Big Data: Homework 2

Will Clark & Matthew DeLio
41201-01

April 16, 2015

1 Data Visualization

We identified three sets of covariates that affect housing prices:

1. Income/education level: home price tends to rise with income (and with education, which correlates highly with income);
2. Home/neighborhood quality: nice homes and nice neighborhoods demand a price premium; and
3. First-time status: first-time home buyers tend to purchase less expensive homes.

1.1 Income & Education

Unsurprisingly, buyers with higher income and more education are able to afford more expensive homes. We see in Figure 1 on the following page that the purchase price of a home increases with each additional level of educaiton achieved. In Figure 2 on the next page, we see that home value rises with income, but much of the variation in income is explained by having a college degree. Red dots, signifying purchasers with a college or graduate degree, are concentrated in the northeast corner, signifying high incomes and expensive homes. Blue dots, signifying purchasers without a college degree, have lower incomes and (expectedly) less expensive homes.

In addition to more expensive homes, more education correlates with neighborhood quality. Figure 3 on page 3 shows the share of households by terminal degree level in good and bad neighborhoods. More educated buyers are more likely to have homes in good neighborhoods; we expect that this is really an income story, as education increases income which affords higher neighborhood quality.

1.2 Neighborhood & Home

The value of a home is also driven by the quality of the home and the surrounding neighborhood. We expect that nicer homes in nicer neighborhoods will cost more. We see in Figure 4 on page 3

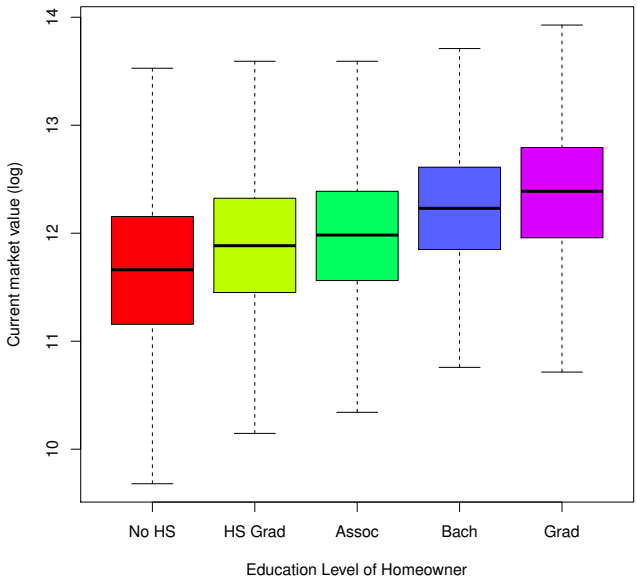

Figure 1: Home Value and Education

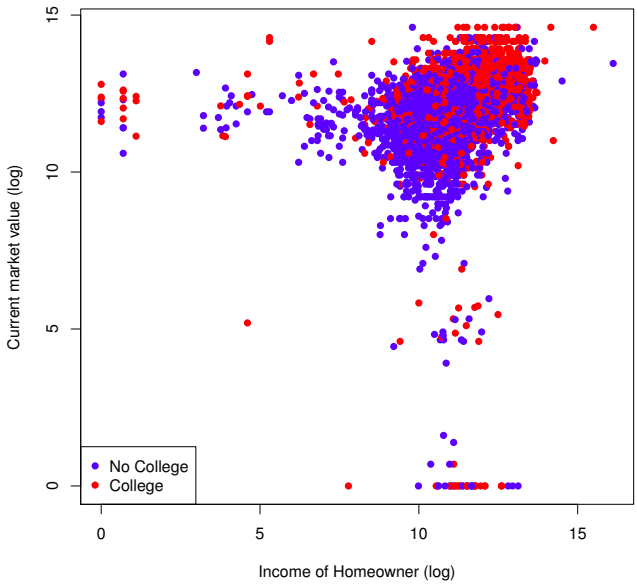

Figure 2: Home Value and Income

and Figure 5 on page 4 that two proxies for neighborhood quality—the presence of junk in the street and proximity to abandoned buldings—both lower home value.

As we would expect, homes that are identified as "good" quality cost more than those that are

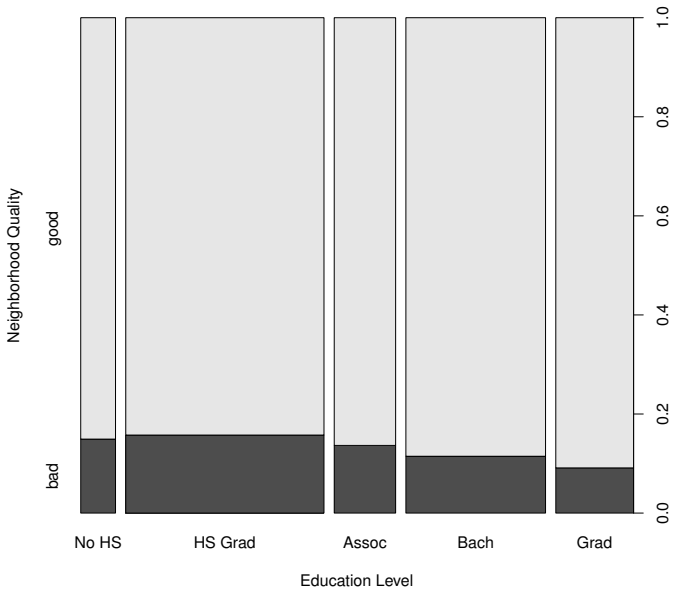

Figure 3: Neighborhood Quality and Income

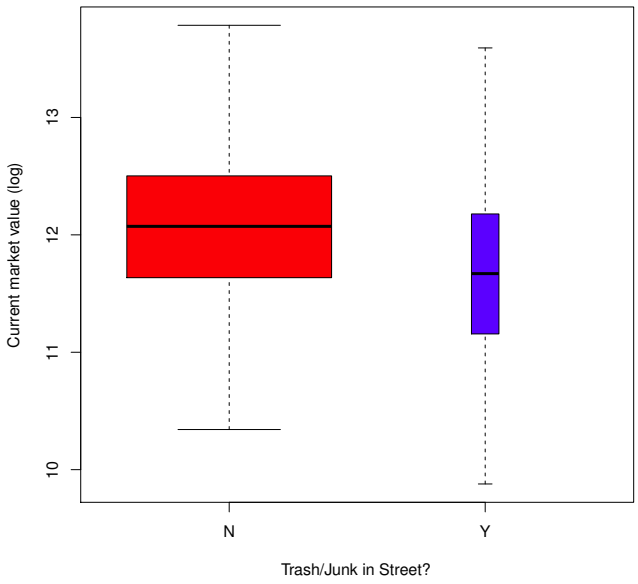

Figure 4: Near Street Trash

not, and good neighborhoods carry a premium over bad neighborhoods (Figure 6 and Figure 7, respectively).

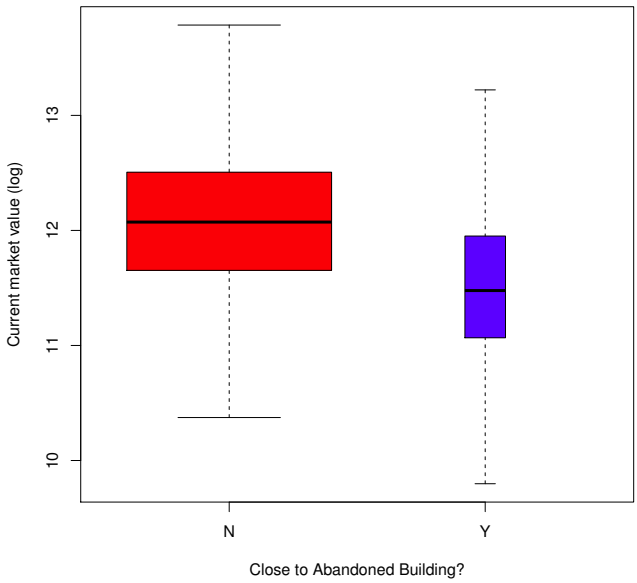

Figure 5: Near Abandoned Buildings

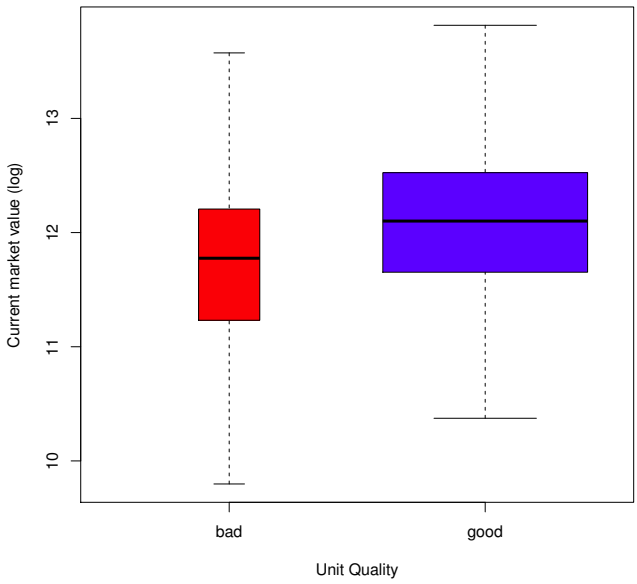

Figure 6: Home Quality

1.3 First-time Buyers & Financing Source

Finally, we examine the difference paid for first-time home buyers versus repeat home buyers. We have two different data sources that tell a similar story. First, buyers who are purchasing their first

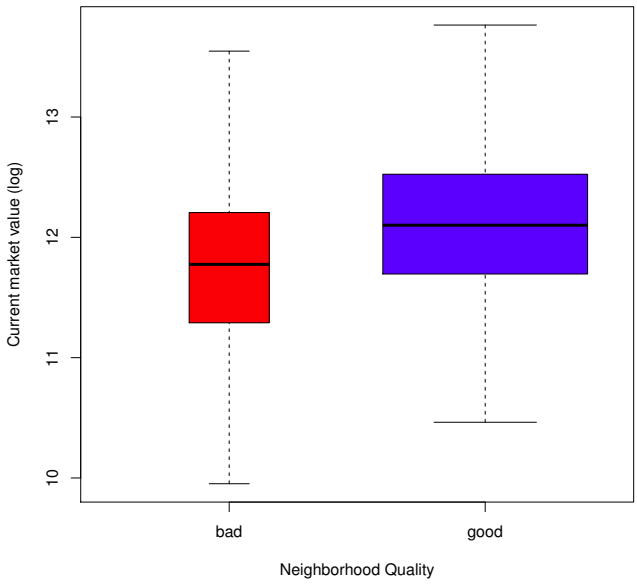

Figure 7: Neighborhood Quality

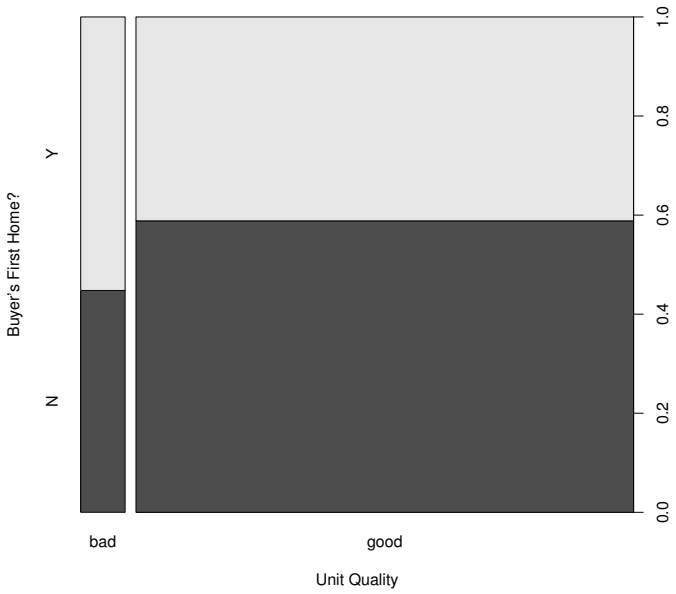

Figure 8:

home tend to pay less than those who are buying a second (or third) home (see Figure 9 on the next page). We expect that this is due to age and income differences; first-time home buyers will tend to be younger and have less accumulated wealth than repeat buyers.

We see that the source of financing for a down payment tells a very similar story. Buyers that use a previous home to pay down their purchase buy more expensive homes (see Figure 10). We expect that this is capturing the same effect as described above: namely, that if a previous home is used as a source of financing, that signals an older buyer with higher income and greater wealth.

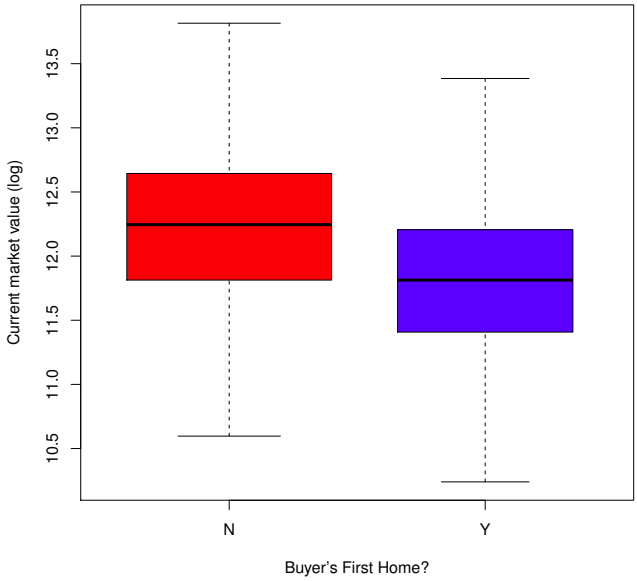

Figure 9: First Time Buyers

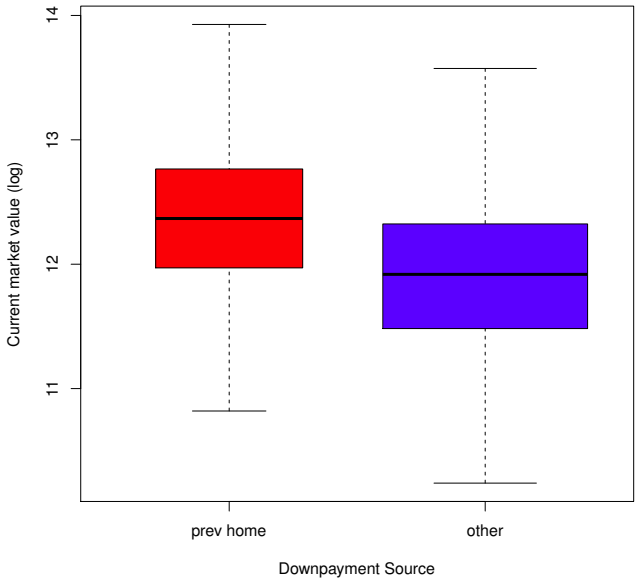

Figure 10: Mortgage Financing

2 Linear Model

We begin by regressing the log of home value onto all possible co-variates (excluding mortgage amount and purchase price, which are known functions of home value). If we apply a false discover rate of 10 percent, we find that there are 34 true discoveries. We regress log of home value on these 34 covariates; the results are displayed in Table 1 on page 9.

The model fit does not change meaningfully between estimates; the R-squared of the first model is 0.3053 and the R-squared of the second model is 0.3050. Removing the extraneous covariates should give us more accurate estimates for the effects of each covariate on home value.

Some notes on the linear regression model:

- Every state included in the model (there were 13) is a significant covariate. California is the most expensive state; Texas is the cheapest.
- More education increases the value of home purchased, but income seems to have very little effect (that is not already included in educaiton).
- Number of bedrooms and bathrooms both increase home value, but the relationship is stronger for number of bathrooms (i.e. it is the best proxy of home size).
- The neighborhood and home quality measures that we saw in the first section appear in our regression results. Junk in the street lowers home value, as do nearby abandoned buildings. Good homes and neighborhoods cost more than bad ones.

3 Logit Model

3.1 Interpret Effects

3.2 Interpret Interaction

4 Out-of-Sample Prediction

4.1 Predicting Downpayments on Homes <100k

4.2 Sample Predictions on Homes >100k

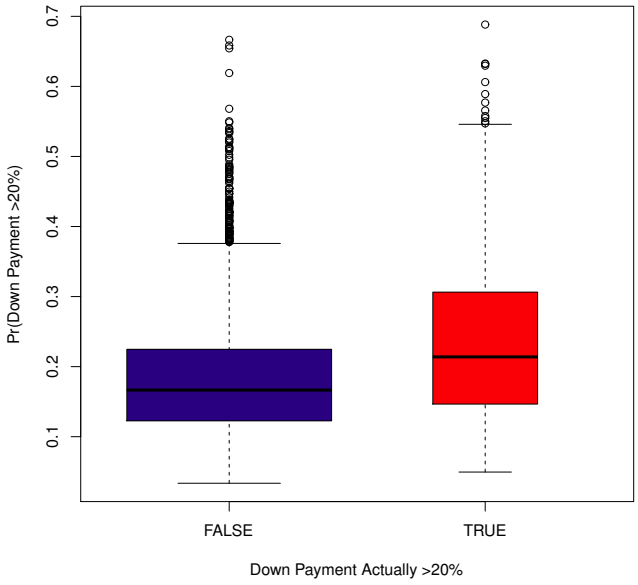

Figure 11:

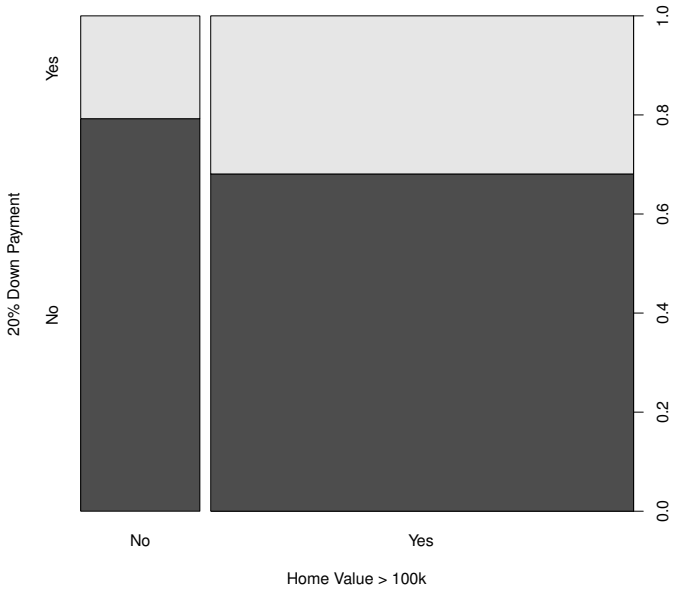

Figure 12:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.5901	0.0606	191.21	0.0000
EAPTBLY	-0.0446	0.0221	-2.02	0.0436
ECOM2Y	-0.0985	0.0470	-2.10	0.0361
EJUNKY	-0.1258	0.0509	-2.47	0.0134
ESFDY	0.2884	0.0292	9.86	0.0000
EABANY	-0.1633	0.0359	-4.55	0.0000
HOWHgood	0.1292	0.0263	4.92	0.0000
HOWNgood	0.1197	0.0219	5.47	0.0000
STRNAY	-0.0391	0.0158	-2.47	0.0136
ZINC2	0.0000	0.0000	11.24	0.0000
HHGRADBach	0.1333	0.0229	5.82	0.0000
HHGRADGrad	0.1979	0.0257	7.69	0.0000
‘HHGRADHS Grad‘	-0.0615	0.0217	-2.84	0.0046
‘HHGRADNo HS‘	-0.1971	0.0318	-6.20	0.0000
NUNITS	-0.0010	0.0005	-1.87	0.0610
INTW	-0.0469	0.0044	-10.66	0.0000
METROurban	0.0837	0.0179	4.67	0.0000
STATECO	-0.2876	0.0290	-9.91	0.0000
STATECT	-0.3444	0.0312	-11.04	0.0000
STATEGA	-0.6555	0.0309	-21.20	0.0000
STATEIL	-0.8624	0.0576	-14.96	0.0000
STATEIN	-0.7779	0.0307	-25.37	0.0000
STATELA	-0.7218	0.0368	-19.63	0.0000
STATEMO	-0.6647	0.0334	-19.89	0.0000
STATEOH	-0.6762	0.0326	-20.73	0.0000
STATEOK	-0.9978	0.0328	-30.43	0.0000
STATEPA	-0.8681	0.0338	-25.67	0.0000
STATETX	-1.0497	0.0343	-30.64	0.0000
STATEWA	-0.1203	0.0309	-3.89	0.0001
BATHS	0.2134	0.0116	18.46	0.0000
BEDRMS	0.0877	0.0094	9.35	0.0000
MATBUYY	-0.0277	0.0136	-2.03	0.0421
‘DWNPAYprev home‘	0.1215	0.0178	6.81	0.0000
FRSTHOY	-0.0829	0.0172	-4.82	0.0000

Table 1: Value of Purchased Homes (in logs)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1925	0.0261	7.37	0.0000
ECOM1Y	-0.0285	0.0096	-2.96	0.0031
ECOM2Y	-0.0439	0.0248	-1.77	0.0763
ESFDY	-0.0566	0.0151	-3.74	0.0002
HOWNgood	0.0166	0.0105	1.59	0.1122
STRNAY	-0.0164	0.0083	-1.97	0.0489
PER	-0.0208	0.0025	-8.34	0.0000
HHGRADBach	0.0354	0.0082	4.31	0.0000
HHGRADGrad	0.0563	0.0103	5.45	0.0000
INTW	-0.0094	0.0023	-4.11	0.0000
STATECT	0.1423	0.0128	11.08	0.0000
STATEGA	-0.0476	0.0130	-3.66	0.0003
STATEIL	0.1045	0.0286	3.66	0.0003
STATEIN	0.0332	0.0127	2.61	0.0091
STATELA	0.0953	0.0166	5.73	0.0000
STATEMO	0.0926	0.0146	6.36	0.0000
STATEOH	0.1342	0.0139	9.63	0.0000
STATEPA	0.1028	0.0148	6.96	0.0000
STATETX	0.0406	0.0148	2.75	0.0059
BATHS	0.0433	0.0058	7.49	0.0000
MATBUYY	0.0493	0.0071	6.94	0.0000
‘DWNPAYprev home‘	0.1605	0.0094	17.15	0.0000
VALUE	0.0000	0.0000	12.89	0.0000
FRSTHOY	-0.0579	0.0090	-6.43	0.0000

Table 2: Probability of Down Payment > 20% (No Interaction Terms)

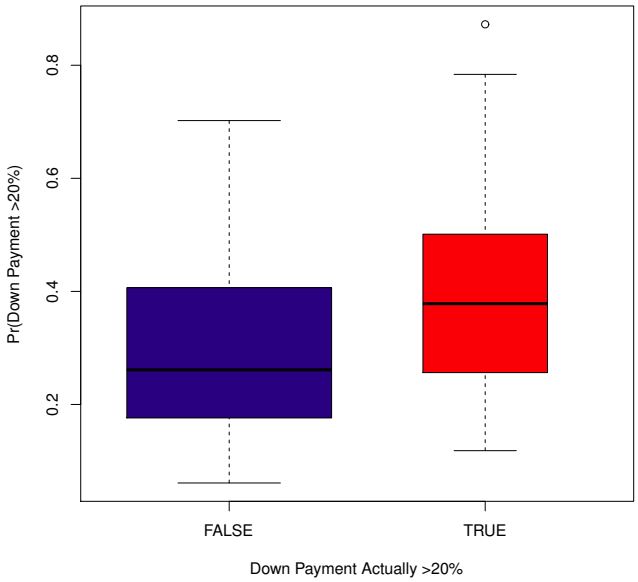

Figure 13:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.1766	0.0255	6.92	0.0000
ECOM1Y	-0.0284	0.0096	-2.94	0.0032
ECOM2Y	-0.0449	0.0248	-1.81	0.0700
ESFDY	-0.0573	0.0151	-3.79	0.0002
HOWNgood	0.0174	0.0105	1.67	0.0957
STRNAY	-0.0166	0.0083	-1.99	0.0462
PER	-0.0208	0.0025	-8.35	0.0000
HHGRADBach	0.0358	0.0082	4.36	0.0000
HHGRADGrad	0.0566	0.0103	5.48	0.0000
INTW	-0.0097	0.0023	-4.22	0.0000
STATECT	0.1411	0.0128	11.00	0.0000
STATEGA	-0.0473	0.0130	-3.64	0.0003
STATEIL	0.1024	0.0286	3.59	0.0003
STATEIN	0.0324	0.0127	2.55	0.0109
STATELA	0.0955	0.0166	5.74	0.0000
STATEMO	0.0910	0.0145	6.25	0.0000
STATEOH	0.1324	0.0139	9.50	0.0000
STATEPA	0.0999	0.0148	6.77	0.0000
STATETX	0.0397	0.0147	2.69	0.0071
BATHS	0.0556	0.0058	9.52	0.0000
MATBUYY	0.0493	0.0071	6.95	0.0000
‘DWNPAYprev home‘	0.1553	0.0092	16.84	0.0000
VALUE	0.0000	0.0000	12.35	0.0000
‘BATHS:FRSTHOY‘	-0.0367	0.0047	-7.75	0.0000

Table 3: Probability of Down Payment > 20% (With Interaction Term)