

How to use dataset in OpenDOSM

Using python programming

Requirements



Internet access



Gmail account (login
to GMAIL account)



URL data sources
from OpenDOSM

Step by steps

Notebook setup

1. Login to your Gmail Account
2. Open new tab and go to <https://github.com/booluckgmie/training/tree/main/python/OpenDOSM/opendosm.ipynb>
3. Add “tocolab”
<https://github.com/tocolab/booluckgmie/training/tree/main/python/OpenDOSM/opendosm.ipynb>
4. Ready to analyse using python



OR create
New
Notebook

Using Google Colab for Python



Google Colab Setup

1. Visit the [Google Colab](https://colab.research.google.com/) page, which will direct you to the [Google Colaboratory Welcome Page](https://colab.research.google.com/).

<http://colab.research.google.com/>

2. Click the **Sign in** button on the right top.

Analysis



LOAD DATA FROM
OPENDOSM



EXPLORE DATASET



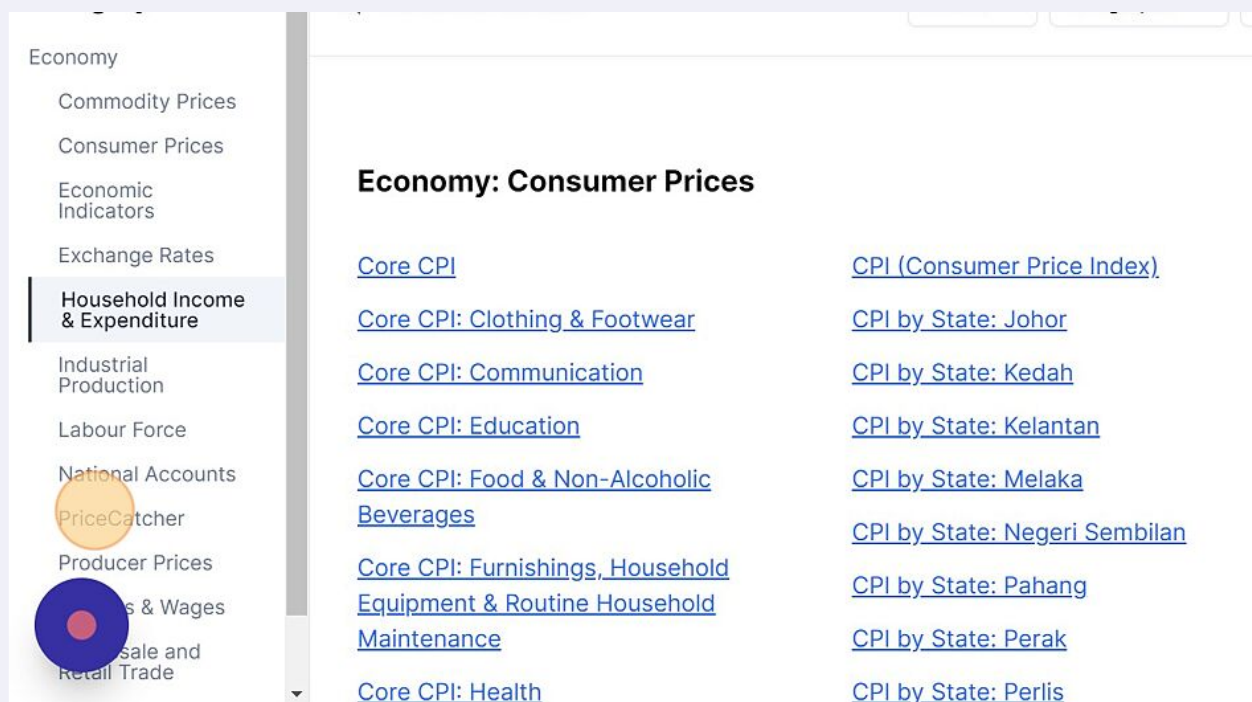
SAVE TO .CSV

Retrieving PriceCatcher Data with Python

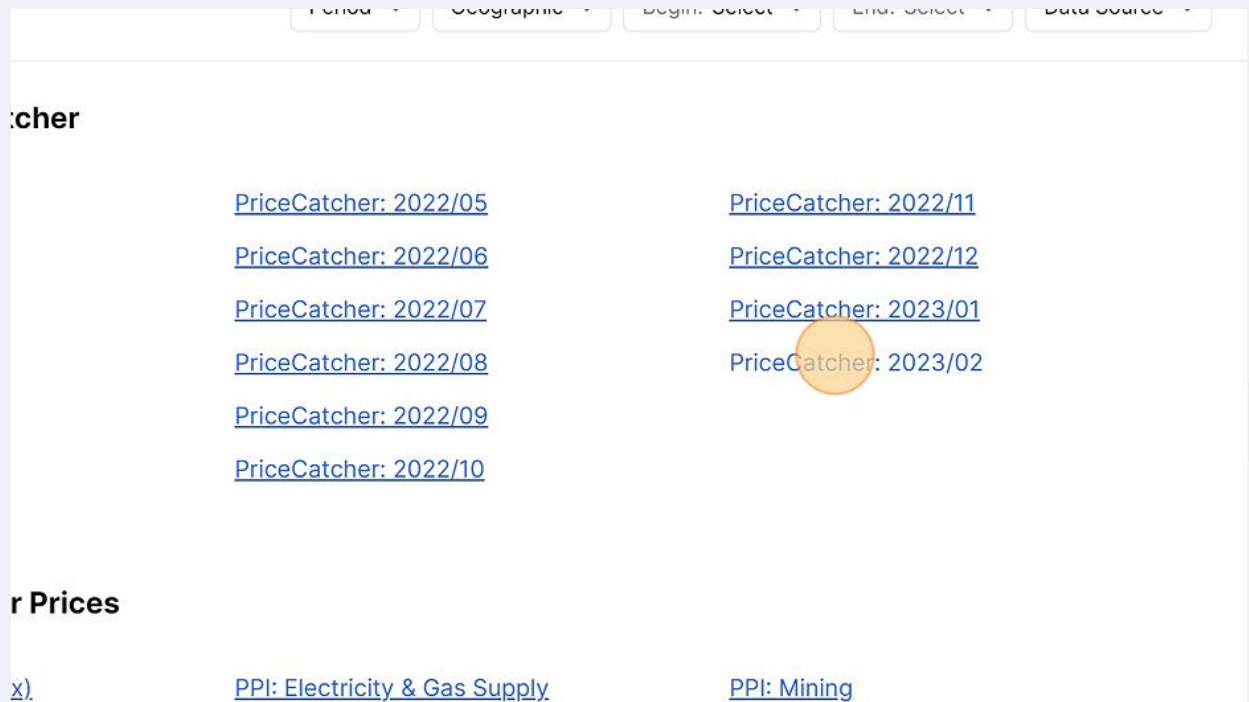
Start

1 Navigate to open.dosm.gov.my/data-catalogue

2 Click "PriceCatcher"

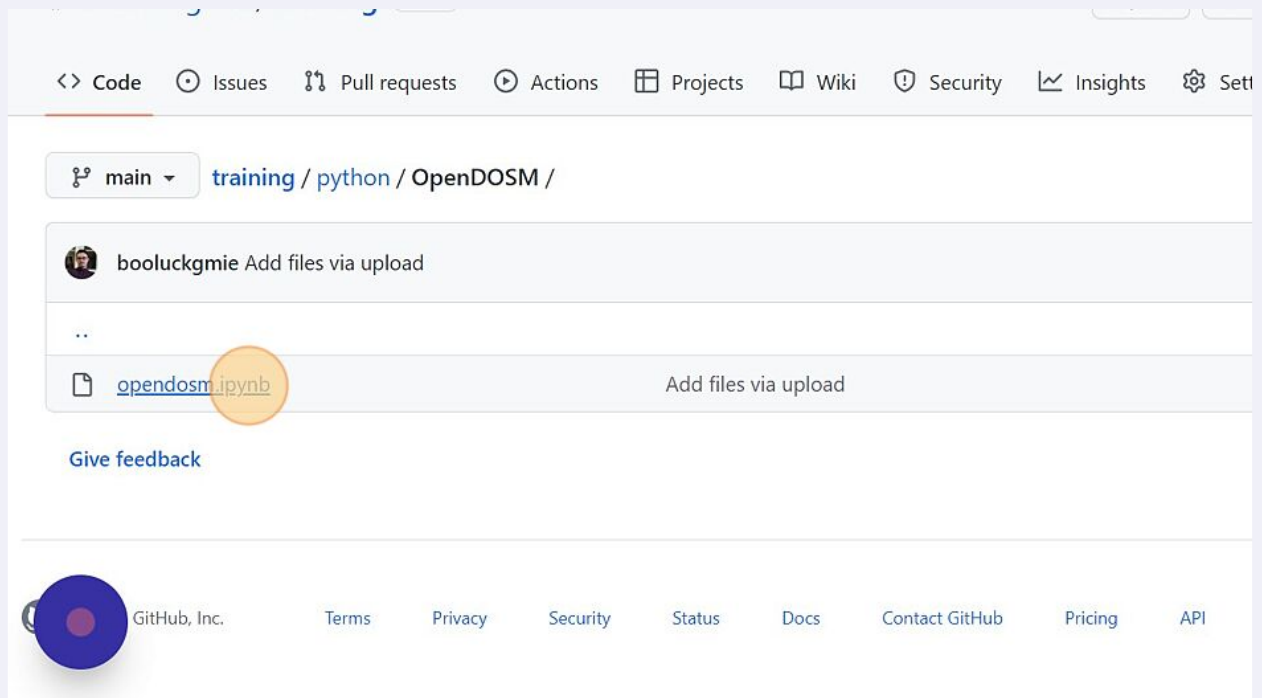


3 Click "PriceCatcher: 2023/02"



4 Switch to tab
"https://github.com/booluckgmie/training/tree/main/python/OpenDOSM"

5 Click "opendosm.ipynb"



6 Add "tocolab" and Navigate to github.com/booluckgmie/tocolab/blob/main/opendosm.ipynb

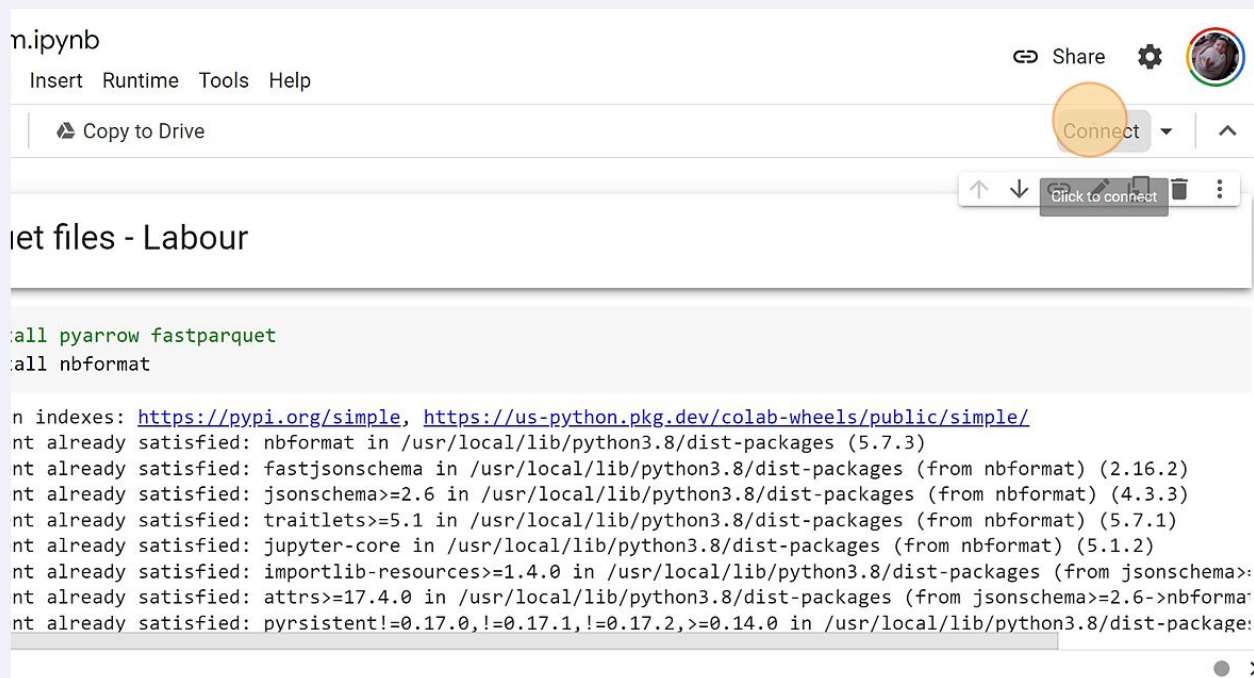


Tip!

Alternatively, login to colab.research.google.com using your Gmail account to access the colab notebook and create new notebook.

Google Colab Environment

7 Click here.



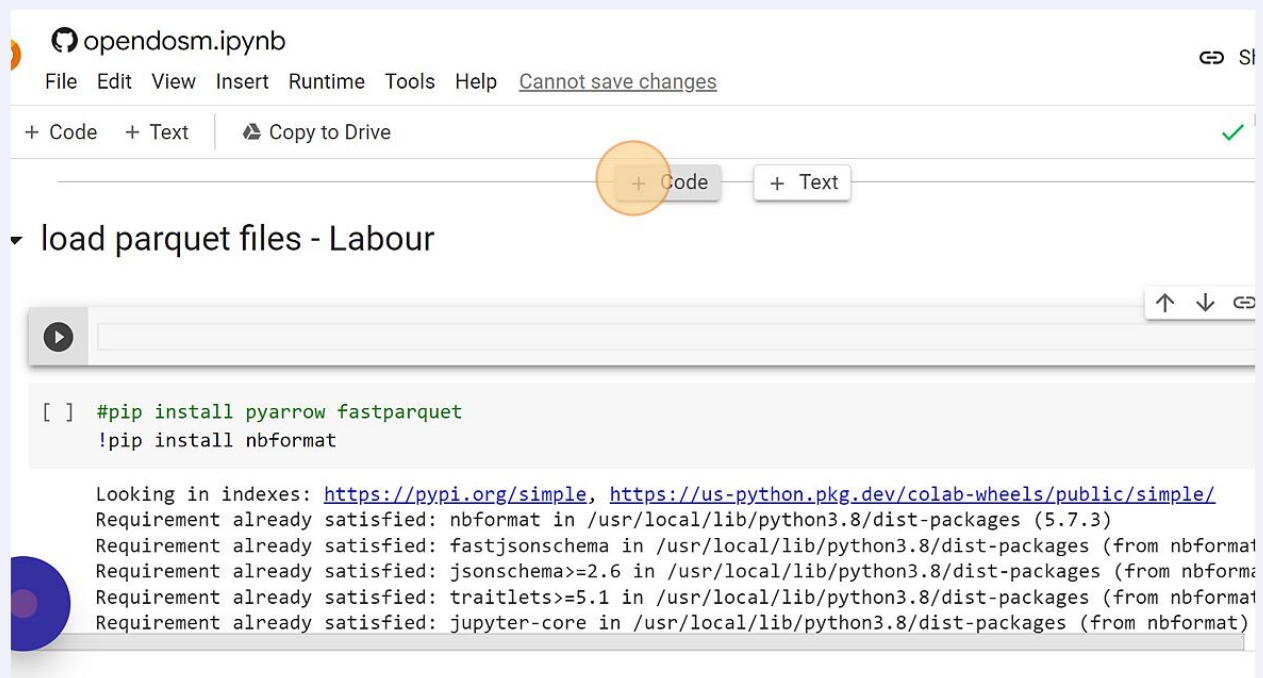
The screenshot shows a Jupyter Notebook interface. At the top, there's a header with "n.ipynb" and a menu bar with "Insert", "Runtime", "Tools", and "Help". A "Share" button and a user profile icon are on the right. Below the menu bar, there's a "Copy to Drive" button. The notebook title is "et files - Labour". The code cell contains the following text:

```
!pip install pyarrow fastparquet
!pip install nbformat
```

Below the code cell, the output shows the results of the installation and dependency checks:

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: nbformat in /usr/local/lib/python3.8/dist-packages (5.7.3)
Requirement already satisfied: fastjsonschema in /usr/local/lib/python3.8/dist-packages (from nbformat) (2.16.2)
Requirement already satisfied: jsonschema>=2.6 in /usr/local/lib/python3.8/dist-packages (from nbformat) (4.3.3)
Requirement already satisfied: traitlets>=5.1 in /usr/local/lib/python3.8/dist-packages (from nbformat) (5.7.1)
Requirement already satisfied: jupyter-core in /usr/local/lib/python3.8/dist-packages (from nbformat) (5.1.2)
Requirement already satisfied: importlib-resources>=1.4.0 in /usr/local/lib/python3.8/dist-packages (from jsonschema>=2.6->nbformat) (5.10.0)
Requirement already satisfied: attrs>=17.4.0 in /usr/local/lib/python3.8/dist-packages (from jsonschema>=2.6->nbformat) (21.4.0)
Requirement already satisfied: pyrsistent!=0.17.0,!0.17.1,!0.17.2,>=0.14.0 in /usr/local/lib/python3.8/dist-packages (from jsonschema>=2.6->nbformat) (0.19.3)
```

8 Click "Code"



The screenshot shows a Jupyter Notebook interface. At the top, there's a header with "opendosm.ipynb" and a menu bar with "File", "Edit", "View", "Insert", "Runtime", "Tools", "Help", and "Cannot save changes". A "Share" button and a user profile icon are on the right. Below the menu bar, there's a "Copy to Drive" button. The notebook title is "load parquet files - Labour". The code cell contains the following text:

```
[ ] #pip install pyarrow fastparquet
!pip install nbformat
```

Below the code cell, the output shows the results of the installation and dependency checks:

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: nbformat in /usr/local/lib/python3.8/dist-packages (5.7.3)
Requirement already satisfied: fastjsonschema in /usr/local/lib/python3.8/dist-packages (from nbformat) (2.16.2)
Requirement already satisfied: jsonschema>=2.6 in /usr/local/lib/python3.8/dist-packages (from nbformat) (4.3.3)
Requirement already satisfied: traitlets>=5.1 in /usr/local/lib/python3.8/dist-packages (from nbformat) (5.7.1)
Requirement already satisfied: jupyter-core in /usr/local/lib/python3.8/dist-packages (from nbformat) (5.1.2)
```

9

Switch previous Tab to open.dosm.gov.my/data-catalogue/dosm-public-pri...

10

Metadata for dataset

enDOSM [Home](#) [Data Catalogue](#) [Dashboards](#) 

Metadata

Dataset description

PriceCatcher is a mobile app developed by the Ministry of Domestic Trade and Cost of Living (KPDNKS/KPDNHEP) to help users compare the prices of key items in their area. Prices are collected and verified daily basis, with over 2 million prices collected every month. This dataset makes that wealth of data available for analysis.

Variable definitions

Name in Dataset	Variable	Definition
date (Date)	Date	The date in YYYY-MM-DD format
		Integer representing the premium, to be mapped using the Premium L

11 Ready API URL for dataset

[/dosm-public-pricecatcher/pricecatcher_2023-02.csv](#)
[/dosm-public-pricecatcher/pricecatcher_2023-02.parquet](#)

Creative Commons Attribution 4.0 International License (CC BY 4.0). A human-readable [link](#).

12 Ready python code to display the data in IDE



OpenDOSM



Home



Data Catalogue



Dashboards ▾

Code

Connect directly to the data with Python.



Python ▾

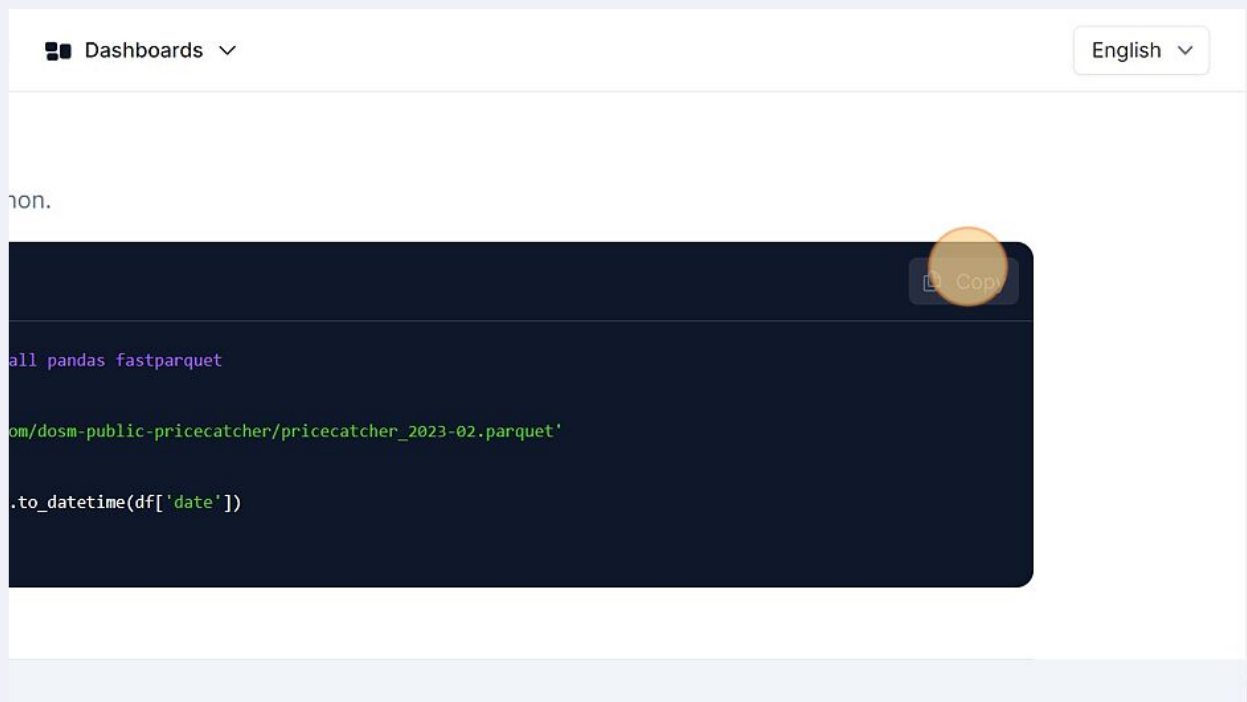
```
# If not already installed, do: pip install pandas fastparquet
import pandas as pd

URL_DATA = 'https://storage.googleapis.com/dosm-public-pricecatcher/pricecatcher_2023-02.parquet'

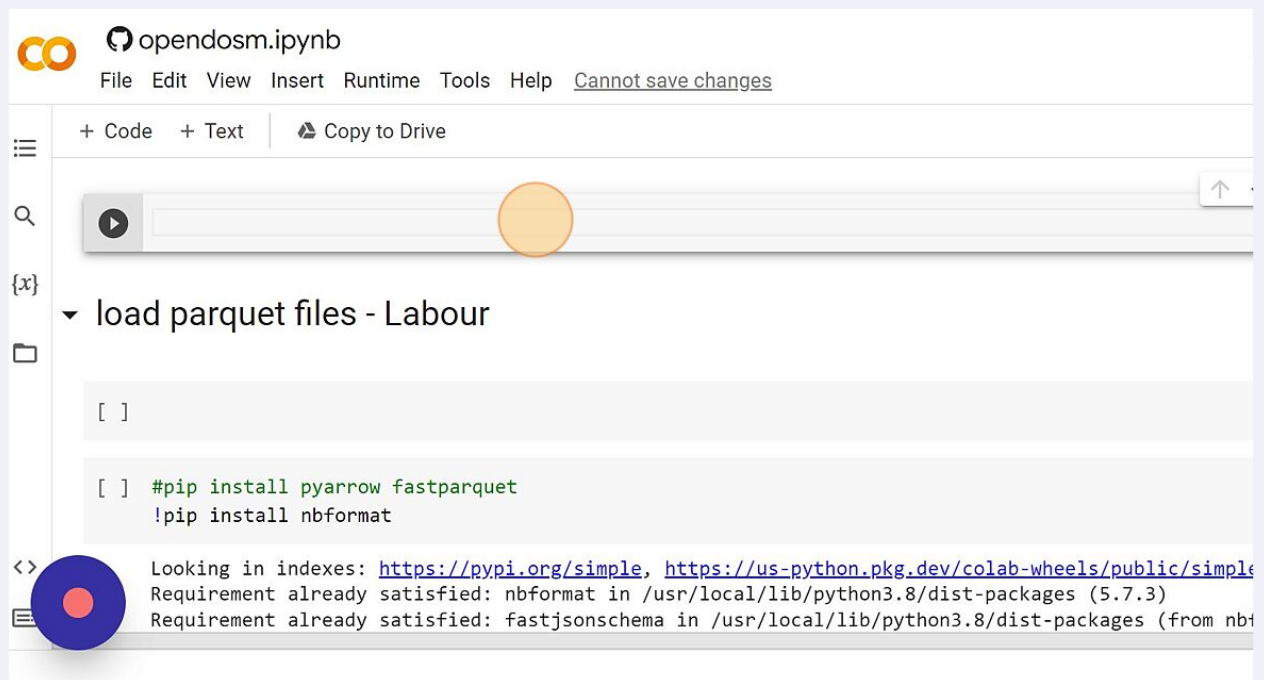
df = pd.read_parquet(URL_DATA)
if 'date' in df.columns: df['date'] = pd.to_datetime(df['date'])

print(df)
```

13 Click "Copy"

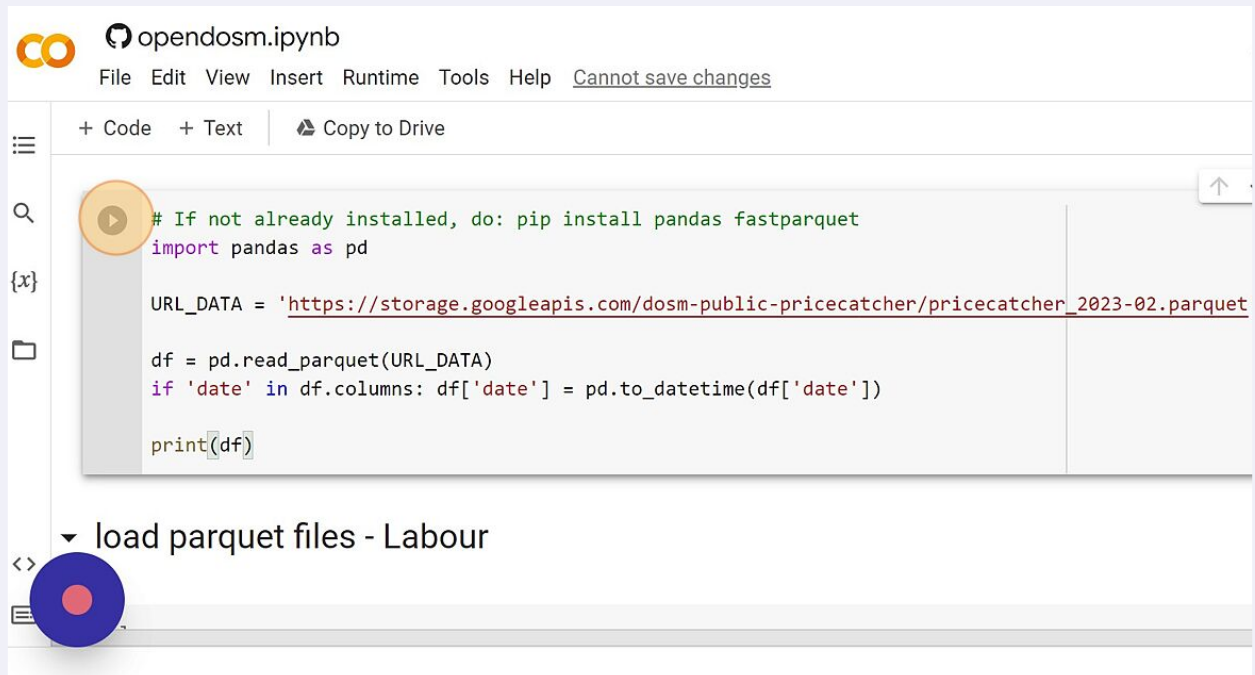


14 Insert new blank code.



15 Press `[[ctrl]] + [[v]]`

16 Click "Run cell"



The screenshot shows the OpenDOSM IPYNB web interface. At the top, there's a header with the OpenDOSM logo and the text "opendosm.ipynb". Below this is a menu bar with "File", "Edit", "View", "Insert", "Runtime", "Tools", and "Help", followed by a link "Cannot save changes". The main area has a toolbar with "+ Code", "+ Text", and "Copy to Drive". A code cell is active, containing the following Python code:

```
# If not already installed, do: pip install pandas fastparquet
import pandas as pd

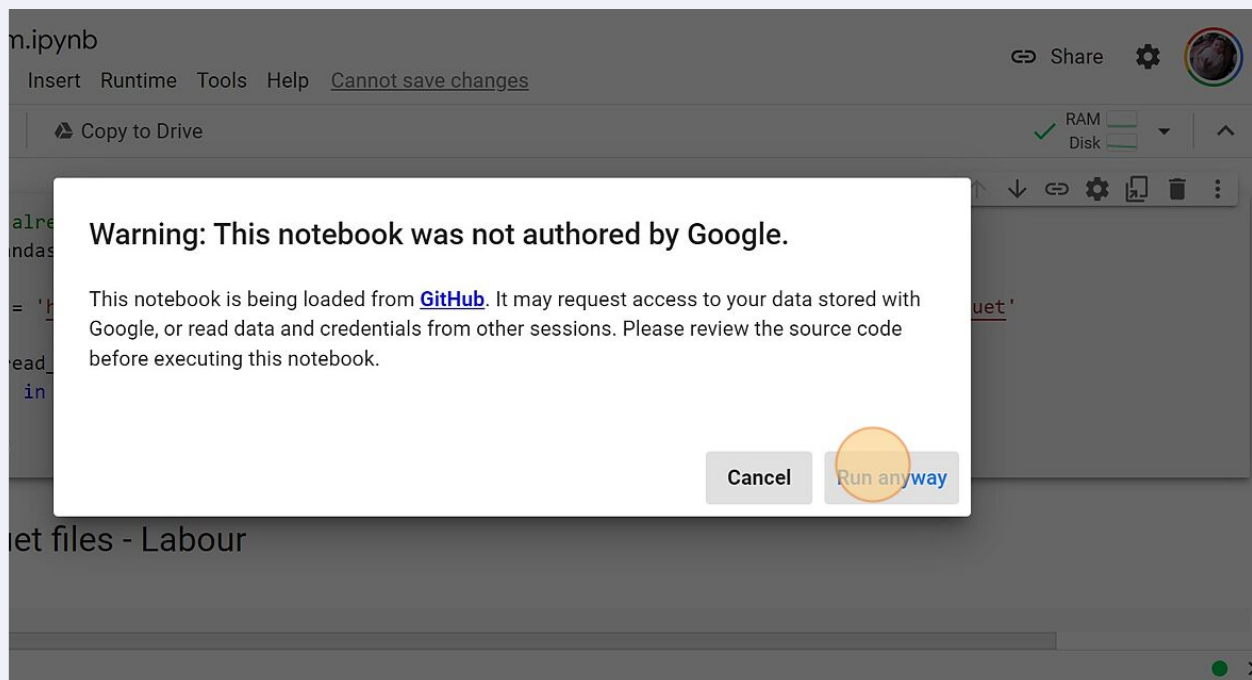
URL_DATA = 'https://storage.googleapis.com/dosm-public-pricecatcher/pricecatcher_2023-02.parquet'

df = pd.read_parquet(URL_DATA)
if 'date' in df.columns: df['date'] = pd.to_datetime(df['date'])

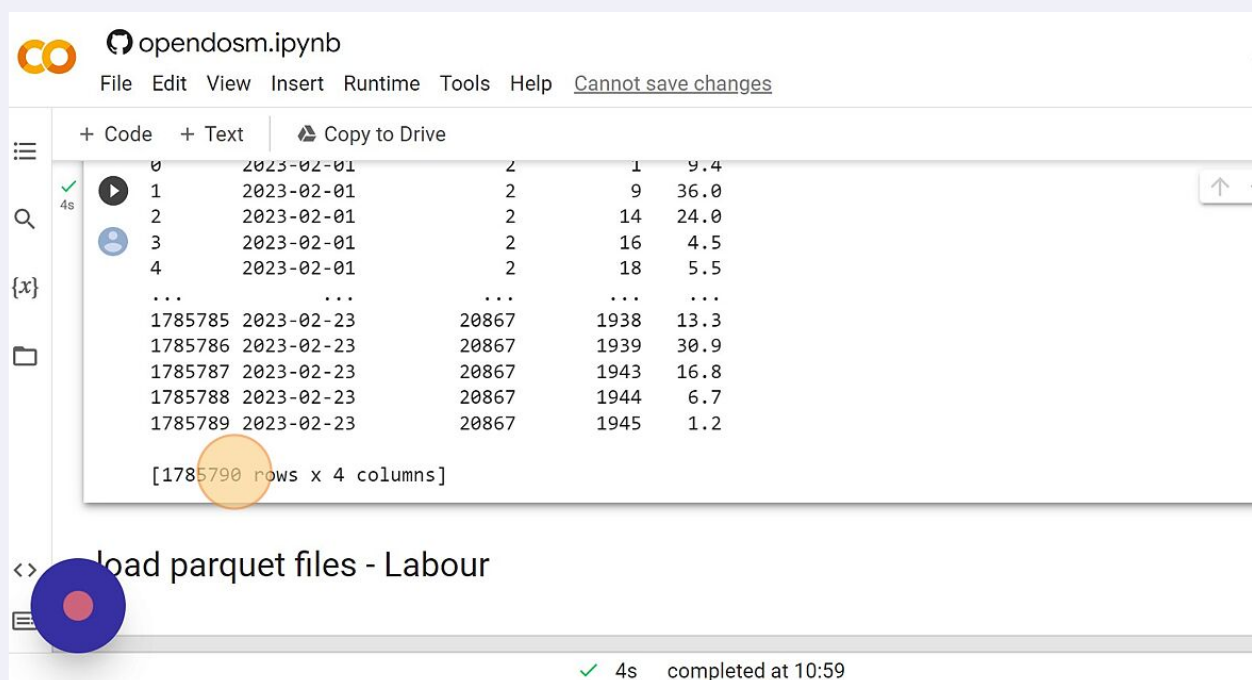
print(df)
```

Below the code cell, there's a section titled "load parquet files - Labour" with a blue circular cursor pointing to the "Run cell" button.

17 Click "Run anyway"



18 Output : Total number of rows



19 Click here.

opendosm.ipynb

File Edit View Insert Runtime Tools Help [Cannot save changes](#)

+ Code + Text Copy to Drive

[3]

1785785	2023-02-23	20867	1938	13.3
1785786	2023-02-23	20867	1939	30.9
1785787	2023-02-23	20867	1943	16.8
1785788	2023-02-23	20867	1944	6.7
1785789	2023-02-23	20867	1945	1.2

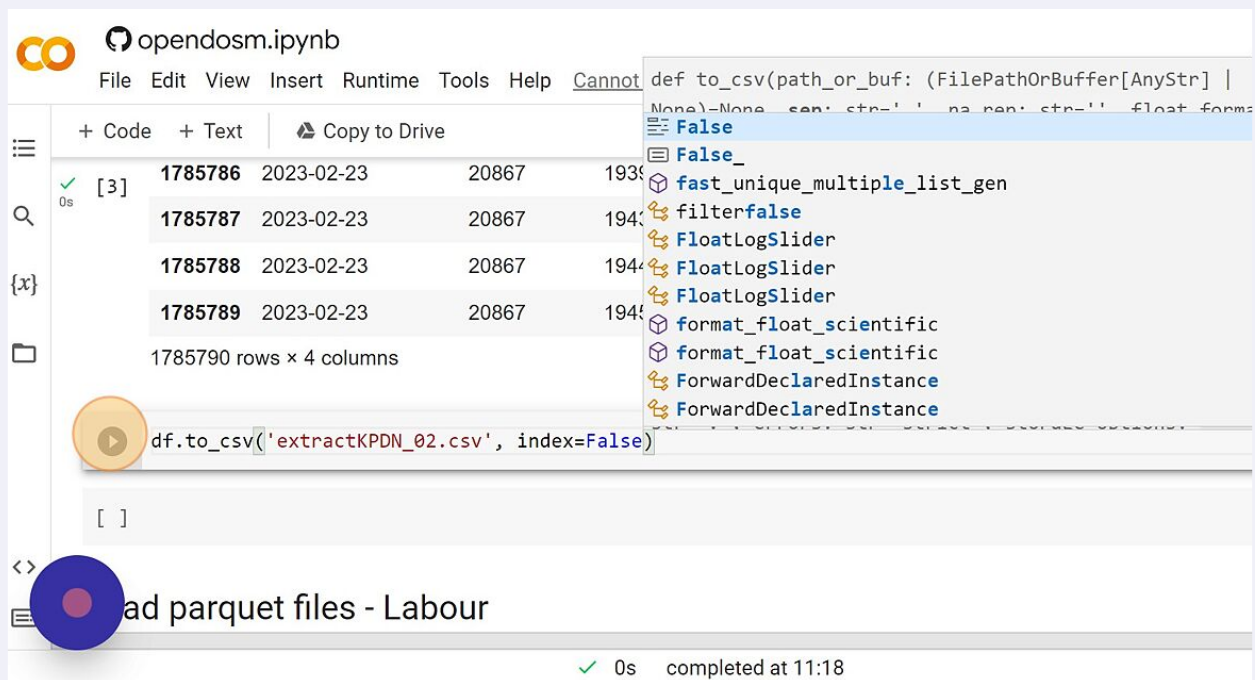
1785790 rows x 4 columns

completed at 11:18

Save dataframe to .csv

20 Type `df.to_csv('extractKPDN_02.csv' , index=False)`

21 Click "Run cell"



opendosm.ipynb

File Edit View Insert Runtime Tools Help Cannot

+ Code + Text Copy to Drive

[3] 1785786 2023-02-23 20867 1938

1785787 2023-02-23 20867 1943

1785788 2023-02-23 20867 1944

1785789 2023-02-23 20867 1945

1785790 rows × 4 columns

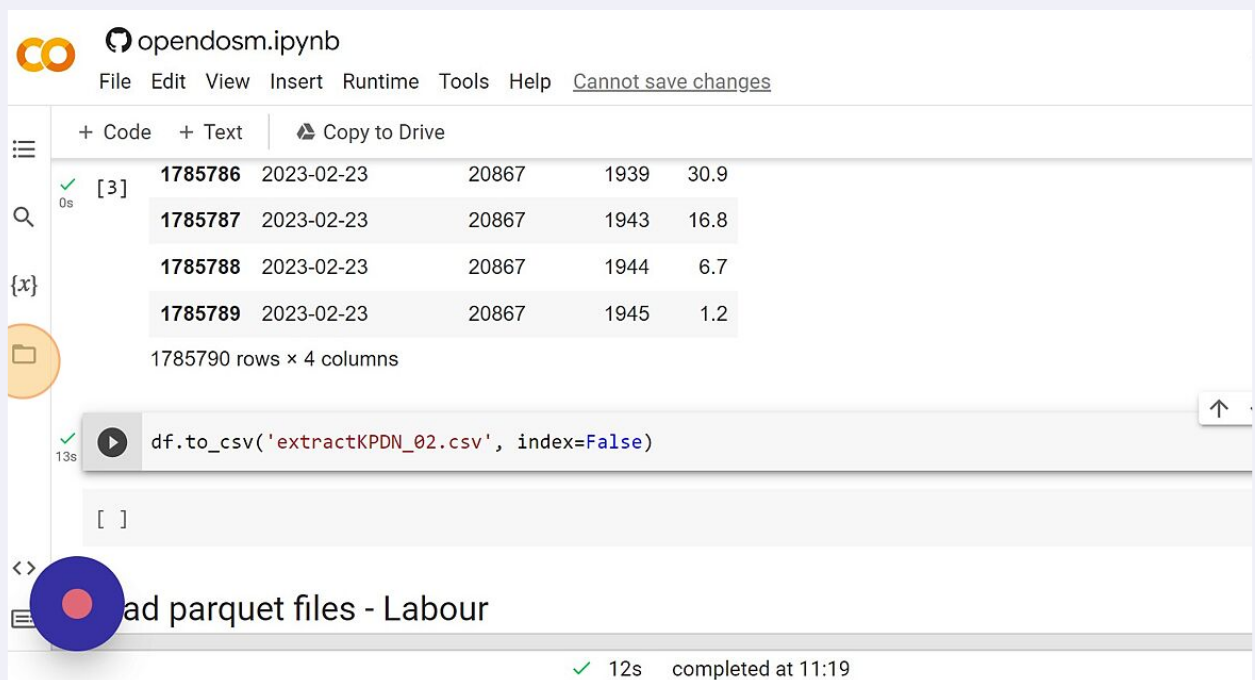
df.to_csv('extractKPDN_02.csv', index=False)

[]

ad parquet files - Labour

✓ 0s completed at 11:18

22 Click "Files"



opendosm.ipynb

File Edit View Insert Runtime Tools Help Cannot save changes

+ Code + Text Copy to Drive

[3] 1785786 2023-02-23 20867 1939 30.9

1785787 2023-02-23 20867 1943 16.8

1785788 2023-02-23 20867 1944 6.7

1785789 2023-02-23 20867 1945 1.2

1785790 rows × 4 columns

df.to_csv('extractKPDN_02.csv', index=False)

[]

ad parquet files - Labour

✓ 12s completed at 11:19

23 Click here.

opendosm.ipynb

File Edit View Insert Runtime Tools Help [Cannot save changes](#)

Files

sample_data

extractKPDN_02.csv

0s [3]

1785786	2023-02-23	20867	1939	30.9
1785787	2023-02-23	20867	1943	16.8
1785788	2023-02-23	20867	1944	6.7
1785789	2023-02-23	20867	1945	1.2

1785790 rows x 4 columns

13s

```
df.to_csv('extractKPDN_02.csv', index=False)
```

[]

load parquet files - Labour

12s completed at 11:19

83.28 GB available

24 Click "Download"

opendosm.ipynb

File Edit View Insert Runtime Tools Help [Cannot save changes](#)

Files

sample_data

extractKPDN_02.csv

0s [3]

1785786	2023-02-23	20867	1939	30.9
1785787	2023-02-23	20867	1943	16.8
1785788	2023-02-23	20867	1944	6.7
1785789	2023-02-23	20867	1945	1.2

1785790 rows x 4 columns

13s

```
df.to_csv('extractKPDN_02.csv', index=False)
```

[]

load parquet files - Labour

12s completed at 11:19

83.28 GB available

- Download
- Rename file
- Delete file
- Copy path
- Refresh

Done