House Price Prediction

Capstone Presentation

Introduction / Business Understanding

Goals:

- Predict sale price of houses

- Understand coefficients in model

Data Understanding

Data from Kaggle

Original dataset describes the houses, including factors like the age and quality of the house, the total size of both the lot and the house itself, the number of bedrooms and bathrooms, and additional features like any porch, garage, basement, etc.

Data Cleaning

A few **categorical variables** were reduced down to the most popular category vs not.

Future analysis will include expanding the feature set to include more of the categorical variables.

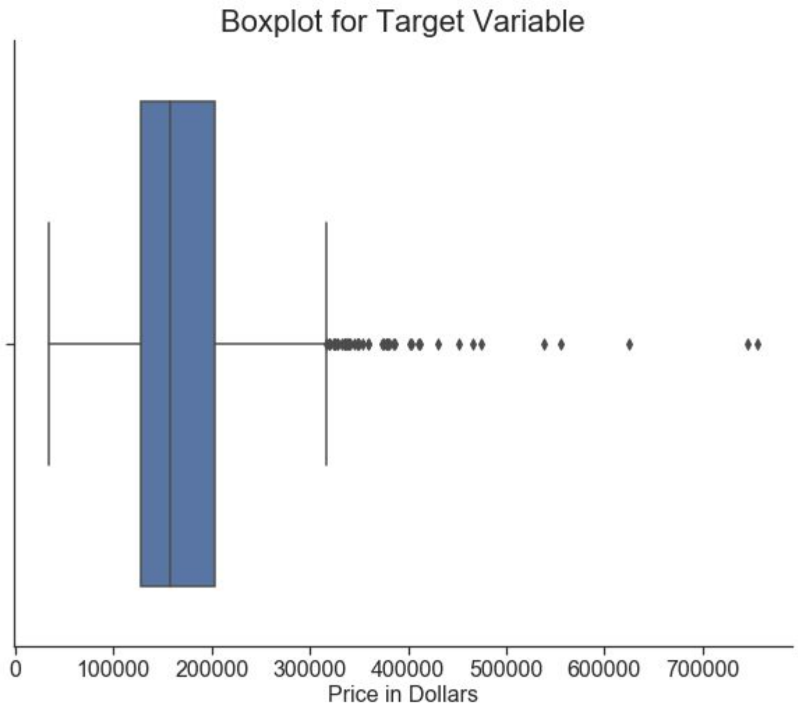
The **resulting dataset** is a mixture of continuous, ordinal, and binary variables with **no missing observations.**

All variables are **numeric**

Many new variables created

- HasPool: There are only 6 pools in the dataset. I dropped PoolQC and transformed PoolArea into HasPool
- HasFireplace: reduced Fireplaces, which is a count of fireplaces in the home, to a binary yes/no if the home has a fireplace at all.
- HasFence: reduced Fence, a list of fence condition descriptors, to a binary yes/no if the home came with a fence.
- HasGarage: if GarageArea equals zero, then the property has no garage. It also explains why some observations have missing variables for garage attributes. Upon investigation, those observations have zero garage area.
- GarageAreaPerCar: GarageArea divided by GarageCars. With this created I could then drop GarageArea. It's assumed that the more cars, the larger the garage, so I did not want both variables in there violating regression assumptions. Having the ratio will control for multi-car garages that actually have very little space.
- BsmtPerFinished: The percent of the basement that is finished
- HasCentralAir: When CentralAir equals "Yes"
- GasAirHeat: Is set to 1 when the house uses a Gas forced warm air furnace for heating
- SBboxElectric: Is set to 1 when Standard Circuit Breakers & Romex is how the electrical is wired
- HasDeck: If the squarefootage of deck area is greater than 0
- HasRemod: If a renovation happened at all in house history
- HouseAge: They age of the house at time of sale
- TimeSinceRemodel: The years since the remodel
- RemodFiveYrs: If the remodel happened within the 5 years prior to the sale
- GaragebuiltWHouse: Was the garage built at the same time as the house or later
- AverageRoomSize: Divided total living above ground space by total above ground room count
- HasFinishedBsmt: If any part of the basement is finished

Target: Sale Price of Houses

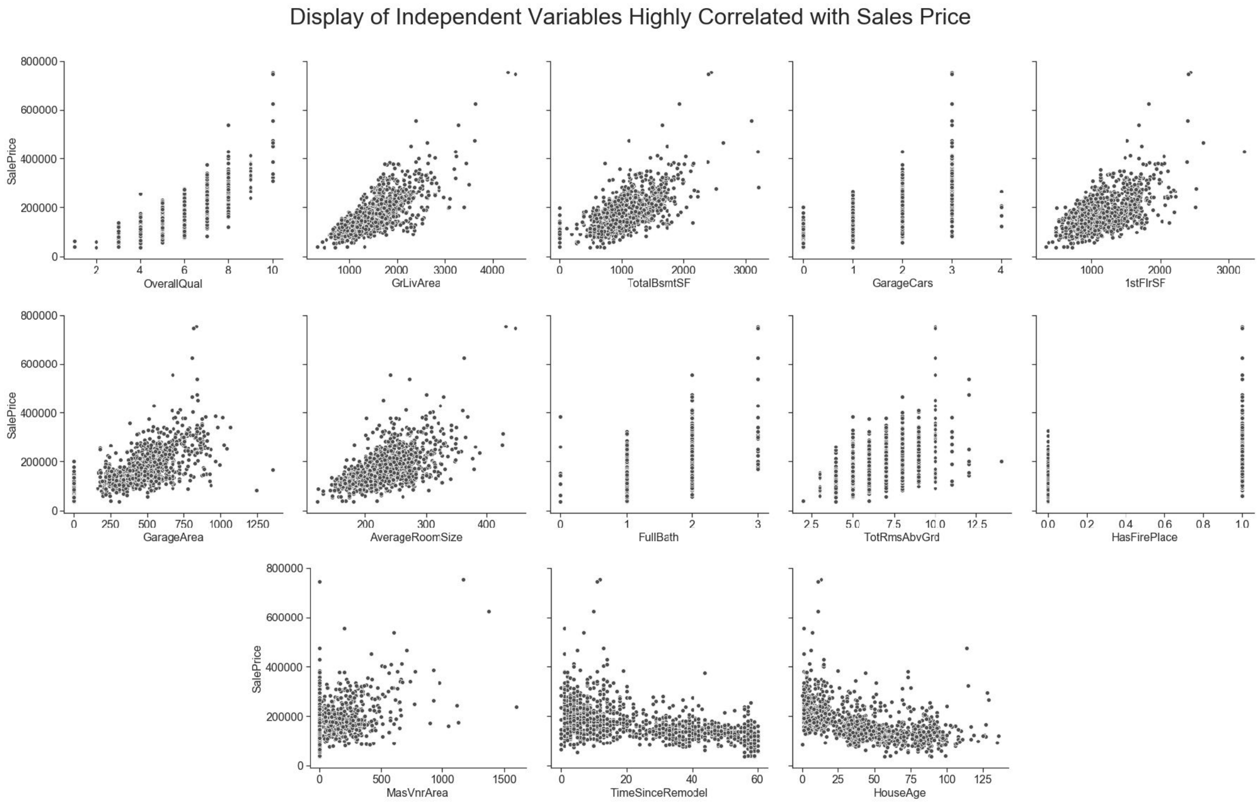


The *median* price of a home sold in this data set is $157,950.00

The *mean* price of a home sold in this data set is $173,294.63

While the bulk of Sale Price distribution appears fairly normal, the target variable shows that there are some outliers in the data. These higher priced homes skew the distribution. I kept these outliers in, however may consider doing a transformation at some point to normalize the target variable. Alternatively, I could train my model without those outliers, with the understanding that then my model would only be accurate at homes that would be sold at prices less than a certain amount.

Variables Compared with Target



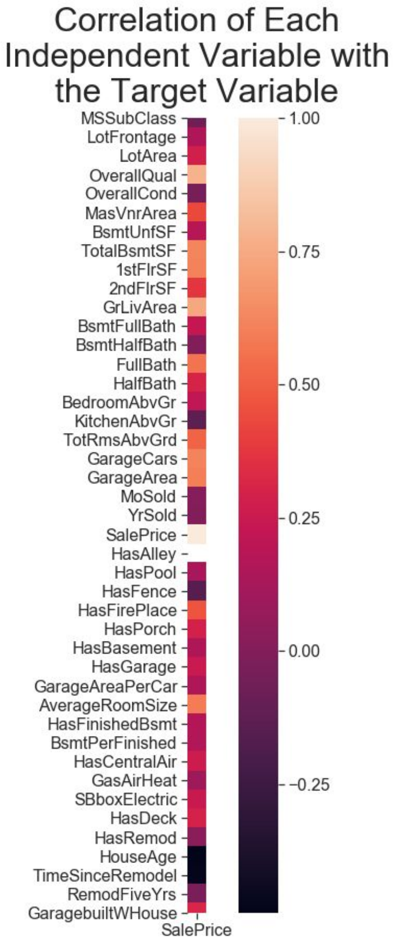
Data Preparation and Cleaning

Used 42 numeric columns for analysis to capture the quantitative aspects of homes being sold, after excluding categorical features.

Many of the initial correlations seem counter-intuitive. Why would having a fence be negatively correlated with sales price?

Why isn’t total square footage showing a larger correlation?

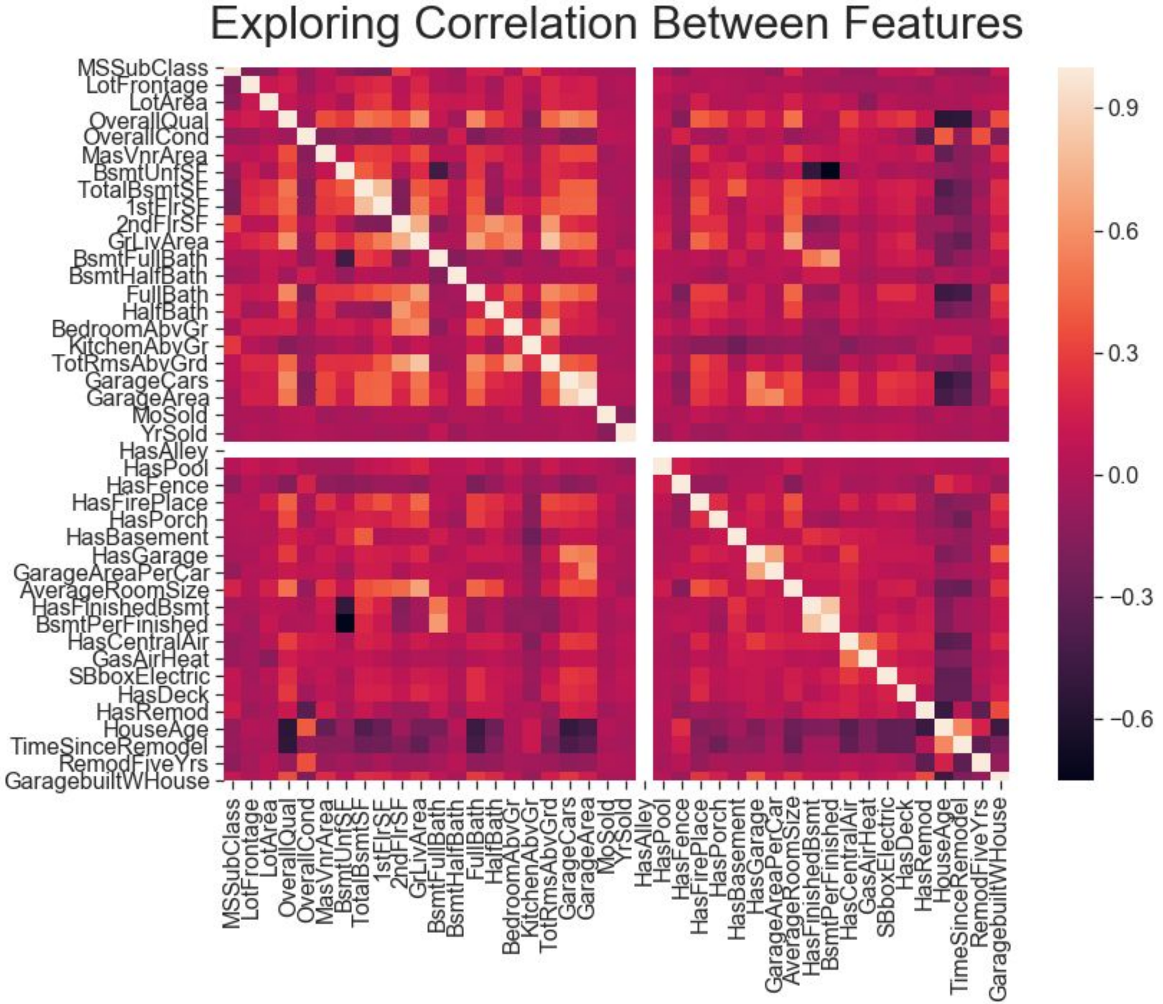
Reason: Correlation is not a robust measure of relationship



Modeling

- **Baseline**: predicts the mean house price from the training data for each house
- **Multiple linear regression** model to improve upon baseline
- **Ridge regression** model to manage the multicollinearity between my features

Multicollinearity



Evaluation

My preferred model is a Ridge Regression model

- R2 Score: .89
- MAE: 1801

	Training R2	Testing R2	Training MAE	Testing MAE
Baseline Model	0.00%	-0.96%	$51,797.20	$52,952.53
OLS Model	89.40%	86.56%	$17,481.87	$18,076.08
Ridge Regression	89.39%	86.62%	$17,462.86	$18,011.05

Weight	Feature
2.0422 ± 0.2634	GrLivArea
0.6316 ± 0.0540	TotRmsAbvGrd
0.3785 ± 0.0292	AverageRoomSize
0.3780 ± 0.0564	TotalBsmtSF
0.1554 ± 0.0200	BsmtUnfSF
0.1182 ± 0.0157	OverallQual
0.1049 ± 0.0189	HouseAge
0.0702 ± 0.0030	2ndFlrSF
0.0563 ± 0.0129	BsmtPerFinished
0.0380 ± 0.0079	BedroomAbvGr
0.0299 ± 0.0042	OverallCond
0.0154 ± 0.0032	LotArea
0.0134 ± 0.0023	HasBasement
0.0124 ± 0.0043	GarageCars
0.0096 ± 0.0074	1stFlrSF
0.0059 ± 0.0045	FullBath
0.0059 ± 0.0042	KitchenAbvGr
0.0041 ± 0.0029	MSSubClass
0.0029 ± 0.0033	MasVnrArea
0.0021 ± 0.0011	HalfBath
0.0020 ± 0.0018	BsmtFullBath
0.0018 ± 0.0015	GarageArea
0.0010 ± 0.0010	GaragebuiltWHouse
0.0007 ± 0.0005	SBboxElectric
0.0007 ± 0.0013	HasPool
0.0006 ± 0.0016	LotFrontage
0.0005 ± 0.0006	HasFence
0.0005 ± 0.0005	HasPorch
0.0004 ± 0.0004	HasFirePlace
0.0002 ± 0.0005	RemodFiveYrs
0.0002 ± 0.0006	HasRemod
0.0000 ± 0.0002	BsmtHalfBath
0.0000 ± 0.0002	HasFinishedBsmt
0.0000 ± 0.0000	HasAlley
-0.0000 ± 0.0050	HasGarage
-0.0001 ± 0.0006	YrSold
-0.0001 ± 0.0002	MoSold
-0.0002 ± 0.0004	TimeSinceRemodel
-0.0003 ± 0.0006	GasAirHeat
-0.0003 ± 0.0005	HasCentralAir
-0.0005 ± 0.0010	GarageAreaPerCar
-0.0006 ± 0.0013	HasDeck

Next Steps

In the future I would like to first use more of the categorical features, and perhaps encode some of the discrete features I used in my final model. I would also like to then only use the most important features, perhaps by regularizing using both LASSO and Ridge through an ElasticNet model. I could also only use the top 5-10 features based on Permutation Importance.

Coefficients

MSSubClass	-47.360701
LotFrontage	36.130818
LotArea	0.550012
OverallQual	13385.754531
OverallCond	7425.838372
MasVnrArea	25.746881
BsmtUnfSF	-42.110121
TotalBsmtSF	73.371413
1stFlrSF	14.246192
2ndFlrSF	32.987960
GrLivArea	139.474943
BsmtFullBath	4003.013587
BsmtHalfBath	-652.279737
FullBath	-5937.087482
HalfBath	-2836.097245
BedroomAbvGr	-10062.078577
KitchenAbvGr	-13777.434459
TotRmsAbvGrd	-21791.479670
GarageCars	6497.169230
GarageArea	6.98801[illegible]

MoSold	-115.735481
YrSold	411.339213
HasAlley	0.000000
HasPool	20217.132597
HasFence	-2566.104747
HasFirePlace	1480.721584
HasPorch	2642.702510
HasBasement	-21519.240631
HasGarage	-12740.468310
GarageAreaPerCar	-8.554247
AverageRoomSize	-641.192879
HasFinishedBsmt	-402.500025
BsmtPerFinished	-31551.363176
HasCentralAir	1411.089038
GasAirHeat	5229.303467
SBboxElectric	-1761.628554
HasDeck	4001.467561
HasRemod	-2165.235041
HouseAge	-511.525974
TimeSinceRemodel	-17.323424
RemodFiveYrs	3362.526696
GaragebuiltWHouse	-5818.363776

Contact

https://www.linkedin.com/learn-co-curriculum

https://github.com/learn-co-curriculum