Part-I-Writeup

Team-FP03

2019/12/7

Introduction

In this project, we are going to explore what factors drove the price of paintings in 18th century Paris, and thus to identify possible overvalued and undervalued paintings.

The dataset we are going to analyze is a series of auction transactions of paintings in Paris, ranging from 1764 to 1780. This dataset mainly contains the following information:

1. Sale data, this include basic information about painters, dealers, end buyers, transaction dates and prices;

2. Characteristics of paintings, such as their sizes, materials, number of figures and themes.

To address our problem, we devide this project into two parts:

1. In the first part, we carried out an exploratory data analysis. The target of this section is to understand the composition of our dataset and identify potential important variables.

2. In the second part, a simple linear regression model was fit to the data, aiming to confirm important variables and interactions from the model selection process and to prepare for fitting a more complex model.

Exploratory Data Analysis

In this section, we are going to explore our dataset in the following way: we first investigate the variables in the dataset to find their characteristics and possible relationships among each other; then we check the scatterplots between the response and each variable to identify potential important predictors.

Variable investigation

First of all, we can remove a few variables from the list of potential predictors simply based on their definitions:

Variable `price` is just the exponetial form of our target response `logprice`, and thus needs removing;
Variable `count` is the same for all observations, therefore there's no point to use it in the model fitting.

Besides these two, there exist quite a number of variables of interest:

Variables to impute

We've found that NA's exist in a lot of variables, and these NA's do not always indicate values missing completely at random. For example, from the R output below, we can see that `Surface` is not missing at random. Thus, instead of simply discarding observations containing NA's, we choose to impute the missing values with the observed ones.

For variables with a lot of blank values such as `endbuyer`, `type_intermed`, `material` and `mat`, we impute `n/a` into them to create a new category.

```
##
## Call:
## lm(formula = paintings_train$logprice ~ is.na(paintings_train$Surface))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9691 -1.3316 -0.0978  1.2455  5.5980
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         4.96915    0.04969 100.002   <2e-16
## is.na(paintings_train$Surface)TRUE -1.86766    0.21383  -8.734   <2e-16
##
## (Intercept)                         ***
## is.na(paintings_train$Surface)TRUE ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.872 on 1498 degrees of freedom
## Multiple R-squared:  0.04846,    Adjusted R-squared:  0.04782
## F-statistic: 76.29 on 1 and 1498 DF,  p-value: < 2.2e-16
```

Variables to manipulate

Variable `position` indicates the position of lot in the catalogue and is expressed as percentages. However, the maximum value of it in the dataset can be as large as 10.82, which are obviously typos. Similarly, there are observations with a series of size variables such as `Surface` all equal to 0. As a result, observations with impossible `position` and `Surface` values are dropped.

Besides, `Shape` variable has some weird values, such as `oval` vs. `ovale`, and `ronde` vs. `round`, which are probably typos and thus need fixing.

Addtionally, if variables `origin_author` and `origin_cat` are known, the value of `diff_origin` is 100% certain. Also, `type_intermed` incorporates all information of `Interm`. Thus, we decide to drop `diff_origin` and `Interm`.

In a similar manner, `Surface` should be known if `Diam_in`, or `Height_in` and `Width_in` are known at the same time. Also, note that `Surface` is the combination of `Surface_Rnd` and `Surface_Rect`. Thus, among all these variables mentioned, we keep just `Surface` in the model fitting process.

Variables `authorstandard`, `author`, `subject`, `sale`, `lot`, and `material` have way too many distinct values. Also, the possible values for these variables are too complicated and we decide not to use them in this simple model. When fitting a more complex model, it may be a good idea to convert them into new variables.

At last, in the dataset there exist strong correlations among some pairs of variables. For example, there is correlation between `Interm` & `type_intermed`, and `mat` & `materialCat`. In **Table 1** and **Table 2**, we display the contingency table for `Interm` vs. `type_intermed`, and as we can see, when `Interm` takes 0 `type_intermed` always takes `n/a`; when `Interm` takes 1, `type_intermed` takes other values. Thus, we decide to remove `Interm` and `materialCat`.

Table 1: Interm vs type_intermed

	B	D	E	EB	n/a
0	0	0	0	0	960
1	11	94	39	1	0

Table 2: mat vs materialCat

	canvas	copper	n/a	other	wood
al	0	0	0	1	0
ar	0	0	0	1	0
b	0	0	0	0	409
br	0	0	0	2	0
c	0	131	0	0	0
ca	0	0	0	2	0
co	0	0	0	5	0
g	0	0	0	1	0
h	0	0	9	0	0
m	0	0	0	1	0
mi	0	0	0	4	0
n/a	0	0	143	0	0
o	0	0	0	1	0
p	0	0	0	10	0
pa	0	0	0	4	0
t	731	0	0	0	0
ta	0	0	39	0	0
v	0	0	0	6	0

Important predictor identification

In this section we are going to evaluate scatter plots between our response `logprice` and each varaible after the manipulation from the previous part.

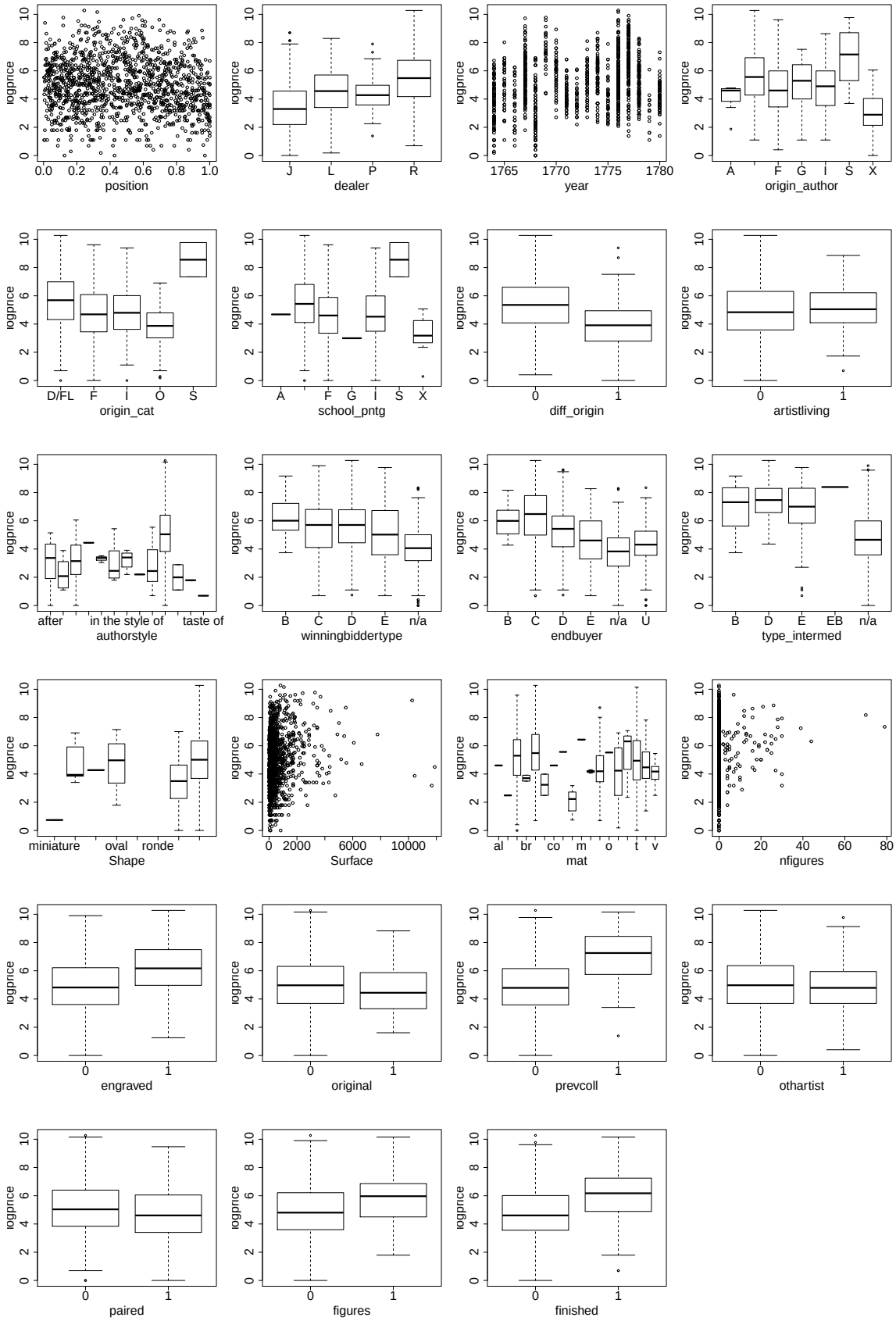

Figure 1: Plots of predictors versus logprice (1 to 23)

Figure 1 above displays the scatter plots between `logprice` and the first 23 variables in the dataset. Our target is to identify variables that show a strong relationship with the response. Bearing this in mind, it is easy to notice that variables `dealer`, `year`, `origin_author`, `winningbiddertype`, `endbuyer`, `type_intermed` and `finished` appear to have the strongest relationship with `logprice`. Also, there seem to be a very weak relationship between `position` and `logprice` as well. In addition, variables such as `Surface` are clustered near the beginning of x axis, and thus we decide to apply log transformations on them and have a closer look afterwards.

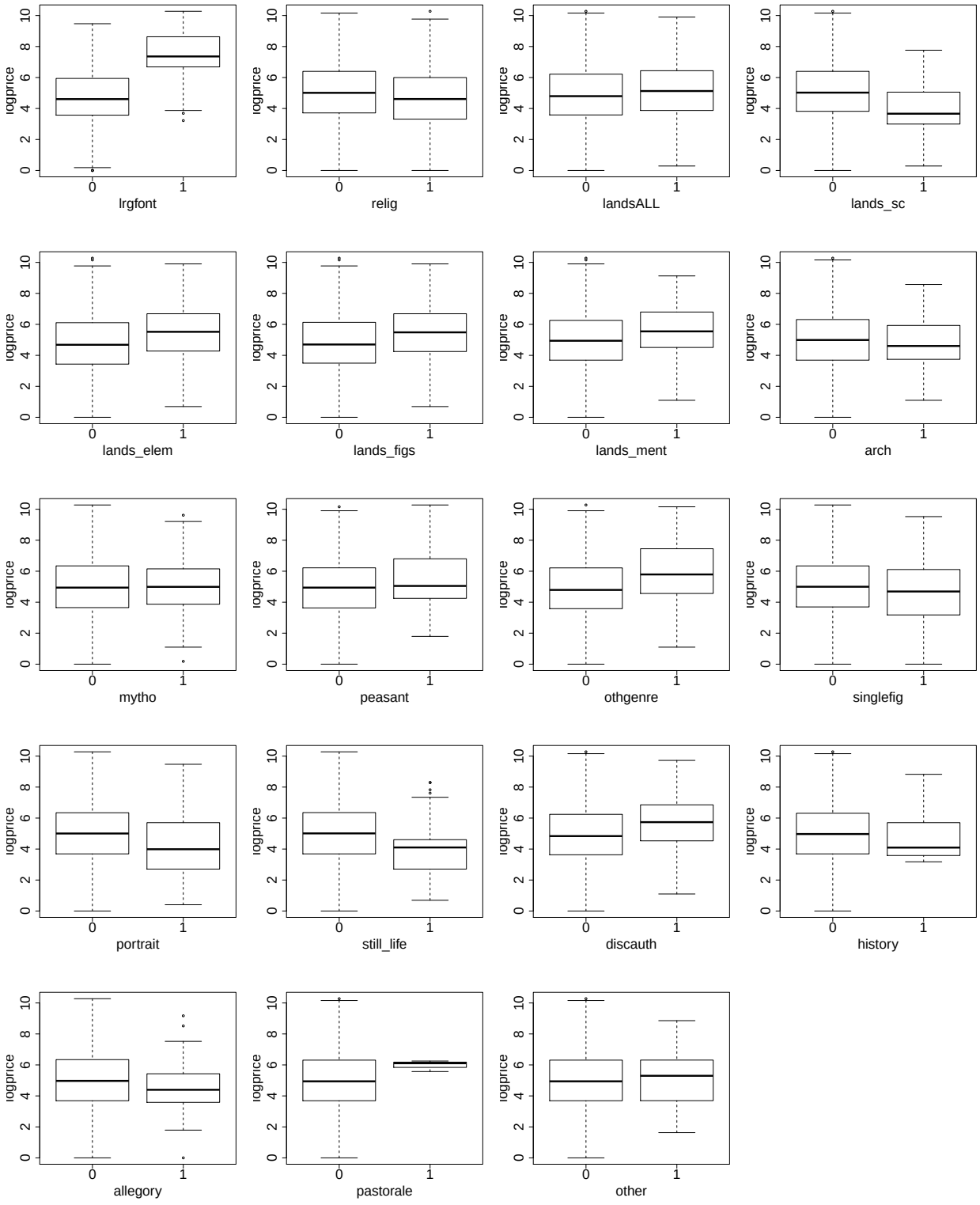

Figure 2: Plots of predictors versus logprice (24 to 42)

Figure 2 above display the scatter plots between `logprice` and the rest of the variables in the dataset. As we can see, most of the binary categorical variables fail to present a strong relationship with the response. The only exception is `lrgfont`, which corresponds to quite different response values at the two different levels.

For `Surface`, we can do log transformation to the corresponding predictors to see their relationship with `logprice` at a greater detail in **Figure 3**.

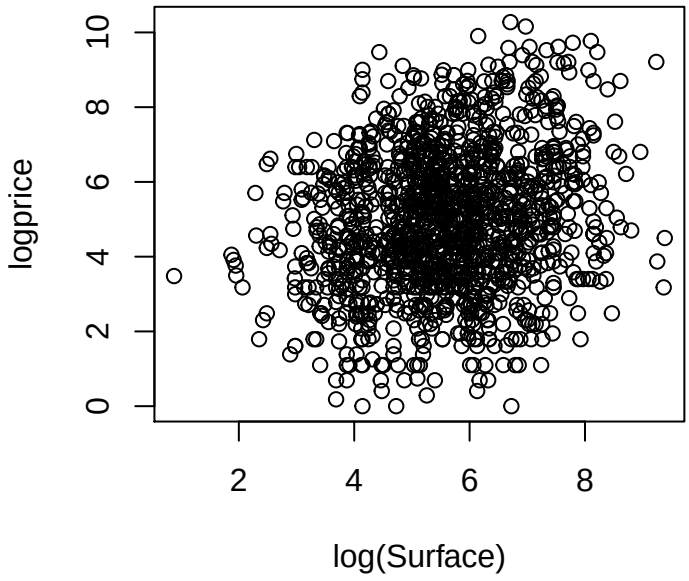

Figure 3: Plots of log Surface versus logprice

As we can see from **Figure 3**, there seem to be a weak relationship between `logprice` and log-transformed `Surface`. Intuitively, the surface of paintings should indeed be correlated to their prices.

Also, notice that `winningbiddertype` has too many levels, which may result in difficulties both in model fitting and in interpretation. Thus, we decide to apply the following transformation on `winningbiddertype`:

Observations with levels `B,BB,BC` are combined to have level `B`;
Observations with levels `C` remains untouched;
Observations with levels `D,DB,DC,DD` are combined to have level `D`;
Observations with levels `E,EB,EBC,EC,ED` are combined to have level `E`;
Blank space and unknown observations are combined to have level `n/a`.

The rationale for the above transformation is that, the bidder who actually attended the auction had the most important influence on the sale price.

In conclusion, after our manipulation of the dataset and inspection of the relationships between response and each variable, we reckon that variables `position`, `dealer`, `year`, `origin_author`, `winningbiddertype`, `endbuyer`, `type_intermed`, `finished`, `lrgfont` and the log transformation of `Surface` are the most important variables in terms of scatter plots and their definitions. However, we need formal model fitting and selection process to decide the variables and interactions to use.

Model fitting

In this section, we are going to present the development and assessment of our simple model.

Above all, we display the summary and anova table for our final model

```
##
## Call:
## lm(formula = logprice ~ dealer + year + origin_author + endbuyer +
##     log_Surface + finished + lrgfont + winningbiddertype + year:winningbiddertype,
##     data = paintings_train_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6733 -0.7864 -0.0191  0.8240  3.7817
##
## Coefficients: (1 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.571e+02  1.500e+02  -1.047   0.2952
## dealerL                     1.379e+00  1.352e-01  10.198   <2e-16 ***
## dealerP                     7.128e-04  1.650e-01   0.004   0.9966
## dealerR                     1.694e+00  1.121e-01  15.106   <2e-16 ***
## year                        8.967e-02  8.481e-02   1.057   0.2906
## origin_authorD/FL           4.021e-01  4.603e-01   0.874   0.3825
## origin_authorF             -1.354e-01  4.601e-01  -0.294   0.7686
## origin_authorG             -3.998e-02  5.160e-01  -0.077   0.9383
## origin_authorI             -3.106e-01  4.683e-01  -0.663   0.5073
## origin_authorS             -1.459e-01  5.998e-01  -0.243   0.8079
## origin_authorX             -9.973e-01  4.721e-01  -2.112   0.0349 *
## endbuyerC                   3.865e-01  4.837e-01   0.799   0.4244
## endbuyerD                  -7.258e-01  4.969e-01  -1.461   0.1444
## endbuyerE                  -7.515e-01  5.332e-01  -1.409   0.1589
## endbuyern/a                -2.506e+01  1.511e+02  -0.166   0.8683
## endbuyerU                  -2.456e+01  1.511e+02  -0.163   0.8709
## log_Surface                 3.568e-01  2.704e-02  13.196   <2e-16 ***
## finished1                   9.461e-01  9.379e-02  10.088   <2e-16 ***
## lrgfont1                    1.054e+00  1.242e-01   8.485   <2e-16 ***
## winningbiddertypeC         -1.096e+02  1.532e+02  -0.716   0.4744
## winningbiddertypeD          1.028e+01  1.513e+02   0.068   0.9458
## winningbiddertypeE         -2.102e+02  1.535e+02  -1.369   0.1712
## winningbiddertypen/a              NA         NA      NA       NA
## year:winningbiddertypeC     6.162e-02  8.663e-02   0.711   0.4770
## year:winningbiddertypeD    -5.489e-03  8.557e-02  -0.064   0.9489
## year:winningbiddertypeE     1.188e-01  8.685e-02   1.368   0.1715
## year:winningbiddertypen/a   1.351e-02  8.539e-02   0.158   0.8743
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.197 on 1304 degrees of freedom
## Multiple R-squared:  0.6086, Adjusted R-squared:  0.6011
## F-statistic: 81.09 on 25 and 1304 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
```

```
## Response: logprice
##                            Df  Sum Sq Mean Sq  F value    Pr(>F)
## dealer                      3  687.41  229.14 159.9782 < 2.2e-16 ***
## year                        1  689.24  689.24 481.2062 < 2.2e-16 ***
## origin_author               6  355.64   59.27  41.3837 < 2.2e-16 ***
## endbuyer                    5  360.03   72.01  50.2722 < 2.2e-16 ***
## log_Surface                 1  353.64  353.64 246.9050 < 2.2e-16 ***
## finished                    1  217.43  217.43 151.8026 < 2.2e-16 ***
## lrgfont                     1  141.39  141.39  98.7181 < 2.2e-16 ***
## winningbiddertype           3   47.97   15.99  11.1647 3.092e-07 ***
## year:winningbiddertype      4   51.00   12.75   8.9022 4.310e-07 ***
## Residuals                1304 1867.73    1.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The following is the process of building model:

1. for our initial model, we decide to incorporate all the important predictors in EDA.
2. we put the chosen main predictors and all their interactions into a full model. Then we use BIC to choose important predictors and interactions for us.
3. in the simply model, we have 8 main predictors and 1 interactions. Roughly 61% variation of dependent variables are explained by this model. By looking at the ANOVA table of the model, all of the variables are significant at the 5% level, which indicates that the variables in the model are reasonable.

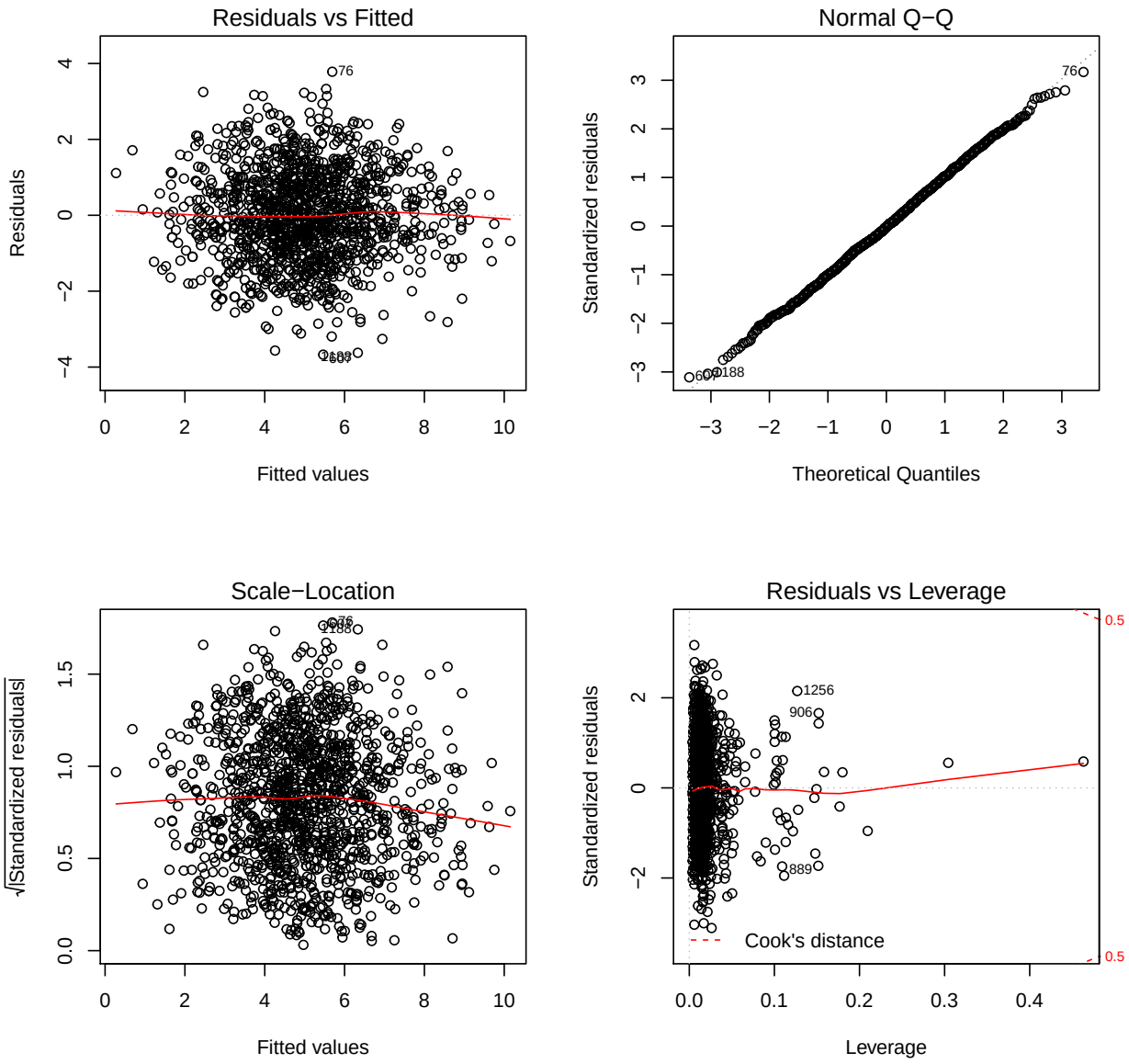

Figure 4: Diagnostic Plots

In the Residual vs Fitted plot, the zero level horizontal line is nearly flat and almost all points are randomly distributed around 0, indicating no violation for the linearity assumption.

In the Normal Q-Q plot, a slight light tail exists, but it's not a serious problem. Most of the residuals do seem to follow a normal distribution.

In the Scale-Location plot, the red line suggests that there is a small pattern for the resuduals. The residuals will increase first and then decrease a little bit. However, the points are very well randomly distributed around the zero level horizontal line, so the violation of constant variance is not significant enough to be very concerning.

In the Residuals vs Leverage plot, there are neither actually influential nor potentially influential points.

Generally, the diagnostic plots tell us that the linear model we get fit the training data very well and no model assumptions are violated.

Table 3: Summary of coefficients and confidence intervals

	Estimate	CI_Low	CI_Up
(Intercept)	-157.1314	-451.2282	136.9655
dealerL	1.3789	1.1139	1.6439
dealerP	0.0007	-0.3227	0.3242
dealerR	1.6935	1.4738	1.9133
year	0.0897	-0.0766	0.2559
origin_authorD/FL	0.4021	-0.5001	1.3044
origin_authorF	-0.1354	-1.0372	0.7664
origin_authorG	-0.0400	-1.0514	0.9714
origin_authorI	-0.3106	-1.2284	0.6072
origin_authorS	-0.1459	-1.3214	1.0297
origin_authorX	-0.9973	-1.9226	-0.0719
endbuyerC	0.3865	-0.5616	1.3346
endbuyerD	-0.7258	-1.6998	0.2482
endbuyerE	-0.7515	-1.7965	0.2935
endbuyern/a	-25.0599	-321.1651	271.0454
endbuyerU	-24.5559	-320.6575	271.5458
log_Surface	0.3568	0.3038	0.4098
finished1	0.9461	0.7623	1.1299
lrgfont1	1.0537	0.8103	1.2971
winningbiddertypeC	-109.5954	-409.7742	190.5834
winningbiddertypeD	10.2786	-286.2375	306.7947
winningbiddertypeE	-210.1942	-511.1240	90.7356
year:winningbiddertypeC	0.0616	-0.1082	0.2314
year:winningbiddertypeD	-0.0055	-0.1732	0.1622
year:winningbiddertypeE	0.1188	-0.0514	0.2890
year:winningbiddertypen/a	0.0135	-0.1538	0.1809

In **Table 3** above, we can see that part of the variables have high estimated coefficients compared to others. It may indicate the importance of the variables or that there are some potential problems in the linear model. There also exist several variables whose condifence intevals contain 0. These variables may either have positive or negative effects on the price, which means the model is still not very satisfactory. we will use a more complicated model in the next part to improve the performance of the model.

Summary and Conclusions

In our final model, the baseline price is e^{-157} livres, which is approximately 0 livres. It represents the price of a painting under baseline categories for all categorical variables, such as `dealer`, `origin_author` and `endbuyer`,etc.

According to the coefficients table above, predictors `year` and `winningbiddertype` have huge impact on the price sale. For `year`, althouth its coefficient is not large compared to others, the big numeric value itself will have impact on the price. Besides `year`, for the dummy variables, `winningbiddertype` is another important predictor that affects the price most.

Thus, the two most important variables are `year` and `winningbiddertype`. And the only interaction we have is the one between `year` and `winningbiddertype`. So it's natural to say that the interaction is also improtant.

Our model also has limitations. We choose all the main predictors from EDA, so we may actually ignore some relatively important predictors that are not identified through EDA. In our simple model, predictors `year`, `winningbiddertype`, and `year:winningbiddertype` look a little bit overly important compared to all other variables. It's questionalble for such a large data set. Besides, we only use the linear model to fit the data, resulting in a few large coefficents and standard deviations. Furthermore, the big estimated coefficients make us hard to interpret the model to the art historian. Thus, we may use nonlinear model to shrink the coefficients in the next part. There may even exist some more complicated relationships in the data such as polynomial, which still needs to be explored.

In our model, for every one year after the previous year, we expect that the price of the painting will be $e^{0.09}$ times higher, and we are 95% confident that the fluction is between $e^{-0.08}$ to $e^{0.26}$, which is from 0.92 to 1.30.

Given all other conditions unchanged (eg: same dealer, same year, same origin, etc.), we expect the price of the painting will be e^{-110} times higher if the type of winning bidder is a collector. And we are 95% confident that the price fluction will be between e^{-410} and e^{191} times higher.

Given all other conditions unchanged (eg: same dealer, same year,same origin,etc.), we expect the price of painting will be e^{10} times higher if the type of winning bidder is a dealer. And we are 95% confident that the price fluction will be between e^{-286} and e^{307} times higher.

Given all other conditions unchanged (eg: same dealer, same year,same origin,etc.), we expect the price of painting will be e^{-210} times higher if the type of winning bidder is an expert organizing the sale. And we are 95% confident that the price fluction will be between e^{-511} and e^{91} times higher.

So we suggest the art historians that the painting bid by dealer with a larger year will have a high value.