

HOW TO SELL WELL ON SHOPEE?

**A Simple Data Science
Project**

Table Of Contents

Overview

- What is this project?
- Why ?
- Dataset

Data Crawler

Some great tips to crawl data

Data exploration & preprocessing

Missing values, invalid value, wrong type features, skewed output, handle object type, scaling.

Data Modeling

How to find a good model to predict sales

Wrap up

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

- Sell well on the internet

to datasets

- Hi Shopee!

How to get it?

- I prefer to use API
- But shoppe have no public 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)

Kênh Người Bán | Tải ứng dụng | Kết nối

Shopee

thoi trang nam

Hoodie Nam Sandal Nữ Áo Nữ Dép Nam Balo Nữ Quần Nam Tất Nữ Váy Trẻ Em

Thời trang nam

Áo ngắn tay không cổ Quần Áo

BỘ LỌC TÌM KIẾM

Theo Danh Mục

- ☐ Áo ngắn tay không cổ (358k+)
- ☐ Quần (347k+)
- ☐ Áo khoác & Áo vest (224k+)
- ☐ Phụ kiện nam (187k+)

Thêm

Sắp xếp theo **Liên Quan** Mới Nhất Bán Chạy Giá

Nơi Bán

- ☐ Hà Nội
- ☐ TP. Hồ Chí Minh
- ☐ Thái Nguyên

<https://banhang.shopee.vn>

Shop liên quan đến "thoi trang nam"

thoi_trang_nam
thoi_trang_nam
2,4k Người Theo Dõi | 25 Đang Theo

Kết quả tìm kiếm cho từ khoá 'thoi trang nam'

Sắp xếp theo **Liên Quan** Mới Nhất Bán Chạy Giá

Yêu thích **50% GIẢM** **Yêu thích** **45% GIẢM** **Yêu thích**

Elements Console Sources **Network** Performance

Filter ☐ Hide data URLs All **XHR** JS CSS Img Media Font Doc WS Manifest Other

☐ Has blocked cookies ☐ Blocked Requests

50000 ms 100000 ms 150000 ms 200000 ms

Name	Status	Type	Initiator	Size	Time	Waterfall
vi.col2b.160630063/.json	200	fetch	bundle.503a...	(prefe...	215 ms	
vi.col96.1593685186.json	200	fetch	bundle.503a...	(prefe...	98 ms	
vi.col34.1600851804.json	200	fetch	bundle.503a...	(prefe...	99 ms	
?set_ids=2	200	fetch	bundle.503a...	(prefe...	456 ms	
vi.col24.1604371407.json	200	fetch	bundle.503a...	(prefe...	3 ms	
tr	200	fetch	vendors~bun...	145 B	183 ms	
get_all	304	fetch	bundle.503a...	894 B	525 ms	
log?key=AlzaSyCx80ru6-R...	200	fetch	firebase.c932...	143 B	2.30 s	
t	200	fetch	bundle.503a...	102 B	239 ms	
tr	200	fetch	vendors~bun...	144 B	66 ms	
tr	200	fetch	vendors~bun...	145 B	219 ms	

55 / 169 requests 113 kB / 1.4 MB transferred 742 kB / 4.2 MB resources

Console What's New

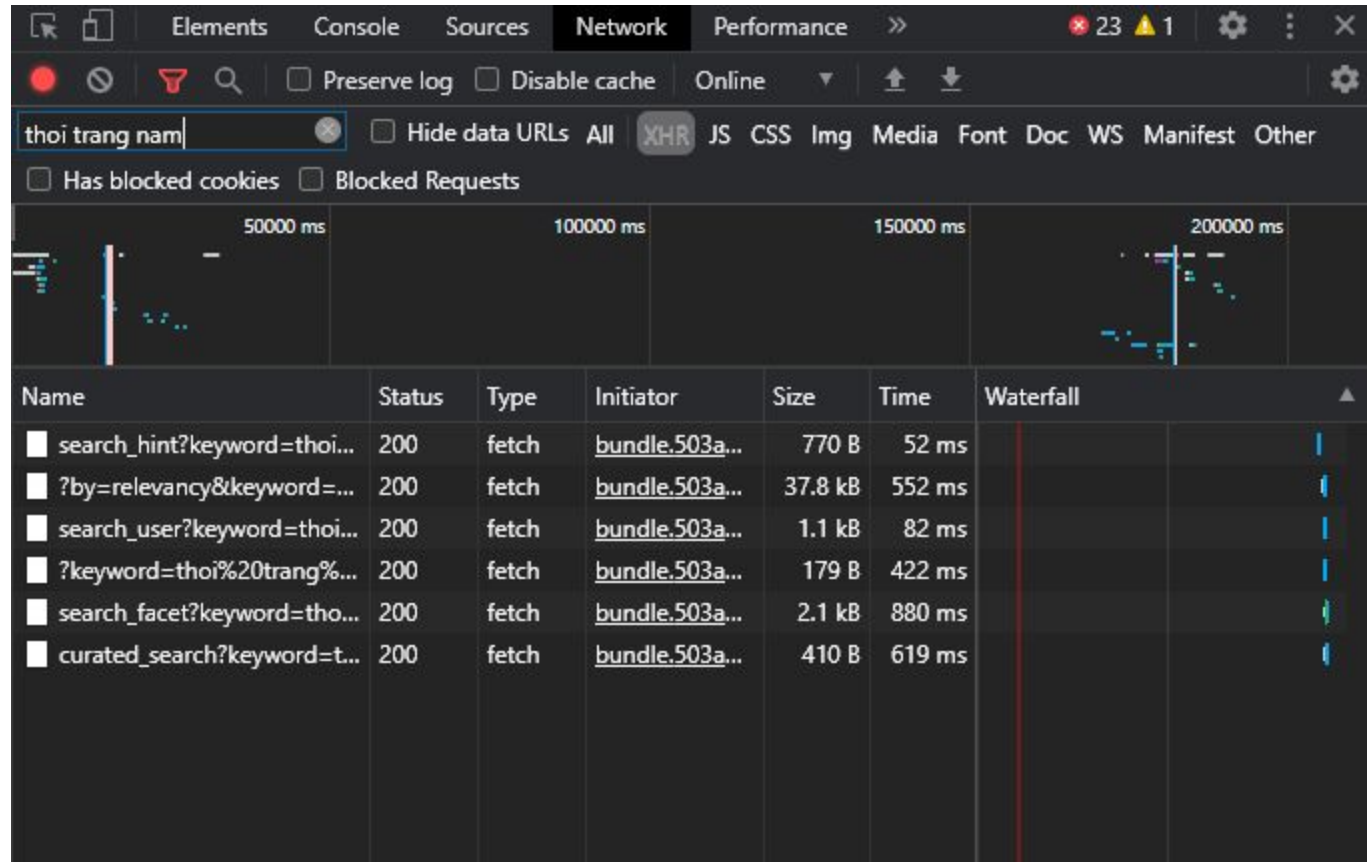
Highlights from the Chrome 87 update

New CSS Grid debugging tools
Debug and inspect CSS Grid with the new CSS Grid debugging tools.

New WebAuthn tab
Emulate authenticators and debug the Web Authentication API with the new WebAuthn tab.

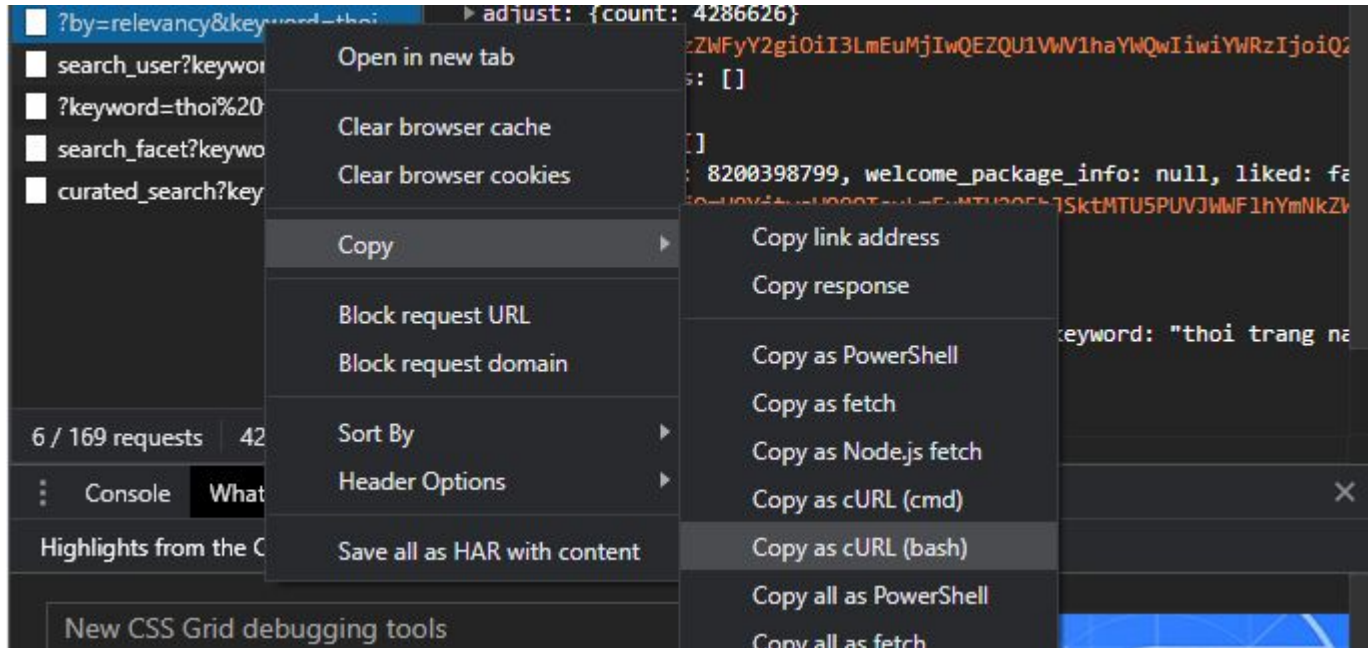
Activate Windows
Go to Settings to activate Windows.

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



Name	Headers	Preview	Response	Initiator	Timing	Cookies
search_hint?keyword=thoi%20t...			adjust: {count: 4280020}			
?by=relevancy&keyword=thoi...			algorithm: "eyJzZWYyZgiOiI3LmEuMjIwQEZQU1VWV1haYWQwIiw1YWRzIjoiQ2			
search_user?keyword=thoi%20t...			disclaimer_infos: []			
?keyword=thoi%20trang%20nam			error: null			
search_facet?keyword=thoi%20...			hint_keywords: []			
curated_search?keyword=thoi%...			▼ items: [{itemid: 8200398799, welcome_package_info: null, liked: fa			
			▶ 0: {itemid: 8200398799, welcome_package_info: null, liked: false			
			▶ 1: {itemid: 9505975325, welcome_package_info: null, liked: false			
			▶ 2: {itemid: 1417601269, welcome_package_info: null, liked: false			
			▶ 3: {itemid: 7770387249, welcome_package_info: null, liked: false			
			▶ 4: {itemid: 7045148454, welcome_package_info: null, liked: false			
			▶ 5: {itemid: 4101478378, welcome_package_info: null, liked: false			
			▶ 6: {itemid: 1687034671, welcome_package_info: null, liked: false			
			▶ 7: {itemid: 1492650885, welcome_package_info: null, liked: false			
			▶ 8: {itemid: 3467039516, welcome_package_info: null, liked: false			
			▶ 9: {itemid: 3052604433, welcome_package_info: null, liked: false			
6 / 169 requests			42.3 kB / 1.4 MB			

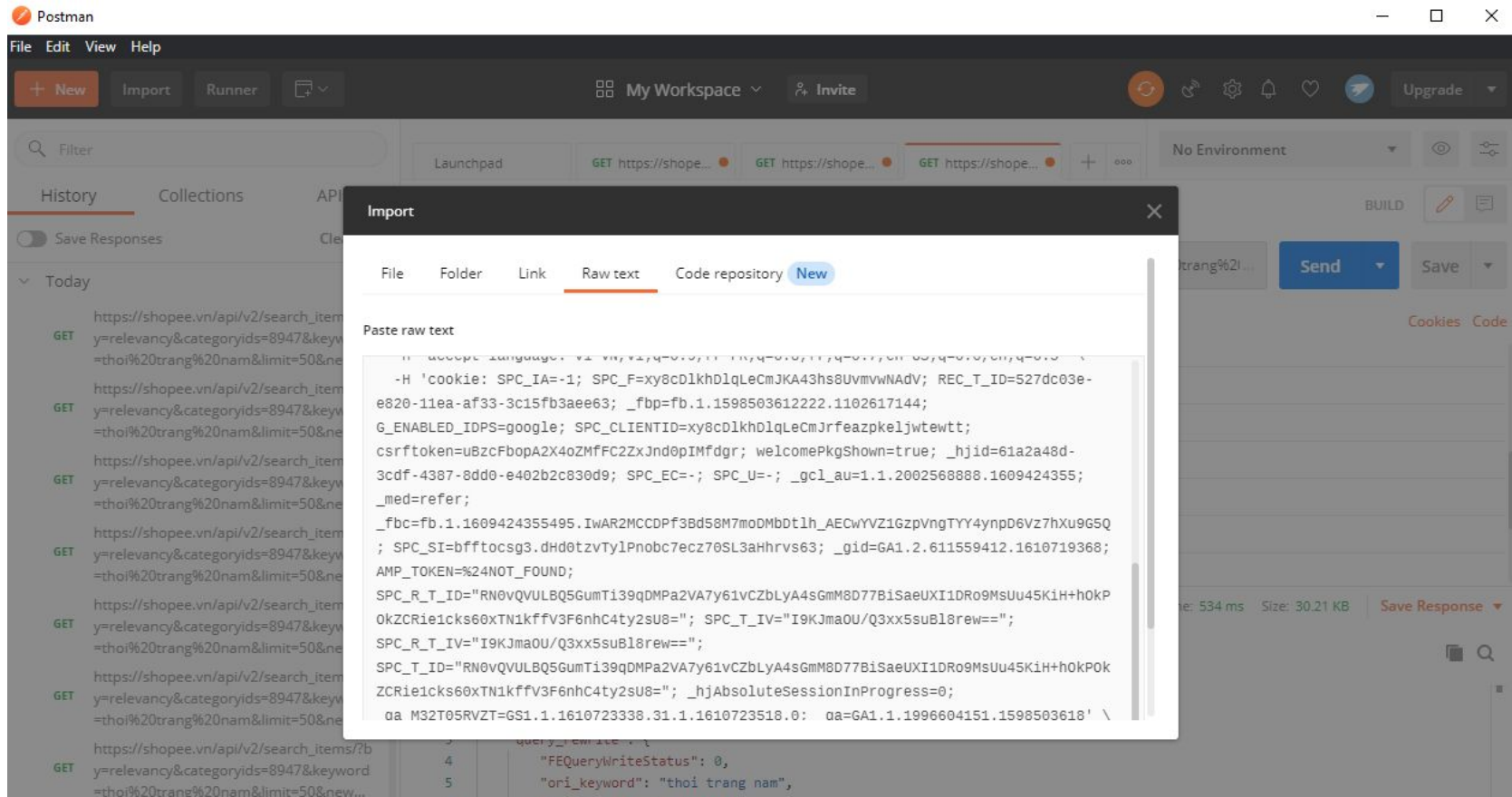
It seems like the call we need.



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%20nam

Send

Save

Params

Authorization

Headers (23)

Body

Pre-request Script

Tests

Settings

Cookies Code

Query Params

	KEY	VALUE	DESCRIPTION	...	Bulk Edit
<input checked="" type="checkbox"/>	by	relevancy			
<input checked="" type="checkbox"/>	categoryids	8947			
<input checked="" type="checkbox"/>	keyword	thoi%20trang%20nam			
<input checked="" type="checkbox"/>	limit	50			
<input checked="" type="checkbox"/>	offset	50			

Result:

The screenshot shows a web browser's developer tools interface. At the top, a checkbox is checked next to the word "newest", and the number "50" is displayed. Below this, the "Body" tab is selected, showing a JSON response. The status is "200 OK", the time is "534 ms", and the size is "30.21 KB". A "Save Response" button is visible. The JSON response is displayed in a "Pretty" format, showing a list of items. The first item is a JSON object with the following properties: "has_group_buy_stock": null, "match_type": null, "preview_info": null, "welcome_package_type": 0, "exclusive_price_info": null, "name": "áo thời trang nam", "distance": null, and "adsid": null. The response is displayed in a table with line numbers 127 to 134. At the bottom of the browser window, there is a "Bootcamp" button and a "Build" button. A "Browse" button is also visible. A watermark "Activate Windows" is present in the bottom right corner.

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,
```

Activate Windows
Go to Settings to activate Windows



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.



Untitled Request BUILD  

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

Params ● Authorization Headers (23) Body Pre-request Script Tests Settings Cookies Code





KEY	VALUE	DESCRIPTION	***	Bulk Edit	Presets
<input type="checkbox"/> authority	shopee.vn				
<input type="checkbox"/> sec-ch-ua	"Google Chrome";v="87", " Not;A Brand";v="99",...				
<input type="checkbox"/> x-shopee-language	vi				
<input type="checkbox"/> x-requested-with	XMLHttpRequest				
<input type="checkbox"/> if-none-match-	55b03-3fd0f4bf7f1503951d3409cbf31998e0				
<input type="checkbox"/> sec-ch-ua-mobile	?0				

Body Cookies (11) Headers (24) Test Results  Status: 200 OK Time: 15.12 s Size: 2.17 KB Save Response ▼

Pretty Raw Preview Visualize HTML ▼  

1

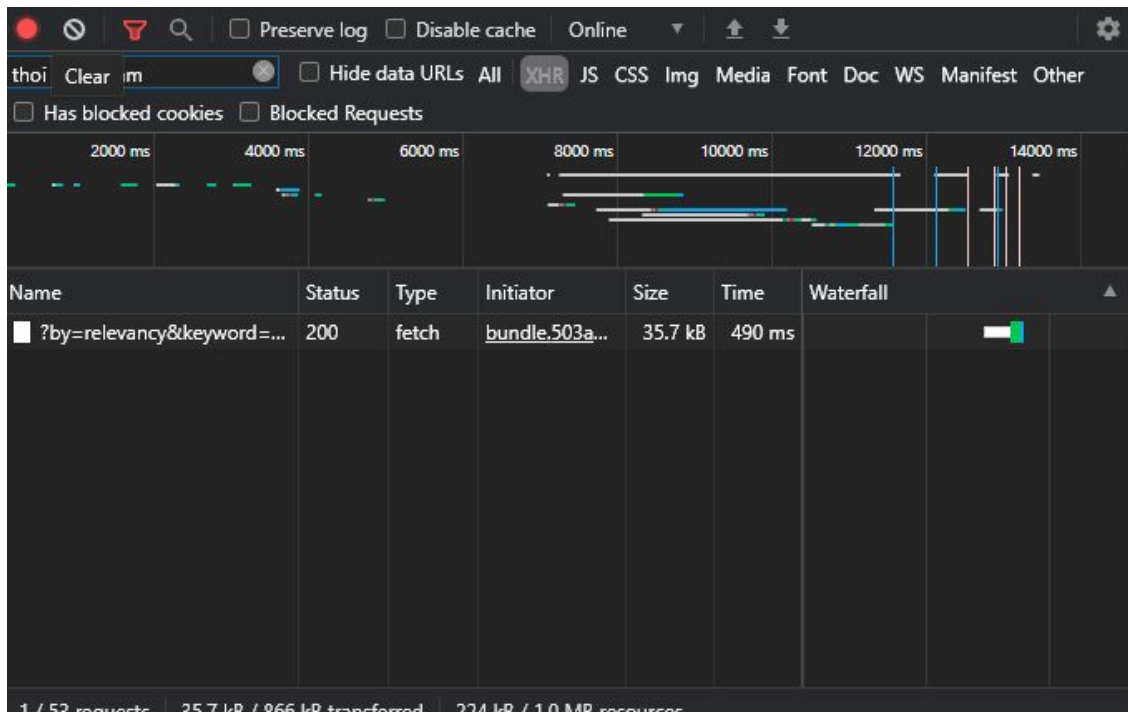
Activate Windows
Go to Settings to activate Windows.

Bootcamp Build Browse    

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,...

The screenshot shows a Postman interface with a GET request to the Shopee API. The request parameters are: by=relevancy, categoryids=8947, keyword=thoi%20trang%20nam, limit=50, newest=150, and order=desc. The response status is 200 OK, with a time of 2.43 s and a size of 27.79 KB. The response body is displayed in JSON format, showing product details for a shirt.

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

Params Authorization Headers (23) Body Pre-request Script Tests Settings Cookies Code

KEY	VALUE	DESCRIPTION
<input checked="" type="checkbox"/> by	relevancy	
<input checked="" type="checkbox"/> categoryids	8947	
<input checked="" type="checkbox"/> keyword	thoi%20trang%20nam	
<input checked="" type="checkbox"/> limit	50	
<input checked="" type="checkbox"/> newest	150	
<input checked="" type="checkbox"/> order	desc	

Body Cookies (12) Headers (20) Test Results Status: 200 OK Time: 2.43 s Size: 27.79 KB Save Response

Pretty Raw Preview Visualize JSON

```
141 "preview_info": null,  
142 "welcome_package_type": 0,  
143 "exclusive_price_info": null,  
144 "name": "Áo Thun Bông Rổ Ngắn Tay Dáng Rộng Phối Lớp Ngoài Thời Trang Nam Cá Tính",  
145 "distance": null,  
146 "adsid": null,  
147 "ctime": 1610175346,  
148 "wholesale_tier_list": [],
```

Activate Windows
Go to Settings to activate Windows.

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:

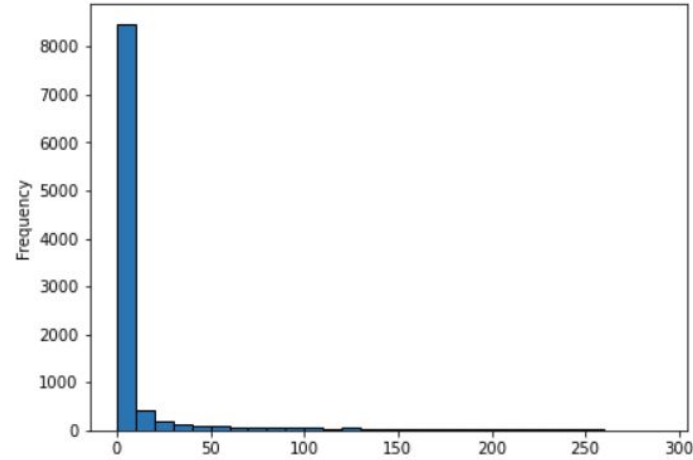
- Predictions of trained model are more accurate for lower values of output
- Skewed distribution converge much slower than symmetric one

Some approaches:

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

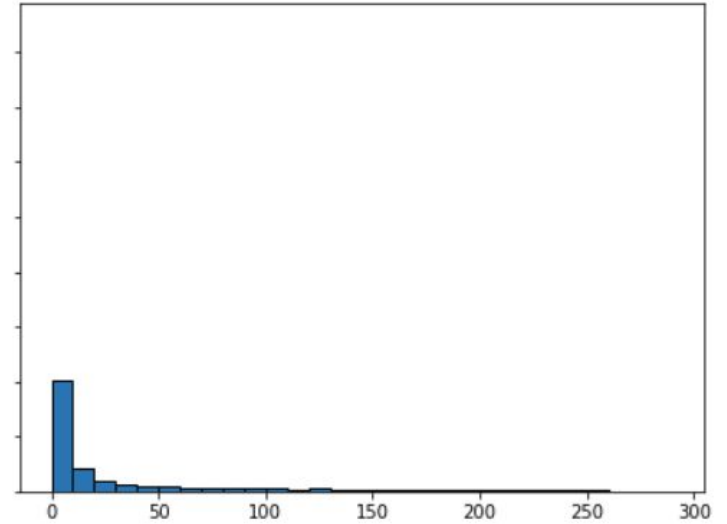
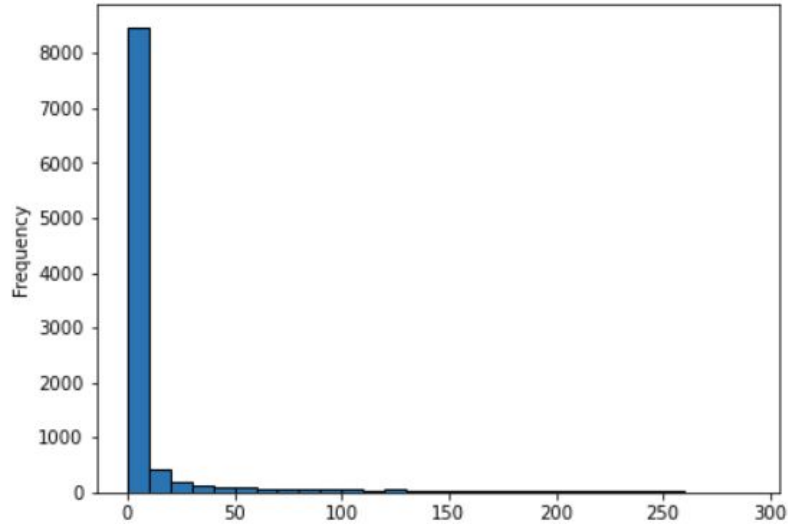
Our approaches:

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



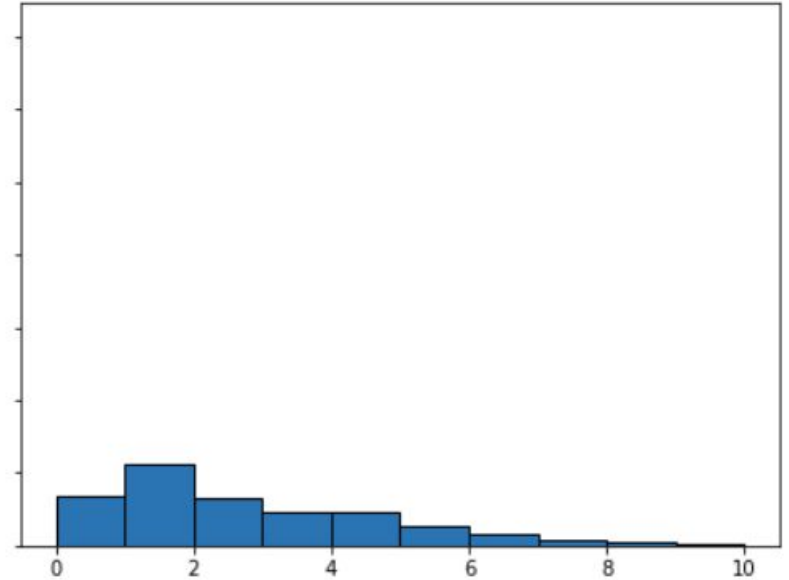
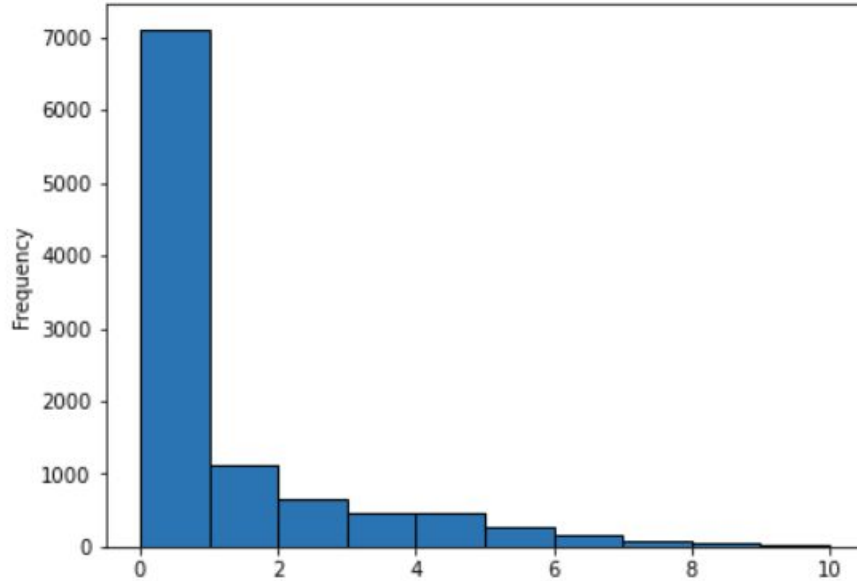
There are more than 60% of zero values in the output of this dataset

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