Project 2: Ames Housing Data and Kaggle Challenge

Author: Artem Lukinov

Problem Statement

There are many variables that can affect house prices. Using Ames Housing Data from Kaggle, let's determine the most influential variables and build the best model to predict house sale prices.

What was looked at?

train.csv - all of the training data for the model

test.csv - the test data for the model. This data does not have a Sale Price column that needed to be predicted

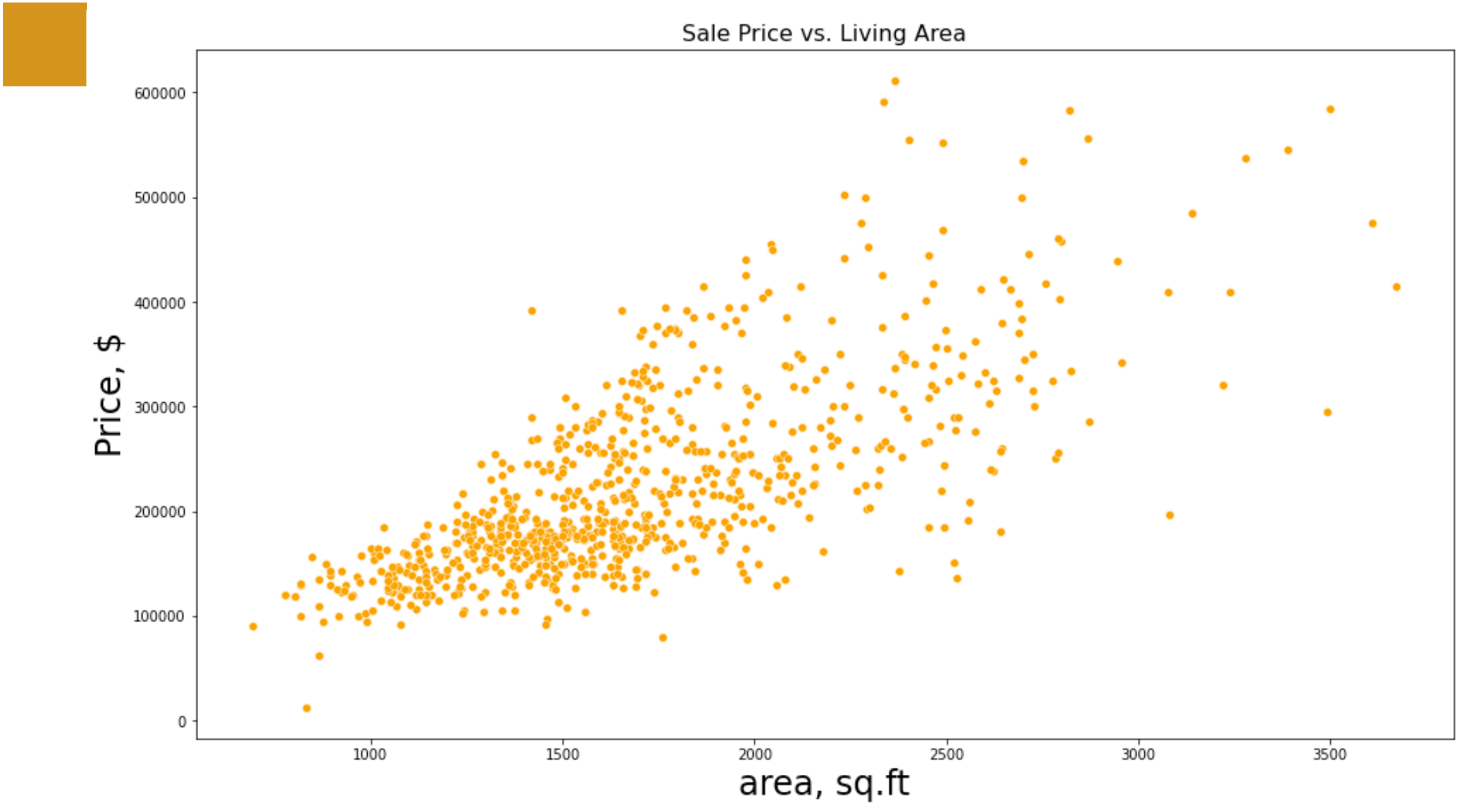
How was the data cleaned?

- column names were renamed for consistency
- columns with more than 60% missing data were dropped
- all missing values were cleaned by replacing with np.nan
- rows with data not containing residential transactions were dropped

Introducing new columns

- Finished Basement Area =
  Total Basement Area - Unfinished Basement Area
- Livable Space =
  Finished Basement Area + Ground Living Area
- House Age =
  Year Sold – Year Built


Sale Price Distribution
1e−6
Density
0
1
2
3
4
5
6
7
8
0
100000
200000
300000
400000
500000
600000
700000
Price, $


Sale Price vs. Living Area
Price, $
600000
500000
400000
300000
200000
100000
0
1000
1500
2000
2500
3000
3500
area, sq.ft

Strongest correlations

Negative:

Home_Age	-0.542197
PID	-0.263199
Overall_Cond	-0.208291
Enclosed_Porch	-0.166815
MS_SubClass	-0.115623
Bsmt_Half_Bath	-0.114152
Id	-0.073480
Kitchen_AbvGr	-0.067277
Misc_Val	-0.027985
Low_Qual_Fin_SF	-0.025623

Positive:

Overall_Qual	0.808871
Liv_SF	0.750161
Garage_Area	0.716803
Gr_Liv_Area	0.713988
Garage_Cars	0.704219
Total_Bsmt_SF	0.695277
1st_Flr_SF	0.665168
Mas_Vnr_Area	0.565146
Garage_Yr_Blt	0.555269
Year_Remod/Add	0.552506

Identifying the features

Data Dictionary

Column Name	Description	Data type
Garage_Area	Size of garage in square feet	float64
TotRms_AbvGrd	Total rooms above grade (does not include bathrooms)	int64
Liv_SF	Total living area (includes finished basement)	float64
Home_Age	Age of the house (either from year built or year remodeled)	int64
Full_Bath	Full bathrooms above grade	int64
Garage_Cars	Size of garage in car capacity	float64
Garage_Area	Size of garage in square feet	float64
Overall_Qual	Rates the overall material and finish of the house	int64

Model effectiveness

- OLS regression score: 84.5% on the train data
- Cross-validation score: 84% on the train data
- RMSE after Kaggle submission: 37002.80

Conclusions and recommendations

- The linear model created can accurately predict the sale price
- There need to be further tweaks to the model
- Further feature engineering, regularization and testing different models is recommended
- As of now, the project helped to learn how to implement the basic functionality of linear regression.

Thank you

Q&A