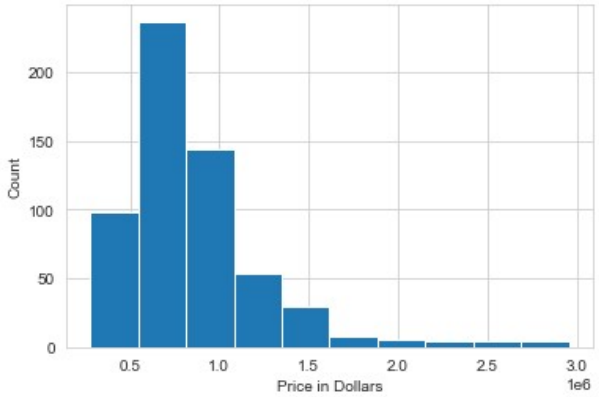
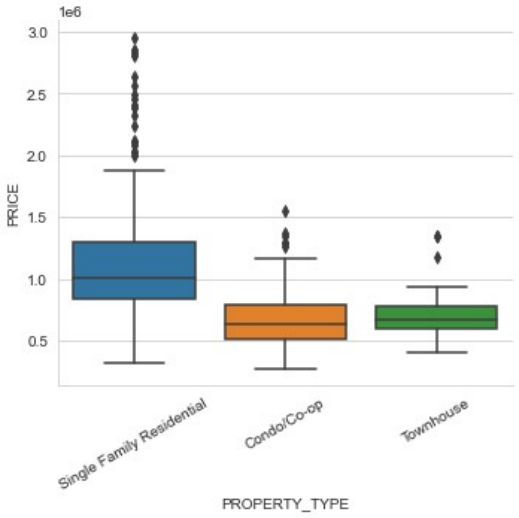
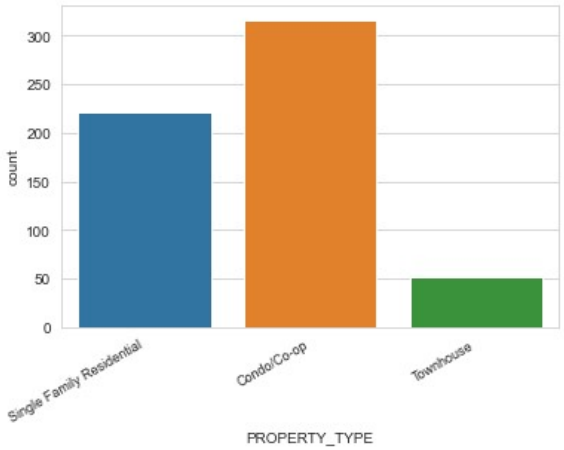
Logistic House Screenshots

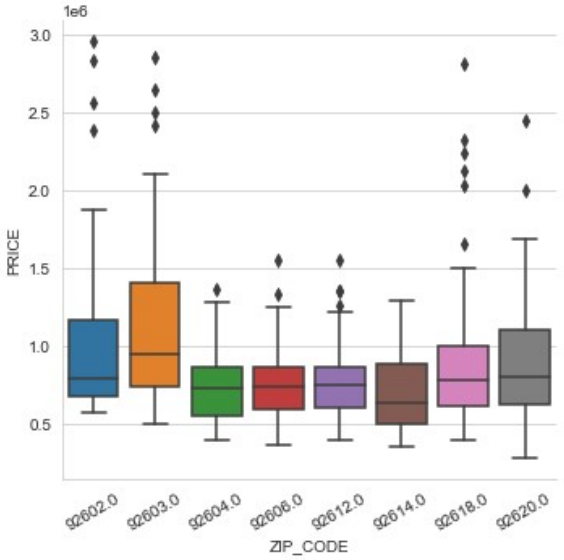
Friday, March 27, 2020 10:48 AM

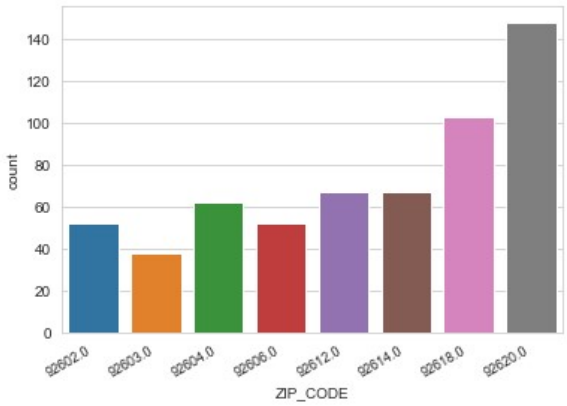
I had access to about 600 houses.
For these visualizations, I removed houses that cost more than 3 million because it condensed the other data. There were 8 houses that cost 3 million or more.

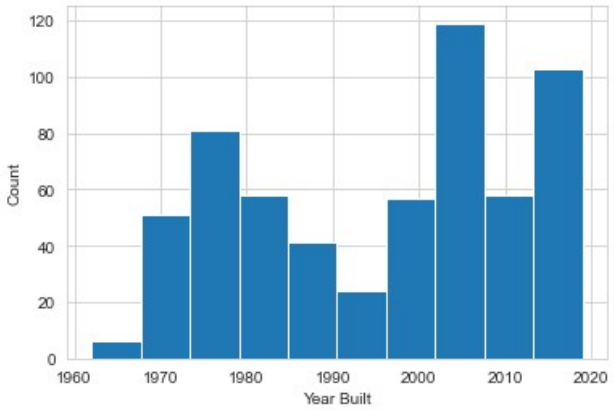


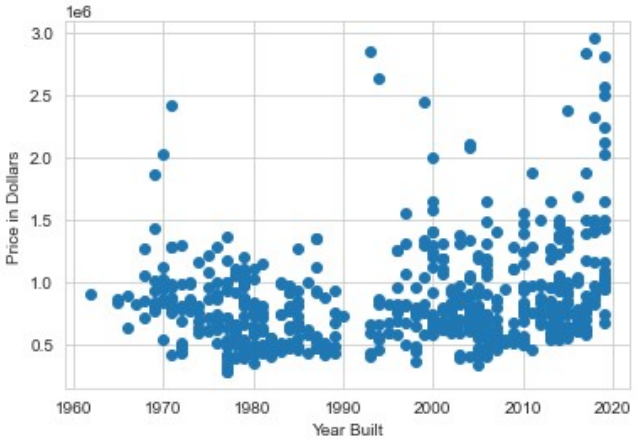


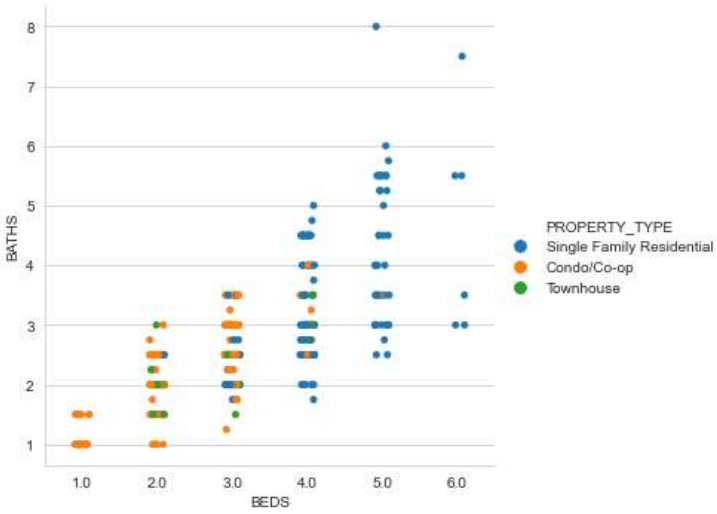

300
250
200
150
100
50
0
count
Single Family Residential
Condo/Co-op
Townhouse
PROPERTY_TYPE


1e6
3.0
2.5
2.0
1.5
1.0
0.5
PRICE
92602.0
92603.0
92604.0
92606.0
92612.0
92614.0
92618.0
92620.0
ZIP_CODE


140
120
100
80
60
40
20
0
count
92602.0
92603.0
92604.0
92606.0
92612.0
92614.0
92618.0
92620.0
ZIP_CODE


Count
Year Built
0
20
40
60
80
100
120
1960
1970
1980
1990
2000
2010
2020


1e6
Price in Dollars
Year Built
0.5
1.0
1.5
2.0
2.5
3.0
1960
1970
1980
1990
2000
2010
2020


BATHS
BEDS
1
2
3
4
5
6
7
8
1.0
2.0
3.0
4.0
5.0
6.0
PROPERTY_TYPE
Single Family Residential
Condo/Co-op
Townhouse

This is the model where I try to fit the model based off Square Footage, Price, and Location. This model performed very poorly and most of the features were useless.

Results: Logit

Model:	Logit	Pseudo R-squared:	0.223
Dependent Variable:	SELLER_HOUSE	AIC:	701.7204
Date:	2020-03-27 10:44	BIC:	837.9217
No. Observations:	598	Log-Likelihood:	-319.86
Df Model:	30	LL-Null:	-411.68
Df Residuals:	567	LLR p-value:	5.4067e-24
Converged:	0.0000	Scale:	1.0000
No. Iterations:	35.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	2.8108	1.3317	2.1108	0.0348	0.2008	5.4208
C(LOCATION)[T.AA - Airport Area]	-0.3186	1.3132	-0.2426	0.8083	-2.8923	2.2552
C(LOCATION)[T.CG - Columbus Grove]	2.2549	1.2940	1.7426	0.0814	-0.2813	4.7910
C(LOCATION)[T.CV - Cypress Village]	2.8699	1.1363	2.5257	0.0115	0.6429	5.0970
C(LOCATION)[T.EASTW - Eastwood]	2.0308	1.5988	1.2702	0.2040	-1.1027	5.1644
C(LOCATION)[T.EC - El Camino Real]	2.1875	1.1414	1.9164	0.0553	-0.0497	4.4247
C(LOCATION)[T.GP - Great Park]	1.9404	1.1957	1.6229	0.1046	-0.4030	4.2839
C(LOCATION)[T.IRSP - Irvine Spectrum]	-13.5080	2681.5155	-0.0050	0.9960	-5269.1818	5242.1658
C(LOCATION)[T.LGA - Laguna Altura]	-17.5462	23147.1962	-0.0008	0.9994	-45385.2171	45350.1247
C(LOCATION)[T.NK - Northpark]	1.8654	1.1514	1.6201	0.1052	-0.3913	4.1220
C(LOCATION)[T.NW - Northwood]	2.0841	1.1062	1.8841	0.0595	-0.0839	4.2522
C(LOCATION)[T.OC - Oak Creek]	0.2345	1.5293	0.1534	0.8781	-2.7628	3.2319
C(LOCATION)[T.OH - Orchard Hills]	0.9151	1.5102	0.6059	0.5446	-2.0449	3.8751
C(LOCATION)[T.OT - Orangetree]	1.6069	1.2842	1.2513	0.2108	-0.9101	4.1238
C(LOCATION)[T.PS - Portola Springs]	2.2351	1.1205	1.9947	0.0461	0.0389	4.4313
C(LOCATION)[T.QH - Quail Hill]	2.1841	1.2653	1.7261	0.0843	-0.2959	4.6640
C(LOCATION)[T.SH - Shady Canyon]	-12.7712	34706.9129	-0.0004	0.9997	-68037.0705	68011.5281
C(LOCATION)[T.SJ - Rancho San Joaquin]	-19.7960	35734.5829	-0.0006	0.9996	-70058.2915	70018.6994
C(LOCATION)[T.STG - Stonegate]	1.9981	1.2904	1.5484	0.1215	-0.5311	4.5272
C(LOCATION)[T.Stonegate]	-13.0883	2672.5018	-0.0049	0.9961	-5251.0955	5224.9189
C(LOCATION)[T.TR - Turtle Rock]	1.0944	1.2810	0.8543	0.3929	-1.4164	3.6051
C(LOCATION)[T.TRG - Turtle Ridge]	2.0252	1.2913	1.5684	0.1168	-0.5057	4.5561
C(LOCATION)[T.UP - University Park]	1.9223	1.1544	1.6652	0.0959	-0.3402	4.1849
C(LOCATION)[T.UT - University Town Center]	2.3439	1.2721	1.8426	0.0654	-0.1493	4.8371
C(LOCATION)[T.WB - Woodbridge]	1.5407	1.1069	1.3919	0.1640	-0.6288	3.7102
C(LOCATION)[T.WD - Woodbury]	3.7453	1.1314	3.3102	0.0009	1.5277	5.9628
C(LOCATION)[T.WI - West Irvine]	1.0027	1.2665	0.7917	0.4285	-1.4796	3.4851
C(LOCATION)[T.WN - Walnut (Irvine)]	2.6916	1.1957	2.2511	0.0244	0.3481	5.0351
C(LOCATION)[T.WP - Westpark]	2.4560	1.1195	2.1937	0.0283	0.2617	4.6502
SQFT_PER	-0.0089	0.0016	-5.4898	0.0000	-0.0121	-0.0057
PRICE	-0.0000	0.0000	-3.0629	0.0022	-0.0000	-0.0000

Accuracy of logistic regression classifier on test set: 0.528

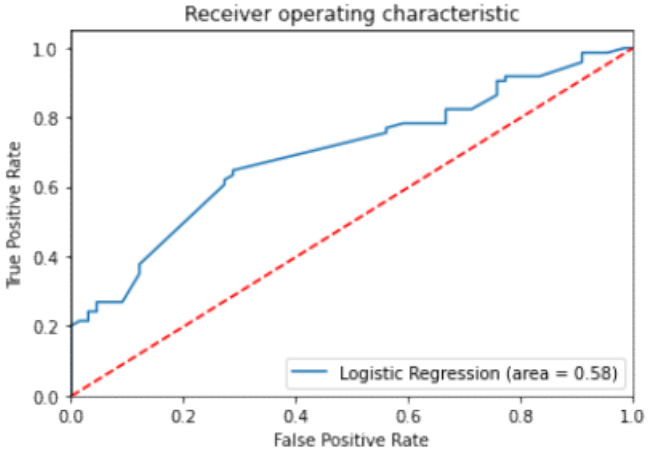
I used recursive feature elimination to create a model.

```
                           Results: Logit
============================================================================
Model:                   Logit              Pseudo R-squared:   0.158
Dependent Variable:      SELLER_HOUSE       AIC:                583.7447
Date:                    2020-03-27 10:53   BIC:                666.6284
No. Observations:        466                Log-Likelihood:     -271.87
Df Model:                19                 LL-Null:            -323.01
Df Residuals:            446                LLR p-value:        2.0756e-13
Converged:               1.0000             Scale:              1.0000
No. Iterations:          7.0000
----------------------------------------------------------------------------
                              Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
----------------------------------------------------------------------------
BEDS                         -0.0652   0.1718 -0.3798 0.7041 -0.4019  0.2714
BATHS                         0.4964   0.2067  2.4009 0.0164  0.0912  0.9016
SQFT_PER                     -0.0015   0.0009 -1.5392 0.1237 -0.0033  0.0004
PT_Condo/Co-op                0.5588   0.5355  1.0435 0.2967 -0.4908  1.6085
PT_Single Family Residential -0.1616   0.6142 -0.2631 0.7924 -1.3655  1.0422
PT_Townhouse                 -0.1426   0.6871 -0.2076 0.8356 -1.4894  1.2041
ZIP_92602.0                  -1.4553   0.5473 -2.6593 0.0078 -2.5280 -0.3827
ZIP_92604.0                  -0.3795   0.5051 -0.7515 0.4524 -1.3695  0.6104
ZIP_92606.0                   0.0187   0.4864  0.0385 0.9693 -0.9347  0.9722
ZIP_92612.0                  -1.1405   0.7650 -1.4909 0.1360 -2.6398  0.3588
ZIP_92614.0                  -0.3246   0.5626 -0.5770 0.5639 -1.4272  0.7780
ZIP_92618.0                  -1.1782   0.4331 -2.7201 0.0065 -2.0271 -0.3292
ZIP_92620.0                  -0.4223   0.4210 -1.0032 0.3158 -1.2475  0.4028
LOC_699 - Not Defined        -2.4740   1.1092 -2.2304 0.0257 -4.6480 -0.3000
LOC_AA - Airport Area        -1.6403   0.9863 -1.6630 0.0963 -3.5735  0.2929
LOC_TR - Turtle Rock         -3.0033   1.0980 -2.7354 0.0062 -5.1553 -0.8514
LOC_UP - University Park      0.1713   0.9032  0.1897 0.8496 -1.5989  1.9415
LOC_WB - Woodbridge          -0.8580   0.4661 -1.8409 0.0656 -1.7715  0.0555
LOC_WD - Woodbury             1.4933   0.4622  3.2307 0.0012  0.5874  2.3993
LOC_WI - West Irvine         -0.9054   1.1764 -0.7696 0.4415 -3.2110  1.4003
============================================================================
```

Then I only kept the variables that were statistically significant. The model is better than the first model but it's still really bad.

```
                        Results: Logit
====================================================================
Model:                Logit              Pseudo R-squared: 0.087
Dependent Variable:   SELLER_HOUSE       AIC:              601.5420
Date:                 2020-03-27 10:56   BIC:              626.4071
No. Observations:     466                Log-Likelihood:   -294.77
Df Model:             5                  LL-Null:          -323.01
Df Residuals:         460                LLR p-value:      6.4991e-11
Converged:            1.0000             Scale:            1.0000
No. Iterations:       7.0000
--------------------------------------------------------------------
                       Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
--------------------------------------------------------------------
BATHS                  0.0405   0.0438  0.9237 0.3556 -0.0454  0.1263
ZIP_92602.0           -0.9550   0.4014 -2.3793 0.0173 -1.7416 -0.1683
ZIP_92618.0           -0.3999   0.2797 -1.4300 0.1527 -0.9480  0.1482
LOC_699 - Not Defined -2.2732   1.0659 -2.1326 0.0330 -4.3624 -0.1840
LOC_TR - Turtle Rock  -2.8302   1.0414 -2.7177 0.0066 -4.8712 -0.7891
LOC_WD - Woodbury      1.6857   0.4260  3.9569 0.0001  0.8507  2.5207
====================================================================

Accuracy of logistic regression classifier on test set: 0.59
```



I made another model where I only kept the statistically significant variables from the new model but it kept saying a variable that used to be statistically significant, was no longer significant. Then I was just left with the 3 location variables.

I'm not sure what else I could have done to make the model better other than having more data from neighboring cities.

I had 13 houses appear multiple times and I wasn't sure whether to leave them in or throw them out. When I took them out, it didn't make a difference so I left them in.