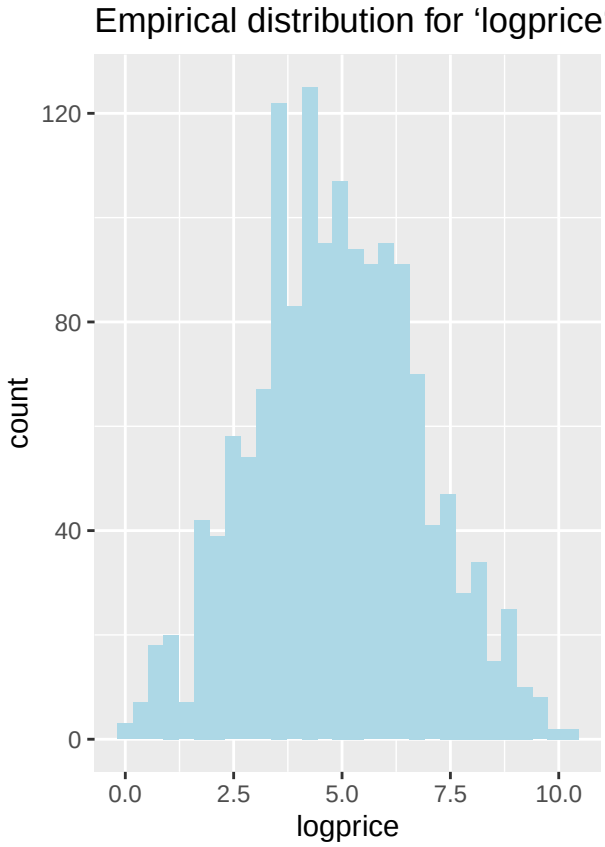
Part-I Simple Model

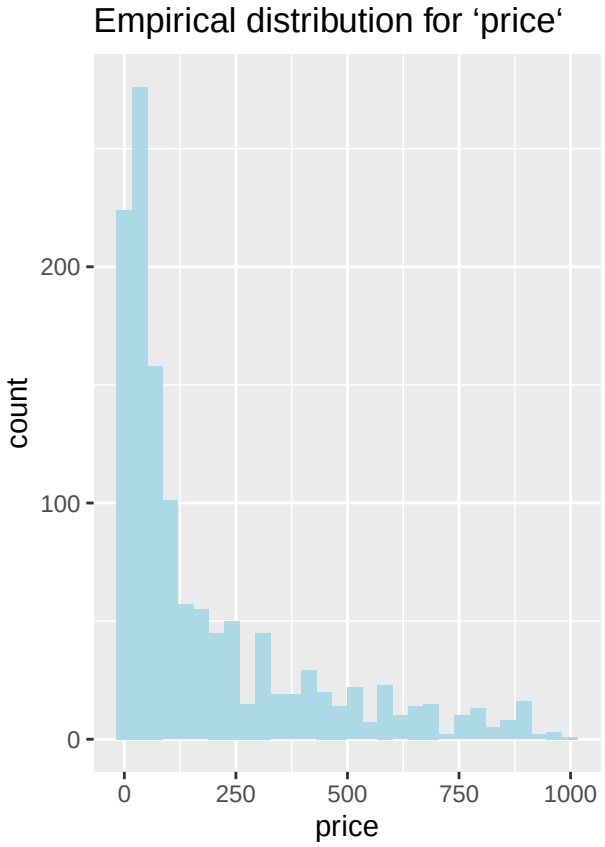
Chenxi Wu

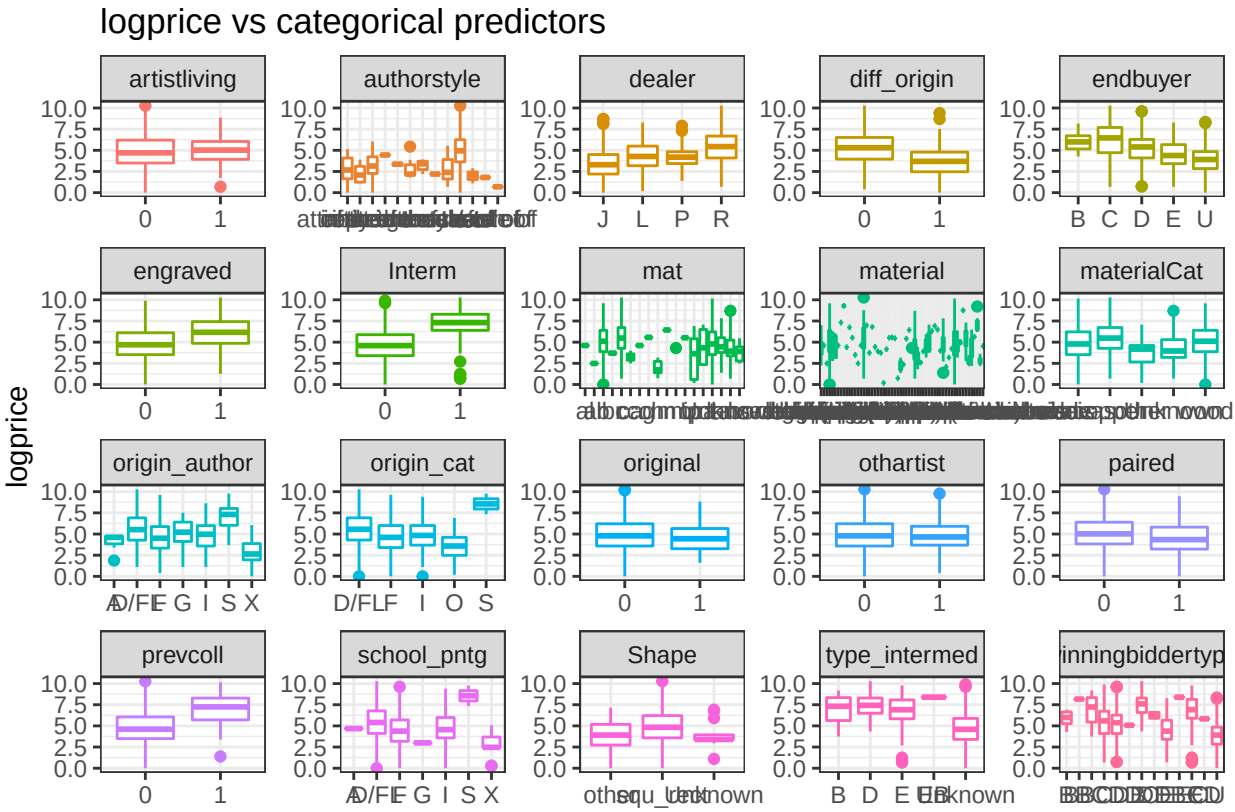
11/30/2019

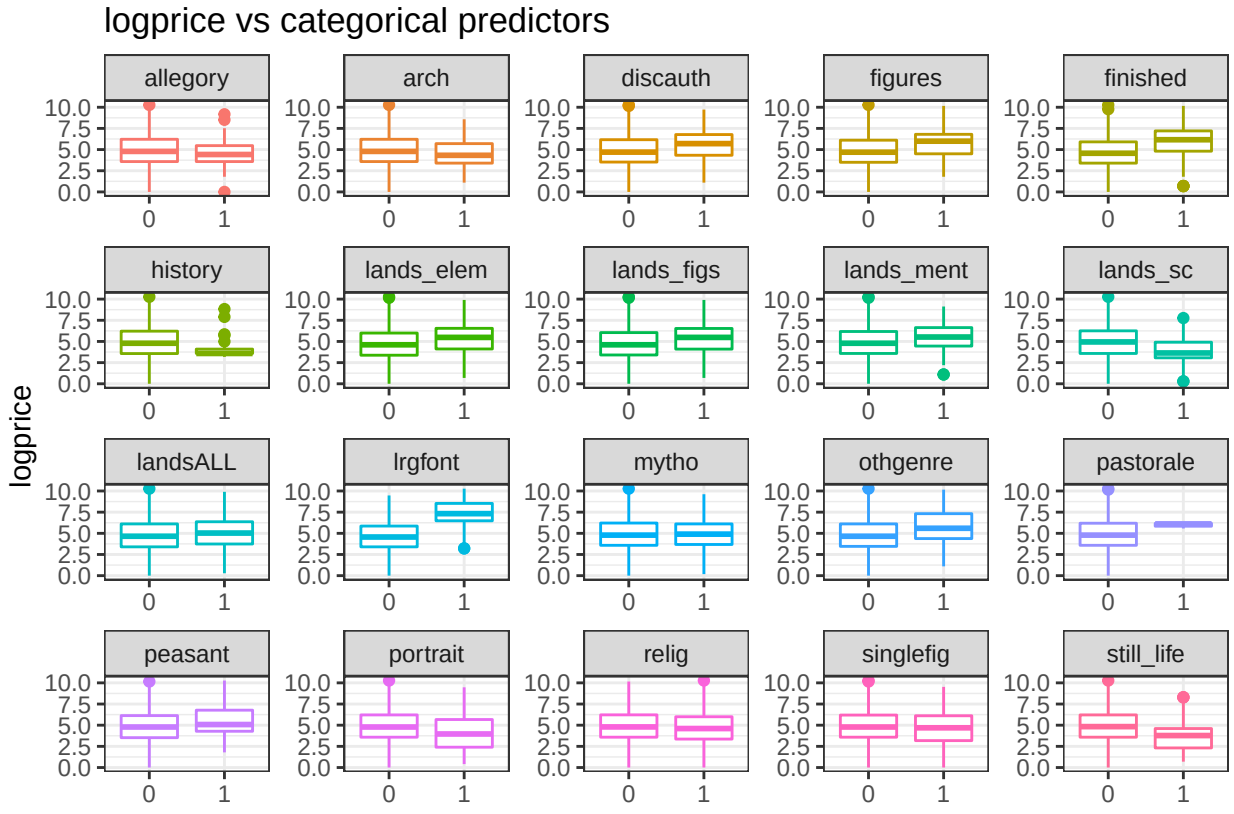
Introduction

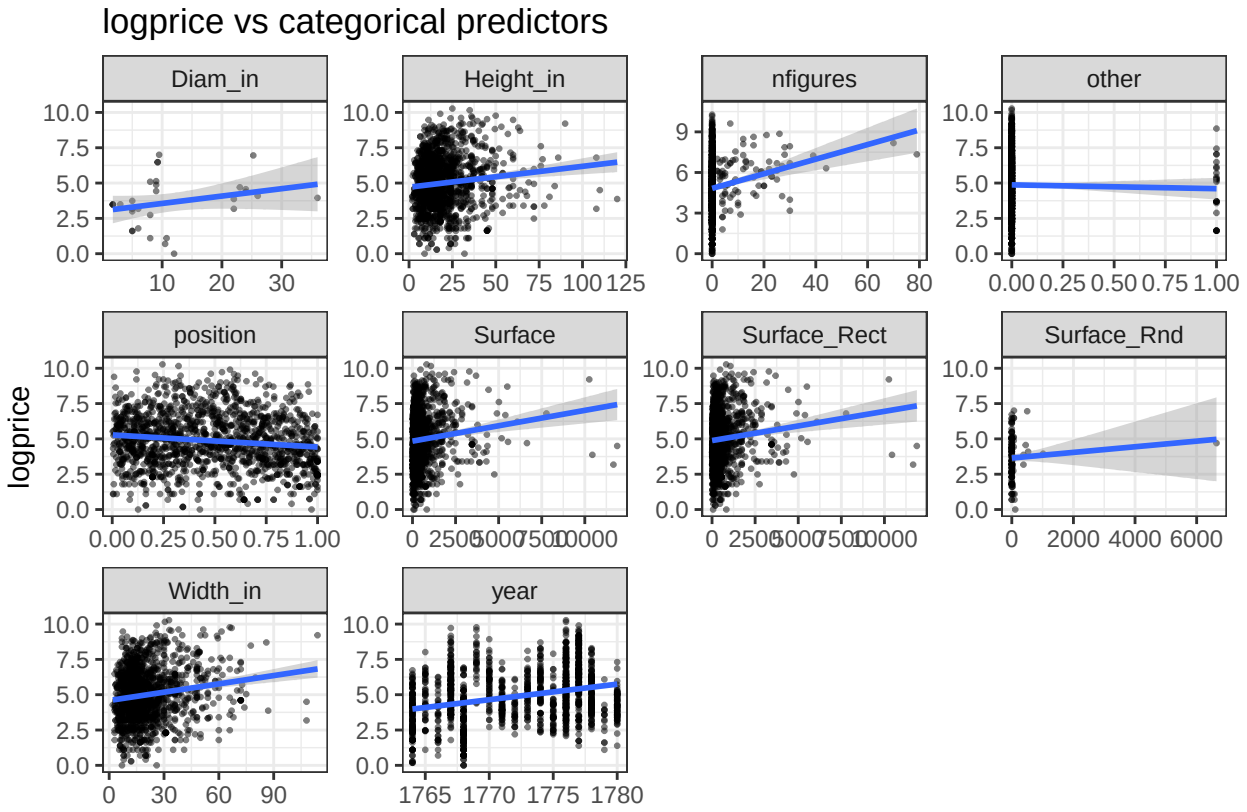
EDA






logprice vs categorical predictors
logprice
artistliving
authorstyle
dealer
diff_origin
endbuyer
engraved
Interm
mat
material
materialCat
origin_author
origin_cat
original
othartist
paired
prevcoll
school_pntg
Shape
type_intermed
inningbiddertyp


logprice vs categorical predictors
logprice
allegory
arch
discauth
figures
finished
history
lands_elem
lands_figs
lands_ment
lands_sc
landsALL
lrgfont
mytho
othgenre
pastorale
peasant
portrait
relig
singlefig
still_life



To start with, we first check on the emprirical distribution of the response variable. There are 2 variables, *logprice* and *price*. From the histogram, 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.

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

For categorical variables:

- *history*, *original*, *type_intermed*, *Shape* 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.
- *mat*, *material* and *materialCat* record duplicate features. We will only keep *materialCat* in future analysis for convenience.
- *origin_author* and *origin_cat* record similar features which is suspecious of highly dependent on each other. We implement Chi-squared test to test the hypothesis, which yield a p-value of 0. 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*.

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.

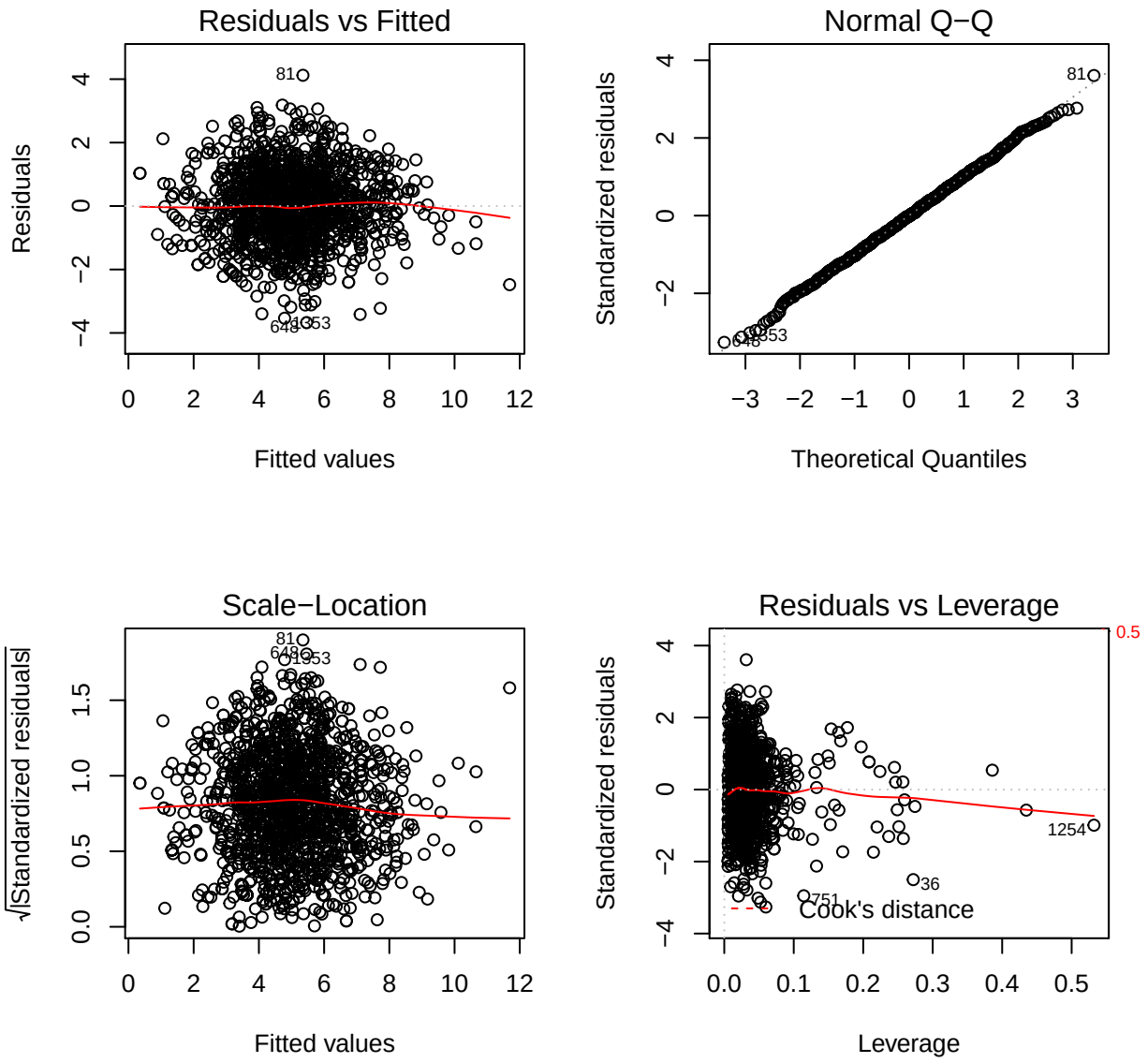
- *Surface* and *Surface_Rect* contain duplicate information, and *Surface* contains all the information in *Surface_Rect*. So we will only keep *Surface*. *Height_in* and *Width_in* are discarded for similar reasoning.

- *nfigures* might need transformations, so I will exponentiate it in the model.

Model Selection

Table 1: Coefficient Summary for Initial Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-200.883	13.951	-14.399	0.000
year	0.116	0.008	14.744	0.000
exp(nfigures)	0.000	0.000	0.694	0.488
origin_authorD/FL	-0.041	0.479	-0.086	0.931
origin_authorF	-0.269	0.493	-0.546	0.585
origin_authorG	-0.015	0.516	-0.029	0.977
origin_authorI	-0.369	0.517	-0.714	0.475
origin_authorS	-0.130	0.650	-0.200	0.842
origin_authorX	-0.944	0.477	-1.979	0.048
dealerL	1.237	0.139	8.877	0.000
dealerP	0.403	0.177	2.279	0.023
dealerR	1.630	0.112	14.605	0.000
discauth1	0.394	0.143	2.760	0.006
endbuyerC	-0.361	0.342	-1.055	0.292
endbuyerD	-0.496	0.339	-1.462	0.144
endbuyerE	-0.930	0.351	-2.647	0.008
endbuyerU	-1.108	0.341	-3.249	0.001
Interm1	0.730	0.136	5.366	0.000
Surface	0.000	0.000	7.369	0.000
materialCatcopper	-0.081	0.122	-0.661	0.508
materialCatother	-0.210	0.231	-0.909	0.364
materialCatUnknown	-0.447	0.122	-3.661	0.000
materialCatwood	-0.104	0.085	-1.233	0.218
diff_origin1	-0.434	0.167	-2.594	0.010
engraved1	0.721	0.147	4.908	0.000
prevcoll1	0.820	0.149	5.504	0.000
origin_catF	-0.251	0.161	-1.557	0.120
origin_catI	-0.296	0.216	-1.371	0.171
origin_catO	-0.034	0.189	-0.178	0.859
origin_catS	0.677	0.986	0.687	0.492
paired1	-0.204	0.071	-2.874	0.004
portrait1	-0.449	0.180	-2.492	0.013
figures1	0.214	0.134	1.593	0.111
finished1	0.681	0.095	7.137	0.000
lrgfont1	0.967	0.121	7.972	0.000
lands_sc1	-0.540	0.127	-4.256	0.000
lands_elem1	0.046	0.075	0.613	0.540
still_life1	-0.245	0.184	-1.334	0.182



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 #1324 and # 751. I will delete these two points to refit the model.

After refitting the data and apply AIC and BIC variable respectively, the new best model we get has an increased R-squared value of 0.6333742. The slope also changed a little. Thus we will set this model which excludes the two influential points to our final model.

Save predictions and intervals.

Once you are satisfied with your model, provide a write up of your data analysis project in a new Rmd file/pdf file: `Part-I-Writeup.Rmd` by copying over salient parts of your R notebook. The written assignment consists of five parts:

1. Introduction: Summary of problem and objectives (5 points)

Table 2: Coefficient Summary for Final Model

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-603.473	211.130	-2.858	0.004	-1017.646	-189.300
year	0.343	0.119	2.879	0.004	0.109	0.577
origin_authorD/FL	364.142	153.558	2.371	0.018	62.909	665.376
origin_authorF	315.577	153.603	2.054	0.040	14.254	616.900
origin_authorG	110.540	184.430	0.599	0.549	-251.257	472.336
origin_authorI	361.808	155.869	2.321	0.020	56.040	667.575
origin_authorS	462.488	209.110	2.212	0.027	52.277	872.700
origin_authorX	223.077	156.171	1.428	0.153	-83.284	529.438
dealerL	190.099	49.984	3.803	0.000	92.045	288.154
dealerP	-11.627	108.561	-0.107	0.915	-224.592	201.338
dealerR	28.495	43.479	0.655	0.512	-56.798	113.788
discauth1	0.363	0.139	2.605	0.009	0.090	0.637
endbuyerC	-48.012	141.999	-0.338	0.735	-326.572	230.547
endbuyerD	97.821	141.511	0.691	0.490	-179.782	375.423
endbuyerE	-148.454	144.041	-1.031	0.303	-431.019	134.112
endbuyerU	-9.841	141.306	-0.070	0.944	-287.041	267.359
Interm1	81.172	52.246	1.554	0.121	-21.320	183.663
Surface	-0.026	0.014	-1.877	0.061	-0.052	0.001
materialCatcopper	-0.075	0.119	-0.627	0.531	-0.309	0.159
materialCatother	-0.288	0.224	-1.286	0.199	-0.728	0.151
materialCatUnknown	-0.480	0.127	-3.776	0.000	-0.729	-0.231
materialCatwood	-0.085	0.082	-1.046	0.296	-0.246	0.075
diff_origin1	66.352	37.081	1.789	0.074	-6.390	139.095
engraved1	0.813	0.143	5.699	0.000	0.533	1.092
prevcoll1	0.794	0.147	5.401	0.000	0.505	1.082
paired1	-0.210	0.069	-3.050	0.002	-0.345	-0.075
portrait1	-0.308	0.176	-1.749	0.080	-0.653	0.037
figures1	62.674	40.011	1.566	0.117	-15.816	141.164
finished1	0.693	0.094	7.375	0.000	0.508	0.877
lrgfont1	0.908	0.119	7.606	0.000	0.674	1.142
lands_sc1	-0.532	0.120	-4.443	0.000	-0.766	-0.297
year:origin_authorD/FL	-0.206	0.087	-2.373	0.018	-0.376	-0.036
year:origin_authorF	-0.179	0.087	-2.059	0.040	-0.349	-0.008
year:origin_authorG	-0.063	0.104	-0.602	0.547	-0.267	0.141
year:origin_authorI	-0.205	0.088	-2.327	0.020	-0.377	-0.032
year:origin_authorS	-0.261	0.118	-2.214	0.027	-0.493	-0.030
year:origin_authorX	-0.127	0.088	-1.435	0.151	-0.299	0.046
year:dealerL	-0.106	0.028	-3.777	0.000	-0.162	-0.051
year:dealerP	0.007	0.061	0.110	0.912	-0.113	0.127
year:dealerR	-0.015	0.025	-0.614	0.539	-0.063	0.033
year:endbuyerC	0.027	0.080	0.335	0.738	-0.130	0.184
year:endbuyerD	-0.055	0.080	-0.694	0.488	-0.212	0.101
year:endbuyerE	0.083	0.081	1.024	0.306	-0.076	0.243
year:endbuyerU	0.005	0.080	0.061	0.951	-0.152	0.161
year:Interm1	-0.045	0.029	-1.541	0.124	-0.103	0.012
year:Surface	0.000	0.000	1.896	0.058	0.000	0.000
year:diff_origin1	-0.038	0.021	-1.805	0.071	-0.079	0.003
year:figures1	-0.035	0.023	-1.559	0.119	-0.079	0.009

2. Exploratory data analysis (10 points): must include three correctly labeled graphs and an explanation that highlight the most important features that went into your model building.
3. Development and assessment of an initial model (10 points)

- Initial model: must include a summary table and an explanation/discussion for variable selection and overall amount of variation explained.
- Model selection: must include a discussion
- Residual: must include residual plot(s) and a discussion.
- Variables: must include table of coefficients and CI

4. Summary and Conclusions (10 points)

What is the (median) price for the "baseline" category if there are categorical or dummy variables in the model (add CI's)? (be sure to include units!) Highlight important findings and potential limitations of your model. Does it appear that interactions are important? What are the most important variables and/or interactions? Provide interprations of how the most important variables influence the (median) price giving a range (CI). Correct interpretation of coefficients for the log model desirable for full points.

Provide recommendations for the art historian about features or combination of features to look for to find the most valuable paintings.

Points will be deducted for code chunks that should not be included, etc.

Upload write up to Sakai any time before Dec 7th

Evaluation on test data for Part I

Once your write up is submitted, your models will be evaluated on the following criteria based on predictions on the test data (20 points):

- Bias: Average (Yhat-Y) positive values indicate the model tends to overestimate price (on average) while negative values indicate the model tends to underestimate price.
- Maximum Deviation: Max |Y-Yhat| - identifies the worst prediction made in the validation data set.
- Mean Absolute Deviation: Average |Y-Yhat| - the average error (regardless of sign).
- Root Mean Square Error: Sqrt Average (Y-Yhat)^2
- Coverage: Average(lwr < Y < upr)

In order to have a passing wercker badge, your file for predictions needs to be the same length as the test data, with three columns: fitted values, lower CI and upper CI values in that order with names, *fit*, *lwr*, and *upr* respectively such as in the code chunk below.

You will be able to see your scores on the score board. They will be initialized by a prediction based on the mean in the training data.