

Sticker Shock: Bay Area Housing Market

REDFIN

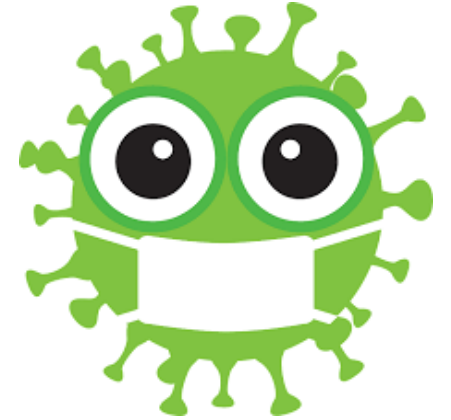
Predicting Home Prices with
Linear Regression

trial

R

Why Home Prices?

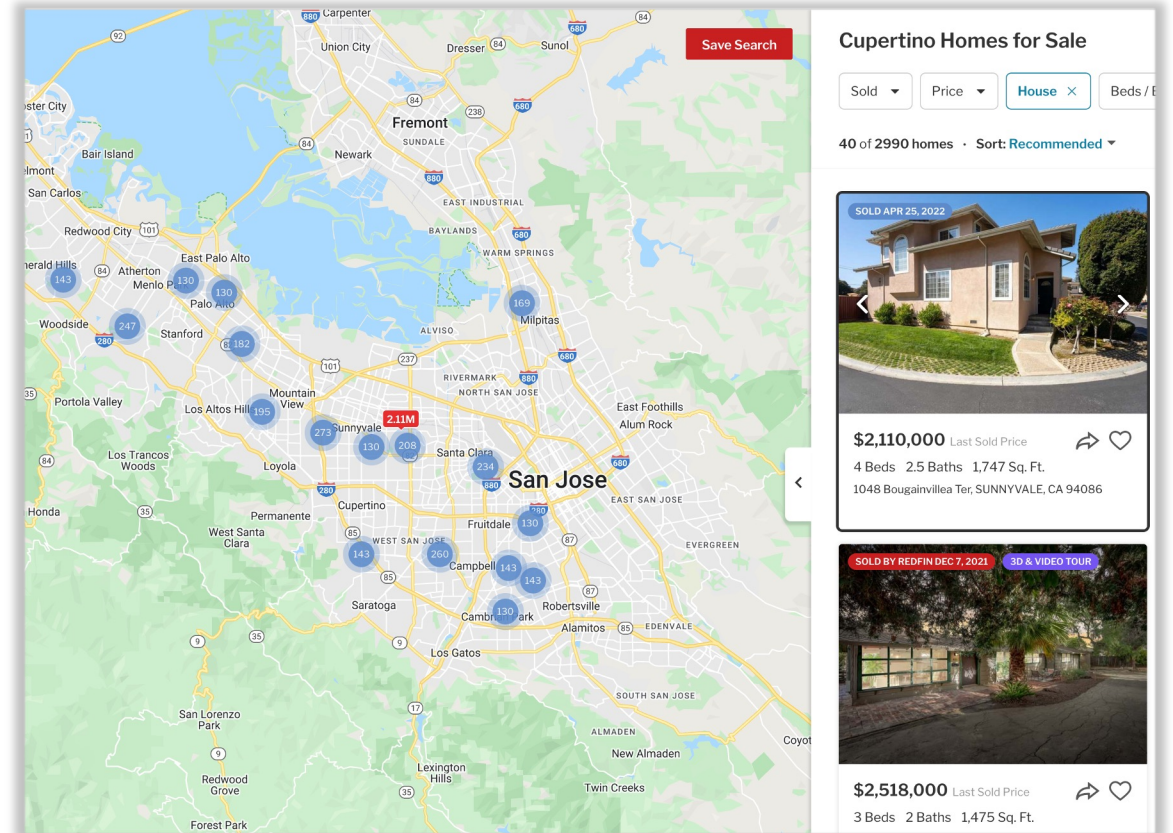
- Covid-19
- Inflation
- Economic Downturn
- Supply chain Shortages





Why Redfin?

- Web Scraping-Friendly
- Extensive Real Estate Listings
- Feature-Rich Data and Filters





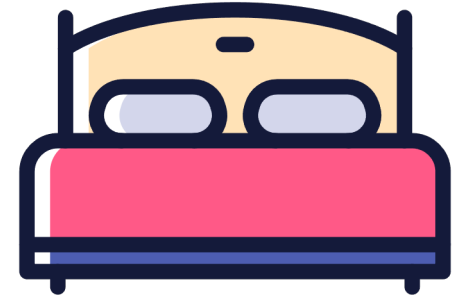
EDA



- Single Family Homes
- Sold within last 6 months
- 1600+ Data points
- 13 Features
- 7 Cities
- 18 Zip Codes

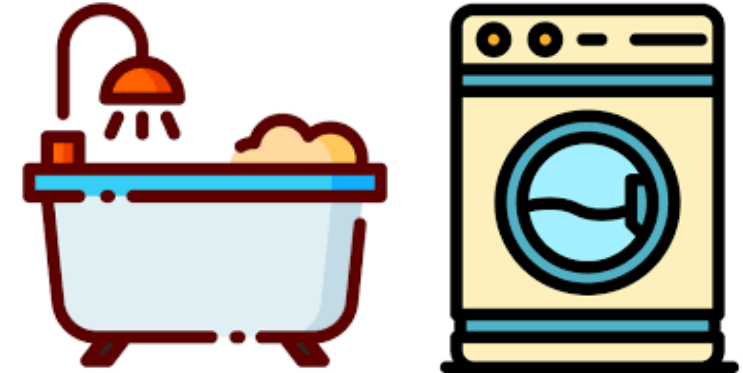


Data Point Example



Sold Price

- Beds
- Baths
- Floors
- Garage
- Lot Size
- Home Size
- Year Built
- School Score
- Laundry
- Heating
- A/C
- Pool
- City





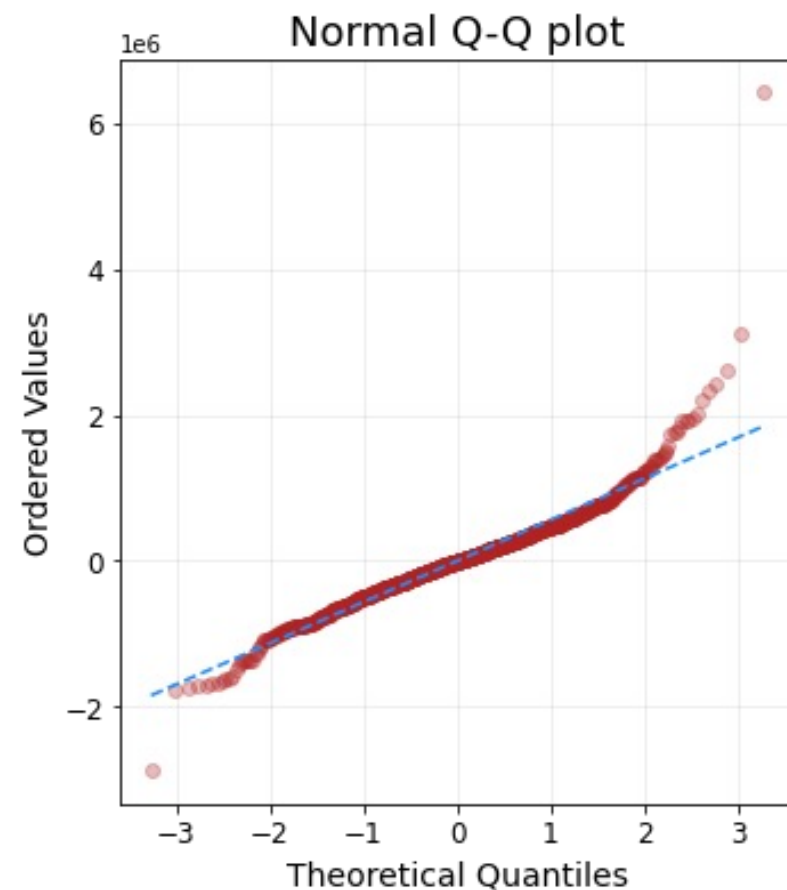
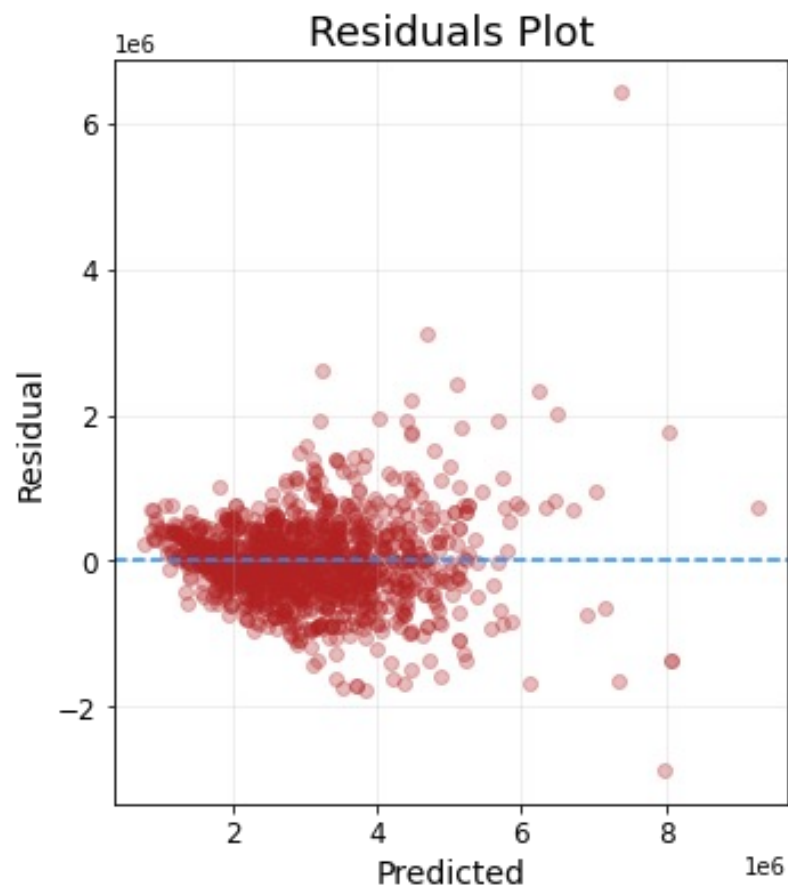
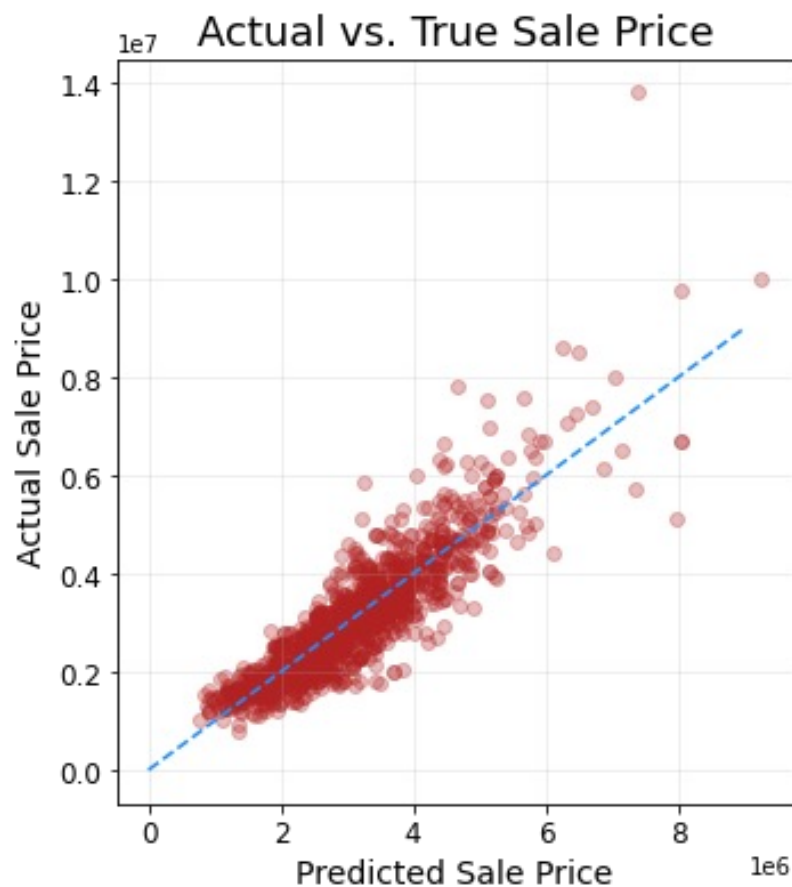
Model Performance (Before)

Baseline Linear Regression Results

R-Squared:	0.783
Adjusted R-Squared:	0.782
Mean Absolute Error:	\$416,244
Root Mean Squared Error:	\$586,488



Model Evaluation (Before)



* Baseline Linear Regression - Untouched Dataset



Feature Engineering

```
Data columns (total 14 columns):  
#      Column      Non-Null Count  Dtype  
---  -  
0     sold price    1292 non-null   float64  
1     beds          1292 non-null   float64  
2     baths         1292 non-null   float64  
3     floors        1292 non-null   float64  
4     garage spaces  1292 non-null   int64  
5     lot size       1292 non-null   float64  
6     home size      1292 non-null   float64  
7     school score avg 1292 non-null   float64  
8     laundry        1292 non-null   bool  
9     heating        1292 non-null   bool  
10    air conditioning 1292 non-null   bool  
11    pool           1292 non-null   bool  
12    city           1292 non-null   object  
13    age of house    1292 non-null   float64
```



```
Data columns (total 32 columns):  
#      Column      Non-Null Count  Dtype  
---  -  
0     sold price    1292 non-null   float64  
1     lot size       1292 non-null   float64  
2     home size      1292 non-null   float64  
3     school score avg 1292 non-null   float64  
4     laundry        1292 non-null   int64  
5     heating        1292 non-null   int64  
6     air conditioning 1292 non-null   int64  
7     pool           1292 non-null   int64  
8     age of house    1292 non-null   float64  
9     beds_3.0       1292 non-null   uint8  
10    beds_4.0       1292 non-null   uint8  
11    beds_5.0       1292 non-null   uint8  
12    beds_6+       1292 non-null   uint8  
13    baths_1.5      1292 non-null   uint8  
14    baths_2.0      1292 non-null   uint8  
15    baths_2.5      1292 non-null   uint8  
16    baths_3.0      1292 non-null   uint8  
17    baths_3.5      1292 non-null   uint8  
18    baths_4.0      1292 non-null   uint8  
19    baths_4.5      1292 non-null   uint8  
20    baths_5.0      1292 non-null   uint8  
21    baths_6+       1292 non-null   uint8  
22    floors_2.0     1292 non-null   uint8  
23    floors_3.0     1292 non-null   uint8  
24    garage spaces_1 1292 non-null   uint8  
25    garage spaces_2 1292 non-null   uint8  
26    garage spaces_3+ 1292 non-null   uint8  
27    city_LOS ALTOS  1292 non-null   uint8  
28    city_MOUNTAIN VIEW 1292 non-null   uint8  
29    city_PALO ALTO  1292 non-null   uint8  
30    city_SANTA CLARA 1292 non-null   uint8  
31    city_SUNNYVALE  1292 non-null   uint8
```



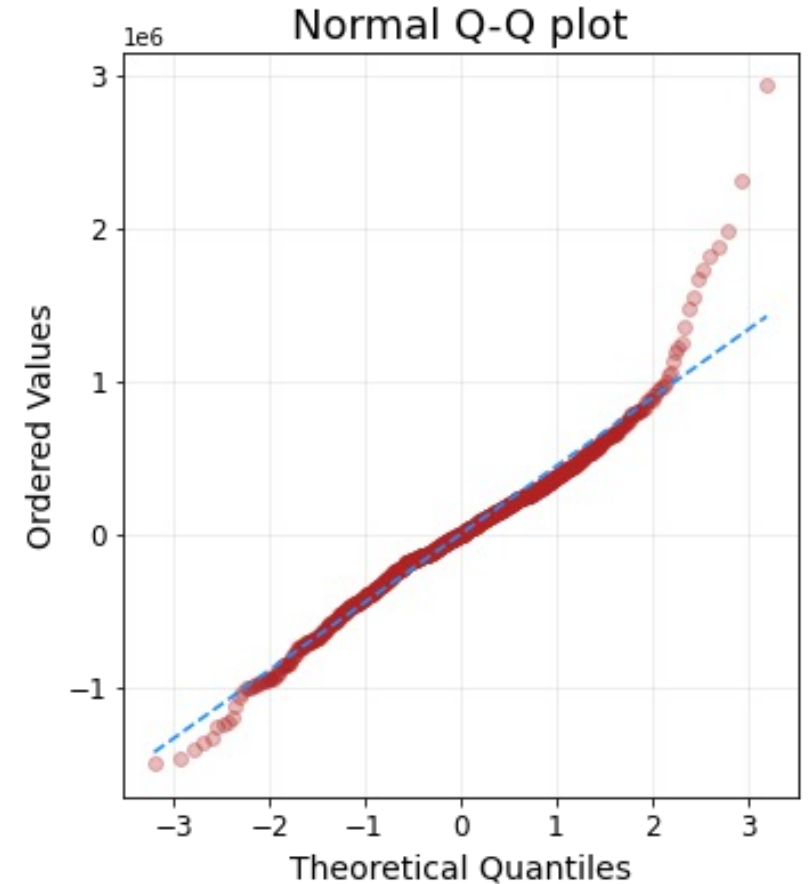
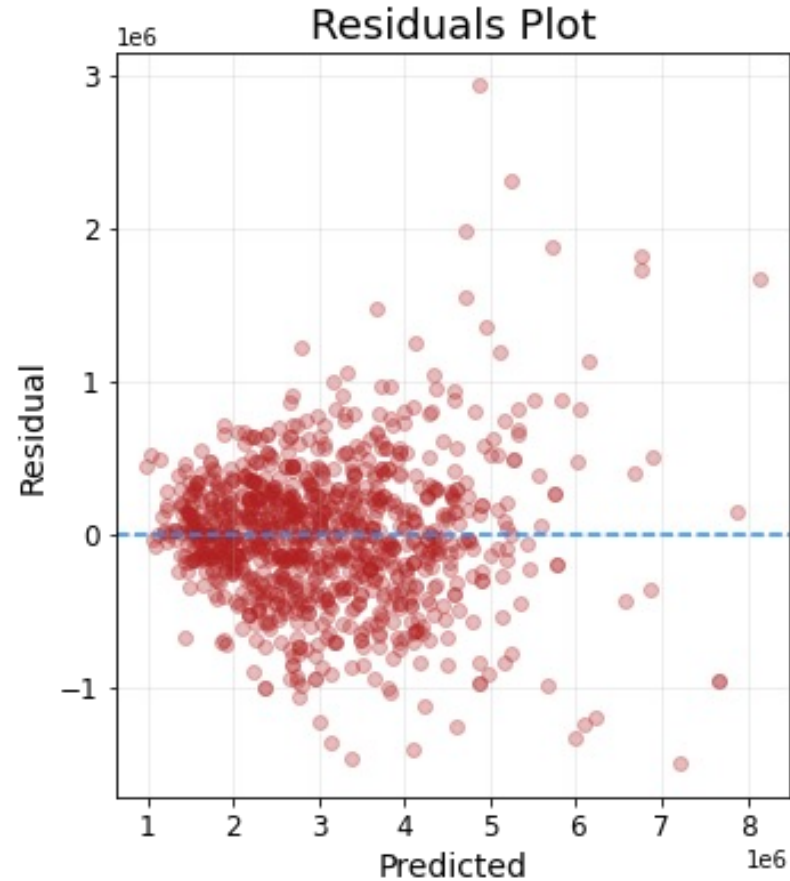
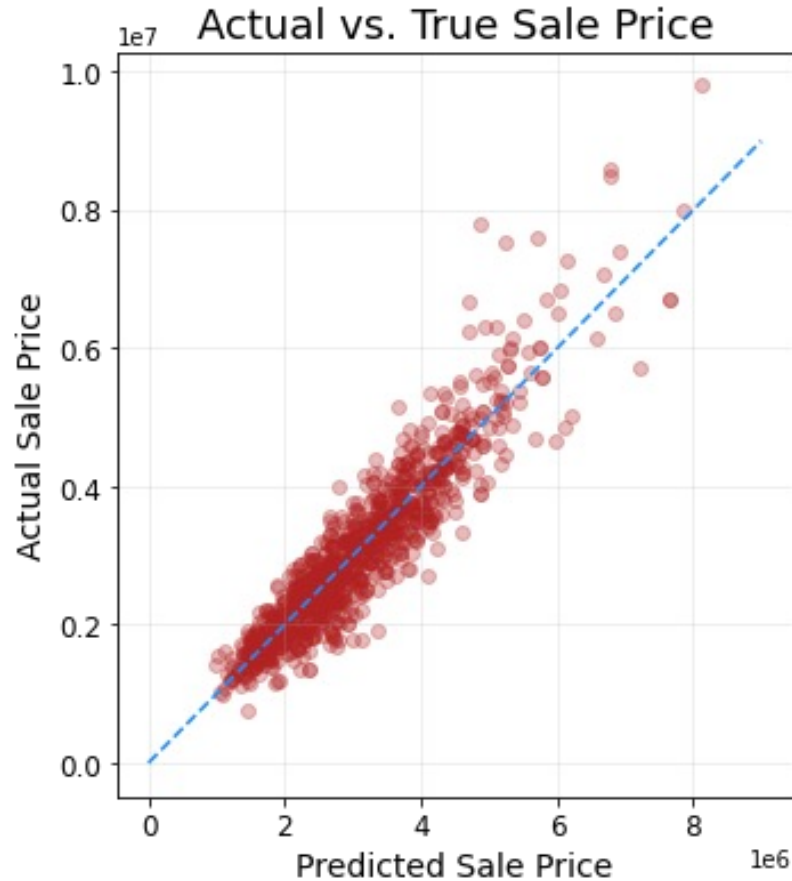

Model Performance (After)

Baseline Linear Regression Results

R-Squared:	0.801
Adjusted R-Squared:	0.774
Mean Absolute Error:	\$398,667
Root Mean Squared Error:	\$543,227



Model Evaluation (After)



* Elastic Net Regression – Test Dataset

R

Conclusion

R^2
0.783



R^2
0.801

MAE
\$416K



MAE
\$398K

RMSE
\$586K



RMSE
\$543K

* Elastic Net Regression – Test Dataset



Conclusion

Interesting Feature Coefficients

home size:	773471.44
school score avg:	269845.16
lot size:	130766.26
*city - palo alto:	336679.56
*city – los altos:	205731.06
age of house:	0.00

* City – Cupertino as reference



Thank You

Questions?

Metis May 2022

Thank You

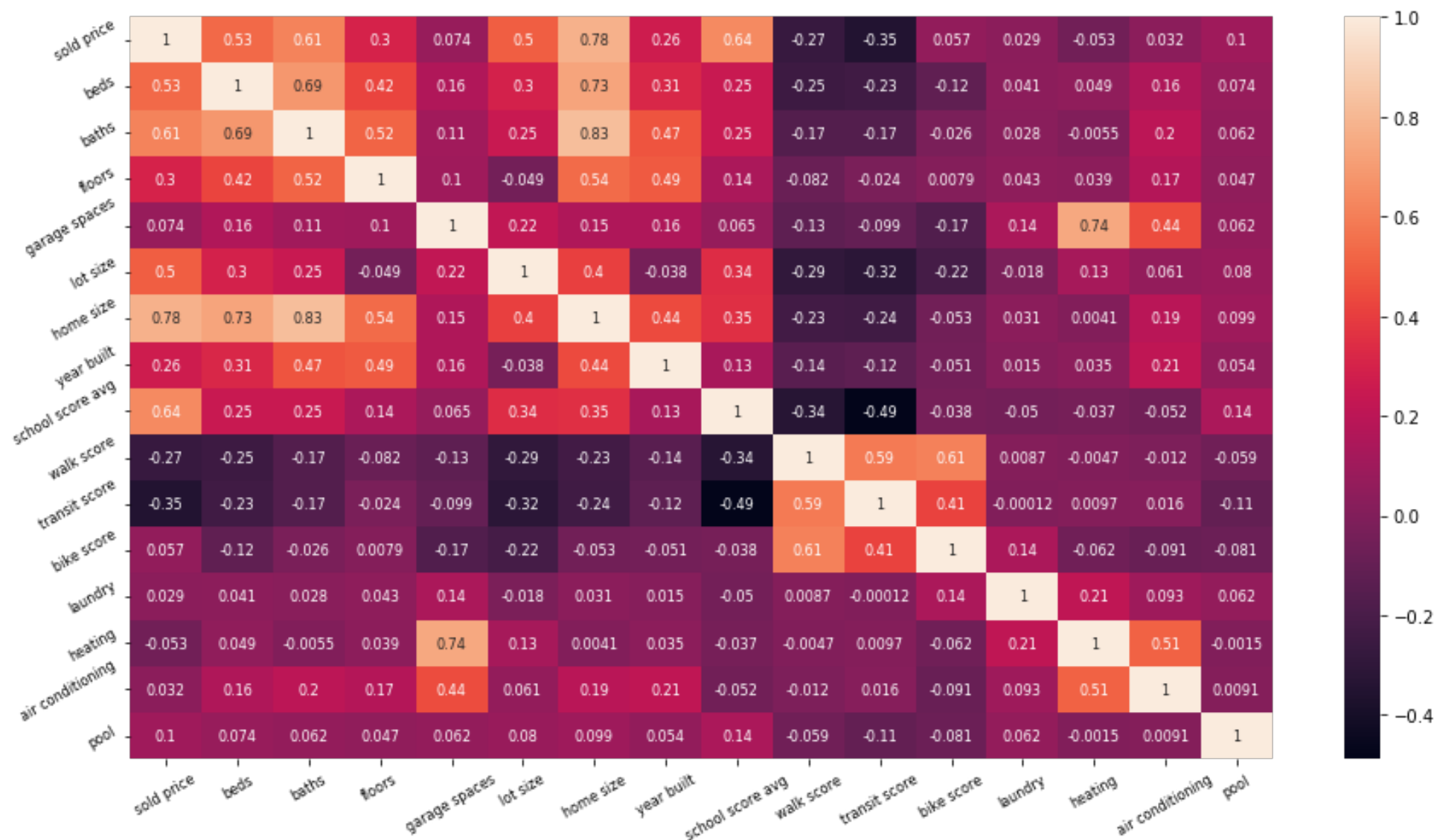
REDFIN

Questions?

metis 2022 may

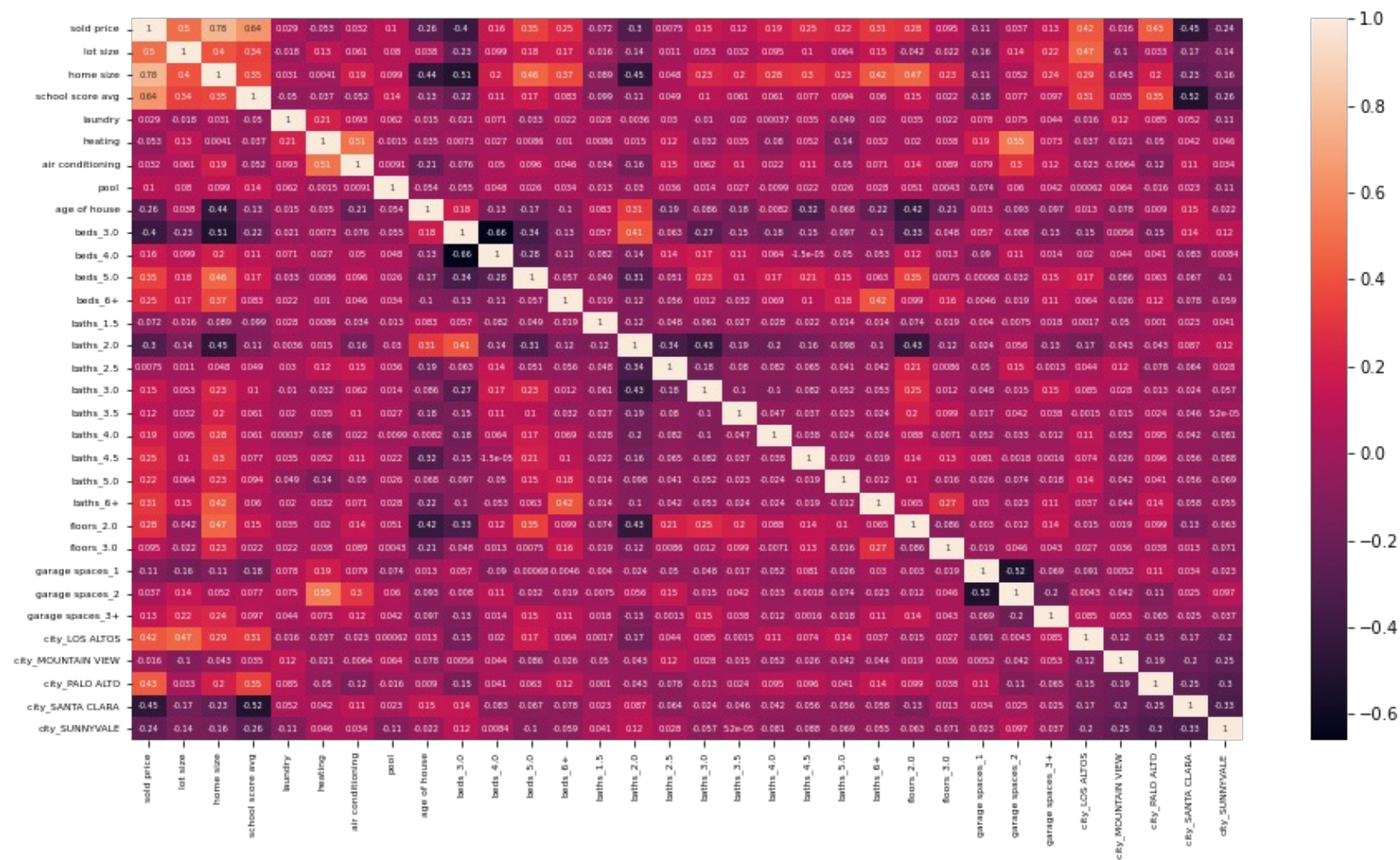
R

Appendix



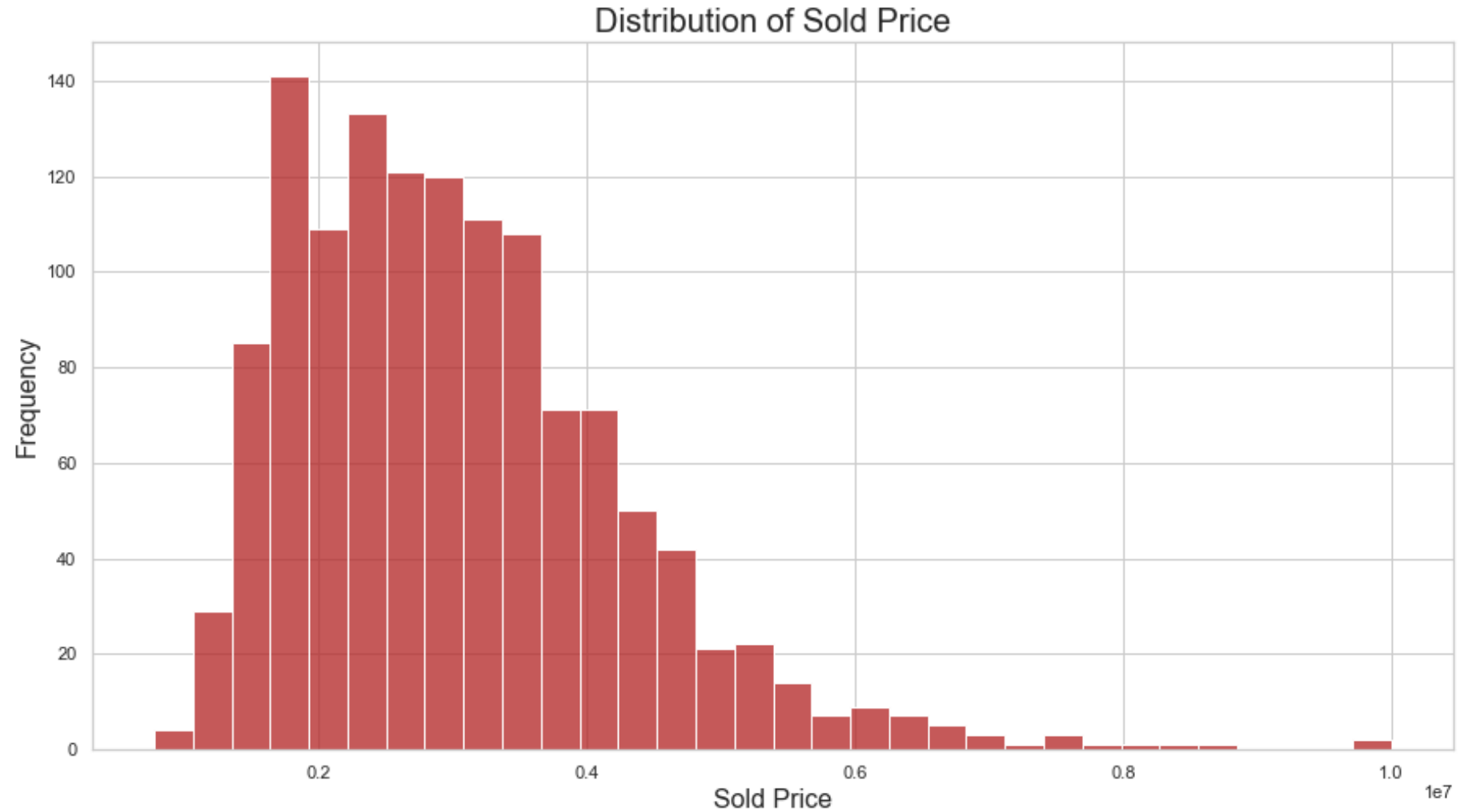
R

Appendix





Appendix



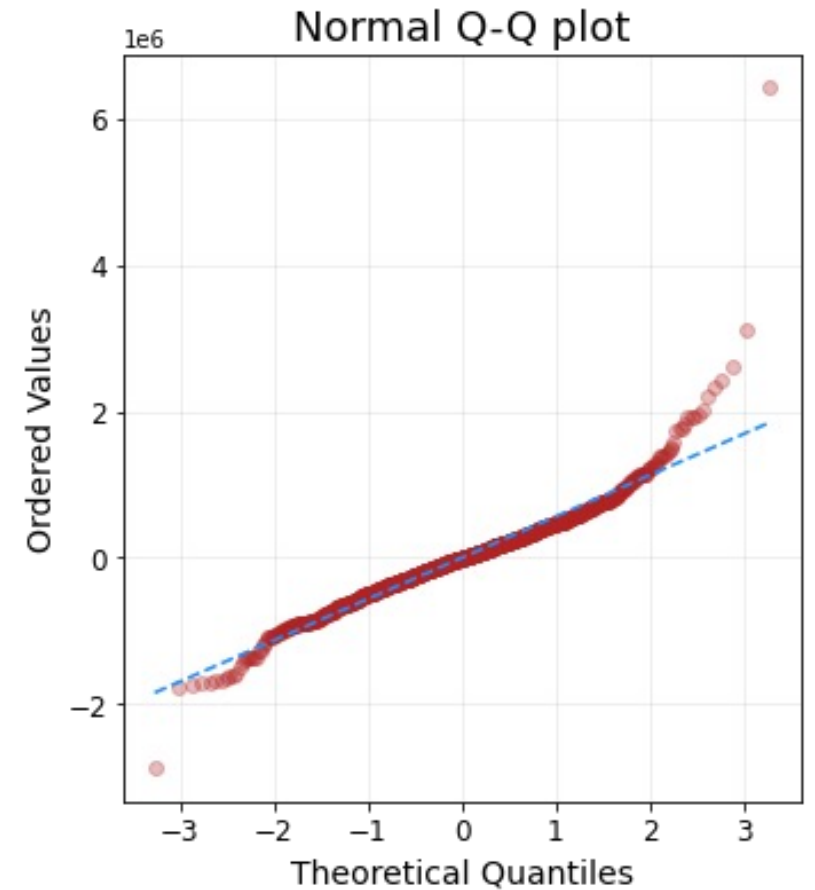
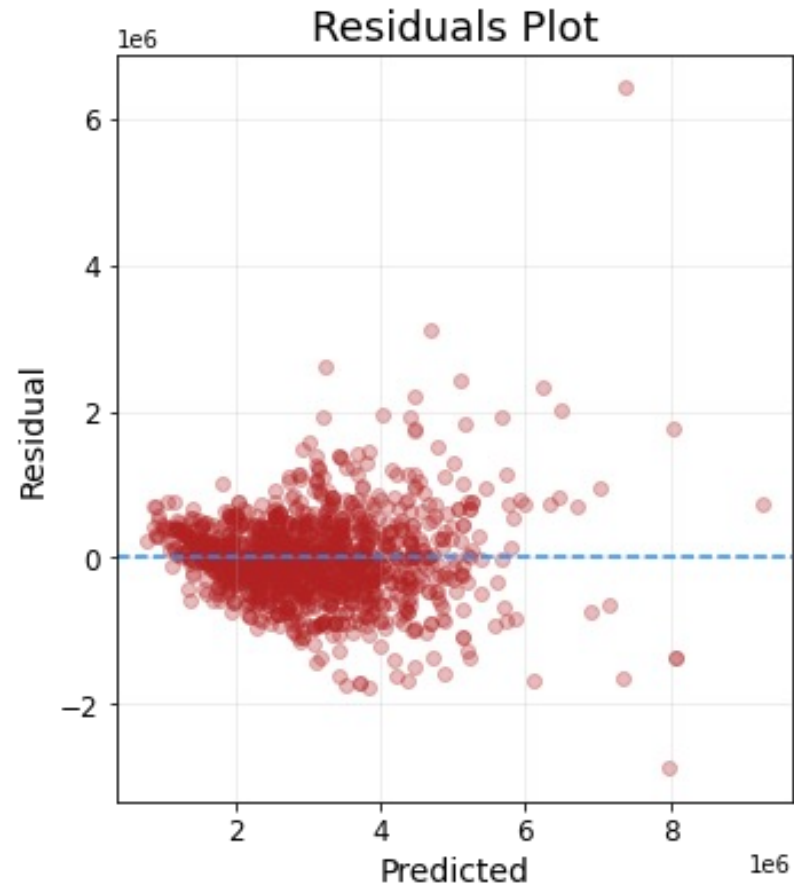
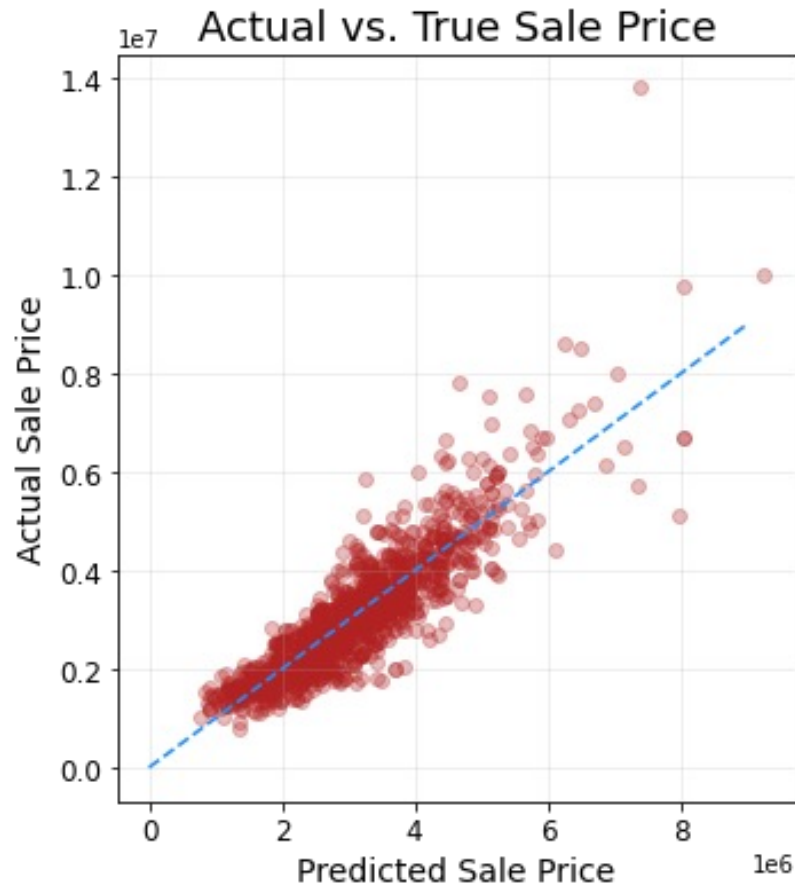


Appendix





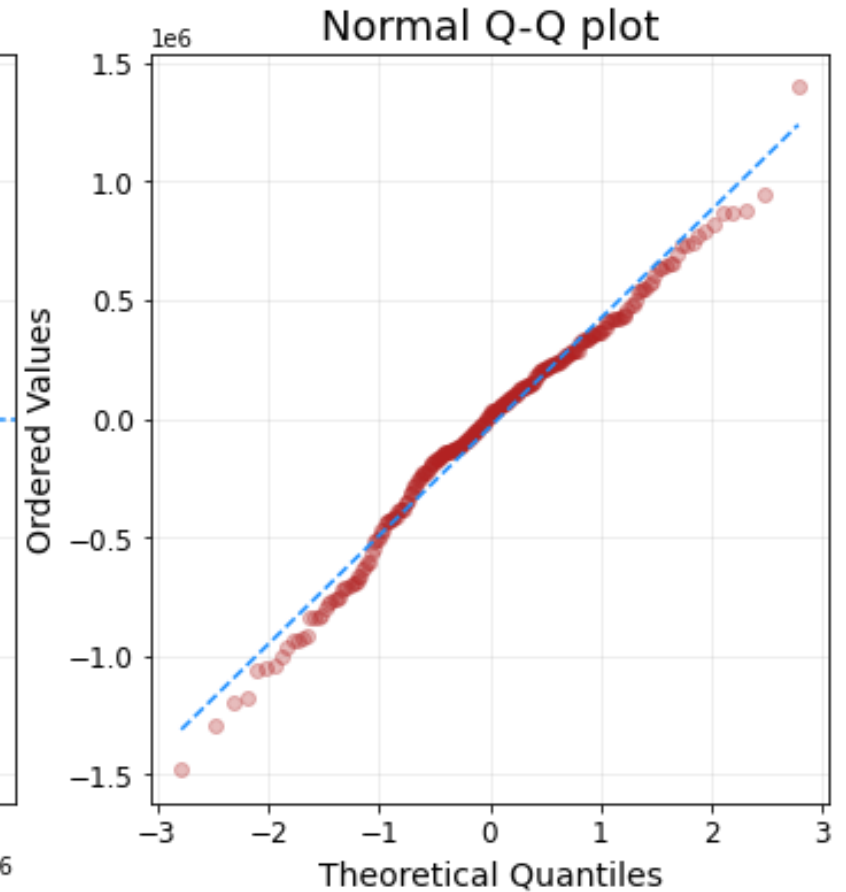
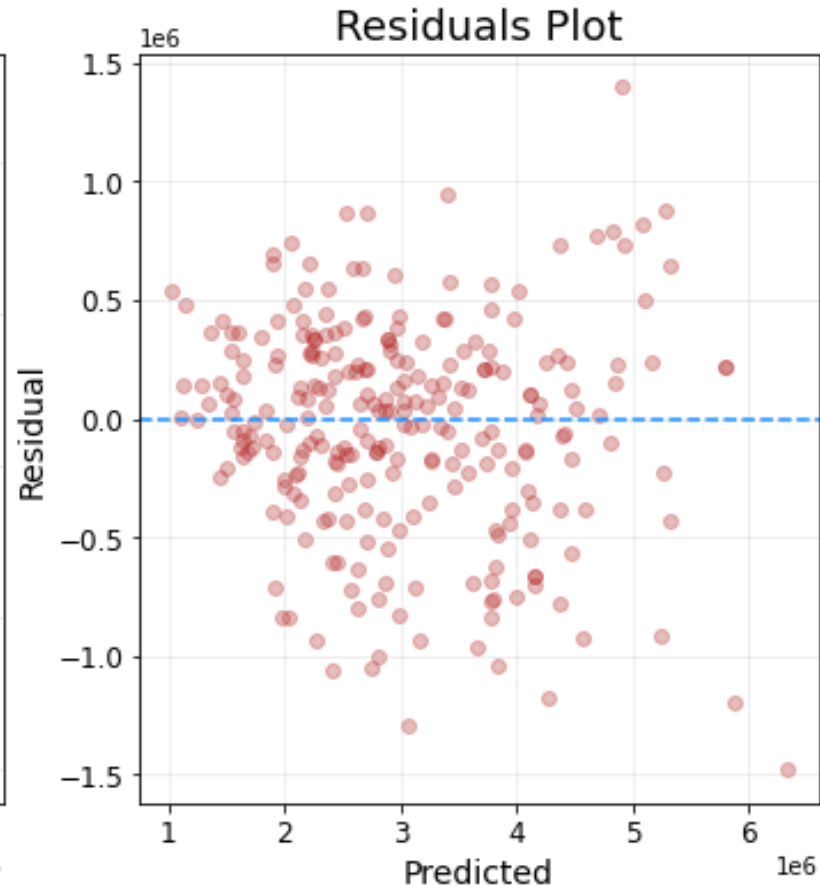
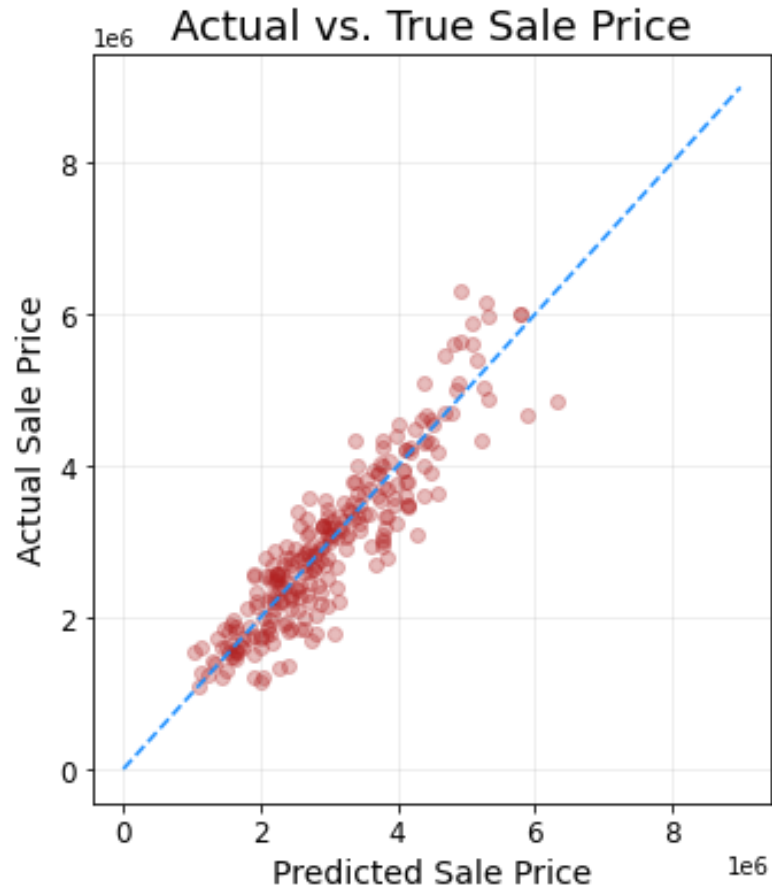
Appendix



* Baseline Linear Regression – Untouched Dataset



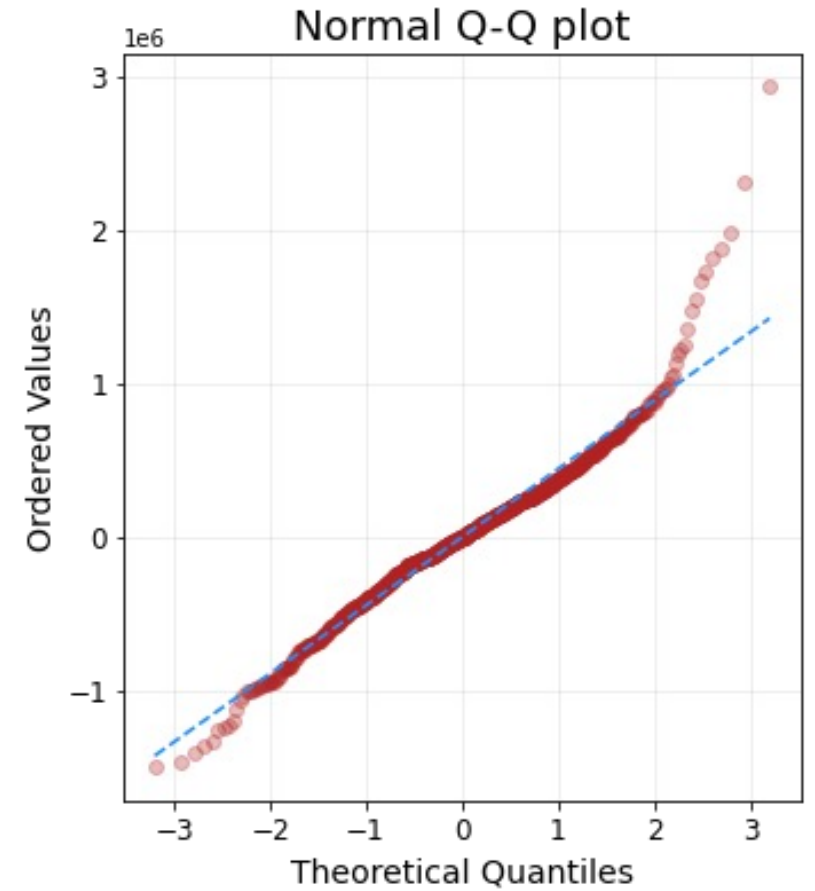
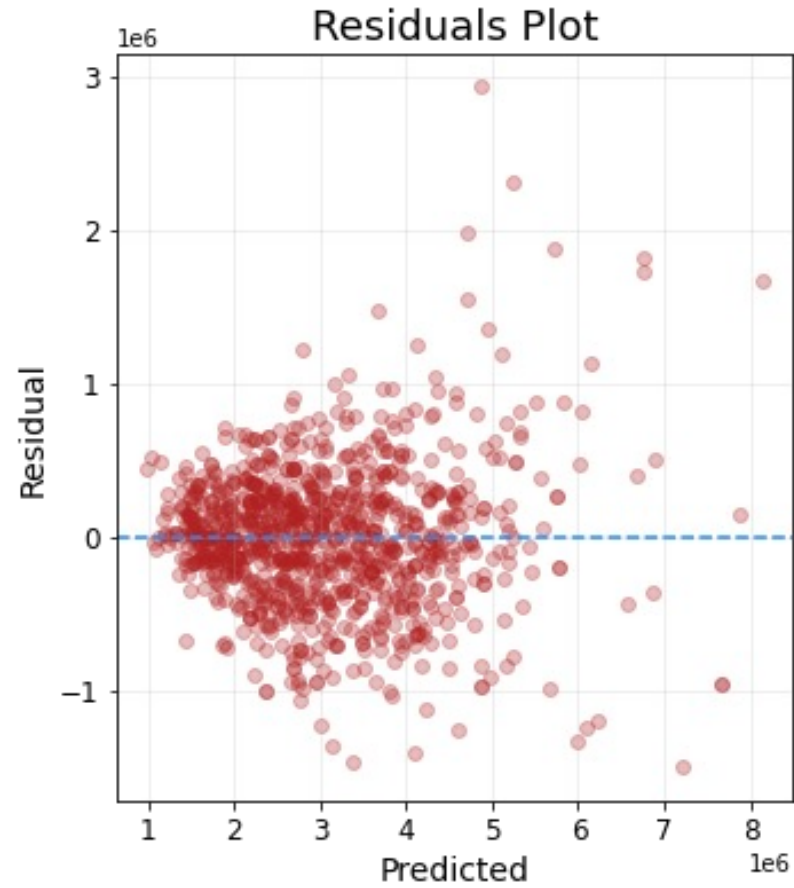
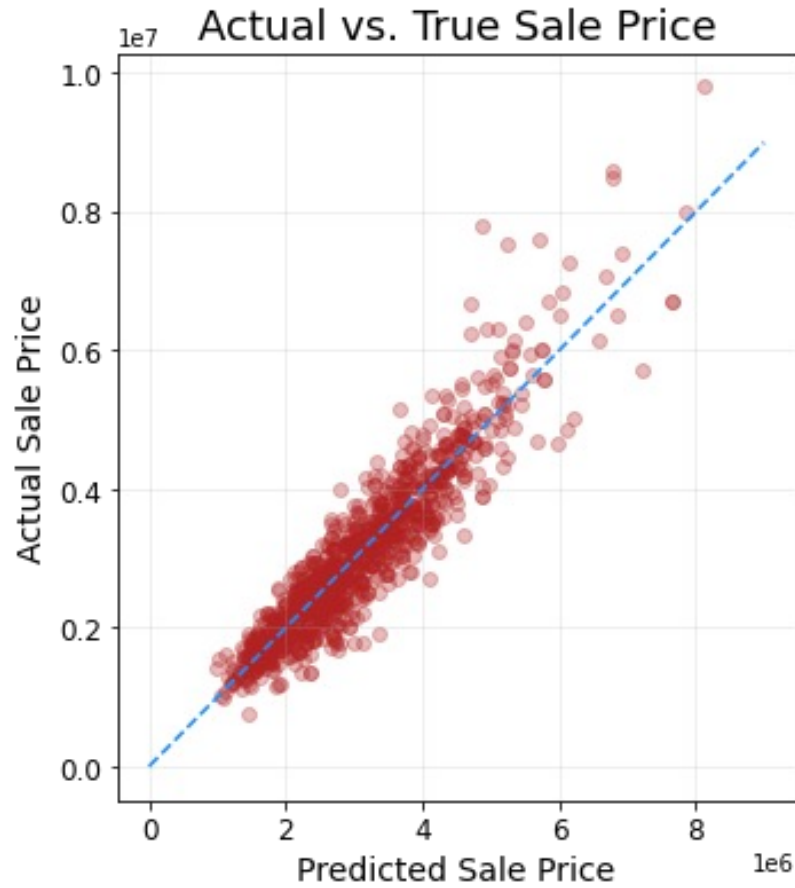
Appendix



* Elastic Net Regression – Validation Dataset



Appendix



* Elastic Net Regression – Test Dataset



Appendix

```
lot size : 130766.26
home size : 773471.44
school score avg : 269845.16
laundry : 3076.32
heating : -39490.85
air conditioning : -18532.17
pool : 15432.62
age of house : 0.00
beds_3.0 : 21291.99
beds_4.0 : 0.00
beds_5.0 : -31173.91
beds_6+ : -71899.62
baths_1.5 : -0.00
baths_2.0 : 0.00
baths_2.5 : -28218.86
baths_3.0 : -0.00
baths_3.5 : 0.00
baths_4.0 : -43796.74
baths_4.5 : 20622.92
baths_5.0 : 37970.59
baths_6+ : 52139.99
floors_2.0 : -36938.23
floors_3.0 : -96698.00
garage spaces_1 : 17165.13
garage spaces_2 : 42698.15
garage spaces_3+ : -37222.90
city_LOS ALTOS : 205731.06
city_MOUNTAIN VIEW : 110700.70
city_PALO ALTO : 336679.56
city_SANTA CLARA : -73019.63
city_SUNNYVALE : 29654.19
```

* Elastic Net Regression – Feature Coefficients