Part-I-Writeup

Jingyi 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:

```
##
##      lot        Length:1500        Class :character   Mode  :character
##      sale       Length:1500        Class :character   Mode  :character
##     price       Length:1500        Class :character   Mode  :character
##      count      Min.   :1          1st Qu.:1          Median :1
##    subject      Length:1500        Class :character   Mode  :character
## authorstandard Length:1500        Class :character   Mode  :character
##     author      Length:1500        Class :character   Mode  :character
## winningbidder  Length:1500        Class :character   Mode  :character
##      other      Min.   :0.000      1st Qu.:0.000      Median :0.000
##
##      lot
##      sale
##     price
##      count      Mean   :1       3rd Qu.:1       Max.   :1
##    subject
## authorstandard
##     author
## winningbidder
##      other      Mean   :0.016   3rd Qu.:0.000   Max.   :1.000
```

variables	unique_values
lot	308
sale	27
price	546
count	1

variables	unique_values
subject	1258
authorstandard	519
author	831
winningbidder	312
other	2

a. The first step we did was to get rid of intuitivelly useless variables to reduce dimention, including: `lot`, `sale`, `price`, `count`, `subject`, `authorstandard`, `author`, `winningbidder`, and `other`. From the summary table, the `count` variable has all 1's; the `other` variable does not convey useful information; the other variables, such as `names` and `subjects`, are not useful in predicting the response variable (such as names). From the table of unique values we can see that some variables have thousands of unique values. Therefore, we can remove them in the first step.

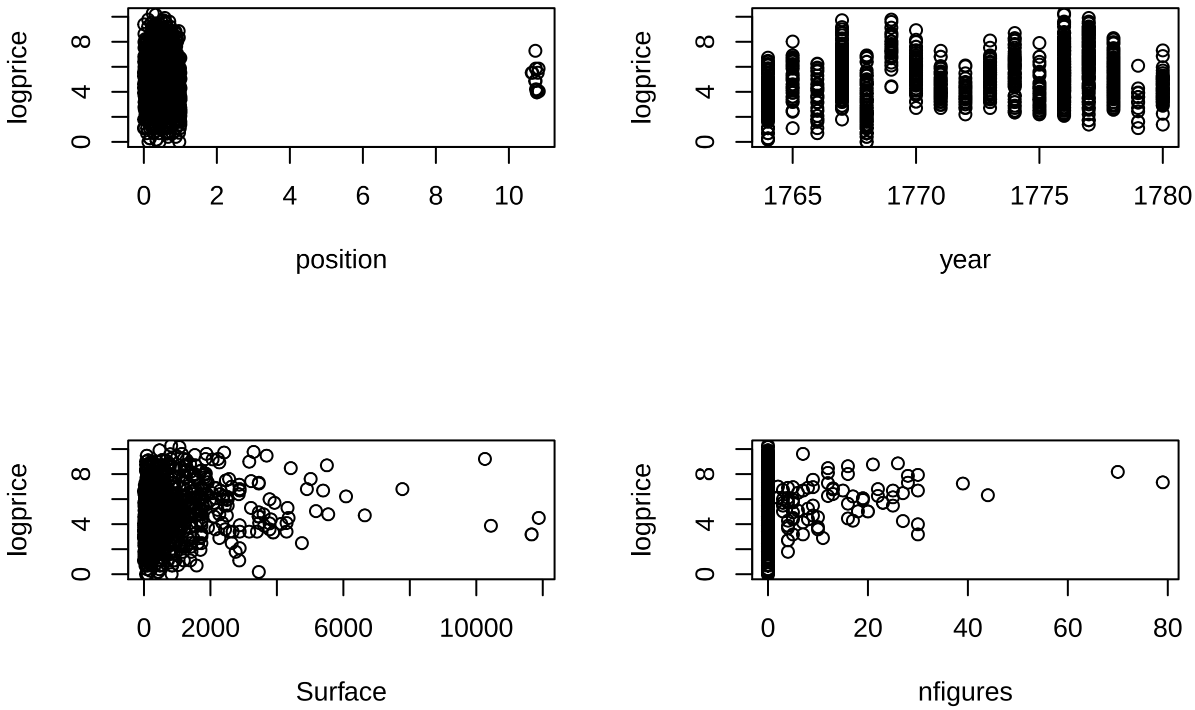
b. By further screening the variables, we found out that `Surface` and `Surface_Rnd`, `Surface_Rect` are corerlated, which are based on the value of `Height_in`, `Width_in`, and `Diam_in`. We decided to use `Surface` in our initial model. The same issue happened to `material`, `mat`, and `materialCat`. 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.

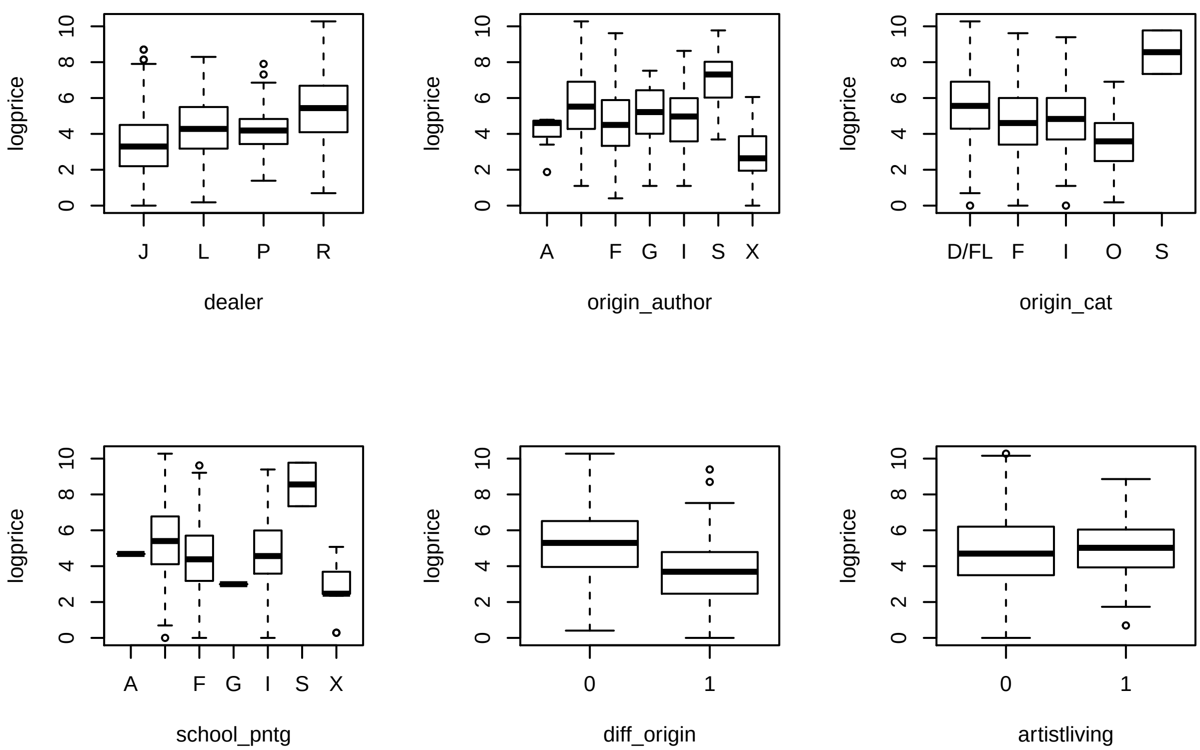
d. Then we dealt with the missing values in `Surface` and `Interm`. We used the package "mice" to address this problem, which uses the observed values in the dataset to impute the missing values. It prevents directly throwing away the missing values, which results in lossing a large amont of information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

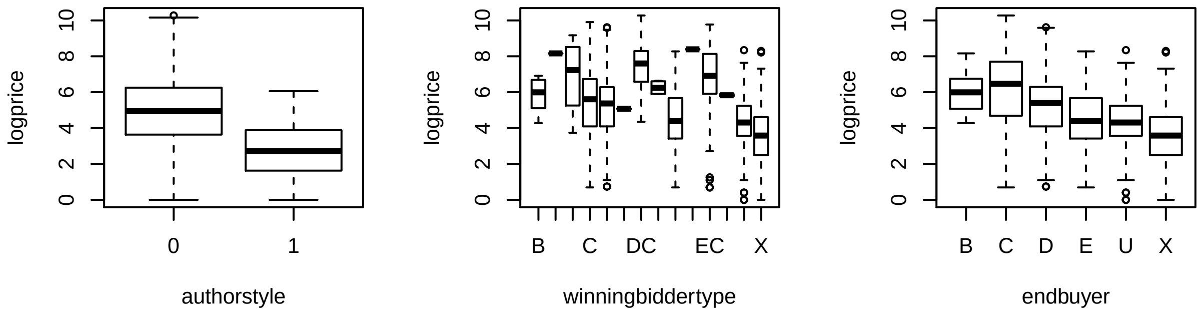
B). Plots

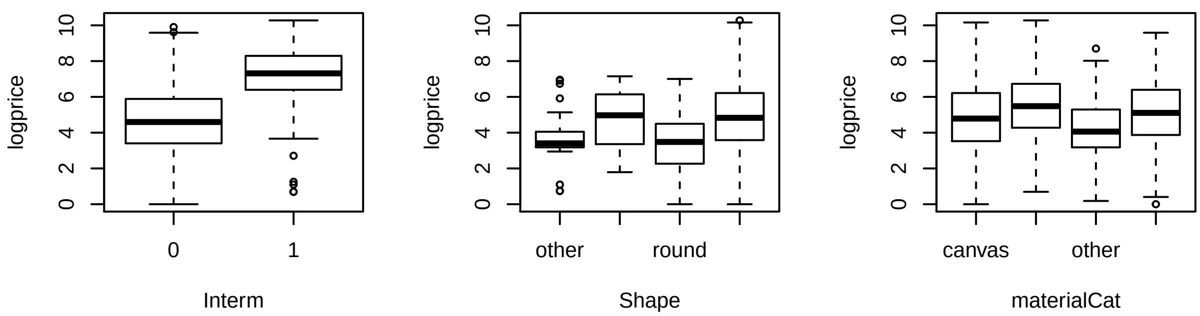
Then we analyed the relationship between those left features and the response variable. With the scatter plots, we can roughly determine which variables can be put into the initial model. For categorical variables, we want to check if the `logprice` spans different ranges in different levels. For numeric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

3. Development and assessment of an initial model

Initial Model

JZS prior

```
##  [1] "Intercept"            "dealerL"              "dealerR"
##  [4] "year"                 "origin_authorX"       "school_pntgD/FL"
##  [7] "school_pntgG"         "school_pntgX"         "diff_origin1"
## [10] "artistliving1"        "authorstyle1"         "winningbiddertypeC"
## [13] "winningbiddertypeDC" "winningbiddertypeE"   "winningbiddertypeU"
## [16] "endbuyerD"            "endbuyerX"            "Interm1"
## [19] "Shaperound"           "Surface"              "materialCatother"
## [22] "engraved1"            "prevcoll1"            "paired1"
## [25] "finished1"            "lrgfont1"             "arch1"
## [28] "peasant1"             "portrait1"            "still_life1"
## [31] "discauth1"            "pastorale1"
```

g-prior

```
##  [1] "Intercept"           "dealerL"             "dealerP"
##  [4] "dealerR"             "origin_authorF"      "school_pntgD/FL"
##  [7] "diff_origin1"        "winningbiddertypeC"  "winningbiddertypeDC"
## [10] "winningbiddertypeE"  "winningbiddertypeEC" "winningbiddertypeX"
## [13] "endbuyerC"           "endbuyerE"           "endbuyerX"
## [16] "Interm1"             "Shapeoval"           "Shaperound"
## [19] "Shapesqu_rect"       "materialCatother"    "engraved1"
## [22] "prevcoll1"           "othartist1"          "figures1"
## [25] "lrgfont1"            "relig1"              "peasant1"
## [28] "history1"            "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
## -4.3859 -0.7578  0.0275  0.8011  4.7970
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.602e+00  1.339e+00   3.437 0.000605 ***
## dealerL          1.537e+00  1.349e-01  11.394  < 2e-16 ***
## dealerP          5.260e-01  1.654e-01   3.180 0.001502 **
## dealerR          1.179e+00  1.081e-01  10.899  < 2e-16 ***
## school_pntgD/FL -6.132e-01  1.261e+00  -0.486 0.626849
## school_pntgF    -1.354e+00  1.262e+00  -1.073 0.283284
## school_pntgG    -3.313e+00  1.777e+00  -1.864 0.062525 .
## school_pntgI    -1.272e+00  1.264e+00  -1.006 0.314433
## school_pntgS    -3.038e-01  1.551e+00  -0.196 0.844761
## school_pntgX    -1.966e+00  1.275e+00  -1.542 0.123317
## diff_origin1    -6.488e-01  9.170e-02  -7.075 2.31e-12 ***
## artistliving1    6.715e-01  1.051e-01   6.389 2.24e-10 ***
## endbuyerC       -1.370e-01  3.448e-01  -0.398 0.691040
## endbuyerD       -1.283e-01  3.417e-01  -0.376 0.707294
## endbuyerE       -8.006e-01  3.557e-01  -2.251 0.024557 *
## endbuyerU       -6.279e-01  3.515e-01  -1.787 0.074215 .
## endbuyerX       -1.306e+00  3.493e-01  -3.740 0.000191 ***
```

```
## authorstyle1    -1.053e+00  1.577e-01  -6.676 3.47e-11 ***
## Interm1          1.011e+00  1.403e-01   7.205 9.25e-13 ***
## Shapeoval        4.606e-01  3.866e-01   1.191 0.233662
## Shaperound      -3.538e-01  3.529e-01  -1.003 0.316161
## Shapesqu_rect    6.133e-01  2.579e-01   2.377 0.017560 *
## Surface          2.110e-04  3.338e-05   6.321 3.45e-10 ***
## engraved1        3.224e-01  1.517e-01   2.125 0.033737 *
## prevcoll1        1.218e+00  1.504e-01   8.093 1.21e-15 ***
## paired1         -3.443e-01  7.082e-02  -4.862 1.29e-06 ***
## finished1        6.275e-01  9.682e-02   6.481 1.24e-10 ***
## lrgfont1         1.093e+00  1.247e-01   8.765  < 2e-16 ***
## portrait1       -6.504e-01  1.765e-01  -3.684 0.000238 ***
## discauth1        3.746e-01  1.451e-01   2.582 0.009916 **
## still_life1     -7.291e-01  1.715e-01  -4.251 2.26e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.25 on 1469 degrees of freedom
## Multiple R-squared:  0.584,  Adjusted R-squared:  0.5755
## F-statistic: 68.73 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 $artistling * endbuyer$).

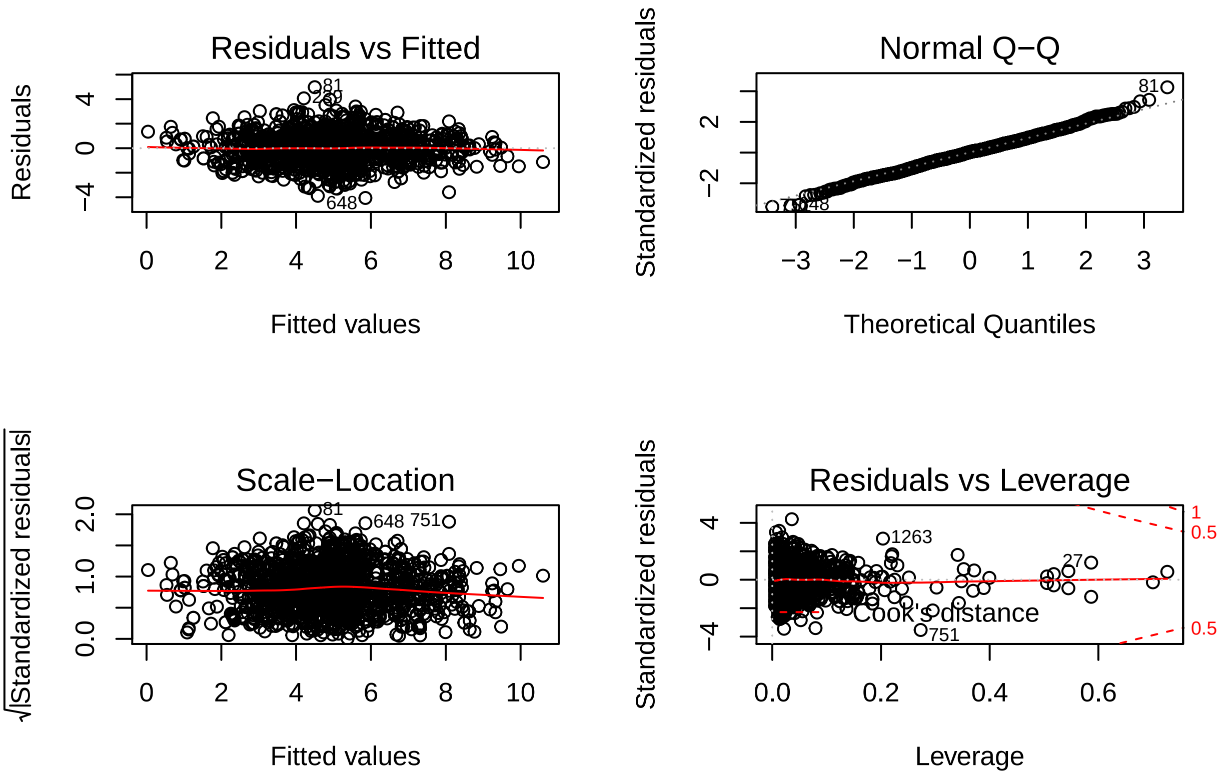
Ultimately, the following variables were selected using the above methods and were fit using OLS regression. The R^2 reduces to 0.6315, which is expected. All the estimated coefficients do not contain NAs.

```
##
## 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
## -4.0599 -0.7246  0.0421  0.7497  4.9748
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.023e+00  1.315e+00   3.059 0.002259 **
## Shapeoval              5.765e-01  3.792e-01   1.520 0.128615
## Shaperound            -4.354e-01  3.475e-01  -1.253 0.210350
## Shapesqu_rect          5.782e-01  2.584e-01   2.237 0.025408 *
## school_pntgD/FL       -6.230e-01  1.221e+00  -0.510 0.609990
## school_pntgF          -1.303e+00  1.222e+00  -1.066 0.286633
## school_pntgG          -3.315e+00  1.718e+00  -1.930 0.053800 .
## school_pntgI          -1.253e+00  1.225e+00  -1.023 0.306486
## school_pntgS          -7.077e-01  1.516e+00  -0.467 0.640581
## school_pntgX          -1.856e+00  1.236e+00  -1.501 0.133611
## dealerL                2.682e+00  2.170e-01  12.359  < 2e-16 ***
## dealerP                1.340e+00  2.880e-01   4.653 3.58e-06 ***
## dealerR                1.869e+00  1.887e-01   9.906  < 2e-16 ***
## Interm1               -1.239e+00  4.822e-01  -2.570 0.010278 *
## Surface                8.438e-04  1.973e-04   4.278 2.02e-05 ***
## paired1                1.988e-01  1.927e-01   1.031 0.302625
## finished1              1.165e+00  2.222e-01   5.241 1.83e-07 ***
## discauth1              9.286e-01  3.913e-01   2.373 0.017759 *
## diff_origin1          -5.523e-01  1.067e-01  -5.176 2.60e-07 ***
## portrait1             -9.320e-01  2.056e-01  -4.533 6.30e-06 ***
## artistliving1         -1.138e-01  8.365e-01  -0.136 0.891829
## endbuyerC             -3.310e-01  3.739e-01  -0.885 0.376064
## endbuyerD             -3.333e-01  3.696e-01  -0.902 0.367363
## endbuyerE             -1.105e+00  3.825e-01  -2.889 0.003919 **
## endbuyerU             -8.314e-01  3.804e-01  -2.186 0.029011 *
## endbuyerX             -1.574e+00  3.780e-01  -4.163 3.32e-05 ***
## authorstyle1          -1.775e+00  4.217e-01  -4.208 2.73e-05 ***
## lrgfont1               1.784e+00  1.762e-01  10.126  < 2e-16 ***
## still_life1           -1.298e-01  2.437e-01  -0.533 0.594312
## prevcoll1              1.526e+00  1.705e-01   8.946  < 2e-16 ***
## dealerL:Interm1        1.507e+00  8.524e-01   1.768 0.077246 .
## dealerP:Interm1        1.861e+00  1.346e+00   1.383 0.166918
## dealerR:Interm1        2.507e+00  4.947e-01   5.067 4.57e-07 ***
## dealerL:Surface       -6.261e-04  2.078e-04  -3.014 0.002625 **
## dealerP:Surface       -3.786e-04  3.426e-04  -1.105 0.269376
## dealerR:Surface       -5.699e-04  1.994e-04  -2.857 0.004334 **
## dealerL:paired1       -1.221e+00  2.556e-01  -4.777 1.96e-06 ***
## dealerP:paired1       -6.733e-01  3.565e-01  -1.889 0.059150 .
## dealerR:paired1       -3.061e-01  2.110e-01  -1.451 0.146977
## dealerL:finished1     -6.422e-01  4.987e-01  -1.288 0.198069
## dealerP:finished1     -6.967e-01  4.083e-01  -1.706 0.088197 .
## dealerR:finished1     -5.909e-01  2.423e-01  -2.439 0.014868 *
## dealerL:discauth1      3.725e-02  9.203e-01   0.040 0.967717
## dealerP:discauth1     -1.719e+00  9.488e-01  -1.812 0.070242 .
## dealerR:discauth1     -8.857e-01  3.620e-01  -2.446 0.014552 *
## Surface:diff_origin1  -1.210e-04  9.693e-05  -1.249 0.212028
```

```
## diff_origin1:portrait1      8.969e-01  4.097e-01   2.189 0.028741 *
## artistliving1:endbuyerC     1.030e+00  8.104e-01   1.271 0.203812
## artistliving1:endbuyerD     1.202e+00  8.020e-01   1.499 0.134178
## artistliving1:endbuyerE     2.127e+00  8.603e-01   2.473 0.013513 *
## artistliving1:endbuyerU     1.242e+00  8.272e-01   1.501 0.133545
## artistliving1:endbuyerX     1.421e+00  8.229e-01   1.727 0.084406 .
## artistliving1:authorstyle1  1.456e+00  8.954e-01   1.626 0.104177
## Interm1:Surface             4.693e-04  1.549e-04   3.029 0.002496 **
## Interm1:lrgfont1           -9.044e-01  2.692e-01  -3.359 0.000801 ***
## Surface:lrgfont1           -2.249e-04  1.144e-04  -1.966 0.049439 *
## Surface:still_life1        -6.215e-04  2.465e-04  -2.521 0.011799 *
## Surface:discauth1          -9.433e-05  2.248e-04  -0.420 0.674857
## finished1:prevcoll1        -1.075e+00  3.253e-01  -3.305 0.000973 ***
## paired1:lrgfont1           -7.640e-01  2.543e-01  -3.004 0.002713 **
## paired1:discauth1          -4.278e-01  3.434e-01  -1.246 0.213076
## diff_origin1:authorstyle1   9.634e-01  4.513e-01   2.135 0.032946 *
## diff_origin1:still_life1   -7.323e-01  3.488e-01  -2.100 0.035945 *
## finished1:discauth1         5.747e-01  3.101e-01   1.854 0.064001 .
## discauth1:lrgfont1         -5.485e-01  4.368e-01  -1.256 0.209449
## finished1:artistliving1    -4.360e-01  2.838e-01  -1.536 0.124744
## Interm1:portrait1          -8.514e-01  5.854e-01  -1.454 0.146100
## dealerL:artistliving1      -1.022e+00  3.588e-01  -2.849 0.004450 **
## dealerP:artistliving1      -6.709e-01  4.347e-01  -1.543 0.122993
## dealerR:artistliving1      -5.191e-01  2.980e-01  -1.742 0.081692 .
## authorstyle1:prevcoll1     -1.696e+00  1.270e+00  -1.335 0.182014
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.191 on 1429 degrees of freedom
## Multiple R-squared:  0.6328, Adjusted R-squared:  0.6148
## F-statistic: 35.18 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)	4.023	1.444	6.602
Shapeoval	0.577	-0.167	1.320
Shaperound	-0.435	-1.117	0.246
Shapesqu_rect	0.578	0.071	1.085
school_pntgD/FL	-0.623	-3.018	1.772
school_pntgF	-1.303	-3.701	1.095
school_pntgG	-3.315	-6.684	0.054
school_pntgI	-1.253	-3.655	1.150
school_pntgS	-0.708	-3.681	2.265
school_pntgX	-1.856	-4.281	0.570
dealerL	2.682	2.256	3.108
dealerP	1.340	0.775	1.905
dealerR	1.869	1.499	2.239
Interm1	-1.239	-2.185	-0.293
Surface	0.001	0.000	0.001

	Coefficient	2.5%	97.5%
paired1	0.199	-0.179	0.577
finished1	1.165	0.729	1.601
discauth1	0.929	0.161	1.696
diff_origin1	-0.552	-0.762	-0.343
portrait1	-0.932	-1.335	-0.529
artistliving1	-0.114	-1.755	1.527
endbuyerC	-0.331	-1.064	0.402
endbuyerD	-0.333	-1.058	0.392
endbuyerE	-1.105	-1.856	-0.355
endbuyerU	-0.831	-1.578	-0.085
endbuyerX	-1.574	-2.315	-0.832
authorstyle1	-1.775	-2.602	-0.947
lrgfont1	1.784	1.438	2.130
still_life1	-0.130	-0.608	0.348
prevcoll1	1.526	1.191	1.860
dealerL:Interm1	1.507	-0.165	3.179
dealerP:Interm1	1.861	-0.779	4.501
dealerR:Interm1	2.507	1.536	3.477
dealerL:Surface	-0.001	-0.001	0.000
dealerP:Surface	0.000	-0.001	0.000
dealerR:Surface	-0.001	-0.001	0.000
dealerL:paired1	-1.221	-1.723	-0.720
dealerP:paired1	-0.673	-1.373	0.026
dealerR:paired1	-0.306	-0.720	0.108
dealerL:finished1	-0.642	-1.620	0.336
dealerP:finished1	-0.697	-1.498	0.104
dealerR:finished1	-0.591	-1.066	-0.116
dealerL:discauth1	0.037	-1.768	1.843
dealerP:discauth1	-1.719	-3.580	0.142
dealerR:discauth1	-0.886	-1.596	-0.175
Surface:diff_origin1	0.000	0.000	0.000
diff_origin1:portrait1	0.897	0.093	1.701
artistliving1:endbuyerC	1.030	-0.559	2.620
artistliving1:endbuyerD	1.202	-0.371	2.775
artistliving1:endbuyerE	2.127	0.440	3.815
artistliving1:endbuyerU	1.242	-0.381	2.864
artistliving1:endbuyerX	1.421	-0.193	3.035
artistliving1:authorstyle1	1.456	-0.301	3.212
Interm1:Surface	0.000	0.000	0.001
Interm1:lrgfont1	-0.904	-1.433	-0.376
Surface:lrgfont1	0.000	0.000	0.000
Surface:still_life1	-0.001	-0.001	0.000
Surface:discauth1	0.000	-0.001	0.000
finished1:prevcoll1	-1.075	-1.713	-0.437
paired1:lrgfont1	-0.764	-1.263	-0.265
paired1:discauth1	-0.428	-1.102	0.246
diff_origin1:authorstyle1	0.963	0.078	1.849
diff_origin1:still_life1	-0.732	-1.416	-0.048
finished1:discauth1	0.575	-0.033	1.183
discauth1:lrgfont1	-0.549	-1.405	0.308
finished1:artistliving1	-0.436	-0.993	0.121
Interm1:portrait1	-0.851	-2.000	0.297

	Coefficient	2.5%	97.5%
dealerL:artistliving1	-1.022	-1.726	-0.318
dealerP:artistliving1	-0.671	-1.524	0.182
dealerR:artistliving1	-0.519	-1.104	0.065
authorstyle1:prevcoll1	-1.696	-4.187	0.795

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 3: 95% Confidence Interval

fit	lwr	upr
70.489	5.379	923.723

Table 4: 95% Prediction Interval

fit	lwr	upr
70.489	2.183	2276.12

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:

- dealer: the type of dealer that the auction went through significantly affects the price of the painting. For example, compared with dealer J, the average selling price from dealer L is `268.2% 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 `123.9% 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-2%` livres.
- finished: if the painting is finished, the selling price on average is `116.5% higher` than when the painting is not finished.
- portrait: if the painting is described as a portrait, the selling price on average is `93.2% lower` times lower than when the painting is not described as a portrait.

- endbuyer: the type of endbuyer will significantly affect the level of price. For instance, compared with the endbuyer type B, the average selling price is `110.5% lower` when the endbuyer is type E.
- prevcoll: when the previous owner is mentioned, the average selling price is `152.6% higher` than when the previous owner is not mentioned.
- lrgfont: when the dealer devotes an additional paragraph, the average selling price is `178.4% higher` than when there is no additional paragraph.
- authorstyle: when the author's name is introduced, the average selling price is expected to be `177.5% lower` than when the author's name is not introduced.
- dealer&Interm interaction: when an intermediary is present, which the price of the auctioned painitngs differs significantly among different dealers. For instance, if the dealer is R and with an intermediary existed, the average selling price is `250.7% higher` than when the dealer is J with an intermediary.
- dealer&discauth: when the dealer engaged with authenticity of the painting, which type of dealer involves the auction has significant difference. For instance, if dealer R engages with authenticity, the average selling is `88.57 % lower` than the price when dealer J engages with authenticity.
- finihsed*prevcoll: given that the painting is finished, when the previous owner is mentioned, the average price is expected to be `107.5% lower` than when the previous owner is not mentioned.

Recommendations

In order to understand the auction prices of 18th century paintings and predict prices of paintings with certain features, we recommend historians focusing on the characteristics mentioned above associated with the painitings (just to mention some, not a complete list). For example, in order to find out highly priced pieces, 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).