Main objective of the analysis that specifies whether your model will be focused on prediction or interpretation.

Main objective is to explore different variants of Linear Regression models to forecast Sale Price of properties offered by Ames Housing, with an emphasis on interpretation to identify key attributes consumers value the most and the least to help guide Ames' business strategy.

Brief description of the data set you chose and a summary of its attributes.

Given Ames Housing data on 1379 properties sold during 2006-2010, with 80 feature measurements covering the following:

1) construction history (e.g. year built, year remodeled, lot area, lot shape);
2) infrastructure (e.g. #total rooms, garage area, utility, pool);
3) assessments & styles (e.g. overall quality, overall condition, neighborhood, house style);
4) sale information (e.g sale type, sale condition, year/month sold, sale price)

Brief summary of data exploration and actions taken for data cleaning and feature engineering.

1) Identify target variable: SalePrice
   - mean: $185,500
   - standard deviation: $79,000
   - min/max range: $35,300 ~ $755,000

2) Focus on 15 key contributing features:
   - categorical: LotShape, GarageType, Neighborhood, HouseStyle, SaleCondition, SaleType
   - numerical: OverallCond, OverallQual, LotArea, TotRmsAbvGrd, GrLivArea, GarageArea, YearBuilt, YearRemodAdd, YrSold

3) Data Cleaning:
   - missing values: none, no imputation needed.
   - outliers: a few properties with extremely high LotArea can be removed if prediction is emphasized. In our case, no action taken to remove/mask/reevaluate those outliers.

4) Feature Engineering
   - skew transformation: 3 numerical features (LotArea, GrLivArea, GarageArea) and the target variable (SalePrice) are positively skewed (>0.75). Log/BoxCox transformation conducted to convert them to be more Normally distributed and improve the performance of Linear Regression models.
   - scaling: numerical features are on drastically different scales. Standard Scaling conducted to standardize numerical variables to comparable scales.
   - encoding: 6 categorical features One-Hot Encoded into 52 columns, with the first dummy column for each feature dropped to prevent multi-colinearity.

- polynomial features: OverallQual^2 and GarageArea^2 added due to the curved positive correlation observed in their pair plots against SalePrice.

- feature interactions: OverallQual*OverallCond and LotArea/TotRmsAbvGrd added to empirically capture “Overall Assessment” and “Average Area per Room” that could be insightful.

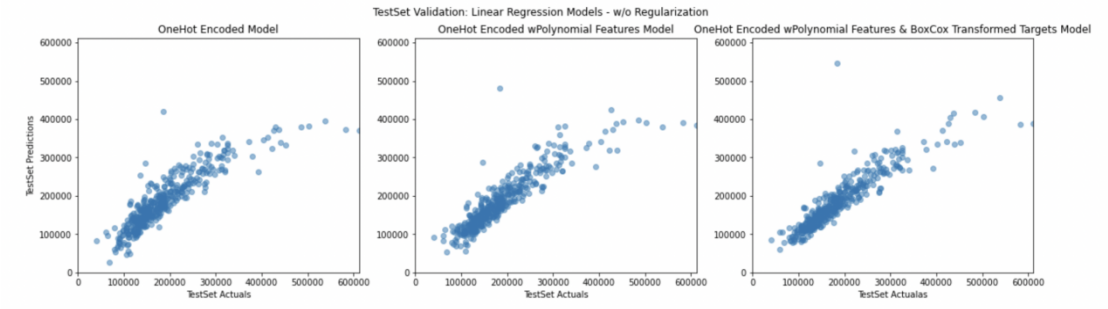
Summary of training at least three linear regression models which should be variations that cover using a simple linear regression as a baseline, adding polynomial effects, and using a regularization regression. Preferably, all use the same training and test splits, or the same cross-validation method.

1) Linear Regression without Regularization

Same train-test splits were used for validating and comparing across different variants of Baseline Linear Regression Models shown below.

– Scaling has no effect on plain vanilla Linear Regression.
– Polynomial terms have added significant predictive power to SalePrice forecast, in that it’s able to reduce train-set error by 18%, and test-set error by 12%.
– Boxcox transformation on target variable has further improved the forecast with a 13% error reduction on train-set and 4% on test-set, and also works better with outliers.

	train	test
OneHot Encoded Baseline LR Model	1.052933e+09	1.419160e+09
OneHot Encoded w/ Scaling Baseline LR Model	1.052933e+09	1.419160e+09
OneHot Encoded w/ Polynomial Features Baseline LR Model	8.641747e+08	1.252315e+09
OneHot Encoded w/ Polynomial Features and Target BoxCox Transformed Baseline LR Model	7.546521e+08	1.206515e+09



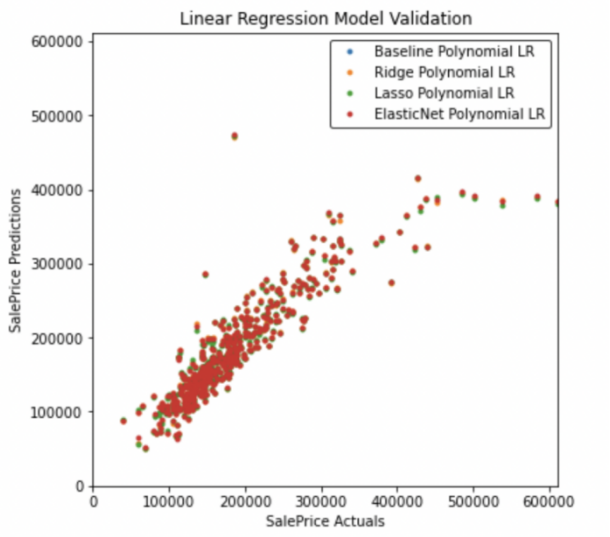
2) Linear Regression with Regularization

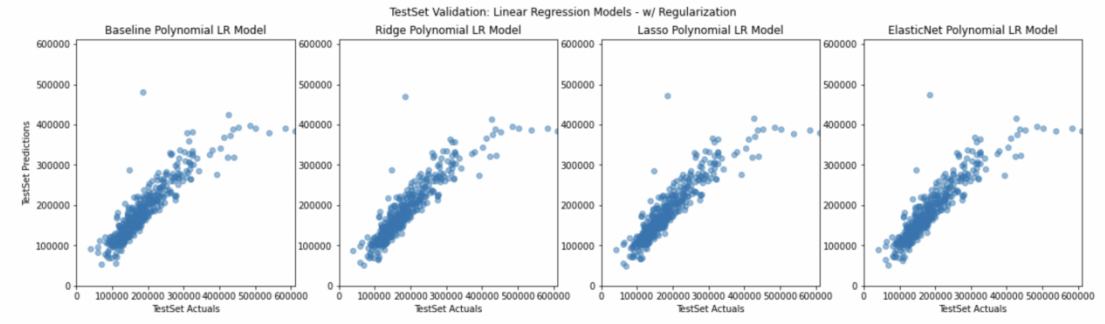
Same KFold cross-validation method was used for validating and comparing across different variants of Regularized Linear Regression Models shown below.

– Ridge, Lasso, ElasticNet regression perform comparably in terms of test-set accuracy (measured using RMSE: root-mean-square error)
– While Ridge & Lasso run pretty efficiently, ElasticNet could take some time for hyperparameter tuning on alpha and L1_ratio. Depending on the pre-specified GridSearch ranges and max iterations,

ElasticNet could end up performing worse than Ridge/Lasso with enormously more computational power consumed.

	RMSE
Baseline LR	37671.746307
Baseline Polynomial LR	35388.063017
Ridge Polynomial LR	35054.637584
Lasso Polynomial LR	35189.753337
ElasticNet Polynomial LR	35092.585685





A paragraph explaining which of your regressions you recommend as a final model that best fits your needs in terms of accuracy and explainability.

Ridge Regression model with polynomial features added will be recommended as the final model for our purpose. Reasons being that:

1) It provides the highest test-set accuracy with affordable run time, while empirically added polynomial features / feature interactions are proved to have noticeable predictive power and serve as good source for business insights to suit our interpretation needs.

2) ElasticNet has the potential to create a more accurate model with significantly more computational and/or analytical resources that could easily outweigh the gain on interpretation here.

Summary Key Findings and Insights, which walks your reader through the main drivers of your model and insights from your data derived from your linear regression model.

Key attributes that consumers value the most:

– OverallCond, OverallQual^2: overall condition and quality of the property
– LotArea/TotRmsAbvGrd, GrLivArea: spaciousness of the property, esp. Living area
– YearBuilt: whether property was built in recent years
– SaleType_CWD/New/Con: specific sale types are more popular than others
– Neighborhood_MeadowV/StoneBr/NoRidge: specific neighborhoods are more popular than others

Key attributes that consumers value the least:

– LotShape_IR3: IR3 may not be a popular lot shape
– GarageType_CarPort: CarPort may not be a popular garage type
– SaleCondition_Partial: Partial may be a rather unacceptable sale condition
– Neighborhood_Edwards/OldTown/NWAmes: specific neighborhoods are less popular than others

While there are definitely some sale types or neighborhoods that are more or less popular among consumers, consumers tend to value the high–level aggregated assessments (such as OverallCond, OverallQual, YearBuilt etc.) as reflected in settled sale price. However, it usually takes only a specific "stylistic" thing to "break the deal" – such as the LotShape being IR3, GarageType being CarPort etc.

Suggestions for next steps in analyzing this data, which may include suggesting revisiting this model adding specific data features to achieve a better explanation or a better prediction.

This data sets contains a sufficiently large sample and a wide variety of feature measurements of descent data quality for analytical purposes. One major issue with the data quality may be its lack of timeliness, as it stores data on properties sold more than 10 years ago and may not reflect a trend that's still applicable to the housing industry today.

For that matter, additional data on properties sold between 2010 and now would definitely prove useful. And any information on property purchaser should also help with consumer segmentation and improve the predictive power and the interpretability of our model further. This in turn will lead to analyses / business recommendations that are more tailored and accurate.