Prediction of property sale price in Seattle

Machine Learning

Course Project

Goriaev Nickolai Islamov Ayrat Logviniuk Mikhail

Mityaev Vladimir Terentyev Egor

FCS HSE MDS, 2025

Business problems

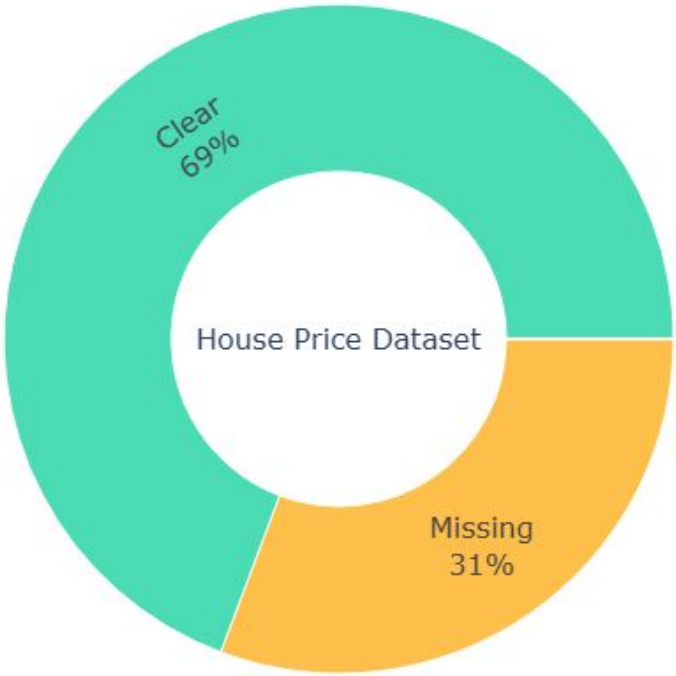
Uncertainty in pricing

Incorrect pricing = longer time on market or loss of potential income

Buyer mistrust or confusion

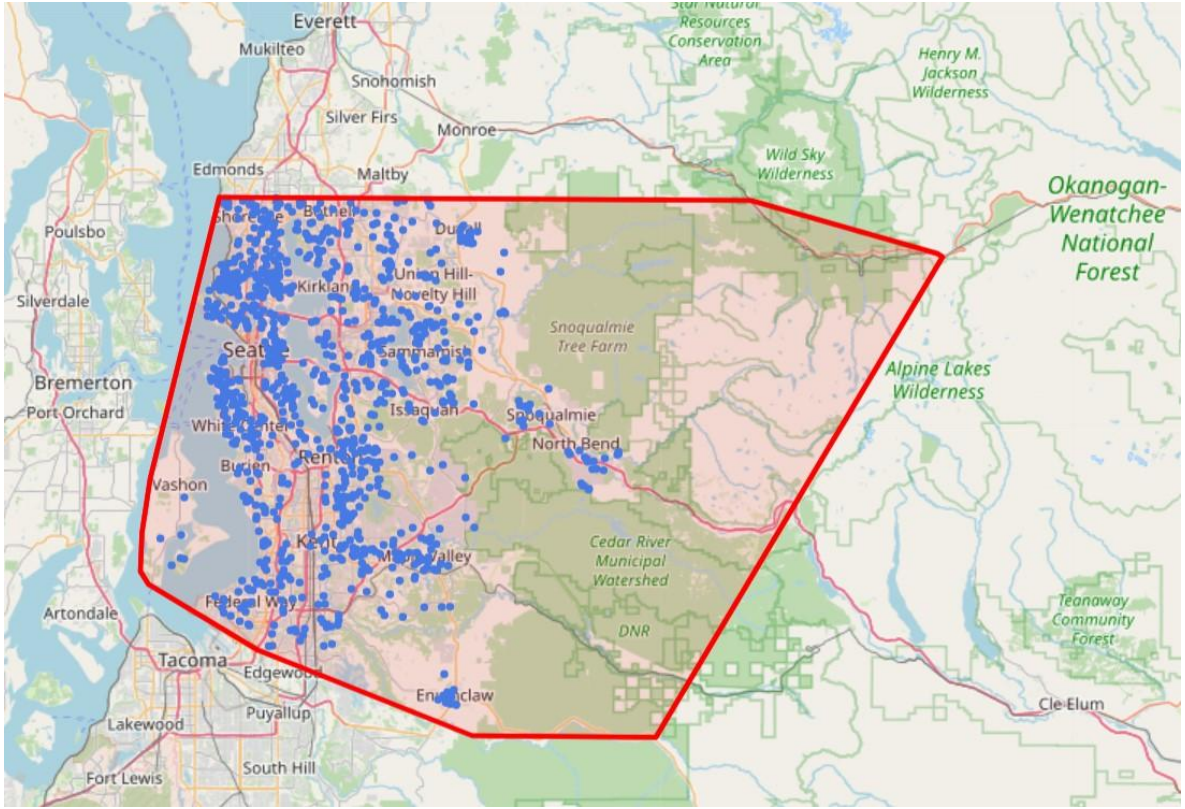
Buyers struggle to evaluate whether a listed price is fair => fewer real estate purchases

Data Quality

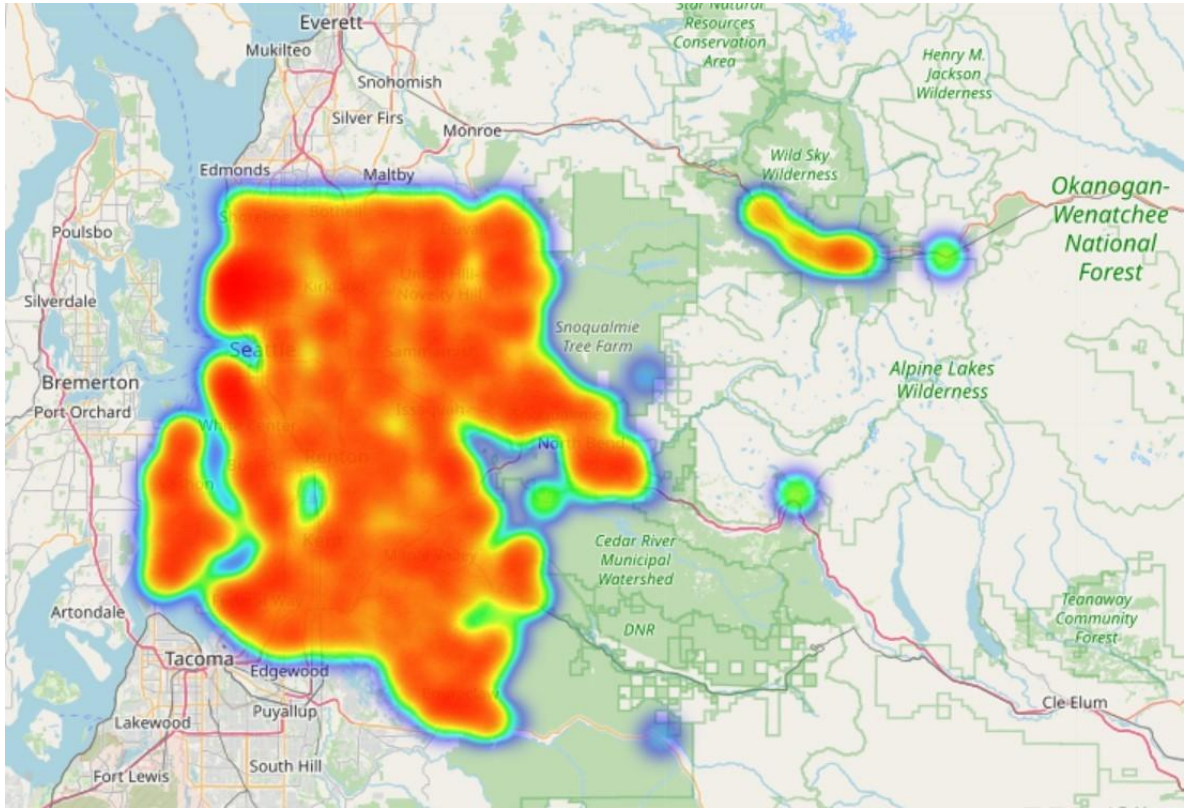


Total: 200,000 rows

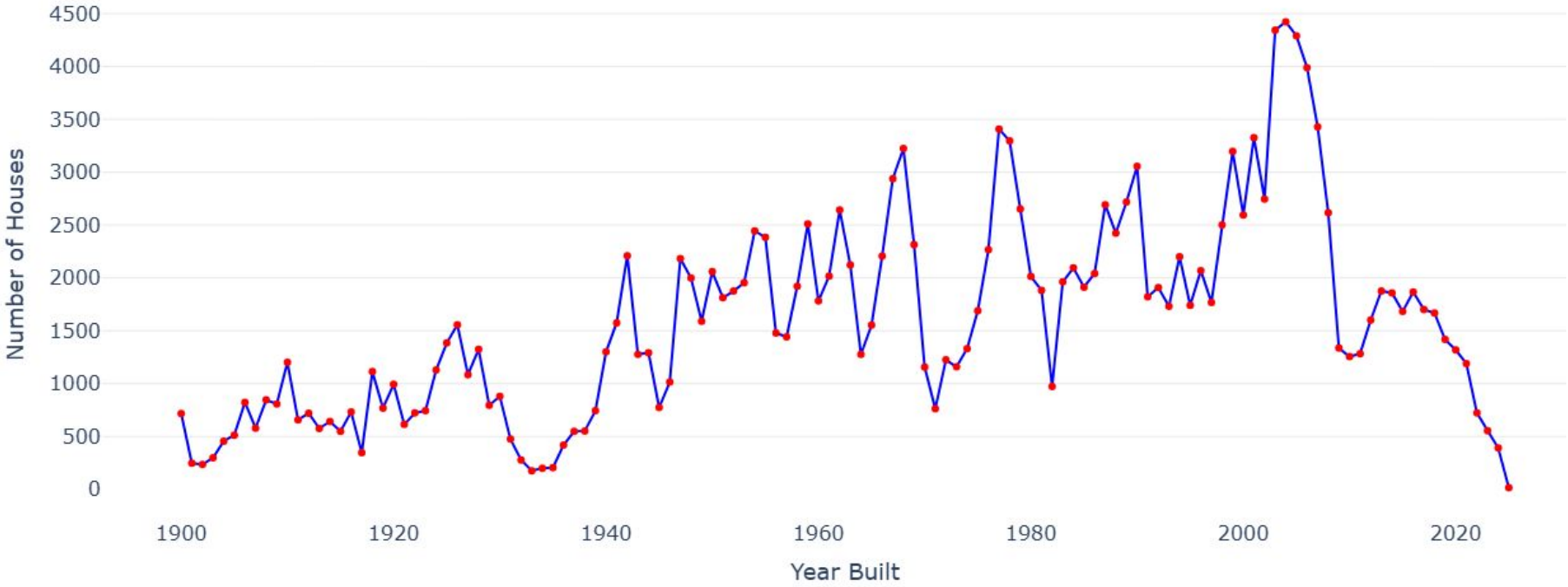
Region


Everett
Mukilteo
Snohomish
Silver Firs
Monroe
Edmonds
Maltby
Poulsbo
Silverdale
Kirkland
Union Hill-
Novelty Hill
Sammamish
Bremerton
Port Orchard
Issaquah
Snoqualmie
North Bend
Renton
Vashon
Kent
Artondale
Tacoma
Edgewood
Puyallup
Lakewood
Fort Lewis
South Hill
Cle Elum
Resources
Conservation
Area
Henry M.
Jackson
Wilderness
Wild Sky
Wilderness
Okanogan-
Wenatchee
National
Forest
Snoqualmie
Tree Farm
Alpine Lakes
Wilderness
Cedar River
Municipal
Watershed
DNR
Teanaway
Community
Forest

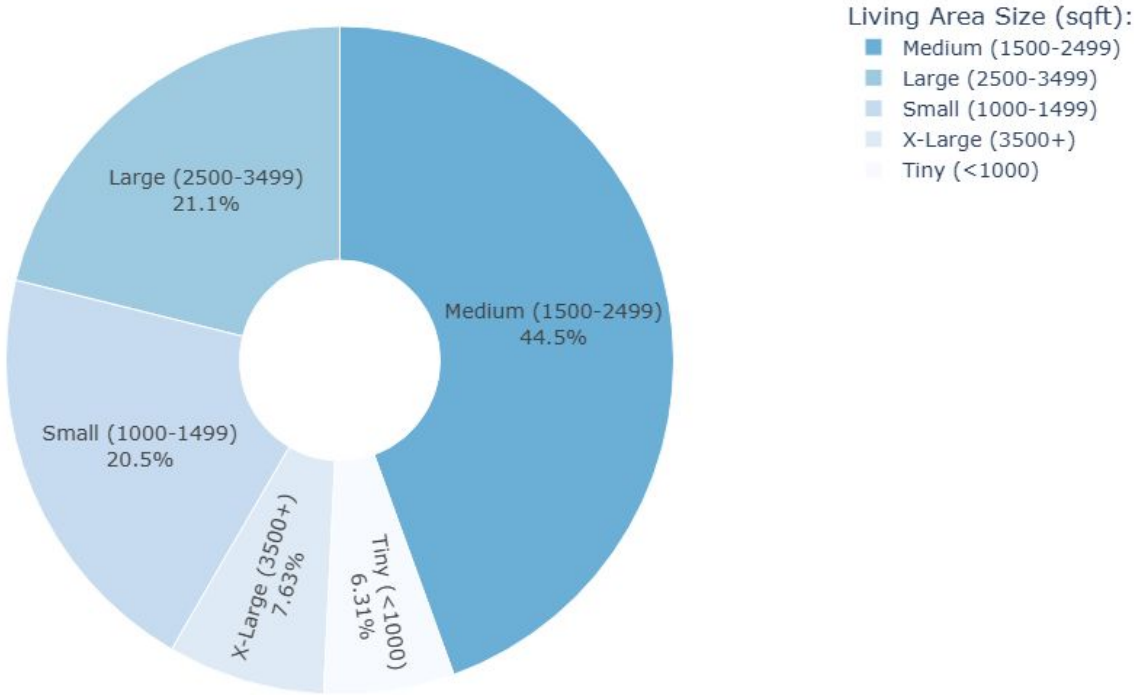
Property Density



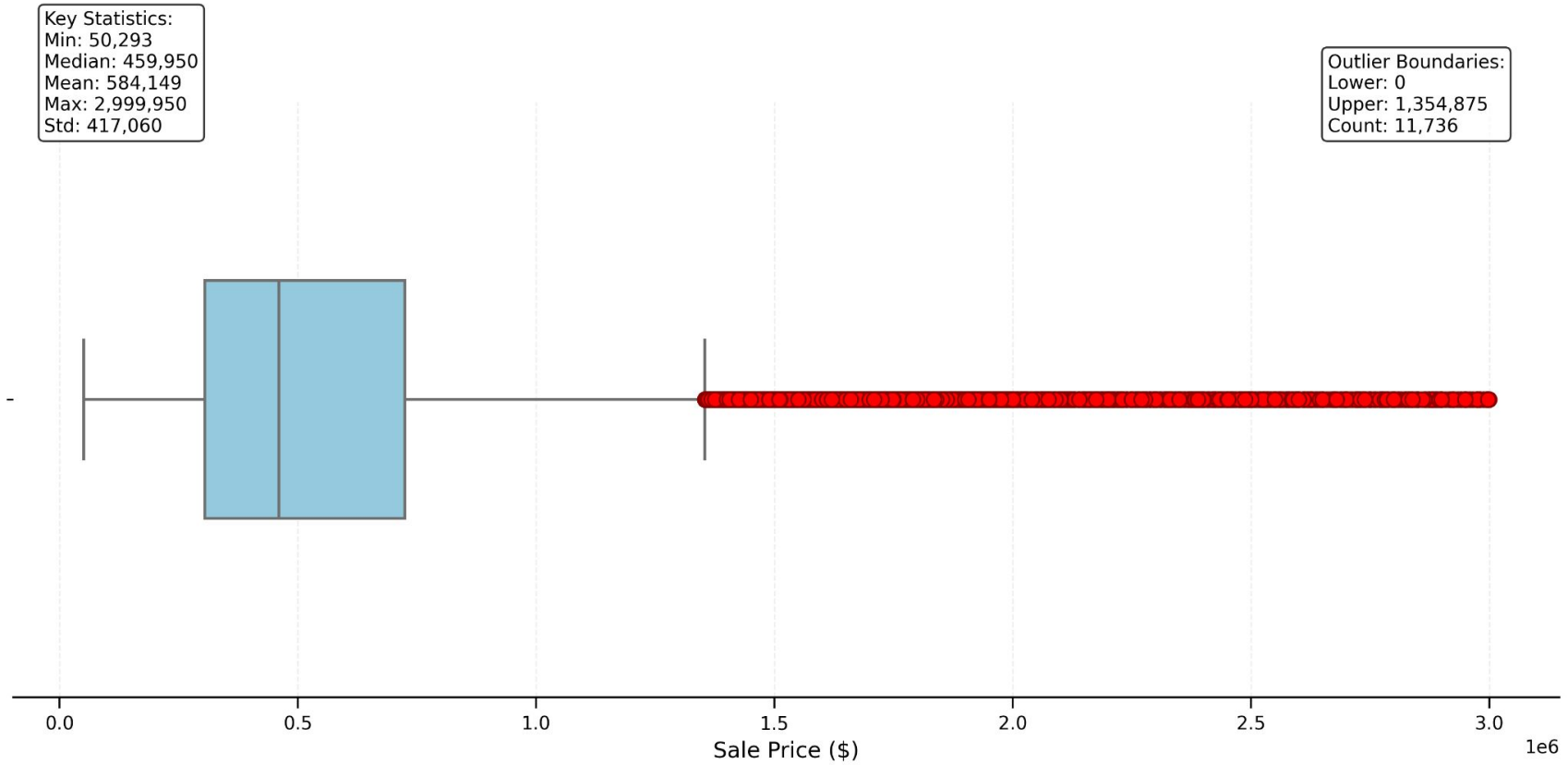
Year Built


4500
4000
3500
3000
2500
2000
1500
1000
500
0
Number of Houses
1900
1920
1940
1960
1980
2000
2020
Year Built

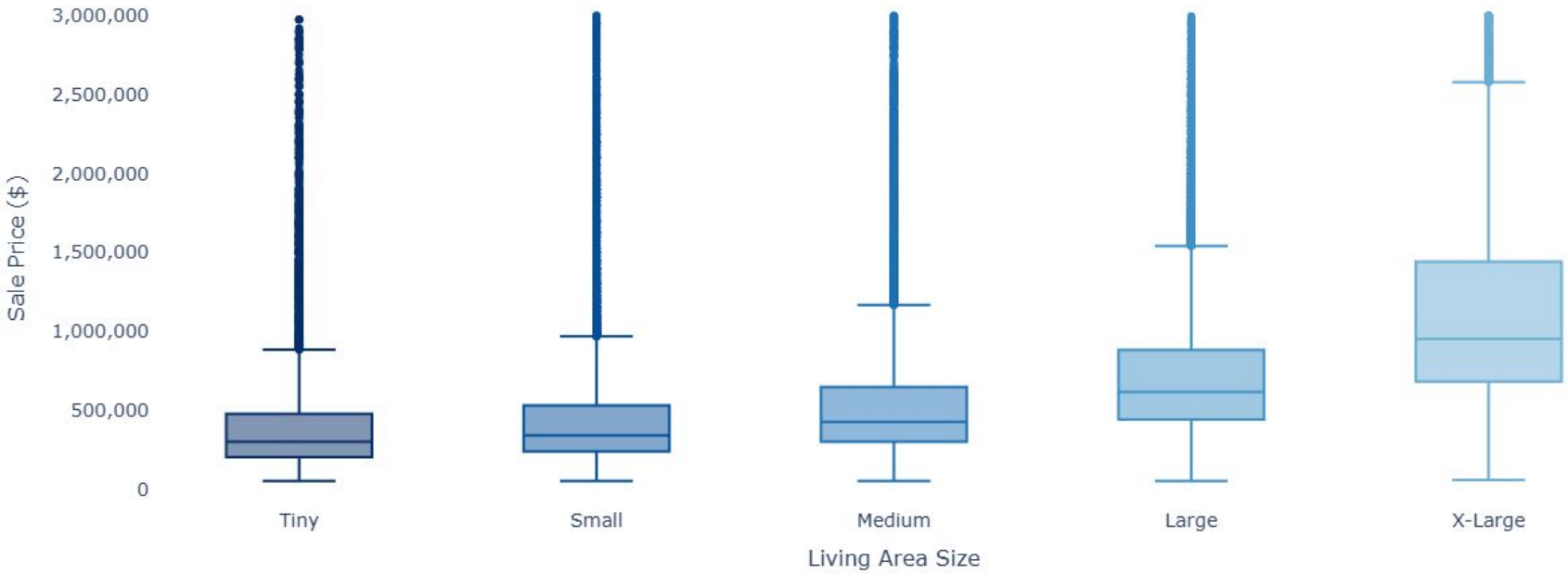
Living Area Size



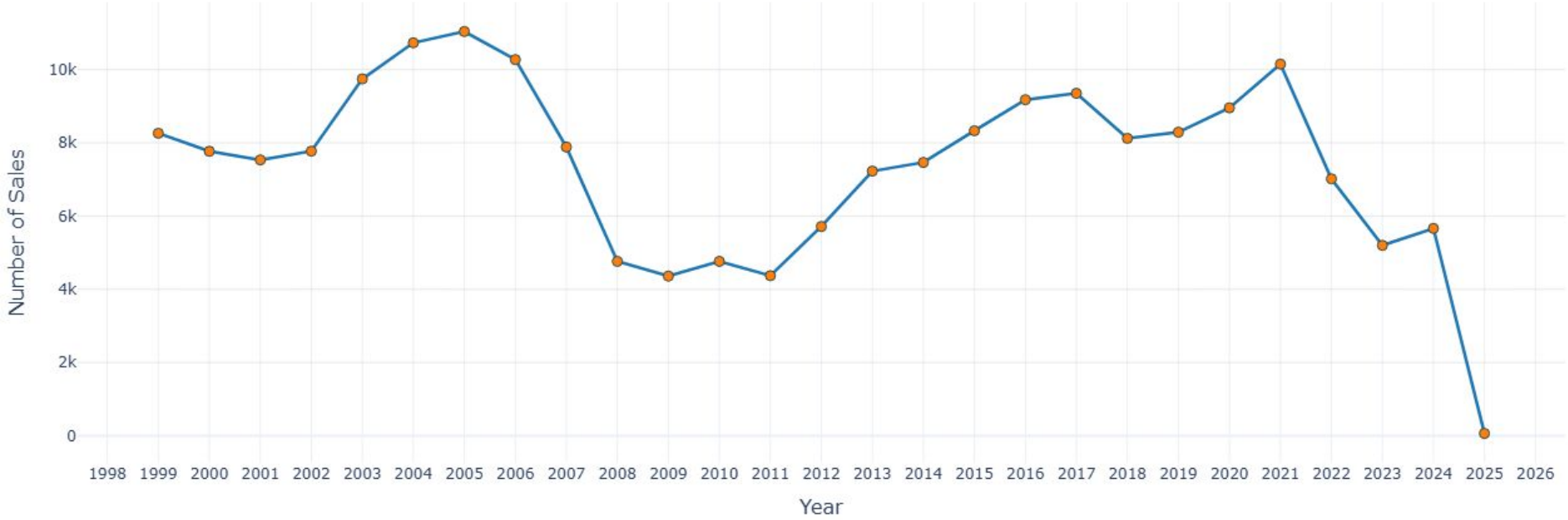
Sale Price


Key Statistics:
Min: 50,293
Median: 459,950
Mean: 584,149
Max: 2,999,950
Std: 417,060
Outlier Boundaries:
Lower: 0
Upper: 1,354,875
Count: 11,736
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Sale Price ($)
1e6

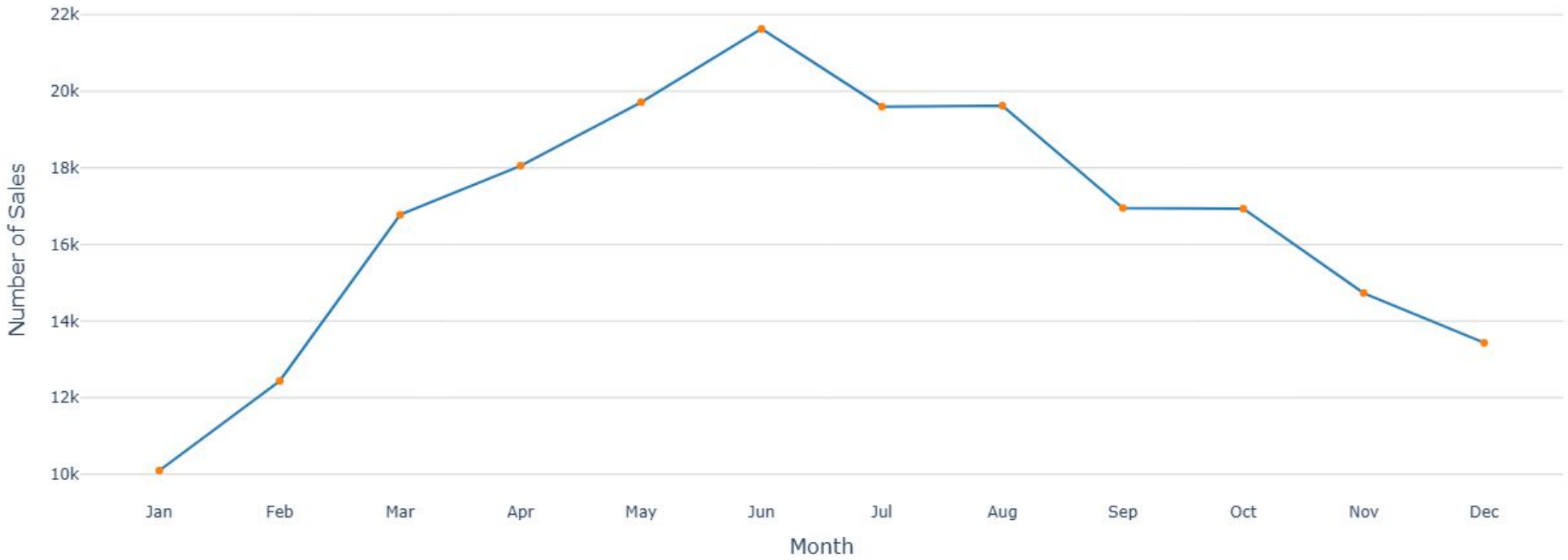
Sale Price by Living Area Size



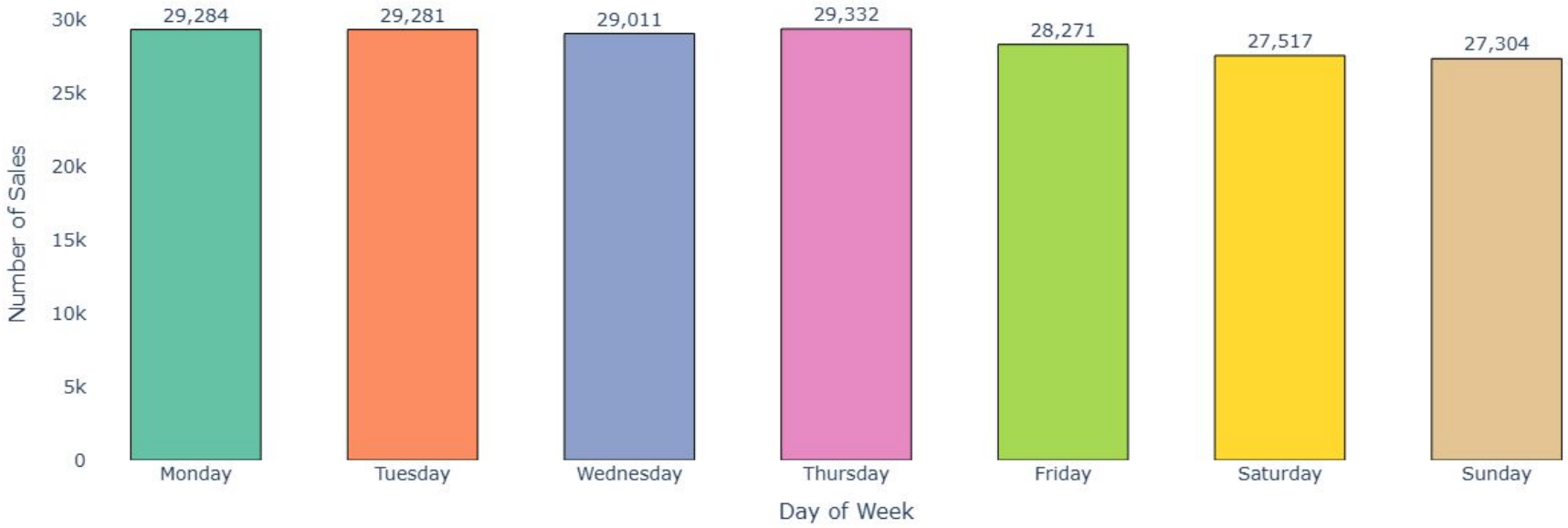
Sales Count By Year (1999-2025)



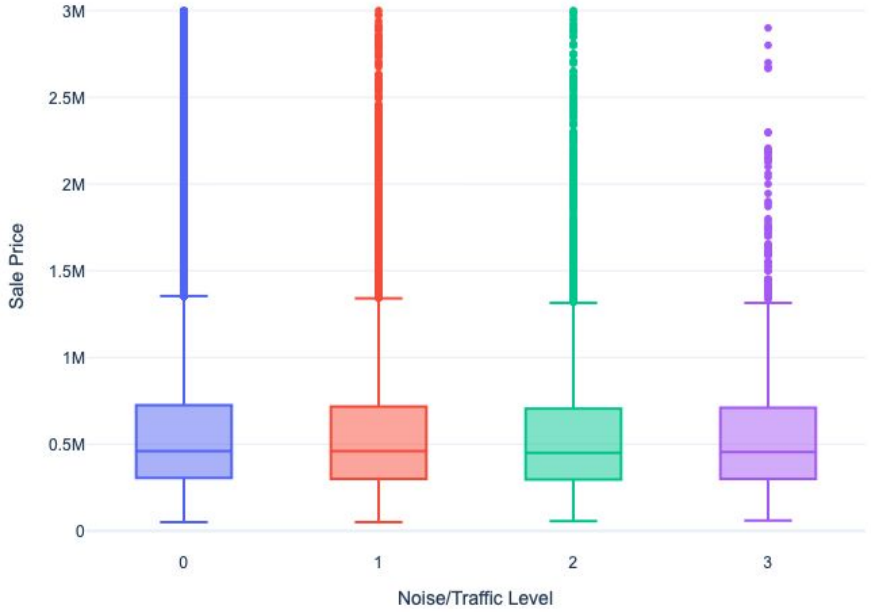
Sales Count By Month (1999-2025)

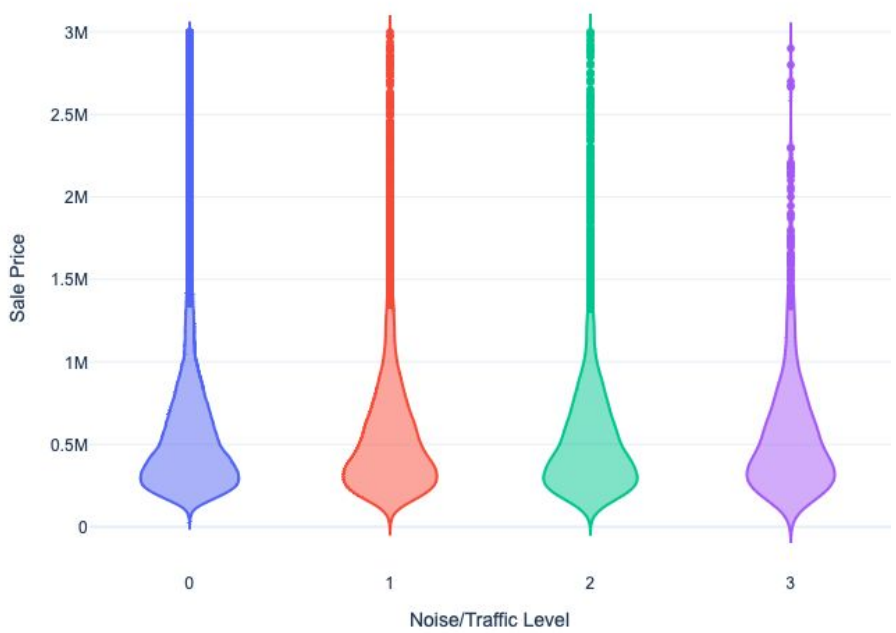


Sales Count By Week Day (1999-2025)



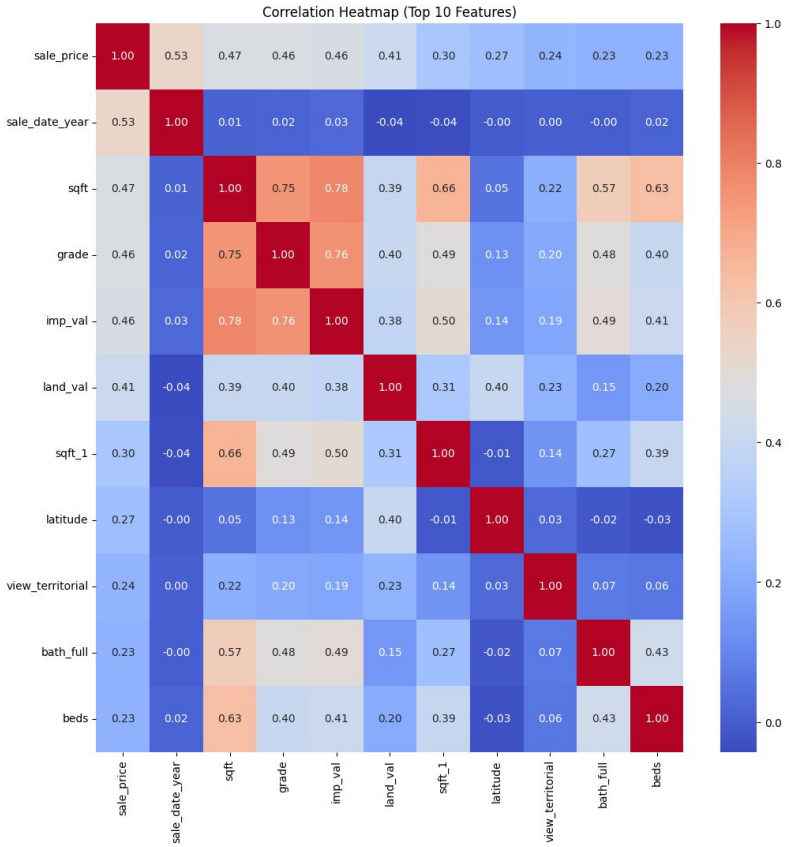
Distribution of Sale Price by Noise/Traffic Level





Top 15 Cities by Average Price

Correlation



ML 1 – Linear Regression and Ridge LR

Baseline LR

All features included
Target Log Transformation
For **Categorical** features:
- Label Encoding (for binary)
- One-Hot Encoding (for multilevel)

For **Numerical** Feature:
- MinMaxScale

RMSE (log): 0,4184
R^2 (log): 0.5422

RMSE (original): 307266
R^2 (original): 0.4582

LR – VIF-based correction and control of multicollinearity

Feature exclusion based on:
- Extremely high VIF
- Low correlation with Target

Result:
- All features VIF < 200
- Decreased multicollinearity risk
- Excluded perfect multicollinear features

RMSE (log): 0,4193
R^2 (log): 0.5405

RMSE (original): 307660
R^2 (original): 0.4568

many features

multicollinearity risk

Ridge LR

Generated key features:
sqft*grade,
sqft*beds,
sqft*bathrooms_total,
condition*grade,
sqft^2,
sqft^3,
Cyclical Time Transformation
for day/week/month

ML 1 – Linear Regression and Ridge LR

Ridge LR

Previously selected features
Tuned hyperparameter:
alpha: 0.02976

- Stability of coefficients
- Robustness to multicollinearity
- Reduction of overtraining

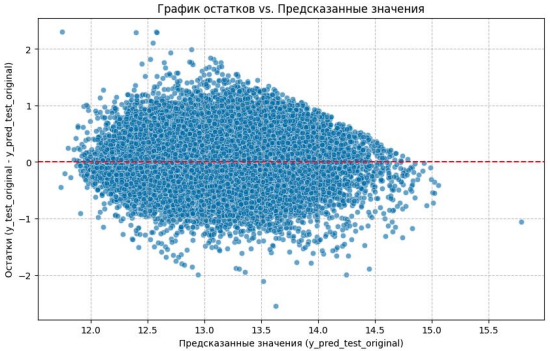
RMSE (log): 0,4193
R^2 (log): 0.5405

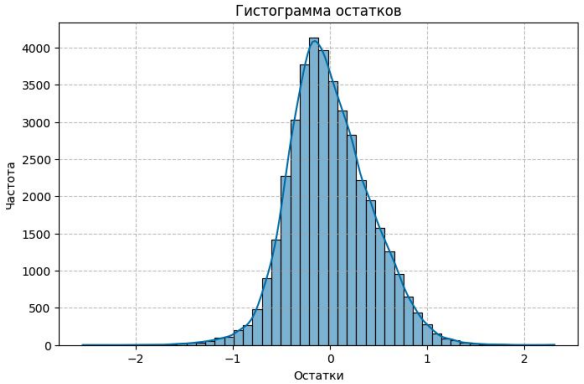
RMSE (original): 307660
R^2 (original): 0.4568

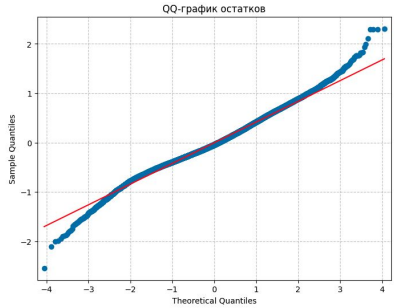
Coefficients (top 10)

sqft	6.406928
sqft*grade	-4.742752
sqft^2	-2.765552
sqft*beds	2.637684
grade	2.269869
sqft_lot	2.203943
sqft*bathrooms_total	-2.043469
imp_val	1.789285
garb_sqft	1.330570
bathrooms_total	1.316744

Linear Regression Assumptions:



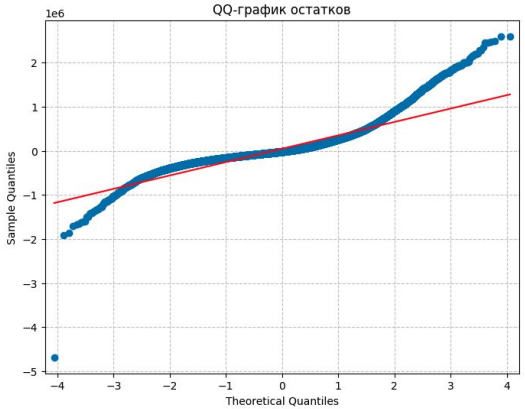


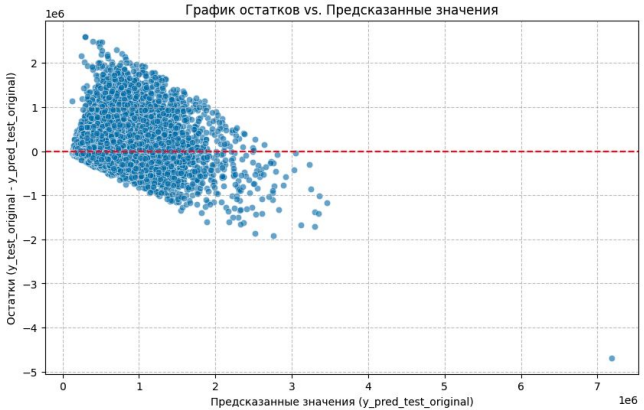


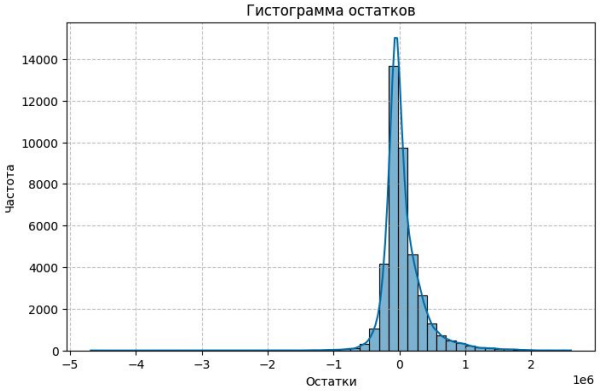
ML 1 - Linear Regression and Ridge LR

Linear Regression Assumptions on Restored Target - **Failed**

- prediction errors in dollars increase with increasing house value.
- there is a heavy tail of large negative errors (overestimation).
- presence of significant outliers: the model is highly erroneous on individual properties, especially underestimating them







ML 2 – Decision tree

DecisionTreeRegressor

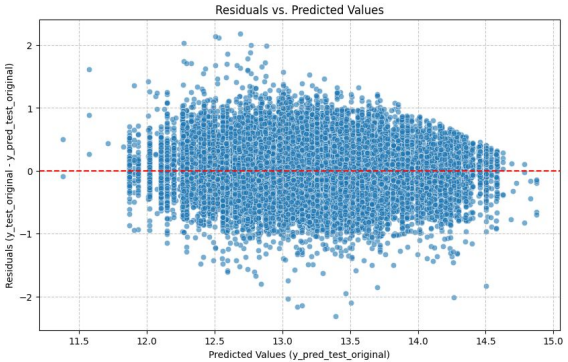
Significant added features

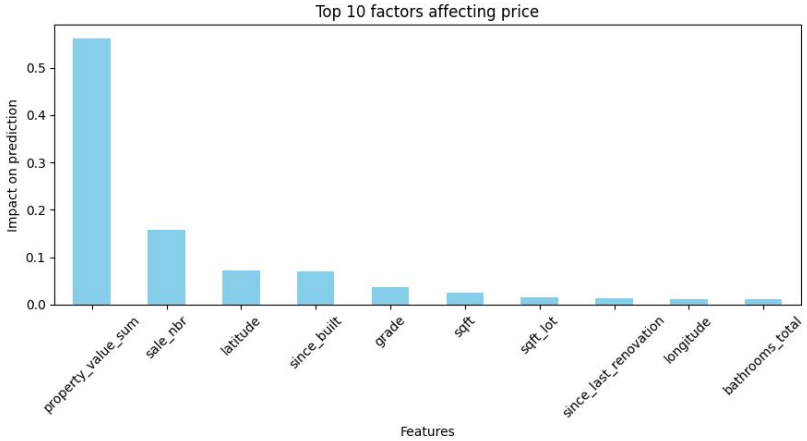
- House age,
- Facilities: waterfront, golf, greenbelt
- Distance to city center

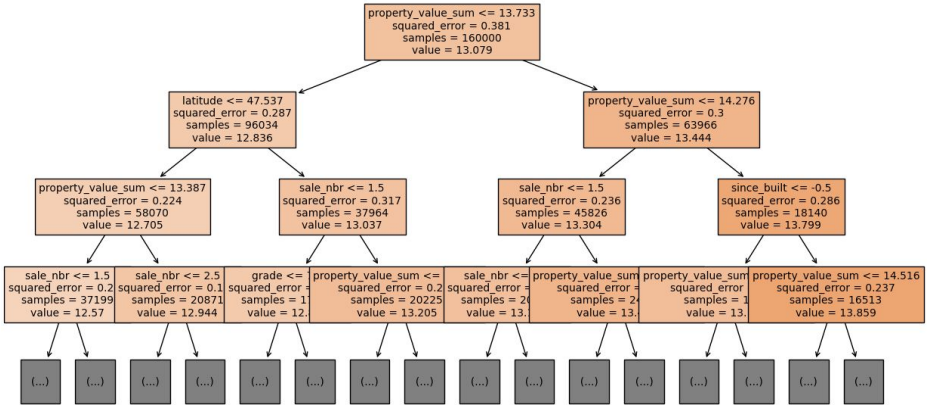
All numeric features included

Log Transformation

Train RMSE (original): 256389
R^2: 0.6219
Test RMSE (original): 268463
R^2: 0.5864

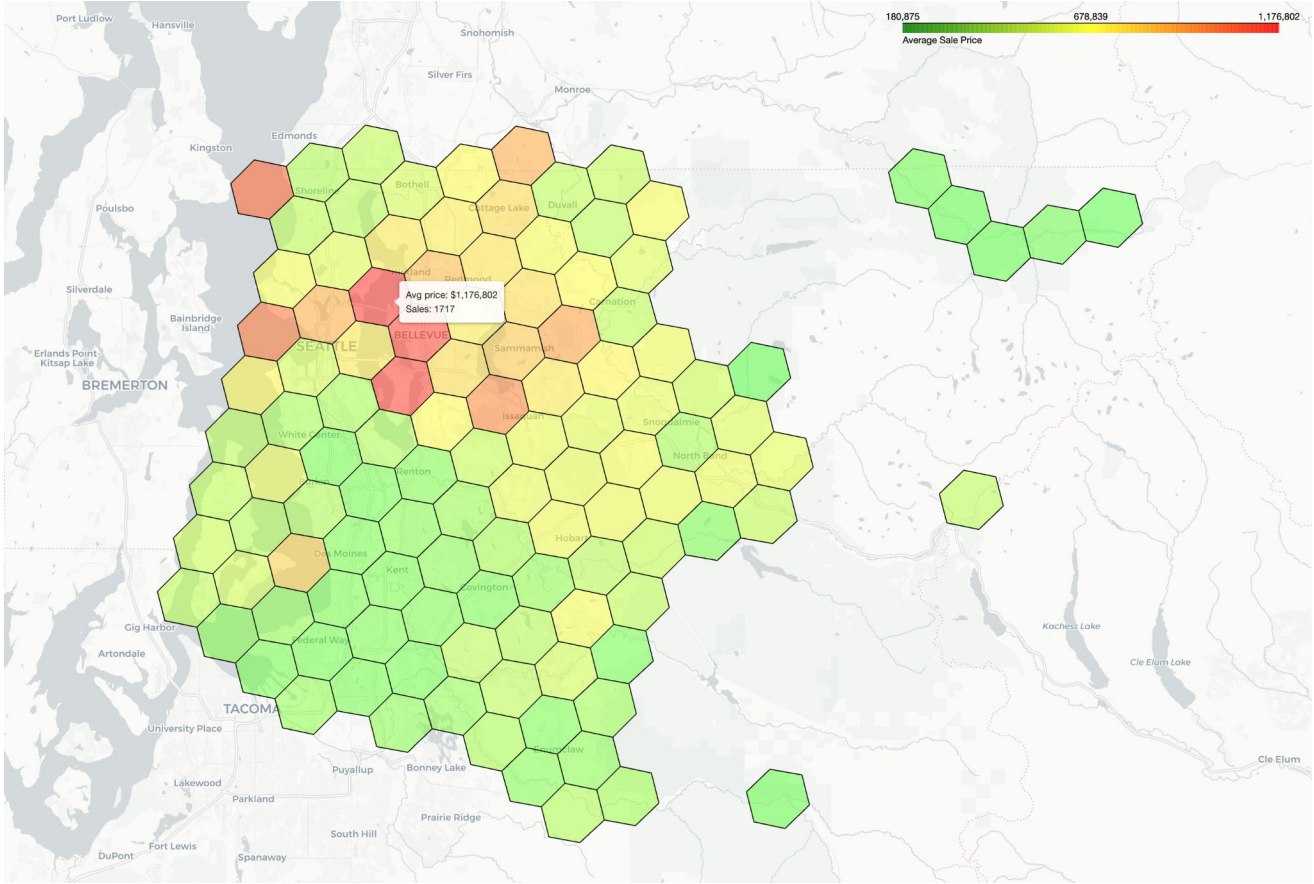




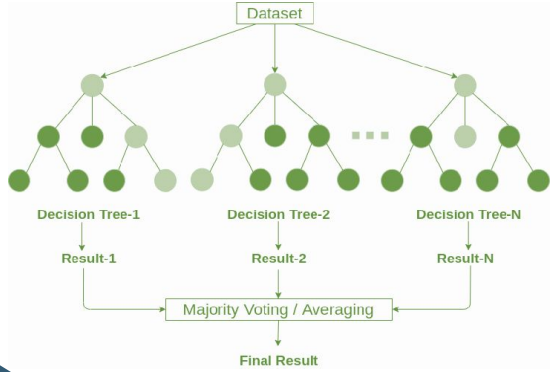


RANDOM EDA MOMENT

Map regions with average sale price and # properties



ML 3 – Random Forest



Reiterated training and removed bottom 20 features by importance

For **Categorical** features:

- Ordinal encoding
- High cardinality - target Encoding

rmse: 270,756
R^2: 0.58
mse: 73,309,088,633

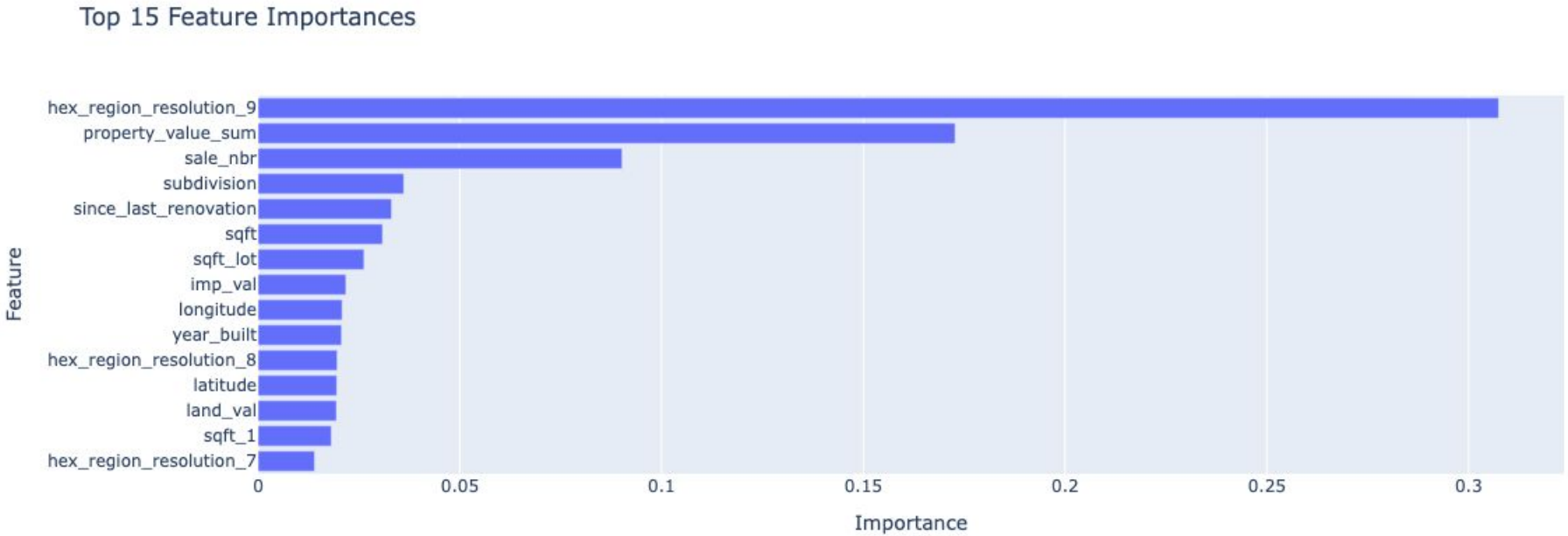
After hyperparameters tuning

params = {'n_estimators': 310, 'max_depth': 23, 'min_samples_split': 5, 'min_samples_leaf': 2, 'max_features': 0.7426985171903291}

mse: 72,683,692,676
R^2: 0.59
rmse: 269,599

ML 3

since_last_renovation - # years since last renovation (or age if never renovated)



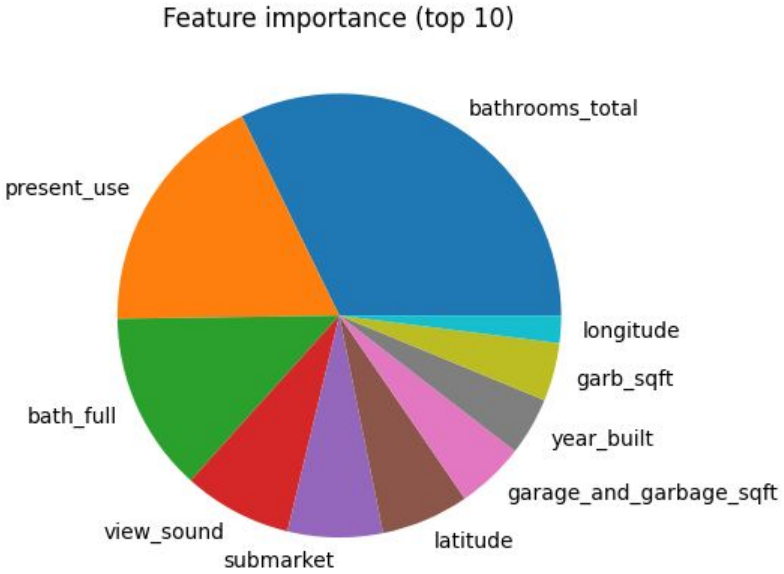
ML 4 – Neural Network

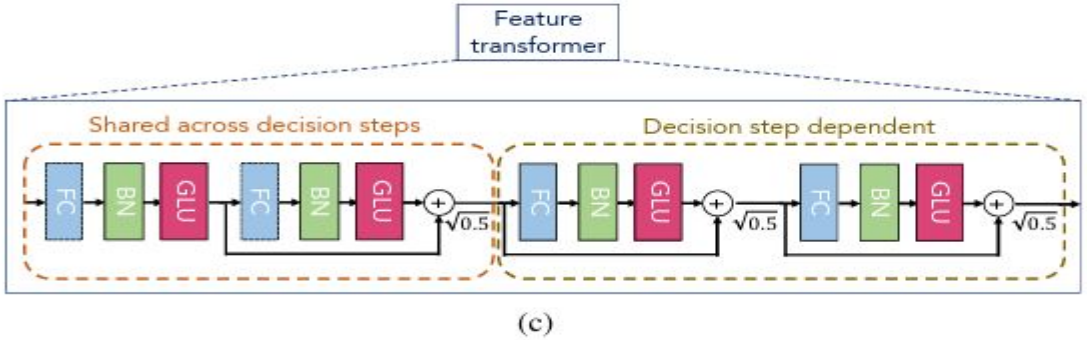
TabNet*

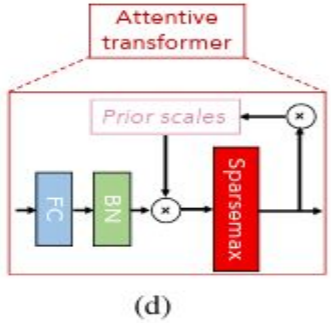
All features included
Target Log Transformation
For **Categorical** features:
- Label Encoding

* Sercan O. Arık, Tomas Pfister, 2019

Final validation metrics :
RMSE: 0.3768
MSE(log): 0.142
R^2 (log): 0.63







Business benefits

1. **Pricing Uncertainty → Data-Driven Confidence**

Our ML model provides accurate, explainable price predictions based on real market data.

- Sellers receive data-backed price guidance, reducing over- or underpricing.
- This leads to shorter time on market and more completed transactions.
- As a result, the platform earns more through transaction volume and seller retention.
- Internal teams spend less time on manual valuation — saving costs.

2. **Buyer Mistrust → Transparent Benchmarking**

By surfacing model-based price estimates alongside listings, we give buyers a reference point.

- Transparent price estimates boost buyer confidence and decision speed.
- Buyers are more likely to engage, inquire, and purchase — increasing conversion rates.
- The platform gains a reputation for fairness and reliability, improving customer loyalty.