# Housing Project Suggestions

Data science is not a linear process. In this project, in particular, you will likely find that EDA, data cleaning, and exploratory visualizations will constantly feed back into each other. Here’s an example:

1. During basic EDA, you identify many missing values in a column/feature.
2. You consult the data dictionary and use domain knowledge to decide *what* is meant by this missing feature.
3. You impute a reasonable value for the missing value.
4. You plot the distribution of your feature.
5. You realize what you imputed has negatively impacted your data quality.
6. You cycle back, re-load your clean data, re-think your approach, and find a better solution.

Then you move on to your next feature. *There are dozens of features in this dataset.*

Figuring out programmatically concise and repeatable ways to clean and explore your data will save you a lot of time.

The outline below does not necessarily cover every single thing that you will want to do in your project. You may choose to do some things in a slightly different order. Many students choose to work in a single notebook for this project. Others choose to separate sections out into separate notebooks. Check with your local instructor for their preference and further suggestions.

## EDA

* **Read the data dictionary.**
* Determine *what* missing values mean.
* Figure out what each categorical value represents.
* Identify outliers.
* Consider whether discrete values are better represented as categorical or continuous. (Are relationships to the target linear?)

## Data Cleaning

* Decide how to impute null values.
* Decide how to handle outliers.
* Do you want to combine any features?
* Do you want to have interaction terms?
* Do you want to manually drop collinear features?

## Exploratory Visualizations

* Look at distributions.
* Look at correlations.
* Look at relationships to target (scatter plots for continuous, box plots for categorical).

## Pre-processing

* One-hot encode categorical variables.
* Train/test split your data.
* Scale your data.
* Consider using automated feature selection.

## Modeling

* **Establish your baseline score.**
* Fit linear regression. Look at your coefficients. Are any of them wildly overblown?
* Fit lasso/ridge/elastic net with default parameters.
* Go back and remove features that might be causing issues in your models.
* Tune hyperparameters.
* **Identify a production model.** (This does not have to be your best performing Kaggle model, but rather the model that best answers your problem statement.)
* Refine and interpret your production model.

## Inferential Visualizations

* Look at feature loadings.
* Look at how accurate your predictions are.
* Is there a pattern to your errors? Consider reworking your model to address this.

## Business Recommendations

* Which features appear to add the most value to a home?
* Which features hurt the value of a home the most?
* What are things that homeowners could improve in their homes to increase the value?
* What neighborhoods seem like they might be a good investment?
* Do you feel that this model will generalize to other cities? How could you revise your model to make it more universal OR what date would you need from another city to make a comparable model?

# Example Directory Structure

Here’s how you might structure a project with multiple notebooks.

project-2  
|\_\_ code  
| |\_\_ 01\_EDA\_and\_Cleaning.ipynb   
| |\_\_ 02\_Preprocessing\_and\_Feature\_Engineering.ipynb   
| |\_\_ 03\_Model\_Benchmarks.ipynb  
| |\_\_ 04\_Model\_Tuning.ipynb   
| |\_\_ 05\_Production\_Model\_and\_Insights.ipynb  
| |\_\_ 06\_Kaggle\_Submissions.ipynb   
|\_\_ data  
| |\_\_ train.csv  
| |\_\_ test.csv  
| |\_\_ submission\_lasso.csv  
| |\_\_ submission\_ridge.csv  
|\_\_ images  
| |\_\_ coefficients.png  
| |\_\_ neighborhoods.png  
| |\_\_ predictions.png  
|\_\_ presentation.pdf  
|\_\_ README.md