# Predicting Housing Prices

Stephen Chen, Jenny Conde, Andy Tertzakian
W207 Fall 2021

# Project Steps

1. Exploratory Data Analysis
2. Feature Engineering
3. Linear Models
4. Nonlinear Models
5. Final Model Selection
6. Submit to Kaggle Competition

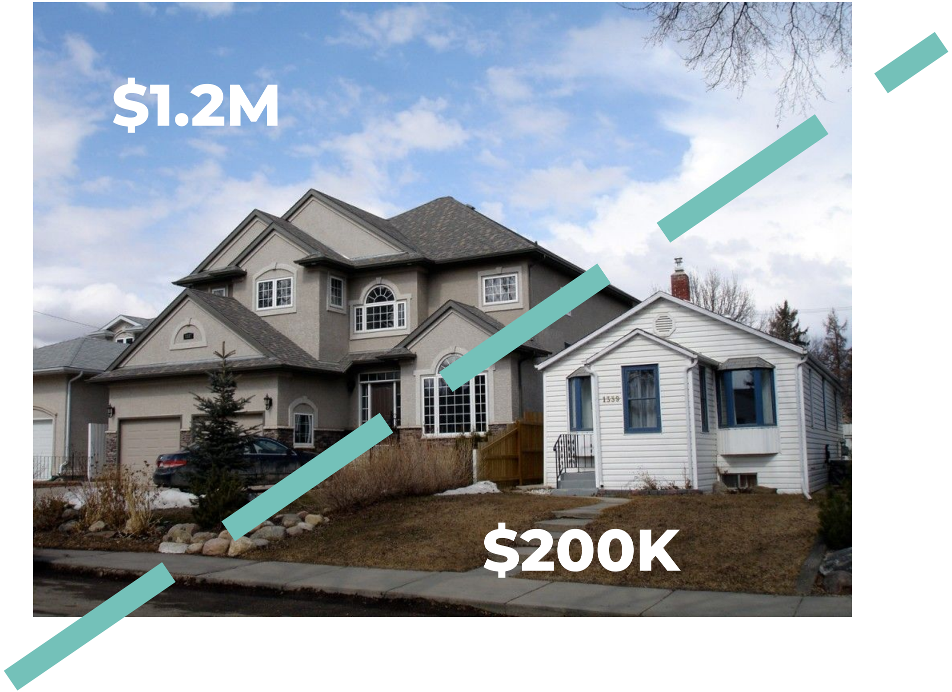
# Intro & Context

- Know your dream house?...

  ... predict the price!

- Want to make money?...

  ...play the market!

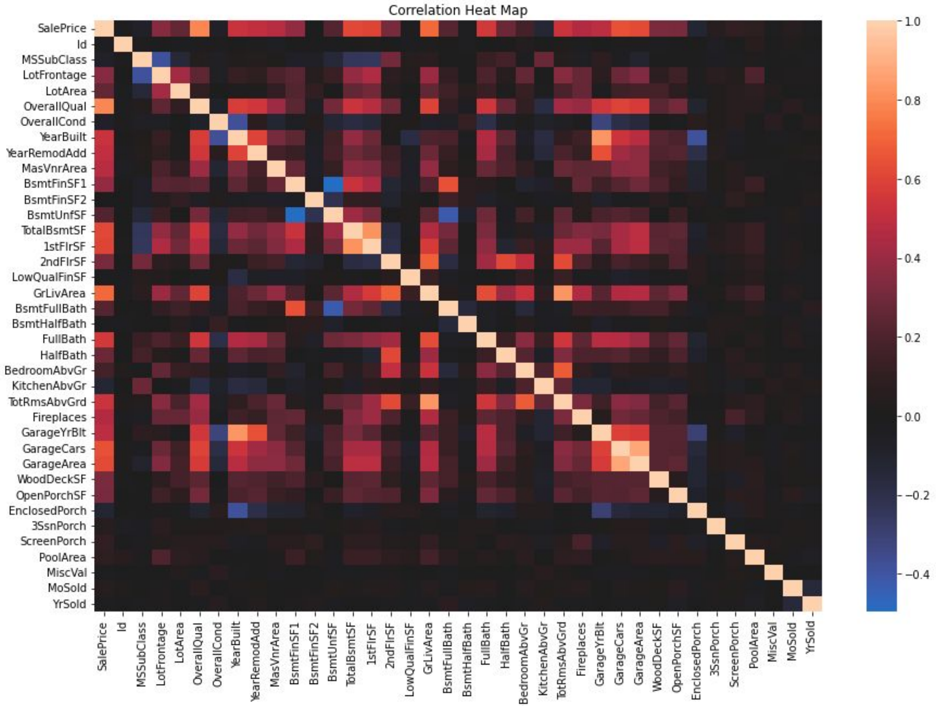


# What's in the Data?

- 80 variables containing thousands of property sales made in Ames, Iowa between 2006 - 2010
- Most variables are pertinent information for a typical home buyer who would want to know more about a potential home
- Continuous variables
    - Lot size
    - Square footage inside the home
- Discrete variables
    - Number of kitchens, bedrooms
- Nominal variables
    - Types of materials used
- Ordinal variables
    - Quality of materials

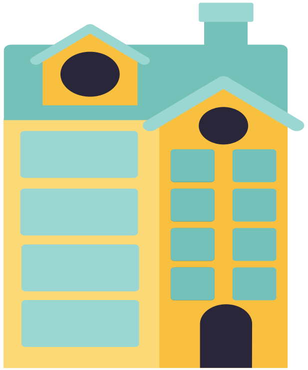
# EDA

- We want to be predicting SalePrice based on features of previously sold houses
- SalePrice has a mean of $180,921 but distribution is very skewed (log-transform helps make distribution look more normal)
- Highly correlated variables with SalePrice: OverallQual, TotalBsmtSF, 1stFlrSF, GrLivArea, GarageCars, GarageArea
- Handling outliers



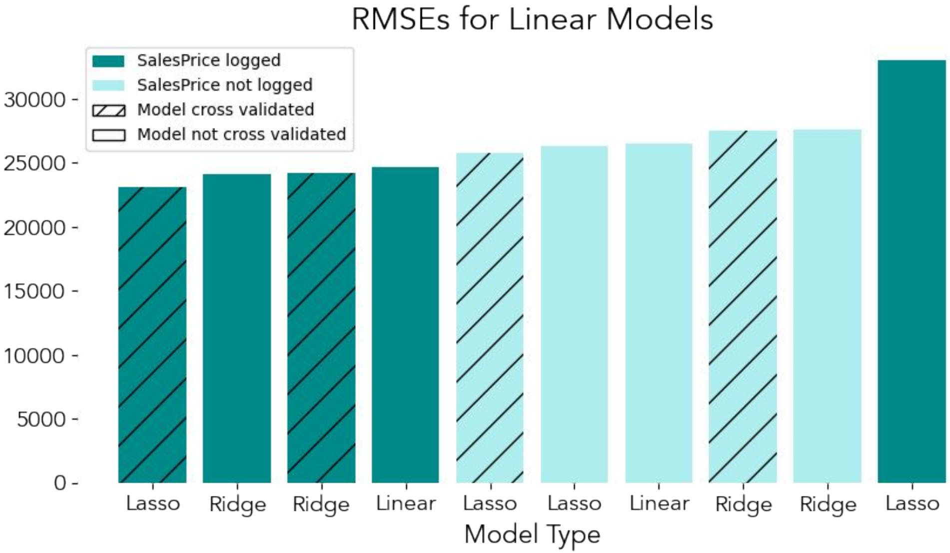
# Feature Engineering

- Lots of missing values - resolved after consulting documentation
  - Quantitative: NA -> 0 if feature that is measured wasn't in the property (ex. porch)
  - Qualitative: NA -> None, feature does not exist
- Normalized quantitative variables
- Many of our variable are qualitative, so we applied one-hot encoding to the qualitative variables so that we could include qualitative data in regression models
- Experimented with creating new variables by combining similar variables together
  - Total Surface Area = Basement + 1st Floor + 2nd Floor
  - N bedrooms above ground per 1000 sq ft above ground



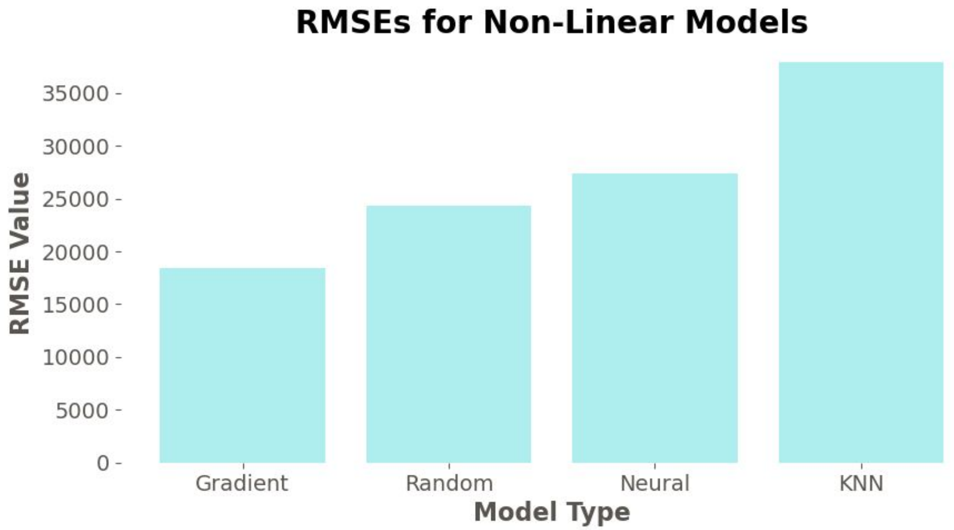
# Linear Models

- Train-dev-test split
- Linear Regressions with and without penalty terms
- Used Cross Validation to find optimal hyperparameters
- Varied SalePrice transformation
- Tested on many random seeds to confirm results
- Best Model: Lasso Regression on log(SalesPrice), cross validated
    - Dev RMSE = 23,282




# Nonlinear Models

- Random Forest
  - GridsearchCV for 2 Hyperparams
  - Best RMSE: **22,493**
- KNN
  - Iterated over 4 Hyperparams
  - Best RMSE: **37,908**
- Neural Network
  - Played with Tensorflow
  - Best RMSE: **27,422**
- Gradient Boost
  - Iterated over 5 Hyperparams
  - Best RMSE: **18,390**




# Conclusions

Key Takeaways

- Gradient Boosted model had best performance on dev and test data
- Important to try a wide variety of models
    - Ex. Nonlinear models performed significantly worse using log(SalePrice)

Kaggle Submission

- Gradient Boosted model had best performance on Kaggle
    - RMLSE of 0.129 (72nd percentile, 1417/5194)
- Neural Networks and KNN had worst performance

Ideas for Future Work

- Look more into outliers
- Work more with Neural Networks
- Additional feature engineering?
- Generalizability to other housing markets and times

# Questions?

# THANKS