

Hustar Homework #1

서울시 아파트 실거래 데이터 전처리

포항공과대학교 산업경영공학과

조현재

email : present@postech.ac.kr

Contents

1. Introduction
2. Data collection & preprocessing
3. Data description
4. Future works

데이터 수집



- 01 네이버 부동산
- 아파트 특성



- 02 국토교통부
- 기간별 아파트 실거래가

Data description

데이터 분석: 네이버 부동산

```
> dim(apt_info)
[1] 43001    27
> apt_info %>% select(아파트.코드) %>% unique() %>% nrow()
[1] 8299
```

- 총 아파트 정보(개별 아파트, 전용면적): 43,001개
- 아파트 관련 변수 수: 27개
- 서울시 개별 아파트 단지 수: 8,299단지

<변수 정보>

```
> str(apt_info)
'data.frame':   43001 obs. of  27 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  "청운백산빌리지" "청운백산빌리지" "청운백산빌리지" "청운백산빌리지" ...
 $ 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  "백산" "백산" "백산" "백산" ...
 $ 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" "서울시 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

1. Numeric 변수 정보

Character 변수 정보



- 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

Data collection

데이터 수집: 국토교통부 실거래가

▶ **조건별 검색** 국토교통부 실거래가 공개시스템을 이용하시면 쉽고 편리하게 이용하실 수 있습니다.

<조건별 자료제공 이용시 유의사항>

- ☐ 본 서비스에서 제공하는 정보는 법적인 효력이 없으므로 참고용으로만 활용하시기 바라며, 외부 공개시에는 반드시 신고일 기준으로 집계되는 공식통계를 이용하여 주시기 바랍니다.
- ☐ 신고정보가 실시간 변경, 해제되어 제공시점에 따라 공개건수 및 내용이 상이할 수 있는 점 참고하시기 바랍니다.
- ☐ 본 자료는 계약일 기준입니다. (※ 7월 계약, 8월 신고건 → 7월 거래건으로 제공)
- ☐ **시도별 자료제공 계약일자 범위를 최대 1년으로 개선하였으니 이용에 참고하시기 바랍니다.**

• 계약일자 2014-01-01 ~ 2019-12-31

• 파일구분 EXCEL ▼

• 실거래가구분 아파트(매매)

• 주소구분 지번주소 도로명주소

• 시도 서울특별시

• 시군구 전체

• 읍면동 전체

• 면적 전체

• 금액선택 (만원) ~ (만원)

다운로드

<변수 정보>

```
> str(trading_all)
'data.frame':  569331 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 ...
```

<To-Do>

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

데이터 병합: 국토교통부 실거래가

- 데이터 병합 -> join key가 필요
- Join key 후보
 1. 아파트 이름+전용면적 크기
 - ① 다른 지역의 동일한 아파트 이름인 경우 (코드번호 다름)
 2. 주소+전용면적 크기

→ 사용하기 어려움

<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

Data Description

최종 데이터셋

I. 데이터 크기

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

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    NA's:454709     Mode :character    Mode :character    Median : 3426
Mean   :2016   Mean   : 6.515                      3rd Qu.:2018   3rd Qu.: 9.000                      Mean   :20603
3rd Qu.:2018   3rd Qu.:2018   3rd Qu.:2018   3rd Qu.:2018   3rd Qu.:2018   3rd Qu.:2018   3rd Qu.:2018   3rd Qu.:22570
Max.   :2019   Max.   :12.000                      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   :1.144   Mean   :1.144   Mean   : 287   Mean   :24.68   Mean   :19.11   Mean   :12.13   Mean   :37.55
3rd Qu.: 1.290   3rd Qu.: 1.290   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.   :11.950   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   1st Qu.:34000   1st Qu.: 46900   1st Qu.: 8.000   1st Qu.:10.00   1st Qu.:391.007   1st Qu.: 634.04   1st Qu.:214.92
Median :127.0   Median :46900   Median :57329   Median :15.32   Median :15.00   Median :596.144   Median : 970.38   Median :322.94
Mean   :127.0   Mean   :57329   Mean   :67000   Mean   :15.32   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.: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
```

- 원본 거래 데이터 수: 569,331개
- 병합 후 데이터 수: 454,709개 (Loss 20.1%)

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 생성된 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 😊
