Residential Real Estate Price Prediction — Feature Engineering

Real Estate Investment Trust • Investment Analytics Project

Overview

- This stage focuses on transforming raw property data into meaningful features that better represent home value.
- New derived features were created and validated to improve model readiness while preserving the full dataset for business intelligence use.

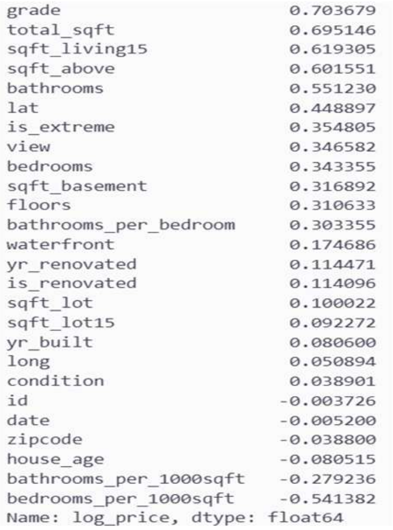
Objective

- The objective of this stage is to engineer and validate predictive features, reduce redundancy caused by overlapping variables, and identify a stable feature set suitable for price modeling without removing source columns.

Results

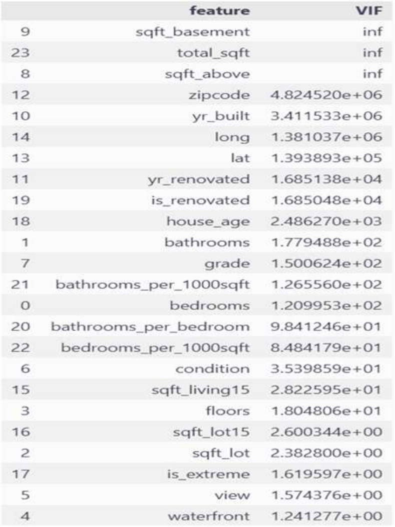
- Correlation analysis identified living area, construction quality, and location as the strongest drivers of price.
- Derived ratio-based features (such as bathrooms per bedroom and size-normalized measures) added additional explanatory value.
- Strong multicollinearity was observed among size-related variables, guiding feature selection decisions rather than column removal.

Correlation



```
grade                       0.703679
total_sqft                  0.695146
sqft_living15               0.619305
sqft_above                  0.601551
bathrooms                   0.551230
lat                         0.448897
is_extreme                  0.354805
view                        0.346582
bedrooms                    0.343355
sqft_basement               0.316892
floors                      0.310633
bathrooms_per_bedroom       0.303355
waterfront                  0.174686
yr_renovated                0.114471
is_renovated                0.114096
sqft_lot                    0.100022
sqft_lot15                  0.092272
yr_built                    0.080600
long                        0.050894
condition                   0.038901
id                         -0.003726
date                       -0.005200
zipcode                    -0.038800
house_age                  -0.080515
bathrooms_per_1000sqft     -0.279236
bedrooms_per_1000sqft      -0.541382
Name: log_price, dtype: float64
```

Multicollinearity



	feature	VIF
9	sqft_basement	inf
23	total_sqft	inf
8	sqft_above	inf
12	zipcode	4.824520e+06
10	yr_built	3.411533e+06
14	long	1.381037e+06
13	lat	1.393893e+05
11	yr_renovated	1.685138e+04
19	is_renovated	1.685048e+04
18	house_age	2.486270e+03
1	bathrooms	1.779488e+02
7	grade	1.500624e+02
21	bathrooms_per_1000sqft	1.265560e+02
0	bedrooms	1.209953e+02
20	bathrooms_per_bedroom	9.841246e+01
22	bedrooms_per_1000sqft	8.484179e+01
6	condition	3.539859e+01
15	sqft_living15	2.822595e+01
3	floors	1.804806e+01
16	sqft_lot15	2.600344e+00
2	sqft_lot	2.382800e+00
17	is_extreme	1.619597e+00
5	view	1.574376e+00
4	waterfront	1.241277e+00

Next Steps

- Select a final feature subset for modeling and build baseline regression models.
- Evaluate model performance and interpret results to guide further refinement.