

아파트 가격 예측

Aiffel DS Campus 3y1y
2024-02-26

목 차

1. 서막: "도시의 심장을 듣다"
2. 배경 이야기: "왜 아파트 가격인가?"
3. 데이터의 미로: "탐험가들의 도구상자"
4. 마법의 주문: "예측의 마법사들"
5. 첫 번째 시험: "실험실의 비밀"
6. 다음 목적지: "미지의 영역으로"
7. 결론: "예측의 마법, 현실을 바꾸다"
8. 호기심의 시간: "궁금증을 풀어드립니다"

1. 서막: "도시의 심장을 듣다"

- **목표 소개:** "우리는 도시의 심장박동, 바로 아파트 가격의 비밀을 파헤치려 합니다. DATATHon 여정에 오신 것을 환영합니다!"
- **팀 소개:** "여기, 데이터를 무기로 삼은 용감한 탐험가들이 모였습니다. 각자의 역할은 고유하지만, 우리의 목표는 하나입니다. 바로 최고의 예측 모델을 만드는 것이죠."

팀장



🍒 윤진영

INFP

응

IF NOT NOW, WHEN?

데이터 분석가

분석



📋 양동영

ENFP

응

물음표 살인마

행복하게 살자

프로덕트 매니저

기술



📋 이영우

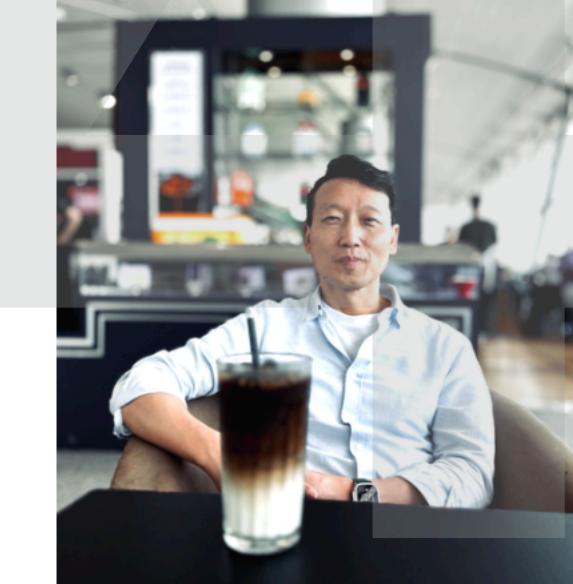
ISFP

힌튼

해보자

데이터 분석가

정리



🏃 정광용

INFJ-A (옹호자) [https://www.16|](https://www.16personalities.com/infp)

응

늦었지만 배움에는 끝이 없다고 생각하

오늘이 남은날 중 가장 젊은 날이다. 흰

프로덕트 매니저

업계 최고 전문가 그룹 구성

준비



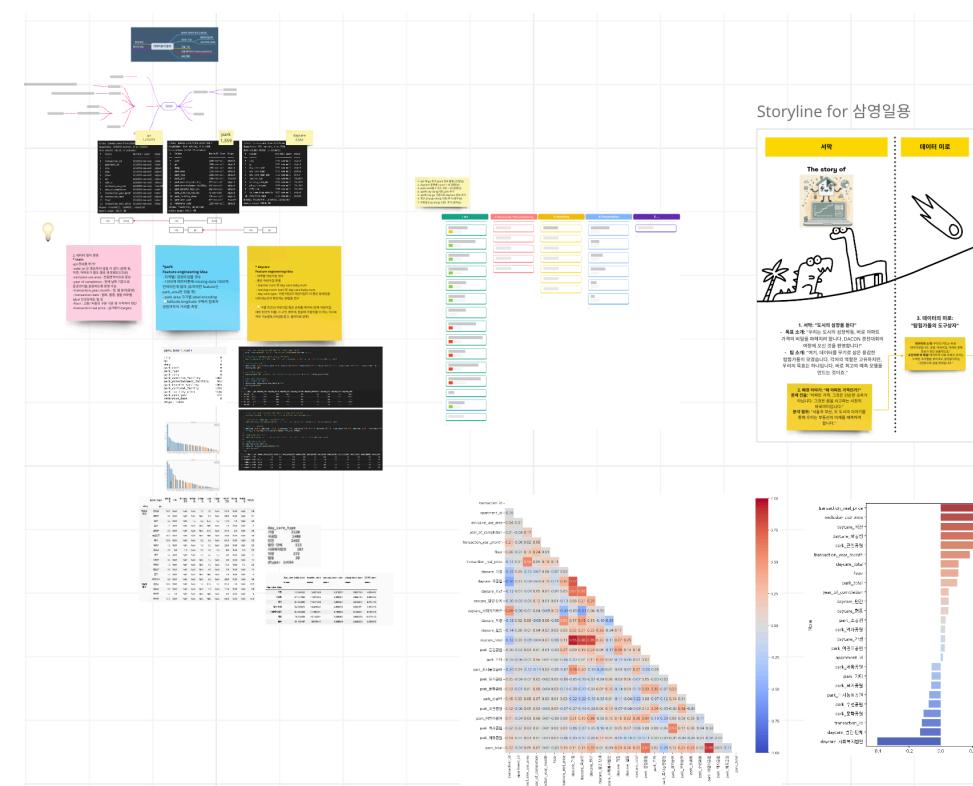
삼영일용

3 팀
윤진영
양동영, 이영우, 정광용
Empty
Empty
Empty
February 22, 2024
+ Add a property
Add a comment...

팀 프로젝트 계획

A Miro board screenshot showing a project plan for a data analysis pipeline. The plan includes steps like '데이터 수집', '데이터 가공', '데이터 전처리', '피처 엔지니어링', '모델 구현', '모델 해석하기 (=interpretability)', and '모델 평가'. A 'Storyline for 삼영일용' section is also present. The URL is https://miro.com/app/board/uXjVNpjHWDo=/?share_link_id=995620072422.

브레인스토밍



협업



Orderlee / aiffel_datathon

Issues Pull requests Actions Projects Security

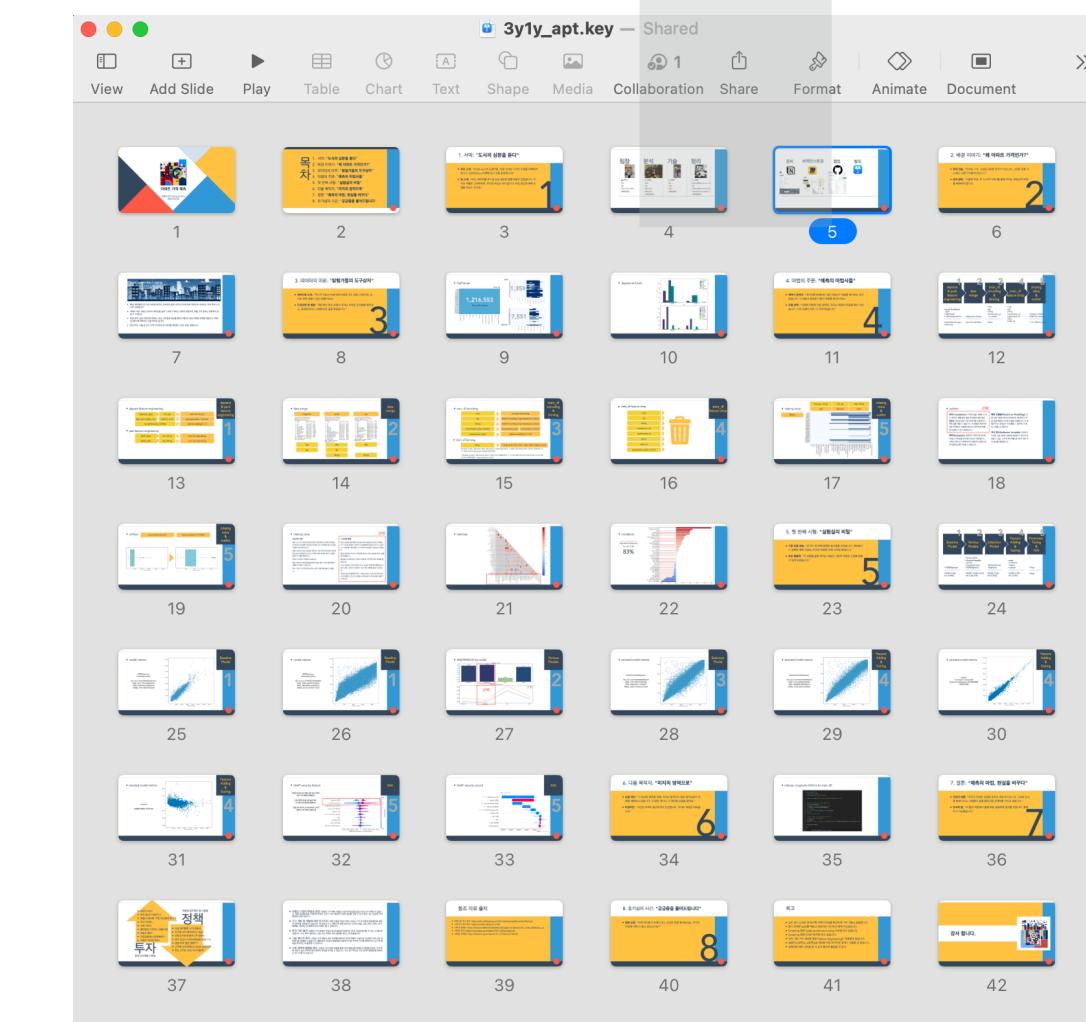
aiffel_datathon Public Watch 1

main Go to file + < Code

Orderlee datathon f0c56ae · 1 hour ago 43 Commits

- data datathon 1 hour ago
- 양동영 test 3 days ago
- 윤진영 commit 3 days ago
- 이영우 datathon 8 hours ago
- 정광용 subway_data 3 days ago
- .DS_Store updated notebook 3 days ago
- .gitignore datathon 1 hour ago
- 3y1y_apt.ipynb datathon 1 hour ago
- 3y1y_apt_legacy.ipynb datathon 1 hour ago
- 3y1y_apt_test.ipynb datathon 1 hour ago

발표

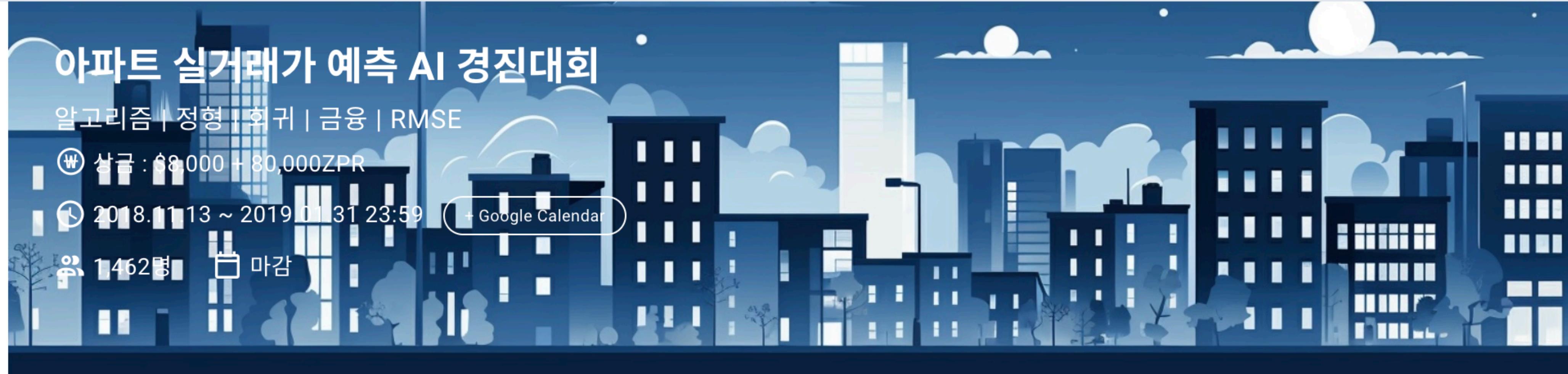


공동작업을 위한 첨단 환경 적용

2. 배경 이야기: "왜 아파트 가격인가?"

- **문제 진술:** "아파트 가격, 그것은 단순한 숫자가 아닙니다. 그것은 꿈을 사고파는 시장의 바로미터입니다."
- **분석 범위:** "서울과 부산, 두 도시의 이야기를 통해 우리는 부동산의 미래를 예측하려 합니다.





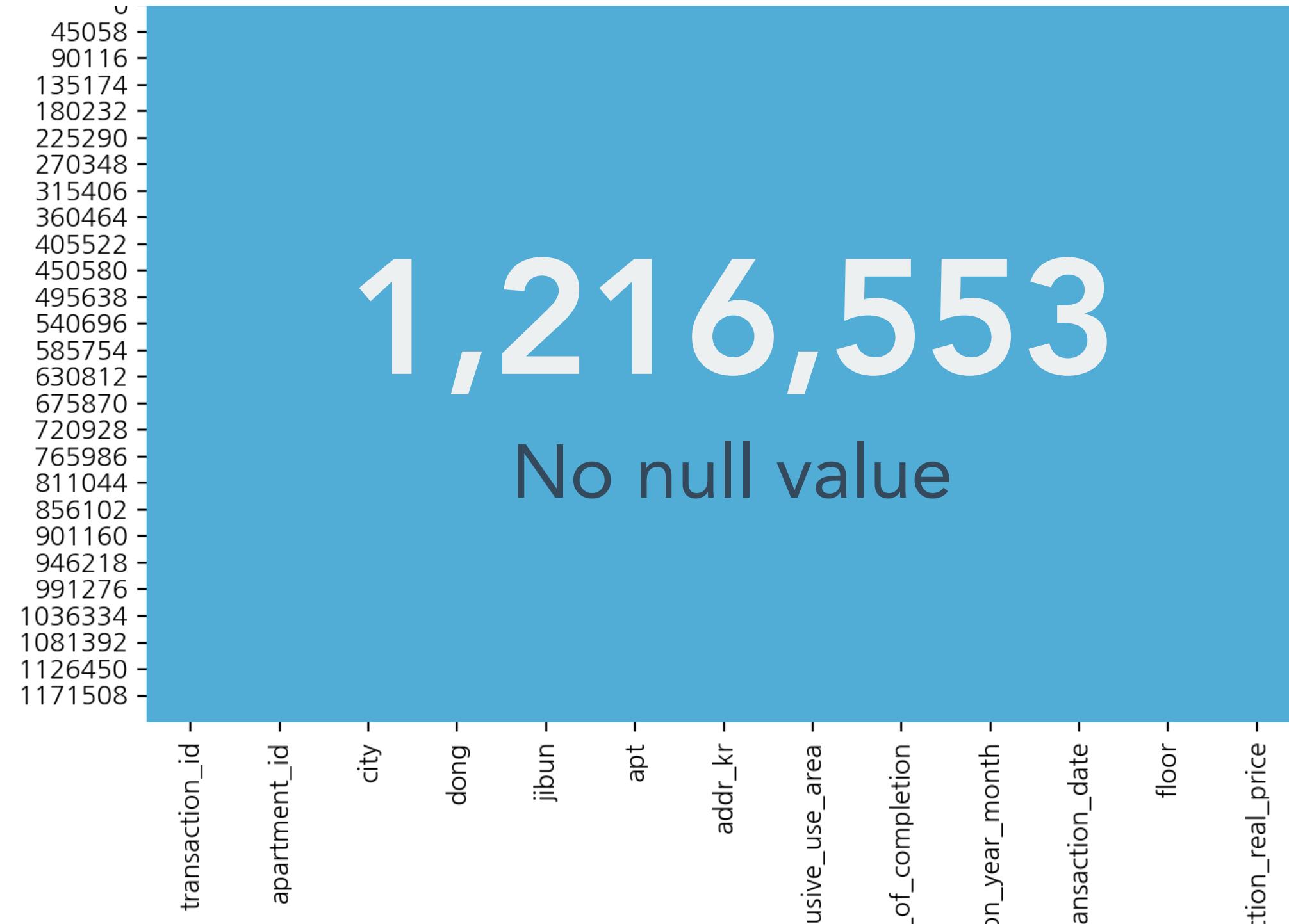
- 배경: 통계청의 2015년 자료에 따르면, 한국인의 절반 가까이가 아파트에 거주하며, 아파트는 부의 축적 수단으로 선호됩니다.
- 데이터 제공: 부동산 정보의 투명성을 높이기 위해 노력하는 직방이 제공하며, 매물 가격 정보는 정확하지 않을 수 있습니다.
- 대회 목적: 실제 거래가와 아파트, 학교, 지하철역 정보를 통해 구매자의 정보 비대칭 문제를 해결하고, 미래 실거래가를 예측하는 것을 목표로 합니다.
- 분석 목적: 서울 및 부산 지역 아파트의 실거래가를 예측할 수 있는 모델 개발입니다.

3. 데이터의 미로: "탐험가들의 도구상자"

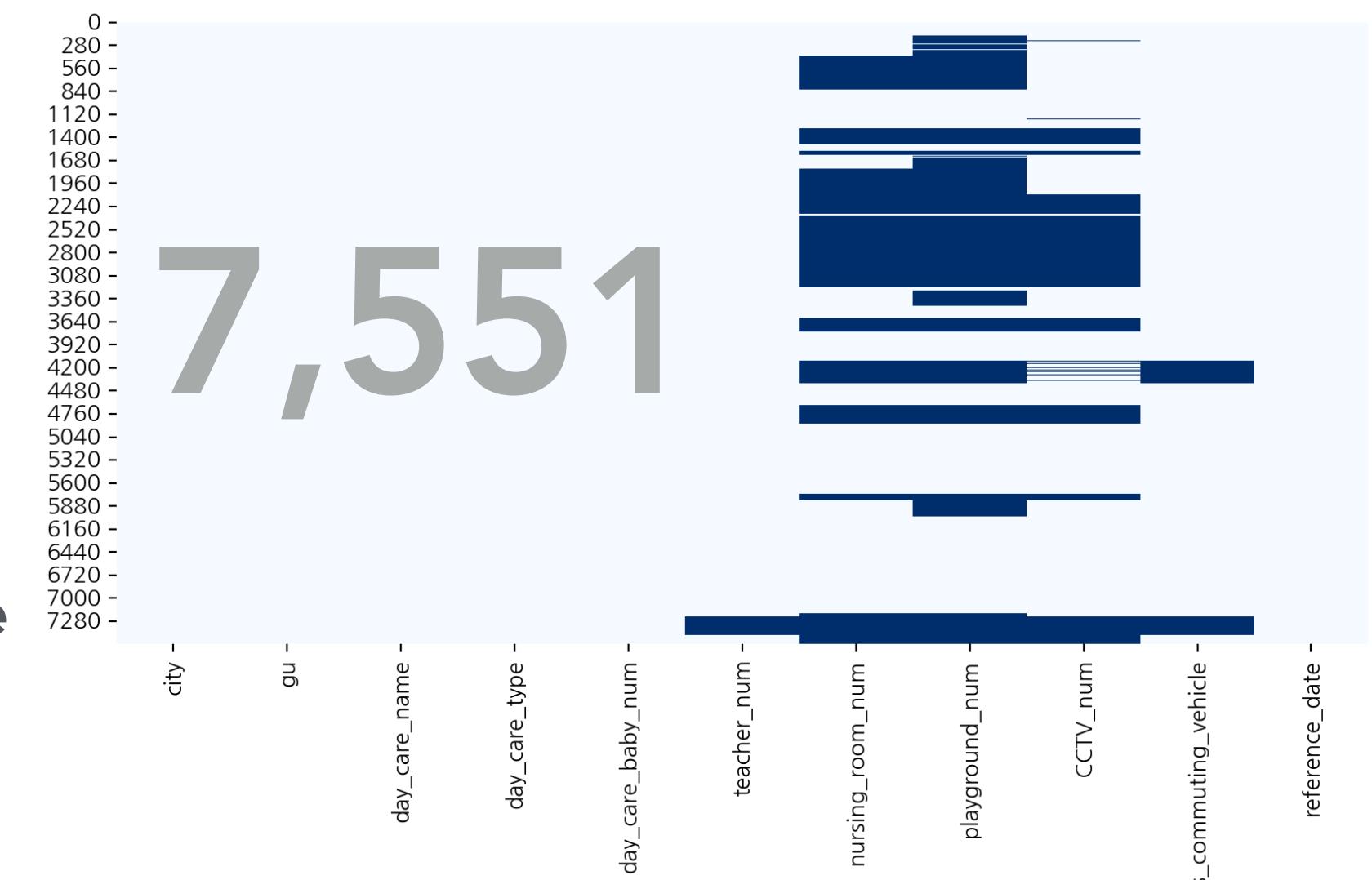
- **데이터셋 소개:** "우리의 지도는 바로 데이터셋입니다. 공원, 어린이집, 아파트 판매 정보가 담긴 보물지도죠."
- **도전과제 및 해결:** "데이터의 미로 속에서 우리는 누락된 조각들을 찾아내고, 중앙값이라는 나침반으로 길을 찾았습니다."

3

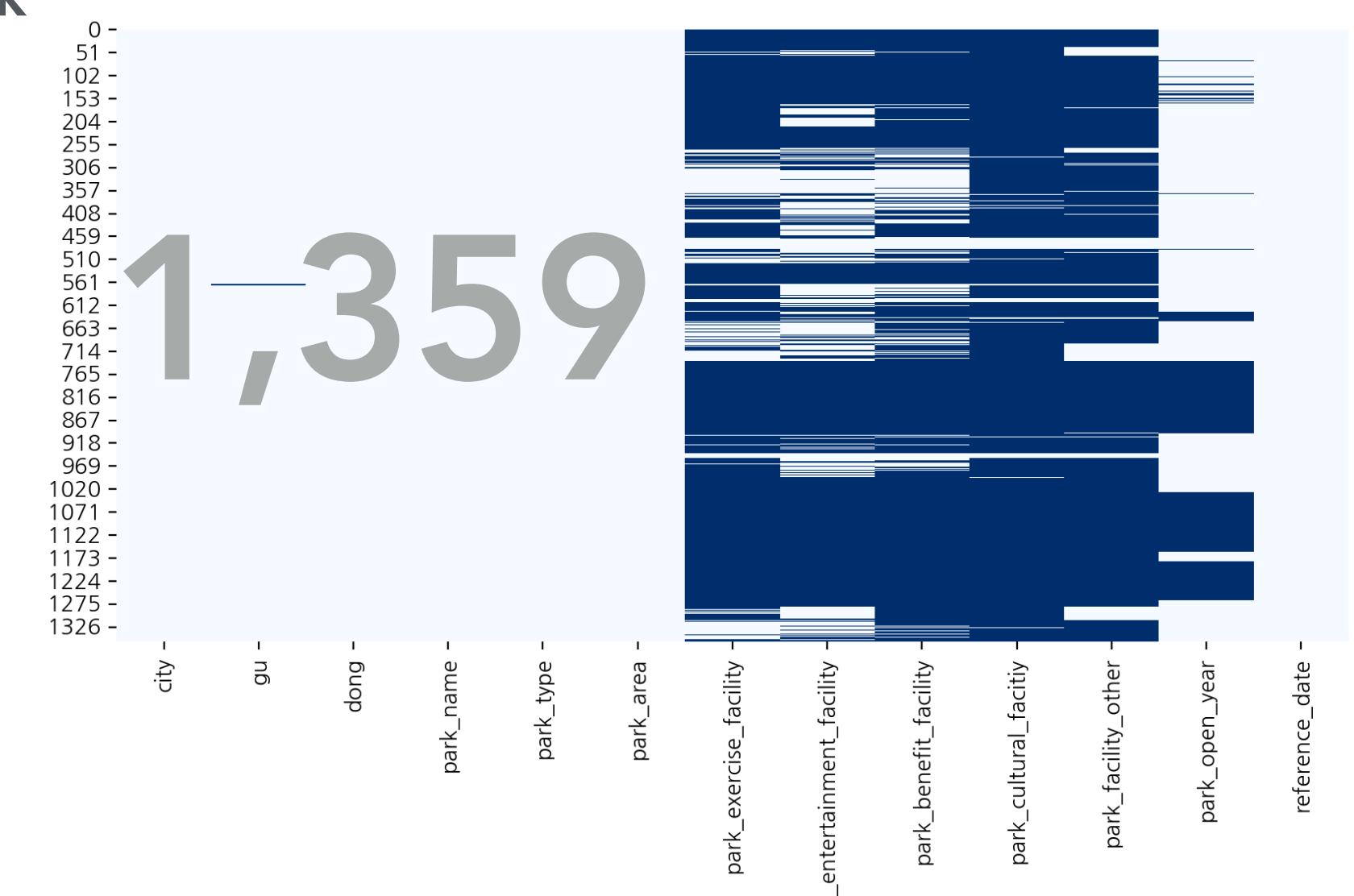
- Null Values



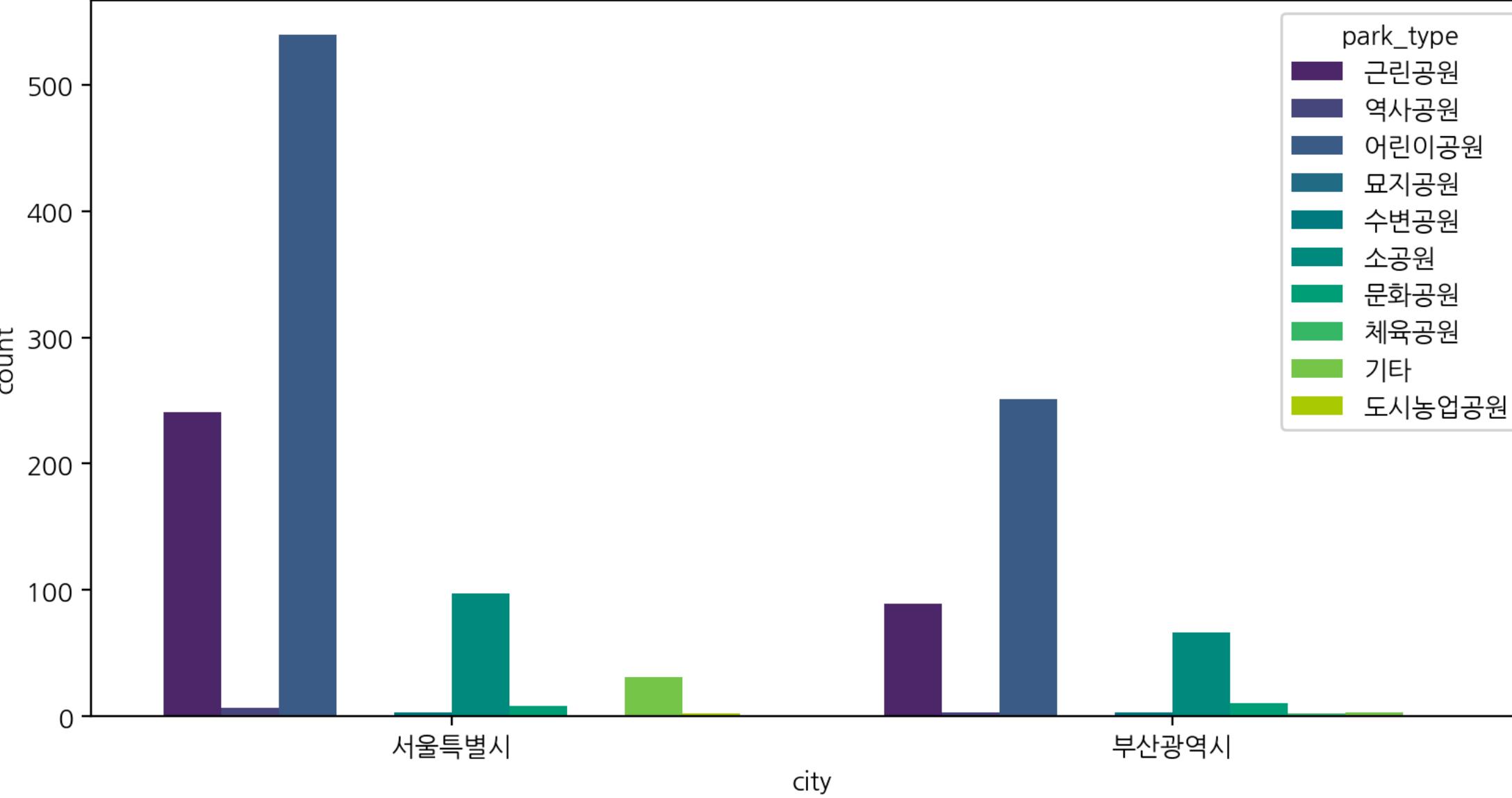
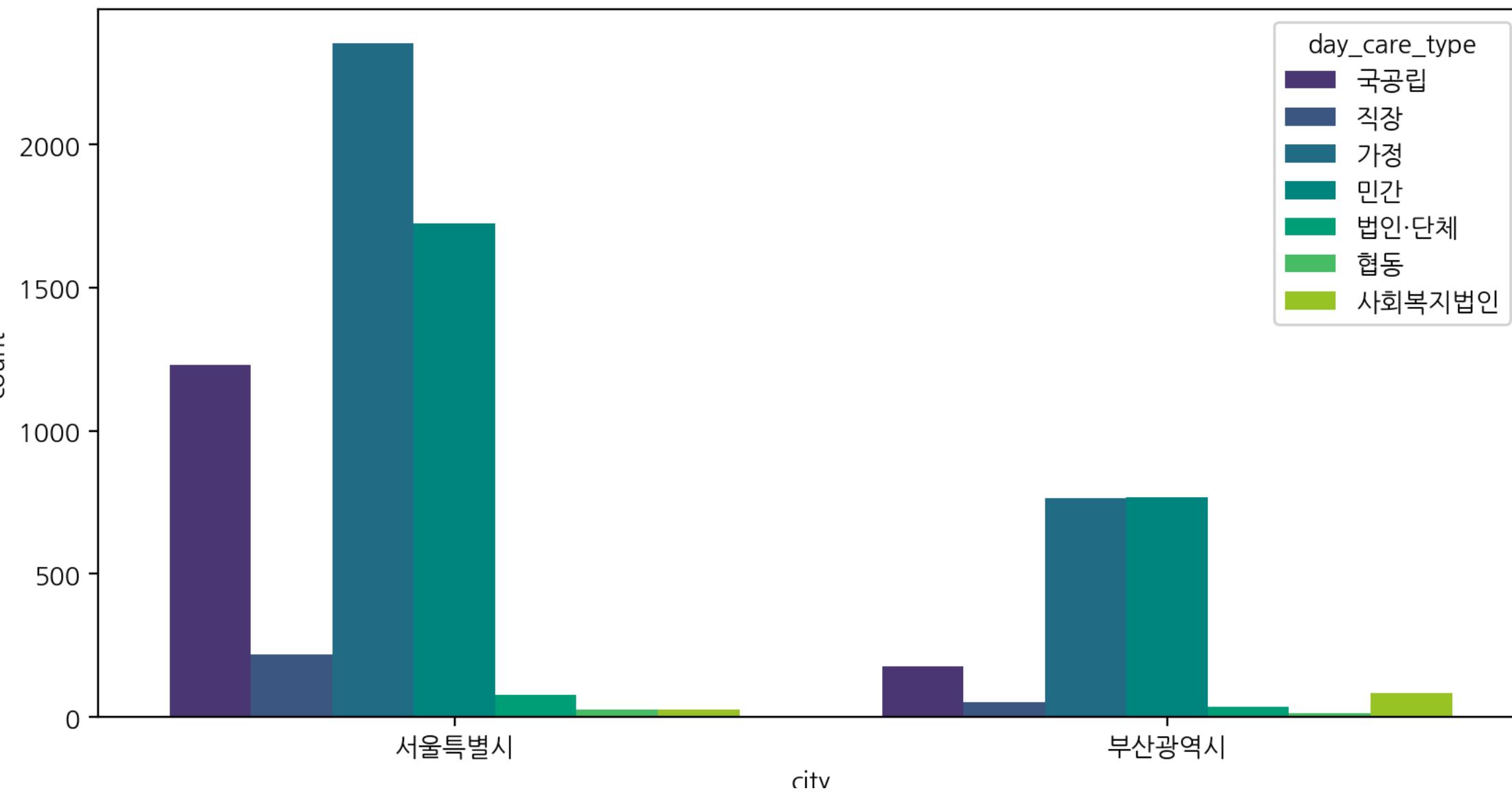
daycare



park



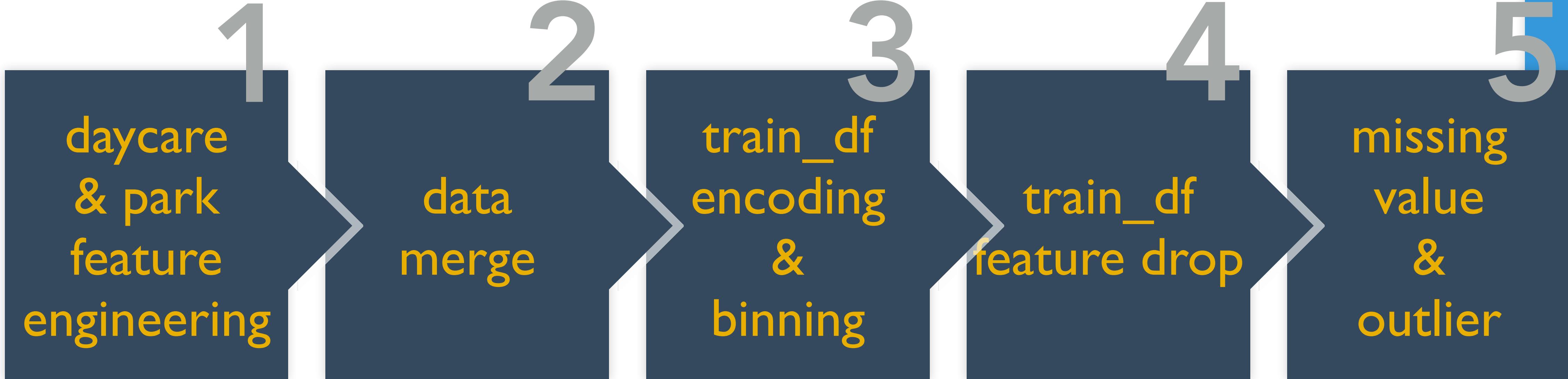
● daycare and park



4. 마법의 주문: "예측의 마법사들"

- **데이터 전처리:** "데이터를 정제하는 것은 마법사가 재료를 준비하는 것과 같습니다. 각 재료가 중요한 마법의 주문을 완성시키죠."
- **모델 선택:** "다양한 마법의 주문 중에서, 우리는 최강의 주문을 찾아 나섰습니다. 기준 모델이 바로 그 시작이었습니다."

4



- group-by daycare type
- baby/Teach
- commuting vehicle

-
- group-by part type
 - part area

-
- daycare by city-gu

-
- part by city-dong

- city
- gu
- dong
- transaction_ye
- ar_month

- floor

- city
- gu
- dong
- transaction_id
- apartment_id
- jibun
- addr_kr

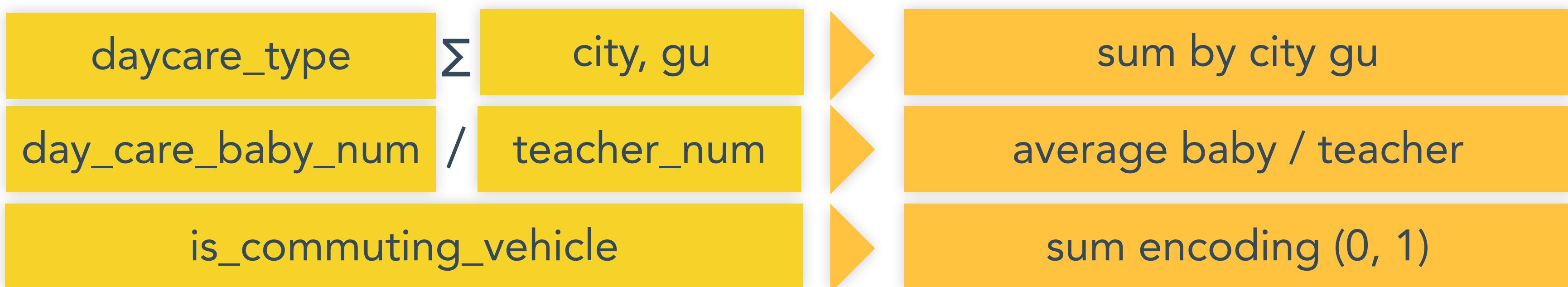
-
- infinity to NaN
 - NaN to median

-
- 1.5 x IQR in price

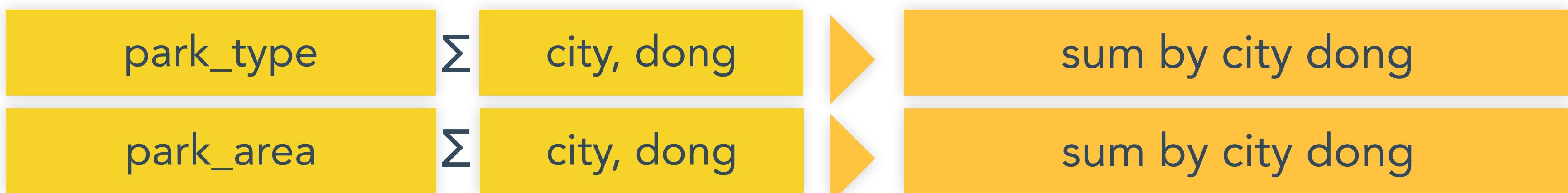
daycare & park feature engineering

1

- daycare feature engineering



- park feature engineering



data merge

- data merge

daycare

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7551 entries, 0 to 7550
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   city              7551 non-null    object  
 1   gu                7551 non-null    object  
 2   day_care_name     7551 non-null    object  
 3   day_care_type     7551 non-null    object  
 4   day_care_baby_num 7551 non-null    int64  
 5   teacher_num       7326 non-null    float64 
 6   nursing_room_num 4352 non-null    float64 
 7   playground_num    3626 non-null    float64 
 8   CCTV_num          5280 non-null    float64 
 9   is_commuting_vehicle 7055 non-null  object  
 10  reference_date    7551 non-null    object  
dtypes: float64(4), int64(1), object(6)
memory usage: 649.0+ KB
```

city

gu

park

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1359 entries, 0 to 1358
Data columns (total 13 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   city              1359 non-null    object  
 1   gu                1356 non-null    object  
 2   dong               1359 non-null    object  
 3   park_name         1359 non-null    object  
 4   park_type         1359 non-null    object  
 5   park_area          1359 non-null    float64 
 6   park_exercise_facility 277 non-null  object  
 7   park_entertainment_facility 435 non-null  object  
 8   park_benefit_facility 266 non-null    object  
 9   park_cultural_facitiy 72 non-null    object  
 10  park_facility_other 175 non-null    object  
 11  park_open_year    937 non-null    float64 
 12  reference_date    1359 non-null    object  
dtypes: float64(2), object(11)
memory usage: 138.1+ KB
```

city

gu

dong

apt

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1216553 entries, 0 to 1216552
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   city              1216553 non-null  object  
 1   dong               1216553 non-null  object  
 2   apt                1216553 non-null  object  
 3   exclusive_use_area 1216553 non-null  float64 
 4   year_of_completion 1216553 non-null  int64  
 5   transaction_year_month 1216553 non-null  int64  
 6   transaction_date    1216553 non-null  int64  
 7   floor               1216553 non-null  int64  
 8   transaction_real_price 1216553 non-null  int64  
 9   transaction_year     1216553 non-null  int64  
 10  transaction_month   1216553 non-null  int64  
 11  floor_category      1216553 non-null  int64  
dtypes: float64(1), int64(8), object(3)
memory usage: 111.4+ MB
```

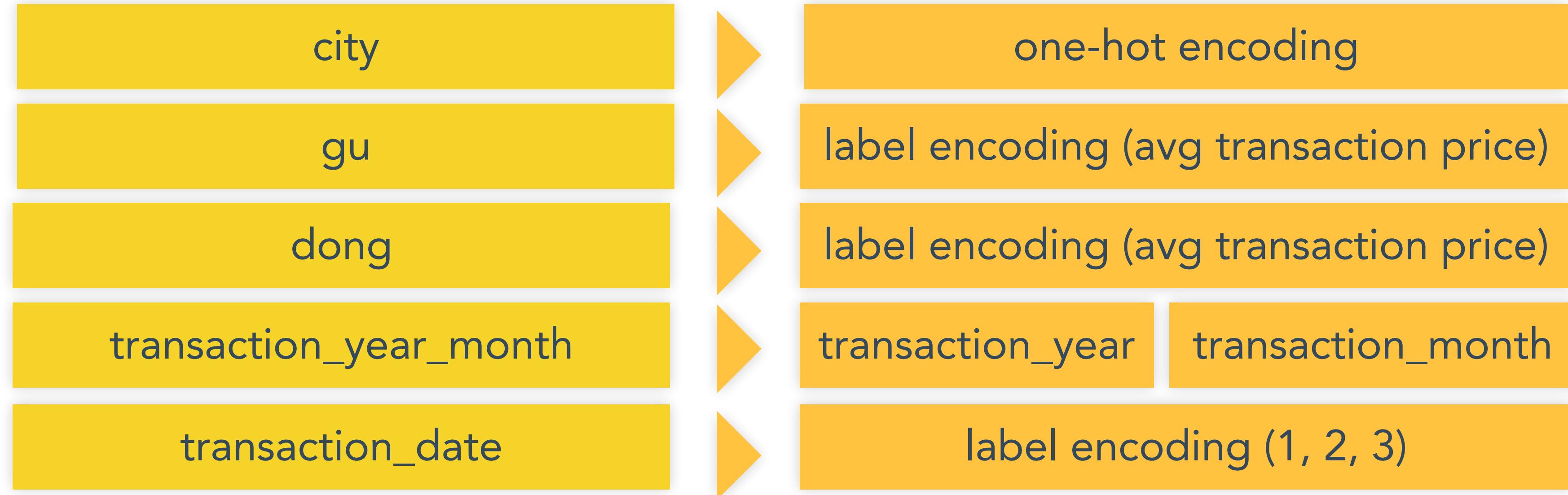
city

dong

train_df encoding & binning

3

● train_df encoding



● train_df binning

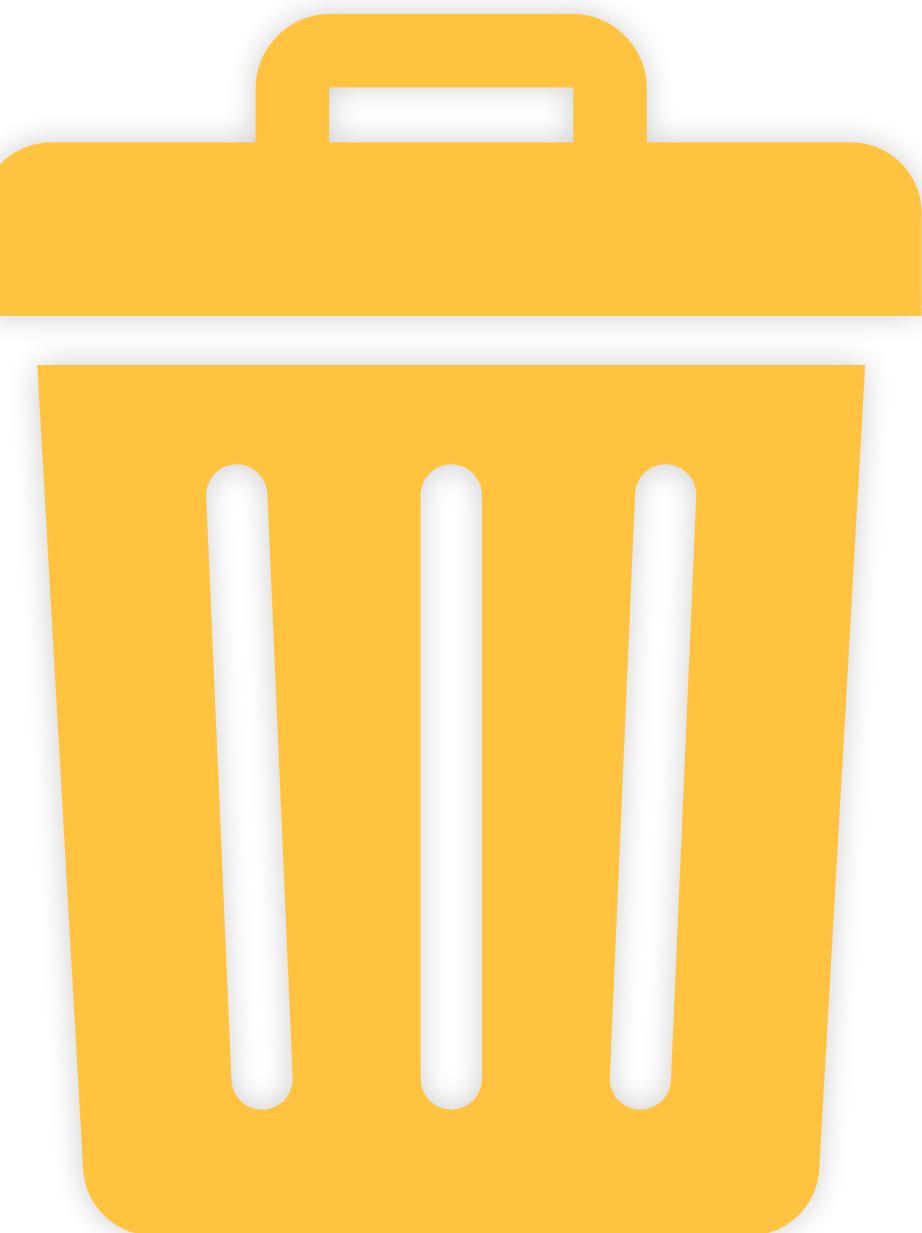
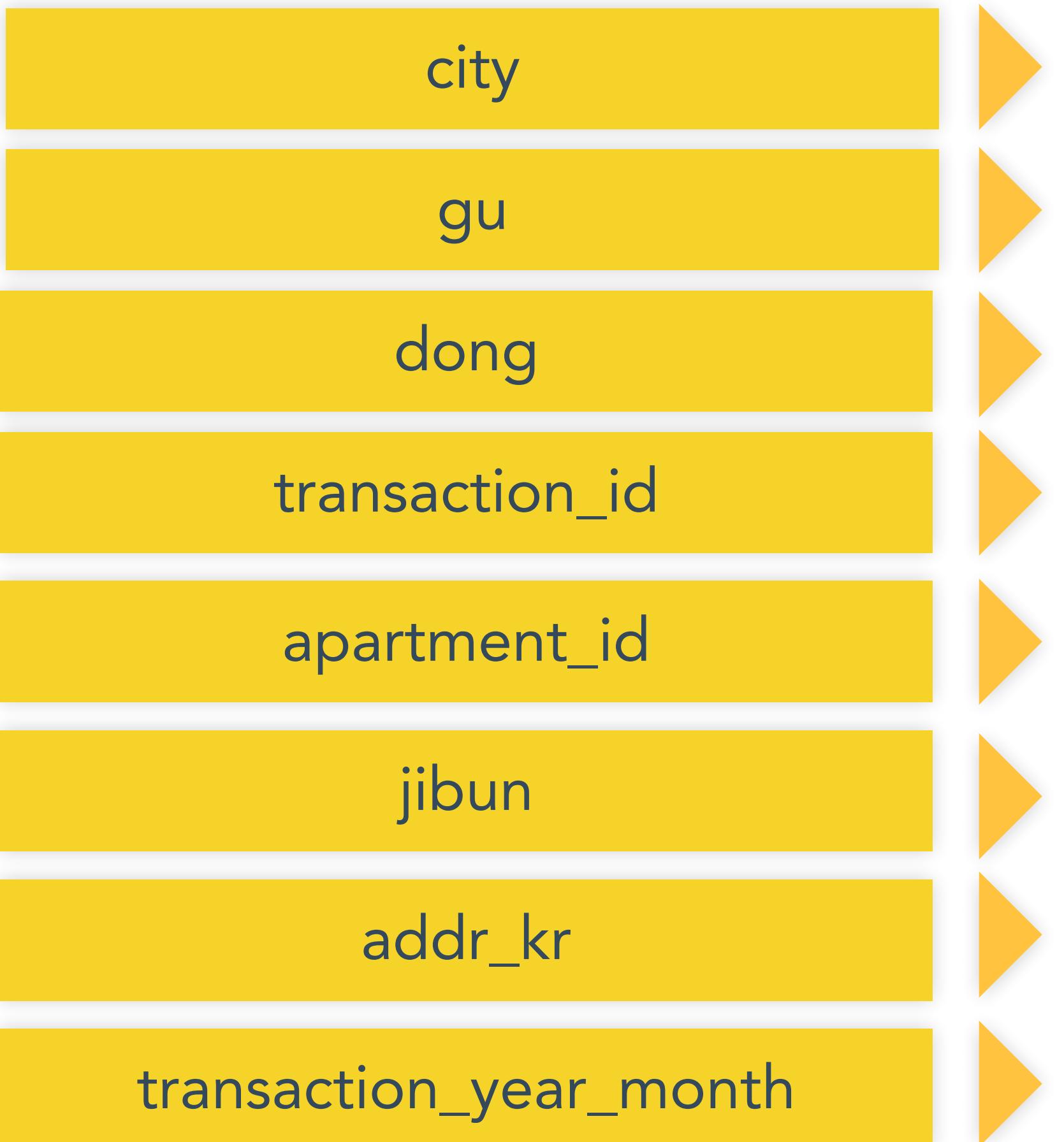


층 아파트는 3-4층이, 12층 아파트는 8층이, 15층 아파트는 10-13층이 20층 아파트는 11-18층이 25층 아파트는 20층이 가장 높은 가격임을 알 수 있다. -<https://www.hankyung.com/article/2016042785871>

“고층건축물”이란 층수가 30층 이상이거나 높이가 120미터 이상인 건축물을 말한다. 15. “초고층 건축물”이란 층수가 50층 이상이거나 높이가 200미터 이상인 건축물을 말한다. -<https://safetyman.co.kr/23>

train_df feature drop

- train_df feature drop



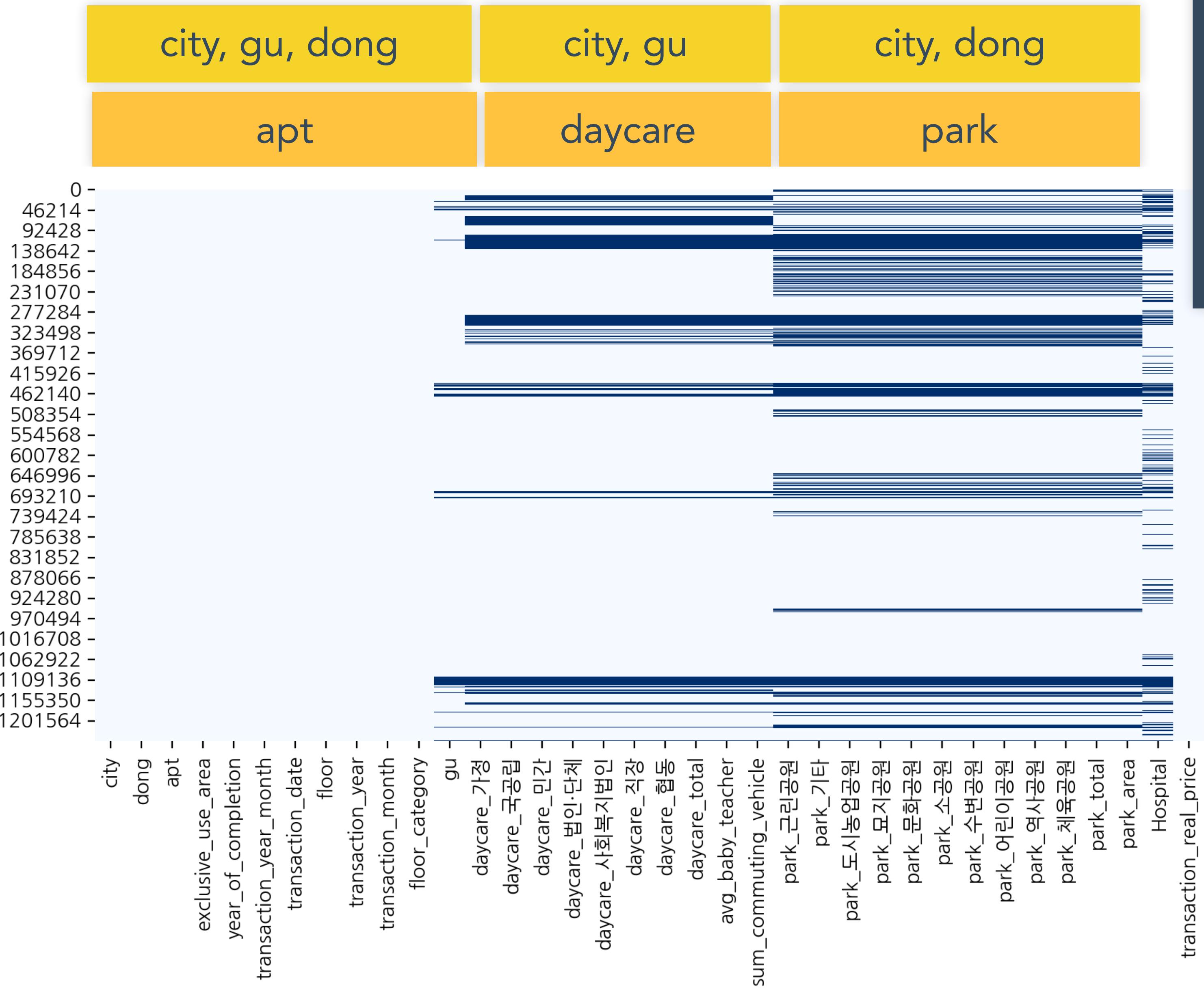
4

missing value & outlier

17

- missing value

중앙값



- outliers

선택!

대치(Imputation): 누락된 값을 대체할 수 있는 데이터, 예를 들어 해당 변수들에 대한 평균, 중앙값, 최빈값 같은 기준 데이터를 사용하여 누락된 값을 채울 수 있습니다. 이 방법은 직관적이지만 주의해서 사용해야 합니다. 왜냐하면 편향을 도입할 수 있기 때문입니다.

제외(Exclusion): 공원이나 어린이집 데이터가 없는 아파트를 분석에서 단순히 제외합니다. 누락된 데이터가 체계적으로 편향되지 않았다면, 이 방법이 실행 가능할 수 있습니다.

예측 모델링(Predictive Modeling): 다른 사용 가능한 데이터 포인트에 기반하여 누락된 값을 예측하기 위해 모델을 사용합니다. 이 방법은 단순 대치보다 더 정확할 수 있지만, 더 많은 노력을 요구합니다.

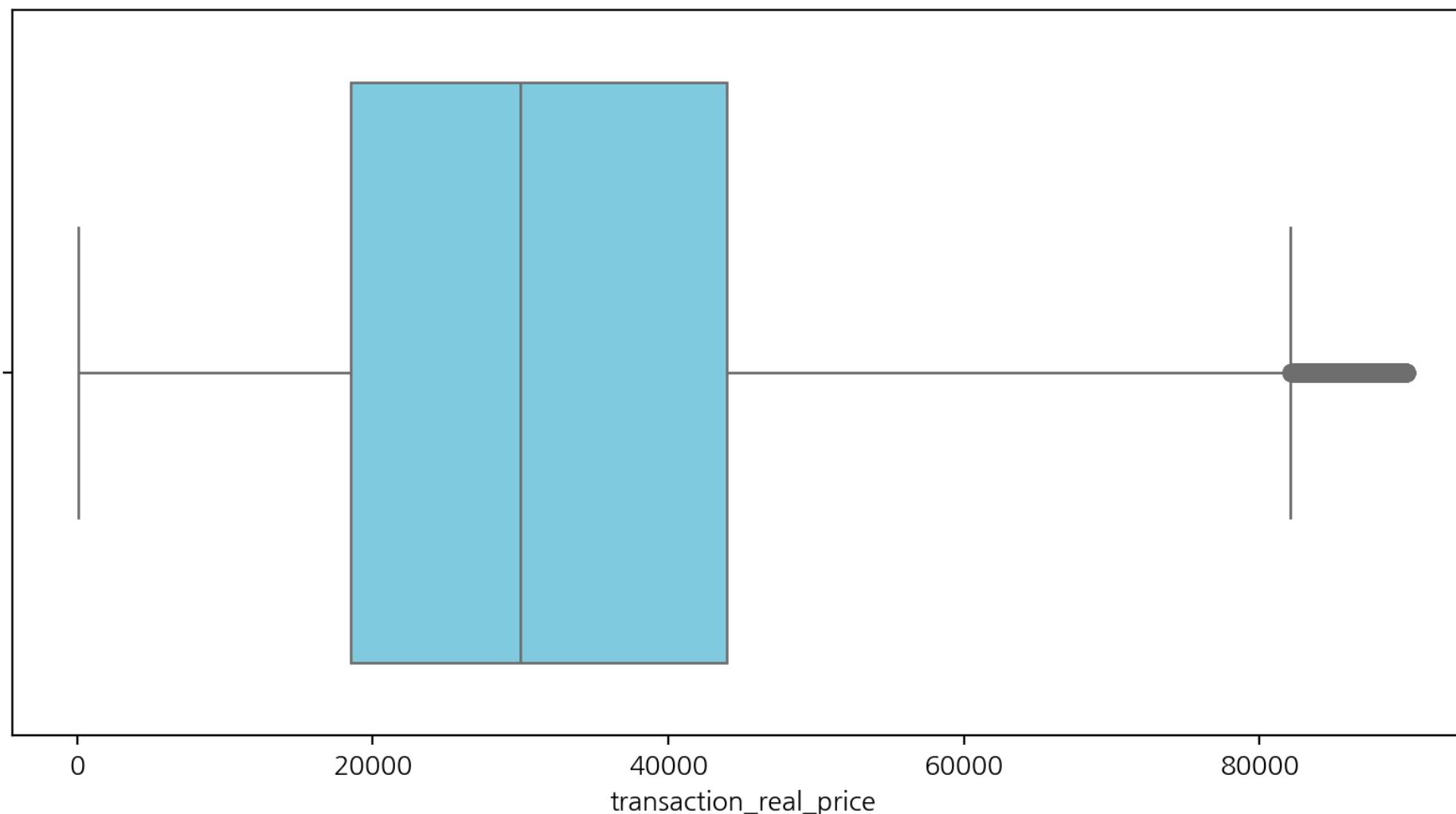
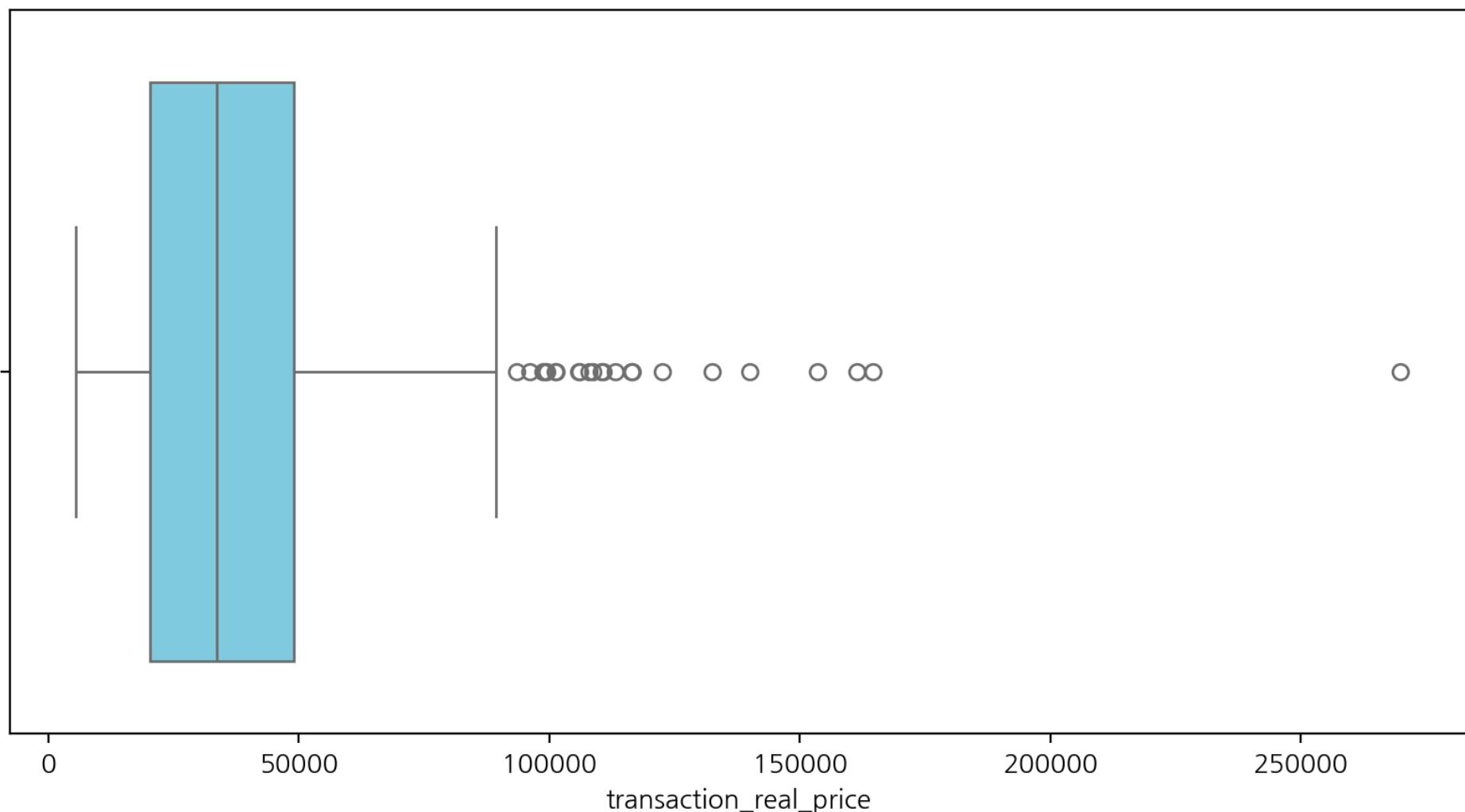
지시 변수(Indicator Variable): 데이터가 누락된 사실 자체가 정보를 제공할 수 있다면 유용할 수 있는, 누락된 데이터를 표시하기 위한 지시 변수를 생성합니다.

missing
value
&
outlier

- outliers

log transformation

drop outliers ($1.5 \times IQR$)



5

● missing value

선택!

3시그마 규칙

정의: 3시그마 규칙은 평균으로부터 표준편차의 세 배 이상 떨어진 데이터 포인트를 이상치로 간주합니다. 이 방법은 정규 분포를 가정할 때 효과적입니다.

장점: 데이터가 정규 분포를 따른다는 가정 하에, 대부분의 데이터 (99.7%)가 평균에서 ± 3 시그마 범위 안에 있어야 한다는 강력한 통계적 근거를 제공합니다.

계산이 간단하고 이해하기 쉽습니다.

단점: 데이터가 정규 분포를 따르지 않는 경우, 이상치를 제대로 식별하지 못할 수 있습니다.

특히, 꼬리가 무거운 분포에서는 많은 이상치를 놓칠 수 있습니다.

1.5IQR 방법

정의: IQR은 제3사분위수(Q3)와 제1사분위수(Q1)의 차이입니다. 1.5IQR 방법은 Q1에서 1.5IQR을 뺀 값보다 작거나, Q3에서 1.5IQR을 더한 값보다 큰 데이터 포인트를 이상치로 간주합니다.

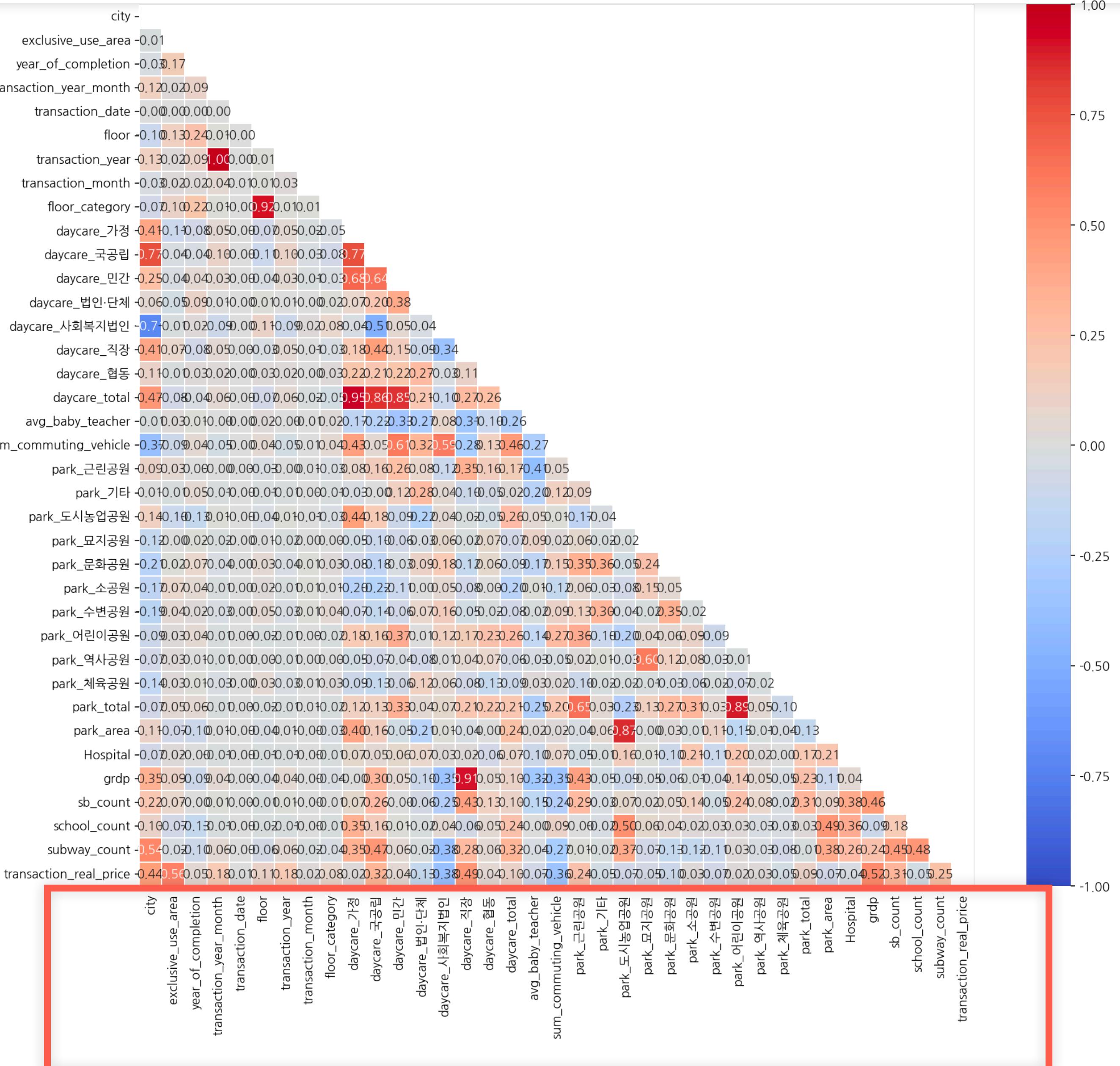
장점: 데이터의 분포가 비대칭일 때 또는 정규 분포를 따르지 않을 때 유용합니다.

중앙값과 사분위수에 기반하기 때문에, 극단적인 값의 영향을 덜 받습니다.

단점: 이상치의 존재 자체가 Q1과 Q3을 계산할 때 영향을 줄 수 있어, 매우 극단적인 이상치가 있는 경우 영향을 받을 수 있습니다.

데이터 분포의 형태에 따라, 1.5라는 계수는 너무 보수적이거나 너무 관대할 수 있으며, 때로는 3IQR 같은 다른 계수를 사용하기도 합니다.

• heatmap

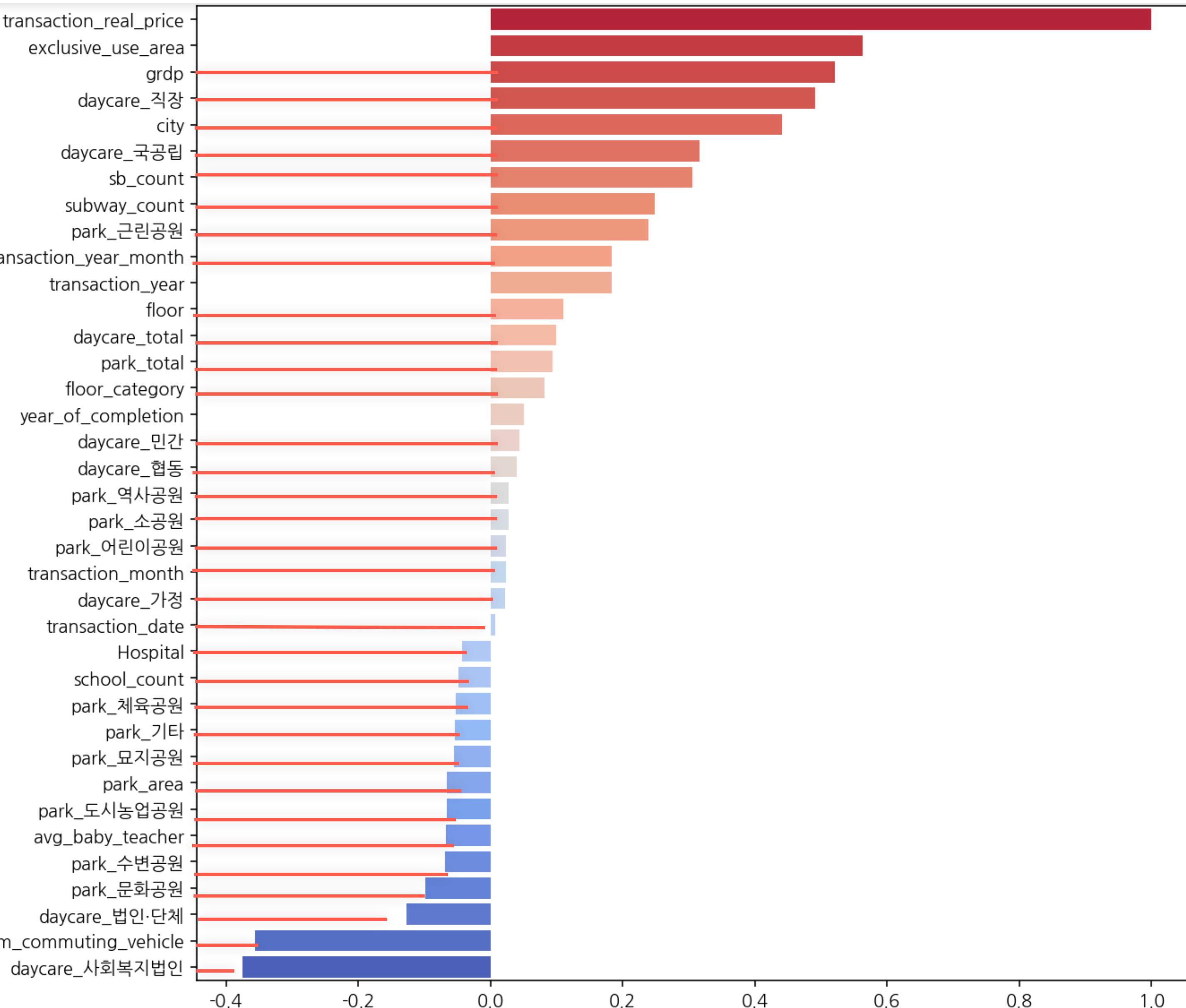


● correlation

engineered features

31 out of 37

83%



5. 첫 번째 시험: "실험실의 비밀"

- **기준 모델 성능:** "우리의 첫 번째 실험은 놀라움을 안겼습니다. 예상보다 더 정확한 예측 이었죠. 하지만 여정은 이제 시작일 뿐입니다."
- **얻은 통찰력:** "이 실험을 통해 우리는 부동산 시장의 미묘한 신호를 읽을 수 있게 되었습니다."





- DecisionTree
- RandomForestRe gressor
- LinerRegression
- XGBRegressor

- RMSE: 3,203
- R²: 0.9706

- RandomForest Regressor

- RMSE: 3,203
- R²: 0.9706

- grdp
- starbucks
- school
- subway

- RMSE: 2,799
- R²: 0.9775

- Shap

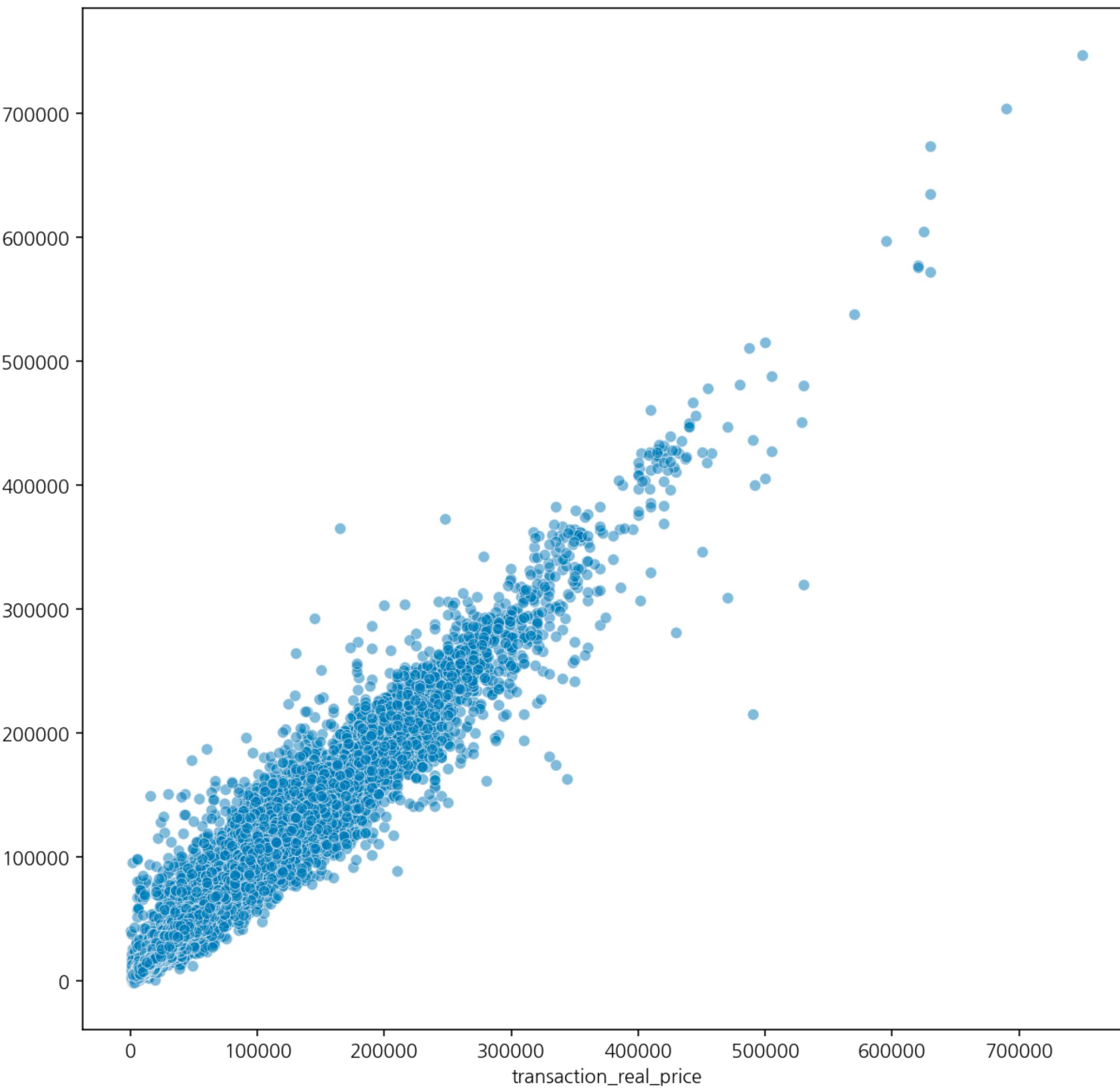
- Shap

Baseline Model

- model metrics

XGBRegressor
including outliers

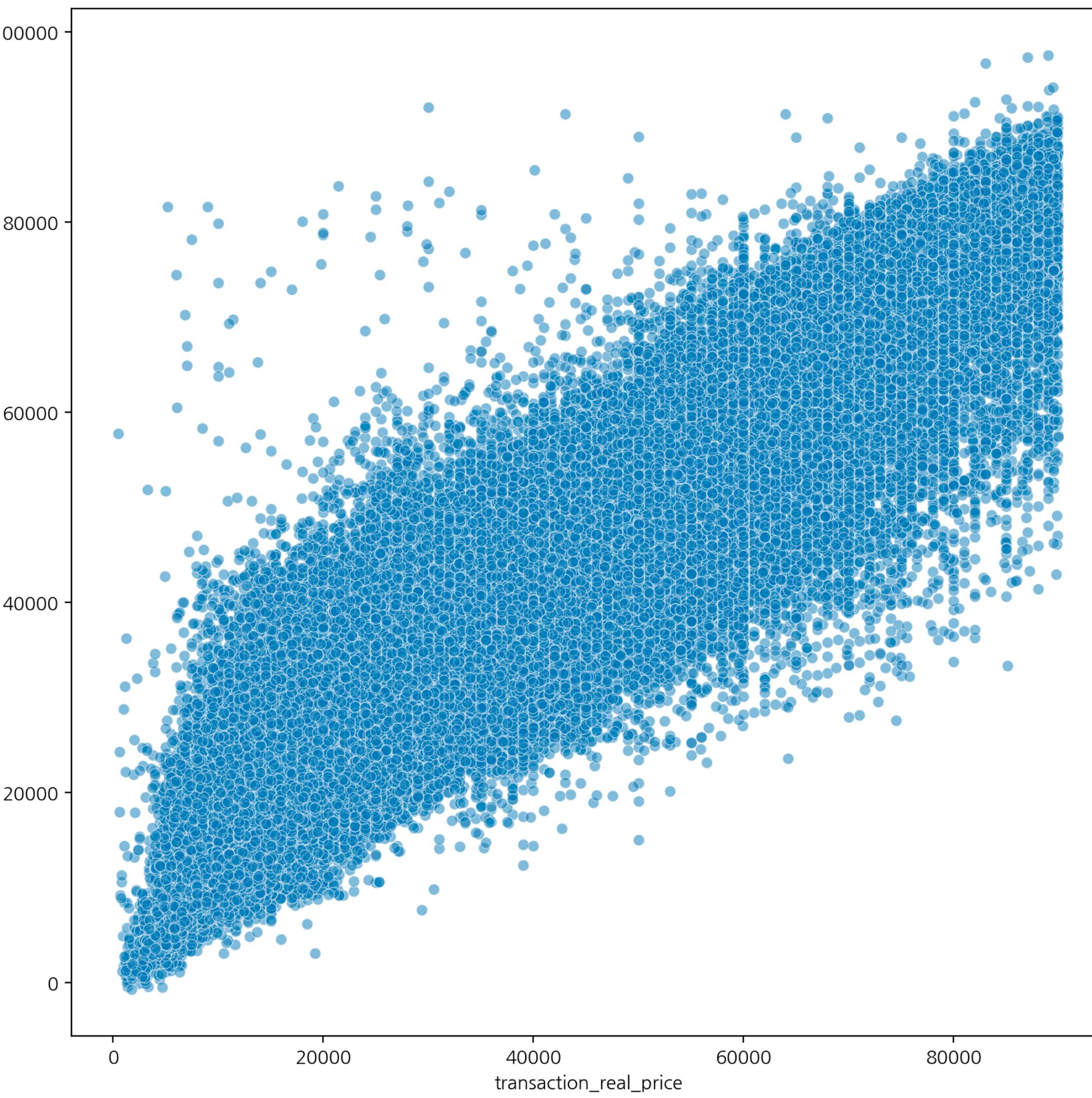
R2 score: 0.96
MAE: 3531.71
MSE: 35184594.5
RMSE: 5931.66



- model metrics

XGBRegressor
excluding outliers

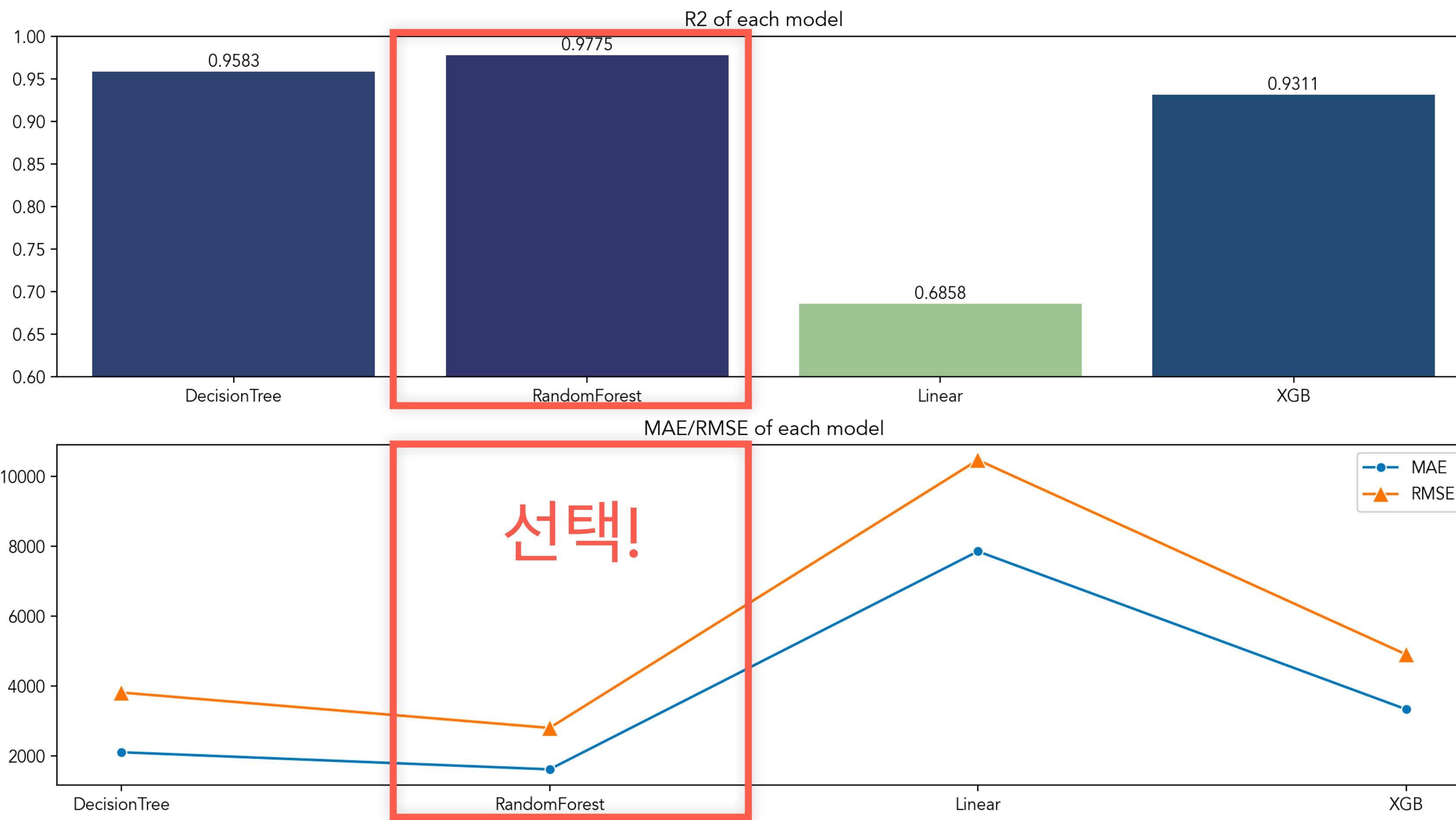
R2 score: 0.95
MAE: 2791.11
MSE: 16951295.45
RMSE: 4117.19



Baseline
Model

1

- MAE/RMSE/R² by model



Selection Model

3

- selected model metrics

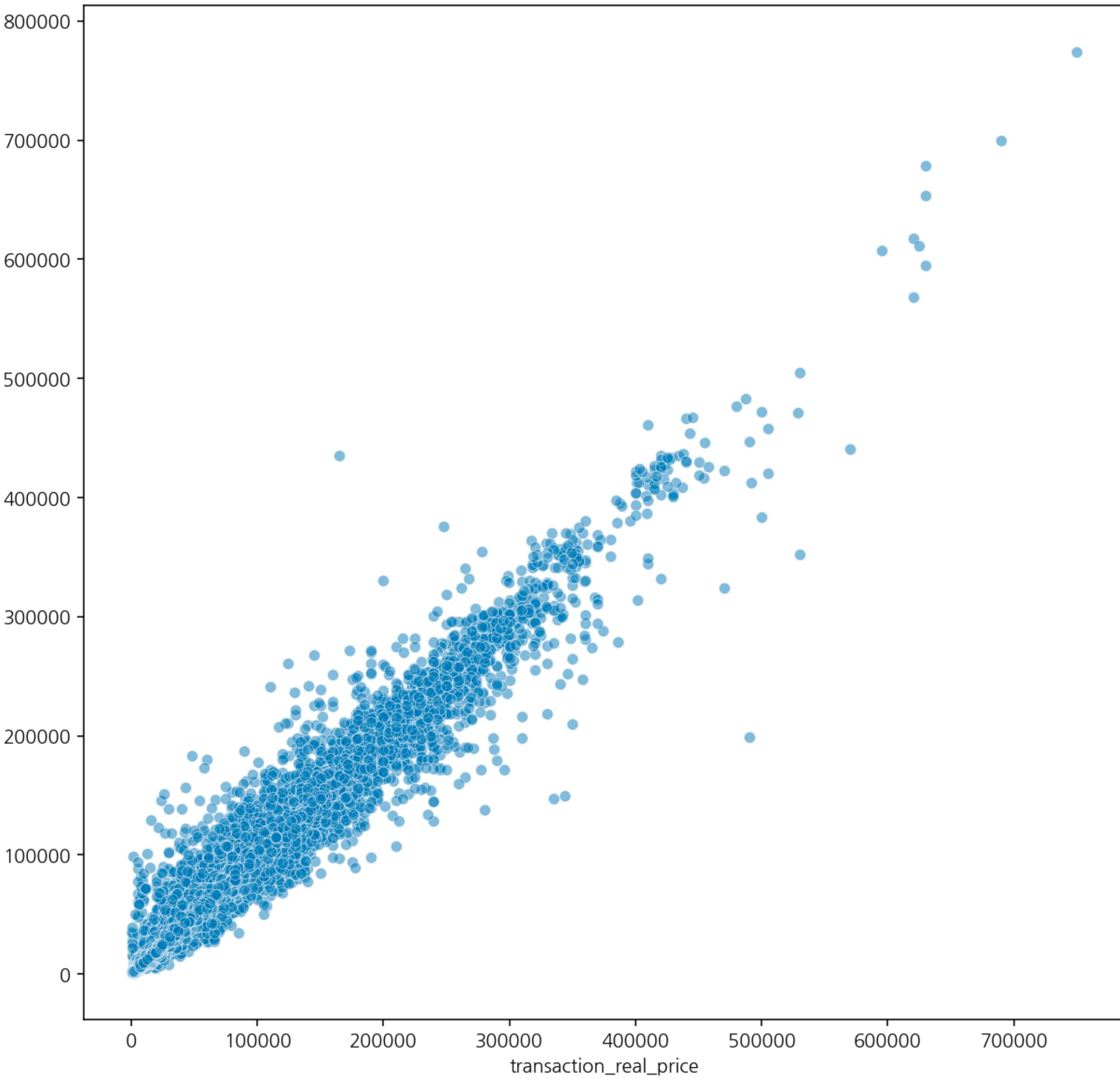
RandomForestRegressor

R2 score: 0.9811

MAE: 1928.26

MSE: 17748711.73

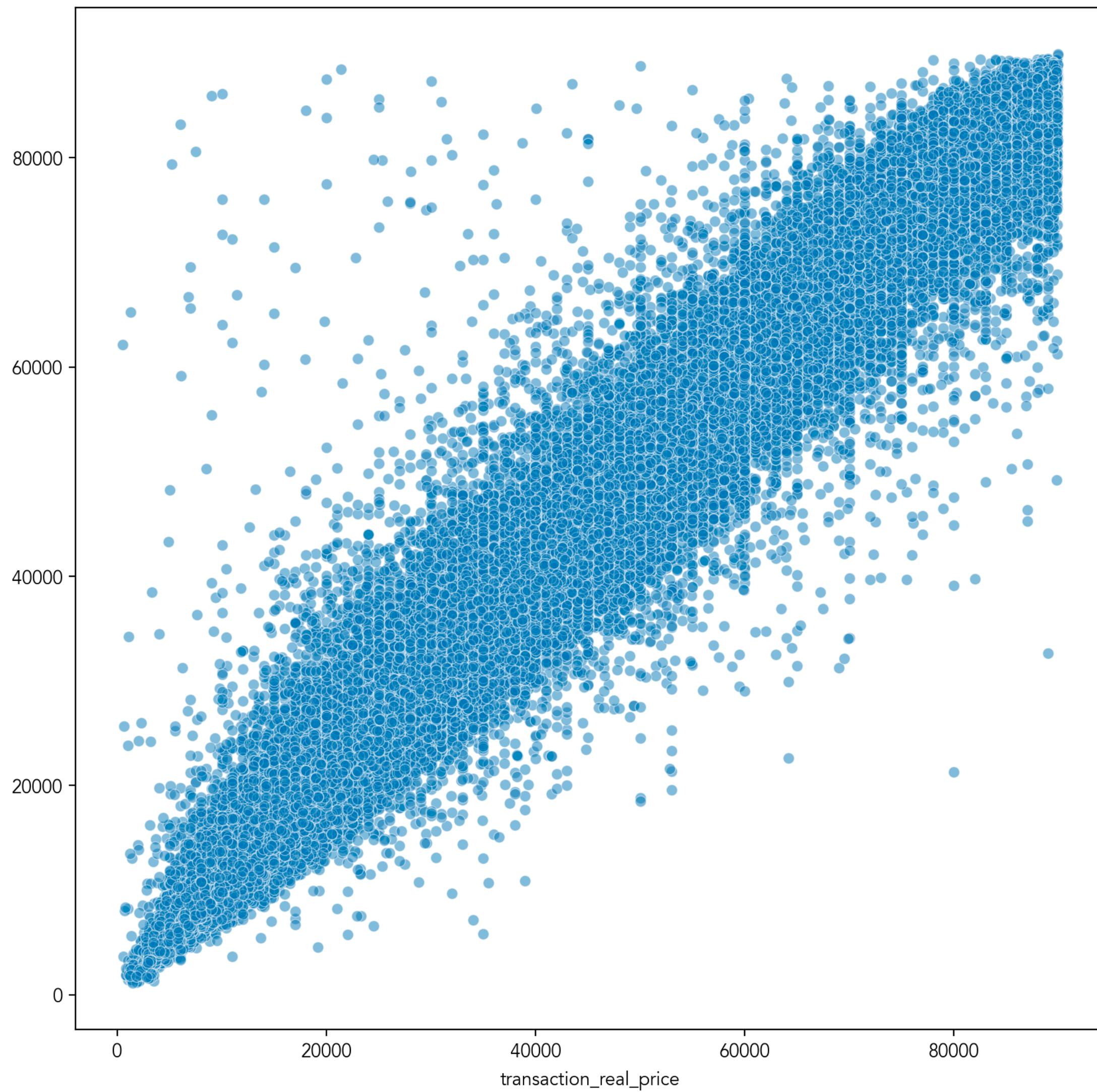
RMSE: 4212.92



- selected model metrics

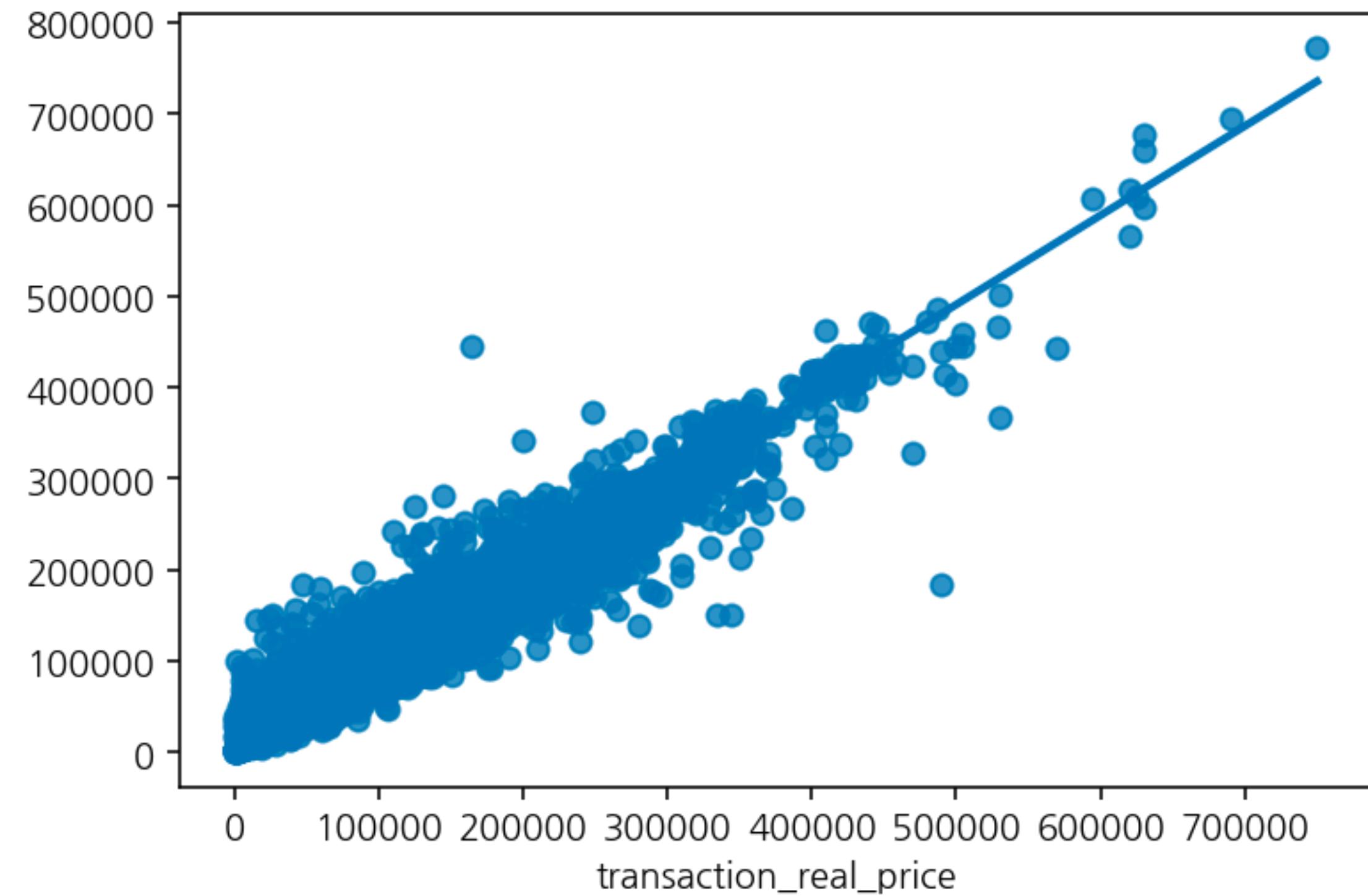
RandomForestRegressor

R2 score: 0.9775376903204372
MAE: 1615.232065443592
MSE: 7830809.020698616
RMSE: 2798.358272398053



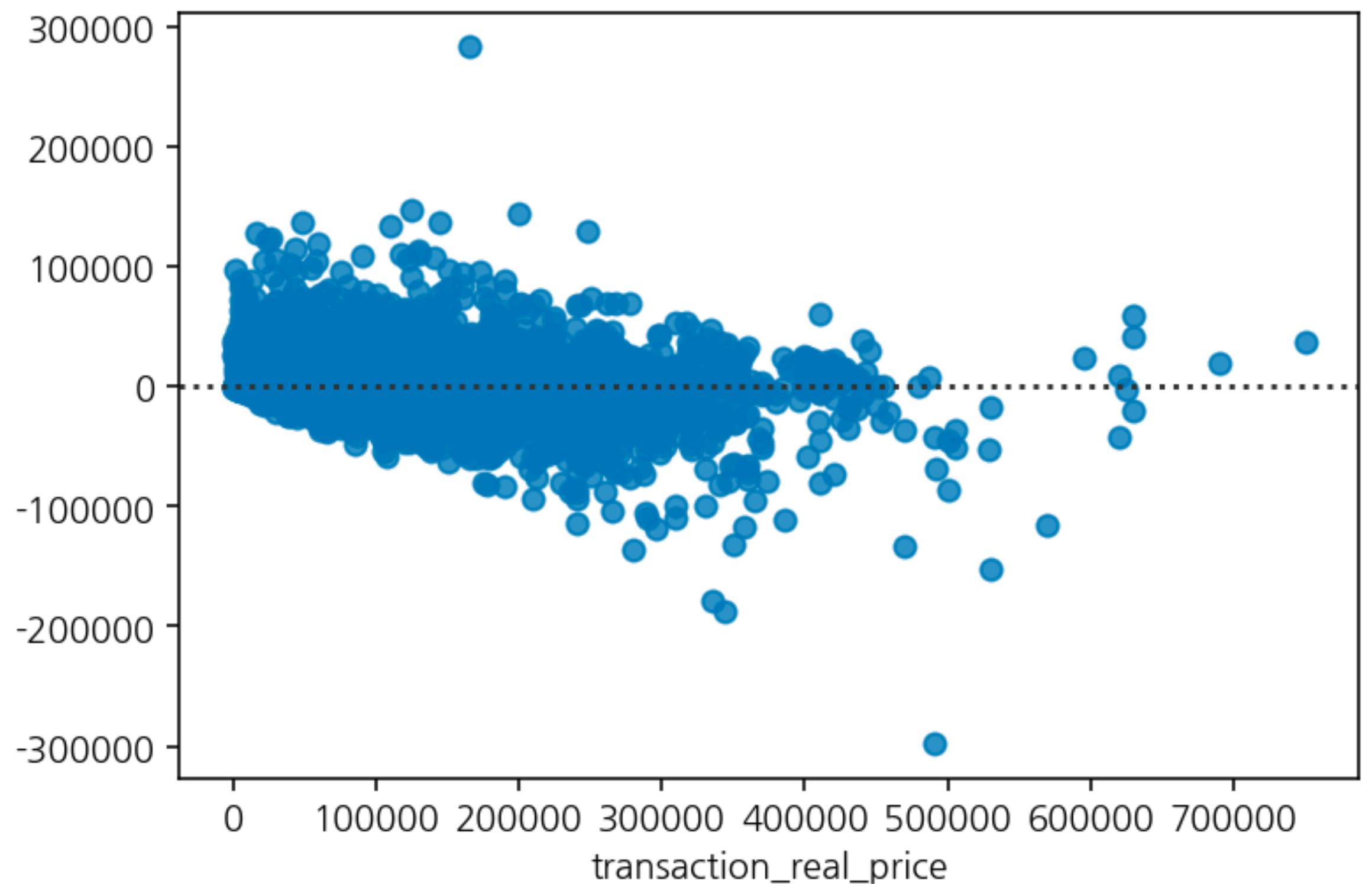
- selected model metrics

regplot
transaction_real_price와
RandomForestRegressor에 의해 예측된
가격간의 관계



- selected model metrics

residplot
실제값과 예측값 사이의 차이

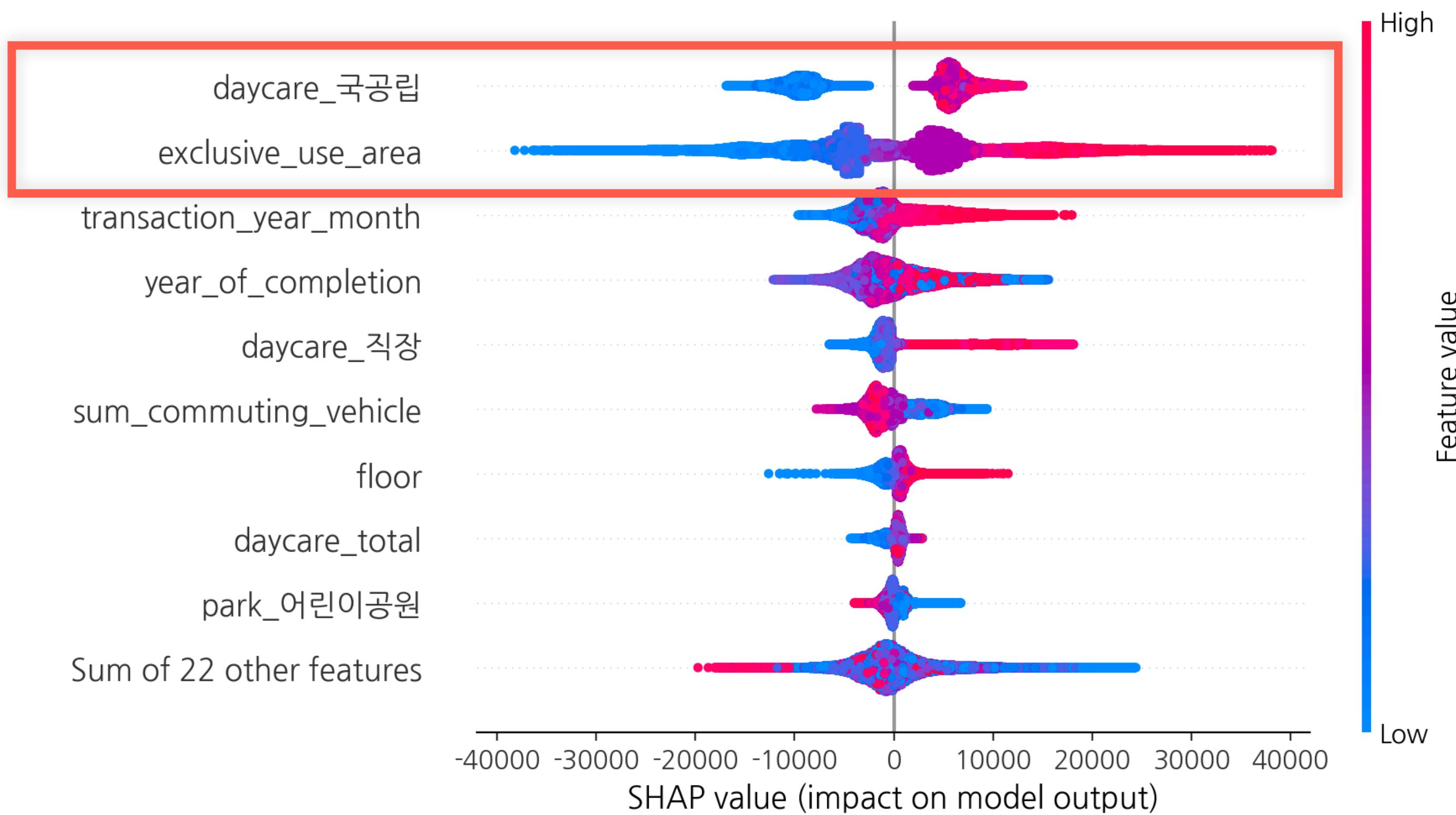


● SHAP value by feature

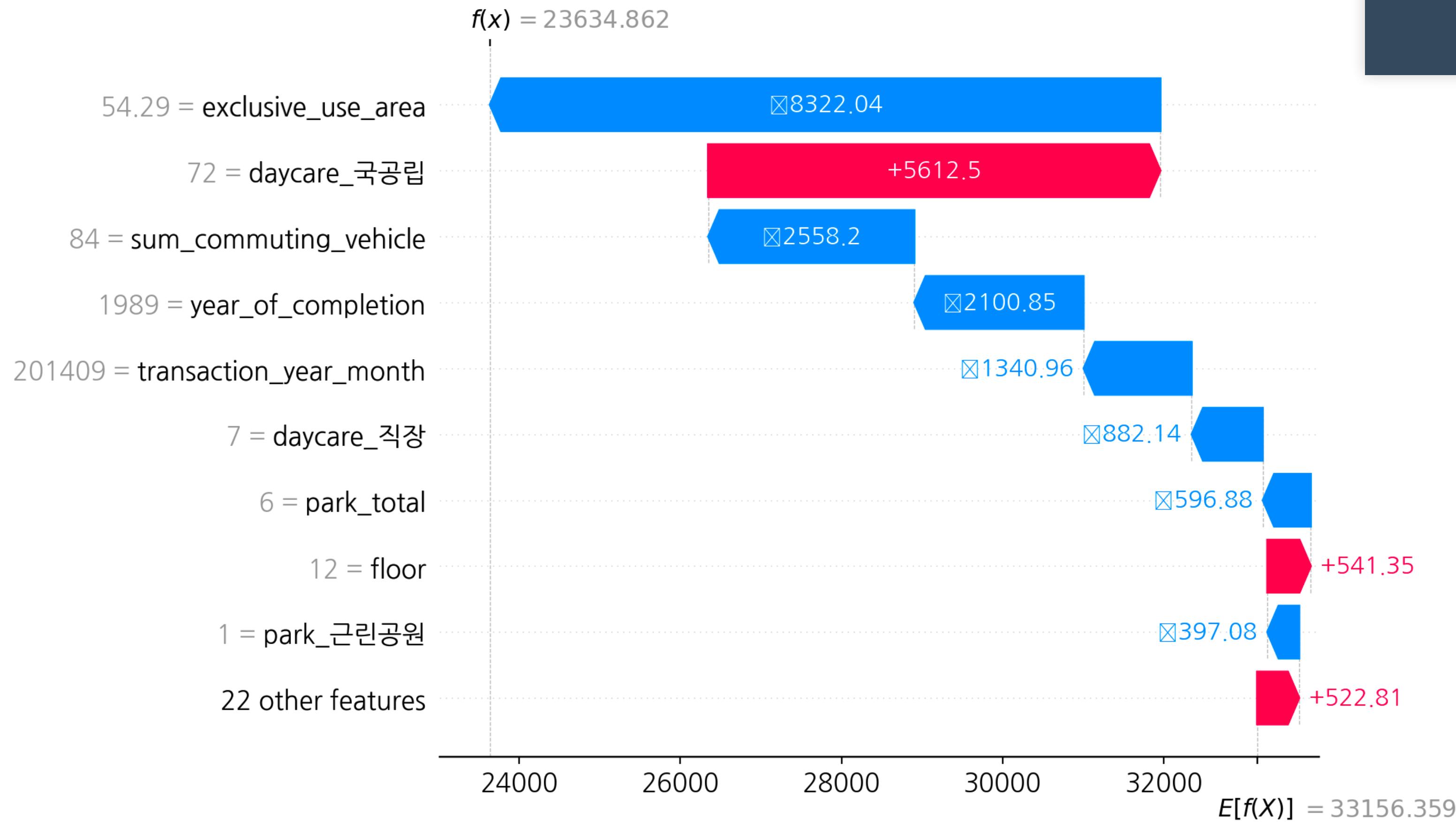
국공립 유치원 수가 많을 수록 높은 아파트
거래 가격 예측에 영향을 줌

전용 면적이 넓을 수록 높은 아파
트 거래 가격 예측에 영향을 줌

전용 면적 변수의 크기에 따라서 고가의
아파트 거래 가격이 결정이 됨



- SHAP value by record



6. 다음 목적지: "미지의 영역으로"

- **모델 개선:** "더 정교한 예측을 위해, 우리는 알려지지 않은 영역으로의 여정을 계획하고 있습니다. 더 많은 데이터, 더 복잡한 모델로 말이죠."
- **타임라인:** "시간은 우리의 동반자이자 도전입니다. 우리의 여정은 계속됩니다."

- latitude, longitude (343 hrs for train_df)

```
1 import pandas as pd
2 from tqdm import tqdm
3 from geopy.geocoders import Nominatim
4
5 train_df = pd.read_csv('/Users/kenny_jung/aiffel/data/apt/train.csv')
6
7
8 # Geopy를 사용하여 위도와 경도를 조회하는 함수
9 def get_lat_lon(address):
10     geolocator = Nominatim(user_agent="kenny")
11     location = geolocator.geocode(address)
12     if location:
13         return location.latitude, location.longitude
14     else:
15         return None, None
16
17 # 'addr_kr' 열의 각 주소에 대해 위도와 경도 조회
18 # tqdm을 사용하여 진행 상황을 표시
19 tqdm.pandas()
20 train_df[['latitude', 'longitude']] = train_df['addr_kr'][:].progress_apply(get_lat_lon).apply(pd.Series)
21
22 print(train_df.head())
23
```

⌚ 31.0s
0% | 29/1216553 [00:28<343:31:09, 1.02s/it]

7. 결론: "예측의 마법, 현실을 바꾸다"

- **작업의 영향:** "우리의 작업은 단순한 숫자의 게임이 아닙니다. 그것은 도시를 변화시키고, 사람들의 삶을 향상시킬 잠재력을 가지고 있습니다."
- **감사의 말:** "이 짧은 여정에서 함께 해준 분들에게 감사를 표합니다. 함께 라서 가능했습니다."

- 큰집이 최고!
- 부자 동네가 비싸구나.
- 맞벌이 해야해. 직장/국공립에 맡기자.
- 역시 역세권...
- 거주? 투자!
- 별다방은 아무데나 않들어와!
- 새집이 좋아!
- 근린공원에서 운동해야지
- 고층이 럭셔리 하지!

투자

할때 고려해볼 사항들

만들때 참조해야 할 사항들

정책

- 사설 유치원은 가기 힘들어.
- 자가용 보다 통학버스는 필수!
- 보육교사당 원생이 너무 많아!!
- 땅이 넓으니 수변/문화공원 많음 ㅎ ㅎ
- 병원 바로 옆은 글쎄???
- 근처에 묘지/체육/도시농업 공원????
- 학교 근처는 조금 시끄러울까?

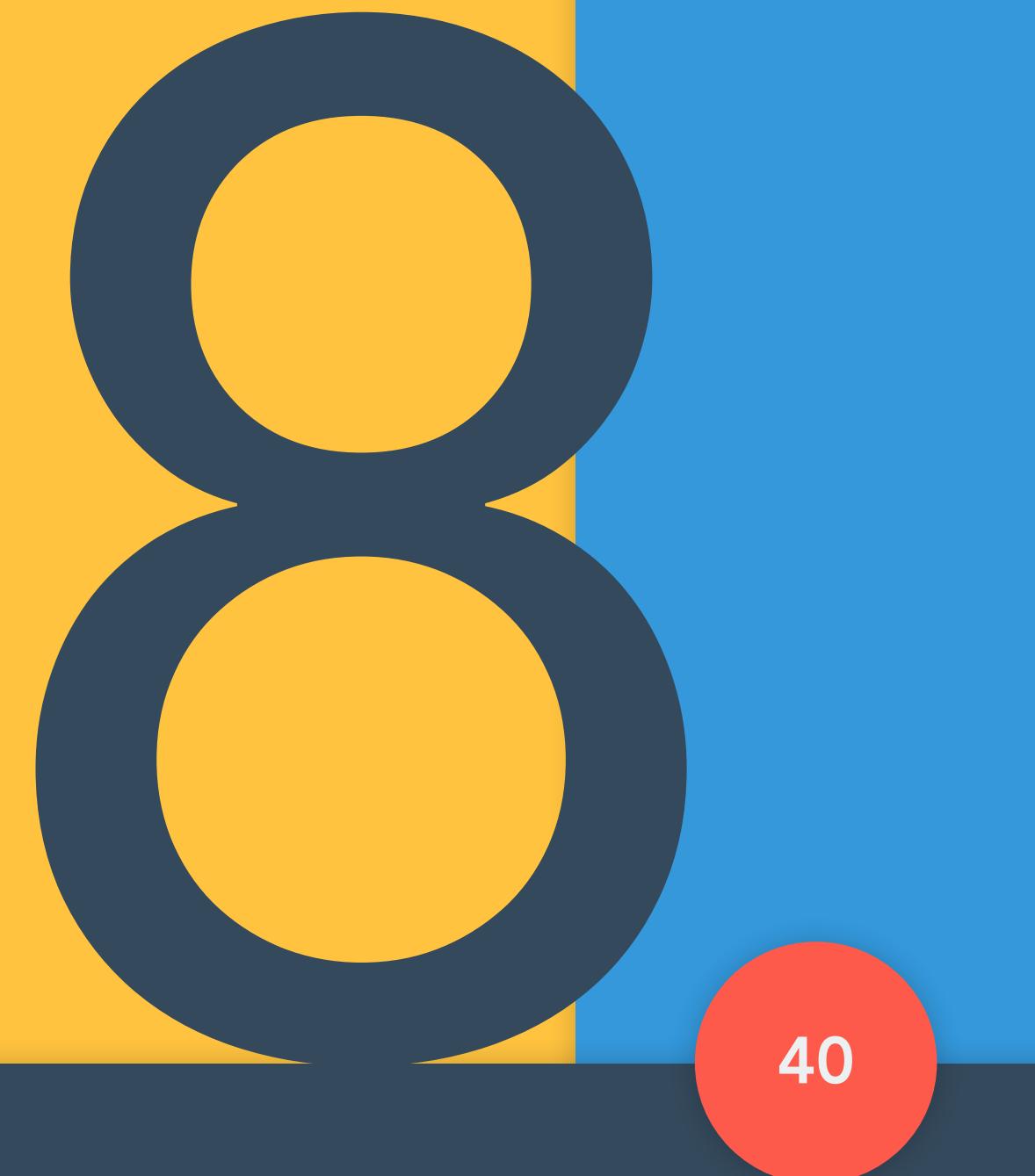
- **부동산 시장의 투명성 증진:** 아파트 가격 예측 모델은 시장의 투명성을 증진시키는데 기여할 수 있습니다. 예측 결과를 통해 구매자와 판매자 모두가 보다 정보에 기반한 결정을 내릴 수 있게 되며, 이는 공정한 가격 형성에 도움이 됩니다.
- **도시 계획 및 개발에 대한 인사이트:** 예측 모델은 특정 지역의 부동산 가치가 어떻게 변화할지에 대한 인사이트를 제공할 수 있습니다. 이 정보는 도시 계획가와 정책 입안자가 인프라 개발, 공공 서비스 배치, 주거 정책을 수립하는 데 중요한 참고 자료가 될 수 있습니다.
- **투자 기회 발견:** 정확한 가격 예측은 부동산 투자자들에게 잠재적인 투자 기회를 발견할 수 있는 근거를 제공합니다. 이는 특히 변화하는 시장 조건 하에서 투자 결정을 내리는 데 유용합니다.
- **기술 혁신의 촉진:** 아파트 가격 예측과 같은 문제를 해결하기 위해 개발된 기술과 알고리즘은 다른 산업 분야에서도 응용될 수 있습니다. 예를 들어, 비슷한 방법론을 사용하여 다른 자산의 가치를 예측하거나 소비자 행동을 분석하는 데 활용할 수 있습니다.
- **사회 경제적 불평등 해소:** 부동산 가격 예측 모델을 통해 가격 변동성을 이해하고 예측함으로써, 주거 비용 부담이 높은 지역에 대한 정책적 개입을 모색할 수 있습니다. 이는 장기적으로 사회 경제적 불평등을 완화하는데 기여할 수 있습니다.

참조 자료 출처

- 어린이집 주소 정보: <https://info.childcare.go.kr/info/oais/openapi/OpenApilInfoSl.jsp>
- 스타벅스 점포 정보: <https://uincity.tistory.com/299>
- 지역내 총생산: https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1C65_03E&conn_p
- 지역별 학교: <https://www.data.go.kr/data/15021148/standard.do>
- 지역별 지하철: https://data.kric.go.kr/rips/M_01_01/detail.do?id=32

8. 호기심의 시간: "궁금증을 풀어드립니다"

- **질문 요청:** "이제 여러분의 차례입니다. 궁금한 점을 물어보세요. 우리의 여정에 대해 더 알고 싶으신가요?"



회고

- 모두 같이 노력한 결과(비록 부족하지만)를 확인하여 너무 기쁘고 보람찹니다.
- 좀더 상세한 EDA를 해보고 싶었지만 시간적인 제약이 있었습니다.
- Sampling 통한 hyper-parameter tuning 시도해 보고 싶습니다.
- Sampling 통한 SHAP 분석해 보고 싶습니다.
- 위치 기반 거리 계산을 통한 Feature Engineering이 유효할것 같습니다.
- 초등학교,중학교,고등학교로 세분화 하면 유의미한 결과가 나왔을 것 같습니다.
- 결측치에 대한 고민을 좀 더 깊게 했으면 좋았을 것 같다.
- Outlier 처리 방안에 대해서 다양한 시도를 해보면 좋을것 같다.
- 각자 역할에 충실하여 주어진 시간에 결론을 만들었습니다. 모두 고생 하셨습니다.
- 또 같이 일해요!

감사 합니다.

