통합 단지/Issue-1) 한 단지 내에 도로명 주소가 여러 개 존재

Sample : 한가람 아파트

|  |  |  |
| --- | --- | --- |
| 한가람 | 이촌로 | 00201 |
| 한가람 | 이촌로65가길 | 00051 |
| 한가람 | 이촌로65가길 | 00023 |
| 한가람 | 이촌로71길 | 00010 |
| 한가람 | 이촌로69길 | 00030 |
| 한가람 | 이촌로65길 | 00008 |

대상 단지 및 비율 : 1853/39125 = 4.74%

이슈설명 : 단지와 주소가 1:1이 아니라 1:N인 경우가 존재.

통합시 고려 필요, 추후 후속 Data에 따라 이슈가 커질 수도 아닐 수도 있음.

해당 단지는 동별로 도로명 주소가 다르며 실거래가가 특정동에 없는 경우, 주소가 누락될 수 있음. 타시스템이 누락된 주소로 단지를 표현한다면, 주소로 같은 단지인지 식별이 불가능함.

SQL]Sample 추출 :

SELECT DISTINCT real\_deal\_cmpx\_nm, road\_nm, ROAD\_NM\_CMPX\_ORGL\_NUM\_CD

FROM KRED\_DEAL\_DTL

WHERE SEQ = '11170-66'

SQL]대상단지추출

select count(\*)

from (

SELECT distinct SEQ

FROM (

SELECT DISTINCT SEQ, ROAD\_NM\_CMPX\_ORGL\_NUM\_CD

FROM KRED\_DEAL\_DTL

) A

GROUP BY SEQ

HAVING COUNT(\*) > 1

) B

;





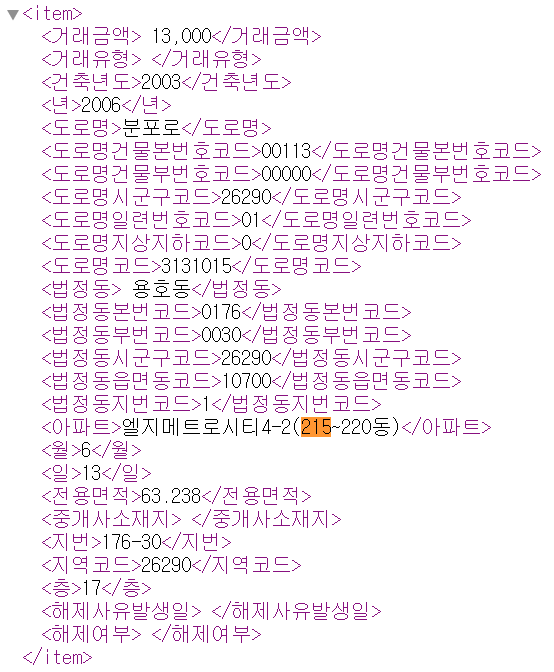
통합 단지/Issue-2) 아파트 실거래가 상세 단지일련번호 미존재 Data 존재

|  |  |  |
| --- | --- | --- |
| Sample : 엘지메트로시티4-2(215~220동) |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

대상 단지 및 비율 : 3362/39125 = 8.6%

이슈설명 : 실거래가 응답전문의 일련번호(단지일련번호) 필드가 없는 경우가 존재.

실거래가 응답 전문은 단지를 식별할수 있는 key값이 없음. 같은 단지의 거래를 실거래마다 다른 단지명으로 입력한 Data가 존재. 최근 일련번호가 추가되어 식별성이 높아졌으나 100%가 아님.



대상단지SQL] select distinct real\_Deal\_cmpx\_nm from kred\_deal\_dtl

where seq is null

통합 단지/Issue-3) 한 번지 내에 여러 단지가 존재

Sample : 광장동 218-1, 극동1차, 극동2차

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | | 200601 | 11215-4 | 극동2 | 75.55 | 7 | 1989 | 광장동 | | 200601 | 11215-3 | 극동1 | 156.92 | 1 | 1985 | 광장동 | | 200601 | 11215-3 | 극동1 | 126.94 | 9 | 1985 | 광장동 | | 200601 | 11215-4 | 극동2 | 75.55 | 12 | 1989 | 광장동 | |  |  |  |

대상 단지 및 비율 : 870/39125 = 2.22% (1개단지 리모델링, 재건축, 혹은 단순 실거래가 dATA이슈로 중복집계 포함이라)

이슈설명 : 한 번지내에 여러 단지 존재.

주소로 단지를 완전 식별하는 것은 불가능. 번지로 1차 필터 후 여러 방법을 조합해서 활용하는게 가장 좋지 않나 싶음.

SQL] SELECT DD.DEAL\_YYMM, DD.SEQ, DD.REAL\_DEAL\_CMPX\_NM, DD.ONLY\_AREA, DD.FLR, DD.CMPL\_YY, DD.LEGL\_DONG\_NM, DD.LEGL\_DONG\_HNUM\_CD, DD.DEAL\_AMT, DD.LEGL\_DONG\_ORGL\_NUM\_CD, DD.LEGL\_DONG\_VICE\_NUM\_CD

FROM KRED\_DEAL\_DTL DD

INNER JOIN (

select LEGL\_DONG\_NM, LEGL\_DONG\_ORGL\_NUM\_CD, LEGL\_DONG\_VICE\_NUM\_CD, LEGL\_DONG\_SGG\_CD, LEGL\_DONG\_UMD\_CD, LEGL\_DONG\_HNUM\_CD, count(\*)

from (

select distinct LEGL\_DONG\_NM, LEGL\_DONG\_ORGL\_NUM\_CD, LEGL\_DONG\_VICE\_NUM\_CD, LEGL\_DONG\_SGG\_CD, LEGL\_DONG\_UMD\_CD, LEGL\_DONG\_HNUM\_CD, SEQ

from KRED\_DEAL\_DTL

where seq is not null

) a

group by LEGL\_DONG\_NM, LEGL\_DONG\_ORGL\_NUM\_CD, LEGL\_DONG\_VICE\_NUM\_CD, LEGL\_DONG\_SGG\_CD, LEGL\_DONG\_UMD\_CD, LEGL\_DONG\_HNUM\_CD

having count(\*) > 1

) A

ON DD.LEGL\_DONG\_NM = A.LEGL\_DONG\_NM

AND DD.LEGL\_DONG\_ORGL\_NUM\_CD = A.LEGL\_DONG\_ORGL\_NUM\_CD

AND DD.LEGL\_DONG\_VICE\_NUM\_CD = A.LEGL\_DONG\_VICE\_NUM\_CD

AND DD.LEGL\_DONG\_SGG\_CD = A.LEGL\_DONG\_SGG\_CD

AND DD.LEGL\_DONG\_UMD\_CD = A.LEGL\_DONG\_UMD\_CD

AND DD.LEGL\_DONG\_HNUM\_CD = A.LEGL\_DONG\_HNUM\_CD

ORDER BY DD.LEGL\_DONG\_SGG\_CD, DD.LEGL\_DONG\_UMD\_CD, DD.LEGL\_DONG\_NM, DD.LEGL\_DONG\_HNUM\_CD, DD.DEAL\_YYMM, DD.LEGL\_DONG\_ORGL\_NUM\_CD, DD.LEGL\_DONG\_VICE\_NUM\_CD

;

