Part-I-Writeup

Rebecca Zhang, Jonathan Klus, Bin Han

1. Introduction:

In this study, we are looking at the auction prices of paintings in 18th century Paris. Specifically, through the assistance of model built based on existing training data, we wish to understand the factors that drive the prices of the paintings, and then be able to predict auction prices based on characteristics of a certain painting. After fitting appropriate model, we also intend to detect specific paintings that are either underpriced or overpriced based on the selected model.

One of the main task and challenge is to narrow down the number of potential predictors from 59 to less than 20 while maintaining a high performance of the model. But being able to explain the results and provide some recommendations to indivisuals without statistical background is equally important and challenging. Therefore, we aim at balancing the performance of model prediction, closeness to true model, simplicity, and interprebility.

2. Exploratory data analysis:

A) Data summary & cleaning

To start with, we looked at the summary of the original trainig data. There are few numeric variables and a lot of binary variables. Some variables, such as `Interm`, `Surface`, `Height_in` etc. have mising values, which need to be taken care of. The followings steps are how we cleaned the data:

a. The first step we did was to get rid of intuitivelly useless variables, including: `lot`, `sale`, `price`, `count`, `subject`, `authorstandard`, `author`, `winningbidder`, and 'other. The are not useful in predicting the response variable (such as names)

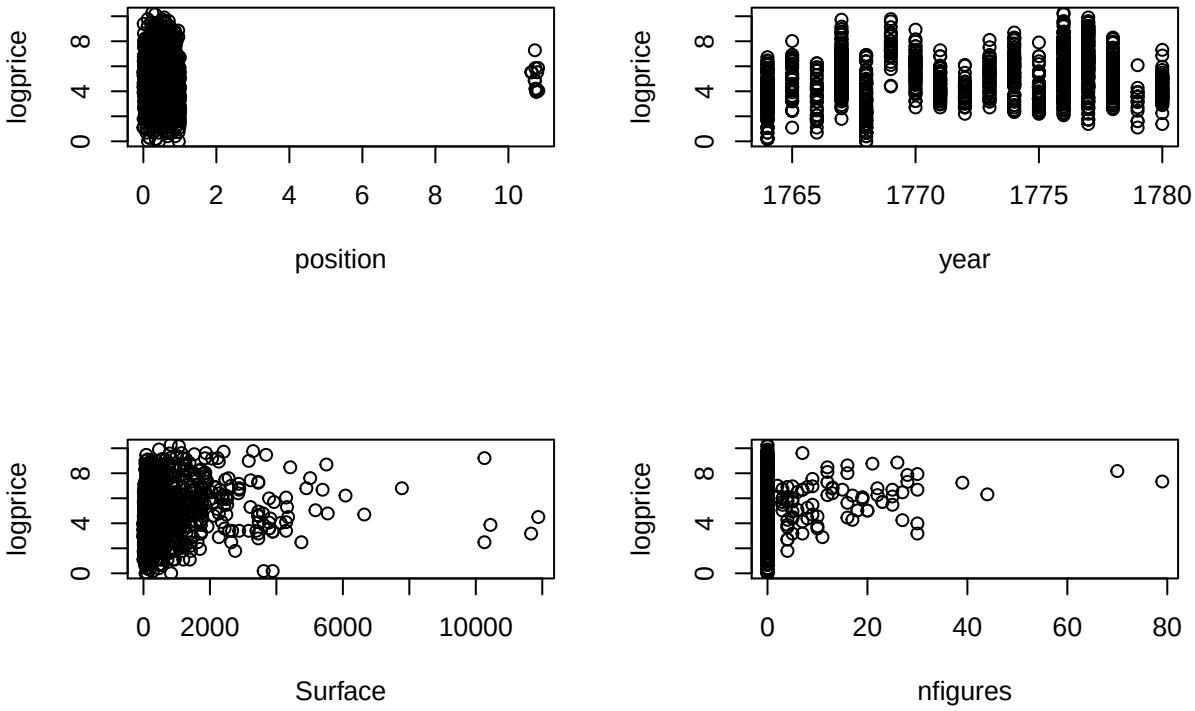
b. By further screening the variables, we found out that `Surface` and `Surface_Rnd`, `Surface_Rect` are corerlated, which are based on the value of `Height_in`, `Width_in`, and `Diam_in`. We decided to use `Surface` in our initial model. The same issue happened to `material`, `mat`, and `materialCat`. The latter one recodes the previous one. Therefore, we used `materialCat`. We applied the same strategy to keep `landsALL` and get rid of other variables related with landscape.

c. For those variables that have multiple levels, to be consistent with how the data was originally coded, we recoded the missing levels as "X", which stands for "no information". For `materialCat` and `Shape`, since there are so many levels, we grouped some levels with few observations together, coded as "other" group. The rest binary vairables are changed into factor.

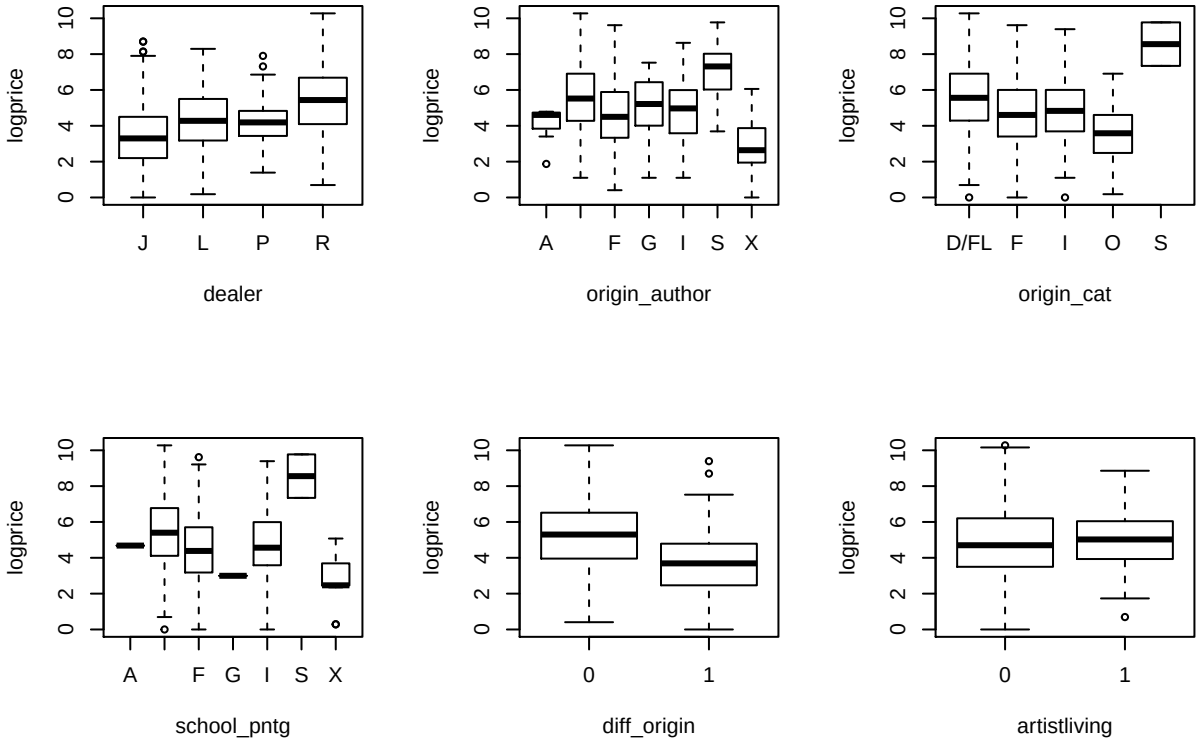
d. Then we dealt with the missing values in `Surface` and `Interm`. We used the package "mice" to address this problem, which uses the observed values in the dataset to impute the missing values. It prevents directly throwing away the missing values, which results in lossing a large amont of information for prediction.

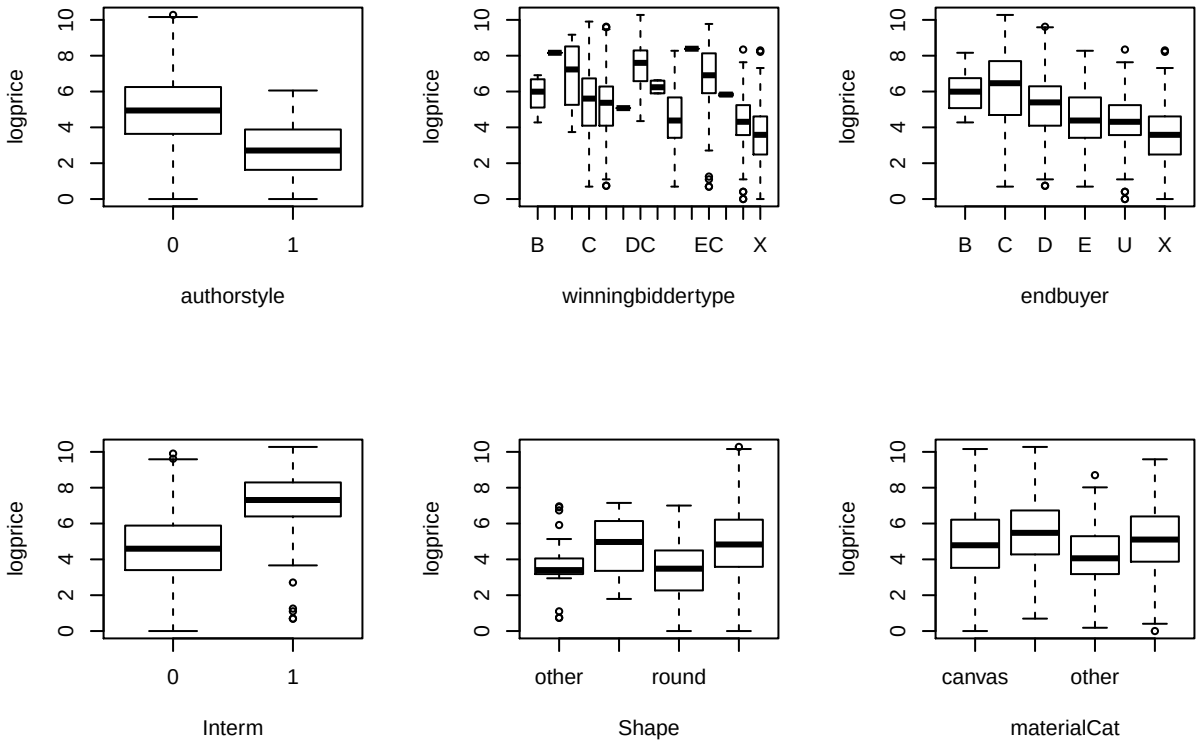
B). Plots

Then we analyed the relationship between those left features and the response variable. With the scatter plots, we can roughly determine which variables can be put into the initial model. For categorical variables, we want to check if the `logprice` spans different ranges in different levels. For numeric variables, we want to check if there is a clear relationship between them and `logprice`.



For numeric variables, we see that `Surface` and `nfigures` seem to show some weak but positive relationship with `logprice`. Since there are several extremely large values in `position` (potentially outliers), it is hard to see that real pattern between the majority of points and `logprice`. But we'll keep it in the model first.





Since there are 33 categorical variables, we don't show the boxplots for all of them. But applied the same method to check all the categorical variables. The following variables show some differences in `logprice` at different levels (not considering the magnitude of the difference at this time): `dealer`, `origin_author`, `origin_cat`, `school_pntg`, `diff_origin`, `authorstyle`, `endbuyer`, `Interm`, `Shape`, `materialCat`, `engraved`, `prevcoll`, `figures`, `finished`, `Irgfont`, `othgenre`, `discauth`, and `still_life`.

If we were to choose 10 best predictive variables for predicting, we would consider the magnitude of differences and the strength of relationships. The 10 variables we choose are: `Surface`, `dealer`, `school_pntg`, `diff_origin`, `authorstyle`, `endbuyer`, `Interm`, `prevcoll`, `engraved`, `Irgfont`.

3. Development and assessment of an initial model

Initial Model

JZS prior

```
##  [1] "Intercept"           "dealerL"             "dealerR"
##  [4] "year"                "origin_authorG"      "origin_authorI"
##  [7] "origin_authorS"      "school_pntgD/FL"     "diff_origin1"
## [10] "artistliving1"       "authorstyle1"        "winningbiddertypeD"
## [13] "winningbiddertypeDC" "winningbiddertypeU"  "winningbiddertypeX"
## [16] "endbuyerE"           "Interm1"             "Shaperound"
## [19] "Surface"             "materialCatother"    "engraved1"
## [22] "prevcoll1"           "paired1"             "finished1"
## [25] "lrgfont1"            "portrait1"           "still_life1"
## [28] "discauth1"
```

g-prior

```
##  [1] "Intercept"           "dealerL"             "dealerR"
##  [4] "origin_catF"         "origin_catO"         "diff_origin1"
##  [7] "artistliving1"       "authorstyle1"        "winningbiddertypeD"
## [10] "winningbiddertypeDC" "winningbiddertypeE"  "winningbiddertypeX"
## [13] "endbuyerC"           "endbuyerE"           "endbuyerX"
## [16] "Interm1"             "Shapeoval"           "Shapesqu_rect"
## [19] "materialCatcopper"   "materialCatother"    "materialCatwood"
## [22] "figures1"            "lrgfont1"            "relig1"
## [25] "landsALL1"           "peasant1"            "singlefig1"
## [28] "portrait1"           "pastorale1"
```

The EDA process gives us an initial idea of which variables to drop out to reduce the dimension, and which variables might be significant in explaining the variation in logprice. But before we built the initial model, we applied BMA, Bayesian Model Averaging, to systemetically choose which base variables that have higher posterior probabilities to be in the initial model. We experimented two modelpriors, "JZS" and "g-prior", which gave us two sets of variables listed above. Then we picked up the common ones from Best Predictive Model(BPM).

Then we fit the linear regression model using the chosen features and all their possible interactions. From the summary table, the $R^2 = 0.5828$, which is fairly high. But we realized that lots of estimated coefficients for interactions are NAs, indicating that some levels in those variables have too few observations to be estimated. Therefore, we need to further reduce the dimention through variable selection.

```
##
## Call:
## lm(formula = logprice ~ dealer + school_pntg + diff_origin +
##     artistliving + endbuyer + authorstyle + Interm + Shape +
##     Surface + engraved + prevcoll + paired + finished + lrgfont +
##     portrait + discauth + still_life, data = paintings_train_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3884 -0.7559  0.0244  0.7966  4.7970
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.620e+00  1.339e+00   3.451 0.000575 ***
## dealerL          1.532e+00  1.349e-01  11.357  < 2e-16 ***
## dealerP          5.203e-01  1.653e-01   3.147 0.001681 **
## dealerR          1.171e+00  1.081e-01  10.835  < 2e-16 ***
## school_pntgD/FL -6.097e-01  1.261e+00  -0.484 0.628701
## school_pntgF    -1.357e+00  1.261e+00  -1.075 0.282364
## school_pntgG    -3.017e+00  1.776e+00  -1.698 0.089639 .
## school_pntgI    -1.264e+00  1.264e+00  -1.001 0.317157
## school_pntgS    -3.095e-01  1.550e+00  -0.200 0.841830
## school_pntgX    -1.960e+00  1.275e+00  -1.538 0.124266
## diff_origin1    -6.456e-01  9.157e-02  -7.050 2.75e-12 ***
## artistliving1    6.789e-01  1.051e-01   6.459 1.43e-10 ***
## endbuyerC       -1.343e-01  3.446e-01  -0.390 0.696843
## endbuyerD       -1.245e-01  3.416e-01  -0.364 0.715645
## endbuyerE       -8.000e-01  3.556e-01  -2.250 0.024596 *
## endbuyerU       -6.266e-01  3.513e-01  -1.784 0.074688 .
## endbuyerX       -1.301e+00  3.491e-01  -3.727 0.000201 ***
```

```
## authorstyle1    -1.073e+00  1.576e-01  -6.809 1.43e-11 ***
## Interm1          1.014e+00  1.403e-01   7.231 7.66e-13 ***
## Shapeoval        4.447e-01  3.864e-01   1.151 0.250064
## Shaperound      -3.729e-01  3.528e-01  -1.057 0.290670
## Shapesqu_rect    5.932e-01  2.581e-01   2.299 0.021664 *
## Surface          2.138e-04  3.337e-05   6.407 2.00e-10 ***
## engraved1        3.234e-01  1.517e-01   2.132 0.033134 *
## prevcoll1        1.215e+00  1.504e-01   8.078 1.36e-15 ***
## paired1         -3.440e-01  7.079e-02  -4.859 1.31e-06 ***
## finished1        6.287e-01  9.679e-02   6.496 1.13e-10 ***
## lrgfont1         1.091e+00  1.246e-01   8.757  < 2e-16 ***
## portrait1       -6.486e-01  1.765e-01  -3.675 0.000246 ***
## discauth1        3.770e-01  1.450e-01   2.600 0.009425 **
## still_life1     -7.235e-01  1.715e-01  -4.219 2.60e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.249 on 1469 degrees of freedom
## Multiple R-squared:  0.5843, Adjusted R-squared:  0.5758
## F-statistic: 68.82 on 30 and 1469 DF,  p-value: < 2.2e-16
```

Model Selection

After completing the initial exploratory data analysis, methods including Stepwise Best Subset Selection using both AIC and BIC were used in order to assess more systematically which covariates were most important for predicting the logprice of paintings. While the number of relevant covariates was initially thinned by examining the data and determining which variables were best suited for modeling (e.g. via dimension reduction, elimination or recoding of categorical variables with too many levels or too few observations for a given level to be useful in estimating a coefficient), there still remained a large number of covariates from which to choose. The goal in using the above described methodology was to demonstrate among several methods, both frequentist and Bayesian, which covariates were routinely deemed to be the most important for modeling logprice.

The variable selection methods described above remain computationally intensive, particularly given the number of variables and potential two-way interactions that must be considered. In order to begin the analysis, The two-way interactions were considered using stepwise selection (AIC & BIC). The goal of this penalized selection process was to avoid overfitting and to deliver a model that was both interpretable and performed well in prediction. Then we compared the results from two methods and filtered out interactions that have NAs as coefficients, that are not significant, and that do not make sense to be interacted (such as $artistling * endbuyer$).

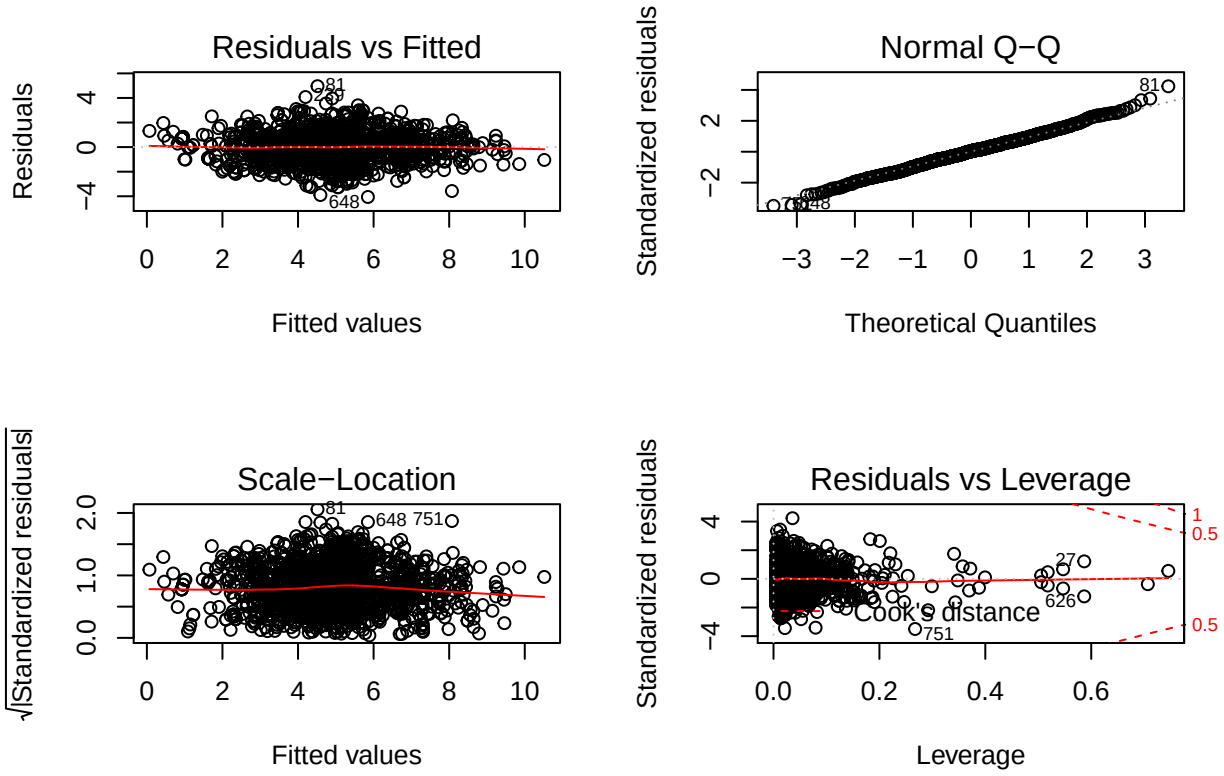
Ultimately, the following variables were selected using the above methods and were fit using OLS regression. The R^2 reduces to 0.6315, which is expected. All the estimated coefficients do not contain NAs.

```
##
## Call:
## lm(formula = logprice ~ Shape + school_pntg + dealer * Interm +
##     dealer * Surface + dealer * paired + dealer * finished +
##     dealer * discauth + diff_origin * Surface + diff_origin *
##     portrait + artistliving * endbuyer + artistliving * authorstyle +
##     Interm * Surface + Interm * lrgfont + Surface * lrgfont +
##     Surface * still_life + Surface * discauth + prevcoll * finished +
##     paired * lrgfont + paired * discauth + diff_origin * authorstyle +
##     diff_origin * still_life + finished * discauth + lrgfont *
##     discauth + artistliving * finished + Interm * portrait +
```

```
##     dealer * artistliving + authorstyle * prevcoll, data = paintings_train_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0581 -0.7264  0.0329  0.7495  4.9569
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4.120e+00  1.316e+00   3.131 0.001777 **
## Shapeoval                   5.079e-01  3.798e-01   1.337 0.181391
## Shaperound                 -5.016e-01  3.476e-01  -1.443 0.149260
## Shapesqu_rect               5.109e-01  2.591e-01   1.972 0.048857 *
## school_pntgD/FL            -6.291e-01  1.223e+00  -0.514 0.607095
## school_pntgF               -1.307e+00  1.224e+00  -1.067 0.286075
## school_pntgG               -3.086e+00  1.717e+00  -1.798 0.072436 .
## school_pntgI               -1.256e+00  1.227e+00  -1.024 0.305922
## school_pntgS               -7.027e-01  1.518e+00  -0.463 0.643542
## school_pntgX               -1.856e+00  1.238e+00  -1.499 0.134132
## dealerL                     2.676e+00  2.161e-01  12.386  < 2e-16 ***
## dealerP                     1.323e+00  2.866e-01   4.616 4.27e-06 ***
## dealerR                     1.838e+00  1.864e-01   9.861  < 2e-16 ***
## Interm1                    -1.038e+00  4.817e-01  -2.155 0.031336 *
## Surface                     8.175e-04  1.973e-04   4.143 3.63e-05 ***
## paired1                     1.738e-01  1.918e-01   0.906 0.365064
## finished1                   1.168e+00  2.225e-01   5.249 1.76e-07 ***
## discauth1                   9.916e-01  3.913e-01   2.534 0.011370 *
## diff_origin1               -5.652e-01  1.061e-01  -5.329 1.15e-07 ***
## portrait1                  -9.124e-01  2.060e-01  -4.428 1.02e-05 ***
## artistliving1              -1.256e-01  8.379e-01  -0.150 0.880824
## endbuyerC                  -3.243e-01  3.745e-01  -0.866 0.386626
## endbuyerD                  -3.244e-01  3.702e-01  -0.876 0.381029
## endbuyerE                  -1.107e+00  3.832e-01  -2.889 0.003928 **
## endbuyerU                  -8.326e-01  3.811e-01  -2.185 0.029059 *
## endbuyerX                  -1.570e+00  3.787e-01  -4.145 3.60e-05 ***
## authorstyle1               -1.775e+00  4.223e-01  -4.204 2.79e-05 ***
## lrgfont1                    1.774e+00  1.767e-01  10.038  < 2e-16 ***
## still_life1                -1.545e-01  2.452e-01  -0.630 0.528719
## prevcoll1                   1.522e+00  1.708e-01   8.913  < 2e-16 ***
## dealerL:Interm1             1.335e+00  8.560e-01   1.559 0.119197
## dealerP:Interm1             1.707e+00  1.349e+00   1.265 0.206094
## dealerR:Interm1             2.320e+00  4.965e-01   4.672 3.26e-06 ***
## dealerL:Surface            -6.386e-04  2.074e-04  -3.079 0.002114 **
## dealerP:Surface            -3.676e-04  3.416e-04  -1.076 0.282034
## dealerR:Surface            -5.452e-04  1.991e-04  -2.739 0.006247 **
## dealerL:paired1            -1.206e+00  2.554e-01  -4.722 2.56e-06 ***
## dealerP:paired1            -6.620e-01  3.567e-01  -1.856 0.063668 .
## dealerR:paired1            -2.836e-01  2.103e-01  -1.348 0.177745
## dealerL:finished1          -6.609e-01  4.996e-01  -1.323 0.186141
## dealerP:finished1          -7.056e-01  4.090e-01  -1.725 0.084703 .
## dealerR:finished1          -5.949e-01  2.427e-01  -2.451 0.014356 *
## dealerL:discauth1           1.795e-02  9.218e-01   0.019 0.984463
## dealerP:discauth1          -1.778e+00  9.497e-01  -1.872 0.061366 .
## dealerR:discauth1          -9.274e-01  3.618e-01  -2.563 0.010474 *
## Surface:diff_origin1       -8.858e-05  9.470e-05  -0.935 0.349739
```

```
## diff_origin1:portrait1      8.747e-01  4.107e-01   2.130 0.033359 *
## artistliving1:endbuyerC     1.017e+00  8.117e-01   1.252 0.210634
## artistliving1:endbuyerD     1.185e+00  8.032e-01   1.476 0.140199
## artistliving1:endbuyerE     2.141e+00  8.616e-01   2.484 0.013090 *
## artistliving1:endbuyerU     1.241e+00  8.285e-01   1.498 0.134345
## artistliving1:endbuyerX     1.437e+00  8.242e-01   1.743 0.081519 .
## artistliving1:authorstyle1  1.446e+00  8.972e-01   1.612 0.107230
## Interm1:Surface             4.265e-04  1.579e-04   2.702 0.006973 **
## Interm1:lrgfont1           -8.757e-01  2.704e-01  -3.239 0.001229 **
## Surface:lrgfont1           -2.100e-04  1.149e-04  -1.827 0.067881 .
## Surface:still_life1        -5.724e-04  2.535e-04  -2.258 0.024103 *
## Surface:discauth1          -1.499e-04  2.260e-04  -0.663 0.507128
## finished1:prevcoll1        -1.077e+00  3.258e-01  -3.307 0.000968 ***
## paired1:lrgfont1           -7.755e-01  2.549e-01  -3.043 0.002386 **
## paired1:discauth1          -4.329e-01  3.452e-01  -1.254 0.210000
## diff_origin1:authorstyle1   9.487e-01  4.519e-01   2.099 0.035949 *
## diff_origin1:still_life1   -7.804e-01  3.485e-01  -2.239 0.025282 *
## finished1:discauth1         5.871e-01  3.109e-01   1.889 0.059153 .
## discauth1:lrgfont1         -4.969e-01  4.384e-01  -1.133 0.257225
## finished1:artistliving1    -4.431e-01  2.843e-01  -1.558 0.119342
## Interm1:portrait1          -8.631e-01  5.866e-01  -1.471 0.141388
## dealerL:artistliving1      -1.001e+00  3.604e-01  -2.778 0.005533 **
## dealerP:artistliving1      -6.788e-01  4.353e-01  -1.559 0.119141
## dealerR:artistliving1      -5.012e-01  2.984e-01  -1.680 0.093268 .
## authorstyle1:prevcoll1     -1.788e+00  1.268e+00  -1.410 0.158885
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.193 on 1429 degrees of freedom
## Multiple R-squared:  0.6316, Adjusted R-squared:  0.6135
## F-statistic:    35 on 70 and 1429 DF,  p-value: < 2.2e-16
```

Residuals & Diagnostics Analysis



After fitting the model, we created the four model diagnostic plots. The overall appearances of all four plots seem acceptable, with no obvious outlier or highly influential points shown. The model also does not violate the normality assumption. The constant variance of residuals assumption seems to be satisfied. However, there are 2 cases that are dropped from the plots because they both have leverage of 1, indicating that they could potentially be the outlying cases of underpriced/overpriced paintings that we will later on investigate in, or have extreme price values. It is worth our attention to specifically look at these cases.

Variables

	Coefficient	2.5%	97.5%
(Intercept)	61.562	4.659	813.467
Shapeoval	1.662	0.789	3.501
Shaperound	0.606	0.306	1.198
Shapesqu_rect	1.667	1.003	2.771
school_pntgD/FL	0.533	0.048	5.872
school_pntgF	0.271	0.025	2.990
school_pntgG	0.046	0.002	1.325
school_pntgI	0.285	0.026	3.158
school_pntgS	0.495	0.025	9.733
school_pntgX	0.156	0.014	1.774
dealerL	14.530	9.510	22.200
dealerP	3.754	2.140	6.586
dealerR	6.281	4.358	9.054
Interm1	0.354	0.138	0.911
Surface	1.001	1.000	1.001

	Coefficient	2.5%	97.5%
paired1	1.190	0.817	1.733
finished1	3.216	2.078	4.975
discauth1	2.696	1.251	5.807
diff_origin1	0.568	0.461	0.700
portrait1	0.402	0.268	0.602
artistliving1	0.882	0.170	4.563
endbuyerC	0.723	0.347	1.507
endbuyerD	0.723	0.350	1.495
endbuyerE	0.331	0.156	0.701
endbuyerU	0.435	0.206	0.918
endbuyerX	0.208	0.099	0.437
authorstyle1	0.169	0.074	0.388
lrgfont1	5.893	4.167	8.334
still_life1	0.857	0.530	1.386
prevcoll1	4.584	3.279	6.408
dealerL:Interm1	3.799	0.709	20.367
dealerP:Interm1	5.511	0.391	77.731
dealerR:Interm1	10.172	3.841	26.939
dealerL:Surface	0.999	0.999	1.000
dealerP:Surface	1.000	0.999	1.000
dealerR:Surface	0.999	0.999	1.000
dealerL:paired1	0.299	0.181	0.494
dealerP:paired1	0.516	0.256	1.038
dealerR:paired1	0.753	0.499	1.138
dealerL:finished1	0.516	0.194	1.376
dealerP:finished1	0.494	0.221	1.102
dealerR:finished1	0.552	0.343	0.888
dealerL:discauth1	1.018	0.167	6.210
dealerP:discauth1	0.169	0.026	1.089
dealerR:discauth1	0.396	0.195	0.804
Surface:diff_origin1	1.000	1.000	1.000
diff_origin1:portrait1	2.398	1.072	5.368
artistliving1:endbuyerC	2.764	0.562	13.583
artistliving1:endbuyerD	3.272	0.677	15.817
artistliving1:endbuyerE	8.505	1.569	46.102
artistliving1:endbuyerU	3.459	0.681	17.572
artistliving1:endbuyerX	4.207	0.835	21.188
artistliving1:authorstyle1	4.246	0.731	24.681
Interm1:Surface	1.000	1.000	1.001
Interm1:lrgfont1	0.417	0.245	0.708
Surface:lrgfont1	1.000	1.000	1.000
Surface:still_life1	0.999	0.999	1.000
Surface:discauth1	1.000	0.999	1.000
finished1:prevcoll1	0.340	0.180	0.645
paired1:lrgfont1	0.460	0.279	0.759
paired1:discauth1	0.649	0.330	1.277
diff_origin1:authorstyle1	2.582	1.064	6.265
diff_origin1:still_life1	0.458	0.231	0.908
finished1:discauth1	1.799	0.978	3.310
discauth1:lrgfont1	0.608	0.257	1.438
finished1:artistliving1	0.642	0.368	1.121
Interm1:portrait1	0.422	0.133	1.333

	Coefficient	2.5%	97.5%
dealerL:artistliving1	0.367	0.181	0.745
dealerP:artistliving1	0.507	0.216	1.191
dealerR:artistliving1	0.606	0.337	1.088
authorstyle1:prevcoll1	0.167	0.014	2.014

In the linear model we selected, we included `Shape`, `school_pntg`, `dealer`, `Interm`, `Surface`, `paired`, `finished`, `discauth`, `diff_origin`, `portrait`, `artistliving`, `endbuyer`, `authorstyle`, `lrgfont`, `still_life`, and `prevcoll` as our base predictors. Interactions selected by the model selection process and, for the sake of interpretation, those that are reasonable and interpretable are kept in the model as well. Since the response variable was orginally log-transformed, the exponentiated coefficients and confidence intervals are shown in the table.

4. Summary and Conclusions

a. The median price predicted is `exp(4.401232) = 81.55128` livres. The 95% confidence interval is (6.248, 1064.357) livres. The prediction interval is (2.532, 2626.714) livres.

Table 2: 95% Confidence Interval

fit	lwr	upr
76.767	5.839	1009.309

Table 3: 95% Prediction Interval

fit	lwr	upr
76.767	2.365	2491.383

Interpretation

From the final model we fitted, we found out that the following variables are statistically significant: `dealer`, `Interm`, `Surface`, `finished`, `discauth`, `diff_origin`, `portrait`, `endbuyer (E,U, X)`, `authorstyle`, `lrgfont`, and `prevcoll`. Some of the interactions are statistically important, such as: `dealer*Interm`, `dealer*paried`, `Interm*lrgfont`, `diff_origin*portrait` etc. We picked the most important ones and interpreted as following:

- dealerL: compared with dealer J, the average selling price from dealer L is `exp(2.526) = 12.50339` times higher. (Same interpretation for dealer P and R, with different coefficients)
- Interm: when there is an intermediary involved in the transaction, the selling price is `exp(-1.523) = 0.218` times lower than when there is no intermediary involved.
- Surface: for every one squared inches increase in the painting surface, the selling price is expected to increase 8.779e-4 livres.
- finished: if the painting is finished, the selling price on average is `exp(1.087) = 2.965` times higher than when the painting is not finished.
- portrait: if the painting is described as a portrait, the selling price on average is `exp(-0.9246) = 0.3967` times lower than when the painting is not described as a portrait.

- dealerR&Interm1: if the dealer is R and with an intermediary existed, the average selling price is 1.854 times higher when the dealer is J with an intermediary.

Recommendations

In order to find the most valuable paintings, we recommend historians focusing on several features (just to mention some, not a complete list). For example, they might want to look for dealer R, with an intermediary involved; they might want to look for dealer L with the painting sold as a pair with another one; they should look for larger, finished paintings; they should focus on paintings whose authors' names are mentioned during the auction.

Limitations

1. As mentioned in the data cleaning process, some of the variables have too many levels to be fitted and some levels have few observations that are not sufficient for estimating coefficients. Therefore, we grouped some of the variables, leading to the result that our model cannot investigate the true effects of those combined levels. If we had a larger data set, we could potentially release more levels.
2. As our goal is to find a balance between the prediction accuracy and interpretability, our model will not predict the response variable very accurately (as a sacrifice on interpretability).