# PROJECT PROPOSAL

# Phân tích yếu tố tác động đến giá nhà tại Mỹ và dự báo giá nhà

## Mục đích và Kết quả mong đợi

### Mục đích: - Nghiên cứu dữ liệu bất động sản nhằm đánh giá các yếu tố ảnh hưởng đến giá nhà. - Xác định xu hướng và các yếu tố quan trọng tác động đến giá nhà. - Dự báo giá nhà từ dữ liệu đặc điểm ngôi nhà.

### Kết quả mong đợi: - Cung cấp insight quan trọng về mối quan hệ giữa các yếu tố (số phòng ngủ, diện tích, chất lượng, vị trí, v.v…) và giá nhà. - Đưa ra các đề xuất định giá, hỗ trợ cho quyết định mua-bán, đầu tư, môi giới và xây dựng.

## Nguồn dữ liệu Nguồn: Kaggle.com Link dataset: <https://www.kaggle.com/datasets/harlfoxem/housesalesprediction>

## Mô tả: Dữ liệu bất động sản bao gồm các đặc điểm của căn nhà, thông tin địa lý, tầm nhìn, kích thước và giá bán, … Tổng số: 21,613 dòng với 21 cột, bao gồm các thông tin:

|  |  |
| --- | --- |
| **Cột** | **Mô tả** |
| id | ID bất động sản |
| date | Ngày bán |
| price | Giá bán |
| bedrooms | Số phòng ngủ |
| bathrooms | Số phòng tắm |
| sqft\_living | Diện tích sử dụng |
| sqft\_lot | Diện tích đất |
| floors | Số tầng |
| waterfront | Có view mặt nước |
| view | Chất lượng view (0-4) |
| condition | Tình trạng nhà |
| grade | Chất lượng xây dựng (1-13) |
| sqft\_above | Diện tích trên mặt đất |
| sqft\_basement | Diện tích tầng hầm |
| yr\_built | Năm xây |
| yr\_renovated | Năm cải tạo |
| zipcode | Mã vùng |
| lat | Vĩ độ |
| long | Kinh độ |
| sqft\_living15 | Diện tích trung bình vùng |
| sqft\_lot15 | Diện tích đất trung bình vùng |

**Kế hoạch phân tích dữ liệu**

1️⃣ Làm sạch dữ liệu (Data Cleaning):

Xử lý các giá trị bị thiếu (missing values).

Chuyển đổi kiểu dữ liệu phù hợp (ví dụ: bathrooms → float, date → datetime).

Chuẩn hóa định dạng ngày (date, yr\_built) để phục vụ phân tích thời gian.

Loại bỏ các dòng dữ liệu bị trùng lặp hoặc sai sót.

2️⃣ Phân tích dữ liệu khám phá (EDA):

Tính toán các thống kê mô tả cho các biến định lượng (trung bình, min, max, phân vị...).

Phân tích xu hướng giá nhà theo thời gian (năm xây dựng, năm cải tạo).

Kiểm tra phân phối giá nhà theo các đặc điểm như:  
• Diện tích sử dụng (sqft\_living)  
• Diện tích đất (sqft\_lot)  
• Số phòng ngủ, phòng tắm (bedrooms, bathrooms)  
• Chất lượng nhà (grade, condition)  
• Vị trí đặc biệt (waterfront, view)

3️⃣ Phân tích mô hình:

Xây dựng mô hình hồi quy tuyến tính đa biến, Xgboost,… để dự báo giá nhà (price).

Xác định các biến độc lập quan trọng ảnh hưởng đến giá.

Đánh giá mô hình bằng các chỉ số:  
• R² – Đo mức độ giải thích của mô hình  
• p-value – Đo độ tin cậy của từng biến  
• RMSE – Sai số dự báo trung bình

So sánh độ chính xác giữa mô hình huấn luyện và kiểm tra.

4️⃣ Trực quan hóa dữ liệu (Visualization):

Biểu đồ cột: So sánh giá nhà theo số phòng, chất lượng, khu vực (zipcode).

Biểu đồ đường (line): Thể hiện xu hướng giá theo thời gian.

Bản đồ (map): Phân tích vị trí địa lý bằng lat, long

Biểu đồ heatmap: Trực quan hóa tương quan giữa các biến đầu vào và giá (price).

5️⃣ Kể chuyện bằng dữ liệu (Data Storytelling):

Tóm tắt những phát hiện chính từ phân tích EDA và mô hình hồi quy.

Nêu bật các yếu tố tác động mạnh nhất đến giá nhà (ví dụ: diện tích, chất lượng, vị trí).

Đưa ra các đề xuất chiến lược thực tế cho:  
• Người mua (chọn khu vực, đặc điểm hợp lý)  
• Người bán (biết yếu tố tăng giá để cải tạo)  
• Nhà đầu tư (dự báo giá nhà, phân vùng sinh lời)

Gắn liền trực quan hóa với insight để tăng tính thuyết phục khi trình bày.

## Lộ trình thực hiện 1. Thu thập và làm sạch dữ liệu bằng Python. 2. Phân tích dữ liệu khám phá (EDA) và trực quan hóa bằng Python 3. Xây dựng mô hình hồi quy tuyến tính đa biến. 4. Phân tích thống kê và rút ra insight. 5. Hoàn thiện báo cáo cuối cùng và chuẩn bị thuyết trình.

## Thuyết trình - Trình bày kết quả phân tích và các insight quan trọng từ dữ liệu. - Đưa ra đề xuất giúp tối ưu hóa giá trị bất động sản và hỗ trợ quyết định kinh doanh.