



Property Price Prediction

Using the Ames Housing dataset

Content

- Problem statement
- Data preprocessing
- Feature selection
- Prediction models
- Key findings
- Conclusion and Recommendation

Problem statement

- What are the factors that will affect housing prices in Ames?
- By knowing these factors, one can provide insights or advice to home owners:
 - who are considering selling their property.
 - who are considering renovating their property.

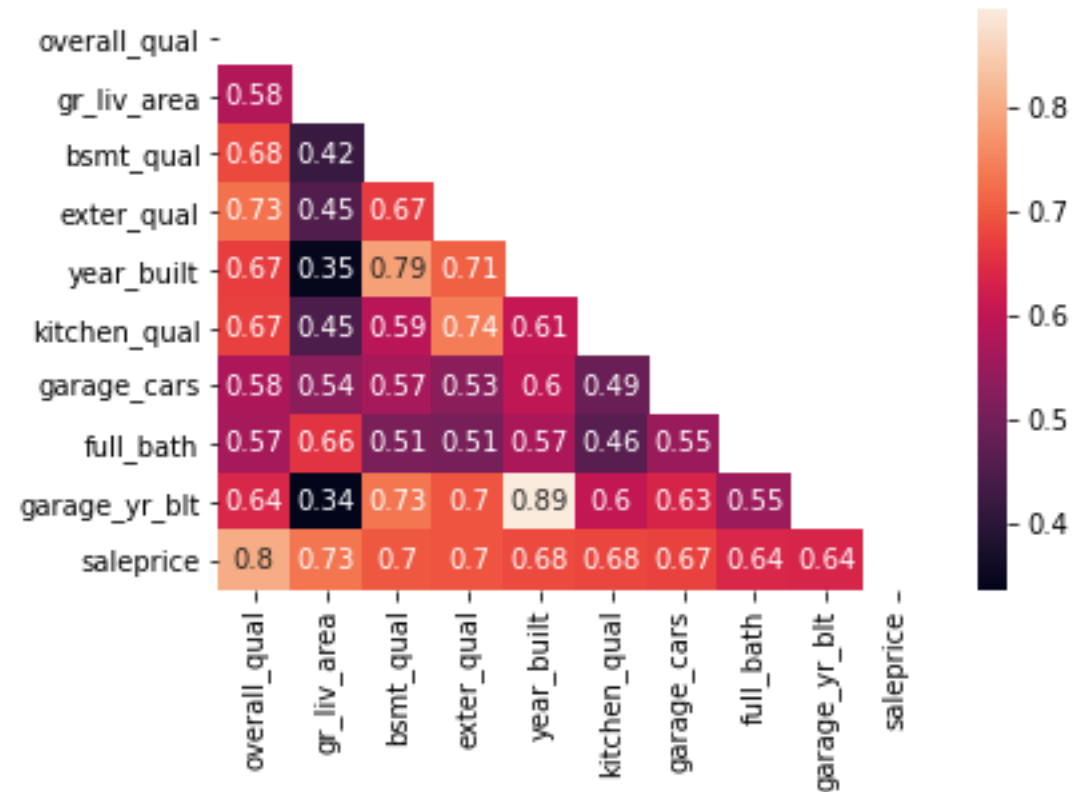
Data preprocessing

- Determine null values
 - Refer to data documentary
 - Null values really 'NA'?
 - Referred from other related variables?
 - Drop missing values?
- Remove outliers?



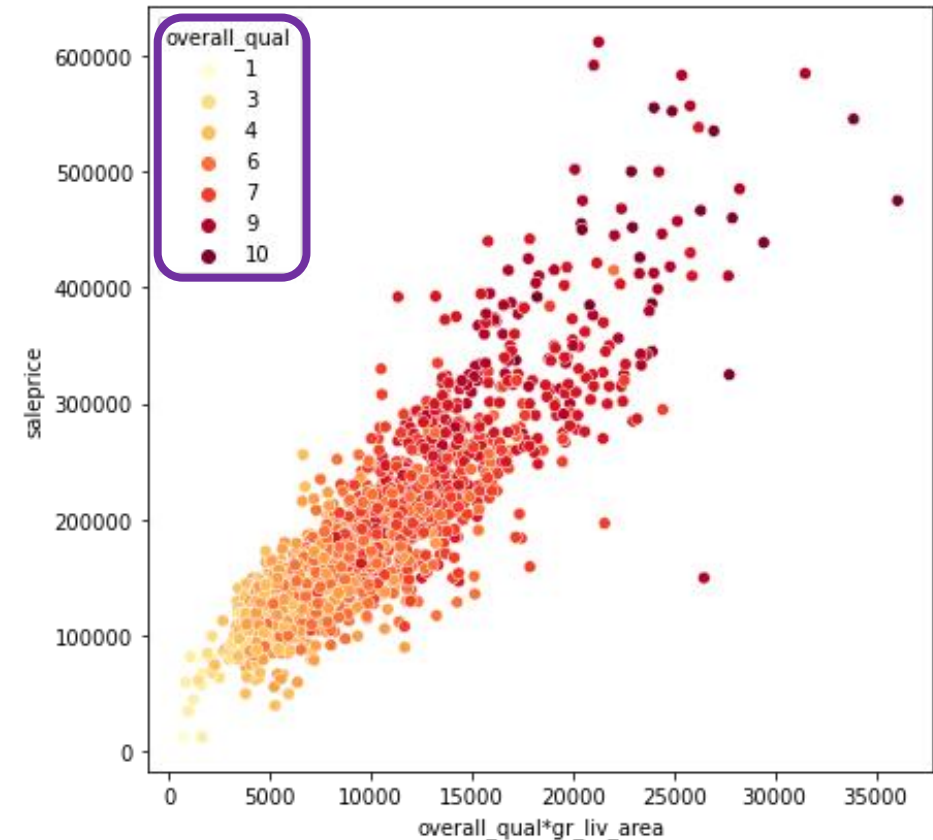
Feature selection – numerical features

- Dependent on data type
- Features were selected based on correlations
- Factors found:
 - **Overall quality** (includes condition of basement, kitchen, exteriors, etc)
 - **Size of living area** (total area, etc)
 - **Age of property** (year that it was built, or remodelled)



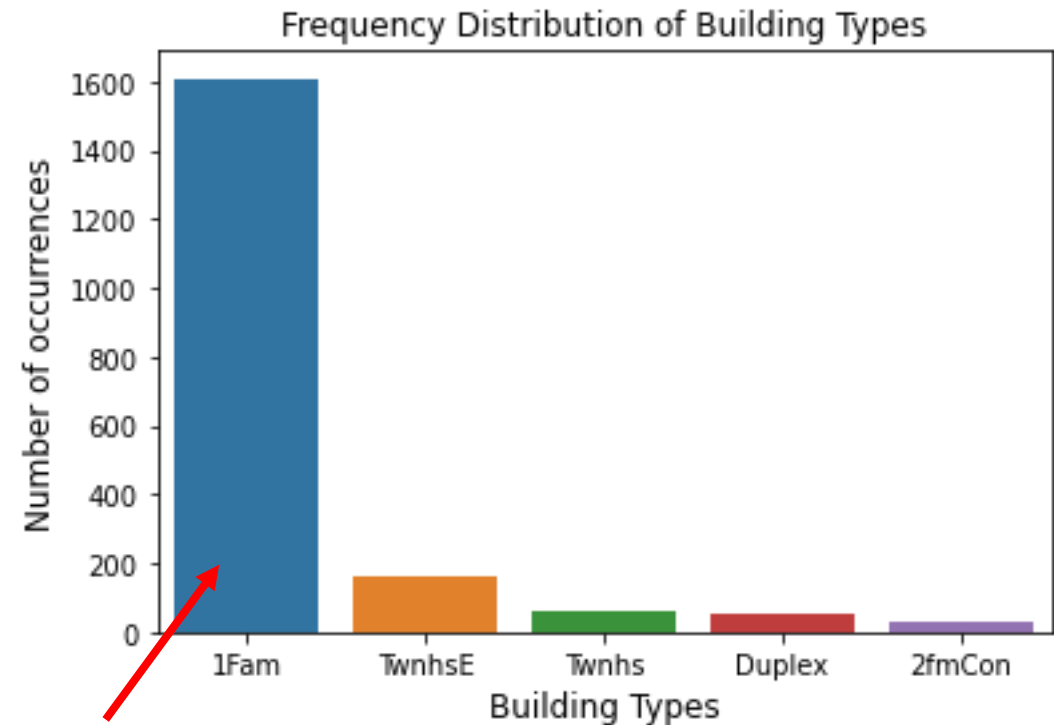
Feature selection – numerical features

- To increase predictive power, some **interaction terms** were created
- Features were selected based on correlations of interaction terms and sale price



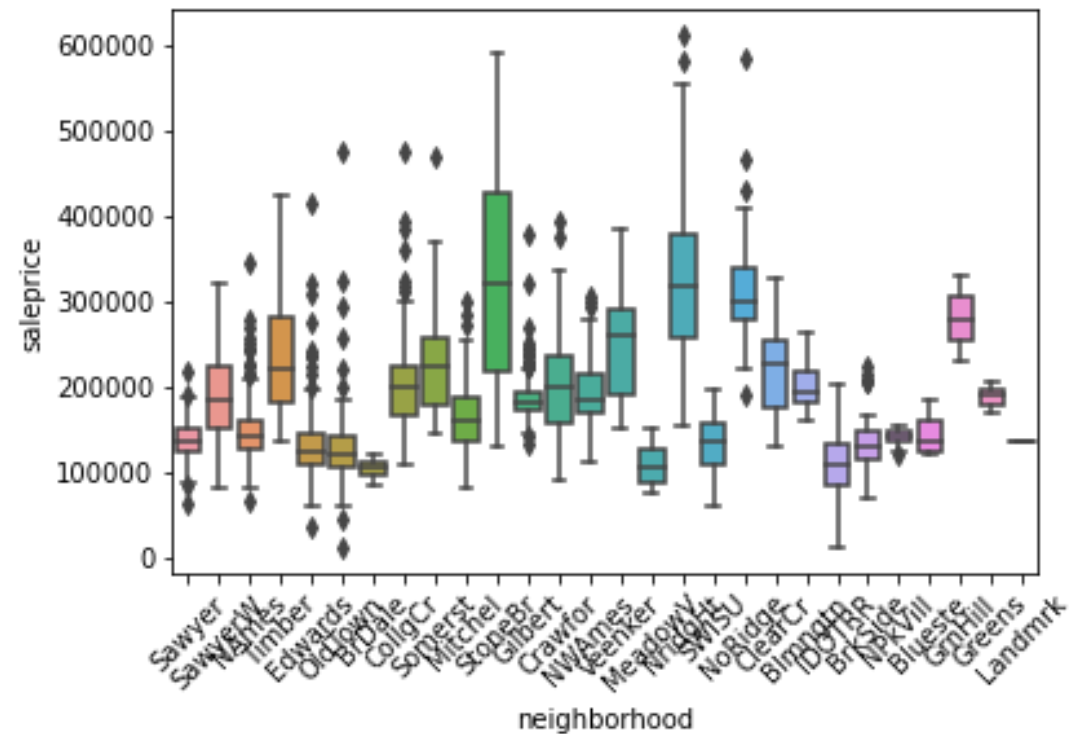
Feature selection – nominal features

- Features were selected based on variability of categories in each feature
- Feature dropped if there is one dominant category



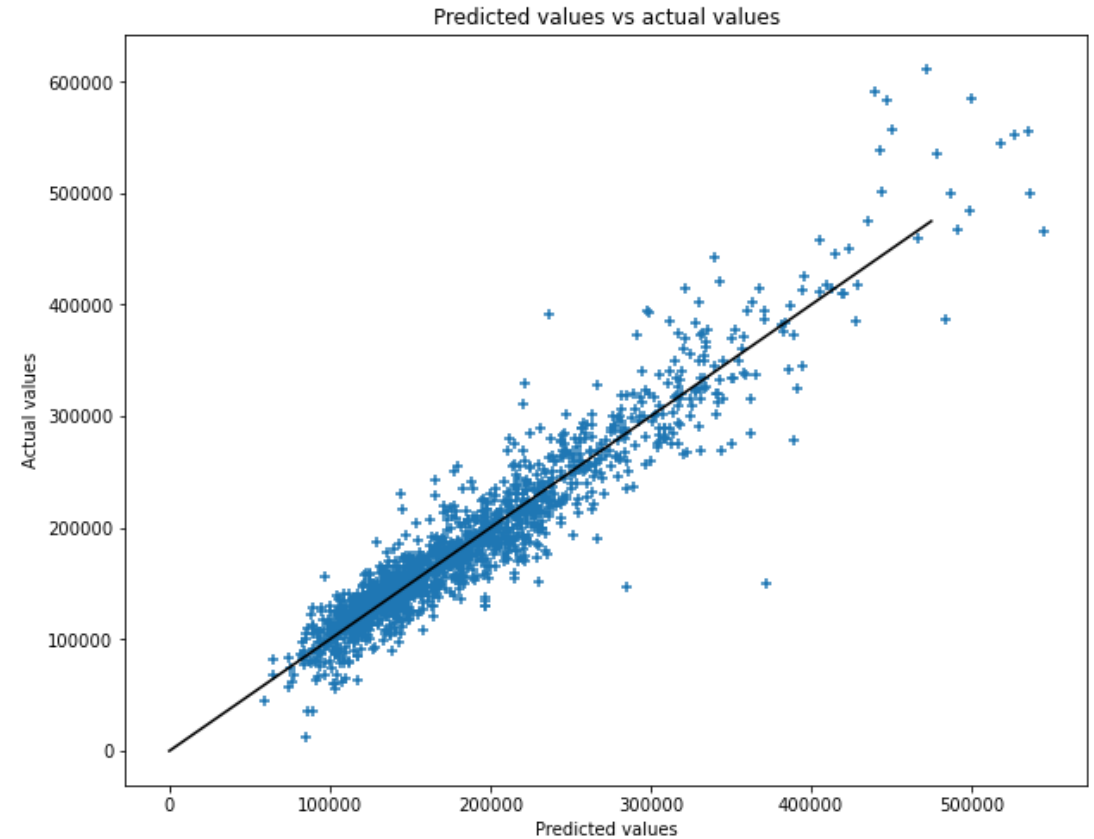
Feature selection – challenges

- Features were selected based on variability of categories in each feature
- Yet, some features with many categories have only a handful of observations each category



Prediction models

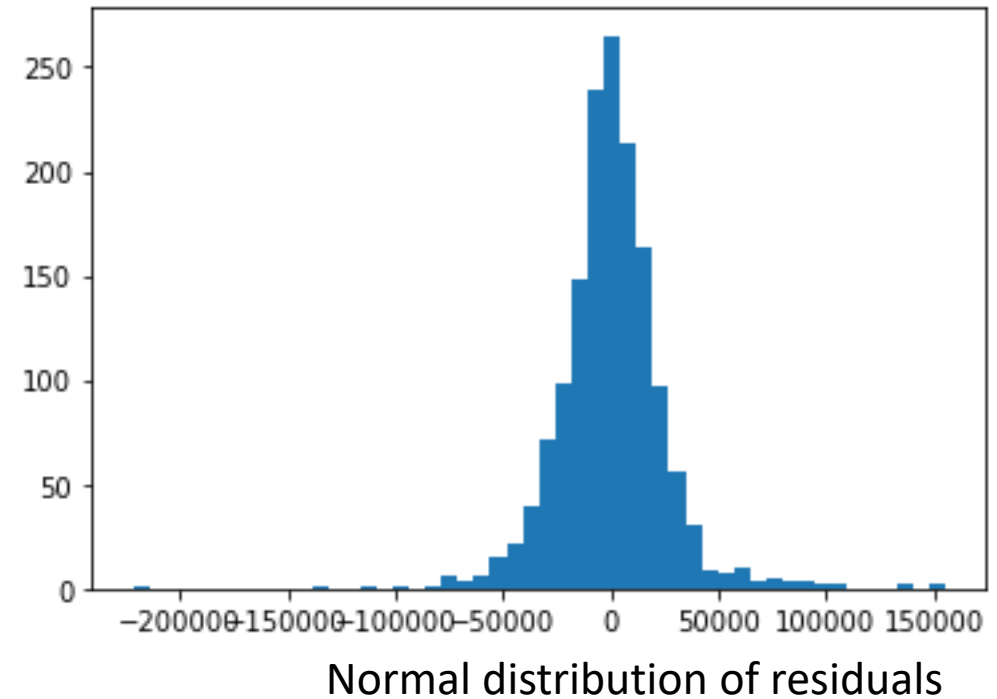
- OLS Linear regression to remove/ add features
- Ridge and Lasso regression on final selected features



Linear regression model

Prediction models - challenges

- Replaced null values may be incorrect
- Many features are related to each other
- Assumptions in a linear regression model
 - Linear relationship between features and sale price
 - Features are independent
 - Normality of errors



Key findings

	variable	coef
13	overall_qual*gr_liv_area	64128.563343
14	exter_qual*1st_flr_sf	56336.405866
15	total_bsmt_sf*kitchen_qual	35058.581132
18	lot_area	8059.870923
19	bsmtfin_type_1	5356.105234
8	total_bath	4383.464034
11	garage_cars	3438.049425
10	fireplace_qu	3428.652251
3	bsmt_qual	3292.464007
5	heating_qc	2785.455078
17	mas_vnr_area	2536.604750
16	open_porch_sf	640.658976
12	overall_qual*overall_qual	-2065.567268
1	age	-5229.924334
9	kitchen_qual	-6869.163424

- Overall quality and size have the highest positive effect on sale price.
- Age has a negative effect on sale price

Conclusion and Recommendations

- Most important features are:
 - Overall quality
 - Area (living area, lot area, masonry veneer area, etc)
 - Age
- Remodeling works!
 - Extending living areas
 - Upgrading kitchen, bathroom, garage, etc
- Forget about the porch, swimming pool

