Data Summary:

The dataset describes data about housing properties in four cities in Saudi Arabia, which are: Riyadh, Jeddah, Dammam, and al Khobar. It contains features such as city, property age, district and the size of the property. The dataset has 24 columns and 3718 rows with 80 rows missing in the detail's column, and 2197 duplicated rows. The target column is the price column.

Description of the columns:

Column	Description	Data Type
city	City where the property is located	Categorical (string)
district	District/neighborhood of the city	Categorical (string)
front	Property orientation (e.g., North, South, West)	Categorical (string)
size	Property size in square meters	Numerical (integer)
property_age	Age of the property in years	Numerical (integer)
bedrooms	Number of bedrooms	Numerical (integer)
bathrooms	Number of bathrooms	Numerical (integer)
livingrooms	Number of living rooms	Numerical (integer)
kitchen	Whether the property has a kitchen (1 = Yes, 0 = No)	Binary (0/1)
garage	Whether the property has a garage (1 = Yes, 0 = No)	Binary (0/1)
roof	Whether the property includes roof access (1 = Yes, 0 = No)	Binary (0/1)
pool	Whether the property has a swimming pool (1 = Yes, 0 = No)	Binary (0/1)
frontyard	Whether the property has a front yard (1 = Yes, 0 = No)	Binary (0/1)
basement	Whether the property includes a basement (1 = Yes, 0 = No)	Binary (0/1)
duplex	Whether the property is a duplex (1 = Yes, 0 = No)	Binary (0/1)
stairs	Whether the property has stairs (1 = Yes, 0 = No)	Binary (0/1)
elevator	Whether the property has an elevator (1 = Yes, 0 = No)	Binary (0/1)
fireplace	Whether the property has a fireplace (1 = Yes, 0 = No)	Binary (0/1)
price	Rental price of the property (likely per year, in SAR)	Numerical (integer)
details	Free-text description of the property	Text (string)

Objective of the analysis:

The aim of the analysis is to build a predictive model to estimate rental housing property prices in Saudi Arabia to help real estate companies set their properties at a fair and competitive price, based on key features in the dataset, such as city, district and property size and other relevant features.

Model Comparison:

Model / Pipeline	**Train R^2 Score**	**Test R^2 Score**
Polynomial Features + Linear Regression + Scaling	~0.56	~0.35
Polynomial Features + Scaling + Lasso regression (with CV = 5)	~0.43	~0.33
Polynomial Features + Scaling + Ridge regression (with CV = 10)	~0.49	~0.36

Key Findings:

- Overall performance gap: All three models show a notable drop from training to test R^2, suggesting potential overfitting and that the models struggle to generalize to unseen data.
- Best-performing model: Polynomial Features + Scaling + Ridge Regression (CV=10) achieved the highest test R^2 (~0.36), slightly outperforming the plain Linear Regression (~0.35) and Lasso (~0.33).
- Regularization impact: Both Lasso and Ridge (with cross-validation) reduced overfitting compared to standard Linear Regression (train R^2

dropped from ~0.56 to ~0.43–0.49), confirming the value of regularization on this dataset.

- Data/feature limitations: The relatively low R^2 scores across all pipelines indicate that important predictive features may be missing or that the relationships in the data are highly nonlinear and not fully captured by polynomial expansions.

Limitations and Next steps:

Consider richer feature engineering (e.g., location-specific variables, amenities), experimenting with non-linear models (Random Forest, Gradient Boosting), and performing more thorough hyperparameter tuning to improve predictive power. Also getting a richer dataset to help the model in performing better on training data.