total talk: 15 min

DSI Project-2

Eziz Abuduaizezi

Roadmap Agenda

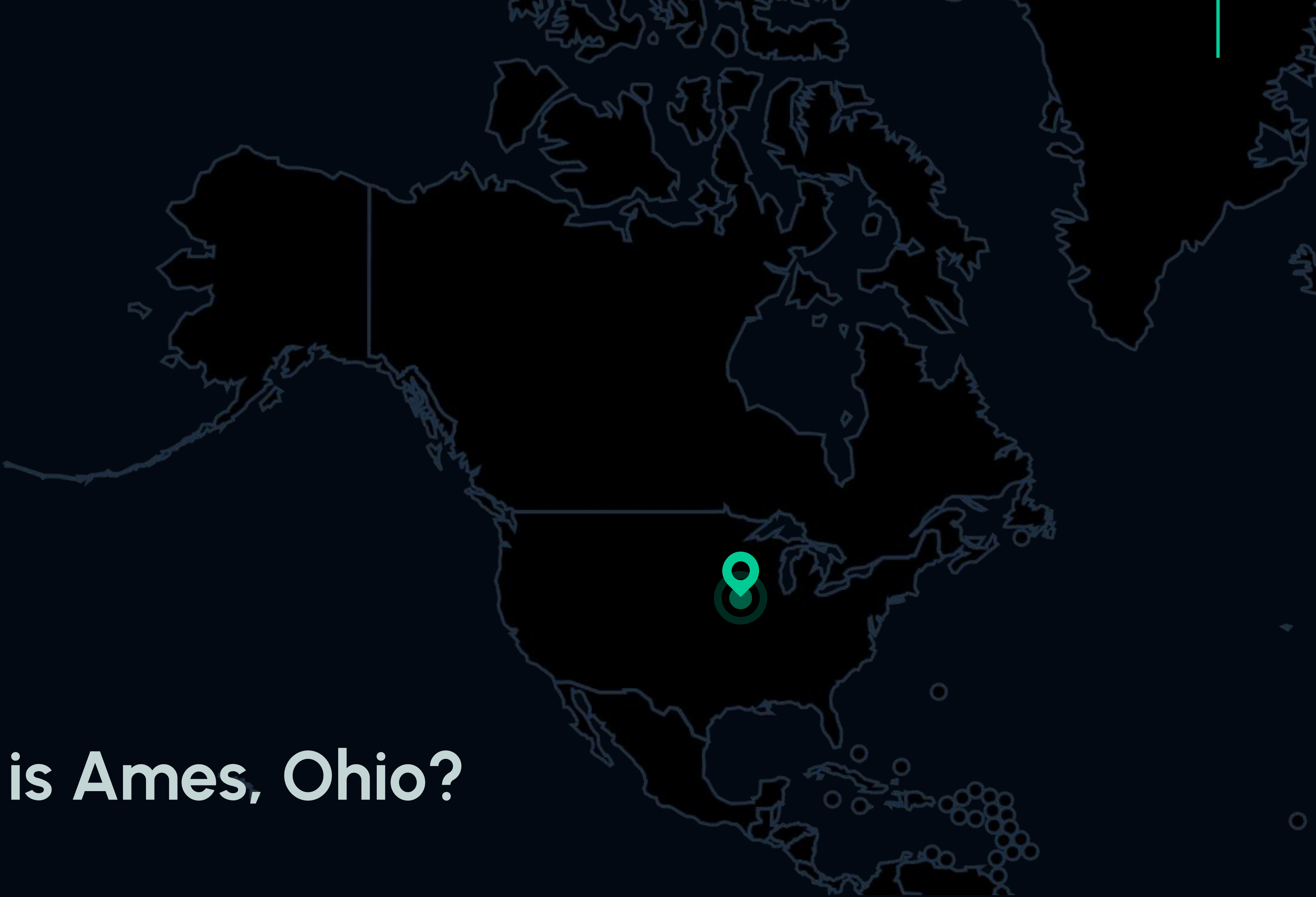
01	EDA
02	Cleaning
03	Feature Engineering
04	Model Benchmarks
05	Conclusion

0 min talk

01 / EDA

0 min talk

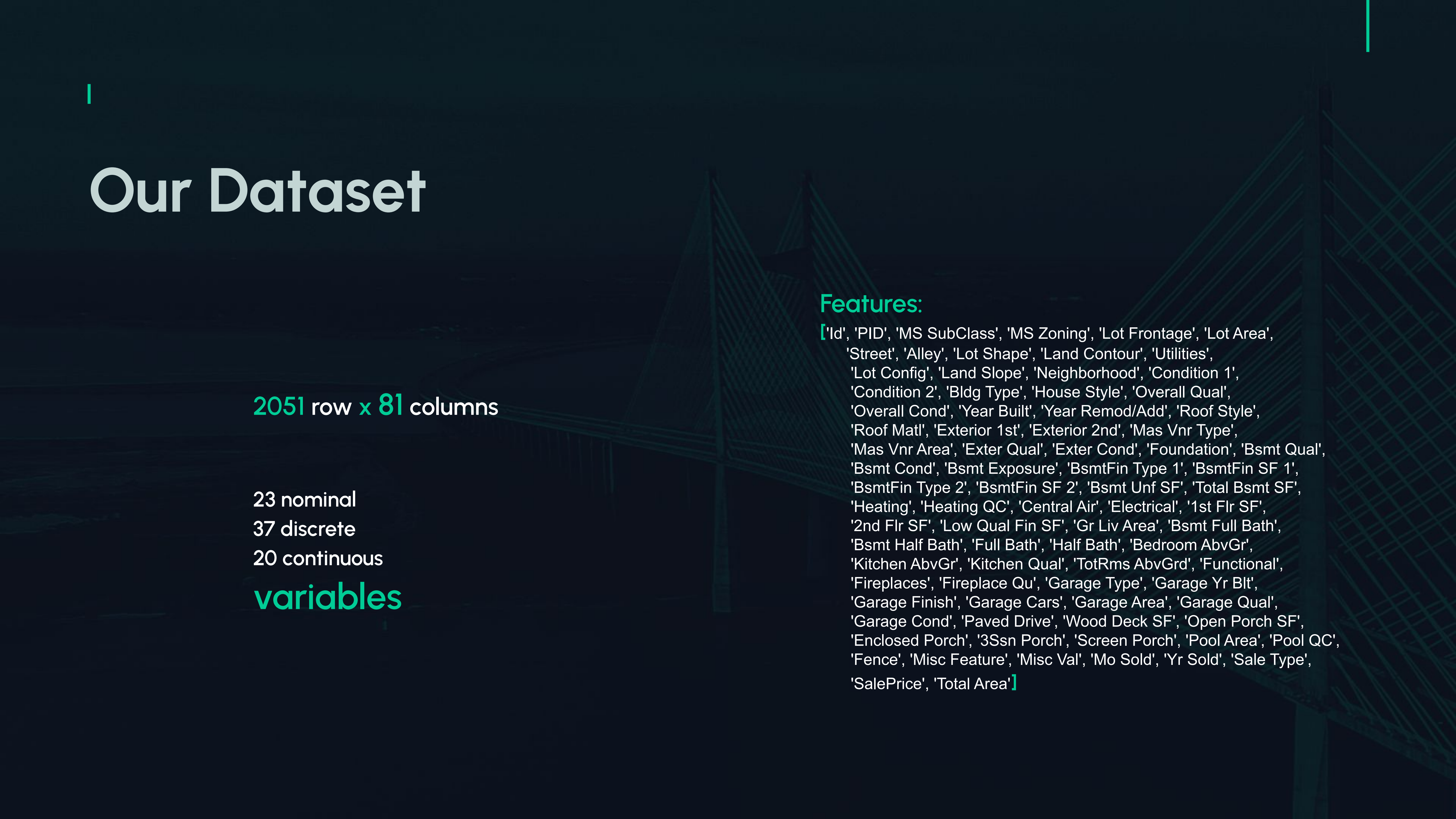
Housing Price in Ames, Ohio



Where is Ames, Ohio?

Problem Statement

What neighbourhood has the most expensive houses in Ames?



Our Dataset

2051 row x 81 columns

23 nominal
37 discrete
20 continuous

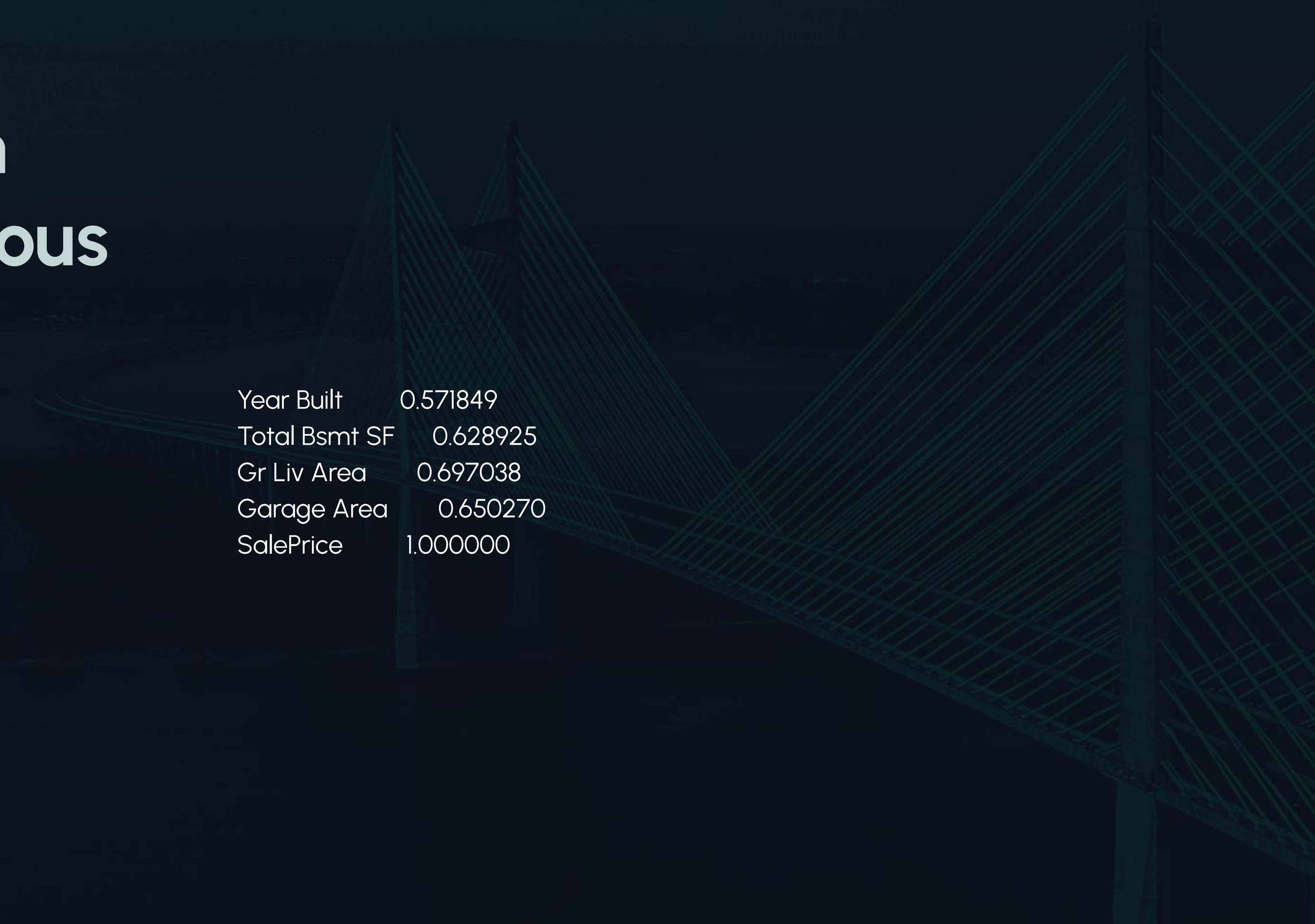
variables

Features:

```
['Id', 'PID', 'MS SubClass', 'MS Zoning', 'Lot Frontage', 'Lot Area',
     'Street', 'Alley', 'Lot Shape', 'Land Contour', 'Utilities',
      'Lot Config', 'Land Slope', 'Neighborhood', 'Condition 1',
      'Condition 2', 'Bldg Type', 'House Style', 'Overall Qual',
      'Overall Cond', 'Year Built', 'Year Remod/Add', 'Roof Style',
      'Roof Matl', 'Exterior 1st', 'Exterior 2nd', 'Mas Vnr Type',
      'Mas Vnr Area', 'Exter Qual', 'Exter Cond', 'Foundation', 'Bsmt Qual',
      'Bsmt Cond', 'Bsmt Exposure', 'BsmtFin Type 1', 'BsmtFin SF 1',
      'BsmtFin Type 2', 'BsmtFin SF 2', 'Bsmt Unf SF', 'Total Bsmt SF',
      'Heating', 'Heating QC', 'Central Air', 'Electrical', '1st Flr SF',
      '2nd Flr SF', 'Low Qual Fin SF', 'Gr Liv Area', 'Bsmt Full Bath',
      'Bsmt Half Bath', 'Full Bath', 'Half Bath', 'Bedroom AbvGr',
      'Kitchen AbvGr', 'Kitchen Qual', 'TotRms AbvGrd', 'Functional',
      'Fireplaces', 'Fireplace Qu', 'Garage Type', 'Garage Yr Blt',
      'Garage Finish', 'Garage Cars', 'Garage Area', 'Garage Qual',
      'Garage Cond', 'Paved Drive', 'Wood Deck SF', 'Open Porch SF',
      'Enclosed Porch', '3Ssn Porch', 'Screen Porch', 'Pool Area', 'Pool QC',
      'Fence', 'Misc Feature', 'Misc Val', 'Mo Sold', 'Yr Sold', 'Sale Type',
      'SalePrice', 'Total Area']
```

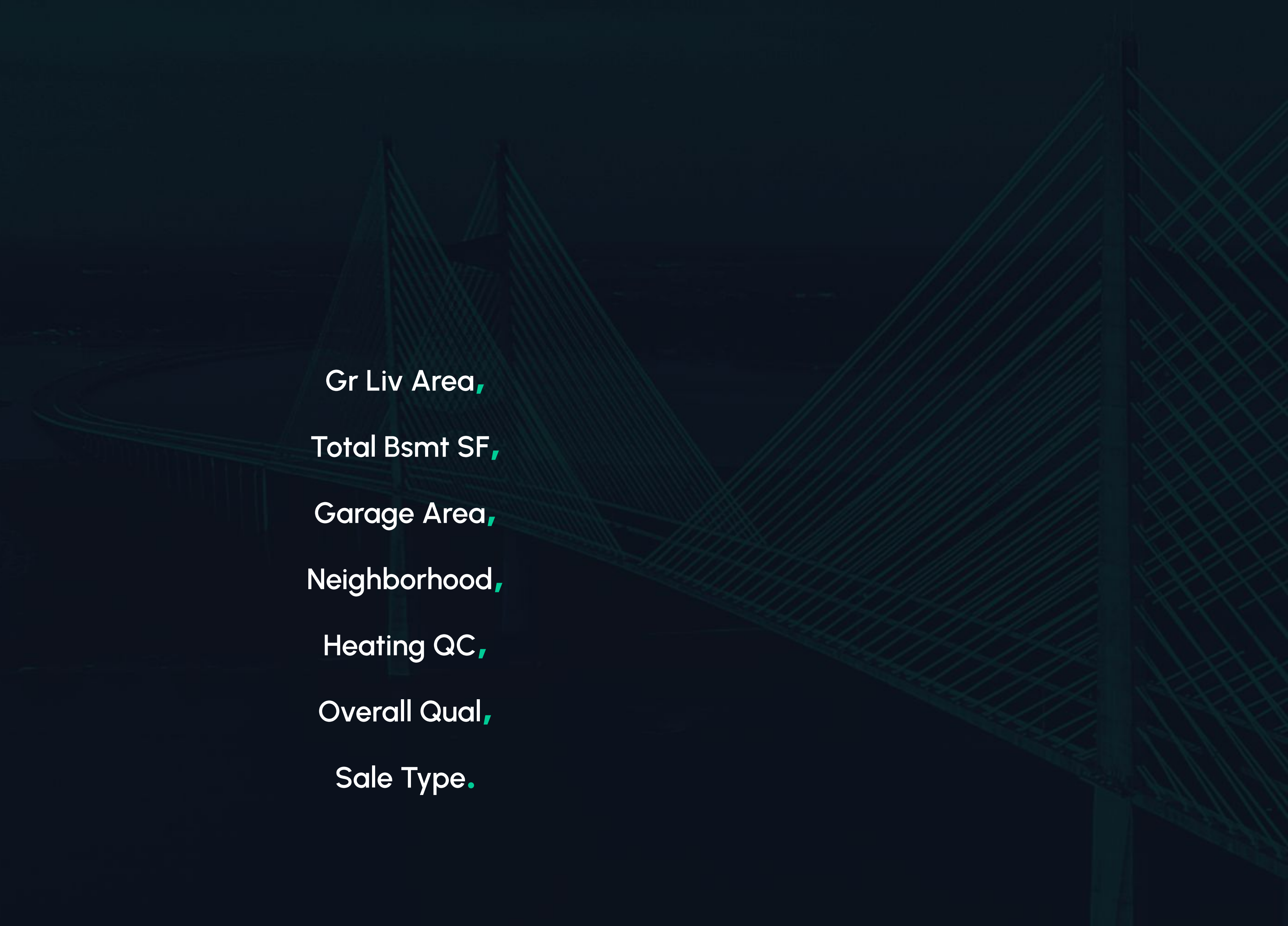
Correlation for continuous variables

TO Sale Price



Year Built	0.571849
Total Bsmt SF	0.628925
Gr Liv Area	0.697038
Garage Area	0.650270
SalePrice	1.000000

Other columns



Gr Liv Area,

Total Bsmt SF,

Garage Area,

Neighborhood,

Heating QC,

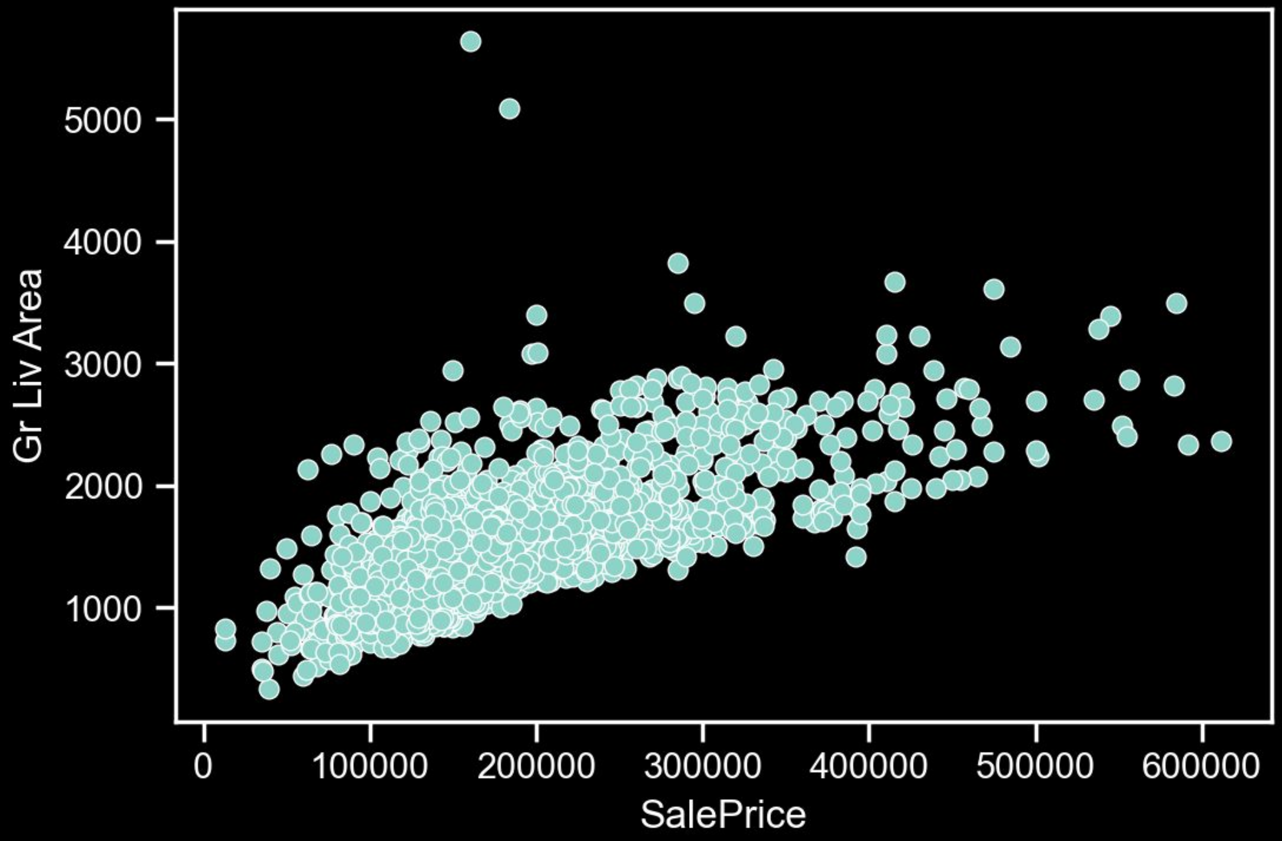
Overall Qual,

Sale Type.

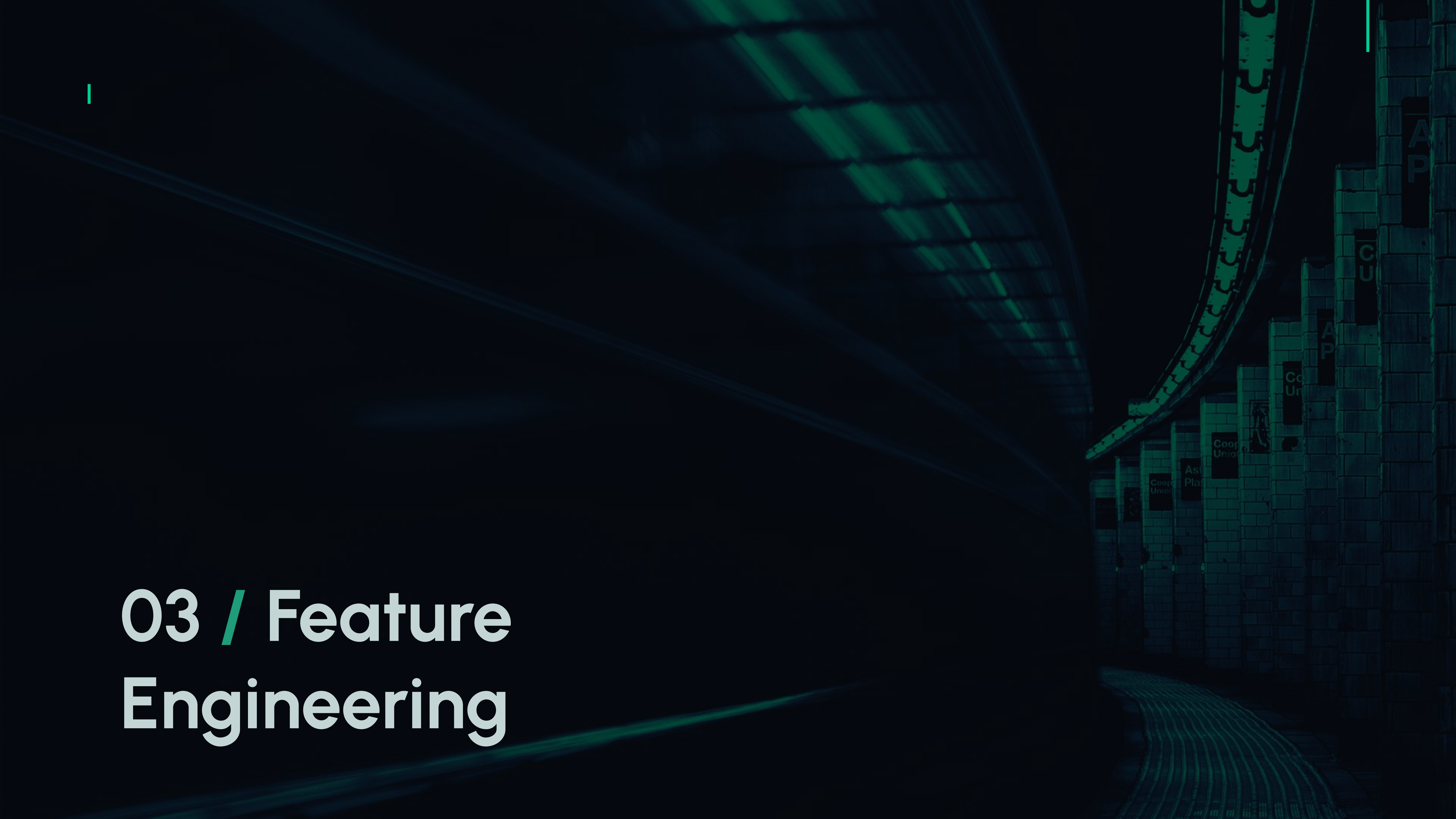
0 min talk

02 / Cleaning

Cleaning

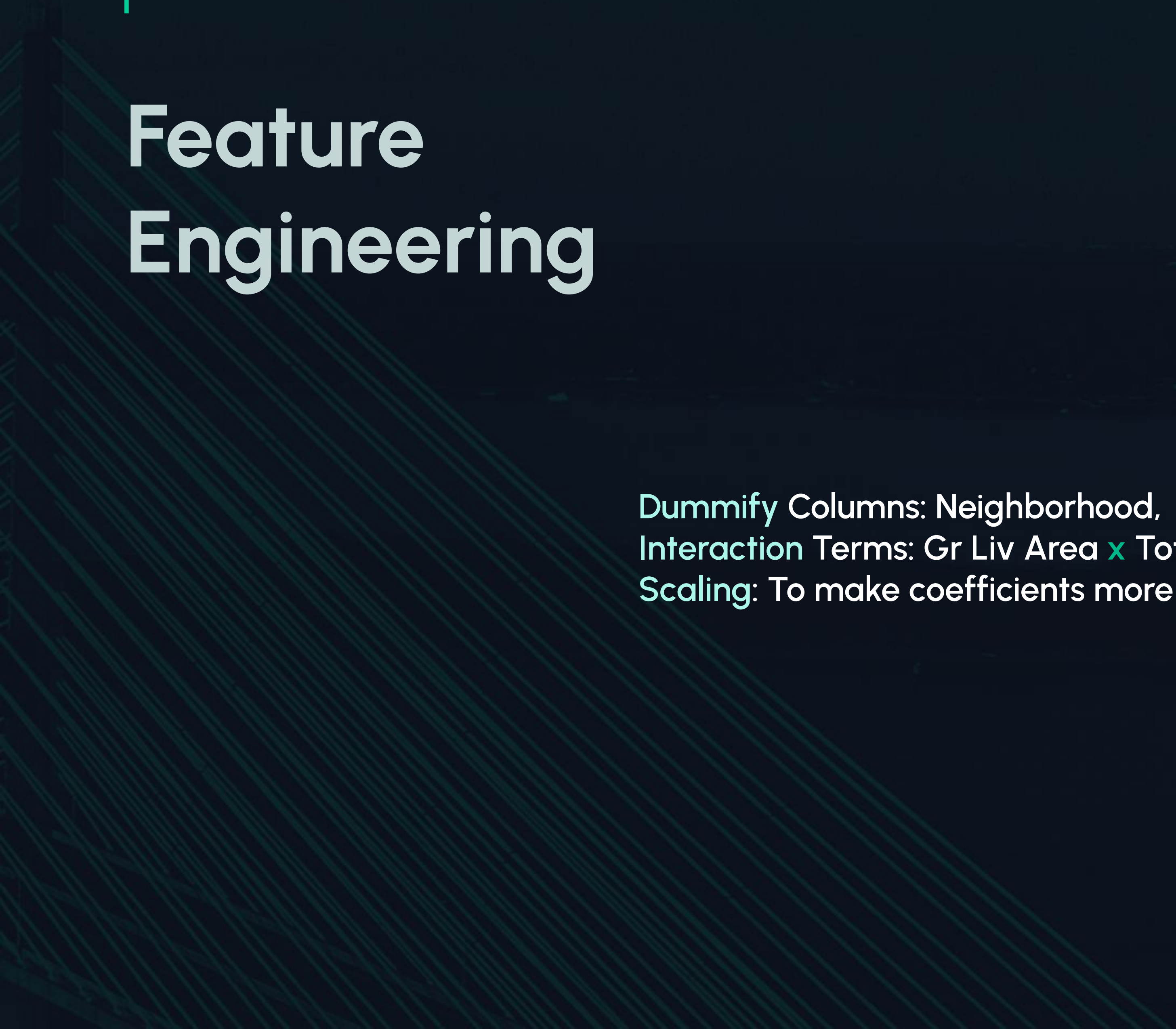


- Removing outliers
- Drop unused columns
- Drop rows with NaN values (2 rows)



03 / Feature Engineering

Feature Engineering



Dummify Columns: Neighborhood, Sale Type, Heating QC, Overall Qual
Interaction Terms: Gr Liv Area x Total Bsmt SF x Garage Area
Scaling: To make coefficients more comparable

0 min talk

04 / Model Fitting

Model Benchmarks

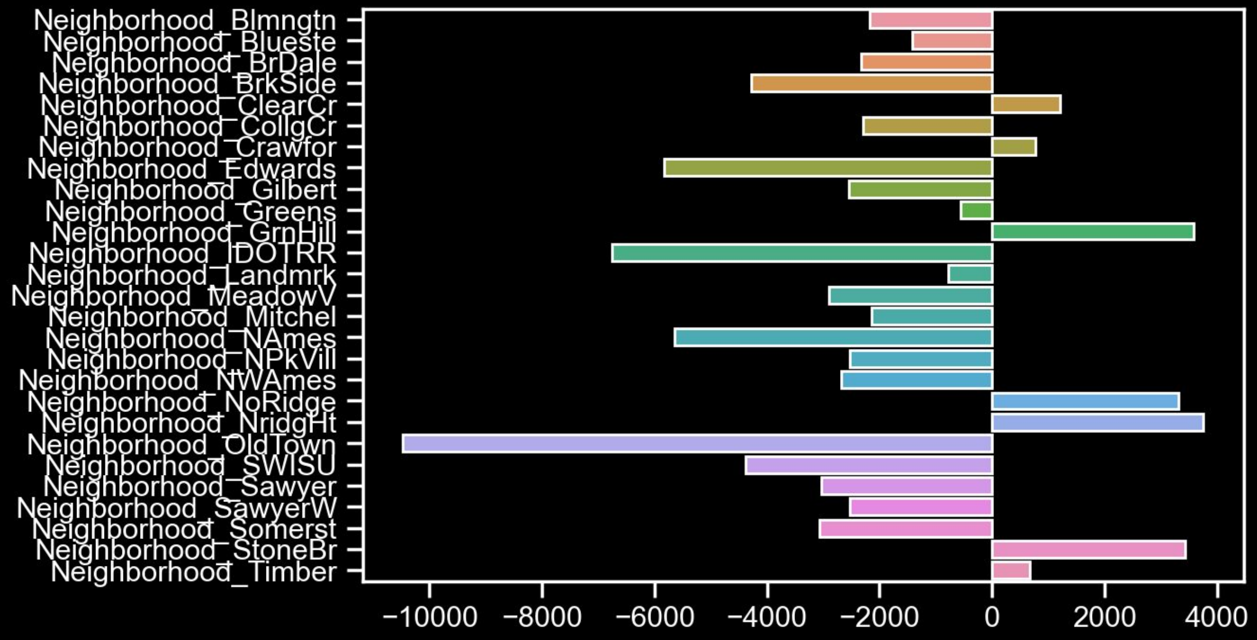
0.858 R2

Linear Regression

Test model score: 0.858

Train model score: 0.869

to Answer our question



Most expensive neighborhoods

- Northridge Heights 3,762 x
- Green Hills 3577 x
- Stone Brook 3435 x

<compare to Veenker>

Thank you!