


() 가

- Connecticut Hartford -

2

2021155151

- 
1. &
 2. &
 - 3.
 - 4.
 5. &

	데이터	변수(Type)	최고 성능
House Price Prediction via Improved Machine Learning Techniques(2019)	Housing Price in Beijing (2009~2018) [231962행, 19열]	위치 → 5개 재무 → 4개 건물 → 10개	Stacked Reg → RMSE 0.16350
House Prices and relative location(2019)	City of Oslo(2007-2015) [40,019행,8열]	위치 → 8개 (위치 변수에 대한 공간적 분석, 통계적 전처리)	Adjusted R2 → 0.815 (예측 X, 요인분석)
Using Machine Learning Algorithms for Housing Price Prediction: The Case of Islamabad Housing Data(2021)	크롤링 진행 [행 개수 불명,22열]	건물 → 13개 위치 → 9개 재무 → 1개	SVM Reg → MAPE 1918.4957
House Price Prediction Based on Multiple Linear Regression(2021)	Boston Housing Prices [506행,13열]	경제→2개,환경→1개,재무→2개 범죄→1개,인구→1개,건물→2개 위치→4개	Linear Reg → 기록X (그림으로만 설명)
Housing Prices Prediction with a Deep Learning and Random Forest Ensemble(2019)	크롤링 진행 [12,223,582행,24열]	이미지→1개,시간→2개,건물→10개 위치→7개,재무→3개	RF Reg → RMSE 0.23847
House Price Prediction using a Machine Learning Model: A Survey of Literature(2020)	X [리뷰 논문]	지역 > 건물 > 경제 변수 순서로 주택 가격의 영향력 정렬	ANN(지역변수) →RMSE 0.0581

-
- 가 (, ,)가 가 (ex. , ,)
- & &



-
-
-

가
가

가

-
-
-

Connecticut Harford
가

가 6

1. [Real Estate Sales 730 Days | Data | City of Hartford](#) → Hartford
2. API → Socrata Open Data API [JSON] → Sodapy → Socrate
3. Hartford → Sodapy → 가

HartfordData

Catalog Developers Help Sign In

Real Estate Sales 730 Days

Financial

These data represents City of Hartford real estate sales for the past 2 years from the current date. Updated nightly

About this Dataset

Updated
November 24, 2021

Data Last Updated November 23, 2021 Metadata Last Updated November 24, 2021

Date Created August 26, 2020

Views Downloads

Topics

Category	Financial
Tags	hartford, real estate, assessor

Licensing and Attribution

License

1. API

Access this Dataset via SODA API

The Socrata Open Data API (SODA) provides programmatic access to this dataset including the ability to filter, query, and aggregate data.

2. API Docs Developer Portal

API Endpoint

<https://data.hartford.gov/resource/3cs2-w2b6> JSON Copy

jQuery Python Pandas PowerShell RSocrata SAS soda-ruby SC

Python package using Pandas to easily work with JSON data

```
#!/usr/bin/env python

# make sure to install these packages before running:
# pip install pandas
# pip install sodapy

import pandas as pd
from sodapy import Socrate

# Unauthenticated client only works with public data sets. Note 'None'
# in place of application token, and no username or password:
client = Socrate("data.hartford.gov", None)

# Example authenticated client (needed for non-public datasets):
# client = Socrate(data.hartford.gov,
#                  MyAppToken,
#                  username="user@example.com",
#                  password="AFakePassword")

# First 2000 results, returned as JSON from API / converted to Python list of
# dictionaries by sodapy.
results = client.get("3cs2-w2b6", limit=2000)

# Convert to pandas DataFrame
results_df = pd.DataFrame.from_records(results)
```

```

from sodapy import Socrata
client = Socrata("data.hartford.gov", None)
results = client.get("3cs2-w2b6", limit=2835) # API 공유키 (data.hartford.gov에서 공공데이터 한정으로 제공 )

# Convert to pandas DataFrame
import pandas as pd
df = pd.DataFrame.from_records(results)
df.head()

WARNING:root:Requests made without an app_token will be subject to strict throttling limits.

```

stname	primarygrantor	saledate	saleprice	totalappraisedvalue	legalreference	xrsalesvalidityid	xrdeedid	ownerfirstname	apartmentunitnumber
P & P EMENT #2 LLC	P & P MANAGEMENT LLC	2019-11- 25T00:00:00.000Z	0	151700	07554-0275	5	8	NaN	NaN
GRON ORTIZ	MARTINEZ- CRUZ MELBIN ELI	2019-11- 25T00:00:00.000Z	156000	135500	07554-0299	1	10	BRENDA	NaN
IPBELL	CAMPBELL CLAUDETTE A	2019-11- 25T00:00:00.000Z	1	185600	07554-0220	5	8	CLAUDETTE	NaN
IPBELL	ROBERTS CLAUDETTE	2019-11- 25T00:00:00.000Z	0	185600	07554-0219	5	13	CLAUDETTE	NaN
DEAN	RODRIGUEZ DANIEL R	2019-11- 25T00:00:00.000Z	229000	162900	07554-0324	1	10	RAYVON	NaN

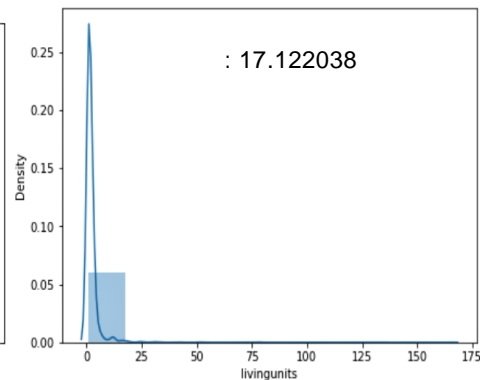
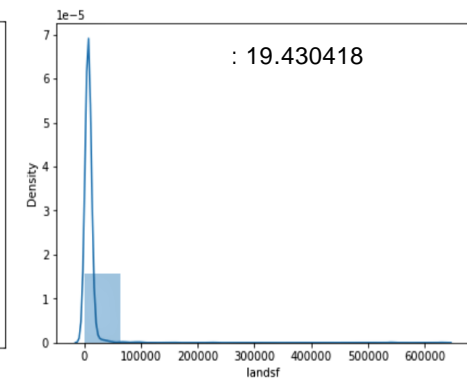
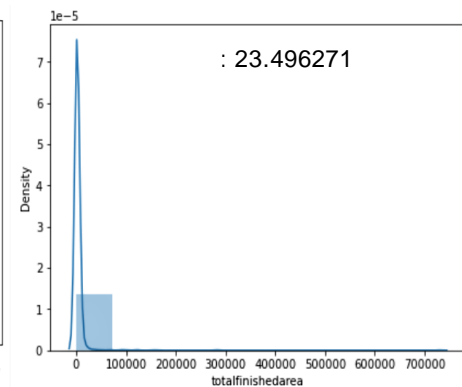
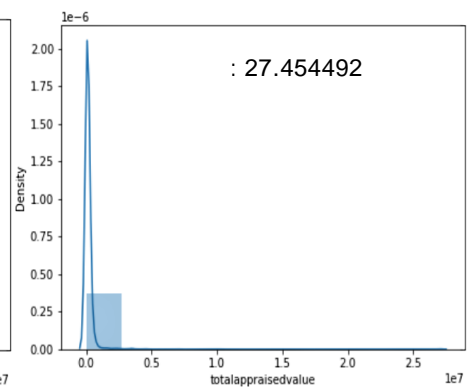
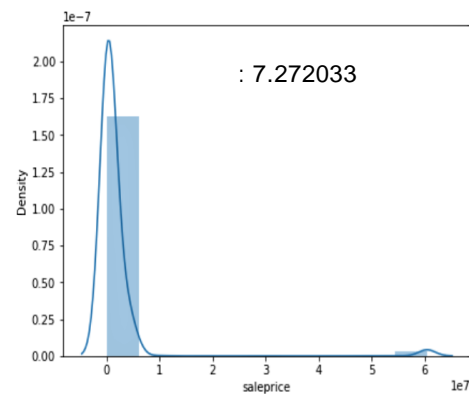
- 2019.11 ~ 2021.01 730 () 가 [2835 20]

- 5 , 15 [가]
- & [ex. saleprice = 1,10,100]
-

-
- à Living Units(),
- à [2835,20] [2835,504]

	변수	정의	Type
종속변수	Sale Price	주택(아파트) 매도 가격	float64
독립변수	Living units	주거단위 (아파트, 연립주택에서 그 전체를 구성하는 가구 수)	float64
	Total Appraised Value	감정평가액	float64
	Total Finished Area	주택(아파트) 면적	float64
	Land SF	토지 면적	float64
	Street Name and Way	도로명	object
	XrComposite Land Use ID	토지 이용 ID(종류)	object
	XrBuilding Type ID	건물 ID(종류)	object
	Sale Date	매도 날짜	object

	landsf	totalfinishedarea	livingunits	saleprice	totalappraisedvalue
count	2835.000000	2835.000000	2835.000000	2.835000e+03	2.835000e+03
mean	9564.044797	3972.886025	2.229982	1.773713e+06	1.838580e+05
std	25983.355077	23490.287584	5.575492	7.729612e+06	8.053519e+05
min	432.000000	239.000000	1.000000	2.100000e+01	3.300000e+03
25%	7000.000000	697.500000	1.000000	2.132500e+05	4.000000e+04
50%	7250.000000	1353.500000	1.000000	2.550000e+05	1.221000e+05
75%	7500.000000	3222.000000	2.000000	3.500000e+05	1.691500e+05
max	630313.000000	730457.920000	165.000000	6.050000e+07	2.702970e+07



1.

가 O

- (Baseline Model)
- L1(Lasso) Penalty à alpha : 1
- L2(Ridge) Penalty à alpha : 0.001
- ElasticNet(L1+L2) à alpha : 0.05

가 X

- Reg à n_estimators=100,max_depth=10
- GBM Reg à n_estimators=100,max_depth=10 ,subsample =0.8
- XGBoost Reg à n_estimators=100,max_depth=10

2. 가

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}} \quad \ddot{u} \quad ()$$

$$MSE = \frac{1}{n} \sum (y - \hat{y})^2$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - x|$$

$$MAPE = \frac{100\%}{n} \sum \left| \frac{y - \hat{y}}{y} \right| \quad \ddot{u} \quad ()$$

Baseline

&

1. Stepwise

(P-value < 0.05)

```
X = df.drop('saleprice',axis=1,inplace=False)
y = df['saleprice']
```

```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.3,random_state=2021)
```

```
step_features = ['Month 12','streetnameandway SOUTH MARSHALL ST','xrcompositelal
len(step_features) # 503 → 77개로 1차 변수 축소 (속도증가 & 과적합 방지 목적)
```

OLS Regression Results

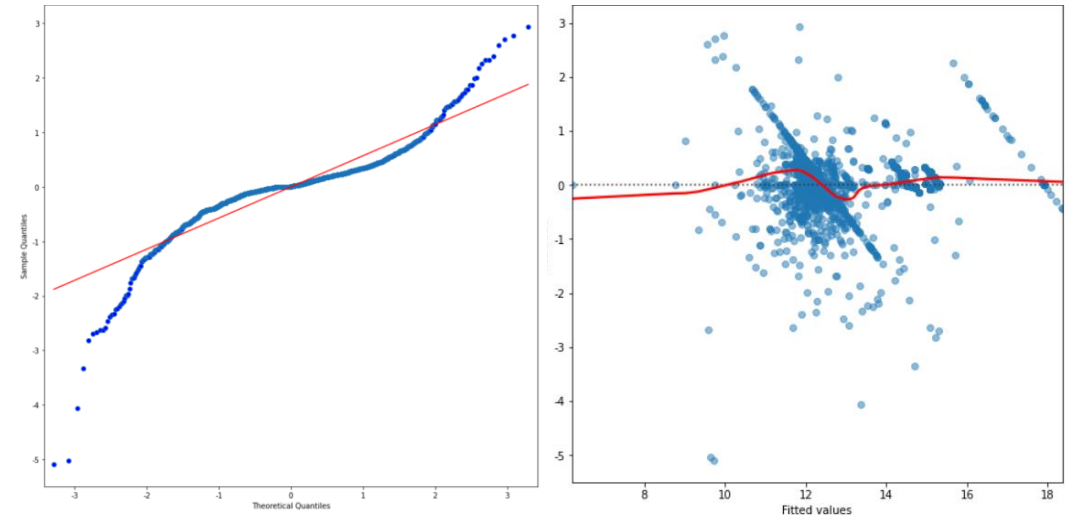
Dep. Variable:	saleprice	R-squared:	0.806
Model:	OLS	Adj. R-squared:	0.798
Method:	Least Squares	F-statistic:	111.6
Date:	Mon, 29 Nov 2021	Prob (F-statistic):	0.00
Time:	19:55:16	Log-Likelihood:	-1867.0
No. Observations:	1984	AIC:	3878.
Df Residuals:	1912	BIC:	4281.
Df Model:	71		
Covariance Type:	nonrobust		

0.788 → 0.798

가

2.

&



3. L1,L2,ElasticNet Penalty

→ Grid Search

4. RF, GBM, XG-Boost

→ Grid Search

(K=10 Fold)

1. Baseline

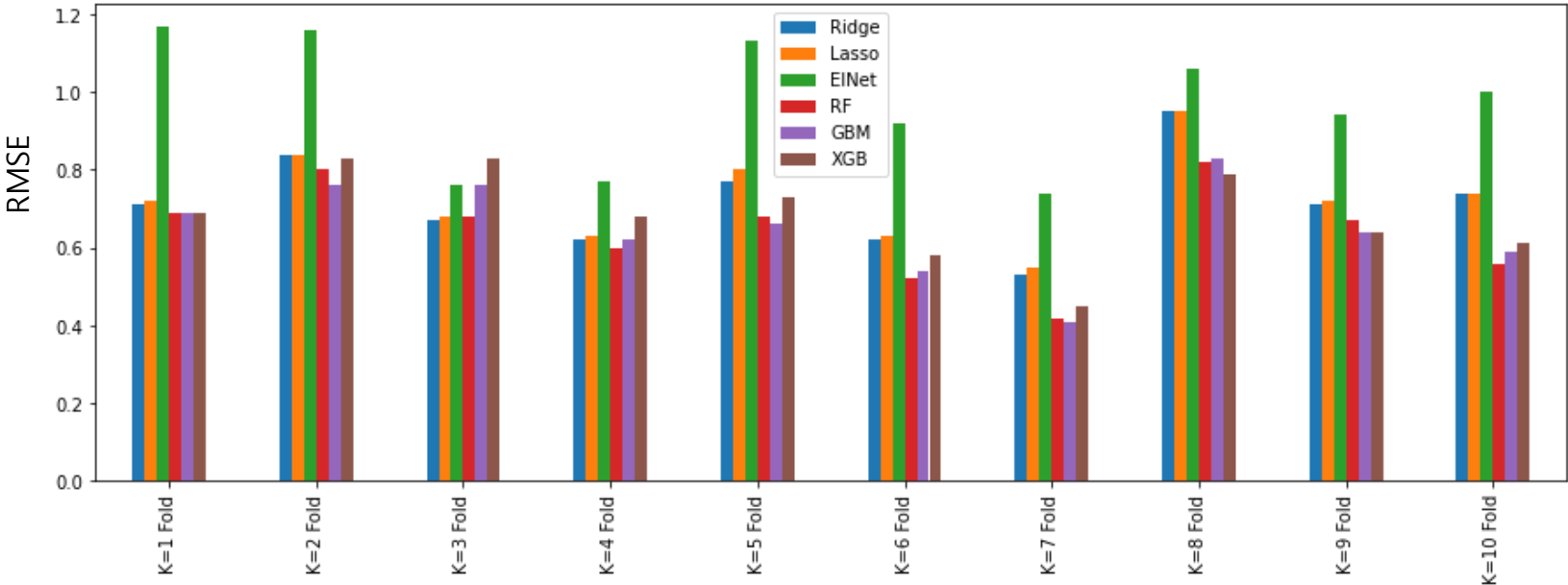
LinearRegression K=10 RMSE List:[3.53829671e+07 8.40000000e-01 5.16760965e+10 3.59974457e+06
3.52617449e+07 1.38568000e+03 5.30000000e-01 1.34107332e+11
1.26423055e+06 5.13198000e+03]

LinearRegression K=10 Mean RMSE List:18585894378.3

1.

- RMSE

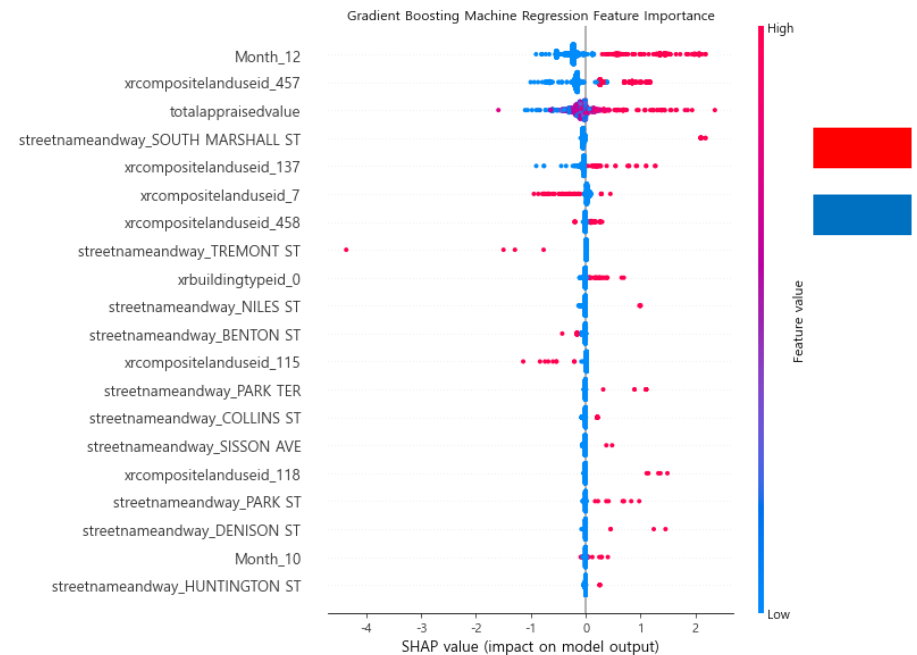
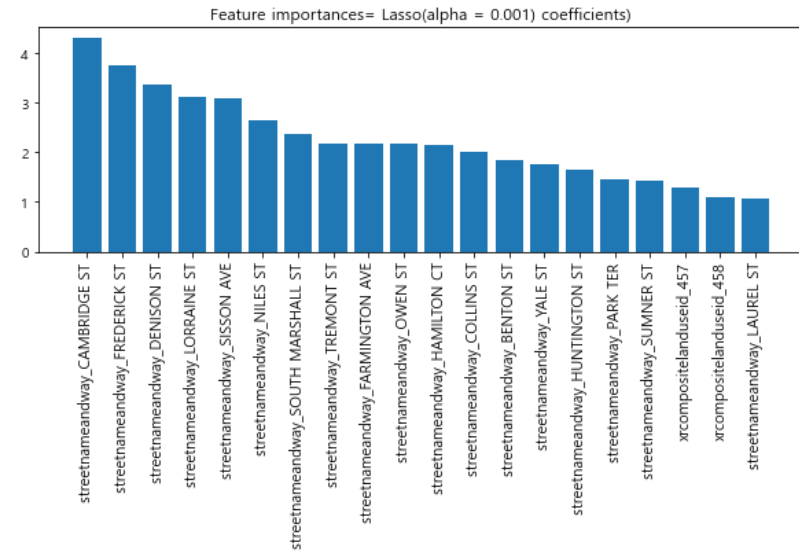
	mean RMSE	rmse std
Ridge	0.716	0.119648
Lasso	0.726	0.115682
ElNet	0.965	0.166950
RF	0.644	0.122583
GBM	0.650	0.121381
XGB	0.683	0.119727



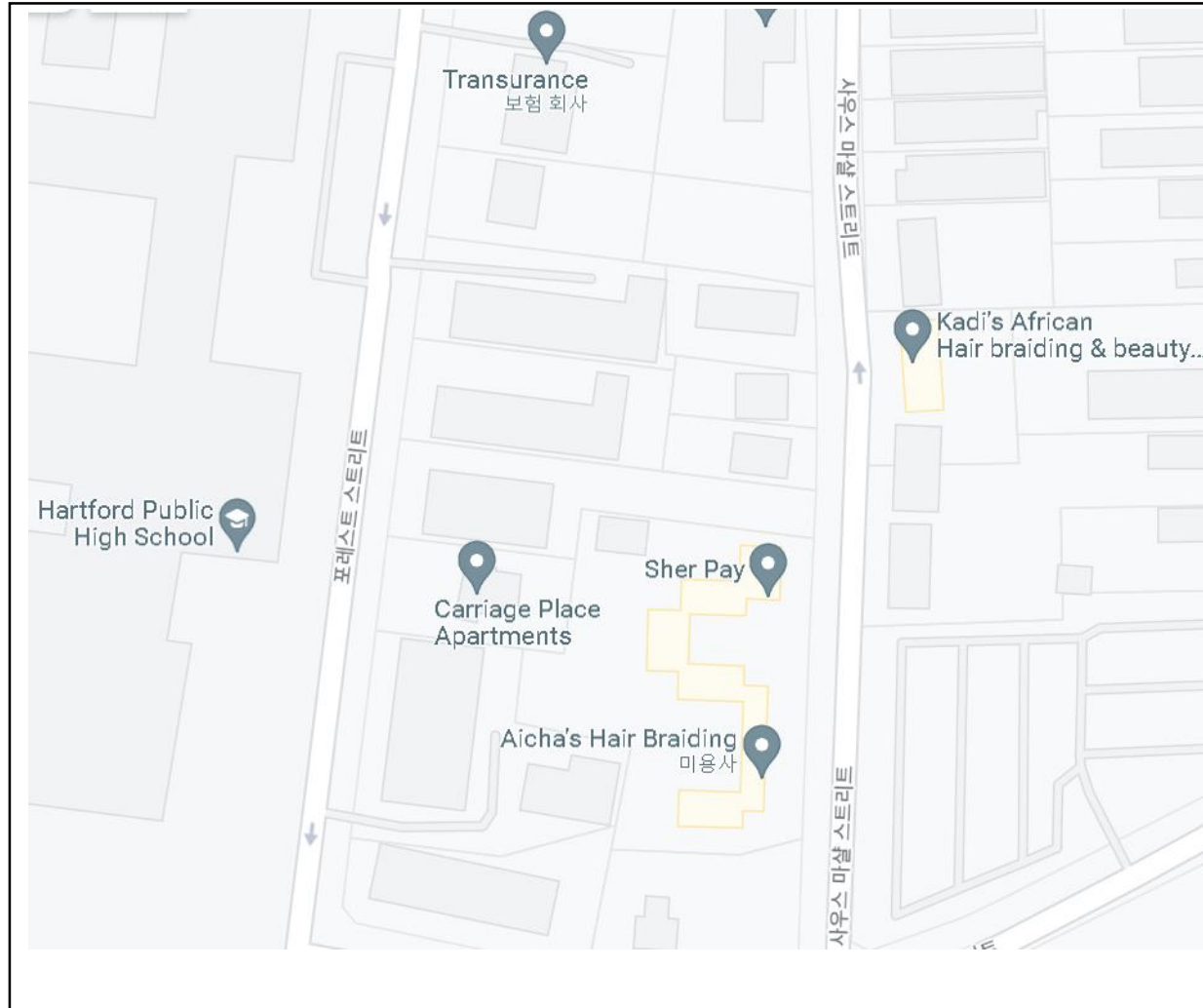
1. RMSE & MAPE

	MSE	RMSE	MAE	MAPE
Ridge	0.50	0.71	0.45	3.88
Lasso	0.48	0.70	0.45	3.85
El_Net	0.89	0.94	0.64	5.24
RF	0.33	0.58	0.33	2.82
GBM	0.35	0.60	0.32	2.75
XGB	0.42	0.64	0.34	2.93

2.



. South Marshall_ST()



- , ,
-

가

- ()
-
- 가 ()
- 가

- [South Marshall_ST Carriage Place Apartment 1.5/5]
- (2835)
[가 4000 (CSV) 가 가 X]
- RMSE
- , Connecticut

- &
-
-