HOW TO SELL WELL ON SHOPEE?

A Simple Data Science Project

Table Of Contents

Overview	Data Crawler	Data exploration & preprocessing	Data Modeling	Wrap up
• What is this project? • Why ? • Dataset	Some great tips to crawl data	Missing values, invalid value, wrong type features, skewed output, handle object type, scaling.	How to find a good model to predict sales	Wrap up

About Us

We are students from University Of Science

18120184 Nguyễn Nguyên Khang

18120189 Trần Đăng Khoa

Thanks to our teacher, Trần Trung Kiên.

HOW TO SELL WELL ON SHOPEE?

A Simple Data Science Project

Nguyen Nguyen Khang | Tran Dang Khoa

HOW TO SELL WELL ON SHOPEE?

> Overview
> Crawl data
> Data exploration & preprocessing
> Data Modeling
> Wrap Up

Overview: What and Why?

- Practical Value
- Impact to real life, relatable: Economic, Medical,..
- Economic -> E-commerce -> For Seller or Customer? -> Seller

  -> How to sell well on the internet?

- Narrow the scope
- The internet -> Shopee -> Men fashion on shopee.vn

  -> How to sell Men Fashion well on shopee.vn

Overview: Dataset

- 10851 samples, 34 features
- features

HOW TO SELL WELL ON SHOPEE?

A Simple Data Science Project

Nguyen Nguyen Khang | Tran Dang Khoa

HOW TO SELL WELL ON SHOPEE?

> Overview
> Crawl data
> Data exploration & preprocessing
> Data Modeling
> Wrap Up

Crawl Data

From idea → to datasets → How to get it?

- Sell well on the internet

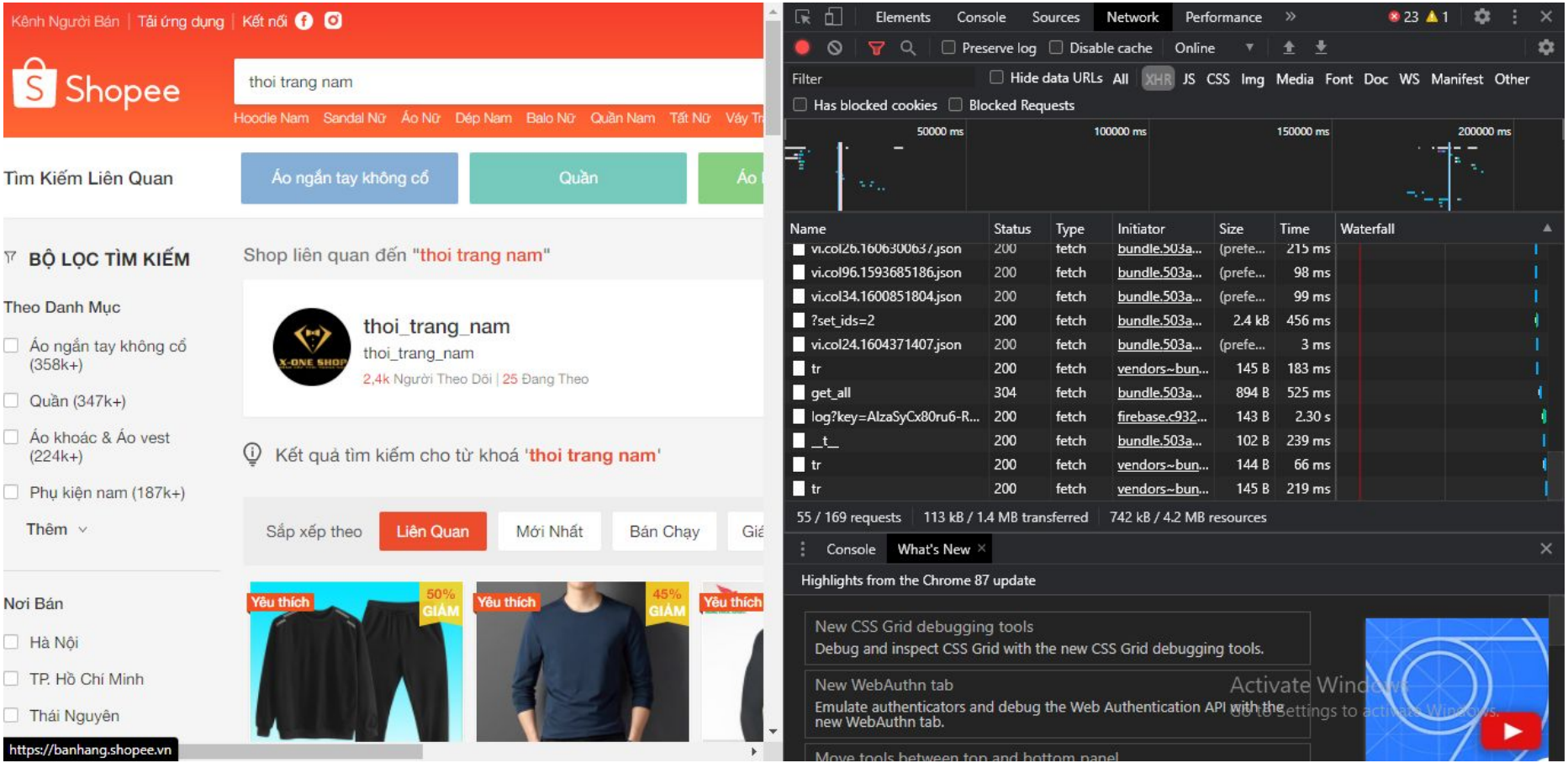
- Hi Shopee!

- I prefer to use API
- But shoppe have no puplic API!

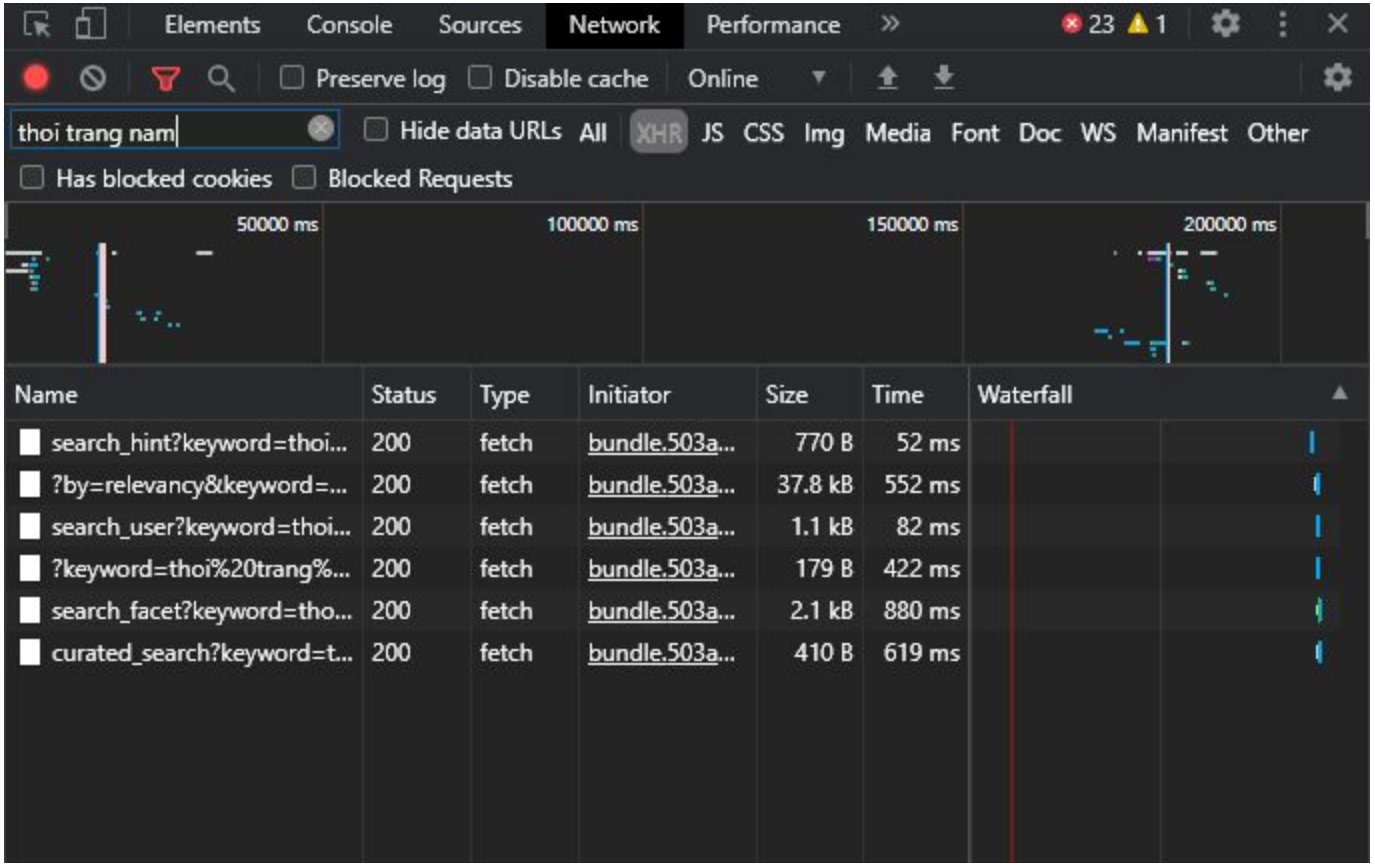
Crawler

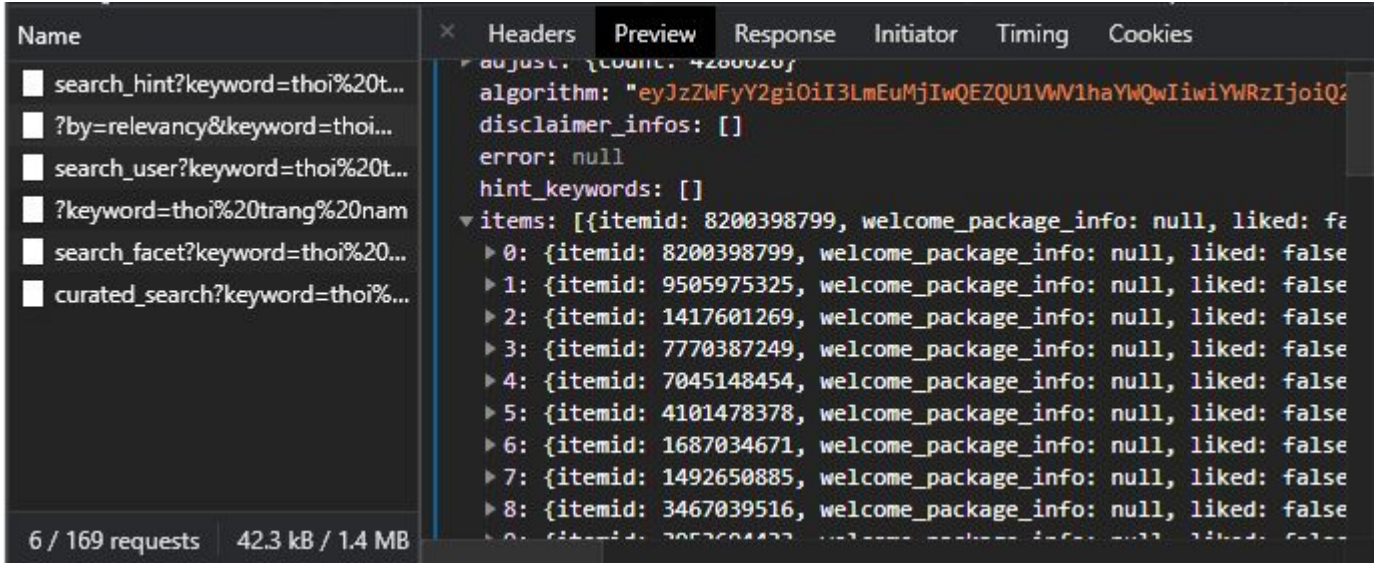
- Based on the knowledge about back-end and front-end
- I have a belief that it must be an API call to return data
- How to catch this?
- We need: Chrome/ FireFox, Postman, Jupyter Notebook

Type “thoi trang nam” to search bar
Use Developer tools/ Network/XHR (In Chrome/FireFox)

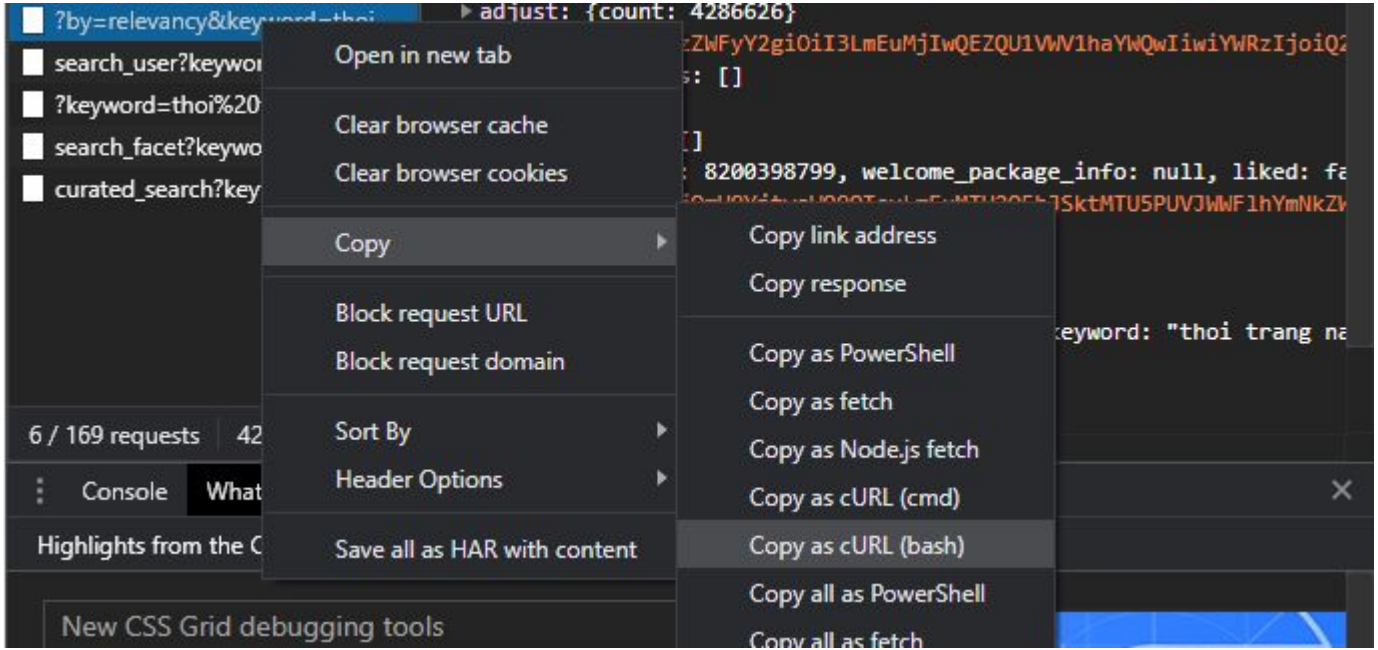


It should be a call have “thoi trang nam”, we search for this and check each request using the data it returns.



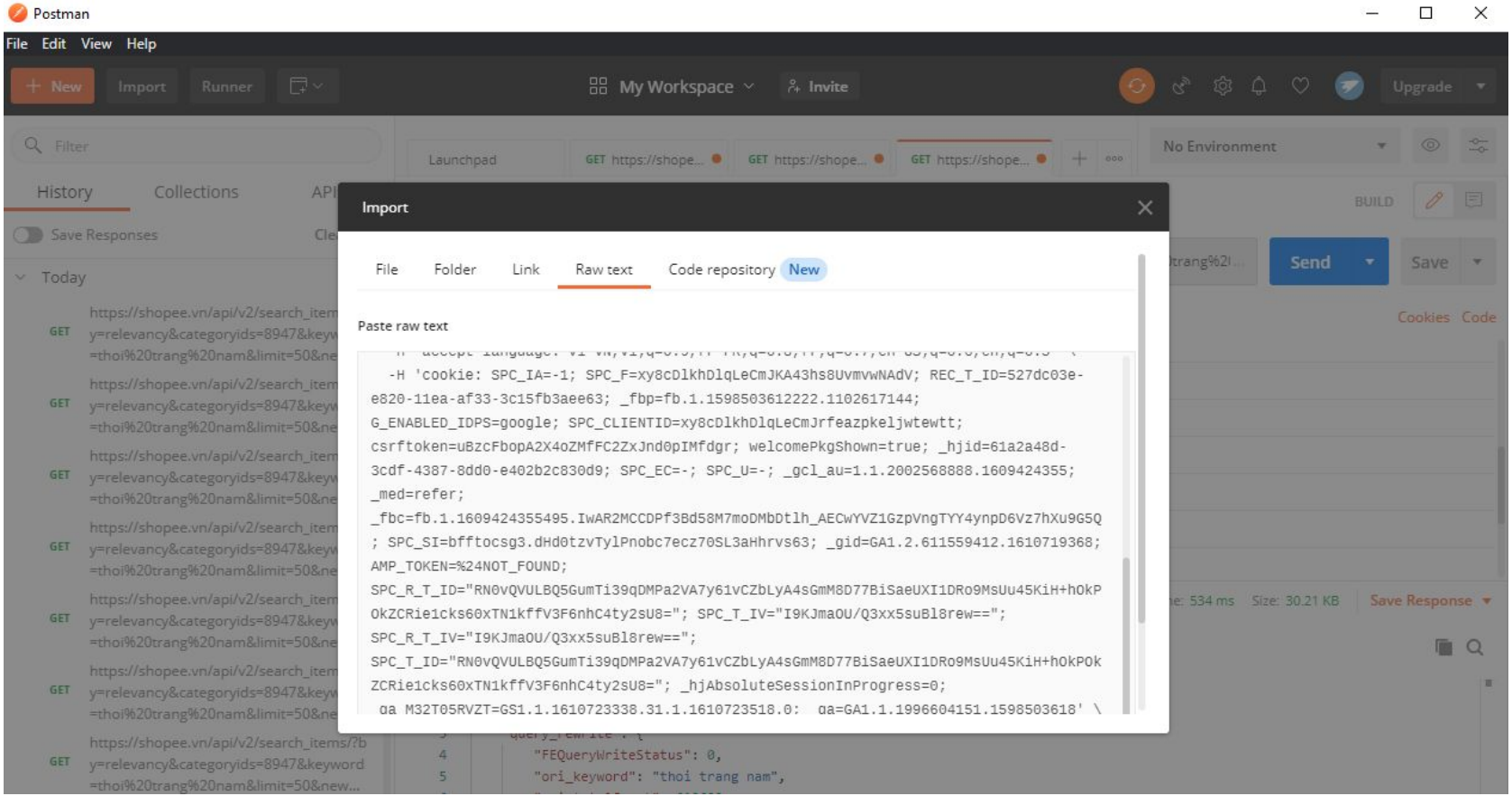


It seems like the call we need.



Copy/ Copy as cURL(bash) (Chrome)
In FireFox: Copy/ Copy as cURL

Next, we use Postman to test this request
Import/ Raw text: Paste what we copied before to this



After Import:

Click Send to send this GET request

Untitled Request

BUILD

GET

https://shopee.vn/api/v2/search_items/?by=relevancy&categoryids=8947&keyword=thoi%20trang%2...

Send

Save

Params

Authorization

Headers (23)

Body

Pre-request Script

Tests

Settings

Cookies

Code

Query Params

	KEY	VALUE	DESCRIPTION
☑	by	relevancy	
☑	categoryids	8947	
☑	keyword	thoi%20trang%20nam	
☑	limit	50	
☑	newest	50	

Bulk Edit

Result:

newest 50

Body Cookies (5) Headers (19) Test Results Status: 200 OK Time: 534 ms Size: 30.21 KB Save Response

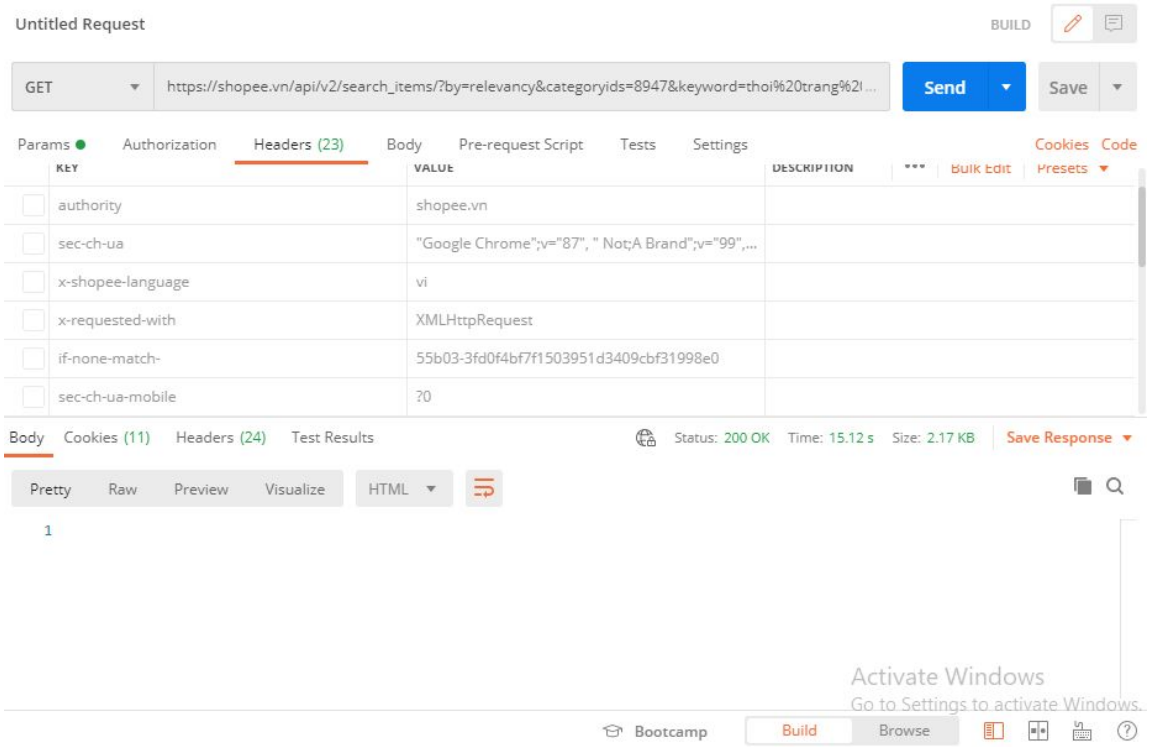
Pretty Raw Preview Visualize JSON


```
127        "has_group_buy_stock": null,
128        "match_type": null,
129        "preview_info": null,
130        "welcome_package_type": 0,
131        "exclusive_price_info": null,
132        "name": "áo thời trang nam",
133        "distance": null,
134        "adsid": null,
```


Bootcamp Build Browse

It definitely the result we need

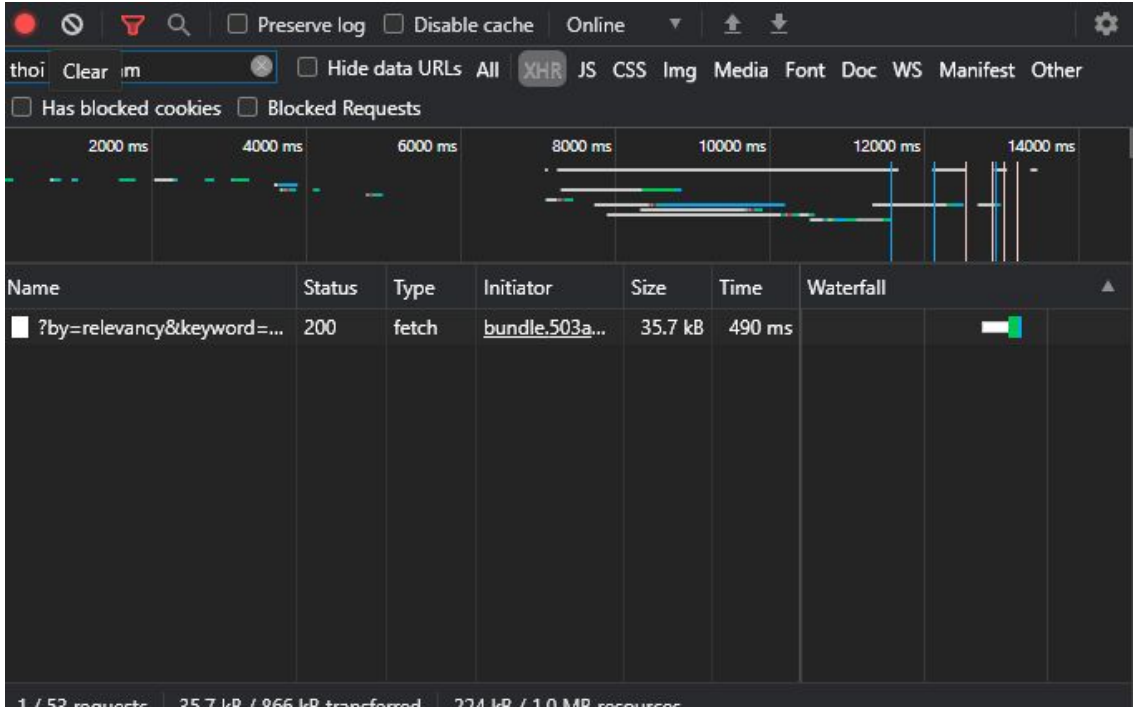
Why we did not just type it in the URL bar?
-> The Headers
Let see what will happen if we uncheck these request headers in Postman.



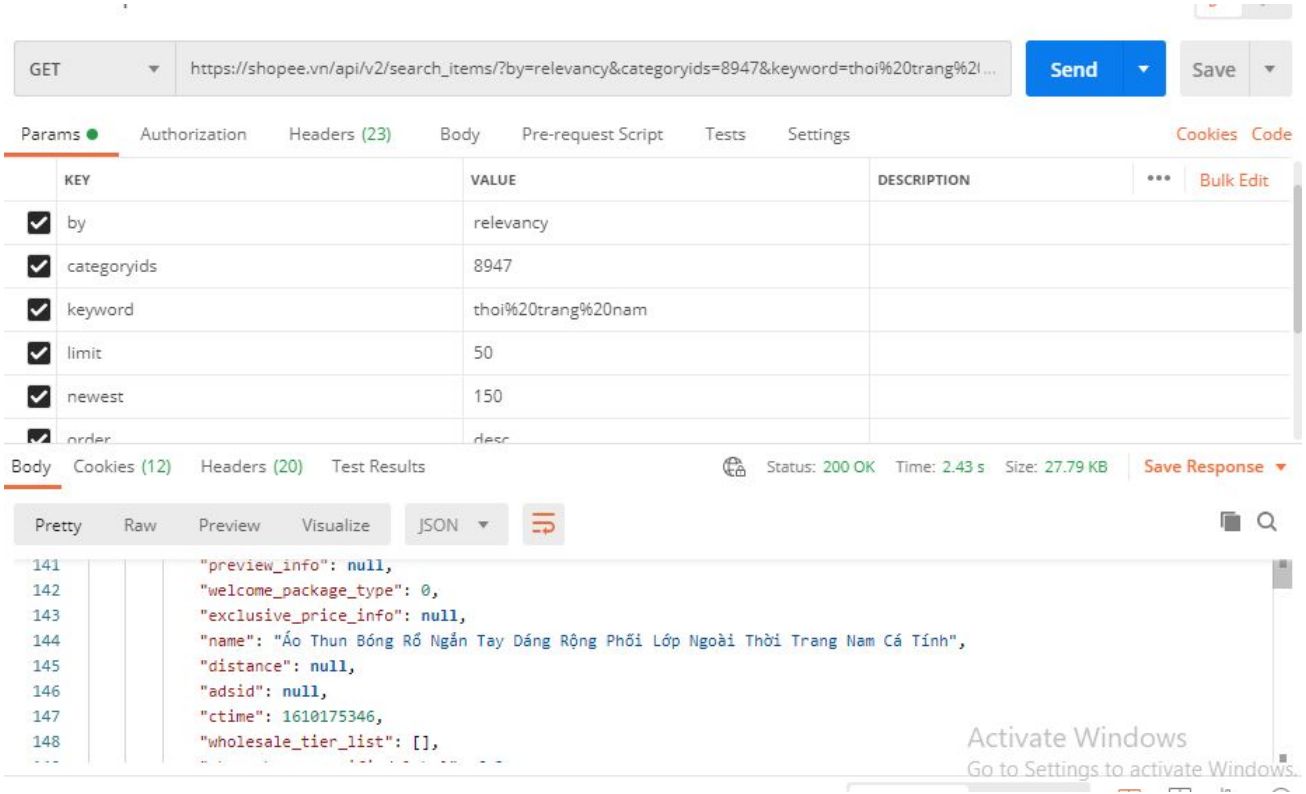
Keep checking till we find this is what we need in headers: cookie, if-non-match, if-non-match-

How to get data in new page?

-> **Clear,** then go to the next page. We will see this request appeared again but with a different value of params: newest = 50 instead of 0



Back to Postman, we will check with the newest = 50,100,150 and see if it return the data on page 1, page 2, page 3.
The answer is Yes, it is the data of page 1, page 2, page3.
It mean page 0: newest = 0, page 1: newest = 50,...



After knowing about Shopee API, it is easy to create a crawler.
We need 3 files:

- Model: Keep the data of 1 sample
- APIService: Where we get data from shopee
- GetData: Use APIService to get Data through page and page and parse it to Model, then save to a file

Demo

HOW TO SELL WELL ON SHOPEE?

A Simple Data Science Project

Nguyen Nguyen Khang | Tran Dang Khoa

HOW TO SELL WELL ON SHOPEE?

> Overview
> Crawl data
> Data exploration & preprocessing
> Data Modeling
> Wrap Up

Initial exploration and preprocessing

Duplicated samples (Solved by removing)

Invalid samples

- Negative time
- Min price > max price

Solution: Removing those samples

- Negative price before discount (valid if there is no discount)

Solution: Remove those samples if there is a discount

Initial exploration and preprocessing

Skewed output

Impact:

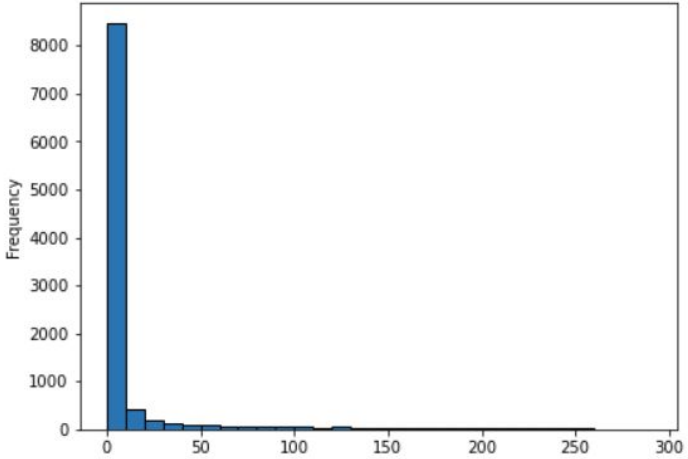
- Predictions of trained model are more accurate for lower values of output
- Skewed distribution converge much slower than symmetric one

Some approaches:

- Acquiring more non-zero output samples (Oversampling)
- Removing most of zero output samples (Undersampling)
- Transforming output to more balanced data

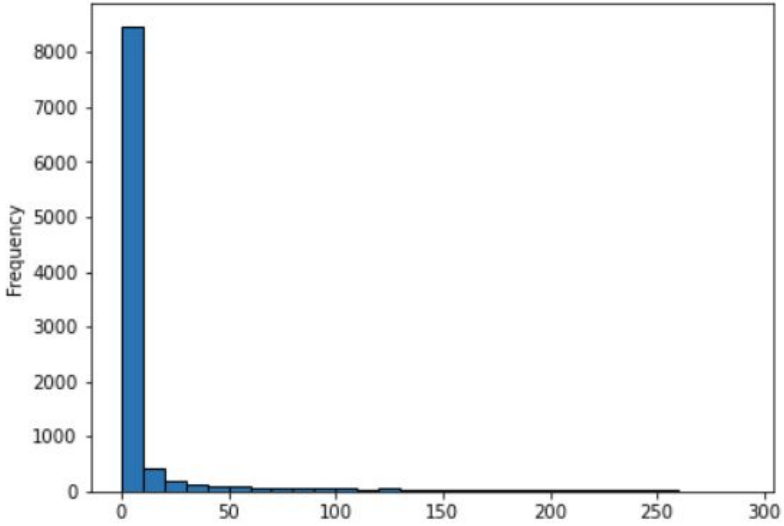
Our approaches:

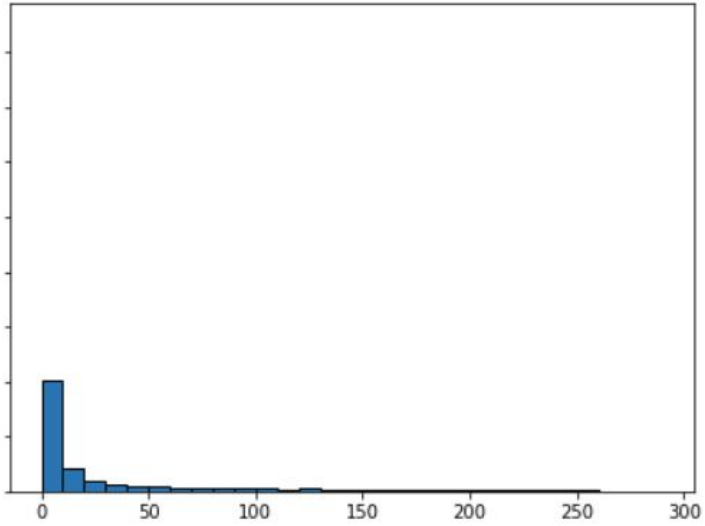
Undersampling lower values (0, 1, 2,...) then Log-transforming



There are more than 60% of zero values in the output of this dataset

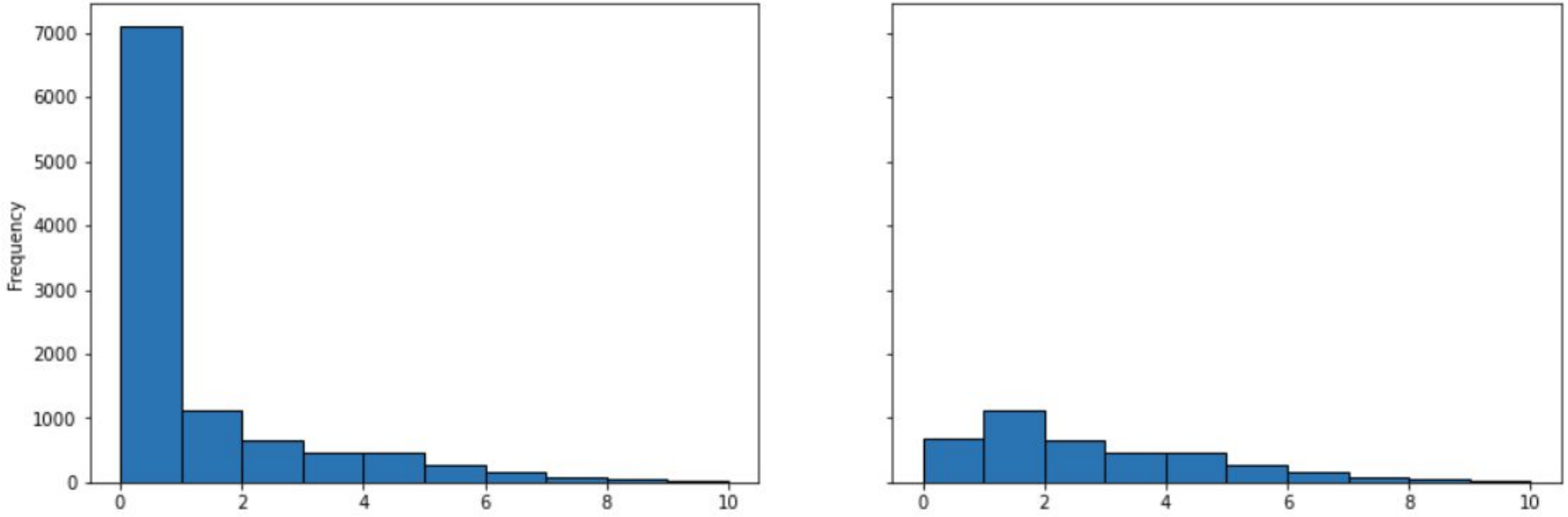
Initial exploration and preprocessing





Before and after undersampling (No transformation)

Initial exploration and preprocessing



Before and after undersampling (with Log-transformation)

Exploration (training set)

Inappropriate data types

- Type of *discount* is string instead of int ('1%', '15%',...)
- Some features have 'None' values instead of missing one
- Type of *show_official_shop_label_in_title* is string instead of boolean ('False', 'True')

Redundant features

- *flash_sale*, *upcoming_flash_sale* and *coin_earn_label* has 100% of missing values
- *is_adult* has 100% of 'False' values

Preprocessing (training set)

- Replacing ‘None’ with np.nan
- Dropping redundant features
- Dropping *name* for it has too many values, which can make model prone to be overfitting
- Selecting only top locations in *shop_location*, the remaining values are replaced with ‘Others’ because it has too many values (too)
- Replacing np.nan with 0 in *discount* because missing value in *discount* implies there is no discount at all, removing ‘%’ and convert to int
- Converting True and False to 0 and 1
- Filling numeric values with mean
- Filling categorical values with mode
- Encoding nominal values by One-hot technique
- Normalising data with z-score

HOW TO SELL WELL ON SHOPEE?

A Simple Data Science Project

Nguyen Nguyen Khang | Tran Dang Khoa

HOW TO SELL WELL ON SHOPEE?

> Overview
> Crawl data
> Data exploration & preprocessing
> Data Modeling
> Wrap Up

Modelling

Regression is a good choice to analyse dataset with continuous output

For simplicity, we used R-squared as measure for regression models (method *score* of scikit-learn computes R-squared value)

In this assignment, we trained two model:

- Linear Regression
- Neural Net Regression

Modelling

Linear Regression

- Score: 0.4948941558639116

Neural Net Regression

- Hyperparameter: hidden neuron = 70, solver = adam, max_iter = 10000

num_top_location/alpha	0.001	0.01	1
1	0.71005751	0.736797	0.75531786
3	0.75942068	0.69650857	**0.7625119**
5	0.61632231	0.57589607	0.74740973

Score table for each hyperparameter combination

Modelling

We chose Neural Net Regression to predict test set:

- Score: 0.7659745153630606
- Error rate: 23.40%

Neural Net is not only suitable for classification but also for regression. However it has some drawbacks:

- Too many hyperparameter
- Takes longer time to train

HOW TO SELL WELL ON SHOPEE?

A Simple Data Science Project

Nguyen Nguyen Khang | Tran Dang Khoa

HOW TO SELL WELL ON SHOPEE?

> Overview
> Crawl data
> Data exploration & preprocessing
> Data Modeling
> Wrap Up

100%

Wrap Up

- It is a flow from the idea to the model which we can use to predict
- Be aware of the target, it could be skewed

Thanks for watching