3RD PROJECT

관악구 부동산의 현재와 미래: 데이터 분석과 예측 모델링

이인의 데이터 분석 프로젝트

2024.10.25

목차

01. 프로젝트 개요

01.프로젝트 개요

주제 관악구 부동산 예측 모델링 프로젝트

목표

- 서울시 관악구의 부동산 실거래가 데이터를 활용하여 미래 부동산 가격을 예측하고 의사결정을 지원.
- part 1. 데이터 기반 분석 및 part 2. 머신러닝 모델 개발을 통해 시장의 주요 요인 파악과 가격 예측을 수행.

데이터 출저

서울특별시 도시계획국 토지관리과
부동산 실거래가 데이터
(서울 부동산 정보광장 제공)

01.프로젝트 개요

진행 절차

01	프로젝트 목표 정의	(09.28 ~ 09.30)
02	데이터 수집 및 전처리	(09.30 ~ 10.03)
03	탐색적 데이터 분석 (EDA)	(10.04 ~ 10.08)
04	모델 선택 및 학습	(10.09 ~ 10.14)
05	모델 평가 및 개선	(10.15 ~ 10.20)
06	결과 시각화 및 ppt 작업	(10.21 ~ 10.25)

01.프로젝트 개요

분석 도구

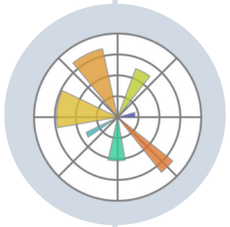
개발 환경

Jupyter Notebook
Google Colab

데이터 분석

Python (pandas, numpy)

데이터 시각화



matplotlib, seaborn

모델 학습

Scikit-learn

01.프로젝트 개요

데이터 개요

총 관측치 및 특징 개수

총 76,894개의
거래 데이터

18개의 특징
(features)

데이터 설명

서울특별시 관악구의
부동산 실거래 데이터
(주로 봉천동, 남현동, 신림
동, 상도동 등)

2006년부터 2024년
까지의 실거래 기록

01.프로젝트 개요

변수 목록

주요 변수

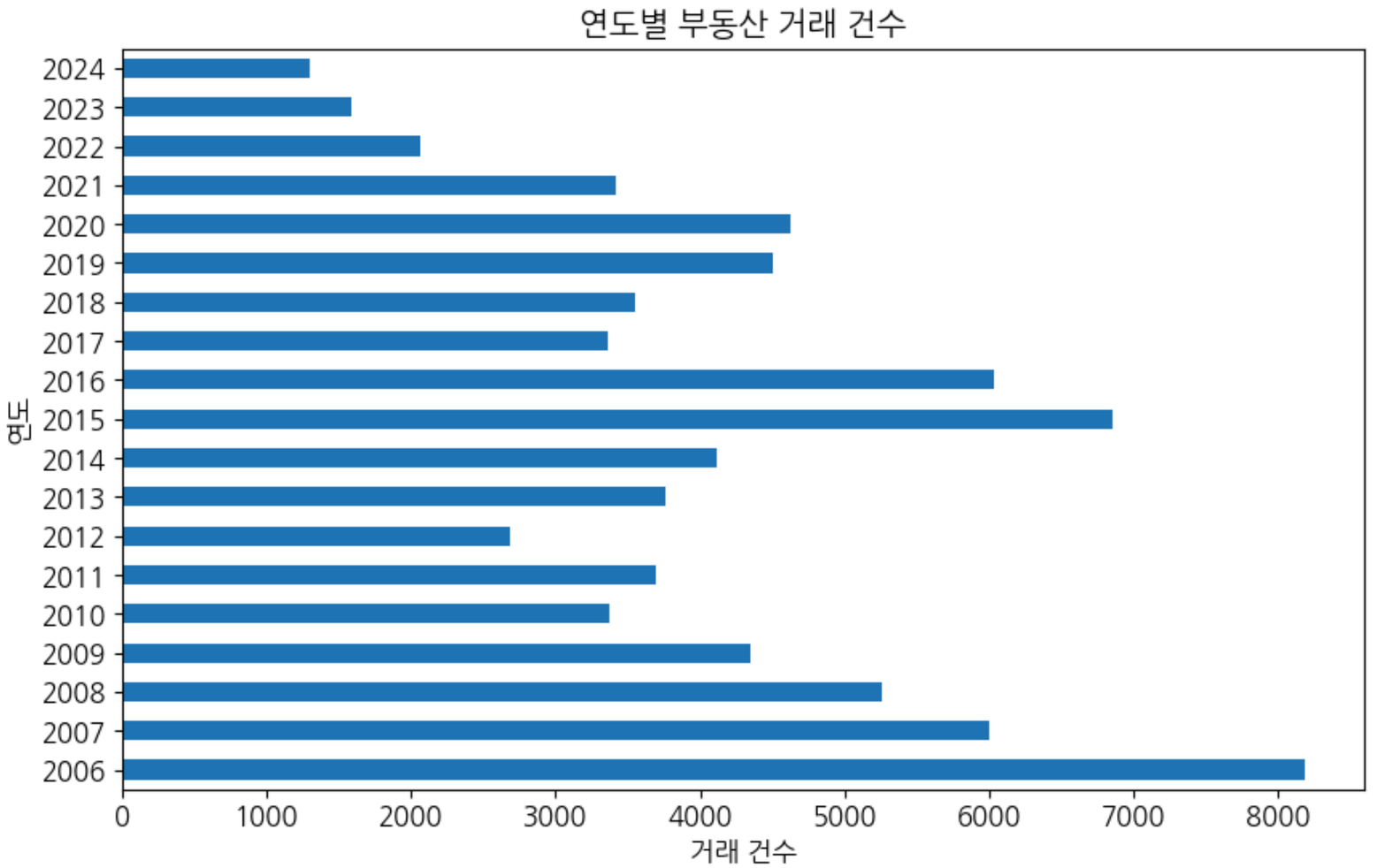
Integer Object Date

법정동코드	법정동명	지번구분명	건물명
계약일	물건금액(만원)	건물면적(㎡)	토지면적(㎡)
층	건축년도	건물용도	연도,월,일
분기	요일	건물면적(평)	평당가격(만원)

Part 1. 데이터 분석

Part 1.
데이터 분석

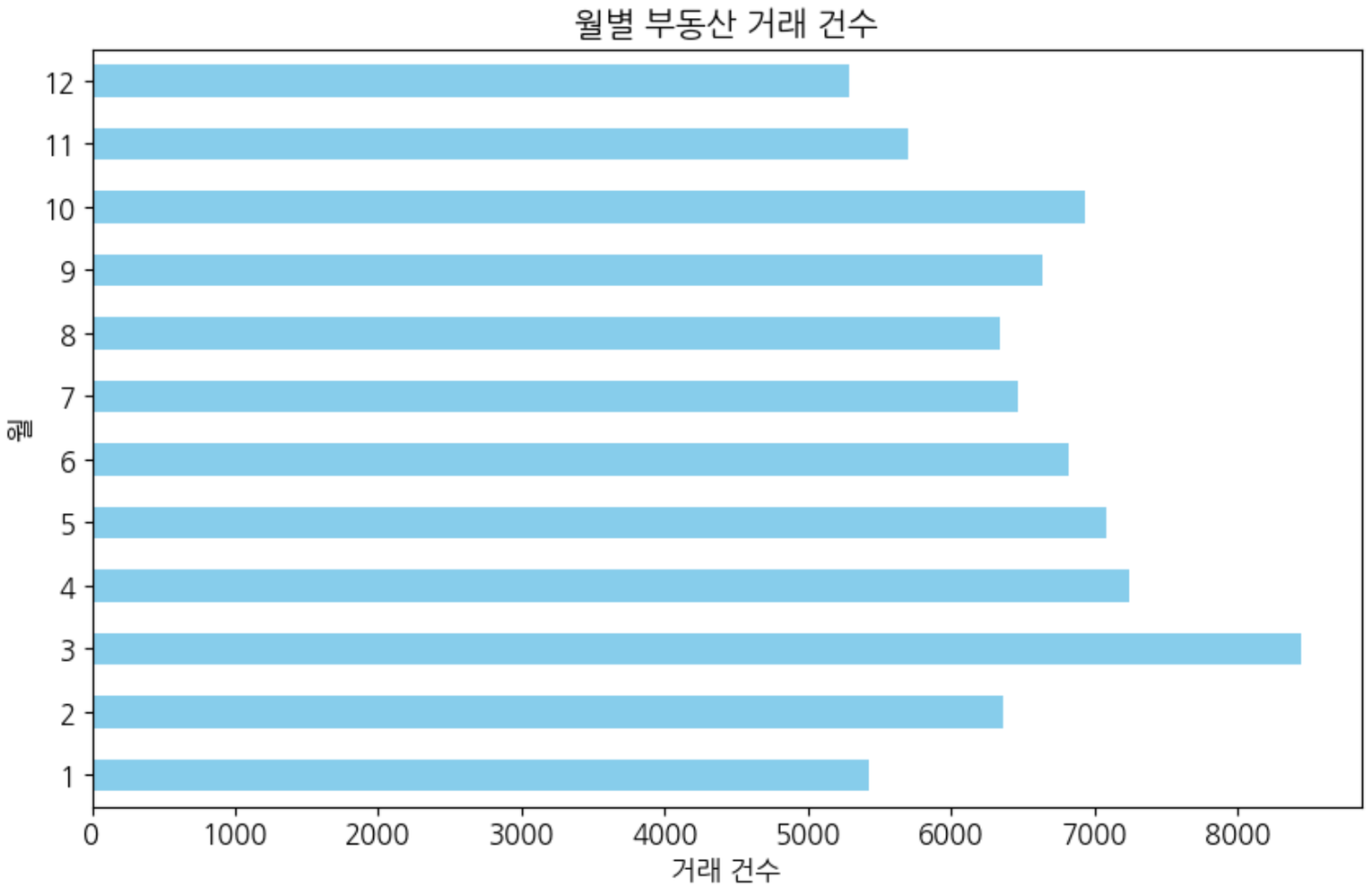
(차트 1)
연도별 부동산 거래 건수 분석



- **2006년**에 거래 건수가 가장 많았으며, **8,000건 이상** 기록.
- 최근 몇 년간 **거래량**이 **감소하는 추세**를 보였으며, 특히 **2022~2024년** 동안 **거래량 감소**가 두드러짐.
- 이러한 감소는 **경제적 요인(금리 인상, 대출 규제)**과 **정책적 요인**의 영향을 반영함.

Part 1.
데이터 분석

(차트 2)
월별 부동산 거래 건수 분석

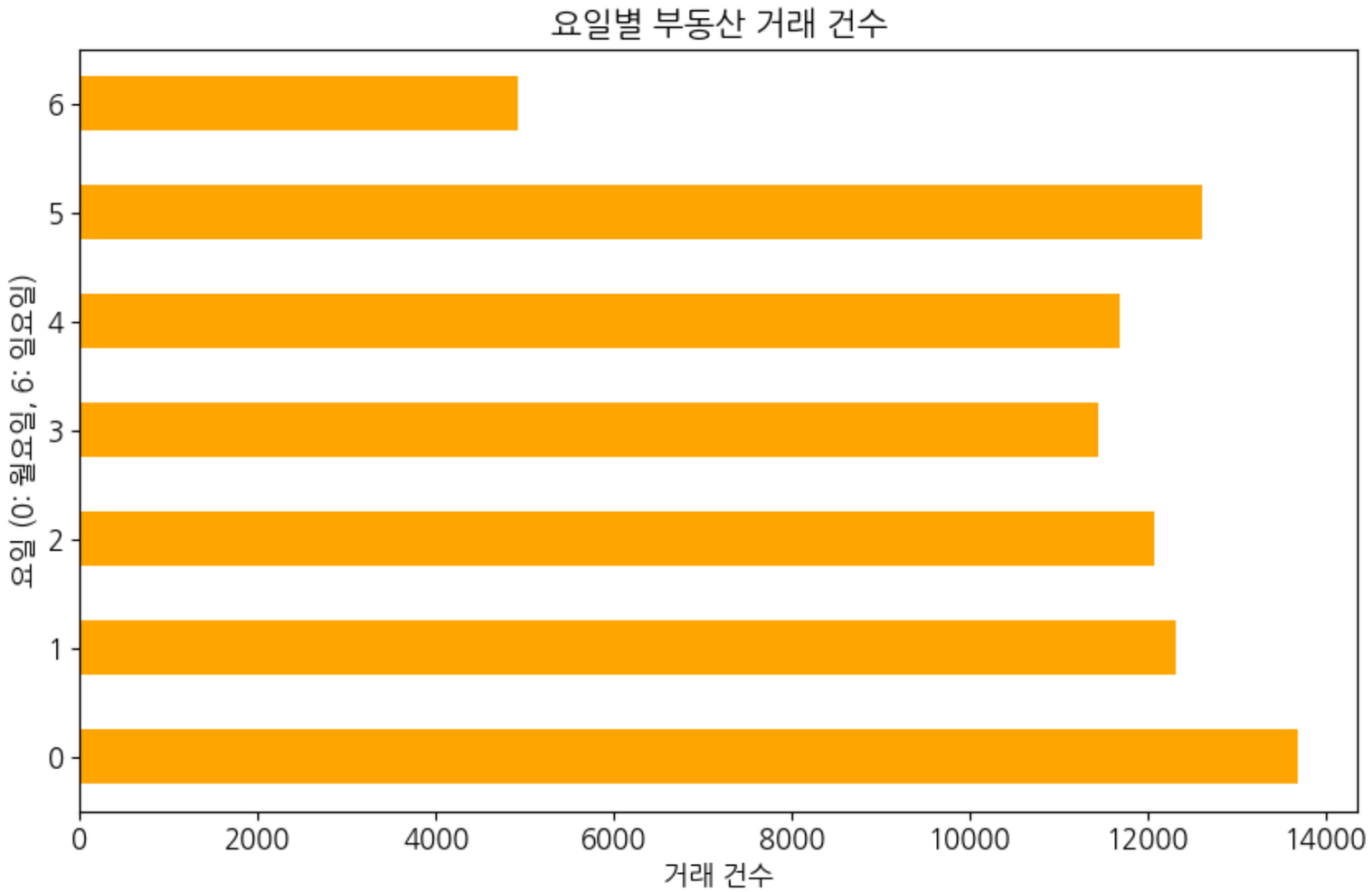


- **3월**에 거래량이 가장 많으며, **8,000건 이상** 거래.
- 겨울철인 **12월**에 거래가 가장 적음.
- 봄과 여름이 부동산 시장에서 **활발한 거래 시기**임을 나타냄.

Part 1.
데이터 분석

(차트 3)

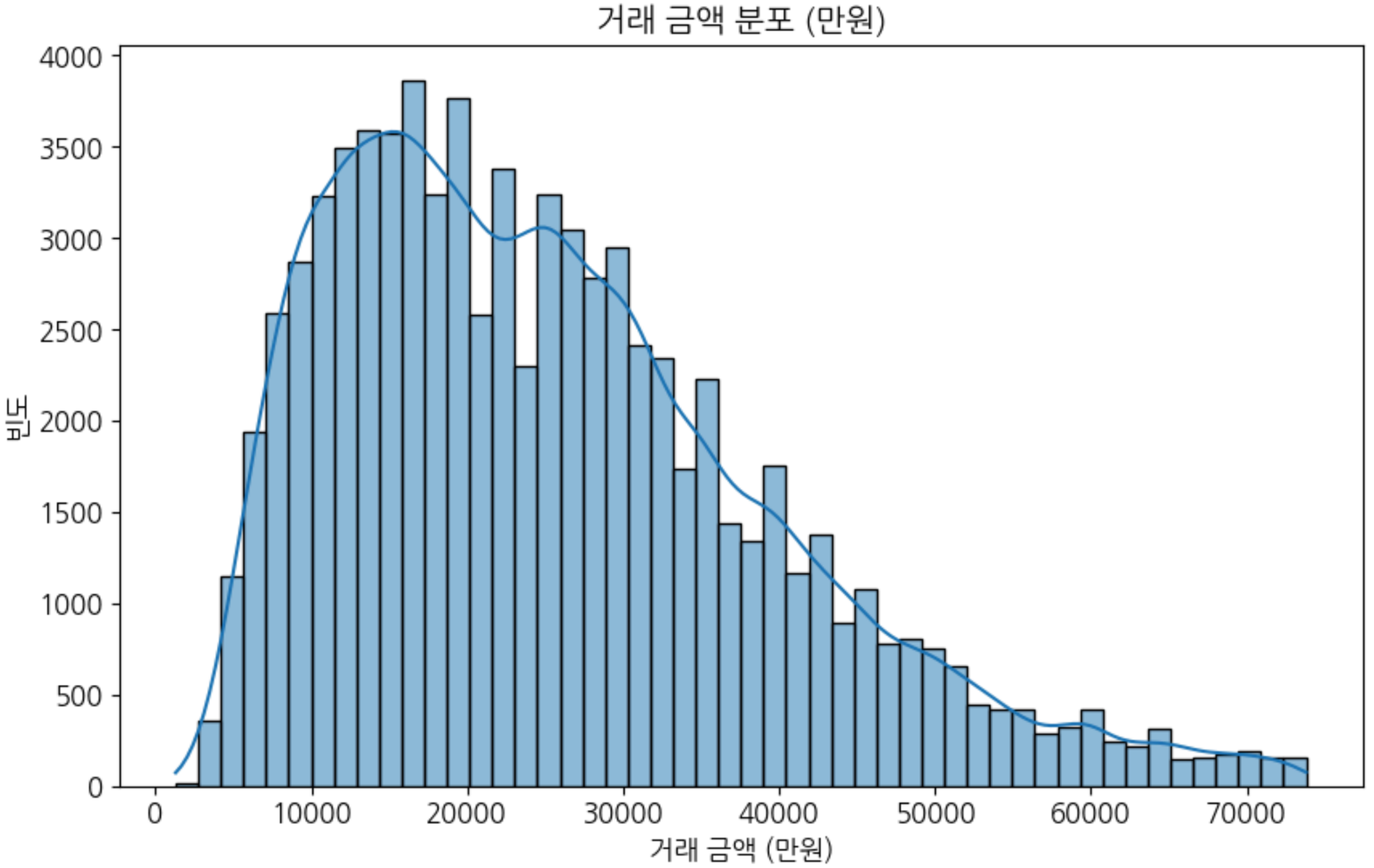
평일과 주말의 거래 패턴 비교



- **평일(월~금)**에 거래가 집중되며, 특히 **월요일**에 거래가 가장 활발.
- **토요일**에는 거래 건수가 월요일 만큼 **활발한 편**.
- **일요일**에 거래 건수가 급격히 **줄어듦**.

Part 1.
데이터 분석

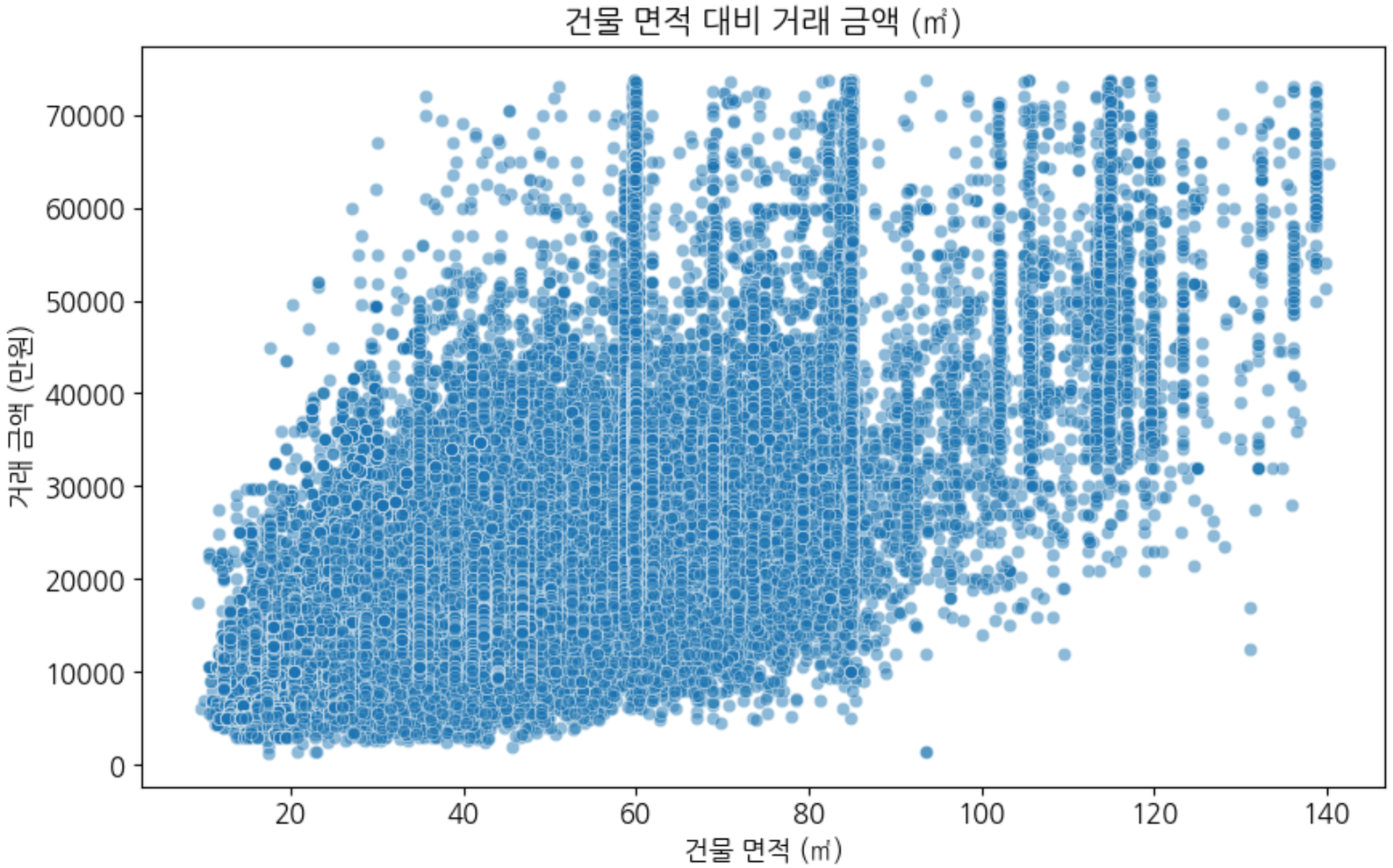
(차트 4)
거래 금액 분포 분석



- **거래 금액**은 **10,000만원 ~ 30,000만원** 구간에 집중되어 있음.
- 고가의 **거래(30,000만원 이상)**는 상대적으로 **적음**.
- 대다수 거래가 **중간 가격대**에서 이루어짐을 확인.

Part 1.
데이터 분석

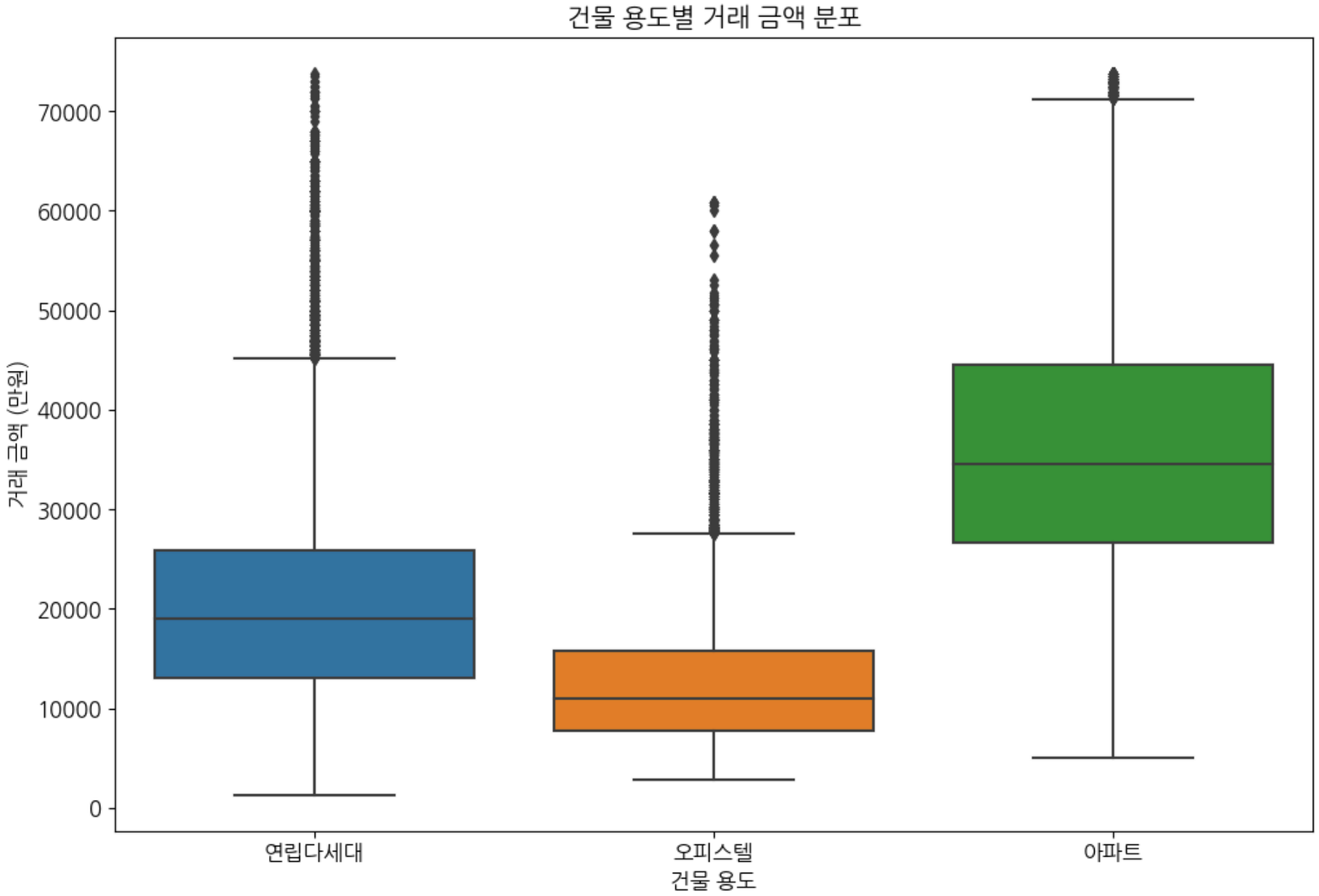
(차트 5)
건물 면적 대비 거래 금액 분석



- **건물 면적**이 **넓을수록** 거래 금액이 **증가**하는 경향.
- **특정 면적 구간(예: 60~80㎡)**에서 높은 거래 빈도.

Part 1.
데이터 분석

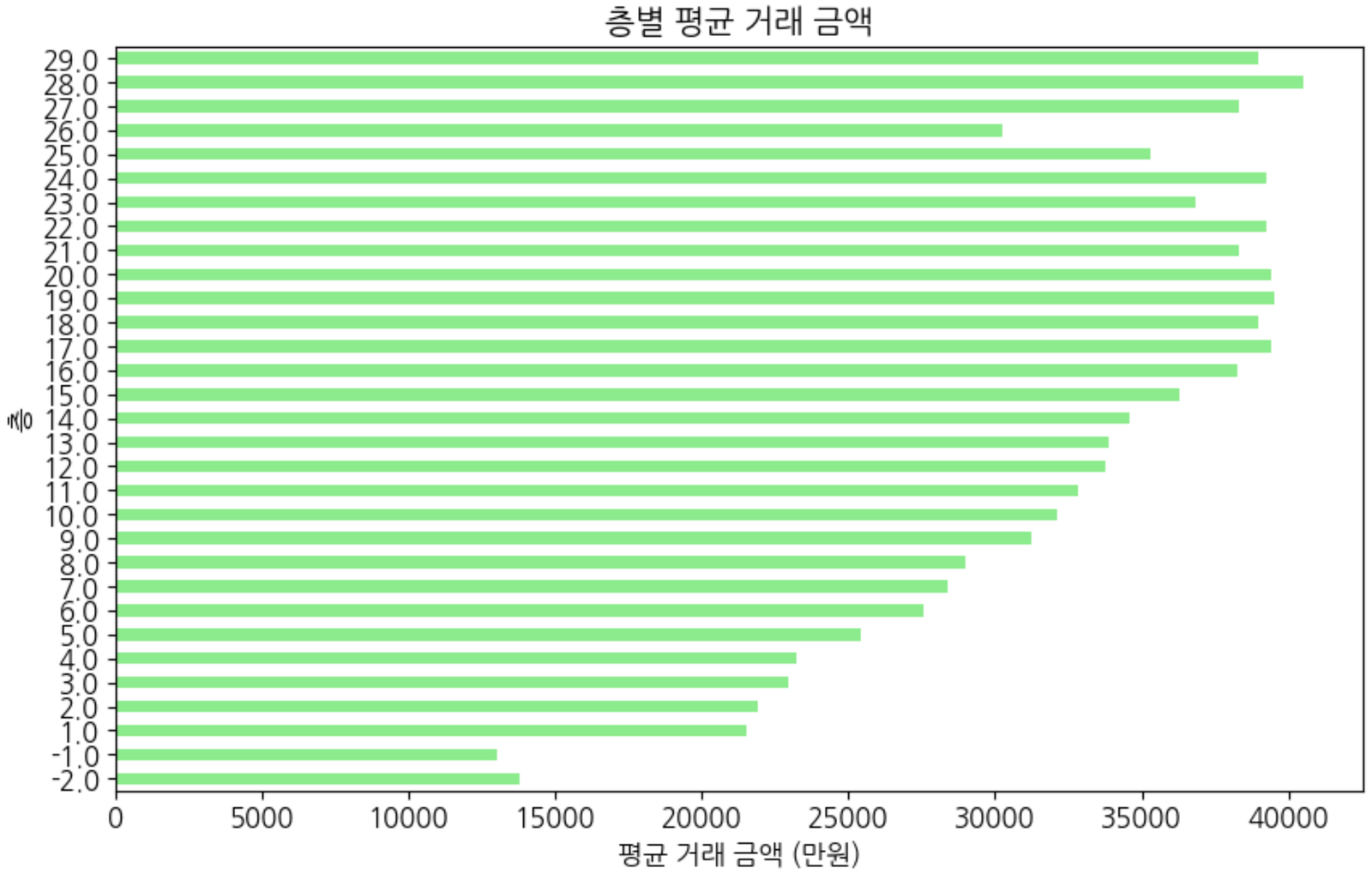
(차트 6)
건물 용도별 거래 금액 분포



- **아파트**가 다른 건물 용도(연립다세대, 오피스텔)에 비해 **평균 거래 금액**이 **높음**.
- **연립다세대**와 **오피스텔**은 거래 금액의 **분산이 큼**.

Part 1.
데이터 분석

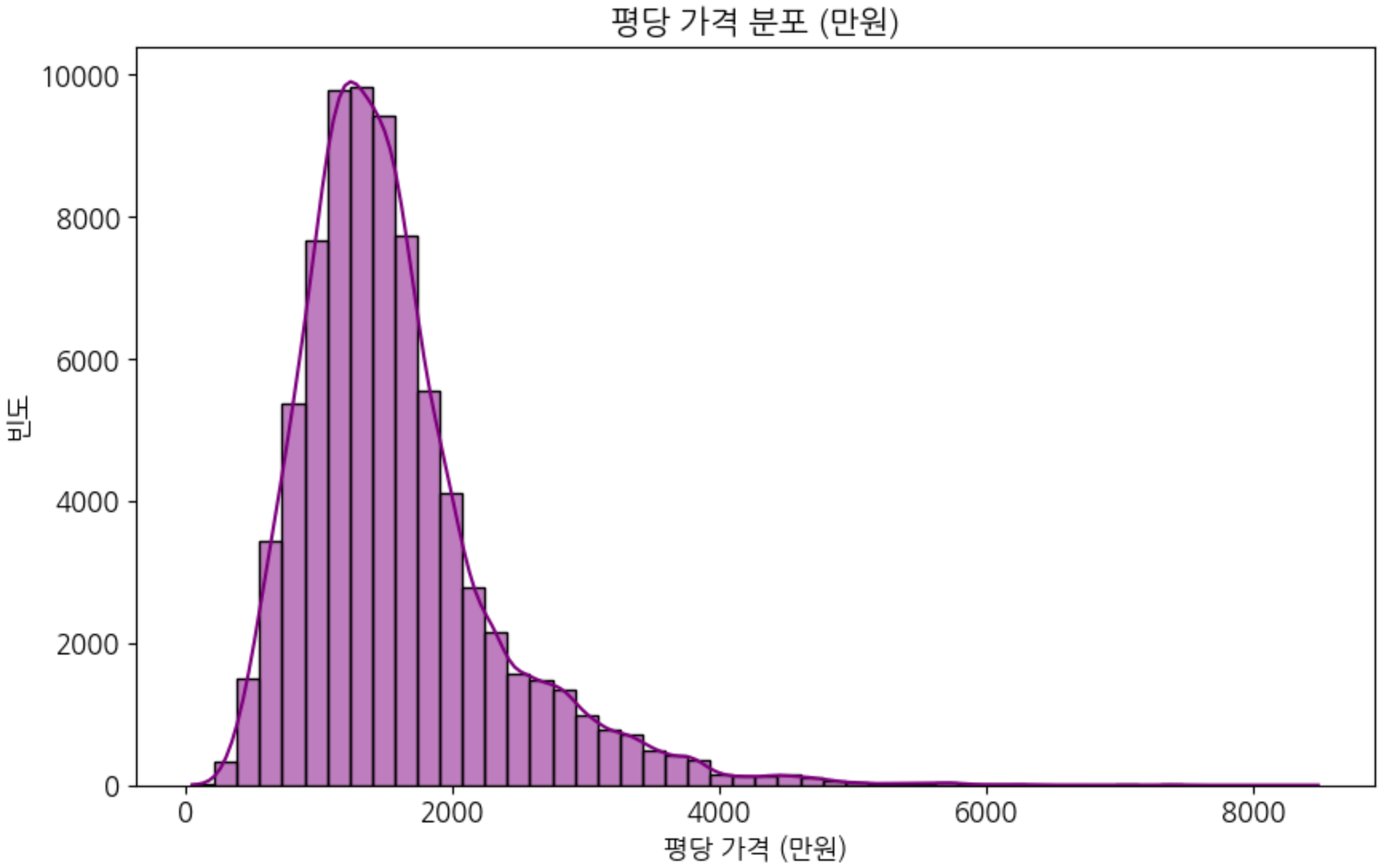
(차트 7)
층별 평균 거래 금액 분석



- **고층(25층 이상)**에서 거래 금액이 상대적으로 **높음**.
- 저층보다는 **중층 이상**에서 **더 높은 거래 금액**을 형성.

Part 1.
데이터 분석

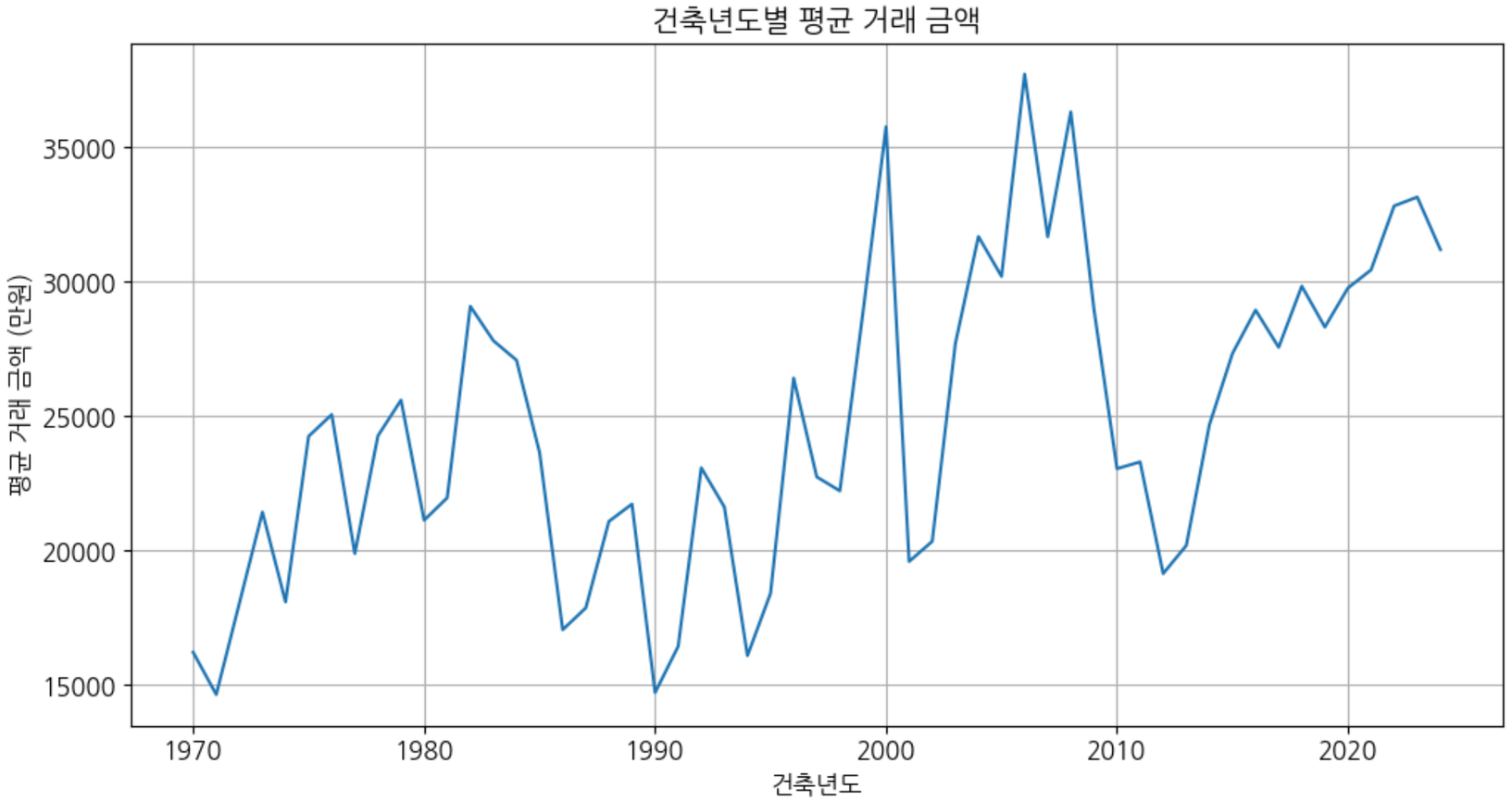
(차트 8)
평당 가격 분포 분석



- **평당 1,500만원 ~ 2,000만원 구간**에서 거래가 가장 **활발**.
- **고가 평당** 거래(5,000만원 이상)는 **거의 드뭄**.

Part 1.
데이터 분석

(차트 9)
건축년도별 평균 거래 금액 분석



- **2000년대 이전** 건물은 평균 거래 금액이 상대적으로 **낮은 편**.
- **2000년대** 건물은 평균 거래 금액이 상대적으로 **높음**.
- **2010년대 이후** 건축된 건물도 **높은 평균 거래 금액**을 유지 중.

Part 1.
데이터 분석

결론

주요 분석 결과:

- 부동산 거래는 주로 건물 면적, 층수, 건축 연도, 그리고 건물의 용도에 따라 크게 달라짐.
- 시장은 봄철에 큰 활동을 보임.
- 아파트 유형이 다른 건물 유형보다 높은 가치를 지니고 있음.
- 고층 건물은 전망과 프라이버시 때문에 더 높은 거래 가격을 형성함.
- 최근에 건축된 건물일수록 거래 금액이 높음.
- 평당 가격 분석 결과, 중간 가격대(1,500만 원에서 2,000만 원 사이)에서 거래가 가장 활발하게 이루어지는 것으로 나타남.

총평: 전체 분석 결과는 일반적인 부동산 시장의 동향과 일치하며, 특별한 사례는 발견되지 않았음.

Part 2. 데이터 모델링

Part 2.
데이터 모델링

데이터 전처리 과정

데이터 전처리 개요:
데이터 모델링을 위한 데이터 품질 개선을 통해 모델의 성능을 높이도록 함.

결측치 처리 및 열 제거 → 범주형 변수 인코딩 → 이상치 제거 → 데이터 정규화 및 타겟 설정

Part 2.
데이터 모델링

데이터 전처리 과정

결측치 처리 및 열 제거 | 이상치 제거 | 범주형 변수 인코딩 | 데이터 정규화 및 타겟 설정

In [72]:
```
# 0. 불필요한 열 제거
# 분석에 불필요한 열인 '법정동코드', '일', '건물명'은 제거
columns_to_drop = ['법정동코드','건물명','일']
df2.drop(columns=columns_to_drop, axis=1, inplace=True)
```

In [73]:
```
# 1. 결측치 처리
# 수치형 변수 결측치는 중앙값으로 대체하고, 범주형 변수 결측치는 최빈값으로 대체
numeric_features = ['건물면적(㎡)', '토지면적(㎡)', '층', '건축년도']
categorical_features = ['건물용도']

numeric_imputer = SimpleImputer(strategy='median')
categorical_imputer = SimpleImputer(strategy='most_frequent')
```

In [74]:
```
# 수치형 변수 결측치 대체
df2[numeric_features] = numeric_imputer.fit_transform(df2[numeric_features])
# 범주형 변수 결측치 대체
df2[categorical_features] = categorical_imputer.fit_transform(df2[categorical_features])
```

- 열 제거: 분석에 불필요한 열인 '법정동코드', '일', '건물명'은 제거
- 결측 데이터 처리: 수치형 변수 결측치는 중앙값으로 대체하고, 범주형 변수 결측치는 최빈값으로 대체

Part 2.
데이터 모델링

데이터 전처리 과정

결측치 처리 및 열 제거 | **이상치 제거** | 범주형 변수 인코딩 | 데이터 정규화 및 타겟 설정

In [75]:
```python
# 2. 이상치 처리
# IQR 방식을 사용해 이상치 제거
def remove_outliers(df2, column):
    Q1 = df2[column].quantile(0.25)
    Q3 = df2[column].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return df2[(df2[column] >= lower_bound) & (df2[column] <= upper_bound)]
```

In [76]:
```python
# 거래 금액과 건물면적에서 이상치 제거
df2 = remove_outliers(df2, '물건금액(만원)')
df2 = remove_outliers(df2, '건물면적(㎡)')
```

- 이상치 탐지 및 제거: IQR 방식을 사용해 ‘거래 금액’ 과 ‘건물면적’ 의 비정상적인 데이터 포인트 제거.

Part 2.
데이터 모델링

데이터 전처리 과정

결측치 처리 및 열 제거 → 이상치 제거 → **범주형 변수 인코딩** → 데이터 정규화 및 타겟 설정

```
In [78]: # 4. 범주형 변수 인코딩
         # '건물용도', '법정동명', '지번구분명'을 OneHotEncoding으로 변환
         encoder = OneHotEncoder()
         encoded_features = encoder.fit_transform(df2[['건물용도', '법정동명', '지번구분명']]).toarray()
         encoded_feature_names = encoder.get_feature_names_out(['건물용도', '법정동명', '지번구분명'])
         encoded_df = pd.DataFrame(encoded_features, columns=encoded_feature_names)
```

```
In [79]: # 기존 데이터프레임에 추가하고 '건물용도', '법정동명', '지번구분명' 열 삭제
         df2 = pd.concat([df2, encoded_df], axis=1)
         df2.drop(['건물용도', '법정동명', '지번구분명'], axis=1, inplace=True)
```

- 범주형 변수 인코딩: ‘건물 용도', ‘법정동명', ‘지번구분명')는 One-Hot Encoding을 통해 수치형으로 변환.

Part 2.
데이터 모델링

데이터 전처리 과정

결측치 처리 및 열 제거 → 이상치 제거 → 범주형 변수 인코딩 → **데이터 정규화 및 타겟 설정**

In [80]:
```
# 5. 데이터 정규화
# '건물면적(㎡)'와 '토지면적(㎡)' 등의 수치형 데이터를 정규화
scaler = StandardScaler()
numeric_scaled = scaler.fit_transform(df2[numeric_features])
scaled_df = pd.DataFrame(numeric_scaled, columns=numeric_features)
```

In [81]:
```
# 기존 데이터프레임에 정규화된 값 반영
df2[numeric_features] = scaled_df
```

- 데이터 정규화: ‘건물면적' 과 ‘토지면적’ 등의 수치형 데이터를 정규화.

```
# 6. 타겟과 피처 설정
features = df2.drop(['물건금액(만원)','계약일'], axis=1)
target = df2['물건금액(만원)']
```

- 최종으로 ‘물건금액(만원)’ 을 타겟 변수로 설정

Part 2.
데이터 모델링

모델링을 위해
전처리 완료된 데이터

모델링을 위해 준비된 피처와 타겟 데이터

In [97]:
```
# 전처리된 데이터 확인
print(features.head())
print(target.head())
```

```
    건물면적(㎡)    토지면적(㎡)         층       건축년도       연도    월   분기   요일    건물면적(평)  \
0 -0.252120  0.282069 -0.830766 -1.381709  2024.0  10.0  4.0  1.0  14.890909
1 -0.669029 -0.730008 -0.639389  1.684247  2024.0  10.0  4.0  1.0  11.896970
2 -1.235317  0.149579 -0.448013  0.151269  2024.0  10.0  4.0  0.0   7.830303
3 -1.177085 -1.926099 -0.256636 -0.615220  2024.0  10.0  4.0  5.0   8.248485
4 -0.767349  1.387074 -0.065260  0.589263  2024.0  10.0  4.0  5.0  11.190909

     평당 가격(만원)  건물용도_아파트  건물용도_연립다세대  건물용도_오피스텔  법정동명_남현동  법정동명_봉천동  법정동명_상도동  \
0  1490.842491       0.0          1.0         0.0         0.0         1.0         0.0
1  3404.228222       0.0          1.0         0.0         0.0         0.0         0.0
2  1277.089783       0.0          0.0         1.0         0.0         1.0         0.0
3  2024.614254       0.0          1.0         0.0         0.0         0.0         0.0
4  2162.469537       0.0          0.0         1.0         0.0         1.0         0.0

   법정동명_신림동  지번구분명_대지  지번구분명_블럭  지번구분명_산
0         0.0         1.0         0.0         0.0
1         1.0         1.0         0.0         0.0
2         0.0         1.0         0.0         0.0
3         1.0         1.0         0.0         0.0
4         0.0         1.0         0.0         0.0
0    22200.0
1    40500.0
2    10000.0
3    16700.0
4    24200.0
Name: 물건금액(만원), dtype: float64
```

Part 2.
데이터 모델링

모델링 과정

사용 모델:
랜덤 포레스트 회귀 (Random Forest Regressor)

랜덤 포레스트 회귀 모델을 사용하여 부동산 가격 예측.

데이터 분리:
Training(학습) 데이터와 Testing(테스트) 데이터로 분리하여 모델의 성능 평가

비율: 학습 데이터 70%, 테스트 데이터 30%.

교차 검증 (Cross Validation)

5-Fold 교차 검증으로 모델의 일반화 성능 확인.

Part 2.
데이터 모델링

모델링 절차

사용 모델

랜덤 포레스트 회귀
(Random Forest Regressor)

01

데이터 분리:
학습 데이터 70%,
테스트 데이터 30%

02

모델 학습

03

예측 및 평가

Part 2.
데이터 모델링

모델 평가 결과 분석

67.45

Mean Absolute Error (MAE)

0.9998

R^2 Score

- **MAE**는 **예측값과 실제값** 사이의 **절대 오차의 평균**을 의미.
- 예측값이 실제값에 비해 평균적으로 **67.45만 원 정도의 차이**를 보임을 나타냄.
- 부동산 가격의 단위를 고려했을 때, 비교적 **작은 오차 값**이며 모델이 대부분의 데이터를 잘 학습하고 있다는 것을 보여줌.

- **R^2 값**이 매우 높음.
- 이는 모델이 타겟 변수의 변동성을 **99.98%**까지 설명할 수 있다는 것을 의미하며, **매우 높은 적합도**를 나타냄.

Part 2.
데이터 모델링

모델 검증 및 안정성 평가

01. 교차 검증 (Cross-Validation)

교차 검증 결과:

Cross-Validation R^2 Scores: [0.9739, 0.9995, 0.9995, 0.9991, 0.9962]

Mean Cross-Validation R^2 Score: 0.9936

개별 Fold의 R^2 Scores:

- 각 Fold의 R^2 값은 0.9739에서 0.9995 사이.
- 대부분의 Fold에서 0.99 이상을 기록, 모델이 일관되게 높은 성능을 유지하고 있음을 의미함.

Mean Cross-Validation R^2 Score:

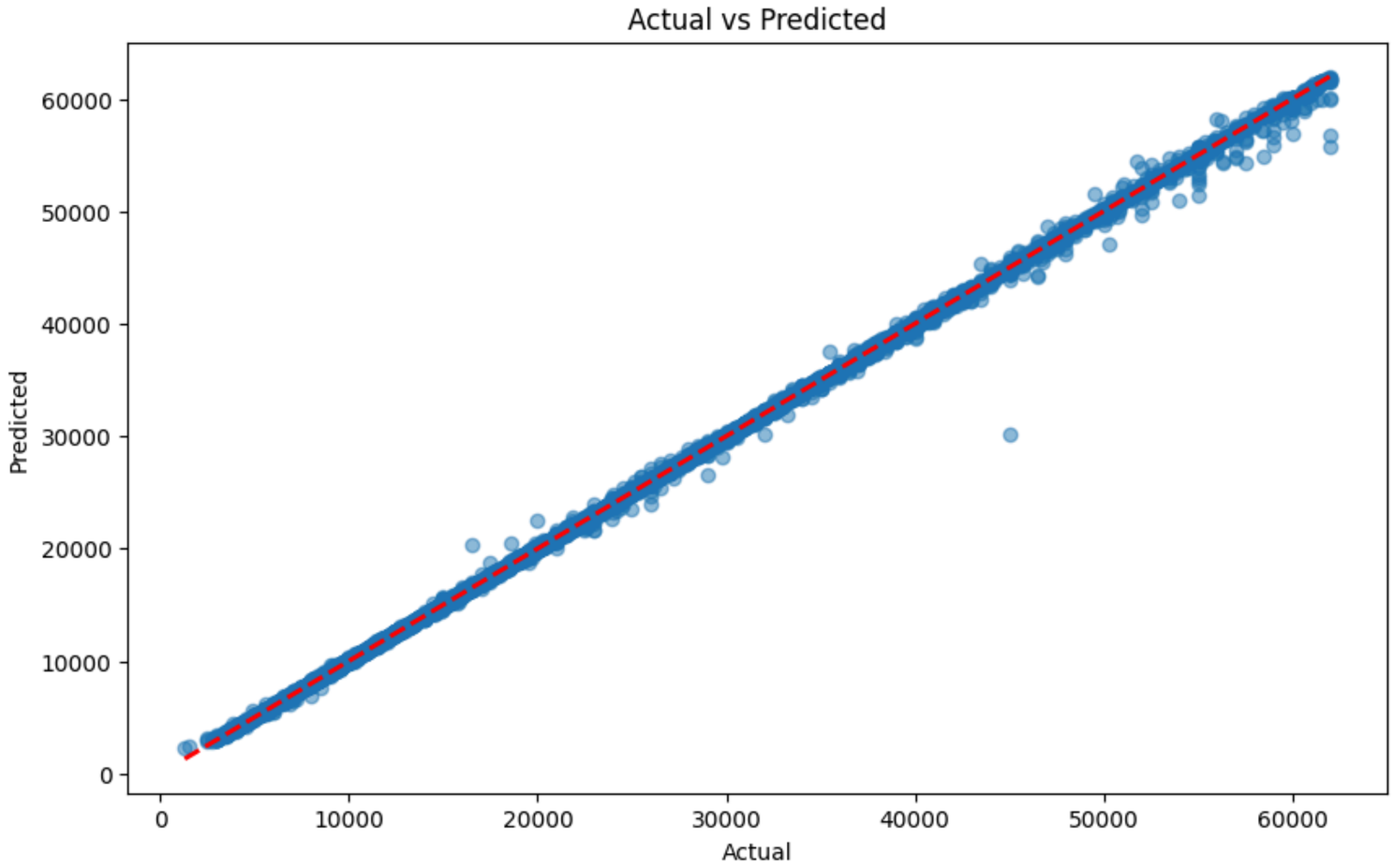
- 평균 R^2 값은 0.9936으로 매우 높음.
- 모델이 데이터 전체에 대해 잘 일반화되고 있으며, 데이터 분할의 차이에 큰 영향을 받지 않음을 의미함.

Part 2.
데이터 모델링

모델 검증 및 안정성 평가

02. 예측 값과 실제 값 비교

점도 및 그래프를 통해 모델의 예측이 얼마나 실제와 일치하는지 평가



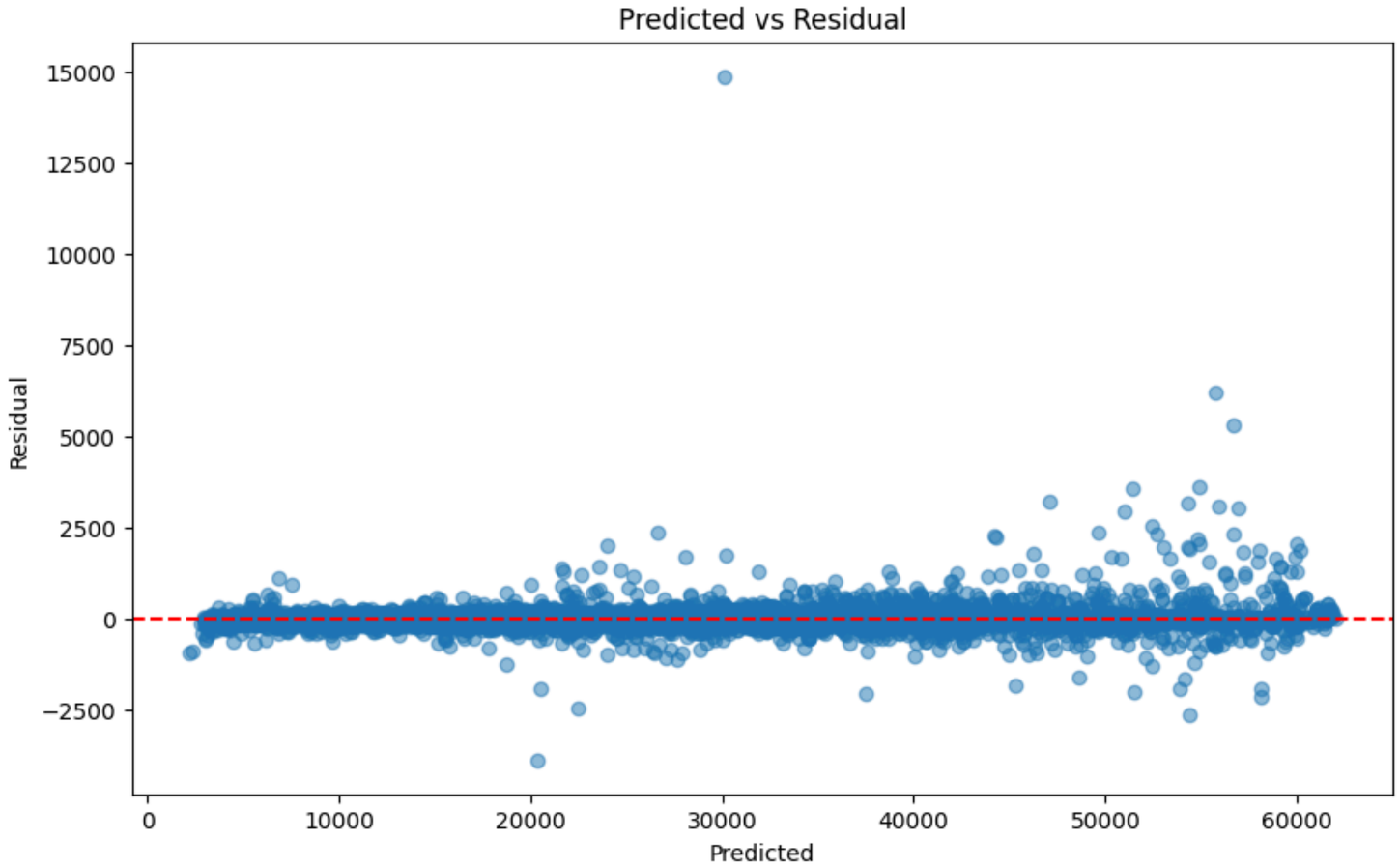
- 대부분의 데이터 포인트가 대각선 근처에 위치하여, 모델이 실제 값에 대해 정확하게 예측하고 있음을 의미함.

Part 2.
데이터 모델링

모델 검증 및 안정성 평가

03. 잔차(residual) 분석

잔차 분포와 예측 값 대비 잔차 그래프를 통해 모델의 편향 여부를 평가



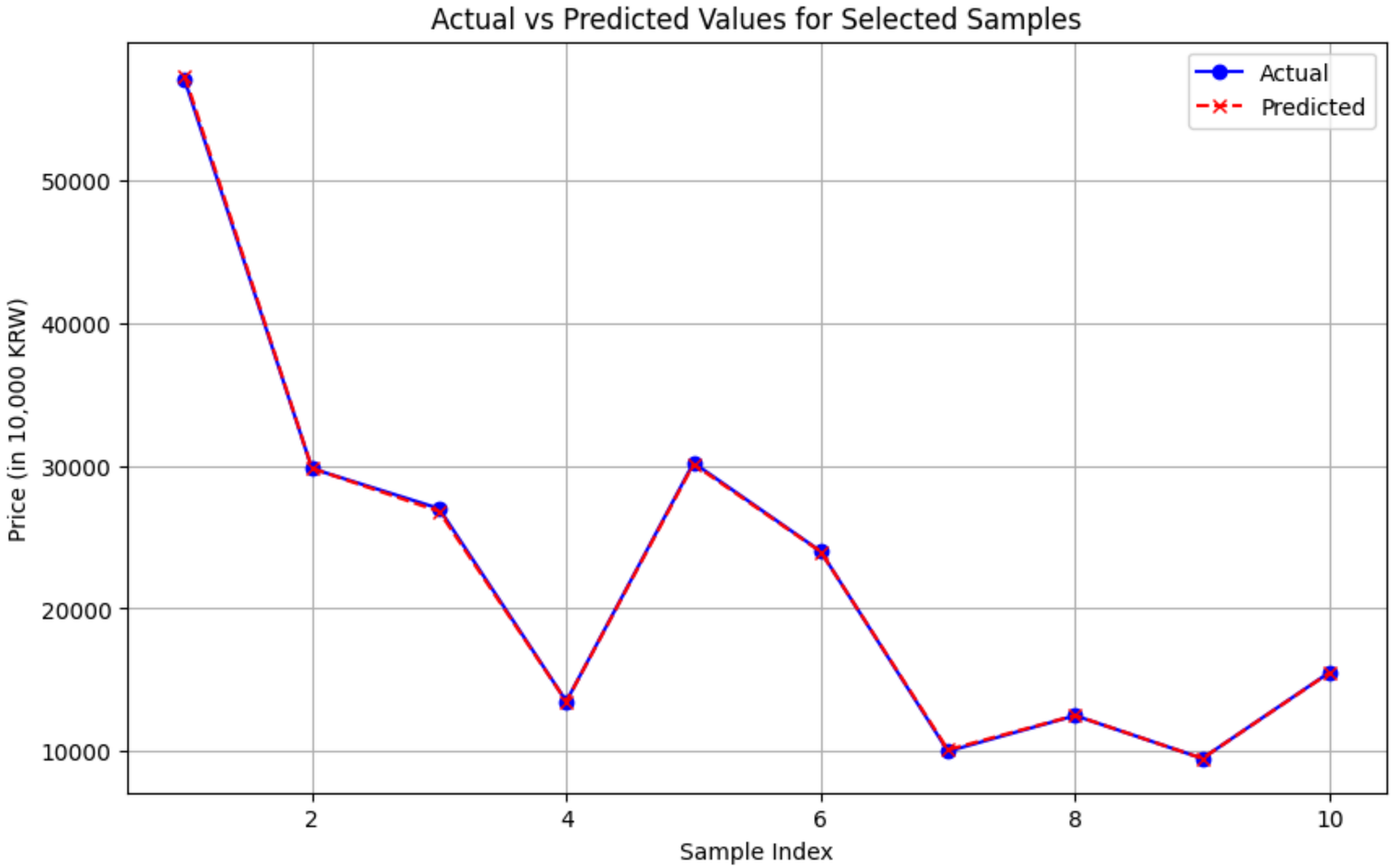
- 매우 낮은 잔차를 보이고 있어, 전반적으로 매우 좋은 성능을 가지고 있다고 평가할 수 있음.
- 잔차가 0에 집중되어 있고, 대체로 무작위로 분포하고 있어 모델이 과적합되지 않고 데이터의 일반적인 패턴을 잘 학습한 것으로 보임.

Part 2.
데이터 모델링

모델 검증 및 안정성 평가

04. 특정 샘플에 대한 예측

일부 샘플에 대해 실제 값과 예측 값을 비교하여 개별 샘플 수준에서 평가



- 실제 값과 예측 값이 거의 동일하게 따라가고 있으며, 모델이 대부분의 샘플에서 정확한 예측을 하고 있다는 것을 보여줌.

Part 2.
데이터 모델링

모델 검증 및 안정성 평가

05. 성능 지표 계산 (테스트 데이터 기준)

MAE, MSE, RMSE 등의 다양한 지표를 계산하여 모델의 예측 정확도를 객관적으로 평가

38,806.76	196.99	67.45
Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)

- **MAE와 RMSE**의 값이 실제 부동산 가격 대비 **매우 작은 값**이라는 점에서, 모델은 전반적으로 **우수한 예측 성능**을 보이고 있음.
- **MSE**가 상대적으로 **높게** 나왔다는 것은 일부 **특정 샘플**에서의 **큰 오차**가 있었음을 의미. 이는 모델이 평균적으로는 잘 예측하지만 **특이값에 대해서**는 **덜 학습**했을 가능성을 시사

Part 2.
데이터 모델링

모델 검증 및 안정성 평가

- 모델의 전반적인 성능이 우수하며, 테스트 데이터와 교차 검증에서도 높은 일관성을 보여줌.

- 현재 상태로도 모델은 실제 사용에 적합한 수준의 성능을 보여주고 있음.

미래 예측

관악구 부동산의 향후 10년간의 미래 가격은 어떻게 될까?

Part 2.
데이터 모델링

미래 예측 전제

시나리오 01 - 봉천동 아파트

- 면적: 50 ~ 150 평방미터 (15 ~ 45 평)
- 평당 가격: 1,000 ~ 4,000 만원
- 특징: 관악구 봉천동 위치한 아파트

시나리오 02 - 남현동 연립다세대

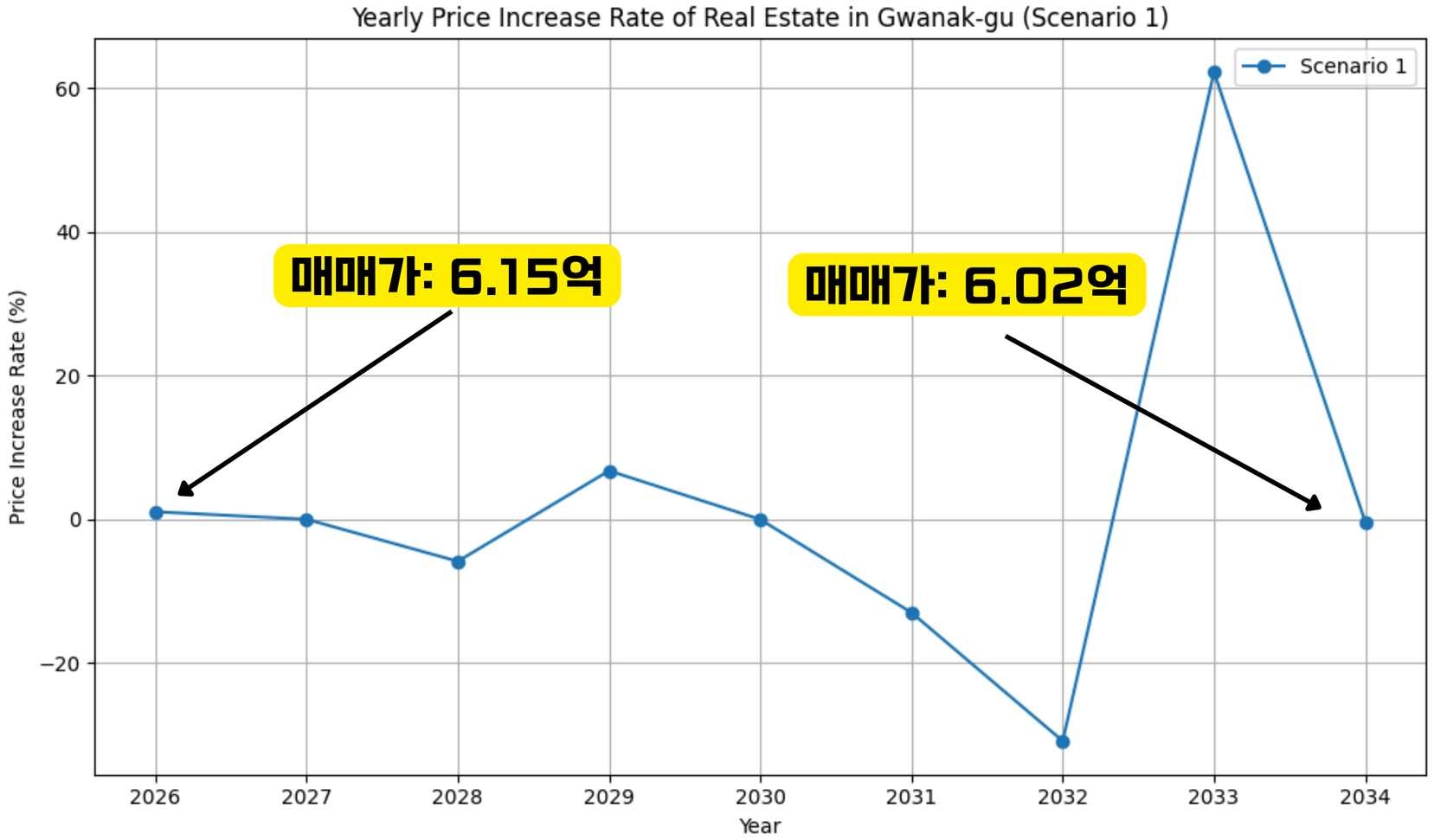
- 면적: 100 ~ 200 평방미터 (30 ~ 60 평)
- 평당 가격: 1,500 ~ 5,000 만원
- 특징: 관악구 남현동에 위치한 연립다세대

시나리오 03 - 신림동 오피스텔

- 면적: 30 ~ 85 평방미터 (10 ~ 35 평)
- 평당 가격: 800 ~ 3,000 만원
- 특징: 관악구 신림동에 위치한 오피스텔

Part 2.
데이터 모델링

시나리오 01-
봉천동 아파트

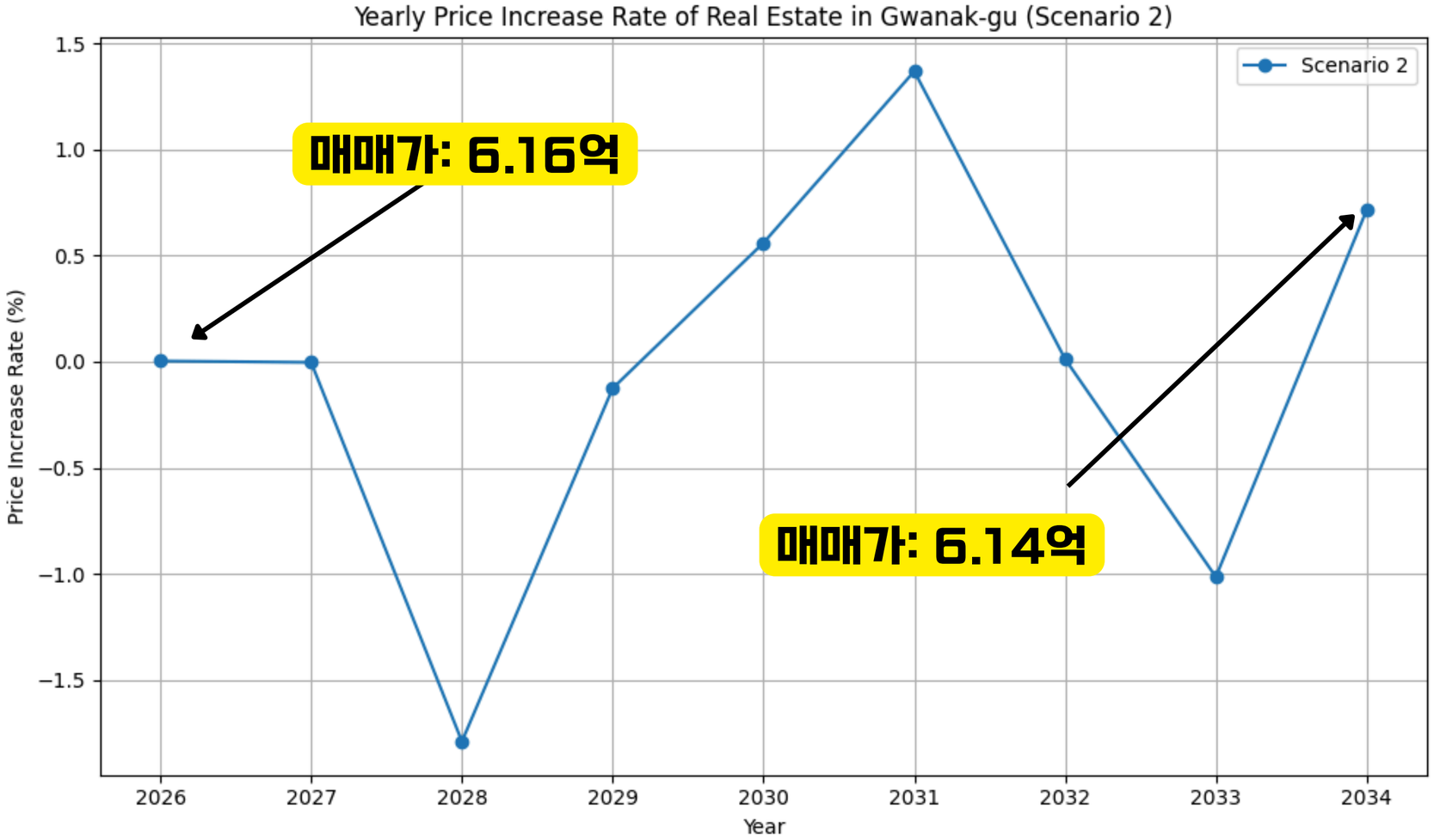


주요 특징:

- 2028년부터 2031년까지 **가격 하락세**를 보임.
- **2031년~ 2032년 구간**에 **30.8%의 급격한 하락**이 나타남.
- **2033년**에 **62.2%의 큰 폭의 가격 상승** 후, 2034년에 소폭 감소.

Part 2.
데이터 모델링

시나리오 02-
남현동 연립다세대

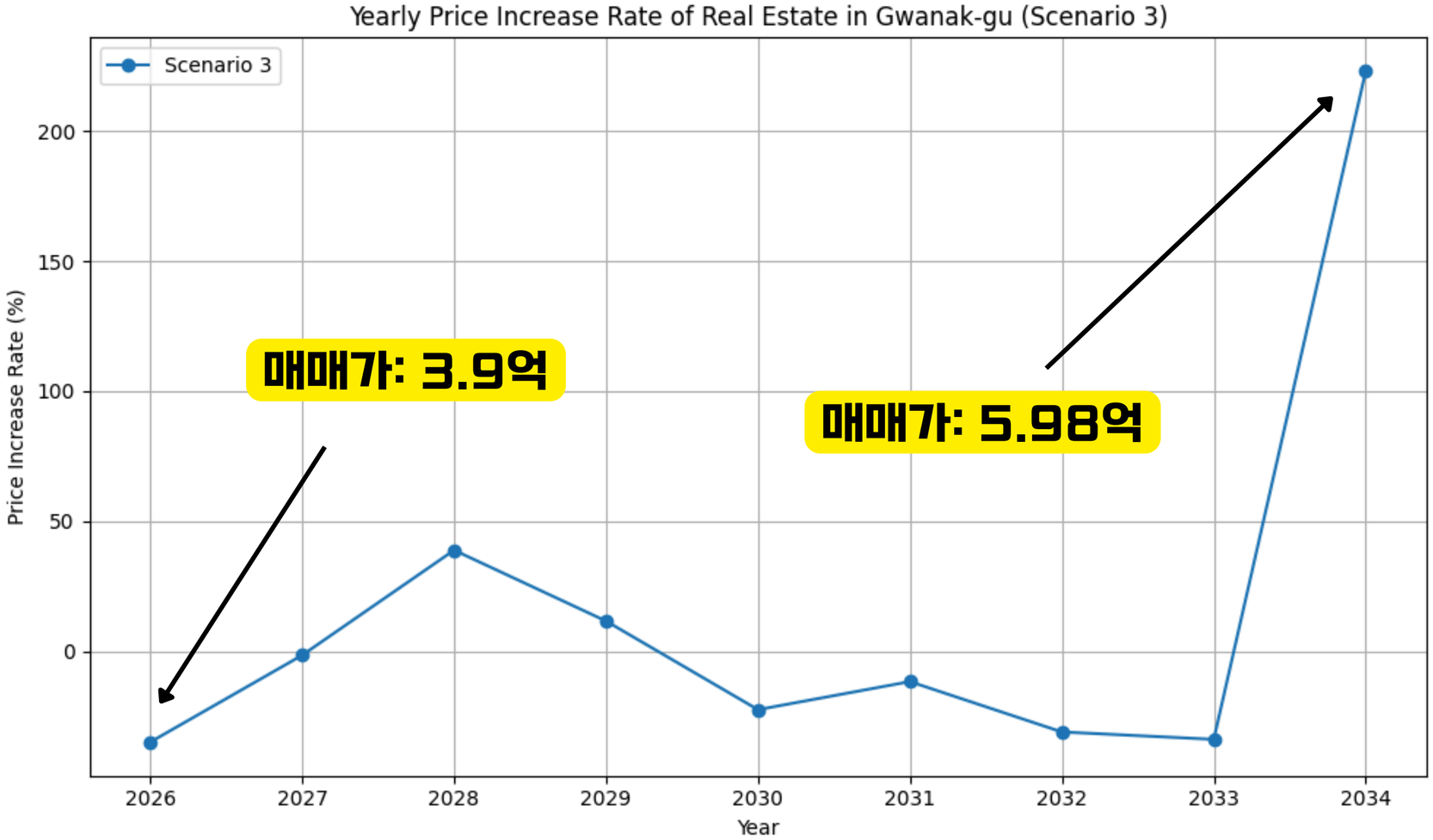


주요 특징:

- **2028년**에 **-1.79%**의 **하락**, **2031년**에 **1.36%의 상승**을 보이는 등, 전반적으로 **안정적인 가격** 변화.

Part 2.
데이터 모델링

시나리오 03-
신림동 오피스텔



주요 특징:

- 초기 몇 년 동안 **상당한 가격 변동**있음.
- **2031년 ~2033년** 구간에 **31.1%, 33.9%**의 **큰폭의 하락**을 기록함.
- **2034년**에 **223.1%**의 **큰 폭의 상승**을 보임.

Part 2.
데이터 모델링

미래 예측 결론

관악구 부동산의 향후 10년 미래 가격 예측 결론

Scenario 1 (중간 크기 아파트, 봉천동)

상승과 하락을 반복하는 큰 변동성을 보여줌. 특히 2033년에 62.2%의 급격한 상승이 관찰되었음. 이는 시장의 외부 요인에 의한 변동 가능성을 시사함.

Scenario 2 (고급 연립다세대, 남현동)

대부분 0%에 가까운 변화로 안정적인 시장 유지. 안정적인 가격 변동을 보이며, 시장의 불확실성에 크게 영향을 받지 않는 모습.

Scenario 3 (저가형 오피스텔, 신림동)

초기 몇 년간 가격 하락세가 강했으나, 2034년에는 223.1%의 큰 상승을 기록하며, 큰 폭의 하락과 급격한 상승을 보여줌. 부동산 관련 정책 변화가 시장에 큰 영향을 준 결과로 해석될 수 있음.

향후 계획

향후 계획

해보면 좋은 것들...

1: 데이터 확장 및 모델 개선

- 다른 자치구와의 비교 분석, 경제 데이터와의 연동.
- 새로운 모델 적용 및 하이퍼파라미터 튜닝 고도화.

2: 상호작용형 분석 도구 개발

- 대시보드와 웹 기반 시각화 도구 개발.
- 예측 시나리오를 사용자가 직접 실험할 수 있도록 지원.

3: 장기 예측 및 미래 시나리오

- 20~30년 장기 예측 및 정책 변화에 따른 시나리오별 분석.

감사합니다

Thank you