



# **Kaggle**

## **(Data Science Community)**



# Table of Content

## What will We Learn Today?

1. Kaggle dan tujuannya
2. Sejarah Kaggle
3. Fitur-fitur Kaggle
4. Mengembangkan skill data melalui Kaggle
5. Kompetisi di Kaggle
6. Eksekusi Ngoding di Kaggle





# Profile



## Professional

- Senior Data Analyst – Kompas (2021 – Present)
- Data Scientist – Rukita (2020 – 2021)
- Research Assistant Analyst – Ensterna (2017 – 2019)

## Educational Background

- Nuclear Engineering – Universitas Gadjah Mada

## Connect with me

-  <https://dataimpact.medium.com/>
-  <https://www.linkedin.com/in/ariprabowo/>
-  <https://github.com/densaiko>



**Ari Sulistiyo Prabowo**





*Kaggle adalah sebuah platform untuk  
berinteraksi, melatih dan menyelesaikan  
tantangan terkait data science dan machine  
learning*

# What is the purpose of Kaggle?

- User dapat menemukan dataset yang sifatnya umum
- Eksplorasi dan membangun model di dalam tampilan web
- Dapat berkoneksi dengan data scientist lain
- Mengikuti kompetisi untuk menyelesaikan suatu permasalahan



# History of Kaggle

- Ditemukan pada **April 2010** oleh **Anthony Goldbloom** dan **Jeremy Howard**
- Pada **8 Maret 2017**, Kaggle diakuisisi oleh **Google**
- Pada bulan **Juni 2017**, Kaggle mengumumkan bahwa mereka telah melampaui 1 juta user dan komunitas yang tersebar di 194 negara







# Features in Kaggle



Kaggle merupakan platform hebat yang memiliki berbagai fitur untuk melatih diri sendiri. Beberapa fitur di antaranya adalah:

- **Menemukan atau memuat dataset**
- **Eksplor codingan dari data scientist lain**
- **Berkolaborasi dari berbagai negara**
- **Kursus belajar mandiri**
- **Mengikuti kompetisi**



# Let's Create an Account

<https://www.kaggle.com/account/login>





# Fitur 1: Menemukan dan memuat dataset

- Home
- Competitions
- Datasets**
- Code
- Discussions
- Courses
- More

Recently Viewed

- Pakistan - Food Prices
- Cryptocurrency Histori...
- Bitcoin Historical Data

## Datasets

Explore, analyze, and share quality data. [Learn more](#) about creating, and collaborating.

+ New Dataset
Your Work

Datasets
Tasks
Computer Science
Education

### Trending Datasets

### Bitcoin Historical Data

Bitcoin data at 1-min intervals from select exchanges, Jan 2012 to March 2021

Zielak • updated 3 months ago (Version 7)

Data
Tasks (1)
Code (284)
Discussion (49)
Activity
Metadata

Download (303 MB)
New Notebook

Usability 10.0

License CC BY-SA 4.0

Tags finance, currencies and foreign exchange, history

Description

#### Context

Bitcoin is the longest running and most well known cryptocurrency, first released as open source in 2009 by the anonymous Satoshi Nakamoto. Bitcoin serves as a decentralized medium of digital exchange, with transactions verified and recorded in a public distributed ledger (the blockchain) without the need for a trusted record keeping authority or central intermediary. Transaction blocks contain a SHA-256 cryptographic hash of previous transaction blocks, and are thus "chained" together, serving as an immutable record of all transactions that have ever occurred. As with any currency/commodity on the market, bitcoin trading and financial instruments soon followed public adoption of bitcoin and continue to grow. Included here is historical bitcoin market data at 1-min intervals for select bitcoin exchanges where trading takes place. Happy (data) mining!

#### Content

bitstampUSD\_1-min\_data\_2012-01-01\_to\_2021-03-31.csv

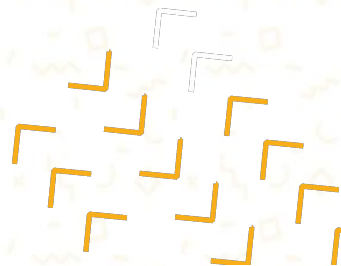
CSV files for select bitcoin exchanges for the time period of Jan 2012 to December March 2021, with minute to minute updates of OHLC (Open, High, Low, Close), Volume in BTC and indicated currency, and weighted bitcoin price. Timestamps are in Unix time. Timestamps without any trades or activity have their data fields filled with NaNs. If a timestamp is missing, or if there are jumps, this may be because the exchange (or its API) was down, the exchange (or its API) did not exist, or some other unforeseen technical error in data reporting or gathering. All effort has been made to deduplicate entries and verify the contents are correct and complete



## Fitur 2: Eksplorasi codingan

The screenshot displays the Kaggle website's 'Code' section. On the left, a sidebar menu includes 'Home', 'Competitions', 'Datasets', 'Code' (highlighted with a red box), 'Discussions', 'Courses', and 'More' (highlighted with a blue box). Below the menu, a 'Recently Viewed' list shows items like 'Bitcoin Historical Data' and 'Pakistan - Food Prices'. The main content area is titled 'Code' and includes a subtitle: 'Explore and run machine learning code with Kaggle Notebooks. Find help in the [Documentation](#).' Below this are buttons for '+ New Notebook' and 'Your work'. A search bar labeled 'Search public notebooks' is present. A row of filter tags includes 'All notebooks', 'Recently Viewed', 'Python', 'R', 'Beginner', 'NLP', 'Finance', 'Random Forest', 'GPU', and 'TF'. The 'Trending' section features three notebook cards: 1. '[SETI E.T.] EfficientNet B4 - Signal Detection' with a heatmap visualization, updated 9 hours ago. 2. 'Basic EDA and XGBoost for TPS July 2021' with a histogram visualization, updated 8 hours ago. 3. 'House Price Advanced Regression - Easy Solution' with a scatter plot visualization, updated 4 hours ago.

- Mengizinkan data scientist untuk **share coding** dan analisis menggunakan Python dan R
- Tempat yang baik untuk **meningkatkan skill** coding dan analisis





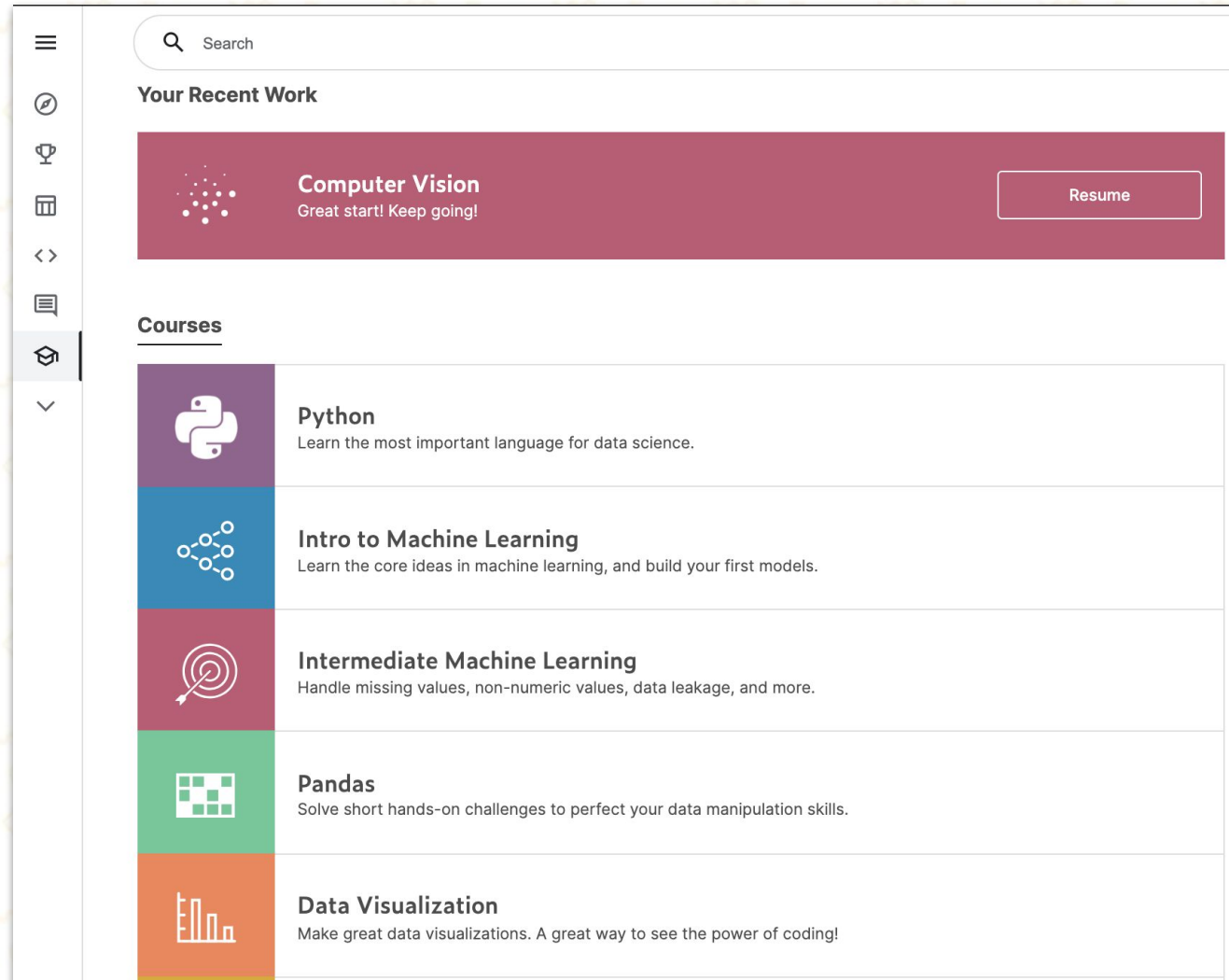
## Fitur 3: Berkolaborasi

#	Team Name	Notebook	Team Members	Score ?	Entries	Last
1	(♡♡♡)/ HP & NVIDIA Global Am...			0.652	130	5h
2	NT			0.641	42	9h
3	RTX 4090			0.641	86	18h
4	10k PCR test per hour			0.634	43	3h
5	Covid go away			0.630	177	1h
6	🤔 [Aillis] closed until 7/6 🤔			0.629	110	4d
7	Mikhail Gurevich			0.627	72	2d
8	Quoc-Hung To			0.626	92	4h
9	Emin Ozturk			0.626	65	7h
10	BabaCondaRabbit			0.626	115	1h
11				0.625	76	5d
12	Schwert			0.625	43	4d
13	D.Imanishi			0.625	64	10h
14	Train4Ever			0.622	71	13h
15	Target 0.63+			0.621	155	16h
16	Ahmed El Bakry			0.620	85	4h
17	kozistr			0.618	77	12h

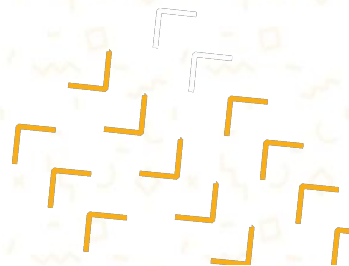




## Fitur 4: Kursus belajar gratis



- **Mempersiapkan** kandidat data scientist dengan kursus yang terpercaya
- Mendapatkan **sertifikat** jika kamu menyelesaikan kursusnya





# Fitur 5: Mengikuti Kompetisi

Featured Code Competition

## SIIM-FISABIO-RSNA COVID-19 Detection

Identify and localize COVID-19 abnormalities on chest radiographs

**\$100,000**  
Prize Money

Society for Imaging Informatics in Medicine (SIIM) · 829 teams · a month to go (a month to go until merger deadline)

Overview Data Code Discussion

Research Prediction Competition

## Google Smartphone Decimeter Challenge

Improve high precision GNSS positioning and navigation accuracy on smartphones

**\$10,000**  
Prize Money

Google · 610 teams · a month to go (25 days to go until merger deadline)

Featured Code Competition

## Optiver Realized Volatility Prediction

Apply your data science skills to make financial markets better

**\$100,000**  
Prize Money

Optiver · 383 teams · 3 months to go (3 months to go until merger deadline)

Overview Data Code Discussion Leaderboard Rules

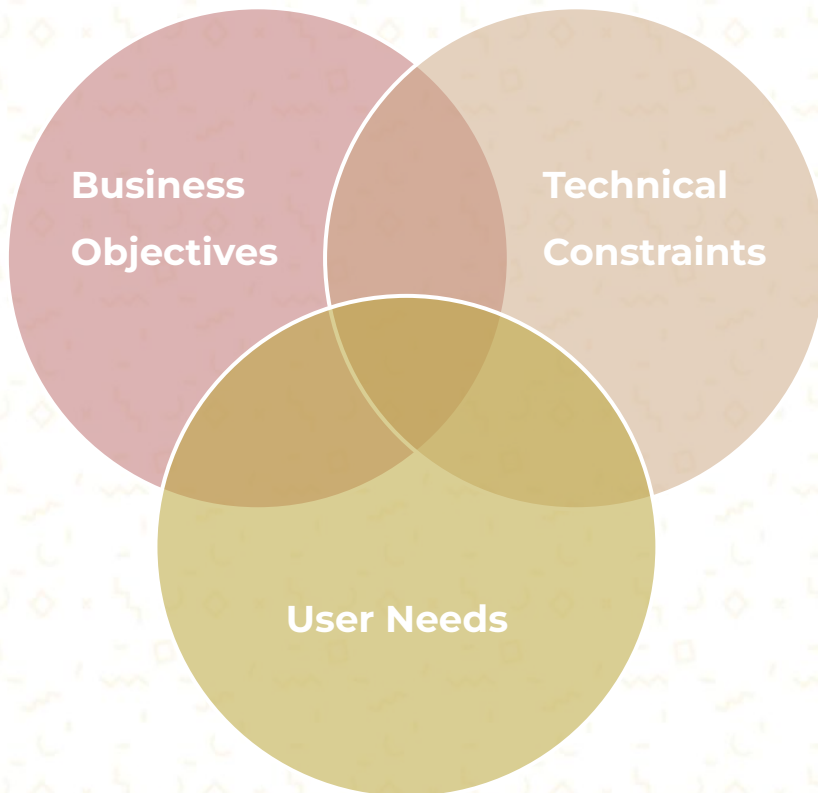
[Join Competition](#) ...

[Join Competition](#) ...





# Manfaat mengikuti kompetisi



## Problem Solving:

- Mempertajam skill untuk menyelesaikan suatu permasalahan dengan (analytical thinking)

## Domain Bisnis:

- Menyelesaikan masalah dengan domain bisnis tertentu akan memperkaya wawasan

## Users:

- Menjadikan kompetisi tersebut sebagai bukti yang dapat ditampilkan di portofolio





# Let's Practice

<https://www.kaggle.com/adityakadiwal/water-potability>



# Water Potability Dataset

Dataset

## Water Quality

Drinking water potability

Aditya Kadiwal • updated 2 months ago (Version 3)

Data Tasks (2) Code (64) Discussion (7) Activity Metadata

Download (513 KB) New Notebook

Usability 10.0 License CC0: Public Domain

Tags earth and nature, beginner, energy, public health, environment and 2 more

## Baca terlebih dahulu tugasnya:

- Pahami dulu permasalahannya seperti apa dan data-nya bagaimana

## Lihat codingan yang dibuat

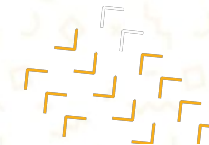
- Melakukan benchmark sebelum memulai analisis data

## Mulai koding

- Mulai notebook



# Water Potability Dataset



## WaterQuality

Draft saved

File Edit View Run Add-ons Help

+ Run All **Markdown**

Draft Session (31m)

### 1. Data Preparation

- Show any null values
- Finding any duplicate values
- Brief statistical analysis

### 2. Exploratory Data Analysis

+ Code + Markdown

## Data Preparation

```
[4]:  
# read the dataset  
data = pd.read_csv('../input/water-potability/water_potability.csv')  
data.head()
```

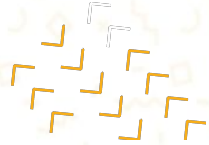
```
[4]:
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

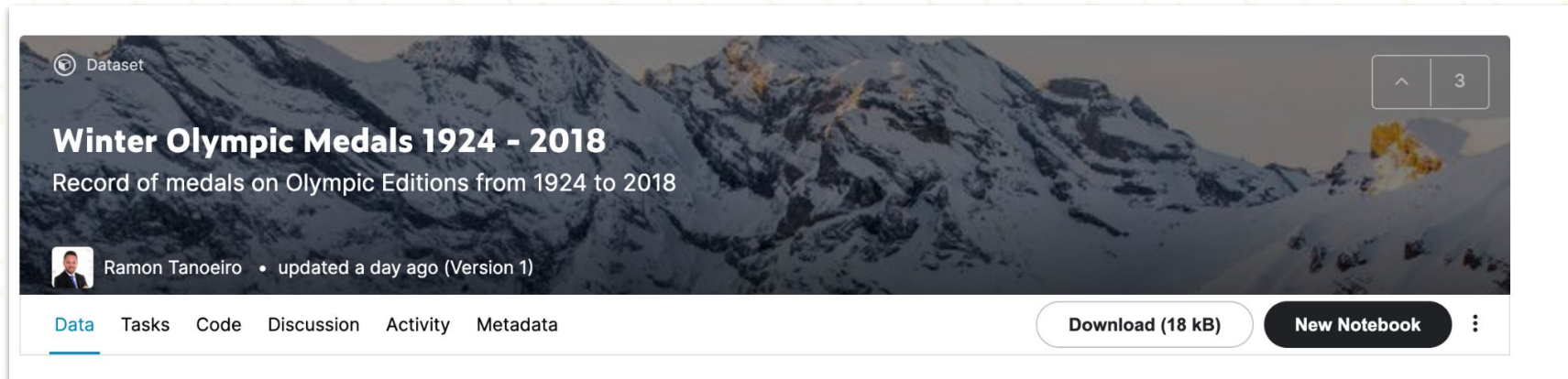




# Tugas



<https://www.kaggle.com/ramontanoeiro/winter-olympic-medals-1924-2018>



## Panduan tugas

1. Berikan deskripsi terkait data ini
2. Membuat notebook langsung di kaggle
3. Eksekusi data di dalam notebook kaggle dengan memunculkan
  - a. Load data
  - b. Informasi singkat terkait tipe datanya
  - c. Informasi singkat terkait statistik dasar

**Thank  
YOU**

