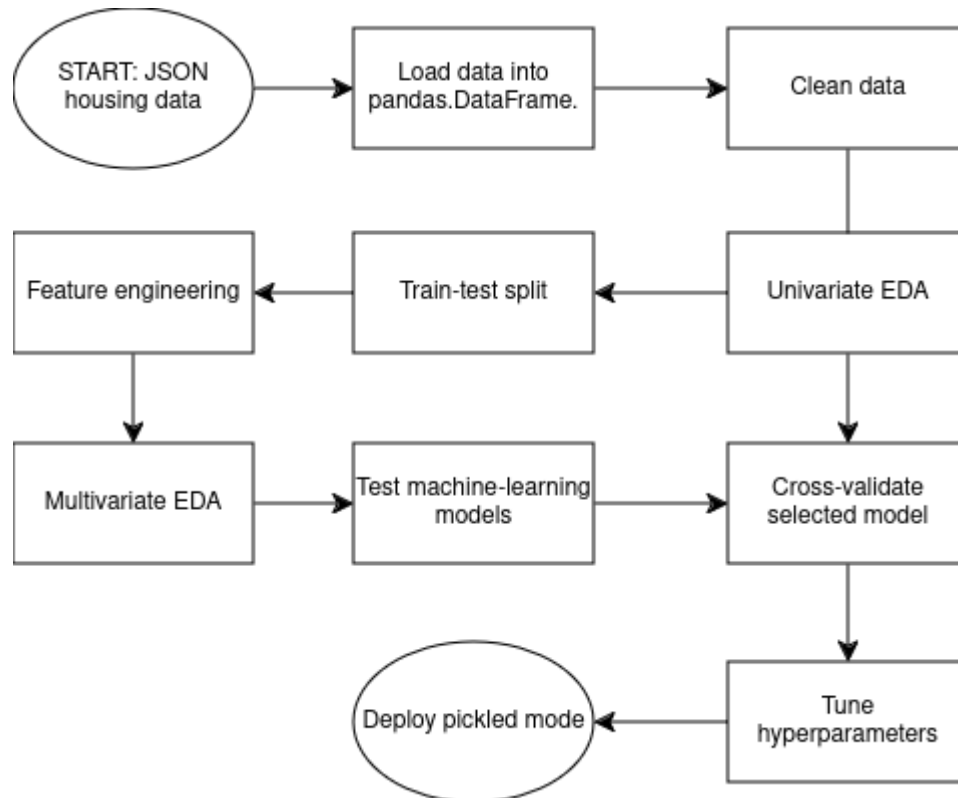# Predicting US Home Sales Prices
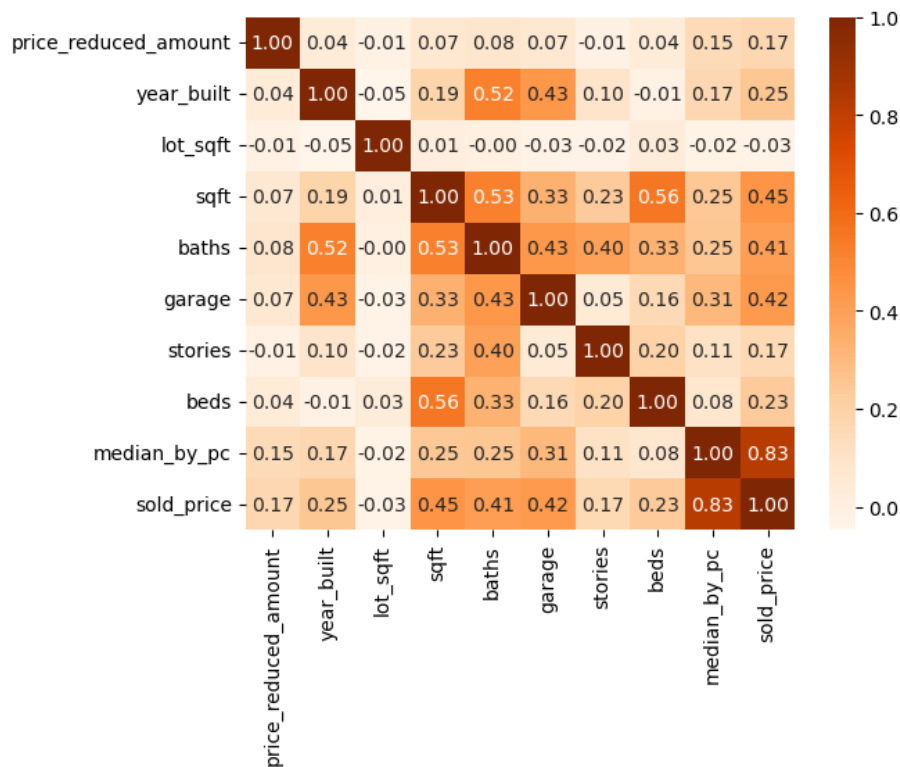
A Lighthouse Labs Data Science Bootcamp Midterm Project

Innocent Byiringiro and Derek Harnett
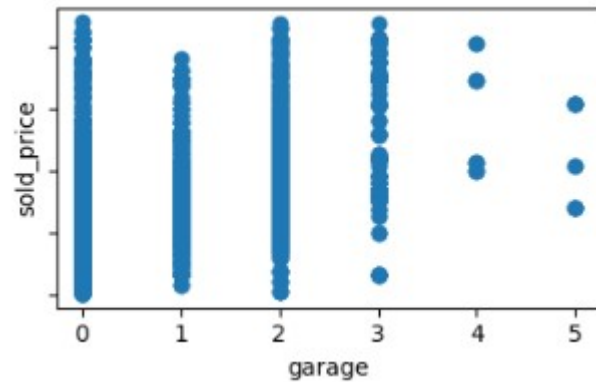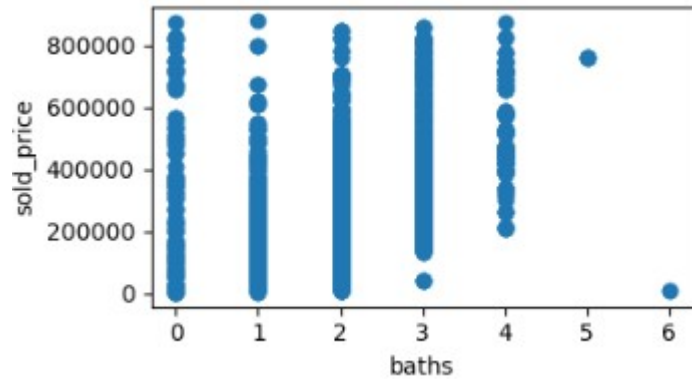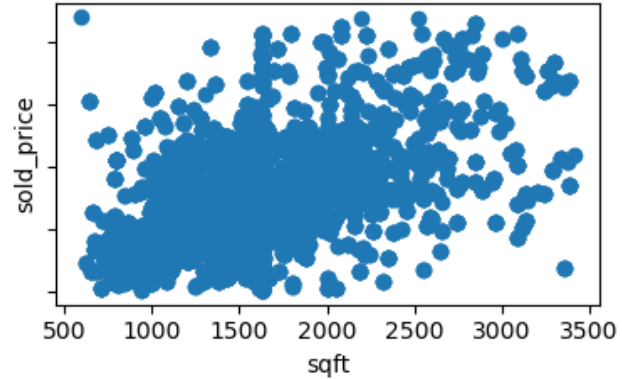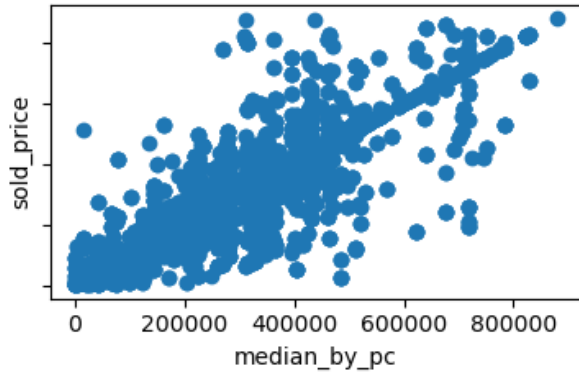
Process Flow Chart

Correlation Heat Map

EDA suggests:

- **Median price by postal code the most significant feature**
- Square footage next most significant
- Number of bathrooms, garage size, year built, and number of bedrooms of interets

Scatter Plots

Model Investigations

Linear (+ Regularization):
- OLS, Ridge, Lasso
- *MinMaxScaler()*
- *SelectKBest(k=8)*
- *PolynomialFeatures()?*
  - No
- Testing adj. R^2: ~0.78

SVM:
- Linear, RBF kernels
- *MinMaxScaler()*
- *SelectKBest(k=8)*
- Testing adj. R^2: ~0.82 (RBF)

XGBoost:
-

Random Forest:
-

Validation, Tuning, Pickling

- Cross-validation & hyperparameter tuning done "by-hand" to **prevent data leakage.**

- Best model pickled to external file