Part-I-Writeup

Team 9

12/2/2019

Introduction

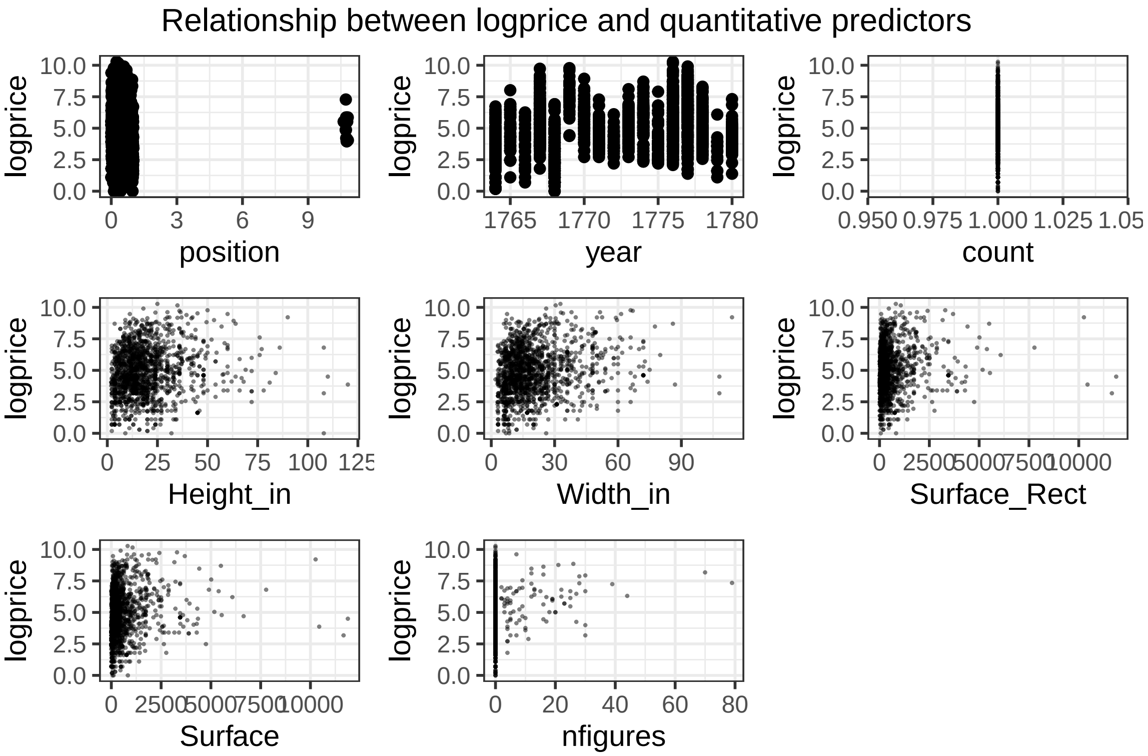
As the second largest city in Europe and the center of art in 18th century, Paris witnessed a lot of auctions on paintings. The most important aspect that all people in auctions would consider is the price. Then the question aries: What could drive the prices of paintings? This project aims to find out what factors could affect the price of paintings in 18th century Paris by using the dataset containing 1500 records with auction price data from 1764-1780 on the sales, painters and other characterisitcs of paintings.

Before we start analyzing, we first clean the dataset and impute missing values for several variables by assuming missing completely at random and using "Mice" package that would assign values of missing cells based on existing data. To conduct the analysis, we use the simple linear model. We take the logarithm of prices of each painting and set it as the response variable in the model. We apply exploratory data analysis by drawing scatterplots and boxplots to find out what variables might be related with the price and put them in the model as predictors. To find out our final model, we set our initial model to contain all selected predictors with all possible two-way interactions. Then we use BIC (Bayesian Information Criterion) to conduct model selection and build a parsimonious model with predictors and interactions that has good performance in terms of RMSE(root mean squared error) and coverage etc.. Predictors and interactions in this final model would give us information about which features or combinations of features would influence the price of paintings so that people could find the most valuable paintings.

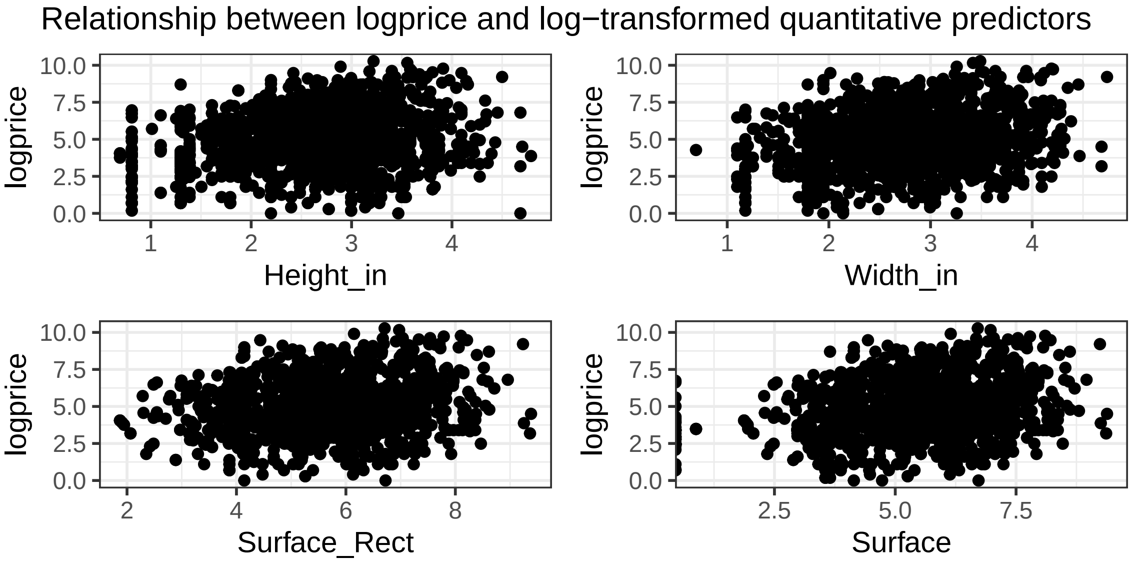
Exploratory Data Analysis

We first checked how many predictors in the original data have missing data. We noticed that `type_intermed`, `Diam_in`, `Surface_Rnd` and `authorstyle` are largely missing, and therefore excluded these predictors from model building. We noticed that many variables have missing or unknown values, and used `mice` to impute the missing values based on the values of the other variables. The following exploratory data analysis is based on the imputed training data set.

To identify the best variables for predicting `logprice`, we want to examine the relationships between `logprice` and the other variables. Since there are too many predictors, we look at quantitative, binary and qualitative predictors separately. We first create a scatterplot matrix for the quantitative predictors. We noticed that `position` is in percentage, but some of the paintings have values larger than 1, suggesting that these were recorded wrong. The above plot also shows that, `Height_in`, `Width_in`, `Surface_Rect`, and `Surface` need transformation, since their distributions are skewed.

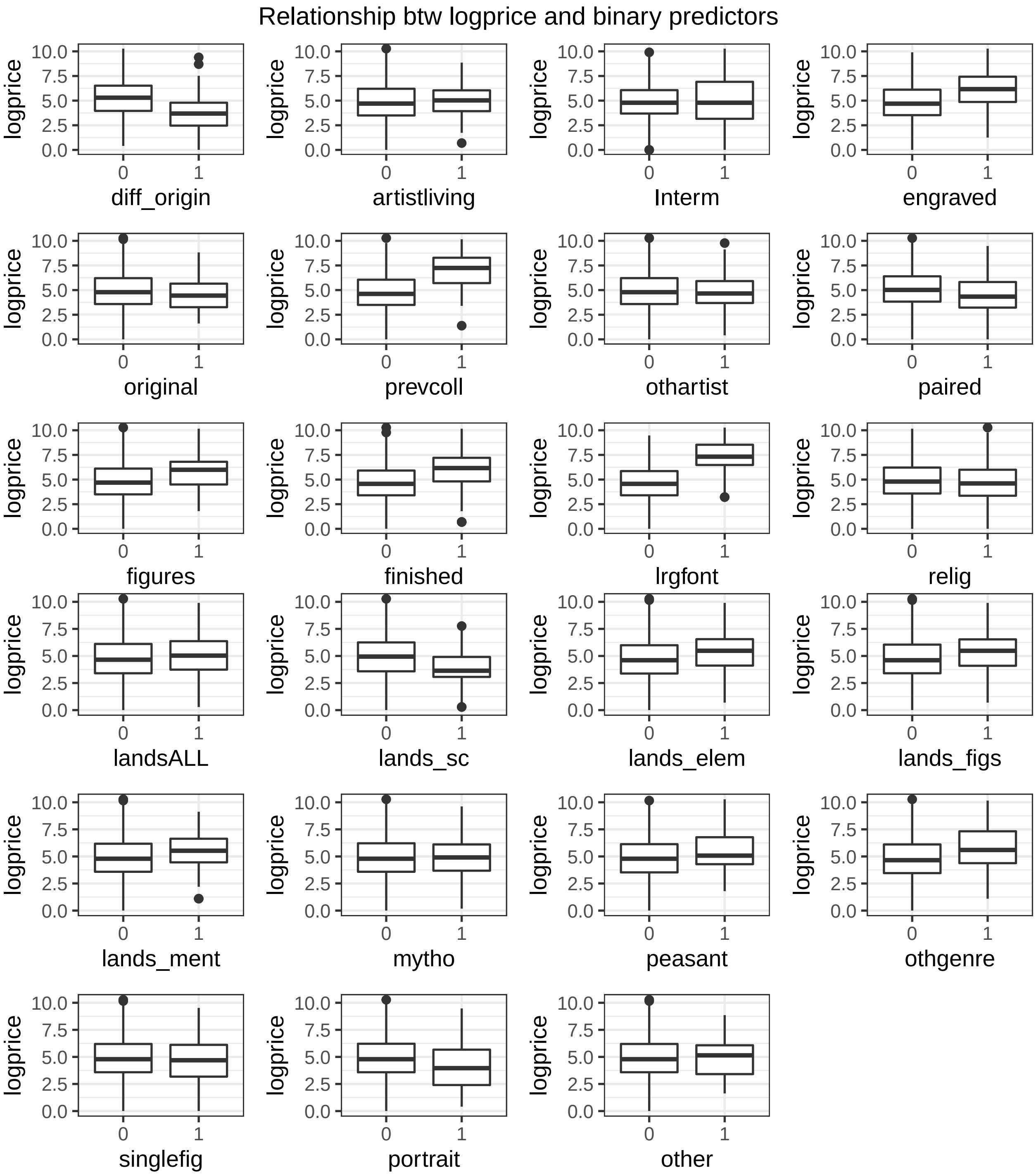


After log-transformation of these quantitative predictors, we obtained the following scatterplot matrix:

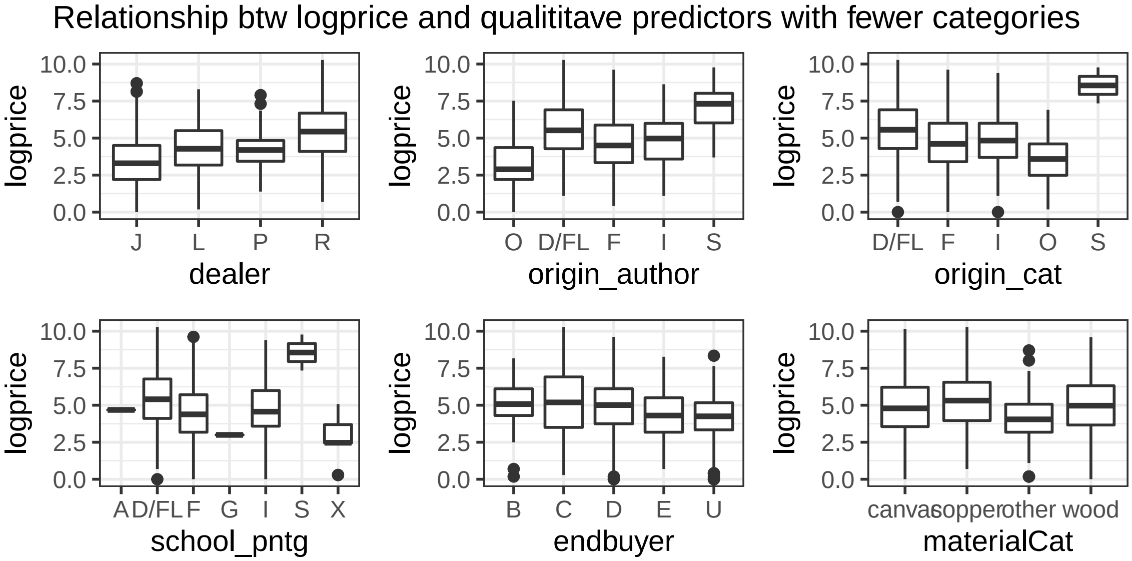


We decided to use `Surface` in our model, because we suspect that paintings with large surface area would be sold at a higher prices. We also wanted to use `year` as a predictor, since the plot shows that `logprice` fluctuates with `year`.

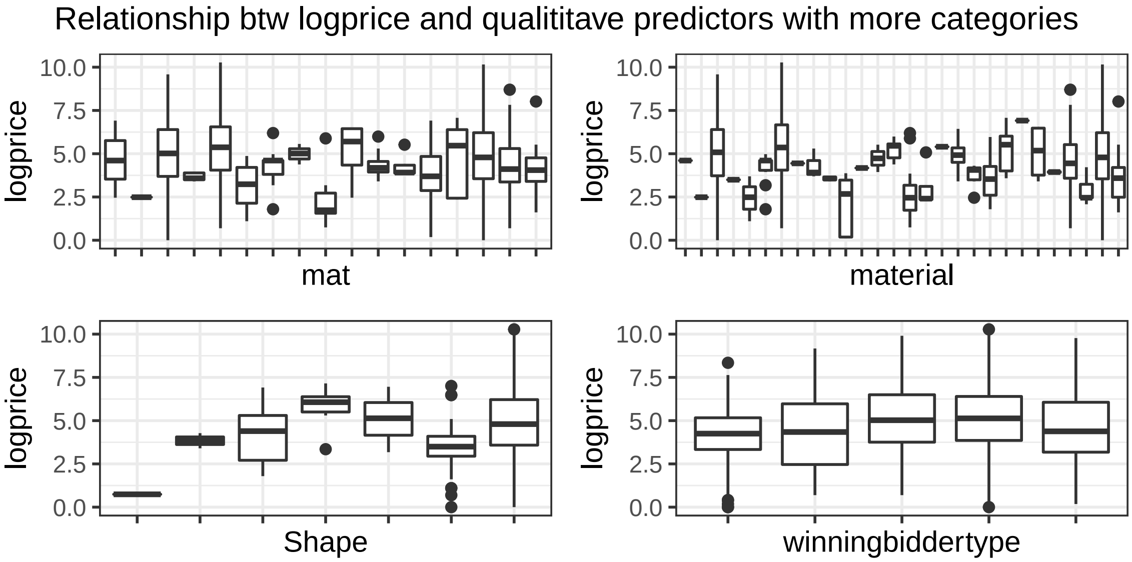
The following is a scatterplot matrix for the binary variables in the data set. The plot shows that we might be able to predict `logprice` based on the variables `diff_origin`, `engraved`, `prevcoll`, `finished`, `lrgfont` and `still_life`, since the difference in `logprice` is large for different levels of the binary variables:



The scatterplot matrix between `logprice` and qualitative variables with fewer categories suggest that we could consider `dealer`, `origin_author` to predict `logprice`.


Relationship btw logprice and qualititave predictors with fewer categories

We further plotted `logprice` v.s. the other qualitative predictors with more categories as shown below. Notice that some factor levels have too few observations, and thus these variables might not be helpful in predicting `logprice`.


Relationship btw logprice and qualititave predictors with more categories

Beyond what plotted above, there are some variables, such as `lot`, `authorstandard`, `winningbidder`, whose relationships to `logprice` are hard to visualize because they contain too many categories. We decided not to use them from our initial model building. Based on the EDA, we selected the 10 best variables for predicting `logprice`:

- Numerical: `Surface`, `year`
- Binary: `diff_origin`, `engraved`, `prevcoll`, `finished`, `lrgfont`, `still_life`
- Other qualitative: `dealer`, `origin_author`

Development and assessment of an initial model

Initial Model

We separated all predicotrs into three categories:numerical, binary and multi-level categorical variables. From the above EDA, we chose two numerical variables:`surface` and `year`, six binary variables: `diff-origin`, `prevcoll`, `engraved`, `lrgfont`, `finished` and `still-life`, and two multi-level categorical variables:`dealer` and `origin_author`.

For numerical variables, from the scatterplot we produced above, we noticed a fluctuated trend between year and logprice. This means that different `year` values would influence the log price of the painting. Although, the correlation between all numerical variables and `logprice` are small, we decided to include the variable `surface`. Since `surface` automatically includes the information from `Height_in`, `Width_in` and `Surface_Rect` and it intuitively makes senes that the size of the painting might effect the log price of the painting.

For categorical variables, we looked at their boxplot against the predictor `logprice` and selected according to their influence level. All those eight variables we selected indicited that the log price of the painting do differ for each levels.

Then, we used all those 10 variables and all possible two-way interactions to predict `logprice` as our initial model.

The summary of the model is the following:

```
## 
## Call:
## lm(formula = logprice ~ .^2, data = paintings_train_fin)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.9013 -0.7878  0.0000  0.7759  3.6515 
## 
## Coefficients: (12 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     -3.650e+02  8.126e+01  -4.491 7.66e-06 ***
## dealerL                          1.273e+02  7.204e+01   1.768 0.077348 .  
## dealerP                          2.701e+02  1.143e+02   2.363 0.018247 *  
## dealerR                          3.143e+01  4.947e+01   0.635 0.525287    
## diff_origin1                     9.109e+01  5.073e+01   1.795 0.072806 .  
## origin_authorD/FL                2.067e+02  6.099e+01   3.389 0.000721 ***
## origin_authorF                   1.240e+02  5.950e+01   2.084 0.037345 *  
## origin_authorI                   1.957e+02  6.911e+01   2.831 0.004706 ** 
## origin_authorS                   4.832e+02  7.506e+02   0.644 0.519884    
## engraved1                       -6.178e+01  7.812e+01  -0.791 0.429173    
## prevcoll1                        6.829e+00  8.046e+01   0.085 0.932370    
## finished1                        8.429e+01  4.666e+01   1.806 0.071057 .  
## lrgfont1                        -4.774e+00  5.690e+01  -0.084 0.933145    
## still_life1                      8.981e+01  8.831e+01   1.017 0.309360    
## Surface                         -2.033e+01  8.392e+00  -2.423 0.015515 *  
## year                             2.065e-01  4.590e-02   4.500 7.36e-06 ***
## dealerL:diff_origin1            -1.471e+00  7.719e-01  -1.906 0.056822 .  
## dealerP:diff_origin1            -1.450e+00  8.093e-01  -1.792 0.073364 .  
## dealerR:diff_origin1            -2.303e+00  6.835e-01  -3.369 0.000774 ***
## dealerL:origin_authorD/FL       -1.068e+00  7.890e-01  -1.353 0.176235    
## dealerP:origin_authorD/FL       -1.582e+00  9.496e-01  -1.666 0.096024 .  
## dealerR:origin_authorD/FL       -1.712e+00  7.384e-01  -2.319 0.020542 *  
## dealerL:origin_authorF          -7.370e-01  7.558e-01  -0.975 0.329688    
## dealerP:origin_authorF          -1.130e+00  9.255e-01  -1.221 0.222350    
## dealerR:origin_authorF          -1.309e+00  7.155e-01  -1.830 0.067525 .  
## dealerL:origin_authorI          -1.226e-01  8.245e-01  -0.149 0.881807    
## dealerP:origin_authorI          -9.588e-01  1.121e+00  -0.856 0.392408    
## dealerR:origin_authorI          -1.395e+00  7.236e-01  -1.929 0.053990 .  
## dealerL:origin_authorS                NA         NA      NA       NA
## dealerP:origin_authorS                NA         NA      NA       NA
## dealerR:origin_authorS                NA         NA      NA       NA
## dealerL:engraved1             -4.812e-01  1.049e+00  -0.459 0.646396
## dealerP:engraved1                     NA         NA      NA       NA
## dealerR:engraved1              2.887e-01  6.577e-01   0.439 0.660714
## dealerL:prevcoll1             -7.971e-02  8.025e-01  -0.099 0.920891
## dealerP:prevcoll1              9.717e-01  9.269e-01   1.048 0.294662
## dealerR:prevcoll1              1.004e-01  5.695e-01   0.176 0.860091
## dealerL:finished1              1.444e-01  5.259e-01   0.275 0.783624
## dealerP:finished1             -9.958e-02  4.369e-01  -0.228 0.819717
## dealerR:finished1             -4.128e-01  3.129e-01  -1.319 0.187268
## dealerL:lrgfont1                      NA         NA      NA       NA
## dealerP:lrgfont1                      NA         NA      NA       NA
## dealerR:lrgfont1                      NA         NA      NA       NA
## dealerL:still_life1           -6.384e-01  9.608e-01  -0.664 0.506547
## dealerP:still_life1           -1.846e+00  1.382e+00  -1.336 0.181765
## dealerR:still_life1           -4.607e-01  5.889e-01  -0.782 0.434199
## dealerL:Surface                9.577e-03  7.889e-02   0.121 0.903394
## dealerP:Surface               -1.447e-01  1.054e-01  -1.373 0.169906
## dealerR:Surface                2.270e-02  6.609e-02   0.344 0.731243
## dealerL:year                  -7.069e-02  4.051e-02  -1.745 0.081219 .
## dealerP:year                  -1.506e-01  6.439e-02  -2.339 0.019478 *
## dealerR:year                  -1.566e-02  2.785e-02  -0.562 0.574093
## diff_origin1:origin_authorD/FL  1.066e-01  2.675e+00   0.040 0.968228
## diff_origin1:origin_authorF     6.306e-01  2.676e+00   0.236 0.813721
## diff_origin1:origin_authorI     3.858e-01  2.694e+00   0.143 0.886136
## diff_origin1:origin_authorS            NA         NA      NA       NA
## diff_origin1:engraved1          4.363e-01  5.797e-01   0.753 0.451800
## diff_origin1:prevcoll1         -3.237e-01  8.007e-01  -0.404 0.686105
## diff_origin1:finished1          2.406e-01  5.512e-01   0.436 0.662572
## diff_origin1:lrgfont1          -9.874e-02  6.909e-01  -0.143 0.886380
## diff_origin1:still_life1       -1.368e+00  9.118e-01  -1.500 0.133821
## diff_origin1:Surface           -4.663e-02  6.198e-02  -0.752 0.451937
## diff_origin1:year              -5.075e-02  2.878e-02  -1.763 0.078116 .
## origin_authorD/FL:engraved1     6.518e-01  1.048e+00   0.622 0.534244
## origin_authorF:engraved1        4.768e-01  1.031e+00   0.463 0.643783
## origin_authorI:engraved1        5.386e-01  1.060e+00   0.508 0.611497
## origin_authorS:engraved1               NA         NA      NA       NA
## origin_authorD/FL:prevcoll1    -1.310e-01  5.145e-01  -0.255 0.799078
## origin_authorF:prevcoll1        1.901e-02  5.902e-01   0.032 0.974311
## origin_authorI:prevcoll1               NA         NA      NA       NA
## origin_authorS:prevcoll1       -7.359e-01  5.438e+00  -0.135 0.892373
## origin_authorD/FL:finished1    -9.379e-01  7.847e-01  -1.195 0.232171
## origin_authorF:finished1       -1.187e+00  7.787e-01  -1.524 0.127617
## origin_authorI:finished1       -9.704e-01  8.129e-01  -1.194 0.232761
## origin_authorS:finished1       -1.772e+00  1.550e+00  -1.143 0.253267
## origin_authorD/FL:lrgfont1     -6.136e-01  1.007e+00  -0.610 0.542240
## origin_authorF:lrgfont1        -9.684e-01  1.018e+00  -0.951 0.341674
## origin_authorI:lrgfont1        -9.215e-01  1.065e+00  -0.865 0.386959
## origin_authorS:lrgfont1         1.650e+00  3.105e+00   0.532 0.595114
## origin_authorD/FL:still_life1  3.058e-01  1.151e+00   0.266 0.790492
## origin_authorF:still_life1     -4.359e-01  1.097e+00  -0.397 0.691093
## origin_authorI:still_life1     -4.704e-01  1.277e+00  -0.368 0.712763
## origin_authorS:still_life1     -1.840e+00  3.045e+00  -0.604 0.545805
## origin_authorD/FL:Surface       1.504e-01  7.662e-02   1.964 0.049783 *
## origin_authorF:Surface          1.355e-01  6.461e-02   2.097 0.036179 *
## origin_authorI:Surface          9.626e-02  1.001e-01   0.962 0.336370
## origin_authorS:Surface         -1.258e-01  6.269e-01  -0.201 0.840983
## origin_authorD/FL:year         -1.161e-01  3.451e-02  -3.363 0.000791 ***
## origin_authorF:year            -6.990e-02  3.369e-02  -2.075 0.038189 *
## origin_authorI:year            -1.103e-01  3.905e-02  -2.823 0.004818 **
## origin_authorS:year            -2.722e-01  4.255e-01  -0.640 0.522540
## engraved1:prevcoll1             1.811e-01  5.905e-01   0.307 0.759085
## engraved1:finished1            -1.568e-01  4.260e-01  -0.368 0.712867
## engraved1:lrgfont1             -1.668e-01  4.708e-01  -0.354 0.723226
## engraved1:still_life1                  NA         NA      NA       NA
## engraved1:Surface               2.041e-01  1.386e-01   1.473 0.141002
## engraved1:year                  3.430e-02  4.406e-02   0.779 0.436391
## prevcoll1:finished1            -1.155e+00  4.208e-01  -2.744 0.006151 **
## prevcoll1:lrgfont1              8.060e-03  4.319e-01   0.019 0.985114
## prevcoll1:still_life1                  NA         NA      NA       NA
## prevcoll1:Surface              -7.704e-03  1.096e-01  -0.070 0.943952
## prevcoll1:year                 -3.208e-03  4.535e-02  -0.071 0.943605
## finished1:lrgfont1              2.965e-01  2.924e-01   1.014 0.310832
## finished1:still_life1          -7.613e-01  8.584e-01  -0.887 0.375337
## finished1:Surface              -1.796e-01  7.493e-02  -2.397 0.016657 *
## finished1:year                 -4.575e-02  2.624e-02  -1.744 0.081383 .
## lrgfont1:still_life1           -2.658e-01  9.717e-01  -0.273 0.784511
## lrgfont1:Surface                2.301e-01  1.048e-01   2.195 0.028297 *
## lrgfont1:year                   3.005e-03  3.209e-02   0.094 0.925408
## still_life1:Surface             7.609e-02  9.921e-02   0.767 0.443271
## still_life1:year               -5.058e-02  4.951e-02  -1.022 0.307176
## Surface:year                    1.154e-02  4.732e-03   2.439 0.014844 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.228 on 1400 degrees of freedom
## Multiple R-squared:  0.6172, Adjusted R-squared:  0.5901
## F-statistic:  22.8 on 99 and 1400 DF,  p-value: < 2.2e-16
```

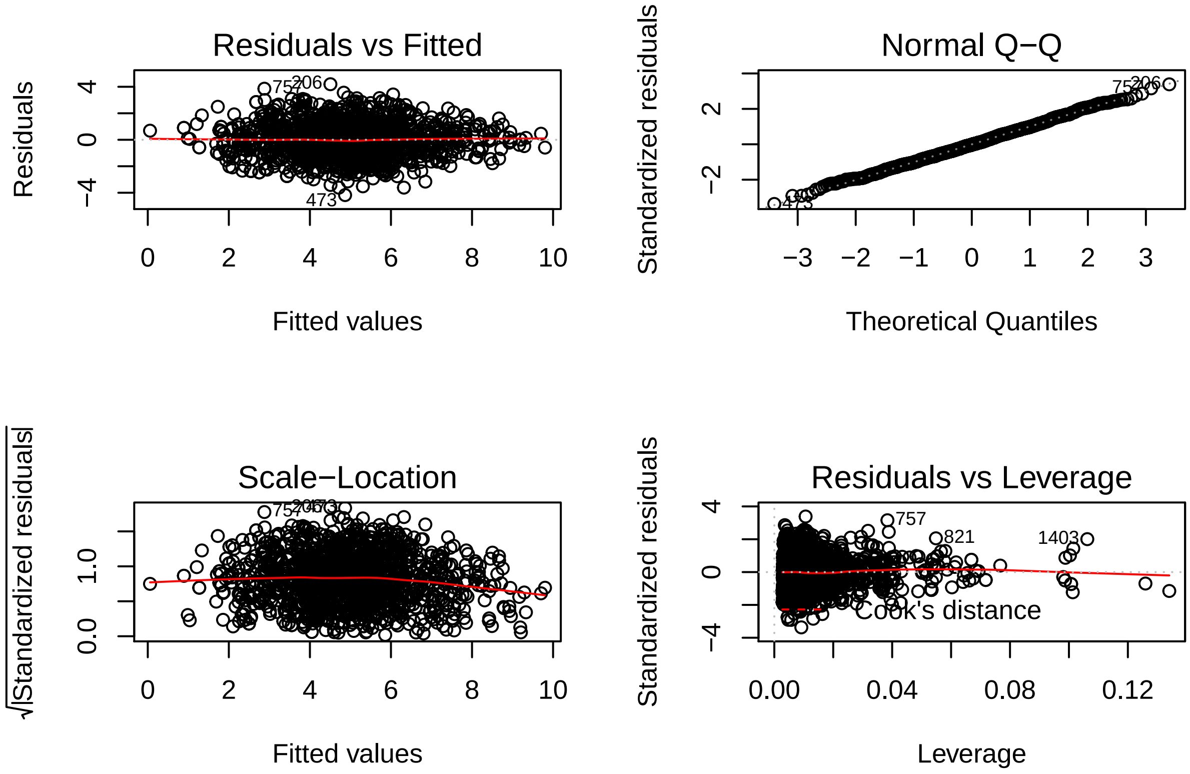
Since the R-squared is 0.6172, our initial model explained about 61% of the variation in the response variable around its mean.

Model Selection

By looking at the summary table for the initial model, a large proportion of the variables are not significant and some predictors showing NAs. Therefore, we need to operate model selection. For linear regression models, we have two criterias for model selection: Akaike Infromation Criterion (AIC) and the Bayes Information Criteria (BIC). Out of those two methods, we chose to use BIC as our selection criteria. The BIC criteria pick the model that has the highest posterior probability. Although, the model might not necessarily be the best predictive model, it is most likely to be true given the data.

As a result, the model selected by the BIC kept all 10 variables and added 4 two-way interactions.

Residual plots



From the above four plots, we can see that our model selected by the BIC criteria performed well. From the residuals plot on the top-left, all residuals are equally spread out indicates linearity. Also, from the normal Q-Q plot we can see that all residuals are perfectly followed a straight line indicates normality. The scale-location plot shows that the residuals are spread equally along the ranges of predicotrs indicated equal variance. Lastly, there are no actually influential points but just a few potential outliers.

Summary

Table 1: Coefficients and 95% Confidence Intervals (Relative Risk

	Coefficients	2.5 %	97.5 %
(Intercept)	0.00	0.00	0.00
dealerL	2.66	1.94	3.64
dealerP	1.36	0.94	1.98
dealerR	9.19	7.30	11.56
diff_origin1	2.89	1.48	5.61
origin_authorD/FL	2.44	1.80	3.29
origin_authorF	1.29	0.97	1.73
origin_authorI	1.20	0.85	1.68
origin_authorS	3.00	1.35	6.69
engraved1	2.36	1.75	3.18
prevcoll1	3.16	2.25	4.44
finished1	2.80	2.31	3.39
lrgfont1	3.55	2.80	4.51
still_life1	0.87	0.58	1.31
Surface	1.32	1.25	1.39
year	1.15	1.13	1.17
dealerL:diff_origin1	1.21	0.69	2.10
dealerP:diff_origin1	0.59	0.29	1.19

	Coefficients	2.5 %	97.5 %
dealerR:diff_origin1	0.35	0.22	0.58
diff_origin1:still_life1	0.29	0.14	0.59
diff_origin1:Surface	0.86	0.80	0.92
prevcoll1:finished1	0.30	0.15	0.57

From the above table, most of the confidence intervals does not contain 0. This means that the overall performance of our BIC selected model is doing pretty good.

Summary and Conclusions

Based on the model results, there're several categorical variables. The "baseline" category of a painting is when the painting is auctioned by dealer "J"; the origin of the painting based on nationality of the author is the same as the origin of painting based on dealers' classification in the catelogue; the origin of the painting based on nationality of artist is others; the dealer doesn't mention engravings done after the painting; the previous owner is not mentioned; the painting is not finished; the dealer doesn't devote an additional paragraph; and the description doesn't indicate still life elements.

The median price with the interval for the "baseline" painting is:

Table 2: Price for the baseline category

fit	lwr	upr
12.05006	1.020443	142.295

We find that the price is 12.05 livres. We're 95% confident that the true price will be in between 1.02 livres to 142.30 livres.

According to the model results, p values for some interactions are very small, indicating they are statistically significant and thus important. It's found that dealer L would affect the price of the painting. When the origin of the painting based on nationality of the author is the same as the origin of painting based on dealers' classification in the catelogue, dealer L could make the price of paintings increase multiplicatively by 2.66, with a 95% confidence interval from 1.94 to 3.64. When the dealer is dealer J, when the description doesn't indicate still life elements, and when the surface area is 1 sqrt inch, the price of the painting when the origin based on author is different from the origin based on dealers' classification would be 2.89 * 0.86 = 2.45 (range from 1.18 to 5.16) times the price of the painting when two origins are the same, holding others constant. If the author's nationality is Dutch or Flemish, the price would be 2.44 (range from 1.80 to 3.29) times the price if the nationality is others, holding other variables constant. Similarly, if the author is from Spain, the price would be 3 (range from 1.35 to 6.69) times the price if not. Moreover, engravtion would make the price increase multiplicatively by 2.36 (range from 1.75 to 3.18); When the painting is not finished, and when the previous owner is mentioned, the price would be 3.16 (range from 2.25 to 4.44) times the price if the previous owner is not mentioned. When the previous owner is not mentioned and the painting is finished, then the price would be 2.80 (range from 2.31 to 3.39) times the price if not. If the dealer devotes an additional paragraph, then the price would multiplivatively increase by 3.55 (range from 2.80 to 4.51). Finally, when the origin based on author is the same as the origin based on dealers' classification, as the surface of paintings increase by 1 squared inch, the price increase multiplicatively by 1.32 with a confidence interval from 1.25 to 1.39. If the year of sale is one more year later, the price could also increse multiplicatively by 1.15, with an interval ranging from 1.13 to 1.17.

Although the model does provide important variables that could affect the price of paintings. There're still several limitations. First of all, we select variables based on the exploratory data analysis. There could be some important variables that we do not choose which might be more related with the log price. In addition, we only consider two-way interactions between predictors. There could be some three-way interactions that are statistically significant. Moreover, the true relationship between log price and predictors might not be linear. Thus, in this case, it's impossible to build a OLS model with good performance.

In conclusion, we suggest that it's better to have dealer L to help with the auction. This positive effect is even stronger when the origin of the painting based on nationality of the author is the same as the origin of painting based on dealers' classification in the catelogue. If the author is from Dutch, Flemish, or Spain, the price of painting would be higher. Engravtion and the mention of the previous owner would both positively contribute to the increase of the price. If the dealer devotes more paragraphs to the painting, it would worth more. Finally, the art historian should look for a painting with as large surface area as possible. It's also better to have a painting that the year of sale is later rather than earlier.