## 머신러닝(XGBoost)을 활용한

# 서울특별시 부동산 시세 상승 원인 분석


김 희 만 : 팀장 (역할 분담, 데이터 수집, 데이터 전처리, 결론 도출)

김 선 오 : (데이터 수집, 데이터 전처리, 데이터 시각화)

이 용 기 : (데이터수집, 데이터 시각화, 결론 도출)

김 지 수 : (데이터수집)



## 더조은 아이티 아카데미

# 요 약 서 (초 록)

| | |
|---|---|
| 과 제 명 | 부동산 시세 예측 모델을 통해 집값 변동에 대한 주요 원인 분석 |
| 소 속 | 더조은 아이티 아카데미 |
| 연구원 | 김선오, 김지수, 김희만, 이용기 |
| 연구기간 | 2024. 11. 19. ~ 2024. 12. 06 |
| Key Word | **집값 상승, 집값 , 부동산 투자, 부동산 시세, 집값 변동** |


## 1. 연구의 필요성 및 목적

- o 2023년도 집값 상승에 대한 원인과 분석을 토대로 대책을 마련하기 위한 필요성
- o 부동산 시세 예측 모델 구축을 통해 집값 변동에 대한 원인 분석

## 2. 연구 내용 및 범위

- o 2015~2023년 부동산 실거래가와 자치구 데이터 등 여러 데이터들을 활용하여 집값 변동 분석 및 모델링 구현
- o 학습된 모델을 통해 Feature Important를 확인하여 집값 상승에 대한 주요 원인 분석
- o 정확한 데이터 추출을 위해 ‘서울특별시 행정 자치구’별 연간 주택 매매 실거래 데이터를 기준 범위로 선정

## 3. 연구 방법

- o 공공 데이터 기관을 통해 데이터 추출 및 머신러닝 학습을 위해 전처리 작업 진행
- o 행정 자치구별 연간 평균 부동산 시세를 예측하기 위해 회귀모델(Random Forest, XGBoost 등) 사용

## 4. 결론

- o 모델 평가지표인 MSE(Mean Squared Error) : 0.39 / R2 Score : 0.85 로 준수한 수준의 모델을 구축함

# 목 차

## 제 Ⅳ 장 데이터 분석 및 시각화



## 제 Ⅴ 장 머신러닝 학습



## 제 Ⅴ 장 결론 및 제언

# <그림목차>

# I

# 서 론

1. 연구 필요성 및 목적
2. 연구방법
3. 연구추진절차
4. 연구의 제한점

# Ⅰ 서론

## 1. 연구 필요성 및 목적

부동산이란, 토지와 그것에 정착된 건물이나 수목 등 움직이거나 옮길 수 없는 재산이다. 그러나 그 재산의 가치는 사회적 이슈, 법률 개정, 당시 소비자의 심리 상태 등 여러 가지 요인들로 인해 매우 유동적으로 변한다.

부동산은 크게 토지와 토지 정착물로 나뉘고 토지 정착물에는 건물, 등기한 입목, 명인방법을 갖춘 수목의 집단, 명인방법을 갖춘 미분리 과실, 농작물 등이 있다. 부동산 매매의 주된 대상은 토지와 건물이며 그중 가장 쉽게 접할 수 있는 것이 주택(아파트, 단독, 오피스텔, 등) 매매이다.

주택 매매의 가장 큰 목적은 투자이다. 단순 내 집 마련이라고 하더라도 대부분의 소비자는 저렴한 금액에 사고, 나중에라도 비싼 금액에 되팔려고 하거나 다소 비싸더라도 나중에 금액이 더 오를 것이라 기대하고 거래하기 때문에 투자라고 볼 수 있다.

그러나 부동산 시세는 예측하기가 쉽지 않다. 여러 정보가 복합적으로 반영돼 유기적인 시세 변동을 보이며 특히 2023년 서울특별시의 부동산 평균 거래 시세가 급상승하는 통계를 보고 안정적인 부동산 투자를 위해 부동산 시세 변동의 주요 원인이 되는 것이 무엇인지 머신러닝 학습을 통해 관심을 가져야 할 정보가 무엇인지 제시하고자 한다.

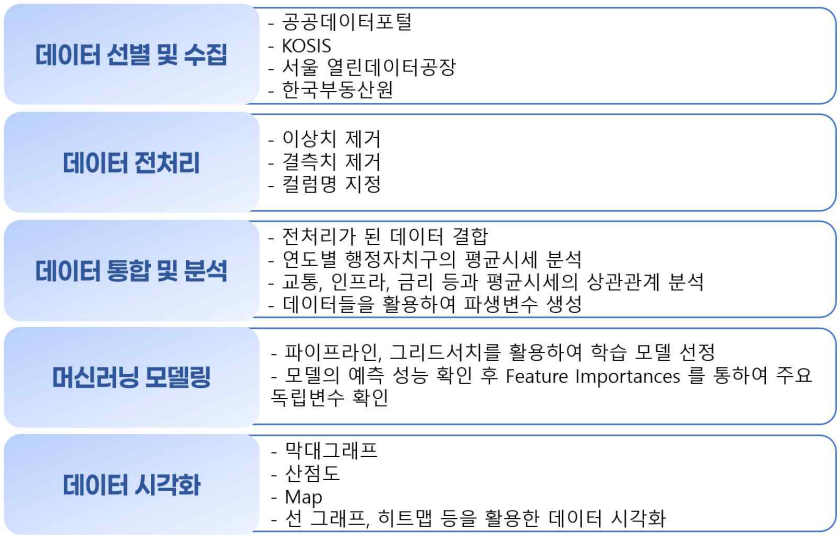
## 2. 연구 방법

○ 공공 데이터 기관 (공공데이터포털, KOSIS, 서울 열린 데이터 광장, 한국부동산원 등)에서 공개된 데이터를 이용해서 파이썬 라이브러리의 pandas를 활용하여 데이터 분석

○ 전처리된 데이터를 이용하여 행정 자치구별 연간 평균 부동산 매매 실거래가, 교통, 인프라, 총인구수 등 기타 자료들을 정리 및 시각화

○ 머신러닝 알고리즘 중 회귀모델인 XGBoost를 기반으로 평균 시세를 예측하고, 독립변수 중요도를 통해 부동산 투자 시 중점적으로 고려해야 할 핵심 정보 도출

## 3. 연구추진절차




## 4. 연구의 제한점

**첫째,** 본 영구는 서울특별시의 행정 자치구별 평균 시세 변동에 대한 부동산 투자 시 중점적으로 고려해야 할 핵심 정보 도출에 의의가 있으며, 향후 주택 매매 투자 시 참고가 될 만한 자료들을 제시한다.

**둘째,** 주택 매매 시세 예측을 위해 2015년 ~ 2023년 자료들을 교통수단, 인프라(병원, 공원), 개발계획, 인구수, 유통업체 등에 따라 조사 및 분석하고 그 결과를 토대로 평균 시세를 예측하기 위해 회귀모델을 사용하며, 모델에 적용 할 수 있는 수치형 데이터를 기준으로 진행

**셋째,** 주 2015 ~ 2023년 자료 외에도 2024년 자료도 몇몇 있으나 아직 모든 자료가 최신화 되어있지 않으므로 2015년부터 2023년 자료를 기준으로 분석을 실시함.

# Ⅱ

# 데이터 수집

1. 데이터 분석 범위
2. 각종 데이터 수집

# Ⅱ 데이터 수집

## 1. 데이터 분석 범위

미래의 주택 매매 시세 예측에는 다양한 변수가 많고, 거래 시세는 여러 변수들에 맞춰 유동적으로 변하기에 정확한 금액을 예측하기 매우 까다롭다. 그럼에도 실제로 많은 투자와 부동산 거래가 이루어지고 있기에 방대한 데이터 중 '서울특별시의 행정 자치구'의 데이터로 한정하여 주택 거래 실거래가에 많이 반영되는 병원, 공원, 교통수단, 유통업체, 해당 행정 자치구의 인구수, 개발계획, 주택담보대출 금리, 부동산 실거래가 와 거래량의 2015년부터 2023년의 데이터를 대상으로 지정하였다.

## 2. 각종 데이터 수집

### 2-1) 공공데이터포털

○ **데이터 명 : 서울시 지역별 지하철역 정보_2015.csv ~ 서울시 지역별 지하철역 정보_2023.csv**

-데이터내용 : 2015년 ~ 2023년 각 행정 자치구 별 지하철 역 개수

-제공기관 : 서울교통공사

-저작권자 : 서울교통공사 영업계획처

-출처 :
https://www.data.go.kr/data/15081868/fileData.do#layer_data_infomation

### 2-2) KOSIS

○ **데이터 명 : 개발계획_2015.csv ~ 개발계획_2023.csv**

-데이터내용 : 2015년 ~ 2023년 각 행정 자치구 별 개발계획 / 개발면적

-제공기관 : KOSIS

-저작권자 : 한국국토정보공사

-출처 :
https://kosis.kr/statHtml/statHtml.do?orgId=460&tblId=TX_315_2009_H1011A&conn_path=I2

○ 데이터 명 : 지역별_매매동향_2015~2023.csv

-데이터내용 : 2015년 ~ 2023년 각 행정 자치구 별 수요와 공급의 비중을 점수화한 수치

-제공기관 : KOSIS

-저작권자 : 한국부동산원

-출처 :
https://kosis.kr/statHtml/statHtml.do?orgId=408&tblId=DT_40803_N0007&conn_path=I2

○ 데이터 명 : 주택의_종류별_주택__읍면동_연도_2015 ~ 2023.csv

-데이터내용 : 2015년 ~ 2023년 각 행정 자치구 별 주택의 종류와 종류별 개수

-제공기관 : KOSIS

-저작권자 : 통계청

-출처 :
https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1JU1501&conn_path=I2

○ 데이터 명 : 인구수_2015 ~ 2023.csv

-데이터내용 : 2015년 ~ 2023년 각 행정 자치구 별 총인구 수와 내국인, 외국인 수

-제공기관 : KOSIS

-저작권자 : 통계청

-출처 :
https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1IN1502&conn_path=I2

○ 데이터 명 : 매매가격지수_주택종합_2015~2023.csv

-데이터내용 : 기준시점과 매기 조사되는 조시시점의 가격비를 이용한 전월 대비 변동률

-제공기관 : KOSIS

-저작권자 : 한국부동산원

-출처 :
https://kosis.kr/statHtml/statHtml.do?orgId=408&tblId=DT_304N_04_00001&conn_path=I2

○ 데이터 명 : 주택담보금리.csv

-데이터내용 : 예금은행 대출금리(신규취급액 기준) - 주택담보대출

-제공기관 : KOSIS

-저작권자 : 한국은행

-출처 :
https://kosis.kr/statHtml/statHtml.do?orgId=301&tblId=DT_121Y006&conn_path=I2

○ 데이터 명 : 행정구역별_건축물거래현황_2015 ~ 202.csv

-데이터내용 : 2015년 ~ 2023년 각 행정 자치구 별 주택 거래량과 총 거래 면적

-제공기관 : KOSIS

-저작권자 : 통계청

-출처 : https://kosis.kr/statHtml/statHtml.do?orgId=408&tblId=DT_408_2006_S0028&vw_cd=MT_ZTITLE&list_id=408_31503_001&seqNo=&lang_

## 2-3) 서울 열린데이터광장

○ 데이터 명 : 부동산_실거래가_2015.csv ~ 부동산_실거래가_2023.csv

-데이터내용 : 2015년 ~ 2023년의 서울특별시 주택 매매 실거래가

-제공기관 : 서울특별시

-저작권자 : 서울특별시

-출처 : https://data.seoul.go.kr/dataList/OA-21275/S/1/datasetView.do

○ 데이터 명 : 공원현황_2015.csv ~ 공원현황_2023.csv

-데이터내용 : 2015년 ~ 2023년 각 행정 자치구 별 공원의 종류와 개수

-제공기관 : 서울특별시

-저작권자 : 서울특별시 푸른도시여가국 공원여가정책과

-출처 : https://stat.eseoul.go.kr:443/statHtml/statHtml.do?orgId=201&tblId=DT_201004_O090019&conn_path=I2

○ 데이터 명 : 버스_정류소현황(2019~2023년).xlsx

-데이터내용 : 2019년 ~ 2023년 각 행정 자치구 별 버스 정류장 이름 및 위치, 개수

-제공기관 : 서울특별시

-저작권자 : 서울특별시

-출처 : http://data.seoul.go.kr/dataList/OA-22193/F/1/datasetView.do

○ 데이터 명 : 유통업체현황_2015.xlsx ~ 유통업체현황_2023.xlsx

-데이터내용 : 2015년 ~ 2023년 각 행정 자치구 별 유통업체의 종류와 개수

-제공기관 : 서울특별시

-저작권자 : 서울특별시 노동공정상생정책관 공정경제담당관

-출처 : http://data.seoul.go.kr/dataList/10128/S/2/datasetView.do

○ **데이터 명 : 의료기관_2015.csv ~ 의료기관_2023.csv**

-데이터내용 : 2015년 ~ 2023년 각 행정 자치구 별 병원의 종류와 개수

-제공기관 : 서울특별시

-저작권자 : 건강보험심사평가원

-출처 : http://data.seoul.go.kr/dataList/173/S/2/datasetView.do

# Ⅲ

# 데이터 전처리

1. 데이터 필터링
2. 머신러닝용 데이터 결합
3. 이상치와 결측치 제거

# Ⅲ 데이터 전처리

## 1. 데이터 필터링

수집한 데이터를 따로 가공하지 않고 사용할 경우 데이터프레임의 컬럼 명이 이상하게 들어가 있거나, 필요가 없는 정보가 들어가서 문제가 발생할 수도 있다. 보통 이러한 데이터를 미가공 데이터(raw data) 라고 하며 작업 목적과 상황에 맞게 컬럼 명 지정, 필요한 데이터만 추출하는 등 데이터 필터링 작업이 필요하다.

```
def precleaning_seoul(df) :
  df = df.groupby(['접수연도', '자치구명'])['물건금액(만원)'].mean().reset_index()
  df = df.rename(columns = {'접수연도' : '연도', '자치구명' : '구', '물건금액(만원)' : '평균시세'}
  df['평균시세'] = round(df['평균시세'], 2)
  return df
```

[그림Ⅲ-1] 데이터 필터링 코드 예시

| | 접수연도 | 자치구코드 | 자치구명 | 법정동코드 | 법정동명 | 지번구분 | 지번구분명 | 본번 | 부번 | 건물명 | ... | 물건금액(만원) | 건물면적(㎡) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2015 | 11650 | 서초구 | 10800 | 서초동 | 1.0 | 대지 | 1337 | 14.0 | 이즈타워 | ... | 21900 | 33.25 |
| 1 | 2015 | 11215 | 광진구 | 10500 | 자양동 | 1.0 | 대지 | 0624 | 19.0 | (624-19) | ... | 12000 | 23.17 |

[그림Ⅲ-2] 데이터 필터링 전 예시

↓

| | 연도 | 구 | 평균시세 |
|---|---|---|---|
| 0 | 2015 | 강남구 | 82473.52 |
| 1 | 2015 | 강동구 | 42617.49 |
| 2 | 2015 | 강북구 | 27129.07 |
| 3 | 2015 | 강서구 | 30024.00 |
| 4 | 2015 | 관악구 | 33197.21 |

[그림Ⅲ-3] 데이터 필터링 후 예시

## 2. 머신러닝용 데이터 결합

여러 개의 데이터프레임으로 분리되어 있는 데이터들을 학습시키고 효율적인 데이터 관리를 하기 위해 한 개의 데이터프레임으로 병합시켜 주어야 한다. 먼저 연도별로 필터링된 개별 데이터들을 각 데이터의 '구' 컬럼을 기준으로 pandas의 merge()를 이용하여 병합시켜 준다. 즉, 2015년 버스 수, 역 개수, 유통업체 수, 평균 시세 등등 각각의 데이터를 merge_2015_df 변수에 병합시켜 주고, 해당 작업을 반복하여 merge_2015_df ~ merge_2023_df 까지 만들어준다.

```
dfs = [park_df, bus_df, train_df, develop_df,
       demandSupply_df, distribute_df, hospital_df,
       population_df, volume_df, abode_house_df, priceRelative_df,
       seoul_area_df]  # 필요한 데이터프레임 추가
result = house_df.copy()

### 데이터 병합
for df in dfs :
  # 병합
  result = result.merge(df, on='구', how='left')

result['연도'] = year
result = result.merge(rate_df, on = '연도', how = 'left')
return result
```

[그림Ⅲ-4] 데이터 병합 함수 예시

```
merge_year_2015 = merge_year_dataFrame(park_2015_df, bus_null_df, train_2015_df, house_2015_df,
                                       develop_2015_df, demandSupply_2015_df, distribute_2015_df,
                                       hospital_2015_df, population_2015_df, volume_2015_df,
                                       abode_house_2015_df, priceRelative_2015_df, seoul_area_df, rate_df,
                                       year = 2015)

merge_year_2016 = merge_year_dataFrame(park_2016_df, bus_null_df, train_null_df, house_2016_df,
                                       develop_2016_df, demandSupply_2016_df, distribute_2016_df,
                                       hospital_2016_df, population_2016_df, volume_2016_df,
                                       abode_house_2016_df, priceRelative_2016_df, seoul_area_df, rate_df,
                                       year = 2016)
```

[그림Ⅲ-5] 데이터 병합 함수 적용 예시

연도별로 병합된 총 9개의 데이터프레임을 다시 결합하여 1개의 데이터 프레임으로 만들어주기 위해 pandas 의 concat()을 이용하여 최종 데이터 프레임 1개를 만들어준다.

```
concat_list = [merge_year_2015, merge_year_2016, merge_year_2017, merge_year_2018,
               merge_year_2019, merge_year_2020, merge_year_2021, merge_year_2022,
               merge_year_2023]
ai_concat = pd.concat(concat_list, ignore_index = True)
```

[그림Ⅲ-6] 데이터 결합 예시

## 3. 이상치와 결측치 제거

이상치와 결측치는 머신러닝 학습에 치명적이다. 이상치(outlier)란, 쉽게 생각해서 이상한 수치 정도로 이해하면 된다. 어린이집 원생의 나이 데이터를 가져왔는데 20대가 있는 것처럼 정상 범주에서 벗어난 데이터를 이상치라고 한다. 이상치를 제거하지 않고 그대로 학습 할 때 이상치가 들어있는 데이터도 학습에 반영되기 때문에 예측값이 크게 달라질 수 있기 때문에 이상치 제거는 필수적인 작업이다.

결측치란, 데이터가 아무것도 들어있지 않은 상태를 의미한다. 가령 ‘ ’ 처럼 공백 같은 경우 이는 결측치가 아닌 ‘공백문자’가 들어간 object 타입 데이터가 된다. 결측치는 말 그대로 데이터가 없는 것이기 때문에 머신러닝 학습 시 큰 방해 요소가 되며 결측치가 있을 경우 학습 자체가 안되는 경우가 많으므로 이상치 제거와 함께 결측치 제거도 필수적인 작업이다.

### 가. 이상치 제거

```
# 1. Q1, Q3 계산
Q1 = ai_concat['평균시세(억)'].quantile(0.25)  # 1분위수
Q3 = ai_concat['평균시세(억)'].quantile(0.75)  # 3분위수
IQR = Q3 - Q1                          # IQR 계산

# 2. 이상치 기준 설정
lower_bound = Q1 - 1.5 * IQR           # 하한
upper_bound = Q3 + 1.5 * IQR           # 상한

# 3. 이상치 제거
df_filtered = ai_concat[(ai_concat['평균시세(억)'] >= lower_bound) & (ai_concat['평균시세(억)'] <= upper_bound)]
```

[그림Ⅲ-7] 이상치 제거 코드

이상치는 보통 상위 25%와 하위 25%의 값을 구하고 해당 값을 기준으로 IQR (InterQuantile Range)을 구한값에 1.5배만큼을 빼거나 더해서 하한값과 상한값 기준을 설정 해주고 그사이에 있는 값만 사용하도록 하여 이상치를 제거해 준다

## 나. 결측치 제거

```python
def fillna_with_neighbor_mean(df, column, year_column):
  # NaN이 있는 연도와 없는 연도 분리
  null_years = df.loc[df[column].isna(), year_column].unique()
  not_null_years = df.loc[~df[column].isna(), year_column].unique()

  # 연도별 평균값 계산
  year_means = df.groupby(year_column)[column].mean()

  # NaN 채우기
  for null_year in null_years:
    # null_year 보다 큰 not_null_year 중 가장 작은 연도
    next_year = not_null_years[not_null_years > null_year].min() if (not_null_years > null_year).any() else None
    # null_year 보다 작은 not_null_year 중 가장 큰 연도
    prev_year = not_null_years[not_null_years < null_year].max() if (not_null_years < null_year).any() else None

    # 평균값 계산
    if next_year is not None:
        fill_value = year_means[next_year]
    elif prev_year is not None:
        fill_value = year_means[prev_year]
    else:
        fill_value = np.nan  # 채울 수 없는 경우 NaN 유지

    # NaN 채우기
    df.loc[(df[year_column] == null_year) & (df[column].isna()), column] = fill_value

  return df
```

[그림Ⅲ-8] 결측치 제거 코드

결측치는 보통 0으로 채우거나 평균값으로 채우거나 데이터가 별로 없을 경우 drop 시켜 결측치가 있는 행을 없애버리며 데이터를 처리한다. 본 연구 같은 경우 데이터를 0 또는 drop으로 처리할 경우 오차가 심하게 발생하기에 연도별 가장 가까운 연도의 평균값으로 대체하도록 했다.

# IV

# 데이터 분석 및 시각화

1. 2015 ~ 2023년 부동산가격 변동 분석

# IV 데이터 분석 및 시각화

## 1. 2015 ~ 2023년 부동산가격 변동 분석

본격적으로 데이터분석과 시각화에 들어가기 전에 데이터를 다룰 때 사용되는 파이썬의 라이브러리 ‘Pandas’, ‘Numpy’, ‘seaborn’, ‘Matplotlib’, ‘wordcloud’과 지도 시각화 할 때 필요한 ‘json’, ‘folium’ 및 폰트 설정에 도움이 되는 ‘rc’도 추가로 사용한다.

```
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings('ignore')
import seaborn as sns
import matplotlib.pyplot as plt

import wordcloud
from wordcloud import WordCloud
import json
import folium
%matplotlib inline
# 한글 설정
# pip install koreanize_matplotlib
plt.rc('font', family='Malgun Gothic')
plt.rc('axes', unicode_minus=False)
```

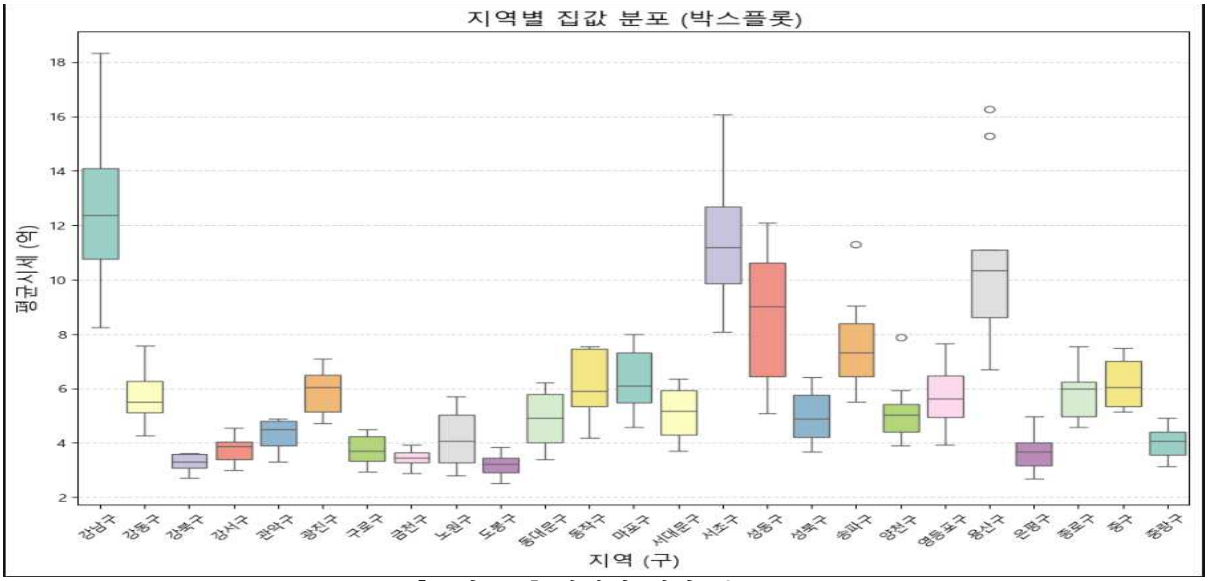
[그림Ⅳ-1] 파이썬 라이브러리

```
from sklearn.model_selection import train_test_spli

from xgboost import XGBRegressor
from sklearn.metrics import mean_squared_error, r2_

import geopandas as gpd

from sklearn.ensemble import RandomForestRegressor
```

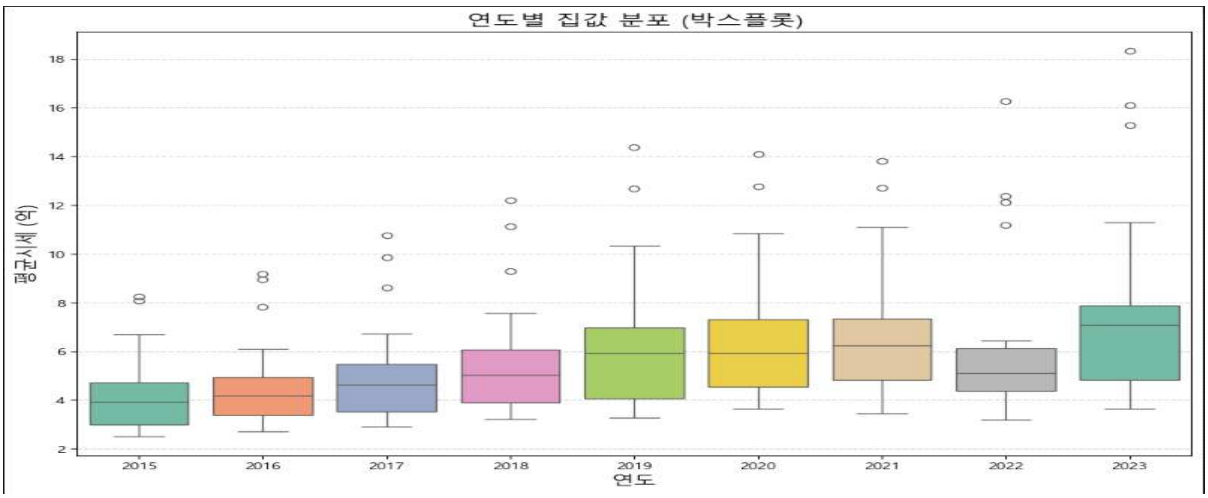
[그림Ⅳ-2] 파이썬 라이브러리

또한 모델링에 필요한 ’train_test_split’, ’XGBRegressor’. ’RandomForestRegressor’와 평가할 때 필요한 ’mean_squared_error’, ’r2_score’를 사용했다.

## 가. 집값 분포도 확인




[그림IV-3] 지역별 집값 분포도

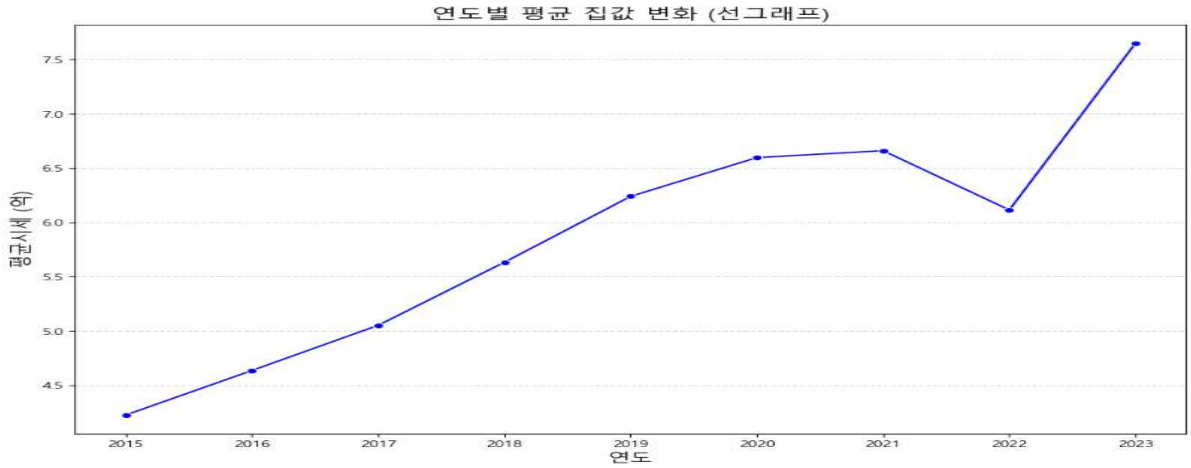



[그림IV-4] 연도별 집값 분포도

데이터 분석을 위해 다양한 지역에 따른 집값의 차이를 나타내고 특정 지역의 집값이 높은지 확인하기 위해 지역별 집값 분포도를 시각화하여 강남구, 서초구, 성동구, 송파구와 용산구의 평균 집값의 변동이 심한 것을 확인할 수 있었다.

또한 연도별 집값 분포도 변동을 시각적으로 나타내는 그래프를 통해 연도에 따른 평균 집값 상승 혹은 하락 추세를 한눈에 볼 수 있도록 시각화하였다.

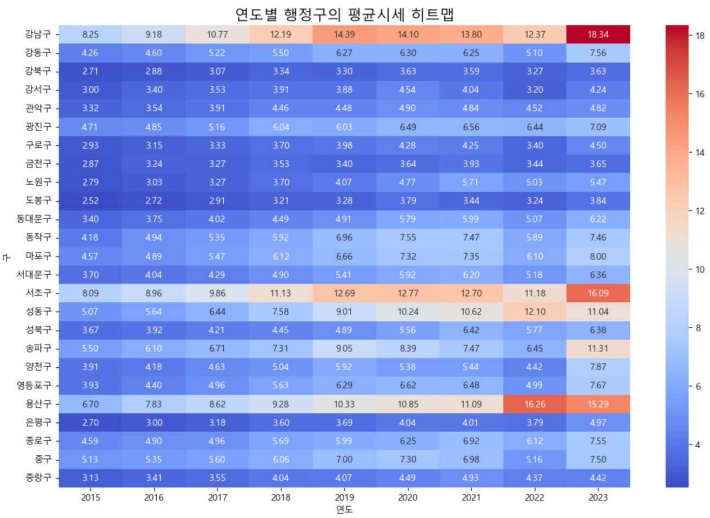
## 나. 집값 변동 패턴 및 추세 확인




[그림Ⅳ-5] 연도별 집값 변화 그래프

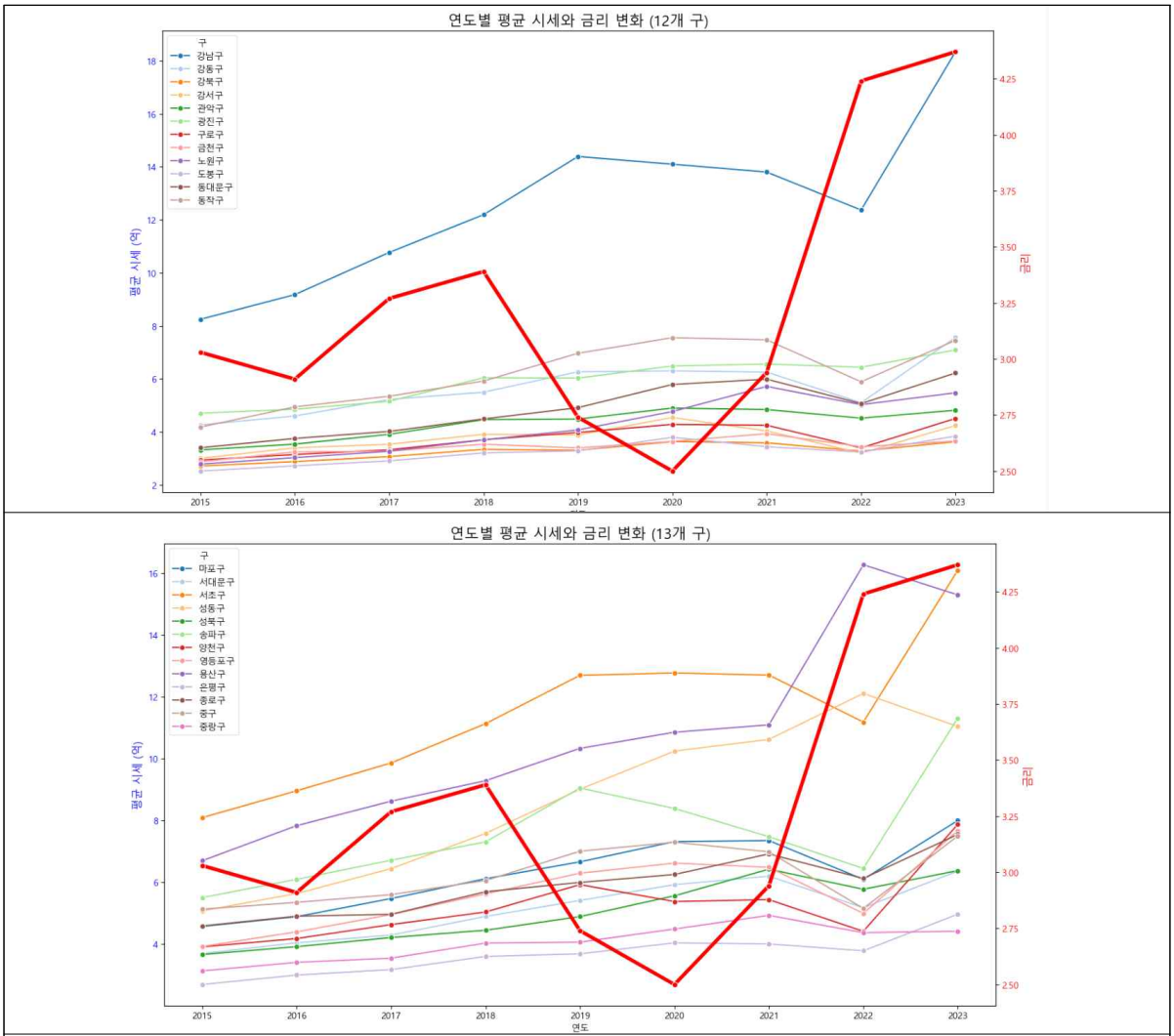
연도별 선 그래프는 각 연도에 해당하는 평균 집값을 시각적으로 연결하여, 집값의 상승 또는 하락 추세를 한눈에 볼 수 있도록 하여 [그림Ⅳ-4]의 연도별 집값 분포도 보다 더욱 직관적인 분석이 가능하다.

이를 통해 장기적인 집값 변동 추세와 특정 시점에서의 급격한 변화(예: 급등, 급락)를 쉽게 식별할 수 있다.




[그림Ⅳ-6] 연도별 지역간 평균시세 히트맵

또한 히트맵으로 연도별 행정구의 평균 시세를 시각화 자료에서 강남구, 서초구, 용산구, 송파구가 다른 행정구에 비해 눈에 띄게 가격이 높고 과하게 가격이 상승하는 것을 정확한 수치와 함께 색상으로 구분하여 더욱 세밀하게 분석이 가능하다.



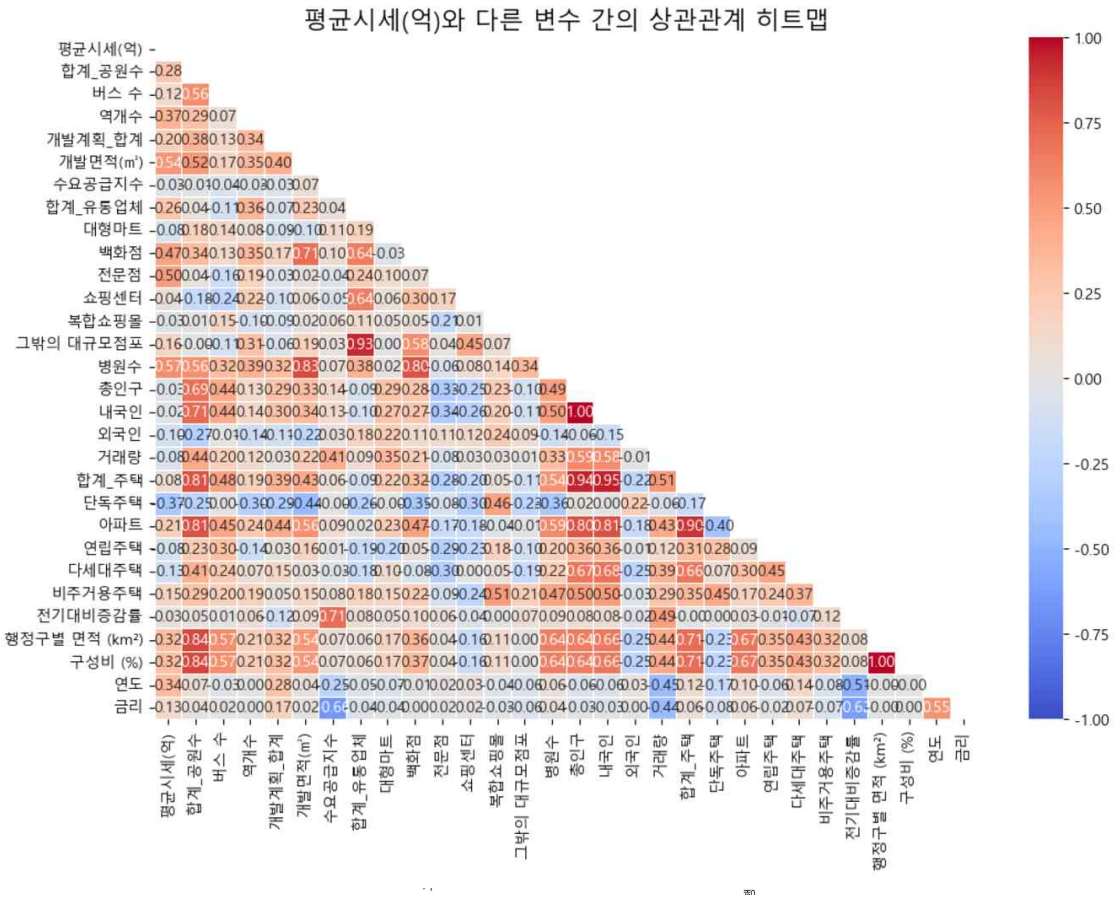

[그림Ⅳ-6] 연도별 지역간 평균시세 와 금리 시각화

일반적으로 금리가 낮아지면 부동산 시세는 올라가고 금리가 높아지면 부동산 시세가 떨어진다. 이는 적은 이자로 인해 부동산 담보 대출하는 소비자들이 많아지며 부동산 거래 수요의 증가로 인한 가격 상승과 반대로 높은 이자로 인해 대출을 하는 소비자들이 적어지고, 부동산 거래 수요가 감소함에 따라 부동산 가격이 떨어지는 원리이다.

실제로 2018년도에서 2020년도까지 금리가 대폭 하락함에 따라 대부분 행정구의 평균 부동산 거래 시세가 상승하는 경향이 보인다.

그러나 2021년도에서 2022년도에 금리가 대폭 상승하여 대부분의 행정구 부동산 시세가 떨어지지만 용산구는 급증, 성동구는 소폭 상승하는 모습을 확인할 수 있어 해당 자치구는 금리 외에도 다른 원인으로 인해 부동산 거래 시세가 바뀌었음을 의미한다.

## 다. 주요 원인과 상관관계 분석




[그림Ⅳ-7] 평균시세와 다른 변수간의 상관관계 히트맵

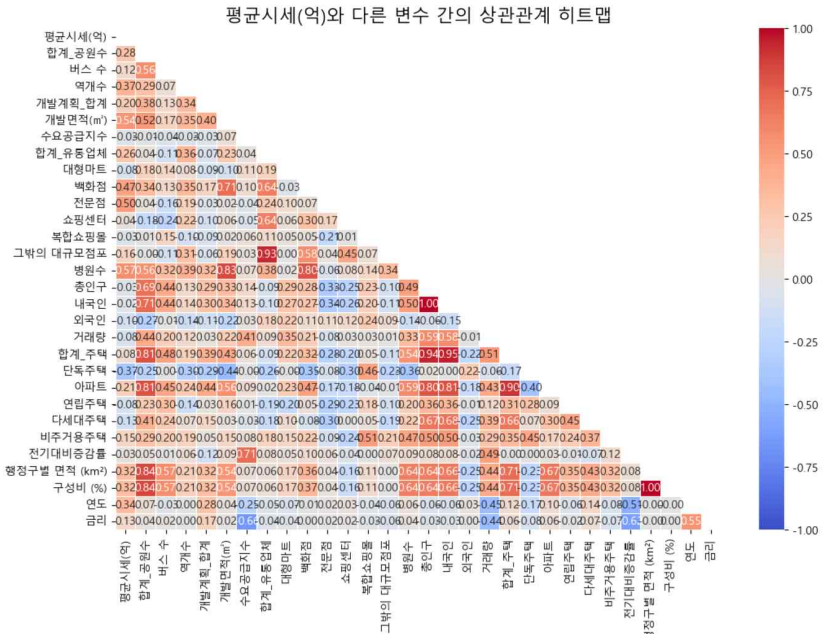
평균 시세와 수집했던 데이터의 변수들을 가지고 상관관계를 나타내기 위한 히트맵을 시각화하여 중요하지 않거나 상관관계가 낮은 데이터는 여러 개의 변수를 결합해서 파생 변수를 생성하고 필요 없는 데이터는 제외시켜 데이터 분석이 더욱 용이하게 한다.

## 라. 파생변수 생성 및 시각화

파생 변수 생성은 기존의 변수들을 활용하여 상관관계가 서로 높은 변수끼리 묶어 하나의 변수로 축소하거나 평균 시세와 변수 간의 상관관계가 떨어질 경우 논리적인 기준으로 더욱 상관관계가 높아지도록 하기 위해 진행한다.

| 변수(컬럼)명 | 이론 설명 (기준 : 각 행정 자치구별 수치) |
|---|---|
| 1인당 공원수 | 공원의 수를 인구 수로 나눠 행정 자치구의 총 인구 수 대비 공원 공급량을 알 수 있다. |
| 1인당 병원수 | 병원의 수를 인구 수로 나눠 행정 자치구의 총 인구 수 대비 병원 공급량을 알 수 있다. |
| 1인당 유통시설 | 유통업체의 합계를 인구 수로 나눠 행정 자치구의 총 인구 수 대비 유통시설의 공급량을 알 수 있다. |
| 교통 접근성 | (버스 수 + 역 개수)를 합한 후 인구 수로 나눠 행정 자치구의 총 인구수 대비 대중교통 접근성을 알 수 있다. |
| 총 인프라 수 | (버스 수 + 역 개수) + 유통업체 합계 + 병원 수를 통해 행정 자치구의 총 인프라 공급량을 알 수 있다. |
| 인프라 밀집도 | 총 인프라 수를 행정구별 총 면적으로 나눠 행정 자치구의 면적 대비 인프라 밀집도를 알 수 있다. |
| 교통_병원 | 교통 접근성에 병원 수를 곱하여 교통 접근성과 병원의 상호작용 수치를 알 수 있다 |
| 주택 밀집도 | 주택 합계에서 비주거용주택을 빼고 행정구별 총 면적을 나눠 행정 자치구의 면적 대비 주택 밀집도를 알 수 있다. |
| 1인당 거래량 | 거래량을 인구 수로 나눠 행정구별 인구 수 대비 거래량을 알 수 있다. |
| 병원 밀집도 | 병원의 수를 행정구별 총 면적으로 나눠 행정 자치구의 면적 대비 병원의 밀집도를 알 수 있다. |
| 거래량 대비 개발계획 | 거래량을 개발계획 합계로 나눠 개발 계획 양에 따라 거래량이 바뀌는지 알 수 있다. |

[표Ⅳ-8] 파생변수 선정기준 및 설명



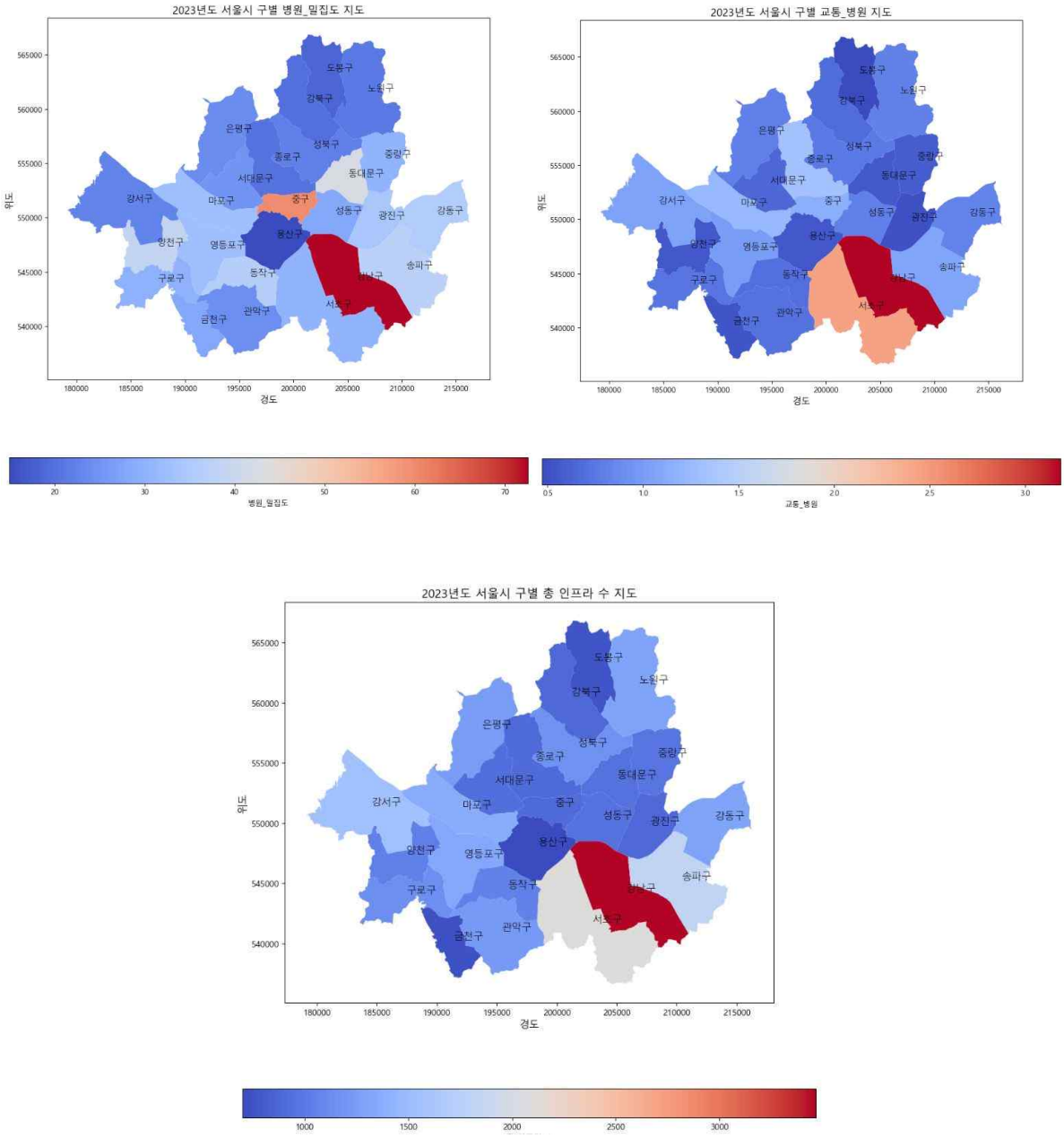

[그림IV-9] 파생변수 생성 후 평균시세와의 상관관계 히트맵

파생 변수 후의 상관관계를 히트맵으로 시각화하였고, 이를 통해 교통_병원과 평균 시세 간에는 강한 양의 상관관계(0.63)를 알 수 있다. 이는 교통_병원 증가가 평균 집값 상승과 밀접한 관계가 있음을 짐작할 수 있다.

또한 금리와 평균 시세 간에는 상관관계가 크지 않으므로, 이는 금리가 상승할 경우 집값이 하락하는 경향이 있음을 알 수 있다.

머신러닝 학습을 위해 모델의 선정도 중요하지만, 독립변수를 어떻게 선정하는가에 따라 성능이 크게 바뀐다. 독립변수가 시험지, 종속변수가 답안지 라고 생각하면 종속변수인 답안지의 답은 고정되어 정해져 있지만 독립변수는 여러 방식으로 바꿀 수가 있다. 독립변수인 시험지의 문제를 얼마나 쉽게, 차근차근 풀어 써가며 문제를 제출했는가? 정도로 생각하면 이해하기 쉽다. 문제의 정답이 3이라면 정답의 문제는 간단하게 1 + 2라고 할 수 있지만, 독립변수를 잘 못 선택하면 한순간에 문제가 (2^3 / 5) * 3 - 1.8 이런 식으로 난해해질 수 있다.

그래서 적절한 독립변수를 선택하기 위해 종속변수인 평균 시세(억) 과 상관관계가 높은 병원_밀집도, 총 인프라 수, 교통_병원 3개의 변수를 시각화하여 분석해 본다.




[그림Ⅳ-10] 높은 상관관계 변수와 평균시세 비교

시각화 결과 평균적으로 23년도에 집값이 많이 올라간 강남구, 송파구, 서초구, 용산구에 상관관계가 높은 변수의 비중이 특정 지역에 몰려 있는 것으로 확인이 된다. 이 말은 즉 해당 변수가 많을수록 집값 상승 요인 중 하나가 될 수 있을 것으로 판단되어 독립변수로 사용하기 적합하다고 판단할 수 있다.

# V

# 머신러닝 학습

1. 독립변수, 종속변수 지정
2. 모델 선정 및 학습
3. 예측 결과 분석

# V 머신러닝 학습

## 1. 독립변수, 종속변수 지정

```python
# 1. 데이터 분리
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# 2. XGBRegressor 모델을 학습할 때 사용되는 최적의 파라미터
best_params = {'colsample_bytree': 1.0,
               'learning_rate': 0.05,
               'max_depth': 7,
               'n_estimators': 200,
               'subsample': 0.8}
model_xgb = XGBRegressor(**best_params)

# 3. 모델 학습
model_xgb.fit(X_train, y_train)

# 4. 예측
y_pred_xgb = model_xgb.predict(X_test)

# 5. 모델 평가
mse_xgb = mean_squared_error(y_test, y_pred_xgb)
r2_xgb = r2_score(y_test, y_pred_xgb)

# 평가 지표 출력
print(f"Best XGBoost Mean Squared Error: {mse_xgb:.2f}")
print(f"Best XGBoost R2 Score: {r2_xgb:.2f}")
```

[그림V-1] 모델링 코드

```
Best XGBoost Mean Squared Error: 0.39
Best XGBoost R2 Score: 0.85
```

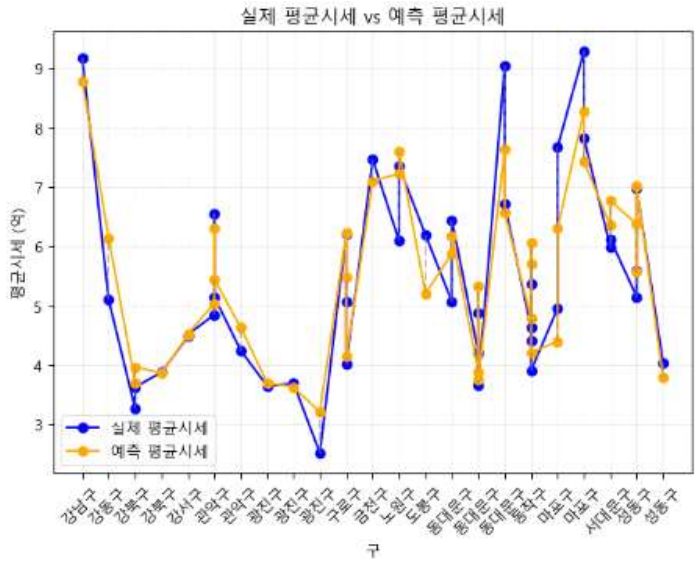
| | 구 | 평균시세(억) | 예측_평균시세 | 오차 |
|---|---|---|---|---|
| 25 | 강남구 | 9.178158 | 8.788880 | 0.389278 |
| 176 | 강동구 | 5.101104 | 6.143574 | 1.042470 |
| 177 | 강북구 | 3.273290 | 3.705835 | 0.432545 |
| 127 | 강북구 | 3.631879 | 3.973297 | 0.341418 |
| 103 | 강서구 | 3.881667 | 3.866060 | 0.015607 |
| 104 | 관악구 | 4.482254 | 4.510640 | 0.028386 |
| 179 | 관악구 | 4.516659 | 4.530098 | 0.013439 |
| 30 | 광진구 | 4.853695 | 5.039029 | 0.185334 |

[그림V-2] 모델링 오차 확인

Pipeline을 이용하여 한 번에 여러 모델을 가지고 모델링한 결과, XGBoost 모델이 가장 적합한 모델로 나왔다.

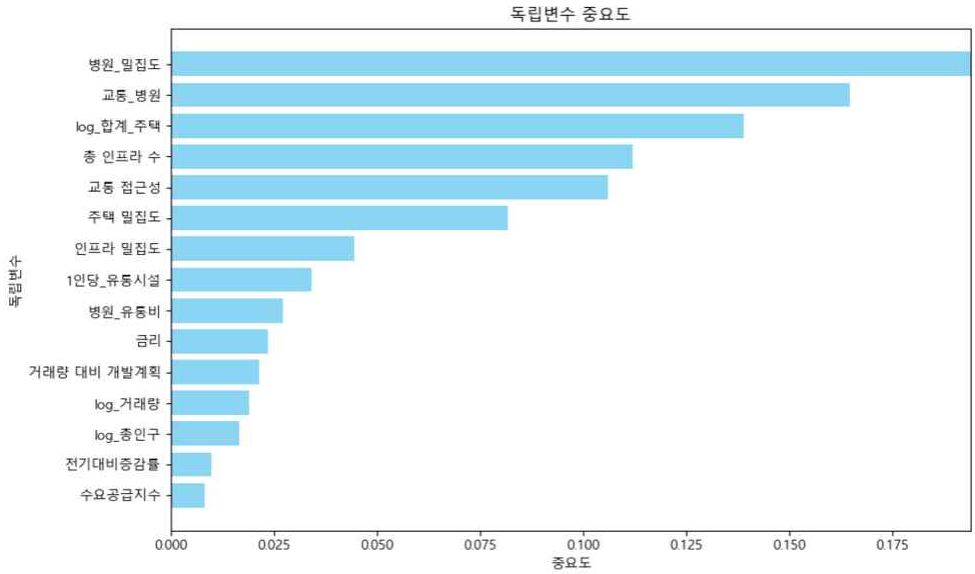
오차를 줄이기 위해 PCA 차원 축소 후 그리드 서치로 최적의 파라미터를 찾은 후 해당 파라미터를 적용 후 feature important 가 0.01 이하인 독립변수 제거한 후 모델링 진행했다.

## 2. 모델 적합도 확인 및 독립변수 중요도 확인



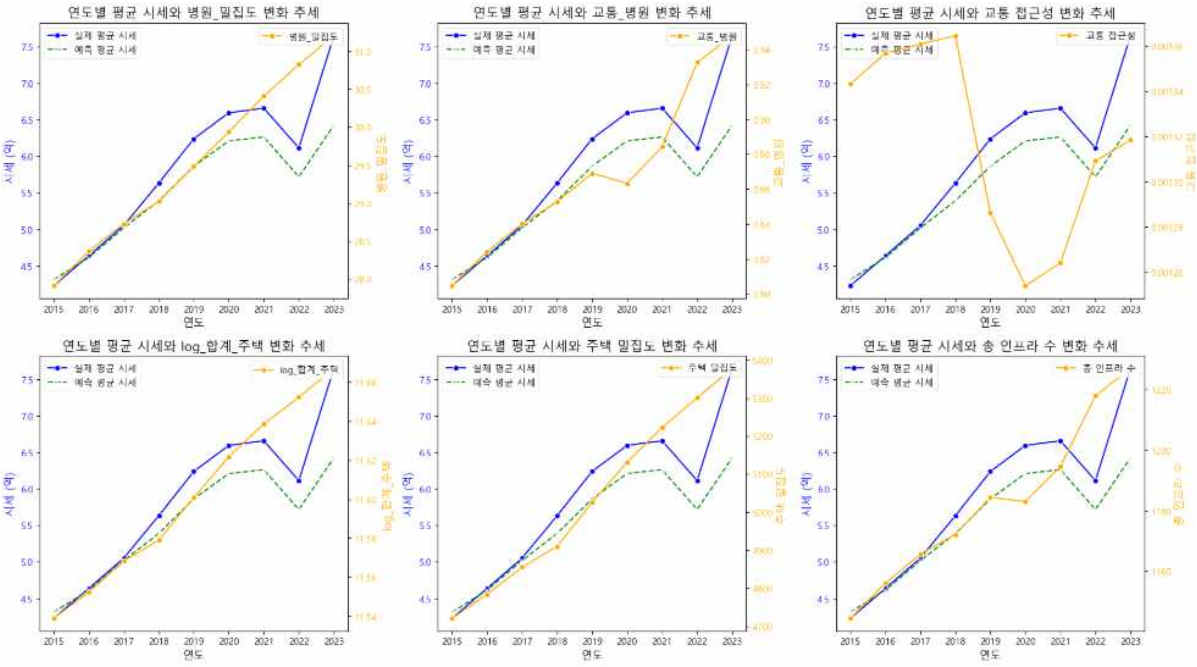

[그림V-3] 모델 적합도 확인 (실제값 vs 예측값)

모델링 한 후 적합한 모델인지 확인을 위한 실제 평균 시세와 예측한 평균 시세를 보기 쉽게 선 그래프로 비교하는 시각화 작업을 하였고, 그림과 같이 모델링이 잘 되어 실제 값과 예측값이 큰 오차 없이 적합한 걸 알 수 있다.




[그림V-4] 변수 중요도 확인 막대그래프

또한 Feature Important로 종속 변수 중 중요도가 높은 변수를 막대그래프로 통해 확인하여 집값과 중요도가 높은 변수와의 관계를 알기 위해 시각화를 진행하여, 중요도가 높은 상위 6개의 독립변수를 알 수 있다.




[그림V-5] 중요도 상위 6개 변수와 예측시세 비교

머신러닝 분석 결과, 교통 접근성, 병원 밀집도, 총 인프라 수, 주택 밀집도, log_합계_주택 그리고 교통_병원이 상위 중요 변수로 도출되었다. 교통 접근성을 제외한 나머지 변수들은 모두 집값 상승에 큰 영향을 미친 것으로 분석되었으며, 특히 병원 밀집도와 총 인프라 수는 주거지의 편의성과 생활 인프라의 질을 향상해 시세에 긍정적인 영향을 미친 주요 요인으로 평가된다. 이러한 변수들의 변화가 예측 값과 실제 집값의 변동과도 밀접하게 일치함을 확인할 수 있다.

# Ⅵ

# 결론 및 제언

1. 결론
2. 제언

# Ⅵ 결론 및 제언

## 1. 결론

최근 서울특별시의 부동산 거래 시세가 급등하고 있다. 20년도에 비해 높은 금리와 경제 침체로 인해 수요가 줄어들어 부동산 거래 시세가 하락하는 추세를 보여야 하지만 다른 이유에서 부동산 시세가 증가한 것이다.

최근 서울특별시의 부동산 거래 시세가 급등하고 있다. 20년도에 비해 높은 금리와 경제 침체로 인해 수요가 줄어들어 부동산 거래 시세가 하락하는 추세를 보여야 하지만 다른 이유에서 부동산 시세가 증가한 것이다.

본 연구로 머신러닝 학습과 데이터 분석을 해본 결과로 그 이유를 유추해 볼 수 있다. 학습시킨 모델로 집값 예측을 해본 결과 평균 오차 약 0.4668(억) 정도로 적은 오차의 예측 결과를 도출할 수 있었다.

해당 모델을 통해 부동산 투자 시 중요하게 봐야 할 정보가 무엇인지 알기 위해 모델의 Feature importance를 확인해 본 결과 상위 6개의 독립변수로 “병원 밀집도, 병원과 교통의 상호작용 지수, 행정구의 주택의 수, 교통 접근성, 총 인프라 수, 주택 밀집도”가 중요한 것으로 나왔다. 실제로 시각화하여 집값 상승에 얼마나 관여하는지 확인해 본 결과 실제로 교통 접근성을 제외한 나머지 변수들의 증가에 따라 집값도 같이 증가하는 경향을 확인했다.

그러나 교통 접근성과 같이 우리가 판단했을 때 교통 접근성이 높아지면 집값이 상승해야 하지만 시각화 결과 교통접근성이 떨어짐에도 집값이 상승하는 등 모든 정보를 하나만 보고 판단해서는 안 된다. 가장 좋은 사례로 금리가 있다. 21년도에서 22년도에 금리가 대폭 상승하여 대부분의 행정구는 부동산 시세가 줄어들었으나 용산구만 급증하는 것을 볼 수 있다. 그 해에 대통령 청사를 용산구로 이전했기 때문에 동시에 용산구의 집값이 급등한 것이다.

이렇듯 부동산 시세는 여러 변수와 소비자의 심리 상태 등 여러 가지 요인들로 인해 부동산 투자를 할 때에는 수많은 정보를 취합하여 결정해야 한다. 다만, 그 수많은 정보 중 본 연구를 통해 얻은 중요 자료(병원 밀집도, 병원과 교통의 상호작용 지수 등)를 제일 먼저 확인해 보고 빠른 판단을 내릴 수 있는 지표로써 활용할 수 있다고 생각한다.

## 2. 제언

머신러닝 학습을 위한 모델링과 데이터 수집 과정에서 많은 아쉬움을 느꼈다. 비교적 많은 데이터를 가져왔다고 생각하고 학습을 진행해 보니 한 개의 독립변수의 의존도만 너무 심하게 잡혀버리거나 오차가 너무 심하게 나와 오차를 해결하는 과정에서 큰 어려움이 있었다.

오차를 잡는 과정에서 단순하게 생각하고 파생 변수를 만들어 버리면 또 그 변수에 대한 의존도만 과하게 잡히는 등 또 다른 문제가 발생해 평균 시세와 상관관계가 낮은 변수들에 대해 더 깊은 분석이 필요하다고 생각한다.

향후에는 더 많은 데이터와 다른 지역의 데이터를 포함해 보다 나은 학습과 분석을 할 필요가 있으며, 이는 보다 신뢰할 수 있는 결과를 도출하는 데 도움이 될 것이라고 예상하며 회귀모델이 아닌 분류모델로도 만들어서 집값이 상승한다, 하락한다, 유지된다는 등 특정 독립변수의 변화에 따라 집값 예측에 더 직관적인 모델을 구현해 보는 것도 좋을 것으로 생각한다.

## 집필진

| | |
|---|---|
| 전체 | 데이터 수집 |
| 김희만,이용기,김선오 | 데이터 전처리 |
| 김희만,이용기 | 데이터 분석 |
| 김희만 | 머신러닝 학습 |

※ 본 연구 결과는 더조은 아이티 아카데미에서 데이터 분석 과정을 참여하며 부동산 시세 변동에 대해서 작성한 분석결과 최종보고서임.