Data Cleaning and EDA

Good inclusion of a code check proving no NaNs after all cleaning is done.

Treatment of NaN Generally

Very good approach by first categorizing the type of features, then addressing the challenges faced.

Particularly, “The tricky thing about creating dummy variables for ordinal variables, is that it can be subjective (eg. no basement is worse than a poor basement). As such, these ordered variables will in general be assigned a range of integers [-2,-1,0,1,2] whereby a positive score denotes a more favorable trait, and vice versa for a negative score.

In variables where there appears to be no real negative(or positive) responses, a non-negative(or non-positive) range can be used to approximate the relationship of the variable (eg. [0,1,2,3] or [-3,-2,-1,0])

Lastly, in my opinion, the weightage of the extreme ends of the variables can be allocated a higher value. (eg.[-5,-2,0,2,5] instead of [-2,-1,0,1,2])”

Distributions Described

Histogram plotted for only SalesPrice. Good discussion though.

Outliers Identified and Addressed

Outliers filtered, displayed and dropped. Good discussion, and testing of model with and without dropping them.

Summary Statistics Provided

Done with good discussion.

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.

Investigate/Manufacture Features with Linear Relationships to Target?

Heatmap for Correlation done.

Scatterplot done for manually selected features, has trend line to visualize the possibility of linear relationships.

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?

Some manual feature selection, most feature selection declared done via Lasso. Still good.

Variety of Models

Linear, Lasso, Ridge, ElasticNet Models.

Defense of Model Choice

Cross_val_score used as main metric across models. No summary table.

Subsequent defense based on R2, RMSE and overfitting. Explanation of variance scoring is a plus here.

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

Explanation of Model and Evaluation of Success/Pitfalls

High coefficient variables not shown nor linked to importance of variables on model.

"The line represents the points where the predited values is equal to the actual values. In other words, this line is the visual representation of the predictions of our model. As can be seen, it appears that the accuracy of the model appears to be more precise in the lower end of the dataset. As was previously mentioned, the dataset has more instances of data occuring in the lower values of SalePrice, and as such would likely have a greater impact on the model.

This would also be part of the reason why the model seems to be less accurate in predicting the higher valued properties." Good.

Evaluation and Conceptual Understanding

Identify and Explain Baseline Score

No explicitly identified baseline model, only reference made to initial best performing model.

Select and Use Appropriate Model Evaluation Metrics

R2, RMSE.

No summary table.

More than one Metric Used?

Yes, but not when doing cross model evaluation.

Interpretation of Results for Inference

Interpretation of results done, generalized discussion made, could be more explicit.

Domain Knowledge Demonstrated

Shown in considerations when dealing with NaN.

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

Executive Summary in readme was more like a table of contents. The material is already there to write a good ES. Please reorganize them into the ES section.