Isaia Pacheco
Isaac Pacheco

Project Progress Report

Introduction

The goal is to predict prices of housing (rentals, single family homes).

Data

Data Sources

We are using two data sources at this time which are:

1) UC Irvine (UCI) Machine Learning Repository
2) Bureau of Economic Analysis (BEA)

We are considering using other data sources, if needed. All datasets are obtained by making API calls which return a dataframe.

We have yet to make progress on BEA datasets but plan to find a dataset to combine with the UCI dataset. The plan is to leverage housing indexes to make better predictions after evaluation of simple linear regression models.

Bureau of Economic Analysis API Key

In order to make calls to the API, you must register for an API key. This is specified in the README.md file of the repository.

UC Irvine Dataset

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). Each record is an apartment listing in the USA. 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. Time was an ambiguous feature as there is no documentation on how the time is formatted and can be interpreted. UCI claimed that the dataset was cleaned such that square_feet and price were never empty, that was true, but they were not integers as stated in the documentation. We had to extensively clean the UCI dataset. Most of the features listed below required cleaning and below specifies how each feature was cleaned.

- Id
    - Casted string to ints, Converted uninterpretable as NaN
- Category

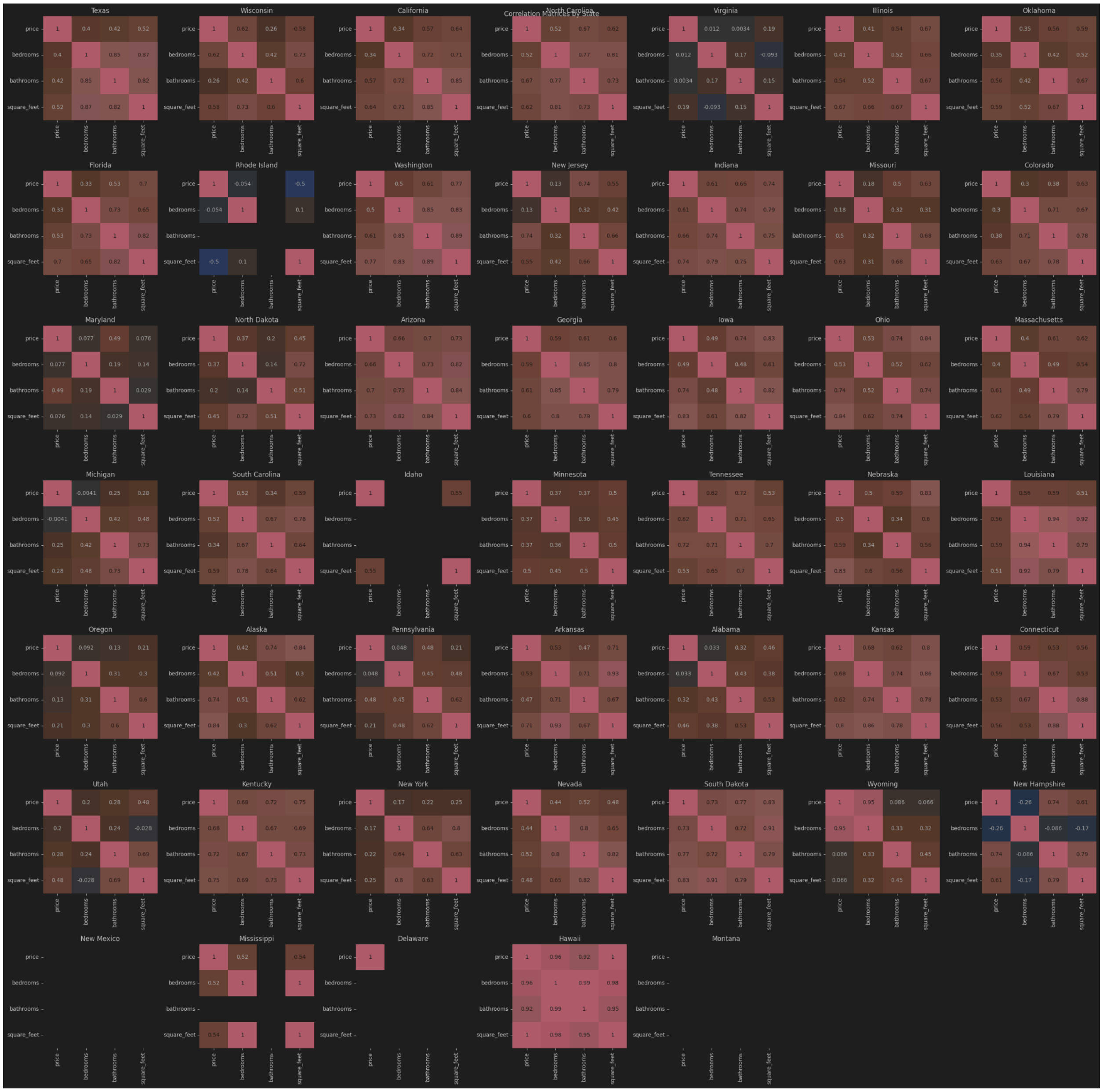
    Fixed typos and inferred 'ousing' to housing category, Converted uninterpretable categories as NaN
- Title
    - Casted to string, Converted uninterpretable as NaN
- Body
    - Casted to string, Converted uninterpretable as NaN
- Amenities
    - Split up the string by / and then flattened to get all 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

Exploratory Data Analysis (EDA)

Correlation Matrices

We began by examining the correlation of the cleaned UCI dataset.



Some states exhibit a linear relationship between price and other apartment features (e.g. number bedrooms and bathrooms, and square footage) however nationally these relationships are not consistent whereas states show a weak linear association. Therefore, this alone is not sufficient to justify developing a linear regression model. So then we took a deeper investigation into our dataset.

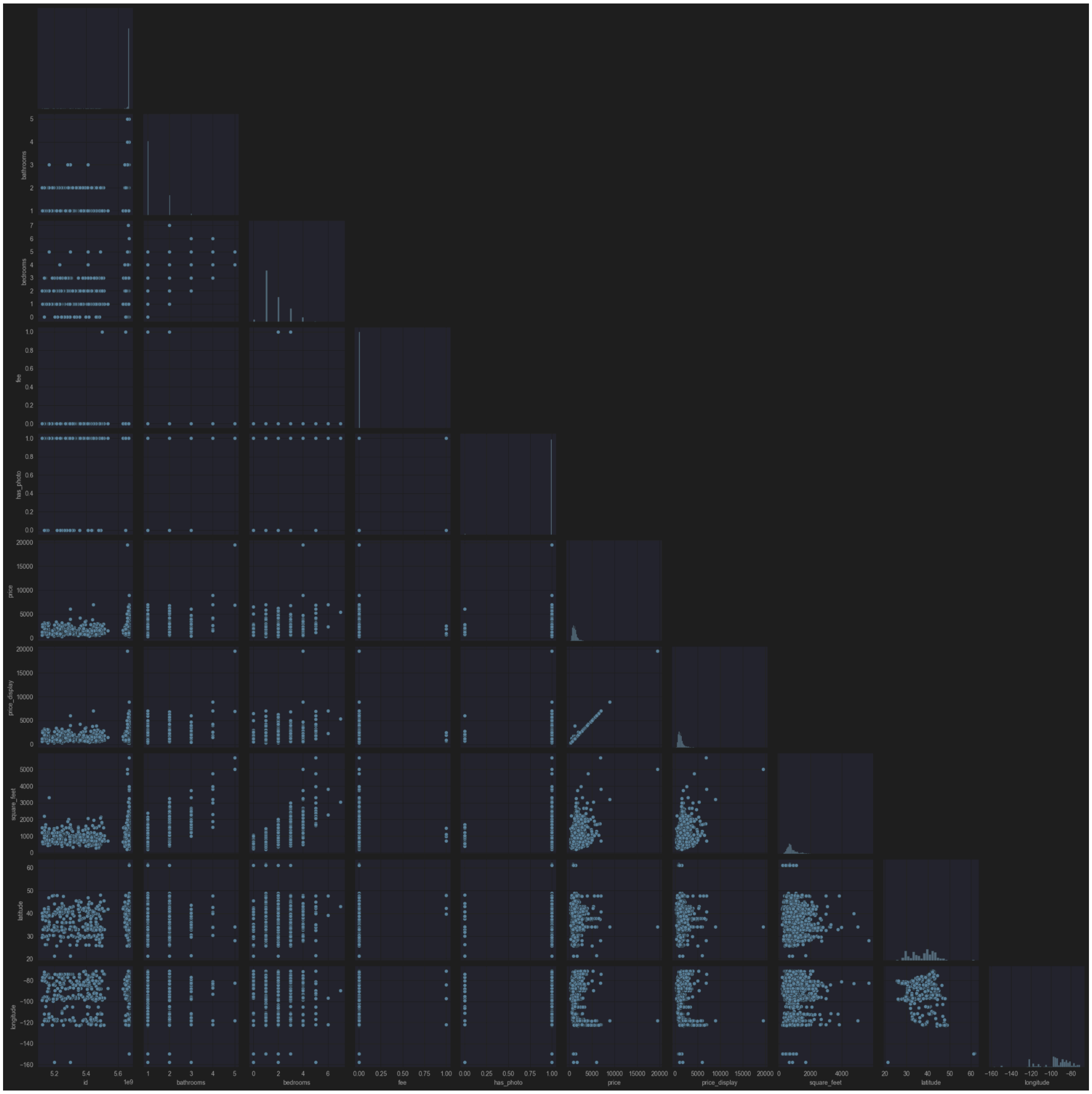
Descriptive Statistics

Most of the dataset is small-to-moderate homes, but it contained a handful of massive properties in terms of square footage.

```
        bathrooms       bedrooms        price     square_feet
count   99633.00        99633.00        99633.00          99633.00
mean    1.42      1.73    1527.40956.02
std     0.53      0.75    899.09  365.19
min     1.00      0.00    100.00  107.00
25%     1.00      1.00    1015.00730.00
50%     1.00      2.00    1350.00900.00
75%     2.00      2.00    1795.001116.00
max     9.00      9.00    52500.00        12000.00
```

After obtaining descriptive statistics we visualized the data in a different way: scatter plot matrices.

Visualization: Scatter Plot Matrices



This scatter matrix visualizes pairwise relationships among bathrooms, bedrooms, price, and square_feet. It captures both discrete-to-continuous and continuous-to-continuous relationships before any transformations. Below were our observations.

Bathrooms and Bedrooms

Strong positive association, homes with more bedrooms generally have more bathrooms.
The relationship appears step-like rather than smooth, since both are discrete integer features.
Sparse points for high values (≥6 bedrooms, ≥4 bathrooms) reflect rare, high-end homes.

Bathrooms and Square Feet

Clear positive trend, as house size increases, bathroom count rises.
The data forms vertical clusters, due to rounding in square footage (e.g., multiples of 500 or 10,00 ft²).
A few extreme points near 10,000 - 12,000 ft² might expand the scale, it also may show luxury properties.

Bedrooms and Square Feet

Also, a strong positive relationship , most homes fall within 2 - 3 bedrooms and 700–15,00 ft².
Vertical and horizontal banding shows how discrete bedroom counts intersect with rounded square footage. Large, rare homes (8 - 9 bedrooms) form isolated points in the top-right corner.

Price and Square Feet

Clear nonlinear positive pattern, price rises with square footage but with large variability.
Most observations lie below 20,00 ft², where prices are tightly clustered.
The upper tail (prices above 20,000) shows a few high-end properties that stretch the scale.

Price and Bedrooms / Bathrooms

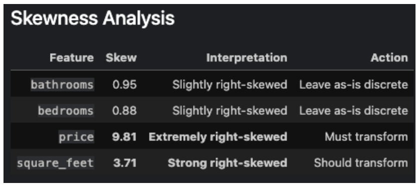
Moderate, less linear relationships compared to square footage.
Considerable overlap: e.g., 2-bedroom and 3-bedroom homes have overlapping price ranges.
Suggests square footage (continuous) is a stronger predictor of price than room counts (discrete).

Axis-Bound Observations

X-axes: discrete groupings (0 - 9) for bedrooms and bathrooms.
Y-axes: continuous scales for price and square footage.
Outliers: vertical streaks and upper-end clusters represent large, expensive homes that distort the visible scale.

At this point we decided to only look at the numerical labels.

Skewness Analysis



Skewness Analysis

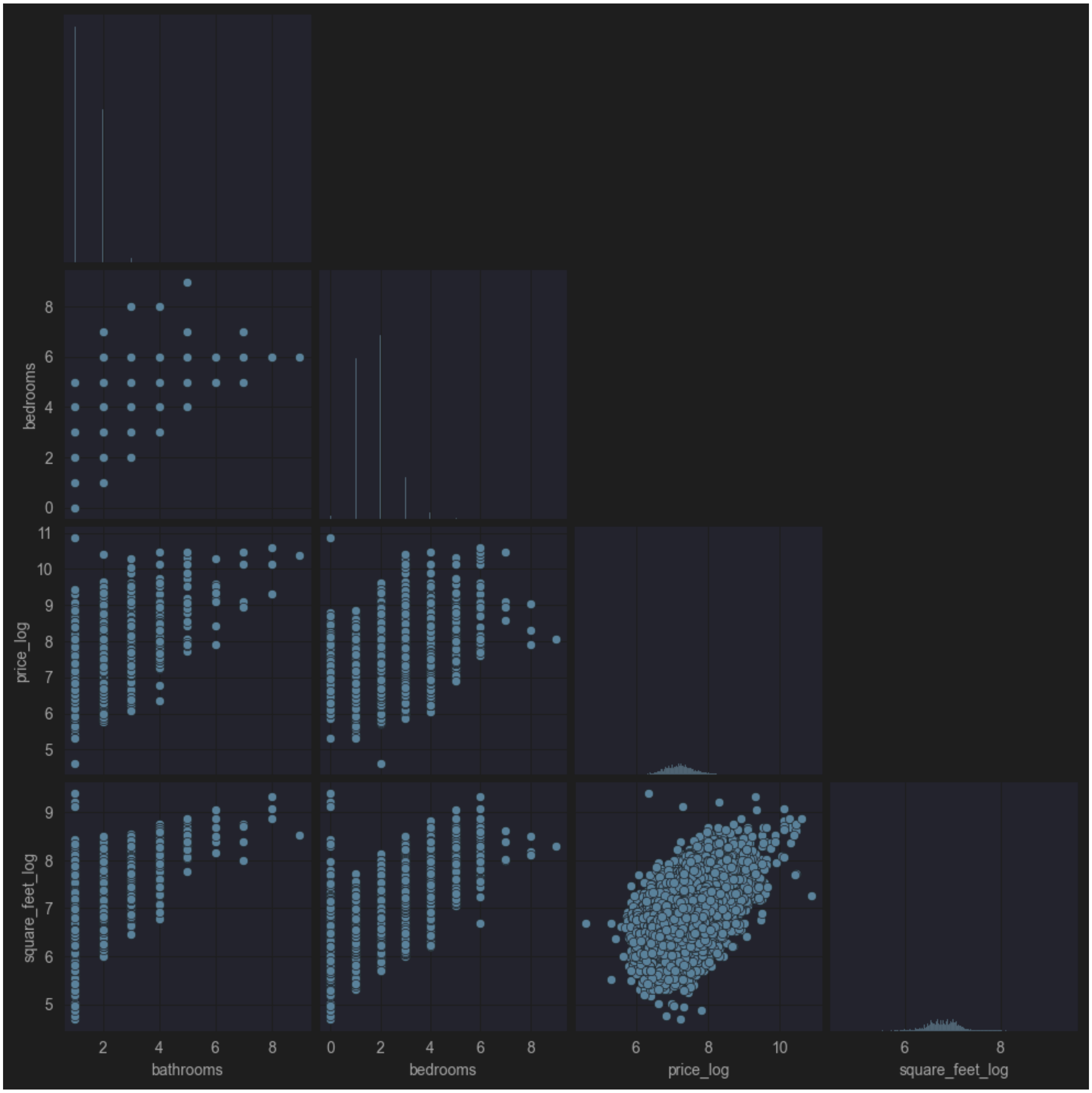
Feature	Skew	Interpretation	Action
bathrooms	0.95	Slightly right-skewed	Leave as-is discrete
bedrooms	0.88	Slightly right-skewed	Leave as-is discrete
price	**9.81**	**Extremely right-skewed**	Must transform
square_feet	**3.71**	**Strong right-skewed**	Should transform

Data ranges from 0 to positive values only, so we will use a log-type transformation to reduce skewness.

Skewness after transformations

```
bathrooms        0.95
bedrooms         0.88
price_log        0.43
square_feet_log  0.12
```

Visualization: Scatter Plots after Transformation



This scatter-matrix shows pairwise relationships among bathrooms, bedrooms, price_log, and square_feet_log. After applying log transformations to continuous features, relationships that were previously nonlinear and skewed now appear smoother and more proportional.

Bathrooms and Bedrooms

Still a strong positive ordinal relationship, homes with more bedrooms generally include more bathrooms.The step-like structure remains since both variables are discrete integers.
Higher bedroom/bathroom combinations (≥6 bedrooms, ≥4 bathrooms) remain sparse and scattered, these are rare, high-value homes.

Bathrooms and Square Feet (log)

Clear linear relationship: as home size increases, the number of bathrooms grows in roughly proportional steps.The vertical clustering seen before has reduced; the log scale compresses large square-footage differences.Outliers are far less extreme, showing improved scaling and distribution balance.

Bedrooms and Square Feet (log)

Still a tight positive association with clearer proportionality than in the raw data. The densest region is centered around 2–3 bedrooms and log(square_feet) ≈ 6.5 - 7.2, corresponding to ~700–1300 ft^2. Larger properties (≥7 bedrooms) appear as upper-end discrete clusters, consistent with rare, luxury homes.

Price (log) and Square Feet (log)

Now the strongest and most linear relationship in the matrix. Points form a clear diagonal cluster, as square_feet_log increases, price_log rises proportionally. The log transform effectively reduces the influence of extreme prices, revealing a consistent scaling trend.

Price (log) and Bedrooms / Bathrooms

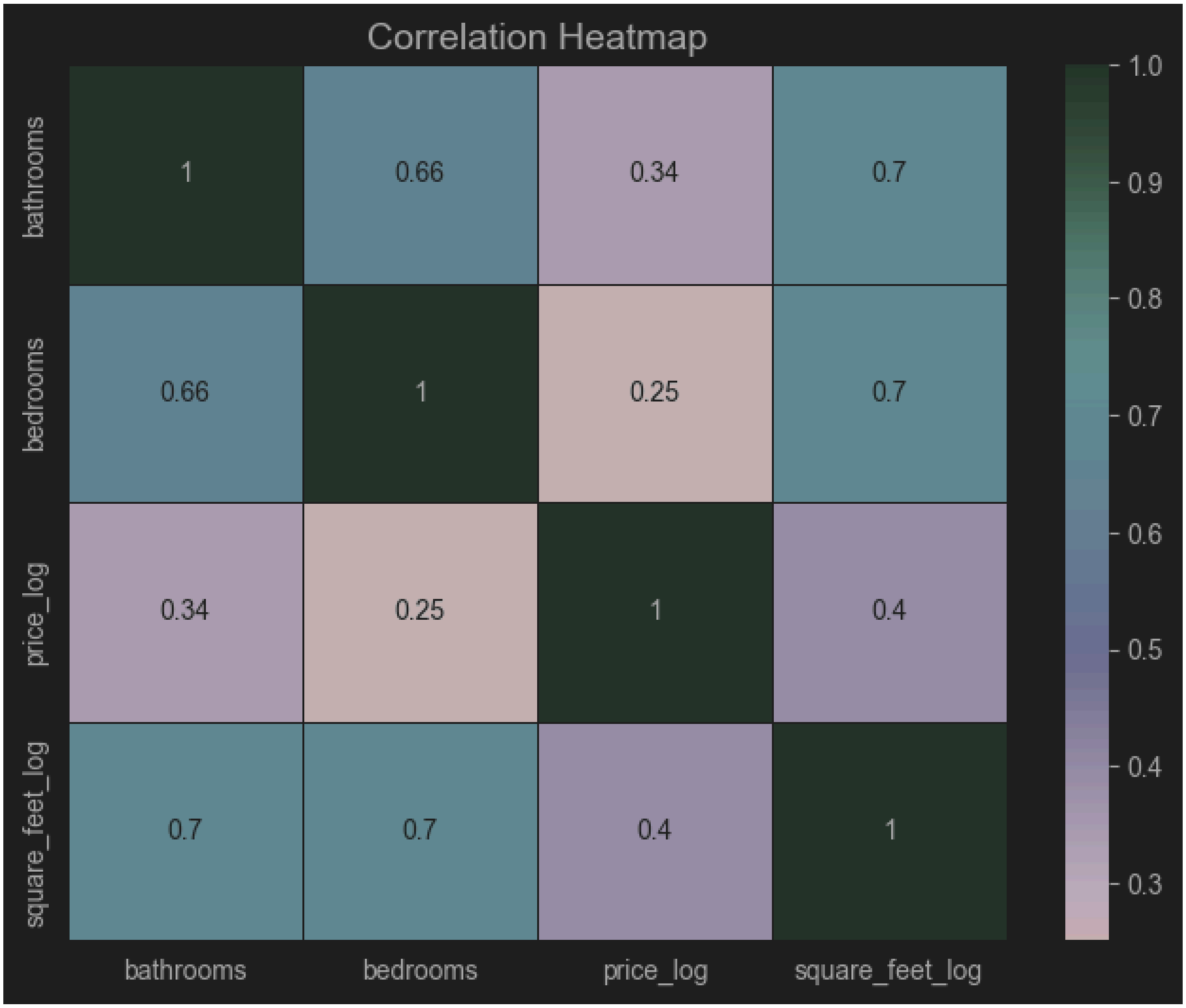
Both show positive but weaker relationships than with square footage. For small to medium homes (1–3 bedrooms), price distributions overlap heavily. Marginal price gains per additional bedroom or bathroom appear smaller at higher counts, indicating diminishing returns in larger homes.

Axis-Bound Observations

x-axes: discrete ranges (0–9) for bedrooms/bathrooms; continuous logs for price and square footage.
y-axes: smoother and compressed compared to the raw plot — extreme high-end outliers no longer dominate.
The clustering along diagonals reflects log-linear scaling, consistent with housing market data patterns.

Key Takeaways of EDA

Log transformation successfully linearized and stabilized relationships among continuous variables. price_log and square_feet_log now display a clear proportional relationship ideal for regression or clustering. Discrete variables (bedrooms, bathrooms) retain ordinal structure and correlate logically with continuous ones. Data are now better balanced, reduced skew, consistent scales, and fewer distortions.

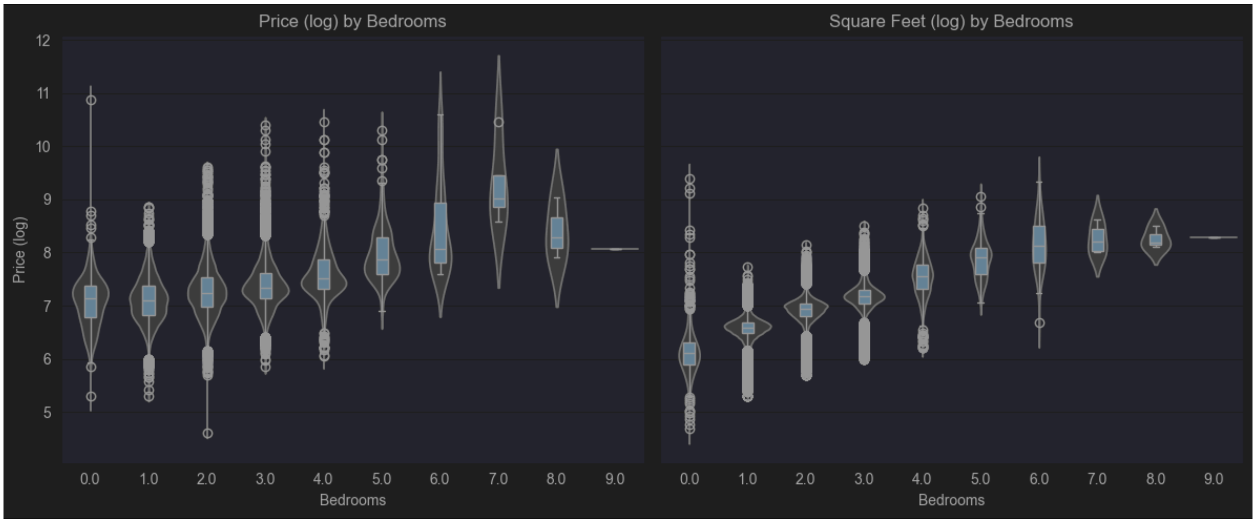


The correlation-matrix shows pairwise correlations among bathrooms, bedrooms, price_log, and square_feet_log. After applying log transformations to continuous features, relationships that were previously nonlinear and skewed now appear smoother and more proportional.

The variables types:

- Discrete / ordinal: bedrooms, bathrooms, integer counts
- Continuous: price_log, square_feet_log, continuous and normalized

We used Pearson correlation; the relationships involving discrete counts are approximate linear associations, not strict parametric correlations. Square_feet_log and price_log show that pairwise alone might be sufficient for a linear correlation.



These plots compare how both price (log) and square feet (log) vary across bedroom counts.
Each violin shows the full distribution shape, while the overlaid boxplots display median and interquartile range (IQR). Both are plotted with the same x-axis to make scale and distribution directly comparable.

Price (log) by Bedrooms, Left Plot

Clear positive association between bedroom count and log-price.
The median price gradually increases up to about 6 bedrooms.
Distribution width (spread) increases with bedrooms, larger homes show more price variability.
For higher bedroom counts (≥ 7), distributions narrow sharply, indicating rarer, high-value homes with consistent pricing. Bimodal hints at 2–3 bedrooms suggest different market tiers within typical homes.

Square Feet (log) by Bedrooms, Right Plot

Strong linear growth of home size with bedroom count.
Distributions are more compact than price, implying size scales more predictably than value.
Variability increases slightly through 5 - 6 bedrooms, then tightens again for the largest homes.
The near-parallel rise of median lines across bedroom categories reinforces the expected structural pattern:

more bedrooms, larger homes, higher prices.

Analysis

We have performed exploratory data analysis but model analysis, We have not chosen an objective function. We may consider a multivariable linear regression, perhaps classification or clustering, but we are not confident yet. We plan on investigating the categorical columns before making a decision.

Difficulties

We had difficulties cleaning the data. A problem was that the dataset is 100k rows, so our computer sometimes runs out of memory. Another problem was once, we were able to see the data, we had mixed data types in columns. In addition some data was corrupted.

We configured pandas settings to bypass the memory issue, and then made user defined methods to clean each column . We attempted to remove the obvious data that didn't represent the column, we tried to fix the data we could.

It has been difficult to find a dataset to pair with the UCI dataset. It appears that time is a very important component in linking it to other datasets. In general, it’s also been challenging to quickly learn the economics of the housing market.

Summary:

Milestones Completed:

1) Read of Data
2) Data Wrangling
3) Data Cleaning
4) Exploratory Data Analysis