AMES HOUSING PROJECT

DSI-09-Samrin

Introduction

- *The project consists of a database with various features that affect the value of the property prices.*
- *Historical data on the current pricing is available in the data set.*

Objective

- The main objective of the project is to develop a model that can provide the use with a estimate of the price taking the features and attributes into consideration.
- The model should work with a performance score that best represents the data in a clear and concise manner.

DATA CLEANING:

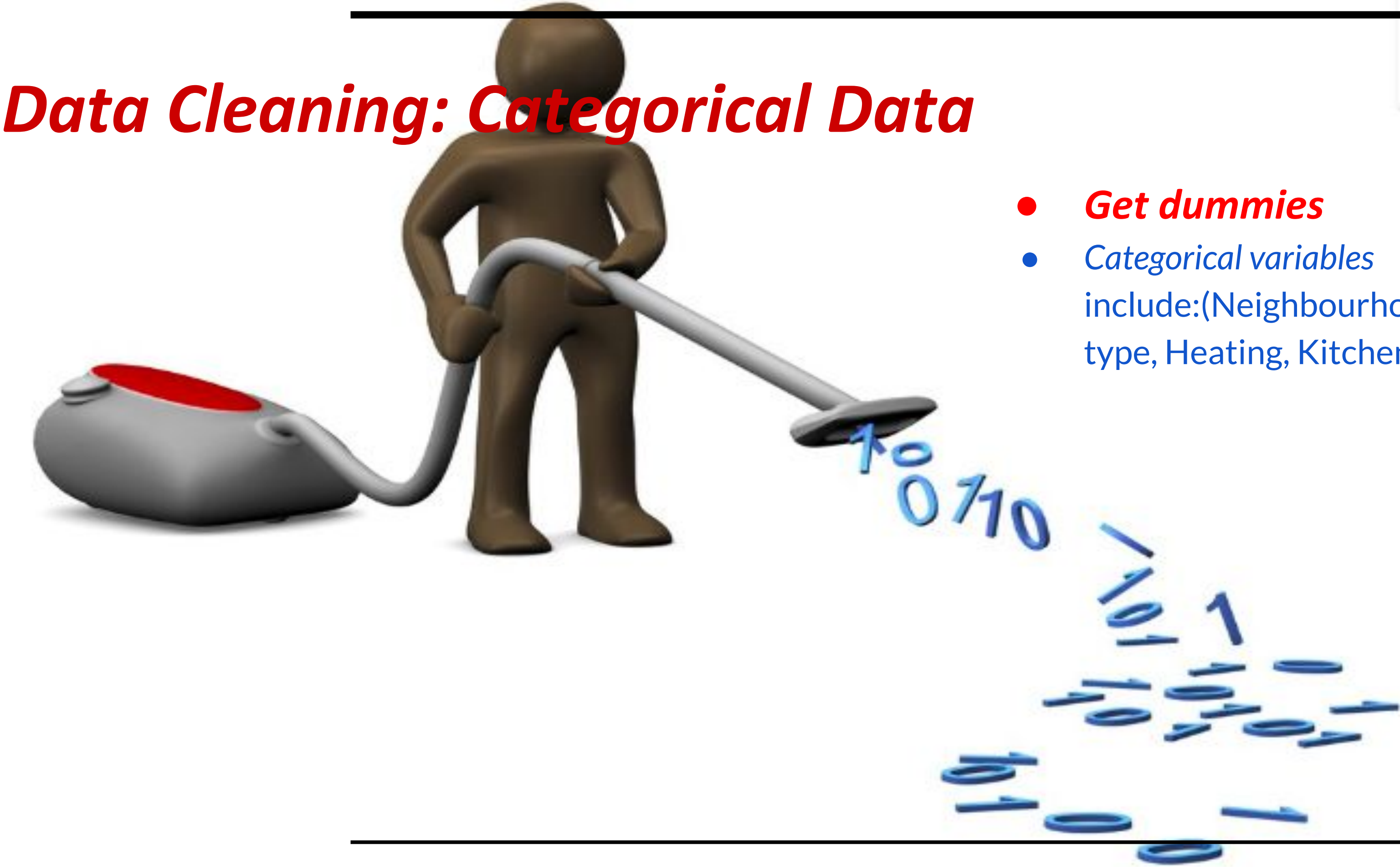
Numeric Data: Null values:drop, fill with 0,or fill with mean???

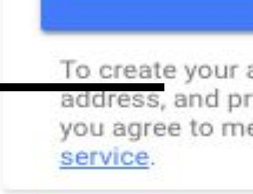
- Drop rows and columns with null values :
- Fill with 0:
- Fit with mean of columns:

(coefficient of determination R-squared) 0.9054337948730329 , Kaggle Score(RMSE:)25757.46358168358Lasso CV model score(train/test) 0.9097482576807232 , 0.9008261948983242

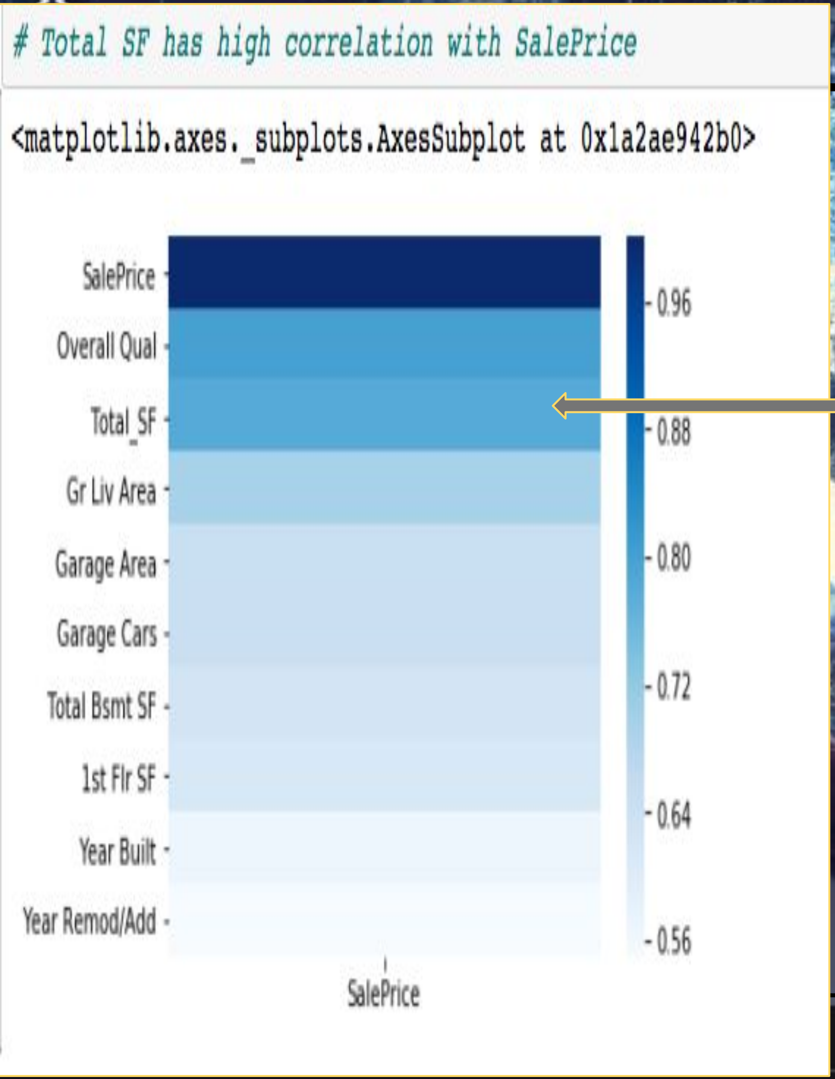
Data Cleaning: Categorical Data

- ***Get dummies***
- *Categorical variables* include:(Neighbourhood, Bldg type, Heating, Kitchen Qual)





To create your a
address, and pr
you agree to me
service.



```
# Total SF has high correlation with SalePrice
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a2ae942b0>
```

Data   Engineering

TotalSF area interaction term:
TotalSF = 'Total Bsmt SF' + '1st Flr SF' + '2nd Flr SF'

#note: data also includes Gr Liv Area (Continuous): Above grade (ground) living area square feet

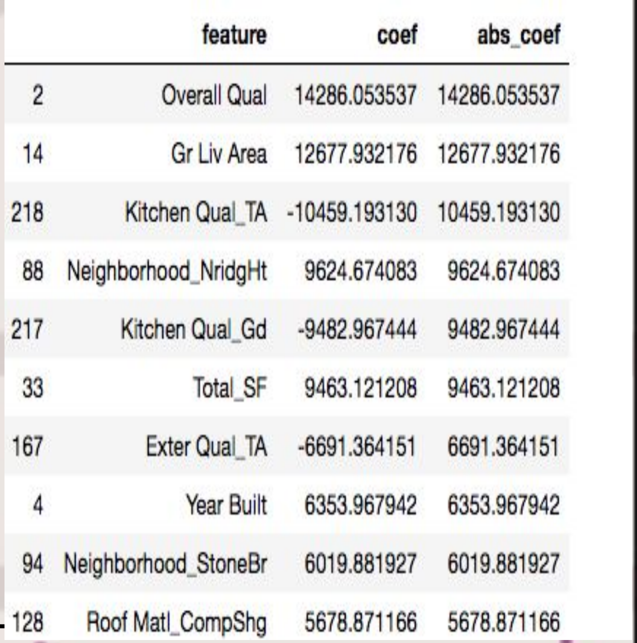
Feature Selection

Lasso Highest co-efficient:

# note: data also includes Gr Liv Area

(Continuous): Above grade (ground)

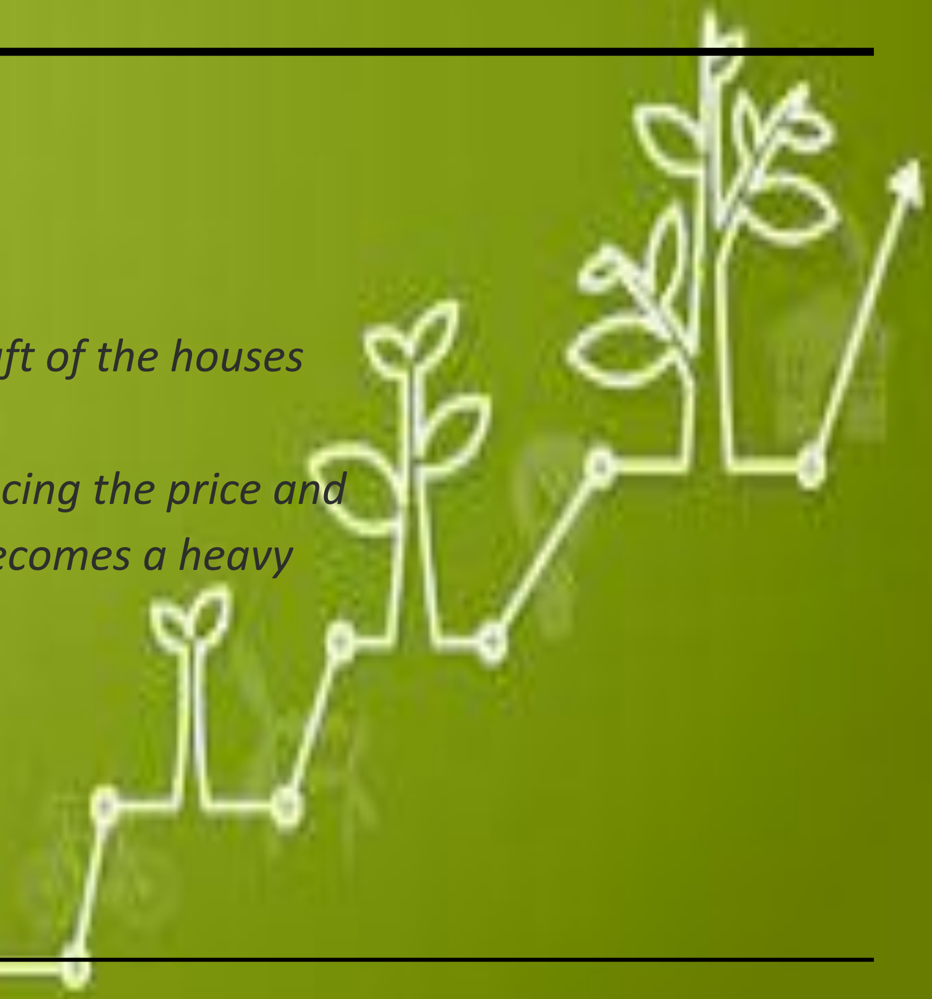
Living area square feet



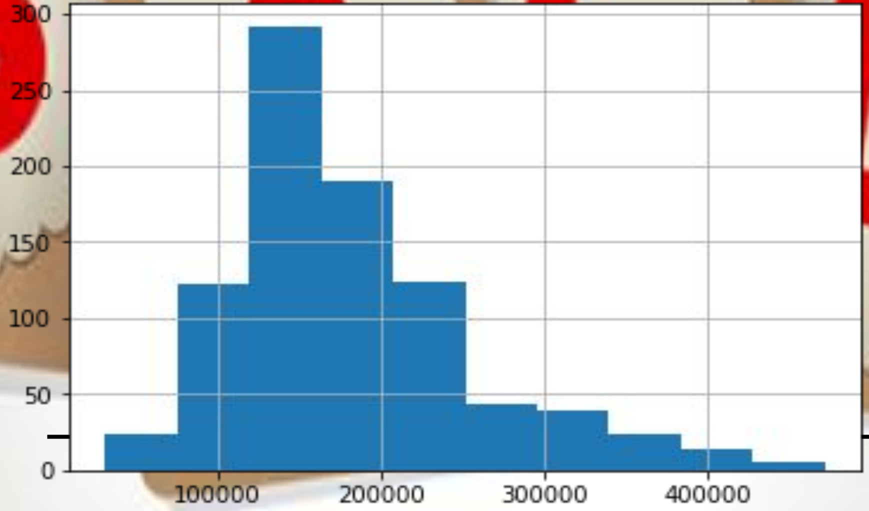
	feature	coef	abs_coef
2	Overall Qual	14286.053537	14286.053537
14	Gr Liv Area	12677.932176	12677.932176
218	Kitchen Qual_TA	-10459.193130	10459.193130
88	Neighborhood_NridgHt	9624.674083	9624.674083
217	Kitchen Qual_Gd	-9482.967444	9482.967444
33	Total_SF	9463.121208	9463.121208
167	Exter Qual_TA	-6691.364151	6691.364151
4	Year Built	6353.967942	6353.967942
94	Neighborhood_StoneBr	6019.881927	6019.881927
128	Roof Matl_CompShg	5678.871166	5678.871166

Relation:

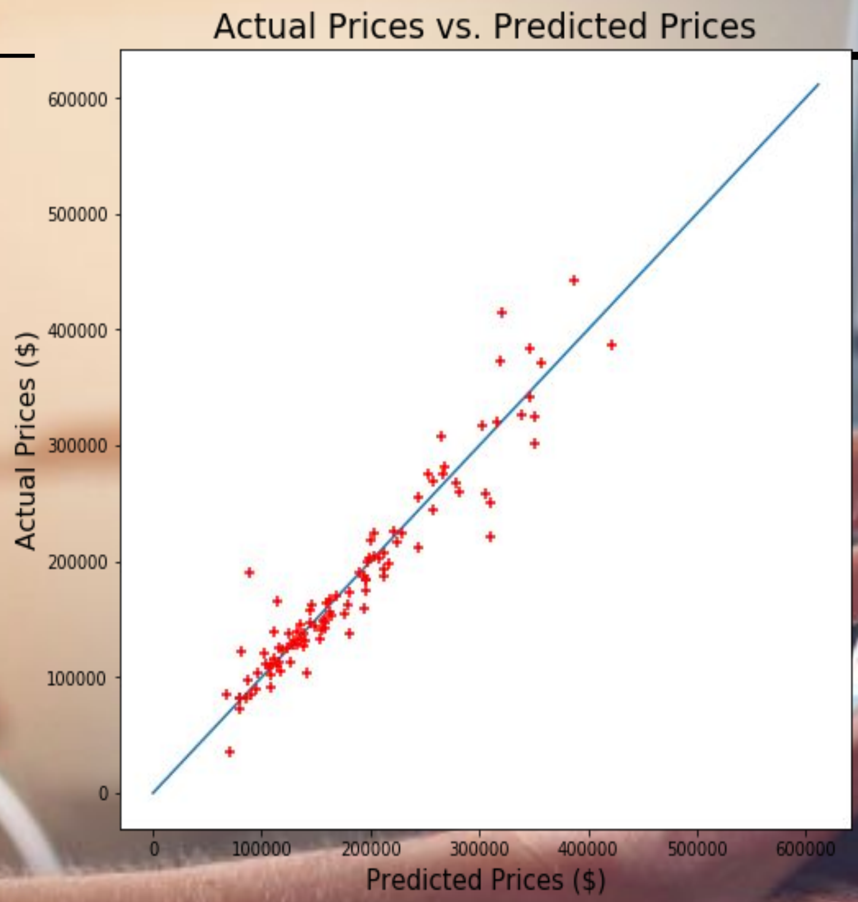
- *Apart from the Quality, the combined sqft of the houses was the second biggest factor.*
- *Location also played a big role in influencing the price and when location combined with pricing, becomes a heavy weighted influencer*



Sale Price Spread



Actual Vs Predicted Prices



920 × 1280

Recommendations:

- Calibration of models regularly to reflect real time change.
- Better interaction of data and inclusion of outliers in a fashion that reflects proper collection of data.
- Optimization can be done through using various modeling techniques and engineering features that can predict prices in a reasonable margin