AMES Housing IOWA, US

By Simran GA DSI 22

01.
PROBLEM STATEMENT

02.
CLEANING AND EDA

03. DATA PREPROCESSING

04.
REGRESSION MODELLING

05.
KAGGLE SUBMISSION

06.
RECOMMENDATIONS

problem STATEMENT

The problem that we are trying to solve is how to predict the sale prices of a property using linear regression models.

DATA CLEANING

HANDLING MISSING VALUES

Changes made:
pool_qc + pool_area = pool

Dropped: misc_feature, alley, fence

FEATURE	MISSING VALUES
pool_qc	2042
misc_feature	1986
alley	1911
fence	1651
fireplace_qu	1000
lot_frontage	330
garage_yr_blt	114
garage_cond	114
garage_qual	114
garage_finish	114
garage_type	113
bsmt_exposure	58
bsmtfin_type_2	56
bsmt_cond	55
bsmt_qual	55
bsmtfin_type_1	55
mas_vnr_type	22
mas_vnr_area	22
bsmt_half_bath	2
bsmt_full_bath	2
garage_cars	1
bsmtfin_sf_1	1
bsmtfin_sf_2	1
bsmt_unf_sf	1
garage_area	1
total_bsmt_sf	1

EDA

Classification: Discrete, Nominal, Ordinal, Continuous

Classification: Nominal, Ordinal, Continuous

Ordinal Data Reclassified to Numerical

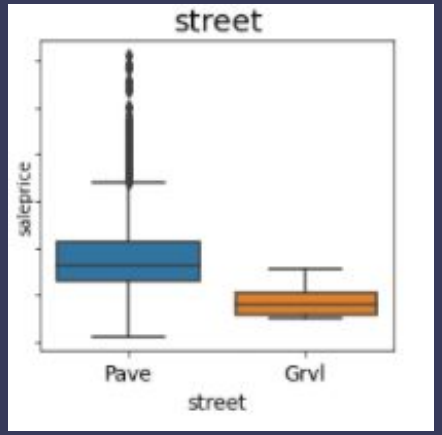
functional: ['Typ' 'Mod' 'Min2' 'Maj1' 'Min1' 'Sev' 'Sal' 'Maj2']

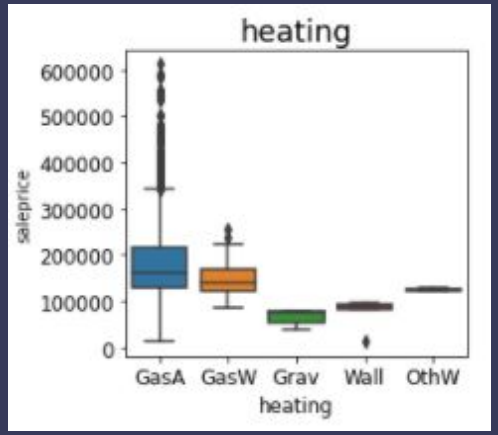
functional: [7 4 5 3 6 1 0 2]

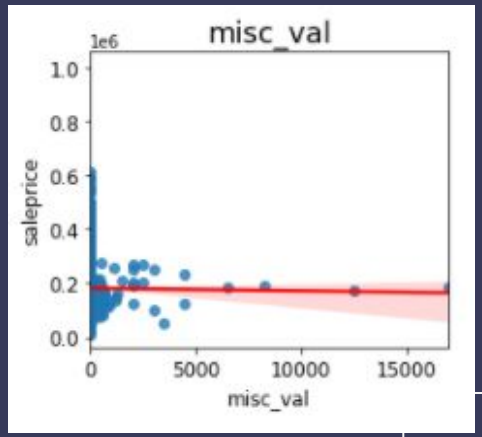
OBSERVATIONS

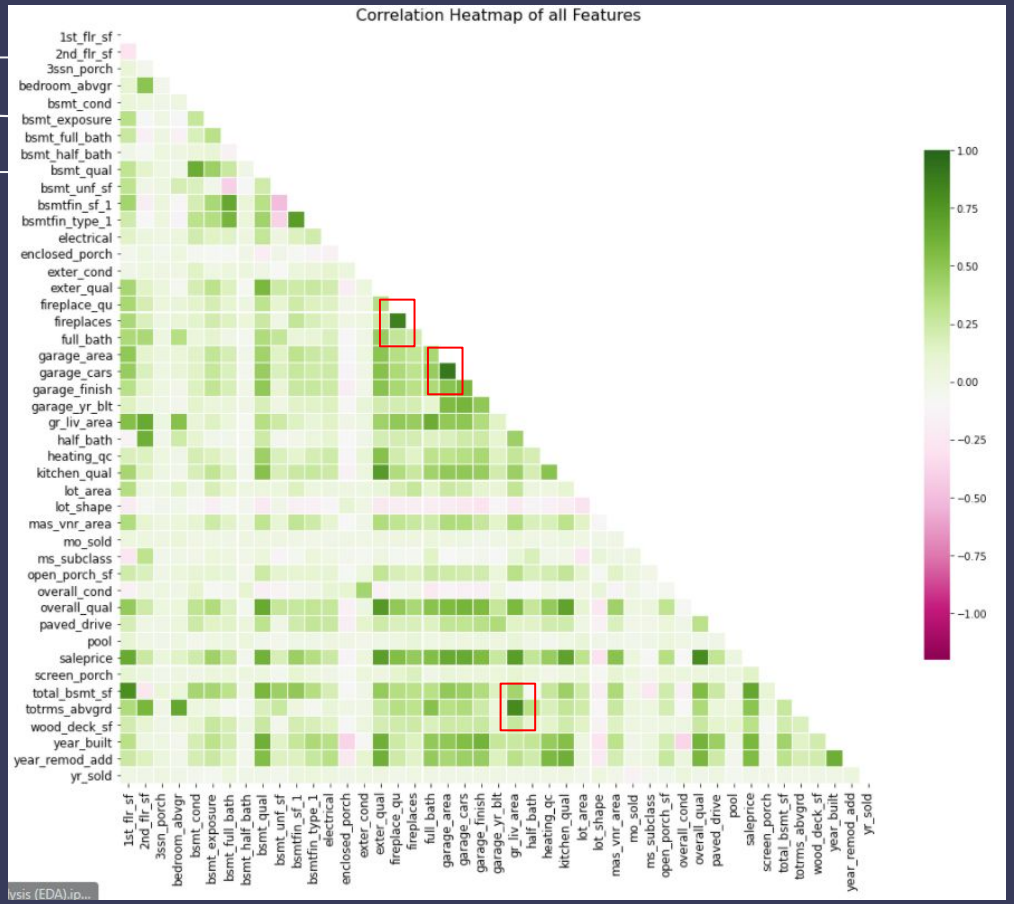
Couple of Features were highly skewed → dropped

Others showed weak to no correlation, too many 0 values → dropped









Multicollinearity Issues:

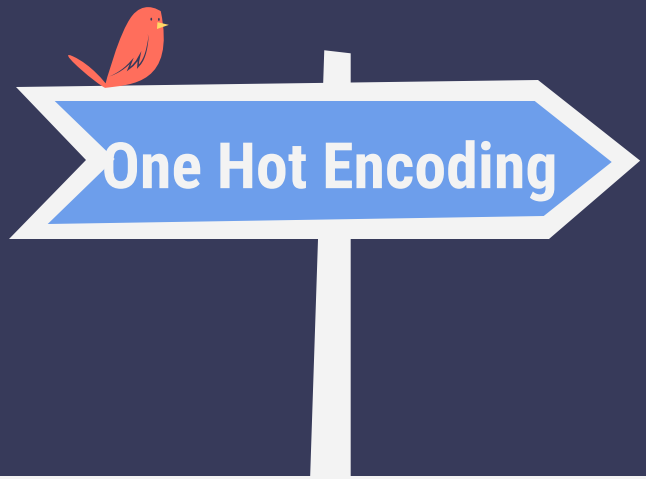
1. 'garage_area' as a strong correlation with 'garage_cars' - the bigger the area, the higher number of cars can be stored

2. 'fireplaces' as a strong correlation with 'fireplaces_qu' - the higher the number of fireplaces, the greater the overall quality of all fireplaces

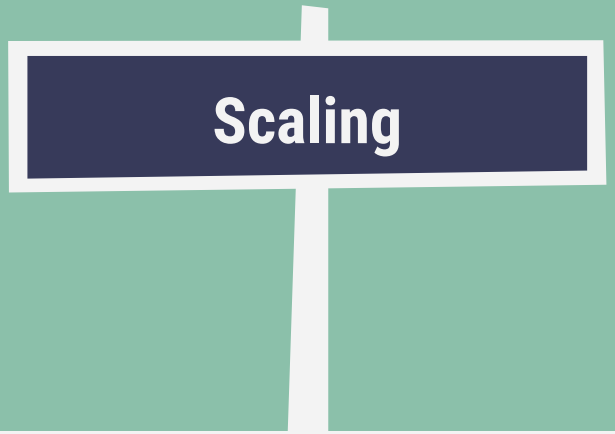
3. 'gr_liv_area' as a strong correlation with 'totrms_abvgrd' - the higher number of rooms above ground, the larger the area.

DATA PREPROCESSING

One Hot Encoding on Nominal Features



Scale Continuous and Ordinal Features



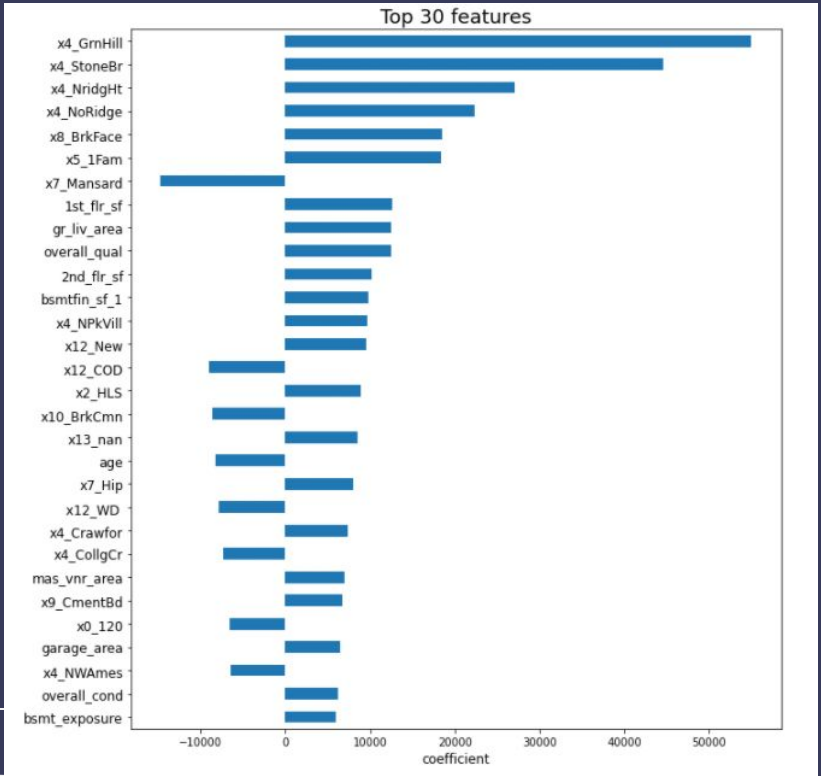
MODELING

COMPARISON BETWEEN MODELS

	Cross Val Score	RMSE
Baseline Model*	79532	78360
Linear Regression	5.23e+15	24896
Ridge Regression	25109	24075
Lasso Regression	25116	23690

*Baseline Model developed using Dummy Regressor

FEATURE SELECTION



Top 30 Features chosen based on highest non-zero coefficient values

Conclusion and Recommendations

	Feature	Coefficient
27	x4_StoneBr	46937
26	x4_NridgHt	38469.4
20	gr_liv_area	24133.2
23	x5_1Fam	21505
25	x4_NoRidge	20942.3
19	overall_qual	19357.5
24	x8_BrkFace	19239.6
15	x12_New	18064.9
8	x4_Crawfor	12843.3
17	bsmtfin_sf_1	11249.4

Top 10 Features

Real Estate Investors/Buyers

1. Property dealers/buyers should consider properties in Stone Brook, Northridge Heights and Northridge Neighbourhoods for investment purposes as they yield higher sale prices.

2. Also look into single-family detached properties/Hillside properties

Home Owners:

3. Consider renovating to Brick Face exterior covering

4. Sell properties earlier as older properties yield lower sales

5. Avoid Masonry veneer types such as Brick Common.