THE EDGE™ MARKETS
MAKE BETTER DECISIONS

Listicle: Open data platforms every Malaysian data scientist should know

A strong data economy is needed not only to support the growth of the digital economy, but also to make the Malaysian Governance Index — used to measure the country's performance — possible.

A strong data economy is built, however, on a foundation of open data, where educators, students and organisations can use it to practise data exercises — generating unique insights that might have been overlooked and commercialising them to kick-start the data economy.

Unfortunately, Malaysia holds one of the lowest ranks in the Global Open Data Index, placing 87th out of a total 94 countries, doing worse than countries such as Zimbabwe and Afghanistan with a score of only 10%. For context, Singapore ranks 17th globally with a score of 60%, with Taiwan and Australia currently taking the top two spots, scoring 90% and 79% respectively.

To take advantage of the nascent open data landscape in Malaysia, here is a compilation of open data portals available to any data enthusiast. This is by no means an exhaustive list. For a more comprehensive list of all the available open data portals globally, readers may visit the Open Data Inception website (opendatainception.io).

1 Malaysia Open Data Portal

data.gov.my

The official Malaysia Open Data Portal serves as the heart of the country's open government data initiative. The site collects datasets from about 400 dataset suppliers, from the various ministries to state governments, police authorities and even the Department of Statistics Malaysia, housing about 12,800 datasets at the time of writing.

The site is being managed by the Malaysian Administrative Modernisation and Management Planning Unit (Mampu). For datasets that are missing, the portal has a feature that allows the general public to request for a particular dataset to be uploaded. There is even a ticketing system that allows users to track the status of the dataset request.

The datasets stored on the website leave much room for improvement, especially when it comes to standardisation. The language may be in either English or Bahasa Malaysia, and some of the data may be duplicated or outdated. Still, it is currently the most comprehensive open government data platform that we have.

2 Kaggle

kaggle.com/datasets

Ask any data scientist, and he or she will probably be familiar with Kaggle. The San Francisco-based subsidiary of Google is an online community for data scientists and machine-learning practitioners. More importantly, the website offers a repository of free datasets for practising and learning, containing both unstructured and structured data sets.

What sets Kaggle apart from other open data platforms is the addition of data science competitions. Through competitions, universities and non-profit organisations can leverage a globally sourced team of data professionals, to accelerate a particular area of research for social good. Businesses can also recruit pre-formed teams of data professionals best fit to address the company's unique challenges.

At the time of writing, the competition with the largest prize pool, of US$100,000, was organised by Amazon.com and the US National Football League (NFL), to develop the best sports injury surveillance and mitigation programme through the use of existing databases of gameplay footage.

Kaggle's inClass competition programme is also free for universities and is being used by institutions such as Harvard, Stanford and Oxford Universities. Students can even use existing codes provided by the Kaggle community and repurpose them for their personal projects.

3 OECD iLibrary

oecd-ilibrary.org

Datasets from the Organisation for Economic Co-operation and Development (OECD) is a must-have for any macroeconomic data analysis. OECD's iLibrary contains books, papers and statistics on the economic conditions of all 38 of its member countries, and its content is widely used by universities, research organisations, think tanks and public administrators.

In addition to economic indicators, the site also contains datasets ranging from energy to education, with content released by the International Energy Agency, Nuclear Energy Agency, Programme for International Student Assessment, and International Transport Forum.

Many of the datasets go as far back as 1998, giving data scientists plenty of wiggle room to conduct any time-series or historical analysis. Even without the raw datasets on the site, the OECD publishes between 300 and 500 books annually — a repository of information useful for any knowledge analysis beyond pure data and should be part of any data scientist's toolkit.

4 National Property Information Centre

napic.jpph.gov.my/portal

The National Property Information Centre's (Napic) online portal contains large databases on the local property market, many of which are not available on the official government open data portal. Most of these databases are not in their rawest format, featuring interpreted results and processed reports, thus making it difficult to repurpose them for machine-learning purposes.

However, these reports are still useful in generating insights and conducting basic analysis. The website contains information such as the status of newly launched residential properties, nationwide construction activities and total property supply.

To obtain more granular data that might involve individual sales transactions, the Valuation and Property Services Department of Malaysia (JPPH), which runs the Napic online portal, also offers unpublished data on a pay-per-use basis. The charges and pricing structure for these databases are not publicly available, but as a reference, the annual property market report itself is priced at RM100.

Currently, these unpublished databases are available only to nationally recognised and registered property valuers and real estate agents, and it is uncertain whether the service will be available to the general public. Parts of the transaction database can be found on the official government portal, but it contains data for only up to 2018 at the time of writing.

5 Asia Open Data Portal

dataportal.asia

Despite being the newest entrant on this list — only recently launched in March — the Asia Open Data Portal has already served more than 8.5 million users, containing about 167,000 datasets. It is Asia's first official open data portal, set up by the Taipei Computer Association in conjunction with Open Data Day 2021.

The catalogues of datasets within the portal can be massive but are neatly organised by country and broad categories, such as environment, health, education and logistics.

It has sourced almost all of its Malaysian datasets from the Malaysian open data portal. The browsing experience is nearly identical to visiting the official portal itself. Still, it provides the added advantage of quickly accessing datasets from around Asia without having to open multiple browser tabs.

https://www.theedgemarkets.com/article/listicle-open-data-platforms-every-malaysian-data-scientist-should-know