

데이터 마이닝 프로젝트 발표자료

House Much?

머신 러닝을 통한 서울시 집값 예측

소프트웨어학과

2015111182 김성현

01

프로젝트 소개와 목표

프로젝트 소개

서울 시 내의 집들의 매매가와 집 정보, 도시 정보 등을 수집하여 학습을 시킨 후, 회귀를 통하여 집값을 예측하는 프로젝트

프로젝트 목표

- 주변 인프라와 주택 정보를 토대로 한 가격 예측을 통한 부동산 가격 예측
- 부동산 가격에 영향을 미치는 요인 분석

데이터 수집



부동산 114에서 데이터 크롤링
서울시 아파트, 도시형생활주택의 매매가 수집
각종 아파트 정보 수집

서울특별시 ▼	<input checked="" type="checkbox"/> 아파트 <input checked="" type="checkbox"/> 도시형생활주택	두산104동
---------	---	--------

42,000 만원

 알짜매물
19.12.06


아파트 방3개 90.95A/66.6㎡ 4층/총15층

T.951 7272 살수록 정드는 살아보고픈 집입니다

서울특별시 노원구 상계동

 우성공인중개사사
무소

02)951-7272

홈페이지

1:1문의

테스트 0

집 정보만 가지고 가격 예측 시도

두산 104동 (매매) 42,000 만원	일파매물 19.12.06		아파트 방3개 90.95A/66.6㎡ 4층/총15층 1.951.7272 할수록 정도는 살아 보고픈 집입니다 서울특별시 노원구 상계동	우성공인중개사사 무소 02)951-7272 홈페이지 1:1문의
자양5차현대 502동 (매매) 135,000 만원	일파매물 19.12.06		아파트 방4개 148.76/114.64㎡ 고층/총23층 남향.로얄층.한강영구조망권.수리원.전세5.7억가고.2020 년10완기 서울특별시 광진구 자양동	세종부동산 02)444-9258 홈페이지 1:1문의
수락파크빌 504동 (매매) 60,000 만원	일파매물 19.12.04		아파트 방3개 109.09A/84.62㎡ 1층/- 저층이아도 햇살이 밝은집. 공기청정아파트.지하철1분초 역세권 서울특별시 노원구 상계동	금빛공인 02)937-5050 홈페이지 1:1문의
장미1차8동 (매매) 225,000 만원	일파매물 19.12.04		아파트 방5개 150.05/120㎡ 고층/총14층 46입주가능한 특물수리남향. 단시중알로알동층 서울특별시 송파구 신천동	제일부동산중개법 인알앤유 02)422-9800 홈페이지 1:1문의
한솔리베르 1동 (매매) 95,000 만원	일파매물 19.12.04		아파트 방3개 109.09A/83.29㎡ 16층/총25층 남향.로얄층.한강조망권.상대강호.입주예물 서울특별시 광진구 자양동	세종부동산 02)444-9258 홈페이지 1:1문의
수락파크빌 503동 (매매) 65,000 만원	일파매물 19.12.03		아파트 방4개 145.45/114.82㎡ 2층/총15층 공기청정아파트. 깨끗이여 입주함의 서울특별시 노원구 상계동	금빛공인 02)937-5050 홈페이지 1:1문의
수락파크빌 503동 (매매) 71,000 만원	일파매물 19.12.03		아파트 방4개 145.45/114.82㎡ 6층/총15층 남향.전망좋은집. 깨끗함. 서울특별시 노원구 상계동	금빛공인 02)937-5050 홈페이지 1:1문의

BeautifulSoup 이용하여
해당 부분만 크롤링

데이터

1. 공급면적
2. 전용면적
3. 방 개수
4. 층

회귀 방식 : Linear Regression

데이터 셋의 개수 : 15000개

03

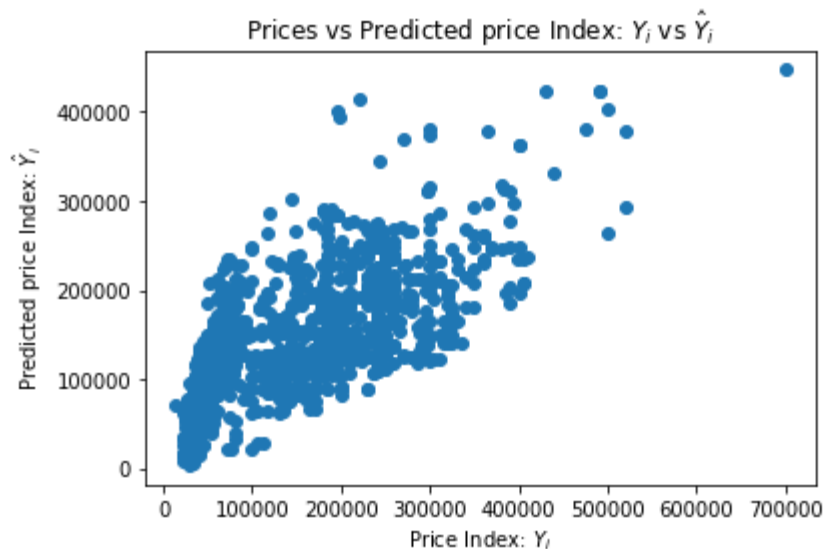
테스트 0 결과 분석

```
<class 'statsmodels.regression.linear_model.RegressionResultsWrapper'>
-----
OLS Regression Results
-----
Dep. Variable:      price      R-squared (uncentered):
0.825
Model:              OLS      Adj. R-squared (uncentered):
0.824
Method:             Least Squares      F-statistic:
1654.
Date:               Sat, 07 Dec 2019      Prob (F-statistic):
0.00
```

1. R-Squared(uncetered)
Y의 평균을 임의적으로 0으로
설정하여 계산한 값

Package's Rsquared : 0.4904033387087402

2. 실제 R-Squared 값은
0.4904



3. Visuallization

X축 : 실제 가격

Y축 : 예측 가격

Y=x 그래프처럼 나오면 잘 예
측한 모델

테스트 1

아파트 단지 정보 추가 한 다음 가격 예측 시도


소재지	서울특별시 노원구 상계동 1110 도로명주소		
공급면적	59.95㎡, 71.15㎡, 78.84㎡, 80.58㎡, 80.58㎡, 82.39㎡, 90.95㎡, 92.08㎡, 106㎡		
단지구모	총 11개동 763가구	층수	총 12층 ~ 15층
주차대수	총 486대 (가구당 0.6대)	난방정보	지역난방, 열병합
입주일	1994.10.01	용적률	224.71%
건폐율	16.12%	내진설계	의무적용 대상 ?
건설회사	두산건설(주)		

데이터

1. 이전 데이터
2. 주차 대수
3. 아파트 층 수
4. 아파트 연식
5. 가까운 지하철과의 거리

회귀 방식 : Linear Regression

데이터 셋의 개수 : 1400개

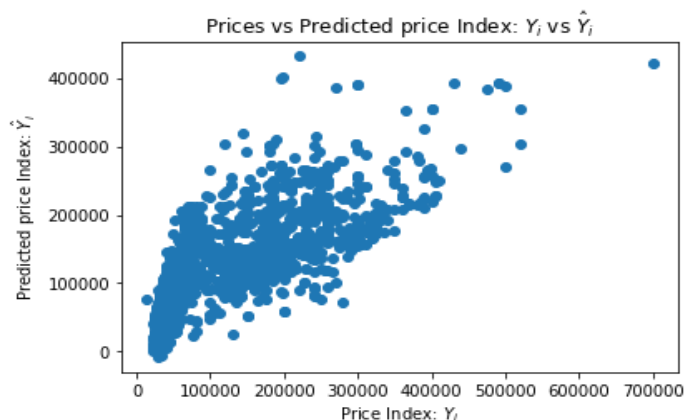
 지하철 ● 7호선 수락산 3출구 도보 8 분 (도로기준 약 523m)

테스트 1 결과 분석

OLS Regression Results

```
=====
Dep. Variable:          price  R-squared (uncentered): 0.840
Model:                  OLS    Adj. R-squared (uncentered): 0.839
Method:                 Least Squares  F-statistic: 1050.
Date:                   Sat, 07 Dec 2019  Prob (F-statistic): 0.00
Time:                   09:50:10  Log-Likelihood: -17673.
No. Observations:       1410  AIC: 3.536e+04
Df Residuals:           1403  BIC: 3.540e+04
Df Model:                7
Covariance Type:        nonrobust
=====
```

Package's Rsquared : 0.5218535337957446



그 전 결과보다 R-squared 값이 증가하였고, 그래프도 $y=x$ 그래프의 형태와 비슷하게 가고 있음을 알 수 있다.

03

테스트 2

아파트 정보보다 아파트 소재지에 따라서 아파트 값의 변동이 커짐을 인지

도시의 발전 지수 : 버거지수로 측정

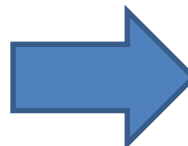
한 도시의 발전 수준은 (버거킹의 갯수+맥도날드의 갯수+KFC의 갯수)/롯데리아의 갯수를 계산하여
높게 나올수록 더 발전된 도시라고 할 수 있다

🔍 📊 ⭐ ...

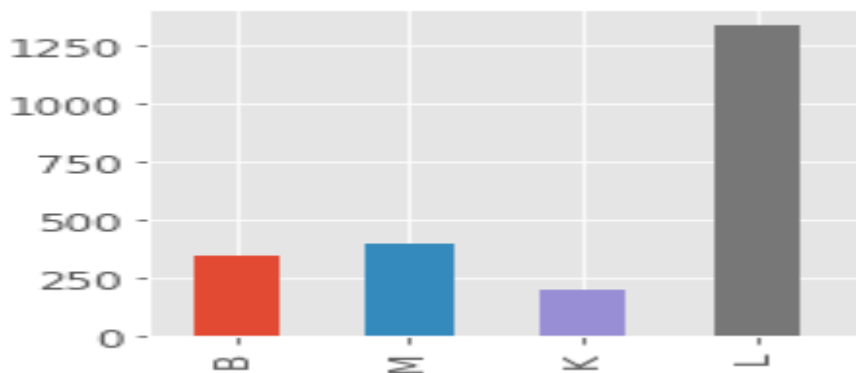
RETWEETS 651
FAVORITES 79



4:17 AM - 21 Sep 2014



$$\xi = \frac{B+M+K}{L}$$



테스트 2



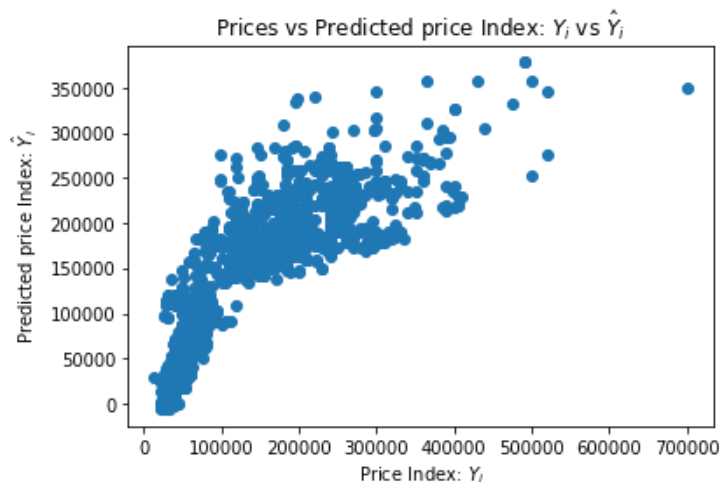
03

테스트 2 결과 분석

```
<class 'statsmodels.regression.linear_model.HegressionResultsWrapper'>
OLS Regression Results
```

```
=====
Dep. Variable:          price  R-squared (uncentered):
0.902
Model:                  OLS  Adj. R-squared (uncentered):
0.901
Method:                 Least Squares  F-statistic:
1427.
Date:                   Sat, 07 Dec 2019  Prob (F-statistic):
0.00
Time:                   10:12:31  Log-Likelihood:
-17329.
No. Observations:       1410  AIC:
468e+04
Df Residuals:           1401  BIC:
472e+04
Df Model:                9
Covariance Type:        nonrobust
=====
```

Package's Rsquared : 0.7319837660936812



전체적으로
예측 값이 정확해 지고 있는 모습

03

테스트 3

구 단위 말고도 동 단위도 집 가격에 큰 영향을 줌을 인지
동 단위 데이터도 입력

소재지

서울특별시 노원구 상계동 1110

도로명주소

전처리를 통하여 해당하는 동에 1, 나머지는 0으로 표기

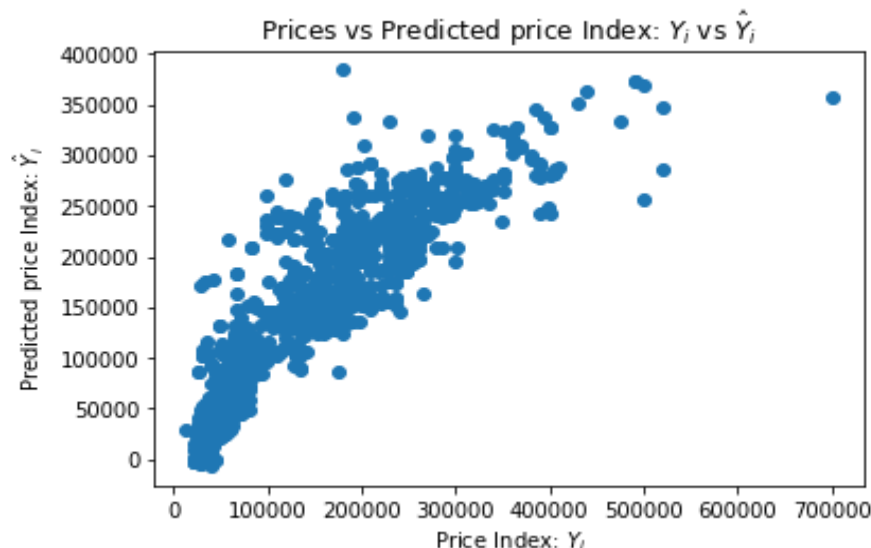
samsung	ilwon	daechi	yeoksam	cheongdam	suseo	nonhyeon	dogok	apgujeong	segok	floor	burger	sanggye	wolgye	junggye	hagye	gongreung	gasan	doksan	siheung
0	1	0	0	0	0	0	0	0	0	1	3.33333	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0	3	3.33333	0	0	0	0	0	0	0	0

테스트 3 결과 분석

```
<class 'statsmodels.regression.linear_model.RegressionResultsWrapper'>
OLS Regression Results
```

Dep. Variable:	price	R-squared:	0.817
Model:	OLS	Adj. R-squared:	0.814
Method:	Least Squares	F-statistic:	309.6
Date:	Sat, 07 Dec 2019	Prob (F-statistic):	0.00
Time:	10:18:19	Log-Likelihood:	-16995.
No. Observations:	1410	AIC:	3.403e+04
Df Residuals:	1389	BIC:	3.414e+04
Df Model:	20		
Covariance Type:	nonrobust		

R-square는 약 8.2 정도로
아까의 0.72와 비교하였을
때, 좋아졌음을 알 수 있었
고 그래프도 $y=x$ 그래프와
비슷해짐을 보임.

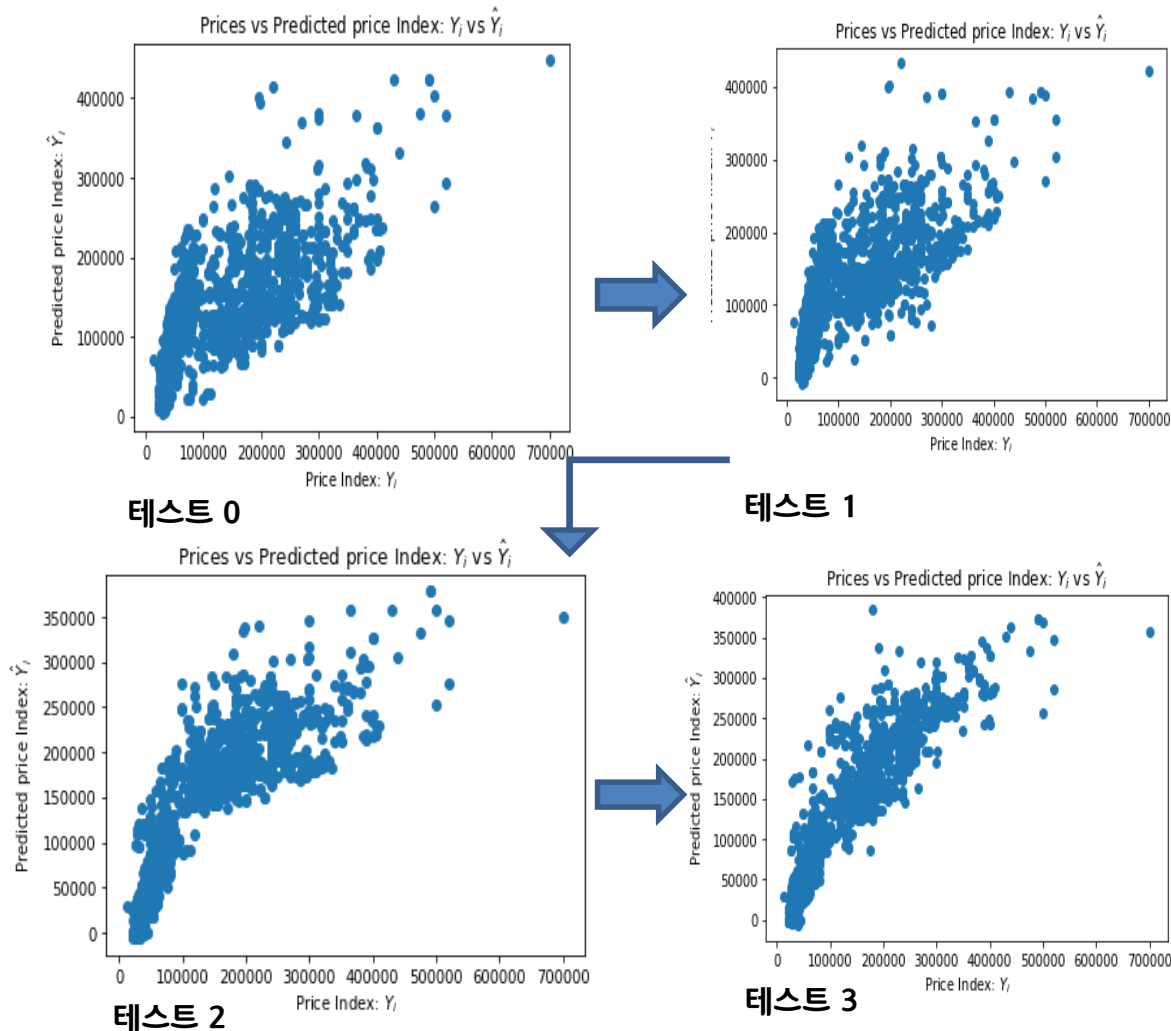


Coefficient 분석

	coef	std err	t	P> t	[0.025	0.975]
larea	3747.1821	465.972	8.042	0.000	2833.097	4661.267
barea	-530.9813	568.407	-0.934	0.350	-1646.010	584.047
age	-666.0931	170.339	-3.910	0.000	-1000.242	-331.944
dist	10.3312	4.146	2.492	0.013	2.198	18.465
units	20.7008	3.919	5.282	0.000	13.013	28.389
room	-3034.9337	2536.014	-1.197	0.232	-8009.765	1939.898
samsung	-3669.1379	4642.194	-0.790	0.429	-1.28e+04	5437.330
burger	1.968e+04	2574.537	7.644	0.000	1.46e+04	2.47e+04
ilwon	7595.5156	9277.817	0.819	0.413	-1.06e+04	2.58e+04
daechi	6.487e+04	3759.161	17.256	0.000	5.75e+04	7.22e+04
yeoksam	-2.852e+04	4338.604	-6.574	0.000	-3.7e+04	-2e+04
cheongdam	-2256.2047	5882.800	-0.384	0.701	-1.38e+04	9283.927
suseo	-1.818e+04	1.35e+04	-1.344	0.179	-4.47e+04	8359.166
nonhyeon	-5.756e+04	7580.251	-7.593	0.000	-7.24e+04	-4.27e+04
dogok	-1.168e+04	4544.479	-2.570	0.010	-2.06e+04	-2763.319
apgujeong	1.251e+05	1.47e+04	8.541	0.000	9.64e+04	1.54e+05
segok	1.617e-11	2.07e-12	7.797	0.000	1.21e-11	2.02e-11
sanggye	-5.172e+04	6388.983	-8.096	0.000	-6.43e+04	-3.92e+04
wolgye	-3.938e+04	7561.769	-5.207	0.000	-5.42e+04	-2.45e+04
junggye	-5.865e+04	6944.995	-8.445	0.000	-7.23e+04	-4.5e+04
hagye	-5.535e+04	6830.593	-8.104	0.000	-6.88e+04	-4.2e+04
gongreung	-5.352e+04	7400.545	-7.232	0.000	-6.8e+04	-3.9e+04
floor	489.7591	209.716	2.335	0.020	78.365	901.153

오래된 건물일 수록 가
격 낮아지고, 강남이 노
원보다 비싸다.

결과 최종 분석



Package's Rsquared : 0.4904033387087402



Package's Rsquared : 0.5218535337957446



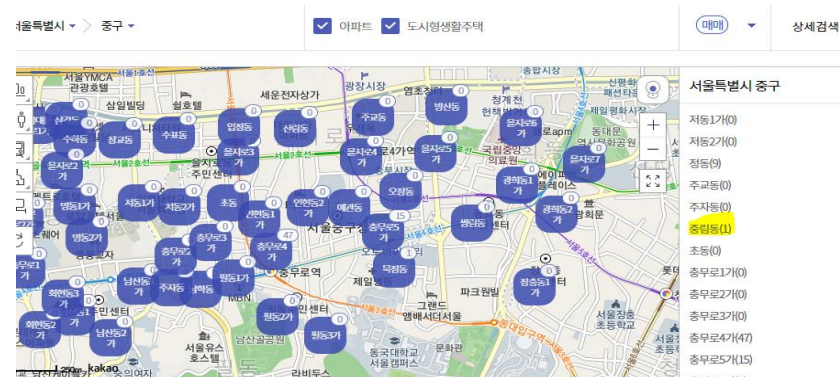
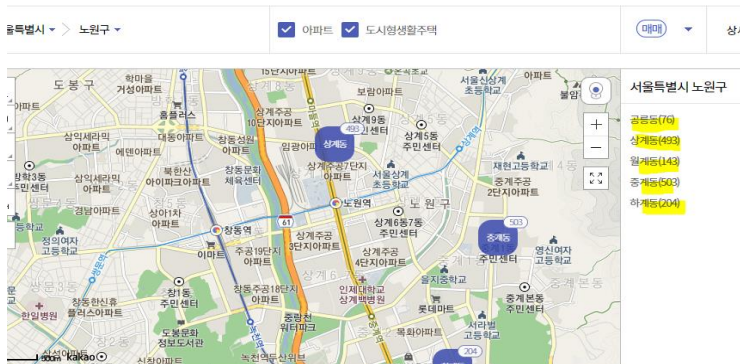
Package's Rsquared : 0.7319837660936812



R-squared:	0.817
Adj. R-squared:	0.814
F-statistic:	900.6

문제점

1. 데이터 크롤링의 시간과 사이트의 정보 누락으로 인해 데이터가 생각보다 많이 모이지 않았음
2. 사이트의 형식이 일정하지 않아서 데이터 전처리 과정에서 문제점이 있었음
3. 사이트의 일정하지 않은 '동 단위'
4. 겹치는 데이터(larea: 공급면적, barea: 전용면적, 공급면적이 커지면 전용면적이 커짐)로 인한 coefficient가 직관과 다르게 나옴



참고사이트

버거지수 2019:

<http://blog.naver.com/PostView.nhn?blogId=idjoopal&logNo=221519294269&parentCategoryNo=68&categoryNo=&viewDate=&isShowPopularPosts=true&from=search>

알고리즘/ 라이브러리

선형 회귀 분석 사용

Regression & 결과 출력 : OLS

느낀점

1. 데이터 Regression 이나 Classification 같은 경우 파이썬의 많은 라이브러리에
서 학습을 지원해줌을 알았다. 결국 중요한 것은 얼마나 유의미한 데이터를 수집
하는가, 그리고 많은 Regression 이나 Classification 모델 중 적절한 모델을 찾
는가 이 두 요소가 좋은 결과 값을 가져다 준다는 것을 알게 되었다.
2. 데이터 수집에 굉장히 많은 노력이 필요함을 알게 되었다. 사이트마다 일정한 형
식으로 되어 있는 것이 아니기 때문에 예외 처리도 굉장히 많이 하여야 했고, 전
처리 과정도 많이 겪어야 했다. 그리고 적절한 데이터를 찾는 것이 좋은 결과를
보여준다는 것도 알게 되었다.
3. 서울시.... 집 값이 많이 비싸다는 것을 알게 되었다.....

**THANK
YOU**