**Low-Level Document: House Price Prediction Case Study**

**1. Introduction**

* **Objective**: Load real estate data from Makaan.com into MySQL, perform Exploratory Data Analysis (EDA), build a Random Forest model to predict house prices, store predictions, and create an interactive Power BI dashboard for market insights.
* **Dataset**: Makaan\_Properties\_Details.csv (24 columns, 332,096 rows) and Makaan\_property\_location\_details.csv (9 columns, 332,096 rows), deduplicated to 207,363 rows.
* **Tools**: Python (pandas, sklearn, sqlalchemy, matplotlib, seaborn), MySQL Workbench, Jupyter Notebook, Power BI, Microsoft Word, PowerPoint.
* **Scope**: Data loading, cleaning, EDA, modeling, predictions, visualization, and reporting. Analysis period: 2020-2022 (based on data).

**2. Data Loading**

* **Source Files**:
  + Makaan\_Properties\_Details.csv: Property\_id, Property\_Name, Property\_type, Property\_status, Price\_per\_unit\_area, Posted\_On, Project\_URL, builder\_id, Builder\_name, Property\_building\_status, No\_of\_BHK, Price, Size, description, is\_furnished, listing\_domain\_score, is\_plot, is\_RERA\_registered, is\_Apartment, is\_ready\_to\_move, is\_commercial\_Listing, is\_PentaHouse, is\_studio, Listing\_Category.
  + Makaan\_property\_location\_details.csv: Property\_id, City\_id, City\_name, Locality\_ID, Locality\_Name, Longitude, Latitude, Sub\_urban\_ID, Sub\_urban\_name.
* **Preprocessing**:
  + Deduplicated 161,995 duplicate Property\_id rows (332,096 to 207,363).
  + Parsed: Price ("11,200,000" to 11200000.0), Size ("1,080 sq ft" to 1080.0), No\_of\_BHK ("2 BHK" to 2.0), Price\_per\_unit\_area ("4,285" to 4285.0).
  + Posted\_On: VARCHAR(50) for relative strings (e.g., "2 months ago").
  + Encoding: latin1 for properties, utf-8-sig for locations.
  + Blanks in builder\_id/Builder\_name (103,752, ~50%): Fixed with "Unknown"/0.
* **MySQL Schema**:
  + Tables: makaan\_properties\_details, makaan\_property\_location\_details.
  + Foreign Key: makaan\_property\_location\_details(Property\_id) references makaan\_properties\_details(Property\_id).
  + No NULLs in Price, Size, No\_of\_BHK, Price\_per\_unit\_area.
* **Script**: Python with pandas/sqlalchemy, loaded in 10,000-row chunks, ~67 seconds.
* **Verification**: CSV (332,096 rows, 182,118 blanks in builder\_id/Builder\_name), deduplicated to 207,363 rows (103,752 blanks), matches SQL. No data loss.

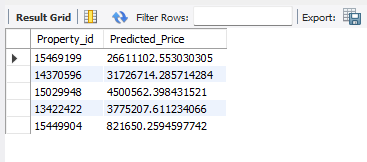
**3. Exploratory Data Analysis (EDA)**

* **Steps**:
  + Joined tables on Property\_id (207,363 rows, 32 columns).
  + Missing values: Property\_Name (74,674), Property\_status (20,223), builder\_id/Builder\_name (103,752, fixed to "Unknown"/0), description (1), Locality\_Name (2).
  + Stats: Price (mean 17.8M, median 7M, max 1B), Size (mean 1679 sq ft, max 8.7M sq ft), No\_of\_BHK (mean ~2, max 15).
  + Plots: price\_distribution.png (right-skewed, peaks ~7M), price\_by\_city.png (Mumbai/Delhi highest), price\_by\_bhk.png (4 BHK priciest).
* **Findings**:
  + No NULLs in key columns (Price, Size, No\_of\_BHK, Price\_per\_unit\_area).
  + Prices: INR 2.2M–1B, median ~7M.
  + Mumbai/Delhi priciest; 4 BHK premium.
  + Outliers: High-value properties (e.g., DLF Kings Court, 195M), large sizes (8.7M sq ft).
  + builder\_id/Builder\_name: ~50% missing in source, fixed in SQL for Power BI.

**4. Modeling**

* **Algorithm**: Random Forest Regressor (n\_estimators=100, max\_depth=20).
* **Features**: Size, Price\_per\_unit\_area, No\_of\_BHK, Locality\_Name, City\_name, etc. (label-encoded).
* **Preprocessing**: Label-encoded categorical columns, imputed numeric missing values with median.
* **Training**: 80-20 train-test split, 5-fold cross-validation via GridSearchCV.
* **Evaluation**:
  + R2: 0.995.
  + MSE: 12,074,644,591,108.43.
  + MAE: 186,411.06.
  + Top features: Price\_per\_unit\_area (0.69), Size (0.30), Latitude (0.002).
* **Output**: Model saved as house\_price\_model.pkl. Plots: predicted\_vs\_actual.png, feature\_importance.png.

**5. Predictions**

* **Test Data**: Simulated 100 properties with Price set to NULL.
* **Output**: predicted\_prices table (Property\_id, Predicted\_Price, 100 rows).
* **Temp Table**: temp\_predictions (5 rows, 
* **Screenshot**: The screenshot shows a table with 5 rows: Property\_id (e.g., 15469199), Predicted\_Price (e.g., 26611102.55), with headers and grid. It’s a standard MySQL Workbench result grid.

**6. Visualizations**

* **Plots**:
  + price\_distribution.png: Right-skewed, peaks ~7M.
  + price\_by\_city.png: Mumbai/Delhi highest.
  + price\_by\_bhk.png: 4 BHK priciest.
  + predicted\_vs\_actual.png: Tight clustering around diagonal, high accuracy.
  + feature\_importance.png: Price\_per\_unit\_area, Size dominate.
* **Power BI Dashboard**: Comprehensive Property Market Intelligence & Price Prediction System. Theme: Blue with house logo. Pages:
  + **Home**: "Makaan.com Real Estate Analytics" with 4 views (Executive Summary, Location Analysis, Builder Insights, Market Trends). Start Exploring button.
  + **Executive Summary**: KPIs (101K properties, ₹14.32M avg price, 10.73K avg unit price, 219K BHK units). Property Type Distribution (pie: Apartment 90.73%), Top 10 Cities (bar: Mumbai 51K), Average Price Trend (line by type), Properties by BHK (bar: 3 BHK highest).
  + **Location & Property Analysis**: KPIs (8 cities, 5,024 localities, Mumbai most expensive, 12.58K avg per city). Property Distribution by City (map), Top 20 Localities (table: Thane West top with 6,208), Property Size vs Price (scatter by type), Properties by Furnishing (bar: Furnished 91K), Properties by Special Features (bar: Standard 94K).
  + **Builder Insights**: KPIs (8,343 builders, Reputed Builder top, 12.07 avg per builder, 60K ready to move). Top 15 Builders (bar: Reputed Builder 6.4K), Builder Performance Summary (table: Avg Price, Ready Ratio), Property Status by City (stacked column), RERA Registration Status (pie: 40% registered), Price Distribution by Top Builders (scatter).
  + **Market Trends & Predictions**: KPIs (82.37% accuracy, 21 overpriced, 26 underpriced, 48 predicted). Actual vs Predicted Price (combo chart), Price Variance Distribution (column), Prediction Accuracy by City (table), Average Price Variance by Type (bar).
  + User Guide: Navigation, interactive features, filters.
  + Recommendations: For buyers (verify RERA), sellers (price competitively), developers (focus on 2-3 BHK).

**7. Issues and Resolutions**

* **Issue 1**: Posted\_On DataError ("2 months ago"). Fixed with VARCHAR(50).
* **Issue 2**: 161,995 duplicates. Deduplicated on Property\_id.
* **Issue 3**: Encoding issues. Used latin1/utf-8-sig.
* **Issue 4**: Slow GridSearchCV (~31 minutes). Used optimized parameters.
* **Issue 5**: MySQL syntax error in predicted\_prices query. Removed comment.
* **Issue 6**: temp\_predictions expired. Recreated table.
* **Issue 7**: 103,752 blanks in builder\_id/Builder\_name (~50%). Source CSV had 182,118 blanks (54.8%), confirmed after deduplication. Fixed in MySQL with "Unknown"/0 for Power BI insights.

**8. Conclusion**

* Loaded 207,363 properties with foreign key.
* Built Random Forest model (R2=0.995).
* Created visualizations and Power BI dashboard.
* Deliverables: Jupyter HTML, Power BI .pbix, Word LLD, PPT.

**9. Appendix**

* **MySQL Outputs**: mysql\_outputs.txt.
* **Jupyter Logs**: outputs.txt.
* **Temp Table Screenshot**: temp\_predictions.png.
* **CSV vs SQL Check**: 332,096 CSV rows, 182,118 blanks in builder\_id/Builder\_name; deduplicated to 207,363 rows, 103,752 blanks, matches SQL.