Part-I-Writeup

Team-FP03

2019/12/4

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 Thus, since the former two variables contain more specific information, we decide to drop `diff_origin`.
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 the following table, 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`.

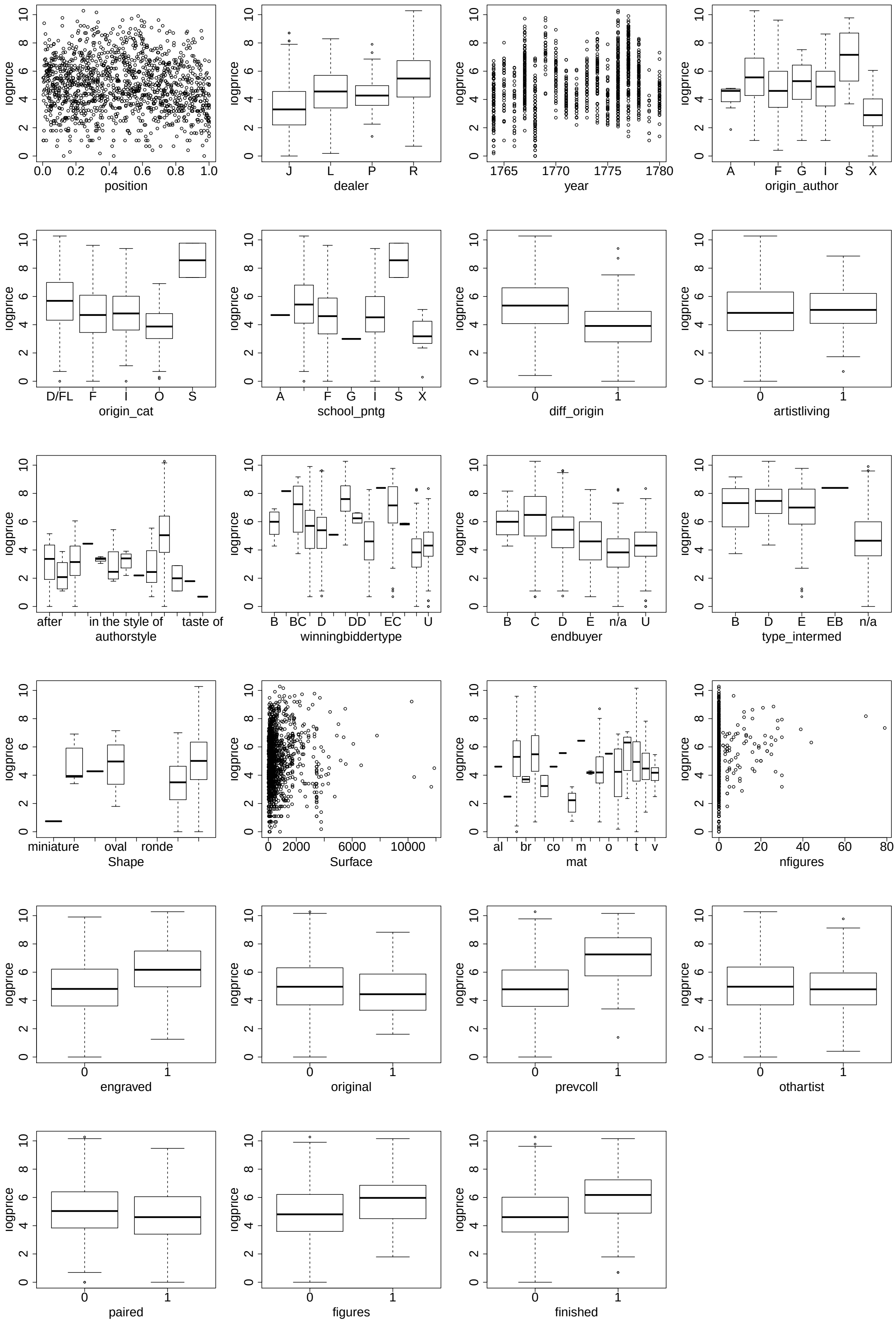
```
##
##        B   D   E  EB n/a
##   0    0   0   0   0 960
##   1   11  94  39   1   0
```

```
##
##        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.



logprice
0 2 4 6 8 10
0.0 0.2 0.4 0.6 0.8 1.0
position
J L P R
dealer
1765 1770 1775 1780
year
A F G I S X
origin_author
D/FL F I O S
origin_cat
A F G I S X
school_pntg
0 1
diff_origin
0 1
artistliving
after in the style of taste of
authorstyle
B BC D DD EC U
winningbiddertype
B C D E n/a U
endbuyer
B D E EB n/a
type_intermed
miniature oval ronde
Shape
0 2000 6000 10000
Surface
al br co m o t v
mat
0 20 40 60 80
nfigures
0 1
engraved
0 1
original
0 1
prevcoll
0 1
othartist
0 1
paired
0 1
figures
0 1
finished

The **Figure 1** above displays the scatter plots between `logprice` and the first 24 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`, `prevcoll`, `endbuyer`, `type_intermed` and `Shape` appear to have the strongest relationship with `logprice`. 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 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**.

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.

In conclusion, after our manipulation with the dataset and inspection of the relationships between response and each variable, we reckon that variables `dealer`, `year`, `origin_author`, `prevcoll`, `endbuyer`, `type_intermed`, `Shape`, `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.

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

```
##
## Call:
## lm(formula = logprice ~ dealer + year + origin_author + endbuyer +
##     type_intermed + log_Surface + finished + lrgfont + prevcoll +
##     year:endbuyer + finished:prevcoll, data = paintings_train_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6752 -0.7790 -0.0252  0.7760  3.8489
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -1.689e+02  1.414e+02  -1.195   0.2324
## dealerL             1.440e+00  1.347e-01  10.686  < 2e-16 ***
## dealerP             5.041e-02  1.666e-01   0.303   0.7622
## dealerR             1.700e+00  1.123e-01  15.128  < 2e-16 ***
## year                9.687e-02  7.991e-02   1.212   0.2256
## origin_authorD/FL   3.936e-01  4.581e-01   0.859   0.3904
## origin_authorF     -1.073e-01  4.580e-01  -0.234   0.8148
## origin_authorG     -1.579e-01  5.140e-01  -0.307   0.7587
## origin_authorI     -3.533e-01  4.667e-01  -0.757   0.4492
## origin_authorS     -3.832e-01  5.986e-01  -0.640   0.5222
## origin_authorX     -1.066e+00  4.703e-01  -2.267   0.0236 *
## endbuyerC          -5.725e+01  1.432e+02  -0.400   0.6893
## endbuyerD           4.182e+01  1.433e+02   0.292   0.7705
## endbuyerE          -2.171e+02  1.460e+02  -1.487   0.1372
```

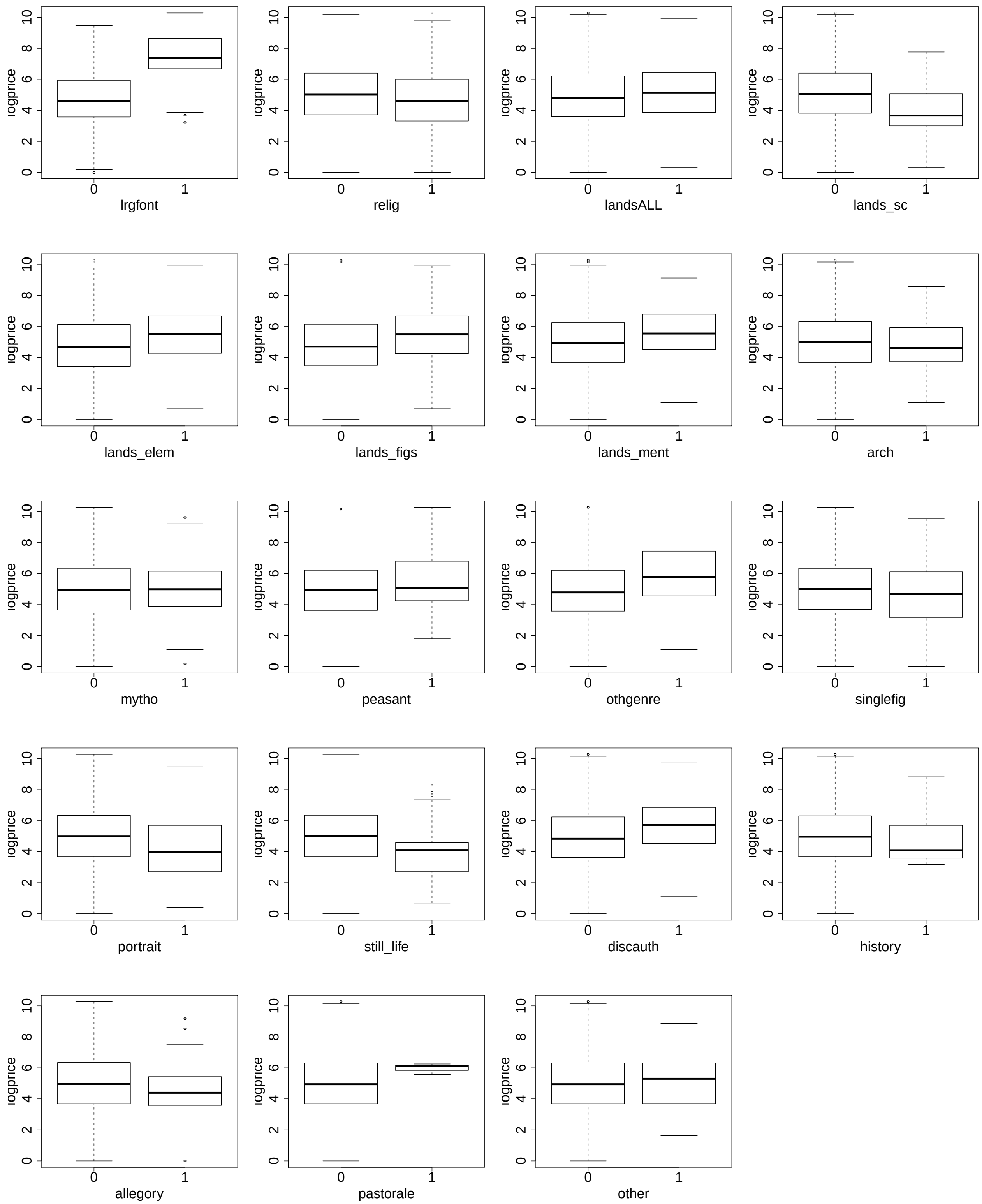


Figure 1: Plots of predictors versus logprice

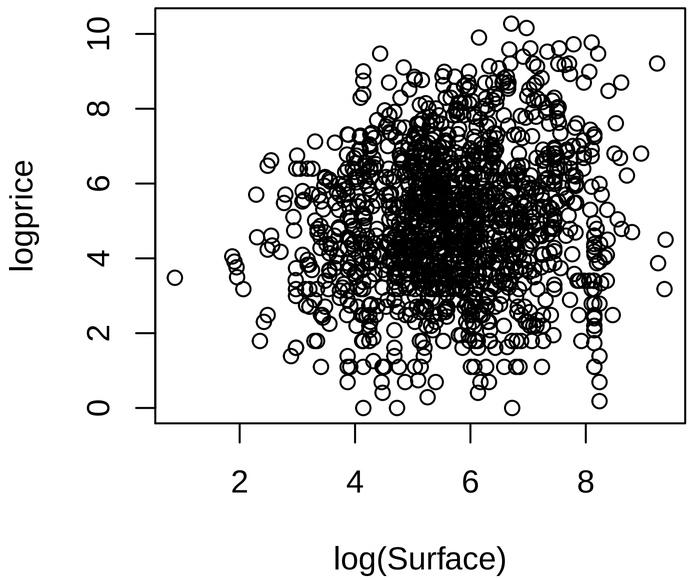


Figure 2: Plots of log Surface versus logprice

```
## endbuyern/a         -4.393e+00  1.429e+02  -0.031   0.9755
## endbuyerU           -1.615e+01  1.455e+02  -0.111   0.9117
## type_intermedD       5.673e-02  3.968e-01   0.143   0.8863
## type_intermedE      -2.171e-01  4.173e-01  -0.520   0.6030
## type_intermedEB      2.709e+00  1.249e+00   2.169   0.0303 *
## type_intermedn/a    -7.860e-01  3.775e-01  -2.082   0.0375 *
## log_Surface          3.097e-01  2.677e-02  11.569  < 2e-16 ***
## finished1            9.982e-01  9.716e-02  10.274  < 2e-16 ***
## lrgfont1             1.047e+00  1.250e-01   8.379  < 2e-16 ***
## prevcoll1            1.123e+00  1.776e-01   6.324 3.51e-10 ***
## year:endbuyerC       3.229e-02  8.090e-02   0.399   0.6899
## year:endbuyerD      -2.368e-02  8.097e-02  -0.292   0.7700
## year:endbuyerE       1.223e-01  8.248e-02   1.483   0.1384
## year:endbuyern/a     1.820e-03  8.072e-02   0.023   0.9820
## year:endbuyerU       8.795e-03  8.223e-02   0.107   0.9148
## finished1:prevcoll1 -8.895e-01  3.218e-01  -2.764   0.0058 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.192 on 1300 degrees of freedom
## Multiple R-squared:  0.613,  Adjusted R-squared:  0.6043
## F-statistic:    71 on 29 and 1300 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  229.14 161.3002 < 2.2e-16 ***
## year                  1  689.24  689.24 485.1827 < 2.2e-16 ***
## origin_author         6  355.64   59.27  41.7256 < 2.2e-16 ***
## endbuyer              5  360.03   72.01  50.6877 < 2.2e-16 ***
## type_intermed         4  118.75   29.69  20.8976 < 2.2e-16 ***
## log_Surface           1  275.20  275.20 193.7262 < 2.2e-16 ***
```

```
## finished             1  209.52  209.52 147.4916 < 2.2e-16 ***
## lrgfont              1  115.32  115.32  81.1822 < 2.2e-16 ***
## prevcoll             1   46.28   46.28  32.5794 1.416e-08 ***
## year:endbuyer        5   56.51   11.30   7.9557 2.138e-07 ***
## finished:prevcoll    1   10.85   10.85   7.6377  0.005797 **
## Residuals         1300 1846.74    1.42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The following is the process of model building:

First, for our initial model, we decide to incorporate all the important predictors identified in EDA, and then add some extra predictors for the following reasons:

1.`origin_cat`: when a painting was created by artists who were not well-known, then the origin of paintings based on dealers' classification is the only way bidders get to know the origin of paintings, therefore we think origin_cat would be helpful for our model beside origin_author. And we use an anova test to check that.

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## origin_cat     4    509  127.32   39.58 <2e-16 ***
## Residuals   1325   4262    3.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, we set the `logprice` to multiple groups according to `origin_cat`, and use anova to compare whether their group means are the same or not. Since the p_value is smaller than 0.05. we can state that the means are differnet across groups. So we decide to use `origin_author`.

2.`finished`: we believe that whether a painting is finished or not will affect the the price of the sale. Also, the plot of `finished` in EDA can also prove our thinking. So we choose this predictor based on both commen sense and plot analysis.

Secondly, we put the chosen main predictors and all their interactions into a full model. Then we use BIC to choose the important predictors and interactios for us.

Finally, in the simple model, we have 8 main predictors and 1 interaction. Roughly 60% 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 indicate the variables in the model are reasonable.

In **Figure 4**, the Residual vs Fitted plot, there is only a slight curve at the begining of the 0-level horizontal line, which is not a serious problem. Almost all points are randomly distributed around 0, indicating no significant violation for the linearity assumption.

The Normal Q-Q plot indicates nearly perfect distribution. Almost all residuals follow a normal distribution.

In the Scale-Location plot, the red line suggest that there is a small pattern for the resuduals. The absolute value of residuals will increase first and then decrease. But the points are very well randomly distributed around the 0 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 ones.

Generally, the diagnostic plots tell us that the linear model we get fits the training data very well and does not violate any assumptions.

	estimate	CI_Low	CI_Up
(Intercept)	-168.9314731	-446.0907986	108.2278524
dealerL	1.4396339	1.1755829	1.7036849
dealerP	0.0504092	-0.2760757	0.3768942
dealerR	1.6995196	1.4793318	1.9197075
year	0.0968714	-0.0597442	0.2534870

	estimate	CI_Low	CI_Up
origin_authorD/FL	0.3935798	-0.5042193	1.2913789
origin_authorF	-0.1073060	-1.0049392	0.7903273
origin_authorG	-0.1579195	-1.1654184	0.8495795
origin_authorI	-0.3532631	-1.2679576	0.5614315
origin_authorS	-0.3832501	-1.5566022	0.7901020
origin_authorX	-1.0659396	-1.9876675	-0.1442118
endbuyerC	-57.2507264	-337.8905824	223.3891295
endbuyerD	41.8161400	-239.0602036	322.6924836
endbuyerE	-217.0834770	-503.1859160	69.0189621
endbuyern/a	-4.3927393	-284.4026726	275.6171941
endbuyerU	-16.1483385	-301.4068983	269.1102213
type_intermedD	0.0567337	-0.7209861	0.8344534
type_intermedE	-0.2170562	-1.0349595	0.6008471
type_intermedEB	2.7088949	0.2605330	5.1572567
type_intermedn/a	-0.7860038	-1.5259355	-0.0460720
log_Surface	0.3096905	0.2572231	0.3621580
finished1	0.9982296	0.8078024	1.1886567
lrgfont1	1.0470983	0.8021764	1.2920202
prevcoll1	1.1233417	0.7751648	1.4715186
year:endbuyerC	0.0322902	-0.1262760	0.1908564
year:endbuyerD	-0.0236815	-0.1823770	0.1350141
year:endbuyerE	0.1222996	-0.0393544	0.2839536
year:endbuyern/a	0.0018196	-0.1563878	0.1600271
year:endbuyerU	0.0087954	-0.1523704	0.1699613
finished1:prevcoll1	-0.8894655	-1.5202830	-0.2586480

In the table above, we can see that part of the variables have high estimates compared to others. It may indicate the importance of the variables or the potential problems exist in the linear model. There also exist several variables whose condifence inteval contain 0. These variables may either have positive or negative effects on the price. we will use 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^{-150} livres, which is approximately 0 livres. It represents the price of a painting under baseline categories for all categorical variables, such as `dealer` and `endbuyer`,etc.

According to the coefficient table we get above, predictor `year` and `endbuyer` 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. Beside `year`, for the dummy variables, `endbuyer` is another important predictor that affect the price most.

The two most important variables are `year` and `endbuyer`. And the only interaction we have is the interaction between `year` and `endbuyer`. So it's natural to say the interaction is also improtant.

Our model also has limitations. we choose the main predictors mostly from EDA and by ourselves so we may ignore some important predictors. In our simple model, predictors `year`, `endbuyer`, and `year:endbuyer` look a little overly important compard to all other variables, Which means to a certain degree we can just predict the price by using 3 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 interpreate the model to the art historian. Thus, we may use nonlinear

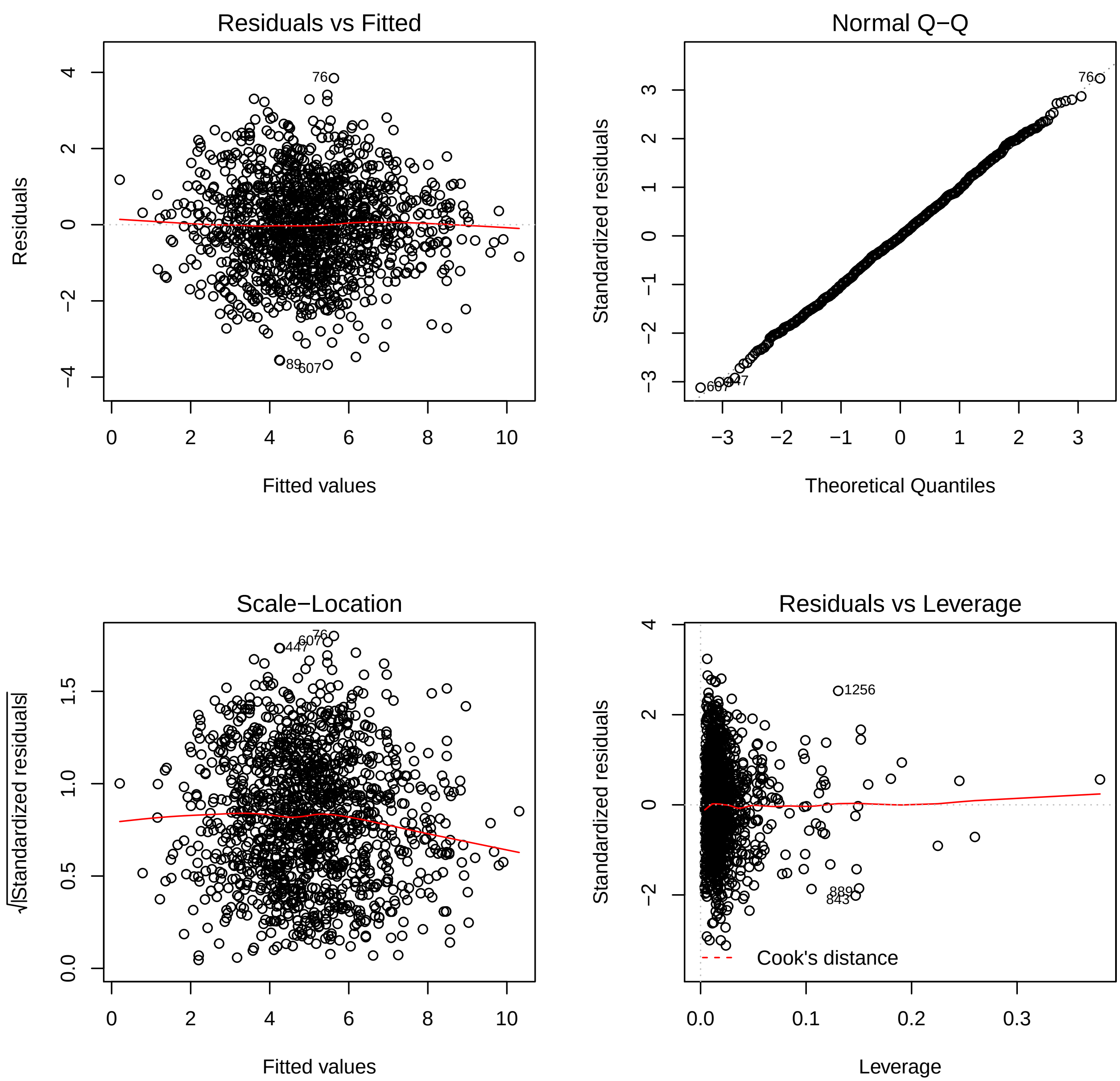


Figure 3: Diagnostic Plots

model to shrink the coefficients in the next part. There may even exist some more complicated relationships in the data such as polynomial. We still need to explore that.

For every one year after the previous year, we expect price of the painting will be $e^{0.09}$ times higher, and we are 95% confident that the fluction is between $e^{-0.07}$ to $e^{0.24}$, which is from 0.93 to 1.27.

Given all other conditions unchanged(eg:same dealer, same year,same origin,etc.), we expect the price of painting will be e^{-91} times higher if the buyer is a collector. And we are 95% confident that the price fluction will be between e^{-375} and e^{193} times higher.

Given all other conditions unchanged(eg:same dealer, same year,same origin,etc.), we expect the price of painting will be e^{13} times higher if the buyer is a dealer And we are 95% confident that the price fluction will be between e^{-272} and e^{398} times higher.

Given all other conditions unchanged(eg:same dealer, same year,same origin,etc.), we expect the price of painting will be e^{-247} times higher if the buyer is expert organizing the sale. And we are 95% confident that the price fluction will be between e^{-537} and e^{43} times higher.

Given all other conditions unchanged(eg:same dealer, same year,same origin,etc.), we expect the price of painting will be e^{-34} times higher if the buyer is unknown person. And we are 95% confident that the price fluction will be between e^{-318} and e^{250} times higher.

Given all other conditions unchanged(eg:same dealer, same year,same origin,etc.), we expect the price of painting will be e^{-40} times higher if the buyer is person without information. And we are 95% confident that the price fluction will be between e^{-329} and e^{248} times higher.

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

```
##
##  iter imp variable
##   1   1  Surface*
##   1   2  Surface*
##   1   3  Surface*
##   1   4  Surface*
##   1   5  Surface*
##   2   1  Surface*
##   2   2  Surface*
##   2   3  Surface*
##   2   4  Surface*
##   2   5  Surface*
##   3   1  Surface*
##   3   2  Surface*
##   3   3  Surface*
##   3   4  Surface*
##   3   5  Surface*
##   4   1  Surface*
##   4   2  Surface*
##   4   3  Surface*
##   4   4  Surface*
##   4   5  Surface*
##   5   1  Surface*
##   5   2  Surface*
##   5   3  Surface*
##   5   4  Surface*
##   5   5  Surface*
##  * Please inspect the loggedEvents
```