Price Prediction: Ames Housing

A data-centric approach to price predictions

Background

Our team has been approached by a property consultancy firm to predict the sales prices of properties in Ames, Iowa. Their aim is to involve a data science perspective in formulating prediction as they plan to apply it to other areas as well.

Project Scope

Context	To derive predictions of sale prices of housing in Ames, Iowa
Requirements	Overcome the challenges of property valuation considering the multiple factors that affect the price
Vision	Using data science, build a model that objectively determines the price based on 2006 - 2010 data
Outcome	A model that accurately computes the housing prices as well as insights into considerations that impact prices

Stakeholders

Primary Stakeholders

Front-end Property Consultants

- Insights into parameters that strongly affect value
- Best advise clients on increasing the value of their properties

Secondary Stakeholders

Back-end Software Engineer

- Model best used to predict values
- Insights on outliers and how to factor that into model building

Process Flow

Data cleaning	Feature Engineering	Exploratory Data Analysis	Model Selection & Evaluation
• Validate Data • Removal / imputation for null values	• Work on ordinal object variables	• Correlation analysis • Graphical review of variables	• Construct regression models • Evaluate models and select • Key findings

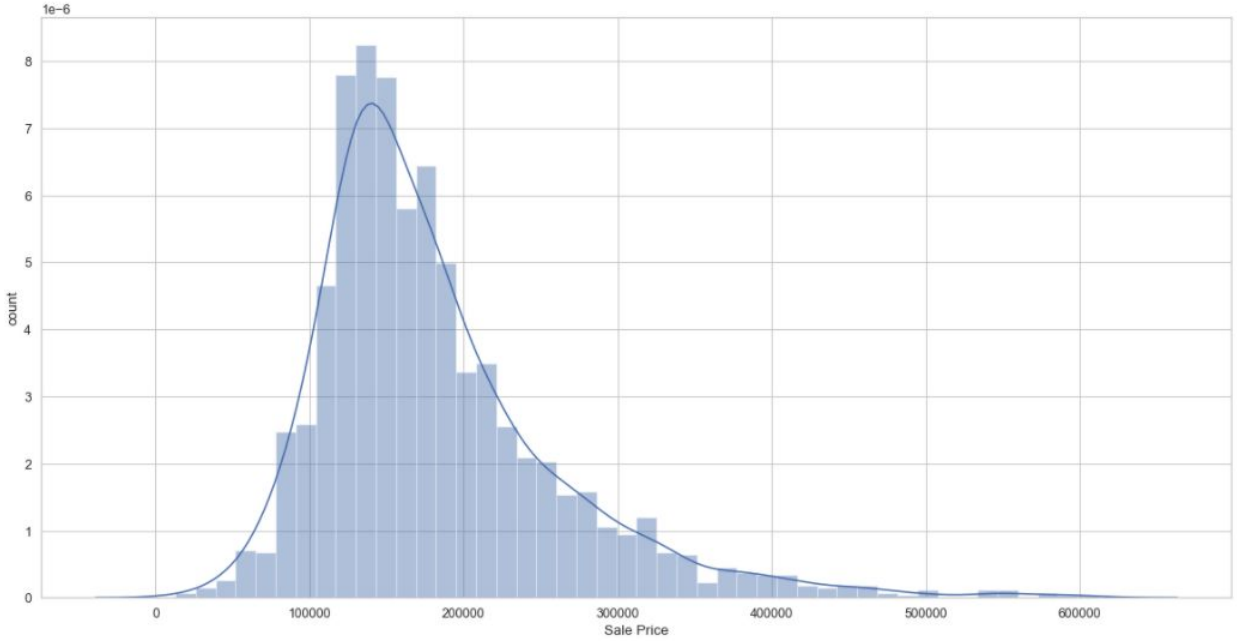
Analysis & Insights from 2006 to 2010

Our hypothesis

We assumed that location, size and overall condition of the property are the most important factors

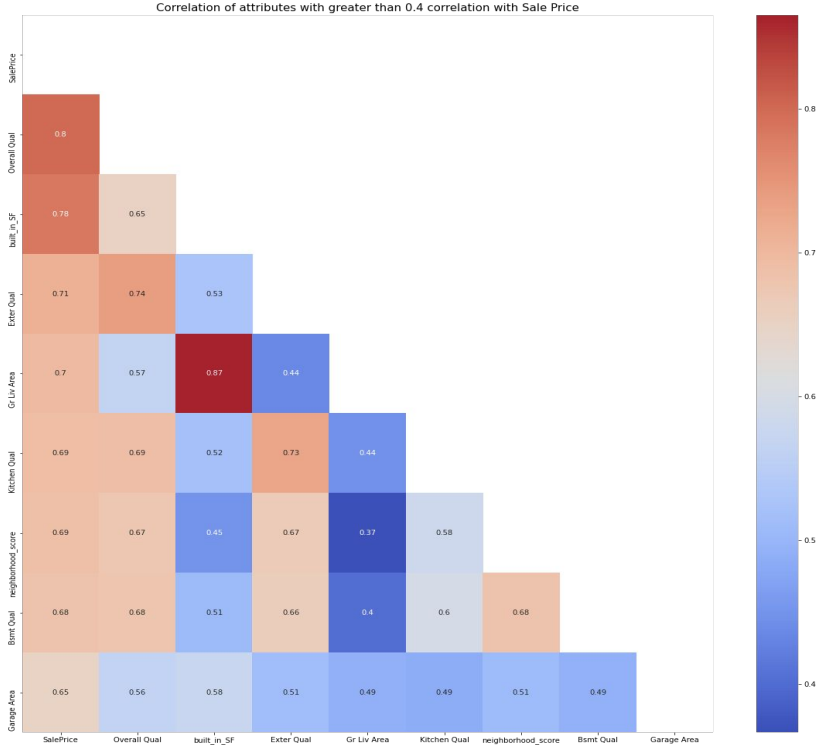
- Built-in square feet (total living areas)
- Scoring of location by neighborhood
- Overall Quality (already within data)

Sale Price Distribution



- Unimodal
- Right Skewed
- Median price of 130,000

Strongest Correlation of Variables with Sale Price



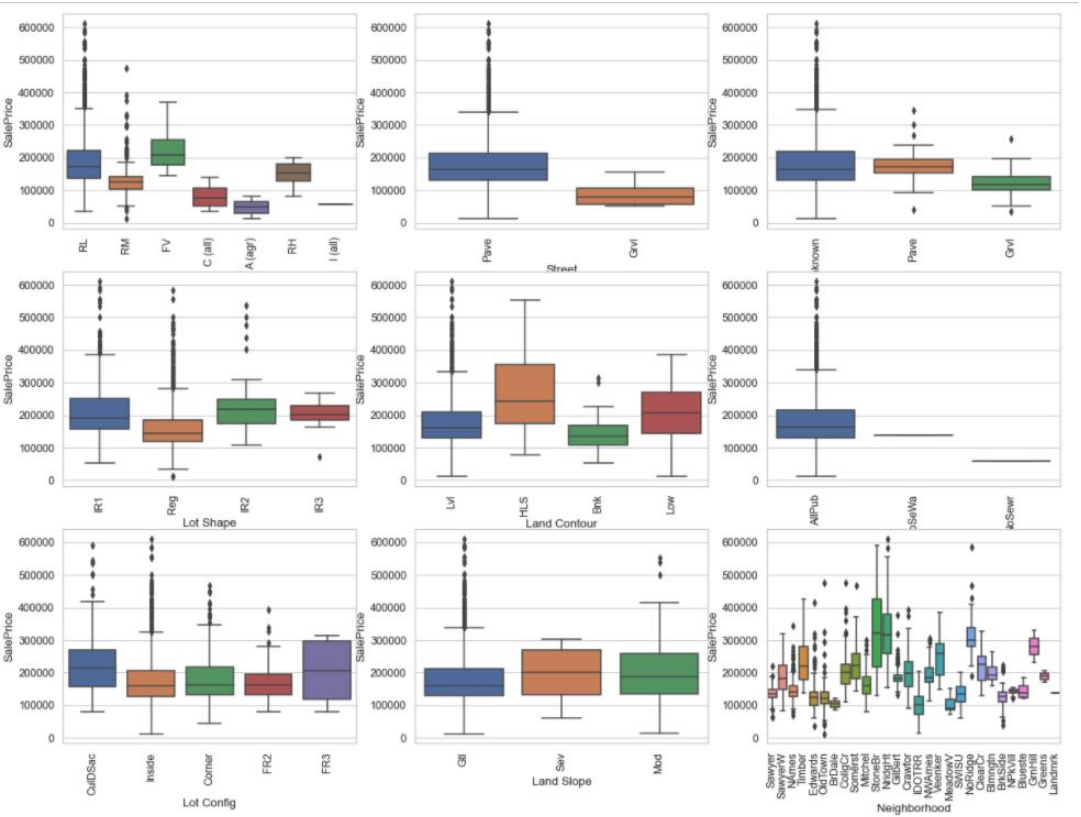
- Heatmap of variable with strongest correlation with Sale Price

  1st - Overall Qual

  2nd - Built in SquareFeet (engineered feature)

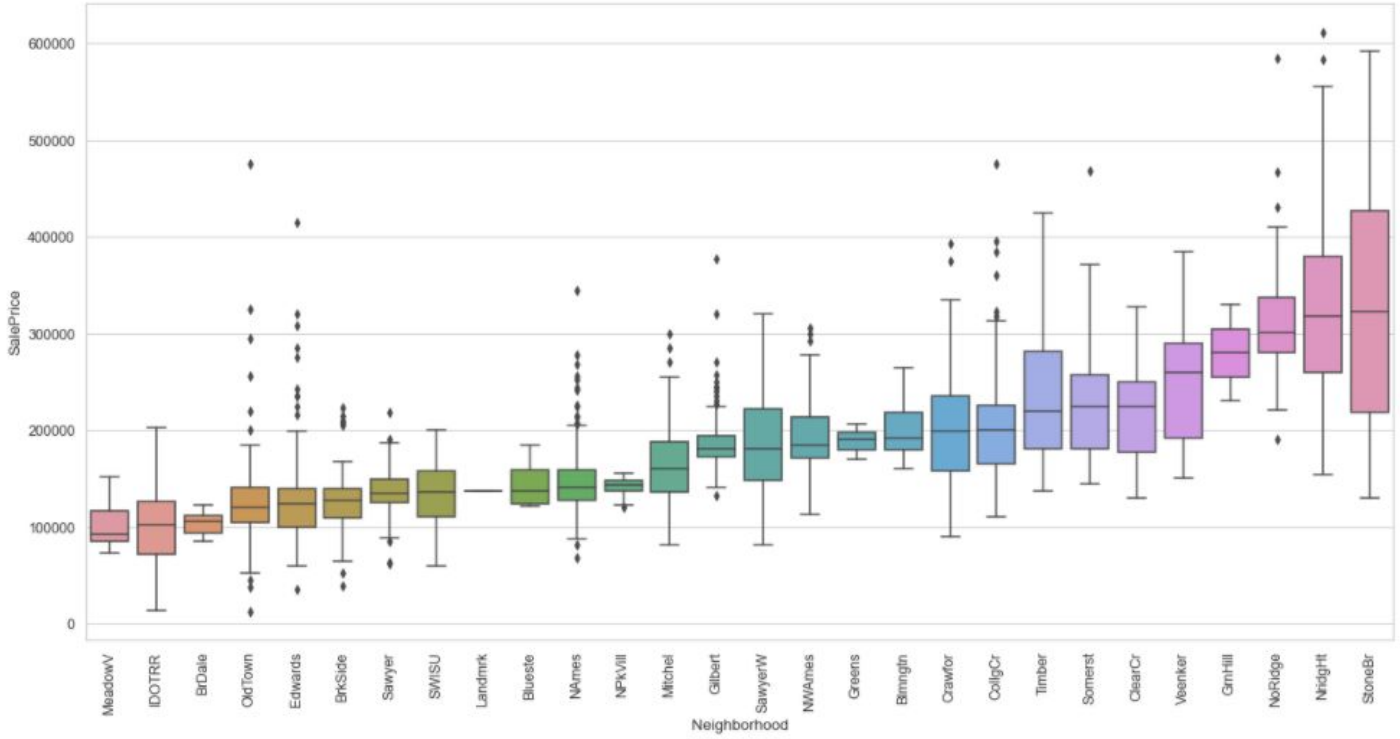
  3rd - Neighborhood score

Box Plot of Categorical Data with Sale Price



- Different distribution within each subcategories
- Neighborhood has the most complex distribution

Neighborhood Sorted by Median Sale Prices



- Cheapest neighborhood - median price of 100,000
- Most expensive neighborhood - median price of >300,000
- Variation in dispersion depending on neighborhood

Model Selection/Evaluation

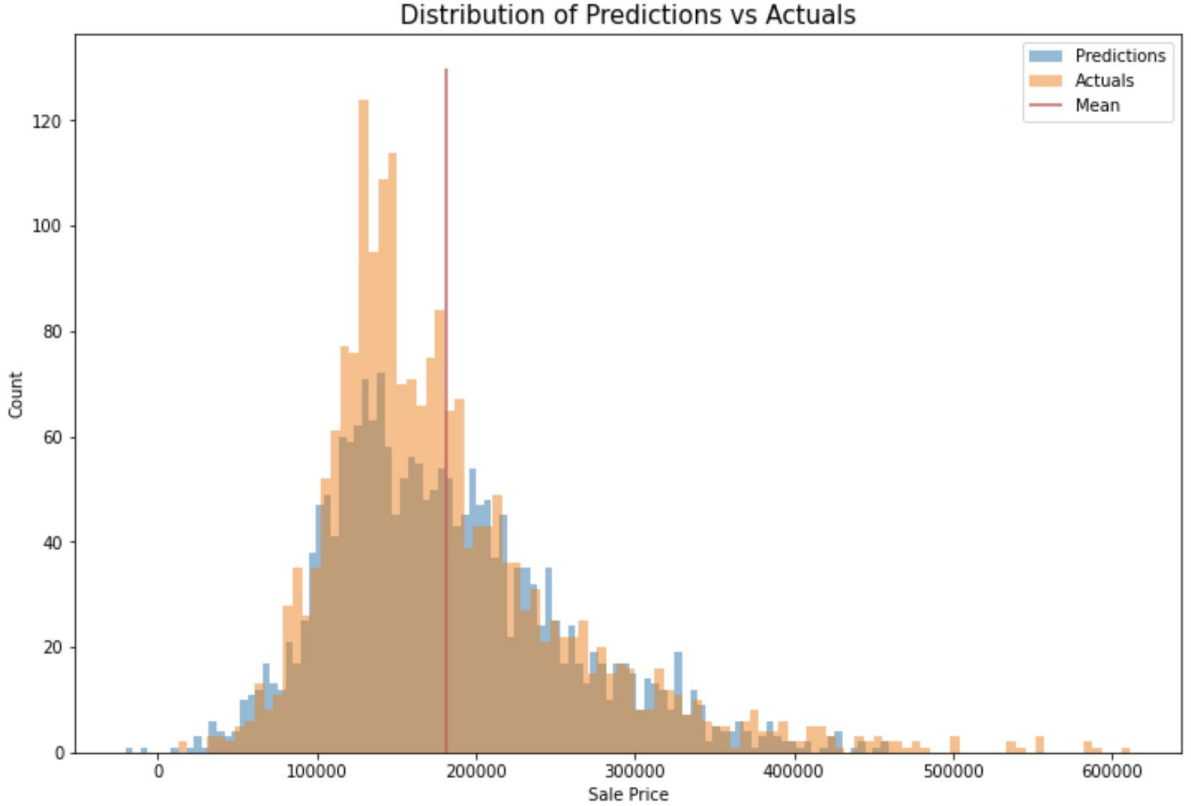
Lasso model selected

- Lasso eliminates many features, and reduce overfitting in the linear model.

RMSE

- Root Mean Square Error shows how concentrated the data is around the line of best fit.
- The lower the score, the better the model fit
- Model achieved low RMSE score

RMSE of 21,024



Key Takeaways & Recommendations

Key Takeaways: Valuing a Property

How to objectively value a property in Ames

- Location
- Size, specifically built-in area
- Overall Quality

Built-in price per sq ft ($/sq ft)	Neighborhood
50 to 55	Iowa DOT and Rail Road, Meadow Village, Old Town, South & West of Iowa State University
55 to 60	Brookside, Edwards
60 to 65	Briardale, Mitchell, North Ames, Northwest Ames, Northpark Villa, Sawyer
65 to 70	Bloomington Heights, Bluestem, Clear Creek, Landmark, Sawyer West
70 to 75	College Creek, Crawford, Gilbert, Greens, Timberland, Veenker
75 to 80	Northridge Heights, Stone Brook
95 to 100	Green Hills

Key Takeaways: Way to Increase Value

Given the main factors affecting value, what advice can be given to clients to **maximise** the value of their property?

Key Takeaways: Way to Increase Value

External Facade

Kitchen

Basement

Modeling Recommendations:

- Create features based on aggregated size such as built-in area
- Ordinally rank locations based on neighborhoods
- Lasso model useful in dataset where there are many parameters provided

Limitations

- Location metric relies on historic data.
- *“What will the value of my property be in 10 years?”*
- Can it be replicated?

Conclusion

- **Location**, **size**, and **house quality** are the main factors
- People value **first impressions**, a nice **kitchen**, and a nice **basement**
- Model needs to create features centered around the 3 main factors
- Model does not take **time** into account
- Location score should not be a by-product of price

Thank you!