AMES, IOWA HOUSING PROJECT

PROBLEM STATEMENT

Given an Ames Housing Dataset design a Regression Model to accurately predict the sale price of houses in Ames, Iowa

DATA

Ames Housing Dataset containing 78 features of buyer are interested in knowing about before purchasing a home/house, which was used to train and test my Models

Test Dataset which was used to evaluate my Model's performance

ACTION

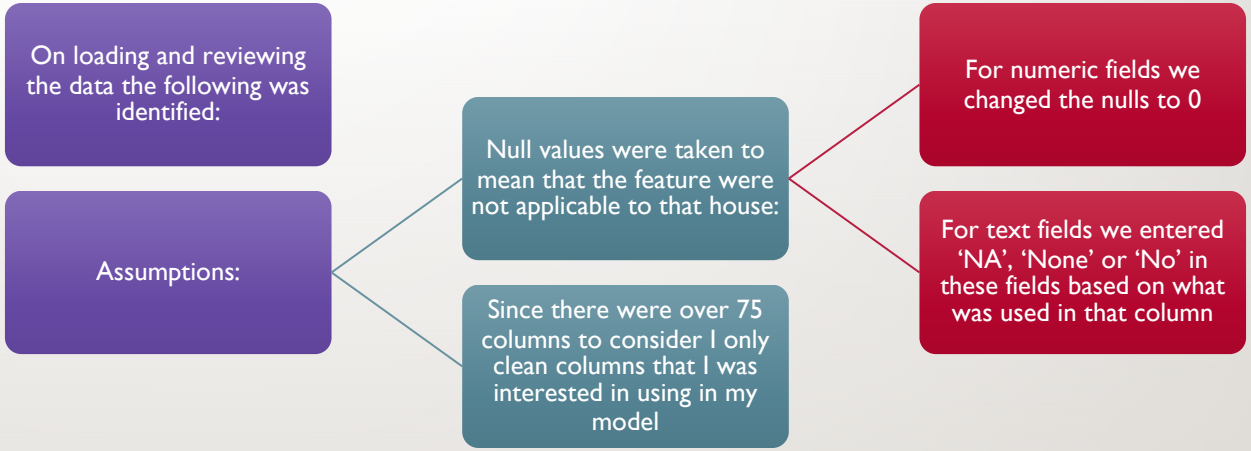
I loaded the data and explored it there alongside the extended data dictionary give. The following approach was then taken using a test size of 10%:

Design a Model based on the features I will look for in a house

Fill all the null values with 0 (assumption here is that these where all 0 because the where applicable). Then design a Model based on the features with abs|corr()| > .4

Submit both results and use the one with the better Kaggle results as my starting point

DATASET & THE CLEANING PROCESS

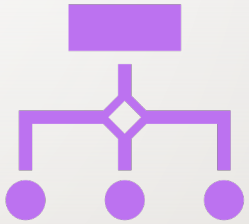


DATA CLEANING FUNCTION

```
def clean_data(df):
    df['lot_frontage'].fillna(0, inplace=True)
    df['garage_area'].fillna(0, inplace=True)
    df['garage_cars'].fillna(0, inplace=True)
    df['total_bsmt_sf'].fillna(0, inplace=True)
    df['garage_yr_blt'].fillna(0, inplace=True)
    df['mas_vnr_area'].fillna(0, inplace=True)
    df['bsmtfin_sf_1'].fillna(0, inplace=True)
    df['bsmtfin_sf_1'].fillna('None', inplace=True)
    df['bsmt_qual'].fillna('NA', inplace=True)
    df['mas_vnr_type'].fillna('NA', inplace=True)
    df['bsmt_cond'].fillna('NA', inplace=True)
    df['bsmt_exposure'].fillna('No', inplace=True)
    df['bsmtfin_type_1'].fillna('NA', inplace=True)
    df['bsmtfin_type_2'].fillna('NA', inplace=True)
    df['bsmtfin_sf_2'].fillna('NA', inplace=True)
    df['bsmt_unf_sf'].fillna(0, inplace=True)
    df['central_air'] = np.where(df['central_air'] == 'Y', 1, 0)
    df['bsmt_full_bath'].fillna(0, inplace=True)
    df['garage_finish'].fillna('NA', inplace=True)
    df['garage_cars'].fillna('NA', inplace=True)
    df['garage_qual'].fillna('NA', inplace=True)
    df['garage_cond'].fillna('NA', inplace=True)
    df['bsmt_half_bath'].fillna(0, inplace=True)
    return df
```

DATASET & THE CLEANING PROCESS cont'd

On attempting to run my first Model with neighborhood I was getting an error message, upon investigation I realized that Landmrk and GrnHill was not in the test dataset so I initially dropped columns: 'neighborhood_Landmrk', 'neighborhood_GrnHill' after running get_dummies() on my columns



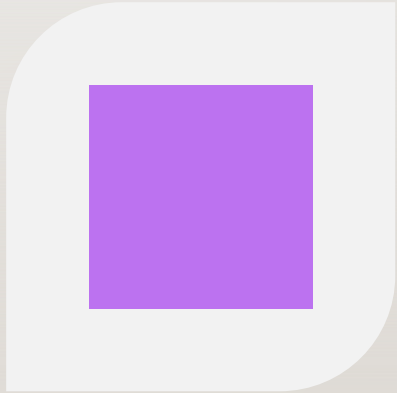
The 2 files were merged (test.csv and train.csv) and cleaned and get_dummies() was executed on the merged file. The file was the split before modeling

Which was accomplished using the following steps (which proved to be faster and give better results):

- test['SalePrice'] = 0
- final = pd.concat([train, test])

FIRST LINEAR REGRESSION MODEL

THIS MODEL WAS GENERATED USED 14 FEATURES THAT I LOOK AT IN A HOUSE:



"LOT_AREA", "STREET", "LOT_SHAPE", "1ST_FLR_SF", "2ND_FLR_SF", "YR_SOLD", "CONDITION_1", "NEIGHBORHOOD", "OVERALL_COND", 'GARAGE_AREA', 'YEAR_REMOD/ADD', 'FULL_BATH', 'TOTRMS_ABVGRD'

RESULTS – BASED ON MY FEATURES

Model Evaluation:

Train:

- R2 Score: 0.8034292369955875;
- RMSE: 35244.14239234582

Test:

- R2 Score: 0.831334523033619 9;
- RMSE: 31511.070831860678

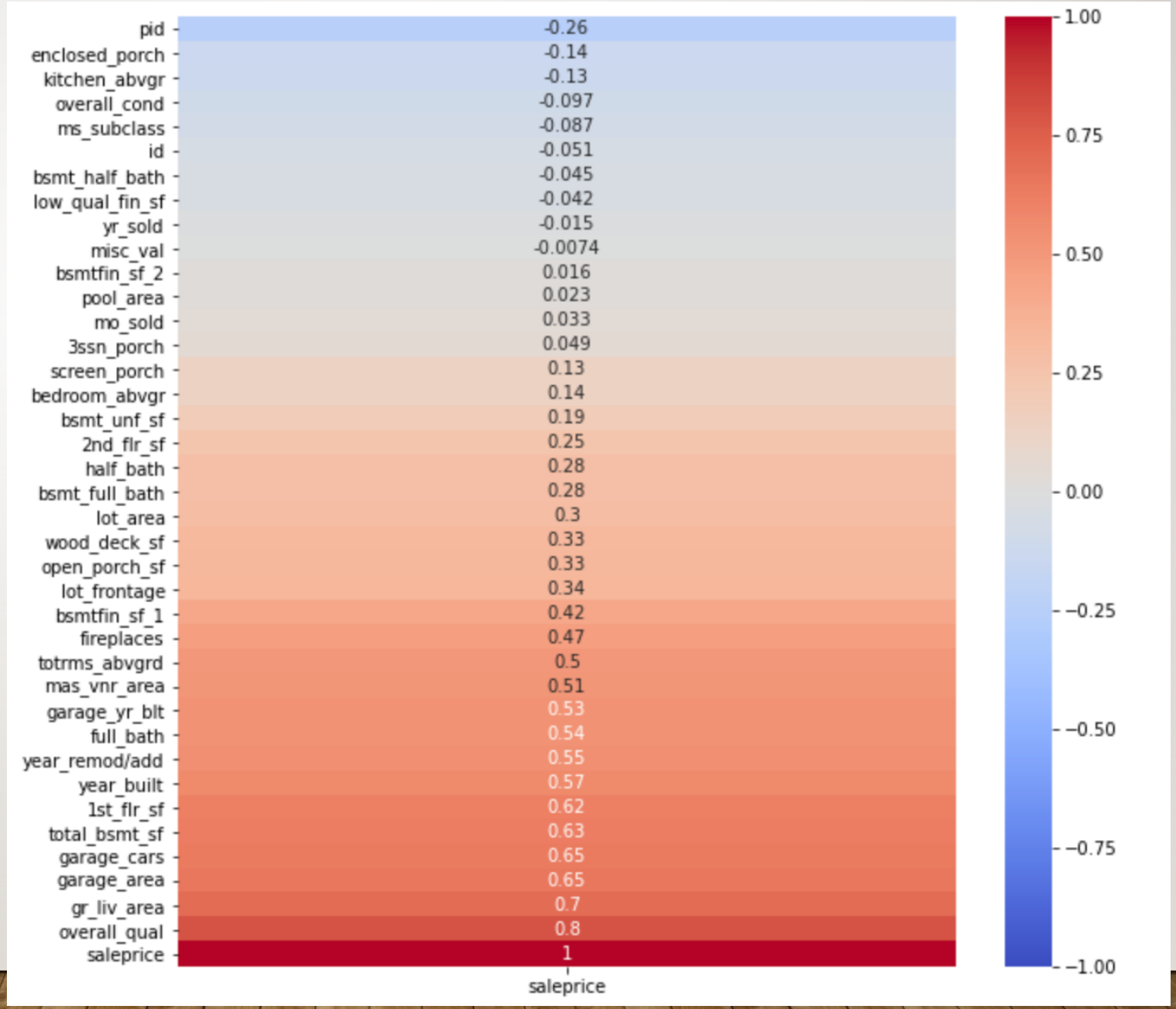
On submitting to Kaggle I got a mean square of over 38,954.85

EDA

EDA were then plotted with the numerical fields to to visualize the data and determine which fields were appropriate for using in the model:

- Heatmap
- Scatter plot
- histograms

Heatmap showing the correlation Between the Sale prices and the numeric values



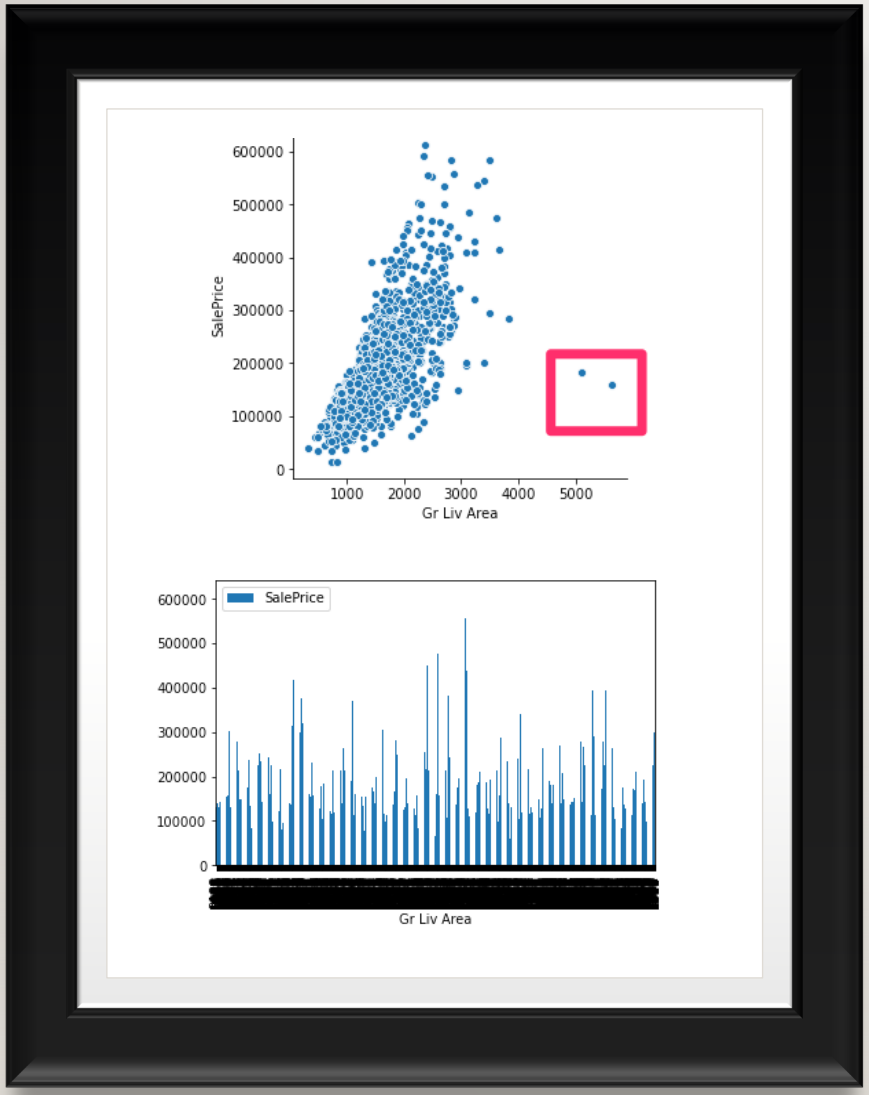
EDA CONT'D

The features with an absolute value of >= 0.4 was identified and would be used in all our models unless otherwise discovering from my modeling process

overall_qual	0.800207
year_built	0.571849
year_remod/add	0.550370
mas_vnr_area	0.512230
bsmtfin_sf_1	0.423519
total_bsmt_sf	0.628925
1st_flr_sf	0.618486
gr_liv_area	0.697038
full_bath	0.537969
totrms_abvgrd	0.504014
fireplaces	0.471093
garage_yr_blt	0.533922
garage_cars	0.648220
garage_area	0.650270

	SalePrice	Lot Area	Overall Qual	TotRms AbvGrd	Gr Liv Area
count	2051.000000	2051.000000	2051.000000	2051.000000	2051.000000
mean	181469.701609	10065.208191	6.112140	6.435885	1499.330083
std	79258.659352	6742.488909	1.426271	1.560225	500.447829
min	12789.000000	1300.000000	1.000000	2.000000	334.000000
25%	129825.000000	7500.000000	5.000000	5.000000	1129.000000
50%	162500.000000	9430.000000	6.000000	6.000000	1444.000000
75%	214000.000000	11513.500000	7.000000	7.000000	1728.500000
max	611657.000000	159000.000000	10.000000	15.000000	5642.000000

EDA CONT'D

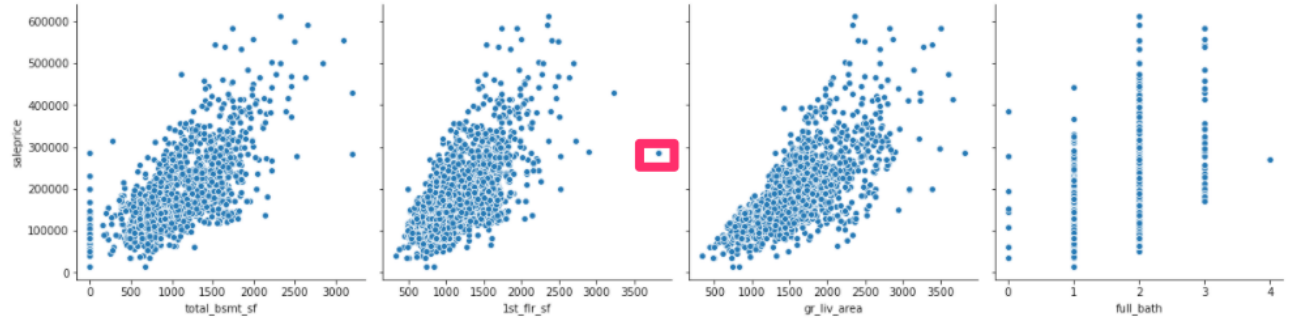


- A scattered plot on 'SalePrice' vs 'Gr Liv Area' identified 2 outlier both having a sq ft over 5,000 and a sale price of under 200k rows which will be deleted from the train dataset

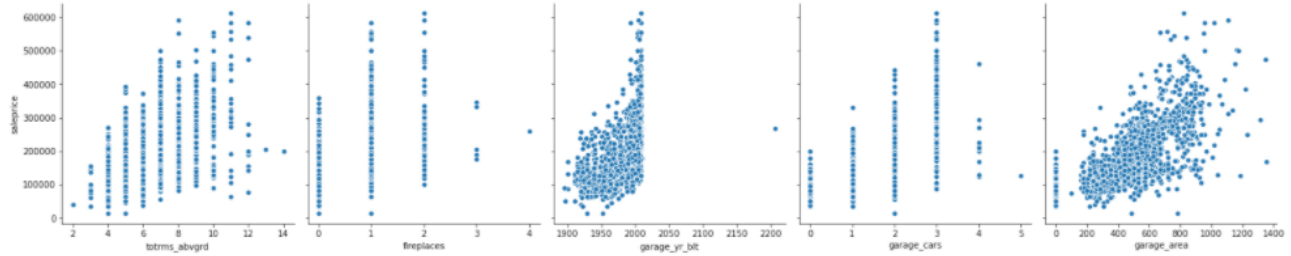
SCATTERED PLOT AGAINST SALE PRICE VS NUMERIC FEATURES

SCATTERED PLOTS CONT'D

```
sns.pairplot(df, y_vars=['saleprice'], x_vars=["total_bsmt_sf", "1st_flr_sf", "gr_liv_area", "full_bath"], height=4);
```




```
sns.pairplot(df, y_vars=['saleprice'], x_vars=["totrms_abvgrd", "fireplaces", "garage_yr_blt", "garage_cars", "garage_a
```



EDA BASED MODEL

Model Based on the EDA:

- Which produced a mean square of over 6,925,364.90 (although the R2 score was better)

The following features were then added to my model to if it will produce better results:

'lot_area', 'street', 'lot_shape', '1st_flr_sf', '2nd_flr_sf', 'yr_sold', 'condition_1', 'neighborhood', 'overall_qual', 'gr_liv_area', 'garage_area', 'garage_cars', 'total_bsmt_sf', '1st_flr_sf', 'year_remod/add', 'year_built', 'full_bath', 'mas_vnr_area', 'garage_yr_blt', 'totrms_abvgrd', 'fireplaces', 'bsmtfin_sf_1', 'overall_cond', 'overall_qual', 'gr_liv_area', 'total_bsmt_sf', '1st_flr_sf'

MODEL RESULT (with some other features added)

- I got a mean square of over 437,850.76 although my R2 and RMSE scores were better on my train dataset

Model Evaluation:

Train:

- R2 Score: 0.8730860914461 47
- RMSE: 28319.27267172839

Test:

- R2 Score: 0.8642674713640728
- RMSE: 28267.801467065 11

ACTION

On comparing both scores I decided to use the Model with my 14 features and add/removed other features based on my results

After Several attempt at massaging my data the best results were produced when I added just a few more features to my initial features set

Best score: 33122.74 (with only had 21 features

CONCLUSION

My model is in no way the ideal Model, but although I was not happy with my unscaled and unregularized version it produced the best results; therefore, it will be the model I am going to Productionize today.

In the further I will spend less time on my Linear Regression Model and let Lasso and Ridge do the work for me.

NEXT STEPS

Apply StandardScaler to the data set before modeling

Work with several other combination of feature, also use Lasso and Ridge to regularize and improve the model