VIETNAM NATIONAL UNIVERSITY-HO CHI MINH CITY

HO CHI MINH UNIVERSITY OF SCIENCE

FACULTY OF INFORMATION TECNOLOGY

KNOWLEDGE ENGINEERING DEPARTMENT

FINAL PROJECT

Course: Programming for Data Science

Student:

23127004 - Le Nhat Khoi

23127165 - Nguyen Hai Dang

23127248 - Nguyen Huu Phuc

Lecturer:

Mr. Le Nhut Nam

Ngày 26 tháng 12 năm 2025

Mục lục

1 Team Information & Roles

Student Name	Student ID	Primary Role	Key Responsibilities
Le Nhat Khoi	23127004	Statistician & Analyst	Advanced Data Cleaning Logic, Statistical Profiling, Hypothesis Testing, Econometric Modeling (Q2, Q3).
Nguyen Hai Dang	23127165	ML Engineer & Team Lead	Pipeline Architecture, Machine Learning in Q5, Q6 (Clustering & Regression), Project Documentation, Workflow Management.
Nguyen Huu Phuc	23127248	Data Engineer & Analyst	Data Acquisition, Raw Data Parsing (Regex), Exploratory Data Analysis (EDA), Visualization (Q1, Q4).

2 Detailed Work Breakdown

All team members contributed **100% effort**, with tasks distributed based on individual strengths and technical complexity.

Le Nhat Khoi (Statistician & Analyst)

- **Phase 1: Advanced Data Processing**
  - Engineered the **Cross-Swap algorithm** in `03_data_processing.ipynb` to automatically detect and correct misaligned data (e.g., swapping values when transaction data appeared in the floor column).
  - Standardized complex categorical variables including `transaction`, `furnishing`, and `status`.
- **Phase 2: Statistical Profiling**
  - Generated correlation matrices and analyzed categorical inconsistencies.
- **Phase 3: Statistical Modeling**
  - **Q2 (Neighborhood Premium):** Implemented hedonic regression models to isolate locality-specific price premiums while controlling for structural confounders.
  - **Q3 (Pricing Uncertainty):** Applied IQR analysis and Levene's Test to quantify market volatility and pricing risk across BHK segments.

Nguyen Hai Dang (ML Engineer & Team Lead)

- **Phase 1: Architecture & Integration**
  - Designed the project directory structure.
  - Integrated individual modules into a unified pipeline, standardizing outputs into `surat_cleaned.np`
- **Phase 2: Machine Learning**
  - **Q5 (Market Segmentation):** Implemented unsupervised learning using K-Means clustering with the Elbow Method.
  - **Q6 (Price Prediction):** Trained and tuned predictive models (Linear Regression, XGBoost, Random Forest) and analyzed feature importance.
- **Phase 3: Documentation & Final Delivery**
  - Authored `06_project_summary.ipynb`.
  - Compiled `README.md` and conducted final code review.

Nguyen Huu Phuc (Data Engineer & Analyst)

- **Phase 1: Data Ingestion & Parsing**
  - Sourced and evaluated the dataset from Kaggle and defined the domain context (Surat, India).
  - Authored `01_data_collection.ipynb`.
  - Developed critical **Regex functions** to parse unstructured text fields, including conversion of Indian numbering units (e.g., `Lac`, `Cr`) into numerical values and standardization of area units.
- **Phase 2: Exploratory Data Analysis (EDA)**
  - Led `02_data_exploration.ipynb`: conducted data type analysis, missing value detection, and visualization of feature distributions.
  - Performed outlier detection and removal for key numerical variables such as `area_sqft` and `price`.
- **Phase 3: Domain Analysis**
  - **Q1 (Unit Price Efficiency):** Analyzed the non-linear relationship between property size and unit price to identify diminishing returns.
  - **Q4 (Floor Effect Analysis):** Investigated the "Floor Premium" paradox by comparing buyer behavior in primary (new) versus secondary (resale) markets.

3 Collaboration Process

- **Version Control:** GitHub was used with feature-branch workflow and Pull Requests for code review.
- **Data Strategy:** A "Golden Source" dataset (`surat_cleaned.npy`) was finalized in Week 2 to enable parallel analysis without data conflicts.
- **Communication:** Weekly sync meetings were conducted to resolve data issues, such as handling high missing rates in the `description` column.

4 Project Timeline

Week	Phase	Key Activities
Week 1	Initiation	Dataset selection (Kaggle); formulation of six research questions; role assignment.
Week 2	Data Engineering	Deep EDA; development of cleaning pipeline (Regex and Cross-Swap); finalization of `03_data_processing.ipynb`.
Week 3	Analysis	Parallel execution of Q1–Q4; visualization and statistical testing.
Week 4	Modeling & Closing	Machine learning implementation (Q5, Q6); writing conclusions; final report assembly and code refactoring.

5 Project Directory Structure

Project Architecture

```
surat-housing-analysis/
│
├── 01_data_collection.ipynb          # Data sourcing & licensing
├── 02_data_exploration.ipynb         # EDA & integrity checks
├── 03_data_preprocessing.ipynb       # 🔧 Cleaning pipeline
│
├── ❓ 04_question_formulation.ipynb     # Question statement and motivation
├── 05_data_analysis_Q1.ipynb         # Q1: Unit Price Efficiency
├── 05_data_analysis_Q2.ipynb         # Q2: Neighborhood Premiums
├── 05_data_analysis_Q3.ipynb         # Q3: Pricing Uncertainty
├── 05_data_analysis_Q4.ipynb         # Q4: Floor Effect Analysis
├── 05_data_analysis_Q5.ipynb         # Q5: Market Segmentation
├── 05_data_analysis_Q6.ipynb         # Q6: Price Prediction Models
│
├── 06_project_summary.ipynb          # Final report & conclusions
│
├── data/
│   ├── raw/surat_uncleaned.csv       # Original dataset
│   └── processed/surat_cleaned.npy   # Processed data (NumPy)
│
├── requirements.txt                   # Python dependencies
├── TEAM_PLAN.md                      # Collaboration strategy
└── README.md                         # This file
```