SUPERVISED LEARNING PROJECT

Predicting Housing Price

Prepared by Jinny Kwon

OVERVIEW

BACKGROUND

Since the 2008 recession, which resulted in a 33% drop in U.S. housing prices, the real estate market has been steadily recovering.

While housing prices in the coastal cities (SF Bay Area, Seattle, NYC) have seen the most intense growth, the real estate markets in the Sun Belt cities (Austin, Raleigh, Atlanta, and Orlando) have also been steadily growing. Experts credit this growth to **an abundance of jobs, an increase in population, and a positive outlook of the economy**.

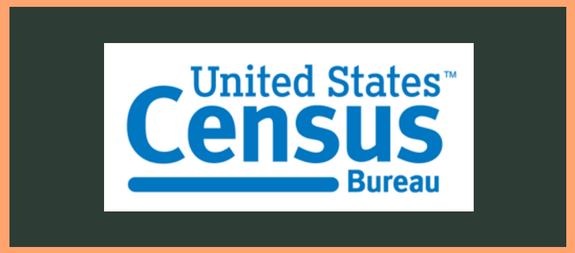
RESEARCH QUESTION

The recent growth in housing markets is due to a strong demand and a lack of supply. The cities that attract professionals will continue to see appreciation in the housing prices. Therefore, I want to see how the population change affects the housing prices.

By using open data from the U.S. Census Bureau and Zillow, I want to do the following:

1. **Predict 2017 home sale prices using population data**
2. **Determine the age groups that most influence the market**

SOURCES FOR DATASETS



County population by age group
2010-2017

	STATE	COUNTY	STNAME	CTYNAME	YEAR	AGEGRP	TOT_POP	TOT_MALE	MALE_RTO	region
114	1	1	Alabama	Autauga County	7	0	54864	26785	0.488207	Alabama Autauga County
115	1	1	Alabama	Autauga County	7	1	3190	1663	0.521317	Alabama Autauga County

Median home sale price by county

	RegionID	region	SizeRank	14_med	15_med	16_med	17_med	15_comp	16_comp	17_comp
0	3101	California Los Angeles County	1	432600.0	461900.0	494500.0	514200.0	6.773000	7.057805	3.983822
1	139	Illinois Cook County	2	207600.0	214400.0	193300.0	201900.0	3.275530	-9.841418	4.449043

FEATURE ENGINEERING

1. CENSUS DATA

In order to see the population change by age group over the years, I had to prepare data as follows:

a. Use GROUPBY and RESET_INDEX to aggregate data by County, Year, and Age Group
b. Calculate the population difference in the last 3 years.

```
pop_diff=pd.DataFrame(population.groupby(['region','YEAR','AGEGRP'])['TOT_POP'].sum()).unstack()
```

```
pop_diff_pct=pd.DataFrame(pop_diff_pct.sum())

pop_diff_abs=pd.DataFrame(df.diff(axis=0).sum())

pop_df=pd.DataFrame(population['region'].unique(), columns=['region'])

pop_df=pd.merge(pop_df, pop_diff_abs, on='region')
pop_df=pd.merge(pop_df, pop_diff_pct, on='region')
```

Reference:

The key for the YEAR variable is as follows:
1 = 4/1/2010 Census population
2 = 4/1/2010 population estimates base
3 = 7/1/2010 population estimate
4 = 7/1/2011 population estimate
5 = 7/1/2012 population estimate
6 = 7/1/2013 population estimate
7 = 7/1/2014 population estimate
8 = 7/1/2015 population estimate
9 = 7/1/2016 population estimate
10 = 7/1/2017 population estimate

0 = Total
1 = Age 0 to 4 years
2 = Age 5 to 9 years
3 = Age 10 to 14 years
4 = Age 15 to 19 years
5 = Age 20 to 24 years
6 = Age 25 to 29 years
7 = Age 30 to 34 years
8 = Age 35 to 39 years

9 = Age 40 to 44 years
10 = Age 45 to 49 years
11 = Age 50 to 54 years
12 = Age 55 to 59 years
13 = Age 60 to 64 years
14 = Age 65 to 69 years
15 = Age 70 to 74 years
16 = Age 75 to 79 years
17 = Age 80 to 84 years
18 = Age 85 years or older

FEATURE ENGINEERING

2. ZILLOW DATA

Calculate the changes in median home sale prices from 2014 to 2017.

	RegionID	region	SizeRank	14_med	15_med	16_med	17_med	15_comp	16_comp	17_comp
0	3101	California Los Angeles County	1	432600.0	461900.0	494500.0	514200.0	6.773000	7.057805	3.983822
1	139	Illinois Cook County	2	207600.0	214400.0	193300.0	201900.0	3.275530	-9.841418	4.449043

3. MERGE TWO DATASETS

```
pop_df=pop_df.merge(sale_price,on='region')
pop_df=pop_df.drop(['14_med','RegionID'],axis=1)
```

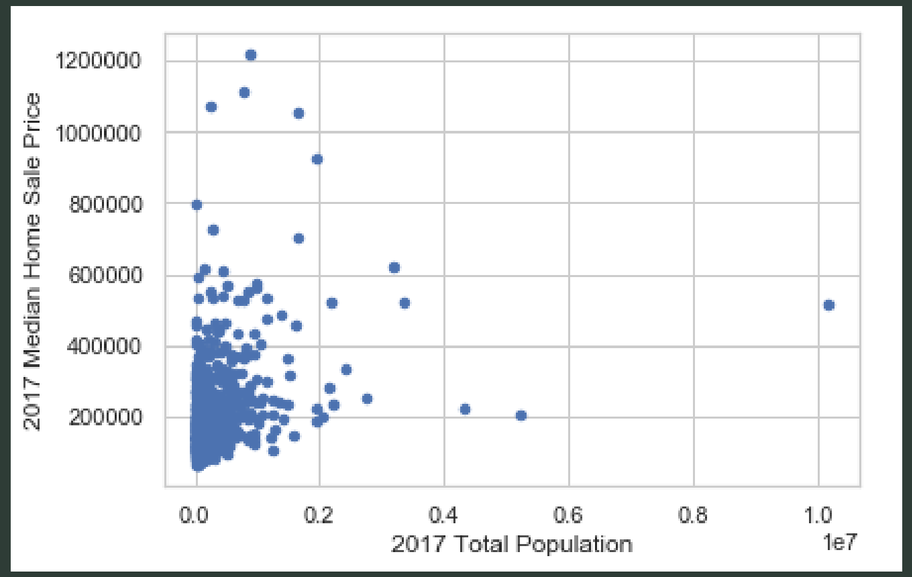
```
population['YEAR']=population['YEAR'].apply(lambda x: '{0:0>2}'.format(x))
population['AGEGRP']=population['AGEGRP'].apply(lambda x: '{0:0>2}'.format(x))

population['age_yr'] =population['YEAR'].astype(str) +'_' + population['AGEGRP'].astype(str)
```

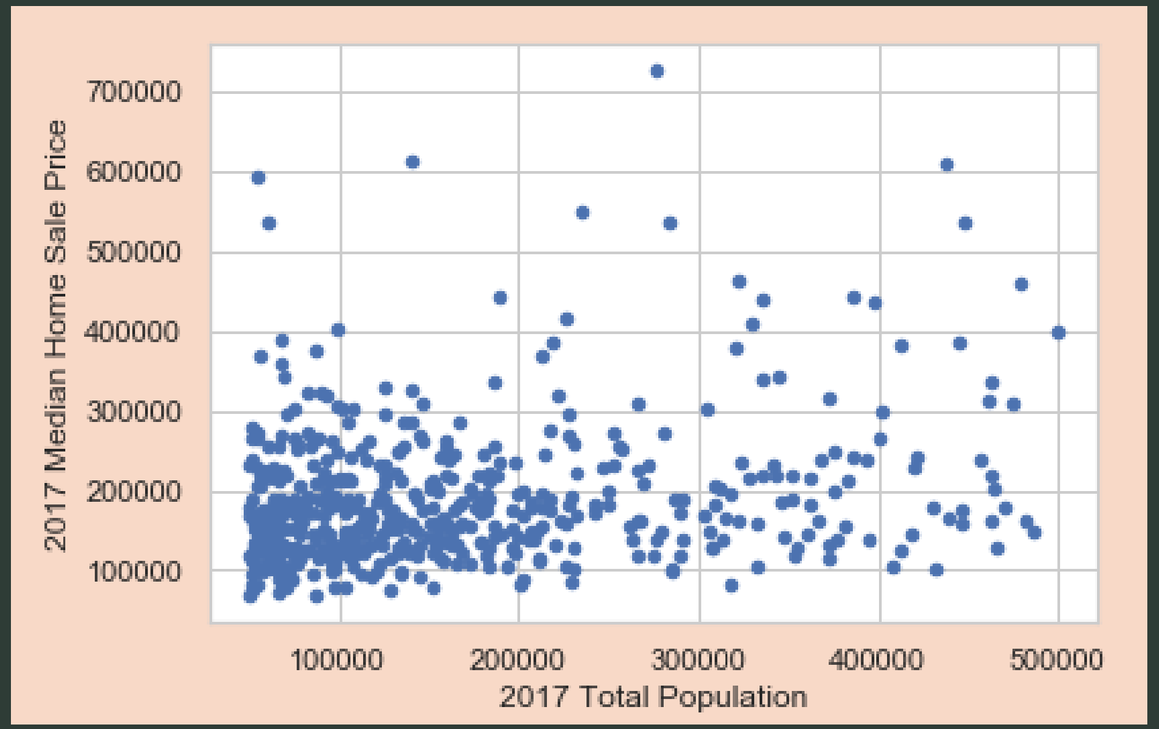
```
df_ageyr = population.loc[population['AGEGRP']!=0].pivot(
                                        index='region', columns='age_yr', values='TOT_POP')
```

EXPLORE DATA

TOTAL POPULATION VS. MEDIAN HOME SALE PRICE

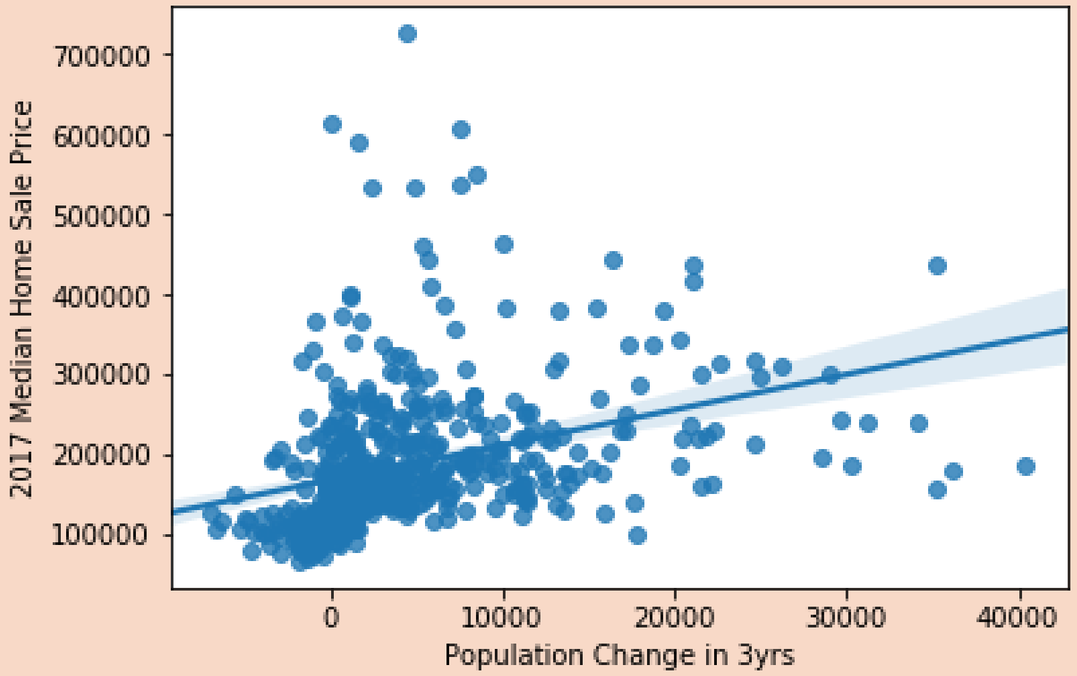


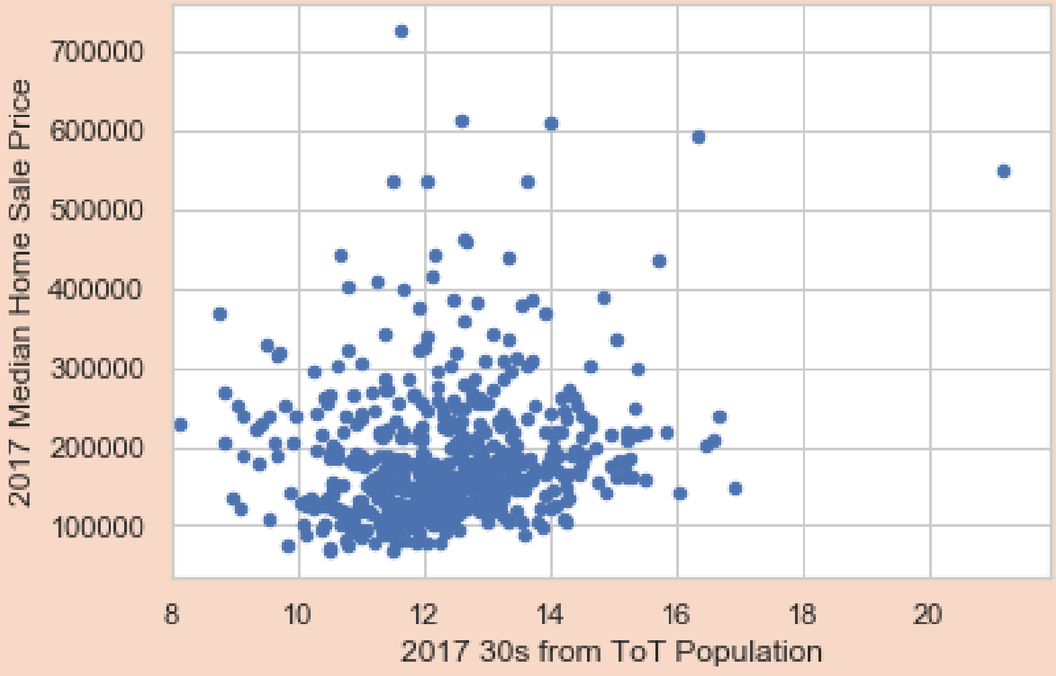
I expected the graph to be linear, which wasn't the case. There were some outliers that I needed to remove.




```
pop_df1=pop_df.loc[(pop_df['TOT_POP_17']>50000)&(pop_df['TOT_POP_17']<500000)&(pop_df['17_med']<1000000)]
pop_df1.plot.scatter(x='TOT_POP_17', y='17_med')
```

EXPLORE DATA





POPULATION CHANGE IN LAST 3YRS VS. MEDIAN HOME SALE PRICE

% OF PEOPLE IN THEIR 30S FROM TOT POPULATION VS. MEDIAN HOME SALE PRICE

2017 HOUSING PRICE

AS **DEPENDENT VARIABLE**

&

51

INDEPENDENT VARIABLES

POPULATION IN 30S

30s_14
30s_15
30s_16
2015_30s_pct
2016_30s_pct

HOME SALE HISTORY

SizeRank
15_med ,16_med
15_comp, 16_comp

POPULATION CHANGE BY AGE GROUP (%)

2014:
'07_01', '07_06', '07_07', '07_08', '07_09', '07_10', '07_11', '07_12', '07_13', '07_14', '07_15', '07_16', '07_17', '07_18'

2015:
'08_01', '08_06', '08_07', '08_08', '08_09', '08_10', '08_11', '08_12', '08_13', '08_14', '08_15', '08_16', '08_17'

2016:
'09_01', '09_06', '09_07', '09_08', '09_09', '09_10', '09_11', '09_12', '09_13', '09_14', '09_15', '09_16', '09_17'

MODEL DATA

1

HOW I CHOSE MODELS

My data consisted of continuous variables.
I used different regression models to determine the model of best fit.

2

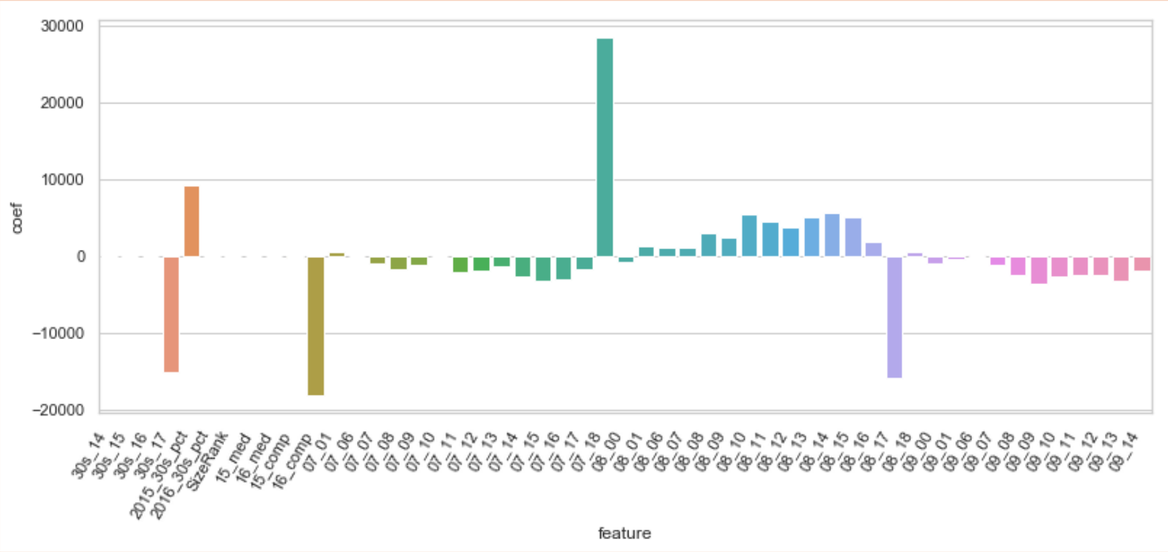
EVALUATE MODELS

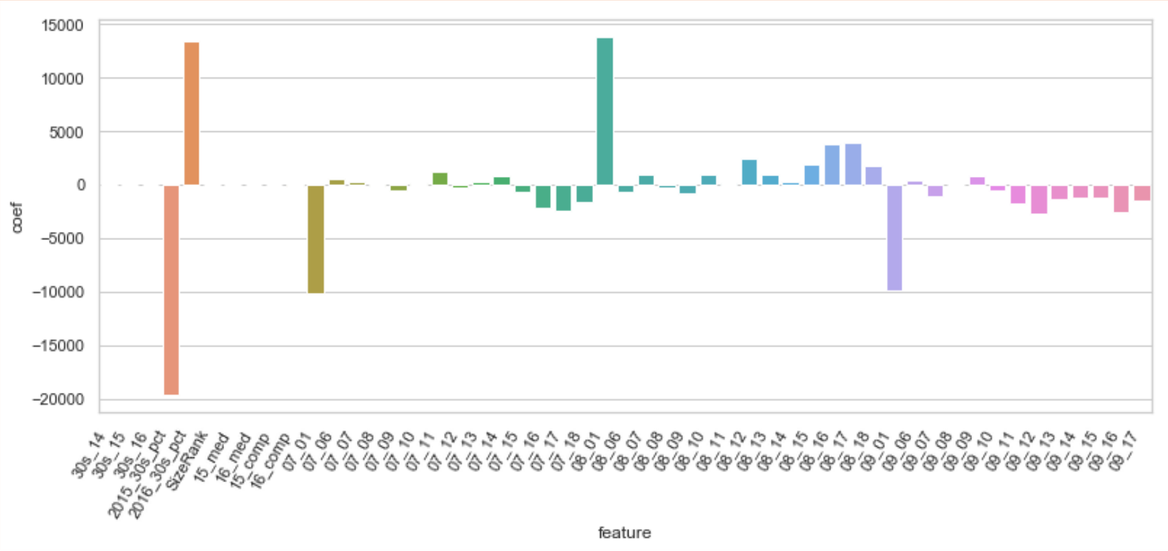
train_test_split
cross validation
root-mean-square error (RMSE)
feature_importances_

REGRESSION MODEL SCORES

	Train	Test (size=0.4)	Cross Val (cv=5)	Cross Val Mean Score
Linear	0.977	0.953	[0.987, 0.846, 0.970, 0.845, 0.952]	0.920
Ridge	0.976	0.955	[0.988, 0.846, 0.970, 0.840, 0.952]	0.919
Lasso	0.976	0.953	[0.987, 0.848, 0.970, 0.844, 0.952]	0.920

REGULARIZATION

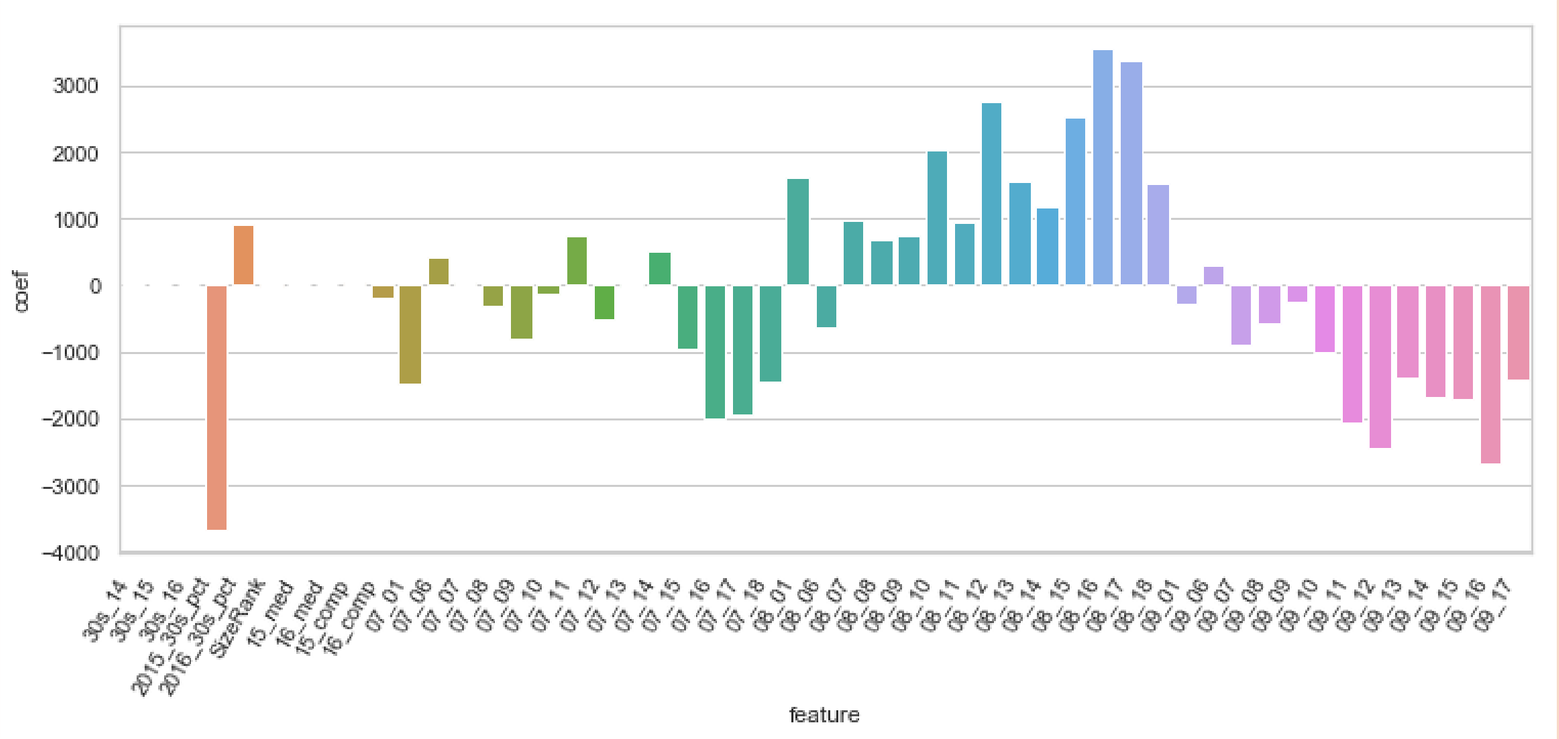


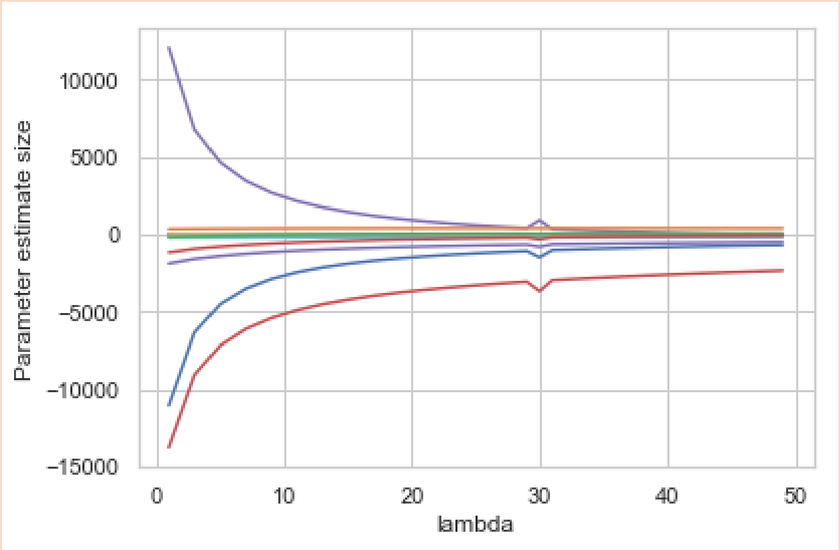


LINEAR REGRESSION VS. LASSO REGRESSION

The magnitude of coefficients is slightly reduced in lasso regression (alpha=15).

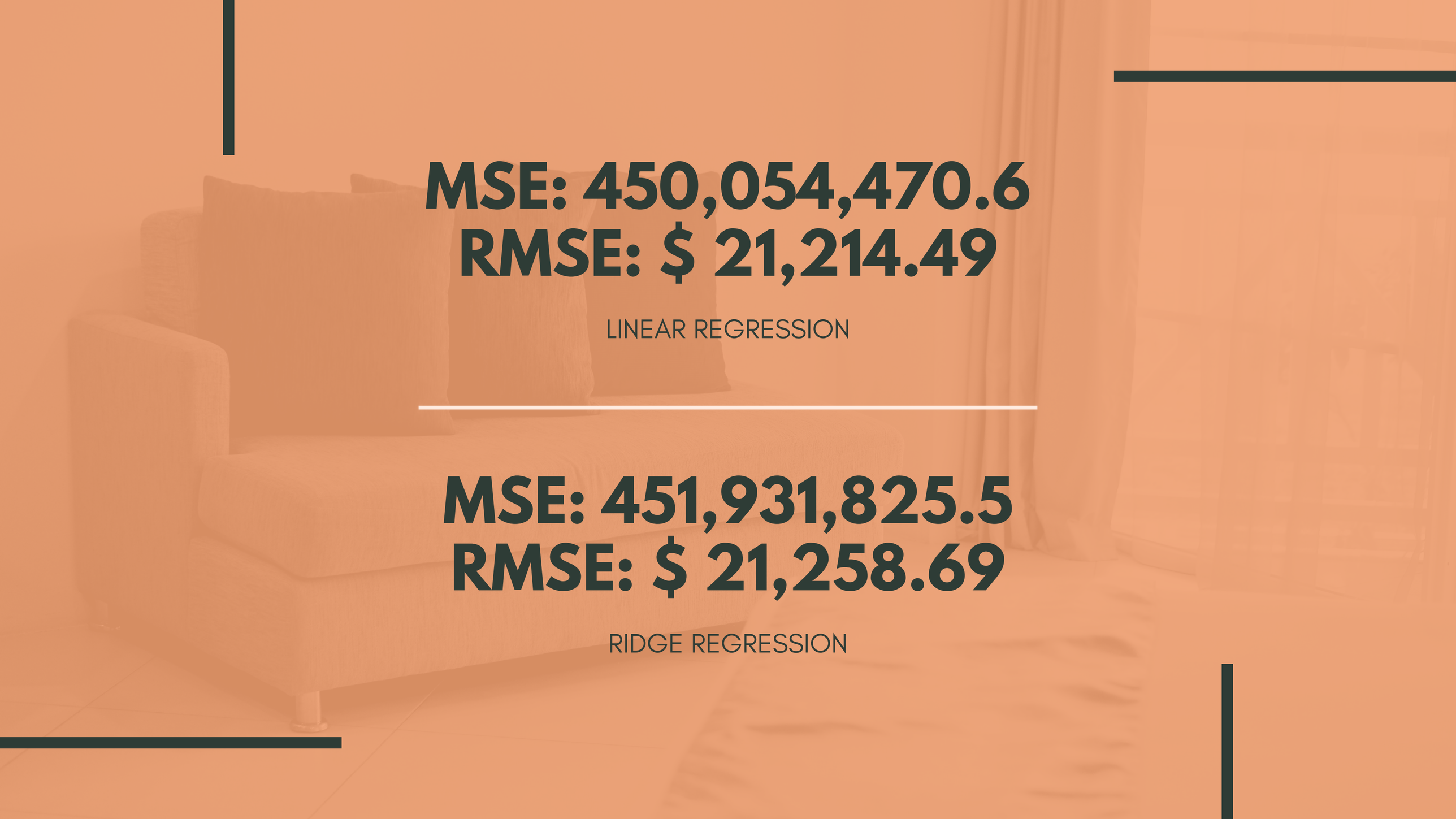
REGULARIZATION





RIDGE REGRESSION (ALPHA=15)

The coefficients are much closer to each other, evening out the degree to which they affect the outcome variable.



MSE: 450,054,470.6
RMSE: $ 21,214.49

LINEAR REGRESSION

MSE: 451,931,825.5
RMSE: $ 21,258.69

RIDGE REGRESSION

GRADIENT BOOSTING

GRADIENT BOOSTING R-SQUARED

0.9682

** Regression R-squared: 0.977*

MSE & RMSE

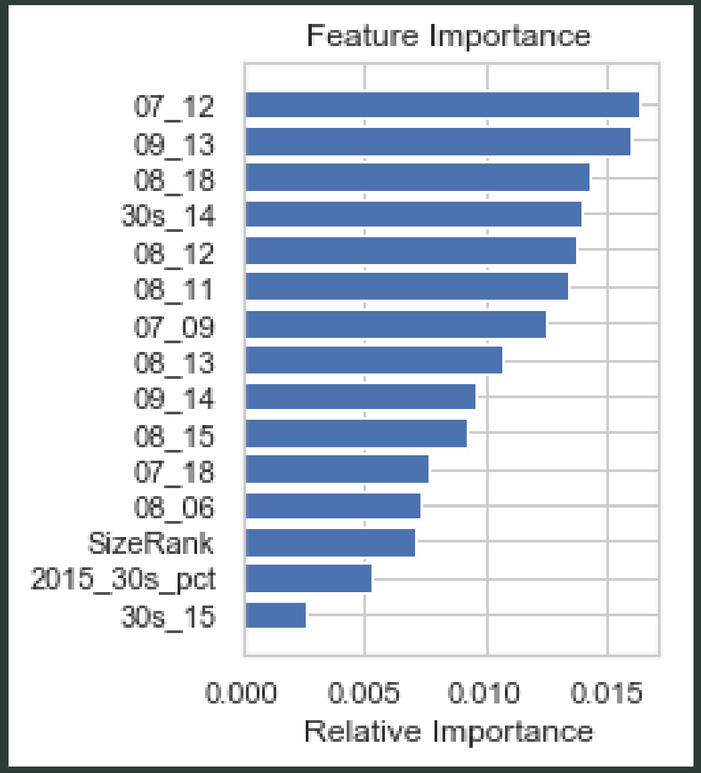
Gradient Boosting MSE: 438,825,672.1

Gradient Boosting RMSE: $ 20,948.17

FEATURE IMPORTANCE

It appears that the age groups 11-13** are the most influential groups to housing prices.

***11 = Age 50 to 54 years | 12 = Age 55 to 59 years | 13 = Age 60 to 64 years | 14 = Age 65 to 69 years*



NEXT STEP

1

LOOK INTO CORE AGE GROUP

I verified that the working age group is the most influential toward housing price. I want to look into how each segment of the working age group affects the market.

2

DIFFERENT MODEL

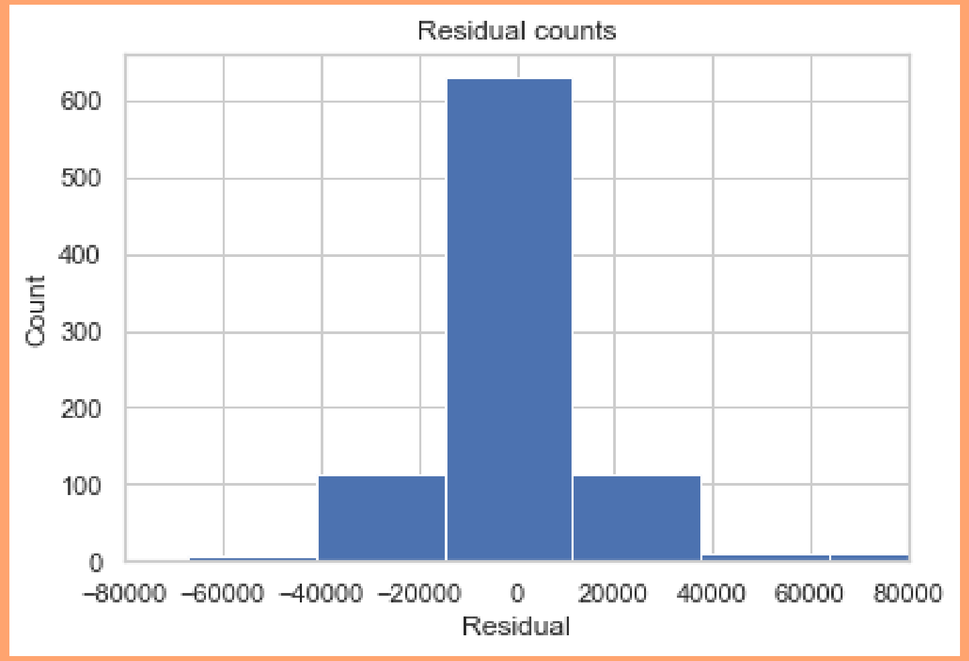
Using time series or mixed-effects models might bring different insights as the datasets are made through observations over time.

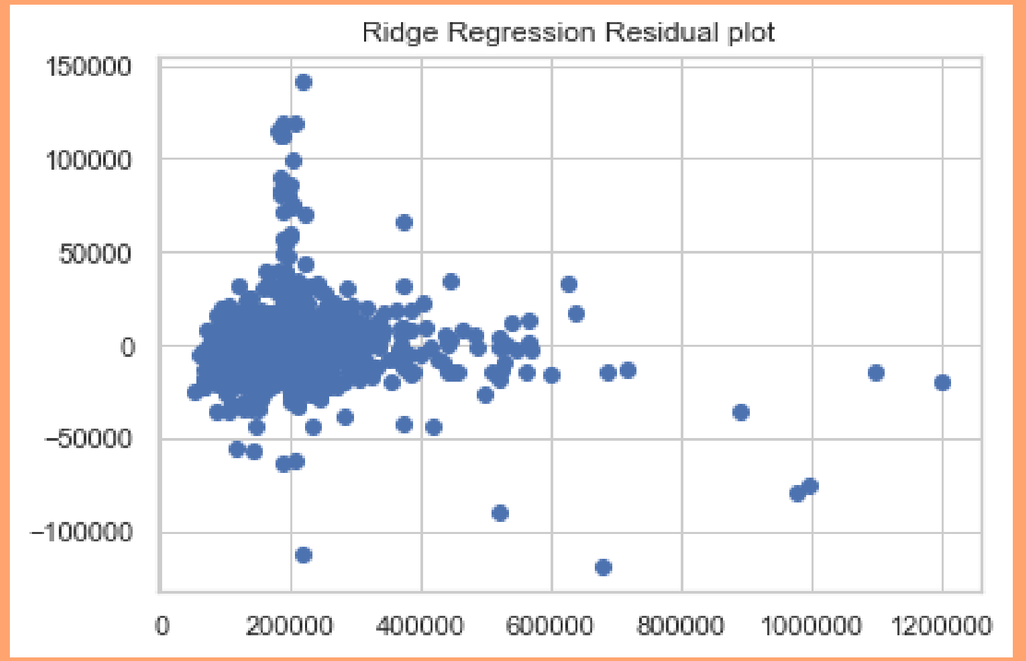
3

PREDICT THE FUTURE

I hope to predict future home sale prices without the Y.

THE WEAKNESS





RESIDUAL Although the residual counts are normally distributed, the error terms are not. There are significant outliers that require some attention. In next iteration, I want to take a closer look at the outliers to improve the model.

SOURCES

HTTPS://WWW.CENSUS.GOV

HTTPS://WWW.ZILLOW.COM/RESEARCH/DATA/

HTTPS://WWW.ZILLOW.COM/RESEARCH/HOME-SALES-METHODOLOGY-7733/

THANK YOU!

HTTPS://GITHUB.COM/NOMADSJOURNAL/THINKFUL