Data Cleaning and EDA

Good inclusion of a code check proving no NaNs after all cleaning is done, together with confirming their replacement value of 0 or 'None'.

Treatment of NaN Generally

"First convert null in int and flt columns to 0 (there are already many numeric columns with large numbers of 0s) and convert null in object columns to 'None' (there are already 1218 'None' values in masvnrtype column). Which columns to drop will be decided later based on how stongly correlated the features are with saleprice"

Rationale above works for many cases, yet it is too broad an application without further investigation. NaN values could also be the result of typos, or columns with data lost when combining datasets.

"Grouping of porch and pool columns into single column indicating presence of features

Porch columns (3ssnporch, enclosedporch, openporch, screenporch) have many None or 0 values. However, they have relatively high correlation with sale price, so they should not be dropped."

This is a very interesting approach. However, rather than dropping related columns such as "poolarea, poolqc", consider breaking up the dataset into two subdatasets, and hence 2 models:

Subdataset 1 based on rows with pools built and with all pool related columns still there.

Subdataset 2 based on rows with no pools built and all pool columns removed.

Models could then be done for each different subset to isolate better and extract insights from such variables.

Predictions could even be made on classification off each sub-model, then combined for a final result.

Distributions Described

Histograms well plotted, with brief overall observations described. Can be more detailed.

Outliers Identified and Addressed

Outliers filtered, and dropped. Would have been good to display their summarized details before dropping.

Summary Statistics Provided

Done with highlights.

Student Addressed Likelihood of Answering Problem Statement with EDA Discoveries

Not done at the EDA stage.

Pre-processing and Modelling

Categorical Variables One-hot Encoded?

Yes.

"train_dummies = pd.get_dummies(train, drop_first = True)" should use drop_first = False instead, as this keeps all sub-values of a column. This will prevent columns from being lost when multiple columns are being dummied. Drop_first = True is more applicable for single column being dummied for multi-classification problems.

Investigate/Manufacture Features with Linear Relationships to Target?

Heatmaps for Correlation done.

Scatterplots have trend line to visualize the possibility of linear relationships.

Discussion of multilinearity of variables and treatment done.

Data Scaled Appropriately?

Standard Scaler was applied.

Train/Test Split and Validation/Training

Train_test_split and cross_val used.

Feature Selection to Remove Noisy or Multi-Collinear Features?

Yes, well done. Addressed and justifications given.

Variety of Models

Linear, Lasso, Ridge, ElasticNet Models.

Defense of Model Choice

R2, RMSE were consistently extracted for different models. Comparison was done off-page, not explicitly shown in notebook.

Should have table summary of scores of models against each other, and against baseline too.

Explicit explanation that a higher or lower score for a metric will demonstrate better or worse performance, would be better too.

"Further verification by measuring the train and test set r2 scores and RMSE of Elastic Net showed that it has the lowest r2 and RMSE of the 4 models." It has to be clear that r2 closest to 1.0 is desired, rather than having lowest numerical value for both metrics.

Explanation of Model and Evaluation of Success/Pitfalls

High coefficient variables were linked to importance of variables on model.

Should link in reasons why variables selected as high coefficient by model makes sense (it was briefly mentioned that square footage has high impact, but should clearly state high positive impact and why so)

Explain limitations of prediction by the model, and reasons why. (eg. economic crisis in some years skewing the sales price, which are not in datasets provided, thus skewing the model)

<u>Evaluation and Conceptual Understanding</u>

Identify and Explain Baseline Score

Baseline set was that of naïve model based on all predictions having same mean value. Clear and Succinct. Good.

Select and Use Appropriate Model Evaluation Metrics

R2, RMSE.

No summary table.

More than one Metric Used?

Yes

Interpretation of Results for Inference

Interpretation of results done, some inference made on how if certain features specific to Ames were removed, it could be tested on other cities.

Domain Knowledge Demonstrated

External research done to tie in to insights from model. Good.

Interpretation of Descriptive and Inferential Statistics

Descriptive Statistics are commented on throughout the notebook, but not organized in a summary section on its own.

No significant Inferential Statistics.

Final Comments

Excellent visuals, especially for subplots, colour choice was clear and distinct, and important, given that there were many plots concentrated together. However, some nested plots, due to small size, had axis labels that were too small. Consider breaking up plots across different code blocks so size can be larger.

Please try to use snake case where possible, lower case letters is great, but use underscore to separate words.

Very good feature engineering!