

# Predicting House Sale Prices for Ames, Iowa

---

Rashidi

SG-DSI-27

# Content

---

- Background
- Feature Selection
  - Data Cleaning
  - Exploratory Data Analysis
- Simple Model
- Model Tuning/Feature Engineering
- Model Benchmarks
- Production Model
- Insights
- Kaggle Submission Score



# Background

---

## **Problem Statement**

As a new Data Scientist in ABC-XYZ Corp., a real estate agency, I was tasked to create a website that can estimate a property sale price for the whole of USA, starting with Ames, Iowa (where our HQ is based).

# Feature Selection

Data Columns = 81

Which to choose? How to clean?

year\_built  
heating central\_air overall\_cond heating\_qc  
bsmtfin\_sf\_2  
condition\_2 paved\_drive bsmt\_qual roof\_matl enclosed\_porch  
yr\_sold garage\_type saleprice totrms\_abvgrd mas\_vnr\_area pid  
condition\_1 mo\_sold gr\_liv\_area half\_bath fireplace\_qu fireplaces garage\_qual open\_porch\_sf  
bsmt\_full\_bath bsmt\_exposure pool\_qc sale\_type garage\_finish lot\_shape misc\_feature bsmtfin\_type\_1 bsmt\_unf\_sf  
garage\_cond kitchen\_qual pool\_area year\_remod roof\_style exterior\_2nd low\_qual\_fin\_sf garage\_cars utilities mas\_vnr\_type lot\_frontage bldg\_type foundation fence lot\_area extender\_cond bsmtfin\_sf\_1 neighborhood bsmt\_half\_bath house\_style land\_slope street wood\_deck\_sf garage\_yr\_blt lot\_config land\_contour exterior\_1st kitchen\_abvgr bedroom\_abvgr ms\_zoning full\_bath ms\_subclass alley bsmt\_cond ms\_zoning full\_bath ms\_subclass bsmtfin\_sf\_1 neighborhood bsmt\_half\_bath house\_style land\_slope street wood\_deck\_sf garage\_yr\_blt lot\_config land\_contour exterior\_1st kitchen\_abvgr bedroom\_abvgr

# Initial Feature Filtering

---

- Filter out those that describes the same thing as another
  - 'garage\_cars' < 'garage\_area'
- Filter out those that is a subset of another (\*Except ordinal features)

$$\begin{array}{l} \text{'bsmtfin_sf_2'} \\ \text{'bsmtfin_sf_1'} \\ \text{'bsmt_unf_sf'} \end{array} \subseteq \text{'total_bsmt_sf'}$$

- Filter out those that are identification features



# Data Cleaning

---

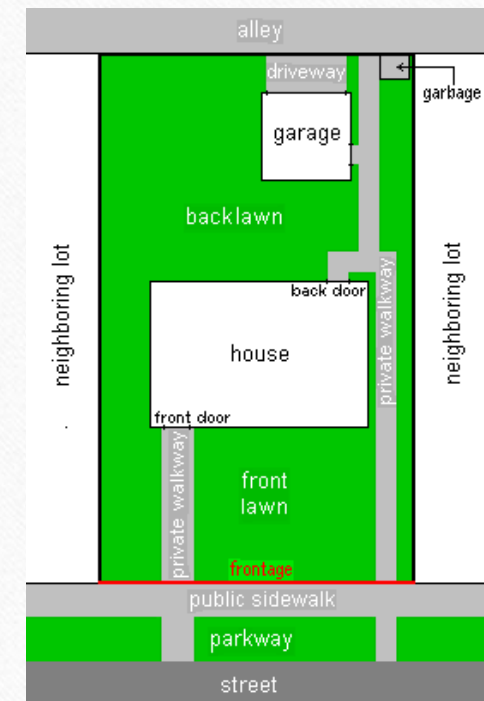
- Numerical features -> Check for null values
- Ordinal features -> Change to numerical scale
- Categorical features -> Manually Dummify
  - Done in order to determine easily which dummy column was dropped
    - Example: '150' was dropped from 'ms\_subclass'.

# Data Cleaning

'lot\_frontage'

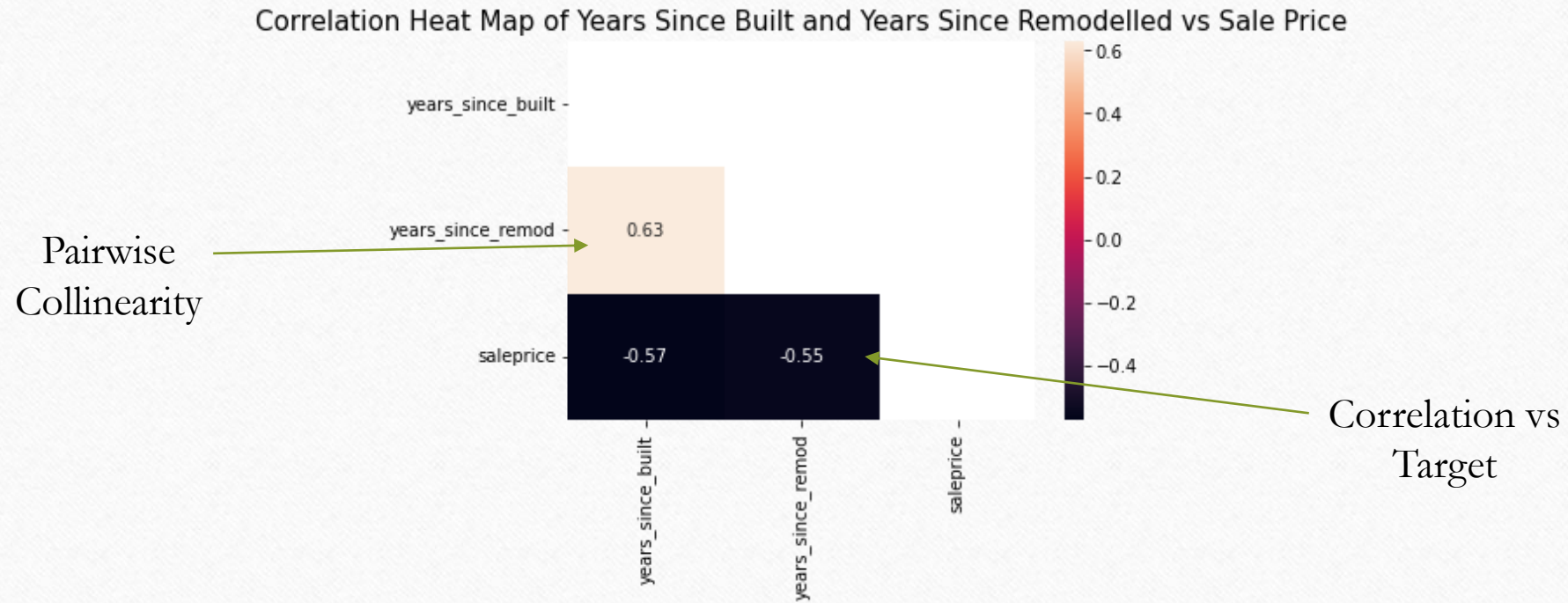
- Check any commonalities for 'ms\_subclass' or 'lot\_config' vs. 'lot\_frontage'
- Since there aren't any, set NaN values as the mean value of each 'ms\_subclass'

'ms_subclass'	mean('lot_frontage')
20	77.03
30	61.04
40	51.75
45	54.82
50	63.00
60	78.27
70	64.32
75	70.47
80	79.87
85	73.33
90	69.40
120	44.82
150	44.82
160	27.59
180	26.60
190	71.60



(from: [https://upload.wikimedia.org/wikipedia/commons/b/bc/Lot\\_map.PNG](https://upload.wikimedia.org/wikipedia/commons/b/bc/Lot_map.PNG))

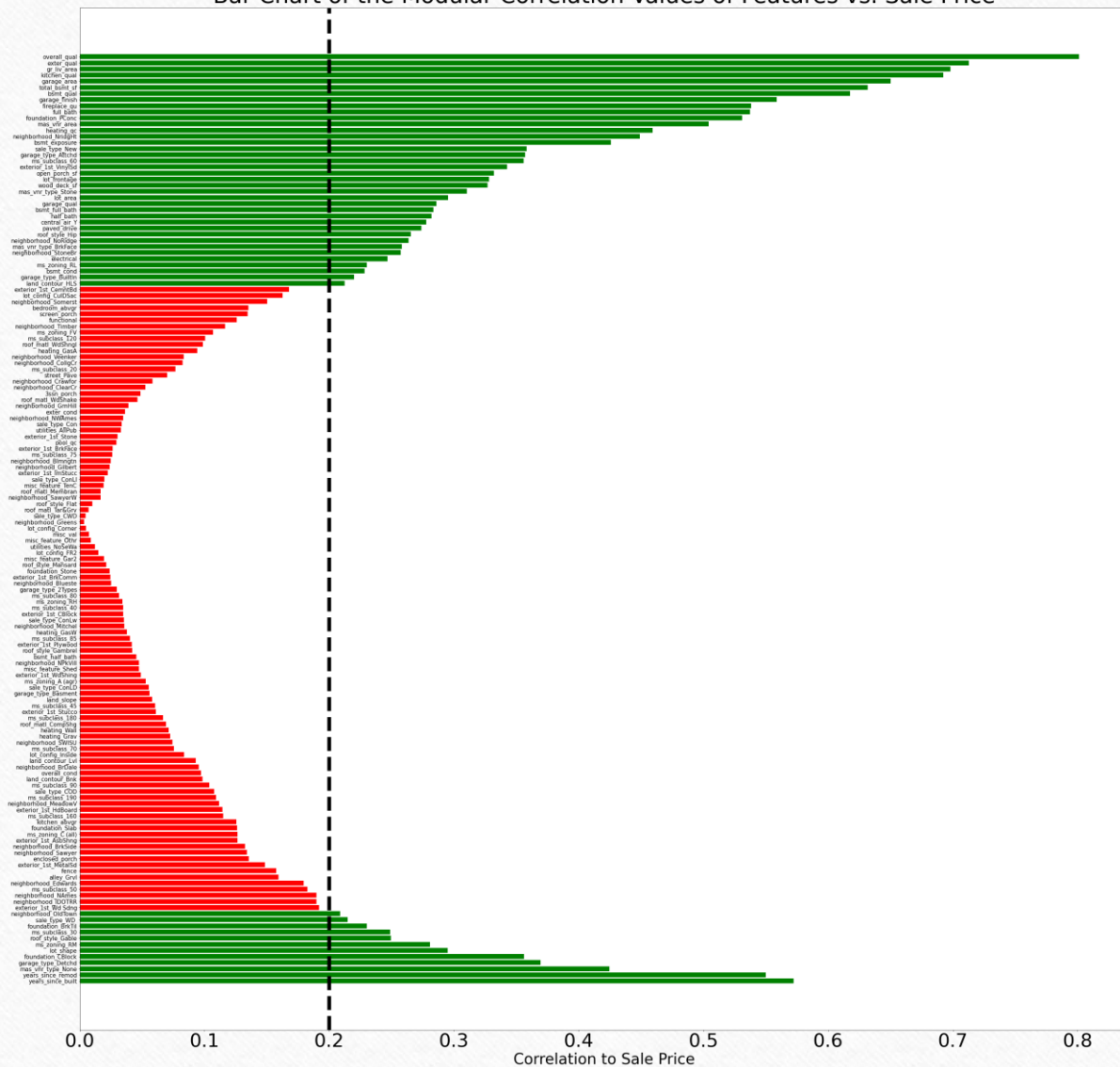
# Exploratory Data Analysis





Pairs	Pairwise Correlation	First Feature Correlation Vs. Sale Price	Second Feature Correlation Vs. Sale Price
✓ ms_subclass_90 vs. bldg_type_Duplex	1.000000	-0.103817	-0.103817
✓ ms_subclass_80 vs. house_style_SLvl	0.954549	-0.031484	-0.042176
✓ garage_qual vs. garage_cond	0.950118	0.285858	0.265517
✓ ms_subclass_50 vs. house_style_1.5Fin	0.942502	-0.182567	-0.196051
✓ pool_area vs. pool_qc	0.904689	0.023115	0.029289
✓ ms_zoning_FV vs. neighborhood_Somerst	0.874843	0.106749	0.150167
✓ ms_subclass_45 vs. house_style_1.5Unf	0.869662	-0.060391	-0.066877
✓ fireplaces vs. fireplace_qu	0.859621	0.470091	0.538252
✓ gr_liv_area vs. totrms_abvgrd	0.812723	0.698046	0.502909

Bar Chart of the Modular Correlation Values of Features vs. Sale Price



Where do  
we draw the  
line?

# Simple Model

---

99.Co



# Simple Model

99.co

99.co Buy Rent Sell | New Launch Explore Mortgage Commercial

Home > Property Value Tool > Your property

HDB [redacted] Edit

6 [redacted] Singapore [redacted]

Property type: HDB [redacted]

Floor & Unit no.: [redacted] - [redacted]

Floor area:

(source: 99.co, <https://www.99.co/>)



Home > Tracked properties > HDB [redacted]

HDB-[redacted], [redacted] **\$S7xx,xxx**  
Est. Sale value

[redacted] • [redacted] 1,517.72 sqft

Neighborhood

MS Subclass

Gr Liv Area

Linear Regression

# Model Tuning/Feature Engineering

---

Steps Taken:

- 1) Lasso Regression ( $\alpha=874.0802078515503$ )
- 2) Linear Regression after dropping Lasso Zero Coefficient Features
- 3) Ridge Regression after dropping Lasso Zero Coefficient Features  
( $\alpha=335.1602650938841$ )

# Model Tuning/Feature Engineering

## Lasso Regression

---

**Pairwise Collinearity of lot\_frontage vs. gr\_liv\_area: 0.360696**

Feature	lasso_coef Value	Saleprice Correlation
lot_frontage	-0.000000	0.328149
gr_liv_area	24273.462134	0.69804

**Number of Features dropped: 97**



# Model Benchmarks

Model	Train MSE	Test MSE	Cross Val Score
drop_0_coeff Ridge Regression	702796769	619778333	950438300
drop_0_coeff Linear Regression	651192793	666638781	1022462959
Lasso Regression	701584300	649567599	1015959823
99co Linear Regression	1464088294	1152320467	1576989381



~40%  
improvement

# Production Model

---

## Production Model Attributes

<b>Train MSE</b>	673341543.1
<b>Cross Val MSE</b>	833273475.8
<b>Ridge Regression Alpha</b>	335.16
<b>Total Features Used</b>	73
<b>Kaggle Public Score</b>	33203.03021

# Insights

---



# Insights

## Production Model

---

- Overfit (Cross Validation MSE  $\gg$  Train MSE)

### **Future Works:**

- 1) Eliminating outliers from deep-diving into model predicted residuals.
- 2) Explore pairwise interactions.
- 3) Explore different cutoffs to see which will effectively eliminate poorly correlated features (vs. Target) and produce the best model.

# Insights

## Project

---

- Features used in model  $\neq$  features easily known by layman
  - It will be better to get data on which features are easily known/accessible by our platform users.
- Strive between simplicity (like 99.co) versus accuracy.
  - No one would want to sit down and complete a form with 70+ blanks to fill up.