當你拿到一堆資料時...

想成爲真正的數據科學家，除了資歷你還需要這四個技能

http://bangqu.com/56VS5p.html

- 不要僅僅質疑自己的結論，**還要質疑自己的數據**
- 能簡潔地表達自己的發現的價值
- 瞭解自己公司的業務
- 管理自己的上級


Research Prediction Competition
WSDM - KKBox's Music Recommendation Challenge
Can you build the best music recommendation system?
$5,000
Prize Money
KKBOX · 1,081 teams · 25 days ago

Solving Problem

In this task, you will be asked to **predict the chances of a user listening to a song repetitively** after the first observable listening event within a time window was triggered.

Data Set

train	test	sample_submission	Members	songs	song_extra_info
user_id	row_id	id	user_id	song_id	song_id
song_id	user_id	target	city	song_length	song_name
source_system_tab	song_id		age	genre_ids	song_identity_code
source_screen_name	source_system_tab		gender	artist_name	
source_type	source_screen_name		registered_via	composer	
target	source_type		registration_date	lyricist	
			expiration_date	language	

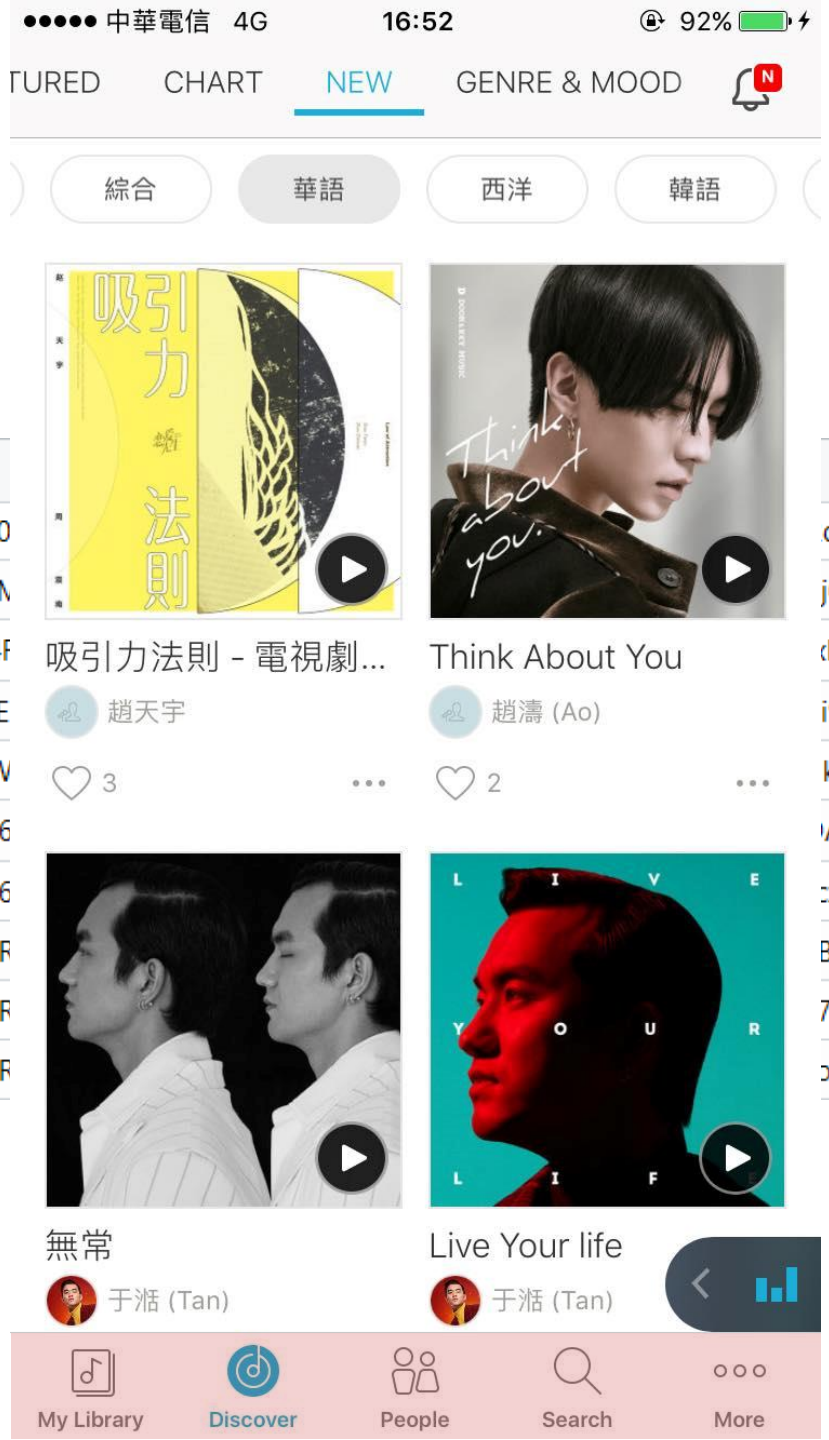
Data Set

- 如何收集到這組 Data Set？
- 這組 Data Set 是母體或抽樣得到的？抽樣的方式是否合理？
- 這組 Data Set 是否有經過任何形式的轉換？
- 這組 Data Set 是否有潛在 (已知) 的問題？

train

user_id	song_id	source_system_tab	source_screen_name	source_type	target
UsJ0TVPX7D+XNIS+bD8v/MQ+rw+ZbjW0PuU5c7tYymM=	uvhnnHYdmRoKrjqIZDtkJlrjyA5Ar3yZOG2aaV2q/EI=	my library	Local playlist more	local-library	1
AXKTVBbdr2/z7m5WL1IXzlriNbun8B+7wM41MsWCYDw=	WTQ1doDa8xAetUIXq+G/3AmF8/GPtJfRtbjaaj0vGqQ=	discover	Discover Feature	album	0
XIpjLNs4f/OHphnrSPoNmsrd45AI8pL0v74FgzDZHnQ=	8OsXpp32oqorzAN4NNkVouaV0I81ddPLZ8DxMRAGJdQ=	my library	Local playlist more	local-playlist	1
hKwxGNnAnVLbafv2sDSj8hIIf88SbDfnrSeE+R+bqFc=	rFLawM8j/vW78o/S3P2GTsmy6HwkTI15P7NF/i9CkX4=	my library	Local playlist more	local-playlist	1
hOyJGmY7tciqGf1EhDj5UT86IfSSnaTHP5JWAPxYN6s=	evaPiUW4ZoRCp27WhCmBCebbpvDtSW7IBNkhQ6M/IdM=	radio	Radio	radio	0
ocdN6b9BP9Qr0B8q6+UNpcBvfErNMk8P6Ny5CK3CyKA=	dSkGPyBKM97xpD3mOsCmhwSLO42eLuqBaDAamdZNBkI=	search	Online playlist more	online-playlist	0
ocdN6b9BP9Qr0B8q6+UNpcBvfErNMk8P6Ny5CK3CyKA=	F2d52gKSwWcIHb9TAQRBtj2uJ/ly5b8p5+cEucxV9aw=	my library	Artist more	top-hits-for-artist	1
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgRnpn3ZY/abw=	r01XmbdwG4sqS8ZA7tlx+VAuxrXNuDu7siNQB+Q2jrs=	my library	Album more	album	0
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgRnpn3ZY/abw=	u70RByc176rGRQLAmXdf7WNM0vU2RrkkYnY7z1wOPyA=	my library	Album more	album	0
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgRnpn3ZY/abw=	ijDKxOgX49hGdPBoC0S7ScN7g2TbuV1C9jfmoQMI6QM=	my library	Search	online-playlist	0

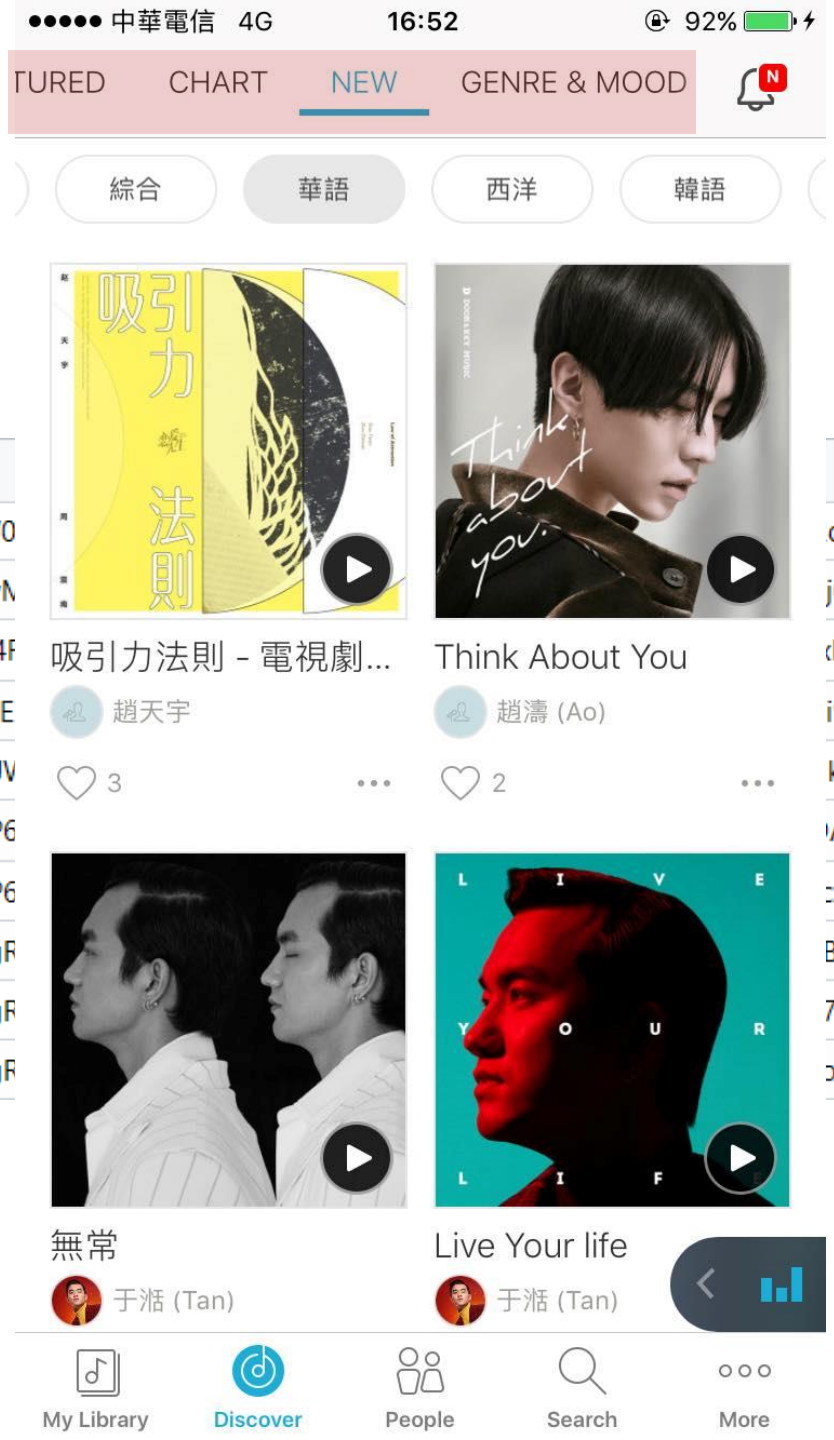
train



user_id
UsJ0TVPX7D+XNIS+bD8v/MQ+rw+ZbjW0
AXKTVBbdr2/z7m5WL1IXzlriNbun8B+7wM
XIpjLNs4f/OHphnrSPoNmsrd45AI8pL0v74F
hKwxGNnAnVLbafv2sDSj8hIIf88SbDfnrSeE
hOyJGmY7tciqGf1EhDj5UT86IfSSnaTHP5JV
ocdN6b9BP9Qr0B8q6+UNpcBvfErNMk8P6
ocdN6b9BP9Qr0B8q6+UNpcBvfErNMk8P6
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgR
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgR
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgR

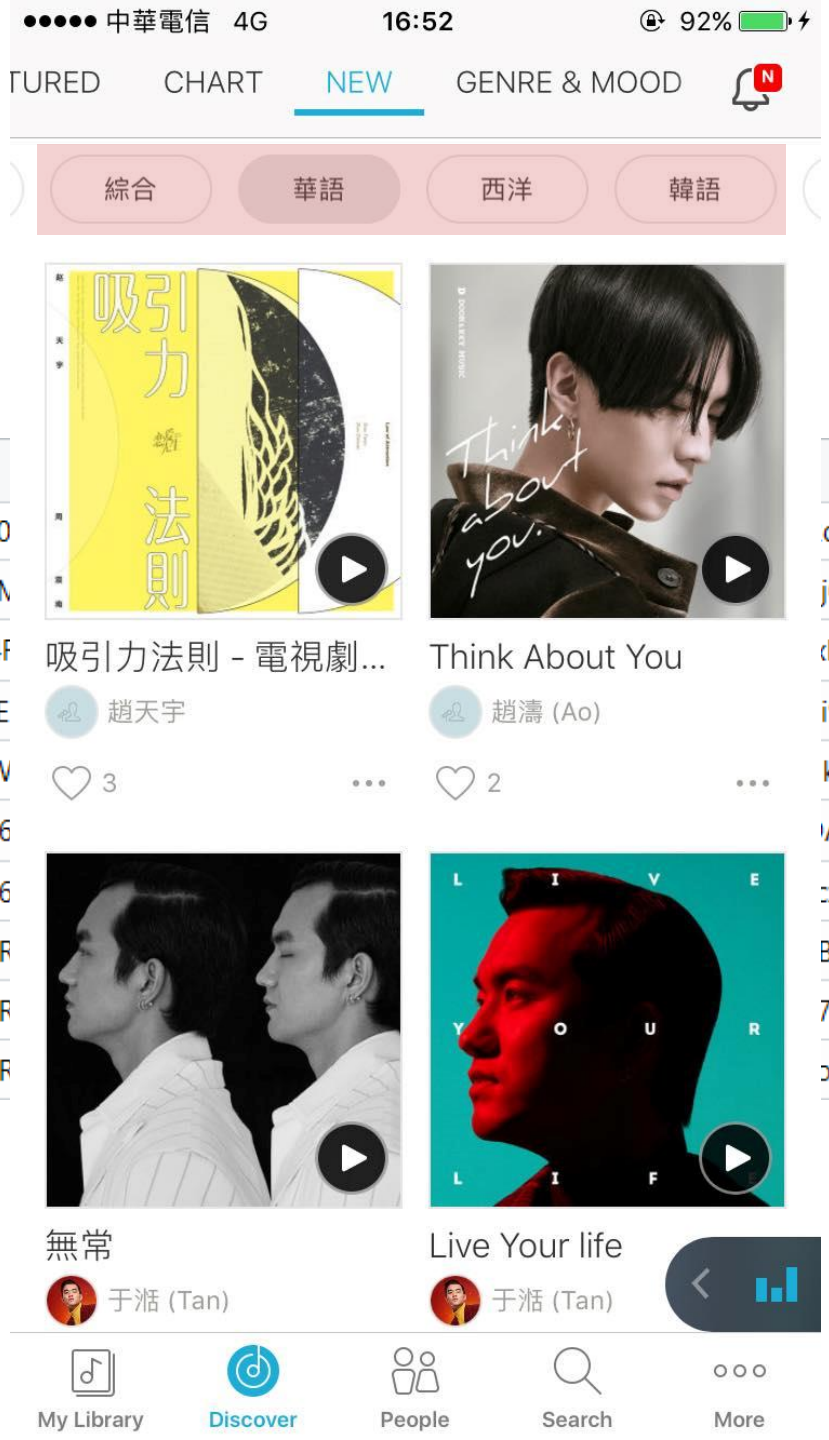
	source_system_tab	source_screen_name	source_type	target
q/EI=	my library	Local playlist more	local-library	1
j0vGqQ=	discover	Discover Feature	album	0
MRAGJdQ=	my library	Local playlist more	local-playlist	1
i9CkX4=	my library	Local playlist more	local-playlist	1
khQ6M/IdM=	radio	Radio	radio	0
AamdZNBkI=	search	Online playlist more	online-playlist	0
xV9aw=	my library	Artist more	top-hits-for-artist	1
B+Q2jrs=	my library	Album more	album	0
7z1wOPyA=	my library	Album more	album	0
QMI6QM=	my library	Search	online-playlist	0

train



user_id		source_system_tab	source_screen_name	source_type	target
UsJ0TVPX7D+XNIS+bD8v/MQ+rw+ZbjW0	q/EI=	my library	Local playlist more	local-library	1
AXKTVBbdr2/z7m5WL1IXzIriNbun8B+7wM	j0vGqQ=	discover	Discover Feature	album	0
XIpjLNs4f/OHphnrSPoNmsrd45AI8pL0v74F	MRAGJdQ=	my library	Local playlist more	local-playlist	1
hKwxGNnAnVLbafv2sDSj8hIIf88SbDfnrSeE	i9CkX4=	my library	Local playlist more	local-playlist	1
hOyJGmY7tciqGf1EhDj5UT86IfSSnaTHP5JV	khQ6M/IdM=	radio	Radio	radio	0
ocdN6b9BP9Qr0B8q6+UNpcBvfErNMk8P6	AamdZNBkI=	search	Online playlist more	online-playlist	0
ocdN6b9BP9Qr0B8q6+UNpcBvfErNMk8P6	xV9aw=	my library	Artist more	top-hits-for-artist	1
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgR	B+Q2jrs=	my library	Album more	album	0
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgR	7z1wOPyA=	my library	Album more	album	0
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgR	oQMI6QM=	my library	Search	online-playlist	0

train



user_id		source_system_tab	source_screen_name	source_type	target
UsJ0TVPX7D+XNIS+bD8v/MQ+rw+ZbjW0	q/EI=	my library	Local playlist more	local-library	1
AXKTVBbdr2/z7m5WL1IXzIriNbun8B+7wM	j0vGqQ=	discover	Discover Feature	album	0
XIpjLNs4f/OHphnrSPoNmsrd45AI8pL0v74F	MRAGJdQ=	my library	Local playlist more	local-playlist	1
hKwxGNnAnVLbafv2sDSj8hIIf88SbDfnrSeE	i9CkX4=	my library	Local playlist more	local-playlist	1
hOyJGmY7tciqGf1EhDj5UT86IfSSnaTHP5JV	khQ6M/IdM=	radio	Radio	radio	0
ocdN6b9BP9Qr0B8q6+UNpcBvfErNMk8P6	AamdZNBkI=	search	Online playlist more	online-playlist	0
ocdN6b9BP9Qr0B8q6+UNpcBvfErNMk8P6	xV9aw=	my library	Artist more	top-hits-for-artist	1
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgR	B+Q2jrs=	my library	Album more	album	0
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgR	7z1wOPyA=	my library	Album more	album	0
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgR	oQMI6QM=	my library	Search	online-playlist	0

Data Set

train	test	sample_submission	Members	songs	song_extra_info
user_id	row_id	id	user_id	song_id	song_id
song_id	user_id	target	city	song_length	song_name
source_system_tab	song_id		age	genre_ids	song_identity_code
source_screen_name	source_system_tab		gender	artist_name	
source_type	source_screen_name		registered_via	composer	
target	source_type		registration_date	lyricist	
			expiration_date	language	

members

registration method

user_id	city	age	gender	registered_via	registration_date	expiration_date
Ro4esuhkwPIVgArTpe7MIGJhIJy9MW7mFEWxuM5dSXE=	1	0		7	20160119	20161227
1XjK/MwTOEShVHyWLZN3cXjwfW4QCGjWuCAszXkjLHw=	5	16	female	4	20151109	20170708
t2K3zE/WEI84LuZDshjmzCXmdCP5L1eDCavjrCm0vGU=	1	0		7	20161213	20170913
vyoxo0c6XWEPvhndjGPS1unkM5HiCCFyI0wvXISSBek=	8	21	female	9	20161124	20180115
NNm80OCBAO6WHxJSWVhXII8TAsd9HjFXTa3uSgbDG2c=	4	16	male	3	20170218	20170221
9n7Yef1vL3Z8ZN7IDCOCBzKmLuVW/viQIFSu/7DdKdU=	1	0		4	20160822	20160825
zrPEmpkTqgI3MmbtKQ6gtmfI5df8JskI4NC/A6bI6iU=	4	24	male	3	20150502	20170911
bA2m/XPUtYDVjqjP9QIpY4YJ+Lbcx+4xcxb390JU+uI=	1	0		9	20150721	20170717
FtCJaSTuuhthFdPUkx6916iNmrGRw/mLDR3doFj/NCw=	1	0		7	20151112	20170926
D9dD56KuXsMxkWsYw/o1FgNwBopA2KO4qN662hNwKo8=	1	0		3	20130614	20170127

songs

↗ in millisecond

song_id	song_length	genre_ids	artist_name	composer	lyricist	language
m86acAJ3Bj2fL+4vbwWPHmOlrxF1Bk9F34/6kj5JpjI=	293511	786|947	寶寶水晶音樂			-1
v91olyybN9ksMslVNo8ZdbOwj/gtBv4VzK35LEk0B/M=	253492	465	周杰倫 (Jay Chou)	周杰倫	方文山	3
7HZRpauo3ediOwo5NUyJEFaybOpu/9FLdOskLIBwi70=	176170	1609	Martin Garrix			52
YdVtBho7uBMUYTmdFGoP8JRevrF98a5dFkEmSIjr8/s=	362788	465	楊乃文 (Naiwen Yang)			3
q0uoD+Tb1VDtc5Srw0G3wJnxFeUTM5rVoPzvwMszedE=	323605	465	Yui Aragaki (新垣結衣)	Kenji Kubo| Kubo Kenji	Iwasato Yuho| Yuho Iwasato	17
9wIALxtR+VQV/Gyjpw5ze9/fJveCVIvKj+KIAP4IL7o=	172245	242	Peggy Lee & Benny Goodman			52
ow9h9K+CIS0SzFZ0ipgI9davwVpxLmB3Fz2Ct4cL9SI=	240349	359	Imagine Dragons			52
zyO10vXfMjUGAE7XqKbK2pS8VGN8dWZ6s2PsD8gbKJk=	299026	465	蕭煌奇 (Ricky Hsiao)	蕭煌奇	胡如虹	3
veeW5CZtHxWRT6rznk1m0uUZs1Zp0/UKKdpz64dP9Vo=	244088	458	林宥嘉 (Yoga Lin)	林家謙	黃偉文	3
R+NIpdw3uc5e71o5+BhPyPidezdJeLGzDNDuxoiYWPE=	207934	465	BY2	林宇中	林怡鳳+毛凌琦	3

song_extra_info

theoretically can be used as an identity of a song

song_id	name	isrc
t0pC0urSmJanB9SC1iKfk7955AUoevuSLub3vm7xoxE=	Open Heaven (River Wild)	AUHS01507447
Lk0inRVUzoMeRVuVer4v863GodpDYDu68IRmGiFJkUU=	你在看什麼	JPSR09403290
jRHgqyJM5J89+ePki8OKSgTZFfNIWOi3yg5F5wwhynY=	唯一 (Wei Yi)	TWA470128001
5Bodh/4SCpTLJiOSS5tHPXR7ra+Nk6FmrDHiHVEQNNI=	Shining Day	
bNNgwWLzKX1Ox8n0sOICsGEoxq9fEqSeTyHF+r0yA4o=	擔心	
8ayqIorzfVcpM0419pUH+Qwv7UNfIRn1S7L40DLan/I=	我們的總和	TWA471603001
u3d60xvx0by7NOV/UZfvuGALrC5qRIeL1U1jNcBh3FM=	鴛鴦路	
RQpdpIPws8dPYtGEI33C6sFD1KPR25sWoZsdUIfNcJc=	Waiting Outside The Lines (勇敢守候)	USUM71025942
NAsdU76r5Z74ZbfsqFIHTFRvJNmp9zH0nZkEgyTThDQ=	最美的痕跡 (The Most Beautiful Scar)	TWA470708008
HKf0oR8pmsbszqQtuYZC8gSGDpgyD9GQoUKJ3R8h4jw=	Coming Home	USUM71200003

Data Set

- 如何收集到這組 Data Set？
- 這組 Data Set 是母體或抽樣得到的？抽樣的方式是否合理？
- 這組 Data Set 是否有經過任何形式的轉換？
- 這組 Data Set 是否有潛在 (已知) 的問題？

Solving Problem

Data Set

Data Profiling

當你拿到一堆資料時，

怎麼做 Data Profiling？

1. 掌握資料概況

2. 找出資料與目標 (問題) 間的關係

Data Profiling

Column Profiling	值域分析、類別分析、資料分布、欄位型態偵測、波動偵測、異常值偵測
Cross-column Profiling	相關性分析
Table Profiling	資料筆數檢視、主鍵唯一性分析
Cross-table Profiling	

Data Profiling

Column Profiling	值域分析、類別分析、資料分布、欄位型態偵測、波動偵測、異常值偵測
Cross-column Profiling	相關性分析
Table Profiling	資料筆數檢視、主鍵唯一性分析
Cross-table Profiling	

Table Profiling

資料筆數檢視

有多少筆資料

有多少個 feature

Table Profiling

資料筆數檢視

members

7 features

34,403 members

user_id	city	age	gender	registered_via	registration_date	expiration_date
Ro4esuhkwPIVgArTpe7MIGJhIJy9MW7mFEWxuM5dSXE=	1	0		7	20160119	20161227
1XjK/MwTOEShVHyWLZN3cXjwfW4QCGjWuCAszXkjLHw=	5	16	female	4	20151109	20170708
t2K3zE/WEI84LuZDshjmzCXmdCP5L1eDCavjrCm0vGU=	1	0		7	20161213	20170913
vyoxo0c6XWEPvhndjGPS1unkM5HiCCFyI0wvXISSBek=	8	21	female	9	20161124	20180115
NNm80OCBAO6WHxJSWVhXII8TAsd9HjFXTa3uSgbDG2c=	4	16	male	3	20170218	20170221
9n7Yef1vL3Z8ZN7IDCOCBzKmLuVW/viQIFSu/7DdKdU=	1	0		4	20160822	20160825
zrPEmpkTqgI3MmbtKQ6gtmfI5df8JskI4NC/A6bI6iU=	4	24	male	3	20150502	20170911
bA2m/XPUtYDVjqjP9QIpY4YJ+Lbcx+4xcxb390JU+uI=	1	0		9	20150721	20170717
FtCJaSTuuhthFdPUkx6916iNmrGRw/mLDR3doFj/NCw=	1	0		7	20151112	20170926
D9dD56KuXsMxkWsYw/o1FgNwBopA2KO4qN662hNwKo8=	1	0		3	20130614	20170127

⋮

```
> dim(members)
[1] 34403     7
```

Table Profiling

主鍵唯一性分析

主鍵 (Primary Key)

由一個或多個欄位組成

Table Profiling

主鍵唯一性分析

主鍵需唯一 (不重複)

不唯一串表格會產生錯誤資料

Table Profiling

主鍵唯一性分析

train

user_id + song_id

user_id	song_id	source_system_tab	source_screen_name	source_type	target
UsJ0TVPX7D+XNIS+bD8v/MQ+rw+ZbjW0PuU5c7tYymM=	uvhnnHYdmRoKrjqIZDtkJlrjyA5Ar3yZOG2aaV2q/EI=	my library	Local playlist more	local-library	1
AXKTVBbdr2/z7m5WL1IXzlriNbun8B+7wM41MsWCYDw=	WTQ1doDa8xAetUIXq+G/3AmF8/GPtJfRtbjaaj0vGqQ=	discover	Discover Feature	album	0
XIpjLNs4f/OHphnrSPoNmsrd45AI8pL0v74FgzDZHnQ=	8OsXpp32oqorzAN4NNkVouaV0I81ddPLZ8DxMRAGJdQ=	my library	Local playlist more	local-playlist	1
hKwxGNnAnVLbafv2sDSj8hIIf88SbDfnrSeE+R+bqFc=	rFLawM8j/vW78o/S3P2GTsmy6HwkTI15P7NF/i9CkX4=	my library	Local playlist more	local-playlist	1
hOyJGmY7tciqGf1EhDj5UT86IfSSnaTHP5JWAPxYN6s=	evaPiUW4ZoRCp27WhCmBCebbpvDtSW7IBNkhQ6M/IdM=	radio	Radio	radio	0
ocdN6b9BP9Qr0B8q6+UNpcBvfErNMk8P6Ny5CK3CyKA=	dSkGPyBKM97xpD3mOsCmhwSLO42eLuqBaDAamdZNBkI=	search	Online playlist more	online-playlist	0
ocdN6b9BP9Qr0B8q6+UNpcBvfErNMk8P6Ny5CK3CyKA=	F2d52gKSwWcIHb9TAQRBtj2uJ/ly5b8p5+cEucxV9aw=	my library	Artist more	top-hits-for-artist	1
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgRnpn3ZY/abw=	r01XmbdwG4sqS8ZA7tlx+VAuxrXNuDu7siNQB+Q2jrs=	my library	Album more	album	0
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgRnpn3ZY/abw=	u70RByc176rGRQLAmXdf7WNM0vU2RrkkYnY7z1wOPyA=	my library	Album more	album	0
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgRnpn3ZY/abw=	ijDKxOgX49hGdPBoC0S7ScN7g2TbuV1C9jfmoQMI6QM=	my library	Search	online-playlist	0

Table Profiling

主鍵唯一性分析

members

user_id

user_id	city	age	gender	registered_via	registration_date	expiration_date
Ro4esuhkwPIVgArTpe7MIGJhIJy9MW7mFEWxuM5dSXE=	1	0		7	20160119	20161227
1XjK/MwTOEShVHyWLZN3cXjwfW4QCGjWuCAszXkjLHw=	5	16	female	4	20151109	20170708
t2K3zE/WEI84LuZDshjmzCXmdCP5L1eDCavjrCm0vGU=	1	0		7	20161213	20170913
vyoxo0c6XWEPvhndjGPS1unkM5HiCCFyl0wvXISSBek=	8	21	female	9	20161124	20180115
NNm80OCBAO6WHxJSWVhXII8TAsd9HjFXTa3uSgbDG2c=	4	16	male	3	20170218	20170221
9n7Yef1vL3Z8ZN7IDCOCBzKmLuVW/viQIFSu/7DdKdU=	1	0		4	20160822	20160825
zrPEmpkTqgI3MmbtKQ6gtmfI5df8JskI4NC/A6bI6iU=	4	24	male	3	20150502	20170911
bA2m/XPUtYDVjqjP9QIpY4YJ+Lbcx+4xcxb390JU+uI=	1	0		9	20150721	20170717
FtCJaSTuuhthFdPUkx6916iNmrGRw/mLDR3doFj/NCw=	1	0		7	20151112	20170926
D9dD56KuXsMxkWsYw/o1FgNwBopA2KO4qN662hNwKo8=	1	0		3	20130614	20170127

Table Profiling

主鍵唯一性分析

```
> pk <- group_by(train, user_id, song_id) %>%
+         summarise(cnt = n()) %>%
+         filter(cnt > 1)
> pk
Source: local data frame [0 x 3]
Groups: user_id [0]

# ... with 3 variables: user_id <chr>, song_id <chr>, cnt <int>
```

PK 只能有一筆

Data Profiling

Column Profiling	值域分析、類別分析、資料分布、欄位型態偵測、波動偵測、異常值偵測
Cross-column Profiling	相關性分析
Table Profiling	資料筆數、主鍵唯一性分析
Cross-table Profiling	

Column Profiling

欄位型態偵測

欄位型態是否符合資料實際定義

Column Profiling

欄位型態偵測

members

user_id	city	age	gender	registered_via	registration_date	expiration_date
Ro4esuhkwPIVgArTpe7MIGJhIJy9MW7mFEWxuM5dSXE=	1	0		7	20160119	20161227
1XjK/MwTOEShVHyWLZN3cXjwfW4QCGjWuCAszXkjLHw=	5	16	female	4	20151109	20170708
t2K3zE/WEI84LuZDshjmzCXmdCP5L1eDCavjrCm0vGU=	1	0		7	20161213	20170913
vyoxo0c6XWEPvhndjGPS1unkM5HiCCFyl0wvXISSBek=	8	21	female	9	20161124	20180115
NNm80OCBAO6WHxJSWVhXII8TAsd9HjFXTa3uSgbDG2c=	4	16	male	3	20170218	20170221
9n7Yef1vL3Z8ZN7IDCOCBzKmLuVW/viQIFSu/7DdKdU=	1	0		4	20160822	20160825
zrPEmpkTqgI3MmbtKQ6gtmfI5df8JskI4NC/A6bI6iU=	4	24	male	3	20150502	20170911
bA2m/XPUtYDVjqjP9QIpY4YJ+Lbcx+4xcxb390JU+uI=	1	0		9	20150721	20170717
FtCJaSTuuhthFdPUkx6916iNmrGRw/mLDR3doFj/NCw=	1	0		7	20151112	20170926
D9dD56KuXsMxkWsYw/o1FgNwBopA2KO4qN662hNwKo8=	1	0		3	20130614	20170127

Column Profiling

欄位型態偵測

members

user_id	character
city	integer
age	integer
gender	character
registered_via	integer
registration_date	integer
expiration_date	integer

```
> str(members)
'data.frame':   34403 obs. of  7 variables:
 $ user_id          : chr  "XQxgAYj3klVKjR3ox ...
 $ city             : int  1 1 1 1 1 13 ...
 $ age              : int  0 0 0 0 0 43 ...
 $ gender           : chr  "" "" "" "" ...
 $ registered_via   : int  7 7 4 9 4 9 4 7 7 7 ...
 $ registration_date: int  20110820 20150628 ...
 $ expiration_date  : int  20170920 20170622 ...
```

Column Profiling

欄位型態偵測

members

user_id	character
city	integer
age	integer
gender	character
registered_via	integer
registration_date	integer
expiration_date	integer

同樣都是 *category*，有的欄位型態是 *character*，有的卻是 *integer* ？

Column Profiling

欄位型態偵測

members

user_id	character
city	integer
age	integer
gender	character
registered_via	integer
registration_date	integer
expiration_date	integer

日期欄位的資料型態是 *integer* ？

Column Profiling

欄位型態偵測

欄位型態轉換

```
> members <- transform(members,
+                      city = as.character(city),
+                      registered_via = as.character(registered_via),
+                      registration_date = as.Date(as.character(registration_date), format = "%Y%m%d"),
+                      expiration_date = as.Date(as.character(expiration_date), format = "%Y%m%d"))

> str(members)
Classes ‘data.table’ and 'data.frame':  34403 obs. of  7 variables:
 $ user_id          : chr  "XQxgAYj3klVKjR3oxPPXYYFp4soD4TuBghkhMTD4oTw=" ...
 $ city             : chr  "1" "1" "1" "1" ...
 $ age              : int  0 0 0 0 0 43 0 0 0 0 ...
 $ gender           : chr  "" "" "" "" ...
 $ registered_via   : chr  "7" "7" "4" "9" ...
 $ registration_date: Date, format: "2011-08-20" "2015-06-28" "2016-04-11" "2015-09-06" ...
 $ expiration_date  : Date, format: "2017-09-20" "2017-06-22" "2017-07-12" "2015-09-07" ...
```

Column Profiling 值域分析

分析欄位值的統計量

(max, min, mean, median,…)

Column Profiling

値域分析

members

age

user_id	city	age	gender	registered_via	registration_date	expiration_date
Ro4esuhkwPIVgArTpe7MIGJhIJy9MW7mFEWxuM5dSXE=	1	0		7	20160119	20161227
1XjK/MwTOEShVHyWLZN3cXjwfW4QCGjWuCAszXkjLHw=	5	16	female	4	20151109	20170708
t2K3zE/WEI84LuZDshjmzCXmdCP5L1eDCavjrCm0vGU=	1	0		7	20161213	20170913
vyoxo0c6XWEPvhndjGPS1unkM5HiCCFyl0wvXISSBek=	8	21	female	9	20161124	20180115
NNm80OCBAO6WHxJSWVhXII8TAsd9HjFXTa3uSgbDG2c=	4	16	male	3	20170218	20170221
9n7Yef1vL3Z8ZN7IDCOCBzKmLuVW/viQIFSu/7DdKdU=	1	0		4	20160822	20160825
zrPEmpkTqgI3MmbtKQ6gtmfI5df8JskI4NC/A6bI6iU=	4	24	male	3	20150502	20170911
bA2m/XPUtYDVjqjP9QIpY4YJ+Lbcx+4xcxb390JU+uI=	1	0		9	20150721	20170717
FtCJaSTuuhthFdPUkx6916iNmrGRw/mLDR3doFj/NCw=	1	0		7	20151112	20170926
D9dD56KuXsMxkWsYw/o1FgNwBopA2KO4qN662hNwKo8=	1	0		3	20130614	20170127

Column Profiling

値域分析

members *age*

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-43.00	0.00	0.00	12.28	25.00	1051.00

```
> summary(members$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-43.00    0.00    0.00   12.28   25.00 1051.00
```

Column Profiling

異常值偵測

欄位是否包含異常資料或數值

Column Profiling

異常值偵測

members *age*

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-43.00	0.00	0.00	12.28	25.00	1051.00

最小值還在媽媽肚子裡？

最大值長命千歲？

Column Profiling

異常值偵測

members *age*

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-43.00	0.00	0.00	12.28	25.00	1051.00

至少一半（19932 筆）的會員是 0 歲？

（比較有可能是遺漏值）

Column Profiling

異常值偵測

members *age*

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-43.00	0.00	0.00	12.28	25.00	1051.00

會員平均年紀 12 歲？

(受大量遺漏值影響)

Column Profiling

異常值偵測

去除異常值或補遺漏值

```
> summary(members[which(members$age > 0 & members$age <= 100), "age"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0    22.0    27.0    28.9    34.0    97.0
```

Column Profiling

類別分析

分析欄位中不同類別的個數

Column Profiling

類別分析

members

city

user_id	city	age	gender	registered_via	registration_date	expiration_date
Ro4esuhkwPIVgArTpe7MIGJhIJy9MW7mFEWxuM5dSXE=	1	0		7	20160119	20161227
1XjK/MwTOEShVHyWLZN3cXjwfW4QCGjWuCAszXkjLHw=	5	16	female	4	20151109	20170708
t2K3zE/WEI84LuZDshjmzCXmdCP5L1eDCavjrCm0vGU=	1	0		7	20161213	20170913
vyoxo0c6XWEPvhndjGPS1unkM5HiCCFyl0wvXISSBek=	8	21	female	9	20161124	20180115
NNm80OCBAO6WHxJSWVhXII8TAsd9HjFXTa3uSgbDG2c=	4	16	male	3	20170218	20170221
9n7Yef1vL3Z8ZN7IDCOCBzKmLuVW/viQIFSu/7DdKdU=	1	0		4	20160822	20160825
zrPEmpkTqgI3MmbtKQ6gtmfI5df8JskI4NC/A6bI6iU=	4	24	male	3	20150502	20170911
bA2m/XPUtYDVjqjP9QIpY4YJ+Lbcx+4xcxb390JU+uI=	1	0		9	20150721	20170717
FtCJaSTuuhthFdPUkx6916iNmrGRw/mLDR3doFj/NCw=	1	0		7	20151112	20170926
D9dD56KuXsMxkWsYw/o1FgNwBopA2KO4qN662hNwKo8=	1	0		3	20130614	20170127

Column Profiling

類別分析

members *city*

city	count
1	19445
3	204
4	1732
5	2634
…	

```
> group_by(members, city) %>%
+   summarise(cnt = n())
# A tibble: 21 × 2
    city   cnt
   <int> <int>
1      1 19445
2      3   204
3      4  1732
4      5  2634
5      6   913
6      7    93
7      8   289
8      9   309
9     10   216
10    11   285
# ... with 11 more rows
```

Column Profiling 類別分析

members *city*

city	count
1	19445
3	204
4	1732
5	2634
…	

有超過一半的會員住在相同城市？

（把它視為遺漏值？）

Column Profiling 類別分析

members *city*

city	count
1	19445
3	204
4	1732
5	2634
…	

處理遺漏值

one-hot encoding

Column Profiling

資料分布

分析每個欄位值的分布情況

Column Profiling

資料分布

欄位值分布是否符合預期

(過度集中在特定值或偏態)

Column Profiling

資料分布

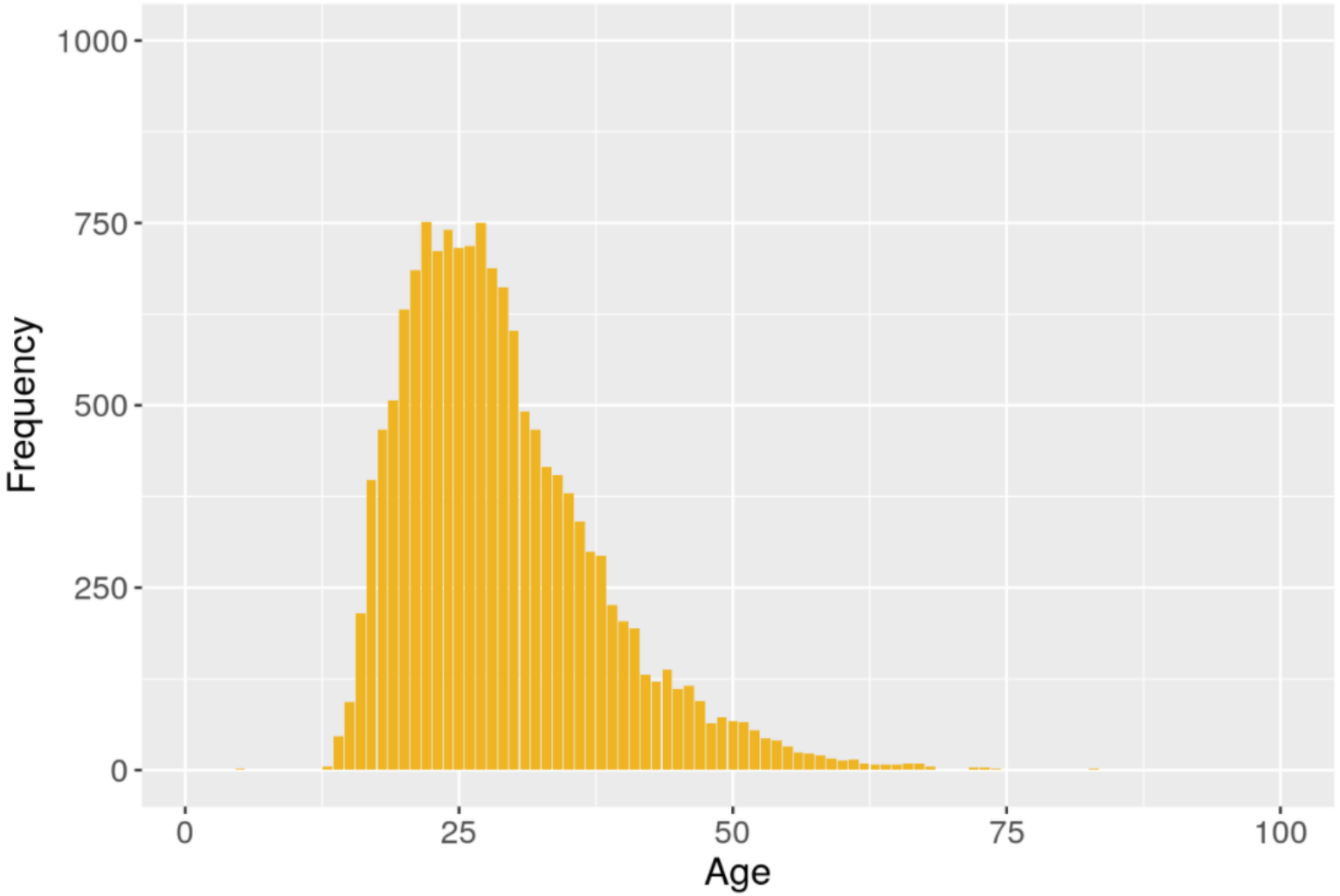
members

age

user_id	city	age	gender	registered_via	registration_date	expiration_date
Ro4esuhkwPIVgArTpe7MIGJhIJy9MW7mFEWxuM5dSXE=	1	0		7	20160119	20161227
1XjK/MwTOEShVHyWLZN3cXjwfW4QCGjWuCAszXkjLHw=	5	16	female	4	20151109	20170708
t2K3zE/WEI84LuZDshjmzCXmdCP5L1eDCavjrCm0vGU=	1	0		7	20161213	20170913
vyoxo0c6XWEPvhndjGPS1unkM5HiCCFyl0wvXISSBek=	8	21	female	9	20161124	20180115
NNm80OCBAO6WHxJSWVhXII8TAsd9HjFXTa3uSgbDG2c=	4	16	male	3	20170218	20170221
9n7Yef1vL3Z8ZN7IDCOCBzKmLuVW/viQIFSu/7DdKdU=	1	0		4	20160822	20160825
zrPEmpkTqgI3MmbtKQ6gtmfI5df8JskI4NC/A6bI6iU=	4	24	male	3	20150502	20170911
bA2m/XPUtYDVjqjP9QIpY4YJ+Lbcx+4xcxb390JU+uI=	1	0		9	20150721	20170717
FtCJaSTuuhthFdPUkx6916iNmrGRw/mLDR3doFj/NCw=	1	0		7	20151112	20170926
D9dD56KuXsMxkWsYw/o1FgNwBopA2KO4qN662hNwKo8=	1	0		3	20130614	20170127

Column Profiling

資料分布




```
> tmp_age <- group_by(members, age) %>%
+               summarise(cnt = n()) %>%
+               arrange(desc(cnt))
>
> tmp_age %>% ggplot(aes_string("age", "cnt")) +
+                geom_col(fill = "goldenrod2") +
+                labs(x = "Age", y = "Frequency") +
+                xlim(0, 100) +
+                ylim(0, 1000) +
+                readable_labs
```

Column Profiling

資料分布

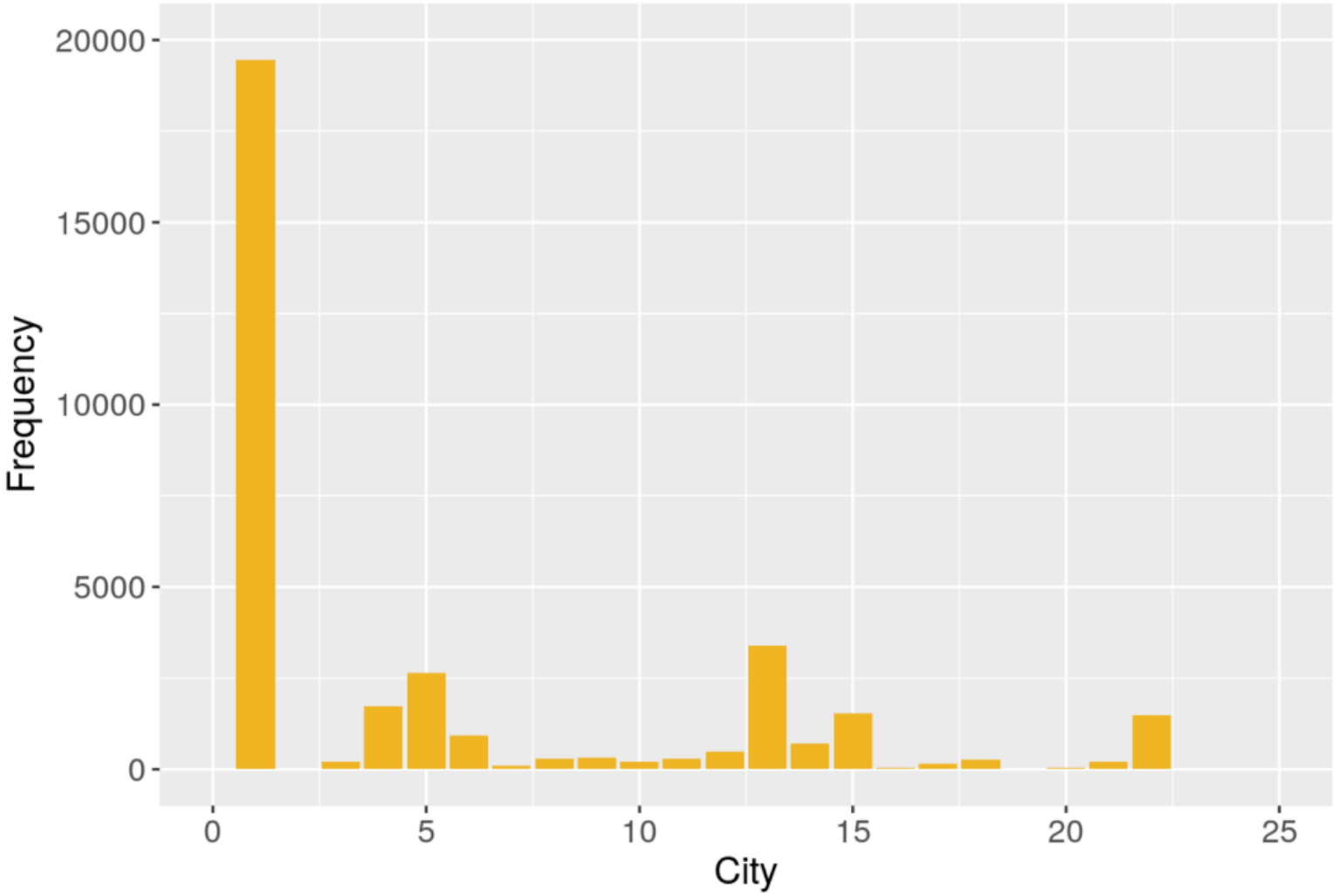
members

city

user_id	city	age	gender	registered_via	registration_date	expiration_date
Ro4esuhkwPIVgArTpe7MIGJhIJy9MW7mFEWxuM5dSXE=	1	0		7	20160119	20161227
1XjK/MwTOEShVHyWLZN3cXjwfW4QCGjWuCAszXkjLHw=	5	16	female	4	20151109	20170708
t2K3zE/WEI84LuZDshjmzCXmdCP5L1eDCavjrCm0vGU=	1	0		7	20161213	20170913
vyoxo0c6XWEPvhndjGPS1unkM5HiCCFyl0wvXISSBek=	8	21	female	9	20161124	20180115
NNm80OCBAO6WHxJSWVhXII8TAsd9HjFXTa3uSgbDG2c=	4	16	male	3	20170218	20170221
9n7Yef1vL3Z8ZN7IDCOCBzKmLuVW/viQIFSu/7DdKdU=	1	0		4	20160822	20160825
zrPEmpkTqgI3MmbtKQ6gtmfI5df8JskI4NC/A6bI6iU=	4	24	male	3	20150502	20170911
bA2m/XPUtYDVjqjP9QIpY4YJ+Lbcx+4xcxb390JU+uI=	1	0		9	20150721	20170717
FtCJaSTuuhthFdPUkx6916iNmrGRw/mLDR3doFj/NCw=	1	0		7	20151112	20170926
D9dD56KuXsMxkWsYw/o1FgNwBopA2KO4qN662hNwKo8=	1	0		3	20130614	20170127

Column Profiling

資料分布



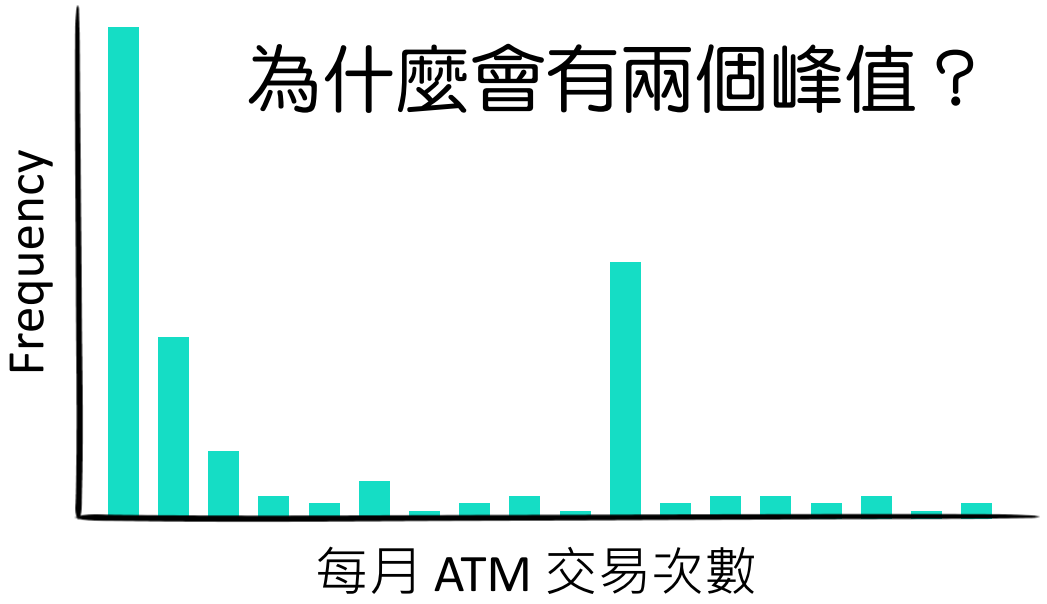

```
> tmp_city <- group_by(members, city) %>%
+                 summarise(cnt = n()) %>%
+                 arrange(desc(cnt))
>
> tmp_city %>% ggplot(aes_string("city", "cnt")) +
+                  geom_col(fill = "goldenrod2") +
+                  labs(x = "City", y = "Frequency") +
+                  xlim(0, 25) +
+                  ylim(0, 20000) +
+                  readable_labs
```

Column Profiling

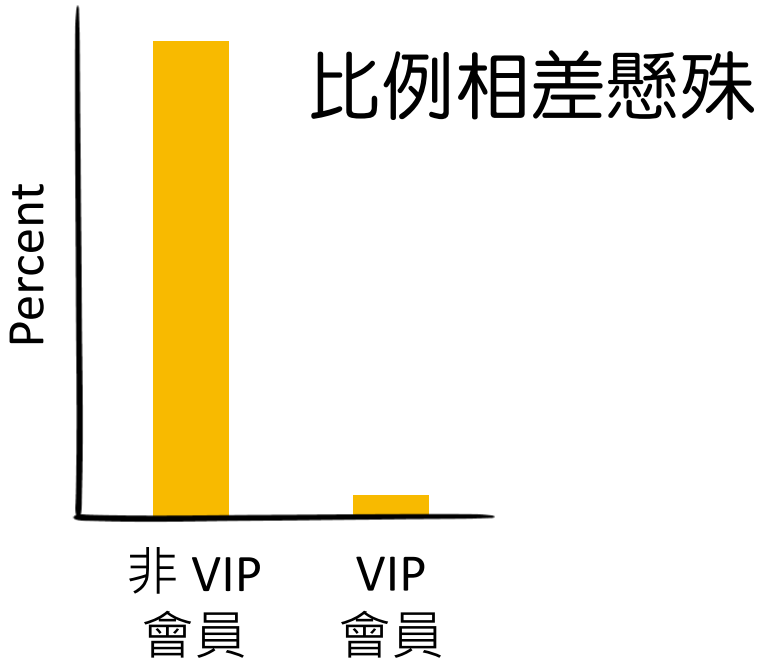
資料分布

為什麼集中在少次數的交易？

為什麼會有兩個峰值？



比例相差懸殊



Data Profiling

Column Profiling	值域分析、類別分析、資料分布、欄位型態偵測、波動偵測、異常值偵測
Cross-column Profiling	相關性分析
Table Profiling	資料筆數、主鍵唯一性分析
Cross-table Profiling	

Cross-column Profiling

相關性分析

分析欄位之間的相關程度

Cross-column Profiling

相關性分析

分析面向

feature 之間的關係

feature 與 target 間的關係

Cross-column Profiling

相關性分析

分析目的

是否滿足業務邏輯

找出潛在未知的關係

Cross-column Profiling

相關性分析

相關係數 / 統計檢定是一種方式

Pearson correlation coefficient

Chi-squared test

Spearman's rank correlation coefficient

Cross-column Profiling

相關性分析

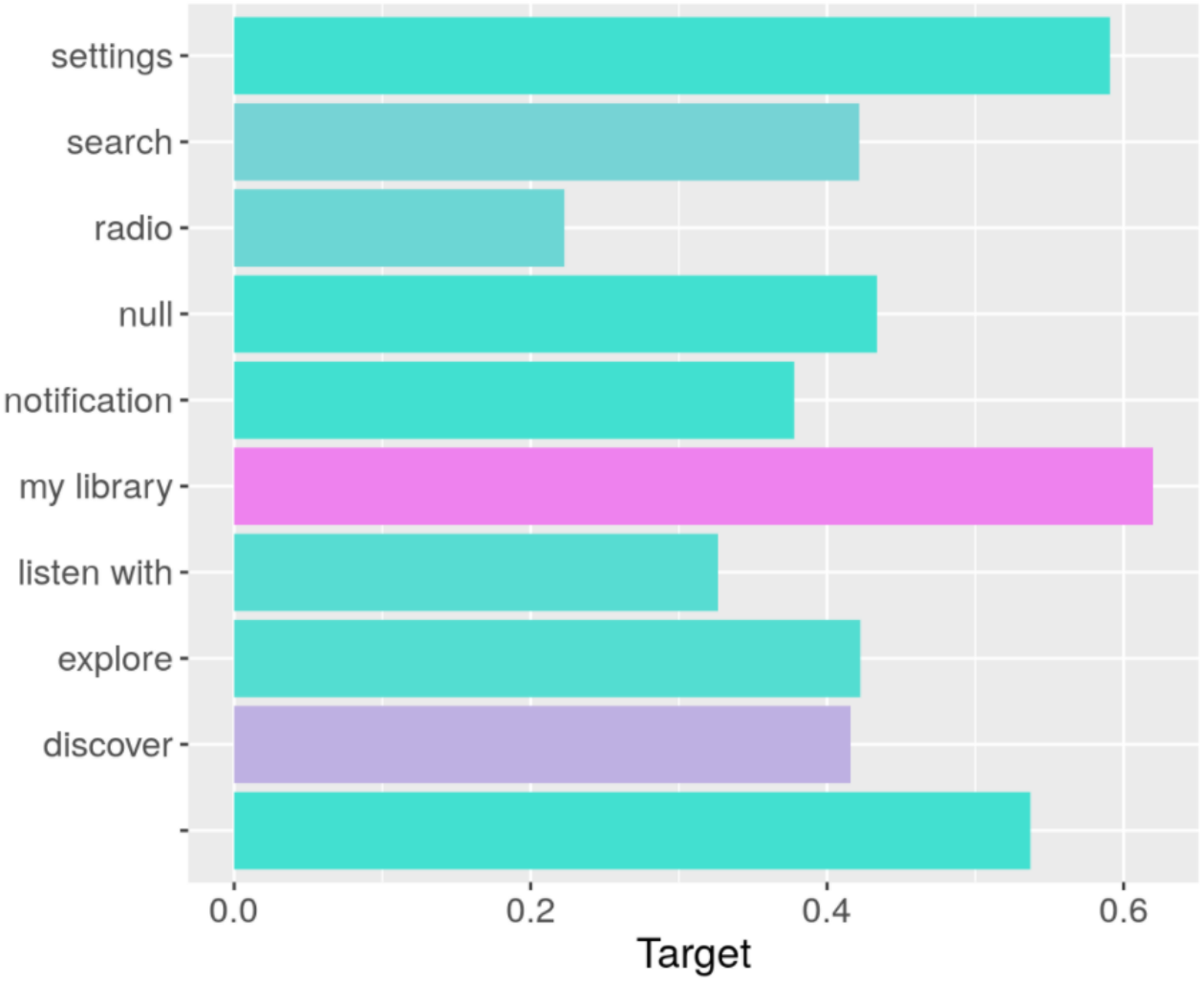
train

source_system_tab vs. target

user_id	song_id	source_system_tab	source_screen_name	source_type	target
UsJ0TVPX7D+XNIS+bD8v/MQ+rw+ZbjW0PuU5c7tYymM=	uvhnnHYdmRoKrjqIZDtkJlrjyA5Ar3yZOG2aaV2q/EI=	my library	Local playlist more	local-library	1
AXKTVBbdr2/z7m5WL1IXzlriNbun8B+7wM41MsWCYDw=	WTQ1doDa8xAetUIXq+G/3AmF8/GPtJfRtbjaaj0vGqQ=	discover	Discover Feature	album	0
XIpjLNs4f/OHphnrSPoNmsrd45AI8pL0v74FgzDZHnQ=	8OsXpp32oqorzAN4NNkVouaV0I81ddPLZ8DxMRAGJdQ=	my library	Local playlist more	local-playlist	1
hKwxGNnAnVLbafv2sDSj8hIIf88SbDfnrSeE+R+bqFc=	rFLawM8j/vW78o/S3P2GTsmy6HwkTI15P7NF/i9CkX4=	my library	Local playlist more	local-playlist	1
hOyJGmY7tciqGf1EhDj5UT86IfSSnaTHP5JWAPxYN6s=	evaPiUW4ZoRCp27WhCmBCebbpvDtSW7IBNkhQ6M/IdM=	radio	Radio	radio	0
ocdN6b9BP9Qr0B8q6+UNpcBvfErNMk8P6Ny5CK3CyKA=	dSkGPyBKM97xpD3mOsCmhwSLO42eLuqBaDAamdZNBkI=	search	Online playlist more	online-playlist	0
ocdN6b9BP9Qr0B8q6+UNpcBvfErNMk8P6Ny5CK3CyKA=	F2d52gKSwWcIHb9TAQRBtj2uJ/ly5b8p5+cEucxV9aw=	my library	Artist more	top-hits-for-artist	1
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgRnpn3ZY/abw=	r01XmbdwG4sqS8ZA7tlx+VAuxrXNuDu7siNQB+Q2jrs=	my library	Album more	album	0
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgRnpn3ZY/abw=	u70RByc176rGRQLAmXdf7WNM0vU2RrkkYnY7z1wOPyA=	my library	Album more	album	0
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgRnpn3ZY/abw=	ijDKxOgX49hGdPBoC0S7ScN7g2TbuV1C9jfmoQMI6QM=	my library	Search	online-playlist	0

Cross-column Profiling

相關性分析




```
> tmp_tab <- group_by(train, source_system_tab) %>%
+               summarize(cnt = n(),
+                         avg = mean(target)) %>%
+               arrange(desc(avg))
>
> tmp_tab %>% ggplot(aes_string("source_system_tab",
+                               "avg")) +
+                geom_col(aes(fill = cnt)) +
+                scale_fill_gradient(low = "turquoise",
+                                    high = "violet") +
+                coord_flip() +
+                labs(x = "", y = "mean_target") +
+                readable_labs
```

Cross-column Profiling

相關性分析

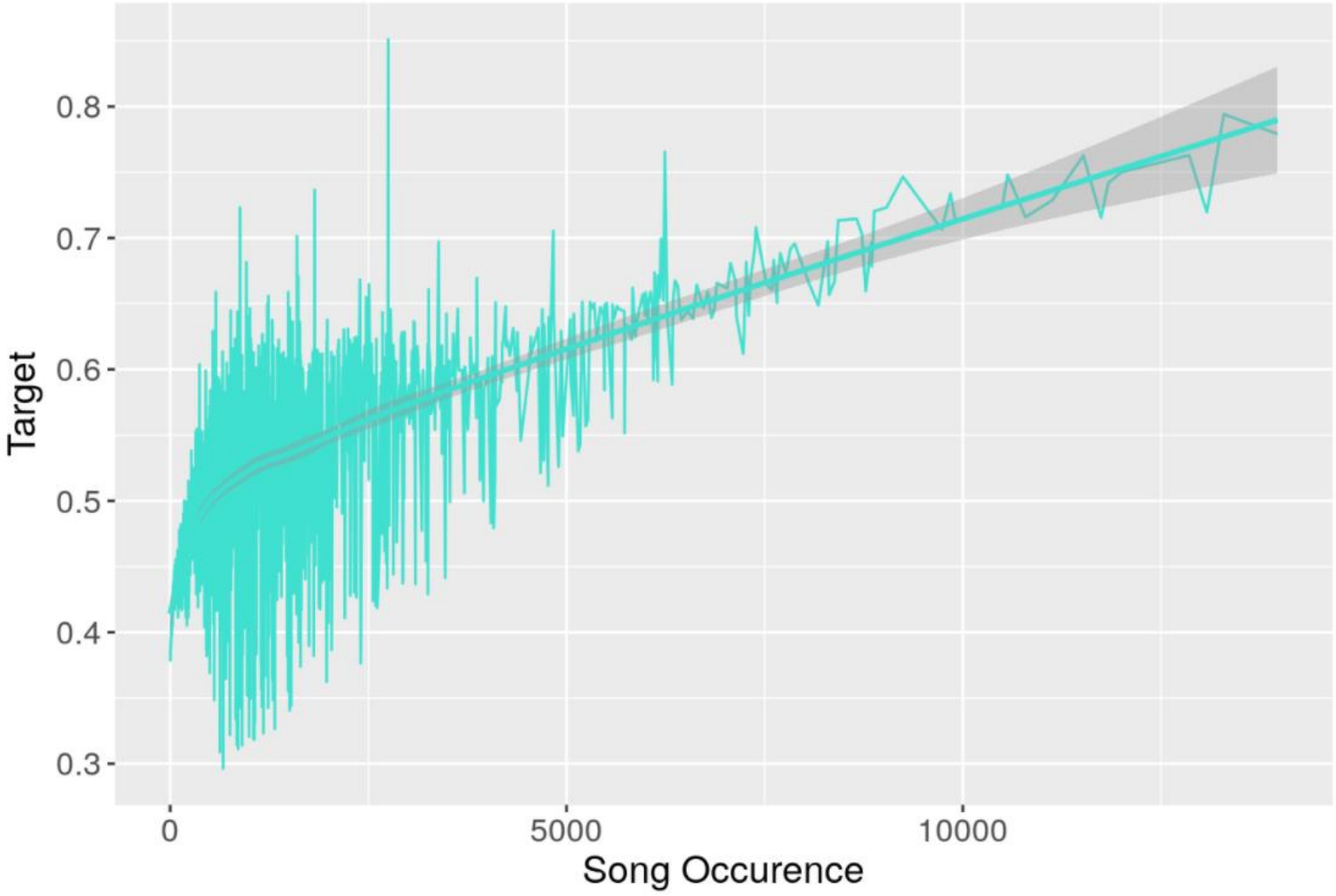
train

song_id vs. target

user_id	song_id	source_system_tab	source_screen_name	source_type	target
UsJ0TVPX7D+XNIS+bD8v/MQ+rw+ZbjW0PuU5c7tYymM=	uvhnnHYdmRoKrjqIZDtkJlrjyA5Ar3yZOG2aaV2q/EI=	my library	Local playlist more	local-library	1
AXKTVBbdr2/z7m5WL1IXzlriNbun8B+7wM41MsWCYDw=	WTQ1doDa8xAetUIXq+G/3AmF8/GPtJfRtbjaaj0vGqQ=	discover	Discover Feature	album	0
XIpjLNs4f/OHphnrSPoNmsrd45AI8pL0v74FgzDZHnQ=	8OsXpp32oqorzAN4NNkVouaV0I81ddPLZ8DxMRAGJdQ=	my library	Local playlist more	local-playlist	1
hKwxGNnAnVLbafv2sDSj8hIIf88SbDfnrSeE+R+bqFc=	rFLawM8j/vW78o/S3P2GTsmy6HwkTI15P7NF/i9CkX4=	my library	Local playlist more	local-playlist	1
hOyJGmY7tciqGf1EhDj5UT86IfSSnaTHP5JWAPxYN6s=	evaPiUW4ZoRCp27WhCmBCebbpvDtSW7IBNkhQ6M/IdM=	radio	Radio	radio	0
ocdN6b9BP9Qr0B8q6+UNpcBvfErNMk8P6Ny5CK3CyKA=	dSkGPyBKM97xpD3mOsCmhwSLO42eLuqBaDAamdZNBkI=	search	Online playlist more	online-playlist	0
ocdN6b9BP9Qr0B8q6+UNpcBvfErNMk8P6Ny5CK3CyKA=	F2d52gKSwWcIHb9TAQRBtj2uJ/ly5b8p5+cEucxV9aw=	my library	Artist more	top-hits-for-artist	1
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgRnpn3ZY/abw=	r01XmbdwG4sqS8ZA7tlx+VAuxrXNuDu7siNQB+Q2jrs=	my library	Album more	album	0
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgRnpn3ZY/abw=	u70RByc176rGRQLAmXdf7WNM0vU2RrkkYnY7z1wOPyA=	my library	Album more	album	0
hE9WBNEn/Wv0qSG2hB7/C4oQStAxcGJgRnpn3ZY/abw=	ijDKxOgX49hGdPBoC0S7ScN7g2TbuV1C9jfmoQMI6QM=	my library	Search	online-playlist	0

Cross-column Profiling

相關性分析




```
> tmp_song <-  group_by(train, song_id) %>%
+                summarize(occurence = n(),
+                          mean_target = mean(target)) %>%
+                group_by(occurence) %>%
+                summarize(no_of_items = n(),
+                          avg_target = mean(mean_target)) %>%
+                arrange(desc(avg_target))
>
> tmp_song %>% ggplot(aes(occurence, avg_target)) +
+                geom_line(color = "turquoise") +
+                geom_smooth(color = "turquoise") +
+                labs(x = "Song Occurence", y = "Target") +
+                readable_labs
```

Data Profiling

Column Profiling	值域分析、類別分析、資料分布、欄位型態偵測、波動偵測、異常值偵測
Cross-column Profiling	相關性分析
Table Profiling	資料筆數檢視、主鍵唯一性分析
Cross-table Profiling	

Data Profiling

Feature Engineering

Modeling

Data Warehouse

Data Quality Control

資料應具備

完整性 正確性 一致性 即時性


Thank
You!