Part1 Analysis

Team: SparkR

2017/12/7

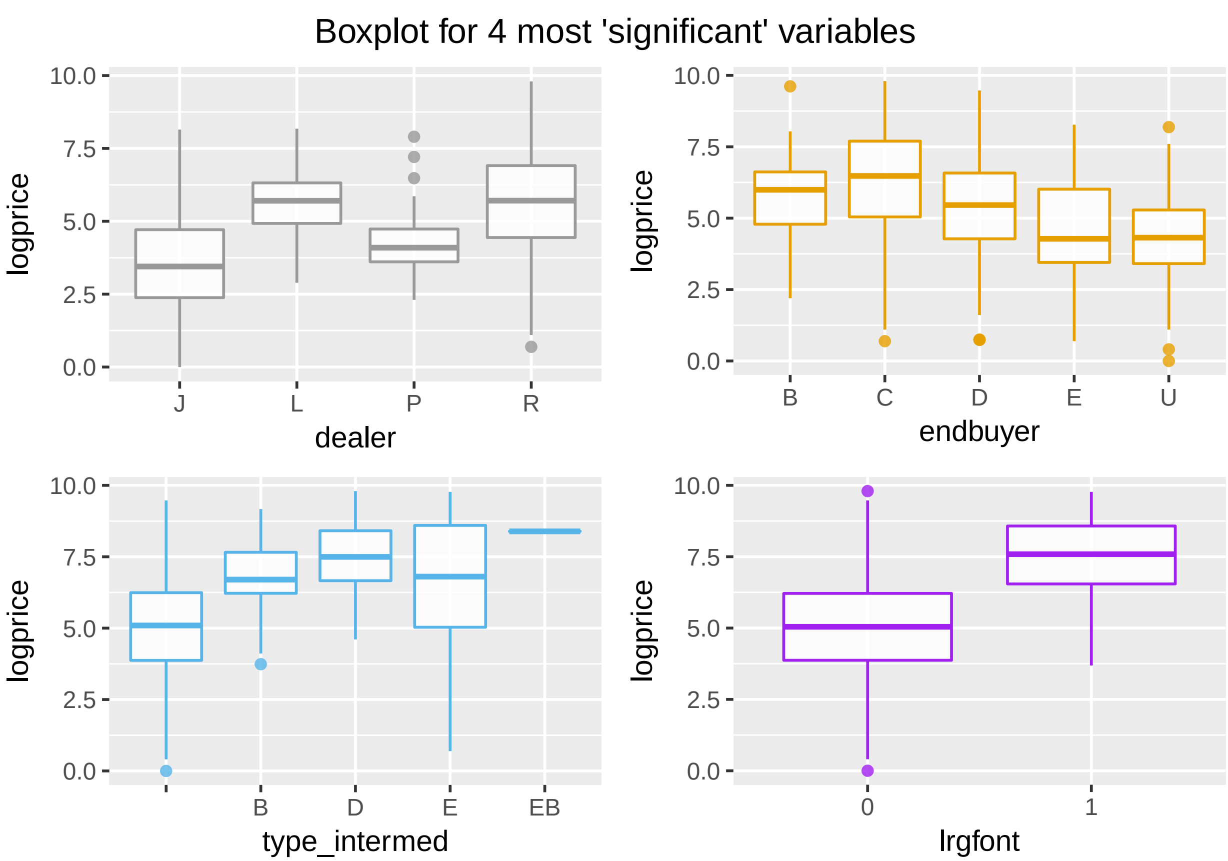
1. Introduction: Summary of problem and objectives (5 points)

Late 18th century Paris has witnessed countless number of auctions on fine arts and paintings from countries across Europe. Our project utilizes this dataset from auction in 18th century Paris with details of paintings, authors and auctions. We hope to build a OLS model to predict the log of price fetched at auction. Since our dataset contains 59 variables, we would want to perform variable selection to build a parsimonious model with good performance in critiria such as RMSE, etc. In addition to that, some variables have a lot of missing values and typos. We would need to perform some sort of missing imputation, e.g. assume MCAR and use MICE package. For typos, we would recode them to appropriate values. From the correlation martices, we find some highly-correlated variables, so we dicide to not include some of the variables in our model to aviod multicollinearity.

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.

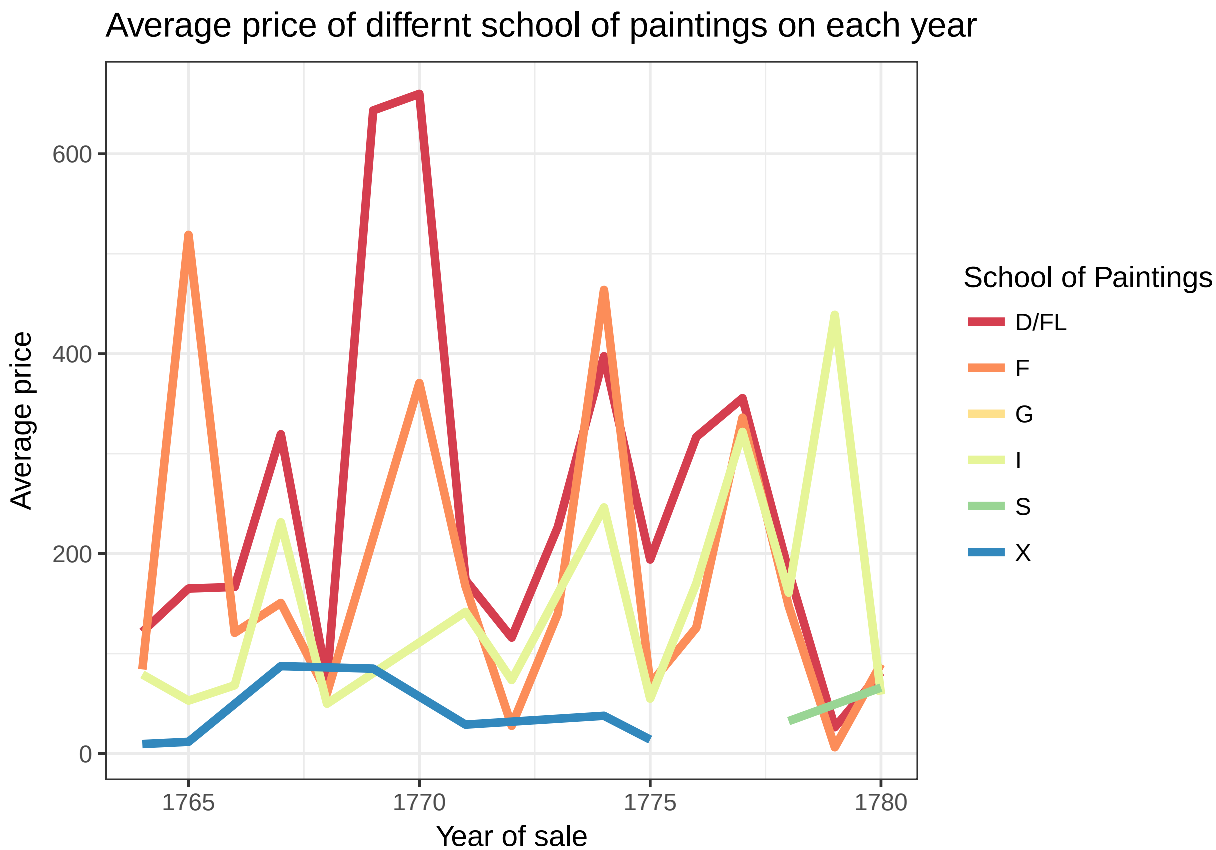
For this part, we intially have found some interaction between predictors from graph perspective; however, in our further model selection process and trials and error.

First, we choose to plot boxplots for the variables that are highly correlated with response `logprice`



As we can see, for these four factor variables, each level has differernt means and distributions across, for examle, doing business with dealer R may not be a good deal, since they tend to have higher price, but it could also be the case that they tend to sell higher value painting; also, collectors tend bid higher price compared with other end buyers; from type of inttype of intermediary perspective, experts tend to have higher price in mind,(becasue they know the true intrinsic value?), lastly, if dealer have more information about the paintings, the final price will be higher as well. Therefore, it is likely human factors(especially type of dealer) play an important role on deciding the final bidding price, and we will expect they will "survive" after model selection process.

Second, we try to find some interaction effect between predictors using tree models.



Since plotting average logprice, all the line lie between each other, we decide to plot price in original scale instead. As you can see from the plot, Dutch/Flemish and French school of paintings tend to have a higher average price in the year between 1765 and 1778, with the maximum across the group, Dutch/Flemish has the average price over 600 dollars for the paintings in the year around 1770(possibly becasue of the decease of some well-known artists);Even though Italian has lowever in the earlier year, they do have higher price in post-1777 era. With all these observations, it suggests there will be "some" interaction effect between `school_pntg` and `year` that affect the overall price of paintings.

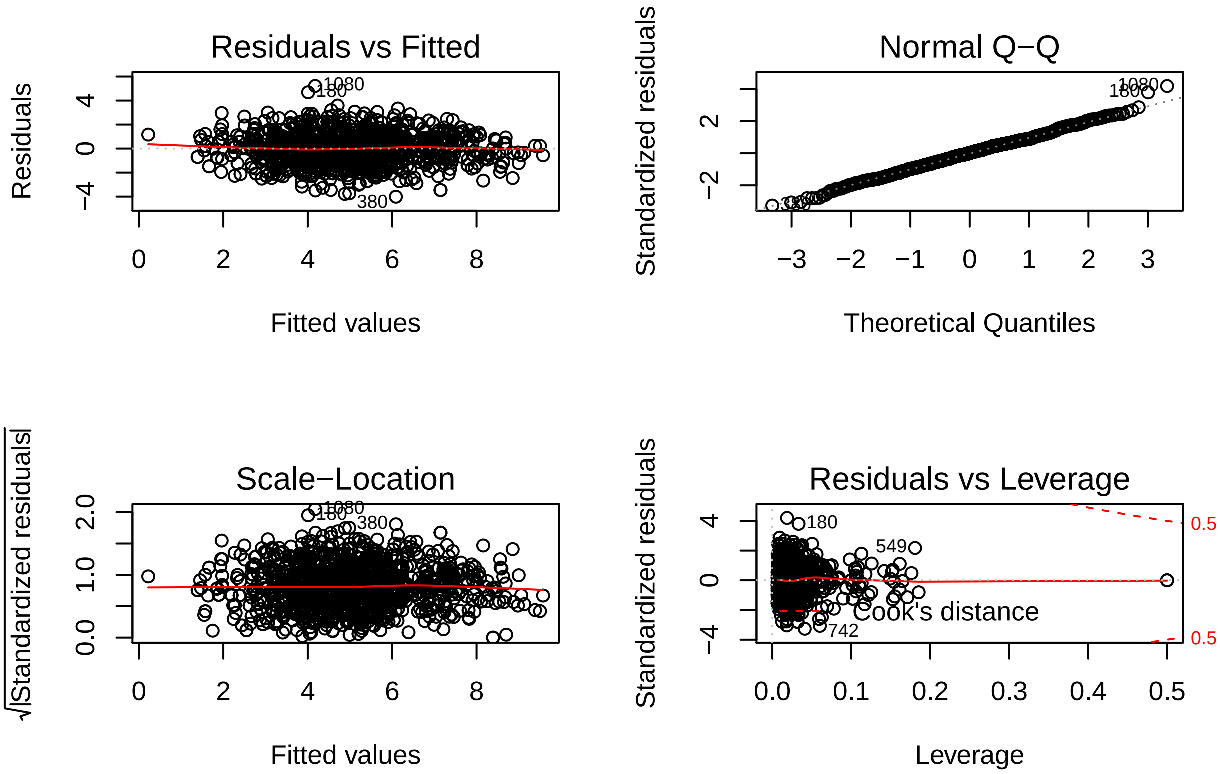
3. Development and assessment of an initial model (10 points)

- 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

```
##
## Call:
## lm(formula = logprice ~ Surface + dealer + endbuyer + diff_origin +
##     type_intermed + engraved + mat_recode + school_pntg_recode +
```

```
##     paired + lrgfont + lands_sc + othgenre + discauth + othartist +
##     still_life, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0094 -0.8462  0.0153  0.7973  5.2178
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.27680    0.21222  10.728  < 2e-16 ***
## Surface                  0.21815    0.02466   8.847  < 2e-16 ***
## dealerL                  1.44048    0.15694   9.178  < 2e-16 ***
## dealerP                  0.48516    0.21439   2.263 0.023833 *
## dealerR                  1.36311    0.12916  10.554  < 2e-16 ***
## endbuyerB                0.95852    0.33223   2.885 0.003989 **
## endbuyerC                0.93528    0.15769   5.931 4.03e-09 ***
## endbuyerD                1.06328    0.13044   8.152 9.70e-16 ***
## endbuyerE                0.34638    0.17272   2.005 0.045158 *
## endbuyerU                0.49085    0.15535   3.160 0.001623 **
## diff_origin1            -0.89560    0.09728  -9.206  < 2e-16 ***
## type_intermedB           0.48464    0.41456   1.169 0.242645
## type_intermedD           1.27266    0.17161   7.416 2.41e-13 ***
## type_intermedE           0.53758    0.24307   2.212 0.027197 *
## type_intermedEB          0.92571    0.90905   1.018 0.308746
## engraved1                0.45289    0.17848   2.537 0.011303 *
## mat_recodemetal          0.42415    0.15193   2.792 0.005334 **
## mat_recodena             0.31973    0.13537   2.362 0.018353 *
## mat_recodeother          0.98612    0.48591   2.029 0.042656 *
## mat_recodepaper         -0.31038    0.39225  -0.791 0.428945
## mat_recodewood           0.04942    0.10534   0.469 0.639059
## school_pntg_recodeF     -0.63543    0.09337  -6.806 1.65e-11 ***
## school_pntg_recodeI     -0.76431    0.12312  -6.208 7.59e-10 ***
## school_pntg_recodeother -0.53954    0.48638  -1.109 0.267540
## school_pntg_recodeX     -1.03756    0.30138  -3.443 0.000598 ***
## paired1                 -0.21257    0.08320  -2.555 0.010759 *
## lrgfont1                 1.30613    0.13283   9.833  < 2e-16 ***
## lands_sc1               -0.31239    0.14530  -2.150 0.031775 *
## othgenre1                0.62601    0.13191   4.746 2.35e-06 ***
## discauth1                0.64973    0.16042   4.050 5.48e-05 ***
## othartist1              -0.34668    0.14976  -2.315 0.020802 *
## still_life1             -0.63116    0.20661  -3.055 0.002306 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.255 on 1099 degrees of freedom
## Multiple R-squared:  0.5899, Adjusted R-squared:  0.5783
## F-statistic: 50.99 on 31 and 1099 DF,  p-value: < 2.2e-16
```

From the summary of our model, we can see the adjusted R-squared is 0.5783, which suggests 57.83% of variability has been explained by our dependent variables when accounted for number of variables used. From our EDA, we can see dealer, type of end buyer(endbuyer), type of intermediary(type_intermed) and whether the dealer devotes an additional paragraph(lrgfont). Those variables appear to be important in the boxplots so we would include them in our model. Starting with the full model, we perform stepwise selection based on BIC.



The Residual vs Fitted plot shows consistent residuals across different fitted values, which means constant variance assumption is satisfied. In addition to that, the Normal Q-Q plot displays essentially normal distributed residuals. And from Cook's distance plot, we can see all observations have small Cook's distance and there is no sign of influential points. There are is a high leverage point but since the standardized residual is very small, it's not an influential point. We suspect observation #180, #1080, #380 are potential outliers, but since their leverage are very low, it would not affect the model much. Hence, we decide that all assumptions for this model have been met and we will use this model for prediction.

The coefficients and confidence interval of variables used are as follows:

	coefficient	2.5%	97.5%
(Intercept)	2.2768035	1.8604006	2.6932064
Surface	0.2181537	0.1697687	0.2665388
dealerL	1.4404780	1.1325325	1.7484236
dealerP	0.4851640	0.0644968	0.9058312
dealerR	1.3631051	1.1096816	1.6165286
endbuyerB	0.9585241	0.3066527	1.6103954
endbuyerC	0.9352802	0.6258633	1.2446971
endbuyerD	1.0632820	0.8073511	1.3192130
endbuyerE	0.3463829	0.0074856	0.6852803
endbuyerU	0.4908454	0.1860280	0.7956629
diff_origin1	-0.8955977	-1.0864791	-0.7047163
type_intermedB	0.4846368	-0.3287884	1.2980621
type_intermedD	1.2726610	0.9359438	1.6093783
type_intermedE	0.5375815	0.0606478	1.0145151
type_intermedEB	0.9257095	-0.8579552	2.7093742
engraved1	0.4528924	0.1026905	0.8030944
mat_recodemetal	0.4241521	0.1260403	0.7222640
mat_recodena	0.3197257	0.0541215	0.5853299
mat_recodeother	0.9861187	0.0326983	1.9395390

	coefficient	2.5%	97.5%
mat_recodepaper	-0.3103834	-1.0800258	0.4592590
mat_recodewood	0.0494201	-0.1572717	0.2561120
school_pntg_recodeF	-0.6354344	-0.8186382	-0.4522306
school_pntg_recodeI	-0.7643115	-1.0058836	-0.5227394
school_pntg_recodeother	-0.5395400	-1.4938701	0.4147900
school_pntg_recodeX	-1.0375641	-1.6289123	-0.4462159
paired1	-0.2125701	-0.3758270	-0.0493131
lrgfont1	1.3061340	1.0455137	1.5667543
lands_sc1	-0.3123892	-0.5974837	-0.0272947
othgenre1	0.6260104	0.3671855	0.8848352
discauth1	0.6497342	0.3349739	0.9644945
othartist1	-0.3466792	-0.6405260	-0.0528323
still_life1	-0.6311594	-1.0365483	-0.2257704

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.

The "baseline" category is auctioned by dealer J, with no information about end buyer nor the type of intermidary. The matrial of painting is canvas, school of painting is Dutch/Flemish, sold not in pairs, and dealer did not devote large paragraph nor engaged with the authenticity of the painting. The painting is not about land scape and did not contain still life elements. And the painting is not linked with the work of another artists or style and its description did not mention a genre scene.

The median price for the "baseline" category is:

```
## [1] 31.65096
```

	coefficient	2.5%	97.5%
Surface	0.2181537	0.1697687	0.2665388
dealerL	1.4404780	1.1325325	1.7484236
dealerP	0.4851640	0.0644968	0.9058312
dealerR	1.3631051	1.1096816	1.6165286
endbuyerB	0.9585241	0.3066527	1.6103954
endbuyerC	0.9352802	0.6258633	1.2446971
endbuyerD	1.0632820	0.8073511	1.3192130
endbuyerE	0.3463829	0.0074856	0.6852803
endbuyerU	0.4908454	0.1860280	0.7956629
type_intermedB	0.4846368	-0.3287884	1.2980621
type_intermedD	1.2726610	0.9359438	1.6093783
type_intermedE	0.5375815	0.0606478	1.0145151
type_intermedEB	0.9257095	-0.8579552	2.7093742
school_pntg_recodeF	-0.6354344	-0.8186382	-0.4522306
school_pntg_recodeI	-0.7643115	-1.0058836	-0.5227394
school_pntg_recodeother	-0.5395400	-1.4938701	0.4147900
school_pntg_recodeX	-1.0375641	-1.6289123	-0.4462159

It's interesting that we find some interaction effects during EDA when we fitted some tree models, however, after fitting linear model, we decide to drop those interaction effects to improve model performance. This result suggets although those interactions might exist, they are not good predictors for logprice. Then, we can take a look at some important variables and their confidence intervals. From the coefficients, we can see dealer is very important to the logprice. For example, if auctioned by dealer L, we would expect our price to increase by 1.44%, and we are 95% confident that the increase in price will be between 1.13% to 1.75%. All other categorical variables would have similar interpretation. For the only numerical variable surface, if we increase surface area by 1%, the price would increase 0.0022 livres.

My suggestion for art historian would be look for a painting with as large surface area as possible, and ideally auctioned by dealer L, bought by a dealer in the auction. And the origin of the painting should be correctly classified by the dealer. The painting should went through a dealer intermediary. In addition, if the dealer mentioned engraving done after the painting, it's likely to worth more. The material of the painting should be other to increase its value. Last, if a dealer devoted a large paragraph to the painting, it will worth more.