Real Estate Price Prediction

By Koller Melanie Turinabo

Introduction

- **Dataset:** Real estate properties with attributes like square footage, number of rooms, garden/pool presence, location score, and distance from the city center.
- **Goal:** Predict property price using these features.
- **Type:** Regression problem (continuous target: Price).

Feature	Description
Square_Feet	Property area (m²)
Num_Bedrooms	Bedrooms count
Num_Bathrooms	Bathrooms count
Num_Floors	Floors count
Year_Built	Year constructed
Has_Garden / Has_Pool	1 = Yes, 0 = No
Garage_Size	Garage area (m²)
Location_Score	Neighborhood quality (0–10)
Distance_to_Center	Distance from city center (km)
Price	Target variable

Dataset Overview

ID	Square_Feet	Num_Bedrooms	Num_Bathrooms	Num_Floors	Year_Built	Has_Garden	Has_Pool	Garage_Size	Location_Score	Distance_to_Center	Price
1	143.6350297	1	3	3	1967	1	1	48	8.297631203	5.93573364	602134.8167
2	287.6785766	1	2	1	1949	0	1	37	6.061465649	10.8273922	591425.1354
3	232.9984855	1	3	2	1923	1	0	14	2.911442478	6.904599073	464478.6969
4	199.664621	5	2	2	1918	0	0	17	2.070949182	8.284018511	583105.656
5	89.00466011	4	3	3	1999	1	0	34	1.523277857	14.6482773	619879.1425
6	88.99863008	5	3	2	1959	1	1	36	8.994552125	17.63324965	670386.8044
7	64.52090304	4	3	1	1938	0	1	32	7.101354318	2.429907822	523827.1256
8	266.5440364	5	1	3	1973	1	1	39	9.373784382	12.69278513	875352.5452
9	200.2787529	5	1	1	1988	1	1	32	6.032918057	11.64287615	738269.8523
10	227.0181444	3	2	1	1917	0	0	29	4.734008815	2.368300784	490552.6812
11	55.14612357	5	2	2	1918	1	0	27	0.59716489	4.237886609	488942.1312
12	292.477463	2	1	3	1935	1	1	27	6.716422815	6.220049895	677623.4668
13	258.1106602	3	2	3	1973	1	0	11	5.822087598	3.338621416	685123.5341
14	103.0847777	5	1	2	1996	0	1	13	9.902088936	4.87458965	658652.7893
15	95.4562418	1	3	1	1954	0	1	39	2.391587778	5.352015977	413729.0743
16	95.85112746	2	2	1	2013	1	0	42	5.307373553	6.713476661	523527.9747
17	126.0605607	2	2	1	1968	0	1	45	6.509573351	14.98780906	527769.813
18	181.1891079	2	1	1	1946	1	1	27	6.445911747	11.43824408	514260.995
19	157.9862547	3	2	3	1948	1	0	46	5.612935429	19.72493344	581392.9358
20	122.807285	5	2	3	1998	1	1	28	7.769280103	8.294716539	750687.6199
21	202.9632237	5	3	2	2007	1	1	37	2.049812095	15.07730148	832883.8728
22	84.87346516	1	2	2	1997	1	0	49	7.256272003	17.42936843	479876.3646
23	123.0361621	1	3	3	1932	0	1	29	6.197345714	13.57956986	437751.643
24	141.5904608	2	1	3	2018	1	0	42	8.960751118	8.237294808	629162.356
25	164.0174961	1	2	1	1920	0	1	21	8.47194439	15.07189261	394665.751
26	246.2939903	3	3	1	1980	0	0	24	3.151470391	4.073342755	651471.4065
27	99.91844554	5	1	2	1921	1	0	48	2.76237325	6.551043304	563531.4769
28	178.5586096	2	2	3	1947	1	0	11	9.816533239	10.9337051	528557.486
29	198.1036422	1	1	2	1936	0	1	38	5.272651809	11.50818775	466357.4418
30	61.61260318	3	2	3	1914	0	0	19	9.064209639	0.442829509	428376.7208
31	201.886213	3	1	1	1934	0	0	11	1.638810238	18.90228378	438805.8696
32	92.63103092	1	1	2	1991	1	0	20	6.428189148	13.93254542	408507.9528
33	66.26289825	5	3	2	1970	0	0	22	8.098601874	9.914075005	592796.9627
34	287.2213843	1	1	3	1931	1	0	19	9.473952496	18.66254469	506721.2732
35	291.4080083	2	2	1	1910	1	1	23	3.37340141	8.781911305	595676.9793
36	252.099337	1	2	2	1909	1	1	27	7.249811699	19.60871141	504981.6036
37	126.1534423	3	3	1	1972	0	1	18	5.38813009	10.24569514	584118.0468
38	74.4180285	1	2	2	1977	1	1	48	0.606289261	2.2857154	493819.2725
39	221.0582566	5	2	1	1976	1	1	37	6.899997373	15.29601945	749855.0594

- ***Red*** : *Independent Features*
- ***Green***: *Dependent Features*

Data Cleaning

Excel & SQL

Practiced Excel formulas:

- XLOOKUP, TRIM, DATEVALUE, TEXT.

Practiced SQL operations:

- Joins (inner, outer, left, right)

Checking for NULLs and duplicates

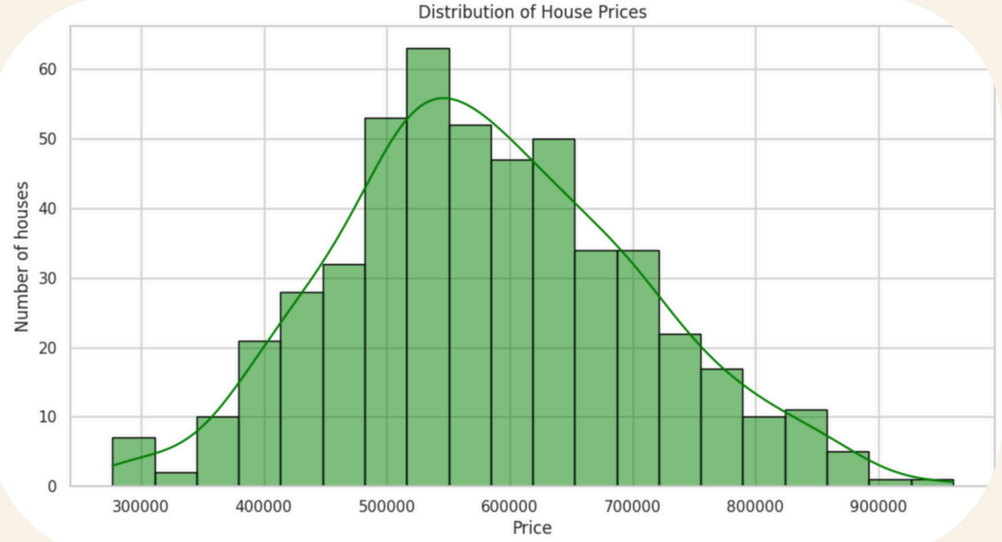
- Simulated “messy” data (missing values & categorical text) for extra practice.
- **Learned**: Cleaning synthetic dirty data improves skill understanding.

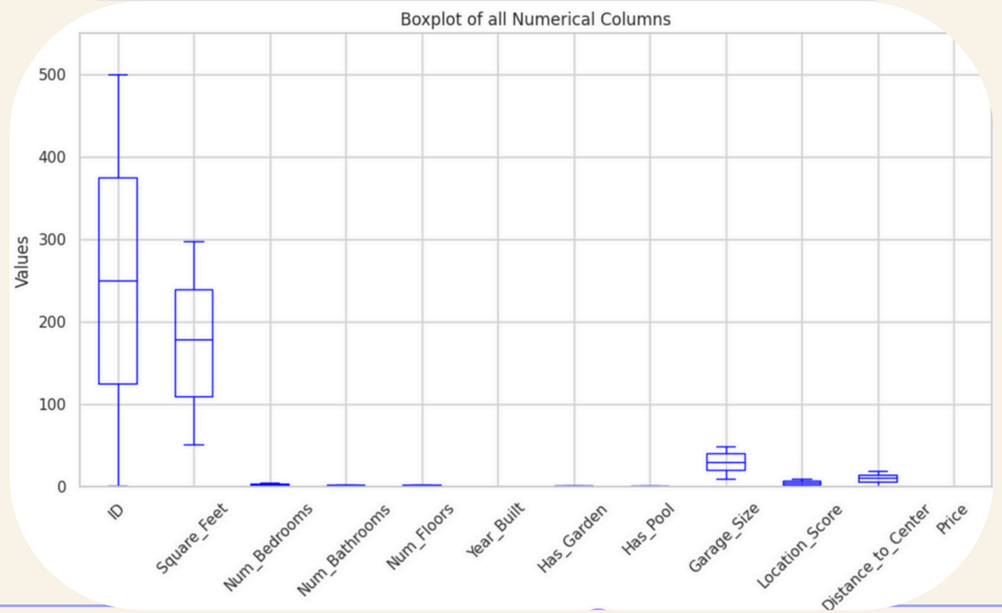
```sql
CREATE DATABASE real_estate_db;
SHOW DATABASES;
USE real_estate_db;
SHOW tables;
DESCRIBE real_estate_dataset;

CREATE TABLE trial(
    name character,
    age int
);

CREATE TABLE trial_1(
    ID int,
    Square_Feet double,
    Num_Bedrooms int
);

SHOW tables;

INSERT into  trial_1 ( ID, Square_Feet, Num_Bedrooms)
SELECT ID, Square_Feet, Num_Bedrooms
FROM real_estate_dataset;
SELECT * From trial_1;

CREATE TABLE trial_2 AS
SELECT ID, Square_Feet, Num_Bedrooms
FROM real_estate_dataset;
```

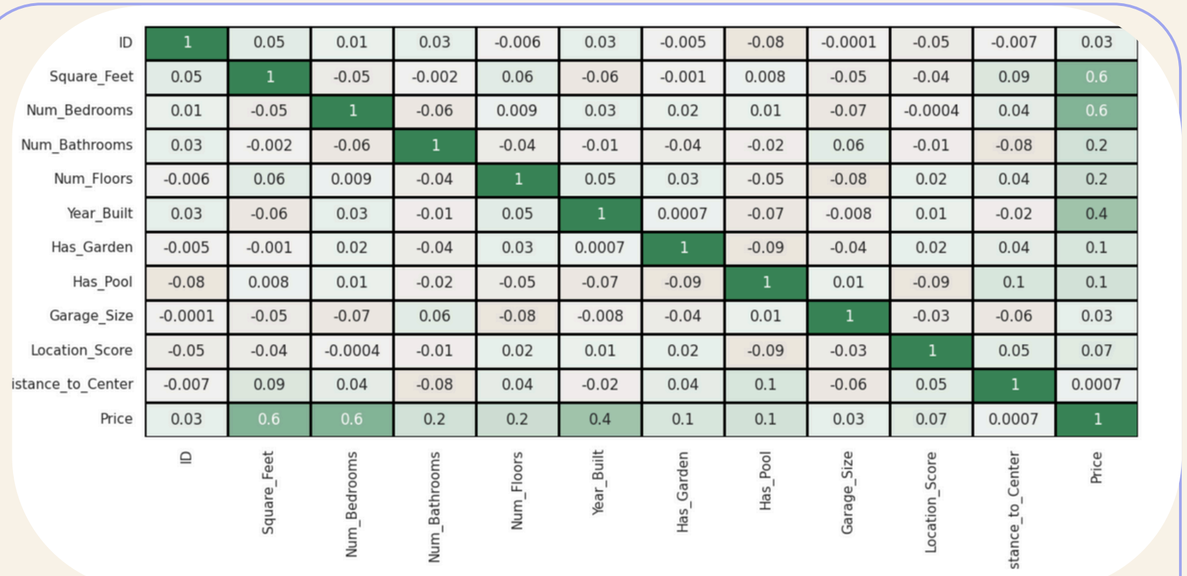
Exploratory Data Analysis

- Explored dataset using head(), info(), describe(), shape, and sample().
- Detected and handled:
  - Missing values
  - Duplicates
  - Outliers
- Created a cleaned DataFrame to prepare for modeling.
- Performed feature engineering and encoding/normalization.

Python Visualisations - EDA







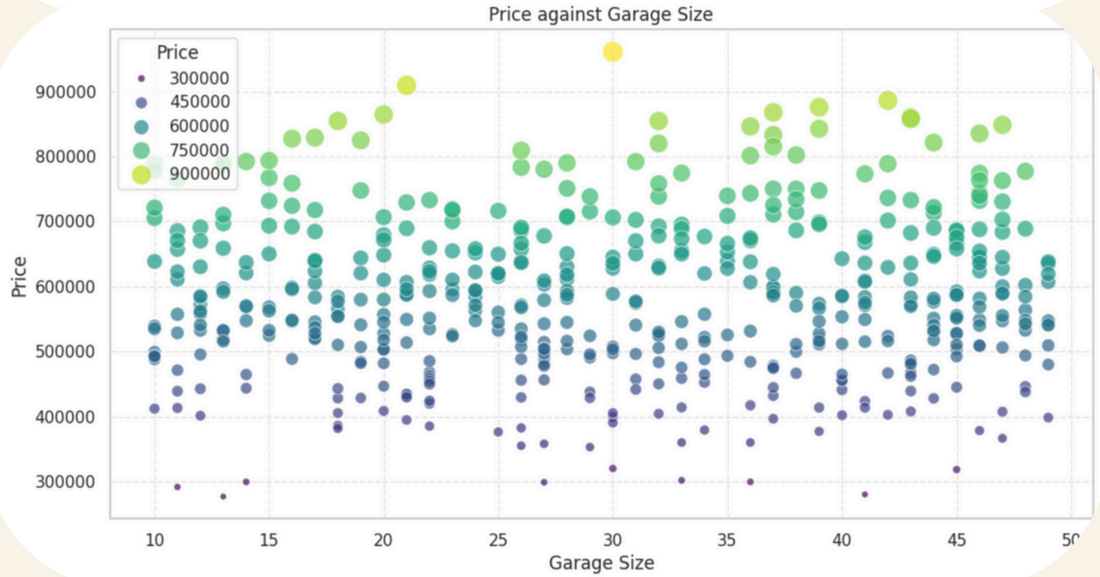
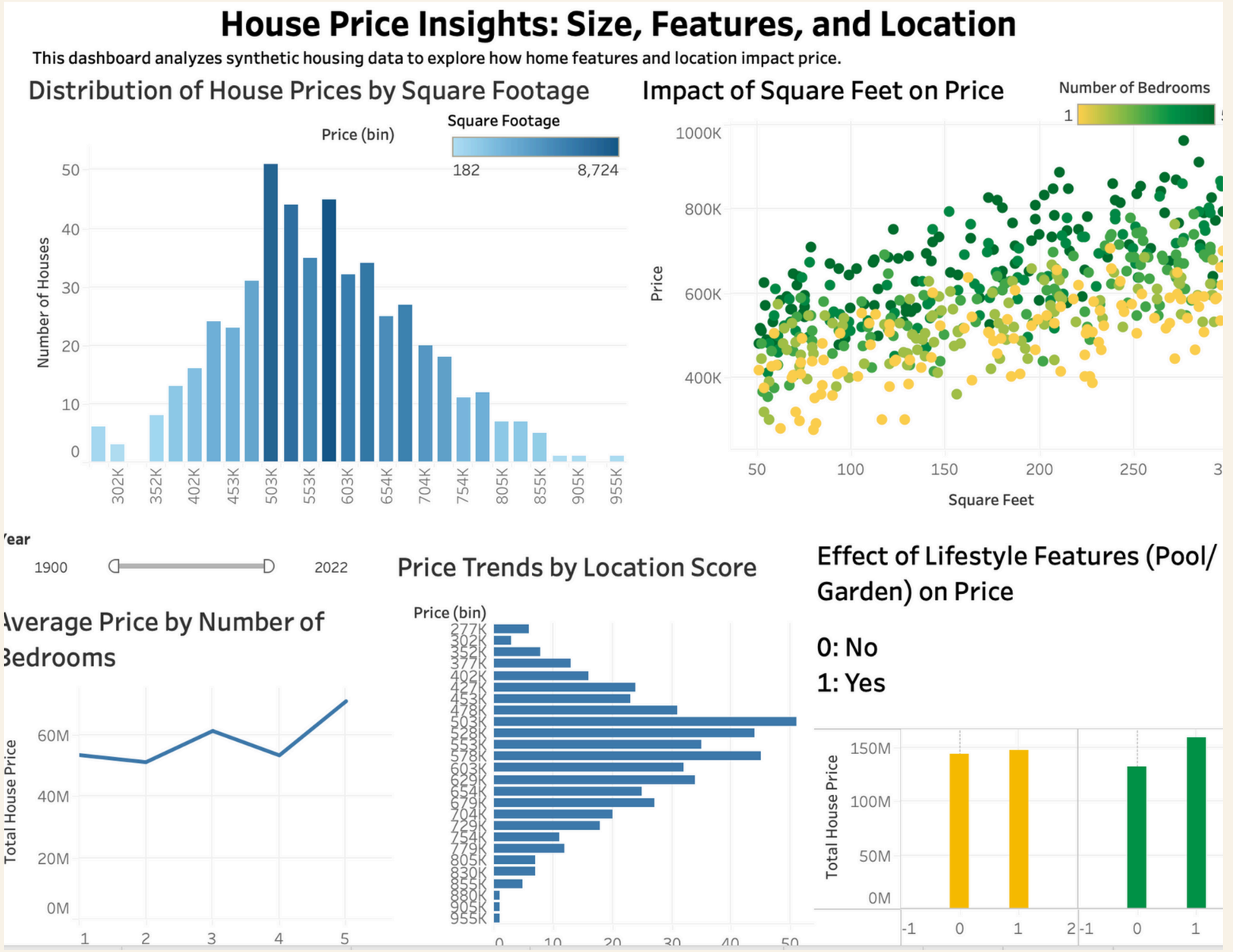


Tableau Visualizations

Key Charts:

- *Histogram: Distribution of House Prices*
- *Scatterplot: Price vs Square Feet*
- *Average Price by Number of Bedrooms*
- *Price Trends by Location Score*
- *Boxplots: Lifestyle Features (Pool/Garden) vs Price*
- *Tableau Dashboard: Combined visual view of trends*



Predictive Modeling

Machine Learning Models

Models Trained:

- Linear Regression – Baseline numeric model
- XGBoost Regressor – Sequential ensemble (strong learner)
- Random Forest Regressor – Multiple decision trees for robustness

Data Split:

- train_test_split(test_size=0.2, random_state=42)

Metrics: MAE, MSE, RMSE, R^2, RMSLE

Model	RMSE (Train)	RMSE (Test)	R^2 (Test)	Notes
Linear Regression	18,508	20,922	0.971	Best generalization
XGBoost	13,041	35,568	0.916	Overfitting
Random Forest	17,571	51,372	0.825	Strong overfitting

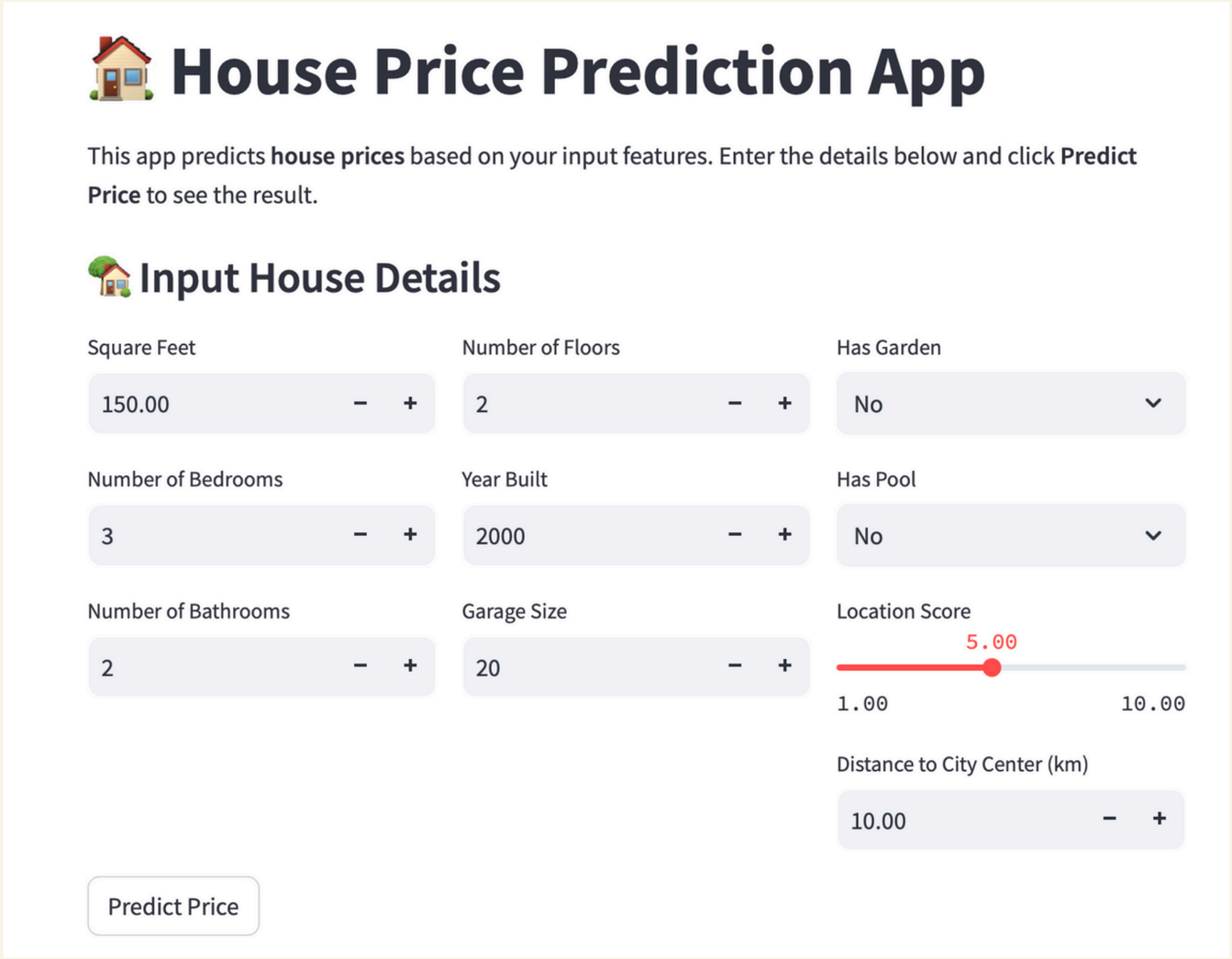
Model Choice

Final Model: Linear Regression

- Lowest generalization error
- Minimal overfitting
- Stable performance on unseen data

Model Deployment

- Built a simple Streamlit app to input property details and view predicted price in real time.



Why Predict House Prices?

- Helps real estate agencies, buyers, and investors estimate fair market values.
- Enables data-driven pricing and investment analysis.

Real-World Importance

Lessons Learned

Technical Skills

- Excel & SQL cleaning
- Applied end-to-end ML workflows (Python, scikit-learn, Streamlit)
- Cleaned and prepared real-world datasets
- Evaluated models using MSE, MAE, RMSE, RMSLE, R^2

Professional Skills

- Documented progress on GitHub & presented findings clearly
- Integrated analytics with Tableau to turn data into insights

Data analysis facilitates predictive modeling and forecasting

Conclusion

Overall Impact:

- Bridged theory and practical application
- Gained confidence in delivering data-driven insights
- Understood how data science supports business decisions

Future Work:

- Would like to test a Deep Learning model and compare with regression, XGBoost, and Random Forest.

Thank You!!

Koller Melanie
Turinabo