Zillow: Ames home prices
(Kaggle competition)

•••

Sarah Scolnik
DSI-US-06

Data Cleaning: numeric data

Null values: drop, fill with 0, or fill with mean?

- **Drop** rows & columns with null values: linear regression model score (coefficient of determination R^2): **0.83358**
  Kaggle score (root mean squared error RMSE): **33,455**
- **Fill with 0:** linear regression model score: **0.83686**, Kaggle score: **33,847**
- **Fill with mean** of column: LassoCV model score (train/test): **0.91776 / 0.91691**, Kaggle score: **30,391**

Data Cleaning: categorical data

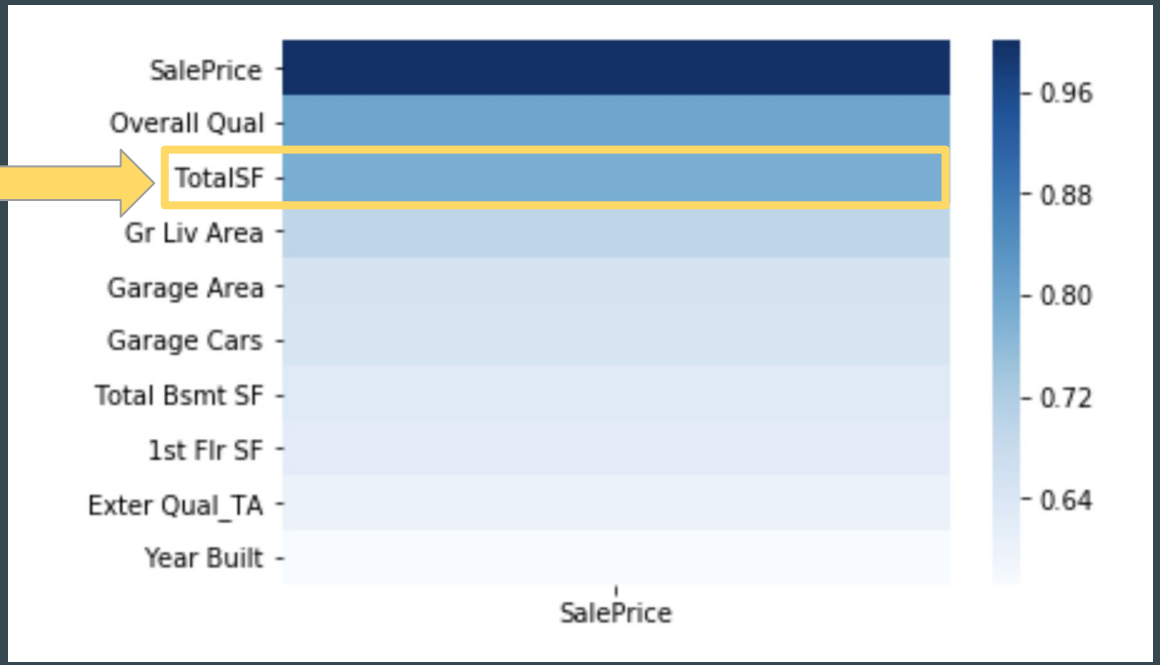
get_dummies

Categorical variables include: Neighborhood, Bldg type, Heating, Kitchen Qual

Feature Engineering

top correlation coefficients

Interaction term:
TotalSF = Total Bsmt SF
+ 1st Flr SF + 2nd Flr SF



Feature Selection

Lasso: highest coefficients

	feature	coef
2	Overall Qual	13359.192353
14	Gr Liv Area	12925.660932
35	TotalSF	11376.525162
215	Kitchen Qual_TA	-10184.602111
131	Roof Matl_CompShg	9828.136102
214	Kitchen Qual_Gd	-9343.919244
32	Misc Val	-9019.613443
91	Neighborhood_NridgHt	8837.626932
169	Exter Qual_TA	-8206.179174
133	Roof Matl_Tar&Grv	7079.694895
4	Year Built	7005.155254
135	Roof Matl_WdShngl	6589.373284
168	Exter Qual_Gd	-5882.595136
97	Neighborhood_StoneBr	5721.925777
180	Bsmt Qual_Gd	-5216.019408
90	Neighborhood_NoRidge	4985.859838

Model

- Linear regression:
  - using numeric data only:
  Kaggle score: 33,846
  - after creating dummies:
  R^2: (train/test): 0.9386 / 0.9254
  model predicts negative sale prices,
  mean: - $463,362

- RidgeCV:
  after scaling data
  Kaggle score: 47,954

- LassoCV:
  after scaling data
  R^2: (train/test): 0.9386 / 0.9254
  Kaggle score: 29,372

Recommendations / Next steps

- fancier models
- more EDA and data engineering