AAI 500: Final Project

Lauren Taylor
Ruben Velarde
Jeffrey Thomas

Project Overview

Data set: Redfin data about house sales

<u>Our goal</u>

↓

Predict the price of a house using
multiple linear regression

Data Set Overview

Data Set Size: 300 rows, 27 columns

Data Types

Categorical: Zip Code, City, Location, Property Type, State

Numerical: Price, Beds, Baths, Square Feet, Lot Size, Year Built

Data Set Details

Data Set Size: 300 rows, 27 columns

Categoricals	
Zip	**5 zips:** 92037, 92127, 91942, 92122, 92067
City	**5 cities:** La Jolla, San Diego, La Mesa, Rancho Santa Fe, Rancho Bernardo
Location	Provided by two columns: longitude and latitude
Property Type	5 types: Single Family, Condo, Townhouse, Vacant Land, Multi-Family
State	All in California

Data Set Details

Data Set Size: 300 rows, 27 columns

Numericals				
Category	**Mean**	**Min**	**Max**	**Std Deviation**
Price	$3,237,747	$369,000	$45,000,000	$4,735,400
Sq Feet	2806.9	432	22,897	2410.73
Beds	3.49	0	10	1.62
Baths	3.2	1	12.5	1.85
Year Built	1984	1920	2022	23.16

Data Cleaning

Dropping Data

- Drop columns with data that is irrelevant to the price (eg. time of next open house, state)
- Drop columns missing information (eg. Sold Date)
- Drop columns that have redundant information (eg. Address, since location is the same information, and easier to work with)
- Drop instances of vacant land, as it is a different type of asset. All other properties include a dwelling.

Data Cleaning

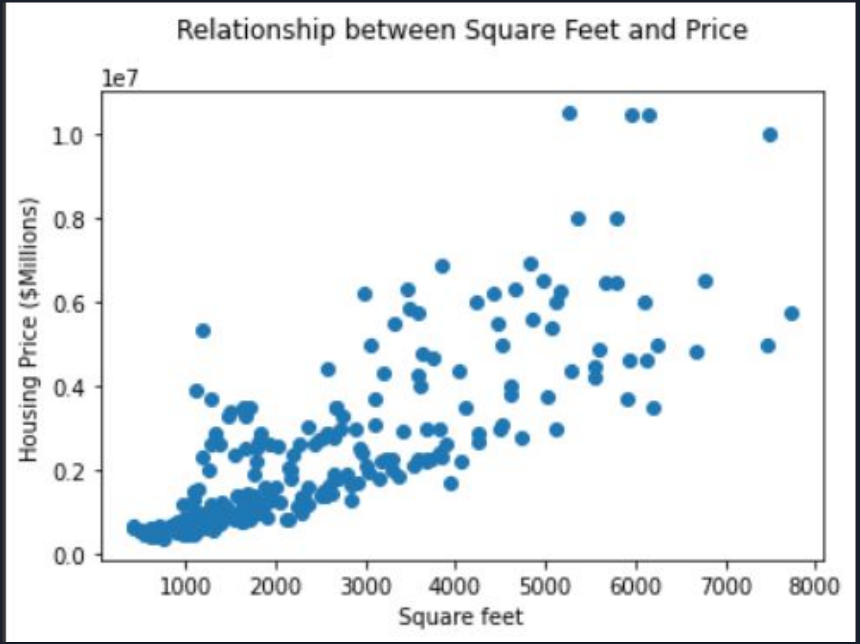
Removing outliers

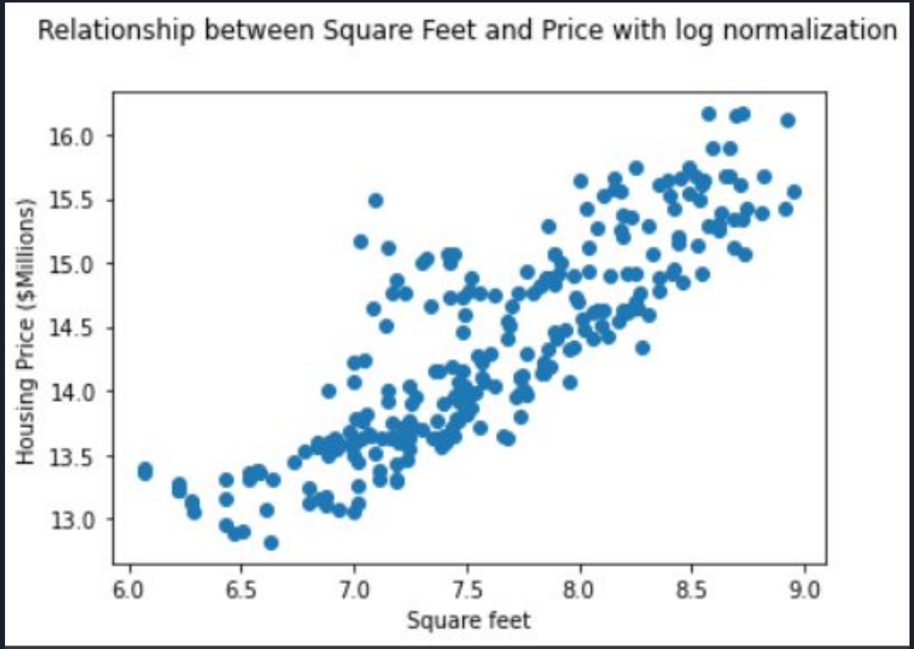
In order to account for outliers we removed all observations where the value was greater than 2.5 standard deviations from the mean

Normalizing data

We normalized the price and square feet columns by using Numpy log function

Square Feet vs Price





Data Set Details After Cleaning

Data Set Size: 300 rows, 27 columns

Numericals				
Category	**Mean**	**Min**	**Max**	**Std Deviation**
Price	$2,338,521	$369,000	$10,500,000	$2,006,316
Sq Feet	2473.64	432	7,722	1604.33
Beds	3.41	0	7	1.39
Baths	3	1	7.5	1.47
Home Age	36.71	0	90	22.82

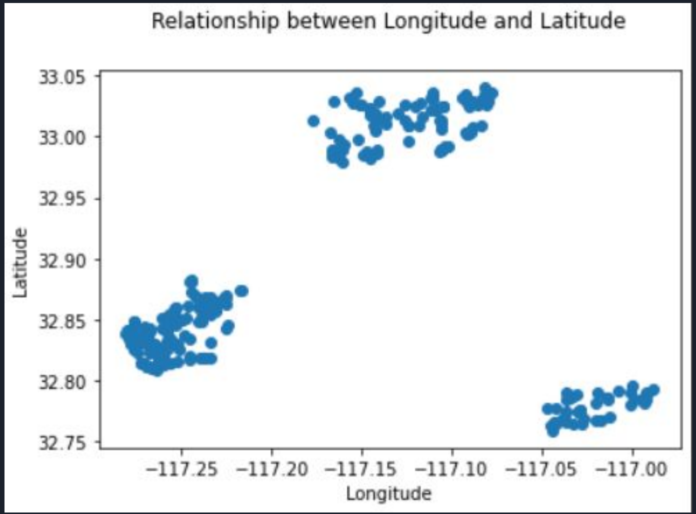
Data Cleaning

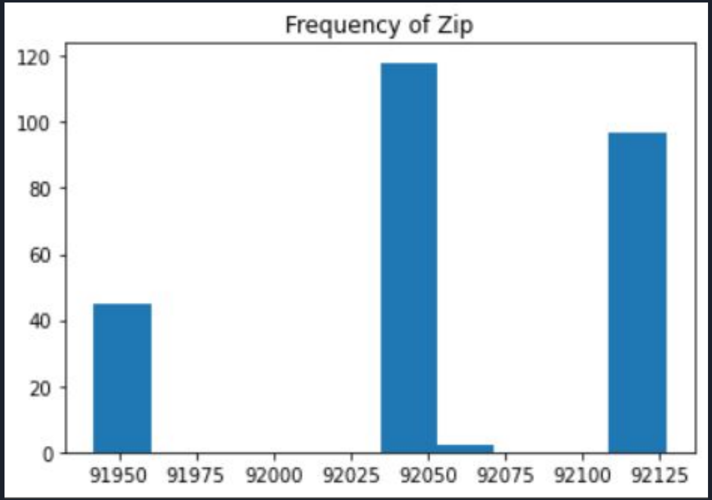
Variable Transformations:

-Transform year built into house age by subtracting the value from 2022.

-Create dummy categorical variable for house age → New houses and old houses.

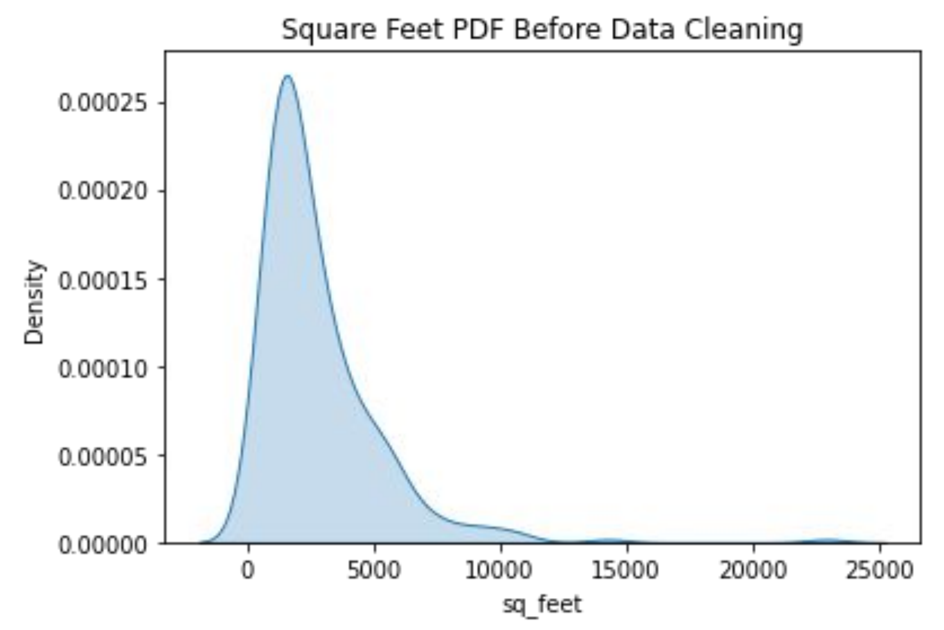
-Create dummy categorical variable for longitude and latitude→3 categories

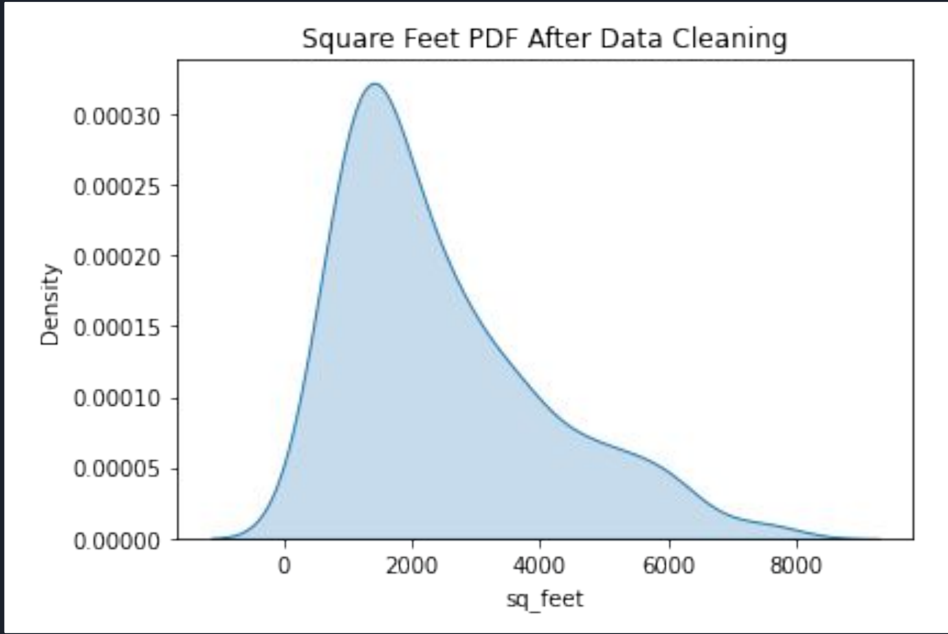




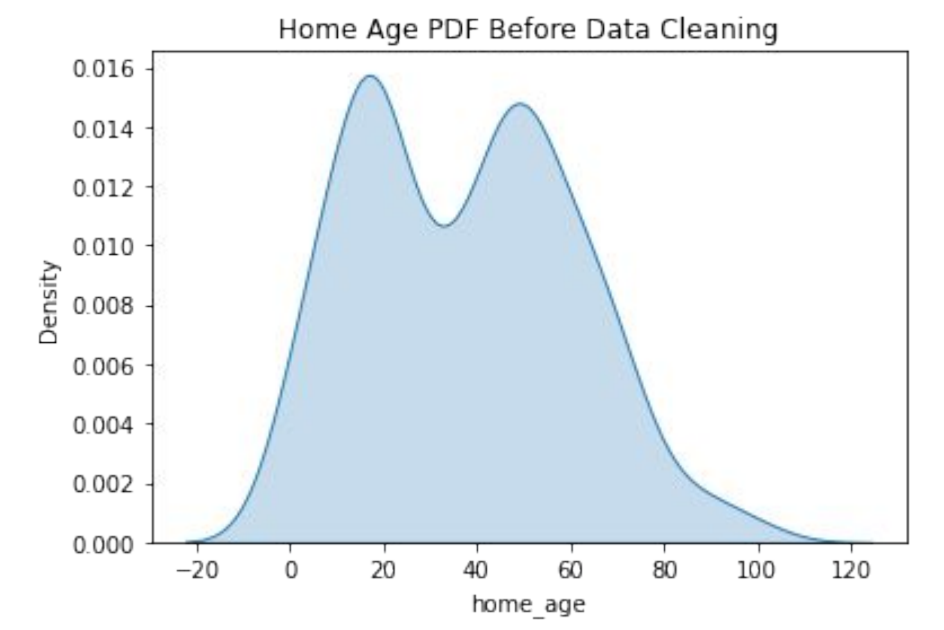
DESCRIPTIVE DATA

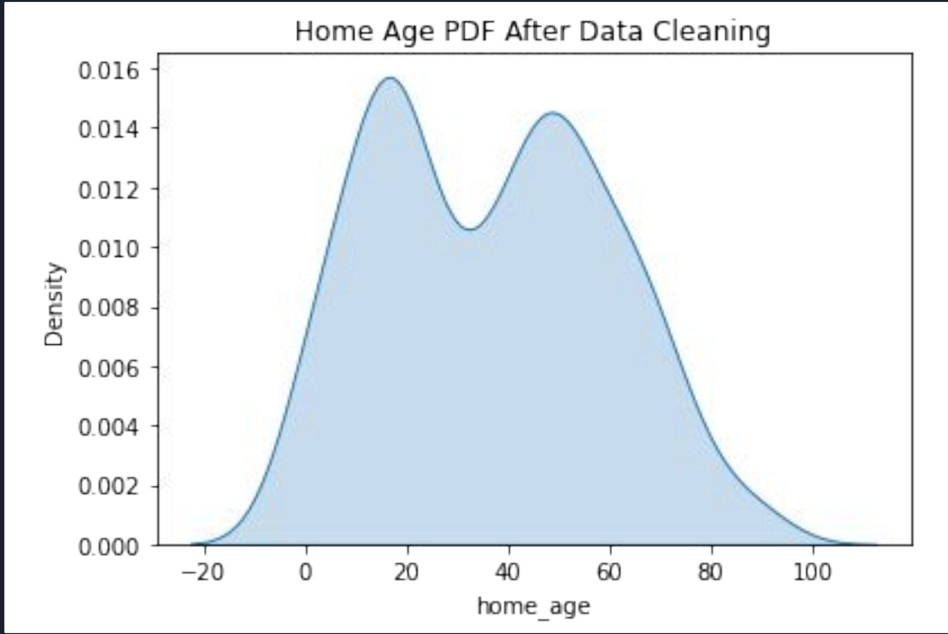
Square Feet


Square Feet PDF Before Data Cleaning
Density
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
5000
10000
15000
20000
25000
sq_feet

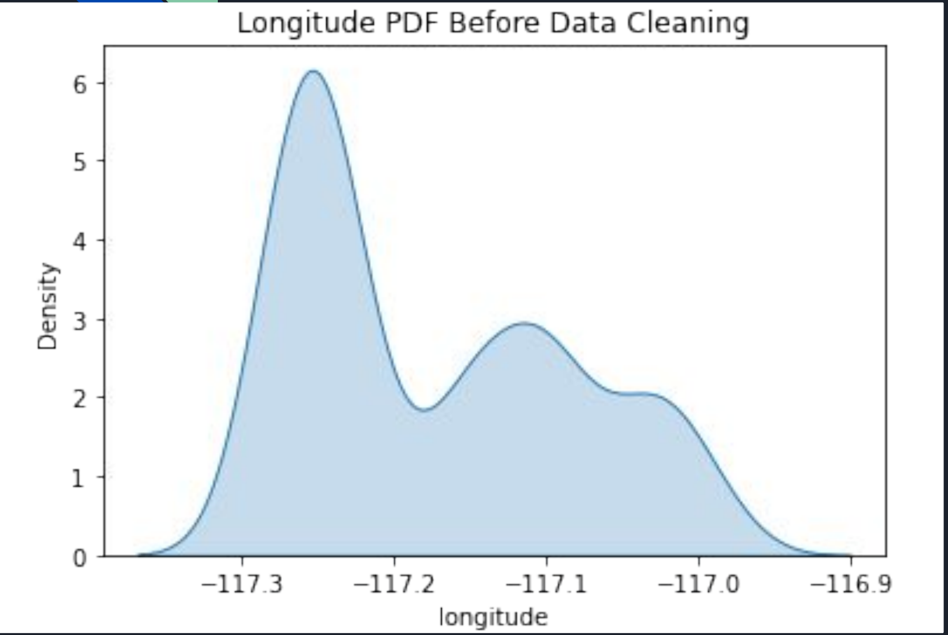

Square Feet PDF After Data Cleaning
Density
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
2000
4000
6000
8000
sq_feet

Home Age

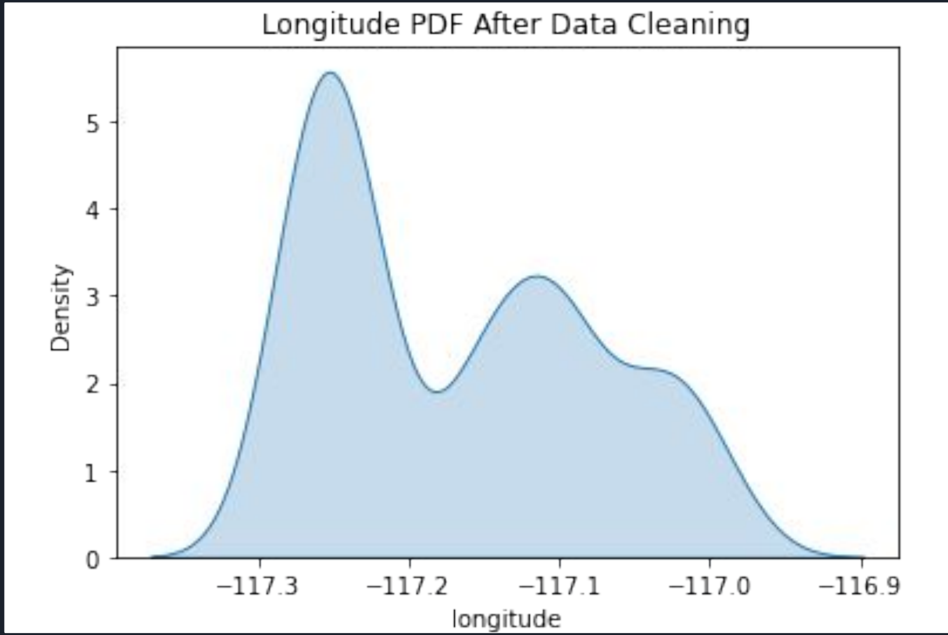

Home Age PDF Before Data Cleaning
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
−20
0
20
40
60
80
100
120
home_age


Home Age PDF After Data Cleaning
0.016
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
Density
−20
0
20
40
60
80
100
home_age

Longitude

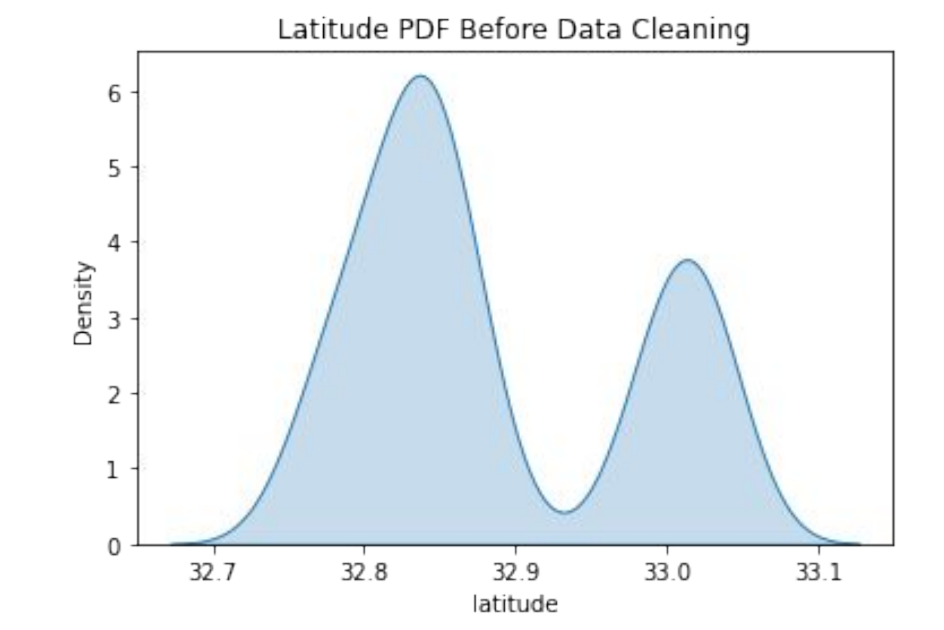


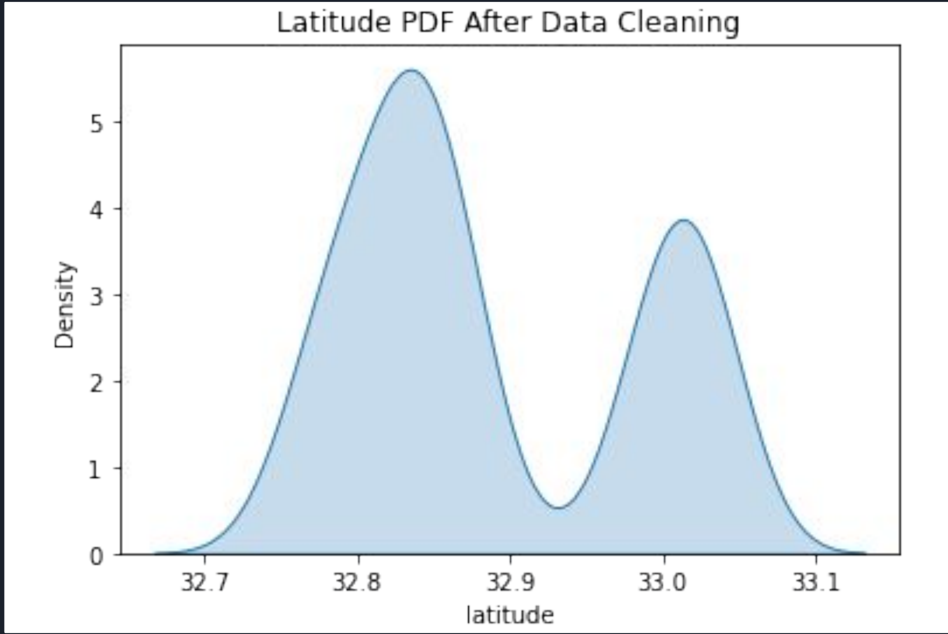
Longitude PDF Before Data Cleaning
Density
0
1
2
3
4
5
6
−117.3
−117.2
−117.1
−117.0
−116.9
longitude



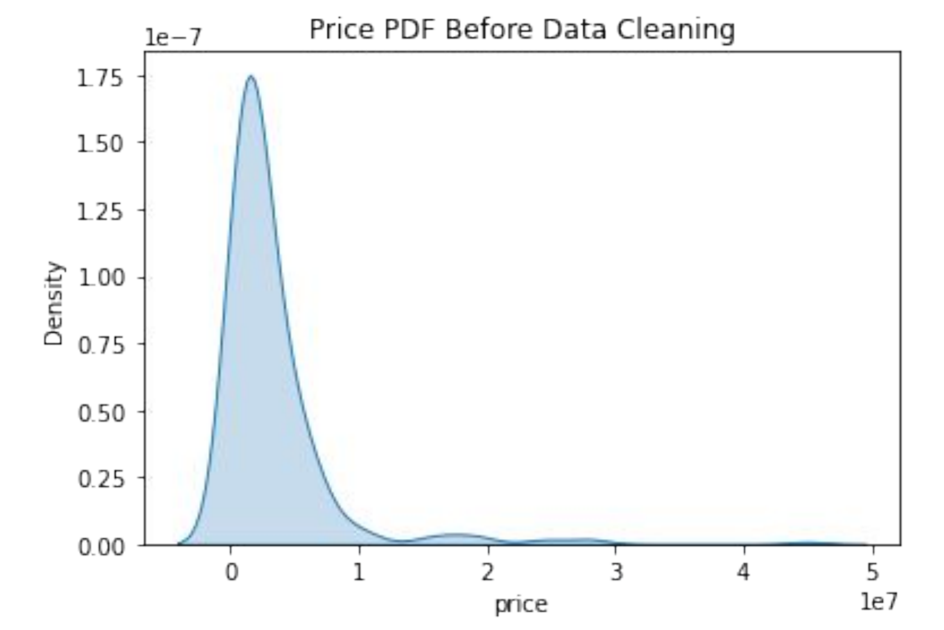
Longitude PDF After Data Cleaning
Density
0
1
2
3
4
5
−117.3
−117.2
−117.1
−117.0
−116.9
longitude

Latitude

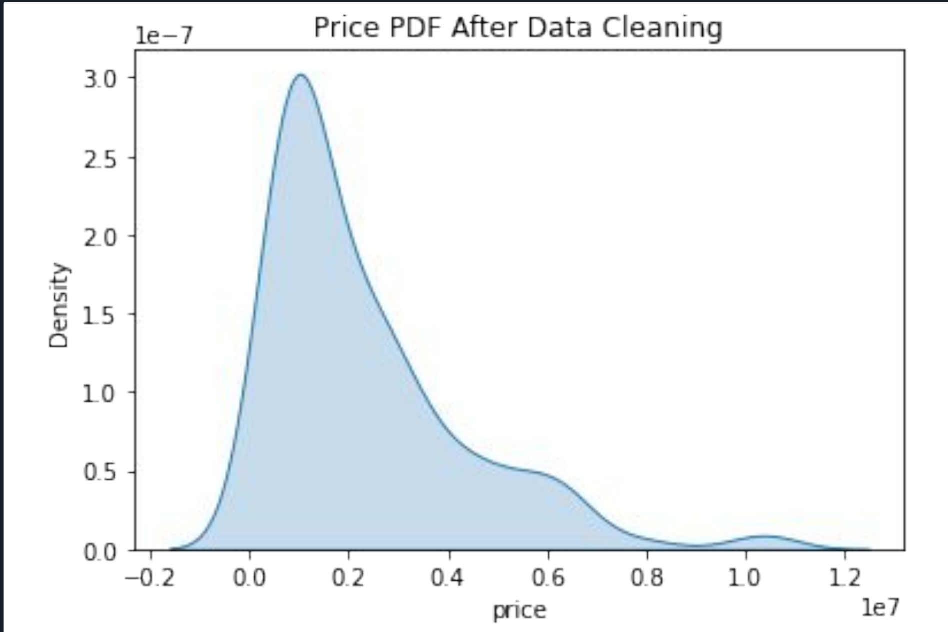

Latitude PDF Before Data Cleaning
Density
latitude
0
1
2
3
4
5
6
32.7
32.8
32.9
33.0
33.1


Latitude PDF After Data Cleaning
Density
latitude
0
1
2
3
4
5
32.7
32.8
32.9
33.0
33.1

House Price



Price PDF Before Data Cleaning
1e−7
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Density
0
1
2
3
4
5
price
1e7



Price PDF After Data Cleaning
1e−7
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Density
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
price
1e7

Type of Distributions

Multimodal:

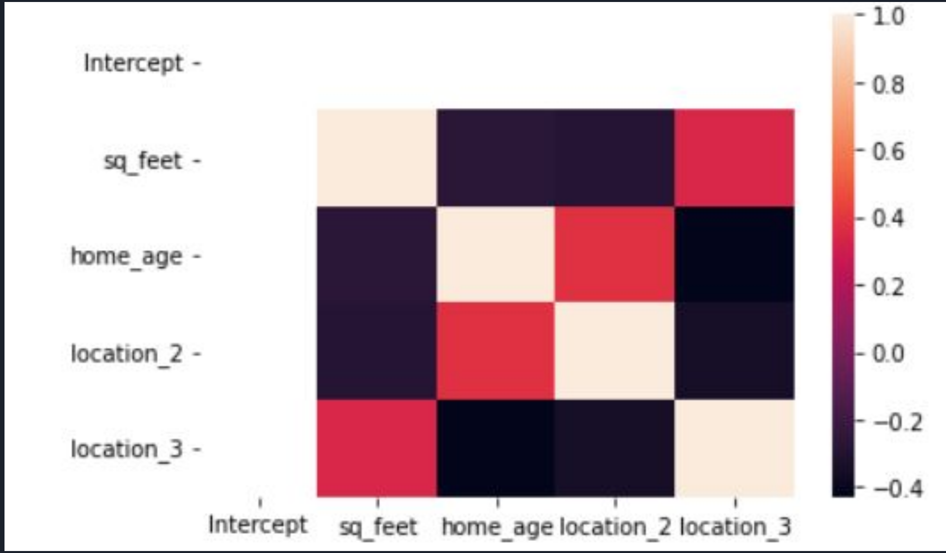
- Home Age (Bimodal),
- Longitude
- Latitude (Bimodal)

Log-Normal:

- Square Feet
- Price

Variable Relationships

Correlation Heat-Map



VIF Scores

	VIF	variable
0	157.620893	Intercept
1	1.189228	sq_feet
2	1.340798	home_age
3	1.259221	location_2
4	1.356345	location_3

Based on these results we can conclude that there is little multicollinearity and that the independent variables have a relationship to the dependent variable for our final model.

Model Results

Model Summary

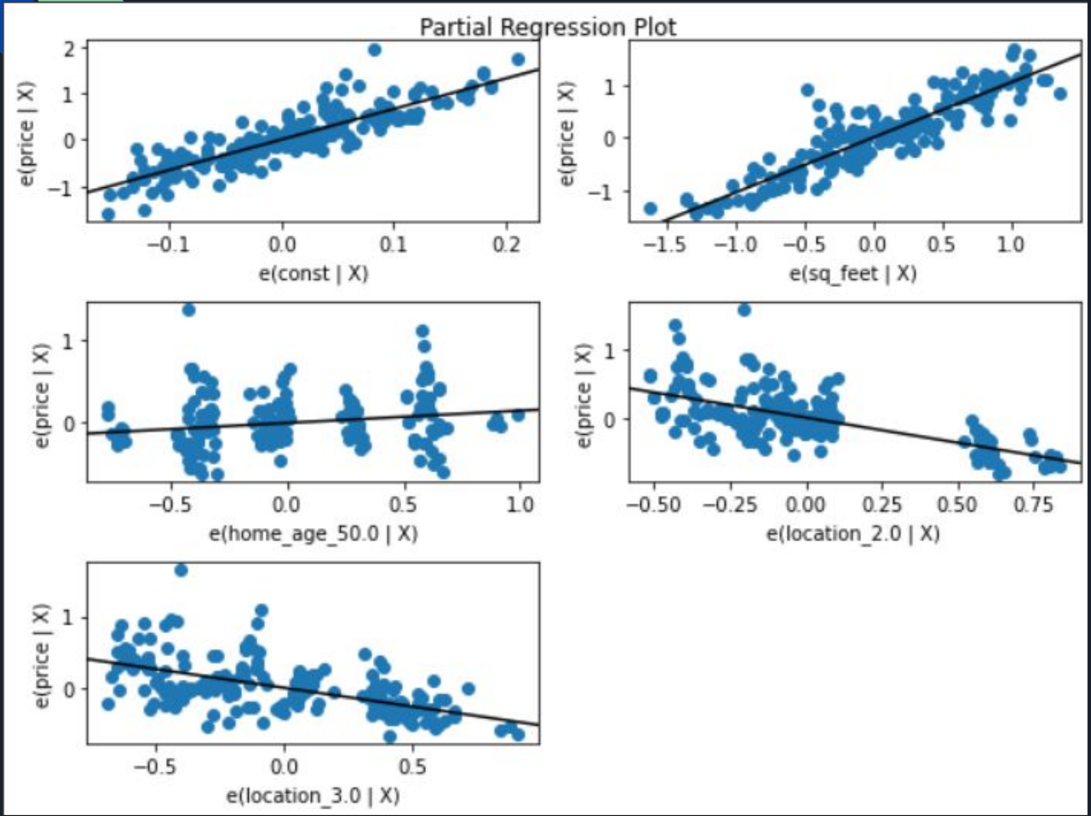
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.868
Model:                            OLS   Adj. R-squared:                  0.865
Method:                 Least Squares   F-statistic:                     334.5
Date:                Sun, 23 Oct 2022   Prob (F-statistic):           2.26e-88
Time:                        05:36:51   Log-Likelihood:                -39.368
No. Observations:                 209   AIC:                             88.74
Df Residuals:                     204   BIC:                             105.4
Df Model:                           4
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const             6.5647      0.257     25.566      0.000       6.058       7.071
sq_feet           1.0569      0.034     31.463      0.000       0.991       1.123
home_age_50.0     0.1484      0.050      2.967      0.003       0.050       0.247
location_2.0     -0.7354      0.059    -12.483      0.000      -0.852      -0.619
location_3.0     -0.5243      0.050    -10.424      0.000      -0.623      -0.425
==============================================================================
Omnibus:                       48.556   Durbin-Watson:                   1.815
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              111.105
Skew:                           1.058   Prob(JB):                     7.48e-25
Kurtosis:                       5.877   Cond. No.                         97.6
==============================================================================
```

- $R^2 = 0.868$
- P-values are less than 0.05 for all variables

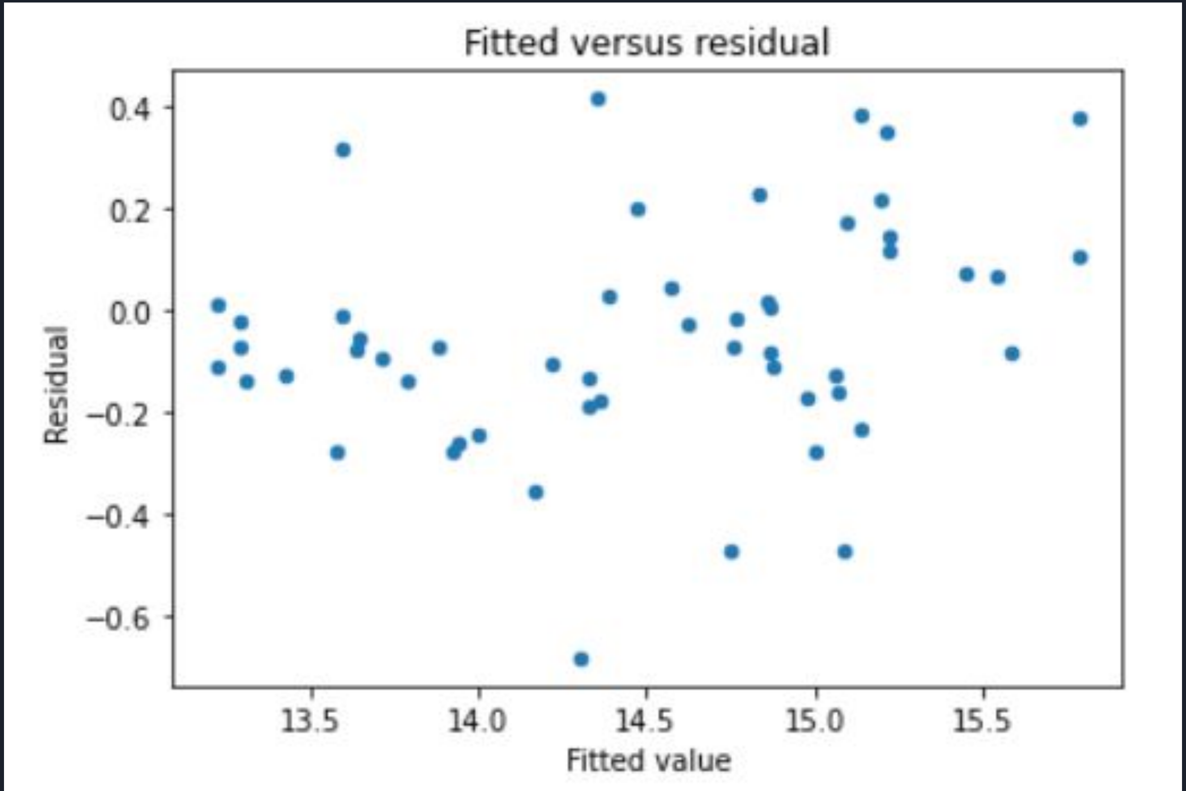
Equation:

$$y = X_1 1.0569 + X_2 0.1484 - X_3 0.7354 - X_4 0.5243 + 6.5647$$

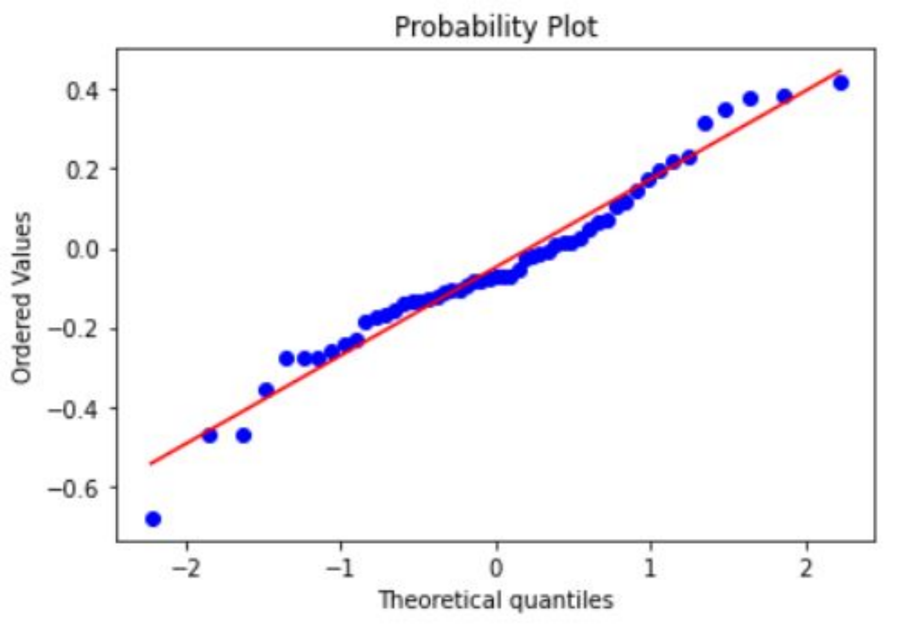
Partial Regression Graph

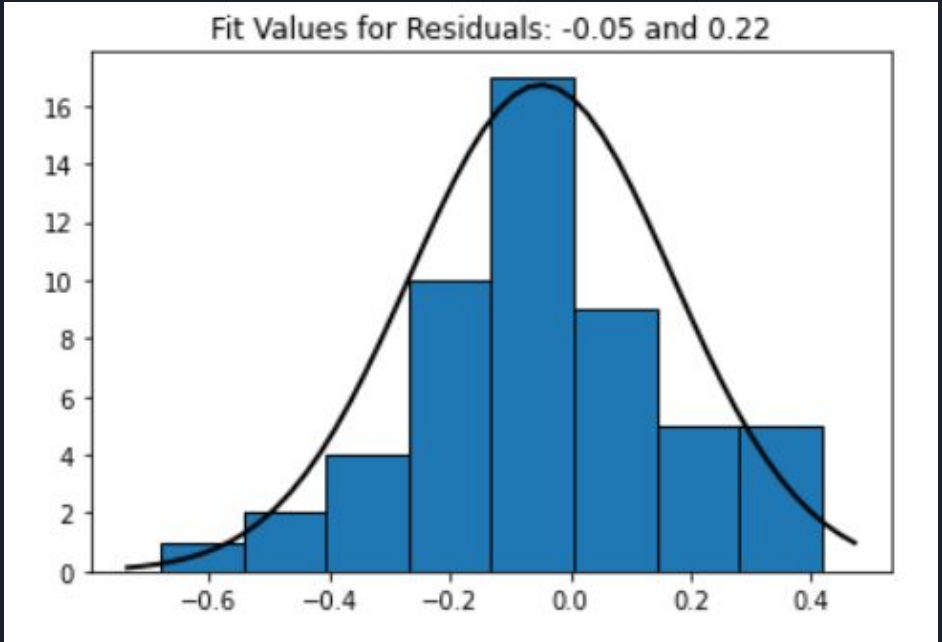


Model Residuals

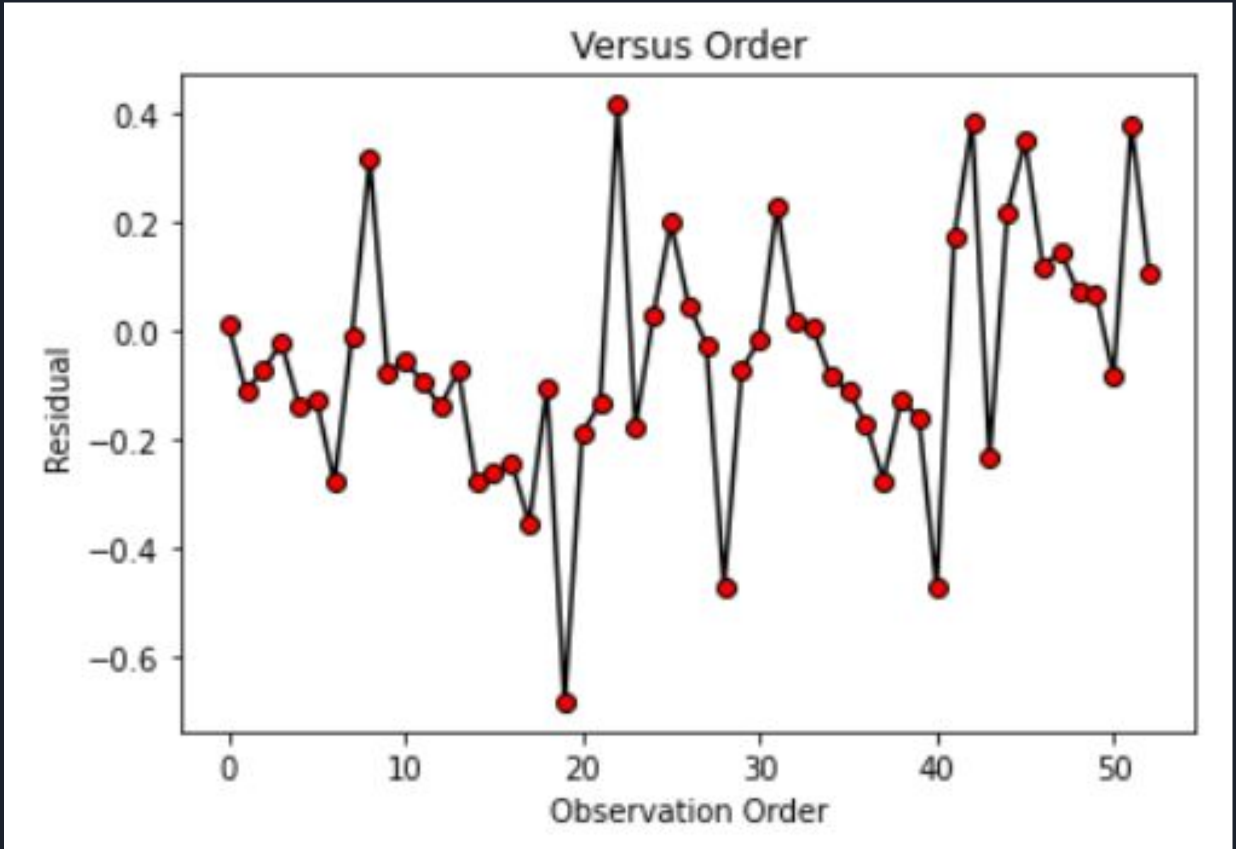


Model Residuals

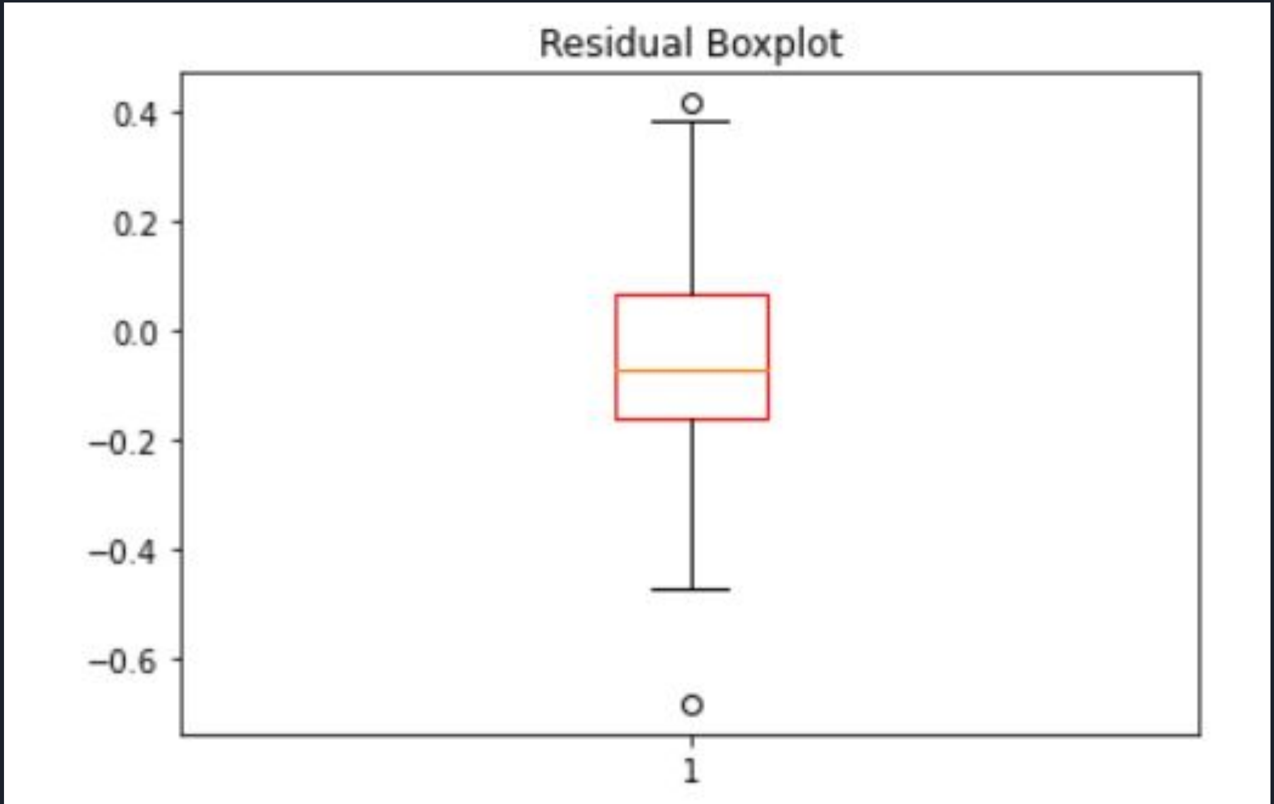




Model Residuals



Model Residuals



Conclusion

Interpretation

-Overall the high R^2 and the low p-values indicate the statistical significance of the model.

- However outliers evident in the residual plots have decreased the predictive power of the model in some situations.

Improvements

-In order to improve this model there needs to be a closer examination of outliers and our filtering of those extreme values

-Analyzing for leverage on certain variables could help reduce outliers since these values could have a negative impact on our model

-One type of predictor that would have been valuable is economic data (eg. inflation rates), this would help improve accuracy overtime because changing economic conditions can affect house prices.

Thank You For Listening!