# Final Data Analysis Project

*See Parts for Write-Up due Dates*

1. Introduction (1 point if improved from before)

We were hired as statistical consultants by an art historian to explore (1) what drove the prices of 18th century paintings in France, and (2) which paintings may be overvalued or undervalued. They have provided us with auction price data from 1764-1780 on the sales (seller/buyer), painter, and other characteristics of paintings.

In part 1, we perform exploratory data analysis to gain high-level insights into the art data, to help us inform further portions, and run a simple model based on these observations. Our objectives of part 1 include:

1. Subsetting data to only consider variables that are both relevant and not redundant of other explanatory variables. For instance, the data contains the variables `material`, `mat`, `materialCat`. These three variables are all conveying similar information, but expressed differently. Which is most appropriate to include in the model?
2. Transforming key text variables to be standardized and in a form that we can pass into regressions and other predictive modeling algorithms. For instance, if we wanted to include the painter (`authorstandard`) in our model, we first may want to remove stop words in the variable such as ("in the taste of"), and may want to count a collaboration between multiple painters (separated by a semicolon in the data) into each individual painter level, rather than count it into a new level of the variable expressing the collaboration separately from the individual painter levels.
3. Impute values of data when they are missing. For instance, the surface area (`Surface`) does not always exist, so we impute the missing data conditional on variables such as `Shape`, `Height_in`, `Width_in`, and `Diam_in` that may give insight into `Surface`.
4. Examining what sources of variation are associated with variations in price for 18th century art sold in France. For instance, given a specific painter, do prices vary significantly? Is the variation for paintings within-painter variation, or across-painter variation?
5. When controlling for all other variables in the dataset, what which variables are most important to predicting the log of the price of paintings. For instance, what is the marginal impact in driving price if it is an Adrien van de Velde painting, holding all other variables constant?
6. Based on the EDA, run an initial model, and examine the in-sample residuals and coverage.

In part 2, we move beyond just using linear models, and try to evaluate art prices and understand the importance of different variables with more complex methods. Specifically, our added objectives for part 2 include:

7. Examine if we should make any variable transformations, interactions, or further refined text cleaning/grouping for our final dataset in part 1.
8. Implement and evaluate model types aside from linear regression to the data, with the intention of further improving test set RMSE and coverage while still maintaining a lot of the interpretability in part 1 (for the purpose of communicating trends and big pictures ideas to our client).
9. Exploratory data analysis (1 point if improved from before):

We once again explain in Parts (a)-(e) what we did for cleaning in part 1 of the assignment. If you read part 1, feel free to skip these parts and instead look at the update.

**Part (a) Remove redundancies**

There are many redundant variables in the data, so for the purposes of illustration, we show below a sample variable selection choice to remove a redundancy. Note that we used a similar process to remove further redundant variables.

Table 1: Winning Bidder Type vs. End Buyer

| | | B | C | D | E | U |
|---|---|---|---|---|---|---|
| | 395 | 0 | 0 | 0 | 0 | 0 |
| B | 0 | 12 | 0 | 0 | 0 | 0 |
| BB | 0 | 1 | 0 | 0 | 0 | 0 |
| BC | 0 | 0 | 10 | 0 | 0 | 0 |
| C | 0 | 0 | 189 | 0 | 0 | 0 |
| D | 0 | 0 | 0 | 464 | 0 | 0 |
| DB | 0 | 1 | 0 | 0 | 0 | 0 |
| DC | 0 | 0 | 89 | 0 | 0 | 0 |
| DD | 0 | 0 | 0 | 4 | 0 | 0 |
| E | 0 | 0 | 0 | 0 | 127 | 0 |
| EBC | 0 | 0 | 1 | 0 | 0 | 0 |
| EC | 0 | 0 | 37 | 0 | 0 | 0 |
| ED | 0 | 0 | 0 | 2 | 0 | 0 |
| U | 0 | 0 | 0 | 0 | 0 | 168 |

We see that the last digit of `winningbiddertype` always corresponds to the `endbuyer`. We also see, however, that many of the `winningbiddertype` categories are scarcely populated (`EBC`, for instance, only has 1 value in the training data). Therefore, we use `endbuyer`, and exclude `winningbiddertype` after this step.

The following are the variables we excluded (with the variable that conveys similar information in the parentheses) : `origin_author`, `school_pntg`, `diff_origin` (`origin_cat`); `author` (`author_standard`); `winningbiddertype` (`endbuyer`); `type_intermed` (`Interm`); `Surface_Rect`, `Surface_Rnd`, `Height_in`, `Width_in`, `Diam_in` (`Surface`); `materialCat`, `mat` (`material`); `nfigures` (`singlefig`, `figures`); `landsALL` (`lands_*` variables); `lands_ment` (`lands_sc`, `lands_figs`); `lands_elem` (`lands_sc`, `lands_figs`).

**Part (b) Cleaning text data**

For the purposes of illustration, we only go through the cleaning process for the `authorstandard` variable, though a similar process was applied to `material`, `year`, and `Shape`.

Without any cleaning, it is a good idea to see a set of sample values of the text variable. Below is the 10 most common values for the authors, uncleaned:

Table 2: 10 most Frequent Authors, Uncleaned

| authorstandard | total |
|---|---|
| Teniers II the Younger, David | 44 |
| Anonymous | 43 |
| Wouwerman, Philips | 28 |
| Boucher (F), Franois | 26 |
| Rijn, Rembrandt Harmenszoon van | 21 |
| La Fosse, Charles de | 19 |
| Bourdon, Sbastien | 17 |
| French | 17 |
| Machy, Pierre-Antoine de | 17 |
| Patel I, Pierre | 17 |

To start, we noticed that some `authorstyle`s were included in the `authorstandard` description, where an `n/a` value corresponds to an original by the said author, and all other values correspond to derivations of the author's original work:

Table 3: Types of Derivations of Painters' Work

| Author Style | Frequency |
|---|---|
| after | 26 |
| attributed to | 7 |
| copy after | 10 |
| esteem of | 1 |
| in the genre of | 3 |
| in the manner of | 7 |
| in the style of | 7 |
| in the taste | 1 |
| in the taste of | 17 |
| n/a | 1417 |
| school of | 2 |
| sketch of | 1 |
| taste of | 1 |

For the latter group, we marked all such rows as a copy (by creating a variable `is_copy`), and removed these keywords from the author. Additionally, there are some instances where multiple authors contributed to a painting. As a result, we created a binary variable for each other who contributed to at least 10 paintings, where a 1 indicates contribution. Below, observe the sample counts for the most frequent set of authors:

Table 4: 10 most Frequent Authors, Cleaned

| authorstandard | total |
|---|---|
| teniers_ii_the_younger_david | 51 |
| anonymous | 43 |
| wouwerman_philips | 35 |
| boucher_f_fran_u_fffd_ois | 30 |
| breughel_i_the_elder_j_jan | 26 |
| rijn_rembrandt_harmenszoon_van | 24 |
| la_fosse_charles_de | 19 |
| patel_i_pierre | 19 |
| machy_pierre_antoine_de | 18 |
| poelenberch_cornelis_van | 18 |

While a lot of the most frequent are the same, we see that some, such as Jan Breughel the Elder, now appear in the data (he contributed with many others in the dataset), and others have an increased count, such as David Teniers the Younger.

### Part (c) Imputing missing data

It is a good idea to examine which variables contain missing data, and how frequently:

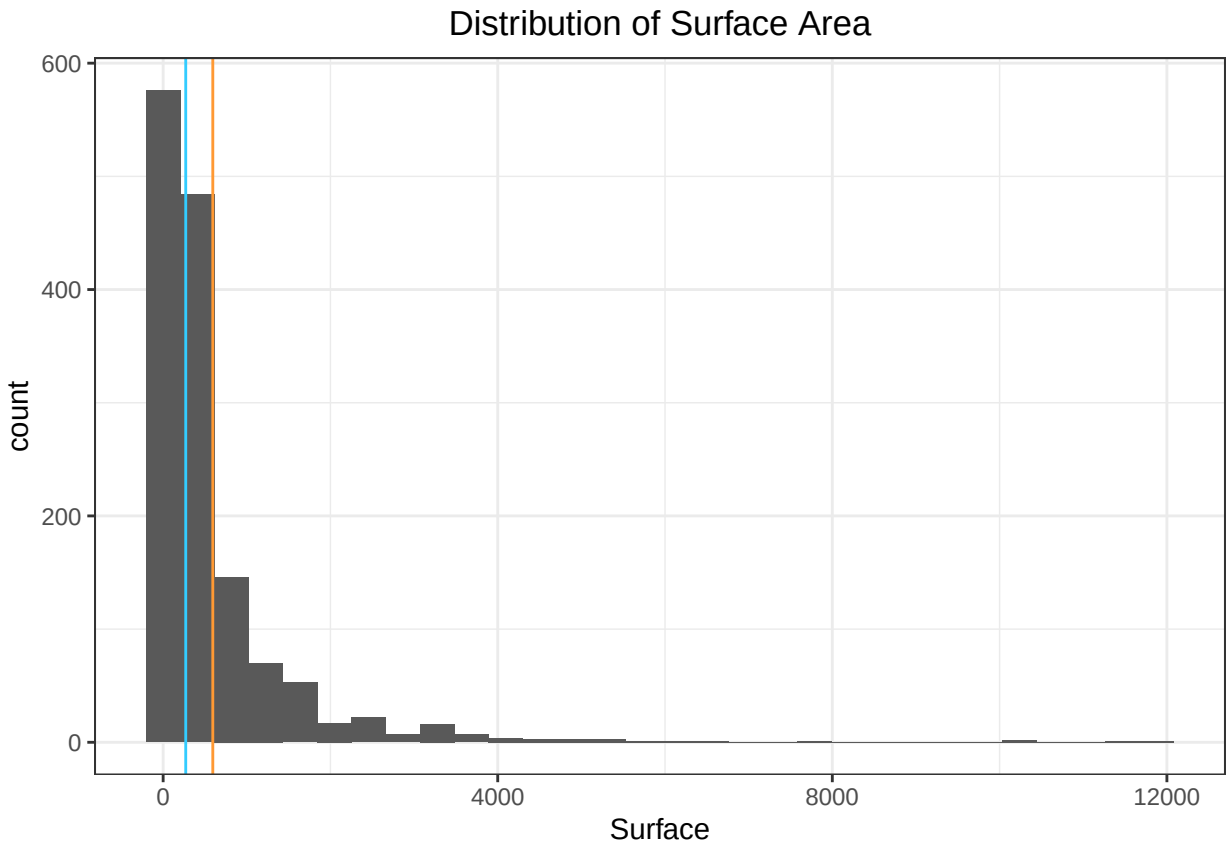
Table 5: Missing Variables

| Variables | Missing.Rate |
|---|---|
| Diam_in | 0.979 |
| authorstyle | 0.945 |
| Surface_Rnd | 0.916 |
| type_intermed | 0.903 |
| winningbidder | 0.263 |
| winningbiddertype | 0.263 |
| endbuyer | 0.263 |
| Interm | 0.263 |
| materialCat | 0.127 |
| mat | 0.095 |
| Width_in | 0.077 |
| Surface_Rect | 0.077 |
| Height_in | 0.075 |
| material | 0.073 |
| Surface | 0.054 |
| Shape | 0.013 |

Since we already removed many of these variables from consideriation in the data redundancy step, we only need to impute values of `Interm`, `endbuyer`, `Surface`.

Since `Interm` and `endbuyer` are categorical, we created a separate category for the missing data.

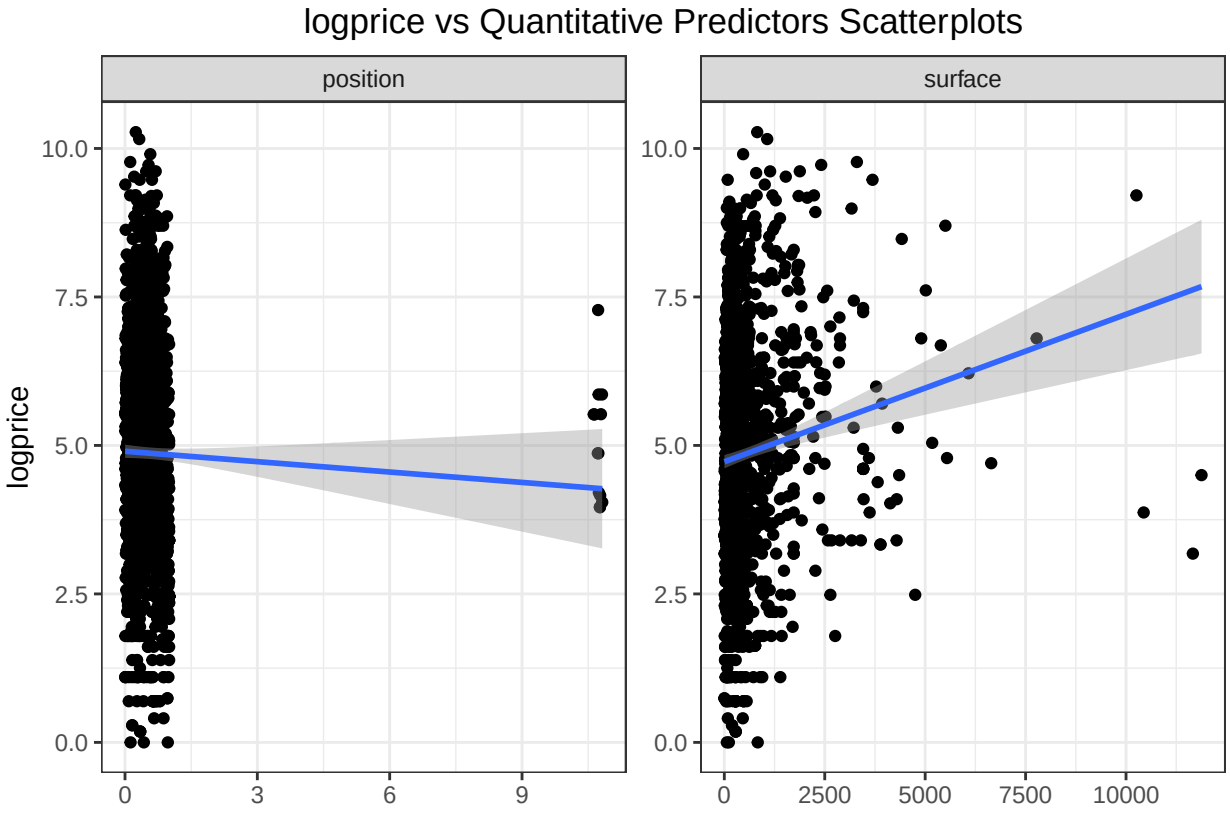
Since `Surface` is continuous, we used median imputation. We chose median imputation because the median is less sensitive to outliers than the mean, and we can see that the data is skewed (blue line represents the median, orange line represents the mean):




Additionally, we noticed that some covariates were not missing when `Surface` is missing, such as `Shape`, `Height_in`, `Width_in`, and `Diam_in`, so we conditioned our imputed values on the non-missing of these variables. For instance, observation 745 is a square rectangle painting with a height of 17 inches, so our imputed value for the surface area is 340 sq. inches, instead of the imputed 283.5 sq. inches for a square rectangle painting with no height or width filled in.

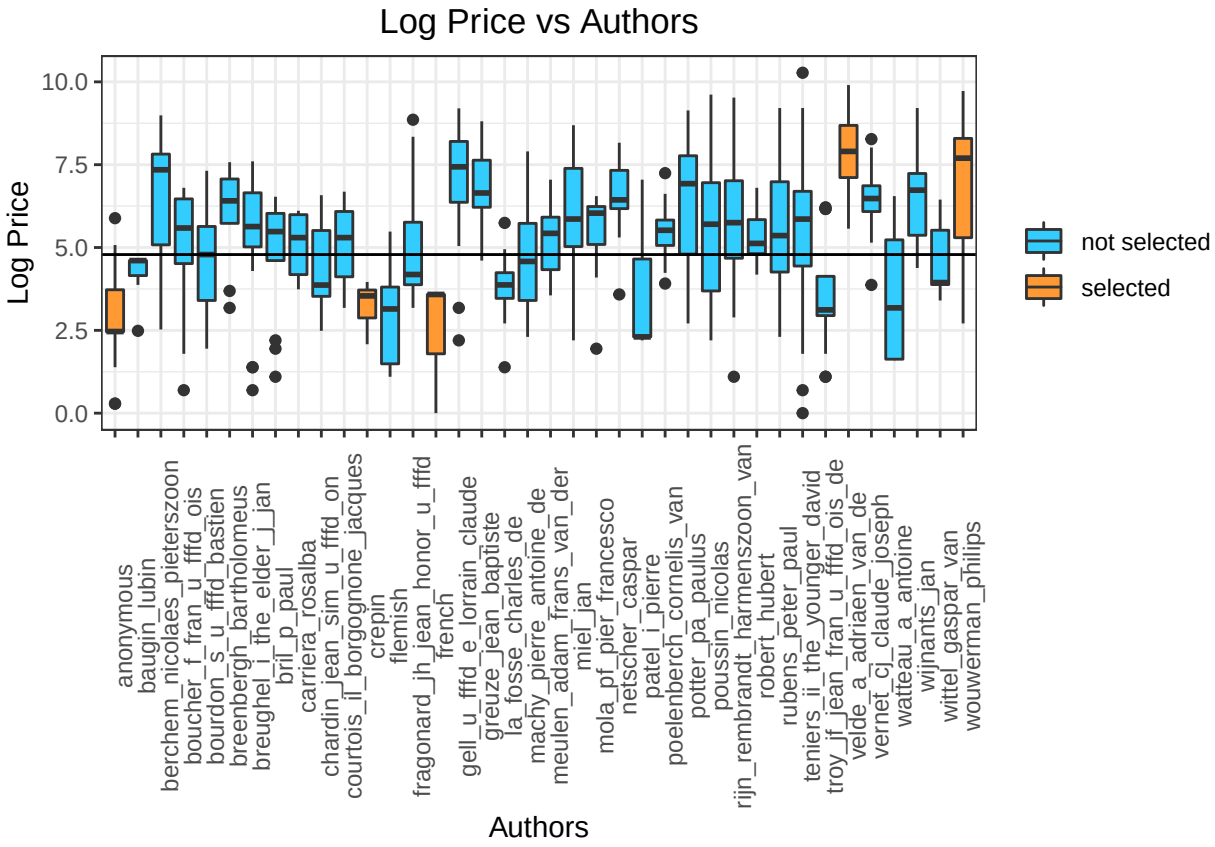
**Part (d) Variation of variables**

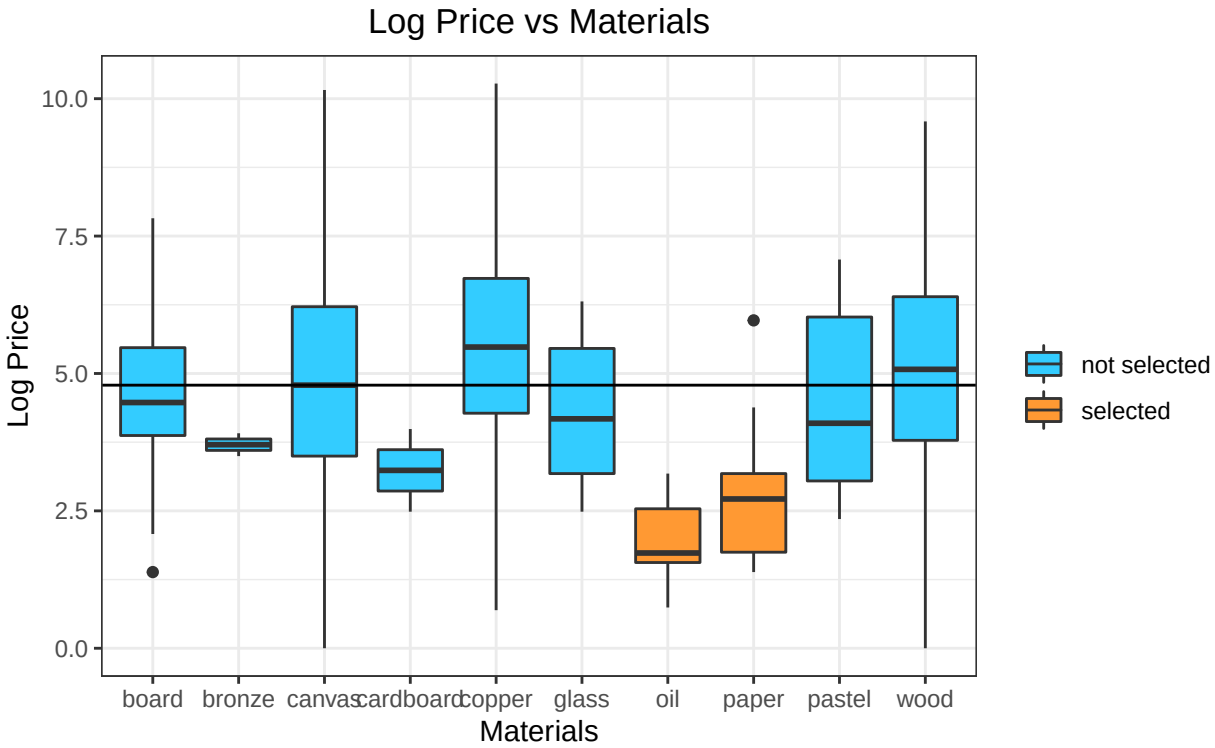
We start by examining the continuous variables:

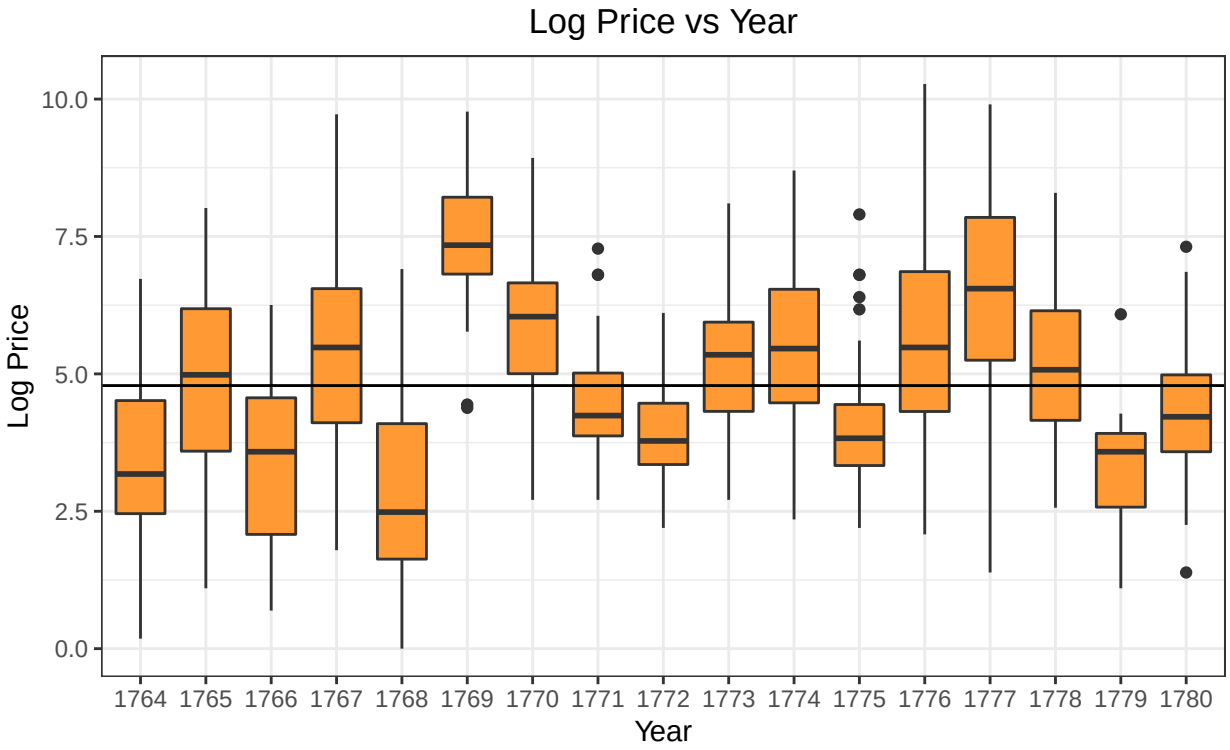



We see above that position doesn't seem to be strongly related to price, whereas there may be some relationship between surface and price.

Next, we consider categorical variables. Specifically, we want to examine the sources of variation of key discrete variables (within-variable variation or between-variable variation). We look at painters who appear at least 10 times in the dataset, the material of the painting, and the year the painting was sold:


Log Price vs Authors
Log Price
10.0
7.5
5.0
2.5
0.0
anonymous
baugin_lubin
berchem_nicolaes_pieterszoon
boucher_f_fran_u_fffd_ois
bourdon_s_u_fffd_bastien
breenbergh_bartholomeus
breughel_i_the_elder_j_jan
bril_p_paul
carriera_rosalba
chardin_jean_sim_u_fffd_on
courtois_il_borgognone_jacques
crepin
flemish
fragonard_jh_jean_honor_u_fffd
french
gell_u_fffd_e_lorrain_claude
greuze_jean_baptiste
la_fosse_charles_de
machy_pierre_antoine_de
meulen_adam_frans_van_der
miel_jan
mola_pf_pier_francesco
netscher_caspar
patel_i_pierre
poelenberch_cornelis_van
potter_pa_paulus
poussin_nicolas
rijn_rembrandt_harmenszoon_van
robert_hubert
rubens_peter_paul
teniers_ii_the_younger_david
troy_jf_jean_fran_u_fffd_ois_de
velde_a_adriaen_van_de
vernet_cj_claude_joseph
watteau_a_antoine
wijnants_jan
wittel_gaspar_van
wouwerman_philips
Authors
not selected
selected


Log Price vs Materials
Log Price
10.0
7.5
5.0
2.5
0.0
board
bronze
canvas
cardboard
copper
glass
oil
paper
pastel
wood
Materials
not selected
selected



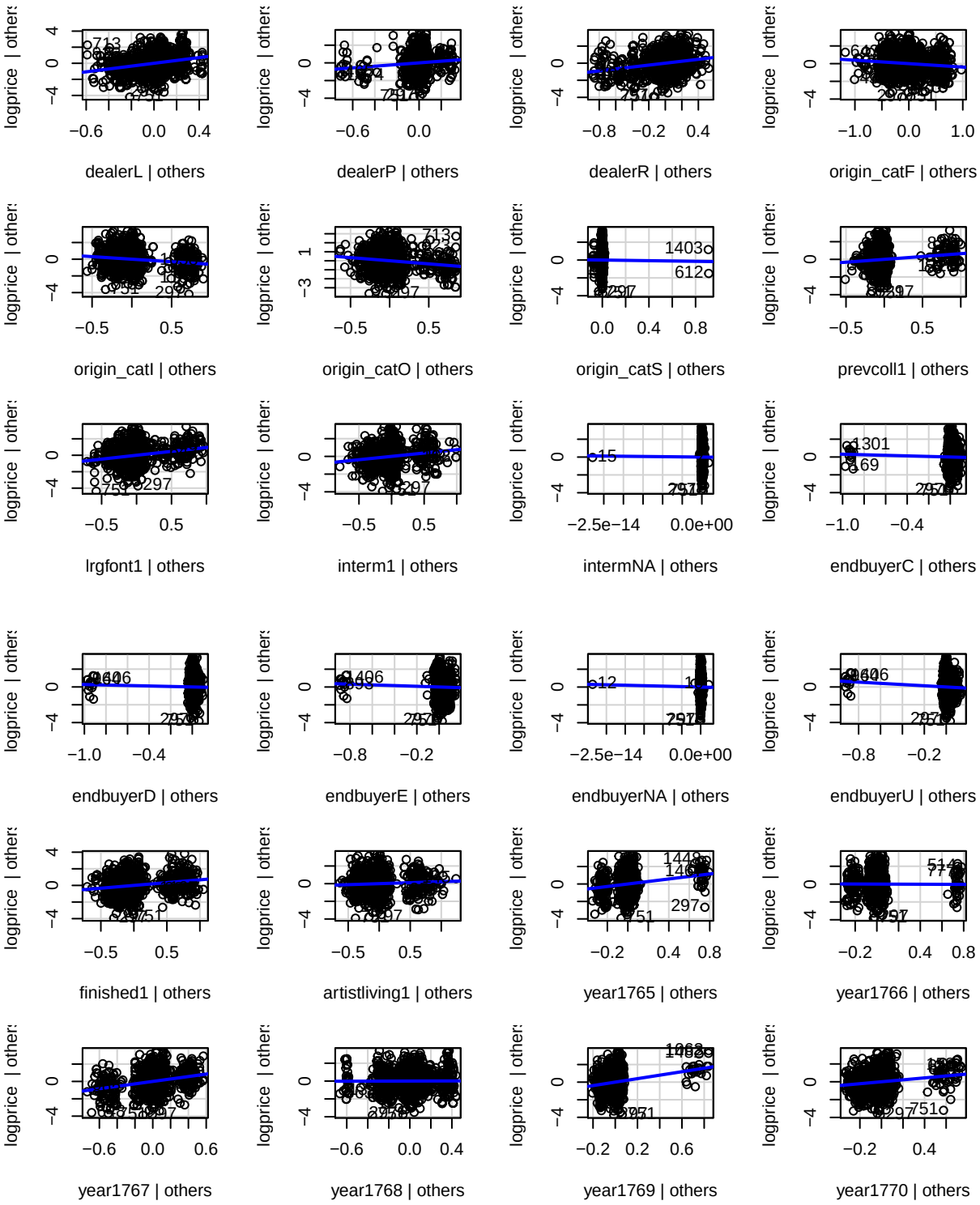

We can see that there exists both within-class and between-class variation for all three variables, indicating that we may want to include some of the individual levels of the variables (we highlighted the levels that we eventually use in the model in part (3) in orange), but will likely also want to control for other sources of variation.

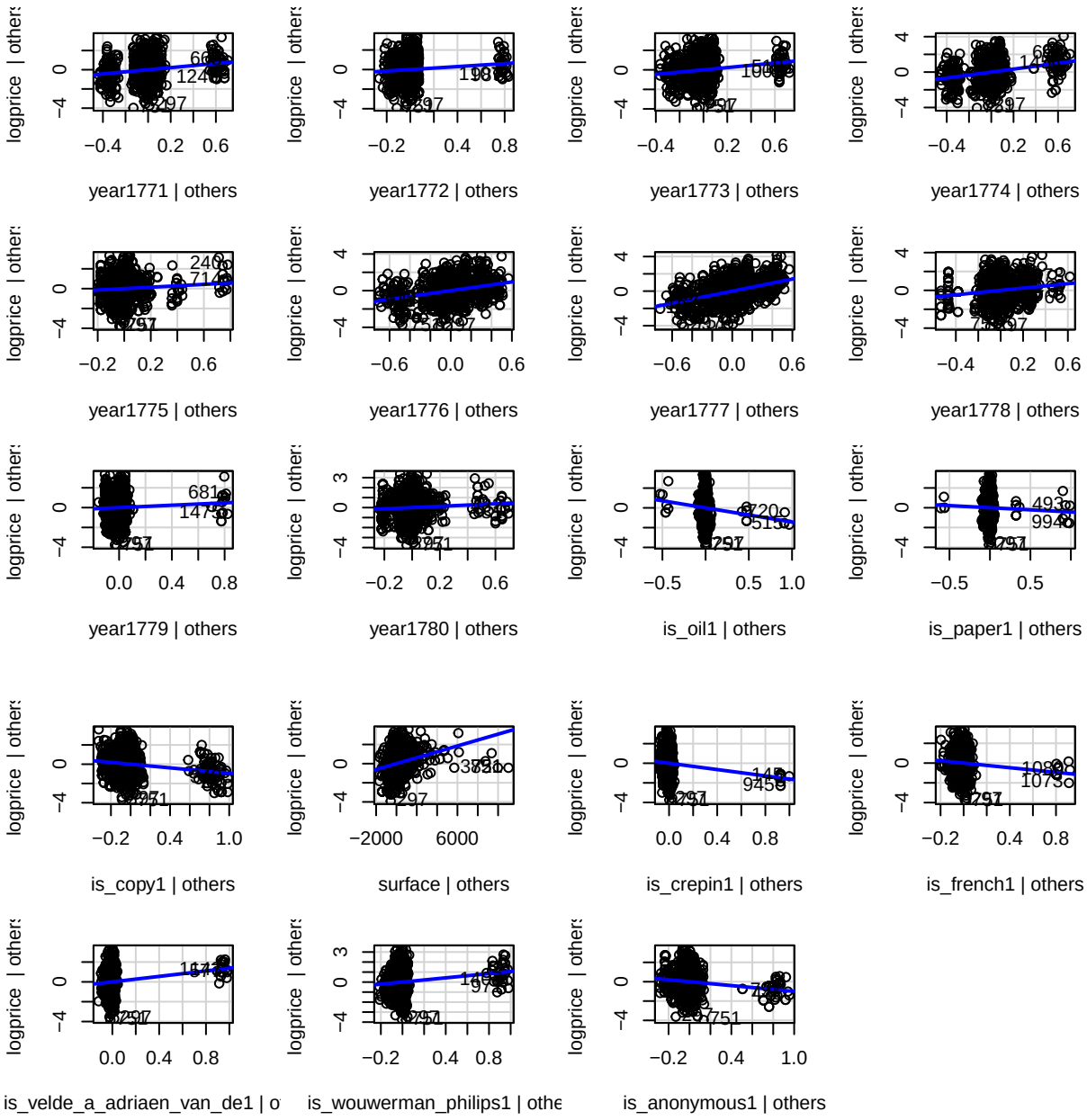
A final part we wanted to consider in sources of variation is in potential interactions. Specifically, we thought that dealers may specialize in different kinds of art, and therefore have different variation in prices for different kinds of paintings. Below, we consider the dealer's prices conditional on the origin of the work (`origin_cat`):


We can see that dealer "R" is the exclusive dealer in Spanish paintings in this dataset, and they sold for much more than the other pieces he sold. Additionally, it looks like dealer "J" sold a small amount of highly priced works of non-Spanish, non-Dutch, non-French, and non-Italian descent, while the other dealers did not, indicating that perhaps dealer "J" had exclusive access to a specific painter/set of painters outside of western Europe.

**Part (e) AV-plots**

As we saw in part (e), looking at individual boxplots shows some but not all of the variation of key categorical variables. We wanted to further this analysis by robustly considering the marginal impact a variable has holding all other variables constant. Below, we plot add-variable plots of all variables we consider for our model:


logprice | others
dealerL | others
dealerP | others
dealerR | others
origin_catF | others
origin_catI | others
origin_catO | others
origin_catS | others
prevcoll1 | others
lrgfont1 | others
interm1 | others
intermNA | others
endbuyerC | others
endbuyerD | others
endbuyerE | others
endbuyerNA | others
endbuyerU | others
finished1 | others
artistliving1 | others
year1765 | others
year1766 | others
year1767 | others
year1768 | others
year1769 | others
year1770 | others



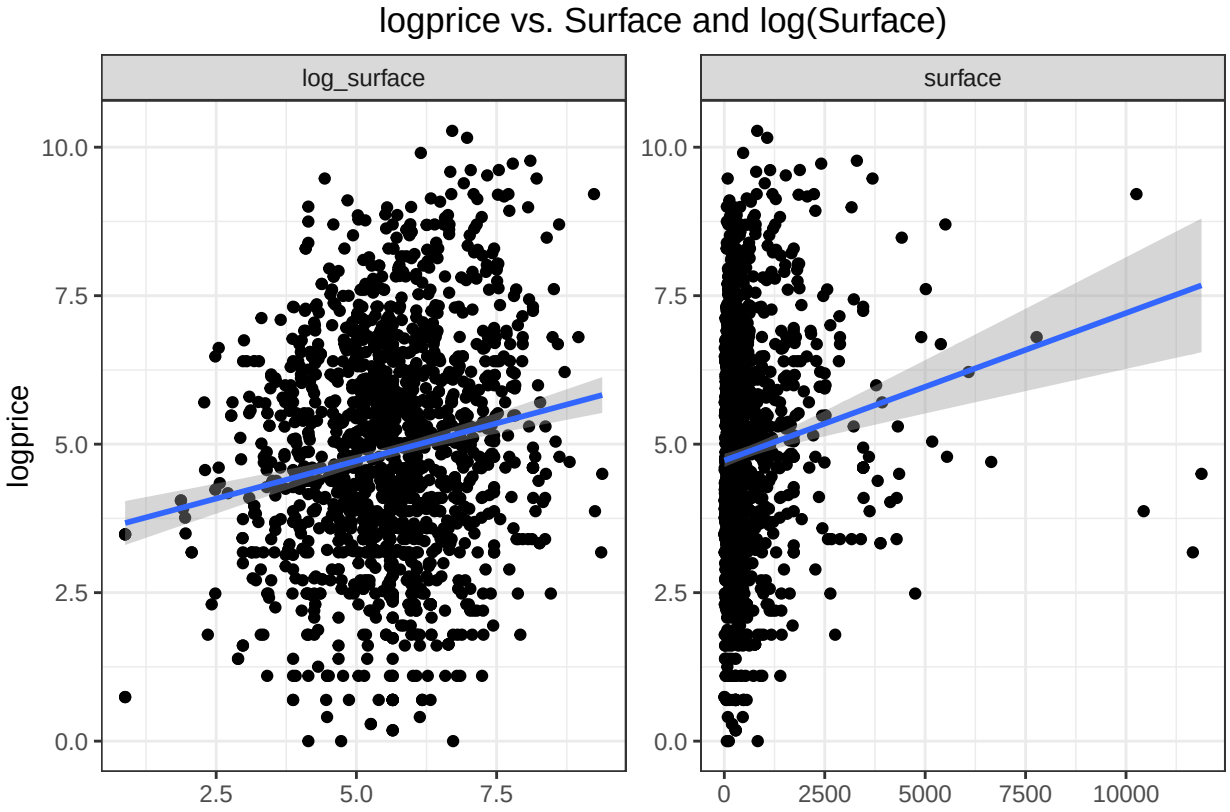

Now, we describe the additional EDA in part 2 of the project:

**Part (f) Variable transformation examination**

Now that we are moving into more complex models, we wanted to look at potential variable transformations

or other variables to add to our model.

We noticed specifically that for our model in part 1, paintings with a large surface area also had very high predicted prices. We may have estimated their prices too optimistically if the the trend is in actuality non-linear but we incorrectly used a linear trend. As a result, we wanted to see if we changed the variable scaling if there is any difference. Below is a plot of a log version of `surface` against log price, compared to the original version we used in part 1.




We see that this transformation is likely a more appropriate way to use `surface`, as the error region (colored in grey) for the best fit line is more homoscedastic. However, we also see that the slope is smaller in magnitude, indicating that even though the logged version of `surface` may be more appropriate, it may not be important in the model.

**Part (g) Other considerations**

Since all of the other variables in our previous model were discrete variables, the other types of additions to consider in complex models are interactions between different variables. A naive way to go about this is to use a modeling method that can automatically explore interactions (such as trees and GAMs). We do indeed do that, but to gain an intuition why, we wanted to briefly examine a potential interaction.

3. Discussion of preliminary model Part I (5 points)

In part 1, we used a linear regression (chosen through a stepwise-selection process with AIC penalty) to arrive at a model. One thing we did to make our regression model potentially more flexible for out-of-sample data is that we created a lot of discrete variables (such as dummy variables for key authors) and fitted them into our preliminary linear model. The performance of this model had mixed results (note that at the time, the wercker was outputting the wrong result), with a coverage of 95.73% and rmse of 2593.86. The maximum deviation reaches 47622, which indicates that the difference between our predicted price and actual price is

very high for at least 1 observation.

The large maximum deviation was an opportunity for further EDA, as it is not good for our model to be off by as much as $47,622 from the perspective of the art historian. After investigating our initial model, we found that the variable `surface` was specifically the driver behind the model predicting extremely large values sometimes. As a result, we tried transforming it. However, as we have shown in our updated EDA, this variable does not appear to be very significant after transformation.

Another thing that we believe might have not helped is the new dummy variables we created for authors. Since each author is observed sparsely, and encoded in a way that can be messy (such as multiple authors working together), we found it hard to properly express all of the information contained in that variable while not introducing a multitude of sparsely populated binary variables. We tried briefly in this portion of the project to use a single `author` variable and express observations with k authors as k separate rows each weighted 1/k (where only the author differs between them), but found that this did not improve on the predictions. As a result, we simplified our model to not include specific information about the artist.

4. Development of the final model (20 points)

### MARS

We wanted to use a modeling technique that allowed for more flexibility of variables (such as interactions and non-linear relationships) than linear regression, as well as have a method that can still have interpretable results and is easy to compute prediction intervals. As a result, we wanted to work with GAMs, since they are natural extensions to linear regressions (by allowing for knots and polynomial transformations). Specifically, we found that multivariate adaptive regression splines (MARS) to work well for this problem.

MARS can be seen as a regression extension of CART. It considers all possible binary splits of the categories for a qualitative predictor (or the support for a continuous predictor) and then uses forward selection to chooses the best pairs of piecewise indicator functions until a stopping criterion is met. Leveraging MARS requires minimal specification from the user to execute feature engineering (e.g., feature scaling) and automatically does feature selection, since non-informative features will not be chosen. Furthermore, highly correlated predictors do not impede predictive accuracy as much as they do with OLS models, and the algorithm of MARS allows for easier interpretation and more tractable prediction intervals than tree methods.

### Final model:

Below is the final model we use:

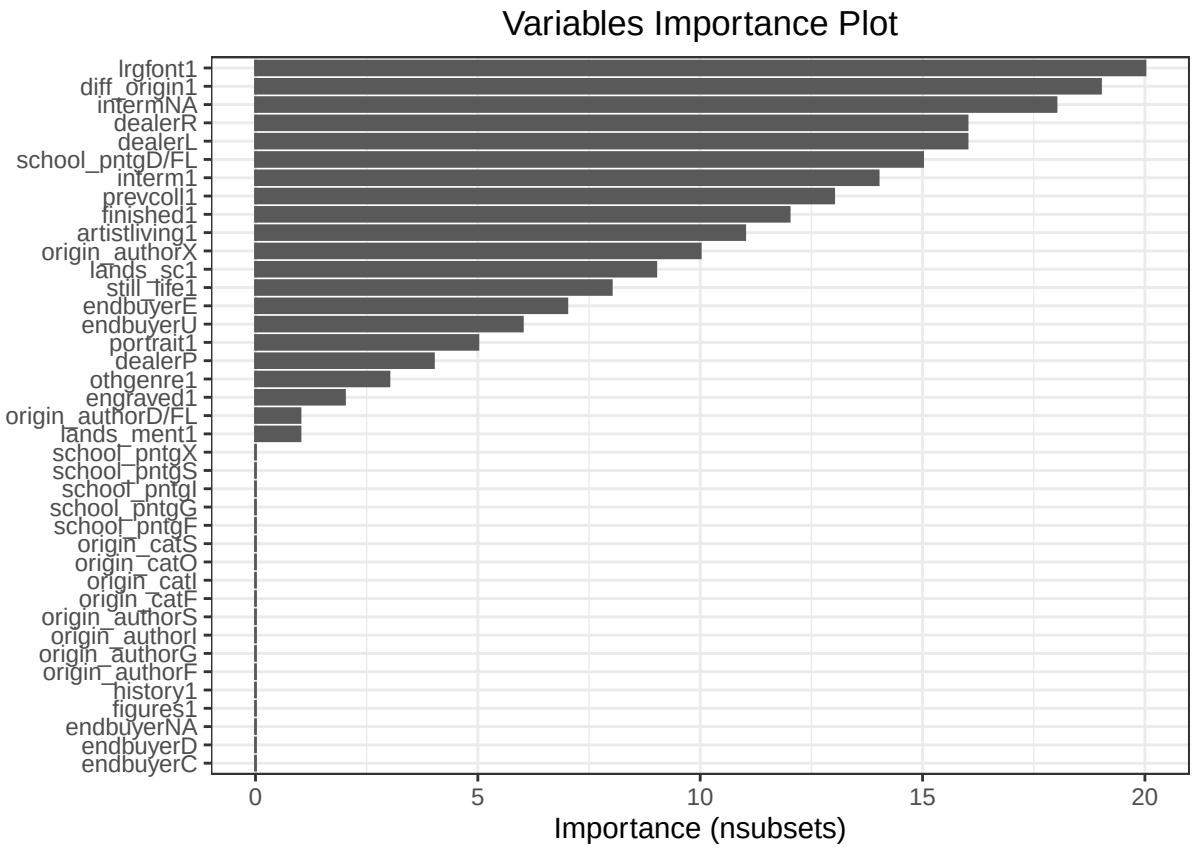
```
# mars model
mars <- earth(
  logprice ~ dealer+ diff_origin+ interm+ lrgfont+ origin_author+
    origin_cat+ prevcoll+ school_pntg+ endbuyer+ engraved+ figures+
    finished+ history+ lands_ment+ lands_sc+ othgenre+ portrait+
    still_life+artistliving,
  data=paintings_train2,
  nfold=10, ncross=30, varmod.method="lm"
)
```

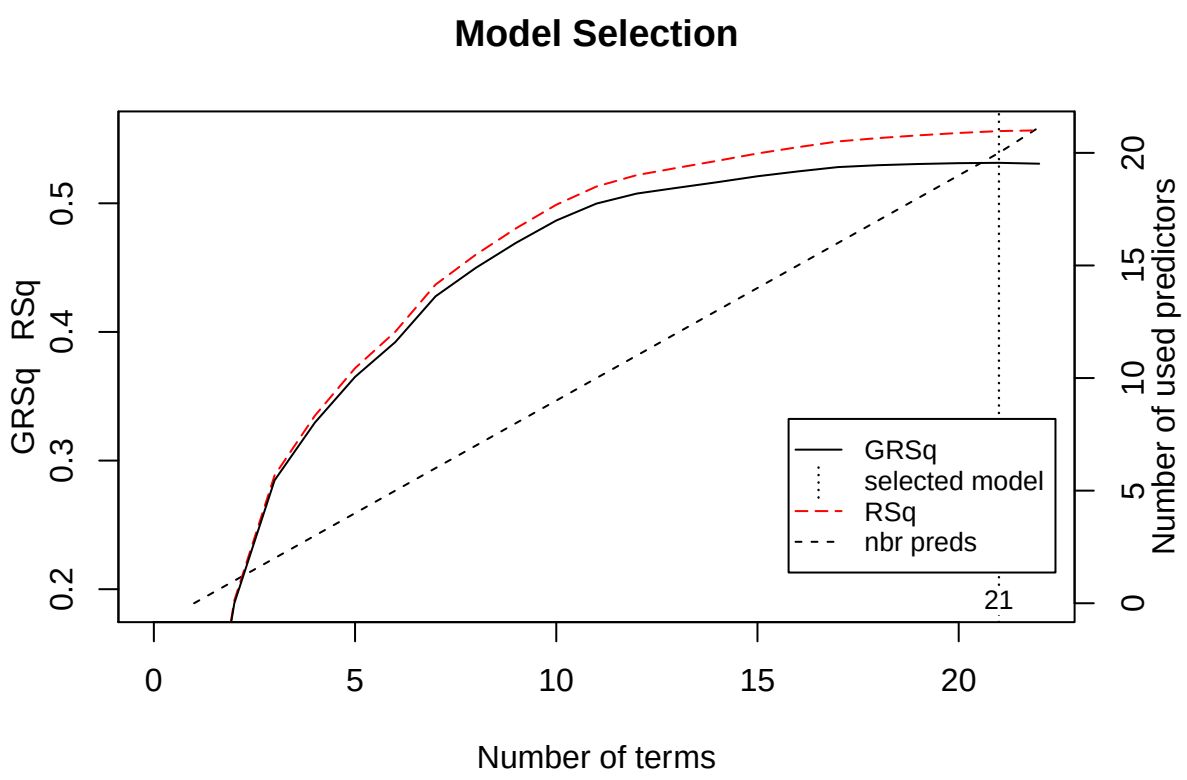
### Variables:

We used a similar set of variables in the model as we considered in part I, with the obvious exceptions of no longer including `surface`, and any `is_*` variable we created. Additionally, we added `diff_origin` and `origin_author`, which we previously excluded from linear regression due to being redundant with `origin_cat` (despite looking significant in the EDA), but can include here since MARS performs selection. `Dealer`, `diff_origin`, `interm`, `lrgfont`, `origin_author`, `origin_cat`, `prevcoll`, `school_pntg`, `endbuyer`,

`engraved`, `figures`, `finished`, `history`, `lands_ment`, `lands_sc`, `othgenre`, `portrait`, `still_life` are the 18 variables we considered to fit the MARS model. These variables are the ones seem significant in the EDA part and are chosen from testing.

We show the vaiable selection results Variables Importance Plot below. MARS's variable selection procedure is built upon the GCV (Generalized Cross Validation) $R^2$ value.






We can see specifically that `lrgfont`, `diff_origin`, and when `interm` is NA have a large influence in this model. Additionally, we can observe that MARS did not select some variables entirely, such as `origin_cat`, likely as a result of introducing the redundant `origin_author`.

**Summary:**

Below, we provide a coefficients summary table.

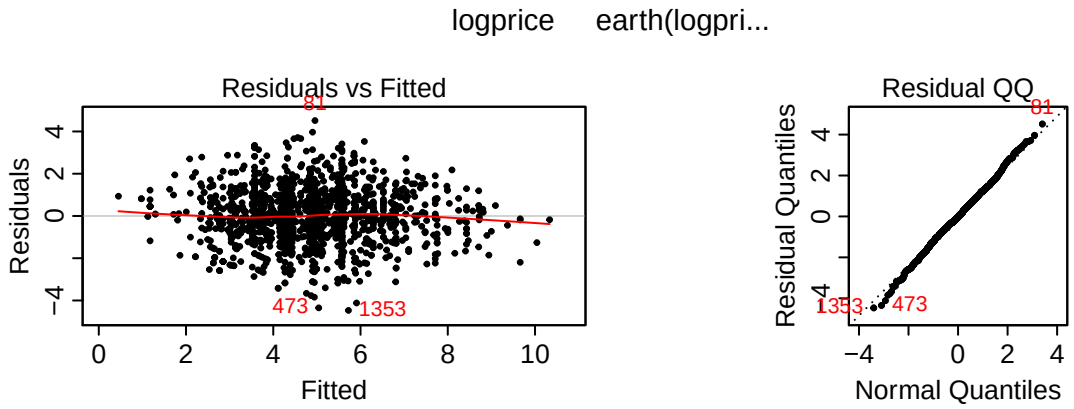
Table 6: MARS Coefficients Summary

| | logprice |
|---|---|
| (Intercept) | 3.7159024 |
| lrgfont1 | 1.1880634 |
| diff_origin1 | -0.5794961 |
| intermNA | -1.1564742 |
| interm1 | 0.9504021 |
| prevcoll1 | 1.2297329 |
| dealerL | 1.5927794 |
| dealerR | 1.1853237 |
| finished1 | 0.7173269 |
| artistliving1 | 0.5801272 |
| origin_authorX | -0.8034653 |
| lands_sc1 | -0.5530999 |
| still_life1 | -0.7246167 |
| endbuyerE | -0.6320670 |
| portrait1 | -0.6181558 |

| | logprice |
|---|---|
| endbuyerU | -0.4710025 |
| school_pntgD/FL | 0.6734971 |
| dealerP | 0.5176411 |
| othgenre1 | 0.2910657 |
| engraved1 | 0.3833605 |
| lands_ment1 | 0.4787292 |

We see that many of the variables that reduced the RSS the most (shown in the variable importance plots) also have larger coefficients here. Additionally, seeing some of these coefficients gives us insight into some levels of variables that were not selected. For instance, the model did not select `endbuyerC` or `endbuyerD`, but did select `endbuyerE` and `endbuyerU`. These two variables have negative coefficients, meaning that it it could be the case that `endbuyers` who are not in group `E` (expert organizing the sale) or `U` (unknown) may be associated with higher prices of paintings.

**Residuals:**




As shown in the residual plots, the residuals are generally centered around zero, are normally distributed, and are homoscedastic, indicating that our model is not obviously biased towards an identifiable subset of the population.

**Prediction Intervals:**

As for the predictions intervals, we used the interval provided by the earth package when fitting mars model. Specifically, in our initial function call, we used the argument `varmod.method="lm"`, indicating that once

`earth()` builds the model, it will then internally apply an `"lm"` (linear regression) to its absolute residuals to estimate the variance conditional on each x. To read more about it in depth, there is a great article here describing the methodology: http://www.milbo.org/doc/earth-varmod.pdf. In the next section, we show the implications of using a linear model for mapping a variance function versus another MARS model for mapping a variance function.

5. Assessment of the final model (25 points)

**Model evaluation**

In order to fit a MARS model, generalized cross-validation is used within the `earth` package, meaning that it is likelier the model will not overfit the training data than a linear regression may. However, it is still a good idea to investigate the in-sample coverage and RMSE to make sure we are within a reasonable range for our model in-sample.

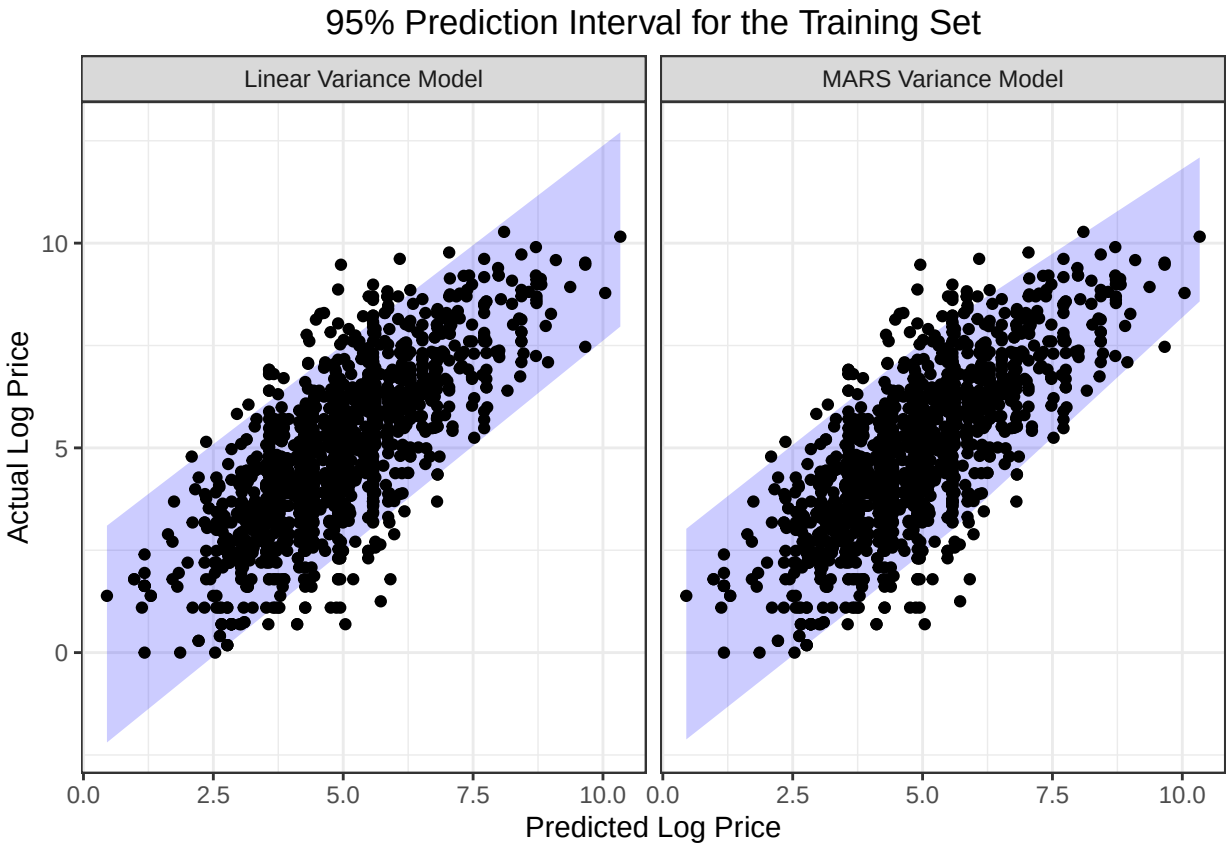



Table 7: RMSE of the Training Set

| RMSE |
|---|
| 1690.57 livres |

From the coverage plots of the training set, we can see that the majority of the observations falls between our prediction interval under either set up, which indicates that our method of computing coverage seems appropriate and not too stringent on the functional form. Because we saw the errors of the MARS model are generally homoscedastic, we use the linear model here.

The RMSE of the training set is 1690.57 livres, which was lower than the linear regressions we ran in part I.

**Model testing**

Table 8: Model Performance for test data

| Bias | Coverage | Maximum Deviation | Mean Absolute Deviation | RMSE |
|---|---|---|---|---|
| 254.32 | 0.948 | 14558.89 | 489.12 | 1305.28 |

We tested our model using the test data in the wercker. The resulting coverage is about 94.8%. This value is pretty close to 95%, which further indicates that our model is doing good in appropriately accounting for the variance of the data. The test RMSE is about 1305, which is even smaller than the training RMSE, indicating that we are not overfitting our training data. The maximum deviation and mean absolute deviation are 14558.89 and 489.12, respectively. These two values indicate that our estimation is slightly off, and for some observations the difference between the actual price and predicted price is very large. This could happen for the paintings that are significantly overvalued or undervalued. However, it is important to remember that both of these values are smaller in the test set than the training set, so it may be part of variation in the response that we cannot explain with the predictor variables we have right now.

**Model Result**

Table 9: Top 10 Valued Paintings in the Validation Set

| fit | dealer | year | authorstandard | lrgfont |
|---|---|---|---|---|
| 20979.776 | R | 1769 | Rijn, Rembrandt Harmenszoon van | 1 |
| 15681.681 | R | 1769 | Breenbergh, Bartholomeus | 1 |
| 9900.203 | R | 1769 | Dou, Gerrit | 1 |
| 8999.657 | R | 1776 | Dujardin, Karel | 1 |
| 8101.764 | R | 1767 | Vernet (CJ), Claude-Joseph | 1 |
| 7653.525 | R | 1776 | Breenbergh, Bartholomeus | 1 |
| 7653.525 | R | 1769 | Dijk (P), Philip van | 1 |
| 7320.537 | R | 1770 | Cantarini (il Pesarese), Simone | 1 |
| 6726.943 | R | 1767 | Dou, Gerrit | 1 |
| 6133.871 | R | 1769 | Metsu, Gabriel | 1 |

From the table above, we can see that the top 10 valued paintings are all sold by dealer R and the majority of them are sold in 1769. As we may have expected based on the variable importance plots, having the dealer devote an extra paragraph to a painting is associated with higher value paintings.

It also seems like the paintings painted by Gerrit Dou and Bartholomeus Breenbergh have high values since both of them have two paintings that are valued top 10. Going back to our choice to exclude individual painters, we still are okay with this choice given the limited number of observations of each painter in the data. However, it may be worth in the future trying to set up a model with a Bayesian hierarchical random effects framework if we really wanted to leverage author names, even if they are sparsely populated.

6. Conclusion (10 points): must include a summary of results and a discussion of things learned. Optional what would you do if you had more time.

During the model building procedure, we found that many models (tree models, Ridge and Lasso) don't have tidy built-in methods to calculate prediction intervals, and when using bootstrap to calculate it by ourselves, the results often lead to bad coverage since the bootstrap method can underestimate uncertainty (note that we did not have time to re-submit all of these methods to github after the wercker test data was updated on Thursday night, so the intervals may be better with the updates).

Additionally, when using Bayesian Model Averaging, we found that while coverage was good, the root mean squared error was fairly high, indicating to us that it was necessary to move beyond linearity.

For us, using MARS, a GAM method that mimicks trees, was a nice compromise between uncovering nonlinearities and kinks in the data with appropriate prediction intervals and digestible interpration. This allowed for good coverage, and great RMSE.

Another key learning moment is that despite our intuition of the author mattering, we could not find a clever way to have it be valuable in our models. Intuitively, famous painters' work is more desired, regardless of the content. However, with little data, it may be hard to assess the level of respect different painters command, especially considering that some may have commanded more respect as they became more known. In a time series dataset like we have here, this is likely too difficult to do.

Looking more directly at the model itself, it seems like the best indicators of a painting's value are the dealer's marketing of it, and the level of respect that that dealer commands. Seeing that so many of the paintings that are predicted to be most valuable in our model come from one dealer when he specifically uses an extra paragraph to advertise it may indicate that people in Paris heavily relied on the word of the dealer to assess the worth of the painting. It would be interesting to see if adjusted for inflation, people today in the art buying business would be as reliant on the dealer's information, or if they could evaluate art more independently. Perhaps it would be worth it to ask the art historian questions like: to the best of your knowledge, did the interaction between buyers, dealers, and painters change from when this data was observed to now? If so, how? Being able to distinguish between paintings that were valuable due to market dynamics at the time and paintings that were valuable because they were actually that good is an important distinction to make, and it is important for us to not assume correlation is causation here.

If we had more time, there are a few things we could consider further:

1. Try doing a bagged MARS model, using the `bagEarth()` function in the `caret` library. This could potentially lead to lower variance predictions, since we can fit multiple MARS models.
2. Change the framework completely to do a Bayesian hierarchical random effects model, incorporating the author, dealer, and potentially even the bidder as random effects. This may stabilize estimates for painters who are sparsely observed in the dataset.

## Reference

Mars: http://uc-r.github.io/mars