Hustar Homework #1
서울시 아파트 실거래 데이터 전처리

포항공과대학교 산업경영공학과

조현재

email : present@postech.ac.kr

Contents

데이터 수집

01 네이버 부동산

- 아파트 특성

국토교통부
Ministry of Land, Infrastructure and Transport

02 국토교통부

- 기간별 아파트 실거래가

데이터 분석: 네이버 부동산

41107 rows × 22 columns

length of unique apartment: 8297

- 총 아파트 정보(개별 아파트, 전용면적): 41,107개
- 아파트 관련 변수 수: 22개
- 서울시 개별 아파트 단지 수: 8,297단지

〈변수 정보〉

```
> str(apt_info)
'data.frame':   41107 obs. of  22 variables:
 $ 도          : chr  "서울특별시" "서울특별시" "서울특별시" "서울특별시" ...
 $ 지역        : chr  "종로구" "종로구" "종로구" "종로구" ...
 $ 구          : logi  NA NA NA NA NA NA ...
 $ 법정동      : chr  "청운동" "청운동" "청운동" "청운동" ...
 $ 아파트.이름: chr  "청운벽산빌리지" "청운벽산빌리지" "청운벽산빌리지" "청운벽산빌리지" ...
 $ 아파트.코드: int  12076 12076 12076 12076 12076 12076 12076 12076 12076 12076 ...
 $ 전용면적    : num  82.9 83.6 137 141.4 140.5 ...
 $ 현관구조    : chr  "계단식" "계단식" "계단식" "계단식" ...
 $ room        : int  3 3 4 4 4 4 5 5 5 5 ...
 $ toilet      : int  1 1 2 2 2 2 2 2 2 2 ...
 $ 세대.수     : int  126 126 126 126 126 126 126 126 126 126 ...
 $ dong        : int  9 9 9 9 9 9 9 9 9 9 ...
 $ y_c         : int  1988 1988 1988 1988 1988 1988 1988 1988 1988 1988 ...
 $ con_n       : chr  "벽산" "벽산" "벽산" "벽산" ...
 $ park        : int  NA NA NA NA NA NA NA NA NA NA ...
 $ per_park    : num  NA NA NA NA NA NA NA NA NA NA ...
 $ heat        : chr  "개별난방" "개별난방" "개별난방" "개별난방" ...
 $ heat_mat    : chr  "도시가스" "도시가스" "도시가스" "도시가스" ...
 $ 용적률      : int  NA NA NA NA NA NA NA NA NA NA ...
 $ 건폐율      : int  NA NA NA NA NA NA NA NA NA NA ...
 $ f_high      : int  3 3 3 3 3 3 3 3 3 3 ...
 $ f_low       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ 초등학교    : chr  "서울청운초등학교" "서울청운초등학교" "서울청운초등학교" "서울청운초등학교" ...
 $ 도로명      : chr  "서울시 종로구 청운동 1" "서울시 종로구 청운동 1" "서울시 종로구 청운동 1" "서울시 종로구 청운동 1" ...
 $ 지번주소    : chr  "서울시 종로구 자하문로36길 16-14" "서울시 종로구 자하문로36길 16-14" "서울시 종로구 자하문로36길 16-14"
종로구 자하문로36길 16-14" ...
 $ 위도        : num  37.6 37.6 37.6 37.6 37.6 ...
 $ 경도        : num  127 127 127 127 127 ...
```

<To-Do>

1. **데이터를 불러와 출력**
   경로: ./Dataset/서울특별시_unique_apt_list.csv

2. **아파트 내에서 unique한 아파트 단지(아파트 코드) 수를 확인할 것**

데이터 전처리: 네이버 부동산

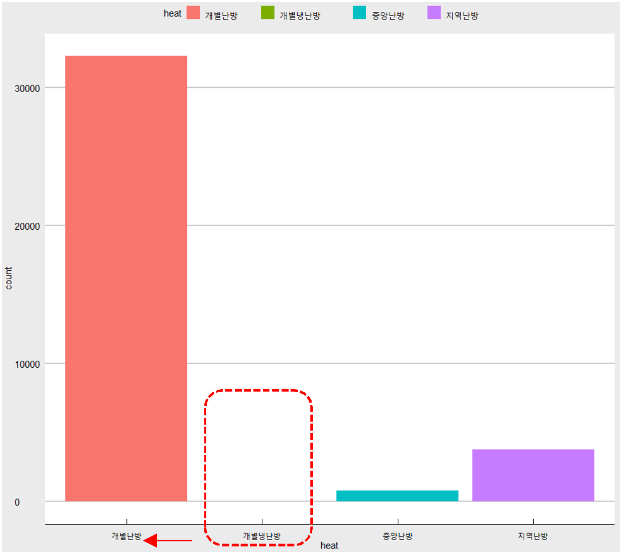
이상치(Outlier) 확인

1. Numeric 변수 정보

〉0 〉0 〈1500 〈100

```
> apt_info %>% select_if(negate(is.character)) %>% select(-아파트.코드, -위도, -경도) %>% summary()
     구            전용면적          room          toilet          세대.수             dong               y_c            park            per_park          용적률              건폐율            f_high          f_low
 Mode:logical   Min.   :  0.00   Min.   :0.000   Min.   :0.000   Min.   :   1.0   Min.   :  1.000   Min.   :1932   Min.   :    1.0   Min.   : 0.020   Min.   :     2.0   Min.   :   2.00   Min.   : 2.00   Min.   :-1.000
 NA's:43001     1st Qu.: 59.22   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:  19.0   1st Qu.:  1.000   1st Qu.:2001   1st Qu.:   19.0   1st Qu.: 0.880   1st Qu.:   225.0   1st Qu.:  26.00   1st Qu.: 7.00   1st Qu.: 5.000
                Median : 80.26   Median :3.000   Median :2.000   Median :  49.0   Median :  1.000   Median :2004   Median :   48.0   Median : 1.080   Median :   255.0   Median :  44.00   Median :10.00   Median : 7.000
                Mean   : 83.38   Mean   :2.845   Mean   :1.654   Mean   : 244.9   Mean   :  3.672   Mean   :2003   Mean   :  299.8   Mean   : 1.193   Mean   :   324.3   Mean   :  40.99   Mean   :12.02   Mean   : 8.719
                3rd Qu.: 97.06   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.: 208.0   3rd Qu.:  3.000   3rd Qu.:2009   3rd Qu.:  253.0   3rd Qu.: 1.330   3rd Qu.:   309.0   3rd Qu.:  55.00   3rd Qu.:15.00   3rd Qu.:11.000
                Max.   :462.94   Max.   :8.000   Max.   :5.000   Max.   :9510.0   Max.   :124.000   Max.   :2019   Max.   :12456.0   Max.   :11.950   Max.   :215041.0   Max.   :2457.00   Max.   :69.00   Max.   :54.000
                                                                                                    NA's   :37     NA's   :3027     NA's   :3034     NA's   :5158       NA's   :6108
```

Character 변수 정보



Numeric 변수 전처리
- 전용면적, 방 수가 0인 데이터 제거
- 용적률과 건폐율 수치가 비이상적인 데이터 제거

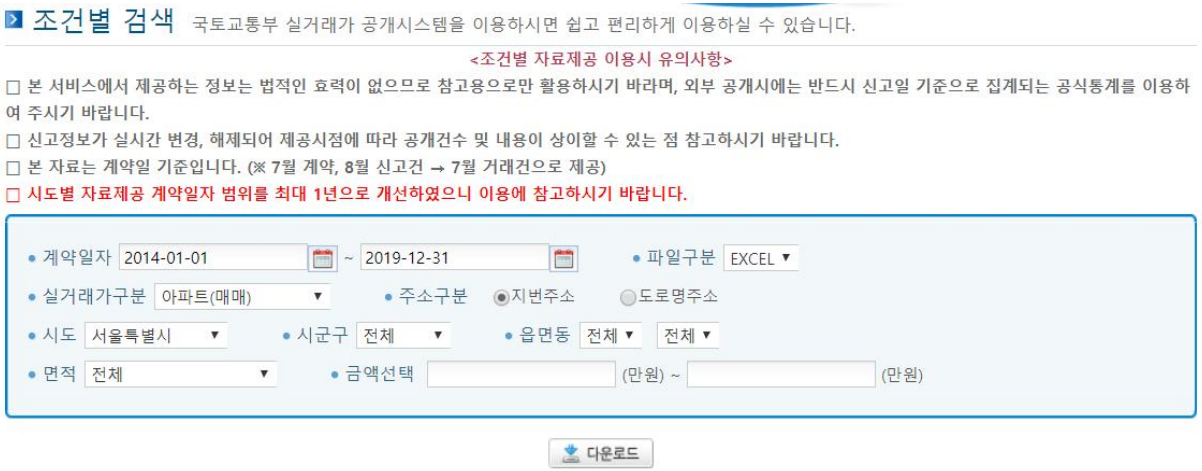
Character 변수 전처리
- 개별냉난방은 개별난방으로 변경 (의미 동일)

<To-Do>

1. 위 기준을 적용하여 이상치 데이터를 제거/수정할 것

POSTECH
POHANG UNIVERSITY OF SCIENCE AND TECHNOLOGY

데이터 수집: 국토교통부 실거래가



〈변수 정보〉

```
> str(trading_all)
'data.frame':  813026  obs. of  11 variables:
 $ YYYY      : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
 $ MM        : int  1 4 2 2 3 3 4 5 6 6 ...
 $ 시군구    : chr  "서울특별시 강남구 개포동" "서울특별시 강남구 개포동" "서울특별시 강남구 개포동" "서울특별시 강남구 개포동" ...
 $ 번지      : chr  "655-2" "658-1" "652" "652" ...
 $ 단지명    : chr  "개포2차현대아파트(220)" "개포6차우성아파트1동~8동" "개포우성3차" "개포우성3차" ...
 $ area      : num  77.8 80 161 133.5 133.5 ...
 $ price     : num  55000 67000 115000 110000 100000 78500 78000 104000 76000 103000 ...
 $ floor     : num  7 4 11 9 15 6 15 6 13 5 ...
 $ birth_year: num  1988 1987 1984 1984 1984 ...
 $ 도로명    : chr  "언주로" "언주로" "개포로" "개포로" ...
 $ age_c     : num  26 27 30 30 30 30 30 30 30 30 ...
```

<To-Do>

1. 국토교통부 실거래가 데이터를 불러와 데이터 개관을 확인
경로: ./Dataset/서울특별시_매매가_info.csv

데이터 병합: 국토교통부 실거래가

- 데이터 병합 -〉 **join key**가 필요

- Join key 후보
  1. 아파트 이름+전용면적 크기
     ① 다른 지역의 동일한 아파트 이름인 경우 (코드번호 다름)

     → 사용하기 어려움

  2. 주소+전용면적 크기

<To-Do>

1. **주소 + 전용면적을 합친 데이터를 기준으로 국토교통부 데이터와 네이버 부동산 데이터를 병합할 것**

데이터 병합: 국토교통부 실거래가

<To-Do>

1. **방 개수, 화장실 개수에 대해서 NA값에 대해 방개수는 평균값으로, 화장실 개수는 최빈값(mode)으로 Impute할 것**

81
813026

최종 데이터셋

I. 데이터 크기

```
> dim(final_info)
[1] 454709     39
```

- 병합 후 데이터 수: 813,026 개
- 병합 후 변수 수: 33개

II. 데이터 요약

```
> summary(final_info)
      YYYY            MM              도                지역               구              법정동            단지명           아파트.코드
 Min.   :2014   Min.   : 1.000   Length:454709      Length:454709      Mode:logical   Length:454709      Length:454709      Min.   :     3
 1st Qu.:2015   1st Qu.: 4.000   Class :character   Class :character   NA's:454709    Class :character   Class :character   1st Qu.:   842
 Median :2016   Median : 7.000   Mode  :character   Mode  :character                  Mode  :character   Mode  :character   Median :  3426
 Mean   :2016   Mean   : 6.515                                                                                             Mean   : 20603
 3rd Qu.:2018   3rd Qu.: 9.000                                                                                             3rd Qu.: 22570
 Max.   :2019   Max.   :12.000                                                                                             Max.   :126215

      area             room            toilet         현관구조            세대.수           dong          birth_year      con_n                park
 Min.   : 10.32   Min.   :1.000   Min.   :1.000   Length:454709      Min.   :   1   Min.   :  1.00   Min.   :1970   Length:454709      Min.   :   1
 1st Qu.: 59.87   1st Qu.:3.000   1st Qu.:1.000   Class :character   1st Qu.: 261   1st Qu.:  3.00   1st Qu.:1996   Class :character   1st Qu.:  278
 Median : 84.57   Median :3.000   Median :2.000   Mode  :character   Median : 642   Median :  7.00   Median :2001   Mode  :character   Median :  674
 Mean   : 79.70   Mean   :2.966   Mean   :1.659                      Mean   :1018   Mean   : 11.77   Mean   :2001                      Mean   : 1154
 3rd Qu.: 84.97   3rd Qu.:3.000   3rd Qu.:2.000                      3rd Qu.:1335   3rd Qu.: 15.00   3rd Qu.:2007                      3rd Qu.: 1415
 Max.   :424.32   Max.   :8.000   Max.   :5.000                      Max.   :9510   Max.   :124.00   Max.   :2019                      Max.   :12456
                                                                                                                                      NA's   :13721
    per_park          heat            heat_mat            용적률            건폐율          f_high          f_low          초등학교             위도
 Min.   : 0.020   Length:454709      Length:454709      Min.   :   2   Min.   : 2.00   Min.   : 2.00   Min.   : 1.00   Length:454709      Min.   :37.43
 1st Qu.: 0.970   Class :character   Class :character   1st Qu.: 225   1st Qu.:18.00   1st Qu.:15.00   1st Qu.: 8.00   Class :character   1st Qu.:37.51
 Median : 1.130   Mode  :character   Mode  :character   Median : 257   Median :21.00   Median :18.00   Median :12.00   Mode  :character   Median :37.55
 Mean   : 1.144                                        Mean   : 287   Mean   :24.68   Mean   :19.11   Mean   :12.13                      Mean   :37.55
 3rd Qu.: 1.290                                        3rd Qu.: 308   3rd Qu.:26.00   3rd Qu.:23.00   3rd Qu.:15.00                      3rd Qu.:37.60
 Max.   :11.950                                        Max.   :1477   Max.   :96.00   Max.   :69.00   Max.   :54.00                      Max.   :37.69
 NA's   :13721
      경도          지번주소             price              floor             age_c          dis_sub          dis_park         dis_ele
 Min.   :126.8   Length:454709      Min.   :   700   Min.   :-4.000   Min.   : 0.00   Min.   :   2.646   Min.   :  55.74   Min.   :   6.78
 1st Qu.:126.9   Class :character   1st Qu.: 34000   1st Qu.: 5.000   1st Qu.:10.00   1st Qu.: 391.007   1st Qu.: 634.04   1st Qu.: 214.92
 Median :127.0   Mode  :character   Median : 46900   Median : 8.000   Median :15.00   Median : 596.144   Median : 970.38   Median : 322.94
 Mean   :127.0                      Mean   : 57329   Mean   : 9.418   Mean   :15.32   Mean   : 772.755   Mean   :1035.23   Mean   : 337.12
 3rd Qu.:127.1                      3rd Qu.: 67000   3rd Qu.:13.000   3rd Qu.:20.00   3rd Qu.: 946.882   3rd Qu.:1357.78   3rd Qu.: 435.89
 Max.   :127.2                      Max.   :700000   Max.   :69.000   Max.   :49.00   Max.   :5583.588   Max.   :3268.24   Max.   :1810.04

    dis_mid          dis_high         dis_univ         dis_muse          dis_Gu
 Min.   :   2.587   Min.   :  24.62   Min.   :  25.14   Min.   :  35.34   Min.   :  13.49
 1st Qu.: 289.105   1st Qu.: 329.44   1st Qu.: 967.51   1st Qu.:1037.49   1st Qu.:1203.52
 Median : 436.379   Median : 505.46   Median :1574.95   Median :1634.26   Median :1839.61
 Mean   : 470.248   Mean   : 575.54   Mean   :1860.67   Mean   :1843.09   Mean   :1931.14
 3rd Qu.: 614.382   3rd Qu.: 762.57   3rd Qu.:2499.60   3rd Qu.:2487.65   3rd Qu.:2510.16
 Max.   :2130.155   Max.   :2837.36   Max.   :7111.54   Max.   :6839.08   Max.   :6521.69
```

III. 변수 정보

변수 특성	변수명
시계열	거래년월 (YYYY, MM)
주택특성	면적, 방/화장실 갯수, 현관구조
이웃특성	용적률, 건폐율, 최고층, 최저층, 세대 수, 단지 크기, (세대 당) 주차 대수, 난방 방식(개별/중앙), 연료
입지특성	법정동
종속변수	거래가격
보충 특성	도, 지역, 구, 단지명, 아파트 코드, 지번주소, 위경도, 아파트 준공년도, 건축사 명, 주변 초등학교

<To-Do>

1. 데이터 전처리 후 최종 샘플링된 데이터 출력

과제 제출

- 데이터 전처리에 사용한 코드파일(.ipynb 형식)

- 코드에 대한 설명(.ipynb 파일 내 주석을 이용하여 설명할 것)

- 데이터 전처리 후 생성된 최종 데이터 셋

위 결과들을 압축하여 #1_이름.zip 형식으로 LMS에 제출할 것
(과제 기한: 2020-09-08 23:59)

Thank You ☺