Part-I-Writeup

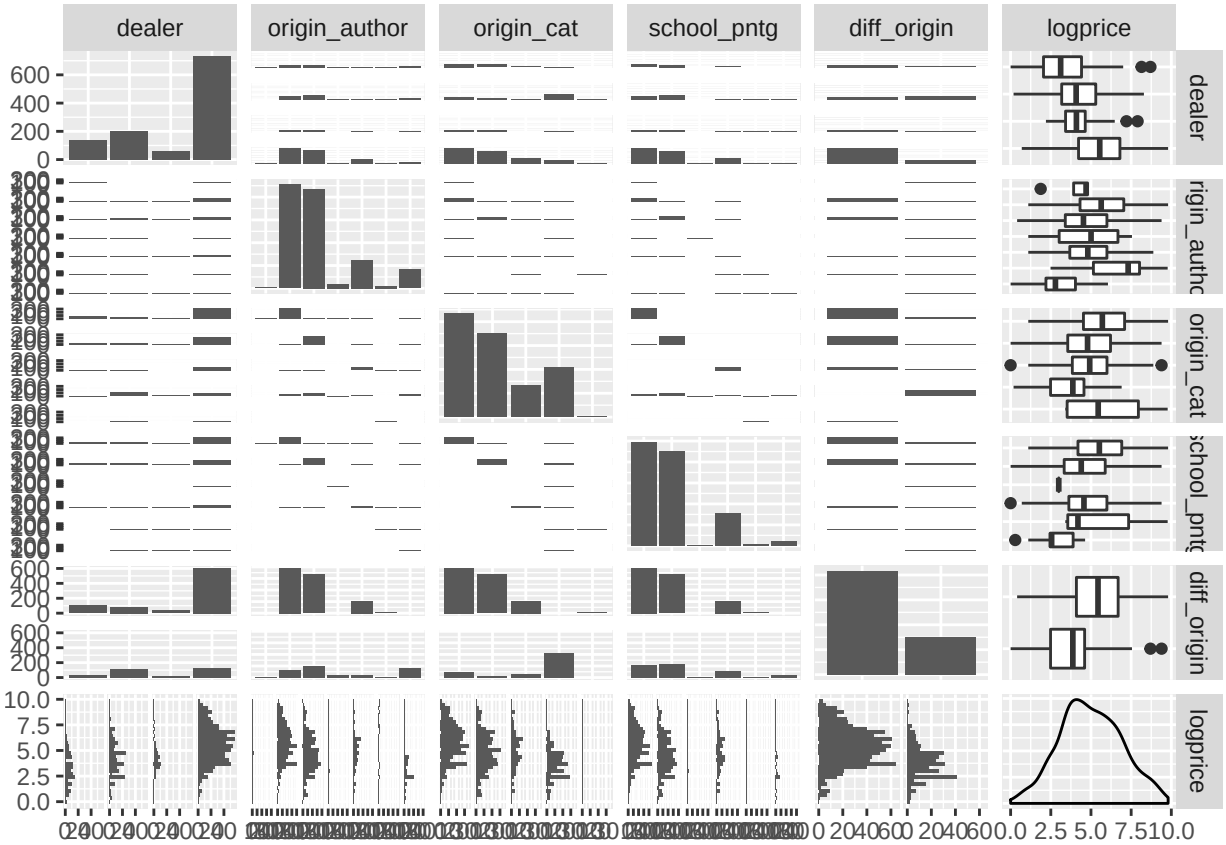
Anonymous to Everyone

12/8/2017

1. Introduction
   For artwork, there is no intrinsic, objective value. The price of paintings might depend on a large range of factors, includes the artists, style of painting, dealers, buyers and so on. In this project, we will help art historian understand what factors drove prices of painting and decide whether paintings might be overvalued or undervalued. We have the original data with 59 possible variables and 1131 observations. The objective of this project is to find the best model that can predict the price of the paintings.

2. Exploratory data analysis

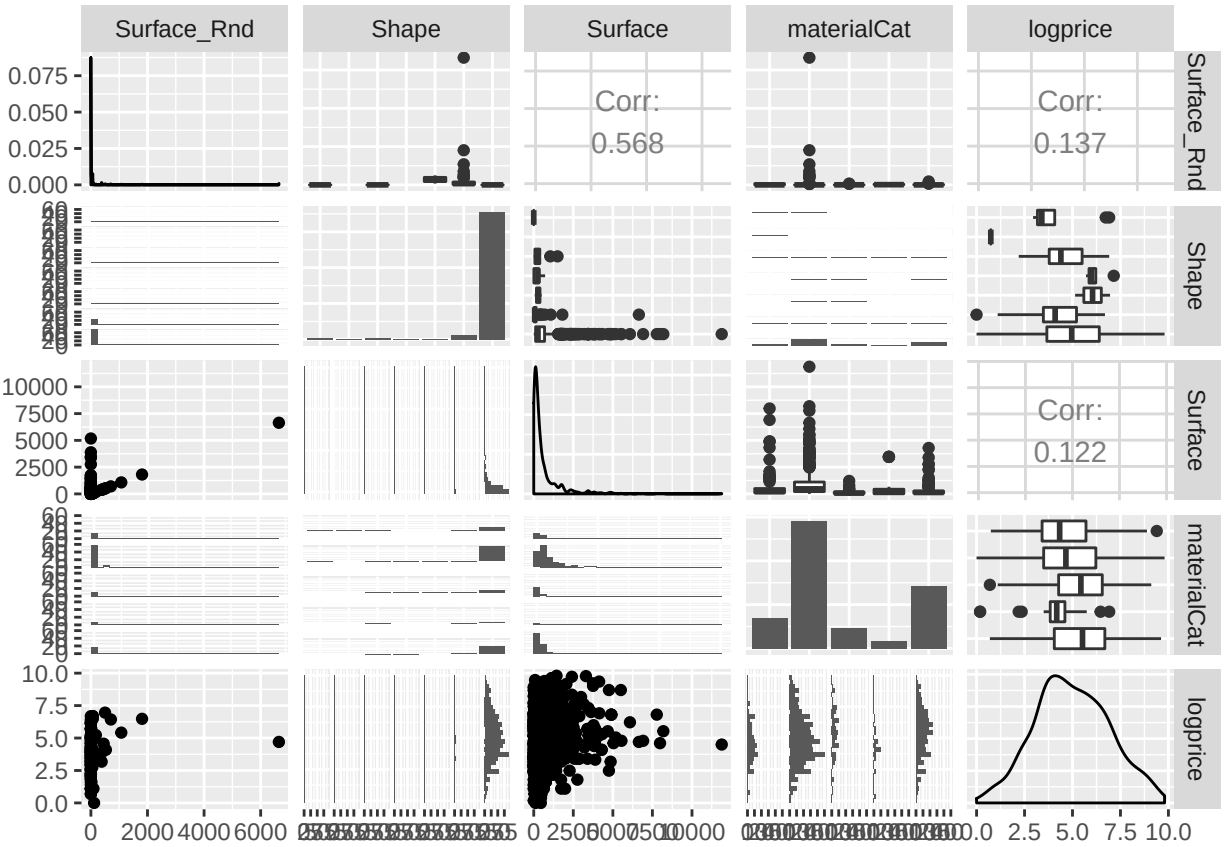
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We first use ggpairs to see the relationship between different variables and log price. The above graph shows us the distribution of categorical variables and their relationship with logprice. The graph in last row is a feature that we focus on, as the last row represents the relationship between log price and other variables. We want to find the variable that have distinct different mean of logprice in each level. For example, it is very obvious that different type of dealer(four levels) has data gathered into different center. The same with Origin_author. As a result, we include those into our model selection process

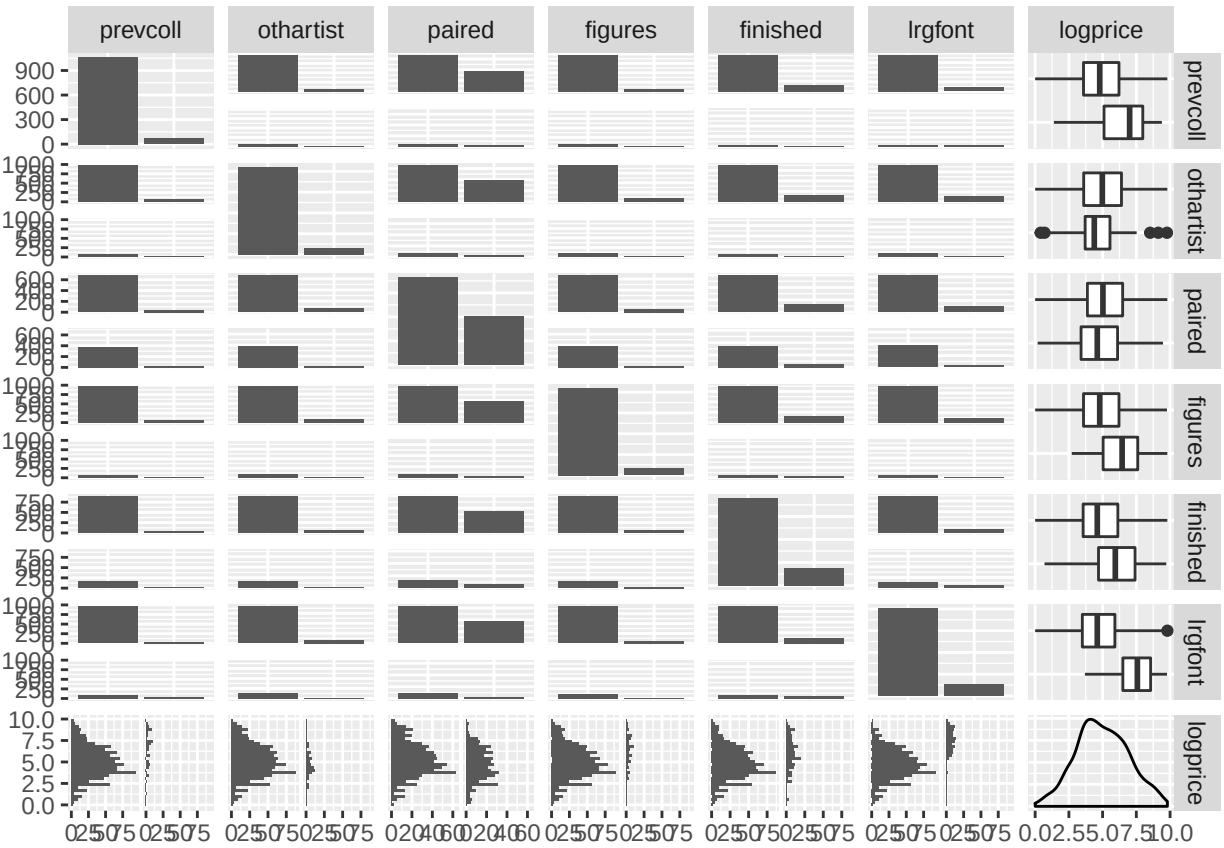
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

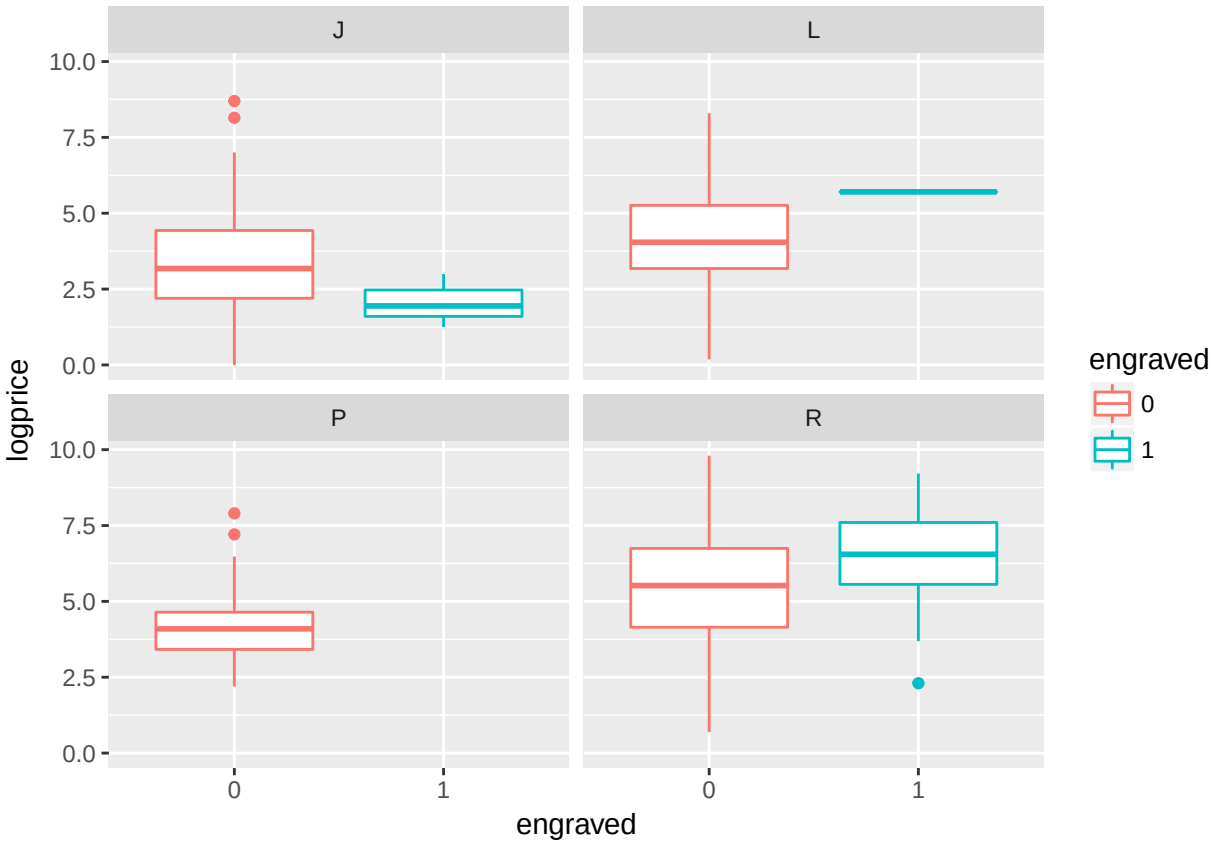


This ggpair graph includes a continuous variable, Surface. Correlation index is the feature that we look at. For example, the correlation of log price and surface here is relative large, so we propose that surface may is a good predictor in predicting price, so we include it into our model selection process. We can also see a correlated trend between Surface and log price in the last row. The log price is significant different for different levels of materialCat, so we also include that variable into our initial model selection.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This ggpairs graph helps us to explore dummy variables. The last row also represents important features. For example, we can identify that the log price is centered at a place very different for having figures versus not having figure. lrgfont also shows the significant difference between two groups. As a result, we also include figures and lrgfont in the model.



We use ggplot with different facets to indentify interaction. The boxplots here show that the mean and quantile of log price for painting that is engraved or not engraved is different for different dealers. As a result, we speculate that there is a interaction between this two varaibles. So we add it into our model selection process.

3. Development and assessment of an initial model (10 points)

* Initial model: must include a summary table and an explanation/discussion for variable selection and overall amount of variation explained.

```
##
## Call:
## lm(formula = logprice ~ year + dealer + origin_cat + lrgfont +
##     Surface + diff_origin + engraved + endbuyer + lands_sc +
##     finished + paired + discauth + type_intermed + origin_cat:arch +
##     dealer:engraved + dealer:paired + shape_recode, data = paintings_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0419 -0.7080 -0.0467  0.6828  3.1942
##
## Coefficients: (2 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.890e+00  6.066e-01   4.763 2.18e-06 ***
## year1765                1.478e+00  2.531e-01   5.841 6.95e-09 ***
## year1766                2.345e-01  3.139e-01   0.747 0.455307
## year1767                1.246e+00  1.598e-01   7.798 1.54e-14 ***
## year1768                1.196e-01  1.917e-01   0.624 0.532904
## year1769                2.386e+00  2.903e-01   8.221 6.03e-16 ***
## year1770                1.277e+00  2.182e-01   5.854 6.46e-09 ***
```

```
## year1771              9.529e-01  1.857e-01   5.133 3.41e-07 ***
## year1772              6.939e-01  2.696e-01   2.574 0.010193 *
## year1773              1.109e+00  2.389e-01   4.642 3.89e-06 ***
## year1774              1.918e+00  1.997e-01   9.606  < 2e-16 ***
## year1775              5.568e-01  3.958e-01   1.407 0.159799
## year1776              1.686e+00  1.544e-01  10.917  < 2e-16 ***
## year1777              2.466e+00  1.647e-01  14.974  < 2e-16 ***
## year1778              5.374e-01  2.472e-01   2.174 0.029925 *
## year1779              7.179e-01  3.928e-01   1.828 0.067913 .
## year1780              3.409e-01  3.512e-01   0.971 0.331927
## dealerL               2.500e+00  2.372e-01  10.540  < 2e-16 ***
## dealerP               1.665e+00  3.906e-01   4.262 2.22e-05 ***
## dealerR               1.513e+00  1.923e-01   7.866 9.25e-15 ***
## origin_catF          -6.662e-01  9.175e-02  -7.261 7.57e-13 ***
## origin_catI          -6.853e-01  1.200e-01  -5.711 1.47e-08 ***
## origin_catO          -8.458e-01  1.849e-01  -4.575 5.34e-06 ***
## origin_catS          -1.376e+00  5.726e-01  -2.403 0.016429 *
## lrgfont1              8.588e-01  1.288e-01   6.666 4.29e-11 ***
## Surface               2.994e-04  3.457e-05   8.658  < 2e-16 ***
## diff_origin1         -5.096e-01  1.342e-01  -3.797 0.000155 ***
## engraved1            -4.009e-01  6.614e-01  -0.606 0.544526
## endbuyerB             8.365e-01  3.056e-01   2.737 0.006309 **
## endbuyerC             8.223e-01  1.474e-01   5.579 3.10e-08 ***
## endbuyerD             7.530e-01  1.247e-01   6.039 2.17e-09 ***
## endbuyerE             4.347e-01  1.586e-01   2.741 0.006230 **
## endbuyerU             3.807e-01  1.458e-01   2.611 0.009154 **
## lands_sc1            -4.040e-01  1.347e-01  -2.999 0.002778 **
## finished1             5.157e-01  1.024e-01   5.034 5.68e-07 ***
## paired1              -1.131e-01  2.014e-01  -0.562 0.574556
## discauth1             4.122e-01  1.503e-01   2.742 0.006205 **
## type_intermedB        4.258e-01  3.721e-01   1.144 0.252816
## type_intermedD        8.682e-01  1.564e-01   5.553 3.58e-08 ***
## type_intermedE        3.246e-01  2.227e-01   1.457 0.145327
## type_intermedEB       4.245e-01  8.281e-01   0.513 0.608350
## shape_recodeoval     -5.082e-01  6.466e-01  -0.786 0.432051
## shape_recoderound    -1.252e+00  6.070e-01  -2.062 0.039459 *
## shape_recodesqu_rect -7.814e-01  5.774e-01  -1.353 0.176273
## origin_catD/FL:arch1 -3.191e-01  2.984e-01  -1.069 0.285218
## origin_catF:arch1     3.373e-01  2.528e-01   1.334 0.182389
## origin_catI:arch1     5.357e-01  5.684e-01   0.942 0.346173
## origin_catO:arch1    -1.499e-01  5.089e-01  -0.295 0.768382
## origin_catS:arch1            NA         NA      NA       NA
## dealerL:engraved1     2.629e+00  1.289e+00   2.040 0.041641 *
## dealerP:engraved1            NA         NA      NA       NA
## dealerR:engraved1     1.243e+00  6.832e-01   1.819 0.069205 .
## dealerL:paired1      -4.929e-01  2.817e-01  -1.749 0.080511 .
## dealerP:paired1      -2.709e-01  3.636e-01  -0.745 0.456371
## dealerR:paired1      -2.954e-01  2.212e-01  -1.335 0.182125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.1 on 1027 degrees of freedom
##   (51 observations deleted due to missingness)
## Multiple R-squared:  0.6814, Adjusted R-squared:  0.6653
```

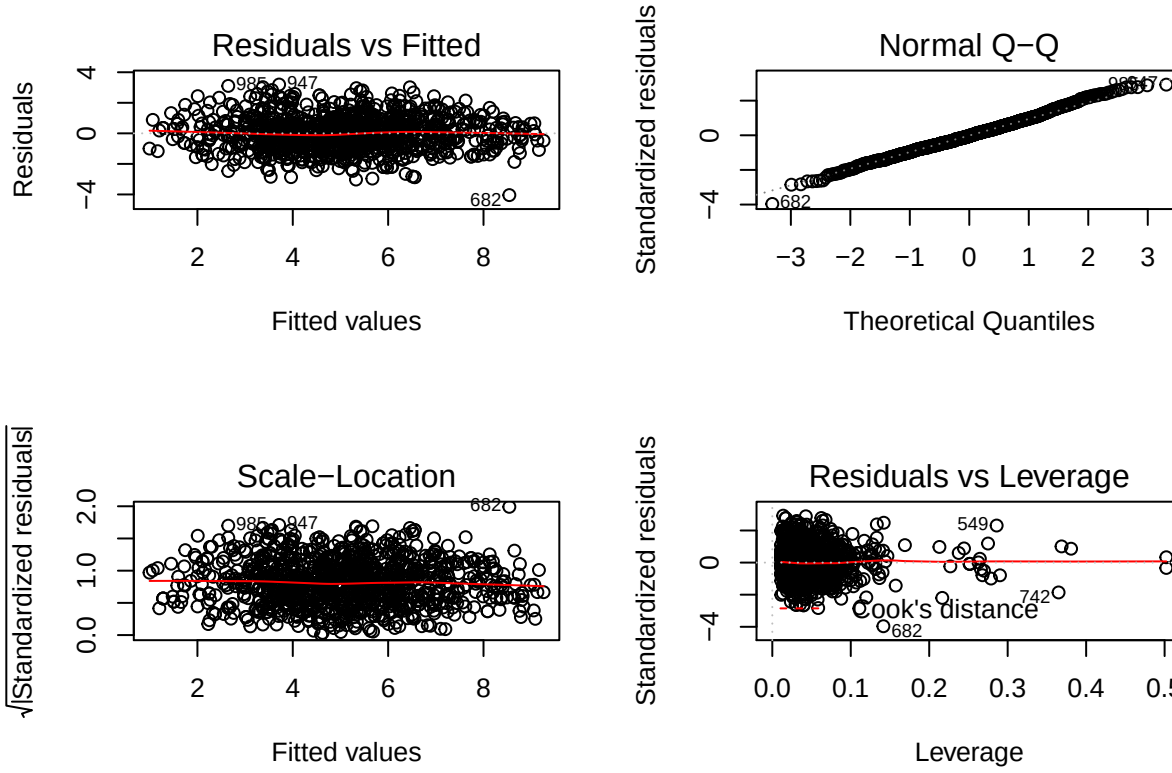
```
## F-statistic: 42.25 on 52 and 1027 DF,  p-value: < 2.2e-16
```

The multiple r squared is 0.6814 and the adjusted r-squared is 0.6653. It means that the model explain 66.53% of the data. As can be seen from the summary, we choose dealer, year, origin_cat, diff_origin, engraved, shape, lrgfont, Surface, endbuyer, lands_sc, finished, paired, discauth, type_intermed and shape as variables and include three interactions: origin_cat:arch, dealer:engraved, dealer:paired. All of the single variables are important according to its p-value. Year of sale influences the price because price depends on the economy trend of the year. Dealers and buyers are the two key marketing forces within the art world and thus we include endbuyers and dealers in the variable. Variables such as origin_cat, diff_origin, engraved, lrgfont, surface, finished are also very important based on the AIC selection. We also add discauth, land_sc, shape,and paired to our model. We treid to include several interactions into our model. Origin_cat might have interaction with variable such as arch, relig and so on. Interactions that can improve the model also include paired & dealer, engraved & dealer and origin_cat &arch interactions.

* Model selection

We first use ggpairs plot to find the variables that are relatively highly correlated to response and include them in our first model. We also include categorical variables that we think might have influence on the reponse. In addition, we use ggplot with different facet to identify interaction. Besides,we also choose variables based on the some information online and the meaning of different variables. Take Surface as an example, we found that for paintings that are not created by famous artist, it is possible that the price is greatly influenced by the surface area of the paintings since people bought them for decoration in the living room or bedroom, different size can lead to different prices. After the first round chosen, we pick about 20 variables includes year, dealer, origin_cat, diff_origin, nfigures, engraved, prevcoll, lrgfont, finished, figures,shape, othgenre, artistliving, type_intermed, Surface, discauth, paired, othartist, endbuyer,Interm etc. We also choose significant variables and tried to see if their interaction is also significant for the model:original:history, Shape:Surface, origin_cat:arch, paired:dealer, engraved:dealer.

Then, we use AIC to further help us choose variables. Variables such as origin_cat, diff_origin, lrgfont, engraved, finished, dealer, year are very important in the AIC selection. We build our model base on these important variables and gradually add more variables that we consider as fit into our model.



- Residual Analysis

  Looking at the Residuals vs Fitted plot, we see that the red line is perfectly flat. There is no discernible

non-linear trend to the residuals. The data appear to be well modeled by a linear relationship between response and predictors, and the points appear to be randomly spread out about the line, with no discerninle non-linear trends or changes in variability. For the Normal Q-Q, residuals are lined well on the straight dashed line, indicates that residuals are normally distributed. The third graph is the Scale-Location plot. This plot indicates that the residuals are spread equally along the ranges of predictors. The forth graph is Residuals vs Leverage. It helps us to find influential cases.There are no points outside the cook's distance and thus there are no influencial points.

- Variables
  Here is the table of our the coefficient of our variables and their confidence intervals.

```
##                        coefficient        2.5 %       97.5 %
## (Intercept)          2.8896422523  1.699269548  4.0800149569
## year1765             1.4783920545  0.981747532  1.9750365765
## year1766             0.2344766913 -0.381563378  0.8505167609
## year1767             1.2459403344  0.932414428  1.5594662407
## year1768             0.1196098120 -0.256650785  0.4958704089
## year1769             2.3862656416  1.816714192  2.9558170911
## year1770             1.2773730768  0.849172943  1.7055732106
## year1771             0.9529344750  0.588624237  1.3172447134
## year1772             0.6938539611  0.164895377  1.2228125454
## year1773             1.1091709463  0.640324954  1.5780169386
## year1774             1.9180826660  1.526273682  2.3098916499
## year1775             0.5567521431 -0.219853268  1.3333575539
## year1776             1.6855911815  1.382620913  1.9885614504
## year1777             2.4657772623  2.142649276  2.7889052490
## year1778             0.5374166904  0.052362262  1.0224711190
## year1779             0.7179040557 -0.052941463  1.4887495742
## year1780             0.3408696864 -0.348202350  1.0299417233
## dealerL              2.4998724635  2.034448912  2.9652960151
## dealerP              1.6647010118  0.898170430  2.4312315940
## dealerR              1.5129687636  1.135526685  1.8904108420
## origin_catF         -0.6662273726 -0.846272755 -0.4861819904
## origin_catI         -0.6853381081 -0.920834403 -0.4498418135
## origin_catO         -0.8457836874 -1.208549546 -0.4830178293
## origin_catS         -1.3761521483 -2.499821043 -0.2524832537
## lrgfont1             0.8587973285  0.605988402  1.1116062547
## Surface              0.0002993593  0.000231515  0.0003672035
## diff_origin1        -0.5096375905 -0.773029535 -0.2462456463
## engraved1           -0.4009058336 -1.698674127  0.8968624603
## endbuyerB            0.8364813497  0.236750963  1.4362117368
## endbuyerC            0.8222946416  0.533049646  1.1115396369
## endbuyerD            0.7529701799  0.508286262  0.9976540981
## endbuyerE            0.4346801235  0.123502612  0.7458576355
## endbuyerU            0.3806528185  0.094596638  0.6667089993
## lands_sc1           -0.4039744311 -0.668339182 -0.1396096799
## finished1            0.5156814235  0.314651561  0.7167112861
## paired1             -0.1130703367 -0.508194019  0.2820533453
## discauth1            0.4122230511  0.117267451  0.7071786515
## type_intermedB       0.4257923852 -0.304444528  1.1560292985
## type_intermedD       0.8682028813  0.561391351  1.1750144119
## type_intermedE       0.3245973687 -0.112464170  0.7616589075
## type_intermedEB      0.4244536773 -1.200428928  2.0493362824
## shape_recodeoval    -0.5081968115 -1.776933375  0.7605397516
## shape_recoderound   -1.2516418098 -2.442751375 -0.0605322448
```

```
## shape_recodesqu_rect -0.7813840374 -1.914425429  0.3516573540
## origin_catD/FL:arch1 -0.3190954495 -0.904708065  0.2665171656
## origin_catF:arch1     0.3373443637 -0.158754313  0.8334430409
## origin_catI:arch1     0.5357195903 -0.579678601  1.6511177815
## origin_catO:arch1    -0.1499111045 -1.148541575  0.8487193663
## origin_catS:arch1               NA           NA            NA
## dealerL:engraved1     2.6290901546  0.099731706  5.1584486032
## dealerP:engraved1               NA           NA            NA
## dealerR:engraved1     1.2426895813 -0.097892162  2.5832713247
## dealerL:paired1      -0.4929005732 -1.045762662  0.0599615159
## dealerP:paired1      -0.2708957275 -0.984302600  0.4425111453
## dealerR:paired1      -0.2953798466 -0.729503271  0.1387435775
```

4. Summary and Conclusions (10 points)

What is the (median) price for the "baseline" category if there are categorical or dummy variables in the model (add CI's)? (be sure to include units!) Highlight important findings and potential limitations of your model. Does it appear that interactions are important? What are the most important variables and/or interactions? Provide interprations of how the most important variables influence the (median) price giving a range (CI). Correct interpretation of coefficients for the log model desirable for full points.

Provide recommendations for the art historian about features or combination of features to look for to find the most valuable paintings.
Centered Surface Model

The median price for the "baseline" category (i.e., when year=1764, dealer=J, origin_cat=D/FL, endbuyer=X, type_intermed="", and all other dummy variables are at level 0) and centered continous variable Surface is eˆ3.08 equls 21.76 livre, got from the value of intercept after the continous variable is centered.The median price in this scenerio ranges from 6.613 livre to 71.88 livre.
Important findings: many variables influence the log price of the paintings, such as the year of sale. At some years, paintings are more popular, while at some years paintings are sold at a lower price. Different type of dealer or the origin of paintings also influence a painting's price. We also notice that bigger painting do will have higher price.
Potential limitations: we only considered linear relationships in this model; we didn't transfer any predictors. This may lead to predictions that are not accurate enough.
Interaction: We find one interaction which is the the interaction between arch,whether architectural construction are mentioned in the painting and origin_cat, which represents origin of painting based on dealers' classification in the catalogue.Another interaction is between dealer and whether the painting is engraved or not. The third interaction is between dealer and whether the painting is paird or not.
Most important variables:Most of our variables in the model are highly significant, indicating that they are all very important in predictions. Specifically,lrgfont has p values less than 2*10ˆ-16.

Interpretation of most important variables
Lrgfont is the most important variables because it has a p values less than 2*10ˆ-16. Besides lrgfont, variables such as surface,dealer and origin_cat are also very important as they have small p values.
According to the coefficient and CI, one increase in surface of painting in squared inches will increase the price by 1.37 livre, the range of increase is from 1.28 livre to 1.48 livre. If the dealer devotes an additional paragraph for the painting, the price will increase by 2.36 livre, the range of increase is from 1.83 livre to 3.039 livre. For dealers, if take dealer J as the baseline, dealer L increases the price by 12.18 livre,dealer P increases the price by 5.28 livre and dealer R increase the price by 4.54 livre.

As a result, we recommend the art historian to look at factors such as the dealer, Surface,lrgfont and the origin of the author when evaluate paintings.

Below are the important coefficients after adjusting from the log model to normal model.

```
##             coefficient2      2.5 %     97.5 %
## dealerL       12.1809403 7.64803622 19.4004452
## dealerP        5.2840931 2.45510721 11.3728802
## dealerR        4.5401896 3.11281258  6.6220888
## origin_catF    0.5136427 0.42901098  0.6149699
## origin_catI    0.5039198 0.39818665  0.6377290
## origin_catO    0.4292209 0.29863011  0.6169188
## origin_catS    0.2525485 0.08209969  0.7768692
## lrgfont1       2.3603203 1.83306312  3.0392363
## Surface        1.3766499 1.28044737  1.4800803
```