Hogwarts School of
Data Science Magic

Data Driven Approach to Real Estate

Predicting King Country House-Prices

Presented by House of Gryffindor

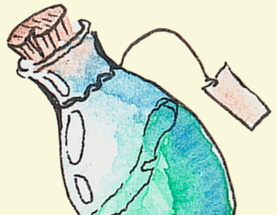
Our Witches

Marta

Vanuhi

Annie

Faniia



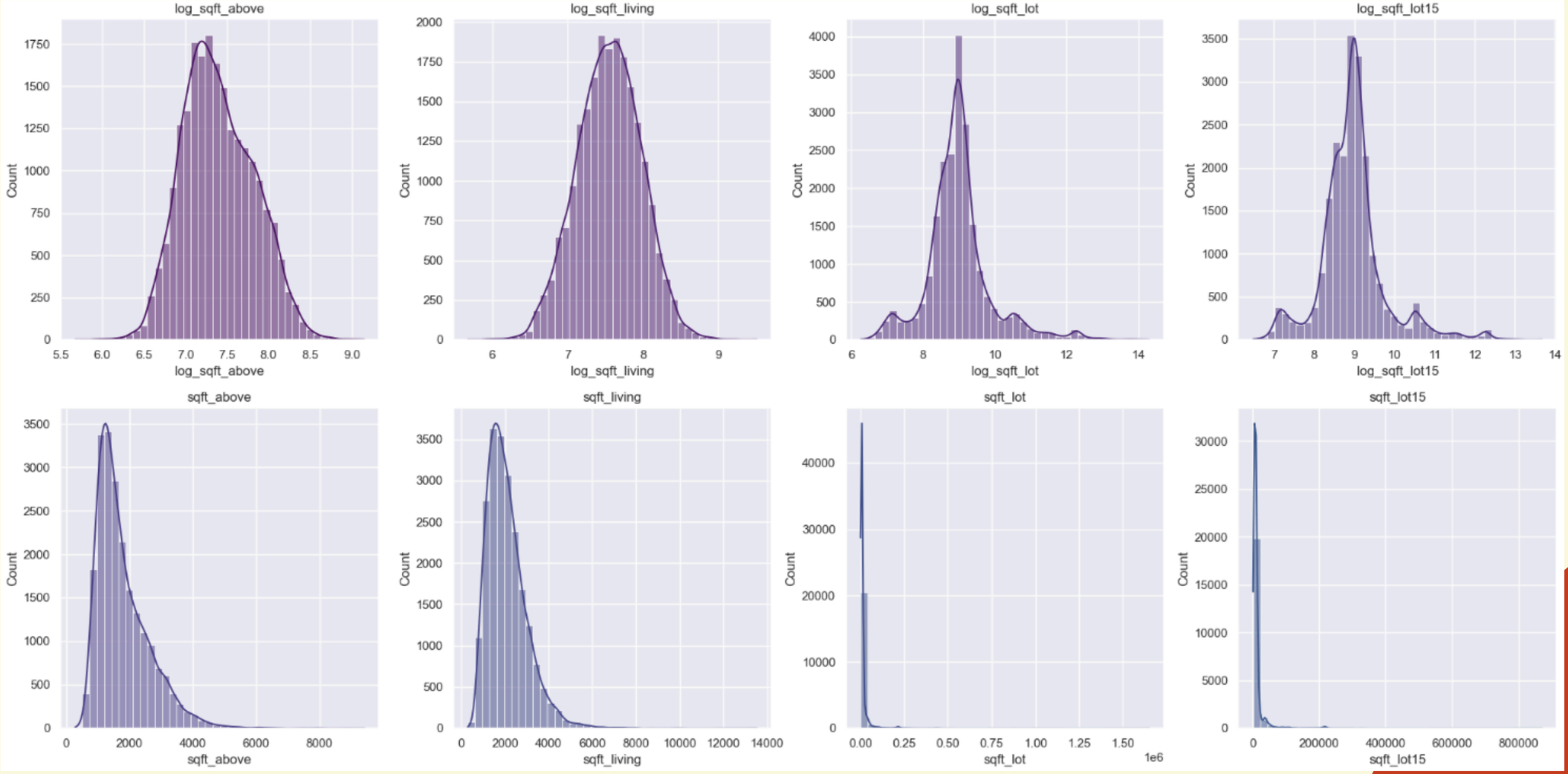
Purpose of the Project

In this presentation, we explore how data-driven models and machine learning techniques can predict housing prices in King County. By analyzing key features we uncover key insights that help improve price estimation accuracy—benefiting buyers, sellers, and investors alike.

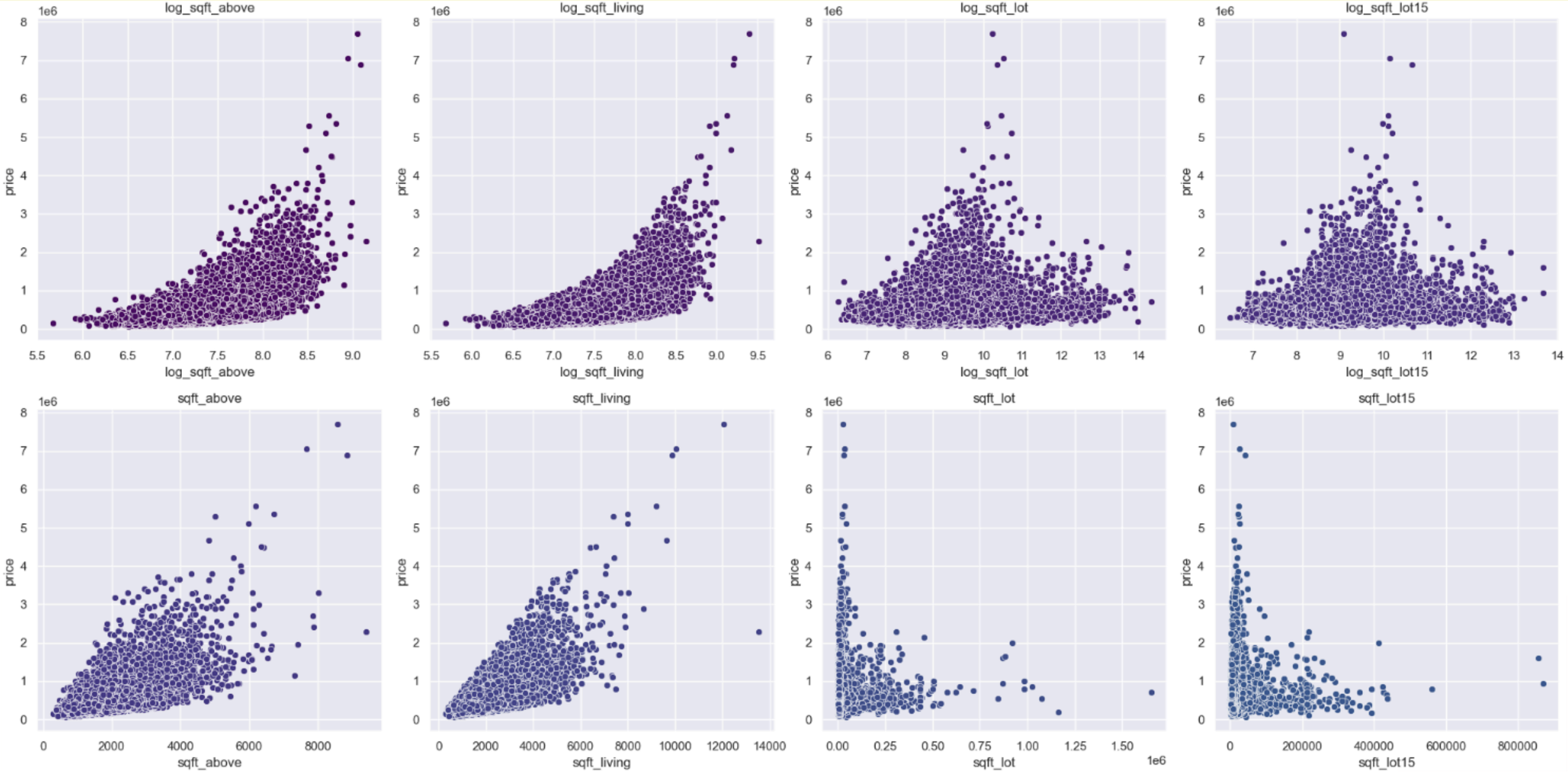
Investigated Regression Models

- LINEAR
- RIDGE
- LASSO
- XGBOOST
- DECISION TREE
- RANDOM FOREST

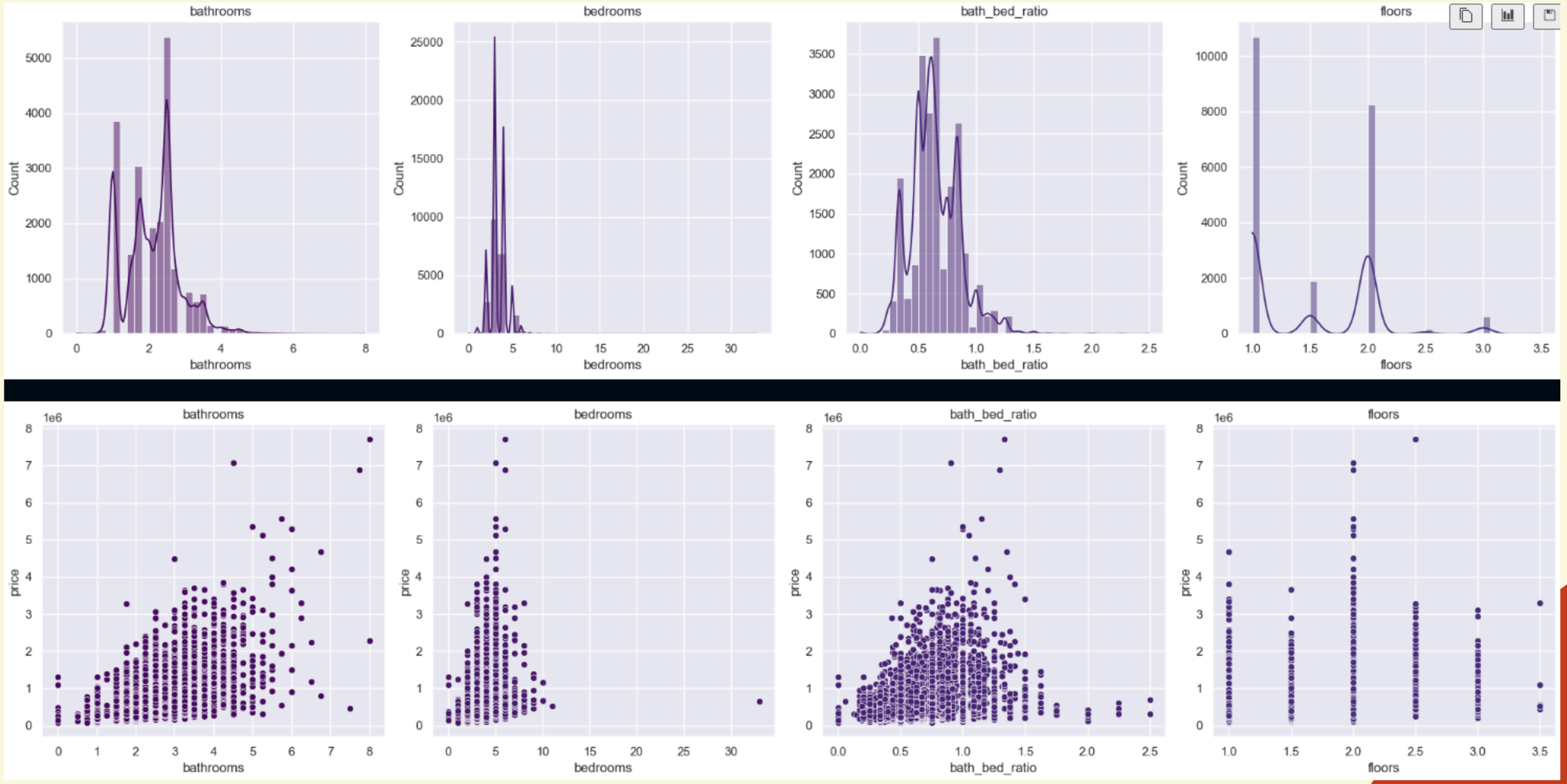
Linear regression



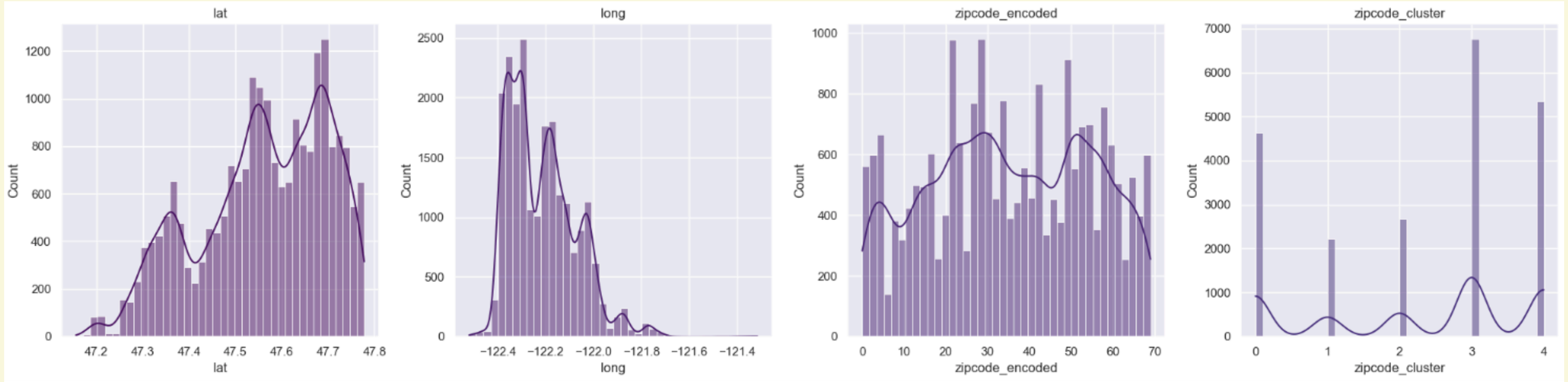
Linear regression

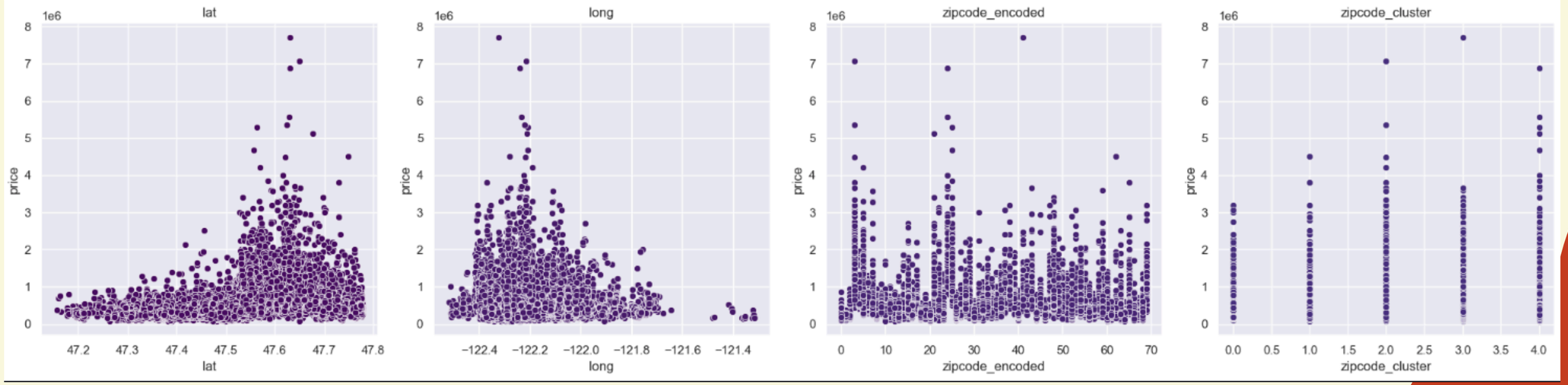


Linear regression

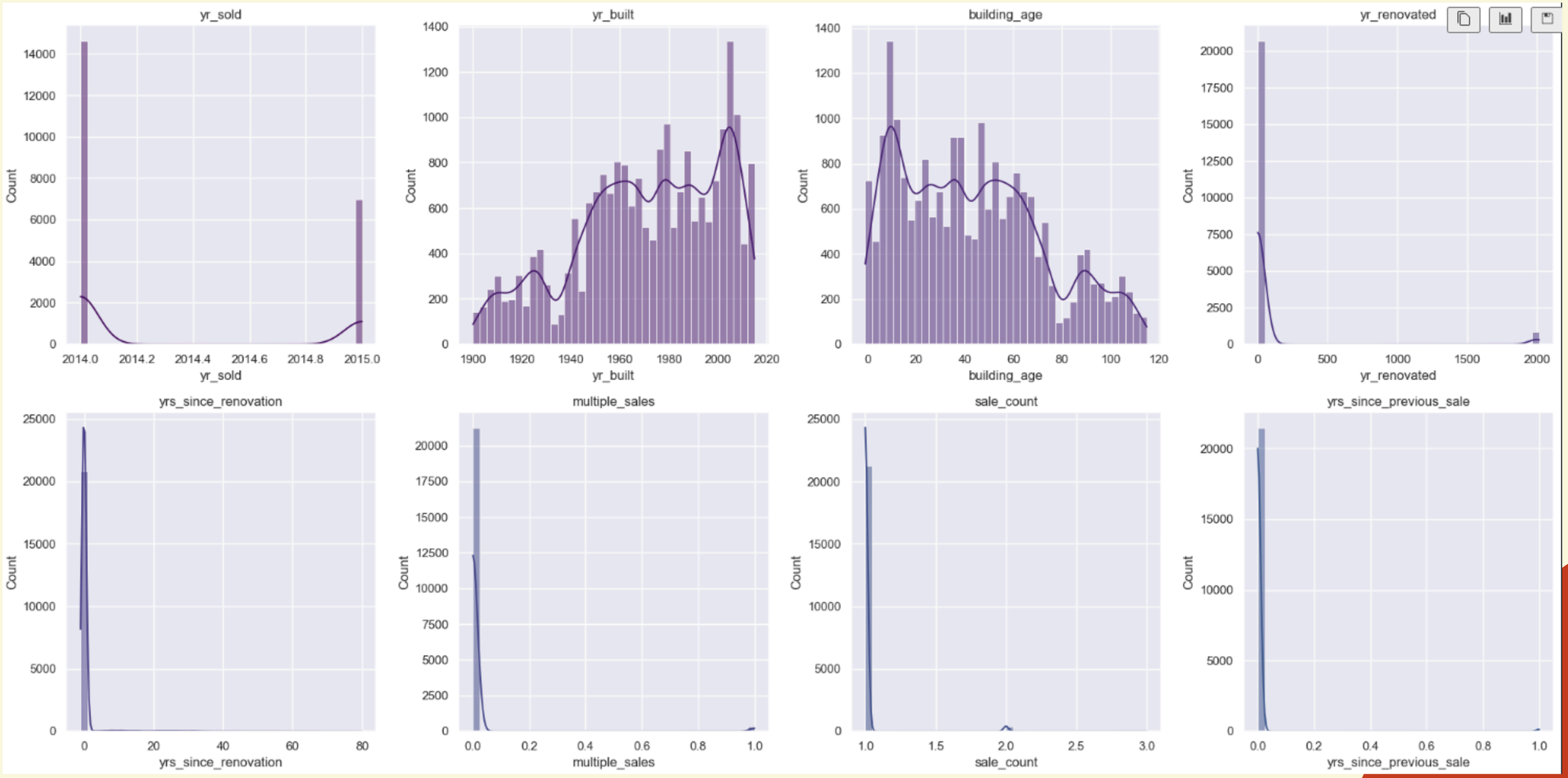


Linear regression

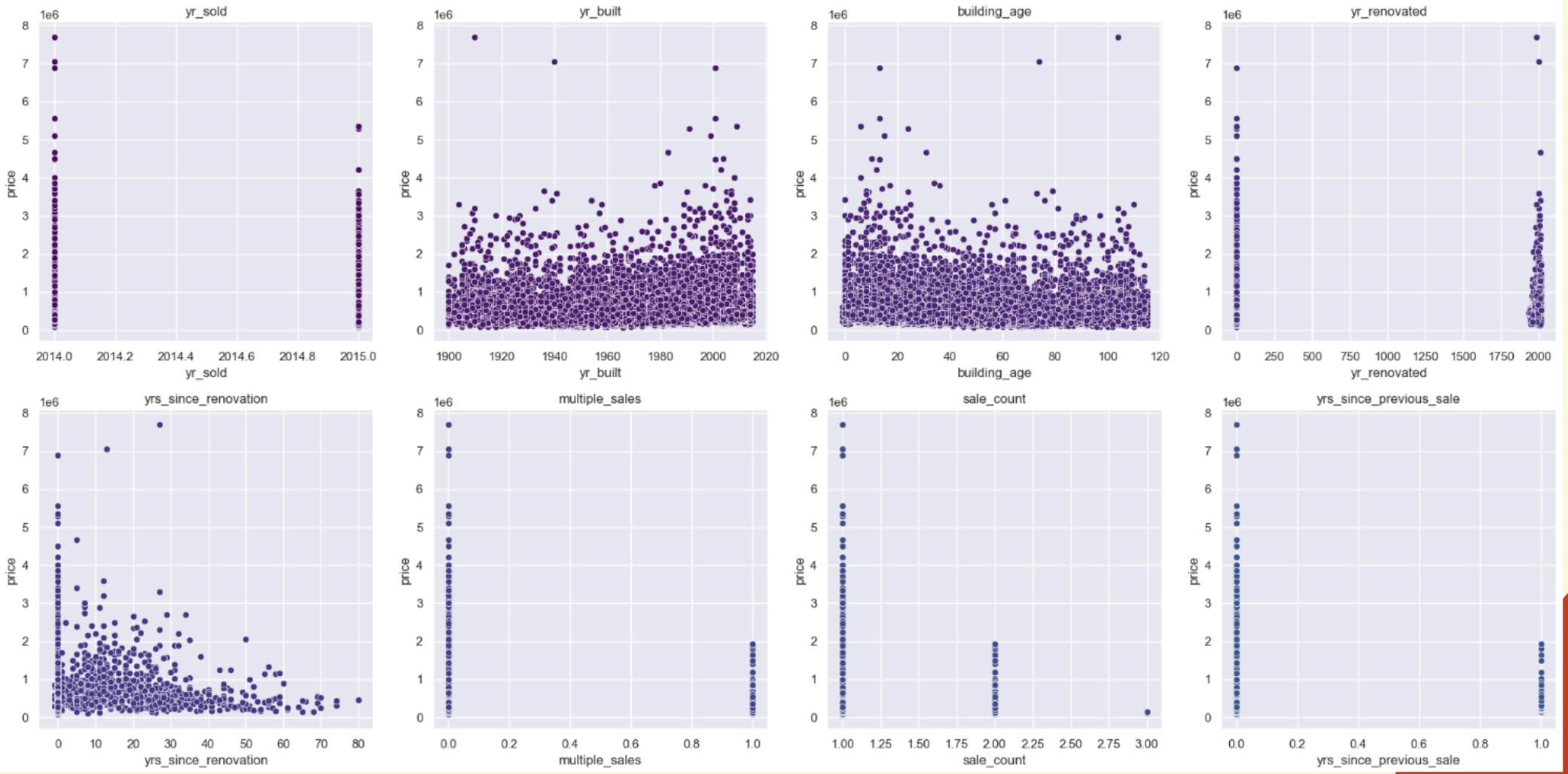




Linear regression



Linear regression



Linear regression

	test_size	random_state	R2	MAE	RMSE	MSE
LR_all_numeric_ts0.2_rs15	0.2	15.0	0.7128	124402.4665	203006.7754	41211750865.0296
LR_all_numeric_ts0.3_rs15	0.3	15.0	0.7027	127105.1038	207959.6066	43247197966.1649
LR_all_numeric_ts0.4_rs15	0.4	15.0	0.6937	128083.5874	208299.5182	43388689297.3495
LR_all_numeric_ts0.4_rs100	0.4	100.0	0.6922	129307.3639	215034.9612	46240034517.3156
LR_all_numeric_ts0.3_rs100	0.3	100.0	0.6848	130815.2597	214304.7771	45926537508.0975
LR_all_numeric_ts0.4_rs42	0.4	42.0	0.6846	127676.4889	198895.7666	39559525982.6036
LR_all_numeric_ts0.3_rs42	0.3	42.0	0.6796	128027.7197	198387.5351	39357614087.2996
LR_all_numeric_ts0.2_rs100	0.2	100.0	0.6761	129810.0785	209166.0131	43750421022.2316
LR_all_numeric_ts0.2_rs42	0.2	42.0	0.6627	128602.0373	200743.4795	40297944556.7745
LR_continuous_ts0.2_rs15	0.2	15.0	0.6572	137242.1804	221760.4922	49177715896.3756
LR_continuous_ts0.3_rs15	0.3	15.0	0.6502	138718.2273	225559.0020	50876863388.7365
LR_continuous_ts0.4_rs100	0.4	100.0	0.6472	139688.7847	230201.4899	52992725961.4172
LR_continuous_ts0.4_rs15	0.4	15.0	0.6462	138498.3839	223886.1184	50124994008.5620
LR_continuous_ts0.3_rs100	0.3	100.0	0.6426	140417.2894	228178.5771	52065463028.2362
LR_continuous_ts0.2_rs100	0.2	100.0	0.6377	138160.5922	221212.7521	48935081677.0367
LR_continuous_ts0.4_rs42	0.4	42.0	0.6320	137786.0745	214846.2522	46158912082.8384

Linear regression

LR_continuous_ts0.3_rs42	0.3	42.0	0.6288	138024.3623	213539.9954	45599329635.6152
LR_continuous_ts0.2_rs42	0.2	42.0	0.6088	139763.5222	216210.2223	46746860243.5492
LR_categorical_ts0.2_rs100	0.2	100.0	0.5295	164809.6148	252091.4834	63550116008.7113
LR_categorical_ts0.3_rs42	0.3	42.0	0.5237	165267.7804	241889.8501	58510699559.4682
LR_categorical_ts0.4_rs42	0.4	42.0	0.5211	164724.6863	245102.2935	60075134263.5881
LR_categorical_ts0.2_rs42	0.2	42.0	0.5126	164985.0523	241331.7015	58240990164.5579
LR_categorical_ts0.3_rs100	0.3	100.0	0.5106	166959.9829	267013.8741	71296408966.6561
LR_categorical_ts0.4_rs100	0.4	100.0	0.5021	166878.7962	273486.3469	74794781935.2175
LR_categorical_ts0.4_rs15	0.4	15.0	0.4992	168310.8498	266367.2559	70951515019.2717
LR_categorical_ts0.3_rs15	0.3	15.0	0.4915	170113.1040	271981.0294	73973680372.0369
LR_categorical_ts0.2_rs15	0.2	15.0	0.4717	168868.3993	275309.2053	75795158510.8610
LR_discrete_ts0.3_rs42	0.3	42.0	0.4308	180545.5931	264417.2734	69916494456.5061
LR_discrete_ts0.4_rs100	0.4	100.0	0.4248	183299.6189	293942.0284	86401916064.0133
LR_discrete_ts0.4_rs42	0.4	42.0	0.4246	180281.6286	268673.4676	72185432207.6200
LR_discrete_ts0.2_rs15	0.2	15.0	0.4235	181996.1981	287600.1775	82713862098.4244
LR_discrete_ts0.2_rs42	0.2	42.0	0.4224	180437.1649	262713.0499	69018146605.9900
LR_discrete_ts0.3_rs15	0.3	15.0	0.4218	183724.5359	290006.3022	84103655336.8933
LR_discrete_ts0.3_rs100	0.3	100.0	0.4206	184607.5193	290535.0400	84410609484.2011
LR_discrete_ts0.2_rs100	0.2	100.0	0.4189	184068.1326	280134.0187	78475068427.5596
LR_discrete_ts0.4_rs15	0.4	15.0	0.4141	183391.9570	288094.3237	82998339356.6180

Linear regression

- All numeric features
- Test Size 20%
- Random state 15

Metrics	Values
R2	0.7128
RMSE	203006.7754
MAE	124402.4665

Ridge regression

Model fit with the dataset:

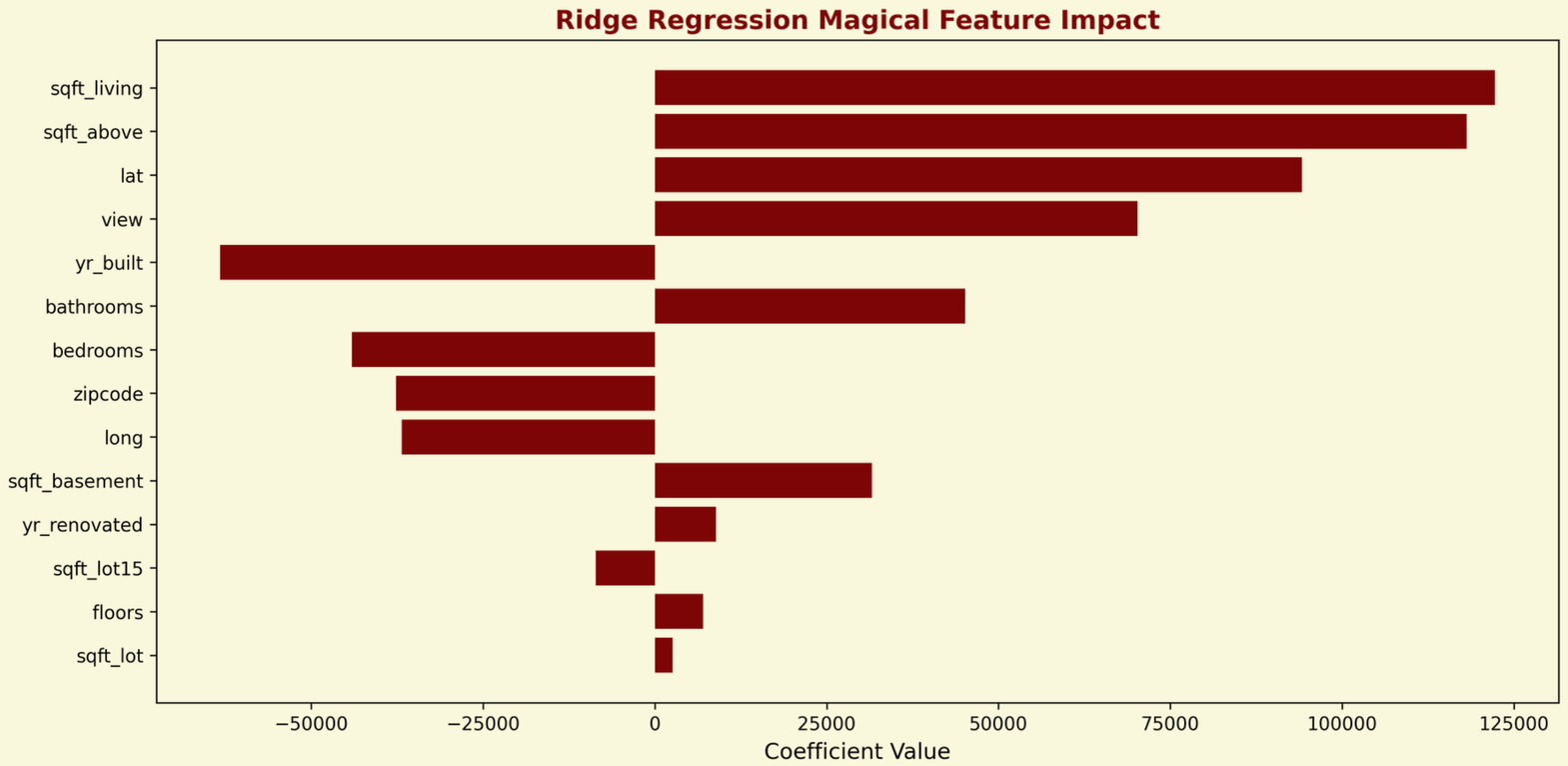
Investigating correlation of correlations with the model results

Idea:

target-feature correlation

intercorrelation

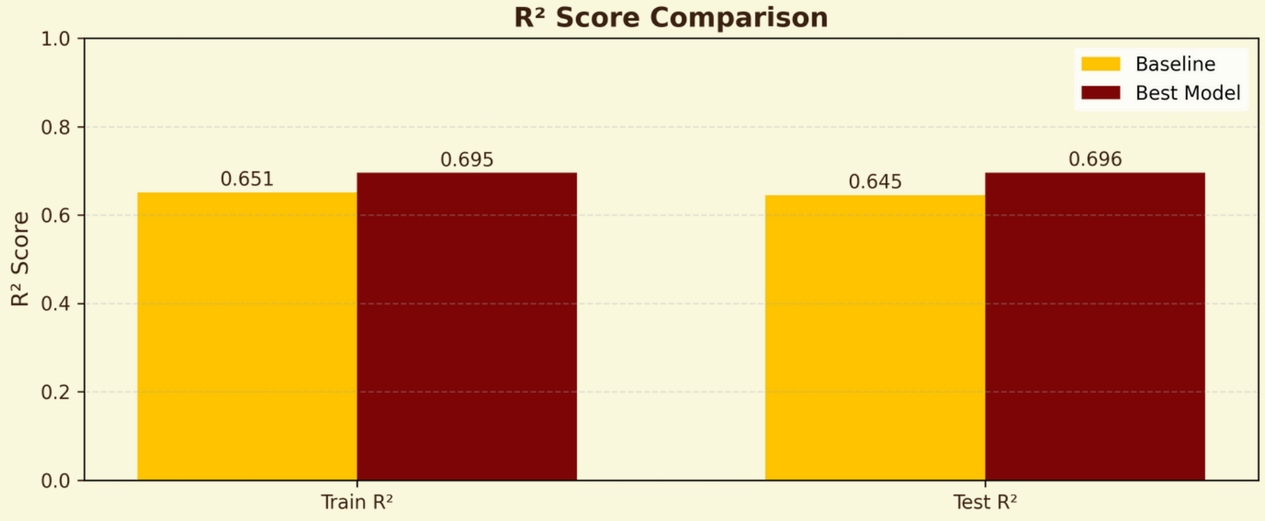
Ridge regression:

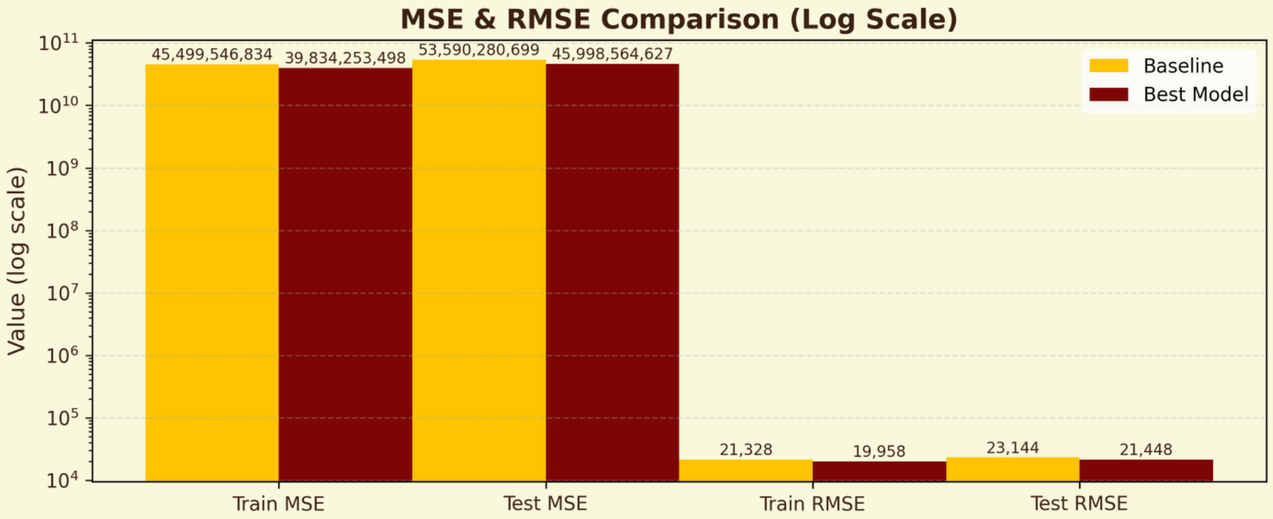


Feature selection model:

Train R2 Score: 0.516
Test R2 Score: 0.518
Train MSE: 63232150589.40324
Test MSE: 72841879071.61089

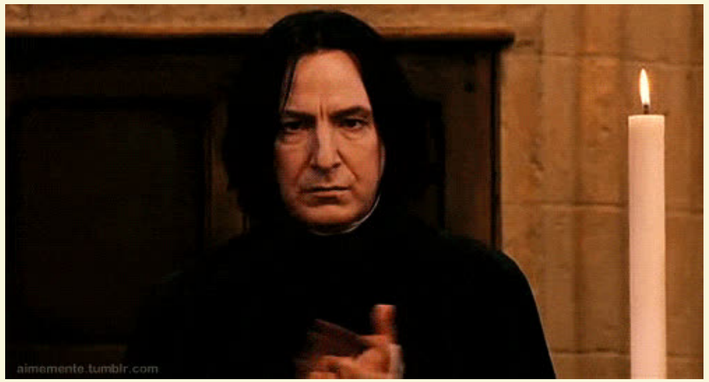
Little wins





Best configuration:

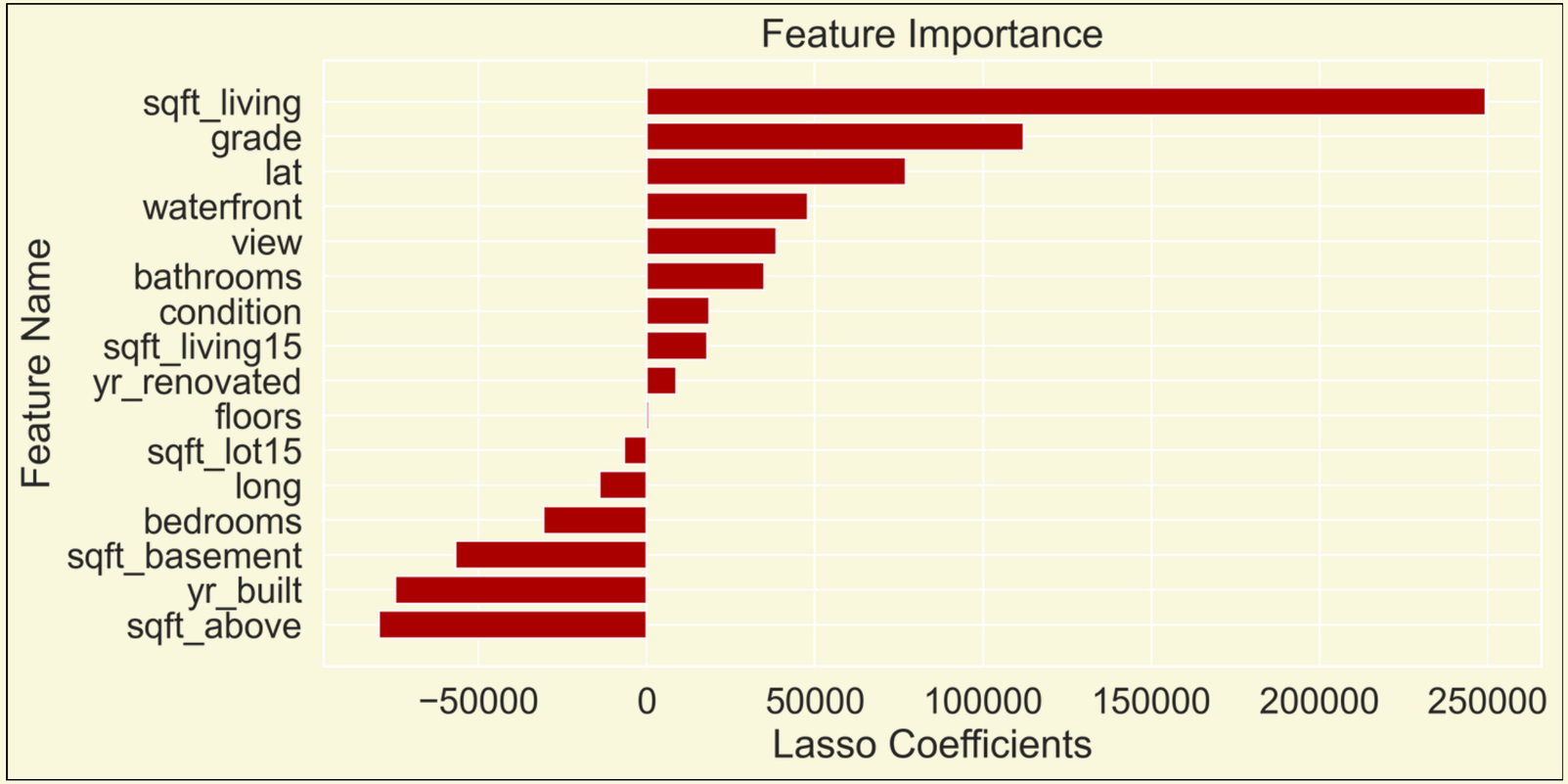
Noise N: 0.01, Noise M: 0.1, Alpha: 0.01
Best R^2 Score: 0.6957298254445554



Lasso Regression

20% Test Size

- Lasso regression is effective in reducing the number of features by setting some coefficients exactly to zero.
- A larger value of **alpha** leads to stronger regularization. Prevents overfitting.
- In this case higher alpha lead to lower R^2



R^2 train = 0.69
R^2 test = 0.69

XGBoost Regression

10% Test Size

- **Gradient Boosting**: a method that builds models sequentially, with each model trying to correct the errors of the previous one.
- **XGBoost**: incorporates both **speed** and **performance** improvements.
- By tuning the **hyperparameters**, you can optimize its performance and prevent overfitting.
  - max_depth
  - learning_rate
  - n_estimators
  - subsample
  - colsample_bytree

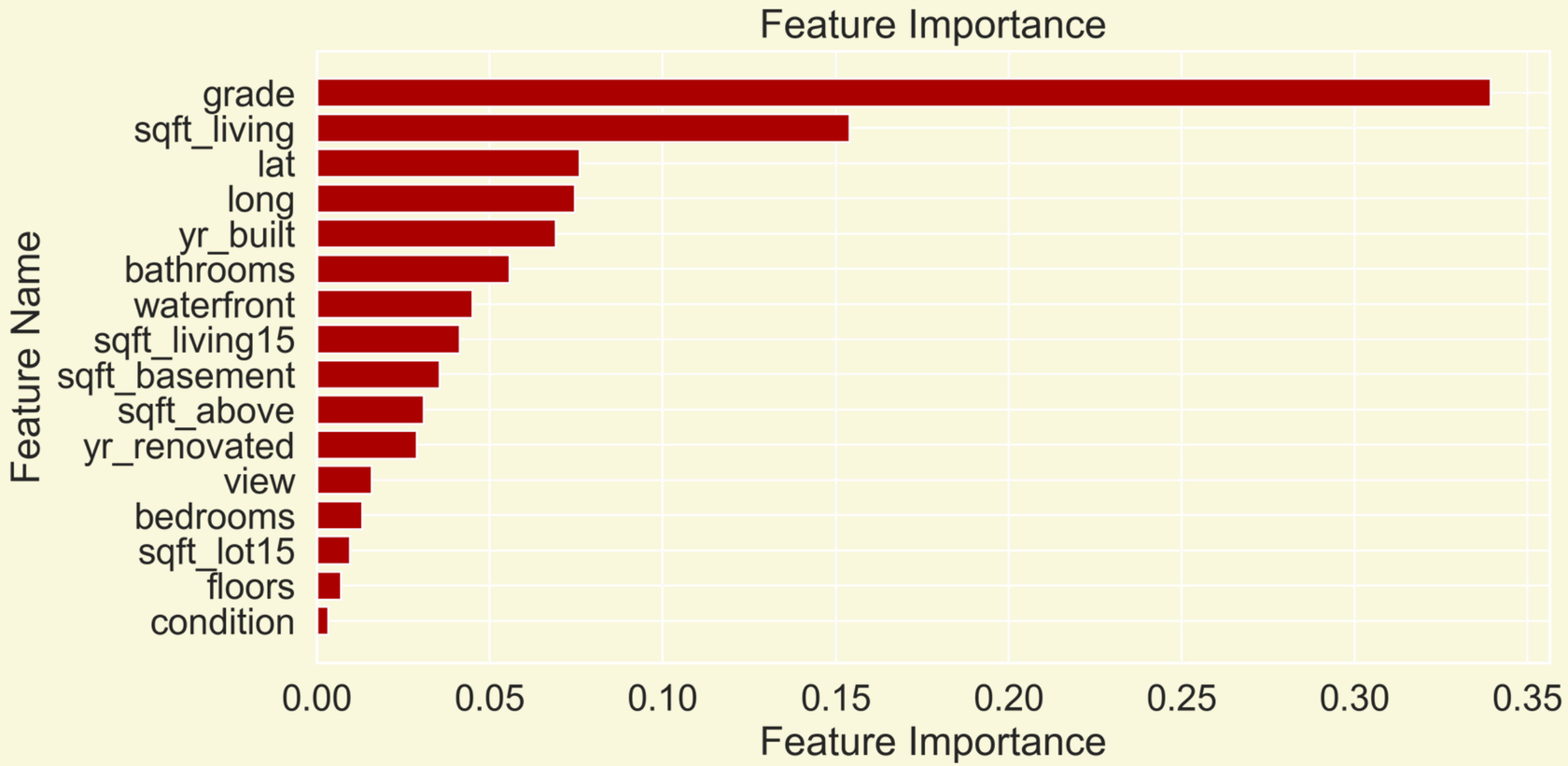
- Cross-validation is a technique used to asses how well a model generalizes to unseen data:

cv=3 ⟶ R^2=0.7964
cv=5 ⟶ R^2=0.8011
cv=10 ⟶ R^2=0.8056

XGBoost Regression

10% Test Size

Metrics	Values
R2	0.806
RMSE	168286.8527
MAE	99342.2758



R^2 train = 0.846
R^2 test = 0.806

Decision Tree

- Flowchart-like model that splits data into branches based on feature values.
- At each split, it selects the feature that best separates the data, continuing until reaching a stopping criterion.
- It makes predictions by following the branches down to a leaf node, which contains the predicted value.
- Sensitive to:
  - **Overfitting**, especially with deep trees.
  - Small **changes in data** (can lead to different splits).

Fit the data

Using features with corr. > 0.2

🔹 Test size 40%

Model Metrics: | **R^2 = 0.6416** | RMSE = 231455.5397 | MAE = 125950.7574

⚠ The model is moderately good, but there's room for improvement.

Cross-Validation: Average Training **R^2 0.9993** | Average Test **R^2 0.6408**

Decision Tree

- Flowchart-like model that splits data into branches based on feature values.
- At each split, it selects the feature that best separates the data, continuing until reaching a stopping criterion.
- It makes predictions by following the branches down to a leaf node, which contains the predicted value.
- Sensitive to:
  - **Overfitting**, especially with deep trees.
  - Small **changes in data** (can lead to different splits).

Fit the data

Using features with corr. > 0.2

🔹 Test size 40%

Model Metrics: | **R^2 = 0.6416** | RMSE = 231455.5397 | MAE = 125950.7574

⚠ The model is moderately good, but there's room for improvement.

Cross-Validation: Average Training **R^2 0.9993** | Average Test **R^2 0.6408**

Hyperparameter Tuning

Model Metrics: | **R^2 = 0.758** | RMSE = 190199.5037 | MAE =103219.8816

✅ The model performs well! It explains a large portion of the variance.

Cross-Validation: Average Training **R^2 0.8442** | Average Test **R^2 0.7507**

max_depth	-> None	-> 10
min_sample_leaf	-> 1	-> 4
min_sample_split	-> 2	-> 10

Decision Tree

- Flowchart-like model that splits data into branches based on feature values.
- At each split, it selects the feature that best separates the data, continuing until reaching a stopping criterion.
- It makes predictions by following the branches down to a leaf node, which contains the predicted value.
- Sensitive to:
  - **Overfitting**, especially with deep trees.
  - Small **changes in data** (can lead to different splits).

Fit the data

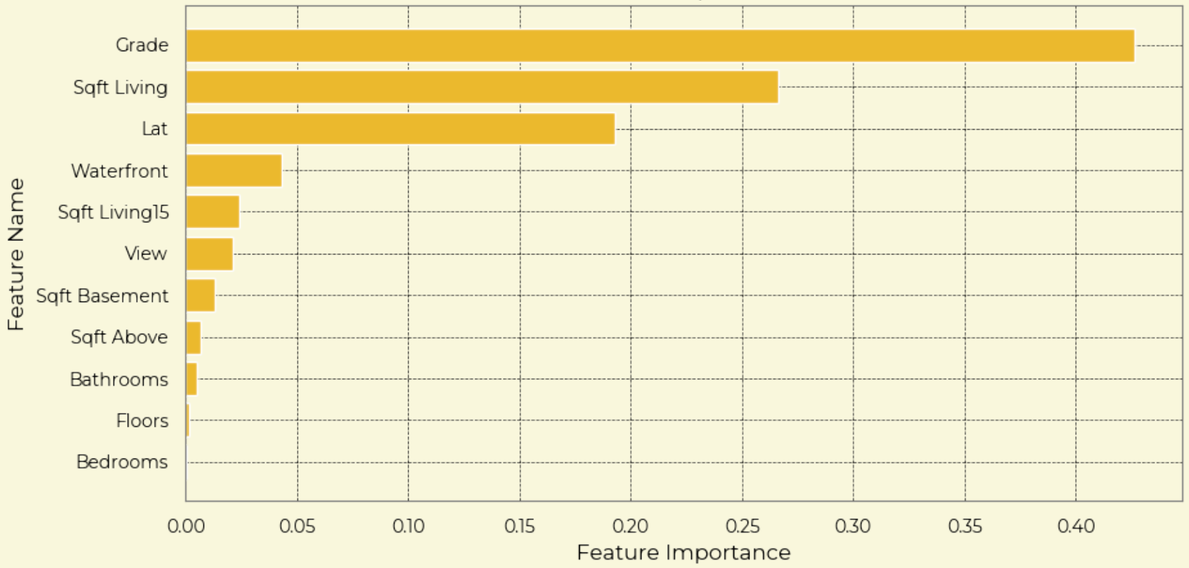
Using features with corr. > 0.2

🔹 Test size 40%

Model Metrics: | **R^2 = 0.6416** | RMSE = 231455.5397 | MAE = 125950.7574

⚠ The model is moderately good, but there's room for improvement.

Cross-Validation: Average Training **R^2 0.9993** | Average Test **R^2 0.6408**



Hyperparameter Tuning

Model Metrics: | **R^2 = 0.758** | RMSE = 190199.5037 | MAE =103219.8816

✅ The model performs well! It explains a large portion of the variance.

Cross-Validation: Average Training **R^2 0.8442** | Average Test **R^2 0.7507**

max_depth	-> None	-> 10
min_sample_leaf	-> 1	-> 4
min_sample_split	-> 2	-> 10

Feature Selection

🔹 Recursive Feature Selection (top 5)

Model Metrics: | **R^2 =0.7643** | RMSE = 187710.7007 | MAE = 103895.9869

✅ The model performs well! It explains a large portion of the variance.

Cross-Validation: Average Training **R^2 0.8347** | Average Test **R^2 0.7641**

Decision Tree

- Flowchart-like model that splits data into branches based on feature values.
- At each split, it selects the feature that best separates the data, continuing until reaching a stopping criterion.
- It makes predictions by following the branches down to a leaf node, which contains the predicted value.
- Sensitive to:
  - **Overfitting**, especially with deep trees.
  - Small **changes in data** (can lead to different splits).

Fit the data

Using features with corr. > 0.2

🔹 Test size 40%

Model Metrics: | **R^2 = 0.6416** | RMSE = 231455.5397 | MAE = 125950.7574

⚠ The model is moderately good, but there's room for improvement.

Cross-Validation: Average Training **R^2 0.9993** | Average Test **R^2 0.6408**

Hyperparameter Tuning

Model Metrics: | **R^2 = 0.758** | RMSE = 190199.5037 | MAE =103219.8816

✅ The model performs well! It explains a large portion of the variance.

Cross-Validation: Average Training **R^2 0.8442** | Average Test **R^2 0.7507**

max_depth	-> None	-> 10
min_sample_leaf	-> 1	-> 4
min_sample_split	-> 2	-> 10

Feature Selection

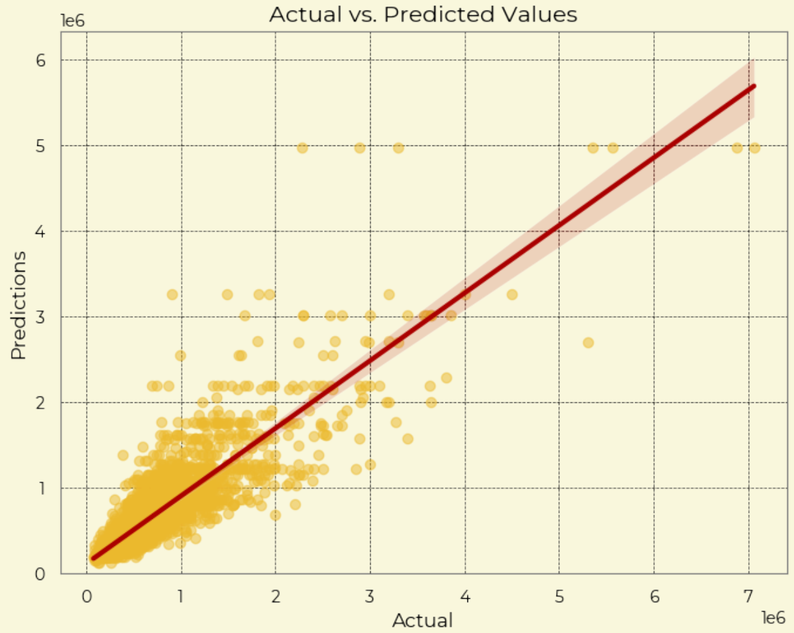
🔹 Recursive Feature Selection (top 5)

Model Metrics: | **R^2 =0.7643** | RMSE = 187710.7007 | MAE = 103895.9869

✅ The model performs well! It explains a large portion of the variance.

Cross-Validation: Average Training **R^2 0.8347** | Average Test **R^2 0.7641**

Model Evaluation



Random Forest

- Ensemble learning method that builds multiple Decision Trees and averages their predictions reducing overfitting
- Each tree is trained on a random subset of the data and considers a random subset of features at each split.
- Sensitive to:
  - Choice of **hyperparameters.**
  - Computational cost.

Fit the data

Using features with corr. > 0.2

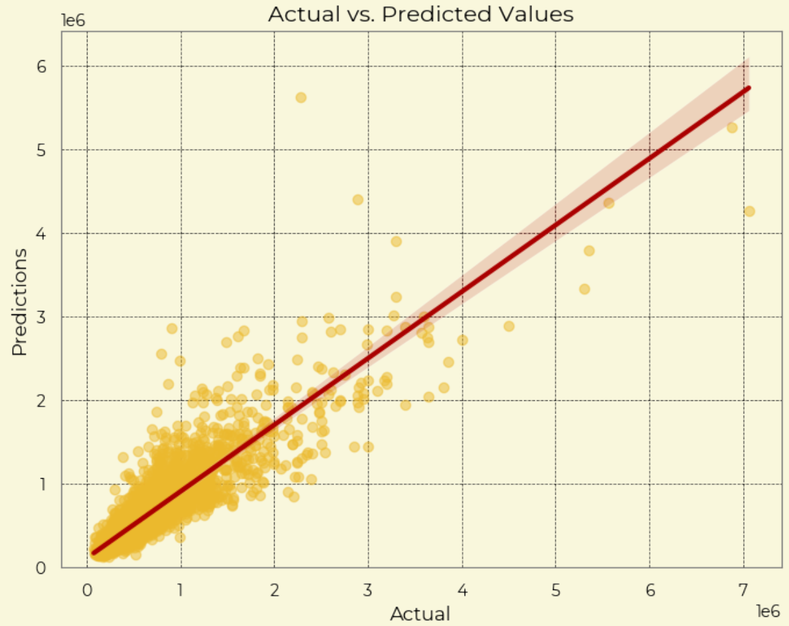
🔹 Test size 40%, n_estimators = 100
Model Metrics: | **R^2 = 0.7999** | RMSE = 172942.0715 | MAE = 91992.6531
✅ The model performs well! It explains a large portion of the variance.
Cross-Validation: Average Training **R^2 0.9721** | Average **Test R^2 0.81**

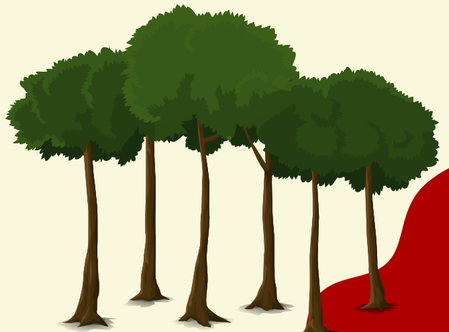
Hyperparameter Tuning

🔹 Test size 40%
Model Metrics: | **R^2 = 0.8004** | RMSE = 172748.9559 | MAE = 91815.6874
✅ The model performs well! It explains a large portion of the variance.
Cross-Validation: Average Training **R^2 0.9711** | Average Test **R^2 0.8145**

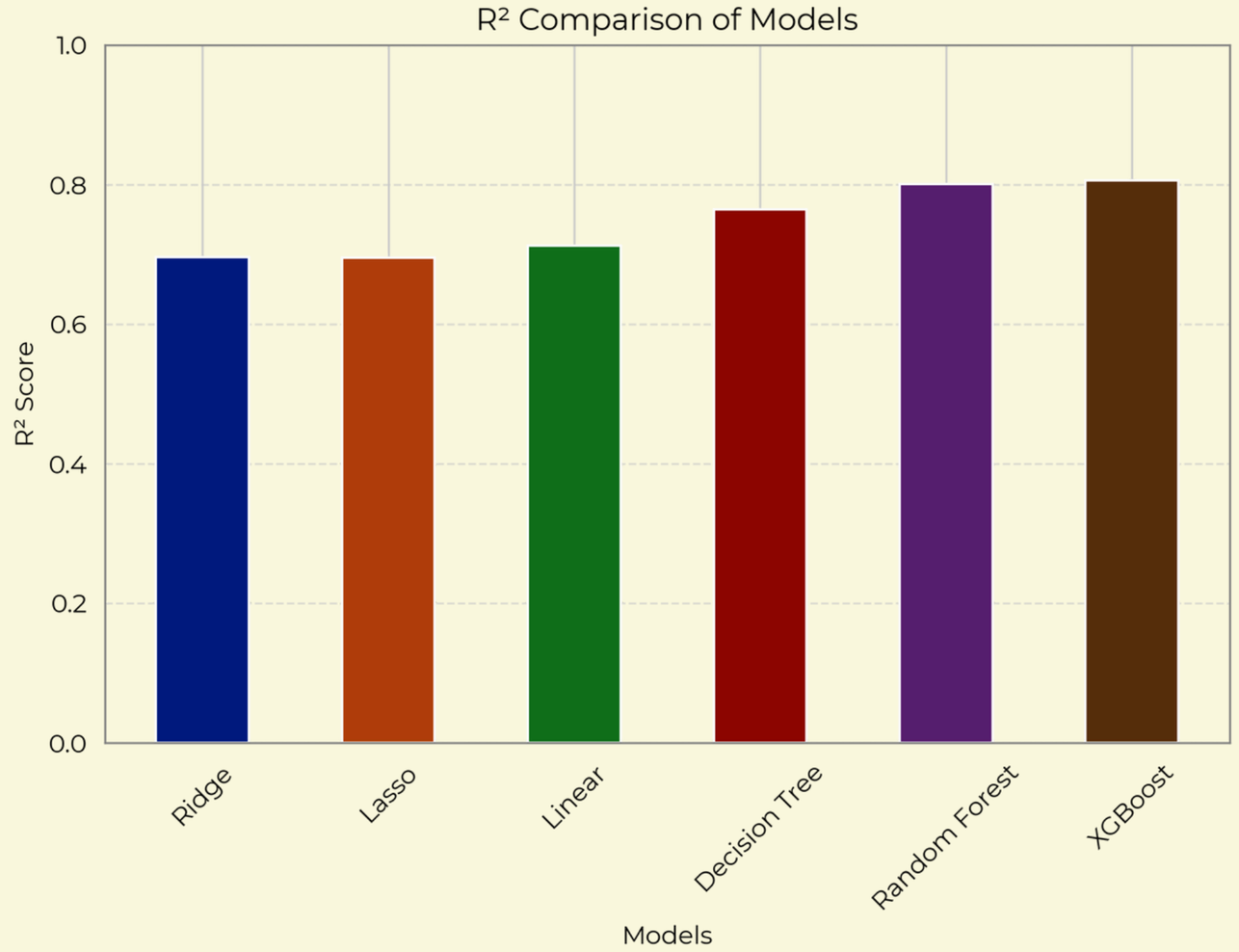
max_depth	-> None	-> 20
min_sample_leaf	-> 1	-> 1
min_sample_split	-> 2	-> 2
n_estimators	-> 100	-> 200

Model Evaluation





Battle of the Models. Which Predicts Best?



Hogwarts School of
Data Science Magic

Thank You

"IT IS OUR CHOICES THAT SHOW WHAT WE TRULY ARE, FAR MORE THAN OUR ABILITIES."

Albus Dumbledore

Presented House of Gryffindor