Final Data Analysis Project

See Parts for Write-Up due Dates

For this project you will take the role of a consultant hired by an Art historian to explore what drove prices of paintings in 18th century Paris. They have provided you with auction price data from 1764-1780 on the sales (seller/buyer), painter, and other characteristics of paintings.

About the Data Analysis Project

The art historian would like to know what factors drove prices of painting, which paintings might be overvalued and which are undervalued. It is up to you to decide what methods you want to use (frequentist or Bayesian or a combination) to answer these questions, and implement them to help to identify undervalued and overvalued paintings, as well as which features and possible interactions are at play.

You will have three data sets: a subset for training, a subset for testing, and a third subset for validation. You will be asked to do data exploration and build your model (or models) initially using only the training data. Then, you will test your model on the testing data, and finally validate using the validation data. We are challenging you to keep your analysis experience realistic, and in a realistic scenario you would not have access to all three of these data sets at once. You will be able to see on our scoreboard how well your team is doing based on its predictive performance on the testing data. After your project is turned in you will see the final score on the validation set.

All members of the team should contribute equally and may be asked to answer any questions about the analysis at the final presentation.

For your analysis create a new Rmd named "project-I.Rmd" for part I and update accordingly rather than editing this. Your write up should not have any of the instructions, for example. Figures should be labeled appropriately and report numbers using significant digits. This file may be updated so do not edit this document.

Code:

In your write up your code should be hidden (`echo = FALSE`) so that your document is neat and easy to read. However your document should include all your code such that if I re-knit your Rmd file I should be able to obtain the results you presented. If there is any code that you wish to highlight you may included it, but it should contribute significantly to your write up that should be directed to the art historian.

see Due dates in Sakai/Calendar for submissions

Read in Training Data

To get started read in the training data:

```
load("paintings_train.Rdata")
load("paintings_test.Rdata")
```

The Code Book is in the file `paris_paintings.md` provides more information about the data.

Part I: Simple Model

EDA

Using EDA and any numerical summaries get to know the data - identify what you might consider the 10 best variables for predicting `logprice` using scatterplots with other variables represented using colors or symbols, scatterplot matrices or conditioning plots.

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

```
##
##           canvas copper other wood
##        74      0      0     0    0
##   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      0    131     0    0
##   ca    0      0      0     2    0
##   co    0      0      0     5    0
##   g     0      0      0     1    0
##   h     9      0      0     0    0
##   m     0      0      0     1    0
##   mi    0      0      0     4    0
##   n/a  69      0      0     0    0
##   o     0      0      0     1    0
##   p     0      0      0    10    0
##   pa    0      0      0     4    0
##   t     0    731      0     0    0
##   ta   39      0      0     0    0
##   v     0      0      0     6    0
```

Therefore, we can drop `Interm` and `materialCat` due to perfect collinearity.

```
# missing randomly or not
missing <- lm(paintings_train$logprice ~ paintings_train$winningbiddertype=="")
summary(missing)
```

```
##
## Call:
## lm(formula = paintings_train$logprice ~ paintings_train$winningbiddertype ==
##     "")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2954 -1.2167  0.0029  1.1832  4.9796
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                     5.29543    0.05357   98.85
## paintings_train$winningbiddertype == ""TRUE -1.62203    0.10439  -15.54
##                                                Pr(>|t|)
## (Intercept)                                      <2e-16 ***
```

```
## paintings_train$winningbiddertype == ""TRUE   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.781 on 1498 degrees of freedom
## Multiple R-squared:  0.1388, Adjusted R-squared:  0.1382
## F-statistic: 241.4 on 1 and 1498 DF,  p-value: < 2.2e-16
```

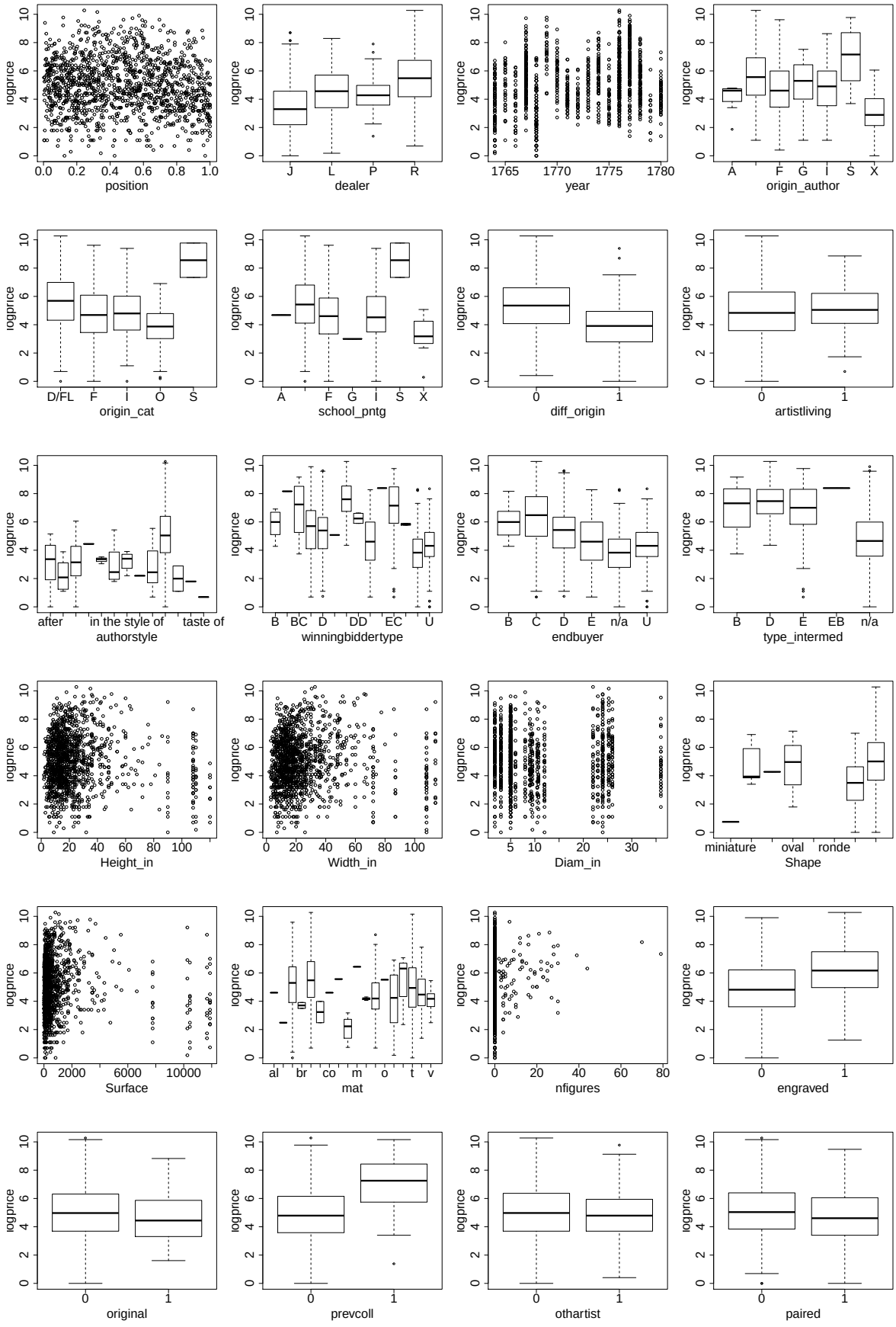
This indicates that `winningbiddertype` is not missing purely randomly and that there is a strong case for selective non-response. Similarly, `endbuyer`, `type_intermed`, `material` and `mat` are not missing purely randomly either. So we impute "n/a" into `winningbiddertype`, `endbuyer`, `type_intermed`, `Shape`, `material` and `mat` to replace "" and "-" as a new category.

We can drop `price` and `count` based on their meanings, with the former being another type of our response and the latter being completely useless since it is the same for all observations.

Besides, `sale`, `lot`, `subject`, `author` and `material` have too many levels, and thus they are not selected.

In addition, we have alrealy built the variables that contains the useful information of `authorstandard` and `winningbidder`, so we can drop them as well. Similarly, we can drop `Height_in`, `Width_in`, `Surface_Rect`, `Diam_in` and `Surface_Rnd`.

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


logprice
position
dealer
year
origin_author
origin_cat
school_pntg
diff_origin
artistliving
authorstyle
winningbiddertype
endbuyer
type_intermed
Height_in
Width_in
Diam_in
Shape
Surface
mat
nfigures
engraved
original
prevcoll
othartist
paired

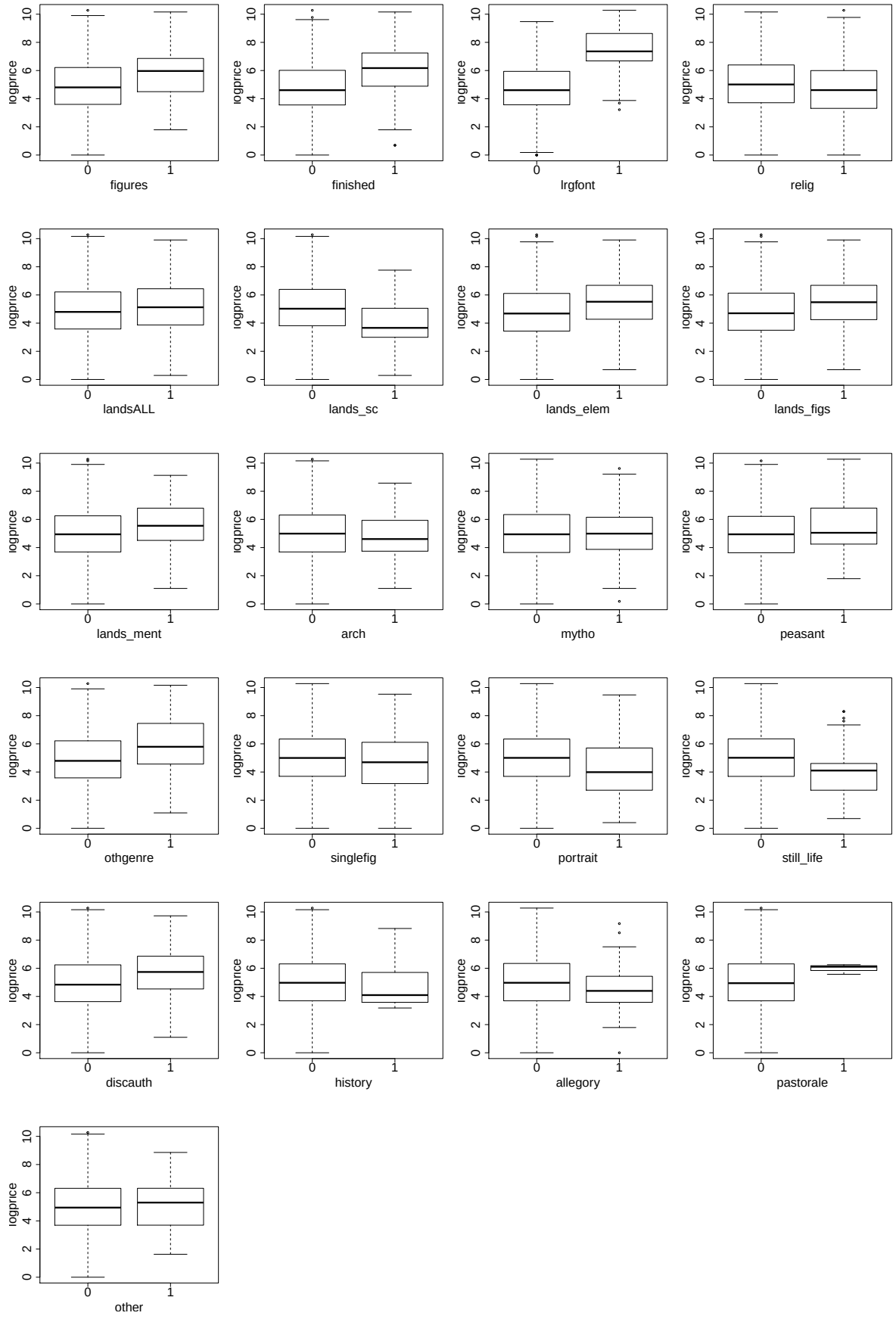


Figure 1: Plots of predictors versus logprice

Since the points of `Surface` are clustered at the beginning of x axis, we can do log transformation to it to see its relationship with `logprice`.

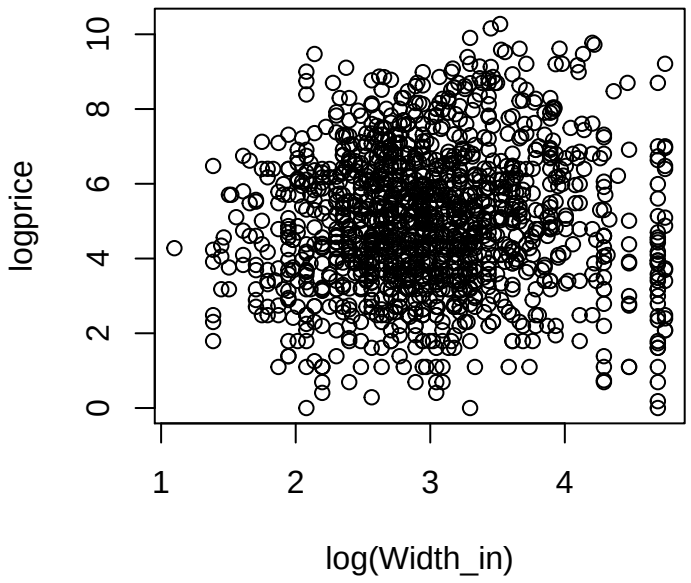


Figure 2: Plots of log Surface versus logprice

Build your first model

In the first model predict the auction price `price` using the transformation `logprice` using at least 10 and up to 20 predictors and any interactions to build a model using linear regression. You may use stepwise model selection to simplify the model using AIC and/or BIC. For reference, we will fit the null model to initialize the leaderboard, but replace model1 with your recommended model.

version 1 (using random forest)

From the result of random forest, we can choose 12 variables. (drop `diff_origin` since it's perfectly collinear with `origin_author` and `origin_cat`; drop `winningbiddertype` and `mat` since there are too many categories in them)

```
# full model
model_full <- lm(logprice ~ (position + dealer + year + origin_author + origin_cat + authorstyle + endbu

# BIC
model_bic <- step(model_full, k = log(nobs(model_full)), direction = "both", trace = F)
summary(model_bic)
```

```
##
## Call:
## lm(formula = logprice ~ dealer + year + origin_author + endbuyer +
##     type_intermed + log_Surface + finished + lrgfont + year:endbuyer,
##     data = paintings_train_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6906 -0.7931 -0.0313  0.8204  3.7759
```

```
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.655e+02  1.457e+02  -1.136   0.2563
## dealerL            1.480e+00  1.384e-01  10.689   <2e-16 ***
## dealerP            6.542e-02  1.716e-01   0.381   0.7030
## dealerR            1.718e+00  1.157e-01  14.849   <2e-16 ***
## year               9.511e-02  8.235e-02   1.155   0.2483
## origin_authorD/FL  4.046e-01  4.722e-01   0.857   0.3916
## origin_authorF    -6.740e-02  4.721e-01  -0.143   0.8865
## origin_authorG    -1.040e-01  5.294e-01  -0.197   0.8442
## origin_authorI    -2.806e-01  4.811e-01  -0.583   0.5599
## origin_authorS    -1.666e-02  6.153e-01  -0.027   0.9784
## origin_authorX    -1.049e+00  4.849e-01  -2.164   0.0306 *
## endbuyerC         -7.361e+01  1.476e+02  -0.499   0.6180
## endbuyerD          3.639e+01  1.477e+02   0.246   0.8054
## endbuyerE         -2.275e+02  1.504e+02  -1.512   0.1307
## endbuyern/a       -1.342e+01  1.472e+02  -0.091   0.9274
## endbuyerU         -1.675e+01  1.500e+02  -0.112   0.9111
## type_intermedD     1.059e-01  4.081e-01   0.259   0.7953
## type_intermedE    -1.917e-01  4.295e-01  -0.446   0.6554
## type_intermedEB    2.650e+00  1.287e+00   2.058   0.0398 *
## type_intermedn/a  -7.684e-01  3.885e-01  -1.978   0.0482 *
## log_Surface        2.420e-01  2.533e-02   9.552   <2e-16 ***
## finished1          9.515e-01  9.634e-02   9.876   <2e-16 ***
## lrgfont1           1.188e+00  1.265e-01   9.391   <2e-16 ***
## year:endbuyerC     4.155e-02  8.337e-02   0.498   0.6183
## year:endbuyerD    -2.060e-02  8.344e-02  -0.247   0.8050
## year:endbuyerE     1.282e-01  8.498e-02   1.509   0.1316
## year:endbuyern/a   6.949e-03  8.318e-02   0.084   0.9334
## year:endbuyerU     9.147e-03  8.474e-02   0.108   0.9141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.229 on 1302 degrees of freedom
## Multiple R-squared:  0.5881, Adjusted R-squared:  0.5795
## F-statistic: 68.84 on 27 and 1302 DF,  p-value: < 2.2e-16
```

```
## Linear Regression
##
## 1330 samples
##    8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1195, 1198, 1196, 1196, 1197, 1198, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   1.238802  0.5760035  0.9914152
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

10-fold cross validation RMSE is 1.243051.

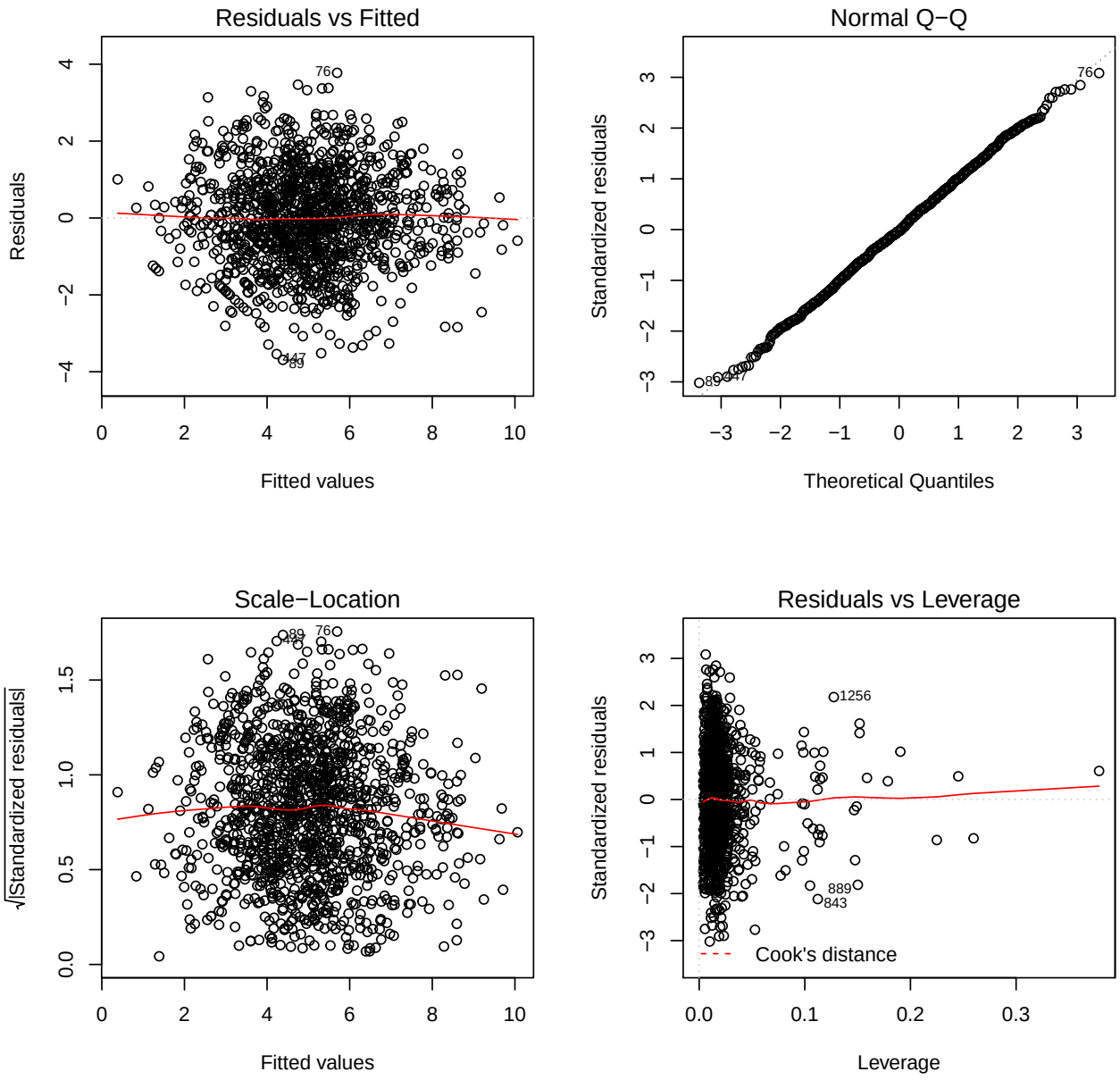


Figure 3: Diagnostic Plots

Save predictions and intervals.

```
##
##  iter imp variable
##    1   1  Height_in*  Width_in*  Diam_in*  Surface*
##    1   2  Height_in*  Width_in*  Diam_in*  Surface*
##    1   3  Height_in*  Width_in*  Diam_in*  Surface*
##    1   4  Height_in*  Width_in*  Diam_in*  Surface*
##    1   5  Height_in*  Width_in*  Diam_in*  Surface*
##    2   1  Height_in*  Width_in*  Diam_in*  Surface*
##    2   2  Height_in*  Width_in*  Diam_in*  Surface*
##    2   3  Height_in*  Width_in*  Diam_in*  Surface*
##    2   4  Height_in*  Width_in*  Diam_in*  Surface*
```

```
##   2   5  Height_in*  Width_in*  Diam_in*  Surface*
##   3   1  Height_in*  Width_in*  Diam_in*  Surface*
##   3   2  Height_in*  Width_in*  Diam_in*  Surface*
##   3   3  Height_in*  Width_in*  Diam_in*  Surface*
##   3   4  Height_in*  Width_in*  Diam_in*  Surface*
##   3   5  Height_in*  Width_in*  Diam_in*  Surface*
##   4   1  Height_in*  Width_in*  Diam_in*  Surface*
##   4   2  Height_in*  Width_in*  Diam_in*  Surface*
##   4   3  Height_in*  Width_in*  Diam_in*  Surface*
##   4   4  Height_in*  Width_in*  Diam_in*  Surface*
##   4   5  Height_in*  Width_in*  Diam_in*  Surface*
##   5   1  Height_in*  Width_in*  Diam_in*  Surface*
##   5   2  Height_in*  Width_in*  Diam_in*  Surface*
##   5   3  Height_in*  Width_in*  Diam_in*  Surface*
##   5   4  Height_in*  Width_in*  Diam_in*  Surface*
##   5   5  Height_in*  Width_in*  Diam_in*  Surface*
##  * Please inspect the loggedEvents
```

Part I Write up ***Last day to submit is Dec 7 by 5; accepted until Dec 6 (5 points off if late)***

Once you are satisfied with your model, provide a write up of your data analysis project in a new Rmd file/pdf file: `Part-I-Writeup.Rmd` by copying over salient parts of your R notebook. The written assignment consists of five parts:

1. Introduction: Summary of problem and objectives (5 points)
2. Exploratory data analysis (10 points): must include three correctly labeled graphs and an explanation that highlight the most important features that went into your model building.
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.
- Model selection: must include a discussion
- Residual: must include residual plot(s) and a discussion.
- Variables: must include table of coefficients and CI

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.

Points will be deducted for code chunks that should not be included, etc.

Upload write up to Sakai any time before Dec 7th

Evaluation on test data for Part I

Once your write up is submitted, your models will be evaluated on the following criteria based on predictions on the test data (20 points):

- Bias: Average (Yhat-Y) positive values indicate the model tends to overestimate price (on average) while negative values indicate the model tends to underestimate price.
- Maximum Deviation: Max |Y-Yhat| - identifies the worst prediction made in the validation data set.
- Mean Absolute Deviation: Average |Y-Yhat| - the average error (regardless of sign).
- Root Mean Square Error: Sqrt Average (Y-Yhat)^2
- Coverage: Average(lwr < Y < upr)

In order to have a passing wercker badge, your file for predictions needs to be the same length as the test data, with three columns: fitted values, lower CI and upper CI values in that order with names, *fit*, *lwr*, and *upr* respectively such as in the code chunk below.

Save predictions and intervals.

You will be able to see your scores on the score board. They will be initialized by a prediction based on the mean in the training data.