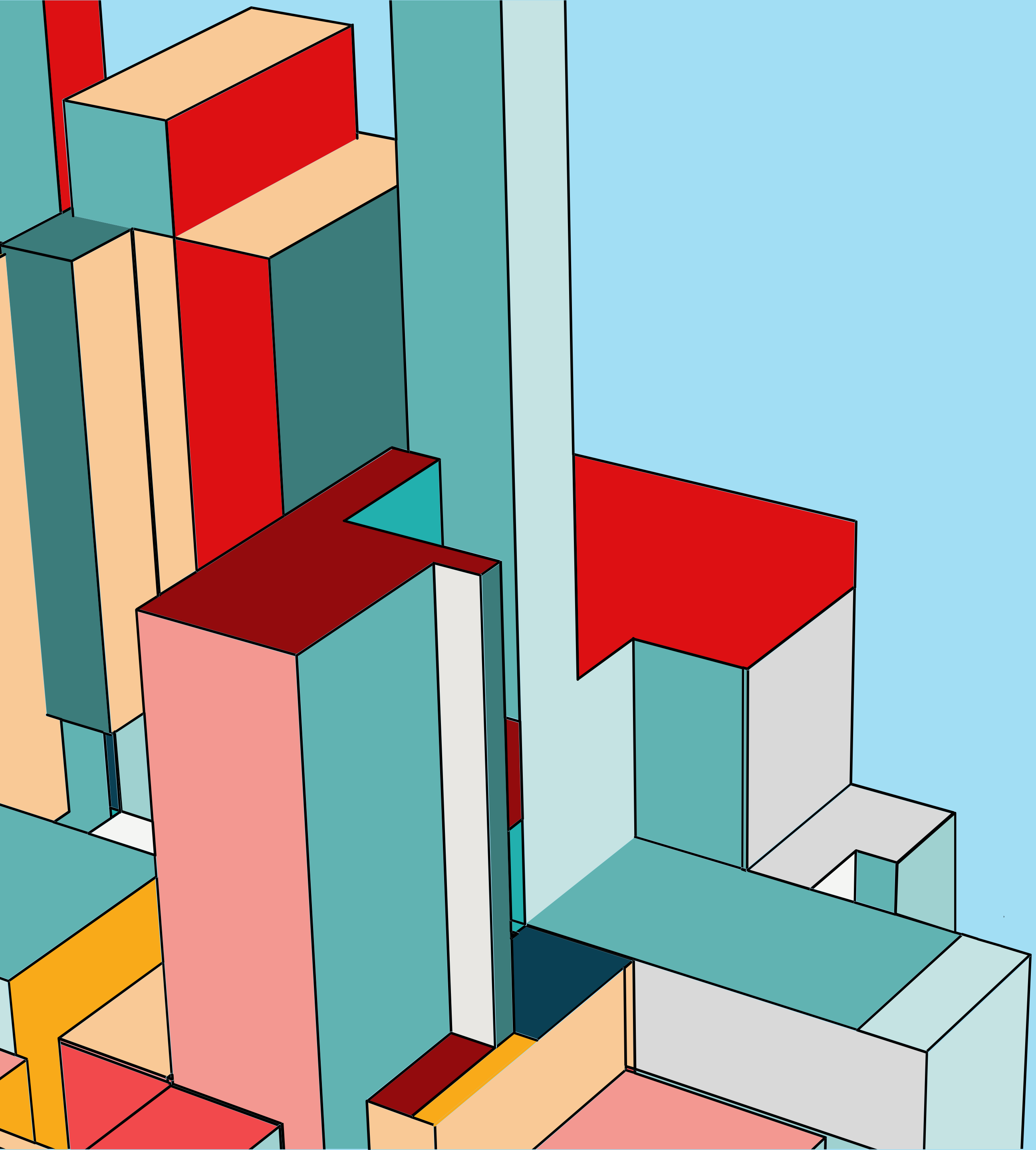


WHERE DATA MEETS DOMAIN EXPERTISE:

PREDICTING HOME SALE PRICES WITH MULTILINEAR REGRESSION MODELS

AN NGUYEN,
DATA SCIENTIST
MARCH 2024

DATA OR INSIGHT?

The rise of real estate data-driven tech companies like Redfin is transforming the local market dynamics where customers are expecting better house price predictions.

Nguyening Deals Agency, a trusted local real estate company facing intense competitions, is questioning:

“should we investing in sophisticated quantitative prediction models or focusing on training our agents to see value beyond the numbers and recognizing qualitative metrics that best predict house sale prices?"

DATA SOURCE:

- Primary data source is a subset of Ames Housing Dataset split into 2 sets:
- 'train' set: 1568 observations, 82 columns including 'SalePrice'
- 'test' set: of 513 observation, 81 columns excluding 'SalePrice'

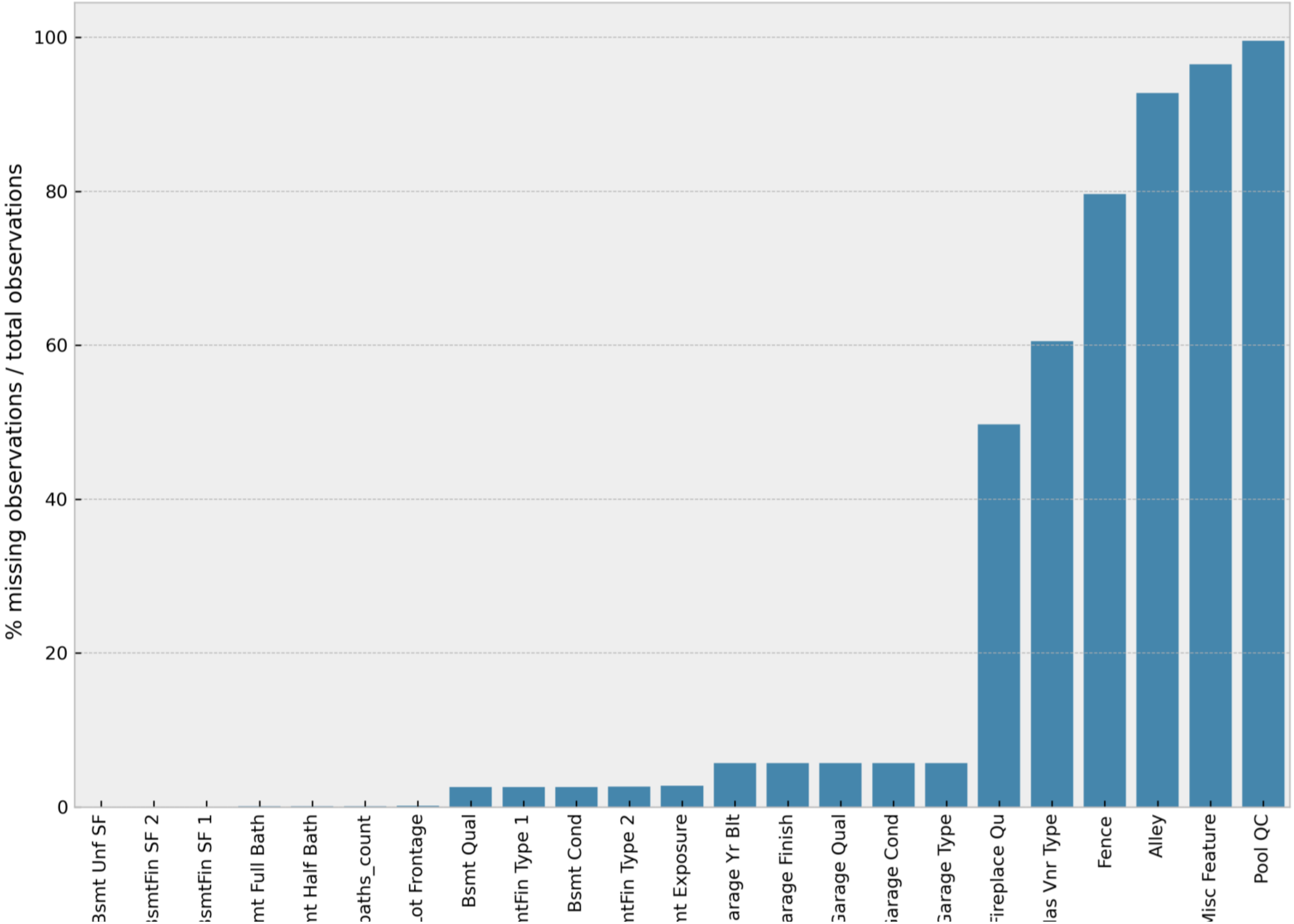
EXPLORATORY DATA ANALYSIS (EDA):

- Separate data columns into numeric, ordinal, and nominal types
- Review Ames Housing Data Documentation to identify potential features
- Correlation heatmap to visualize the relationship between 'numeric' features and 'SalePrice'
- Plots of each feature candidate against "SalePrice"

DATA CLEANING & MODELLING

- Identify and remove outliers
- Fill in null values appropriately for numeric, ordinal, and nominal data types
- Apply the same steps to "train" set and "test" set using a customized data cleaning function
- Encode and transform numerical, nominal, ordinal features accordingly
- Build a linear regression model
- Evaluate model performance metrics (R-squared and RMSE)
- Cross validate model performance between 'train' set and the known 'test' set

PERCENTAGE DISTRIBUTION OF COLUMNS WITH NULLS



29 KEY FEATURES

13 Numeric Features

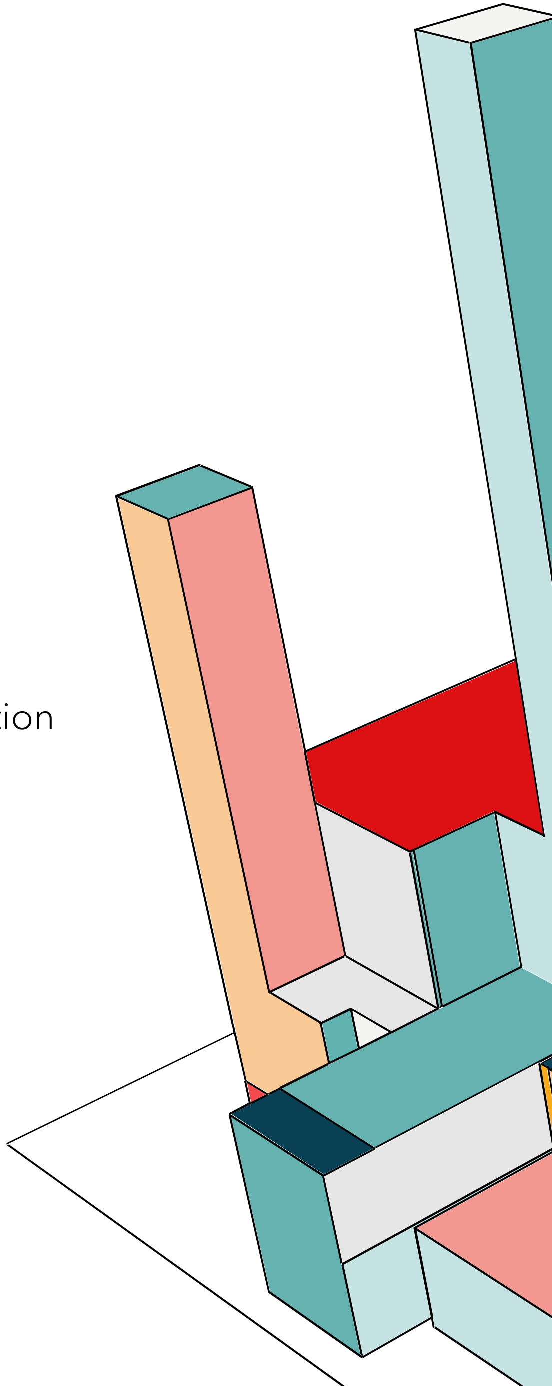
- above ground living area (sqft)
- total basement area (sqft)
- garage cars capacity
- garage area (sqft)
- year built
- year remodeled
- total rooms above ground
- masonry veneer area
- number of fireplaces
- lot frontage (ft)
- lot area (sqft)
- total area of porch and deck (sqft)
- count of full and half baths

7 Nominal Features

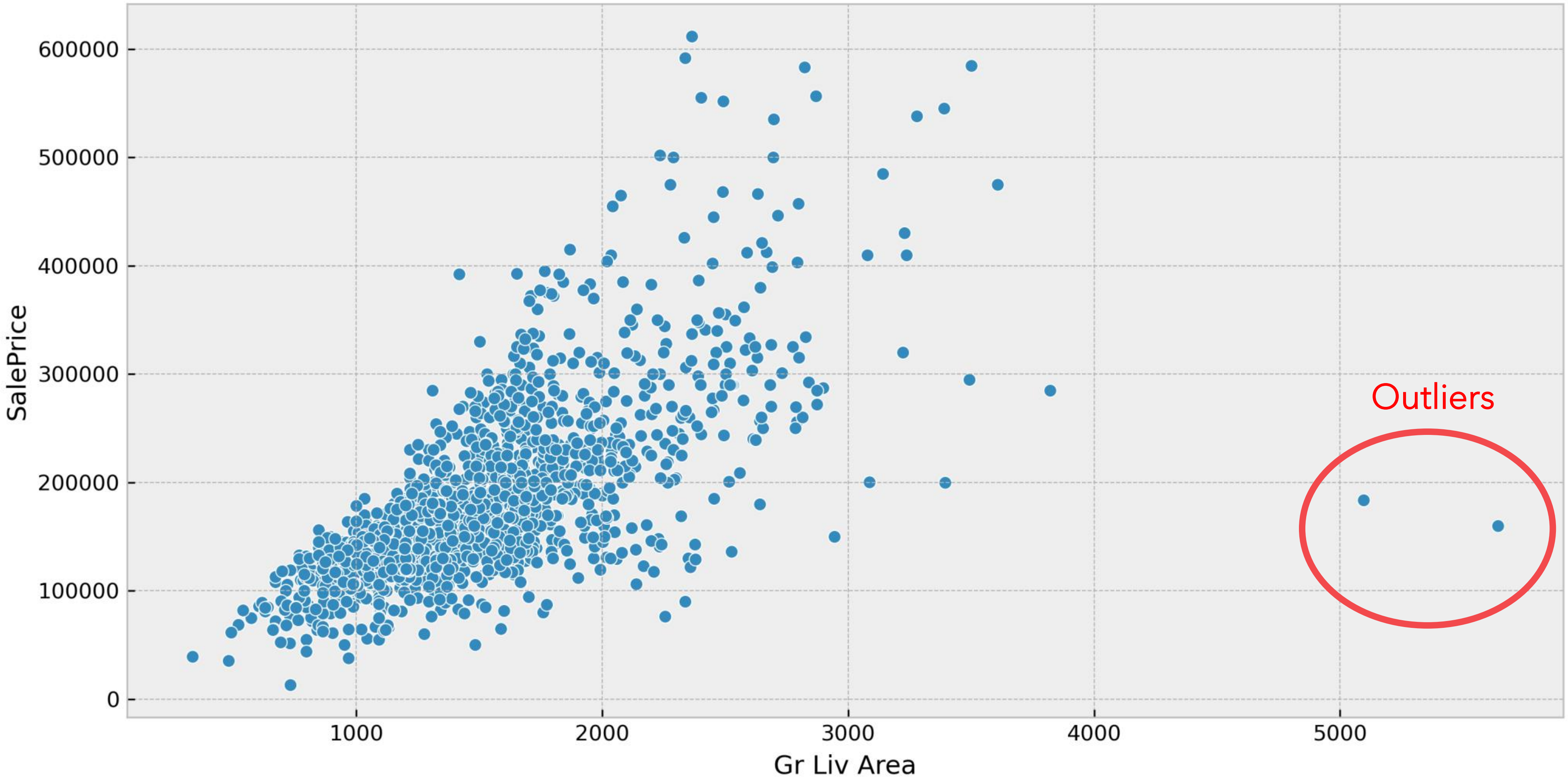
- neighborhood
- MS zoning
- building type
- masonry veneer type
- House style
- foundation
- sale type

13 Ordinal Features

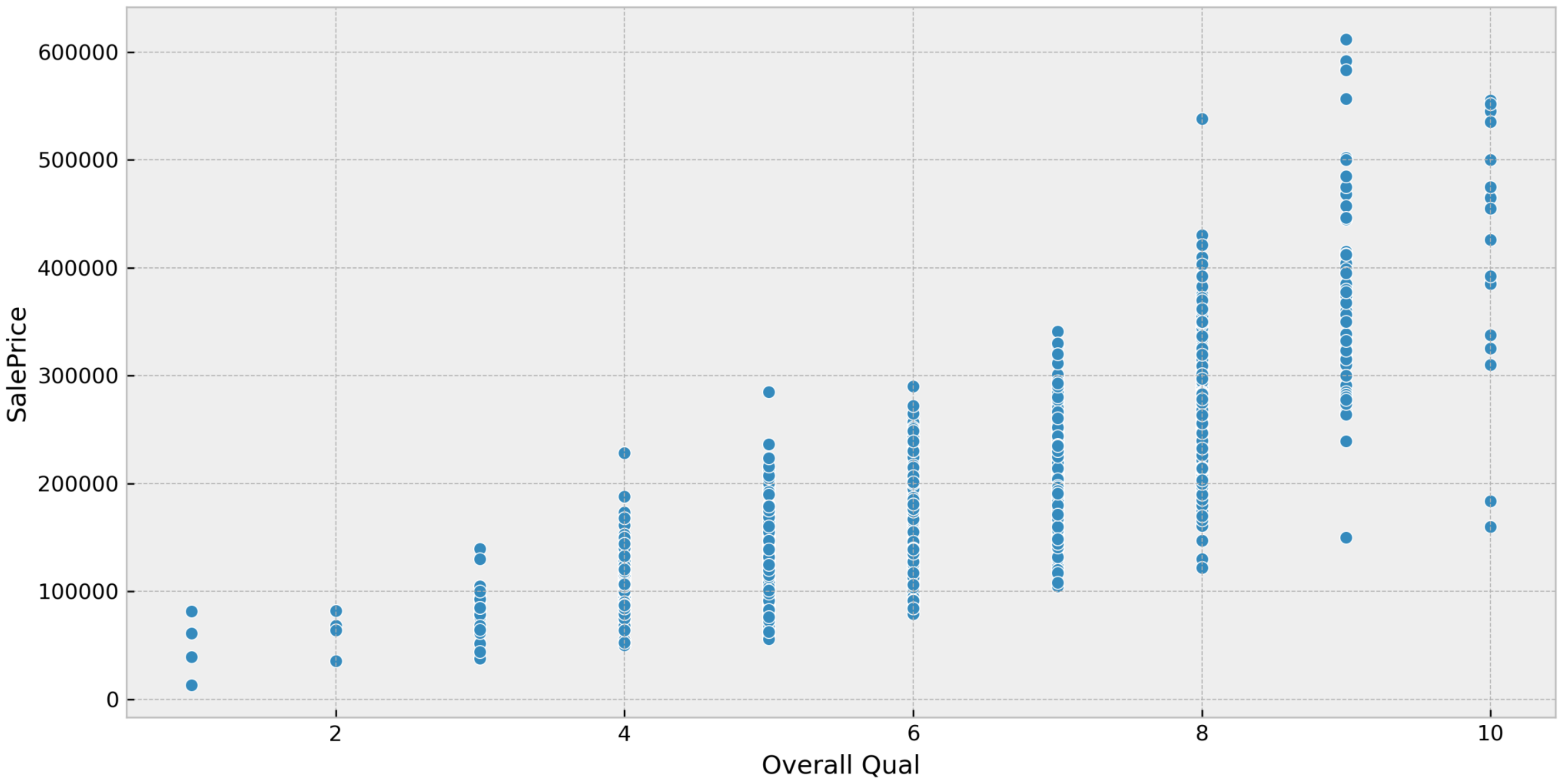
- overall quality
- overall condition
- exterior quality
- basement quality
- heating quality and condition
- kitchen quality
- home functionality
- electrical system type
- garage quality



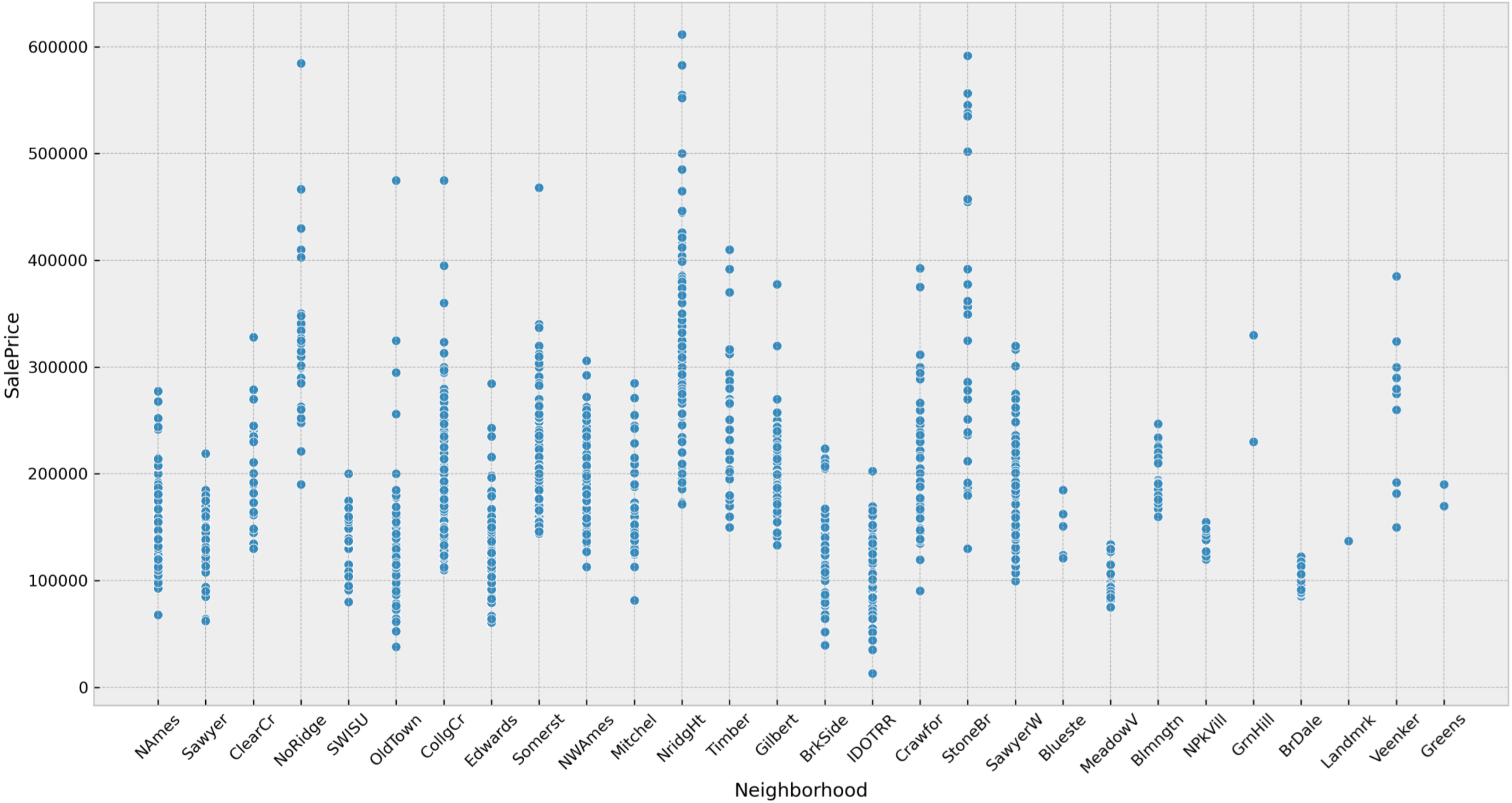
THERE IS A STRONG POSITIVE RELATIONSHIP BETWEEN "ABOVE GROUND LIVING AREA" (SQFT) AND "SALEPRICE".



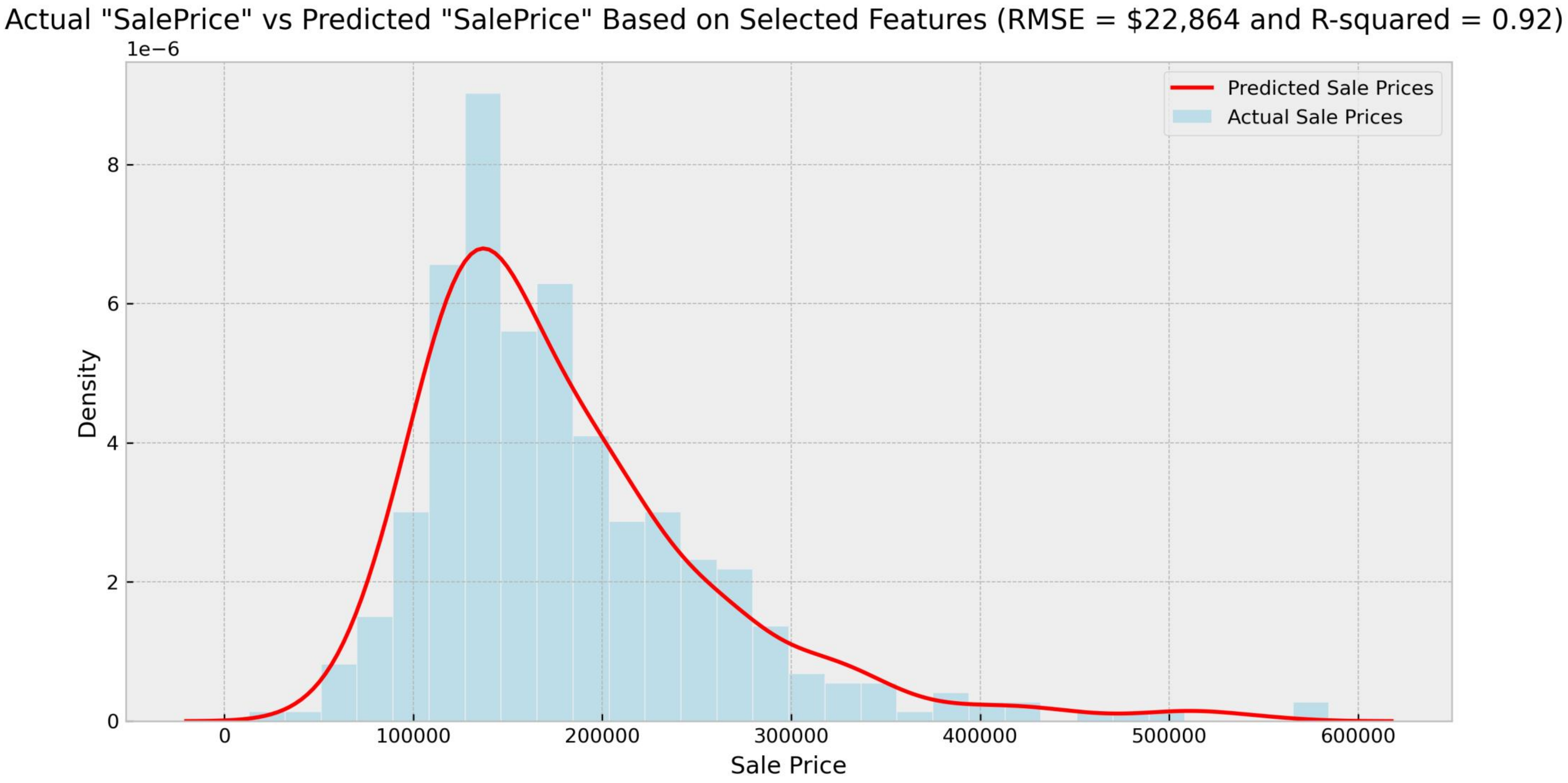
HIGHER "OVERALL QUALITY' SCORE CORRELATES TO HIGHER "SALE PRICE" ON AVERAGE.



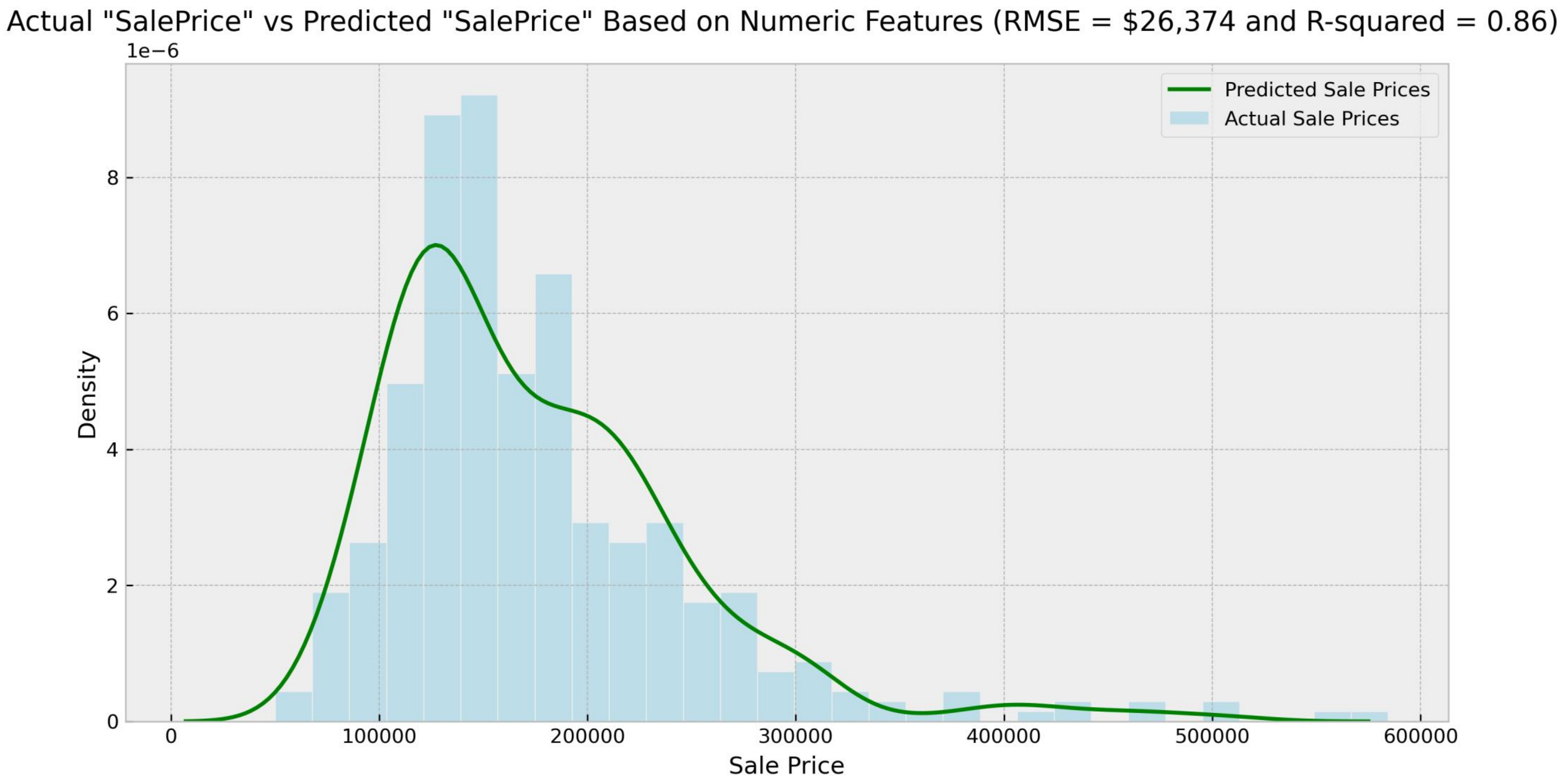
EACH NEIGHBORHOOD IS UNIQUE WITH DIFFERENT RANGES IN "SALE PRICE".



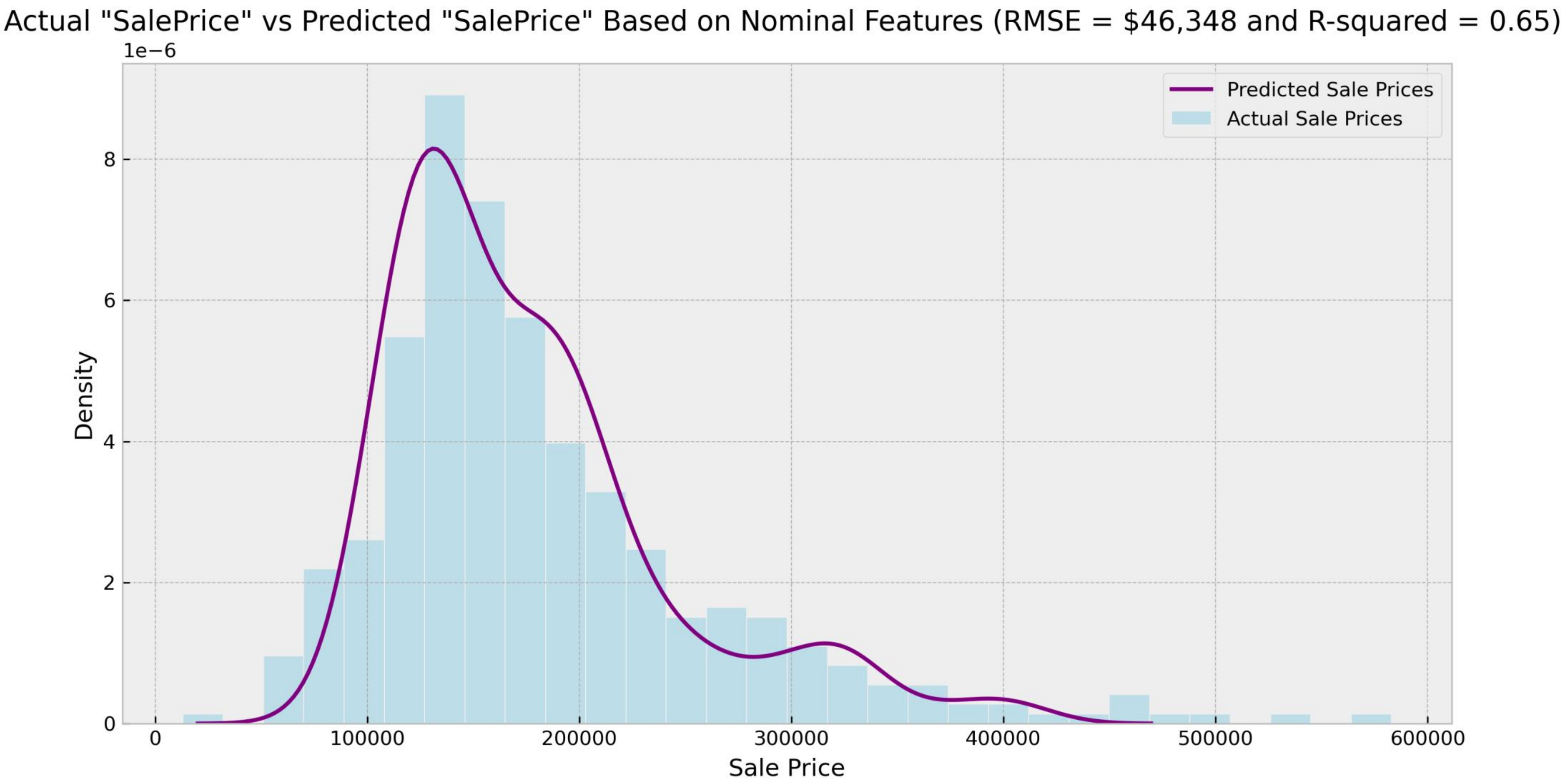
WHAT IS THE OVERALL ACCURACY OF THE HOUSE SALE PRICE PREDICTION MODEL?



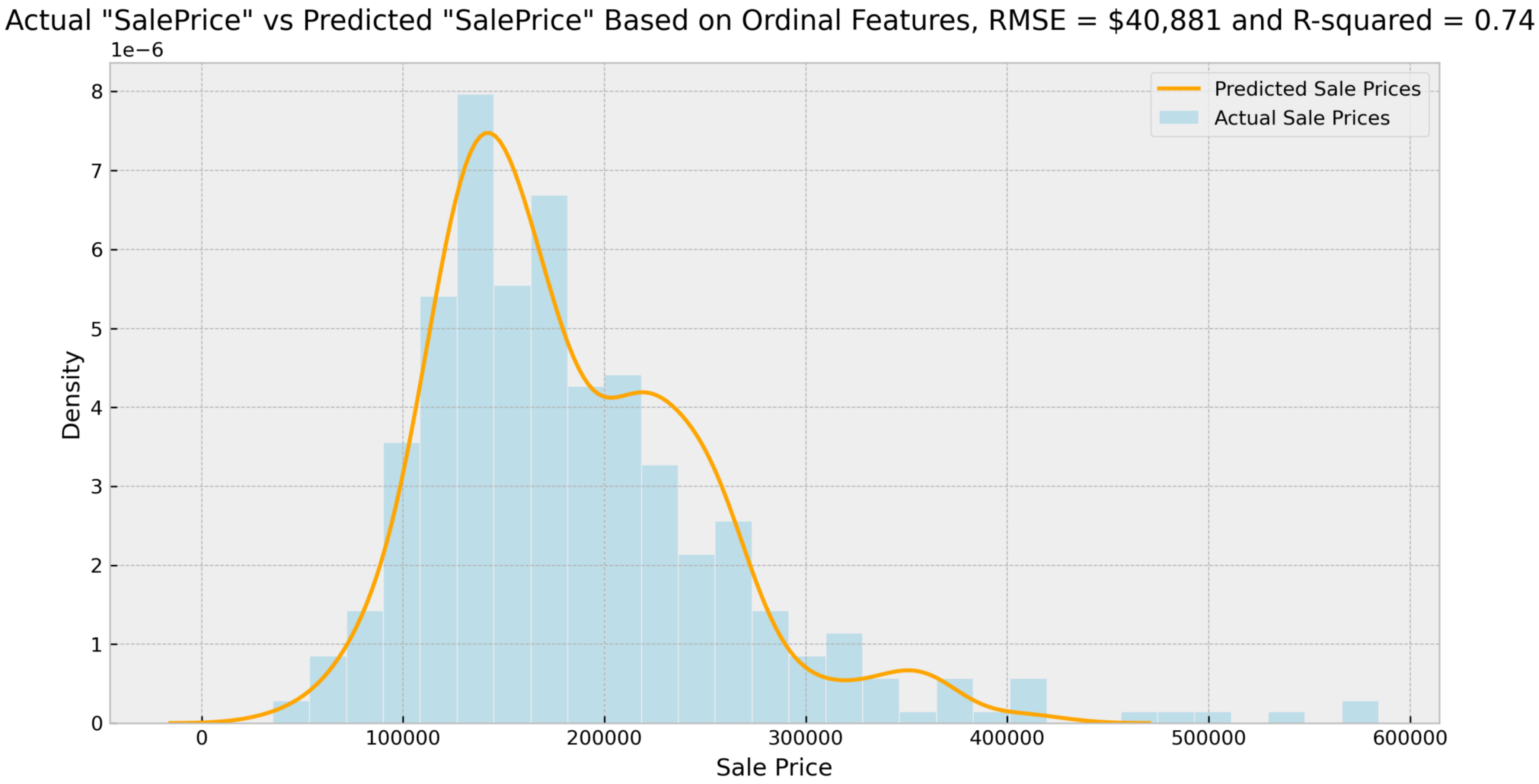
REGRESSION MODEL BASED ON NUMERIC FEATURES ONLY



REGRESSION MODEL BASED ON NOMINAL FEATURES ONLY



REGRESSION MODEL BASED ON ORDINAL FEATURES ONLY



DATA AND INSIGHT

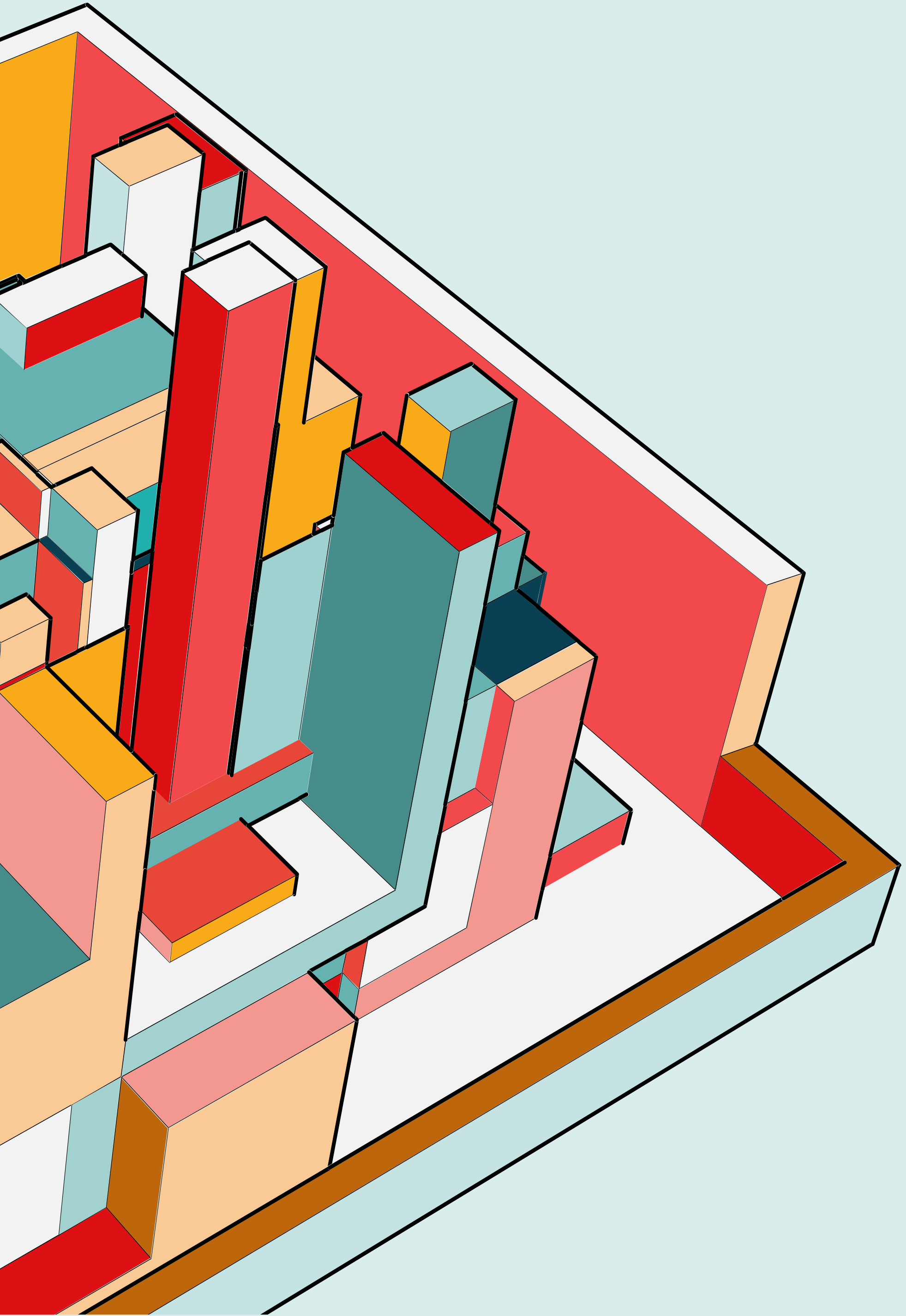
- The best house sale price prediction model was built on **29 features**:

Prediction Model	R-squared	RMSE
29 features	0.92	$22,864
13 numeric features	0.86	$26,374
7 nominal features	0.65	$46,348
9 ordinal features	0.74	$40,881

- The average “SalePrice” in the dataset is $181,061.
- Numeric, ordinal, and nominal metrics are all important when building a regression model to predict sale prices.
- Nguyening Deals Agency should take a hybrid approach:
    1. train their agents on a data-first mindset in house price evaluations,
    2. empower knowledge sharing of subjective insights during biweekly or monthly “state of the market” group discussions, and
    3. start a small-scaled data science project to collect internal data on these metrics, particularly on ordinal features like neighborhoods.

SOURCES

- Ames Housing Dataset: https://jse.amstat.org/v19n3/decock/DataDocumentation.txt
- Ames, Iowa: Alternative to the Bosting Housing Data as an End of Semester Regression Project: https://jse.amstat.org/v19n3/decock.pdf



THANK YOU