

---

[ SQL 프로젝트 ]

# 서울시 부동산 실거래가 분석: 가격에 가장 큰 영향을 미치는 요인은 무엇일까?



빅데이터 분석 19기 5조  
노승욱, 안민영

---

---

# 목차

- 
- 01 데이터 소개
  - 02 분석 방향 설정 ( 회귀분석 )
  - 03 SQL
  - 04 결론
-

# 1. 데이터 소개



01

서울시 부동산 실거래가 현황 [house\_2024]

	접수연도	자치구코드	자치구명	법정동코드	법정동명	지번구분	지번구분명	본번	부번	건물명	계약일	물건금액	건물면적	토지면적	층	권리구분	취소일	건축년도	건물용도	신고구분	공인중개사시군명
1	2024	11650	서초구	10800	서초동	1	대지	1507	0025	세원플러스빌 (1507)	20240511	65000	48.65	23	2	(null)	(null)	2001	연립다세대	중개거래	서울 서초구
2	2024	11440	마포구	10800	대흥동	1	대지	0807	0000	마포그랑자이 (2단지)	20240511	136000	59.96	0	8	(null)	(null)	2020	아파트	중개거래	서울 마포구
3	2024	11200	성동구	10700	행당동	1	대지	0346	0000	행당한진타운	20240511	139000	114.62	0	17	(null)	(null)	2000	아파트	중개거래	서울 강남구
4	2024	11470	양천구	10300	신월동	1	대지	0068	0004	미래팰리스	20240511	21700	36.67	23	2	(null)	(null)	2016	연립다세대	중개거래	서울 강서구
5	2024	11650	서초구	10800	서초동	1	대지	1665	0003	공간6차	20240511	30100	28.79	18	2	(null)	(null)	2017	연립다세대	중개거래	서울 서초구
6	2024	11440	마포구	12500	성산동	1	대지	0620	0006	미란채 (620-6)	20240511	43000	75.09	48	4	(null)	(null)	2003	연립다세대	중개거래	서울 마포구
7	2024	11200	성동구	10200	하왕십리동	1	대지	1066	0000	텐즈힐 (1단지)	20240511	150000	84.88	0	3	(null)	(null)	2015	아파트	중개거래	서울 성동구

자치구명	건물명	계약일	물건금액	...	건물면적	층	건축년도	건물용도
서초구	마포그랑자이(2단지)	20240511	136000	...	59.96	8	2020	아파트

출처: 공공 데이터 포털

# 1. 데이터 소개

02

서울시 대규모 점포 인허가 현황 [shop]

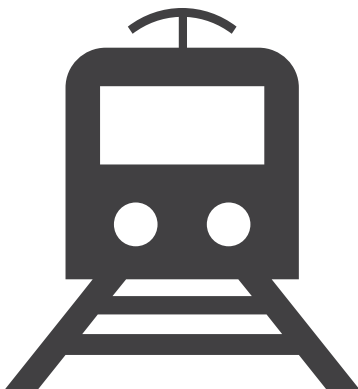


개방자치단체코드	관리번호	인허가일자	인허가취소일자	영업상태코드	영업상태명	상세영업상태코드	상세영업상태명	폐업일자	휴업시작일자	휴업종료일자	재개업일자	전화번호	소재지면적	소재지우편번호	지번주소
1	3100000	1998310...	98/12/31	(null)	1 영업/정상	1	정상영업	(null)	(null)	(null)	(null)	029314825	7863.67	139-748	서울특별시 노원구 상계동 173번지
2	3020000	2008302...	87/11/10	(null)	1 영업/정상	1	정상영업	(null)	(null)	(null)	(null)	02-704-3013	11669.19	(null)	서울특별시 용산구 한강로3가 1번지 1호 나진상가
3	3030000	1962303...	62/01/25	(null)	1 영업/정상	1	정상영업	(null)	(null)	(null)	(null)	00024644426	2861	(null)	서울특별시 성동구 성수동2가 335번지
4	3070000	2023307...	23/06/08	(null)	1 영업/정상	1	정상영업	(null)	(null)	(null)	(null)	02-918-5651	80.49	(null)	서울특별시 성북구 하월곡동 62-1
5	3050000	2000305...	60/06/29	(null)	1 영업/정상	1	정상영업	(null)	(null)	(null)	(null)	02-967-8721	15540.7	(null)	서울특별시 동대문구 제기동 1019 경동시장
6	3060000	2011306...	11/07/04	(null)	1 영업/정상	1	정상영업	(null)	(null)	(null)	(null)	024918546	466.1	131-122	서울특별시 중랑구 중화2동 314번지 1호
7	.....	.....	.....	.....	1 영업/정상	1	정상영업	.....	.....	.....	.....	.....	.....	.....	서울특별시 ..... ..번지 ..호

관리번호	영업상태코드	폐업/휴업일	...	소재지면적	주소	사업장	업태구분
2020318023907500001	1	( null )		84573.34	서울특별시 영등포구 여의대로 108	THE HYUNDAI SEOUL	백화점

출처: 서울 열린데이터 광장

# 1. 데이터 소개



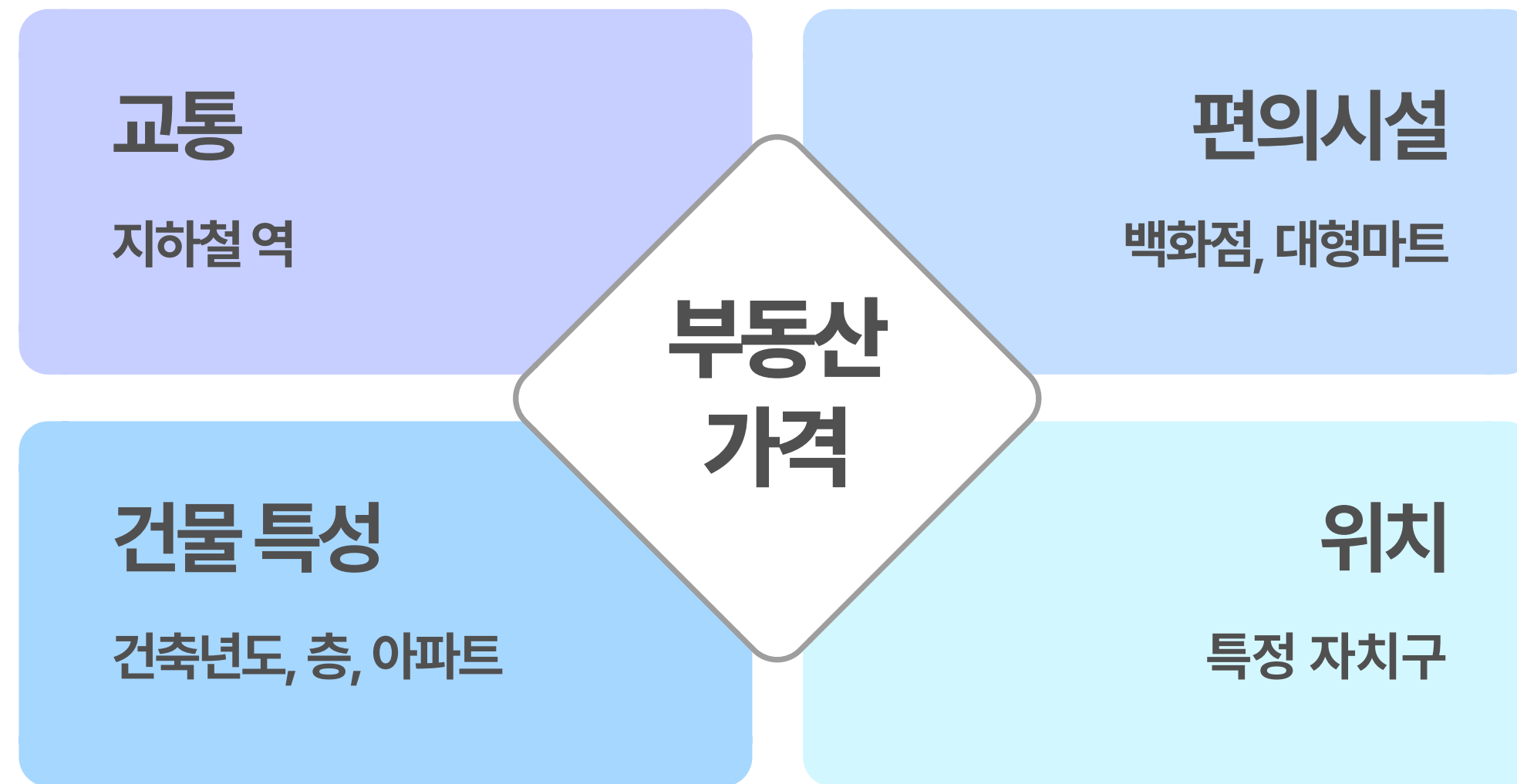
03

서울시 자치구별 지하철역 정보 [subway]

	자치구명	해당역	역개수
1	강남구	삼성 (2), 선릉 (2), 역삼 (2), ...	21
2	강동구	천호 (5), 강동 (5), 길동 (5), ...	14
3	강북구	수유 (4), 미아 (4), 미아사거리 (4)	3
4	강서구	방화 (5), 개화산 (5), 김포공항...	9
5	관악구	낙성대 (2), 서울대입구 (2), 봉...	4
6	광진구	건대입구 (2), 구의 (2), 강변 (2...	11
7	구로구	구로디지털단지 (2), 대림 (2), ...	7
8	금천구	가산디지털단지 (7)	1
9	노원구	당고개 (4), 상계 (4), 노원 (4) ...	13
10	도봉구	창동 (4), 쌍문 (4), 도봉산 (7)	3

- 1 ~ 9호선
- 환승역 중복 카운트

## 2. 분석 방향 : 회귀분석



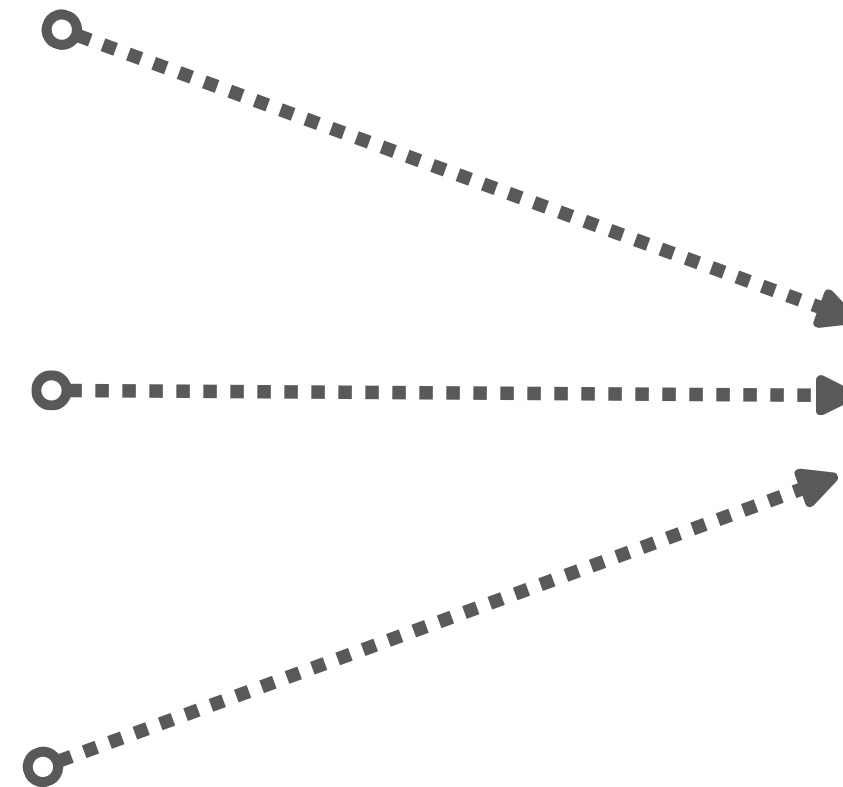
### 3. SQL (1)테이블 구성 및 전처리

DATA

01 [ HOUSE\_2024 ]  
부동산 실거래가 정보

02 [ SHOP ]  
대규모점포 인허가 현황

03 [ SUBWAY ]  
자치구별 지하철역 정보



회귀분석 데이터

04 [ HOUSE\_TEST ]

### 3. SQL (1)테이블 구성 및 전처리

#### DATA

01

[ HOUSE\_2024 ]  
부동산 실거래가 정보

02

[ SHOP ]  
대규모점포 인허가 현황

03

[ SUBWAY ]  
자치구별 지하철역 정보

▶ 테이블 생성  
[ house\_2024]

```
create table house_2024
(
    접수연도          number(38),
    자치구코드        number(38),
    자치구명          varchar2(26),
    법정동코드        number(38),
    법정동명          varchar2(26),
    지번구분          number(38),
    지번구분명        varchar2(26),
    본번              varchar2(26),
    부번              varchar2(26),
    건물명            varchar2(100),
    계약일            number(38),
    물건금액          number(38),
    건물면적          number(38,4),
    토지면적          number(38,6),
    층                number(38),
    권리구분          varchar2(26),
    취소일            number(38),
    건축년도          number(38),
    건물용도          varchar2(26),
    신고구분          varchar2(26),
    공인중개사시군명  varchar2(100) );
```



### 3. SQL (1)테이블 구성 및 전처리

#### DATA

01 [ HOUSE\_2024 ]  
부동산 실거래가 정보

02 [ SHOP ]  
대규모점포 인허가 현황

03 [ SUBWAY ]  
자치구별 지하철역 정보

#### ▶ 층 수 업데이트

```
update house_2024  
set 층 = 1  
where 건물용도 = '단독다가구' ;
```

### 3. SQL (1)테이블 구성 및 전처리

#### DATA

01

[ HOUSE\_2024 ]

부동산 실거래가 정보

02

[ SHOP ]

대규모점포 인허가 현황

03

[ SUBWAY ]

자치구별 지하철역 정보

▶ 테이블 생성  
[ shop ]

▶ PK설정  
- 관리번호 컬럼

```
CREATE TABLE SHOP
```

```
(   개발자치단체코드 NUMBER(38,0),  
    관리번호 NUMBER(38,0) constraint shop_number_pk primary key,  
    인허가일자 DATE,  
    인허가취소일자 VARCHAR2(26),  
    영업상태코드 number(38,0),  
    영업상태명 VARCHAR2(50),  
    상세영업상태코드 varchar2(26),  
    상세영업상태명 VARCHAR2(26),  
    폐업일자 VARCHAR2(26),  
    휴업시작일자 VARCHAR2(26),  
    휴업종료일자 VARCHAR2(26),  
    재개업일자 VARCHAR2(26),  
    전화번호 VARCHAR2(26),  
    소재지면적 NUMBER(38,2),  
    소재지우편번호 VARCHAR2(26),  
    지번주소 VARCHAR2(128),  
    도로명주소 VARCHAR2(200),  
    도로명우편번호 VARCHAR2(26),  
    사업장명 VARCHAR2(200),  
    최종수정일자 DATE,  
    데이터갱신구분 VARCHAR2(26),  
    데이터갱신일자 VARCHAR2(26),  
    업태구분명 VARCHAR2(40),  
    좌표엑스 NUMBER(38,15),  
    좌표와이 NUMBER(38,15),  
    점포구분명 VARCHAR2(26) );
```

### 3. SQL (1)테이블 구성 및 전처리

#### DATA

01

[ HOUSE\_2024 ]  
부동산 실거래가 정보

02

[ SHOP ]  
대규모점포 인허가 현황

03

[ SUBWAY ]  
자치구별 지하철역 정보

▶ regexp\_substr로 자치구명 추출

▶ 자치구 별 점포 개수 count

```
create table shop_cnt
as
select decode(지번주소,null,regexp_substr(도로명주소,'[^ ]+',7,1),regexp_substr(지번주소,'[^ ]+',7,1)) as 자치구명,
       count(*) 대규모점포수
from shop s
where  영업상태코드 = 1
      and 소재지면적 > 3000
      and 업태구분명 in ('백화점', '대형마트', '복합쇼핑몰')
      and 사업장명 != '종로세운상가'
      and 사업장명 != '영풍문고'
      and 사업장명 != '현대화유통기업(주)'
      and 사업장명 != '(주)미소종합유통'
group by decode(지번주소,null,regexp_substr(도로명주소,'[^ ]+',7,1),regexp_substr(지번주소,'[^ ]+',7,1)) ;
```

지번주소

서울특별시 노원구 상계동 173번지

### 3. SQL (1)테이블 구성 및 전처리

#### DATA

01 [ HOUSE\_2024 ]  
부동산 실거래가 정보

02 [ SHOP ]  
대규모점포 인허가 현황

03 [ SUBWAY ]  
자치구별 지하철역 정보

▶ [ shop\_cnt ] 생성

	자치구명	대규모점포수
1	노원구	5
2	도봉구	3
3	서초구	5
4	구로구	8
5	강남구	9
6	강북구	3
7	종구	6
8	용산구	3
9	성동구	2
10	동대문구	3
11	영등포구	10
12	강서구	3
13	송파구	6
14	관악구	5
15	성북구	8
16	양천구	3
17	종로구	2
18	광진구	3
19	중랑구	7
20	은평구	2
21	서대문구	1
22	마포구	4
23	금천구	5
24	강동구	4

### 3. SQL (1)테이블 구성 및 전처리

#### DATA

01 [ HOUSE\_2024 ]  
부동산 실거래가 정보

02 [ SHOP ]  
대규모점포 인허가 현황

03 [ SUBWAY ]  
자치구별 지하철역 정보

#### ▶ 테이블 생성 [ subway ]

```
create table subway  
( 자치구명 varchar2(200),  
  해당역 varchar2(2000),  
  역개수 number(10) ) ;
```

### 3. SQL (1)테이블 구성 및 전처리

DATA

- 01

[ HOUSE\_2024 ]  
부동산 실거래가 정보
- 02

[ SHOP ]  
대규모점포 인허가 현황
- 03

[ SUBWAY ]  
자치구별 지하철역 정보

▶ 경기도 시 삭제

```
delete from subway
where 자치구명 in ('고양시', '광명시', '성남시', '의정부시', '하남시') ;
```

자치구명	해당역	역개수
1 강남구	삼성 (2), 선릉 (2), 역삼 (2...	21
2 강동구	천호 (5), 강동 (5), 길동 (5...	14
3 강북구	수유 (4), 미아 (4), 미아사...	3
4 강서구	방화 (5), 개화산 (5), 김포...	9
5 관악구	낙성대 (2), 서울대입구 (2), ...	4
6 광진구	건대입구 (2), 구의 (2), 강...	11
7 구로구	구로디지털단지 (2), 대림 (2...	7
8 금천구	가산디지털단지 (7)	1
9 노원구	당고개 (4), 상계 (4), 노원...	13
10 도봉구	창동 (4), 쌍문 (4), 도봉산 (7)	3
	(1), 청...	6
	, 동작...	12
	(2), 신...	16
14 서대문구	홍정로 (2), 홍제 (3), 무악...	5
15 서초구	교대 (2), 서초 (2), 방배 (2...	12
16 성동구	상왕십리 (2), 왕십리 (2), ...	14
17 성북구	길음 (4), 성신여대입구 (4), ...	9
18 송파구	잠실나루 (2), 잠실 (2), 잠...	28
19 양천구	양천구청 (2), 신정네거리 (2...	5
20 영등포구	문래 (2), 영등포구청 (2), ...	11
21 용산구	서울 (4), 숙대입구 (4), 삼...	10
22 은평구	구파발 (3), 연신내 (3), 불...	13
23 종로구	종각 (1), 종로3가 (1), 종...	15
24 중구	서울 (1), 시청 (1), 시청 (2...	23
25 중랑구	봉화산 (6), 신내 (6), 먹골...	8

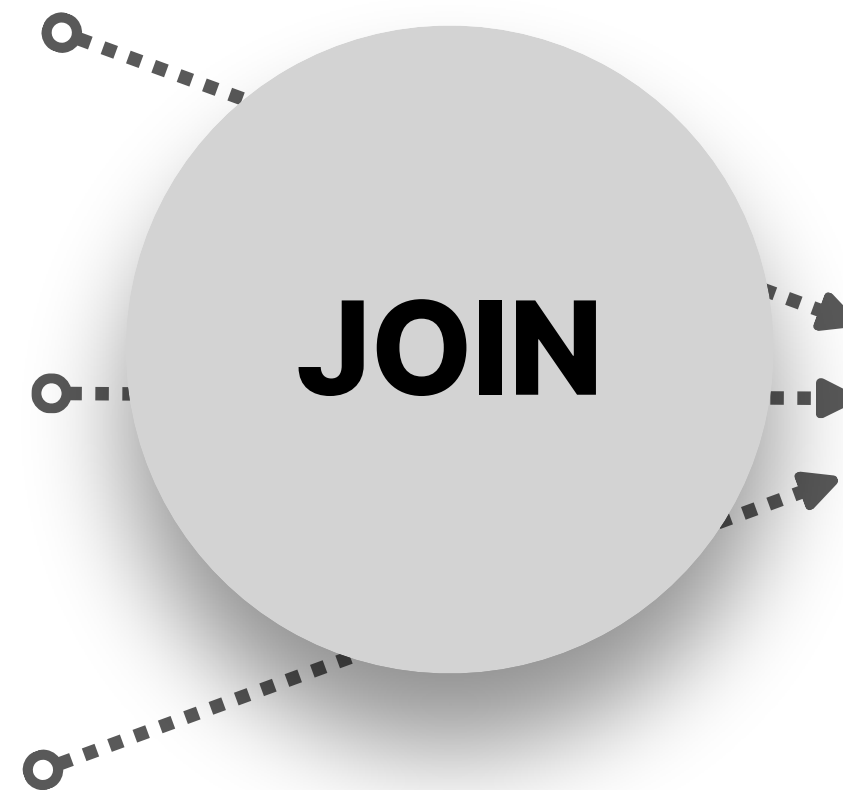
### 3. SQL (2) 회귀분석 데이터 구성

DATA

01 [ HOUSE\_2024 ]  
부동산 실거래가 정보

02 [ SHOP ]  
대규모점포 인허가 현황

03 [ SUBWAY ]  
자치구별 지하철역 정보



회귀분석 데이터

04 [ HOUSE\_TEST ]

### 3. SQL (2) 회귀분석 데이터 구성

#### DATA

01

[ HOUSE\_2024 ]

부동산 실거래가 정보

02

[ SHOP ]

대규모점포 인허가 현황

03

[ SUBWAY ]

자치구별 지하철역 정보

▶ JOIN \*'자치구명' 이 연결고리

```
create table house_test
```

```
as
```

```
select rownum as h_id,
```

```
h.물건금액,
```

```
h.건물면적,
```

```
h.건축년도,
```

```
h.층,
```

```
s.대규모점포수,
```

```
w.역개수
```

```
from house_2024 h, shop_cnt s, sw_cnt w
```

```
where h.자치구명 = s.자치구명 and h.자치구명 = w.자치구명
```

```
and h.최소일 is null
```

```
and h.건축년도 is not null
```

```
and h.계약일 > 20230000;
```



### 3. SQL (2) 회귀분석 데이터 구성

04

[ HOUSE\_TEST ]

	H_ID	물건금액	건물면적	건축년도	층	대규모점포수	역개수
1	1	65000	48.65	2001	2	5	12
2	2	136000	59.96	2020	8	4	16
3	3	139000	114.62	2000	17	2	14
4	4	21700	36.67	2016	2	3	5
5	5	30100	28.79	2017	2	5	12
6	6	43000	75.09	2003	4	4	16
7	7	150000	84.88	2015	3	2	14
8	8	112000	59.76	1997	5	2	14
9	9	85500	84.63	2010	11	3	5
10	10	72000	84.873	2003	7	3	3
11	11	238000	84.8	2008	10	6	28
12	12	13000	19.83	2013	7	7	8
13	13	100000	21.010	2011	2	1	11

#### ▶ COMMENT

물건금액: '건물의 실거래가에 대한 컬럼입니다. (만 원 단위)';

건물면적: '건물의 전용면적에 대한 컬럼입니다.';

대규모점포수: '건물이 속한 자치구 내 대규모점포 개수를 나타낸 컬럼입니다.';

역개수: '건물이 속한 자치구 내 1~9호선 지하철역 개수를 나타낸 컬럼입니다.';

## 4. 회귀분석 모델 실행

변수: 건물면적, 건축년도, 층, 대규모점포수, 역개수, 물건금액

### 01 train, test 테이블 생성

```
create table house_test_train
as select *
  from house_test
 where h_id <= 23965 ;

create table house_test_test
as select *
  from house_test
 where h_id > 23965 ;
```

- [ house\_test ] 총 26628건
- 9 : 1 비율로 train : test 테이블 생성

### 02 머신러닝 모델 환경구성 테이블 생성

```
CREATE TABLE SETTINGS_REG2
AS
SELECT *
  FROM TABLE (DBMS_DATA_MINING.GET_DEFAULT_SETTINGS)
 WHERE SETTING_NAME LIKE '%GLM%';

BEGIN

INSERT INTO SETTINGS_REG2
VALUES (DBMS_DATA_MINING.ALGO_NAME, 'ALGO_GENERALIZED_LINEAR_MODEL');

INSERT INTO SETTINGS_REG2
VALUES (DBMS_DATA_MINING.PREP_AUTO, 'ON');

COMMIT;

END;
/
```

## 4. 회귀분석 모델 실행

변수: 건물면적, 건축년도, 층, 대규모점포수, 역개수, 물건금액

### 03 머신러닝 모델 생성

```
BEGIN
  DBMS_DATA_MINING.CREATE_MODEL(
    MODEL_NAME          => 'MD_REG_MODEL3',
    MINING_FUNCTION      => DBMS_DATA_MINING.REGRESSION,
    DATA_TABLE_NAME     => 'house_test_train',
    CASE_ID_COLUMN_NAME  => 'h_id',
    TARGET_COLUMN_NAME   => '물건금액',
    SETTINGS_TABLE_NAME  => 'SETTINGS_REG2');
END;
/
```

### 04 생성된 머신러닝 모델 확인

```
SELECT MODEL_NAME,
       ALGORITHM,
       MINING_FUNCTION
FROM ALL_MINING_MODELS
WHERE MODEL_NAME = 'MD_REG_MODEL3';
```

# 4. 회귀분석 모델 실행

변수: 건물면적, 건축년도, 층, 대규모점포수, 역개수, 물건금액

05

## 모델의 회귀계수 확인

```
SELECT ATTRIBUTE_NAME, ATTRIBUTE_VALUE, ROUND(COEFFICIENT), P_VALUE
FROM TABLE (DBMS_DATA_MINING.GET_MODEL_DETAILS_GLM ('MD_REG_MODEL3'));
```

	ATTRIBUTE_NAME	ATTRIBUTE_VALUE	ROUND(COEFFICIENT)	P_VALUE
1	(null)	(null)	7084	0.17149604214213521
2	건물면적	(null)	838	0
3	건축년도	(null)	-21	0.0000000000000000095075260389590276
4	대규모점포수	(null)	1157	0.000000000020391217114157501
5	역개수	(null)	2856	0
6	층	(null)	2824	0

# 4. 회귀분석 모델 실행

변수: 건물면적, 건축년도, 층, 대규모점포수, 역개수, 물건금액

06

모델이 예측한 결과 확인

```
SELECT h_id, 건물면적, 건축년도, 층, 대규모점포수, 역개수, 물건금액 as 실제값,
       PREDICTION (MD_REG_MODEL3 USING *) 예측값
FROM house_test_test ;
```

	H_ID	건물면적	건축년도	층	대규모점포수	역개수	실제값	예측값
1	23966	15.55	2014	9	6	23	14400	74920.620415785001
2	23967	59.97	2016	6	5	12	260000	71072.398284173891
3	23968	43.5	2006	2	2	13	21500	45569.698906575926
4	23969	36	2018	5	7	8	28000	39005.638406674669
5	23970	41.77	2019	4	3	3	33300	22089.568617972651
6	23971	59.82	1997	22	3	11	107500	111363.70317924605
7	23972	114.68	2004	10	3	3	64500	100470.00724304686
8	23973	29.96	2023	1	3	10	59500	23621.949583547077
9	23974	21.68	2015	5	10	11	115000	21600.500015000000

## 4. 회귀분석 모델 실행

변수: 건물면적, 건축년도, 층, 대규모점포수, 역개수, 물건금액

07

### 모델의 성능평가 : 정답과 기계가 예측한 예측값과의 상관계수

```
select corr(실제값, 예측값)
from (
  SELECT h_id, 건물면적, 건축년도, 층, 대규모점포수, 역개수, 물건금액 as 실제값,
    PREDICTION (MD_REG_MODEL3 USING *) 예측값
  FROM house_test_test
) ;
```

	⚡ CORR(실제값,예측값)
1	0.7170668891666779956303798412191844336071

## 4. 회귀분석 모델 실행 - 파생변수 추가

변수: 건물면적, 건축년도, 층, 저층, 대규모점포수, 역개수, 아파트 여부, 구, 5구 내 아파트, 물건금액

+

### 파생변수 추가하여 [ house\_test ] 재생성

```
create table house_test
as
select rownum as h_id, h.물건금액, h.건물면적, h.건축년도, h.층, s.대규모점포수, w.역개수,
       case when h.층 <= 2 then 1 else 0 end as 저층,
       case when h.건물용도 = '아파트' then 1 else 0 end as 아파트여부,
       case when h.자치구명 in ('강남구', '서초구', '송파구', '용산구', '성동구') then 1 else 0 end as 구,
       case when h.건물용도 = '아파트' and h.자치구명 in ('강남구', '서초구', '송파구', '용산구', '성동구')
            then 1 else 0 end as 5구내아파트
from house_2024 h, shop_cnt s, subway w
where h.자치구명 = s.자치구명 and h.자치구명 = w.자치구명
      and h.취소일 is null and h.건축년도 is not null and h.계약일 > 20230000;
```

[illegible]



# 4. 회귀분석 모델 실행 - 파생변수 추가

변수: 건물면적, 건축년도, 층, 저층, 대규모점포수, 역개수, 아파트 여부, 구, 5구 내 아파트, 물건금액

06

모델이 예측한 결과 확인



	...	건물면적	건축년도	층	저층	대규모점포수	역개수	아파트여부	구	5구내아파트	실제값	예측값
1	23966	15.55	2014	9	0	6	23	1	0	0	14400	41601.592549231595
2	23967	59.97	2016	6	0	5	12	1	1	1	260000	164183.46660980833
3	23968	43.5	2006	2	1	2	13	0	0	0	21500	24018.07404178664
4	23969	36	2018	5	0	7	8	0	0	0	28000	25239.213859702424
5	23970	41.77	2019	4	0	3	3	0	0	0	33300	23225.978343736926
6	23971	59.82	1997	22	0	3	11	1	0	0	107500	90005.422297770376
7	23972	114.68	2004	10	0	3	3	1	0	0	64500	112526.75019562223
8	23973	29.96	2023	1	1	3	10	0	1	0	59500	37124.490188557866
9	23974	84.98	2015	6	0	10	11	1	0	0	117300	92466.439260602376
10	23975	39.53	1992	12	0	9	21	1	1	1	114000	163450.62497459087

# 4. 회귀분석 모델 실행 - 파생변수 추가

변수: 건물면적, 건축년도, 층, 저층, 대규모점포수, 역개수, 아파트 여부, 구, 5구 내 아파트, 물건금액

07

## 모델의 성능평가 : 정답과 기계가 예측한 예측값과의 상관계수

	 CORR(실제값,예측값) 
1	0.7896029633339748624130716535098208729407

## 5. 결론 및 보완점

	ATTRIBUTE_NAME	ROUND(COEFFICIENT)
1	(null)	24270
2	5구내아파트	71539
3	건물면적	752
4	건축년도	-21
5	구	24329
6	대규모점포수	1087
7	아파트여부	25312
8	역개수	100
9	저층	2267
10	총	1480

가장 영향력이 큰 변수

주요 구 내 아파트

결정적 변수

위치, 아파트 여부

가장 영향력이 작은 변수

건축년도

보완점

- 다양한 변수 : 금리, 학군, 지하철 외 교통 요인 등 추가 고려
- 분석 건수 : 2023년 이전 데이터 추가

---

~ 감사합니다 ~

---