



Predicting House Prices in Ames, Iowa

Friday, April 14th 2023
Martijn de Vries
General Assembly
DSIR-0320 Project 2



Problem Statement



A real estate company in Ames, Iowa is looking for a better way to evaluate the market value of a house. Using the **Ames dataset**, I will build a **predictive model** to predict the sale price of a house.

Benchmark model:

An model using only the **total living area** to predict sale price.

How much can we improve on this?

Data



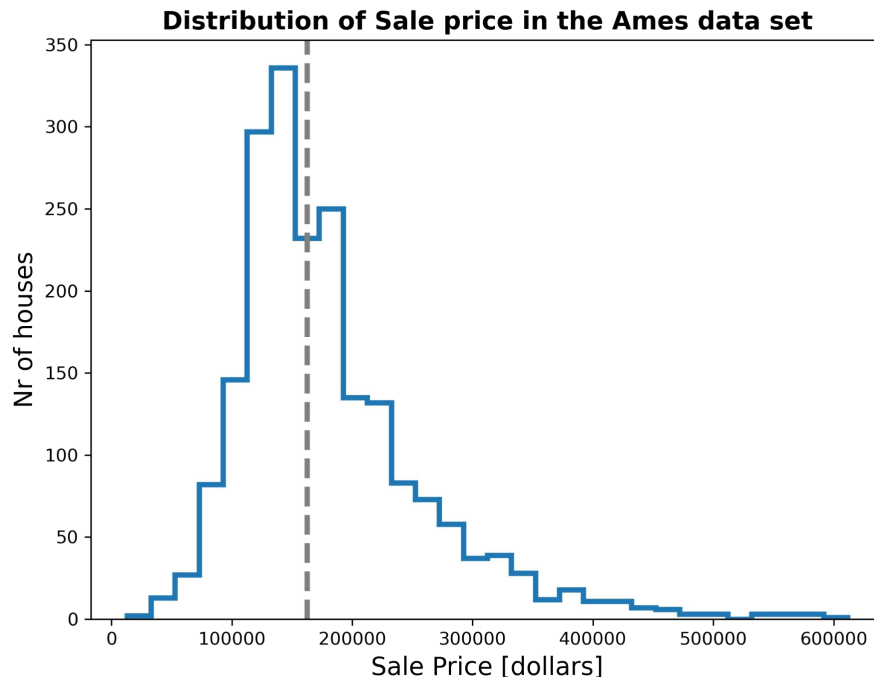
The 'famous' Ames dataset¹

Training data: 2051 houses

Houses sold in Ames, IA from 2006-2010

78 columns (!) + Sale price

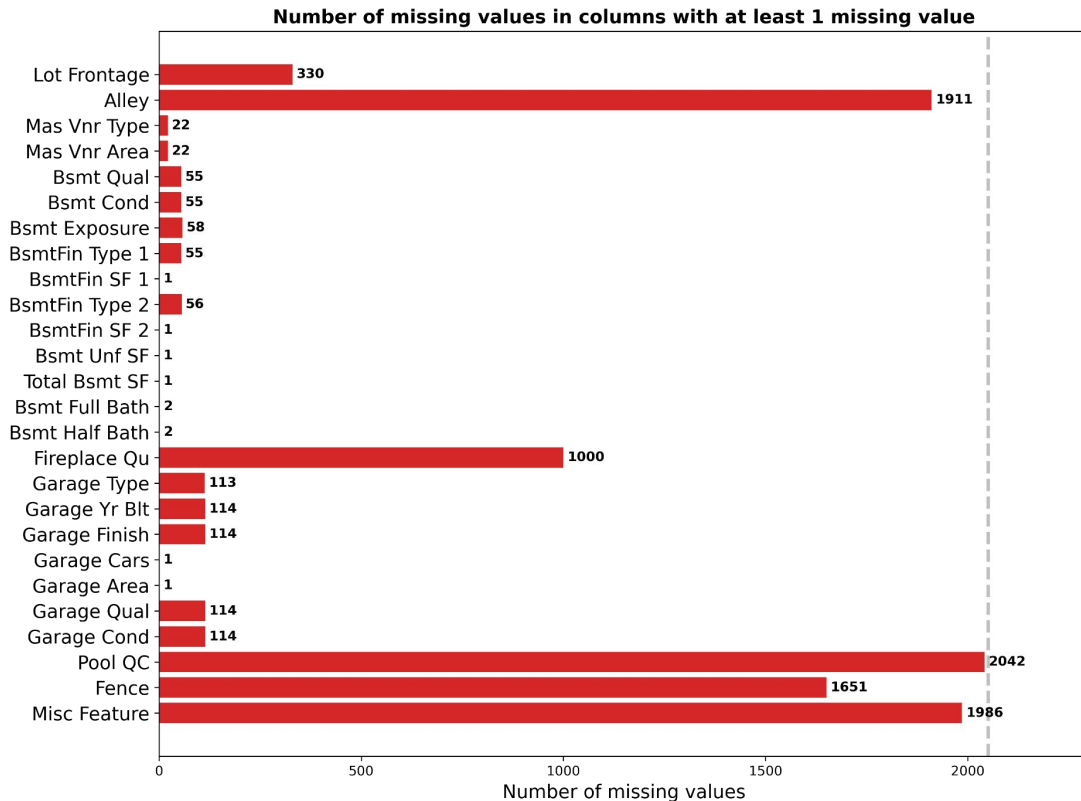
- Numerical (eg: Square footage)
- Ordinal (eg: Quality of the kitchen)
- Nominal (eg: What kind of garage)



¹<https://jse.amstat.org/v19n3/decock.pdf>

Data Cleaning

1. 'NA' values interpreted as missing -> 'NP'
2. Related columns set to zero (eg. Garage Area)
3. Lot Frontage: imputation based on $\sqrt{\text{Lot Area}}$
4. Everything else: impute most frequent value (categories) or median (numerical data)



Building a model

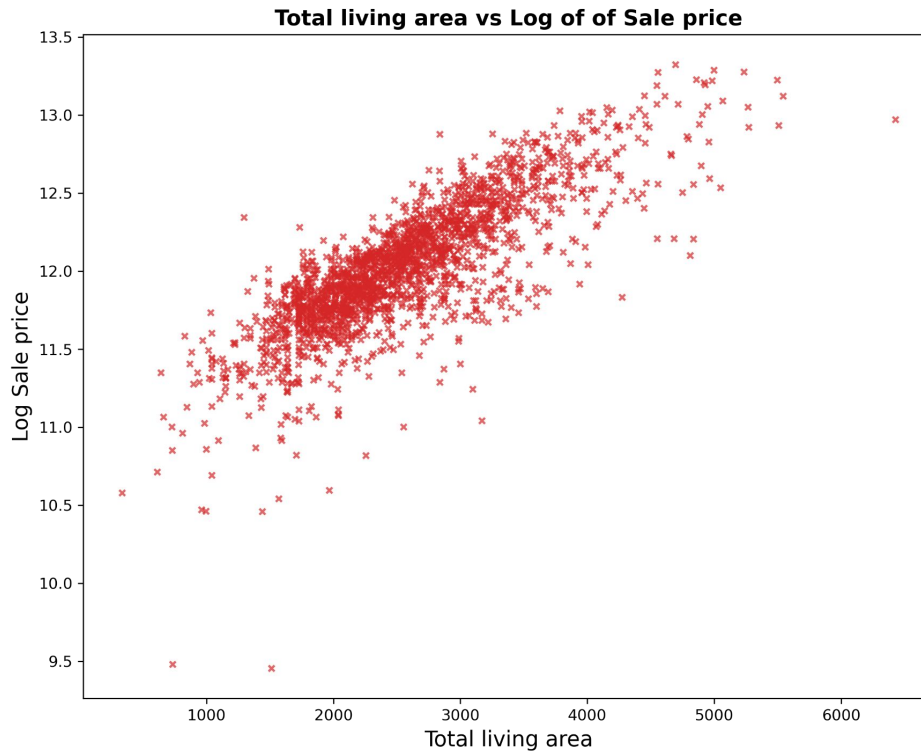


Start simple, then increase complexity

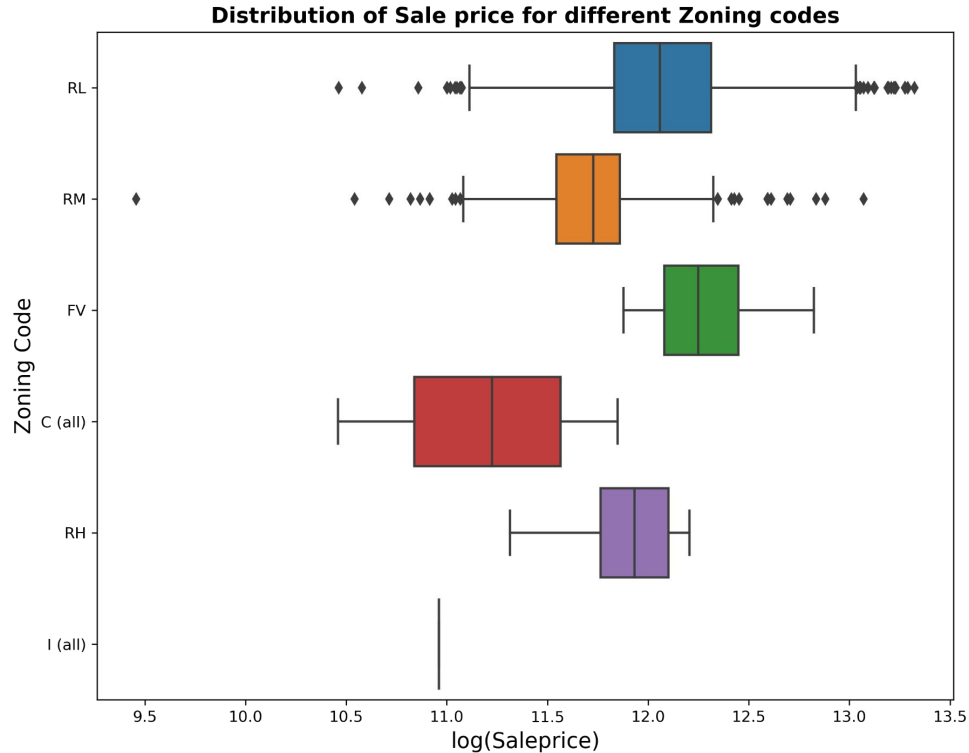
Most important:

Living area above + below ground:
'total living area'

Linear relationships to $\log(\text{Saleprice})$



Building a model



For **Categorical** data, I checked

1) Difference in sale price

2) How many are there in each category?

Model



Best model:

12 numerical features and
9 categorical features,
using linear regression

**How much does this improve over the
benchmark model?**

Benchmark versus best model R2-score:
0.64 vs 0.88

Benchmark versus best model RMSE:
\$44700 vs \$22100



Conclusions



- 1) Using the Ames dataset, I built a model to predict house prices in Ames Iowa using 12 numerical and 9 categorical inputs.
- 2) Model improves over benchmark, explaining more of the sale price variance (88%), and having having better predictions on average
- 3) Model depends on local data and time-related factors
- 4) The model could potentially work with even more complexity

Bonus slide 1: Data dictionary

Feature	Type	Dataset	Description	Encoding Info
ID	int	Ames data	Primary Key -The house ID in the Ames housing data set	
tot_area	float	Feature Engineered	The total living area in square feet	Sum of above-grade SF and basement SF
gar_area	float	Ames data	the Garage area in square feet	
lot_frontage	float	Ames data	the total amount of lot frontage in linear feet	
yr_built	int	Ames data	the year the house was built	
yr_remod	int	Ames data	the year the house was remodeled or added to	
qual	int	Ames data	The overall quality rating of the house	rated 1-10
tot_rooms_abv	int	Ames data	The total number of rooms above ground	
full_bath	int	Ames data	the total number of bathrooms	
mas_vnr_area	float	Ames data	the total surface area of masonry veneer	
fireplaces	int	Ames data	the total number of fireplaces	
cond	int	Ames data	The overall condition rating of the house	rated 1-10
gar_cars	int	Ames data	The number of cars that fit into the garage	
condit_***	category	Feature engineered	Special conditions that apply to the house	Dummified: 'Typ' rating is baseline. Up to two conditions are possible
MS Zoning_***	category	Feature engineered	Zoning code for the house (eg Residential, industrial)	Dummified: 'RL' is the baseline
Func_***	category	Feature engineered	Functionality for the house (eg. deductions, salvage)	Dummified: 'Typ' is the baseline
Neighborhood_***	category	Feature engineered	Neighborhood	Dummified: 'NAMES' is the baseline
Kitchen Qual_***	category	Feature engineered	Quality rating of the kitchen (good, poor, fair, etc)	Dummified: 'TA' is the baseline
Bsmt Qual_***	category	Feature engineered	Quality rating of the basement (good, poor, fair, etc)	Dummified: 'TA' is the baseline
Central Air_***	category	Feature engineered	Whether the house has central air conditioning or not	Dummified: 'Y' is the baseline
Alley_***	category	Feature engineered	Whether the house has an alley and what type	Dummified: 'NP' is the baseline
Paved Drive_***	category	Feature engineered	Is there a paved driveway	Dummified: 'Y' is the baseline
Yr Sold_***	category	Feature engineered	Year the house was sold	Dummified: '2006' is the baseline
Saleprice	int	Ames data	Target variable -the sale price in dollars	I use log(Saleprice) in the actual regression

Bonus slide 2: residuals

