Project 2

Predicting home sale prices in Ames, Iowa

Shauna, Kar how, Willy, Ikhwan

Business Objective

We are property agents from an established real estate firm. We hope to provide value add to our clients by providing **reliable** and **accurate** insights into the **house prices** of Ames, Iowa, through our **prediction** service.

Modelling Approach

- Data cleansing
- Exploratory Data Analysis
- Feature Engineering (Interaction Terms)
- Feature Selection
- Linear Modelling
- Regularization
- Prediction Outcome

Data Cleansing

1. Drop non-features e.g. PID, ID
2. Drop high number of null values
3. Drop high number of zeros
4. Data Types
5. Invalid categorical values
6. Develop utility functions to cross-reference categorical values to data dictionary to highlight problems.

Data Cleansing

Columns 5, 56, 71, 72, 73 have too many **null / zero** values so we drop features:

Alley, Fireplace Qu, Pool Qc,Fence, Misc Feature

```
2: 16.089712335446123%
5: 93.1740614334471%
24: 1.0726474890297415%
24: 1.0726474890297415%
29: 2.681618722574354%
29: 2.681618722574354%
31: 2.827888347147735%
29: 2.681618722574354%
33: 0.04875670404680644%
34: 2.730375426621160 6%
33: 0.04875670404680644%
33: 0.04875670404680644%
33: 0.04875670404680644%
46: 0.09751340809361288%
46: 0.09751340809361288%
56: 48.7567040468064 3%
57: 5.509507557289127%
58: 5.558264261335934%
58: 5.558264261335934%
33: 0.04875670404680644%
33: 0.04875670404680644%
58: 5.558264261335934%
58: 5.558264261335934%
71: 99.5611896635787 4%
72: 80.49731838127742%
73: 96.8308142369575 8%
```

Data Types

1. Categorical features with numeric values e.g. Mo Sold
2. Yr Sold, Year Built represent years, numeric but more categorial.
3. Can be converted to more meaningful data ***i.e. Yr Sold - Year Built = Age @ time of sales.***
4. Discrete numeric columns with float data type e.g. Bsmt Full Bath

Invalid Categorical Values

Reference data documentation

http://jse.amstat.org/v19n3/decock/DataDocumentation.txt

Whitespaces

e.g. Sale Type: 'WD '

Spelling Mistakes

e.g. CmentBd: CemntBd

Invalid values

e.g. MS Zoning: A (agr), C (all), I (all)

External resources

i.e. Google Maps, Ames City Map
e.g. Neighborhood: Greens = Greeensborro, Sommerset district (Somerst)

Utility Functions

```
valid_values = {
    'MS SubClass': [20, 30, 40, 45, 50, 60, 70, 75, 80, 85, 90, 120, 150, 160, 180, 190],
    'MS Zoning': ['RL', 'RM', 'FV', 'C', 'RH', 'A', 'I', 'RP'],
    'Street': ['Pave', 'Grvl'],
    'Alley': ['Pave', 'Grvl', 'NA']...
```

```
# Iterate through all categorical features
and check for invalid values
for index, error in
enumerate(check_invalid_values(df,
valid_values), 1):
    print( str(index) + ': ' + error)
```

1: Invalid Bldg Type value: Twnhs
2: Invalid Exterior 2nd value: Brk Cmn
3: Invalid Exterior 2nd value: CmentBd
4: Invalid Exterior 2nd value: Wd Shng
5: Invalid MS Zoning value: A (agr)
6: Invalid MS Zoning value: C (all)

EDA & Feature Selection

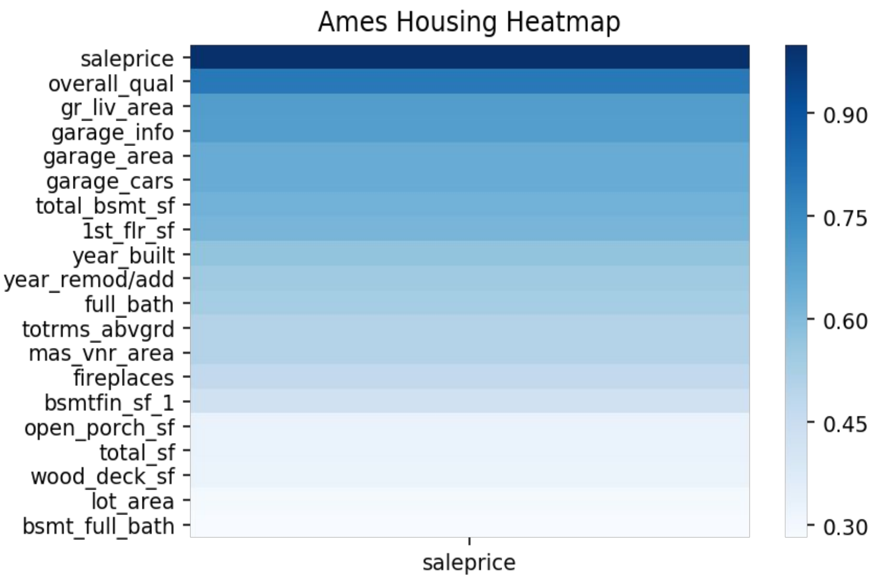
Categorical -> Chi square test (to remove collinearity), followed by the creation of dummy variables

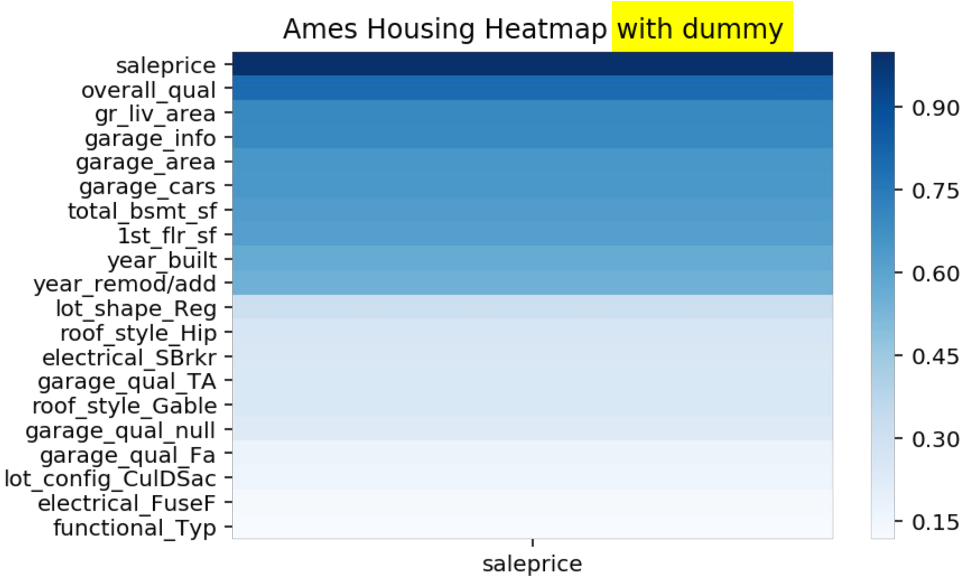
Numeric -> correlation plot

Feature matrix -> categorical +numeric

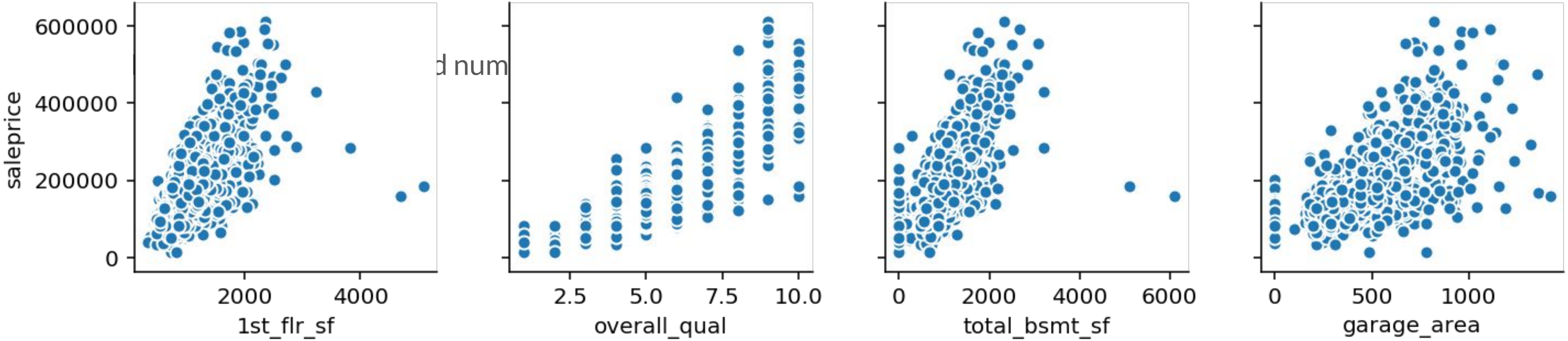
Target vector-> sale price

EDA & Feature Selection

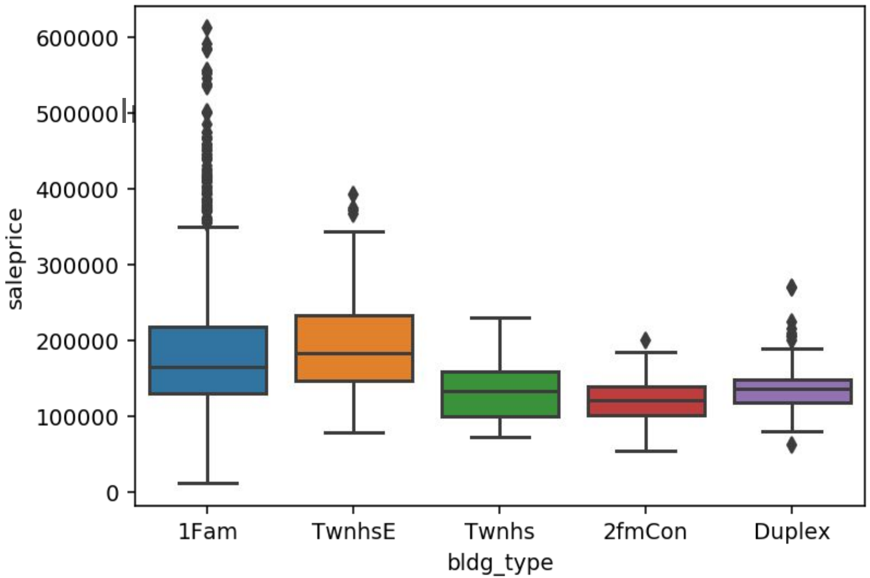


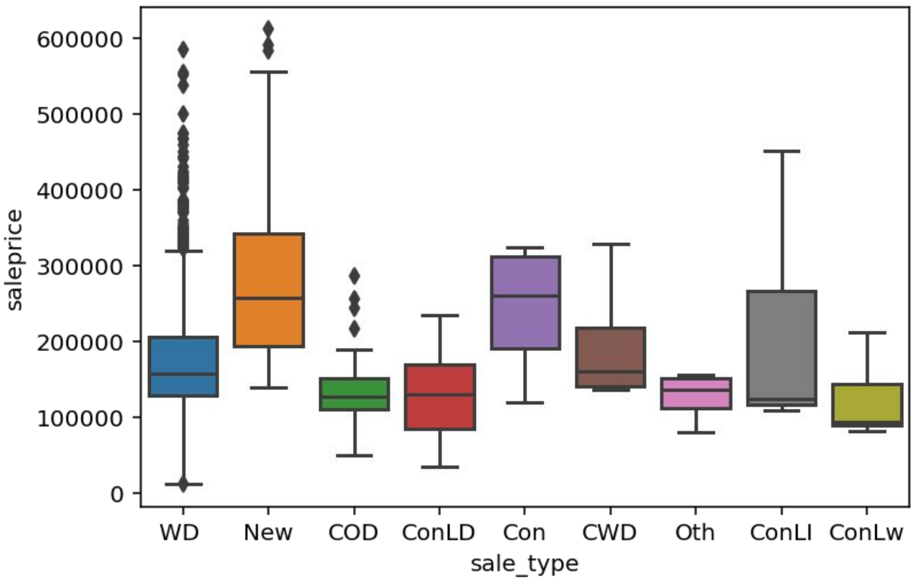


EDA & Feature Selection



EDA & Feature Selection





Feature Engineering

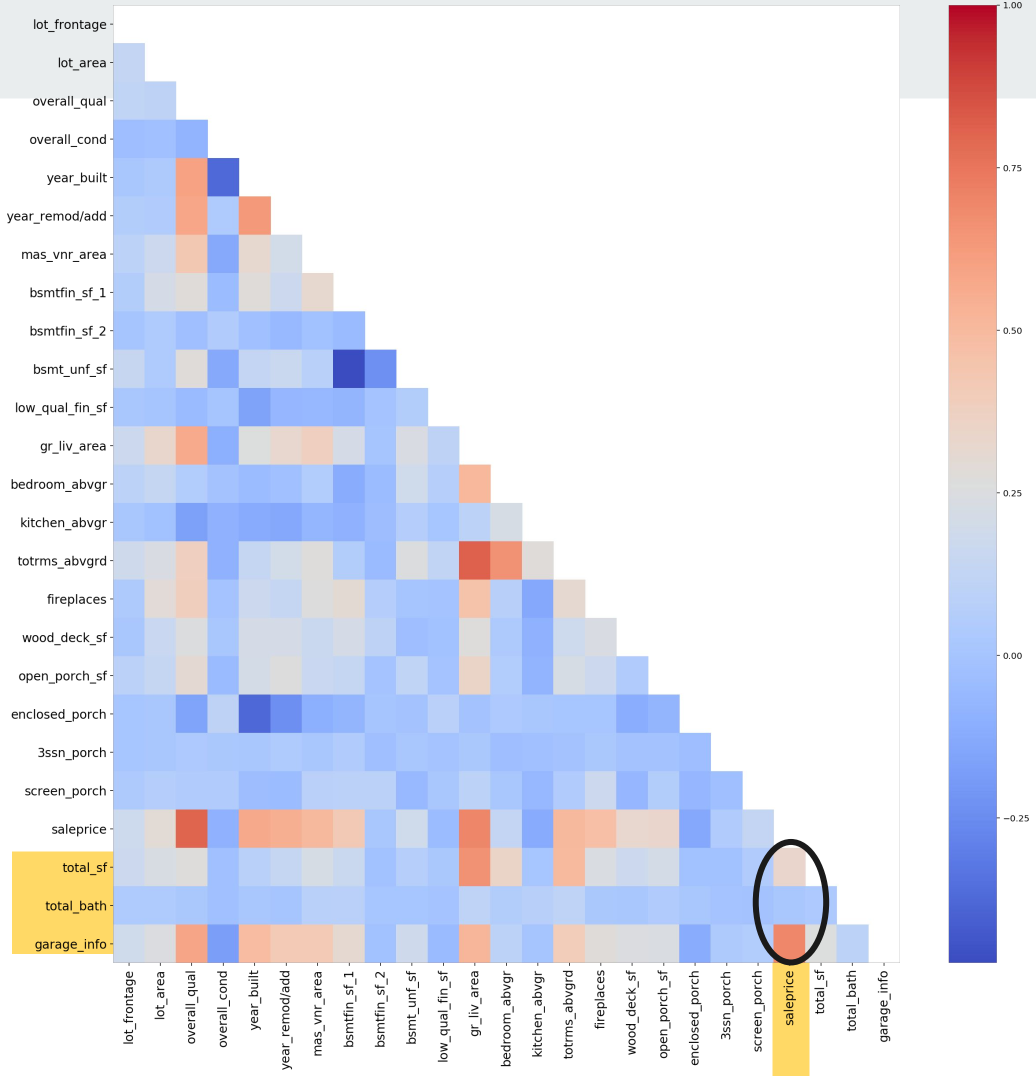
Interaction Terms

Total square foot -> total basement square foot, 1st floor square foot, 2nd floor square foot

Total bath -> basement full bath, basement half bath, full bath, half bath

Garage Info -> garage year built, garage cars, garage area

Useful Interaction Terms



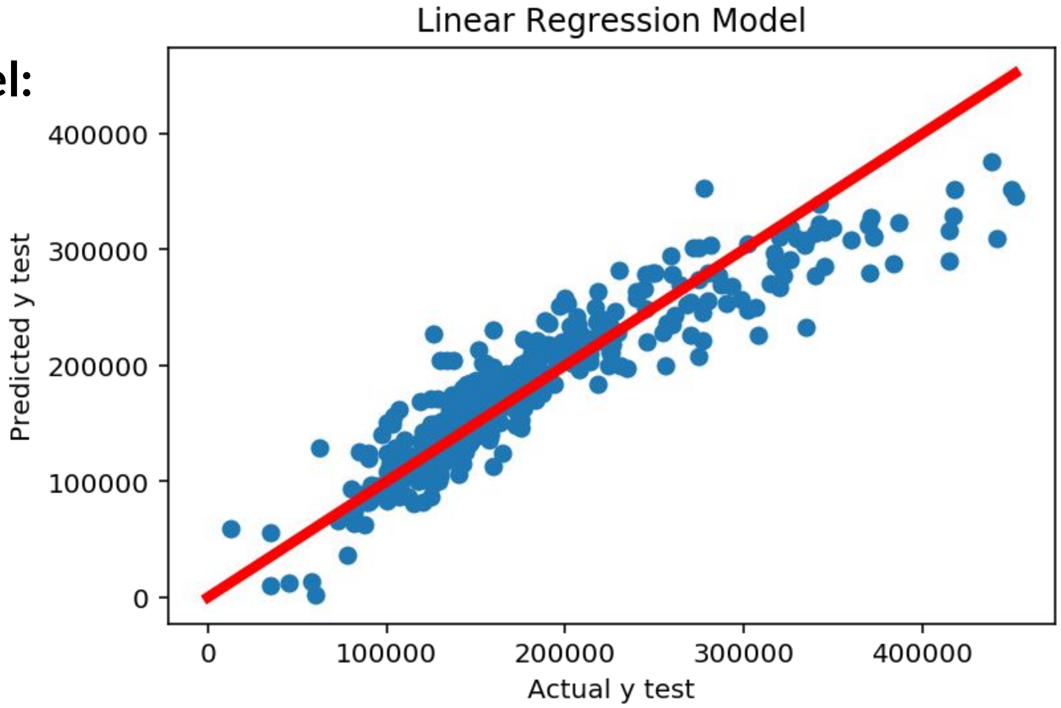
Feature	Correlation Coef
Total Bath	0.0163
Full Bath	0.538
Half Bath	0.283
Basement full bath	0.283
Basement half bath	0.0453
Total Sq Ft	0.333
Basement Sq Ft	0.191
1st Floor Sq Ft	0.618
2nd Floor Sq Ft	0.248
Garage Info	0.695
Garage Year Built	0.259
Garage Cars	0.648
Garage Area	0.650

Regression Model

Baseline Linear Regression Model:
Mean CV score: 0.805

RidgeCV: after scaling data
Mean CV score (Train/Test) : 0.782/0.811

LassoCV: after scaling data
Mean CV Score (Train/Test) : 0.783/0.815



Prediction Outcome

	variable	coef	abs_coef
0	Overall Qual	20597.732368	20597.732368
1	Gr Liv Area	14920.352271	14920.352271
3	Total Bsmt SF	10385.122935	10385.122935
7	Fireplaces	9925.988237	9925.988237
5	Year Remod/Add	8765.427194	8765.427194
4	Year Built	8116.878973	8116.878973
2	Garage Area	7518.689414	7518.689414
6	Mas Vnr Area	3323.082159	3323.082159

CONCLUSION

- With features and coefficients known:
    - We can predict price of House
    - determine which factors affect price most

-Given a particular price , what features of a house can be built to meet that price.

Thank you!

☺