PREDICTORS OF SALE PRICES OF HOMES

AMES, IOWA

PROBLEM STATEMENT

Target: To make accurate predictions of housing prices in Ames, Iowa, given over 2,000 data points that include over 80 features.

Parties involved: Home owners, Real Estate Agents, New Ames Residents

METHODS USED

DATA CLEANING & EDA

FEATURE ENGINEERING & MODELLING

EVALUATION

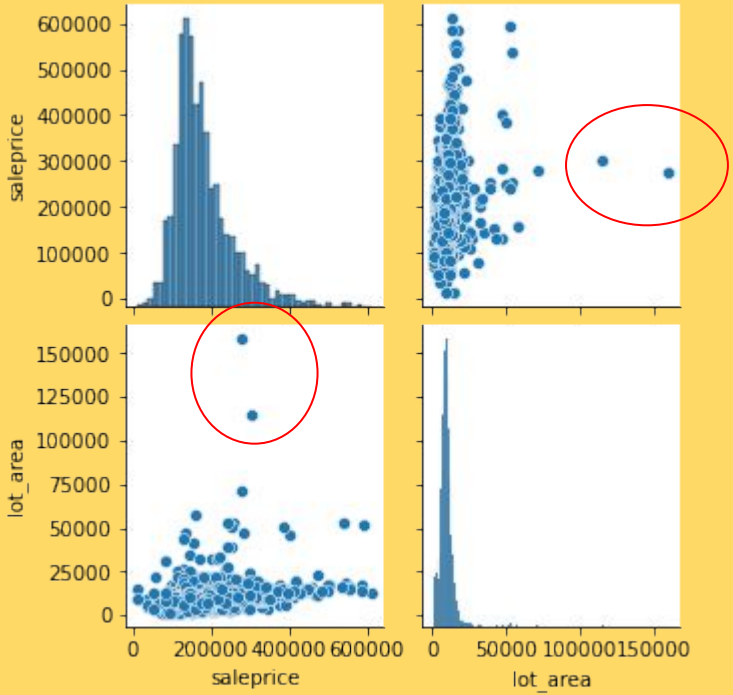
DATA CLEANING

- MISSING VALUES:

  Columns that contained more than 85% null values were dropped due to the lack of information available

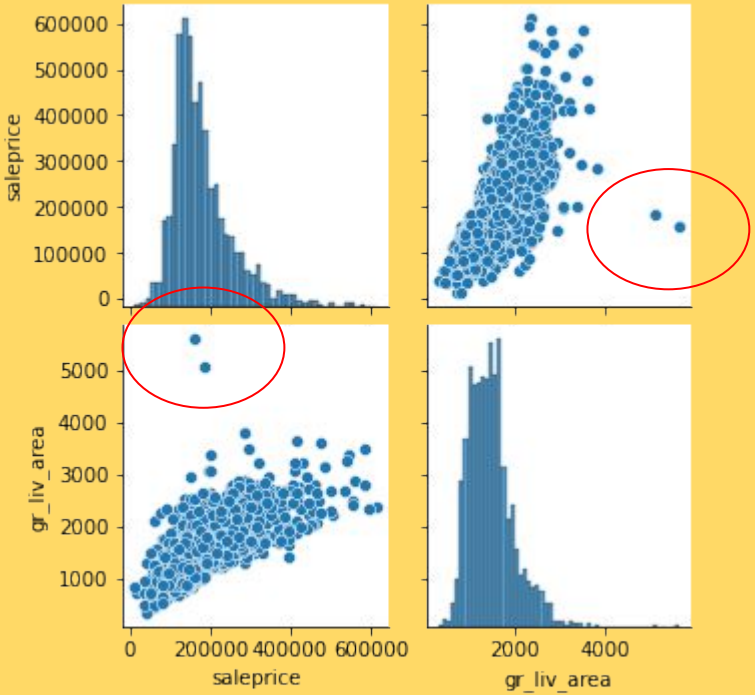
	sum_of_nulls	missing_pct
pool_qc	2042	99.561190
pool_area	2042	99.561190
3ssn_porch	2025	98.732326
low_qual_fin_sf	2018	98.391029
misc_val	1986	96.830814
misc_feature	1986	96.830814
bsmt_half_bath	1925	93.856655
alley	1911	93.174061

OUTLIERS



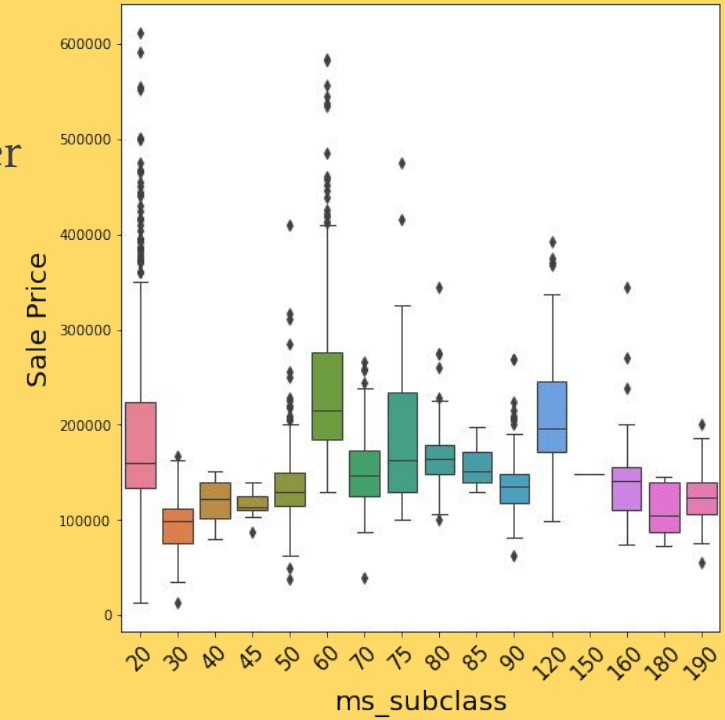
- Outliers identified where lot area > 75,000 sqft

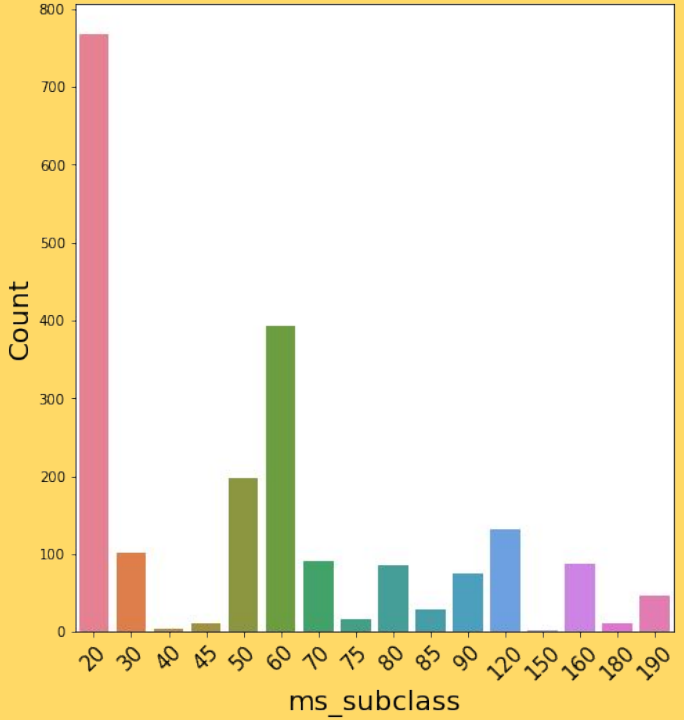
- Living area > 4,000 sqft



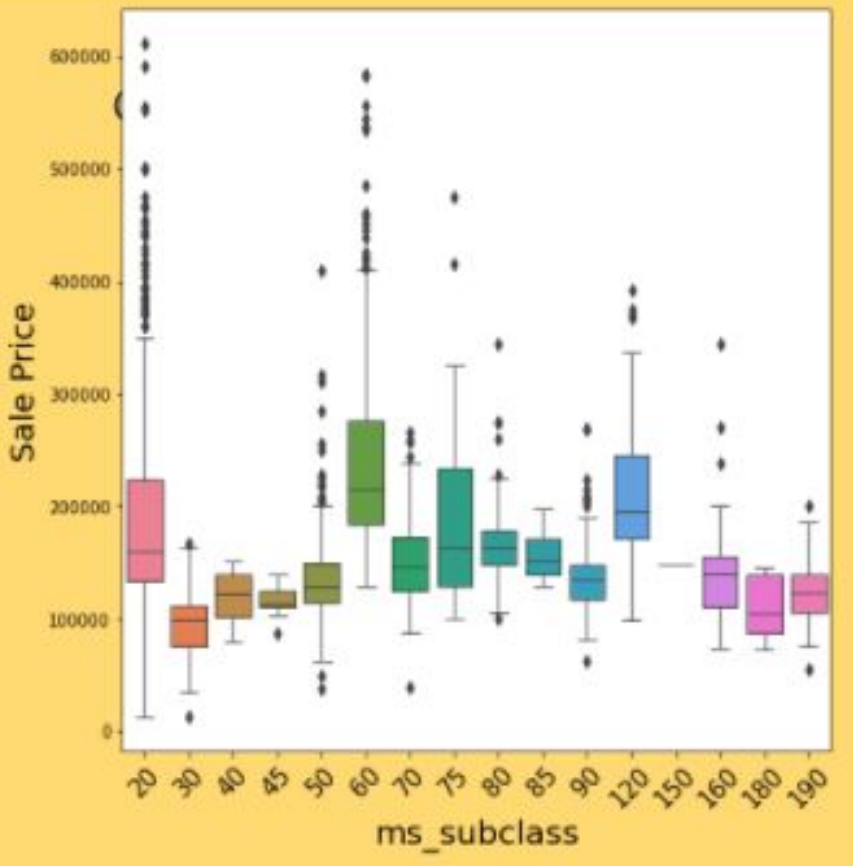
FEATURE ENGINEERING

i.e. individual components for ms_subclass were regrouped into smaller bin sizes based on similar interquartile ranges





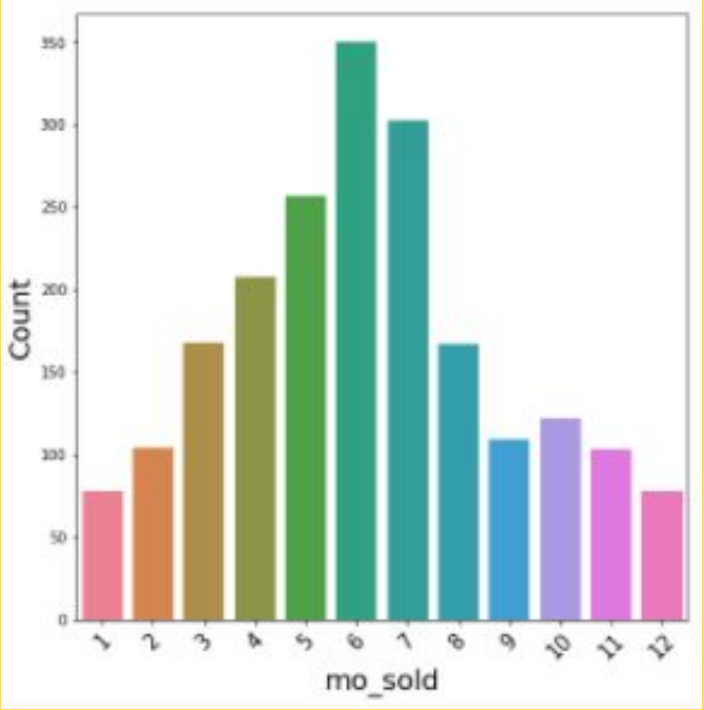
REGROUPING OF FEATURES



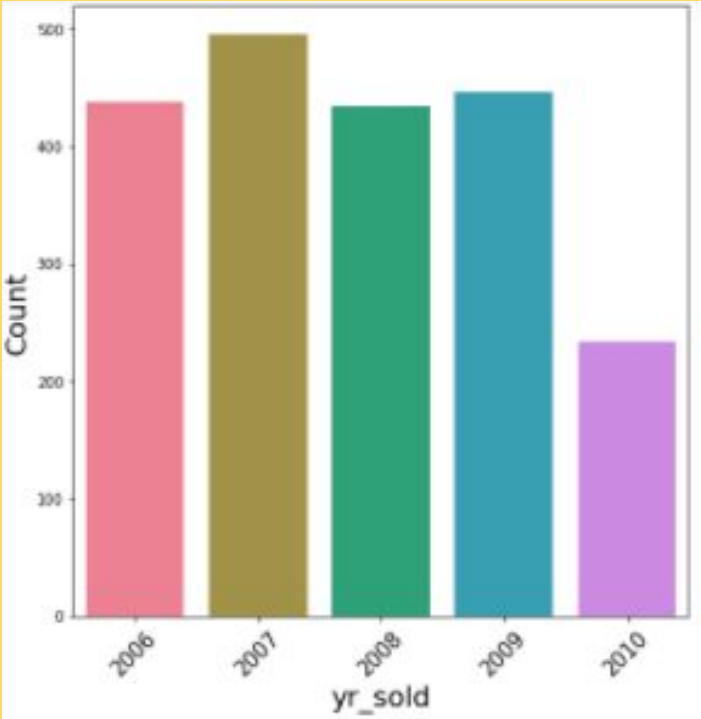
BIN No.	Subclasses
bin1	20 & 75
bin2	30
bin3	60
bin4	70
bin5	80 & 85
bin6	50 & 90
bin7	120
bin8	190 & 40
bin9	160 & 180
bin10	All other subclasses (45 & 150)

SALES VS TIME PERIOD

- High volume of sales during the months leading into summer (April, to Aug)



HIGHEST VOLUME OF SALES DURING SUMMER MONTHS



Volume of sales were consistent throughout the years of 2006 ~ 2009.

Spillover effects from the housing market crash only started to happen in Iowa in 2010.

MODELLING

- NOMINAL FEATURES >>> One Hot Encoding

```
encode_cols = OneHotEncoder(handle_unknown = 'ignore')

X_pre = df[columns_OHE]
encode_cols.fit(X_pre)
X = encode_cols.transform(X_pre).toarray()

# one_hot = pd.DataFrame(X)
one_hot = pd.DataFrame(X,columns =
                       encode_cols.get_feature_names(columns_OHE))

df = pd.concat([df,one_hot],axis=1)
```

- ORDINAL FEATURES >>> Manual Encoding

```
train.replace(to_replace = {
    'lot_shape':{'Reg': 4,'IR1': 3,'IR2': 2,'IR3': 1},
```

METRIC USED FOR SCORING: RMSE

- Linear Regression Model: CV RMSE - 24,574
- Mean predicted sale price: $182,680

	value
neighborhood_GrnHill	106885.550053
neighborhood_StoneBr	38152.917937
house_style_1.5Unf	24742.615760
ms_subclass_bin_10	23890.818350
neighborhood_NridgHt	20879.510658
sale_type_COD	17882.092969
kitchen_abvgr	17519.029536
neighborhood_CollgCr	17176.477307
neighborhood_Veenker	16857.089961
utilities	16773.271750

METRIC USED FOR SCORING: RMSE

- Ridge performed slightly worse than Linear at 24,600.
- Lasso was the best performing on our training set.

```
Ridge CV RMSE: 24600.65450017758
Ridge Model RMSE: 24560.480642002796

Lasso CV RMSE: 23975.16863220038
Lasso Model RMSE: 23935.659844768266
```

Thank you!