



**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**  
**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**  
**KHOA CÔNG NGHỆ THÔNG TIN**

**BÁO CÁO QUÁ TRÌNH LÀM VIỆC NHÓM**  
**MÔN: NHẬP MÔN KHOA HỌC DỮ LIỆU**

| Giáo viên hướng dẫn |

**Trần Trung Kiên**

| Sinh viên thực hiện |

**Trần Thanh Tùng - 18120258**

**Trần Hữu Chí Bảo – 18120288**

**Thành phố Hồ Chí Minh - 2021**

---

## MỤC LỤC

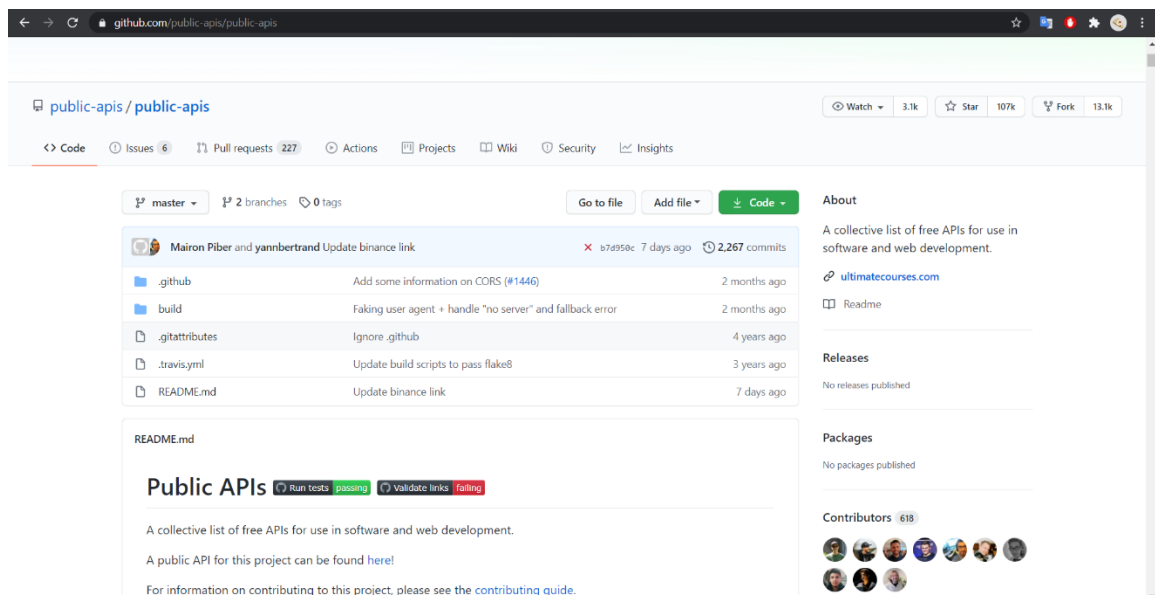
---

MỤC LỤC .....	2
I. CÁC MỐC THỜI GIAN .....	3
1. Từ ngày 12-12 tới ngày 26-12: .....	3
2. Từ ngày 27-12 tới ngày 2-1:.....	6
3. Từ ngày 3-1:.....	6
II. NHÌN LẠI QUÁ TRÌNH LÀM ĐỒ ÁN.....	8
III. TÀI LIỆU THAM KHẢO.....	9

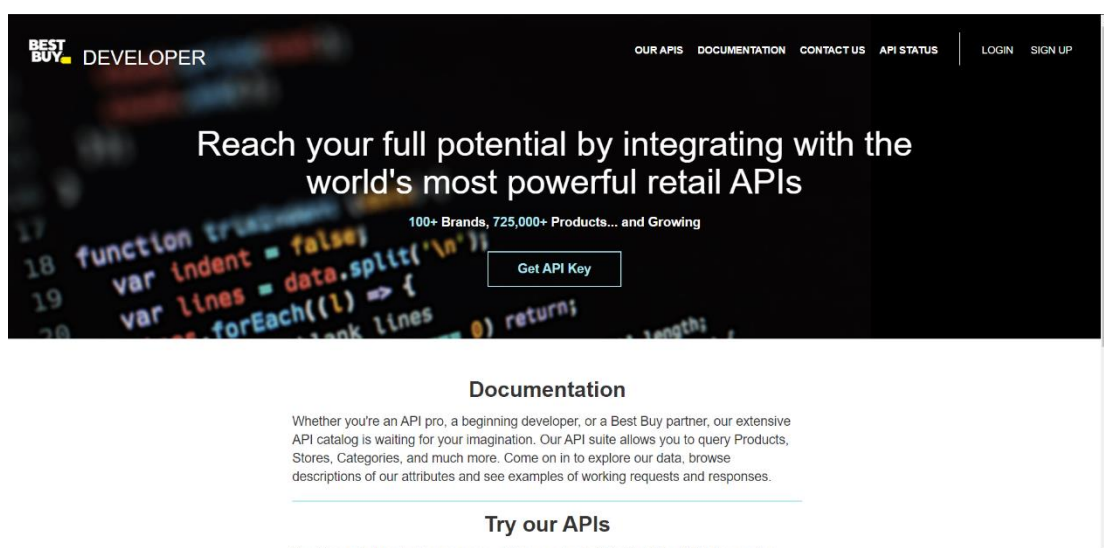
# I. CÁC MỐC THỜI GIAN

## 1. Từ ngày 12-12 tới ngày 26-12:

- Tìm hiểu các nguồn dữ liệu từ internet, đầu tiên vào link tổng hợp các public API trên github : <https://github.com/public-apis/public-apis>.




- Đầu tiên định làm đề tài shopping, thu thập sản phẩm trên các trang thương mại điện tử bestbuy.com hoặc eBay.com.
- Tuy nhiên bestbuy.com yêu cầu API key mà API key chỉ có thể có qua việc đăng ký tài khoản



- Nhưng bestbuy lại không cho đăng ký tài khoản:

### Create an Account



Sorry, something went wrong.  
Please try again.

Trần


Tùng


ianmoonepro@gmail.com


Show Passwords

.....

.....

 Your passwords match!

239-429-2420 



Create an Account

- Quyết định chuyển sang ebay.com. Ebay không yêu cầu API key, tuy nhiên Ebay chỉ là 1 cái “chợ” ở đó người dùng đăng sản phẩm họ cần bán và tự đặt mức giá à dữ liệu không đáng tin cậy.

Chuyển hướng đề tài sang thị trường tiền ảo, dùng api  
<https://api.exchange.bitcoin.com/> để thu thập thông tin về thị trường bitcoin.

api.exchange.bitcoin.com/api2/public/trades/BTCUSD

[{"id":1067124014,"price":"39214.53","quantity":"0.00319","side":"buy","timestamp":"2021-01-09T06:18:46.227Z"}, {"id":1067124006,"price":"39213.94","quantity":"0.03960","side":"buy","timestamp":"2021-01-09T06:18:45.891Z"}, {"id":1067123992,"price":"39188.95","quantity":"0.21500","side":"buy","timestamp":"2021-01-09T06:18:44.858Z"}, {"id":1067123971,"price":"39188.42","quantity":"0.00059","side":"buy","timestamp":"2021-01-09T06:18:43.604Z"}, {"id":1067123968,"price":"39184.20","quantity":"0.00100","side":"sell","timestamp":"2021-01-09T06:18:43.516Z"}, {"id":1067123942,"price":"39219.18","quantity":"0.01109","side":"buy","timestamp":"2021-01-09T06:18:41.890Z"}, {"id":1067123853,"price":"39214.36","quantity":"0.00017","side":"buy","timestamp":"2021-01-09T06:18:37.203Z"}, {"id":1067123844,"price":"39197.98","quantity":"0.00500","side":"sell","timestamp":"2021-01-09T06:18:36.975Z"}, {"id":1067123822,"price":"39235.29","quantity":"0.00025","side":"sell","timestamp":"2021-01-09T06:18:36.361Z"}, {"id":1067123816,"price":"39228.20","quantity":"0.04600","side":"sell","timestamp":"2021-01-09T06:18:36.210Z"}, {"id":1067123815,"price":"39235.28","quantity":"0.00001","side":"sell","timestamp":"2021-01-09T06:18:36.213Z"}, {"id":1067123814,"price":"39235.28","quantity":"0.01000","side":"sell","timestamp":"2021-01-09T06:18:36.205Z"}, {"id":1067123813,"price":"39235.28","quantity":"0.02206","side":"sell","timestamp":"2021-01-09T06:18:36.180Z"}, {"id":1067123812,"price":"39235.57","quantity":"0.01002","side":"sell","timestamp":"2021-01-09T06:18:36.180Z"}, {"id":1067123811,"price":"39240.29","quantity":"0.00025","side":"sell","timestamp":"2021-01-09T06:18:36.180Z"}, {"id":1067123810,"price":"39240.30","quantity":"0.00004","side":"sell","timestamp":"2021-01-09T06:18:36.180Z"}, {"id":1067123809,"price":"39247.06","quantity":"0.00000","side":"sell","timestamp":"2021-01-09T06:18:36.180Z"}, {"id":1067123767,"price":"39267.94","quantity":"0.00005","side":"buy","timestamp":"2021-01-09T06:18:32.586Z"}, {"id":1067123722,"price":"39267.94","quantity":"0.03960","side":"sell","timestamp":"2021-01-09T06:18:28.560Z"}, {"id":1067123681,"price":"39274.36","quantity":"0.03960","side":"buy","timestamp":"2021-01-09T06:18:25.890Z"}, {"id":1067123648,"price":"39250.03","quantity":"0.00100","side":"sell","timestamp":"2021-01-09T06:18:23.583Z"}, {"id":1067123596,"price":"39297.45","quantity":"0.00059","side":"buy","timestamp":"2021-01-09T06:18:20.718Z"}, {"id":1067123583,"price":"39297.45","quantity":"0.03240","side":"buy","timestamp":"2021-01-09T06:18:20.400Z"}, {"id":1067123575,"price":"39273.58","quantity":"0.01586","side":"buy","timestamp":"2021-01-09T06:18:20.046Z"}, {"id":1067123574,"price":"39271.58","quantity":"0.00794","side":"buy","timestamp":"2021-01-09T06:18:20.046Z"}, {"id":1067123573,"price":"39271.57","quantity":"0.01620","side":"buy","timestamp":"2021-01-09T06:18:20.045Z"}, {"id":1067123572,"price":"39271.58","quantity":"0.09206","side":"buy","timestamp":"2021-01-09T06:18:20.015Z"}, {"id":1067123570,"price":"39273.58","quantity":"0.00174","side":"buy","timestamp":"2021-01-09T06:18:20.015Z"}, {"id":1067123559,"price":"39254.77","quantity":"0.22000","side":"buy","timestamp":"2021-01-09T06:18:19.988Z"}, {"id":1067123559,"price":"39253.78","quantity":"0.00174","side":"buy","timestamp":"2021-01-09T06:18:19.287Z"}, {"id":1067123558,"price":"39252.39","quantity":"0.02360","side":"sell","timestamp":"2021-01-09T06:18:19.220Z"}, {"id":1067123554,"price":"39258.74","quantity":"0.01620","side":"sell","timestamp":"2021-01-09T06:18:19.002Z"}, {"id":1067123550,"price":"39266.14","quantity":"0.10671","side":"sell","timestamp":"2021-01-09T06:18:18.722Z"}, {"id":1067123548,"price":"39270.86","quantity":"0.01620","side":"sell","timestamp":"2021-01-09T06:18:18.722Z"}, {"id":1067123547,"price":"39270.86","quantity":"0.06000","side":"sell","timestamp":"2021-01-09T06:18:18.722Z"}, {"id":1067123546,"price":"39273.58","quantity":"0.04600","side":"sell","timestamp":"2021-01-09T06:18:18.722Z"}, {"id":1067123536,"price":"39300.00","quantity":"0.37760","side":"sell","timestamp":"2021-01-09T06:18:22.292Z"}, {"id":1067123535,"price":"39301.60","quantity":"0.27840","side":"sell","timestamp":"2021-01-09T06:18:18.296Z"}, {"id":1067123534,"price":"39302.12","quantity":"0.19540","side":"sell","timestamp":"2021-01-09T06:18:18.296Z"}, {"id":1067123533,"price":"39304.17","quantity":"0.00860","side":"sell","timestamp":"2021-01-09T06:18:18.284Z"}, {"id":1067123532,"price":"39304.17","quantity":"0.20000","side":"sell","timestamp":"2021-01-09T06:18:18.284Z"}, {"id":1067123519,"price":"39305.52","quantity":"0.06812","side":"sell","timestamp":"2021-01-09T06:18:18.035Z"}, {"id":1067123518,"price":"39306.77","quantity":"0.15470","side":"sell","timestamp":"2021-01-09T06:18:18.035Z"}, {"id":1067123517,"price":"39307.14","quantity":"0.11500","side":"sell","timestamp":"2021-01-09T06:18:18.035Z"}, {"id":1067123516,"price":"39307.14","quantity":"0.13460","side":"sell","timestamp":"2021-01-09T06:18:18.035Z"}, {"id":1067123515,"price":"39307.14","quantity":"0.12880","side":"sell","timestamp":"2021-01-09T06:18:18.035Z"}, {"id":1067123514,"price":"39307.14","quantity":"0.06000","side":"sell","timestamp":"2021-01-09T06:18:18.035Z"}, {"id":1067123513,"price":"39315.88","quantity":"0.00025","side":"sell","timestamp":"2021-01-09T06:18:18.035Z"}, {"id":1067123504,"price":"39327.00","quantity":"0.03240","side":"sell","timestamp":"2021-01-09T06:17:17.712Z"}, {"id":1067123434,"price":"39292.42","quantity":"0.03111","side":"sell","timestamp":"2021-01-09T06:18:12.474Z"}, {"id":1067123402,"price":"39282.47","quantity":"0.03240","side":"buy","timestamp":"2021-01-09T06:18:08.740Z"}, {"id":1067123355,"price":"39299.35","quantity":"0.02575","side":"buy","timestamp":"2021-01-09T06:18:05.687Z"}, {"id":1067123352,"price":"39294.43","quantity":"0.01405","side":"buy","timestamp":"2021-01-09T06:18:05.675Z"}, {"id":1067123350,"price":"39272.51","quantity":"0.04600","side":"buy","timestamp":"2021-01-09T06:18:05.273Z"}, {"id":1067123325,"price":"39256.14","quantity":"0.00097","side":"buy","timestamp":"2021-01-09T06:17:59.688Z"}, {"id":1067123244,"price":"39275.19","quantity":"0.06843","side":"buy","timestamp":"2021-01-09T06:17:59.633Z"}, {"id":1067123243,"price":"39261.10","quantity":"0.00001","side":"buy","timestamp":"2021-01-09T06:17:59.633Z"}, {"id":1067123236,"price":"39257.58","quantity":"0.00509","side":"sell","timestamp":"2021-01-09T06:17:59.403Z"}, {"id":1067123232,"price":"39259.95","quantity":"0.01405","side":"sell","timestamp":"2021-01-09T06:17:59.103Z"}, {"id":1067123231,"price":"39261.60","quantity":"0.00001","side":"sell","timestamp":"2021-01-09T06:17:59.103Z"}, {"id":1067123222,"price":"39272.76","quantity":"0.04346","side":"sell","timestamp":"2021-01-09T06:17:58.879Z"}, {"id":1067123221,"price":"39272.76","quantity":"0.00254","side":"sell","timestamp":"2021-01-09T06:17:58.879Z"}, {"id":1067123219,"price":"39294.31","quantity":"0.01465","side":"sell","timestamp":"2021-01-09T06:17:58.783Z"}, {"id":1067123218,"price":"39301.53","quantity":"0.00471","side":"sell","timestamp":"2021-01-09T06:17:58.783Z"}, {"id":1067123217,"price":"39302.11","quantity":"0.00116","side":"sell","timestamp":"2021-01-09T06:17:58.783Z"}, {"id":1067123158,"price":"39308.94","quantity":"0.00179","side":"sell","timestamp":"2021-01-09T06:17:54.389Z"}, {"id":1067123103,"price":"39327.07","quantity":"0.03860","side":"sell","timestamp":"2021-01-09T06:17:52.333Z"}, {"id":1067122943,"price":"39366.83","quantity":"0.00001","side":"buy","timestamp":"2021-01-09T06:17:28.252Z"}, {"id":1067122757,"price":"39336.44","quantity":"0.03960","side":"sell","timestamp":"2021-01-09T06:17:21.176Z"}, {"id":1067122735,"price":"39377.04","quantity":"0.03240","side":"sell","timestamp":"2021-01-09T06:17:25.406Z"}, {"id":1067122721,"price":"39380.20","quantity":"0.03941","side":"buy","timestamp":"2021-01-09T06:17:23.242Z"}, {"id":1067122693,"price":"39370.53","quantity":"0.00033","side":"buy","timestamp":"2021-01-09T06:17:23.242Z"}, {"id":1067122648,"price":"39331.25","quantity":"0.00153","side":"sell","timestamp":"2021-01-09T06:17:18.754Z"}, {"id":1067122644,"price":"39331.60","quantity":"0.02091","side":"sell","timestamp":"2021-01-09T06:17:18.751Z"}, {"id":1067122636,"price":"39329.75","quantity":"0.00833","side":"sell","timestamp":"2021-01-09T06:17:18.476Z"}, {"id":1067122627,"price":"39331.51","quantity":"0.00030","side":"sell","timestamp":"2021-01-09T06:17:18.106Z"}, {"id":1067122482,"price":"39362.72","quantity":"0.03240","side":"buy","timestamp":"2021-01-09T06:17:07.515Z"}, {"id":1067122472,"price":"39357.00","quantity":"0.02985","side":"buy","timestamp":"2021-01-09T06:17:06.691Z"}, {"id":1067122465,"price":"39355.81","quantity":"0.05309","side":"buy","timestamp":"2021-01-09T06:17:06.243Z"}, {"id":1067122464,"price":"39349.93","quantity":"0.04600","side":"buy","timestamp":"2021-01-09T06:17:06.243Z"}, {"id":1067122456,"price":"39329.07","quantity":"0.03960","side":"buy","timestamp":"2021-01-09T06:17:05.839Z"}, {"id":1067122448,"price":"39326.71","quantity":"0.00004","side":"buy","timestamp":"2021-01-09T06:17:05.111Z"}, {"id":1067122419,"price":"39373.22","quantity":"0.03732","side":"buy","timestamp":"2021-01-09T06:17:03.472Z"}, {"id":1067122346,"price":"39311.57","quantity":"0.22000","side":"sell","timestamp":"2021-01-09T06:17:02.028Z"}, {"id":1067122344,"price":"39315.86","quantity":"0.03115","side":"sell","timestamp":"2021-01-09T06:16:58.887Z"}, {"id":1067122343,"price":"39315.87","quantity":"0.02339","side":"sell","timestamp":"2021-01-09T06:16:58.712Z"}]

- Tuy nhiên không đưa ra được câu hỏi từ dữ liệu này.
- Tiếp tục chuyển hướng sang nhiều lĩnh vực khác như thời tiết, bóng đá, game nhưng cũng không đưa được ra câu hỏi vừa ý.
- Chuyển sang làm về nông nghiệp, tuy nhiên rất nhiều dữ liệu nông nghiệp phải mua chứ không miễn phí.
- Vô tình tìm được website ricepedia.org, có tổng hợp khá chi tiết và đầy đủ dữ liệu về tình hình sản xuất gạo của các quốc gia thành các bảng:

Not secure | ricepedia.org/vietnam

Basic Statistics

Element	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979	1980	19
Arable land (000 Ha)	5,550.00	5,550.00	5,550.00	5,550.00	5,550.00	5,550.00	5,570.00	5,590.00	5,600.00	5,630.00	5,630.00	5,650.00	5,680.00	5,700.00	5,700.00	5,900.00	5,980.00	5,999.00	5,970.00	5,940.00	5.9
Rice area (000 Ha)	4,744.00	4,888.86	4,496.52	4,987.80	4,826.30	4,681.30	4,795.80	4,893.80	4,930.00	4,724.40	4,692.10	4,900.00	5,030.00	5,111.92	4,855.90	5,297.30	5,468.70	5,462.50	5,485.20	5,600.20	5.6
Paddy yield (t/ha)	1.90	1.99	2.14	1.94	1.94	1.81	1.92	1.71	1.79	2.15	2.23	2.19	2.21	2.16	2.12	2.23	1.94	1.79	2.07	2.08	
Paddy Production (000 t)	8,997.40	9,747.04	9,622.67	9,697.03	9,369.70	8,463.50	9,188.40	8,366.15	8,815.00	10,173.30	10,447.00	10,748.20	11,125.00	11,023.29	10,293.60	11,827.20	10,597.10	9,789.90	11,362.90	11,647.40	12.4
Milled production (000 t)	6,001.27	6,501.28	6,418.32	6,467.92	6,249.59	5,645.15	6,128.66	5,580.22	5,879.60	6,785.59	6,968.15	7,169.05	7,420.38	7,352.53	6,965.83	7,888.74	7,068.27	6,529.86	7,579.05	7,768.82	8.2
Rice imports (000 t)	18.50	211.50	450.00	420.00	329.59	894.19	1,250.00	1,230.00	1,030.00	1,260.00	690.00	880.00	850.00	910.00	350.00	147.70	197.00	70.00	250.00	201.40	
Rice exports (000 t)	182.25	89.76	229.32	59.58	2.96	12.52	3.44	2.36	20.08	18.48	6.00	3.00	2.00	2.00	22.00	5.60	4.00	17.00	5.00	33.30	
Total rice consumption (000 t)	5,833.23	6,256.78	6,336.93	6,529.92	6,584.77	6,631.98	6,874.22	6,839.97	6,977.51	7,443.89	7,516.12	7,709.83	7,982.71	8,181.67	8,228.56	8,342.41	7,618.28	6,683.86	7,598.67	8,326.06	8.7
Fertilizer usage (NPK) (000 t)	16.04	17.12	18.92	16.58	14.05	8.29	20.91	20.30	49.21	55.29	51.10	45.69	43.73	49.79	57.89	45.93	70.17	56.89	27.34	26.12	

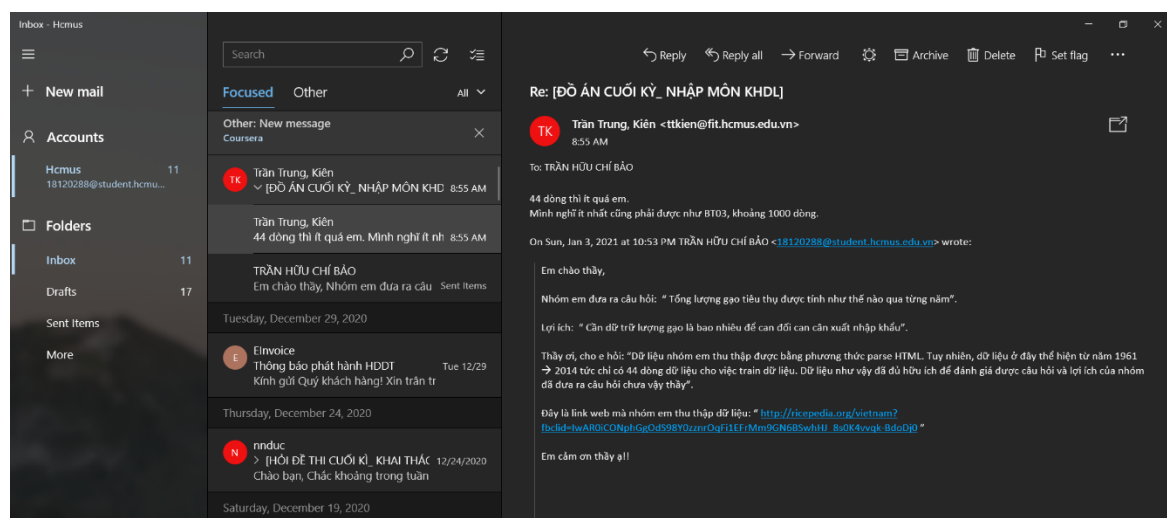
Quyết định làm về đề tài sản xuất gạo ở Việt Nam và parse trang <http://ricepedia.org/vietnam> để lấy dữ liệu về.

## 2. Từ ngày 27-12 tới ngày 2-1:

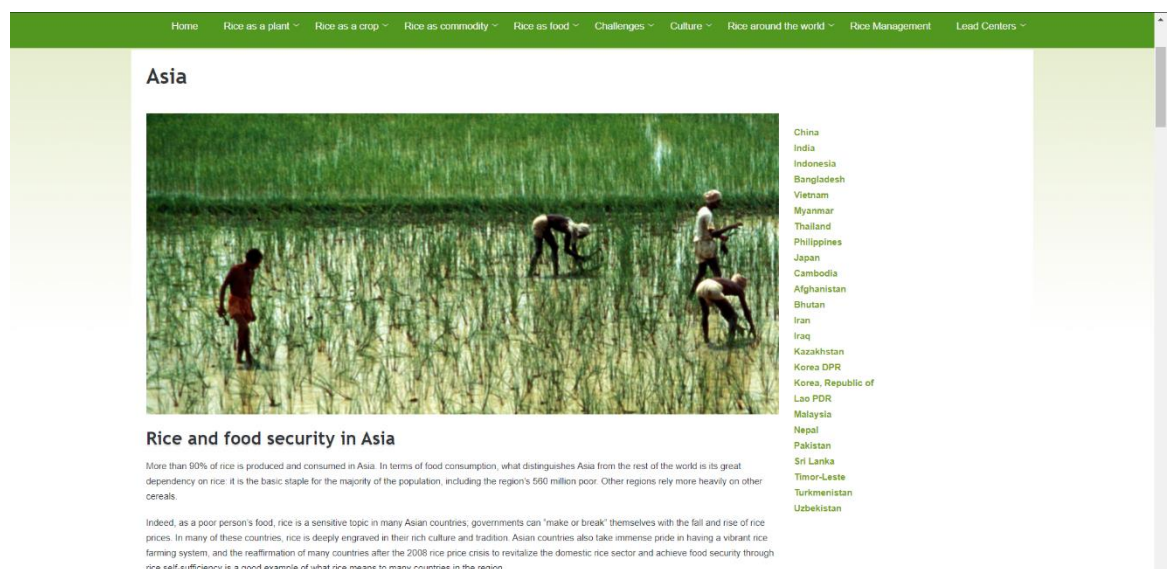
- Áp dụng quy trình khoa học dữ liệu đã được học, tiền xử lý dữ liệu raw ban đầu parse được. Đồng thời thử nghiệm một số mô hình để dự đoán:
  - Linear Regression
  - MLP

## 3. Từ ngày 3-1:

- Ghi nhận góp ý của thầy:



Thay đổi đề tài từ sản xuất gạo ở Việt Nam thành sản xuất gạo ở các quốc gia châu Á và mỹ latin. Sau đó parse dữ liệu của các quốc gia này trên ricepedia về làm.





- countryeconomy.com
Countries ▾ Data ▾ Reports ▾
€ \$ Search
Follow us

GDP VIETNAM 2018

## GDP improves in Vietnam

**Gross Domestic Product of Vietnam** grew 7.1% in 2018 compared to last year. This rate is 2-tenths of one percent higher than the figure of 6.9% published in 2017.

The GDP figure in 2018 was \$241,272 million, Vietnam is number 49 in the ranking of GDP of the 196 countries that we publish. The absolute value of GDP in Vietnam rose \$20,896 million with respect to 2017.

The **GDP per capita of Vietnam in 2018** was \$2,525, \$195 higher than in 2017, it was \$2,330. To view the evolution of the GDP per capita, it is interesting to look back a few years and compare these data with those of 2008 when the GDP per capita in Vietnam was \$1,154.

If we order the countries according to their GDP per capita, **Vietnam** is in 139th position, its population has a low level of affluence compare to the 196 countries whose GDP we publish.

Here we show you the progression of the GDP in Vietnam. You can see GDP in other countries in GDP and see all the economic information about Vietnam in Vietnam's economy.

**LATEST UPDATED DATA**
  - Euro zone: Unemployment Rate
  - Germany: Unemployment Rate

Evolution: Annual GDP Vietnam				Evolution: GDP per capita Vietnam			
Date	Annual GDP		GDP Growth (%)	Date	GDP per capita		GDP P.C. Annual Growth
2018	241.272M \$		7.1%	2018	2.525\$		8.4%
2017	220.376M \$		6.9%	2017	2.330\$		7.3%
2016	201.326M \$		6.7%	2016	2.172\$		4.1%
2015	191.288M \$		7.0%	2015	2.086\$		1.9%
2014	185.759M \$		6.4%	2014	2.047\$		7.8%
2013	170.444M \$		5.6%	2013	1.899\$		8.5%
2012	155.483M \$		5.5%	2012	1.751\$		14.3%
2011	134.598M \$		6.4%	2011	1.532\$		18.1%
2010	112.771M \$		6.4%	2010	1.297\$		9.8%
2009	101.634M \$		5.4%	2009	1.161\$		2.3%
2008	98.269M \$		5.7%	2008	1.154\$		25.4%
2007	77.520M \$		7.1%	2007	.920\$		15.5%
2006	66.393M \$		7.0%	2006	.797\$		13.5%

- [illegible]

- Trang 7 / 9

---

## II. NHÌN LẠI QUÁ TRÌNH LÀM ĐỒ ÁN

---

### **Khó khăn:**

- Khó khăn trong việc đi tìm nguồn dữ liệu chính thống với thông tin chính xác, ít nhiễu, ít thô.
- Khó khăn trong việc tìm hiểu knowledge domain của dữ liệu.

### **Những điều hữu ích học được:**

- Được làm việc hoàn chỉnh trên một mô hình khoa học dữ liệu.
- Kỹ năng khám phá và tiền xử lý dữ liệu.
- Kỹ năng đọc và tìm hiểu dữ liệu, cũng như nghiên cứu các mô hình.



---

### III. TÀI LIỆU THAM KHẢO

---

1. <https://likegeeks.com/python-correlation-matrix>
2. <https://worldbank.github.io/debt-data/api-guide/ids-api-guide-python-1.html>
3. <https://scikitlearn.org/stable/modules/generated/sklearn.impute.KNNImputer.html>