Snaphomz Trial: EDA, Baseline Modeling, and Local RAG

Notebook: notebooks/trial.ipynb *Data: data/listings_sample.csv*

Highlights

- No missing values in key fields; derived target `price_per_sqft`.
- Distributions for `price` and `sqft` are right-skewed; city medians show clear pps differences.
- Engineered features: keyword flags (pool/garage/quiet/updated/backyard), ratios, logs, and city priors.
- Baseline regression (test): **RandomForest** R^2 $\approx$ 0.61, better than Linear.
- LLM feature: local TF-IDF retriever + rule-based synthesis; returns answer with supporting contexts.

Key Metrics

Model	MAE	RMSE	R^2
Linear Regression	111.4	22746.9	0.584
Random Forest Regressor	**107.9**	**21361.2**	**0.610**

Takeaway: size, city effects, and lightweight text features capture moderate signal; nonlinearity helps.

Representative Screenshots

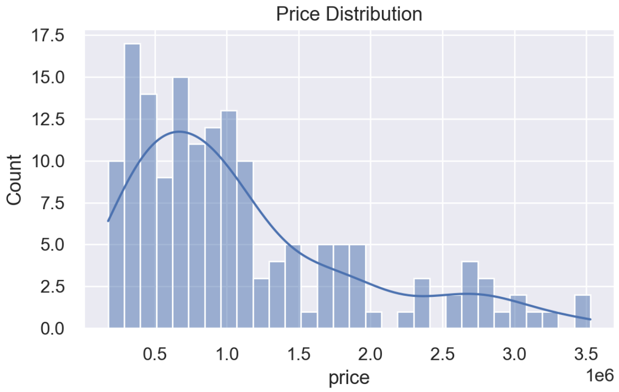

Figure 1: Price distribution (right-skewed)

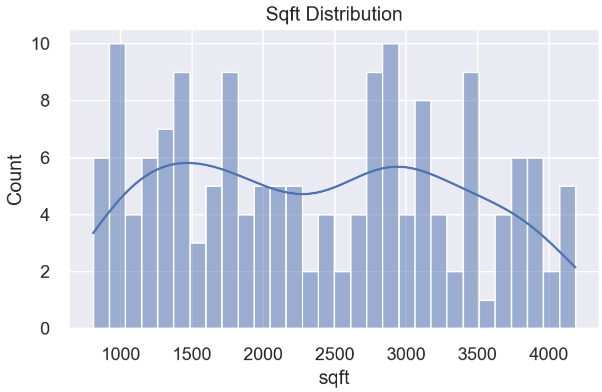

Figure 2: Sqft distribution (right-skewed)

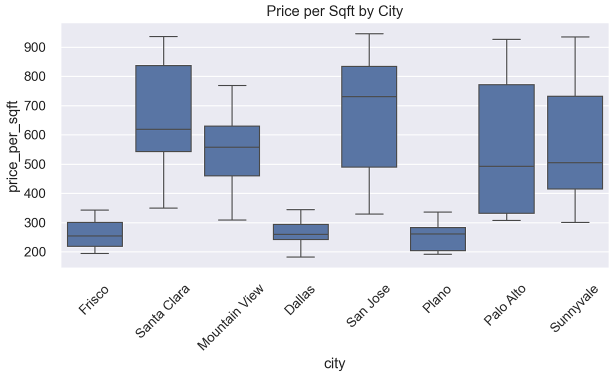

Figure 3: Price per sqft by city (boxplot)

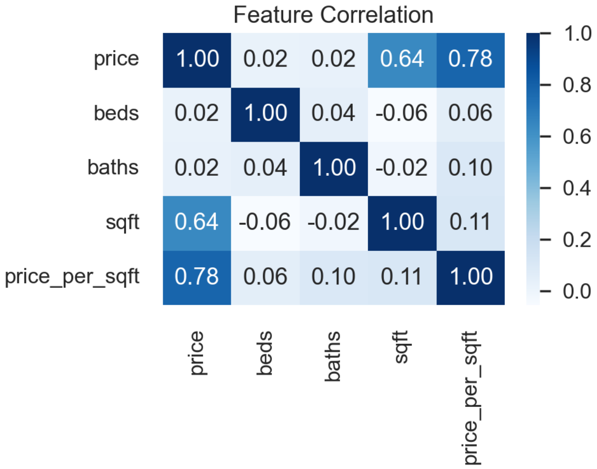

Figure 4: Correlation heatmap

LLM Mini Q&A (Local)

Retriever: TF-IDF + NearestNeighbors (cosine).
Answering: Rule-based synthesis; returns short answer + top-k contexts (`id` + remark snippet).
Example:

> Q: Do these listings typically have a pool?
> A: Here is what I found from similar listings: Pool mentioned: yes.

How to Reproduce

1. Create and activate a virtual environment; install deps:

```
python -m venv .venv
&& source .venv/bin/activate
&& pip install -r requirements.txt
```

2. Open and run `notebooks/trial.ipynb` from top to bottom.

3. Export figures using the last cell (they will be saved to `results/`).

4. Compile this one-pager to PDF:

```
cd results
&& pdflatex one_pager.tex
```

Notes (Limits & Upgrades)

- Small dataset; minimal tuning. Consider per-city models or interactions.
- Upgrade RAG: swap TF-IDF for open embeddings (e.g., BAAI/bge-small-en-v1.5) and optional reranker.
- Add local generation (e.g., FLAN-T5 small) for richer answers or 2–3 sentence summaries.

Contact: KaheiLam/kaheilam973@gmail.com. Repo: add URL.