Isaia Pacheco
Isaac Pacheco
December 3rd, 2025

Apartment Rental Predictions

1. Housing Crisis

- California home prices and rents are among the highest in the country.
- Affordability has declined for many households.
- Predicting housing prices helps understand market trends and support better decision-making.

2. Problem Statement

- We set a goal to predict apartment rental listing prices in the U.S. to understand how property features and location influence market value and to build a model that can assist renters, property managers/owners in making data-driven decisions.

3. UC Irvine Machine Learning Repository API

- This dataset contains information on apartments and homes for rent across the USA.
- It contains 10,000 rows and 22 features and there is a mixture of qualitative variables, nominal quantitative variables, discrete and continuous (e.g., state, and bedrooms, price respectively).
- It contains the following features: id , category, title, body, amenities, bathrooms, bedrooms, currency, fee, has_photo, pets_allowed, price, price display, price type, square_feet, address, city_name, state, latitude, longitude, source, and time.
- Each record is a listing in the USA.

4. Data Quality

4.1 Accuracy and Validity

- The data reflected real world values. We checked each unique source to verify the accuracy and validity of the dataset.
- We found that most of the sources were from various listing websites and few of the sources were discontinued.
- It is unclear whether the listing is a rental or selling but we have reason to believe that it represents both by the name of the source.
- Time was an ambiguous feature as there was no documentation on how the time is formatted and can be interpreted.

- And because of this, we could not combine economic data on housing based on time of listing which is what we originally planned to do.

4.2 Completeness

- Roughly 99% of the data contained records with complete fields.
- UCI Irvine indicate that the column price and square_feet is never empty, which was true, however other features had mixed values (data types)

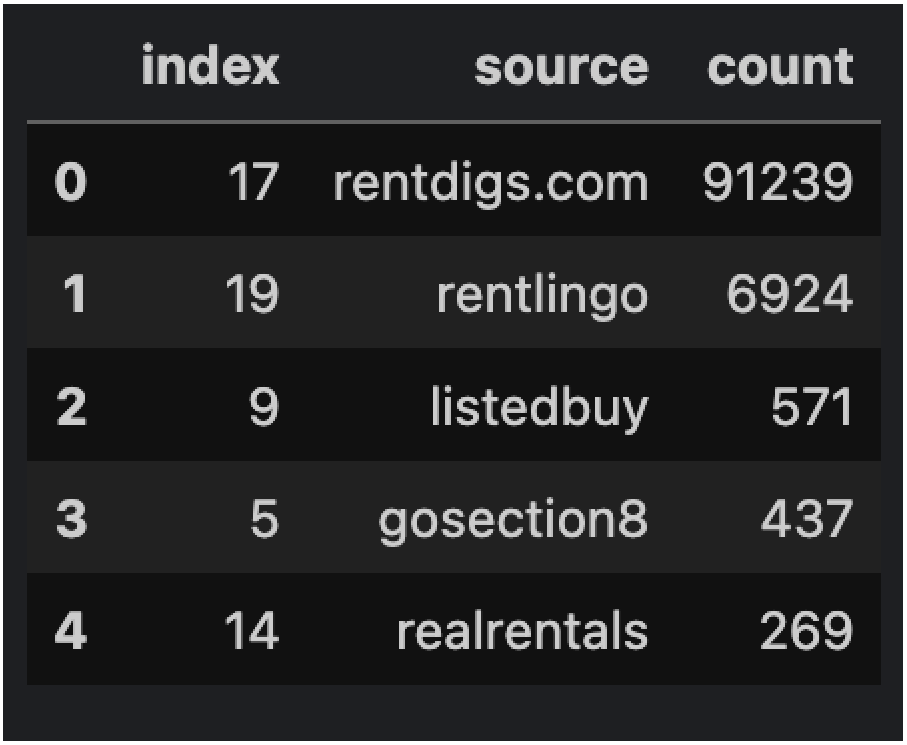
4.3 Representative

- The dataset contained records from all 50 states.
- When grouped by state, we discovered there exists class imbalance in the data.

4.4 Bias

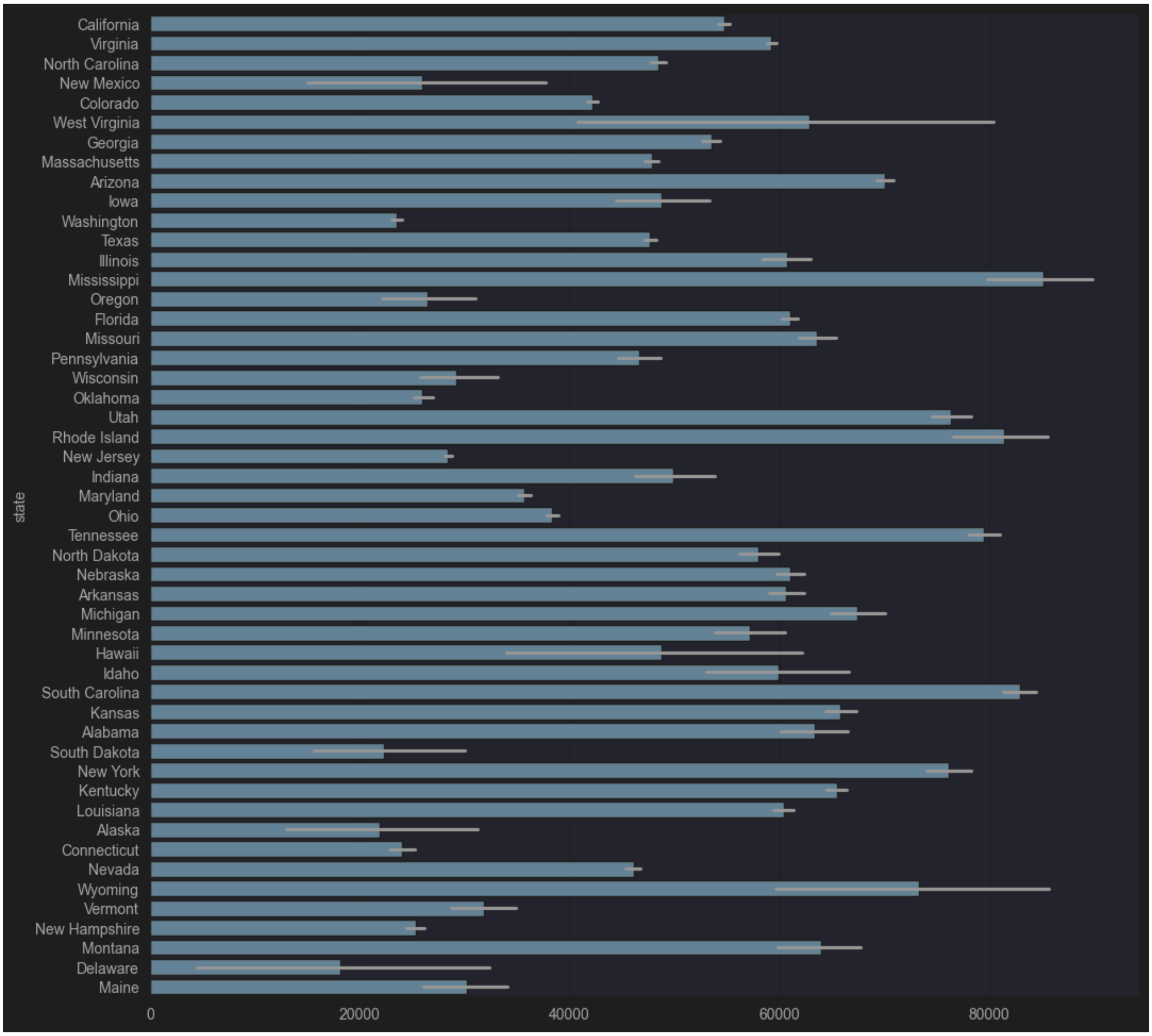
- Based on the data sources for each apartment listing, we have reason to believe that data doesn't capture the entire market for apartment listings. There are other sources like redfin that capture other apartment listings.
- Imbalance by state

Figure A: Top 5 sources with the most count of records



	index	source	count
0	17	rentdigs.com	91239
1	19	rentlingo	6924
2	9	listedbuy	571
3	5	gosection8	437
4	14	realrentals	269

Figure B: Bar Plot of Count of Records by State



5. Data Cleaning

- At a high level, all records that were not complete, were dropped.
- Id
  - Casted string to ints, Converted uninterpretable as NaN
- Category
  - Fixed typos and inferred ‘ousing’ to housing category, Converted uninterpretable categories as NaN
- Title
  - Casted to string, Converted uninterpretable as NaN
- Body
  - Casted to string, Converted uninterpretable as NaN

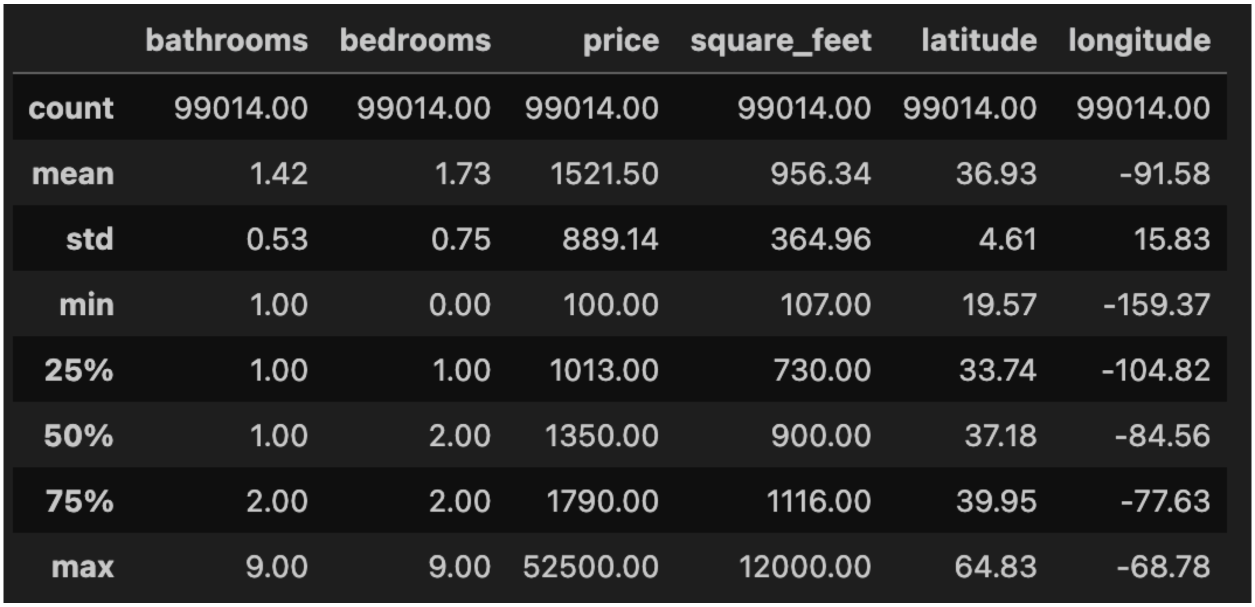
- Amenities
    - Split up the string by / and then flattened to get all the amenities as a list
- Bathrooms
    - Casted to floats the ints, Converted uninterpretable as NaN
- Bedrooms
    - Casted to floats the ints, Converted uninterpretable as NaN
- Currency
    - Found only USD, Converted uninterpretable as NaN
- Fee
    - Mapped yes and no to true/false, Converted uninterpretable as NaN
- Has Photo
    - Mapped yes and thumbnail and no to true/false, Converted uninterpretable as NaN
- Pets Allowed
    - Mapped entries to cats&dogs, cats, dog, x indicating Not allowed, Converted uninterpretable as NaN
- Price
    - Casted to float , Converted uninterpretable as NaN
- Price Display
    - Extracted the price, if there was a range took average of the prices, Converted uninterpretable as NaN
- Price Type:
    - Mapped to monthly and weekly, Converted uninterpretable as NaN
- Square_feet
    - Casted to float, Converted uninterpretable as NaN
- Address
    - In progress
- City Name:
    - Mapped to US Cities only , and Fixed abbreviations, Converted uninterpretable as NaN
- State:
    - Mapped to US States only , and Fixed abbreviations, Converted uninterpretable as NaN
- Latitude:
    - latitude is an angle that ranges from −90° at the south pole to 90° at the north pole, with 0° at the Equator. Converted uninterpretable as NaN
- Longitude
    - latitude is an angle that ranges from −180° to 180, with 0° at the Equator. Converted uninterpretable as NaN
- Source
    - Casted to string, Converted uninterpretable as NaN
- Time
    - Casted to string, Converted uninterpretable as NaN

6. Exploratory Data Analysis (EDA)

6.1 Initial Observations

- Price - max is 35x the mean, likely luxury or anomalous listings and long tailed
- Bathrooms & Bedrooms - (IQR 1–2), suggests the dataset overwhelmingly, consists of 1–2 bedroom units, likely rentals or small homes. Low variance
- Bathrooms & Bedrooms - Max 9, possibly mansions, multiunits or mislabeled
- Sqft - Mean 956, Med(50%) 900, Max 12k is far beyond normal residential sizes, long tailed, high variance
- Latitude range**:** [19.57 , 64.83], Longitude range**:** [–159, –68], spans HI, AK, US,

Figure C: Statistical Summary of Data

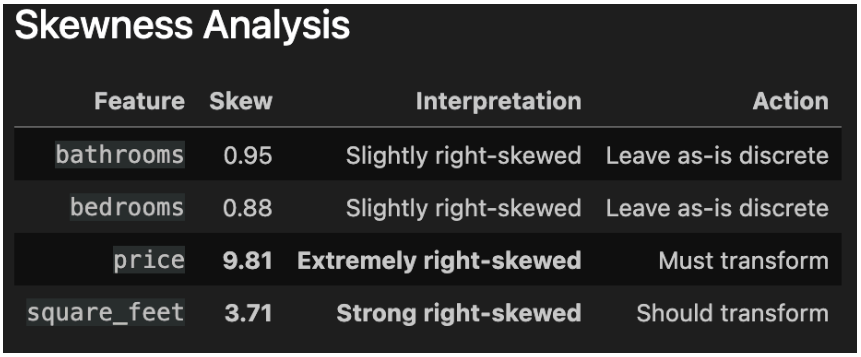


	bathrooms	bedrooms	price	square_feet	latitude	longitude
count	99014.00	99014.00	99014.00	99014.00	99014.00	99014.00
mean	1.42	1.73	1521.50	956.34	36.93	-91.58
std	0.53	0.75	889.14	364.96	4.61	15.83
min	1.00	0.00	100.00	107.00	19.57	-159.37
25%	1.00	1.00	1013.00	730.00	33.74	-104.82
50%	1.00	2.00	1350.00	900.00	37.18	-84.56
75%	2.00	2.00	1790.00	1116.00	39.95	-77.63
max	9.00	9.00	52500.00	12000.00	64.83	-68.78

6.2 Data Transformation: Skewness

- Price and square feet were right skewed
    - Basis for logarithmic transformations
- Expensive and massive homes were distorting scales.

Figure D: Skewness of Data



Skewness Analysis

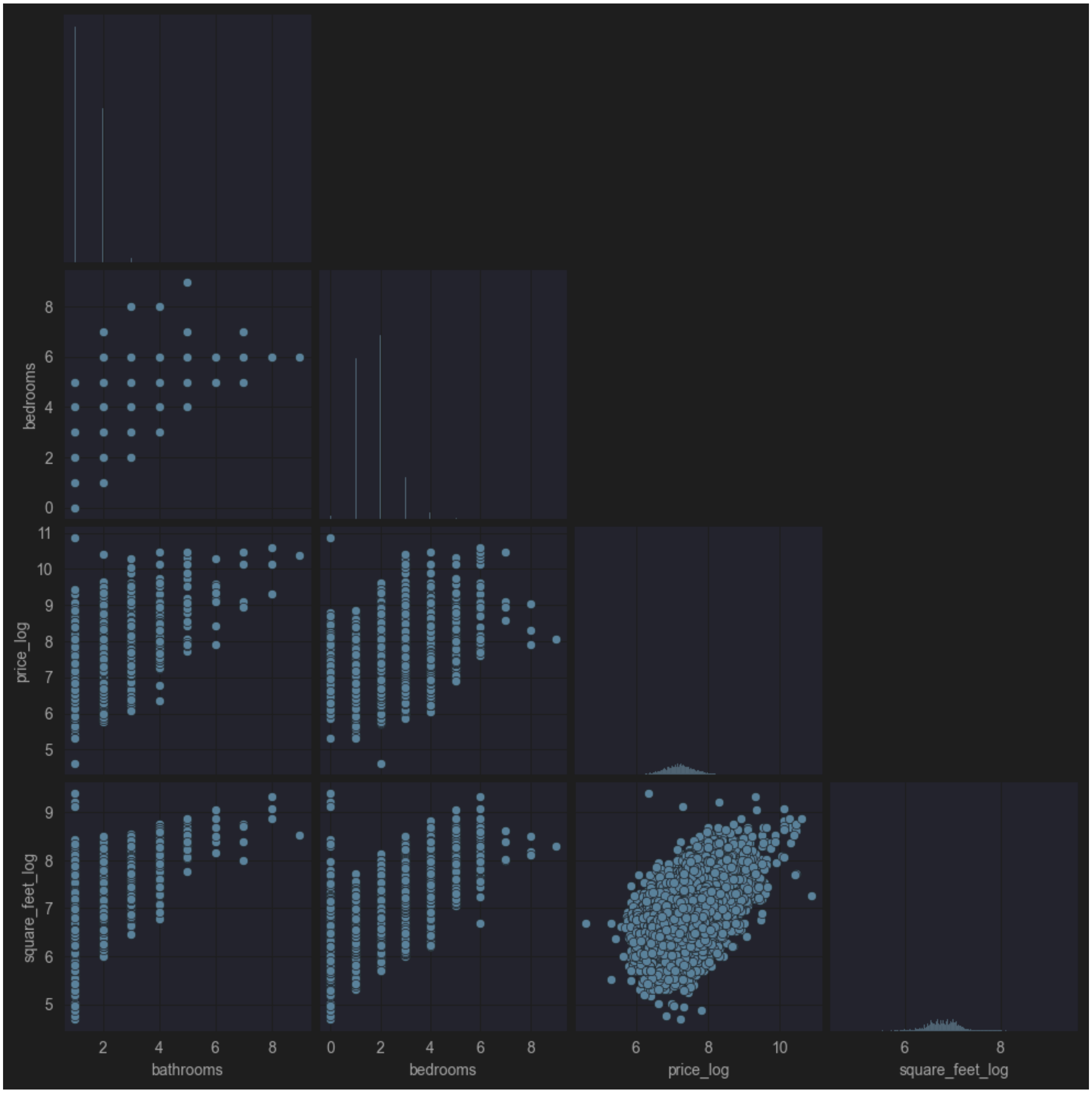
Feature	Skew	Interpretation	Action
bathrooms	0.95	Slightly right-skewed	Leave as-is discrete
bedrooms	0.88	Slightly right-skewed	Leave as-is discrete
price	**9.81**	**Extremely right-skewed**	Must transform
square_feet	**3.71**	**Strong right-skewed**	Should transform

6.3 Linear Relationships

6.3.1 Scatter Plots

- price_log generally increases as square_feet_log increases, but the pattern is not strongly linear; though not strong enough to draw firm conclusions from the scatterplots alone.
- The relationships between price_log and the discrete features (bedrooms and bathrooms) appear to have an upward trend.
- However, because these are discrete and tightly grouped, the scatterplots alone do not provide strong evidence for or against nonlinear transformations.

Figure E: Scatter Plot

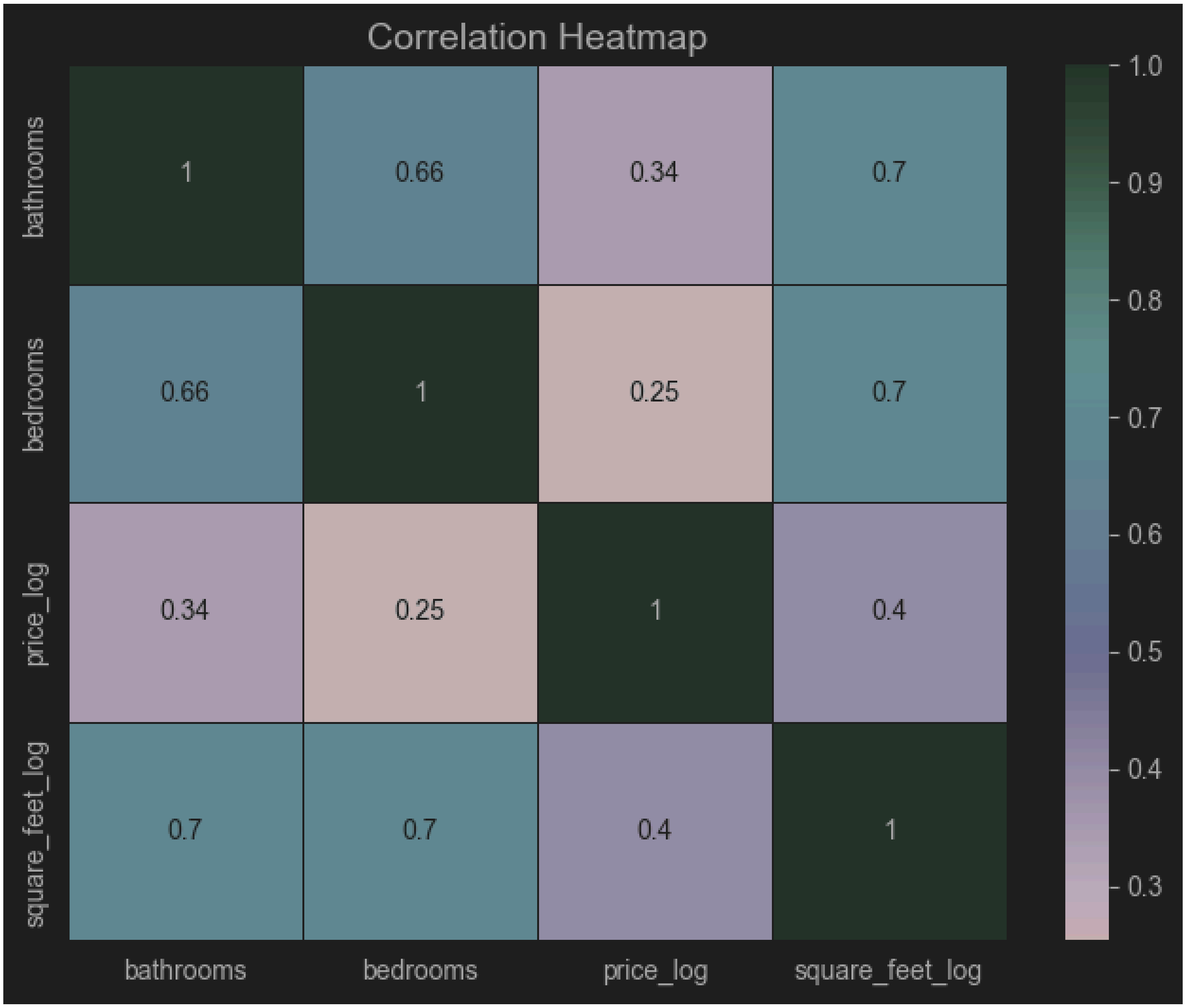


6.3.2 Correlation Matrix of All Apartment Listings

- square_feet_log showed moderate linear association with price_log (0.40) and is also highly correlated with bedrooms and bathrooms (0.70 each).
- Since square_feet_log is the only continuous predictor and that is not a near linear, including quadratic term might be reasonable.
- Bedrooms, bathrooms, and square footage are all moderately correlated (0.66–0.70), which shows they are associated in a consistent way.

- However, these correlations are not high enough to cause multicollinearity problems, so it is reasonable to include all three features in the model at the same time.
- Bedrooms (0.25) and bathrooms (0.34) have relatively weak linear correlations with price_log, and the scatterplots do not show clear nonlinear patterns. Because there is no strong evidence of curvature or a more complex relationship, the simplest reasonable choice is to model these features with linear terms.

Figure F: Heatmap of Correlation of all data

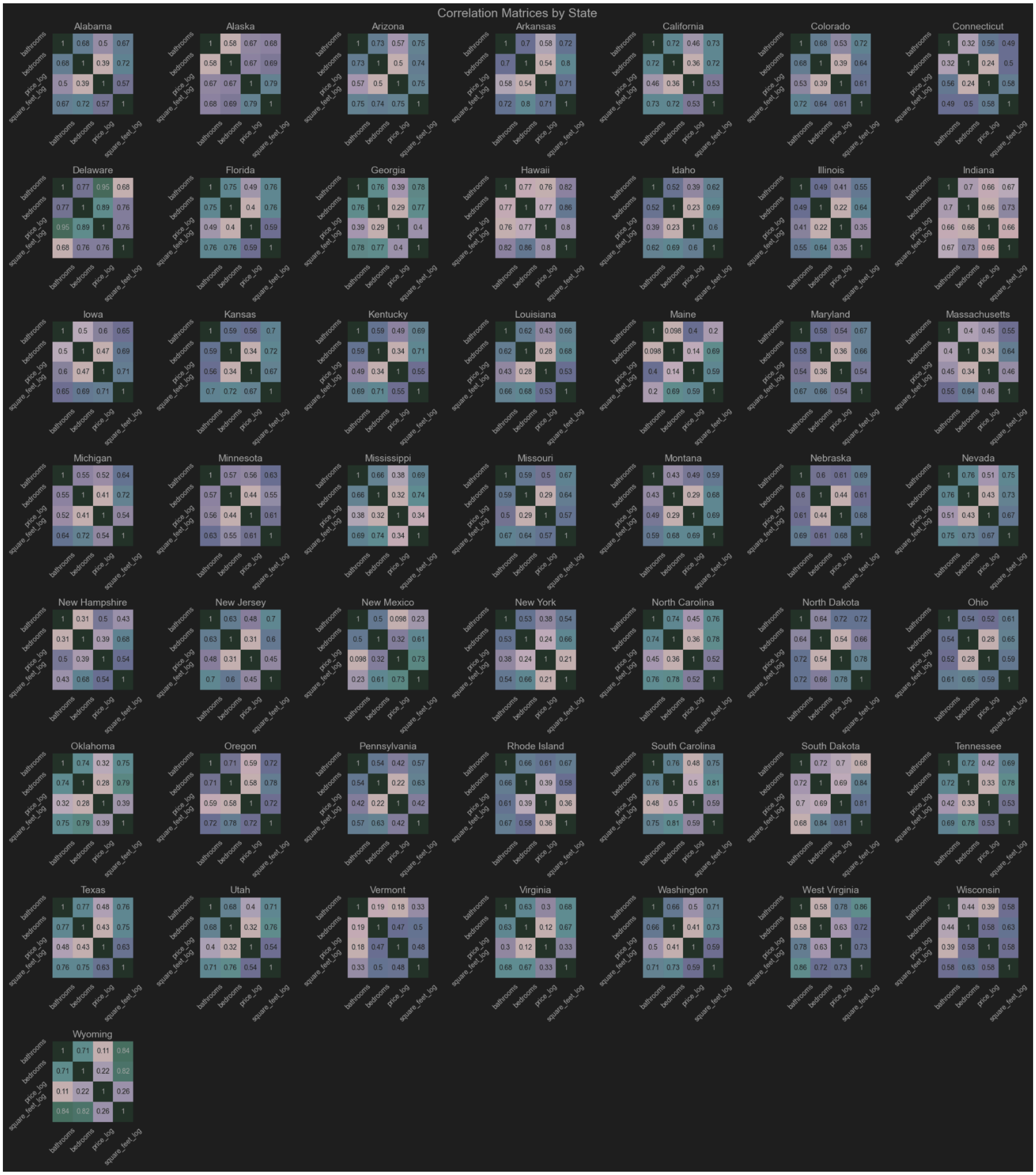


6.3.3 Correlation Matrices by State

- Grouping by state and visualizing correlation matrices of each state showed various strengths in correlation depending on the state.
- Strong correlation
    - Alaska, Hawaii
- Moderate correlation

    - California, Minnesota, Texas
- Weak correlation
    - Delaware, Indiana, Virgina

Figure G: Correlation Matrices by State



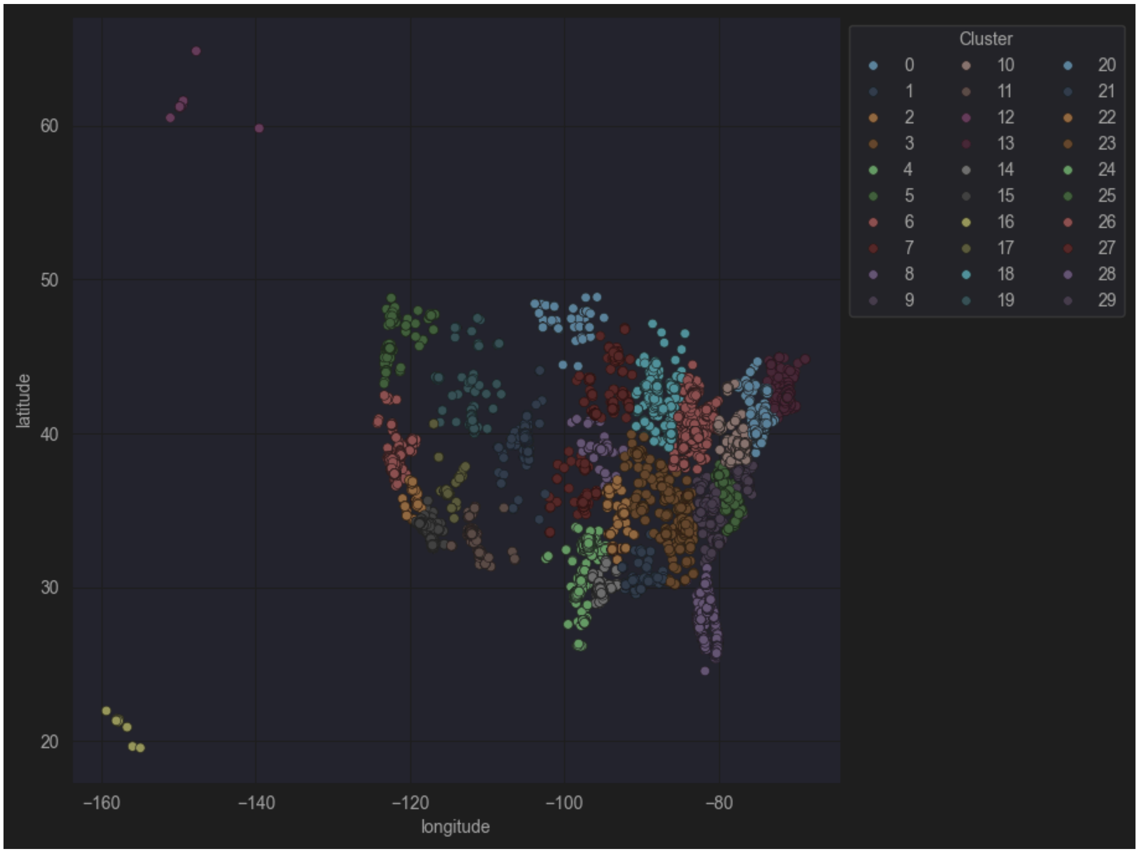
6.4 Key Insights

6.4.1 Structural Features

- Square footage, bedrooms, and bathrooms aren't enough: These structural features might explain some variation in price, but they cannot capture location-driven effects, for example, homes with identical size and room counts can differ in price depending on neighborhood, or city, or state.

6.4.2 Spatial Features

- Macro Level
    - City/State provides geopolitical boundaries, which capture broad market differences such as labor markets, tax structures, regulations, and overall cost-of-living.
- Micro Level
    - Why clustering is needed: By clustering latitude and longitude, we created neighborhood-like groups that capture these location-specific price differences.
- The cluster labels introduce spatial structure and numerical features enabling a model to adjust baseline price levels across regions and more accurately represent real-world housing markets.



7. K-Means Clustering with Linear Regression

7.1 Model

- price_log_hat = β0 + β1 * square_feet_log + β2 * square_feet_log^2 + β3 * bedrooms + β4 * bathrooms + c(CityEffects) + c(NeighborhoodClusterEffects)

7.2 Loss Function

7.2.1 Decision context

- Our aim is to estimate log-price based on apartment characteristics and Ordinary Least Squares (OLS) is designed to capture central trends rather than extreme cases.

7.2.2 Consequences of Prediction Errors

- Since squared residuals heavily penalize large mistakes, OLS naturally emphasizes reducing large pricing errors, which aligns with the practical costs of mispricing homes which could result in financial losses of renters and owners.

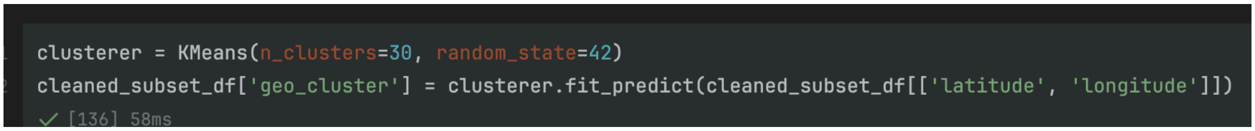
7.2.3 Performance Measurement

- OLS connects directly to standard regression metrics (RMSE, R^2) and provides interpretable coefficients, making performance assessment straightforward.

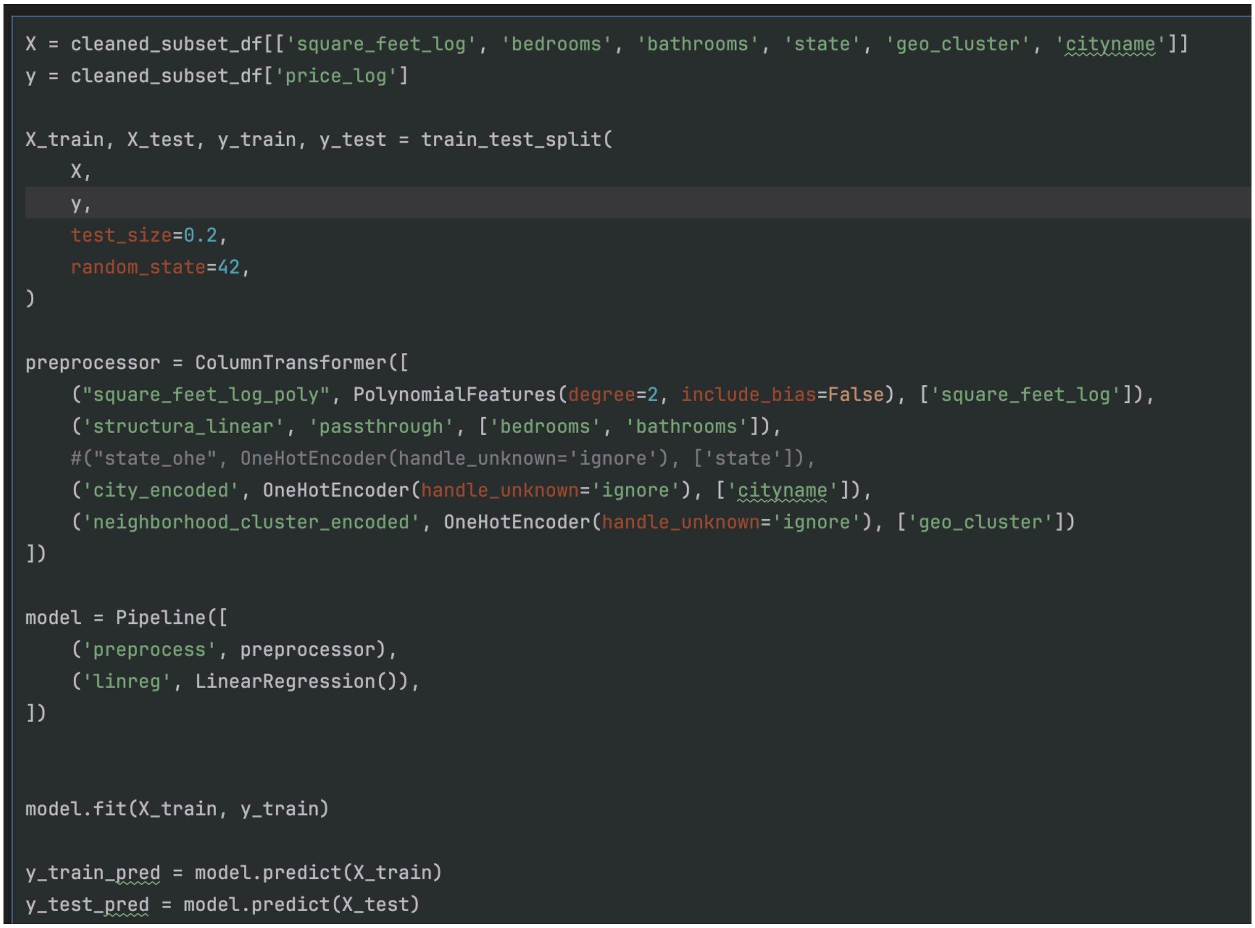
7.2.4 Heteroskedasticity & Outlier Mitigation

- We performed log transformation to price and square foot

7.3 Methodology



```
clusterer = KMeans(n_clusters=30, random_state=42)
cleaned_subset_df['geo_cluster'] = clusterer.fit_predict(cleaned_subset_df[['latitude', 'longitude']])
✓ [136] 58ms
```



```
X = cleaned_subset_df[['square_feet_log', 'bedrooms', 'bathrooms', 'state', 'geo_cluster', 'cityname']]
y = cleaned_subset_df['price_log']

X_train, X_test, y_train, y_test = train_test_split(
    X,
    y,
    test_size=0.2,
    random_state=42,
)

preprocessor = ColumnTransformer([
    ("square_feet_log_poly", PolynomialFeatures(degree=2, include_bias=False), ['square_feet_log']),
    ('structura_linear', 'passthrough', ['bedrooms', 'bathrooms']),
    #("state_ohe", OneHotEncoder(handle_unknown='ignore'), ['state']),
    ('city_encoded', OneHotEncoder(handle_unknown='ignore'), ['cityname']),
    ('neighborhood_cluster_encoded', OneHotEncoder(handle_unknown='ignore'), ['geo_cluster'])
])

model = Pipeline([
    ('preprocess', preprocessor),
    ('linreg', LinearRegression()),
])


model.fit(X_train, y_train)

y_train_pred = model.predict(X_train)
y_test_pred = model.predict(X_test)
```

7.4 Model Assessment

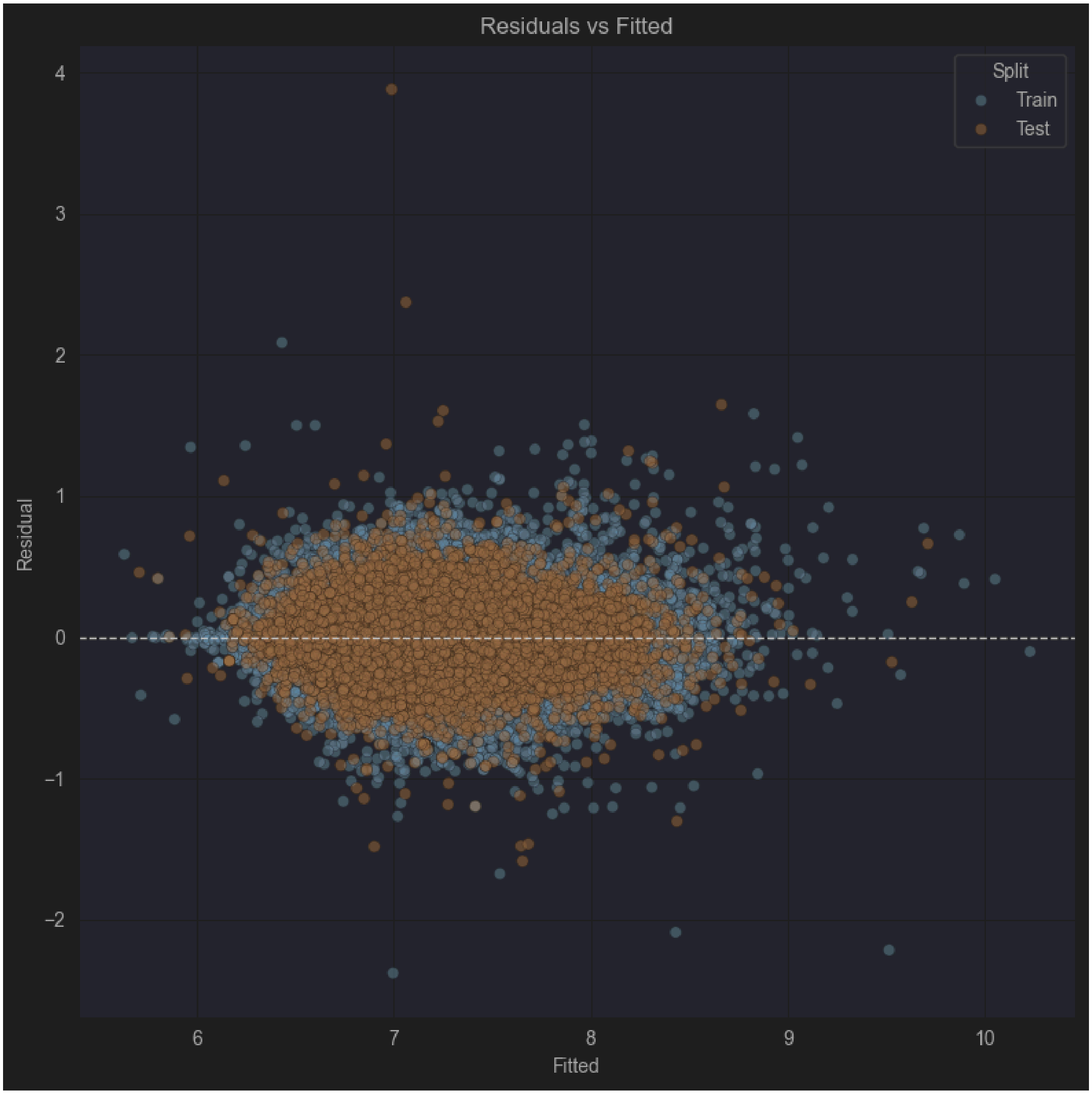
Metric	Train Value	Test Value	Interpretation
R²	0.7838	0.7581	% of variance explained. The model captures ~78% of training variation and ~76% of unseen variation. strong fit, low overfitting.
MSE	0.0411	0.0463	Average squared error. Test MSE only slightly higher. good generalization.
RMSE	0.2027	0.2151	Typical prediction error magnitude. ~20–21% deviation in normalized/log scale.
MAE	0.1488	0.1572	Average absolute error. Predictions off by ~15–16%.

MAPE	0.0206	0.0219	Mean Absolute Percentage Error. ~2% if the target is scaled, or ~2% error in/log space. Very low. Consistent accuracy.
Median Abs Error	0.1114	0.1185	Half of the predictions are within 11–12% error

8. Real World Interpretation

Error Type	Train Factor	Test Factor	Meaning
Typical Error (RMSE)	x 1.225	x 1.240	Predictions typically off by 22.5% (train) and 24.0% (test).
Average Error (MAE)	x 1.160	x 1.170	On average, predictions are off by 16.0% (train) and 17.0% (test).

Figure I: Residuals vs Fitted Plot



- Residuals stay centered around zero, no systematic bias.
- Spread is fairly uniform, heteroskedasticity largely reduced by log transform.
- Train and test points overlap, good generalization, no obvious overfitting.
- No visible curvature model form is reasonably appropriate.
- A few large residuals appear, but not enough to distort the overall pattern.
- Slight right skewed

9. Conclusion

- We tried to predict log space price values based on home structure features and geo spatial features
- We used log space square foot of the home, number of bedrooms, number of bathrooms, the city hot encoded and 30 hot encoded clusters
- We were able to account for +75% of the variance of the test set.
- Overall, due to the fact that states were not equally represented, biasness in the dataset, and assumptions made in the model, we would be hesitant to recommend this model to end-users.

10. Q&A