AMES HOUSING ANALYSIS

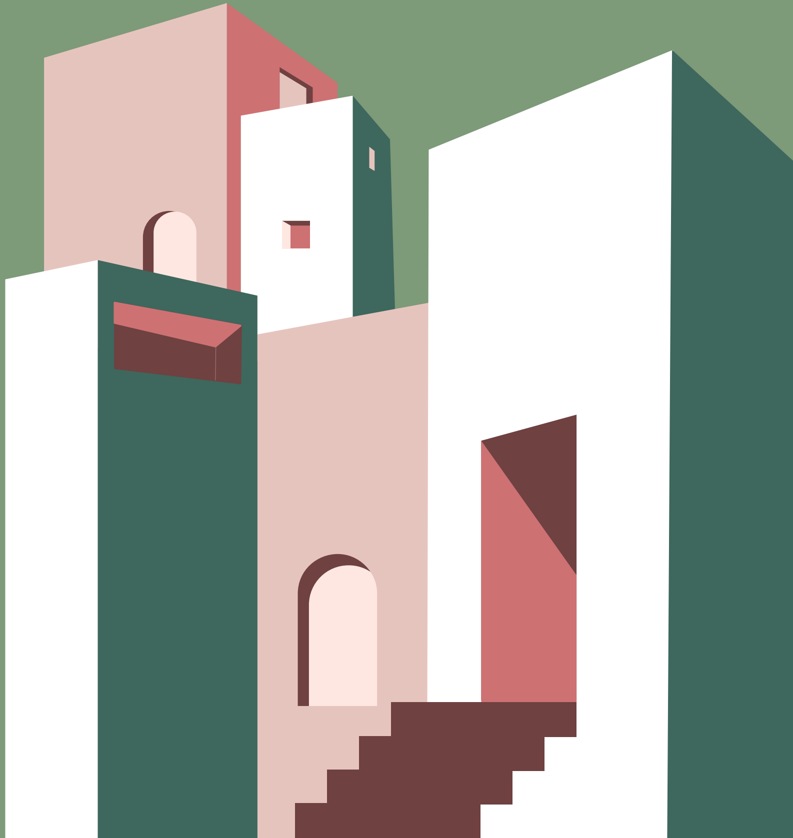
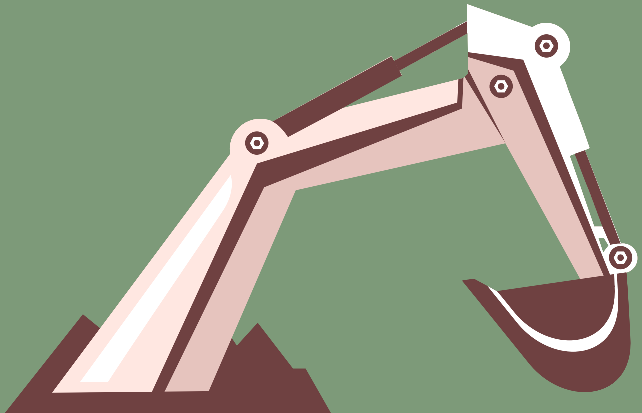


TABLE OF CONTENTS

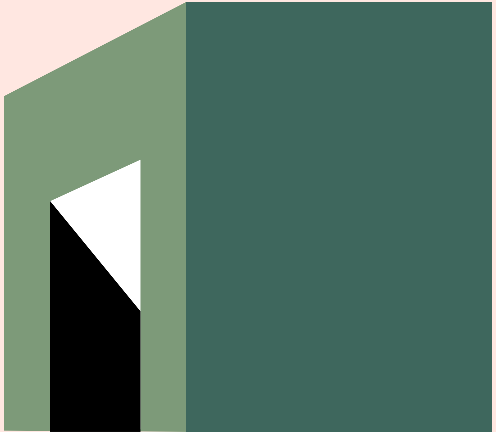
PROBLEM

Iowa's housing market is seeing a slow down. Appropriately priced houses get sold faster, but the process of getting a property valuation is long and tedious



GOAL

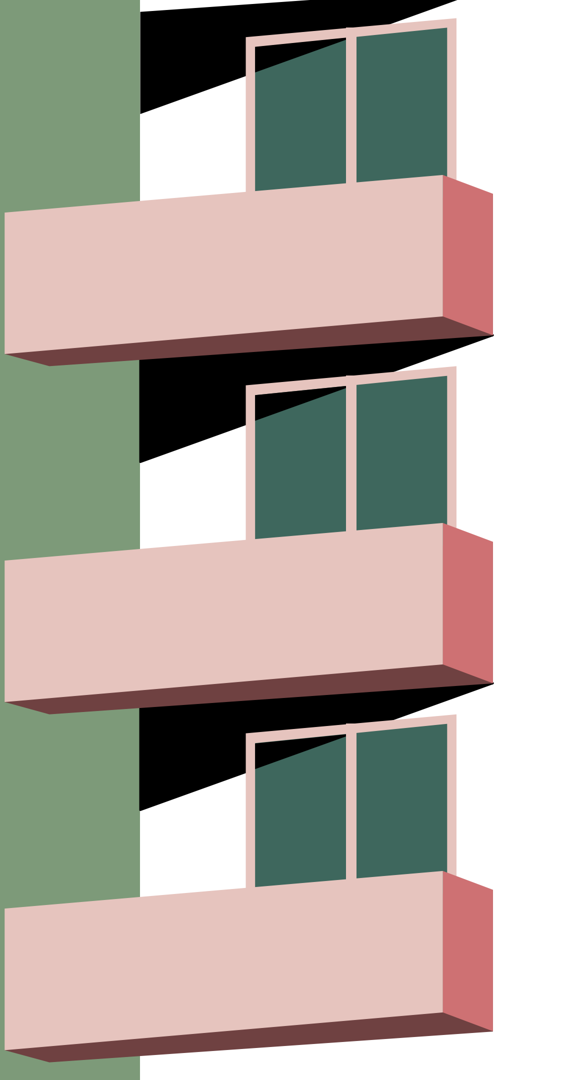
To automate pricing of houses through machine learning to expedite valuation and sale of houses for those looking to sell their property



02

WORKFLOW

Data cleaning, feature engineering and selection



PROCESS BEFORE MODELLING

CLEANING

Null values
i) Absence of features
ii) Impute based on most common
Check data types

EDA

Sale price - right skewed
Heatmap - correlation with target
Scatterplots of price vs living area
to select features

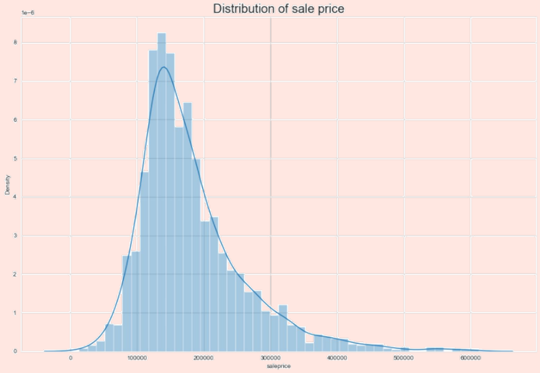
FEATURE ENGINEERING

Feature creation (unification)
Convert ordinal, nominal (OHE)
Scaling, polynomial features
Feature selection based on coefficients

MODEL PREPARATION

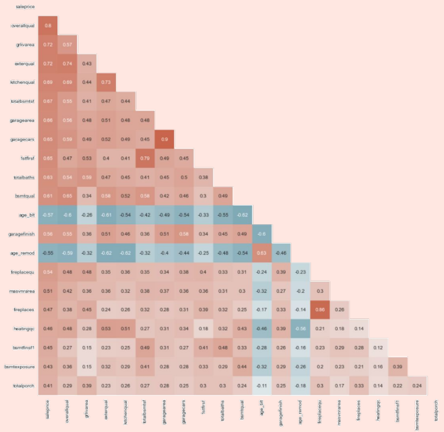
Train, test split (20% test)
Find optimal alphas

EDA OBSERVATIONS



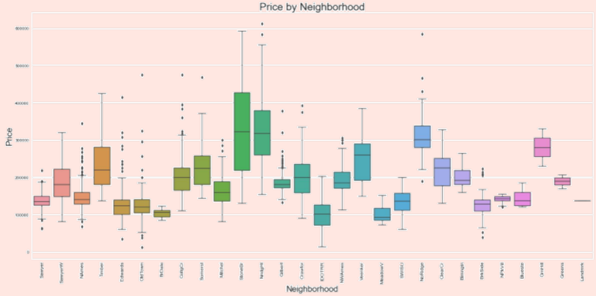
SALE PRICE

Right skewed, not normally distributed



TOP 20 HEAT MAP

Overall quality, living area, external quality, kitchen quality are amongst the top



BOX PLOTS

Various neighborhoods showed significant differences in mean prices

03

MODELLING

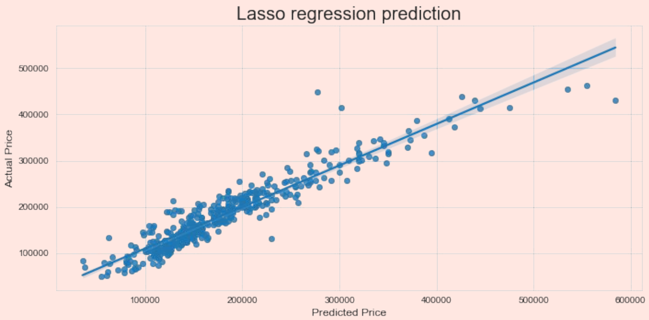
Summary and final model chosen

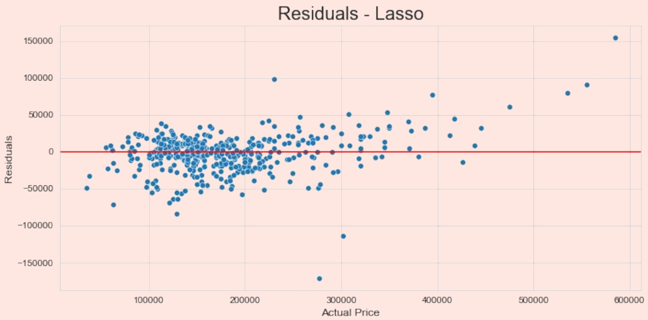
SUMMARY

	BASELINE	LINEAR (OVERFIT)	LASSO (W POLYFIT)	LASSO (DROP COLLINEAR)
# Features	92	92	30	26
Training score	0 %	90.5 %	89.7%	89.4%
Testing Score	0 %	89.8 %	89.1%	88.4%
RMSE	80,473	25,746	26,562	27,000

Add 5 interaction terms
Pick top 30 cf

MODEL EVALUATION





ACTUAL VS PRED

Line of best fit goes through most points, except at higher sale prices (>300K) but this is only 8% of data

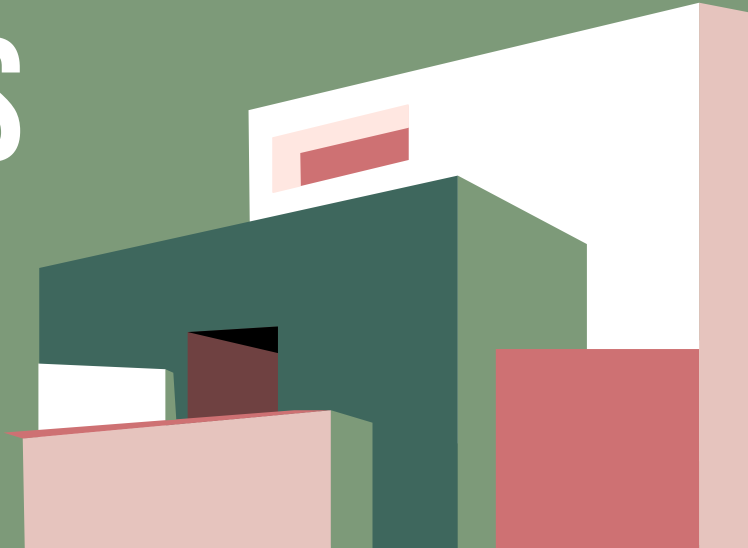
RESIDUALS

The inequality in variance is especially prominent at the extreme ends of prices

04

RECOMMENDATIONS

Limitations and improvements



RECOMMENDATIONS

FOR VALUATION

- Model is able to explain 88.7% of variation in price
- For better negotiation power, add 27K USD on top of predicted valuation
- If property is high value (>300K), it is best to get a realtor to do a physical valuation instead

FOR INVESTMENT

BONUS! Here are some things to consider for best returns:

- Properties in Northridge Heights
- Stone masonry veneer type houses
- Ensure garages are not attached
- Houses with a porch
- Newer houses
- Foundation made from poured concrete
- Central air conditioning

LIMITATIONS

- Many null values imputed based on highest value counts, might not be representative of true data
- Model would be more accurate with more features, but this would make it less interpretable to home owners

FUTURE IMPROVEMENTS

- Consider a separate model dropping the prices above 300K, and examine the features important to this group of properties
- Take into account inflation as these prices were collected over a few years