**1. Data Acquisition**

* Explain how the dataset was obtained from *realestate.com.au* (manual HTML saving + BeautifulSoup scraping).
* Mention the raw columns: address, suburb, state, postcode, date\_sold, bedroom, bathroom, carspace, size, type, price.

**2. Data Preprocessing & Exploratory Data Analysis (EDA)**

* **Cleaning:** handling missing values, converting dates, parsing numbers, removing commas.
* **Feature Engineering:** added time-based features (sale\_year, sale\_month, sale\_dow, is\_weekend), structural features (rooms\_total, size\_per\_room, carspace\_per\_rm), frequency encodings (suburb\_freq, postcode\_freq).
* **Encoding & Standardization:** used OneHotEncoder for categorical features, StandardScaler for numerics in a ColumnTransformer pipeline.
* **Visualizations:**
  + Price distribution (normal and log).
  + Correlation heatmap.
  + Outlier detection with boxplots.
  + Price trend analysis across suburbs.

**3. Model Development**

* Built three models: Linear Regression, Random Forest, Gradient Boosting.
* Used **5-fold cross-validation** with **MAE, RMSE, and R²** metrics.
* Show performance table and justify choice of best model (lowest RMSE / best R²).
* Explain underfitting vs overfitting risk.

**4. Feature Importance**

* Extracted **feature importances** from tree models.
* Plotted top 20 features.
* Used **SHAP values** to analyze contributions of features like suburb\_freq, size, bedroom, sale\_year.
* Discussion of interpretability: tree models vs linear regression.

**5. Model Deployment**

* Built a **Streamlit web app (app.py)**.
* Allows users to enter property details and sale date.
* App computes engineered features and feeds them into the trained pipeline.
* Returns predicted price with simple UI.

A computer screen shot of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.