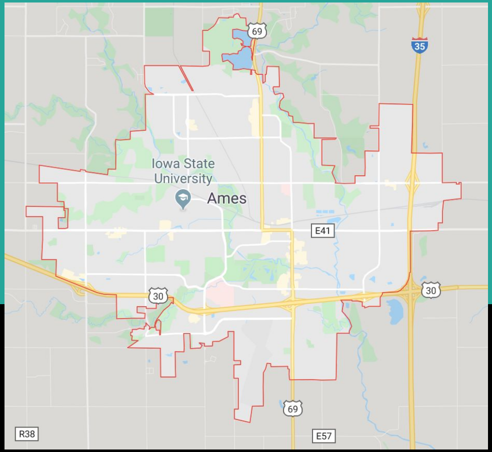



# AMES IOWA HOUSING

## MODELLING FOR PRICE PREDICTION

*Jerome, Aishah, Jia Chi, Munish*

# Business Objective

- Deciding on the price of a house is a very subjective matter, that varies from person to person.

- Our team is tasked to create a model that *objectively* and *reproducibly* predicts the price of housing in Ames, Iowa, USA, from past transaction data obtained in 2006-2010

# Data Set

Descriptive Abstract:

- Data set contains information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010.

Sources:

- Ames, Iowa Assessor's Office

Files:

- train.csv (2051 rows, 81 columns)
- test.csv (879 rows, 80 columns)

# Technical Overview

1. EDA / Data Cleaning → to get a good dataset to work with
2. EDA / Feature Engineering → to make features as balanced as possible for the model to train on
3. Modelling Process
4. Conclusion

# EDA/Cleaning

1. Read and internalize data documentation thoroughly

    Understand the different types of variables available and what they mean

2. EDA & Cleaning (original 2051, 81)
    a. Drop columns with obvious and extensive **(> 300) missing values** (5)
    b. Drop non-numeric columns whose variation of data is minimal (occurence of **SAME category > 85%**) (18)
    c. Drop numeric columns whose variation of data is minimal (occurence of **SAME category > 85%**) (8)
    d. Drop rows with **NA (184)** → (1867, 49)

| Street | Alley | Lot Shape |
|---|---|---|
| Pave | NaN | IR1 |
| Pave | NaN | IR1 |
| Pave | NaN | Reg |
| Pave | NaN | Reg |
| Pave | NaN | IR1 |

```
df['Utilities'].value_counts()
```

```
AllPub    2049
NoSeWa       1
NoSewr       1
```

# EDA/Cleaning

## CONTINUOUS VARIABLES

- drop weakly correlated variables with SalePrice ($< 0.30$)
- drop variables strongly corr with each other ($> 0.70$), keep stronger corr with SalePrice

## Other Variables by Observation / Intuition

| | BsmtFin SF 1 | Bsmt Unf SF | Total Bsmt SF |
|---|---|---|---|
| 0 | 533.0 | 192.0 | 725.0 |

+ =

| | 1st Flr SF | 2nd Flr SF | Gr Liv Area |
|---|---|---|---|
| 0 | 725 | 754 | 1479 |
| 1 | 913 | 1209 | 2122 |
| 2 | 1057 | 0 | 1057 |
| 3 | 744 | 700 | 1444 |
| 4 | 831 | 614 | 1445 |

98.8% Match!

drop!

drop!

`Year Built` vs `Year Remod/Add`

`MS Subclass` similar to `House Style`

drop!

| Corr with SalePrice | |
|---|---|
| Gr Liv Area | 0.707080 |
| Garage Cars | 0.653924 |
| Garage Area | 0.646488 |
| 1st Flr SF | 0.626024 |
| Total Bsmt SF | 0.623269 |
| Full Bath | 0.551443 |
| TotRms AbvGrd | 0.536021 |
| Mas Vnr Area | 0.504794 |
| Fireplaces | 0.444809 |
| BsmtFin SF 1 | 0.401466 |
| Open Porch SF | 0.332883 |
| Wood Deck SF | 0.311658 |
| Lot Area | 0.296810 |
| Bsmt Full Bath | 0.268839 |
| Half Bath | 0.264706 |
| Bedroom AbvGr | 0.139310 |
| Enclosed Porch | -0.128909 |

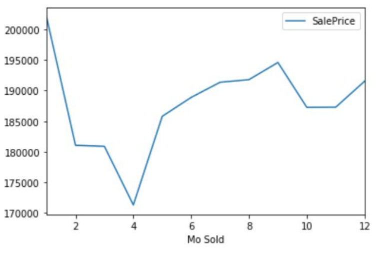
drop!

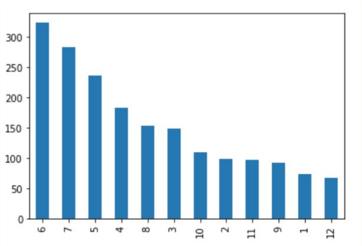
| | SalePrice | Lot Area | Mas Vnr Area | BsmtFin SF 1 | Total Bsmt SF | 1st Flr SF | Gr Liv Area |
|---|---|---|---|---|---|---|---|
| Lot Area | 0.3 | 1 | 0.17 | 0.23 | 0.32 | 0.36 | 0.32 |
| Mas Vnr Area | 0.5 | 0.17 | 1 | 0.31 | 0.41 | 0.4 | 0.4 |
| BsmtFin SF 1 | 0.4 | 0.23 | 0.31 | 1 | 0.53 | 0.48 | 0.22 |
| Total Bsmt SF | 0.62 | 0.32 | 0.41 | 0.53 | 1 | 0.89 | 0.47 |
| 1st Flr SF | 0.63 | 0.36 | 0.4 | 0.48 | 0.89 | 1 | 0.55 |
| Gr Liv Area | 0.71 | 0.32 | 0.4 | 0.22 | 0.47 | 0.55 | 1 |
| Bsmt Full Bath | 0.27 | 0.12 | 0.14 | 0.63 | 0.3 | 0.28 | 0.064 |
| Full Bath | 0.55 | 0.12 | 0.24 | 0.069 | 0.35 | 0.36 | 0.62 |
| Half Bath | 0.26 | 0.046 | 0.19 | -0.029 | -0.11 | -0.14 | 0.44 |
| Bedroom AbvGr | 0.14 | 0.14 | 0.061 | -0.13 | 0.028 | 0.057 | 0.49 |
| TotRms AbvGrd | 0.54 | 0.23 | 0.3 | 0.072 | 0.32 | 0.39 | 0.81 |
| Fireplaces | 0.44 | 0.28 | 0.25 | 0.28 | 0.32 | 0.39 | 0.46 |
| Garage Cars | 0.65 | 0.2 | 0.36 | 0.24 | 0.47 | 0.45 | 0.52 |
| Garage Area | 0.65 | 0.26 | 0.39 | 0.32 | 0.53 | 0.52 | 0.51 |
| Wood Deck SF | 0.31 | 0.15 | 0.15 | 0.2 | 0.21 | 0.23 | 0.27 |
| Open Porch SF | 0.33 | 0.13 | 0.17 | 0.14 | 0.27 | 0.26 | 0.35 |
| Enclosed Porch | -0.13 | 0.011 | -0.1 | -0.073 | -0.078 | -0.034 | -0.033 |

drop!

drop!

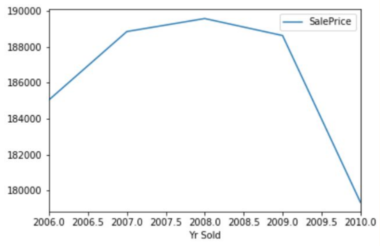
# Data Transformations

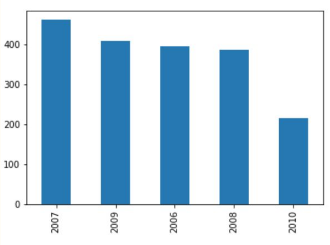


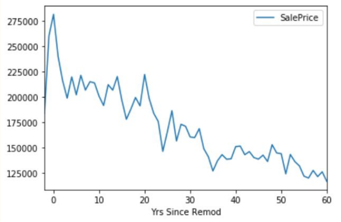
TIME VARIABLES

- Reengineer to continuous with respect to sale price
- Keep ‘mo sold’ as categorical and created dummy
- Keep ‘yr sold’ as continuous

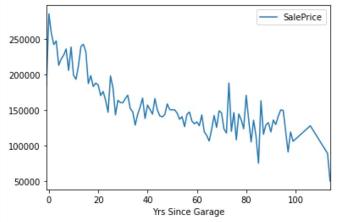


```
def remake_remod(df):
    df = df.copy()
    df['Yrs Since Remod'] = df['Yr Sold'] - df['Year Remod/Add']
```




```
def remake_garageyrblt(df):
    df = df.copy()
    df['Yrs Since Garage'] = df['Yr Sold'] - df['Garage Yr Blt']
```



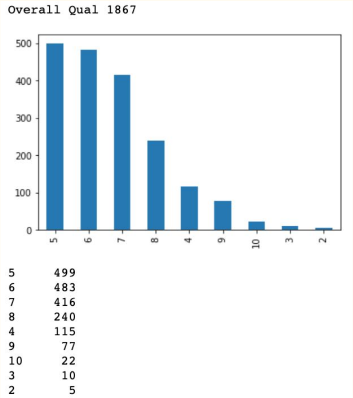
# Data Transformations

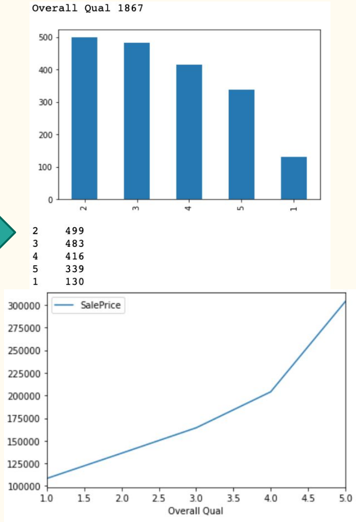
## ORDINAL VARIABLES

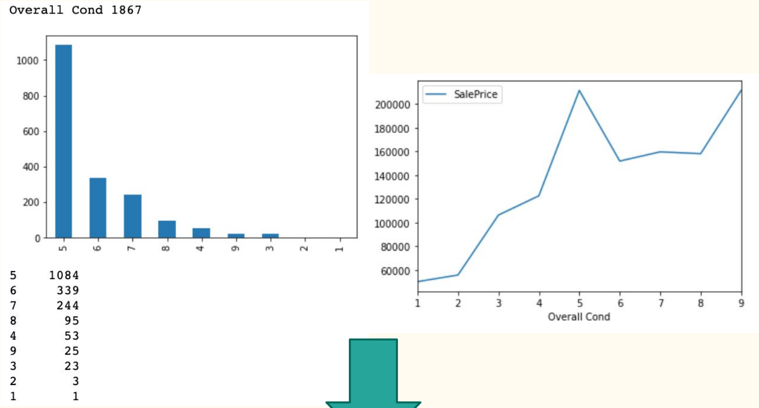
- check relations, re-balance

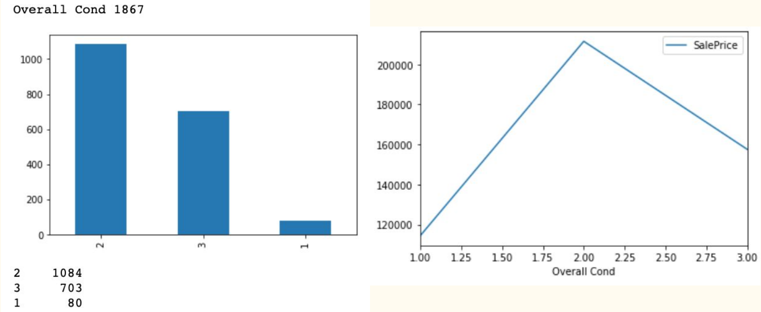
## CATEGORICAL VARIABLES

- balance & encode/create dummies



# EDA/Cleaning/Engg (55) → 19 → 17 (ALL adjR2 = 0.826)

**FINAL VARIABLES**

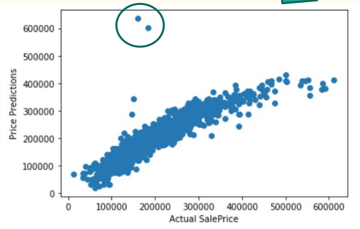
‘Exter Qual’,
‘Bsmt Qual’, ‘Overall Qual’,
‘Overall Cond’, ‘Kitchen Qual’,
‘Gr Liv Area’, *‘Mas Vnr Area’*,
‘Fireplaces’, ‘BsmtFin SF 1’,
‘Bldg Type’, ‘House Style_1+’,
‘Garage Area’, ‘Garage Cars’,
‘Garage Type_BuiltIn’,
‘Mo Sold_7’, ‘Roof Style’,
‘MS Zoning_RM’,

Create ‘Pipeline’ of transformation functions done on train set to apply on test set and modelling later

```
Dep. Variable:               SalePrice   R-squared:                       0.831
Model:                             OLS   Adj. R-squared:                  0.826
Method:                  Least Squares   F-statistic:                     178.7
Date:                 Sun, 24 Nov 2019   Prob (F-statistic):               0.00
Time:                         19:06:56   Log-Likelihood:                -22036.
No. Observations:                 1867   AIC:                         4.417e+04
Df Residuals:                     1816   BIC:                         4.446e+04
Df Model:                           50
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1.871e+05    758.721    246.540      0.000    1.86e+05    1.89e+05
x1           882.4270   1122.607      0.786      0.432   -1319.310    3084.164
x2           714.8367   1008.418      0.709      0.478   -1262.945    2692.619
x3          8477.2801   1248.686      6.789      0.000    6028.268    1.09e+04
x4          1.099e+04   1390.021      7.906      0.000    8263.048    1.37e+04
x5          1.116e+04   1324.777      8.427      0.000    8565.691    1.38e+04
x6          1.194e+04   1368.924      8.724      0.000    9257.696    1.46e+04
x7          4409.4884    895.571      4.924      0.000    2653.032    6165.945
x8         -1176.6176    786.640     -1.496      0.135   -2719.432     366.197
x9         -2075.1727   1283.525     -1.617      0.106   -4592.514     442.168
x10         3328.4664   1521.151      2.188      0.029     345.077    6311.856
x11         2.008e+04   2227.182      9.017      0.000    1.57e+04    2.45e+04
x12         6475.7612   1658.710      3.904      0.000    3222.581    9728.942
x13         5116.3030   1664.231      3.074      0.002    1852.295    8380.311
x14          620.6696   2419.764      0.257      0.798   -4125.145    5366.484
x15         2360.7268   1969.375      1.199      0.231   -1501.751    6223.205
x16         -401.0821   1171.985     -0.342      0.732   -2699.663    1897.499
x17         2527.2501   1412.382      1.789      0.074    -242.814    5297.315
x18         6234.0027    923.749      6.749      0.000    4422.280    8045.725
x19         4598.0525    942.604      4.878      0.000    2749.350    6446.755
x20         7200.4488    985.256      7.308      0.000    5268.094    9132.804
x21         -414.4853    860.498     -0.482      0.630   -2102.155    1273.185
x22         1512.0740    833.269      1.815      0.070    -122.192    3146.340
x23          558.6934    800.318      0.698      0.485   -1010.948    2128.335
x24         5672.6902    887.261      6.393      0.000    3932.531    7412.849
x25        -3047.5508    843.930     -3.611      0.000   -4702.726   -1392.376
x26        -1310.5248    849.293     -1.543      0.123   -2976.219     355.169
x27         -987.3533    768.021     -1.286      0.199   -2493.650     518.944
x28         1179.9300    741.110      1.592      0.112    -273.588    2633.448
x29          103.5481    763.603      0.136      0.892   -1394.085    1601.181
x30          621.7910    499.217      1.246      0.213    -357.308    1600.890
x31        -1249.1237    601.015     -2.078      0.038   -2427.877     -70.370
x32         2801.3284    741.198      3.779      0.000    1347.638    4255.019
x33        -2417.9927    889.861     -2.717      0.007   -4163.252    -672.734
x34         -942.0044    772.201     -1.220      0.223   -2456.501     572.492
x35        -1334.6303    840.175     -1.589      0.112   -2982.441     313.180
x36          -28.6085    642.508     -0.045      0.964   -1288.740    1231.523
x37          860.8367    751.639      1.145      0.252    -613.332    2335.005
x38        -1305.6603    795.929     -1.640      0.101   -2866.693     255.372
x39         -163.3471    552.263     -0.296      0.767   -1246.485     919.791
x40         -203.3982    786.853     -0.258      0.796   -1746.630    1339.834
x41         2132.4565    807.373      2.641      0.008     548.979    3715.934
x42          -36.1087    777.124     -0.046      0.963   -1560.260    1488.042
x43         -684.3806    687.691     -0.995      0.320   -2033.129     664.368
x44         -303.6845    735.584     -0.413      0.680   -1746.363    1138.994
x45         -753.5535    723.084     -1.042      0.297   -2171.718     664.611
x46          136.8354    702.115      0.195      0.846   -1240.202    1513.873
x47         -446.6520    698.567     -0.639      0.523   -1816.732     923.428
x48          360.7650    670.926      0.538      0.591    -955.102    1676.632
x49         -224.2597    649.118     -0.345      0.730   -1497.356    1048.837
x50         1635.8042    661.156      2.474      0.013     339.099    2932.510
x51          217.3407    701.802      0.310      0.757   -1159.083    1593.764
x52         -236.0728    732.086     -0.322      0.747   -1671.891    1199.746
x53        -1117.3720    722.050     -1.548      0.122   -2533.507     298.763
x54         -277.0694    725.779     -0.382      0.703   -1700.518    1146.379
x55           95.3494    737.269      0.129      0.897   -1350.635    1541.334
==============================================================================
Omnibus:                     1115.307   Durbin-Watson:                   1.968
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           157635.670
Skew:                          -1.799   Prob(JB):                         0.00
Kurtosis:                      47.871   Cond. No.                     1.17e+16
==============================================================================
```

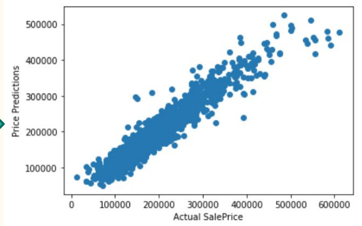
```
                            OLS Regression Results
==============================================================================
Dep. Variable:               SalePrice   R-squared:                       0.828
Model:                             OLS   Adj. R-squared:                  0.826
Method:                  Least Squares   F-statistic:                     522.6
Date:                 Sun, 24 Nov 2019   Prob (F-statistic):               0.00
Time:                         19:06:57   Log-Likelihood:                -22054.
No. Observations:                 1867   AIC:                         4.414e+04
Df Residuals:                     1849   BIC:                         4.424e+04
Df Model:                           17
Covariance Type:             nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                1.871e+05    759.420    246.313      0.000    1.86e+05    1.89e+05
Kitchen Qual         9044.5450   1193.867      7.576      0.000    6703.076    1.14e+04
Exter Qual           1.149e+04   1328.505      8.650      0.000    8886.697    1.41e+04
Bsmt Qual            1.198e+04   1194.701     10.029      0.000    9638.581    1.43e+04
Overall Qual         1.284e+04   1295.319      9.909      0.000    1.03e+04    1.54e+04
Overall Cond         4522.7676    834.677      5.419      0.000    2885.759    6159.776
Gr Liv Area          2.321e+04   1149.860     20.186      0.000     2.1e+04    2.55e+04
Garage Cars          6959.8478   1578.866      4.408      0.000    3863.301    1.01e+04
Mas Vnr Area         6440.0474    899.379      7.161      0.000    4676.143    8203.952
Fireplaces           4993.3644    902.456      5.533      0.000    3223.424    6763.305
BsmtFin SF 1         8183.5638    889.090      9.204      0.000    6439.838    9927.289
Bldg Type            6073.6798    821.518      7.393      0.000    4462.479    7684.880
Roof Style          -3364.7259    814.691     -4.130      0.000   -4962.536   -1766.915
House Style_1+       6018.0506    901.243      6.678      0.000    4250.491    7785.610
Garage Area          4167.6170   1550.527      2.688      0.007    1126.648    7208.586
Garage Type_BuiltIn  2123.1981    834.341      2.545      0.011     486.848    3759.548
Mo Sold_7            1926.8344    763.615      2.523      0.012     429.195    3424.474
MS Zoning_RM        -2615.9747    823.211     -3.178      0.002   -4230.496   -1001.453
==============================================================================
Omnibus:                     1050.173   Durbin-Watson:                   1.964
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           134884.000
Skew:                          -1.644   Prob(JB):                         0.00
Kurtosis:                      44.510   Cond. No.                         6.39
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

**KEEP**



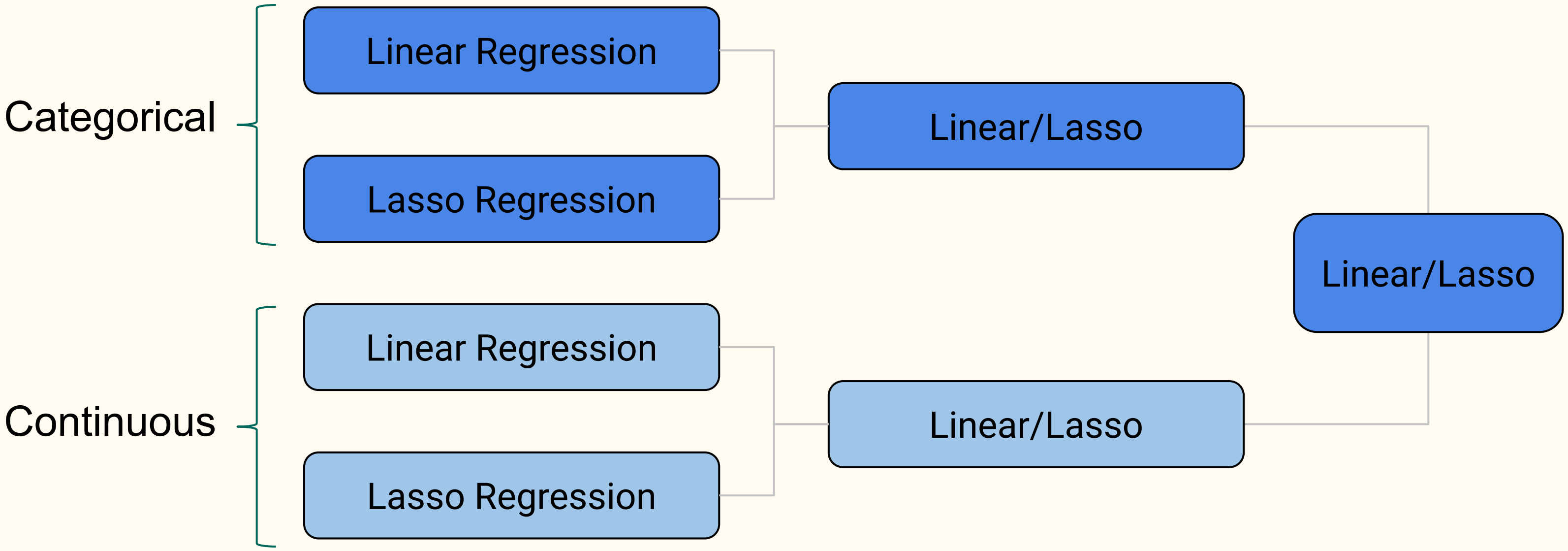

R2 = 0.826

poly




R2 = 0.863

# Modelling Process:



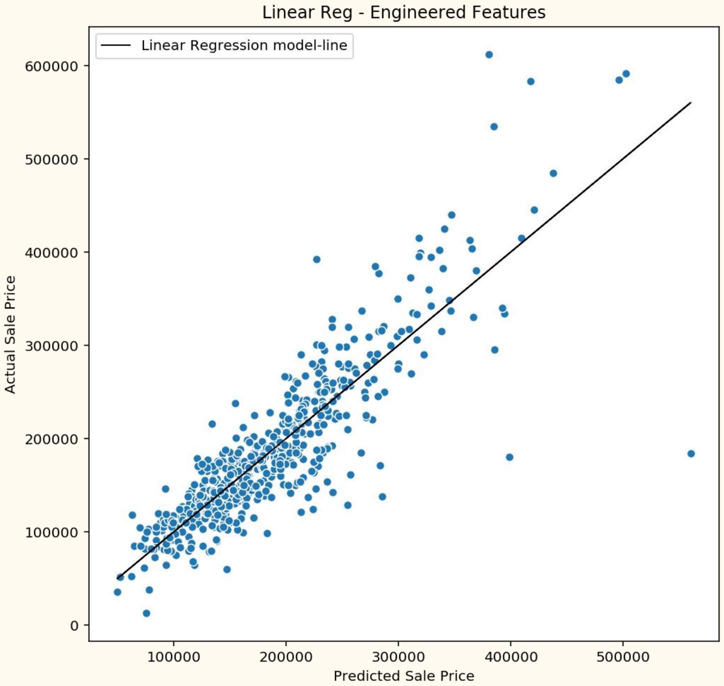

**Stage 1:**
- Train-test-split
- Feat. Eng
- Dum.Var

**Stage 2:**
- Cross-val
- R2-score

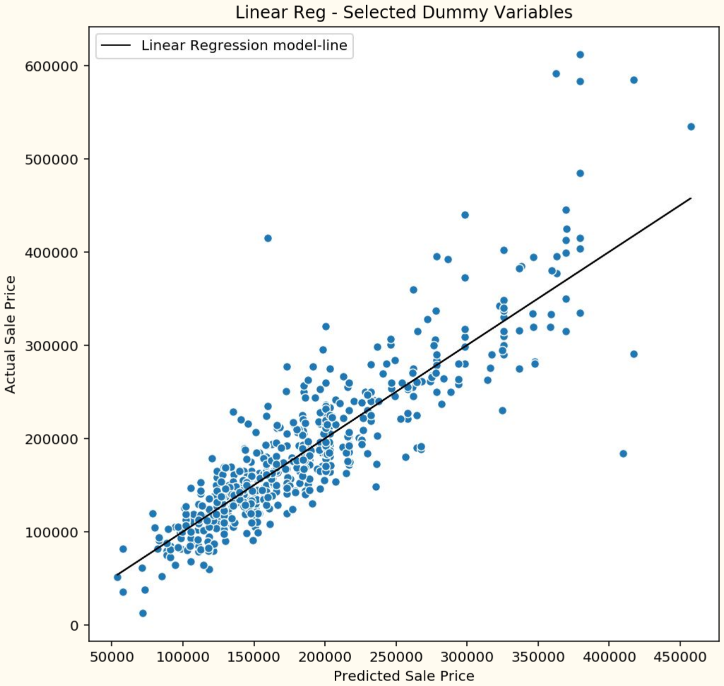
**Stage 3:**
- Combine Feat
- R2-score

**Stage 4:**
- Final
- Kaggle Test

# Model Evaluation (Individual)

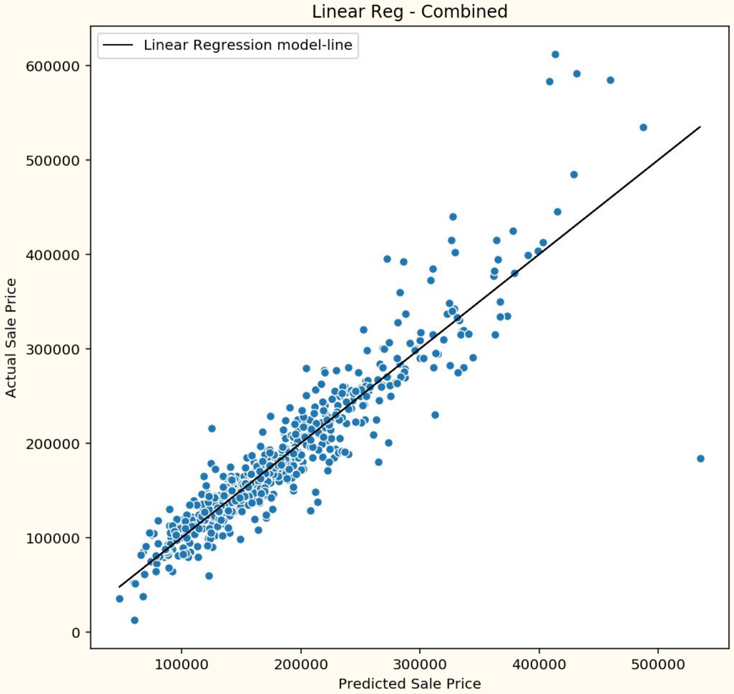



**Continuous Features**
R2-score: 0.687
RMSE: 39647.45




**Categorical Features**
R2-score: 0.698
RMSE: 37771.02

# Model Evaluation (Combined)




**Combined Features**
R2-score: 0.824
RMSE: 30864.89

| | Predictors | Coeffs. |
|---|---|---|
| 47 | Overall Qual_10 | 131754.019070 |
| 29 | Neighborhood_GrnHill | 102370.455690 |
| 55 | Overall Qual_9 | 92271.974348 |
| 54 | Overall Qual_8 | 70048.499174 |
| 53 | Overall Qual_7 | 53156.259879 |
| 52 | Overall Qual_6 | 41775.848841 |
| 51 | Overall Qual_5 | 32727.480697 |
| 44 | Neighborhood_StoneBr | 30685.501542 |
| 23 | Neighborhood_ClearCr | 27874.277119 |
| 50 | Overall Qual_4 | 26069.262578 |
| 3 | Gr Liv Area | 22631.793502 |
| 25 | Neighborhood_Crawfor | 20651.731751 |
| 49 | Overall Qual_3 | 20641.100360 |
| 37 | Neighborhood_NoRidge | 20020.889372 |

**Top 15 positive predictors**

| | Predictors | Coeffs. |
|---|---|---|
| 28 | Neighborhood_Greens | -14513.151549 |
| 60 | Bsmt Qual_Gd | -14693.750663 |
| 35 | Neighborhood_NPkVill | -15319.336894 |
| 57 | Exter Qual_Gd | -15925.999486 |
| 30 | Neighborhood_IDOTRR | -16326.259845 |
| 39 | Neighborhood_OldTown | -16601.759971 |
| 64 | Bsmt Qual_UnKn | -16992.136110 |
| 59 | Bsmt Qual_Fa | -19490.351089 |
| 58 | Exter Qual_TA | -21061.428090 |
| 63 | Bsmt Qual_TA | -21206.505257 |
| 32 | Neighborhood_MeadowV | -22788.035580 |
| 66 | Kitchen Qual_Gd | -26383.688639 |
| 56 | Exter Qual_Fa | -32661.475977 |
| 67 | Kitchen Qual_TA | -35635.269316 |
| 65 | Kitchen Qual_Fa | -41310.676463 |

**Top 15 negative predictors**

# Test data preprocessing & Kaggle Submission

**Kaggle Test Data**

1. Linear Reg
2. Lasso Reg

**Cleaning**
**Pre-processing**
**Prediction**

## Kaggle scores:

1. Linear Regression (RMSE)
   - Private: 30262.80
   - Public: 28869.88

2. Lasso Regression (RMSE)
   - Private: 31505.63
   - Public: 30562.34

# Conclusion

1. The model is sufficiently robust to predict housing prices in Ames, IA.
   - Area of the house (sq-feet)
   - Condition of the house
   - Neighbourhood

   “general” features

2. Ability for the model to be applied in future housing predictions.
   - Property evaluation
   - Housing development projects

# Future Recommendations

- Train on 2006-2009 data, test on 2010 data
- Income bracket of the person buying/ selling
- Crime rate of area
- Distance to offices/ amenities around the neighbourhood

# Thanks!