PREDICTING THE PRICE OF HOUSES IN AMES, IA

OUR PROBLEM STATEMENT

- As a member of a firm based in neighbouring Des Moines that intends to set up in Ames, Iowa, our boss has asked our team to come up with an independent analysis of the market there which can be used to map to our existing assumptions about real estate.
  - Location is important
  - Quality of the house
  - Question: Are there other features that are important?
- As a data analyst within the firm, we have been tasked to explore the dataset that our procurement team got, and report our findings and limitations.

Target audience: Immediate project lead and Board.

EXPLORATORY DATA ANALYSIS (EDA)

- 2051 points of data
- 81 different variables

Square Footage	Location	Quality	Exterior	Basement	Other features
Total Footage	MS - Zoning	Date of house built	Style of house	Quality and Condition	Land Slope
1st Floor Area	Neighborhood	Date of recent modifications	Style of roof	Exposure to Grade level	Land Contour
Garage Area	Proximity to various conditions (e.g. Arterial street)	Garage – Quality, Condition	Exterior coverings of house	Rating of finished area	Alley access
Garage space (by car)		Fireplace- No., Quality	Type of Dwelling	Basement Space available	Shape of Lot
		Foundation Type	Masonry Veneer – Area and Type	Presence of half or full baths	
		Kitchen Quality	External- Qual and Condition		
		Pool	Fence		
		Porch	Type of Driveway		
		Heating			
		Electrical System			

EDA - THINGS TO CONSIDER

- Date of House, Renovation and Garage built
  - Convert that to "Time Since" or "Age"
  - Cleaning of erroneous data (e.g. Garage that was built 200 years before house was sold)
- Exploring of data that have null values

Data	Missing Data
lot_frontage	330
alley	1909
mas_vnr_type	22
mas_vnr_area	22
bsmt_qual	55
bsmt_cond	55
bsmt_exposure	58
bsmtfin_type_1	55
bsmtfin_sf_1	1
bsmtfin_type_2	56
bsmtfin_sf_2	1
bsmt_unf_sf	1
total_bsmt_sf	1

bsmt_full_bath	2
bsmt_half_bath	2
fireplace_qu	1000
garage_type	113
garage_finish	114
garage_cars	1
garage_area	1
garage_qual	114
garage_cond	114
pool_qc	2040
fence	1649
misc_feature	1985
age_garage	114

EDA - THINGS TO CONSIDER

- Date of House, Renovation and Garage built
  - Convert that to "Time Since" or "Age"
  - Cleaning of erroneous data (e.g. Garage that was built 200 years before house was sold)
- Exploring of data that have null values
  - Lot Frontage with 330 null values: No linear connection to street
  - Alley with 1909 null values: No access to the alley
  - Fire Quality with 1000 null values: No Fireplace
  - Pool Quality with 2040 null values: No Pool
  - Fence with 1649 null values: No Fence

EDA – FEATURE SELECTION

- Numerical Data vs Categorical Data
  - Overall Quality, Presence of Baths vs MS Zoning, Alley Access

EDA – FEATURE SELECTION

- Numerical Data vs Categorical Data
  - Overall Quality, Presence of Baths vs MS Zoning, Alley Access

EDA – FEATURE SELECTION (NUMERICAL)

saleprice	1.000000
overall_qual	0.800718
gr_liv_area	0.707159
garage_area	0.650068
garage_cars	0.646957
total_bsmt_sf	0.643262
1st_flr_sf	0.636332
full_bath	0.539232
mas_vnr_area	0.516955
totrms_abvgrd	0.509614
fireplaces	0.472419
bsmtfin_sf_1	0.430006
open_porch_sf	0.331283
wood_deck_sf	0.329713
lot_area	0.298608
half_bath	0.280414
2nd_flr_sf	0.251278
bsmt_unf_sf	0.188752
lot_frontage	0.180196

bedroom_abvgr	0.140195
screen_porch	0.138037
3ssn_porch	0.049944
mo_sold	0.025703
pool_area	0.023797
bsmtfin_sf_2	0.017764
misc_val	-0.009608
yr_sold	-0.011662
low_qual_fin_sf	-0.041083
id	-0.054207
ms_subclass	-0.085452
overall_cond	-0.092714
kitchen_abvgr	-0.126112
enclosed_porch	-0.137827
pid	-0.251008
recent_remod/add	-0.550562
age_house	-0.571165

EDA – FEATURE SELECTION (NUMERICAL)

saleprice	1.000000
overall_qual	0.800718
gr_liv_area	0.707159
garage_area	0.650068
garage_cars	0.646957
total_bsmt_sf	0.643262
1st_flr_sf	0.636332
full_bath	0.539232
mas_vnr_area	0.516955
totrms_abvgrd	0.509614
fireplaces	0.472419
bsmtfin_sf_1	0.430006
open_porch_sf	0.331283
wood_deck_sf	0.329713
lot_area	0.298608
half_bath	0.280414
2nd_flr_sf	0.251278
bsmt_unf_sf	0.188752
lot_frontage	0.180196

bedroom_abvgr	0.140195
screen_porch	0.138037
3ssn_porch	0.049944
mo_sold	0.025703
pool_area	0.023797
bsmtfin_sf_2	0.017764
misc_val	-0.009608
yr_sold	-0.011662
low_qual_fin_sf	-0.041083
id	-0.054207
ms_subclass	-0.085452
overall_cond	-0.092714
kitchen_abvgr	-0.126112
enclosed_porch	-0.137827
pid	-0.251008
recent_remod/add	-0.550562
age_house	-0.571165

EDA – FEATURE SELECTION (NUMERICAL)

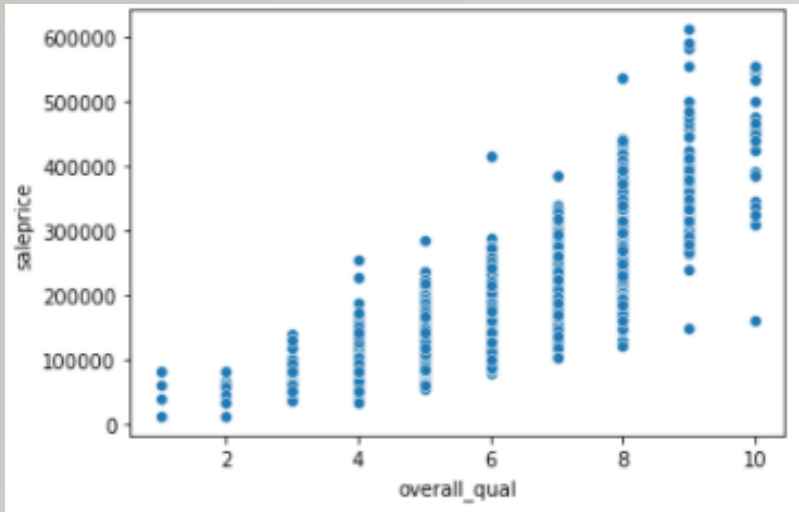
saleprice	1.000000
overall_qual	0.800718
gr_liv_area	0.707159
garage_area	0.650068
garage_cars	0.646957
total_bsmt_sf	0.643262
1st_flr_sf	0.636332
full_bath	0.539232
mas_vnr_area	0.516955
totrms_abvgrd	0.509614
fireplaces	0.472419
bsmtfin_sf_1	0.430006
open_porch_sf	0.331283
wood_deck_sf	0.329713
lot_area	0.298608
half_bath	0.280414
2nd_flr_sf	0.251278
bsmt_unf_sf	0.188752
lot_frontage	0.180196

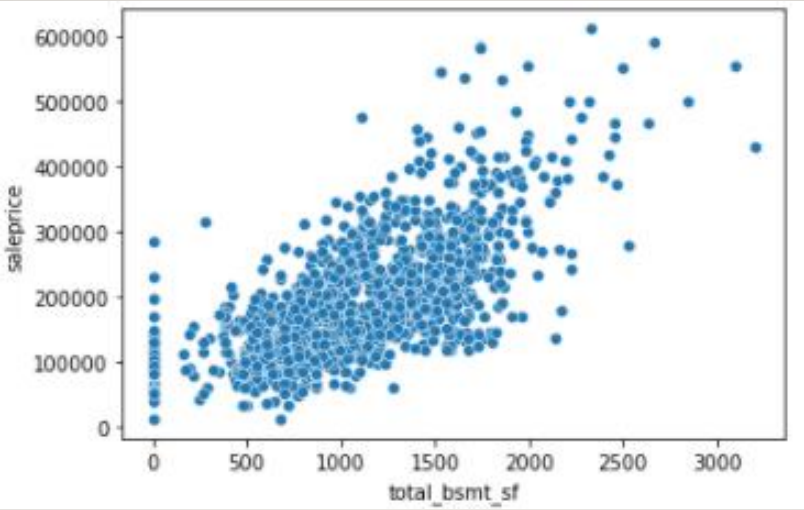
13 Numerical Variables Chosen

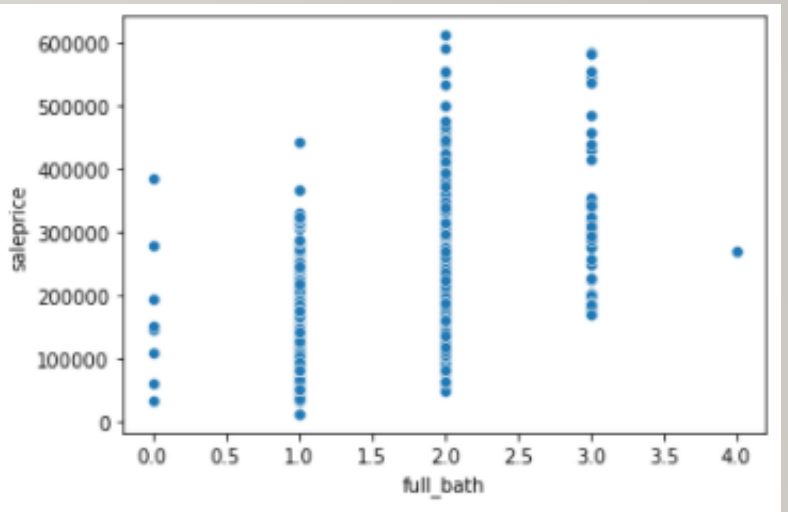
bedroom_abvgr	0.140195
screen_porch	0.138037
3ssn_porch	0.049944
mo_sold	0.025703
pool_area	0.023797
bsmtfin_sf_2	0.017764
misc_val	-0.009608
yr_sold	-0.011662
low_qual_fin_sf	-0.041083
id	-0.054207
ms_subclass	-0.085452
overall_cond	-0.092714
kitchen_abvgr	-0.126112
enclosed_porch	-0.137827
pid	-0.251008
recent_remod/add	-0.550562
age_house	-0.571165

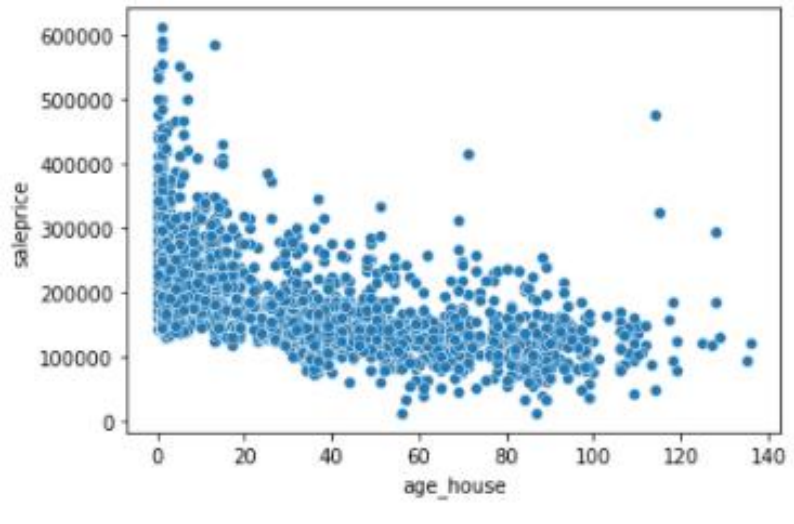
EDA – FEATURE SELECTION (NUMERICAL)

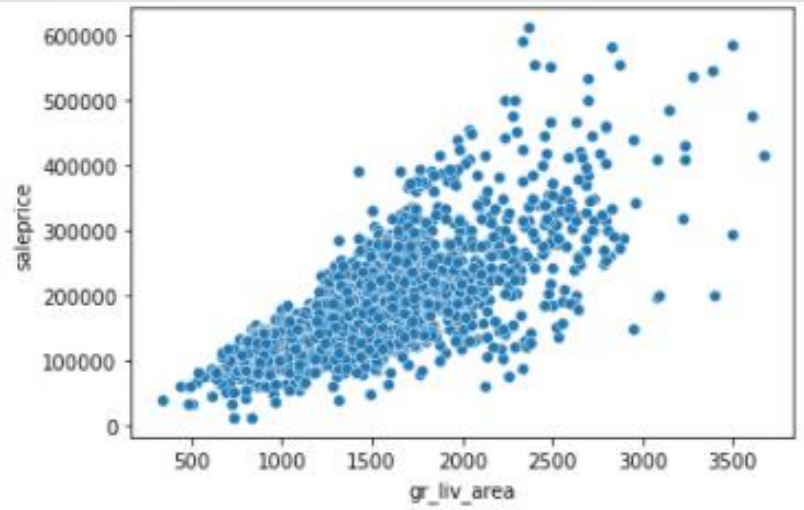
- Numerical Data vs Categorical Data
  - Overall Quality, Presence of Baths vs MS Zoning, Alley Access
- Numerical Data

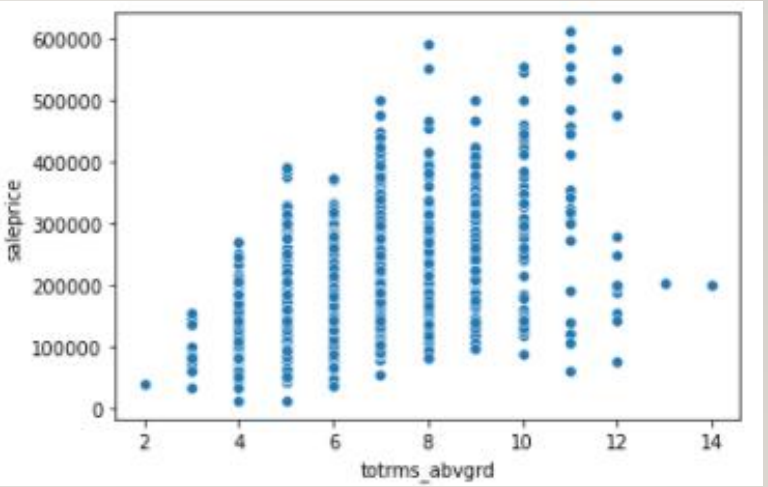

saleprice
overall_qual

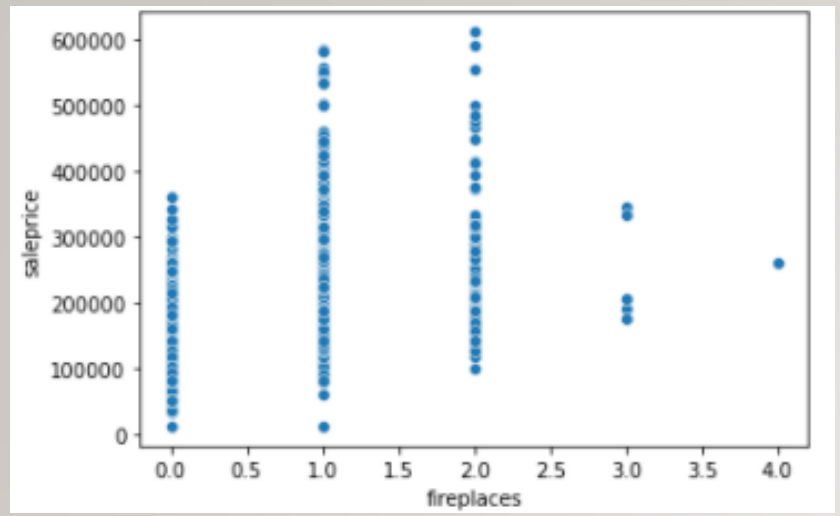

saleprice
total_bsmt_sf

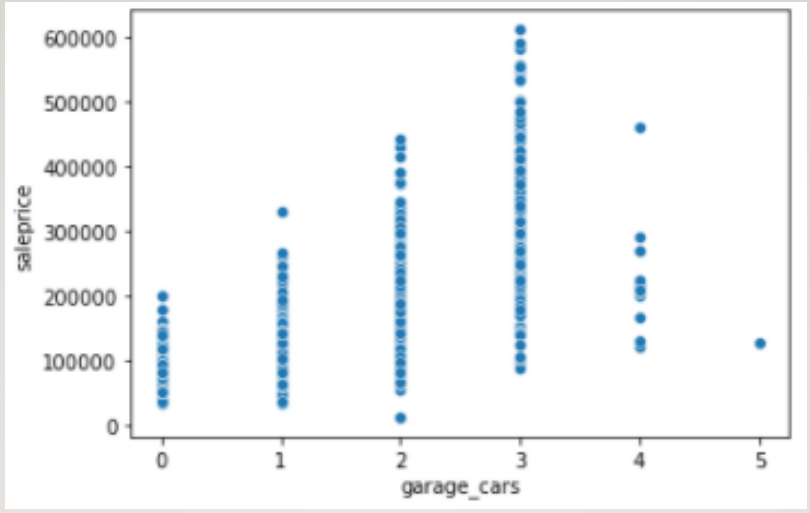

saleprice
full_bath

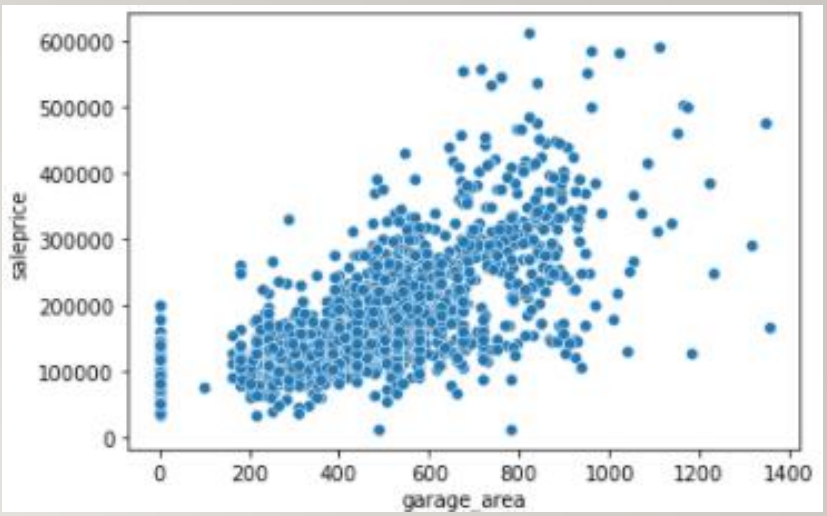

saleprice
age_house

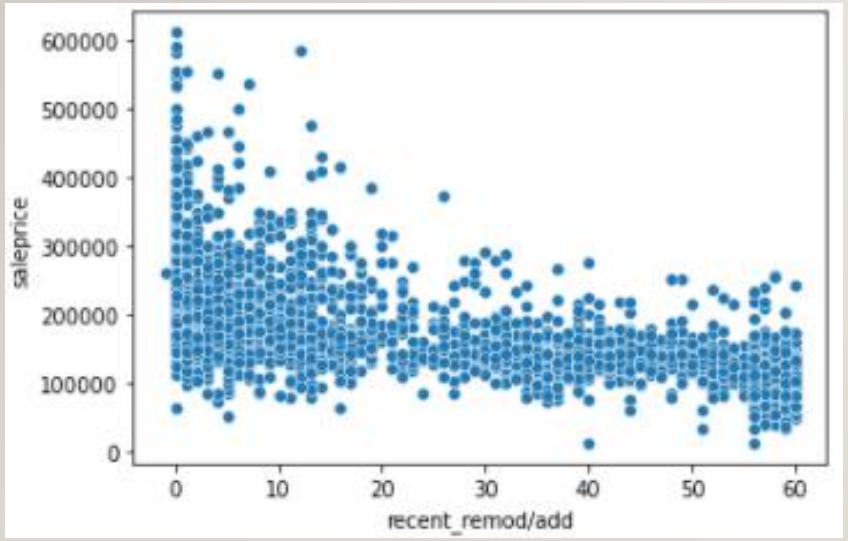

saleprice
gr_liv_area

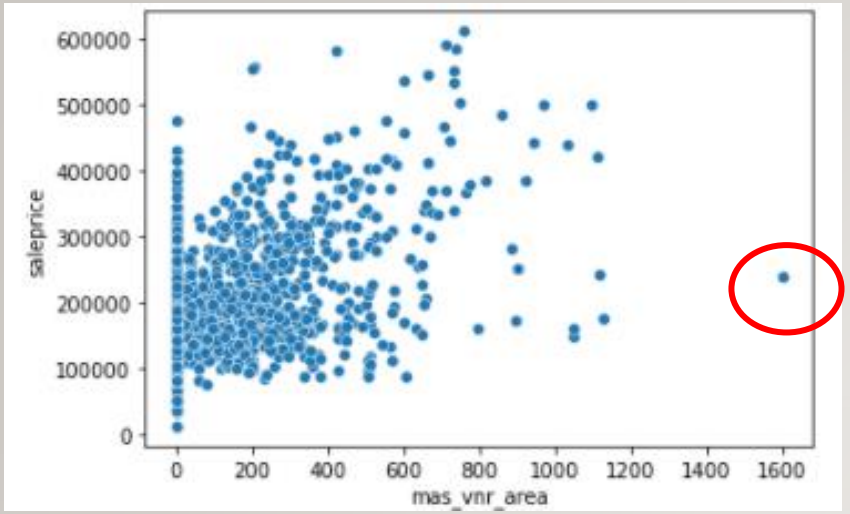

saleprice
totrms_abvgrd

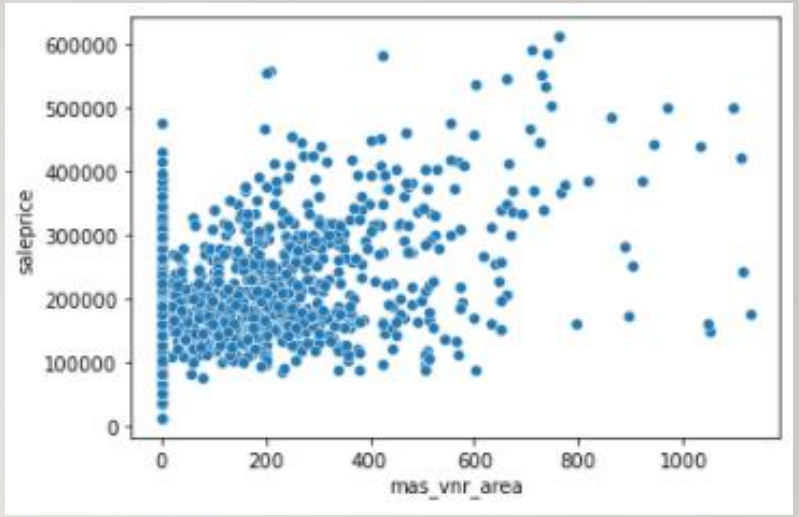

saleprice
fireplaces
0
100000
200000
300000
400000
500000
600000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

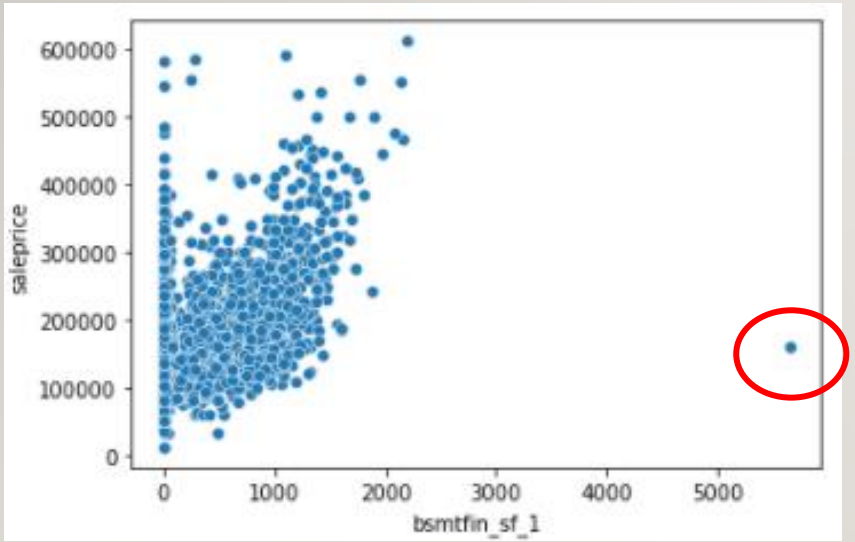

saleprice
garage_cars
0
100000
200000
300000
400000
500000
600000
0
1
2
3
4
5

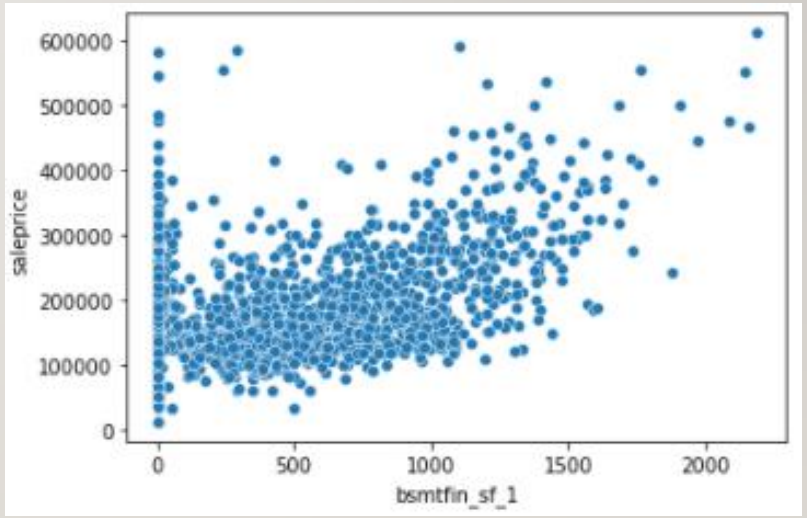

saleprice
garage_area
0
100000
200000
300000
400000
500000
600000
0
200
400
600
800
1000
1200
1400

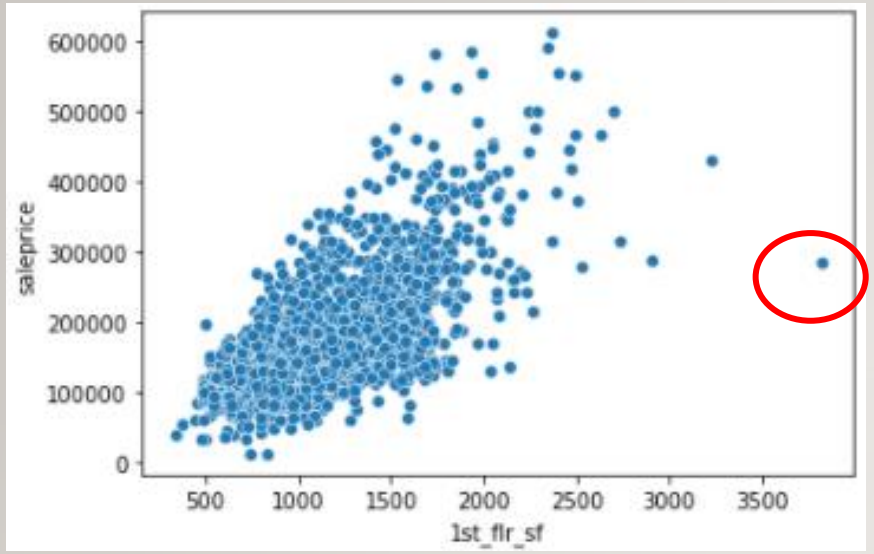

saleprice
recent_remod/add
0
100000
200000
300000
400000
500000
600000
0
10
20
30
40
50
60

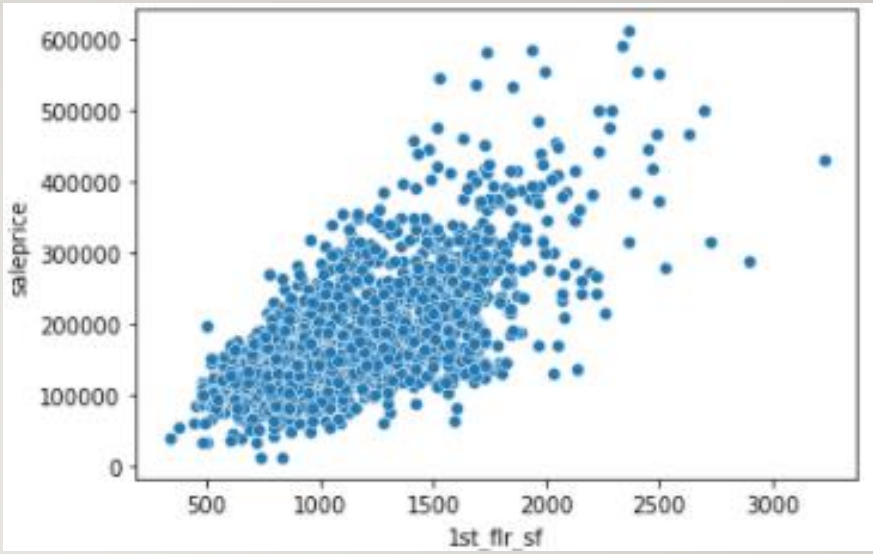

saleprice
mas_vnr_area
0
100000
200000
300000
400000
500000
600000
0
200
400
600
800
1000
1200
1400
1600


saleprice
mas_vnr_area
0
100000
200000
300000
400000
500000
600000
0
200
400
600
800
1000


saleprice
bsmtfin_sf_1
0
100000
200000
300000
400000
500000
600000
0
1000
2000
3000
4000
5000


saleprice
bsmtfin_sf_1
0
100000
200000
300000
400000
500000
600000
0
500
1000
1500
2000


saleprice
600000
500000
400000
300000
200000
100000
0
500
1000
1500
2000
2500
3000
3500
1st_flr_sf


saleprice
600000
500000
400000
300000
200000
100000
0
500
1000
1500
2000
2500
3000
1st_flr_sf

EDA – FEATURE SELECTION (CATEGORICAL)

- Numerical Data vs Categorical Data
  - Overall Quality, Presence of Baths vs MS Zoning, Alley Access

```
(['ms_zoning', 'street', 'alley', 'lot_shape', 'land_contour',
  'utilities', 'lot_config', 'land_slope', 'neighborhood', 'condition_1',
  'condition_2', 'bldg_type', 'house_style', 'roof_style', 'roof_matl',
  'exterior_1st', 'exterior_2nd', 'mas_vnr_type', 'exter_qual',
  'exter_cond', 'foundation', 'bsmt_qual', 'bsmt_cond', 'bsmt_exposure',
  'bsmtfin_type_1', 'bsmtfin_type_2', 'heating', 'heating_qc',
  'central_air', 'electrical', 'kitchen_qual', 'functional',
  'fireplace_qu', 'garage_type', 'garage_finish', 'garage_qual',
  'garage_cond', 'paved_drive', 'pool_qc', 'fence', 'misc_feature',
  'sale_type', 'saleprice'],
 dtype='object')
```

- 43 Variables

```
(['ms_zoning', 'street', 'alley', 'lot_shape', 'land_contour',
  'utilities', 'lot_config', 'land_slope', 'neighborhood', 'condition_1',
  'condition_2', 'bldg_type', 'house_style', 'roof_style', 'roof_matl',
  'exterior_1st', 'exterior_2nd', 'mas_vnr_type', 'exter_qual',
  'exter_cond', 'foundation', 'bsmt_qual', 'bsmt_cond', 'bsmt_exposure',
  'bsmtfin_type_1', 'bsmtfin_type_2', 'heating', 'heating_qc',
  'central_air', 'electrical', 'kitchen_qual', 'functional',
  'fireplace_qu', 'garage_type', 'garage_finish', 'garage_qual',
  'garage_cond', 'paved_drive', 'pool_qc', 'fence', 'misc_feature',
  'sale_type', 'saleprice'],
 dtype='object')
```

- Underwent changing of ordinal data to values

  e.g. Lot Shape

```
['Reg'],4)
['IR1'],3)
['IR2'],2)
['IR3'],1)
```

- Dropped variables that are too skewed to one category in data points e.g. Street

```
Pave    0.996533
Grvl    0.003467
Name: street, dtype: float64
```

- Dropped 8 variables:
  - Street
  - Utilities
  - Lot Configuration
  - Roof Style
  - Heating
  - Functional
  - Pool
  - Miscellaneous Features

FEATURES TO CONSIDER

- Did three different models based on:
  - Square footage and Location
  - Quality of house and Exterior
  - Basement and Other Features

FEATURES TO CONSIDER

- Sq Footage
  - '1st_flr_sf',
  - 'gr_liv_area'
  - 'garage_cars'
  - 'garage_area'
- Location
  - Ms_zoning,
  - Neighbourhood,
  - Condition_1,
  - Condition_2
- Quality
  - 'overall_qual'
  - 'overall_cond
  - 'age_house',
  - 'recent_remod/add',
  - 'full_bath',
  - 'totrms_abvgrd',
  - 'fireplaces',
- Exterior
  - House_style
  - Roof_style
  - Exterior_1st
  - Exterior_2nd
  - Bldg_type
  - 'mas_vnr_area',
  - Mas_vnr_type
  - Exter_qual
  - Exter_cond
- Basement
  - Bsmt_qual
  - Bsmt_cond
  - Bsmt_exposure
  - Bsmtfin_type_1
  - Bsmtfin_type_2
  - 'bsmtfin_sf_1',
  - 'total_bsmt_sf',
- Other Features
  - Alley
  - Lot-shape
  - Land-contour
  - Land slope

Dummified variables, in total there were 117 columns

SQ FOOTAGE, LOCATION

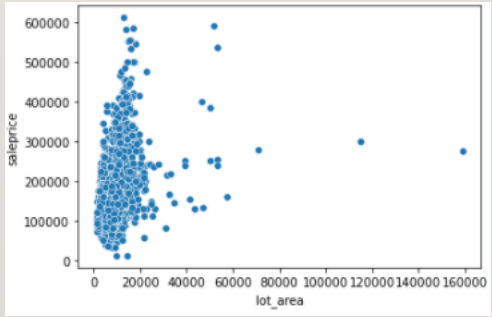
- Sq footage, Location: RSME of 36131
- + Quality and Exterior: RSME of 28218
- + Basement and Other Features: RSME of 25180

SQ FOOTAGE, LOCATION

- Sq footage, Location: RSME of 36131
- + Quality and Exterior: RSME of 28218
- + Basement and Other Features: RSME of 25180

Further EDA and Cleaning to minimise our error

- Took away outliers in garage cars, rooms above ground, and fireplaces
- Log-transformed lot area

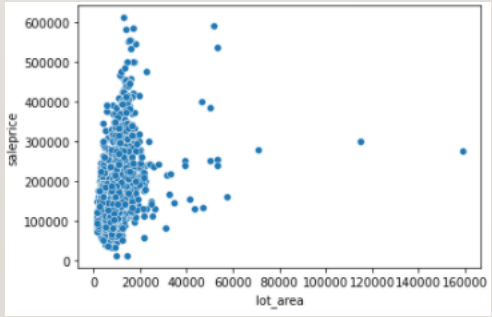


SQ FOOTAGE, LOCATION

- Sq footage, Location: RSME of 36131
- + Quality and Exterior: RSME of 28218
- + Basement and Other Features: RSME of 25180

Further EDA and Cleaning to minimise our error

- Took away outliers in garage cars, rooms above ground, and fireplaces
- Log-transformed lot area
- Ridge, Lasso and Elastic Net
- RMSE of 24969



IMPORTANT VARIABLES

- Above Ground living area
- Garage area
- Overall Quality of the house
- Overall Condition of the house
- Age of the house
- Exterior Quality of the house
- Basement Square Feet
- Basement Exposure

CONCLUSIONS / RECOMMENDATIONS

- There are other factors that can affect the price of the house, that might not be present in this dataset as can be seen in the high amount variability and error in RMSE.

1) Quality of local schools

2) Employment opportunities

3) Proximity to shoppping, entertainment and recreational centers

https://www.opendoor.com/w/blog/factors-that-influence-home-value

Other factors include:

4) Individual fittings and quality

https://www.yopa.co.uk/blog/how-does-an-estate-agent-value-a-property/

CONCLUSIONS / RECOMMENDATIONS

- Other factors that we can consider adopting and exploring:

- Location factors:
- Presence of good schools within a 10 minute driving distance
- Distance / Time taken by car to local supermarket
- Presence of greenery and parks
- Distance to city Business District

- Quality of the fittings inside the house:
- Window fittings
- Renovation of toilets and bedrooms

THANK YOU