STA 521 - Final Project Part I

FP-Team 01: Qianyin Lu, George Lindner, Chenxi Wu, Yi Mi

December 7th, 2019

1. Introduction: Summary of problem and objectives

Our team of esteemed statisticians was recently hired by a prestigious art historian for a consulting project. We were asked to help build a predictive model in exchange for an A on our STA 521 Final Exam. After much discussion, our team accepted the historian's offer.

We were given the task of predicting paintings' selling prices at auctions in 18th century Paris. To accomplish this, we used a dataset containing information about each painting's buyer, seller, painter, and characteristics of the painting. These variables were all possible predictor variables in modeling the response variable, the selling price of a painting.

There were two primary objectives in our analysis:

1) To determine which variables (or interactions) drove the price of a painting
2) To determine which paintings were overpriced or and which were underpriced.

After arriving at a final model, we are able to answer these primary questions. Any variables that appear in the model will be important in driving painting prices, and observing residuals will enable us to determine if a painting was over or underpriced.

We had 1,500 observations to train the model on, along with 750 observations held out as a testing set. There were a total of 59 variables in the dataset, both categorical and continuous.

2. Exploratory Data Analysis:

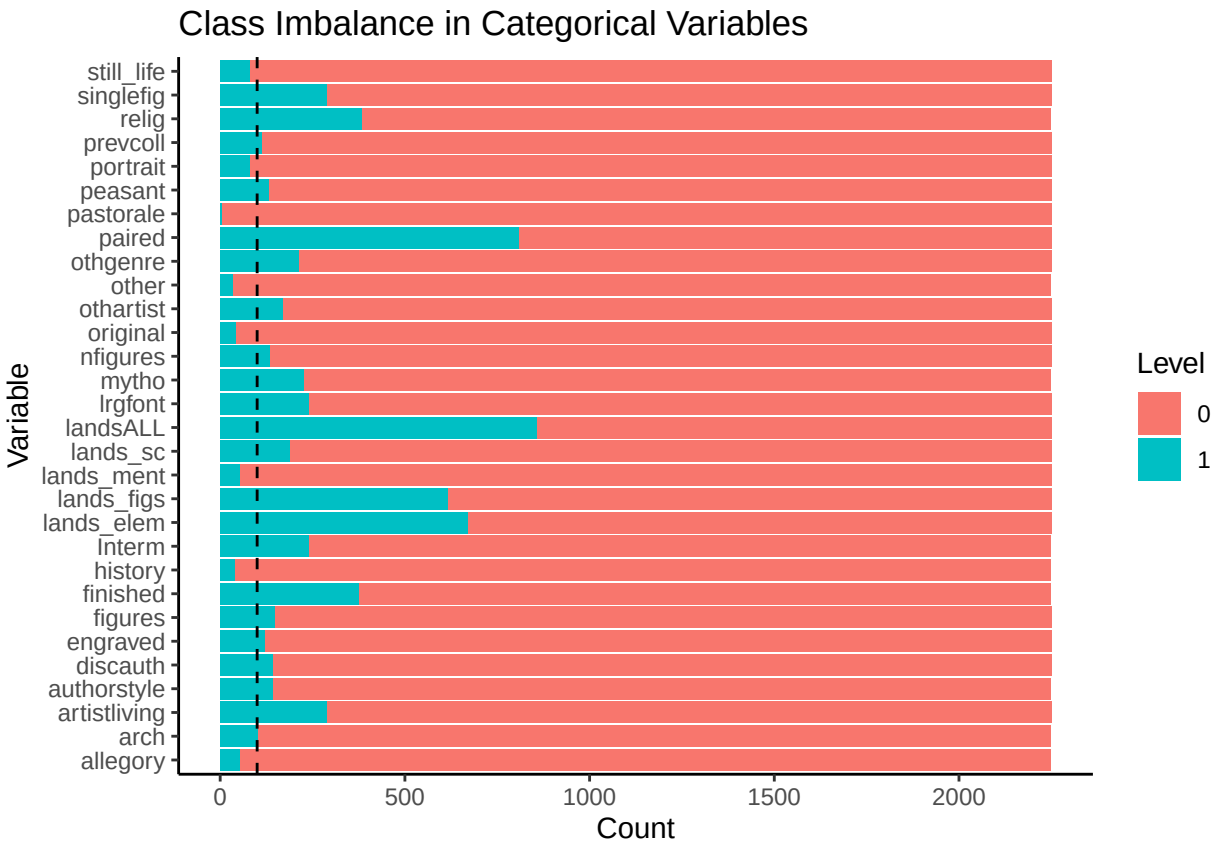
Initial Data Cleaning

We began our data cleaning process by reading the codebook for a better understanding of what each variable in the data represented. Several predictors in the dataset were redundant and therefore removed to avoid high correlation among the predictors. Examples of this include the variable *sale*, which is a combination of *dealer* and *year*. Additionally, there were other predictors that we deemed would not be useful for prediction, such as *count* which was 1 for every observation, or *subject* which was a short description of the content in the painting. We simplified the data by eliminating unnecessary predictors.

Categorical Variables

We recoded each categorical variable to be a factor. We created a visualization of the binary categorical variables to observe the balance between classes below.

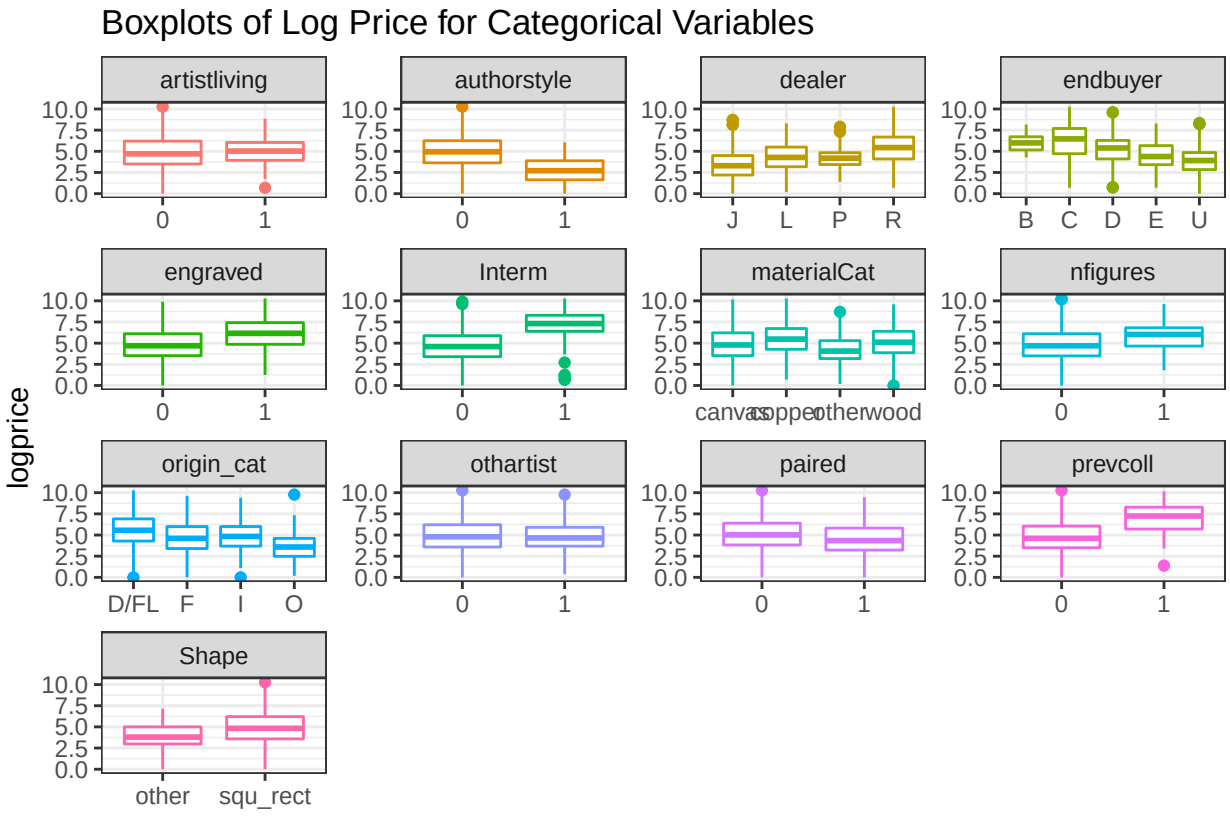
Plot 1

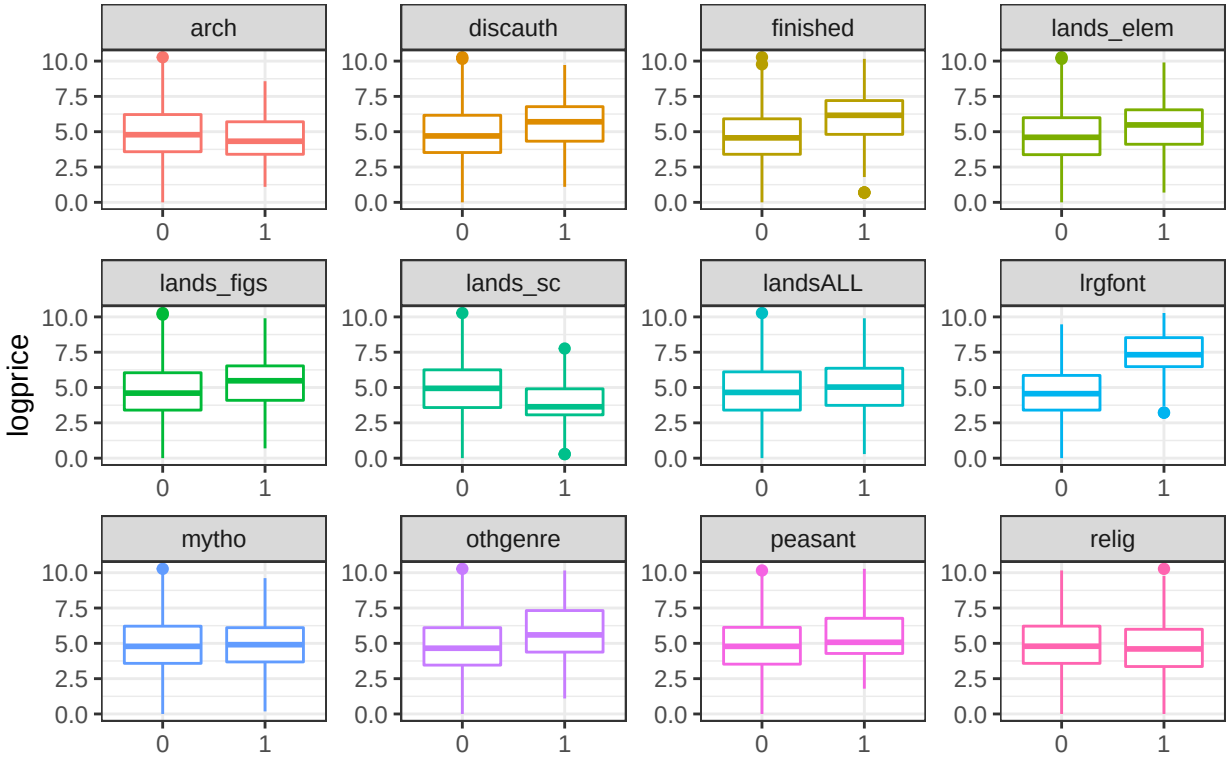


Imbalanced classes can lead to poor β estimates if the underrepresented class does not have enough data. This was our motivation to remove any variable that had less than an arbitrary 100 observations in a class, which is denoted by the dotted black line in our visualization above.

To identify important categorical variables, we created a boxplot for each variable that compared the distribution of *logprice* over every level of the factor. The results are shown below.

Plot 2


Boxplots of Log Price for Categorical Variables
logprice
artistliving
authorstyle
dealer
endbuyer
engraved
Interm
materialCat
nfigures
origin_cat
othartist
paired
prevcoll
Shape
0
1
J
L
P
R
B
C
D
E
U
canvas
copper
otherwood
D/FL
F
I
O
other
squ_rect
10.0
7.5
5.0
2.5
0.0


Boxplots of Log Price for Categorical Variables (continued)
logprice
arch
discauth
finished
lands_elem
lands_figs
lands_sc
landsALL
lrgfont
mytho
othgenre
peasant
relig
0
1
10.0
7.5
5.0
2.5
0.0

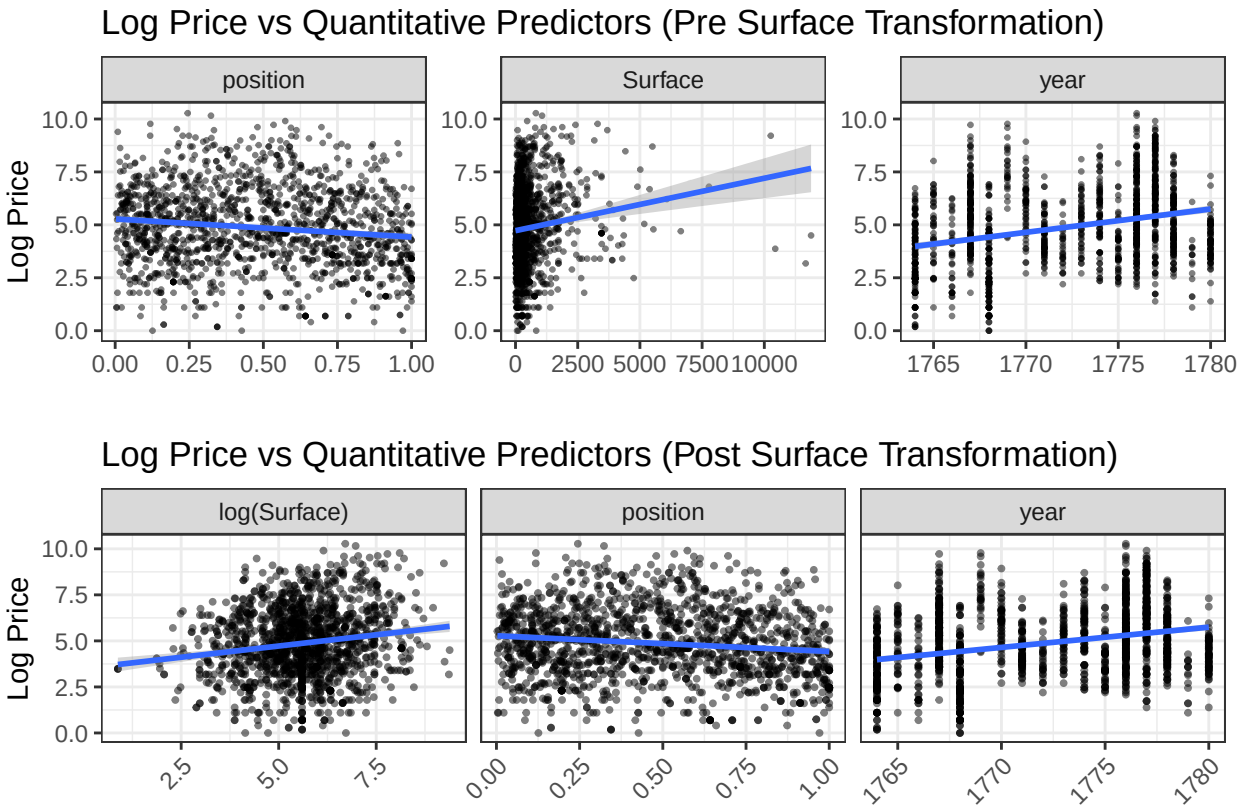
The boxplots above help us identify which variables could be important in predicting a painting's price. They also help us in our variable selection process by displaying variables that have similar prices in all of their categories. After inspecting the boxplots, we determined that *mytho*, *landsALL*, *relig*, and *othartist* were not useful for prediction. Variables that may be important include, but are not limited to, *lrgfont*, *Interm*, *authorstyle*, and *prevcoll*.

Quantitative Variables

There are also quantitative variables in our data that could be used for prediction. Like the categorical variables, many of these predictors were redundant. For example, we were given the surface area of a painting. Additionally, we were given a variable for surface area if the painting was round and a surface area variable if the painting was rectangular. We also were given the height, the width, and the diameter of the painting. We determined that all this information could be condensed to a single variable, *Surface*.

There was missing data in *Surface* that we had to address. Surface area intuitively seems like it could drive the price of a painting, so we had to develop a strategy for handling the missing observations. With the help of the plot below, we determined that imputing the median surface area size of the dataset would be a good estimation for missing values. Since the distribution of *Surface* is skewed, we wanted an imputation strategy that would be robust to outliers. Thus, we opted for the median over the mean.

Plot 3



We created scatterplots to observe the relationship between our three quantitative predictor variables and the log price of a painting. The distribution of *Surface* was skewed right and a log transformation was necessary. We plot the relationship of logprice and the transformed Surface column in the lower graph.

After considering our EDA plots, we determined the 10 variables that we thought would be most useful in predicting *logprice*.

Table 1: 10 Most Important Predictor Variables, from EDA

Rank	Variable
1	log(Surface)
2	lrgfont
3	Interm
4	authorstyle
5	prevcoll
6	origin_cat
7	engraved
8	finished
9	discauth
10	dealer

With the data cleaned and important variables identified, we move to the next step of the process: modeling the data.

3. Development and Assessment of Initial Model:

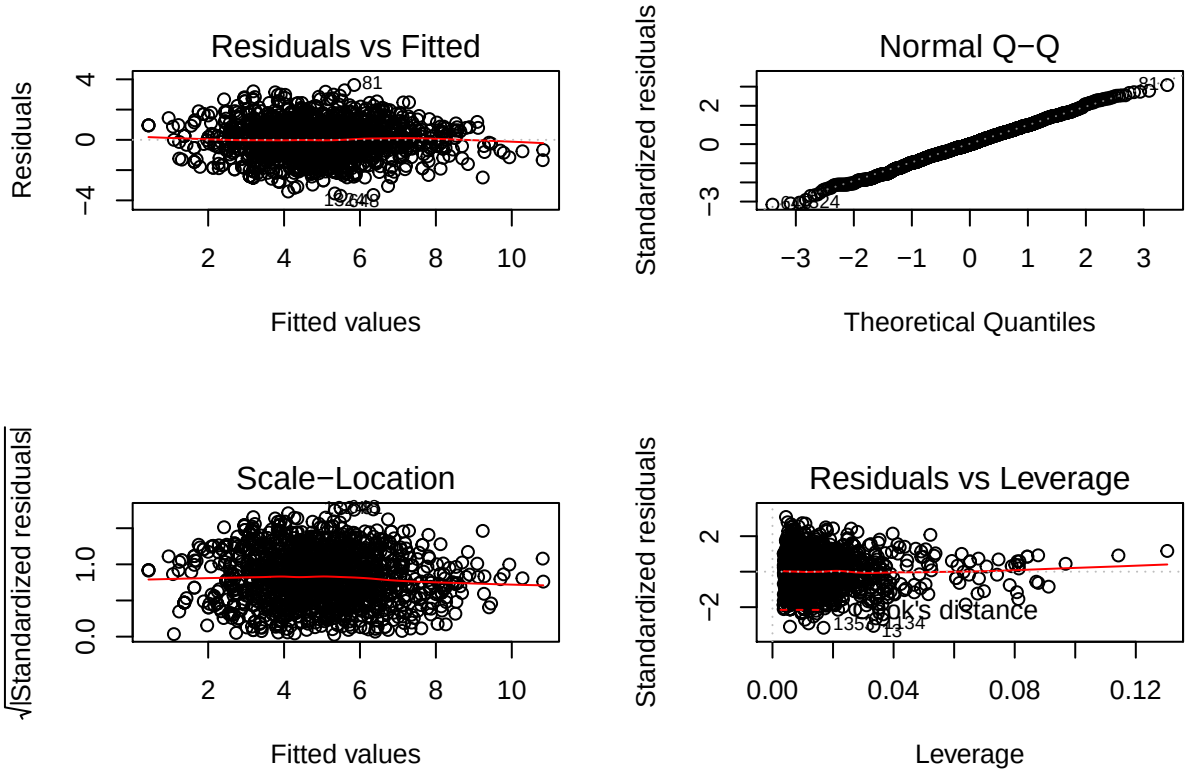
Model Development

We considered three additional variables in addition to the ten most important variables identified through EDA. These variables were *lands_sc*, *endbuyer*, and *year*. We chose to include these three variables because the EDA suggested that they might add some predictive power to our model. Approximately 62.2% of the variation in *logprice* can be explained by the predictors in our initial model, according to the summary output below. Next, we used step-wise variable seleciton with AIC as our criteria to ensure that each variable reduced RSS enough to justify including the variable in the model. Step-wise AIC selection returned our full model, indicating that we did a good job selecting predictor variables through the EDA.

Next, we considered interaction terms for our predictor variables. We again used intuition as our method for introducing interactions in the model. Interactions that we considered were *authorstyle* with *log(Surface)*, *Interm* with *log(Surface)*, and *discauth* with *log(Surface)*. We chose the first term because both the style of a painting and the size of the painting could be important in determining the price. Various styles of paintings might increase in value at different rates as the size of the painting changes. The second interaction term considers the dealer engaging with the authenticity of the painting and the size of the painting. Authentic paintings could increase in value as the size increases at a different rate than non-authentic paintings. We also thought that whether an intermediary was involved could be an important interaction with the size of the painting.

We fit our initial model on these predictors and interaction terms. Included below are our model plots and the summary of the model.

Model Plots



Model Summary

```
##
## Call:
## lm(formula = logprice ~ log(Surface) + lrgfont + Interm + authorstyle +
##     prevcoll + origin_cat + engraved + finished + discauth +
##     dealer + lands_sc + endbuyer + year + authorstyle:log(Surface) +
##     Interm:log(Surface) + log(Surface):discauth, data = paint_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6963 -0.7402 -0.0144  0.7876  3.6259
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -2.210e+02  1.290e+01 -17.130  < 2e-16 ***
## log(Surface)                3.530e-01  2.871e-02  12.294  < 2e-16 ***
## lrgfont1                    8.354e-01  1.207e-01   6.919 6.76e-12 ***
## Interm1                     7.832e-02  4.722e-01   0.166 0.868284
## authorstyle1               -7.890e-01  6.314e-01  -1.250 0.211633
## prevcoll1                   8.159e-01  1.426e-01   5.720 1.29e-08 ***
## origin_catF                -6.358e-01  7.643e-02  -8.318  < 2e-16 ***
## origin_catI                -7.276e-01  1.064e-01  -6.838 1.17e-11 ***
```

```
## origin_cat0               -8.452e-01  1.103e-01  -7.665 3.23e-14 ***
## engraved1                  7.141e-01  1.438e-01   4.964 7.70e-07 ***
## finished1                  8.011e-01  9.086e-02   8.817  < 2e-16 ***
## discauth1                  1.053e+00  6.603e-01   1.595 0.110873
## dealerL                    1.298e+00  1.287e-01  10.083  < 2e-16 ***
## dealerP                    3.154e-01  1.586e-01   1.988 0.046973 *
## dealerR                    1.791e+00  1.032e-01  17.357  < 2e-16 ***
## lands_sc1                 -4.232e-01  1.161e-01  -3.645 0.000276 ***
## endbuyerC                 -2.950e-01  3.254e-01  -0.907 0.364777
## endbuyerD                 -5.030e-01  3.230e-01  -1.557 0.119642
## endbuyerE                 -8.880e-01  3.351e-01  -2.650 0.008131 **
## endbuyerU                 -1.107e+00  3.247e-01  -3.410 0.000666 ***
## year                       1.261e-01  7.266e-03  17.350  < 2e-16 ***
## log(Surface):authorstyle1 -4.493e-02  1.074e-01  -0.418 0.675698
## log(Surface):Interm1       1.090e-01  8.163e-02   1.335 0.182025
## log(Surface):discauth1    -9.104e-02  1.122e-01  -0.811 0.417352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.18 on 1476 degrees of freedom
## Multiple R-squared:  0.6276, Adjusted R-squared:  0.6218
## F-statistic: 108.1 on 23 and 1476 DF,  p-value: < 2.2e-16
```

From the residual vs. fitted plot, we see that our residuals are randomly distributed with mean 0. There is no heteroskedacity satisfying the constant variance assumption of linear regression. Our QQ plot appears approximately normal, as well. The residuals vs fitted plot shows that there are no high leverage points, influential points, or outliers. We can see from the summary output that approximately 62.8% of the variation in *logprice* can be explained by our model.

4. Summary and Conclusions: