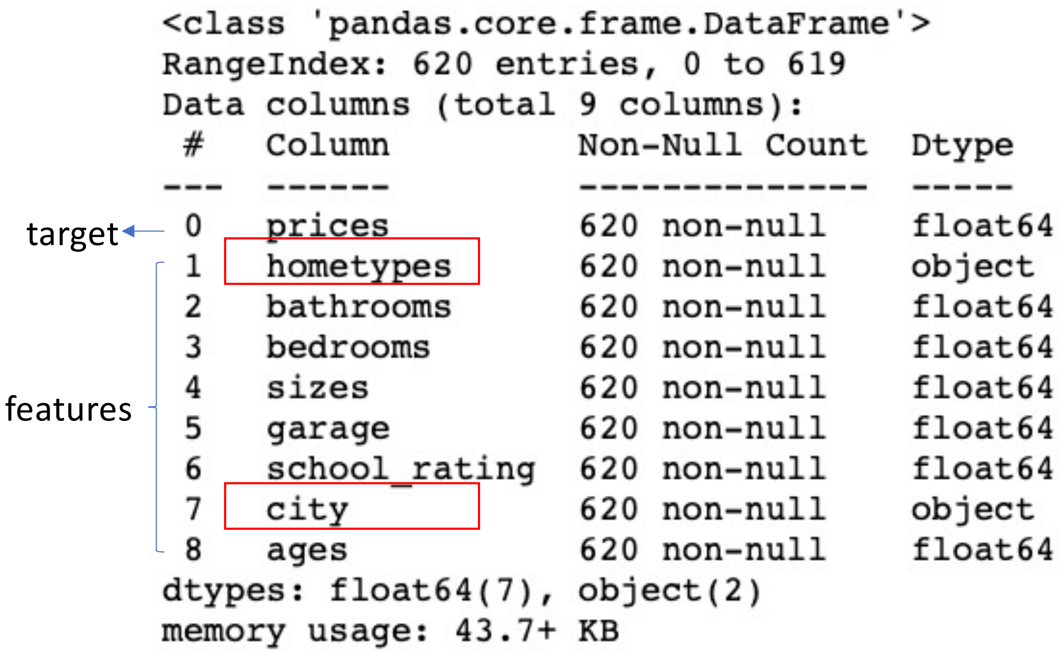
Housing price prediction based on Multiple Linear Regression

Ni-Ting Chiou

Data

- Get pages from the search results of Zillow website.
- Scrape features from multiple houses displayed in a page.
- Scrape more features by entering the links of the individual house scraped from pages.



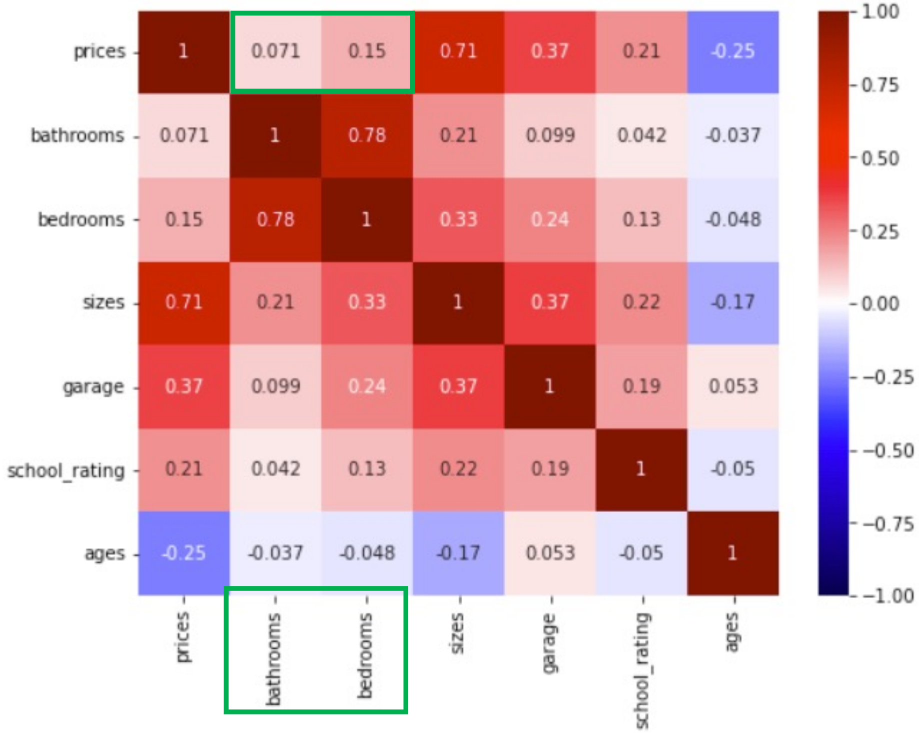
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 620 entries, 0 to 619
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   prices         620 non-null    float64
 1   hometypes      620 non-null    object 
 2   bathrooms      620 non-null    float64
 3   bedrooms       620 non-null    float64
 4   sizes          620 non-null    float64
 5   garage         620 non-null    float64
 6   school_rating  620 non-null    float64
 7   city           620 non-null    object 
 8   ages           620 non-null    float64
dtypes: float64(7), object(2)
memory usage: 43.7+ KB
```

target

features

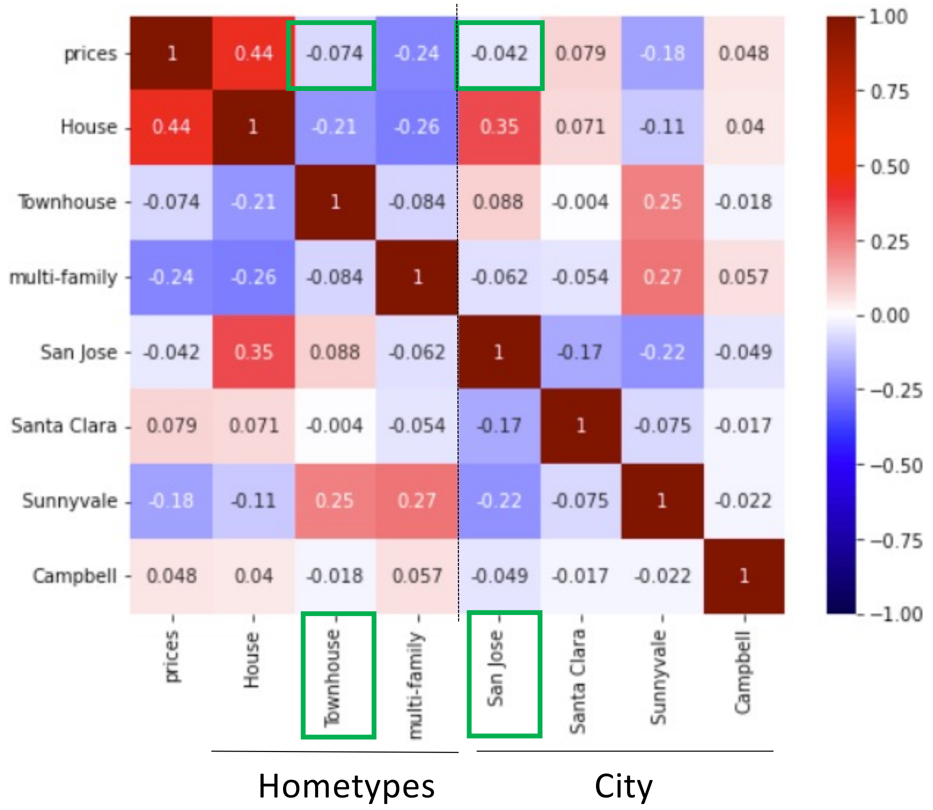
The Pearson correlation of features and prices

6 numeric variables



	prices	bathrooms	bedrooms	sizes	garage	school_rating	ages
prices	1	0.071	0.15	0.71	0.37	0.21	-0.25
bathrooms	0.071	1	0.78	0.21	0.099	0.042	-0.037
bedrooms	0.15	0.78	1	0.33	0.24	0.13	-0.048
sizes	0.71	0.21	0.33	1	0.37	0.22	-0.17
garage	0.37	0.099	0.24	0.37	1	0.19	0.053
school_rating	0.21	0.042	0.13	0.22	0.19	1	-0.05
ages	-0.25	-0.037	-0.048	-0.17	0.053	-0.05	1

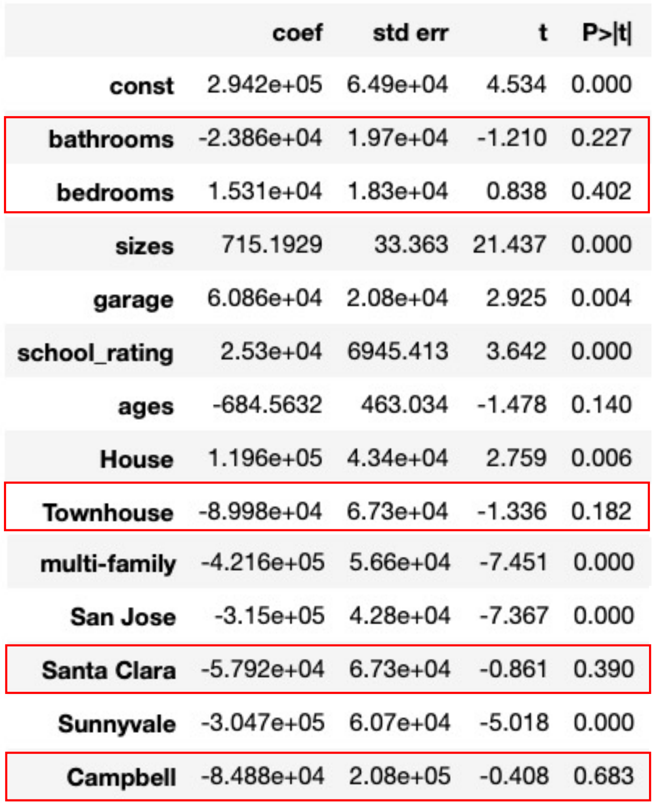
7 dummy variables



	prices	House	Townhouse	multi-family	San Jose	Santa Clara	Sunnyvale	Campbell
prices	1	0.44	-0.074	-0.24	-0.042	0.079	-0.18	0.048
House	0.44	1	-0.21	-0.26	0.35	0.071	-0.11	0.04
Townhouse	-0.074	-0.21	1	-0.084	0.088	-0.004	0.25	-0.018
multi-family	-0.24	-0.26	-0.084	1	-0.062	-0.054	0.27	0.057
San Jose	-0.042	0.35	0.088	-0.062	1	-0.17	-0.22	-0.049
Santa Clara	0.079	0.071	-0.004	-0.054	-0.17	1	-0.075	-0.017
Sunnyvale	-0.18	-0.11	0.25	0.27	-0.22	-0.075	1	-0.022
Campbell	0.048	0.04	-0.018	0.057	-0.049	-0.017	-0.022	1

Use standardized data for model development

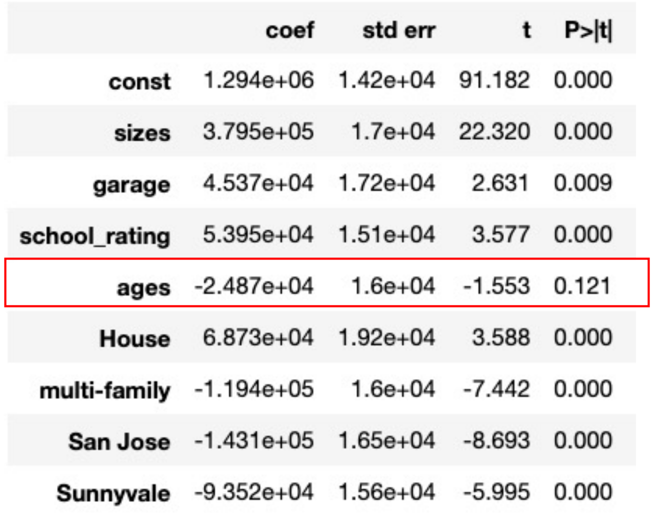
Model I (All features)



	coef	std err	t	P>|t|
const	2.942e+05	6.49e+04	4.534	0.000
bathrooms	-2.386e+04	1.97e+04	-1.210	0.227
bedrooms	1.531e+04	1.83e+04	0.838	0.402
sizes	715.1929	33.363	21.437	0.000
garage	6.086e+04	2.08e+04	2.925	0.004
school_rating	2.53e+04	6945.413	3.642	0.000
ages	-684.5632	463.034	-1.478	0.140
House	1.196e+05	4.34e+04	2.759	0.006
Townhouse	-8.998e+04	6.73e+04	-1.336	0.182
multi-family	-4.216e+05	5.66e+04	-7.451	0.000
San Jose	-3.15e+05	4.28e+04	-7.367	0.000
Santa Clara	-5.792e+04	6.73e+04	-0.861	0.390
Sunnyvale	-3.047e+05	6.07e+04	-5.018	0.000
Campbell	-8.488e+04	2.08e+05	-0.408	0.683

Adj. R-squared: 0.643

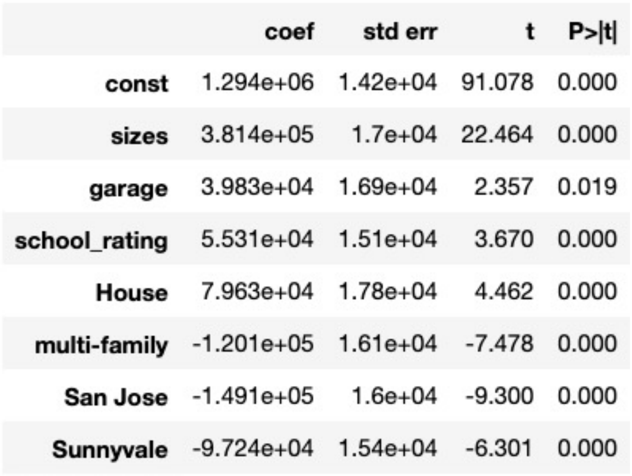
Model II (Remove high-P-value features)



	coef	std err	t	P>|t|
const	1.294e+06	1.42e+04	91.182	0.000
sizes	3.795e+05	1.7e+04	22.320	0.000
garage	4.537e+04	1.72e+04	2.631	0.009
school_rating	5.395e+04	1.51e+04	3.577	0.000
ages	-2.487e+04	1.6e+04	-1.553	0.121
House	6.873e+04	1.92e+04	3.588	0.000
multi-family	-1.194e+05	1.6e+04	-7.442	0.000
San Jose	-1.431e+05	1.65e+04	-8.693	0.000
Sunnyvale	-9.352e+04	1.56e+04	-5.995	0.000

Adj. R-squared: 0.644

Model III (Remove ages feature)

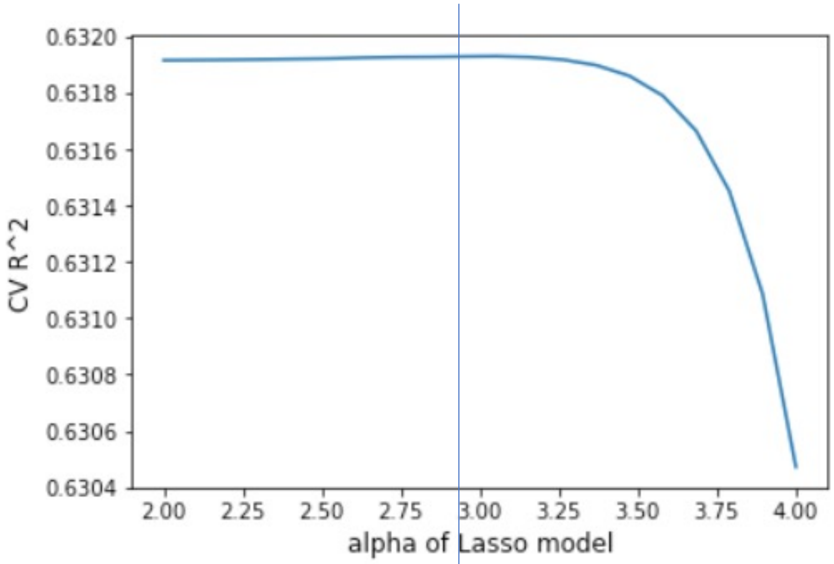


	coef	std err	t	P>|t|
const	1.294e+06	1.42e+04	91.078	0.000
sizes	3.814e+05	1.7e+04	22.464	0.000
garage	3.983e+04	1.69e+04	2.357	0.019
school_rating	5.531e+04	1.51e+04	3.670	0.000
House	7.963e+04	1.78e+04	4.462	0.000
multi-family	-1.201e+05	1.61e+04	-7.478	0.000
San Jose	-1.491e+05	1.6e+04	-9.300	0.000
Sunnyvale	-9.724e+04	1.54e+04	-6.301	0.000

Adj. R-squared: 0.643
Cross-validation R-squared: 0.632

Model III is further optimized by lasso regularization

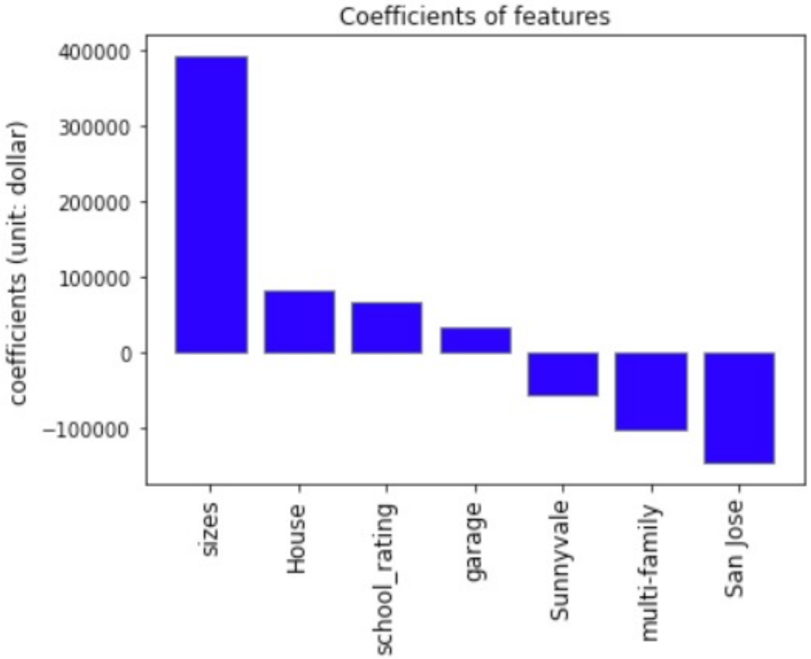
R^2 = 0.632



α = 1132

Lasso model

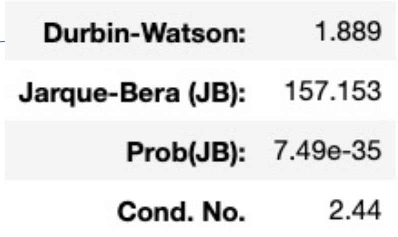
R^2 of train set: 0.646
R^2 of test set: 0.644
MAE of test set: $280,945



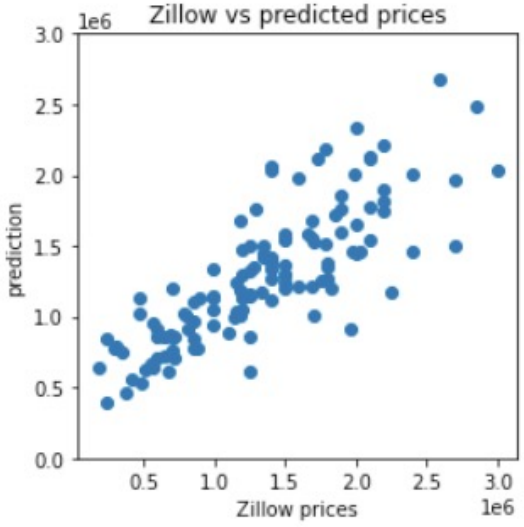
Constant: 1,200,000

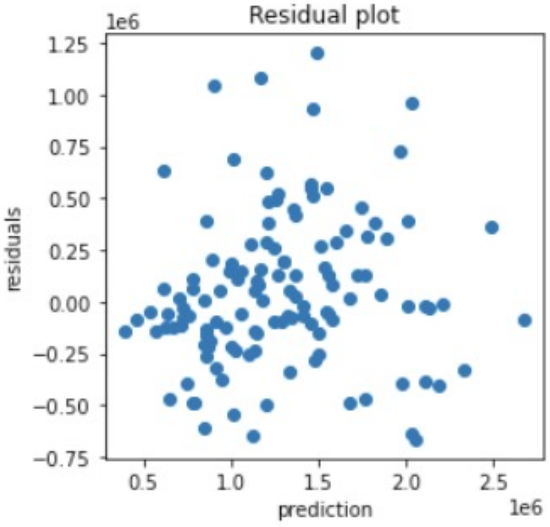
Final lasso model evaluation

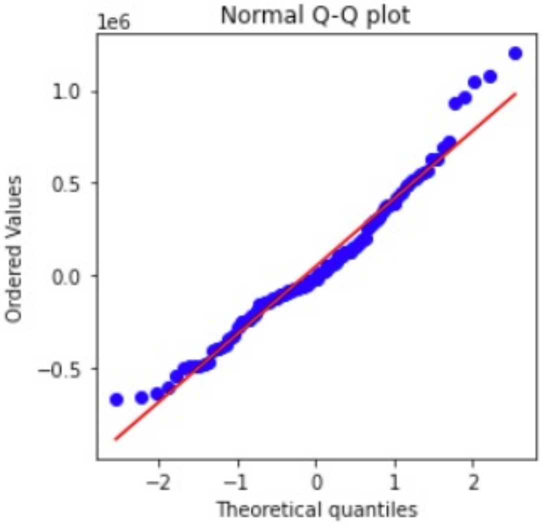
- Residuals are independent to each other
- Residuals are normally distributed



Durbin-Watson:	1.889
Jarque-Bera (JB):	157.153
Prob(JB):	7.49e-35
Cond. No.	2.44







Conclusion and discussion

Final Lasso Model

- The most significant factor that influences the housing prices is size (the model with size feature only has R^2 of 0.5).
- The R^ of the model with 7 features is 0.63 and MAE is ~$ 280,000.

Model improvement

- **Add more features:**
  - Most features scrapped from Zillow website are house factors. Other variables, such as transportation and environmental factors should also be considered.
- **Try other models:**
  - Other models, such as RandomForestRegressor, could be better for the housing price prediction.