Technical Presentation

Understanding House Prices

Using Linear Regression to Find the Predictors of House Prices in Ames, Iowa

Problem Summary

Problem: how to predict a house's final sale price?

House prices are influenced by many factors. Some of them are fairly well-known to the general public, but many factors are not known to most home buyers, impairing their ability to negotiate house prices accurately and fairly.

Technical Approach

Can we accurately predict house prices in Ames, Iowa, by developing a statistically significant regression model based on data on house characteristics?

The goal of this analysis was to develop a predictive model of sale price for houses in Ames, Iowa, based on property characteristics.

Statistical Approach #1

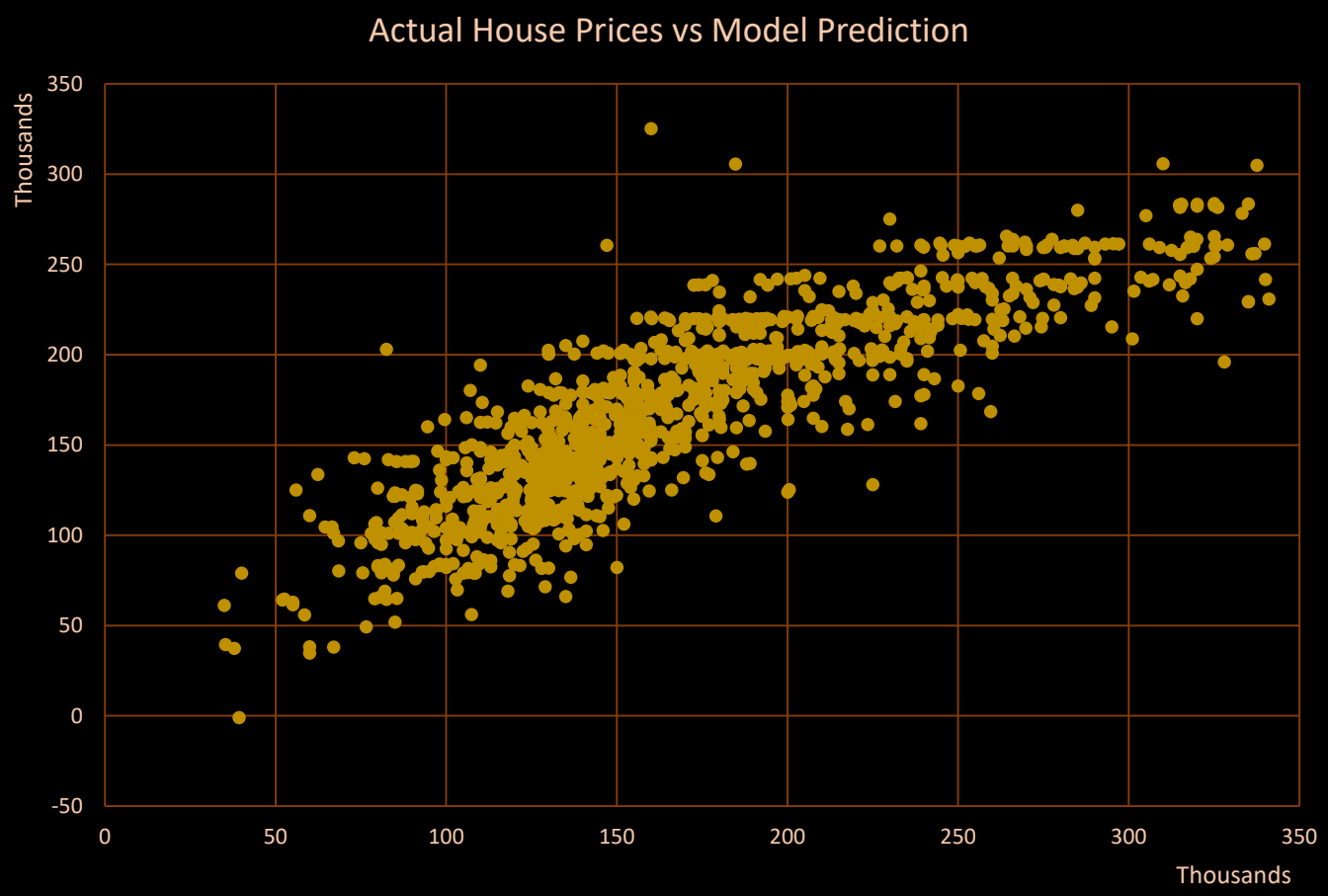
Descriptive statistics of house characteristics in Ames, Iowa, focusing on analysis of spread and distribution.

Statistical Approach #2

Inferential statistical analysis, namely correlational analysis and linear regression, to understand what are the best predictors of sale price

Executive Summary

Analysis showed that the variables overall quality of finish and materials, year built, the year of remodel, number of fireplaces and garage car capacity predict 73.7% of the variation of sale price of houses in Ames, Iowa.



To predict sale price of houses in Ames, Iowa, buyers and sellers should focus on overall quality score, year built, year remodeled, nr. of fireplaces and garage car capacity.

Support Point 1

Regression analysis indicated that these factors are statistically significant predictors of 73.7% of the variation in sale price.

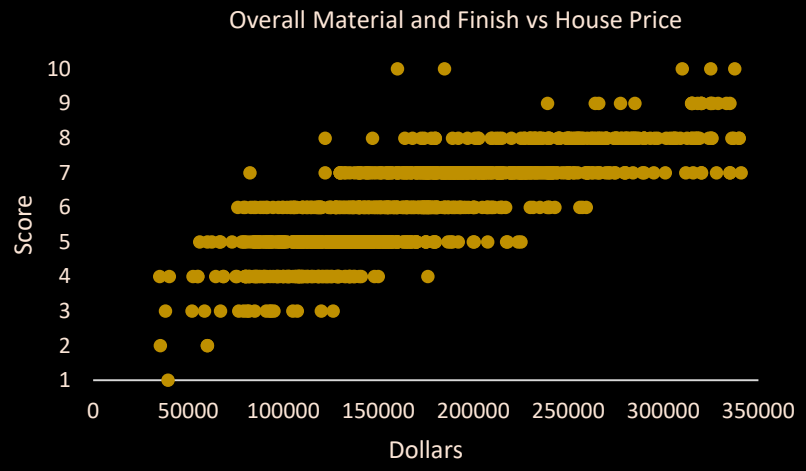
Support Point 2

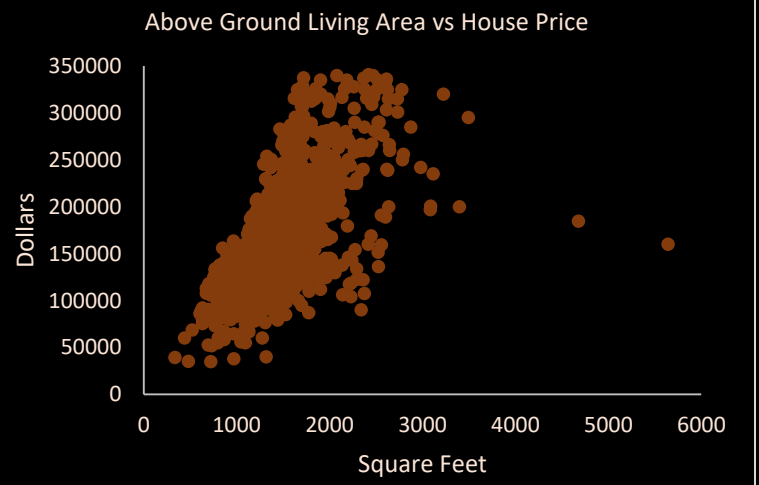
The dataset was mostly focused on variables regarding area built. More information is needed to understand the effect of lot area without any built infrastructure.

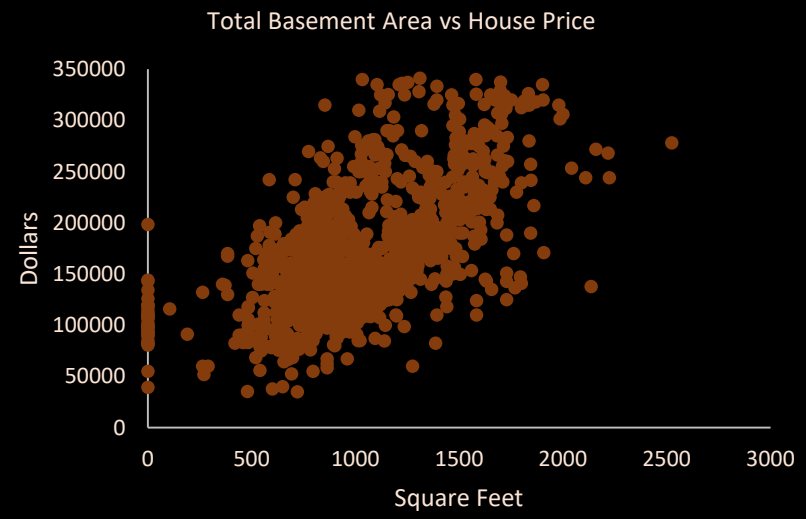
Source: 'House Prices: Advanced Regression Techniques' dataser, available at https://www.kaggle.com/c/house-prices-advanced-regression-techniques

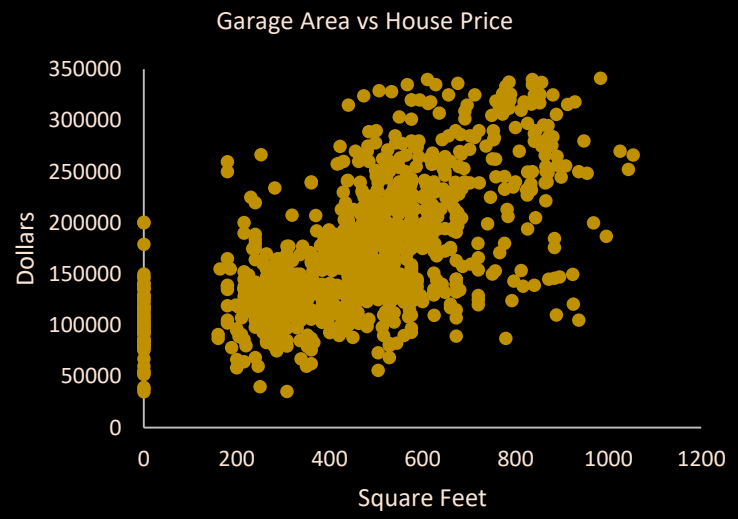
Support Slide

Initial descriptive analysis suggested that some house characteristics might be associated with sale price









Distribution analysis suggested that some house characteristics have a relevant relationship with sale price.

This suggested that it was likely that further statistical analysis would identify potential strong predictors of sale price in this dataset.

In particular, the variables overall quality of finish and materials, year built, basement dimensions, above ground living area and garage area seemed to be significantly associated with sale price.

Source: 'House Prices: Advanced Regression Techniques' dataser, available at https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Support Slide

Correlational analysis confirmed that some house characteristics are strongly correlated with sale price, namely overall quality of finish and materials, ground living area, garage area and garage car capacity.

Relevant correlations w/ Sale Price

LotFrontage	0.322159108
LotFrontage	0.322159108
OverallQual	0.799092563
YearBuilt	0.583471862
YearRemodAdd	0.56033252
MasVnrArea	0.371145829
TotalBsmtSF	0.557955707
1stFlrSF	0.536715388
GrLivArea	0.646156206
FullBath	0.589083179
TotRmsAbvGrd	0.470106284
Fireplaces	0.452492377
GarageYrBlt	0.305002952
GarageCars	0.635767274
GarageArea	0.617895888
WoodDeckSF	0.324491185
OpenPorchSF	0.336827721
Dummy 1 [CentralAir]	0.301981272
Dummy 3 [GarageType - Attached]	0.444912912
Dummy 7 [GarageType - Detached]	-0.386312073

Correlational analysis confirmed that some house characteristics are indeed strongly correlated with house sale price in Ames, Iowa.

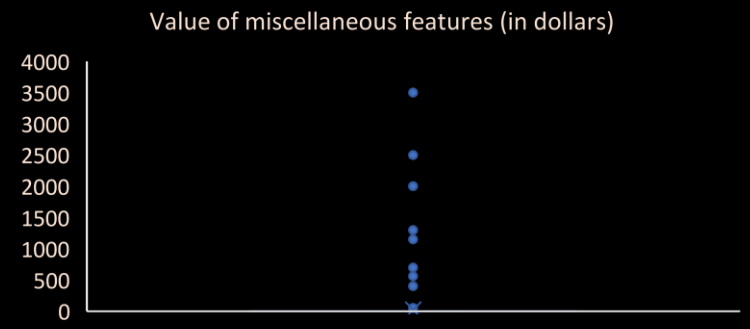
The variables more strongly correlated with sale price matched the variables we had previously identified. However, a total of 20 variables show relevant correlations with sale price.

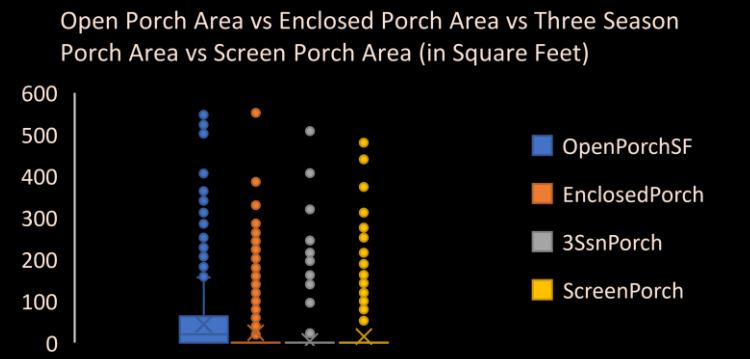
Based on this information, the next step was running a regression analysis to identify the best predictors of sale price

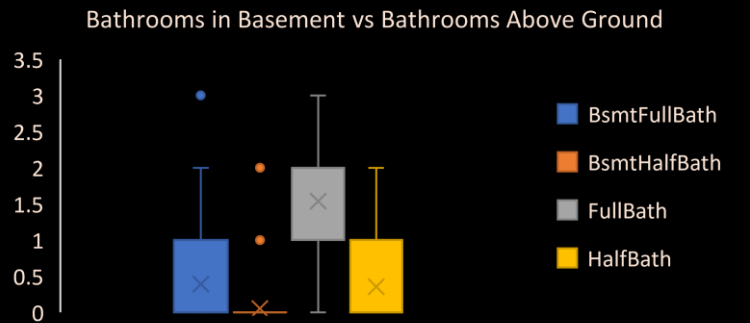
Source: 'House Prices: Advanced Regression Techniques' dataser, available at https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Support Slide

Software limitations forced the reduction of the number of variables in regression; the 16 variables more correlated with sale price were selected for regression.







Relevant correlations w/ Sale Price

LotFrontage	0.322159108
LotFrontage	0.322159108
OverallQual	0.799092563
YearBuilt	0.583471862
YearRemodAdd	0.56033252
MasVnrArea	0.371145829
TotalBsmtSF	0.557955707
1stFlrSF	0.536715388
GrLivArea	0.646156206
FullBath	0.589083179
TotRmsAbvGrd	0.470106284
Fireplaces	0.452492377
GarageYrBlt	0.305002952
GarageCars	0.635767274
GarageArea	0.617895888
WoodDeckSF	0.324491185
OpenPorchSF	0.336827721
Dummy 1 [CentralAir]	0.301981272
Dummy 3 [GarageType - Attached]	0.444912912
Dummy 7 [GarageType - Detached]	-0.386312073

Because Microsoft Excel only runs regressions with 16 independent variables maximum, it was essential to reduce the variables under analysis.

The first step in reducing variables for regression was removing those variables that didn't add much to the mix. That included variables with very low counts, or that had mostly null data.

Since that wasn't enough, the next step was looking at the correlations, and removing the ones more weakly associated with house price.

Our final pool of variables consisted of the 16 variables in this list more strongly correlated with sale price.

Source: 'House Prices: Advanced Regression Techniques' dataser, available at https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Support Slide

Regression analysis showed that 10 house characteristics are strong statistically significant predictors of sale price, predicting 79.4% of its variation.

	Coefficients	Standard Error	t Stat	P-value
OverallQual	15096.23479	1011.338	14.92699	4.08629E-46
YearBuilt	257.3441092	37.87117	6.795251	1.74722E-11
YearRemodAdd	367.9287124	50.84764	7.235905	8.52701E-13
TotalBsmtSF	13.37719159	2.447006	5.466758	5.63987E-08
GrLivArea	29.52114384	2.314022	12.7575	6.44697E-35
Fireplaces	11881.21566	1547.465	7.677857	3.4926E-14
GarageYrBlt	-5.6640354	2.232238	-2.53738	0.011302347
GarageCars	9588.003075	2572.843	3.726618	0.000203685
GarageArea	23.11564583	8.768459	2.636227	0.008498468
WoodDeckSF	32.36539223	7.469734	4.332871	1.60292E-05

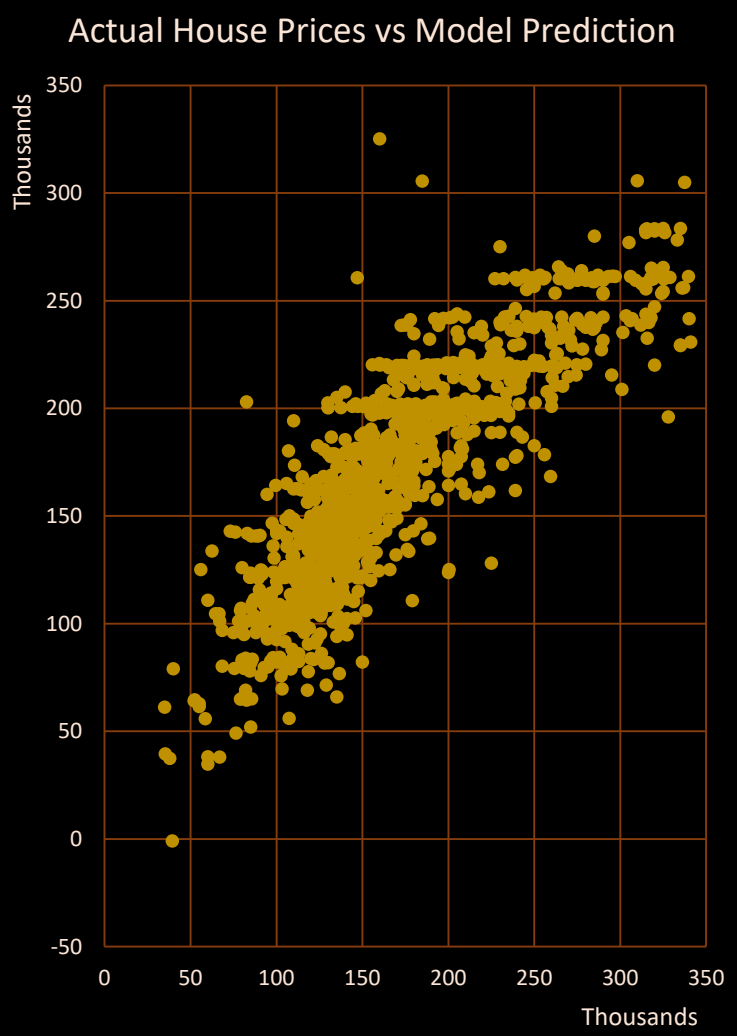
A model consisting of overall quality of finish and materials, year built, year remodeled, total basement area, above ground living area, number of fireplaces, garage year built, garage car capacity, garage area and wood deck area allows buyers and sellers to predict 79.4% of the variation of sale price.

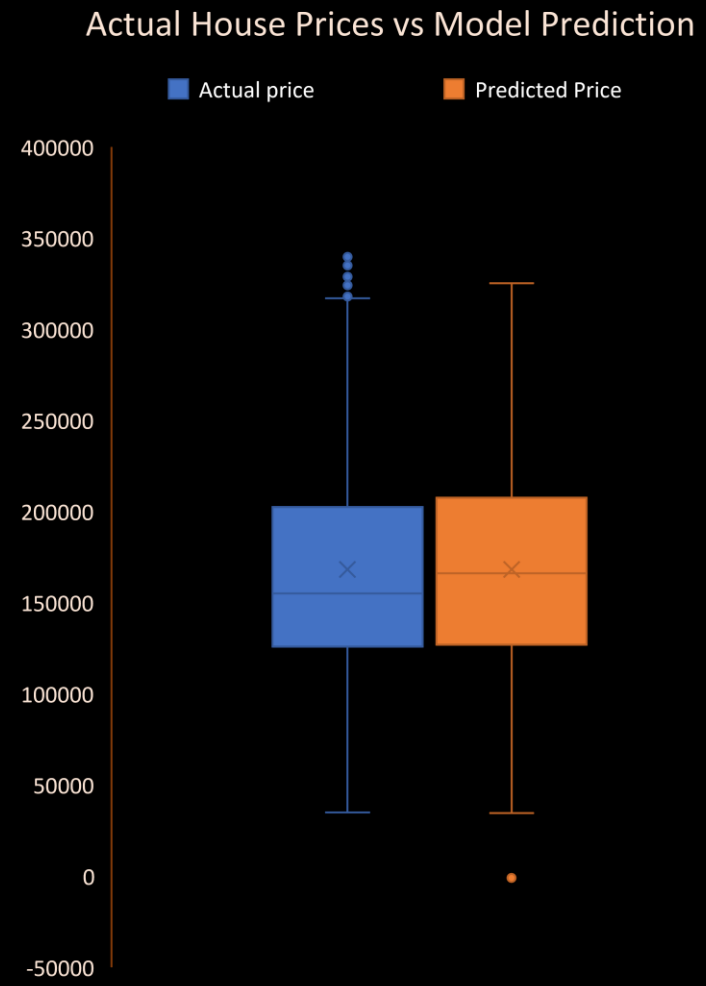
This is a strong predictive model that can be highly useful to buyers and sellers. Not only predicts most of the variation of sale price, but it's also based on 10 house characteristics about which buyers and sellers can easily get information.

Source: 'House Prices: Advanced Regression Techniques' dataser, available at https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Support Slide

Further analysis allowed us to obtain a reduced predictive model of sale price with a R^2 of 0.737 and only 5 variables - overall quality, year built, the year of remodel, number of fireplaces and garage car capacity.





The 10-variable model is already highly predictive of sale price variation, but not all variables add much predictive power. Therefore we developed a reduced model that predicts 73.7% of the variation in sale price using only 5 house characteristics, making it more efficient to use.

Based on this analysis, we recommend that buyers and sellers focus on overall quality score, year built, year remodeled, nr. of fireplaces and garage car capacity when trying to predict sale price of houses in Ames, Iowa.

Source: 'House Prices: Advanced Regression Techniques' dataser, available at https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Understanding House Prices

Thank you.

Source: 'House Prices: Advanced Regression Techniques' dataser, available at https://www.kaggle.com/c/house-prices-advanced-regression-techniques