Ames Housing Set

Predicting House Prices Using Regression

Introduction & Problem Statement

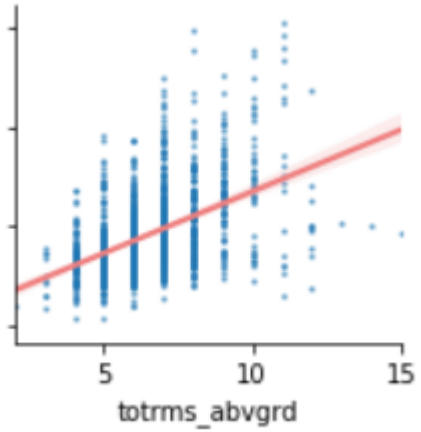
- Ames Housing Dataset
  - >2000 Observations, >80 Features
- Target: Predicting Sale Price
- Construct a multiple linear regression model to predict
- **“How can we best predict the sale price of a house using a linear regression model?”**

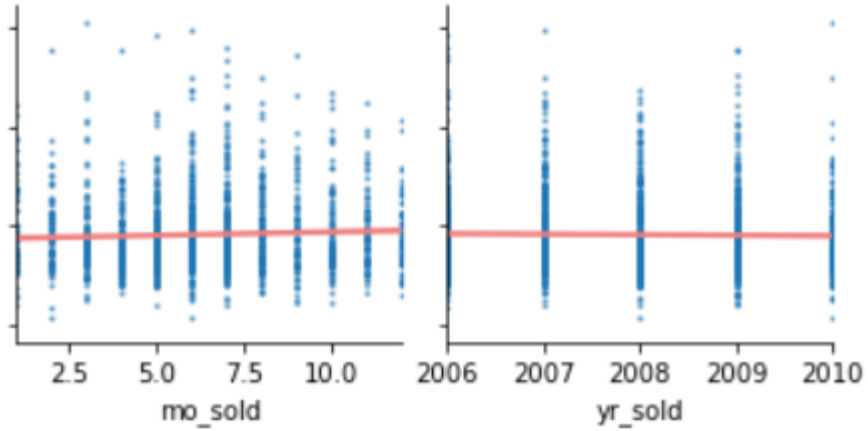
Data Cleaning & Feature Engineering

- Train-validation split
- Dropping of obviously useless features
- Null Imputation
  - Large no of null values (>100)
    - Impute ‘0’ / ‘NA’
  - Intermediate no of null values (between 5 – 100)
    - Impute ‘0’ / ‘NA’
    - Impute by mode
  - Small no of null values (<5)
    - Impute by mode
    - Tried to avoid dropping rows

Classifying Features

- Nominal / Ordinal / Continuous
- Treated differently during EDA & Preprocessing
- Discrete Features: Categorical or Continuous?
  - Range of values
  - Linear relationship with target?





Ordinal Features

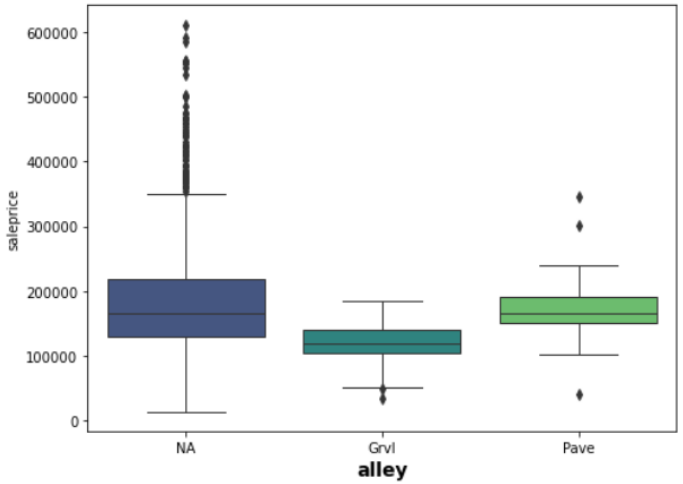
- Ranking numerically vs treating as nominal
- Implied ordering lost when treated as nominal
- Transform ordinal features into ranked numeric values
  - Arranged by order in data dictionary
  - Mode for each feature assigned 0
  - NA values assigned appropriate rank based on median value

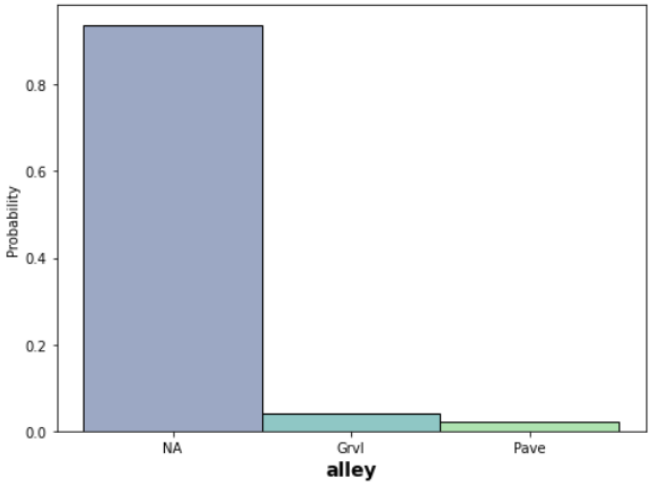
```
# Checking that ordinal features are numeric
for feature in feature_types['ordinal']:
    print(f'{feature}: {train[feature].unique()}')
```

```
lot_shape: [-1  0 -2 -3]
utilities: [ 0 -1 -2]
land_slope: [ 0 -1 -2]
overall_qual: [ 7  8  6  5  4  9 10  3  2  1]
overall_cond: [5 6 7 8 4 9 3 2 1]
exter_qual: [ 1  0  2 -1]
exter_cond: [ 0  1  2 -1 -2]
bsmt_qual: [-2  1  0  2 -1 -3]
bsmt_cond: [-2  0 -1  1 -3  2]
bsmt_exposure: [-1  0  1  3  2]
bsmtfin_type_1: [-6 -5 -1 -4 -3  0 -2]
bsmtfin_type_2: [-1  0  5  1  2  3  4]
heating_qc: [-1 -2  0 -3 -4]
electrical: [ 0 -1 -2 -3 -4]
bsmt_full_bath: [0. 1. 3. 2.]
bsmt_half_bath: [0. 1. 2.]
full_bath: [2 1 3 0 4]
half_bath: [0 1 2]
```

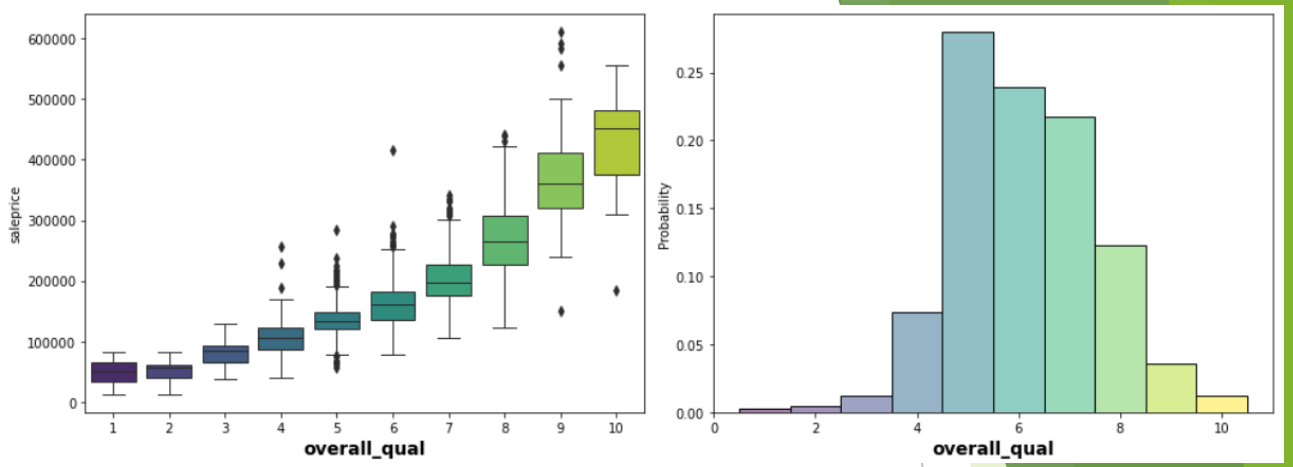
EDA – Nominal Data

- Explore boxplot & bar chart together
- Looking for extreme skew in values, overlapping IQRs

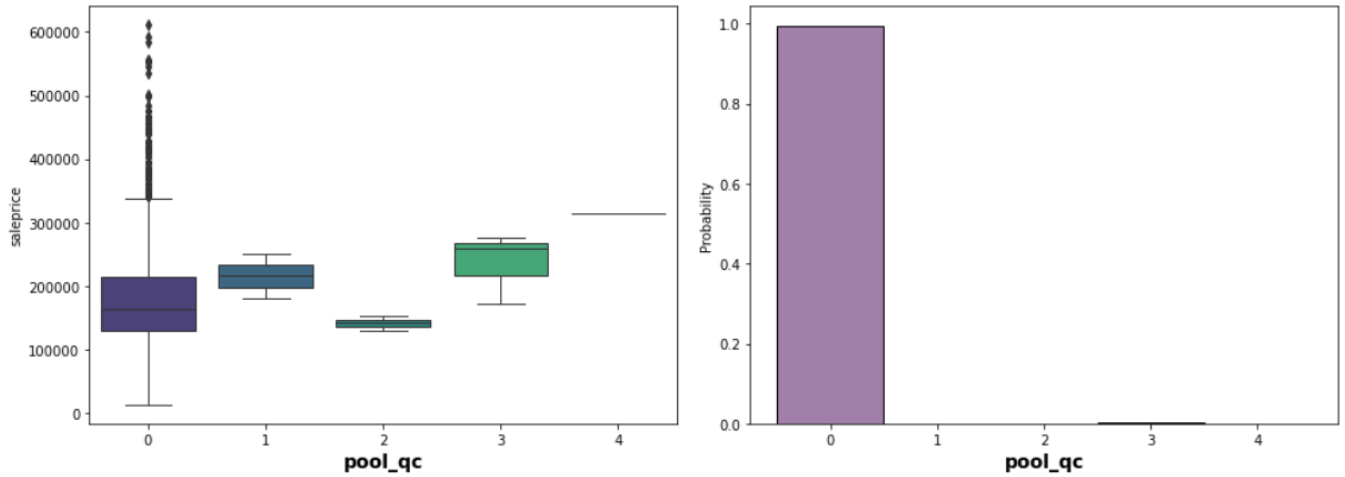




EDA – Ordinal Data

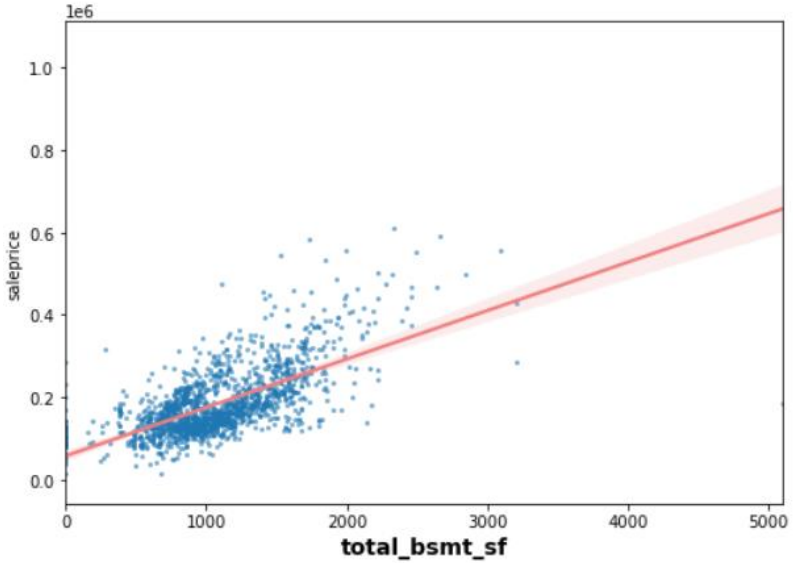


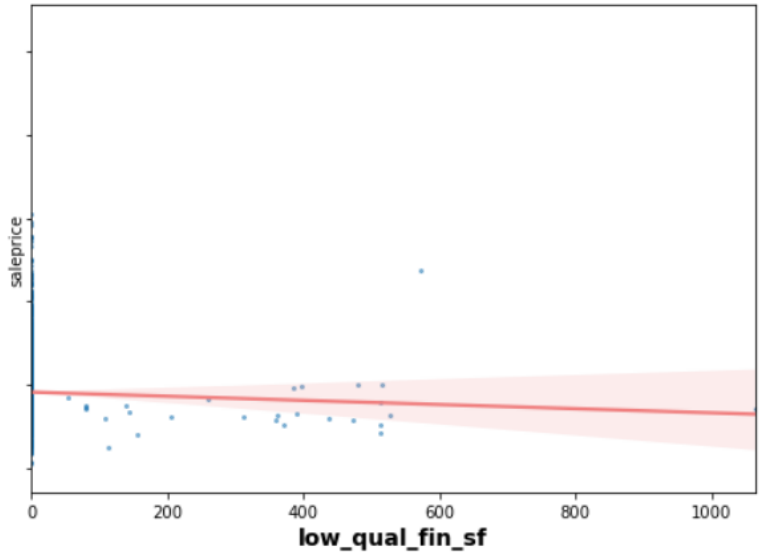
- Explore boxplot & bar chart together
- Looking for extreme skew in values, overlapping IQRs
  - Pay more attention to boxplot
  - Distribution of sale prices should reflect trend following order



EDA – Continuous Data

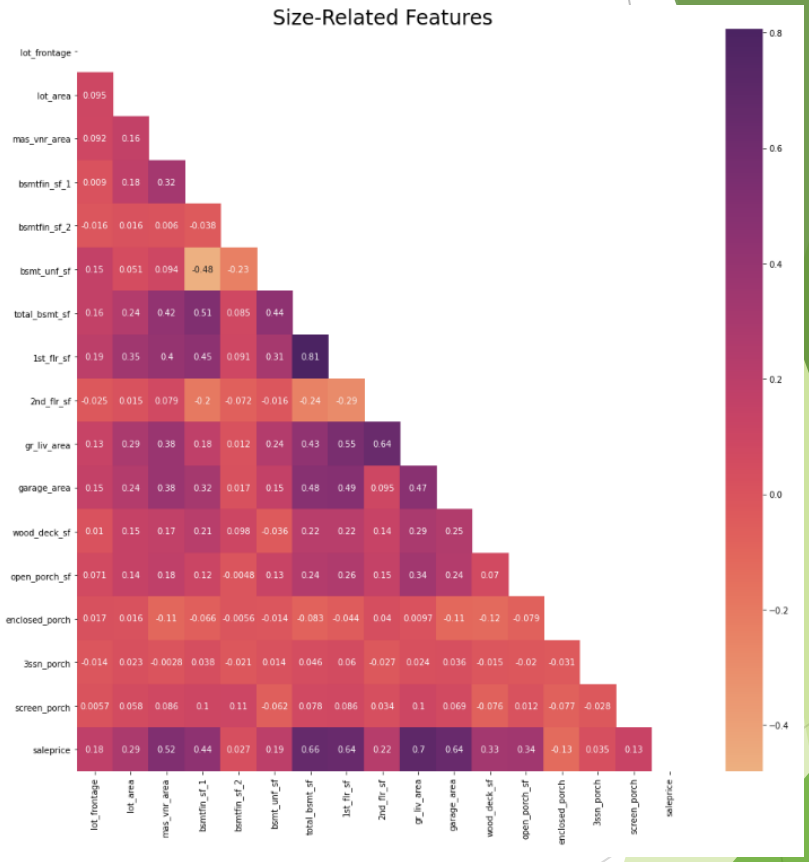
- Study scatterplot against Sale Price
- Find trends (or lack of trends)





Feature Engineering

- Year → Age
- Explore heatmaps (Correlation with one another)
- Combine / Add / Create Interaction Features



Outliers

- Removed a few following data dictionary
- Left the rest
  - Increase variance, avoid overfitting to training data

Preprocessing

- One-hot Encoding
  - Nominal Features
- Scaling
  - Continuous + Ordinal Features

- Apply all changes to:
  - Full train set (incl validation)
  - Validation set
  - Test set

Creating Prediction Model

- Null regression model as baseline
  - Train target mean as predicted value
  - RMSE Score of **83689.75**
  - Minimum score to beat for actual models
- Model Preparation
  - Split into X matrix & y vectors for all datasets

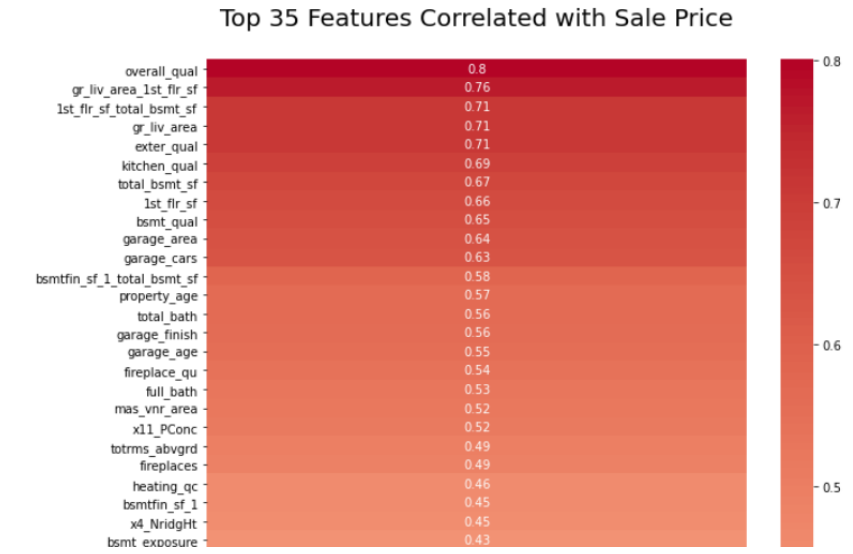
Full Feature Models

- 187 features after cleaning & preprocessing
- Run 3 models using all 187 features
  - Linear Regression
  - Ridge Regression
  - Lasso Regression

Model	No. of Features Input	Cross Val Score	Validation Score (RMSE)
null regression	-	-	83689.75 (Kaggle)
lr_full	187	1.9E+15	5.13E+15
ridge_full	187	23179.81	79571.38
lasso_full	187	23030.92	77536.92

Feature Selection

- Reduce noise fed to the model
- Two ways:
  - Filtering by correlation with sale price
  - Recursive Feature Elimination
    - Using 3 models as well
- Ended up using combination of both methods
  - RFE (Lasso) to 75 features, then filter
  - Filtered to 20 – 35, compared results



Final Model

- Lasso Regression
- 35 Features

No. of Features	LR CVS	LR Validation Score	Ridge CVS	Ridge Validation Score	Lasso CVS	Lasso Validation Score
20	25623.79	87152.86	25615.53	86218.1341	25672.33	85773.27
25	25150.09	88179.55	25129.72	87302.8222	25227.76	86079.82
30	24776.5	85453.67	24750.55	84405.0116	24861.41	83118.74
35	24496.37	80448.76	24460.35	79223.2212	24600.17	78229.86

- Final RMSE Score: **23064.34**

Conclusions

- Features that added the most value included:
  - High value neighbourhoods
  - General size of house
  - Overall quality
- Features that caused biggest drops in value:
  - Property Age
  - Certain types of finishes e.g. wooden sidings
  - Basement size
    - Counter-intuitive
    - Relationship with interaction features which added more value