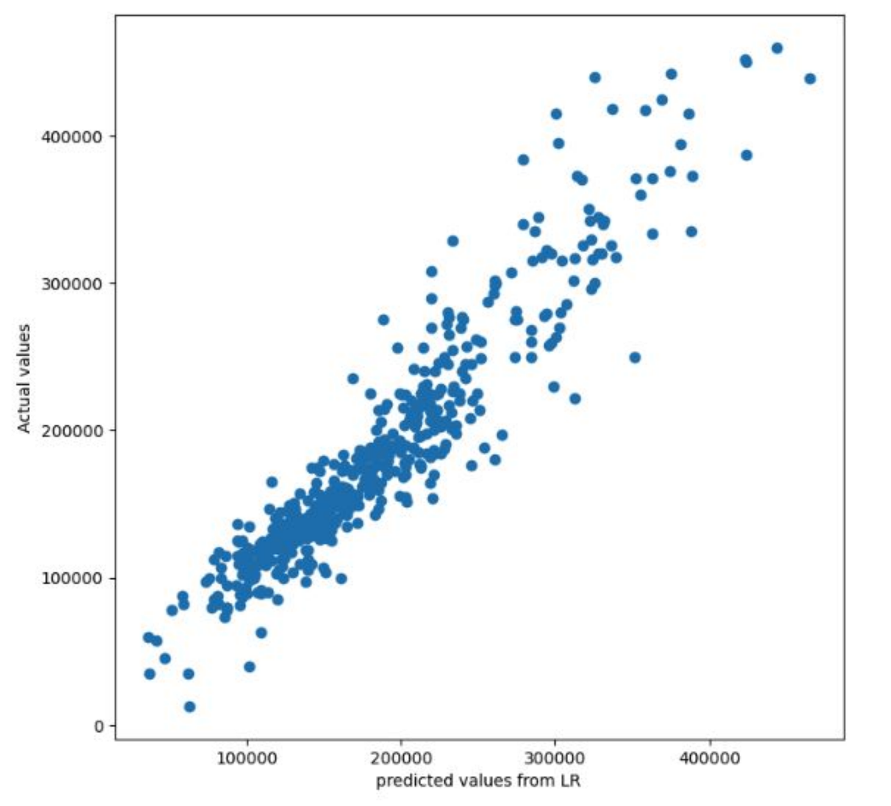
Project 2: Predicting House Prices from Ames Housing Data

Problem Statement

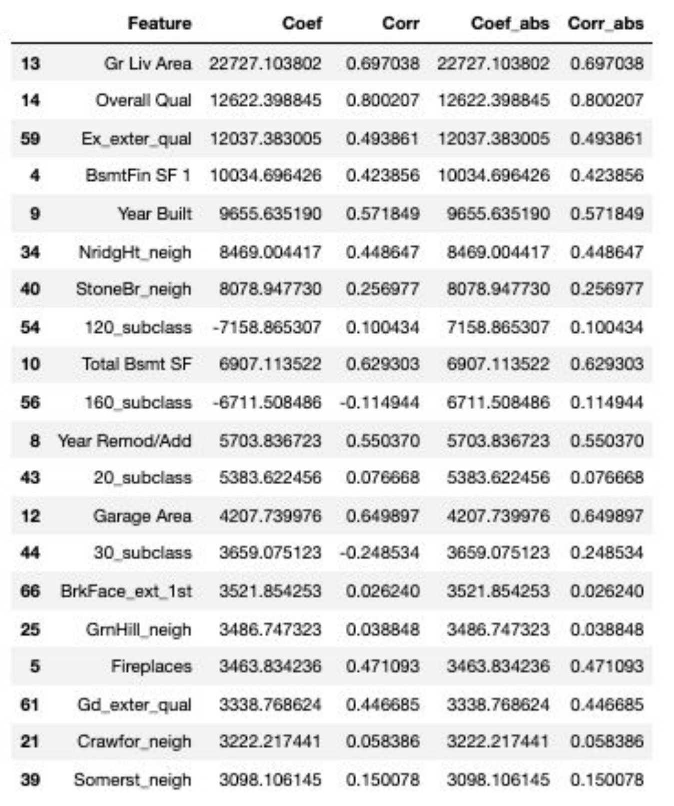
You are an estate agent in Ames, Iowa. You job is to use the dataset to predict Sale Price and also understand which areas are desirable/unattractive. You will also use this to advise clients on how to improve the value of their homes.

I managed to get a test score of 0.902 with Ridge



- Train 0.911, test 0.902, RMSE $23013
- I removed above ground living area above 3000 sq ft to remove outliers, and removed a few collinear features
- From the non continuous variables I added neighborhood, subclass, exterior quality and air conditioning
- After tuning my hyperparameters, I managed to achieve a 0.902 on Ridge which was my best score
- **Kaggle score 37323**

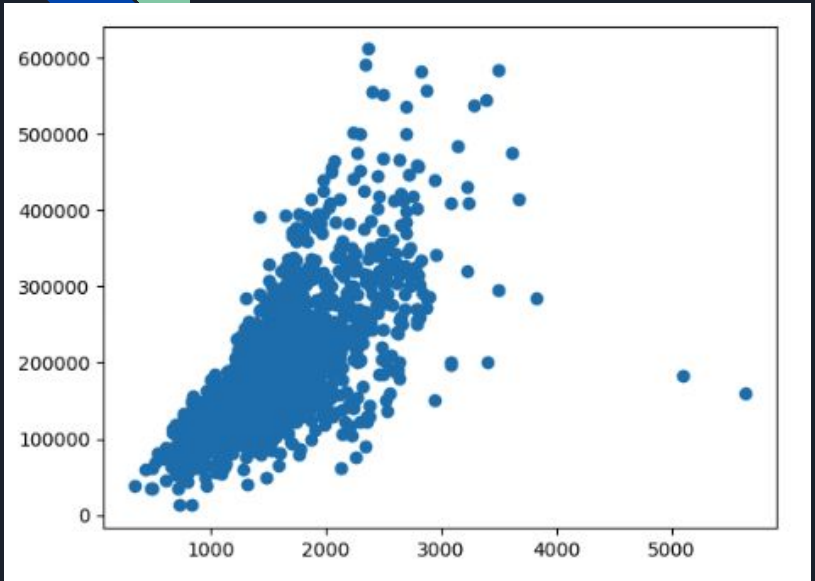
What are the most important features for home value?



	Feature	Coef	Corr	Coef_abs	Corr_abs
13	Gr Liv Area	22727.103802	0.697038	22727.103802	0.697038
14	Overall Qual	12622.398845	0.800207	12622.398845	0.800207
59	Ex_exter_qual	12037.383005	0.493861	12037.383005	0.493861
4	BsmtFin SF 1	10034.696426	0.423856	10034.696426	0.423856
9	Year Built	9655.635190	0.571849	9655.635190	0.571849
34	NridgHt_neigh	8469.004417	0.448647	8469.004417	0.448647
40	StoneBr_neigh	8078.947730	0.256977	8078.947730	0.256977
54	120_subclass	-7158.865307	0.100434	7158.865307	0.100434
10	Total Bsmt SF	6907.113522	0.629303	6907.113522	0.629303
56	160_subclass	-6711.508486	-0.114944	6711.508486	0.114944
8	Year Remod/Add	5703.836723	0.550370	5703.836723	0.550370
43	20_subclass	5383.622456	0.076668	5383.622456	0.076668
12	Garage Area	4207.739976	0.649897	4207.739976	0.649897
44	30_subclass	3659.075123	-0.248534	3659.075123	0.248534
66	BrkFace_ext_1st	3521.854253	0.026240	3521.854253	0.026240
25	GrnHill_neigh	3486.747323	0.038848	3486.747323	0.038848
5	Fireplaces	3463.834236	0.471093	3463.834236	0.471093
61	Gd_exter_qual	3338.768624	0.446685	3338.768624	0.446685
21	Crawfor_neigh	3222.217441	0.058386	3222.217441	0.058386
39	Somerst_neigh	3098.106145	0.150078	3098.106145	0.150078

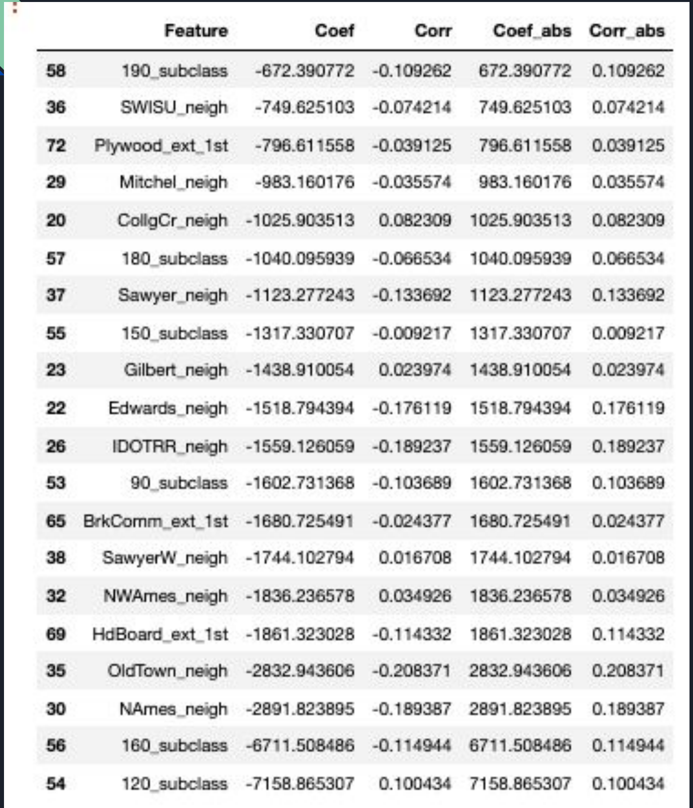
- I ranked the features by coefficient, with correlation as a check
- The most important feature is above ground living area (sq ft)
- Then Overall Quality, followed by External Quality, Basement square footage and Year Built
- Year of remodelling is also important, suggesting there is a return to remodelling work
- All of these have high coefficients and correlation
- Neighborhood is also important which we will explore in the next slide
- To add value, homeowners should
  - increase the living area
  - Increase the exterior quality
  - Increase the garage area
  - Increase number of fireplaces though this is likely correlated to rooms and living area
- However, these features are likely highly correlated to area and overall quality of the house

What are the most important features for home value?



- This shows Gr Liv Area vs Sale Price, which was the most important feature
- Removing the outliers above 3000 sq ft improved the model by ~0.015 as these outliers were far more dispersed

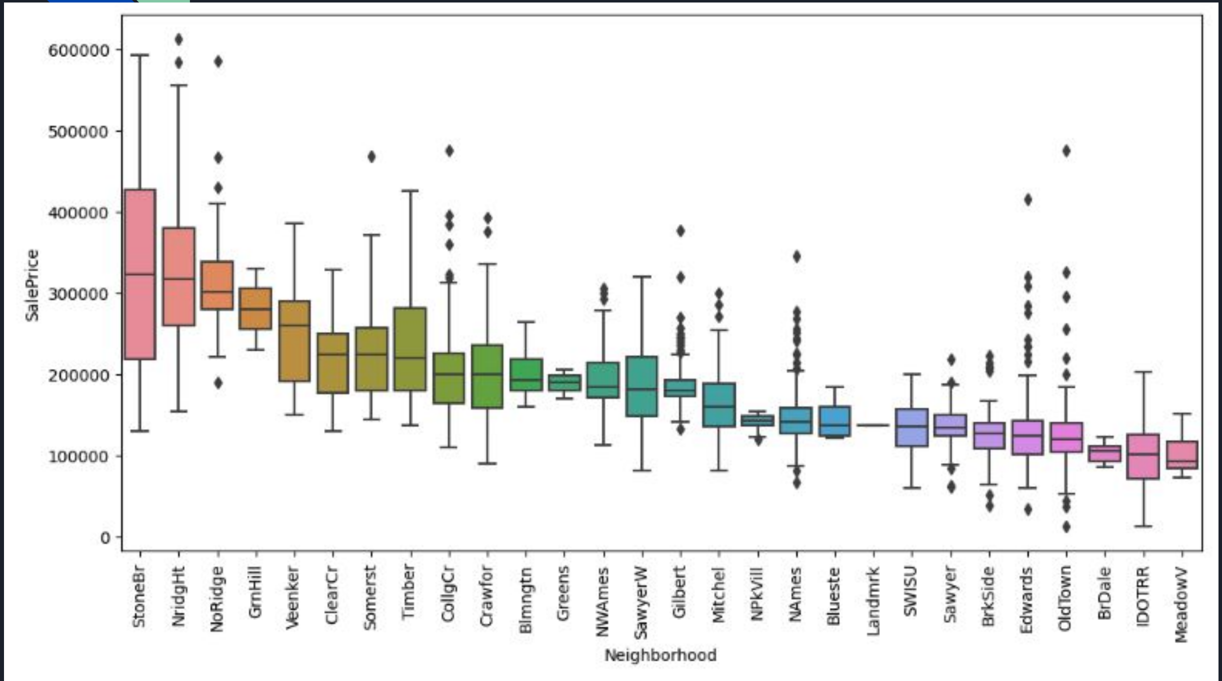
What are the most important features for home value?



	Feature	Coef	Corr	Coef_abs	Corr_abs
58	190_subclass	-672.390772	-0.109262	672.390772	0.109262
36	SWISU_neigh	-749.625103	-0.074214	749.625103	0.074214
72	Plywood_ext_1st	-796.611558	-0.039125	796.611558	0.039125
29	Mitchel_neigh	-983.160176	-0.035574	983.160176	0.035574
20	CollgCr_neigh	-1025.903513	0.082309	1025.903513	0.082309
57	180_subclass	-1040.095939	-0.066534	1040.095939	0.066534
37	Sawyer_neigh	-1123.277243	-0.133692	1123.277243	0.133692
55	150_subclass	-1317.330707	-0.009217	1317.330707	0.009217
23	Gilbert_neigh	-1438.910054	0.023974	1438.910054	0.023974
22	Edwards_neigh	-1518.794394	-0.176119	1518.794394	0.176119
26	IDOTRR_neigh	-1559.126059	-0.189237	1559.126059	0.189237
53	90_subclass	-1602.731368	-0.103689	1602.731368	0.103689
65	BrkComm_ext_1st	-1680.725491	-0.024377	1680.725491	0.024377
38	SawyerW_neigh	-1744.102794	0.016708	1744.102794	0.016708
32	NWAmes_neigh	-1836.236578	0.034926	1836.236578	0.034926
69	HdBoard_ext_1st	-1861.323028	-0.114332	1861.323028	0.114332
35	OldTown_neigh	-2832.943606	-0.208371	2832.943606	0.208371
30	NAmes_neigh	-2891.823895	-0.189387	2891.823895	0.189387
56	160_subclass	-6711.508486	-0.114944	6711.508486	0.114944
54	120_subclass	-7158.865307	0.100434	7158.865307	0.100434

- The biggest detractors are being a planned unit or suplex subclass (120, 160, 90)
- Then the major detractor was neighborhood, much like being in a desirable neighborhood was a positive driver of house price

How does neighborhood affect home price?



- StoneBr and NridgHt were the most desirable areas (though with StoneBr having wide dispersion) and had high coefficients and correlations
- Interestingly, GrnHill had a higher median price, but had far less impact on the model ie coefficient and correlations

- If I had more time I would try to understand features independent of neighborhood to understand the value of home improvements