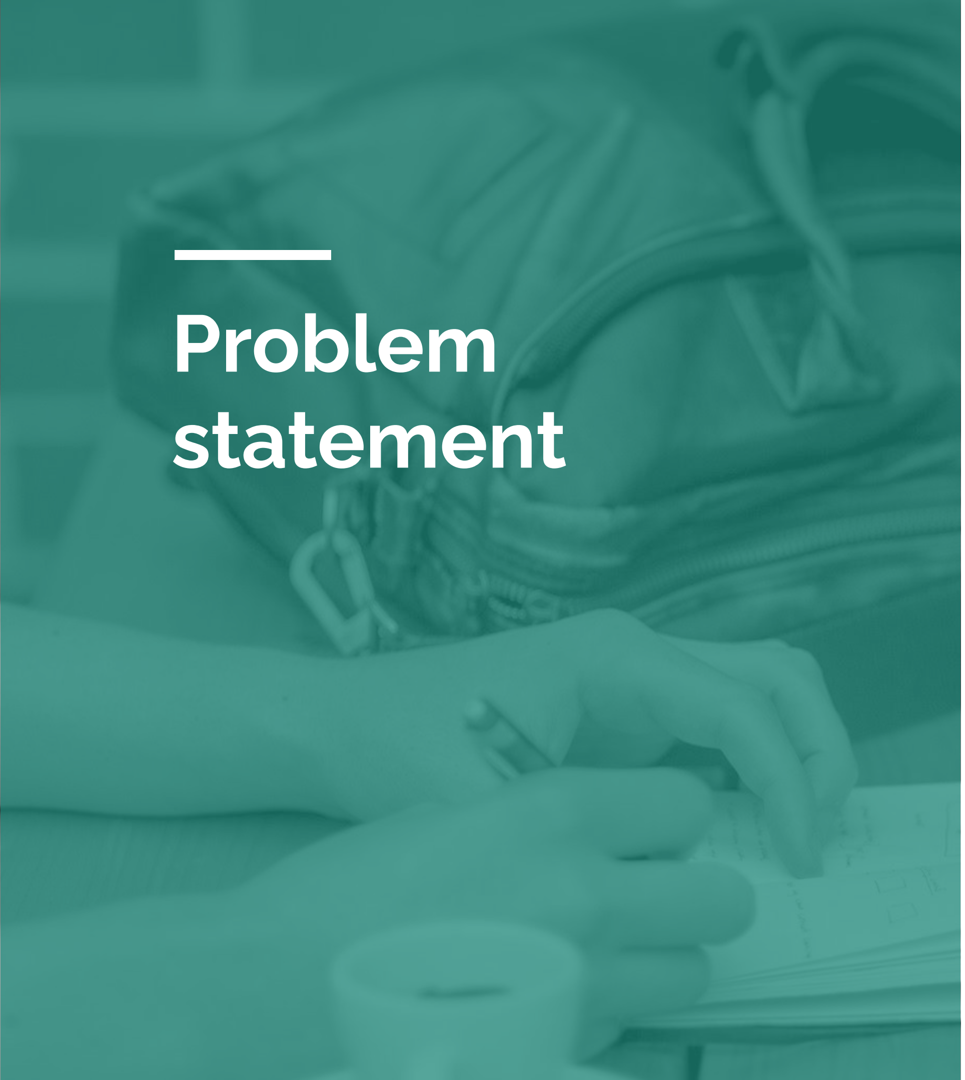
Project 2 Ames Housing Data

Roman Tedeschi



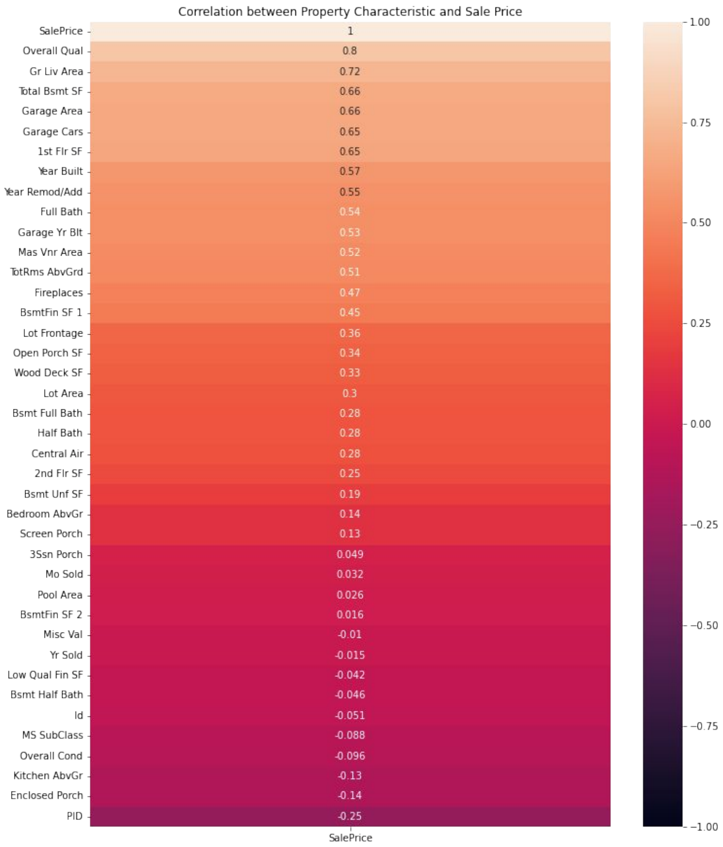
Problem statement

Personal property tax revenue is very important for municipal operations in Ames. Citizens need to have confidence that they will be taxed accurately and fairly, and the city of Ames cannot waste unnecessary resources on determining this assessment every year.

Summary

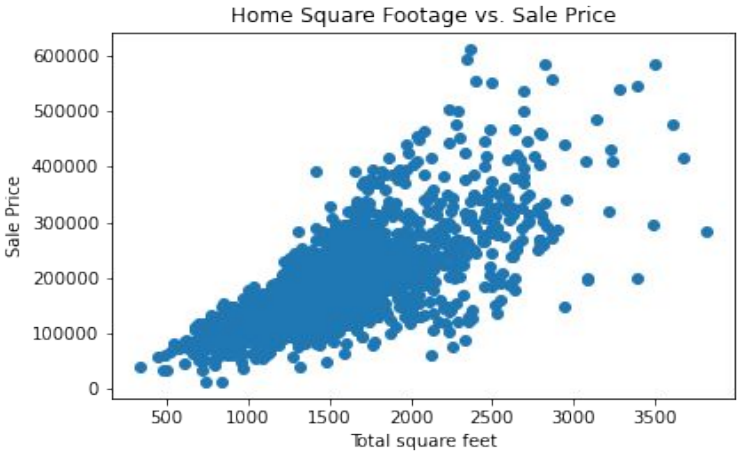
- Data analysis
- Exploration of different models
- Using metrics to find the most appropriate model
- Conclusion

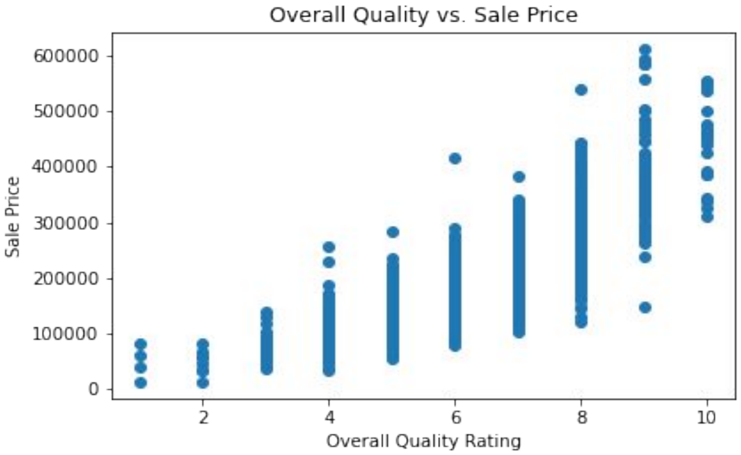
Exploratory Data Analysis



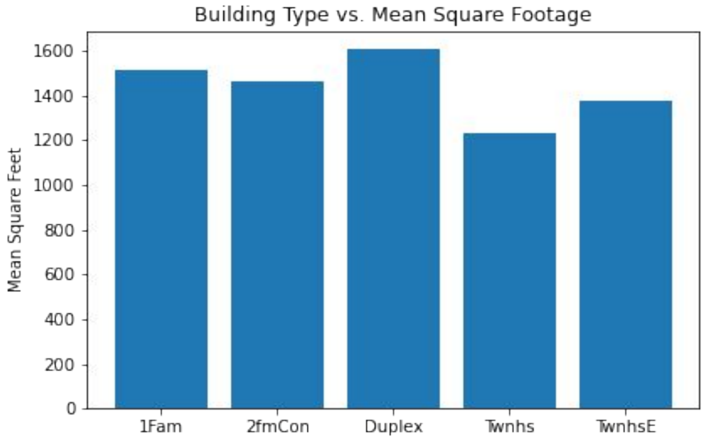
** heatmap taken almost directly from a lesson taught by Katie Silvia*

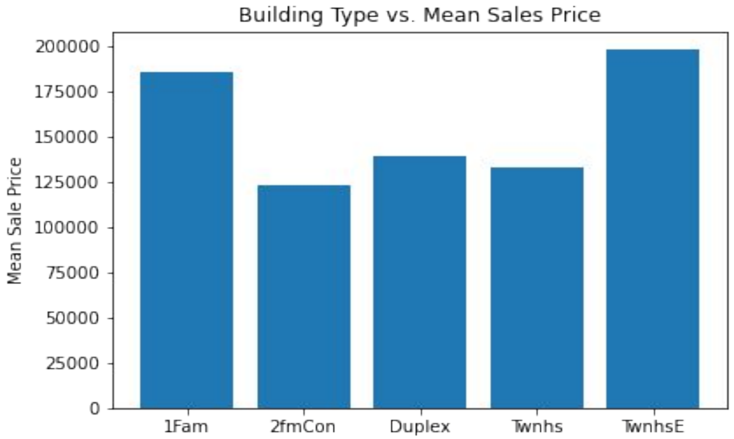
Correlation between quantitative variables





Correlation between categorical variables





Model selection

- ➔ First attempt was a simple linear regression with only two independent variables
- ➔ Second attempt added simple feature engineering
- ➔ Third applied those same features to a ridge regression
- ➔ Lastly, a LASSO was applied and additional feature engineering performed

Analysis metrics

The main metric used to test fit was r^2 (coefficient of determination).

This score, ranging from 0 to 1, reflects the proportion of the variance in Sales Price that can be predicted from the independent variables chosen in that model.

Sample R^2 Scores (train / test)

Linear with two variables: 0.75 / 0.73

Linear with engineering: 0.81 / 0.82

Ridge: 0.82 / 0.76

Final LASSO: 0.92 / 0.93

Chosen model description

The final accepted model is a LASSO that includes a total of 49 features.

It includes several categorical variables one hot encoded into multiple columns. These make up approximately half of the total features and include building type, central AC, and neighborhood.

The other features are quantitative and are generally taken from the highest correlated variables as seen in the heatmap shown previously.

Conclusion

Using r^2 values, multiple models can quickly be designed and tested for accuracy on our dataset. We were easily able to increase model complexity to reach an r^2 value of 92%. As this is being used solely for tax assessment data, this value is within acceptable levels. It would be recommended to round each property's assessed value to the nearest $1,000 or $5,000 for the sake of simplicity and to take out any arbitrary variance in the model which is not accurately reflected in the property data itself.

Conclusion (legal)

As any tax is subject to legal and Constitutional challenges, it should be safe to allow public viewing of this algorithm in the case of any suit or FOIA request. The categories themselves are numerous and were not motivated by any real or perceived bias - rather, they were chosen almost completely on the strength of correlation statistics. Additionally, the weight of each of these numerous categories was done entirely by automated LASSO regression and is consistent with real-life sales figures.

Conclusion (next steps)

Additionally, this model is easy and inexpensive to maintain. With slight adjustments - and depending on the exact format of data-keeping - it can easily be expanded to other similar markets. The main challenges will be checking for incomplete data (which could mess up the algorithm without manual observation and correction) and the varying types and complexities of neighborhood organization and importance.

Questions?