PREDICTING HOUSING PRICES WITH PYTHON

Ben Moss

Data Science Immersive Program

General Assembly

January, 2024

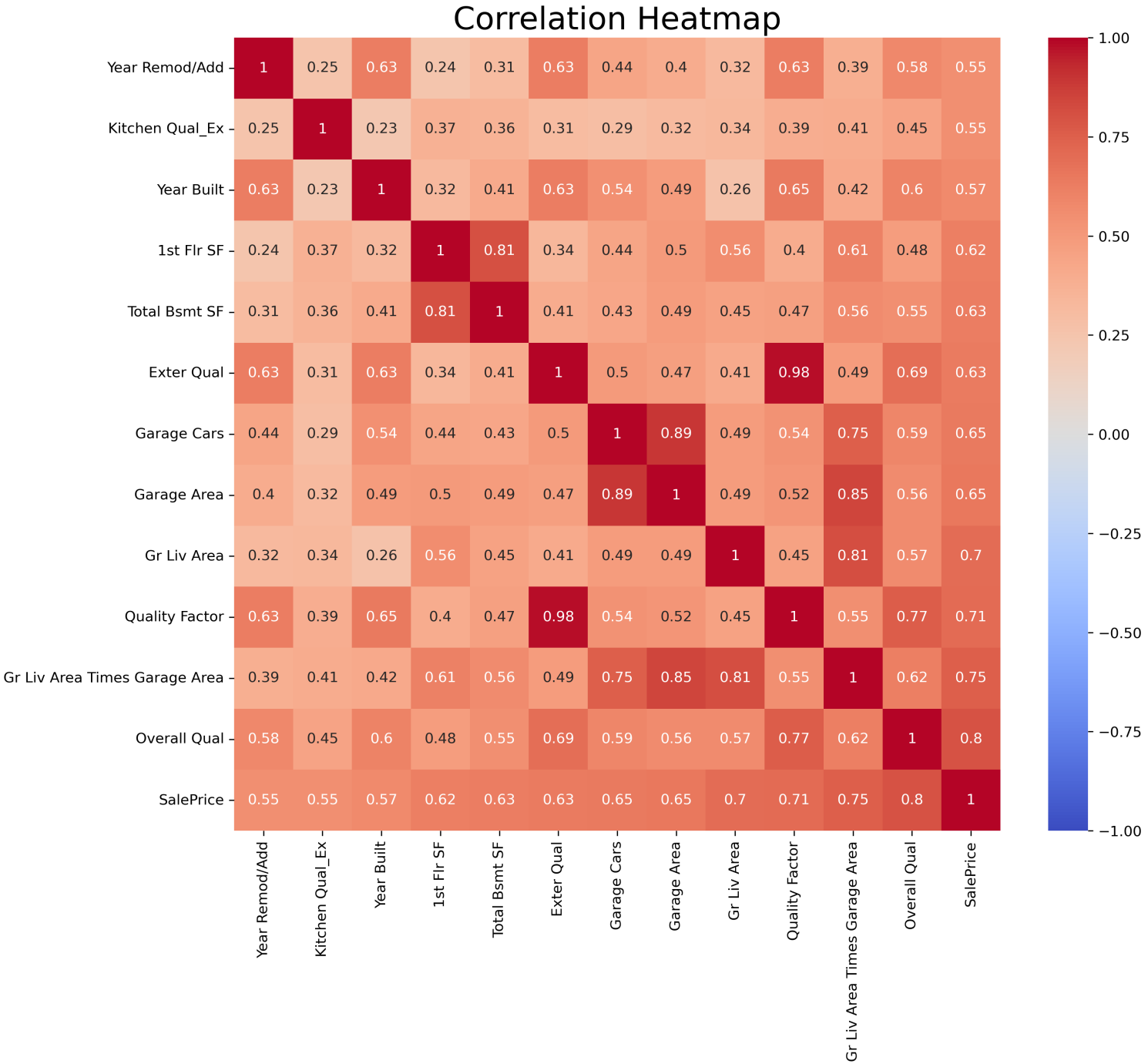
INTRODUCTION

- Two Datasets; train.csv and test.csv
- Objective: Effectively utilize Data Munging, EDA, and modeling techniques to predict the house sales prices from an unknown dataset.
- Identify key variables that play the biggest role in predicting housing prices.
- Provide recommendations to those considering their first time home purchase.

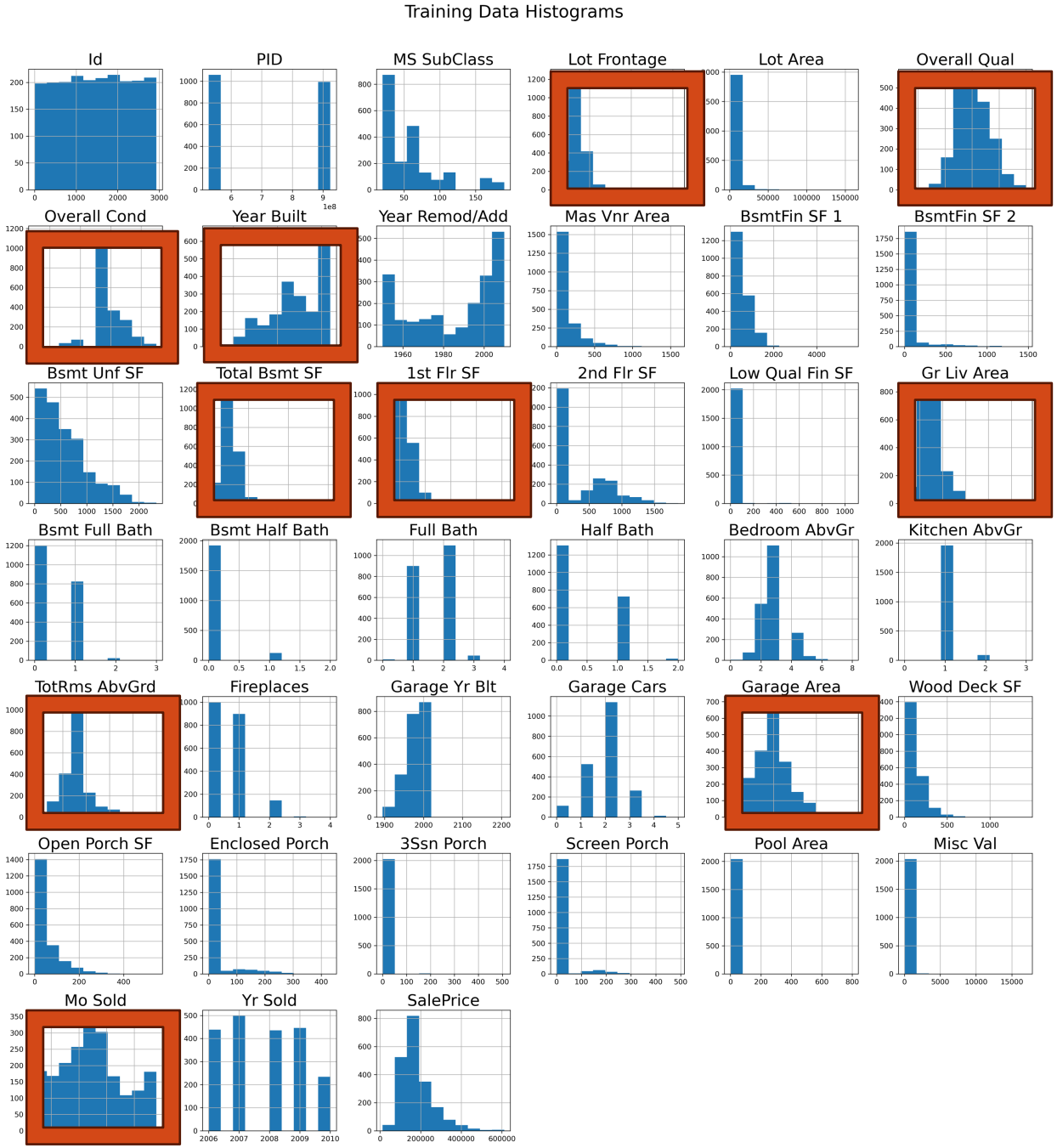
MODELING PIPELINE

- Data Munge; filling NaNs with median values.
- EDA: Heatmaps, histograms, pairplots to discover:
  - Which numerical variables were correlated closely with sales price?
  - Which numerical variables were normally distributed?
  - Which numerical variables had a linear correlation with minimal heteroscedastacity?
- Created Dummy variables and Feature variables that based on commonplace parameters that affected sale price (exterior condition, year build and remodeled etc.)

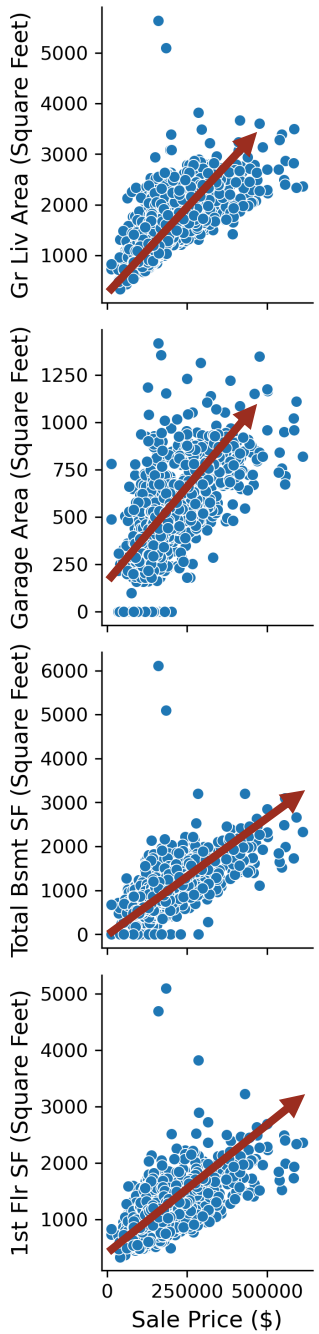
HEATMAP:


Correlation Heatmap
Year Remod/Add 1 0.25 0.63 0.24 0.31 0.63 0.44 0.4 0.32 0.63 0.39 0.58 0.55
Kitchen Qual_Ex 0.25 1 0.23 0.37 0.36 0.31 0.29 0.32 0.34 0.39 0.41 0.45 0.55
Year Built 0.63 0.23 1 0.32 0.41 0.63 0.54 0.49 0.26 0.65 0.42 0.6 0.57
1st Flr SF 0.24 0.37 0.32 1 0.81 0.34 0.44 0.5 0.56 0.4 0.61 0.48 0.62
Total Bsmt SF 0.31 0.36 0.41 0.81 1 0.41 0.43 0.49 0.45 0.47 0.56 0.55 0.63
Exter Qual 0.63 0.31 0.63 0.34 0.41 1 0.5 0.47 0.41 0.98 0.49 0.69 0.63
Garage Cars 0.44 0.29 0.54 0.44 0.43 0.5 1 0.89 0.49 0.54 0.75 0.59 0.65
Garage Area 0.4 0.32 0.49 0.5 0.49 0.47 0.89 1 0.49 0.52 0.85 0.56 0.65
Gr Liv Area 0.32 0.34 0.26 0.56 0.45 0.41 0.49 0.49 1 0.45 0.81 0.57 0.7
Quality Factor 0.63 0.39 0.65 0.4 0.47 0.98 0.54 0.52 0.45 1 0.55 0.77 0.71
Gr Liv Area Times Garage Area 0.39 0.41 0.42 0.61 0.56 0.49 0.75 0.85 0.81 0.55 1 0.62 0.75
Overall Qual 0.58 0.45 0.6 0.48 0.55 0.69 0.59 0.56 0.57 0.77 0.62 1 0.8
SalePrice 0.55 0.55 0.57 0.62 0.63 0.63 0.65 0.65 0.7 0.71 0.75 0.8 1
Year Remod/Add
Kitchen Qual_Ex
Year Built
1st Flr SF
Total Bsmt SF
Exter Qual
Garage Cars
Garage Area
Gr Liv Area
Quality Factor
Gr Liv Area Times Garage Area
Overall Qual
SalePrice
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

NORMALLY DISTRIBUTED VARIABLES



LINEARITY


Linearity of Key Numerical Variables
Gr Liv Area (Square Feet)
5000
4000
3000
2000
1000
Garage Area (Square Feet)
1250
1000
750
500
250
0
Total Bsmt SF (Square Feet)
6000
5000
4000
3000
2000
1000
0
1st Flr SF (Square Feet)
5000
4000
3000
2000
1000
0
250000
500000
Sale Price ($)

TRAIN-TEST-SPLIT

- Small trial and error with training size.
- Feature variables were the best fit correlations, the dummy columns, and feature columns.
- Rescale for uniformity.
- Try Linear, Lasso, Ridge.
- Utilize Polynomial Features for better R^2 and less error.
- Achieved $\mathbf{R^2}$ less than ~2% difference between test and train.
- ***RMSE*** a little over a tenth of the expected mean (181k)

CHARACTERISTIC (POLYNOMIAL) EQUATION

- Second Order Polynomial with units m^3 as leading coefficient of order-2 term.
- Form : $y = ax^2 + bx + c$
- Why did we choose Ridge?
  - Great with multicollinearity
  - Reducing Complexity
- What if hundreds or thousands of variables?
- Big risk of overfitting
- Force simpler model, often defined as smaller and "more regular" (less varying) coefficients... small Euclidean norm (2D distance formula)

ERROR DISCUSSION

- Extremely high $R^2 \approx 1$, very close between train and test, (less than 0.1% difference) and low RMSE error achieved with suitable choice of parameters.
- For this we used Lasso to achieve a low RMSE and high Kaggle score: 23322.66654

RECOMMENDATIONS

- Biggest predictors of Sales Price:
  - Area-based variables
  - Overall and other quality factors.
  - Neighborhoods also played a role.
- Year built and renovations (remodeling), namely after the year 2000, were also a significant factor.

CONCLUSION

- Utilized training-test-split techniques
- Handled a complex multilinear regression model with a variety of regressions
- Worked with fundamental fitting methods
- Effectively predicted housing prices from an unknown dataset.

SOURCES

- Chance, Beth L. Rossman, Allan J. *Investigating Statistical Concepts, Applications, and Methods (Third Edition)* August 2015 Beth Chance and Allan Rossman San Luis Obispo, California
- Craven, Mark and Page, David *Regression*,, Computer Sciences 760 Spring 2018, www.biostat.wisc.edu/~craven/cs760
- Kuiper, Shonda *Introduction to Multiple Regression: How Much Is Your Car Worth?* (2008), Journal of Statistics Education, 16:3, , DOI: 10.1080/10691898.2008.11889579
- Pardoe, Iain *Modeling Home Prices Using Realtor Data*, Lundquist College of Business, University of Oregon, *Journal of Statistics Education* Volume 16, Number 2 (2008), www.amstat.org/publications/jse/v16n2/pardoe.html