(Not So) Extreme Makeover Home Edition

Group 1
Connie, Emi, YongHe

I want to increase my value of my house!

I want to sell my house but the price offered is too low!

I want to do renovation and I wish to have high ROI for future!

I just bought my house and I want a better resale value in future!

I want to buy a house for investment and I not sure whether it’s worth it…

Who Are We?

A group of home improvement consultants

What are we doing?

We provide suggestions on how to refurbish the houses in selected neighborhoods in Ames, Iowa.

How we do it?

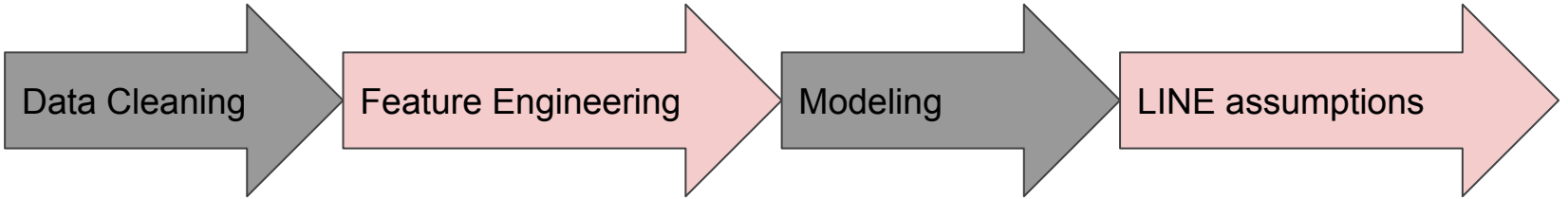
We select the best features for homeowners to renovate, in order to improve the value of their homes in a cost-effective way.

Data Used

Ames Iowa Housing dataset - From Kaggle

[some cost data to complement features selected]

Building Models



-null value
-outliers

-Categorical feature
- Features selection
- Final feature list

-Linear Regression
-Ridge
-LASSO

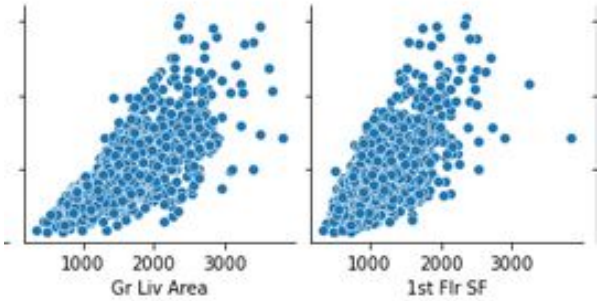
-Linearity (N)
-Independence (Y)
- Nominal distribution (Y)
- Equal variance of residuals (N)

Data Cleaning - Null Values

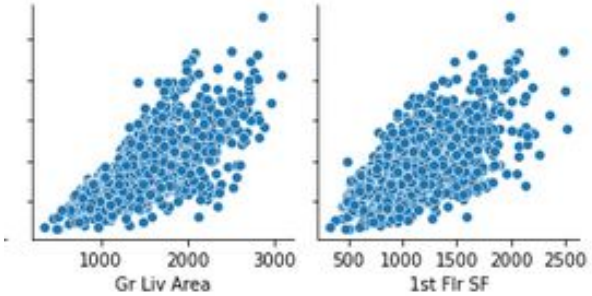
feature	null %	null #
Pool QC	99.56	2042
Misc Feature	96.83	1986
Alley	93.17	1911
Fence	80.50	1651
Fireplace Qu	48.76	1000
Lot Frontage	16.09	330
Garage Finish	5.56	114
Garage Qual	5.56	114
Garage Yr Blt	5.56	114
Garage Cond	5.56	114
Garage Type	5.51	113
Bsmt Exposure	2.83	58
BsmtFin Type 2	2.73	56
BsmtFin Type 1	2.68	55
Bsmt Cond	2.68	55
Bsmt Qual	2.68	55
Mas Vnr Area	1.07	22
Mas Vnr Type	1.07	22

- Drop features with too many null values
- Impute null values with zero or “NA” or “None”, based on data directory

Data Cleaning -Outliers



Before

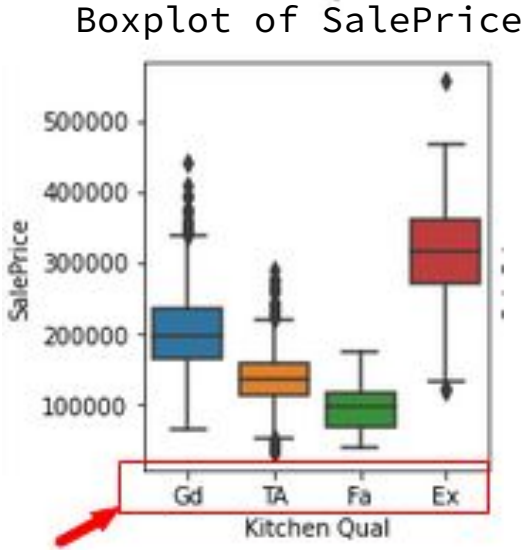


After

- outliers: data that fall outside of Q3 + 2.5*IQR
- 8% of data are identified as outliers

Feature Transforming - Categorical Features

- Target encoding is the process of replacing a categorical value with the mean or median of the target variable



Kitchen Qual	median SalePrice
Ex	317,250.00
Gd	196,000.00
TA	135,000.00
Fa	97,275.00

Target Encoding

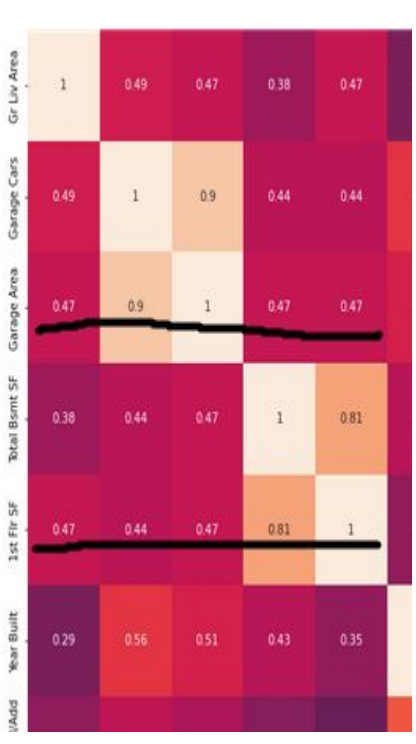
Kitchen Qual
317,250.00
196,000.00
135,000.00
97,275.00

Feature Selection

Corr with SalePrice

feature	corr
Gr Liv Area	0.712
Garage Cars	0.663
Garage Area	0.654
Total Bsmt SF	0.640
1st Flr SF	0.618
Year Built	0.617
Year Remod/Add	0.577
Full Bath	0.550
TotRms AbvGrd	0.498
Mas Vnr Area	0.449
Fireplaces	0.449
Open Porch SF	0.404
BsmtFin SF 1	0.403
Lot Area	0.382
Wood Deck SF	0.318
Garage Yr Blt	0.277
Half Bath	0.273
Bsmt Full Bath	0.259
2nd Flr SF	0.243
Bsmt Unf SF	0.192
Bedroom AbvGr	0.156
Screen Porch	0.082
3Ssn Porch	0.068
Mo Sold	0.048
Pool Area	0.019
BsmtFin SF 2	-0.004
Misc Val	-0.011
Yr Sold	-0.018
Bsmt Half Bath	-0.041
Low Qual Fin SF	-0.075
Kitchen AbvGr	-0.157
Enclosed Porch	-0.159

multicollinearity



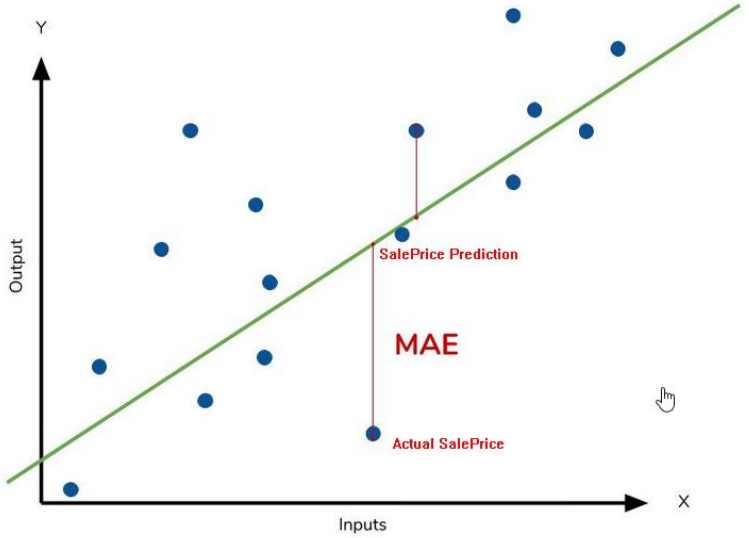
- Select features with an high absolute correlation (>0.4) with "SalePrice"
- Then drop the features when they have high correlation (>0.6) with the others

Final Features List for Modeling

'Exterior 1st',
'Garage Finish',
'Heating QC',
'Garage Cars',
'Foundation',
'Overall Qual',
'BsmtFin SF 1',
'Open Porch SF',
'Fireplaces',
'Overall Cond',
'BsmtFin Type 1',
'MS SubClass',
'Mas Vnr Area',
'Bsmt Exposure',
'Gr Liv Area',
'Total Bsmt SF'

Modeling - Metrics

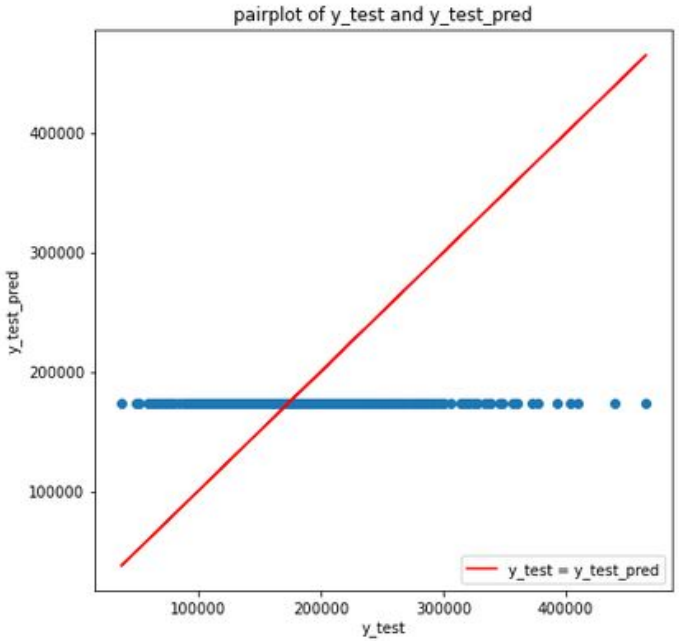
1. MAE: Mean Absolute Error
    - Measure accuracy of our models
    - The smaller, the better
    - Easier to interpret



2. Perc_diff: % of difference of MAE_train and MAE_test

    - Measure generalization of our models
    - Must less than 2% (absolute value)

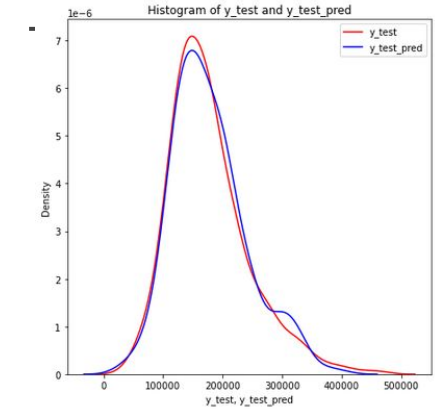
Baseline Model - Mean of y_train

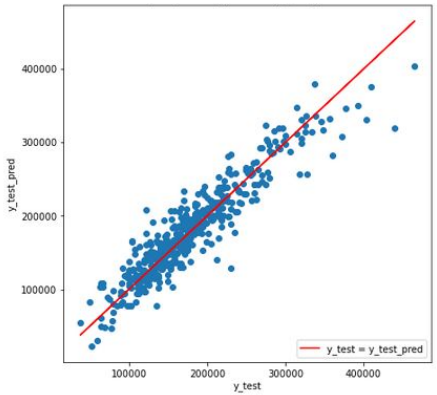


mean of y_train (SalePrice): $173,354

- mae of train data: $52,418 (30% of mean saleprice)
- mae of test data: $49,887 (28% of mean saleprice)
- diff%: 4.83% (did not pass)

Model - **linear regression**

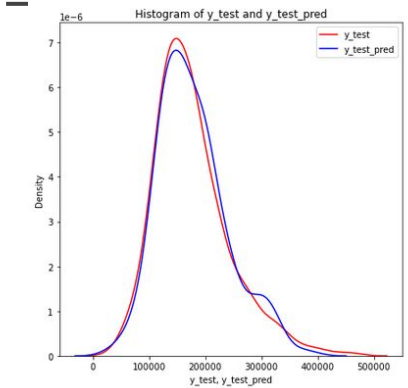


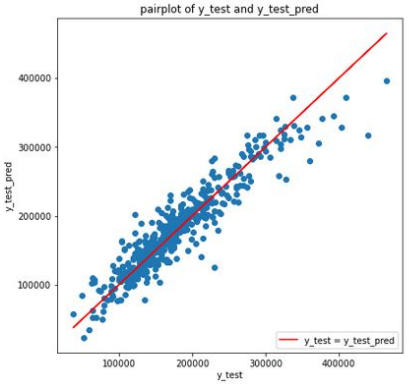


	feature	coef
0	Overall Qual	30,113.773
1	Gr Liv Area	25,962.788
2	BsmtFin SF 1	17,429.760
3	Heating QC	9,742.835
4	Total Bsmt SF	8,672.934
5	Garage Cars	8,368.176
6	Fireplaces	7,726.425
7	Foundation	6,182.984
8	Exterior 1st	5,782.254
9	Mas Vnr Area	2,821.262
10	Garage Finish	2,469.086
11	MS SubClass	1,965.329
12	BsmtFin Type 1	-1,871.241
13	Open Porch SF	1,602.638
14	Bsmt Exposure	1,466.907
15	Overall Cond	-1,254.350

- mae of train data: $17,516 (10% of mean saleprice)
- mae of test data: $17,767 (10% of mean saleprice)
- diff%: -1.43% (pass)

Model - Ridge Regression

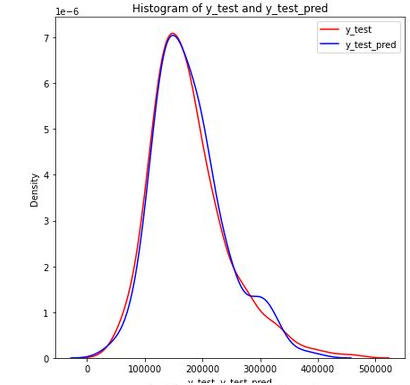


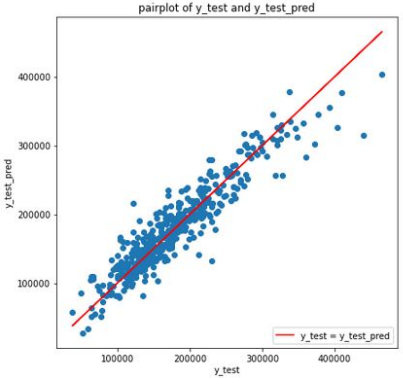


	feature	coef
0	Overall Qual	28,151.862
1	Gr Liv Area	24,144.596
2	BsmtFin SF 1	14,816.584
3	Total Bsmt SF	9,471.839
4	Garage Cars	8,779.481
5	Heating QC	8,665.006
6	Fireplaces	8,142.117
7	Foundation	5,589.669
8	Exterior 1st	5,367.248
9	Mas Vnr Area	3,362.273
10	Garage Finish	3,009.724
11	MS SubClass	2,876.809
12	Open Porch SF	2,117.680
13	Bsmt Exposure	1,627.787
14	BsmtFin Type 1	419.899
15	Overall Cond	-342.822

- mae of train data: $17,482 (10% of mean saleprice)
- mae of test data: $17,429 (10% of mean saleprice)
- diff%: 0.30% (pass)

Model - Lasso Regression





	feature	coef
0	Overall Qual	31,486.069
1	Gr Liv Area	25,590.859
2	BsmtFin SF 1	14,917.827
3	Total Bsmt SF	8,843.634
4	Garage Cars	8,239.459
5	Heating QC	7,327.816
6	Fireplaces	6,149.737
7	Foundation	4,073.732
8	Exterior 1st	3,537.664
9	Garage Finish	2,392.419
10	Mas Vnr Area	2,321.713
11	MS SubClass	1,865.039
12	Bsmt Exposure	1,552.332
13	Open Porch SF	1,145.275
14	Overall Cond	0.000
15	BsmtFin Type 1	0.000

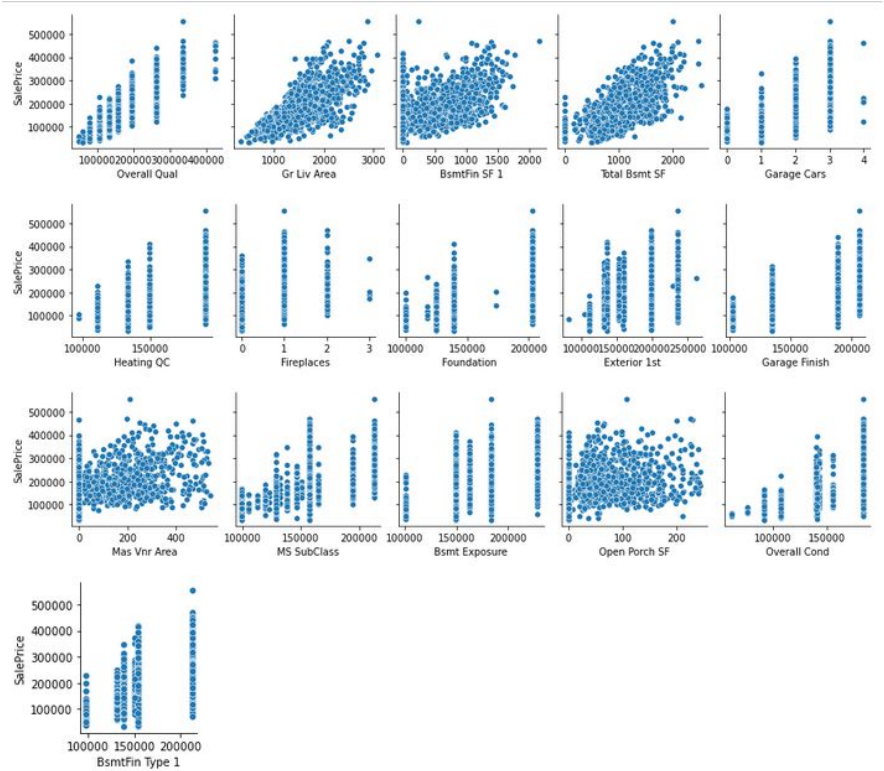
- mae of train data: $17,390 (10% of mean saleprice)
- mae of test data: $17,429 (10% of mean saleprice)
- diff%: -0.22% (pass)

Production Model - Lasso Regression

---

1. 3 models have similar performance, indicating our model might be underfitting
2. Lasso Regression Model has fewer features and is chosen as production model

LINE Assumption - **Linearity**



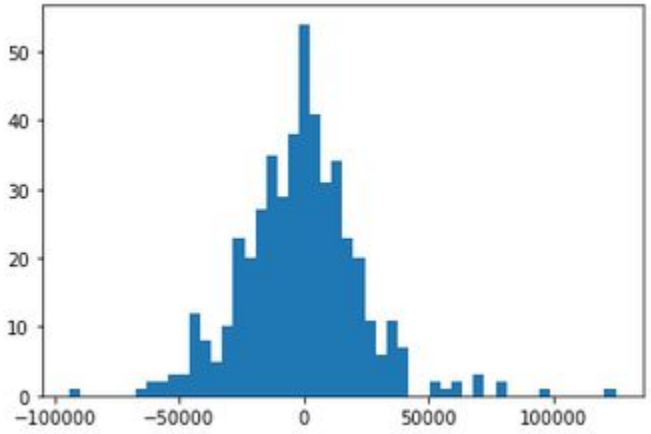
- Most features seem positively correlated but not linearly related with “SalePrice”

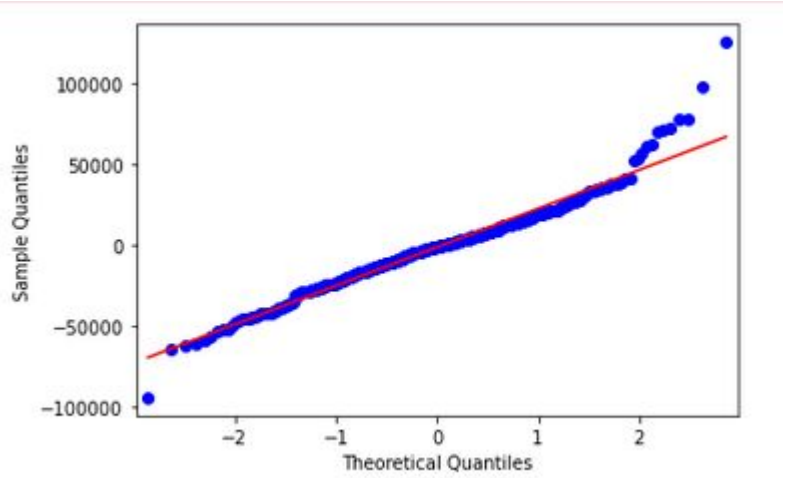
LINE Assumption - **Independence**

	feature	coef	vif
0	Overall Qual	31,481.693	2.702
1	Gr Liv Area	25,590.818	1.926
2	BsmtFin SF 1	14,916.819	1.581
3	Total Bsmt SF	8,846.192	1.905
4	Garage Cars	8,240.426	1.969
5	Heating QC	7,329.821	1.565
6	Fireplaces	6,149.993	1.368
7	Foundation	4,070.590	2.391
8	Exterior 1st	3,539.114	1.591
9	Garage Finish	2,385.688	2.093
10	Mas Vnr Area	2,322.913	1.288
11	MS SubClass	1,871.622	2.001
12	Bsmt Exposure	1,553.231	1.422
13	Open Porch SF	1,144.602	1.261
14	BsmtFin Type 1	0.000	1.901
15	Overall Cond	0.000	1.486

- All features have VIF < 5, indicating they are independent with other other

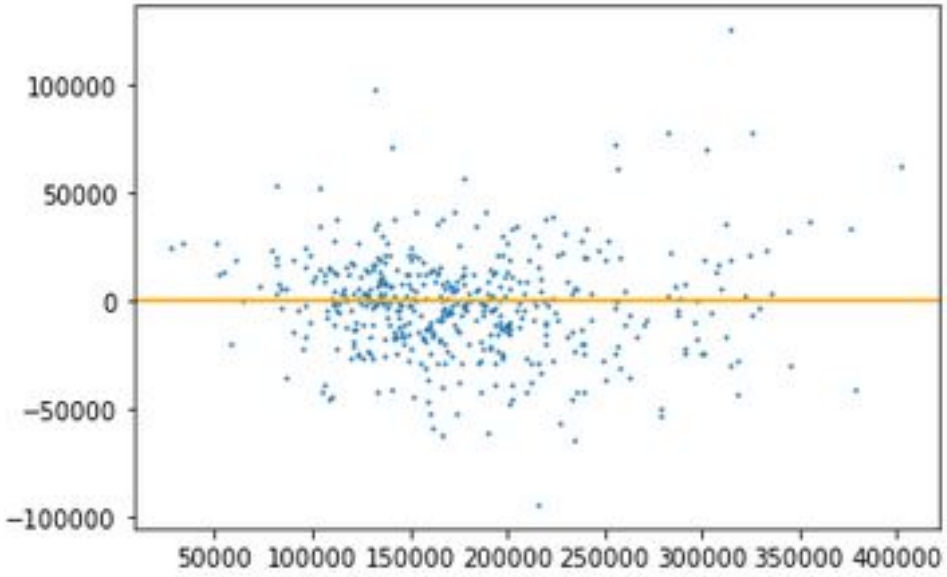
LINE Assumption - **Normality of residuals**





- The residuals form a roughly normal distribution with long right tail

LINE Assumption - Equal Variance of residuals



- the variance of residuals seems not equal at each point

Shortlisted ideal features for home improvement:

	feature	coef
0	Overall Qual	31,334.76
1	Gr Liv Area	25,635.75
2	BsmtFin SF 1	15,102.21
3	Total Bsmt SF	8,827.28
4	Garage Cars	8,250.60
5	Heating QC	7,528.23
6	Fireplaces	6,300.76
7	Foundation	4,186.32
8	Exterior 1st	3,717.97
9	Garage Finish	2,397.19
10	Mas Vnr Area	2,363.46
11	MS SubClass	1,858.14
12	Bsmt Exposure	1,538.50
13	Open Porch SF	1,184.73
14	BsmtFin Type 1	0.00
15	Overall Cond	0.00

Criteria:

- **Features impact saleprice greatly**
- **Features are cost-effective**

- **'Overall Qual'** is an ordinal feature and very subjective, and it doesn't refer to specific part of a house. We do not suggest it as home improvement feature

- **'Gr Liv Area':** Above grade (ground) living area square feet. Most of time, it is very difficult to change it without changing the main structure. We do not recommend it as home improvement feature

- **'BsmtFin SF 1':** basement finished area, third important feature to affect sale price. Very suitable for home improvement

- **'Total Bsmt SF':**Total square feet of basement area, same as 'Gr Liv Area', not recommended for home improvement

- **'Garage Cars':** Size of garage in car capacity, same as 'Gr Liv Area', not recommended for home improvement

- **'Heating QC':** Heating quality and condition, ordinal and subjective feature, not recommended

- **'Fireplaces':** Number of fireplaces, easy to install. It is highly recommended for home improvement

- **'Foundation':** Type of foundation. Change of this feature is very costly, not worthy due to its limited contribution to the saleprice

- the remaining features are not considered because of their limited contribution to saleprice

Best Features for Home Improvement

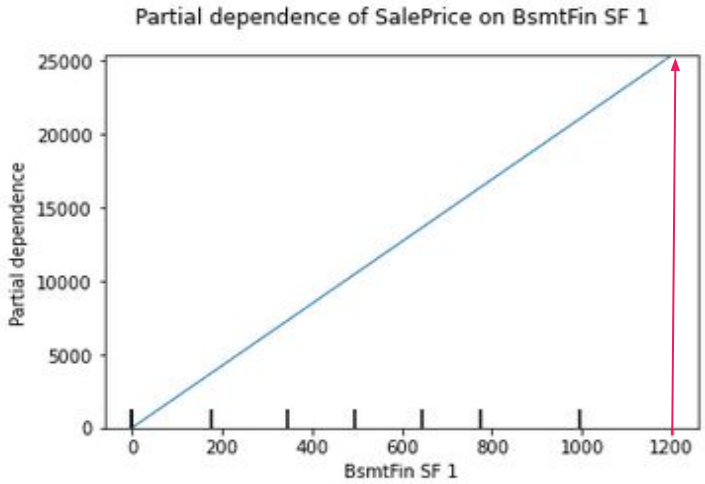
BsmtFin SF 1 : basement finished area

```
BsmtFin Type 1  (Ordinal): Rating of basement finished area

       GLQ      Good Living Quarters
       ALQ      Average Living Quarters
       BLQ      Below Average Living Quarters
       Rec      Average Rec Room
       LwQ      Low Quality
       Unf      Unfinshed
       NA       No Basement

BsmtFin SF 1 (Continuous): Type 1 finished square feet
```

	feature	coef
0	Overall Qual	31,481.693
1	Gr Liv Area	25,590.818
2	BsmtFin SF 1	14,916.819
3	Total Bsmt SF	8,846.192
4	Garage Cars	8,240.426
5	Heating QC	7,329.821
6	Fireplaces	6,149.993
7	Foundation	4,070.590
8	Exterior 1st	3,539.114
9	Garage Finish	2,385.688
10	Mas Vnr Area	2,322.913
11	MS SubClass	1,871.622
12	Bsmt Exposure	1,553.231
13	Open Porch SF	1,144.602
14	BsmtFin Type 1	0.000
15	Overall Cond	0.000

How does ‘BsmtFin SF 1’ impact “SalePrice”

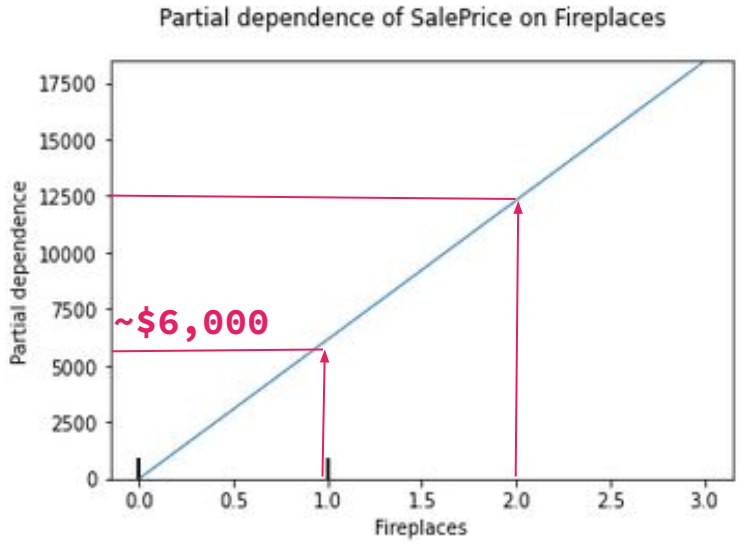
	SalePrice	BsmtFin SF 1
0	164,621.887	0.000
1	179,825.804	719.333
2	195,029.722	1,438.667
3	210,233.639	2,158.000



1200 square feet can increase the house value by $25,000

How does ‘Fireplaces’ impact “SalePrice”

	SalePrice	Fireplaces
0	170,061.026	0.000
1	176,211.020	1.000
2	182,361.013	2.000
3	188,511.006	3.000



Every fireplace can increase the house value by around $6,000

Comparison by neighborhoods

Old Neighborhoods

	Neighborhood	Avg Year Built
20	OldTown	1923
11	IDOTRR	1927
3	BrkSide	1932
21	SWISU	1932
6	Crawfor	1946
7	Edwards	1955
15	NAmes	1959
22	Sawyer	1963

New Neighborhoods

	Neighborhood	Avg Year Built
18	NoRidge	1995
25	StoneBr	1997
26	Timber	1997
5	CollgCr	1998
8	Gilbert	1998
24	Somerst	2004
0	Blmngtn	2005
19	NridgHt	2005

We have selected 1 old and 1 new neighborhood, to compare the features which affect SalePrice in each neighborhood

* Neighbourhoods selected based on data > 100 rows, to allow for a meaningful comparison of features identified.

Old Neighborhood - Oldtown

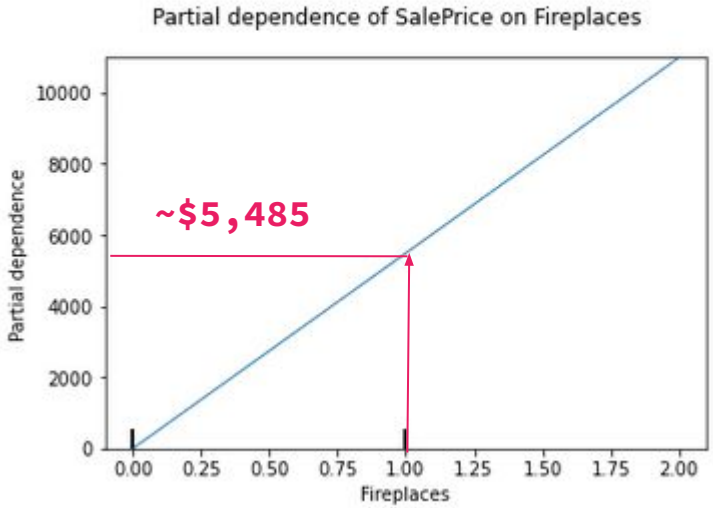
```
searching best model for OldTown
OldTown has 151 rows of data
{'classifier__alpha': 0.1}
mean of y_train: SalePrice   117,503.522
dtype: float64
mae of train data: 12,790.60
mae of test data: 12,550.07
diff%: 1.88%
           feature       coef   vif
0     Overall Qual 17,708.325 1.343
1       Heating QC 14,619.853 1.140
2      Garage Cars  9,354.771 1.524
3      Gr Liv Area  7,471.421 1.556
4      MS SubClass  6,567.059 1.414
5       Fireplaces  5,485.593 1.309
6    Total Bsmt SF  4,215.946 1.682
7     BsmtFin SF 1  2,011.240 1.353
8    Open Porch SF  1,977.464 1.223
9       Foundation  1,723.394 1.099
10    Overall Cond    297.427 1.097
11  BsmtFin Type 1    285.883 2.010
12    Exterior 1st     33.171 1.139
13    Mas Vnr Area     23.565 1.148
14   Garage Finish     -0.015 1.427
15   Bsmt Exposure     -0.010 2.395
y: 151
```

Based on the model for 'OldTown', 'Fireplaces' is also an important feature for home improvement.

In addition, Heating QC also ranks very highly in coefficients that affect SalePrice, and it could be because homes are older in this neighborhood and buyers may want to make sure insulation of these homes are still up to standard.

Old Neighborhood - Oldtown

	SalePrice	Fireplaces
0	116,219.19	0
1	121,704.78	1
2	127,190.38	2



How does 'Fireplaces' impact "SalePrice" in "Old Town"

The partial dependence analysis suggest that for every 1 fireplace increase, the sale price will increase by around 5,485 dollars (lower than model average)

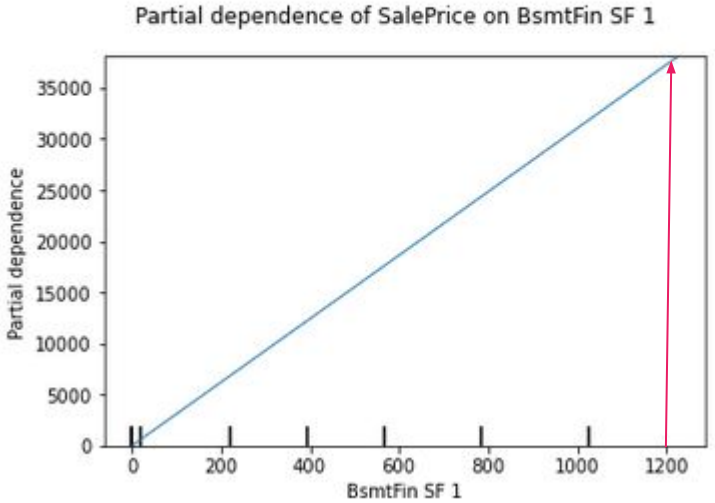
New Neighborhood - Somerst

```
searching best model for Somerst
Somerst has 125 rows of data
{'classifier__alpha': 432.87612810830615}
mean of y_train: 221,901
mae of train data: 13,873
mae of test data: 13,874
diff%: -0.01%
           feature      coef  vif
0      Gr Liv Area 31,191.41 2.42
1     BsmtFin SF 1 17,942.87 2.18
2    Total Bsmt SF 17,885.16 2.38
3     Overall Qual 16,361.24 1.81
4       Fireplaces  8,716.84 1.93
5    Open Porch SF  7,323.97 1.47
6     Mas Vnr Area -5,843.15 1.64
7     Exterior 1st  4,554.35 1.66
8   BsmtFin Type 1  4,165.49 2.03
9      Garage Cars  3,390.27 1.49
10   Garage Finish  3,150.37 1.82
11   Bsmt Exposure    965.46 1.52
12     MS SubClass   -662.49 2.37
13      Foundation     -0.86 1.20
14    Overall Cond     -0.30 1.17
15      Heating QC      0.18 1.05
y: 125
```

Based on the model for 'Somerst', 'BsmtFin SF 1', 'Fireplaces' are also amongst the features recommended for home improvement.

New Neighborhood - Somerst

	SalePrice	BsmtFin SF 1
0	215,802.12	0.00
1	238,132.40	719.33
2	260,462.68	1,438.67
3	282,792.97	2,158.00

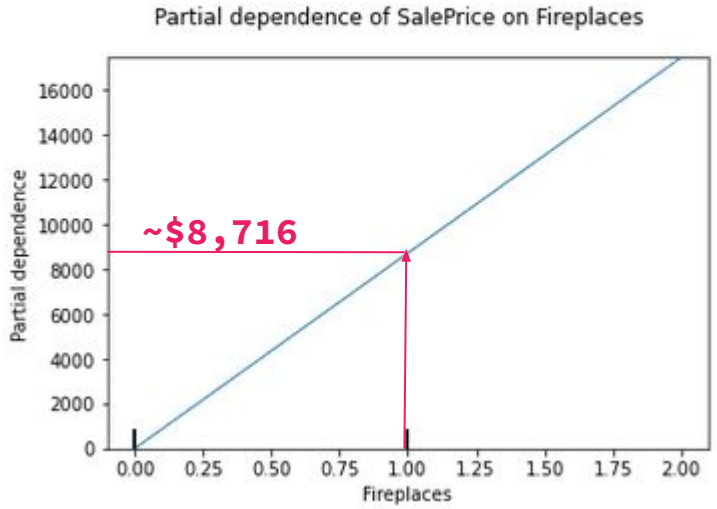


How does ‘BsmtFin SF 1’ impact “SalePrice” in Somerst

1,200 square feet can increase the house value by more than $35,000 (higher than model average)

New Neighborhood - Somerst

	SalePrice	Fireplaces
0	223,031.09	0
1	231,747.93	1
2	240,464.77	2



How does 'Fireplaces' impact "SalePrice" in Somerst

The partial dependence analysis suggest that for every 1 fireplace increase, the sale price will increase by around 8,716 dollars (higher than model average)

Conclusions

- Overall, ‘Bsmt FinS F1’ and ‘Fireplaces’ are recommended features to renovate for homeowners looking to sell their homes, due to their importance in affecting SalePrice and ease of improving these features.

- Home owners in Somerst will have greater value added to their homes if they were to increase fireplaces or quality/size of basement finished area, compared to the average home in Ames, Iowa.

- In addition, home owners in Old Town may want to look into improving their insulation of homes as it seems to have a substantial impact on SalePrice.


Q & A