Predicting Housing Prices
in Ames, Iowa

A presentation from **Room 6**

May Gee | Chee Tzen | Jia Wen | Hong Aik | Nazira

Background

- Housing is commonly seen as an investment tool to flip for handsome profits.
- Homebuyers (primary stakeholder) are keen to identify features which the value of the house is sensitive to in order to **maximise capital appreciation**.
- Homeowners (secondary stakeholder) may also want to renovate their houses to i**mprove its marketability**.

Problem Statement

- How can homebuyers in Ames identify houses which have great appreciation value?
- How can homeowners in Ames improve their house values?

Using a dataset of housing prices in Ames, we have built a regression model to predict housing prices given its features, and in the process identify features which significantly affect housing prices.

Evaluation of the model will be done via root-mean-squared-error (RMSE), which is the average difference between the predicted SalePrice and the actual SalePrice.

Data Source

Data was taken from Kaggle: https://www.kaggle.com/c/dsi-us-11-project-2-regression-challenge/data

2051

Number of houses

80

Number of house features

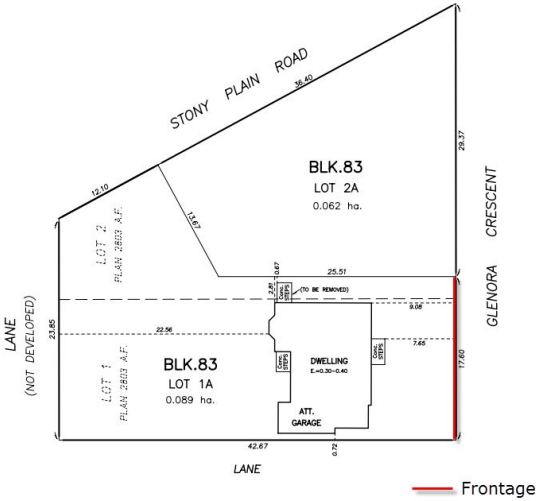
In Singapore...

In Ames...

Data Source

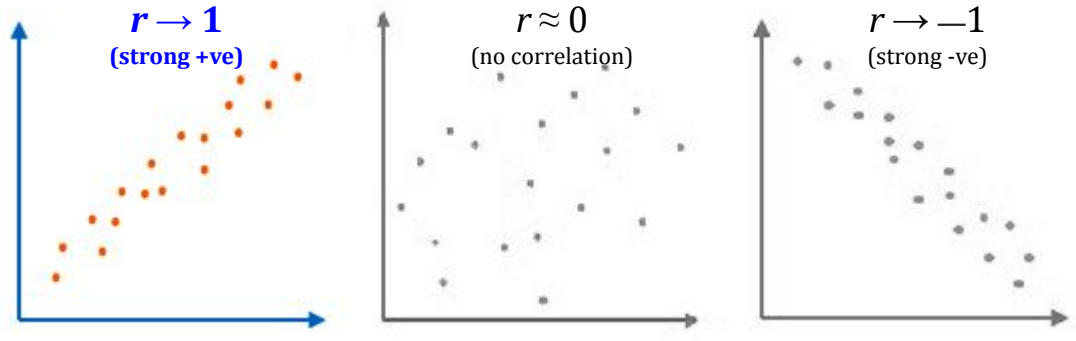
Number of fireplaces

Type of masonry veneer wall



Lot Area/Frontage

Data Exploration

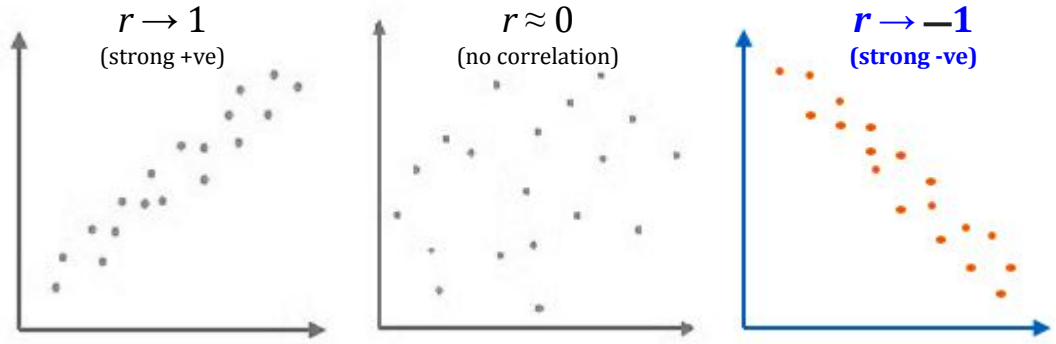


Overall Qual (Rates the overall material and finish of the house): 0.80

Ground Living Area: 0.70

Garage Area: 0.65

Data Exploration



Exterior Material Quality (Average/Typical): — 0.60

Age of house as at sale: — 0.57

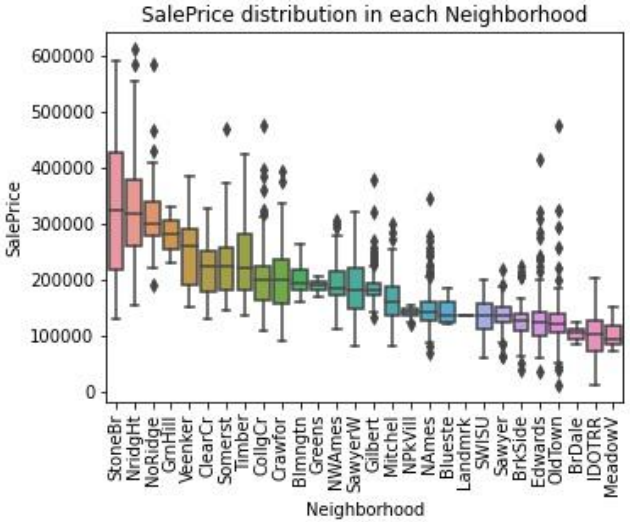
Kitchen Quality (Average/Typical): — 0.54

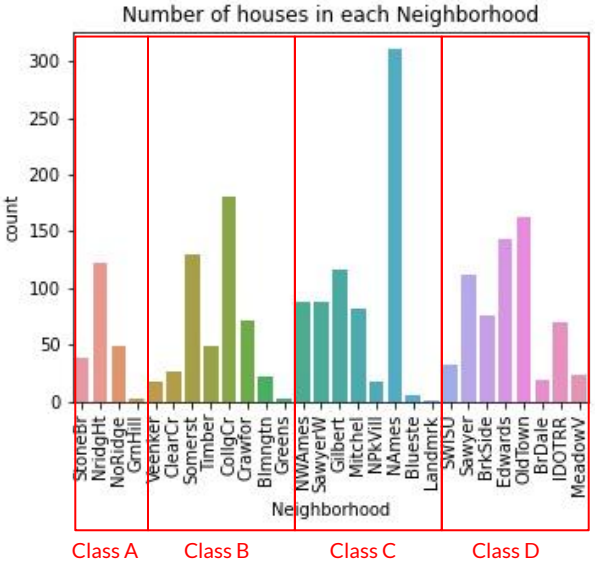
Feature Engineering

Feature	Correlation
Overall Qual	**0.80**
Ground Living Area	**0.70**
Garage Area	0.65

New Feature	Correlation
Overall Qual-Ground Living Area	**0.84**
Overall Qual-Garage Cars	0.82
Overall Qual-Garage Area	0.81

Feature Engineering





Group categorical columns into at most 4 categories based on

- median SalePrice
- number of observations

Model Results

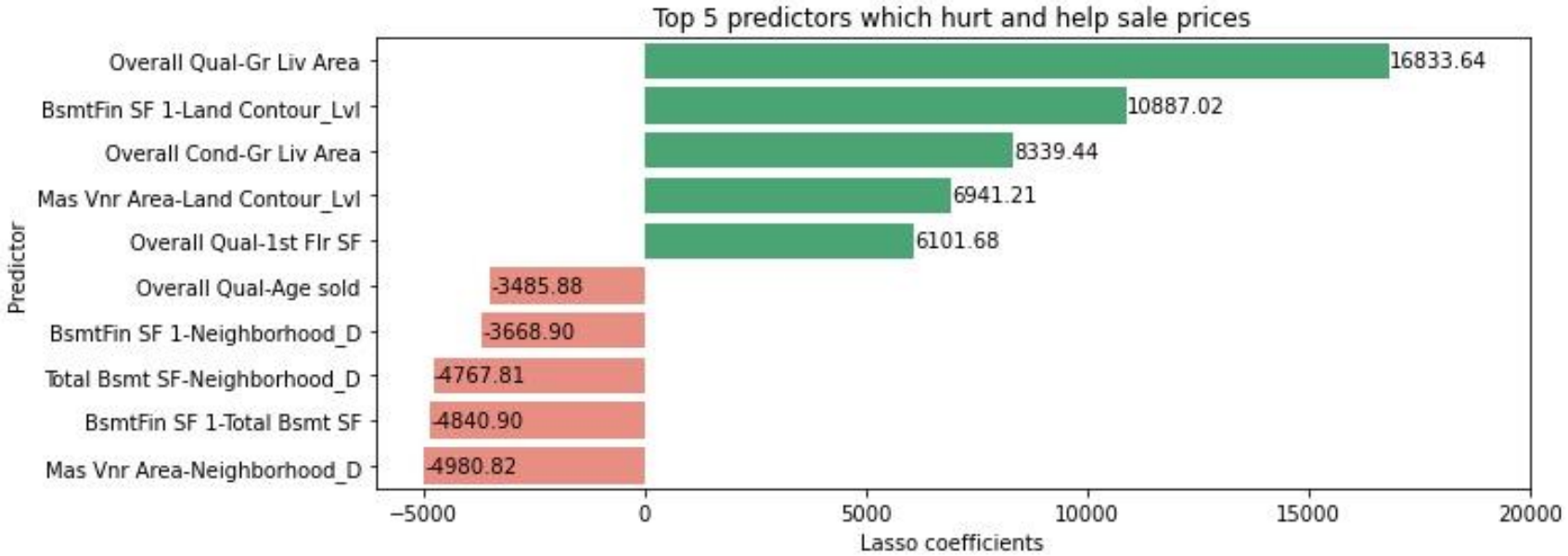
$20,468

Average error

~11%

Error Percentage of Mean SalePrice

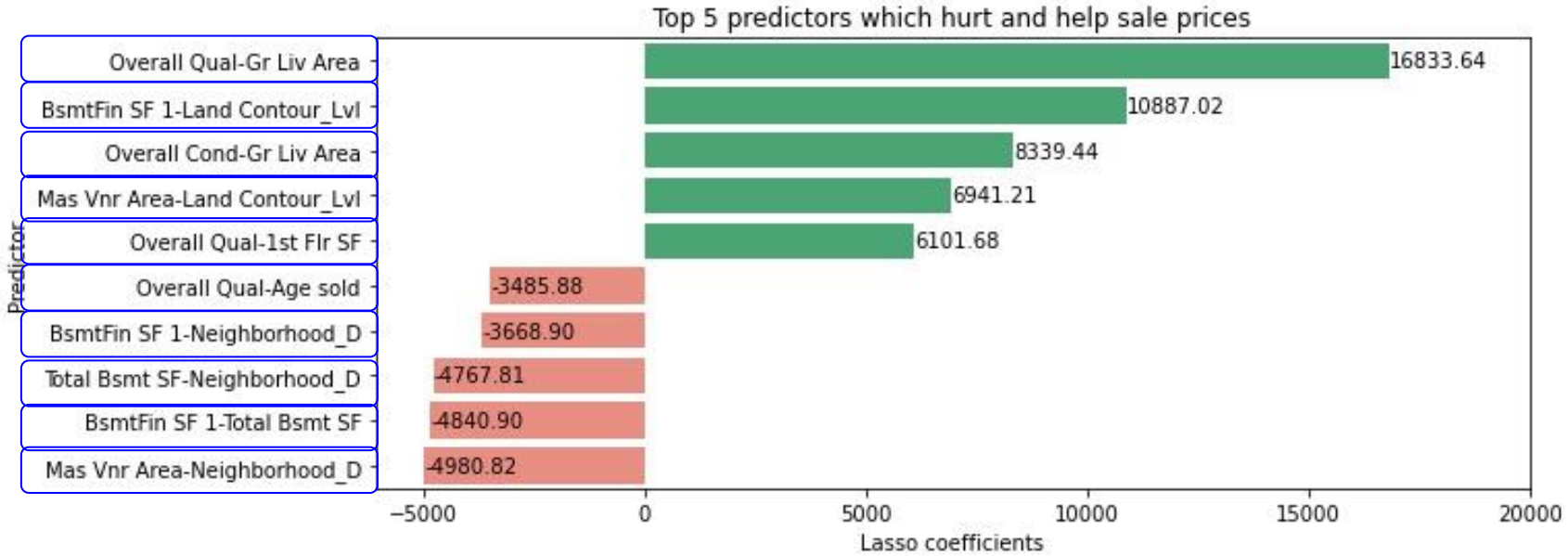
Insights from the Model



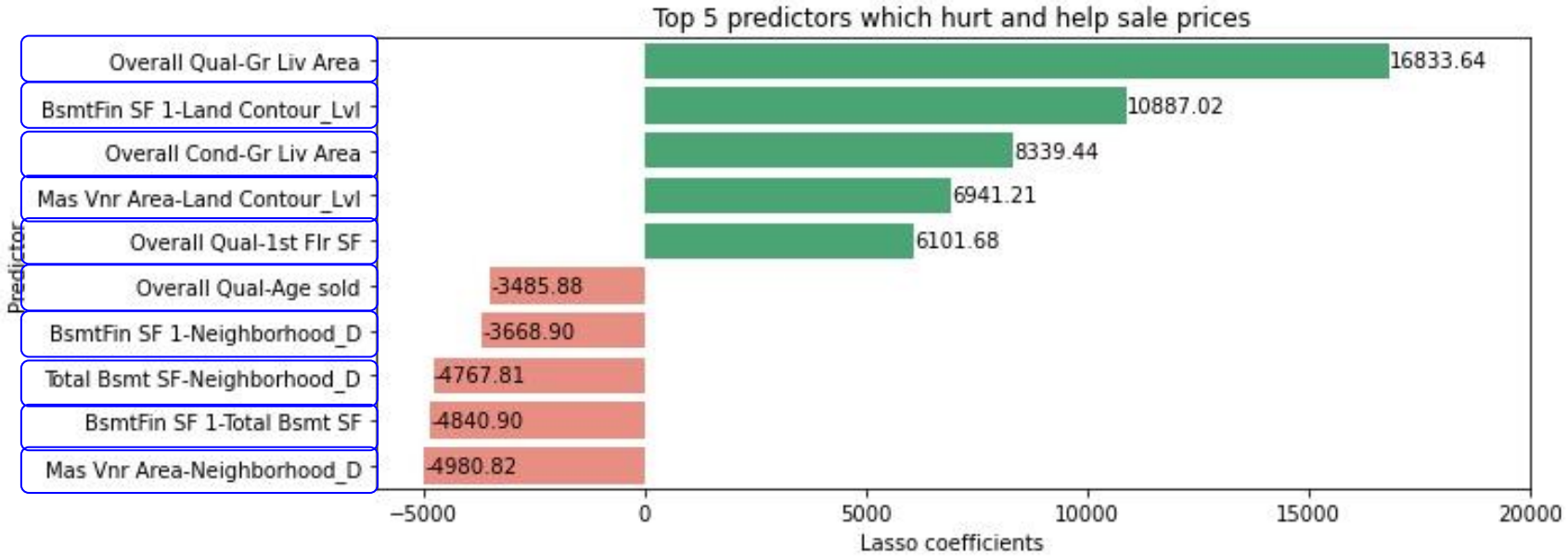
Insights from the Model

Unique house features in top 5 predictors that **improve** house prices	Unique house features in top 5 predictors that **hurt** house prices
1st Floor Area	Neighborhood D
Basement Finished Area	Basement Finished Area
Overall Quality	Overall Quality
Land Contour (Near Flat)	Total Basement Area
Masonry Veneer Area	Masonry Veneer Area
Overall Condition	Age sold
Ground Living Area	-

Recommendations for Homeowners



Recommendations for Homebuyers



Model Caveats

- Sensitive to context; unlikely to be generalizable. This is because two populations may not share the same “taste” for desirable features in a house
- To make the model more generalizable, reduce the number of variables, at the expense of predictive power.

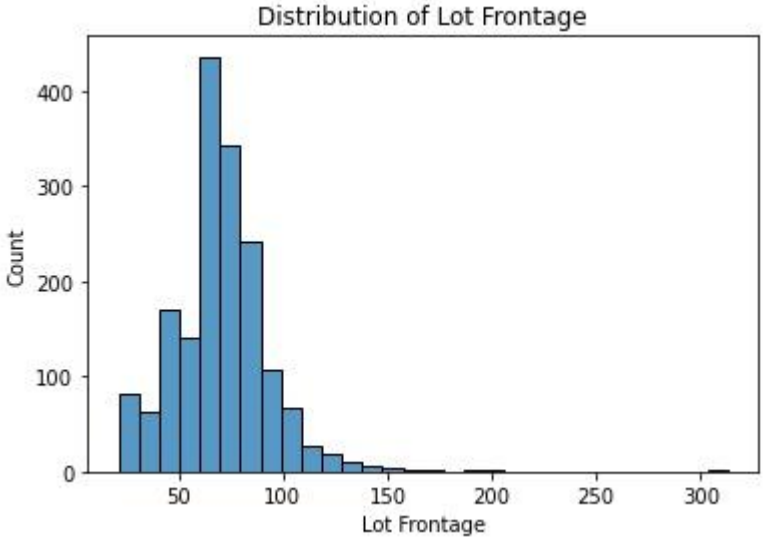
Possible Enhancements

- Trial other predictive models (eg. RandomForest, Support Vector Machine)
- Obtain more data, preferably recent
- Obtain more general features of the house such as crime rate, ethnicity, distance to amenities or facilities

Thank You

Annex

Lot Frontage



Ames Housing Dataset

- Describes the sale of individual residential property in Ames, Iowa from 2006 to 2010
- Dataset contains —
  - 80* explanatory variables to access home values
    - → **23 nominal (* example..), 23 ordinal (* example..), 14 discrete and 20 continuous*
- Variables are —
  - objective & quantitative (*i.e. year built, number of fireplaces*)
  - subjective & qualitative (*i.e. heating quality, exterior quality*)

Data Cleaning (Handling Nulls)

01	Lot Frontage	• Imputed with mean
02	Columns related to Masonry Veneer, Basement and Garage	• Categorical features not present → imputed with a standard string *(eg. 'No Garage')* • Related numerical features → imputed with 0
03	Other columns	• Null values have intended meanings, → imputed with standard string *(eg. 'No Fence')*

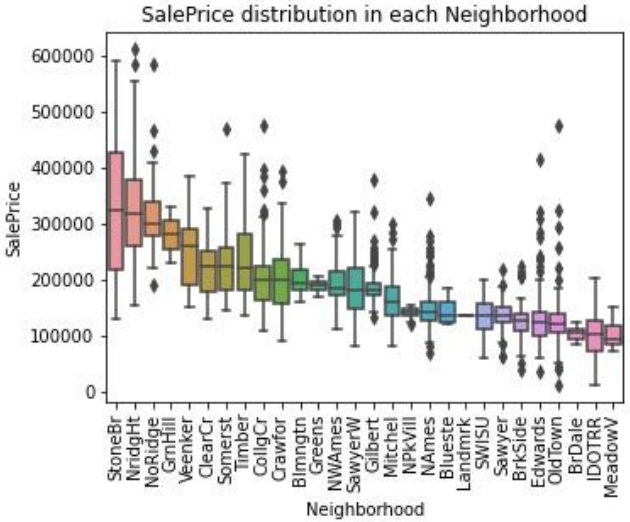
Data Cleaning (Dropping Columns)

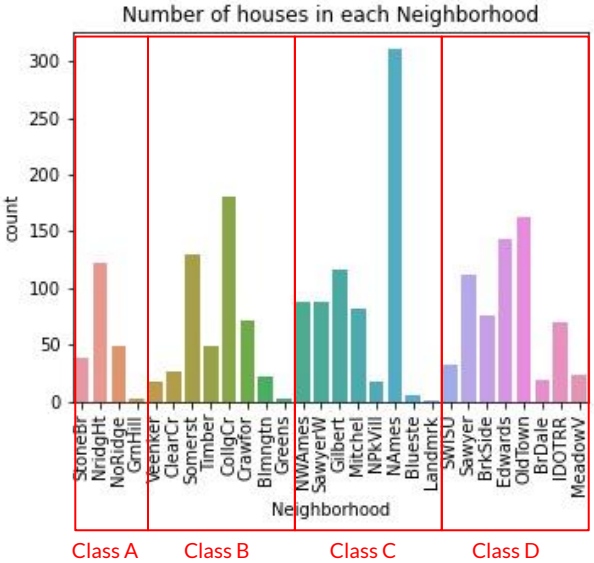
Dropped Columns	Rationale
Id, PID, Sale Type	Not related to SalePrice
Year Built	Replaced with **AgeSold**
Year Remod/Add	Replaced with **AgeRemodelled**
Pool Area, Pool QC	Vast majority of houses have no pool

Data Cleaning (Dropping Columns)

Dropped Columns	Overwhelmingly Single Value	Rationale
Condition 2	‘Norm’ - 98.7%	near-zero variance ⇒ likely contain little valuable predictive information
Utilities	‘AllPub’ - 99.9%	
Roof Matl	‘CompShg’ - 98.7%	
Heating	‘GasA’ - 98.3%	
Misc Feature / Misc Val	‘NA’ - 96.8%	
3Ssn Porch	‘0’ - 98.7%	

Data Cleaning (High Cardinality Categorical Variable)

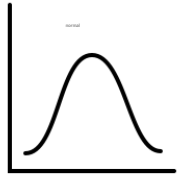


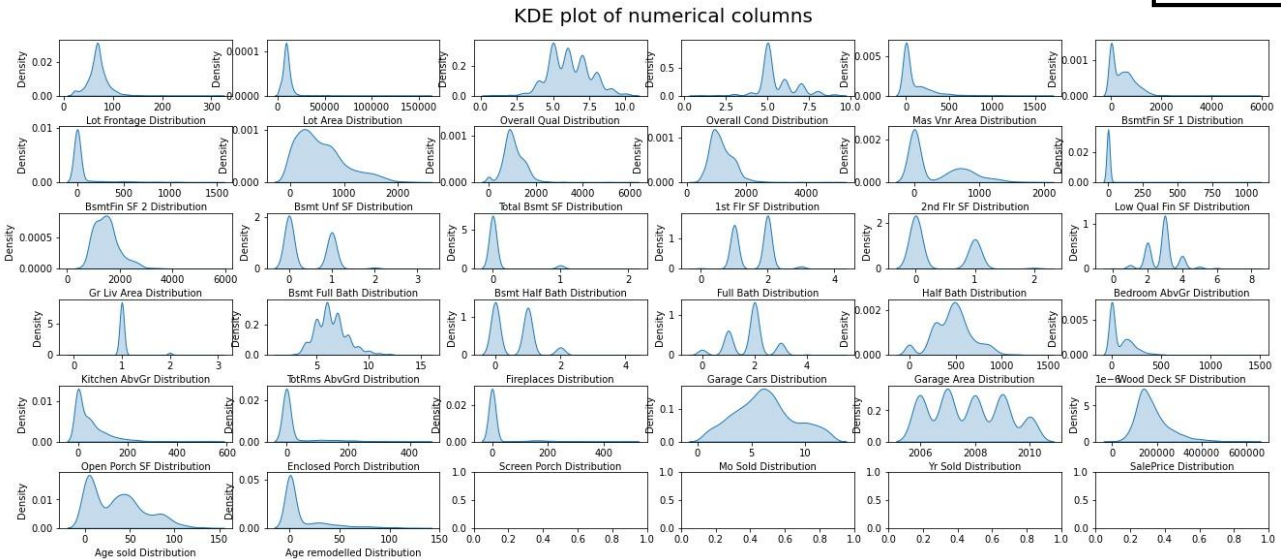


Group categorical columns into at most 4 categories based on

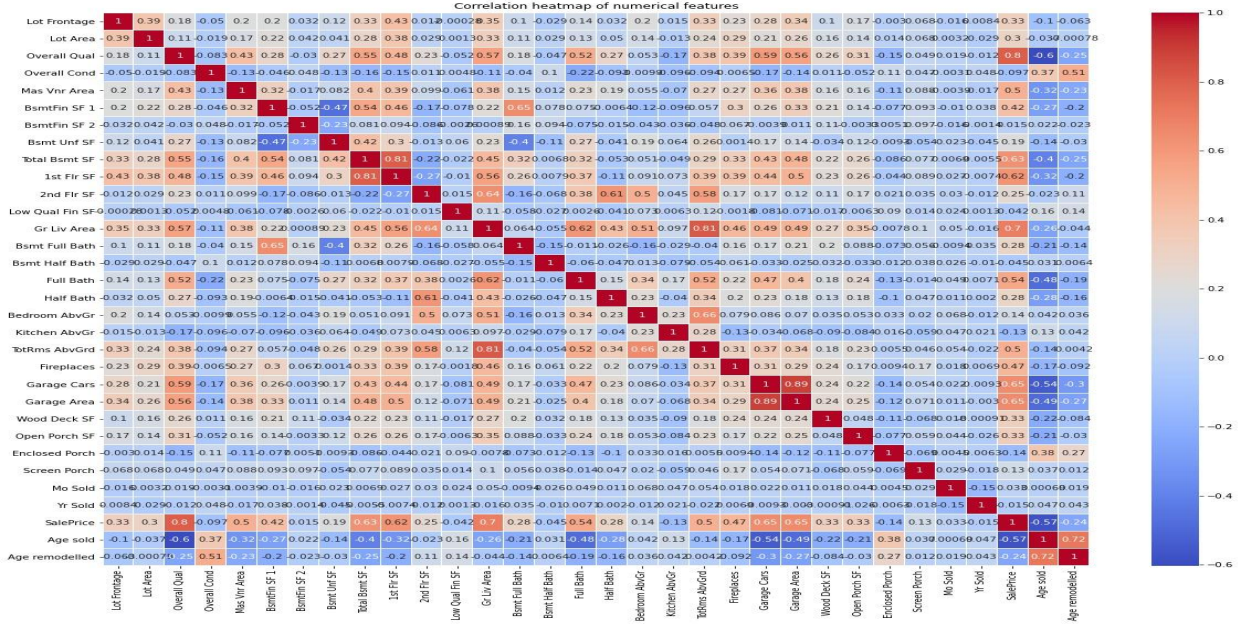
- median SalePrice
- number of observations

Data Cleaning (Numerical Columns)





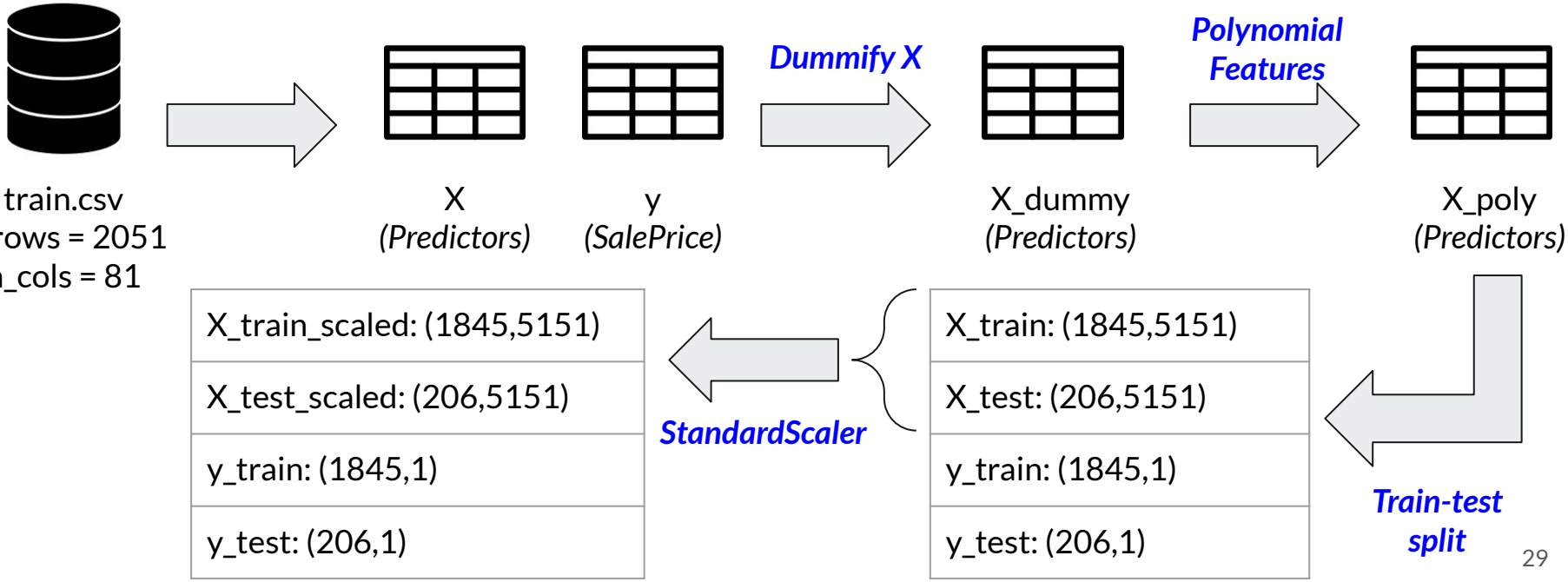
Exploratory Data Analysis



Most (numerical) features are not correlated with one another.

Instead of dropping related columns, we will leverage on interaction terms and regularization.

Model Preprocessing



Baseline Model

	Train R^2	Test R^2	Train RMSE	Test RMSE
Linear Regression	0.99999	-2.1296e-15	65	3.5408e+12
Mean SalePrice	0	-0.00595	79492	76727

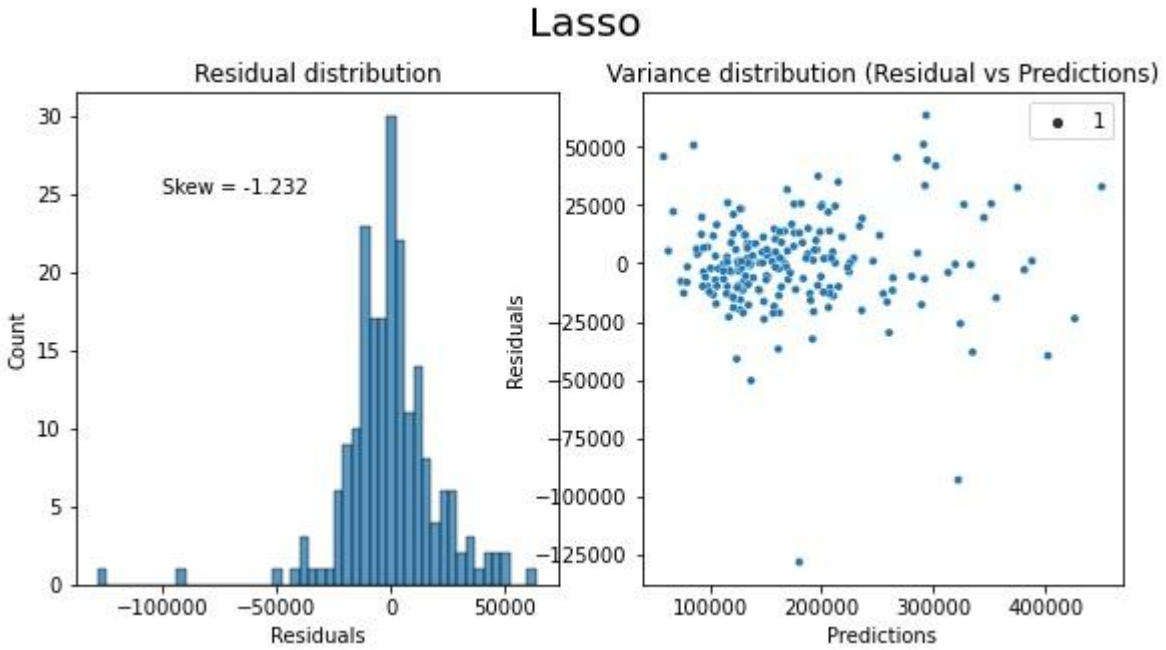
Baseline models, as expected, have performed poorly. Linear Regression has severely overfitted, while the null model has underfitted.

Model Selection

Model	Best alpha	Best L1 ratio	Train R^2	Test R^2	Train RMSE	Test RMSE
RidgeCV	2364	NA	0.95942	0.92596	16,012	20,877
LassoCV	497	NA	0.95688	0.92883	16,505	20,468
ElasticNetCV	31	0.95	0.95511	0.92539	16,842	20,956

The lasso model has performed the best, which is heartening, because it also has the ability to "remove" unnecessary features (94%).

Testing Linearity Assumptions



While the skew value indicates a considerable amount of right skewness, there appears to be no discernible pattern in the residuals.

Hence, we can consider the assumptions satisfied.

Stacking Models (only for Kaggle)