Prediction of Housing Prices

Problem

Social benefits of homeownership:

- Happiness and prosperity
- Civic participation
- Education

What do prospective buyers look for? Not just price, but also:

- Living area
- Number of bedrooms
- Number of bathrooms

Main Clients

Homebuyers

- How much house they’d get for their money
- More informed decisions

Businesses

- Greater stability

Government officials

- Policies that improve housing affordability

About the Data

- The data comes from Kaggle
- Collected in 2011
- 80 variables
- 2,390 properties in Ames, Iowa.

Overview

The steps involved in this analysis include:

- Data cleaning and wrangling
- Feature engineering
- Preprocessing: scaling, one-hot encoding
- Exploratory data analysis
- Machine learning

Steps in the Analysis

- Data Cleaning
  - Isolate useful values and rename them
  - Convert values such as “poor” and “good” to numeric
  - Replace “nan” values with zeros
- Feature Engineering
  - Calculate age: 2011 - year built
  - Combine half and full bathrooms
  - First floor + second floor + basement = overall living area

Steps in the Analysis (cont.)

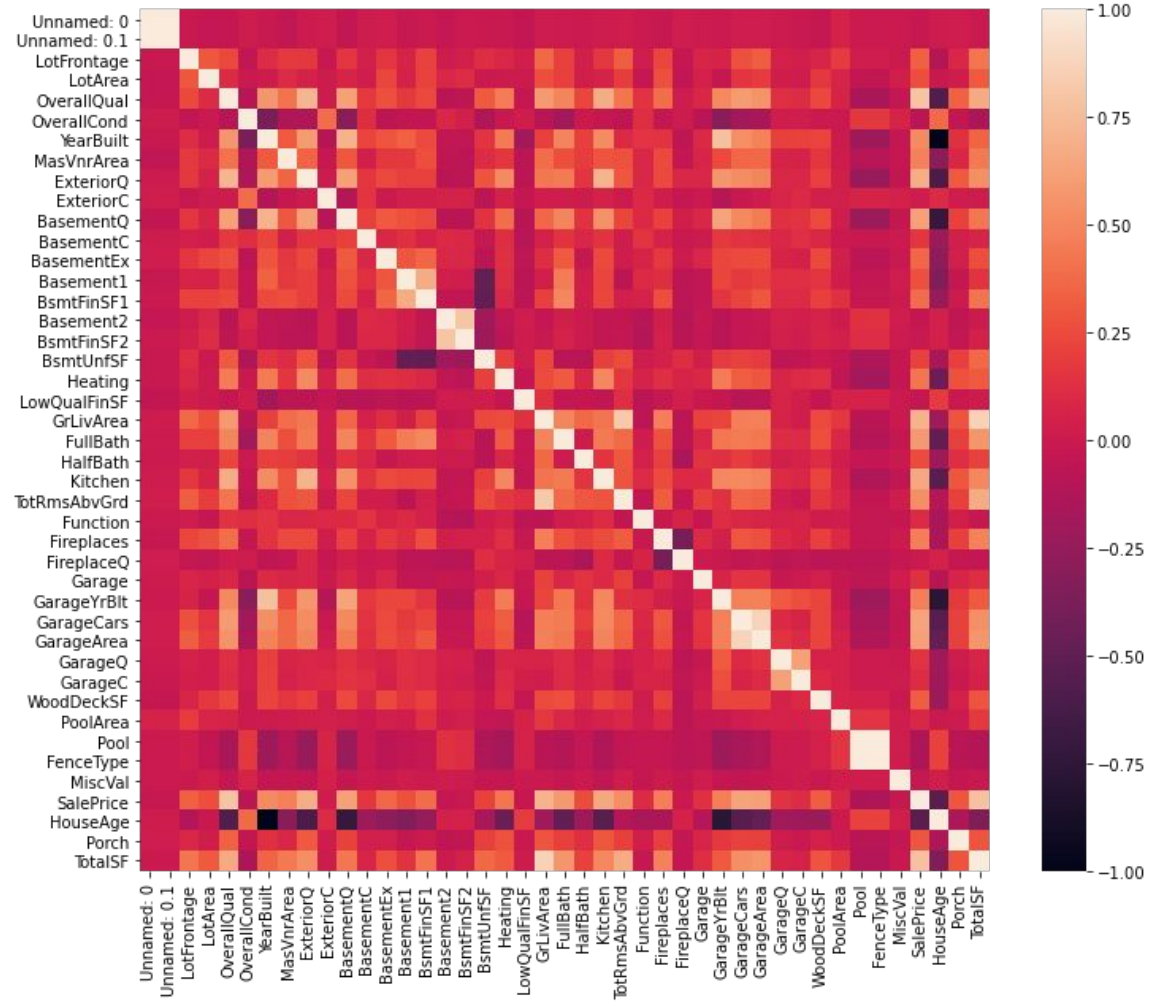
- Preprocessing
  - Scaling
  - One-hot encoding
- Exploratory Data Analysis
  - Checking for correlations with heatmap
  - Plotting sale price against other variables
  - Hypothesis testing

Steps in the Analysis (cont.)

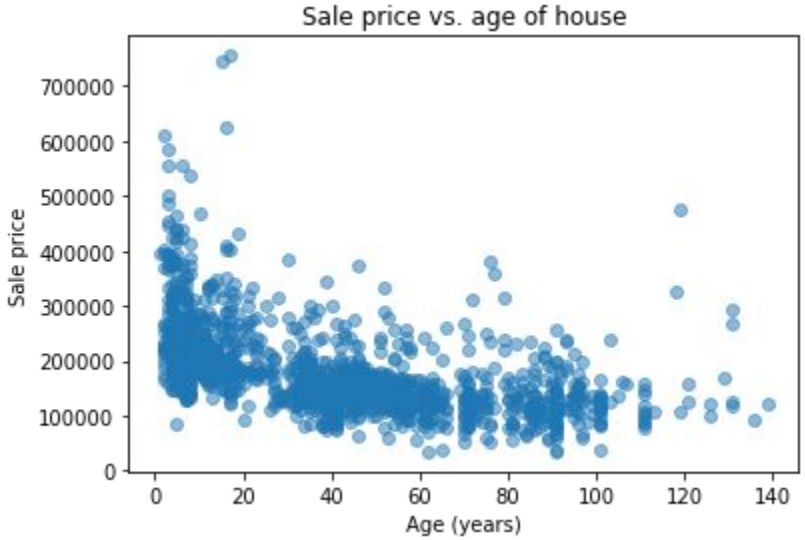
- Machine Learning
    - Run regression analyses using test-train splits

Exploratory Data Analysis

From the correlation heat map, the variables with the most positive correlations are overall quality, kitchen, number of cars that can fit in the garage, and garage area. The variable negatively correlated the most with sale price is the age of the house.



Exploratory Data Analysis (cont.)

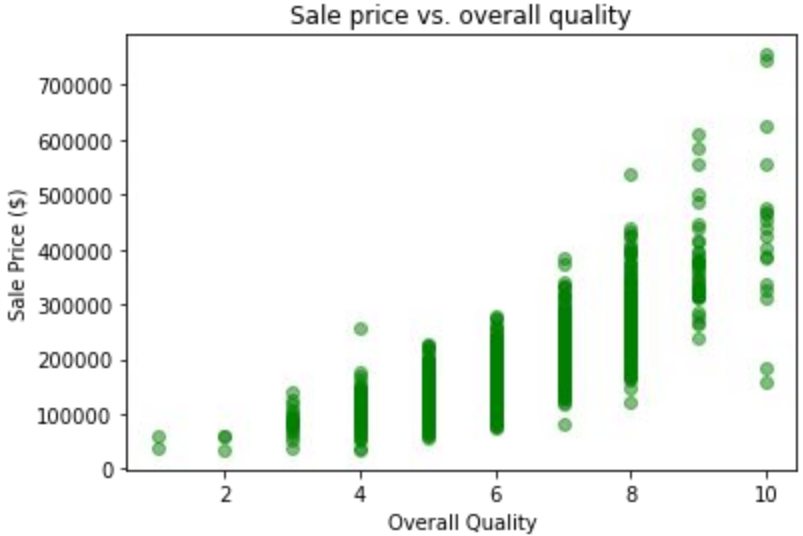


Plotting the age of the house against the sale price. Age of the house is negatively correlated with sale price.

Exploratory Data Analysis (cont.)

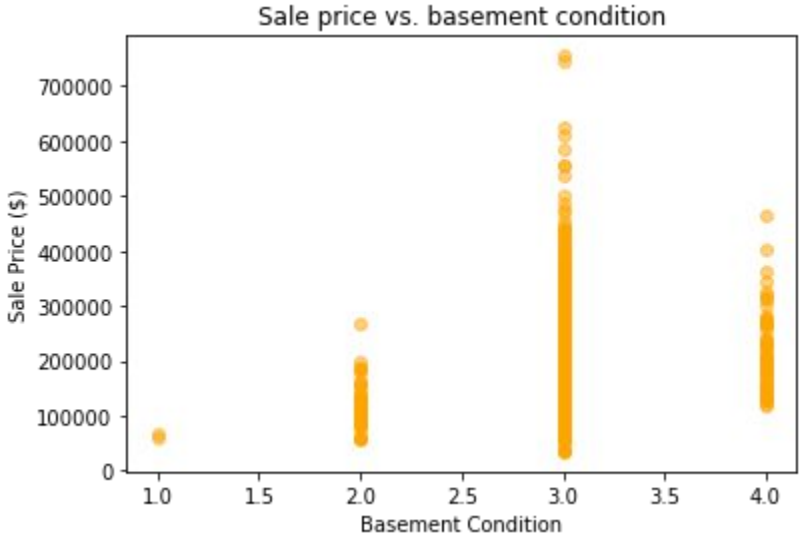
Plotting sale price against greater living area A strongly positive correlation can be seen between the living area and sale price.

Exploratory Data Analysis (cont.)



Plotting overall quality with sale price. Quality is positively correlated with sale price.

Exploratory Data Analysis (cont.)



Plotting basement condition against sale price. Houses with higher sales prices had basements with condition values of 3.

Machine learning

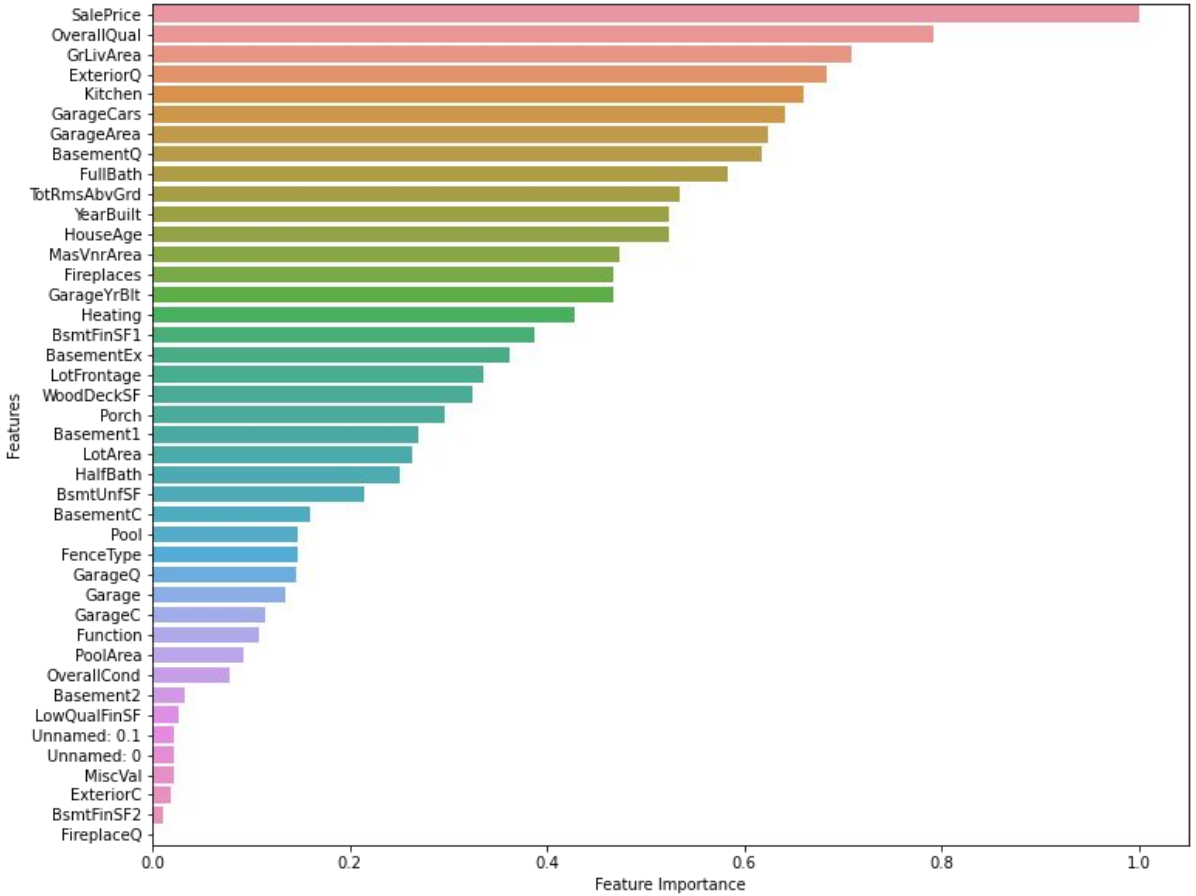
Comparing different regression models:

- Linear Regression
  - R-squared: 83.6%
- Ordinary Least Squares
  - R-squared: 80.0%
- Random Forest Regressor
  - R-squared: 90.0%
- SVM
  - R-squared: -0.03%

Feature Selection

Strongest features:

- Overall quality
- Living area
- External quality
- Kitchen
- Garage area



Conclusions

Most important features

- Overall quality
- Living area
- Exterior quality
- Kitchen
- Garage area

Conclusions (cont.)

Other external factors can influence housing prices:

- Geography
- Population
- Crime rates
- Proximity to schools

Future Recommendations

- Dataset >10 years old, so more recent data may help.
- Data to model after coronavirus pandemic