

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 include 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.

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 `modell` with your recommended model.

```
modell = lm(logprice ~ 1, data=paintings_train)
```

Save predictions and intervals.

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 CIs)? (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 interpretations 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 $(\hat{Y} - Y)$ positive values indicate the model tends to overestimate price (on average) while negative values indicate the model tends to underestimate price.
- Maximum Deviation: $\max |Y - \hat{Y}|$ - identifies the worst prediction made in the validation data set.
- Mean Absolute Deviation: Average $|Y - \hat{Y}|$ - the average error (regardless of sign).
- Root Mean Square Error: $\sqrt{\text{Average } (Y - \hat{Y})^2}$
- Coverage: Average $(\text{lwr} < Y < \text{upr})$

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.

Part II: Complex Model (start Dec 4th ideally!)

In this part you may go all out for constructing a best fitting model for predicting housing prices using methods that we have covered this semester. You should feel free to create any new variables (such as quadratic, interaction, or indicator variables, splines, etc) and try different methods, keeping in mind you should be able to explain your methods and results.

Update your predictions using your complex model to provide point estimates and CI.

You may iterate here as much as you like exploring different models until you are satisfied with your results, however keep in mind you must be able to explain your results to the art historian.

Part II: Write Up

Once you are satisfied with your model, provide a write up of your data analysis project in a new Rmd file/pdf file: **Part-II-Writeup.Rmd** by copying over salient parts of your R notebook and the previous writeup (you should also save the pdf version) The written assignment consists of five parts:

1. Introduction (1 point if improved from before) add previous intro with any edits
2. Exploratory data analysis (1 point if improved from before): add previous EDA
3. Discussion of preliminary model Part I (5 points) Discuss performance based on leader board results and suggested refinements.
4. Development of the final model (20 points)
 - Final model: must include a summary table
 - Variables: must include an explanation
 - Variable selection/shrinkage: must use appropriate method and include an explanation
 - Residual: must include a residual plot and a discussion
 - discussion of how prediction intervals obtained
5. Assessment of the final model (25 points)
 - Model evaluation: must include an evaluation discussion
 - Model testing : must include a discussion
 - Model result: must include a selection and discussion of the top 10 valued paintings in the validation data.
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.

Final Predictions Validation (20 points)

Create predictions for the validation data from your final model using the dataframe `paintings_validation.Rdata` in your repo. You may refit your final model to the combined training and test data. Write predictions out to a file `prediction-validation.Rdata` *This should have the same format as the model output in Part I and II!*

Final: Class Presentations and Peer Evaluation

Each Group should prepare 5 slides in their Github repo: (save as slides.pdf)

- Most interesting graphic *a picture (painting) is worth a thousand words prize!*
- Best Model (motivation, how you found it, why you think it is best)
- Best Insights into predicting Price.
- 3 Best Paintings to purchase (and why) (images are a bonus!)
- Best Team Name/Graphic

We will select winners based on the above criteria and overall performance.

Finally your repo should have: `Part-I-Writeup.Rmd`, `Part-I-Writeup.pdf`, `Part-II-Writeup.Rmd`, `Part-II-Writeup.pdf`, `slides.Rmd` (and whatever output you use for the presentation) and `predict-train.Rdata`, `predict-test.Rdata` `predict-validation.Rdata`.