Sticker Shock:
CA Housing Market, Bay Area

REDFIN

Predicting Home Prices with Linear Regression

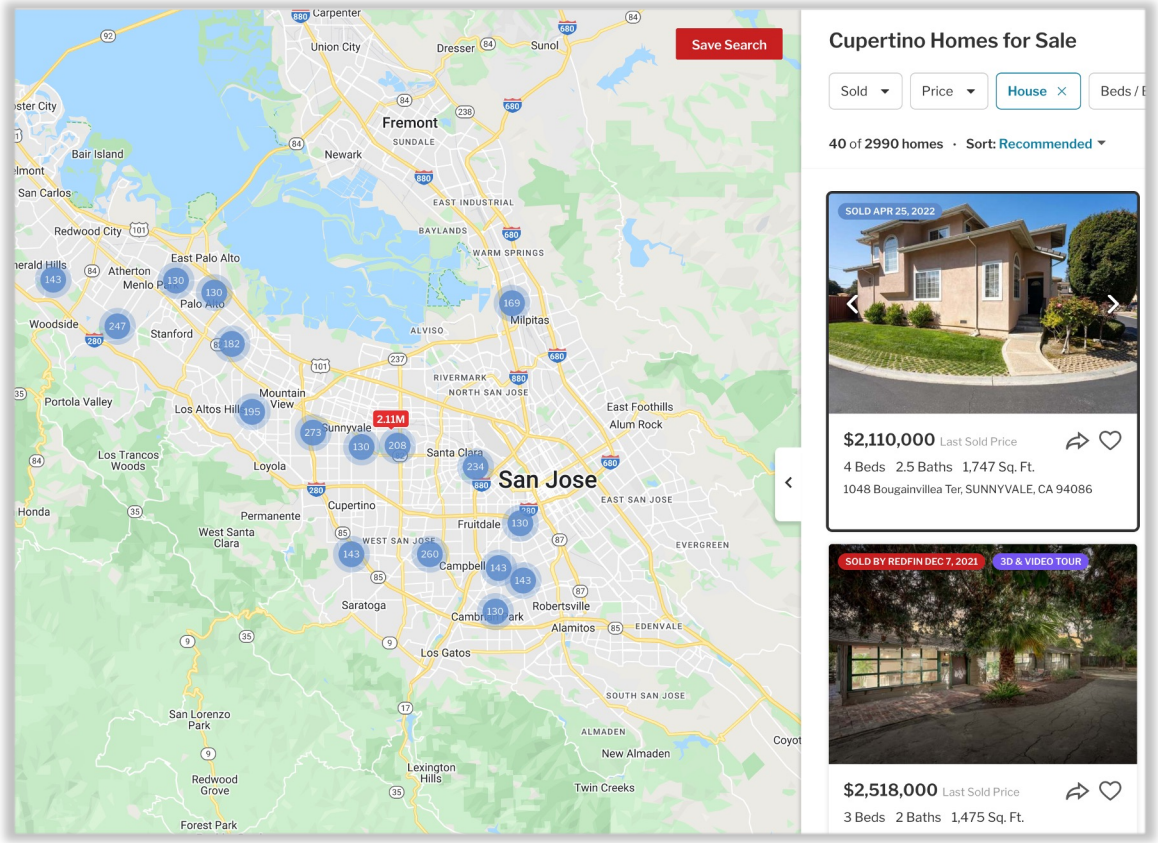
tri le

Why Home Prices?

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

Why Redfin?

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

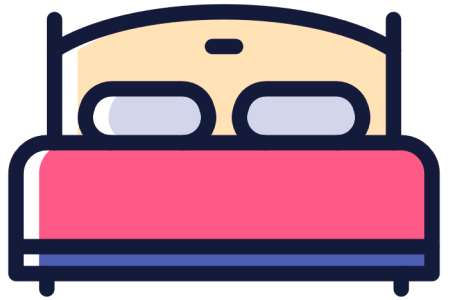


EDA

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

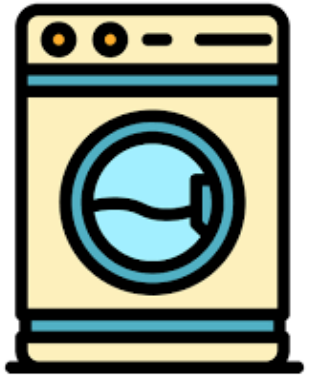
* Palo Alto, Los Altos, Mountain View, Cupertino, Sunnyvale, Santa Clara

Data Point Example



The target: Sold Price

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



Sold Price Stats

Mean:	$3.03 M
Standard Deviation:	$1.23 M
25% Percentile:	$2.10 M
50% Percentile :	$2.83 M
75% Percentile :	$3.70 M

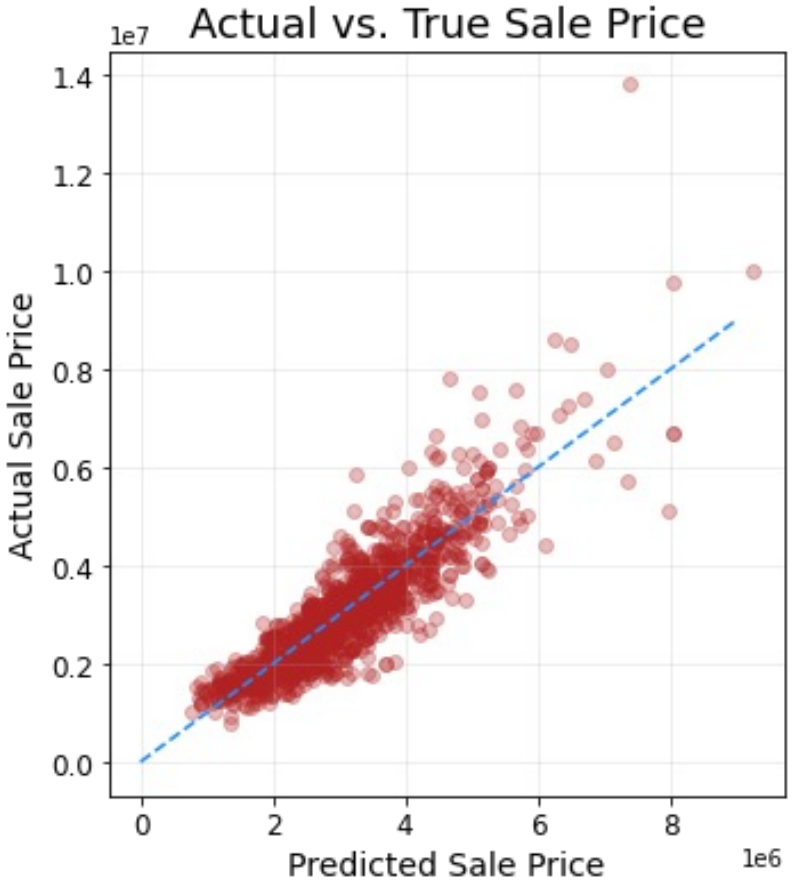
* For Homes in Santa Clara County

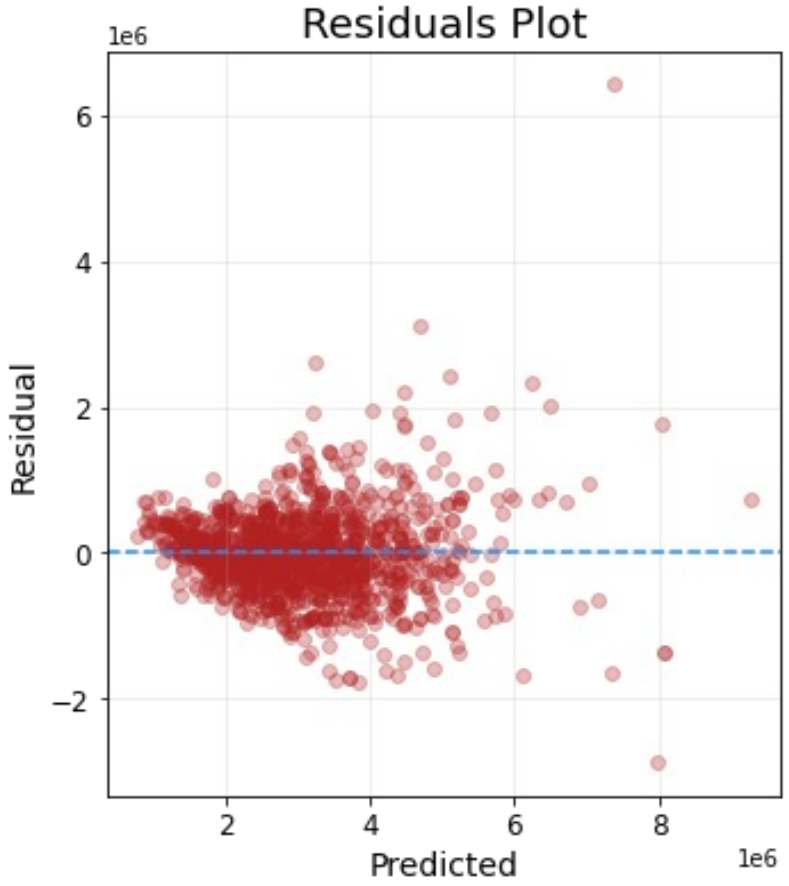
Model Performance (Before)

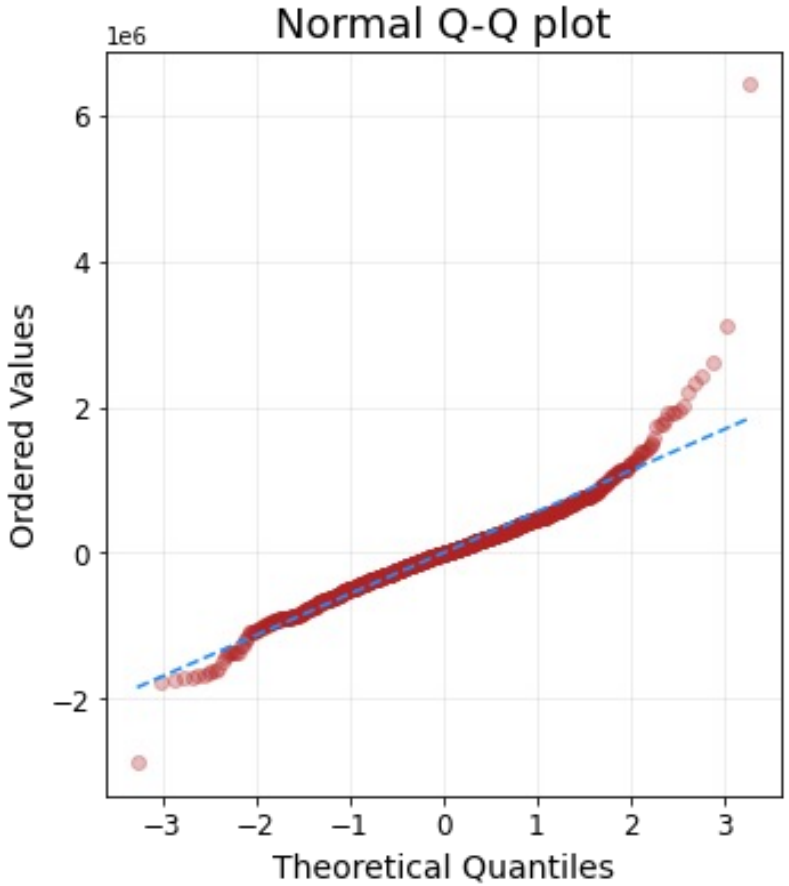
R-Squared:	0.783
Adjusted R-Squared:	0.782
Mean Absolute Error:	$416,244
Root Mean Squared Error:	$586,488

* Baseline Linear Regression - Untouched Dataset

Model Evaluation (Before)







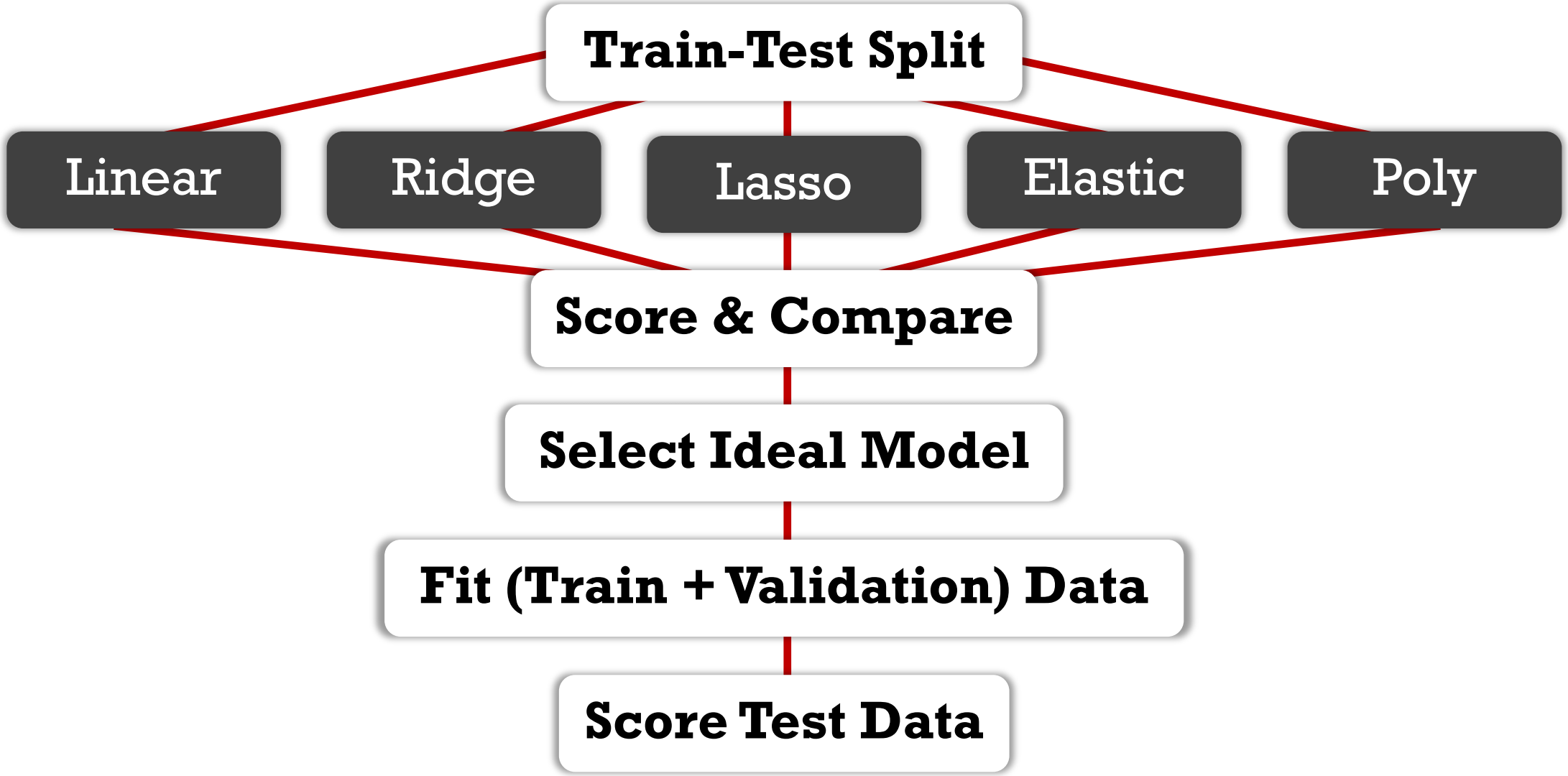
* Baseline Linear Regression - Untouched Dataset

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
```

CV & Regularization

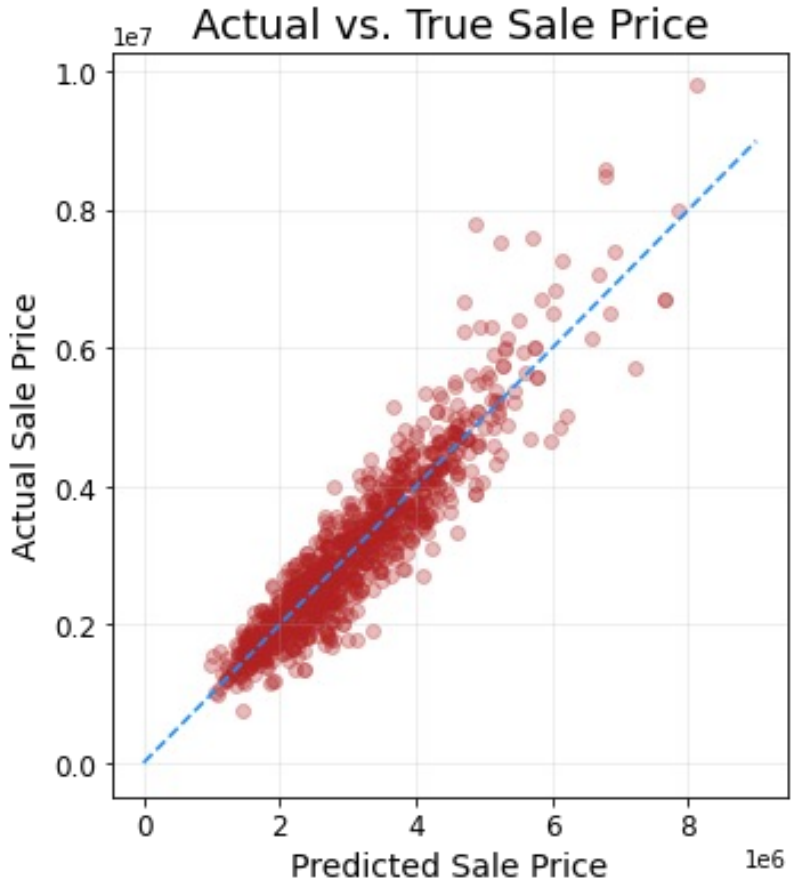


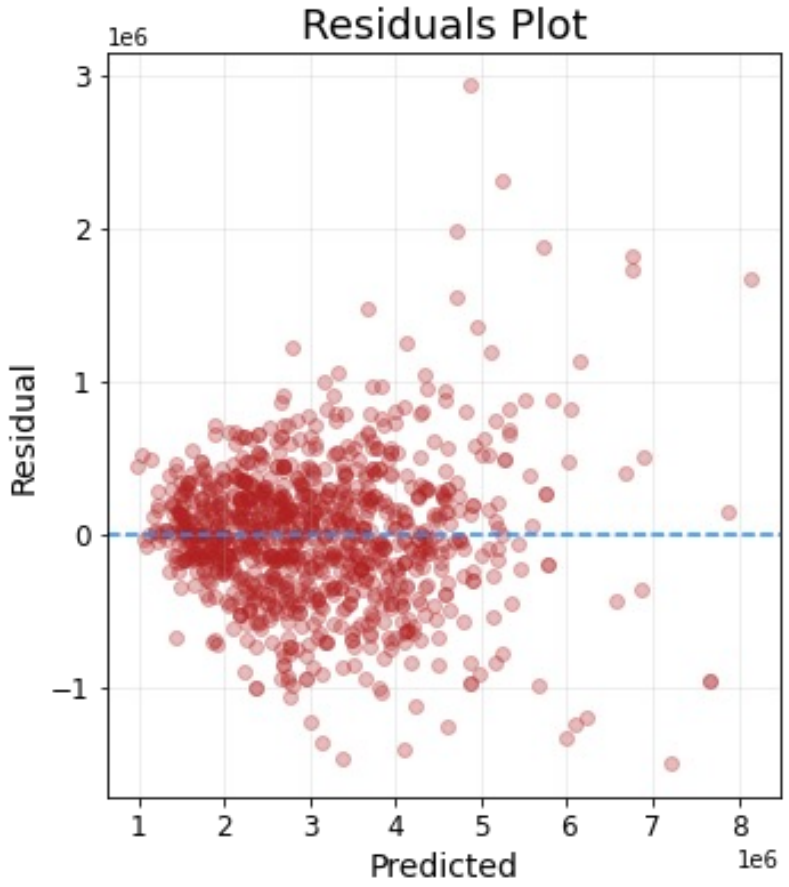
Model Performance (After)

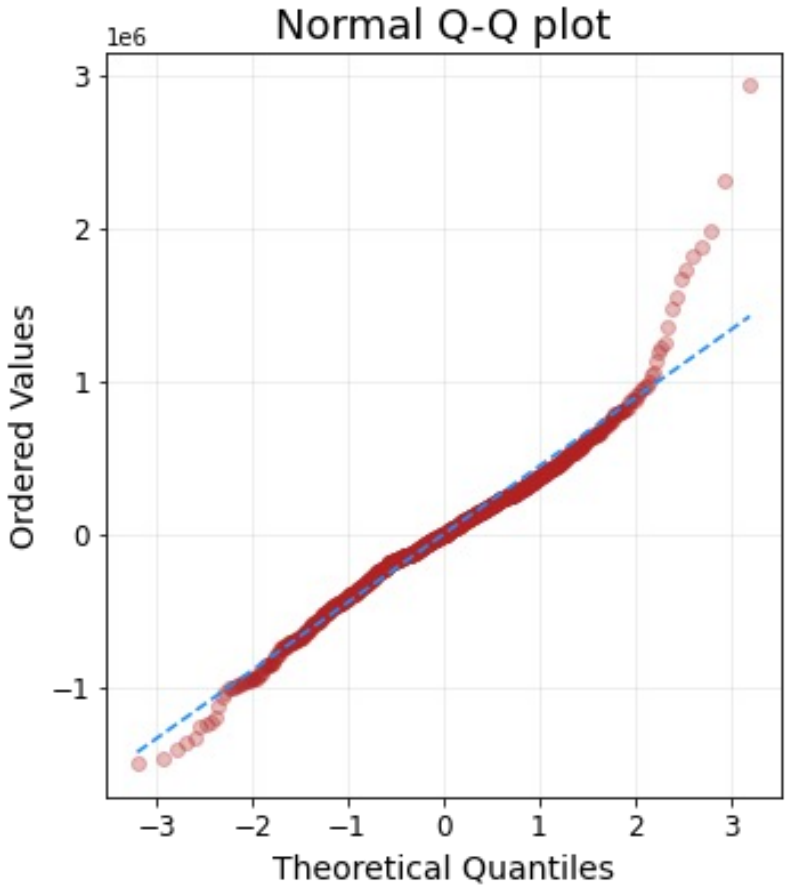
R-Squared:	0.801
Adjusted R-Squared:	0.774
Mean Absolute Error:	$398,667
Root Mean Squared Error:	$543,227

* Elastic Net Regression – Test Dataset

Model Evaluation (After)







* Elastic Net Regression – Test Dataset

Conclusion

* 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

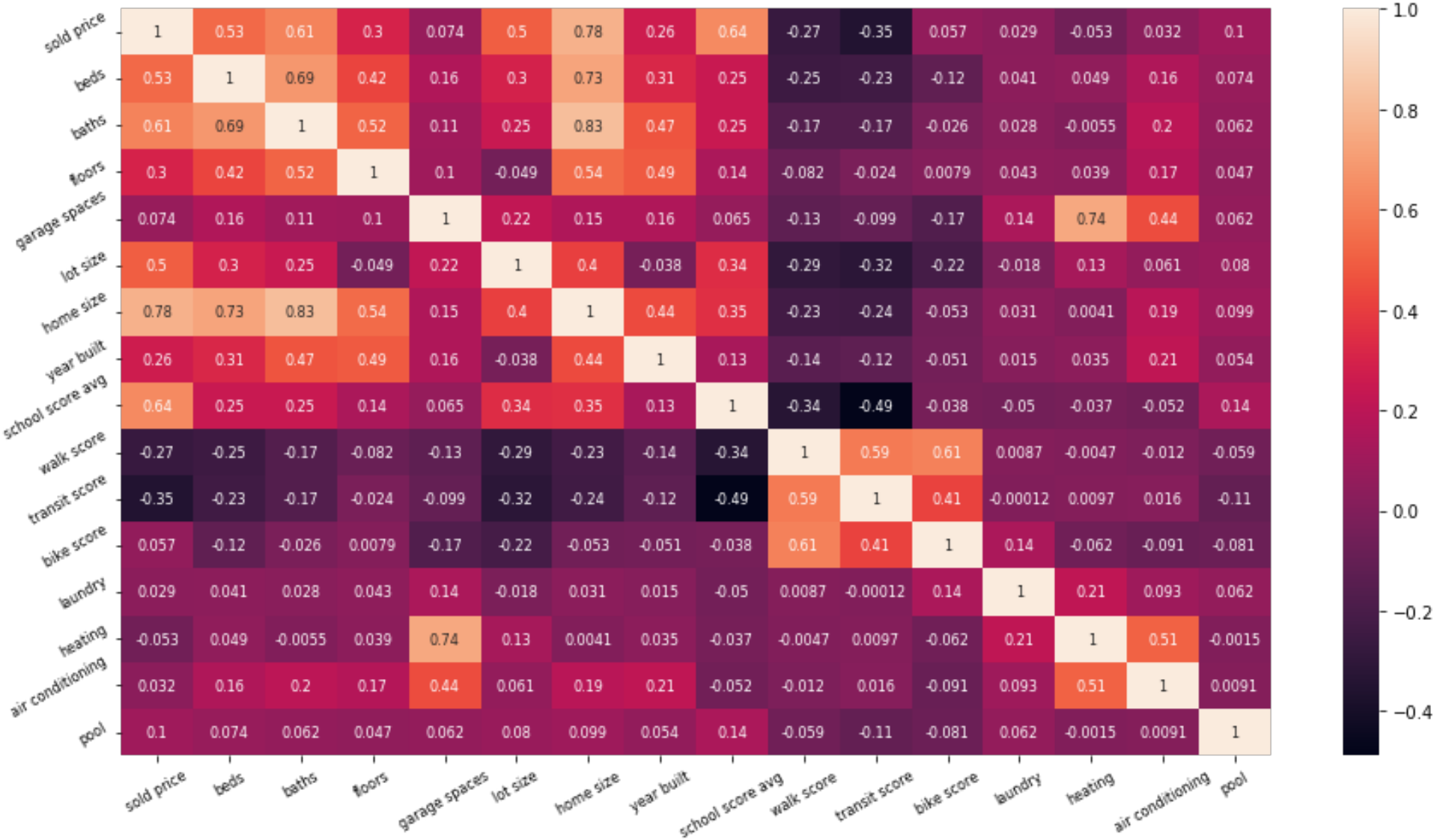
Thank You

REDFIN

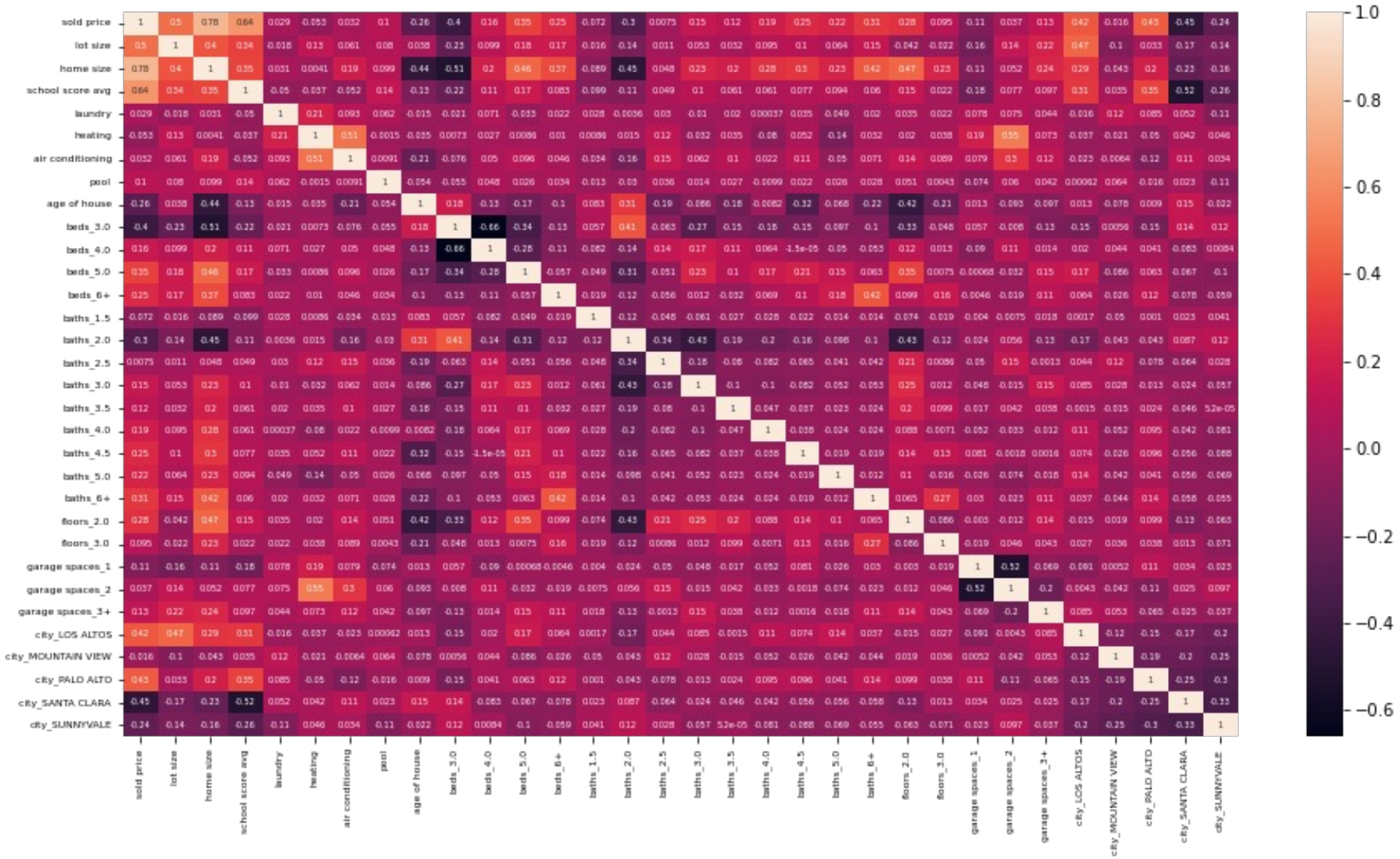
Questions?

metis 2022 may

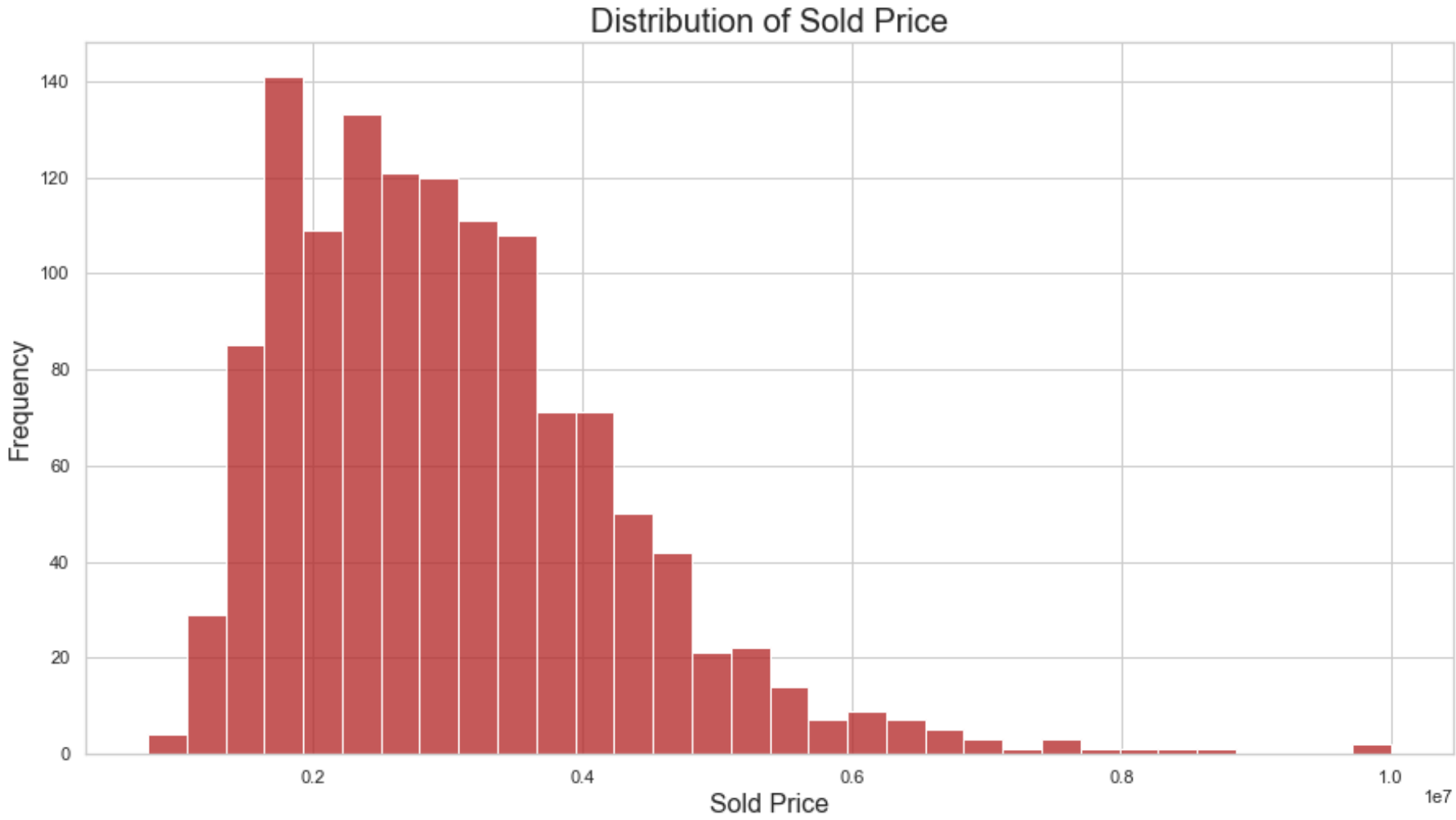
Appendix


sold price 1 0.53 0.61 0.3 0.074 0.5 0.78 0.26 0.64 -0.27 -0.35 0.057 0.029 -0.053 0.032 0.1
beds 0.53 1 0.69 0.42 0.16 0.3 0.73 0.31 0.25 -0.25 -0.23 -0.12 0.041 0.049 0.16 0.074
baths 0.61 0.69 1 0.52 0.11 0.25 0.83 0.47 0.25 -0.17 -0.17 -0.026 0.028 -0.0055 0.2 0.062
floors 0.3 0.42 0.52 1 0.1 -0.049 0.54 0.49 0.14 -0.082 -0.024 0.0079 0.043 0.039 0.17 0.047
garage spaces 0.074 0.16 0.11 0.1 1 0.22 0.15 0.16 0.065 -0.13 -0.099 -0.17 0.14 0.74 0.44 0.062
lot size 0.5 0.3 0.25 -0.049 0.22 1 0.4 -0.038 0.34 -0.29 -0.32 -0.22 -0.018 0.13 0.061 0.08
home size 0.78 0.73 0.83 0.54 0.15 0.4 1 0.44 0.35 -0.23 -0.24 -0.053 0.031 0.0041 0.19 0.099
year built 0.26 0.31 0.47 0.49 0.16 -0.038 0.44 1 0.13 -0.14 -0.12 -0.051 0.015 0.035 0.21 0.054
school score avg 0.64 0.25 0.25 0.14 0.065 0.34 0.35 0.13 1 -0.34 -0.49 -0.038 -0.05 -0.037 -0.052 0.14
walk score -0.27 -0.25 -0.17 -0.082 -0.13 -0.29 -0.23 -0.14 -0.34 1 0.59 0.61 0.0087 -0.0047 -0.012 -0.059
transit score -0.35 -0.23 -0.17 -0.024 -0.099 -0.32 -0.24 -0.12 -0.49 0.59 1 0.41 -0.00012 0.0097 0.016 -0.11
bike score 0.057 -0.12 -0.026 0.0079 -0.17 -0.22 -0.053 -0.051 -0.038 0.61 0.41 1 0.14 -0.062 -0.091 -0.081
laundry 0.029 0.041 0.028 0.043 0.14 -0.018 0.031 0.015 -0.05 0.0087 -0.00012 0.14 1 0.21 0.093 0.062
heating -0.053 0.049 -0.0055 0.039 0.74 0.13 0.0041 0.035 -0.037 -0.0047 0.0097 -0.062 0.21 1 0.51 -0.0015
air conditioning 0.032 0.16 0.2 0.17 0.44 0.061 0.19 0.21 -0.052 -0.012 0.016 -0.091 0.093 0.51 1 0.0091
pool 0.1 0.074 0.062 0.047 0.062 0.08 0.099 0.054 0.14 -0.059 -0.11 -0.081 0.062 -0.0015 0.0091 1
sold price
beds
baths
floors
garage spaces
lot size
home size
year built
school score avg
walk score
transit score
bike score
laundry
heating
air conditioning
pool
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

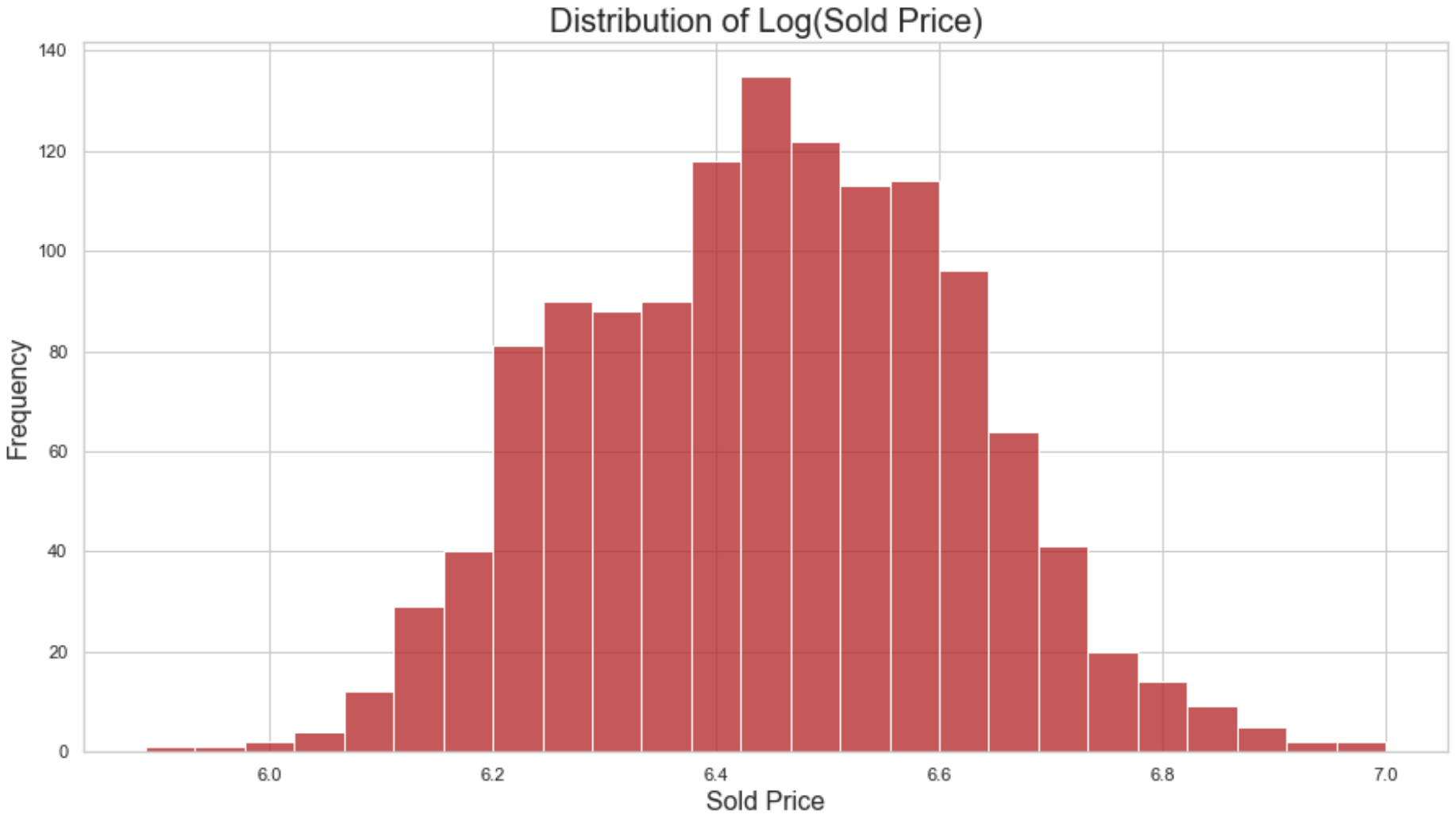
Appendix


sold price
lot size
home size
school score avg
laundry
heating
air conditioning
pool
age of house
beds_3.0
beds_4.0
beds_5.0
beds_6+
baths_1.5
baths_2.0
baths_2.5
baths_3.0
baths_3.5
baths_4.0
baths_4.5
baths_5.0
baths_6+
floors_2.0
floors_3.0
garage spaces_1
garage spaces_2
garage spaces_3+
city_LOS ALTOS
city_MOUNTAIN VIEW
city_PALO ALTO
city_SANTA CLARA
city_SUNNYVALE
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

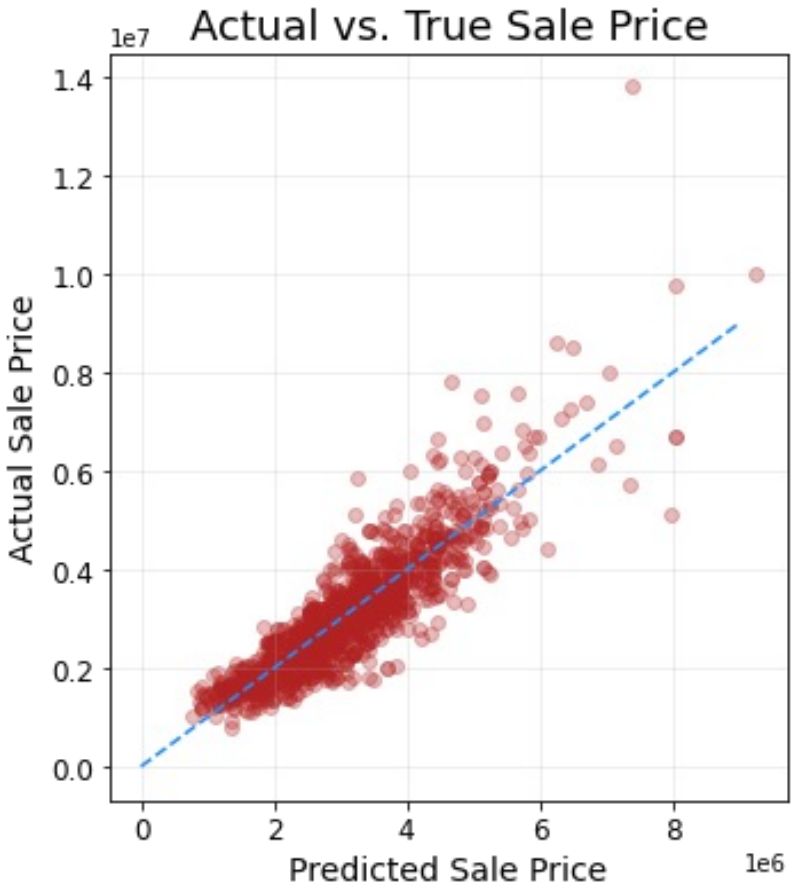
Appendix

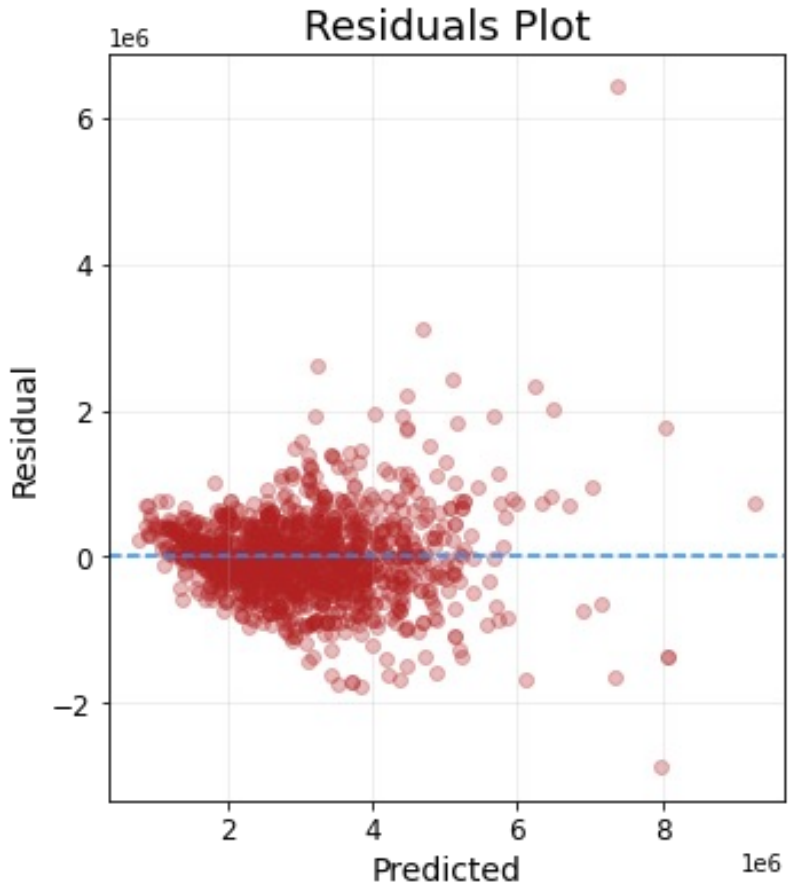

Distribution of Sold Price
Frequency
0
20
40
60
80
100
120
140
0.2
0.4
0.6
0.8
1.0
1e7
Sold Price

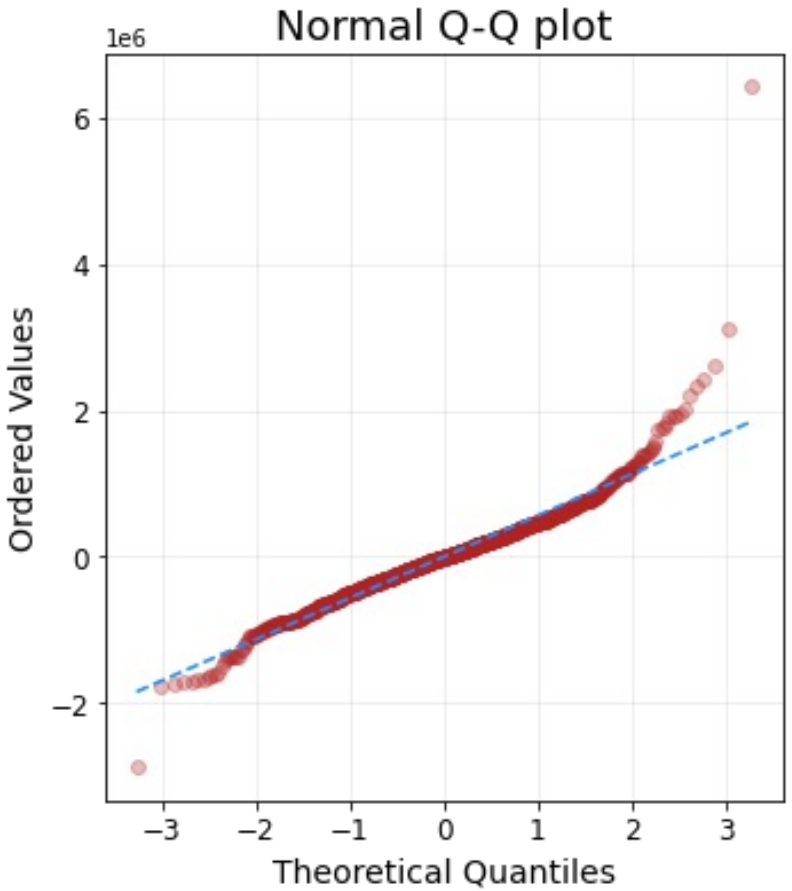
Appendix


Distribution of Log(Sold Price)
Frequency
0
20
40
60
80
100
120
140
6.0
6.2
6.4
6.6
6.8
7.0
Sold Price

Appendix

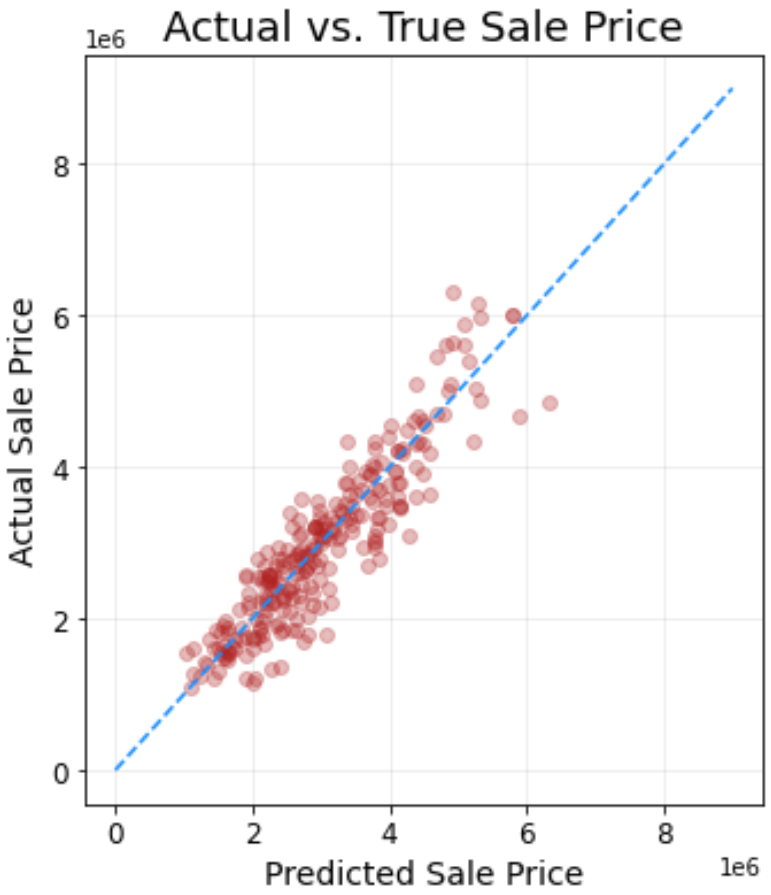


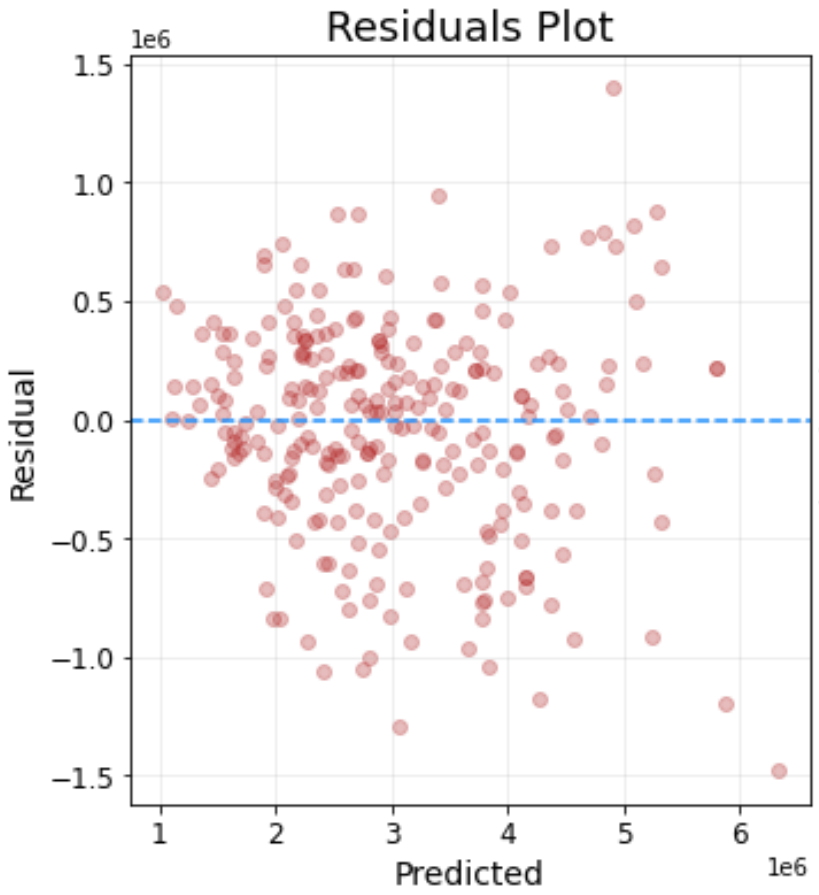


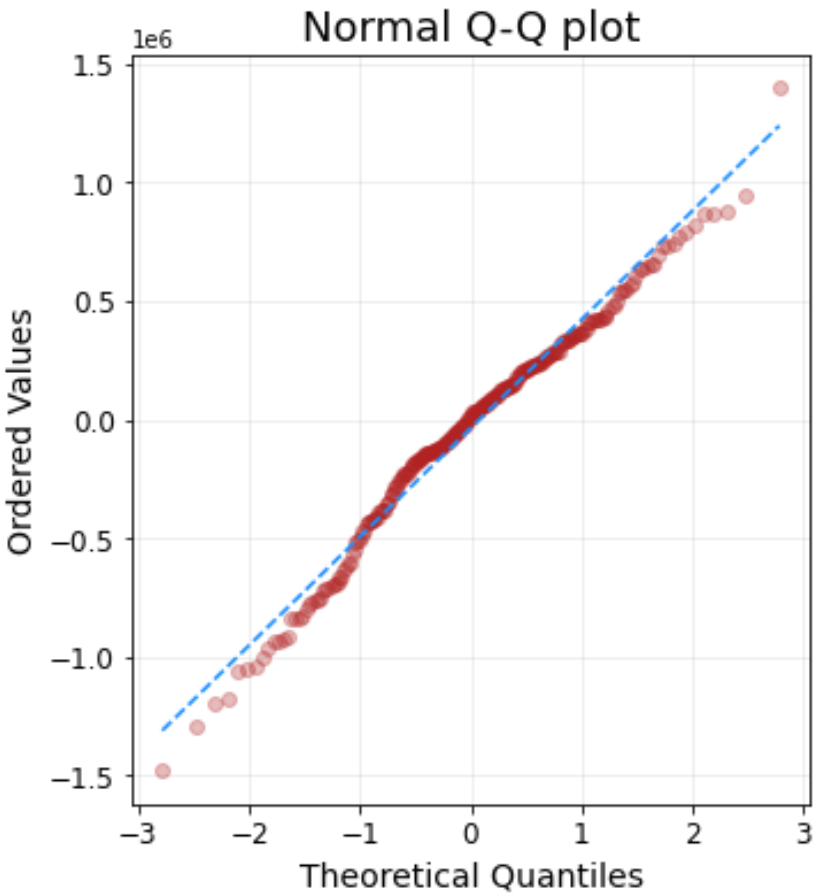


* Baseline Linear Regression – Untouched Dataset

Appendix

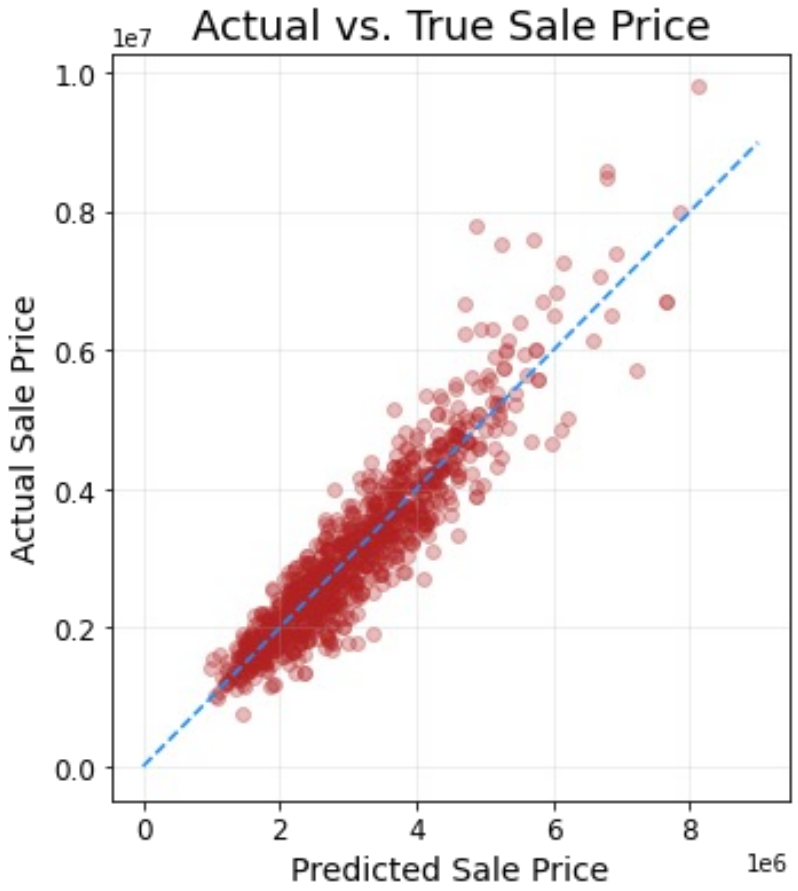


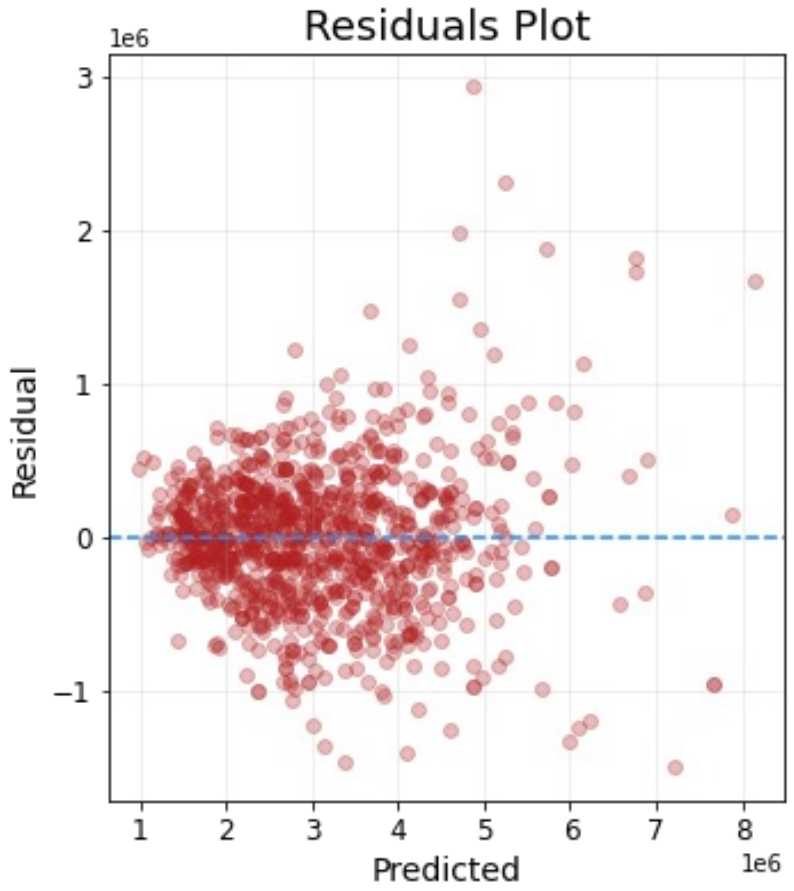


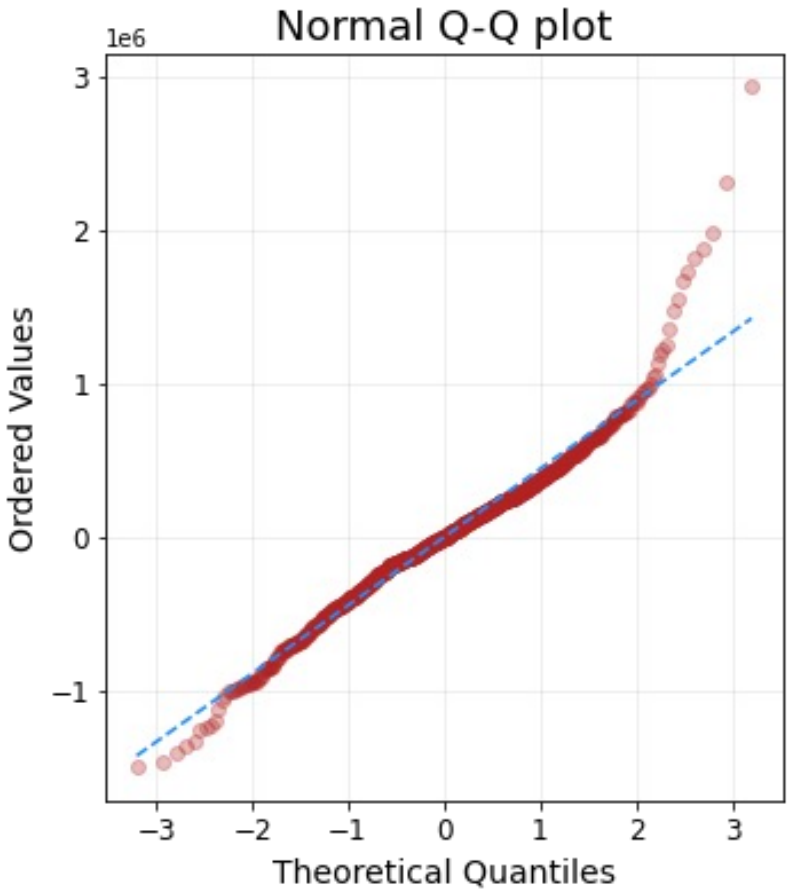


* Elastic Net Regression – Validation Dataset

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