Part I Write-up

Jingyi Zhang, Jonathan Klus, Bin Han

1. Introduction:

In this study, the auction prices of paintings in 18th century Paris were examined. Specifically, we wish to understand the variables which affect the prices of the paintings, and then be able to predict auction prices based on characteristics of a certain painting. By fitting an appropriate model, we will also be creating a tool to help decide whether specific paintings that are either underpriced or overpriced given their realization of the covariates that were included in the model.

One of the main challenges in building this model is to narrow down the number of covariates from the 59 canadidates in the original data set to less than 20 in the final model. This must be done in such a way that an undue amount of bias is not introduced, and overfitting is avoided. The ability to explain the results and provide some recommendations to indivisuals without statistical background is equally important and challenging, since the primary audience for this analysis is intended to be art historians. The goal was therefore to balance predictive performance, model simplicity, and interprebility in order to create a pricing model for artwork in 18th century France.

2. Exploratory data analysis:

A) Data summary & cleaning

The training set consists of a few numeric variables and many categorical variables. Some variables, such as `Interm`, `Surface`, `Height_in` etc. have mising values, which need to be imputed before any analysis can occur. Data cleaning proceeded using the following steps:

```
## 'data.frame':    1500 obs. of  9 variables:
##  $ lot           : chr  "11" "28" "104" "12" ...
##  $ sale          : chr  "L1778b" "J1777b" "R1764" "R1767" ...
##  $ price         : chr  "122.0" "25.0" "14.0" "401.0" ...
##  $ count         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ subject       : chr  "Couches de Sainte Barbe" "(2) Vues de ports de mer d'Italie" "Un riche Paysa
##  $ authorstandard: chr  "Bassano (Jacopo da Ponte), Jacopo" "Lacroix de Marseille, Charles-Fran\u008
##  $ author        : chr  "Jacques Bassan" "N. la Croix" "Francisque Mil\u008e" "Pierre Fran\u008dois 
##  $ winningbidder : chr  "Unknown" "Quenet" "Vaurio" "Renouard, abb\u008e" ...
##  $ other         : int  0 0 0 0 1 0 0 0 0 0 ...
```

```
##
##     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
```

a. In order to reduce the dimensionality of the problem, variables that were deemed intuitively not useful or colinear with other covariates were removed. This included: `lot`, `sale`, `price`, `count`, `subject`, `authorstandard`, `author`, `winningbidder`, and `other`. From the and structure and 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 their present form for predicting the response variable. While its possible that some variables like `subjects` could be recoded using some underlying characteristic (and given some art expertise), we do not attempt to do this here.

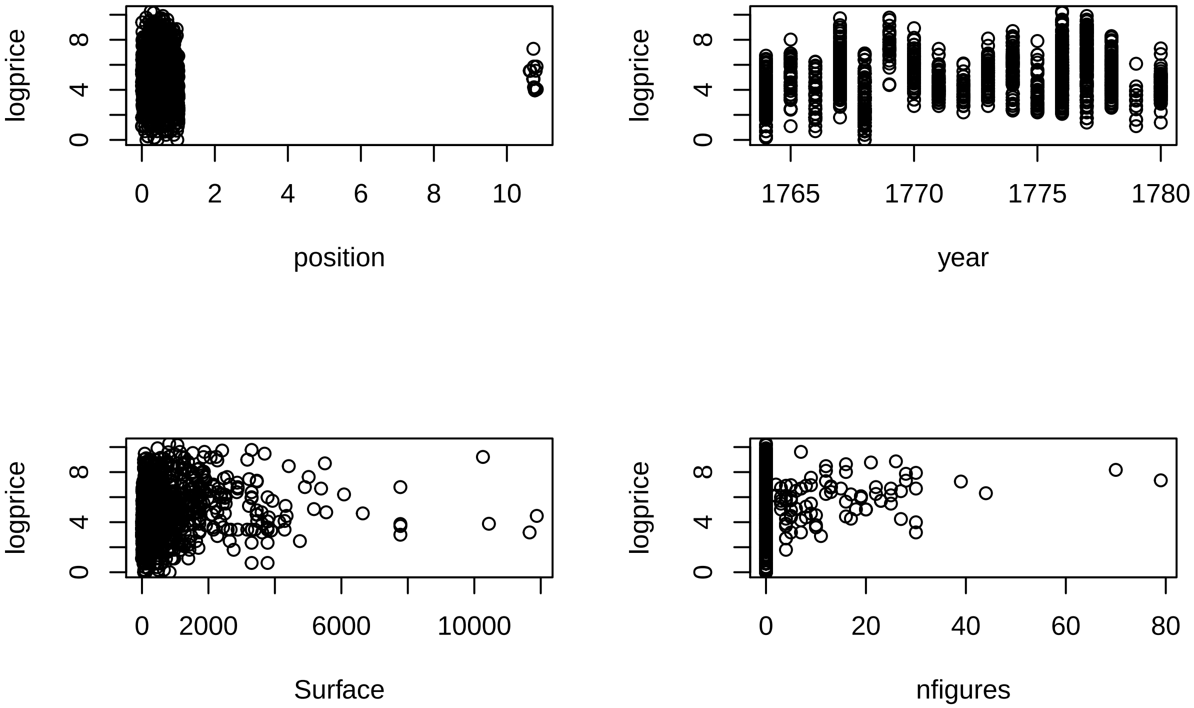
b. It was determined that `Surface` and `Surface_Rnd`, `Surface_Rect` are similar, based on the value of `Height_in`, `Width_in`, and `Diam_in`. We decided to use `Surface` in our initial model as it contained all information in the latter two surface area variables. A similar approach was used for variables `material`, `mat`, and `materialCat`. The latter one recodes the previous one, with fewer levels for simplicity (39 levels in the former versus just 5 in the latter). Therefore, we used `materialCat`. We applied the same strategy to keep `landsALL` and remove other `lands` indicator variables which contained little information and would therefore have been difficult to estimate an accurate coefficient.

c. This data contained a great deal of structurally missing values (i.e. missingness resulting from how the researchers coded the data, rather than truly unavailable or omitted information). 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 either "other" or "no information" in the code book, depending upon the variable in question. For `materialCat` and `Shape`, since there are so many levels, we grouped some levels with few (<10) observations together, coded as the "other" group. The remaining binary vairables were converted into factors.

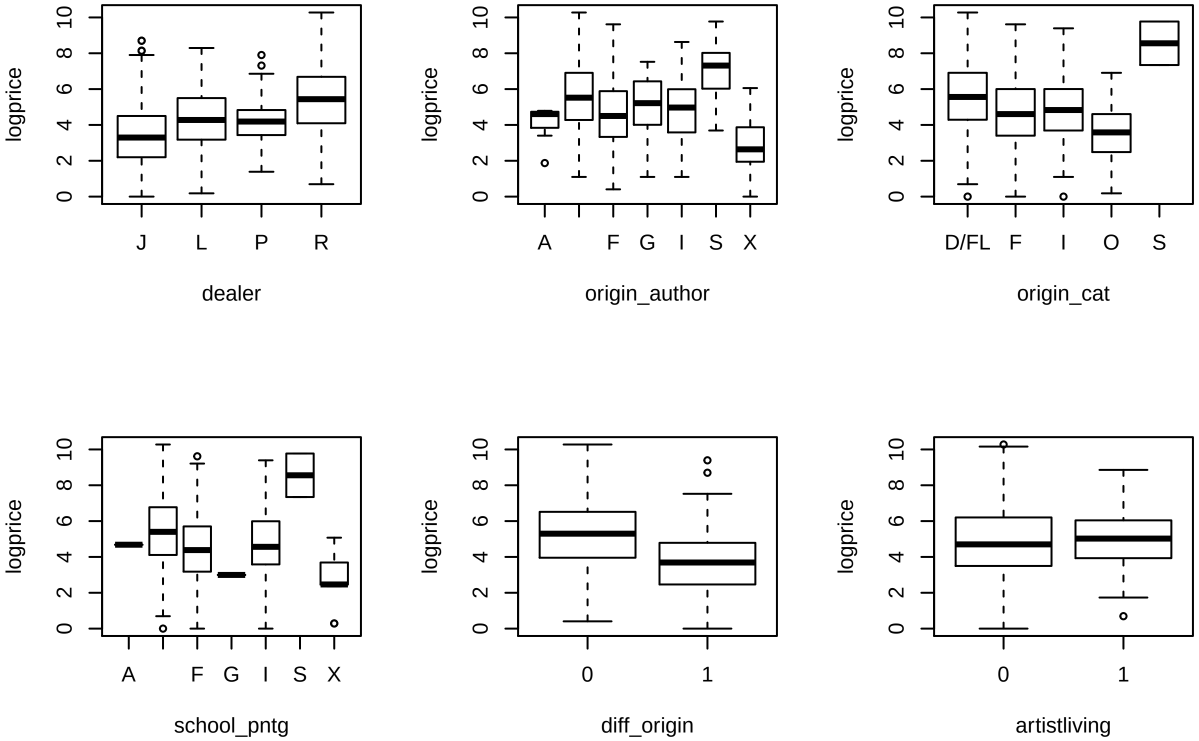
d. The remaining data issue was how to deal with missing values in the numeric continuous variable `Surface` and the binary variable `Interm`. The `mice` package (Multivariate Imputation by Chained Equations) was used to address this problem. It uses the observed values of other covariates in the dataset to create a model to impute the missing values. This method is superior to complete case analysis, which would result in losing an unacceptably large amount of data, as well as simpler imputation methods (i.e. imputing the mean of a given covariate to replace missing values).

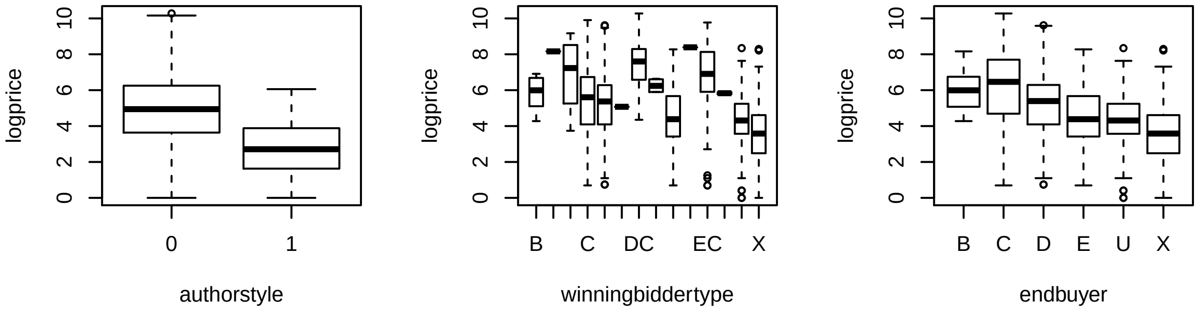
B). Plots

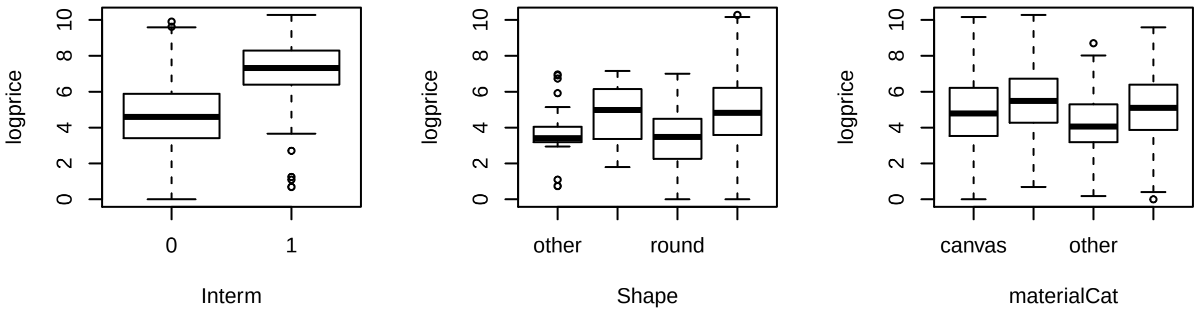
Following the data cleaning and dimension reduction described above, the potential covariates were plotted against the response variable, `logprice`. Scatter plots were used for numeric variables, and allow us to roughly determine which variables may be useful in the initial model. For categorical variables, we use boxplots to check if the range of `logprice` is observably different for each level of the variable. If so, it may be a good predictor. For numeric variables, we want to check if there is a clear linear relationship between the variable and `logprice`.



For numeric variables, we note that `Surface` and `nfigures` appear to have a weak but positive relationship with `logprice`. Since there are several extremely large values in `position` (potential outliers), it is difficult to know if there is a truly useful relationship here between the majority of points and `logprice`. But we will keep it in the initial model for now.







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 variables for prediction at this point, we would consider the magnitude of differences and the strength of relationships. The 10 best variables based upon the above EDA are: `Surface`, `dealer`, `school_pntg`, `diff_origin`, `authorstyle`, `endbuyer`, `Interm`, `prevcoll`, `engraved`, `Irgfont`. They exhibit the strongest relationship with the response variable, `logprice`.

3. Development and assessment of an initial model

Initial Model

JZS prior

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

g-prior

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

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

An OLS regression model was then fit using the chosen features and all their possible interactions. From the summary table, the $R^2 = 0.5828$, was okay, with approximately 58% of the variation in `logprice` explained by the model. But this model suffered from several potential issues, not least among which was that many coefficients could not be estimated (they returned NAs). Despite attempts during the first part of EDA at dimension reduction and removing colinear variables, it appears that multicolinearity is still very clearly an issue, as our design matrix is not full rank. Therefore, we need to again reduce the dimensionality of the model through further 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 
## -4.3981 -0.7703  0.0248  0.8069  4.8291 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      4.369e+00  1.340e+00   3.260 0.001138 ** 
## dealerL          1.566e+00  1.346e-01  11.635  < 2e-16 ***
## dealerP          5.124e-01  1.655e-01   3.097 0.001991 ** 
## dealerR          1.192e+00  1.082e-01  11.020  < 2e-16 ***
## school_pntgD/FL -6.425e-01  1.262e+00  -0.509 0.610656    
## school_pntgF    -1.373e+00  1.263e+00  -1.087 0.277033    
## school_pntgG    -4.542e+00  1.795e+00  -2.530 0.011497 *  
## school_pntgI    -1.301e+00  1.265e+00  -1.028 0.303939    
## school_pntgS    -3.067e-01  1.552e+00  -0.198 0.843336    
## school_pntgX    -2.039e+00  1.276e+00  -1.598 0.110181    
## diff_origin1    -6.407e-01  9.160e-02  -6.995 4.02e-12 ***
## artistliving1    6.626e-01  1.051e-01   6.304 3.82e-10 ***
## endbuyerC       -1.351e-01  3.449e-01  -0.392 0.695236    
## endbuyerD       -1.545e-01  3.420e-01  -0.452 0.651479    
## endbuyerE       -8.036e-01  3.559e-01  -2.258 0.024096 *  
## endbuyerU       -6.286e-01  3.516e-01  -1.788 0.074017 .  
## endbuyerX       -1.310e+00  3.494e-01  -3.748 0.000185 ***
## authorstyle1    -1.035e+00  1.578e-01  -6.556 7.62e-11 ***
## Interm1          1.001e+00  1.404e-01   7.130 1.57e-12 ***
## Shapeoval        7.222e-01  3.886e-01   1.858 0.063311 .
## Shaperound      -1.047e-01  3.542e-01  -0.295 0.767668
## Shapesqu_rect    8.713e-01  2.588e-01   3.367 0.000781 ***
## Surface          2.022e-04  3.258e-05   6.206 7.08e-10 ***
## engraved1        3.188e-01  1.518e-01   2.100 0.035903 *
## prevcoll1        1.221e+00  1.505e-01   8.111 1.05e-15 ***
## paired1         -3.347e-01  7.107e-02  -4.710 2.71e-06 ***
## finished1        6.340e-01  9.689e-02   6.544 8.27e-11 ***
## lrgfont1         1.095e+00  1.247e-01   8.780  < 2e-16 ***
## portrait1       -6.742e-01  1.766e-01  -3.817 0.000141 ***
## discauth1        3.720e-01  1.452e-01   2.562 0.010495 *
## still_life1     -7.492e-01  1.716e-01  -4.366 1.35e-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:  0.5836, Adjusted R-squared:  0.5751
## F-statistic: 68.62 on 30 and 1469 DF,  p-value: < 2.2e-16
```

Model Selection

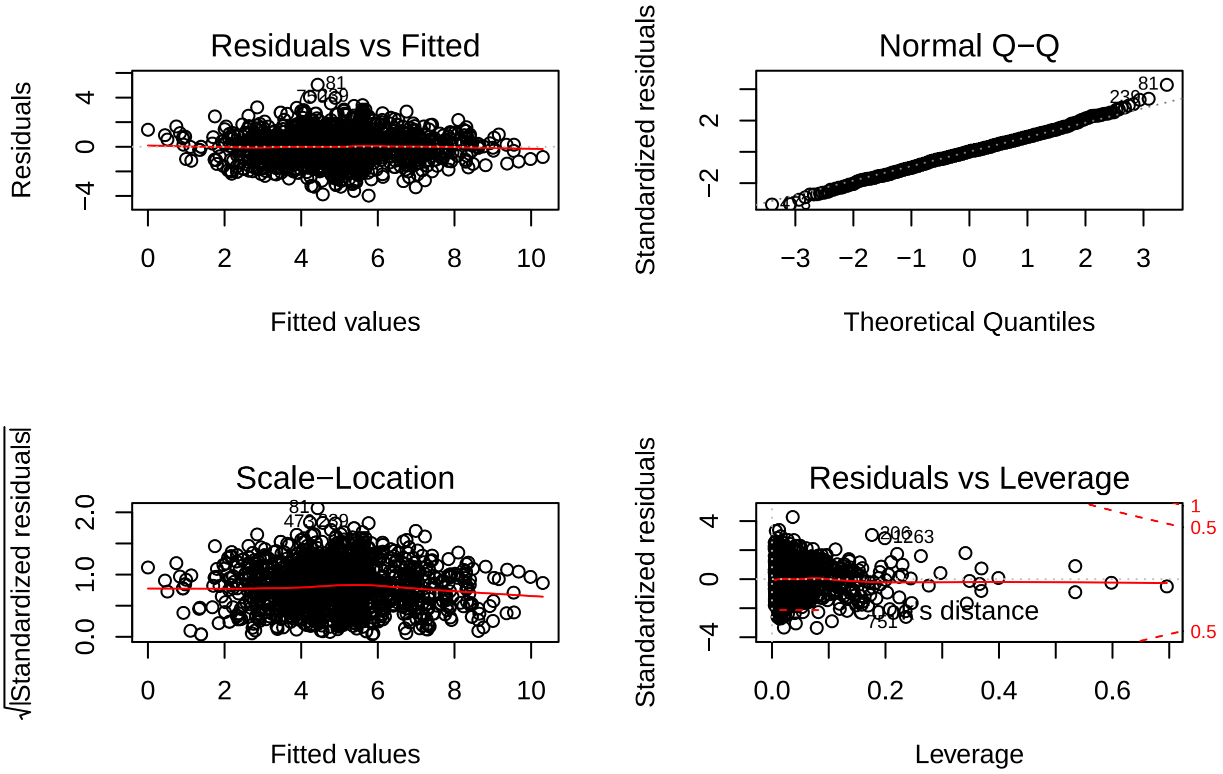
After completing the initial exploratory data analysis, methods including Stepwise Selection using both AIC and BIC penalties were used in order to assess more systematically which covariates and interactions were most important for predicting the `logprice` of paintings. While the number of relevant 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 by reducing the number of covariates included in the final model and to deliver a model that was both interpretable and performed well in prediction. The results of the two methods were then compared, and interactions that were not intuitive were filtered out (e.g. $artistling * endbuyer$).

Ultimately, the following variables were selected using the above methods and were fit using OLS regression. The resulting $Adj - R^2$ was 0.6079. All the included covariates had estimable coefficients (i.e. there were no NAs, as the resulting design matrix was full rank).

```
##
## Call:
## lm(formula = logprice ~ Shape + school_pntg + dealer * Interm +
##     dealer * Surface + dealer * paired + dealer * finished +
##     diff_origin * Surface + diff_origin * portrait + artistliving *
##     endbuyer + 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
## -3.9679 -0.7183  0.0264  0.7532  5.0430
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                4.048e+00  1.326e+00   3.052 0.002312 **
## Shapeoval                  8.865e-01  3.932e-01   2.255 0.024305 *
## Shaperound                -1.410e-01  3.609e-01  -0.391 0.695978
## Shapesqu_rect              8.996e-01  2.739e-01   3.285 0.001045 **
## school_pntgD/FL           -7.452e-01  1.233e+00  -0.604 0.545642
## school_pntgF              -1.423e+00  1.234e+00  -1.153 0.249038
## school_pntgG              -4.071e+00  1.825e+00  -2.231 0.025835 *
## school_pntgI              -1.366e+00  1.236e+00  -1.105 0.269526
## school_pntgS              -8.962e-01  1.528e+00  -0.586 0.557745
## school_pntgX              -1.920e+00  1.249e+00  -1.538 0.124298
## dealerL                    2.445e+00  2.140e-01  11.422  < 2e-16 ***
## dealerP                    1.164e+00  2.631e-01   4.424 1.04e-05 ***
## dealerR                    1.678e+00  1.831e-01   9.167  < 2e-16 ***
## Interm1                   -9.900e-01  4.851e-01  -2.041 0.041465 *
## Surface                    4.502e-04  1.521e-04   2.960 0.003128 **
## paired1                    6.711e-02  1.930e-01   0.348 0.728059
## finished1                  1.303e+00  2.195e-01   5.937 3.64e-09 ***
## diff_origin1              -5.763e-01  1.055e-01  -5.462 5.53e-08 ***
## portrait1                 -1.014e+00  2.100e-01  -4.826 1.54e-06 ***
## artistliving1             -4.766e-02  8.435e-01  -0.057 0.954951
## endbuyerC                 -3.303e-01  3.772e-01  -0.876 0.381289
## endbuyerD                 -3.623e-01  3.731e-01  -0.971 0.331608
## endbuyerE                 -1.081e+00  3.859e-01  -2.802 0.005143 **
## endbuyerU                 -8.163e-01  3.836e-01  -2.128 0.033493 *
## endbuyerX                 -1.598e+00  3.813e-01  -4.192 2.94e-05 ***
## lrgfont1                   1.780e+00  1.772e-01  10.047  < 2e-16 ***
## still_life1               -2.883e-01  2.323e-01  -1.241 0.214768
## discauth1                  1.528e-01  2.466e-01   0.620 0.535642
## prevcoll1                  1.510e+00  1.719e-01   8.781  < 2e-16 ***
## authorstyle1              -2.004e+00  4.134e-01  -4.847 1.39e-06 ***
## dealerL:Interm1            1.305e+00  8.615e-01   1.515 0.130009
## dealerP:Interm1            1.788e+00  1.358e+00   1.316 0.188284
## dealerR:Interm1            2.250e+00  5.002e-01   4.498 7.43e-06 ***
## dealerL:Surface           -2.171e-04  1.730e-04  -1.255 0.209719
## dealerP:Surface           -8.196e-05  2.355e-04  -0.348 0.727839
## dealerR:Surface           -2.492e-04  1.556e-04  -1.602 0.109473
## dealerL:paired1           -1.066e+00  2.594e-01  -4.109 4.20e-05 ***
## dealerP:paired1           -6.404e-01  3.608e-01  -1.775 0.076167 .
## dealerR:paired1           -1.592e-01  2.111e-01  -0.754 0.450911
## dealerL:finished1         -7.555e-01  5.010e-01  -1.508 0.131757
## dealerP:finished1         -8.938e-01  4.050e-01  -2.207 0.027499 *
## dealerR:finished1         -7.601e-01  2.368e-01  -3.210 0.001356 **
## Surface:diff_origin1      -8.729e-05  8.437e-05  -1.035 0.300990
## diff_origin1:portrait1     9.893e-01  4.134e-01   2.393 0.016836 *
## artistliving1:endbuyerC    1.067e+00  8.176e-01   1.305 0.192079
## artistliving1:endbuyerD    1.217e+00  8.091e-01   1.504 0.132711
## artistliving1:endbuyerE    2.179e+00  8.667e-01   2.515 0.012024 *
## artistliving1:endbuyerU    1.238e+00  8.345e-01   1.484 0.138030
## artistliving1:endbuyerX    1.454e+00  8.294e-01   1.753 0.079887 .
## Interm1:Surface            3.692e-04  1.514e-04   2.438 0.014872 *
## Interm1:lrgfont1          -8.494e-01  2.701e-01  -3.144 0.001699 **
## Surface:lrgfont1          -1.324e-04  1.134e-04  -1.167 0.243239
## Surface:still_life1       -3.154e-04  2.009e-04  -1.570 0.116745
## Surface:discauth1          1.318e-04  1.827e-04   0.721 0.470843
## finished1:prevcoll1       -1.104e+00  3.255e-01  -3.391 0.000714 ***
## paired1:lrgfont1          -7.937e-01  2.561e-01  -3.099 0.001980 **
## paired1:discauth1         -2.613e-01  3.336e-01  -0.783 0.433603
## diff_origin1:authorstyle1  1.291e+00  4.436e-01   2.909 0.003679 **
## diff_origin1:still_life1  -6.962e-01  3.587e-01  -1.941 0.052468 .
## finished1:discauth1        7.969e-01  2.923e-01   2.726 0.006486 **
## lrgfont1:discauth1        -1.054e+00  3.900e-01  -2.704 0.006940 **
## finished1:artistliving1   -4.516e-01  2.866e-01  -1.576 0.115314
## Interm1:portrait1         -7.136e-01  5.912e-01  -1.207 0.227618
## dealerL:artistliving1     -9.829e-01  3.563e-01  -2.758 0.005886 **
## dealerP:artistliving1     -8.747e-01  4.529e-01  -1.931 0.053652 .
## dealerR:artistliving1     -5.817e-01  3.001e-01  -1.939 0.052754 .
## prevcoll1:authorstyle1    -1.658e+00  1.267e+00  -1.309 0.190672
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.201 on 1433 degrees of freedom
## Multiple R-squared:  0.6252, Adjusted R-squared:  0.6079
## F-statistic: 36.22 on 66 and 1433 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 appear acceptable, with no obvious outliers or highly influential points shown. The model also does not violate the normality assumption for residuals. The constant variance of residuals assumption appears to be satisfied, and there is no fanning or other obvious pattern in this plot. While there are a two points that are identified as outliers, they were not found to be influential.

Variables

	Coefficient	2.5%	97.5%
(Intercept)	4.048	1.446	6.649
Shapeoval	0.886	0.115	1.658
Shaperound	-0.141	-0.849	0.567
Shapesqu_rect	0.900	0.362	1.437
school_pntgD/FL	-0.745	-3.164	1.673
school_pntgF	-1.423	-3.844	0.998
school_pntgG	-4.071	-7.651	-0.492
school_pntgI	-1.366	-3.791	1.060
school_pntgS	-0.896	-3.894	2.102
school_pntgX	-1.920	-4.370	0.529
dealerL	2.445	2.025	2.865
dealerP	1.164	0.648	1.680
dealerR	1.678	1.319	2.038
Interm1	-0.990	-1.942	-0.038
Surface	0.000	0.000	0.001
paired1	0.067	-0.311	0.446

	Coefficient	2.5%	97.5%
finished1	1.303	0.872	1.733
diff_origin1	-0.576	-0.783	-0.369
portrait1	-1.014	-1.425	-0.602
artistliving1	-0.048	-1.702	1.607
endbuyerC	-0.330	-1.070	0.410
endbuyerD	-0.362	-1.094	0.369
endbuyerE	-1.081	-1.838	-0.324
endbuyerU	-0.816	-1.569	-0.064
endbuyerX	-1.598	-2.346	-0.850
lrgfont1	1.780	1.432	2.128
still_life1	-0.288	-0.744	0.167
discauth1	0.153	-0.331	0.636
prevcoll1	1.510	1.172	1.847
authorstyle1	-2.004	-2.815	-1.193
dealerL:Interm1	1.305	-0.385	2.995
dealerP:Interm1	1.788	-0.877	4.452
dealerR:Interm1	2.250	1.269	3.231
dealerL:Surface	0.000	-0.001	0.000
dealerP:Surface	0.000	-0.001	0.000
dealerR:Surface	0.000	-0.001	0.000
dealerL:paired1	-1.066	-1.575	-0.557
dealerP:paired1	-0.640	-1.348	0.067
dealerR:paired1	-0.159	-0.573	0.255
dealerL:finished1	-0.756	-1.738	0.227
dealerP:finished1	-0.894	-1.688	-0.099
dealerR:finished1	-0.760	-1.225	-0.296
Surface:diff_origin1	0.000	0.000	0.000
diff_origin1:portrait1	0.989	0.178	1.800
artistliving1:endbuyerC	1.067	-0.537	2.671
artistliving1:endbuyerD	1.217	-0.370	2.804
artistliving1:endbuyerE	2.179	0.479	3.880
artistliving1:endbuyerU	1.238	-0.399	2.875
artistliving1:endbuyerX	1.454	-0.173	3.081
Interm1:Surface	0.000	0.000	0.001
Interm1:lrgfont1	-0.849	-1.379	-0.320
Surface:lrgfont1	0.000	0.000	0.000
Surface:still_life1	0.000	-0.001	0.000
Surface:discauth1	0.000	0.000	0.000
finished1:prevcoll1	-1.104	-1.743	-0.465
paired1:lrgfont1	-0.794	-1.296	-0.291
paired1:discauth1	-0.261	-0.916	0.393
diff_origin1:authorstyle1	1.291	0.420	2.161
diff_origin1:still_life1	-0.696	-1.400	0.007
finished1:discauth1	0.797	0.223	1.370
lrgfont1:discauth1	-1.054	-1.820	-0.289
finished1:artistliving1	-0.452	-1.014	0.111
Interm1:portrait1	-0.714	-1.873	0.446
dealerL:artistliving1	-0.983	-1.682	-0.284
dealerP:artistliving1	-0.875	-1.763	0.014
dealerR:artistliving1	-0.582	-1.170	0.007
prevcoll1:authorstyle1	-1.658	-4.143	0.826

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 model is interpreted in terms of exponentiated values below.

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
64.673	4.825	866.916

Table 3: 95% Prediction Interval

fit	lwr	upr
64.673	1.942	2153.491

Interpretation

From the final model, 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. The most important covariates and interactions are interpreted as follows:

- 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 price from dealer L is 244.5% higher. (Same interpretation for dealer P and R, with different coefficients)
- Interm: when there is an intermediary involved in the transaction, on average, the selling price is 99% lower than when there is no intermediary involved.
- Surface: for every one square inch increase in the painting surface, the selling price is, on average expected to increase .045%.
- finished: if the painting is noted for being highly finished, the selling price on average is 130.3% higher than when the painting is not noted for being highly finished.
- portrait: if the painting is described as a portrait, the selling price on average is 101.4% 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 (buyer), the average selling price is -108.1% lower when the endbuyer is type E (expert).
- prevcoll: when the previous owner is mentioned, the average selling price is 151.0% higher than when the previous owner is not mentioned.
- lrgfont: when the dealer devotes an additional paragraph, the average selling price is 178.0% higher than when there is no additional paragraph.

- authorstyle: when the author's name is introduced, the average selling price is expected to be `200.4% lower` than when the author's name is not introduced.
- dealer&Interm interaction: when an intermediary is present, which the price of the auctioned paintings differs significantly among different dealers. For instance, if the dealer is R and an intermediary is used, the average selling price is `225.0% higher` than when the dealer is J with an intermediary.
- finished*prevcoll: given that the painting is noted for being highly finished, when the previous owner is mentioned, the average price is expected to be `110.4% 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 transactions that involved 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. These features were among those that conferred the greatest percent increase in price over the base case in the model presented above.

Limitations

As mentioned in the data cleaning process, some of the variables have so many levels that fitting (and interpreting) such a model would be cumbersome. Many levels of the categorical variables included in this data set have few observations that are not sufficient for estimating coefficients. Therefore, we grouped some of the levels of variables, and grouped some variables via dimension reduction. This reduces the granularity of the analysis presented above, and risks introducing some bias, as the true model would probably be more granular and include the actual realization of each categorical predictor, not just an artbitrary grouping.

As our goal is to find a balance between the prediction accuracy and interpretability, our model may not predict the logprice response variable as accurately as other more advanced methods (which sacrifice interpretability in exchange for improved predictions). In the next phase of this project, we will attempt to fit other models which do just this, then compare the results to our initial work using an OLS regression model.