

AREIX

Everyone Worthy, Everyone Wealthy





Data collection and management

Kenny Fung, Data Engineer Lead

. Data preprocessing

. Data sources





Contents

Data preprocessing

- . Data quality & management**
- . Data understanding**
- . Data cleansing**
- . Data transformation**

Data sources

- . API**
- . Websites**



Data preprocessing

. Data Quality & Management

- . Data understanding
- . Data cleansing
- . Data transformation

“Garbage in, Garbage out”

NaN

Unstructured

Categorical

Redundant

Sparseness

Outliers

Irrelevant Input



Data preprocessing

. Data Quality & Management

- . Data understanding
- . Data cleansing
- . Data transformation

Standards

Validity

Is your data follows a specific format or business rules?

Accuracy

Whether the information represents the reality of the situation?

Completeness

Whether there are any missing elements?

Consistency

Whether the information is same in every instance?

Uniformity

Whether the same information has a same format from different sources?



Data preprocessing

- . Data Quality & Management
- . **Data understanding**
- . Data cleansing
- . Data transformation

Methods

Numerical

Outliers
Minimums
Maximums
Percent missing
Mean
Mode
Median
Ranges
Standard deviations

Graphical

Histogram
Scatter plot
Bar chart
Stem-and-leaf-plot



Data preprocessing

- . Data Quality & Management
- . **Data understanding**
- . Data cleansing
- . Data transformation

Example: Kickstarter Dataset

15 Columns

ID	state
name	backers
category	country
main_category	usd_pledged
currency	usd_pledged_real
deadline	usd_goal_real
goal	
launched	

Number of rows: 378661

Number of null: 0

Name:

375765 Unique values



Data preprocessing

- . Data Quality & Management
- . **Data understanding**
- . Data cleansing
- . Data transformation

Kickstarter: Currency

Unique value & Percentage:

USD	78%
GBP	9%
Others	13%

14 Unique values

Percentage of missing: 0%



Data preprocessing

- . Data Quality & Management
- . **Data understanding**
- . Data cleansing
- . Data transformation

Kickstarter: State

Unique value & Percentage:

failed	52.2%
successful	35.3%
canceled	10.2%
undefined	0.9%
live	0.7%
suspended	0.4%

6 Unique values

Percentage of missing: 0%



Data preprocessing

- . Data Quality & Management
- . **Data understanding**
- . Data cleansing
- . Data transformation

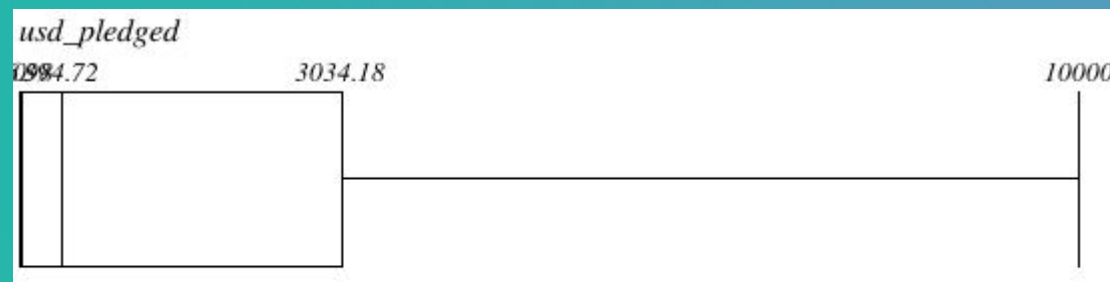
Kickstarter: `usd_pledged`

`usd_pledged`

mean: 7037.7
mode: 0
median: 394.72
min: 0
max: 20338986.27
sd: 78639.74531

`usd_pledged_real`

mean: 9058.9
mode: 0
median: 624.33
min: 0
max: 20338986.27
sd: 90973.34311





Data preprocessing

- . Data Quality & Management
- . Data understanding
- . **Data cleansing**
- . Data transformation

Missing Data/Outliers

To find out:

Percentage List
Data histogram
box plot

To solve:

Drop it (Either a row or column)
Impute it (Using mode/median/mean)
Replace it (-999/"Missing")
keep it



Data preprocessing

- . Data Quality & Management
- . Data understanding
- . **Data cleansing**
- . Data transformation

Unnecessary data

(Uninformative/Irrelevant/Duplicate)

To find out:

Data that doesn't add value

To solve:

Drop it (Either a row or column)

usd_pledged		usd_pledged_real	
mean:	7037.7	mean:	9058.9
mode:	0	mode:	0
median:	394.72	median:	624.33
min:	0	min:	0
max:	20338986.27	max:	20338986.27
sd:	78639.74531	sd:	90973.34311



Data preprocessing

- . Data Quality & Management
- . Data understanding
- . **Data cleansing**
- . Data transformation

Capitalization (Inconsistency)

To find out:

Show unique data

To solve:

Put all letters to lower or upper cases

Unique value & Percentage:

failed	52.2%
successful	35.3%
Failed	1%



Data preprocessing

- . Data Quality & Management
- . Data understanding
- . Data cleansing
- . **Data transformation**

Normalization

1. Removing any redundancies
2. Formatting the data
3. Consolidating data



Data preprocessing

- . Data Quality & Management
- . Data understanding
- . Data cleansing
- . **Data transformation**

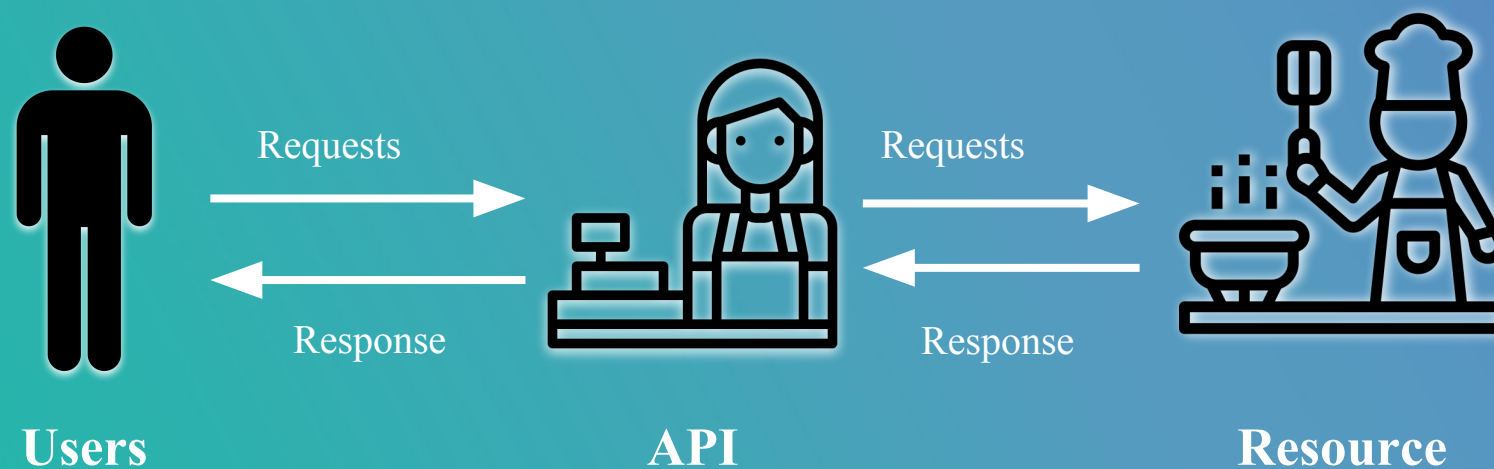
“Garbage in,
Garbage out”



Data Sources

- **API**
- Websites

What is API?



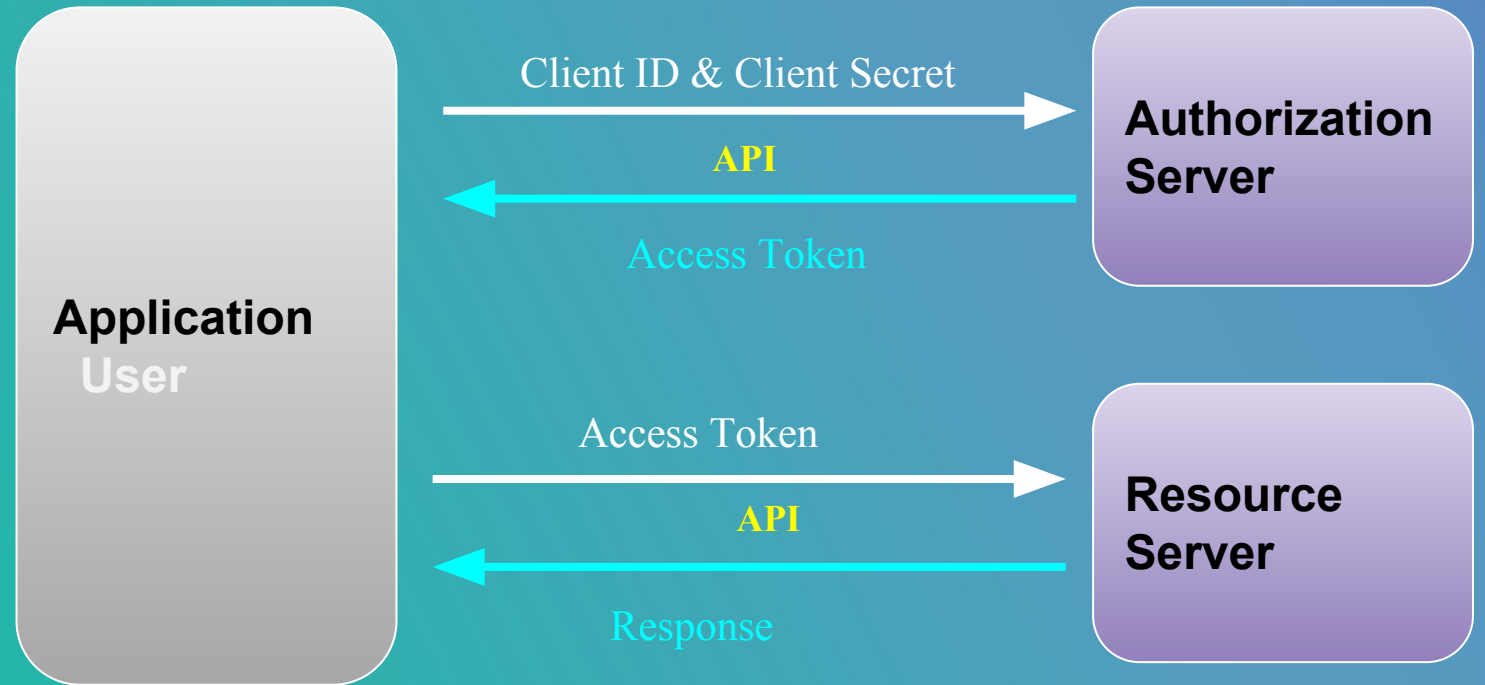


Data Sources

- **API**
- Websites

So.....


OAuth2





Data Sources

- API
- Websites



HONG KONG MONETARY AUTHORITY
香港金融管理局

HomeAbout HKMA's APIDocumentationAPI Data Enquiry

Home / Documentation / Market Data And Statistics / Monthly Statistical Bulletin

Documentation

Market Data And Statistics ▼

Bank & SVF Related Information

Debt Securities Settlement System ▼

Press Releases

inSight Articles

Coin Cart Schedule

Recruitment

Monthly Statistical Bulletin

Categories:

- Financial statistics summary
- Money
- Banking
- Money markets and debt instruments
- Exchange Fund Bills & Notes
- Exchange rates and interest rates
- Monetary market operation
- Exchange Fund and Foreign Currency Reserve Assets
- Government Bond Programme



Data Sources

- . API
- . **Websites**



A screenshot of the Data.gov.hk website. The top navigation bar includes the Data.gov.hk logo and the text "資料一線通 DATA.GOV.HK". Below the navigation bar, there is a search bar and a "Home > Dataset" breadcrumb. A filter menu is open, showing a list of categories: Development, Education, Employment and Labour, Environment, Finance (highlighted), Food, Health, Housing, IT and Broadcasting, and Law and Security. Below the filter menu, there is a dropdown menu for "Format(s)" with the option "-- ALL --". A "Clear Filter" button is visible. The main content area displays search results for "M" Mark Events, including a section for "11-a-side Soccer Pitches (Artificial Turf Pitch)". The results show the data category as "Recreation and Culture" and the format as "CSV" for the first result and "JSON" for the second result.

<https://data.gov.hk/en/>



Data Sources

- API
- Websites

HKSTP DATA STUDIO

The screenshot displays the HKSTP DATA STUDIO website interface. At the top, there is a navigation bar with links: DATA, FINTECH, DEVELOPERS, DATA PUBLISHERS, BLOG, CONTACT US, ABOUT, and FAQ. Below this is a category filter bar with links: SHOW ALL, ECONOMY - (29), EDUCATION - (17), ENVIRONMENT - (28), FINANCE - (163), HEALTH - (27), INFRASTRUCTURE - (14), SOCIETY - (107), TECHNOLOGY - (48), and TRANSPORT - (32). The main content area shows a list of APIs, with the first three visible:

- 2006 Population By-census - Statistical Tables**
Dependency Ratios, 1996, 2001 and 2006 (A108) (English), http://www.byccensus2006.gov.hk/FileManager/EN/Content_981/a108e.xls...
Population
- 2006 Population By-census: Hong Kong Resident Population - Studying Full Time In HK**
Hong Kong Resident Population (Studying Full Time In HK) by Ethnicity and Educational Attainment (Highest Level Attended), 2006 (B113) (English),...
Population
- 2006 Population By-census_Dependency Ratios**
Dependency Ratios, 1996, 2001 and 2006 (A108) (English), http://www.byccensus2006.gov.hk/FileManager/EN/Content_981/a108e.xls...
Population

The bottom row shows the first three of the following APIs:

- 2006 Population Census_Aged 5 and Over Able to Speak Selected Languages**
- 2006 Population Census_Hong Kong Resident Population**
- 2006 Population Census_Hong Kong Resident Population by age group**

<http://datastudio.hkstp.org/>



Data Sources

- API
- Websites



			回報, 風險, 開支, 基金規模及推出日期			費用及收費				
成分基金	受託人	基金類別	推出日期	基金規模 (百萬元)	風險級別 <small>類別 1 2 3 4 5 6 7 8 9 10</small>	最近期基金 開支比率 (%)	● 年率化回報 (% p.a.)		● 累積回報 (%)	
							一年期	五年期	十年期	推出至今
美洲基金	友邦信託	股票基金 - 美國股票基金	23-09-2011	1,648.46	6	0.99	9.79	9.89	n.a.	9.51
亞洲債券基金	友邦信託	債券基金 - 亞洲債券基金	23-09-2011	1,385.96	3	0.79	4.49	3.61	n.a.	1.96
亞洲股票基金	友邦信託	股票基金 - 亞洲股票基金	01-12-2004	4,417.22	6	1.94	13.27	7.86	4.36	6.00
均衡組合	友邦信託	混合資產基金 - 41% 至 60% 股票	01-12-2000	6,054.49	4	1.95	8.27	4.50	3.75	4.22
穩定資本組合	友邦信託	混合資產基金 - 21% 至 40% 股票	01-12-2000	3,807.44	4	1.93	6.18	3.58	2.67	3.64
中港動態資產配置基金	友邦信託	混合資產基金 - 未分類混合資產基金	04-07-2017	700.84	n.a.	1.32	6.49	n.a.	n.a.	3.94
核心累積基金	友邦信託	混合資產基金 - 預設投資策略 - 核心累積基金	01-04-2017	3,241.85	4	0.83	10.50	n.a.	n.a.	7.01
亞歐基金	友邦信託	股票基金 - 未分類股票基金	23-09-2011	404.95	5	0.99	2.28	3.39	n.a.	5.51
歐洲股票基金	友邦信託	股票基金 - 歐洲股票基金	01-01-2002	1,524.41	6	1.91	1.58	1.80	5.05	4.25
富達穩定資本基金	友邦信託	混合資產基金 - 21% 至 40% 股票	01-12-2010	1,239.99	4	1.84	7.69	4.27	n.a.	2.86
富達增長基金	友邦信託	混合資產基金 - 81% 至 100% 股票	01-12-2010	2,620.64	5	1.86	13.17	6.87	n.a.	4.89
富達穩定增長基金	友邦信託	混合資產基金 - 41% 至 60% 股票	01-12-2010	2,635.26	4	1.85	9.92	5.65	n.a.	4.15
環球債券基金	友邦信託	債券基金 - 環球債券基金	01-12-2007	2,524.35	3	0.97	6.35	3.85	1.63	2.34
大中華股票基金	友邦信託	股票基金 - 大中華股票基金	01-12-2004	9,559.16	6	1.92	36.64	10.90	5.39	7.00
綠色退休基金	友邦信託	股票基金 - 環球股票基金	31-03-2006	2,449.81	6	1.64	14.98	7.90	9.20	4.67
增長組合	友邦信託	混合資產基金 - 81% 至 100% 股票	01-12-2000	11,773.96	5	1.97	12.37	6.43	5.84	5.16
保證組合	友邦信託	保證基金	01-12-2000	9,555.81	1	1.60	0.15	0.17	0.64	1.52
中港基金	友邦信託	股票基金 - 香港股票基金	23-09-2011	2,355.79	6	0.98	0.15	5.00	n.a.	5.40
香港股票基金	友邦信託	股票基金 - 香港股票基金	01-01-2002	6,765.62	6	1.93	14.04	5.74	2.75	7.11
日本股票基金	友邦信託	股票基金 - 日本股票基金	01-01-2002	735.03	5	1.91	12.56	5.18	6.31	3.43
基金經理精選退休基金	友邦信託	混合資產基金 - 未分類混合資產基金	01-08-2008	4,774.33	5	1.66	7.63	5.19	5.56	5.41
強積金保守基金	友邦信託	貨幣市場基金 - 強積金保守基金	01-12-2000	6,225.85	1	0.97	0.82	0.50	0.32	0.67
北美股票基金	友邦信託	股票基金 - 美國股票基金	01-01-2002	3,942.23	6	1.90	17.67	10.49	12.43	5.55
全球基金	友邦信託	股票基金 - 環球股票基金	01-12-2007	1,660.20	6	0.99	5.76	7.20	7.13	3.11
安聯亞洲基金 - 單位B	銀聯信託	股票基金 - 亞洲股票基金	04-08-2004	1,464.48	n.a.	1.25	44.51	12.57	8.80	10.34

https://mfp.mpfa.org.hk/tch/mpp_list.jsp



Data Sources

- API
- **Websites**

Web Scraping: Selenium

```
1 from selenium import webdriver
2 from selenium.common.exceptions import NoSuchElementException
3 from selenium.webdriver.chrome.options import Options
4 import time
5
6 CHROMEDRIVER_PATH = r"C:\Program Files\Google\Chrome\Application\chromedriver.exe"
7 browser = webdriver.Chrome(executable_path=CHROMEDRIVER_PATH)
8
9 browser.get("https://mfp.mpfa.org.hk/tch/mpp_list.jsp")
10 time.sleep(2)
11 y = 500
12 for timer in range(8, 44):
13     browser.execute_script("window.scrollTo(0, " + str(y) + ";)")
14     y += 250
15     time.sleep(1)
```

```
browser.find_elements_by_xpath(xpath)
.get_attribute("textContent")
```



Summary

