Part-I Simple Model

Chenxi Wu, George Lindner, Qianyin Lu, Yi Mi

11/30/2019

Introduction

need intro

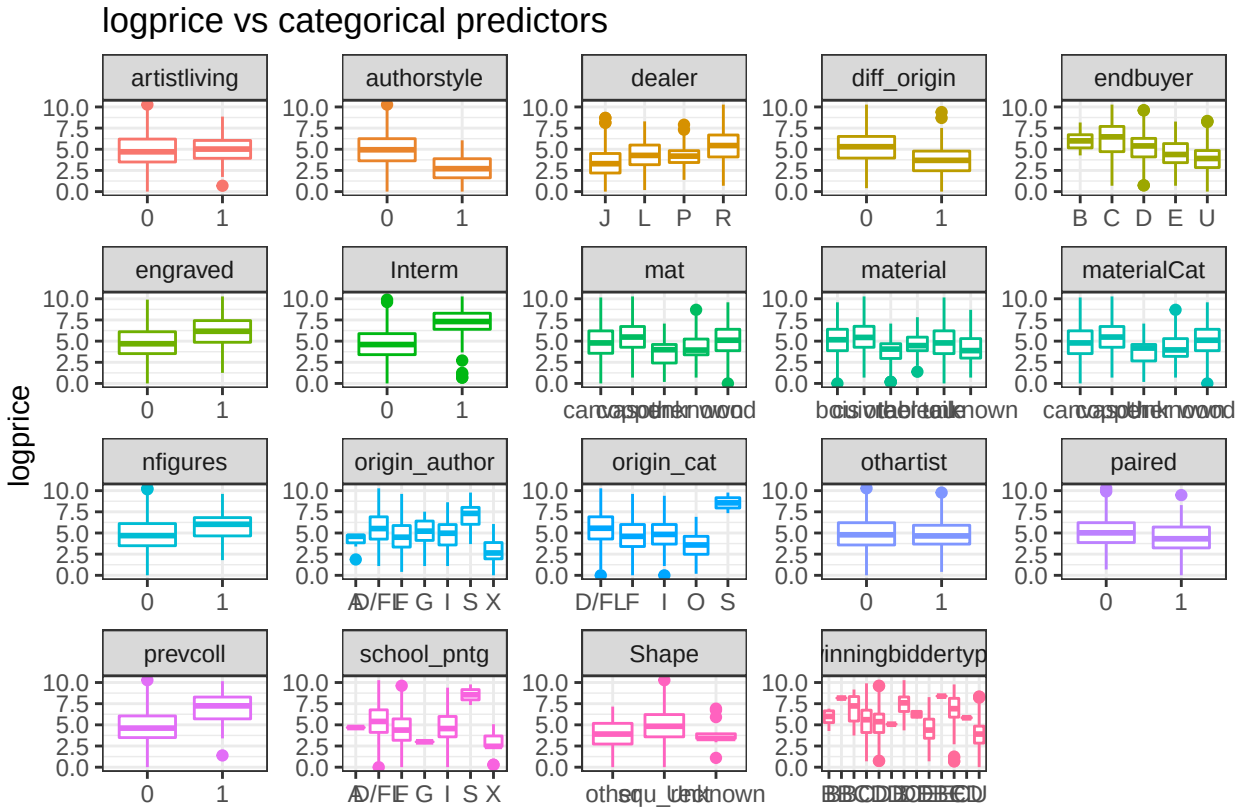
Data Manipulation and EDA

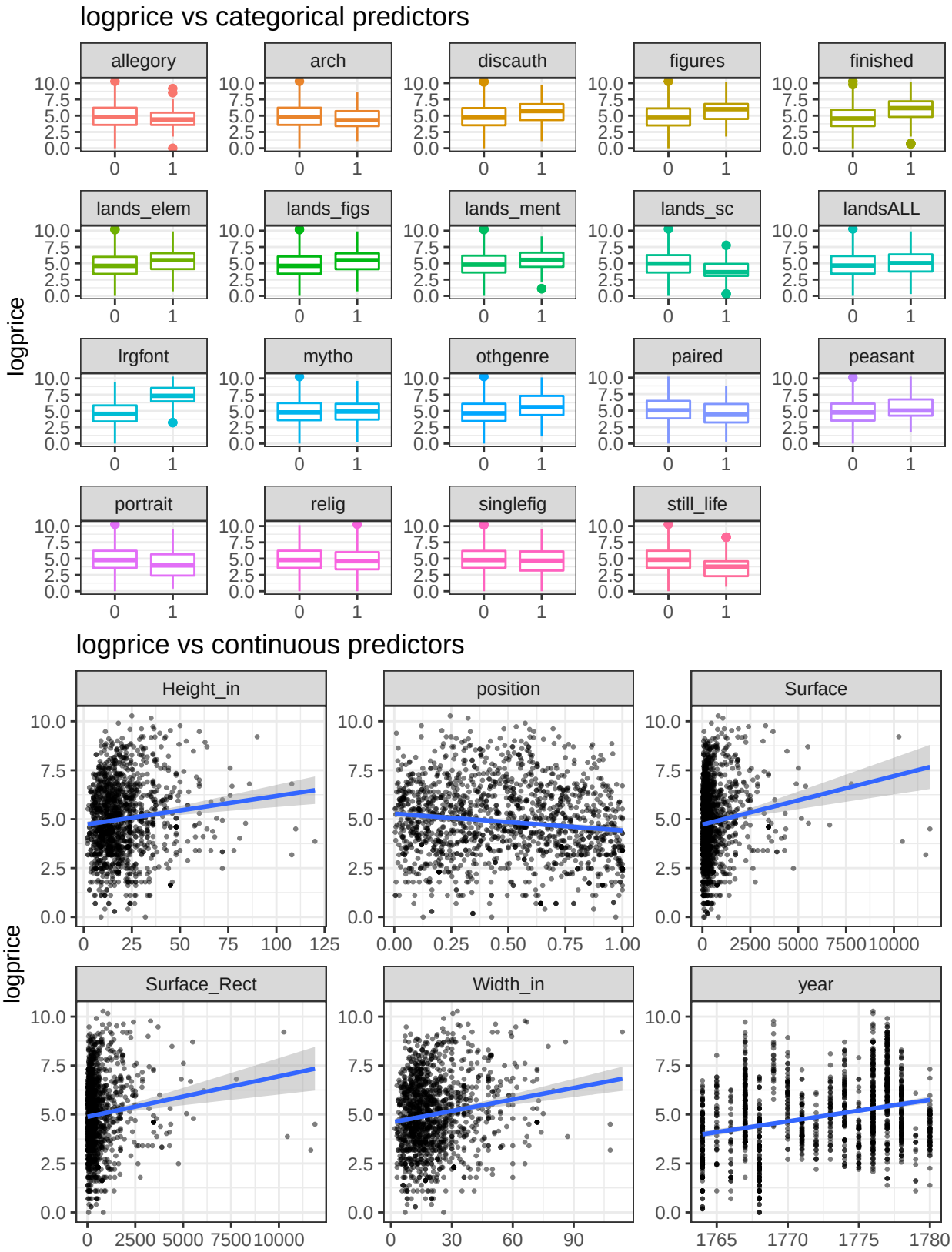
Looking at the summary of the dataset (See Appendix), we made following manipulation to the variables:

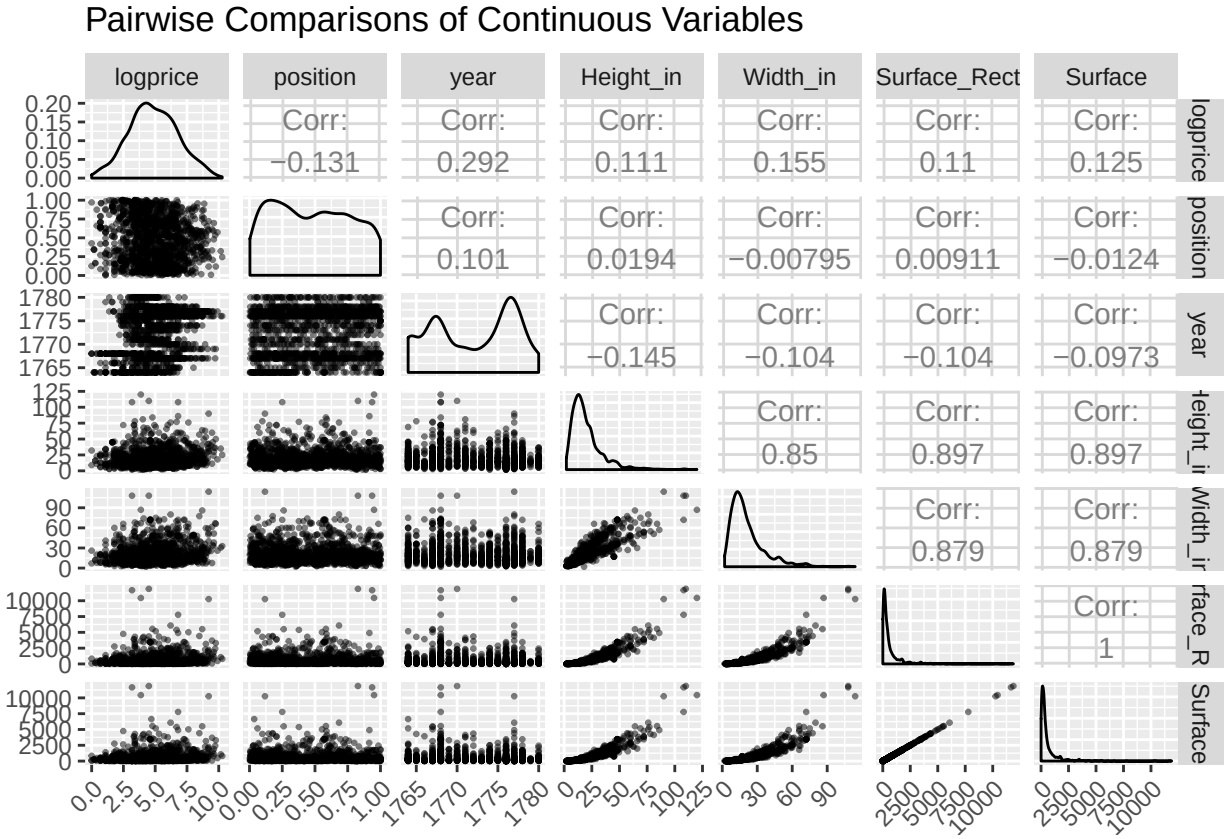
- *history*, *original*, *type_intermed*, and *pastorale* have serious **class imbalance** problem, where certain value is seldomly observed. So we will discard these variables in modeling.

- The variable *authorstyle* contains too many NA's which is not sufficient to do any analysis, which will also be discarded. (need to add description to each variable deleted above, but it actually makes more sense to delete just a few of them before eda, then delete the rest looking at the eda)

To start with, we first check on the emprirical distribution of the response variable. There are 2 variables, *logprice* and *price*. From the histogram (See Appendix), we can see that *logprice*, which is the logarithm of *price*, is more normally-distributed. Consider the normality assumption of linear regression, we will use *logprice* as the response variable.




logprice vs categorical predictors
logprice
allegory
arch
discauth
figures
finished
lands_elem
lands_figs
lands_ment
lands_sc
landsALL
lrgfont
mytho
othgenre
paired
peasant
portrait
relig
singlefig
still_life
logprice vs continuous predictors
logprice
Height_in
position
Surface
Surface_Rect
Width_in
year



Then we plot *logprice* against all continuous and categorical variables respectively. Here are some interesting findings.

For categorical variables:

- After cleaning the data, *mat*, *material* and *materialCat* record duplicate features. The Chi-squared test shows a strong correlation between these variables (See Appendix). We will only keep *matCat* in future analysis since it contains all the information in the other two.
- *origin_author* and *origin_cat* record similar features which is suspecious of highly dependent on each other. We implement Chi-squared test (See Appendix) to test the hypothesis, which yield a p-value of $2.2e-16$. Thus we will only keep *origin_cat* since it contains less levels.
- At glance, strong predictors includde *dealer*, *diff_origin*, *discauth*, *endbuyer*, *materialCat*, *interm*, *finished*, *engraved*, *figures*, *Irgfont*, *origin_cat*, *paired*, *portrait*, *prevcoll*, *lands_sc*, *lands_elem*. (Might need adjustments)

For continuous variables:

- *other* and *position* have most of their values gather around 0 and we cannot observe any obvious pattern.
- *Diam-in* and *Surface_Rnd* contain too many NA's and therefore should not be inlcuded in the model.
- *Surface*, *Surface_Rnd* and *Surface_Rect* contain duplicate information, and *Surface* contains all the information in *Surface_Rect*. So we will only keep *Surface*. *Height_in* and *Width_in* are discarded for similar reasoning.
- *nfigures* might need transformations, so I will exponentiate it in the model.
- Additionally, *figures*, *nfigures* and *siglefig* depict duplicate information. Looking at the empirical distribution, we decide to use

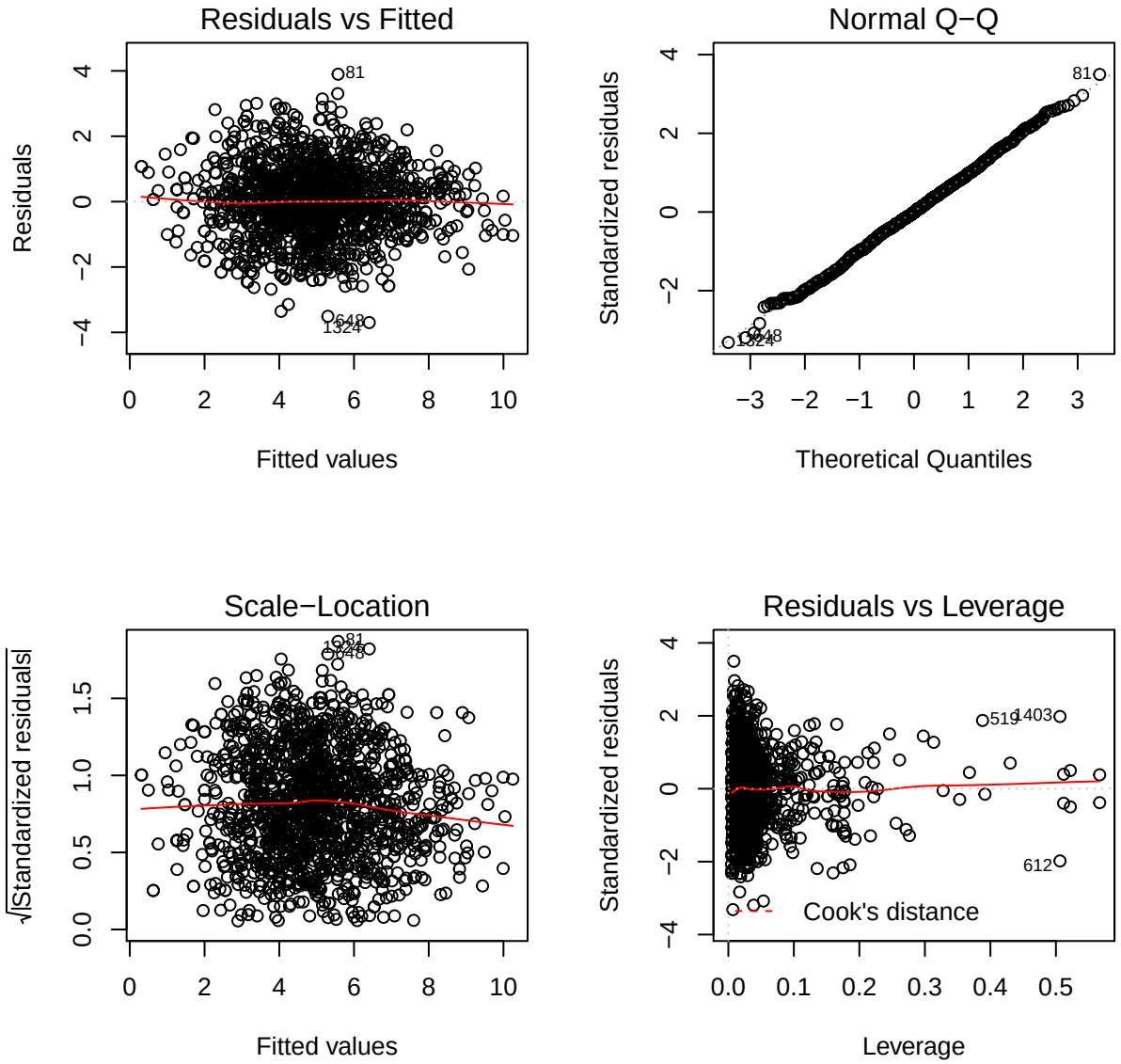
Model Selection

We first remove some variables because they have very similar levels and thus may not have significant influence on the response.

We use BIC to select the important variables. And we use the BIC result to build a new model and construct all interaction terms and then we use AIC to determine the final model.(may not appear in the write-up for bic part)

We deleted some interactions for the follow reasons. We deleted the interactions `lands_sc:discauth`, `paired:lands_sc` and `artistliving:lands_sc`. First of all, for the interactions related to `lands_sc`, we think the `lands_sc`–if described as a plain landscape, is a variable describing a single dimension of painting content and we don't think it will interact with the variables describing the dealers' behaviour, the pairing of a painting, or the living info of the artists. Though the Adjusted R-squared decreased a little, from 0.66 to 0.659, we think it worth to make the model more reasonable.

Model Assessment



looking at the diagnostic plots, our model 1 seems to satisfy the assumptions of linear regression resonablly well. From the Residual vs Fitted plot we can see equally spread residuals around a horizontal line without

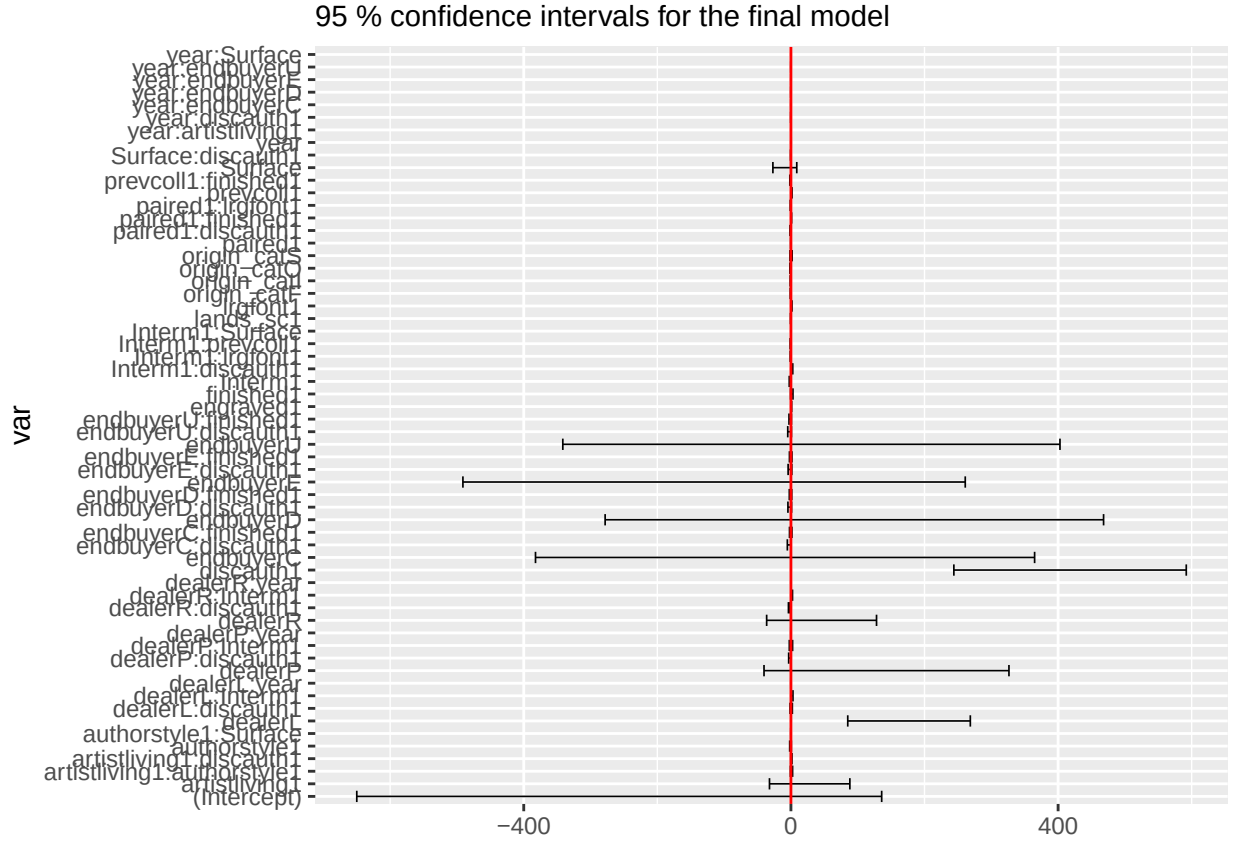
any distinct patterns; The Normal Q-Q plot shows the residuals are almost normally-distributed. The Scale-Location plot shows that homoscedasticity is met. The Residual vs Leverage plot shows that most of the points are not influential. There are only very few points that falls outside of Cook's distance line. The plot identified the influential observation as #218 and #1129. We removed these two points and refitted the model. But removing the two points does not improve the R^2 nor does it help with the residual diagnostics plots (See Appendix the residual plot after refitting) So we will set Model 1 to our final model in part-I.

	rstudent	p	bonf.p	signif	cutoff
81	3.510978	0.0004602	0.6894504	FALSE	0.05

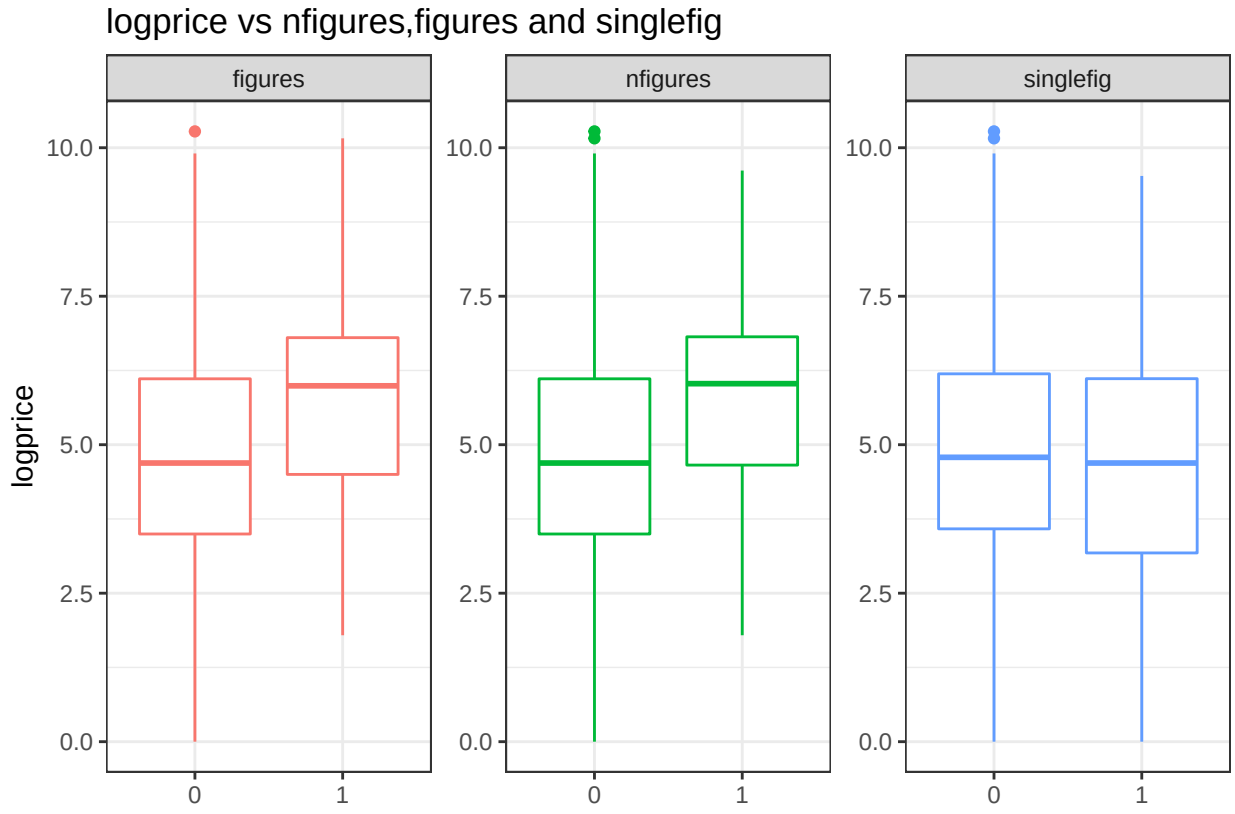
We see that one point has a large rstudent residual, but given the number of points in the model this is not a very wild residual value, as indicated by the Bonferonni-Adjusted P value of 0.69, indicating that this is not a significantly distant outlier.

We also checked the largest Cook's distance, which is much less than one, indicating that there are no points with very large leverage in this model. From the plot wee that the points with large cook's distance do not appear to be significantly distant from the other points in the dataset, and their influence is not undue.

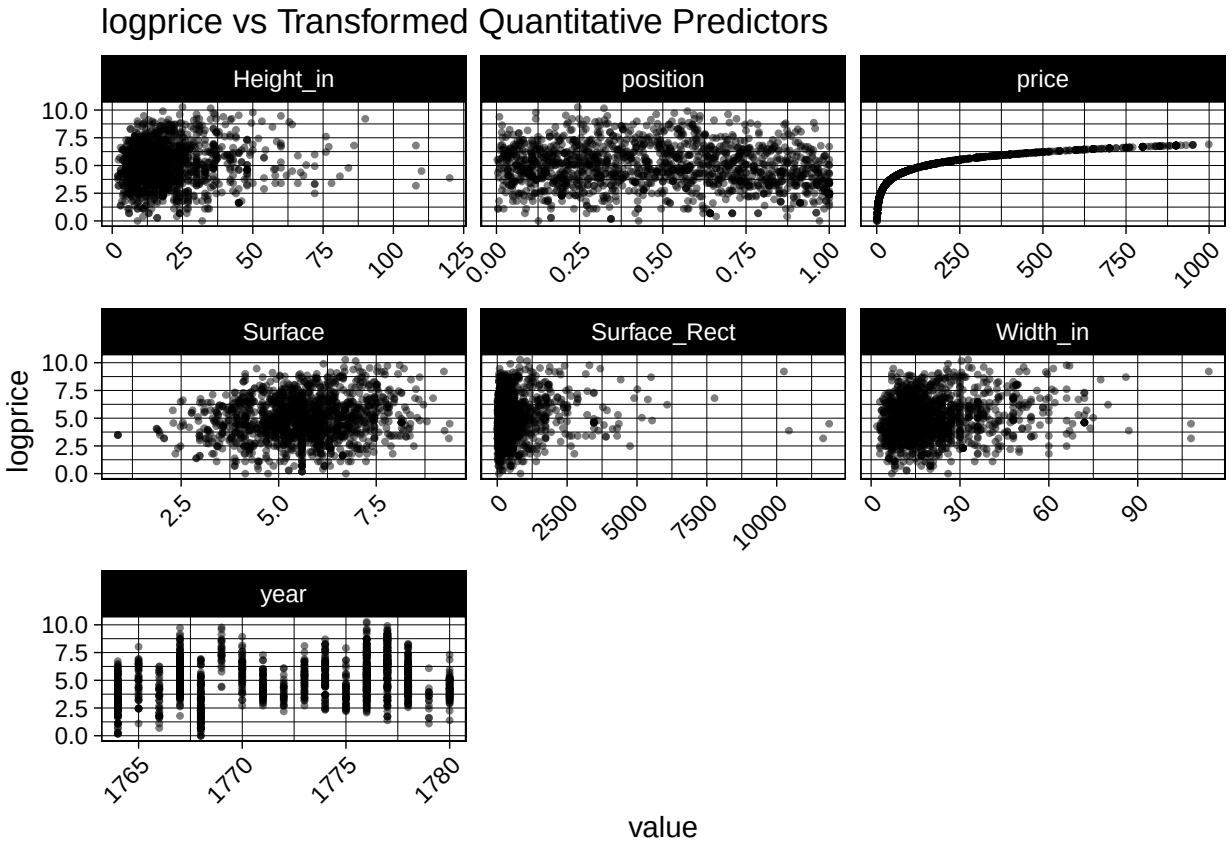
Sumnmary and conclusions



Appendix



	D/FL	F	I	O	S
A	6	0	0	1	0
D/FL	520	4	13	53	0
F	9	451	4	114	0
G	25	0	0	1	0
I	4	7	135	13	0
S	0	0	9	0	2
X	30	21	9	69	0



Call: lm(formula = logprice ~ dealer + year + origin_cat + artistliving + authorstyle + endbuyer + Interm + Surface + engraved + prevcoll + paired + finished + lrgfont + lands_sc + discauth + dealer:year + +dealer:Interm + dealer:discauth + year:artistliving + year:endbuyer + year:Surface + year:discauth + artistliving:authorstyle + artistliving:discauth + authorstyle:Surface + endbuyer:finished + endbuyer:discauth + Interm:Surface + Interm:prevcoll + Interm:lrgfont + Interm:discauth + Surface:discauth + prevcoll:finished + paired:finished + paired:lrgfont + paired:discauth, data = paint_train_trans)

Residuals: Min 1Q Median 3Q Max -3.6999 -0.6999 -0.0084 0.7264 3.8991

Coefficients: Estimate Std. Error t value (Intercept) -257.051573 200.310123 -1.283 dealerL 176.843518 46.773723 3.781 dealerP 143.085605 93.470425 1.531 dealerR 45.901034 41.926343 1.095 year 0.146252 0.113017 1.294 origin_catF -0.652228 0.080201 -8.132 origin_catI -0.697565 0.103330 -6.751 origin_catO -1.132467 0.120543 -9.395 origin_catS 0.195315 0.825809 0.237 artistliving1 28.075785 30.671354 0.915 authorstyle1 -0.645780 0.607575 -1.063 endbuyerC -8.782033 190.464908 -0.046 endbuyerD 94.957746 190.213178 0.499 endbuyerE -115.030753 191.689435 -0.600 endbuyerU 30.642794 189.725766 0.162 Interm1 -1.276699 0.616342 -2.071 Surface -9.119096 9.105733 -1.001 engraved1 0.727258 0.139197 5.225 prevcoll1 1.200293 0.167297 7.175 paired1 -0.259688 0.071736 -3.620 finished1 1.255594 0.901129 1.393 lrgfont1 1.137438 0.149004 7.634 lands_sc1 -0.367156 0.112762 -3.256 discauth1 417.902026 88.689923 4.712 dealerL:year -0.098900 0.026372 -3.750 dealerP:year -0.080285 0.052597 -1.526 dealerR:year -0.024808 0.023656 -1.049 dealerL:Interm1 1.469755 0.810716 1.813 dealerP:Interm1 0.062646 1.242191 0.050 dealerR:Interm1 1.435328 0.478940 2.997 dealerL:discauth1 0.444092 0.865395 0.513 dealerP:discauth1 -1.590138 0.890114 -1.786 dealerR:discauth1 -2.812777 0.465821 -6.038 year:artistliving1 -0.015625 0.017296 -0.903 year:endbuyerC 0.004930 0.107468 0.046 year:endbuyerD -0.053660 0.107326 -0.500 year:endbuyerE 0.064626 0.108164 0.597 year:endbuyerU -0.017720 0.107051 -0.166 year:Surface 0.005332 0.005138 1.038 year:discauth1 -0.232575 0.049850 -4.666 artistliving1:authorstyle1 0.868333 0.822690 1.055 artistliving1:discauth1 0.671771 0.463023 1.451 authorstyle1:Surface -0.047213 0.103586 -0.456 endbuyerC:finished1 -0.323477 0.914014 -0.354 endbuyerD:finished1 -0.595502 0.910978 -0.654 endbuyerE:finished1 -0.339537 0.940883 -0.361 endbuyerU:finished1 -1.069766 0.917325 -1.166 endbuyerC:discauth1 -2.510757 1.441472 -1.742 endbuyerD:discauth1 -1.766001

1.375047 -1.284 endbuyerE:discauth1 -1.428412 1.398124 -1.022 endbuyerU:discauth1 -2.007086 1.393734 -1.440 Interm1:Surface 0.139416 0.087021 1.602 Interm1:prevcoll1 -0.505839 0.337125 -1.500 Interm1:lrgfont1 -0.560115 0.263296 -2.127 Interm1:discauth1 1.637538 0.549741 2.979 Surface:discauth1 -0.320604 0.121026 -2.649 prevcoll1:finished1 -1.015843 0.311686 -3.259 paired1:finished1 0.595253 0.195932 3.038 paired1:lrgfont1 -0.564608 0.235559 -2.397 paired1:discauth1 -0.721266 0.342301 -2.107 Pr(>|t|)
(Intercept) 0.199605
dealerL 0.000163 ***dealerP 0.126036***
dealerR 0.273787
year 0.195850
origin_catF 0.000000000000000899 origin_catI 0.000000000021246902 ***origin_catO < 0.0000000000000002*** origin_catS 0.813068
artistliving1 0.360148
authorstyle1 0.288014
endbuyerC 0.963230
endbuyerD 0.617702
endbuyerE 0.548541
endbuyerU 0.871714
Interm1 0.038498 *
Surface 0.316769
engraved1 0.000000200077269666 ***prevcoll1 0.0000000000011562 75*** paired1 0.000305 ***finished1 0.163727***
lrgfont1 0.000000000000041342 lands_sc1 0.001156 ** discauth1 0.000002690714331104 ***dealerL:year 0.000184*** dealerP:year 0.127125
dealerR:year 0.294488
dealerL:Interm1 0.070054 .
dealerP:Interm1 0.959785
dealerR:Interm1 0.002774 ** dealerL:discauth1 0.607914
dealerP:discauth1 0.074238 .
dealerR:discauth1 0.000000001978434753 ***year:artistliving1 0.366480***
year:endbuyerC 0.963417
year:endbuyerD 0.617172
year:endbuyerE 0.550276
year:endbuyerU 0.868548
year:Surface 0.299596
year:discauth1 0.000003364146190320 artistliving1:authorstyle1 0.291383
artistliving1:discauth1 0.147043
authorstyle1:Surface 0.648612
endbuyerC:finished1 0.723460
endbuyerD:finished1 0.513412
endbuyerE:finished1 0.718249
endbuyerU:finished1 0.243735
endbuyerC:discauth1 0.081757 .
endbuyerD:discauth1 0.199236
endbuyerE:discauth1 0.307112
endbuyerU:discauth1 0.150062
Interm1:Surface 0.109351
Interm1:prevcoll1 0.133718
Interm1:lrgfont1 0.033563 *
Interm1:discauth1 0.002943 ** Surface:discauth1 0.008160 ** prevcoll1:finished1 0.001143 ** paired1:finished1 0.002424 ** paired1:lrgfont1 0.016662 *
paired1:discauth1 0.035280 *
— Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.12 on 1440 degrees of freedom Multiple R-squared: 0.6727, Adjusted R-squared:

0.6593 F-statistic: 50.17 on 59 and 1440 DF, p-value: < 0.00000000000000022

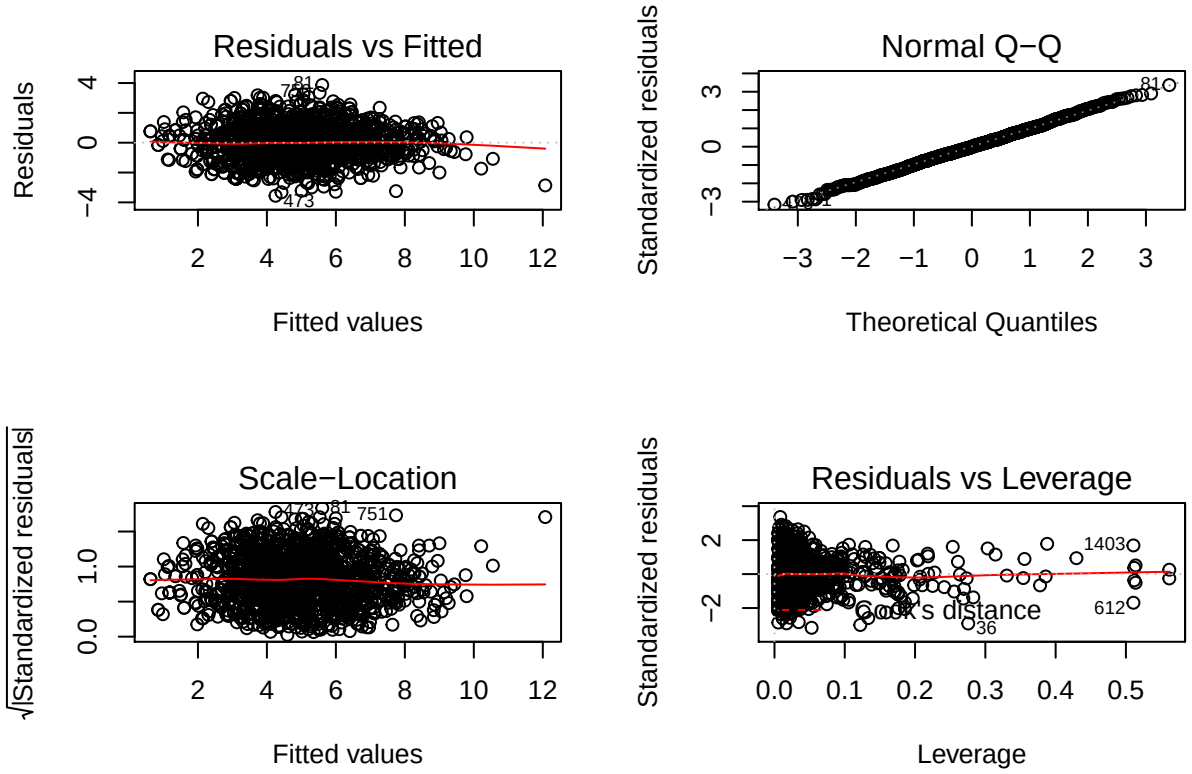


Table 1: Coefficient Summary for Final Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-257.052	200.310	-1.283	0.200	-649.982	135.879
dealerL	176.844	46.774	3.781	0.000	85.092	268.595
dealerP	143.086	93.470	1.531	0.126	-40.267	326.438
dealerR	45.901	41.926	1.095	0.274	-36.342	128.144
year	0.146	0.113	1.294	0.196	-0.075	0.368
origin_catF	-0.652	0.080	-8.132	0.000	-0.810	-0.495
origin_catI	-0.698	0.103	-6.751	0.000	-0.900	-0.495
origin_catO	-1.132	0.121	-9.395	0.000	-1.369	-0.896
origin_catS	0.195	0.826	0.237	0.813	-1.425	1.815
artistliving1	28.076	30.671	0.915	0.360	-32.090	88.241
authorstyle1	-0.646	0.608	-1.063	0.288	-1.838	0.546
endbuyerC	-8.782	190.465	-0.046	0.963	-382.400	364.836
endbuyerD	94.958	190.213	0.499	0.618	-278.167	468.082
endbuyerE	-115.031	191.689	-0.600	0.549	-491.051	260.990
endbuyerU	30.643	189.726	0.162	0.872	-341.526	402.811
Interm1	-1.277	0.616	-2.071	0.038	-2.486	-0.068
Surface	-9.119	9.106	-1.001	0.317	-26.981	8.743
engraved1	0.727	0.139	5.225	0.000	0.454	1.000
prevcoll1	1.200	0.167	7.175	0.000	0.872	1.528
paired1	-0.260	0.072	-3.620	0.000	-0.400	-0.119
finished1	1.256	0.901	1.393	0.164	-0.512	3.023
lrgfont1	1.137	0.149	7.634	0.000	0.845	1.430
lands_sc1	-0.367	0.113	-3.256	0.001	-0.588	-0.146
discauth1	417.902	88.690	4.712	0.000	243.927	591.877
dealerL:year	-0.099	0.026	-3.750	0.000	-0.151	-0.047
dealerP:year	-0.080	0.053	-1.526	0.127	-0.183	0.023
dealerR:year	-0.025	0.024	-1.049	0.294	-0.071	0.022
dealerL:Interm1	1.470	0.811	1.813	0.070	-0.121	3.060
dealerP:Interm1	0.063	1.242	0.050	0.960	-2.374	2.499
dealerR:Interm1	1.435	0.479	2.997	0.003	0.496	2.375
dealerL:discauth1	0.444	0.865	0.513	0.608	-1.253	2.142
dealerP:discauth1	-1.590	0.890	-1.786	0.074	-3.336	0.156
dealerR:discauth1	-2.813	0.466	-6.038	0.000	-3.727	-1.899
year:artistliving1	-0.016	0.017	-0.903	0.366	-0.050	0.018
year:endbuyerC	0.005	0.107	0.046	0.963	-0.206	0.216
year:endbuyerD	-0.054	0.107	-0.500	0.617	-0.264	0.157
year:endbuyerE	0.065	0.108	0.597	0.550	-0.148	0.277
year:endbuyerU	-0.018	0.107	-0.166	0.869	-0.228	0.192
year:Surface	0.005	0.005	1.038	0.300	-0.005	0.015
year:discauth1	-0.233	0.050	-4.666	0.000	-0.330	-0.135
artistliving1:authorstyle1	0.868	0.823	1.055	0.291	-0.745	2.482
artistliving1:discauth1	0.672	0.463	1.451	0.147	-0.237	1.580
authorstyle1:Surface	-0.047	0.104	-0.456	0.649	-0.250	0.156
endbuyerC:finished1	-0.323	0.914	-0.354	0.723	-2.116	1.469
endbuyerD:finished1	-0.596	0.911	-0.654	0.513	-2.382	1.191
endbuyerE:finished1	-0.340	0.941	-0.361	0.718	-2.185	1.506
endbuyerU:finished1	-1.070	0.917	-1.166	0.244	-2.869	0.730
endbuyerC:discauth1	-2.511	1.441	-1.742	0.082	-5.338	0.317
endbuyerD:discauth1	-1.766	1.375	-1.284	0.199	-4.463	0.931
endbuyerE:discauth1	-1.428	1.398	-1.022	0.307	-4.171	1.314
endbuyerU:discauth1	-2.007	1.394	-1.440	0.150	-4.741	0.727
Interm1:Surface	0.139	0.087	1.602	0.109	-0.031	0.310
Interm1:prevcoll1	-0.506	0.337	-1.500	0.134	-1.167	0.155
Interm1:lrgfont1	-0.560	0.263	-2.127	0.034	-1.077	-0.044
Interm1:discauth1	1.638	0.550	2.979	0.003	0.559	2.716
Surface:discauth1	-0.321	0.121	-2.649	0.008	-0.558	-0.083