Import and Cleaning

Feature	Resolution	Justification
Lot Frontage	Fill with mean	Continuous feature with no 0 values, skew is minimal so mean or median filling should be fine
Alley	Fill with 'NA'	On observation of the dataset, it is highly likely that the data which was supposed to be input as an 'NA' string was input as null values
Mas Vnr Type	Fill with 'None'	On observation of the dataset, it is highly likely that the data which was supposed to be input as a 'None' string was input as null values
Mas Vnr Area	Fill with 0	If we fill Vnr Type with 'None', the area should be 0
Bsmt Qual, Bsmt Cond, Bsmt Exposure and BsmtFin Type 1 & 2	Fill with NA	On observation of the dataset, it is highly likely that the data which was supposed to be input as an 'NA' string was input as null values
BsmtFin SF 1 & 2, Bsmt Unf SF and Total Bsmt SF	Fill with 0	It should be ok to set one datapoint to 0
Bsmt Full Bath and Bsmt Half Bath	Fill with 0	Basement details are already missing, as such we set basement baths to 0
Fireplace Qu	Fill with 0	Houses with NA for 'Fireplace Qu' have 0 fireplaces
Garage Type, Finish, Qual and Cond	Fill with 'NA'	We fill in missing garage info as there being no garage. It should be ok to make the assumption for ~5% of the feature data
Garage Yr Built	Fill with mean	We cannot fill with 0s as that will heavily skew the data considering the minimum year is 1895. We can fill with the mean as the skew of the data is minimal
Garage Cars and Area	Fill with 0	As we are setting the entries as having no garages, we set the values to 0. It should be ok to make the assumption for ~5% of the feature data
Pool QC, Fence and Misc Feature	Fill with 'NA'	On observation of the dataset, it is highly likely that the data which was supposed to be input as an 'NA' string was input as null values

Feature	Convert to	Reason
PID	string	Nominal data
Ms SubClass	string	Nominal data
Lot Shape	int64	Ordinal data
Utilites	int64	Ordinal data
Land Slope	int64	Ordinal data
Exter Qual	int64	Ordinal data
Exter Cond	int64	Ordinal data
Bsmt Qual	int64	Ordinal data
Bsmt Cond	int64	Ordinal data
Bsmt Exposure	int64	Ordinal data
BsmtFin Type 1 & 2	int64	Ordinal data
HeatingQC	int64	Ordinal data
Electrical	int64	Ordinal data
KitchenQual	int64	Ordinal data
Functional	int64	Ordinal data
FireplaceQu	int64	Ordinal data
Garage Finish	int64	Ordinal data
Garage Qual	int64	Ordinal data
Garage Cond	int64	Ordinal data
Paved Drive	int64	Ordinal data
Pool QC	int64	Ordinal data
Fence	int64	Ordinal data

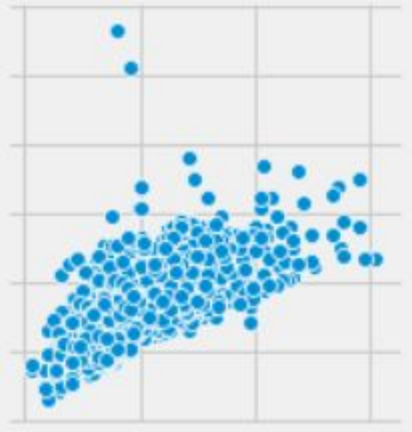
Data Analysis

Feature 1	Feature 2	Corr. Score	Decision
exter_qual	overall_qual	0.74	It is likely that overall quality is highly dependent on exterior quality. As such we choose to drop exterior quality as overall quality should be a sufficient predictor
kitchen_qual	exter_qual	0.73	Kitchen, exterior and overall quality seem to be interdependent. As kitchen and overall quality have a correlation score of 0.69, we choose to drop kitchen quality as well
bsmtfin_type_1	bsmtfin_sf_1	0.70	The basement rating and its area seem to be related. As such we choose to drop 'bsmtfin_sf_1' as its area can be represented by its rating
bsmtfin_type_2	bsmtfin_sf_2	0.78	Similar to the previous case, we choose to drop 'bsmtfin_sf_2'
1st_flr_sf	total_bsmt_sf	0.81	The size of the basement is highly correlated to the size of the first floor, and we choose to drop total_bsmt_sf as 1st_flr_sf would serve as a sufficient predictor
totrms_abvgrd	gr_liv_area	0.81	The 'total rooms above grade and 'above grade living area' features are highly correlated. We choose to drop 'totrms_abvgrd' as 'gr_liv_area' seems to have a better correlation with sale price
fireplace_qu	fireplaces	0.86	Fireplace quality and their quantity are highly correlated. We drop 'fireplaces' as 'fireplace_qu' has a better correlation with sale price
garage_yr_blt	year_built	0.79	When houses and garages are built seem to be highly correlated, as such we drop 'garage_yr_blt' as 'year_built has a better correlation with sale price
garage_area	garage_cars	0.89	The garage area directly affects the number of cars it can store. We drop 'garage_cars' as the area serves as a sufficient predictor
paved_drive	garage_qual	0.95	There is a very high correlation between how the driveway is paved and the quality of the house's garage. We drop 'paved_drive' as 'garage_qual' should serve as a sufficient predictor
pool_qc	pool_area	0.87	The quality of the pool and its area are highly correlated. We drop 'pool_area' as 'pool_qc' has a slightly higher correlation with sale price

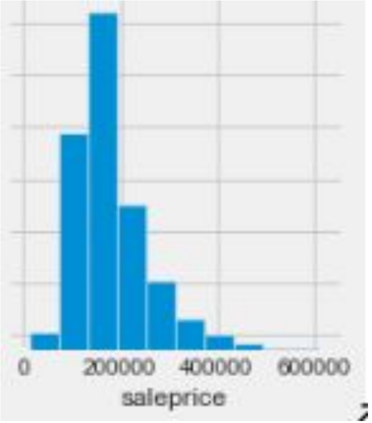
Feature	Corr. Score
utilities	0.03
land_slope	-0.06
overall_cond	-0.10
exter_cond	0.04
bsmtfin_type_2	0.01
bsmtfin_sf_2	0.02
bsmt_unf_sf	0.19
low_qual_fin_sf	-0.04
bsmt_half_bath	-0.05
bedroom_abvgr	0.14
kitchen_abvgr	-0.13
functional	0.13
enclosed_porch	-0.14
3ssn_porch	0.05
screen_porch	0.13
pool_area	0.02
pool_qc	0.03
fence	-0.16
misc_val	-0.01
mo_sold	0.03
yr_sold	-0.02

Data Analysis

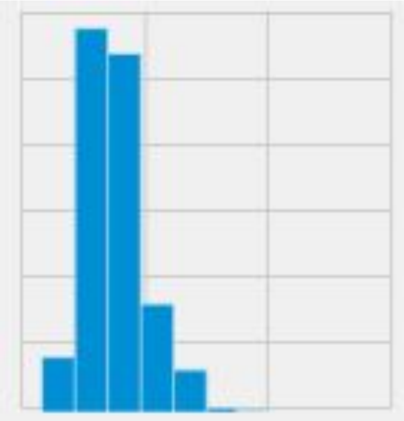
gr_liv_area vs saleprice



saleprice histogram



gr_liv_area histogram



overall_cond vs saleprice

Modelling

	variable	coef	abs_coef
17	gr_liv_area	24134.744732	24134.744732
4	overall_qual	15943.389729	15943.389729
61	neighborhood_NridgHt	9819.480090	9819.480090
0	ms_subclass	-8926.083779	8926.083779
6	year_built	7371.431790	7371.431790
11	bsmt_exposure	7211.733603	7211.733603

Method	Number of Features	Parameters Optimized over	Parameter value	Adjusted R^2 score
Lasso	36	All features	568.038	0.841979
Linear Regression	36	NA	NA	0.841947
Linear Regression	1	NA	NA	0.418358
Ridge	36	36 features	28.660	0.842677
Ridge	37	All features	289.942	0.842217
Elastic Net	36	36 features	$\alpha = 63.749, l_1 ratio = 1.0$	0.842042
Elastic Net	32	All features	$\alpha = 594.531, l_1 ratio = 1.0$	0.842738

Conclusions

We recommend that Elastic Net Regression be used with an alpha parameter of 594.531 and an l1 ratio of 1.0. A set of 32 features should be used as predictors which are found in the 'features.txt' document in our datasets folder.