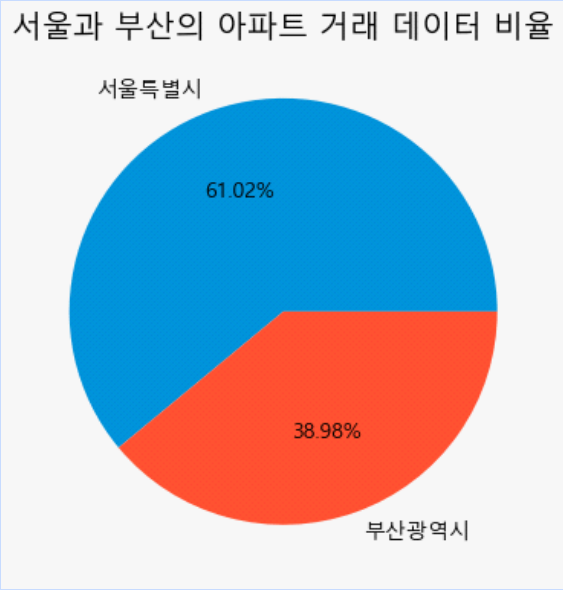
D&A Machine Learning Session

ML Competition 2023

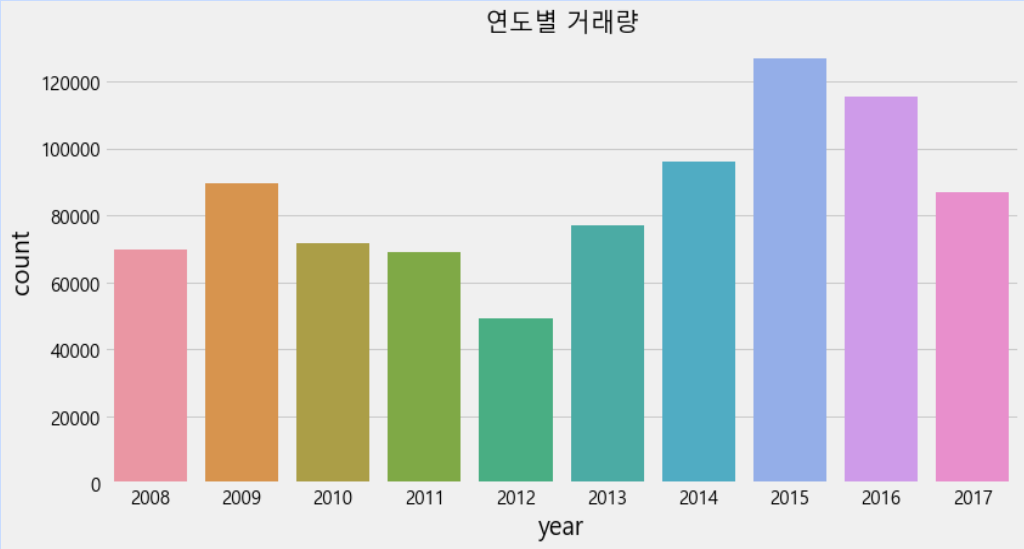
TEAM: 갈길이 머러 **MENTO:** 김지은
MEMBER: 김소현, 민수홍, 송은아, 신지후, 조현식

train 데이터셋 서울/부산 비율 시각화



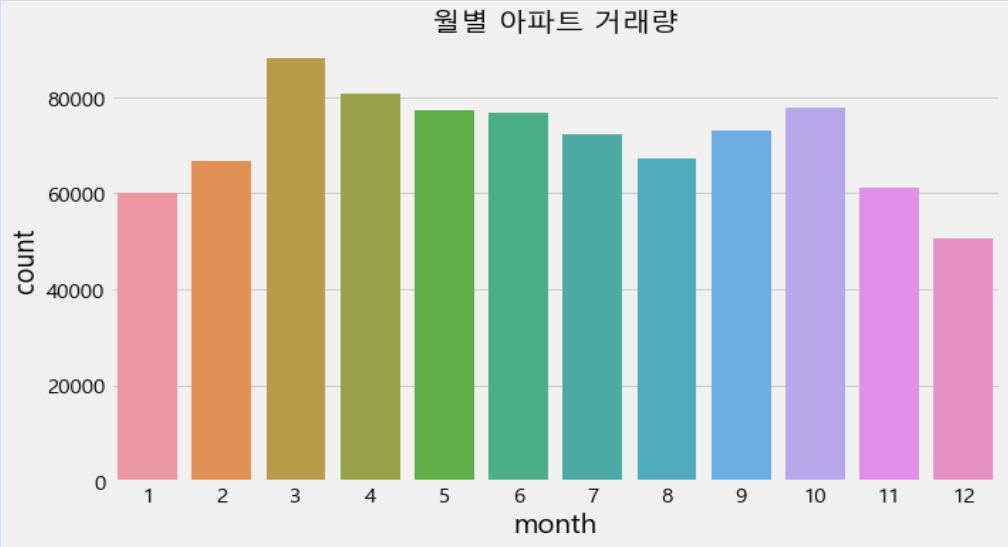
train 데이터셋에 대해
서울과 부산의 데이터 비율은 약 6:4로,
해당 데이터셋은 서울에 대해
좀 더 많은 데이터를 가지고 있음

연도별 거래량 시각화



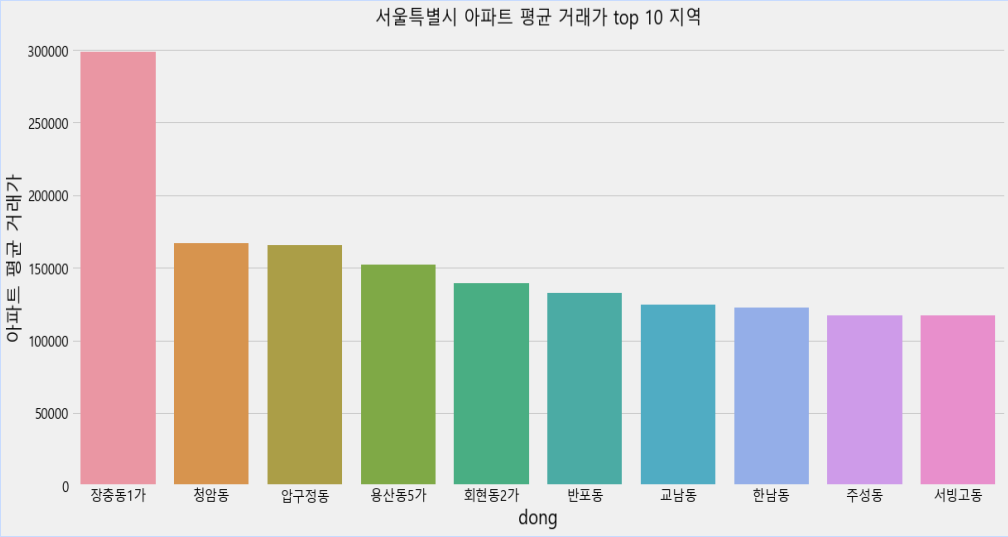
2014년 이후로
거래량이 증가하는
추세를 보여줌

월별 거래량 시각화



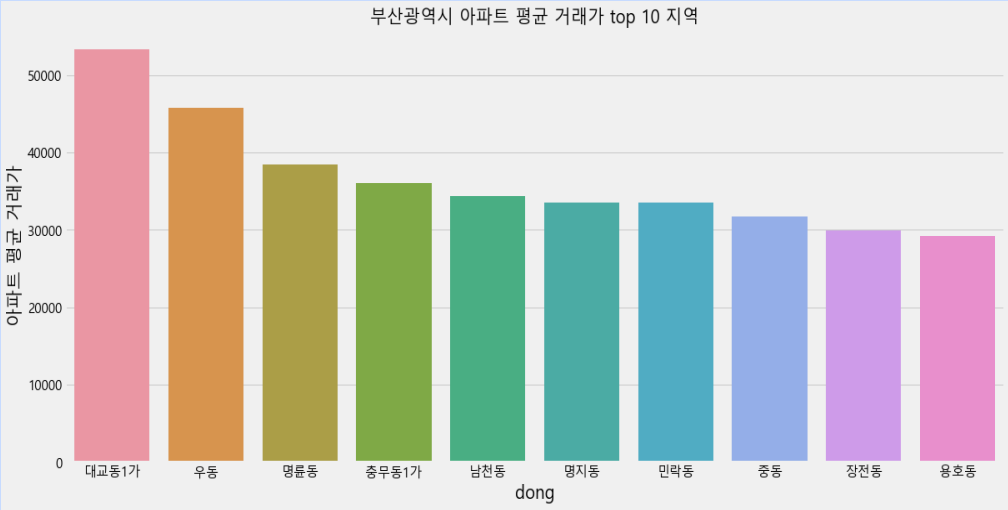
12,1,2월에 비해
3,4월의 거래량이 많음
→ 거래는 추운 겨울보단
따뜻한 봄에 활발함

서울 아파트 평균 거래가 top 10 시각화



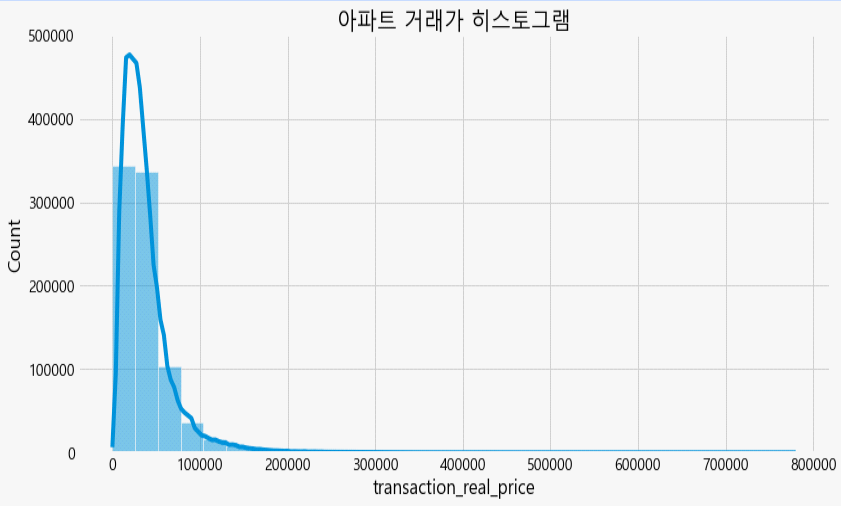
중구에 위치한
'장충동 1가 ' 가
가장 높게 나옴
→ 법정동에 존재하는
아파트가 2개로,
표본이 적어서
평균가가 높게 잡힌 것

부산 아파트 평균 거래가 top 10 시각화



영도구에 위치한
'대교동 1가 ' 가
가장 높게 나옴
→ 법정동에 존재하는
아파트가 1개로,
표본이 적어서
평균가가 높게 잡힌 것

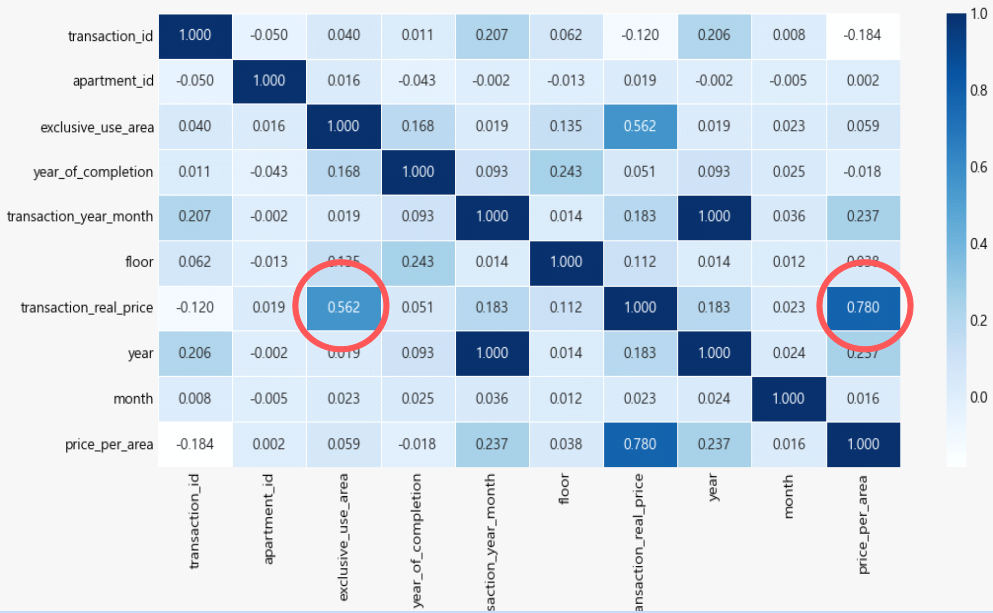
아파트 거래 가격 시각화(로그 전)



거래 가격이 한쪽으로
치우쳐져 있는 형태
→ 정규분포 형태로 만들어주기 위해
가격에 로그를 취해서 학습하고자 함

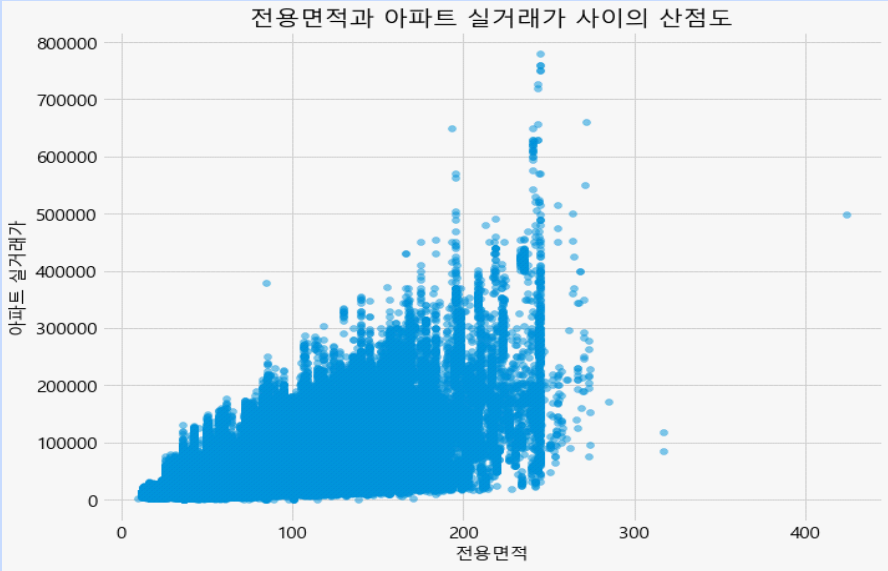
아파트 거래 가격 시각화(로그 후)

상관관계 시각화



전용면적과
아파트 거래가격 사이에
양의 상관관계가
존재함을 알 수 있음

전용면적과 아파트 실거래가 시각화



산점도를 통해 확인해보니
우상향하는 형태를 보임
→ 즉, 전용면적이
넓어지면 거래가격도
올라감

pyung_area

: 'exclusive_use_area'에 0.3025를 곱해 아파트 평수 계산,
이를 바탕으로 'pyung_area' 피처 생성

bucket_area

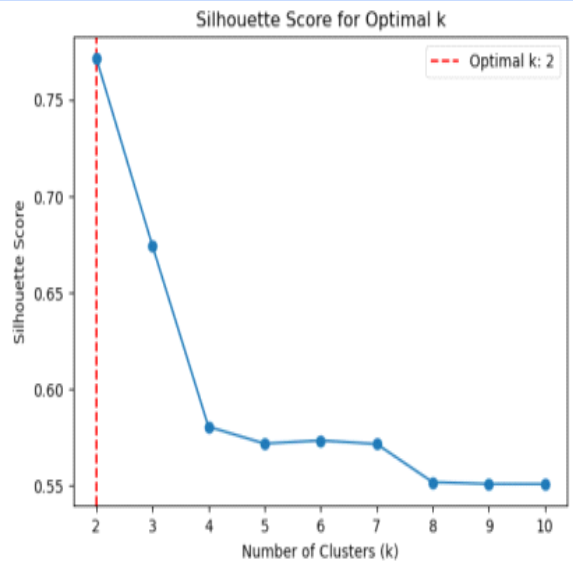
: 'make_area_bucket' 함수를 만들어 아파트 면적을
기준으로 그룹화를 진행해 '0,1,2,3'의 값으로 이루어진
'bucket_area' 피처 생성

new_area

: 'make_new_area' 함수를 만들어 10단위의
새로운 면적 feature인 'new_area'를 생성

가격과의 상관관계가 가장 높았던
전용면적 변수에 중점을 두고
연관된 파생 변수 생성

cluster : 'apartment_id'를 기준으로 군집화 진행



: 1) 아파트 아이디로 그룹화 한 후, 그룹별 거래가격 평균을 계산

2) 거래가격을 numpy 배열로 변환하고 1차원으로 변경

3) Silhouette 계수를 계산해 최적의 군집 개수를 찾음

4) 'Silhouette Score for Optimal k' 그래프 시각화
- 빨간색 점선: 최적의 k 지점을 의미
- 파란색 실선: Silhouette 계수를 의미

5) k를 2로 군집화를 진행한 후, 가격을 기준으로 군집화 한 결과를 데이터에 추가

	apartment_id	cluster
0	0	0
1	1	0
2	2	1
3	3	0
4	4	0
...	...	...
12384	12654	0
12385	12655	0
12386	12656	1
12387	12657	1
12388	12658	1

12389 rows × 2 columns

gu

: '행정구역분류' 외부 데이터를 활용하여 'gu' 피처 생성

	시도	시군구	법정동
0	서울특별시	서울특별시	서울특별시
1	서울특별시	종로구	종로구
2	서울특별시	종로구	청운동
3	서울특별시	종로구	신교동
4	서울특별시	종로구	궁정동

지하철역 개수

: '서울지하철', '부산지하철' 외부 데이터를 활용해 '지하철역 개수' 피처 생성

day

: 'transaction_date' 의 [-2:]로 'day' 피처 생성

age

: '2017 - year_of_completion'을 통해 아파트 나이를 계산한 'age' 피처 생성

금리

: '한국은행 기준금리 데이터' 인 외부 데이터를 활용해 transaction_data와 매핑하여 해당하는 날짜에 대해 금리를 가져와 interest_rate 피처 생성

	날짜	year	month	day	금리
0	2023-01-13	2023	1	13	3.50
1	2022-11-24	2022	11	24	3.25
2	2022-10-12	2022	10	12	3.00
3	2022-08-25	2022	8	25	2.50
4	2022-07-13	2022	7	13	2.25

어린이집 수

: day_care_center 데이터에서 어린이집 수를 계산해 'num_child_per_gu ' 를 만들고, 이를 각각 train과 test 데이터에 병합 이때 발생한 결측치는 0으로 처리

공원 수

: park 데이터에서 공원 수를 계산해 'num_park_per_dong ' 를 만들고, 이를 각각 train과 test 데이터에 병합 이때 발생한 결측치는 0으로 처리

floor

: 최솟값이 -4이므로 4를 더해서 음수를 없앰

top10

: top 10 시공사 아파트 여부를 나타내는 'top10' 피처 생성

transformed

: 빈도수에 따라 대표 아파트 25개를 리스트로 만들고 top10 시공사와 합쳐 ‘apt_name_list’를 만듦.
해당 리스트를 활용하여 ‘transformed’ 피처 생성

dong

: 가격기준으로 'dong'을 정렬한 리스트를 바탕으로 dong에 대한 Label Encoding 진행

apt, gu

: 'apt'와 'gu'에 대해 Binary Encoding 진행

transformed

: 'True'면 1, 아니면 0으로 Binary Encoding 진행

city

: '서울특별시'이면 1, 아니면 0으로 'city'에 대해 Binary Encoding 진행

transaction_year_month

: 'transaction_year_month'에 대해 연월이 증가하는 순으로 Label Encoding

RandomForestRegressor

고차원 데이터 대응 앙상블 효과

안정적인 예측 다양성과 중요도

1. 데이터 분할

```
X,y= train_X, train_y = train.drop('log_price', axis=1), train['log_price']

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

2. 모델 학습 후 예측

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error

rf = RandomForestRegressor(random_state=42)

rf.fit(X_train, y_train)

y_pred = rf.predict(X_test)

mse = mean_squared_error(y_test, y_pred)

rmse = np.sqrt(mse)

predict = rf.predict(test)
predict_final = np.expm1(predict)
```

RMSE: 0.08995829580857266

1. 파라미터 그리드 정의

```
rf_params_grid ={
    'n_estimators':[100,150,200,250],
    'max_depth':[10,20,25],
    'max_features': [0.33,1],
    'n_jobs':[-1]
    }
```

2. 그리드 탐색 수행

```
rf_grid_cv = GridSearchCV(rf, param_grid=rf_params_grid, cv=3,
                          scoring='neg_root_mean_squared_error')
rf_grid_cv.fit(X_train, y_train)
y_pred = rf_grid_cv.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
rmse = np.sqrt(mse)
print(rmse)
```

최적의 파라미터 값

```
best parameters : {'max_depth': 25, 'max_features': 0.33, 'n_estimators': 250, 'n_jobs': -1}
best estimator : RandomForestRegressor(max_depth=25, max_features=0.33, n_estimators=250,
                      n_jobs=-1, random_state=42)
```

RMSE: 0.08805205776664891

→ 튜닝 후 약 0.002 감소

Public

3	갈길이머러		4,202.70054	31	15h

Private

3	—	갈길이머러		4,173.75765	31	15h

D&A Machine Learning Session

ML Competition 2023

감사합니다 !

TEAM : 갈길이 머러 MENTO : 김지은
MEMBERS : 김소현, 민수홍, 신지후, 송은아, 조현식