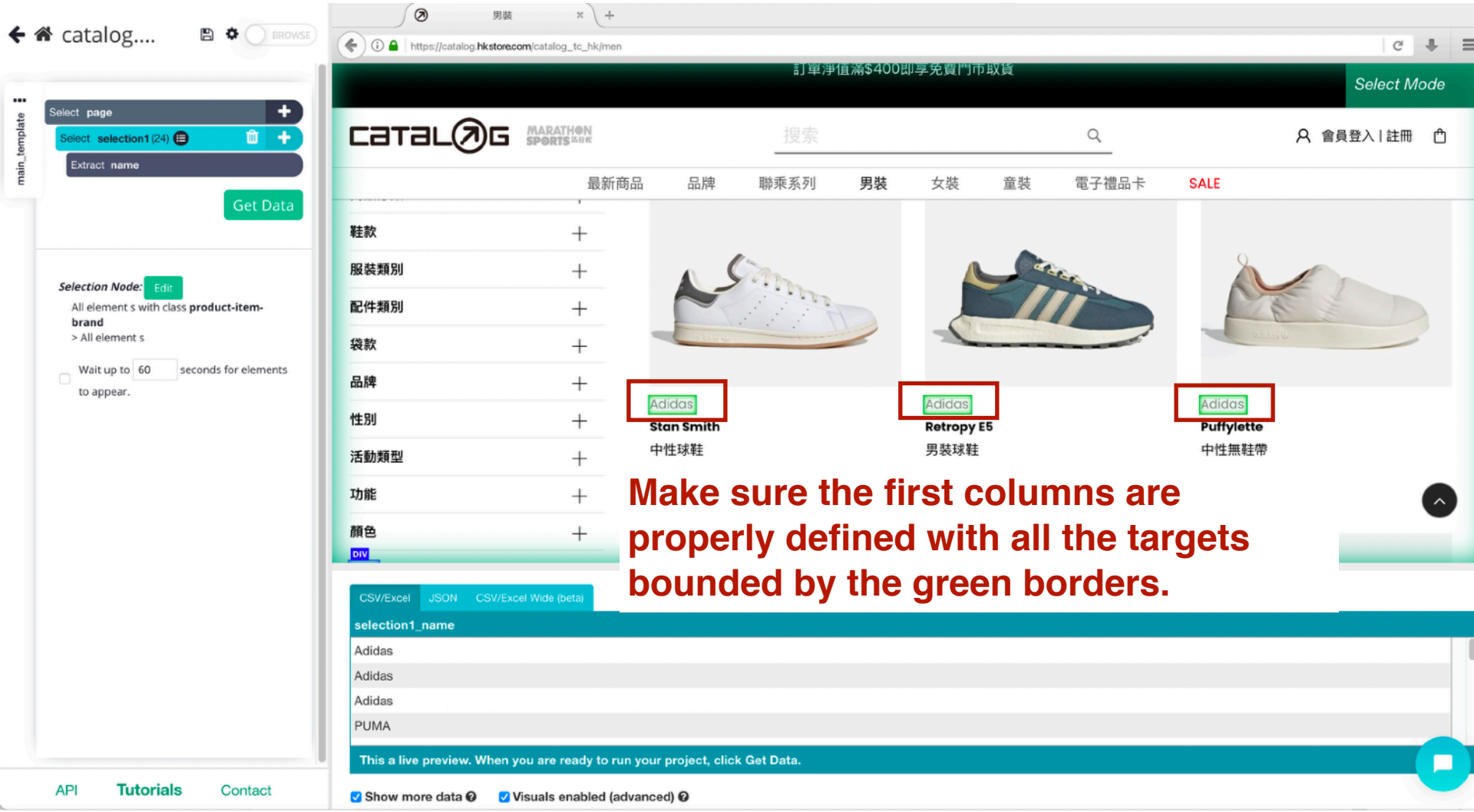

bernard@intechni...
Projects
Runs
My Account
Integrations
Plans & Billing
Tutorials
Documentation
API
Contact
Log Out
My Projects
New Project
Import Project
Export All
Search Projects
catalog.hkstore.com demo 1
basketball-reference.com Proj...
OpenRice
startupbeat.hkej.com Project

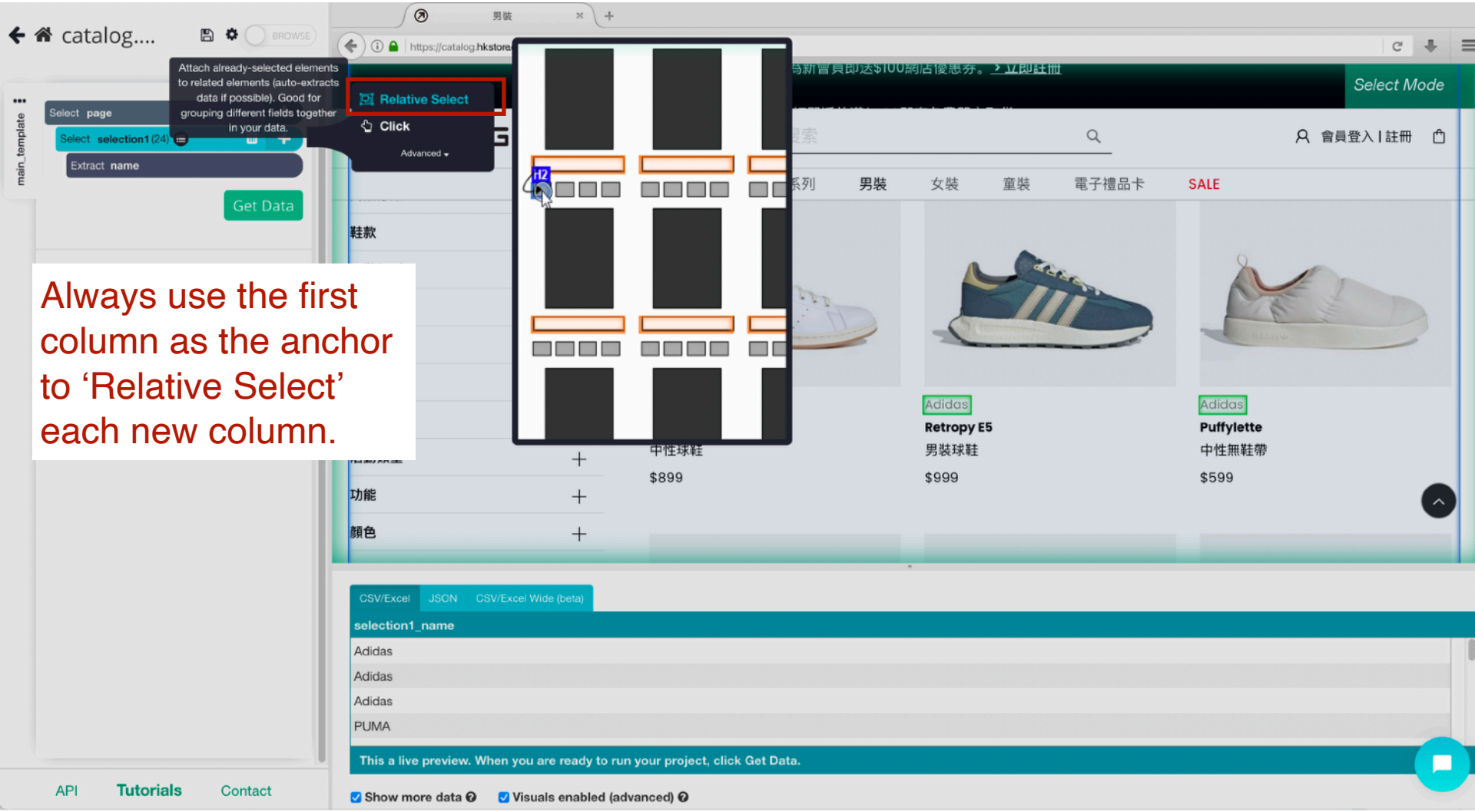


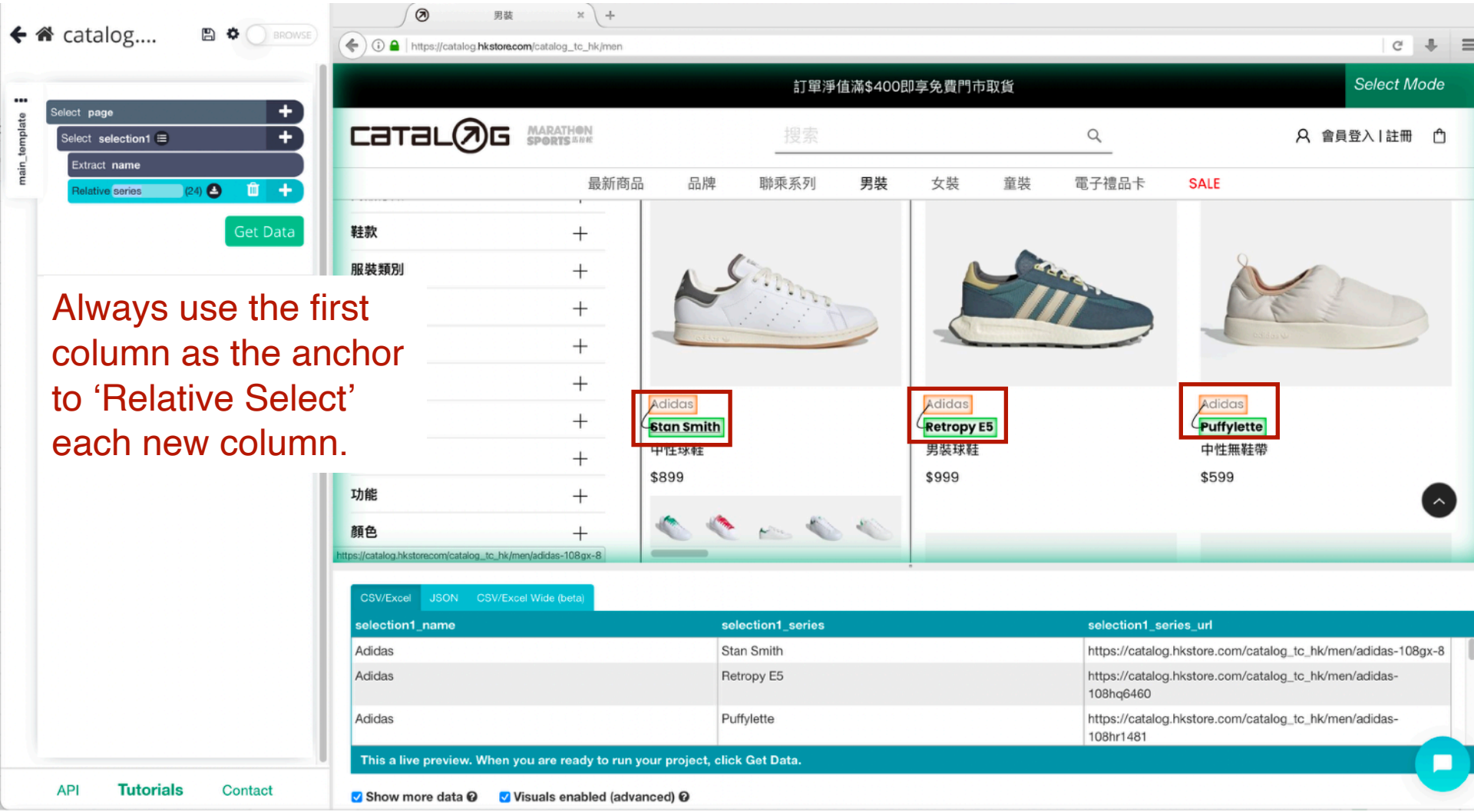
Enter URL and click ‘Start project on this URL’ .

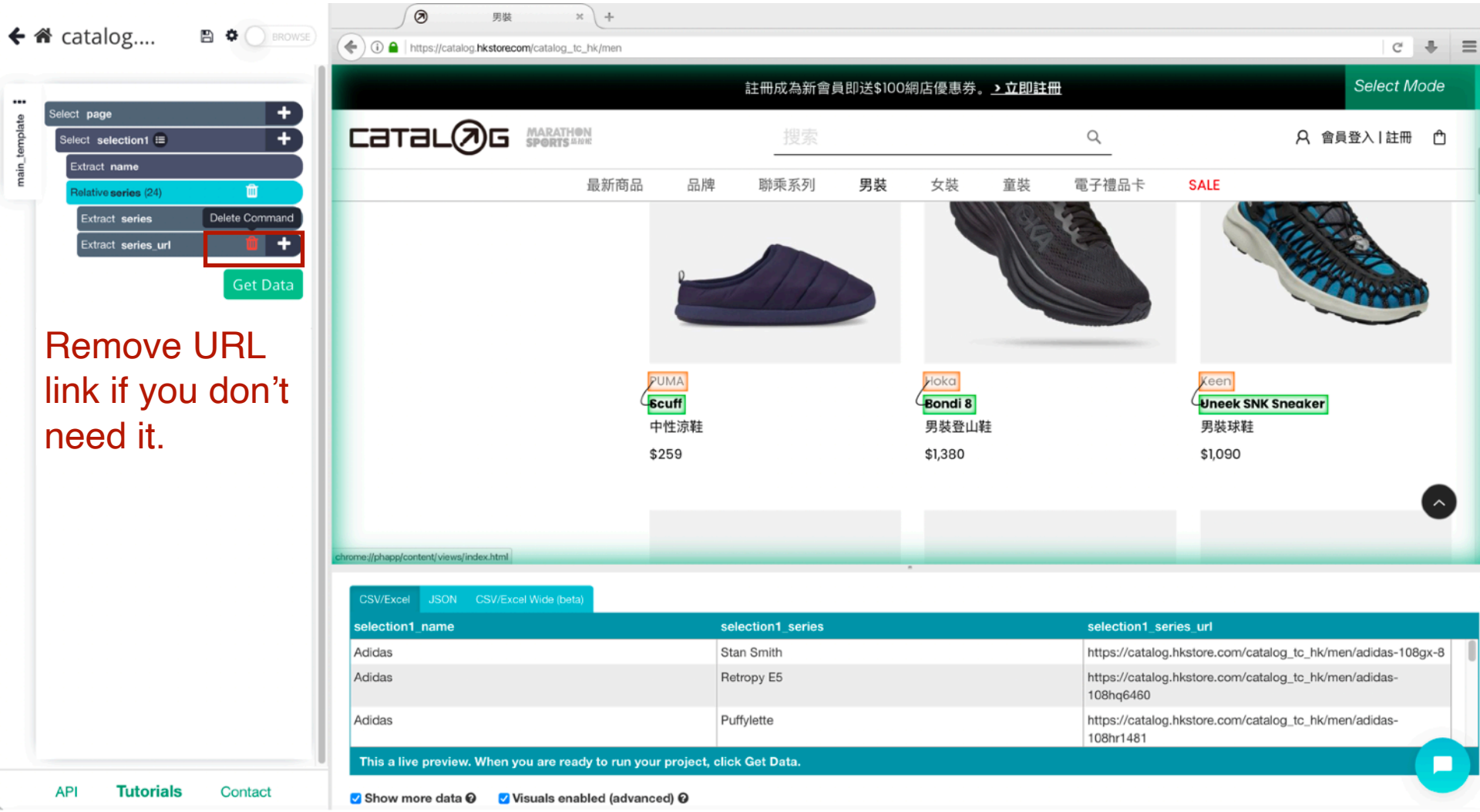
https://catalog.hkstore.com/catalog_tc_hk/men

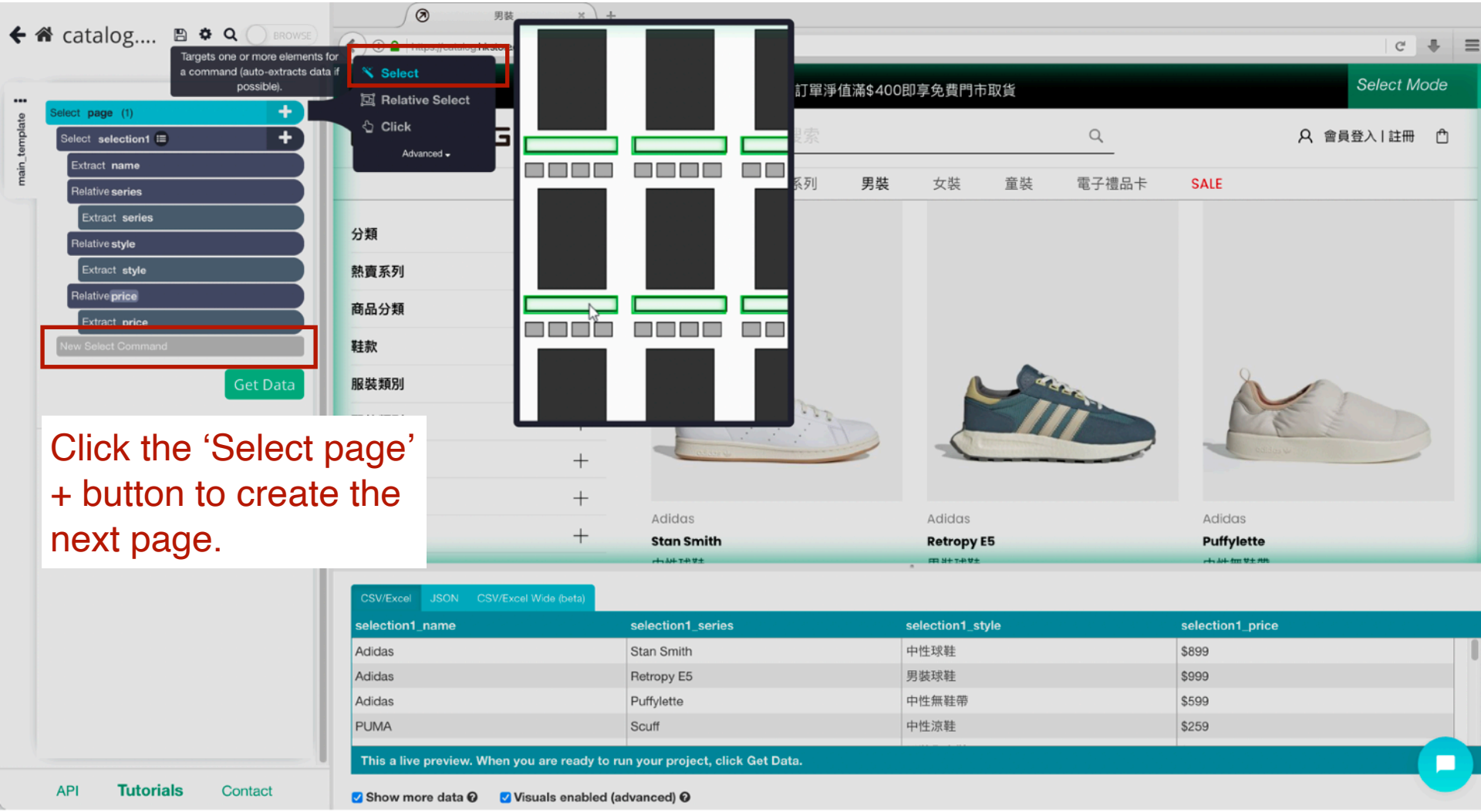

catalog....
BROWSE
main_template
Select page
Select & Extract selection1 (1)
Get Data
Selection Node: Edit
1st li
All element s with class product-item-brand
> All element s
Wait up to 60 seconds for elements to appear.
API
Tutorials
Contact
男裝
https://catalog.hkstore.com/catalog_tc_hk/men
註冊成為新會員即送$100網店優惠券。› 立即註冊
Select Mode
CATALOG
MARATHON SPORTS
搜索
會員登入 | 註冊
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
$999
Adidas
Puffylette
中性無鞋帶
$599
加入購物車
加入喜愛清單
Click first target as anchor column.
CSV/Excel
JSON
CSV/Excel Wide (beta)
selection1
Adidas
This a live preview. When you are ready to run your project, click Get Data.
Show more data
Visuals enabled (advanced)

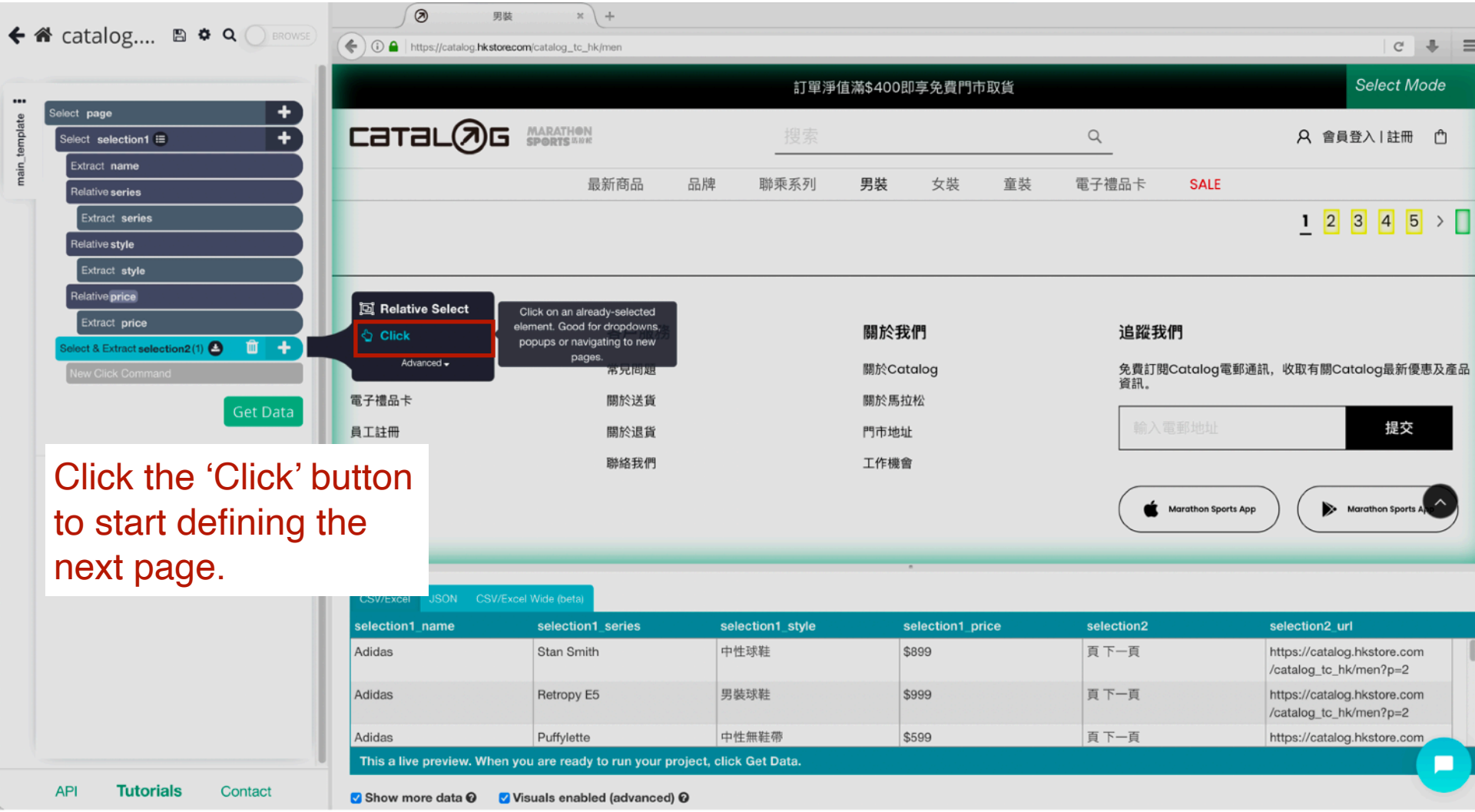

catalog....
BROWSE
main_template
Select page
Select selection1 (24)
Extract name
Get Data
Selection Node: Edit
All element s with class product-item-brand
> All element s
Wait up to 60 seconds for elements to appear.
API
Tutorials
Contact
https://catalog.hkstore.com/catalog_tc_hk/men
訂單淨值滿$400即享免費門市取貨
Select Mode
CATALOG
MARATHON SPORTS
搜索
會員登入 | 註冊
最新商品
品牌
聯乘系列
男裝
女裝
童裝
電子禮品卡
SALE
鞋款
服裝類別
配件類別
袋款
品牌
性別
活動類型
功能
顏色
DIV
Adidas
Stan Smith
中性球鞋
Adidas
Retropy E5
男裝球鞋
Adidas
Puffylette
中性無鞋帶
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)

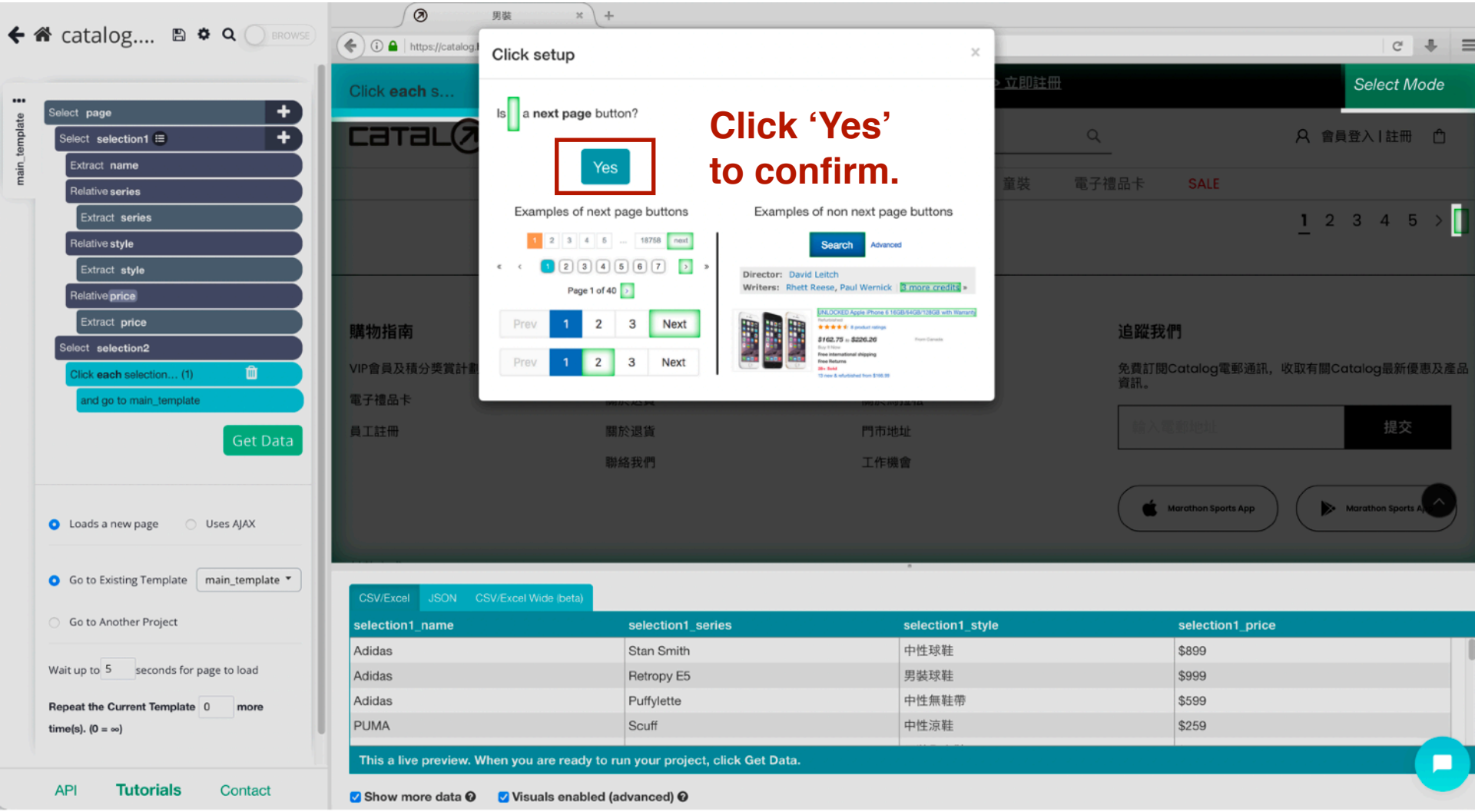

catalog....
BROWSE
Attach already-selected elements to related elements (auto-extracts data if possible). Good for grouping different fields together in your data.
Relative Select
Click
Advanced
Select page
Select selection1 (24)
Extract name
Get Data
main_template
Always use the first column as the anchor to 'Relative Select' each new column.
h2
Select Mode
會員登入 | 註冊
男裝
女裝
童裝
電子禮品卡
SALE
鞋款
功能
顏色
Adidas
Retropy E5
男裝球鞋
$999
Adidas
Puffylette
中性無鞋帶
$599
中性球鞋
$899
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)
API
Tutorials
Contact

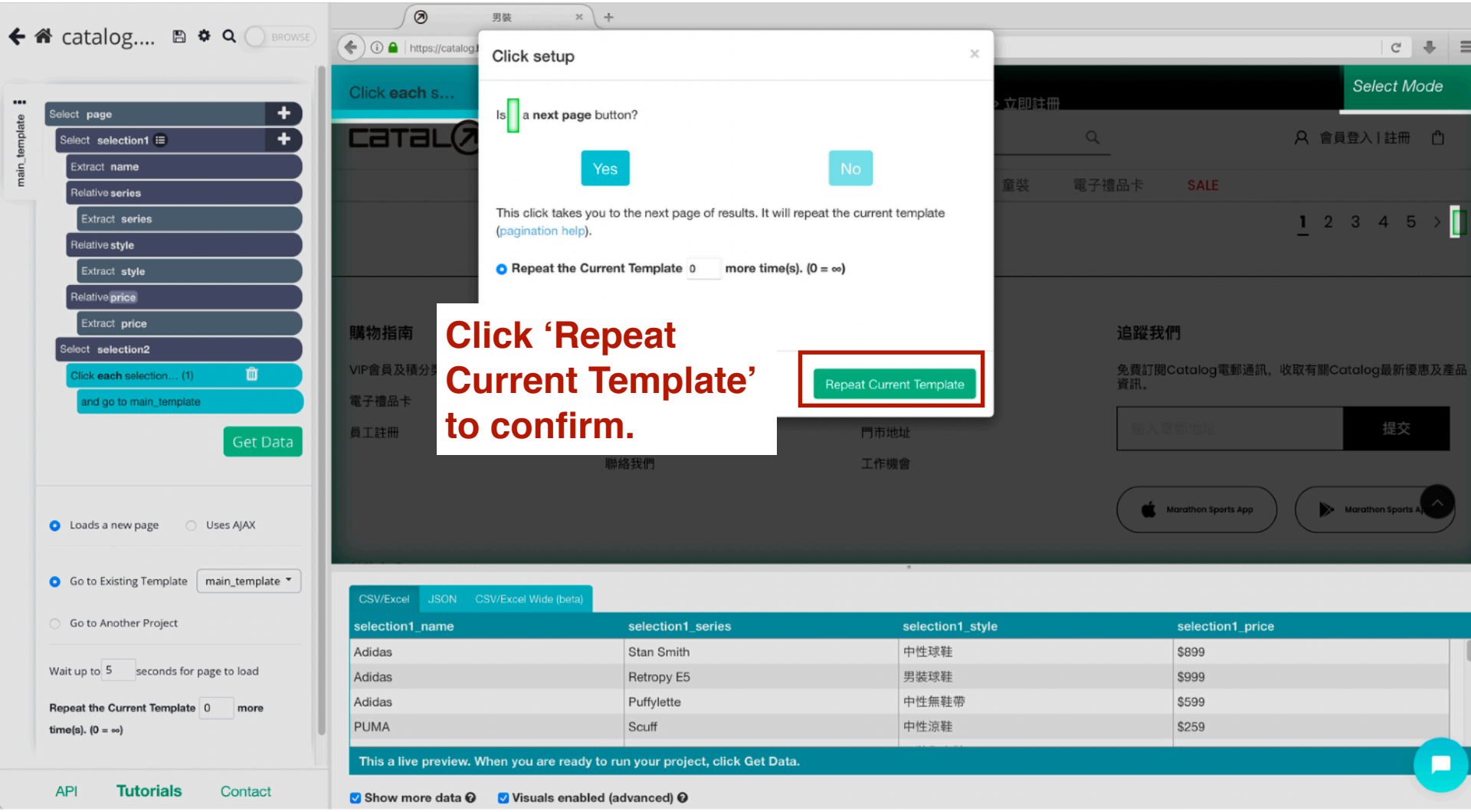

catalog....
BROWSE
main_template
Select page
Select selection1
Extract name
Relative series (24)
Get Data
Always use the first column as the anchor to 'Relative Select' each new column.
API
Tutorials
Contact
https://catalog.hkstore.com/catalog_tc_hk/men
訂單淨值滿$400即享免費門市取貨
Select Mode
CATALOG
MARATHON SPORTS
搜索
會員登入 | 註冊
最新商品
品牌
聯乘系列
男裝
女裝
童裝
電子禮品卡
SALE
鞋款
服裝類別
功能
顏色
Adidas
Stan Smith
$899
Adidas
Retropy E5
男裝球鞋
$999
Adidas
Puffylette
中性無鞋帶
$599
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 a live preview. When you are ready to run your project, click Get Data.
Show more data
Visuals enabled (advanced)

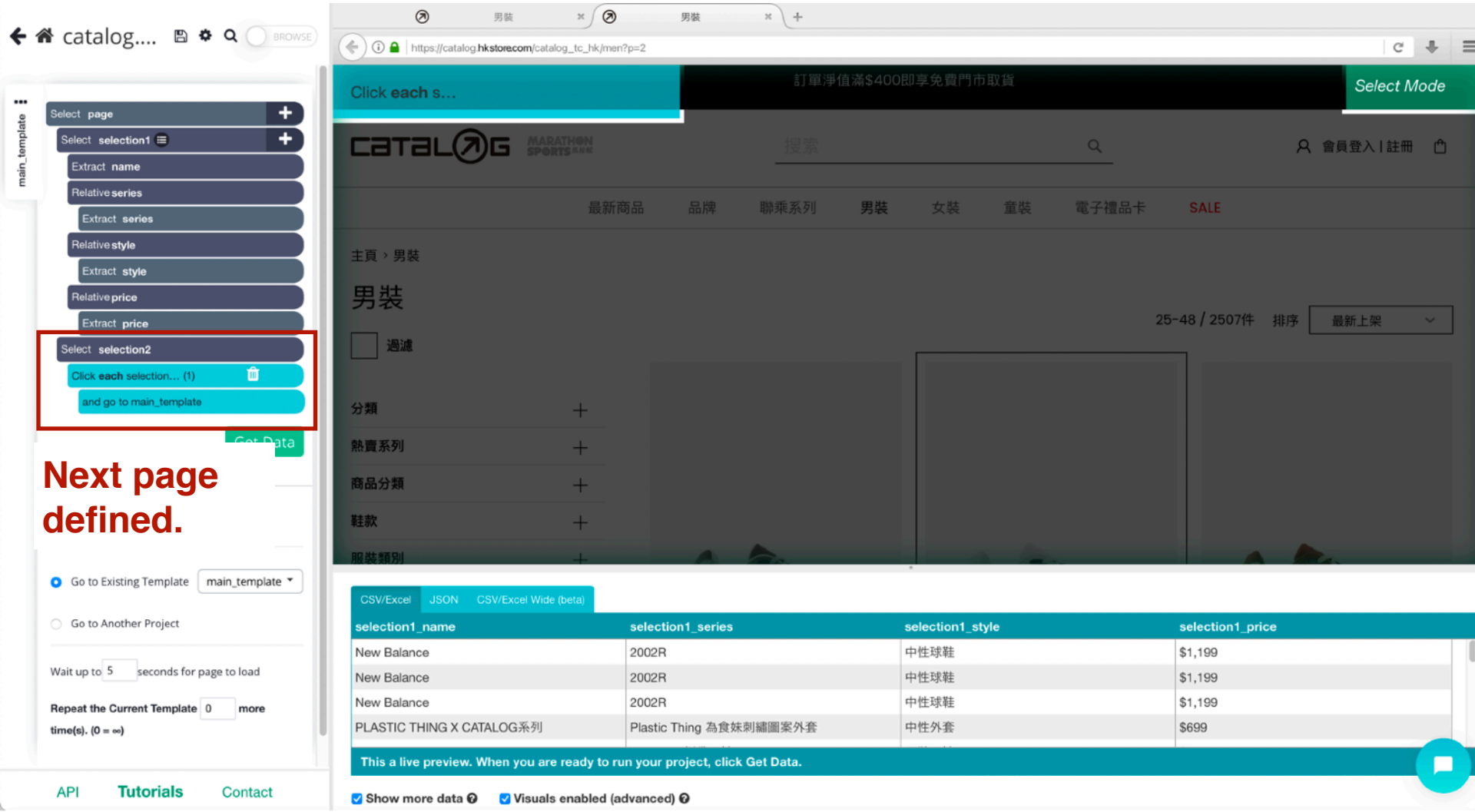

catalog....
BROWSE
main_template
Select page
Select selection1
Extract name
Relative series (24)
Extract series
Delete Command
Extract series_url
Get Data
Remove URL link if you don't need it.
API
Tutorials
Contact
男裝
https://catalog.hkstore.com/catalog_tc_hk/men
註冊成為新會員即送$100網店優惠券。› 立即註冊
Select Mode
CATALOG
MARATHON SPORTS
搜索
會員登入 | 註冊
最新商品
品牌
聯乘系列
男裝
女裝
童裝
電子禮品卡
SALE
PUMA
6cuff
中性涼鞋
$259
Hoka
Bondi 8
男裝登山鞋
$1,380
Keen
Uneek SNK Sneaker
男裝球鞋
$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
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 a live preview. When you are ready to run your project, click Get Data.
Show more data
Visuals enabled (advanced)

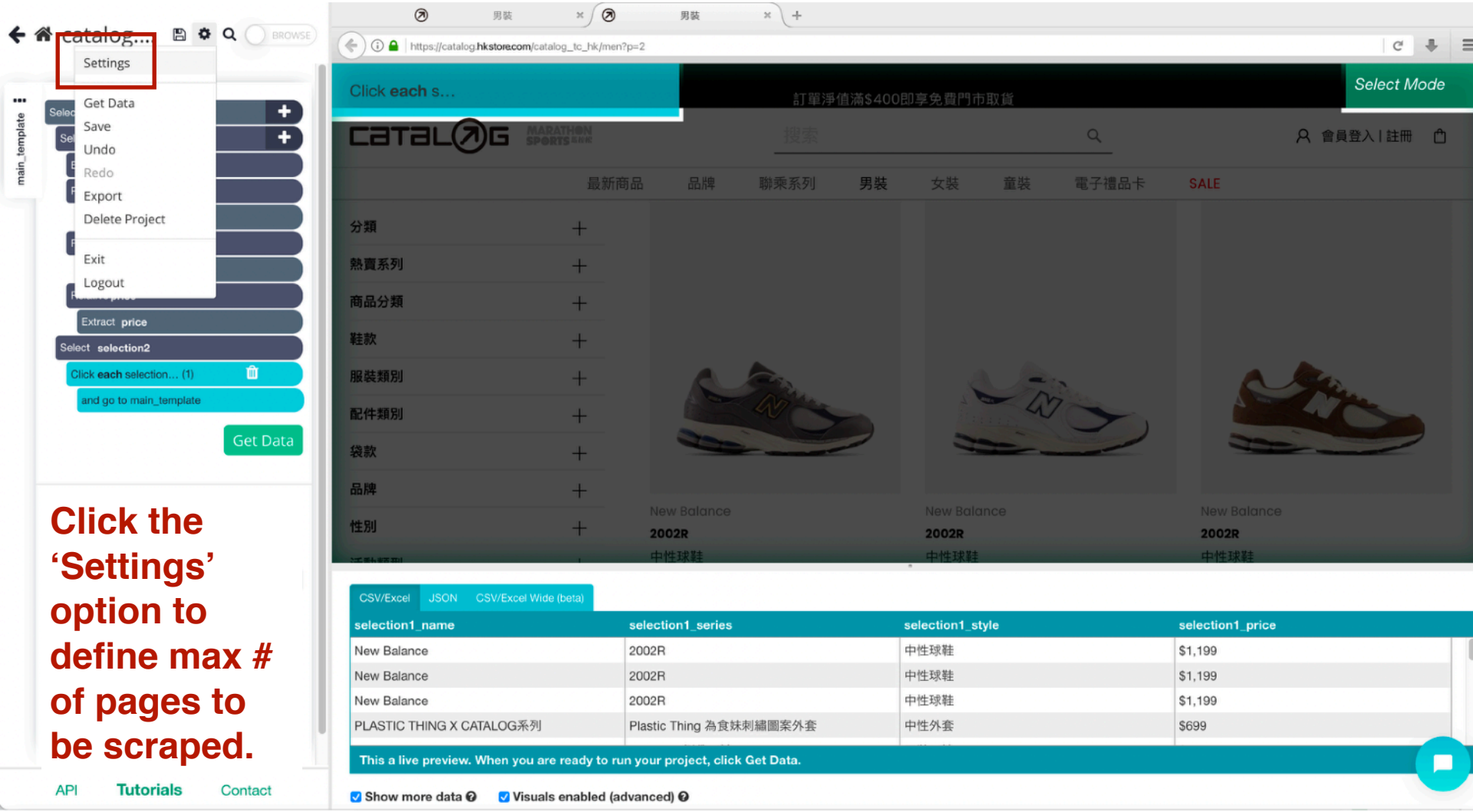

catalog....
Targets one or more elements for a command (auto-extracts data if possible).
Select
Relative Select
Click
Advanced
Select page (1)
Select selection1
Extract name
Relative series
Extract series
Relative style
Extract style
Relative price
Extract price
New Select Command
Get Data
main_template
Click the ‘Select page’ + button to create the next page.
訂單淨值滿$400即享免費門市取貨
Select Mode
會員登入 | 註冊
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
Adidas
Stan Smith
Adidas
Retropy E5
Adidas
Puffylette
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
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.
Click the ‘Click’ button to start defining the next page.
https://catalog.hkstore.com/catalog_tc_hk/men
訂單淨值滿$400即享免費門市取貨
Select Mode
CATALOG
MARATHON SPORTS
搜索
會員登入 | 註冊
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
電子禮品卡
員工註冊
關於送貨
關於退貨
聯絡我們
關於我們
關於Catalog
關於馬拉松
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。
輸入電郵地址
提交
Marathon Sports App
CSV/Excel
JSON
CSV/Excel Wide (beta)
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
This a live preview. When you are ready to run your project, click Get Data.
Show more data
Visuals enabled (advanced)
API
Tutorials
Contact

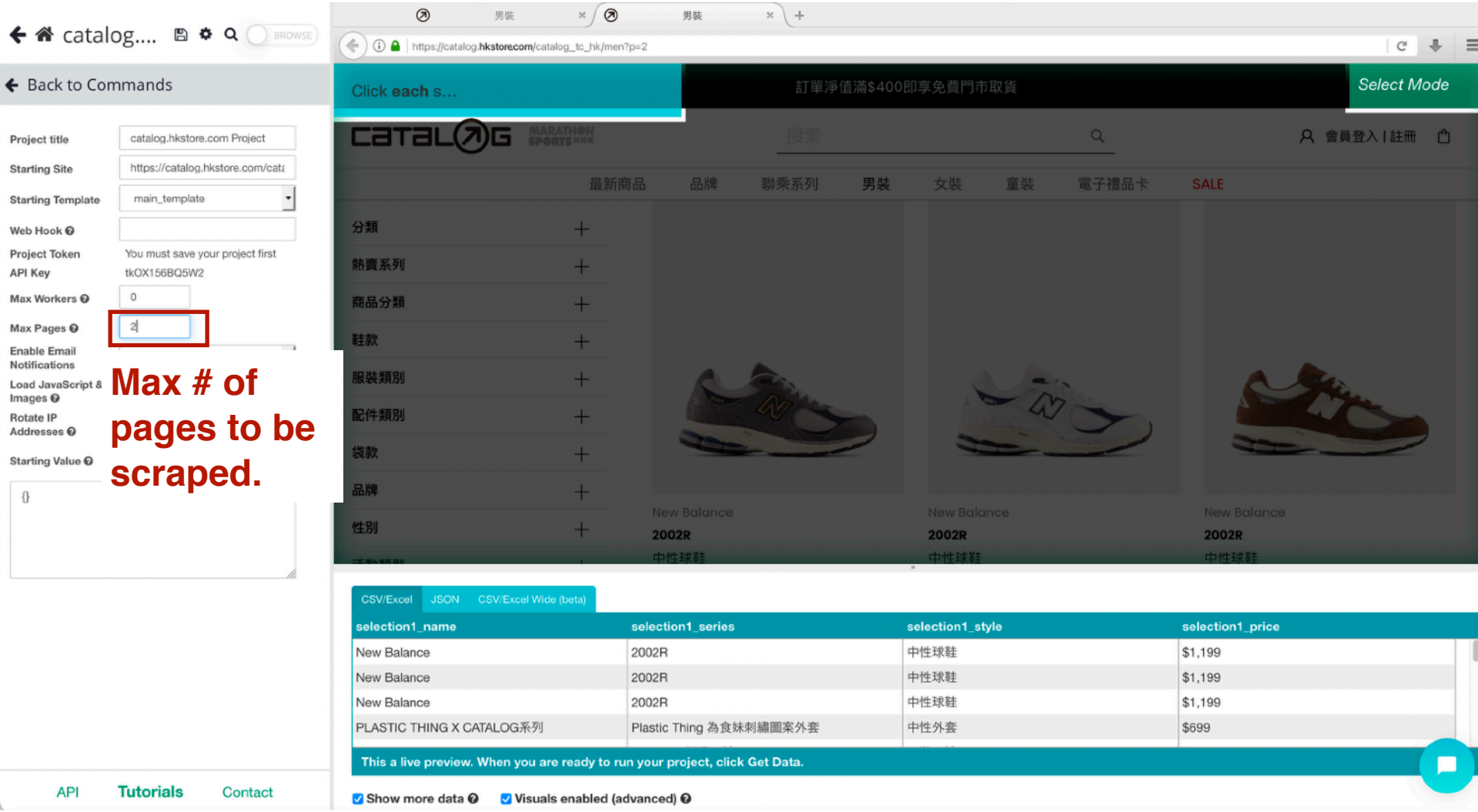

catalog....
BROWSE
main_template
Select page
Select selection1
Extract name
Relative series
Extract series
Relative style
Extract style
Relative price
Extract price
Select selection2
Click each selection... (1)
and go to main_template
Get Data
Loads a new page
Uses AJAX
Go to Existing Template
main_template
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 setup
Is a next page button?
Yes
Click 'Yes' to confirm.
Examples of next page buttons
Examples of non next page buttons
Page 1 of 40
Prev
Next
Search
Advanced
Director: David Leitch
Writers: Rhett Reese, Paul Wernick
3 more credits
Select Mode
購物指南
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。
提交
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)

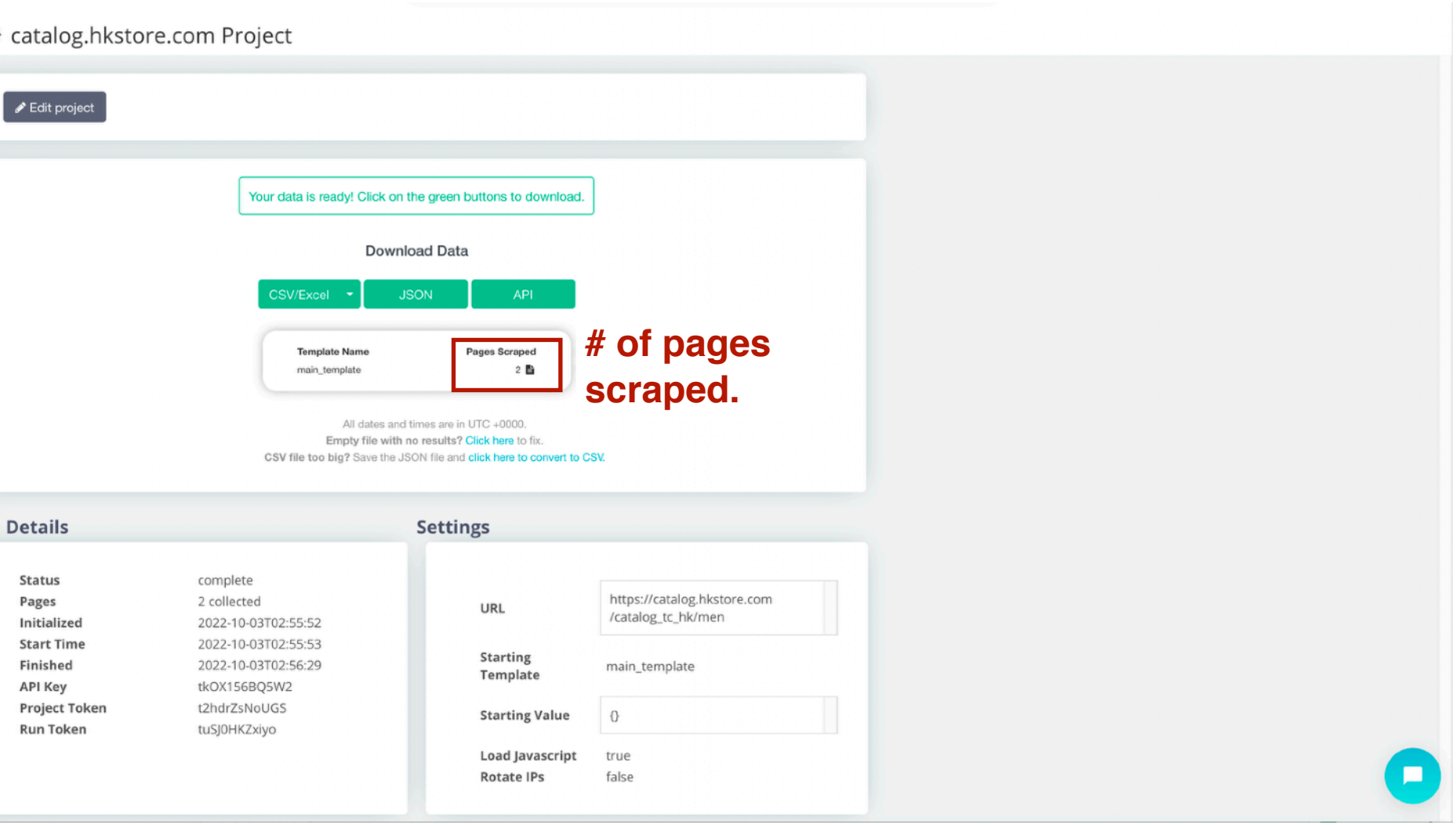

catalog....
BROWSE
main_template
Select page
Select selection1
Extract name
Relative series
Extract series
Relative style
Extract style
Relative price
Extract price
Select selection2
Click each selection... (1)
and go to main_template
Get Data
Loads a new page
Uses AJAX
Go to Existing Template
main_template
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 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 = ∞)
Repeat Current Template
Click 'Repeat Current Template' to confirm.
Select Mode
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)


catalog....
BROWSE
main_template
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
https://catalog.hkstore.com/catalog_tc_hk/men?p=2
Click each s...
訂單淨值滿$400即享免費門市取貨
Select Mode
CATALOG
搜索
會員登入 | 註冊
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)

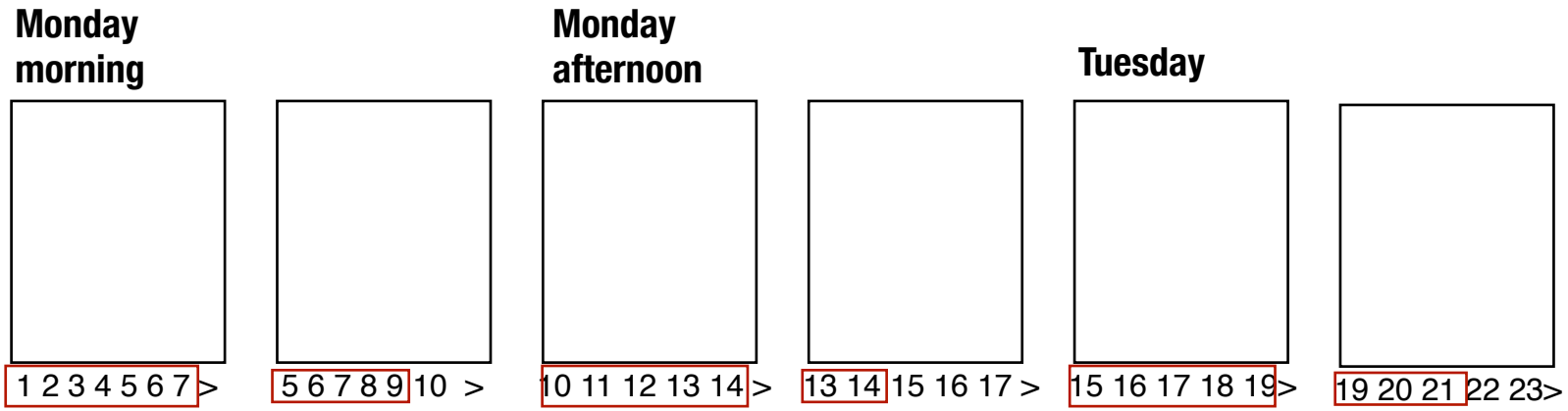


Click the ‘Settings’ option to define max # of pages to be scraped.


catalog....
BROWSE
Back to Commands
Project title
catalog.hkstore.com Project
Starting Site
https://catalog.hkstore.com/cata
Starting Template
main_template
Web Hook
Project Token
You must save your project first
API Key
tkOX156BQ5W2
Max Workers
0
Max Pages
2
Enable Email Notifications
Load JavaScript & Images
Rotate IP Addresses
Starting Value
{}
Max # of pages to be scraped.
API
Tutorials
Contact
https://catalog.hkstore.com/catalog_tc_hk/men?p=2
Click each s...
訂單淨值滿$400即享免費門市取貨
Select Mode
CATALOG
搜索
會員登入 | 註冊
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
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 Scraped
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 https://catalog.hkstore.com /catalog_tc_hk/men
Starting Template main_template
Starting Value {}
Load Javascript true
Rotate IPs false

Periodic scraping in smaller batches.


Monday
morning
Monday
afternoon
Tuesday
1 2 3 4 5 6 7 >
5 6 7 8 9 10 >
10 11 12 13 14 >
13 14 15 16 17 >
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.

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

WRITTEN BY
James Densmore
Data Science and Data Engineering Consultant at Data Liftoff
https://www.datalifto ff.com

Follow

The Ethical Scraper

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

- 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 it 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

OpenRefine

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

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.

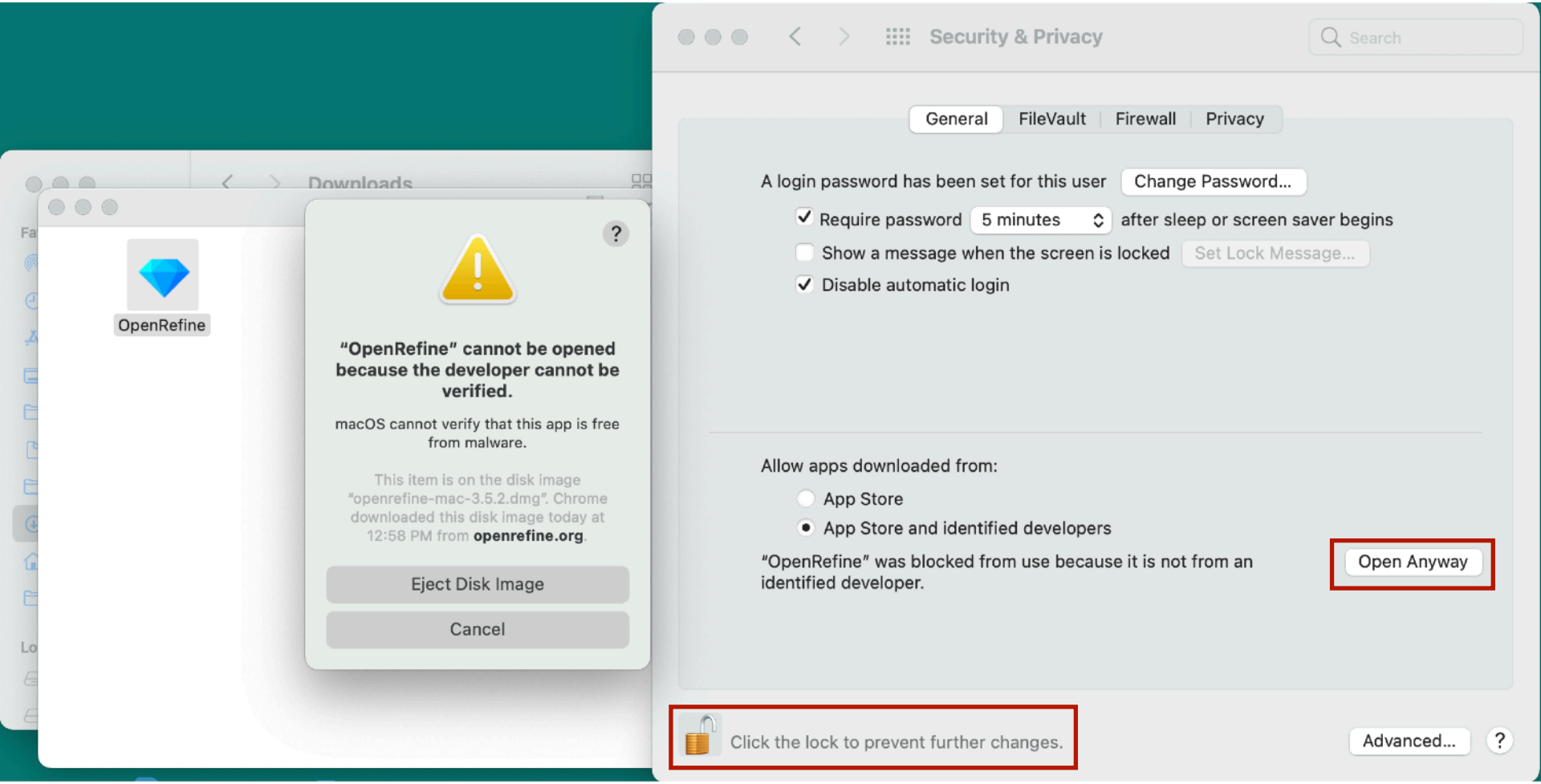
- **Windows kit**, This requires Java to be installed on your computer. Download, unzip, and double-click on *openrefine.exe* or *refine.bat* if the former does not work.
- **Windows kit with embedded Java**, includes OpenJDK Java, available under the GPLv2+CE license. Download, unzip, and double-click 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


Downloads
OpenRefine
"OpenRefine" cannot be opened because the developer cannot be verified.
macOS cannot verify that this app is free from malware.
This item is on the disk image "openrefine-mac-3.5.2.dmg". Chrome downloaded this disk image today at 12:58 PM from openrefine.org.
Eject Disk Image
Cancel
Security & Privacy
Search
General
FileVault
Firewall
Privacy
A login password has been set for this user
Change Password...
Require password
5 minutes
after sleep or screen saver begins
Show a message when the screen is locked
Set Lock Message...
Disable automatic login
Allow apps downloaded from:
App Store
App Store and identified developers
"OpenRefine" was blocked from use because it is not from an identified developer.
Open Anyway
Click the lock to prevent further changes.
Advanced...

Most Basic Data Cleaning Tasks Part 1

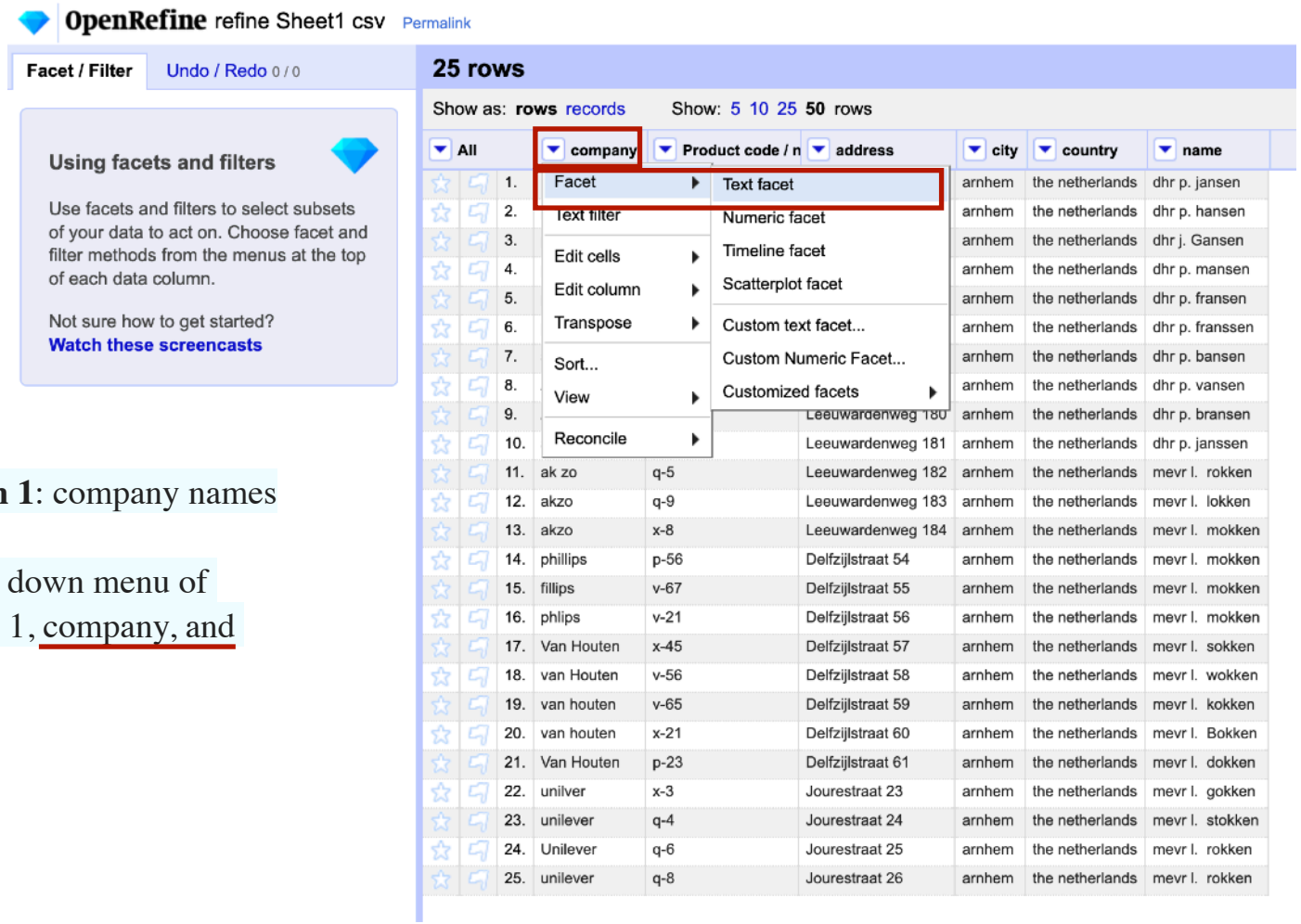
1. Resolve inconsistent entries.
2. Rename, combine and split columns for making further data presentation and processing easier.
3. Transform incorrect code (e.g. wrong data types such as text instead of numeric and calculations)

Tutorial

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

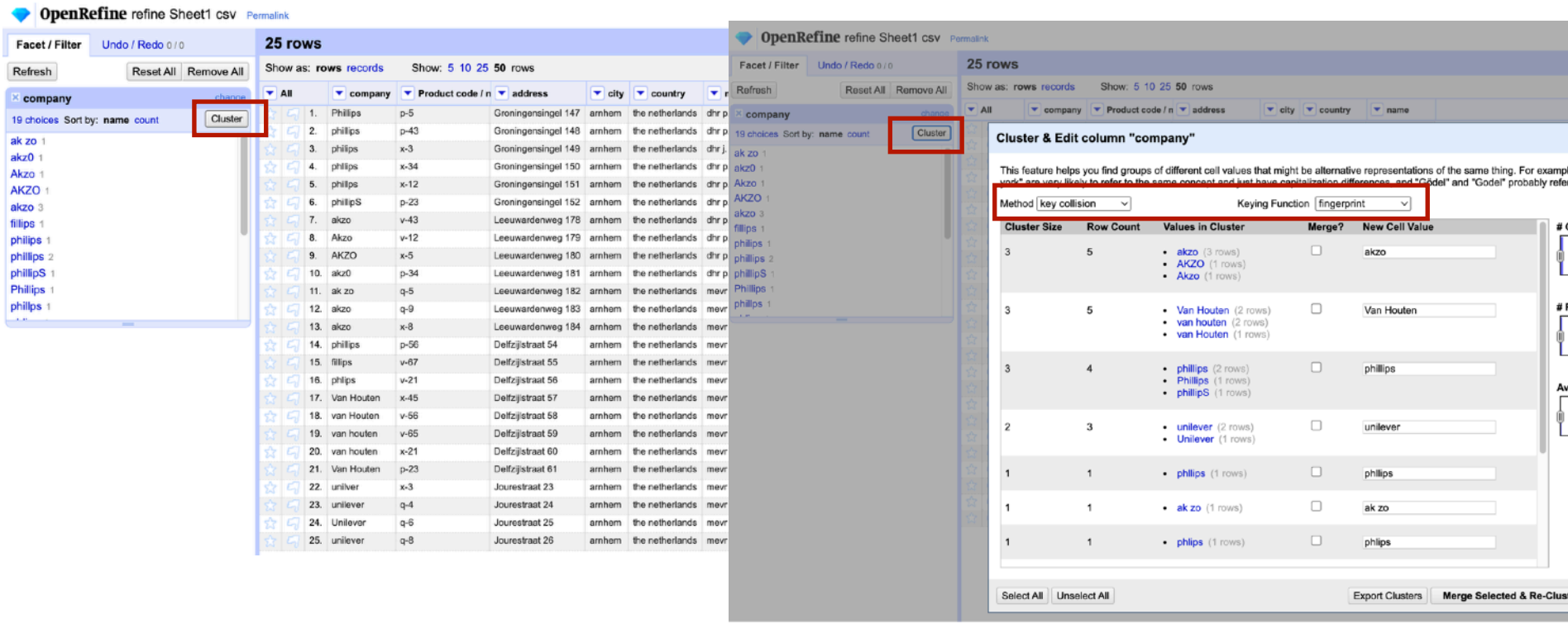
Practice Dataset

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

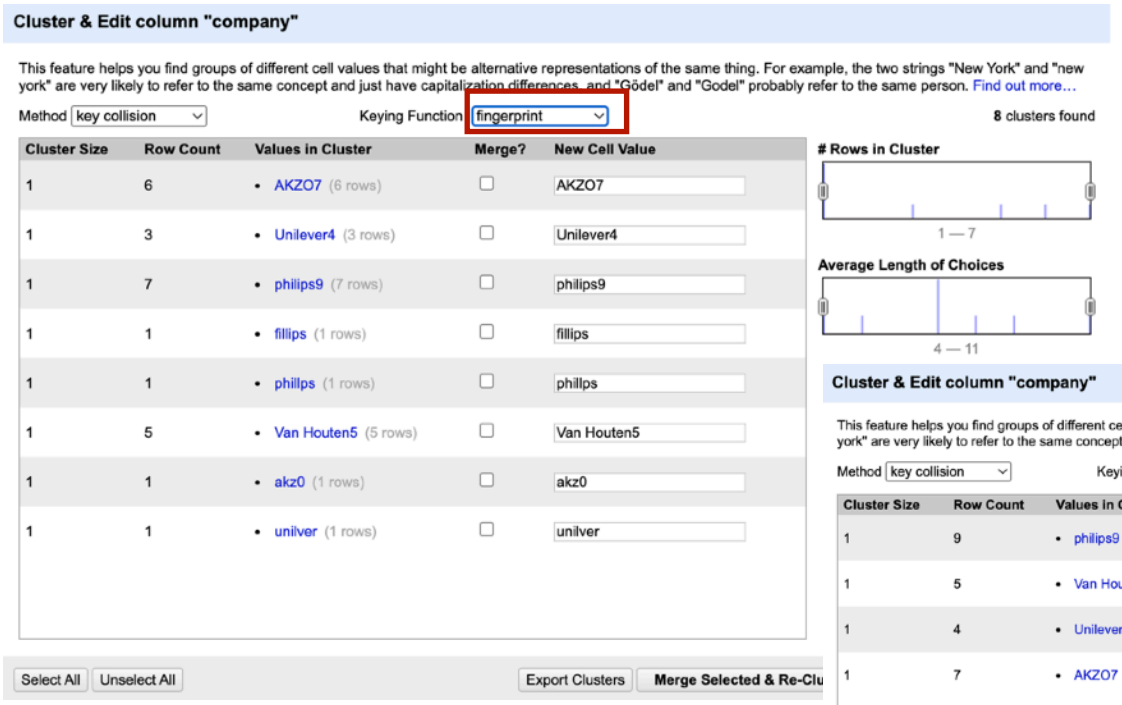


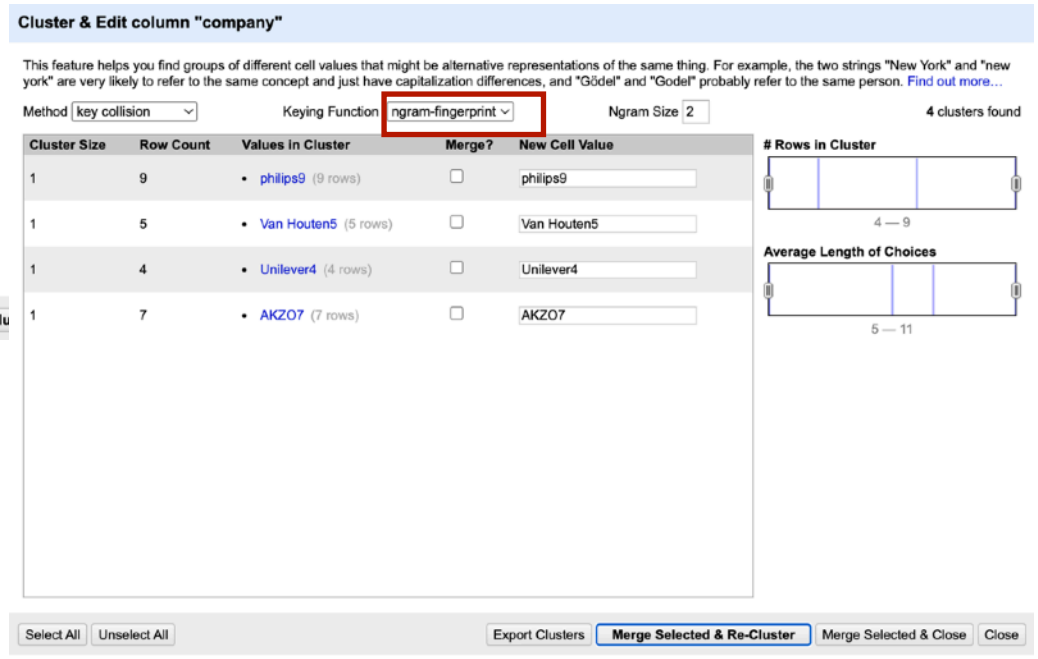
Cleaning column 1: company names

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



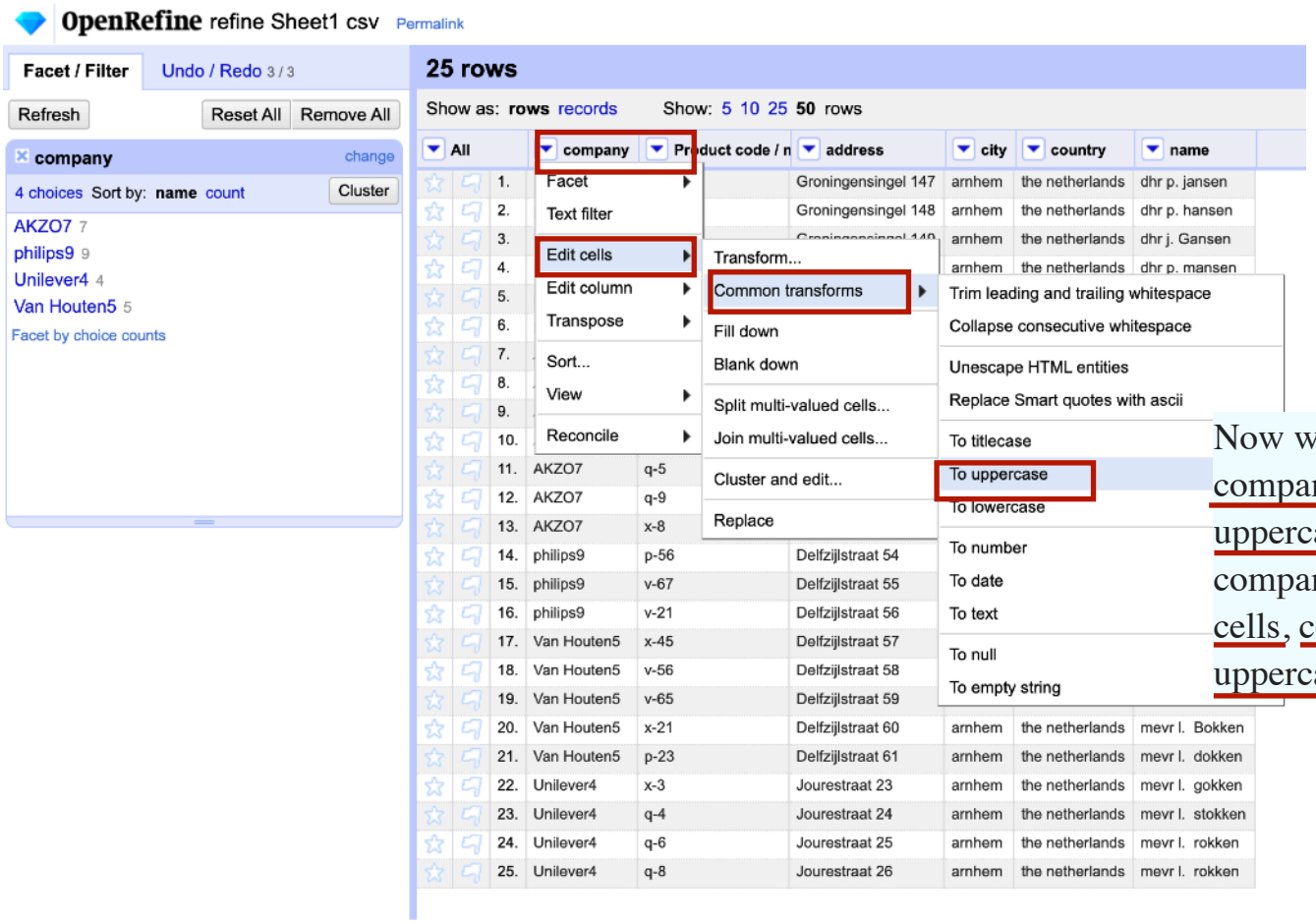
On the left you see the companies listed. Click on cluster and choose 'key collision' and 'finger print' for clustering; that is correcting the names. Choose 'merge selected & re cluster'.



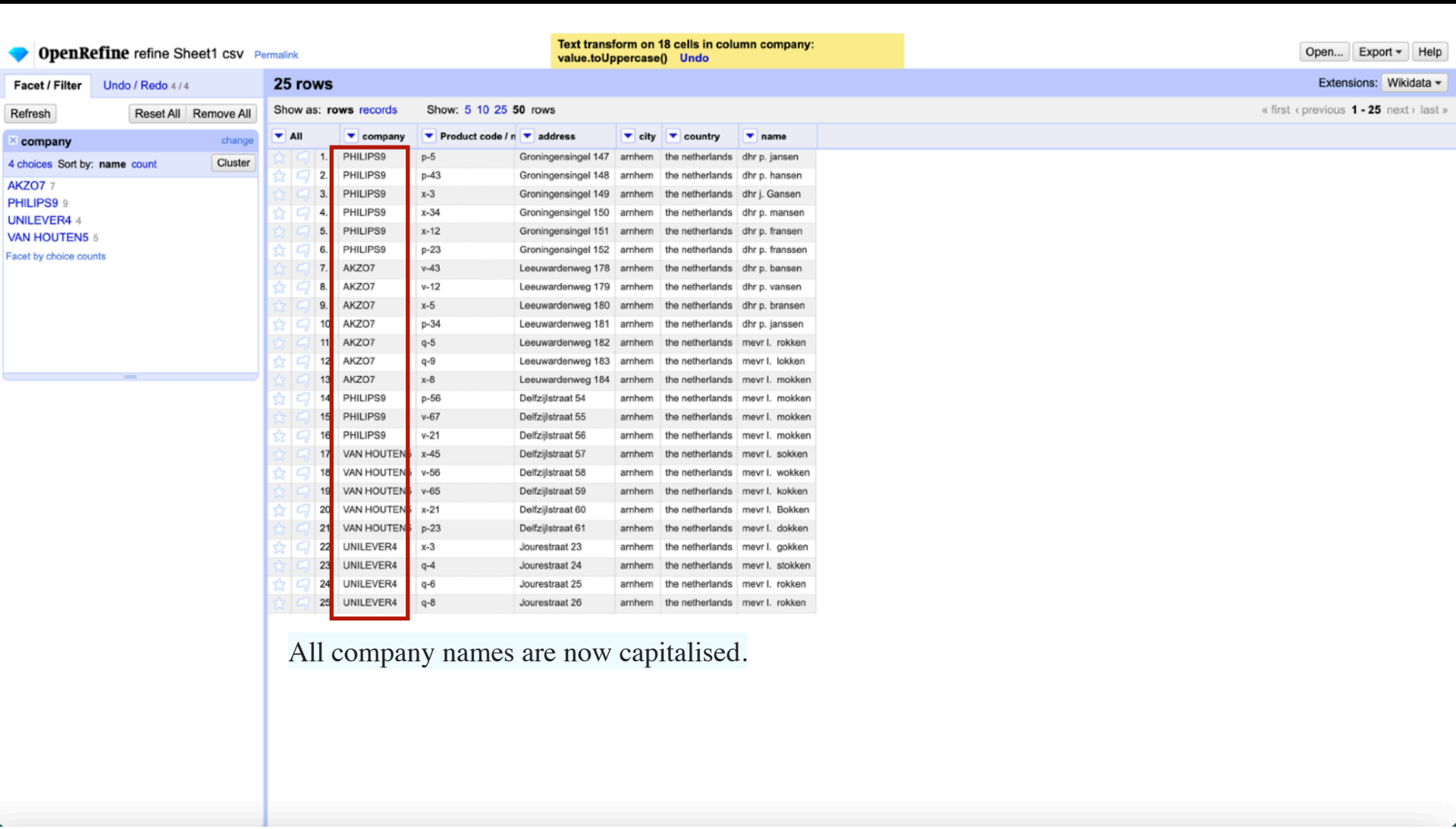


Continue clustering with a new algorithm: ngram-fingerprint. Continue till you have the following names:

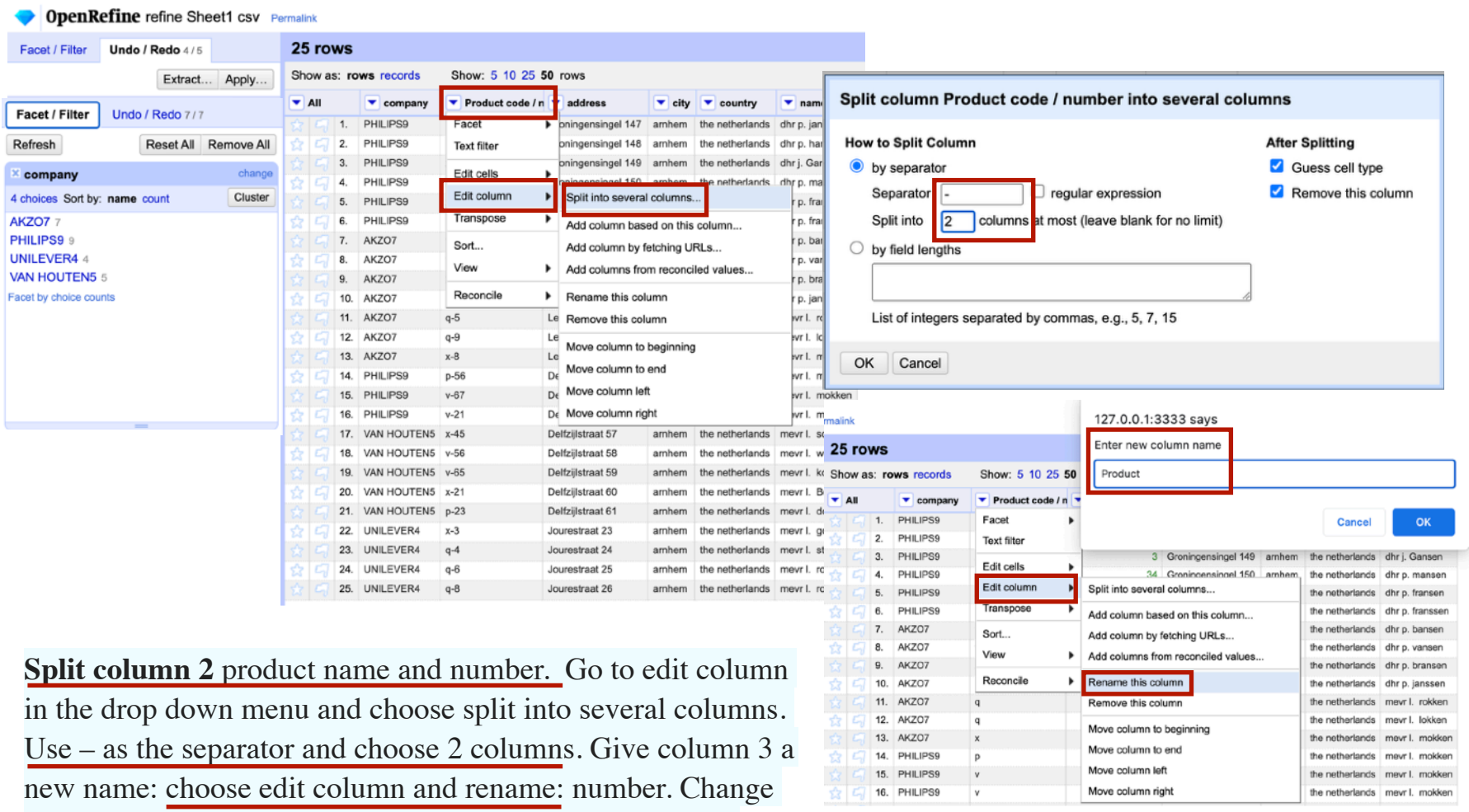
AKZO7
philips9
Unilever4
Van Houten5



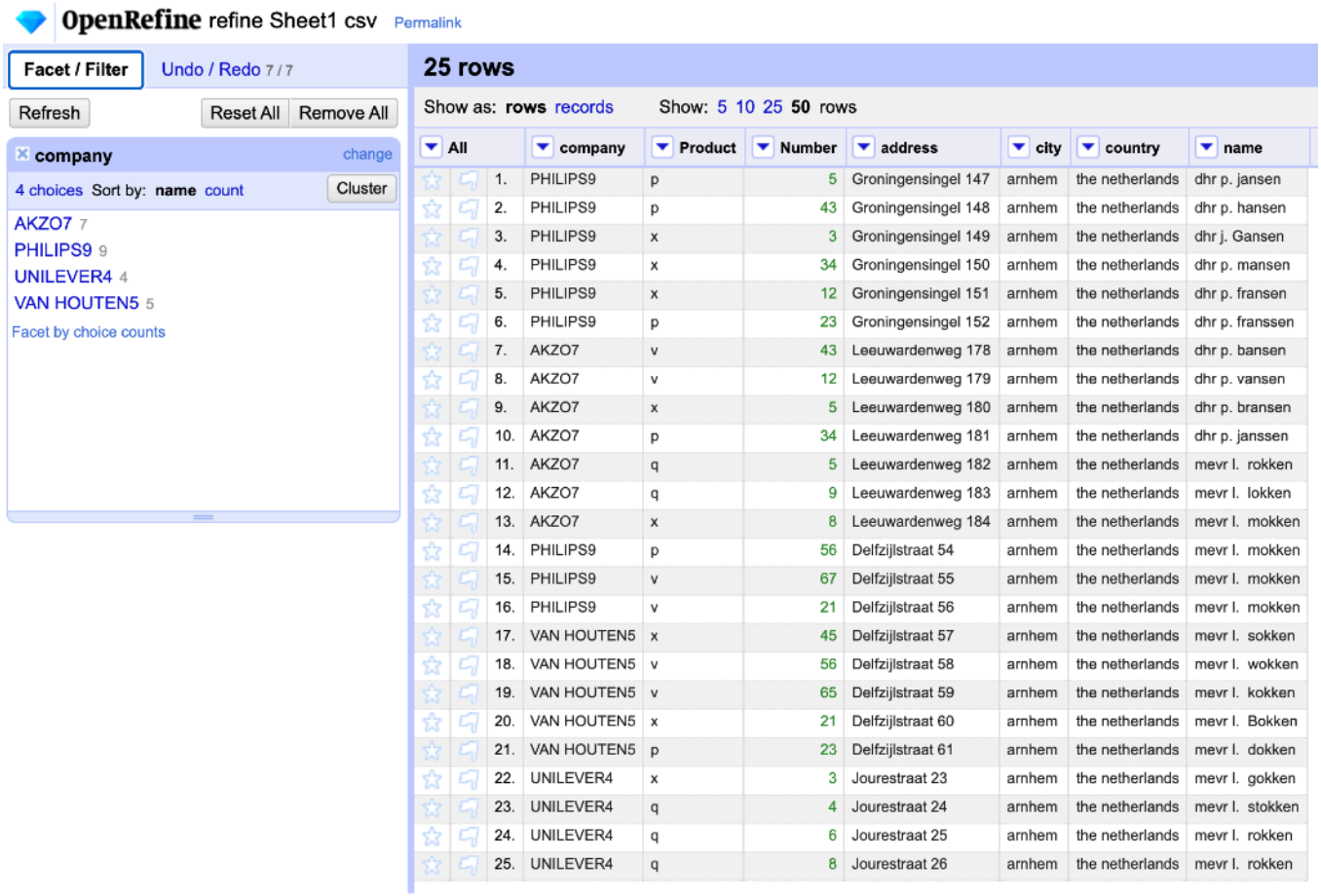
Now we have the correct spelling of the company names; let's turn them into uppercase. Choose from column company in the drop down menu edit cells, common transformations, into uppercase.



All company names are now capitalised.



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.



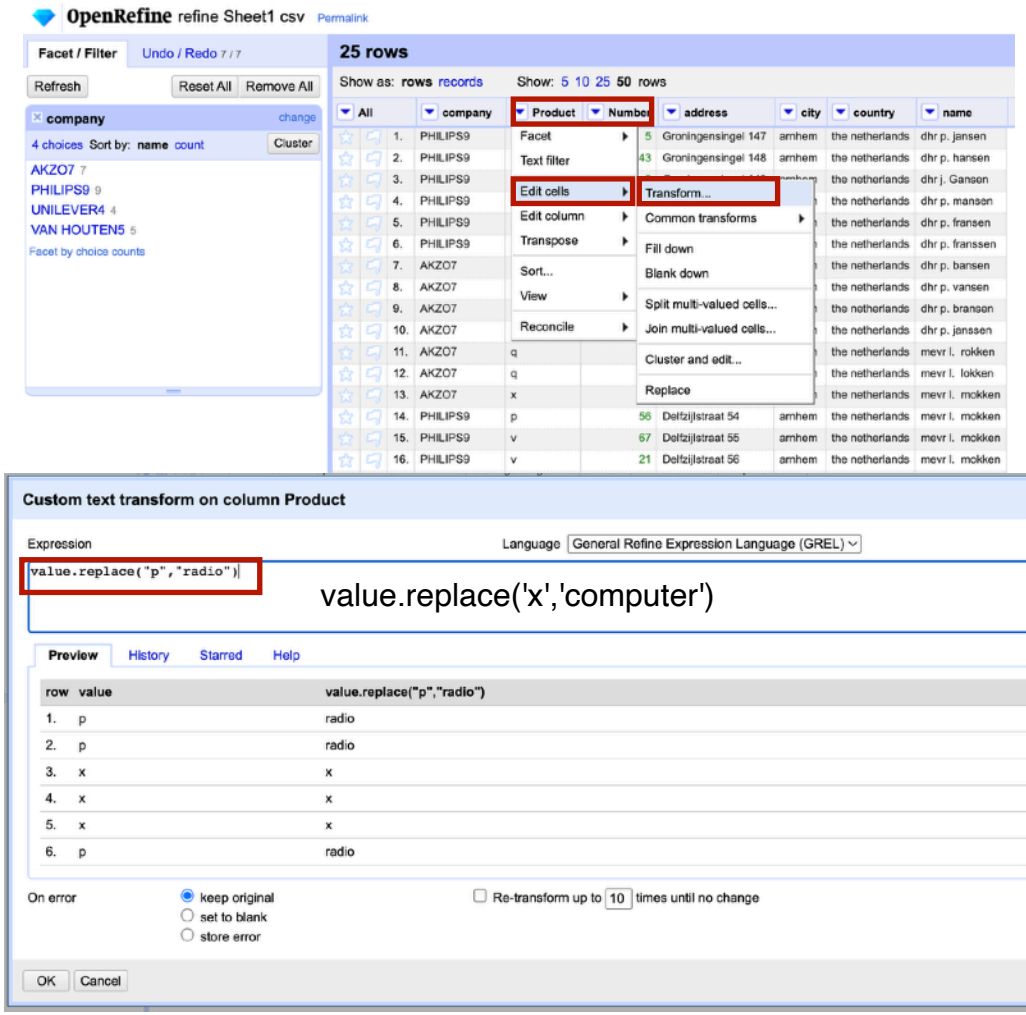
Product code and number are now separated.

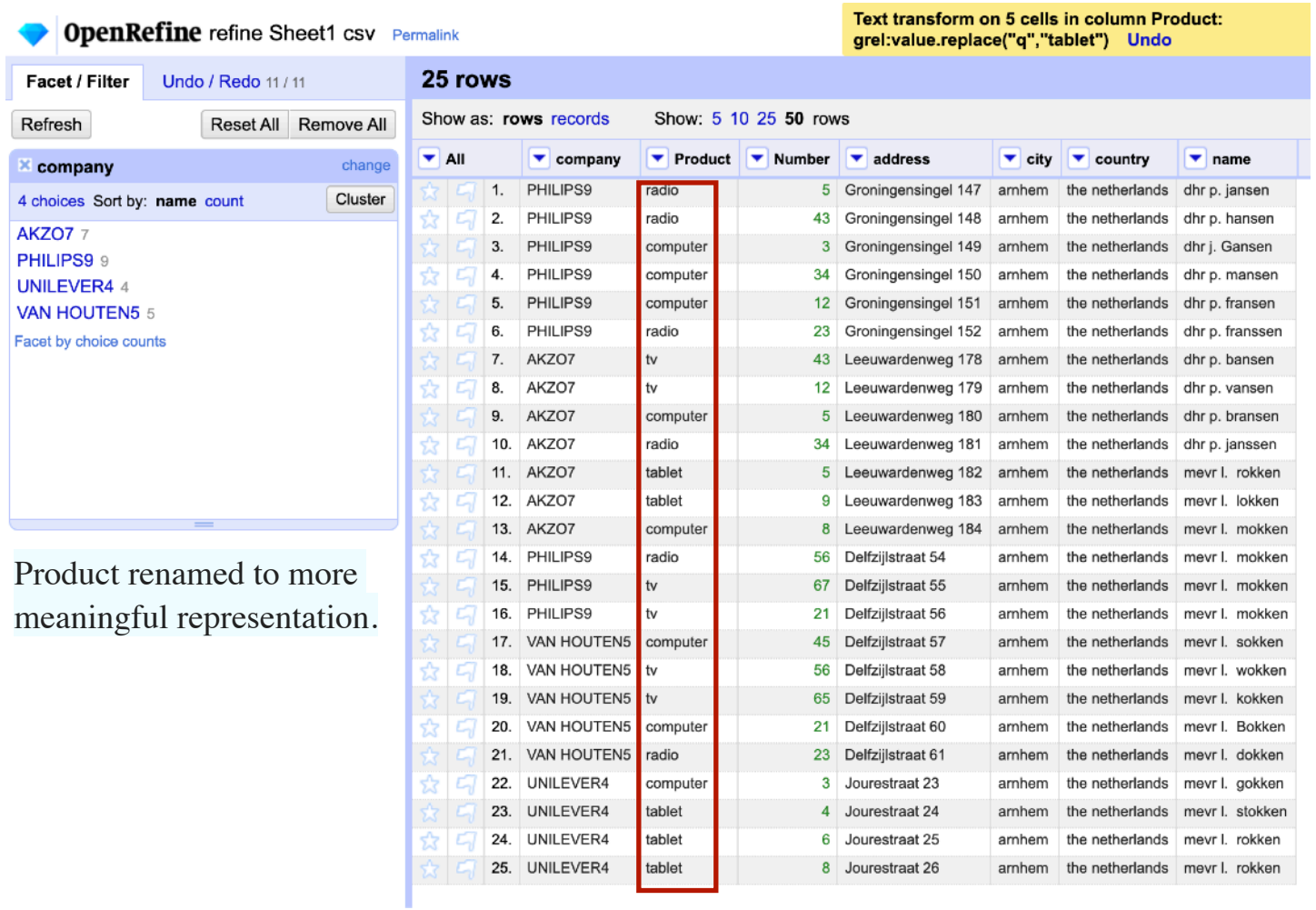
In column 2 we have a code for the product but not the name of the **product**. Here is the conversion:
p=radio
v=tv
x=computer
q=tablet

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:

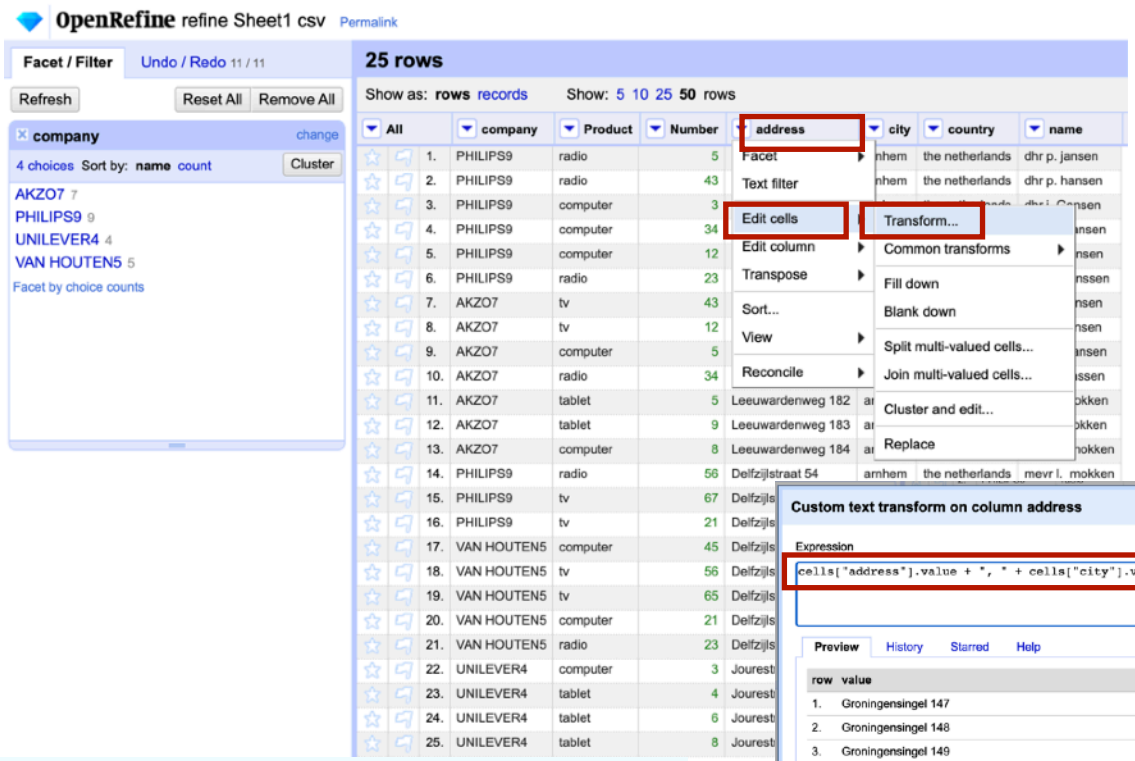
```
value.replace("p","radio")
```

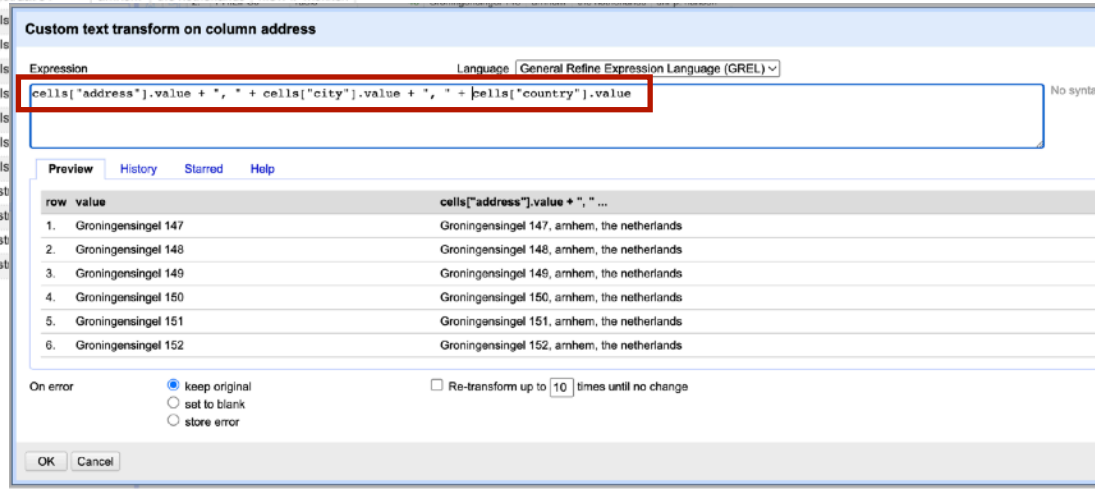
The left column show the original and the right the result of the transformation.
Do this for all different products.





Product renamed to more meaningful representation.





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:

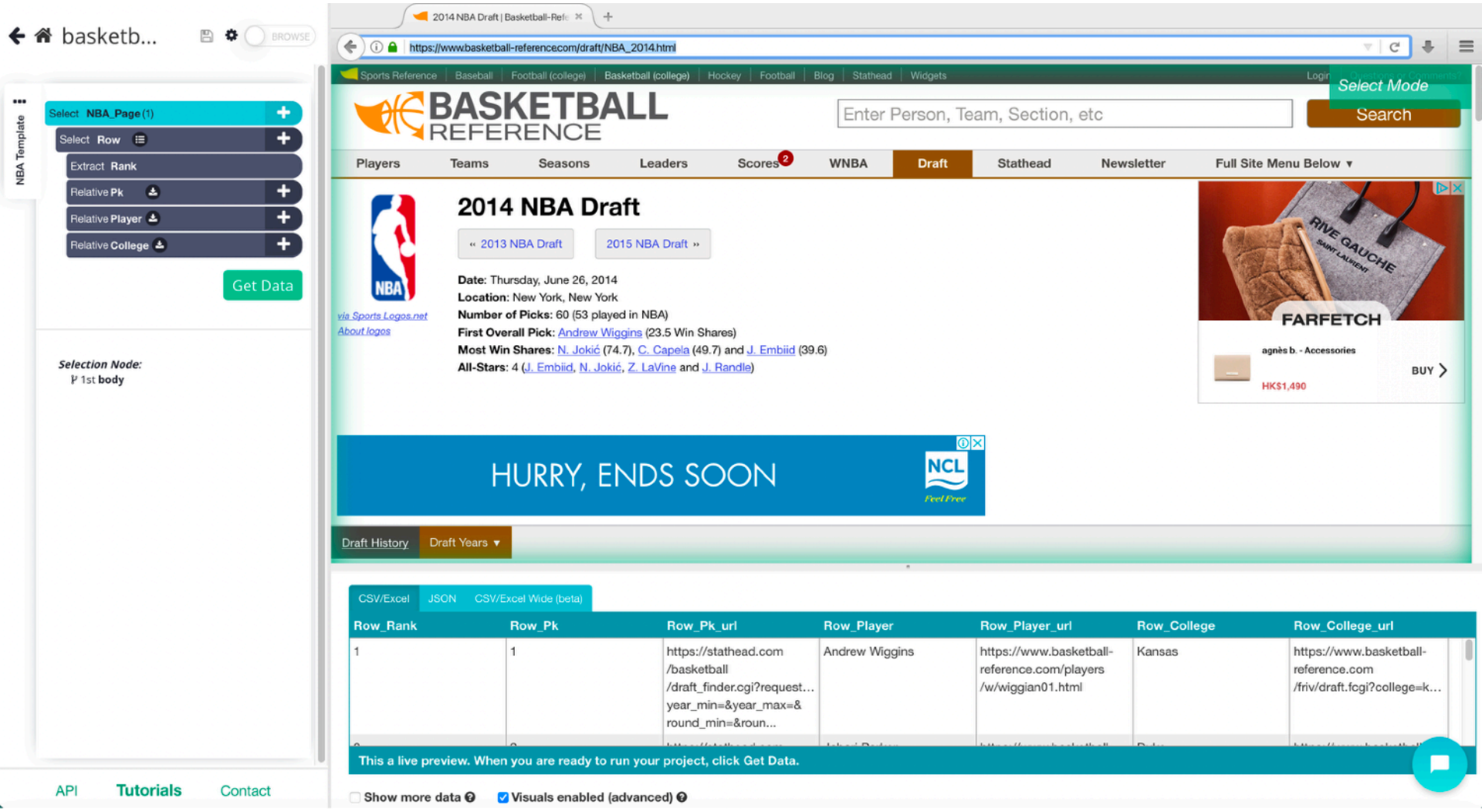
```
cells["address"].value + ", " + cells["city"].value + ", " + cells["country"].value
```

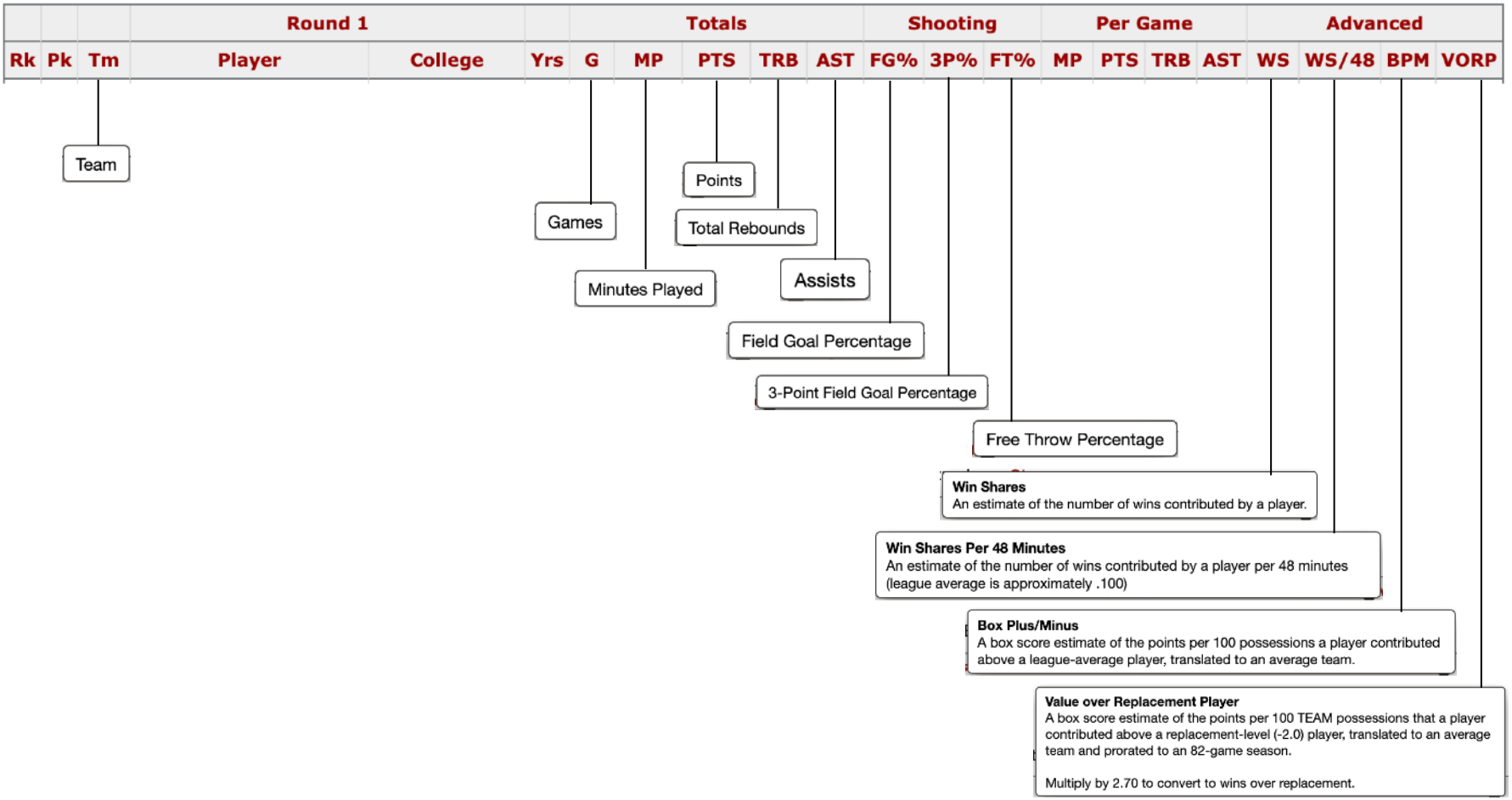
cells['address'].value + ', '+ cells['city'].value + ", " + cells['country'].value

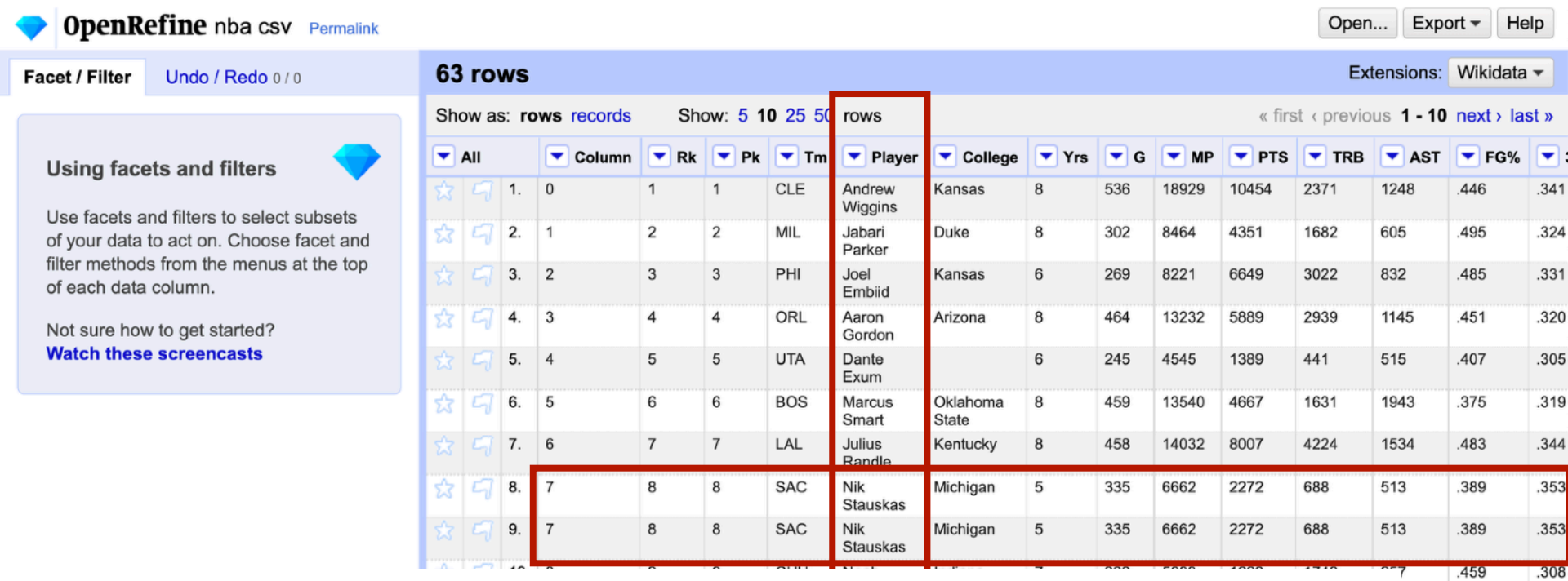
After combining the address, city, and country columns into one column, we can now remove the city and country columns.

Most Basic Data Cleaning Part 2

4. Remove duplicates.
5. Remove blanks.

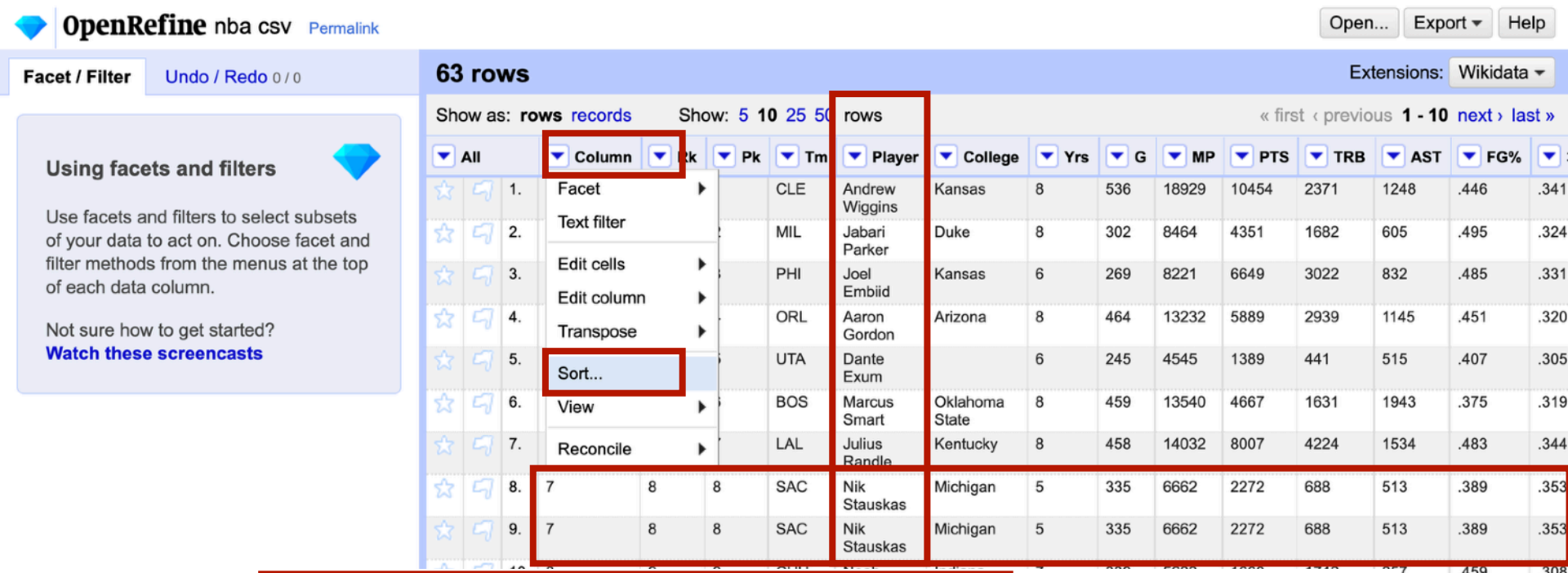

basketb...
BROWSE
NBA Template
Select NBA_Page (1)
Select Row
Extract Rank
Relative Pk
Relative Player
Relative College
Get Data
Selection Node:
1st body
API
Tutorials
Contact
2014 NBA Draft | Basketball-Ref
https://www.basketball-reference.com/draft/NBA_2014.html
Sports Reference | Baseball | Football (college) | Basketball (college) | Hockey | Football | Blog | Stathead | Widgets
Login
Select Mode
Search
BASKETBALL REFERENCE
Enter Person, Team, Section, etc
Players | Teams | Seasons | Leaders | Scores | 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
FARFETCH
agnès b. - Accessories
HK$1,490
BUY
HURRY, ENDS SOON
NCL
Feel Free
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
https://stathead.com/basketball/draft_finder.cgi?request...year_min=&year_max=&round_min=&roun...
Andrew Wiggins
https://www.basketball-reference.com/players/w/wiggian01.html
Kansas
https://www.basketball-reference.com/friv/draft.fcgi?college=k...
This a live preview. When you are ready to run your project, click Get Data.
Show more data
Visuals enabled (advanced)


Round 1
Totals
Shooting
Per Game
Advanced
Rk
Pk
Tm
Player
College
Yrs
G
MP
PTS
TRB
AST
FG%
3P%
FT%
MP
PTS
TRB
AST
WS
WS/48
BPM
VORP
Team
Games
Minutes Played
Points
Total Rebounds
Assists
Field Goal Percentage
3-Point Field Goal Percentage
Free Throw Percentage
Win Shares
An estimate of the number of wins contributed by a player.
Win Shares Per 48 Minutes
An estimate of the number of wins contributed by a player per 48 minutes (league average is approximately .100)
Box Plus/Minus
A box score estimate of the points per 100 possessions a player contributed above a league-average player, translated to an average team.
Value over Replacement Player
A box score estimate of the points per 100 TEAM possessions that a player contributed above a replacement-level (-2.0) player, translated to an average team and prorated to an 82-game season.
Multiply by 2.70 to convert to wins over replacement.



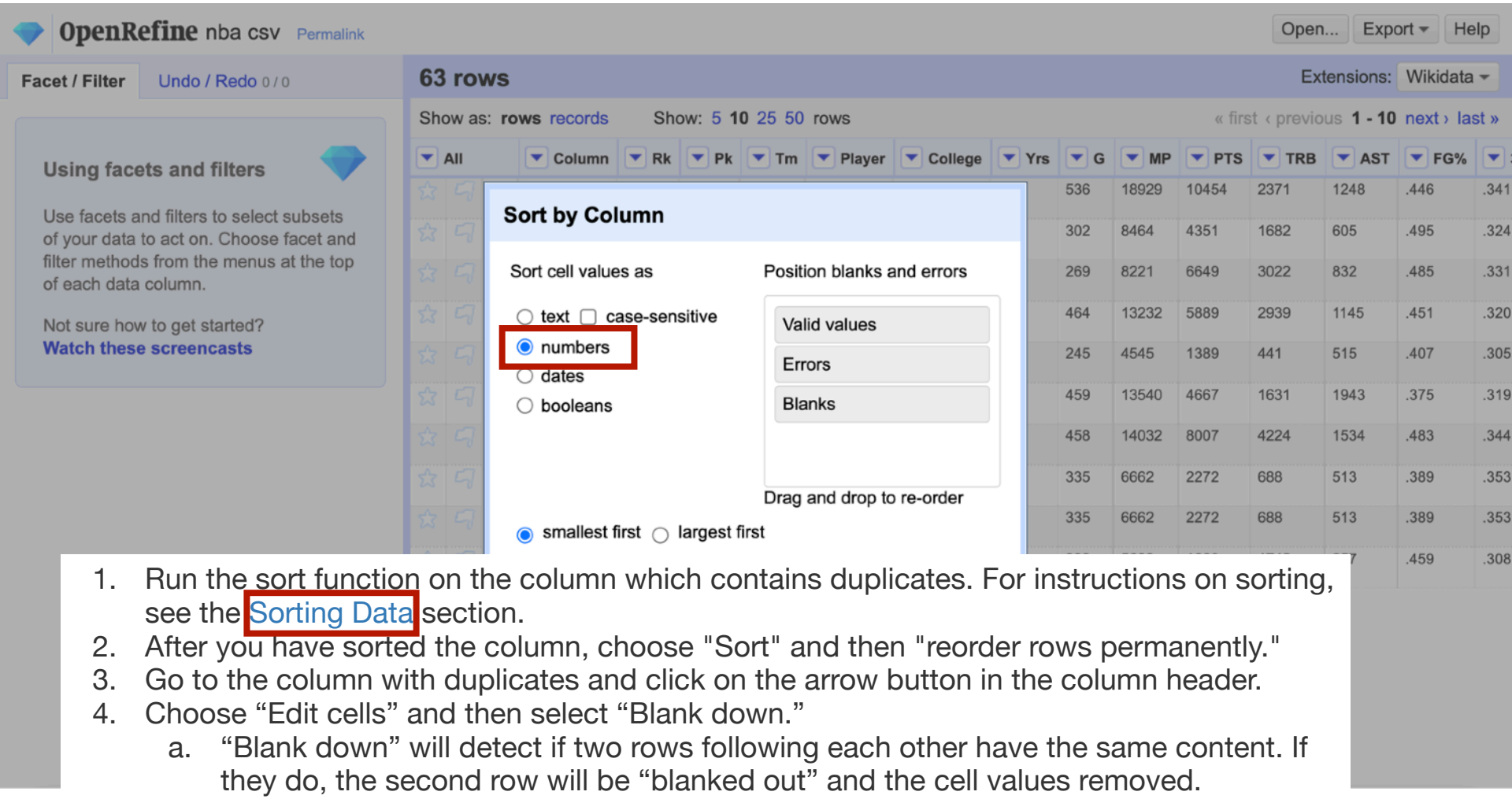
1. Run the sort function on the column which contains duplicates. For instructions on sorting, see the Sorting Data section.
2. After you have sorted the column, choose "Sort" and then "reorder rows permanently."
3. Go to the column with duplicates and click on the arrow button in the column header.
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



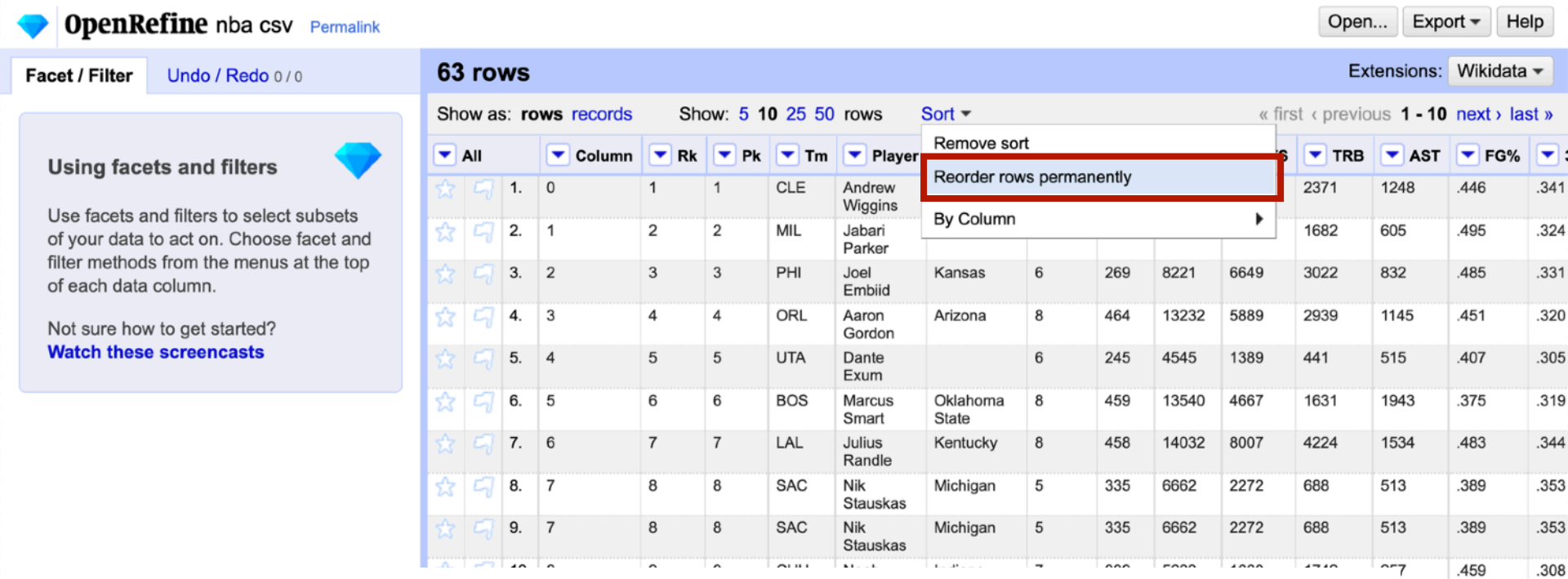
1. Run the sort function on the column which contains duplicates. For instructions on sorting, see the Sorting Data section.
2. After you have sorted the column, choose "Sort" and then "reorder rows permanently."
3. Go to the column with duplicates and click on the arrow button in the column header.
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



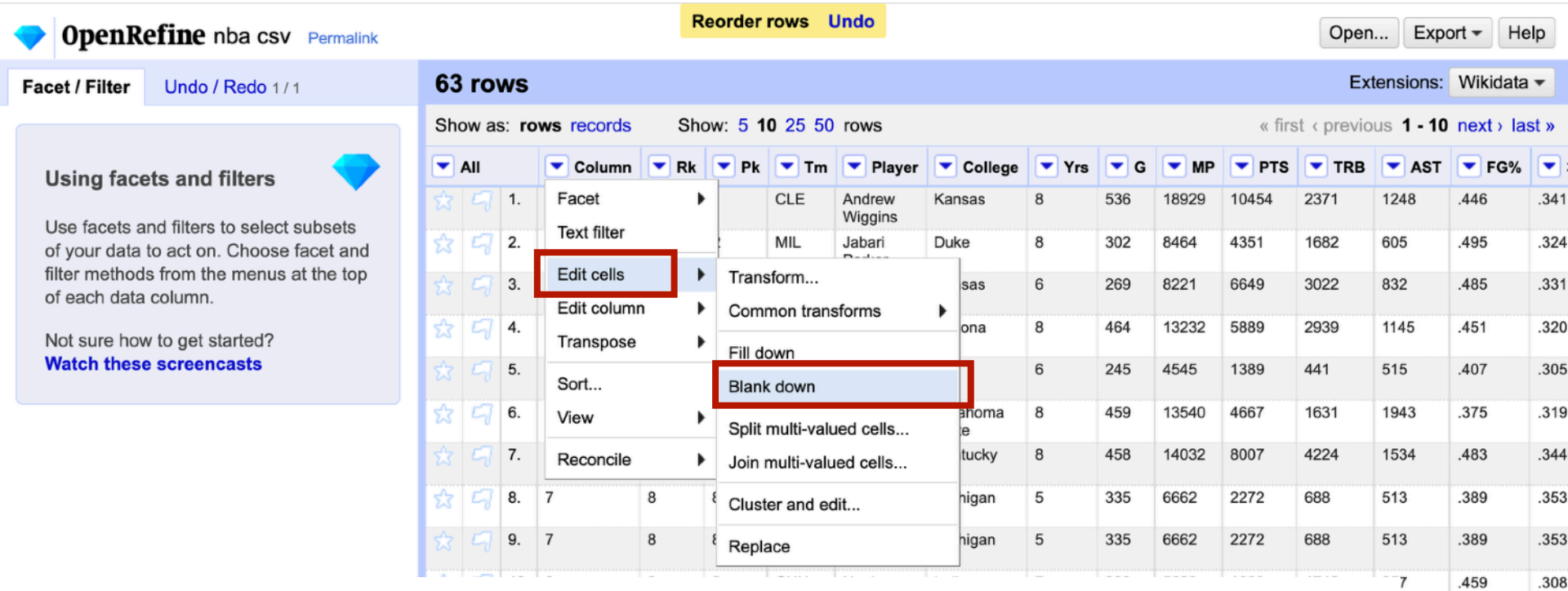
1. Run the sort function on the column which contains duplicates. For instructions on sorting, see the Sorting Data section.
2. After you have sorted the column, choose "Sort" and then "reorder rows permanently."
3. Go to the column with duplicates and click on the arrow button in the column header.
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



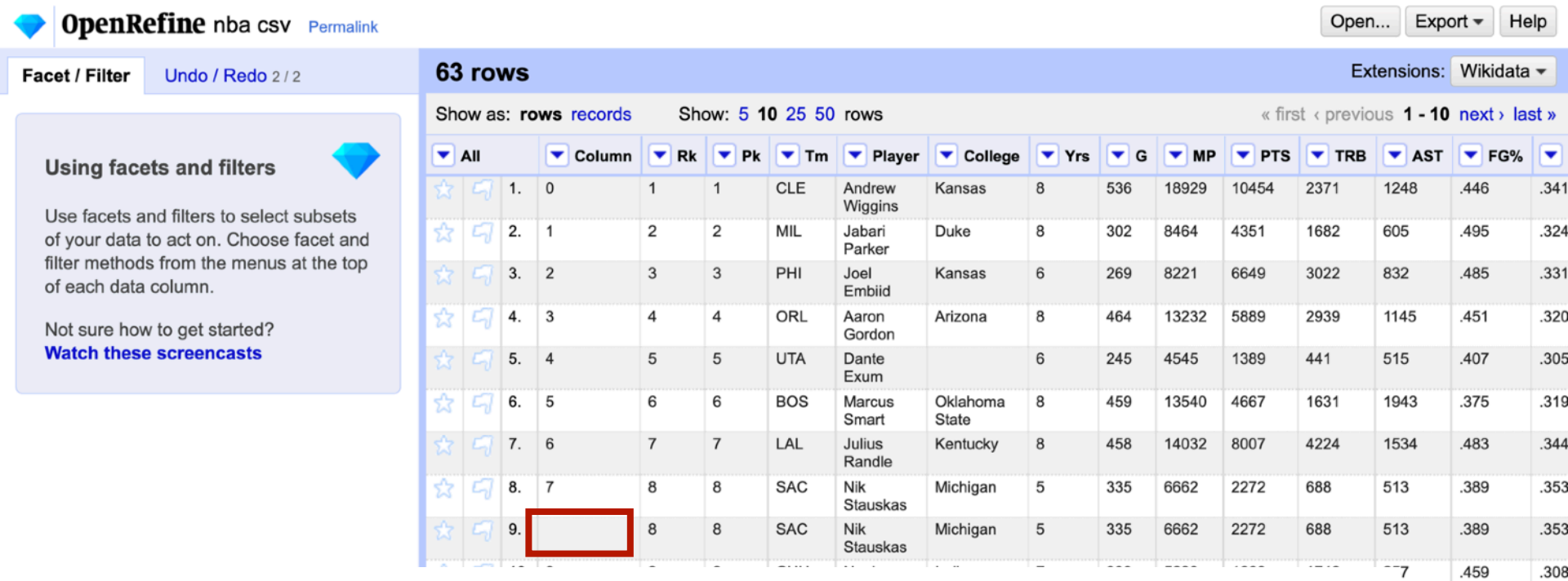
1. Run the sort function on the column which contains duplicates. For instructions on sorting, see the Sorting Data section.
2. After you have sorted the column choose "Sort" and then "reorder rows permanently."
3. Go to the column with duplicates and click on the arrow button in the column header.
4. 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



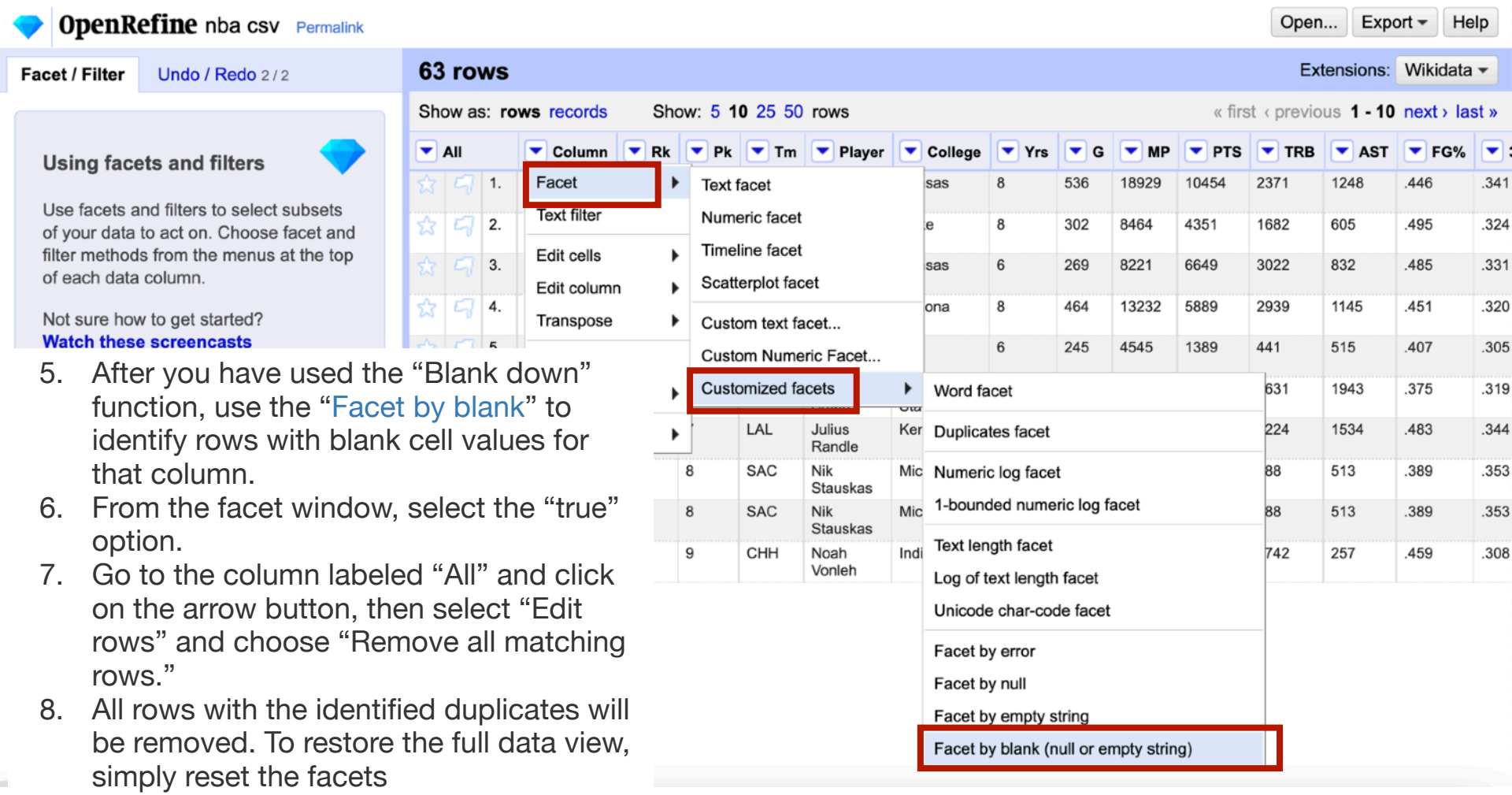
1. Run the sort function on the column which contains duplicates. For instructions on sorting, see the Sorting Data section.
2. After you have sorted the column, choose "Sort" and then "reorder rows permanently."
3. Go to the column with duplicates and click on the arrow button in the column header.
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



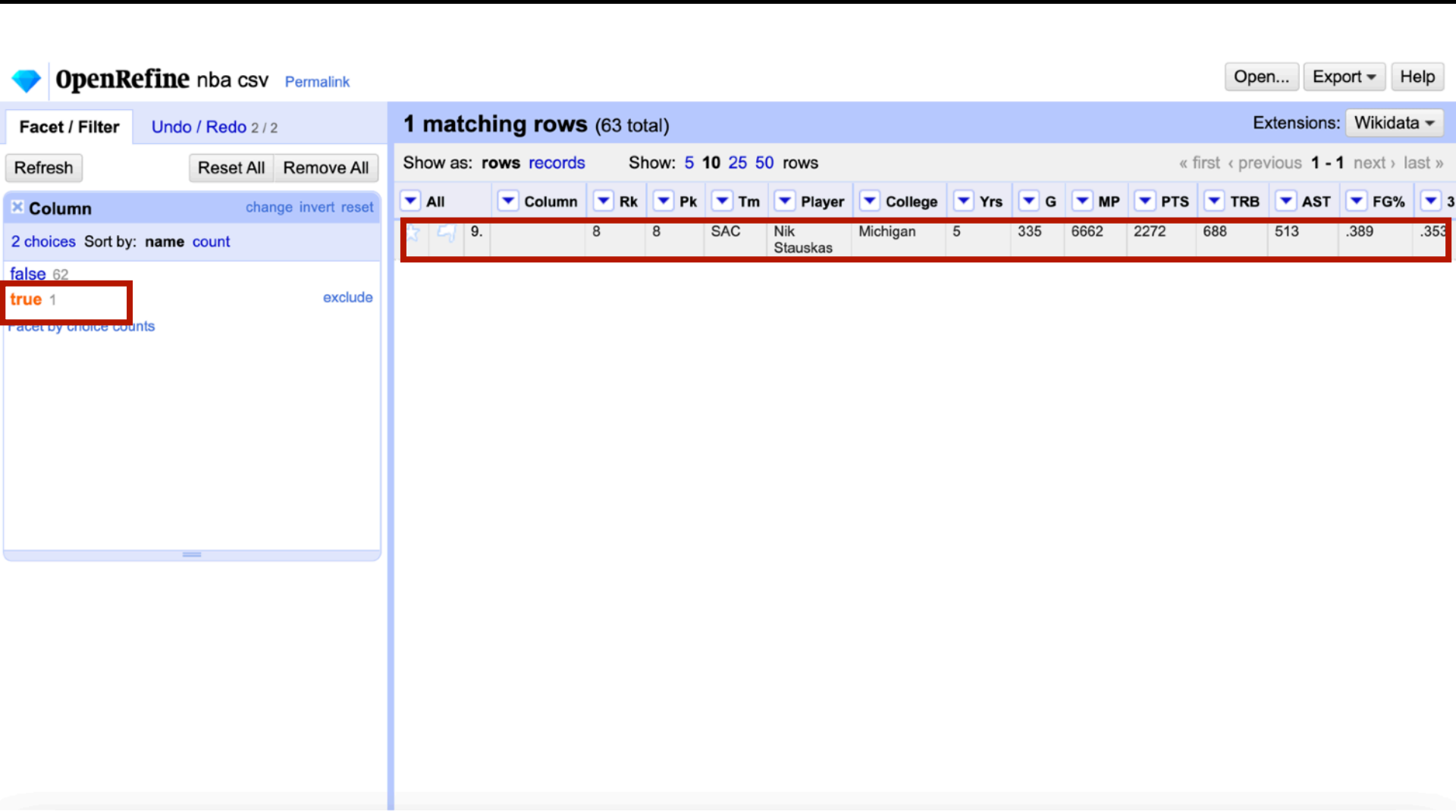
1. Run the sort function on the column which contains duplicates. For instructions on sorting, see the Sorting Data section.
2. After you have sorted the column, choose "Sort" and then "reorder rows permanently."
3. Go to the column with duplicates and click on the arrow button in the column header.
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

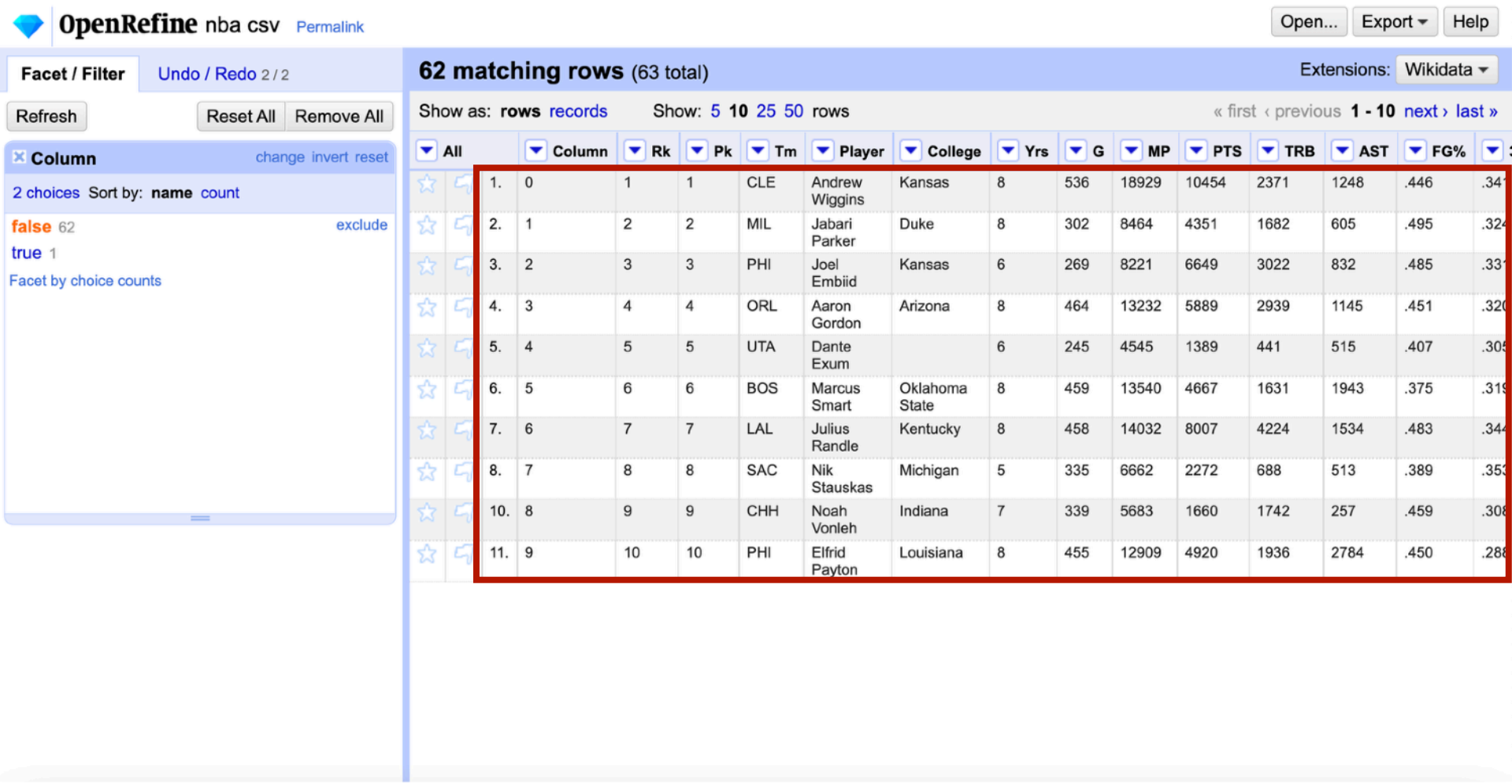


5. After you have used the “Blank down” function, use the “Facet by blank” to identify rows with blank cell values for that column.
6. From the facet window, select the “true” option.
7. Go to the column labeled “All” and click on the arrow button, then select “Edit rows” and choose “Remove all matching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

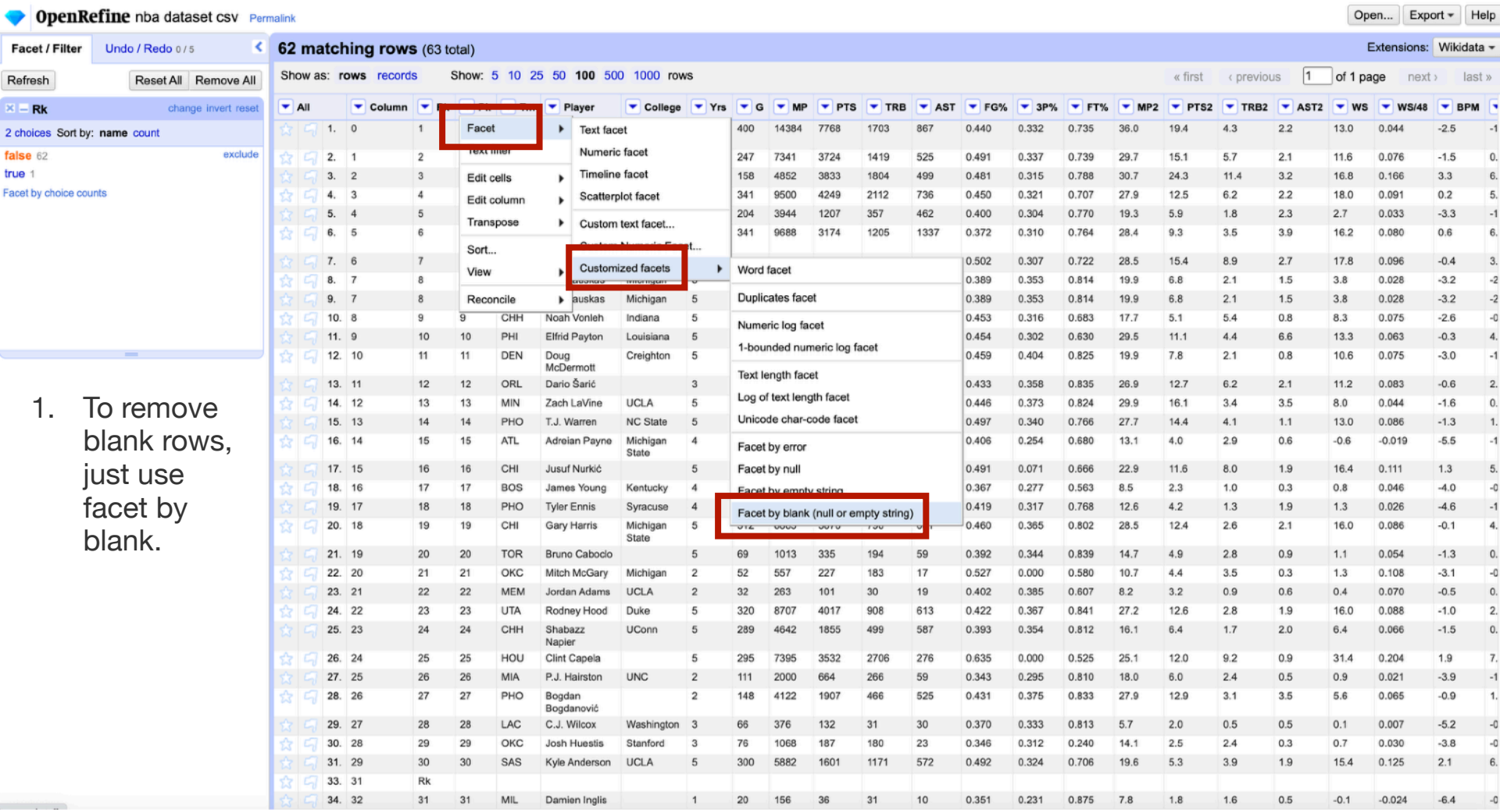
Source: library.illinois.edu



Source: library.illinois.edu



Source: library.illinois.edu

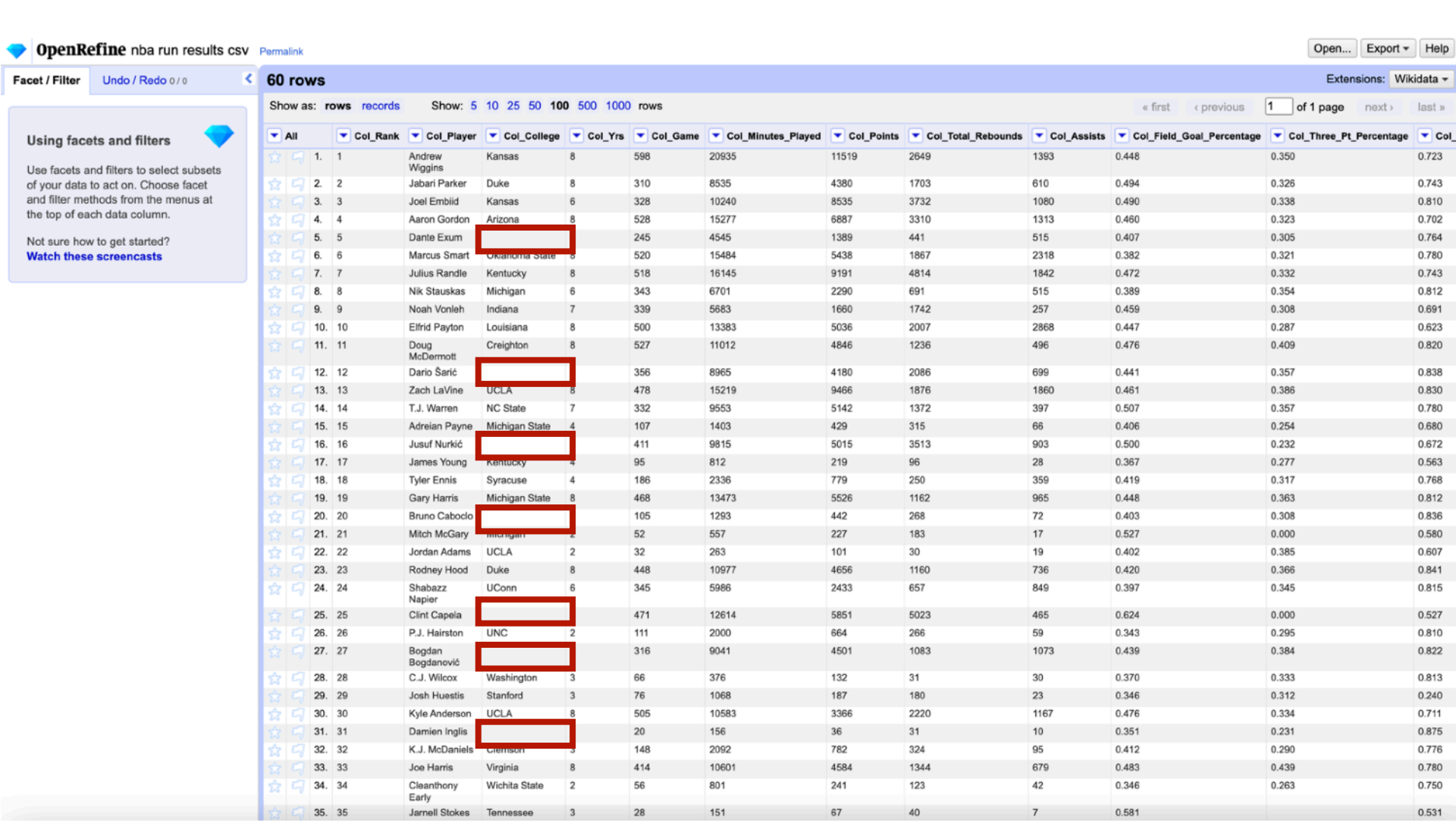


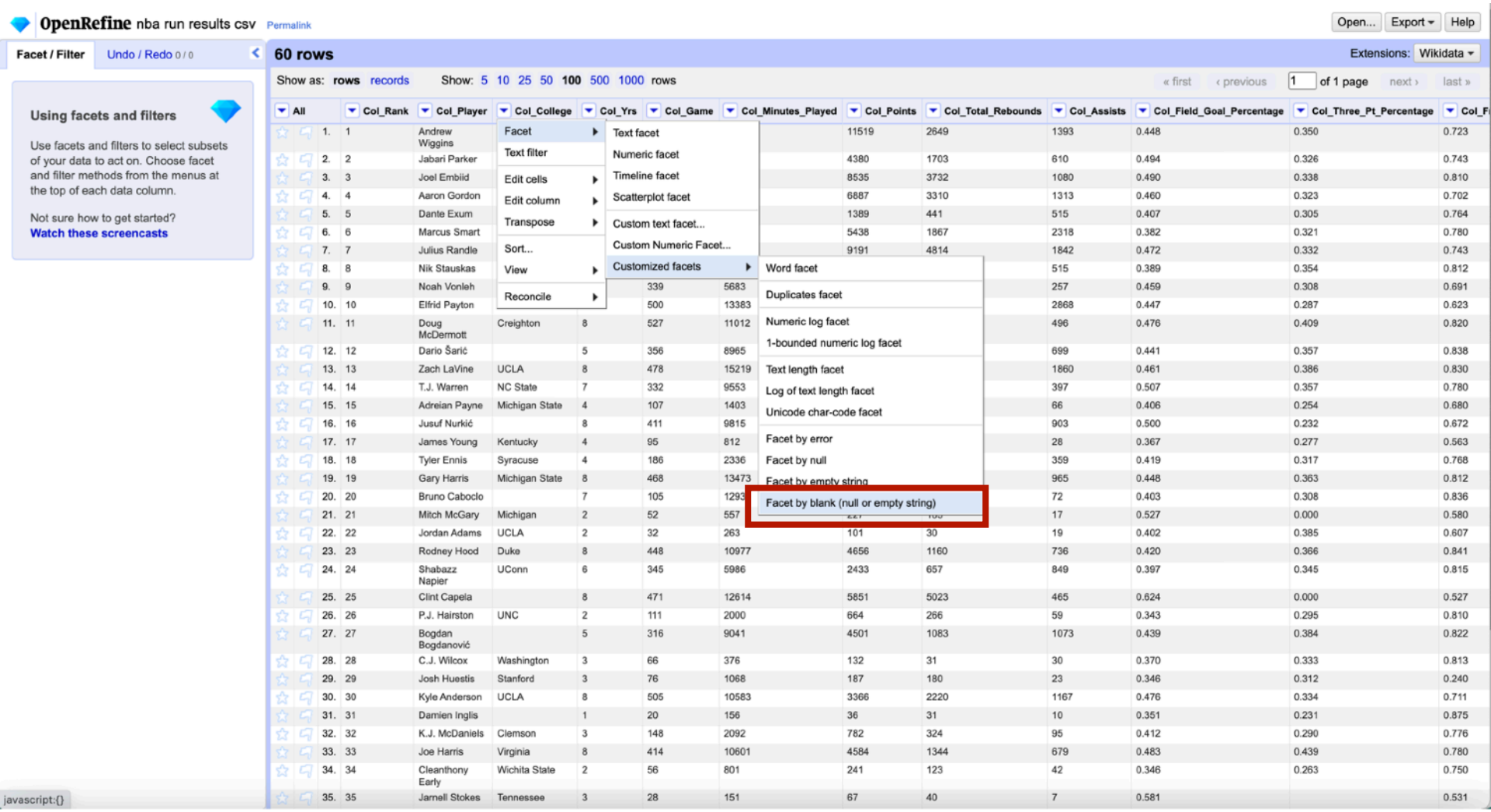
1. To remove blank rows, just use facet by blank.

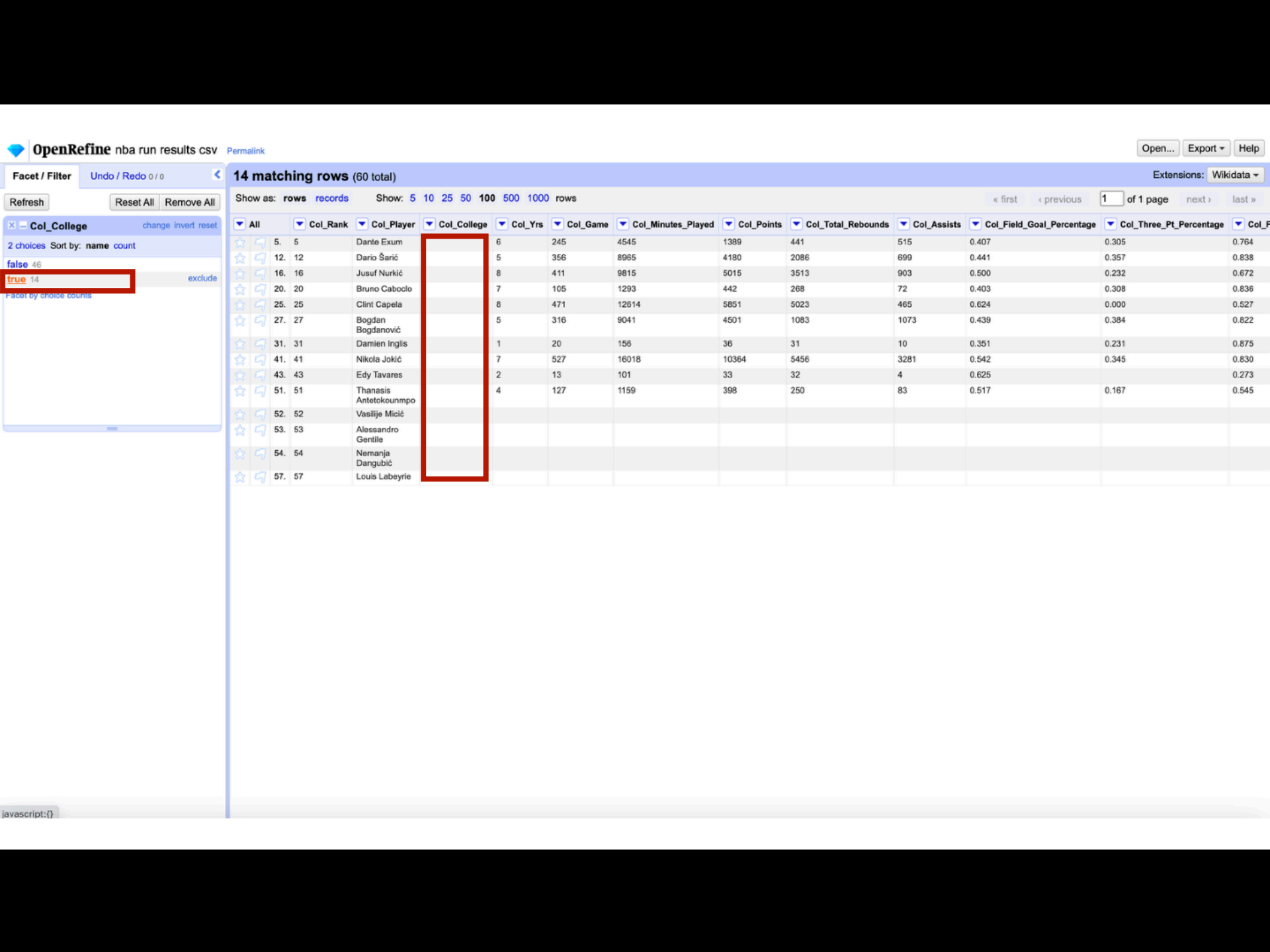
Most Basic Data Cleaning and Wrangling Tasks

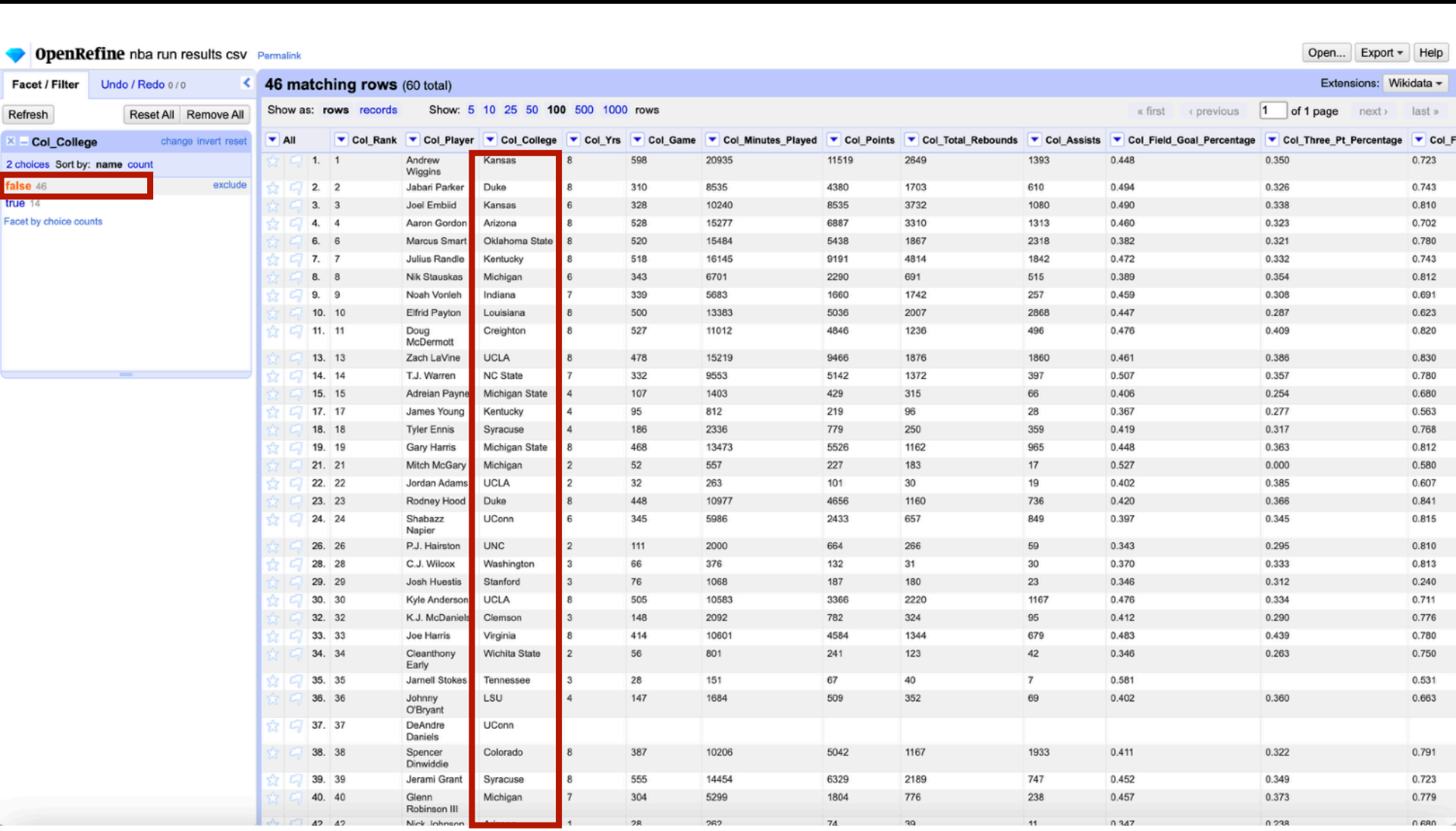
1. Remove blanks
2. Remove duplicates
3. Resolve inconsistent entries
4. Transform incorrect data (e.g. wrong data types such as text instead of numeric and calculations)
5. Rename, combine and split columns for making further data presentation and processing easier
6. Impute missing entries (need expertise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

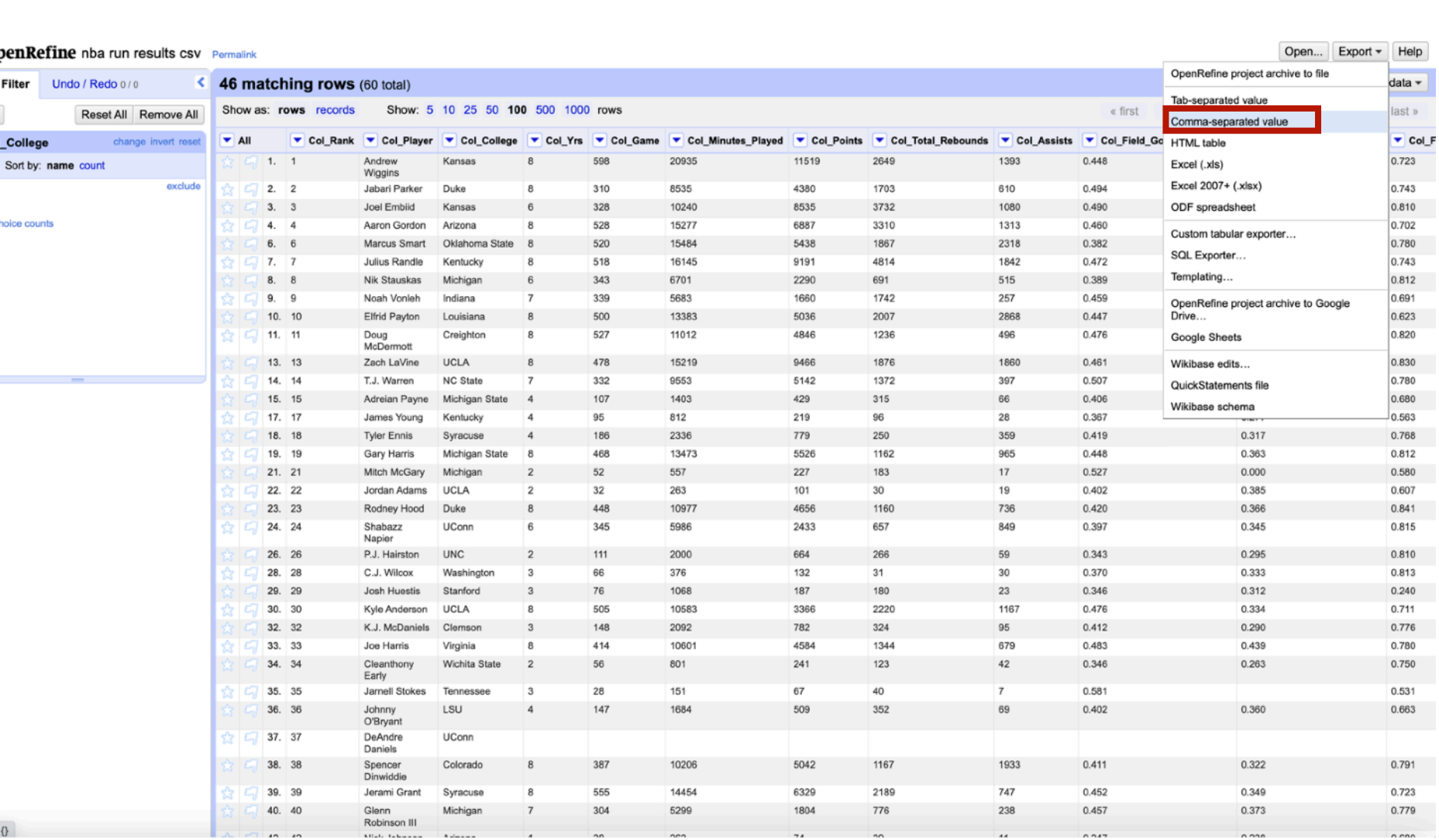
Analysing Cleaned Data Using SQL


OpenRefine nba run results csv
Permalink
Open...
Export
Help
Facet / Filter
Undo / Redo 0 / 0
60 rows
Extensions: Wikidata
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
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
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 245 4545 1389 441 515 0.407 0.305 0.764
6. 6 Marcus Smart 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 356 8965 4180 2086 699 0.441 0.357 0.838
13. 13 Zach LaVine 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 411 9815 5015 3513 903 0.500 0.232 0.672
17. 17 James Young 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 105 1293 442 268 72 0.403 0.308 0.836
21. 21 Mitch McGary 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 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 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 Inglis 20 156 36 31 10 0.351 0.231 0.875
32. 32 K.J. McDaniels 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
Open...
Export
Help
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
60 rows
Extensions: Wikidata
Show as: rows records
Show: 5 10 25 50 100 500 1000 rows
« first ‹ previous 1 of 1 page next › last »
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
Facet
Text filter
Edit cells
Edit column
Transpose
Sort...
View
Reconcile
Text facet
Numeric facet
Timeline facet
Scatterplot facet
Custom text facet...
Custom Numeric Facet...
Customized facets
Word facet
Duplicates facet
Numeric log facet
1-bounded numeric log facet
Text length facet
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)
1. 1 Andrew Wiggins 11519 2649 1393 0.448 0.350 0.723
2. 2 Jabari Parker 4380 1703 610 0.494 0.326 0.743
3. 3 Joel Embiid 8535 3732 1080 0.490 0.338 0.810
4. 4 Aaron Gordon 6887 3310 1313 0.460 0.323 0.702
5. 5 Dante Exum 1389 441 515 0.407 0.305 0.764
6. 6 Marcus Smart 5438 1867 2318 0.382 0.321 0.780
7. 7 Julius Randle 9191 4814 1842 0.472 0.332 0.743
8. 8 Nik Stauskas 515 0.389 0.354 0.812
9. 9 Noah Vonleh 339 5683 257 0.459 0.308 0.691
10. 10 Elfrid Payton 500 13383 2868 0.447 0.287 0.623
11. 11 Doug McDermott Creighton 8 527 11012 496 0.476 0.409 0.820
12. 12 Dario Šarić 5 356 8965 699 0.441 0.357 0.838
13. 13 Zach LaVine UCLA 8 478 15219 1860 0.461 0.386 0.830
14. 14 T.J. Warren NC State 7 332 9553 397 0.507 0.357 0.780
15. 15 Adreian Payne Michigan State 4 107 1403 66 0.406 0.254 0.680
16. 16 Jusuf Nurkić 8 411 9815 903 0.500 0.232 0.672
17. 17 James Young Kentucky 4 95 812 28 0.367 0.277 0.563
18. 18 Tyler Ennis Syracuse 4 186 2336 359 0.419 0.317 0.768
19. 19 Gary Harris Michigan State 8 468 13473 965 0.448 0.363 0.812
20. 20 Bruno Caboclo 7 105 1293 72 0.403 0.308 0.836
21. 21 Mitch McGary Michigan 2 52 557 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 8 471 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 5 316 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 Inglis 1 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
javascript:{}


OpenRefine nba run results csv
Permalink
Open...
Export
Help
Facet / Filter
Undo / Redo 0 / 0
Refresh
Reset All
Remove All
Col_College
change invert reset
2 choices Sort by: name count
false 46
true 14
exclude
Facet by choice counts
14 matching rows (60 total)
Extensions: Wikidata
Show as: rows records
Show: 5 10 25 50 100 500 1000 rows
« first ‹ previous 1 of 1 page next › last »
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
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 0.167 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
Open...
Export
Help
Facet / Filter
Undo / Redo 0 / 0
Refresh
Reset All
Remove All
Col_College
change invert reset
2 choices Sort by: name count
false 46
exclude
true 14
Facet by choice counts
46 matching rows (60 total)
Extensions: Wikidata
Show as: rows records
Show: 5 10 25 50 100 500 1000 rows
« first ‹ previous 1 of 1 page next › last »
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
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
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
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
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
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 0.531
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



All	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		0.723
2.	2	Jabari Parker	Duke	8	310	8535	4380	1703	610	0.494		0.743
3.	3	Joel Embiid	Kansas	6	328	10240	8535	3732	1080	0.490		0.810
4.	4	Aaron Gordon	Arizona	8	528	15277	6887	3310	1313	0.460		0.702
6.	6	Marcus Smart	Oklahoma State	8	520	15484	5438	1867	2318	0.382		0.780
7.	7	Julius Randle	Kentucky	8	518	16145	9191	4814	1842	0.472		0.743
8.	8	Nik Stauskas	Michigan	6	343	6701	2290	691	515	0.389		0.812
9.	9	Noah Vonleh	Indiana	7	339	5683	1660	1742	257	0.459		0.691
10.	10	Elfrid Payton	Louisiana	8	500	13383	5036	2007	2868	0.447		0.623
11.	11	Doug McDermott	Creighton	8	527	11012	4846	1236	496	0.476		0.820
13.	13	Zach LaVine	UCLA	8	478	15219	9466	1876	1860	0.461		0.830
14.	14	T.J. Warren	NC State	7	332	9553	5142	1372	397	0.507		0.780
15.	15	Adreian Payne	Michigan State	4	107	1403	429	315	66	0.406		0.680
17.	17	James Young	Kentucky	4	95	812	219	96	28	0.367		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
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		0.531
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

In [3]:

```
import pandas as pd
import sqlite3
con = sqlite3.connect('nba_db.db')
sql = "SELECT * FROM nba_run_results WHERE Col_Game > 400"
df = pd.read_sql_query(sql,con)
con.close()
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	

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

```
# CAST(ROUND(Column_Name, 2) AS DECIMAL(10,2)
import pandas as pd
import sqlite3
con = sqlite3.connect('nba_db.db')
# sql = "SELECT Col_Player, Col_College, Col_Game, Col_Yrs, ROUND((Col_Game*1.0/Col_Yrs),2) AS [Games/Years], ROUN
sql = "SELECT Col_Player, Col_College, Col_Game, Col_Yrs, ROUND(CAST(Col_Game AS FLOAT)/Col_Yrs,2) AS [Games/Years
df = pd.read_sql_query(sql,con)
con.close()
df
```

	Col_Player	Col_College	Col_Game	Col_Yrs	Games/Years	Avg Poits
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!