PREDICTING HOUSE PRICES

+Analysis of continuous and non-continuous features

Naveen Parthasarathy

TABLE OF CONTENTS

PROBLEM STATEMENT

1. **How do we accurately predict house prices?**
2. **Do non-continous variables affect the price of a house?**
    A. Continous variables - Square footage, Garage Area
    B. Non-Continous variables - Year Built, Full Baths

For the purposes of this analysis,

1. Numeric variables were considered continuous,
2. Nominal variables were omitted entirely and
3. Others were considered non-continuous (categorical)

INITIAL STATISTICS

Lot Frontage	253
Mas Vnr Area	18
BsmtFin SF 1	1
BsmtFin SF 2	1
Bsmt Unf SF	1
Total Bsmt SF	1
Wood Deck SF	0
Pool Area	0
Screen Porch	0
3Ssn Porch	0
Enclosed Porch	0
Open Porch SF	0
Low Qual Fin SF	0
Garage Area	0
Gr Liv Area	0
Lot Area	0
2nd Flr SF	0
1st Flr SF	0
Misc Val	0
dtype: int64	

Numeric Nulls
19 Columns

Garage Yr Blt	88
Bsmt Full Bath	2
Bsmt Half Bath	2
Year Built	0
Year Remod/Add	0
Full Bath	0
Half Bath	0
Bedroom AbvGr	0
Kitchen AbvGr	0
TotRms AbvGrd	0
Fireplaces	0
Garage Cars	0
Mo Sold	0
Yr Sold	0
dtype: int64	

Discrete Nulls
14 Columns

Fireplace Qu	765
Garage Cond	88
Garage Qual	88
Garage Finish	88
Bsmt Exposure	43
BsmtFin Type 2	41
BsmtFin Type 1	40
Bsmt Qual	40
Bsmt Cond	40
Electrical	0
Functional	0
Kitchen Qual	0
Lot Shape	0
Heating QC	0
Utilities	0
Exter Cond	0
Exter Qual	0
Overall Cond	0
Overall Qual	0
Land Slope	0
Paved Drive	0
dtype: int64	

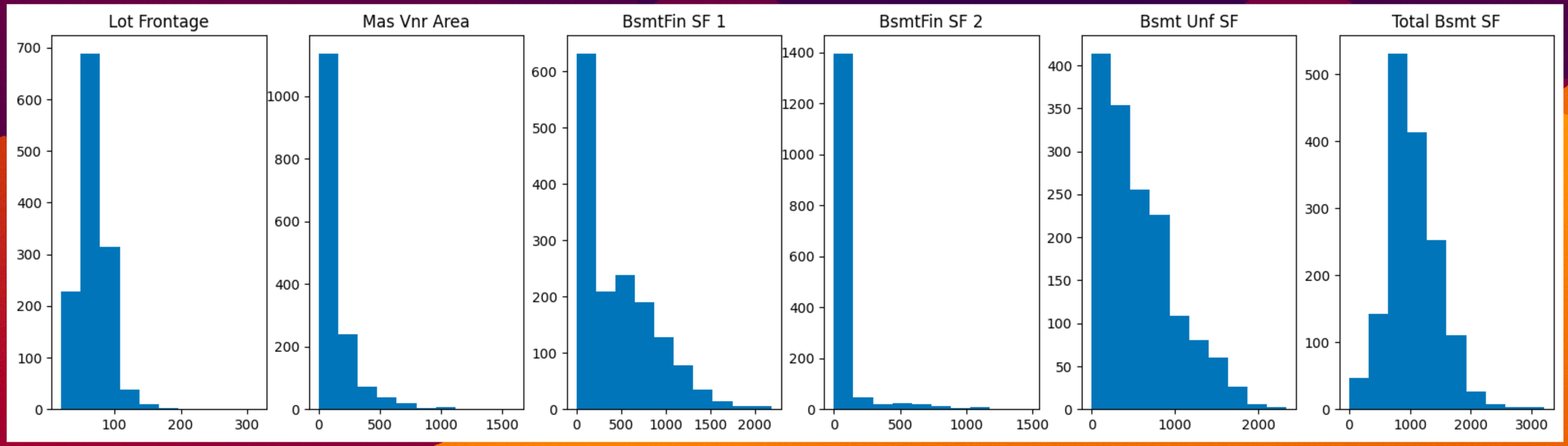
Ordinal Nulls
21 Columns

Numeric Columns that have nulls will be imputed based on their distribution.

Ordinal Columns have a large number of nulls. These nulls have been label encoded.

EXPLORATORY DATA ANALYSIS

Distribution of numeric columns with missing



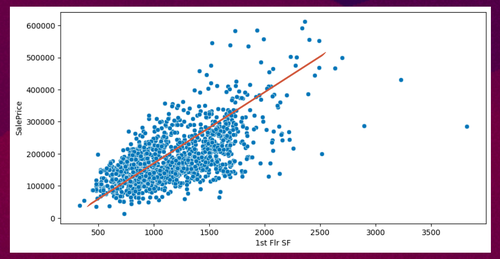
4 of these columns are skewed to the right, so median was used to impute missing

Garage Area had a missing value in Test, so median was used there as well.

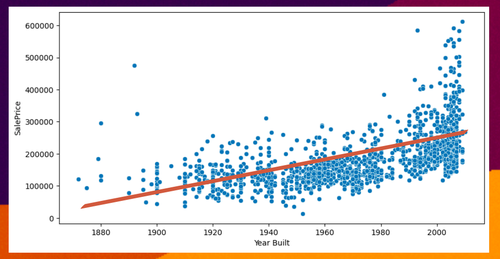
EDA + SELECTING VARIABLES

Columns correlated to Sale Price

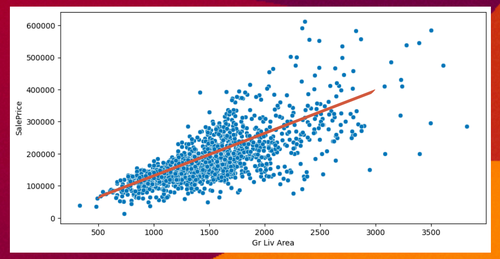
Mas Vnr Area
BsmtFin SF 1
Total Bsmt SF
1st Flr SF
Gr Liv Area
Garage Area
Year Built
Year Remod/Add
Full Bath
TotRms AbvGrd
Fireplaces
Garage Cars
Garage Yr Blt
Overall Qual
Exter Qual
Bsmt Qual
Heating QC
Kitchen Qual
Fireplace Qu
Garage Finish



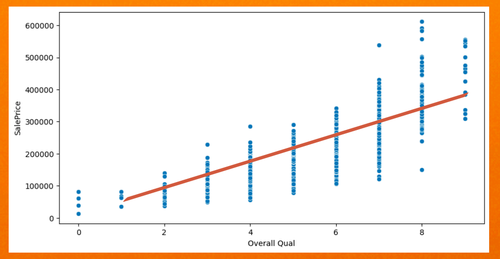
1st Flr SF



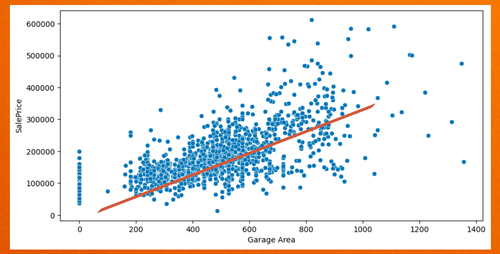
Year Built



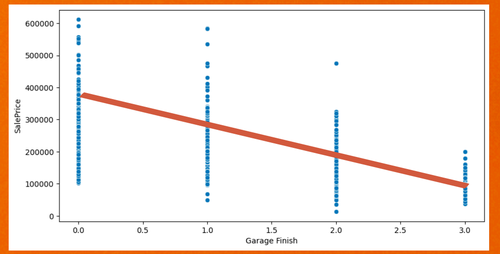
Gr Liv Area



Overall Qual



Garage Area



Garage Finish

MODEL BUILDING

Only Continuous

1. Instantiated model
2. Split data
3. Scaled columns
4. Fit model
5. Made predictions
6. Obtained RMSE

RMSE 42457.310

Continuous + Categorical

1. Instantiated model
2. Split data
3. Transformed columns
4. Fit model
5. Made predictions
6. Obtained RMSE

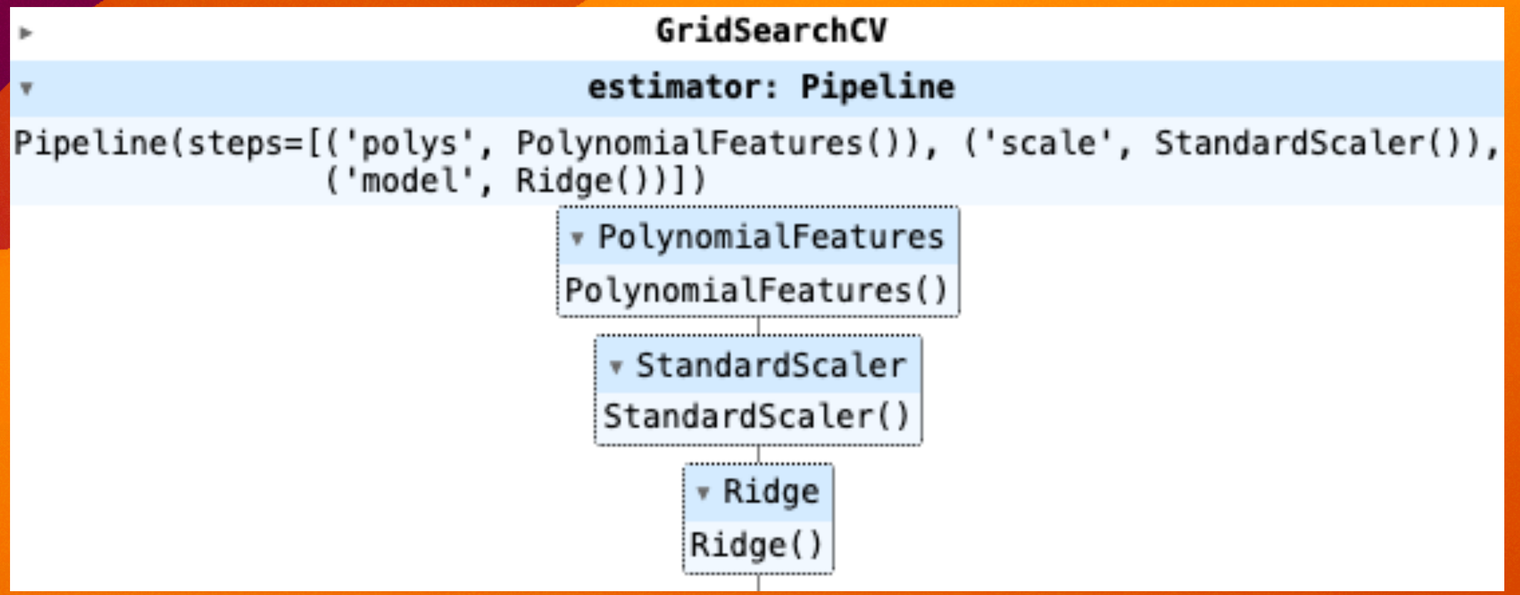
RMSE 28873.763

One Hot Encoding + Scaling

MODEL BUILDING

Best Model

1. Removed Outliers
2. Instantiated model
3. Split data
4. Transformed columns
    A. Polynomial Features
    B. Scaled
    C. Ridge Regularization
5. Fit model with Grid
6. Made predictions
7. Obtained RMSE



RMSE 22507.251

CONCLUSIONS

1. By utilizing polynomial features, scaling and regularization, we were able to obtain a model that had an RMSE of 22507.
2. For this analysis, continuous variables alone did not perform as well as continuous variables combined with other variables.
3. Linear relationships between predictor variables and target variable imply better model performance.

NEXT STEPS

1. Perform feature engineering to develope better variables and remove redundant ones.
2. Explore other more complex models (SGDRegressor).
3. Normalize numerical values to aid with skewed data.

THANK YOU