

2023 DATA ANALYSIS

for APT price prediction in Seoul

Team 'K2S2'



CONTENTS

1

프로젝트 배경 및 목적

프로젝트 주제, 개요

2

데이터 분석

회귀분석을 통한 분석 과정

3

결론

결론과 향후 과제

4

팀소개 & 느낀점

파트 분배, 과제 후 느낀점



01

프로젝트 배경 및 목적

- 주제 선정
- 프로젝트 목적
- 프로젝트 진행 과정

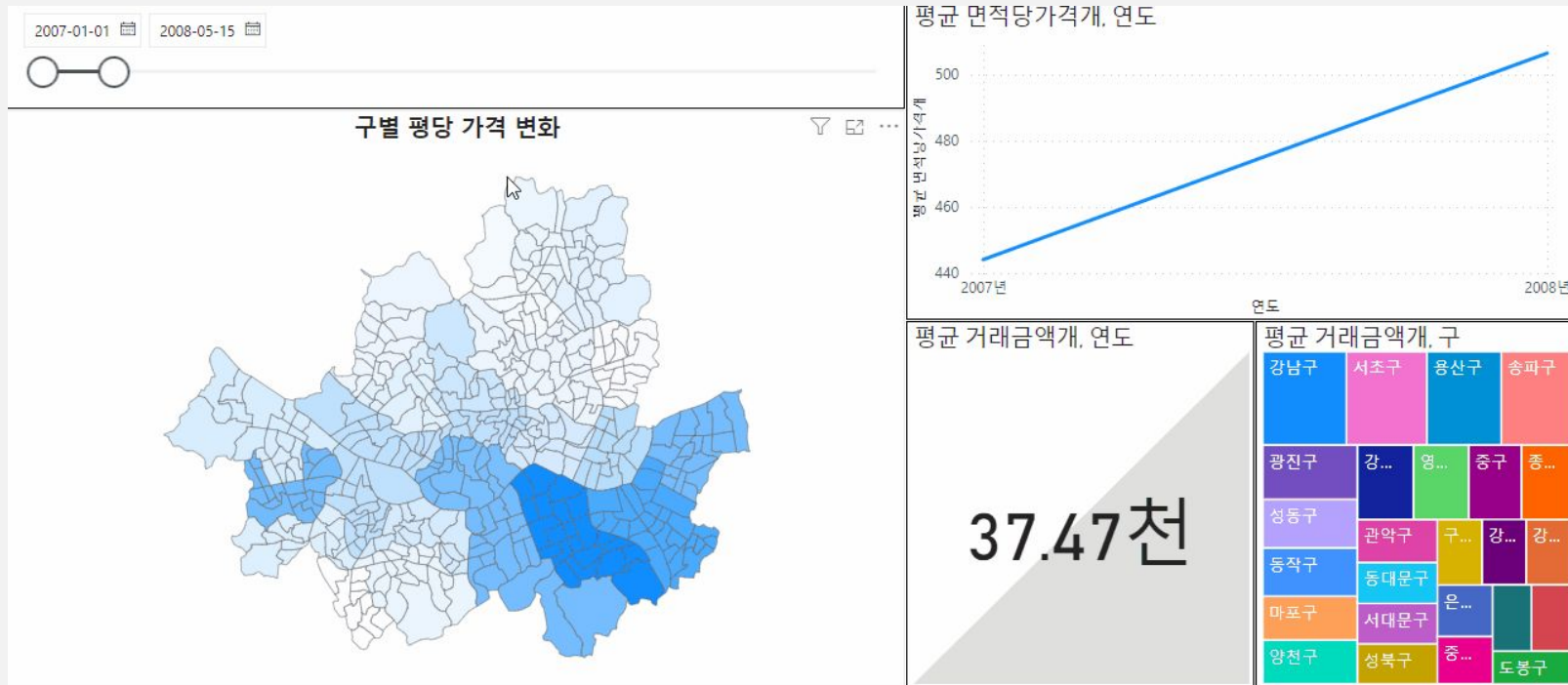


서울시 아파트 가격의 변화 예측

지난 15년 간의 데이터를 이용하여 서울시 아파트 가격에 대한 예측을 하고자 했다.



2007년부터 현재까지의 서울시 아파트 가격 변동



가격과 연관 있는 변수 탐색

연관성 높은 독립변수 탐색

↳ 다중공선성, p value 비교

미래 가격 변동 추이 제시

독립변수의 변화에 따른 예측

↳ 적합한 회귀 모델 선택



프로젝트 개요

주제 선정 이유

서울시 아파트 가격 변동을 예측하여,
적당한 매입, 매도 시기를 도출하고자 했다.

프로젝트 진행 순서

1. 주제 선정 및 데이터 수집
2. 데이터 분석 설계
3. 데이터 분석
4. 결론 도출 및 해석





02

데이터 분석

- 데이터 소개
- 데이터 분석 과정

- 데이터 분석 결과

종속변수

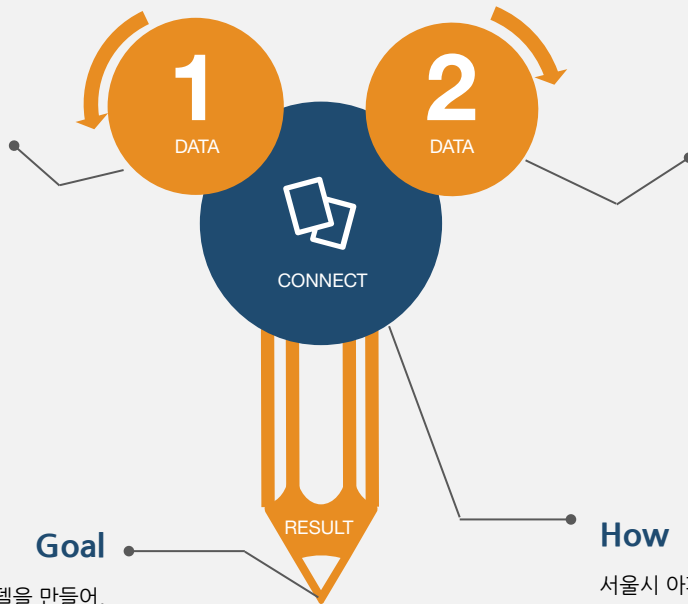
• 개요

2007.01 ~ 2023.02

서울시 아파트 매매 실거래가 데이터
(주소(구,동), 전용면적, 거래금액,
층, 건축년도, 계약일)

• 데이터 파일

아파트_실거래가.csv



독립변수

- ① 주택담보 대출금리
- ② 달러, 엔화, 유로
- ③ 소비자물가지수
- ④ 코스피, 코스닥
- ⑤ 무담보콜금리, KORIBOR(3개월), CD금리(91일), 국고채금리(1년/10년)

• 데이터 파일

finaldata.csv

Goal

매매가 변동 예측 모델을 만들어,
적당한 서울시 아파트 매매 시기를 알아본다.

How

서울시 아파트 매매가에 영향을 주는 변수들을 탐색하여,
앞으로의 매매가 변동을 예측하기 위한 분석을 진행.



전처리 과정


① 종속변수 (서울시 아파트 매매 실거래가)


- 데이터 분리
2007~2022년 데이터 : 분석 및 학습 진행
2023년 데이터 : 추후 학습모델의 예측가와 비교할 데이터셋
- 그룹화
서울시 / 구별로 groupby, 계약일을 년-월 단위로 묶음.
- 컬럼 추가
'면적당 가격 평균' = 거래금액/전용면적


②

독립변수

- 시간 단위 통일
년-월 (일 단위로 기록된 변수들은 평균가로 변환)

 Add your title

 Add your title

 Add your title

1

서울시 전체 (전처리 전) /
서울시 전체 (전처리 후) /
행정구별 데이터
3가지 경우의 OLS 분석
(R-squared, P-value 값)

2

독립변수 간 상관성 분석 (상관계수, 다중공선성),
R-squared 값을 높여주는
연관성 높은 변수만 남기기

3

구별 데이터로
회귀모델 학습 & 평가

4

예측 결과 확인
(2023년 데이터랑 값 비교)



1. OLS 분석

① 전체 데이터

: 별도의 전처리 없이 한 경우

R-squared : 0.805

OLS Regression Results						
Dep. Variable:	면적당가격	R-squared (centered):	0.805			
Model:	OLS	Adj. R-squared (centered):	0.805			
Method:	Least Squares	F-statistic:	2.716e+05			
Date:	Thu, 23 Mar 2023	Prob (F-statistic):	0.00			
Time:	16:46:49	Log-Likelihood:	-8.2106e+06			
No. Observations:	11117750	AIC:	1.644e+07			
DF Residuals:	11117733	BIC:	1.644e+07			
DF Model:	17					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
전통면적	0.4169	0.012	34.076	0.000	0.393	0.441
층	7.1531	0.061	116.842	0.000	7.033	7.273
건축년도	-2.8416	0.040	-71.787	0.000	-2.919	-2.764
개역년	1.4230	0.040	35.167	0.000	1.344	1.502
개역월	4.5677	0.120	38.175	0.000	4.333	4.802
주택담보대출	-16.2201	2.043	-7.938	0.000	-20.225	-12.215
미국달러	0.4166	0.011	39.397	0.000	0.396	0.437
일본엔	4.0709	0.452	9.001	0.000	3.164	4.957
유로	0.4711	0.007	69.607	0.000	0.458	0.484
소비자물가지수	15.7082	0.217	72.354	0.000	15.264	16.135
KOSPI_종가	0.1190	0.003	35.913	0.000	0.113	0.126
KOSDAQ_종가	0.9063	0.006	140.002	0.000	0.894	0.919
무담보대출	-351.8827	3.938	-89.363	0.000	-359.600	-344.165
KORIBOR	442.8011	10.416	42.492	0.000	422.166	463.016
CD	53.4213	10.237	5.219	0.000	33.358	73.485
국고채_1년	-124.3806	4.029	-30.869	0.000	-132.278	-116.483
국고채_10년	-34.3710	1.759	-19.541	0.000	-37.618	-30.924
Omnibus:	570708.181	Durbin-Watson:	1.758			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5469432.321			
Skew:	2.269	Prob(JB):	0.00			
Kurtosis:	12.841	Cond. No.	1.59e+05			

② 서울시 평균 데이터

: 서울시 전체를 평균 낸 경우

R-squared : 0.885

Dep. Variable:	면적당가격	R-squared:	0.885			
Model:	OLS	Adj. R-squared:	0.877			
Method:	Least Squares	F-statistic:	105.9			
Date:	Mon, 20 Mar 2023	Prob (F-statistic):	1.06e-76			
Time:	15:10:03	Log-Likelihood:	-1158.5			
No. Observations:	193	AIC:	2345.			
DF Residuals:	179	BIC:	2391.			
DF Model:	13					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	-2944.2723	262.714	-11.207	0.000	-3462.688	-2425.857
주택담보대출관리	28.6341	38.501	0.744	0.458	-47.339	104.608
미국달러	0.4307	0.190	2.268	0.024	0.056	0.805
일본엔	0.4190	8.554	0.470	0.639	-12.861	20.699
유로	0.3242	0.127	2.549	0.012	0.073	0.575
소비자물가지수	19.8063	3.648	5.429	0.000	12.607	27.005
소비자물가지수증감률	21.0612	23.813	0.884	0.378	-25.929	68.052
KOSPI_종가	0.0629	0.066	0.958	0.340	-0.067	0.192
KOSDAQ_종가	0.9946	0.146	6.833	0.000	0.707	1.282
무담보대출	-173.0088	64.524	-2.681	0.008	-300.334	-45.684
KORIBOR	451.0353	193.137	2.335	0.021	69.917	832.154
CD	-198.5257	192.000	-1.034	0.303	-577.400	180.348
국고채_1년	-104.5263	67.306	-1.553	0.122	-237.342	28.290
국고채_10년	-2.1603	33.082	-0.065	0.948	-67.442	63.121
Omnibus:	6.610	Durbin-Watson:	0.347			
Prob(Omnibus):	0.037	Jarque-Bera (JB):	9.334			
Skew:	0.175	Prob(JB):	0.00940			
Kurtosis:	4.019	Cond. No.	1.08e+05			

③ 행정구별 평균 데이터

: 구별로 평균 낸 경우

•Top 4

중랑구 R-squared : 0.642
 성북구 R-squared : 0.637

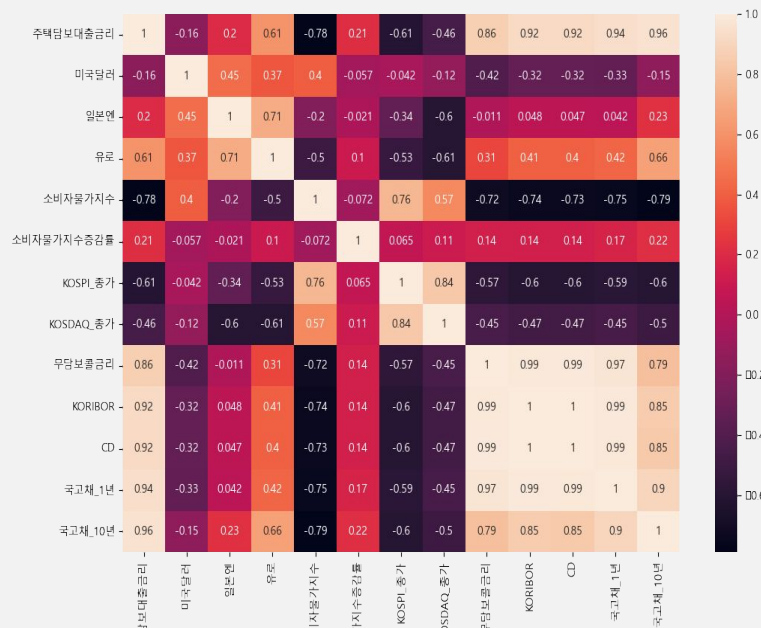
은평구 R-squared : 0.631
 서대문구 R-squared : 0.624

행정구	R-squared
중랑구	0.642
성북구	0.637
은평구	0.631
서대문구	0.624
성동구	0.608
노원구	0.602
강북구	0.601
동대문구	0.585
마포구	0.552
관악구	0.549
강서구	0.548
중구	0.543

금천구	0.532
도봉구	0.525
광진구	0.522
동작구	0.512
영등포구	0.488
종로구	0.474
양천구	0.462
구로구	0.44
용산구	0.408
강남구	0.405
서초구	0.405
강동구	0.373
송파구	0.339

2. 독립변수 상관성 분석

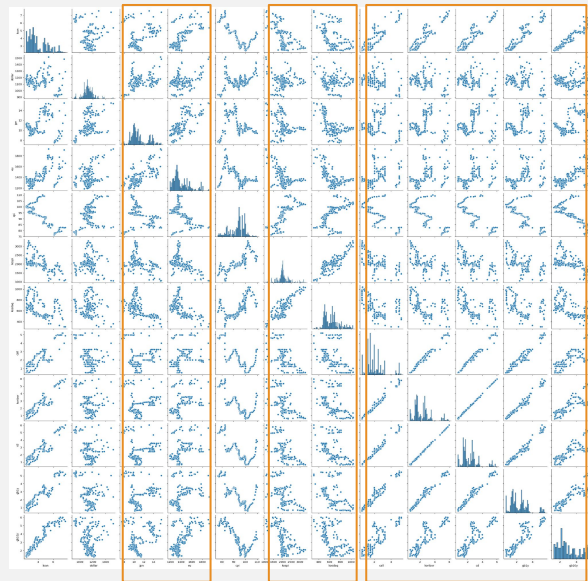
① 상관계수



② 다중공선성

	VIF	변수
0	1.403638	계약월
1	2.246945	총
2	3.137945	전용면적
3	3.662097	건축년도
4	5.713912	일본엔
5	6.436517	미국달러
6	7.329334	유로
7	8.184322	KOSDAQ_총가
8	10.957126	KOSPI_총가
9	26.357654	계약년
10	43.806864	국고채_10년
11	49.841558	주택담보대출
12	143.171527	무담보콜금리
13	154.514508	국고채_1년
14	1261.865944	CD
15	1326.901343	KORIBOR

- 1) 환율 (엔화, 유로)
- 2) 주가 (kospi, kosdaq)
- 3) 금리 (주담대, 콜금리, koribor, cd, 국고채(1,10년))
(cpi는 3번 그룹과 상관계수 높지만, 그래프 다름)





2. 독립변수 상관성 분석

③ 구별 데이터 사용 & 연관성 높은 변수만 남기기

| VIF 순위가 구마다 다름
따라서, 구별로 분석.

| 적합한 변수 도출 과정
1. VIF가 높은 순서대로 변수를 제거
2. OLS 분석을 반복하며 r, p값을 비교

| VIF 10 이하인 변수를 사용해
회귀 분석 진행

은평구

	VIF	변수
0	1.006628	층
1	1.030393	전용면적
2	1.056982	건축년도
3	1.334102	계약월
4	4.615210	일본엔
5	5.377316	미국달러
6	6.307534	KOSDAQ_종가
7	7.797948	유로
8	8.686680	KOSPI_종가

서대문구

	VIF	변수
0	1.010470	전용면적
1	1.040466	층
2	1.095460	건축년도
3	1.372340	계약월
4	4.445222	일본엔
5	5.179834	KOSDAQ_종가
6	5.737198	미국달러
7	7.188304	유로
8	7.851818	KOSPI_종가

성북구

	VIF	변수
0	1.016487	층
1	1.024922	전용면적
2	1.050183	건축년도
3	1.379832	계약월
4	4.310482	일본엔
5	5.447995	미국달러
6	5.526794	KOSDAQ_종가
7	7.579139	유로
8	7.783923	KOSPI_종가

중랑구

	VIF	변수
0	1.034254	층
1	1.043395	전용면적
2	1.076151	건축년도
3	1.388016	계약월
4	4.758556	일본엔
5	5.285470	KOSDAQ_종가
6	6.238184	미국달러
7	7.092226	유로
8	8.035585	KOSPI_종가



3. 회귀 모델 학습 및 평가, 실제값과 비교

① 회귀 모델 선택

1) LinearRegression
: 오차범위가 가장 적음. 추천

2) DecisionTreeRegressor
: 과적합, 사용하기에
적합하지 않음

3) RandomForestRegressor
: 오차 범위가 1보다 큼

*

decisionTreeregressor과
randomForestRegressor는
예측값 1000개로 평균을 냈다.

② 모델별 결과

LinearRegression



중량구

train : 0.890점
test : 0.842점
실거래가: 827.834
예측가: 790.506



성북구

train : 0.872점
test : 0.789점
실거래가: 1094.741
예측가: 1096.834



은평구

train : 0.929점
test : 0.835점
실거래가: 989.959
예측가: 950.821



서대문구

train : 0.897점
test : 0.810점
실거래가: 1088.542
예측가: 1083.869

DecisionTreeRegressor

중량구

train : 1.0점
test : 0.776점
실거래가: 827.834
예측가: 946.068

성북구

train : 1.0점
test : 0.637점
실거래가: 1094.741
예측가: 1011.991

은평구

train : 1.0점
test : 0.795점
실거래가: 989.959
예측가: 1075.836

서대문구

train : 1.0점
test : 0.533점
실거래가: 1088.542
예측가: 1168.184

RandomForestRegressor

중량구

train : 0.984점
test : 0.903점
실거래가: 827.834
예측가: 781.347

성북구

train : 0.987점
test : 0.903점
실거래가: 1094.943
예측가: 1038.58

은평구

train : 0.988점
test : 0.834점
실거래가: 989.959
예측가: 1023.471

서대문구

train : 0.981점
test : 0.799점
실거래가: 1088.542
예측가: 1022.65



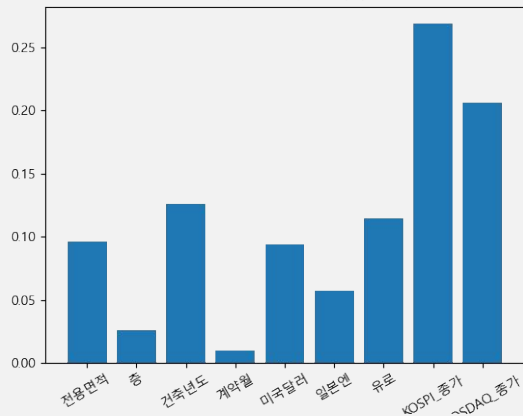
4. 연관성 높은 변수 시각화

분석 내용

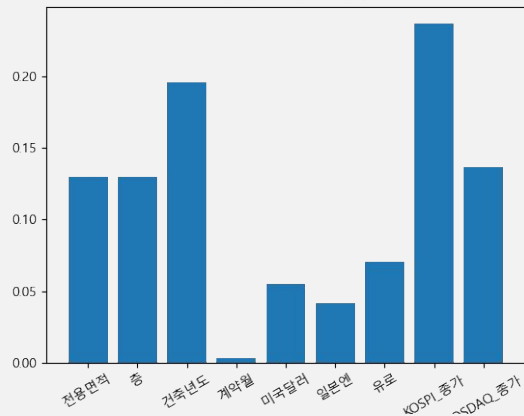
GridSearchCV를 이용하여
최적의 하이퍼 파라미터를
알아보고자 했다.

결과값을 통해 구마다
독립변수들의 기여도 순위가
다르다는 것을 알 수 있다.

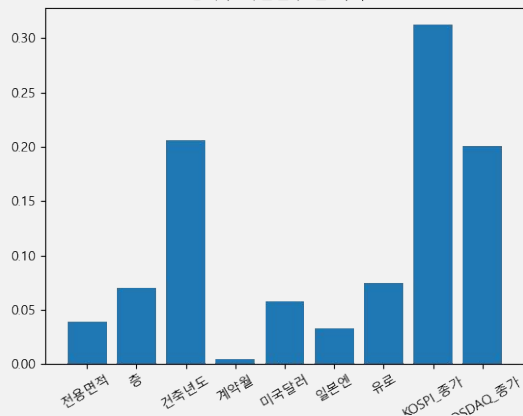
중랑구 독립변수 별 기여도



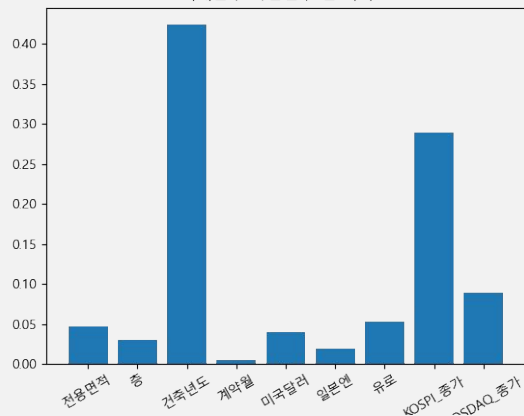
은평구 독립변수 별 기여도



성북구 독립변수 별 기여도



서대문구 독립변수 별 기여도





데이터 분석 결과 정리

연관성 높은 변수

같은 서울시라도 지역마다 연관성 있는 변수가 달라지는 것으로 확인됐다. 이번 프로젝트에서 상세히 알아본 네 지역에서는 건물 자체 특성(층, 전용면적, 건축년도)와 계약 월, 환율(일본엔, 미국달러,유로), 주가(KOSDAQ, KOSPI)의 영향을 많이 받는 것으로 보인다.

아파트 매매가 예측

회귀 분석을 통해 모델링한 결과로 만들어낸 예측 모델 역시 지역에 따라 적합한 모델이 다른 것으로 나타났다.

LinearRegression 모델이 모든 지역에서 가장 오차범위가 적어서 적합하다고 보여지고, RandomForestRegressor 모델도 오차 범위가 LinearRegression보다는 크지만 사용하기엔 나쁘지 않은 정도로 나타났다. 다만, DecisionTreeRegressor 모델은 모든 지역에서 과적합이 발생하기 때문에 아파트 가격 예측에 사용하기에는 맞지 않는 것으로 유추된다.

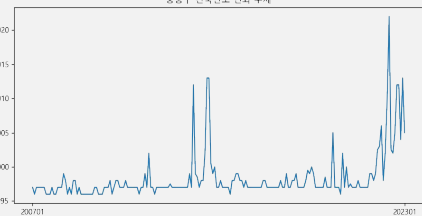
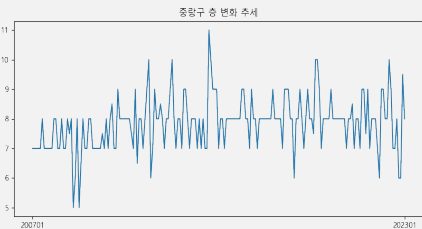
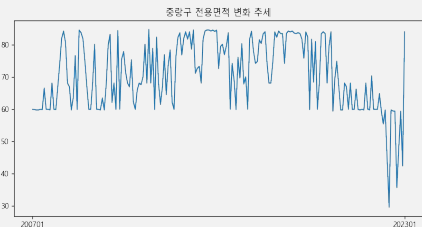


2024 서울시 구별 평단가 예측하기

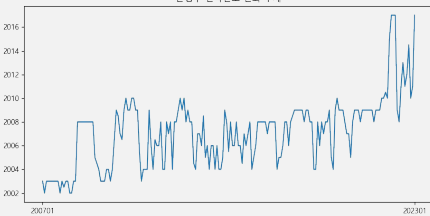
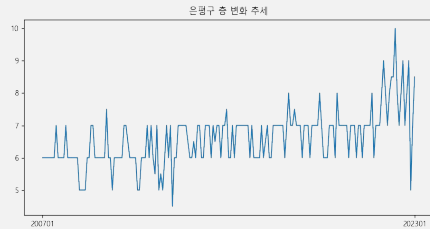
'전용면적', '층', '건축년도', 'KOSPI', '미국달러', '일본엔', '유로', 'KOSDAQ', '계약월'

■그래프 변화 : 전용면적(감소 또는 유지), 층(일정한 중간값으로 유지), 건축년도(상승하는 추세)

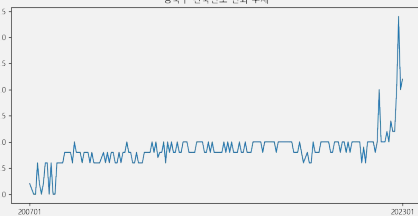
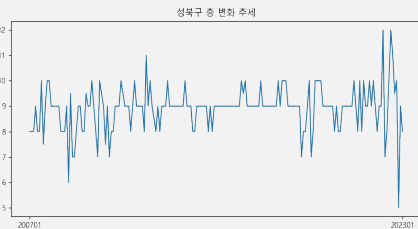
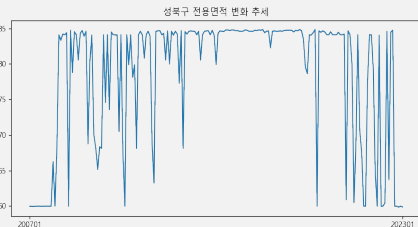
■중랑구



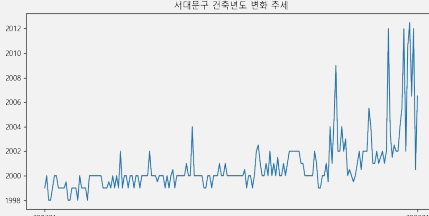
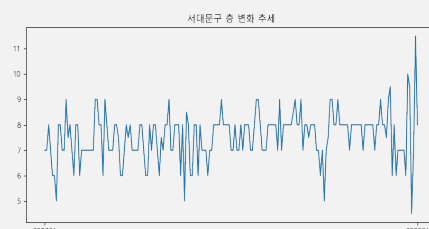
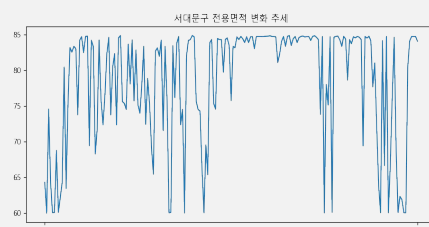
■은평구



■성북구



■서대문구



2024 서울시 구별 평단가 예측하기

▪사용 데이터 : [2024년 예측 데이터](#)

▪2024년 1월 예측 데이터 (평균값)

'전용면적', '층', '건축년도', 'KOSPI', '미국달러', '일본엔', '유로', 'KOSDAQ', '계약월' = '구마다 상이', '구마다 상이', '구마다 상이', 2380, 1301, 9.92, 1397.85,827.69 , 1

▪**중랑구** [45, 8.86, 2003.26,2380, 1301, 9.92, 1397.85,827.69 , 1]

	2023	2024
전용면적	69.286400	감소 = 45
층	8.864286	유지 = 8.86
건축년도	2001.261905	증가 = 2003.26

▪**성북구** [77.269,9.539,2005.4,2380, 1301, 9.92, 1397.85,827.69 , 1]

	2023	2024
전용면적	77.269716	유지 = 77.269
층	9.539216	유지 = 9.539
건축년도	2003.441176	증가= 2005.4

▪**은평구** [73.86,7.26, 2007,2380, 1301, 9.92, 1397.85,827.69 , 1]

	2023	2024
전용면적	78.869703	감소 = 73.86
층	7.266246	유지 = 7.26
건축년도	2005.116464	증가 = 2007

▪**서대문구** [81,8.327496,2002.5848,2380, 1301, 9.92, 1397.85,827.69 , 1]

	2023	2024
전용면적	77.136663	유지 = 77.136
층	8.327496	유지 = 8.327496
건축년도	2000.584828	증가 = 2002.5848

2024 서울시 구별 평단가 예측하기

▪2024년 1월 구별 평균 예상 면적당 가격

	2023 (만원)	2024 (만원)	증감
중랑구	827.83	982.11	155만원 ↑
성북구	1094.74	1003.58	91만원 ↓
은평구	989.96	861.47	128만원 ↓
서대문구	1094.74	1012.91	82만원 ↓



03

결론

- 데이터 분석 결과 정리
- 결론 및 향후 과제



결론 및 향후 과제

분석 결과, 서울시 아파트 매매가와 연관이 있을 것으로 예상한 데이터 중에서 연관성이 높은 변수들을 알아낼 수 있었고, 회귀 분석을 통해 매매가 예측 모델을 학습시켰다.

아쉬운 점은 우리가 생각한 요인 이외에도 집값 변동에 영향을 미치는 요인들은 많기 때문에, 서울시 내에 있는 지역들 각각의 특성을 잘 반영하지 못한 것 같다. 다른 영향 요인을 더 구했더라면 더 정확한 데이터 예측을 할 수 있었을 것이다. 또한, 전처리 과정에서 월평균 데이터로 정리를 함으로써 데이터의 양이 현저히 줄어들었기에, 추후에는 일평균 데이터로 변환을 하여 분석을 진행해보아야 할 것이다.

변수들을 대입했을 때의 매매가 예측 모델을 만들 수 있었기에, 적당한 매입 시기까지 알기 위해서는 더 정확한 미래 예측 데이터들을 독립변수에 넣고 회귀 모델을 돌려보면 앞으로의 집값 변동을 예측 할 수 있을 것이라고 판단된다.



04

팀원 소개 및 느낀점

- 팀원 역할
- 느낀점

» 팀원 역할 및 느낀 점

느낀 점

수업 시간에 배운 이론들을 적용해봄으로, 전체적인 분석 과정을 이해하는데 도움이 되었다. 아직 완벽히 이해한 것은 아닌 것 같지만,,, 내용을 듣기만 할 때보다 조금 정리가 된 것 같다.

확실히 수업시간에 배우기만 하는것보다 실전으로 적용해보니 데이터 전처리 과정이나 데이터를 구할때 각 변수들을 이용해 어떤 결과가 나오는지 분석할 수 있어서 좋은 경험이 된 것 같다.

데이터를 수집하고 전처리하는 과정에서 원 데이터를 기준으로 보다 나은 결과를 만들어 내는 것이 중요하며 모델을 학습 하는데 있어 더 많은 고민이 필요할 것 같다.

프로젝트를 처음 시작할 땐 갈 길도 멀고 막막했는데, 팀원과 같이 협력하여 하나씩 차근차근 해나가다보니 이렇게 성과를 낼 수 있었다.



김영경

- 데이터 수집
- 데이터 설계
- 데이터 분석 및 시각화
- ppt 제작



김창균

- 데이터 수집
- 데이터 설계
- 데이터 분석 및 시각화
- ppt 제작



송찬의

- 데이터 수집
- 데이터 설계, 의견 조율
- 데이터 분석 및 시각화
- ppt 제작



신진섭

- 데이터 수집
- 데이터 설계
- 데이터 분석 및 시각화
- ppt 제작

Q&A

THANK YOU!

DATA ANALYSIS POWERPOINT

Team 'K2S2'

