Final Project Part 1

Team_FP04

12/6/2019

Introduction

The 18th century is often refered to as "le siècle, des Lumières" (the century of the lights) in reference to the philosophers that emerged early on the 1700s leading the way towards the French Revolution. In addition to its important societal evolutions, the 18th century was also a major period for art in France and it is therefore of interest to understand painting trading during that period, especially before the French Revolution (1789). The aim of our analysis is to explore the factors that drove painting prices in 18th century Paris. The painting prices will be predicted from auction price data between 1764-1780 containing information on the sale (seller/buyer), the artist, and other characteristics of the painting. This analysis will also allow us to assess which paintings were overvalued or undervalued. We will first explore the effects of potential predictors and their interactions on painting prices by conducting an Exploratory Data Analysis. This will also allow us to prepare the data for the next phase. We will then build a linear model using stepwise regression method with akaike information criterion (AIC) and a training subset to select a robust model predicting the auction price (using the log transformation `logprice`). We will then validate our model on a test subset.

EDA

We start by exploring variables in the training dataset and understand their meaning. We first implement the required data pre-processing:

- transform empty string and "n/a" character to NA
- Delete duplicate rows
- Transform binary and charcter variables to factors
- Change `position` values not bounded from 0 to 1 to NA
- Reconcile `Shape` coding (ovale = oval, round = ronde)
- Impute the missing data

Imputation: We impute the mean of each quantitative column except Surface (heavily skewed right) for which we impute its median. For the binary and multiple level factor variables, we choose to impute the mode of each column.

we then see that some variables were used to classify each painting and therefore cannot be used in our analysis: `sale`, `lot`, `count`, `subject`, `authorstandard`, `author`, `subject`, `authorstyle`, `winningbidder`

In addition, since we are predicting `logprice`, we will not used variable `price`.

We can classify the remaining variables in two different ways, either by the way they are coded (quantitative, dumy, multiple level factors) or by the information they provide (i.e. sale, author, size & material or characteristics). First, we goup variables according to the way they are coded:

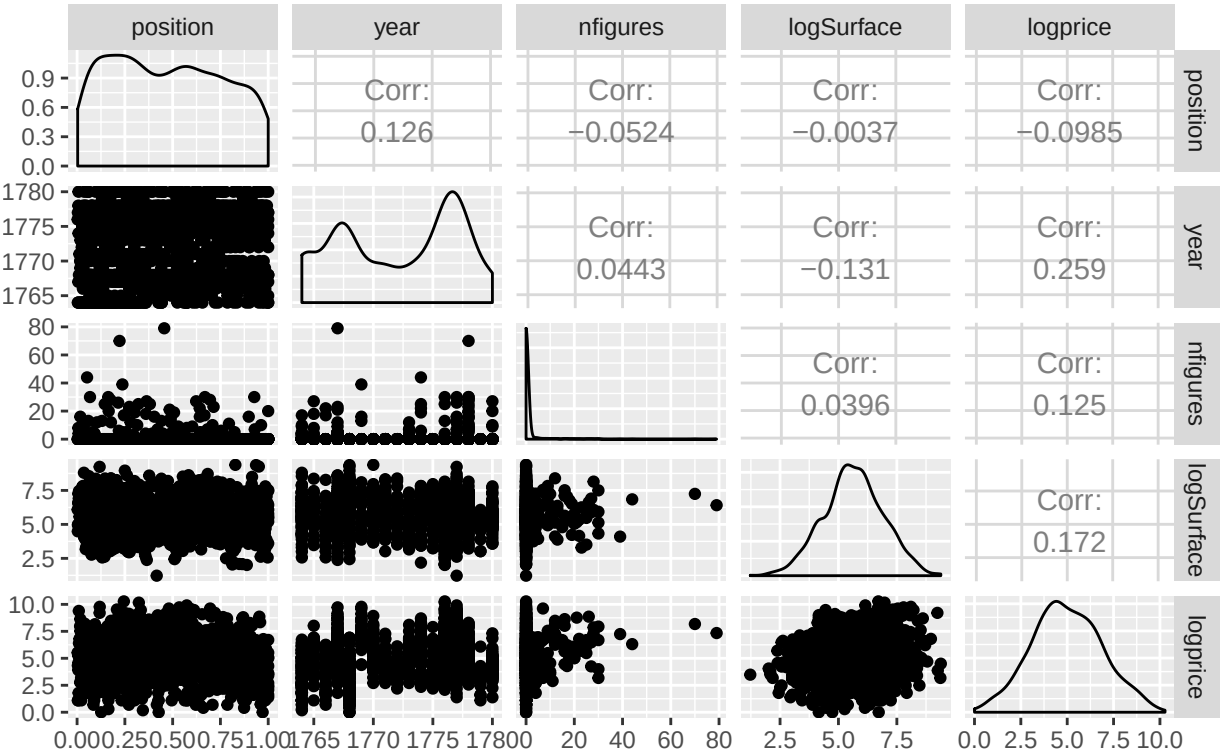
Quantitative Variables

`position`, `year`, `Height_in`, `Width_in`, `Surface_Rect`, `Diam_in`, `Surface_Rnd`, `Surface`, `nfigures`

We decide to classify `year` as numeric in our analysis as it is spread around more than twenty years. Using our intuition, we choose to only use `Surface` and drop `Height_in`, `Width_in`, `Surface_Rect`, `Diam_in` and `Surface_Rnd` as they are extremely correlated and would not necessarily bring any additional information. We might want to later investigate the relationship of `Surface` with other variables such as `Shape` or the type

of material. We use a scatterplot matrix to investigate the relationship between these quantitative variables and `logprice` (FIG 1).

FIG 1: Relationship of relevant quantitative variables

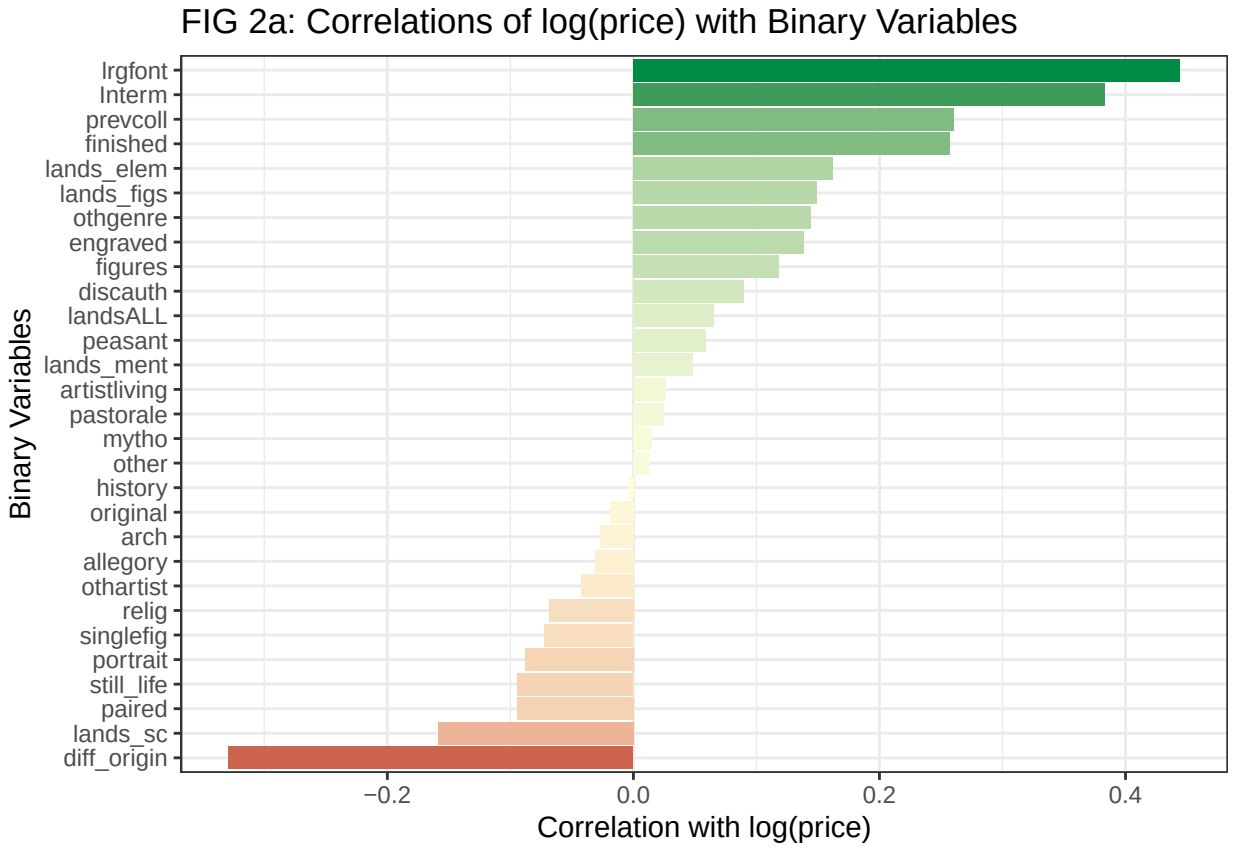


We see that the quantitative predictors plotted here are not really correlated with each other. Variable `position` only has a small negative and not necessarily linear correlation with the response variable. `year` has a stronger effect on `logprice` (.26) and we could consider a sort of overall "inflation" on paintings even though it is not linear nor monotonic. `Surface` is also positively correlated with `logprice` (.17). Note here that we used a log transformation on `Surface` in order to make its relationship with `logprice` linear. Finally, variable `nfigures` behave in a strange way. While having no figure does not seem to give any information on `logprice`, we can see that for paintings with at least one figure, more figures is correlated with higher price. It will be interresting to explore the interraction of `nfigures` with some of the binary predictor variables that we explore next.

Binary Variables

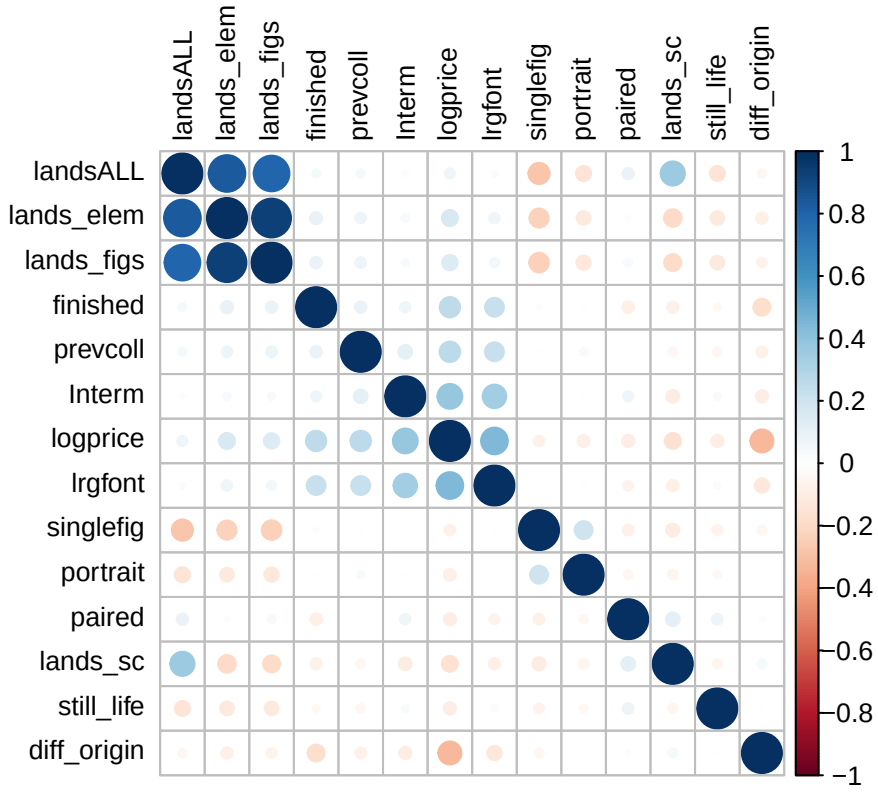
`diff_origin`, `artistliving`, `Interm`, `figures`,`engraved`, `original`, `prevcoll`, `othartist`, `paired`, `finished`, `lrgfont`, `relig`, `landsALL`, `lands_sc`, `lands_elem`, `lands_figs`, `lands_ment`, `arch`, `mytho`, `peasant`, `othgenre`, `singlefig`, `portrait`, `still_life`, `discauth`, `history`, `allegory`, `pastorale`, `other`

We present correlations between all the binary variables and log(price) (FIG 2a). Unsurprisingly, an additional paragraph in a larger font has a high correlation with price, suggesting that these paintings were the highlights in the different auctions. An intermediary also suggests a higher price, maybe because these individuals are involved mostly in high stakes sales. A mention of the previous owner and having a highly polished finishing are also factors that seem to drive up the price. Factors the drive the prices down are different origin of author and painting, if the content of the painting includes a "plain landscape", if the painting is just a "pairing" of another art work, and if the content is still life.



Once we have an idea of the top binary variable candidates to include in our model, it is important to take a look at their correlations within themselves and with other variables. Fig 2b is a correlation matrix of select binary variables. From this plot, it is clear that there is no need to include more than one variable among `landsALL`, `lands_elem` and `lands_figs` as they are all related in content, highly correlated and have similar effect on prices. `singlefig` is strongly negatively correlated with the above three, and negatively correlated with price, so we should also consider dropping it if we select one of the others. Regarding `lrgfont` and `interm`, although both of them have a high correlation with prices, they are also correlated with each other, so we might consider choosing one of them.

FIG 2b: Correlation plot of log(price) and select binary variables

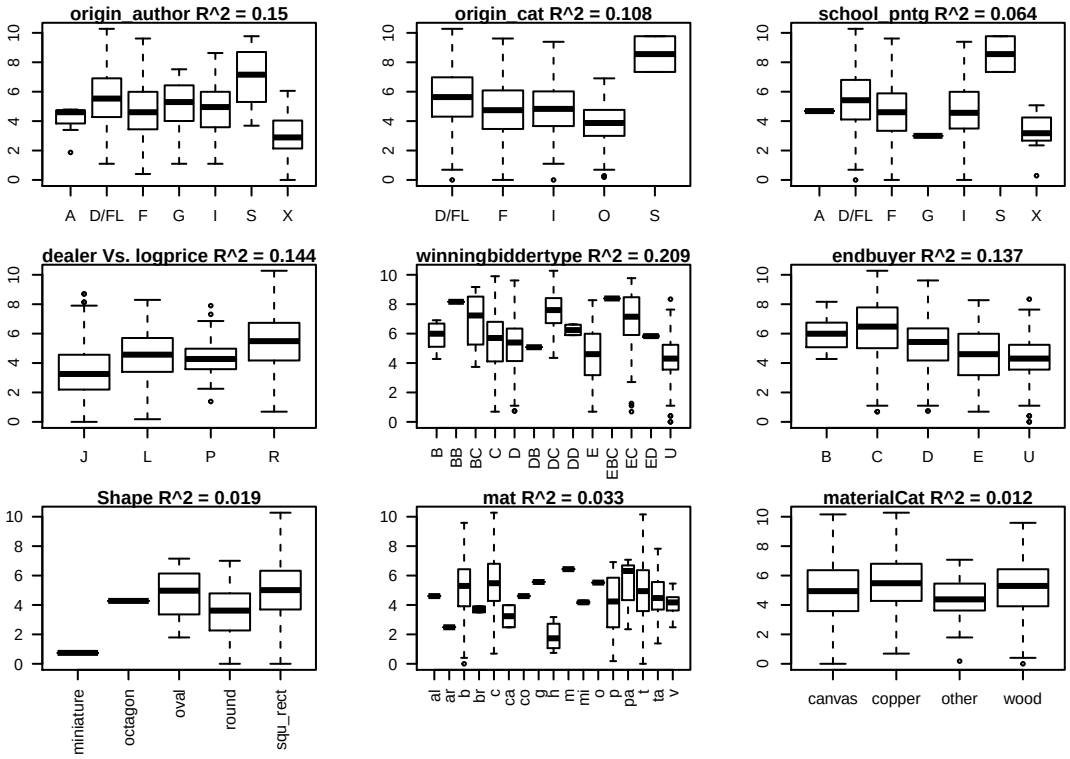


Multiple Level Factor Variable

`origin_author`, `origin_cat`, `school_pntg`, `Shape`, `material`, `mat`, `materialCat`, `dealer`, `winningbiddertype`, `endbuyer`, `type_intermed`

When investigating the material variable, we decide to exclude `material` as it has too many levels. As it has an important number of NA values, `type_intermed` is also ignored for now but will be considered later for interactions. (FIG 3)

FIG 3: Distribution of log(price) across multilevel factors



Using intuition, we could classify the first three plots regarding `origin_author`, `origin_cat` and `school_pntg` as information about the author. We expect these variables to be correlated with each other and therefore only using one of them would most likely give us enough information. We decide here to select `origin_author` as it has the highest r.squared (.15). The next three plots give us information about the sale of each painting. Looking at the `dealer` plot, we see that sale prices seem to be a little different across dealers. This might be explained by the kind of painting they each sale or by the kind of client they reach to. Now focusing on the `winningbiddertype` and `endbuyer` plots, we can deduce that these two variables inform on the buyer and so are probably highly correlated with each other. We can observe some differences accross buyers. We think that an explaination for these differences might be the intervention of intermediaries. Finally, the last three plots can be categorized as shape and material. These plots only explain an insignificant amount of the variance in `logprice`. However, the effect of a different shape as well as the interaction between the `Shape` and `Surface` might be worth looking at in our model building. Regarding material, the interaction with `Surface` could be of interest.

From our EDA, we are able to extract what we consider the 10 most important variables to predict `logprice`. Looking at the quantitative variables `year` and `Surface`, one can see that they are quite strongly correlated with our response variable (r = .26 and .17) and bring information about price evolution across years and accross the overall size of the painting. For the variables giving us information on the author, `diff_origin` has a relatively strong correlation (about .33) with `logprice`. Variables `origin_author`, `origin_cat` and `school_pntg` are correlated with each other and some of their imformation is already carried by `diff_origin`. Therefore, we decide to choose origin_author as it has the strongest r-squared. When looking at the sale of the paintings, one can observe that `winningbiddertype` has an r-squared of .21 when regressed on `logprice`. However, we can assume that it will be strongly correlated to `interm` and therefore select `endbuyer` (r.sq = .14) instead. Variable `interm` (presence of an intermediary) is also quite strongly correlated with the response (about .38). Despite its high correlation with other variables included in this list, we still believe that it provides important information. Finally, when investigating characteristics of paintings, variables `lrgfont` is the one with the highest correlation with the response (>.4) and despite its correlation with `interm`, we consider it as an important predictor. Variables `lands_sc`, `prevcoll`, `finished` are also considered important

in predicting `logprice` with correlations between .15 and .25 and a low correlation with the other variables selected.

Our 10 variables:
`lands_sc`, `prevcoll`, `finished`, `lrgfont`, `origin_author`, `endbuyer`, `interm`, `Surface`, `year`, `diff_origin`

Model Development and Assessment

Initial Model

Table 1: Initial Model Summary

	# Coefficients	Res Sd Error	df (n-k)	Adjusted r.sq	F stat
value	69	1.1676	1284	0.6191	33.3206

In order to select our initial model with 10 to 20 variables, we first rely on our EDA and include the variables we selected as the 10 most important. We then look at what kind of new information some additional variables could bring about the response even with a low correlation.
Regarding the sale of the painting, we believe that knowing if the artist was alive at the time of the sale, what dealer sold the painting or if they engage with the authenticity of the painting might provide us with new information regarding `logprice`. Focusing on `Shape` and material, we imagine that knowing if a painting is round, oval or recangular may impact `logprice` at least a little bit and so would the use of different material. We decide to use `material_Cat` here as `mat` has a lot of different levels that could produce NA coefficients and irregularities with the test set. Now, exploring predictors related to the characteristics of the paintings, we know that `lands_figs` and `lands_elem` are correlated with each other and `logprice`. Therefore, we decide to only choose `lands_fig` as `lands_elem` also provides information aout `lands_ment`, a variable not so correlated with the response. Finally, `peasant`, `engraved`, `portrait` and `still-life` are all quite correlated with `logprice` and give us more specifics about the paintings that we have not covered in our model yet. Variable `paired` was not used as we think that information it provides might already be covered by variable such as `lrgfont`, `dealer` or `endbuyer`. We also believe that most of the predicting power brought by `othrgenre` is supported by `peasant`.

When thinking about the interactions, we try to use our intuition. The effect of `Shape` on the response might depend on the material used. In addition, the effect of an intermediary on prices might be different depending on the `dealer` or the `endbuyer`. In the same way, prices might change depending on who a specific dealer treats with or the origin of the author of a painting a specific dealer is trading. Finally, The price of a non finished painting might be different if the previous owner is known or not, for transparency reasons. After investigation, no interaction with `nfigures` seemed relevant in the model.

This initial model explains 61.91% of the variance in `logprice`

Model Selection

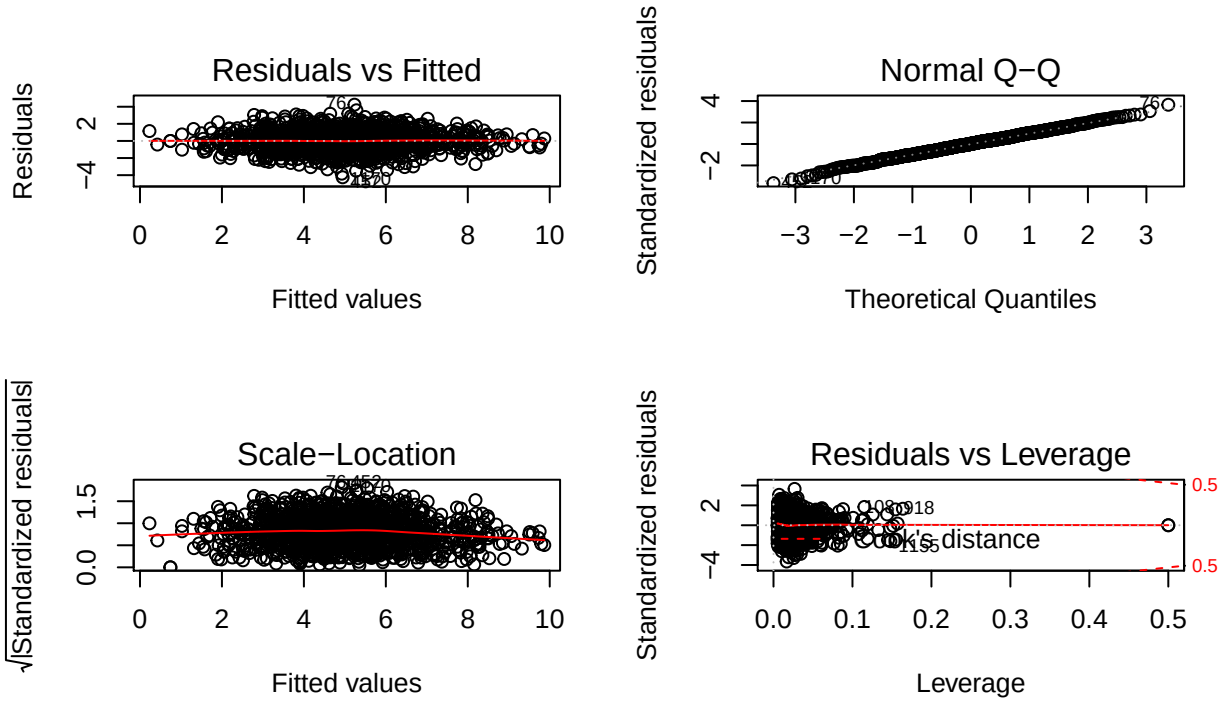
Table 2: AIC-selected Model Summary

	# Coefficients	Res Sd Error	df (n-k)	Adjusted r.sq	F stat
value	36	1.1753	1317	0.6141	62.4715

From this initial model, we run a stepwise selection algorithm (both directions) to reduce the number of coefficients, using AIC as our selection criteria. Since our initial model relies on exploratory data analysis to determine which covariates seem significant, this procedure gives us a more rigorous way to select the

most important covariates and interactions in the dataset. Furthermore, given the amount of coefficients in the initial model, it would not have been feasible to check all subsets, so a stepwise procedure is the most efficient option. This selection successfully simplifies the model, reducing the number of coefficients from 69 to 36; the adjusted R^2 also decreases slightly, from 0.6191 to 0.6141, so this much simpler model is basically explaining the same amount of variability in price.

Residual Plots



Residual plots for AIC−selected model

From these residual plots of the simplified model, it appears that all the assumptions of linear regression are met. The first plot shows residuals are independent and homoskedastic, and the second plot shows the residuals are very close to normally distributed. There are a few possible outliers and two leverage points (at about $h_{ii} \approx 0.5$ and $h_{ii} \approx 1$). From the plot, the first point does not appear influential, but the second point might be. However, the Cook's distance for both is under 0.015 (and removing these points does not really change the model), so for the sake of this analysis, we do not removed any observation.

Included Coefficients

Below are the coefficient estimates and respective confidence intervals for our final model:

Table 3: Coefficients for AIC-selected model

	Estimate	P-value	2.5 %	97.5 %
(Intercept)	-208.2675	0.0000	-235.2387	-181.2964
year	0.1179	0.0000	0.1027	0.1331
diff_origin1	-0.4232	0.0001	-0.6324	-0.2139
Interm1	0.7567	0.0000	0.4908	1.0227
engraved1	0.6935	0.0000	0.4028	0.9841
prevcoll1	1.0802	0.0000	0.7414	1.4190

	Estimate	P-value	2.5 %	97.5 %
finished1	0.8744	0.0000	0.6834	1.0654
lrgfont1	0.9185	0.0000	0.6793	1.1578
lands_sc1	-0.4697	0.0002	-0.7148	-0.2246
peasant1	0.2038	0.1500	-0.0737	0.4813
portrait1	-0.4711	0.0066	-0.8108	-0.1314
still_life1	-0.5207	0.0039	-0.8738	-0.1677
discauth1	0.4393	0.0018	0.1633	0.7152
dealerL	1.1381	0.0000	0.8742	1.4019
dealerP	0.3487	0.0367	0.0215	0.6758
dealerR	1.8321	0.0000	1.6200	2.0441
materialCatcopper	0.4287	0.0010	0.1736	0.6838
materialCatother	0.0373	0.8622	-0.3839	0.4584
materialCatwood	0.2809	0.0014	0.1083	0.4535
endbuyerC	-0.2310	0.4781	-0.8699	0.4078
endbuyerD	-0.5913	0.0658	-1.2213	0.0387
endbuyerE	-0.7388	0.0286	-1.4000	-0.0776
endbuyerU	-0.7595	0.0226	-1.4124	-0.1067
Shapeoctagon	2.1598	0.1386	-0.6992	5.0189
Shapeoval	1.4405	0.1068	-0.3107	3.1917
Shaperound	0.6565	0.4574	-1.0761	2.3891
Shapesqu_rect	1.2109	0.1572	-0.4674	2.8891
origin_authorD/FL	-0.1252	0.7854	-1.0275	0.7770
origin_authorF	-0.7019	0.1284	-1.6068	0.2030
origin_authorG	-0.1196	0.8136	-1.1146	0.8754
origin_authorI	-0.8515	0.0692	-1.7699	0.0670
origin_authorS	-0.3891	0.5102	-1.5480	0.7698
origin_authorX	-1.2035	0.0094	-2.1112	-0.2958
artistliving1	0.3687	0.0005	0.1605	0.5768
log(Surface + 1)	0.4056	0.0000	0.3432	0.4681
prevcoll1:finished1	-1.1479	0.0003	-1.7729	-0.5229

Summary and Conclusions

By observing the p-values of the coefficients, it is clear that most of the them are highly statistically significant. Some, mostly related to the shape of the painting and the origin of the painter, are not statistically significant. We decide to keep them in the model due to their possible correlation with other variables. Some of the most important variables and interactions in our model are `year`, `diff_origin`, `Interm`, `engraved`, `prevcoll`,`finished`, `lrgfont`, `dealerL`, `log(Surface + 1)` and `prevcoll1:finished1`.

The interpretation of the effects should take note that the response is in log. So, for a few examples: an additional year in the date of the sale is relatd to a price increase of 10 to 13 percent, keeping other variables constant. For another example, When the origin is different from the origin based on dealers' classification,the painting price would be 0.65 times the price of the painting when two origins are the same, giving a range from 0.54 to 0.81, keeping other variables constant.The price of painting with an intermediary (`Interm`) would be 2.13 times the price of the painting without an intermediary giving a range from 1.63 to 2.77, keeping other variables constant. Lastly, the price of painting with engravtion would be 1.99 times the price of the painting without engravtion giving a range from 1.49 to 2.66, keeping other variables constant. Regarding interactions, only one was kept in the selection process, and therefore the ones we chose were not that helpful in predicting price.

The regression includes dummies and categorical factors, while dropping the "baseline" dummies. The baseline group in our model is paintings where:

- diff_origin = 0
- Interm = 0
- engraved = 0
- prevcoll = 0
- finished = 0

- lrgfont = 0
- lands_sc = 0
- peasant = 0
- portrait = 0
- still_life = 0
- discauth = 0
- dealer = J
- materialCat = canvas
- endbuyer = B
- Shape = miniature
- origin_author = A
- artistliving = 0

The median year in our sample is 1773 and the median `log(surface+1)` = 5.6. For a "median" observation, the predicted price is $e^{(-208.26+0.1179*1773+0.405*5.6)} = 21$ livres.

Our model is limited in various ways. First, it is a linear model and therefore cannot take into account more complex relations between the variables. Second, since most of the variable selection was done manually, it might not be the best selection. Probably more interactions and other variables can improve the fit. Lastly, the predictions are not limited in any way. While the range of the prices in the training data is between 1 and 29K, the upper limit bound of our prediction is 196K. More sophisticated models might be able to improve on that.

Meanwhile, our recommendation for the art historian is that paintings that are sold in later years, that had a second paragraph in a bigger font, and an intermediary was involved, were probably sold for a higher price. The identity of the dealer is important as well, and so is the content of the painting. French buyers in the 18th century, it turns out, did not really value "plain landscapes".