Ames Housing Data Kaggle Challenge

Neil Yap

General Assembly Data Science Immersive, June 2021

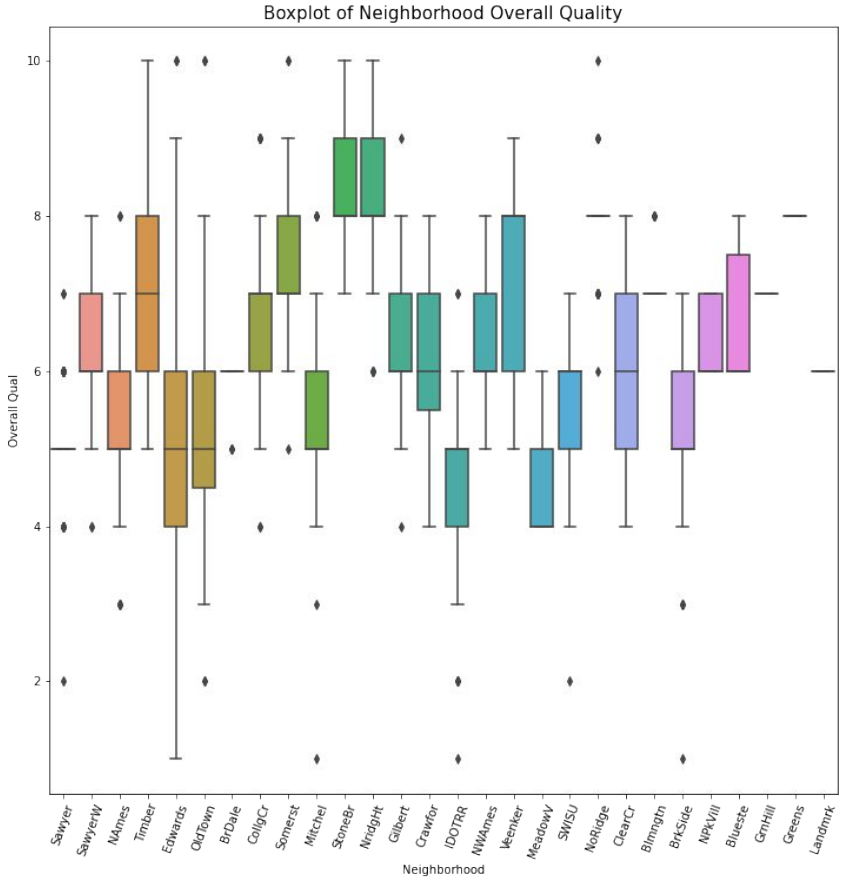
Problem

- Predicting house prices based on data such as
  - Square feet area of each floor, garage, basement, pool
  - Quality of finishing
  - Year built or renovated
  - Amount of rooms, bathrooms, fireplaces
  - Neighborhood
  - Physical lot features such as contour, shape and foundation material
  - Many other features

Feature Selection

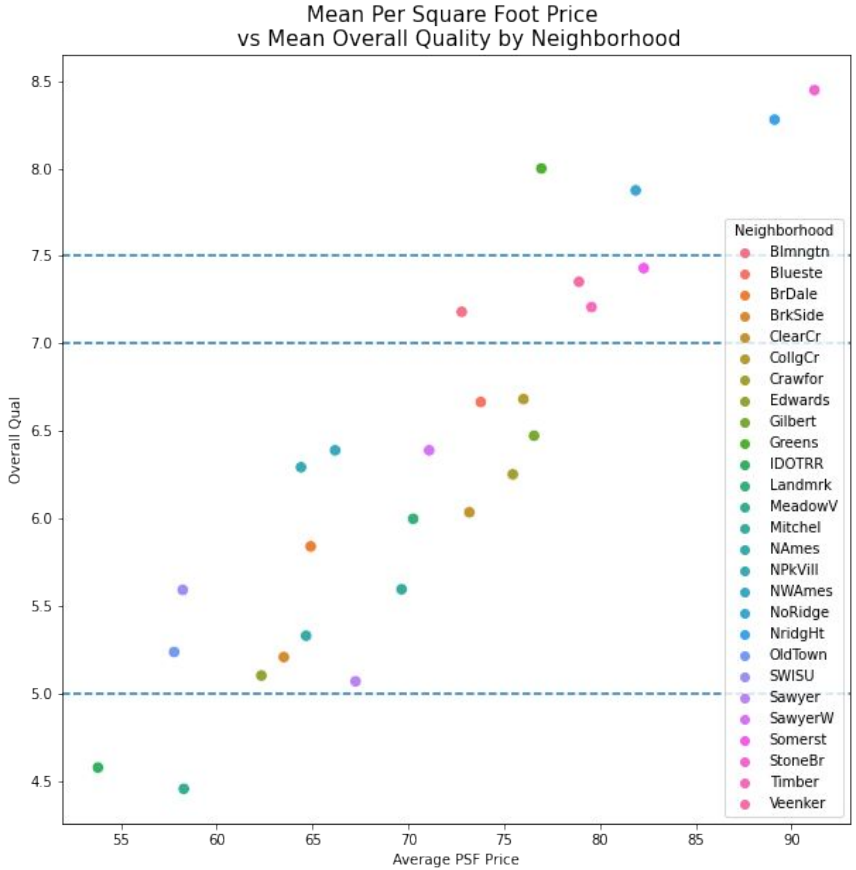
Segmenting Neighborhoods

- Boxplot of neighbourhoods based on overall quality of finishing
- Can roughly group them, with the two highest boxes being the most obvious
- Should also take into account sales price per square foot to determine class of neighbourhood

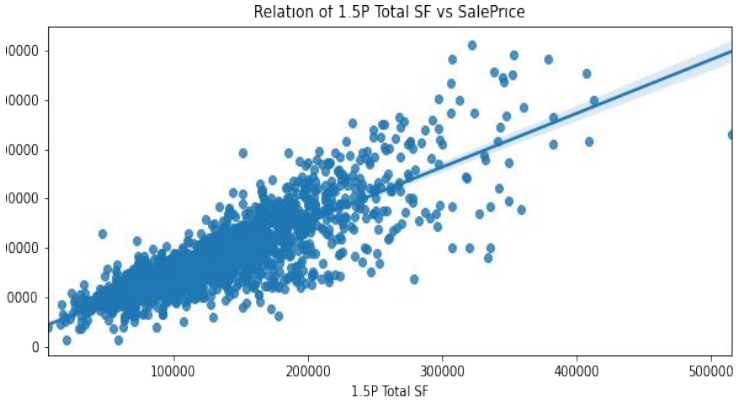


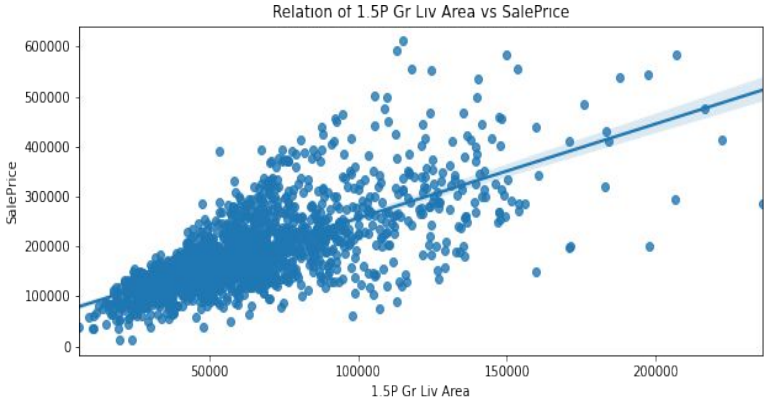
Segmenting Neighborhoods

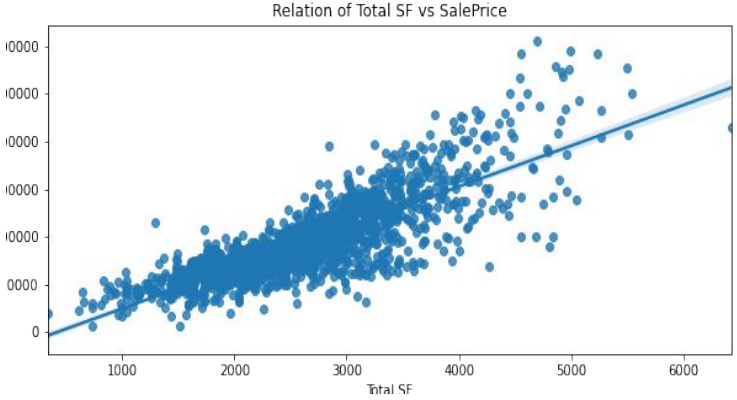
- Segmenting based average overall finishing quality and on average price per square foot
- Dotted lines represent boundary segmenting neighbourhoods
- 4 segments with GreenHill (outlier) in own segment

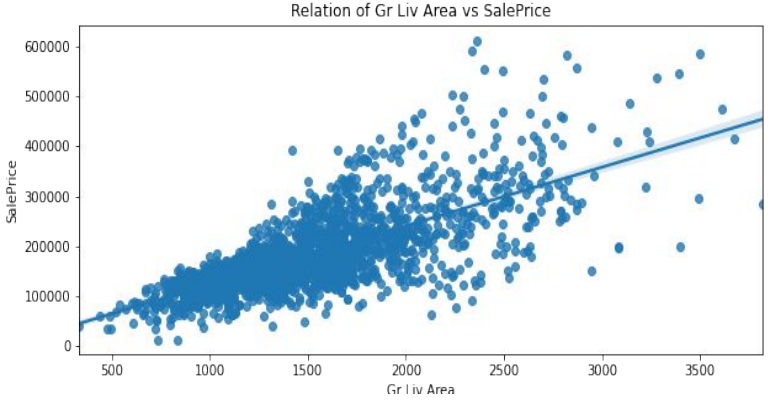


Feature Engineering for Floor Area









Results and Analysis

	model_group	model_type	model_params	train_R2	train_RMSE	val_R2	val_RMSE	comments	features	feature_set
0	1	Linear Regression		0.92	24482.04	0.91	23708		53	features_max
1	1	Lasso	alpha=67.09	0.91	23275.27	0.9	24193.28		53	features_max
2	1	Ridge	alpha=1.0	0.92	23078.05	0.9	23869.37	Kaggle RMSE: 24727.75	53	features_max
3	2	Linear Regression		0.89	26621.22	0.88	26424.24	Non-CV R2 on train set is 0.9.	53	features_max_powerless
4	2	Lasso	alpha=65.87	0.9	25427.73	0.88	26713.18	max_iter = 10000	53	features_max_powerless
5	2	Ridge	alpha=10.0	0.9	25405.1	0.88	26575.5		53	features_max_powerless
6	3	Linear Regression		0.9	25433.04	0.9	24795.49	Non-CV R2 on train set is 0.91.	45	features_drop_weak
7	3	Lasso	alpha=117.24	0.91	24342.21	0.89	25283.04	max_iter = 1000	45	features_drop_weak
8	3	Ridge	alpha=1.0	0.91	24068.17	0.9	24910		45	features_drop_weak
9	4	Linear Regression		0.88	27353.21	0.89	25224.94	Non-CV R2 on train set is 0.89.	24	features_lite
10	4	Lasso	alpha=67.09	0.89	26751.63	0.89	25425.04	max_iter = 1000	24	features_lite
11	4	Ridge	alpha=1.0	0.89	26714.61	0.89	25297.28		24	features_lite
12	5	Linear Regression		0.83	32166.15	0.84	30618.22	Non-CV R2 on train set is 0.84.	2	features_min
13	5	Lasso	alpha=71.94	0.84	31964.52	0.84	30622.32	max_iter = 1000	2	features_min
14	5	Ridge	alpha=1.0	0.84	31964.44	0.84	30619.56		2	features_min
15	1	Linear Regression		0.93	22813.1	0.9	24442.4	Logarithmic transformation of target variable.	53	features_max

Strong Features

- Total area was strongest predictor: higher total area, higher price
- Lots at a cul de sac, on a gradient or irregularly shaped related negatively to price
- Low quality finished area strongly negatively related with price
- Larger pool size associated with higher prices

Feature	ridge coef
1.5P Total SF	68050.26
Total Bsmt SF	-45174.13
1st Flr SF	-38907.59
2nd Flr SF	-37510.62
CulDSac	-30230.2
Hillside	-25995.79
Regular Lot Shape	-24515.89
P3 Overall Qual	19130.79
Year Remod/Add	16531.61
Low Qual Fin SF	-16285.95
Pool Area	14749.63

Weak Features

- Structural features such as foundation and veneer not so important

- Number of bathrooms not very predictive of price

- Kitchen quality one of the weakest predictors of price, likely because kitchens do not vary much in quality, or that purchasers would just renovate kitchens after buying

Feature	coef
PConc Foundation	788.82
Has Vnr	773.49
Has Alley Access	661.13
Bath Log	534.78
Excellent Heating	-194.56
3Ssn Porch	174.36
Attached or BuiltIn Garage	-169.04
Kitchen Qual Num	113.12
Fireplace Qu Num	110.25

Thank You