아파트 실거래가 예측

D.N.A 성룡팀

목차

프로젝트 소개

Enjoy your stylish business and campus life with BIZCAM

아파트 실거래가 예측

금융 | 부동산 빅데이터와 AI를 이용하여 실거래가를 예측 분석 | 회귀 | RMSE

상금 : $8,000 + 80,000ZPR

2018.11.13 ~ 2019.01.31 23:59 + Google Calendar

890명 마감

Background

통계청 2015년 자료에 의하면 (https://bit.ly/2SFyzMA)
일반적인 한국인의 절반은 48.1%는 아파트에 살고 있습니다.
그들은 아파트 주거 선호도가 매우 높습니다.
또한 부의 증식 수단으로 생각 하기 때문에 아파트 가격에 관심이 많습니다.

이번 대회의 데이터 제공자는 직방입니다.
직방은 부동산 정보의 비대칭성과 불투명성을 해소하기 위해 노력하며,
중개사와 구매자를 연결하여 부동산정보 서비스 시장의 신뢰도를 높이는데 기여합니다.

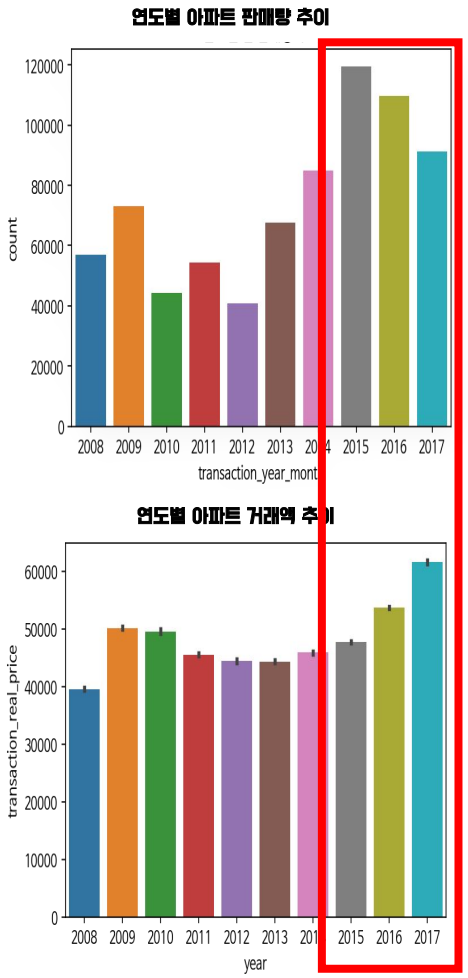
최근 매물 가격 정보는 직방, 다음부동산, 네이버부동산에서 볼 수 있습니다.
하지만 최근 매물 가격은 아직 거래되지 않아 정확하지 않은 정보 일 수 있습니다.

이에따라, 본 대회는 아파트 구매자들의 비대칭성 정보를 해결하기 위해 미래의 실 거래가 예측을 목표로 합니다.

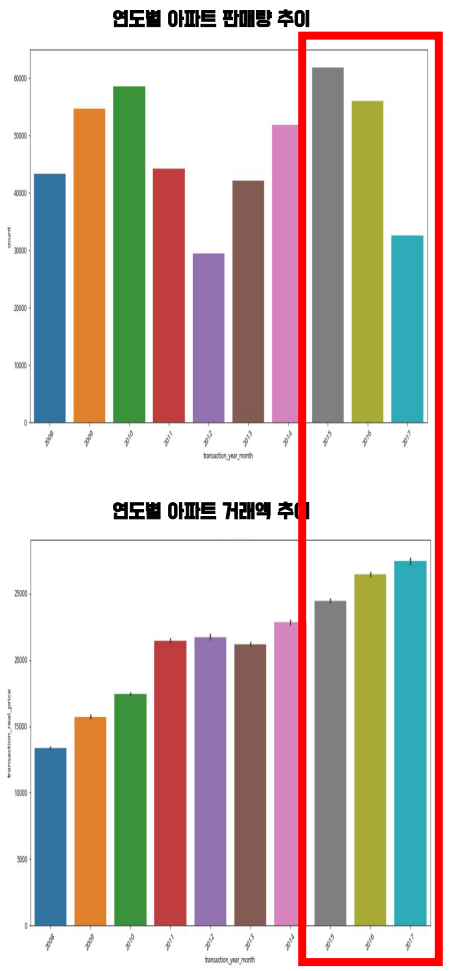
EDA

아파트 EDA

Enjoy your stylish business and campus life with BIZCAM



〈서울〉

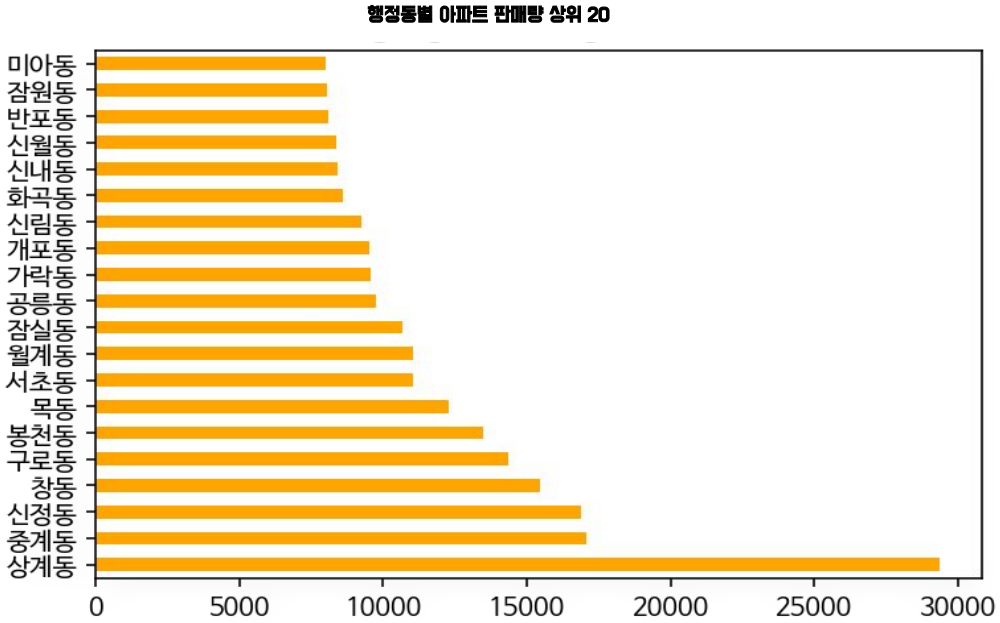


〈부산〉

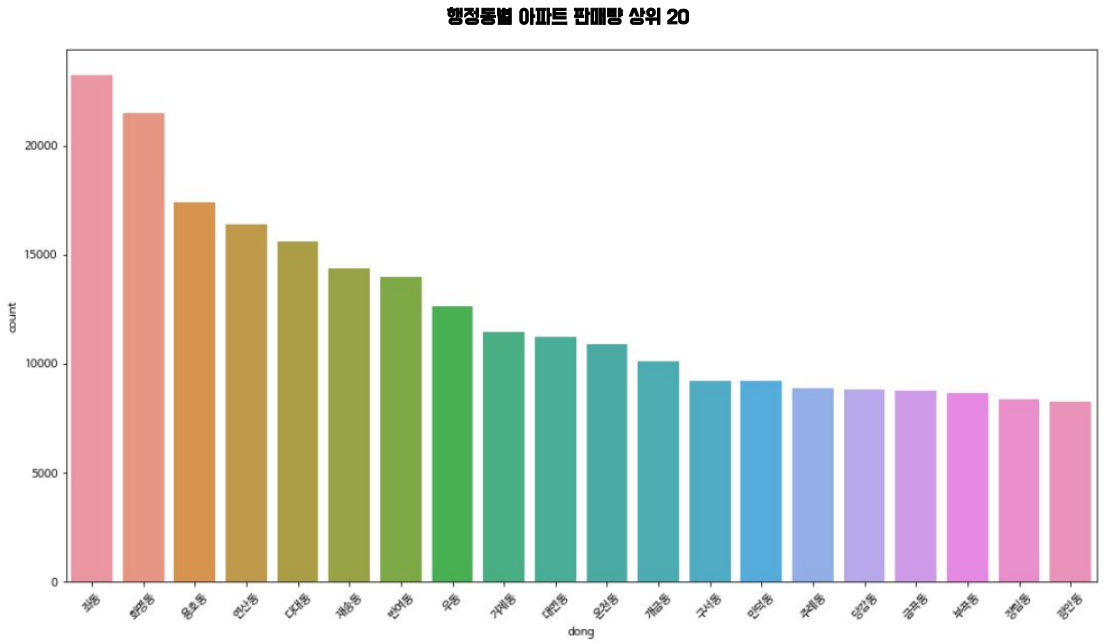
서울과 부산의 아파트 판매량은
2015년 이후 하락, 거래액은 상승

아파트 EDA

Enjoy your stylish business and campus life with BIZCAM



〈서울〉



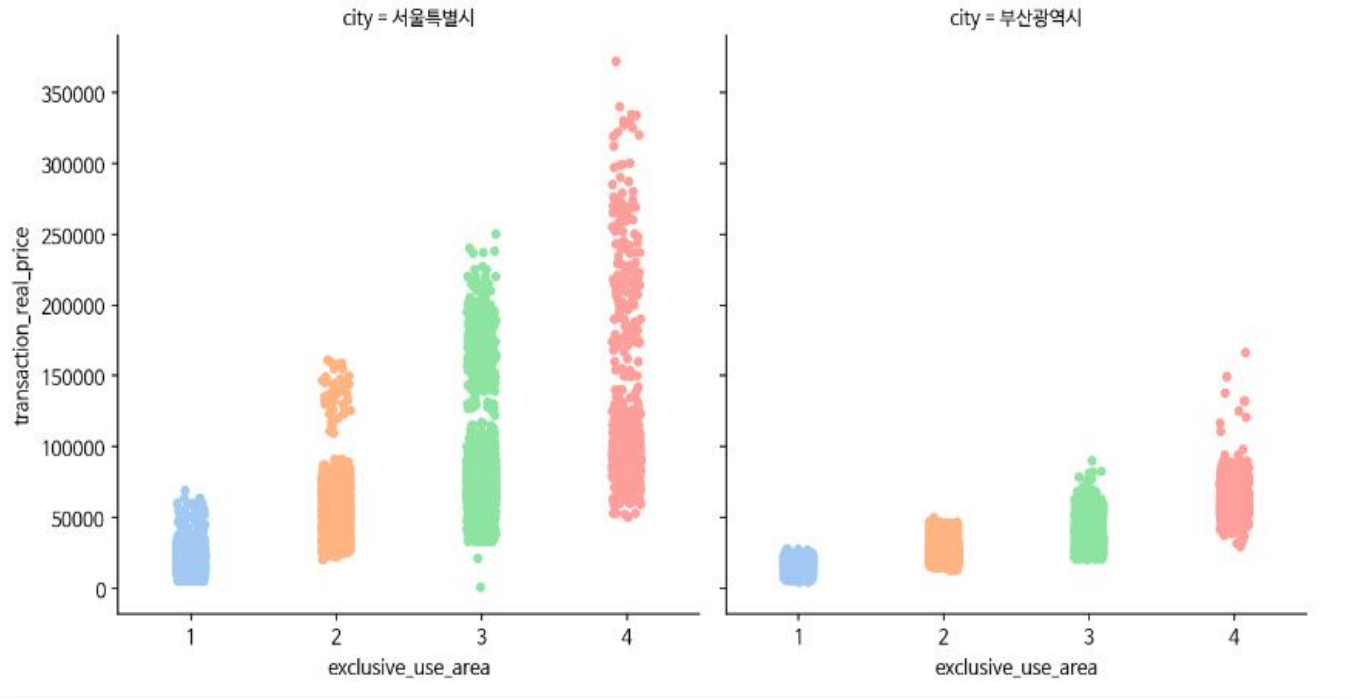
〈부산〉

행정동 아파트 판매량 순위

서울 : 상계동〉 중계동〉 신정동 ..

부산 : 좌동〉 화영동〉 용호동 ..

아파트 EDA



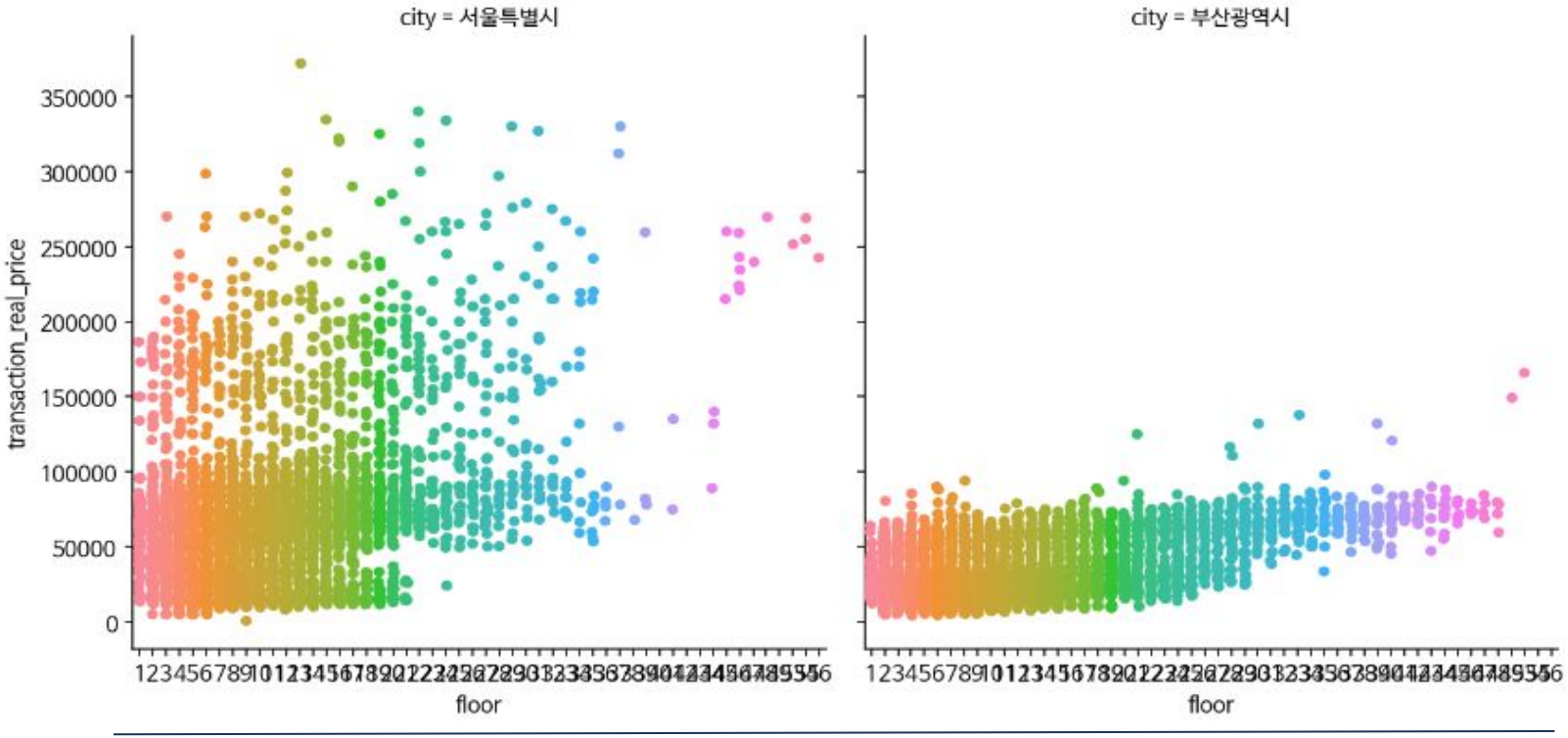
$50m^2 = 1$
$50m^2 \sim 80m^2 = 2$
$80m^2 \sim 100m^2 = 3$
$100m^2 \sim 120m^2 = 4$
$120m^2 \sim = 5$

평수별 가격 차이 분포

-> 전반적으로 평수와 가격이 비례하지만 집값 단위 자체는 서울특별시 월등히 높음

-> 120m^2이상의 아파트는 없는 것으로 판단

아파트 EDA



아파트 층수에 따른 집값 분포

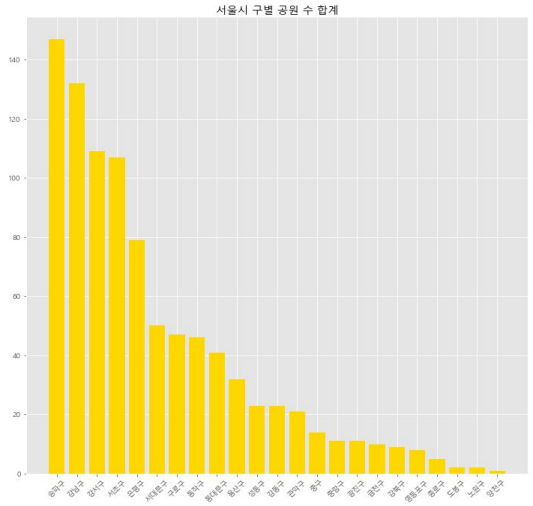
-> 서울의 경우, 고층 저층 관계 없이 다양한 가격

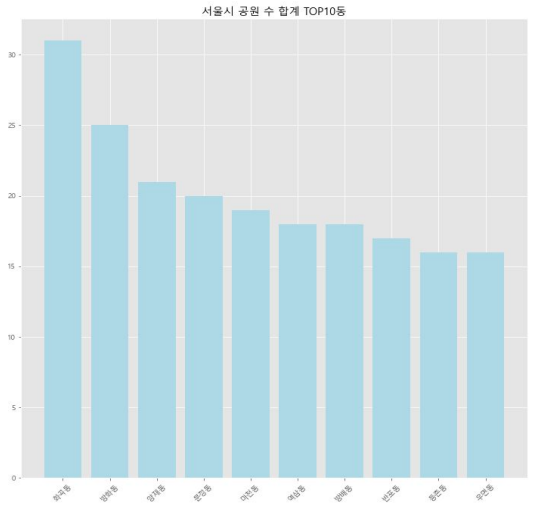
-> 부산의 경우, 대체적으로 고층이 저층보다 높은 가격

-> 층수가 아닌 다른 변수 (위치, 전용면적 등)가 아파트 가격 영향

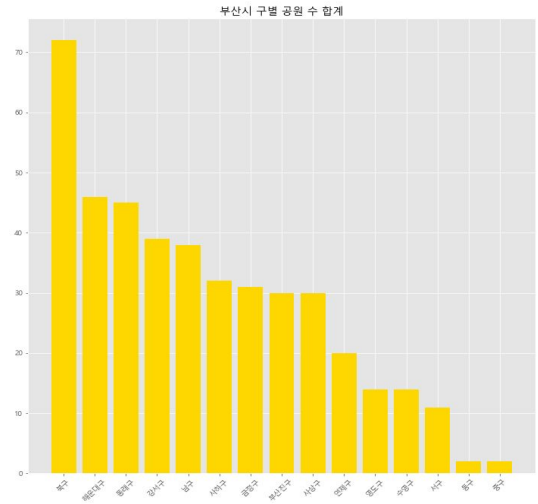
공원 EDA

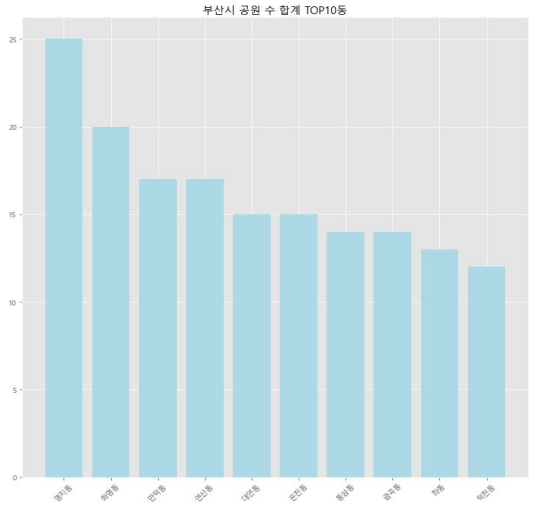
Enjoy your stylish business and campus life with BIZCAM





〈서울〉





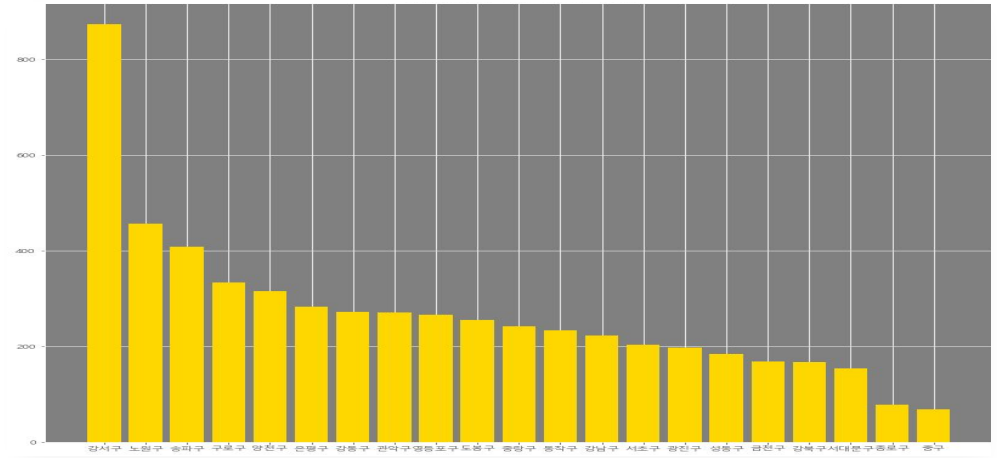
〈부산〉

서울&부산 구,동별 공원 수 합계

부산시는 구별 공원수가 동별 공원수와 비슷한 양상을 보이고 있으나, 서울시는 그렇지 않음.

유치원 EDA

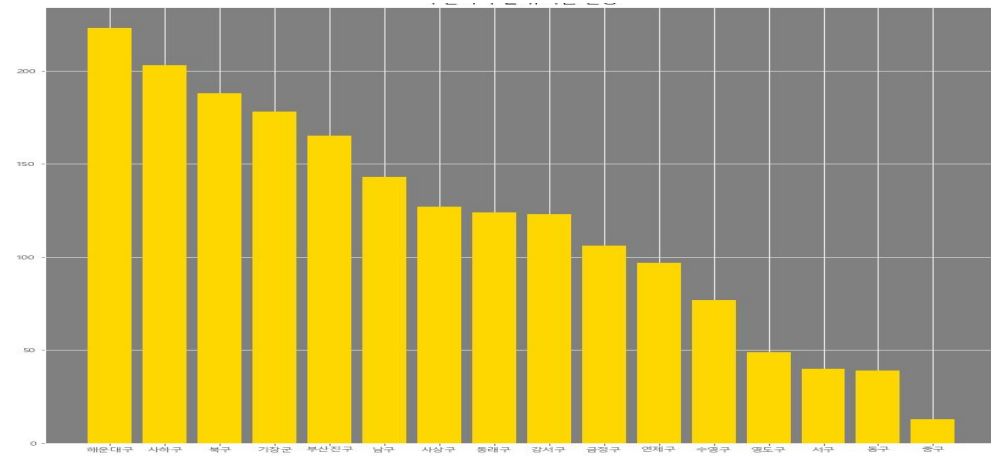
Enjoy your stylish business and campus life with BIZCAM



서울시 구별 유치원 현황

강서구에 특히 많은 유치원이 분포하고 있으며 그 뒤로

노원구, 송파구,구로구 등이 위치



부산시 구별 유치원 현황

해운대구, 사하구, 북구 순으로 유치원이 분포

전처리

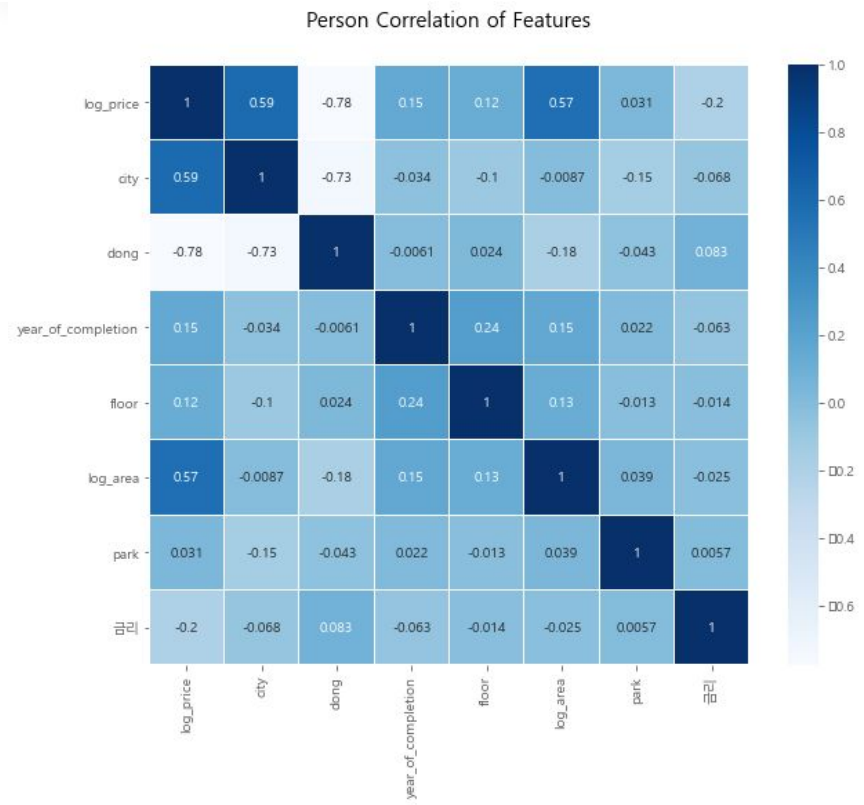
전처리

- 실거래가, 전용면적 로그 변환
- 층수 결측값 처리
- 인덱스 초기화
- 겹치는 동 병합
- 라벨링(건축년도, 도시, 공원, 동)

학원, 금리, 공원, 원자재 가격 등의
파생변수 추가고려

전처리


```
cols=["log_price","city","dong","year_of_completion","floor","log_area","park","금리"]
heatmap_data = train[cols]
colormap = plt.cm.PuBu
plt.figure(figsize=(10, 8))
plt.title("Person Correlation of Features", y = 1.05, size = 15)
sns.heatmap(heatmap_data.astype(float).corr(), linewidths = 0.1, vmax = 1.0, square = True, cmap = colormap, linecolor = "white", annot = True, annot_kws =
{"size" : 16})
```



히트맵을 통해서 상관관계를 분석했을 때 0.9 이상인 상관관계 X

다중공선성의 문제 없음

전처리

Enjoy your stylish business and campus life with BIZCAM

한국은행이 26일 기준금리를 연 0.5%에서 0.75%로 인상하면서 금리 인상이 부동산 시장에 미칠 영향이 관심을 끌고 있다. 부동산 전문가들은 "단기적으로는 집값 상승이 빠르게 꺾이지는 않을 것"이라고 입을 모았다. 다만 추가 금리 인상이 이어지면 중장기적으로 집값이 안정될 가능성이 높다고 분석했다.

심교언 건국대 부동산학과 교수는 "이론적으로 보면 금리 인상은 집값 하락을 불러오지만 지금까지 금리를 올렸다고 집값이 내려간 적은 없는 것 같다"며 "금리 인상 폭이 작은 데다 전세가격 등도 오르고 있어 주택시장의 전반적인 수급 상황 등에 더 큰 영향을 받을 것"이라고 말했다.

학군명성과 교육환경 · 결과 변수가 아파트 가격에 미치는 영향 분석*

Effects of School District Reputation, Educational Input and Outcome Variables on Apartment Price

이 광 현 (Kwang-Hyun Lee)**

< Abstract >

This article analyzes the effect of school district reputation, educational input and outcome variables on apartment prices in Busan, where inter/intra-district choices are implemented.

외부변수 선정 원인

- **금리의 영향을 많이 받는 부동산 시장** — **금리 변수 & 원자재 가격 변수**
- **교육환경의 영향을 받는 부동산 시장** — **학원 변수**

전처리

transaction_year_month	insaction_da	floor	action_real_	log_price	log_area	park	금리	원자재가격
201711	21~30	4	78500	11.2709	4.43687	0	1.25773	nan
201711	11~20	2	88500	11.3908	4.54871	0	1.25773	nan
201711	21~30	3	145000	11.8845	5.03318	0	1.25773	nan
201711	21~30	10	106250	11.5736	4.98989	0	1.25773	nan
201711	21~30	6	113000	11.6351	5.0133	0	1.25773	nan
201711	21~30	3	104500	11.5569	4.97342	0	1.25773	nan
201711	21~30	8	59500	10.9937	4.58813	1	1.25773	nan
201711	1~10	7	52900	10.8762	4.74953	0	1.25773	nan
201711	1~10	7	43000	10.669	4.3804	0	1.25773	nan
201711	11~20	11	36200	10.4968	4.16914	0	1.25773	nan
201711	11~20	12	54000	10.8967	4.66927	0	1.25773	nan
201711	11~20	10	52500	10.8686	4.66927	0	1.25773	nan
201711	21~30	6	35800	10.4857	4.16914	0	1.25773	nan
201711	21~30	1	19500	9.87817	4.08682	0	1.25773	nan
201711	1~10	14	13800	9.53242	2.71403	0	1.25773	nan
201711	11~20	8	49800	10.8158	4.09301	0	1.25773	nan

- 월별 거래 변수를 기준으로 금리와 원자재가격 데이터 병합
- 학원변수의 경우 기존 행정구역 변수를 결합하여 병합해야하는데 데이터가 없으므로 제외

모델링

모델링

	transaction_id	apartment_id	city	dong	jibun	apt	exclusive_use_area	year_of_completion	transaction_year_month	transaction_date	floor	금리	원자재가격	log_price
0	0	7622	서울특별시	신교동	6-13	신현(101동)	84.82	15	200801	21~31	2	4.987619	48.7	10.532096
1	1	5399	서울특별시	필운동	142	사직파크맨션	99.17	44	200801	1~10	6	4.987619	48.7	9.903488
2	2	3578	서울특별시	필운동	174-1	두레엘리시안	84.74	10	200801	1~10	6	4.987619	48.7	10.558414
3	3	10957	서울특별시	내수동	95	파크팰리스	146.39	14	200801	11~20	15	4.987619	48.7	11.678440
4	4	10639	서울특별시	내수동	110-15	킹스매너	194.43	13	200801	21~31	3	4.987619	48.7	11.695247

최종 데이터셋:
transaction_id, apartment_id, city, dong, jibun, apt, exclusive_use_area,year_of_completion, transaction_year_month, transaction_date, floor, 금리,원자재가격 ,log_price(타겟변수_로그변환)

모델링


```
test=test.fillna(1.25773)
```

```
lgb_param = {
        'n_estimators':[300,500,700],
        'min_child_weight': [3,5,7],
        'max_depth': [5,9,15]
        }

grid_lgb = GridSearchCV(model_lgb, param_grid=lgb_param, cv=3)
grid_lgb.fit(X_train, y_train)
print(grid_lgb.best_params_)
15, 5, 700
```

- train data: NaN값 x
- test data: 금리변수 NaN값 존재-> 마지막 달 금리 값 대체 (결측값 근접한 날짜)

- Grid search 를 이용하여 하이퍼 파라미터 조정

모델링

Enjoy your stylish business and campus life with BIZCAM

[모델 성능 평가 지표]
RMSE 채택

사용 모델:
Ridge(alpha값 조정) => RMSE값: 21322

Lasso(alpha값 조정)=> RMSE값: 21978

lightgbm + grid search=> RMSE값: 6588

DecisionTreeRegressor => RMSE값: 19682

RandomForest => RMSE값: 27291

GradientBoosting => RMSE값: 26143

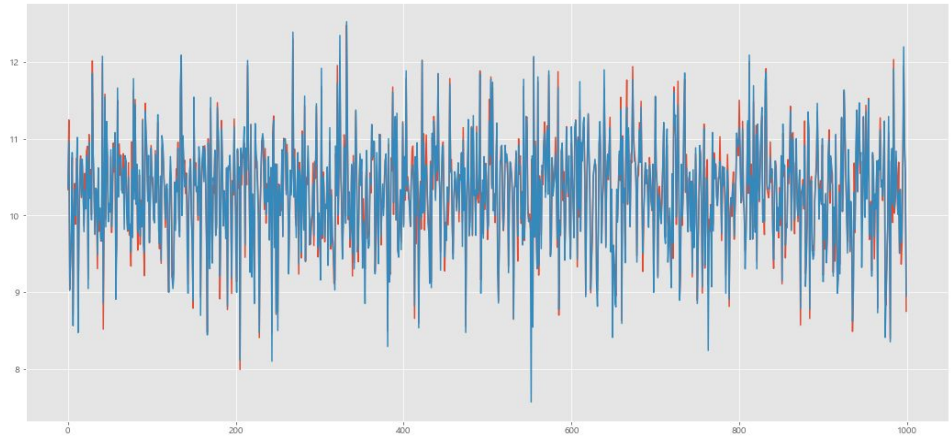
XGBRegressor => RMSE값: 23913

모델링

사용 모델:
Ridge
Lasso

lightgbm
DecisionTreeRegressor
RandomForest
GradientBoosting
XGBRegressor

grid search 조정&lightgbm 채택



grid search조정후 실제값& 예측값

모델링

제출결과: public 4위 (LGBM) & grid search 조정 & 순위 encoding

1	유재성 KADE		4,795.7994	10	2달 전
2	통계청김웅곤+양경성+민성욱EDA김현우	양경 Ex	5,787.87666	139	일 년 전
3	겸댕이	겸댕	6,177.80363	1	3달 전
4	서드임	서드	6,895.6105	1	5시간 전

제출결과: public 29위 (DecisionTreeRegressor) & label encoding

29	ahj	ah	20907.17812	2	한 시간 전

차후 계획 및 소감

차후 계획

〈개선방안〉

연도	시도	행정구역	종류	분야	학원수
2013	서울	종로구	학교교과교습학원	입시검정및보습	78
2013	서울	종로구	학교교과교습학원	국제화	8
2013	서울	종로구	학교교과교습학원	예능	16
2013	서울	종로구	학교교과교습학원	특수교육	1
2013	서울	종로구	학교교과교습학원	기타	5
2013	서울	종로구	학교교과교습학원	소계	108
2013	서울	종로구	평생직업교육학원	국제화	49
2013	서울	종로구	평생직업교육학원	직업기술	72
2013	서울	종로구	평생직업교육학원	인문사회	23
2013	서울	종로구	평생직업교육학원	기예	28
2013	서울	종로구	평생직업교육학원	소계	172

- **다양한 외부변수 활용 :**

  행정구역 데이터 x=〉 지역별 교육 변수, 공원 어린이집 변수 포함x

- **선형 회귀 모델 채택**

- **어린이집, 공원 data 파생변수 도출:**

  행정구역 데이터 x=〉 어린이집& 공원 데이터 활용 불가능

개인 소감

Enjoy your stylish business and campus life with BIZCAM

아현

인코딩과 최적 파라미터 조정이 얼마나 중요한지를 이번 활동을 통해서 느끼게 되었다. 같은 외부변수 추가와 전처리 작업이 같음에도 모델부분과 grid search을 이용한 조정부분에서 나온 차이가 접수 차이를 줬다. 다음 부분에서는 모델링 부분에 신경을 더 써서 성능을 보완해야겠다

형배

다음에 더 열심히 모델링을 해봐야 될 것 같다 아좌좡 O_(

용원

처음 팀장을 맡아보는거라 능숙하지 못한점이 많았는데 팀원들이 끝까지 마무리 잘해서 너무 기분이 좋습니다. 다음에는 모델링 공부를 더 열심히해서 더 좋은 결과 내고 싶습니다.

진혁

처음으로 팀프로젝트에 직접 참여하게 되었는데 많은 공부의 필요성을 느끼게 되었다. EDA와 전처리를 먼저 계속 공부하여 프로젝트에서 내가 할 수 있는 일을 점차적으로 늘려야겠다.

감사합니다