

# **Đồ án xây dựng mô hình dự đoán giá nhà**

Nhóm: Hợp Hoan Tông

Nguyễn Tuấn Công - 22120039

Âu Lê Tuấn Nhật - 22120250

Bùi Trọng Trịnh - 22120309

Nguyễn Lê Phúc Thắng - 221203

Liêu Hải Lưu Danh - 22120459

# Mục tiêu đồ án

Giúp sinh viên thực hành các kỹ thuật phân tích và xử lý dữ liệu bảng (tabular data), đồng thời hiểu thêm về bài toán dự đoán có ràng buộc khoảng tin cậy (prediction interval regression).



# Hoạt động của nhóm

Các công việc chung (ai cũng tham gia): viết báo cáo, làm slides.

| STT | MSSV     | Họ tên               | Phân công   | Mức độ hoàn thành |
|-----|----------|----------------------|---|-------------------|
| 1   | 22120039 | Nguyễn Tuấn Công     | Thực hiện đánh giá các mô hình, thuyết trình video  | 100%              |
| 2   | 22120250 | Âu Lê Tuấn Nhật      | Thực hiện cài đặt và huấn luyện mô hình Quantile Regression, LightGBM, XGBoost, Gradient Boosting         | 100%              |
| 3   | 22120332 | Nguyễn Lê Phúc Thắng | Thực hiện cài đặt và huấn luyện mô hình Ensemble, so sánh các mô hình, phân tích feature importance       | 100%              |
| 4   | 22120390 | Bùi Trọng Trịnh      | Thực hiện EDA và đưa ra dữ liệu đã qua xử lý  | 100%              |
| 5   | 22120459 | Liêu Hải Lưu Danh    | Thực hiện trực quan hóa kết quả dự đoán của mô hình tốt nhất và đưa ra insights, tạo file nộp trên Kaggle | 100%              |

Bảng phân công và đánh giá mức độ  
hoàn thành công việc của đồ án

# Luồng hoạt động

01

**EDA:** Phân tích các đặc trưng của dữ liệu, phân bố, mối quan hệ của các biến...

02

**Xử lý dữ liệu:** xử lý dữ liệu thiếu, outliers, feature engineering, chuẩn hoá dữ liệu... và lưu lại.

03

**Huấn luyện mô hình và đánh giá** trên tập validation (tập validation tách từ tập train).

04

**Chọn mô hình tốt nhất,** huấn luyện trên toàn bộ dữ liệu và dự đoán với tập test.

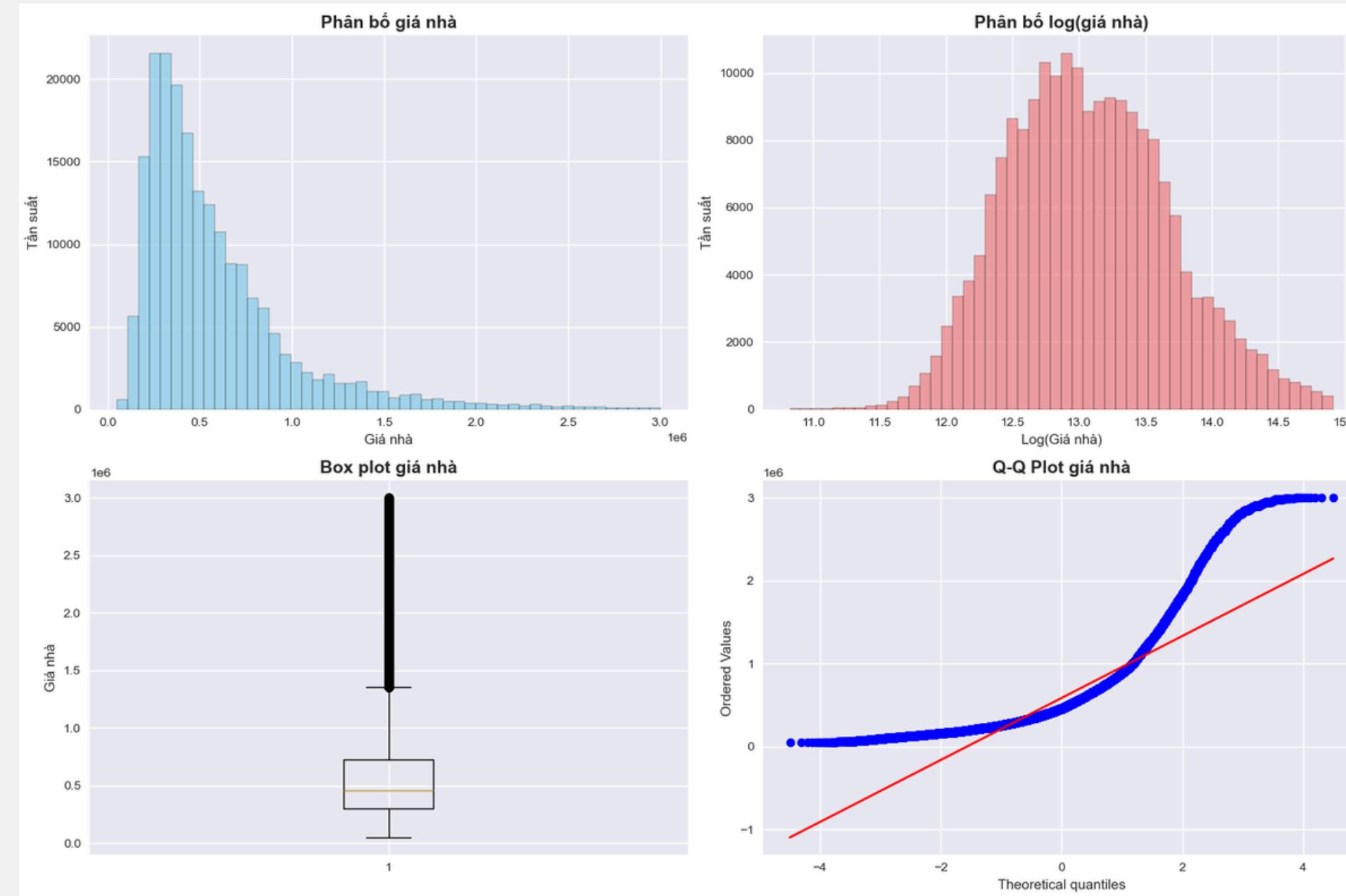
\*Tập train: tập dữ liệu huấn luyện

\*Tập validation: tập dữ liệu kiểm định

\*Tập test: tập dữ liệu kiểm thử

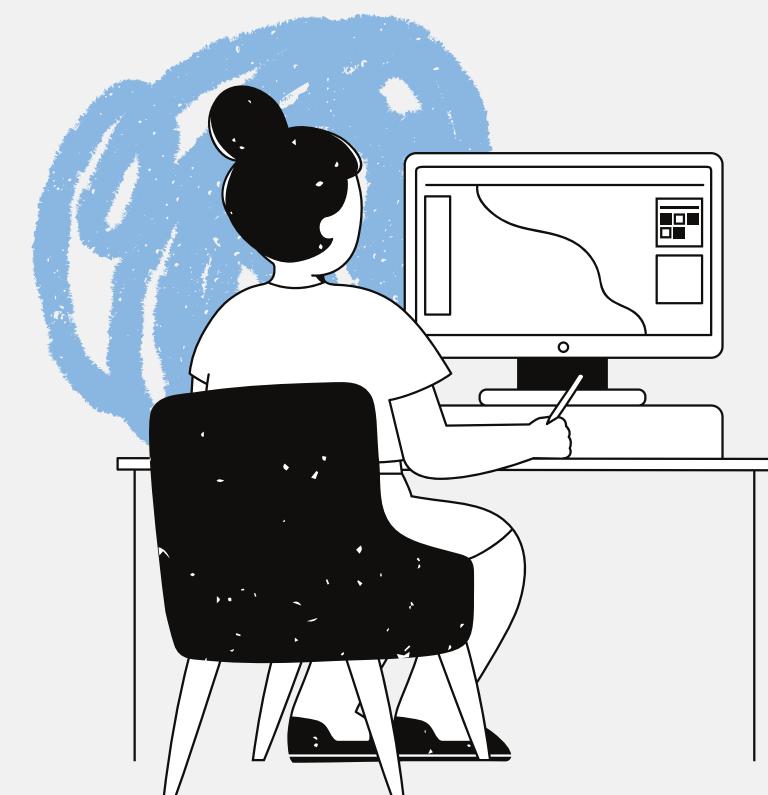
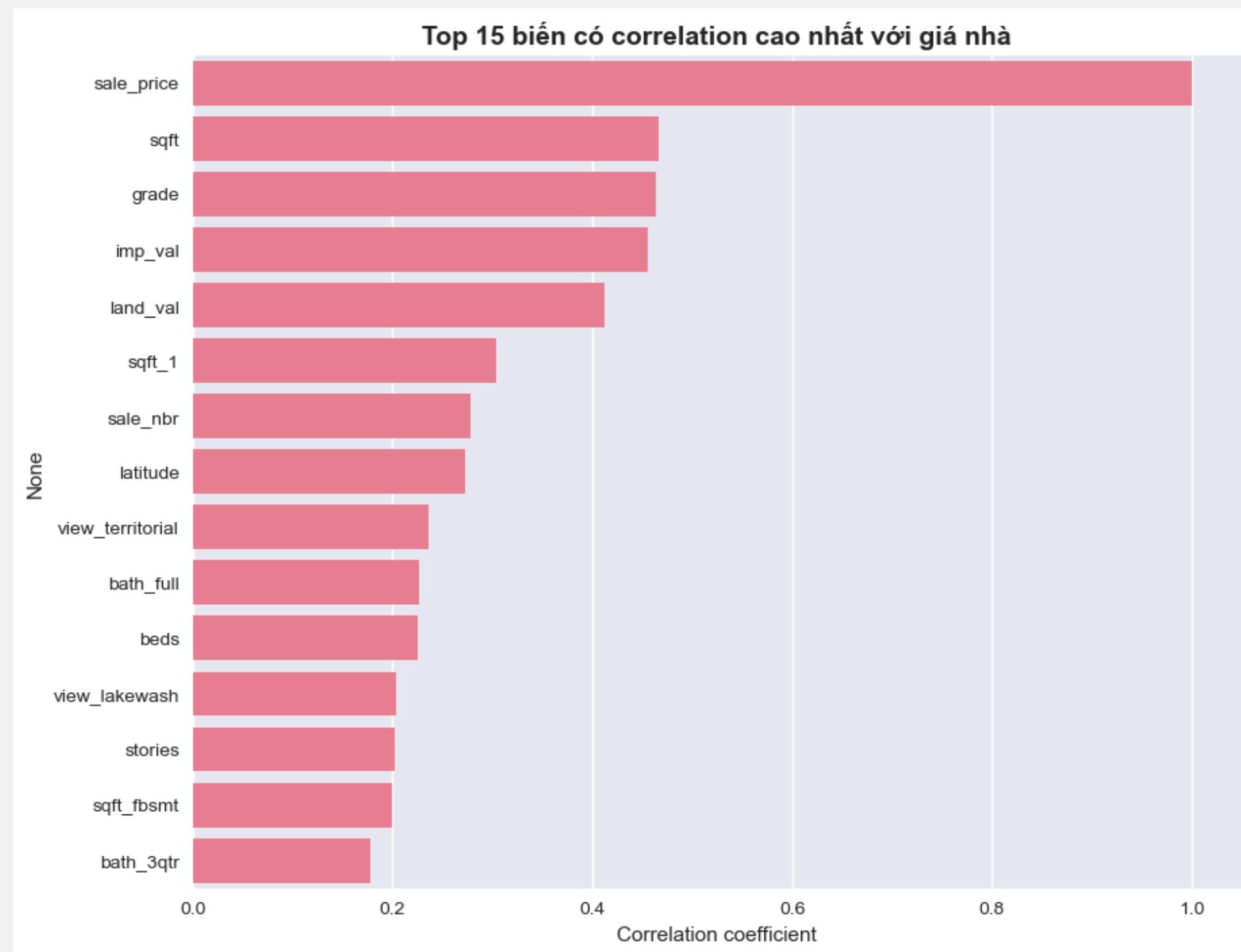
# EDA

- Tổng quan về dữ liệu: Bộ dữ liệu có 200,000 giao dịch bất động sản với 47 đặc trưng cho tập huấn luyện và 46 cho tập kiểm thử. Biến mục tiêu là giá bán nhà.
- Phân tích biến mục tiêu: Giá bán nhà có xu hướng lệch phải với trung bình 584,149.5 USD và giá trị lớn nhất trên 2 triệu USD, có ảnh hưởng của outliers.



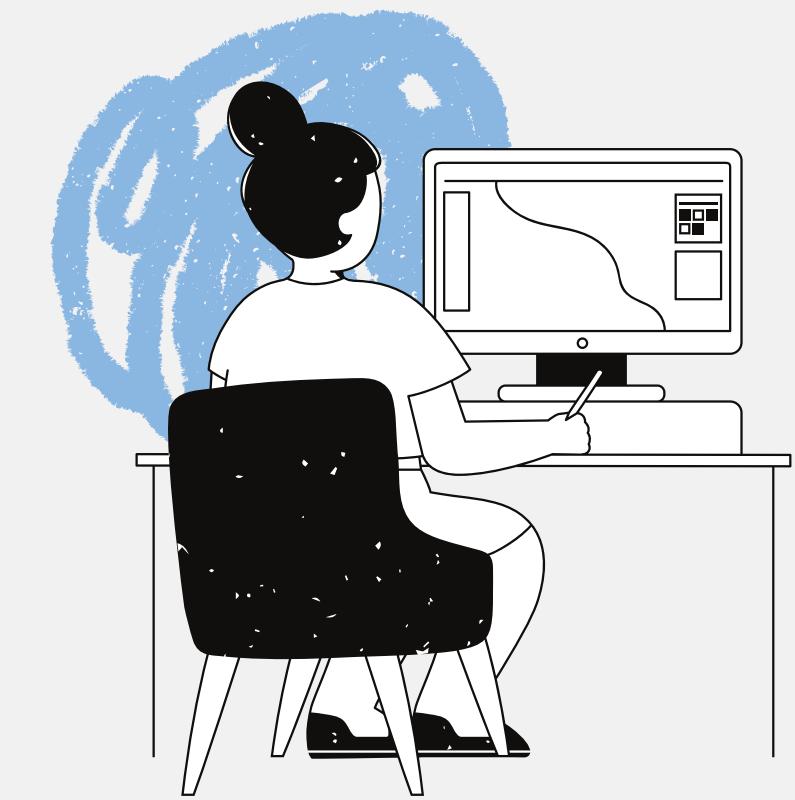
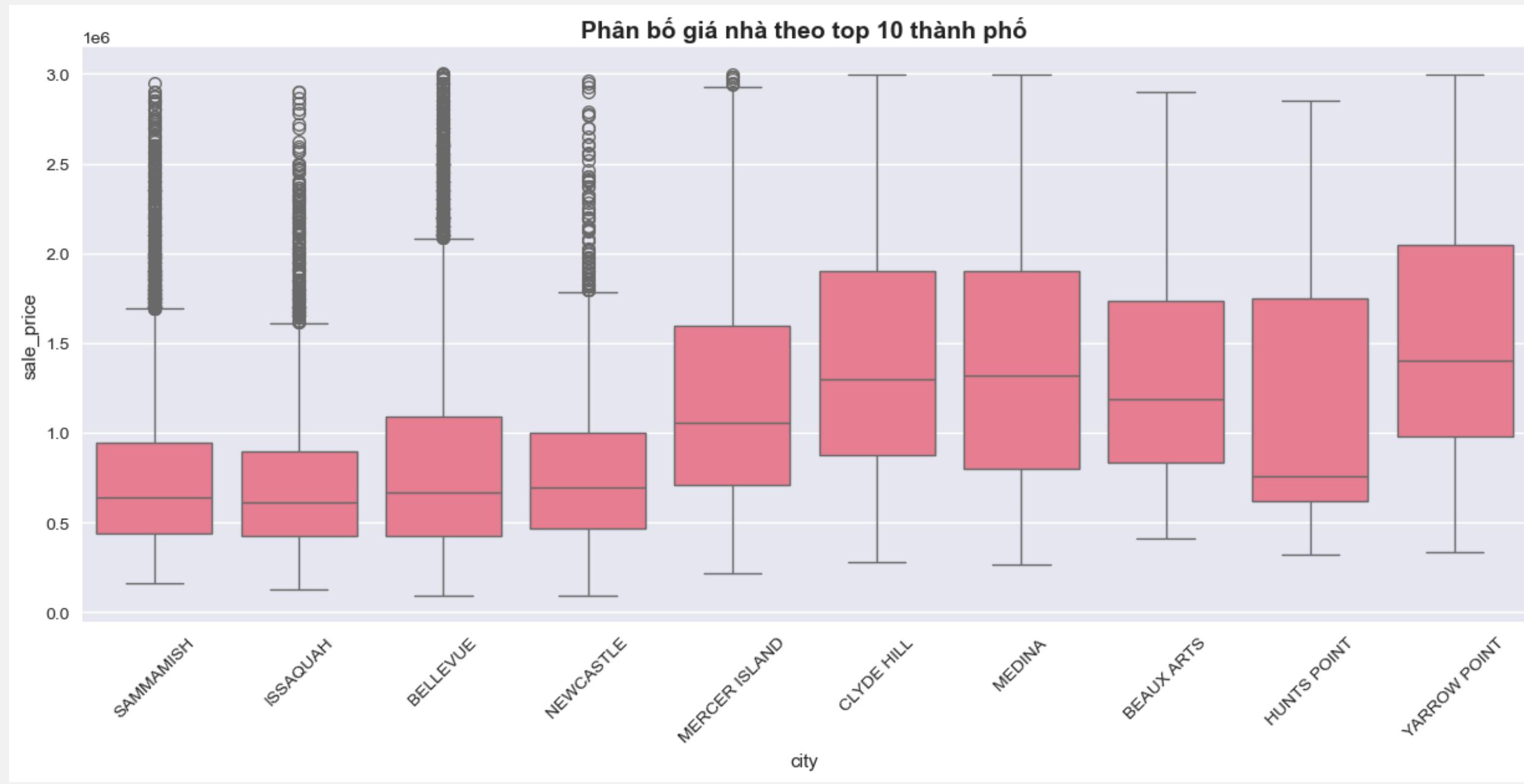
# EDA

- Phân tích tương quan: Diện tích nhà (sqft) và cấp độ xây dựng (grade) là hai yếu tố quan trọng nhất ảnh hưởng đến giá.



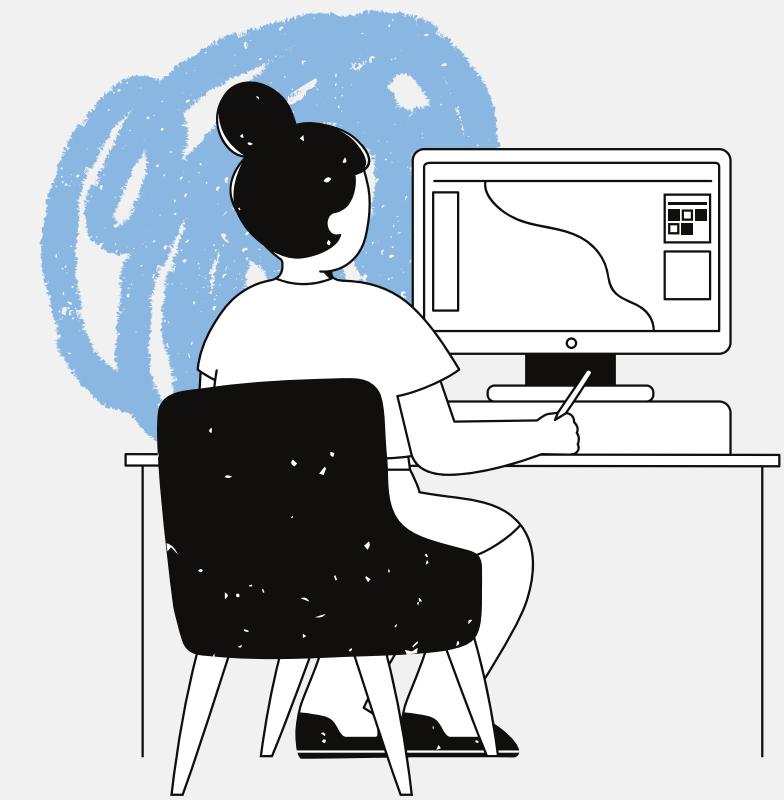
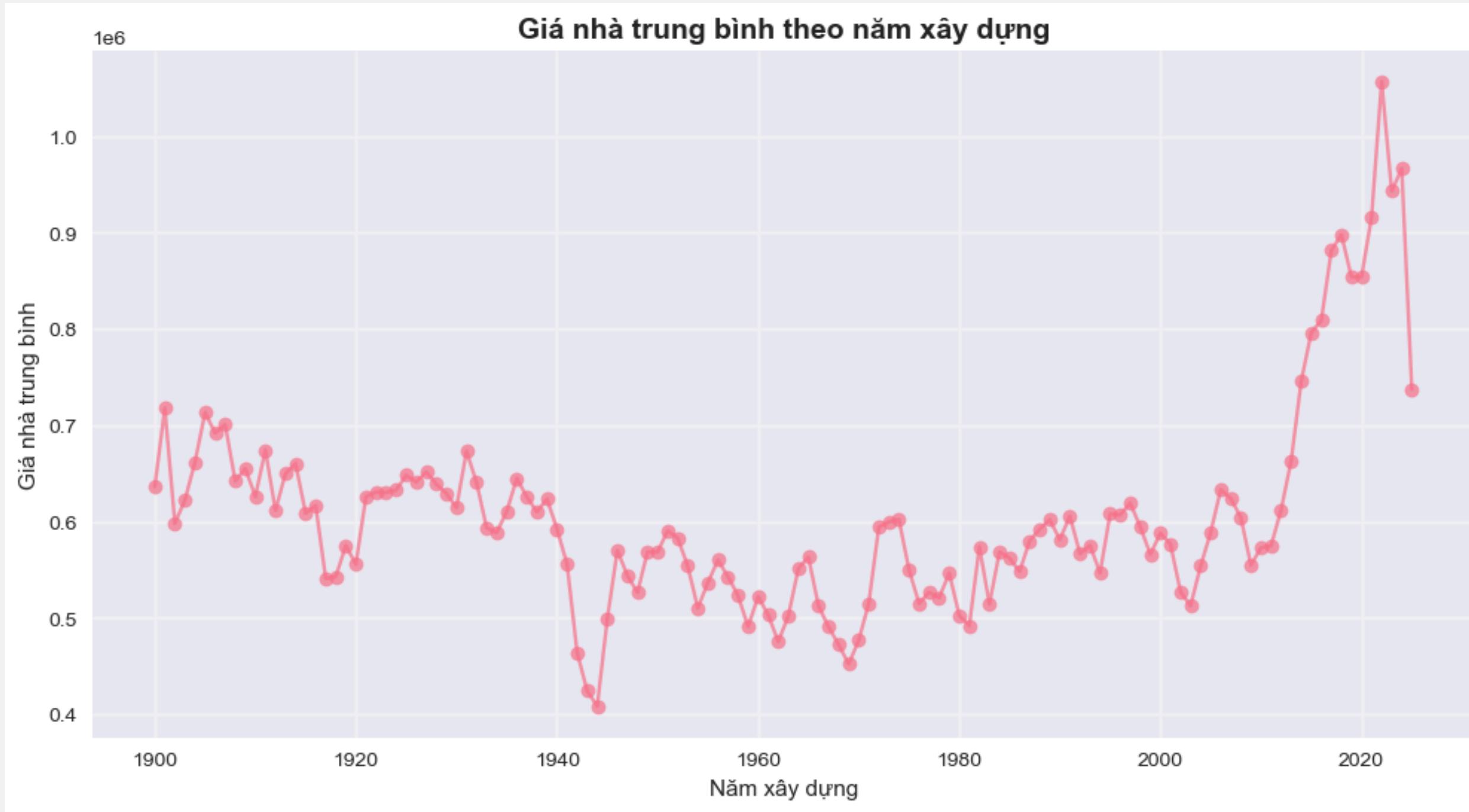
# EDA

- Phân tích theo thành phố: Có sự chênh lệch lớn về giá nhà giữa các thành phố, với Yarrow Point có giá trung bình cao nhất.



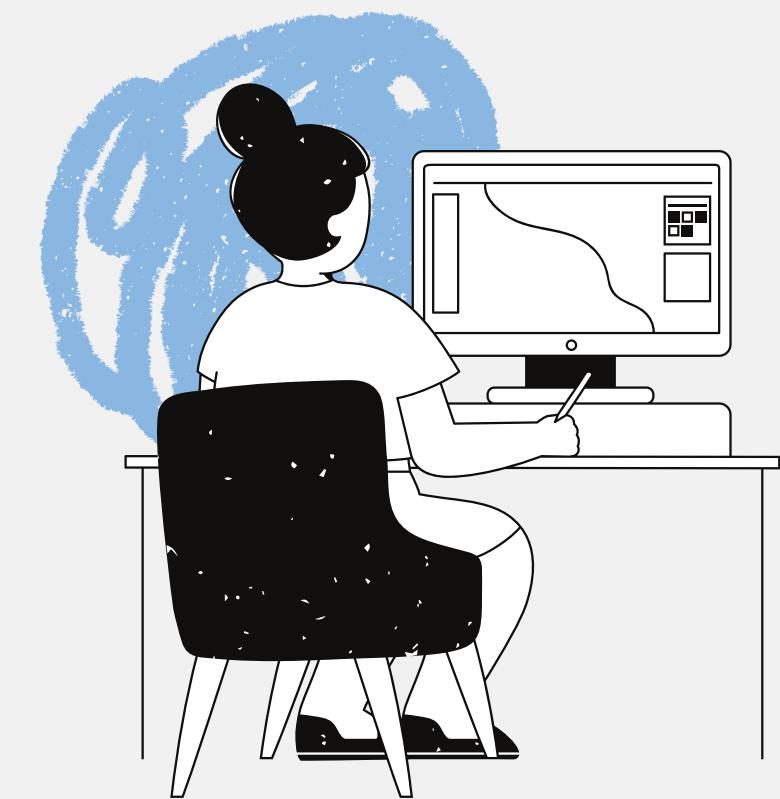
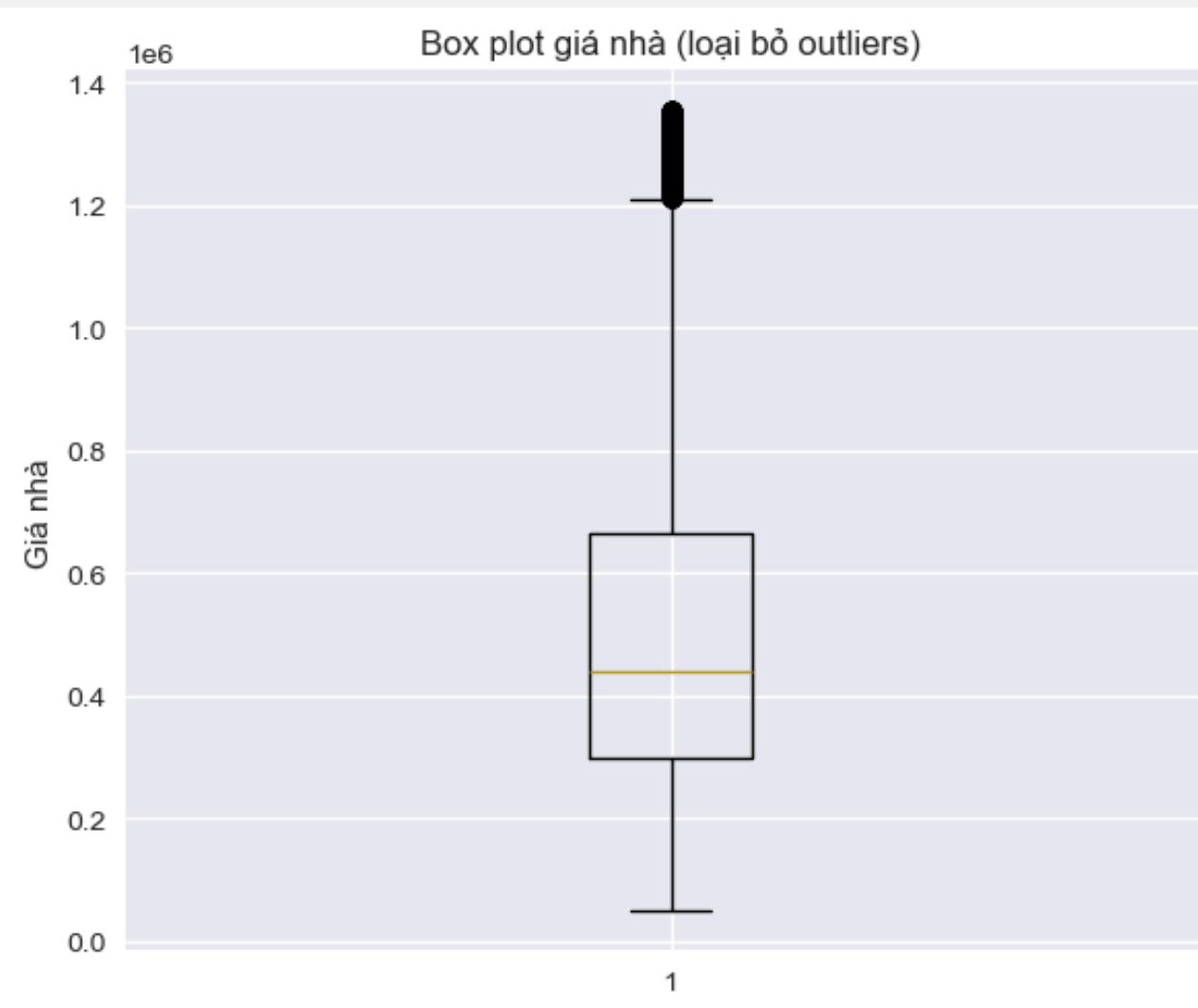
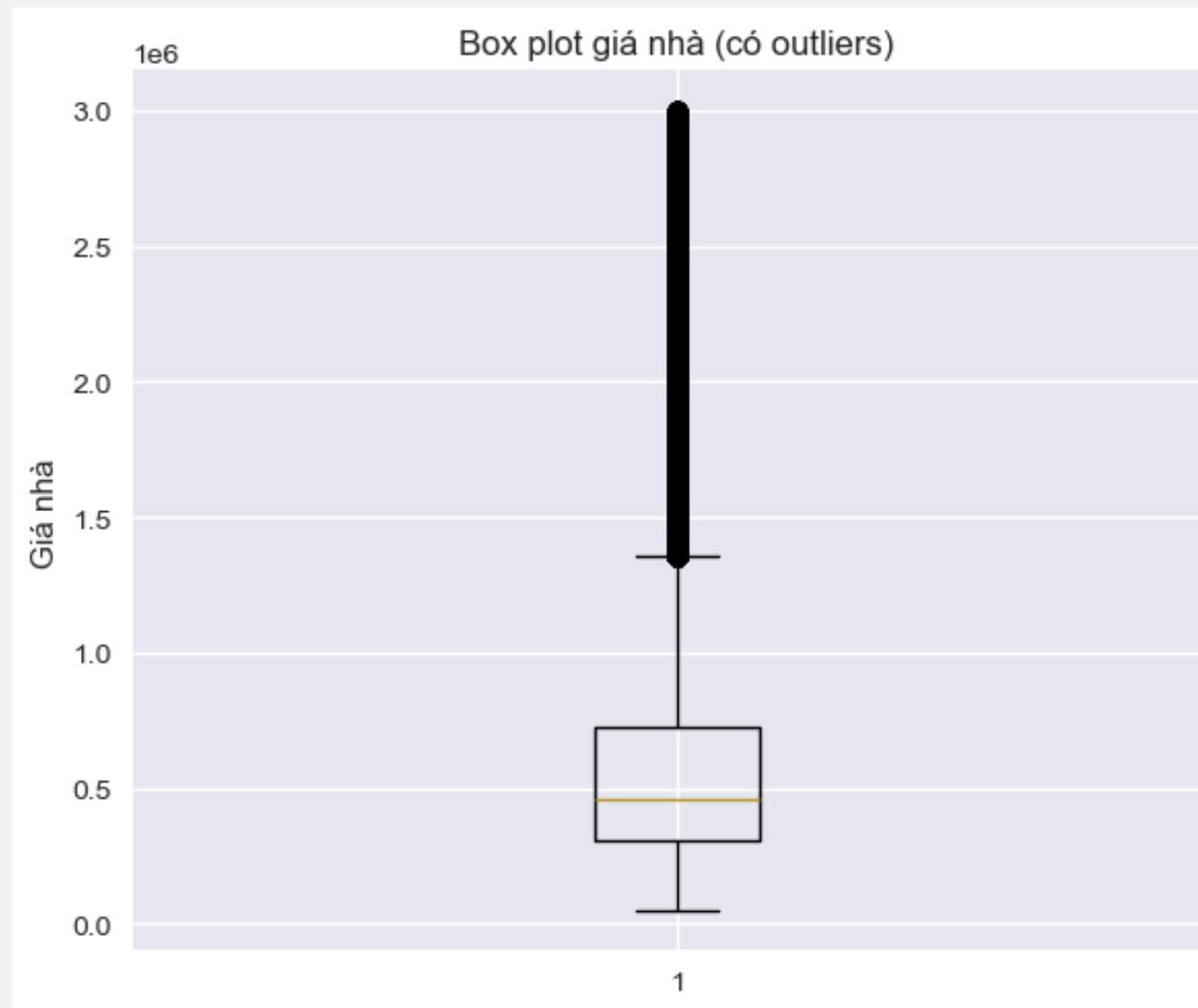
# EDA

- Phân tích theo năm xây dựng: Nhà mới thường có giá cao hơn, trong khi nhà xây trước 1950 thường rẻ hơn.



# EDA

- Phân tích giá trị ngoại lai: Phát hiện 11,736 giá trị ngoại lai, giữ lại những giá trị phù hợp.



# Xử lý dữ liệu

Xử lý dữ liệu thiếu:

- Biến dạng số: Điền giá trị trung vị/trung bình tùy theo phân phối
- Biến phân loại: Điền giá trị phổ biến nhất hoặc tạo nhóm "Không xác định"
- Xử lý cụ thể: sale\_nbr: điền giá trị trung vị theo thành phố; subdivision: tạo nhóm "Không có khu phân lô"

```
# sale_nbr: Fill với median
if 'sale_nbr' in df.columns:
    df['sale_nbr'].fillna(df['sale_nbr'].median(), inplace=True)

# subdivision: Fill với 'Unknown'
if 'subdivision' in df.columns:
    df['subdivision'].fillna('Unknown', inplace=True)

# submarket: Fill với 'Unknown'
if 'submarket' in df.columns:
    df['submarket'].fillna('Unknown', inplace=True)

# Các categorical columns khác
categorical_cols = df.select_dtypes(include=['object']).columns
for col in categorical_cols:
    if col not in ['subdivision', 'submarket']:
        df[col].fillna(df[col].mode()[0] if len(df[col].mode()) > 0 else 'Unknown', inplace=True)

# Các numerical columns
numerical_cols = df.select_dtypes(include=[np.number]).columns
for col in numerical_cols:
    if col not in ['id', 'sale_price']: # Không fill cho id và target
        df[col].fillna(df[col].median(), inplace=True)
```



# Xử lý dữ liệu

Mã hóa (encoding) biến phân loại

- Mã hóa nhãn: city, zoning, subdivision, submarket
- Mã hóa một chiều: sale\_warning, join\_status
- Lý do chọn mã hóa nhãn: Giảm số chiều, phù hợp với mô hình dựa trên cây

Chuẩn hóa đặc trưng số

- Phạm vi: Tất cả 60 đặc trưng số
- Công thức:  $X_{chuẩn hóa} = \frac{X - \mu}{\sigma}$

Chuẩn bị dữ liệu cuối cùng

- X\_train: (200,000, 63) - 63 đặc trưng sau feature engineering
- y\_train: (200,000,) - Biến mục tiêu
- X\_test: (200,000, 63) - Đặc trưng kiểm thử





# Mô hình dự đoán

Nhóm thực hiện cài đặt tổng cộng 05 mô hình, so sánh hiệu năng và chọn ra mô hình tốt nhất để dự đoán.

# Mô hình dự đoán

## Hồi quy phân vị (Quantile Regression)

Đặc điểm:

- Dự đoán các phân vị có điều kiện thay vì giá trị trung bình
- Phù hợp với bài toán yêu cầu khoảng dự đoán
- Bền vững với giá trị ngoại lai

Tham số:

- Phân vị: 0.05 và 0.95 (khoảng tin cậy 90%)
- Alpha = 0.01 (điều chuẩn)
- Bộ giải: 'highs'

Kết quả:

- Độ bao phủ: 90.2%
- Điểm Winkler: 1,743,074.51



# Mô hình dự đoán

## LightGBM

Đặc điểm:

- Gradient boosting của Microsoft
- Phát triển cây theo lá thay vì theo mức
- Tối ưu hóa cho tốc độ và hiệu suất

Tham số chính:

- Tốc độ học: 0.05
- Số ước lượng: 1000
- Độ sâu tối đa: 8
- Tỷ lệ đặc trưng/đóng gói: 0.8
- Hàm mục tiêu: 'quantile' với alpha=[0.05, 0.95]

Kết quả:

- Độ bao phủ: 86.6%
- Điểm Winkler: 340,167.11



# Mô hình dự đoán

## XGBoost

Đặc điểm:

- Triển khai tối ưu của gradient boosting
- Hiệu suất cao trong các cuộc thi học máy
- Có điều chỉnh L1 và L2

Tham số chính:

- Tốc độ học: 0.05
- Độ sâu tối đa: 6
- Số ước lượng: 1000
- Mẫu con/cột: 0.8
- Điều chỉnh alpha: 0.05, lambda: 1.0
- Hàm mục tiêu: reg:quantileerror

Kết quả:

- Độ bao phủ: 87.0%
- Điểm Winkler: 342,390.46



# Mô hình dự đoán

## Gradient Boosting

Đặc điểm:

- Triển khai chuẩn từ scikit-learn
- Tập trung vào tính ổn định
- Phù hợp cho môi trường sản xuất

Tham số chính:

- Số ước lượng: 800
- Tốc độ học: 0.05
- Độ sâu tối đa: 6
- Hàm mất mát: quantile với alpha=[0.05, 0.95]

Kết quả:

- Độ bao phủ: 88.7%
- Điểm Winkler: 357,562.68



# Mô hình dự đoán

## Ensemble (kết hợp mô hình)

Phương pháp:

- Trung bình có trọng số của các mô hình tốt nhất
- Công thức:  $\hat{y}_{kết hợp} = \sum w_i \cdot \hat{y}_i$

Trọng số lựa chọn tối ưu:

- XGBoost: 0.4
- LightGBM: 0.35
- Gradient Boosting: 0.25
- Quantile Regression: 0.0 (loại bỏ)

Kết quả:

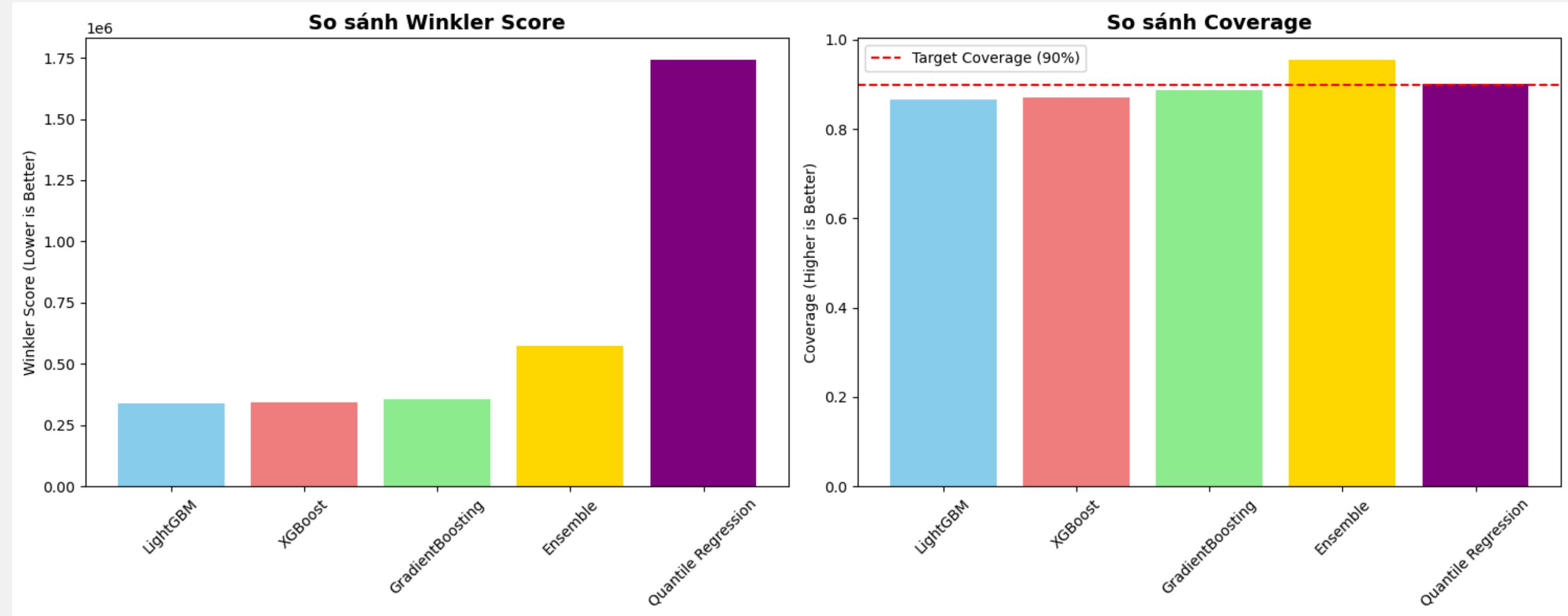
- Độ bao phủ: 95.6%
- Điểm Winkler: 575,624.33



# So sánh hiệu suất các mô hình

| Mô hình           | Độ bao phủ              | Điểm Winkler                 |
|-------------------|-------------------------|------------------------------|
| LightGBM          | 86.6%                   | <b>340,167.11</b> (tốt nhất) |
| XGBoost           | 87.0%                   | 342,390.46                   |
| Gradient Boosting | 88.7%                   | 357,562.68                   |
| Kết hợp mô hình   | <b>95.6%</b> (tốt nhất) | 575,624.33                   |
| Hồi quy phân vị   | 90.2%                   | 1,743,074.51                 |

# So sánh hiệu suất các mô hình



# Kết quả dự đoán trên Kaggle

| Prediction interval competition II: House price |                         |          |      |      |        | Submit Prediction | ...         |           |      |             |     |  |
|---|-------------------------|----------|------|------|--------|-------------------|-------------|-----------|------|-------------|-----|--|
|   |                         | Overview | Data | Code | Models | Discussion        | Leaderboard | Rules     | Team | Submissions |     |  |
| 179   | ThufailAhamedVJ         |          |      |      |        |                   |             | 346942.09 |      | 1           | 3d  |  |
| 180   | jade.lake8852           |          |      |      |        |                   |             | 347062.58 |      | 6           | 1mo |  |
| 181   | Tuệ Nguyễn              |          |      |      |        |                   |             | 347097.10 |      | 3           | 1d  |  |
| 182   | Jagadish Devu           |          |      |      |        |                   |             | 347612.15 |      | 2           | 2mo |  |
| 183   | Sanskar291202           |          |      |      |        |                   |             | 347851.02 |      | 5           | 13d |  |
| 184   | FIT-HCMUS-Hợp Hoan Tông |          |      |      |        |                   |             | 348228.33 |      | 2           | 3d  |  |

**Thank you  
very much!**